

Building Ranking Infrastructure: Data-Driven, Lean, Flexible

Sergii Khomenko, Data Scientist, STYLIGHT
sergii.khomenko@stylight.com, @lc0d3r



APACHECON
EUROPE

CORINTHIA HOTEL
BUDAPEST, HUNGARY
— NOVEMBER 17-21, 2014 —

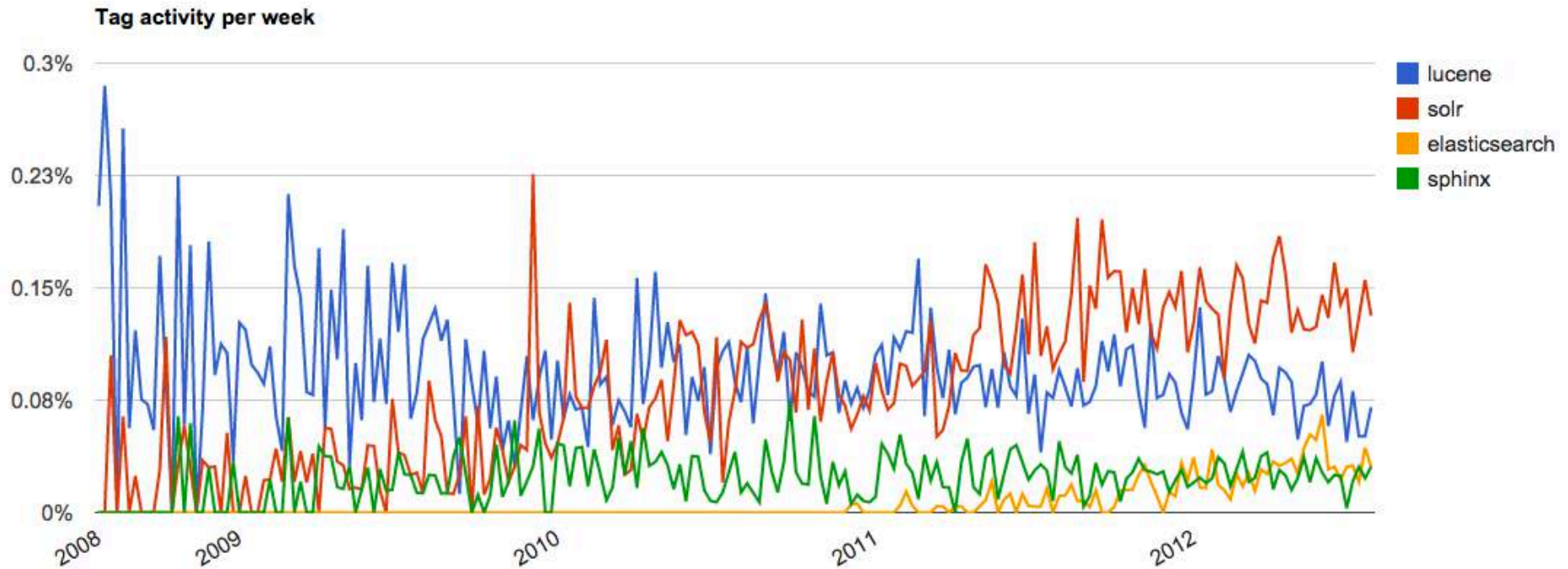


Agenda

1. Problem definition
2. Boosting
3. Lean approach to ranking infrastructure
4. Real-world examples



LUCENE, SOLR, ELASTICSEARCH





SOLR USERS



Instagram

Fast beautiful photo sharing



STYLIGHT





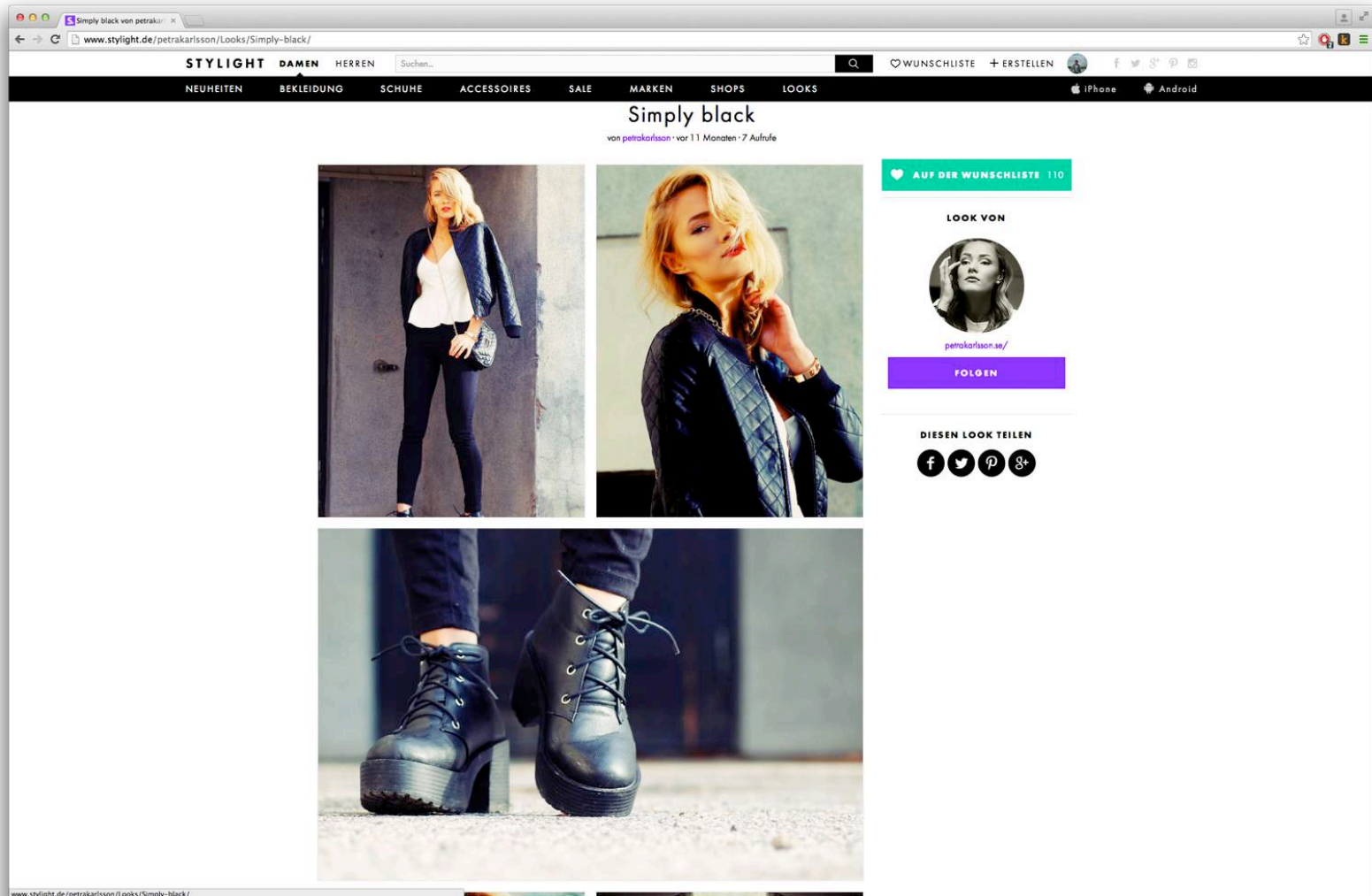
THE BEST PLACE TO DISCOVER FASHION

The screenshot shows the Stylight website interface. At the top, there's a navigation bar with 'STYLIGHT DAMEN HERREN' and a search bar. Below that, a black bar contains categories: 'NEUHEITEN', 'BEKLEIDUNG', 'SCHUHE', 'ACCESSOIRES', 'SALE', 'MARKEN', 'SHOPS', 'LOOKS'. On the right, there are links for 'WUNSCHLISTE', 'REGISTRIEREN', 'LOGIN' and mobile app icons for 'iPhone' and 'Android'. The main content area is titled 'Jeden Tag neu' with the subtitle 'Die angesagteste Auswahl aus über 100 Shops.' Below this, a grid of 24 fashion items is displayed, each with a product image, a title, a price, and a 'Versand: kostenlos' label. Some items have discount tags (e.g., -33%, -58%, -52%, -20%).

Item	Price	Discount
So Nice - KLEIDER - Midikleider ...	144,00 €	
Boho Inspiration von trendy_dreams	119,90 €	
Buffalo - Buffalo Plateau Sandal...	119,90 €	
Iska - Kleid mit Reißverschluss...	28,57 € 42,84 €	-33%
Vilo - Shift Dress Tiny schwarz/I...	34,90 €	
Fall Midi Dress von silviallace	34,90 €	
Hallhuber - Ethikleid silber	119,95 €	
Georgia Rose - Bahoul by Geor...	99,00 €	
Darling - KLEIDER - Lange Kleide...	99,00 €	
FW #2 von iloukylou	99,00 €	
Ganni - KLEIDER - Kurze Kleider ...	54,00 €	
Mai Più Senzo - Plateustiefelette...	149,95 €	
Ethnic Look von multimotion	2,90 € 6,90 €	-58%
Pieces - Ethno-Ohrhänger Orfan...	34,90 €	
Noisy May - Rock im Ethno-Look...	34,90 €	
Glamorous - Cardigan mit Ethno...	59,95 €	
Ethno Jacket von stylight_editor	29,95 €	
Noisy May - Ethno-Muster Blaze...	29,95 €	
(Bottom row items)		-52%, -20%

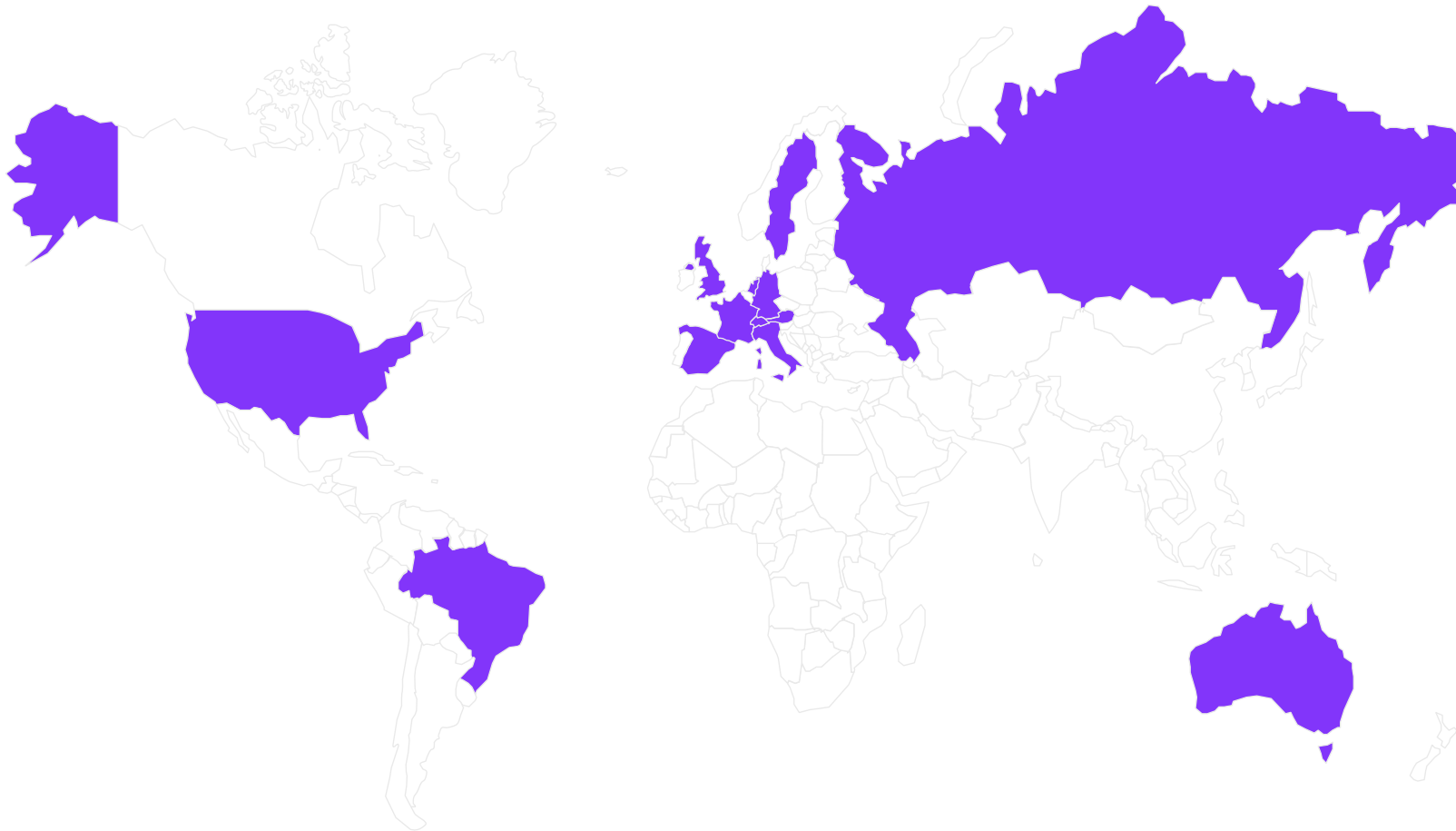


GET INSPIRED BY LOOKS CREATED BY COMMUNITY





STYLIGHT – INTERNATIONAL COMMUNITY

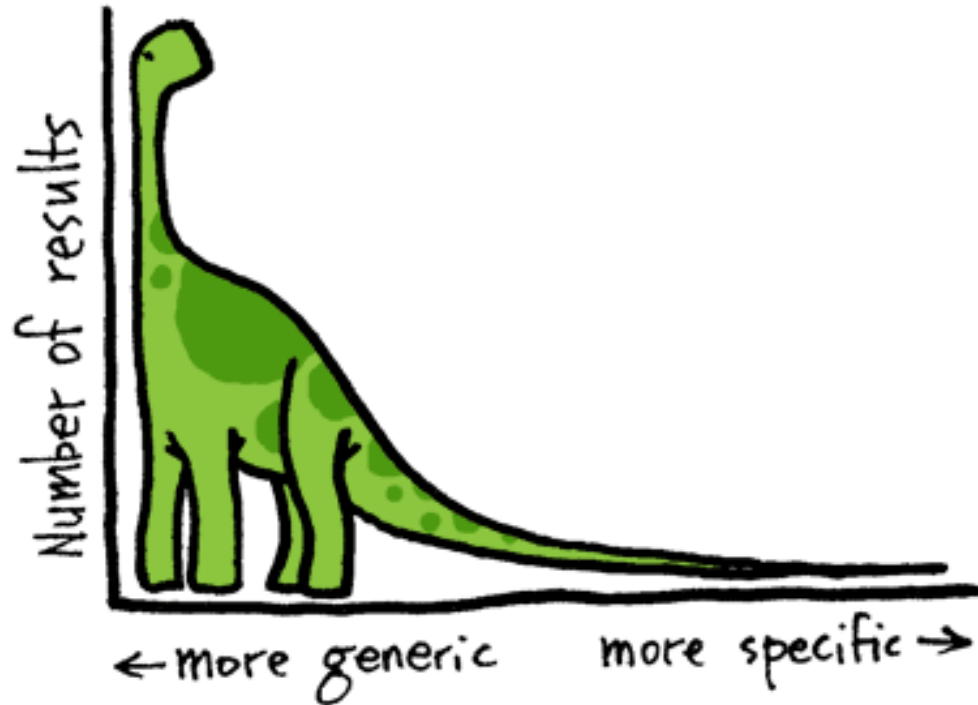




PROBLEM DEFINITION



Problem definition





Problem definition

Ranking specifics:

- Seasonal influence
- Trends
- Cold start of new countries, shops
- Multiple dimensions of ranking model



IMPROVING RELEVANCE



TF-IDF - default scoring model in Lucene/Solr

$$\text{coord}(q, d) \cdot \sum_{t \in q} \left(\text{tf}(t, d) \cdot \text{idf}(t)^2 \cdot \text{norm}(\text{field}(t), d) \right)$$

- matching more query terms is better
- more occurrences of a query term is better
- more novel terms increase doc score more than common terms



Improving relevance

Stages to improve relevance in Solr

- Editorial voting
(QueryEvaluationComponent)
- Indexing time
(analyzing content, text analysis)
- Query-time
(function queries, boosting)



Improving relevance - Solr queries

q = +brand:adidas shop:monshowroom^3

q = +adidas monshowroom

defType = dismax

qf = brand shop^3

sort = user_ratings desc, score desc

qq = adidas

q = {!boost b=\$b defType=dismax v=\$qq}

b = prod(popularity, clicks)



Definition of solr.ExternalFileField

```
<types>
```

```
  <fieldType name="float" class="solr.FloatField"
omitNorms="true"/>
```

```
  <fieldType name="file_delta2"
class="solr.ExternalFileField" keyField="id" defVal="1.0"
indexed="false" stored="false" valType="float" />
```

```
</types>
```

```
<fields>
```

```
  <field name="delta2" type="file_delta2"/>
```

```
</fields>
```



Example of external file with boosting

`\cores\de_DE\products\external_delta2.txt`

`15062471=0.5`

`15062479=0.2`

`15062507=0.41`



LEAN APPROACH TO RANKING



Lean manufacturing, lean enterprise, or lean production, often simply, "***lean***", is a production practice that considers the expenditure of resources for any goal other than the creation of value for the end customer to be wasteful, and thus a target for elimination.

Essentially, lean is centered on ***preserving value with less work.***



Lean approach to Ranking

Requirements:

- Decreasing time to implement new ranking model
- Possibility to use more dynamic ranking models
- Keeping working infrastructure alive



Lean approach to Ranking

Requirements:

- A/B testing without changing entire infrastructure
- Performance level - “still fast” and “transparent”



Python benchmark, consistency checker

- gaid gaid, -g gaid** Google analytics site id.
- gadate gadate** a date to fetch the most popular pages from Google Analytics
- solr solr, -s solr** Solr server to benchmark performance.
- pages number, -p number** a number of top pages from Google Analytics.
- repeats number, -r number** a number of repeats for an every page.
- compare compare, -c compare** Different rankings algorithms to compare.



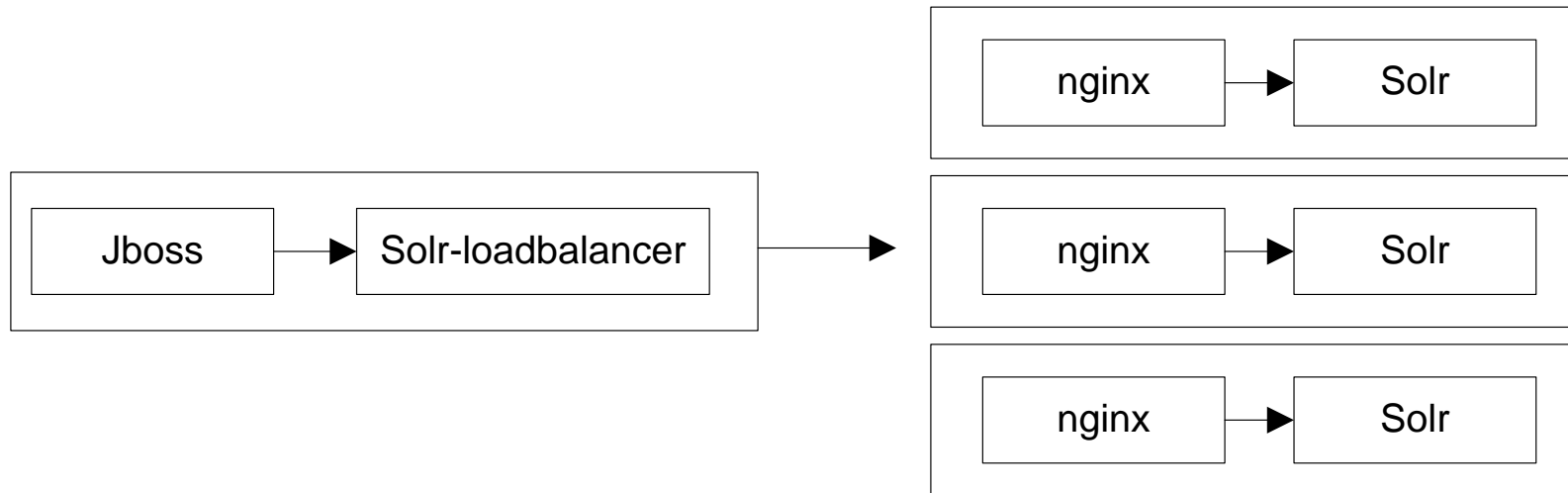
Python benchmark, consistency checker

```
python solr-benchmark\benchmark.py -c  
RankingClassical,RankingDelta2 --cmpmode  
1
```

```
python solr-benchmark\benchmark.py -c  
RankingClassical,RankingDelta2
```

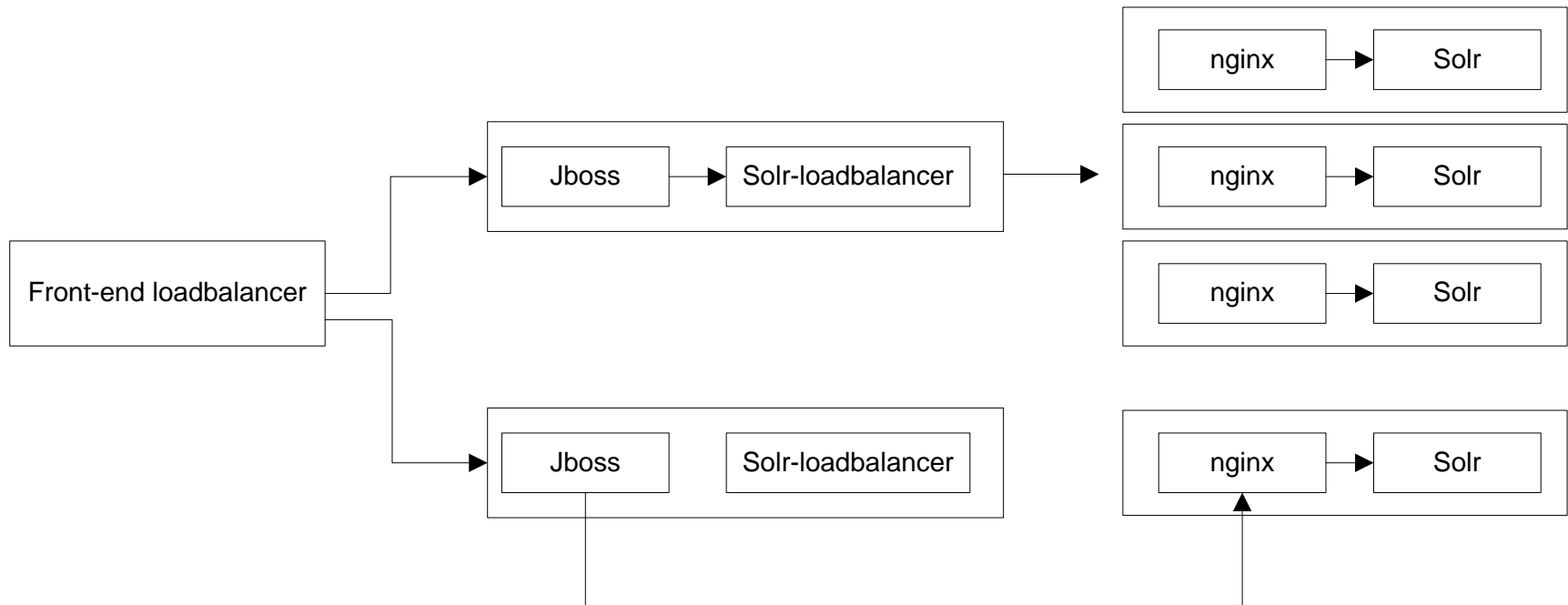


Common search infrastructure





Updated infrastructure





```
include nginx
```

```
nginx::config { "solr_dev": }
```

```
nginx::solr-ranking { "delta2":
```

```
  urls => [
```

```
    "/search.action?
```

```
    gender=women&brand=2271&tag=1161&tag=877&tag  
    =468",
```

```
    "/search.action?
```

```
    gender=men&brand=11235&tag=10203&tag=10299&ta  
    g=10326"
```

```
  ],
```



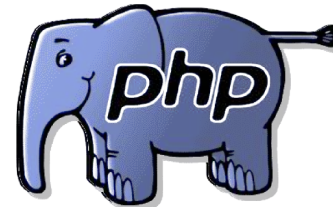
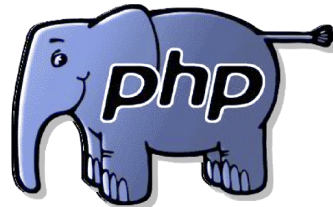
```
<% urls.each do |url| -%>
if ($args ~* <% if url['gender'] > 0 -%>gender_id%3A<%=
url['gender'] %>.*<% end -%><% url['tags'].each do |tag| -
%>tag_id%3A<%= tag %>.*<% end -%><% if url['brand'] > 0 -
%>brand_id%3A%28<%= url['brand'] %>%29<% end -%>) {
  set $orig $args;
  set $args "q={!boost+b=%24b+defType=dismax+v=%24qq}
&qq=id:*";
  rewrite ^(.*)$ "$1?$orig" break;
}
<% end -%>
```



REAL-WORLD EXAMPLES



Elephant-Driven architecture





Simplified Version



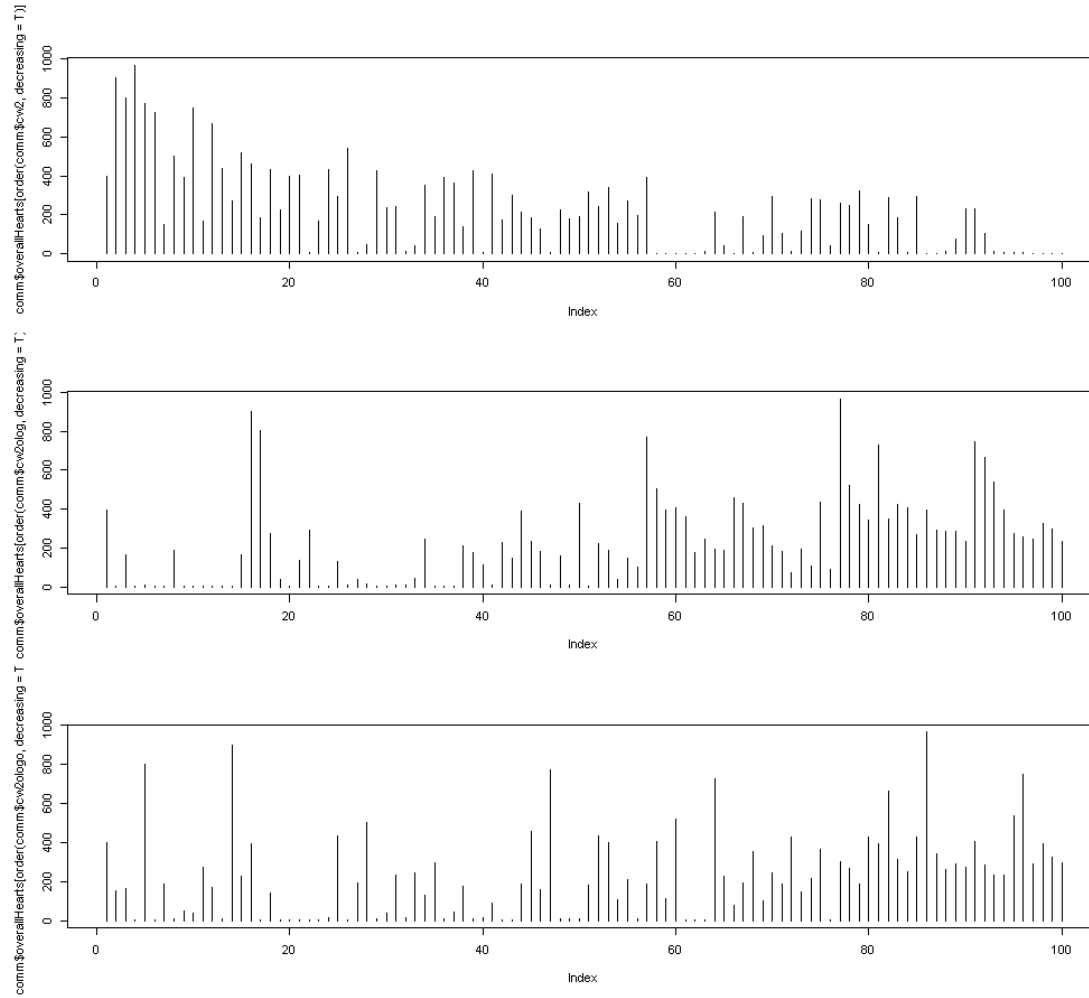


Simplified Version





Real-world examples





Multiple points to evaluate

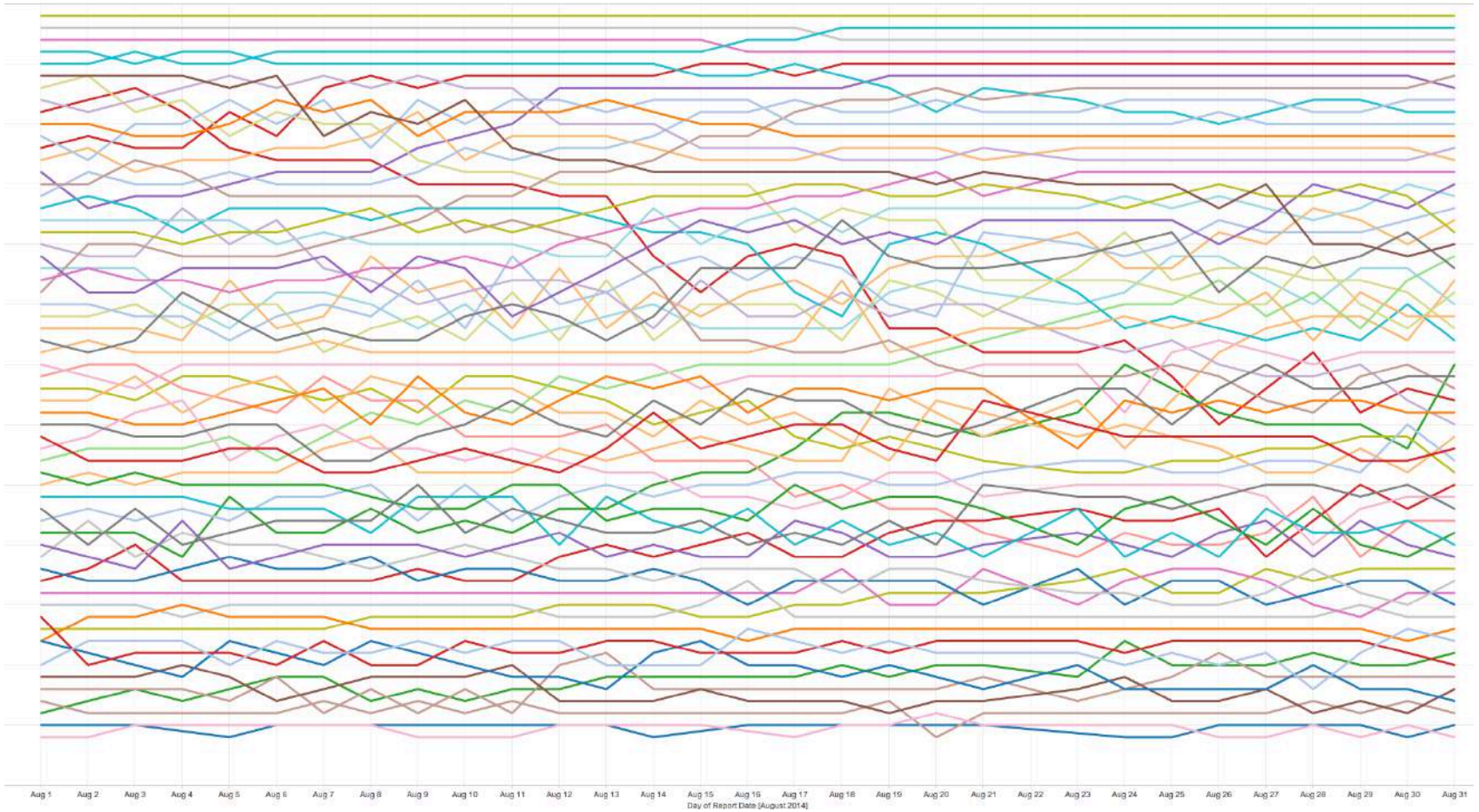
Stages to evaluate the model:

- R ranking model
- Independent Solr-node
 - For internal use-cases
 - Testing for some of pages
 - A/B roll out for % of users
- Production roll out



Real-world examples

< traffic by category rank table for categories **dataviz of subcat changes** badenmode Zehentrenner mantel jacken spain >

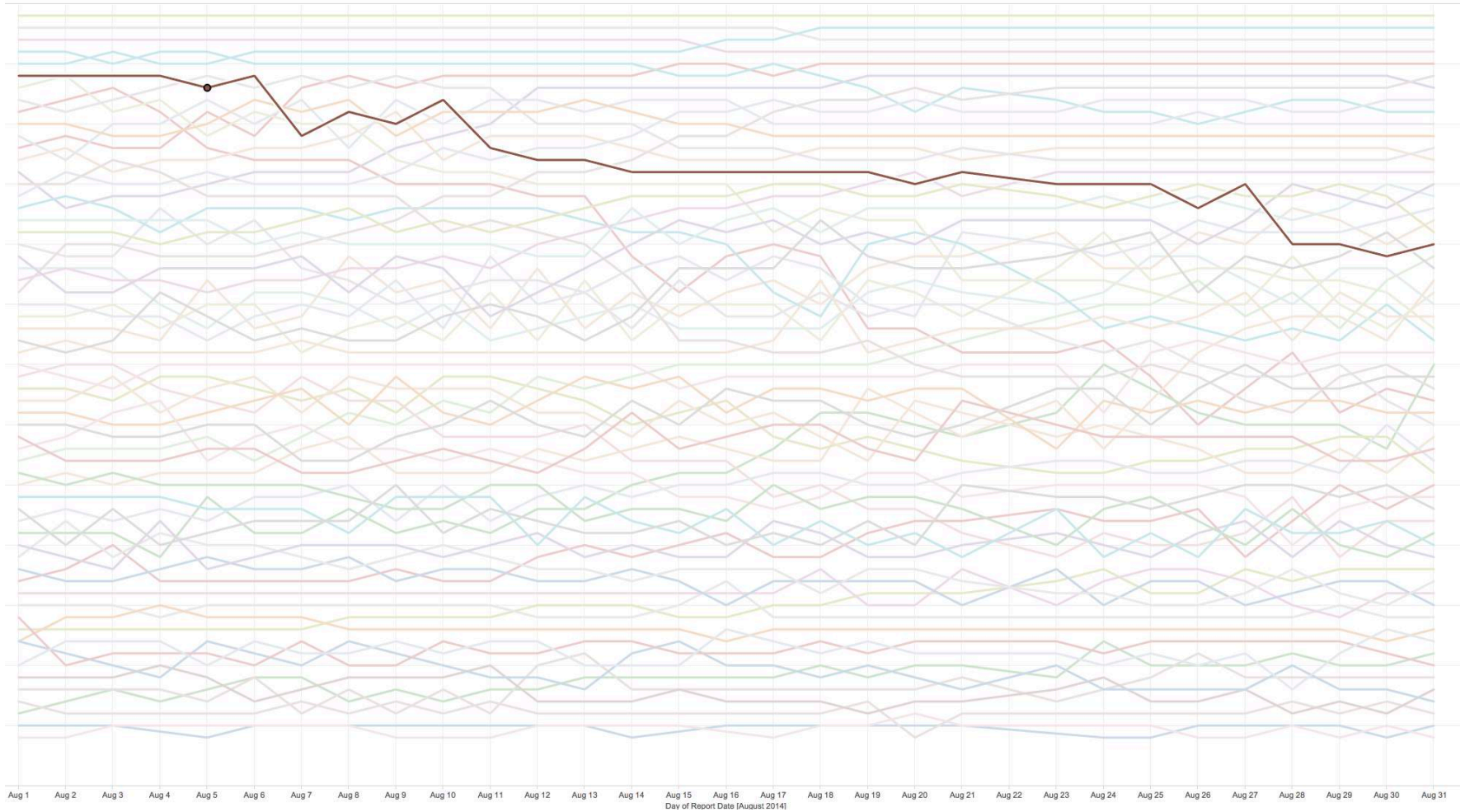




Real-world examples

< traffic by category rank table for categories dataviz of subcat changes **badenmode** Zehentrenner mantel jacken spain >

New Blank Point
Duplicate





Multiple points to evaluate

Stages to evaluate the model:

- R ranking model
- Independent Solr-node
 - For internal use-cases
 - Testing for some of pages
 - A/B roll out for % of users
- Production roll out



Thanks for your attention!

Questions?



STYLIGHT

Sergii Khomenko

Data Scientist

STYLIGHT GmbH

sergii.khomenko@stylight.com

[@lc0d3r](#)

Nymphenburger Straße 86

80636 Munich, Germany

STYLIGHT.COM



REFERENCE LIST

- Stack Overflow Tag Trends
http://hewgill.com/~greg/stackoverflow/stack_overflow/tags/#!lucene+solr+elasticsearch+sphinx
- Public websites using Solr
<http://wiki.apache.org/solr/PublicServers>
- CommonQueryParameters
<http://wiki.apache.org/solr/CommonQueryParameters>
- Thoughts in plain text <http://lc0.github.io/>
- STYLIGHT Engineering
<http://www.stylight.com/Engineering/>