FUNDAMENTOS E TÉCNICAS EM CIÊNCIAS DE DADOS

PROF. JOSENALDE OLIVEIRA

josenalde@eaj.ufrn.br
https://github.com/josenalde/datascience

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

OBJETIVOS

O(a) discente selecionará, com base na origem dos dados, as ferramentas e técnicas para pré-processamento (limpeza), tratamento e processamento de dados, com a respectiva saída gráfica ou textual para auxílio à tomada de decisão. Adquirirá uma visão ampla dos softwares, bibliotecas, métodos estatísticos e linguagens utilizadas na área de ciência de dados, de modo a incorporar tais soluções no desenvolvimento de software.

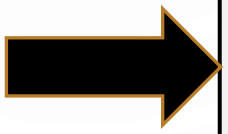
CONTEÚDO

Unidade I: introdução à ciência de dados, big data e suas etapas (produção; armazenamento (nosql); transformação e armazenamento analítico (olap, data warehouse, BI, data discovery), dashboards; **Unidade II**: processamento paralelo, distribuído e tempo real (hadoop, spark, spark streaming); **Unidade III**: análise de dados, métricas estatísticas e visualização, noções de mineração de dados e técnicas de predição (regressão linear simples, logística, séries temporais)

PLANO DE CURSO

DESCRIÇÃO, ARMAZENAMENTO

Introdução a ciência de dados e Big Data; Banco de dados não relacionais (desnormalização); Conceitos e definições sobre Ciência dos Dados; produção de dados; armazenamento; análise de dados; visualização, agrupamento e análise em rede. Caracterização, importância das etapas de modelagem e análise científica de dados. Ferramentas e linguagens para análise científica de dados.



PREDIÇÃO

Aquisição e formatação de dados. Análise estatística de dados. Agrupamento e classificação de dados. Visualização científica de dados. Principais conceitos para gerenciamento de Big Data; Tecnologias para Big Data; Técnicas estatísticas de predição; Data Discovery, OLAP e Visualização de Dados; Processamento paralelo e distribuído de dados; Visualização de dados: estática e interativa. Desenvolvimento de painéis de visualização (dashboards).

VISUALIZAÇÃO, ANÁLISE

Introdução a mineração de dados; Análise de dados de redes sociais. Modelagem de dados em grafos. Introdução a raspagem de dados (scraping). Coleta de dados estruturados e não estruturados. Desenvolvimento de raspadores de dados (scrapers). Ética e legislação sobre a raspagem de dados.

O que define um projeto de ciência de dados? A melhoria de algum aspecto no/do cliente

PLANO DE CURSO

- Referências

Notas de aula .pdf disponibilizadas de autoria do docente (slides)

IGUAL, Laura; SEGUÍ, Santi. Introduction to Data Science: a Python approach to concepts, techniques and applications. Springer, 2017.
https://github.com/DataScienceUB/introduction-datascience-python-book

E-book disponível em: https://www.springer.com/gp/book/9783319500164

AMARAL, Fernando. Introdução à Ciência de Dados: mineração de dados e big data. Rio de Janeiro: Alta Books, 2016.

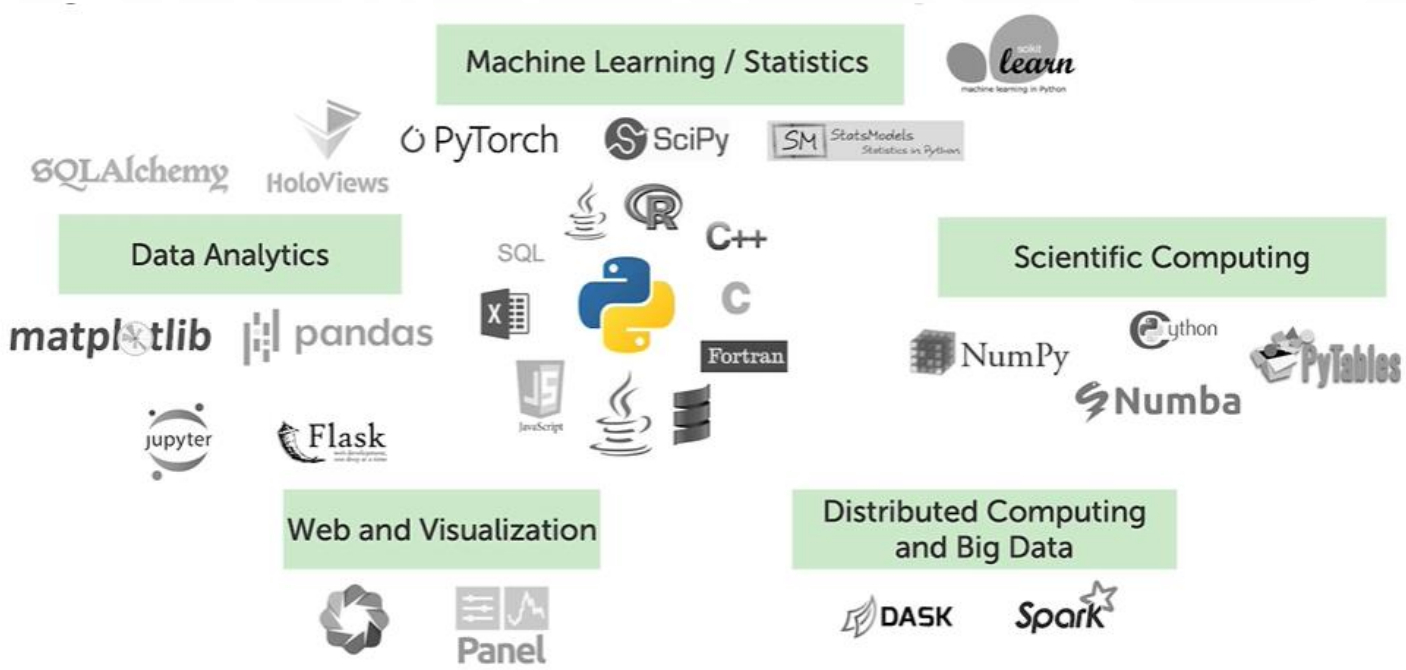
* McKINSEY, Wes. Python para análise de dados. São Paulo: Novatec, 2018.
https://github.com/wesm/pydata-book

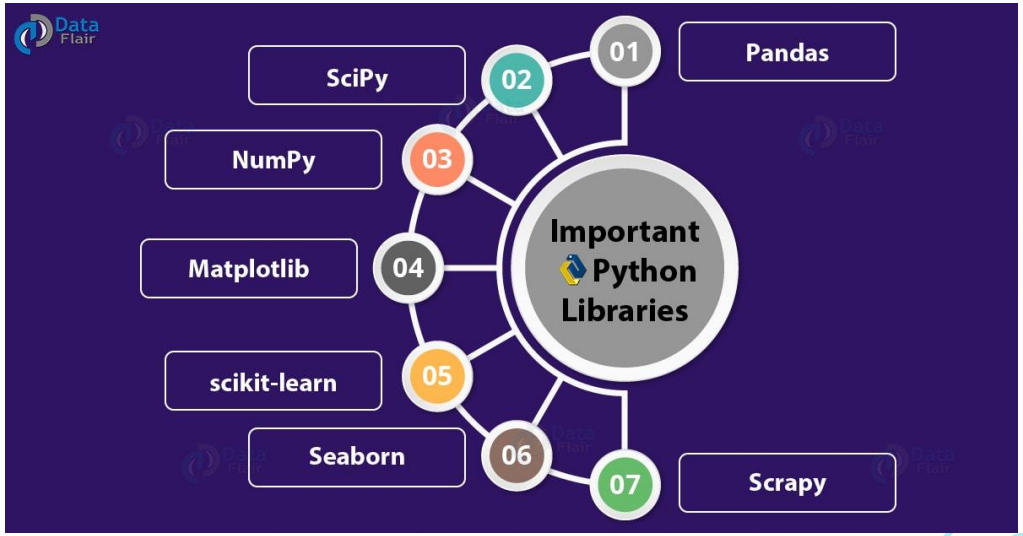
https://github.com/josenalde/datascience
www.kaggle.com

Complementar:
KELLEHER, John D.; TIERNEY, Brendan. Data Science. MIT Press, 2018.

FERRAMENTAS (SUGERIDAS)

- Dentro da miríade (crescente) de ferramentas, é selecionado um ambiente interativo (Jupyter notebooks) baseado em **Python** para desenvolvimento de nossos estudos, a partir do qual bibliotecas e recursos extras vão sendo apresentados/explorados no tempo...





FERRAMENTAS (SUGERIDAS)

- Ganho de produtividade...

Pandas

Dataframe Panda

```
import pandas as pd
acervobczm = pd.read_csv('../datasets/acervoaquisicao.csv', sep=';')
```

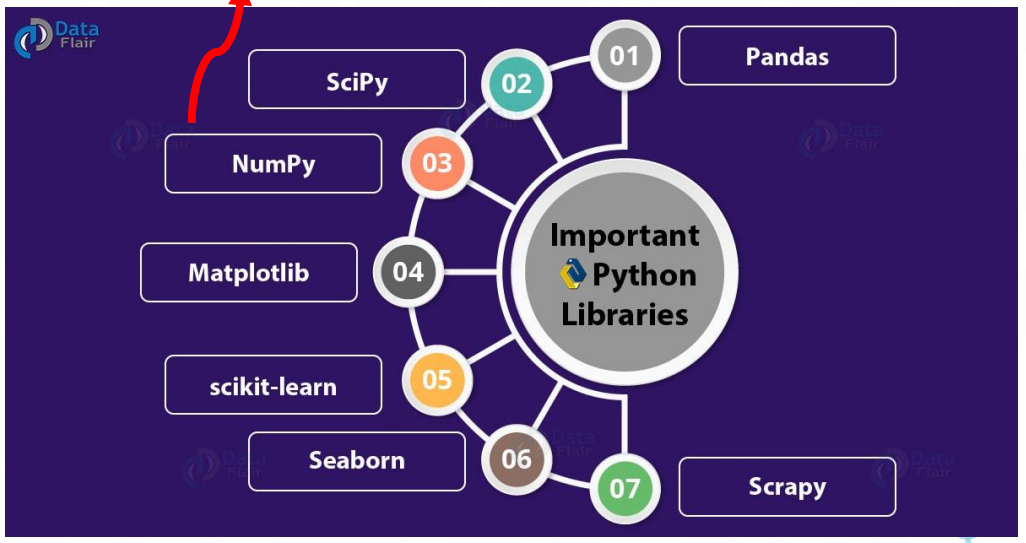
- Provê estruturas de dados de alto nível para dados estruturados ou tabulares (Series, DataFrames)
- Início: 2010 (chave para o Python ganhar espaço em CD
- Facilita reformatação, manipulação, agregação, seleção
- Tratamento flexível para dados ausentes

Dicionário Python

```
import csv

with open('../datasets/acervoaquisicao.csv', encoding='utf8') as f:
    reader = csv.DictReader(f, delimiter=';')
    for row in reader:
        print(row['titulo'])
```

- Provê objeto array multidimensional *ndarray* rápido e eficiente (alto desempenho) que as estruturas built-in
- Funções para processamento dos arrays
- Algebra linear, Fourier, aleatórios etc.
- Escrita e leitura de blocos ndarray em disco
- API C para interoperabilidade C-python numPy

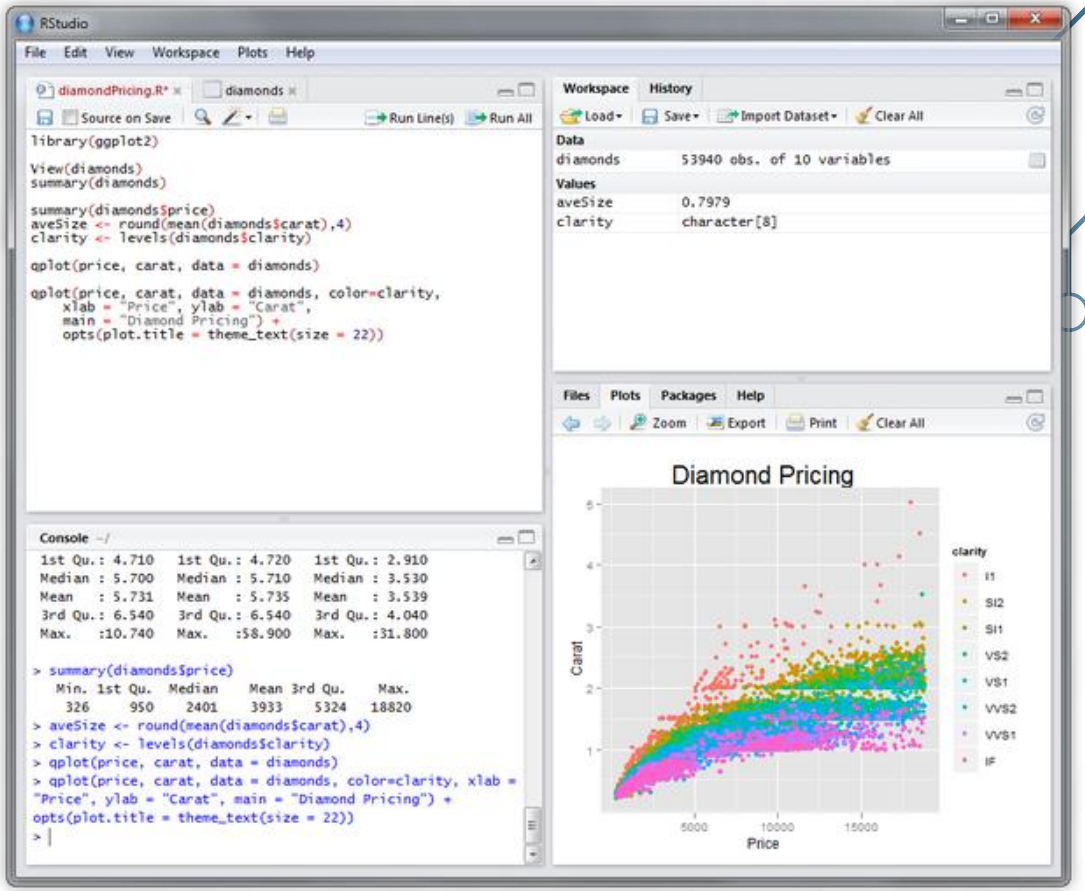


```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import statsmodels as sm
```

FERRAMENTAS (SUGERIDAS)

- Qual linguagem base devo utilizar?

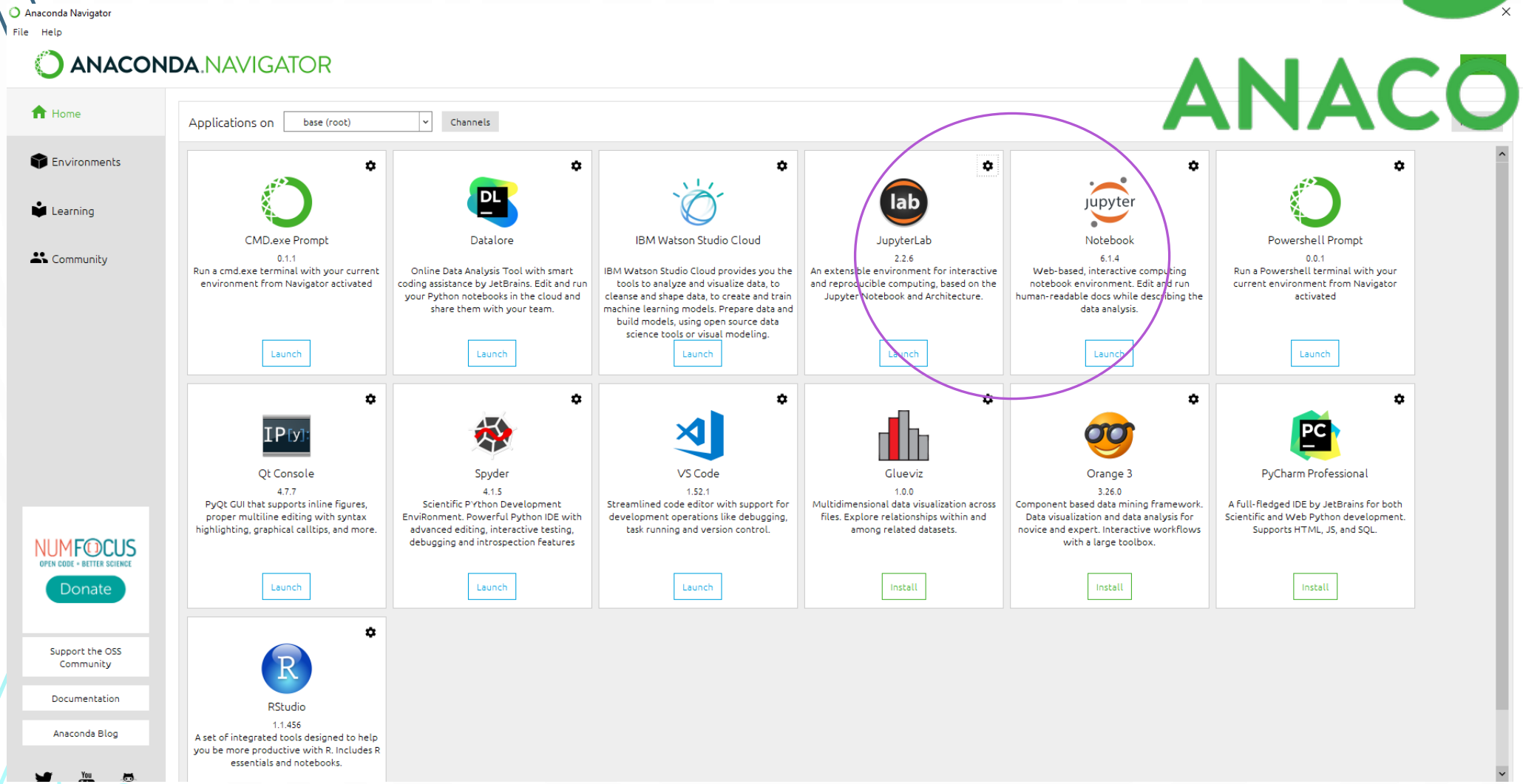
  Ver esta análise [aqui]



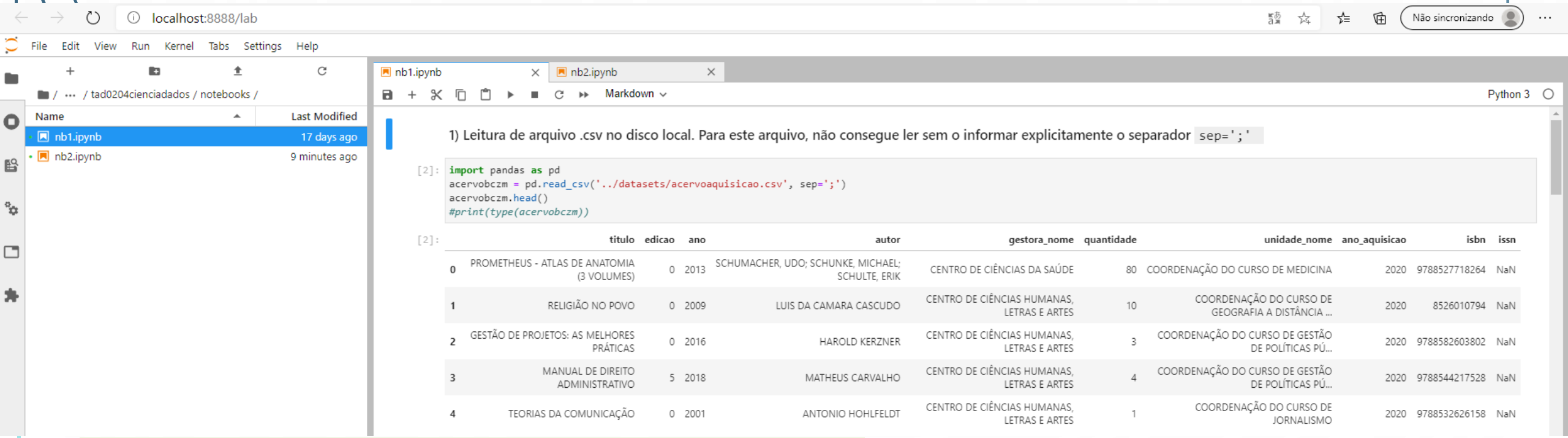
Ambas possuem suas especificidades, qualidades e, em geral, a escolha principal tem sido associada ao background do desenvolvedor, se em computação (*python*) ou de matemática/estatística (*R*)

OBS: *python necessário para a análise de dados (codificação)*

JUPYTER LAB (NOTEBOOK):



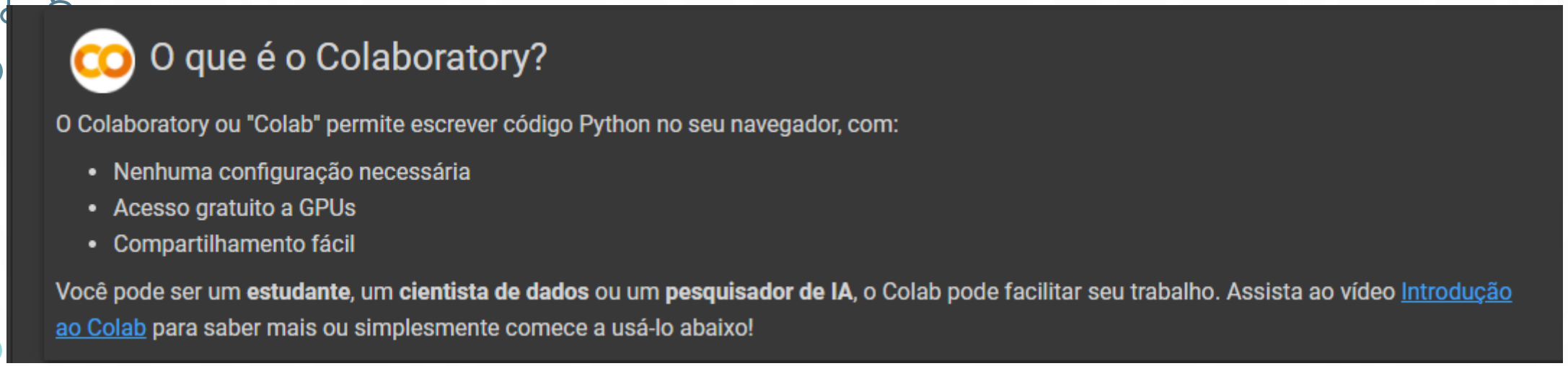
JUPYTER LAB (NOTEBOOK):



- Ambiente para execução interativa (Ipython), com mescl
  código interpretável, texto, imagens, LaTeX – lingua
  Markdown e HTML
- No Anaconda Prompt:
  - conda update jupyter
  - notebook jupyter

- Baseado no projeto Ipython (Fernando Pérez, 2001...)
- Projeto Jupyter (2014...)
- Desenvolvimento baseado no fluxo execução-exploração ao invés de edição-compilação-execução
- Em ciência de dados, as análises envolvem exploração, tentativa e erro e iteração

GOOGLE COLAB



- Desenvolvimento online com integração Gdrive e Github – interessante na modalidade REMOTA

```
Dito isto...
```

The Periodic Table of Data Science

Data-Science-Periodic-Table.pdf

An overview of key companies, resources and tools in data science (as of 4/12/2017)

Symbol → **Dc**
Name → DataCamp

Courses
Boot camps
Conferences
Data
Projects & Challenges, Competitions
Programming Languages & Distributions
Search & Data Management
Machine Learning & Stats
Data Visualization & Reporting
Collaboration
Community & Q&A
News, Newsletters & Blogs
Podcasts

Py Python	**Js** JavaScript	**Vb** Visual Basic	**Pgs** PostgreSQL	**Sli** SQLite	**Ah** Apache Hadoop	**W** Weka	**Bml** BigML	**Kn** Knime	**Sm** Spark MlLib	**Pb** Power BI	**Obi** Oracle BI	**Shn** Shiny	**Ddl** Domino Data Lab	**De** Data Science Experience
R R	**Cp** C++	**Sc** Scala	**Ar** Amazon Redshift	**Bq** Google BigQuery	**Hw** Hortonworks	**O** Oracle	**Dar** DataRobot	**Lib** LibSVM	**Ho** H2O	**Bo** BusinessObjects	**Alt** Alteryx	**Mpl** Matplotlib	**Nt** Nteract	**Rs** Rstudio
S SQL	**Pl** Perl	**Ca** Cassandra	**Hb** HBase	**Td** Teradata	**Cl** Cloudera	**Mss** Microsoft SQL server	**Rm** RapidMiner	**Mat** Mathematica	**Th** Theano	**Sp** Spotfire	**Sav** SAS Visual Analytics	**Ply** Plotly	**Ro** Rodeo	**Be** Beaker Notebook
B Bash	**Mr** Microsoft R Open	**P** Pig	**Mdb** Mongo DB	**To** Toad	**Aem** Amazon Elastic Mapreduce	**Spl** Splunk	**Cho** Chorus	**Mah** Mahout	**Aml** Azure Machine Learning	**Ql** Qlikview	**Po** PowerPivot	**Me** Microsoft Excel	**Spy** Spyder	**Ze** Apache Zeppelin
Mtl Matlab	**Cy** Canopy	**Im** Impala	**K** Kafka	**Ms** MySQL	**Mar** MapR	**Sr** Solr	**Tf** Tensorflow	**St** Stata	**D** D3	**Co** Cognos	**Gch** Google Charts	**Pe** Pentaho	**Dst** Data Science Studio	**Ju** Jupyter
J Java	**An** Anaconda	**Sp** Spark	**Hi** Hive	**Idb** IBM DB2	**Lu** Lucene	**El** ElasticSearch	**Sk** Scikit-Learn	**Da** Dato/Graphlab	**My** Microstrategy	**Aa** Adobe Analytics	**T** Tableau	**B** Bokeh	**Db** Databricks notebook	**Gh** Github

Dw Data.world	**Q** Quandl	**Fte** FiveThirtyEight	**Sa** Socrata	**Gp** Google Public	**Dg** Data.gov	**K** Kaggle
St Statista	**Uci** UCI Machine Learning Repository	**Wb** World Bank	**At** Academic Torrents	**Bf** Buzzfeed	**Dk** DataKind	**Dd** DrivenData

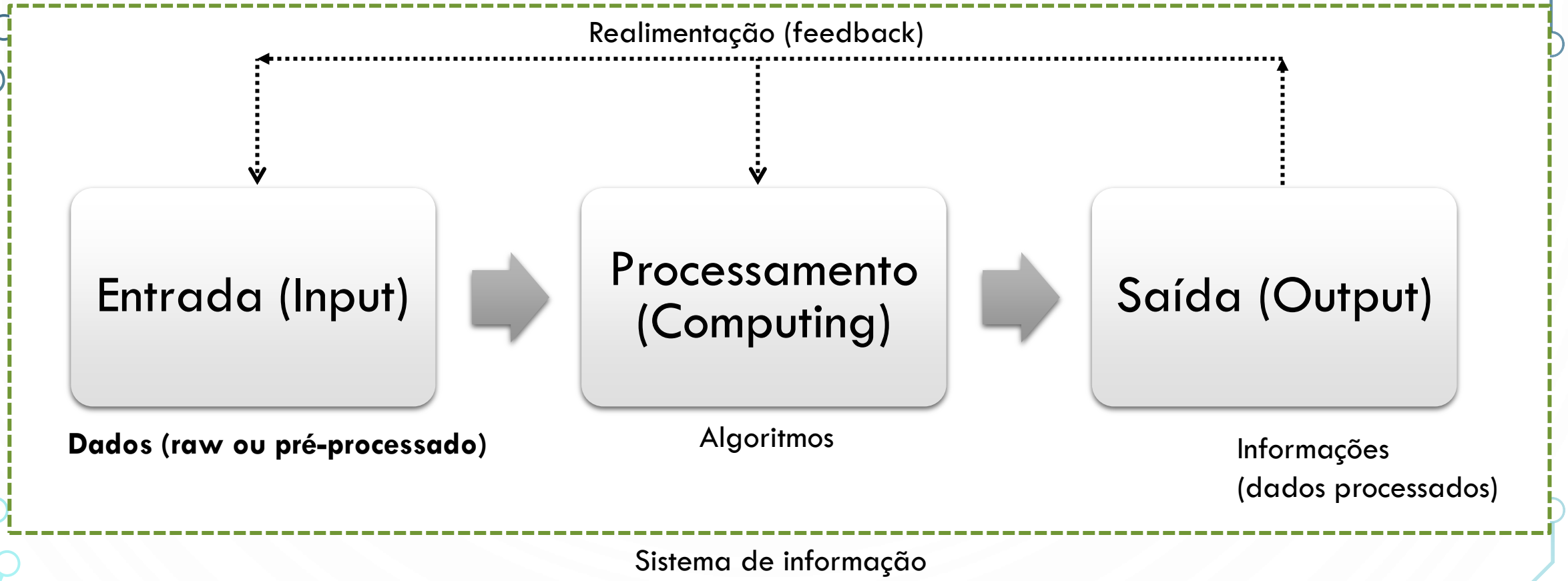
Re Reddit	**So** Stack Overflow	**Cv** Cross Validated	**Qu** Quora	**Av** Analytics Vidhya	**Dse** Data Science Stack Exchange
Mu Meetup	**Rdm** RDataMining				

Kdn KDnuggets	**Ibd** insideBIGDATA
Rb R-Bloggers	**Pp** PlanetPython
Hn HackerNews	**Dt** DataTau
Dsc Data Science Central	**Dsr** Data Science Roundup
Dsw Data Science Weekly	**Or** O'Reilly
Dr Data Elixir	**Pw** Python Weekly
Rw R Weekly	**Pd** Partially Derivative
Bds Becoming a Data Scientist	**Tm** Talking Machines
Ds Data Stories	**Dsk** Data Skeptic
Ld Linear Digressions	**Ns** Not So Standard Deviations

Dc DataCamp	**Ga** General Assembly	**Sd** Strata Data			
Sb SpringBoard	**M** Metis	**Od** ODSC			
Ex Edx	**Di** Data Incubator	**Tc** Tableau Conference			
C Coursera	**In** Insight	**U** UseR!			
Uda Udacity	**Dsa** NYC Data Science Academy	**Pd** PyData			
Ude Udemy	**G** Galvanize	**Paw** Predictive Analytics World			
Ps Pluralsight	**Dsg** Data Science for Social Good	**Kdd** ACM SIGKDD Conference			
Ly Lynda	**Dsy** Data Society	**Tpc** Teradata Partners Conference			
Tt TeamTreeHouse	**Dsj** Data Science Dojo	**Icd** IEEE International Conference on Data Mining			
Bdu Big Data University					

DataCamp
Learn Data Science By Doing

ELEMENTOS BASE: DADO, INFORMAÇÃO, SISTEMA, CONHECIMENTO



Informação: dado analisado e com algum significado X Conhecimento: informação interpretada, entendida e aplicada

- **Sabemos que os dados podem vir de várias FONTES (do ponto de vista de codificação entender os formatos e como atuar sobre os mesmos (ler, manipular)**

- **Exemplo de dado X informação X conhecimento**
  - **Projeto palmaS (UFRN@Tapioca, EMPARN – 11.2019...)**

Processo (algoritmo)

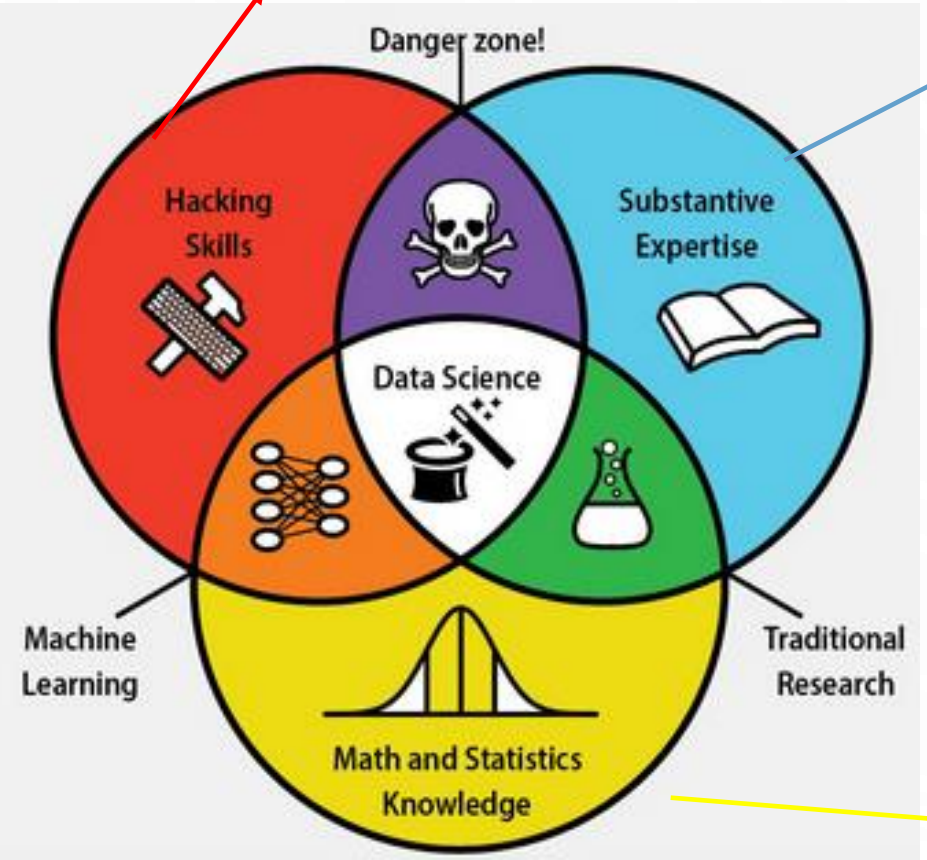
Conhecimento (negócio): tratamento A, em relação ao tratamento B possui eficácia X, devido à ..., sendo recomendado a aplicação da seguinte forma: ...

Dado: imagem da presença de uma praga chamada cochonilha de escama na palma forrageira após aplicação de determinado tratamento (produto etc.)

Informação: contagem das cochonillas fêmeas = 114

- **Portanto, estudos de ciência de dados em nosso contexto:**

Desenvolvedor



Negócio

Técnicas estatísticas

	Devido a sua natureza interdisciplinar, requer intersecção de habilidades (codificação, negócio, matemática e estatística)
	Necessário para trabalhar com massa de dados que precisa ser adquirida, limpa e manipulada
	Permite a escolha de métodos e ferramentas para extrair conhecimento a partir dos dados
	Conhecimento do negócio para demandar questões, estabelecer hipóteses e interpretar os resultados
	Pesquisa tradicional com aplicações estatísticas ao domínio do problema
	Isoladamente não requer conhecimento do negócio, pois pode-se limitar a aplicação de algoritmos, sem propósito individualmente
	Pode levar a análises incorretas, pela falta de métodos matemáticos e estatísticos rigorosos

Fonte: Kirk Bourne

- Na aquisição de dados, independente da fonte, pode comprometer o conhecimento, se os dados não apresentam tais características (não necessariamente todas, mas conjuntos razoáveis e lógicos)
  - **Relevantes**: são importantes para o tomador de decisões. Exemplo: se o preço de um tecido vai subir não é relevante para um fabricante de circuitos integrados
  - **Simples**: dados sofisticados e/ou detalhados podem não ser necessários. Sobrecarga de informações dificulta tomada de decisão ao invés de auxiliar
  - **Apresentadas no momento exato**: saber como foi o tempo semana passada não ajuda a definir que roupa devo usar hoje
  - **Verificáveis:** deve ser possível checar dados para garantir que estejam corretos, checando por exemplo várias fontes
  - **Acessíveis**: fácil acesso para usuários autorizados, com o formato correto e no momento correto
  - **Seguras**: evitar acesso não autorizado

Exemplo de demanda atual: **Organização Internacional da Uva e do Vinho**

oiv-digital-transformation-project-implementation-tor.pdf