INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

# O uso do Web Scraping nas estatísticas oficiais

Departamento de Recolha de Informação
Maria José Fernandes
Porto, Abril de 2017

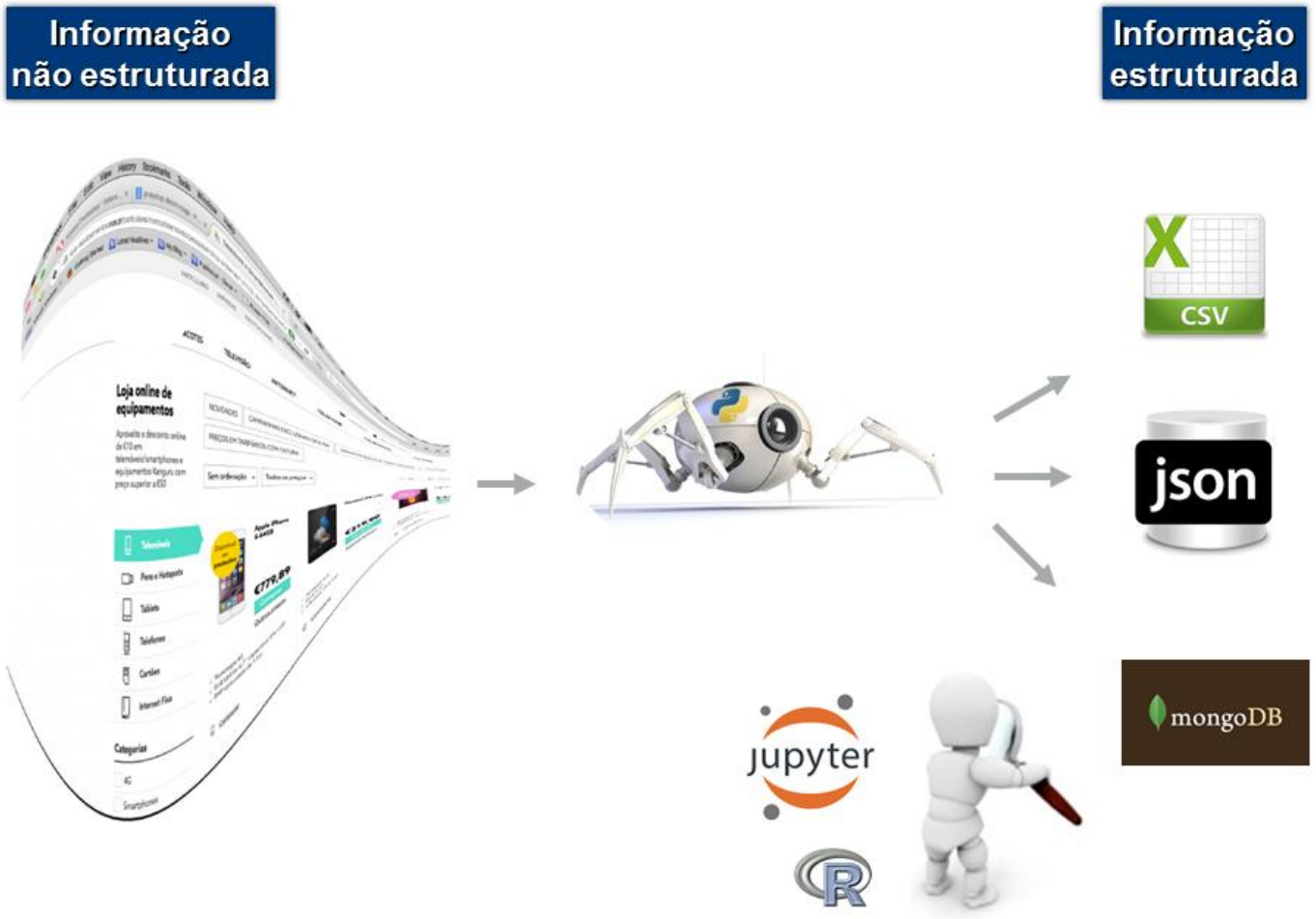
# Web Scraping

- O que é?
- Como funciona?
- Porquê e para quê?
- Como se faz?
- O que é necessário saber?
- Ciclo Web Scraping
- Aplicações em produção

# O que é?

- Web scraping é uma técnica computacional para **extração de informação de páginas web**
- **Automatização da web** (simulação da interação humana com a web quando utilizamos um browser)
- Transforma **informação não estruturada** em **informação estruturada** que pode depois ser armazenada e analisada

# Como funciona?



# Como funciona?

## Arquitetura do Web Scraping


Programas/Scripts usados para fazer web scraping são conhecidos por **web scrapers**, **web spiders**, **web crawlers**, **web robots**, **bots**, etc.

# Porquê?

- Na era da informação a capacidade de automatizar a recolha de informação é crucial
- A recolha manual de informação na internet é enfadonha, lenta e falível
- Possibilita uma redução substancial dos custos da recolha (com a possibilidade de aumentar as amostras observadas)

# Para quê?

- Recolha de preços no comércio eletrónico (comparar, agregadores de preços, etc…)
- Mercados financeiros
- Sector imobiliário (agregadores de imóveis)
- Análise de sentimentos (monitorizar a reação dos utilizadores)
- Media (agregadores de notícias, pesquisa jornalística)
- Atualização de universos/recolha de contactos
- Monitorização meteorológica
- …
- archive.org - Internet Archive - evolução da net (captura dos sites em diferentes momentos)

# Para quê?

## Recolha de preços no comércio eletrónico

- **Produtos de alto valor comercial** (app, sites agregadores de preços - aviação, supermercados, hotéis, etc.)
- **Índice de Preços no Consumidor**
  - Experiências em Países Europeus - Eurostat tem financiado projetos piloto
  - BPP (The Billion Prices Project - MIT - Prof. Alberto Cavallo)
    - 2008 - índices diários de inflação para a Argentina
    - 2010 - índices diários para os EUA
    - Desde 2011 - índices diários para 20 países -> PriceStats "**the leading provider of global inflation statistics**" -> distribuídos pelo State Street Bank para o setor financeiro

# Como se faz?

Nível de conhecimentos de programação

Nível de controlo sobre o processo

Mínimo

Mínimo

- usando ferramentas tipo “point and click” – fácil de usar mas não muito eficiente em sites mais complexos (ex. sites com javascript, etc..)
  - plataformas web - Import.io, dexi.io, octoparse, etc.,
  - browser plugins - Web scraper, Data Miner, Sraper, etc.,
  - softwares dedicados – imacros

Máximo

Máximo

- escrevendo o nosso próprio código (construindo um “Web Scraper”)

# Como se faz no INE

**Escolha das opções tecnológicas para**

**extrair ( + armazenar + analisar) a informação**

- Total controlo sobre o processo **construir um "Web Scraper"**
- Utilizar ferramentas **"free and open-source"**
- Aproveitar a oportunidade para usar e partilhar novas linguagens e novos processos - **utilizar o que nunca tinha sido utilizado**

# O que é necessário saber?

- Linguagem de programação (**Python**, Java, Ruby, Javascript, R, .NET, Perl, PHP)

- Conhecer muito bem a web e estar atento às mudanças
  - **Protocolo HTTP** - linguagem de comunicação na web
  - **HTML** - "linguagem base da internet" - linguagem de marcação para estruturar páginas web
  - **CSS** - linguagem para definir o estilo do documento html (para formatar conteúdos estruturados)
  - **Javascript** – linguagem de programação para tornar as páginas web dinâmicas e interativas
  - **XPATH** - linguagem de consulta para selecionar nós num documento XML)
  - **Expressões regulares** (regex) - define um padrão a ser usado para procurar palavras ou grupos de palavras - "forma muito precisa de fazer buscas de determinada porção de texto"

# O que é necessário saber?

## Python

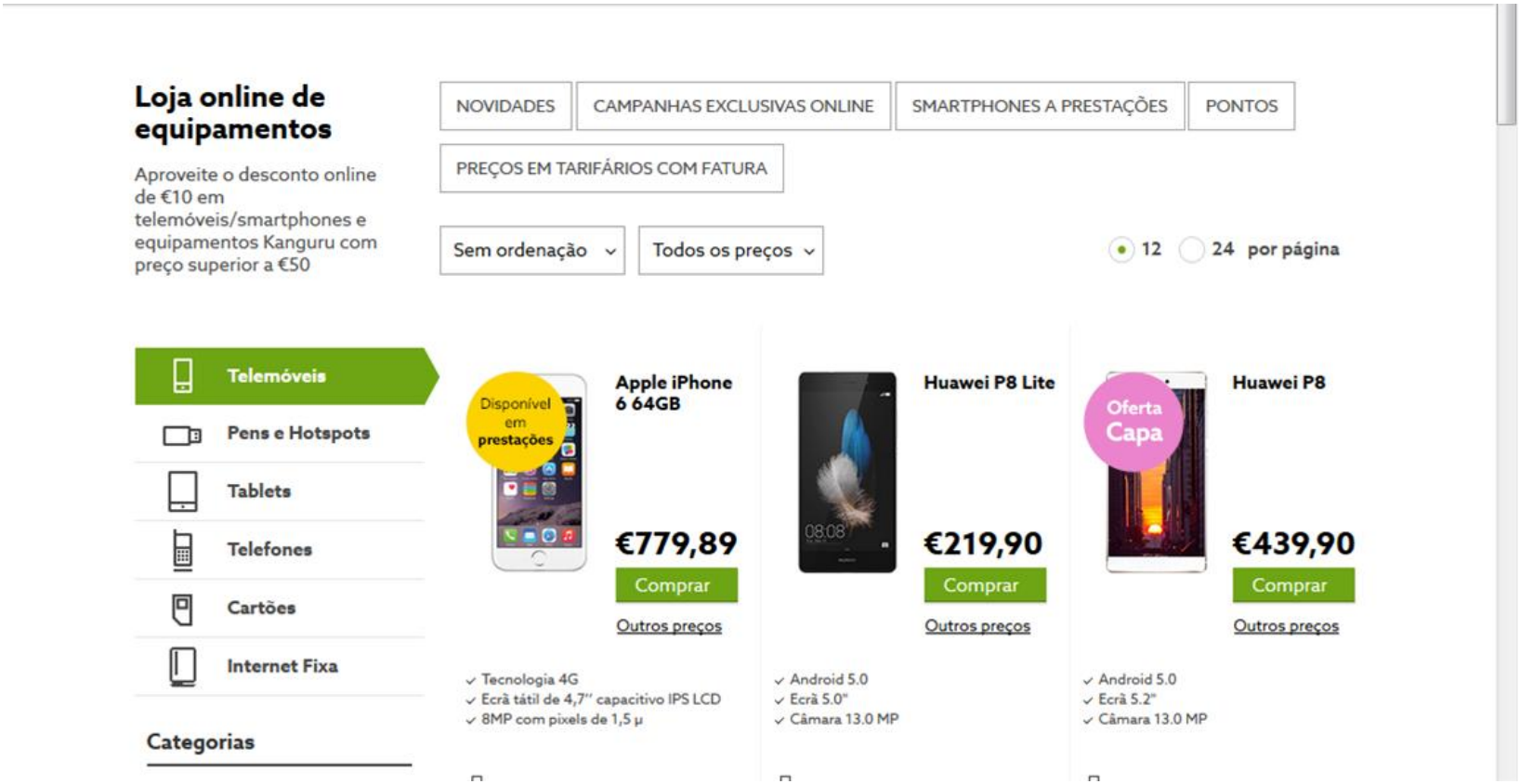
- Linguagem orientada a objetos - robusta, produtiva, fácil aprendizagem, intuitiva, web, muito bem documentada
- Lançada em 1991 por Guido Van Rossum possui atualmente um modelo de desenvolvimento comunitário, aberto e gerido pela organização sem fins lucrativos Python Software Foundation
- “A arma secreta da Google”
- No top 5 das linguagens mais utilizadas
- Muitas frameworks, livrarias, módulos - “one stop shop - from web app to data analysis”

# Ciclo web scraping

- **Estudar** a página web da qual queremos extrair informação (estudar a estrutura, identificar a informação que nos interessa, se é necessária navegação, se é necessária mais do que uma extração, etc.)
- "**Descobrir** o caminho para a informação" usando Xpath, CSS selectors, Regex, etc.
- **Escrever/testar** o script
- **Extrair** e **limpar** a informação
- **Exportar/armazenar/analisar** a informação

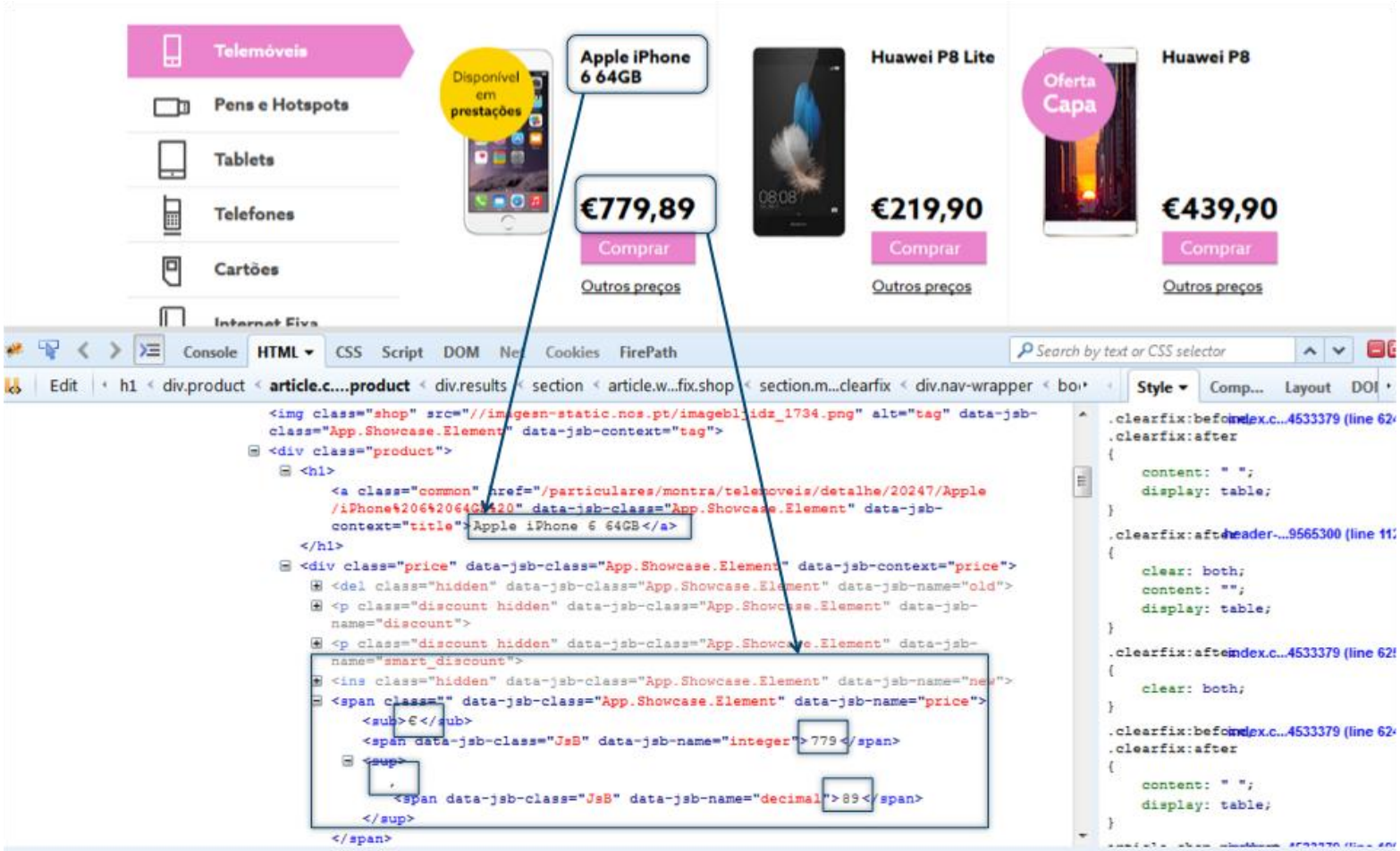
# Ciclo web scraping

## Estudar a página web



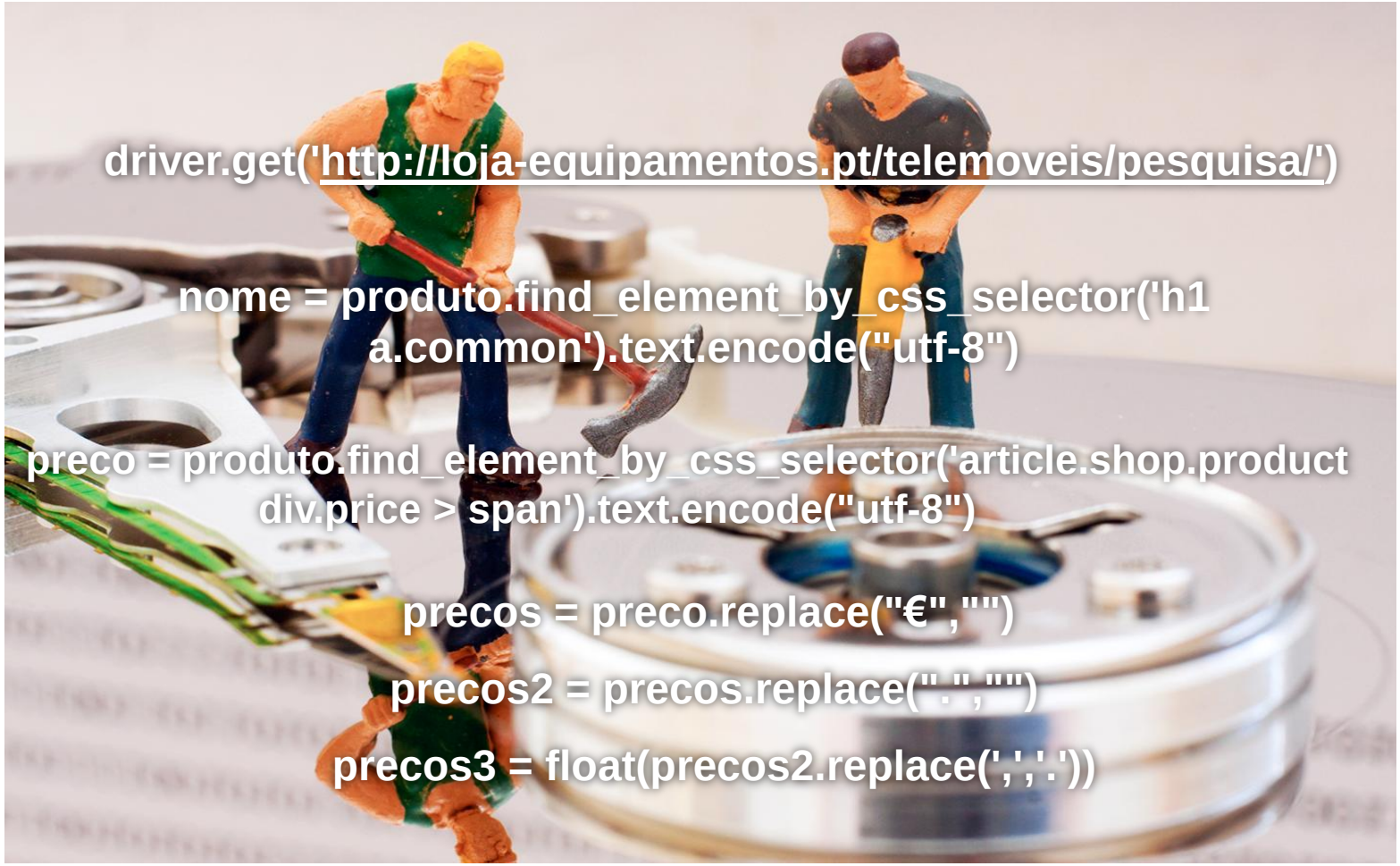
# Ciclo web scraping

**“Descobrir o caminho” para a informação**



# Ciclo web scraping

## Escrever/extrair/limpar



# Ciclo web scraping

**Armazenar/Processar/Analisar/Visualizar**

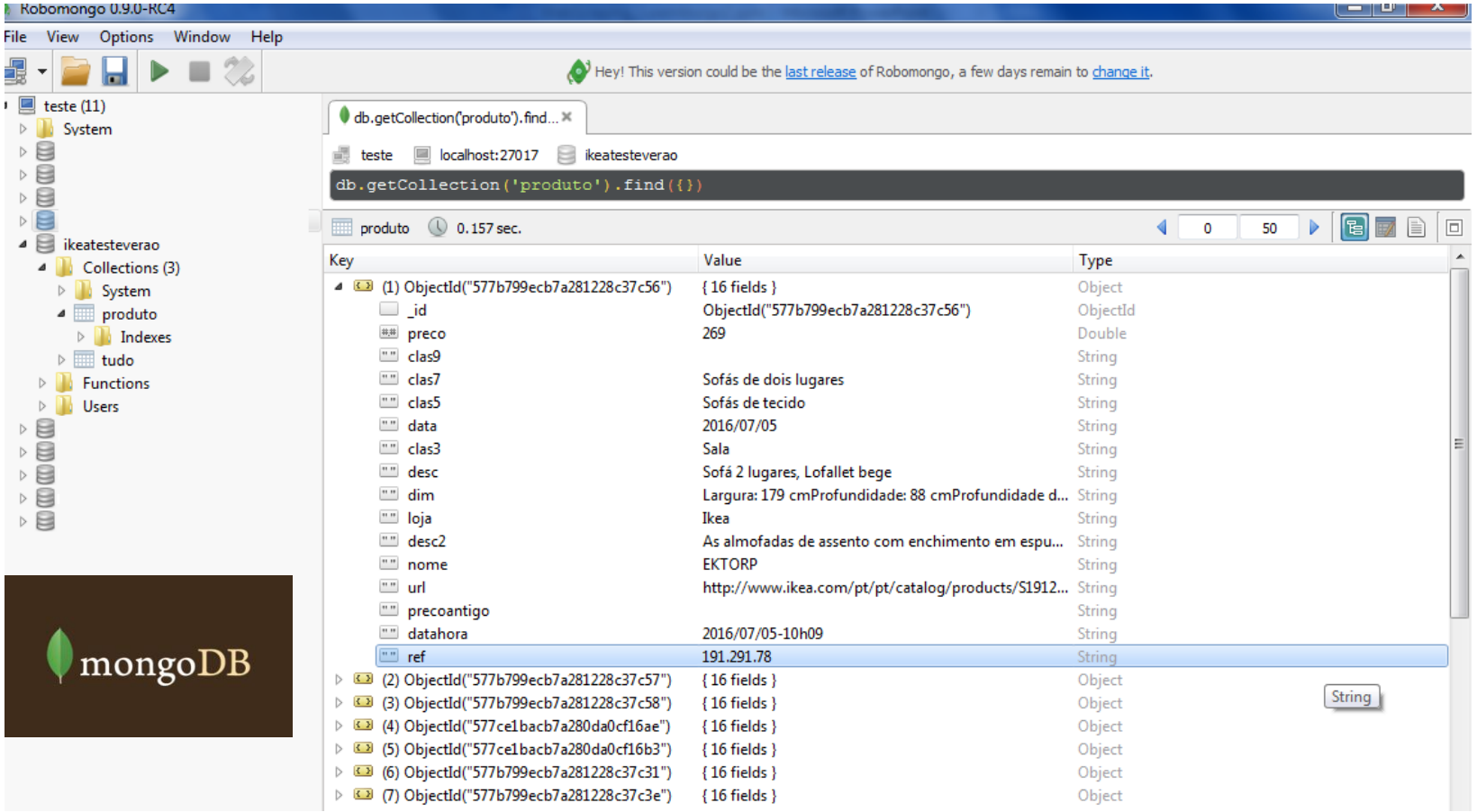
- **Armazenamento – MongoDB** (Base de dados não relacional - NoSQL)
  - orientada a documentos, alta performance, sem esquema
  - suporta Map Reduce e Sharding
  - documentos bson - ficheiros simples, rápidos e leves, muito utilizados para a troca de informação na web - Google, Facebook, Twitter, etc..

- **Processamento, Análise, Visualização – Jupyter Notebook** (Ambiente computacional interativo)
  - combina execução de código, texto explicativo, visualização gráfica, imagem e video;
  - os inputs e outputs podem ser guardados e distribuídos no formato de notebook (.ipynb);
  - exporta em vários formatos (html, latex, markdown, slides, etc..)

**Tudo novo (no INE), tudo "free and open source"**

# Ciclo web scraping

## Armazenar a informação (MongoDB + RoboMongo)



# Ciclo web scraping

## Processar/Analisar/Visualizar

jupyter IKEA_Classification_Analysis (unsaved changes)

File Edit View Insert Cell Kernel Help Python 2

Code CellToolbar

In [ ]:

In [60]:
```
#GEOMETRIC MEAN per CPI classe
mediageofinalbd = juntacorrente.groupby(['ano','mes','data','classifica', 'coicop'], as_index=False).ag
mediageofinalbd.preco = mediageofinalbd.preco.round(decimals=2)
mediageofinalbd
```

Out[60]:

| | ano | mes | data | classifica | coicop | preco |
|---|---|---|---|---|---|---|
| 0 | 2016 | 08 | 2016/08/16 | armários cozinha | 051111301 | 182.94 |
| 1 | 2016 | 08 | 2016/08/16 | cama de casal | 051111201 | 237.75 |
| 2 | 2016 | 08 | 2016/08/16 | cama de jovem | 051111202 | 157.85 |
| 3 | 2016 | 08 | 2016/08/16 | colchão | 051111203 | 148.13 |
| 4 | 2016 | 08 | 2016/08/16 | estante | 051111401 | 71.34 |
| 5 | 2016 | 08 | 2016/08/16 | mesa cozinha | 051111302 | 131.61 |
| 6 | 2016 | 08 | 2016/08/16 | mesa decoração | 051111402 | 82.55 |
| 7 | 2016 | 08 | 2016/08/16 | mesas e cadeiras | 051111101 | 385.40 |
| 8 | 2016 | 08 | 2016/08/16 | roupeiro/cómoda | 051111204 | 145.04 |
| 9 | 2016 | 08 | 2016/08/16 | sofá pele | 051111403 | 490.86 |
| 10 | 2016 | 08 | 2016/08/16 | sofá tecido | 051111404 | 259.50 |

jupyter

IP[y]: IPython

# Aplicações em produção

## Índice de Preços no Consumidor (IPC) - site IKEA

- Substituição da recolha manual online – 110 produtos IPC (10 categorias da nomenclatura IPC)
- Extração e processamento de todos os produtos das mesmas categorias (aprox. 1000 produtos) – em avaliação a incorporação no Índice

## Inquérito aos Transportes Rodoviários de Mercadorias (ITRM) - site IMT

Atualização da amostra – recolha das matrículas canceladas no site do Instituto da Mobilidade e dos Transportes

# Desafios futuros

- Alargar a recolha a mais sites (produtos/serviços)

- Estudar o impacto no IPC da substituição da recolha tradicional (para algumas categorias da nomenclatura)

- Desenvolver um modelo de classificação dos produtos com base nos descritivos recolhidos (machine learning supervisionado)

# E por falar em desafios…

# MOOC & CIA

**Massive Open Online Courses**










INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

# O uso do Web Scraping nas estatísticas oficiais

**Obrigada pela vossa atenção!**

mjose.fernandes@ine.pt