

# Introdução ao Aprendizado de Máquina

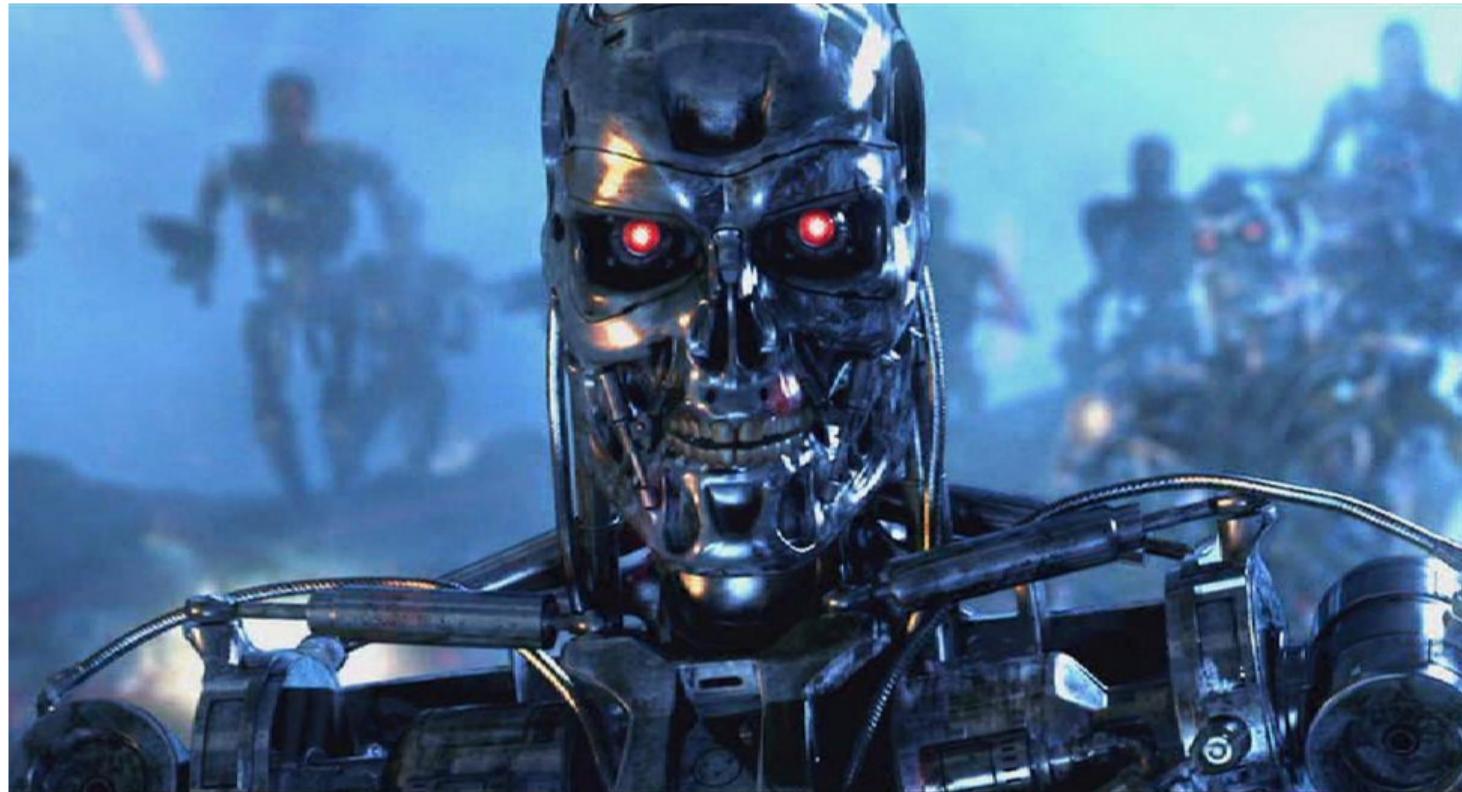
Prof. Erneson A. Oliveira

MBA em Ciência de Dados  
Universidade de Fortaleza

1 de Fevereiro de 2020



# Aula 2 - Desafios de Aprendizado de Máquina



# Como aplicar AM?

# Como aplicar AM?



<http://www.python.org>

# Como aplicar AM?



<http://www.jupyter.org>

# Como aplicar AM?



# ANACONDA®

<http://www.anaconda.com>

# Como aplicar AM?



<https://scikit-learn.org>

# Como aplicar AM?



# TensorFlow

<https://www.tensorflow.org>

# Como aplicar AM?



<https://keras.io/>

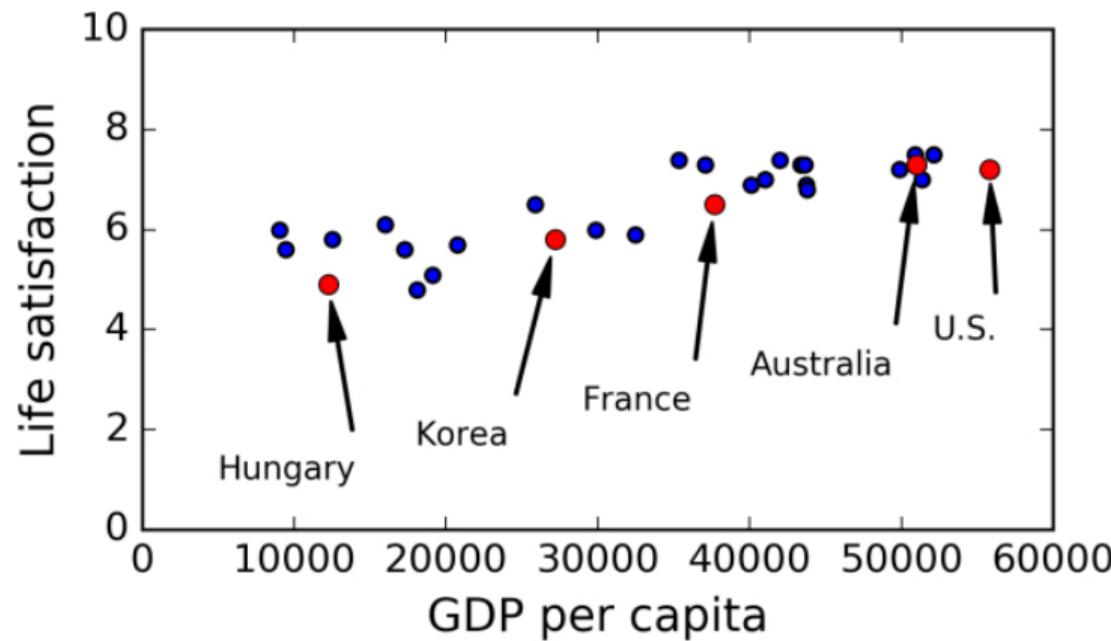
# Exemplo: Dinheiro torna as pessoas felizes?

# Exemplo: Dinheiro torna as pessoas felizes?

- Dados: Renda (FMI) e Satisfação (OCDE).

Country	GDP per capita (USD)	Life satisfaction
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
United States	55,805	7.2

# Exemplo: Dinheiro torna as pessoas felizes?



Existe tendência?

# Exemplo: Dinheiro torna as pessoas felizes?

Sim! A satisfação parece crescer linearmente com a renda.

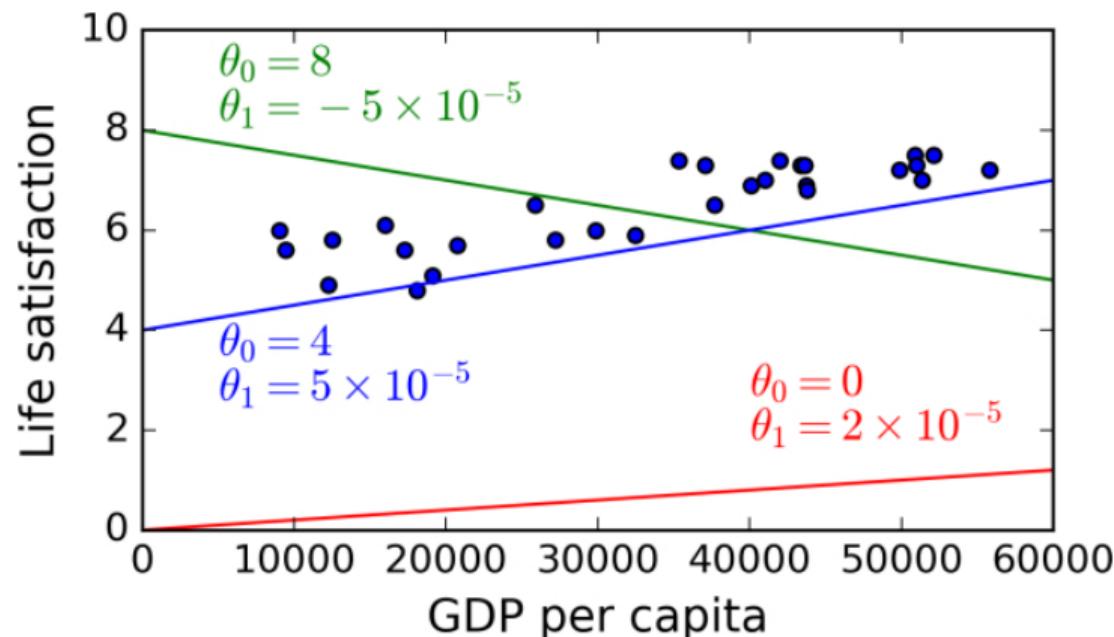
# Exemplo: Dinheiro torna as pessoas felizes?

- Seleção de Modelo: Modelo linear para satisfação ( $S$ ) com apenas o atributo de renda ( $R$ ).

$$S = \theta_0 + \theta_1 R,$$

onde  $\theta_0$  e  $\theta_1$  são parâmetros do modelo.

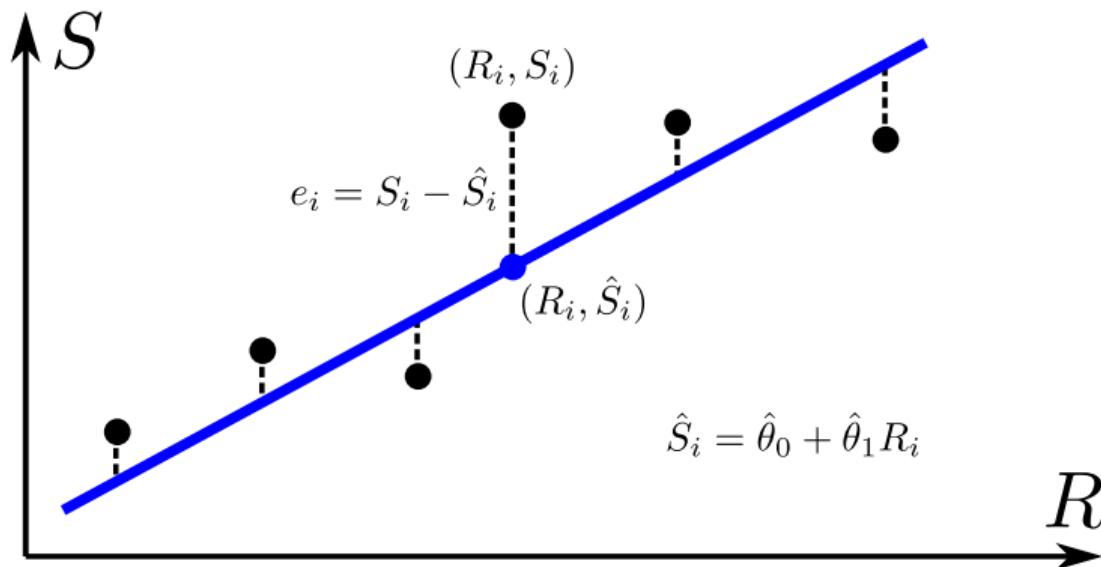
# Exemplo: Dinheiro torna as pessoas felizes?



Existem vários modelos lineares dependendo de  $\theta_0$  e  $\theta_1$

# Exemplo: Dinheiro torna as pessoas felizes?

- ▶ Medida de desempenho: Função Utilidade ou Função Custo.



- ▶ Problema de otimização: Algoritmo de Aprendizagem (AA).

# Exemplo: Dinheiro torna as pessoas felizes?

Para problemas de regressão linear, podemos minimizar a função custo:

$$SSR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (S_i - \hat{S}_i)^2 = \sum_{i=1}^n (S_i - \hat{\theta}_0 - \hat{\theta}_1 R_i)^2,$$

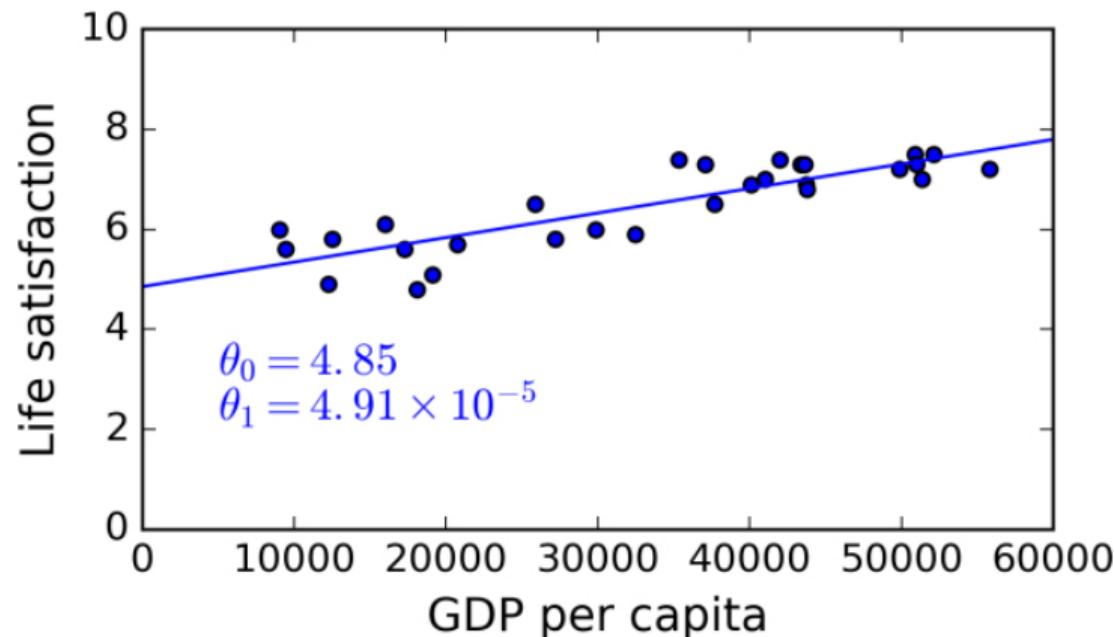
onde  $SSR$  é a soma dos quadrados dos resíduos (Método dos Mínimos Quadrados). Os parâmetros estimados ficam:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (R_i - \langle R \rangle)(S_i - \langle S \rangle)}{\sum_{i=1}^n (R_i - \langle R \rangle)^2} \quad \text{e} \quad \hat{\theta}_0 = \langle S \rangle - \hat{\theta}_1 \langle R \rangle,$$

onde

$$\langle R \rangle = \frac{1}{n} \sum_{i=1}^n R_i \quad \text{e} \quad \langle S \rangle = \frac{1}{n} \sum_{i=1}^n S_i.$$

# Exemplo: Dinheiro torna as pessoas felizes?



Modelo linear que se ajusta melhor ao conjunto de treinamento

Exemplo: Dinheiro torna as pessoas felizes?

Podemos fazer previsões (inferências)?

# Exemplo: Dinheiro torna as pessoas felizes?

Abra o Jupyter Notebook!

Em suma, em AM...

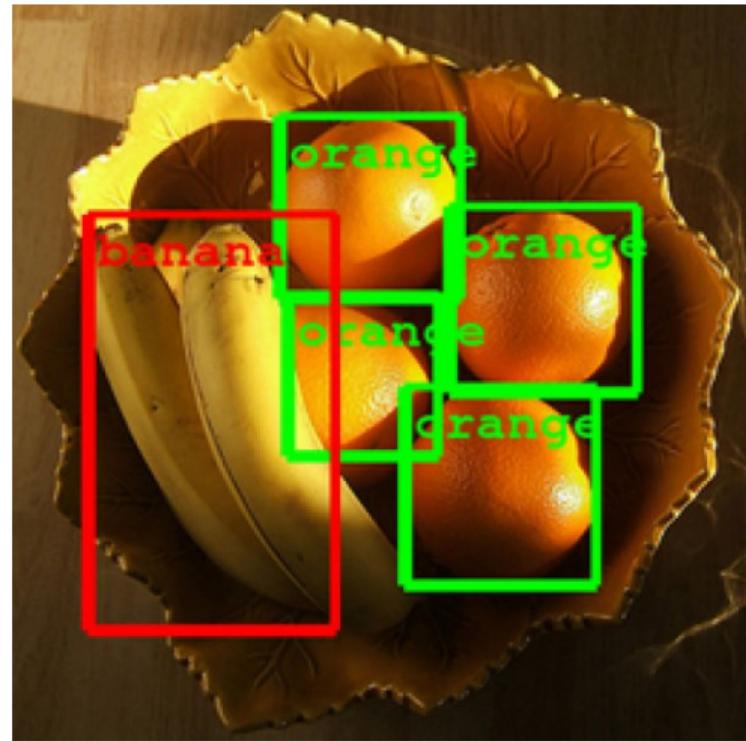
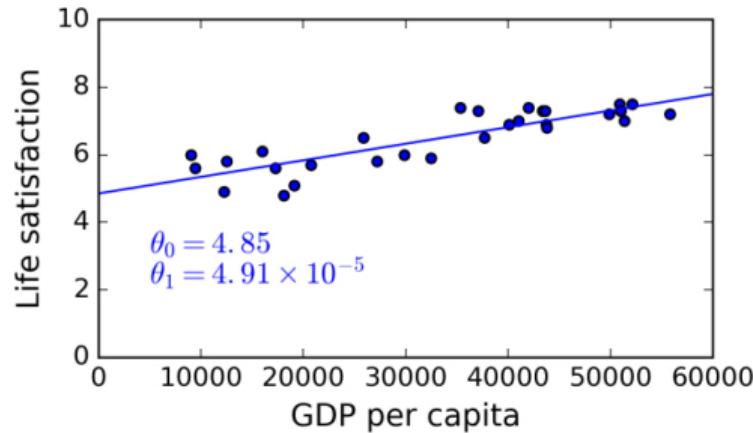
# Em suma, em AM...

- ▶ Estudamos o dado;
- ▶ Definimos como será o SAM;
- ▶ Treinamos o SAM (Encontramos os parâmetros que minimizam/maximiza a função custo/utilidade);
- ▶ Fazemos previsões para novos casos.

# Quais são os principais desafios em AM?

# Quantidade insuficiente de dados no conjunto de treinamento

# Quantidade insuficiente de dados no conjunto de treinamento



Problemas simples × Problemas complicados

# Quantidade insuficiente de dados no conjunto de treinamento

## Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill

Microsoft Research

1 Microsoft Way

Redmond, WA 98052 USA

{mbanko,brill}@microsoft.com

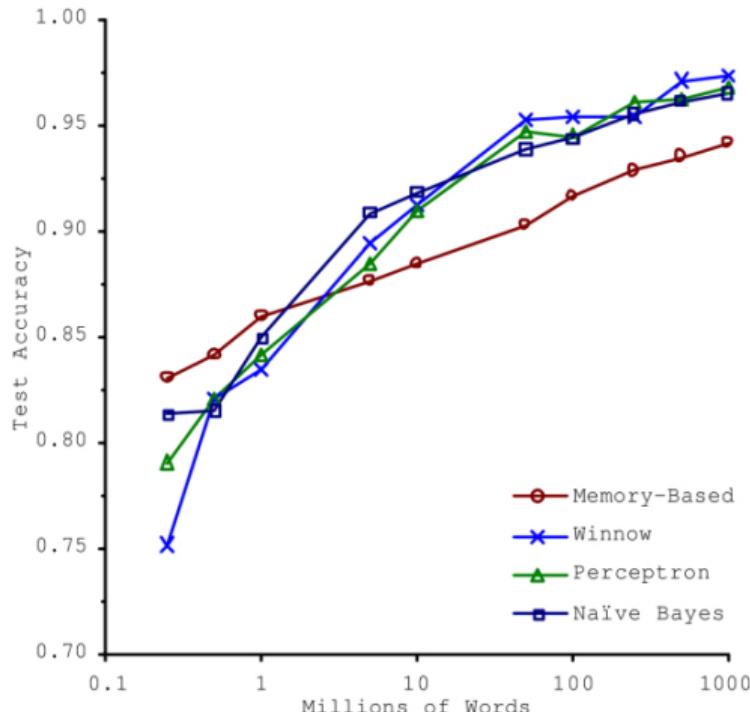
### Abstract

The amount of readily available online text has reached hundreds of billions of words and continues to grow. Yet for most core natural language tasks, algorithms continue to be optimized, tested and compared after training on corpora consisting of only one million words or less. In this paper, we evaluate the performance of different learning methods on a prototypical natural language disambiguation task.

standardization of data sets used within the field, as well as the potentially large cost of annotating data for those learning methods that rely on labeled text.

The empirical NLP community has put substantial effort into evaluating performance of a large number of machine learning methods over fixed, and relatively small, data sets. Yet since we now have access to significantly more data, one has to wonder what conclusions that have been drawn on small data sets may carry over when these learning methods are trained using much larger corpora.

In this paper, we present a study of the



## Dados × Algoritmos

# Quantidade insuficiente de dados no conjunto de treinamento



**EXPERT OPINION**  
Contact Editor: Brian Brannon, bbrannon@computer.org

## The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"<sup>1</sup> examines why so much of physics can be neatly explained with simple mathematical formulas

such as  $f = ma$  or  $e = mc^2$ . Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant math-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

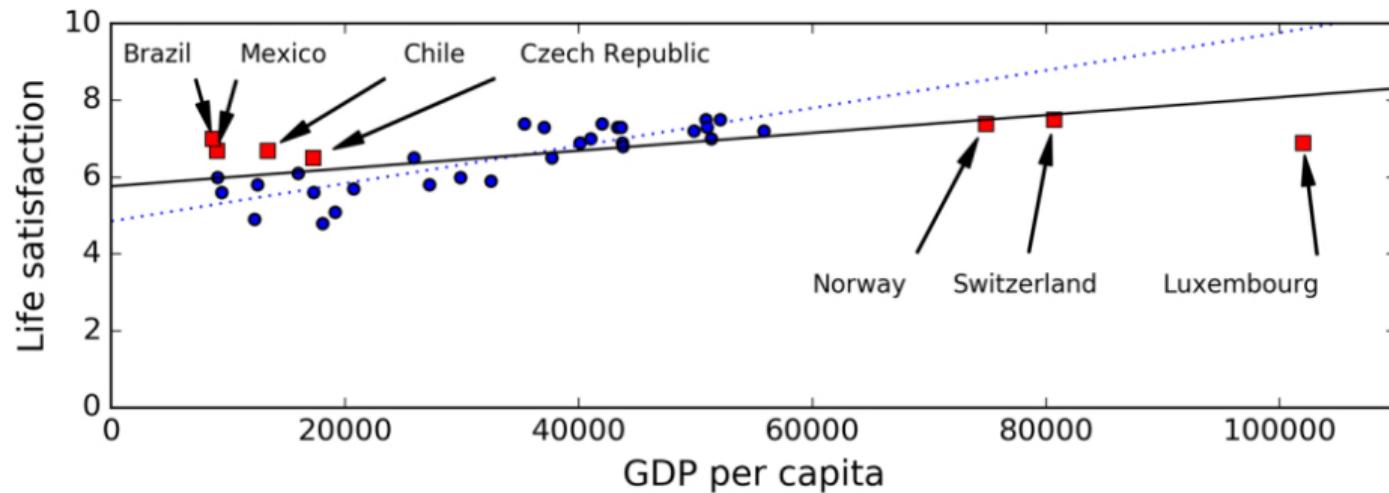
### Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are

## Dados são mais importantes do que algoritmos?

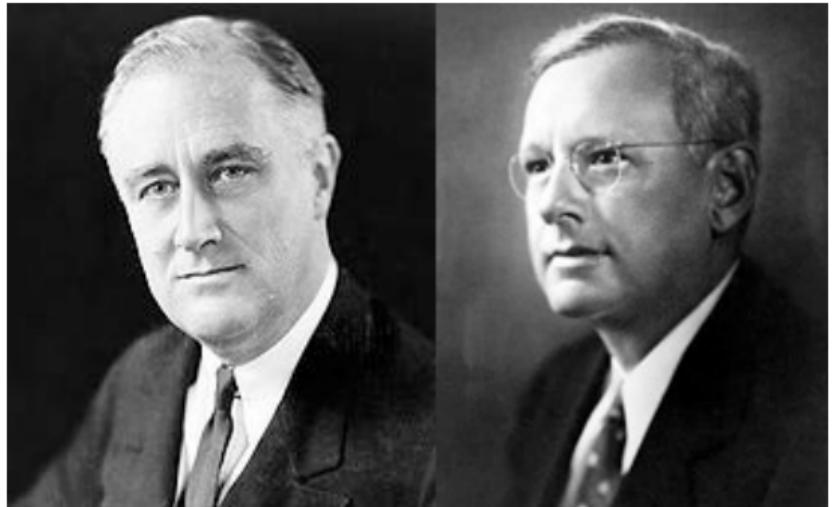
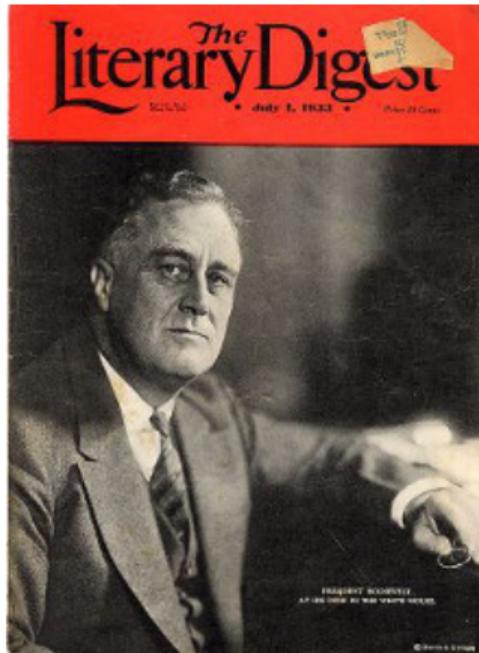
# Dados de treinamento não-representativos

# Dados de treinamento não-representativos



Dados de treinamento devem ser representativos para novos casos (ruído na amostragem)!

# Dados de treinamento não-representativos



Eleições dos EUA (1936): Roosevelt × Landon

- ▶ Viés na amostragem e viés de não-resposta.

# Qualidade baixa dos dados

# Qualidade baixa dos dados

Modelo	Quilometragem (km)	Ano	Marca	...	Valor (R\$)
COROLLA	1.000	2019	Toyota	...	60 000 000
Fusquinha	100 000	1984	Volkswagen	...	2 000
Onix 2017	15,000	2017	Chevrolet	...	30 000
:	:	:	:	:	:
Duster	10 000	2218	Renault	...	45.000

- ▶ Mineração, higienização e padronização dos dados.

# Características irrelevantes

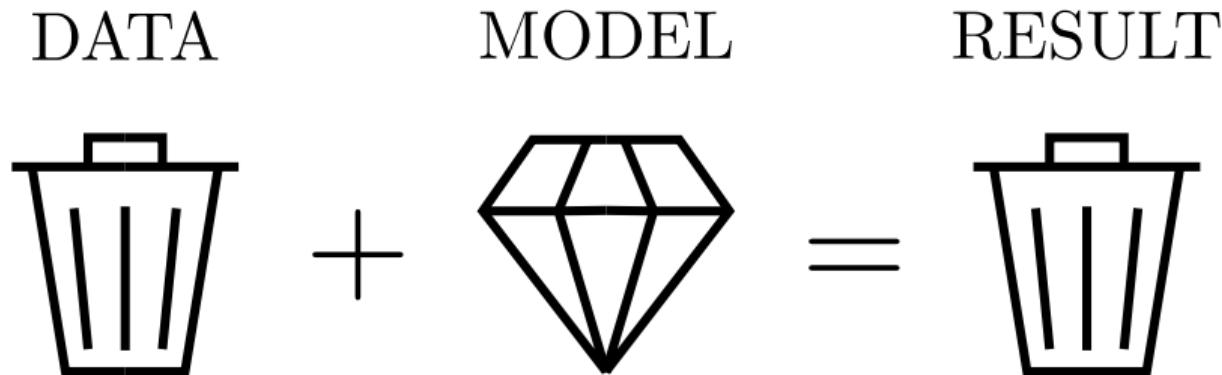
# Características irrelevantes

Modelo	Quilometragem (km)	Ano	Marca	...	Valor (R\$)
Corolla	1 000	2019	Toyota	...	60 000
Fusca	100 000	1984	Volkswagen	...	2 000
Onix	15 000	2017	Chevrolet	...	30 000
:	:	:	:	:	:
Duster	10 000	2018	Renault	...	45 000

- ▶ Engenharia de características: Seleção, extração, criação de características.

Em suma, para dados...

Em suma, para dados...

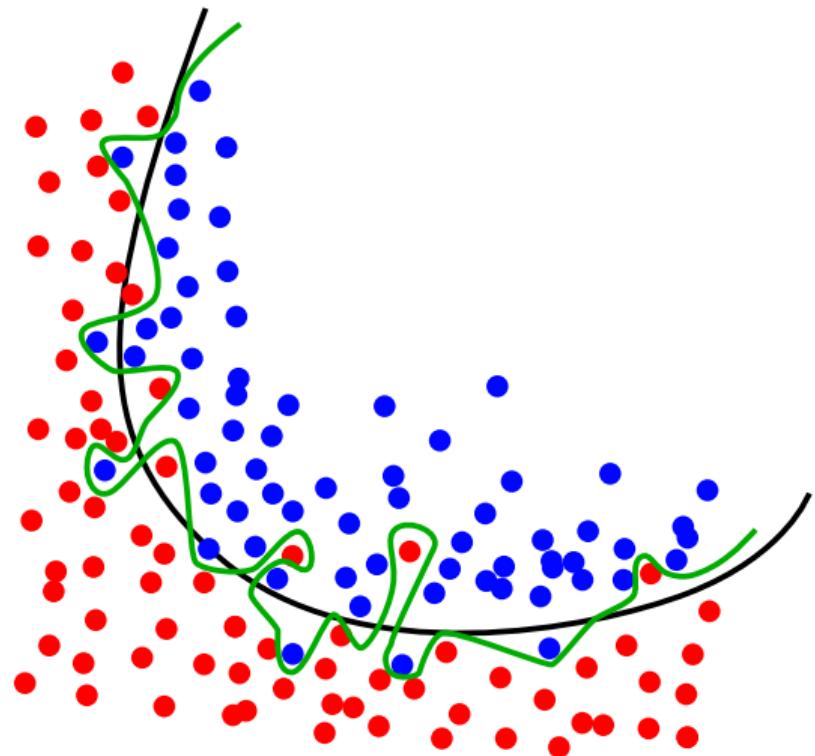


*Garbage in, garbage out!*

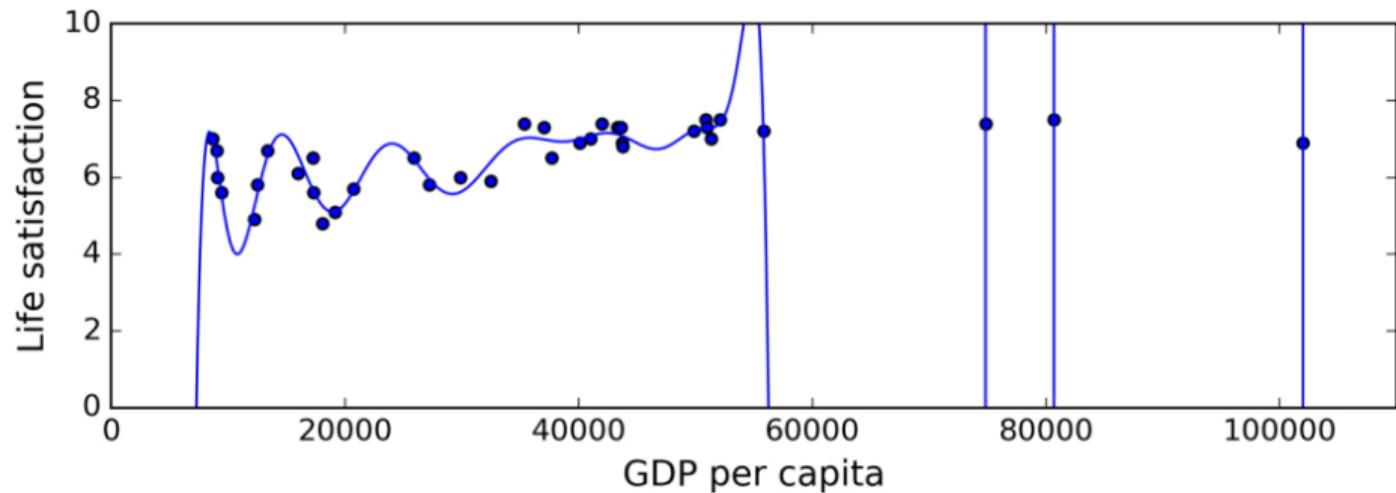
# Sobreajuste do conjunto de treinamento

# Sobreajuste do conjunto de treinamento

- Sobreajuste (ou *Overfitting*):  
Acontece quando o SAM tem bom desempenho apenas no conjunto de treinamento, i.e. quando o SAM é muito complexo em relação ao conjunto de treinamento.



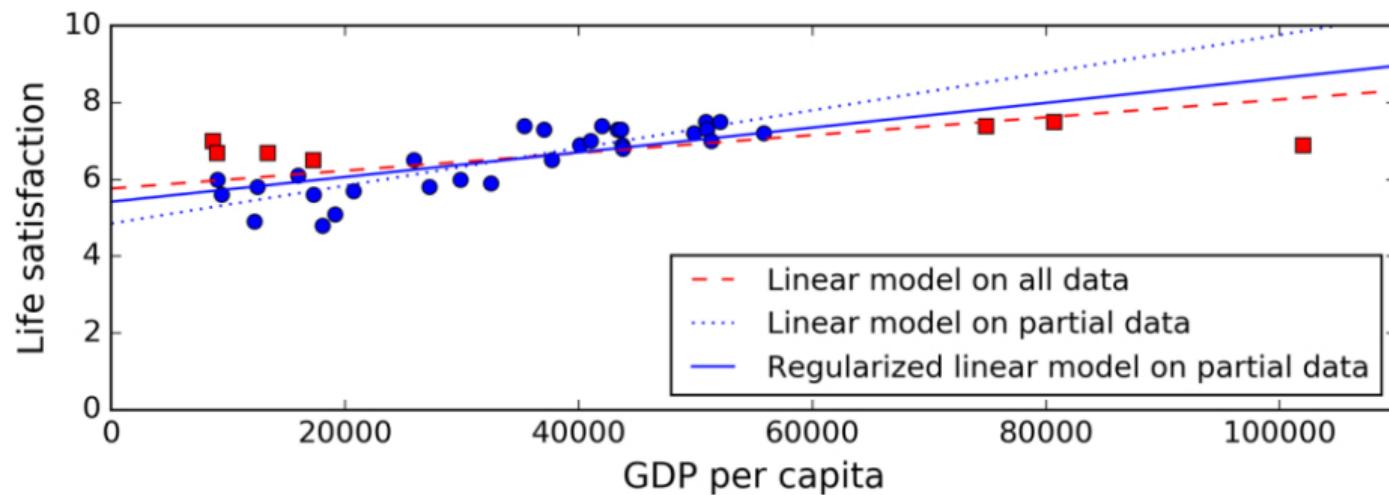
# Sobreajuste do conjunto de treinamento



Sobreajuste do conjunto de treinamento!

# Sobreajuste do conjunto de treinamento

- Regularização: Imposição de restrições, controladas pelos hiperparâmetros, ao AA para tornar o SAM mais simples.



Regularização reduz os riscos de sobreajustes!

# Sobreajuste do conjunto de treinamento

Como resolver o sobreajuste?

# Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;

# Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;

# Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;
3. Restringir o modelo;

# Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;
3. Restringir o modelo;
4. Adquirir mais dados;

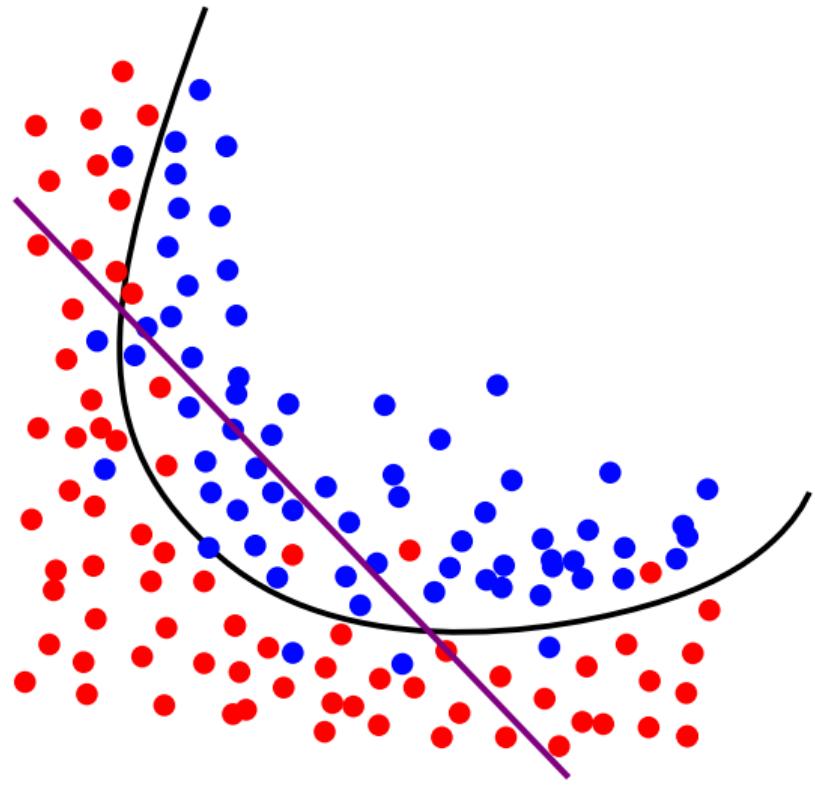
# Sobreajuste do conjunto de treinamento

1. Utilizar um modelo mais simples;
2. Reduzir o número de atributos;
3. Restringir o modelo;
4. Adquirir mais dados;
5. Reduzir o ruído do conjunto de treinamento.

# Subajuste do conjunto de treinamento

# Subajuste do conjunto de treinamento

- Subajuste (*Underfitting*): Acontece quando o SAM não tem bom desempenho nem mesmo no conjunto de treinamento, i.e. quando o SAM é muito simples em relação ao conjunto de treinamento.



# Subajuste do conjunto de treinamento

Como resolver o subajuste?

# Subajuste do conjunto de treinamento

1. Utilizar um modelo mais poderoso;

# Subajuste do conjunto de treinamento

1. Utilizar um modelo mais poderoso;
2. Escolher atributos melhores;

# Subajuste do conjunto de treinamento

1. Utilizar um modelo mais poderoso;
2. Escolher atributos melhores;
3. Reduzir o número de restrições.

# Conjunto de teste e conjunto de validação

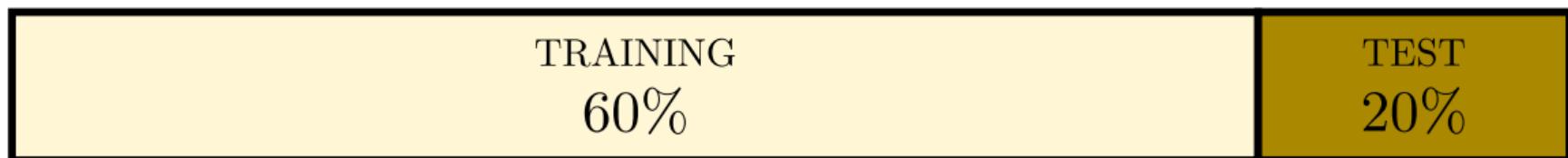
# Conjunto de teste e conjunto de validação

- ▶ Conjunto de treinamento: Dados usados para ajustar os parâmetros do SAM;
- ▶ Erro de treinamento.

TRAINING  
100%

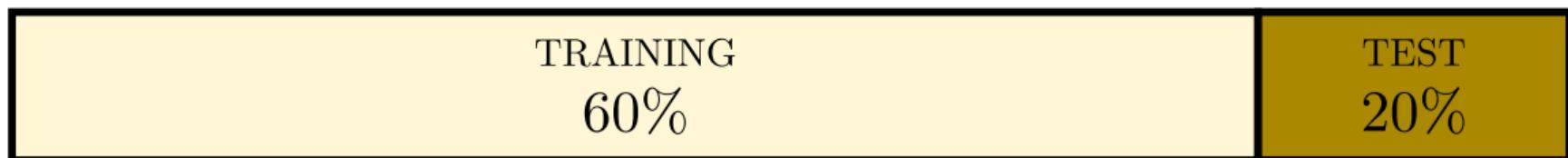
# Conjunto de teste e conjunto de validação

- ▶ Conjunto de teste: Dados usados para fazer a avaliação final do SAM;
- ▶ Erro de generalização.



# Conjunto de teste e conjunto de validação

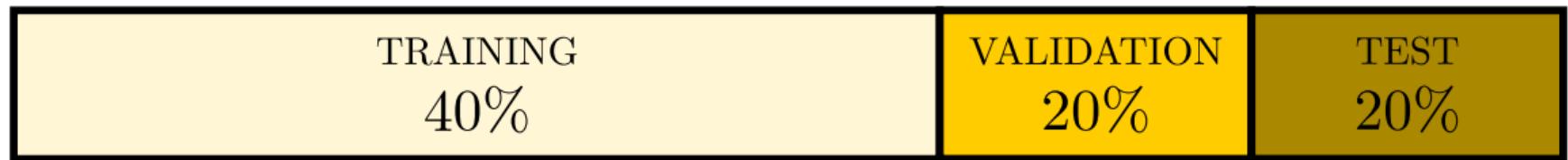
- ▶ Conjunto de teste: Dados usados para fazer a avaliação final do SAM;
- ▶ Erro de generalização.



- ▶ Se o erro de treinamento é baixo e erro de generalização é alto, então o SAM está sobreajustado.

# Conjunto de teste e conjunto de validação

- ▶ Conjunto de validação: Dados usados para ajustar os hiperparâmetros do AA e fazer avaliações intermediárias do SAM;
- ▶ Erro de validação.



# Conjunto de teste e conjunto de validação

- ▶ Conjunto de validação: Dados usados para ajustar os hiperparâmetros do AA e fazer avaliações intermediárias do SAM;
- ▶ Erro de validação.



- ▶ Se o erro de treinamento é baixo e erro de validação é alto, então o SAM está sobreajustado;

# Conjunto de teste e conjunto de validação

- ▶ Conjunto de validação: Dados usados para ajustar os hiperparâmetros do AA e fazer avaliações intermediárias do SAM;
- ▶ Erro de validação.

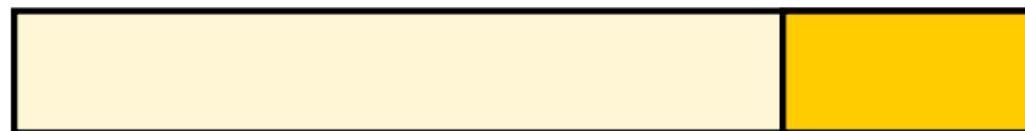


- ▶ Se o erro de treinamento é baixo e erro de validação é alto, então o SAM está sobreajustado;
- ▶ Se os erros de treinamento e validação são baixos e erro de generalização é alto, então o SAM está sobreajustado.

# Conjunto de teste e conjunto de validação

- ▶ Validação cruzada: Avalia a capacidade de generalização de uma análise estatística a partir de um conjunto de dados (e.g. *k-Fold*).

1<sup>st</sup> iteration



$e_1$

2<sup>nd</sup> iteration



$e_2$

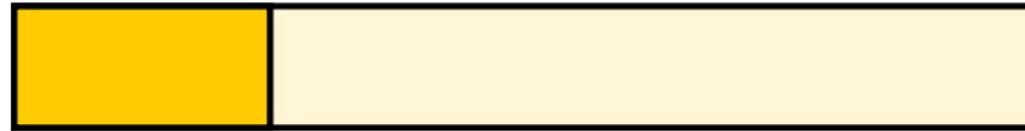
$\Rightarrow \langle e \rangle$

3<sup>rd</sup> iteration



$e_3$

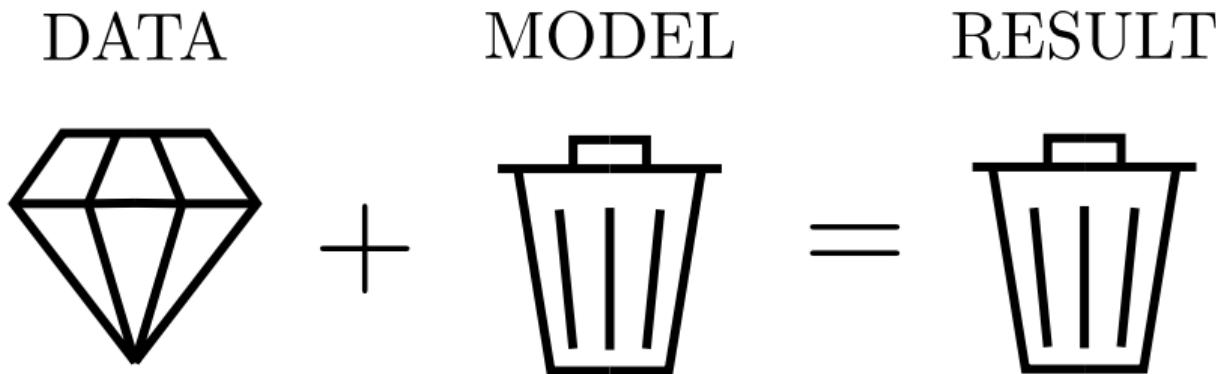
4<sup>th</sup> iteration



$e_4$

Em suma, para modelos...

Em suma, para modelos...



*Garbage in, garbage out!*

# Projeto de AM

The Kaggle logo is displayed in a large, bold, blue sans-serif font. The letters are slightly slanted to the right. The 'k' has a vertical stroke on its left side, and the 'g' has a small horizontal stroke at its bottom.

<http://www.kaggle.com>

# Projeto de AM



<https://www.kaggle.com/edgarhuichen/espn-nba-players-data>

## O salário do Stephen Curry é justo?



Hasta la vista, baby!