

Contexto de Big Data, Ciência de Dados e KDD

Fabício J. Barth

quantidade de informações

1970

1980

1990

2000

2010

2020



Sempre houve:

Produção de informação

Geração de dados

Geração de conteúdo

quantidade de informações

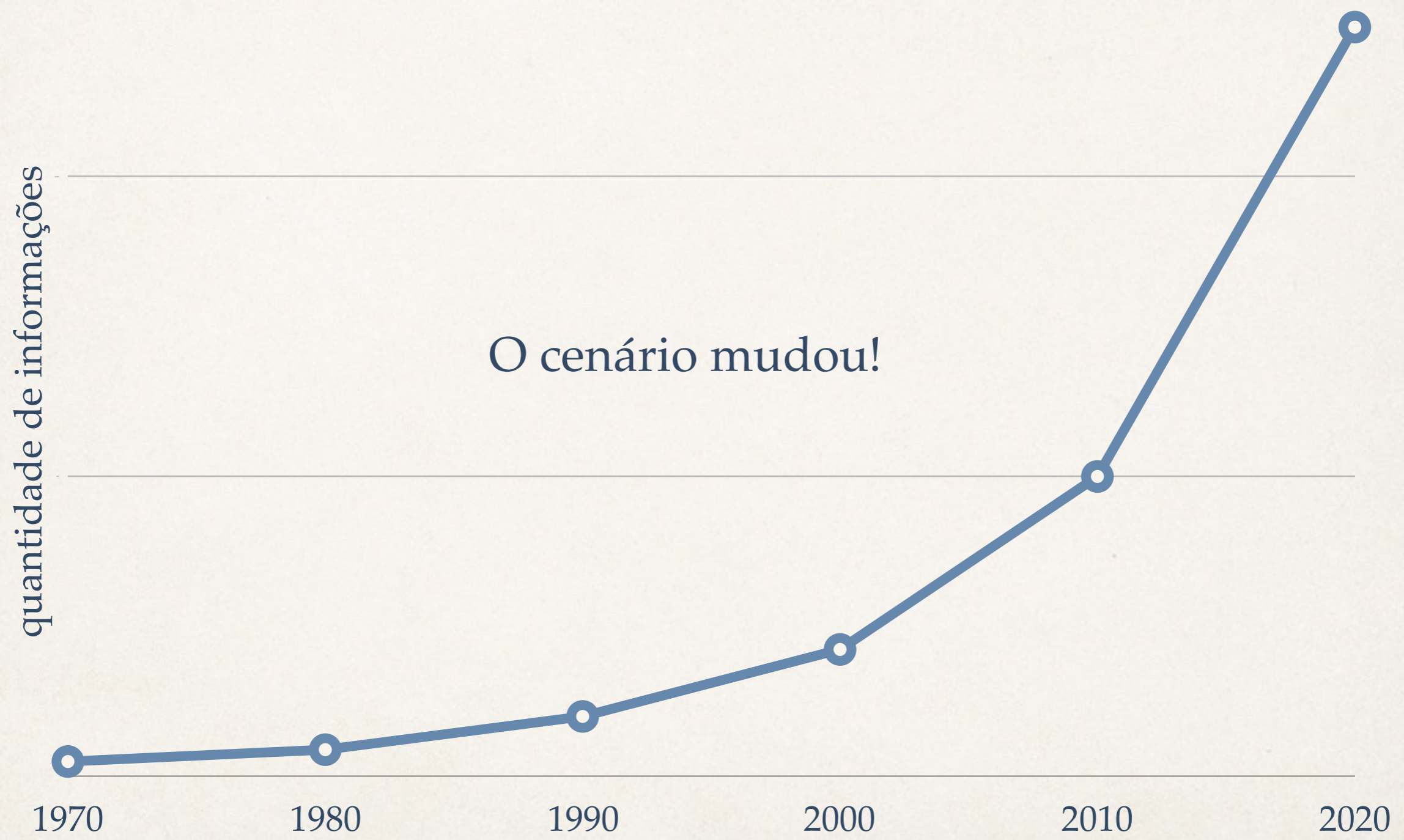


Sempre **desejou-se:**
sintetizar a informação
manter, disseminar, organizar, criar
conhecimento e tomar decisões mais
assertivas com base nos dados.

Métodos, processos e ferramentas

- ❖ Gestão de Conhecimento, Sistemas Especialistas e Mineração de Dados
- ❖ Sistemas Especialistas e Projetos de Mineração de Dados (Processo de Descoberta de Conhecimento) só funcionavam em cenários muito bem delimitados e utilizando dados muito bem tratados e estruturados.
- ❖ Também, manipular dados diferentes dos dados não estruturados não parecia ser algo relevante.





quantidade de informações

O cenário mudou!

1970

1980

1990

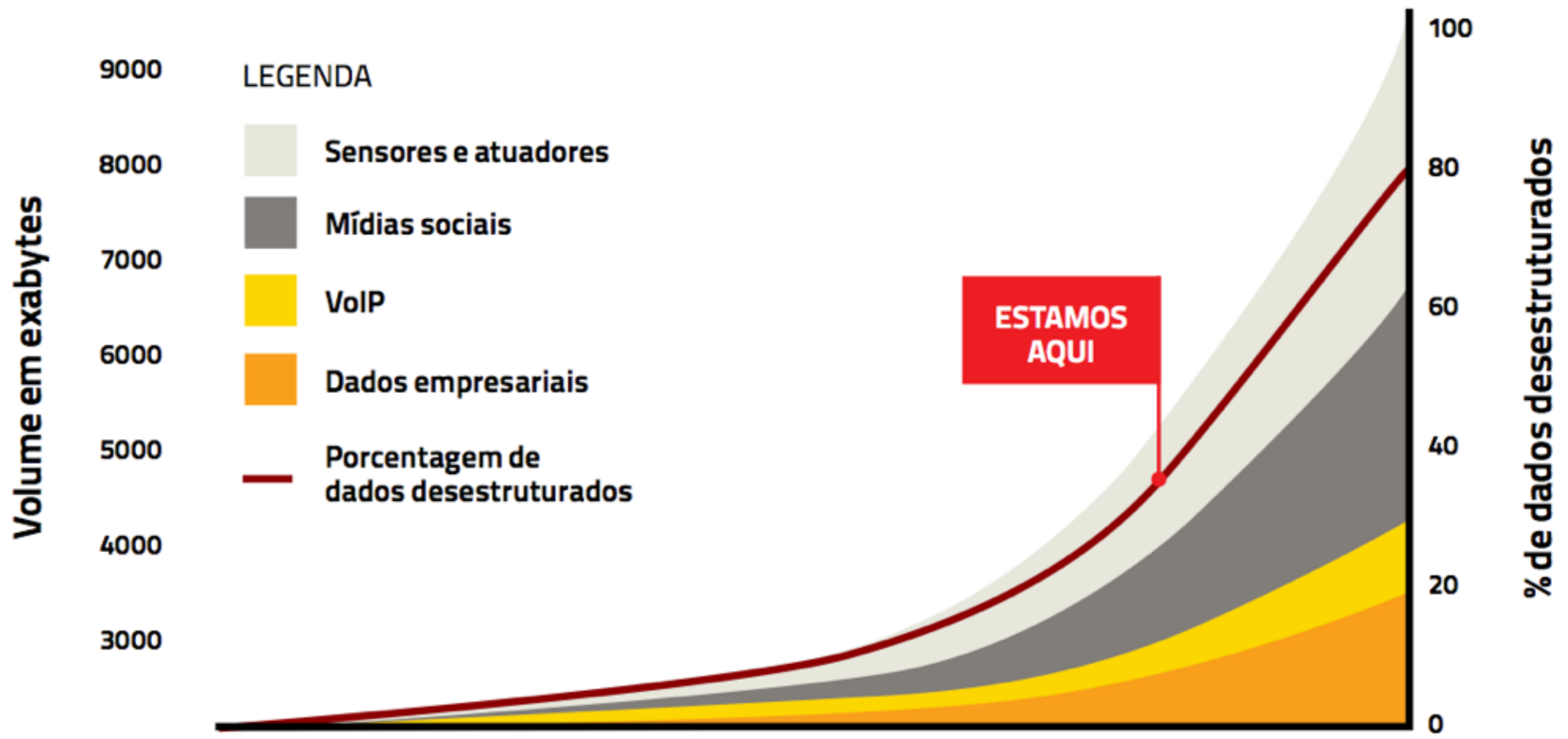
2000

2010

2020

NOVO CENÁRIO | PROGRESSÃO DOS DADOS ARMAZENADOS AO LONGO DO TEMPO

Tendência clara de crescimento dos dados desestruturados, também chamados de dados incertos (*uncertain data*, em inglês)



Gandour, F. O que muda com a computação cognitiva? Revista de ESPM, Set/Out de 2014.

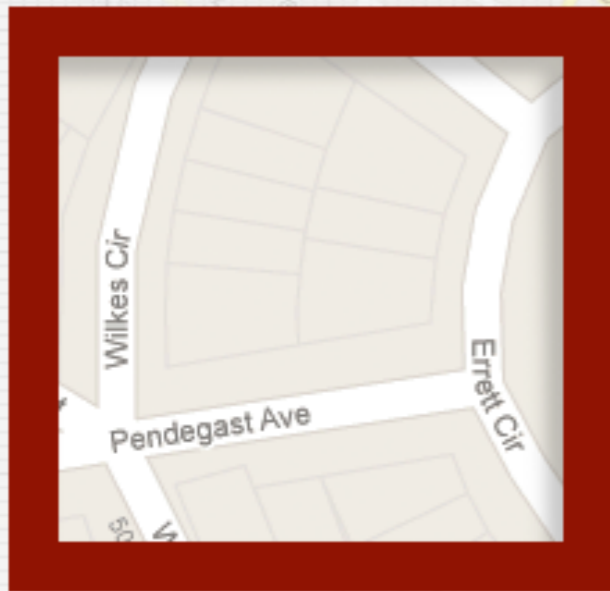
- * O cenário mudou!
- * Mas o desejo ainda continua:
 - * sintetizar
 - * manter
 - * disseminar
 - * organizar
 - * encontrar
 - * tomar decisões baseadas em



Mas o desafio mudou

- ❖ Ficou mais complexo devido as características dos dados, da forma como eles são gerados e das novas necessidades dos usuários.
 - ❖ O **volume** de dados gerados é muito alto.
 - ❖ A **velocidade** com que eles são gerados e perdem a validade é muito rápida.
 - ❖ A **variedade** das fontes é bem diversificada (estruturada + não estruturada)
- ❖ Aparentemente, os usuários não querem mais saber do **passado**. Estão muito interessados no **presente** e **futuro**.

Alguns exemplos



PredPol®

Predict Crime in Real Time®

PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

■ SEE PREDICTIVE POLICING IN ACTION. LOS ANGELES, CALIFORNIA

[LEARN MORE](#)



Preventative Tactics



Field Tested



Easy to Use



In The News

**Entrada: 13 milhões de registros históricos sobre crimes em LA.
Saída: determinar quando um crime irá acontecer.**

Earthquake Hazards Program

Home About Us Contact Us

EARTHQUAKES

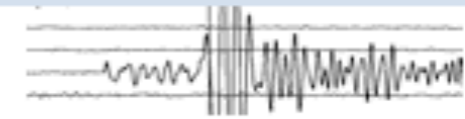
HAZARDS

LEARN

PREPARE

MONITORING

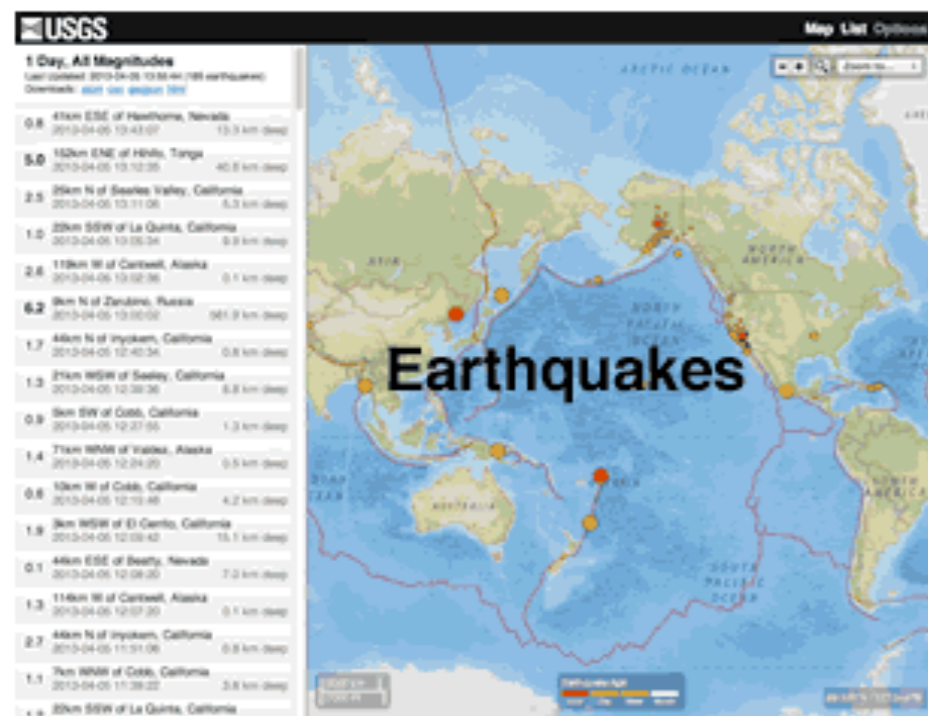
RESEARCH



The USGS Earthquake Hazards Program is part of the [National Earthquake Hazards Reduction Program \(NEHRP\)](#), established by Congress in 1977. We monitor and report earthquakes, assess earthquake impacts and hazards, and research the causes and effects of earthquake.



Latest Earthquakes



Significant Earthquakes

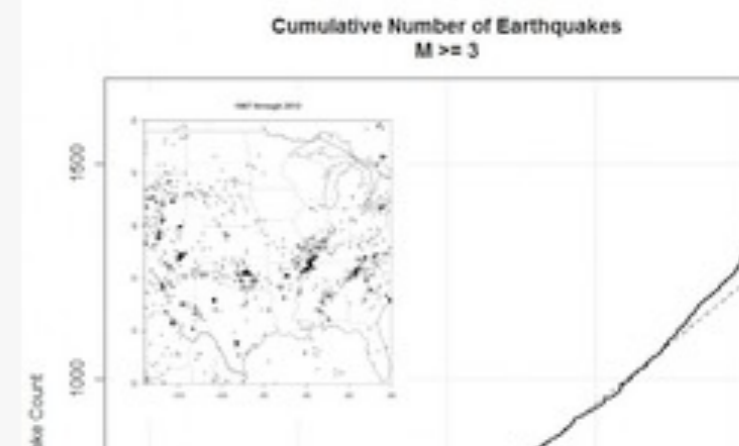
Past 30 Days

- 6.6** [98km WSW of Mutis, Colombia](#)
2013-08-13 15:43:15 UTC 12.0 km deep
- 4.9** [9km SSW of Volcano, Hawaii](#)
2013-08-11 15:54:05 UTC 31.8 km deep
- 5.9** [13km E of Chabu, China](#)
2013-07-21 23:45:56 UTC 9.8 km deep
- 6.5** [46km ESE of Blenheim, New Zealand](#)
2013-07-21 05:09:31 UTC 14.0 km deep

[Significant Earthquake Archive](#)

Featured Items

[Man-Made Earthquakes](#)



View recent events or search for past earthquakes. Optimized for mobile and desktop.



[Real-time Feeds & Notifications](#)

Get real-time earthquake notifications sent to you using a number of popular mediums: Feeds, Email, Twitter, etc...

Entrada: rede de sensores que cobre todo o mundo
Saída: determinar quando um terremoto irá acontecer

Your Amazon.com > Recommended for You > Books

These recommendations are based on [items you own](#) and more.

view: **All** | [New Releases](#) | [Coming Soon](#)

Just For Today

[Browse Recommended](#)

Recommendations Books

- [Arts & Photography](#)
- [Audible Audiobooks](#)
- [Bargain Books](#)
- [Biographies & Memoirs](#)
- [Books on CD](#)
- [Business & Investing](#)
- [Calendars](#)
- [Children's Books](#)
- [Christian Books & Bibles](#)
- [Comics & Graphic Novels](#)
- [Computers & Technology](#)
- [Cookbooks, Food & Wine](#)
- [Crafts, Hobbies & Home](#)
- [Education & Reference](#)
- [Gay & Lesbian](#)
- [Health, Fitness & Dieting](#)
- [History](#)

1.



Machine Learning for Hackers

by Drew Conway (February 22, 2012)
Average Customer Review: ★★★★★ (21)
In Stock

List Price: ~~\$39.99~~
Price: **\$26.48**
[74 used & new from \\$21.90](#)

[Add to Cart](#) [Add to Wish List](#)

I own it Not interested ★★★★★ Rate this item
Recommended because you added **Data Mining with R** to your Wishlist and more (Fix this)

2.



Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media

by Matthew A. Russell (February 8, 2011)
Average Customer Review: ★★★★★ (23)
In Stock

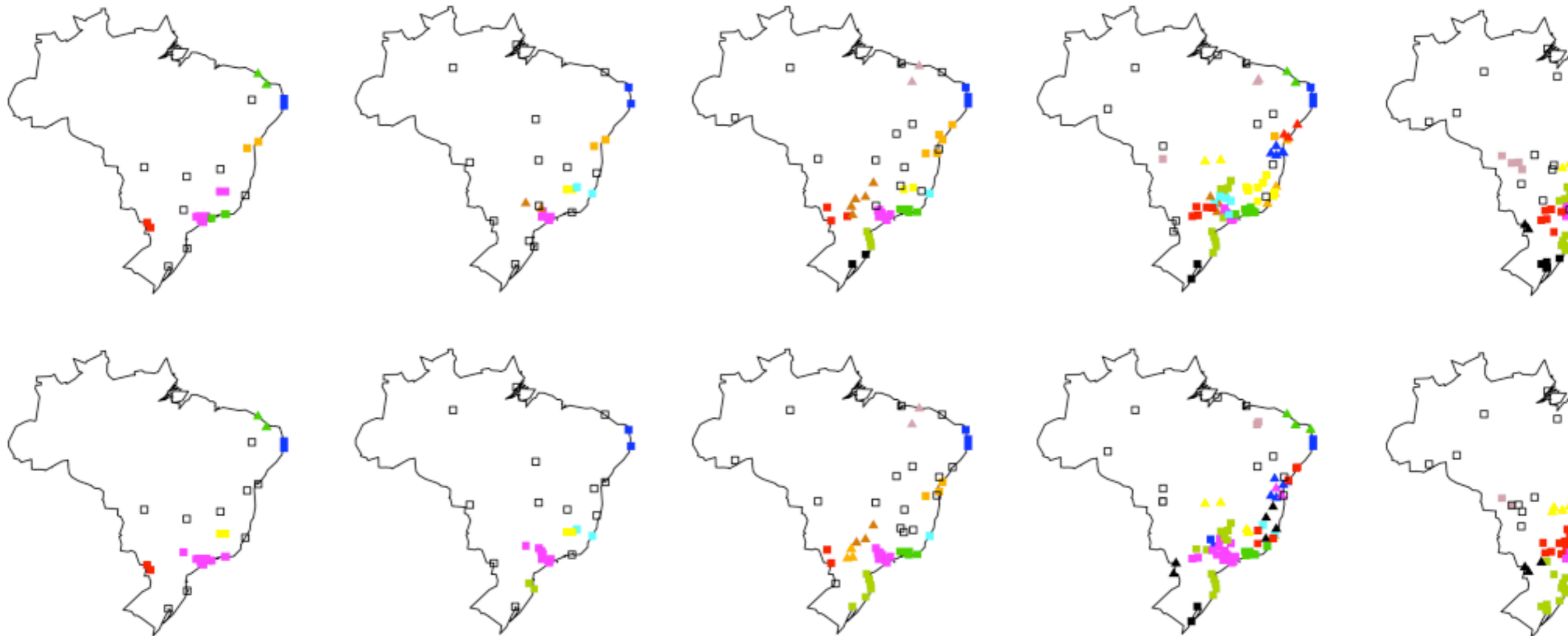
List Price: ~~\$39.99~~
Price: **\$26.48**
[84 used & new from \\$20.00](#)

[Add to Cart](#) [Add to Wish List](#)

I own it Not interested ★★★★★ Rate this item
Recommended because you added **Data Mining with R** to your Wishlist and more (Fix this)

Entrada: milhões de itens e milhões de usuários

Saída: recomendar itens com acurácia alta para os usuários (clientes)





Yelp Recruiting Competition

Finished
Wednesday, March 27, 2013 Jobs • 352 teams Sunday, June 30, 2013

- Dashboard
- Home
- Data
- Make a submission
- Information
- Description
- Evaluation

[Competition Details](#) » [Get the Data](#) » [Make a submission](#)

How many "useful" votes will a Yelp review receive? Show off your skills to land an interview for a position on a Yelp data mining team!

NETFLIX

Português ▼



1 MÊS GRÁTIS

Assista a filmes e séries de TV quando quiser, onde quiser. Apenas R\$16,90 ao mês.

Comece a utilizar seu mês grátis



Meus Sets

- Análise de Investimentos
- Eleições 2010
- Criar Novo

Empresas / Entidades

- BRDESCO
- GERDAU
- PETROBRAS
- USIMINAS
- VALE R DOCE

Feeds

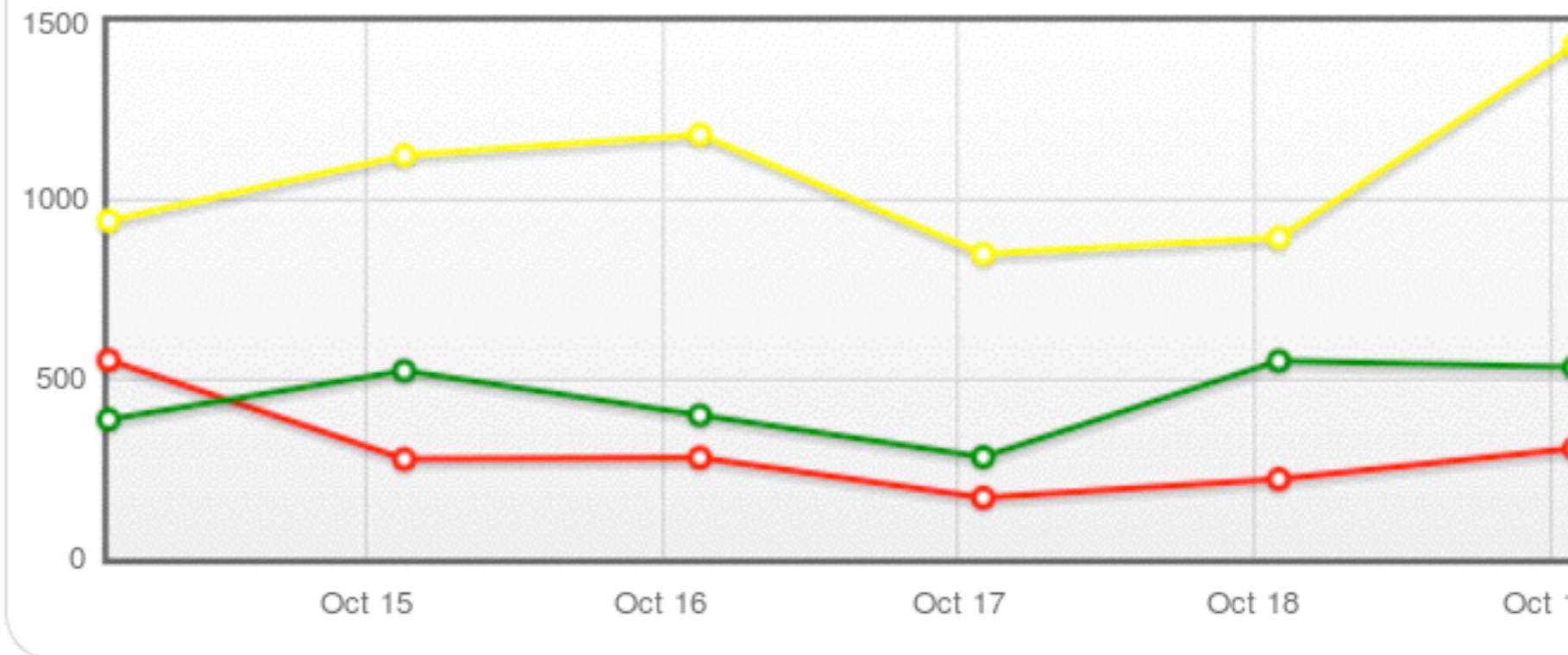
- Abril.com - Economia
- BBCBrasil.com | Tópicos | Economia
- Direto do Pregão
- estadao.com.br - Últimas notícias
- Folha Online - Dinheiro - Principal
- G1 Economia e Negócios
- Google News - Petróleo

Status

Todos Últimas 50 notícias Todas

Sumário

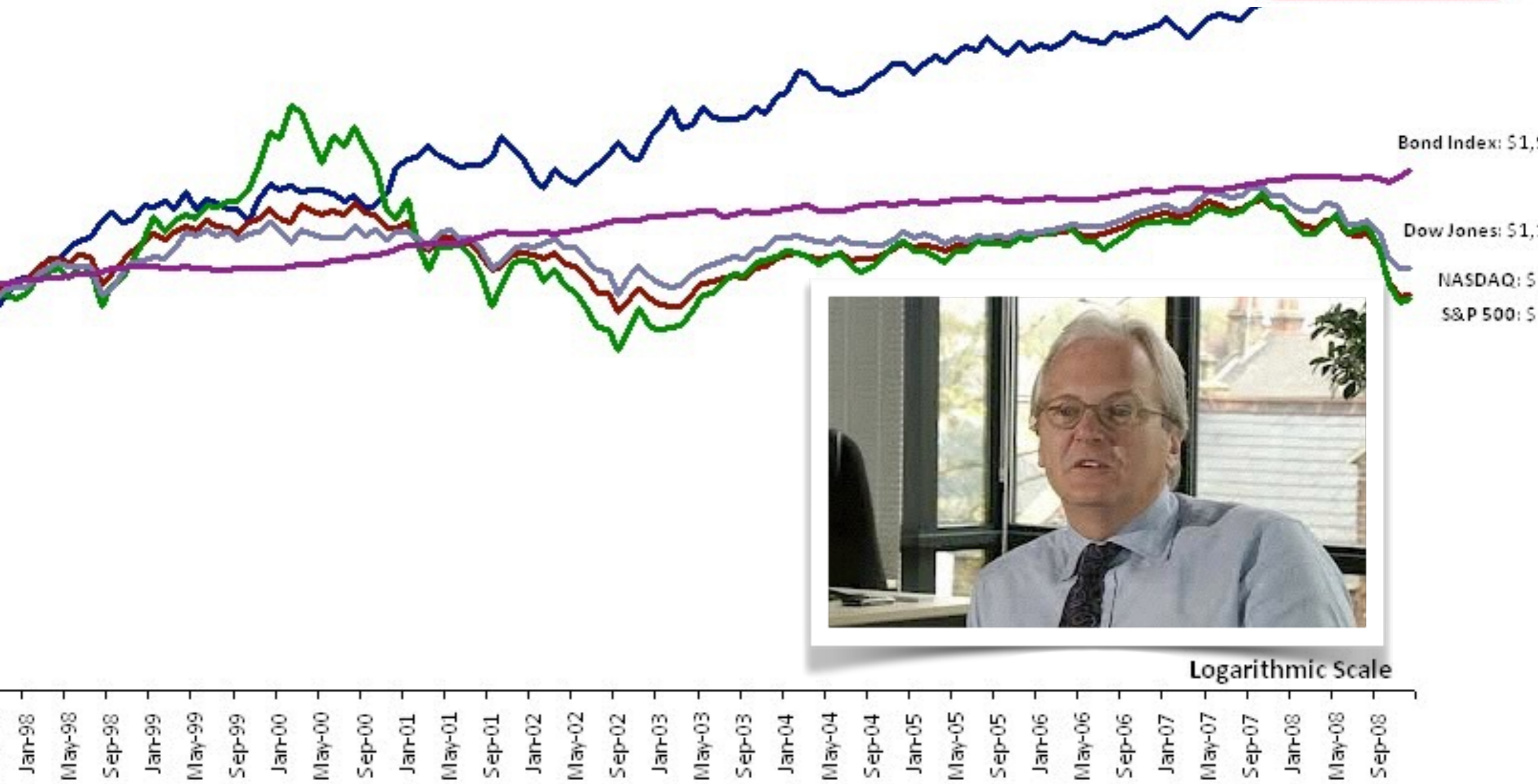
12420 ▲ 24% ● 58% ▼ 16% 7 dias atualizar



Noticias filtradas

Data/Hora	Titulo
-----------	--------

Criar robos que compram e vendem ações



O que estes projetos têm em comum?

- ❖ **Manipulam grandes volumes de informação**
- ❖ **Outros exemplos de grandes volumes de informação:**
 - ❖ **A380: Heathrow - JFK: 640 TBs de log**
 - ❖ **Twitter: 12+ TBs of tweet every day**
 - ❖ **Facebook: 25+ TBs of log data every day**

O que estes projetos têm em comum?

- ❖ **A origem dos dados é muito variada.**



Mobile Sensors



FACEBOOK
GROWS BY
250 MILLION
PHOTOS / DAY

Social Media



Video
Surveillance

Video Rendering



READING METERS
EVERY 15 MINS.
IS 3,000X MORE
DATA INTENSIVE



Smart Grids

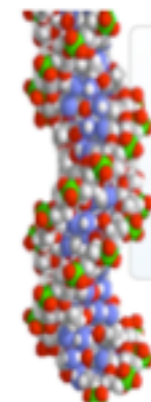


Geophysical
Exploration

Medical Imaging



Gene Sequencing



COST TO SEQUENCE
ONE GENOME
HAS FALLEN FROM
\$100M IN 2001
TO \$10K IN 2011

O que estes
projetos têm
em comum?



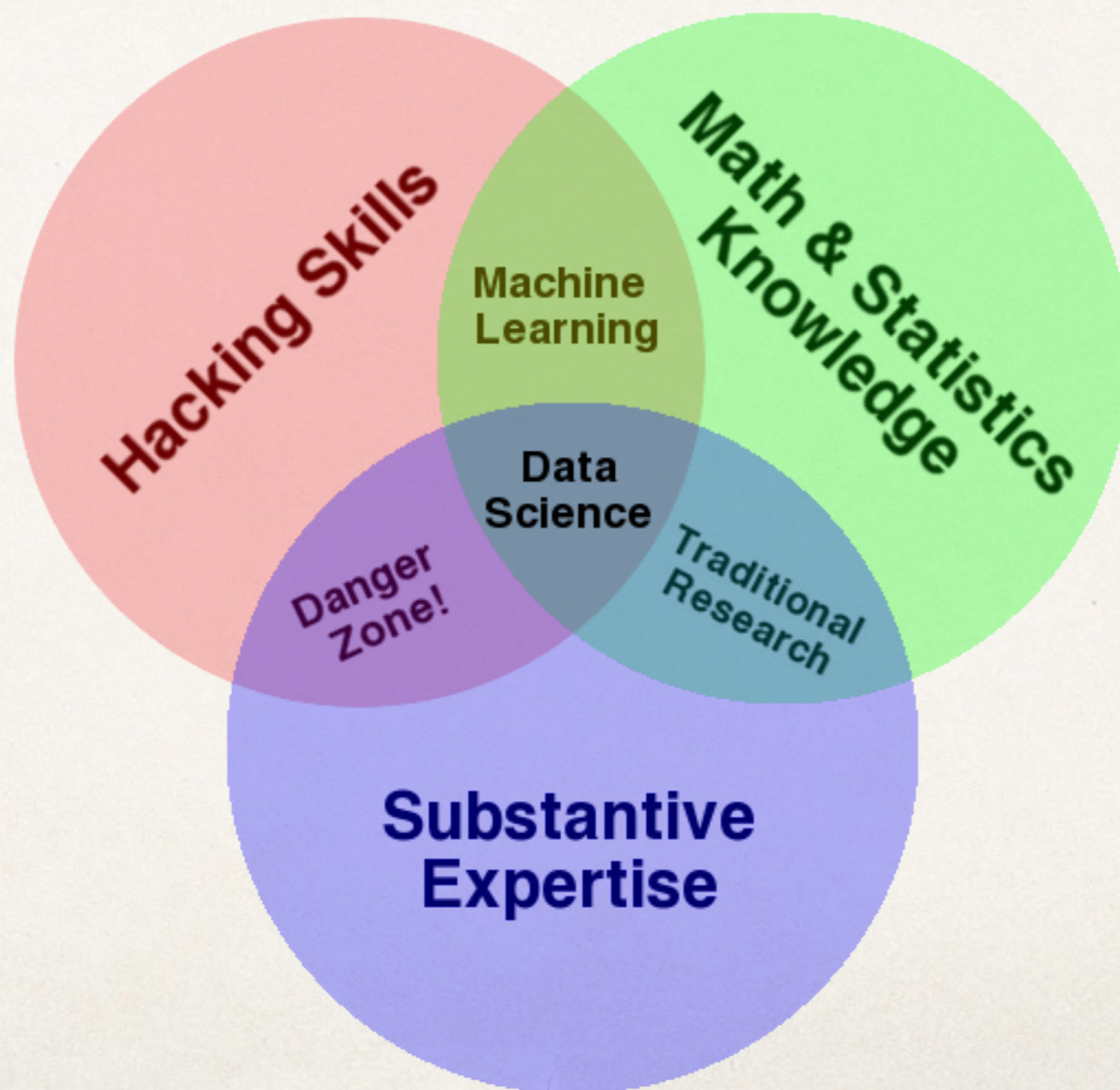


Queremos modelos predictivos

Outros exemplos

- ❖ Identificar comportamento anômalo (i.e., fraudes, falhas)
- ❖ Sumarizar tendências de publicações de artigos e patentes sobre um determinado tema.
- ❖ Sumarizar e filtrar notícias relevantes.
- ❖ Sumarizar a opinião expressa na Web sobre a sua empresa.
- ❖ Identificar padrões de navegação em sites.
- ❖ Identificar conteúdo impróprio em sites.

Ciência de Dados (Data Science)



Cientísta de Datos (Data Scientist)

- ❖ Data Scientist: The sexiest job of the 21st Century. Harvard Business Review.
- ❖ **Data Scientist** applies advanced **analytical** tools and algorithms to generate **predictive insights** and **new** product **innovations** that are a direct result of the data.

Processo de Descoberta de Conhecimento

Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)

Qual é a pergunta?

É possível classificar espécies do gênero iris levando em consideração apenas o tamanho das plantas?



Iris virginica



Iris setosa



Iris versicolor



Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)

Qual é a pergunta?

Aquisição e pré-processamento dos dados

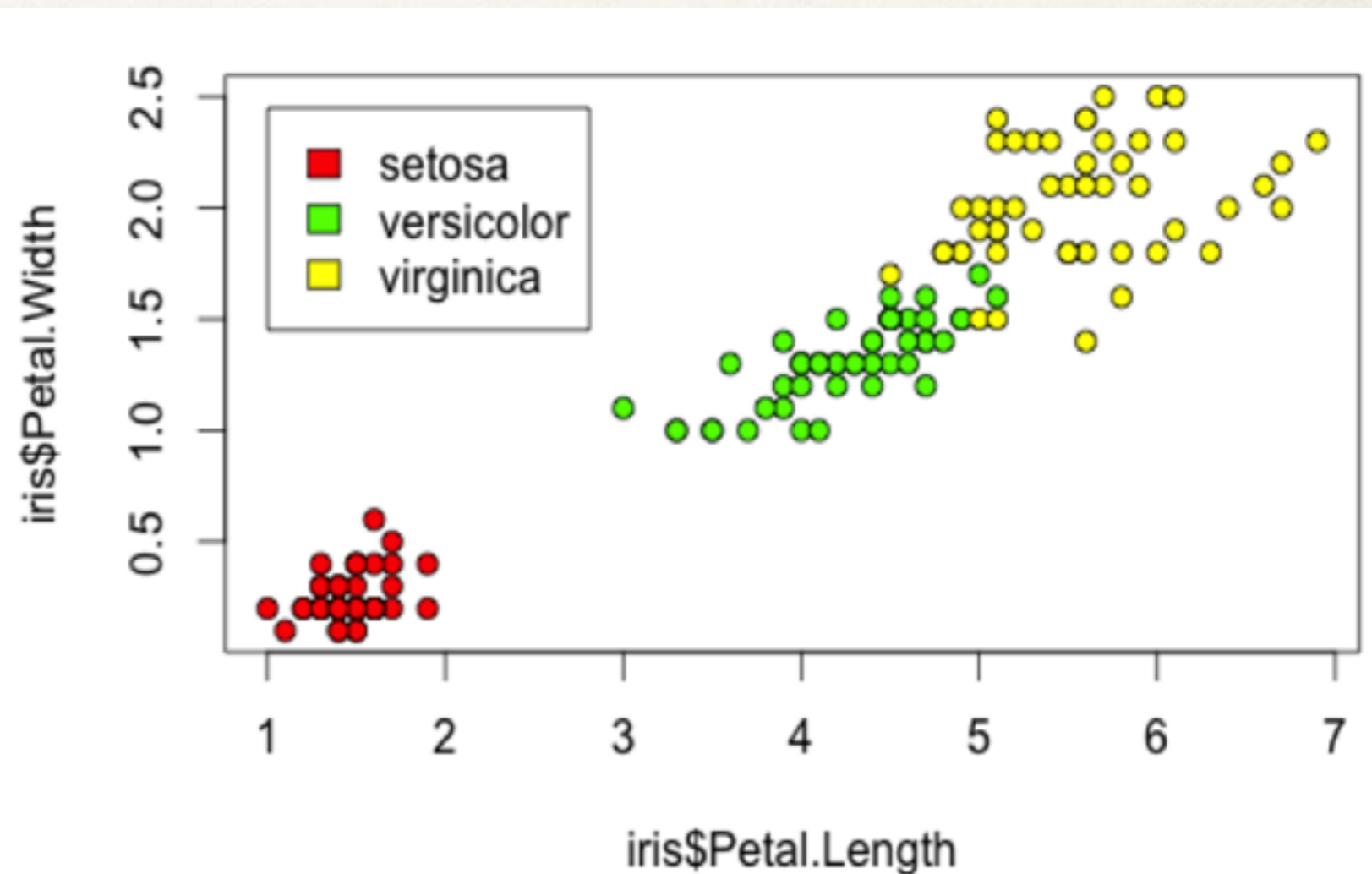
```
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5         1.4         0.2  setosa
2           4.9         3.0         1.4         0.2  setosa
3           4.7         3.2         1.3         0.2  setosa
4           4.6         3.1         1.5         0.2  setosa
5           5.0         3.6         1.4         0.2  setosa
6           5.4         3.9         1.7         0.4  setosa
>
>
> sapply(iris,class)
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
  "numeric"   "numeric"   "numeric"   "numeric"   "factor"
```

Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)

Qual é a pergunta?

Aquisição e pré-processamento dos dados

Análise exploratória



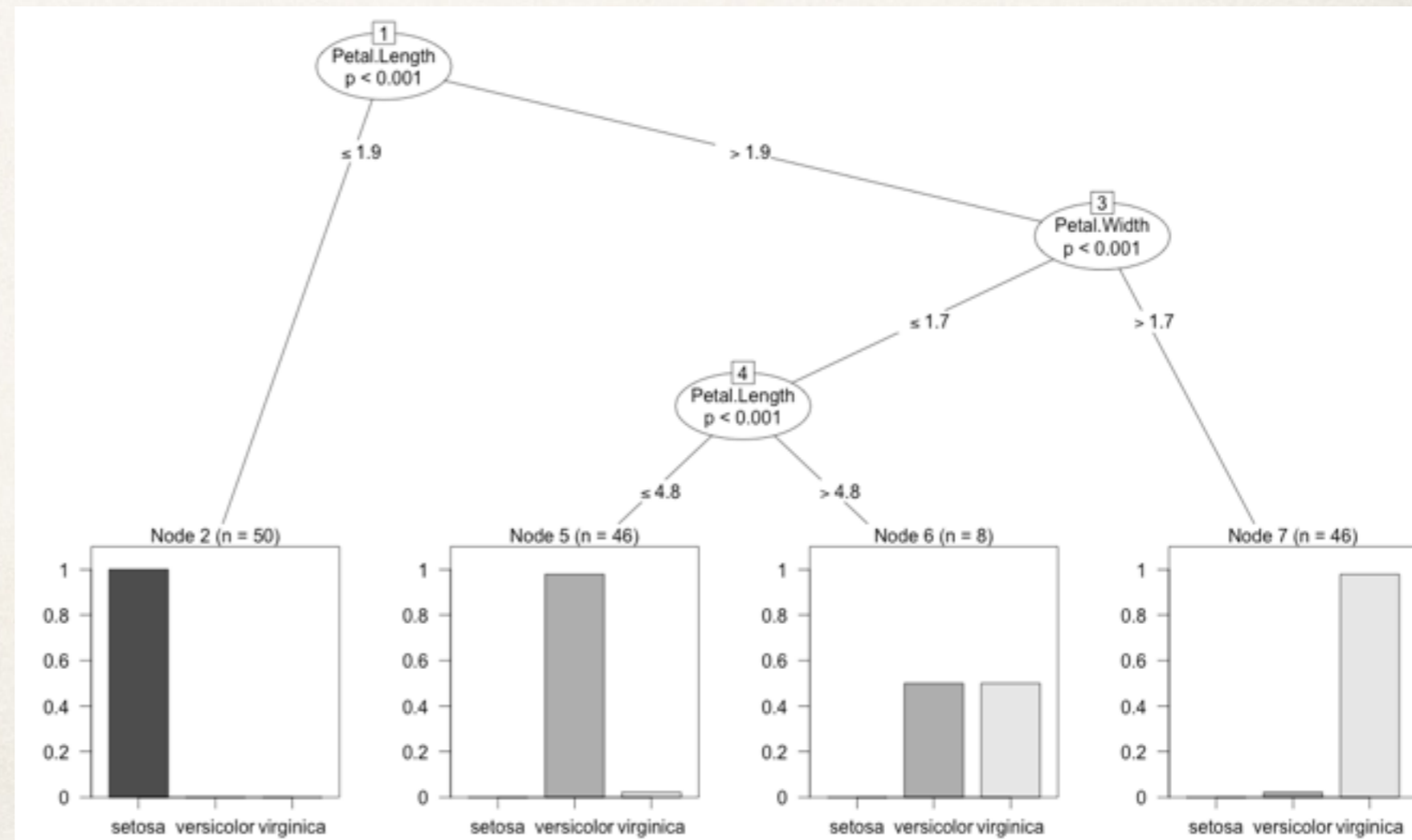
Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)

Qual é a pergunta?

Aquisição e pré-processamento dos dados

Análise exploratória

Modelagem



Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)

Qual é a pergunta?

Aquisição e pré-processamento dos dados

Análise exploratória

Modelagem

Avaliação do modelo

```
> table(predict(model,iris), iris$Species)
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	49	5
virginica	0	1	45

Acurácia do modelo?

Quantidade falsos positivos?

Falsos negativos?

Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)

Qual é a pergunta?

Aquisição e pré-processamento dos dados

Análise exploratória

Modelagem

Avaliação do modelo

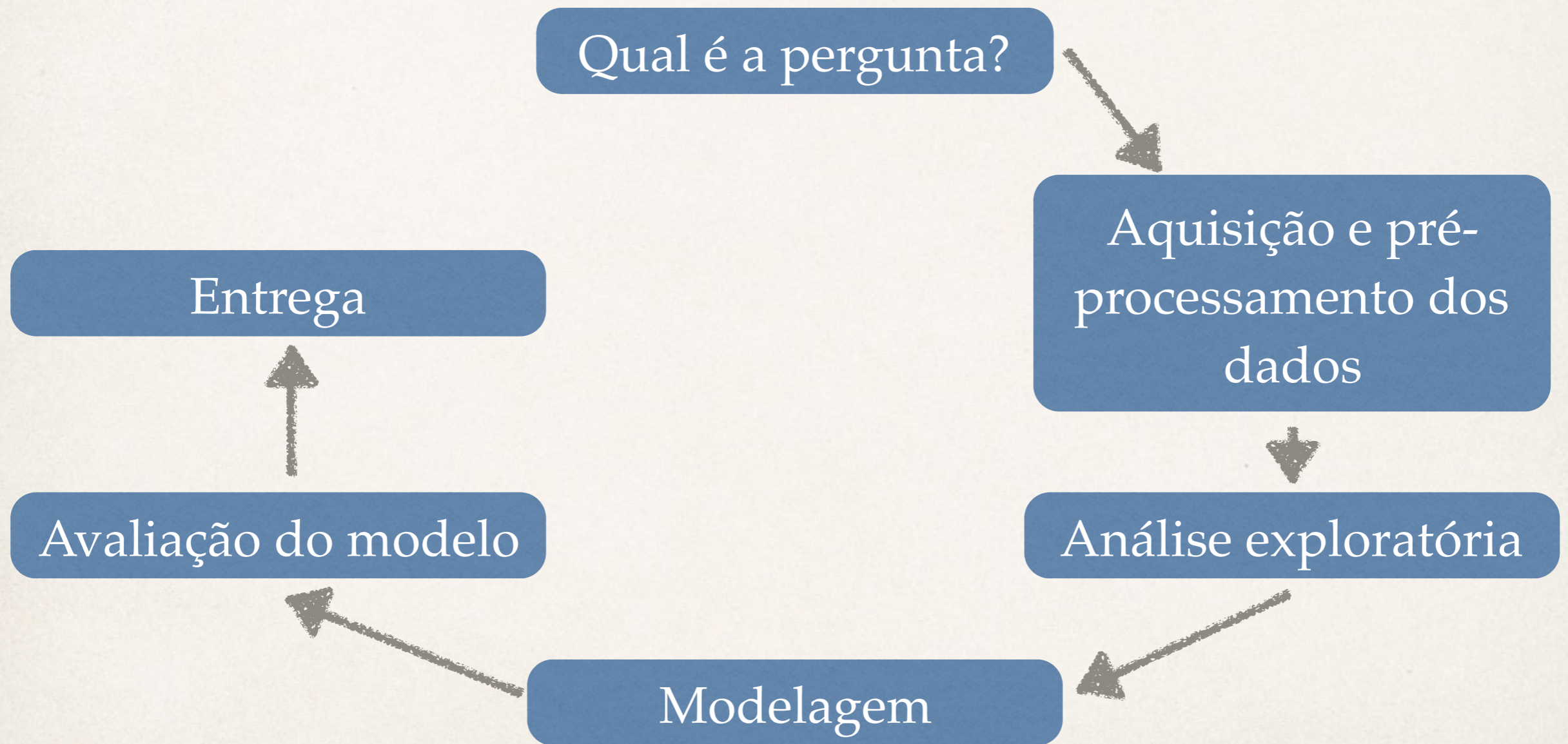
Entrega

Relatórios Estáticos

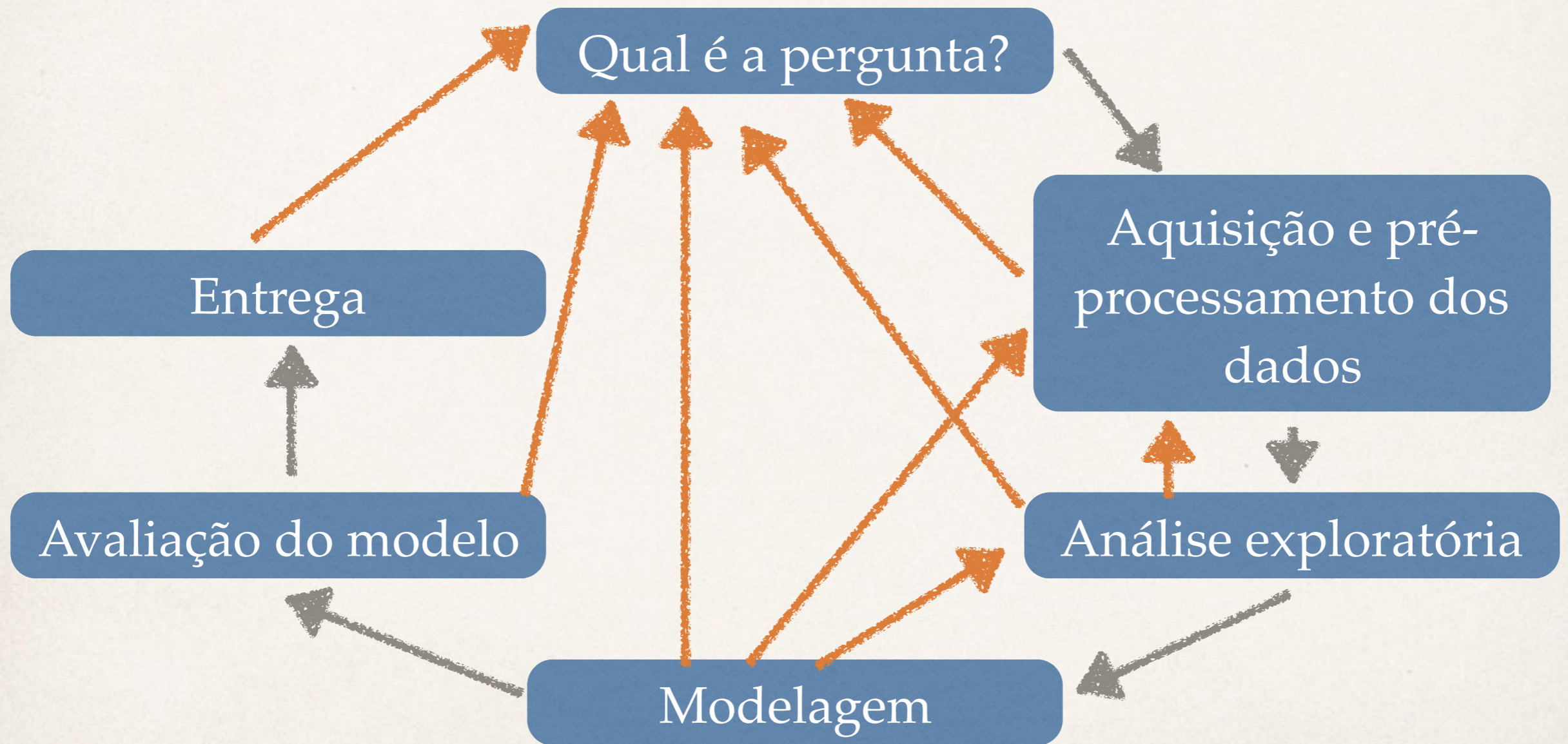
Relatórios Dinâmicos

Aplicativos

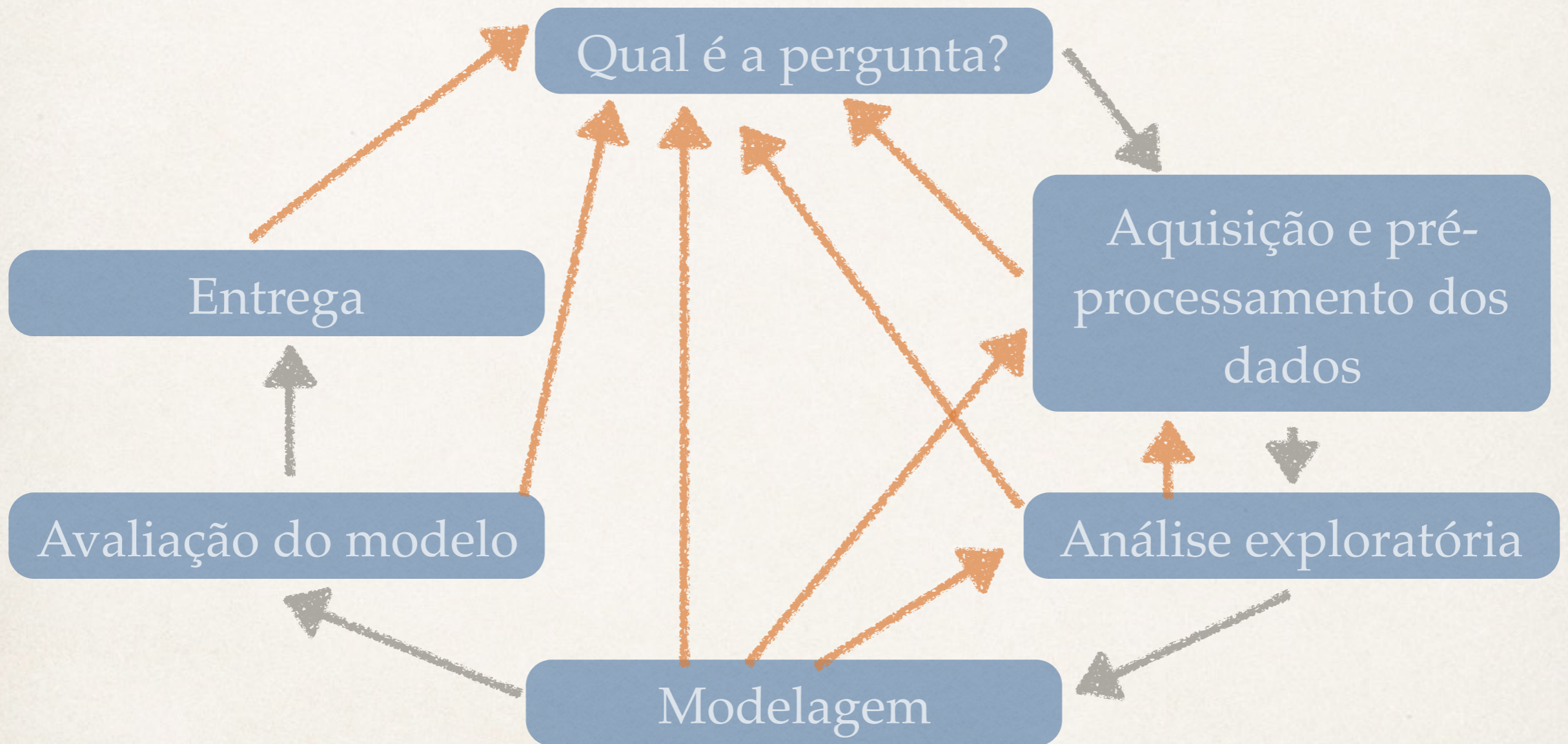
Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)



Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)



Processo de Descoberta de Conhecimento (KDD - Knowledge Discovery in Databases)



Este processo pode ser suportado por diversas ferramentas, entre elas: R, SPSS, RapidMiner, Tableau, Weka, Matlab, Octave, Python, Julia,...

Leitura sugerida

- ❖ Capítulos 1, 2 e 3 do livro EMC Education Services, editor. Data Science and Big Data Analytics: Discovering, Analysing, Visualizing and Presenting Data. John Wiley & Sons, 2015.
- ❖ Demais materiais da disciplina estão em:
 - ❖ <http://fbarth.net.br/materiais/cursoMineracaoDados>