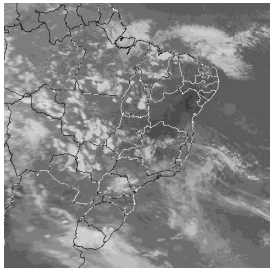


CONCEITOS DE *DATA SCIENCE*

Rafael Santos – rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/

Resumo

Data Science é mais um termo usado para descrever o processo de transformação de dados em conhecimento. É diferente de e ao mesmo tempo expande campos já conhecidos como estatística, *analytics*, mineração de dados, descoberta de conhecimento em bases de dados, com ênfase no desenvolvimento de soluções que integram os processos da transformação de dados heterogêneos, em diferentes escalas, incompletos e possivelmente mal-estruturados em conhecimento.

Neste mini-curso introdutório veremos alguns conceitos de *Data Science*, definições de seus proponentes, conhecimentos técnicos que definem um *data scientist* e como adquiri-los; e exemplos do que é (ou não) *Data Science*.

O que é *Data Science*?

Hype

By 2018, the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge.

Susan Lund et al., "Game Changers: Five Opportunities for US Growth and Renewal," McKinsey Global Institute Report, July 2013.
http://www.mckinsey.com/insights/americas/us_game_changers

Duas definições iniciais

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

- "What is a 'Data Scientist'? An analyst who lives in California." – *were-bycicle*

Envolve dados, mas...

... não somente gerenciamento de bases de dados!

- Aplicações baseadas em dados são comuns.
  - Indispensáveis em algumas atividades!
- Usar (coletar, armazenar, publicar) dados não é *data science*. É preciso agregar valor aos dados e permitir novas formas de uso.
  - Exemplo: base de dados CDDB.
- *Data science* possibilita a criação de produtos de dados.

https://www.oreilly.com/ideas/what-is-data-science

Envolve programação, mas...

... não somente programação e novas tecnologias.

- ***The key word in "Data Science" is not Data, it is Science***
  - "Não é *Big Data*, só tem X gigabytes."
  - "Meus dados são maiores que os seus."
  - "Eu sei Hadoop, você sabe?"
- Menos ênfase em tamanho e tecnologia, mais em aplicação de tecnologias para obter respostas sobre os dados.

http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

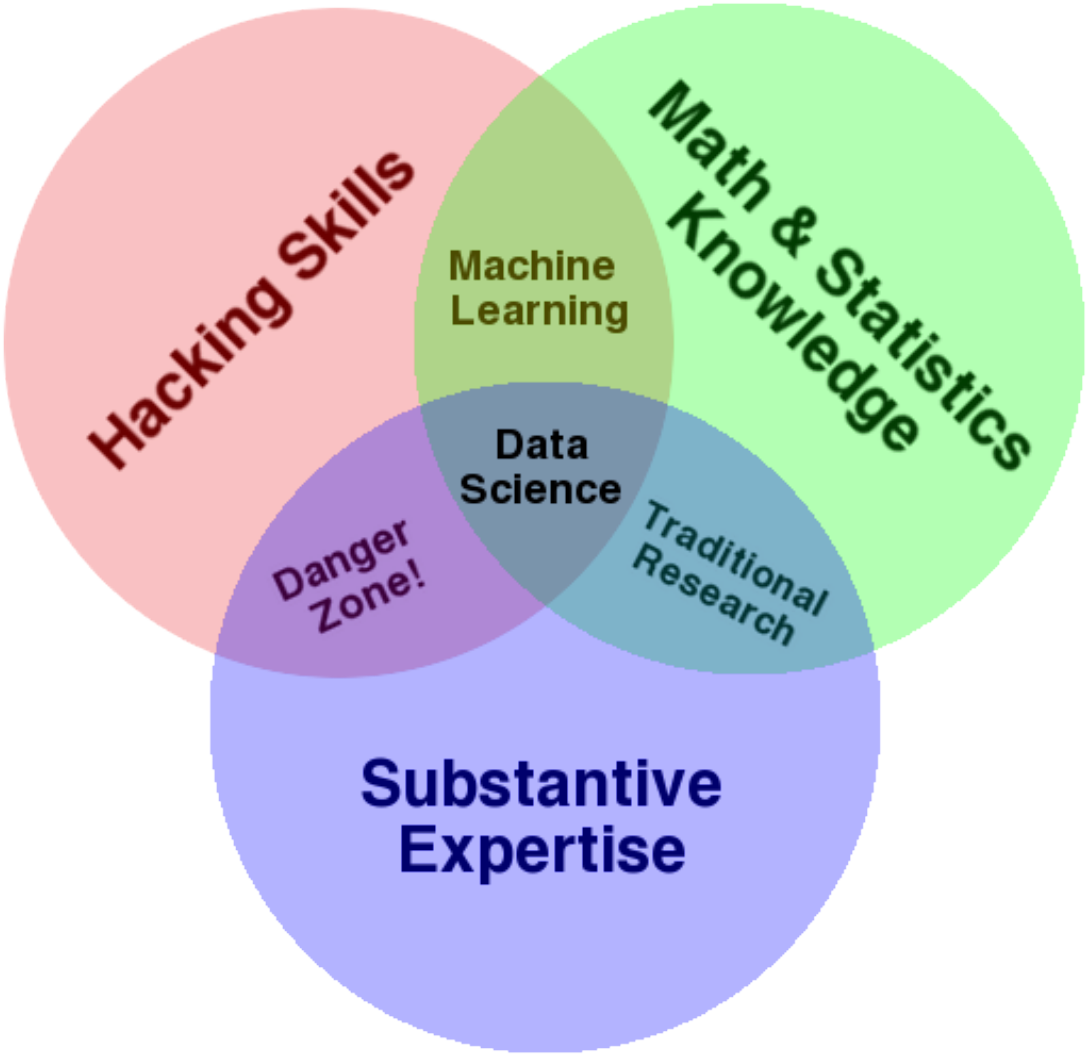
Envolve estatística, mas...

...não puramente estatística tradicional.

- Pode ser necessário escalonar métodos tradicionais.
- É necessário prototipar em linguagens como R e Python.
  - Aplicações *point-and-click* não seriam eficientes.
- São precisos conhecimentos em combinação de fontes de dados, análise exploratória de dados, HPC, visualização, etc.
- É preciso apreciar casos do mundo real!

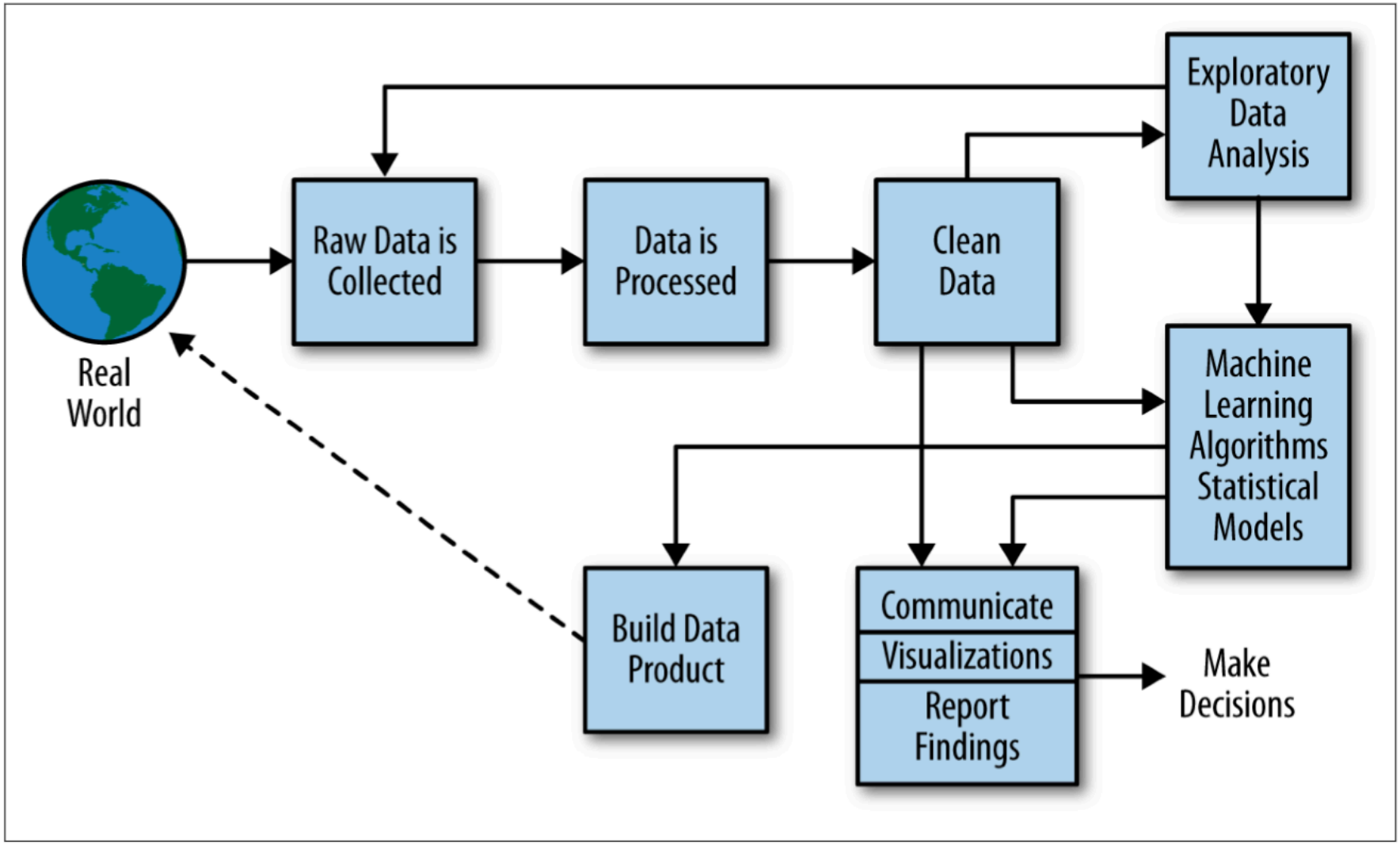
http://magazine.amstat.org/blog/2013/07/01/datascience/

É tudo isto (e ainda mais?)



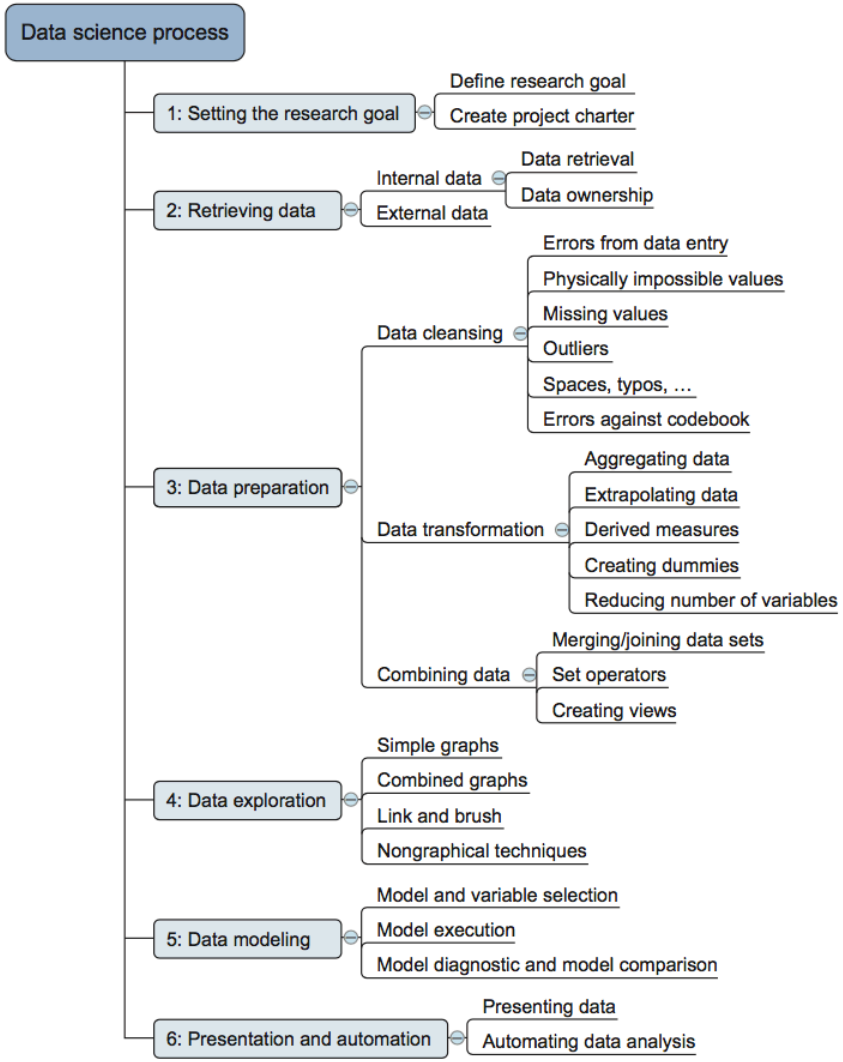
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

É um processo (?)



Doing Data Science, Rachel Shutt and Cathy O'Neil, OReilly, 2014

É um processo (?)



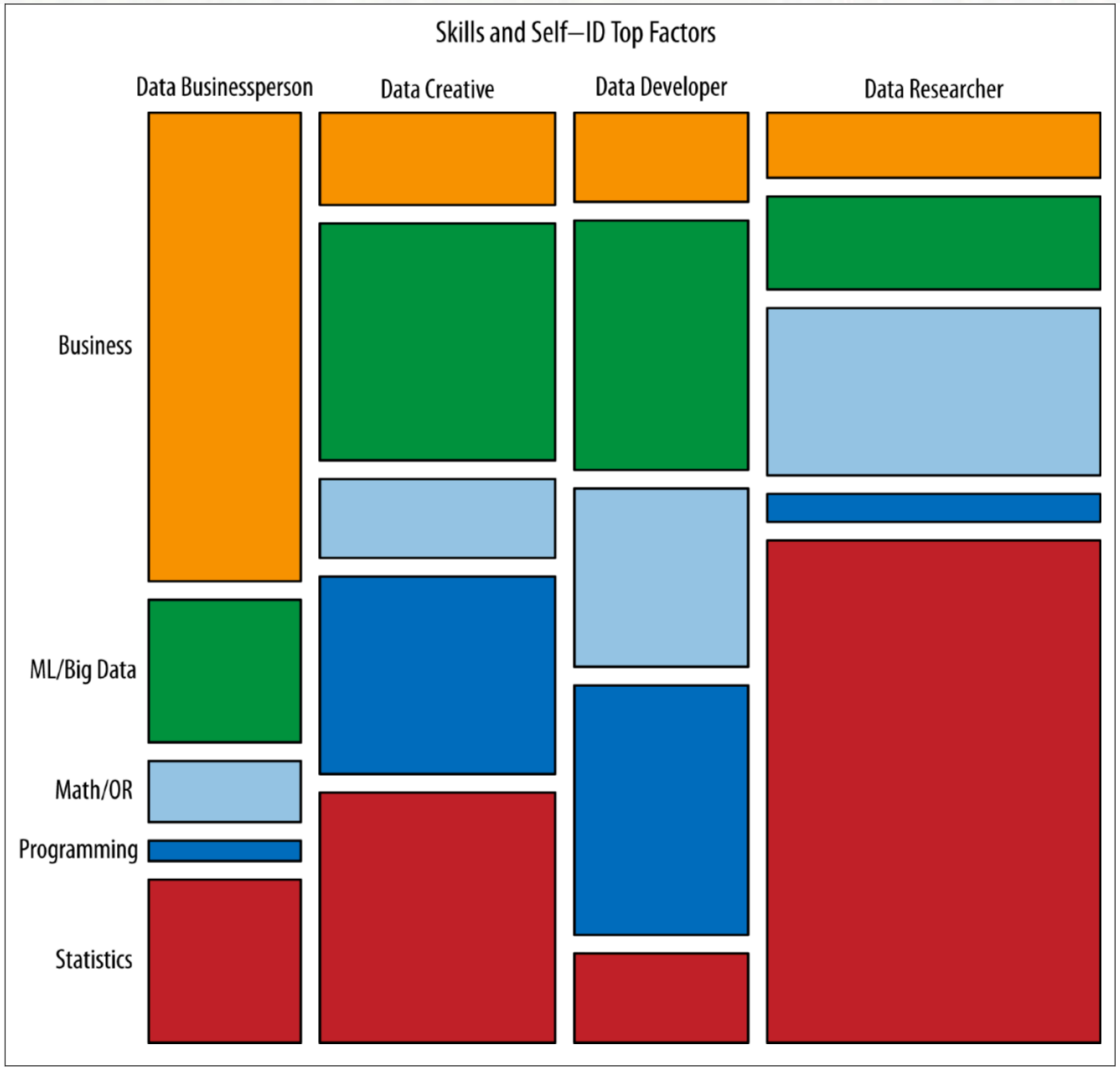
Introducing Data Science, Davy Cielen, Arno Meysman, Mohamed Ali, Manning, 2016

Então você quer ser um *data scientist*…

O que é mesmo um *Data Scientist*?

- *Analyzing the Analyzers*:
  - Alguém que sabe algo sobre estatística, programação e visualização?
  - Alguém com experiência em extrair informações de dados?
  - Precisamos de uma descrição mais específica (“doutor”, “atleta”, “*data scientist*” são termos genéricos)!
  - A definição depende do problema em questão.
- Pesquisa com 250 voluntários selecionados.

O que é mesmo um *Data Scientist*?

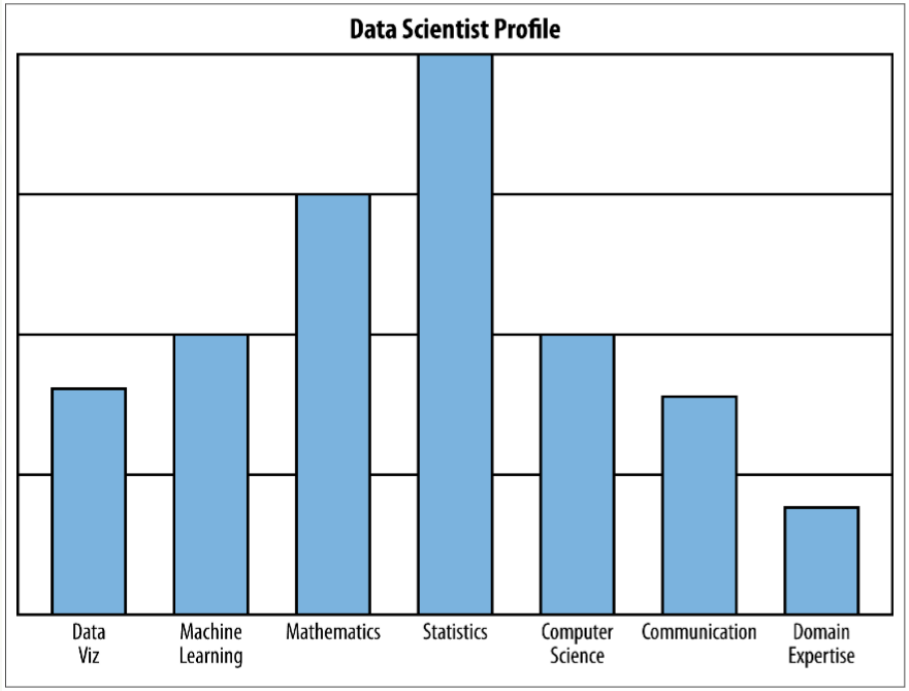


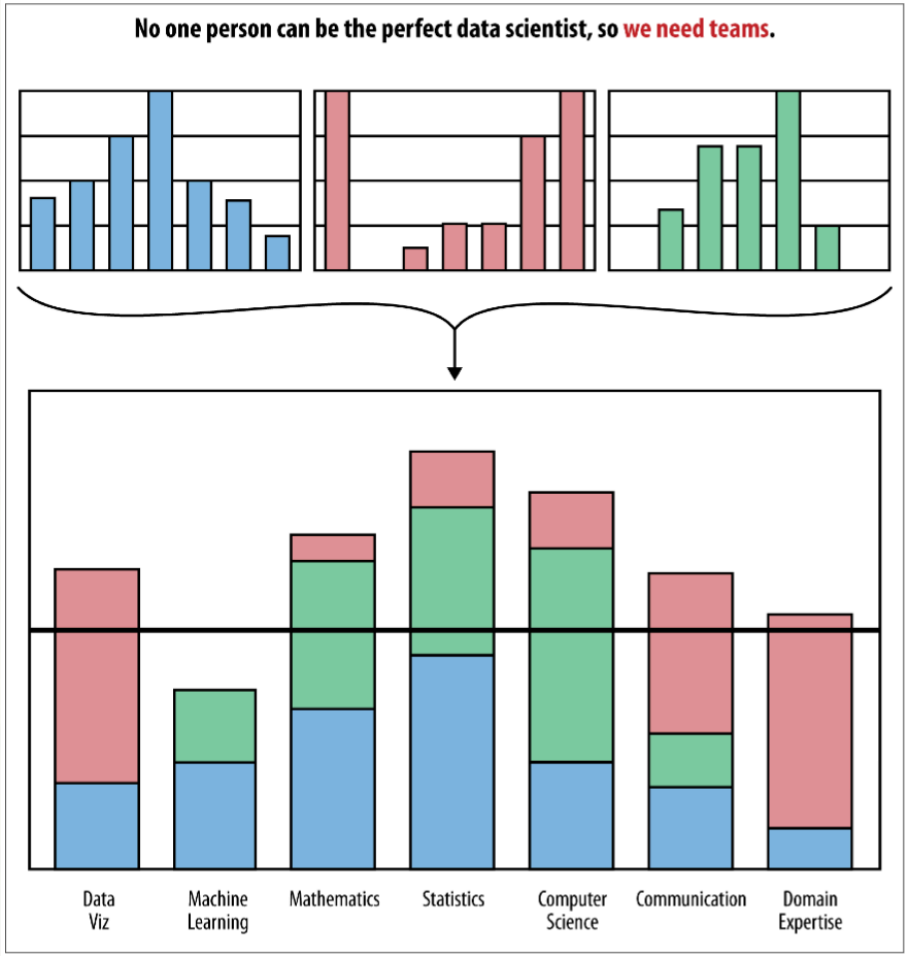
O que é mesmo um *Data Scientist*?

- *Analyzing the Analyzers*: evidência do *T-Shaped Data Scientist*
- Gama abrangente de conhecimentos mas conhecimentos profundos em uma área.
  - Mais aptos para tarefas e grupos interdisciplinares.
  - Mais eficientes em sua área de conhecimento mais profunda.
- Outro levantamento indica três categorias:
  - Curadoria de dados.
  - *Analytics* e visualização.
  - Redes e infraestrutura.

Jeffrey Stanton et al, Interdisciplinary Data Science Education, http://pubs.acs.org/doi/abs/10.1021/bk-2012-1110.ch006

T-Shaped Data Scientist





Doing Data Science, Rachel Shutt and Cathy O'Neil, OReilly, 2014

Então você quer ser um *data scientist*...

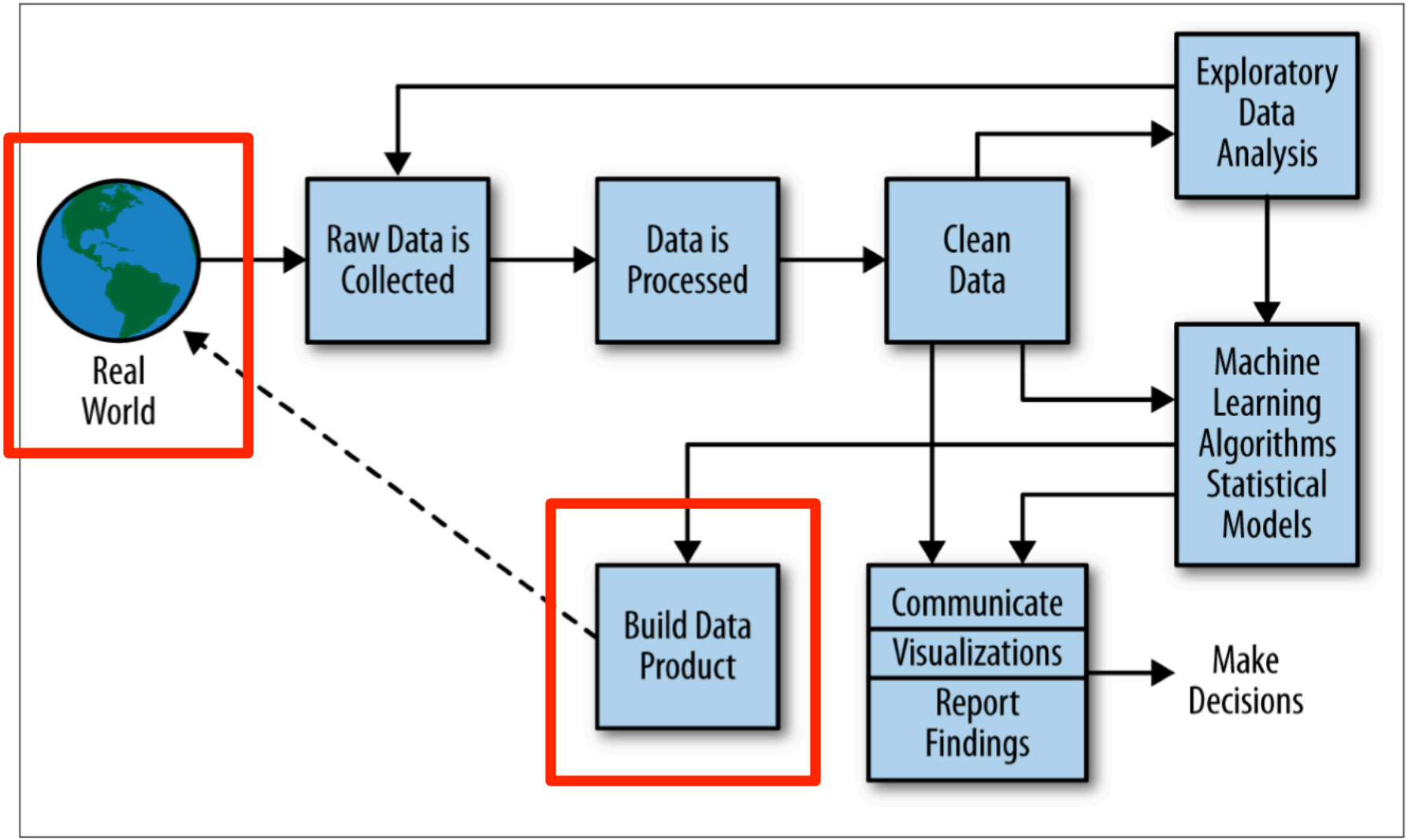
- Você...
  - ☑ Tem acesso (ou pode ter) a coleções de dados temáticos em diferentes graus de organização e/ou
  - ☑ Entende o suficiente de linguagens como R, Python, tecnologias SQL/NoSQL, sistemas distribuídos, etc. e/ou
  - ☑ Entende o suficiente de modelagem, testes e características de algoritmos de análise.
- ...provavelmente já tem por onde começar.

Então você quer ser um *data scientist (como?)*

Skills

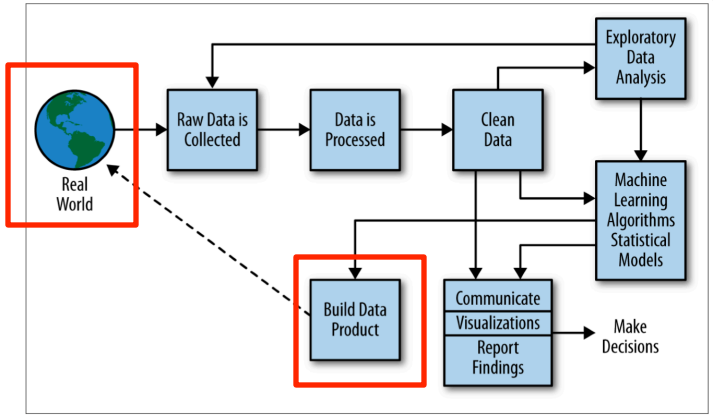
- Uma lista de conhecimentos e capacidades...
  - ...não exclusiva: novas tecnologias aparecem o tempo todo.
  - ...com viés: é saudável questionar algumas ideias.
  - ...potencialmente redundantes: o *data scientist* tem que saber como jogar em várias posições em vários times.
  - ...individualmente impossíveis: *"Rockstar Programmer"*, *"Rockstar SysAdmin"*, *"Rockstar Analyst"?*
  - ...não necessariamente técnicos: *data science* deve envolver aspectos do mundo real.

Skill: Entender o Problema

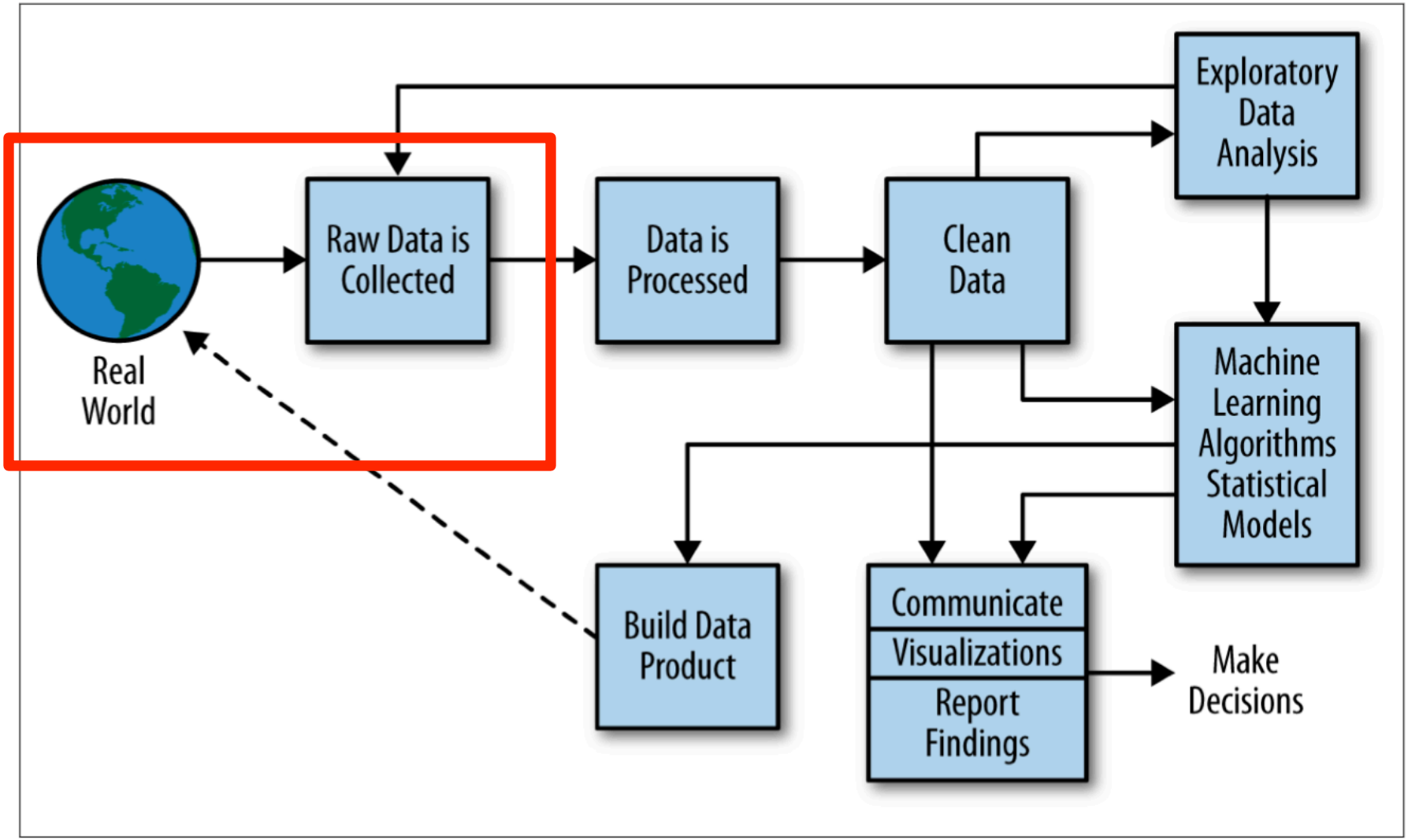


Skill: Entender o Problema

- Ao menos o suficiente para se comunicar com quem tem o problema!
  - DS é inerentemente interdisciplinar!
- Que dados existem?
- Que dados deveriam existir?
  - Produto de Dados!
- Alerta: não devemos fazer *data science* sem entender o problema!

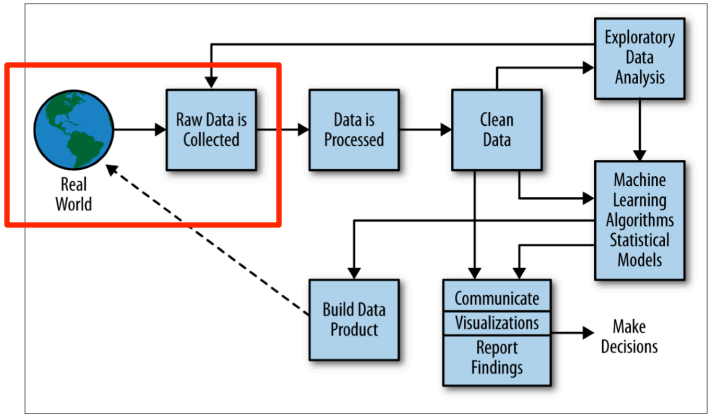


Skill: Achar e Organizar Dados



Skill: Achar Dados

- Achar = localizar, identificar, etc.
- Que dados existem relacionados ao problema em questão?
- Que dados estão disponíveis?
  - É preciso coletar mais/outros?
  - Como acessar os dados?
  - Existem formas prontas?
  - Preciso replicar/amostrar?
  - Qual é o volume destes dados e no que isto impacta a coleta?



Big Data

- O que é *Big Data*?
- Tradicional: qualquer conjunto de dados muito grande...
  - ...para análise simples?
  - ...para processamento efetivo?
  - ...para armazenamento total?
- Medidas em {Gb,Tb,Pb} podem refletir o tamanho dos dados mas não o do problema.

Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

Big Data

- "3 Vs":
  - **Volume**: quanto de armazenamento é necessário. Depende de capacidade tecnológica: armazenamento, capacidade de processamento.
  - **Velocidade**: quão rapidamente os dados devem ser recuperados/ processados/analisados.
    - Em quanto tempo temos que ter respostas?
    - Por quanto tempo temos que ter respostas?
  - **Variedade**: quão heterogêneos os dados são, quantas medidas por registro, como é feita a conexão entre fontes, etc.
    - Forma, formato, estrutura, representação, etc.

Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

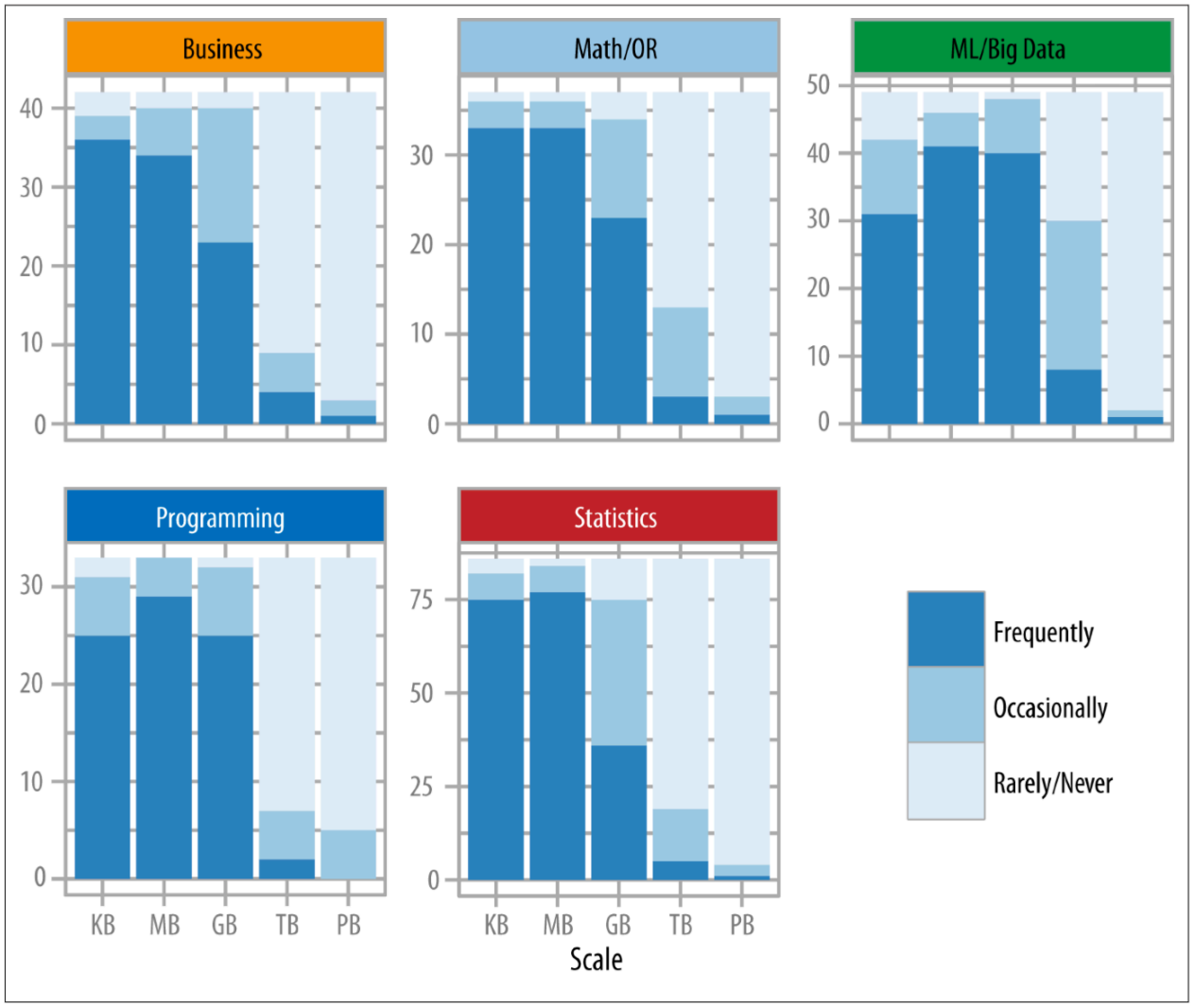
Big Data

- "3 Vs": Volume, Velocidade, Variedade
- **Valor**: se estamos coletando dados científicos é porque é caro e/ou vale a pena!
- **Validade**: os dados são confiáveis?
  - Proveniência, Completude, Metadados... dependem do *valor*.
- **Variabilidade**: significado ou origem mudam com o tempo?
- **Vocabulário**: o que mais é necessário para entender os dados?

Big Data Lessons from the Climate Science Community, Seth McGinnis, 2016

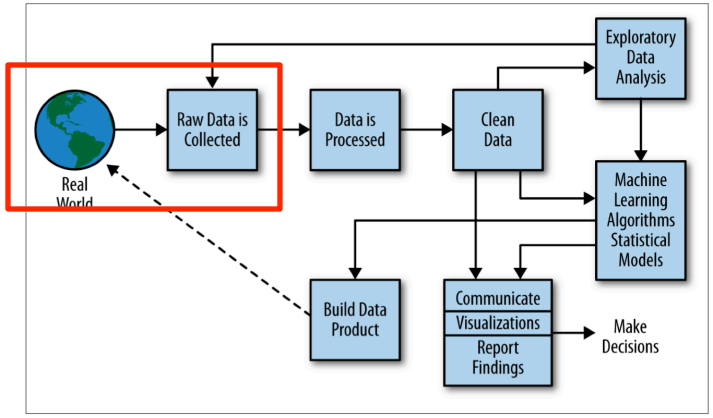
Big Data

- *Analyzing the Analyzers*: e *Big Data*?



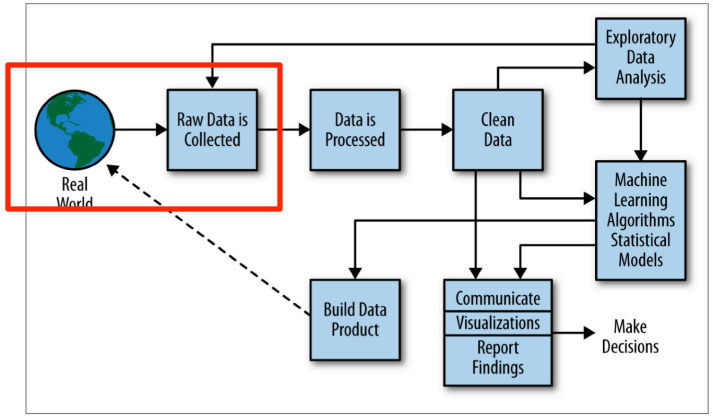
Skill: Entender a Organização dos Dados (1)

- Antes do processamento:
  - Como os dados são representados?
  - Tabelas, documentos, imagens, relações, mistura?
  - Os dados estão em um formato útil para resolver nosso problema?
    - Como transformar?
    - Qual é o tamanho desta tarefa?



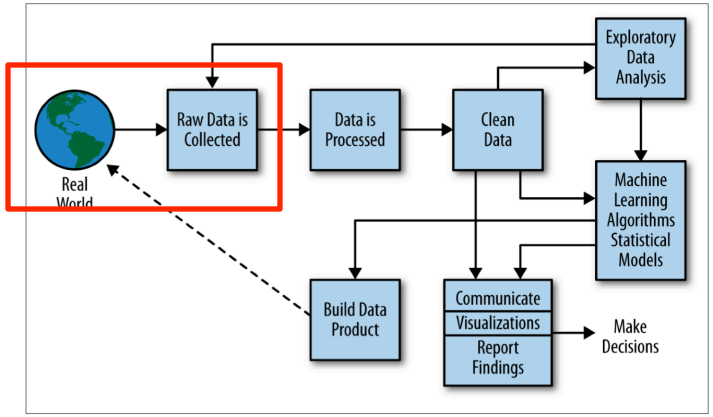
Skill: Entender a Organização dos Dados (2)

- Precisamos destes dados com organização específica?
  - De onde eles vem?
  - Coletaremos repetidamente?
  - Precisamos de proveniência, anotações?
  - O que precisa ser preservado? O que precisa ser aumentado? Como?
  - Terão uma vida à parte das fontes originais?



Skill: Entender a Organização dos Dados (3)

- Se precisamos deles de forma separada, como os organizaremos?
  - Coleções de {documentos, imagens, arquivos, tabelas}?
  - *Big Data?* Que tecnologias de armazenamento e/ou processamento são necessárias?



Que tecnologias são necessárias?

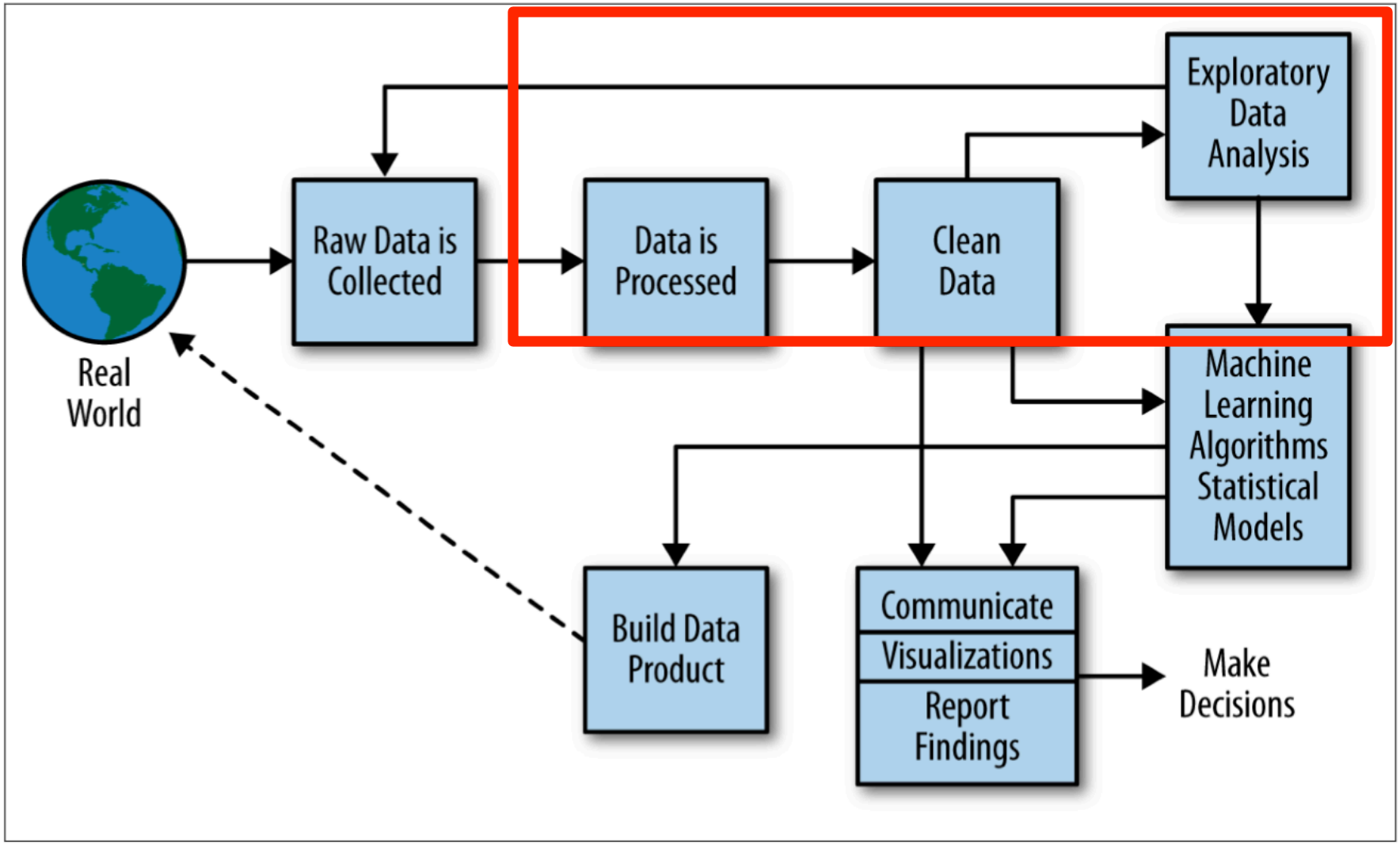
- Muitas opções, cada uma com diferentes capacidades e limitações...
- Ainda estamos falando de *skills*?
  - Conheça SQL: excelente para dados bem estruturados.
    - Na medida em que estrutura deve ser mais versátil tabelas ficam mais complexas...
  - Conheça alguns bancos de dados NoSQL.
    - NoSQL pode ser mais flexível para dados com estruturas diferentes.
    - Várias abordagens/implementações/modelos...

NoSQL

- Baseados em pares chave/valor
  - Arrays associativos, mapas ou dicionários.
  - Redis, Riak, Memcached, etc.
- Baseados em colunas
  - Amplia chave/valor para várias colunas.
  - Cassandra, HBase
- Baseado em Documentos
  - Permite hierarquia de chaves/valores/documentos.
  - Couchbase, CouchDB, MongoDB
- Baseados em Grafos
  - Armazena nós e relações entre nós.
  - Neo4J, OrientDB

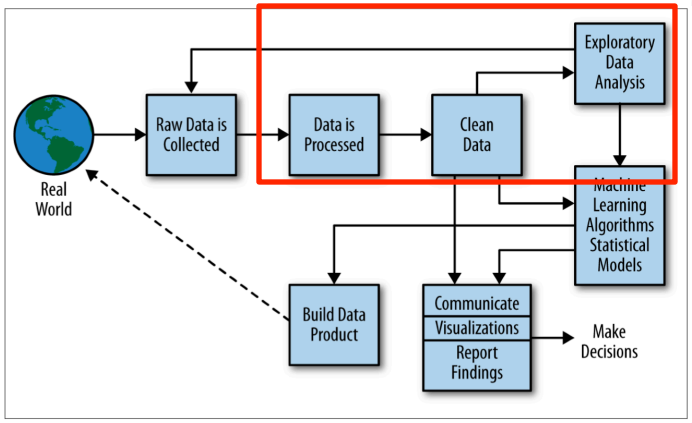
https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models

Skill: Análise (*Hacking*)



Skill: Análise (*Hacking*)

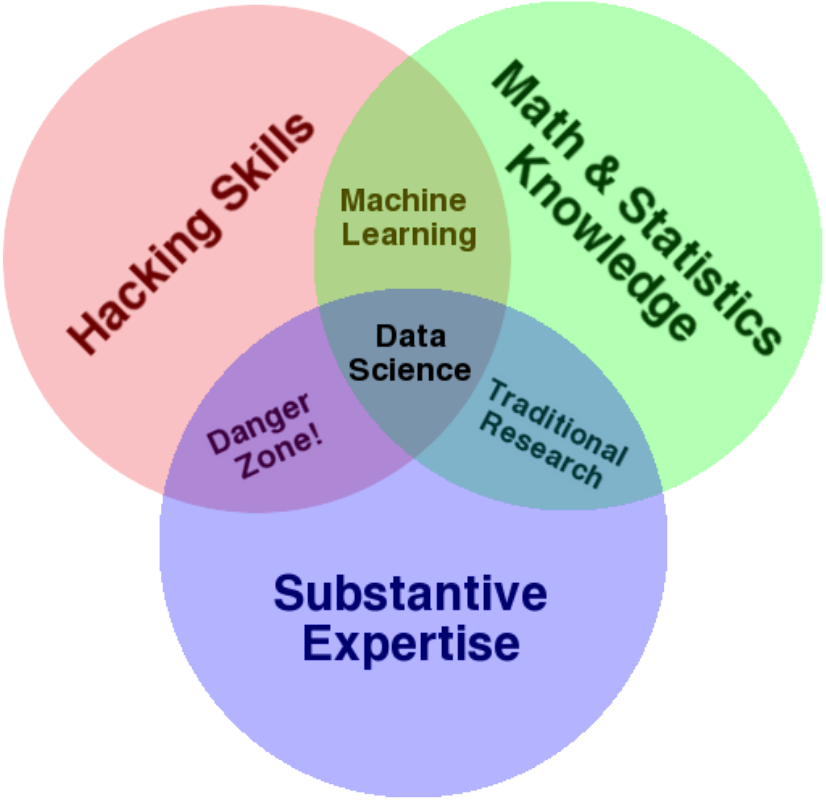
- Temos os dados. O que fazer agora?
  - Sabemos o que queremos achar?
  - Conhecimentos básicos em estatística/modelagem são muito úteis.

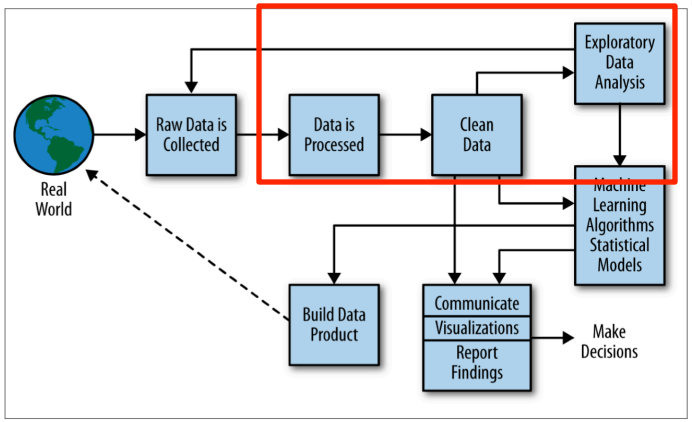


- Em caso de não saber... *explore os dados*!
  - Crie gráficos de vários tipos (de acordo com os dados).
  - Calcule estatísticas básicas.
  - Avalie que tipo de informação pode ser extraída dos dados.

Skill: Análise (*Hacking*)

- Lembrete importante!





Skill: Análise (*Hacking*): Python

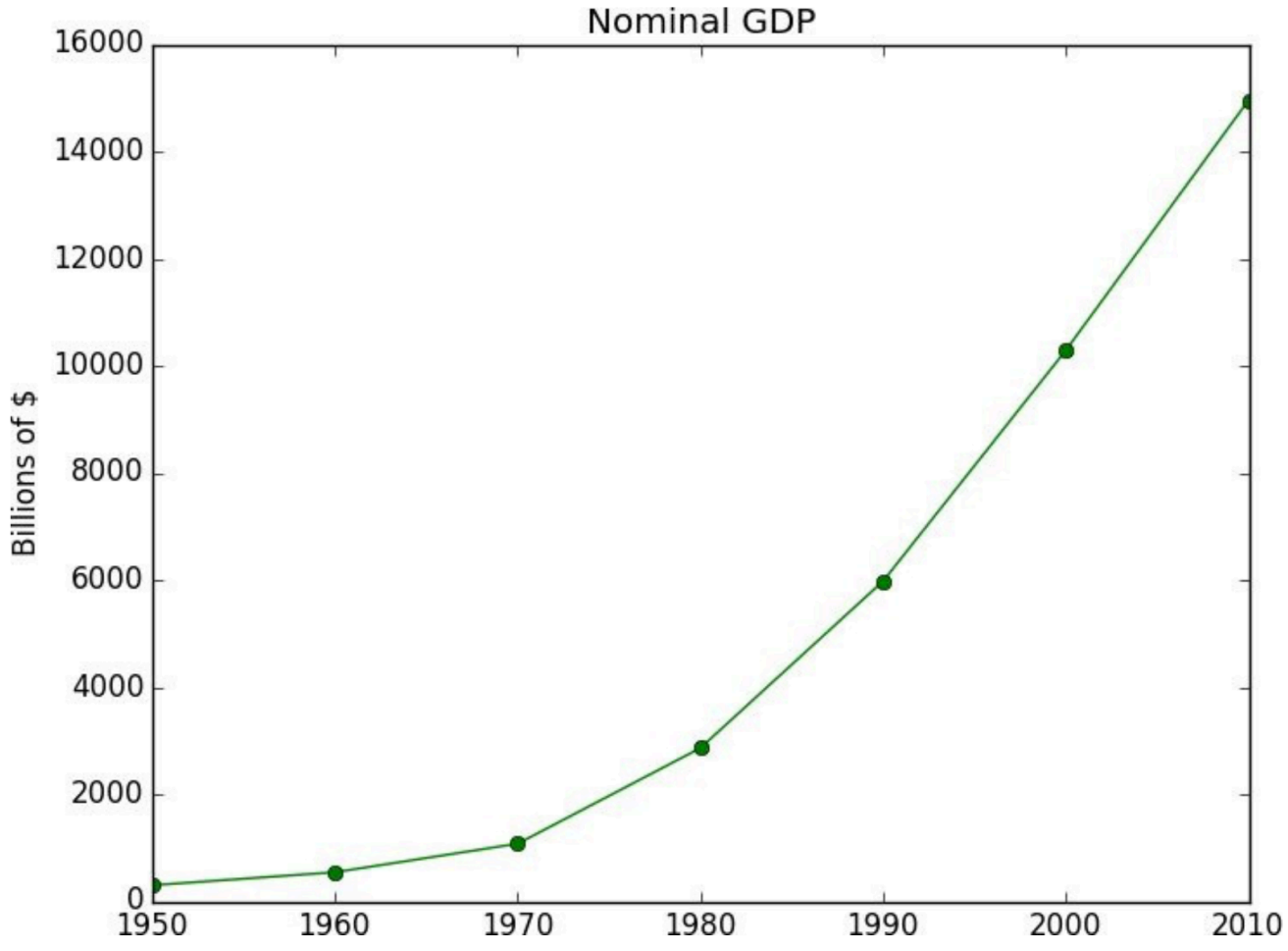
- Exemplo básico

```
from matplotlib import pyplot as plt

years = [1950, 1960, 1970, 1980, 1990, 2000, 2010]
gdp = [300.2, 543.3, 1075.9, 2862.5, 5979.6, 10289.7, 14958.3]

# create a line chart, years on x-axis, gdp on y-axis
plt.plot(years, gdp, color='green', marker='o', linestyle='solid')

# add a title
plt.title("Nominal GDP")

# add a label to the y-axis
plt.ylabel("Billions of $")
plt.show()
```

Data Science from Scratch. Joel Grus, O'Reilly, 2015

Skill: Análise (*Hacking*): Python

- Exemplo básico



Skill: Análise (*Hacking*): Python

- Muitas bibliotecas interessantes:
  - **NumPy**: arrays, operadores, IO, integração com C, C++.
  - **SciPy**: computação científica, matrizes esparsas, processamento de sinais, etc.
  - **pandas**: facilidades para processamento de dados estruturados (ex. tabelas, séries temporais, modificações, seleções, conversões).
  - **matplotlib**: gráficos e visualização.
  - **iPython**: conceito de notebook, facilita prototipagem, documentação e possibilita pesquisa reprodutível.

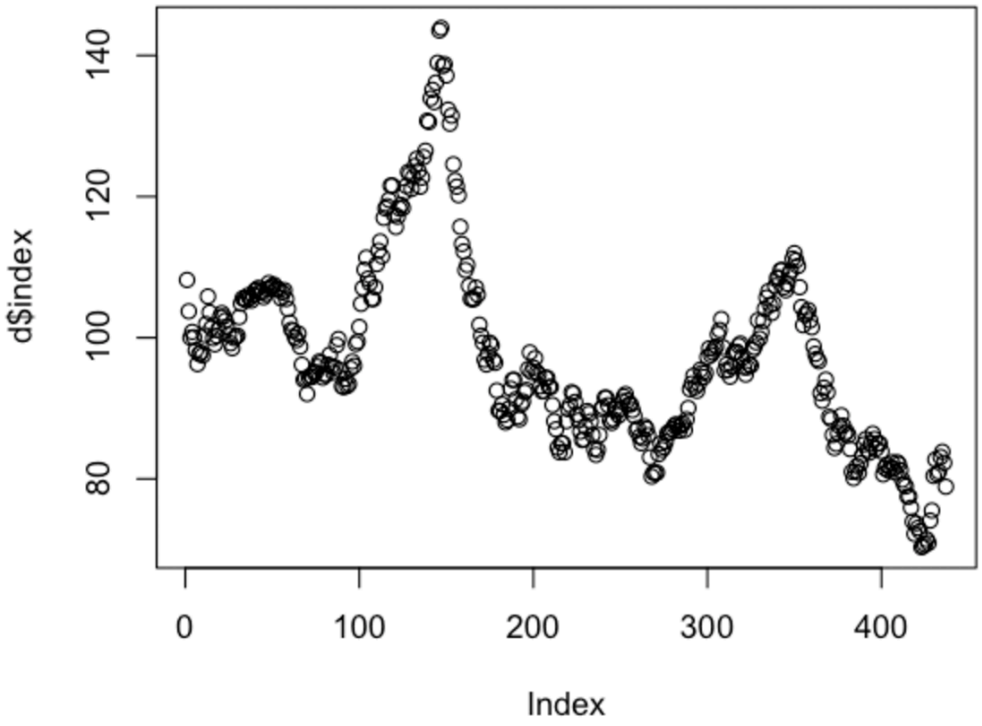
Críticas a Python

- Python 3 é a última versão; não é totalmente compatível com Python 2.7.
  - Muitas bibliotecas interessantes funcionam melhor com 2.7!
- Existe redundância em algumas bibliotecas, e algumas não são mantidas.
  - PyPi ajuda.
- *There should be one – and preferably only one – obvious way to do it.*

Skill: Análise (*Hacking*): R

- Exemplo básico:

```
> d = read.table('dollar_vs_major_currencies_index.txt',
            header=F, sep="t", col.names=c("month", "index"))
> dim(d)
[1] 437   2
> head(d)
     month    index
1 JAN 1973 108.1883
2 FEB 1973 103.7461
3 MAR 1973 100.0000
4 APRimg 1973 100.8251
5 MAY 1973 100.0602
6 JUN 1973  98.2137

> plot(d$index)
```



Skill: Análise (*Hacking*): R

- Longa tradição em estatística e análise.
- Vasta gama de algoritmos de mineração de dados.
- Muitos pacotes organizados no CRAN.
- RStudio!

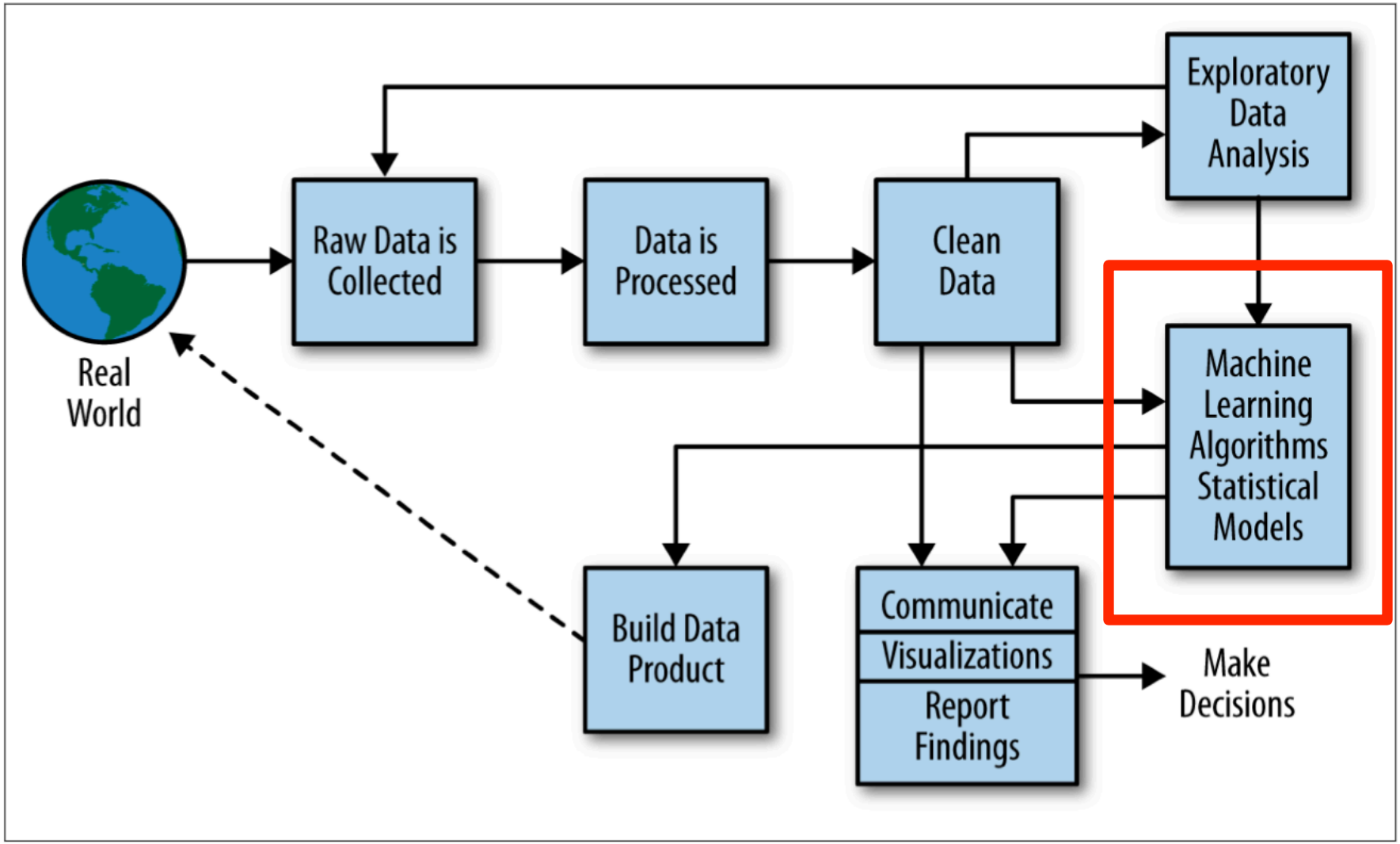
Críticas a R

- Não tão amigável quanto outras linguagens de alto nível.
- Não escala bem.

Só R e Python?

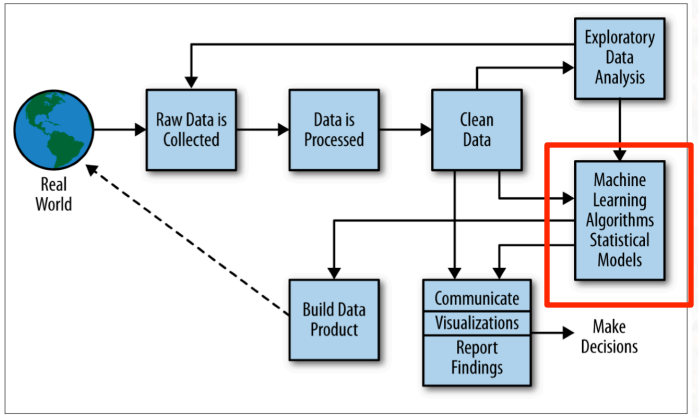
- Nada impede de usar outras, mas...
  - Pacotes existentes (vale a pena reinventar a roda?)
  - Adoção por muitos grupos e empresas.
  - Comunidades existentes (ex. stackoverflow.com).

Skill: *Machine Learning, Models*



Skill: *Machine Learning, Models*

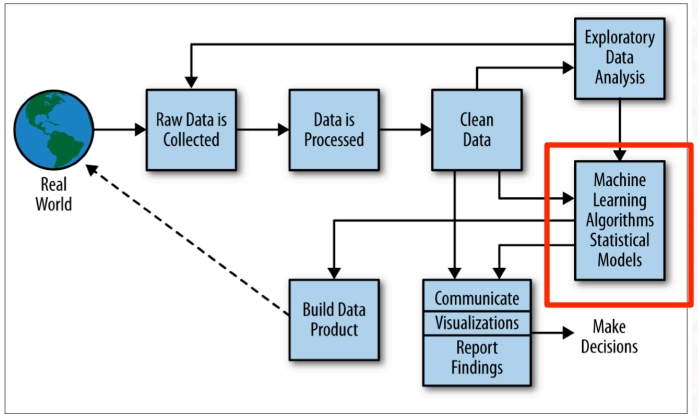
- O que posso aprender a partir de meus dados?
- *Exploratory Data Analysis* deve servir para dar indícios da natureza dos dados e de que conhecimento podemos extrair deles.

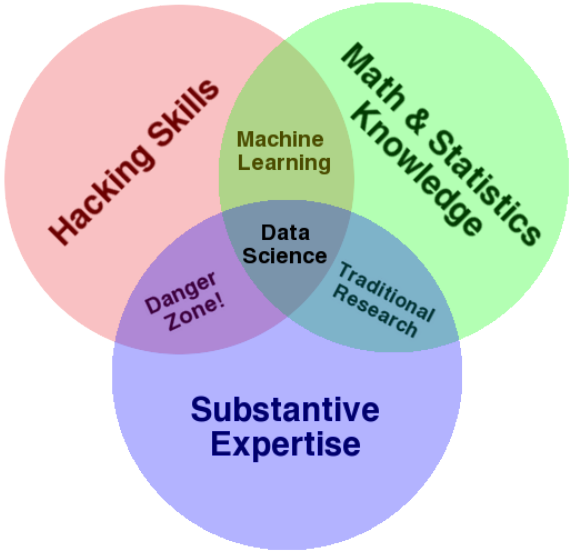


- *Machine Learning*, *Data Mining*, etc. podem servir para criar modelos que descrevam os dados.

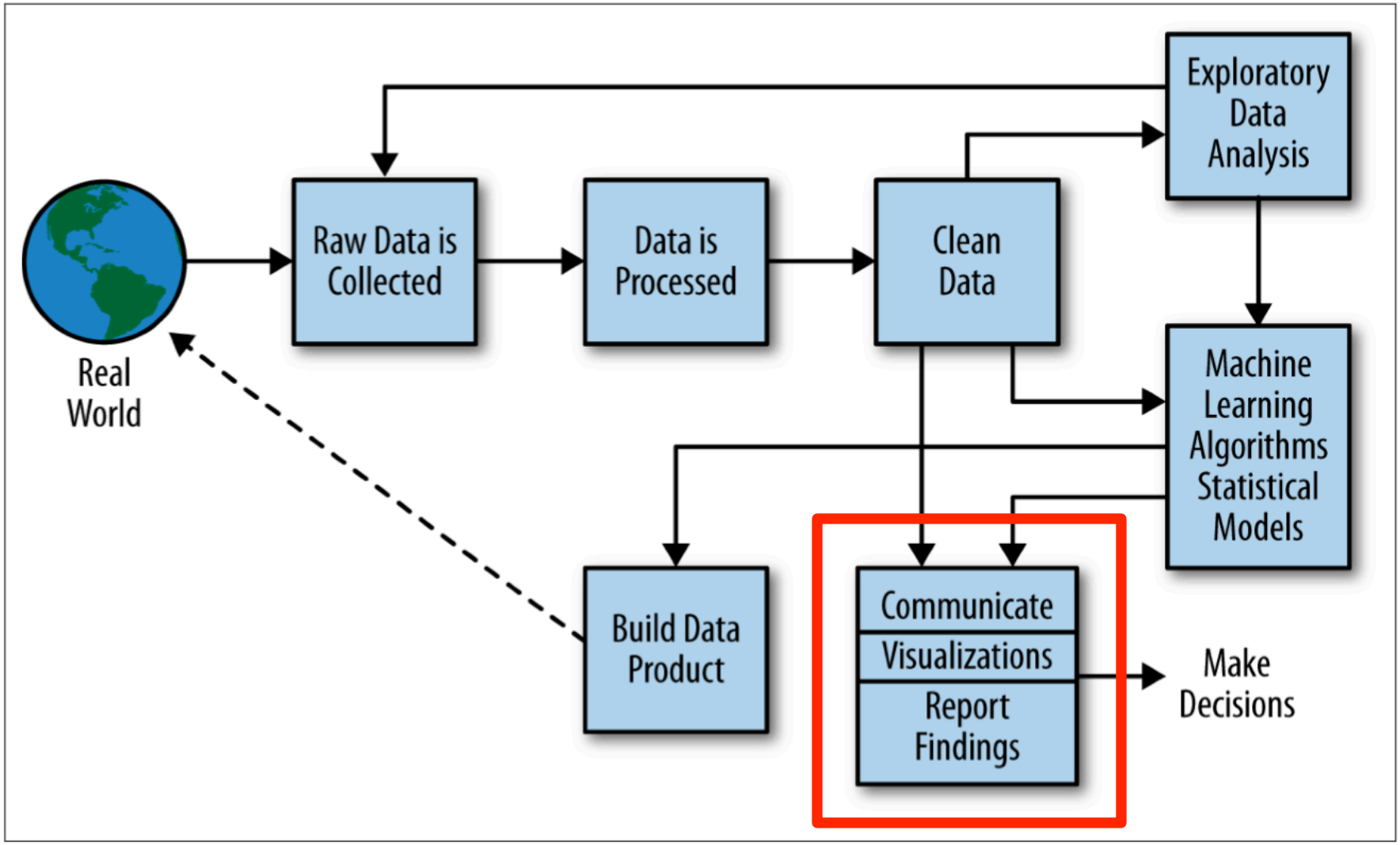
Skill: *Machine Learning, Models*

- **Cuidados**:
  - Modelos podem ser bem mais complexos do que EDA sugere.
  - Existem muitas técnicas, algoritmos, variações.
  - Interpretabilidade e validação de modelos é imprescindível!
  - Escalabilidade pode ser um problema!



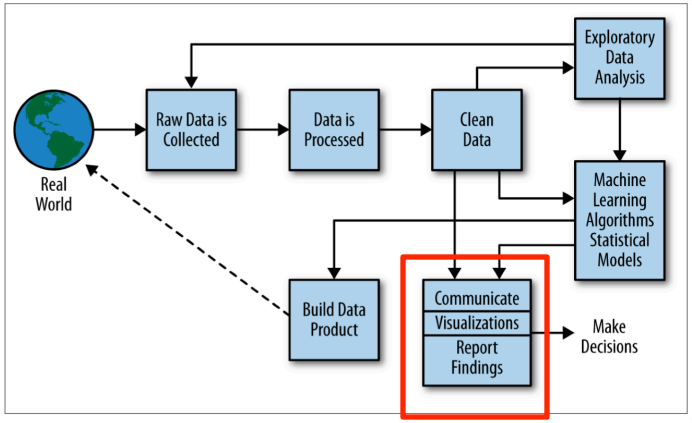


Skill: Comunicação de Resultados



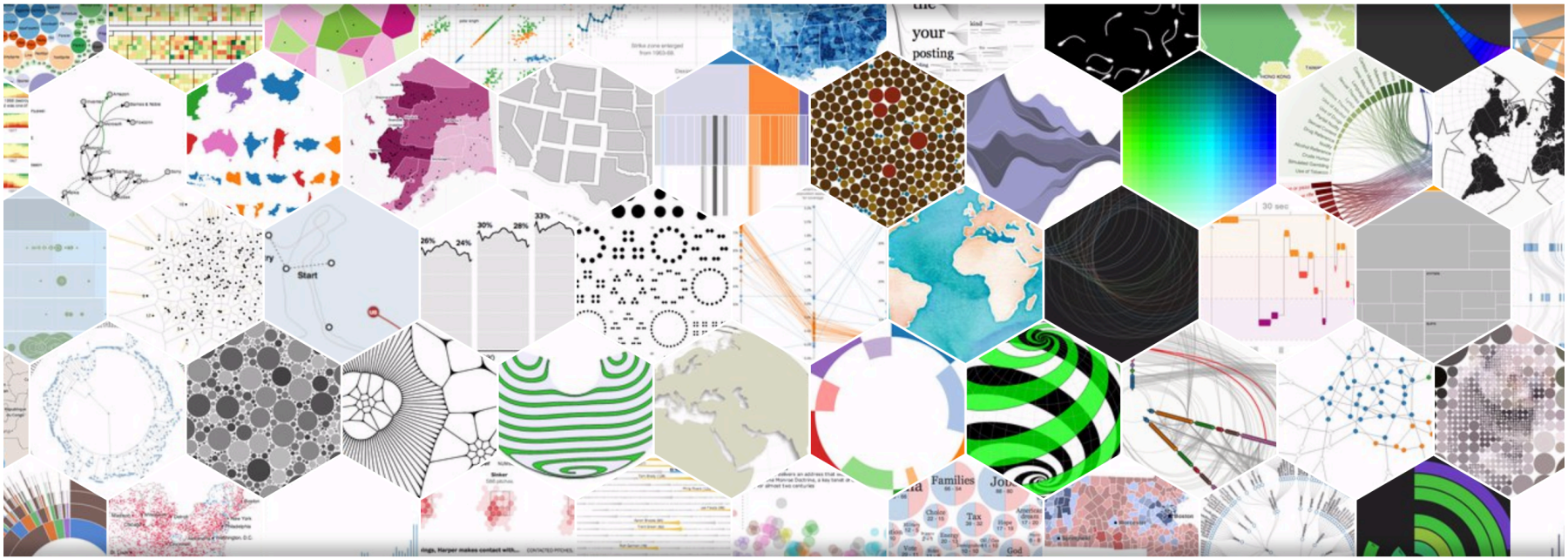
Skill: Comunicação de Resultados

- Outra área interdisciplinar:
  - Visualização: arte e ciência.
  - Design: significado para usuários.
- Ferramentas de análise tem funções para exibição de resultados, visualização, etc.
- Outras ferramentas podem ser parte do seu repertório.



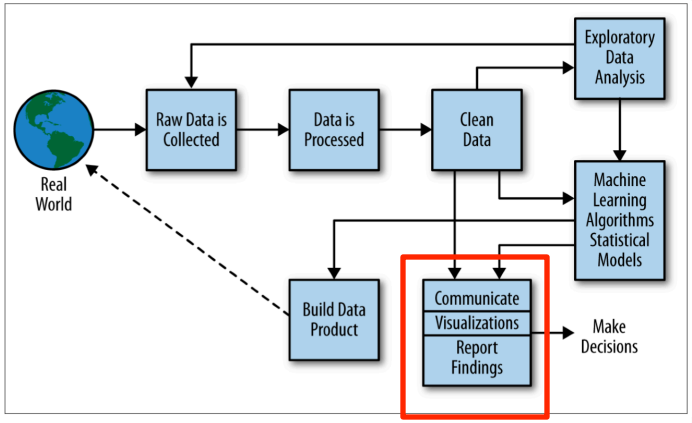
Skill: Comunicação de Resultados

- **D3.js:** ***Data-Driven Documents***
  - Biblioteca em JavaScript para manipulação de DOM (=dados!)



Skill: Comunicação de Resultados

- Aplicações na Web, *Web Services*, etc.
- Notebooks online: **iPython**, **Jupyter** permitem a criação de documentos interativos em várias linguagens de análise.
- *Reproducible Research!*



Jupyter

```
import time
```

```
from numpy import cumprod, linspace, random
```

```
from bokeh.sampledata.stocks import AAPL, FB, GOOG, IBM, MSFT
from bokeh.plotting import figure, output_notebook, show
```

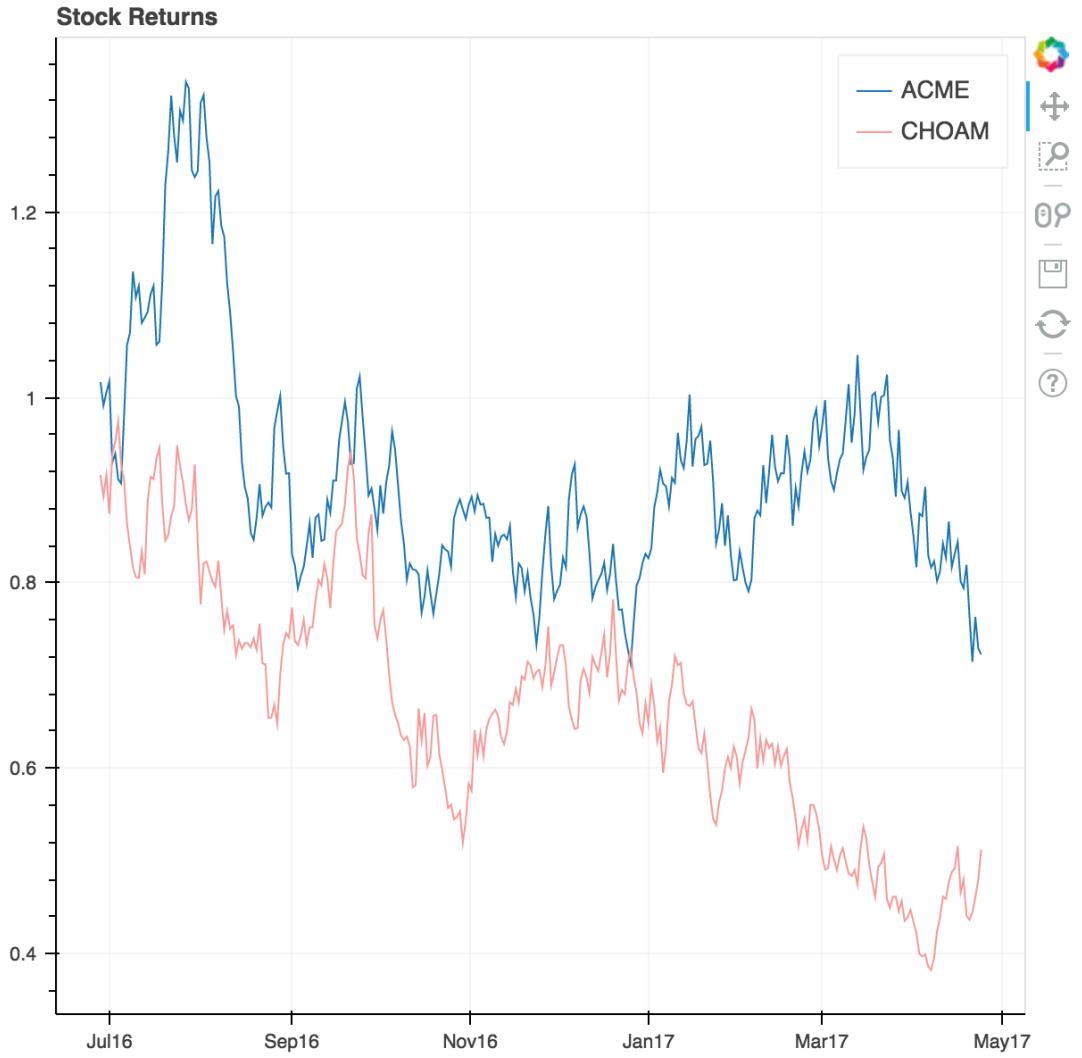
```
num_points = 300

now = time.time()
dt = 24*3600 # days in seconds
dates = linspace(now, now + num_points*dt, num_points) * 1000 # times in ms
acme = cumprod(random.lognormal(0.0, 0.04, size=num_points))
choam = cumprod(random.lognormal(0.0, 0.04, size=num_points))
```

```
output_notebook()
```

BokehJS successfully loaded

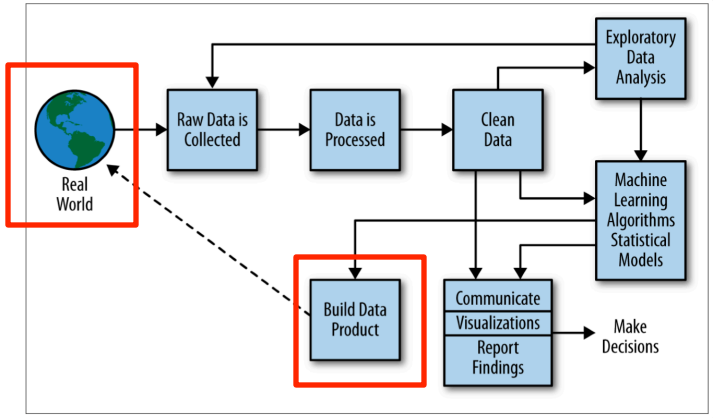
```
p1 = figure(x_axis_type = "datetime")

p1.line(dates, acme, color='#1F78B4', legend='ACME')
p1.line(dates, choam, color='#FB9A99', legend='CHOAM')

p1.title.text = "Stock Returns"
p1.grid.grid_line_alpha=0.3

show(p1)
```

Jupyter


Stock Returns
ACME
CHOAM
1.2
1
0.8
0.6
0.4
Jul16
Sep16
Nov16
Jan17
Mar17
May17

Skill: Entender (melhor) o Problema

- Que dados deveriam existir?
  - Produto de Dados!
- Depois de aplicar estes conhecimentos, processamentos, técnicas, etc., que dados seriam interessantes para:
  - Entender melhor todo o problema?
  - Agregar valor aos existentes?
  - Possibilitar novas aplicações?



Estes devem ser os objetivos principais de um Data Scientist!

Conceitos de *Data Science*

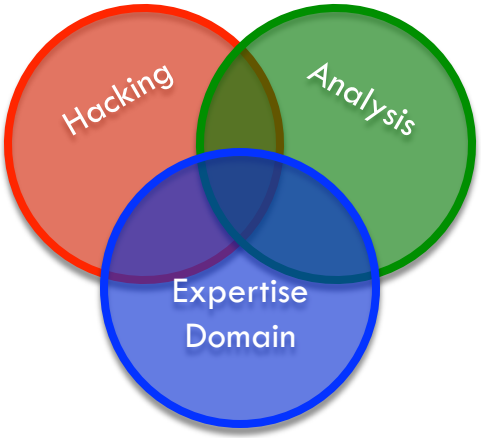
Projetos

LattesLab

- **Dados**: um subconjunto curado, temático de currículo Lattes (armazenados *offline*).
  - Contém informações sobre pesquisadores e alunos, publicações, áreas de atividade, etc.
  - Conexões entre currículos Lattes dependem da precisão dos dados entrados (depende de quem preencheu).
- Como corrigir/complementar/enriquecer estes dados?

LattesLab

- **Expertise**:
  - Conhecimento das necessidades de análises e relatórios baseados no Lattes.
  - Conhecimento da estrutura e problemas com dados organizados no Lattes.
- **Análise**:
  - *Text Mining* e Casamento de Padrões/Grafos.
  - Visualização.
- **Hacking**:
  - Processamento de dados em XML.
  - Correlação com outras fontes de dados (texto): *data munging*.
  - *Text mining* / Casamento de padrões.
  - Ferramentas de visualização (D3).

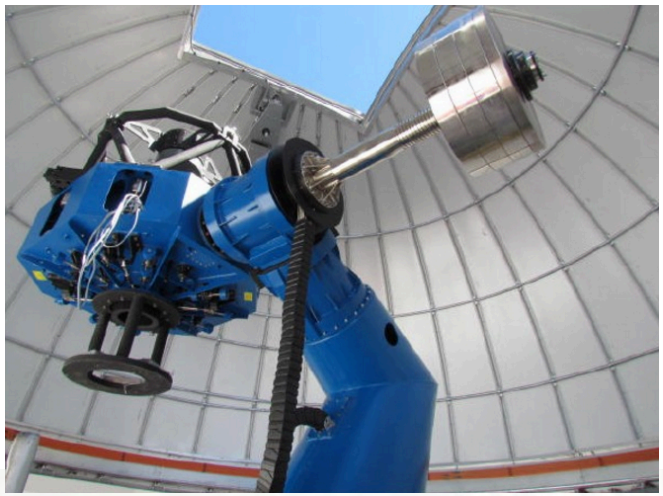


LattesLab

- **Produtos de Dados**:
  - Dicionários de similaridade de nomes e conceitos.
  - Casamento de publicações.
  - Bases de anotações.
  - Bases de grafos.

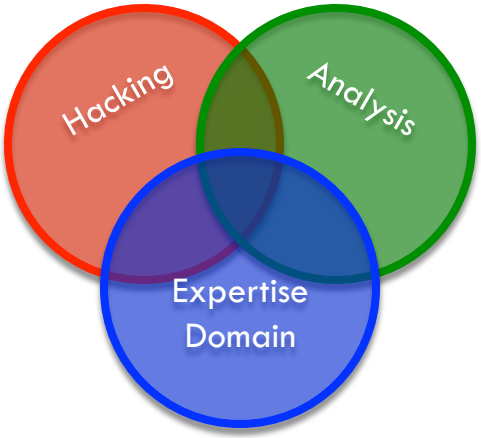
S-Plus Virtual Observatory

- **Dados**: imagens, espectros e parâmetros coletados pelo *Southern Photometric Local Universe Survey* (S-PLUS), organizados em bancos de dados.
- Devem ser criados *data releases* anuais.
- Serão usados pela comunidade de astronomia.



S-Plus Virtual Observatory

- **Expertise**:
  - Conhecimentos básicos de astronomia, complementados pela equipe.
  - Conhecimentos da organização de dados e protocolos dos observatórios virtuais astronômicos (VOs).
- **Análise**:
  - Somente para subprojetos.
- **Hacking**:
  - Processamento de dados no formato FITS.
  - Criação de *web services* para VOs.

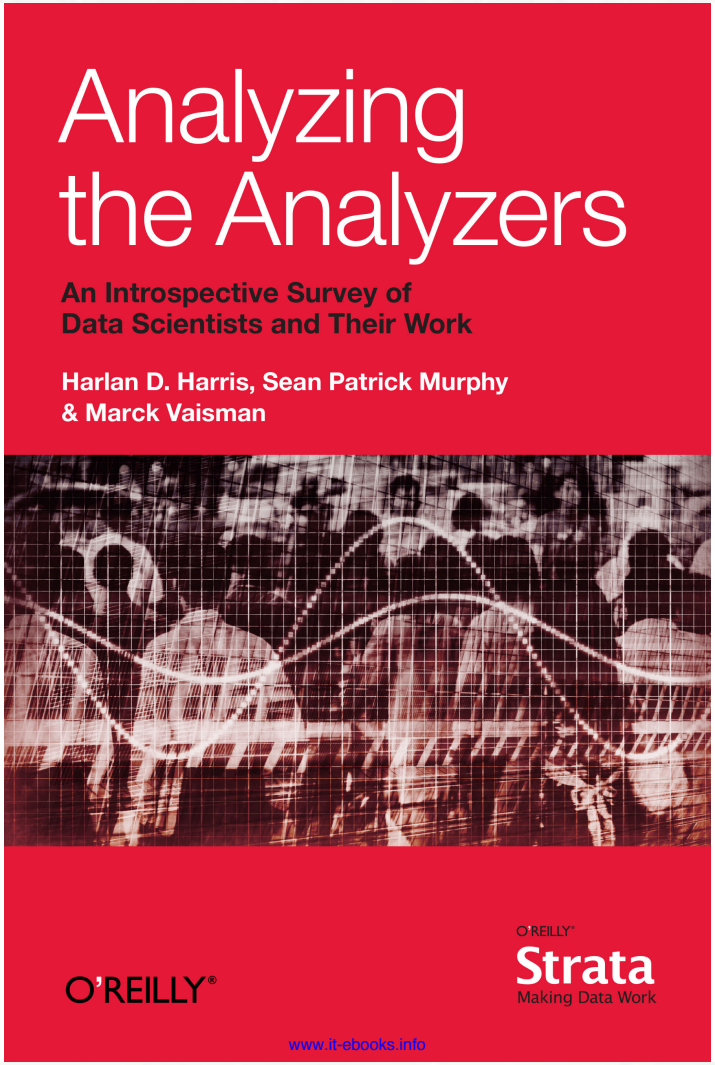


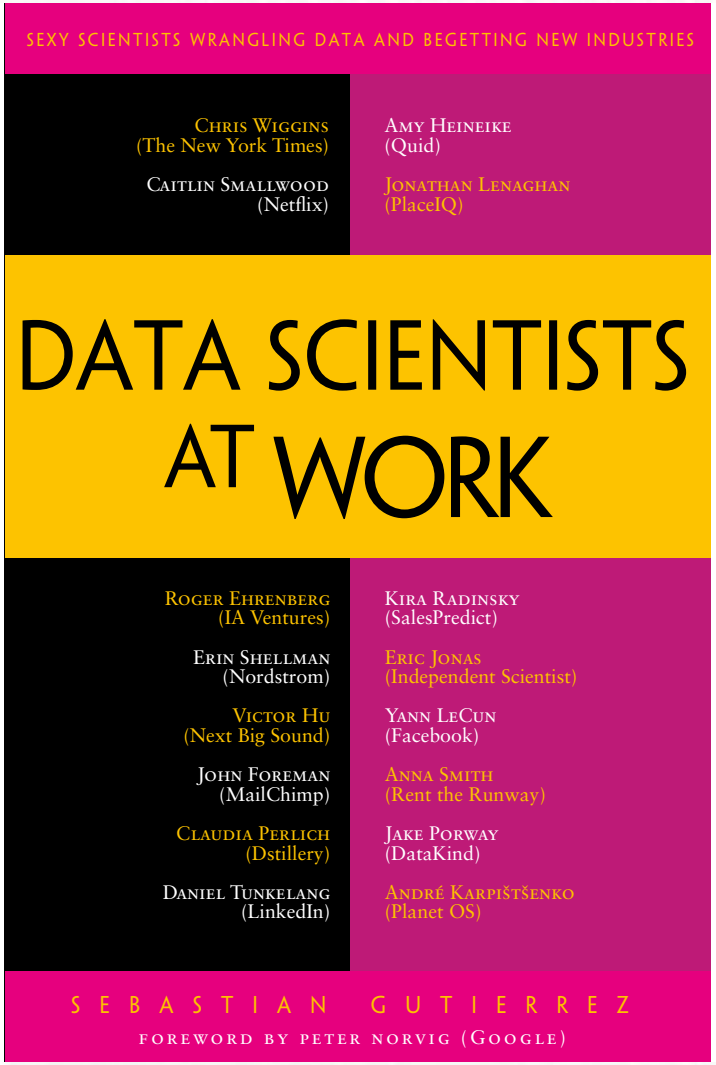
S-Plus Virtual Observatory

- **Produtos de Dados**:
  - Catálogos de objetos e sistemas de busca nos mesmos.
  - Metadados (proveniência).

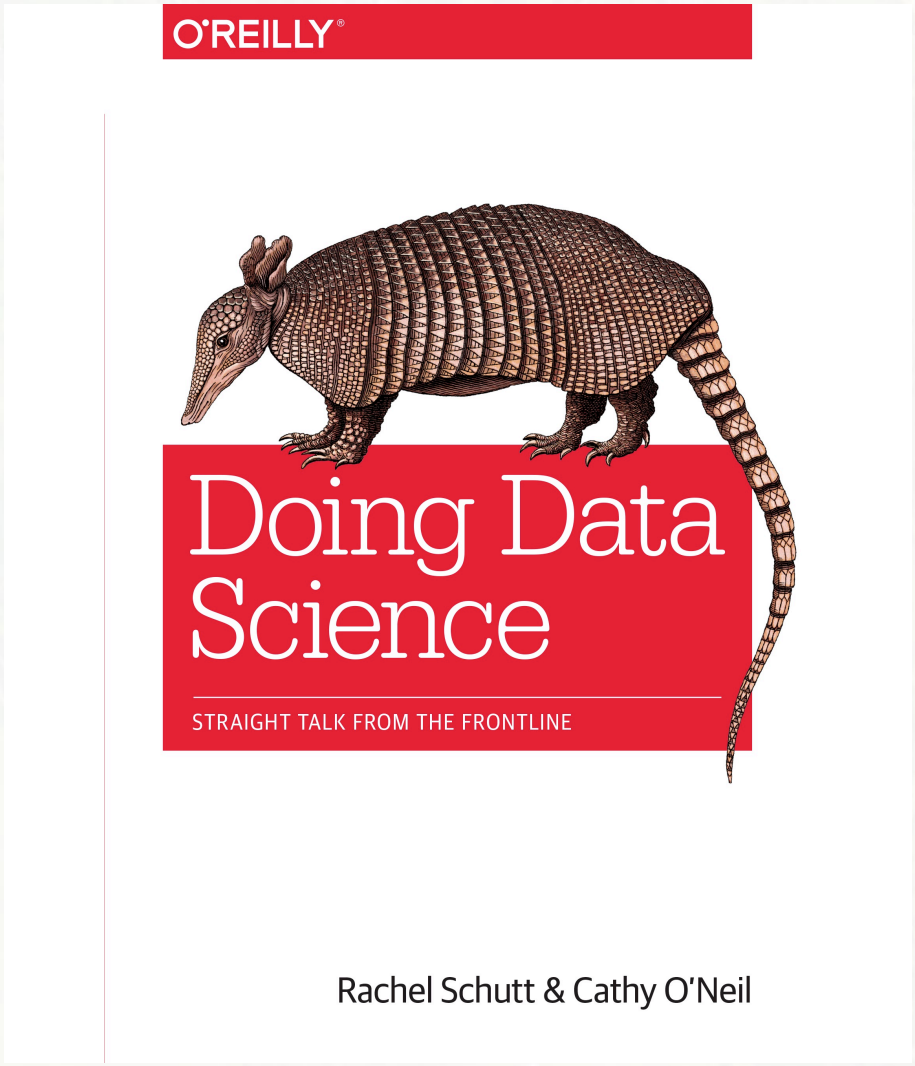
Referências

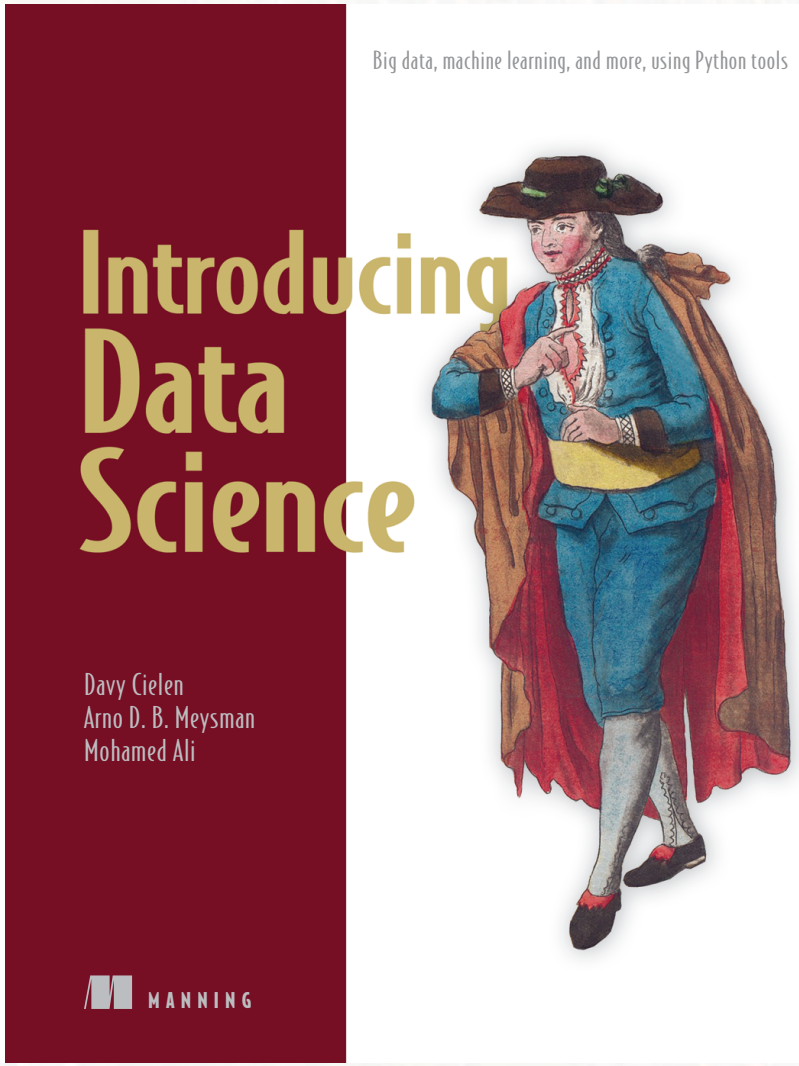
Referências



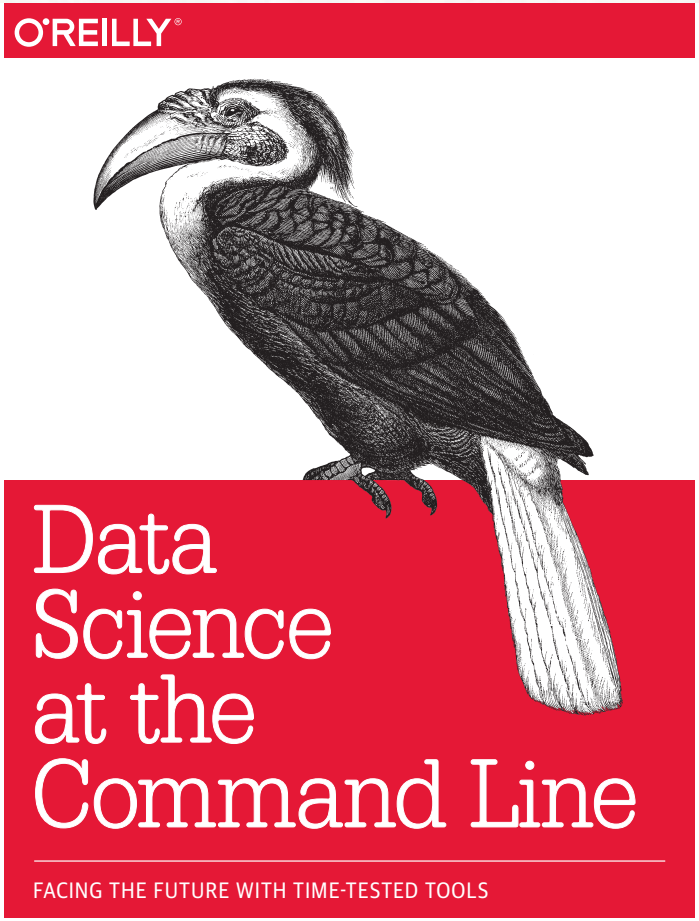


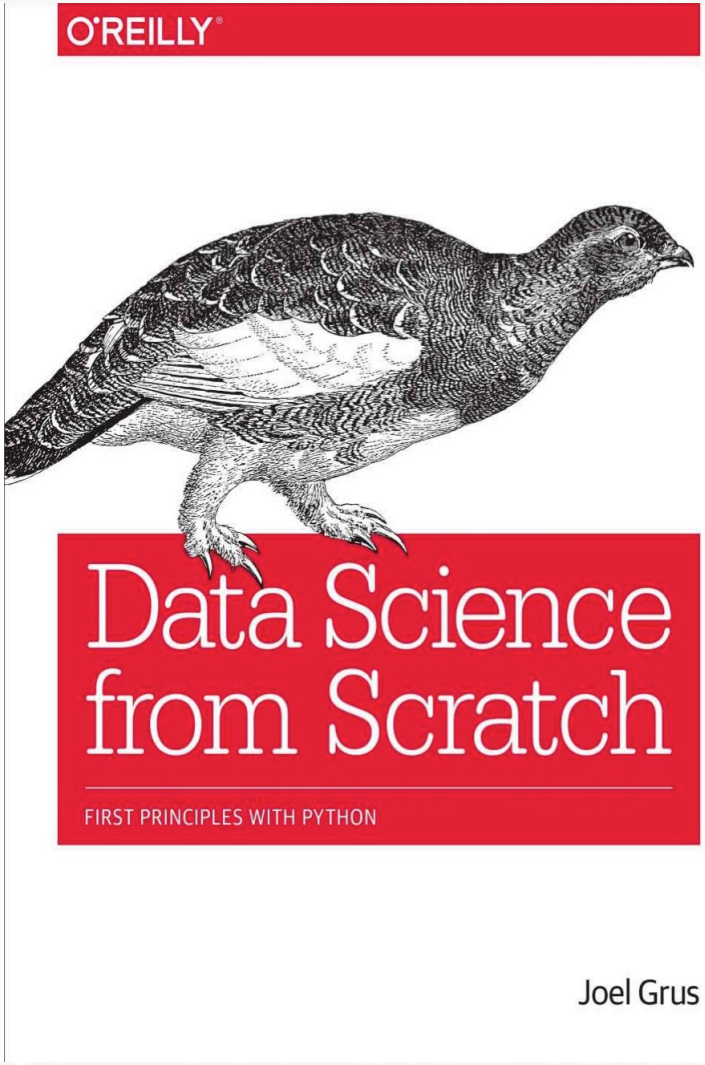
Referências



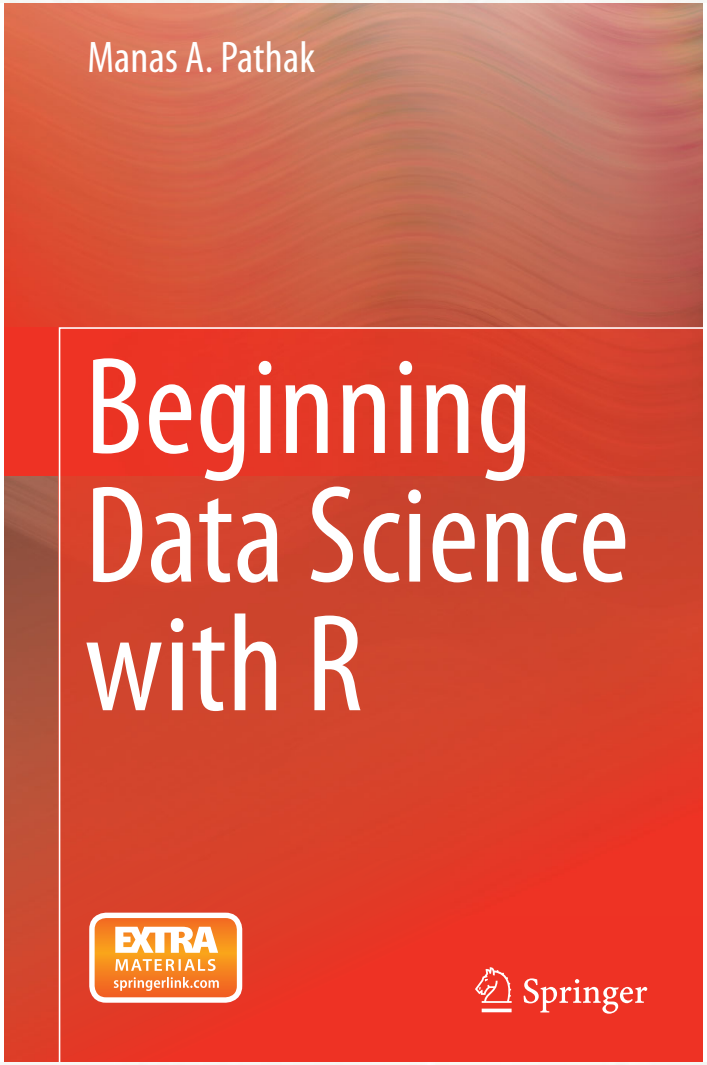


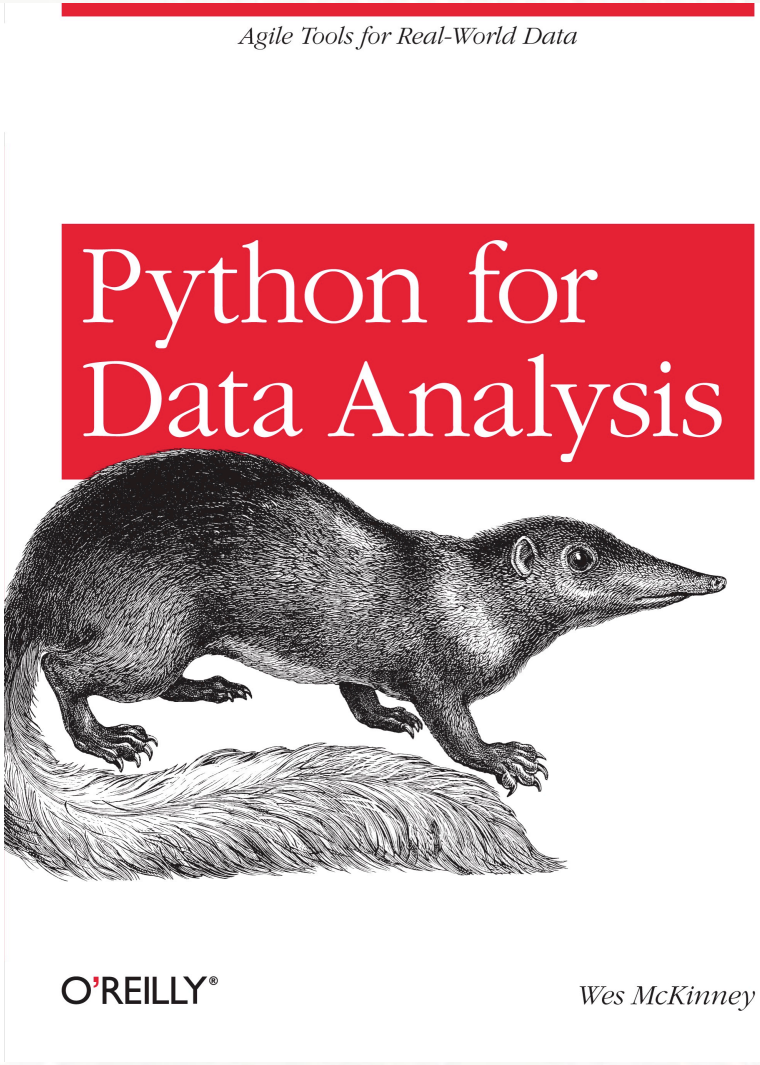
Referências





Referências

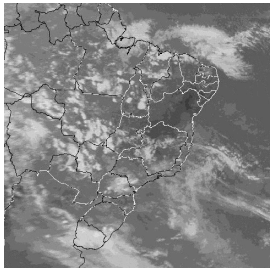




Em Breve!

- Curso **Introdução a *Data Science*** no Programa de Pós-Graduação em Computação Aplicada (a partir de 2017).
  - Contexto mais prático e científico: *Data Science Process*
  - Divisão de conteúdo:
    - IDS: Organização de Dados, EDA, Visualização.
    - PADM: Algoritmos de DM e Aplicações.
  - Ambas focadas em projetos.
- Este material em http://www.lac.inpe.br/~rafael.santos

Doing Data Science, Rachel Shutt and Cathy O'Neil, OReilly, 2014



CONCEITOS DE *DATA SCIENCE*

Rafael Santos – rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/