

# Data Science For Dummies

—

Eduardo Bonet

# Bio

Bonet

Engenharia de Controle e Automação

"Mestrando" em Ciência da Computação

Full Stack, Mobile, Data Science

[Github](#) | [LinkedIn](#)



# O que essa palestra NÃO vai ser

- Não será uma palestra técnica
- Não será uma palestra motivacional
- Você não vai sair daqui um Cientista de Dados

# O que essa palestra vai ser

- O que um Cientista de Dados faz
- As habilidades esperadas de um Cientista de Dados.
- Onde procurar recursos para ir em busca dessas habilidades.

Os slides estarão disponíveis em [github.com/ebonet/presentations/dsfordummies](https://github.com/ebonet/presentations/dsfordummies)

**O que é um Cientista de Dados?**

**Ninguém Sabe!**



**Josh Wills**

@josh\_wills



 Seguir

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

### **Cientista de Dados:**

Uma pessoa que é melhor em estatística do que qualquer Engenheiro de Software e melhor em engenharia de software que qualquer estatístico.



**Big Data Borat**

@BigDataBorat

 Follow

Data Science is statistics on a Mac.

10:32 AM - 27 Aug 2013



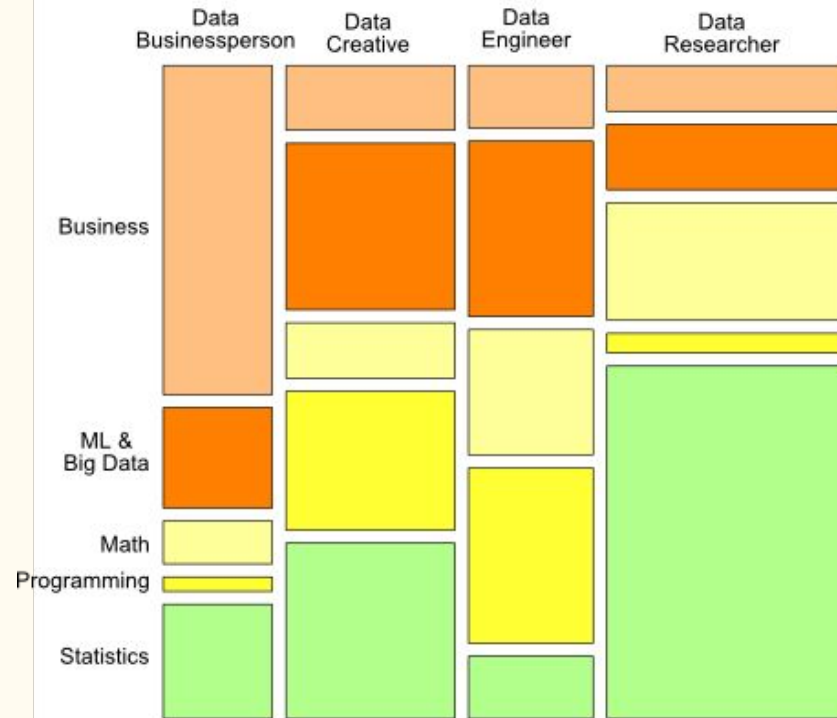
613



274

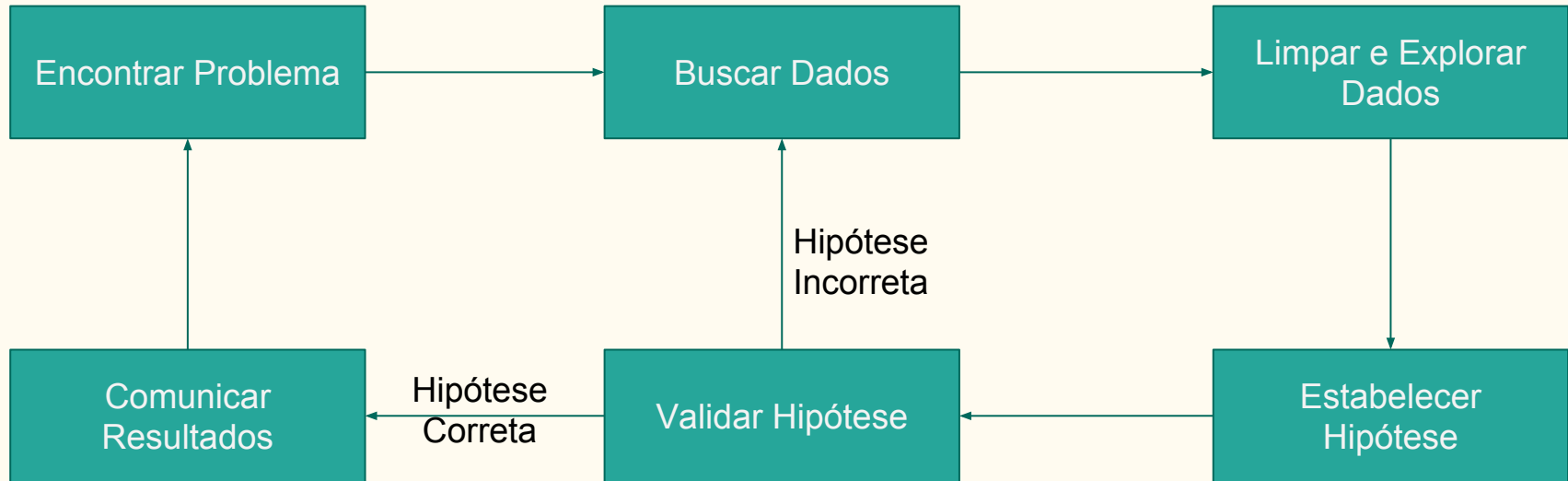
**Cientista de Dados:**  
Um estatístico que usa Mac





Autor: Halan Harris

# Dia a dia de Ciência de Dados, exemplificado



Dia a dia exemplificado

Encontrar um problema

**É possível estimar um aluguel em  
Florianópolis?**

## Dia a dia exemplificado

# Coletar Dados

Viva Real: site de imóveis para aluguel. <http://api.vivareal.com:80/api/1.0/api-docs>

Retorna informações como (Exemplo).

- Bairro
- Latitude, Longitude
- Preço
- Número de banheiro, quartos, garagens, etc...
- Preço de condomínio



Dia a dia exemplificado

# Limpar dados

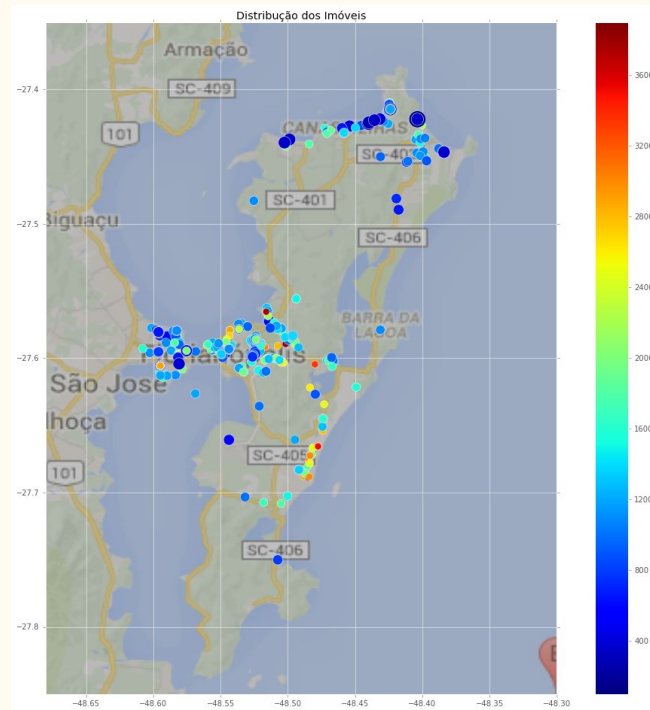
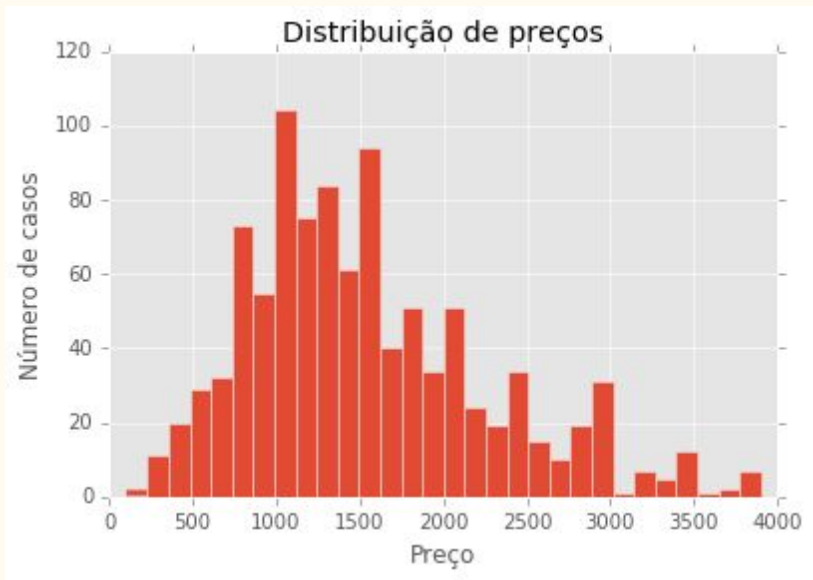
- Corrigir localidades com coordenadas incorretas
- Remover entradas com valores estranhos

	latitude	longitude	garages	area
<b>count</b>	2894.000000	2894.000000	2894.000000	1503.000000
<b>mean</b>	-28.333275	-47.552591	1.080166	97.284358
<b>std</b>	4.398606	4.723807	3.810740	560.715869
<b>min</b>	-48.406512	-50.218856	0.000000	1.000000
<b>25%</b>	-27.593236	-48.525292	1.000000	55.000000
<b>50%</b>	-27.437144	-48.458217	1.000000	74.000000
<b>75%</b>	-27.422163	-48.422870	1.000000	100.000000
<b>max</b>	0.000000	0.000000	203.000000	21723.000000

# Dia a dia exemplificado

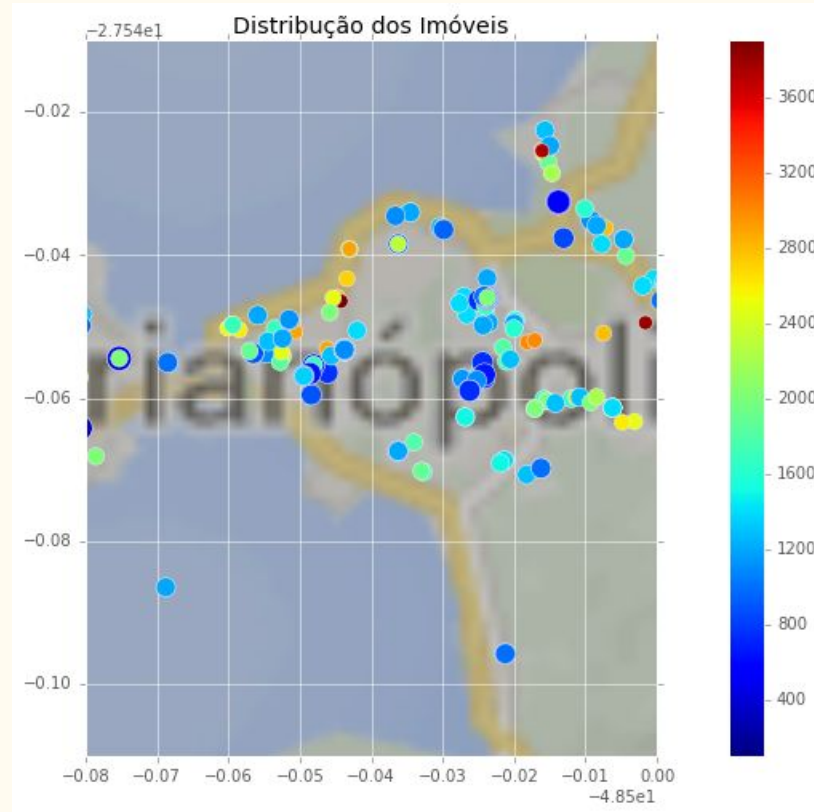
# Explorar Dados

- Verificar distribuições de preços
- Estudar possíveis correlações



Dia a dia exemplificado

# Explorar Dados



Dia a dia exemplificado

# Elaborar Hipótese ou Proposta de solução

O preço do aluguel pode ser calculado com um modelo linear em cima das seguintes variáveis:

Área

Número de quartos

Número de banheiros

Latitude

Longitude

Número de garagens



Dia a dia exemplificado

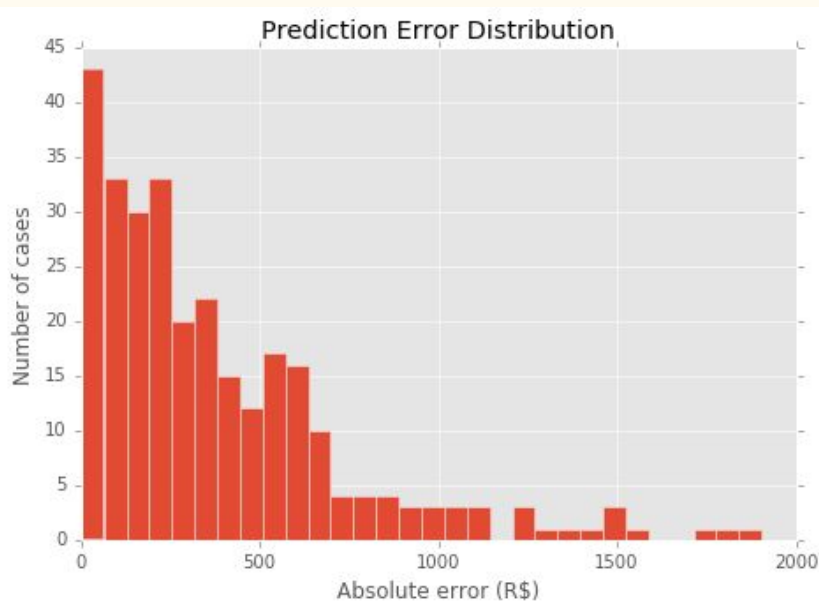
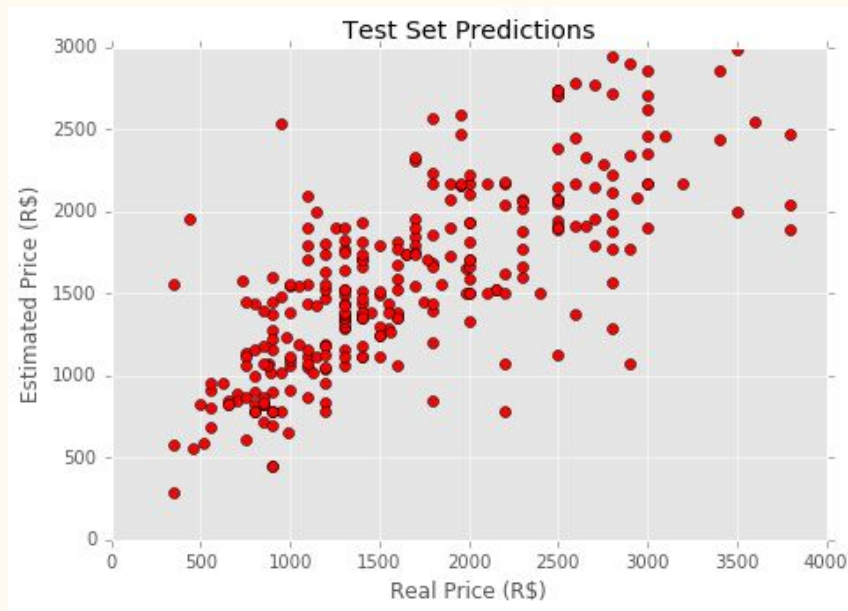
# Testar Hipótese (e falhar miseravelmente)

<b>Dep. Variable:</b>	rentPrice	<b>R-squared:</b>	0.519
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.516
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	152.3
<b>Date:</b>	Wed, 13 Apr 2016	<b>Prob (F-statistic):</b>	1.38e-109
<b>Time:</b>	12:48:49	<b>Log-Likelihood:</b>	-5411.5
<b>No. Observations:</b>	711	<b>AIC:</b>	1.084e+04
<b>Df Residuals:</b>	705	<b>BIC:</b>	1.086e+04
<b>Df Model:</b>	5		
<b>Covariance Type:</b>	nonrobust		

$R^2$  deveria ser perto de 1

Dia a dia exemplificado

# Comunicar Resultados



# Encontrar nova hipótese

- Usar bairro em vez de latitude / longitude
- Cruzar entradas com banco de dados de crimes
- Banco de dados de tráfego
- Usar métodos estatísticos e algébricos para descobrir as variáveis que mais influenciam no preço

Implementar, testar ... e continuar falhando miseravelmente

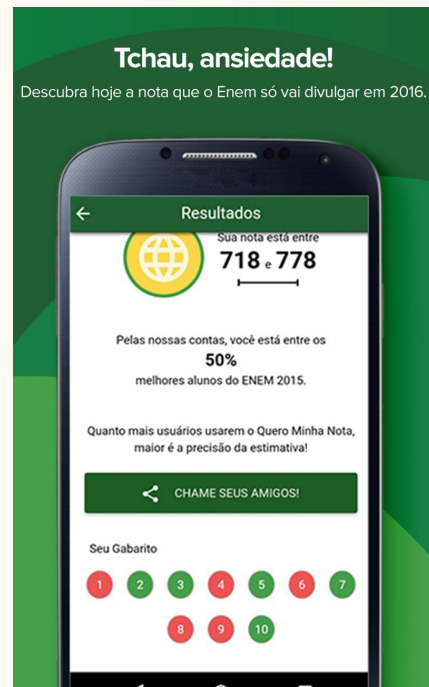


# quero minha **nota**

Problema 1: Inep solta os resultados em 3 meses para os alunos, em 1 ano para download

Problema 2: Como calcular a nota do ENEM (não é trivial)

Solução: Coletar gabaritos de alunos, estimar notas usando a mesma metodologia do Inep, atualizar notas conforme mais alunos entrarem



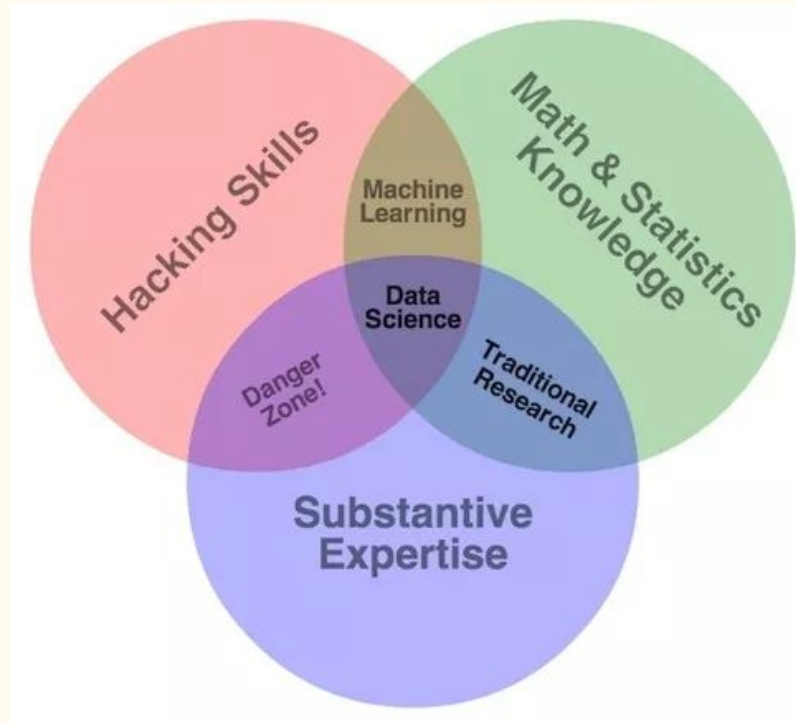
# Alguns aspectos

- Um cientista de dados é um generalista, não um especialista
- CIÊNCIA de dados = Método Científico
- Um bom cientista é quase um unicórnio: precisa ter perfil acadêmico e de mercado, ter noção de visualização, saber garimpar dados, como armazenar esses dados, achar maneira de processar tudo isso, gerar insights e trazer esses insights para produção
- É possível encontrar vagas com graduação, mas muitos pedem pelo menos mestrado

# Quais são as habilidades de um Cientista de Dados?

E onde consigo encontrá-las?

# Diagrama de Venn de Drew Conway



# Habilidades - Programação



# Programação Python

- Versátil e fácil de aprender
- Serve tanto para fazer análise quanto para colocar em produção



[Especialização Python](#) (University of Michigan) | [Python](#) (CodeAcademy) | [Python Class](#) (Google)  
[Codewars](#), [Codility](#), [Hackerrank](#)

# Programação R

- Linguagem feita para estatísticos
- Amplamente usada no meio acadêmico
- Tem entrada forte no mercado, recebendo grandes investimentos.



[Aprendendo R com R](#) | [Udacity + Facebook - Data analysis with R](#) | [R-bloggers](#) | [DataCamp - Intro to R](#)  
[Coursera - R Programming](#) | [R-Cookbook](#)

# Programação

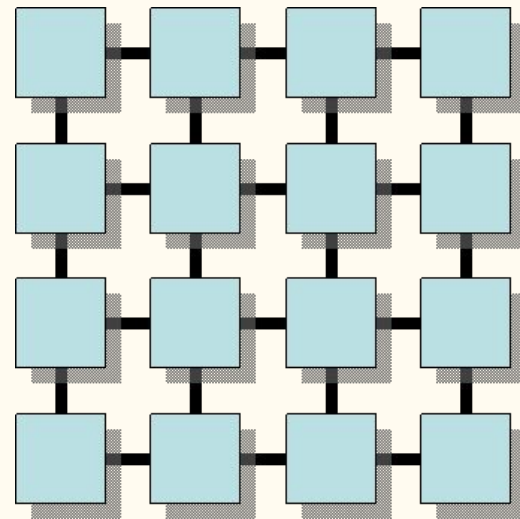
# Bancos de Dados

- Grande disponibilidade de DBMS
- PostgreSQL, MongoDB, MySQL, ....
- SQL vs NoSQL



# Programação Computação Distribuída

- Algumas vezes, um computador apenas não dá conta
- MapReduce, Apache Spark, Hadoop, etc ...



[MapReduce e Hadoop \(Udacity + Cloudera\)](#) | [Intro to Parallel Programming \(Nvidia + Udacity\)](#)

# Aquisição e Limpeza de Dados

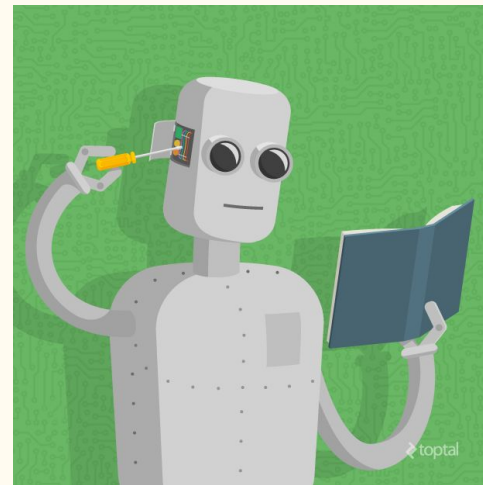
- 80% do processo é gasto entre aquisição e limpeza de dados
- Inconsistência de dados é algo comum
- Sem dados não há Ciência de Dados



[Model Building and Validation \(Udacity\)](#) | [Cleaning Data in R \(pago\) \(Data Camp\)](#)  
[Data Mining \(University of Illinois\)](#)

# Programação Machine Learning

- Permite criar modelos extramente complexos e poderosos
- Sistemas Recomendadores, Busca, Aprendizado Dinâmico
- Redes Neurais, KNN, Máquinas de Vetor Suporte



[Intro to Machine Learning](#), [Supervised ML](#), [Unsupervised ML](#), [Reinforcement Learning](#) (Udacity)

[Machine Learning Specialization](#) (University of Washington)

[Machine Learning](#) (Stanford)

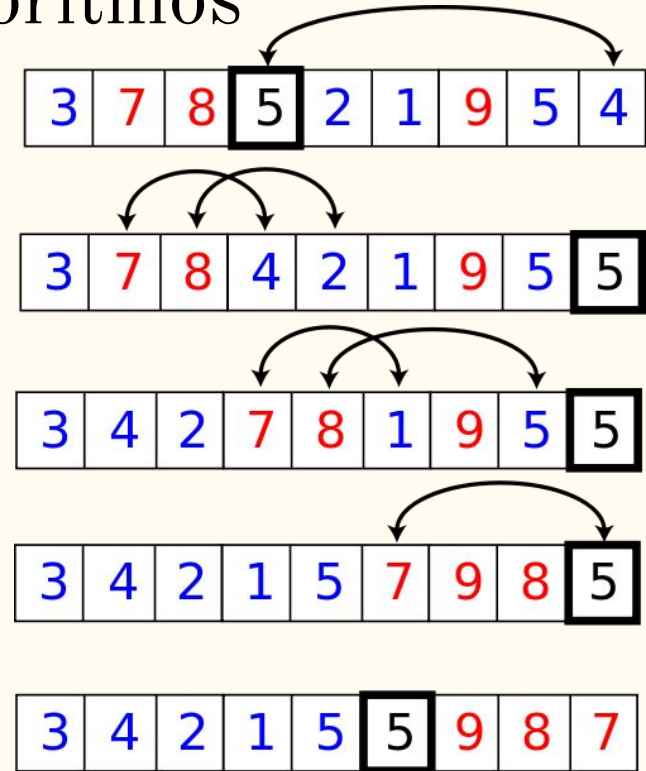
# Programação - Algoritmos

- **Eficiência faz diferença!**
- Saber implementar algoritmos paralelizáveis

Algoritmos Parte I e II (Princeton)

Algoritmos I e II (Stanford)

Especialização (UC San Diego)



# Habilidades - Estatística e Matemática

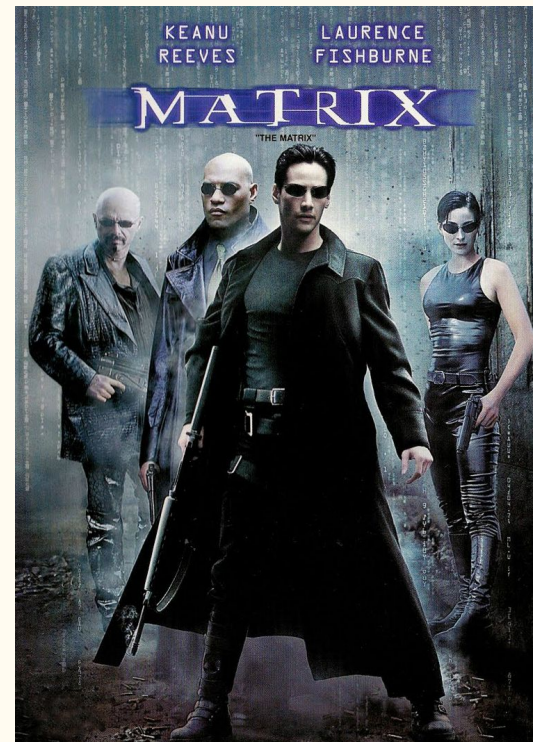


Matemática / Estatística

# Operações Matriciais

- Multiplicação
- Fatoração matricial
- Autovalores e Autovetores

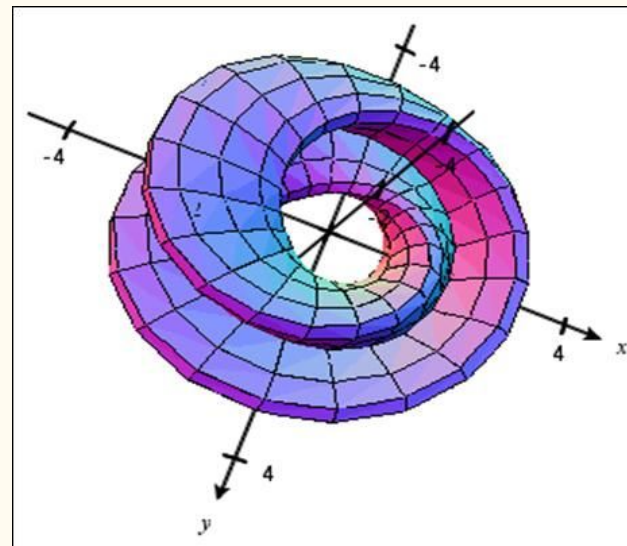
Coursera - [Coding the Matrix](#) | MIT- [Linear Algebra](#)



Matemática / Estatística

# Cálculo Multivariável

- Matrizes Jacobiana e Hessiana
- Base para Modelos Estatísticos, Otimização e Aprendizado de Máquina.

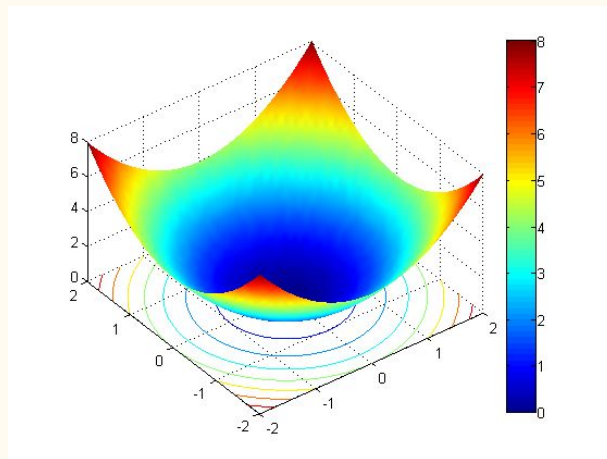


Berkeley - [Multivariate Calculus](#) | MIT - [Multivariate Calculus - 2007](#) | MIT - [Multivariate Calculus - 2010](#)

Matemática / Estatística

# Otimização

- Métodos para minimizar uma função específica
- Base para a maioria dos algoritmos de Aprendizado de Máquina e Regressões Estatísticas

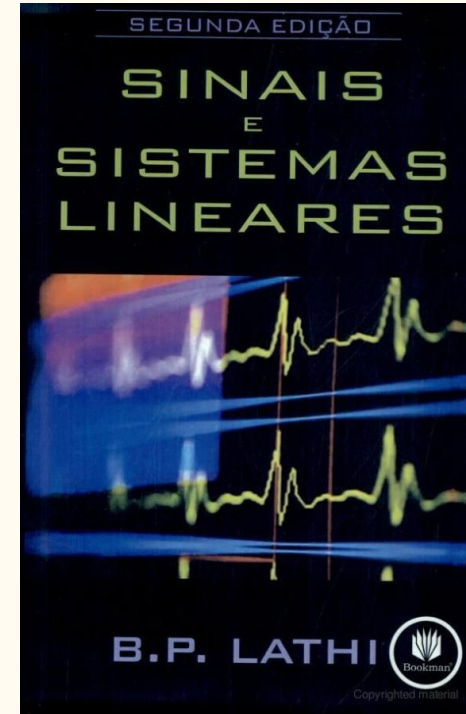


Stanford - [Convex Optimization \(Avançado\)](#) | Cursos de Cálculo Multivariável | [Coursera - Linear Programming](#)

# Matemática - Processamento de Sinais

- Detectar periodicidade em eventos
- Remover ruído nos dados

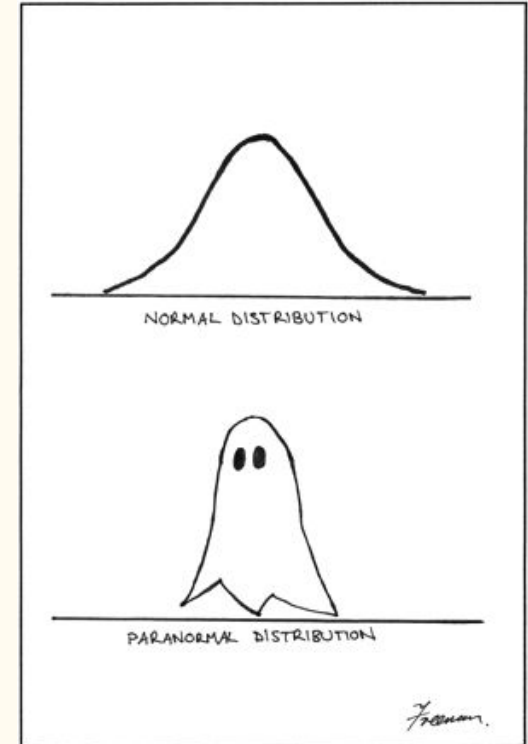
Digital Signal Processing | DSP (MIT - Oppenheim)



# Estatística - Distribuições

- Poisson, Normal, Uniforme, Gama
- Saber qual distribuição melhor representa o problema
- Testes de Hipóteses
- Lidar com amostras enviesadas

[Intro to Statistics](#) | [Intro to Descriptive Statistics](#) | [Intro to Inferential Statistics](#)



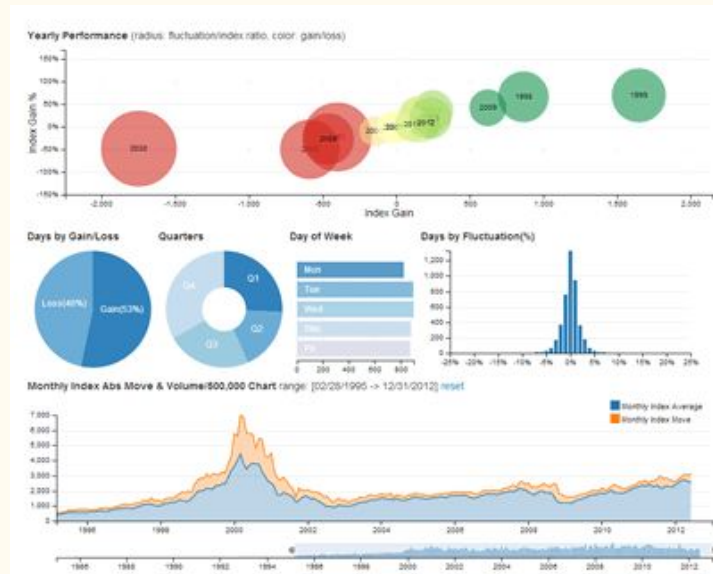
# Testes A/B

- Permite comparar resultados de duas ou mais modificações



# Visualização de dados

- Comunicação de Resultados
- Representar mais do que duas dimensões
- Representação de diferentes variáveis



Udacity - Data Visualization and D3.js | <https://plot.ly> | [Flowing Data](#) | [Edward Tufte - Envisioning Information](#)

Saúde

E-commerce

Ecologia

Personalização

Mercado Financeiro

Marketing

# Conhecimentos de Domínio

Psicologia

Biologia

Otimização de Processos

Logística

Política

Esportes



# Cursos sobre Ciência de Dados

- Udacity - Intro to Data Science
- edX - Data Science Ethics
- John Hopkins - Data Science Especialization
- Big Data University

# Links Diversos

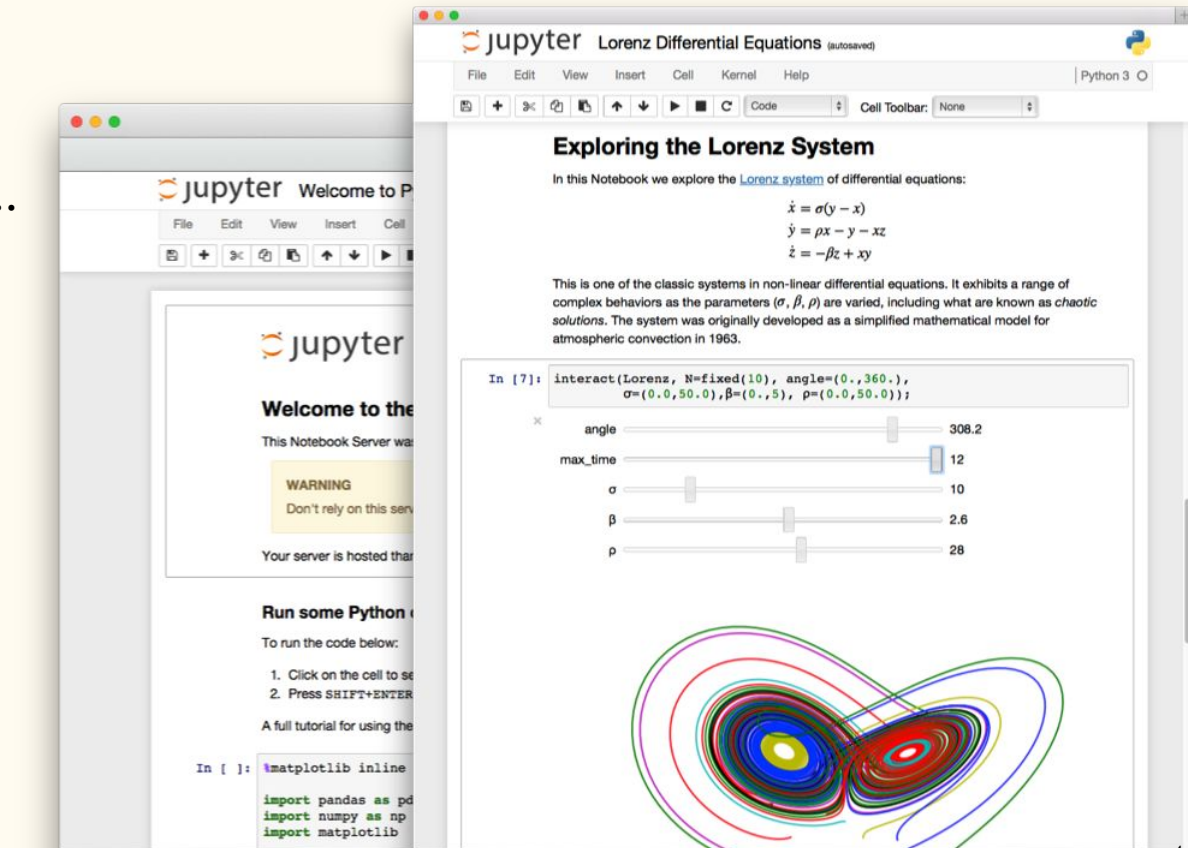
- [Quora Data Science FAQ](#)
- [DataCamp](#) - Cursos bons, alguns free
- [Open Datasets](#)
- [Kaggle](#)
- Coursera - [Lista de Cursos](#)



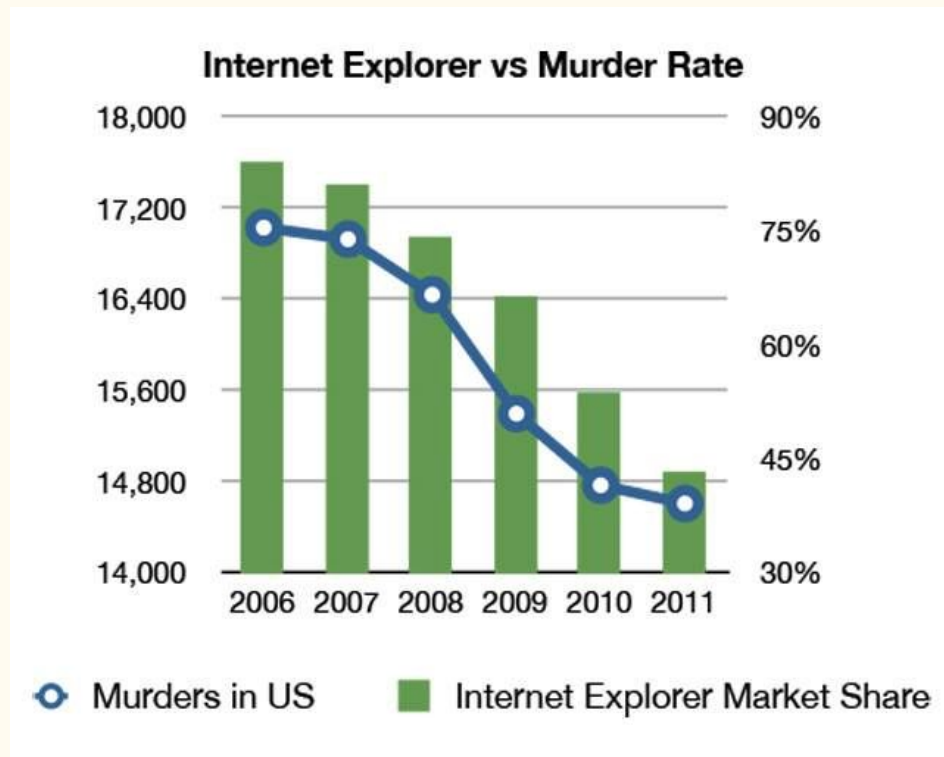
# Tips and Tricks

# Jupyter

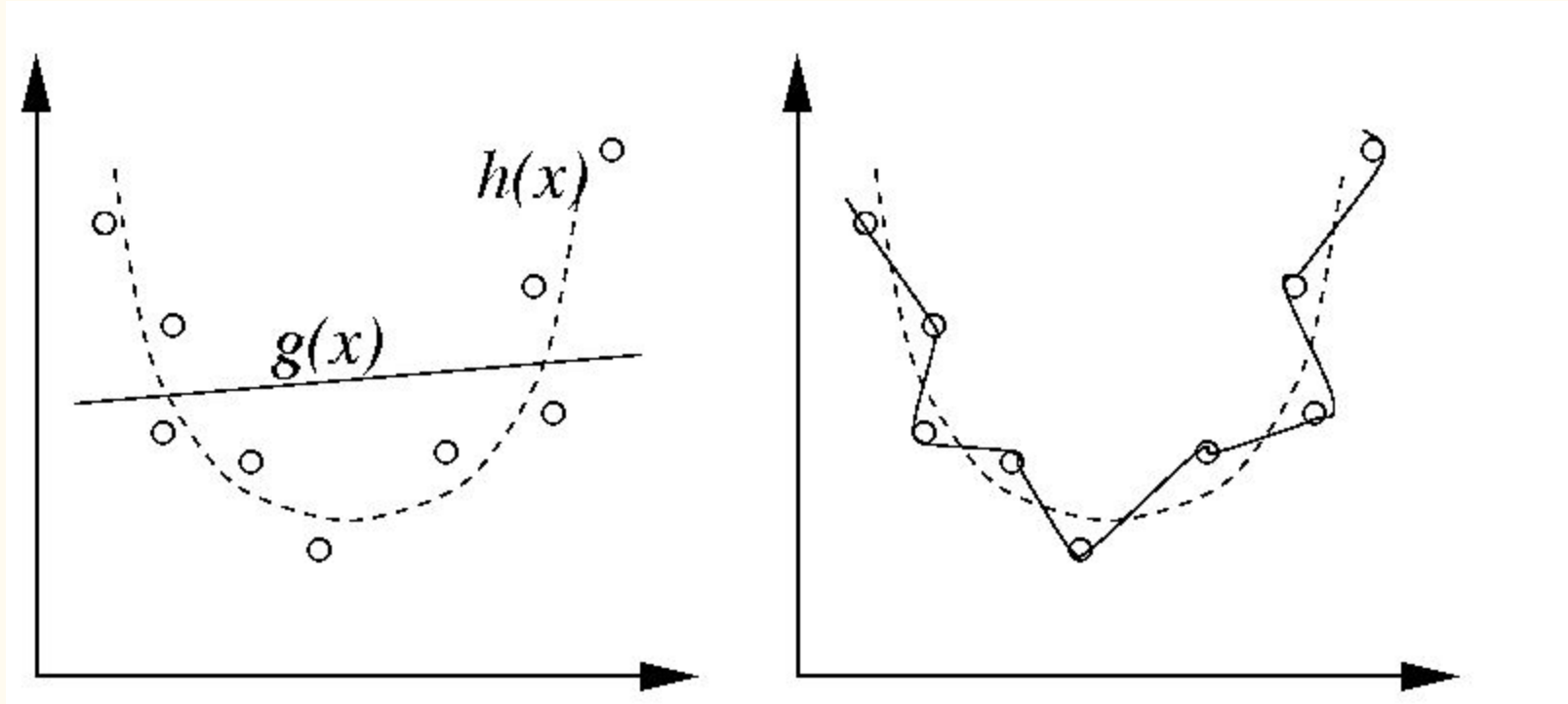
- Código + Texto
- Python, R, Julia, Go .....



# Correlação não é Causalidade



# Cuidado com erro muito baixo



Método Científico significa que resultados devem ser reproduzíveis!

# Curriculum da Automação e Data Science



# Obrigatórias

1	DAS 5334	Introdução à Informática para Automação
	MTM 5161	Cálculo A
2	DAS 5102	Fundamentos da Estrutura da Informação
	MTM 5162	Cálculo B
	MTM 5245	Álgebra Linear
3	DAS 5103	Cálculo Numérico para Controle e Automação
	MTM 5163	Cálculo C
4	INE 5108	Estatística e Probabilidade para Ciências Exatas
	DAS 5114	Sinais e Sistemas Lineares

# Optativas

DAS 5341	Inteligência Artificial Aplicada a Controle e Automação
DAS 5306	Programação Concorrente e Sistemas de Tempo Real
INE 5225	Fundamentos de Sistemas de Bancos de Dados
EMC 5341	Otimização

## **Computação (não está no lista de optativas)**

INE 5433	Reconhecimento de Padrões
INE 5644	Data Mining

# Concluindo

- Um cientista de dados é um generalista, não um especialista
- Muitas habilidades não triviais para se estudar
- Inúmeras oportunidades de criar impacto
- Extremamente desafiador, mas muito divertido
- A graduação não é suficiente, é preciso ir além!

**Obrigado!**

**Github** | **LinkedIn**