




# Learning to Rank

Jasmine Wilkerson

# Overview

- Introduction to Ranking Problem
  - Using Information Retrieval (IR)-Document retrieval as example.
- General setup of a Ranking Model
- Quick recap of loss function and SVM
- Approaches
  - Pointwise
  - Pairwise
  - Listwise
- Further reading recommendation


# Example

[Web](#) [Images](#) [News](#) [Videos](#) [Maps](#) [More ▾](#) [Search tools](#)

---


About 1,360,000,000 results (0.43 seconds)

**Dog - Wikipedia, the free encyclopedia**  
<https://en.wikipedia.org/wiki/Dog>  Wikipedia ▾  
The domestic **dog** (*Canis lupus familiaris* or *Canis familiaris*) is a domesticated canid which has been selectively bred for millennia for various behaviors, ...  
[Origin of the domestic dog](#) - [Man's best friend \(phrase\)](#) - [List of dog breeds](#) - [Breed](#)

**Dog Supplies | Dog Accessories & Dog Products - Dog.com**  
[www.dog.com/](http://www.dog.com/) ▾  
Dog.com is your source for **dog** supplies! We carry high quality **dog** food, **dog** beds, **dog** treats & other **dog** products at great low prices!

**Dog Health Center | Dog Care and Information from WebMD**  
[pets.webmd.com/dogs/](https://pets.webmd.com/dogs/) ▾ WebMD ▾  
Welcome to the new WebMD **Dog** Health Center. WebMD veterinary experts provide comprehensive information about **dog** health care, offer nutrition and ...

**In the news**








**Woman adopts dying dog, treats him to bucket list**  
[CNN International](#) - 4 hours ago  
(CNN) Nicole Elliott was browsing a Georgia animal shelter's Facebook page two weeks ...

[Caught on Camera: UPS Driver Kicks Dog](#)  
[NBC News](#) - 1 hour ago

[6-Year-Old Boy Killed by Dog in Western North Carolina](#)  
[ABC News](#) - 2 hours ago

[More news for dog](#)

**Dog: Dog Breeds, Adoption, Bringing a Dog Home and Care**  
<https://www.petfinder.com/dogs/> ▾ Petfinder ▾



[More images](#)

## Dog






Animal

The domestic dog is a domesticated canid which has been selectively bred for millennia for various behaviors, sensory capabilities, and physical attributes. [Wikipedia](#)

**Scientific name:** *Canis lupus familiaris*  
**Gestation period:** 63 d  
**Lifespan:** 13 y (dying of natural causes, UK population)  
**Height:** 2.2 – 2.3 ft. (At Shoulder)  
**Daily sleep:** 10 h  
**Rank:** Subspecies

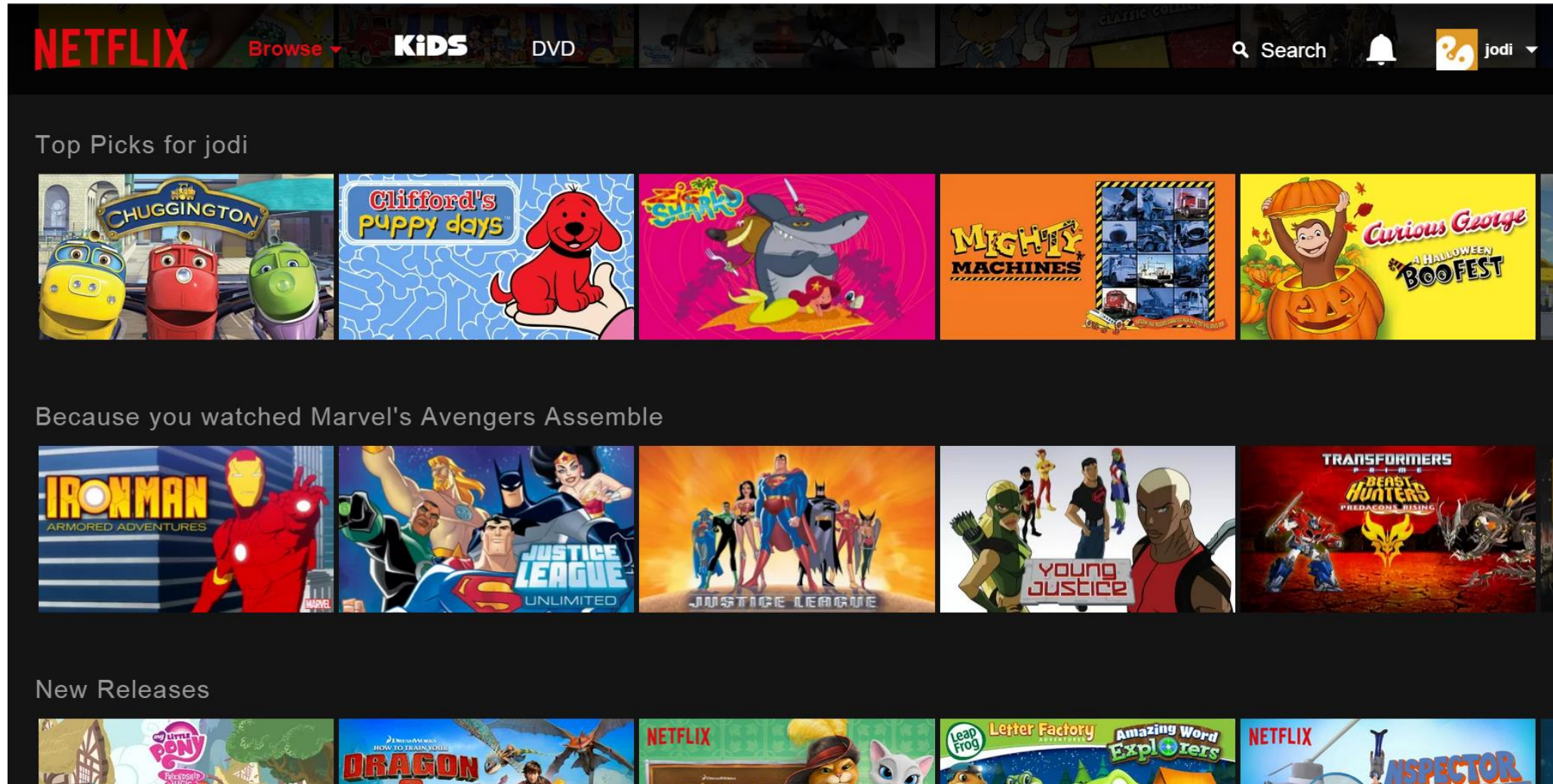
### Breeds

[Labrador Retriever](#) [German Shepherd](#) [Golden Retriever](#) [Yorkshire Terrier](#) [Bulldog](#) [View 15+ more](#)



[Feedback](#)

# Example



# Example

The screenshot shows the LinkedIn search interface for the job title 'data scientist'. The top navigation bar includes the LinkedIn logo, a 'PREMIUM' badge, a search bar with 'data scientist' entered, and an 'Advanced' search link. Below the navigation bar, the search results are displayed. On the left, there is a sidebar with filters: 'Advanced' (with a chevron), 'All', 'Jobs' (selected), and 'More...'. Below these are filter sections for 'Keywords' (with a text input containing 'data scientist'), 'Company' (with an empty text input), 'Title' (with an empty text input), 'Location' (with a dropdown menu showing 'Located in or near:'), 'Country' (with a dropdown menu showing 'United States'), and 'Postal Code' (with an empty text input). The main search results area shows '36,431 results for data scientist' and a 'Sort by Relevance' dropdown. The results list includes three job postings from Amazon and CyberCoders, each with a 'View' button. A filter banner at the bottom of the results list asks 'Only show jobs in Greater Seattle Area?' with a 'Filter' button and a help icon. The bottom of the results list shows a job posting from 'Context Relevant' with a 'View' button.

in PREMIUM data scientist x Advanced

Home Profile Connections Jobs Interests

**Search** 36,431 results for data scientist Sort by Relevance

Advanced >

All  
**Jobs**  
More...

**Keywords**  
data scientist

**Company**

**Title**

**Location**  
Located in or near: ▾

**Country**  
United States ▾

**Postal Code**

**Software Development Engineer, Big Data, NLP, Data Mining**  
Amazon  
US-WA-Seattle • Jul 7, 2015  
[Similar](#) **View**

**Software Development Engineer, Big Data, NLP, Data Mining**  
Amazon  
US-WA-Seattle • Jul 7, 2015  
[Similar](#) **View**

**Data Scientist - Data Mining, Big Data, Machine Learning**  
CyberCoders  
Santa Monica, CA • Jul 8, 2015  
[Similar](#) **View**

**Only show jobs in Greater Seattle Area?** **Filter** ?

**Data Scientist**  
Context Relevant  
Seattle, NYC • Jul 8, 2015  
[Similar](#) **View**

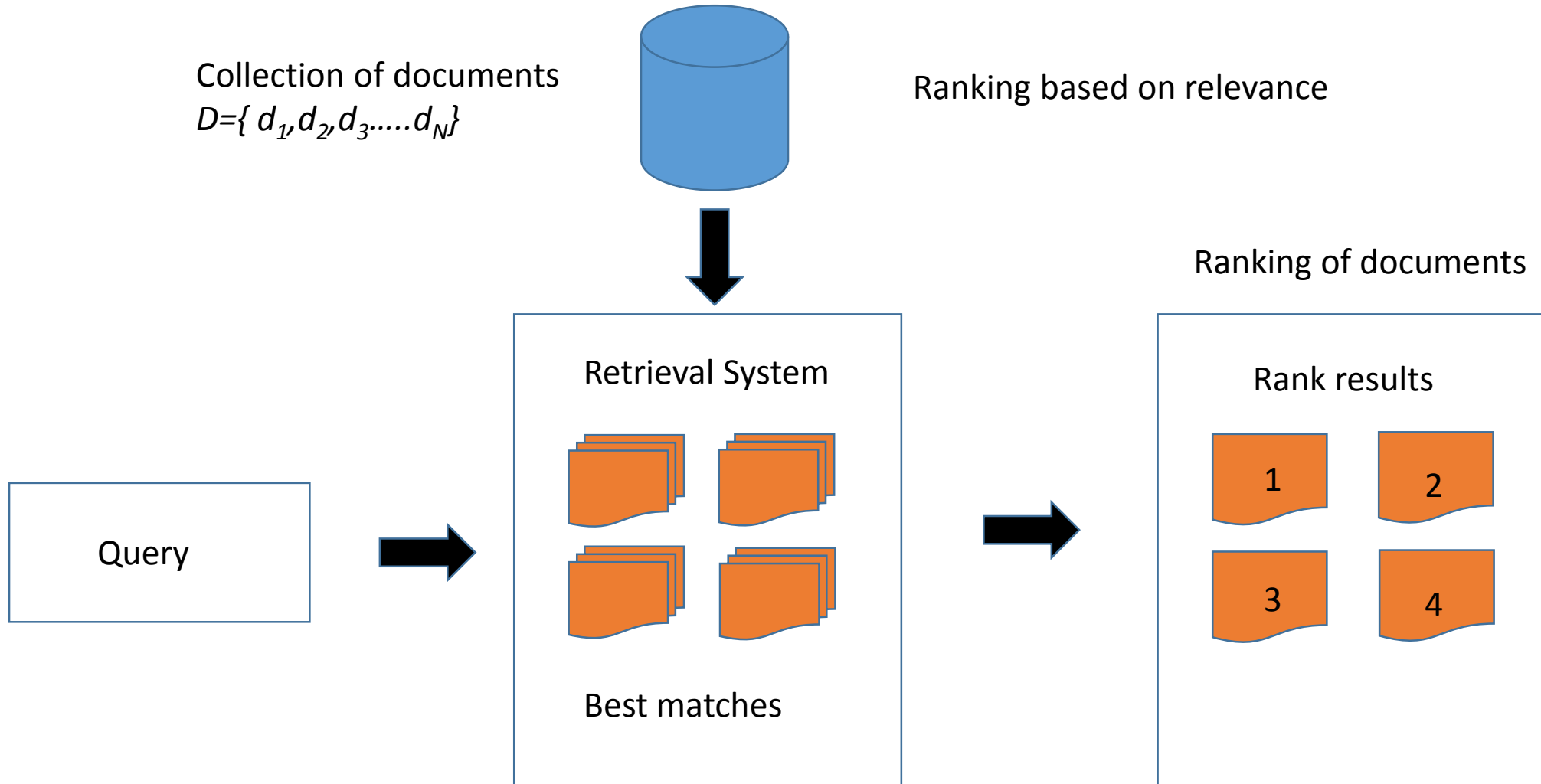
# Introduction

- Learning to rank is a machine learning techniques for training model in ranking task
- Learning to rank can be employed to variety to application such as Information Retrieval (IR), Natural Language Processing (NLP), Data Mining (DM)
- Eg. Document retrieval which will be used as example through-out the presentation

# Task of Retrieval and Ranking

- Given a query, system retrieves documents containing query words from document collection, ranks the documents and returns top ranked documents

# Document Retrieval





# Traditional Ranking Model

- Traditional ranking model  $f(q,d)$  is created without training
- Eg. BM25
  - Conditional probability of relevancy of query to documents  
 $P(r/q,d)$        $r = 1$  when relevant ;  $r = 0$  when irrelevant
- Eg. Language Model for IR (LMIR)
  - conditional probability distribution by calculated the words appearing in the query and document  
 $P(q,d)$

# New Ranking Model Approach

- New trend, particularly in web search is to employ machine learning techniques to automatically construct the ranking model  $f(q, d)$
- Motivated by existence of many signals to represent relevance in web search
  - Anchor text, PageRank score of a webpage
- These features are incorporated into the ranking model.

# Other Features

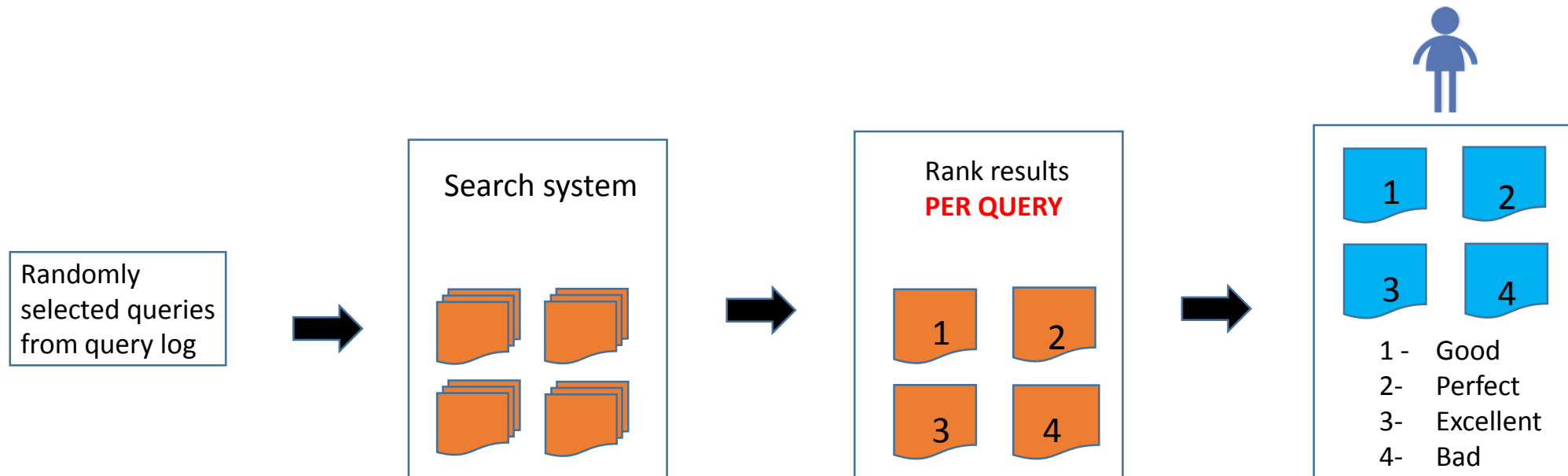
- Modern systems – especially on the Web – use a great number of features:
  - Log frequency of query word in anchor text?
  - Query word in color on page?
  - # of images on page?
  - # of (out) links on page?
  - PageRank of page?
  - URL length?
  - URL contains “~”?
  - Page length?
- The *New York Times* (2008-06-03) quoted Amit Singhal as saying Google was using over 200 such features.

# Learning to Rank Model

- Supervised learning
- Labeled Dataset
  - Split into training and testing sets
- Algorithm
- Evaluate

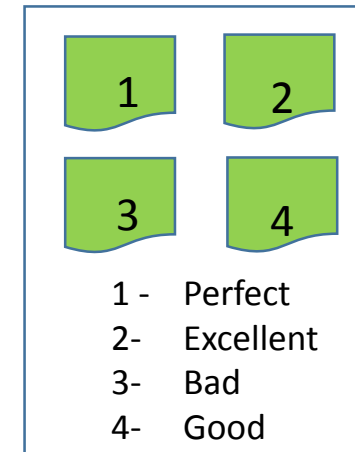
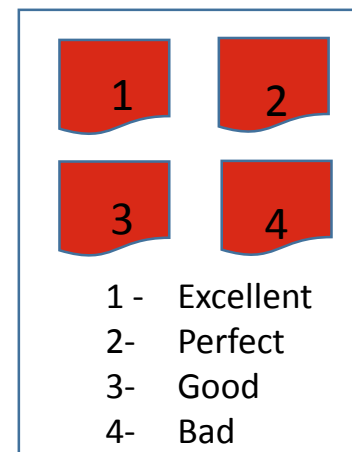
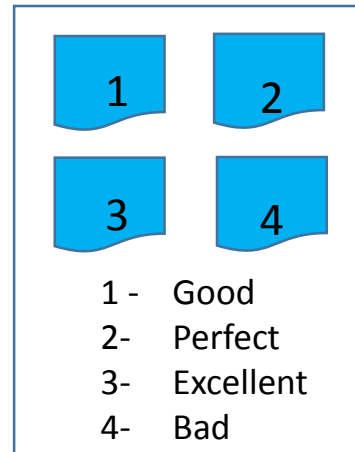
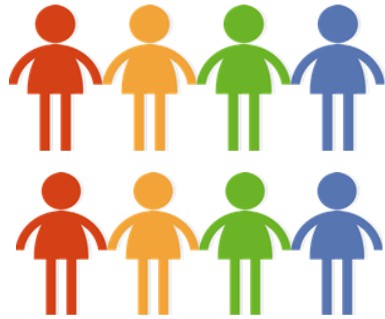
# Obtain Labeled Data

- Explicit feedback by human judgements



# Obtain Labeled Data

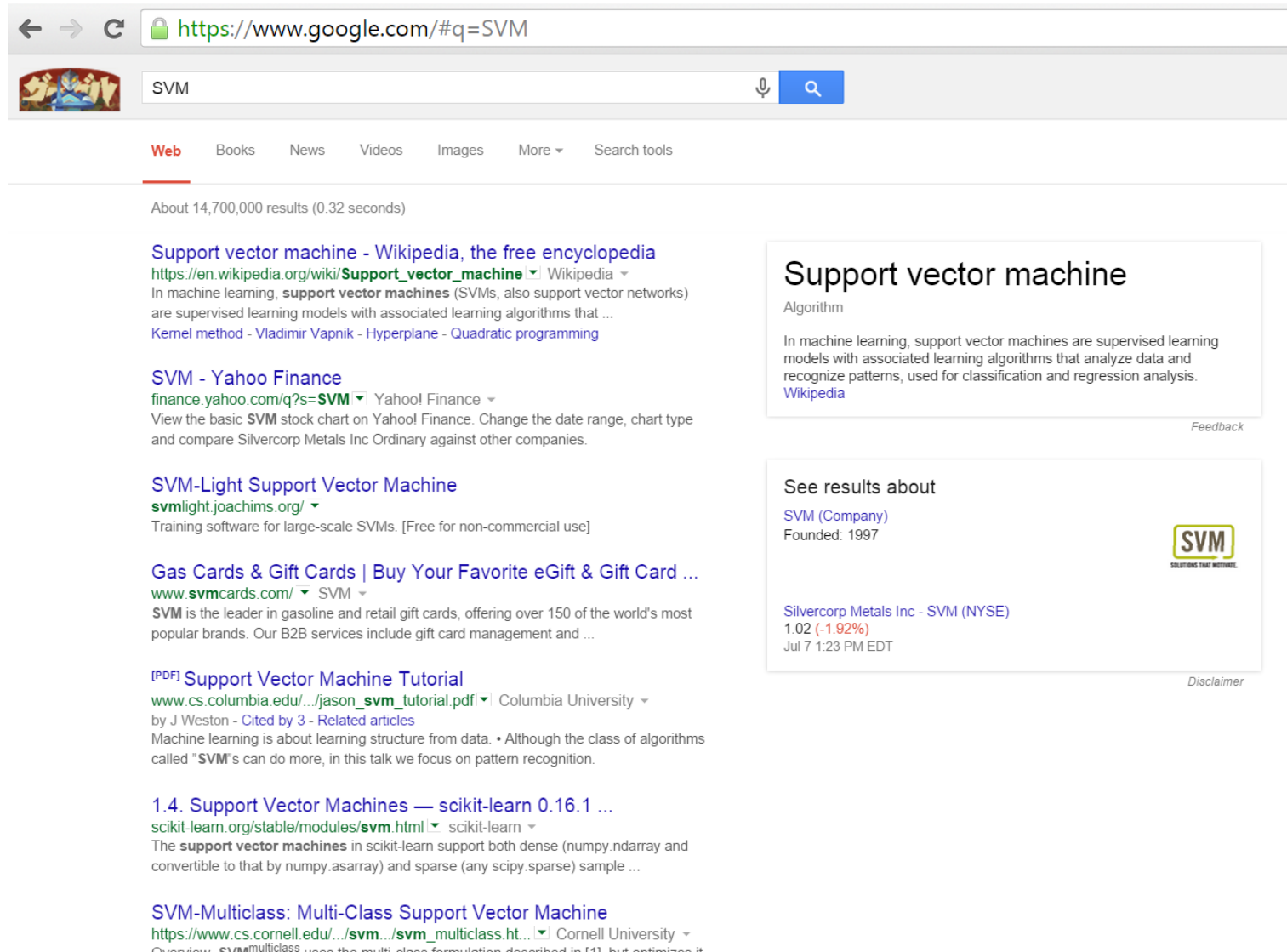
PER QUERY



...

- Labels representing relevance are assigned to the query documents pairs
- Conduct majority voting with multiple judges.
- Expensive
- Overhead for users

# How would you judge the rank?




The screenshot shows a Google search results page for the query "SVM". The browser address bar displays "https://www.google.com/#q=SVM". The search bar contains "SVM" and a microphone icon. Below the search bar, navigation tabs include "Web", "Books", "News", "Videos", "Images", "More", and "Search tools". The results section indicates "About 14,700,000 results (0.32 seconds)".

The search results are as follows:

- Support vector machine - Wikipedia, the free encyclopedia**  
[https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine) | Wikipedia  
In machine learning, **support vector machines** (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that ...  
[Kernel method](#) - [Vladimir Vapnik](#) - [Hyperplane](#) - [Quadratic programming](#)
- SVM - Yahoo Finance**  
<finance.yahoo.com/q?s=SVM> | Yahoo! Finance  
View the basic **SVM** stock chart on Yahoo! Finance. Change the date range, chart type and compare Silvercorp Metals Inc Ordinary against other companies.
- SVM-Light Support Vector Machine**  
<svmlight.joachims.org/>  
Training software for large-scale SVMs. [Free for non-commercial use]
- Gas Cards & Gift Cards | Buy Your Favorite eGift & Gift Card ...**  
<www.svmcards.com/> | SVM  
SVM is the leader in gasoline and retail gift cards, offering over 150 of the world's most popular brands. Our B2B services include gift card management and ...
- [PDF] Support Vector Machine Tutorial**  
[www.cs.columbia.edu/.../jason\\_svm\\_tutorial.pdf](http://www.cs.columbia.edu/.../jason_svm_tutorial.pdf) | Columbia University  
by J Weston - [Cited by 3](#) - [Related articles](#)  
Machine learning is about learning structure from data. • Although the class of algorithms called "**SVM**"s can do more, in this talk we focus on pattern recognition.
- 1.4. Support Vector Machines — scikit-learn 0.16.1 ...**  
[scikit-learn.org/stable/modules/svm.html](http://scikit-learn.org/stable/modules/svm.html) | scikit-learn  
The **support vector machines** in scikit-learn support both dense (numpy.ndarray and convertible to that by numpy.asarray) and sparse (any scipy.sparse) sample ...
- SVM-Multiclass: Multi-Class Support Vector Machine**  
[https://www.cs.cornell.edu/.../svm.../svm\\_multiclass.ht...](https://www.cs.cornell.edu/.../svm.../svm_multiclass.ht...) | Cornell University  
[Overview](#) SVMmulticlass uses the multi-class formulation described in [1] but optimizes it

On the right side of the page, there are two summary cards:

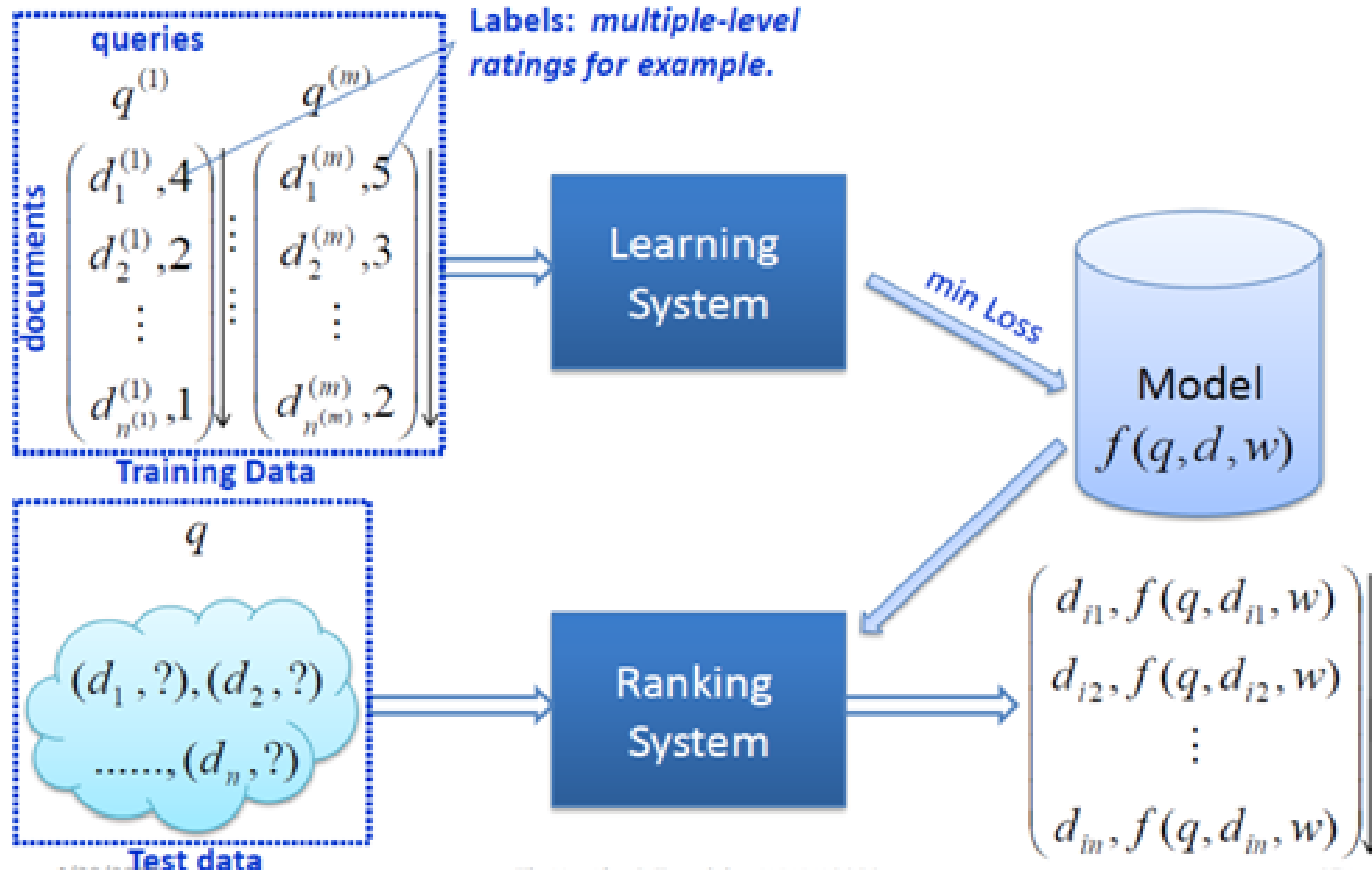
- Support vector machine**  
Algorithm  
In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.  
[Wikipedia](#)  
[Feedback](#)
- See results about**  
[SVM \(Company\)](#)  
Founded: 1997  
  
[Silvercorp Metals Inc - SVM \(NYSE\)](#)  
1.02 (-1.92%)  
Jul 7 1:23 PM EDT  
[Disclaimer](#)

# Obtain Labeled Data

- Implicit feedback derived from log search data
- Eg. Click-Through-Rate, mousing, scrolling, eye-tracking
- No need to hire human judge
- No overhead for users
- More difficult to interpret



# Training and Testing



# Evaluation

- Discounted Cumulative Gain (DCG)
- Normalized Discounted Cumulative Gain (NDCG)

# Discounted Cumulative Gain (DCG)

- Uses graded relevance as a measure usefulness from a document
- Gain is accumulated starting at the top of the ranking, or discounted, at lower ranks
- DCG is the total gain accumulated at a particular rank  $p$ :

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i)}$$

# DCG Example

$$\text{DCG}_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i)}$$

- 10 ranked documents judged on 0-3 relevance scale:  
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- Discounted Gain (DC):
  - 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0  
= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- Discounted Cumulative Gain:
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

# IDCG and nDCG

- Idealized discounted cumulative gain (IDCG)
- The ideal ranking returns document with the highest relevance level, follow by next highest relevance level, etc.
- Normalized Discounted Cumulative Gain (nDCG)

$$\text{nDCG}_p = \frac{DCG_p}{IDCG_p}$$

# Example of nDCG

4 documents:  $d_1, d_2, d_3, d_4$

i	Ground Truth		Ranking Function <sub>1</sub>		Ranking Function <sub>2</sub>	
	Document Order	$r_i$	Document Order	$r_i$	Document Order	$r_i$
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	NDCG <sub>GT</sub> =1.00		NDCG <sub>RF1</sub> =1.00		NDCG <sub>RF2</sub> =0.9203	

$$\text{nDCG}_p = \frac{DCG_p}{IDCG_p}$$

$$DCG_{GT} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left( \frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left( \frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

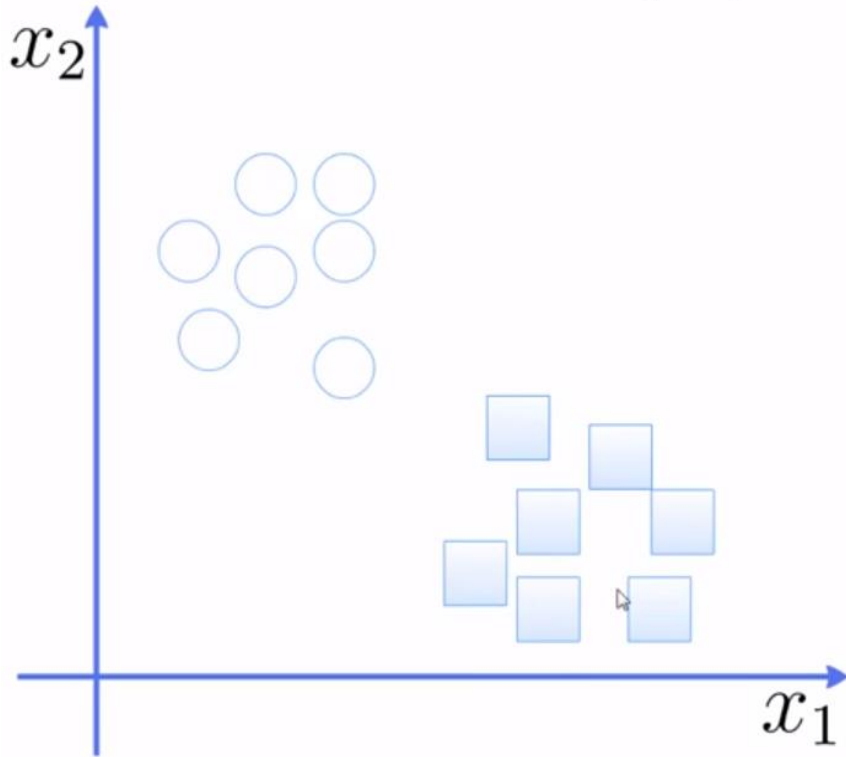
$$\text{MaxDCG} = DCG_{GT} = 4.6309$$

# Recap

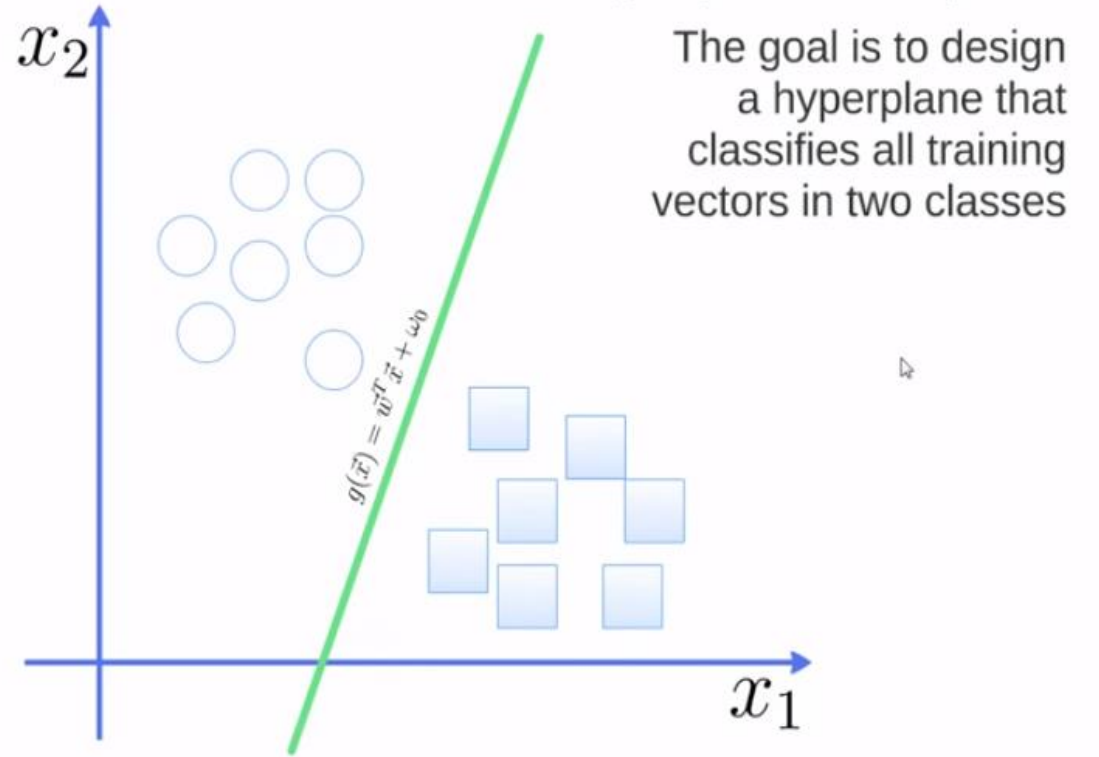
- Relationship of Hypothesis Function and Loss Function
- Recap of SVM

# SVM recap

SVM for linearly separable binary sets



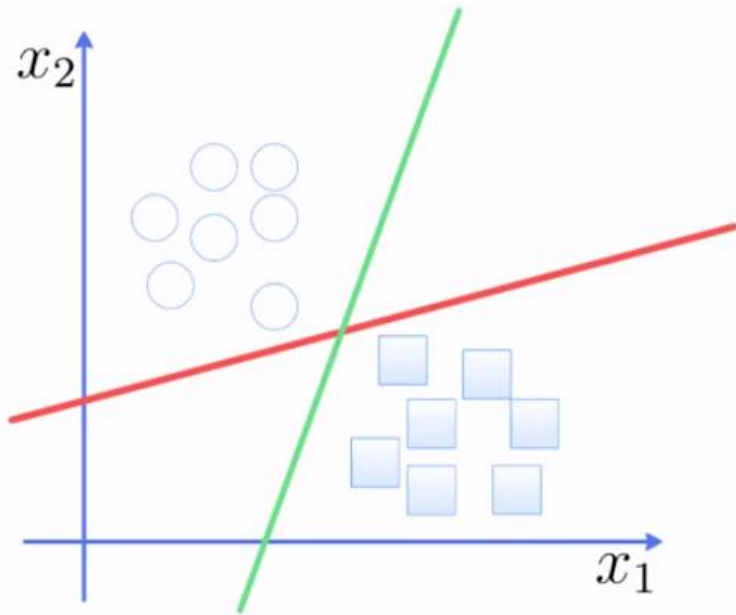
SVM for linearly separable binary sets



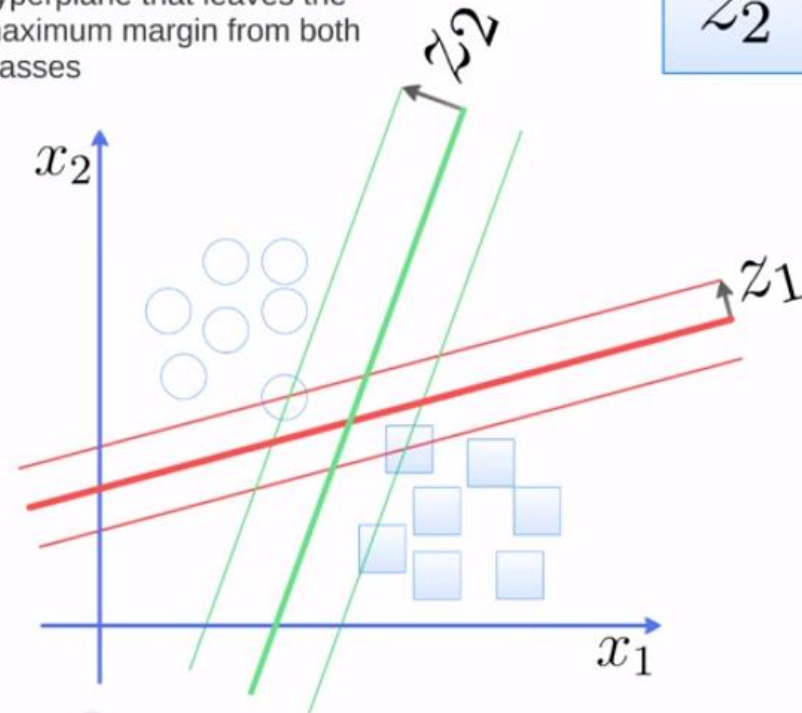


# SVM recap

The best choice will be the hyperplane that leaves the maximum margin from both classes



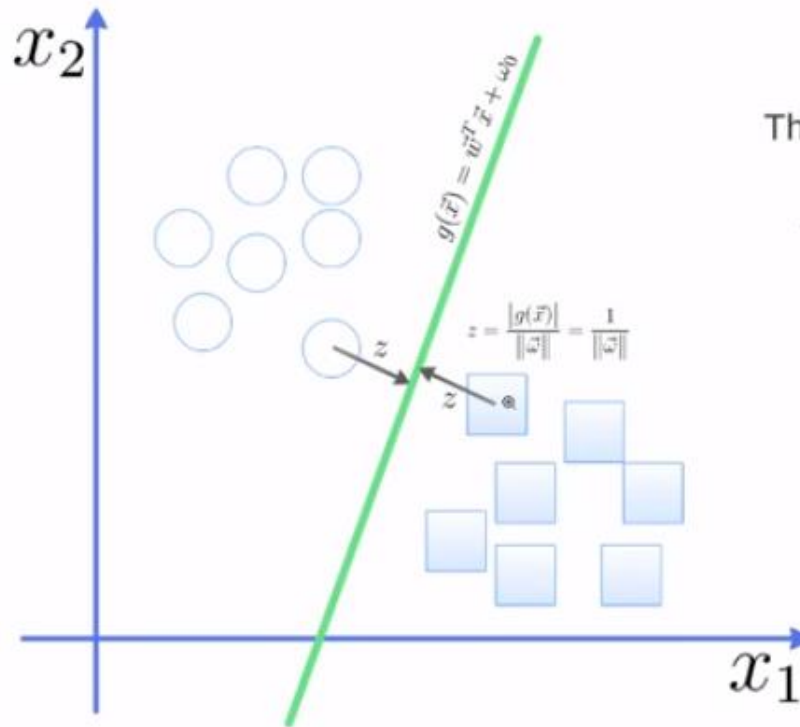
The best choice will be the hyperplane that leaves the maximum margin from both classes



# SVM recap

$$g(\vec{x}) \geq 1, \quad \forall \vec{x} \in \text{class 1}$$

$$g(\vec{x}) \leq -1, \quad \forall \vec{x} \in \text{class 2}$$



The total margin is computed by

$$\frac{1}{\|\vec{w}\|} + \frac{1}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|}$$

Minimizing this term  
will maximize the  
separability

# Learning to Rank Approach

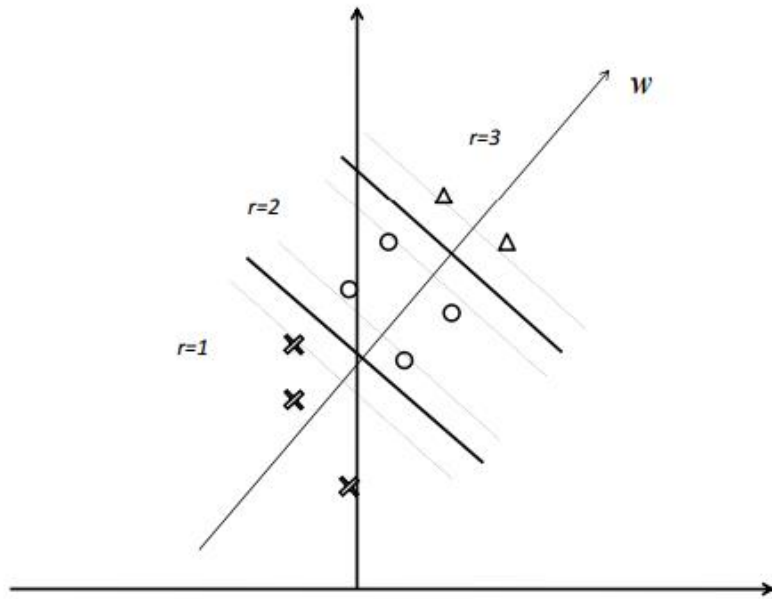
- Pointwise
- Pairwise
- Listwise

# Pointwise Approach

- Transforming ranking to regression, classification, or ordinal regression.
- Technique includes Subset Ranking, McRank, Prank, OC SVM
- Supposed there is a group of objects (documents associated with a query) in the feature space
- Suppose that there are three grades (levels)
- Example:  $x_1$ ,  $x_2$  and  $x_3$  in the first group are at different grades.
- The weight vector  $w$  corresponding to the linear function  $f(x) = \langle w, x \rangle$ , which can score and rank the objects.

# Ordinal Classification SVM

- Learns the parallel hyperplanes by the large margin principle.



## Loss Function

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{r=1}^{l-1} \sum_{i=1}^{m_r} (\xi_{r,i} + \xi_{r+1,i}^*) \\ \text{s. t.} \quad & \langle w, x_{r,i} \rangle + b_r \geq 1 - \xi_{r,i} \\ & \langle w, x_{r+1,i} \rangle + b_r \leq 1 - \xi_{r+1,i}^* \\ & \xi_{r,i} \geq 0, \quad \xi_{r+1,i}^* \geq 0 \\ & i = 1, \dots, m_r, \quad r = 1, \dots, l-1 \\ & m = m_1 + \dots + m_l, \end{aligned}$$

# Pairwise Approach

- Transforms ranking to pairwise classification problem
- Techniques are Ranking SVM, RankBoost, RankNet, IR SVM etc

# Pairwise Approach

- Idea: Learn a ranking function, so that number of violated pair-wise training preferences is minimized.

- What is pairwise?

Given a query  $q \rightarrow$  Document  $d_i$  is more relevant than document  $d_j$

Thus,  $(q, d_i) > (q, d_j)$

- Form of Ranking Function: sort by

$$\begin{aligned} U(q, d_i) = & w_1 * (\text{\#of query words in title of } d_i) + \\ & w_2 * (\text{\#of query words in anchor}) + \\ & \dots + \\ & w_n * (\text{page-rank of } d_i) = w * \phi(q, d_i) \end{aligned}$$

if user prefers  $d_i$  to  $d_j$  for query  $q$ , then  $U(q, d_i) > U(q, d_j)$

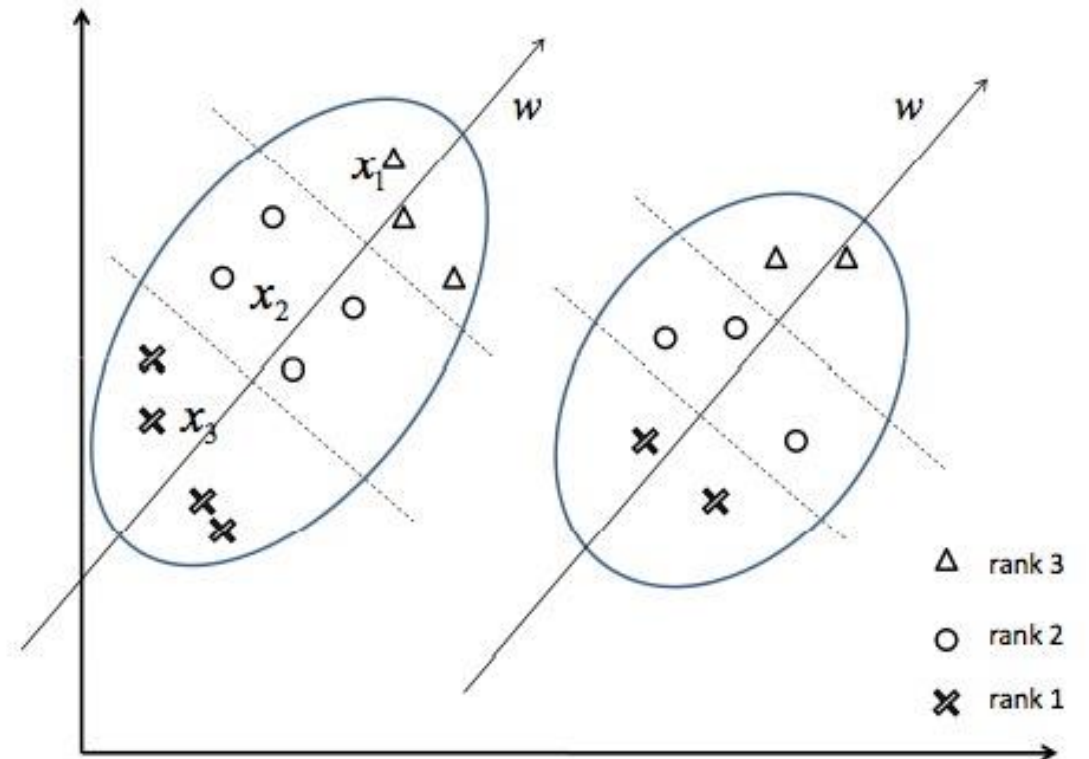
# Pairwise Approach Example

- Supposed there are groups of objects (documents associated with two queries) in the feature space
- Suppose that there are three grades (levels)
- Example:  $x_1$ ,  $x_2$  and  $x_3$  in the first group are at different grades.
- The weight vector  $w$  corresponding to the linear function  $f(x) = \langle w, x \rangle$ , which can score and rank the objects.



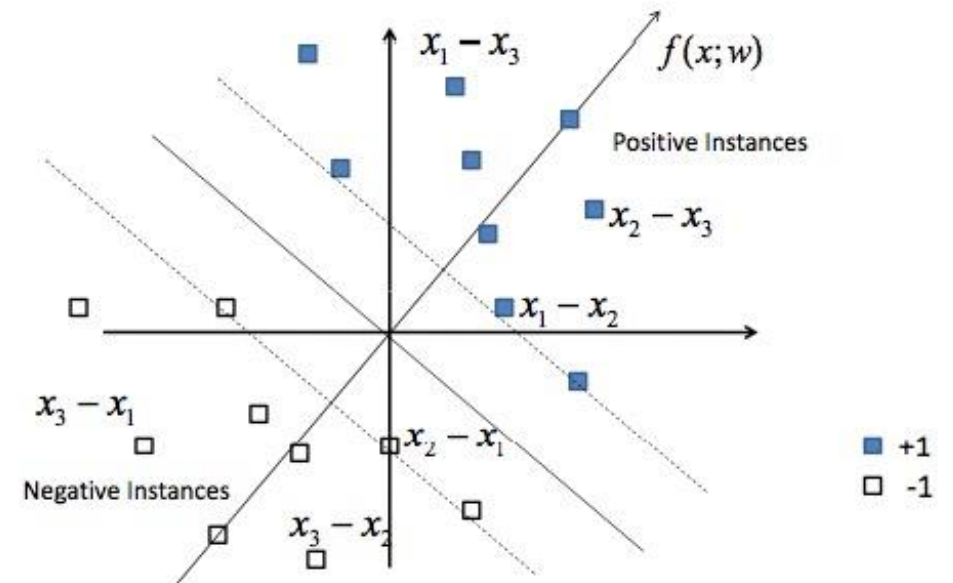
# Pairwise Approach

- For a given query  $q$ , document  $d_1 > d_2 > d_3$
- related,  $x_1, x_2, x_3$  are  $d_1, d_2, d_3$  characteristics
- Using machine learning methods to sort then transformed into a classification problem.



# Pairwise Approach- Ranking SVM

- Define a new training sample such that
  - $x_1 - x_2, x_1 - x_3, x_2 - x_3$  is a positive sample,
  - $x_2 - x_1, x_3 - x_1, x_3 - x_2$  is negative sample
- Training a two classifier with support vector machine to these new training samples.



# Ranking SVM

- using SVM to classify by minimizing loss function

$$\begin{aligned} \min_{w, \xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s. t.} \quad & y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ & i = 1, \dots, m, \end{aligned}$$

where  $w$  is the parameter vector,  $x$  is a feature of the document,  $y$  is the document to the relative correlation,  $\xi$  is slack variable.

# Future Reading

---

## Large Scale Learning to Rank

---

D. Sculley  
Google, Inc.  
dsculley@google.com

### Abstract

Pairwise learning to rank methods such as RankSVM give good performance, but suffer from the computational burden of optimizing an objective defined over  $O(n^2)$  possible pairs for data sets with  $n$  examples. In this paper, we remove this super-linear dependence on training set size by sampling pairs from an implicit pairwise expansion and applying efficient stochastic gradient descent learners for approximate SVMs. Results show orders-of-magnitude reduction in training time with no observable loss in ranking performance. Source code is freely available at: <http://code.google.com/p/sofia-ml>

## Combined Regression and Ranking

D. Sculley  
Google, Inc.  
Pittsburgh, PA USA  
dsculley@google.com

### ABSTRACT

Many real-world data mining tasks require the achievement of two distinct goals when applied to unseen data: first, to induce an accurate preference *ranking*, and second to achieve good *regression* performance. In this paper, we give an efficient and effective Combined Regression and Ranking method (CRR) that optimizes regression and ranking objectives simultaneously. We demonstrate the effectiveness of CRR for both families of metrics on a range of large-scale tasks, including click prediction for online advertisements. Experiments show that CRR often achieves performance equivalent to the best of both ranking-only and regression-only methods. In the case of rare events or skewed distributions, we also find that this combination can actually im-

prove for producing predicted values with the same pairwise ordering  $y'_1 > y'_2$  as the true values  $y_1 > y_2$  for a pair of given examples.

In many settings good performance on both families together is needed. An important example of such a setting is the prediction of clicks for sponsored search advertising. In real-time auctions for online advertisement placement, ads are ranked based on  $bid * pCTR$  where  $pCTR$  is the predicted click-through rate (CTR) of an ad. Predicting a good ranking is critical to efficient placement of ads. However, it is also important that the  $pCTR$  not only give good ranking value, but also give good regression estimates. This is because online advertisements are priced using next-price auctions, in which the price for a click on an ad at rank  $i$

# References

- A Short Introduction to Learning to Rank IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10 OCTOBER 2011
- <https://www.youtube.com/watch?v=1NxnPkZM9bc> lecture on SVM by **Thales Sehn Körting**
- <http://fangdonghao.com/>
- [http://users.cis.fiu.edu/~lzhen001/activities/KDD\\_USB\\_key\\_2010/docs/p979.pdf](http://users.cis.fiu.edu/~lzhen001/activities/KDD_USB_key_2010/docs/p979.pdf)
- <http://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/35662.pdf>