
Tera

Aula #06

Módulo 2

André Silveira, 06/mar/2018

Intro

Backgrounds diversos
Objetivo em comum

“This grand show is eternal. **It is always sunrise somewhere**; the dew is never all dried at once; a shower is forever falling; vapor ever rising. Eternal sunrise, eternal sunset, eternal dawn and gloaming, on seas and continents and islands, each in its turn, as the round earth rolls.” John Muir

Como será?

Parte 1: abordar e resolver um problema na prática
Parte 2: definir os principais traços de ML e sua participação no valor de um data-insight

Colaboração
compartilhe, pergunte, responda, estamos juntos!
Boas práticas
Zele pelo nosso estudo
Use bem o nosso tempo

Agenda

Warmup

Parte 1

1. Intro e expectativas
2. O problema
3. Solução: nossas ferramentas
4. Solução: implementação
5. Solução: compreensão
6. Solução: métrica de resultado

Parte 2

7. Problemas reais para ML
8. Pipeline de ML
9. Classificação e Regressão
10. Fluxo do data-insight
11. Review

Decompression e feedback

O problema

Como determinar a espécie de uma flor a partir de suas medidas?

Ronald Fisher [1]

Solução: nossas ferramentas

- Jupyter notebook
- Scikit-learn
- NumPy
- Pandas

[2]

[3]

[4]

[5]

Solução: implementação

Implementação

Hello world!

Solução: implementação

Hello world dataset

Maçãs e Laranjas

Bumpy = 0
Smooth = 1

Apple = 0
Orange = 1

[6]

Weight	Texture	Label
150g	Bumpy	Orange
170g	Bumpy	Orange
140g	Smooth	Apple
130g	Smooth	Apple
...	...	...

Solução: implementação

Implementação

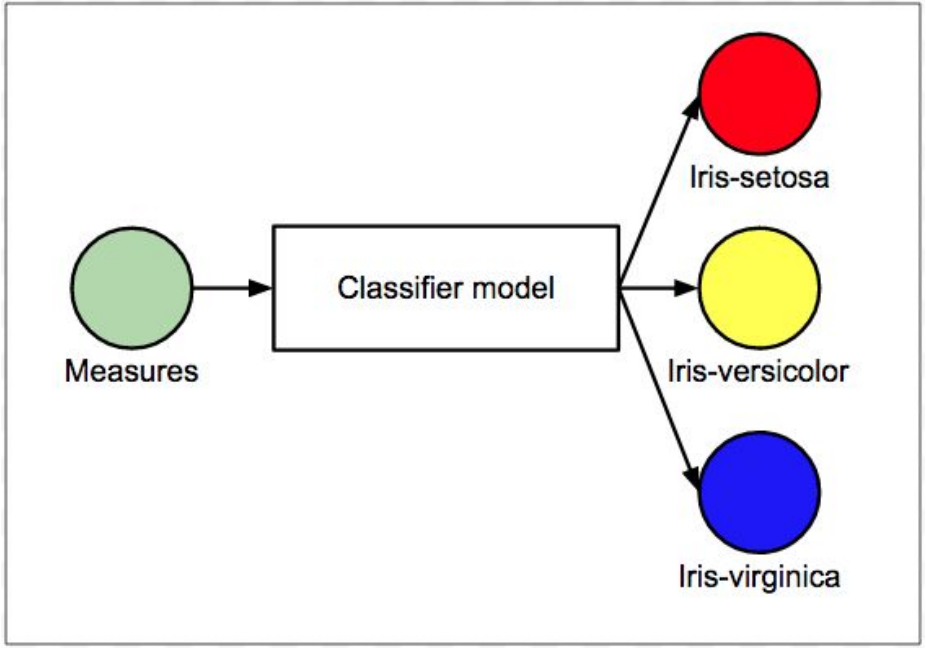
Iris case

Solução: implementação

Nome	Código
Iris-Setosa	0
Iris-Versicolour	1
Iris-Virginica	2

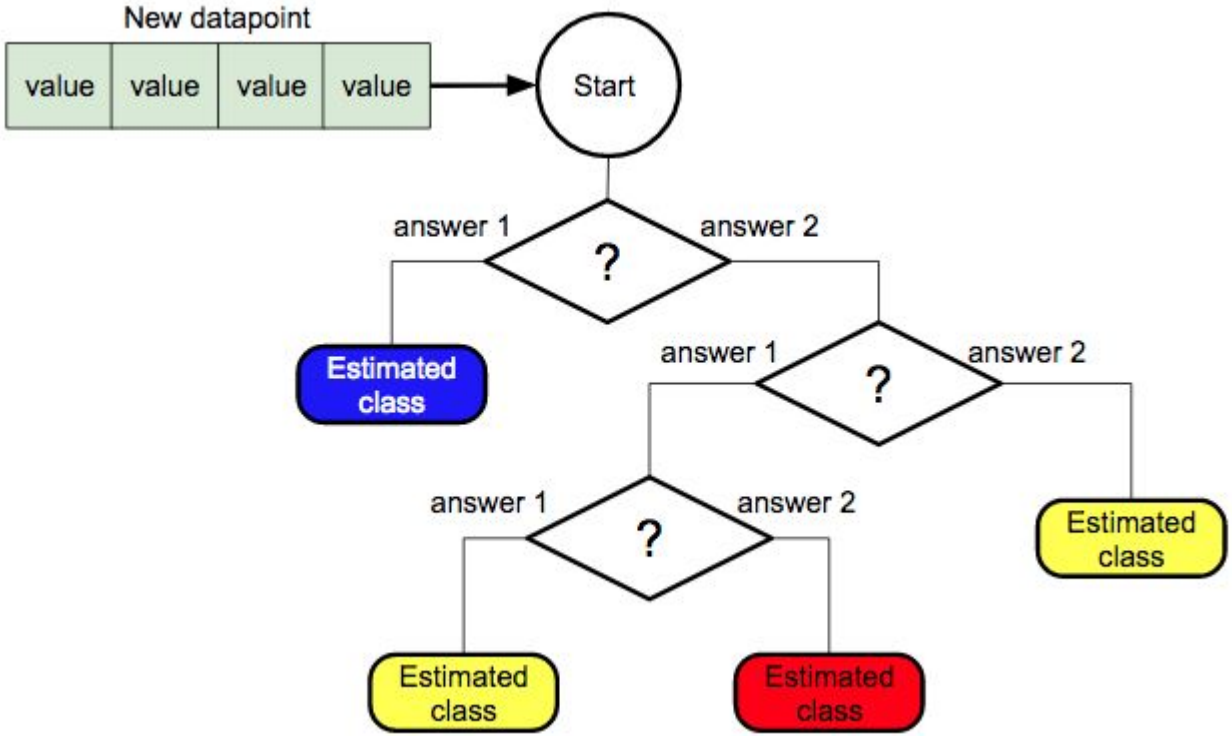
Solução: compreensão

Iris Case



Solução: compreensão

Decision Tree genérica



Solução: compreensão

Datapoints e splits

x1	x2	x3	x4	y
#	#	#	#	
#	#	#	#	
#	#	#	#	
#	#	#	#	
#	#	#	#	

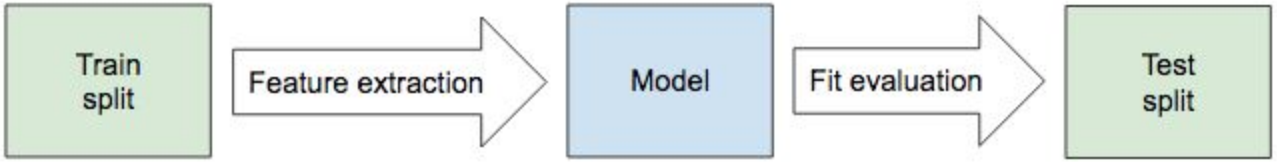
Train split

#	#	#	#	?
#	#	#	#	?

Test split

Solução: compreensão

Modelagem



Solução: compreensão

Algoritmo greedy

Passo 1: Comece com uma árvore vazia

Passo 2: Selecione uma feature para fazer a pergunta

Para cada subset de resposta:

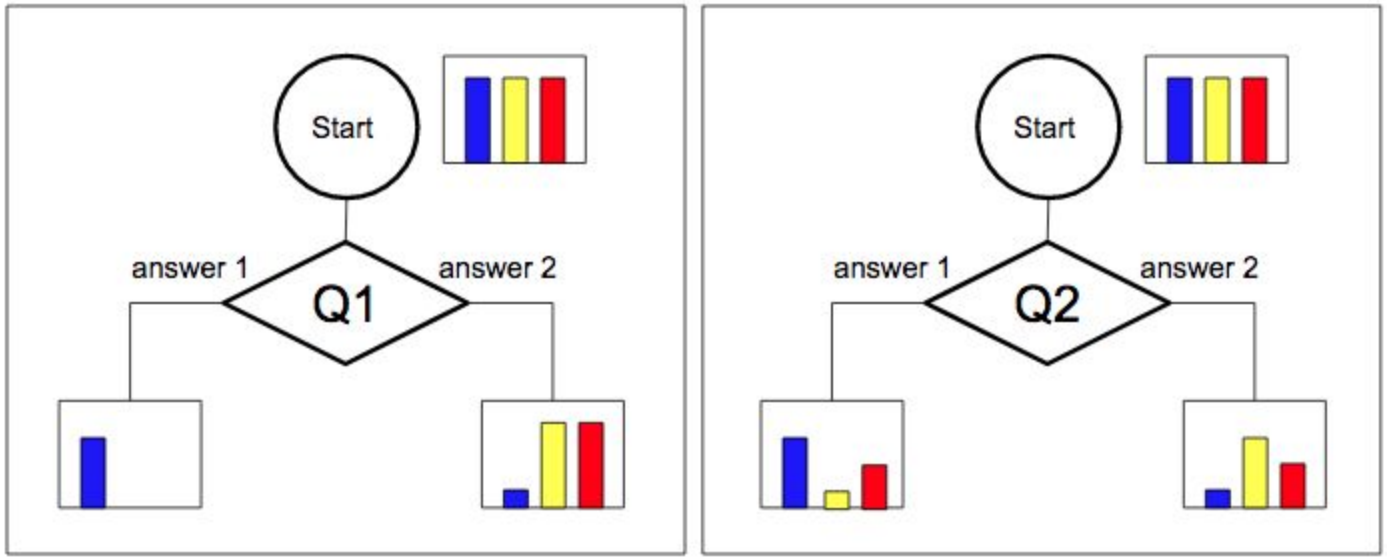
Passo 3: Se não houver mais dúvida entre targets ou não houver novas features para perguntar então faça a estimativa

Passo 4: Senão vá para o passo 2 e continue a partir deste subset

[7, 8]

Solução: compreensão

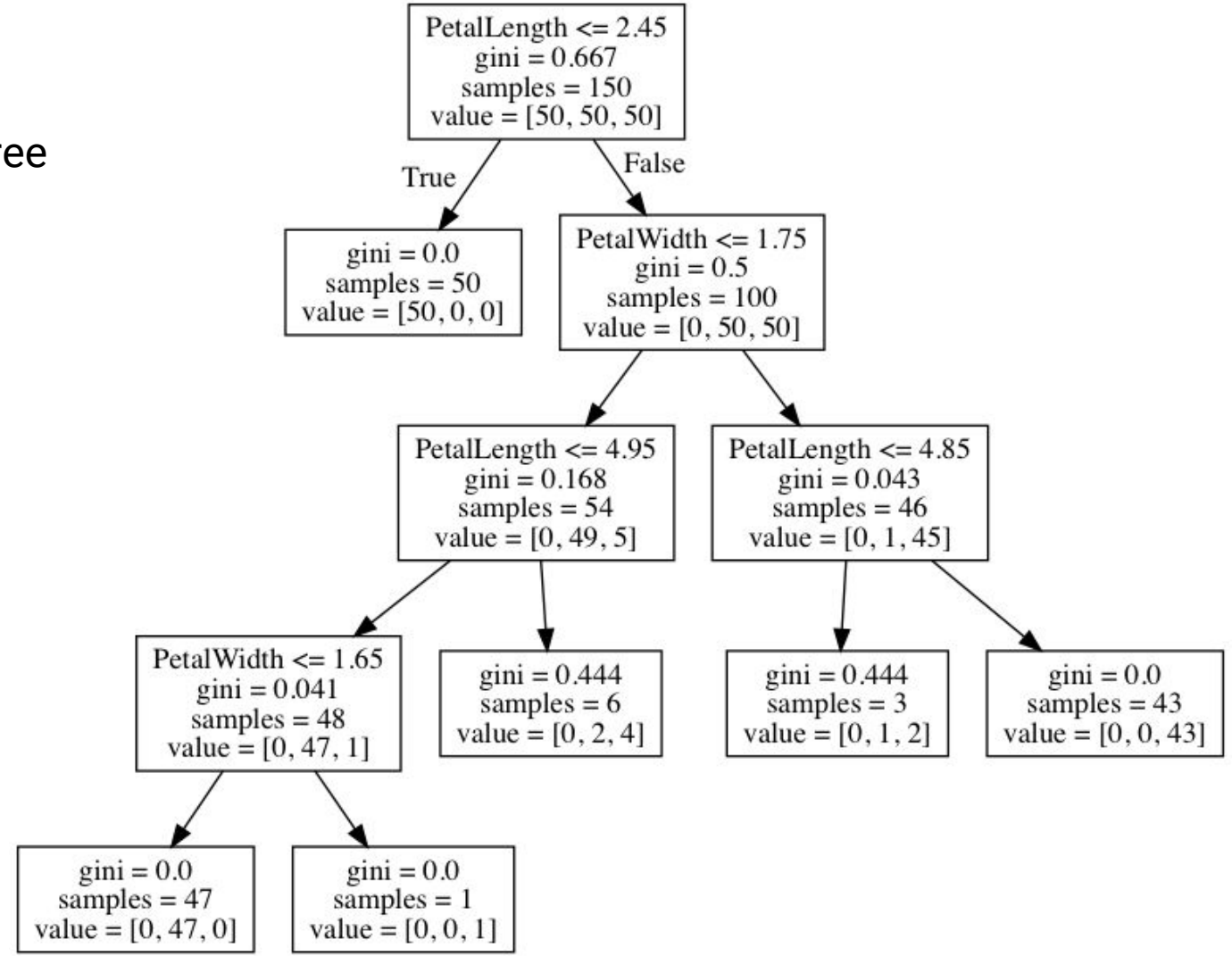
Como escolher a pergunta?



$$\text{Error} = \frac{\text{\# incorrect estimatives}}{\text{\# datapoints in subset}}$$

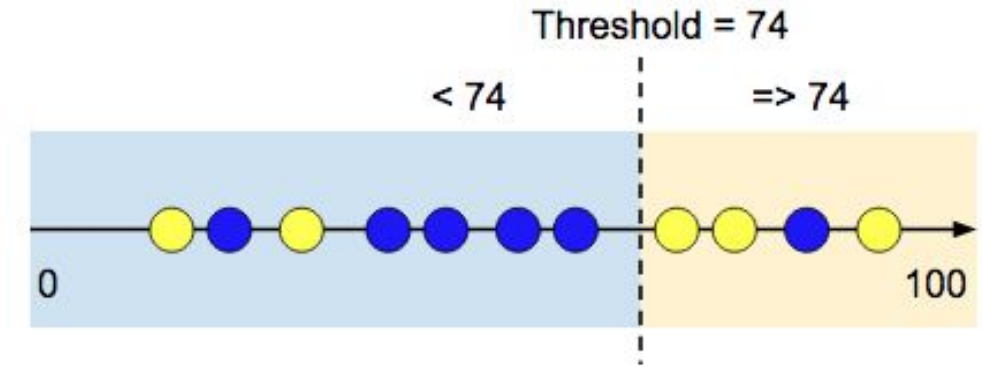
Solução

Iris decision tree



Solução: compreensão

Features com valores numéricos



[Adaptado de 7]

Solução: métrica de resultado

Quão acurado é a estimativa da árvore que construímos?

O que significa acurácia?

Hello world:
maçãs e laranjas

?

Iris case

?

...vamos ver na doc do scikit-learn.

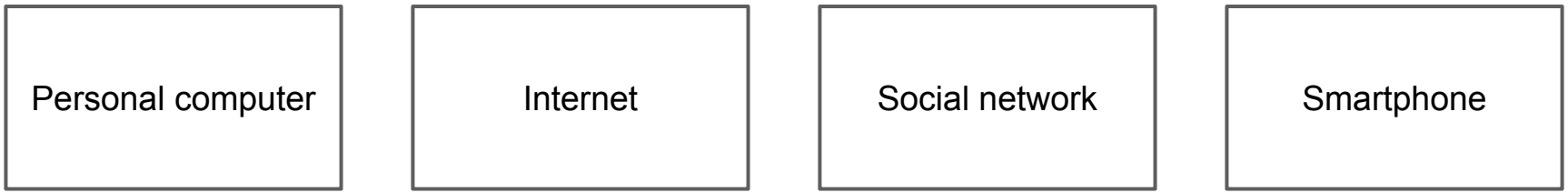
Intervalo

Problemas reais para ML

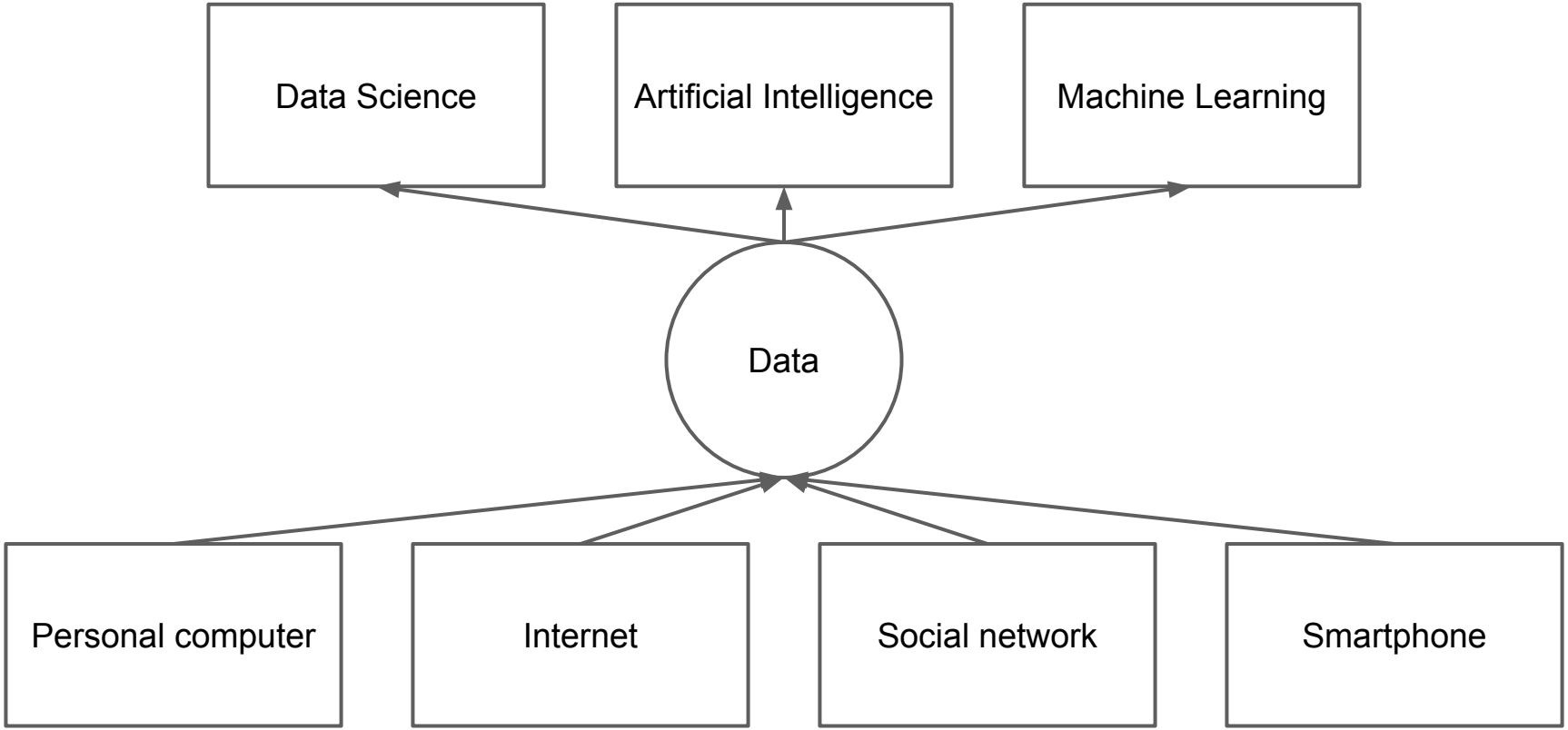
Exemplos?

Problemas reais para ML

Problemas reais para ML



Problemas reais para ML



Problemas reais para ML

Machine Learning

O que faz:

- Learn from data <> explicitly programmed
- Pattern recognition
- Inference, modelling

Recursos:

- “enough” data
- methods: get, prepare, explore, model, infer

Limites:

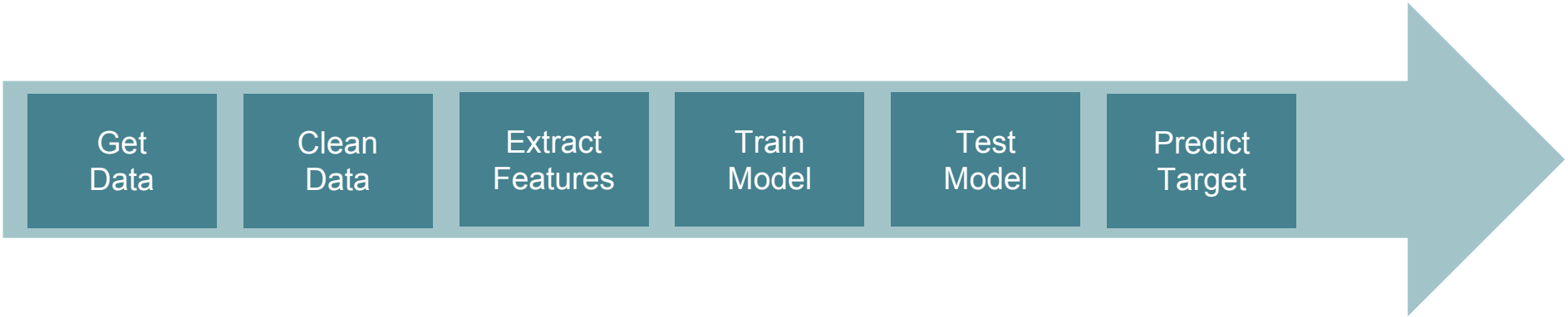
- garbage in, garbage out
- overfitting

Problemas reais para ML

Email filtering, Optical Character Recognition, Learn to Rank, Identify image quality, Identify elements in images, credit card fraud, translation, medical diagnosis, insurance, credit approval, marketing, recommender systems, sentiment analysis, financial market, time series forecasting, ...

Pipeline de ML

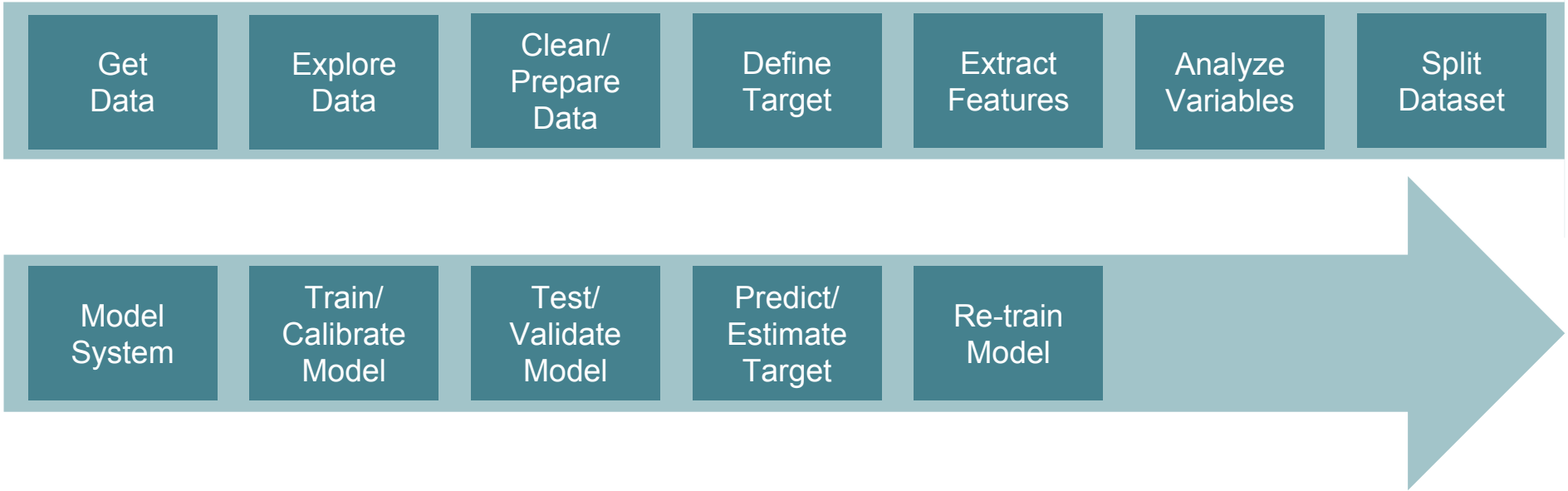
Simples



Pipeline de ML

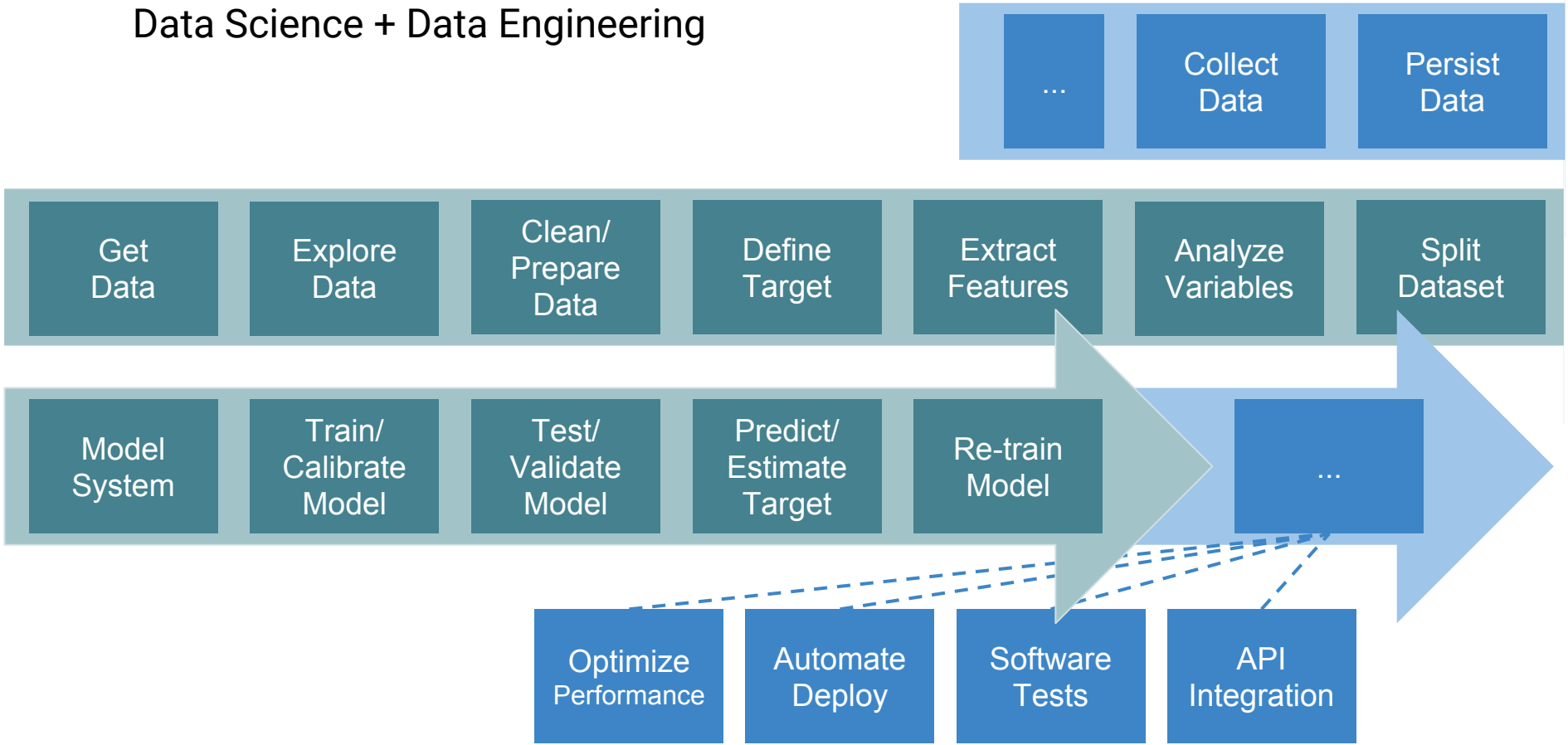
Detalhado*

* Não exaustivamente, em cada contexto novas etapas podem ser destacadas.



Pipeline de ML

Data Science + Data Engineering



Classificação e Regressão

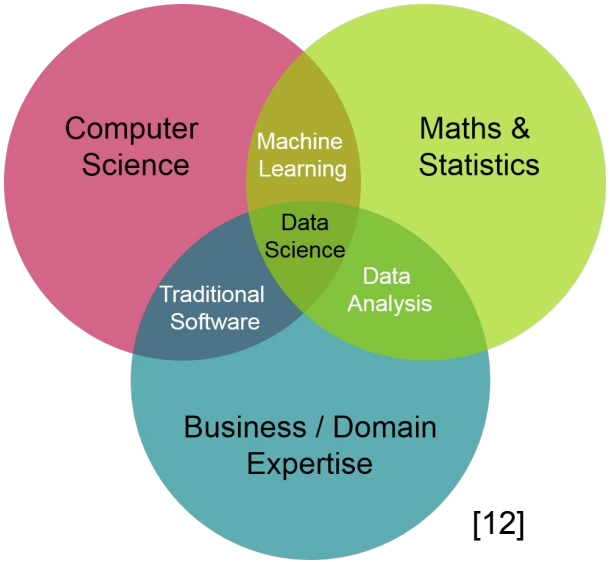
[9]

[10]

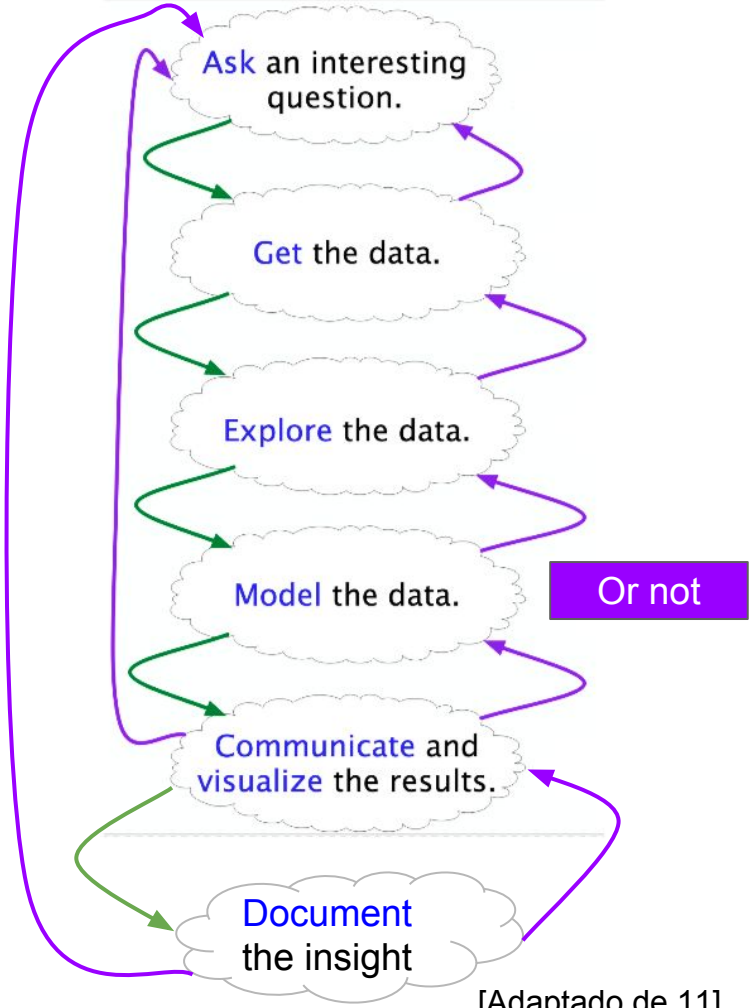
...vamos ver na [doc do scikit-learn](#).

Fluxo do data-insight

"Thus, data science is an act of interpretation—we translate the customer's 'voice' into a language more suitable for **decision-making**." [13]



[12]



[Adaptado de 11]

Fluxo do data-insight

> At Airbnb, Data Science Belongs Everywhere [13]

> How Airbnb Democratizes Data Science With Data University [15]

> Superset: Scaling Data Access and Visual Insights at Airbnb [14]

> Democratizing Data at Airbnb [16] Dataportal

Review

O que aprendemos hoje?

Decompression
e
Feedback

Referências

1. https://en.wikipedia.org/wiki/Iris_flower_data_set
2. http://jupyter.org/
3. http://scikit-learn.org/stable/index.html
4. http://www.numpy.org/#
5. https://pandas.pydata.org/
6. https://www.youtube.com/watch?v=cKxRvEZd3Mw&list=PLOU2XLYxmsIIuiBfYad6rFYQU_jL2ryal
7. https://www.coursera.org/learn/ml-classification/lecture/F8kuT/intuition-behind-decision-trees
8. https://en.wikipedia.org/wiki/Greedy_algorithm
9. http://www.landinfo.com/classification_object-based-image-analysis.htm
10. http://blog.csdn.net/iracer/article/details/50658506
11. http://cs109.github.io/2015/
12. https://ion.icaew.com/itcounts/b/weblog/posts/theaccountinganddatascienceworldsmeet
13. https://medium.com/airbnb-engineering/at-airbnb-data-science-belongs-everywhere-917250c6beba
14. https://medium.com/airbnb-engineering/superset-scaling-data-access-and-visual-insights-at-airbnb-3ce3e9b88a7f
15. https://medium.com/airbnb-engineering/how-airbnb-democratizes-data-science-with-data-university-3eccc71e073a
16. https://medium.com/airbnb-engineering/democratizing-data-at-airbnb-852d76c51770

Obrigado!