DESMISTIFICANDO A CIÊNCIA DE DADOS

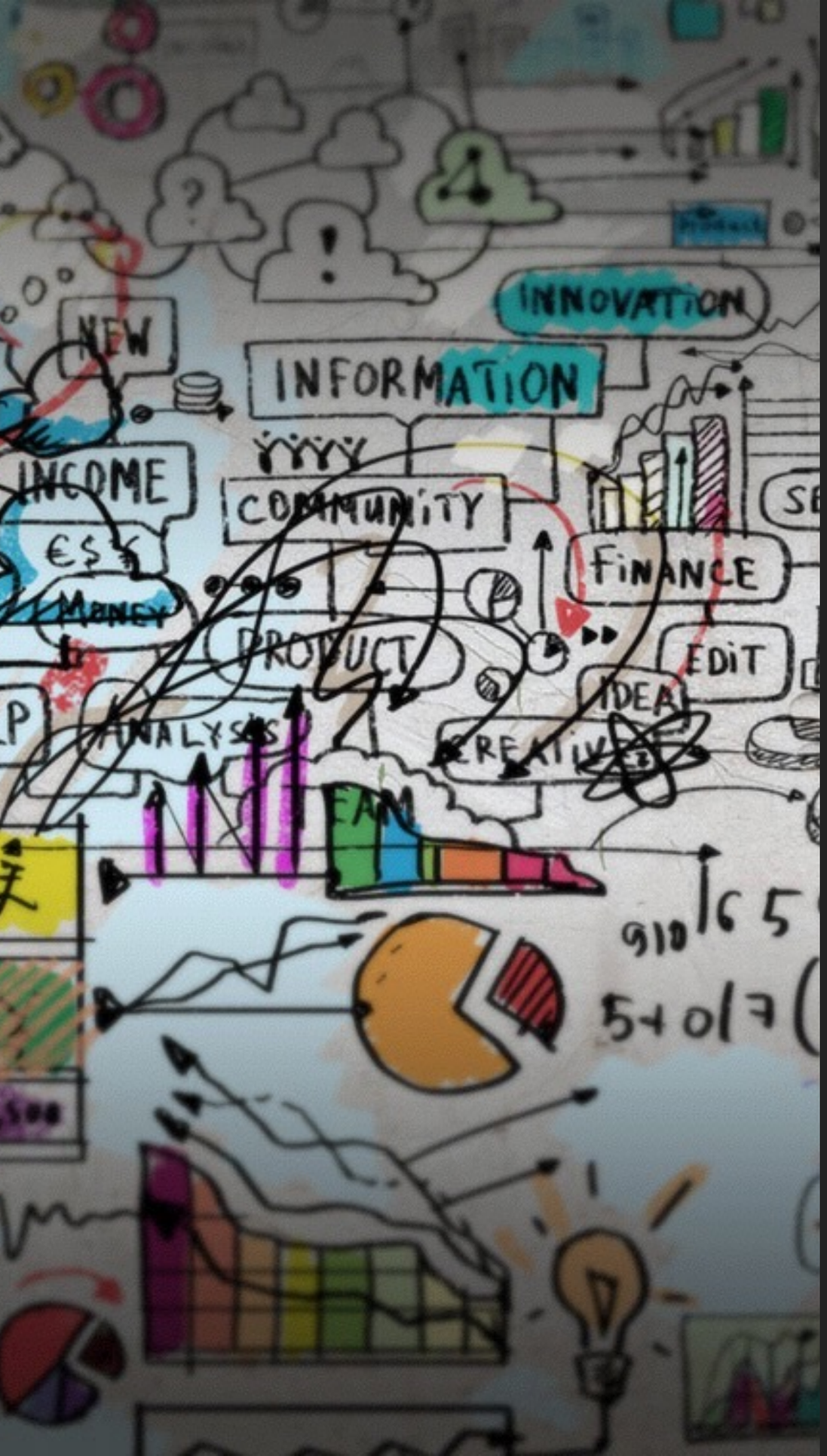
NAUBER GOIS

- Aluno de Doutorado em Informática Aplicada da Unifor
  - Área de Estudo: Reinforcement Learning e metaheurísticas aplicadas a testes de performance.
- Analista de Desenvolvimento do Serpro

AGENDA

- DEFINIÇÃO DE DATA SCIENCE
- TIPOS DE MODELOS
- EXEMPLOS DE APLICAÇÃO
- RELATÓRIOS
- TESTES A/B

O QUE É CIÊNCIA DE DADOS?



DATA SCIENCE É MAIS UM TERMO USADO PARA DESCREVER O PROCESSO DE TRANSFORMAÇÃO DE DADOS EM CONHECIMENTO.

(LOUKIDES, 2016)

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

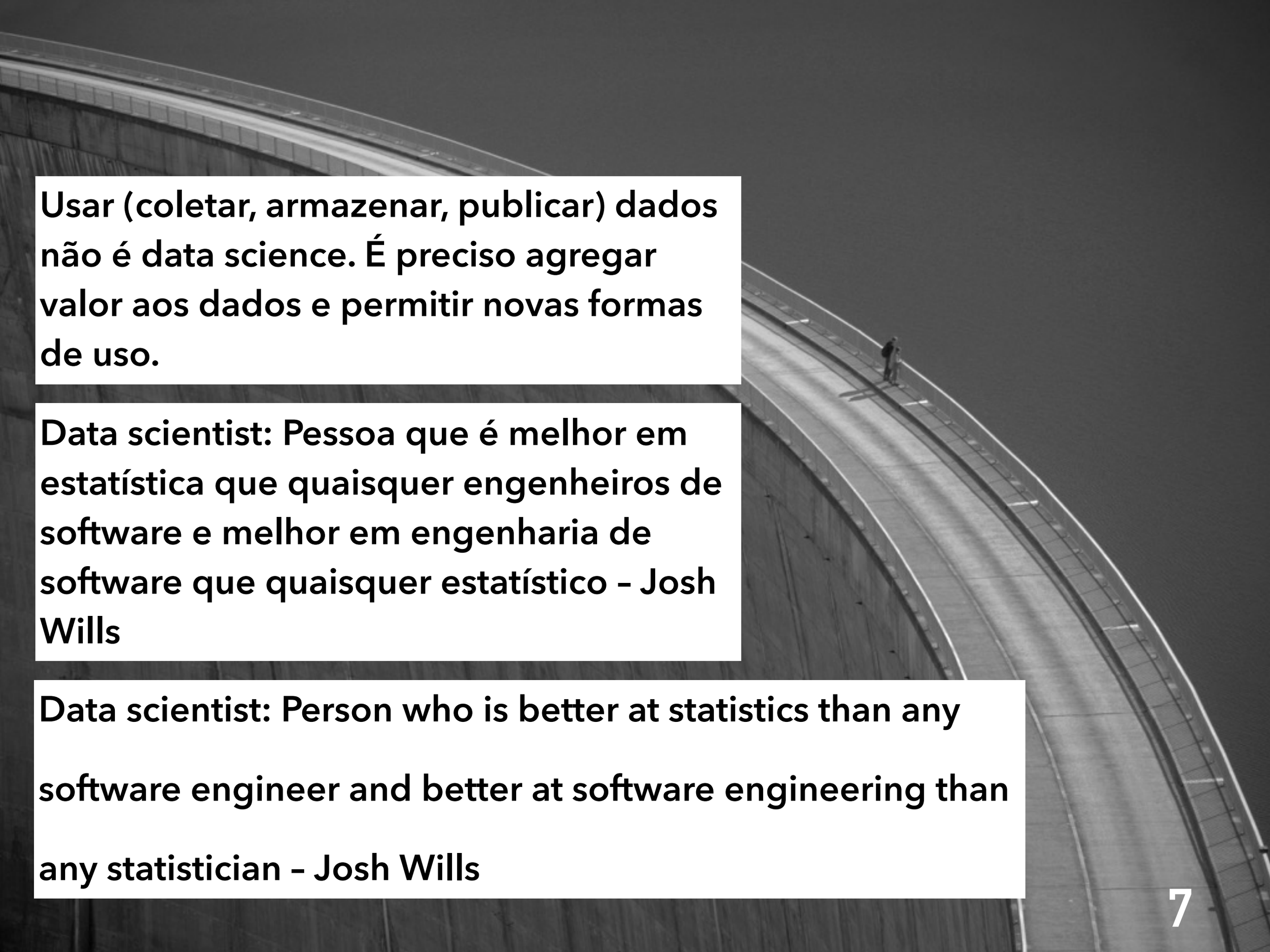
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

CIENTISTA DE DADOS

- Matemática e estatística
- Banco de Dados e Programação
- Conhecimento de Negócio
- Comunicação



Usar (coletar, armazenar, publicar) dados não é data science. É preciso agregar valor aos dados e permitir novas formas de uso.

Data scientist: Pessoa que é melhor em estatística que quaisquer engenheiros de software e melhor em engenharia de software que quaisquer estatístico – Josh Wills

Data scientist: Person who is better at statistics than any software engineer and better at software engineering than any statistician – Josh Wills

TYPES OF ANALYTICS

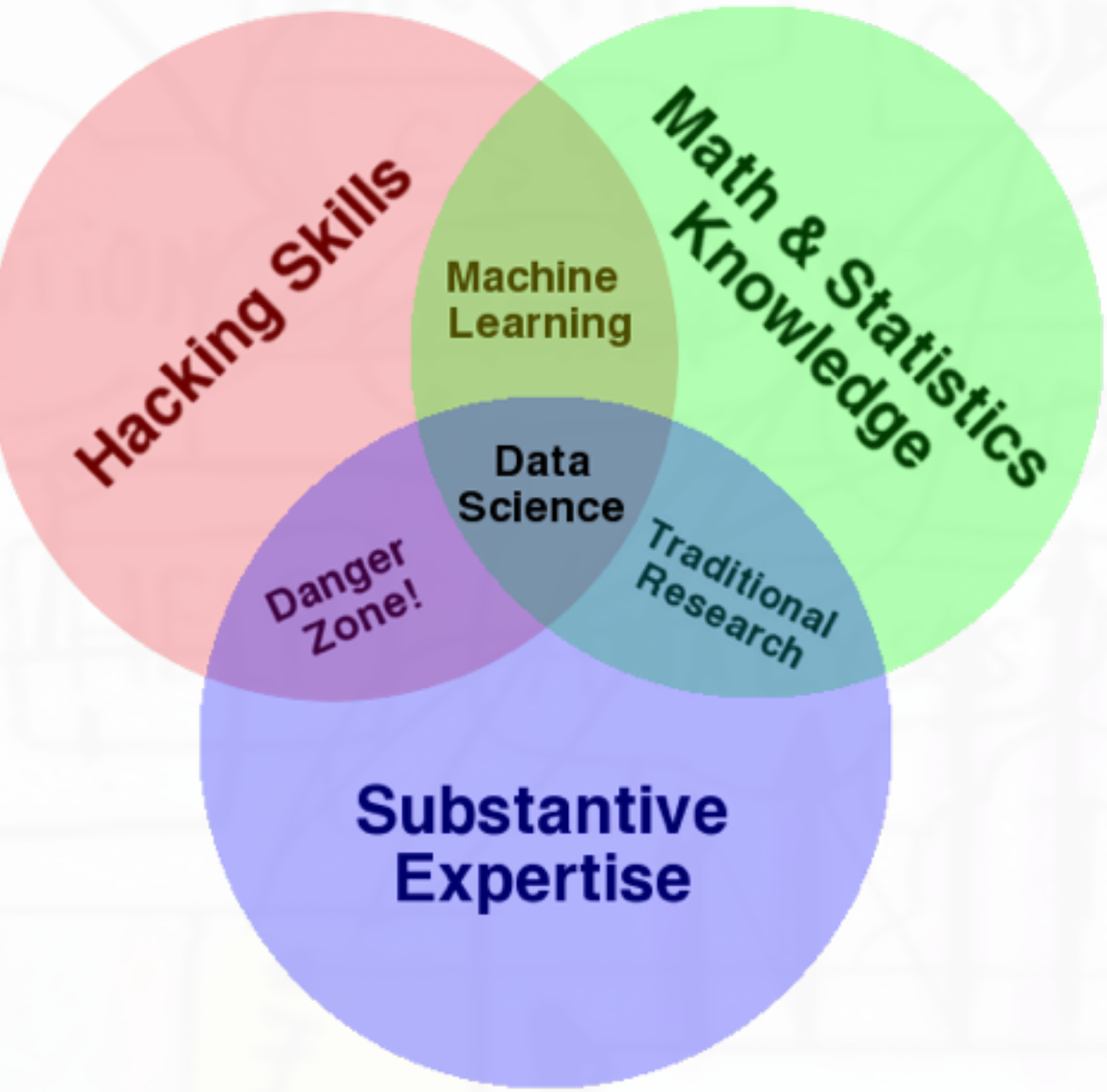
Investigative Analytics

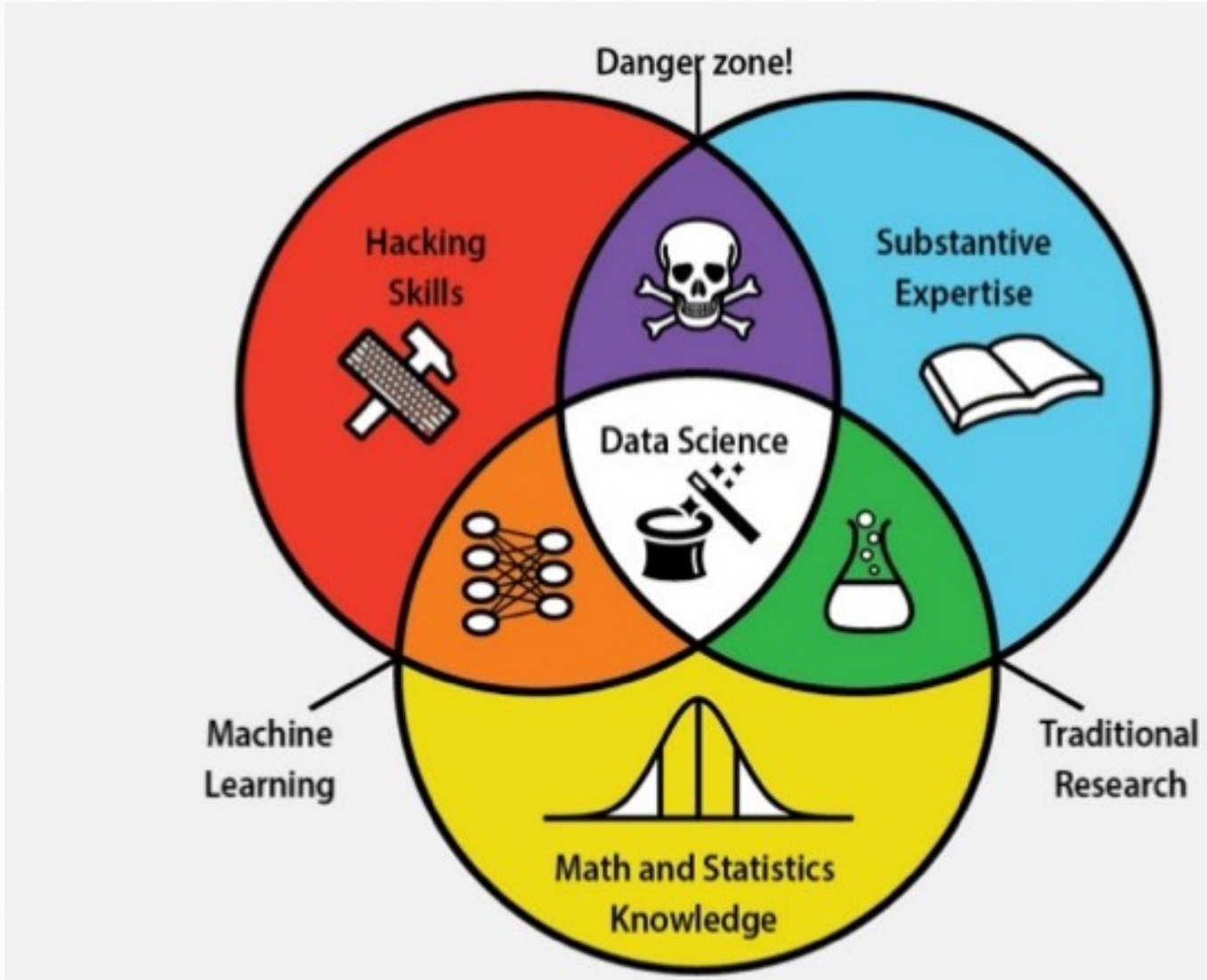
Consumers: Humans

Operational Analytics

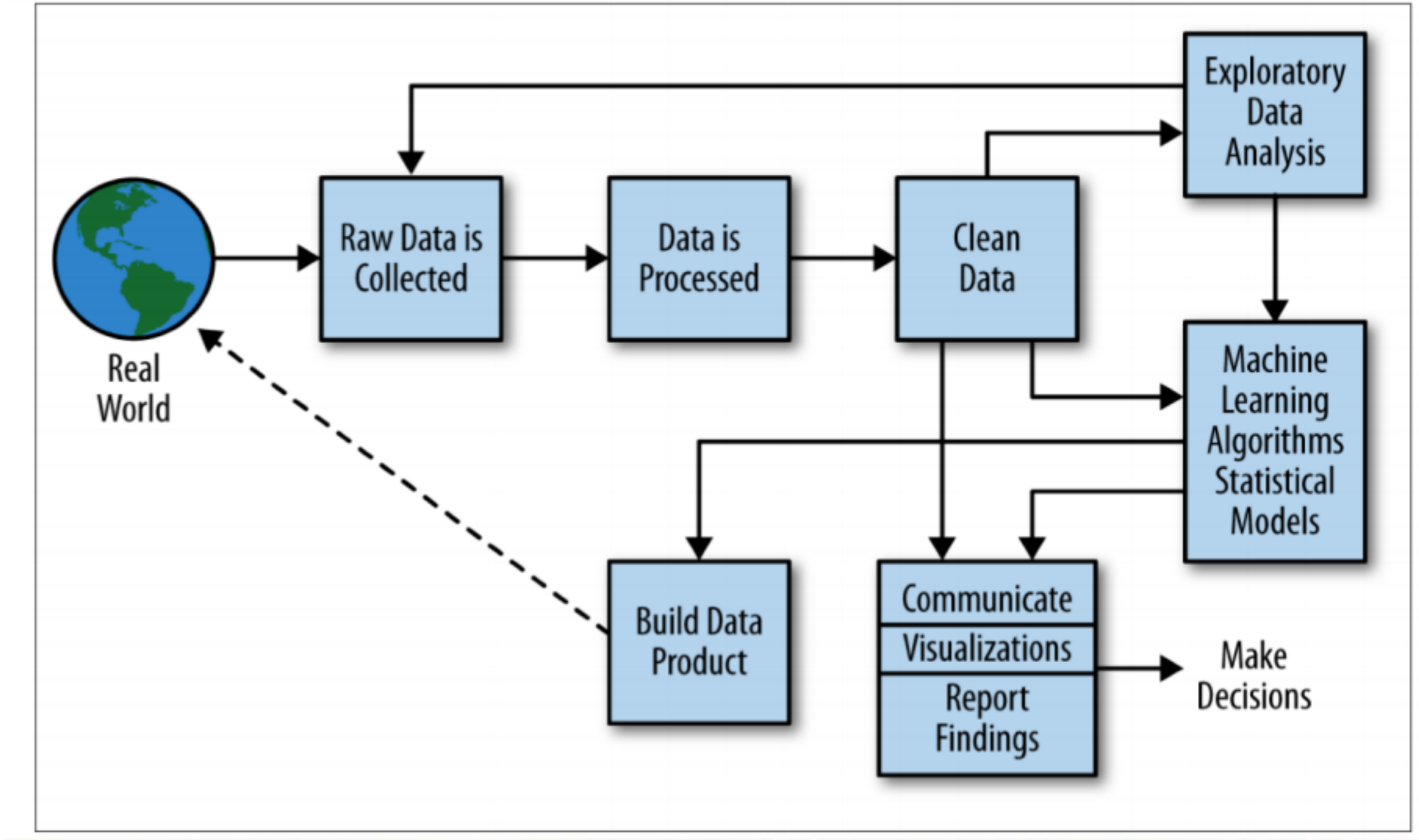
Consumers: Machines

DATA SCIENCE VENN DIAGRAM

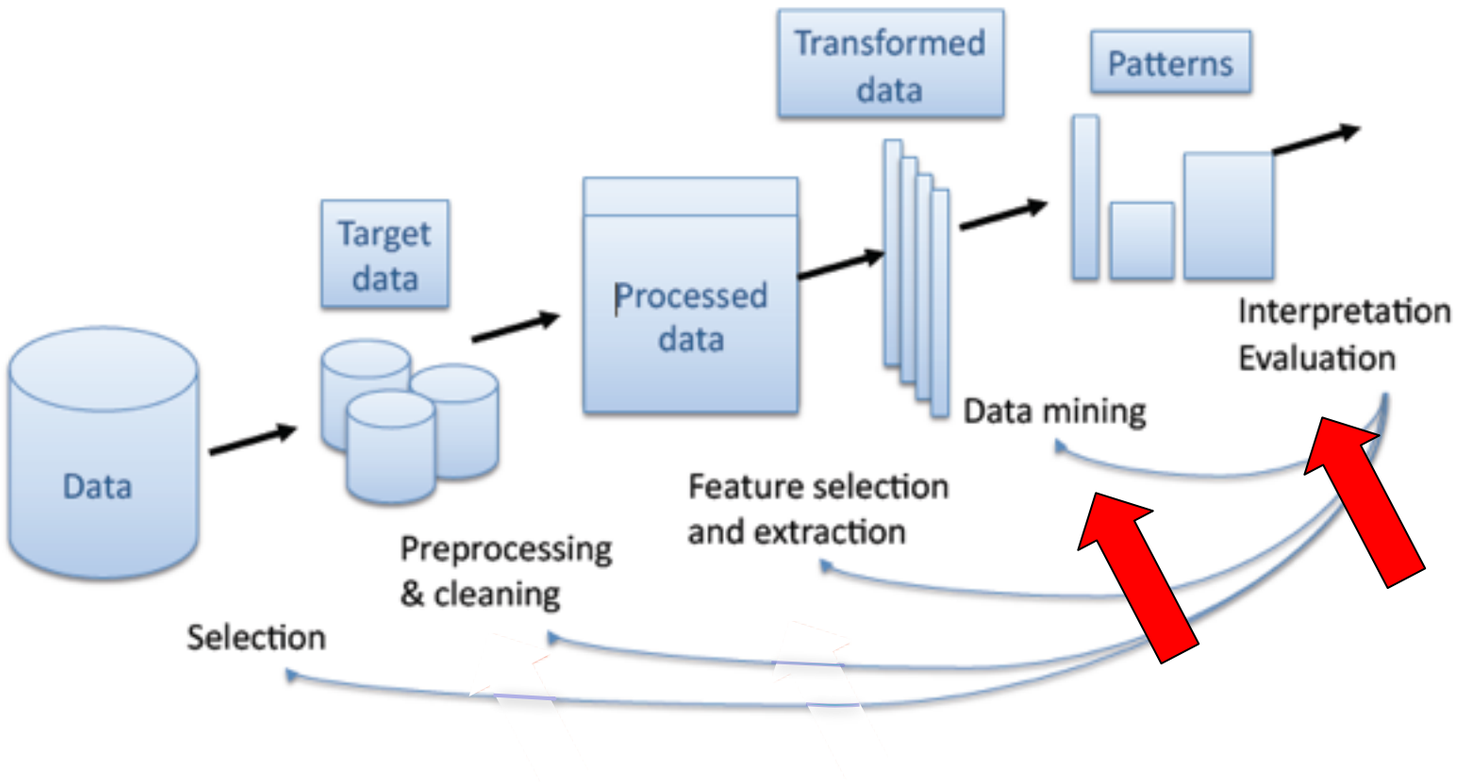




É um processo (?)

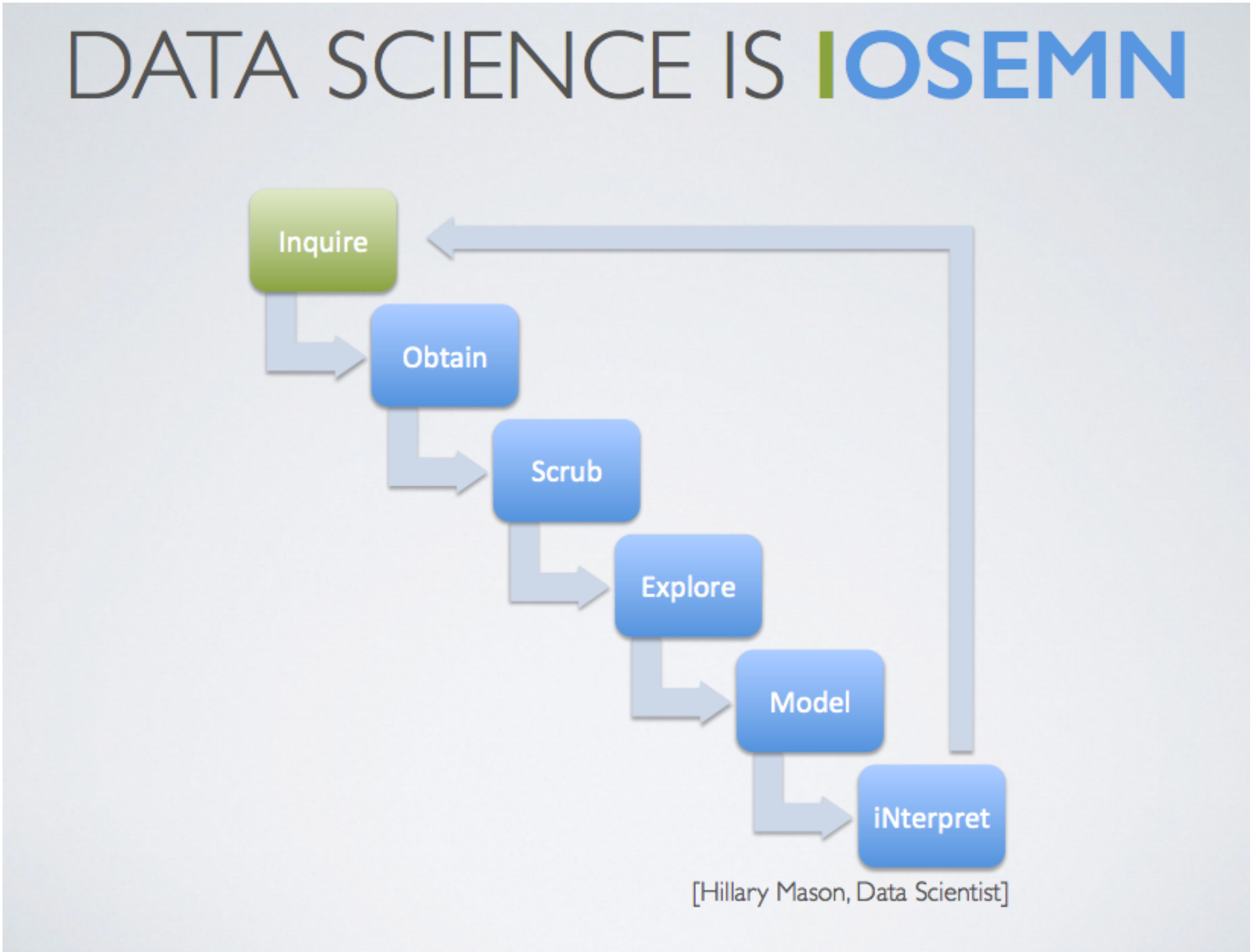


Seleção de Features e Redução de Dimensionalidade



U.M.Fayyad, G.Patetsky-Shapiro and P.Smyth (1995)

PROCESSO?

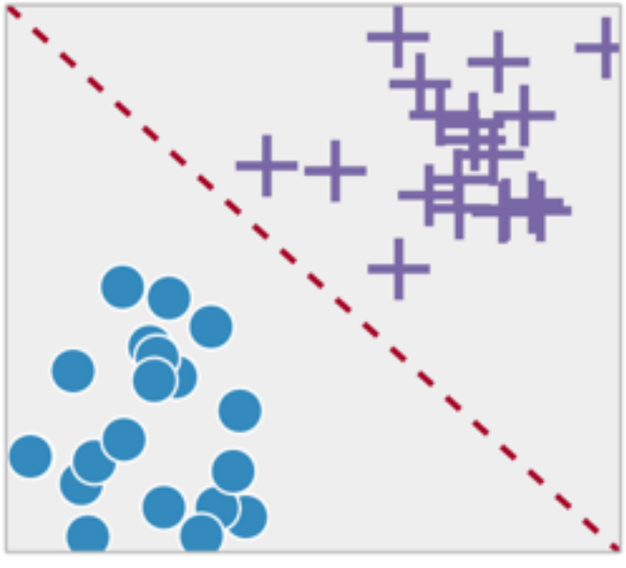

DATA SCIENCE IS IOSEMN
Inquire
Obtain
Scrub
Explore
Model
iNterpret
[Hillary Mason, Data Scientist]

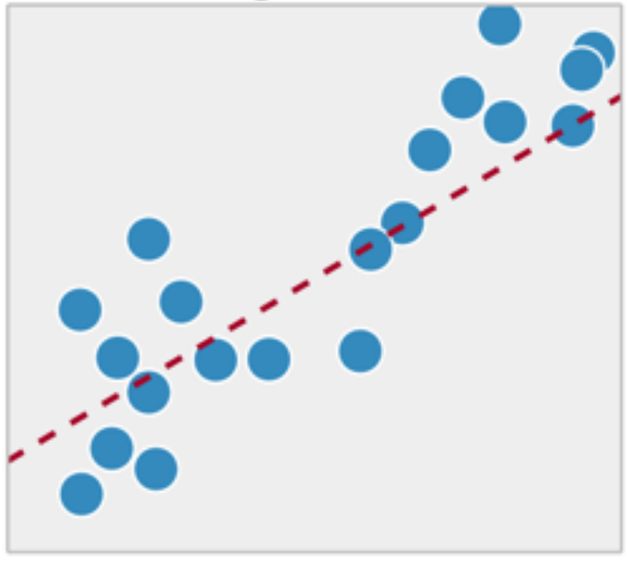
COLETAR E PROCESSAR OS DADOS

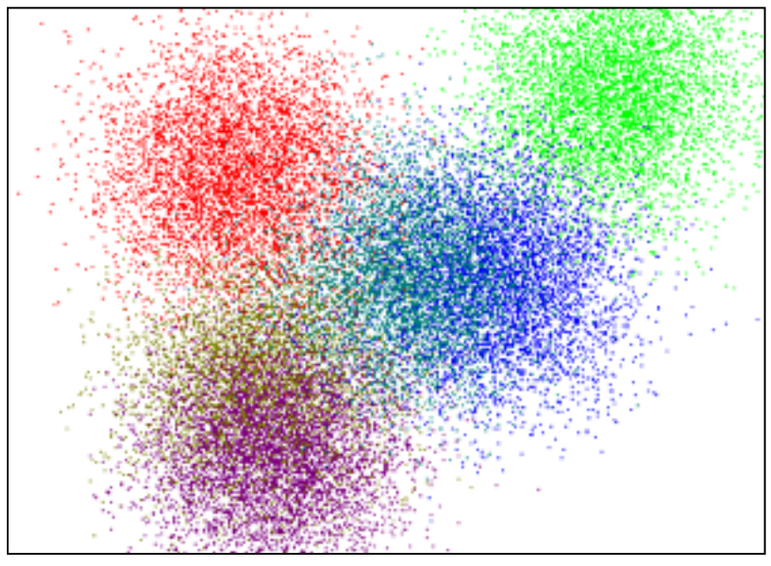
- Conduzir experimento de pesquisa.
- Coletar amostras de uma população.
- Transformar , filtrar e sumarizar os dados.
- Preparar os dados para o modelo escolhido

FORMULAÇÃO DE UM PROBLEMA

Identificação de uma área de interesse e o tipo de modelo



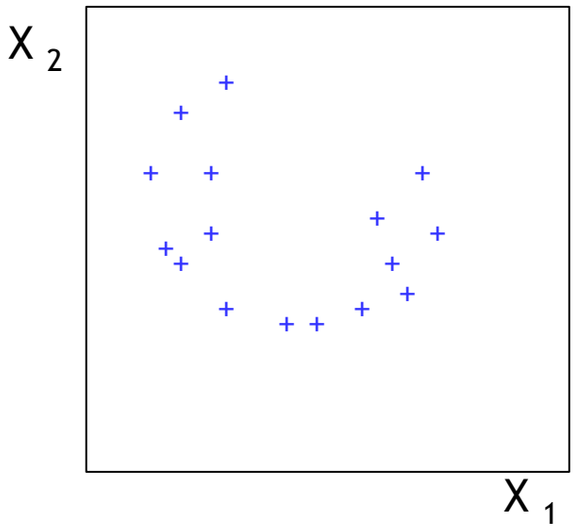




ESCOLHA DO TIPO DE MODELO A SER APLICADO

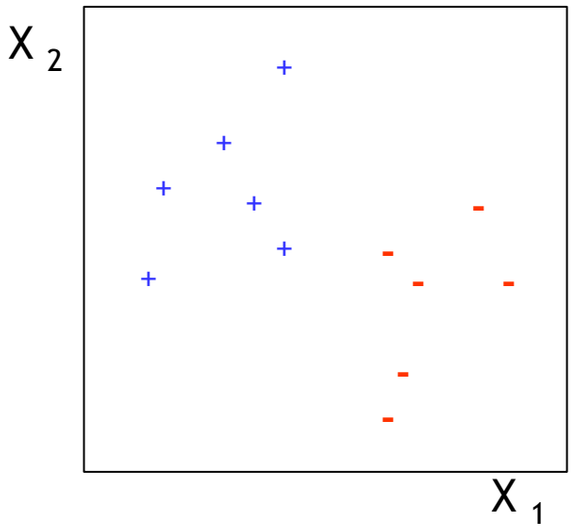
Common Data Mining tasks

Clustering



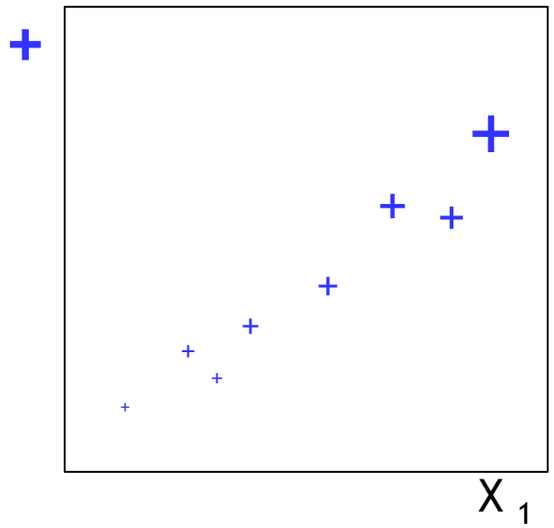
- k-th Nearest Neighbour
- Parzen Window
- Unfolding, Conjoint Analysis, Cat-PCA

Classificação



- Linear Discriminant Analysis, QDA
- Logistic Regression (Logit)
- Decision Trees, LSSVM, NN, VS

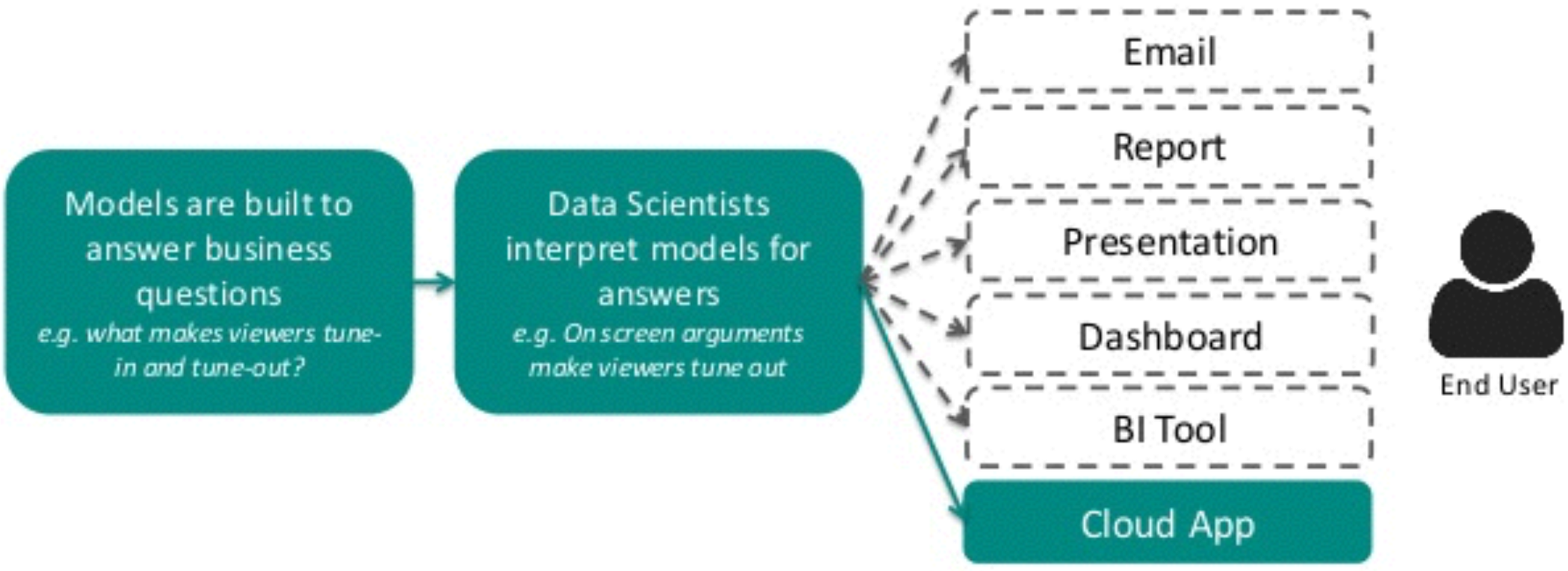
Regressão



- Classical Linear Regression
- Ridge Regression
- NN, CART

Models → Insights → Actions

A good insight drives action that will generate value for stakeholders



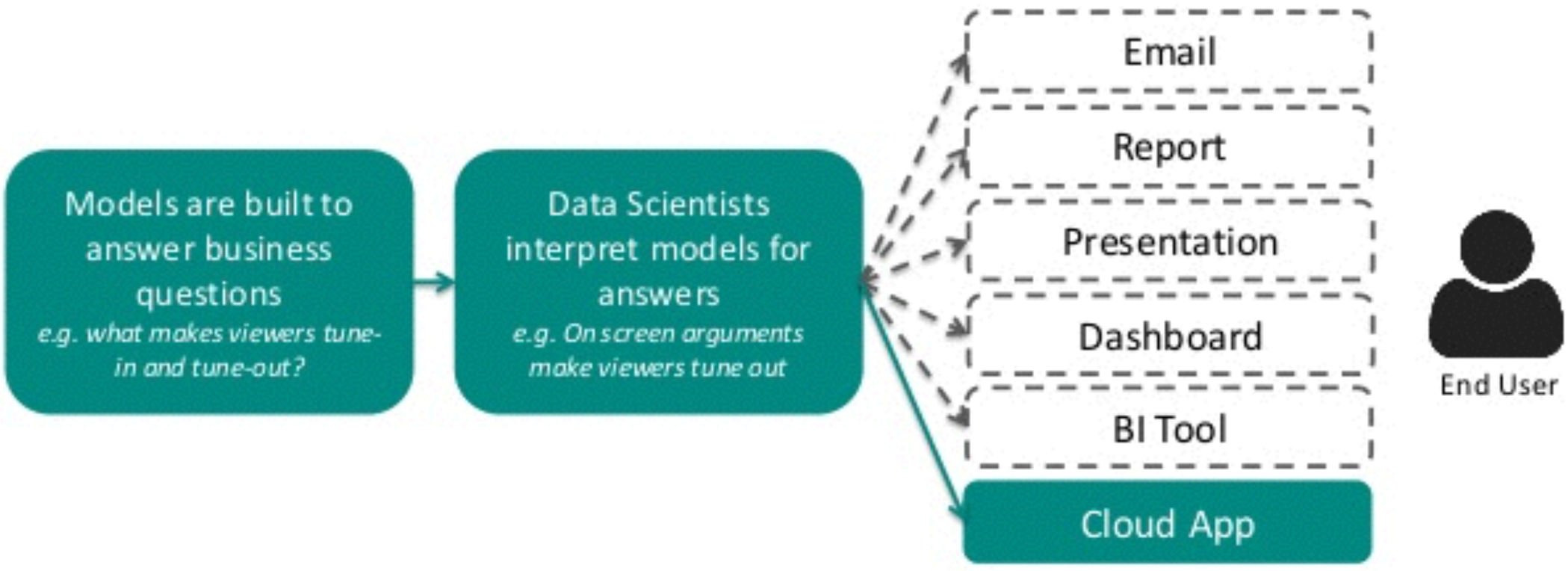
NAB SHOW

Pivotal

etc

Models → Insights → Actions

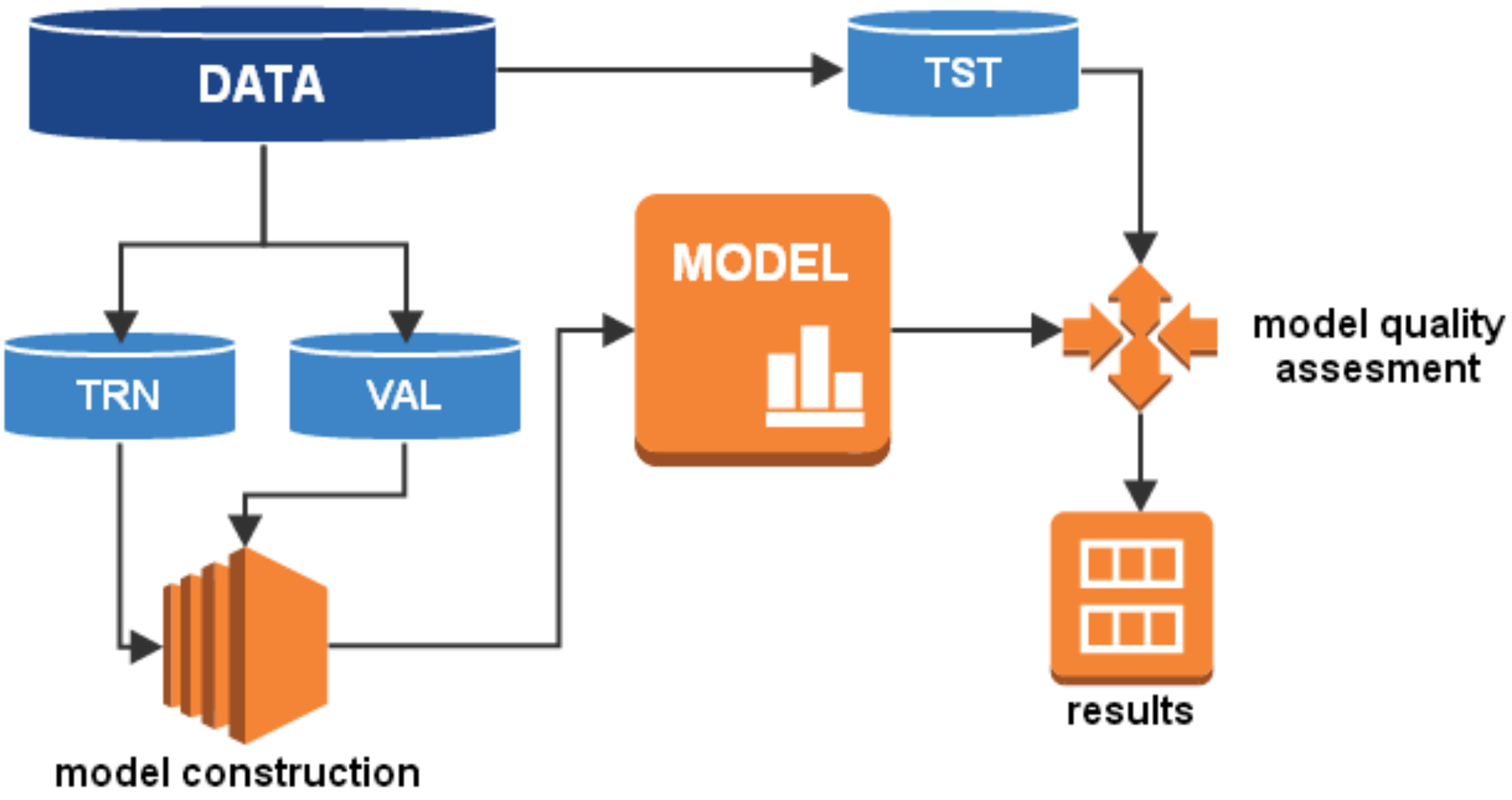
A good insight drives action that will generate value for stakeholders

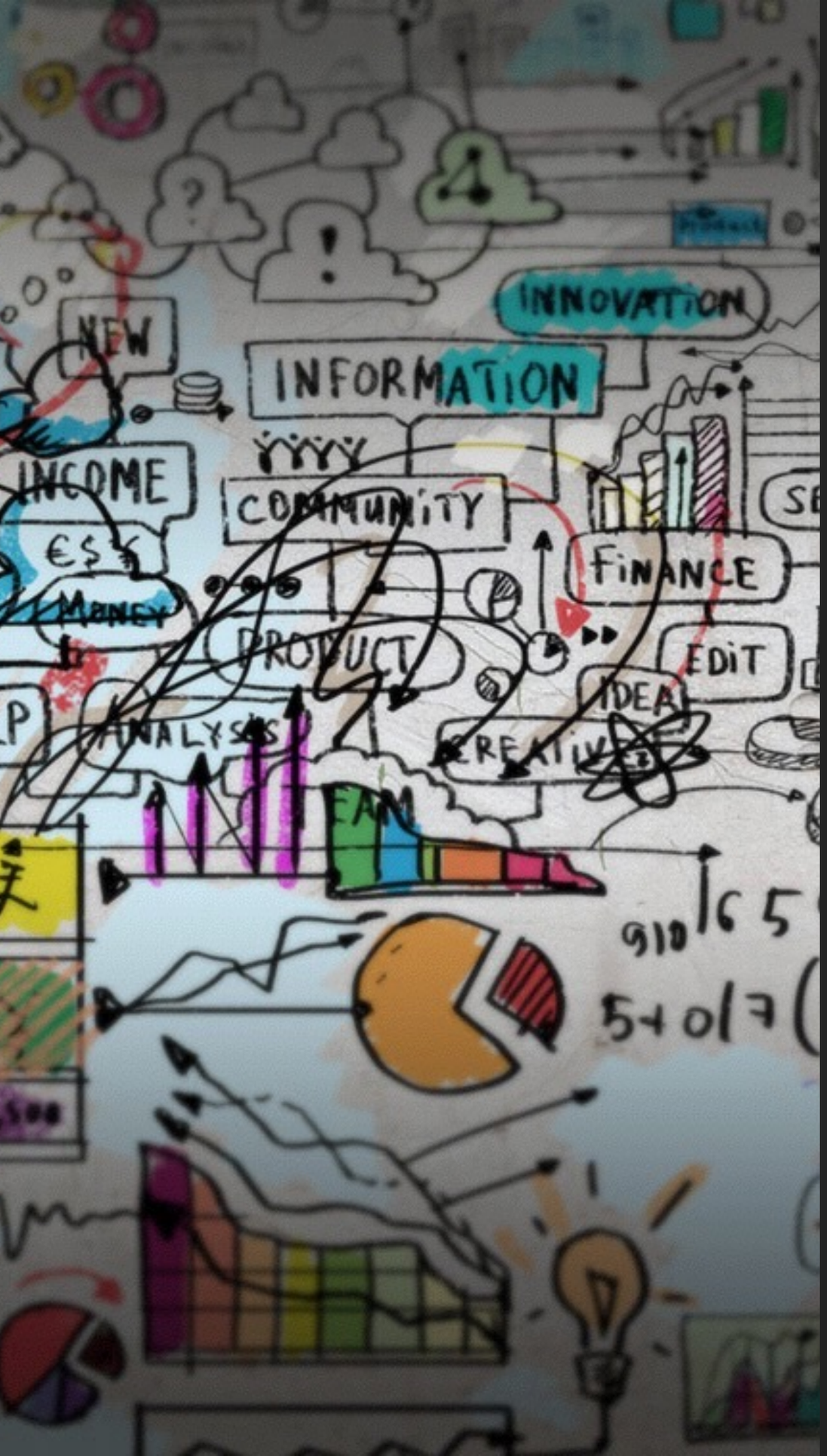


NAB SHOW

Pivotal

etc


DATA
TST
TRN
VAL
MODEL
model quality
assesment
model construction
results

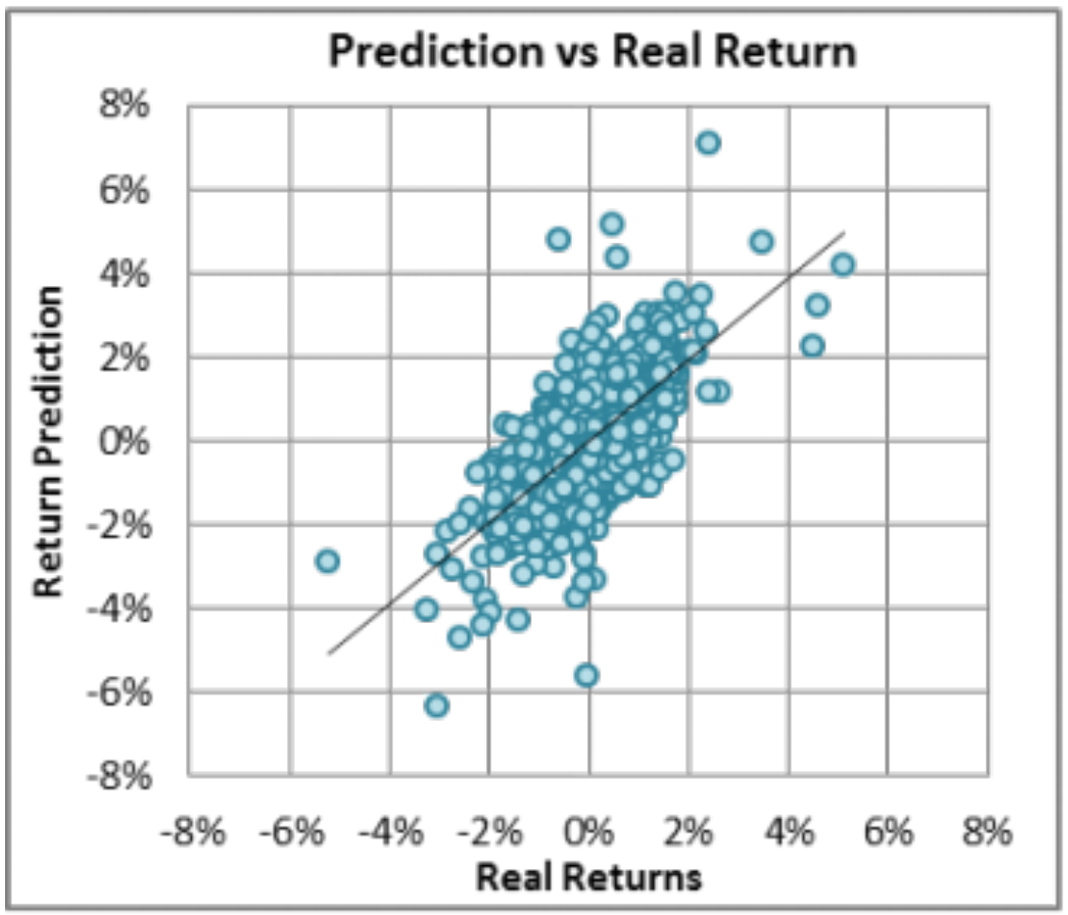


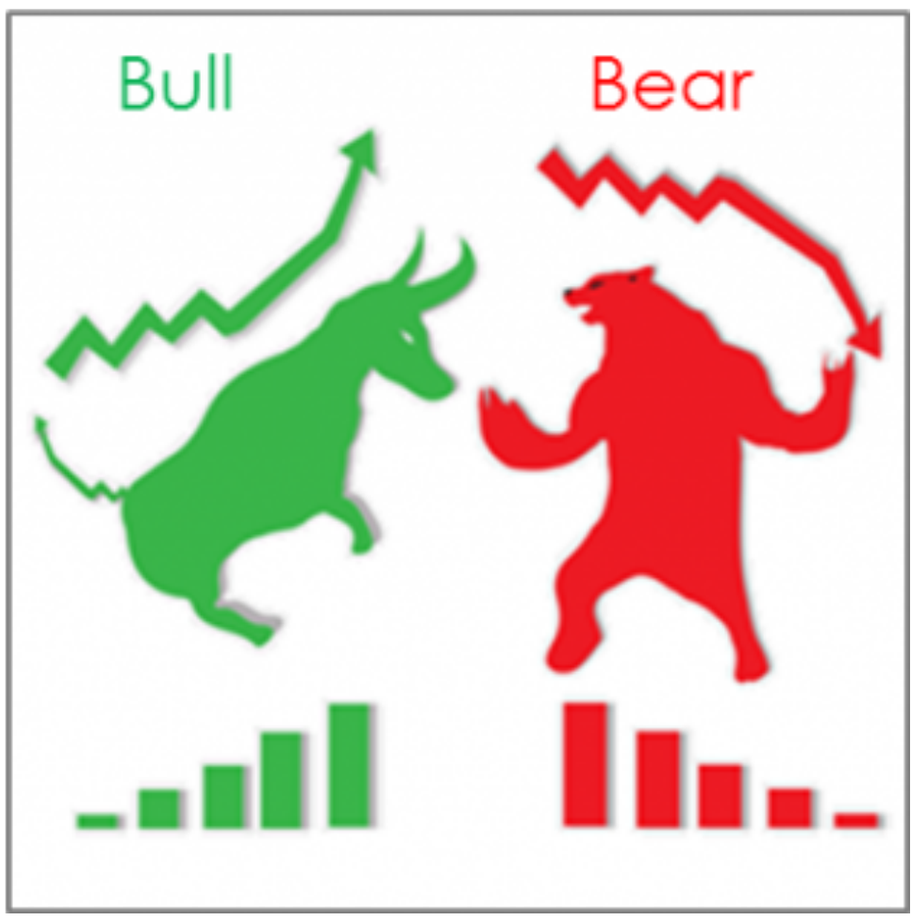
TIPOS DE MODELOS

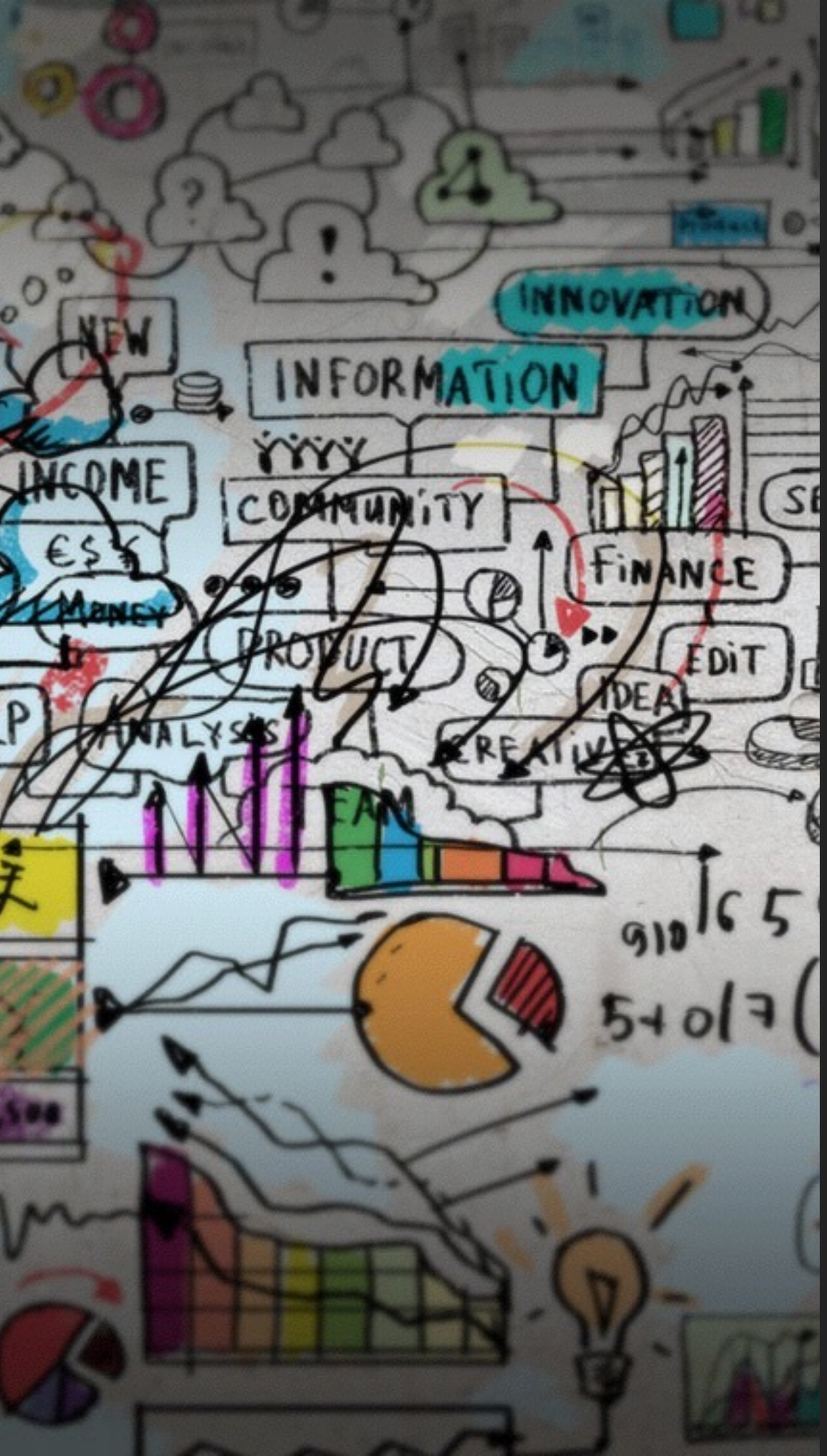
- REGRESSÃO
- CLASSIFICAÇÃO
- AGRUPAMENTO

CLASSIFICAÇÃO VS REGRESSÃO

Regression vs Classification



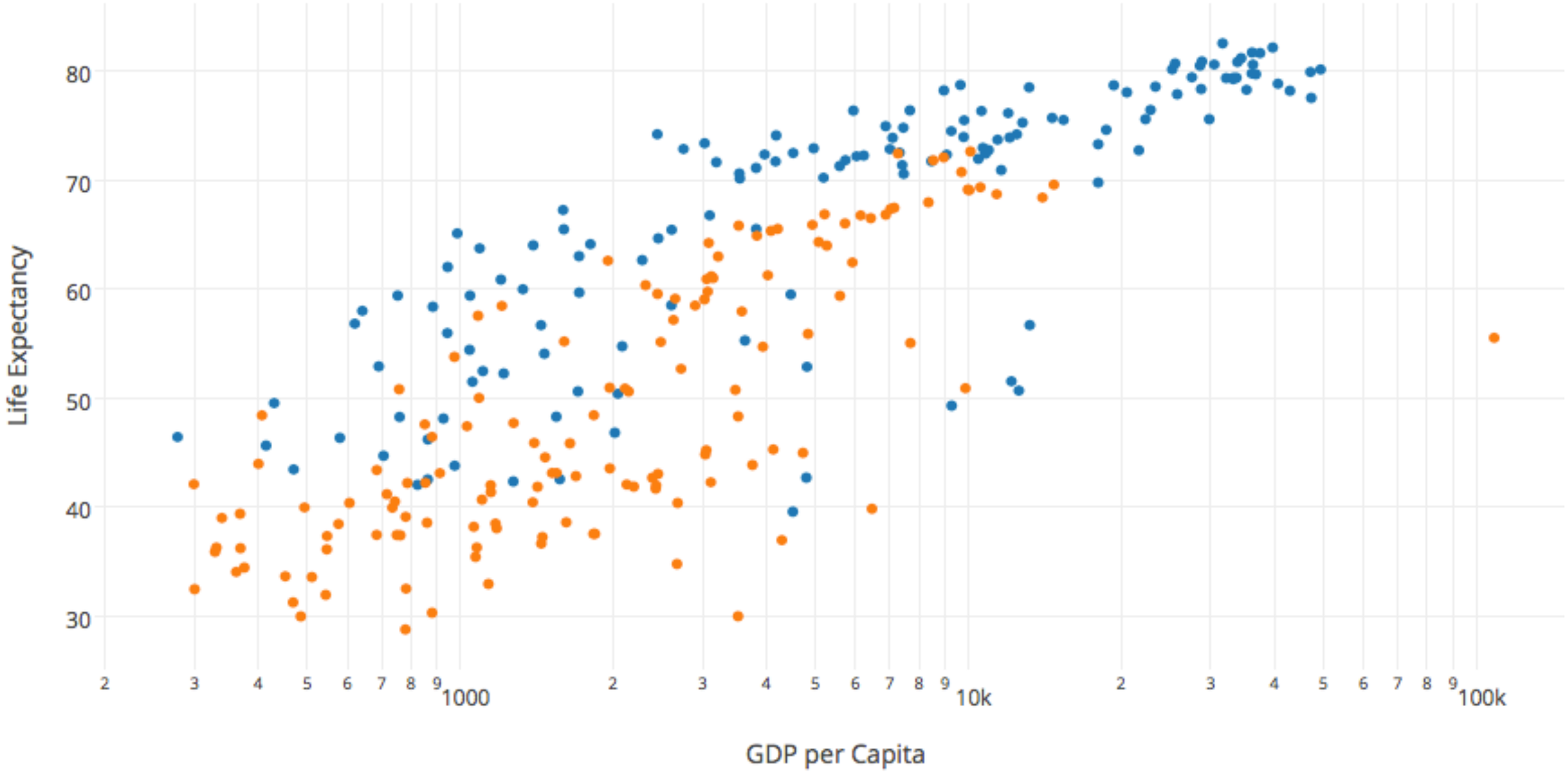




O QUE É REGRESSÃO ?

REGRESSÃO

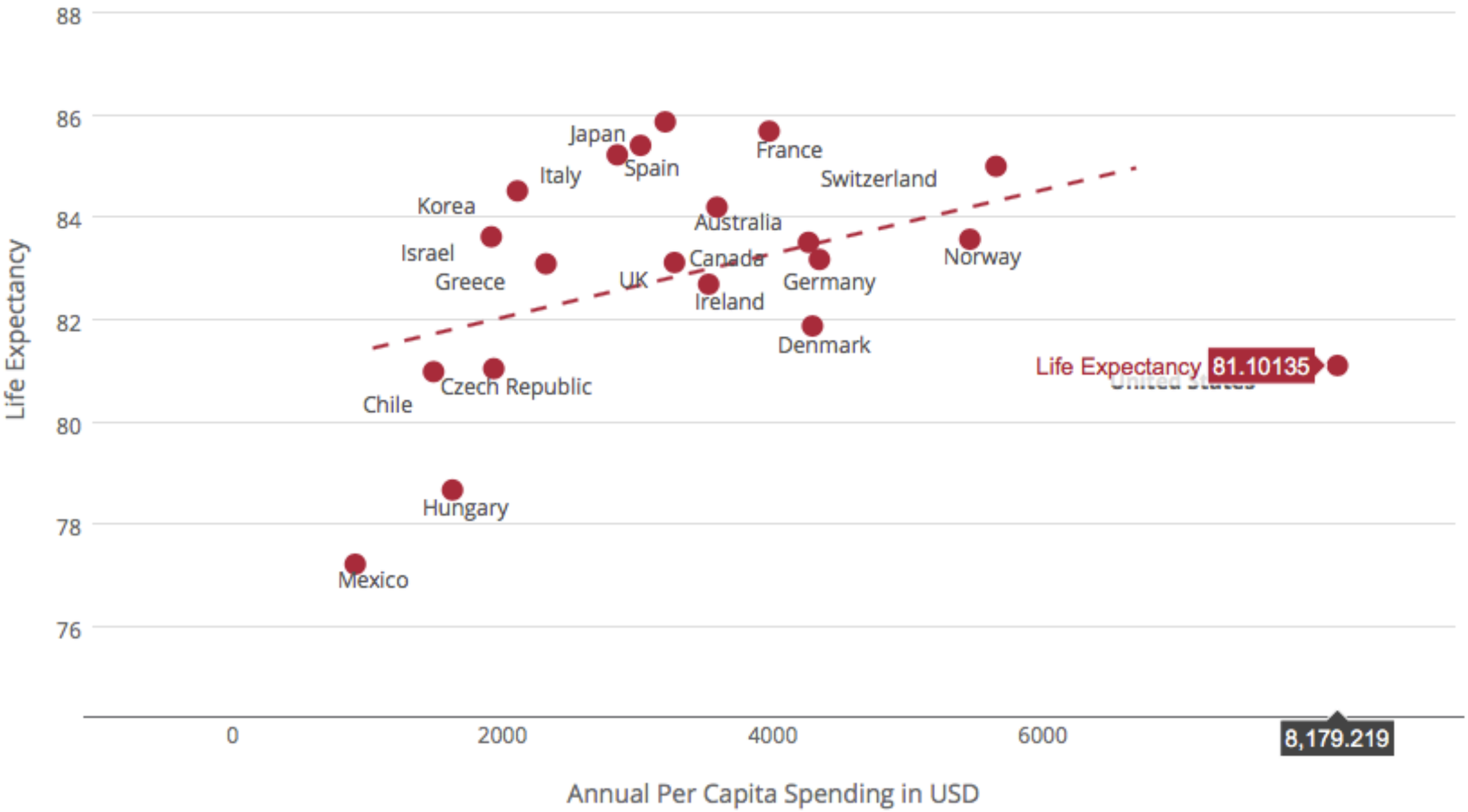
https://plot.ly/pandas/line-and-scatter/



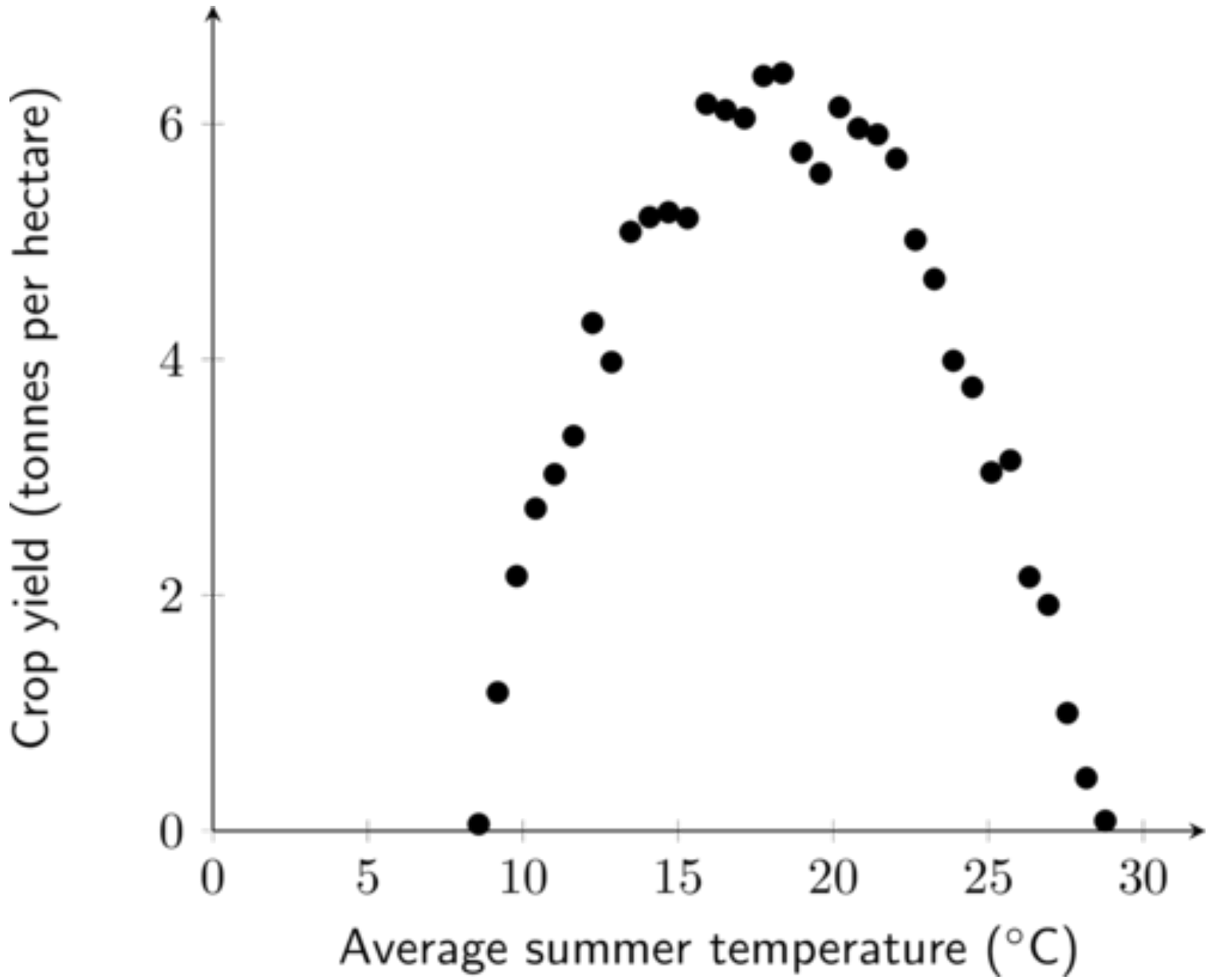
REGRESSÃO

https://plot.ly

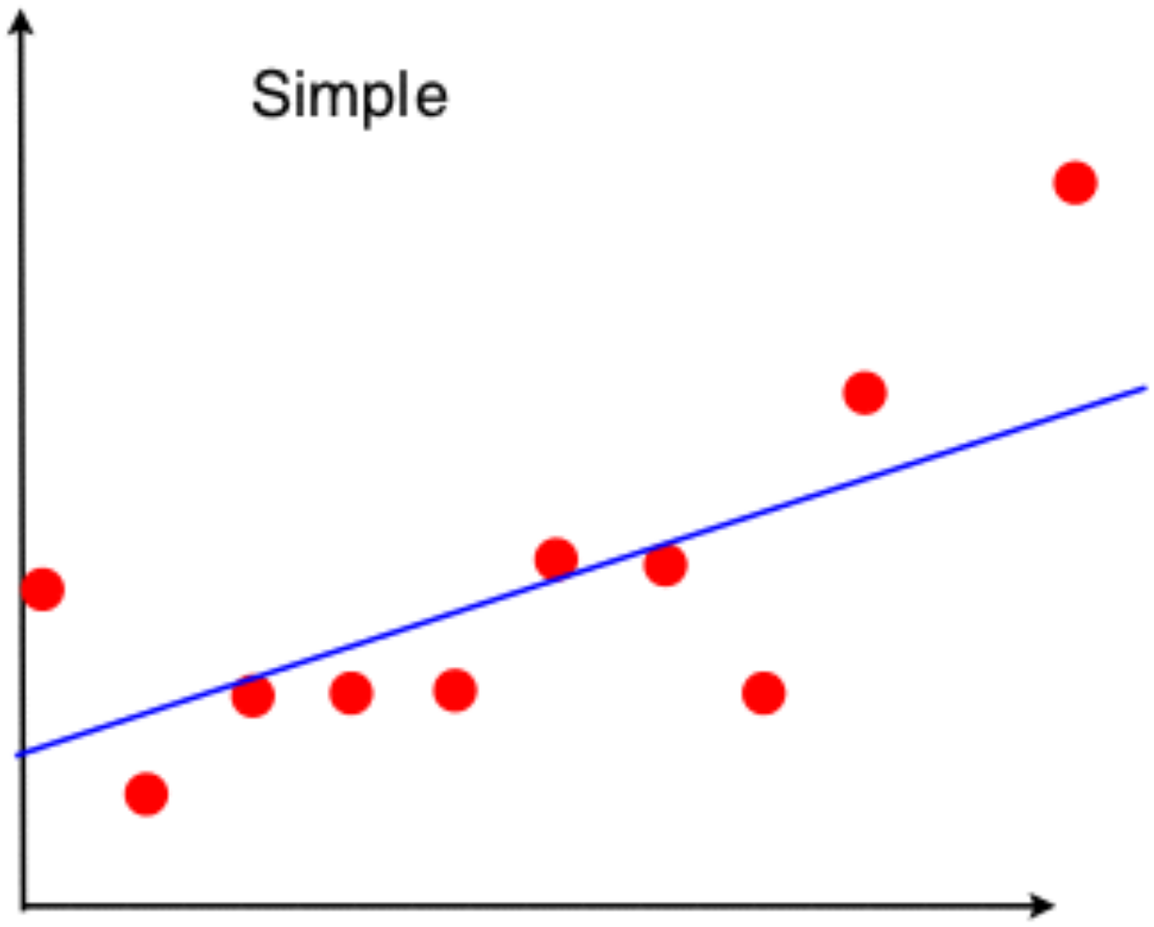
```
('Coefficients: \n', array([ 0.00118801]))
Mean squared error: 9.71
Variance score: -2.38
```

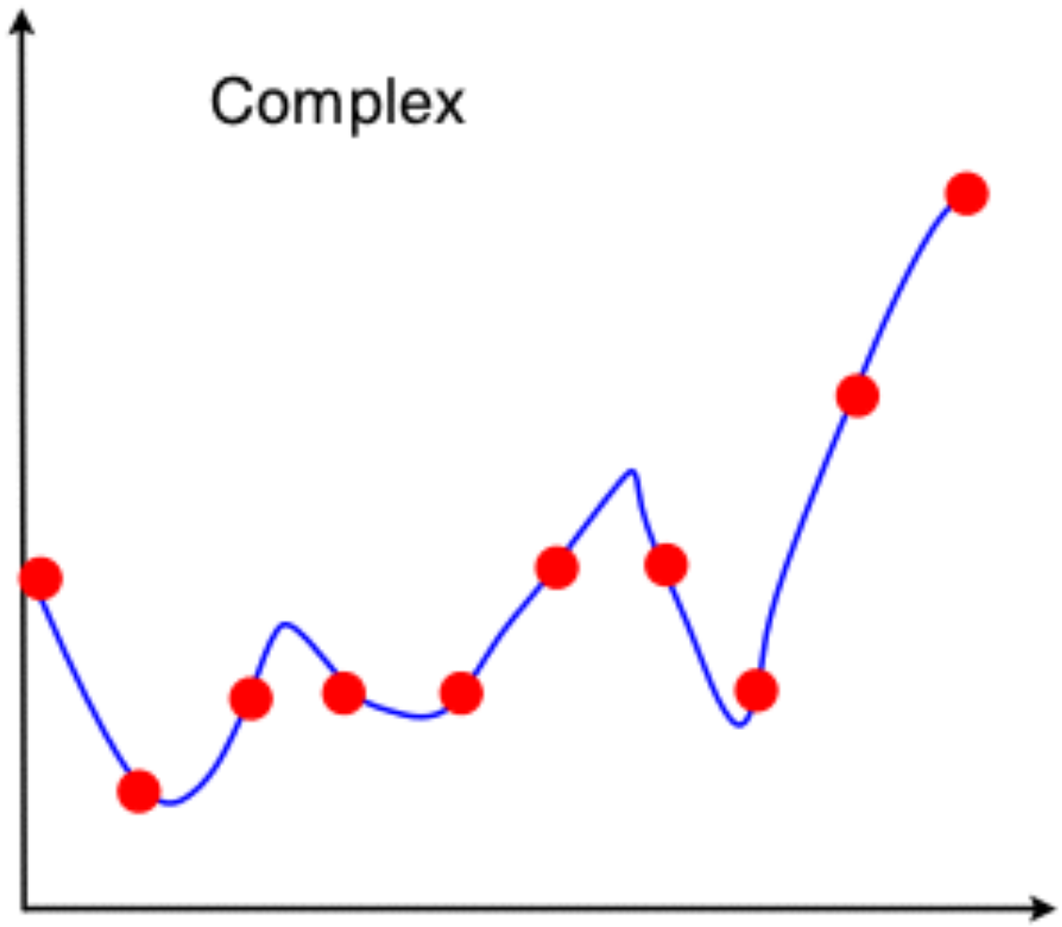


REGRESSÃO

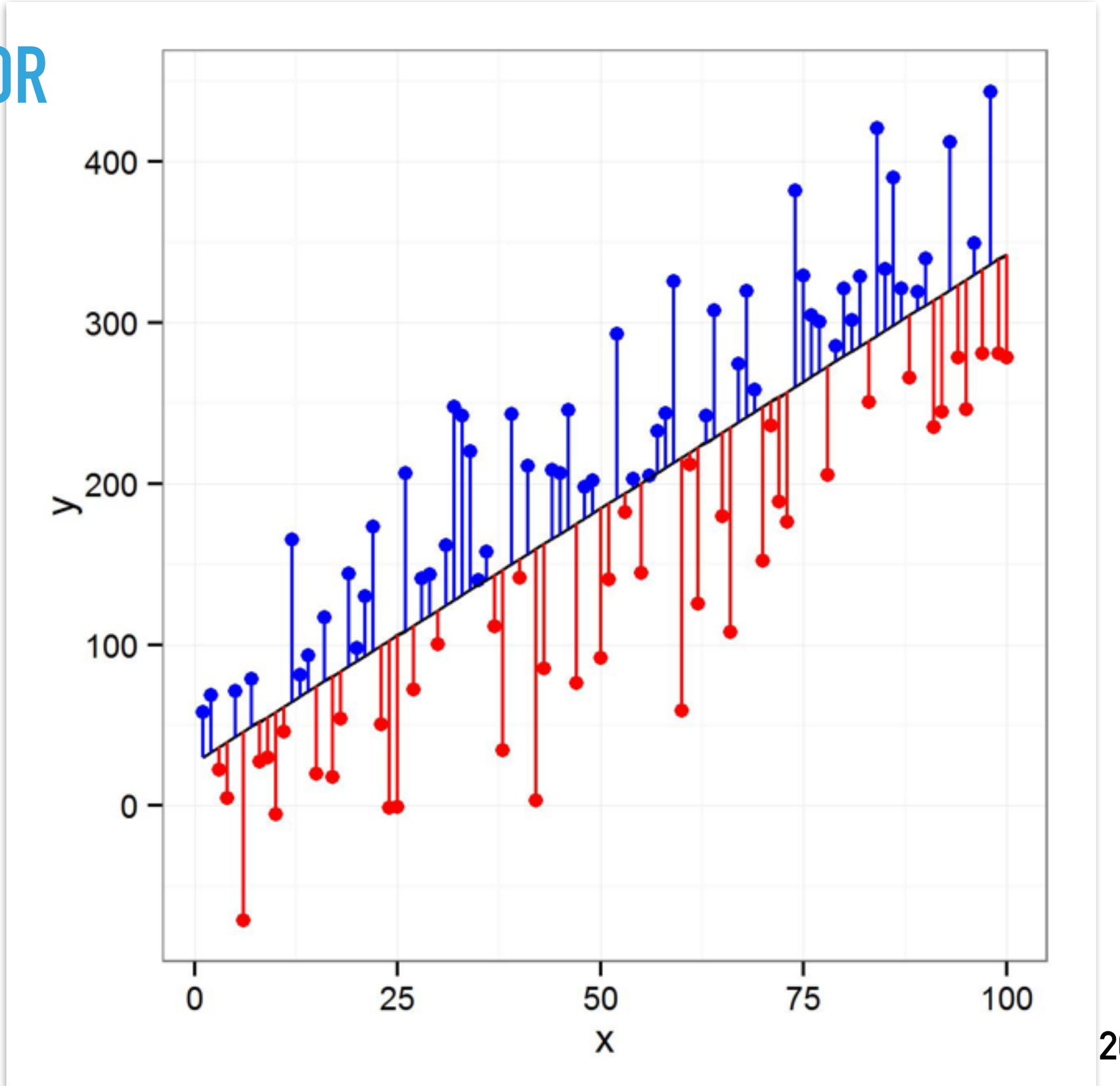

Crop yield (tonnes per hectare)
0
2
4
6
0
5
10
15
20
25
30
Average summer temperature (°C)

QUAL A MELHOR PREDIÇÃO

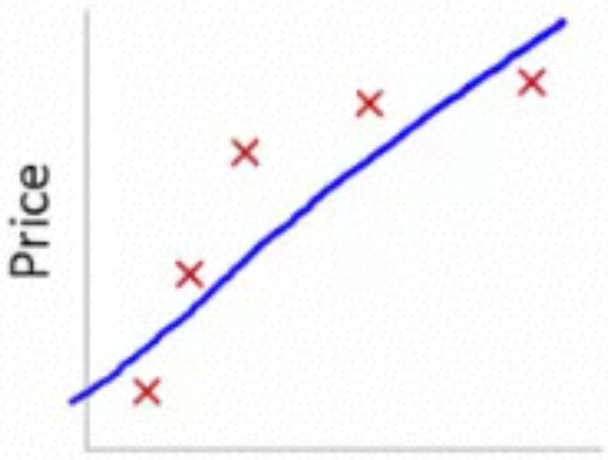




REGRESSION ERROR

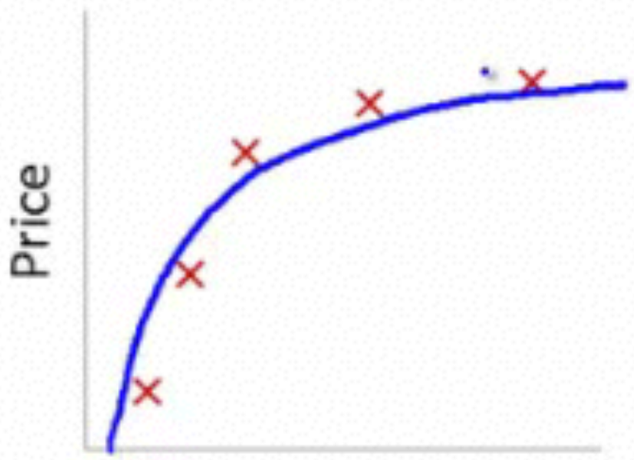


BIAS E OVERFITTING



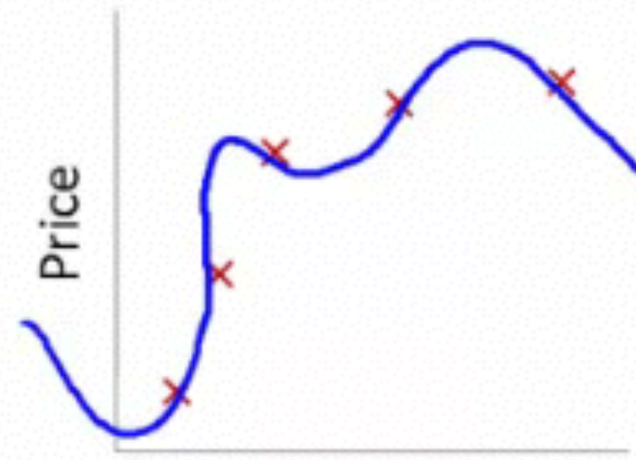
$\theta_0 + \theta_1 x$

High bias
(underfit)



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

"Just right"



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

High variance
(overfit)

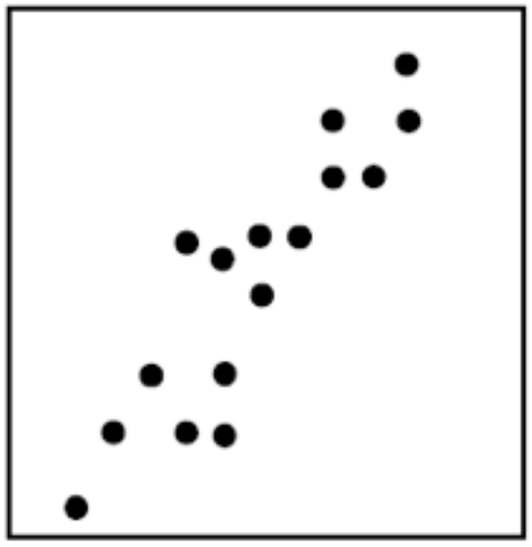
OCCAM'S RAZOR

- ***Se os resultados forem semelhantes escolha a solução mais simples.***
- ***Em Data Science prefira sempre o modelo mais simples***

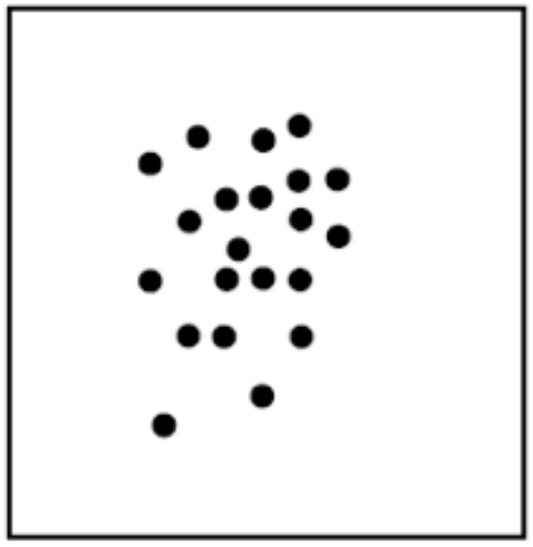
14th-century English logician William of Ockham

CORRELAÇÃO DOS DADOS

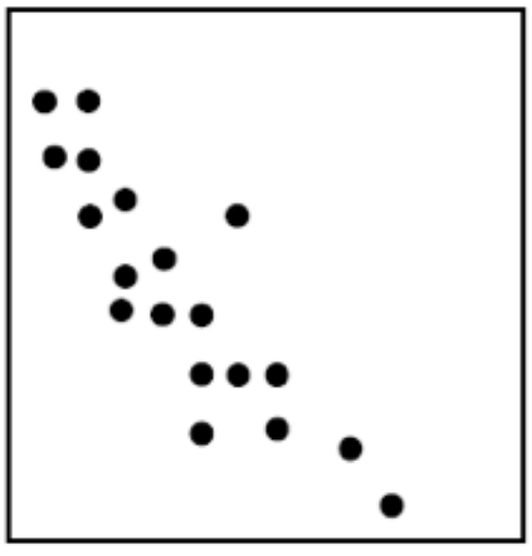
Strong positive correlation



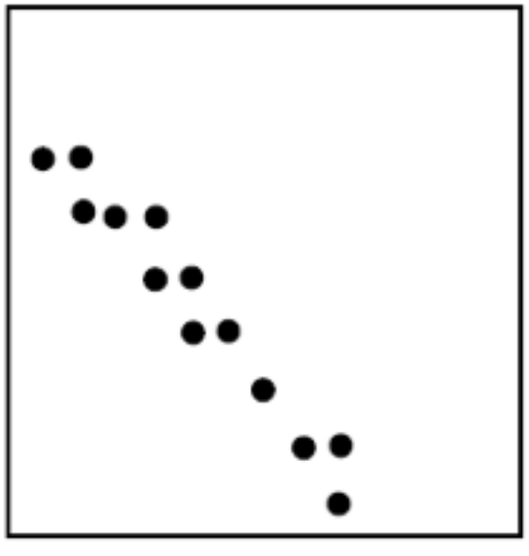
Moderate positive correlation



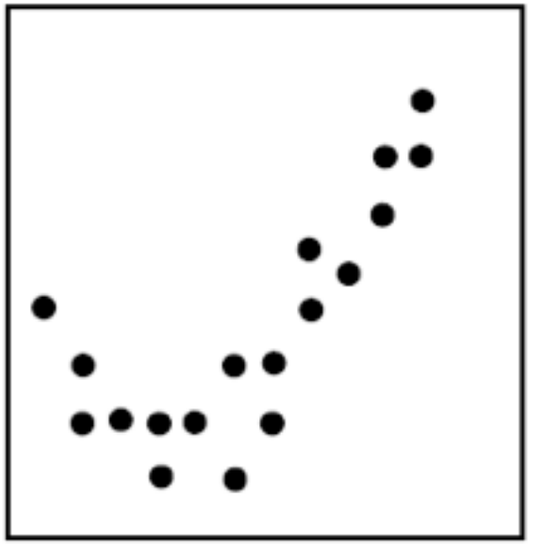
No correlation



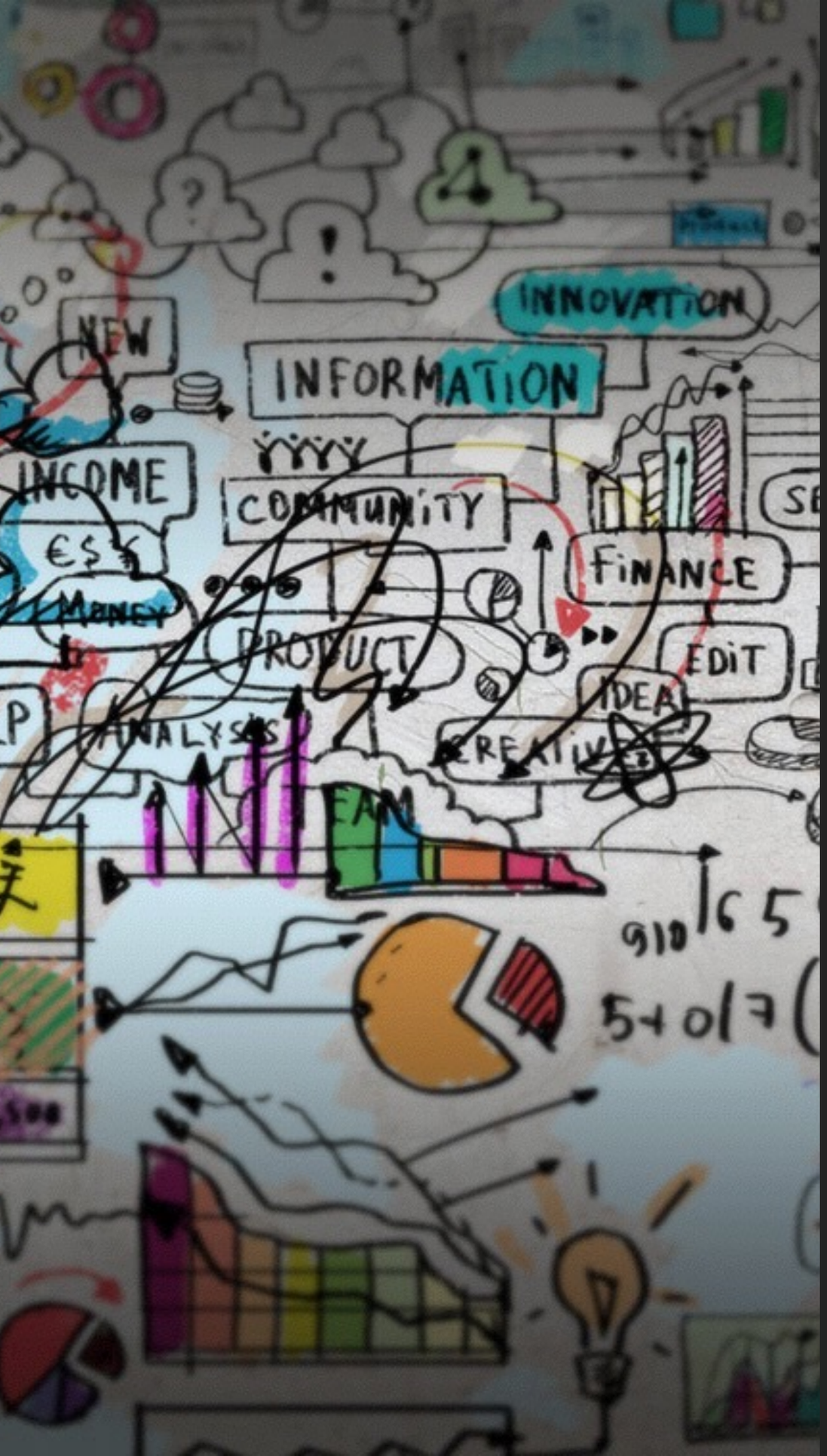
Moderate negative correlation



Strong negative correlation

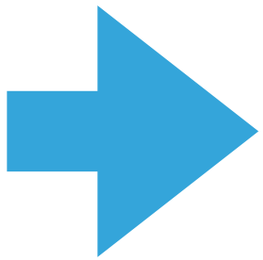


Curvilinear relationship

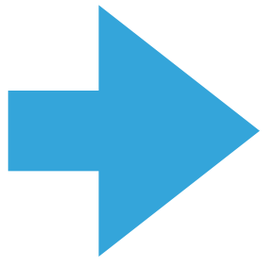


O QUE É CLASSIFICAÇÃO ?

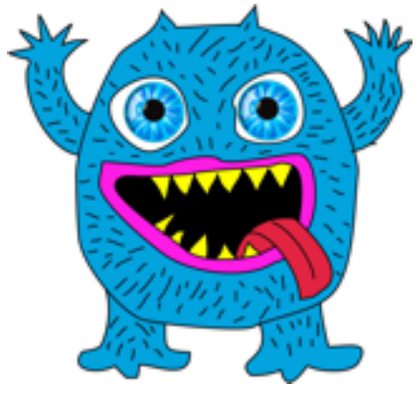
CLASSIFICAÇÃO

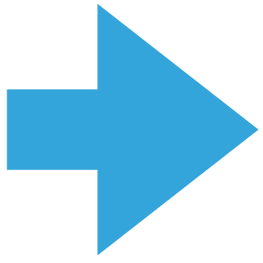


A

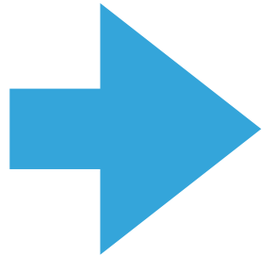


A



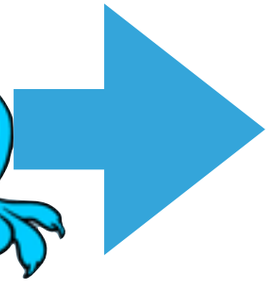


B

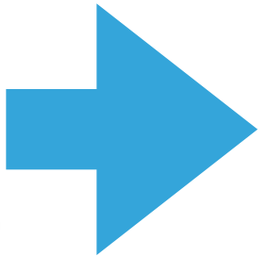


A

?

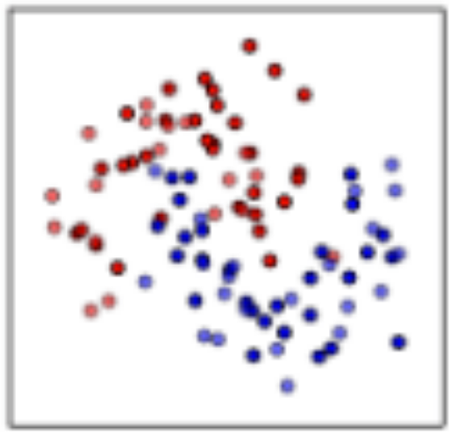


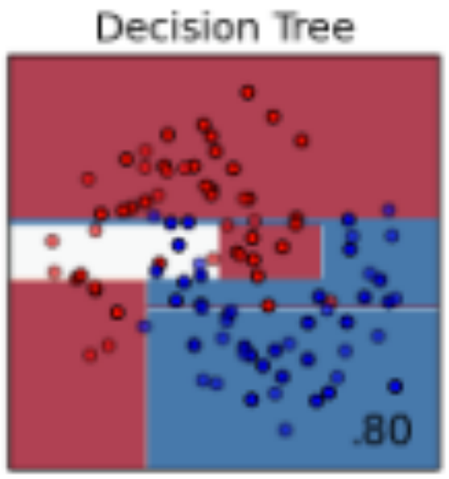
B

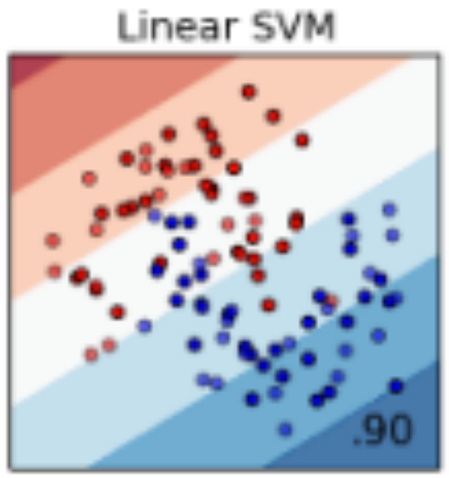


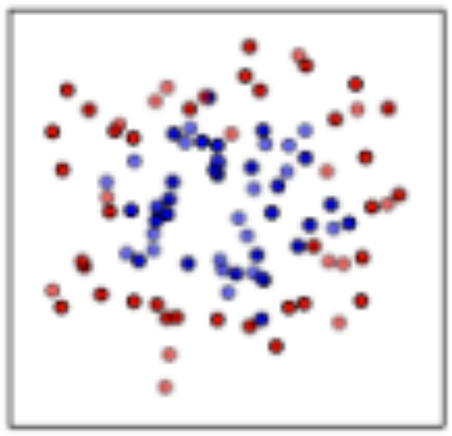
B

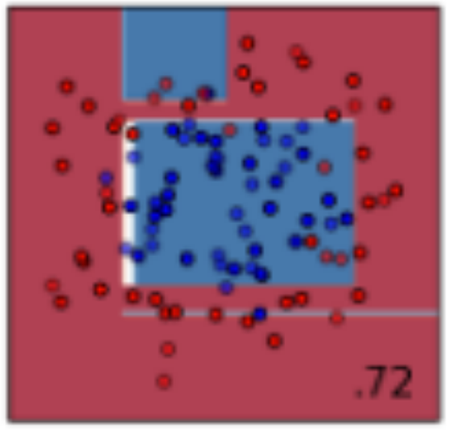
CLASSIFICADORES

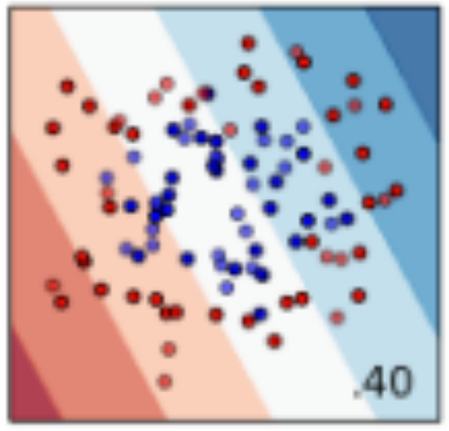


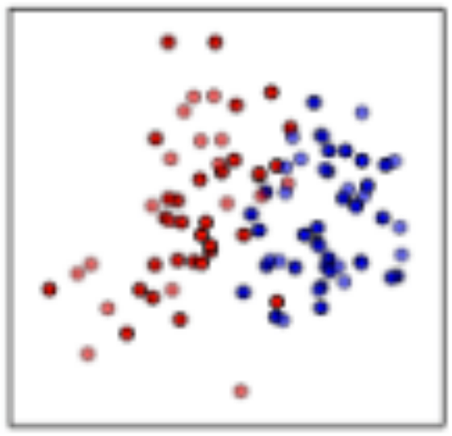


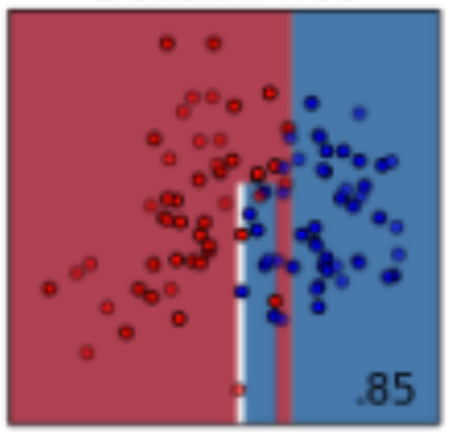


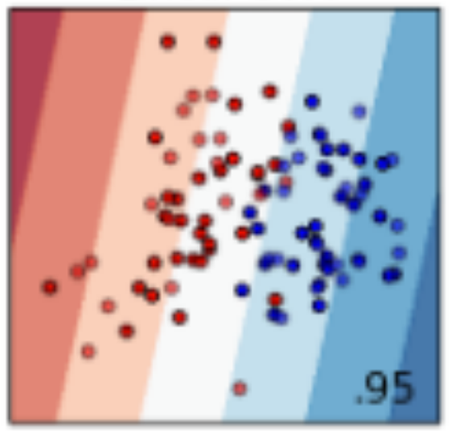




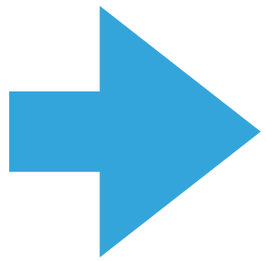




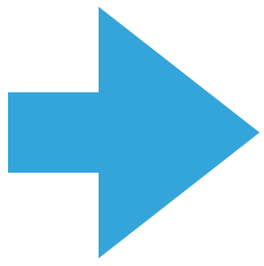




QUAL A TAG A APLICAR?



?

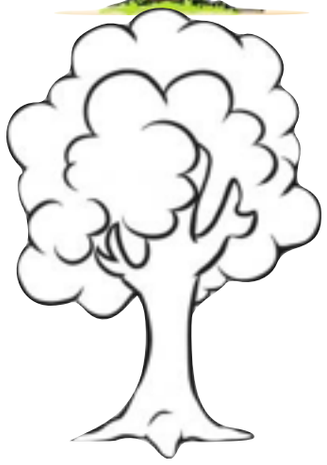


?

QUAL A TAG A APLICAR?

A

B



A

?

B

C

DADOS DE TREINO VS DADOS DE TESTE

- ***Dados de Treino***
  - Usados para treinar um modelo
  - Exemplos

- ***Dados de Teste***
  - Usados para testar a performance do modelo
  - Dados de validação.

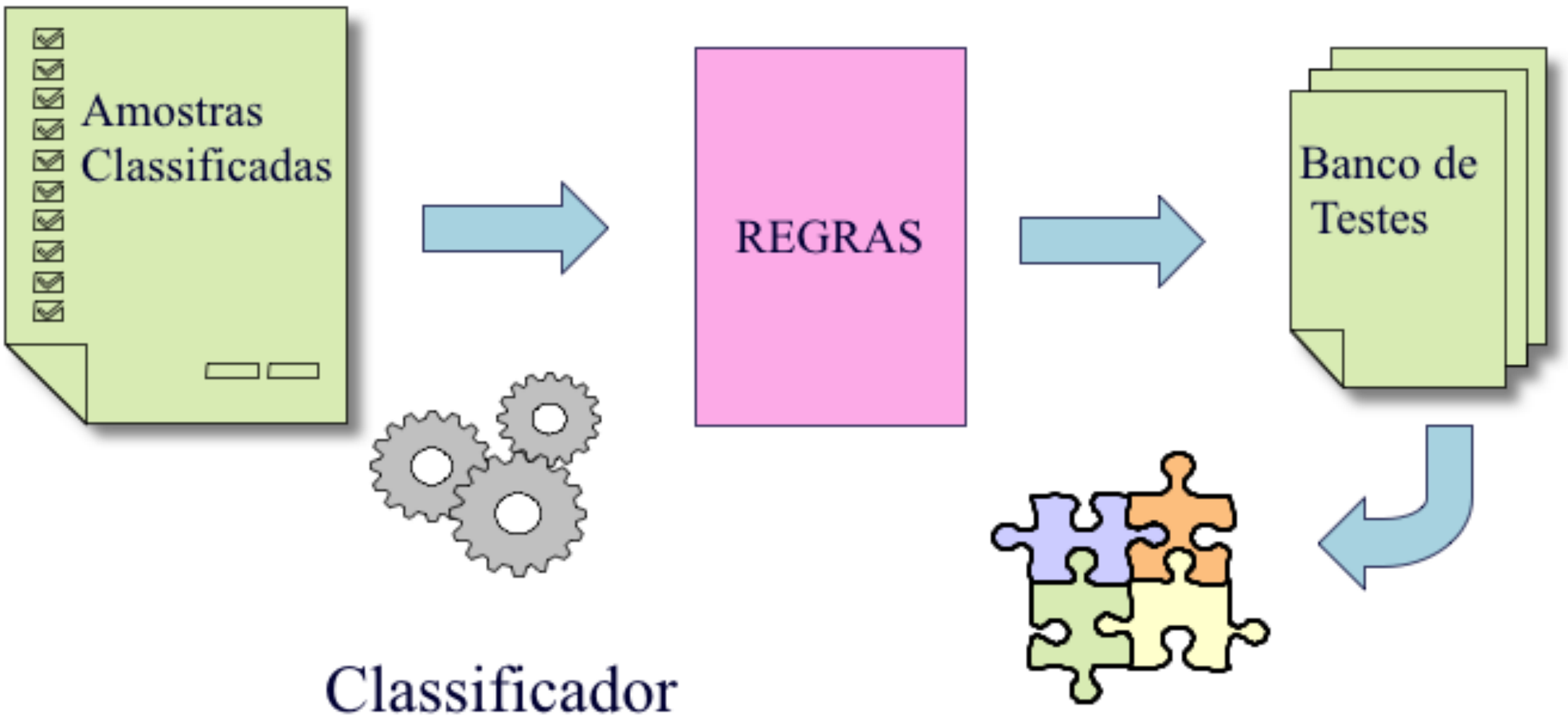
e.g. facial gender classification

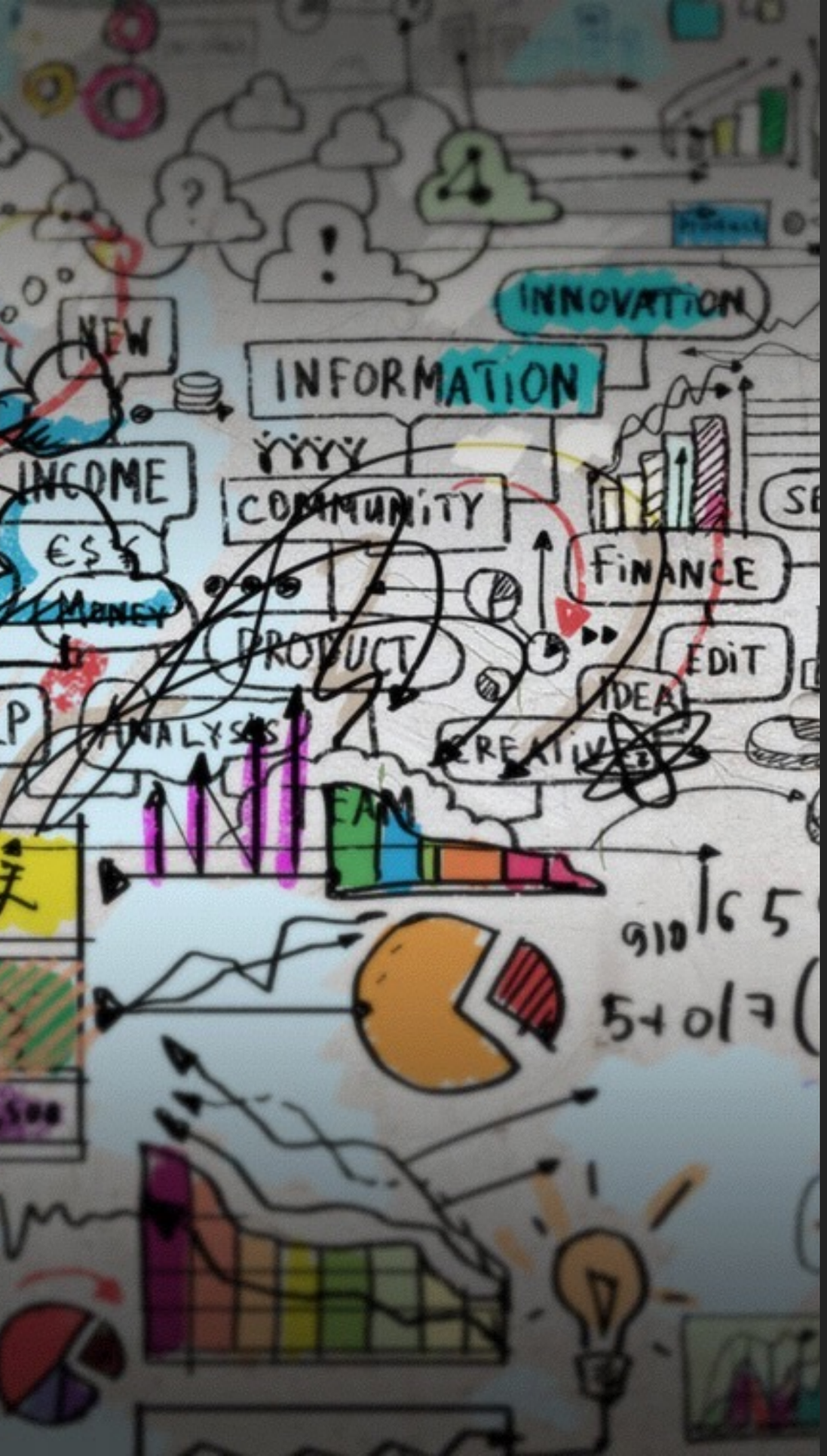
Classificação

Nome	Idade	Renda	Profissão	Classe
Daniel	≤ 30	Média	Estudante	Sim
João	31..50	Média-Alta	Professor	Sim
Carlos	31..50	Média-Alta	Engenheiro	Sim
Maria	31..50	Baixa	Vendedora	Não
Paulo	≤ 30	Baixa	Porteiro	Não
Otavio	> 60	Média-Alta	Aposentado	Não

SE. Idade ≤ 30 E Renda é Média ENTÃO Compra-Produto-Eletrônico = SIM.

Etapas do Processo

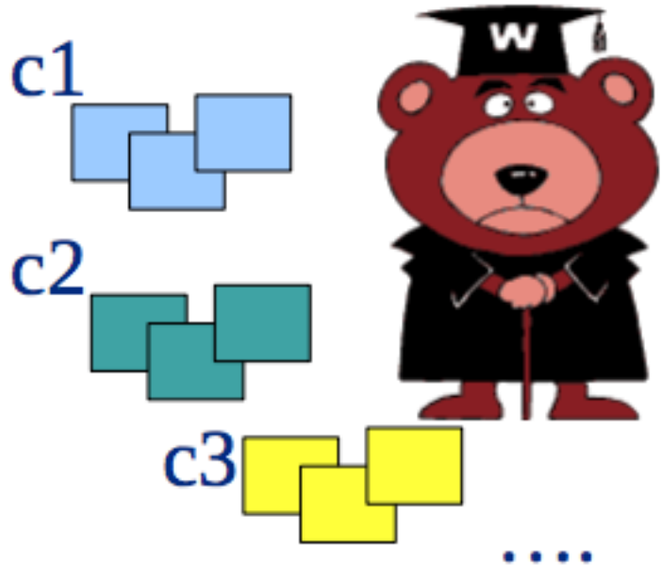




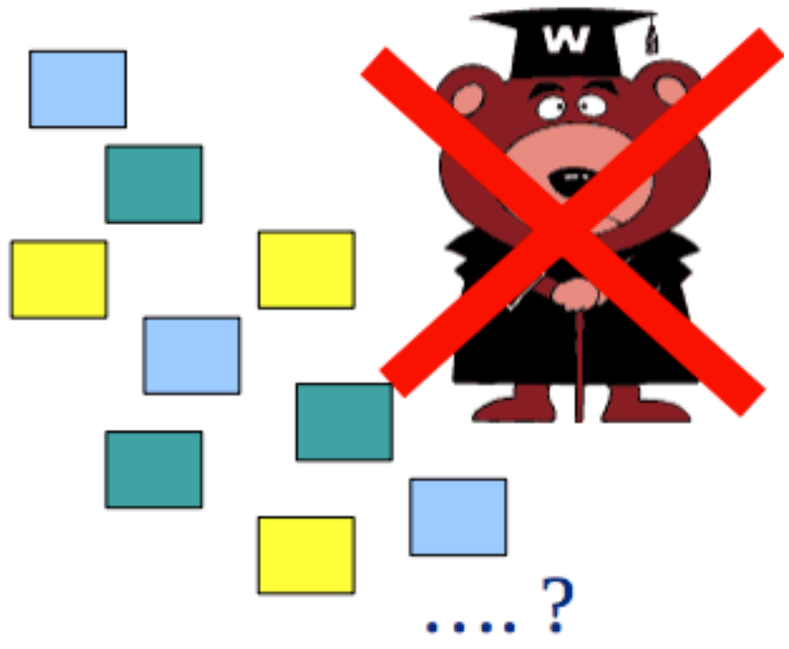
APRENDIZADO SUPERVISIONADO VS NÃO SUPERVISIONADO

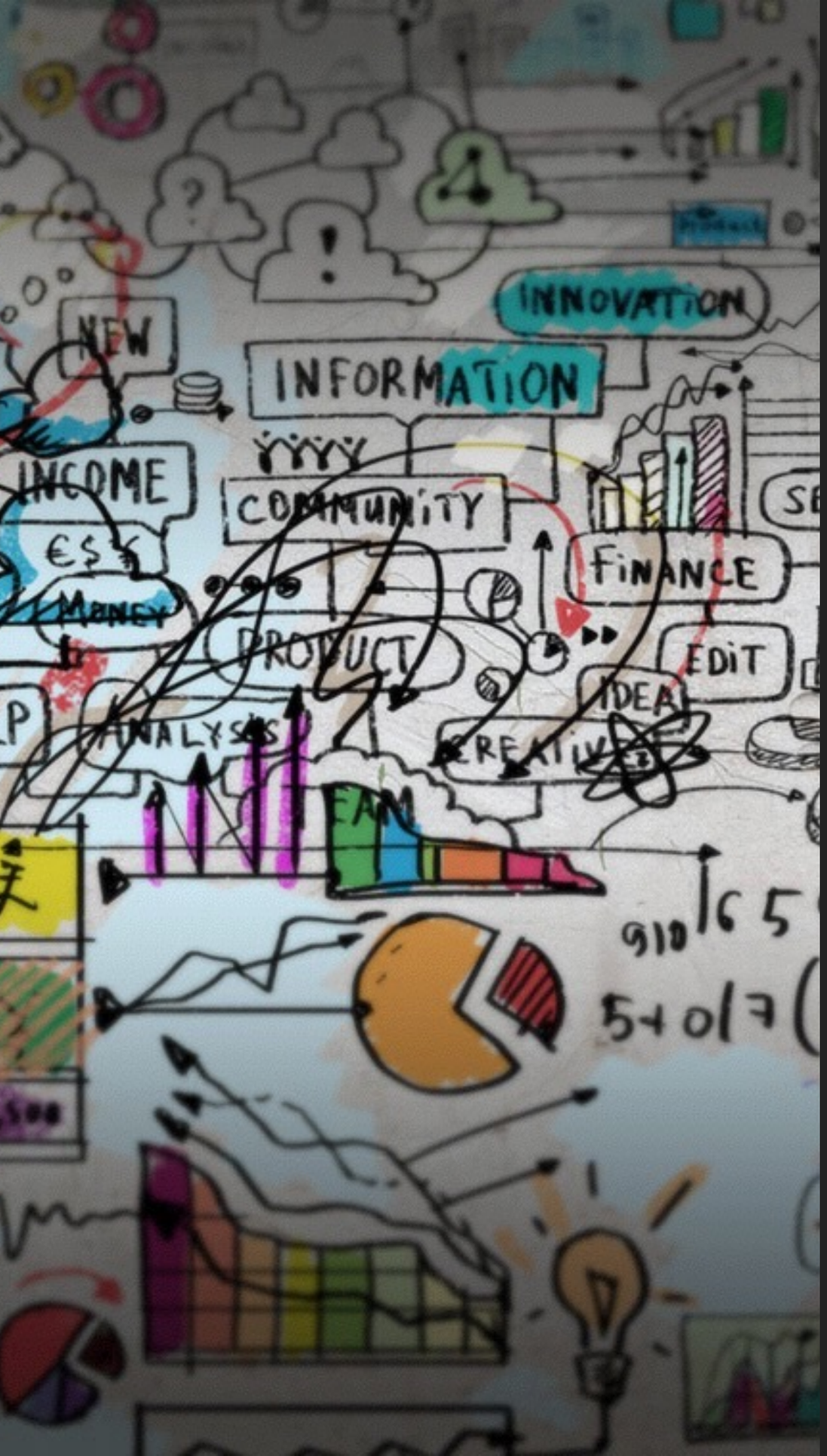
APRENDIZADO SUPERVISIONADO VS NÃO SUPERVISIONADO

- ***Supervisionado***
  - Conhecimento das entradas e saídas de dados
  - Os dados possuem um label
  - O objetivo é predizer a classe ou o label do dado

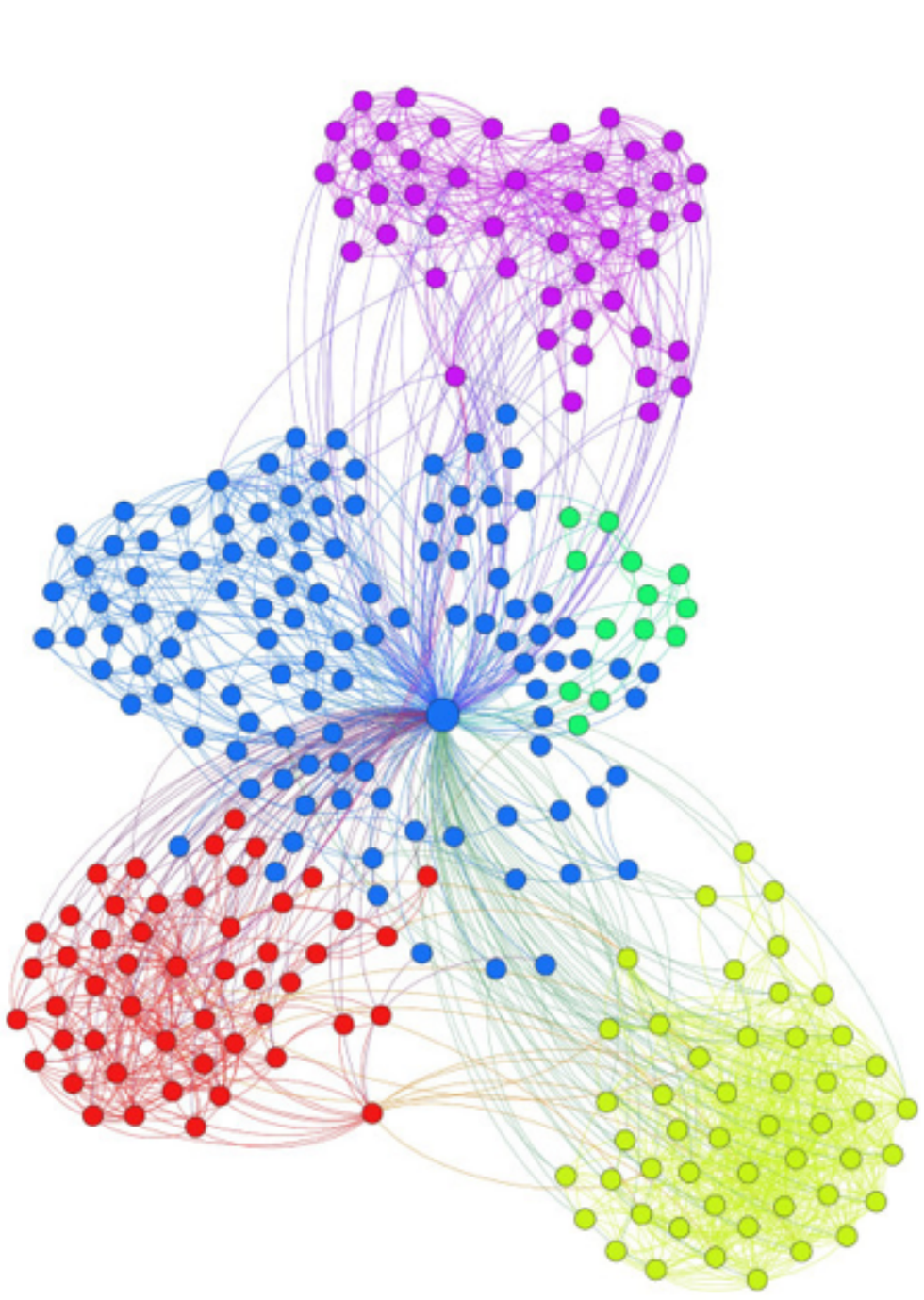


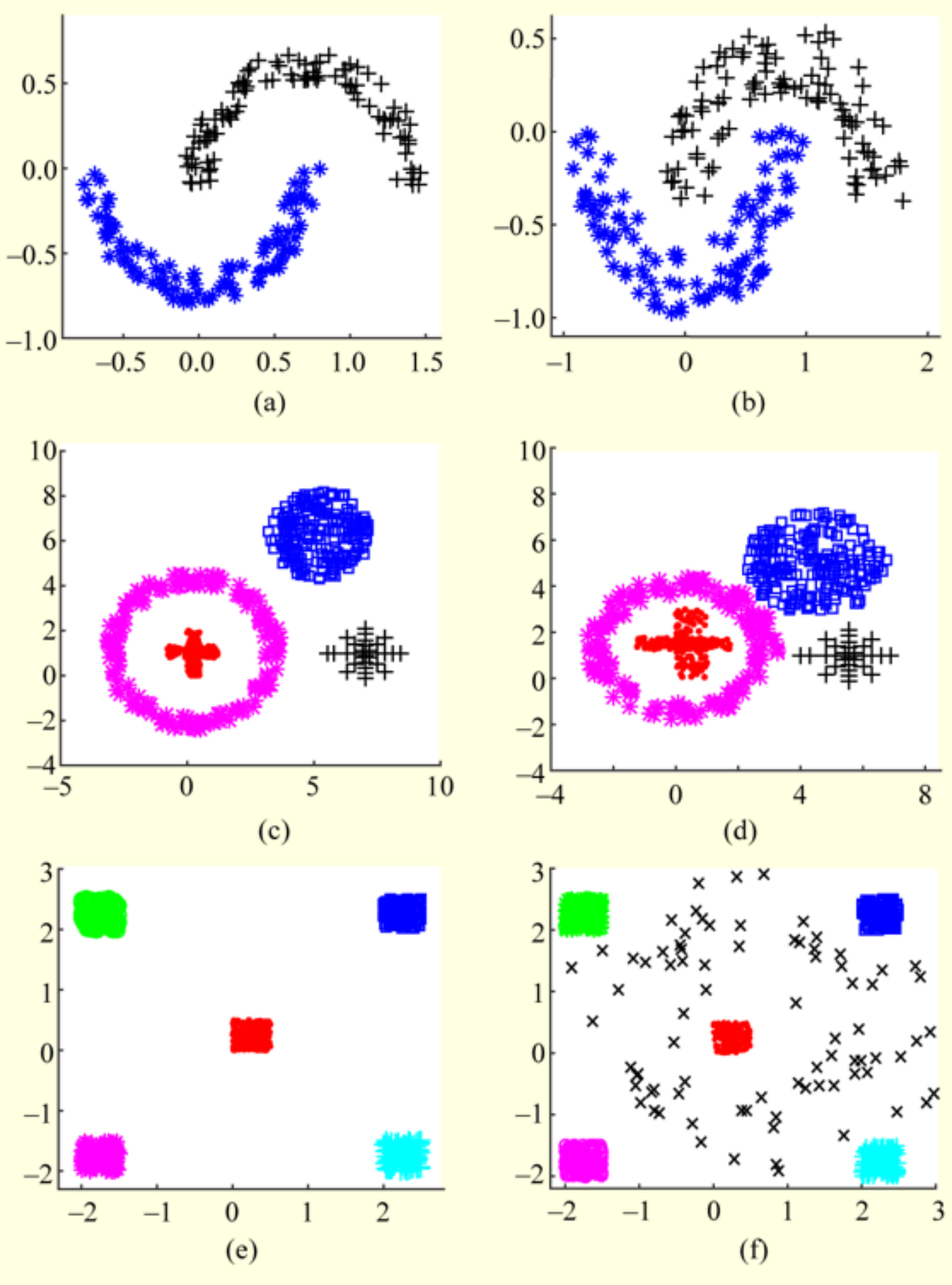
- ***Não Supervisionado***
  - Sem conhecimento prévio dos dados
  - O objetivo é determinar padrões





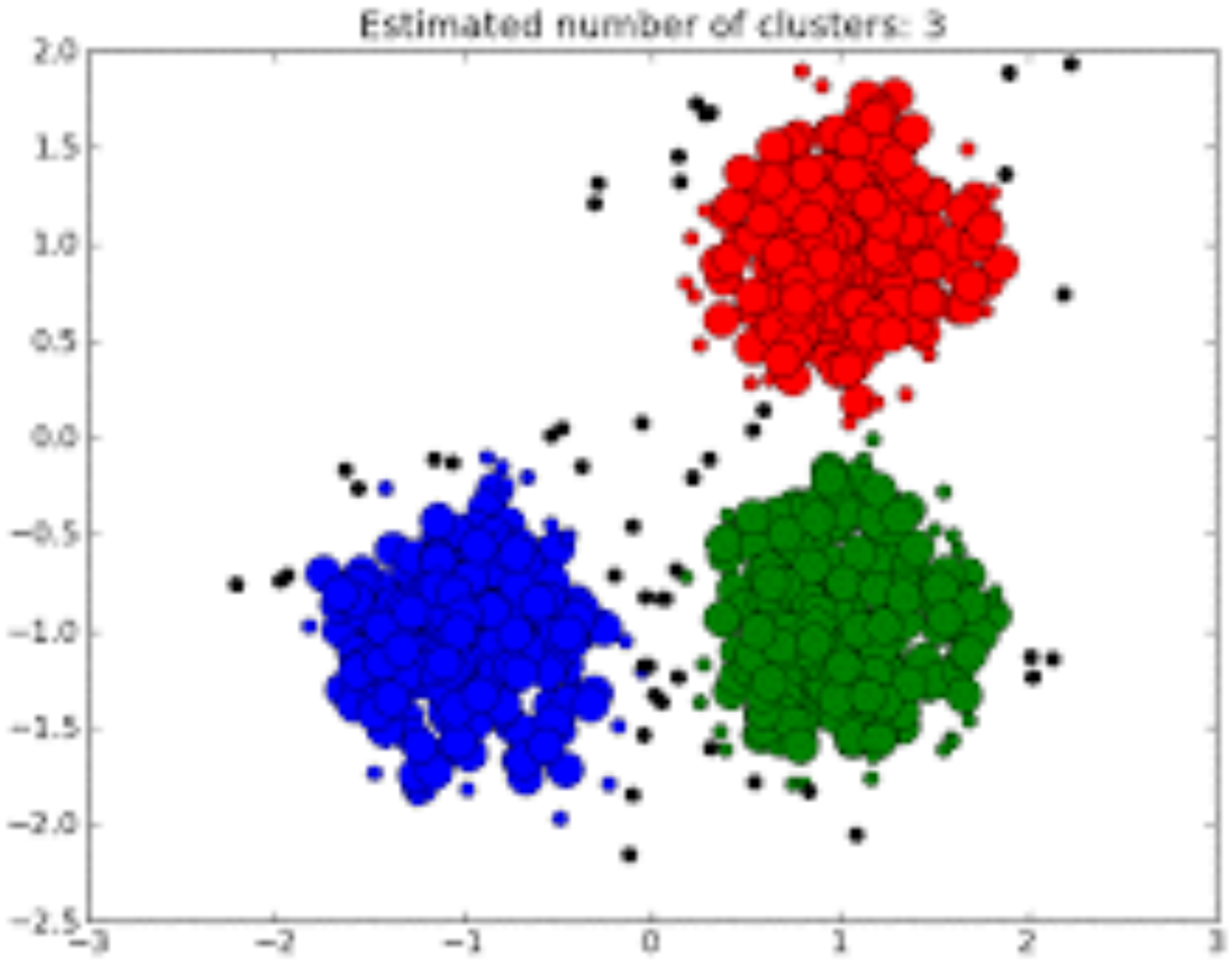
O QUE É AGRUPAMENTO ?




0.5
0.0
−0.5
−1.0
−0.5 0.0 0.5 1.0 1.5
(a)
0.5
0.0
−0.5
−1.0
−1 0 1 2
(b)
10
8
6
4
2
0
−2
−4
−5 0 5 10
(c)
10
8
6
4
2
0
−2
−4
−4 0 4 8
(d)
3
2
1
0
−1
−2
−2 −1 0 1 2
(e)
3
2
1
0
−1
−2
−2 −1 0 1 2 3
(f)

AGRUPAMENTO

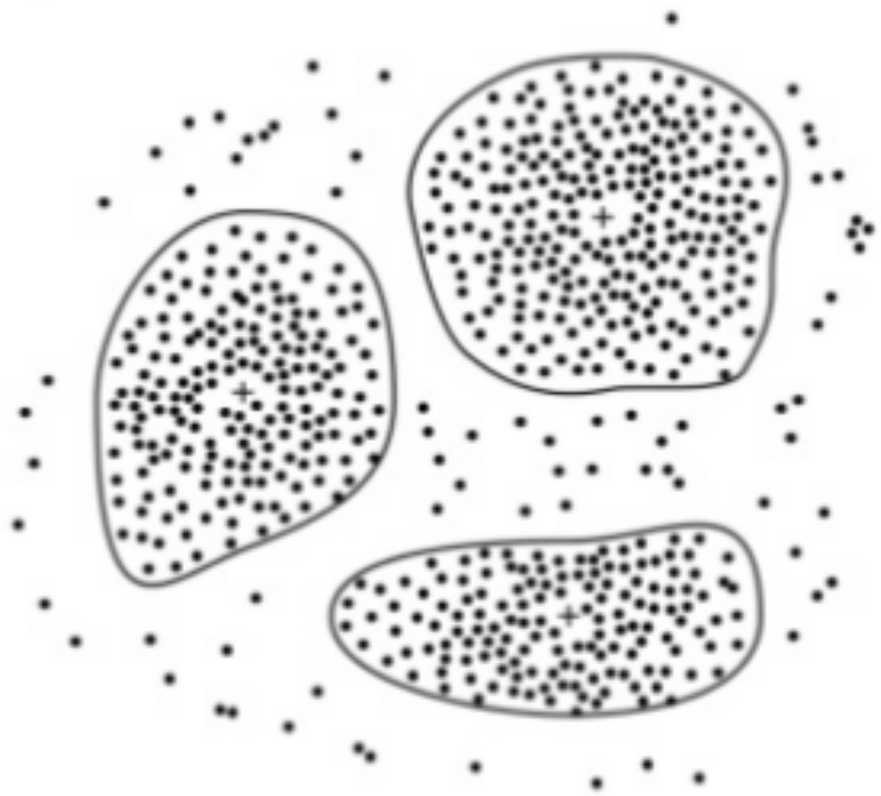
Processo de agrupar objetos com características semelhantes



CLUSTER

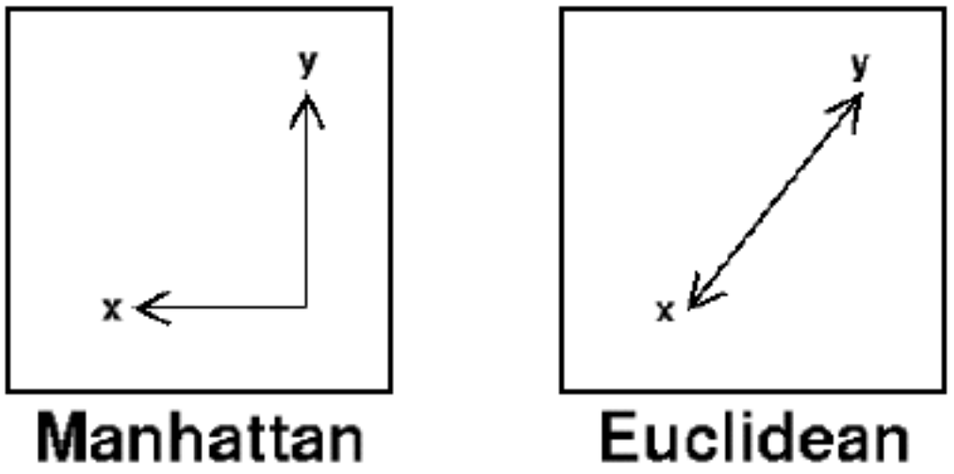
Cluster

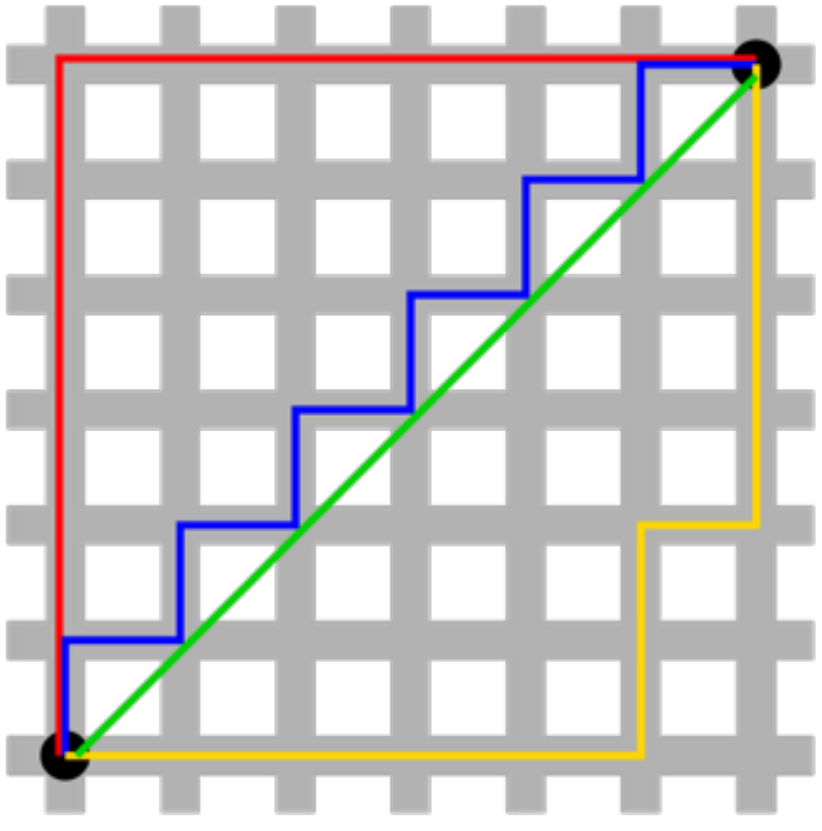
- Uma coleção de objetos que são similares entre si, e diferentes dos objetos pertencentes a outros clusters.



COMO AGRUPAR?

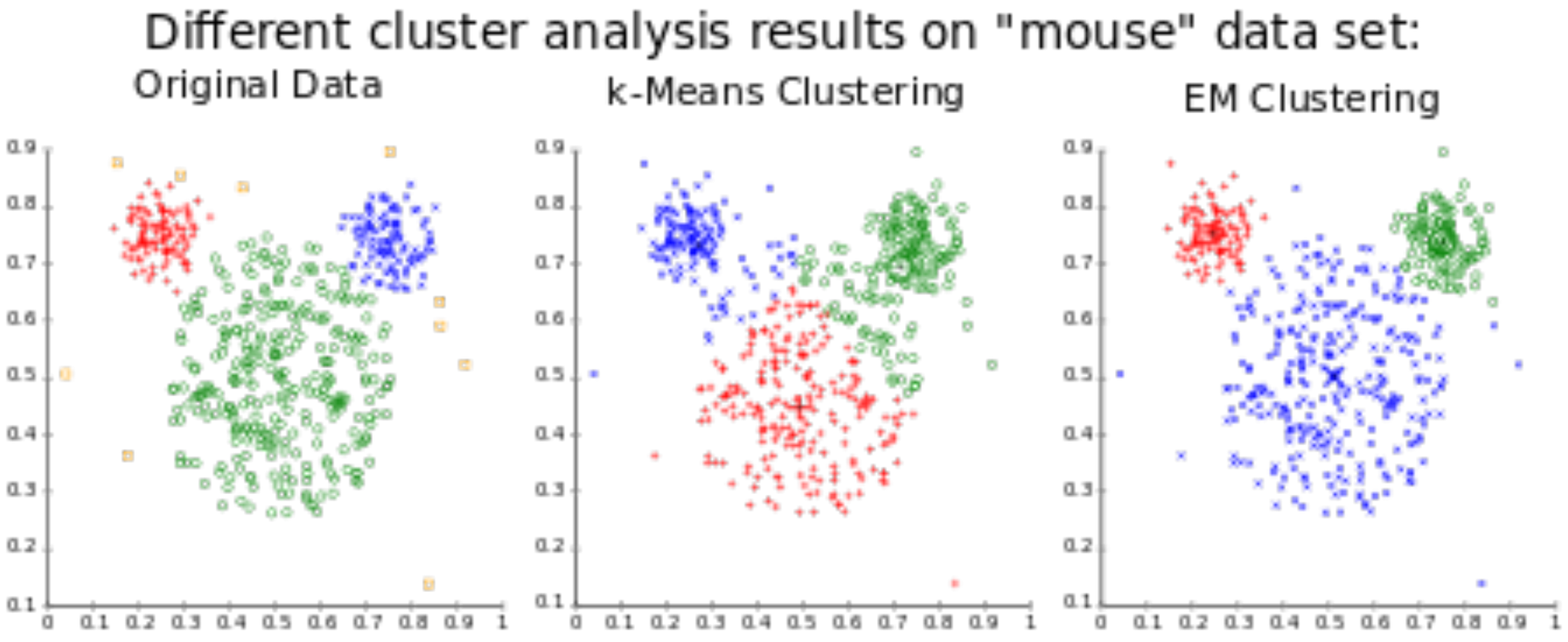
DISTÂNCIA

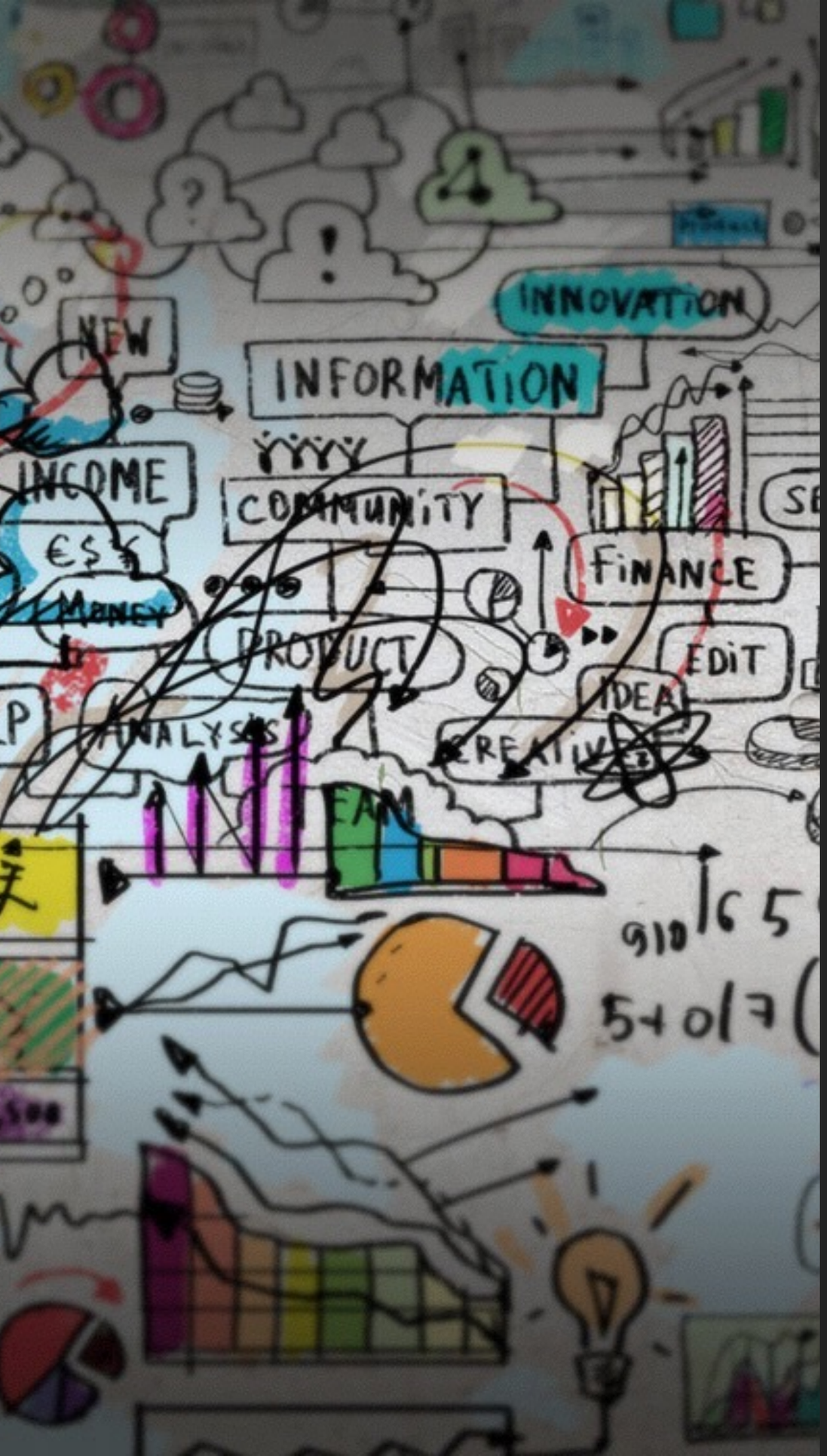

y
x
Manhattan
y
x
Euclidean




Clinton
Korea Town
Murray Hill
Kips Bay
Midtown East

MODELOS DE AGRUPAMENTO





ANÁLISE DE TEXTO

ANALISANDO TEXTO

Passados quase três meses do final dos Jogos, o Comitê **Rio-2016** ainda deve reembolso a **8.000 torcedores** que utilizaram sua plataforma online para revender ingressos.

A entidade reduziu o contingente de consumidores a quem devia pagamentos, que chegou a 140 mil em 19 de outubro, data até a qual prometeu quitar os débitos. Mas ainda não deu fim ao problema.

A entidade afirmou que tem dificuldades para ressarcir o restante. Alega problemas para encontrar os credores e inconsistência nos dados bancários fornecidos —muitos depósitos não foram completados.

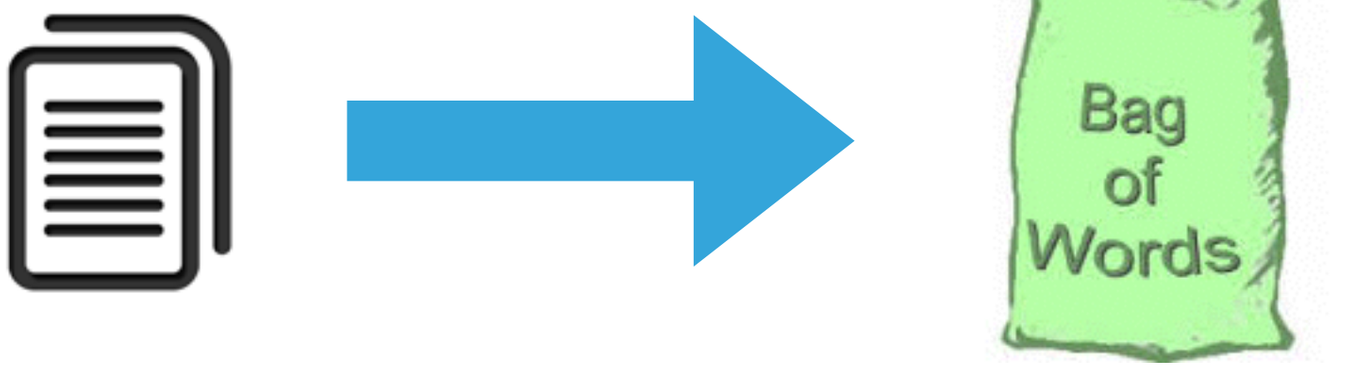
De acordo com o **comitê**, 3.500 pessoas foram procuradas mas não responderam às mensagens eletrônicas, 2.500 até deram retorno, porém as informações repassadas continham algum erro e 2.000 devem receber o reembolso até esta segunda (12), após terem dados checados.

Uma mutação aparentemente insignificante no **DNA** dos ancestrais da humanidade pode ter contribuído para que nosso cérebro alcançasse o tamanho descomunal que tem hoje (três vezes maior que o dos grandes macacos).

Bastou inserir o gene que contém essa mutação em fetos de camundongo para que dobrasse o número de células que dão origem aos neurônios do córtex, a área cerebral mais "nobre".

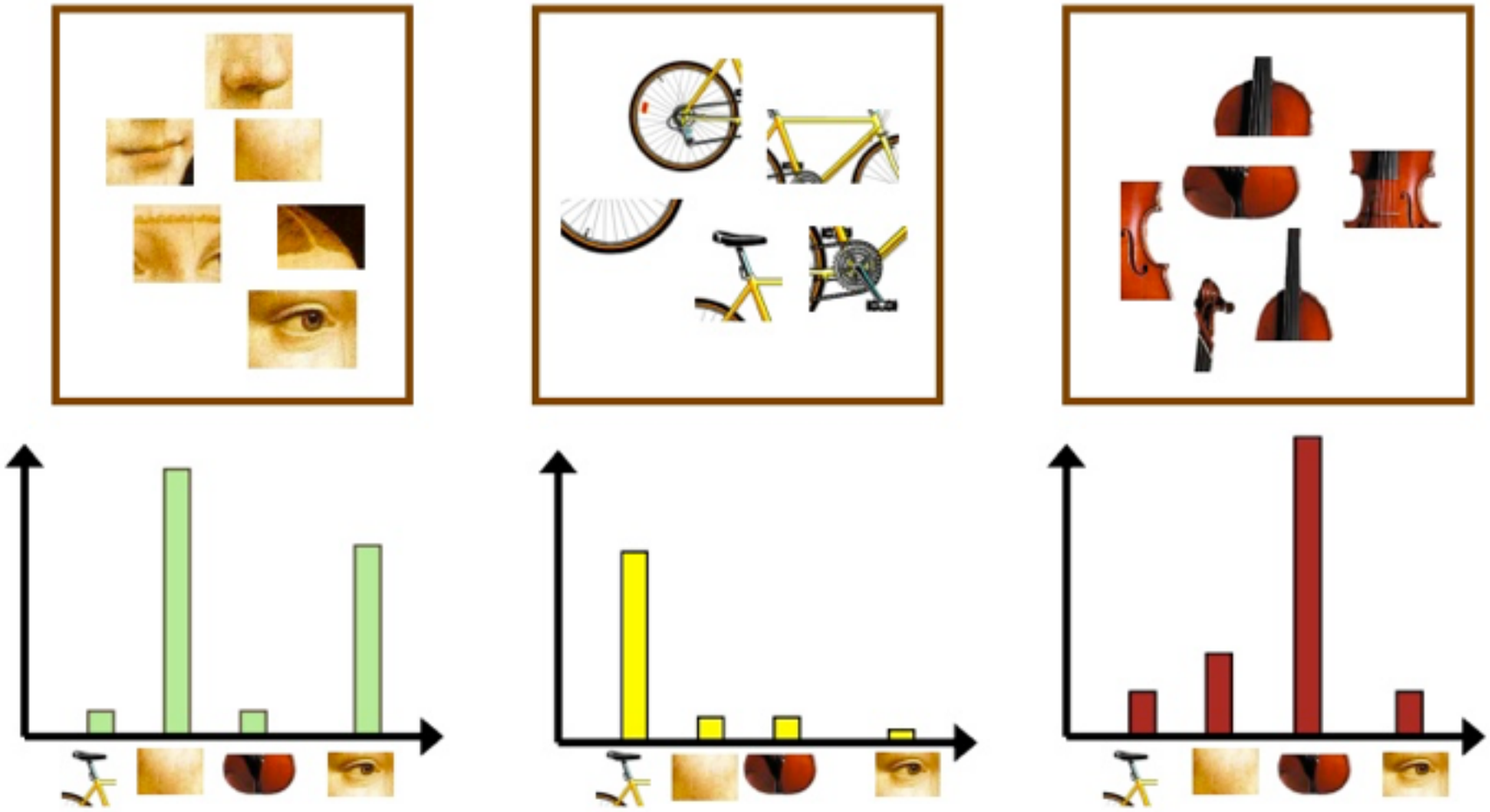
A **pesquisa,** conduzida por cientistas do **Instituto** Max Planck (Alemanha), é um dos primeiros frutos da tentativa de usar o genoma para entender como a evolução humana se desenrolou. Por enquanto, isso não tem sido fácil –tanto que o gene analisado pelos pesquisadores no novo estudo, designado pela indigesta sigla ARHGAP11B, é o único específico da linhagem humana a ser associado com a proliferação das tais células do córtex cerebral.

ANÁLISE DE SENTIMENTO




```
>>> docs_new = ['God is love', 'OpenGL on the GPU is fast']
>>> X_new_counts = count_vect.transform(docs_new)
>>> X_new_tfidf = tfidf_transformer.transform(X_new_counts)

>>> predicted = clf.predict(X_new_tfidf)

>>> for doc, category in zip(docs_new, predicted):
...     print('%r => %s' % (doc, twenty_train.target_names[category]))
...
'God is love' => soc.religion.christian
'OpenGL on the GPU is fast' => comp.graphics
```

REPRESENTAÇÃO DE TEXTO

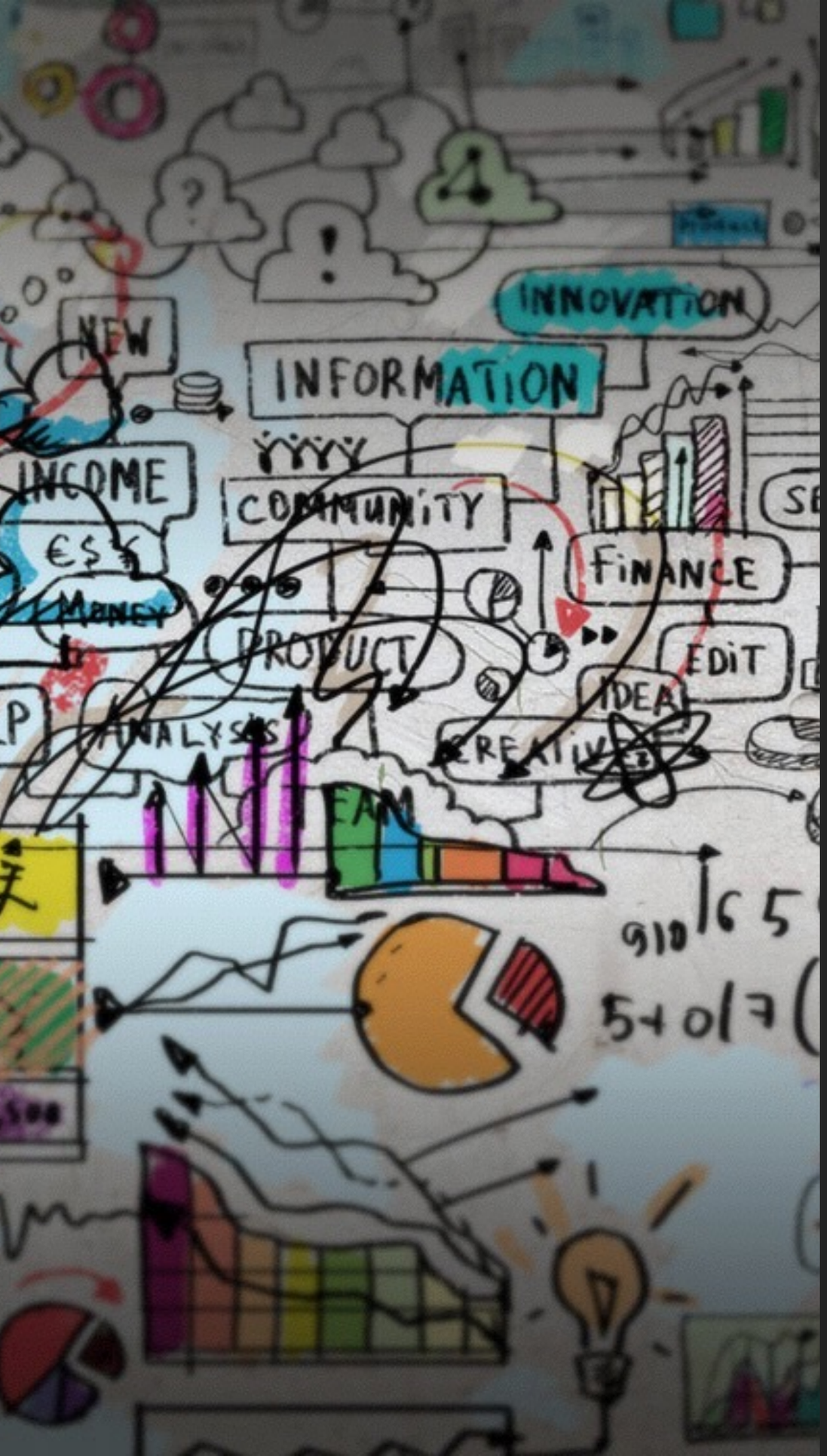


Document

In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
upon the face of the waters.
And God said, Let there be
light: and there was light.

Representation

beginning	1
earth	2
God	3



ANALISANDO A PERFORMANCE DE UM MODELO

ACURÁCIA

Acurácia – Taxa de erros

- Acc(M) = porcentagem das tuplas dos dados de teste que sao corretamente classificadas.
- Err(M) = 1 – Acc(M)
- Matriz de Confusão

		Classes Preditas	
		C1	C2
Classes Reais	C1	**Positivos verdadeiros**	**Falsos Negativos**
	C2	**Falsos Positivos**	**Negativos verdadeiros**

Outras medidas mais precisas

- **Exemplo** : acc(M) = 90%
  C1 = tem-câncer (4 pacientes)
  C2 = não-tem-câncer (500 pacientes)

Classificou corretamente 454 pacientes que não tem câncer

Não acertou nenhum dos que tem câncer

Pode ser classificado como "bom classificador" mesmo com acurácia alta ?

- **Sensitividade** = $\frac{\text{true-pos}}{\text{pos}}$
- **Especificidade** = $\frac{\text{true-neg}}{\text{neg}}$

% pacientes classificados **corretamente com câncer** dentre todos os que **realmente tem câncer**

- **Precisão** = $\frac{\text{true-pos}}{\text{true-pos} + \text{falso-pos}}$

% pacientes classificados **corretamente com câncer** dentre todos os que foram classificados **com câncer**

PERFORMANCE DE UM CLASSIFICADOR

- Acuracia = classificados corretamente /total de exemplos
- Erro = 1-Acuracia

PERFORMANCE DE UMA REGRESSÃO

Uma das formas de avaliar a qualidade do ajuste do modelo é através do coeficiente de determinação. Basicamente, este coeficiente indica quanto o modelo foi capaz de explicar os dados coletados. O coeficiente de determinação é dado pela expressão

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \frac{\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

Razão entre a soma de quadrados da regressão e a soma de quadrados total.

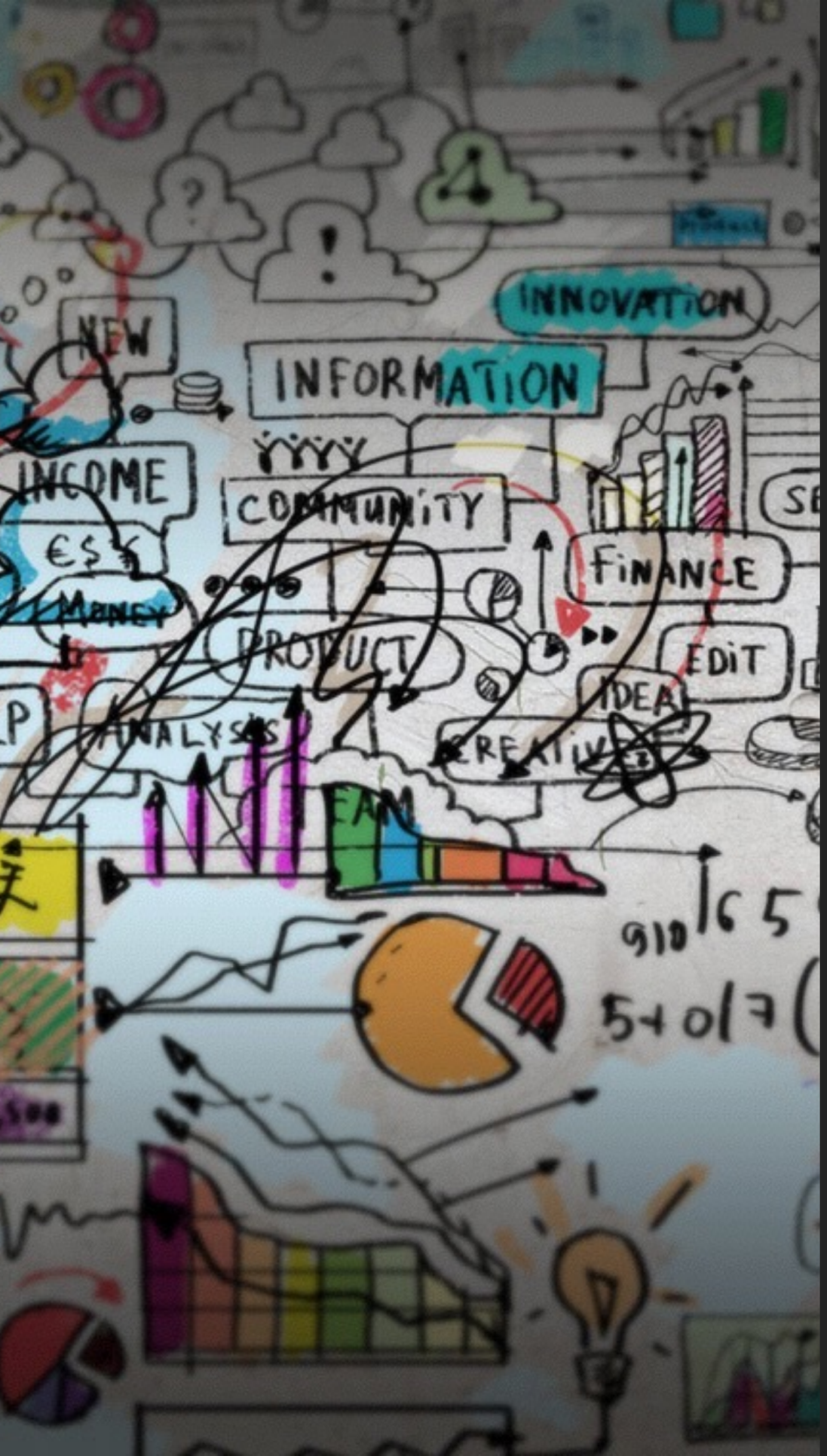
$$0 \leq R^2 \leq 1.$$

TABELA CONFUSÃO

		Classe real	
		p	n
Classe predita	p	Verdadeiro Positivo	Falso Positivo
	n	Falso Negativo	Verdadeiro Negativo

MEAN SQUARE ERROR

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$



FERRAMENTAS E LINGUAGENS

- NUMPY
- PANDAS
- SKLEARN
- R STUDIO

NUMPY

Biblioteca em python para manipulação de arrays e matrizes

PANDAS

Biblioteca de Manipulação de dados e analise em python

NUMPY E PANDAS- LENDO ARQUIVO CSV

```
import numpy as np
import pandas as pd
import visuals as vs # Supplementary code
from sklearn.cross_validation import ShuffleSplit
```

```
# Load the Boston housing dataset
data = pd.read_csv('housing.csv')
prices = data['MEDV']
features = data.drop('MEDV', axis = 1)
```

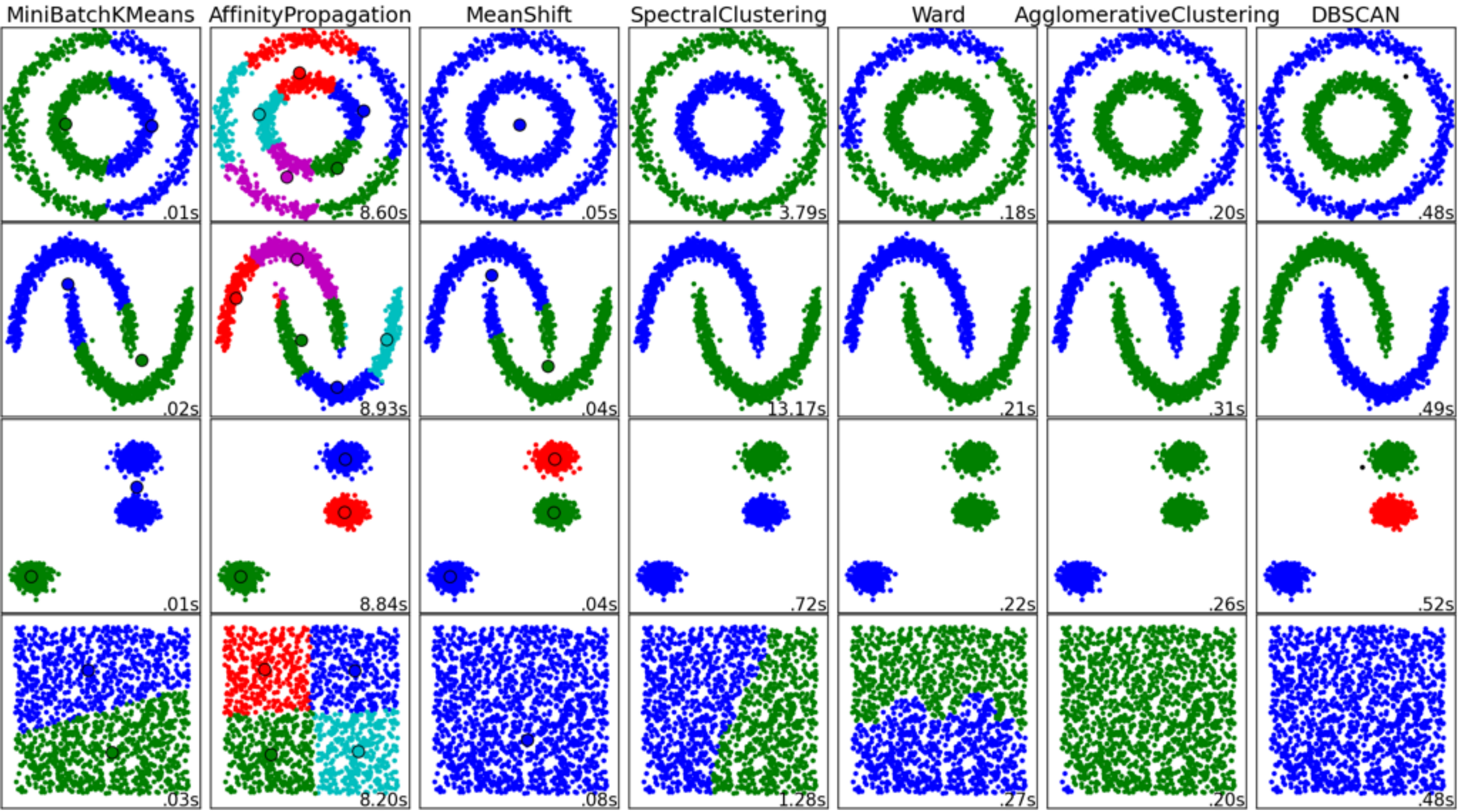
NUMPY E PANDAS- MÉDIA, MEDIANA E DESVIO PADRÃO

```
# TODO: Mean price of the data
mean_price = np.mean(prices)

# TODO: Median price of the data
median_price = np.median(prices)

# TODO: Standard deviation of prices of the data
std_price = np.std(prices)

# Show the calculated statistics
print "Statistics for Boston housing dataset:\n"
print "Mean price: ${:,.2f}".format(mean_price)
print "Median price ${:,.2f}".format(median_price)
print "Standard deviation of prices: ${:,.
2f}".format(std_price)
```

SKLEARN

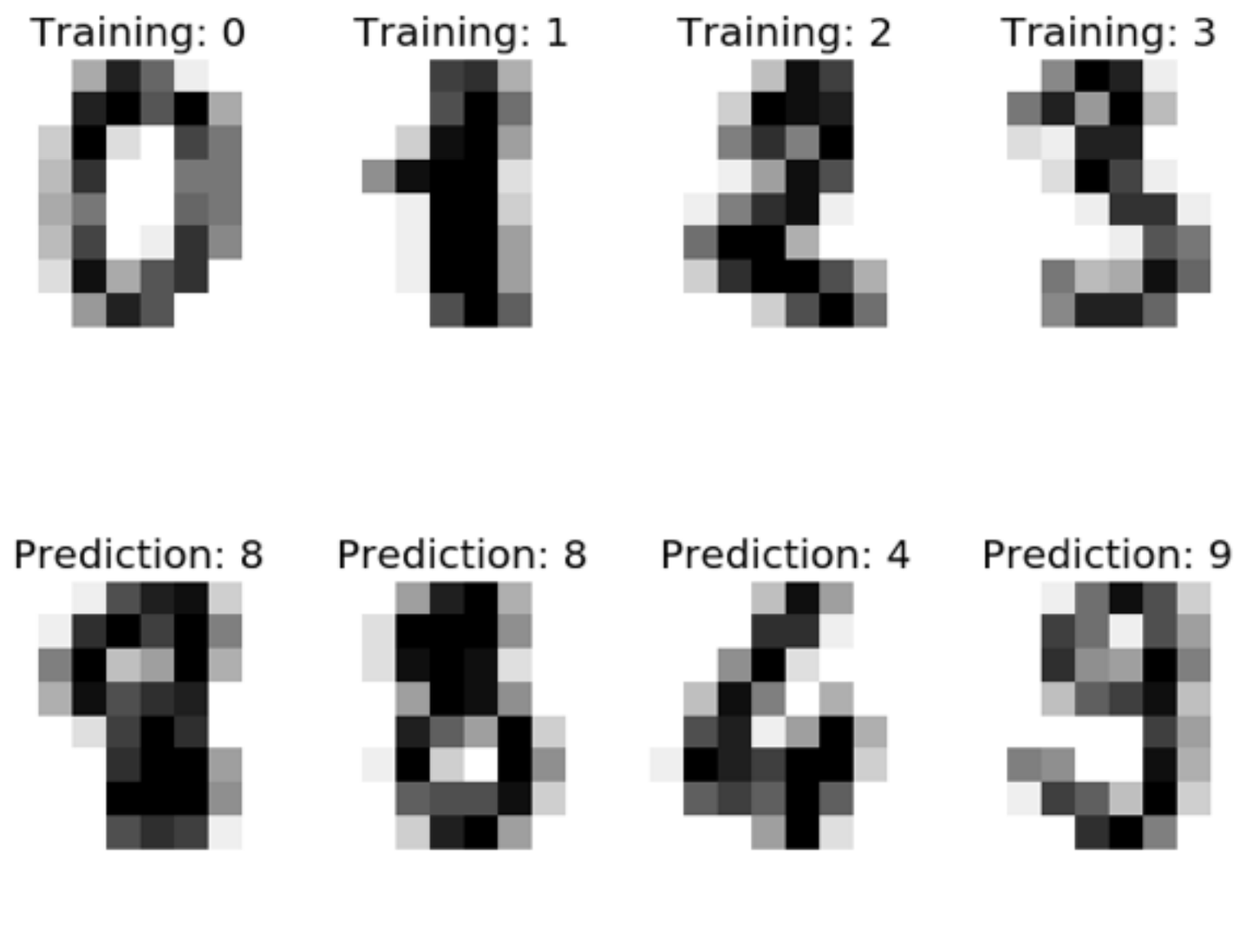
- Aplicação simples e eficiente para data mining e data analysis
- Feito com NumPy, SciPy, e matplotlib
- Open source, commercially usable – BSD license

ALGORITMOS DE AGRUPAMENTO DO SKLEARN



CLASSIFICAÇÃO NO SKLEARN

```
# Create a classifier: a support vector classifier
classifier = svm.SVC(gamma=0.001)

# We learn the digits on the first half of the digits
classifier.fit(data[:n_samples / 2], digits.target[:n_samples /
2])
```

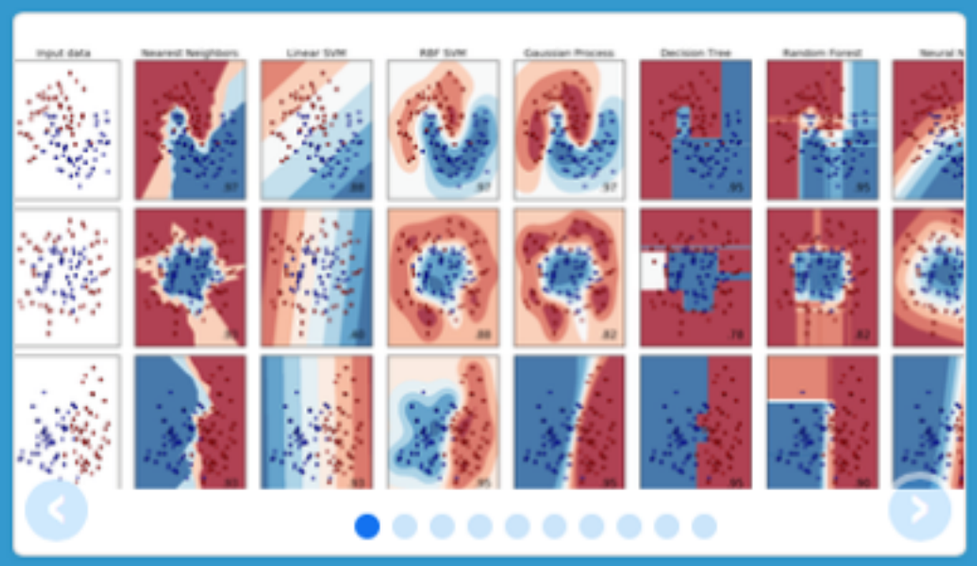


CLASSIFICANDO EMAILS

	I	Linux	tomorrow	today	Viagra	Free
ham	319	619	123	67	0	50
spam	233	3	42	432	291	534

```
from sklearn.naive_bayes import MultinomialNB
classifier = MultinomialNB()
…
classifier.fit(counts, targets)

examples = ['Free Viagra call today!', "I'm going to
attend the Linux users group tomorrow."]
….
predictions = classifier.predict(example_counts)
predictions # [1, 0]
```

HTTP://SCIKIT-LEARN.ORG/STABLE/



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Regression

Predicting a continuous-valued attribute associated with an object.

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Group

LINGUAGEM R

R é uma linguagem e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos.

Foi criada originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia.

EXEMPLOS DE APLICAÇÃO

EXEMPLO DE PROBLEMA – ÁREA (BIOLOGIA) – TIPO: REGRESSÃO

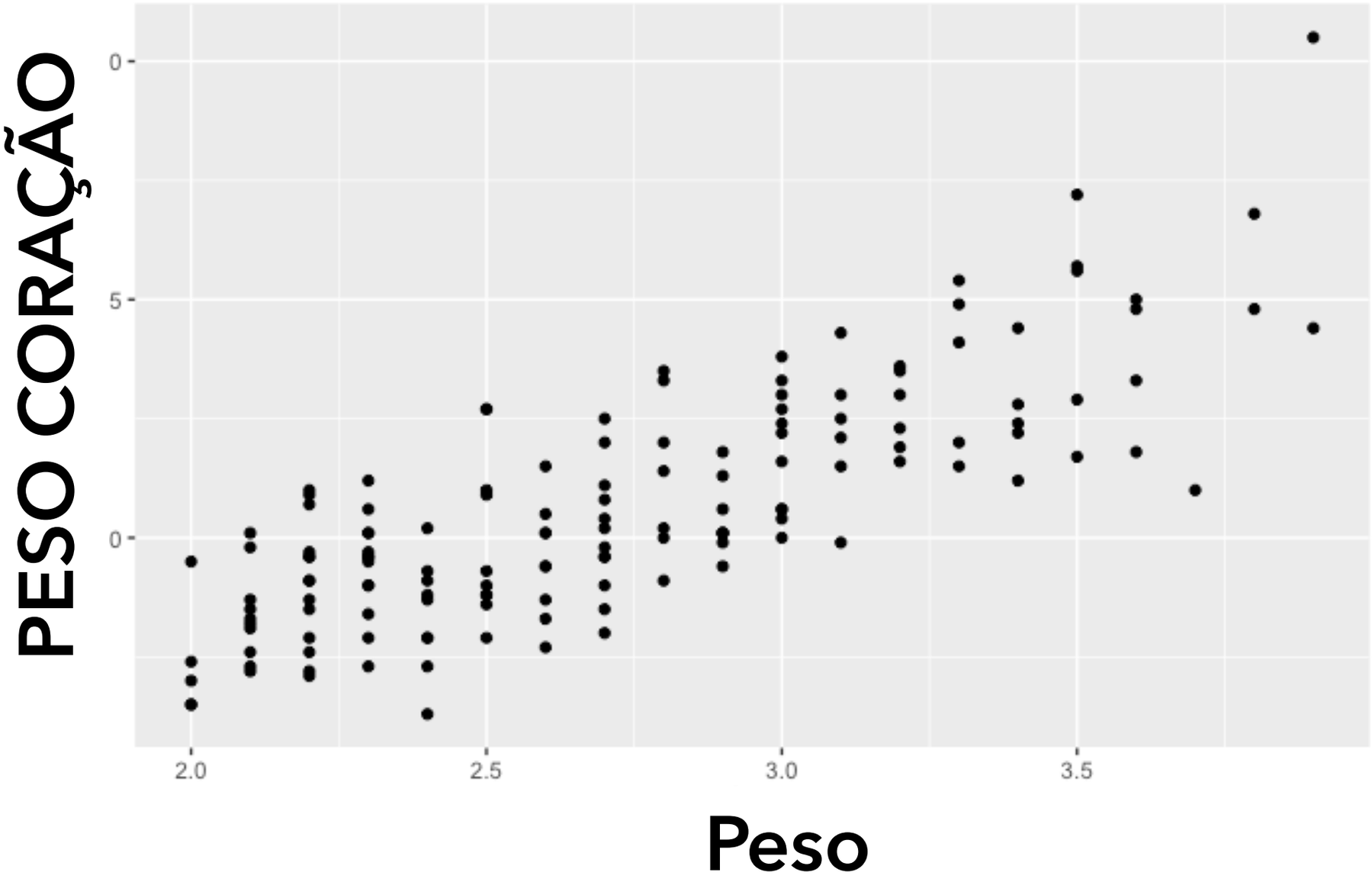
ID	SEXO	CORAÇÃO	PESO
1	F	2.0	7.0
2	F	2.0	7.4
3	F	2.0	9.5
4	F	2.1	7.2
5	F	2.1	7.3
6	F	2.1	7.6
7	F	2.1	8.1
8	F	2.1	8.2

library("MASS")

data(cats)

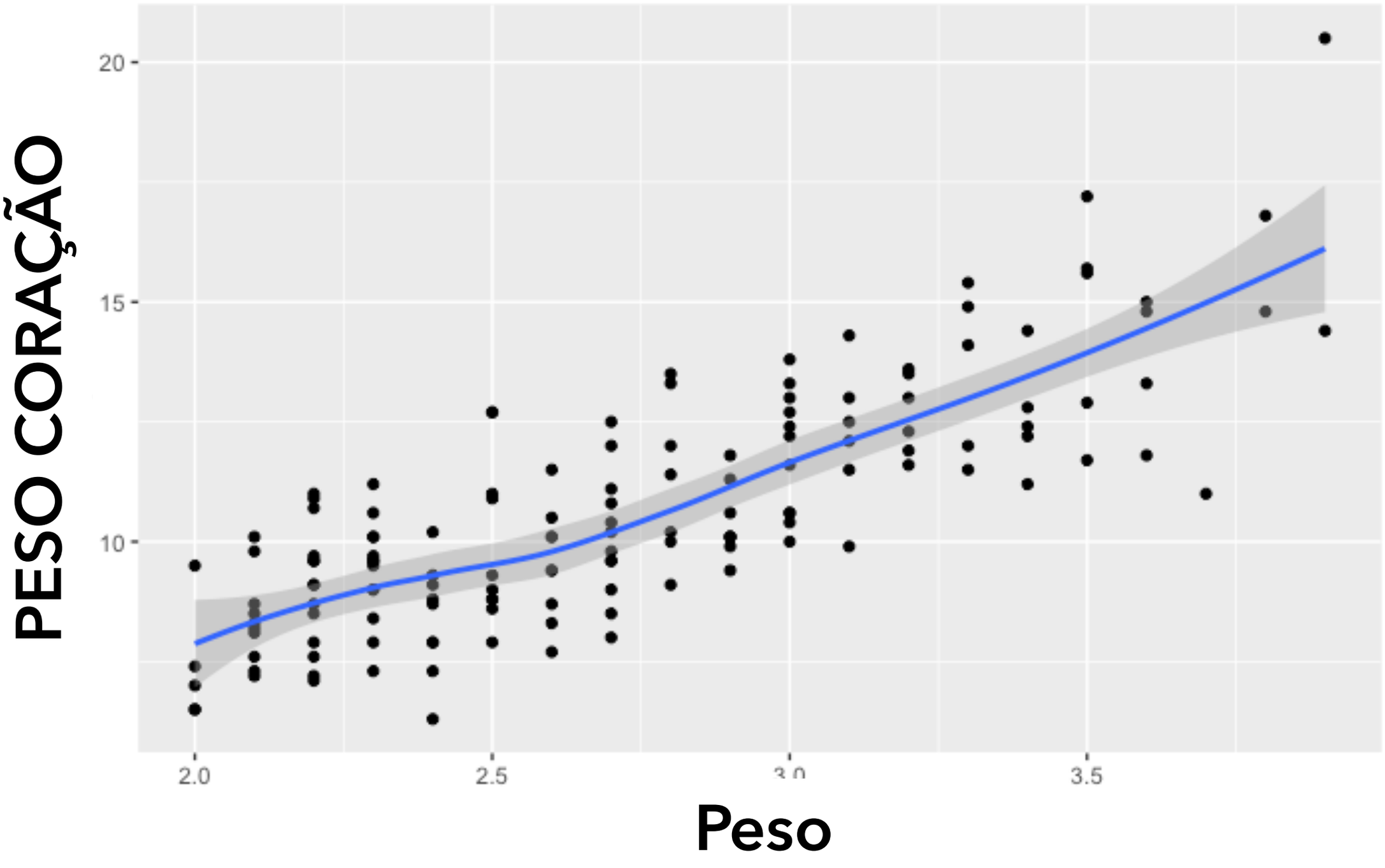
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

EXEMPLO DE PROBLEMA – ÁREA (BIOLOGIA) – TIPO: REGRESSÃO

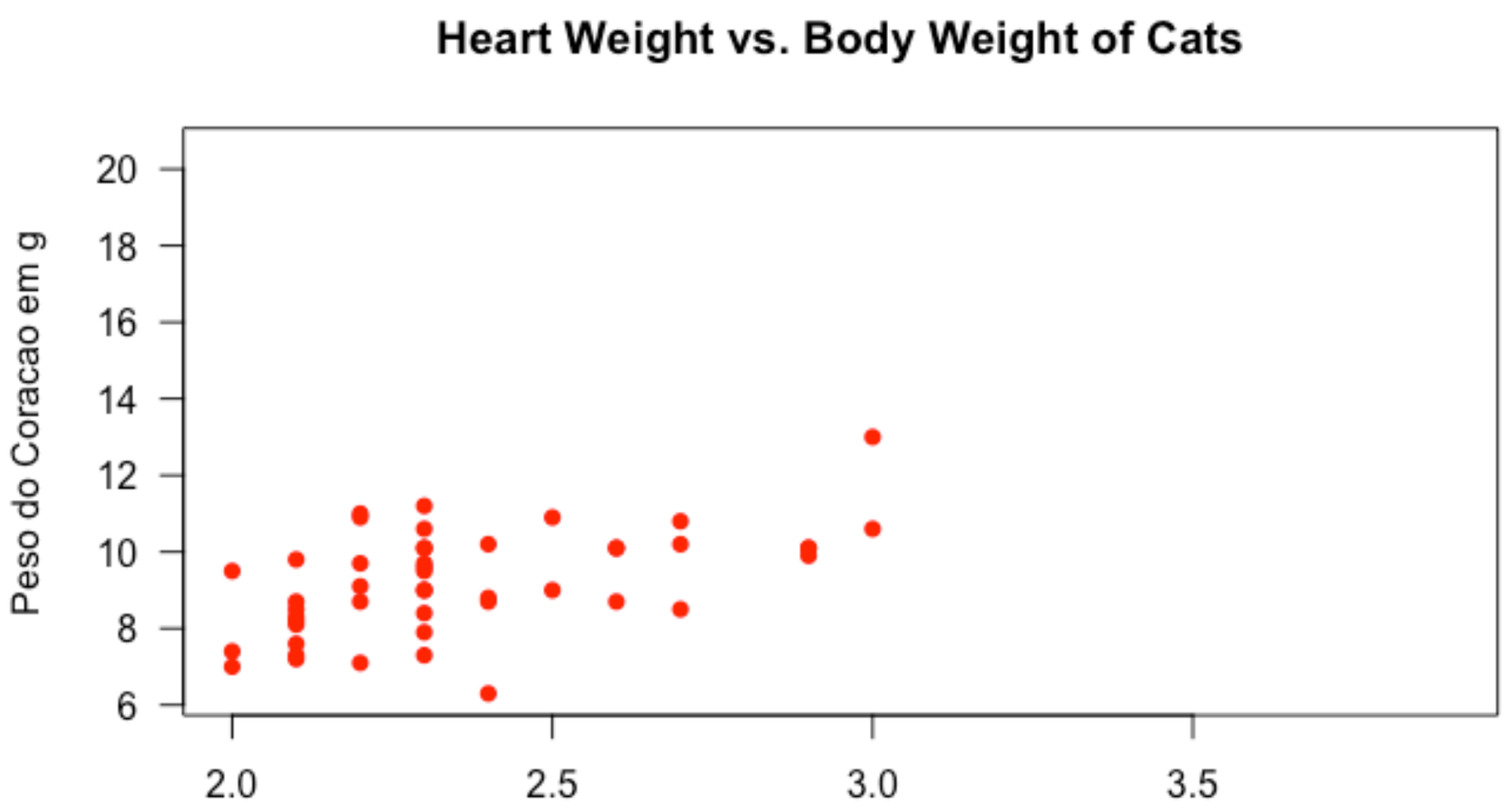


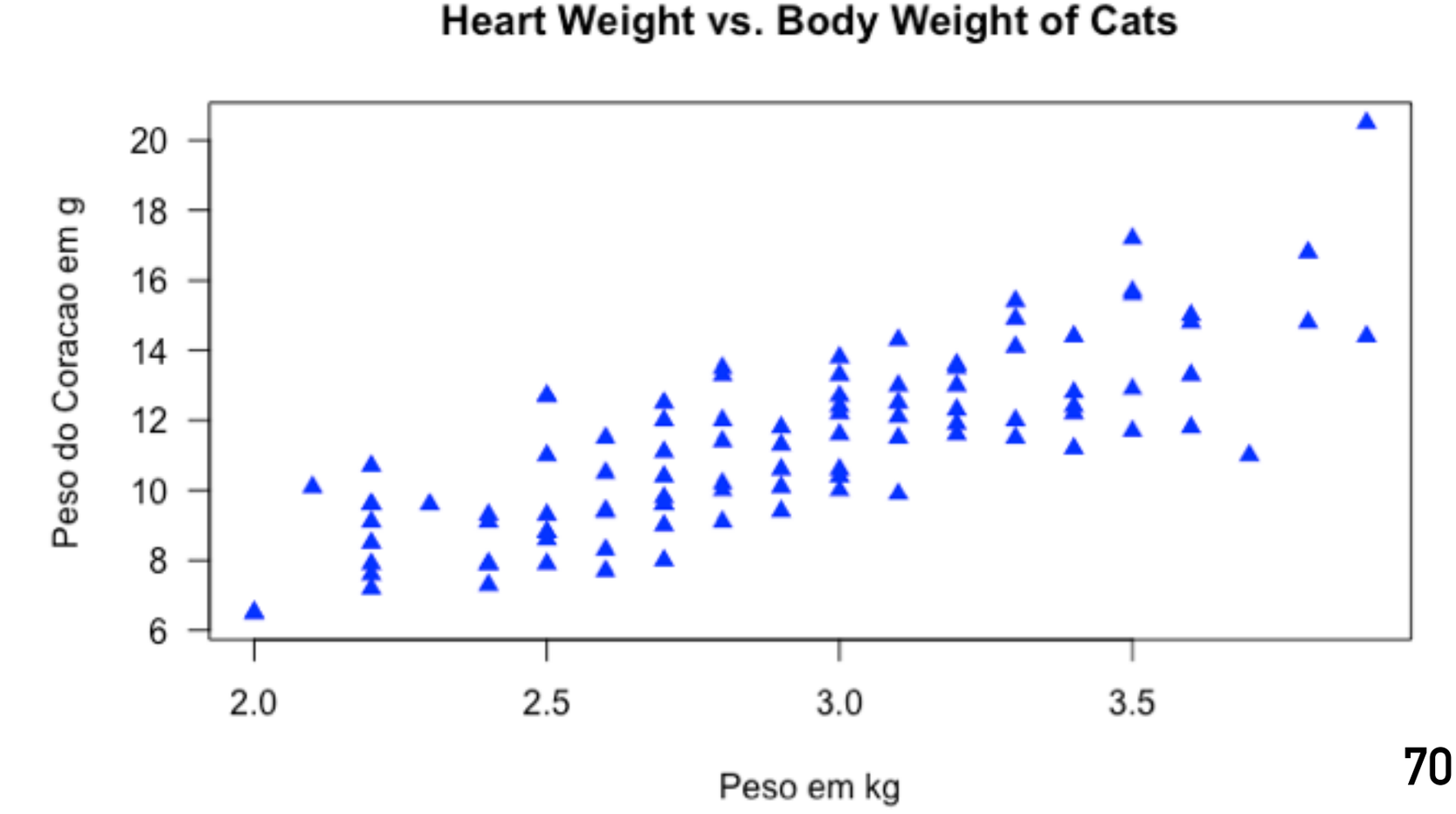
ggplot(a,aes(aBwt,aHwt))+geom_point()

EXEMPLO DE PROBLEMA – ÁREA (BIOLOGIA) – TIPO: REGRESSÃO



ggplot(a,aes(aBwt,aHwt))+geom_point()+geom_smooth()


Heart Weight vs. Body Weight of Cats
Peso do Coracao em g
20
18
16
14
12
10
8
6
2.0
2.5
3.0
3.5


Heart Weight vs. Body Weight of Cats
Peso do Coracao em g
20
18
16
14
12
10
8
6
2.0
2.5
3.0
3.5
Peso em kg

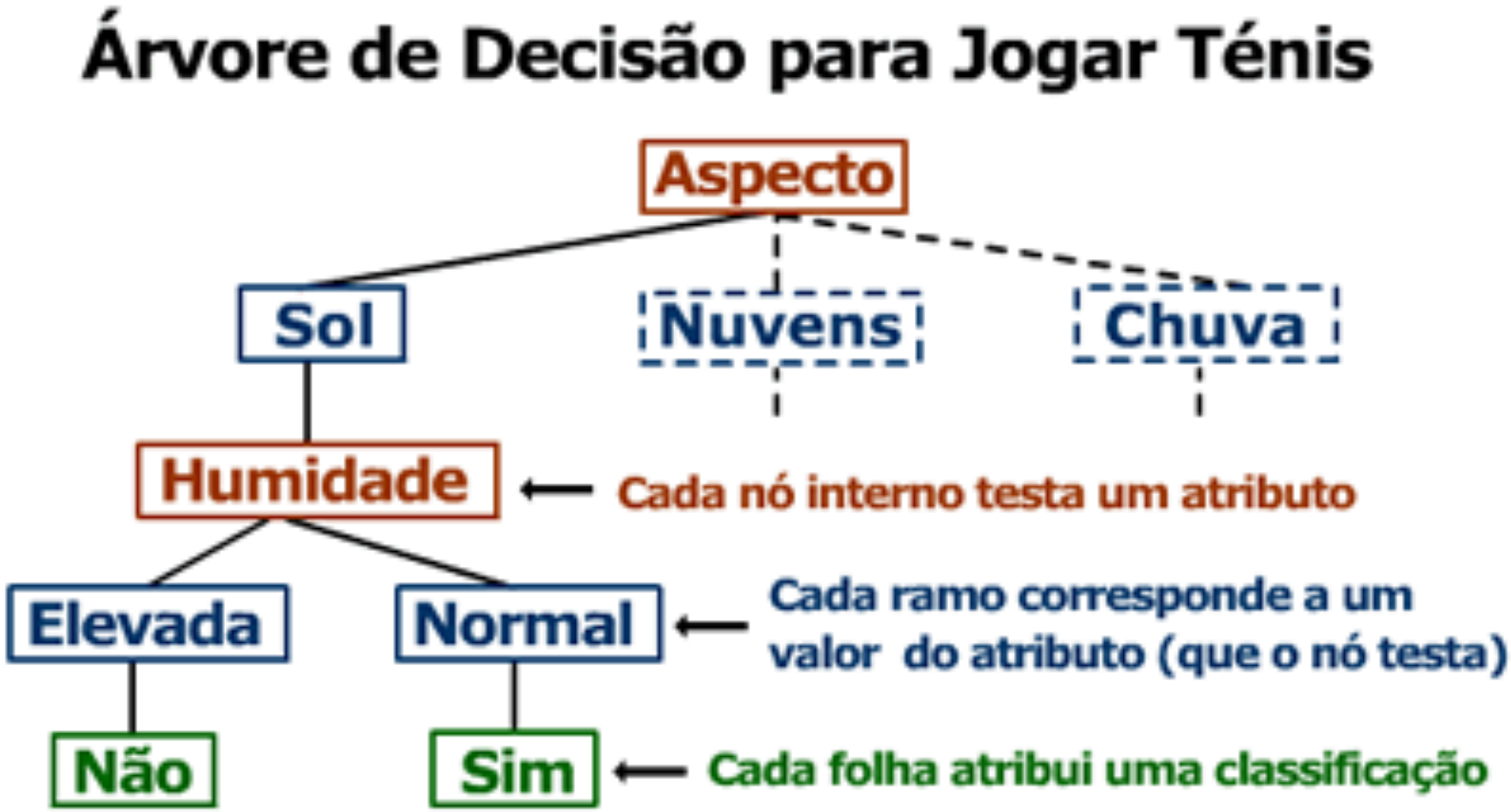
EXEMPLO PREDIZER PREÇOS DE CASAS EM BOSTON

EXEMPLO DE PROBLEMA – PREÇO DE IMÓVEIS EM BOSTON

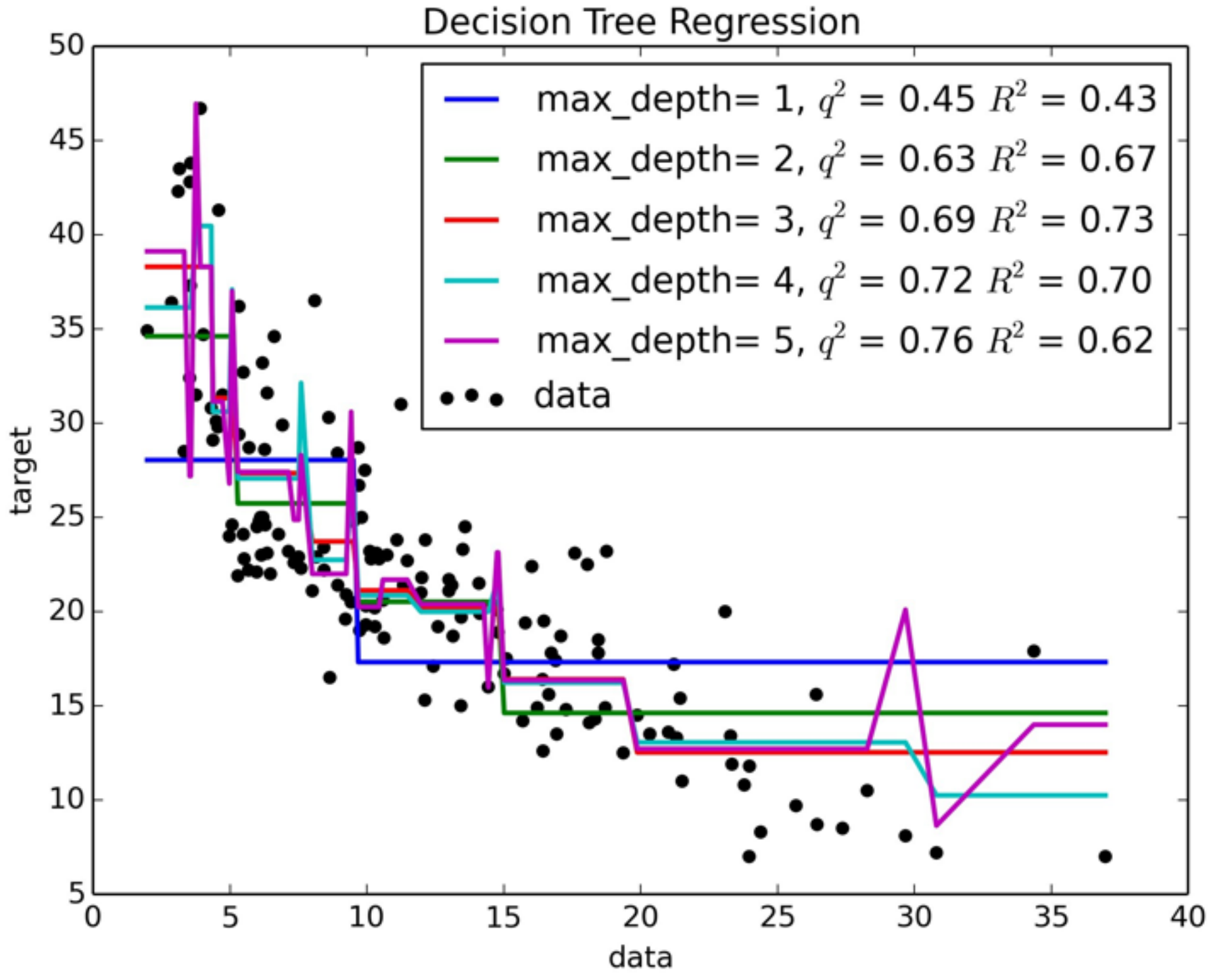
- **`RM'` -Média do número de quartos**
- **`'LSTAT'` percentual de proprietários considerados "lower class" (working poor).**
- **`'PTRATIO'` razão do número de estudantes por professor no bairro**

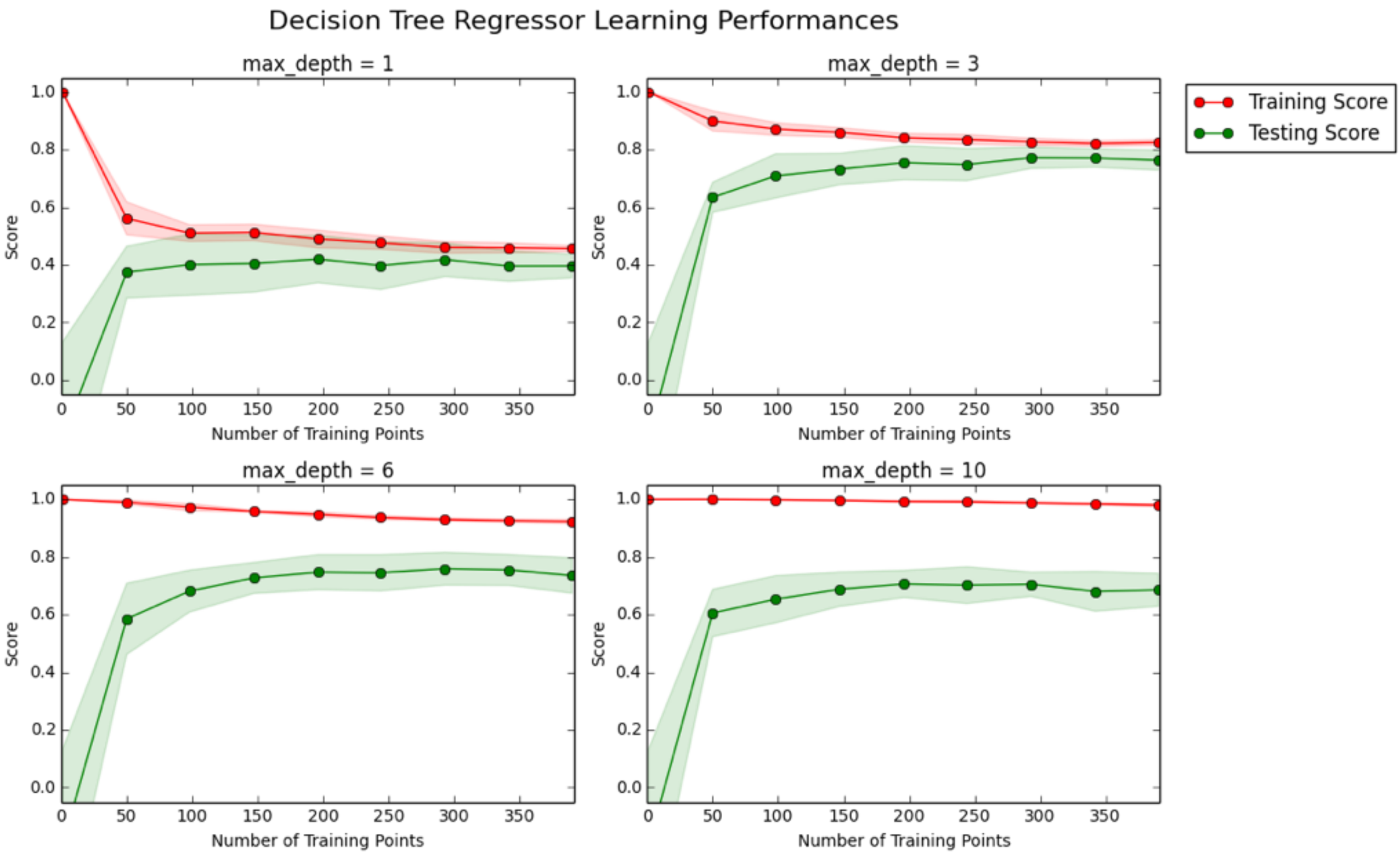
RM	LMRATIO	PTRATIO	PREÇO
6.575	4.98	15.3	504000
6.421	9.14	17.8	453600
7.185	4.03	17.8	728700
6.998	2.94	18.7	701400
7.147	5.33	18.7	760200
6.43	5.21	18.7	602700
6.012	12.43	15.2	480900
6.172	19.15	15.2	569100
5.631	29.93	15.2	346500
6.004	17.1	15.2	396900
6.377	20.45	15.2	315000
6.009	13.27	15.2	396900
5.889	15.71	15.2	455700
5.949	8.26	21	428400
6.096	10.26	21	382200

ARVORES DE DECISÃO



PREÇOS DAS CASAS EM BOSTON




Decision Tree Regressor Learning Performances
max_depth = 1
max_depth = 3
max_depth = 6
max_depth = 10
Score
Number of Training Points
Training Score
Testing Score
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350

CÓDIGO DE REGRESSÃO NO SKLEARN

```
    regressor = DecisionTreeRegressor(random_state=42)
```

```
   params =  {"max_depth": [1,2,3,4,5,6,7,8,9,10]}
```

```
    # TODO: Create the grid search object
    grid = GridSearchCV(regressor,
                                    param_grid=params, scoring =
scoring_fnc, cv = cv_sets)

    # Fit the grid search object to the data to compute the optimal
model
    grid = grid.fit(X, y)
```

PREDICTING

Feature	Client 1	Client 2	Client 3
Total number of rooms in home	5 rooms	4 rooms	8 rooms
Neighborhood poverty level (as %)	17%	32%	3%
Student-teacher ratio of nearby schools	15-to-1	22-to-1	12-to-1

```
# Produce a matrix for client data
client_data = [[5, 17, 15], # Client 1
               [4, 32, 22], # Client 2
               [8, 3, 12]]  # Client 3

# Show predictions
for i, price in
enumerate(reg.predict(client_data)):
```

Predicted selling price for Client 1's home: $414,473.68
Predicted selling price for Client 2's home: $214,302.44
Predicted selling price for Client 3's home: $910,700.00

EXEMPLO INTERVENÇÃO DE ESTUDANTES

INTERVENÇÃO DE ESTUDANTES

```
Feature values:
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher
1     GP   F   17       U     GT3       T     1     1  at_home     other
2     GP   F   15       U     LE3       T     1     1  at_home     other
3     GP   F   15       U     GT3       T     4     2   health  services
4     GP   F   16       U     GT3       T     3     3    other     other

    ...    higher internet  romantic  famrel  freetime goout Dalc Walc health  \
0   ...       yes       no        no       4         3     4    1    1       3
1   ...       yes      yes        no       5         3     3    1    1       3
2   ...       yes      yes        no       4         3     2    2    3       3
3   ...       yes      yes       yes       3         2     2    1    1       5
4   ...       yes       no        no       4         3     2    1    2       5

  absences
0        6
1        4
2       10
3        2
4        4
```

PREPARANDO OS DADOS

```
# If data type is non-numeric, replace all yes/no values with 1/0
  if col_data.dtype == object:
      col_data = col_data.replace(['yes', 'no'], [1, 0])
```

SEPARANDO DADOS DE TREINO E TESTE

X_train, X_test, y_train, y_test = **train_test_split**(X_all, y_all, stratify=y_all, train_size=train_size,test_size=0.24)

INICIALIZANDO MODELOS

```
clf_A = svm.SVC(random_state=42)

clf_B = tree.DecisionTreeClassifier(random_state=42)

clf_C = AdaBoostClassifier(tree.DecisionTreeClassifier(max_depth=1),
                           algorithm="SAMME",
                           n_estimators=300,random_state=42)
clf_D=KNeighborsClassifier(n_neighbors=3)

clf_E= GaussianNB()

clf_F=RandomForestClassifier(n_estimators=100,random_state=42)

clf_G=LogisticRegression(random_state=42)
```

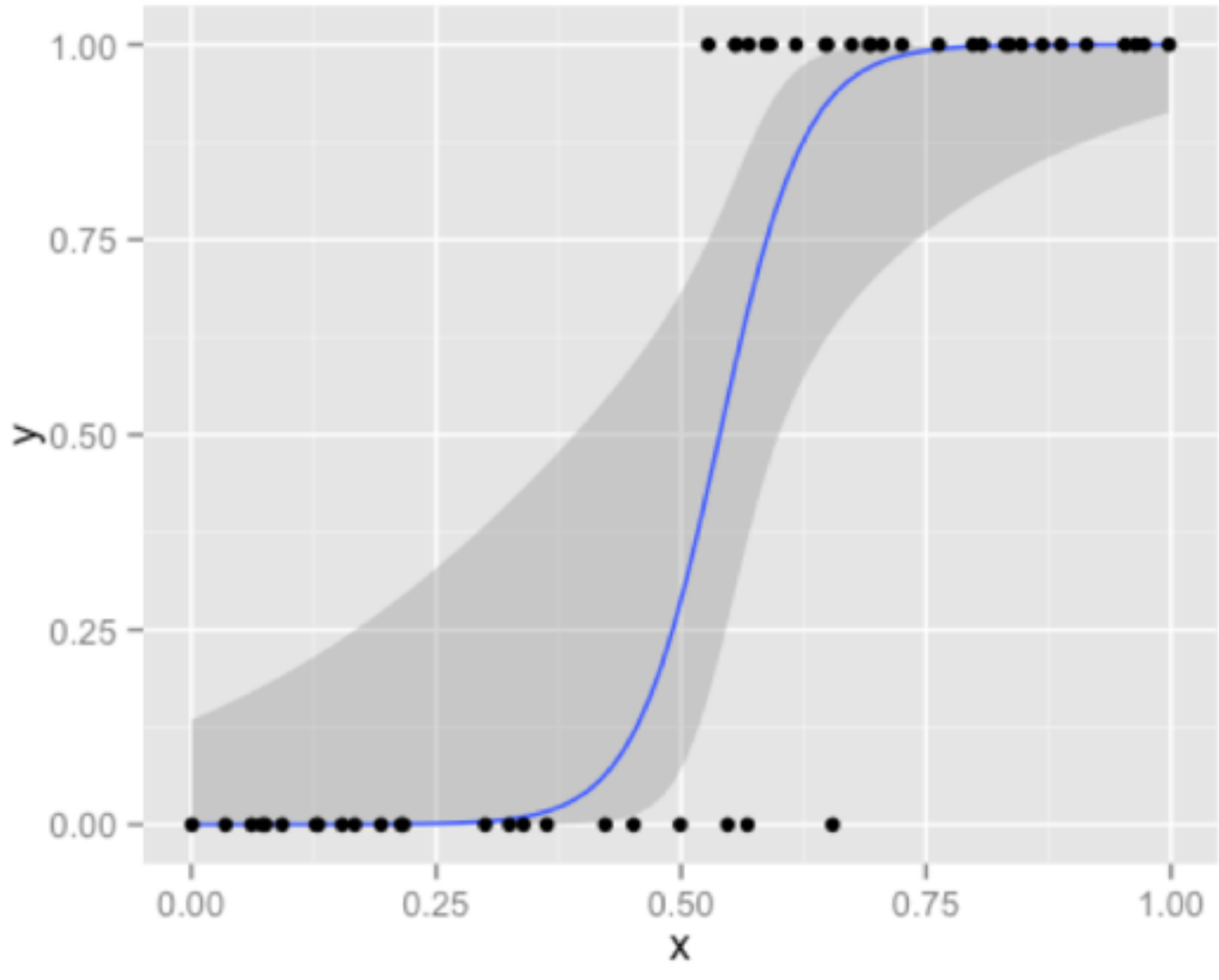
COMPARAÇÃO DE MODELOS

Logistic Regression Pros:

- **Implementação eficiente**

Logistic Regression Cons:

- **Não é performático com muitas features**



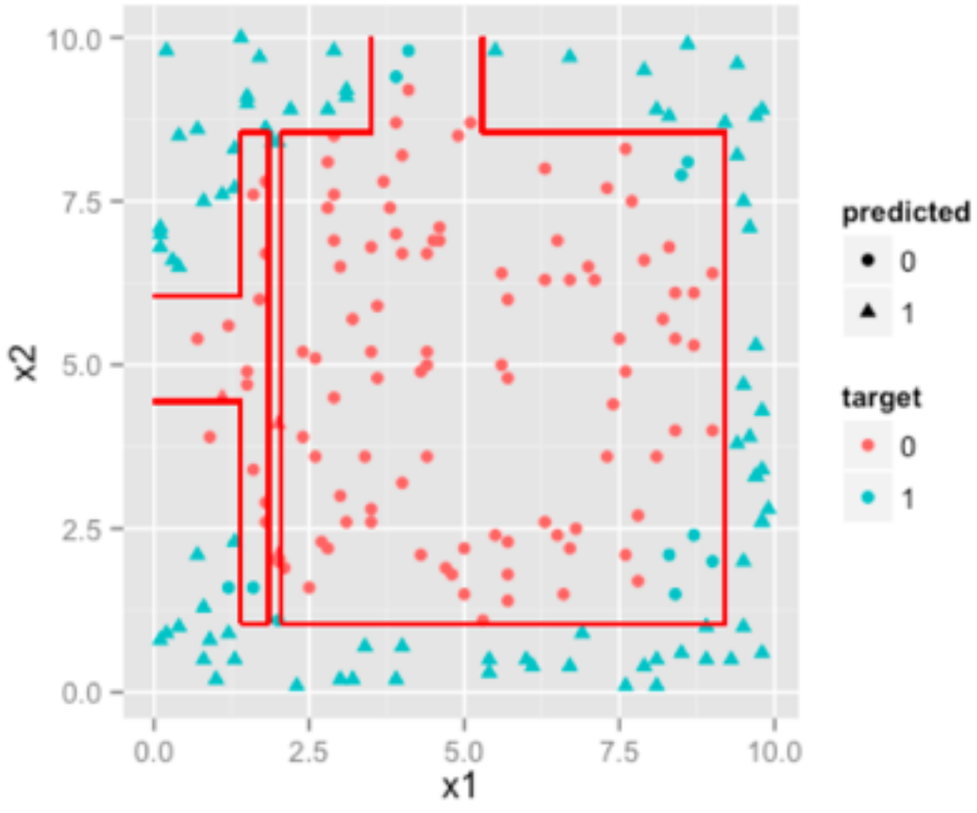
COMPARAÇÃO DE MODELOS

Decision Trees Pros:

- **Regras de decisão intuitivas**
- **Pode utilizar campos não lineares**

Decision Trees Cons:

- **Alto Bias [Random Forests pode ser a solução]**
- **Sem ranking score**



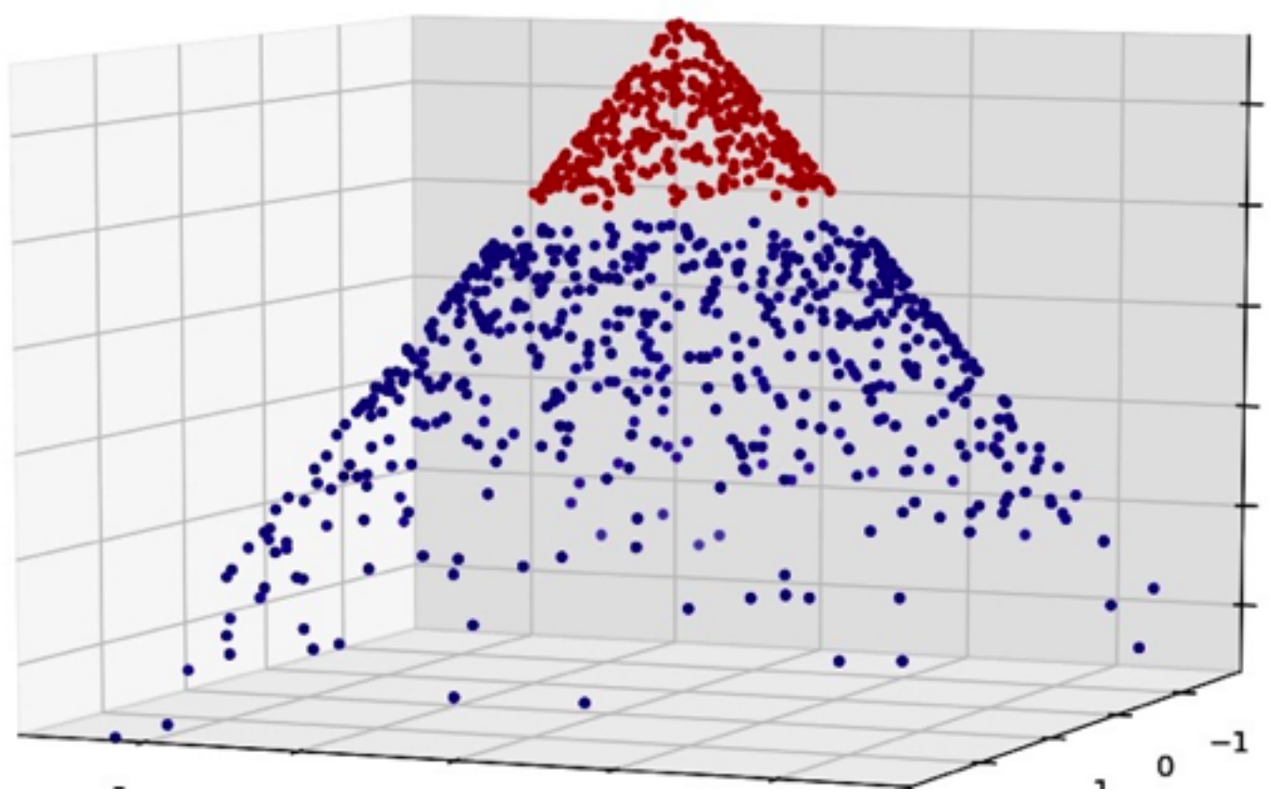
COMPARAÇÃO DE MODELOS

SVM Pros:

- **Pode lidar com um grande número de features**

SVM Cons:

- **Não é performático em um dataset com um maior número de linhas**



PERFORMANCE DOS CLASSIFICADORES

Classifer 1 - SVM

Training Set Size	Training Time	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0019	0.0007	0.7952	0.8050
200	0.0045	0.0013	0.7879	0.8050
300	0.0066	0.0016	0.8024	0.8050

Classifer 2 - Logistic Regression

Training Set Size	Training Time	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0011	0.0001	0.7952	0.8050
200	0.0013	0.0001	0.7879	0.8050
300	0.0017	0.0001	0.8024	0.8050

Classifer 3 - RandomForestClassifier

Training Set Size	Training Time	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.1594	0.0088	1.0000	0.7639
200	0.1543	0.0109	1.0000	0.7500
300	0.1876	0.0113	1.0000	0.7552

SEGMENTANDO FORNECEDORES

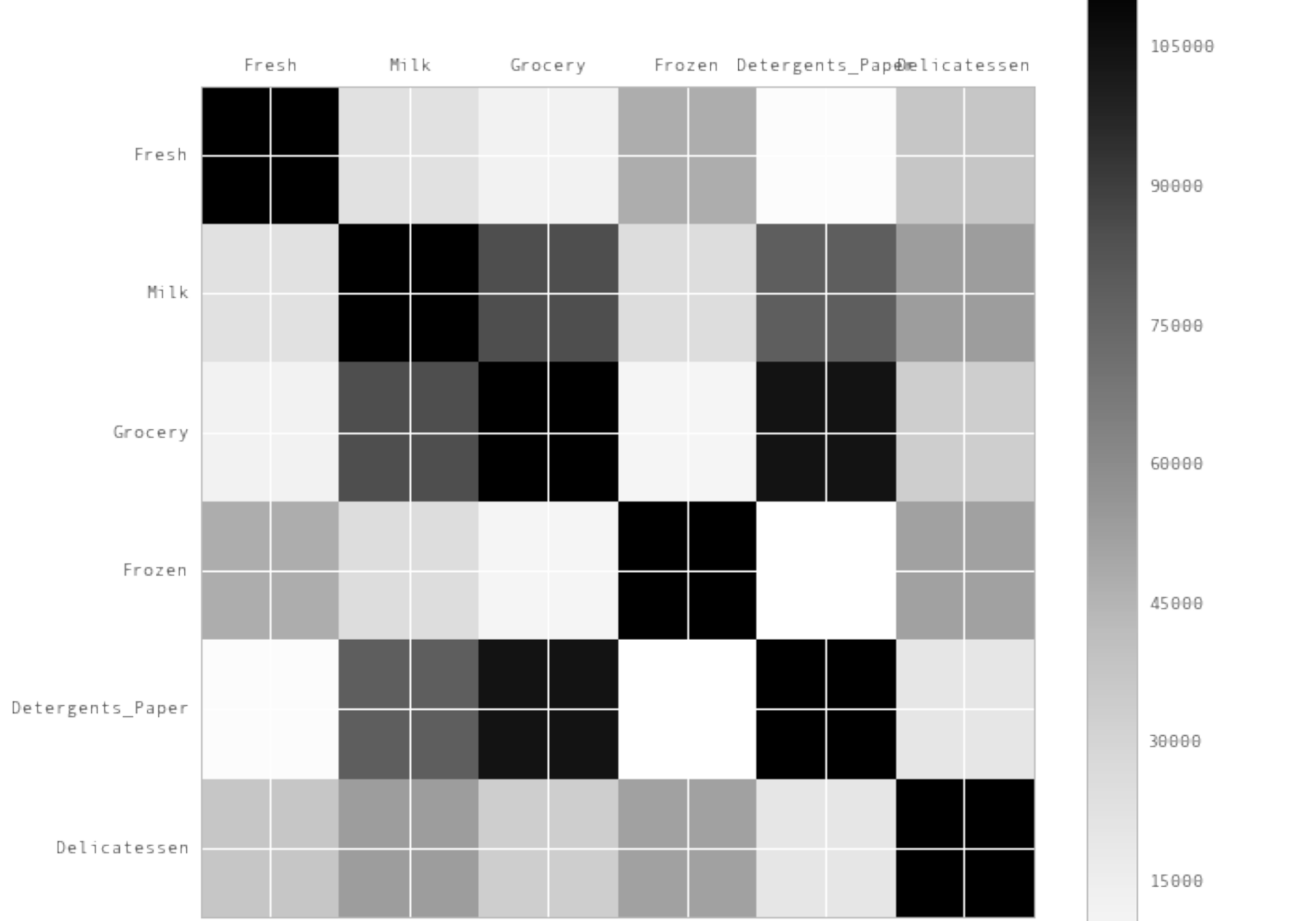
BANCO DE DADOS DE PRODUTOS

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

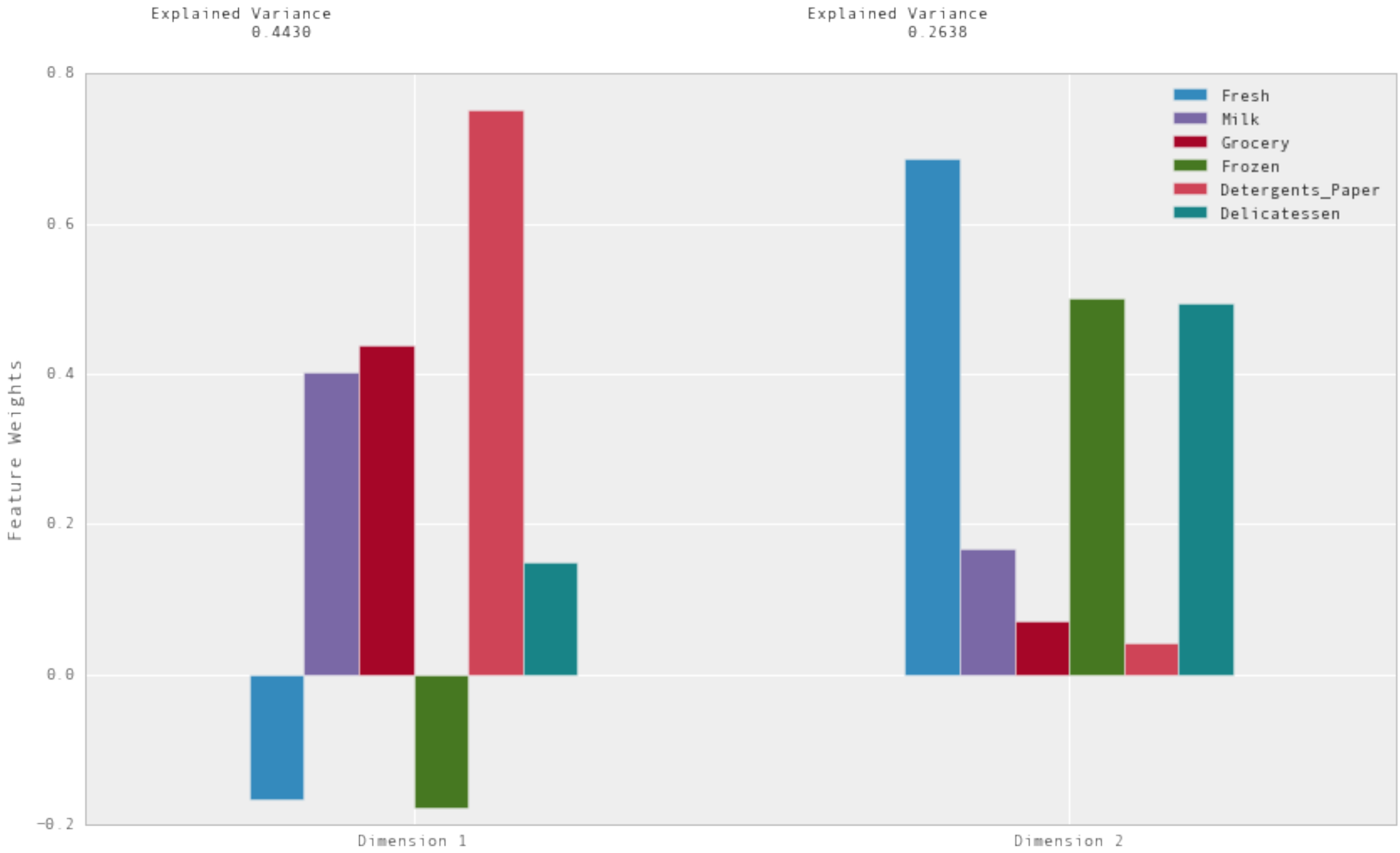
VERIFICANDO CORRELACIONAMENTO

```
for product in products:

    y_array=data[product]
    new_data = data.drop([product],axis=1)


    X_train, X_test, y_train, y_test = train_test_split( new_data, y_array, test_size=0.25, random_state=

    for reg in regressors:
        reg.fit(X_train,y_train)
        score = reg.score(X_test, y_test)
        print('Product '+product+' Score is '+ str(score))
        parameters=regressors_parameters.get(reg)
        clf = GridSearchCV(reg, parameters)
        clf.fit(X_train,y_train)
        score = clf.score(X_test, y_test)
        print('Product '+product+' Score is '+ str(score)+' after GridSearchCV ')
```

```
Product Fresh Score is -0.333070533605
Product Fresh Score is -0.329449950604 after GridSearchCV
Product Milk Score is 0.173438009379
Product Milk Score is 0.205871721893 after GridSearchCV
Product Grocery Score is 0.699248196675
Product Grocery Score is 0.699248196675 after GridSearchCV
Product Detergents_Paper Score is 0.348777454691
Product Detergents_Paper Score is 0.348777454691 after GridSearchCV
Product Frozen Score is -0.278249148824
Product Frozen Score is -1.30732144534 after GridSearchCV
Product Delicatessen Score is -11.0236279005
Product Delicatessen Score is -9.55743305081 after GridSearchCV
```


Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicatessen
105000
90000
75000
60000
45000
30000
15000

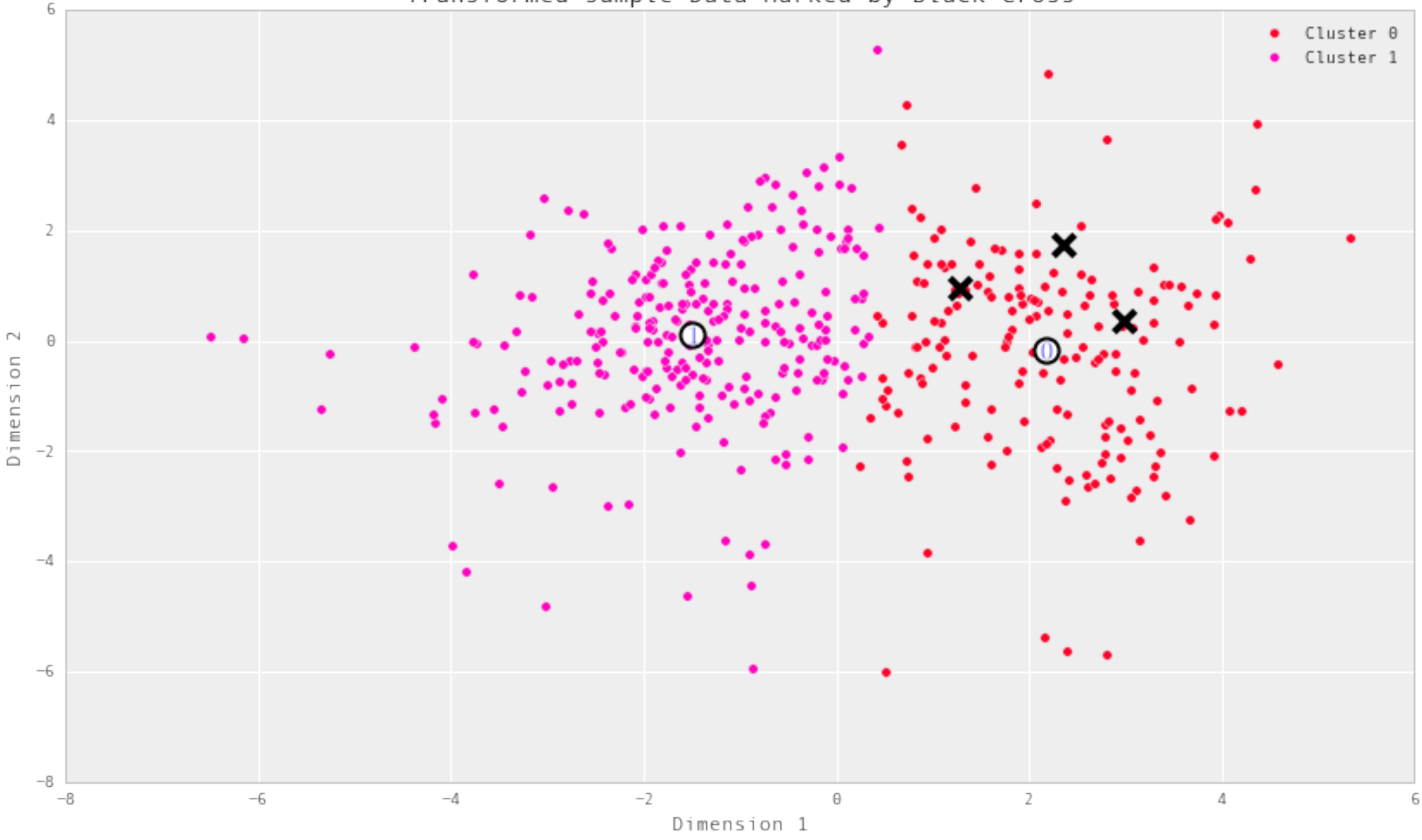
PCA –REDUZINDO DIMENSÕES



GERANDO OS CLUSTERS

```
clusterer = KMeans(n_clusters=i,
random_state=29).fit(reduced_data)

preds = clusterer.predict(reduced_data)
```

Cluster Learning on PCA-Reduced Data - Centroids Marked by Number
Transformed Sample Data Marked by Black Cross



SITES ONDE CONSEGUIR INFORMAÇÃO

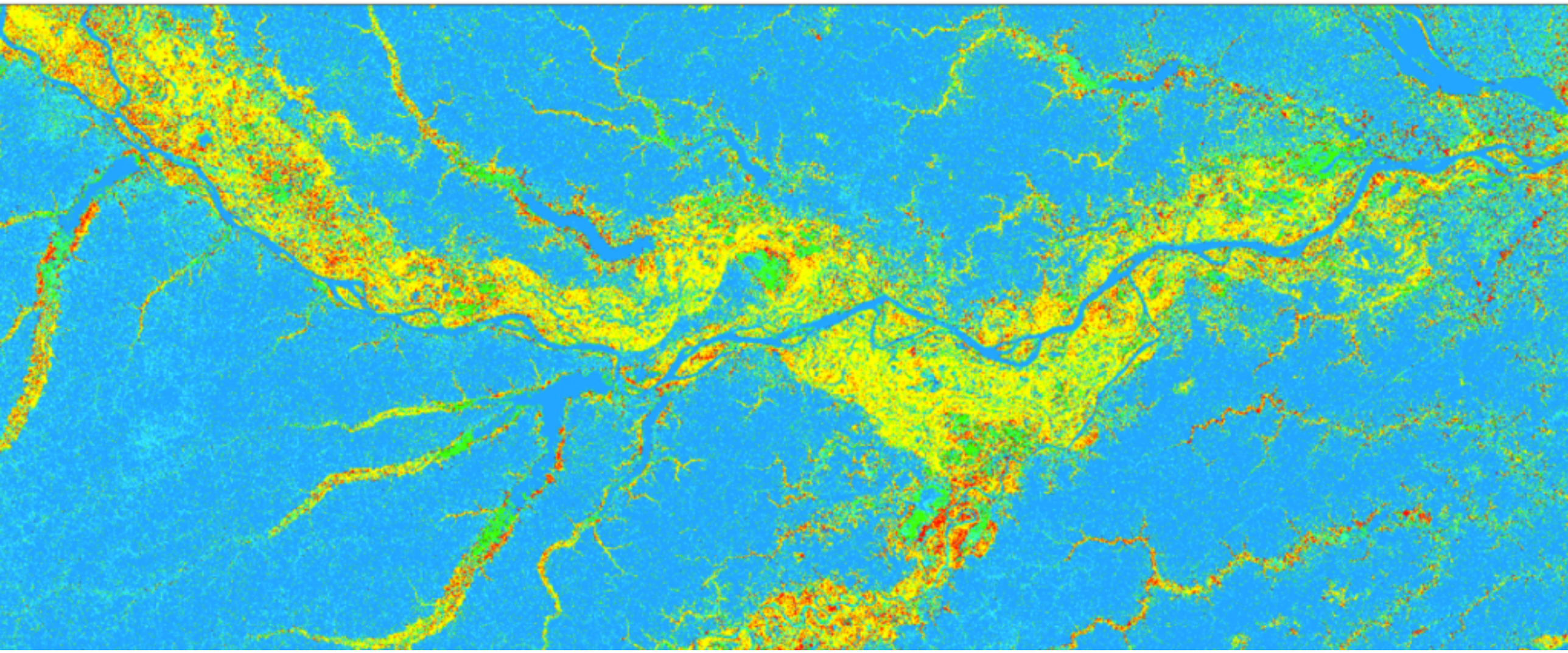
DESMISTIFICANDO A CIÊNCIA DE DADOS

HTTPS://ENSINANDOMAQUINASBLOG.WORDPRESS.COM

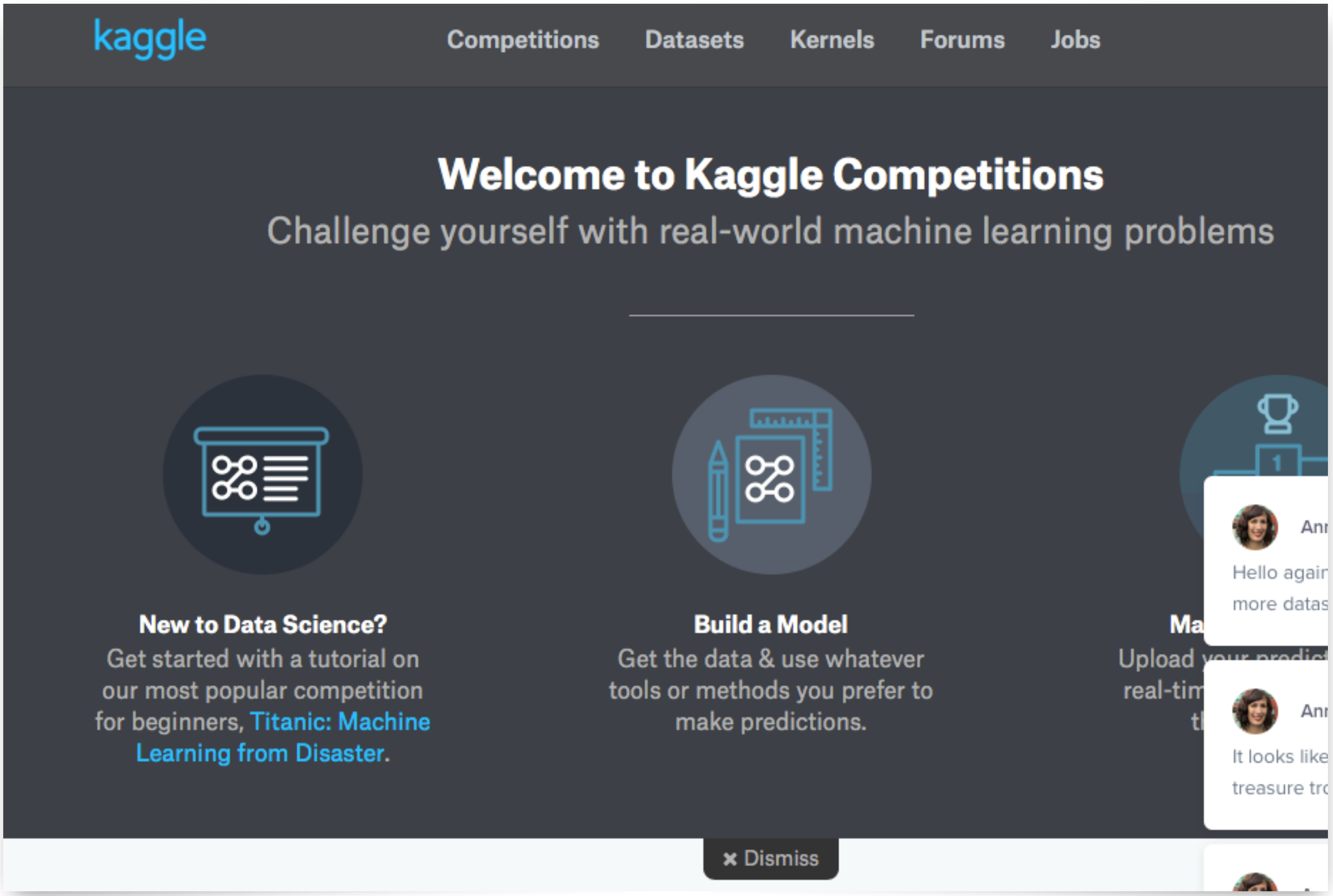
INTELIGÊNCIA COMPUTACIONAL PARA MINERAÇÃO DE DADOS

ENSINANDO MÁQUINAS

SOBRE



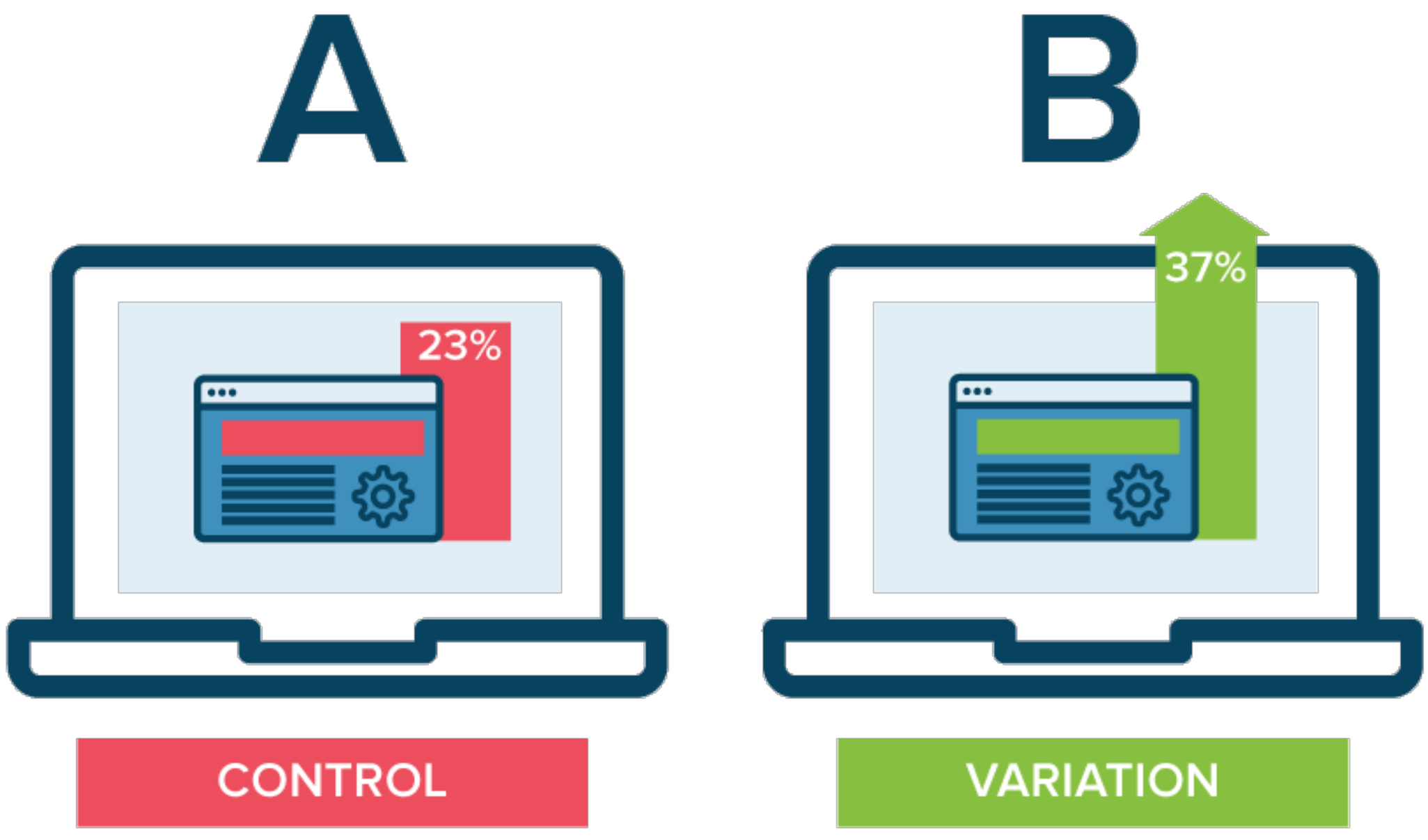
KAGGLE


kaggle
Competitions
Datasets
Kernels
Forums
Jobs
Welcome to Kaggle Competitions
Challenge yourself with real-world machine learning problems
New to Data Science?
Get started with a tutorial on our most popular competition for beginners, Titanic: Machine Learning from Disaster.
Build a Model
Get the data & use whatever tools or methods you prefer to make predictions.
Hello again
more datas
It looks like
treasure tr
Dismiss

RELATÓRIOS

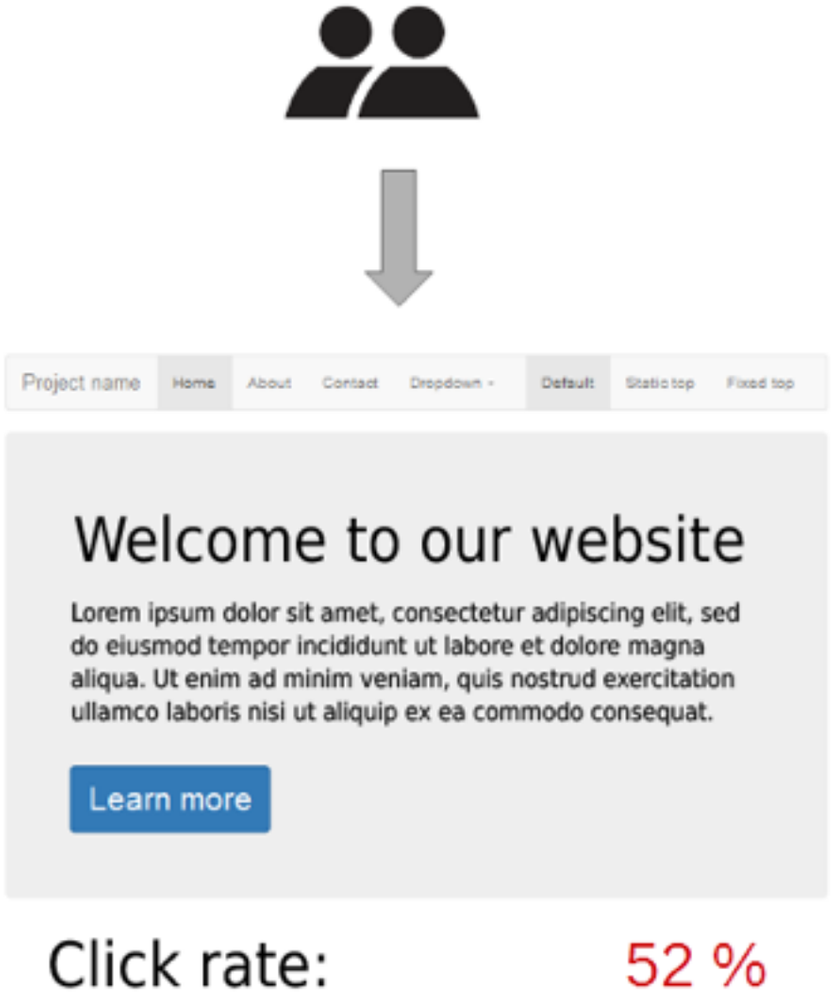
- Notebooks online: iPython, Jupyter
- permitem a criação de documentos
- interativos em várias linguagens de análise.
- ¨ **Reproducible Research!**

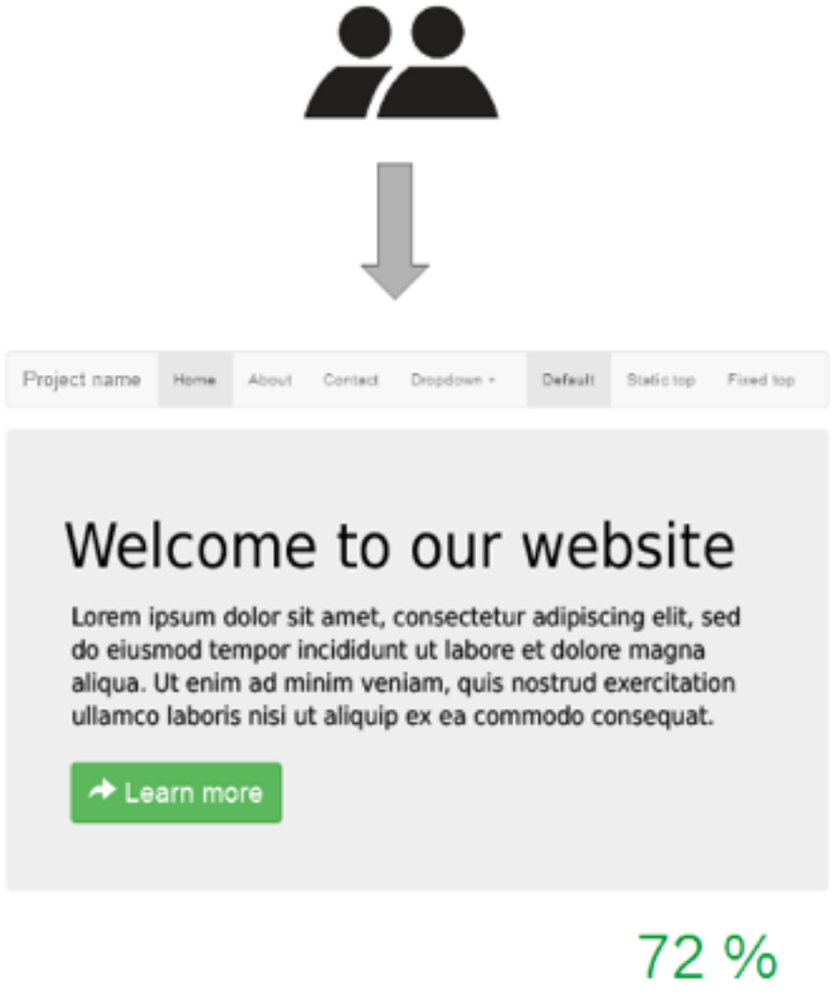
TESTE A/B



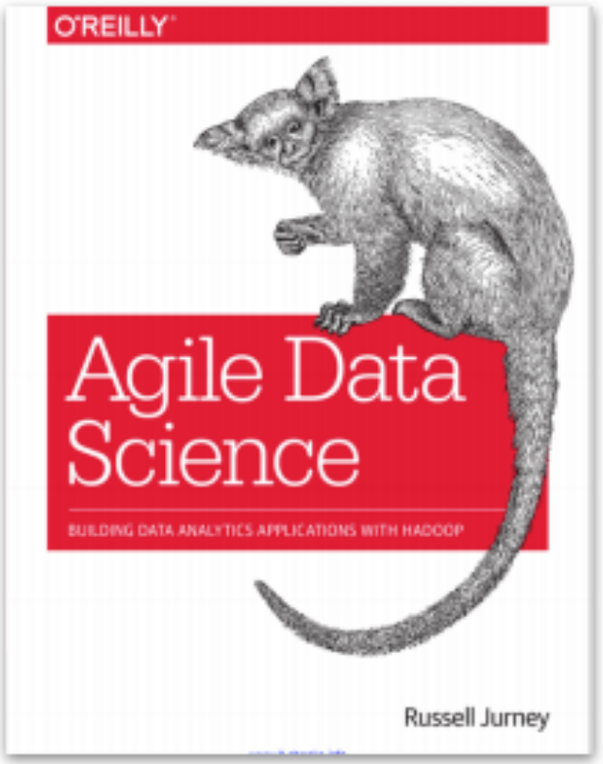
TESTE A/B

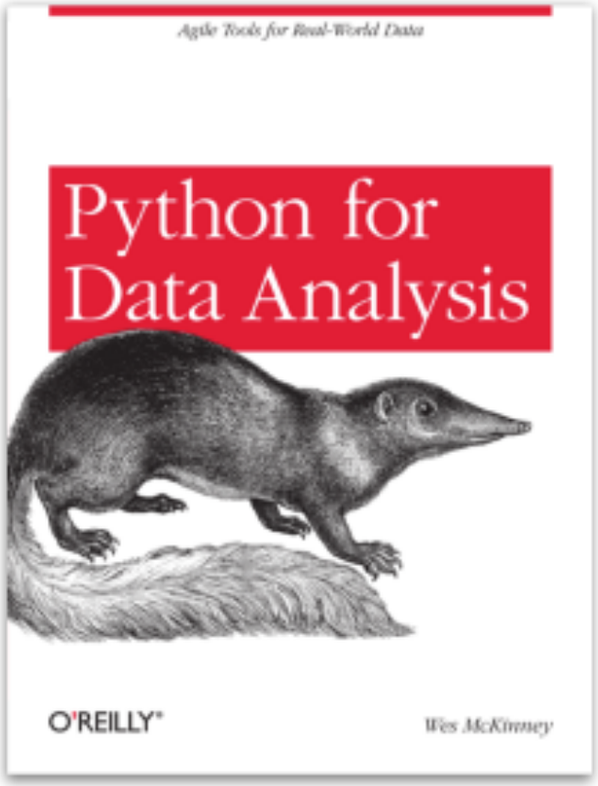
Teste A/B é um método de teste onde se comparam duas práticas, A e B, em que estes são o controle e o tratamento de uma experiência controlada, com o objetivo de melhorar a percentagem de aprovação.

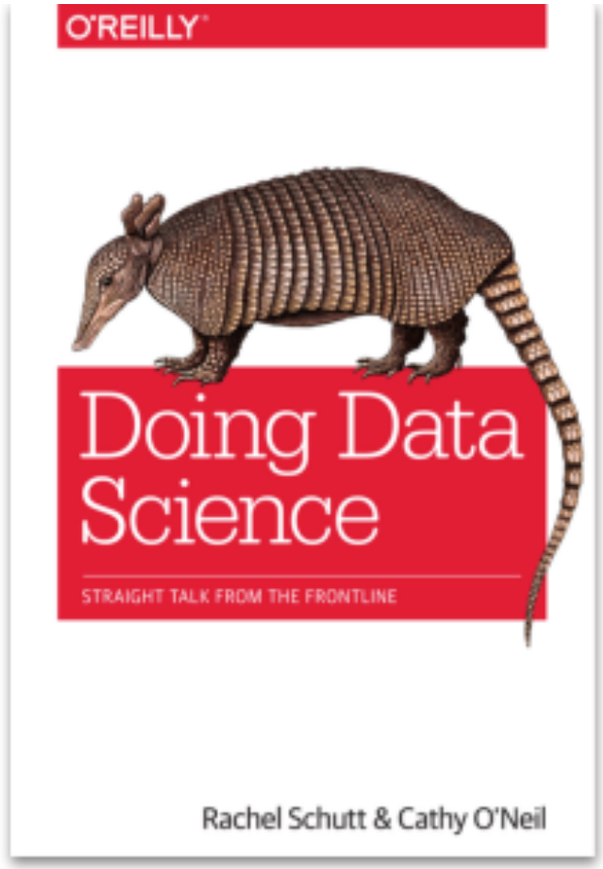


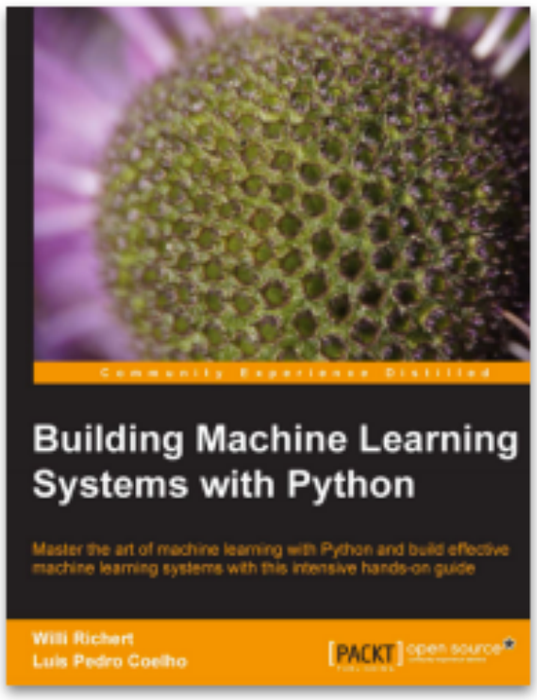


LIVROS


O'REILLY
Agile Data Science
BUILDING DATA ANALYTICS APPLICATIONS WITH HADOOP
Russell Jurney


Agile Tools for Real-World Data
Python for Data Analysis
O'REILLY
Wes McKinney


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Rachel Schutt & Cathy O'Neil


Community Experience Distilled
Building Machine Learning Systems with Python
Master the art of machine learning with Python and build effective machine learning systems with this intensive hands-on guide
Willi Richert
Luis Pedro Coelho
PACKT open source


Quick answers to common problems
Python Data Visualization Cookbook
Over 60 recipes that will enable you to learn how to create attractive visualizations using Python's most popular libraries
Igor Milovanović
PACKT open source

- Dúvidas: naubergois@gmail.com