Pythonzão: uma saga do Python no mundo de Big Data

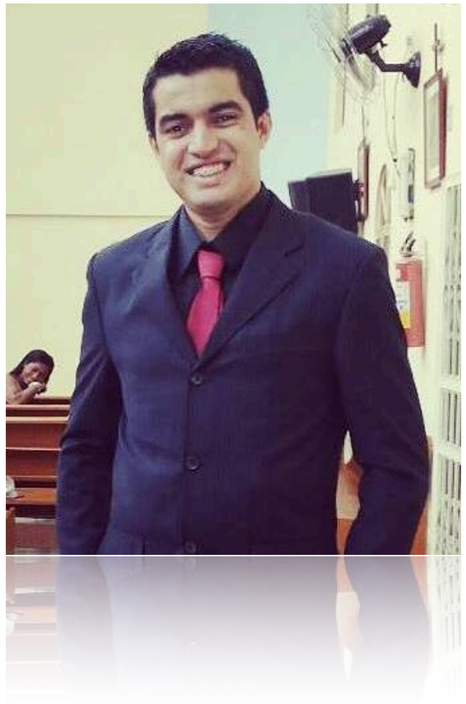
Bruno Ábia

bruno.abia@icomp.ufam.edu.br

12/08/2017

Pycon Amazônia 2017

Quem sou eu?



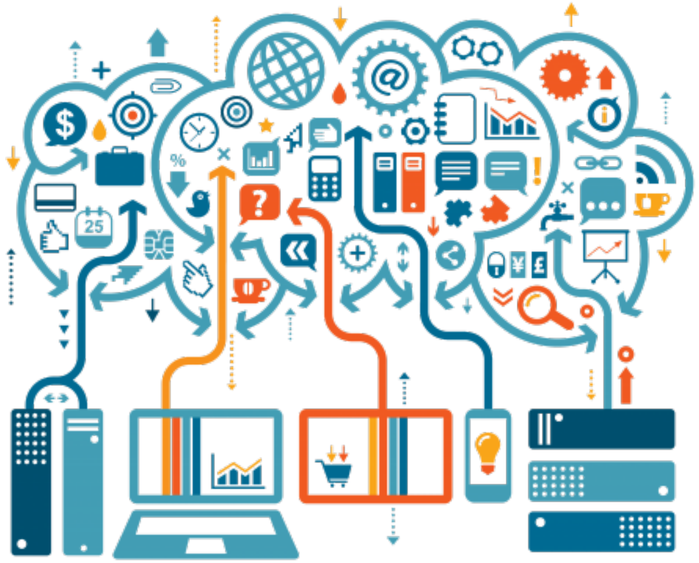
- Engenheiro de Computação (Faculdade Fucapi);
- Mestre em Informática (UFAM);
- Doutorando em Informática (UFAM);

- Líder de Desenvolvimento (UNASUS/AM) (UEA);
  - Algoritmos de Recomendação;

- Pesquisador do Grupo DNS LAB(UFAM);
  - Ciência de dados;
  - Machine Learning;
  - Redes Sociais;
  - Análise de Sentimentos

DNS lab
Data and Network Science Lab

Introdução



Acessórios

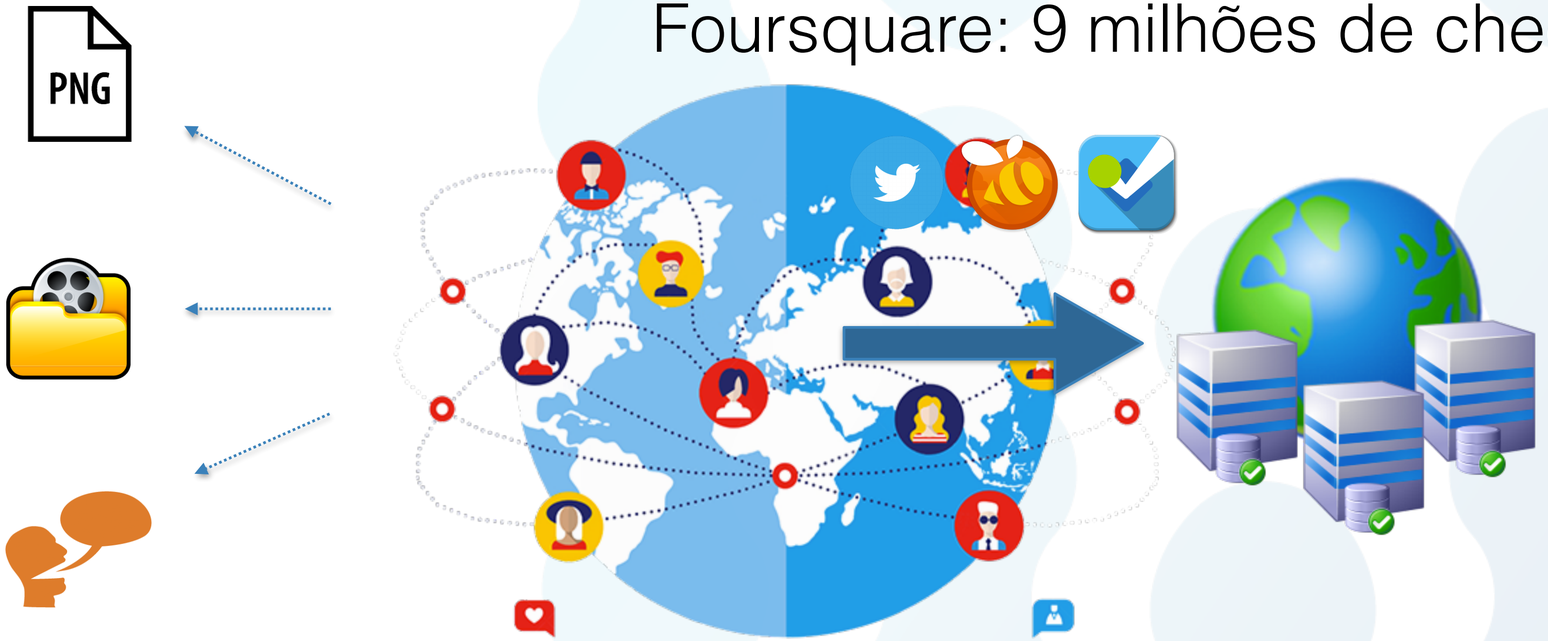
Redes Sociais

Smart Cities

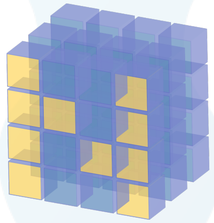
DNS lab
Data and Network Science Lab

Introdução

Twitter: 1 Bilhão de tweets mensais
Foursquare: 9 milhões de checkins por mês

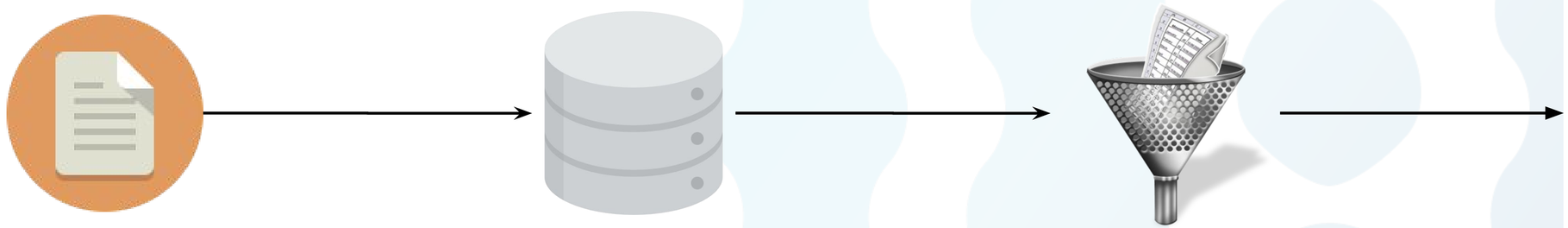


Introdução



NumPy

Introdução



Documentos da WEB

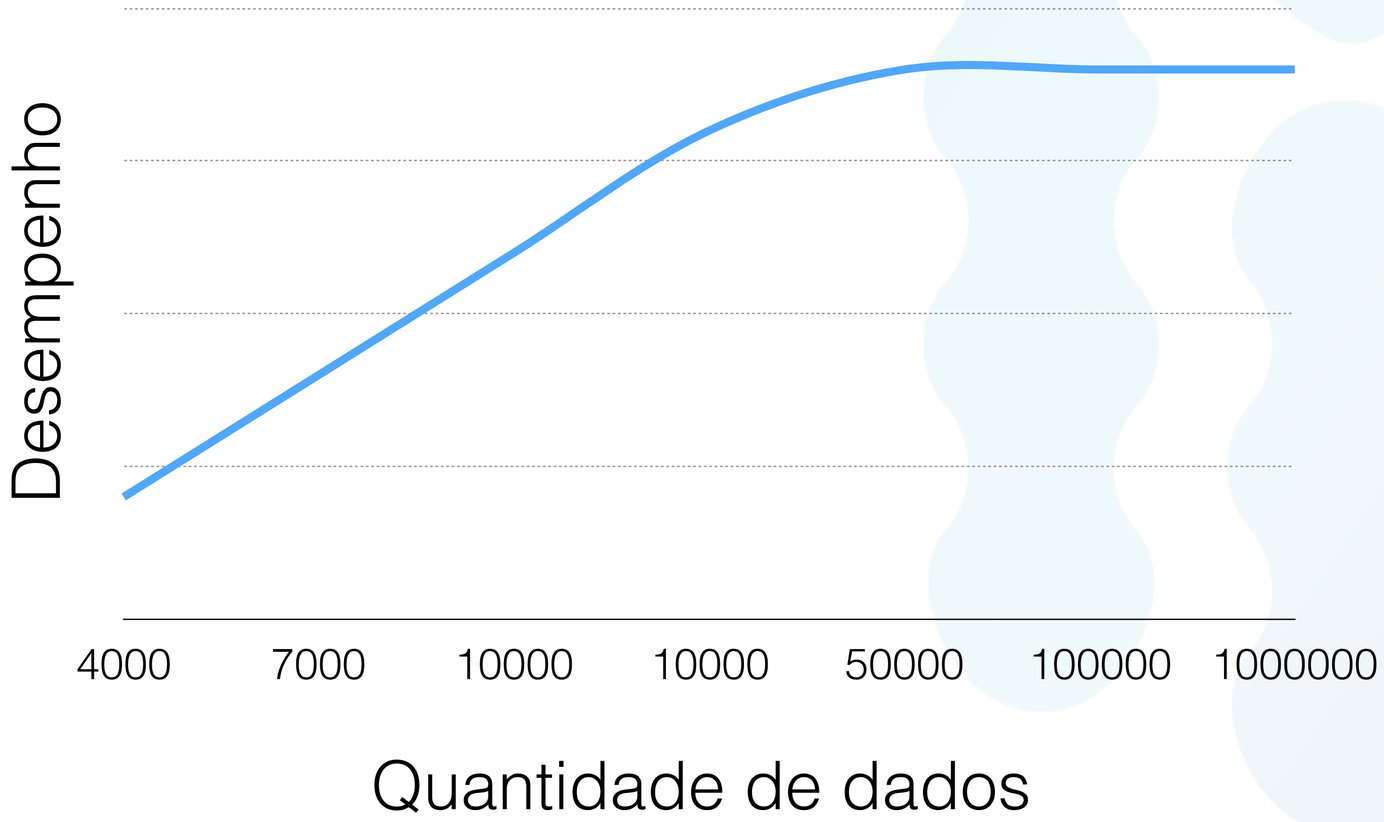
Coleta de dados

Pré-processamento

Classificação

DNS lab
Data and Network Science Lab

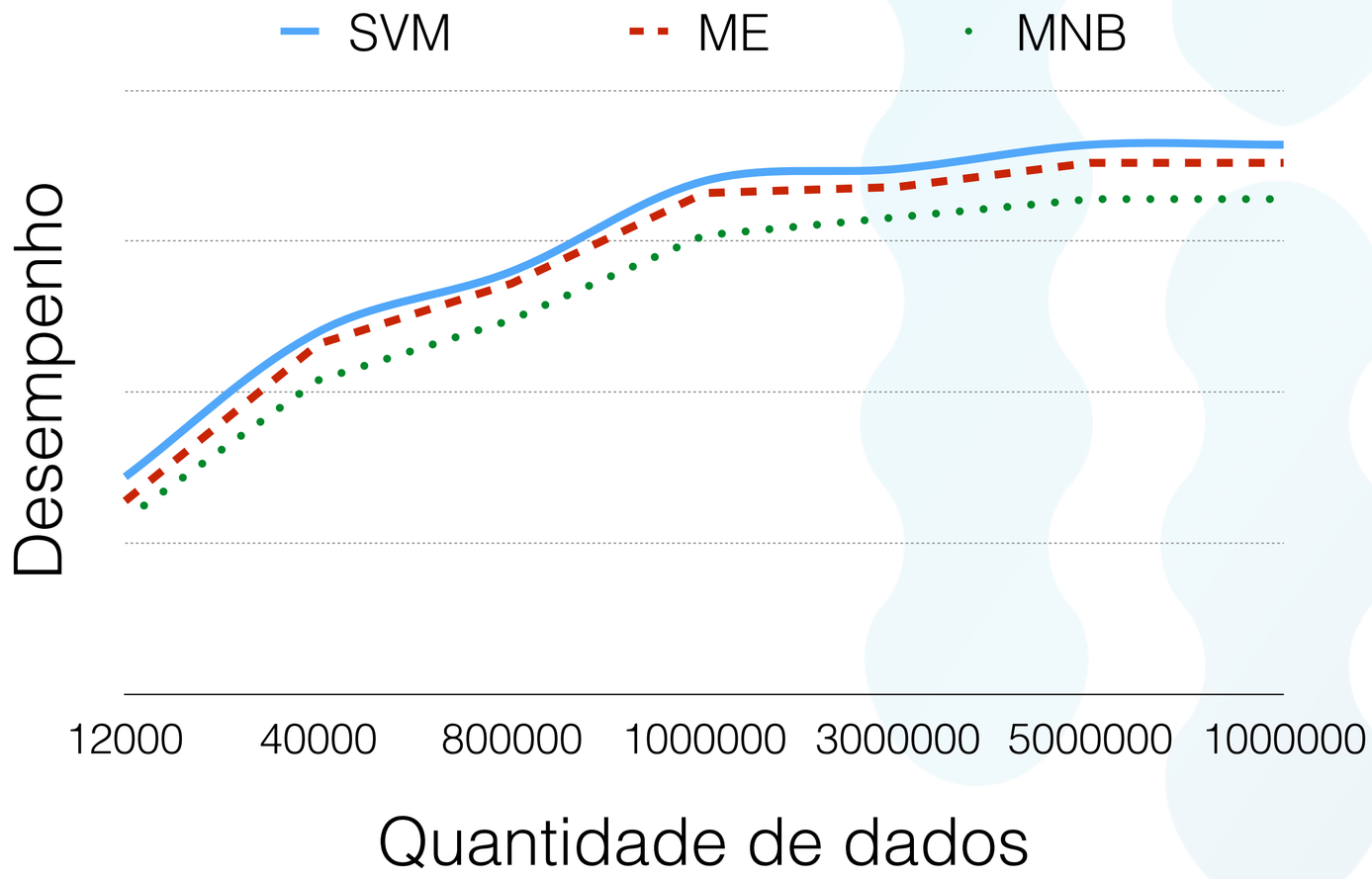
Problema



O desempenho dos métodos de aprendizagem diminui.

DNS lab
Data and Network Science Lab

Problema



Reduzindo o problema de variância e aumentando generalização

DNS lab
Data and Network Science Lab

Soluções

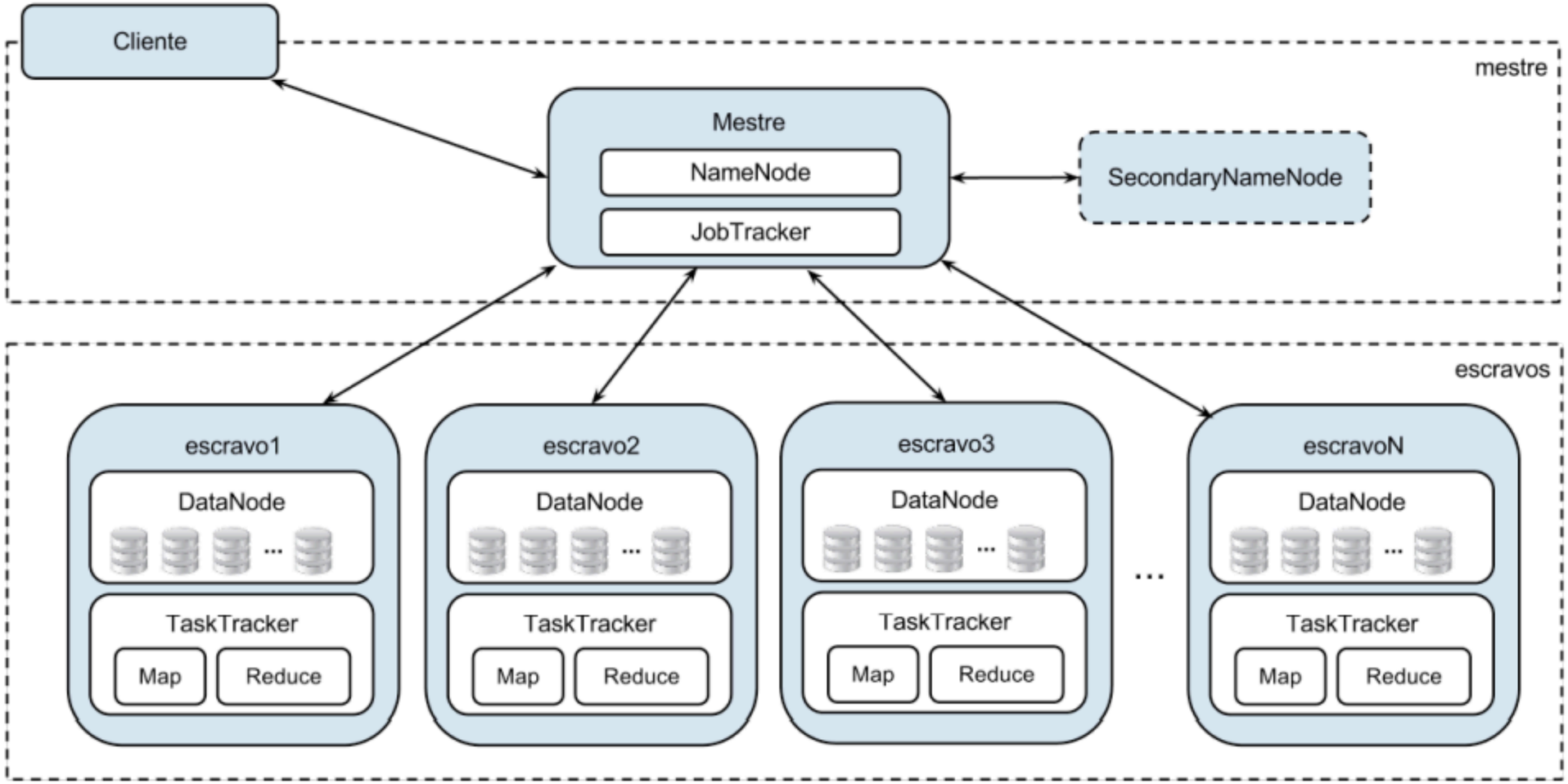
DNS Lab
Data and Network Science Lab

Soluções

- Plataforma de computação distribuída;
- Custo Baixo e em grande escala;
- Pode analisar dados complexos;

DNS lab
Data and Network Science Lab

Soluções


Cliente
mestre
Mestre
NameNode
JobTracker
SecondaryNameNode
escravos
escravo1
DataNode
TaskTracker
Map
Reduce
escravo2
DataNode
TaskTracker
Map
Reduce
escravo3
DataNode
TaskTracker
Map
Reduce
...
escravoN
DataNode
TaskTracker
Map
Reduce

APACHE HBASE

Soluções

- São variações do Haddop;
- Se baseiam nos dados salvos no HDFS;
- Tem como objetivo facilitar o acesso aos dados;

Desvantagem:

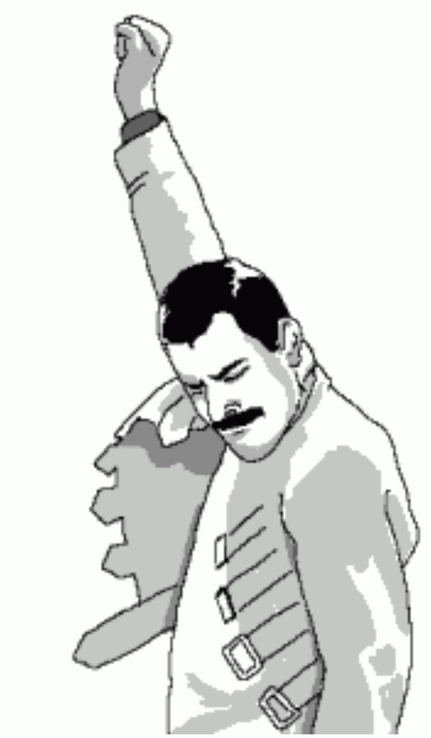
- Só funciona para dados estáticos;
- Necessário conhecer Java;
- Podemos escrever códigos em Python, porém precisamos da Interface **Jython;**

DNS Lab
Data and Network Science Lab

Soluções

- Framework que utliza Resilient Distributed Datasets (RDDs);
- Opera o conjunto de dados de uma só vez;
- 100 vezes mais rápido que o Hadoop para análises in-memory;

DNS lab
Data and Network Science Lab



Soluções

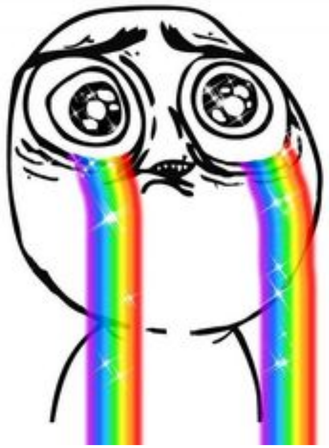
E sim!!!! Temos um versão pra Python!!.

PySpark

DNS Lab
Data and Network Science Lab

Recursos Disponíveis

PySpark



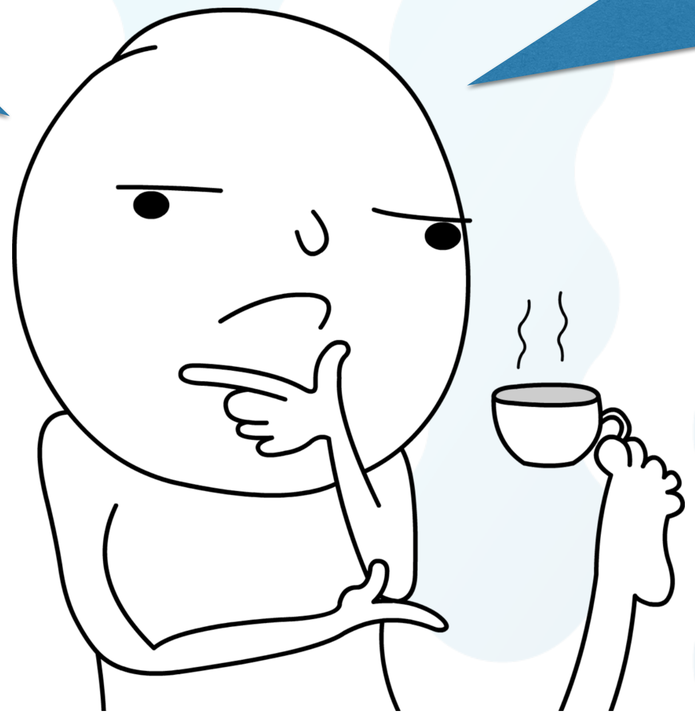
- Machine Learning
- Deep Learning
- Complex Networks
- Natural Process Language
- Data Mining

DNS Lab
Data and Network Science Lab

Dúvidas Comuns

Onde posso aplicar?

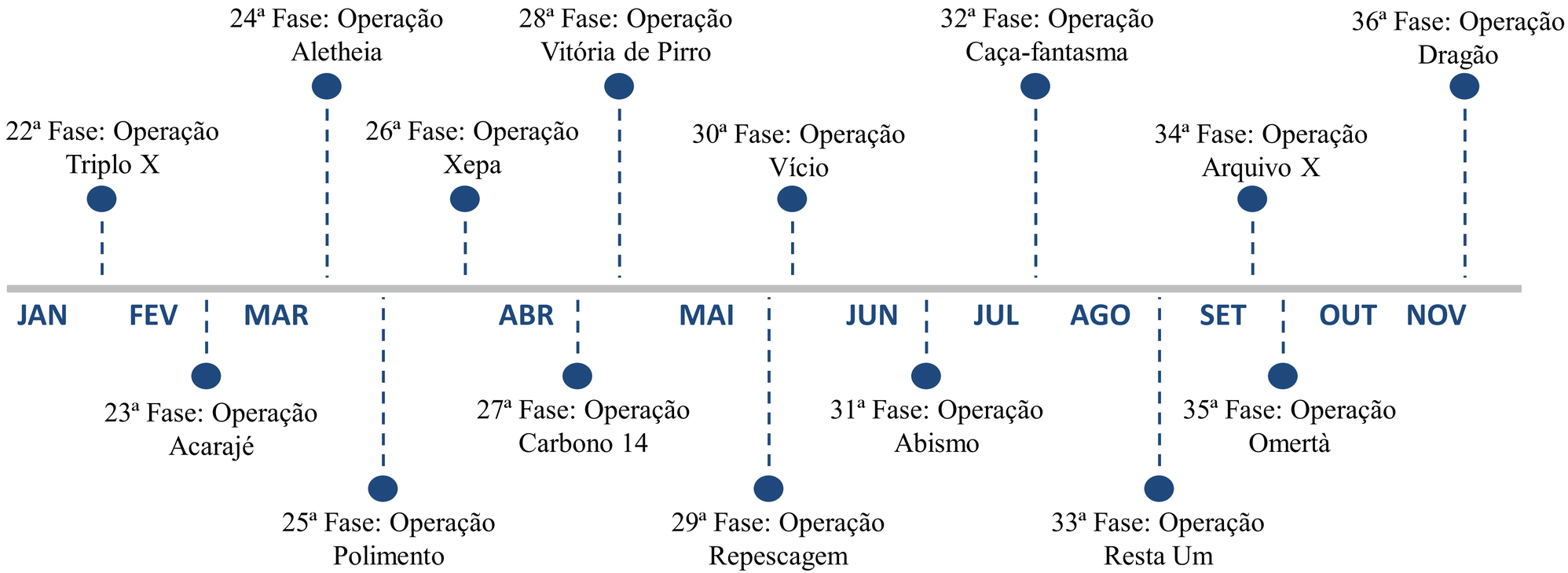
É complexo usar o Pyspark?



DNS lab
Data and Network Science Lab

Exemplo de Aplicação

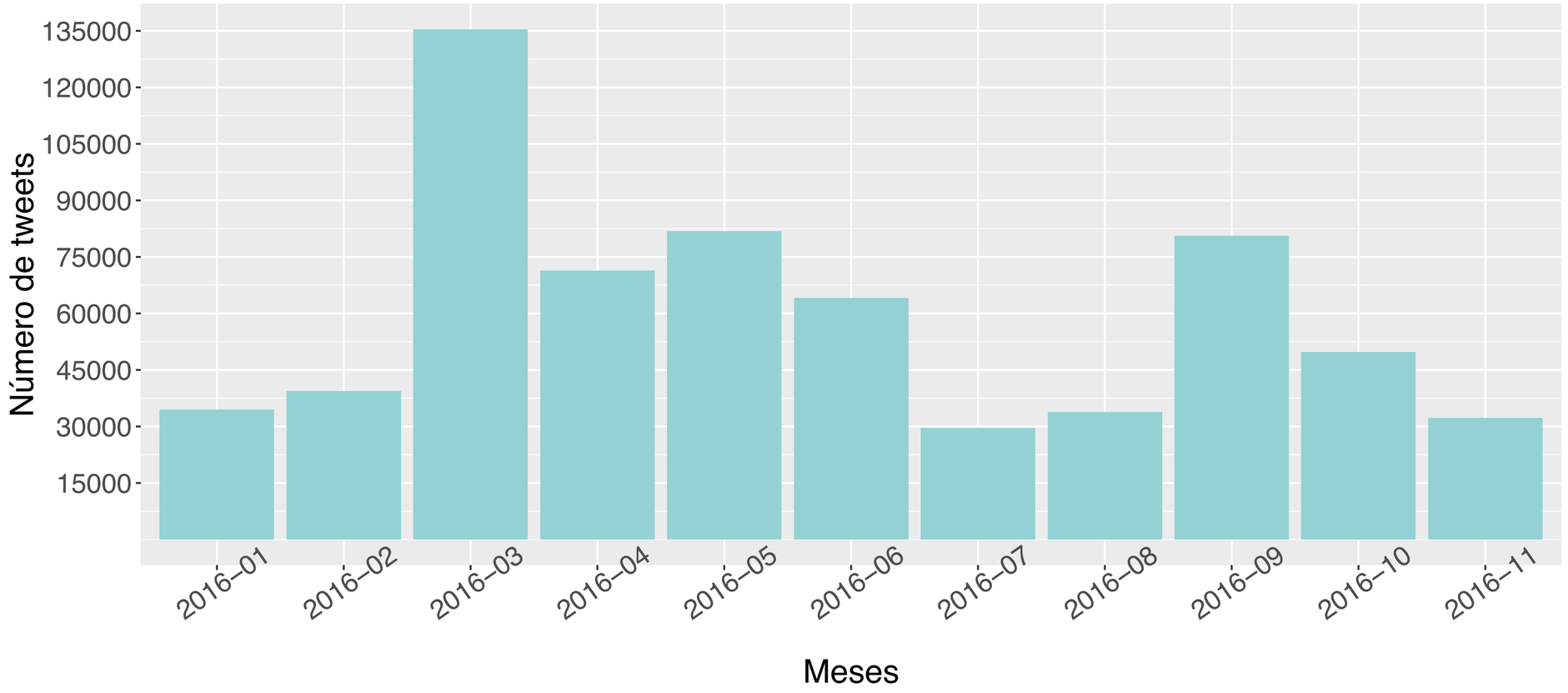
Operação Lava Jato da Polícia Federal



Exemplo de Aplicação

Base de Dados

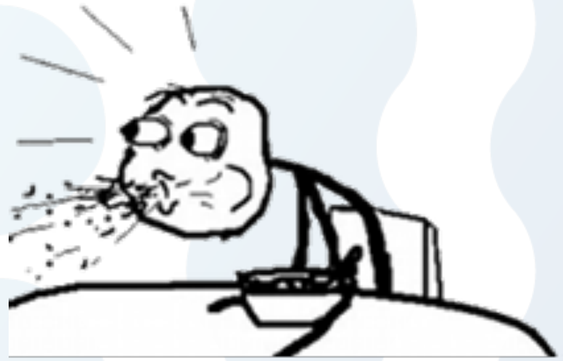
- Período Coletado 01/01/2016 a 20/11/2016;
- Total de 652.210 tweets;



DNS | Data and Network Science Lab

Como iniciar?

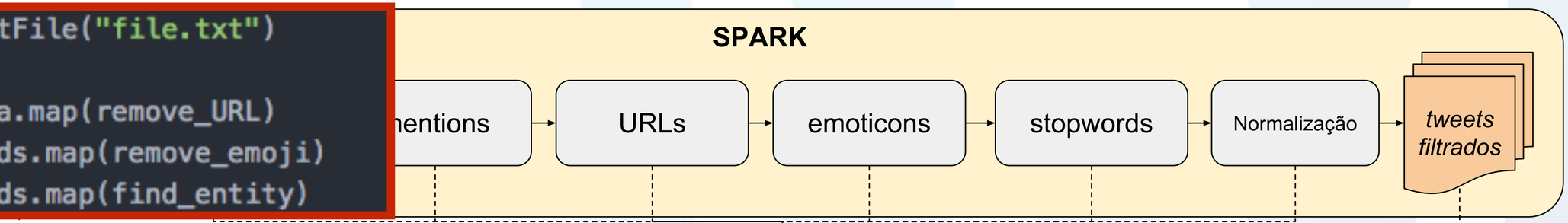
```
os.environ['SPARK_HOME'] = "/home/aluno/spark-1.6.1-bin-hadoop1/"

# # Append pyspark  to Python Path
sys.path.append("/home/aluno/spark-1.6.1-bin-hadoop1/python/")

#Create Pyspark import
try:

    from pyspark import SparkContext
    from pyspark import SparkConf

    print("Successfully imported Spark Modules")
    sc = SparkContext("local", "Simple Language Model Computing")

except ImportError as e:
    print("Can not import Spark Modules", e)

    sys.exit(1)
```



DNS lab
Data and Network Science Lab

Pré-Processamento

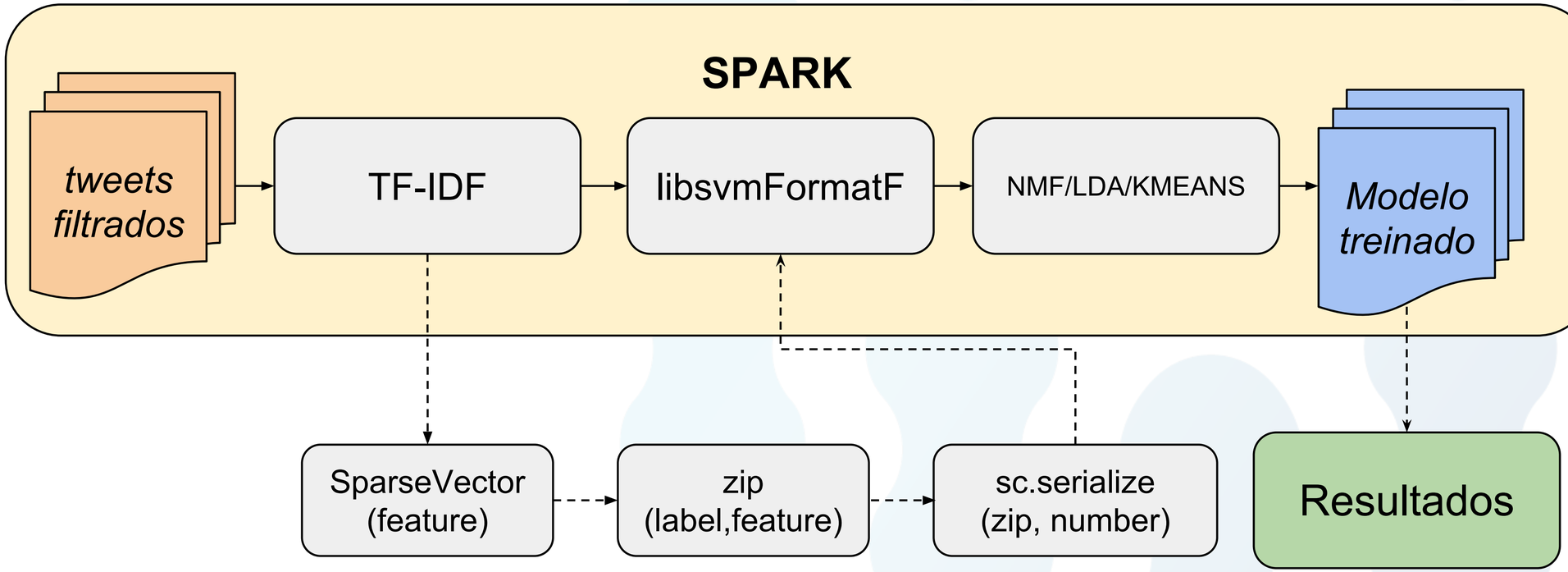
```
data=sc.textFile("file.txt")

words = data.map(remove_URL)
words = words.map(remove_emoji)
words = words.map(find_entity)
```




```
def remove_emoji(text):
    d = []

    d = re.compile("["u"\U0001F600-\U0001F64F"  # emoticons
                   u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                   u"\U0001F680-\U0001F6FF"  # transport & map symbols
                   u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                   "]+", flags=re.UNICODE)
    d = d.sub(r'', text)

    return d
```

DNS lab
Data and Network Science Lab

Treinamento dos Algoritmos de M.L.



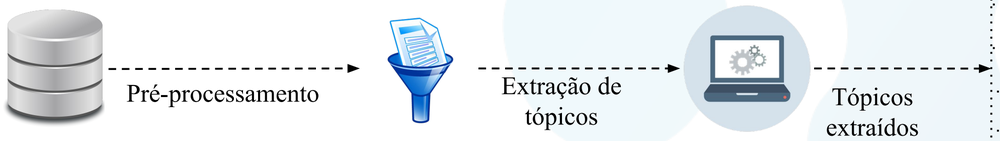
DNS lab
Data and Network Science Lab

Resultados

Notícia compartilhada (SET/2016)

Exemplo de *tweet*:
O **Paulo** Bernardo e **Gleisi** Hoffman são reús na Lava Jato.. Justiça sendo feita!!! Fora **PT**!!

$$\text{Percentual} = \frac{\text{Tópicos correspondentes}}{\text{Total de Notícias}}$$



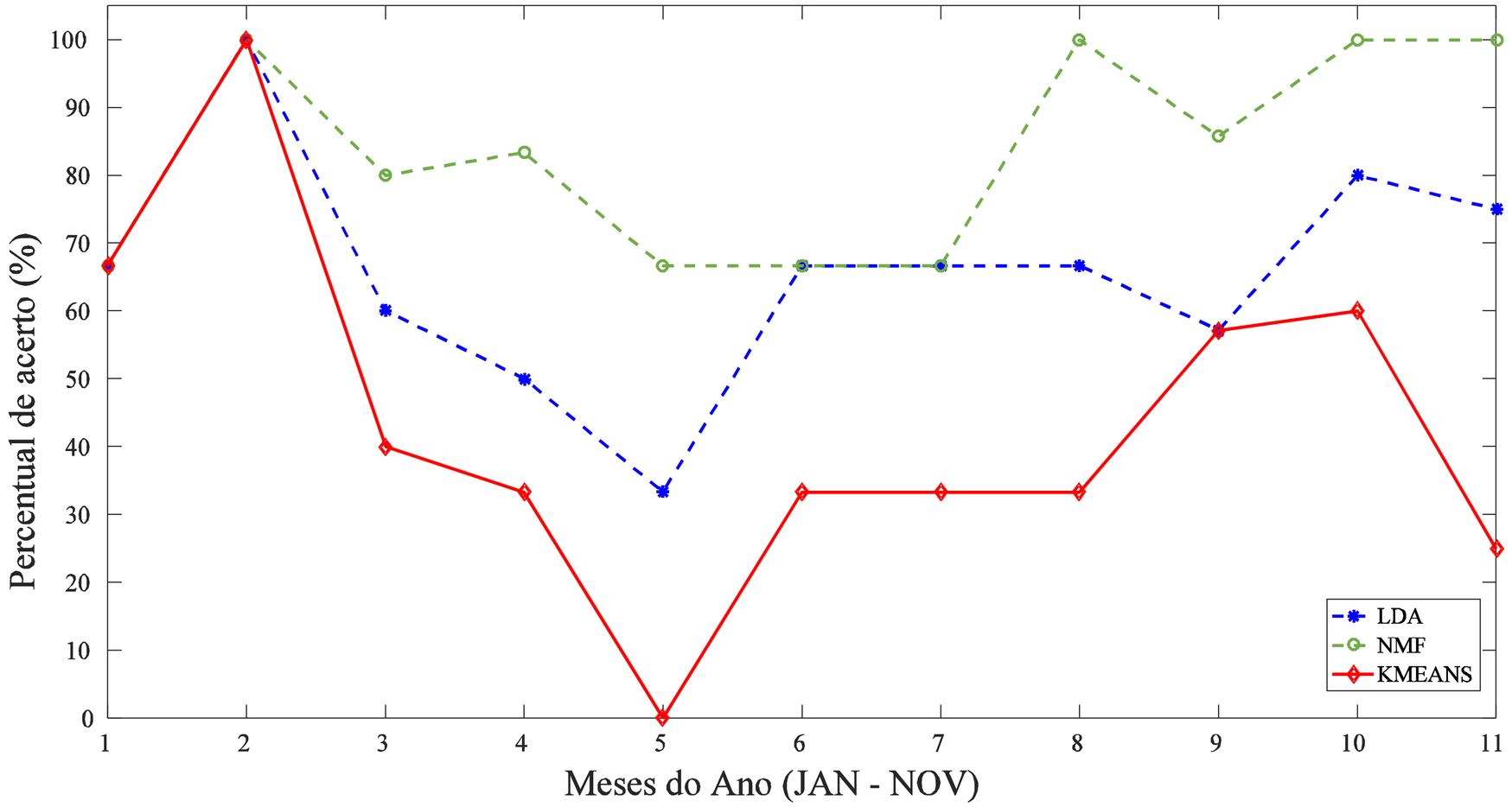
Petrobras Mantega STF
Lula Marisa MPF
Sergio Moro Bumlai
Palocci Antonio Dilma
PT Paulo Gleisi

(Zhao et al., 2011)

DNS lab
Data and Network Science Lab

Resultados

Percentual de Acerto por mês



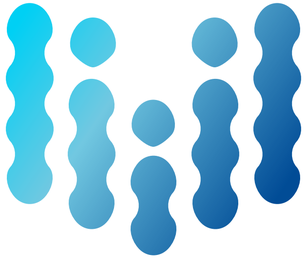
DNS Lab
Data and Network Science Lab

Conclusão

- Big Data é uma realidade em nosso dia a dia;
- Precisamos pensar como cientistas de dados;
- Python é realmente a linguagem para resolver esses problemas;
- O céu é o limite;

DNS lab
Data and Network Science Lab

Obrigado!



DNS lab
Data and Network Science Lab

Bruno Ábia Souza
bruno.abia@icomp.ufam.edu.br

DNS lab
Data and Network Science Lab

DNS Lab
Data and Network Science Lab