European Commission

Uses of web scraping for official statistics

ESTP course on Big Data Sources – Web, Social Media and Text Analytics, Day 1

Olav ten Bosch, Statistics Netherlands

THE CONTRACTOR IS ACTING UNDER A FRAMEWORK CONTRACT CONCLUDED WITH THE COMMISSION

Outline

- Introduction
- Web scraping and official statistics
- Examples at Statistics Netherlands
- How it works
- Challenges
- Legal
- Wrap up

Introduction (1)

- Web scraping is automatically retrieving (and processing) information from websites.
- Web scraping is as old as the internet itself. There would not be any search engine without web scraping.
- Web scrapers use the same internet techniques that browsers use to visit web sites.
- Web scrapers use the structure and contents of web pages to identify and scrape relevant information.

Introduction (2)

- There are many different flavours of web scraping, varying from
  - harvesting very diverse information from multiple sites to very dedicated scrapers focussed on particular items on a single site
  - fully automatic scrapers operating silently to programs that require user interaction (an assistant)
  - one-time scraping for a research project to scraping for use in production
- Web scraping has many names:
  - Crawlers, harvesting, spiders, bots, internet robots

Web scraping and official statistics

- Web sites contain detailed and frequently updated information that may be useful to official statistics.
- We can enrich the results of the traditional methods with automatically collected information.
- Web data may enable new types of indicators that are not feasible with traditional methods.
- The collection and analysis time is much faster than performing these tasks manually.

Faster, better, more efficient

Administrative sources

- Tax, social security
- Municipalities/ Provinces
- Supermarkets
- …

Internet sources

- …
- Surveys **Less!!!**

New indicators

New

14-06-2013	Exports shrinks
14-06-2013	Calendar
13-06-2013	Retail turnover 0.6 percent lower
13-06-2013	Large drop in turnover for car and motorcycle trade
10-06-2013	More than half of employees commute to work

Competitors?

PriceStats®

HOME | APPROACH | INDICES | NEWS | ABOUT US | CONTACT US

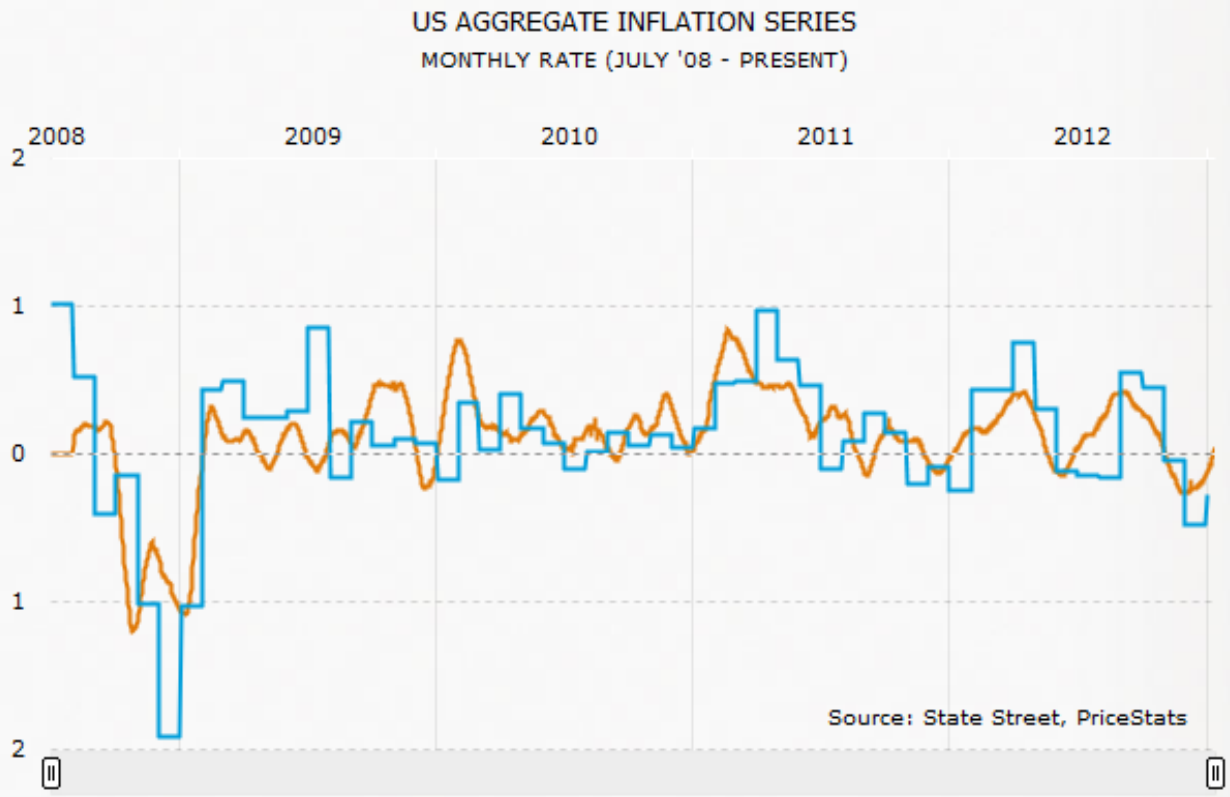
Overview | US Series | Argentina Series | Social Responsibility

US INFLATION SERIES

PriceStats estimates aggregate inflation in the US using online prices. The objective of this series is to anticipate major changes in US inflation trends, but not to forecast monthly CPI announcements. At any point in time, our index can be different from the CPI. Our data anticipates changes in inflation trends not only because we observe prices sooner, but also because online prices tend to react to shocks more quickly.

- Official CPI
- PriceStats Index

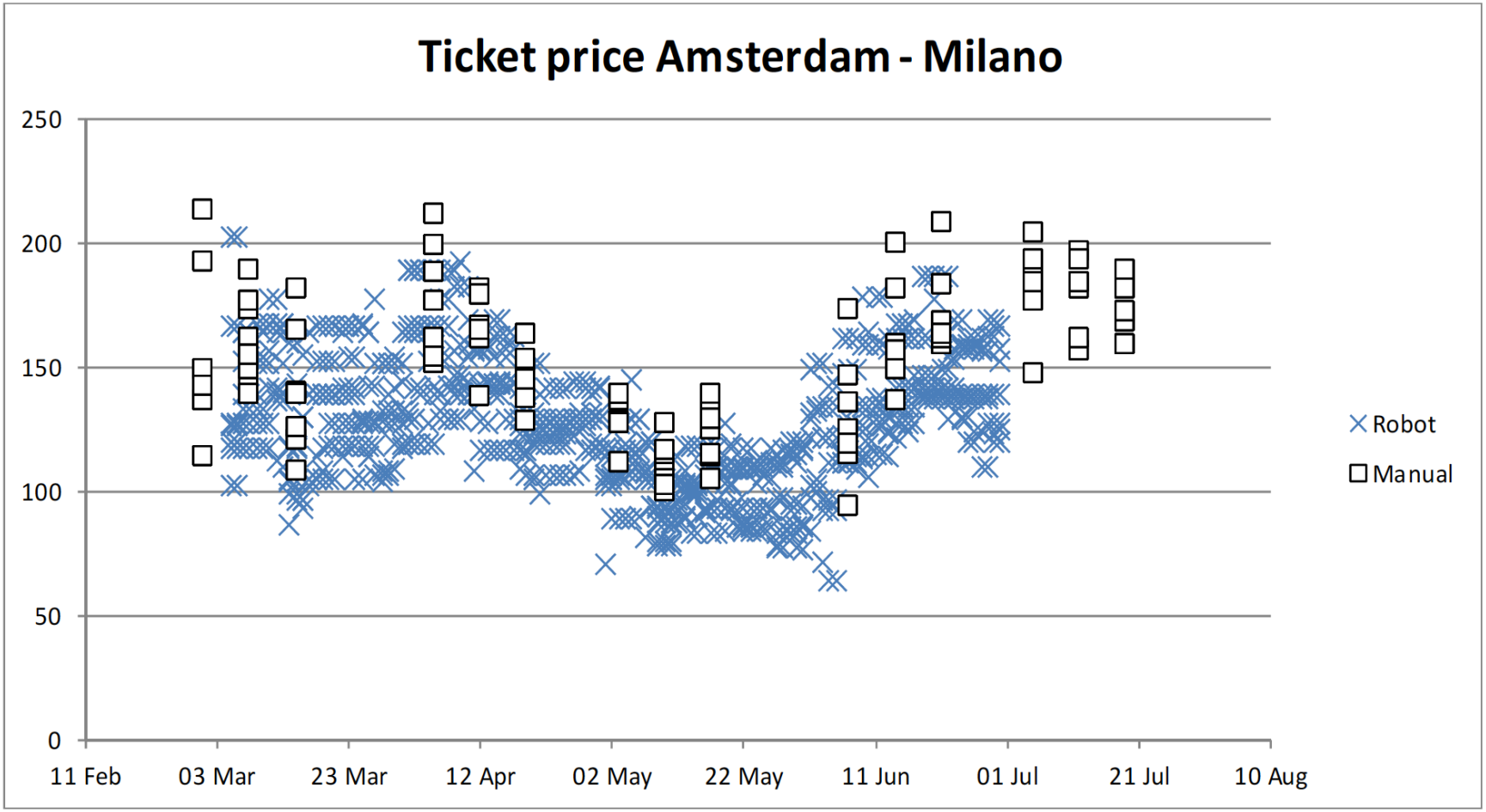
Visit State Street Research Portal to download the data.



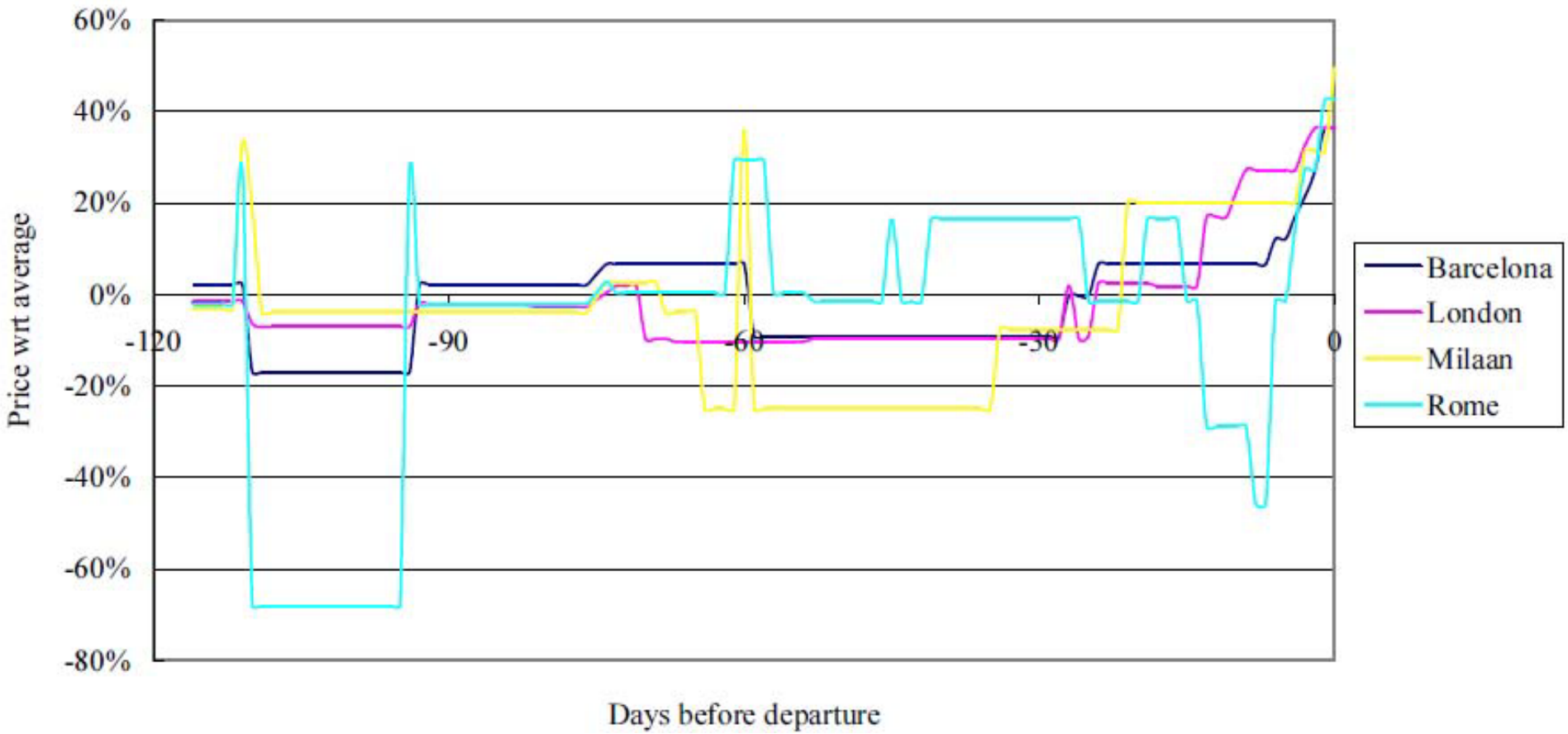
Some use cases

Internet sources

Prices of airline tickets
Real estate sites for housing statistics
Internet vacancies for job statistics
Social media sentiment for consumer confidence
Trade in second-hand goods as economic indicators
Use wikipedia for improving the business register
Internet prices for CPI



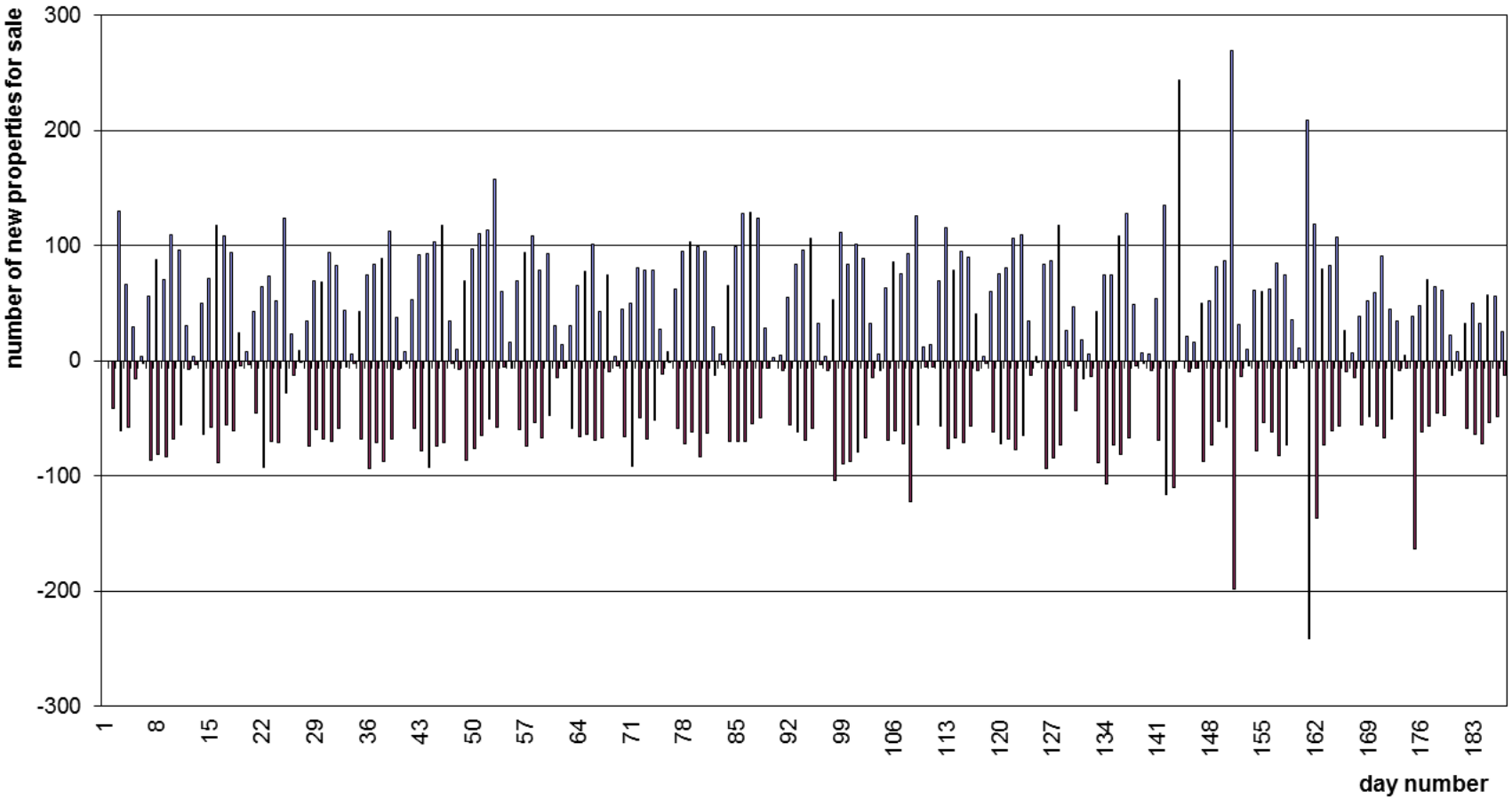
Ticket prices according to robot and manual collection from web sites



New Analysis: Volatility of flight prices of four destinations starting 116 days before departure.

Tracking housing market (2011)

- Identifying the source web sites and understanding the characteristics of the data they provide was a challenge.
- We scraped 5 sites, about 250.000 observations / week, for 2 years
- Dutch property market statistics use results of the research using these robots.
- After research phase CBS obtained data directly



Volatility of the content of one of the housing websites.
Positive bars are properties added, negative are properties deleted.

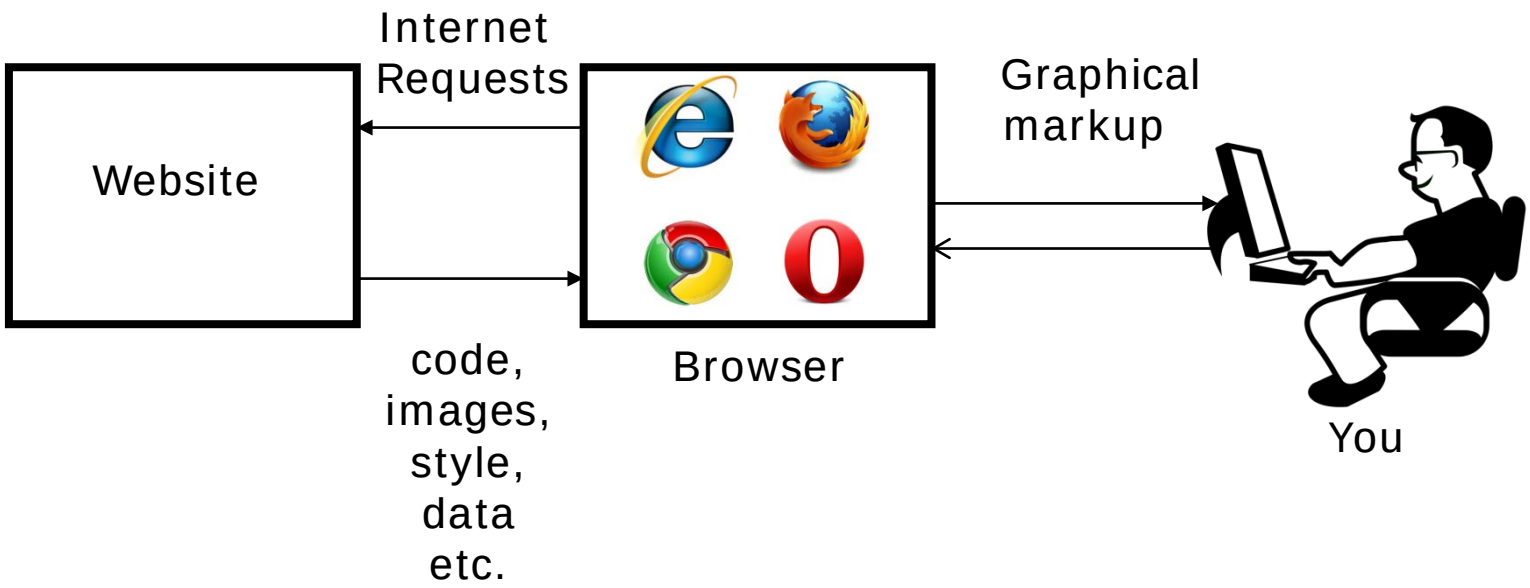
Bulk price collection for CPI (2012->)

- Mainly clothing
- Software scrapes *all* prices and product data (id, name, description, category, colour, size,...)

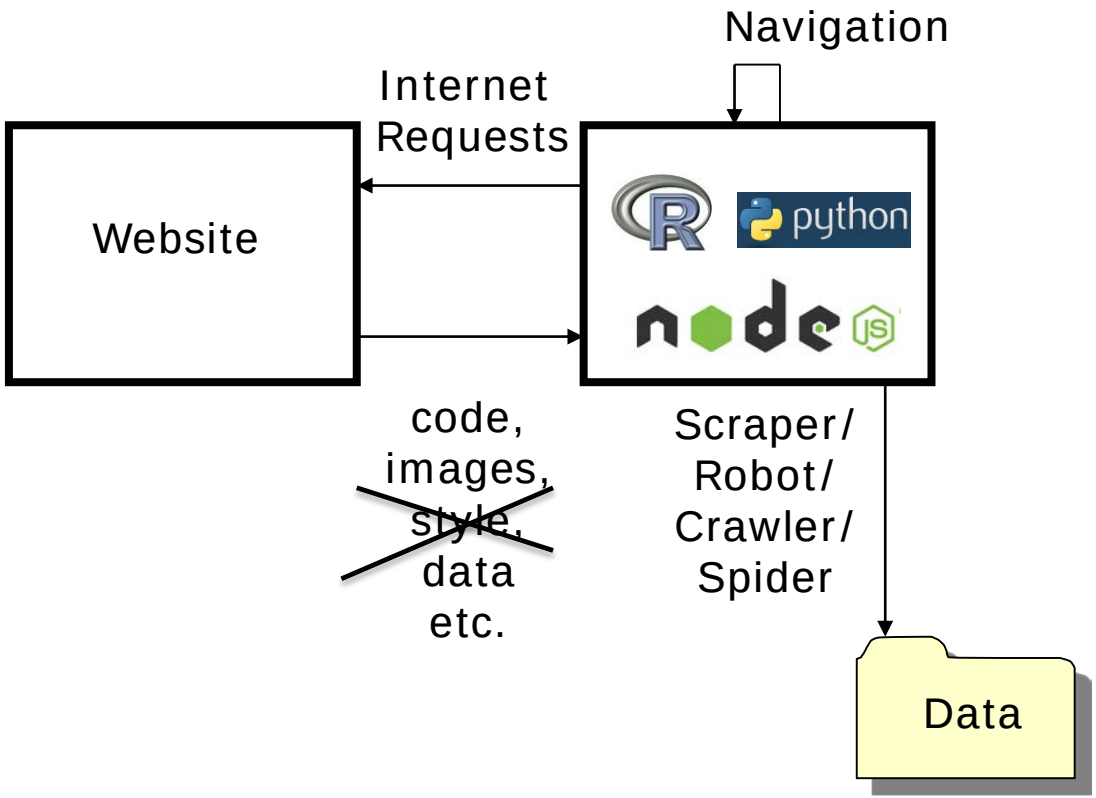
2016:

- About 500.000 price observations daily from 10 sites
- Data from 3 sites used in production of Dutch CPI
- Price collection process embedded in organisation
- Plans to extend to > 20 sites and other domains

How it works (1)

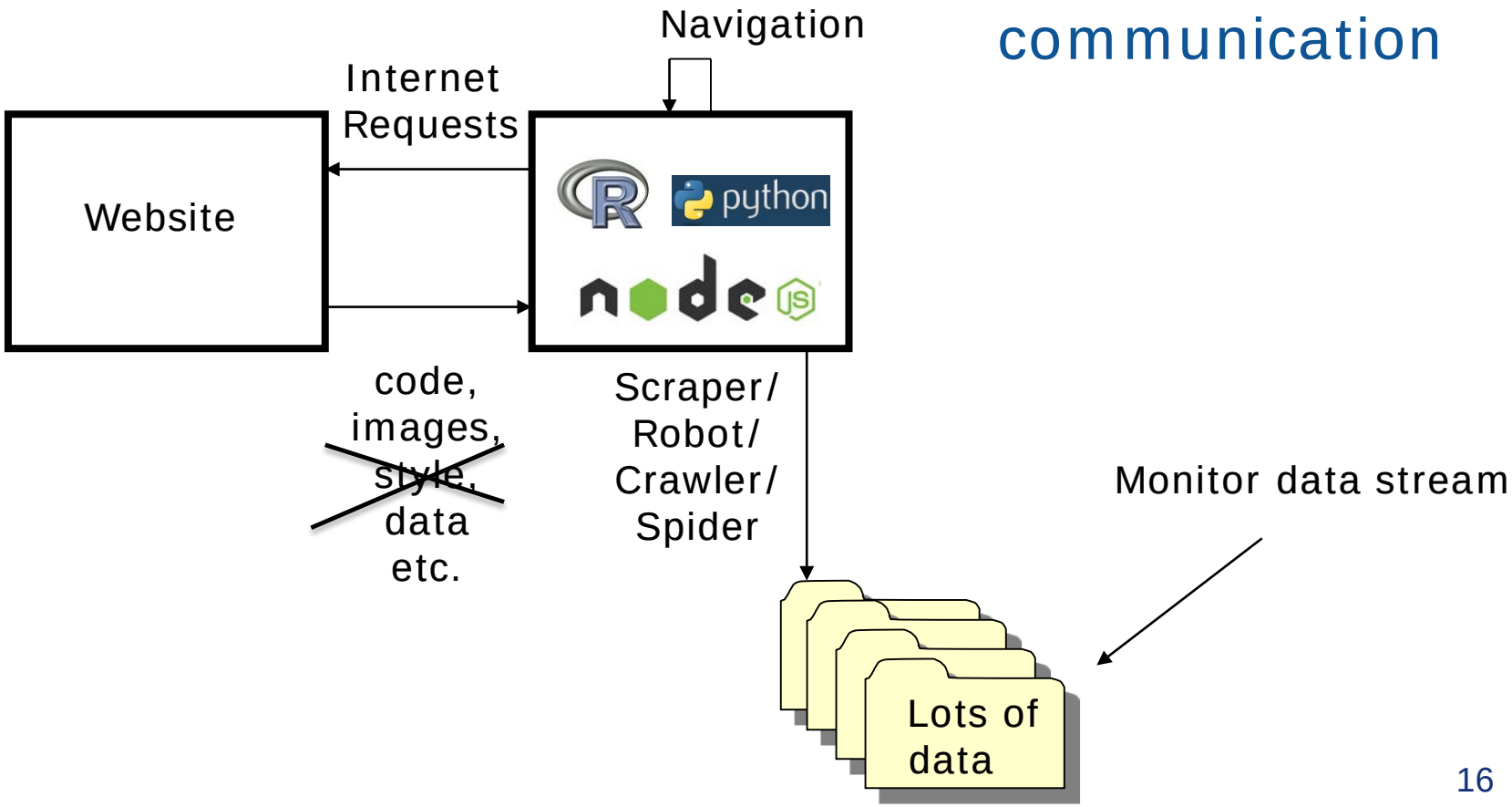


How it works (2)



How it works (3)

Live demo of website communication



Challenges in web scraping (1)

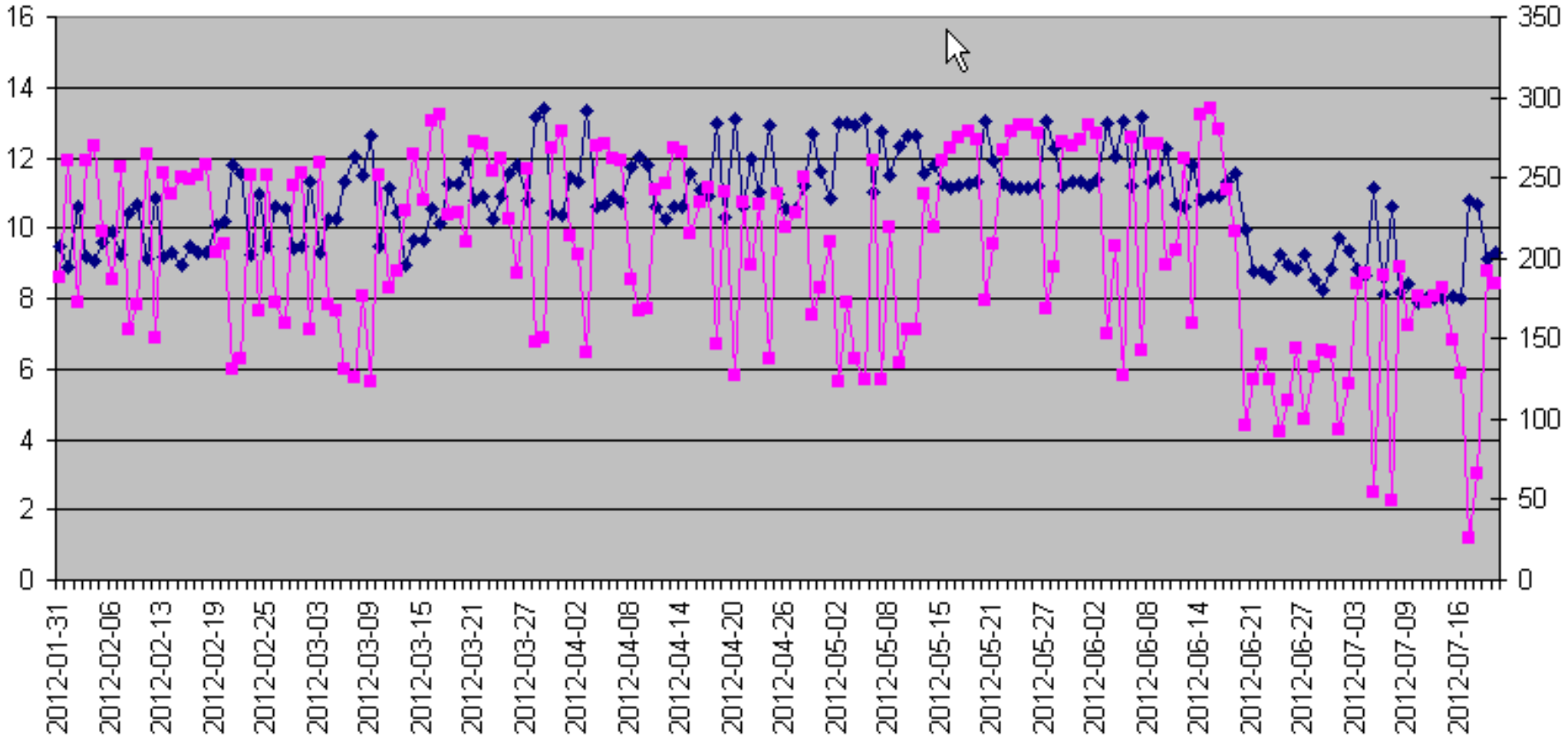
- Which data?
  - **It is not always easy to know which site to scrape**
  - **It starts with detecting the most up-to-date and complete sources**
  - **The internet is full of copies of data**
  - **Which data is relevant, up to date, reliable?**
  - **Sometimes choose between scraping the data owner or an aggregator site**
  - **Can we get the data from the data owner directly, without scraping?**
  - **Advice: explore the data flows among sites**

Challenges in web scraping (2)

- The internet is dynamic
  - **Each web site has a particular structure, which may be subject to changes anytime**
  - **The technology used on websites changes continuously (example: infinite scroll)**
  - **Sometimes websites do not comply with standards**
  - **Advise: try to build scrapers as robust as possible**
- Data is volatile
  - **Be aware of changing data patterns over time**
  - **Advise: monitor data frequently**

Example of volatile data

Number of products per product category (right axis, red line) and the average price based on these products (left axis, blue line)



Challenges in web scraping (3)

- Legal issues and informing web site owners
- When using internet data in production, how to organise your scraping process?
- Advice: Manage the organisation of the collection process and transform collected data into a standardized format.

Legal (1)

- Legislation for scraping may be country specific
  - **Below is inspired by the Dutch situation**
- From National Statistics Law:
  - **Enterprises have to provide data to the NSI on request.**
  - **Scraping may reduce response burden**
- Database legislation:
  - **Commercial re-use of scraped intellectual property from internet sources is forbidden**
  - **NSI's usually uses data for official statistics only**

Legal (2)

- Privacy:
  - **Legislation on protection of personal information**
  - **At this moment we only scrape public sources and process data within NSI's safe environment, just as with other (privacy-sensitive) data internally**
- Netiquette (practical):
  - **respect the Robots Exclusion Protocol also known as the robots.txt (example)**
  - **identify yourself (user-agent)**
  - **do not overload servers, use some idle time between requests, run crawlers at night / morning**
  - **Inform website owners if feasible**

Wrap up

- We have explained the basics of web scraping, retrieving information from the internet
- Web scraping has many different flavours, we have seen a few
- Web scraping is useful for official statistics in different ways, not only as primary source
- There are still challenges in scraping, data processing and also legal issues

European Commission

Thanks for listening!

Any comment or question?