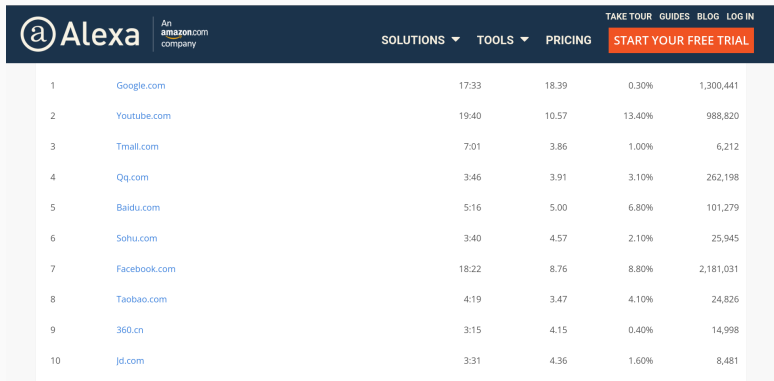


# User-facing ML or LTR is everywhere

---

Senior Machine Learning Programmer at Proxet / Giphy

Take a popular service and review Machine Learning use-cases.




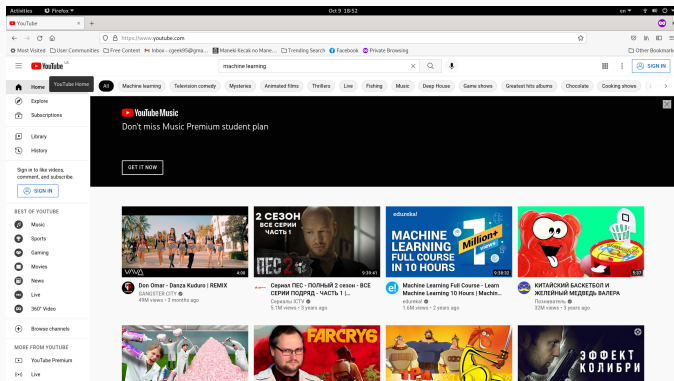
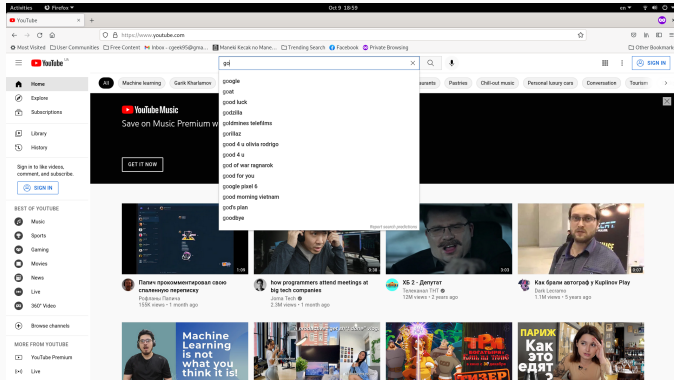
|  <small>An amazon.com company</small> |                              | SOLUTIONS ▾ | TOOLS ▾ | PRICING | <a href="#">START YOUR FREE TRIAL</a> | TAKE TOUR | GUIDES | BLOG | LOG IN |
|--|------------------------------|-------------|---------|---------|---------------------------------------|-----------|--------|------|--------|
| 1  | <a href="#">Google.com</a>   | 17:33       | 18.39   | 0.30%   | 1,300,441                             |           |        |      |        |
| 2  | <a href="#">Youtube.com</a>  | 19:40       | 10.57   | 13.40%  | 988,820                               |           |        |      |        |
| 3  | <a href="#">Tmall.com</a>    | 7:01        | 3.86    | 1.00%   | 6,212                                 |           |        |      |        |
| 4  | <a href="#">Qq.com</a>       | 3:46        | 3.91    | 3.10%   | 262,198                               |           |        |      |        |
| 5  | <a href="#">Baidu.com</a>    | 5:16        | 5.00    | 6.80%   | 101,279                               |           |        |      |        |
| 6  | <a href="#">Sohu.com</a>     | 3:40        | 4.57    | 2.10%   | 25,945                                |           |        |      |        |
| 7  | <a href="#">Facebook.com</a> | 18:22       | 8.76    | 8.80%   | 2,181,031                             |           |        |      |        |
| 8  | <a href="#">Taobao.com</a>   | 4:19        | 3.47    | 4.10%   | 24,826                                |           |        |      |        |
| 9  | <a href="#">360.cn</a>       | 3:15        | 4.15    | 0.40%   | 14,998                                |           |        |      |        |
| 10   | <a href="#">jd.com</a>       | 3:31        | 4.36    | 1.60%   | 8,481                                 |           |        |      |        |

Figure 1: Top websites according to Alex.com

# Youtube - Home Page after 1 search query



# Youtube - Autocomplete



# Youtube - Search

The screenshot shows a web browser window with the address bar displaying 'https://www.youtube.com/results?search\_query=machine+learning'. The search bar contains the text 'machine learning'. The left sidebar shows the YouTube navigation menu with options like Home, Explore, Subscriptions, Library, and History. The main content area displays search results for 'machine learning'. The first result is 'Artificial Intelligence Course - Machine Learning, TensorFlow' by the European IT Certification Framework, with a 'VISIT SITE' button. The second result is 'Machine Learning | Courses' by Coursera, also with a 'VISIT SITE' button. The third result is 'Machine Learning Techniques - Start Learning Today' by edX, with a 'VISIT SITE' button. Below these are video thumbnails. The first video is 'Machine Learning Basics | What is Machine Learning? | Introduction To Machine Learning | Simplilearn' with 2.1M views and a 'VISIT SITE' button. The second video is 'Neural Networks From the ground up' by 3Blue1Brown, with 4.5M views and a 'VISIT SITE' button. The bottom of the page has a small disclaimer about additional funding provided by Google Partners.

Activities Firefox Oct 9 16:52

machine learning - YouTube

https://www.youtube.com/results?search\_query=machine+learning

machine learning

Filters

Artificial Intelligence Course - Machine Learning, TensorFlow

New! ETH Zurich Artificial Intelligence Academy in the European IT Certification Framework.

Web Development Computer Graphics Information Security Business Information

https://www.ethz.ch/en/ VISIT SITE

Machine Learning | Courses

Launch your career with a Machine Learning Certificate from a top program! Andrew Ng's popular introduction to Machine Learning Fundamentals. Learn whether AI? Customer Support.

https://www.coursera.org/ VISIT SITE

Machine Learning Techniques - Start Learning Today

Learn The Most Techniques to Create Machine Learning Algorithms With Data Science Experts. Join...

https://www.edx.org/course/machine-learning/online-course VISIT SITE

2 MILLION+ VIEWS

BASICS OF MACHINE LEARNING

Machine Learning Basics | What is Machine Learning? | Introduction To Machine Learning | Simplilearn

2.1M views · 3 years ago

Simplilearn

Below topics are explained in this Machine Learning basics video: 1. What is Machine Learning? | 2. Types of Machine...

Neural Networks

From the ground up

3BLUE1BROWN SERIES: SS - 01

But what is a neural network? | Chapter 1, Deep learning

4.5M views · 4 years ago

3Blue1Brown

Additional funding for this project provided by Google Partners. Type correction: at 1:14, misspelled 'the last index on'...

# Youtube - Comments

Activities Firefox Oct 9 19:54

Day in the Life of a Data Analyst

https://www.youtube.com/watch?v=0d3jy6E...

Most Visited User Communities Free Content Inbox: ogeed5@yna... Morel Keçan no Mane... Trending Search Facebook Private Browsing Other Bookmarks

Search

160 Comments SORT BY

Add a public comment...

Libba Person 2 months ago  
This is so cool! I'm glad that you look so calm and free despite the busy day. Excited for you to travel the world! There's nothing better than working freelance or having your own business - you own your life. :)

👍 1 🗨️ 1 REPLY

Osair Bull 1 month ago  
Much more relatable than someone who wakes up at 5 or 6, makes it more entertaining and useful insight into not being the only one!

👍 1 🗨️ 1 REPLY

View reply from Stefanovic

Key Brown 1 month ago  
I recently started watching you after your review and I can definitely see a huge difference. This was absolutely entertaining, relatable, and we are starting to see more of your personality. I'm not part of Maths course group (it's been explained prior through a friend) and I gotta say your fruits of your labor are showing! Great work!

Read more

👍 1 🗨️ 1 REPLY

View reply from Stefanovic

Joan Steward 1 month ago  
That's really cool how you documented all of this. I came in just looking for data analyst information but you made me want to get more done throughout the day

👍 1 🗨️ 1 REPLY

Quinn Parker 1 month ago  
I love your videos and the professional quality they have. You seem to have a nice life, and a great diet/fitness regime. It's good to see a more personal side to your channel. You deserve way more than 3K subs.

👍 1 🗨️ 1 REPLY

rebecca06 1 month ago  
I love your videos and the professional quality they have. You seem to have a nice life, and a great diet/fitness regime. It's good to see a more personal side to your channel. You deserve way more than 3K subs.

👍 1 🗨️ 1 REPLY

Shashank Kulkarni  
862K views · 8 months ago

Day in the life of a data analyst  
Luka Barotovic  
33K views · 7 months ago

A Day in the Life of a Software Engineer... WTF  
Pavle Katic  
1.7M views · 1 month ago

How to Become a Data Analyst  
360 Data Science  
400K views · 2 years ago

Day in the Life of a Millionaire Day Trader  
Live Trader  
174K views · 3 months ago

Best advice I have EVER heard  
Stefanovic  
1.9K views · 5 months ago

a day in the life of an engineer working from home  
Jenna Tech  
8.8M views · 4 months ago

Get a JOB w/ Google Data Analytics Certificate!!! (It...  
Luka Barotovic  
119K views · 3 months ago

Day in My Life as a CORPORATE LAWYER IN...  
Liam Perini  
238K views · 1 month ago

# Classification of Machine Learning Problems [Part 1]

- Learning Problems
- Hybrid Learning Problems
- Statistical Inference
- Learning Techniques

# Classification of Machine Learning Problems [Part 2]

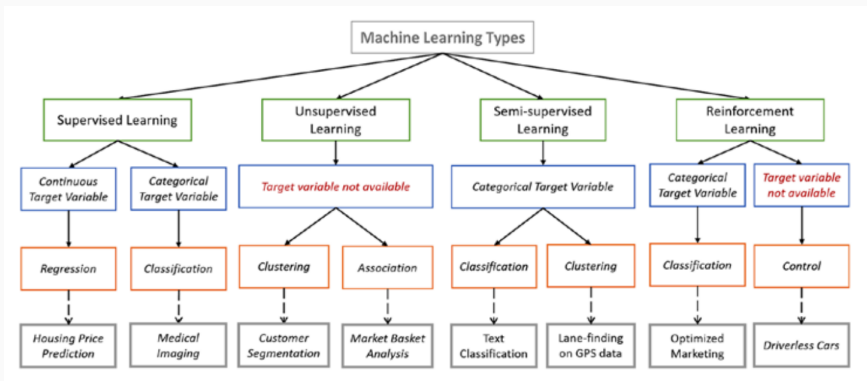


Figure 2: Types of Machine Learning Algorithms



# Youtube Services

- [Personalized] Home Page Content Selection
- [Personalized] Global Search
- [Personalized] Channel Search
- [Personalized] Search Autocomplete
- [Personalized] Related Videos
- [Personalized] Playlists
- [Personalized] Trending
- [Personalized] Notifications
- [Personalized] Comments selection
- Copyright Violation
  - By Performer
  - By Music
  - By Text
  - Abusive speech
  - Violence
- [Personalized] ADs
- Security

# Learning to Rank

LTR is everywhere, where we have a list of elements.



# Approaches to Learning to Rank

- Pointwise (Regression/Classification)
- Pairwise (LambdaRank IR-SM, Lambda Rank)
- Listwise (Soft Rank, SmoothRank, AdaRank, ListNet, BoltzRank)

- Candidate Generation
- Offline ranking
- Online ranking
- Data Collection
- Data debiasing
- A/B testing

# What is LTR

Learning to rank or machine-learned ranking (MLR) is the application of machine learning, typically supervised, semi-supervised or reinforcement learning, in the construction of ranking models for information retrieval systems

- Candidate Generation
- Offline ranking
- Online ranking
- Data Collection
- Data debiasing
- A/B testing

## Search Dataset - Click-Based

| session_id | query            | document_id | relevance |
|------------|------------------|-------------|-----------|
| 1          | machine learning | 1           | 0.0       |
| 1          | machine learning | 2           | 0.0       |
| 1          | machine learning | 3           | 0.0       |
| 1          | machine learning | 4           | 0.0       |
| 1          | machine learning | 5           | 1.0       |

## Search Dataset - Click-Based

| session_id | query            | document_id | relevance | position |
|------------|------------------|-------------|-----------|----------|
| 1          | machine learning | 1           | 0.0       | 1        |
| 1          | machine learning | 2           | 0.0       | 2        |
| 1          | machine learning | 3           | 0.0       | 3        |
| 1          | machine learning | 4           | 0.0       | 4        |
| 1          | machine learning | 5           | 1.0       | 5        |

# Search Dataset - Human Relevance

| session_id | query            | document_id | relevance |
|------------|------------------|-------------|-----------|
| 1          | machine learning | 1           | 3.0       |
| 1          | machine learning | 2           | 2.0       |
| 1          | machine learning | 3           | 1.0       |
| 1          | machine learning | 4           | 4.0       |
| 1          | machine learning | 5           | 5.0       |



$$PairAccuracy = \sum_{i < j} [rel_i > rel_j] \quad (1)$$

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)} \quad (2)$$

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (3)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (4)$$

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}. \quad (5)$$

## Listing 1: Train CatBoost LTR with YetiRankPairwise loss

---

```
import catboost
cb_pool_train = catboost.Pool(
    [[0.0], [0.0], [0.0], [1.0]],
    label=[0.0, 0.0, 0.0, 1.0],
    group_id=[1, 1, 1, 1]
)

cb_pool_eval = catboost.Pool(
    [[0.0], [0.0], [0.0], [2.0]],
    label=[0.0, 0.0, 0.0, 1.0],
    group_id=[1, 1, 1, 1]
)
params = {
    "loss_function": "YetiRankPairwise",
    "custom_metric": ["NDCG"]
}
cb_model_ltr = catboost.CatBoost(params=params)
cb_model_ltr.fit(cb_pool_train, eval_set=cb_pool_eval)
```

---

1. split by query
2. backtesting

# Clicks vs Explicit feedback

1. Explicit feedback is not vulnerable to spam.
2. Explicit feedback can be outdated.
3. Explicit feedback has biases, which you can control with a user's manual.
4. Explicit feedback has fewer contradictions, but almost all the time, the size of a dataset is much smaller..
5. You must mimic query distribution in your dataset.
6. When you start a new product, you don't have enough clicks.
7. Explicit feedback is expensive, and you have to update the test questions very often.

1.  $f_1(\text{document\_id})$  - knows about document modalities
2.  $f_2(\text{session\_id})$  - knows about user id, location, previous actions
3.  $f_3(\text{query})$
4.  $f_4(\text{query}, \text{document\_id})$
5.  $f_5(\text{query}, \text{session\_id})$
6.  $f_6(\text{session\_id}, \text{document\_id})$

# Search Dataset - Feature Examples

1. textual similarity (BM25, BM25 + Stemming, vector-space models)
2. document-level CTR from search logs
3. document-level CTR from recommendations logs
4. different counter aggregations
5. Word2Vec on clicks (StartSpace implementation)
6. 'PageRank'
7. fraud detection
8. OCR + textual similarity
9. speech recognition + textual similarity

The gradient boosting with YetiRankPairwise loss function gives the best results. Neural Nets are not as good, but you can build a lot of great features with Siamese neural networks. You can get ideas from Question-Answering models. Pinterest uses a simple Convolutional Neural Net, but they haven't tried the CatBoost with YetiRankPairwise loss.

You need at least two table:

1. Requests and Responses (timestamp, query, document\_id, position, response\_id, location, user\_id, device\_id, session\_id)
2. Actions (timestamp, response\_id, action\_id, document\_id)

You can use Column-oriented database and a queue like kafka to collect the logs. Also, you can duplicate the data on S3.

- you can train different models for different regions or encode a location in your features
- you can have different search algorithm for different user buckets
- user-based personalization kills caching
- you can cache heavy features from the head of your search log



# Search vs Recommendations

- almost identical approach is applicable for Recommendation Systems, but the document id plays the role of a query.
- you should reuse document-level features
- you can recompute the same feature, but on different datasets
- recommendations are almost identical to search, but for relatively small collections you can use a heavy artillery, because you can store predictions in a relatively small matrix
- you can use top search predictions to generate recommendations

## LTR for every product

| Task                | Query        | Document           | Personalizable | A/B test Target |
|---------------------|--------------|--------------------|----------------|-----------------|
| Search              | Query String | Document Id        | +              | CTR/MRR         |
| Recommendations     | Document Id  | Document Id        | +              | CTR/MRR         |
| Home Page Documents | DateTime     | Document Id        | +              | CTR/MRR         |
| Home Page Queries   | DateTime     | Query String       | +              | CTR/MRR         |
| Search Autocomplete | Query String | Query String       | +              | CTR/MRR         |
| Ranking Comments    | Document Id  | Comment + Metadata | +              | Likes           |

1. [Search Quality Rating Guidelines](#)
2. [Introduction to Information Retrieval](#)
3. [Information Retrieval: Implementing and Evaluating Search Engines](#)
4. [Demystifying Core Ranking in Pinterest Image Search](#)
5. [Catboost usage example.](#)
6. [StarSpace](#)
7. [Personalized Trending Search Suggestions](#)
8. [Winning The Transfer Learning Track of Yahoo!'s Learning To Rank Challenge with YetiRank](#)
9. [On NDCG Consistency of Listwise Ranking Methods](#)
10. [GeoTrend: Spatial Trending Queries on Real-time Microblogs](#)
11. [Finding Trending Local Topics in Search Queries](#)
12. [Challenges in building large-scale information retrieval systems: invited talk](#)
13. [SIMD-Based Decoding of Posting Lists](#)