



Fundamentos de Data Science, Data Mining e Análise Preditiva

Especialização em Ciência de Dados com Big Data, Bile Data Analytics

Prof. Dr. Carlos Barros

Prof. Dr. Carlos Barros

- Formação
 - Doutor em Informática Aplicada- UNIFOR
 - Optimum Path Snakes: Novo Método de Contornos Ativos Adaptativo e Não-paramétrico, Ano de obtenção: 2016.
 - Mestre em Engenharia de Teleinformatica– UFC
 - Analise comparativa de Técnicas de Detecção de Movimento e Rastreamento de Objetos em Video Digital
 - Especialista em Tecnologia da Informação - UFC
 - Dispositivos Móveis usando Visão Artificial com processamento Nativo.
 - Bacharel em Sistema de Informação - UNERSA
 - Negócios sud - uma plataforma para gerenciamento e divulgação de anúncios online
 - Tecnologo em Mecatronica – IFCE
 - Biblioteca de lógica fuzzy para implementação de controladores.

Prof. Dr. Carlos Barros

- Experiencia

- CEO –Sumplus Projetos e Inovação
- Coordenador dos projetos
 - Construção de Jogo Educacional Utilizando Visão Computacional
 - Sistema de Percepção de Risco e Reconhecimento de Comportamentos de Motoristas
 - Raw2data - Sistema de coleta e pré-processamento para mineração de dados
 - Easy carbo diabetes
- Integrante dos seguintes projetos
 - Projeto Portátil
 - Dsaem Desenvolvimento De Softwares Para Ambientes Embarcado
 - Samsung Movement Tracking And Research Group – Smtrg
 - Controle E Automação De Gaseificadores Em Leito Fluidizado De Casca De Castanha De Caju (Programa De Apoio A Equipes De Pesquisa
 - Sistemas Inteligentes Aplicados Na Área Biomédica
 - Desenvolvimento De Sistema Web 3d Para Auxílio Ao Diagnóstico Médico De Doenças Usando Conceitos De Iot E Deep Learning
 - Desenvolvimento De Novas Técnicas E Aplicações De Processamento Digital De Imagens Para Aplicações Em Imagens Médicas
 - Desenvolvimento De Softwares Para Dispositivos Móveis Para Aplicações De Telemetria
 - Obsekium - Identificação E Rastreamento De Objetos Codificados Por Visão Computacional Em Sistemas Mobile
 - Composição Corporal Mobile (Visão Computacional Em Dispositivos Móveis)
 - Locktec - Plataforma De Controle Inteligente De Condomínios Com Portaria Virtual

Prof. Dr. Carlos Barros

- Atuação
 - Magistério Público Federal - Unilab
 - Professor de pós-graduação nas áreas de Data Science e Aprendizado de Máquina
 - Pesquisador no LAPISCO com ênfase em IIoT, Visão Computacional e Aprendizado de Máquina
 - Possui seis patentes em Sistemas Embarcados (INPI)
 - Empreendedor em IIoT com ênfase em:
 - Sistemas
 - Computação Embarcada

Prof. Dr. Carlos Barros

- Publicações
 - <http://lattes.cnpq.br/3311439131550698>

Metodologia

- Aula expositivas com discursos
- Praticas em laboratorio
- Leituras
- Tarefas Individuais
- Avaliacao (em dupla)



"Fundamentos de Data Science, Data Mining e Análise Preditiva"

Carga Horária: 24 horas

Introdução a Data Mining e Ciência dos Dados. Obtendo informações a partir dos dados. Principais Paradigmas e Modelos para mineração de dados. Dados incertos, com ruídos/outliers e confiança nos dados. Análise de dados exploratória. Introdução ao uso de modelos de predição. Escolha de modelos para mineração de dados. Redução de dimensionalidade e engenharia de dados. Minerando dados complexos.

Trabalhando e limpando os Dados.

Encontros

- Dia 1 – 15/03/2019 - sexta
 - Apresentação
 - Introdução a Data Mining e Ciência dos Dados.
 - Obtendo informações a partir dos dados
 - Pratica 1 – tarefa avaliativa – 20:35H
- Dia 2 – 16/03/2019 - sábado
 - Dados incertos, com ruídos/outliers e confiança nos dados
 - Principais Paradigmas e Modelos para mineração de dados
 - Escolha de modelos para mineração de dados.
 - Introdução ao uso de modelos de predição.
 - Pratica 2 – tarefa avaliativa – 13:00H

Encontros

- Dia 3 – 29/03/2019 - sexta
 - Análise de dados exploratória
 - Trabalhando e limpando os Dados.
 - Pratica 3 – tarefa avaliativa – 20:35H
- Dia 4 – 30/03/2019 - sábado
 - Minerando dados complexos.
 - Redução de dimensionalidade e engenharia de dados.
 - Avaliação – 13:00H



Repositório

<https://github.com/acsbarros/fdsdm>

Introdução a Data Mining e Ciência dos Dados



Princípios data mining

A aplicação de métodos de aprendizado de máquina em grandes bases de dados é chamada de Mineração de Dados (data mining)

O que é Mineração de Dados?

Processo de identificação de padrões válidos, novos, potencialmente úteis e comprehensíveis embutidos nos dados (Fayyad et al, 1996) Encontra informações úteis embutidas em GRANDES volumes de dados

Análise de dados e o uso de técnicas de software para encontrar padrões e regularidades em conjuntos de dados

O computador é responsável por encontrar os padrões por meio da identificação de regras e características implícitas nos dados

É possível “achar ouro” em lugares inesperados na medida em que o software de mineração de dados extrai padrões antes não discerníveis ou tão obvios que ninguém tinha notado antes

Analogia com a mineração

Grandes volumes de dados são “peneirados” na tentativa de se encontrar alguma informação de valor

Exemplos

Qual produto de alta lucratividade venderia mais com a promoção de um item de baixa lucratividade, analisando os dados dos últimos dez anos?

Quais são os clientes potenciais para praticar fraudes?

Quais clientes gostariam de comprar o novo produto X?

Que genes são determinantes para o diagnóstico de um determinado tipo de doença?

Descoberta de Conhecimento

Descoberta de conhecimento ou *Knowledge Discovery in Database (KDD)* é um outro termo para o processo de Mineração de Dados

Alguns autores consideram os termos KDD Mineração de Dados referentes a processos distintos

Mineração de Dados seria uma etapa do processo de KDD

Mineração de Dados - uma área multidisciplinar

Banco de Dados

Estatística

Computação de Alto-desempenho

Aprendizado de Máquina

Visualização

Matemática

Objetivos da Mineração de Dados

Atividades Preditivas: Classificação e Regressão

Sistemas de mineração de Dados aprendem a partir de exemplos como particionar ou classificar os dados (p. ex., gerando regras de classificação)

Exemplo - base de dados de clientes de um banco

Pergunta: Um novo cliente solicitando um empréstimo é um bom ou mau investimento?

Regra típica formulada:

Se STATUS = cassado e RENDA > 2000 e PROPRIETARIO-IMÓVEL = sim
então TIPO-DE-INVESTIMENTO = bom

Objetivos da Mineração de Dados

Atividades Descritivas: Associação, Clustering, Sumarização

Regras de Associação

Regras que associam um atributo de uma relação a outro
Abordagens orientadas a conjuntos são os meios mais eficientes para a descobertas de tais regras

Exemplo - base de dados de um supermercado

72% de todos os registros que contêm itens A e B também contêm item C

A porcentagem específica de ocorrências é o fator de confiança da regra

Estágios do Processo de Mineração de Dados

Identificação do Problema

Quais são as principais metas do processo?

Quais critérios de desempenho são importantes?

O conhecimento extraído deve ser compreensível a seres humanos ou um modelo tipo caixa-preta é apropriado?

Qual a deve ser a relação entre simplicidade e precisão do conhecimento extraído?

Pré-processamento

Extração e Integração

Limpeza

Transformação

Seleção e Redução

Criação de um modelo - Aprendizado de Máquina

Escolha da tarefa - classificação, regressão, associação, clustering, ...

Escolha do(s) algoritmo(s)

Aplicação do(s) algoritmo(s)

Teste do modelo

Interpretação e avaliação

Técnicas de Aprendizado de Máquina

k-NN

Naive Bayesian Learning

Árvores de Decisão

Regras

Redes Neurais Artificiais

Support Vector Machines

Ensembles

Regras de Associação

k-means

Métodos de agrupamento hierárquico

Aplicações de Mineração de Dados

Atribuição de crédito

Predição no mercado financeiro

Diagnóstico de falhas em linhas de produção

Descobertas médicas

Detecção de fraudes

Análise de tendências de compra

Marketing direcionado

....

Ciência dos Dados

Data Science

Data Science é mais um termo usado para descrever o processo de transformação de dados em conhecimento. É diferente 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.

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.

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.
- “ *Data Science* possibilita a criação de produtos de dados.

Envolve programação, mas...

... não somente programação e novas tecnologias.

“**The keyword 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?”

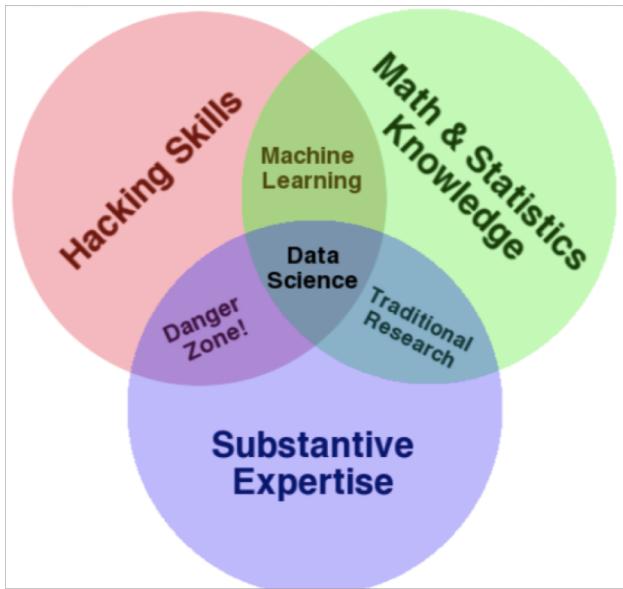
“ Menos ênfase em tamanho e tecnologia, mais em aplicação de tecnologias para obter respostas sobre os dados.

Envolve estatística, mas...

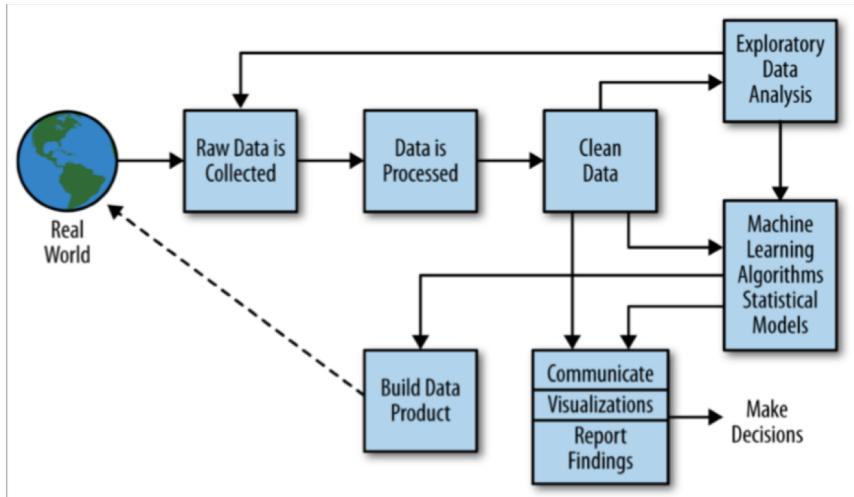
...não puramente estatística tradicional.

- “ Pode ser necessário escalar 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!

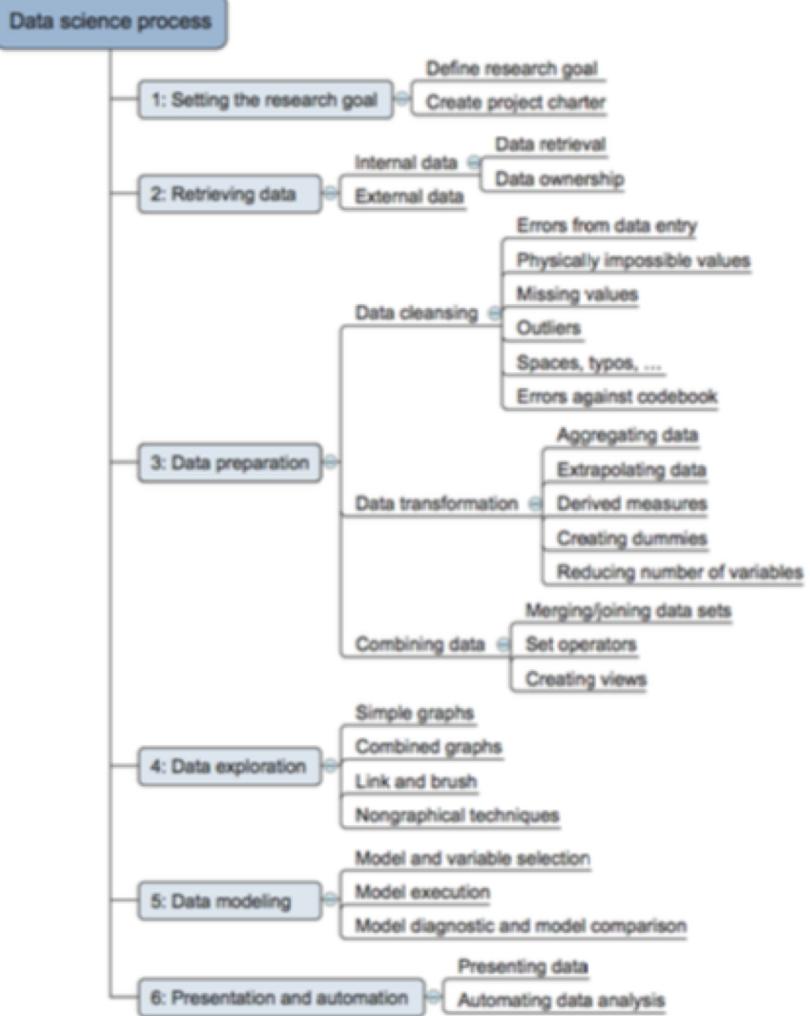
É tudo isto
(e ainda
mais?)



É um
processo (?)



É um
processo
(?)



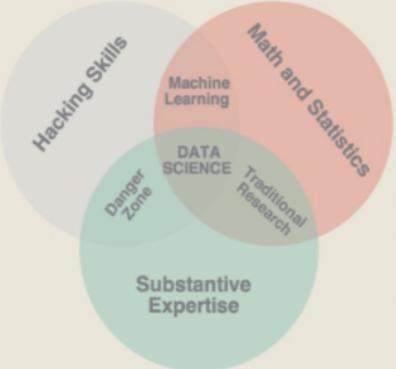
Become a



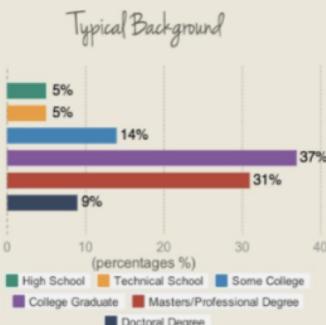
Data Scientist

in 8 easy steps

What's a data scientist?



A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g., R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics and econometrics, data warehousing and big data systems, marketing channel insights in Paid Search, SEO, Social, CRM and brand.

1 Get good at stats, math and machine learning

Math



> Math Track of Khan Academy

> Linear Algebra by MIT OpenCourseware



Stats



> Intro to Statistics by Udacity

> OpenIntro Statistics



ML



> Machine Learning by Andrew NG (Stanford Online)
> Practical Machine Learning by John Hopkins (Coursera)

2 Learn to code



Computer Science Fundamentals

> CS50x on edX



Grasp end-to-end development
The things you build will be integrated
into other systems



Choose a first language

> Open Source: R, Python, etc.
> Commercial: SAS, SPSS, etc.

Learn Interactively

> R: DataCamp, tryR
> Python: Codecademy, Google Class



3 Understand databases

As a data scientist student, you will often work with data in text files. However, once you enter the industry, a database is almost always used to store data. It's going to be stored in MySQL, Postgres, MongoDB, Cassandra, etc.



4 Master data munging, visualization and reporting

Data cleaning and munging



WHAT

Data munging is the process of converting one "raw" form into another format for more convenient consumption



TOOLS

> Getting and Cleaning data by John Hopkins (Coursera)

DataWrangler alpha



data.table
dplyr

Data visualization



WHAT

Data visualization involves the creation and study of the visual representation of data.



TOOLS

ggvis



vega

5 Level up with Big Data

When you start operating with data at the scale of the web, the fundamental approach and process of analysis must change. Most data scientists are working on problems that can't be run on single machines. They have large data sets that require distributed processing.

Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.



MapReduce



MapReduce is this programming paradigm that allows for massive scalability across the servers in a Hadoop cluster.

Apache Spark is Hadoop's speedy Swiss Army knife. It is a fast-running data analysis system that provides real-time data processing functions to Hadoop.



6 Get experience, practice and meet fellow data scientists

Practice makes perfect ...



Join in
competitions



Meet fellow data
scientists



Have a pet
project



Develop your
intuition

7 Internship, bootcamp or get a job

The best way to find out whether you are a true data scientist or not is to take the bull by the horns and to enter the real-life jungle of data-analysis and science with your freshly acquired skill set.

Internship
★ ★ ★
BEGINNER

Bootcamp
★ ★ ★
INTERMEDIATE

Job
★★★★★
ADVANCED



8 Follow and engage with the community

Sites to follow

- > DataTau
- > Kdnuggets
- > fivethirtyeight
- > datascience101
- > r-bloggers

People to follow

- > Hilary Mason
- > David Smith
- > Nate Silver
- > dj patil

Need Data?



Desafios – variedade de dados



Arquivos
(XML, CSV, Excel, JSON, ...)



Banco de Dados
(MySQL, Oracle, ...)



API



Sites



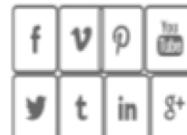
Textos e relatórios



Mapas



Imagen e video



Redes sociais

Desafios – variedade de dados

Obtendo informações a partir dos dados.

Amostragem

Medidas de Centralidade e Variabilidade

Probabilidade

Distribuição Binomial

Distribuição Normal

Estatística não paramétrica

Intervalos de Confiança

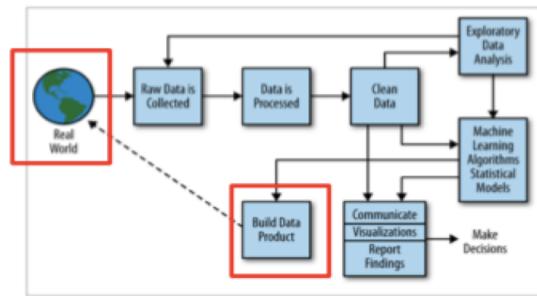
Testes de Hipótese

Distribuição t de Student

- 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.

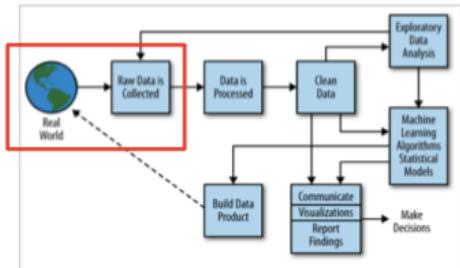
Entender o Problema

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



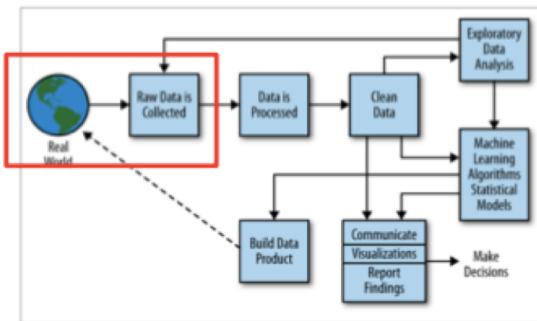
Skill: Achar Dados

- 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?



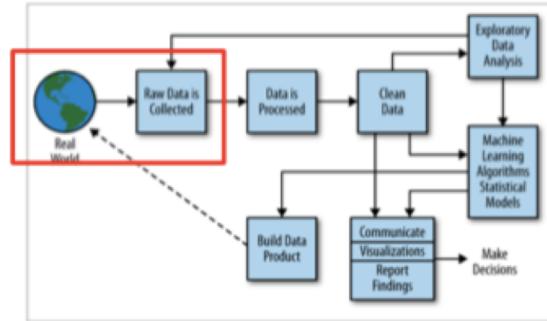
Skill: Entender a Organização dos Dados (1)

- **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?



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...

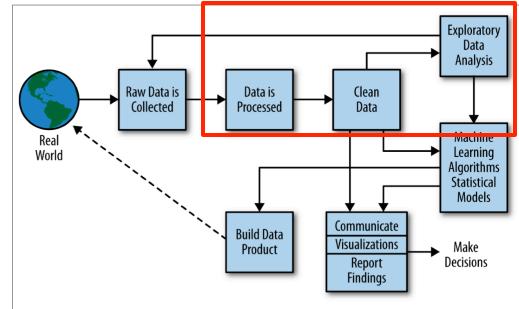
Skill: Análise (Hacking)

□ 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): 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]

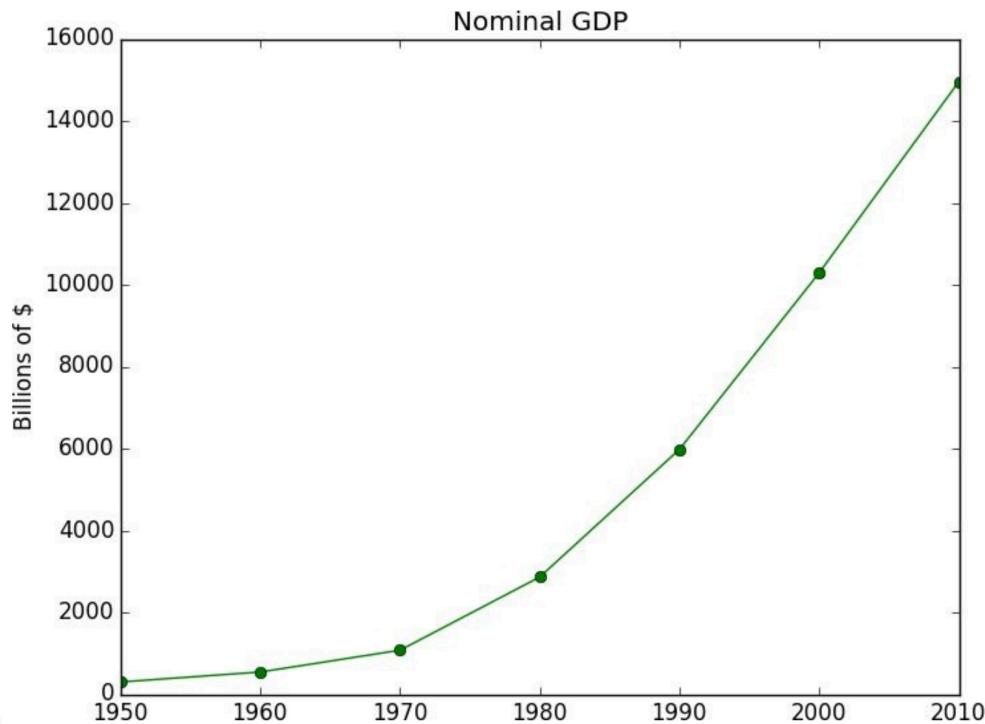
# create a line chart, years on x-axis, gdp on y-axis
plt.plot(years, gdp, color='green', marker='o', linestyle='solid')

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

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

Skill: Análise (Hacking): Python

□ Exemplo básico

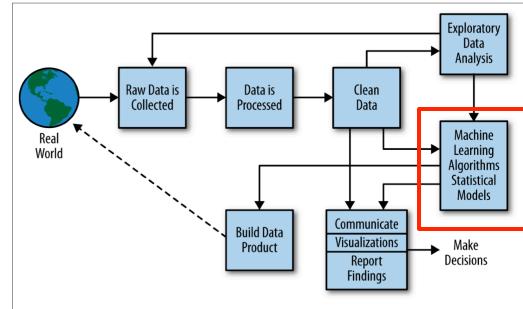


Skill: Análise (Hacking): Python

- 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 reproduzível.

Skill: Machine Learning, Models

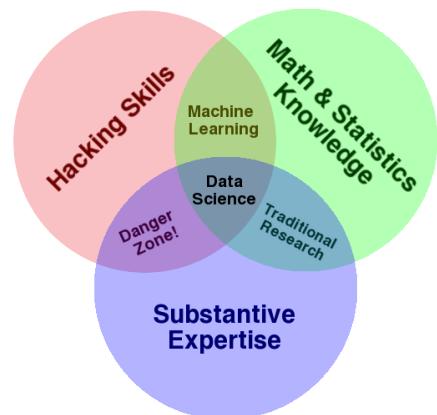
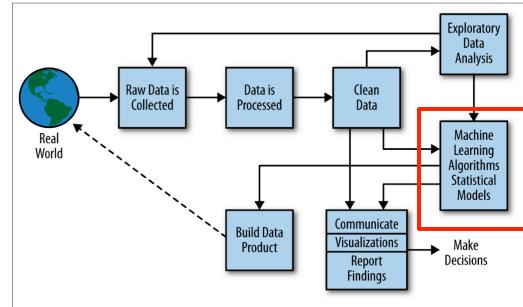
- O que posso aprender a partir de meus dados?
- *Exploratory Data Analysis* deve servir para dar indícios da natureza dos dados e de que conhecimento podemos extrair deles.
- *Machine Learning, Data Mining, etc.* podem servir para criar modelos que descrevam os dados.



Skill: Machine Learning, Models

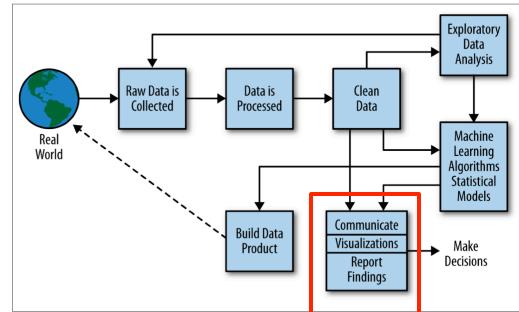
□ Cuidados:

- 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!



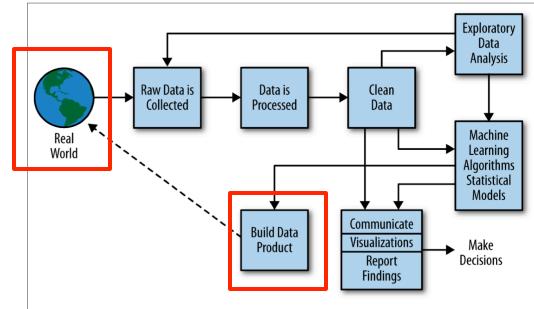
Skill: Comunicação de Resultados

- 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: 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!