

# the hedge algorithm and applications



Virgil Pavlu

# will it rain tomorrow ?

- I say it will rain
  - Nick says it will not rain
  - CNN says it will rain
  - FOX says it will not rain
- 
- TOMORROW : sunny all day...
  - what about Saturday?

# weighted majority algorithm

- Give the “weather-men” weights
  - Initial uniform or according to some prior belief
- Run the forecast for several days. Every day :
  - Make our prediction by weighted-majority-vote
  - Get the real outcome
  - Update weights
    - “penalize” wrong predictors
    - “reward” good predictors

# more problems

- more problems
  - trading stocks
  - IR metasearch
  - disease classification
    - where we need to “train” doctors
- common underlying idea
  - generalization of weighted-majority
- why not exactly WM
  - losses are real numbers instead of discrete 0-1
  - our loss may be a weighted sum of expert losses

- Hedge algorithm for online allocation [schapire, freund '96]
- Applications
  - On classification : Adaboost
  - On IR : metasearch algorithm

# “how to use expert advice” ?

- N experts (or strategies)
- maintain a set of weights over experts
- loop for T episodes  $t=1,2,\dots,T$ 
  - allocate Resources (believe) - based on weights
  - receive loses
  - reweight the experts

# online allocation - hedge algorithm

Start with uniform distribution  $D$

Choose Allocation

Receive loss vector

Update distribution

$\beta \in [0,1]$ ;  $N$  strategies (systems)

initial weights  $w^1 \in [0,1]^N$ ;  $\sum_{i=1}^N w_i^1 = 1$

$$p_i^t = \frac{w_i^t}{\sum_{i=1}^N w_i^t}$$

$$l^t \in [0,1]^N$$

$$L_H^t = p^t \cdot l^t$$

$$w_i^{t+1} = w_i^t \cdot \beta^{l_i^t}$$

$$L_{HEDGE} = \sum_{t=1}^T p^t \cdot l^t$$

# why hedge

- goal : Hedge loss close to the Loss of the best expert (bound)
- proof idea :
  - relate episodic Hedge loss with the sum of weights

$$\sum_{i=1}^N w_i^{t+1} \leq \left( \sum w_i^t \right) \left( 1 - (1 - \beta) L_{Hedge(\beta)}^t \right)$$

- relate cumulate Hedge loss with sum of final weights

$$\ln \left( \sum_{i=1}^N w_i^{T+1} \right) \leq -(1 - \beta) L_{Hedge(\beta)}$$

- relate sum of final weights with loss of the best expert

$$L_{Hedge(\beta)} \leq \frac{\ln(1/\beta)}{1 - \beta} L_{best} + \frac{1}{1 - \beta} \ln(N)$$



# episodic hedge loss

$$\sum_{i=1}^N w_i^{t+1} \leq \left( \sum_{i=1}^N w_i^t \right) \left( 1 - (1 - \beta) L_{Hedge}^t \right)$$

$$\sum_{i=1}^N w_i^{t+1} = \sum_{i=1}^N w_i^t \beta^{l_i^t}$$

$$\leq \sum_{i=1}^N w_i^t (1 - (1 - \beta) l_i^t)$$

$$= \sum_{i=1}^N p_i^t \left( \sum_{i=1}^N w_i^t \right) (1 - (1 - \beta) l_i^t)$$

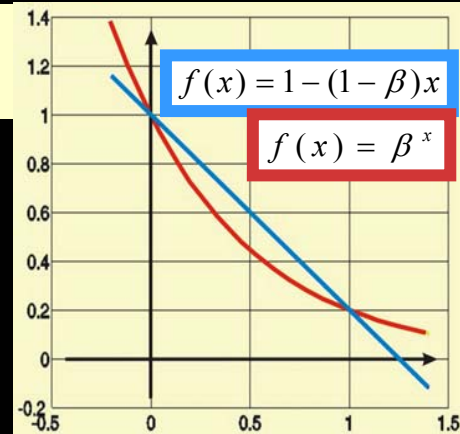
$$= \left( \sum_{i=1}^N w_i^t \right) \sum_{i=1}^N (p_i^t - (1 - \beta) p_i^t l_i^t)$$

$$= \left( \sum_{i=1}^N w_i^t \right) \left( 1 - (1 - \beta) \sum_{i=1}^N p_i^t l_i^t \right)$$

$$= \left( \sum_{i=1}^N w_i^t \right) \left( 1 - (1 - \beta) p^t \cdot 1^t \right)$$

$$\beta, l \in [0, 1]$$

$$\beta^l \leq 1 - (1 - \beta)l$$



$$p_i^t = \frac{w_i^t}{\sum_{i=1}^N w_i^t}$$

$$L_{Hedge}^t = p^t \cdot 1^t$$

$$\ln \left( \sum_{i=1}^N w_i^{T+1} \right) \leq -(1 - \beta) L_{\text{Hedge}(\beta)}$$

cumulative  
hedge loss

$$\sum_{i=1}^N w_i^{t+1} \leq \left( \sum w_i^t \right) \left( 1 - (1 - \beta) L_{\text{Hedge}}^t \right)$$

Applying repeatedly for  $t = 1, \dots, T$  yields

$$\sum_{i=1}^N w_i^{t+1} \leq \left( \sum w_i^1 \right) \prod_{t=1}^T \left( 1 - (1 - \beta) p^t \cdot l^t \right)$$

$$\leq \left( \sum w_i^1 \right) \prod_{t=1}^T \exp \left( - (1 - \beta) p^t \cdot l^t \right)$$

$$1 - x \leq \exp(x)$$

$$\leq \exp \left( - (1 - \beta) \sum_{t=1}^T p^t \cdot l^t \right)$$

$$L_{\text{Hedge}} = \sum_{t=1}^T p^t \cdot l^t$$

# [almost] as good as the best expert

$$\ln \left( \sum_{i=1}^N w_i^{T+1} \right) \leq -(1 - \beta) L_{\text{Hedge}(\beta)}$$

$$\sum_{i=1}^N w_i^{T+1} \geq w_{\text{any}}^{T+1} = w_{\text{any}}^1 \beta^{L_{\text{any}}}$$

$$w_{\text{any}}^1 = \frac{1}{N}$$

$$L_{\text{Hedge}(\beta)} \leq \frac{-\ln\left(\frac{1}{N}\right) - L_{\text{any}} \ln \beta}{1 - \beta}$$

# optimality

$$L_{\text{Hedge}(\beta)} \leq \frac{-\ln\left(\frac{1}{N}\right) - L_{\text{best}} \ln \beta}{1 - \beta} = \frac{\ln(1/\beta)}{1 - \beta} L_{\text{best}} + \frac{1}{1 - \beta} \ln(N)$$

**Theorem 3** *Let  $B$  be an algorithm for the on-line allocation problem with an arbitrary number of strategies. Suppose that there exist positive real numbers  $a$  and  $c$  such that for any number of strategies  $N$  and for any sequence of loss vectors  $\ell^1, \dots, \ell^T$*

$$L_B \leq cL_{\text{best}} + a \ln(N)$$

Then for all  $\beta \in (0, 1)$ , either

$$c \geq \frac{\ln(1/\beta)}{1 - \beta} \quad \text{or} \quad a \geq \frac{1}{(1 - \beta)}$$

# how to choose $\beta$

$$L_{Hedge(\beta)} \leq \frac{\ln(1/\beta)}{1-\beta} L_{best} + \frac{1}{1-\beta} \ln(N)$$

- $\beta$  = confidence parameter
- think of it as a trade-off
- try to make the bound tight
- binary search : perfect expert + ( $\beta=0$ )

# that wasn't so bad

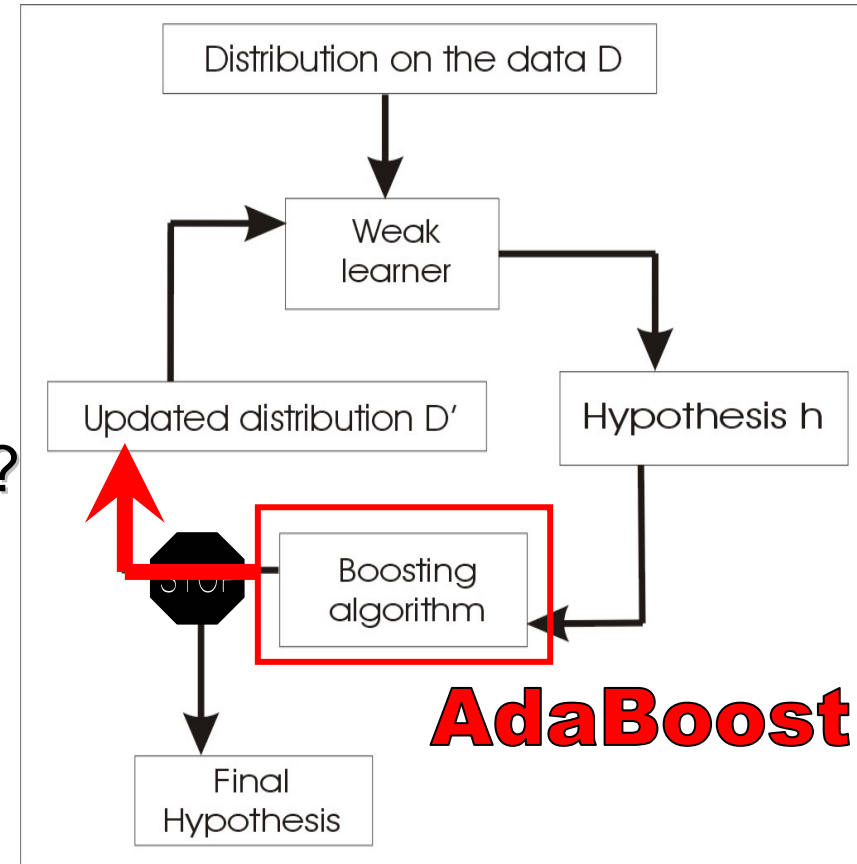
☑ Hedge algorithm for  
online allocation

## ● Applications

- On classification : Adaboost
- On IR : metasearch algorithm

# boosting

- disease classification....
  - get past data
  - “train” a disease-predictor “doctor” (“hypothesis”, “weak learner”)
- how good is it (on training data) ?
  - where is it wrong ?
  - train a new predictor to correct mistakes for the first one



# hedge application : AdaBoost

[schapire, freund '96]

## HEDGE

- given : experts

• incoming : loses

• reweight: experts

## BOOSTING

- given : datapoints
  - experts in weak-learners performance

• incoming: weak learners

- compute error (loss)

- compute “believe” ( $\beta$ )

• reweight : datapoints

## ADABOOST

$$D_1(i) = \frac{1}{m}$$

$$h_t : X \rightarrow \{-1, 1\}$$

$$\varepsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$$

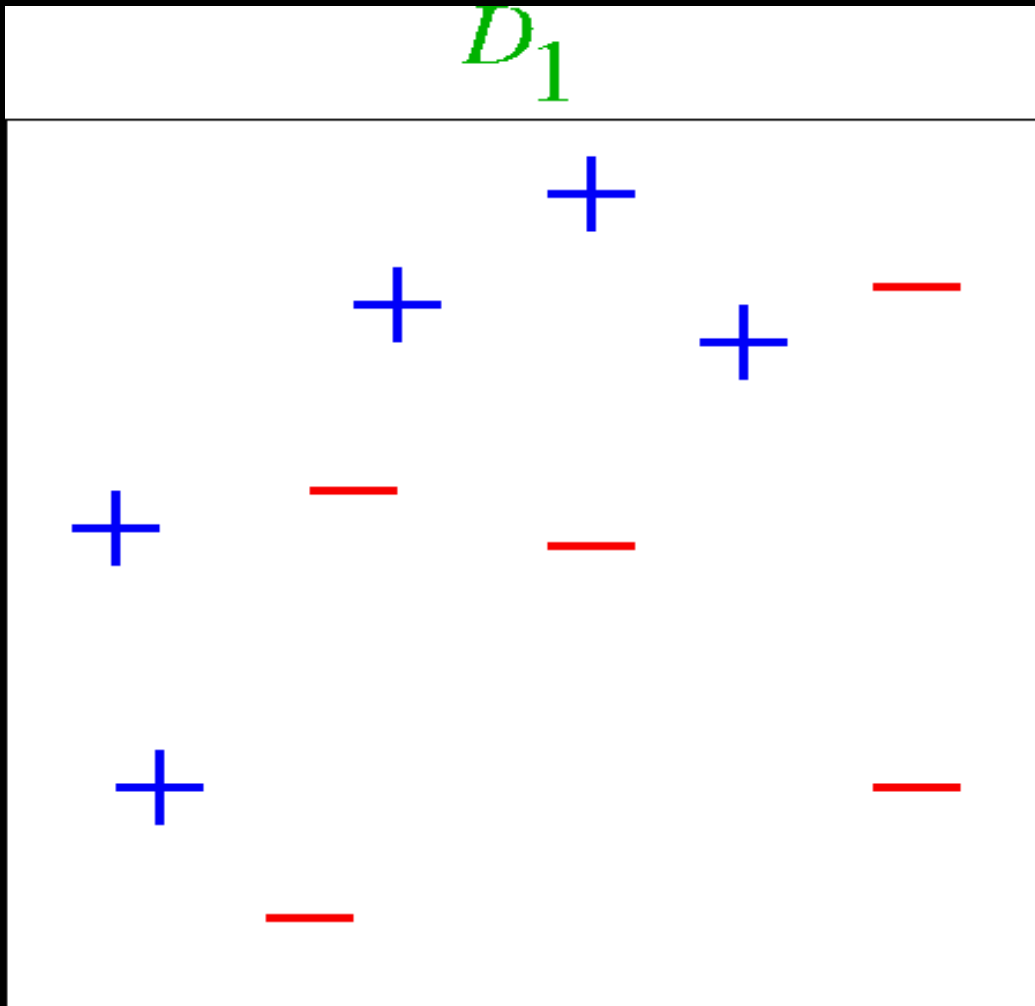
$$\beta_t = \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}} > 1$$

$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} 1/\beta_t & \text{if } y_i = h_t(x_i) \\ \beta_t & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$H_{\text{final}}(x) = \text{sgn} \left( \sum_t \ln(\beta_t) h_t(x) \right)$$

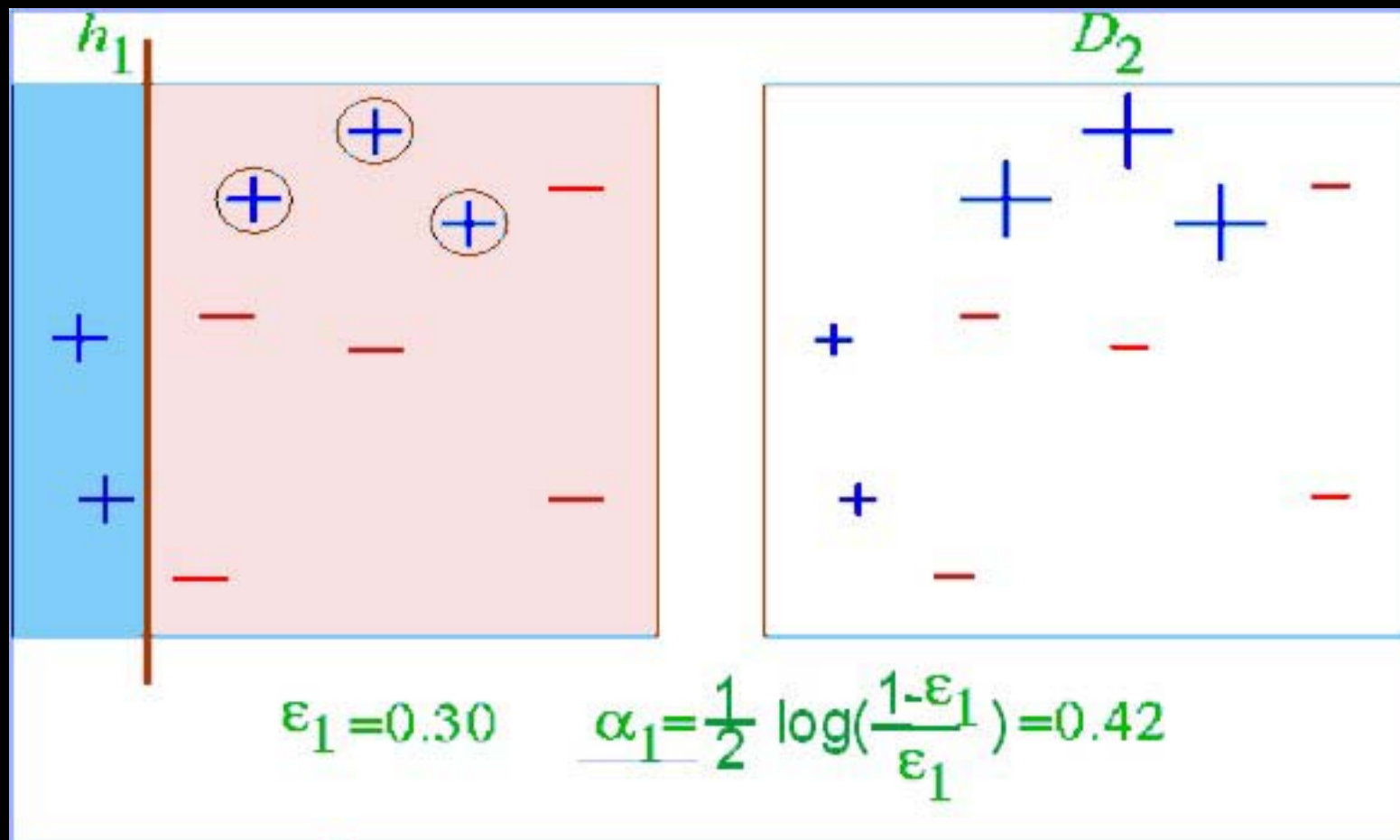


# AdaBoost - example

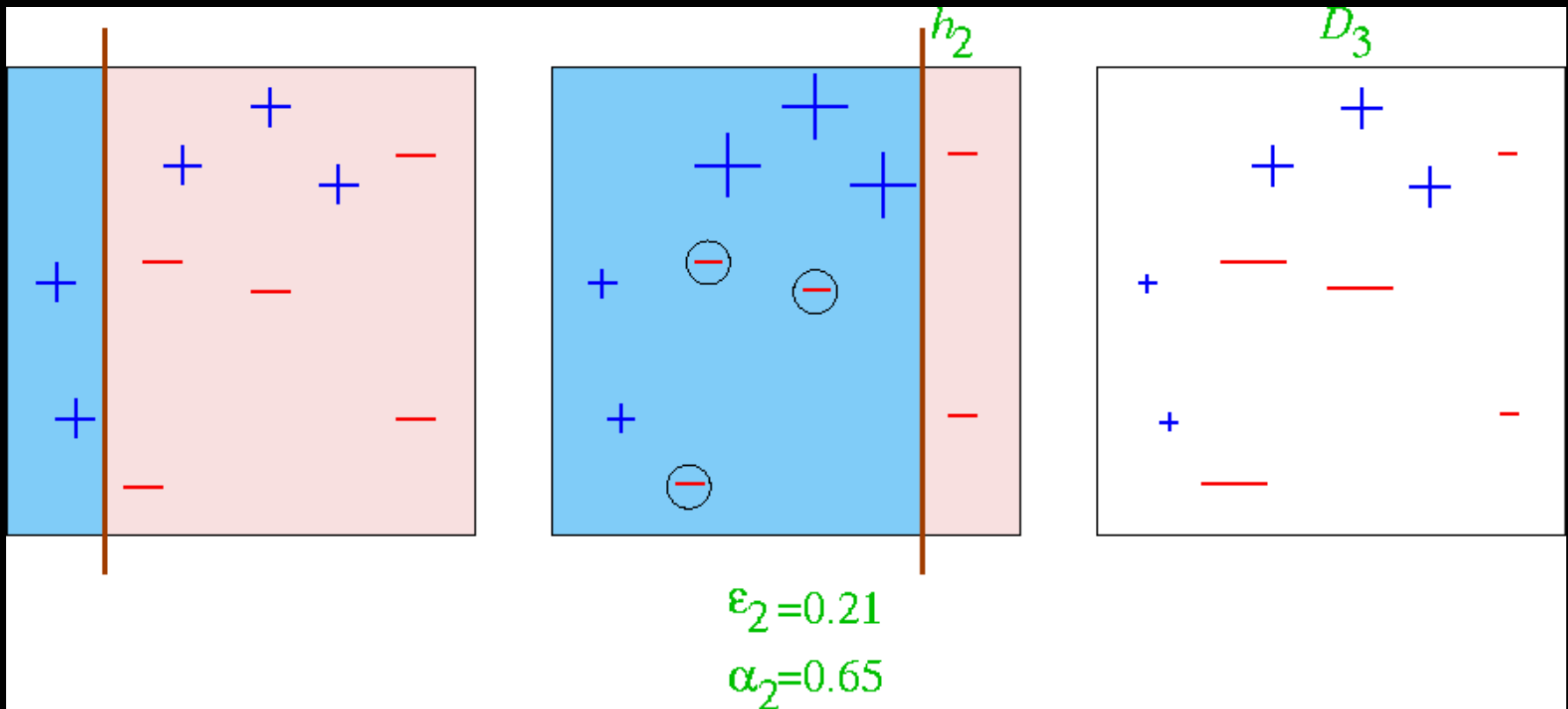


- Start with uniform distribution on data
- Weak learners = halfplanes

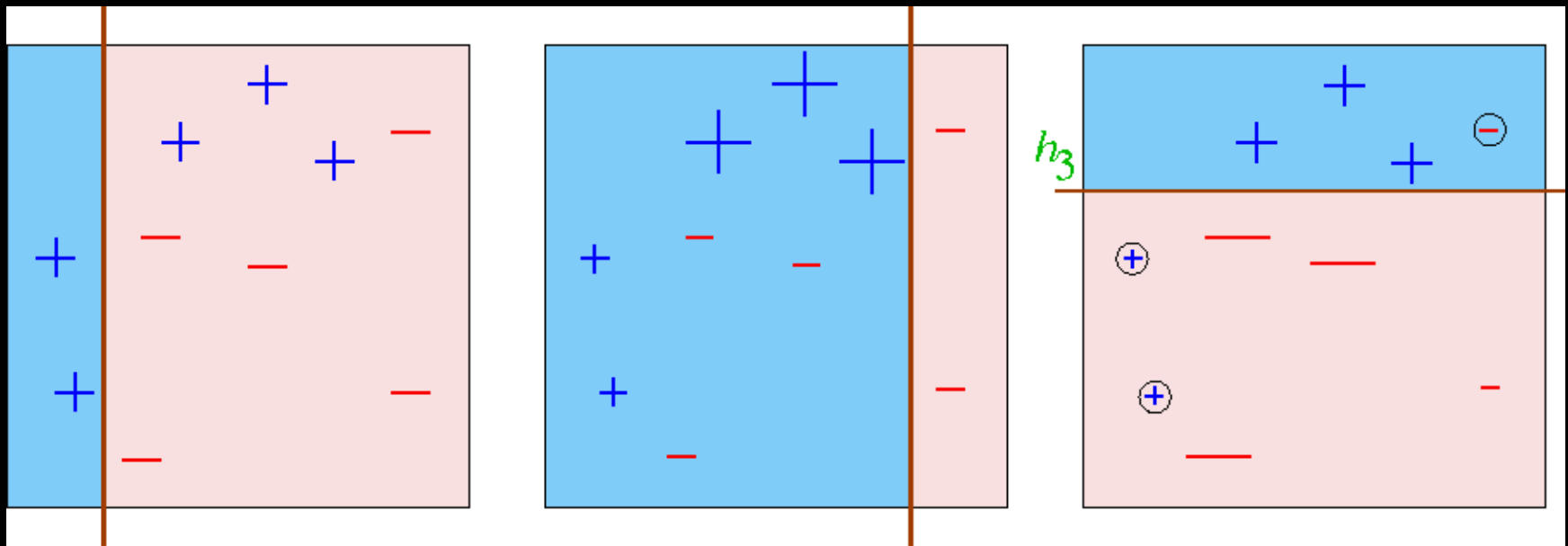
# round 1



# round 2



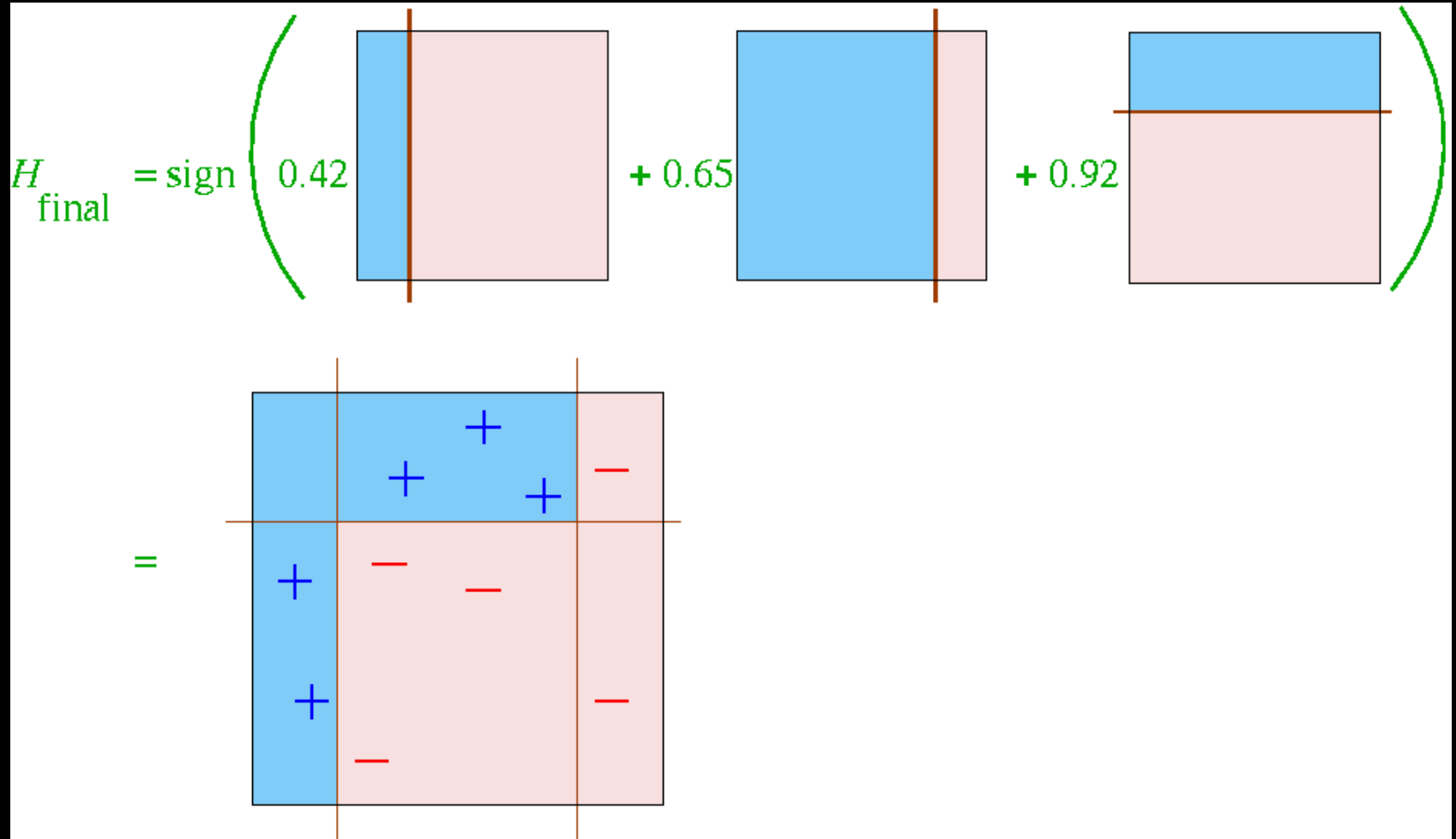
# round 3



$$\epsilon_3 = 0.14$$

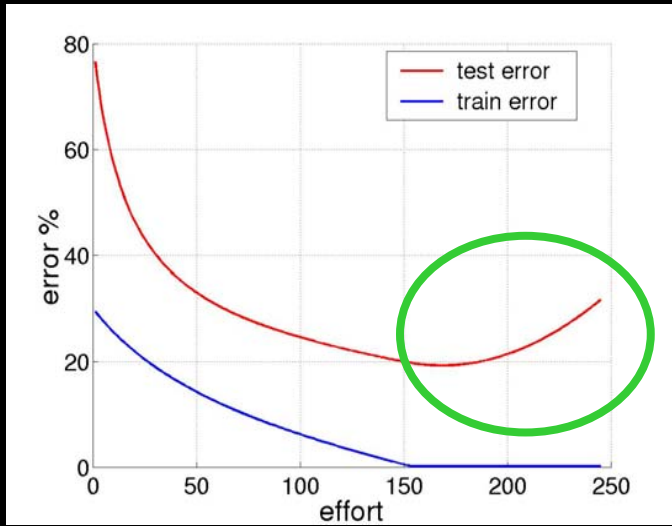
$$\alpha_3 = 0.92$$

# final hypothesis

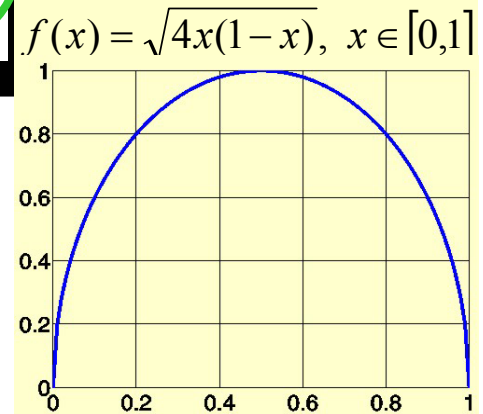
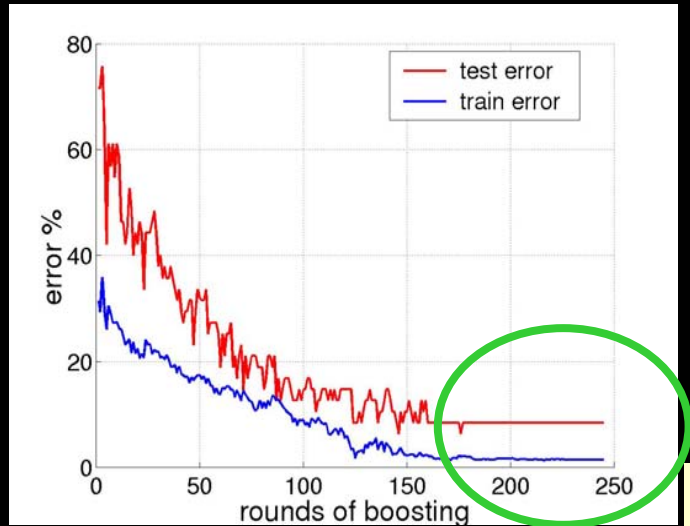


# AdaBoost - analysis

typical curve



AdaBoost



- training error

$$\text{training error}(H_{\text{final}}) \leq \prod_t \sqrt{4\varepsilon_t(1-\varepsilon_t)}$$

- generalization error

$$\mathbf{P}_{\mathcal{D}} [yf(x) \leq 0] \leq \mathbf{P}_{\mathcal{S}} [yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \left(\frac{\log m \log |\mathcal{H}|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right)$$

# 2003 Gödel Prize

## Yoav Freund and Robert Schapire

- “The prize was awarded to Yoav Freund and Robert Schapire for their paper "A Decision Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences 55 (1997), pp. 119-139. This paper introduced AdaBoost, an adaptive algorithm to improve the accuracy of hypotheses in machine learning. The algorithm demonstrated novel possibilities in analysing data and is a permanent contribution to science even beyond computer science. Because of a combination of features, including its elegance, the simplicity of its implementation, its wide applicability, and its striking success in reducing errors in benchmark applications even while its theoretical assumptions are not known to hold, the algorithm set off an explosion of research in the fields of statistics, artificial intelligence, experimental machine learning, and data mining. The algorithm is now widely used in practice. The paper highlights the fact that theoretical computer science continues to be a fount of powerful and entirely novel ideas with significant and direct impact even in areas, such as data analysis, that have been studied extensively by other communities.”

# what we did last year [Aslam, Pavlu, Savell]

Hedge algorithm for  
online allocation

- Applications

- On classification : Adaboost

- On IR : metasearch algorithm



# metasearch problem

Search for: **CIKM 2003**

The image shows three overlapping search engine results pages for the query "CIKM 2003".

- Google (left):** Shows search results for "CIKM 2003 Homepage", "ACM CIKM 2003 Call For Papers", "CIKM-2003 Registration", "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", "Conference on Information and Knowledge Management (CIKM)", "[Asis-] CIKM 2003", "(DBWORLD) CIKM 2003 (Padm)", "bridge-cikm-2003", and "Collaborative Filtering Mailing".
- Altavista (middle):** Shows search results for "CIKM 2003 Homepage", "ACM CIKM 2003 Call For Papers", "CIKM 2002 Homepage", "Conference on Information and Knowledge Management (CIKM)", "CIKM", "SIGIR Information Server", and "Web Caching Publications/Venue".
- Alltheweb (right):** Shows search results for "CIKM 2003 Homepage", "IVML Call For Papers Archive: ACM CIKM", "CIKM 2002 Homepage", "ACM WIDM'2003", "MMDB'03", "D-Lib Workshops and Conferences: 2003", "Invited paper to be presented in CIKM", and "Internet Conferences 2003".

## metasearch

CIKM 2003 Homepage  
ACM CIKM 2003 Call For Paper  
CIKM-2003 Registration  
ACM CIKM 2003 Call For Paper  
CIKM 2002 Homepage  
Conference on Information and Knowledge Management (CIKM)  
[Asis-] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm)  
bridge-cikm-2003  
Collaborative Filtering Mailing  
CIKM Home Page ACM DL: CIKM  
Yahoo! Groups : webir Message Boards  
dBforums - Cfp: Cikm 03  
Mailing List ARL-ERESERVE@arl.nrl.navy.mil  
Received: from cni.org by b.cri.nrl.navy.mil  
ACM WIDM 2003  
ACM - MMDB 2003  
ACM CIKM 2003 PRELIMINARY Selected Publications  
Web Caching Publications/Venue  
cikm '03  
Call for papers  
Mario A. Nascimento - Personal Home Page  
Invited Paper to be Presented in CIKM  
Conferences and Journals on Information and Knowledge Management  
Conferences On Information and Knowledge Management  
Calendrier des manifestations  
(DBWORLD) Final Call for ACM CIKM 2003  
Collaborative Filtering Mailing

# search and metasearch

- Search engines:
  - Provide a ranked list of documents.
  - May provide relevance scores.
- Metasearch engines:
  - Query multiple search engines.
  - May or may not combine results.

# metasearch: Dogpile

The screenshot shows the Dogpile metasearch engine interface. At the top, the browser address bar displays the URL: <http://search.dogpile.com/text?search?q=chili+peppers&geo=no&fs=web&av=custom&engines=goto&engines=looksmart&engines=tl>. The Dogpile logo is prominently displayed, along with navigation links such as Home, Custom Search, Dogpile Remote, Search at Home, Help with Syntax, Tell a Friend, and Go Shopping. A SmarterKids.com banner is visible, promoting an Ask Our Teachers column. Below the banner, there are links for Amazon.com and Electric Library. The search results section shows a table of links related to chili peppers, including Red Hot Chili Peppers, Peppers, Chili, and Hot Peppers. A Dogpile Suggests section recommends ClassMates.com. The search engine used is Looksmart, which found 117 results. The query string sent was +chili +peppers. The results list includes: 1. The Red Hot Chili Peppers (Find photos, lyrics, updates, tour info, and news on alternative-funk-rock band the Red Hot Chili Peppers. Looksmart category - Red Hot Chili Pepper), 2. Red Hot Chili Peppers Audio and Video (Watch videos and listen to music by this rock/funk band. Looksmart category - Red Hot Chili Peppers), 3. Chili and Hot Sauces (Shop for mouth-burning chili sauces, Tabasco, hot salsas and other pepper-inspired sauces. Looksmart category - Chili & Hot Sauces), 4. Chili and Hot Sauces (Find chili and other hot sauce recipes, including salsas, dips, spices, and rubs, and visit the Pepper Fool. Looksmart category - Chili & Hot Sauces), and 5. Red Hot Chili Peppers - Screens and Themes (Promotional screensaver for the funk-rock band features falling chili peppers. LookSmart category - Red Hot Chili Peppers Multimedia). A link for the next set of results from Looksmart is provided. The search engine used is GoTo.com, which found 10 or more results.

File Edit View Go Communicator Help

Bookmarks Location: <http://search.dogpile.com/text?search?q=chili+peppers&geo=no&fs=web&av=custom&engines=goto&engines=looksmart&engines=tl> What's Related

Home  
Custom Search  
Dogpile Remote  
Search at Home  
Help with Syntax  
Tell a Friend  
Go Shopping

 **DOGPILE**  
Web Metasearch Results

 Visit the Ask Our Teachers column at [SmarterKids.com](#).

Spend web time wisely. Shop smarter. Hear My Mail from Weblevy! Find old classmates here! [ClassMates.com](#)

Buy books about "chili peppers" at [Amazon.com](#) Search for "chili peppers" on [Electric Library](#)

Are you looking for:

<a href="#">Red Hot Chili Peppers</a>	<a href="#">Peppers</a>	<a href="#">Chili</a>	<a href="#">Hot Peppers</a>
<a href="#">Red Hot Chili Peppers Tabs</a>	<a href="#">Canning Recipes Peppers</a>	<a href="#">Red Hot Chili Peppers Lyrics</a>	<a href="#">Salsa Recipes Free</a>

 [ClassMates.com](#). Remember your promise to "keep in touch" at graduation? With over 8.5 million registered high school alumni, chances are you'll find your old friends.

Search engine: [Looksmart](#) found 117 results.  
The query string sent was [+chili +peppers](#)

- [1. The Red Hot Chili Peppers](#)  
Find photos, lyrics, updates, tour info, and news on alternative-funk-rock band the Red Hot Chili Peppers.  
[Looksmart category - Red Hot Chili Pepper](#)
- [2. Red Hot Chili Peppers Audio and Video](#)  
Watch videos and listen to music by this rock/funk band.  
[Looksmart category - Red Hot Chili Peppers](#)
- [3. Chili and Hot Sauces](#)  
Shop for mouth-burning chili sauces, Tabasco, hot salsas and other pepper-inspired sauces.  
[Looksmart category - Chili & Hot Sauces](#)
- [4. Chili and Hot Sauces](#)  
Find chili and other hot sauce recipes, including salsas, dips, spices, and rubs, and visit the Pepper Fool.  
[Looksmart category - Chili & Hot Sauces](#)
- [5. Red Hot Chili Peppers - Screens and Themes](#)  
Promotional screensaver for the funk-rock band features falling chili peppers.  
[LookSmart category - Red Hot Chili Peppers Multimedia](#)

[Next set of results from Looksmart](#)

Search engine: [GoTo.com](#) found 10 or more results.

# metasearch: Metacrawler

The screenshot displays the Metacrawler search engine interface. At the top, the browser's address bar shows the URL: `http://search.metacrawler.com/crawler?general=chili+peppers&method=0&redirect=web&pp=20&hp=10&region=0&timeout=0&sort=`. The search bar contains the text "chili peppers" and a "Search" button. Below the search bar, there are navigation options: "any", "all", and "phrase". A "The Web" button is also visible.

Under the heading "Are you looking for:", there are several links: [Red Hot Chili Peppers](#), [Peppers](#), [Chili](#), [Hot Peppers](#), [Red Hot Chili Peppers Tabs](#), [Canning Recipes Peppers](#), [Red Hot Chili Peppers Lyrics](#), and [Salsa Recipes Free](#).

Below this, there are two search boxes: one for "Discovery Health" and another for "Get Auto Insurance answers".

A navigation bar shows "View Related:" with buttons for "Web Pages", "Directory Listings", "Audio/MP3", "Images", "News Groups", and "Auctions".

The main content area displays "Results for 'chili peppers' 1 to 20 of 122 results". It includes a "View by:" section with options for "Relevance", "Site", and "Source", and a link to "Email results to a friend".

A "MetaCrawler Suggests" section recommends "ClassMates.com, where you can find your High School Friends!".

Below this, a section titled "You just searched these search engines:" lists various search engines like AltaVista, Infoseek, LookSmart, Lycos, GoTo, Thunderstone, DirectHit, Google, Internet Keywords, FindWhat, and Sprinks by About.

The search results list includes:

- Red Hot Chili Peppers**: AltaVista: Buy Californication Now. Enter your email address to receive information about The Red Hot Chili Peppers. Email: Go to the RED HOT CHILI PEPPERS... Infoseek: Video clips and photos of the band, plus lyrics and interview. Google: Buy Californication Now. Enter your email address to receive 1000, <http://www.wbr.com/chilipeppers/index.html> (AltaVista, Infoseek, Google) | [More Like This](#)
- Red Hot Chili Peppers**: This web site was created for FREE at [www.homestead.com](http://www.homestead.com). Visit [www.homestead.com](http://www.homestead.com) to get your free web site - no programming required. 424, <http://www.homestead.com/musiccorner/hotchilipeppers.html> (AltaVista) | [More Like This](#)
- redhotchilipeppers online**: click here to skip intro... 424, <http://redhotchilipeppers.com/> (Google) | [More Like This](#)
- The Red Hot Chili Peppers**: Find photos, lyrics, updates, tour info, and news on alternative-funk-rock band the Red Hot Chili Peppers. 424, <http://www.looksmart.com/eas11.../us315010/us312562/uniaq:6986> (LookSmart) | [More Like This](#)
- Lycos Music: Red Hot Chili Peppers - Biography**: Biography - Toting in L.A.'s club circuit for years, the Red Hot Chili Peppers finally pounded their Funk/trap/rock into the mainstream with the 1989 LP Mother's Milk. Making bedfellows of their native L.A. Punk. 424, <http://click.hotbot.com/direct...li+%2Bpeppers&source=LCOSMCW1> (Lycos) | [More Like This](#)
- GlassShop.com - Handmade Blown Glass Chili Peppers**: Our standard offering includes four gorgeous 4-6 inch long blown glass chili peppers. Check 'em out! They make a great gift item. 424, <http://www.glassshop.com/main.html> (FindWhat) | [More Like This](#)
- Chili Pepper smoked Trout**: Chili Pepper smoked Trout recipe. 424, <http://bbq.about.com/ibrasy...raw&terms=%2Bchili+%2Bpeppers> (Sprinks by About) | [More Like This](#)

On the right side of the page, there is a sidebar with several advertisements, including "0% Visa Cards Now! @ YourNextCreditCard" and "Apply Now for an Aria Visa!".

# metasearch: Profusion

The screenshot shows a web browser window displaying the Profusion metasearch engine results for the query "chili peppers". The browser's address bar shows the URL: <http://www.profusion.com/cgi-bin/rph-ProFusion.pl?queryterm=chili+peppers&option=all&display=10&totalverify=0&auto=auto>. The search interface includes a search bar with the text "chili peppers" and a "Web Search" button. The results are displayed in a list format, with each entry showing a relevance score, a title, a summary, and a URL. The results are sorted by relevance, with the highest score being 1.0000.

**Search Results** [Track This Search](#)

Results: 1-10 of 63. << Previous 1 | 2 | 3 | 4 | 5 | 6 | 7 | All Next >>

- 1.0000 Red Hot Chili Peppers** [Track This Page](#)  
Summary: This web site was created for FREE at www.homestead.com. Visit www.homestead.com to get your free web site - no programming required.  
URL: <http://www.homestead.com/musiccorner/hotchilipeppers.html> (AlphaVista)
- 0.9667 Abbey Road: Red Hot Chili Peppers Fansite** [Track This Page](#)  
Summary: For fans of pre-1992 Red Hot Chili Peppers Funk Style!...  
URL: [http://members.aol.com/gld\\_rhcp/index.html](http://members.aol.com/gld_rhcp/index.html) (AlphaVista)
- 0.9333 dotmusic - Single Review - Red Hot Chili ...** [Track This Page](#)  
Summary: Typically groovy bassist Flea fuels the momentum  
URL: <http://www2.dotmusic.com/reviews/Singles/news/2000/reviews/12638...> (AlphaVista)
- 0.9000 Red Hot Chili Peppers** [Track This Page](#)  
Summary: Buy Californication Now. Enter your email address to receive information about The Red Hot Chili Peppers. Email: Go to the RED HOT CHILI PEPPERS...  
URL: <http://www.wbr.com/chilipeppers/index.html> (AlphaVista)
- 0.8781 Chili Pepper Rub for Brisket Recipe** [Track This Page](#)  
Summary: Chili Pepper Rub for Brisket recipe.  
URL: <http://bbq.about.com/brbrv/rec/b170823a.htm> (About)
- 0.8667 Red Hot Chili Peppers - Screens and Themes** [Track This Page](#)  
Summary: Promotional screensaver for the funk-rock band features falling chili peppers. More like this: Red Hot Chili Peppers Multimedia  
URL: <http://www.screensandthemes.com/Previews-Downloads/Screens-Savers/...> (LookSmart)
- 0.8667 Brigu's Red Hot Chili Peppers & John ...** [Track This Page](#)  
Summary: Bootlegs of great classic rock, the Red Hot Chili Peppers, and John Frusciante. I've got stuff from both the Californication era and the pre-heroin...  
URL: <http://members.aol.com/californiandr/boots.html> (AlphaVista)
- 0.8342 Chili Pepper smoked Trout** [Track This Page](#)  
Summary: Chili Pepper smoked Trout recipe.  
URL: <http://bbq.about.com/brbrv/rec/b170816a.htm> (About)
- 0.8333 Chili Peppers** [Track This Page](#)  
Summary: Did You Know...? | Home | The Legend | Legend Form | A few items for those interested in chili peppers: Scoville Test For Capsaicin--A Thermal...  
URL: <http://www.cybersauce.com/knowledge.htm> (AlphaVista)
- 0.8000 Hot Chili Peppers Homepage** [Track This Page](#)  
Summary: Hot Chili Peppers Homepage (German)...

# metasearch algorithms

- Heuristics and hacks:
  - Interleave, average rank, sum scores, etc.
- Principled models:
  - Bayesian inference, election theory, etc.
  - *On-line combination of expert advice.*

# online metasearch

Search for: **CIKM 2003**

The screenshot shows a metasearch engine interface with three search engines: Google, Altavista, and AllTheWeb. The search term is 'CIKM 2003'. The results are aggregated and displayed in a list format. The top results include 'CIKM 2003 Homepage', 'ACM CIKM 2003 Call For Papers', 'CIKM-2003 Registration', and 'CIKM 2002 Homepage'. Each result includes a brief description and a link to the source page. The interface also shows search filters like 'Worldwide' and 'U.S.' for Altavista, and '1 - 10 of 6,293 Results' for AllTheWeb.

## metasearch

CIKM 2003 Homepage  
ACM CIKM 2003 Call For Paper  
CIKM-2003 Registration  
ACM CIKM 2003 Call For Paper  
CIKM 2002 Homepage  
Conference on Information and Knowledge Management (CIKM) 2003  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm  
bridge-cikm-2003  
Collaborative Filtering Mailing  
List ARL-ERESERVE@ar  
Received: from cni.org by b.c  
ACM WIDM 2003  
ACM - MMDB 2003  
ACM CIKM 2003 PRELIMINARY  
Selected Publications  
Web Caching Publications/Ven  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iceved  
Conferences and Journals on T  
Conferences On Information V  
Calendrier des manifestations  
(DBWORLD) Final Call for ACM  
Collaborative Filtering Mailing

## metasearch

CIKM 2003 Homepage  
CIKM-2003 Registration  
CIKM 2002 Homepage  
ACM CIKM 2003 Call For Paper  
ACM CIKM 2003 Call For Paper  
CIKM Home Page ACM DL: CIK  
Conference on Information and  
ACM CIKM 2003 PRELIMINARY  
CIKM Home Page ACM DL: CIK  
ACM CIKM 2003 PRELIMINARY  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm  
bridge-cikm-2003  
Collaborative Filtering Mailing  
List ARL-ERESERVE@ar  
Received: from cni.org by b.c  
ACM WIDM 2003  
ACM - MMDB 2003  
Selected Publications  
Web Caching Publications/Ven  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iceved  
Conferences and Journals on T  
Conferences On Information V  
Calendrier des manifestations

# hedge application : metasearch

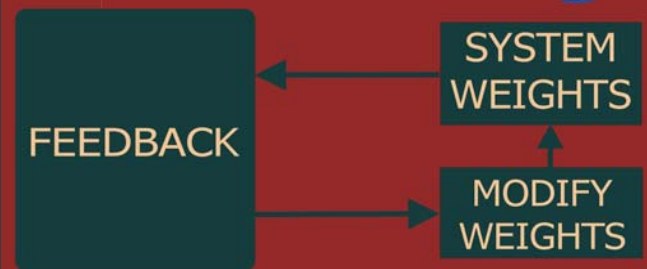
## HEDGE

- given : experts
- incoming : loses
- reweight: experts

## METASEARCH

- given : search engines
- incoming: documents
  - compute losses
- reweight : search engines

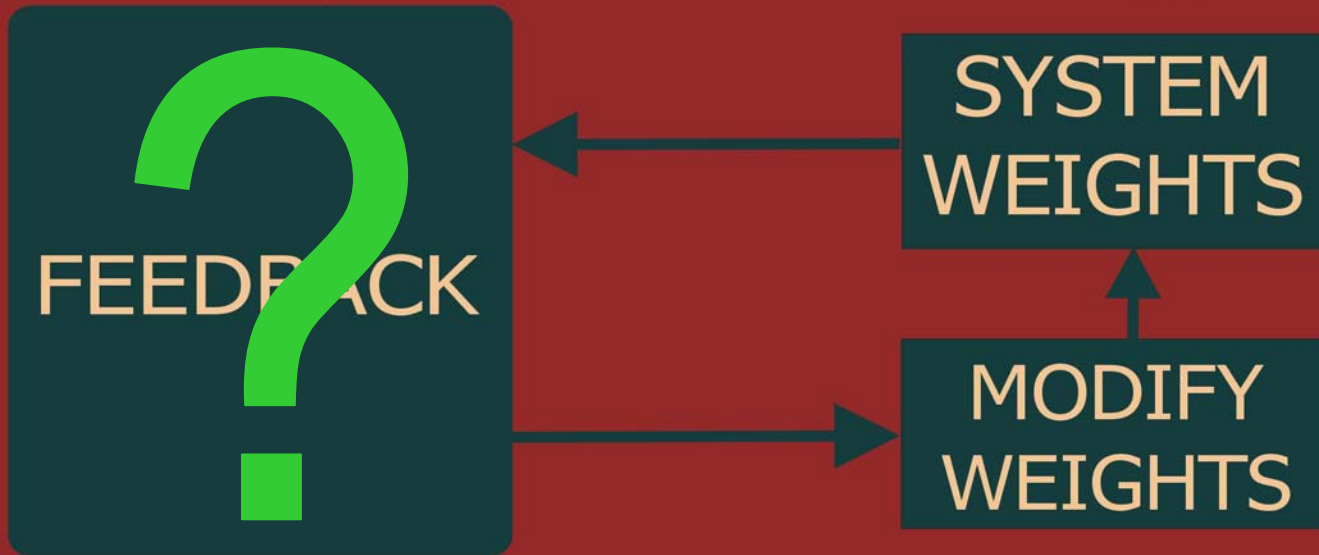
## Core : Hedge





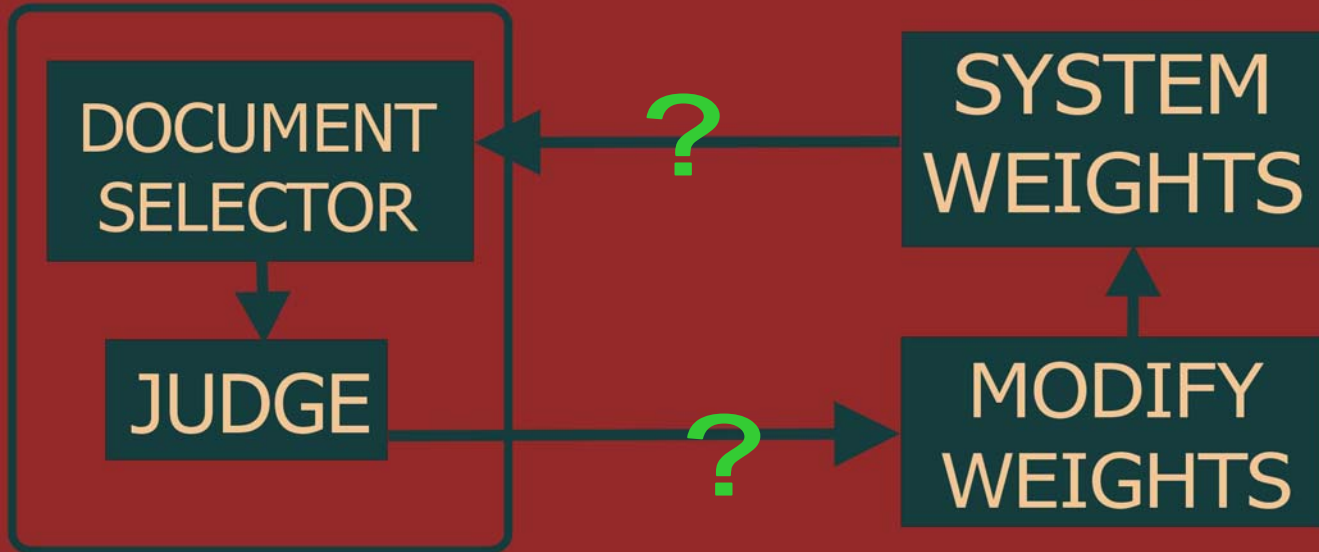
# a unified model

## Core : Hedge

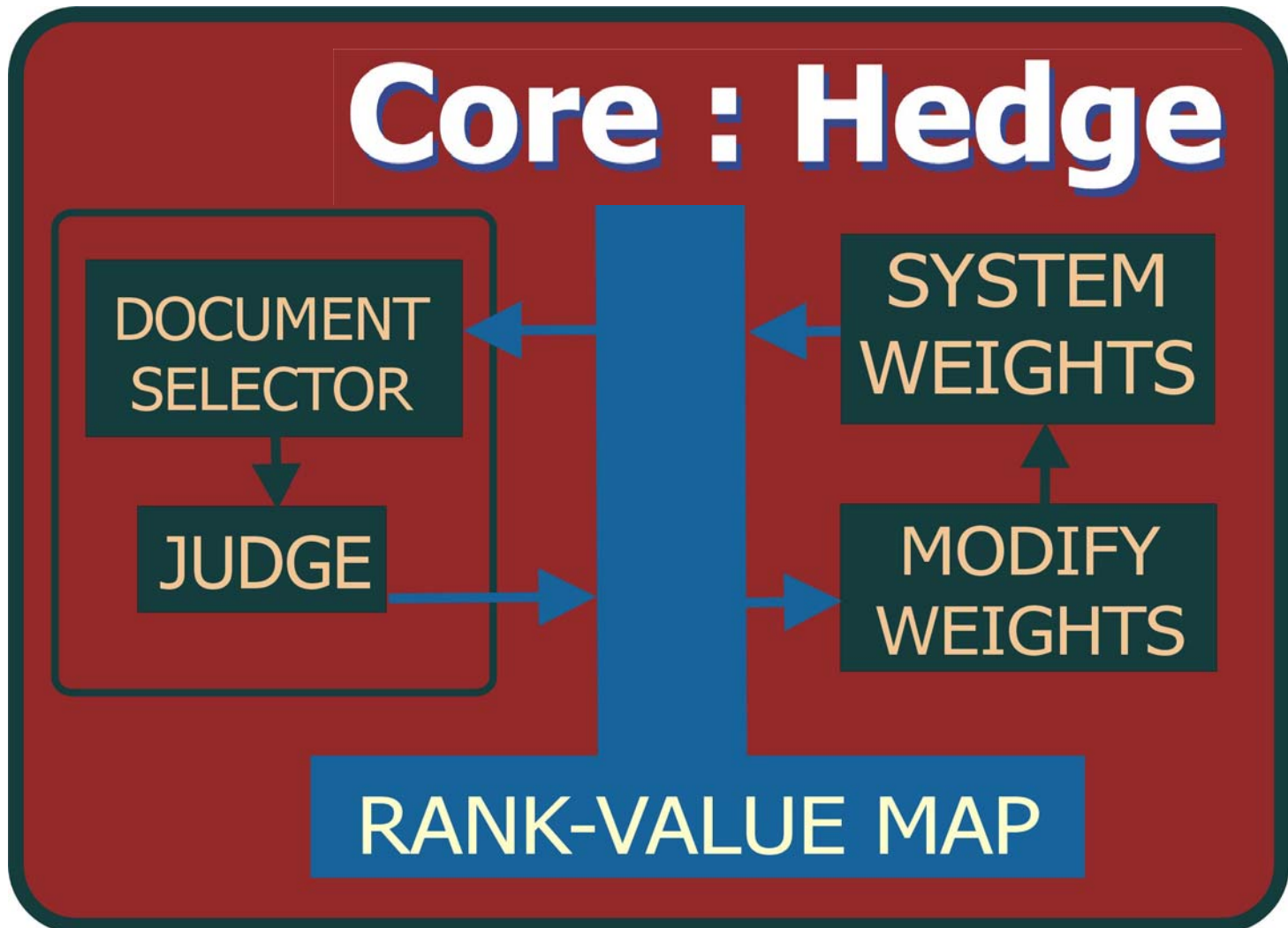


# a unified model

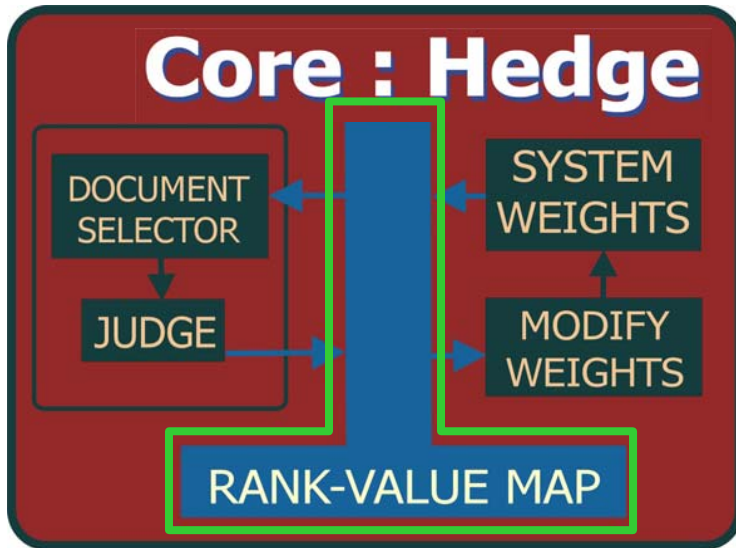
## Core : Hedge



# a unified model



# ranks importance



- map ranks to values

$$value(r) = \frac{1}{r} + \frac{1}{r+1} + \dots + \frac{1}{Z} \approx \ln \frac{Z}{r}$$

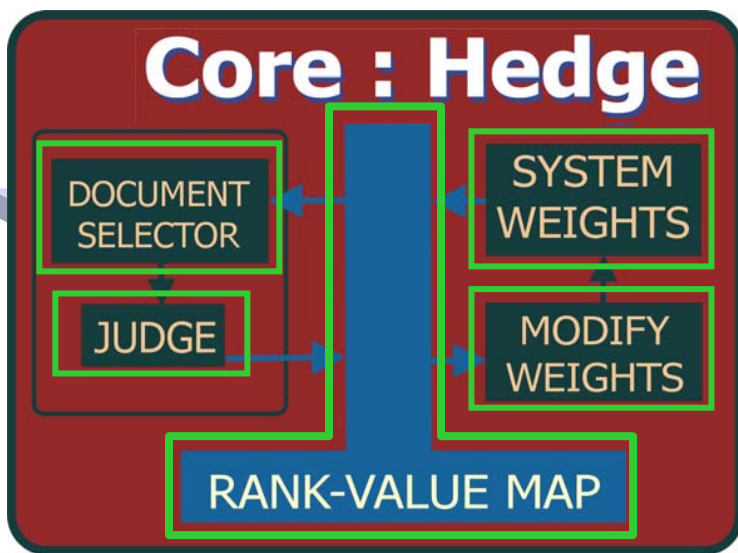
$$RELEVANT = -1$$

$$NONRELEVANT = +1$$

$$LOSS(d, s) = label(d) \cdot value(rank_{d,s}) \approx label(d) \cdot \ln \frac{Z}{r}$$

total loss vs. total precision vs. average precision

# feedback loop



$$average\_value_t(d) = \sum_{s=1}^N w_s^{t-1} \cdot value(rank_{d,s})$$

**metasearch**

- CIKM 2003 Homepage
- CIKM-2003 Registration
- CIKM 2002 Homepage
- ACM CIKM 2003 Call For Paper
- ACM CIKM 2003 Call For Paper
- CIKM Home Page ACM DL: CIK
- Conference on Information an
- ACM CIKM 2003 PRELIMINARY
- CIKM Home Page ACM DL: CIK
- ACM CIKM 2003 PRELIMINARY
- [Asis-] CIKM 2003
- (DBWORLD) CIKM 2003 (Padm
- bridge-cikm-2003
- Collaborative Filtering Mailing
- Yahoo! Groups : webir Messag
- dBforums - Cfp: Cikm 03
- Mailing List ARL-ERESERVE@ar
- Received: from cni.org by b.cr
- ACM WIDM 2003
- ACM - MMDB 2003
- Selected Publications
- Mario A. Nascimento - Person
- Received
- Conferences and Journals on T
- Conferences On Information V
- Calendrier des manifestations



- average\_value at episode t
  - “trust” in systems change with t
- metasearch list : order docs by
- get feedback
- compute losses
- modify weights

$$average\_value_t(d)$$

$$label(d_t)$$

$$LOSS(d_t, s) = label(d_t) \cdot value(rank_{d_t, s})$$

$$w_s^{t+1} = w_s^t \cdot \beta^{LOSS_t(d_t, s)}$$

# experiments

- TREC 3,5,6,7,8
  - 41-129 systems
  - 50 queries per TREC
  - Metasearch combines *all* systems
- Use TREC judgments as user feedback

# metasearch - no feedback (yet)

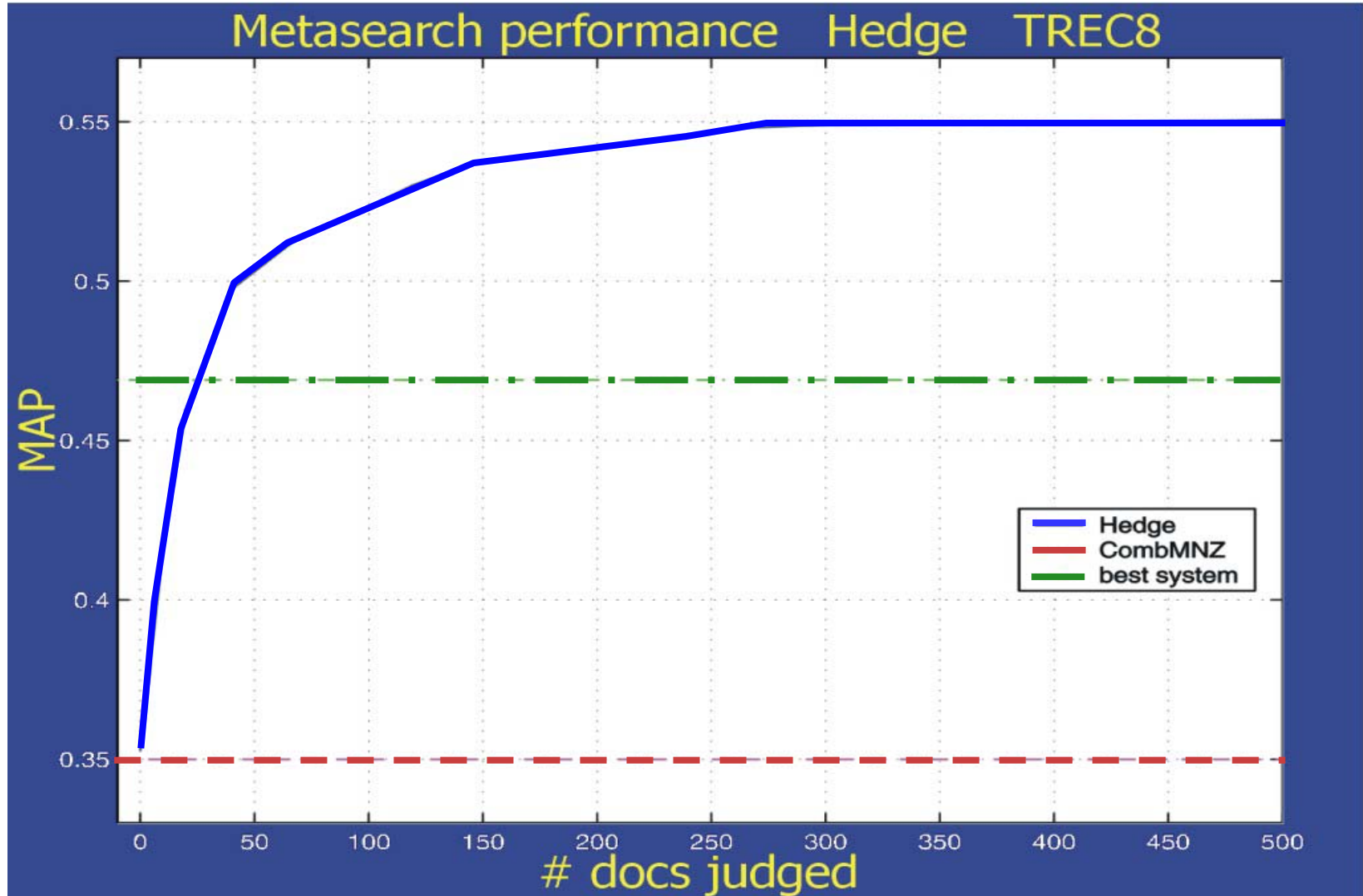
no relevant judgements → uniform weights

TREC	MNZ	COND	Hedge-0	%MNZ	%COND
3	0.423	0.403	0.418	-1.2	+3.7
5	0.294	0.307	0.309	+5.1	+0.6
6	0.341	0.315	0.345	+1.2	+9.5
7	0.320	0.308	0.323	+0.9	+4.9
8	0.350	0.343	0.352	+1.4	+2.6

MNZ=CombMNZ(Fox,Shaw,Lee et al)

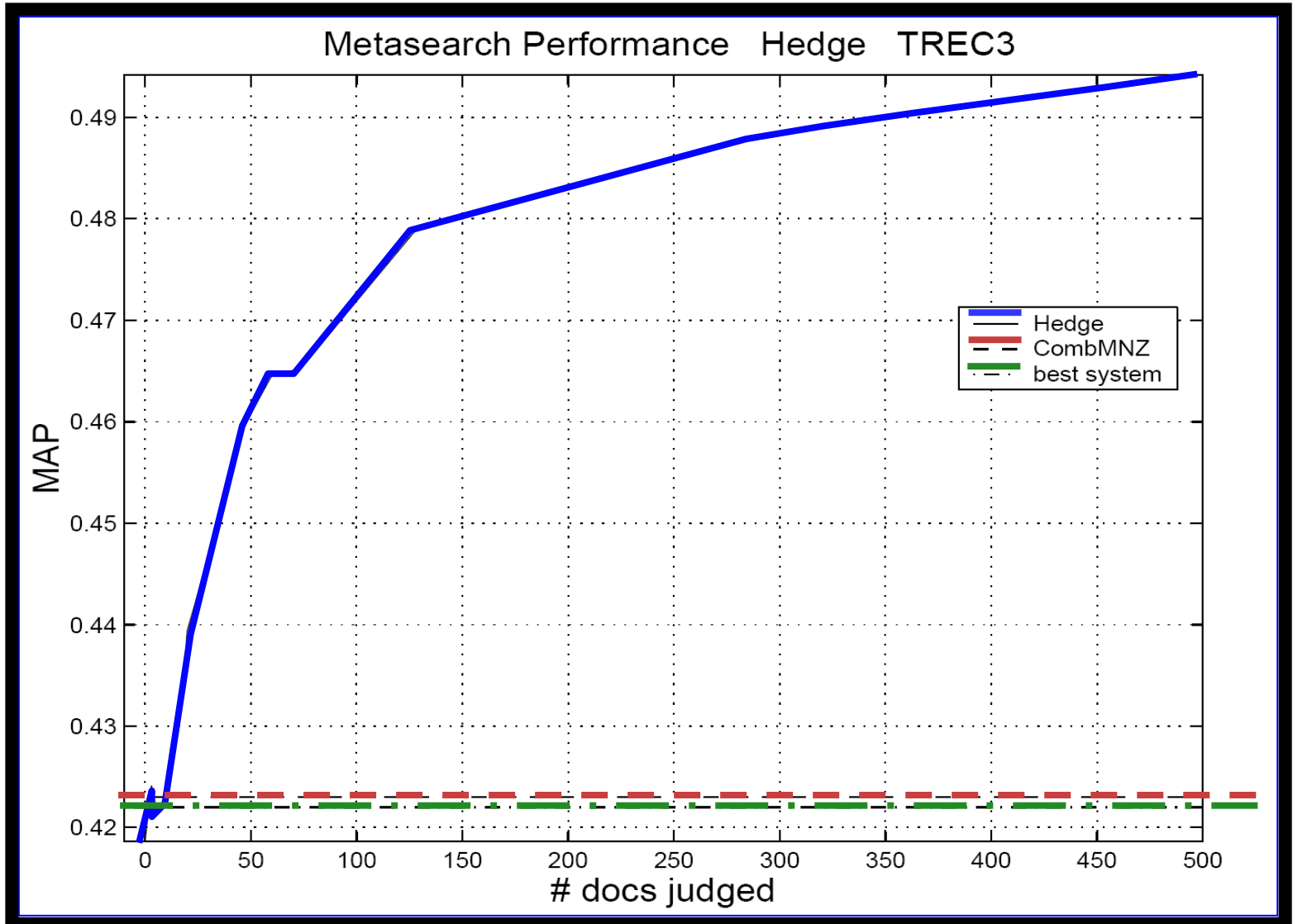
COND=Condorcet(Aslam,Montague)

# metasearch – TREC8

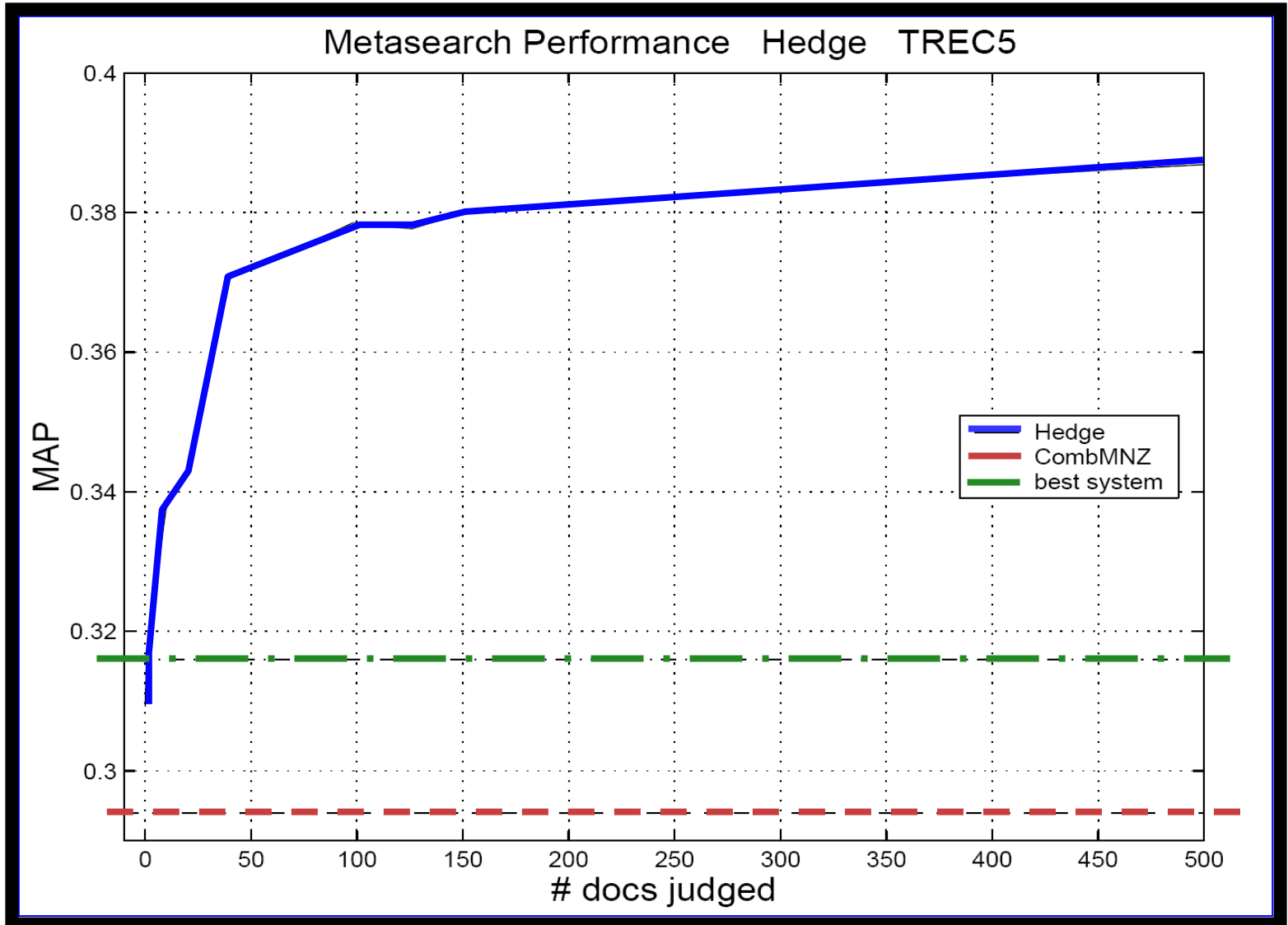




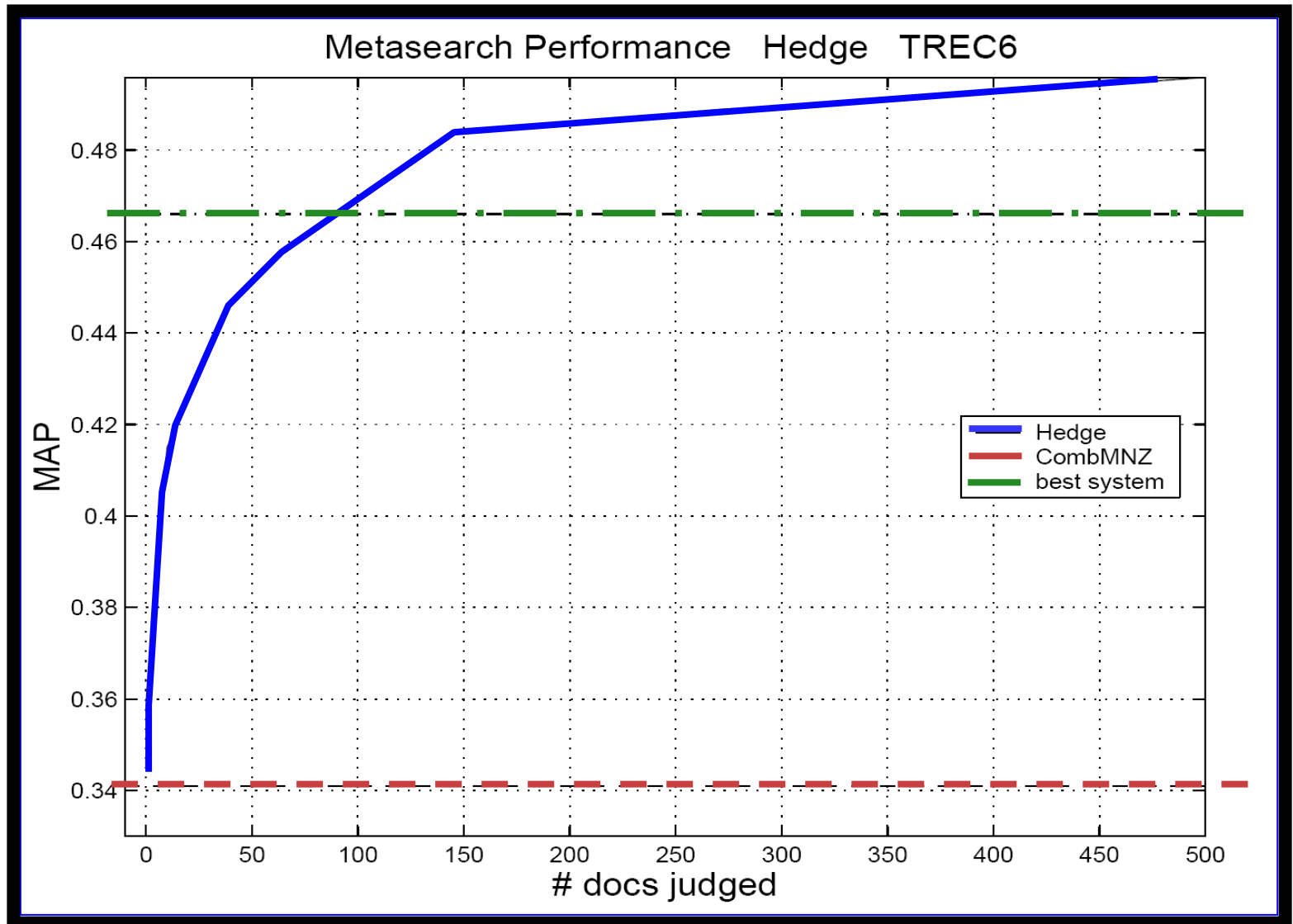
# metasearch – TREC 3



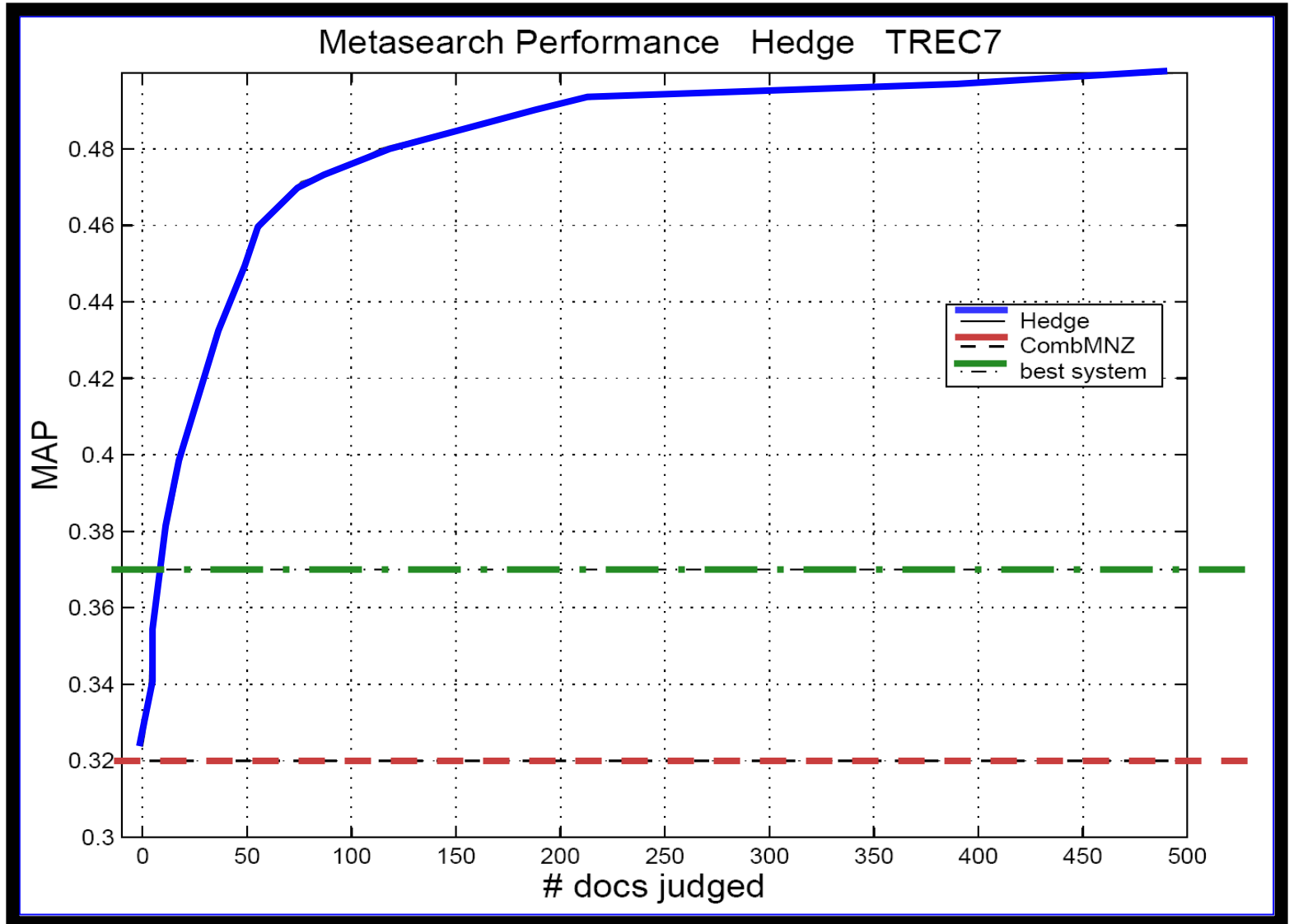
# metasearch – TREC 5



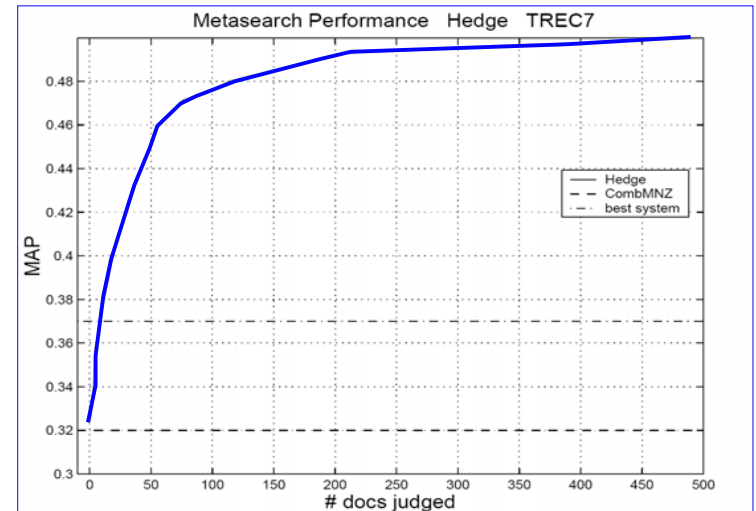
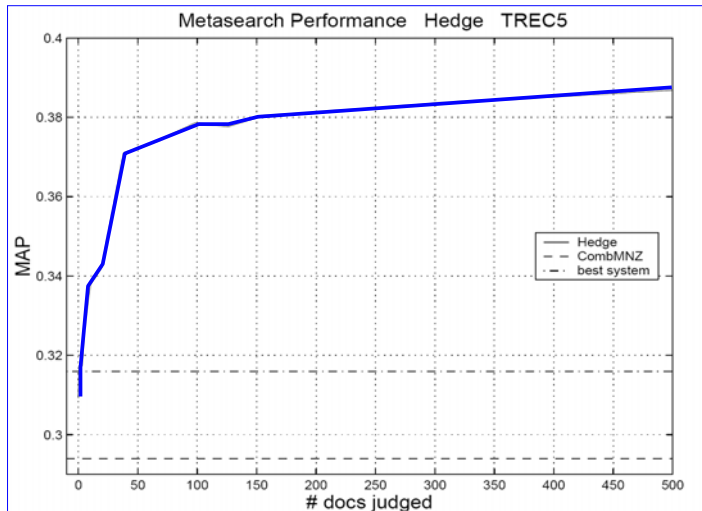
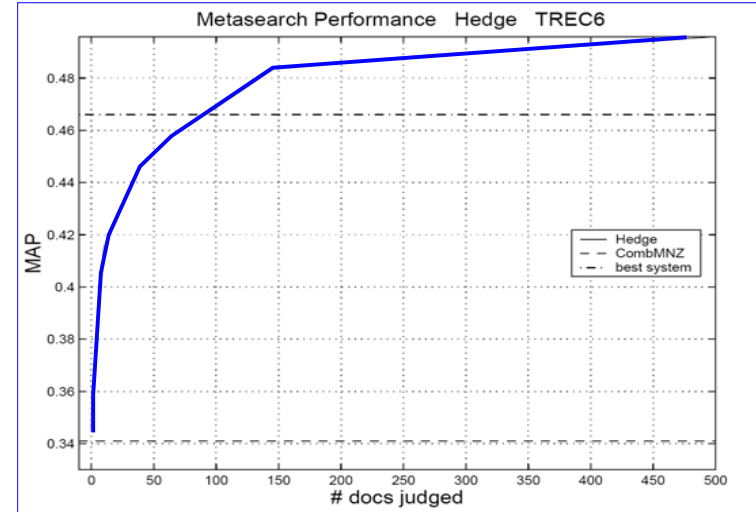
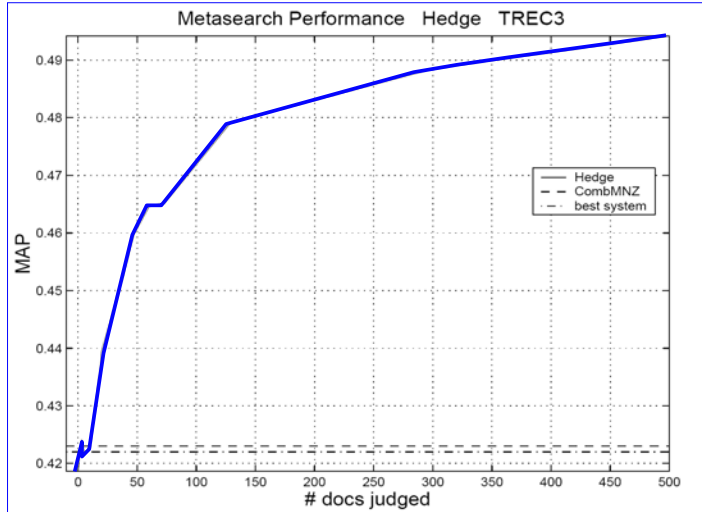
# metasearch – TREC 6



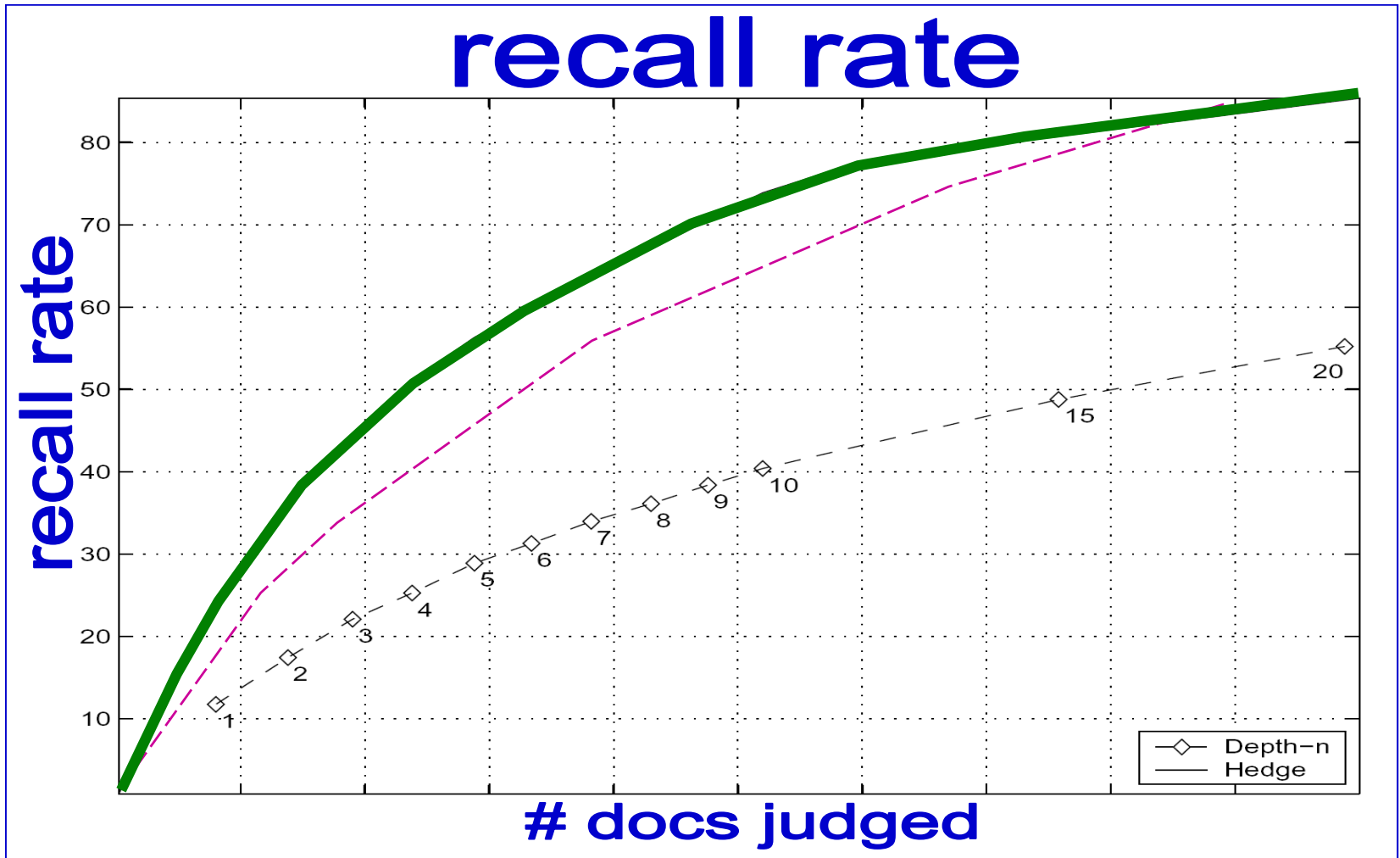
# metasearch – TREC 7



# metasearch – TREC 3,5,6,7

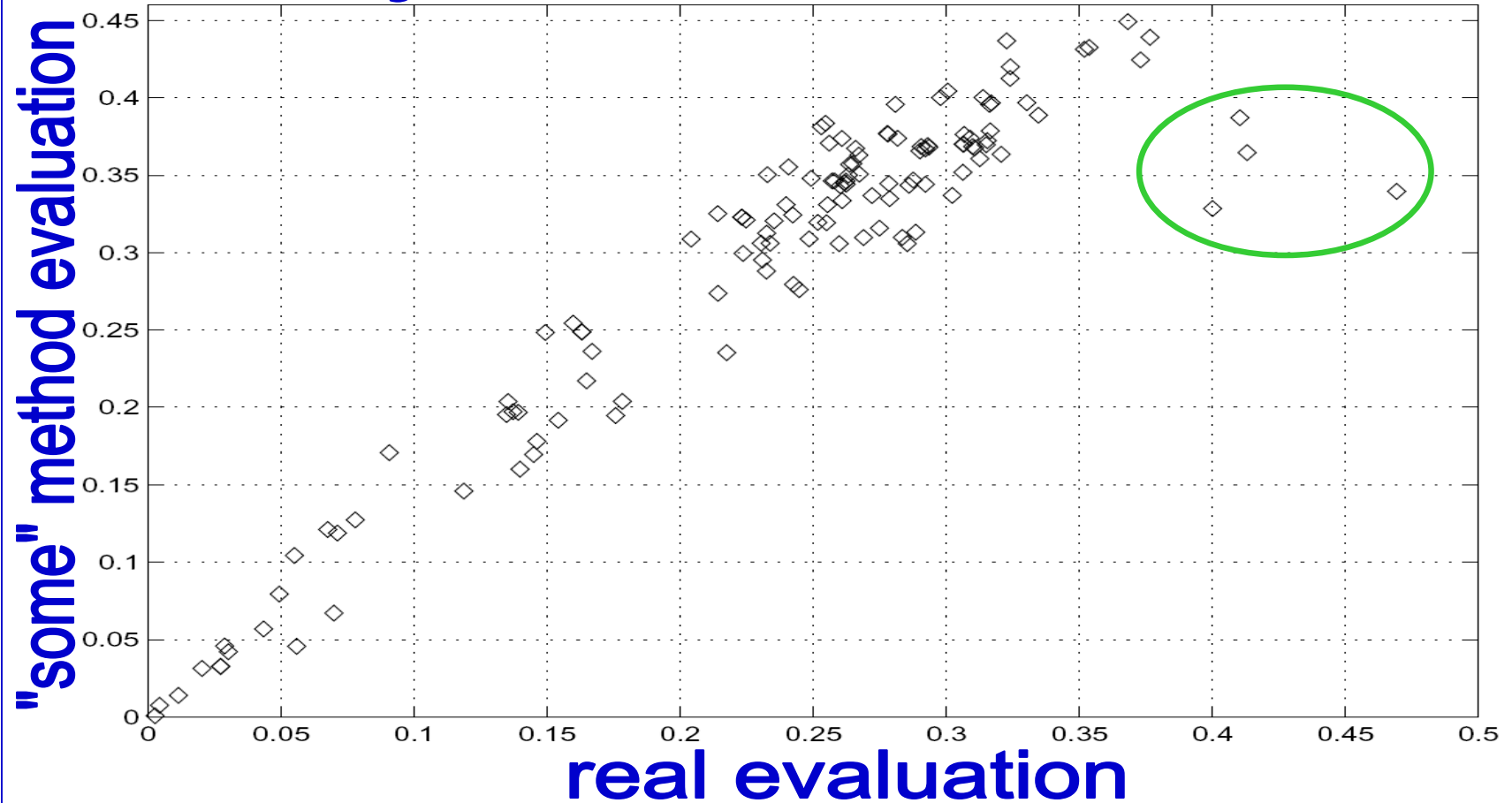


# actually...we do more than metasearch

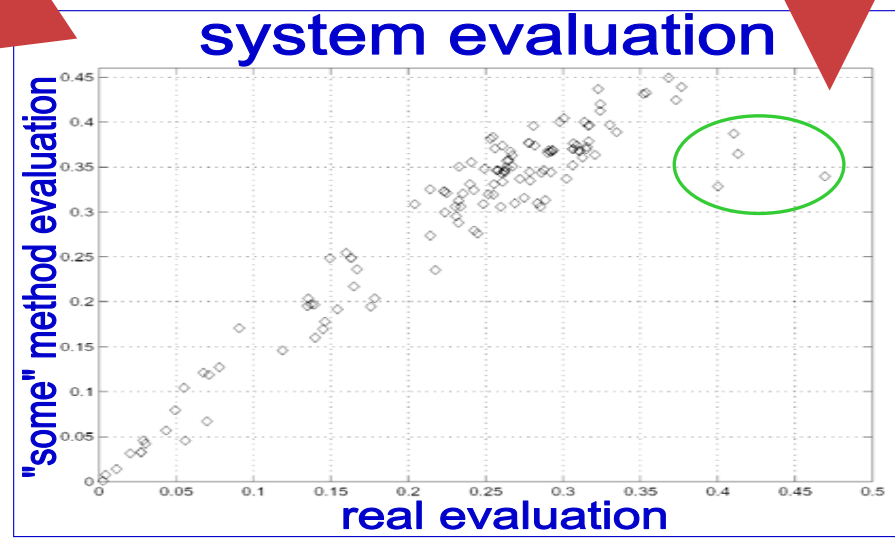
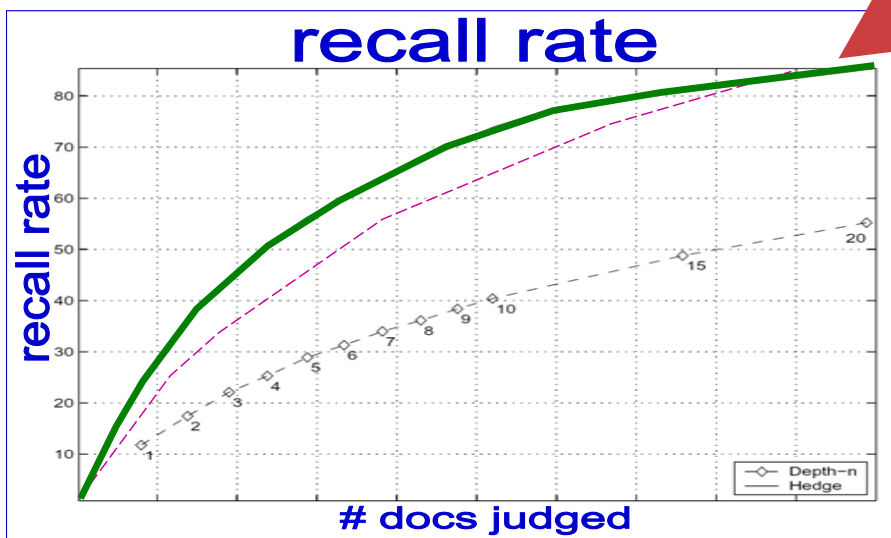
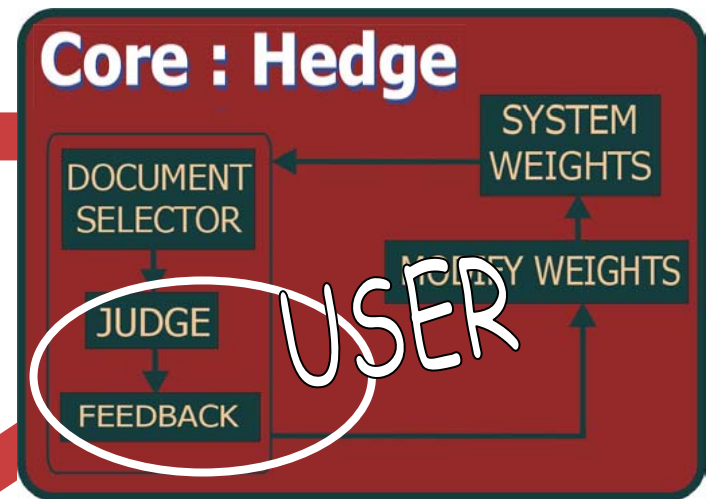
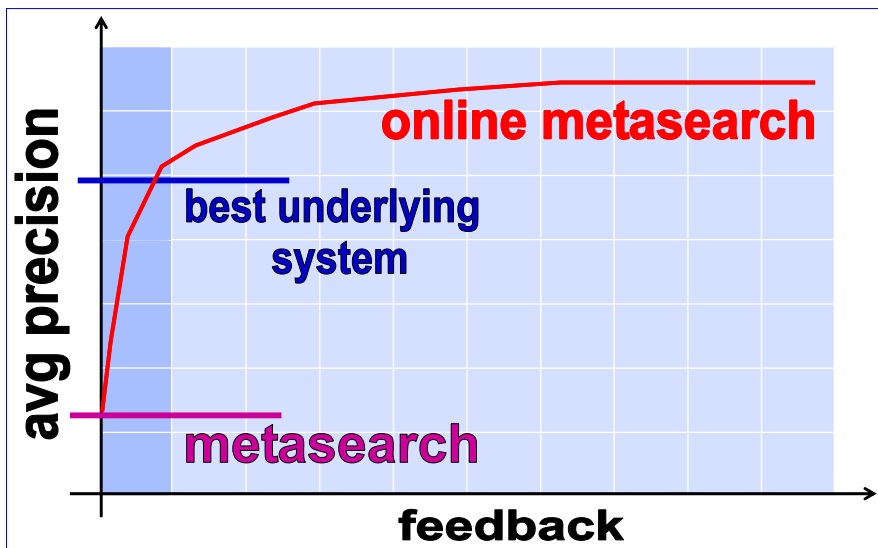


actually...we do more than metasearch

## system evaluation



# "a unified model for metasearch, pooling and system evaluation"





# conclusion

- theoretic explanation of “being adaptive”
- simple, elegant, intuitive
- usually performs much better than the bound

**END**

# AdaBoost - technical

$$D_1(i) = \frac{1}{m}$$

- Start with uniform distrib  $D_1$ :

- At every round  $t=1$  to  $T$

given  $D_t$

find weak hypothesis

with error

compute “belief” in  $h_t$

update distribution

$$h_t : X \rightarrow \{-1, 1\}$$

$$\varepsilon_t = \Pr_{D_t} [h_t(x_i) \neq y_i]$$

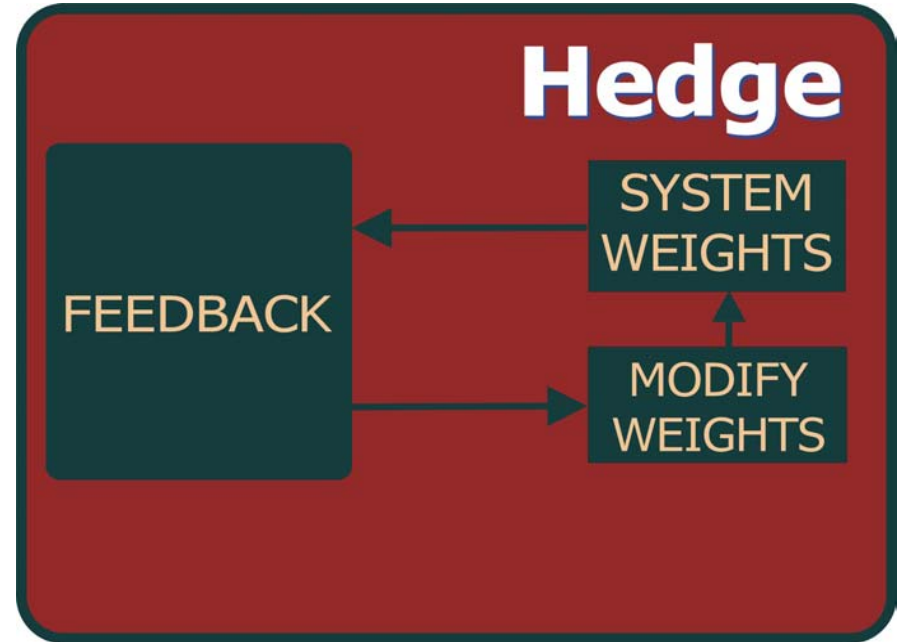
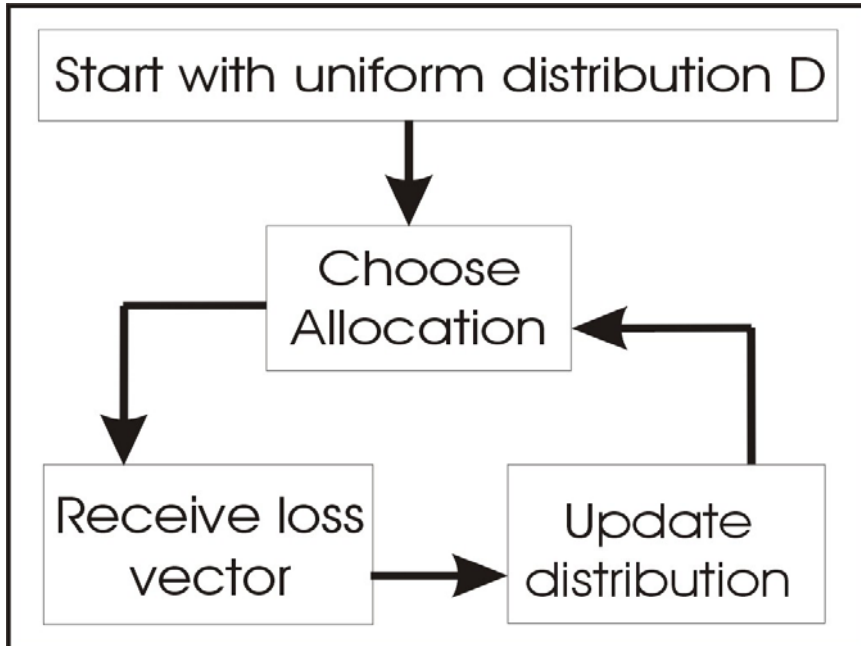
$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) > 0$$

$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

- final hyp

$$H_{\text{final}}(x) = \text{sgn} \left( \sum_t \alpha_t h_t(x) \right)$$

# On-line Setup



# On-line Metasearch [Aslam, Pavlu, Savell]

The screenshot shows a metasearch engine interface with three search engines: Google, altavista, and alltheweb. The search term is 'cikm 2003'. The results are displayed in a list format, with each engine's results appearing side-by-side. The Google results include links to 'CIKM 2003 Homepage', 'ACM CIKM 2003 Call For Papers', and 'CIKM-2003 Registration'. The altavista results show 'Did you mean: ci km 2003' and 'AltaVista found 5,186 results'. The alltheweb results show '1 - 10 of 6,293 Results for cikm 2003' and a list of search results including 'CIKM 2003 Homepage', 'IVML Call For Papers Archive', and 'ACM WIDM2003'.

**metasearch**

CIKM 2003 Homepage

ACM CIKM 2003 Call For Paper

CIKM-2003 Registration

ACM CIKM 2003 Call For Paper

CIKM 2002 Homepage

Conference on Information and Knowledge Management (CIKM)

[Asis-I] CIKM 2003

(DBWORLD) CIKM 2003 (Padm bridge-cikm-2003

Collaborative Filtering Mailing List ARL-ERESERVE@ar

Received: from cni.org by b.cni.org

ACM WIDM 2003

ACM - MMDB 2003

ACM CIKM 2003 PRELIMINARY Selected Publications

Web Caching Publications/Ver

cikm '03

Call for papers

Mario A. Nascimento - Personal

Received

Conferences and Journals on Information and Knowledge Management

Conferences On Information and Knowledge Management

Calendrier des manifestations (DBWORLD) Final Call for ACM

Collaborative Filtering Mailing

**metasearch**

CIKM 2003 Homepage

CIKM-2003 Registration

CIKM 2002 Homepage

ACM CIKM 2003 Call For Paper

ACM CIKM 2003 Call For Paper

CIKM Home Page ACM DL: CIKM

Conference on Information and Knowledge Management (CIKM)

ACM CIKM 2003 PRELIMINARY

CIKM Home Page ACM DL: CIKM

ACM CIKM 2003 PRELIMINARY

[Asis-I] CIKM 2003

(DBWORLD) CIKM 2003 (Padm bridge-cikm-2003

Collaborative Filtering Mailing List ARL-ERESERVE@ar

Received: from cni.org by b.cni.org

ACM WIDM 2003

ACM - MMDB 2003

Selected Publications

Web Caching Publications/Ver

cikm '03

Call for papers

Mario A. Nascimento - Personal

Received

Conferences and Journals on Information and Knowledge Management

Conferences On Information and Knowledge Management

Calendrier des manifestations

# hedge application : metasearch

- problem : metasearch
  - search engines
  - document judgment
  - metaserach
- [hedge] solution
  - search engines are “experts”
  - judgments on documents are “loses”

# an IR setup

# generalization error- based on margins

- training error [Schapire, Freund 1996]

$$\text{training error}(H_{\text{final}}) \leq \prod_t \sqrt{4\varepsilon_t(1-\varepsilon_t)}$$

- generalization error [Schapire, Freund, Barlett, Lee 1998]

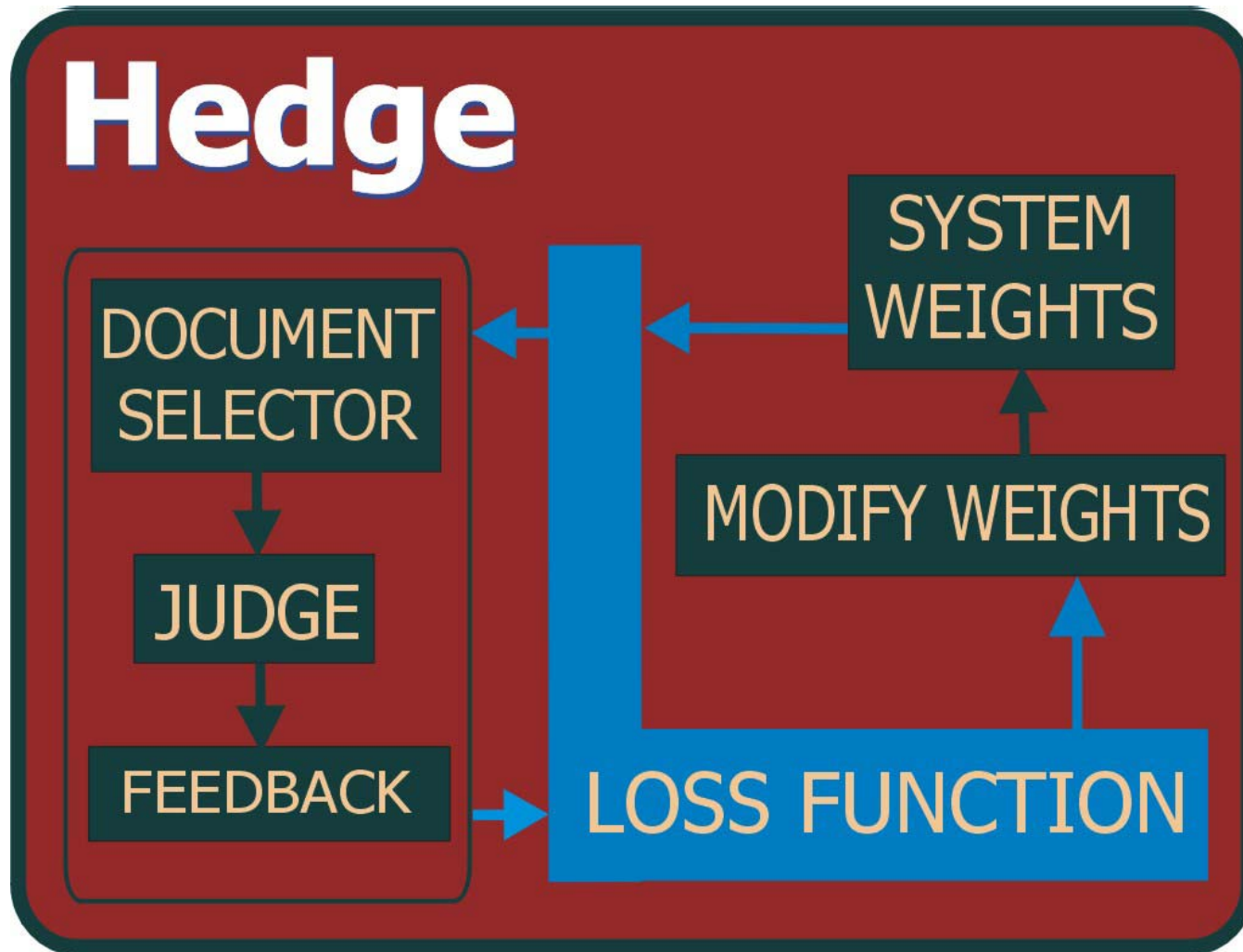
**Theorem 1** Let  $\mathcal{D}$  be a distribution over  $X \times \{-1, 1\}$ , and let  $S$  be a sample of  $m$  examples chosen independently at random according to  $\mathcal{D}$ . Assume that the base-classifier space  $\mathcal{H}$  is finite, and let  $\delta > 0$ . Then with probability at least  $1 - \delta$  over the random choice of the training set  $S$ , every weighted average function  $f \in \mathcal{C}$  satisfies the following bound for all  $\theta > 0$ :

$$\mathbf{P}_{\mathcal{D}} [yf(x) \leq 0] \leq \mathbf{P}_S [yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \left(\frac{\log m \log |\mathcal{H}|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right).$$

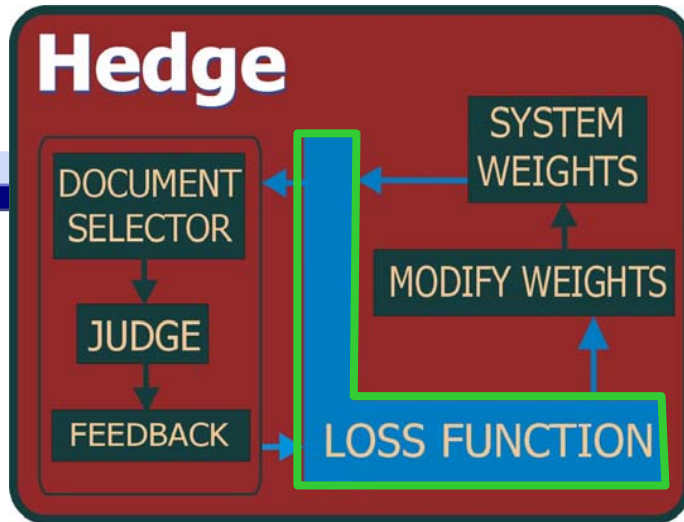
**testing error**                      **training miss-confidence**                      **does not depend on T**



# hedge approach



# loss function



$RELEVANT = -1$   
 $NONRELEVANT = +1$

- map ranks to values

$$value(r) = \frac{1}{r} + \frac{1}{r+1} + \dots + \frac{1}{Z} \approx \ln \frac{Z}{r}$$

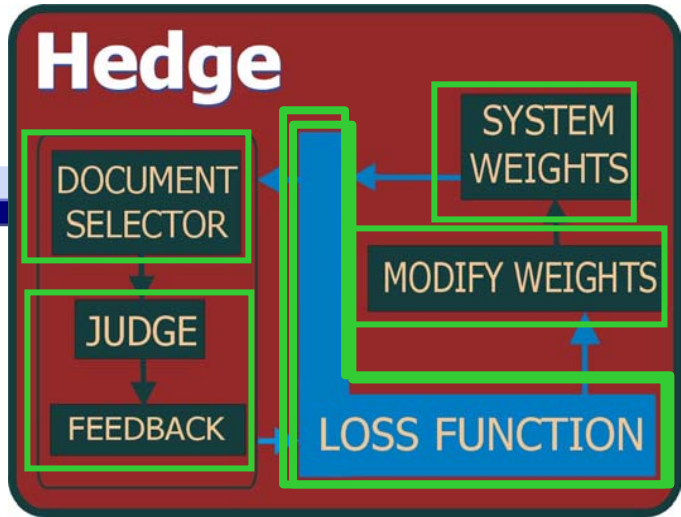
$$LOSS(d, s) = label(d) \cdot value(rank_{d,s}) \approx label(d) \cdot \ln \frac{Z}{r}$$

$TP(s) =$  total precision of  $s =$  average of precision values at all ranks

$$FACT : \sum_{\text{all docs}} LOSS(d, s) = C * (Z - \underline{2 * TP(s)})$$

- Average the precision at **ALL** ranks
- Normalize so ideal system gets  $TP=1$
- math is simpler

ork ?



$$average\_value_t(d) = \sum_{s=1}^N w_s^{t-1} \cdot value(rank_{d,s})$$

**metasearch**

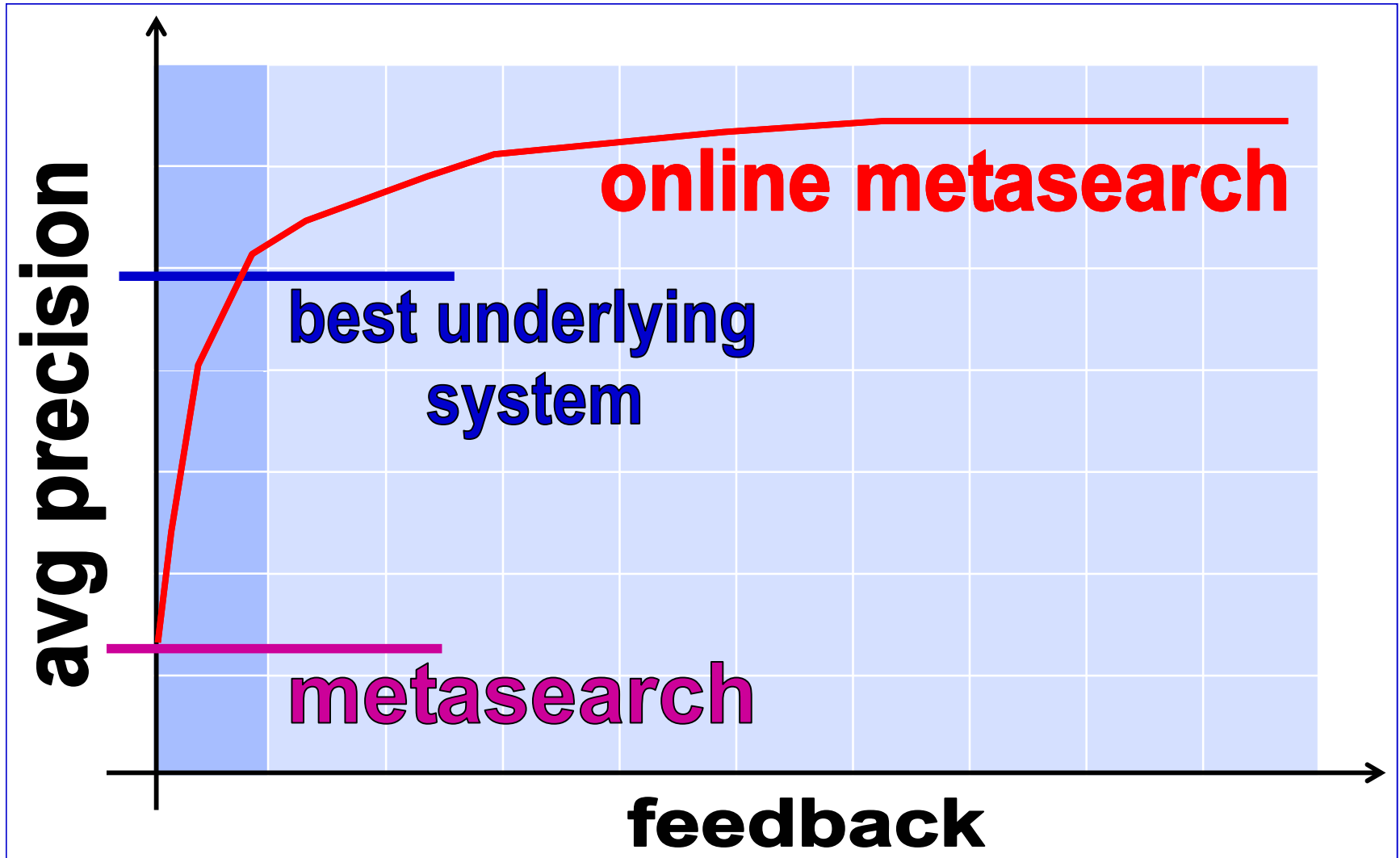
- CIKM 2003 Homepage
- CIKM-2003 Registration
- CIKM 2002 Homepage
- ACM CIKM 2003 Call For Paper
- ACM CIKM 2003 Call For Paper
- CIKM Home Page ACM DL: CIK
- Conference on Information an
- ACM CIKM 2003 PRELIMINARY
- CIKM Home Page ACM DL: CIK
- ACM CIKM 2003 PRELIMINARY
- [Asis-] CIKM 2003
- (DBWORLD) CIKM 2003 (Padm
- bridge-cikm-2003
- Collaborative Filtering Mailing
- Yahoo! Groups : webir Messag
- dBforums - Cfp: Cikm 03
- Mailing List ARL-ERESERVE@ar
- Received: from cni.org by b.cr
- ACM WIDM 2003
- ACM - MMDR 2003

- average\_value at episode t
  - “trust” in systems change with t
- metasearch : include already judged docs
- pooling  $d_t = \arg \max_d (average\_value(d))$ 
  - system evaluation
- get feedback  $LOSS(d_t, s) = label(d_t) \cdot value(rank_{d_t, s})$
- modify weights

$$w_s^{t+1} = w_s^t \cdot \beta^{LOSS_t(d_t, s)}$$

Call for papers  
 Mario A. Nascimento - Person  
 lceived  
 Conferences and Journals on T  
 Conferences On Information V  
 Calendrier des manifestations

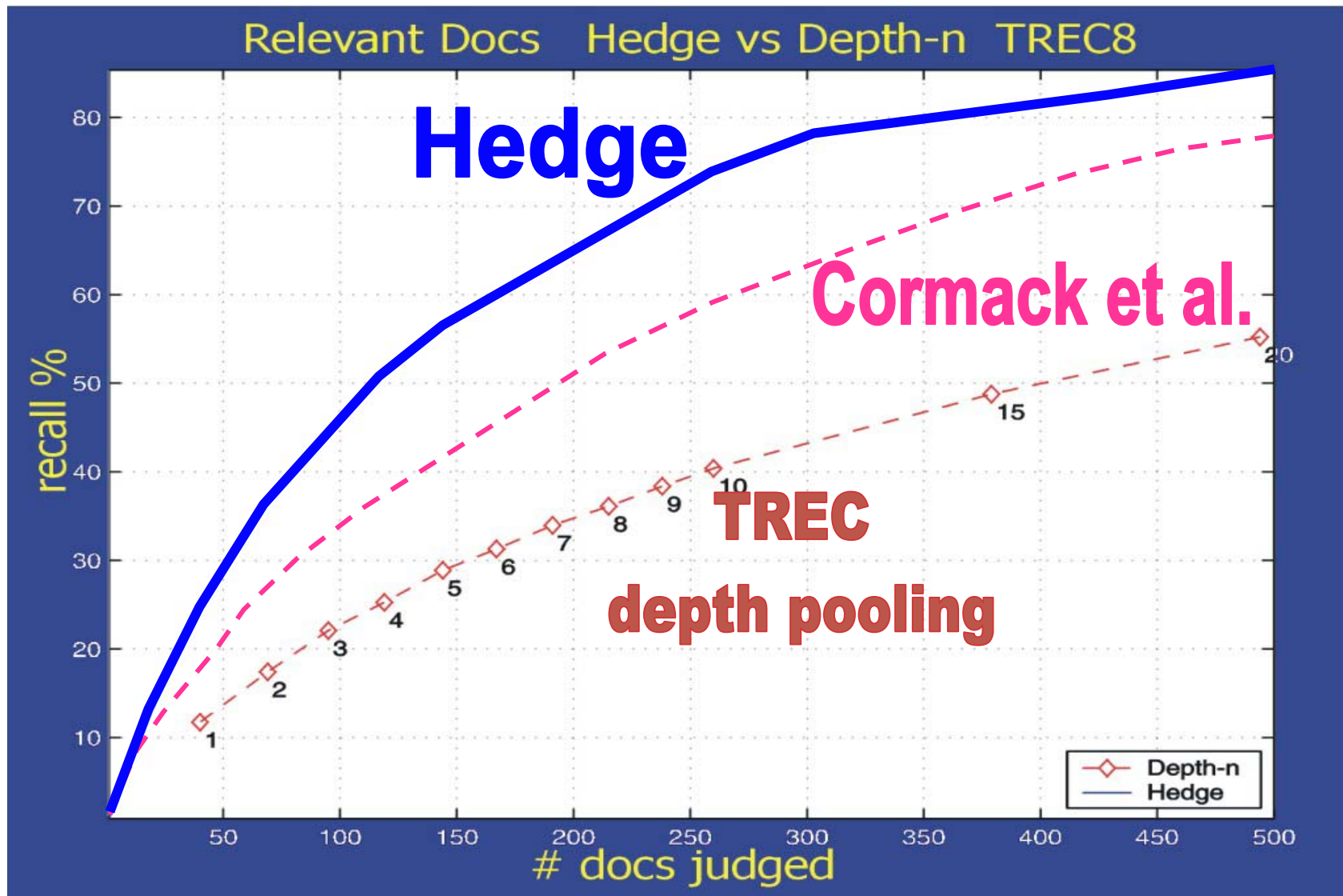
# actually...we do more than metasearch



# experiments

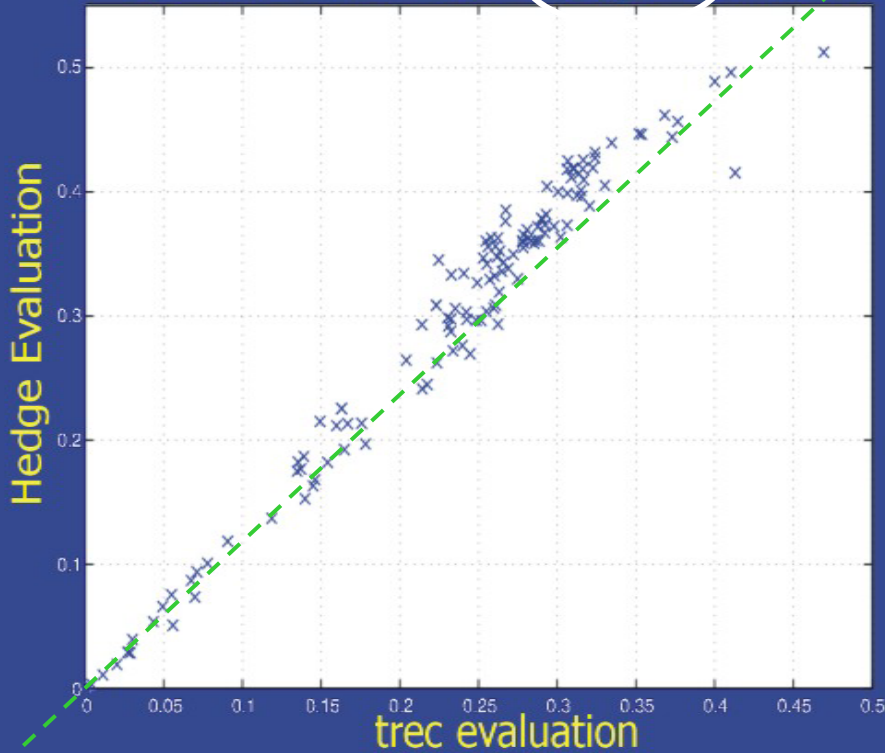
- TREC 3,5,6,7,8
  - 41-129 systems
  - 50 queries per TREC
  - metasearch uses *all* systems
- use TREC judgments as user feedback
- system evaluation : incomplete judgments

# experiments - relevant docs found

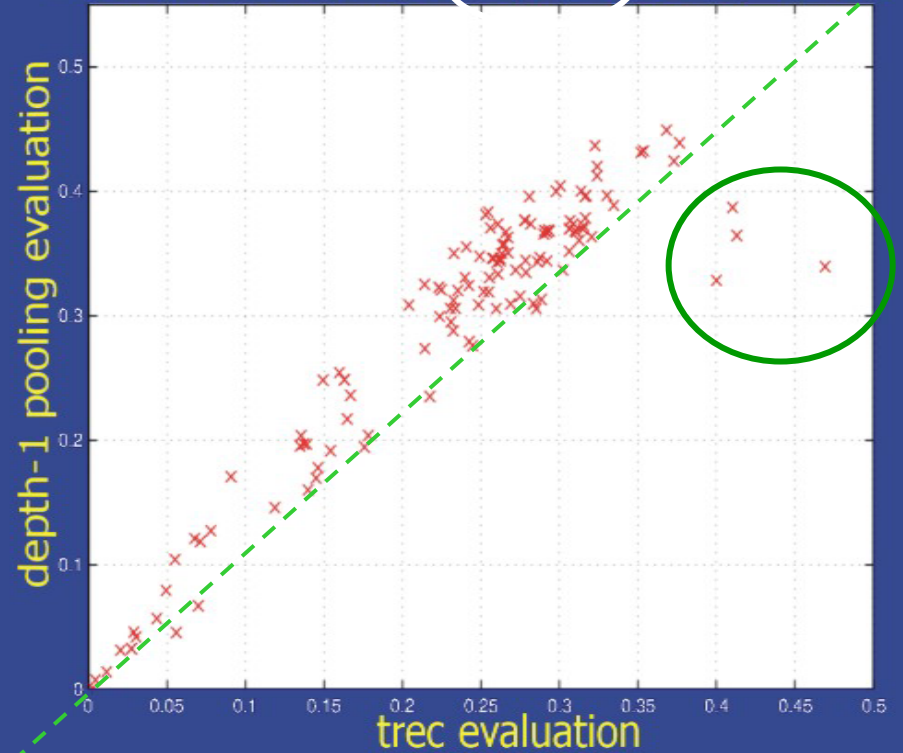


# experiments - system evaluation

SYSTEM EVALUATION Hedge-40 TREC8

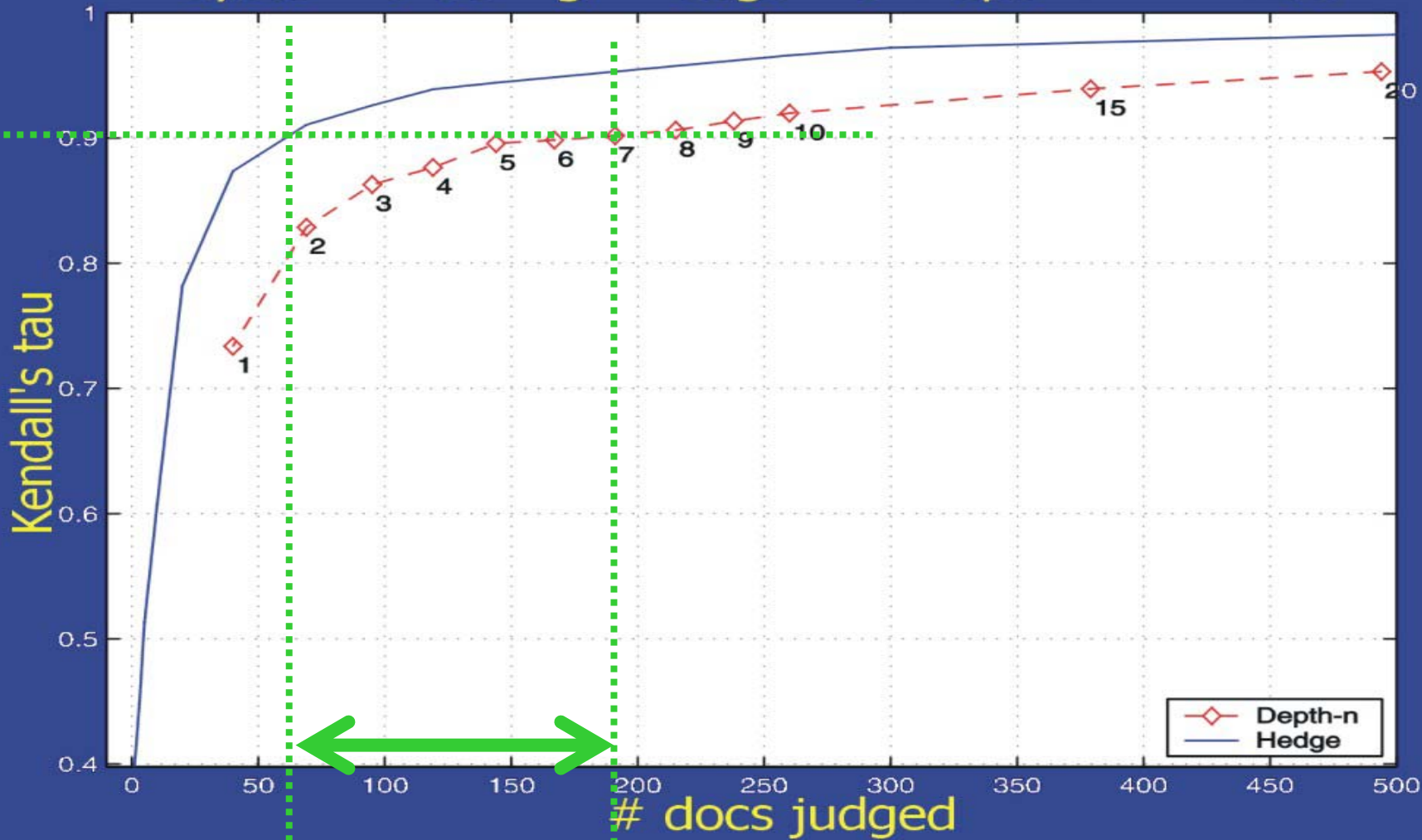


SYSTEM EVALUATION depth-1 pooling TREC8



# system evaluation – kendall's $\tau$

System Ordering Hedge vs Depth-n TREC8





# metasearch - no feedback (yet)

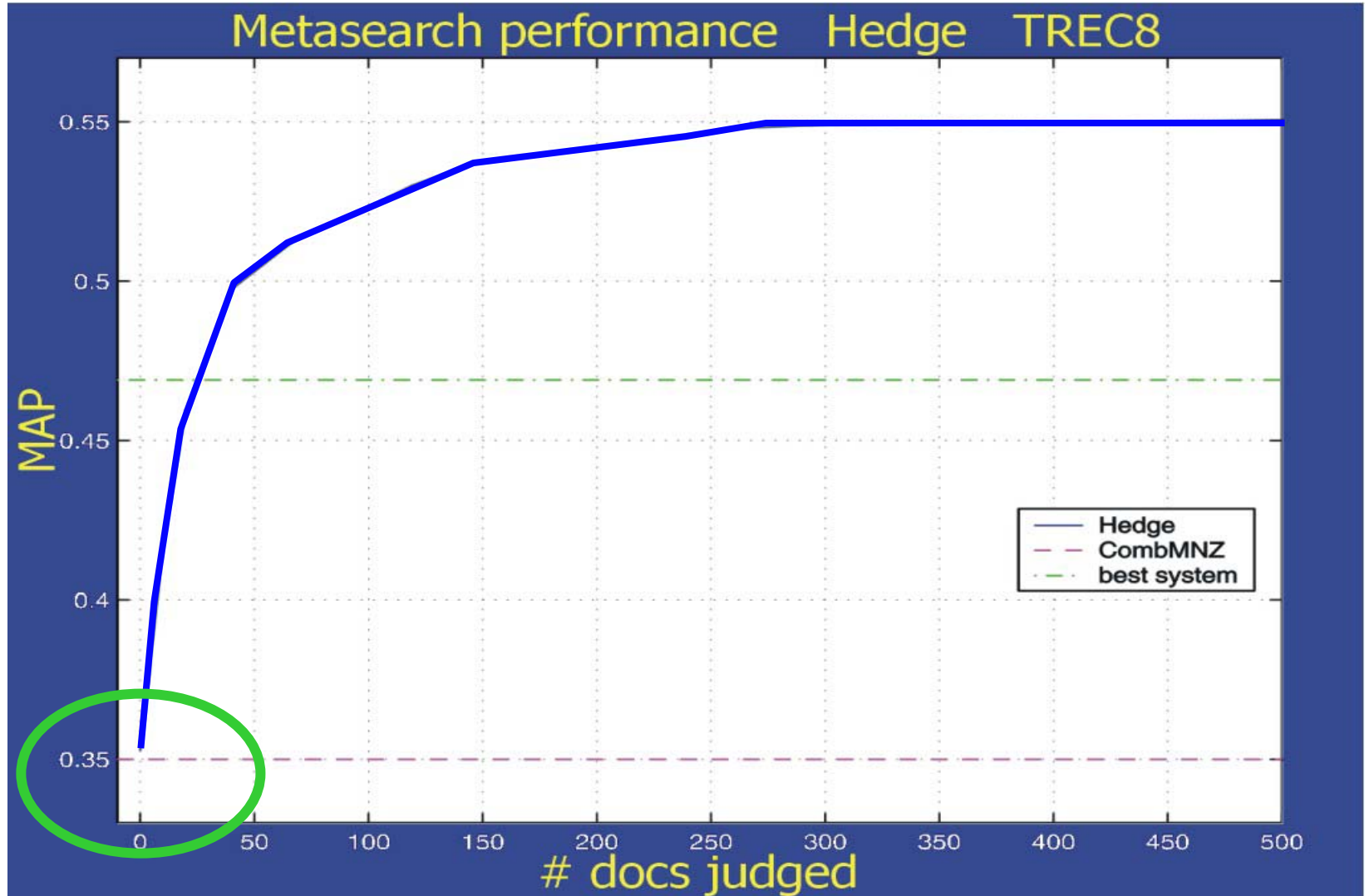
no relevant judgements → uniform weights

TREC	MNZ	COND	Hedge-0	%MNZ	%COND
3	0.423	0.403	0.418	-1.2	+3.7
5	0.294	0.307	0.309	+5.1	+0.6
6	0.341	0.315	0.345	+1.2	+9.5
7	0.320	0.308	0.323	+0.9	+4.9
8	0.350	0.343	0.352	+1.4	+2.6

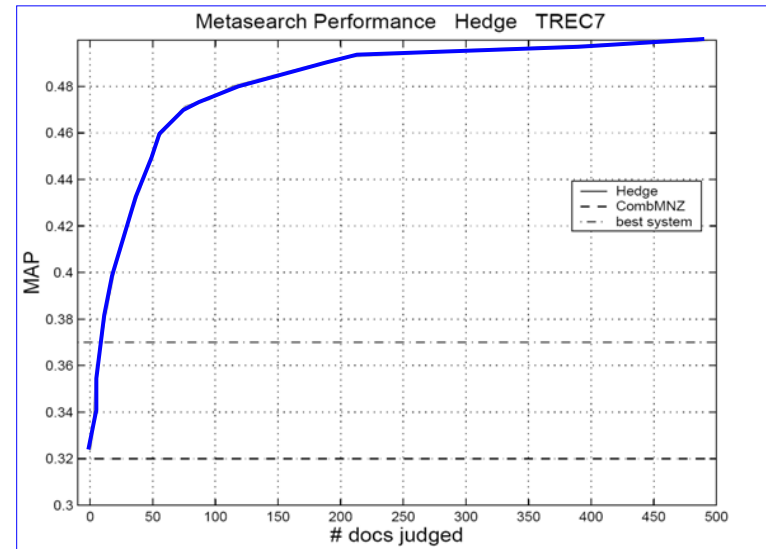
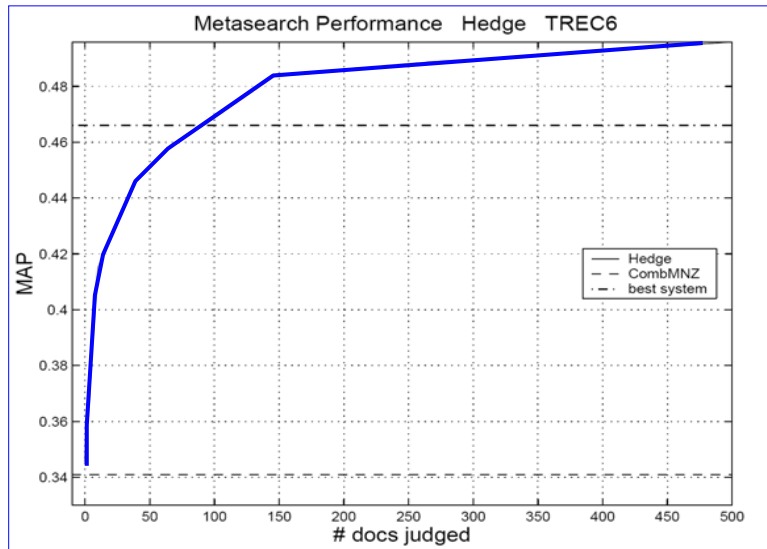
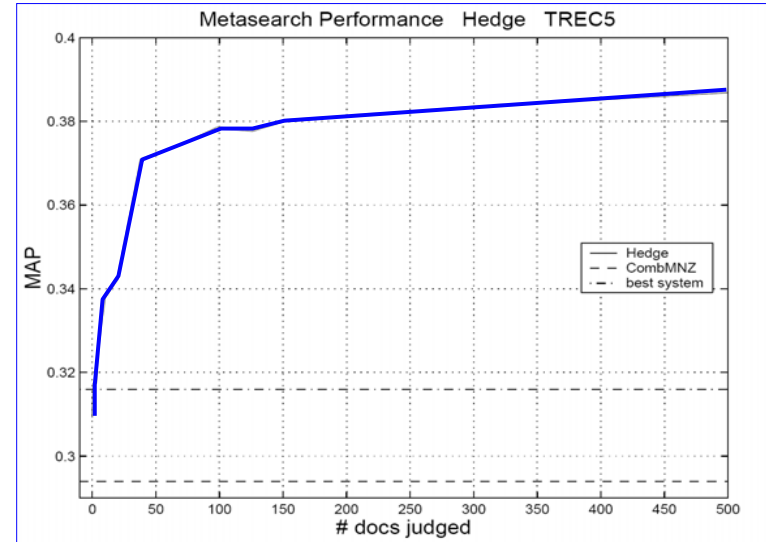
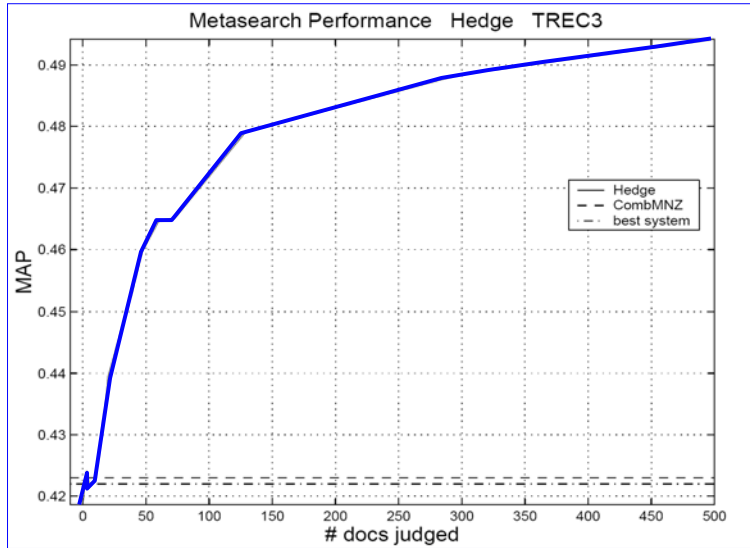
MNZ=CombMNZ(Fox,Shaw,Lee et al)

COND=Condorcet(Aslam,Montague)

# experiments – metasearch – TREC8



# metasearch – TREC 3,5,6,7



# conclusion

- a powerful machine learning approach
  - Hedge = AdaBoost core
- works [usually] better than anything else we've seen
- true, it uses feedback
  - but without feedback there are provable limitations

# Why hedge [schapire, freund '96]

$$\sum_{i=1}^N w_i^{T+1} \leq \left( \sum_{i=1}^N w_i^T \right) \left( 1 - (1 - \beta) p^T \cdot l^T \right) \leq \dots$$

$$\dots \leq \exp(- (1 - \beta) \sum_{t=1}^T p^t \cdot l^t)$$

hedge loss at episode  $T = p^T \cdot l^T$

cummulat loss  $L_{HEDGE}$

$$L_{HEDGE} \leq \frac{\ln\left(\frac{1}{\beta}\right) L_{SYSTEM} + \ln N}{1 - \beta}$$

$L_{Hedge(\beta)}$

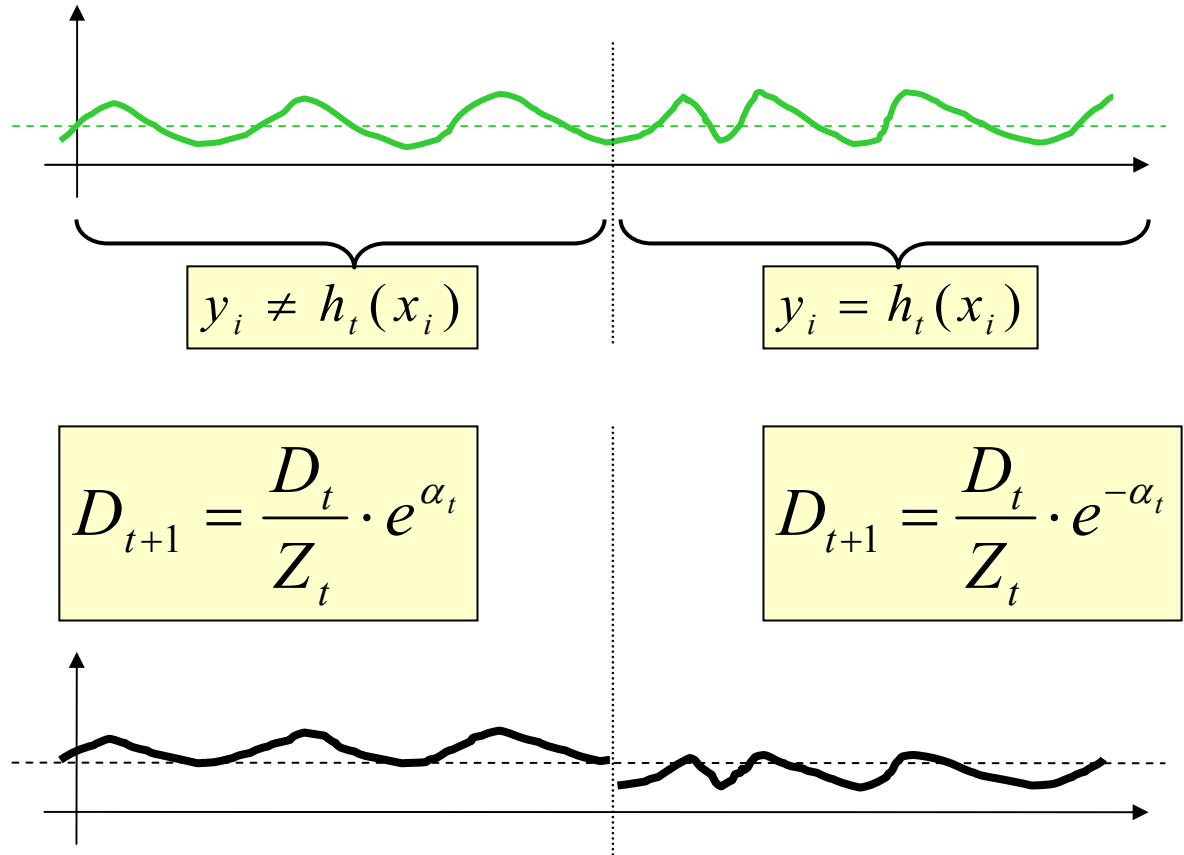
$$\sum_{i=1}^N w_i^{T+1}$$

# AdaBoost – distribution update

$$h_t : X \rightarrow \{-1, 1\}$$

$$\varepsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$$

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) > 0$$



- “neutralize” the last weak hypothesis

# Application - metasearch

Google Advanced Search Preferences Language  
Searched the web for cikm 2003

**CIKM 2003 Homepage**  
... 03, Hotel Inter-Continental, Novem  
Conference on Information and Know  
Description: ACM Conference on Info  
Category: Reference > Knowledge M  
bit.csc.lsu.edu/~cikm2003/ - 9k - Ca

**ACM CIKM 2003 Call Fo**  
... Management (CIKM03). AC  
highest caliber papers submit  
bit.csc.lsu.edu/~cikm2003/CI  
[ More results from bit.csc.lsu.

**CIKM-2003 Registration**  
... All Amounts in USD. Attendee Sta  
one author of an accepted paper mus  
www.cikmreg.org/ - 13k - Cached - S

**ACM CIKM 2003 Call For Pap**  
ACM CIKM 2003 -Call for Industry T  
Sponsored by ACM SIGIR and ACM  
www.cs.wisc.edu/dbworld/messages

**CIKM 2002 Homepage**  
... 03), Las Vegas, Nevada, June 23-  
Conference on Information and Know  
Description: SAIC Headquarters, McLea  
Category: Business > Management  
www.cikm.org/2002/ - 15k - Cached

**Conference on Informati**  
... The CIKM 2003 web page is  
devoted to emerging areas of  
www.cikm.org/ - 6k - Cached

**[Asis-I] CIKM 2003**  
[Asis-I] CIKM 2003. Padmini Srinivas  
Mon, 29 Sep 2003 12:59:36 -0500: P  
mail.asis.org/pipermail/asis-I/2003-S

**(DBWORLD) CIKM 2003 (Pad**  
(DBWORLD) CIKM 2003 (Padmini S  
CIKM 2003; From: "Padmini Srinivas  
wwwiti.cs.uni-magdeburg.de/MailA

**bridge-cikm-2003**  
bit.csc.lsu.edu/~cikm2003.  
www.logistics-2000-versailles.net/site

**Collaborative Filtering Mailing**  
[collab@siml] CFP: CIKM 2003. ...

**altavista**  
cikm 2003  
SEARCH: Worldwide U.S.  
Did you mean: ci km 2003  
AltaVista found 5,186 results About  
**CIKM 2003 Homepage**  
12th international conference on info  
bit.csc.lsu.edu/~cikm2003 • Refresh  
More pages from bit.csc.lsu.edu  
**ACM CIKM 2003 Call For**  
... Management (CIKM03) AC  
caliber papers submitted to C  
bit.csc.lsu.edu/~cikm2003/C  
More pages from bit.csc.lsu.edu  
**CIKM 2002 Homepage**  
... 2002 Advance Technical Program  
Treasurer, Dr. Nicholas, the CIKM02  
www.cikm.org/2002 • Related Pages  
**Conference on Information**  
CIKM CIKM 2003 CIKM Topi  
Knowledge Management The  
www.cikm.org • Related Page  
More pages from www.cikm.o  
**CIKM**  
... Information and Knowledge Mana  
CIKM 2003 Home Page 11. CIKM 2  
www.informatik.uni-trier.de/~ley/db/c  
More pages from www.informatik.un

**7. CIKM 1998: Bethesda,**  
7. CIKM 1998: Bethesda, Ma  
Copyright © by Michael Ley (l  
www.informatik.uni-trier.de/~l  
More pages from www.inform

**SIGIR Information Server**  
... award. Upcoming SIGIR Spons  
2003. JCDL 2004 (Tucson, AZ) - Jur  
www.acm.org/sigir • Refreshed in pa  
More pages from www.acm.org

**Web Caching Publications/Ven**  
2000 Books 2001 Books 2002 Books 2003 net

## metasearch

**CIKM 2003 Homepage**  
ACM CIKM 2003 Call For Paper  
CIKM-2003 Registration  
ACM CIKM 2003 Call For Paper  
CIKM 2002 Homepage  
Conference on Information and  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm  
bridge-cikm-2003  
Collaborative Filtering Mailing  
CIKM Home Page ACM DL: CIK  
Yahoo! Groups : webir Messag  
dbforums - Cfp: Cikm 03  
Mailing List ARL-ERESERVE@a  
Received: from cni.org by b.cr  
ACM WIDM 2003  
ACM - MMDB 2003  
ACM CIKM 2003 PRELIMINARY  
Selected Publications  
Web Caching Publications/Ven  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iced  
Conferences and Journals on T  
Conferences On Information V  
Calendrier des manifestations  
(DBWORLD) Final Call for ACM  
Collaborative Filtering Mailing

## metasearch

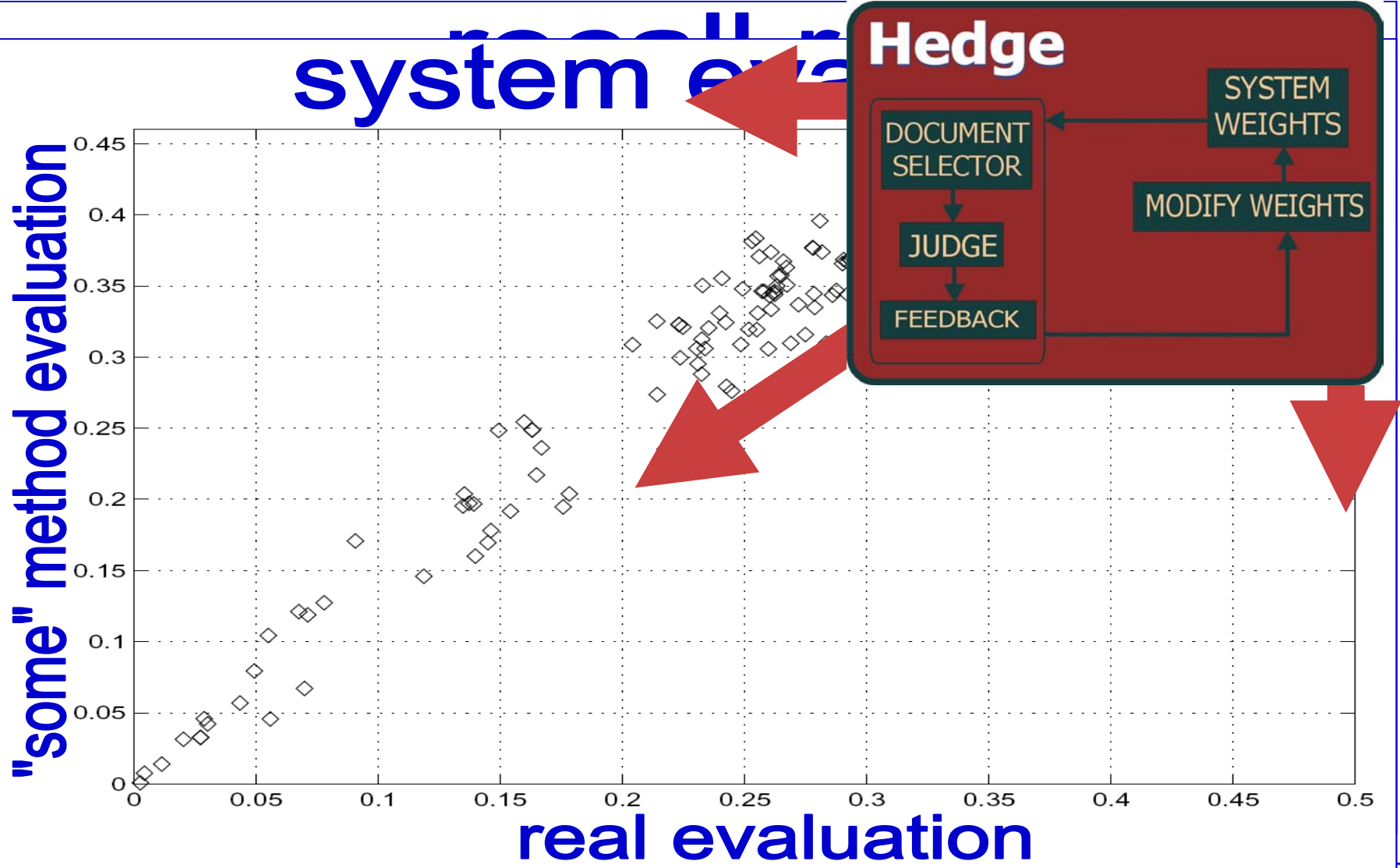
**CIKM 2003 Homepage**  
**CIKM-2003 Registration**  
CIKM 2002 Homepage  
ACM CIKM 2003 Call For Paper  
ACM CIKM 2003 Call For Paper  
CIKM Home Page ACM DL: CIK  
Conference on Information and  
ACM CIKM 2003 PRELIMINARY  
CIKM Home Page ACM DL: CIK  
ACM CIKM 2003 PRELIMINARY  
[Asis-I] CIKM 2003  
(DBWORLD) CIKM 2003 (Padm  
bridge-cikm-2003  
Collaborative Filtering Mailing  
Yahoo! Groups : webir Messag  
dbforums - Cfp: Cikm 03  
Mailing List ARL-ERESERVE@a  
Received: from cni.org by b.cr  
ACM WIDM 2003  
ACM - MMDB 2003  
Selected Publications  
Web Caching Publications/Ven  
cikm '03  
Call for papers  
Mario A. Nascimento - Person  
Iced  
Conferences and Journals on T  
Conferences On Information V  
Calendrier des manifestations

# problem setup

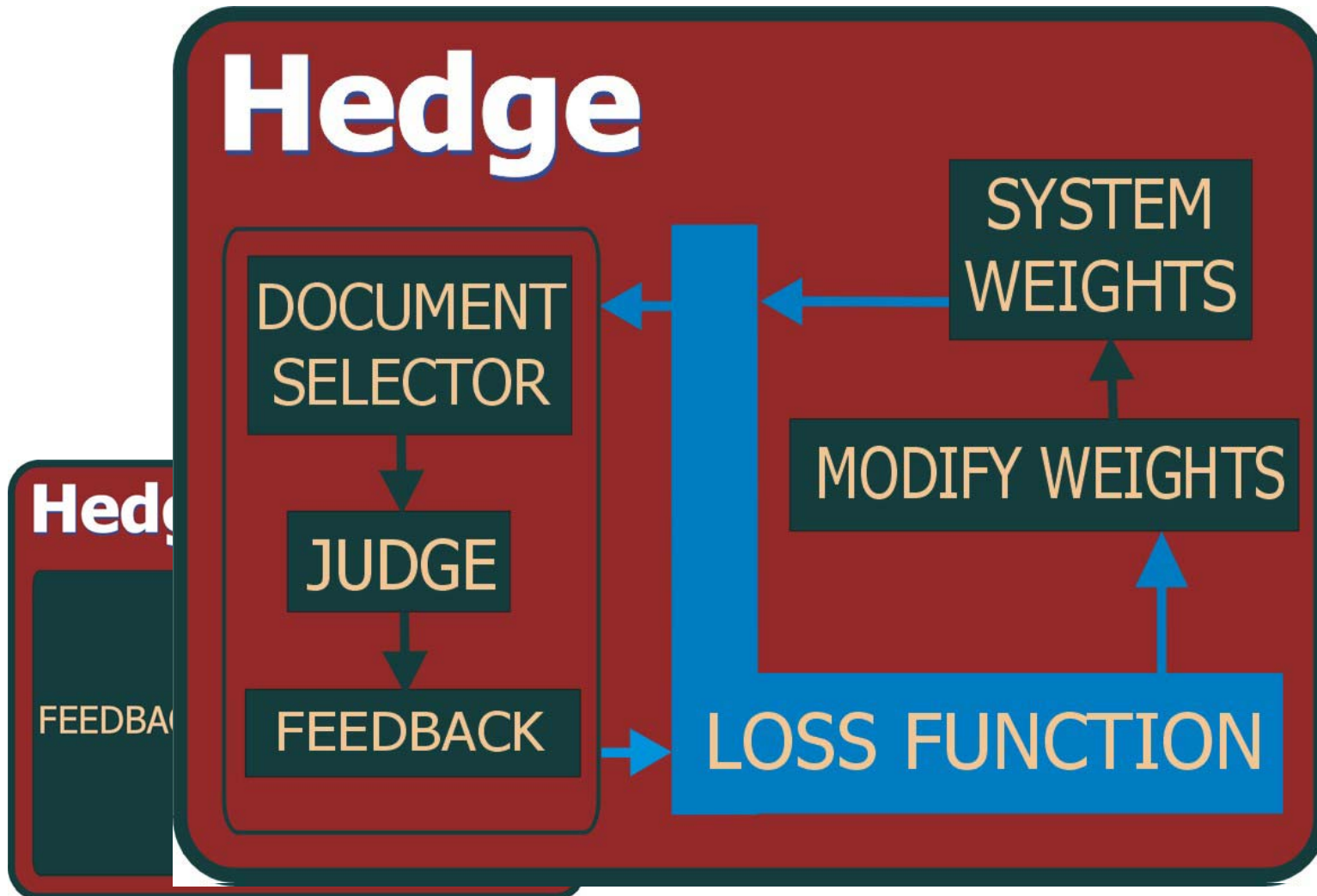
- On the same query
- Set of underlying systems
- User feedback
- Goal
  - Find relevant documents
  - Produce online metasearch lists
  - Perform online system evaluation
- We are looking for an adaptive approach



# motivation



# our unified model



# loss function

## Hedge

SYSTEM

$$LOSS(d, s) = \text{label}(d) * \left( \frac{1}{r} + \frac{1}{r+1} + \dots + \frac{1}{Z} \right) \approx \text{label}(d) * \ln \frac{Z}{r}$$

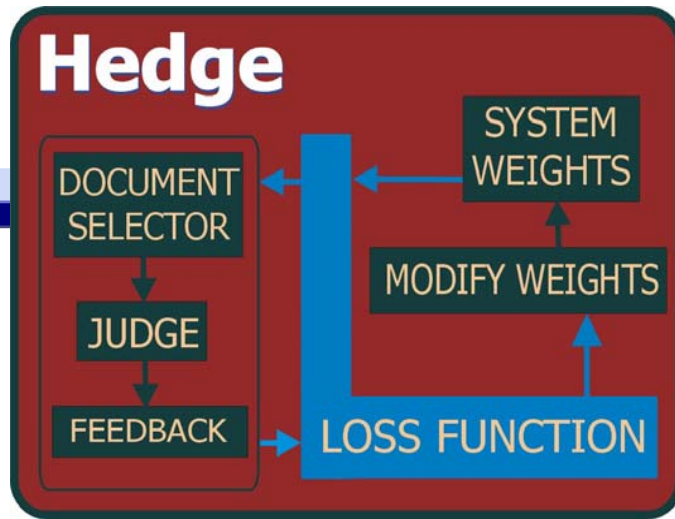
MODIFY WEIGHTS  
magnitude

$TP(s)$  = total precision of  $s$  = average precision at all ranks

FACT:  $\sum_{\text{all docs}} LOSS(d, s) = C * (Z - 2 * \underline{TP(s)})$

- Average the precision at **ALL** ranks
- Normalize so ideal system gets  $TP=1$
- math is more simple

# pooling – howto



pooling value(d)

$$\text{- select } d_t = \arg \max_{d \text{ not labeled}} \left[ \sum_{s=1}^T w_s^{t-1} * LOSS(d, s \mid d = NR) \right]$$

- Naturally “want” top ranks
- If NON RELEVANT, then a NR in top ranks of the system lists
- If RELEVANT, bingo.

# system evaluation – howto

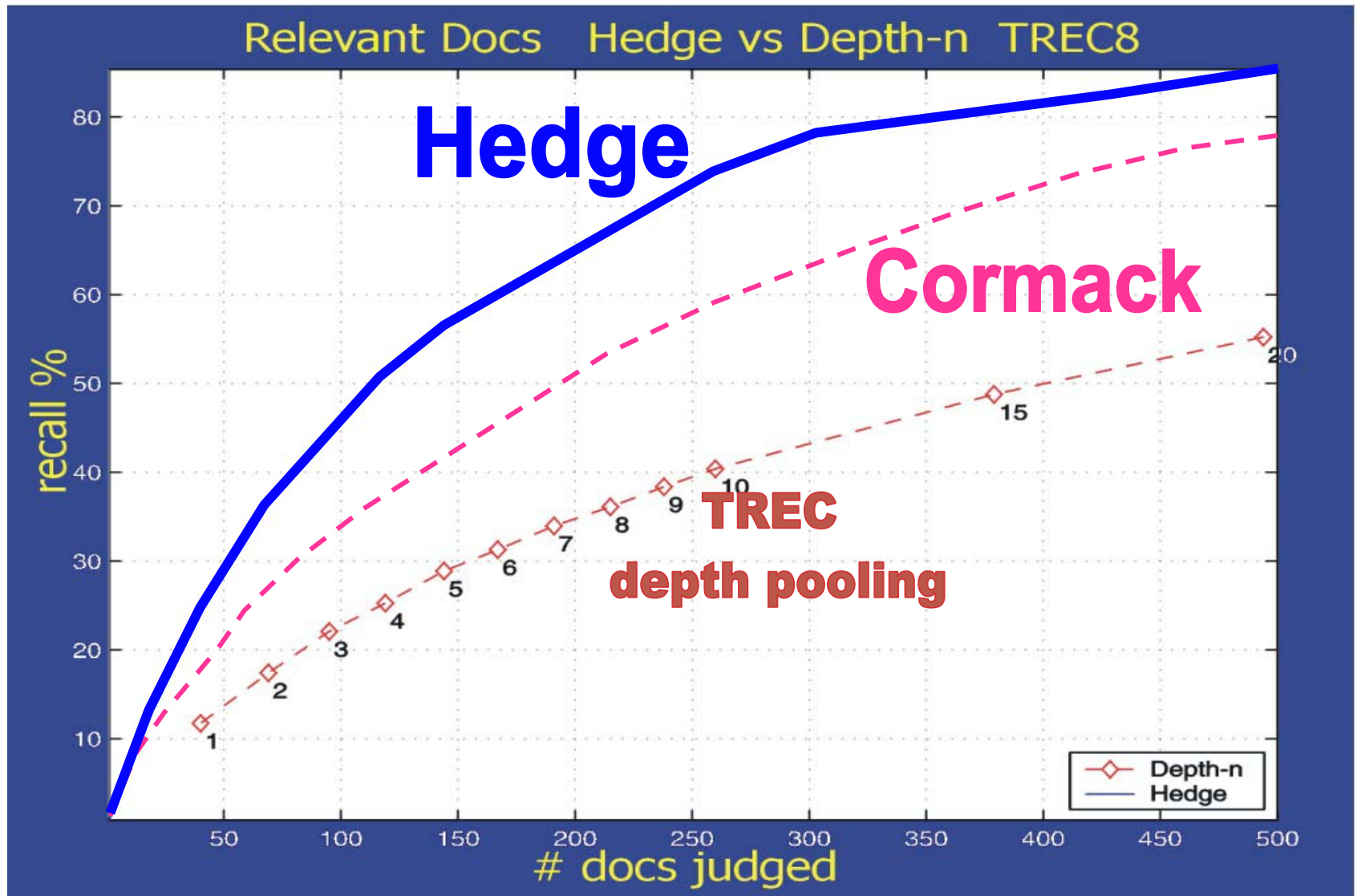
- assume **all** docs not judged (**so many ?**) to be NON RELEVANT
- compute AvegPrecision for every system
- one (or few) very good systems – use small  $\beta$

# metasearch - howto

- Compute “pooling value” for each doc
  - Instead of “select the top doc” for pooling
  - do “select the top 1000 doc” for metasearch
- almost 1000 – docs already pooled are automatically in top of metasearch list

- TREC
  - ~100 systems
  - 50 queries each competition
- Use TREC qrels as user feedback
  - incomplete feedback
- For comparison with depth-pooling we use average number of pools (over queries)

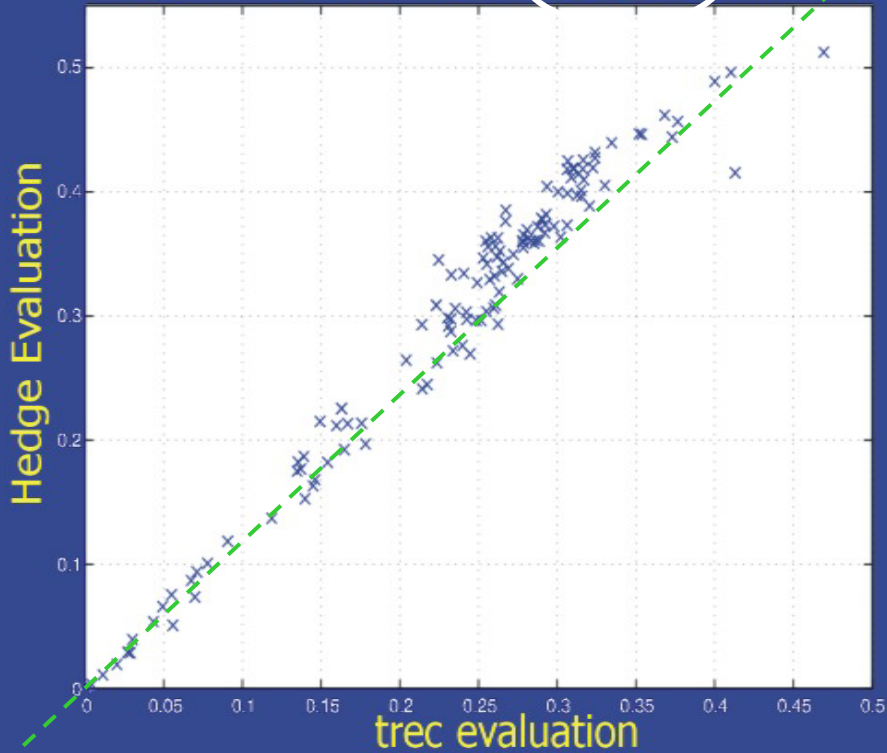
# experiments - relevant docs found



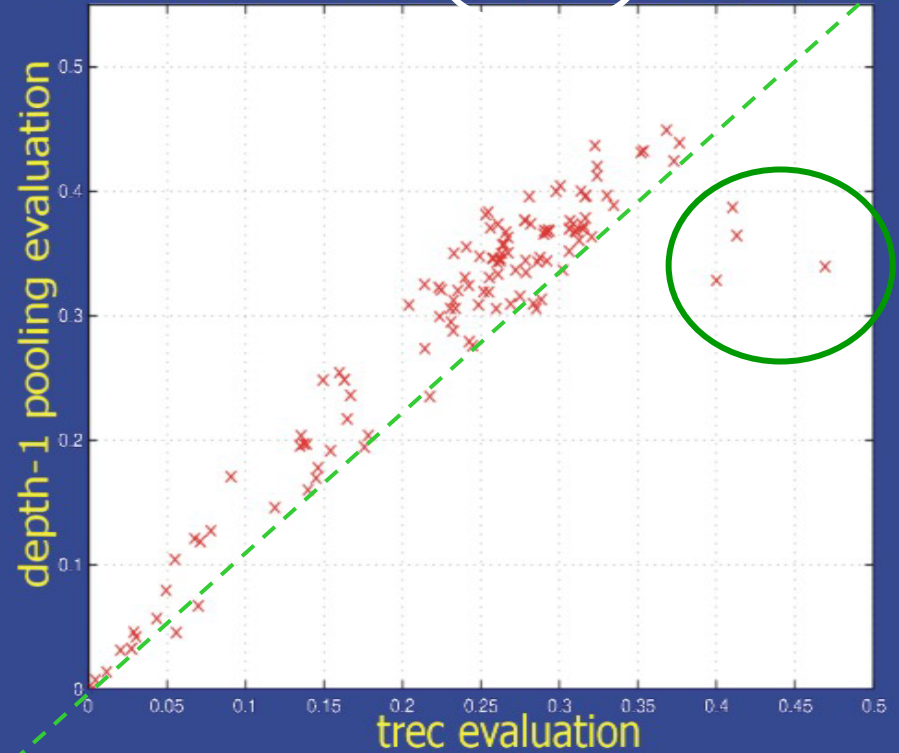


# experiments - system evaluation

SYSTEM EVALUATION Hedge-40 TREC8

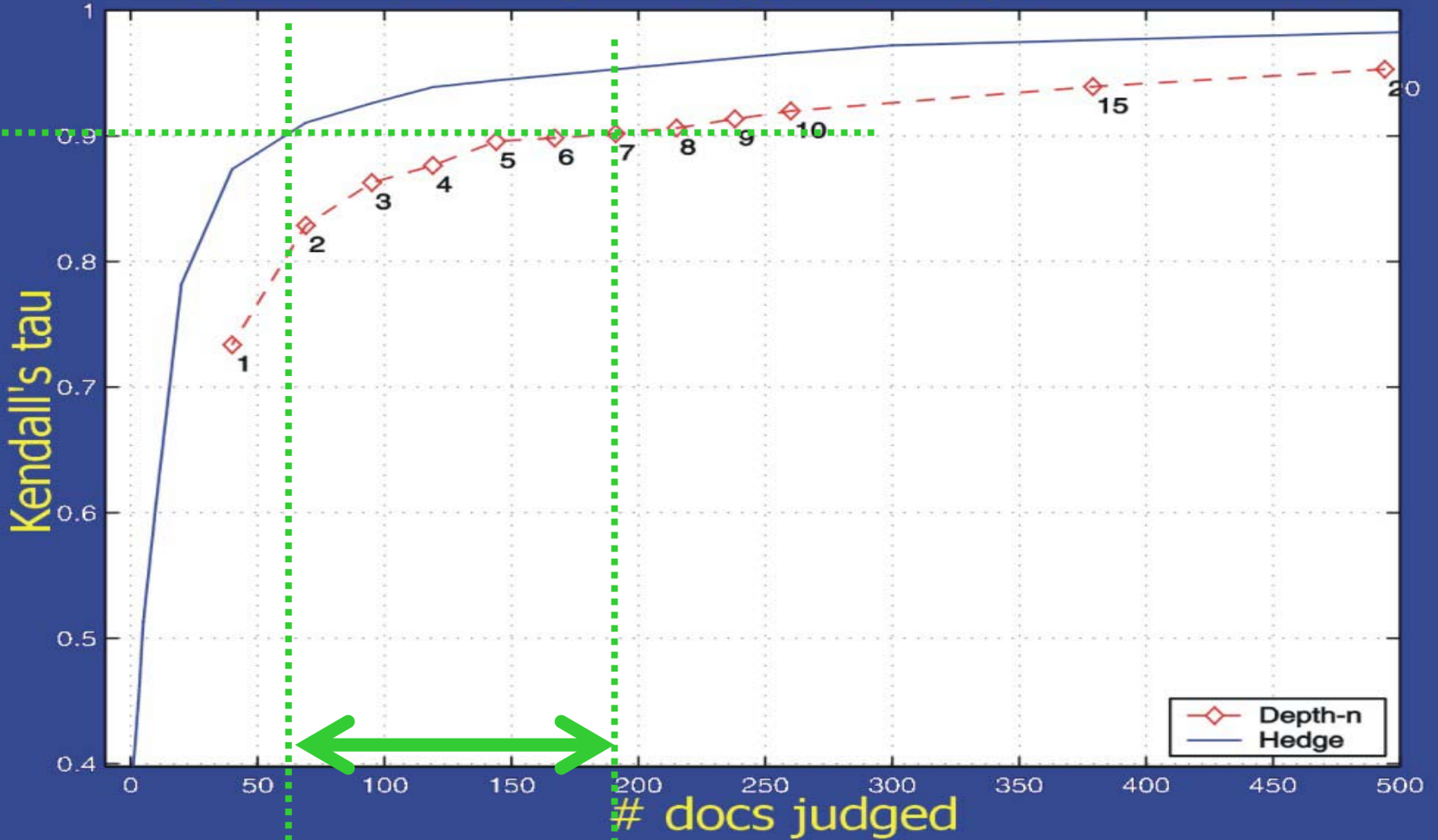


SYSTEM EVALUATION depth-1 pooling TREC8



# system evaluation – kendall's tau

System Ordering Hedge vs Depth-n TREC8

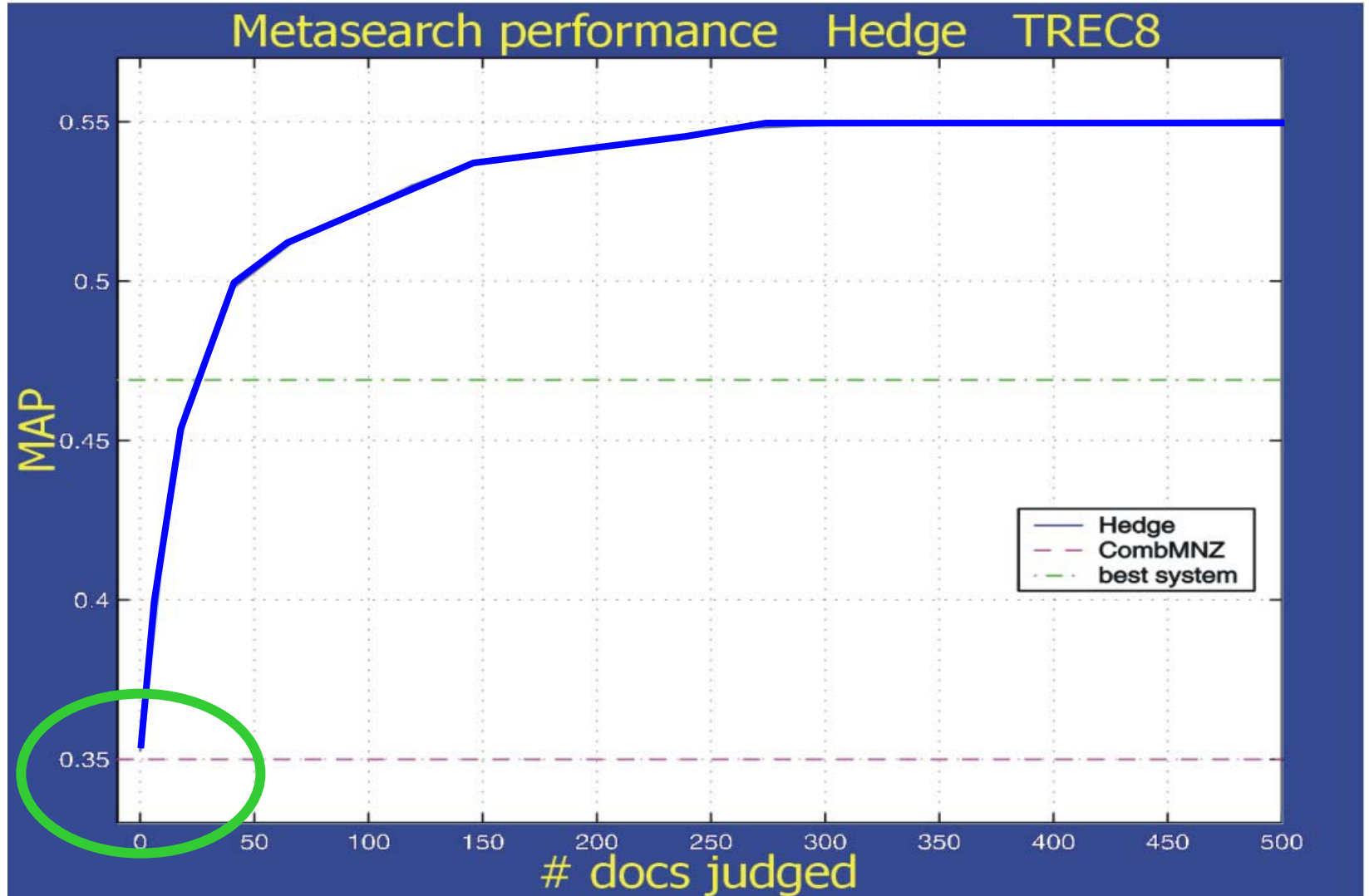


# metasearch - no feedback (yet)

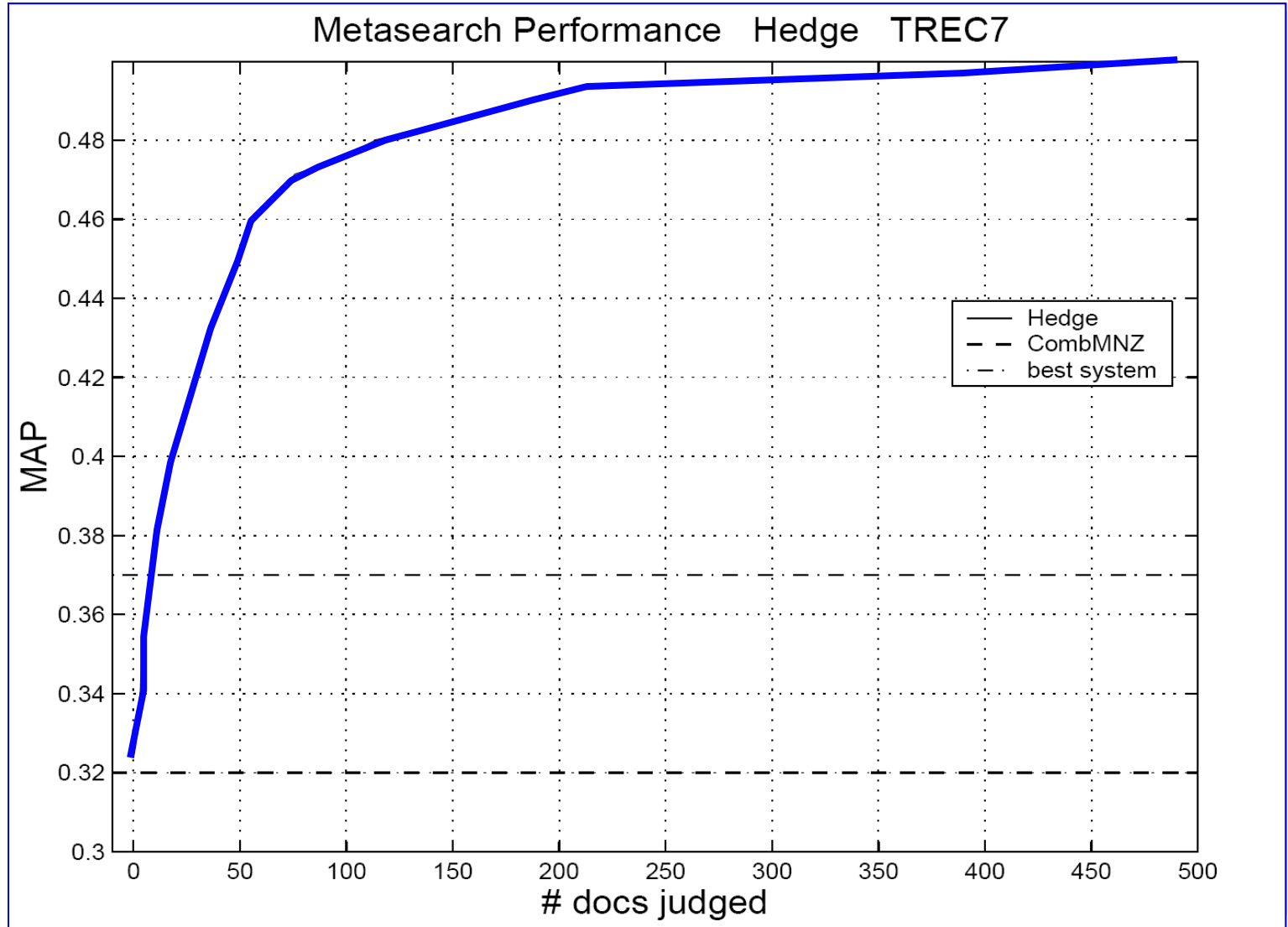
no relevant  
judgements

TREC	MNZ	COND	Hedge-0	%MNZ	%COND
3	0.423	0.403	0.418	-1.2	+3.7
5	0.294	0.307	0.309	+5.1	+0.6
6	0.341	0.315	0.345	+1.2	+9.5
7	0.320	0.308	0.323	+0.9	+4.9
8	0.350	0.343	0.352	+1.4	+2.6

# experiments – metasearch – TREC8



# metasearch – TREC 3,5,6,7



# conclusion

- A powerful machine learning approach
  - Hedge = AdaBoost core
- Works [usually] better than anything else we've seen
- True, it uses feedback
  - But without feedback there are provable limitations
- It is missing a rigorous analysis
  - We are not very far away with that
  - Need a model assumption

# differentiate (classify) the search engines

RANK	SYS1	SYS2	SYS3	SYS129
1	8	111	2	1013
2	2	2	821	
3	1	4	13	
4	134	7284	9	5
5	14215	769	106	2038
6	24	84	439	4
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1000	8521	11003	84	3967

84=RELEVANT      821=NON RELEVANT

## TREC8 competition

129 search engines (systems)

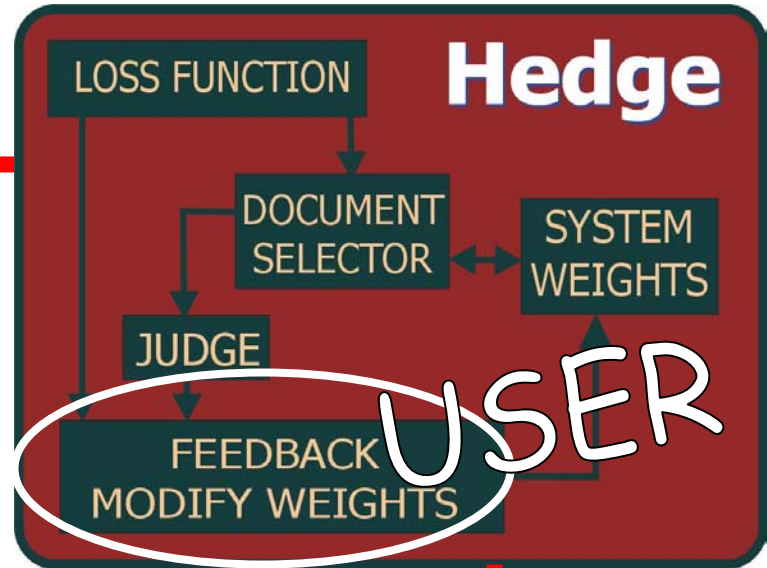
~1000 docs returned /query

TREC Evaluation DEPTH 100

-judge top 100 docs (each sys); use MAP to rank systems

SYSTEM	ORDERING (MAP)
READWARE2	0.469245750504929
orcl99man	0.413044919590457
iit99mal	0.41036111615777
AP	0.40000000658
CL99To	0.3855272016
CL99KT	0.3804322631
CL99SD	0.353710984082113
CL99SDopt2	0.351955930155801
8manexT3D1N0	0.334646677365296
pir9Aatd	0.330293582307575
ok8alx	0.324046294809153
Flab8atdn	0.323972048518037
MITSLstdn	0.32268558626124
pir9Attd	0.320651128642633
ok8amxc	0.31687789001205
att99atde	0.316549112790725
Flab8ax	0.316270864248346
apl8p	0.315425192468865
apl8c221	0.314991172213836
UniNET8Lg	0.313842898958697
apl8c621	0.312624929467966
fub99tt	0.31068000938879
fub99tf	0.309923242027166
att99atdc	0.308935872304471
fub99a	0.306768289435422
GE8MTD2	0.306560441783452
fub99td	0.306376965574526
pir9At0	0.306288595615414
pir9Atd0	0.302238406288713

# After conclusion - don't read



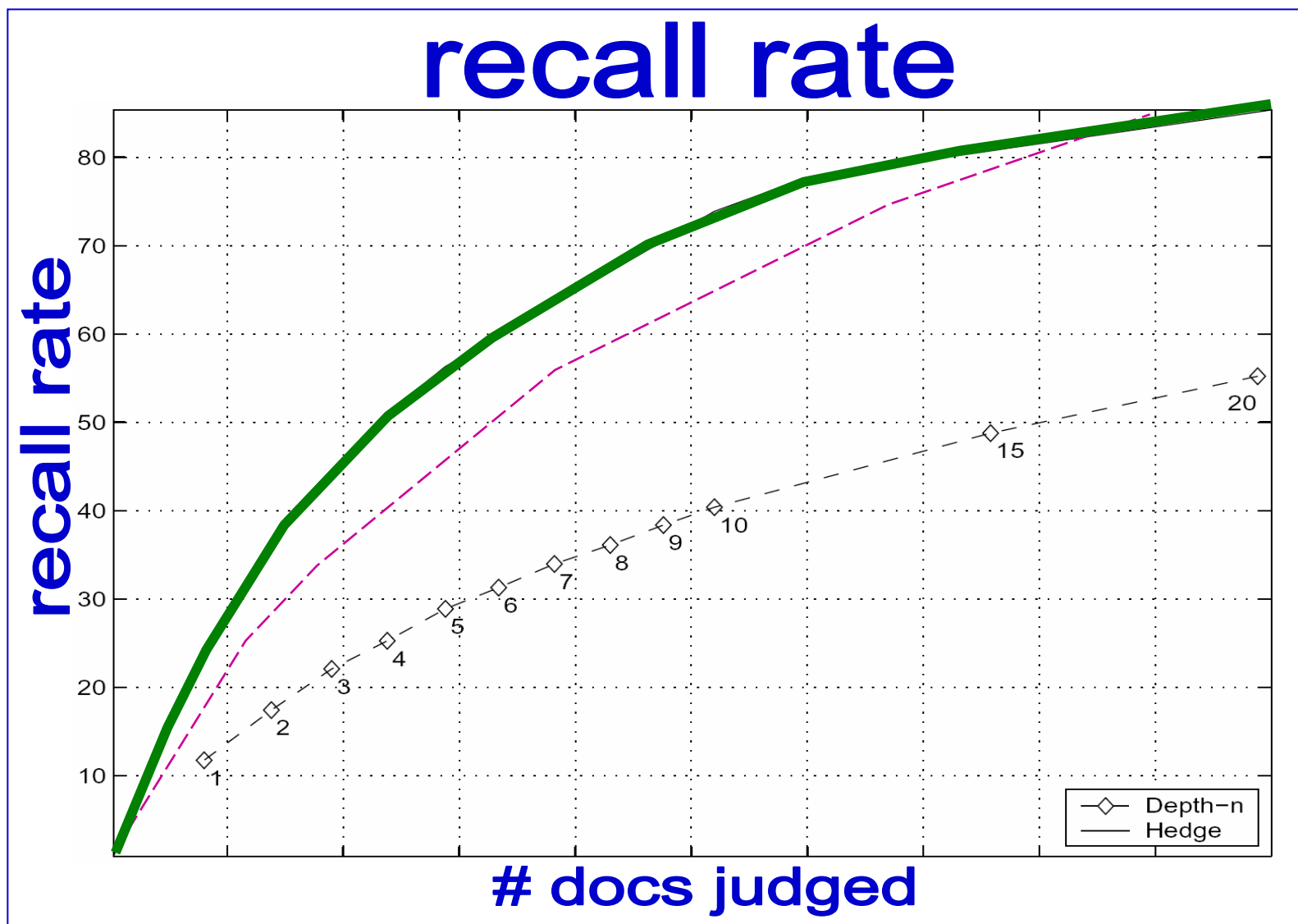
## fuse the lists (metasearch)

toronto sigir



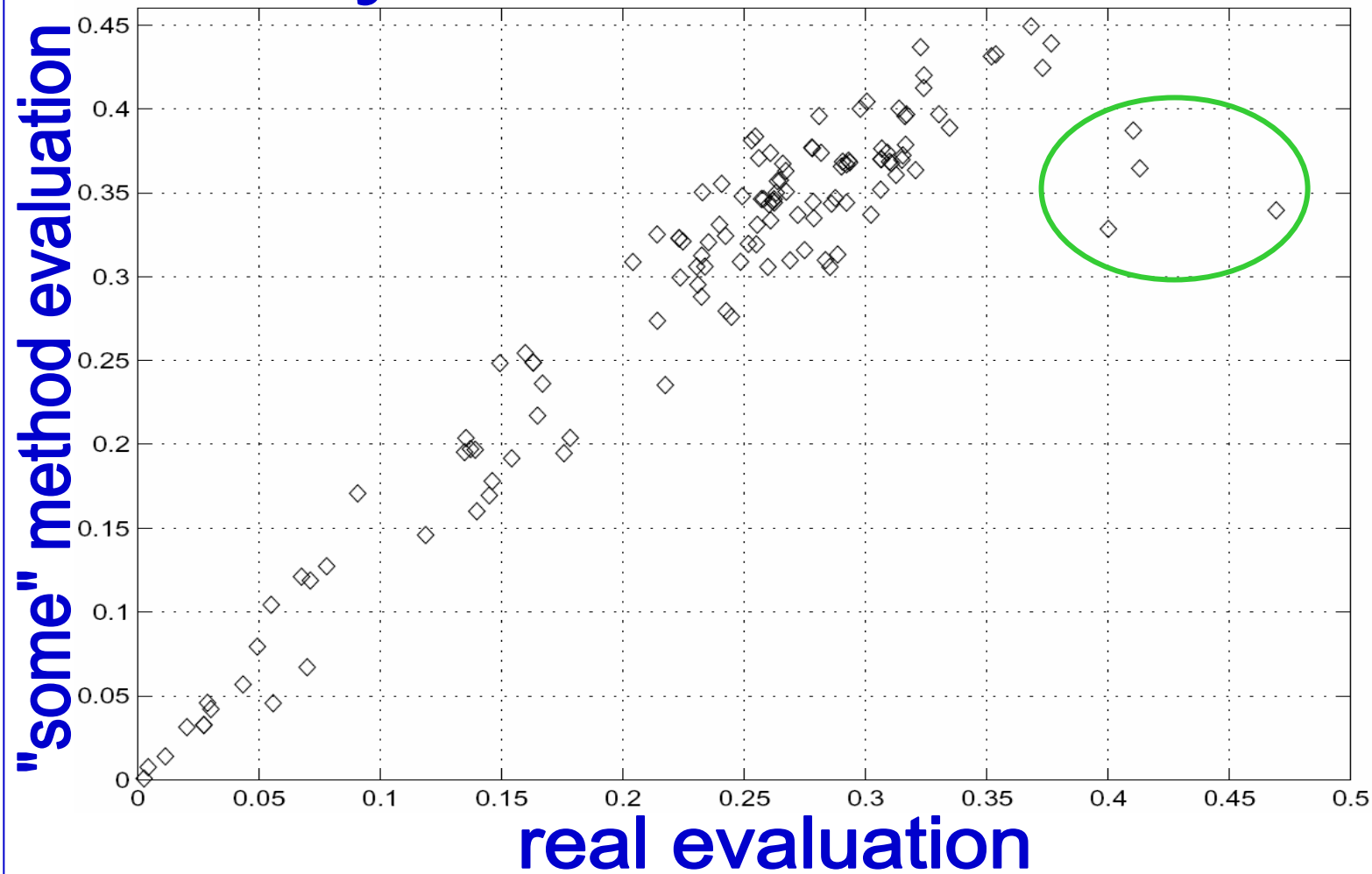
- metasearch list
- 1 SIGIR 2003
  - 2 Information Retrieval Conferences
  - 3 SIGIR Information Server
  - 4 SIGIR Upcoming events
  - 5 PROPOSAL FOR A SIGIR 2003 WORKSHOP
  - 6 Andreas S. WEIGEND, PhD
  - 7 [Sigir-] 2003 SIGIR Conference -
  - 8 CFP : Multimedia Information Retrieval
  - 9 The 25th ACM SIGIR 2002 Conference
  - 10 Mark Grolami's home page -
  - 11 Conferences on Information Retrieval
  - 12 OntoWeb - SIGIR 2003 - 2003-01-
  - 13 Yahoo! Groups : webir Messages
  - 14 DBWorld Message / 2002-
  - 15 Publications and Presentations of Thomas
  - 16 SIGIR 2003 Workshop on the Evaluation
  - 17 David Crow: Upcoming Conferences
  - 18 Upcoming conferences for the WUME Lab
  - 19 SIGIR 99 Preliminary Program
  - 20 SIGIR 99 CFP/PC
  - 21 Mounia Lalmas's list of activities
  - 22 A Partial List of Publications of Hongyuan
  - 23 ACM SIGIR 2002 Tutorial Bibliography
  - 24 SIGIR <http://www.informatik.uni-
  - 25 Information Retrieval Link's
  - 26 iSchool Events

**KEY:**  
FIND RELEVANT  
DOCS

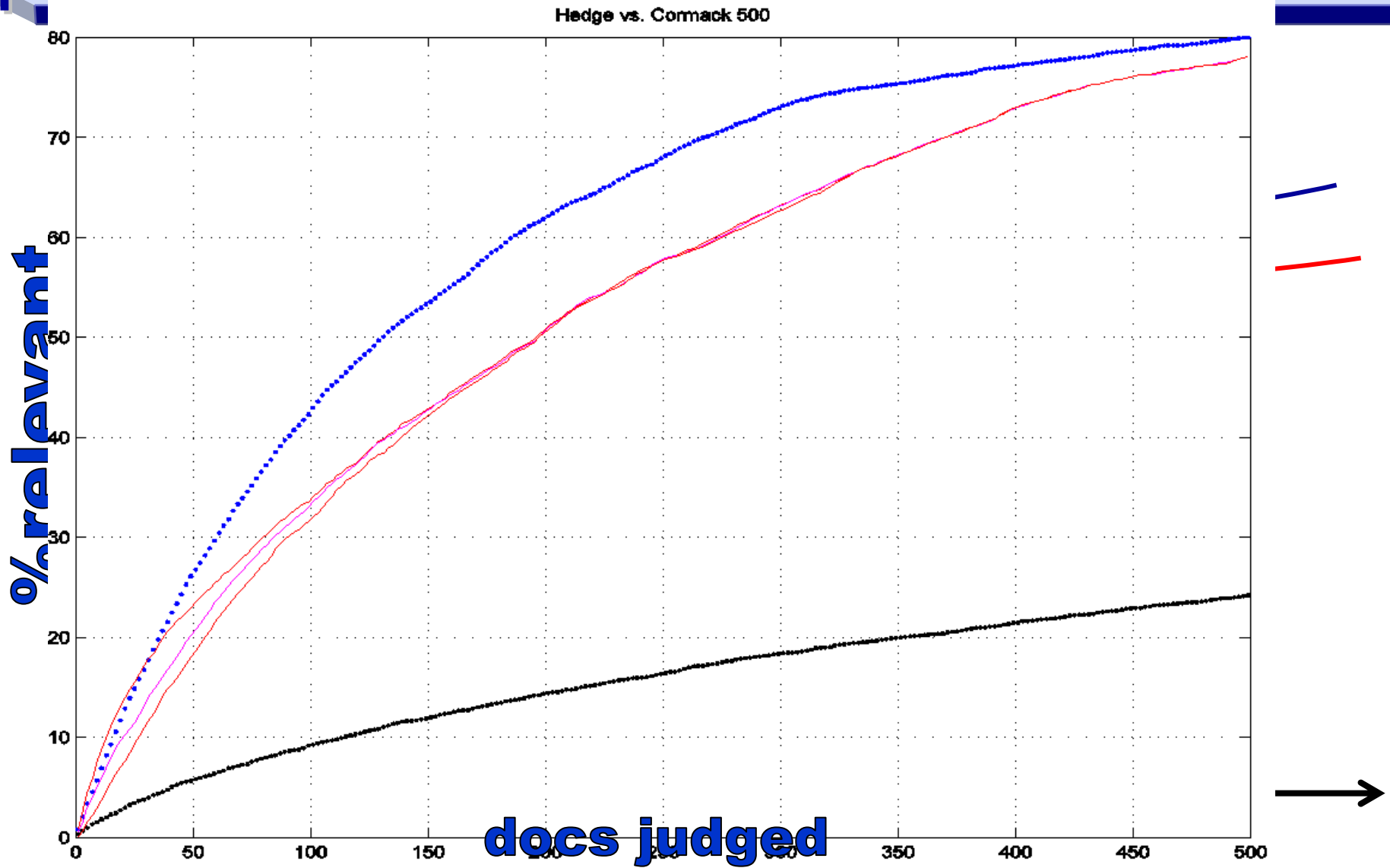




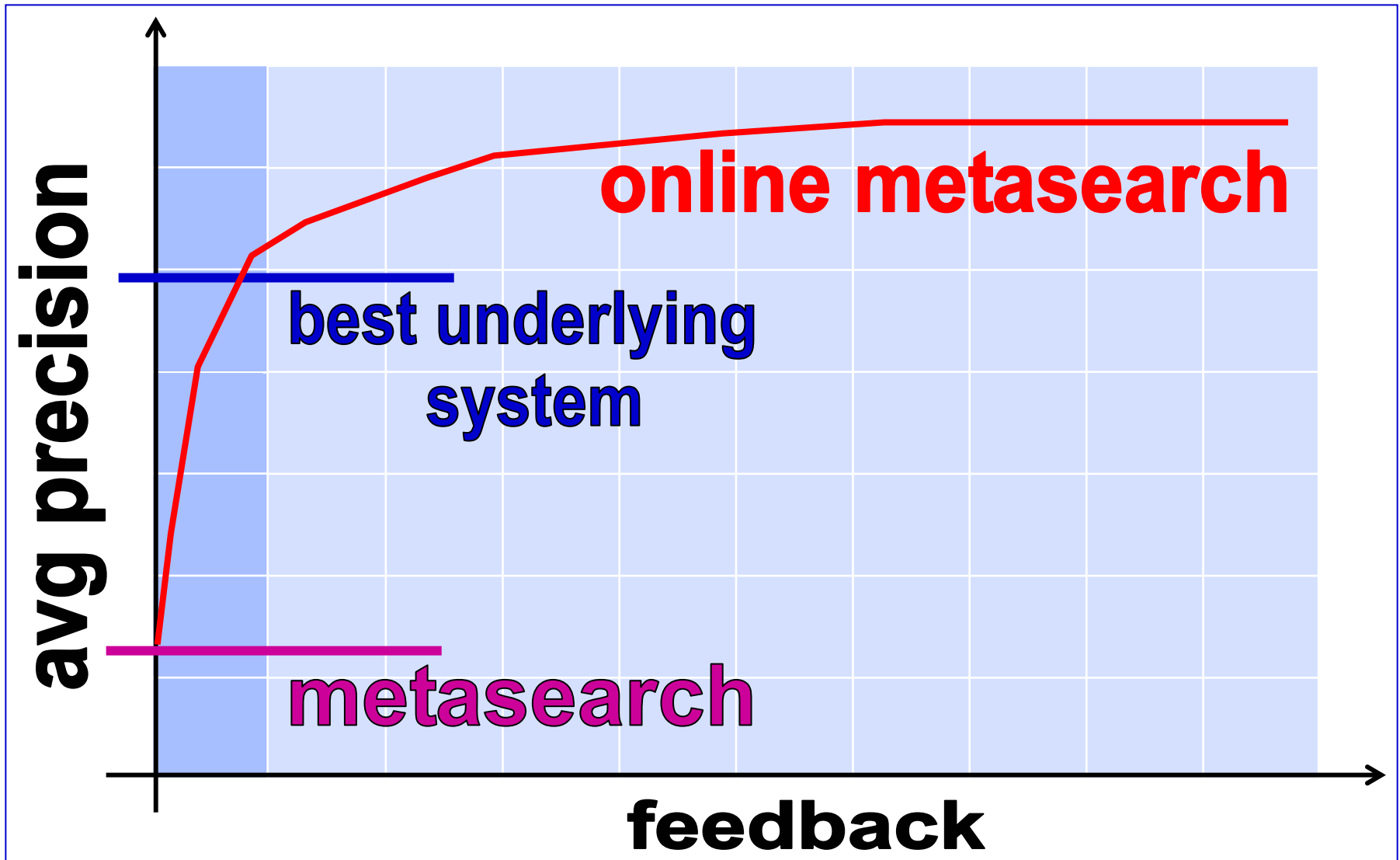
# system evaluation

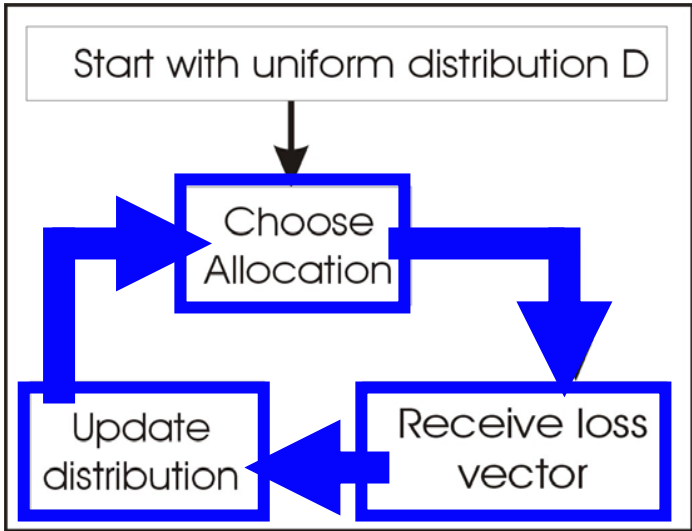


# pooling - comparison with Cormack



# motivation





# pooling - howto

$$\text{- select } d_t = \underset{d \text{ not labeled}}{\text{arg max}} \left[ \sum_{s=1}^N w_s^{t-1} * \text{LOSS}(d, s \mid d = NR) \right]$$

pooling value(d)

$$\text{LOSS}(d, s \mid d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \dots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}$$

- Naturally “want” top ranks
- If NON RELEVANT, then a NR in top ranks of the system lists
- If RELEVANT, bingo.

# metasearch – howto

- Before the next episode

- Compute “pooling value”

for each doc

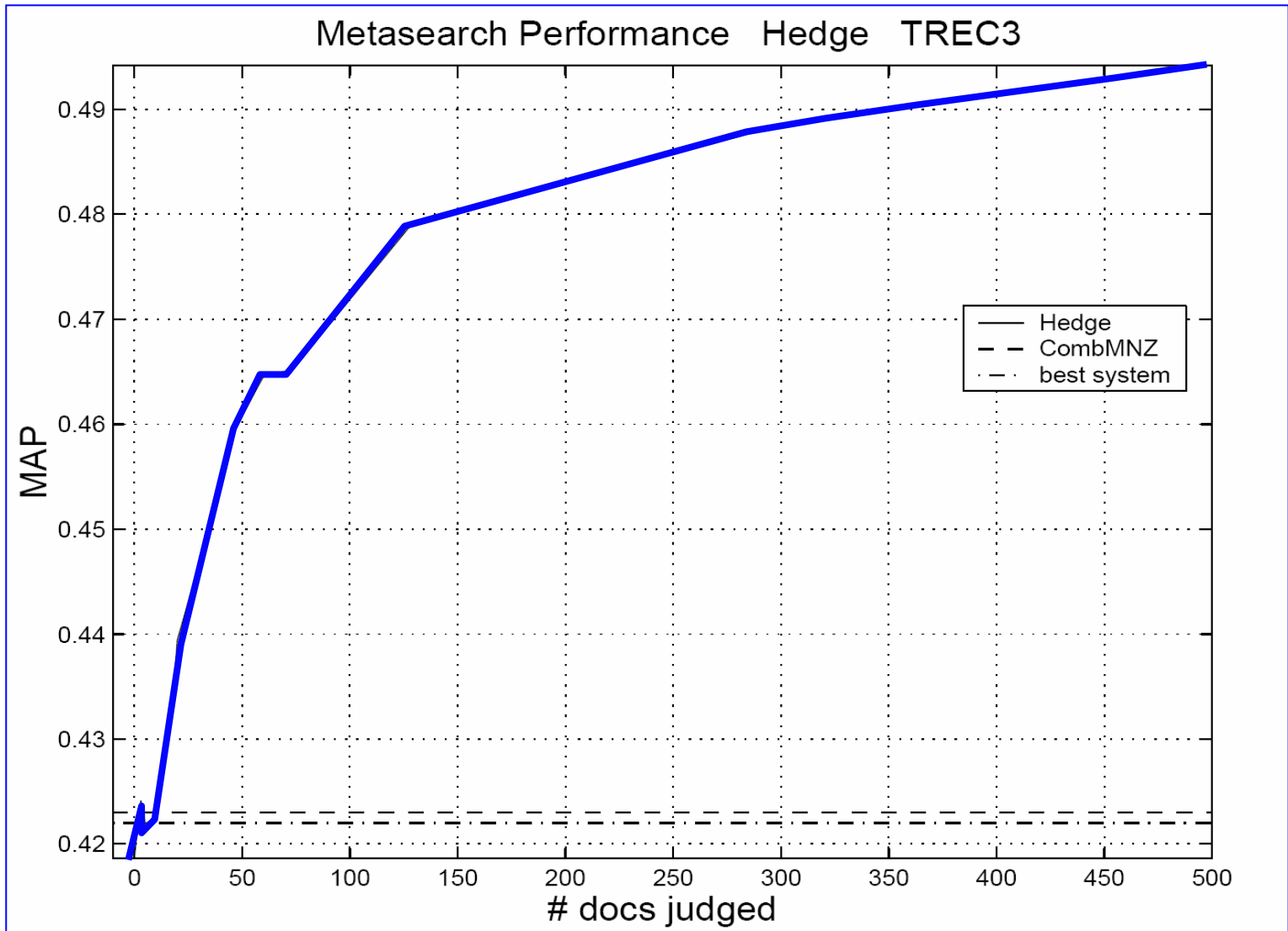
$$\left[ \sum_{s=1}^N w_s^{t-1} * LOSS(d, s | d = NR) \right]$$

- Instead of “select the top doc” for pooling do “select the top 1000 doc” for metasearch
- almost 1000 – docs already pooled are automatically in top of metasearch list

# “total” precision

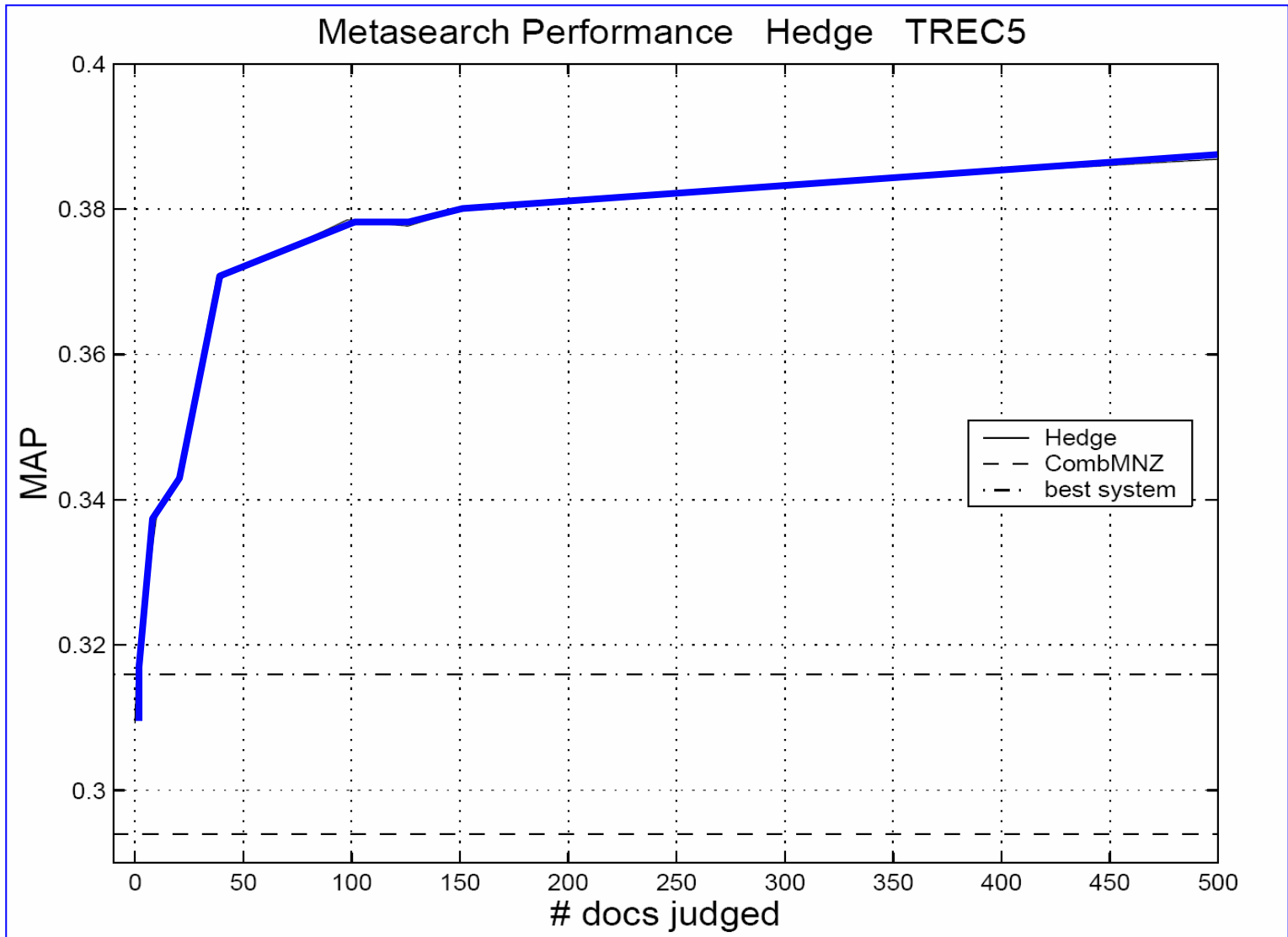
- Average the precision at **ALL** ranks
  - Normalize so ideal system gets  $TP=1$
- math is more simple

# Metas - trec3

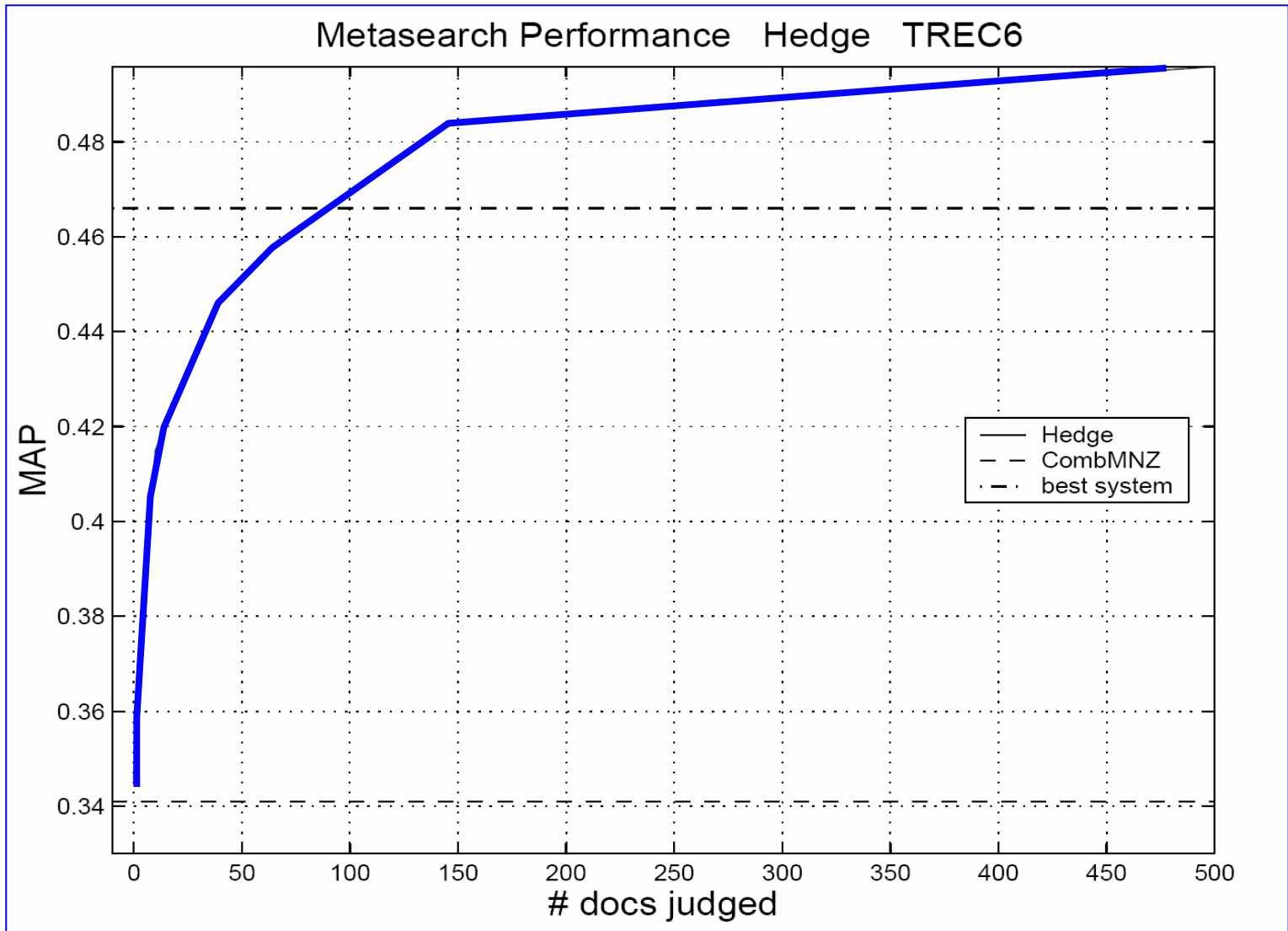




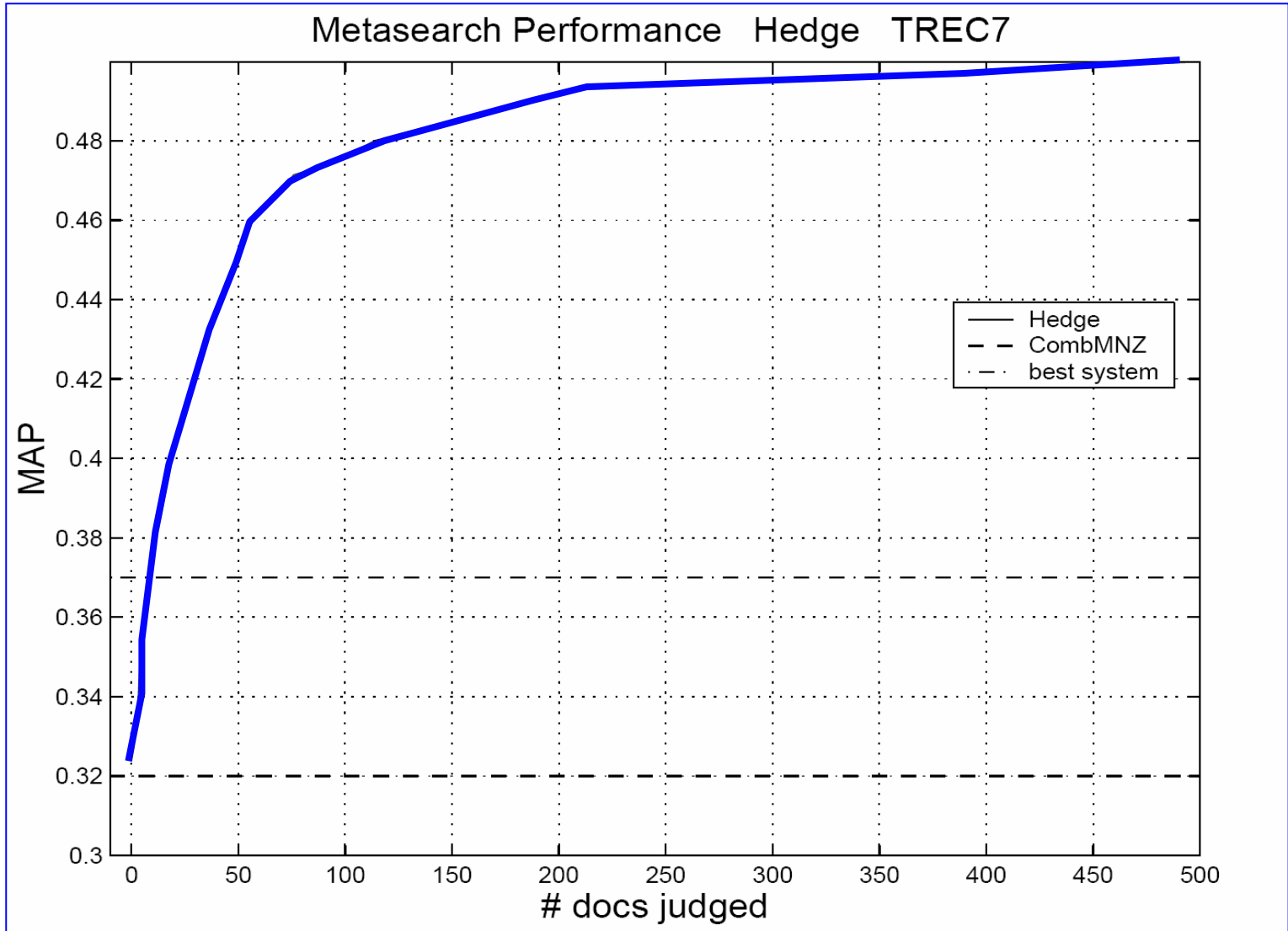
# Metas - trec5



# Metas - trec6



# Metas - trec7



# Metas - trec3

