Web-crawling to gather data and online market surveillance – Nordic and Swedish experiences

Kasper Mogensen - Big2Great

(Troels Fjordbak Larsen) - Big2Great

Background

Why do we want to do online market surveillance?

- Market insights
- Market enforcement
- Policy evaluation

Problem 1

Does the online market cover the entire market (online + physical shop)?

Data and method

- Assumption: GfK sales data on model level represents the entire market
- Method: Compare and match GfK data to scraped data

	Clothes washers	Dishwashers	Refrigerator Only	Refrigerator -Freezer	TV
Dec-13	X	X	X	X	X
Mar-14					X
Apr-14	X	X	X	X	
Aug-14	X	X	X	X	X
Oct-14					X

Issues we needed to correct for

- Normalization of naming (naming depends on context e.g. color)
- GfK registration time (look for match in surrounding periods)

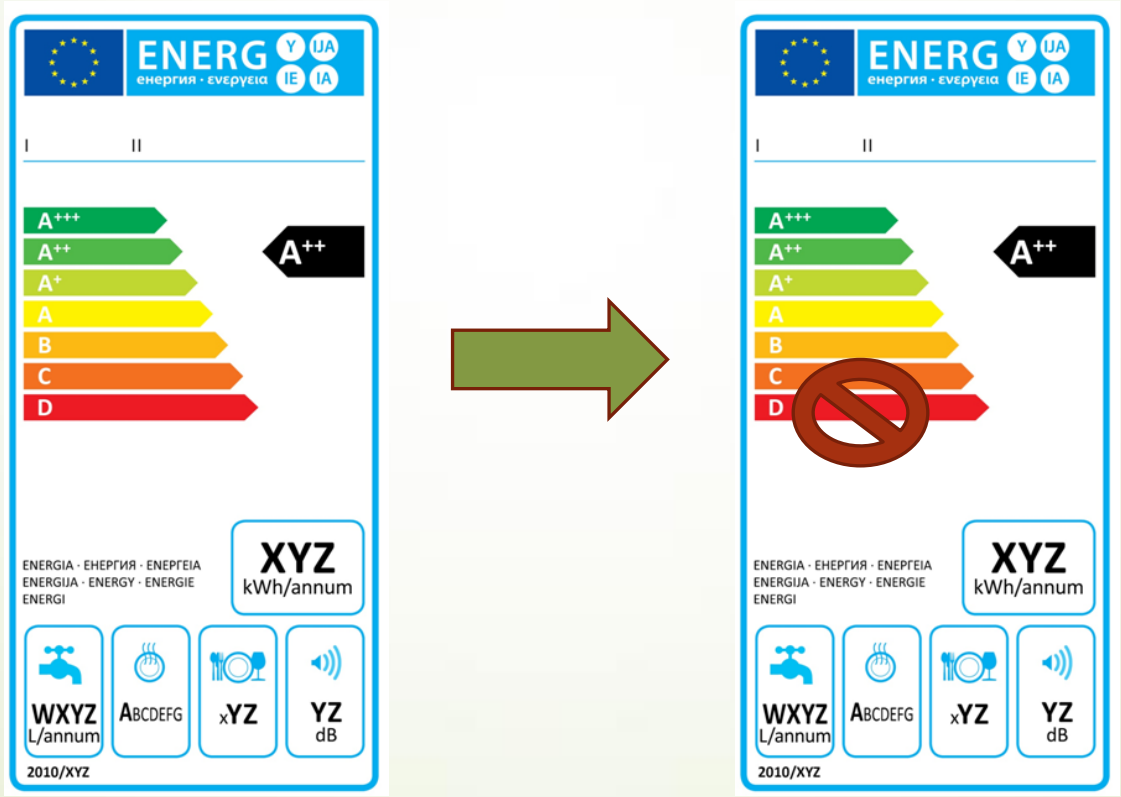
Problem 1 - Results

		December 2013	April 2014	August 2014
Clothes washers	Hit rate	82%	86%	84%
	Hit rate 90% MS	91%	93%	87%
Dishwashers	Hit rate	76%	79%	76%
	Hit rate 90% MS	86%	86%	91%
Refrigerator Only	Hit rate	85%	85%	82%
	Hit rate 90% MS	89%	94%	91%
Refrigerator-Freezer	Hit rate	76%	80%	79%
	Hit rate 90% MS	85%	88%	83%

		December 2013	Marts 2014	August 2014	October 2014
TV	Hit rate	64%	71%	62%	74%
	Hit rate 90% MS	94%	95%	91%	90%

Problem 2

Can scraped data be used to perform policy evaluation?



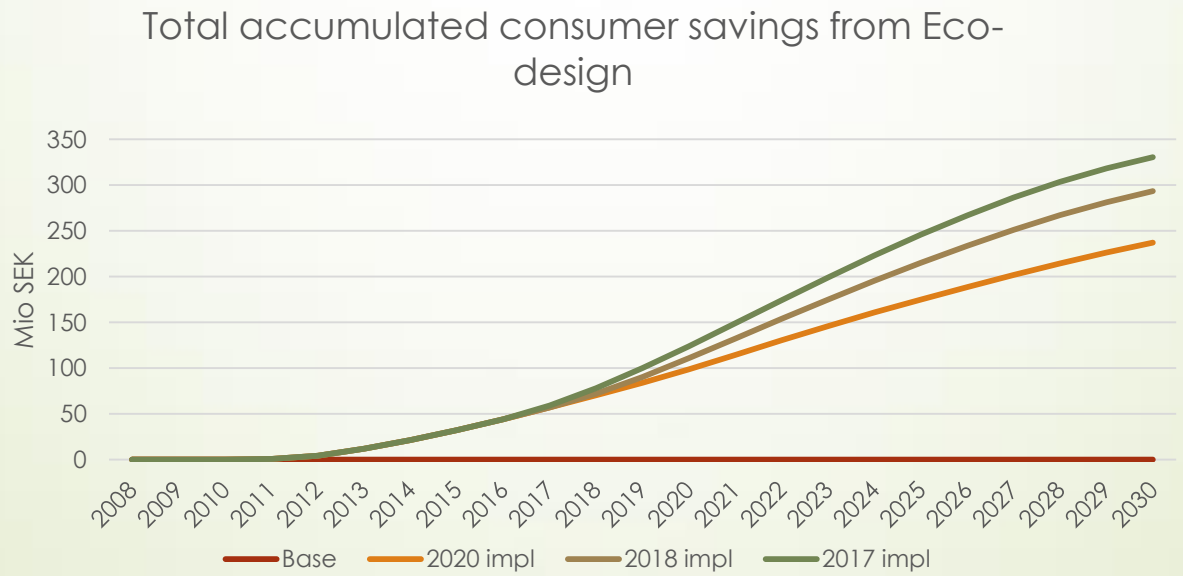
Data and method

- Stock-model based on sales data and energy class distribution
- Prolong historic period with GfK or Scraped data
- Compare the result

16	2013-09	2013-10	2013-11	2013-12	2014-01	2014-02	2014-03	2014-04	2014-05	2014-06	2014-07	2014-08	2014-09	2014-10	2014-11	2014-12	2015-01
17	25100	25100	25100	25100	25600	25600	25600	25600	25600	25600	25600	25600	25600	25600	25600	25600	26112
18																	
19																	
20				Crawl				Crawl				Crawl					
21																	
22	2013-09	2013-10	2013-11	2013-12	2014-01	2014-02	2014-03	2014-04	2014-05	2014-06	2014-07	2014-08	2014-09	2014-10	2014-11	2014-12	
23	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	G
24	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	F
25	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	E
26	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	D
27	1,0	1,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	C
28	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	B
29	19,0	19,0	19,0	22,0	13,0	13,0	13,0	15,0	13,0	13,0	13,0	12,0	13,0	13,0	13,0	13,0	A
30	36,0	36,0	36,0	48,0	49,0	49,0	49,0	47,0	49,0	49,0	49,0	43,0	49,0	49,0	49,0	49,0	A+
31	37,0	37,0	37,0	21,0	31,0	31,0	31,0	27,0	31,0	31,0	31,0	33,0	31,0	31,0	31,0	31,0	A++
32	6,0	6,0	6,0	9,0	7,0	7,0	7,0	11,0	7,0	7,0	7,0	12,0	7,0	7,0	7,0	7,0	A+++
33																	

Problem 2 – Results

- Very little difference between scenario based on GfK data and Crawled data
- Equally uncertain
- Scraped data can be used to replace bought data

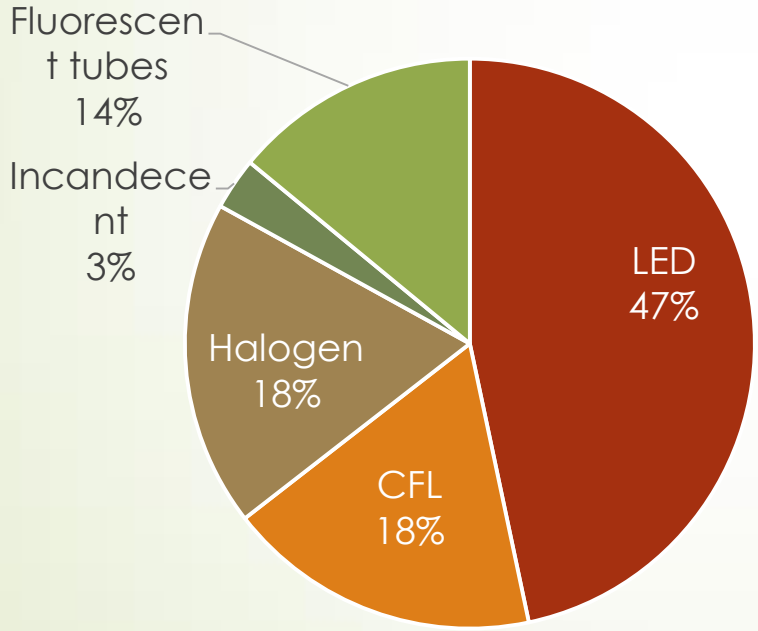


What insights can be gained?
Swedish light market example

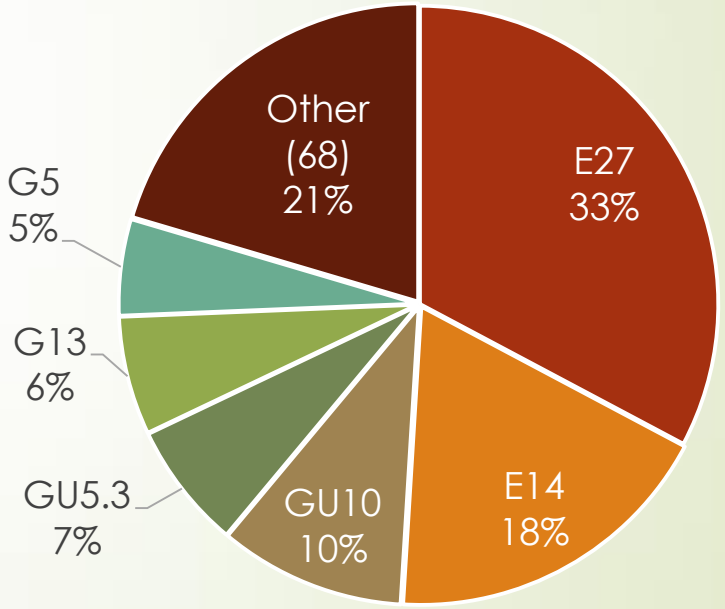
Light data “harvested” from week 43 to week 48

Distributions

Technology distribution

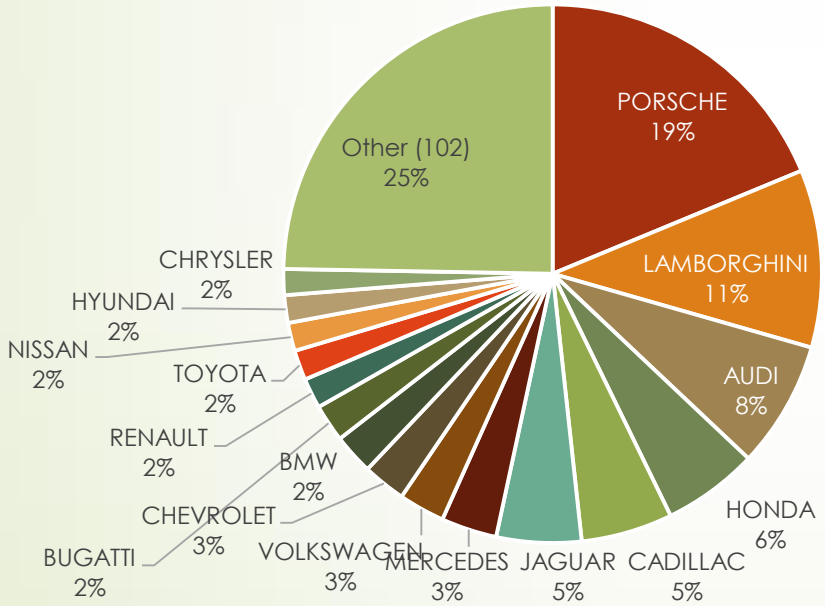


Light bases

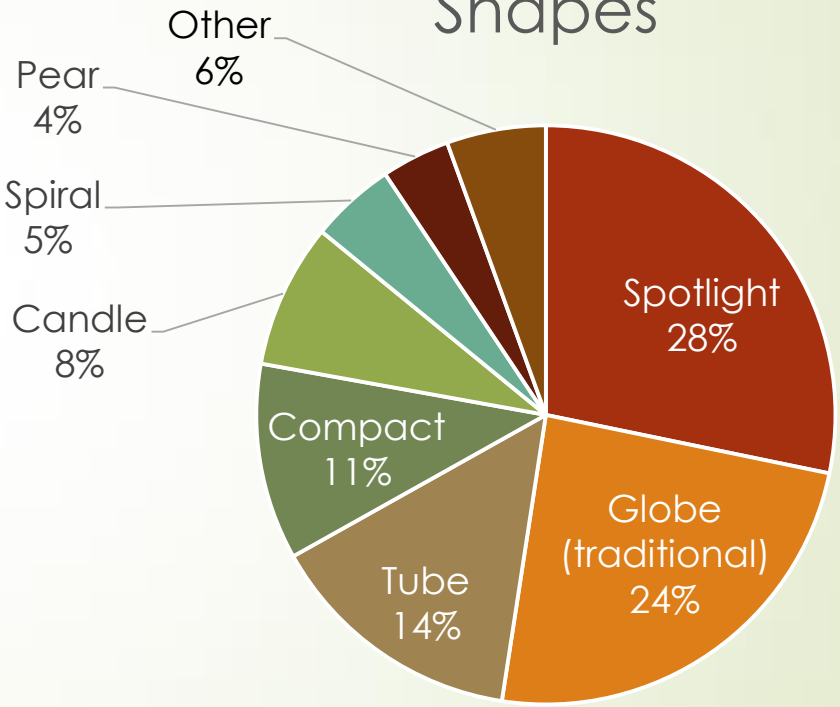


Distributions

Brands (masked)

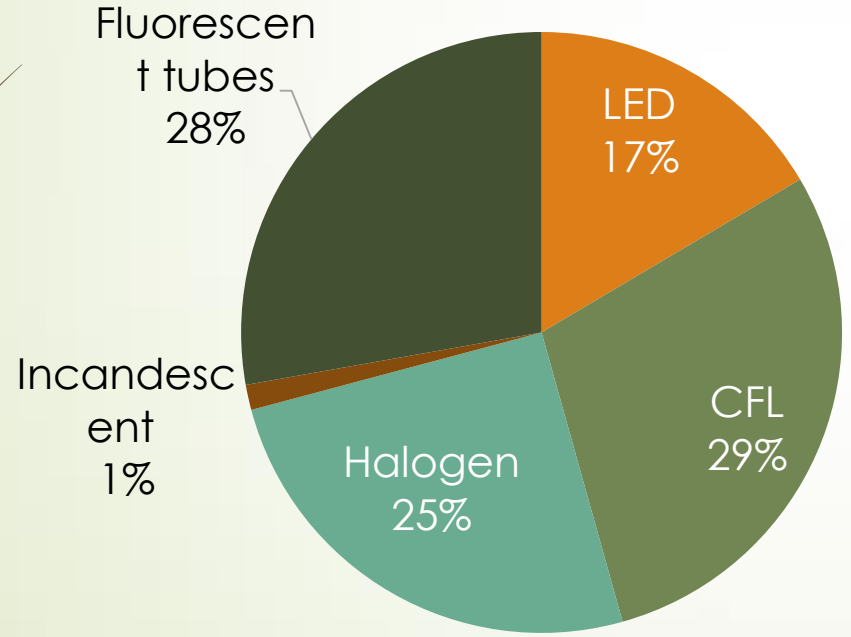


Shapes

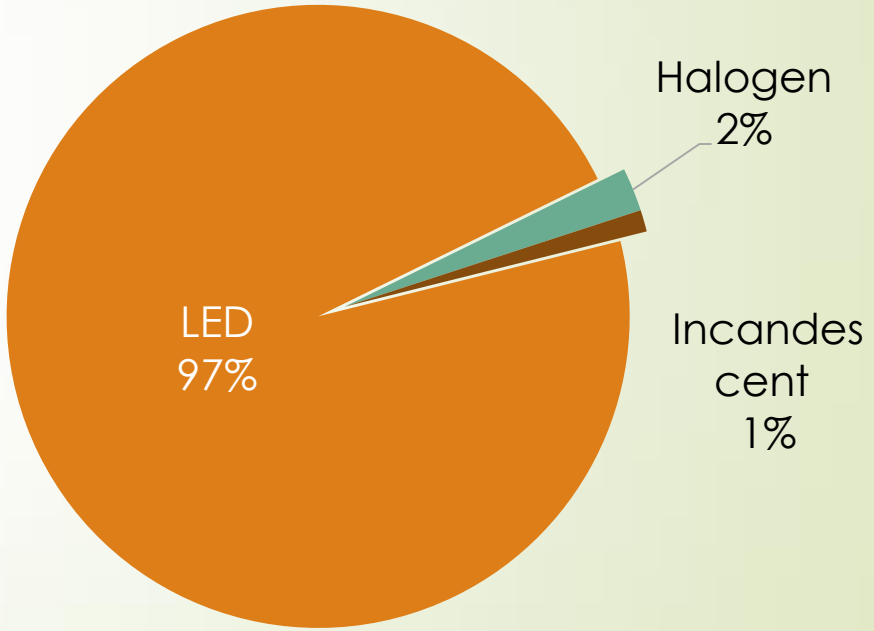


Brands - Technologies

Old brand

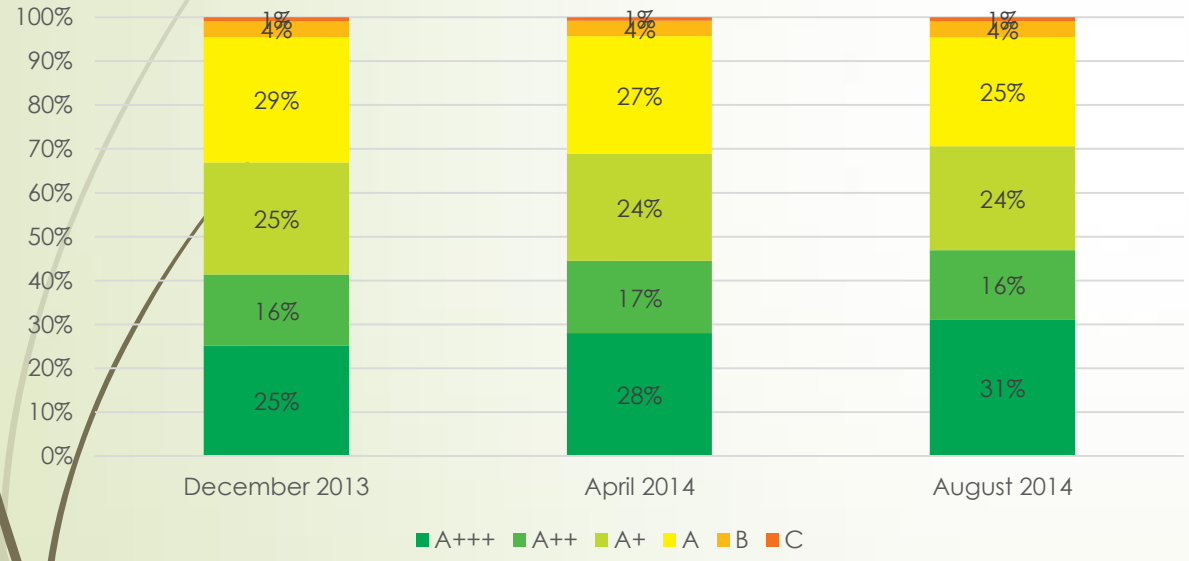


New brand

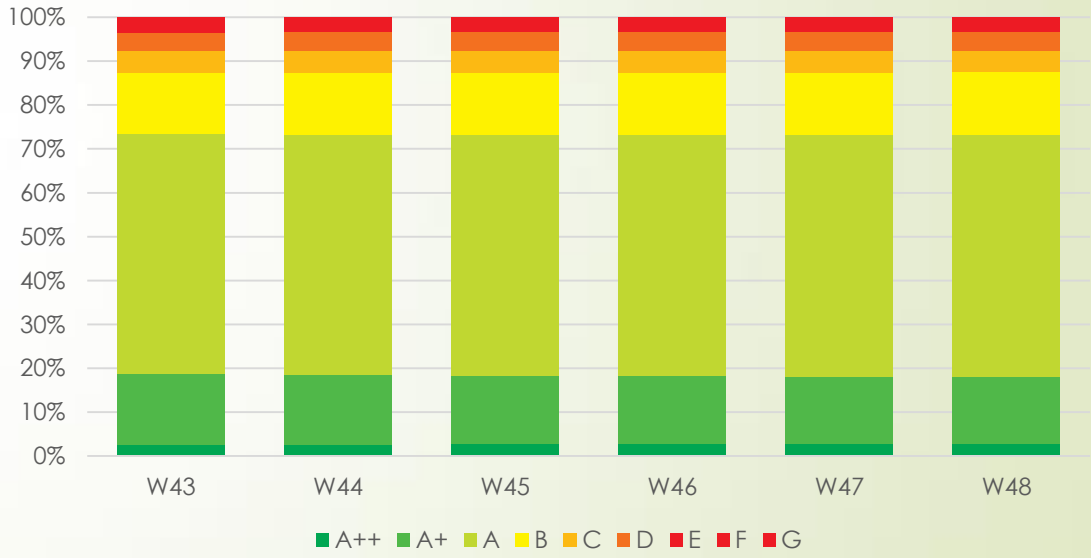


Market transformation

Clothes washers energy label distribution



Light sources energy label distribution (crawl data)



Market enforcement

- Data example (masked)

id_U43	url_U43	Pris_U43	L product	L Manufac	L Antal i fö	L Driftspär	L Effekt_U	L Energikl	L Kvicksilv	L Livslängd	L Motsvar	L Sockel_U	L Typ av la	L Form_U4	L Diamete	L Längd/hö
-254547	http://ww	395:-	CITROEN L	CITROEN	1 st	230 V	10W	A		20 000 h	60 W	E27	LED/Lysdi	Klot (trad	60 mm	110 mm
-254514	http://ww	75:-	CITROEN L	CITROEN	1 st	230 V	3,6W	A		25 000 h	35 W	GU10	LED/Lysdi	Spotlight	50 mm	58 mm
-254473	http://ww	25:-	CITROEN L	CITROEN	1 st	230 V	11W	A	1,3 mg	10 000 h	75 W	G23	Lågenergi	Kompakt	32 mm	17,5 mm
-345435	http://ww	25:-	CITROEN L	FORD	1 st	230 V	9W	A	1,3 mg	10 000 h	60 W	G23	Lågenergi	Kompakt	32 mm	17,5 mm
-261611	http://ww	29:-	FORD Ligh	FORD	1 st	230 V	60W	G		1 200 h	60 W	E27	Glödlamp	Klot (trad	60 mm	105 mm
-345434	http://ww	198:-	FORD Bio	Narva Ljus	1 st	230 V	23W	B	2,5 mg	15 000 h	83 W	E27	Lågenergi	Spiral	60 mm	152 mm
-345433	http://ww	101:-	FORD LED	FORD	1 st	230 V	10W	A		20 000 h	60 W	E27	LED/Lysdi	Klot (trad	60 mm	110 mm
-282521	http://ww	275:-	TOYOTA C	TOYOTA	1 st	231 V	4,7W	A		15 000 h	40 W	E27	LED/Lysdi	Klot (trad	65 mm	113 mm
-282518	http://ww	69:-	CITROEN L	CITROEN	1 st	232 V	4,6W	A		15 000 h	35 W	GU10	LED/Lysdi	Spotlight	50 mm	53 mm
-282513	http://ww	128:-	MERCEDES	MERCEDES	1 st	233 V	12W	A+		25 000 h	75 W	E27	LED/Lysdi	Klot (trad	60 mm	119 mm

NordCrawl project

- Nordic countries
- Building a system on top of scraped data
- Analyze data (e.g. spot trends)
- Generate reports (e.g. compliance reports)

Thanks

Kasper Mogensen

Co-founder, Big2Great ApS

ksm@big2great.dk

Denmark