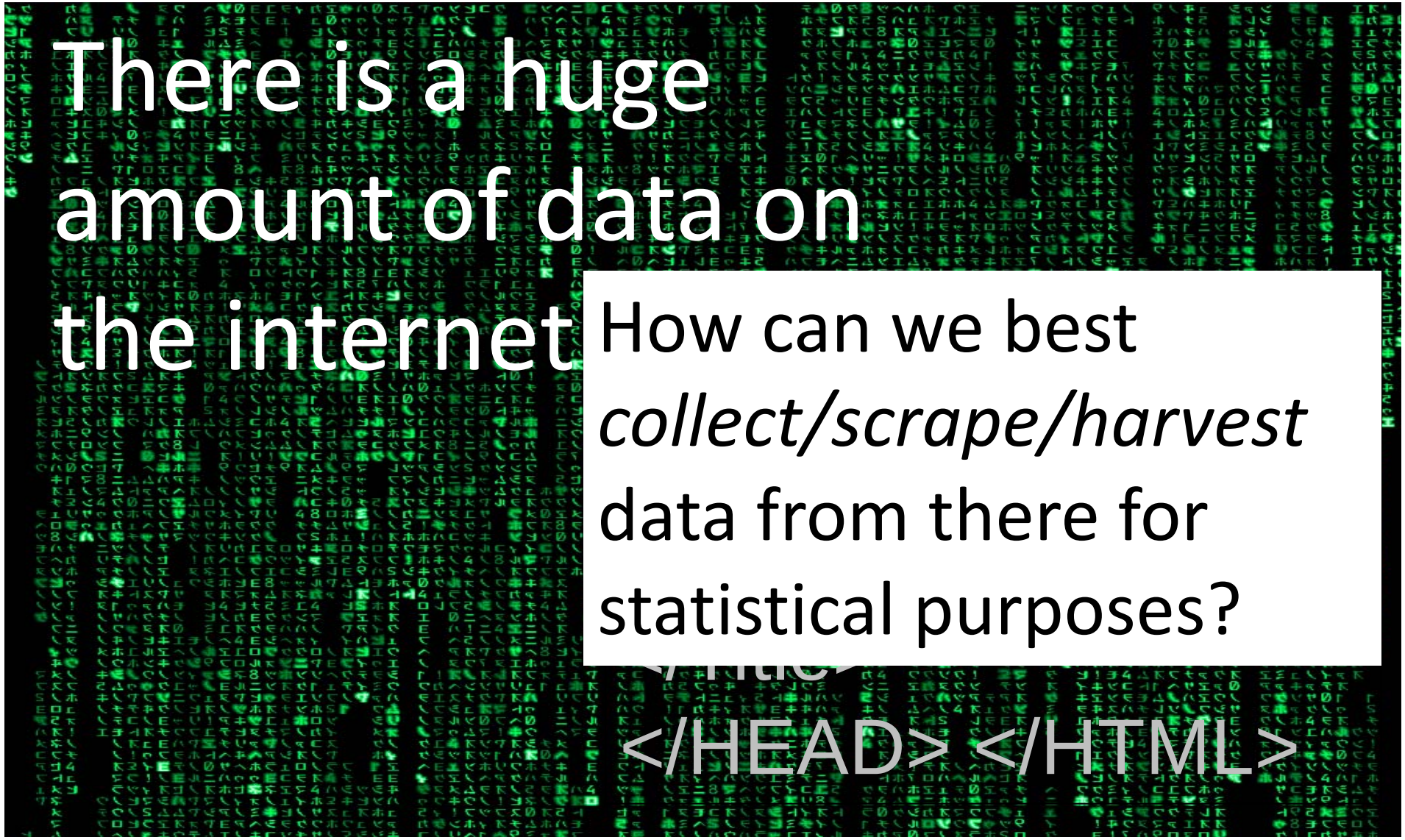
Ingolf Boettcher

Tokio
20. May 2015

Automatic price collection on the internet (web scraping)

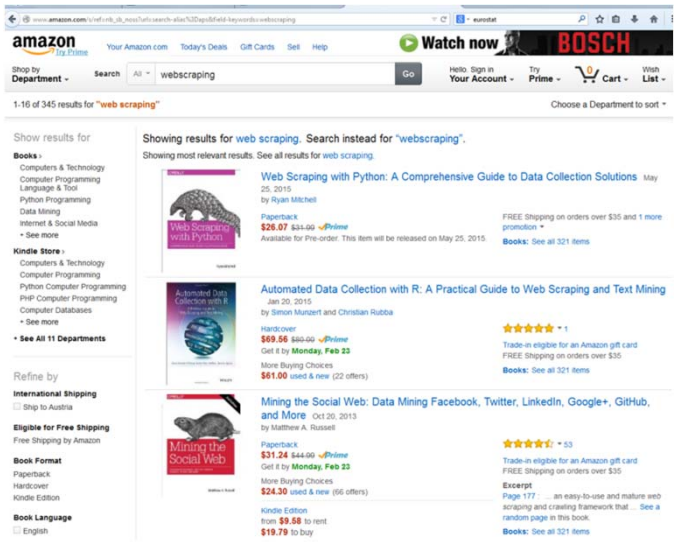
Ottawa Group 2015 –
Topic 1 Alternate data sources and Index number formulas
Session 2 Online Prices and Web Scraping

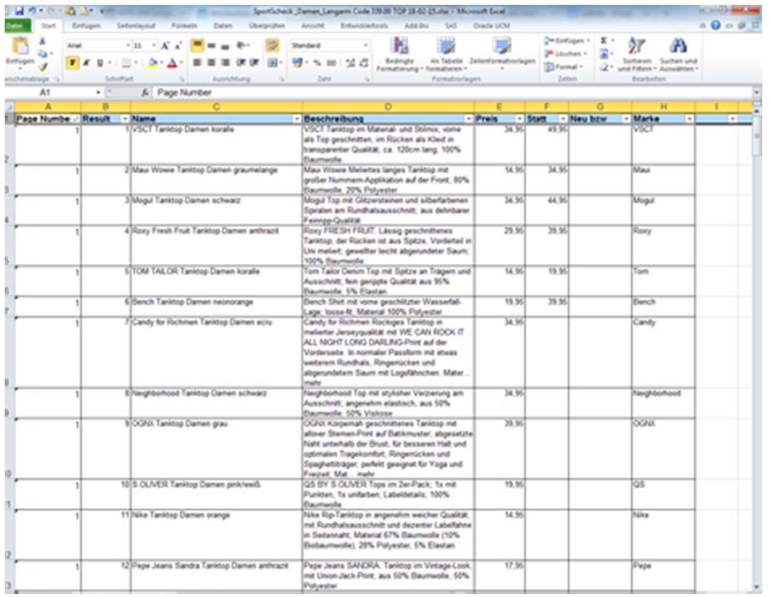
Web scraping



Internet data collection – Minimum goal for (Price) Statistics:

Turn website content into a spreadsheet





Internet data collection

Options:

1. Manual price collection
2. Develop an API /Web scraper
   - 2.1 by writing custom computer code
   - 2.2 by using point and click web tools

Reasons for not writing an own web scraper

```
/* schleife für alle gewünschten jahre */
   %do j=&vonjahr %to &bisjahr %by 1;

       /* ermitteln für welche monate die schleife laufen soll */
       %if &j=&vonjahr %then %let startmonat=&vonmonat;
                       %else %let startmonat=01;
       %if &j=&bisjahr %then %let endmonat=&bismonat;
                       %else %let endmonat=12;

       /* schleife für alle gewünschten monate innerhalb eines jahres */
       %do m=&startmonat %to &endmonat %by 1;
          %if %length(&m)=1 %then %let mm=0&m;
                            %else %let mm=&m;

/* Monat ct vergleich */

proc sql;
create table testct as
select c.agg,c.mzger,h.mzger as mzgerhpi
from vpi.tmatrixct c, vpi.tmatrixhpi h
where  c.basis='201212' and c.jahr='2012' and c.monat='05' and c.aggebene in ('1','0') and c.agg=h.agg an
h.basis='201112' and h.jahr='2012' and h.monat='05' and h.aggebene in ('1','0')
order by c.agg
;
quit;

/* ganze Jahr ct gesamt */

proc sql;
create table testct as
```

IT-developer needed, therefore:

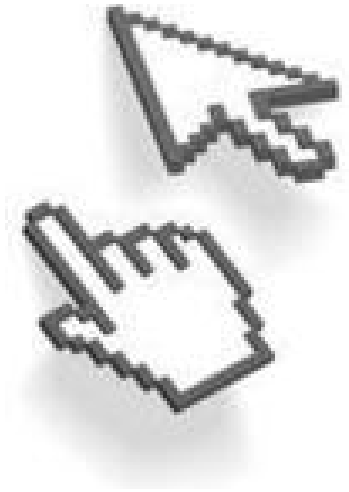
- Expensive
- Inflexible
- Even maintenance cannot be handled by CPI staff

Reasons to use click and point webtools for web scraping:

No IT-developer needed,

therefore:

- Cheap
- Flexible
- No programming skill required



How web scraping with click and point using *import.io* looks like:

- web-platform that allows to structure and extract data from websites

Web scraping with click and point on web-based platform offers solutions to:

- extract data by point-and-click
- record actions on a website
- crawl all the data of a webpage

More issues to be considered:

- Legality to crawl on websites
- Internal IT Security
- Training of staff

Automatic price collection on the internet (web scraping)

Contact:
Ingolf Boettcher

Guglgasse 13, 1110 Wien
Tel: +43 (1) 71128-7917
Fax: +43 (1) 7180718
Ingolf.boettcher@statistik.gv.at

	A	B	C	D	E	F	G	H
1	**Generated on February 20th 2015 at 2:46:51pm by ibot**							
2	**Input**	**Name**	**Description**	**Preis**				
3	webpage/url	Salomon SPEEDCROSS 3 Laufschuhe Damen grau/mint	Salomon SPEEDCROSS 3. Extrem leichter Damen-Trailrunningschuh mit präziser Passform und genialem Grip im Gelände. Salomons Kultschuh ist seit Jahren der	124,95 EUR				
4	webpage/url	Salomon SPEEDCROSS 3 Laufschuhe Damen mint/lila	Salomon SPEEDCROSS 3. Extrem leichter Damen-Trailrunningschuh mit präziser Passform und genialem Grip im Gelände. Salomons Kultschuh ist seit Jahren der	124,95 EUR				
5	webpage/url	Brooks Ravenna 6 Laufschuhe Damen pink/blau	Brooks RAVENNA 6. Der innovative Damen-Laufschuh bewegt sich zwischen Neutral- und Supportschuh und bietet eine ausgewogene Mischung zwischen Dämpfung und	134,95 EUR				
6	webpage/url	Salomon SPEEDCROSS 3 Laufschuhe Damen hellgrün/rot	Salomon SPEEDCROSS 3. Extrem leichter Damen-Trailrunningschuh mit präziser Passform und genialem Grip im Gelände. Salomons Kultschuh ist seit Jahren der	124,95 EUR				
7	webpage/url	Salomon SPEEDCROSS 3 Laufschuhe Damen rot/gelb	Salomon SPEEDCROSS 3. Extrem leichter Damen-Trailrunningschuh mit präziser Passform und genialem Grip im Gelände. Salomons Kultschuh ist seit Jahren der	124,95 EUR				
8	webpage/url	Salomon SPEEDCROSS 3 Laufschuhe Damen blau/mint	Salomon SPEEDCROSS 3. Extrem leichter Damen-Trailrunningschuh mit präziser Passform und genialem Grip im Gelände. Salomons Kultschuh ist seit Jahren der	124,95 EUR				
9	webpage/url	adidas Supernova Sequence 6 W Laufschuhe Damen pink	adidas SUPERNOVA SEQUENCE 6 W. Hochfunktioneller Laufschuh mit nahtlosem Schaft aus atmungsaktivem, mehrlagigem Mesh und stabiliserenden Synthetik-	89,95 EUR				
10	webpage/url	Nike Free 5.0 Laufschuhe Damen grün/orange	Nike FREE 5.0+. Leichter Damen-Laufschuh mit Barfußlauffeeling. Nach Vorbild des Barfußlaufens garantiert dieser Schuh maximal natürliche Beweglichkeit bei	114,95 EUR				
11	webpage/url	Salomon SPEEDCROSS 3 GTX Laufschuhe Damen dunkelrot/mint	Salomon SPEEDCROSS 3 GTX. Leichter, leistungsfähiger Trailrunningschuh für Wettkampf und High Speed Training, dauerhaft wasserdicht und hoch atmungsaktiv	149,95 EUR				
12	webpage/url	Salomon SPEEDCROSS 3 GTX Laufschuhe Damen schwarz/lila	Salomon SPEEDCROSS 3 GTX. Leichter, leistungsfähiger Trailrunningschuh für Wettkampf und High Speed Training, dauerhaft wasserdicht und hoch atmungsaktiv	149,95 EUR				
13	webpage/url	adidas Kanadia 7 tr Laufschuhe Damen schwarz/apricot	adidas KANADIA 7 TR. Traillaufschuh für Damen, mit spezifischem Leisten für optimale Passform. SCHAFT: hoch atmungsaktives Sandwich-Mesh mit	79,95 EUR				
14	webpage/url	Nike AIR MAX 2015 Laufschuhe Damen schwarz/türkis/fuchsia	Nike AIR MAX 2015. Leichter Laufschuh mit nahtlosem Schaft aus Meshmaterial und strategisch plazierten, geschäumten Partien; gibt Halt und bietet hohen Komfort; die	189,95 EUR				
15	webpage/url	Brooks Launch2 Laufschuhe Damen neonpink	Brooks LAUNCH2. Ein höchst funktioneller und schneller Wettkampfschuh undein echter Spezialist für kurze, schnelle Distanzen. Trotz seines geringen Gewichts,	109,95 EUR				
16	webpage/url	ASICS GEL KAYANO 21 Laufschuhe Damen blau/neongelb/pink	ASICS GEL KAYANO 21. Laufschuh der Stabilitätskategorie; für Läuferinnen mit Überpronation. SCHAFT: hoch atmungsaktives Sandwich-Mesh verstärkt mit ECO-	179,95 EUR				
17	webpage/url	Nike Air Zoom Pegasus 31 Laufschuhe Damen pink	Nike AIR Zoom PEGASUS 31. Trainingsschuh für Läuferinnen mit neutralem Abrollverhalten begeistert mit optimaler Passform. SCHAFT: besteht aus	109,95 EUR				
18	webpage/url	Salomon SPEEDCROSS 3 GTX Laufschuhe Damen hellblau	Salomon SPEEDCROSS 3 GTX. Leichter, leistungsfähiger Trailrunningschuh für Wettkampf und High Speed Training, dauerhaft wasserdicht und hoch atmungsaktiv	149,95 EUR				
19	webpage/url	ASICS GEL NOOSA TRI 10 Laufschuhe Damen neongelb/bunt	ASICS GEL NOOSA TRI 10. Laufschuh in farbenfrohem Design. Speziell für den Triathlon und Ironman-Wettkämpfe entwickelt. SCHAFT: Der GEL NOOSA ist aus	139,95 EUR				
	webpage/url	Nike Free 5.0+ Flash W Laufschuhe Damen	Nike FREE 5.0 FLASH. Leichter Laufschuh für Damen. SCHAFT: Das Flywire-	89,95 EUR				