

Web crawling

Technology and benefits

IEA webinar
December 3rd 2020

Kasper Mogensen, Big2Great, Denmark

Big2Great - Kasper

Kasper Mogensen

Denmark

Co-founder of **Big2Great**

(with Troels Fjordbak)

*Spend **24%** of my life crawling the web*

Master of Business Administration and Computer Science,
Copenhagen Business School

Masters thesis about web crawling

- Mission: From big data to great value
- Software and statistics consultants
- Web crawling, inhouse large-scale crawler (BigCrawler)
- Statistical models builder
- Energy efficiency & product safety

Why is web crawling important?



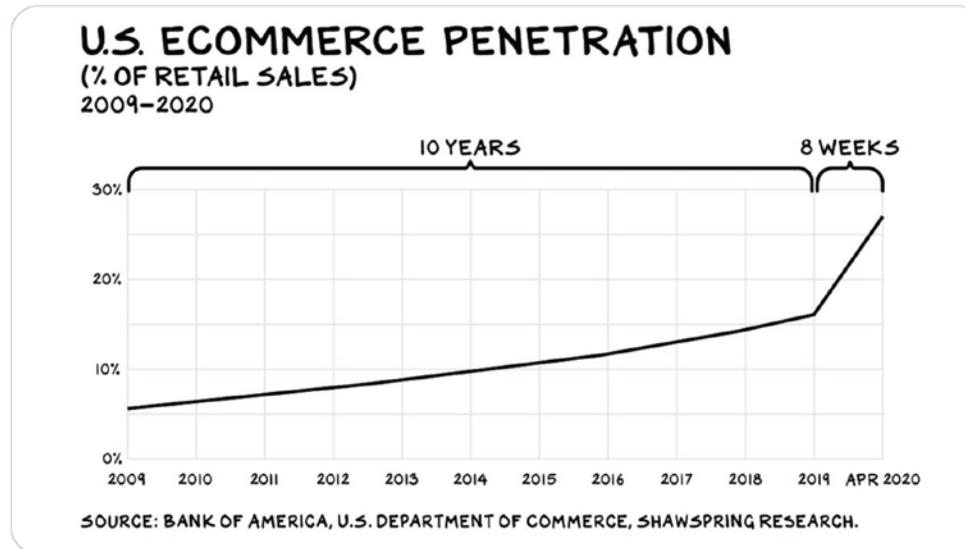
Scott Galloway ✓
@profgalloway

...

from #PostCorona

Online sales grew as much in 8 weeks as it had in the decade before the pandemic

Oversæt Tweet



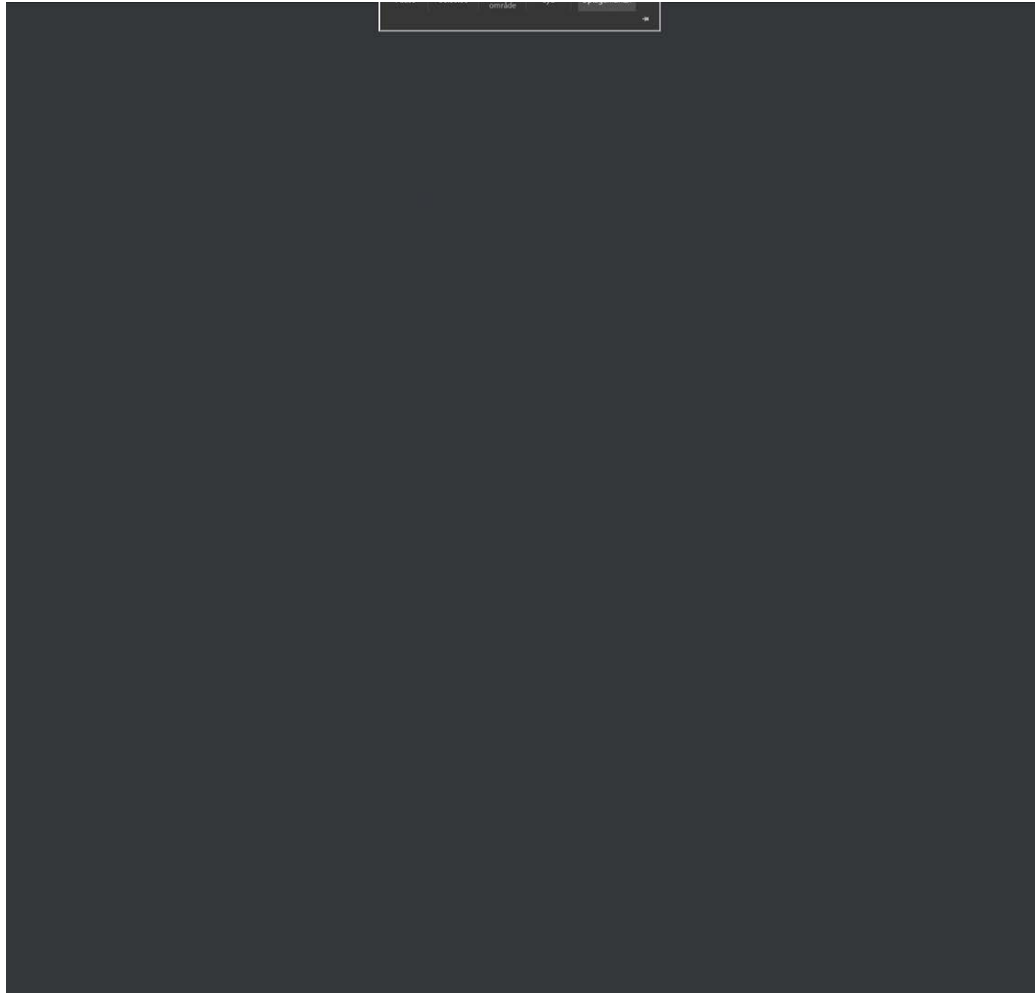
10.16 PM · 29. nov. 2020 · Twitter Web App

(<https://twitter.com/profgalloway/status/1333157834138259456>)













Background

WHAT IS WEB CRAWLING

What is web crawling?



- 1) Page with models
- 2) Visit all links

 <p>★★★★★</p> <p>LG vaskemaskine/tørretumbler F4J6VG0W</p> <p>5 999 Outlet-pris fra 5 099</p> <p>A Datablad</p> <p>✓ Forventet på lager online: 15-01-21</p>	 <p>★★★★★</p> <p>Siemens IQ300 vaskemaskine/tørretumbler</p> <p>6 999 Outlet-pris fra 5 949</p> <p>B Datablad</p> <p>✓ På lager online (-5)</p>	 <p>★★★★★</p> <p>LG vaskemaskine/tørretumbler FH2U2HDM1N</p> <p>4 499 Outlet-pris fra 3 824</p> <p>B Datablad</p> <p>✓ Kolding (-5)</p>	 <p>★★★★★</p> <p>LG vaskemaskine/tørretumbler CM20T5S2E</p> <p>2 999</p> <p>A Datablad</p> <p>✓ På lager online (100+)</p>
 <p>★★★★★</p> <p>CANDY CSWS485TWMCE Washer drye</p> <p>2 999</p> <p>✓ Forventet på lager online: 09-12-20</p>	 <p>★★★★★</p> <p>Samsung vaskemaskine/tørretumbler</p> <p>3 999</p> <p>B Datablad</p> <p>✓ Forventet på lager online: 04-02-21</p>	 <p>★★★★★</p> <p>Beko vaskemaskine/tørretumbler</p> <p>3 999</p> <p>A Datablad</p> <p>✓ Forventet på lager online: 09-12-20</p>	 <p>★★★★★</p> <p>Electrolux PerfectCare 700 vaskemaskine/tørretumbler</p> <p>4 499</p> <p>A Datablad</p> <p>✓ Forventet på lager online: 14-12-20</p>
 <p>★★★★★</p> <p>LG kombineret vaskemaskine/tørretumbler</p> <p>4 999</p> <p>B Datablad</p> <p>✓ Forventet på lager online: 28-12-20</p>	 <p>★★★★★</p> <p>Samsung WD5000T vaskemaskine/tørretumbler</p> <p>4 999</p> <p>B Datablad</p> <p>✓ På lager online (100+)</p>	 <p>★★★★★</p> <p>AEG vaskemaskine/tørretumbler L7WDB861G</p> <p>4 999</p> <p>A Datablad</p> <p>✓ Bestillingsvare. Forventet leveringstid: 35 dage</p>	 <p>★★★★★</p> <p>LG vaskemaskine/tørretumbler CV40V5S2E</p> <p>5 999</p> <p>A Datablad</p> <p>✓ På lager online (50+)</p>

- 3) Download page as text
- 4) Extract info for text

Modelbeskrivelse

Produkttype	Kombineret vaskemaskine/tørretumbler
Type vaskemaskine	Frontbetjent
Placering	Fritstående
Producentens varekode	F4J6VG0W

Kapacitet, forbrug og strøm

Energmærke	A
Vaskeevne (EU)	A
Centrifugeringsøve (EU)	B
Forbrug kWh (vask)	1,17
Forbrug kWh (vask/tørring)	6,12
Forbrug kWh/år (vask)	234,00
Forbrug kWh/år (vask/tørring)	1 224,00
Vandforbrug L/cyklus (vask)	140
Vandforbrug L/cyklus (vask/tørring)	140
Vandforbrug L/år (vask/tørring)	28 000
Nominal kapacitet vaskecyklus (kg)	9,00
Tørrekapacitet (kg)	5,00
Tromle volumen (liter)	59
Centrifugeringshastighed (rpm)	1 400
Lydniveau vask (dB)	55
Lydniveau tørring (dB)	56
Lydniveau centrifugering (dB)	75

Funktioner

Forskudt start	Ja
Tromlebelysning	Nej
Indikator for resterende tid	Ja
15C koldvask	Ja
Hurtig vask (min.)	30
Dampfunktion	Ja
Automatisk dosering af vaskemiddel	Nej
Wi-Fi	Nej

```

1 <div class="tab-data-row tab-specs-row show" data-tab="tab-specs">
2
3 <div class="tab-slot" itemscope="">
4 <div class="spec-wrap">
5 <dl class="spec-section">
6 <dt class="any-1-1">
7 <span class="spec-section-title">Modelbeskrivelse</span>
8 </dt>
9 <dd class="any-1-1">
10 <table class="spec-table 5-1-1">
11 <tbody>
12 <tr>
13 <td class="any-1-4 5-1-2" data-md-name-id="311">
14 <span class="abbr">Produkttype</span>
15 </td>
16 <td class="any-3-4 5-1-2 xh-highlight" data-md-
17 </tr>
18 <tr>
19 <td class="any-1-4 5-1-2" data-md-name-id="315">
20 <span class="abbr">Type vaskemaskine</span>
21 <span class="fa fa-question" data-md-name-id="315">
22 <span>Viser om vaskemaskinen er top-
23 </span>
24 </td>
25 <td class="any-3-4 5-1-2 xh-highlight" data-md-
26 </tr>
27 <tr>
28 <td class="any-1-4 5-1-2" data-md-name-id="311">
29 <span class="abbr">Placering</span>
30 </td>
31 <td class="any-3-4 5-1-2 xh-highlight" data-md-
32 </tr>
33 <tr>
34 <td class="any-1-4 5-1-2" data-md-name-id="314">
35 <span class="abbr">Producentens varekode</span>
36 </td>
37 <td class="any-3-4 5-1-2 xh-highlight" data-md-
38 </tr>
39 </tbody>
40 </table>
41 </dd>
42 </dl>
43 <dl class="spec-section">
44 <dt class="any-1-1">
45 <span class="spec-section-title">Kapacitet, forbrug og str
46 </span>
47 </dt>
48 <dd class="any-1-1">
49 <table class="spec-table 5-1-1">
50 <tbody>
51 <tr>
52 <td class="any-1-4 5-1-2" data-md-name-id="304">
53 <span class="abbr">Energmærke</span>
54 <span class="fa fa-question" data-md-name-id="304">
55 <span>Viser produktets energiklasse, A
56 </span>
57 </td>
58 <td class="any-3-4 5-1-2 xh-highlight" data-md-
59 </tr>
60 <tr>
61 <td class="any-1-4 5-1-2" data-md-name-id="310">
62 <span class="abbr">Vaskeevne (EU)</span>
63 <span class="fa fa-question" data-md-name-id="310">
64 <span>Viser hvor grundigt produktet va
65 </span>
66 </td>
67 <td class="any-3-4 5-1-2 xh-highlight" data-md-
68 </tr>
69 <tr>
70 <td class="any-1-4 5-1-2" data-md-name-id="320">
71 <span class="abbr">Centrifugeringsøve (EU)</span>
72 <span class="fa fa-question" data-md-name-id="320">
73 <span>Viser hvor effektivt produktet c

```


NordCrawl

Nordic Council of Ministers

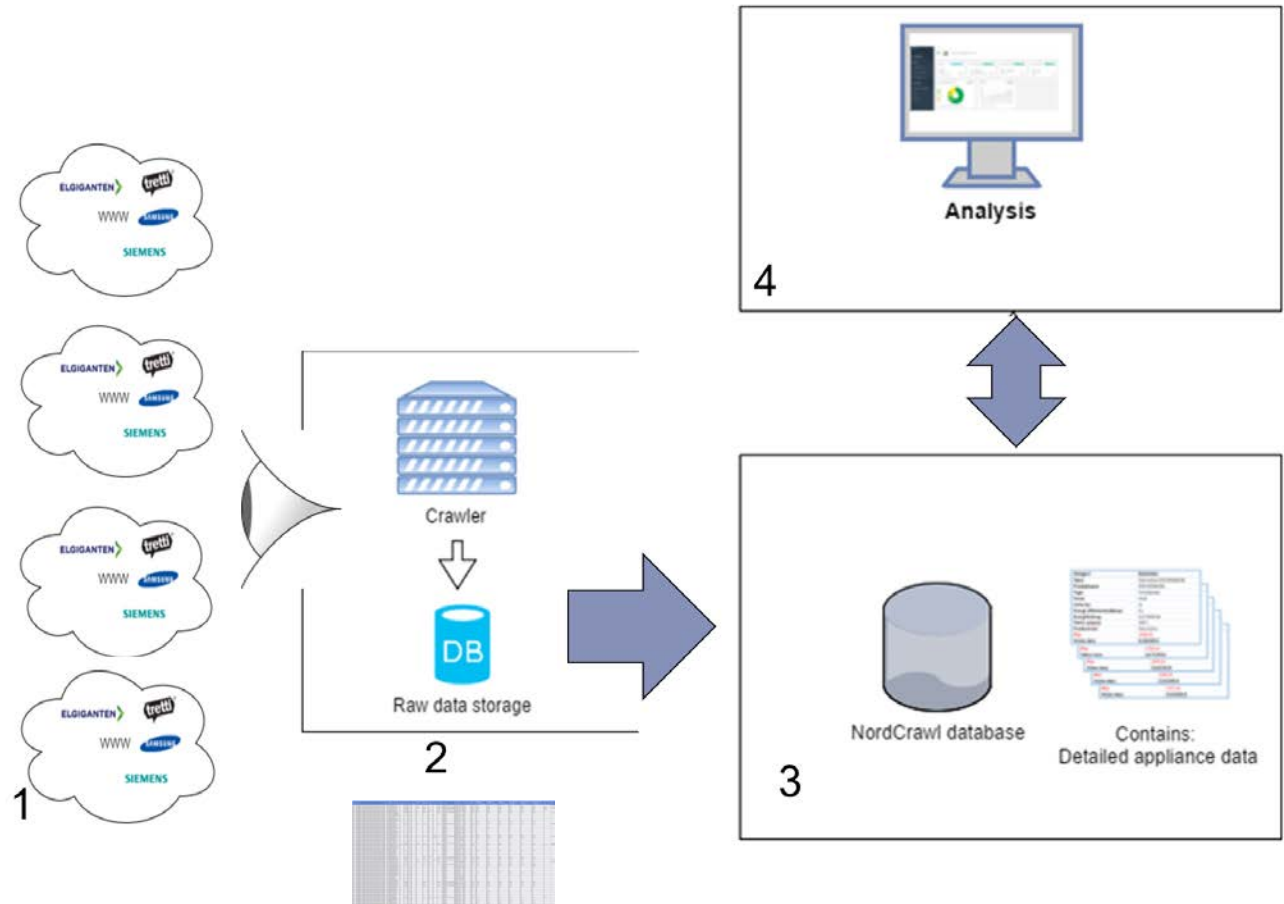


The Nordic energy/market surveillance agencies











System overview

1. Data on products, stored at various publicly available web sites.
2. The web crawler engine and a temporary raw data storage.
3. A product data base containing processed data.
4. The analysis module (a program), which access and display the data in various ways.

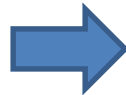


QA – Attribute cleaning & normalization

+ Norm rule
⚡ Perform Normalization Now

Order	Rule type	Rule action	Edit	Delete
1	Remove Text	Remove: dB		
2	Remove Text	Remove: (A)		
3	Remove Text	Remove: [space]		
4	Remove Text	Remove: A		

76
77
77
77,0
78
78



76
77
77.0
78

Examples of

NORDCRAWL FOR MARKET SURVEILLANCE

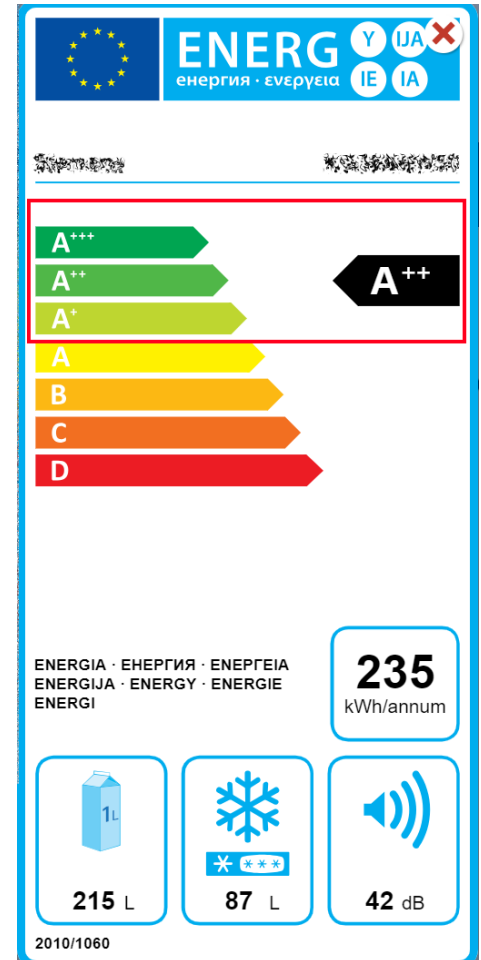
MS – Banned energy classes

Purpose

- The Ecodesign directives sets minimum performance standards for energy consumption
- Identify product below the minimum standard

Method

- Minimum Energy Efficiency Index (EEI) or minimum consumption per year in kWh => minimum energy class
- Search database for recently seen model below minimum energy class
- Remove “false positive” products that is not under the regulation
- Some might be on market before regulation



MS – Banned energy classes



Search for help...

Hello Kasper! Log off

Compliance Rules - Min Energy label

Rule type	Product	Product name	In power	Min value	Active	Find non-compliant	Regulation
Energy class		All Washingmachines	12-2013		✓	Find non-compliant	1015/2010
Energy class		All Vacuum cleaners	1-2014		✓	Find non-compliant	666/2013
Energy class		All range hoods	1-2016		✓	Find non-compliant	66/2014
Energy class		All Ovens	1-2016		✓	Find non-compliant	66/2014
Energy class		All Refrigerator-Freezer	7-2014		✓	Find non-compliant	643/2009
Energy class		Refrigerator only	7-2014		✓	Find non-compliant	643/2009
Energy class		All freezers	7-2014		✓	Find non-compliant	643/2009
Energy class		All Chest Freezers	7-2014		✓	Find non-compliant	643/2009
Rule type	Product	Product name	In power	Min value	Active	Find non-compliant	Directive

MS – Banned energy classes






Compliance Rules - All Vacuum cleaners - min label class: D

<p>Non-compliance rate Rate</p> <p>0.0060 Non-compliance rate</p>	<p>Non-compliant models Number</p> <p>1 Number of non-compliant models</p>	<p>Models Number</p> <p>166 Total number of models</p>
--	---	---

BrandId	Brand Name	Model Name	ModelId	Value	View model
44	Electrolux	Electrolux	60944	F	To model
BrandId	Brand Name	Model Name	ModelId	Value	View model



MS – Banned energy classes

	Minimum	 DK	 SE	 NO	 FI	 ICE
Washing machines	A+	0,75% (4)	1,36% (7)	0,44% (2)	0,67% (1)	0% (0)
Vacuum cleaners	D	1,86% (6)	0,6% (1)	0% (0)	5,22% (7)	0% (0)
Range hoods	F	0% (0)	0% (0)	0% (0)	0% (0)	-
Ovens	C	0% (0)	0% (0)	0% (0)	0% (0)	-
Refrigerator-Freezer	A+	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Refrigerator	A+	1,19% (9)	0% (0)	0,83% (4)	0,36% (1)	0% (0)
Freezers	A+	0,22% (1)	0% (0)	0% (0)	0% (0)	0% (0)
Chest Freezers	A+	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)

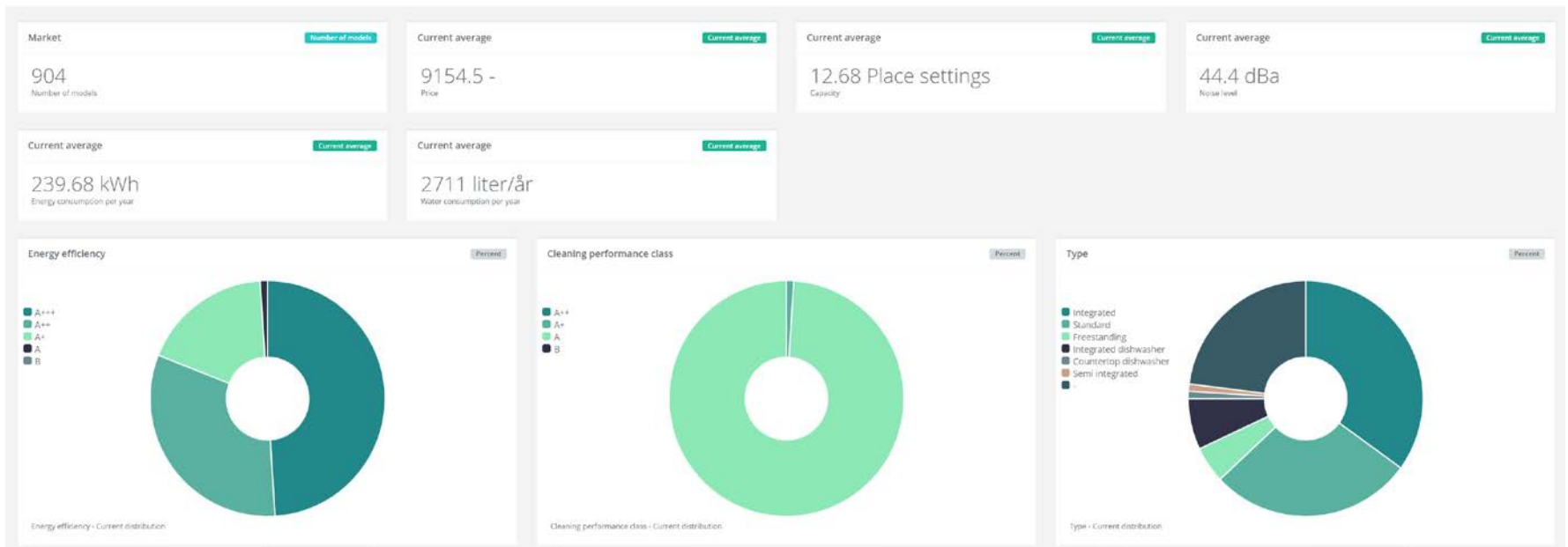
Examples of

NORDCRAWL FOR POLICY EVALUATION

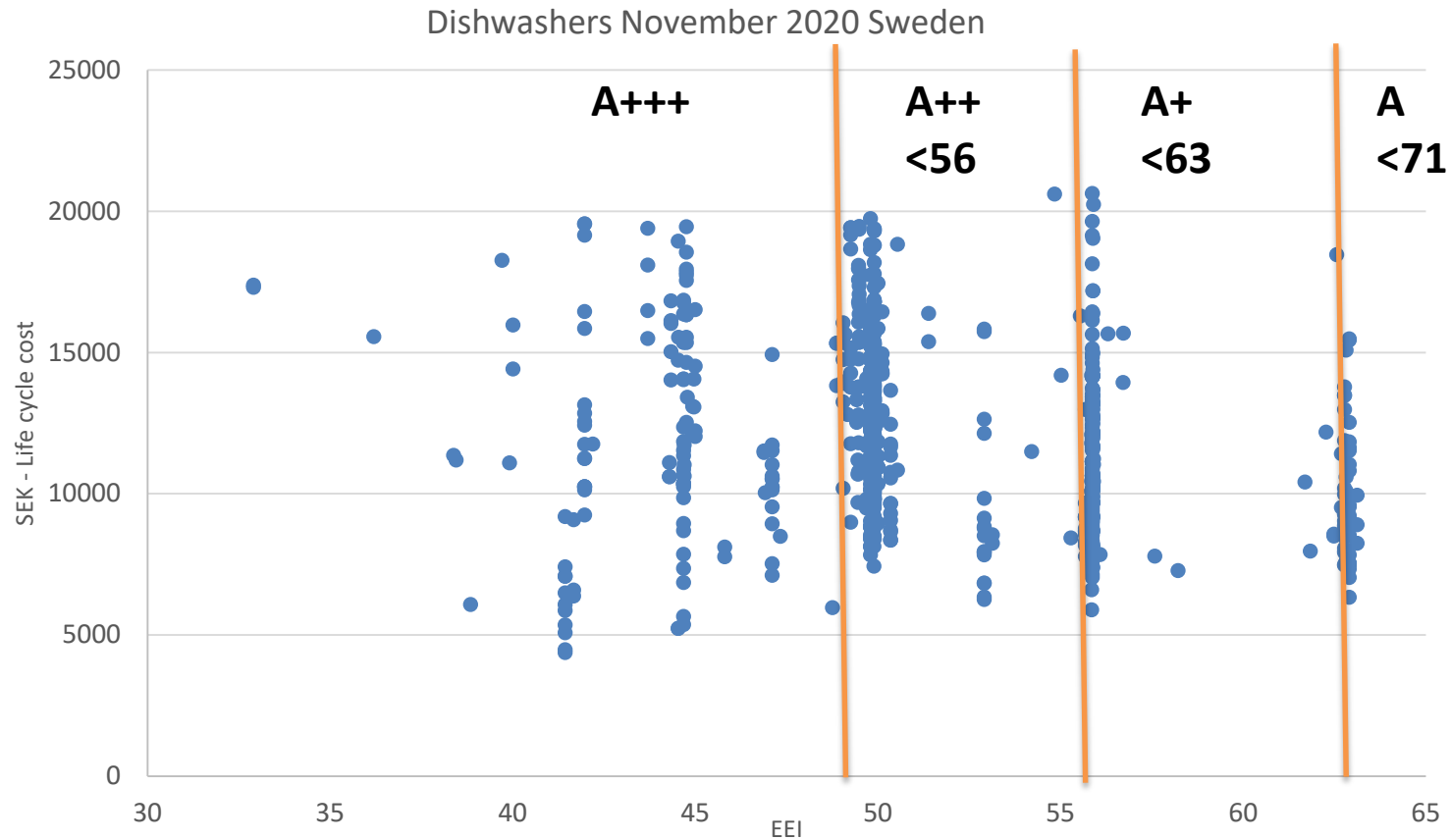


Market snapshot for policy evaluation

☰ ⚙️ 🇸🇪 | All dishwashers

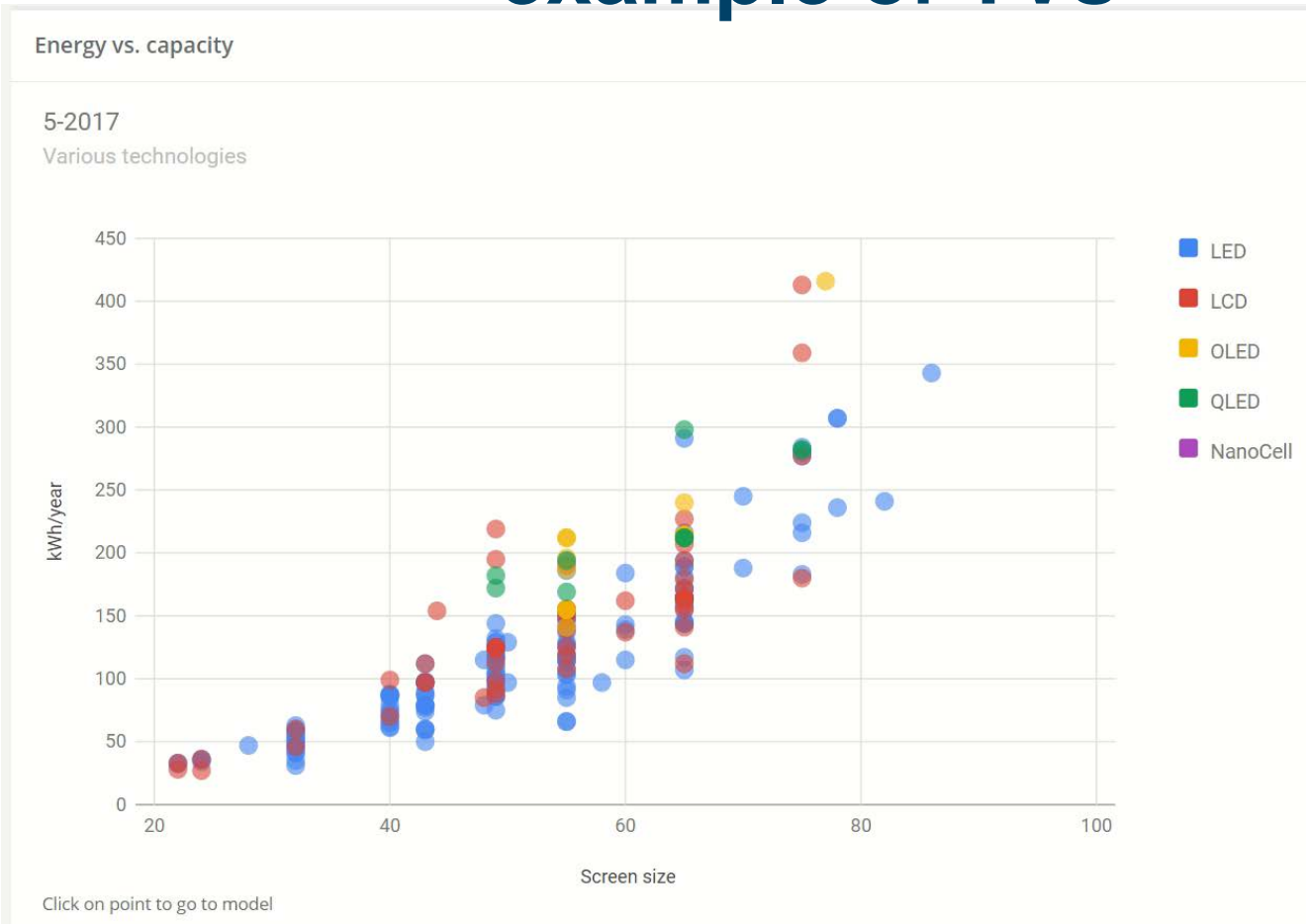


Policy evaluation – trends



Policy evaluation – trends.

Stop motion: market development example of TVs



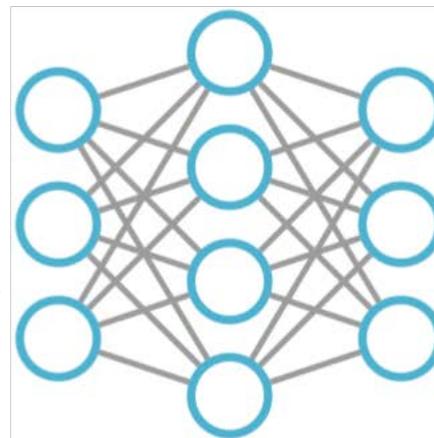
NordCrawl future

- **New energy label tracker**
 - Energy label rescale March 2021 (*Washing machines and washer-dryers, Dishwashers, Electronic displays, and Refrigerators*)

Energy labels
from websites



Machine learning categorizing
energy labels (New or old)



New/Old



BIG 2 GREAT
From big data to great value

Other countries



IEA project air-conditioner and refrigerators in South Africa and Indonesia

- Fall 2019
- Major online retailer and manufacturers (provided by local partners)
- Get, clean and merge data





BIG 2 GREAT
From big data to great value

Information quality

<p>R 11,999.00</p> <p>GET IT ON: Credit from R 603.52 per month x 30 months*</p>	<p>R 6,990.00</p> <p>GET IT ON: Credit from R 378.27 per month x 30 months*</p>	<p>R 14,999.00</p> <p>GET IT ON: Credit from R 732.31 per month x 30 months*</p>
<p>R 3,299.00</p>	<p>R 2,999.00</p>	<p>R 7,799.00</p>

<p>Midea Cool by Alliance Wall Split 18000 Btu/hr Non Inverter Air Conditioner</p> <p>★★★★★ 0 Reviews Write a Review</p> <p>Midea Cool by Alliance Wall Split 18000 Btu/hr Non Inverter Model MHP090A03 Eco-friendly BSL SA.</p> <p>Brand: Midea Category: Air Conditioners</p>	<p>From R5 499.00 at iStock</p> <p>View Offer</p>
<p>Samsung APO000 Wall Split 18000 Btu/hr Non Inverter Air Conditioner</p> <p>★★★★★ 0 Reviews Write a Review</p> <p>Samsung APO000 Wall Split 18000 Btu/hr Non Inverter Model AH18TQ6AAK003 Eco-friendly</p> <p>Brand: Samsung Category: Air Conditioners</p>	<p>From R9 299.00 at iStock</p> <p>View Offer</p>
<p>Midea 2000W One-Touch Induction Cooker Black</p> <p>★★★★★ 0 Reviews Write a Review</p> <p>2000W One-Touch Induction Cooker...</p> <p>Brand: Midea Category: Air Conditioners</p>	<p>From R799.00 at Clicks</p> <p>View Offer</p>
<p>Midea Blue Wall Split 9000 Btu/hr Inverter Air Conditioner</p> <p>★★★★★ 0 Reviews Write a Review</p> <p>Midea Blue Wall Split 9000 Btu/hr Inverter Air Conditioner Model MHA09A1-09R0CA1 260V Ready</p> <p>Brand: Midea Category: Air Conditioners</p>	<p>From R7 999.00 at iStock</p> <p>View Offer</p>



Information quality

Product Details | Reviews (0) | Questions (0) | Stats

Features

Manufacturer Jet-air

Model Number 1332016002
254

Fit your rooms with this AC and enjoy advanced cooling and enjoy fresh and cold. Covering every corner of the room and cooling it evenly is its primary feature. The Jet-Air Portable Air Conditioner is engineered with a classic design and can be operated from a user-friendly remote control. Features: - Remote control - Elegant Modern Panel Design, with LED display - Free Movement - Super-quiet Running - Air flow: 400 m3/h - Applicable Room Area: 13-27 m2 - Noise level: 64 dB(A) - Warranty: 3 years. [T Apply] Dimension(mm): - Net: 480 x 400 x 795 - Gross: 520 x 450 x 885 More Specs: - Rating Cooling/Heating(KW/h): 12000/10000 - Rating Cooling/Heating Current (A): 5,95/4,88 What's in the box 1 x Jet-Air Portable Air conditioner 1 x Remote Control 1 x Exhaust Hose 1 x Window Kit

Spesifikasi Sharp AH-A7SAY AC Split 3/4PK Standard Putih

Produk SKU	AH-A7SAY
Type	AC Standard ⓘ
Warna	Putih
Daya Listrik	595 Watt ⓘ
Dimensi	88 x 29 x 22 cm
Berat	9 Kg
Garansi	1 Tahun Resmi ⓘ
PK	3/4 PK ⓘ
BTU/h	7000 BTU/h ⓘ
Refrigrant	R-32 ⓘ
Uk. Pipa	ø1/4 ø3/8 ⓘ
Dimensi Ou.	60 x 50 x 27 cm
Berat Ou.	19 Kg
Made In	thailand
Customer Reviews	★★★★★ 2 review

Ex. level of information

South Africa – Air-con

Cooling capacity btu/h	90%
Refrigerant	80%
Room size m2	37%
Energy class	17%
Noise Level (dB) outdoor unit	9%
EER	5%

Indonesia – Refrigerators

Price	89%
Capacity total liter(net)	78%
Type of product	17%
Electricity per year (kWh)	4%
Noise level dB (A)	2%

Observations

- Information on energy and energy efficiency is minimal
- Hard for consumers to make well-informed decisions
- “EU like” requirements for consumer information would help

Thank you!

Kasper Mogensen, CTO
Big2Great

ksm@big2great.dk

Feel free to contact me