

Predicting Resource Consumption in a Large-Scale Information Retrieval System

Some ideas on dealing with inherently noisy data in Machine Learning

Who are Meltwater?

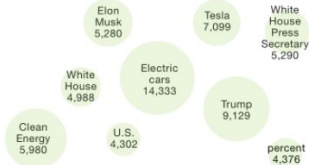


Dashboards

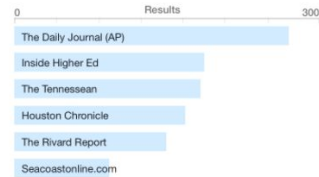
Tesla Dashboard

Jan 1 - Jan 7, 2017

Trending Themes



Top Sources



Sentiment



Heat Map



Top Locations



Web Impact

Understand how news coverage drives traffic to your site



Tuesday, September 12, 2017



Welcome to the Etsy Media Roundup. This is not a comprehensive collection of all articles referencing Etsy, but the most important articles pertaining to Etsy and our favorite campaign! Please click on the headlines to read the articles in full.

Etsy News

These Beauty And The Beast Bridal Heels Will Make You Feel Like The Princess You Truly Are
Cosmopolitan, syndicated to two outlets, including Best Products and True Viral News - 08/11/2017
Every girl wants to feel like a princess on their wedding day, and these shoes will ensure you truly achieve this goal.

3 Ways Software Will Transform The Market For Handcrafted Goods
The Huffington Post - 08/11/2017
In a society where the average person spends ten hours a day staring at one screen or another, consumers are craving a return to handcrafted goods.

15 Stunning Sapphire Engagement Rings To Celebrate September's Birthstone
Love Inc. Mag. - 08/12/2017

As more and more couples push traditional wedding norms to the side, many sapphirees are embracing color in lieu of the classic white diamond when it comes to wedding rings.
Faded Baby Motions: They're Soft, Snuggly, And Perfect For A Nursery
Mantle Magazine - 08/11/2017
For artist Andrea Burnett, taking bags as a way to remember home.

Why Amazon's Whole Foods And Karmora Partnerships Make Perfect Sense For The Brands
Forbes - 08/11/2017

Much debate has gone on regarding Amazon's apparent divergent strategies of acquiring Whole Foods as well as Amazon Fresh.
As Amazon Pushes Forward With Robots, Workers Find New Roles
The New York Times - 08/11/2017
Nissa Scott started working at the cavernous Amazon warehouse in southern New Jersey last year, stacking products like the size of small dinosaurs.

Florida's Bay Online Retailers Let Them Down Ahead Of Time
WFLX - 08/11/2017
Maya Kogu was in California when Hurricane Irma began hitting toward Florida.

Amazon's China Hiring Signals Renewed Ambitions In Alibaba Battle
Bloomberg - 08/11/2017
Amazon.com Inc. is hiring by the hundreds in China to fill jobs ranging from internet software engineers to designers for AWS, positioning the company to nudge some of the market share it lost to Alibaba Group Holding Ltd. in the world's largest online shopping arena.

Alibaba Is Following Amazon's Lead In One Big Way
Business Insider - 08/11/2017
Since April 23, 2016, shares of Amazon's have risen 157%.

Industry News

Zilligse Raises \$18M For Its Fashion E-Commerce Service In Southeast Asia

Communications Page

What we do and who we are
Editorial Communications Dashboard
Core guidelines for all external communications

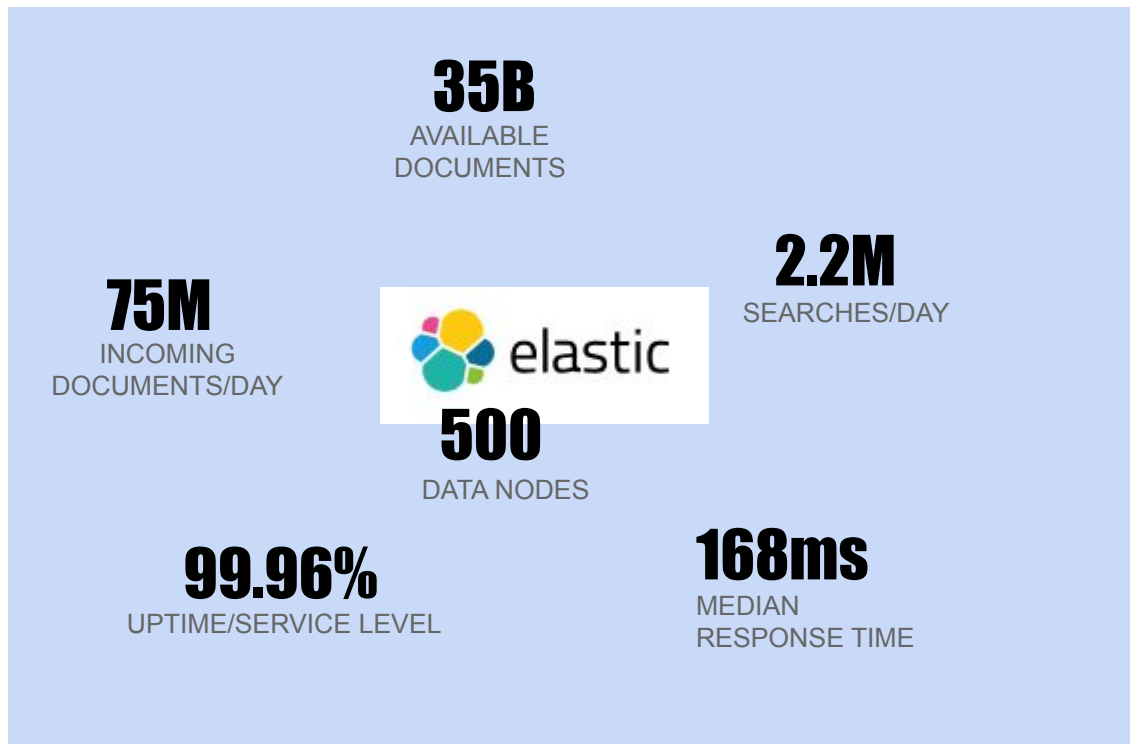
News Blog
Where we publish timely company news, product launches, and other announcements

Speaking Guidelines
A resource for speaking at a conference, at a meet-up, at another company, etc.

Additional questions? Reach out to communications@etsy.com or find us on Slack at #etsy

Want more content news? Join us at [etsy.com/press](https://www.etsy.com/press) or @etsy on Slack

Large-Scale Information Retrieval



Large-Scale Information Retrieval

75M
INCOMING
DOCUMENTS/DAY

35B
AVAILABLE
DOCUMENTS



500
DATA NODES

2.2M
SEARCHES/DAY

99.96%
UPTIME/SERVICE LEVEL

168ms
MEDIAN
RESPONSE TIME

```
{  
  "query": {  
    "notMatchQuery": {  
      "field": "body.content.text",  
      "type": "term",  
      "value": "Greek"  
    },  
    "matchQuery": {  
      "allQueries": [  
        {  
          "field": "body.content.text",  
          "type": "term",  
          "value": "GAIA"  
        },  
        {  
          "field": "body.content.text",  
          "type": "term",  
          "value": "Conference"  
        },  
        {  
          "anyQueries": [  
            {  
              "field": "body.content.text",  
              "type": "term",  
              "value": "Gothenburg"  
            },  
            {  
              "field": "body.content.text",  
              "type": "term",  
              "value": "Göteborg"  
            }  
          ]  
        }  
      ]  
    }  
  }  
}
```

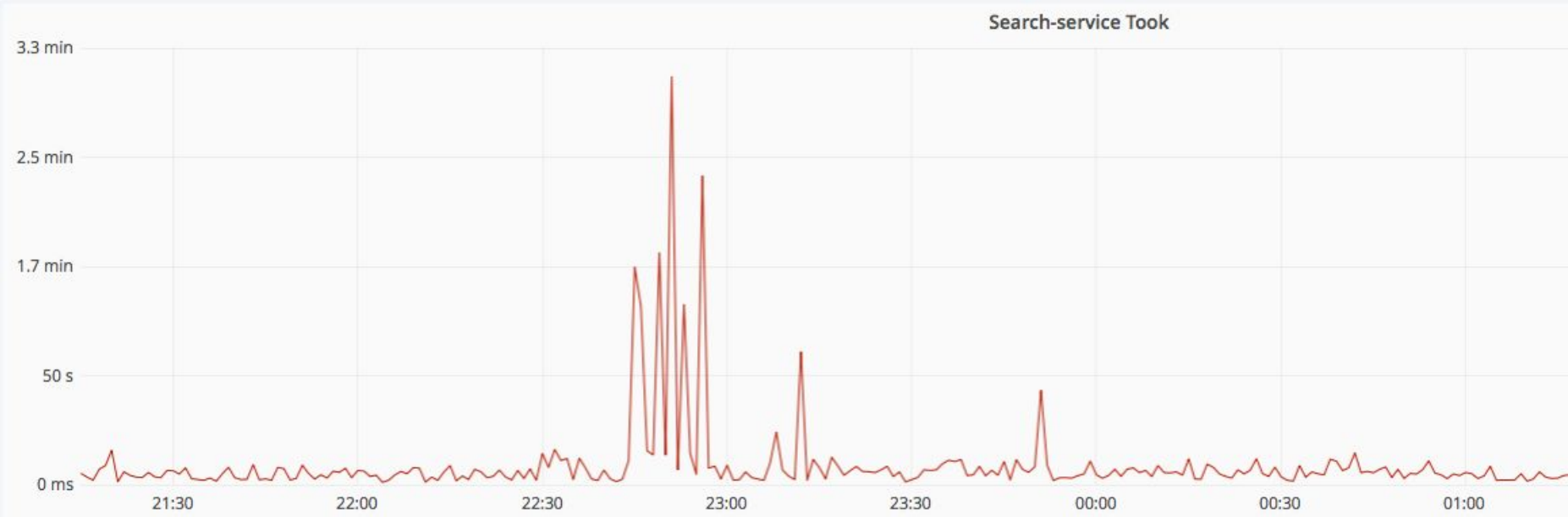
Some fun queries

```
1 {
2   "query": {
3     "anyQueries": [
4       {
5         "field": "body.content.text",
6         "type": "wildcard",
7         "value": "a*"
8       },
9       {
10        "field": "body.content.text",
11        "type": "wildcard",
12        "value": "b*"
13      },
14      {
15        "field": "body.content.text",
16        "type": "wildcard",
17        "value": "c*"
18      },
19      {
20        "field": "body.content.text",
21        "type": "wildcard",
22        "value": "d*"
23      }
24    ],
25    ...
26  }
27 }
```

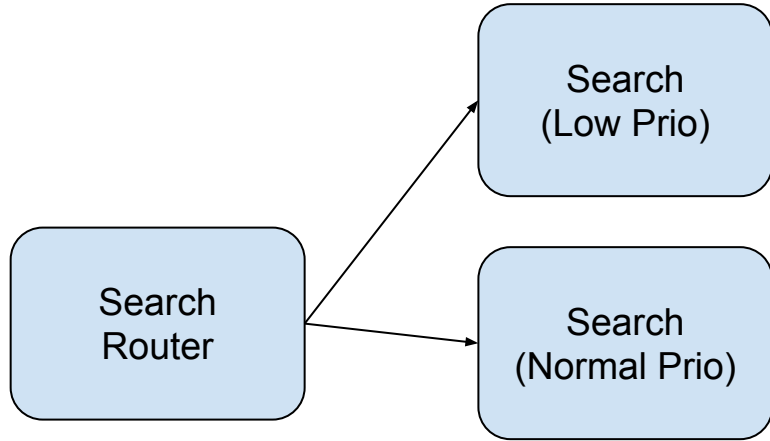
```
1 {
2   "query": {
3 >     "notMatchQuery": {=
8     "matchQuery": {
9       "allQueries": [
10        {
11 >         "anyQueries": [=
223211         "type": "any"
223212       }
223213     ],
223214     "type": "all"
223215   },
223216   "type": "not"
223217 },
223218 "viewRequests": {=
223349 }|
```


Search spikes

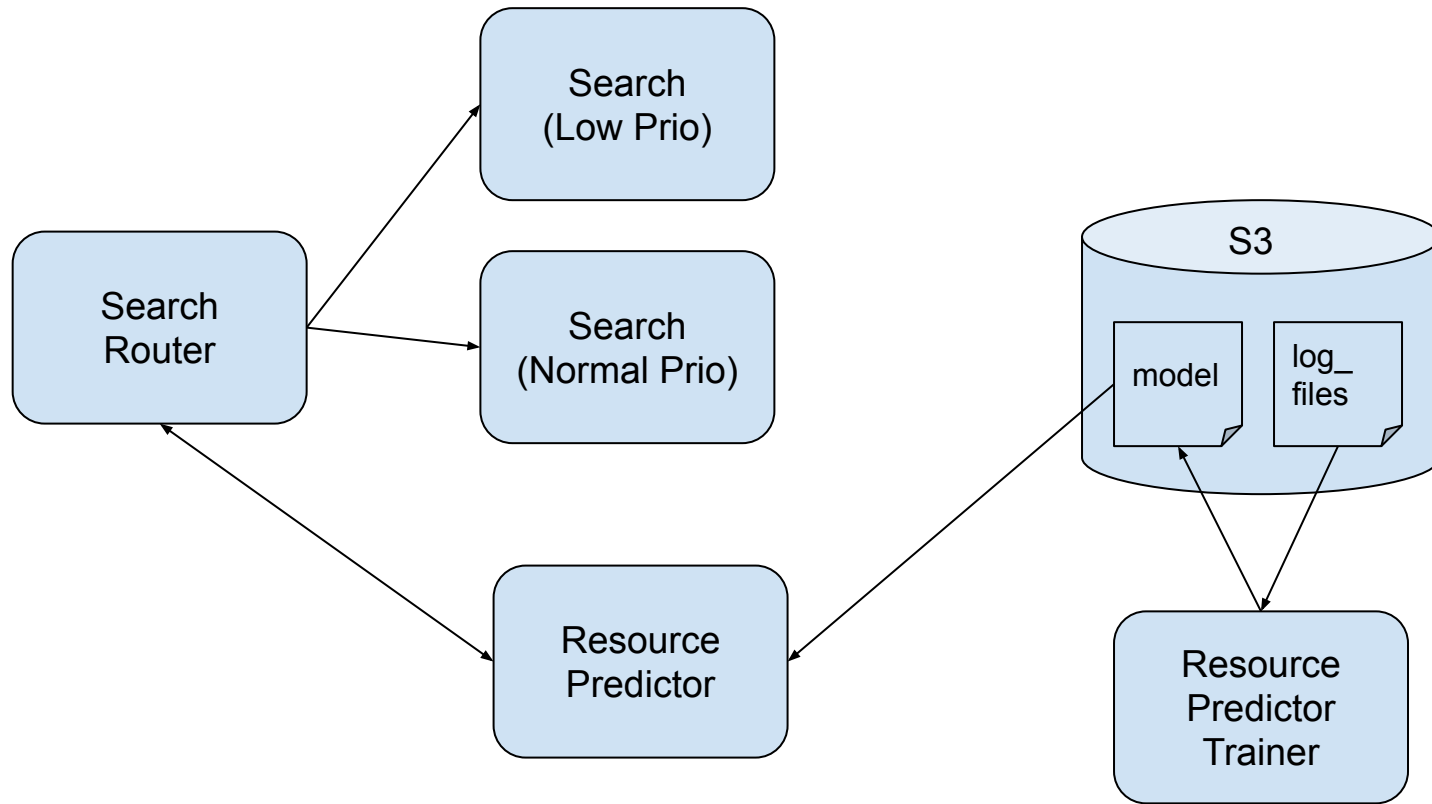
Search Times (average)



Query Resource Prediction for Routing



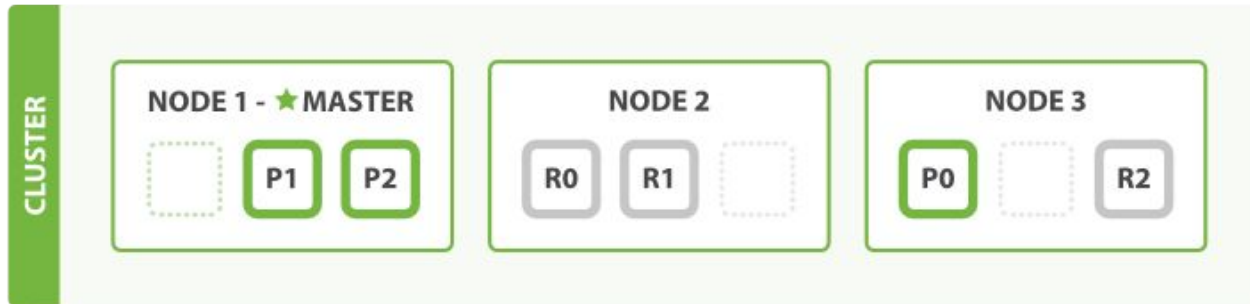
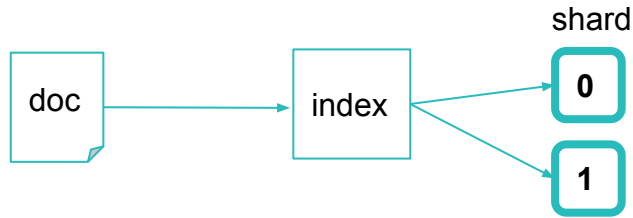
Query Resource Prediction for Routing





elasticsearch

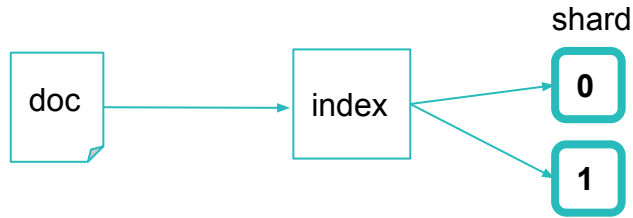
- scalable, distributed search



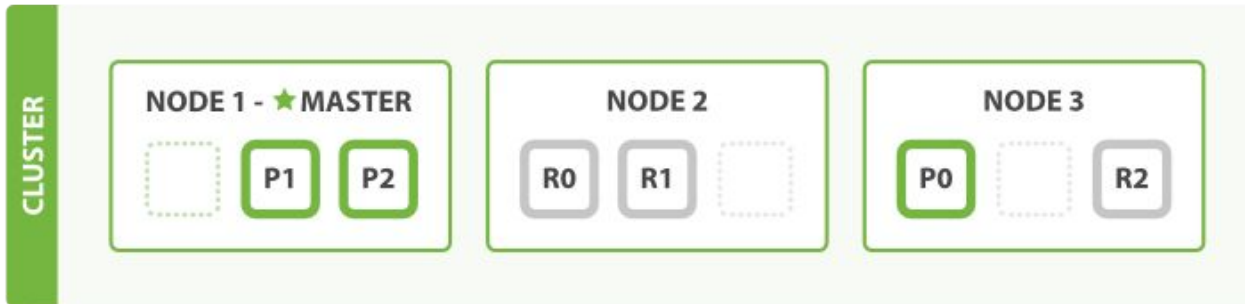


elasticsearch

- scalable, distributed search



Resource consumption
 $\approx \text{query_time} \times \# \text{ shards}$



Building a model

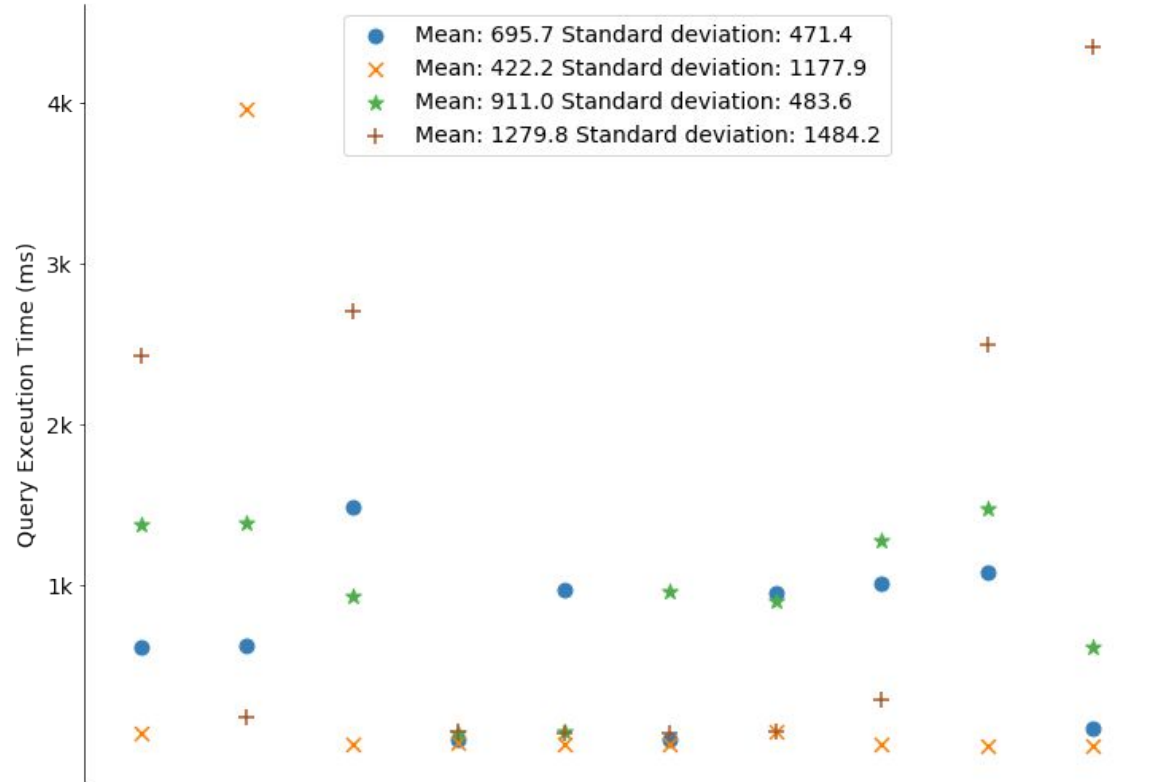
Query	feature ₁	feature ₂	...	feature _n	rc
q ₁	f ₁₁	f ₁₂	...	f _{1n}	r ₁
q ₂	f ₂₁	f ₂₂	...	f _{2n}	r ₂
...					
q _n	f _{n1}	f _{n2}	...	f _{nn}	r _n

Building a model

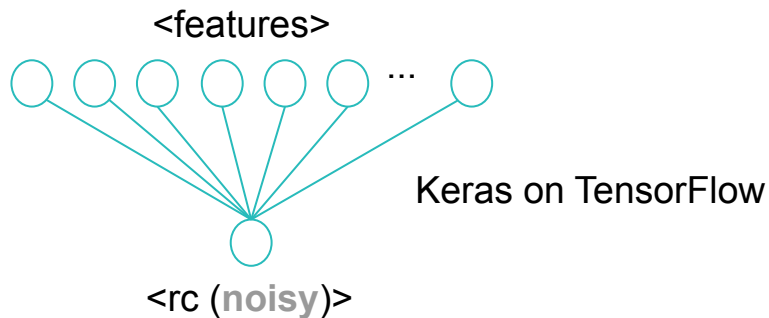
Query	feature ₁	feature ₂	...	feature _n	rc
q ₁	f ₁₁	f ₁₂	...	f _{1n}	r ₁
q ₂	f ₂₁	f ₂₂	...	f _{2n}	r ₂
...					
q _n	f _{n1}	f _{n2}	...	f _{nn}	r _n

Building a model

Query	feature ₁	feature ₂	...	feature _n	rc
q ₁	f ₁₁	f ₁₂	...	f _{1n}	r ₁
q ₂	f ₂₁	f ₂₂	...	f _{2n}	r ₂
...					
q _n	f _{n1}	f _{n2}	...	f _{nn}	r _n



Building a model



Model Strategies

1. Dynamic/online noisy model

```
Features =  
{  
    number_of_words,  
    number_of_wildcards,  
    number_of_ors,  
    number_of_and,  
    number_of_not,  
    ...  
    etc.  
}
```

Building a model

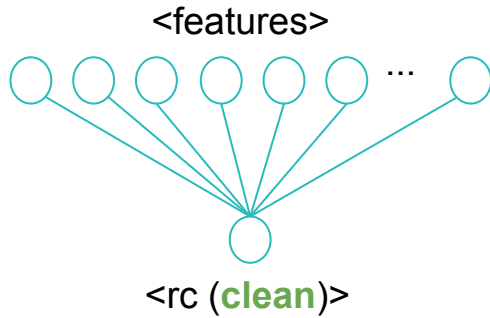
Query	feature ₁	feature ₂	...	feature _n	rc	rc _c
q ₁	f ₁₁	f ₁₂	...	f _{1n}	r ₁	r _{1c}
q ₂	f ₂₁	f ₂₂	...	f _{2n}	r ₂	r _{2c}
...						
q _n	f _{n1}	f _{n2}	...	f _{nn}	r _n	r _{nc}

Clean measurements

Model Strategies

1. Dynamic/online noisy model

Building a model



Model Strategies

1. Dynamic/online noisy model
2. Static/offline clean model

Building a model

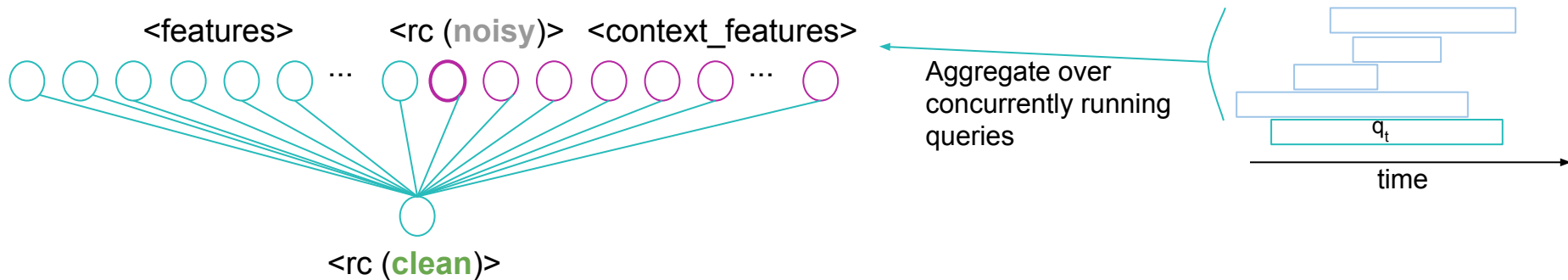
Query	feature ₁	feature ₂	...	feature _n	rc	rc _c	rc _d
q ₁	f ₁₁	f ₁₂	...	f _{1n}	r ₁	r _{1c}	r _{1d}
q ₂	f ₂₁	f ₂₂	...	f _{2n}	r ₂	r _{2c}	r _{1d}
...							
q _n	f _{n1}	f _{n2}	...	f _{nn}	r _n	r _{nc}	r _{1d}

Model Strategies

1. Dynamic/online noisy model
2. Static/offline clean model

Denoised
measurements

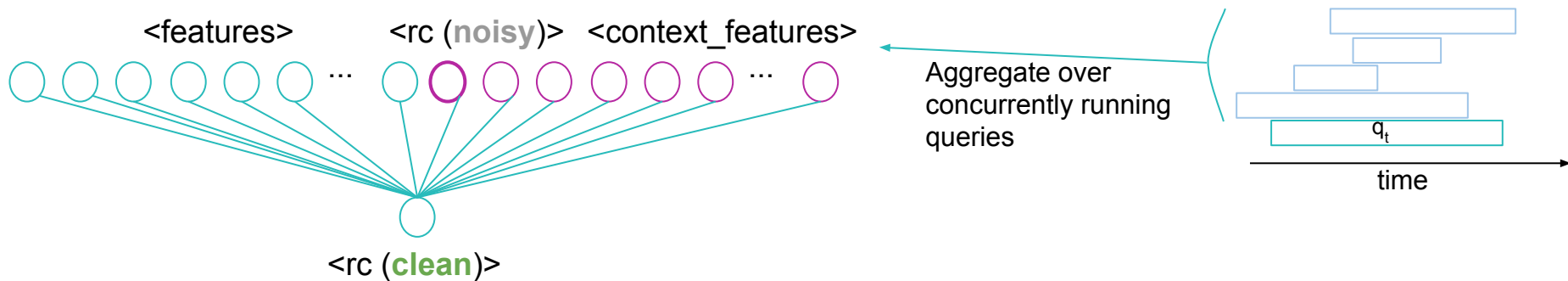
Building a model



Model Strategies

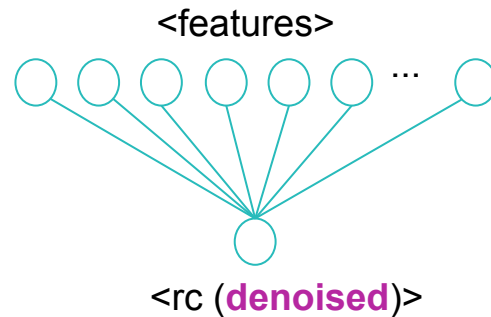
1. Dynamic/online noisy model
2. Static/offline clean model
3. Dynamic/online denoised model
 - Denoise noisy measurements
 - Train on denoised measurements

Building a model

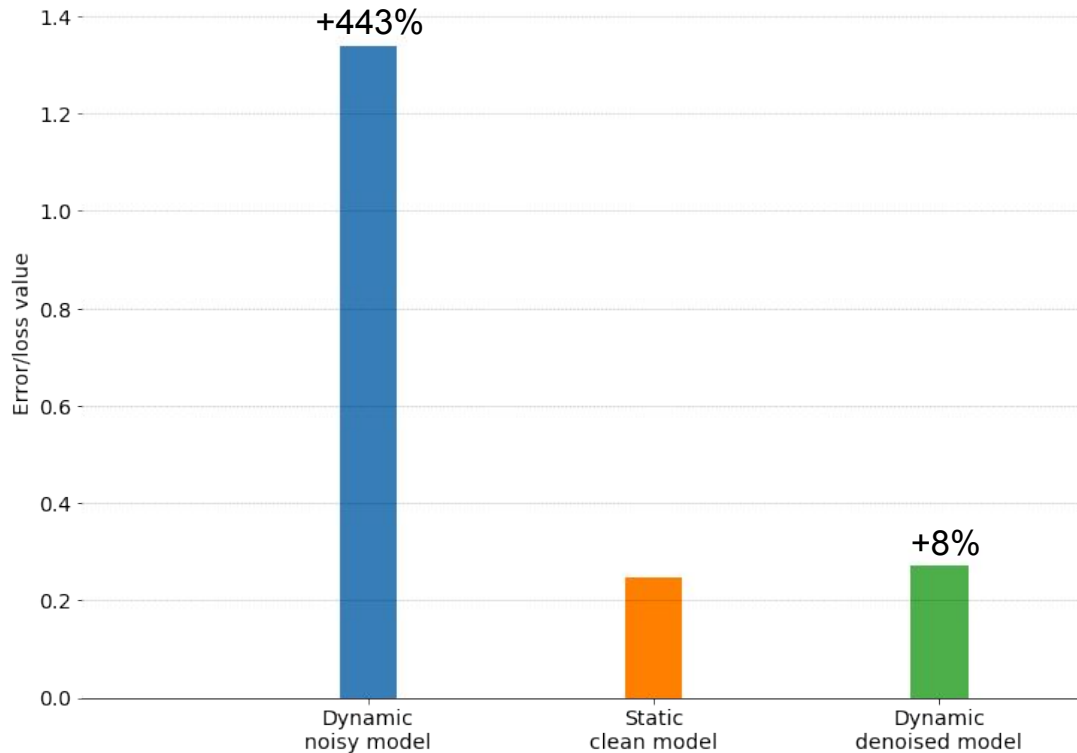


Model Strategies

1. Dynamic/online noisy model
2. Static/offline clean model
3. Dynamic/online denoised model
 - Denoise noisy measurements
 - Train on denoised measurements



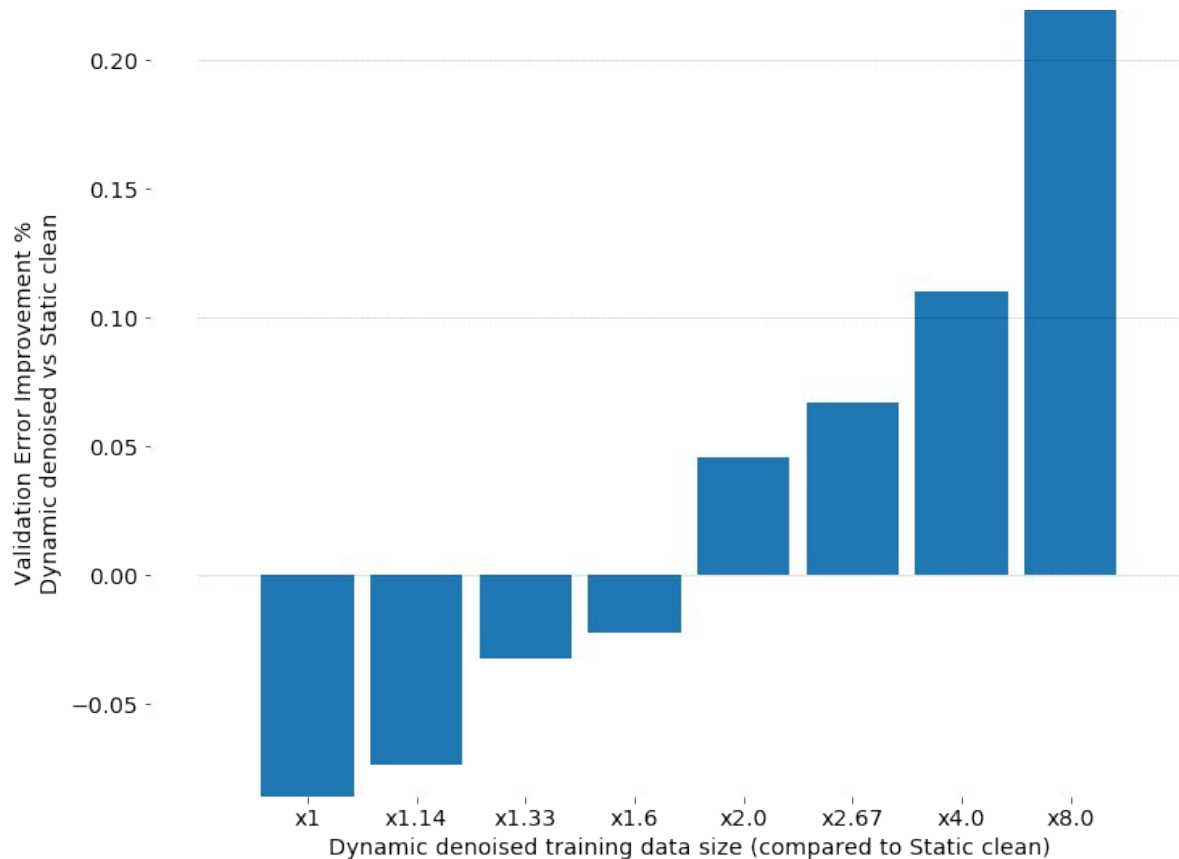
Comparing strategies



Validation Error Comparison

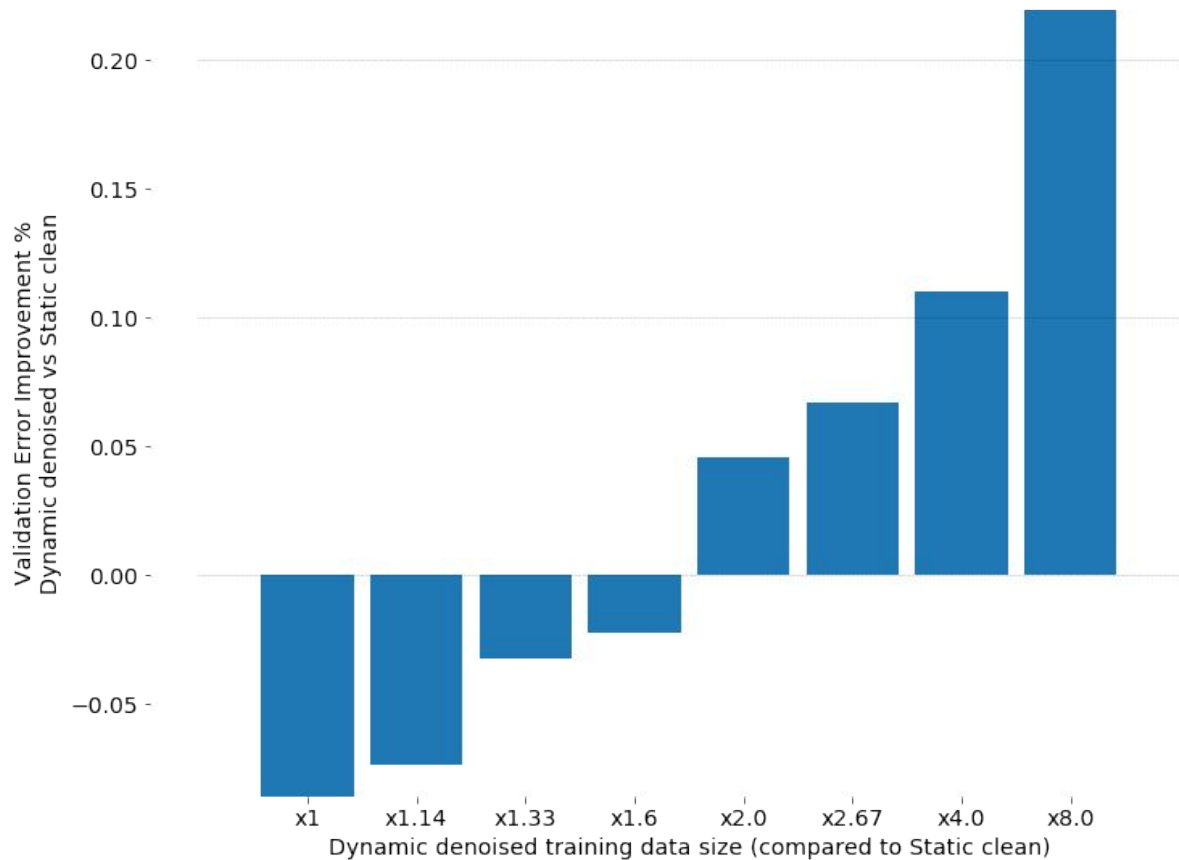
- Fixed training data size
- Fixed number of training cycles/epochs

Comparing strategies



- Static model extremely limited in how much training data can be obtained.
- Dynamic denoised can easily be trained on much larger data sets, to give better results.

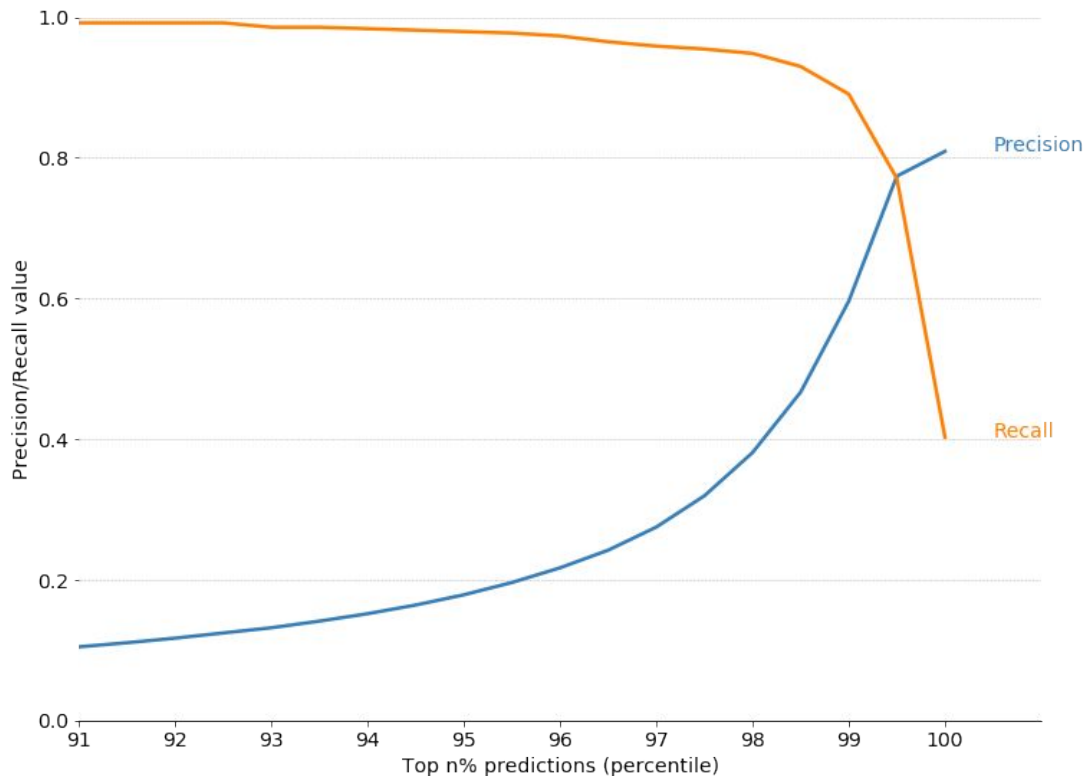
Comparing strategies



Model Strategies

1. Dynamic noisy model
2. Static clean model
3. Dynamic denoised model

Routing decisions



Can we successfully predict the top 1% (heaviest resource consumption) of queries?

Taking the 99th percentile of predictions, gives:
precision = recall = 0.77

- We detect 77% of the heaviest queries

Future work, Less noise, Better results

- New model features
- Network topology
- Robust regression
 - RANdom Sample Consensus (RANSAC)
 - Ensemble methods

Machine Learning in Information Retrieval

- Improved search result relevance based on user-interactions
- Deep ingressor: generation of ingress/summary for documents
- Deep narration: learn what makes a set of documents different from a background set, and generate a narrative that describes these significant differences
- Provide recommendations on top of search results
- Query augmentation and expansion

Acknowledgements

Special thanks to:

- Meltwater Gothenburg
- Team Horace
- [Johan Nilsson Hansen](#)
- [Mikael Johansson](#)



Horace

Q and A

