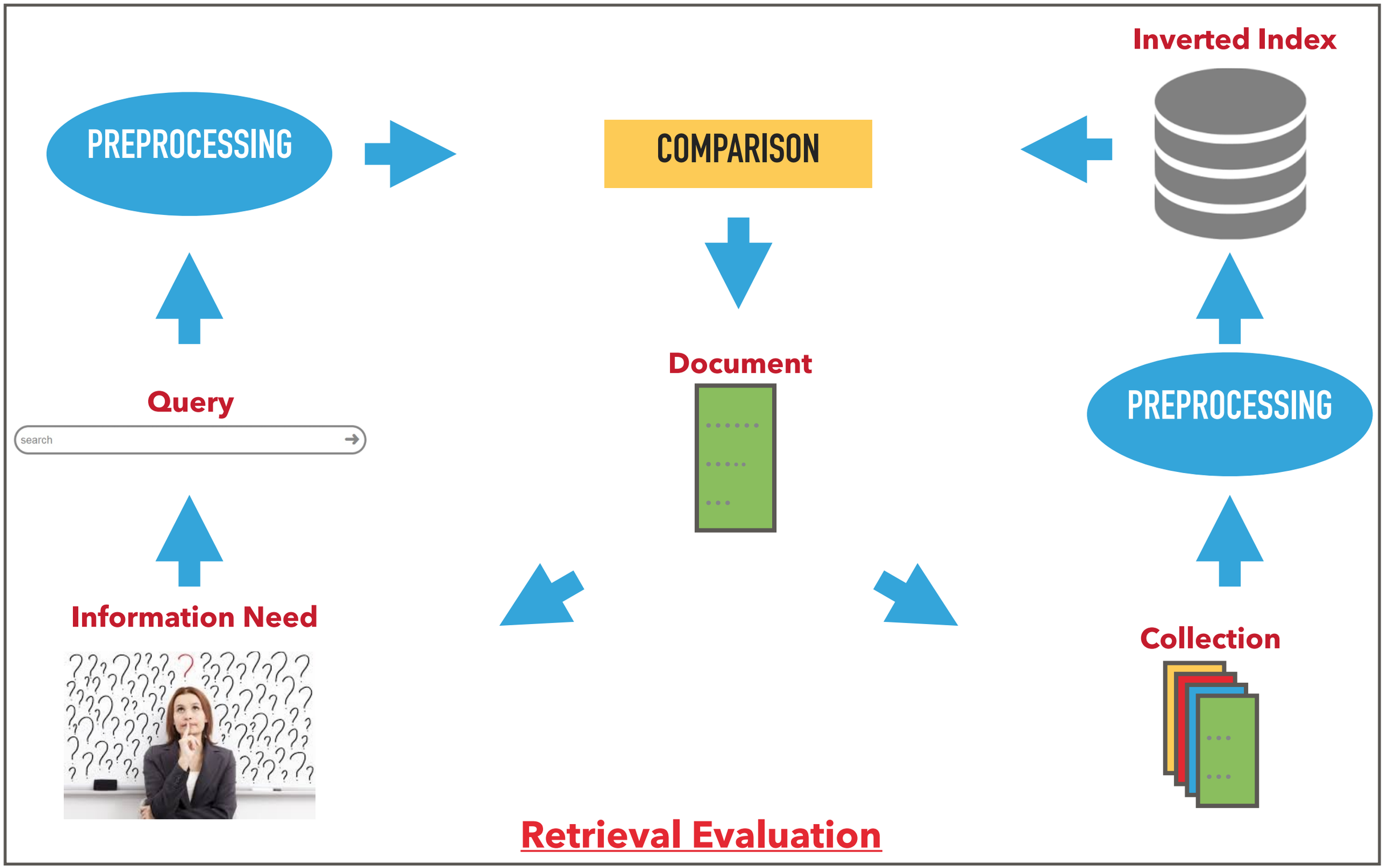
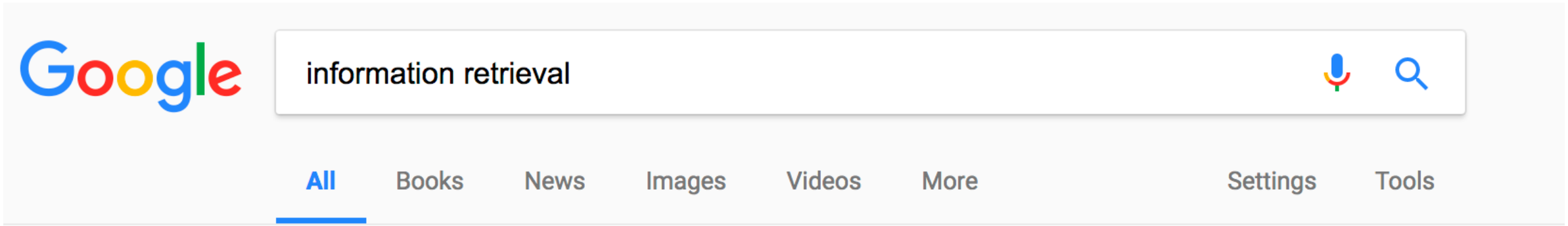
67-300 SEARCH ENGINES

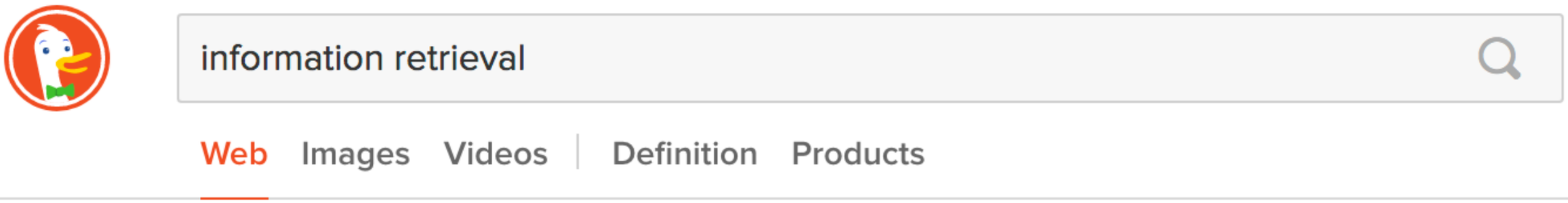
IR EVALUATION

LECTURER: JOAO PALOTTI (JPALOTTI@ANDREW.CMU.EDU)
3RD APRIL 2017

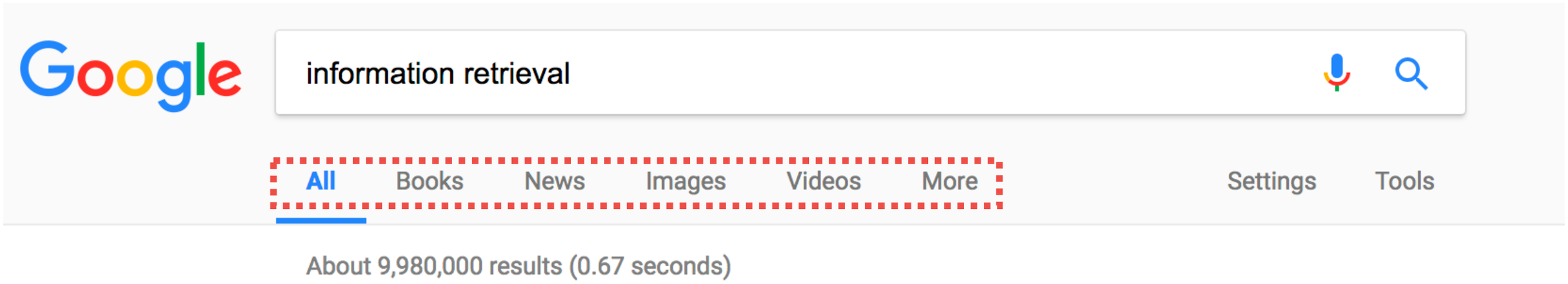

Inverted Index
PREPROCESSING
COMPARISON
Document
PREPROCESSING
Query
search
Information Need
Collection
Retrieval Evaluation

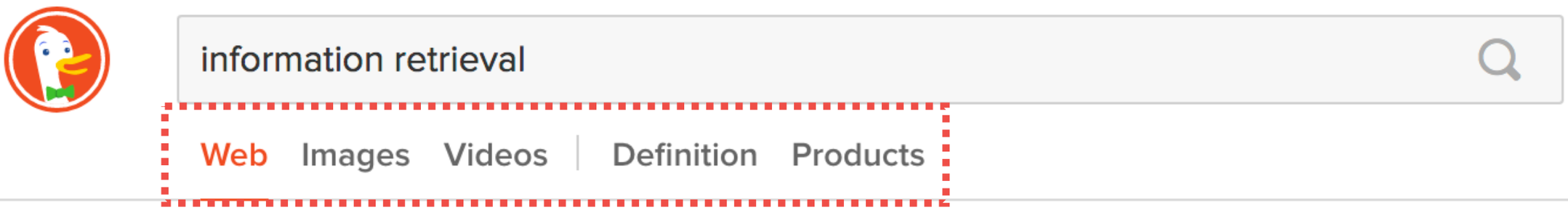
WHICH SEARCH ENGINE DO YOU LIKE THE MOST?



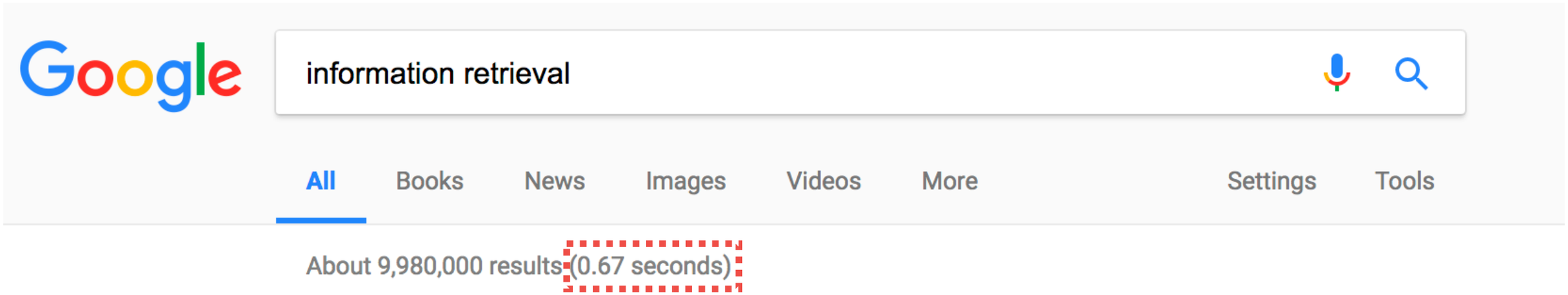


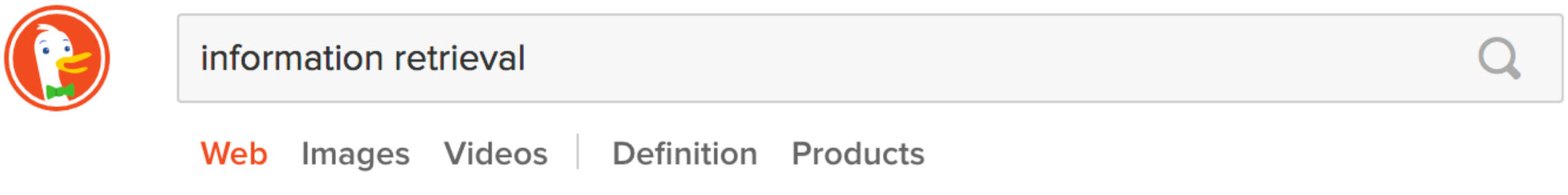
WHICH SEARCH ENGINE DO YOU LIKE THE MOST?



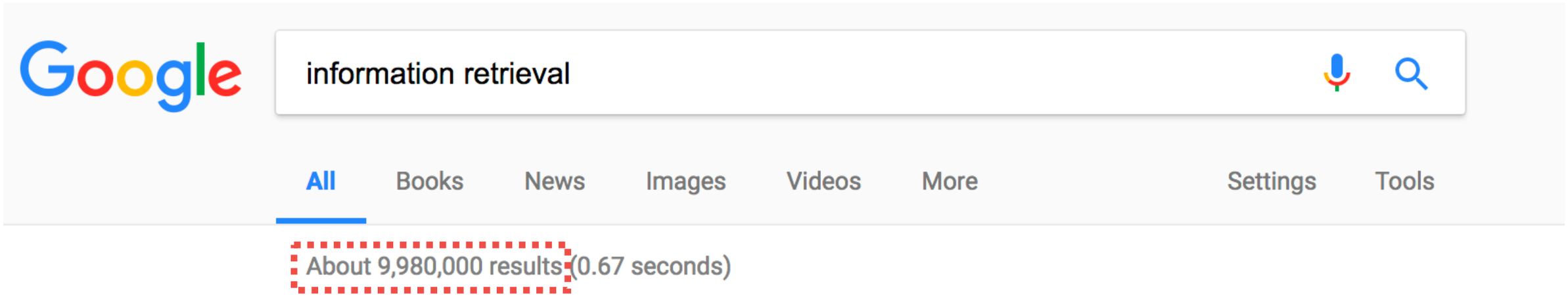


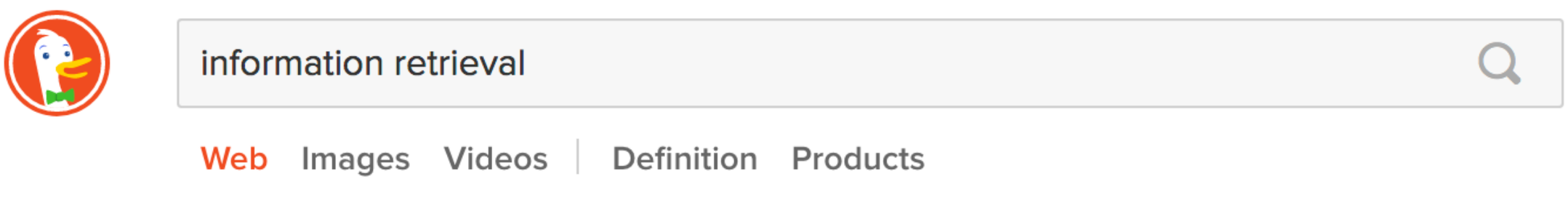
WHICH SEARCH ENGINE DO YOU LIKE THE MOST?



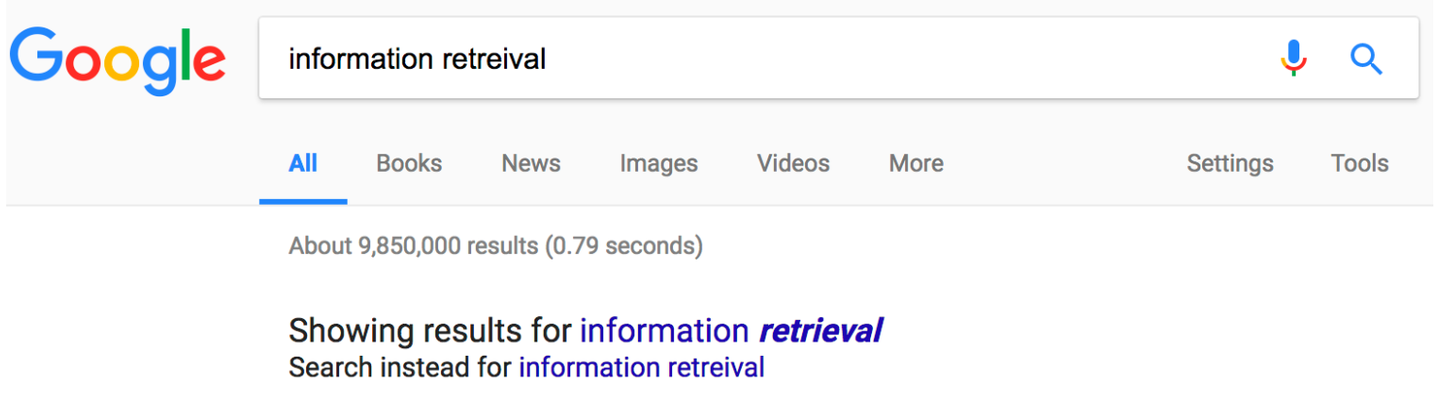


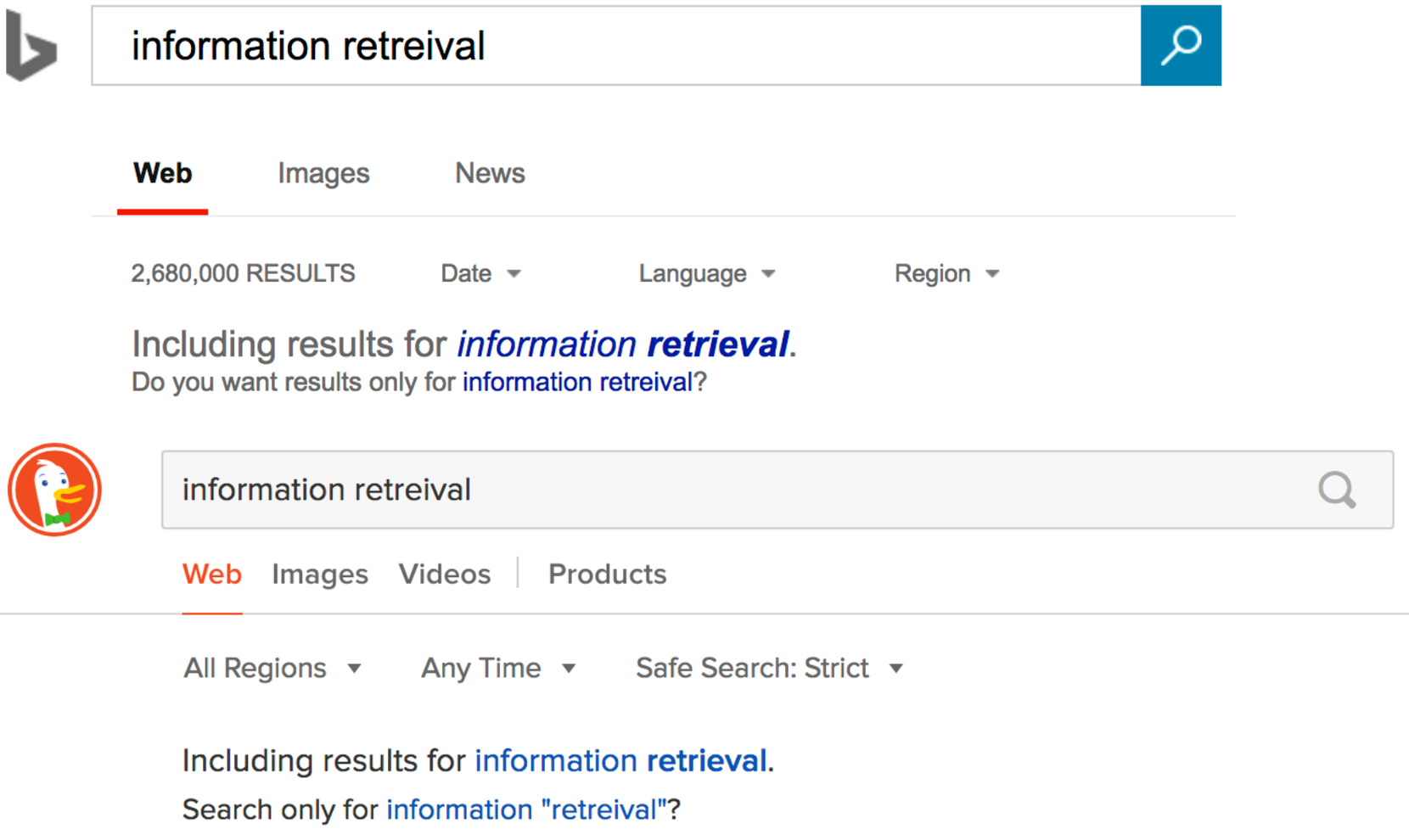
WHICH SEARCH ENGINE DO YOU LIKE THE MOST?



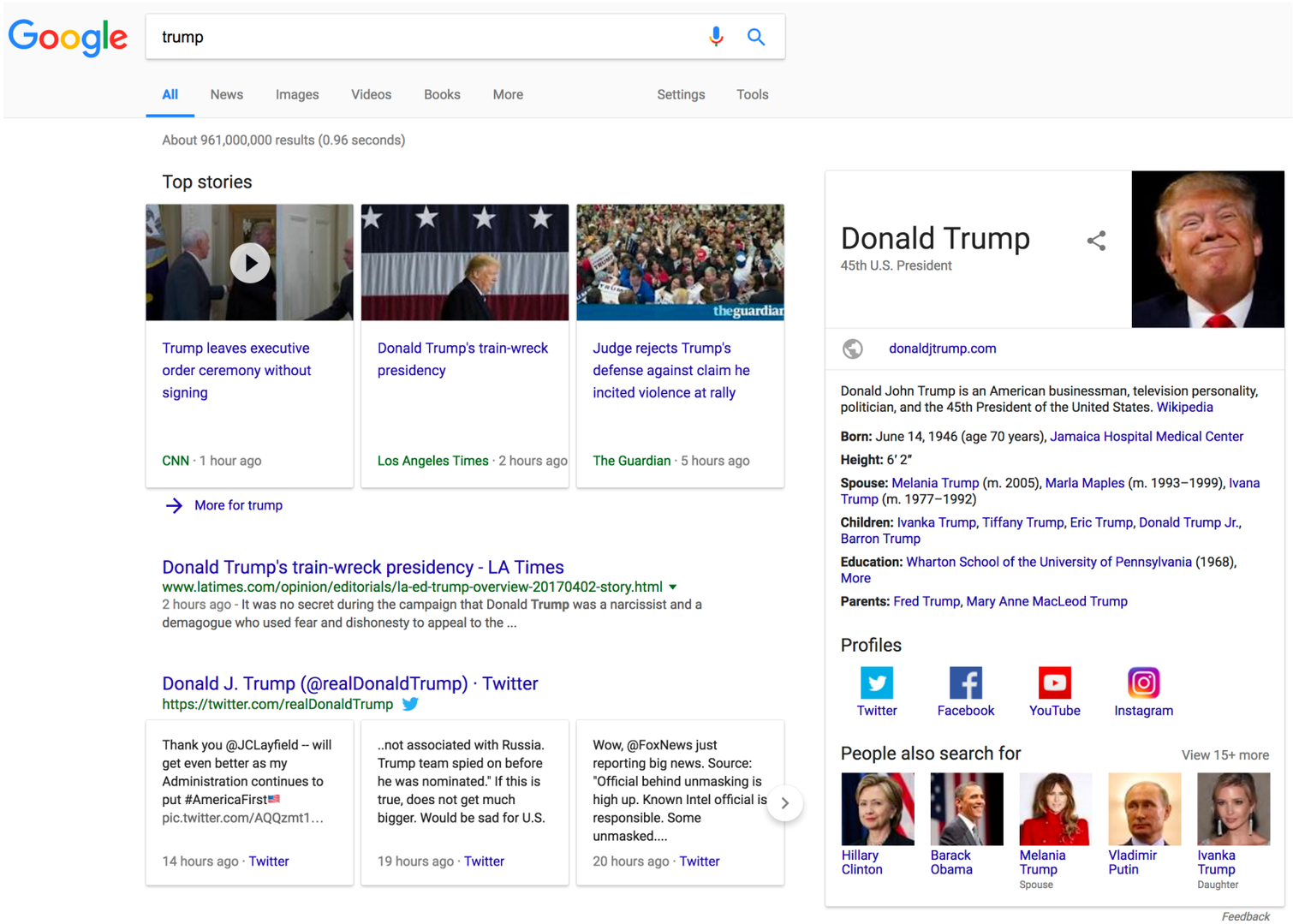


WHICH SEARCH ENGINE DO YOU LIKE THE MOST?

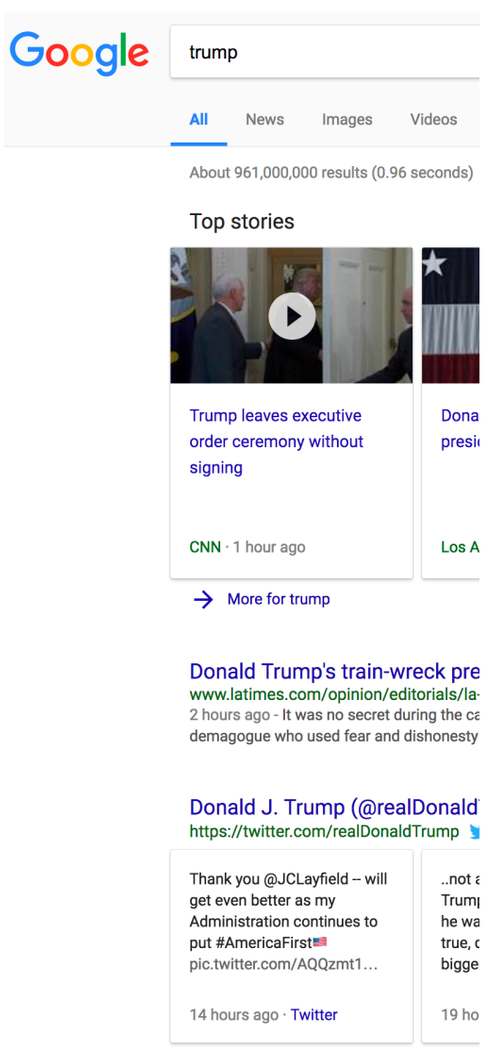


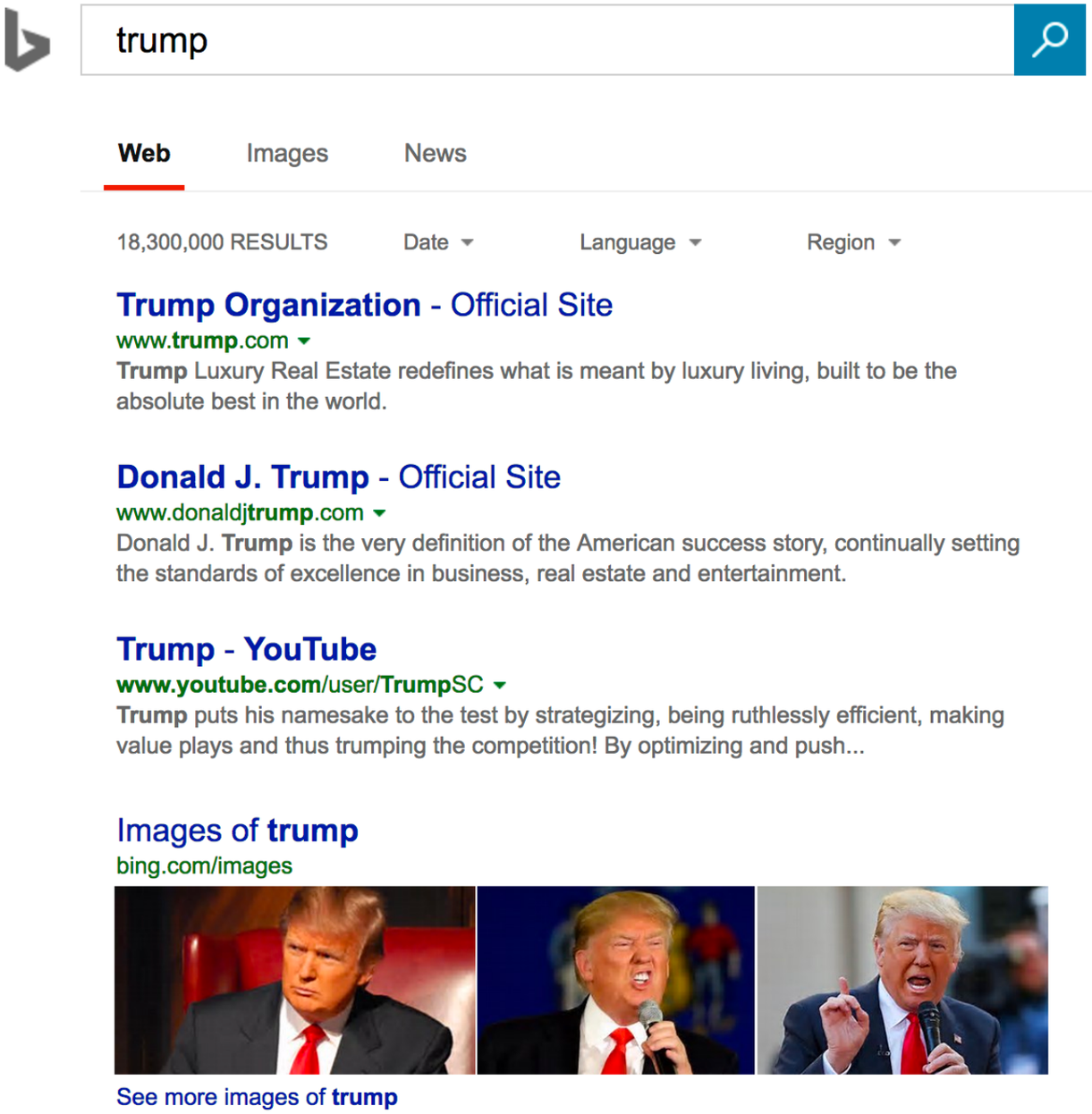


WHICH SEARCH ENGINE DO YOU LIKE THE MOST?

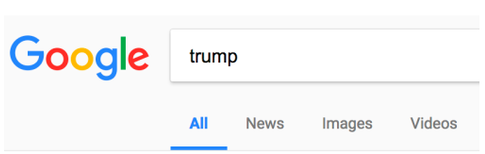


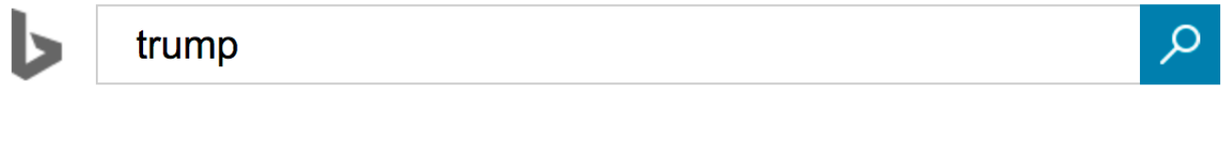
WHICH SEARCH ENGINE DO YOU LIKE THE MOST?

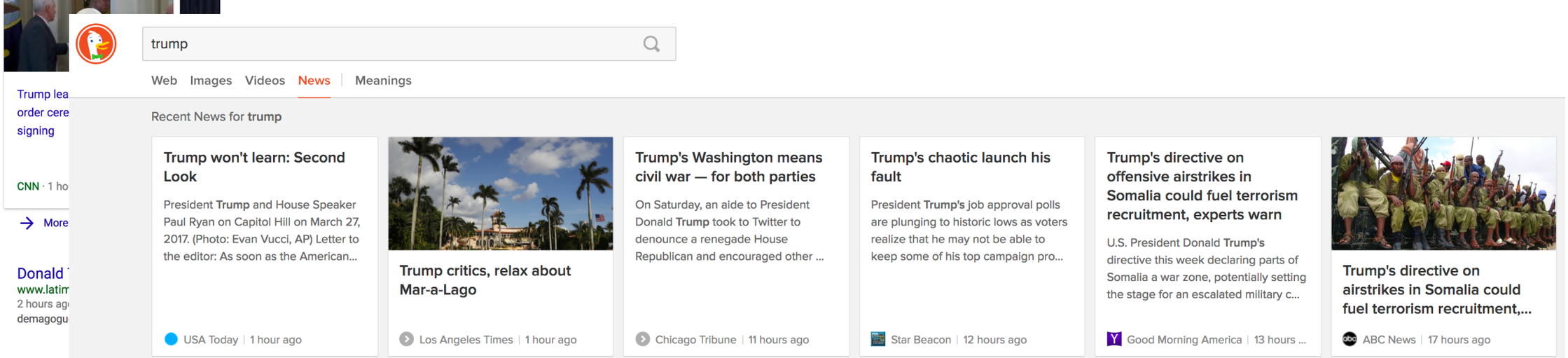




WHICH SEARCH ENGINE DO YOU LIKE THE MOST?







WHICH SEARCH ENGINE DO YOU LIKE THE MOST?

Google

restaurants near me

All Maps News Books More Settings Tools

About 194,000,000 results (0.86 seconds)

Rating Hours

Chef's Garden
4.3 ★★★★½ (51) · Restaurant
950.0 m · Education City, Cycle Path
Quiet · Casual · Locals

Papa Johns
3.9 ★★★★☆ (17) · Pizza
Take-out & delivery pizza chain
400.0 m · Student Center, Education City, Qatar Foundation
Casual · Delivery · Groups

Hardee's هاردیز
3.0 ★★★☆☆ (7) · Fast Food
Fast-food chain for burgers & breakfast
1.8 km
Casual · Delivery · Good for kids

More places

restaurants near me

Web Images News

52,200,000 RESULTS Date Language Region

Best Restaurants Near Me - TripAdvisor
https://www.tripadvisor.com/Restaurants
Find **restaurants near** you from 4 million **restaurants** worldwide with 400 million reviews and opinions from **TripAdvisor** travelers.

Restaurants Near Me - Discover and Enjoy Local **Restaurants**
https://restaurantsnearme.com
Search local **restaurant** listings **near** you that are now open. **Restaurants Near Me** features area **restaurants** within walking distance.

Grubhub - Official Site
https://www.grubhub.com
Free online ordering from **restaurants near** you! With more than 30,000 **restaurants** in

restaurants near me

Web Images Videos

All Regions Any Time Safe Search: Strict

Outback Steakhouse® - Try our New 3-Point Bloom AD
Hurry In And Score Our New 3-Point Bloom. 3 Outback Favorites Together At Last.
www.Outback.com | Order Curbside Take-Away | Nearest Outback Locations

Restaurants Near Me - Discover and Enjoy Local **Restaurants**
Search local **restaurant** listings **near** you that are now open. **Restaurants Near Me** features area **restaurants** within walking distance.
https://restaurantsnearme.com

Food Delivery | **Restaurant** Takeout | Order Food Online | Grubhub
Free online ordering from **restaurants near** you! With more than 30,000 **restaurants** in 500+ cities, food delivery or takeout is just a click away. Because with Grubhub ...
https://grubhub.com

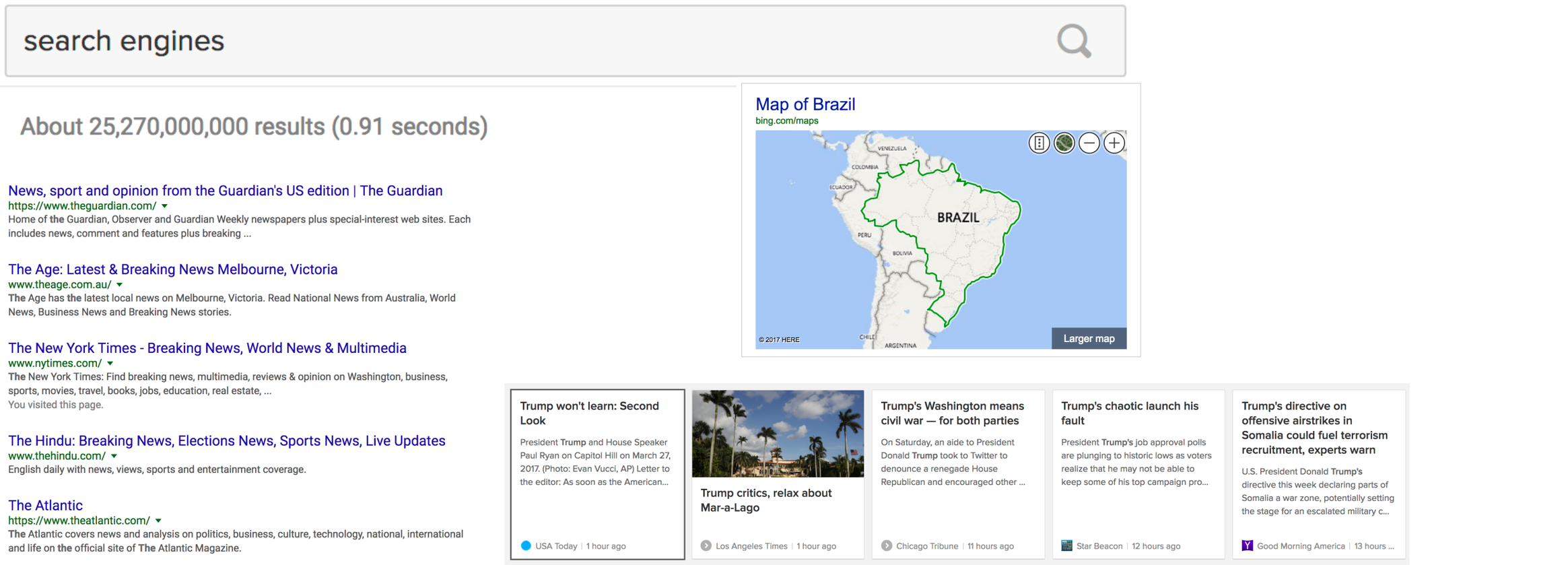
WHICH SEARCH ENGINE DO YOU LIKE THE MOST?

- How to evaluate search engines?
  - Number of results?
  - Time to retrieve results?
  - Spell correction?
  - Query suggestions?
  - Results presentation?
  - Use of user context?
  - Price? Free to use?

ALL?
NONE?

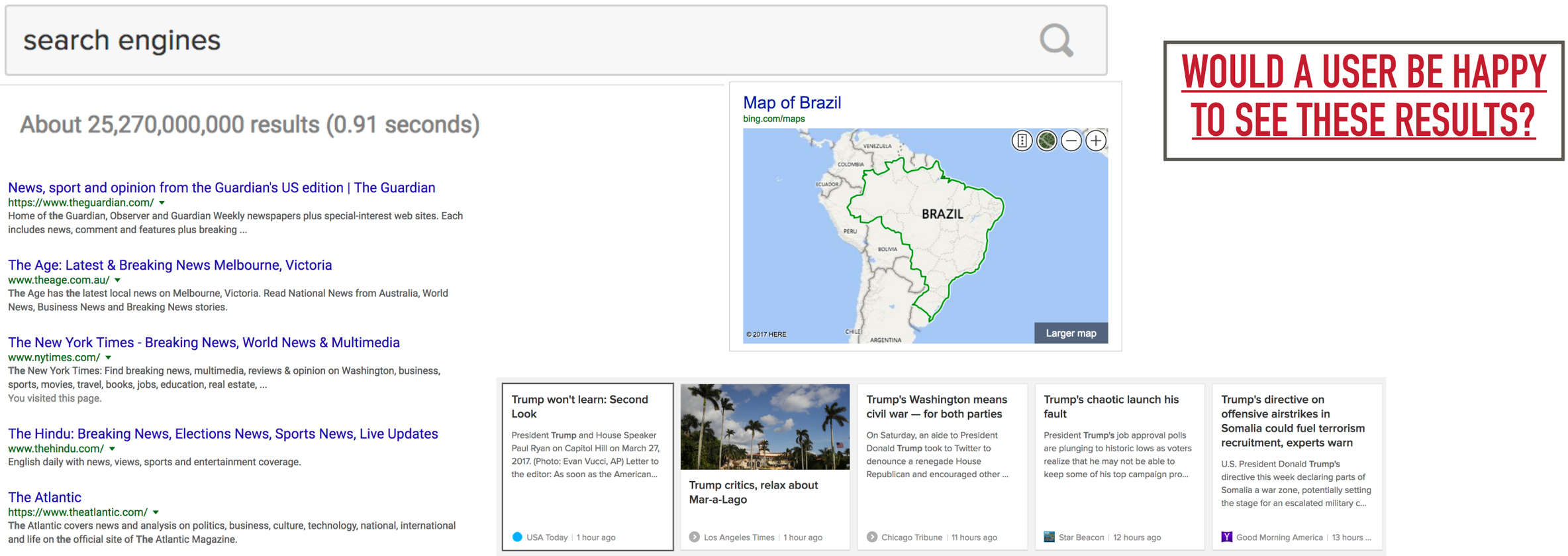
BEST SEARCH ENGINE EVER…

- All metrics cited are important, but is it useful to have a system that instantaneously retrieves a very large number of results and present them in a super fancy way even though all results are **random thrash**?



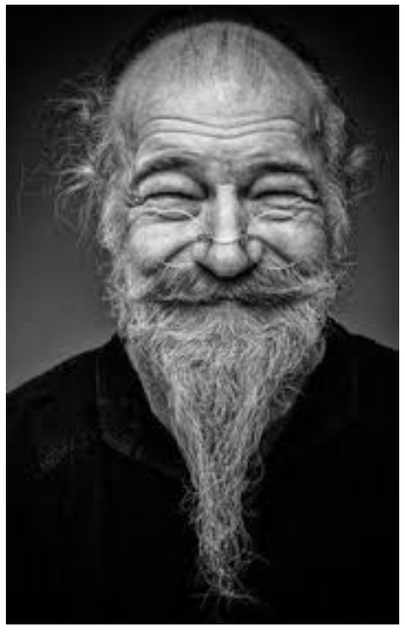
BEST SEARCH ENGINE EVER…

- All metrics cited are important, but is it useful to have a system that instantaneously retrieves a very large number of results and present them in a super fancy way even though all results are **random thrash**?



WOULD A USER BE HAPPY TO SEE THESE RESULTS?

USER HAPPINESS...



USER HAPPINESS...

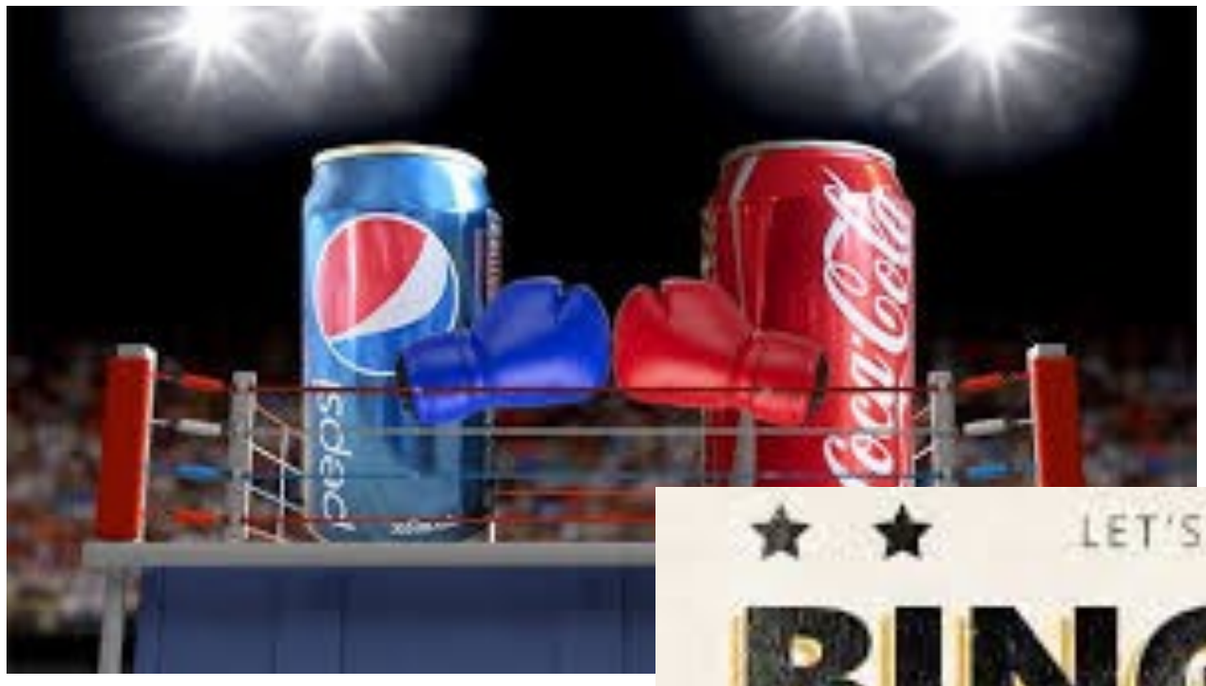
- Hard to define...
- **Simplification:** search engine users are happy when their **information needs** are satisfied:
  - Users are happy when they find what they are looking for...

Vague and boring definition...

- We need to look at the results to decide which search engine provides the results that better satisfy our information needs

USER HAPPINESS...

USER HAPPINESS...



BING IT ON

THE SIDE-BY-SIDE SEARCH OFF

WHICH GAVE YOU THE BEST RESULT FOR:

"internet explorer"

ROUND 3

RESULTS ON LEFT ☐ DRAW ☐ RESULTS ON RIGHT ☐

Internet Explorer - Web Browser for Microsoft Windows
windows.microsoft.com/en-us/**Internet-explorer**/products/ie/home
Download Windows **Internet Explorer** 9 and **Internet Explorer** 8. Watch videos, get help, and learn about the free web browser from Microsoft.

Download | Windows
Windows XP | Pinned Sites
What is Windows 7 | Windows 7
Features | Videos

Internet Explorer downloads - Microsoft Windows
windows.microsoft.com/en-US/**Internet-explorer**/downloads/ie
Internet Explorer 8 · Download
Get downloads for **Internet Explorer**, including **Internet Explorer** 9 and **Internet Explorer** 8.

Internet Explorer - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Internet_Explorer**
History · Features · Architecture · Extensibility · Security
Windows **Internet Explorer** (formerly Microsoft **Internet Explorer**, commonly abbreviated IE or MSIE) is a series of graphical web browsers developed by Microsoft …

Support for Windows **Internet Explorer**
support.microsoft.com/ph/807
Help and support for **Internet Explorer**. Links to customer service and technical solutions, downloads, updates, and answers to top issues.

Microsoft Corporation: Software, Smartphones, Online, Games, …
www.microsoft.com
Download Center · Download
More products. Surface; Windows Phone; Xbox; **Internet Explorer**; Skype; Bing; SkyDrive; Hotmail; PC hardware; MSN

Internet Explorer - CNET Download.com
download.cnet.com/**Internet-Explorer**/3000-2356_4-10013275.html
User rating: 2/5 · 15.35 MB · 19845914 downloads · Download
Update, March 14, 2011: Note that **Internet Explorer** 8 is now for Windows XP only.

Windows **Internet Explorer** 9 - enhanced with Bing and MSN
ie9.discoverbing.com
Windows **Internet Explorer** 9 - enhanced with Bing and MSN **Internet Explorer** 9 is only

Internet Explorer - Web Browser for Microsoft Windows
windows.microsoft.com/en-us/**Internet-explorer**/products/**ie**/home
Download Windows **Internet Explorer** 9 and **Internet Explorer** 8. Watch videos, get help, and learn about the free web browser from Microsoft.

Download Internet Explorer
Get downloads for Internet Explorer, including Internet ...

Internet Explorer 9 + Windows 7
Find out how Internet Explorer 9 and Windows 7 join forces to ...

Discover Internet Explorer 9
Learn about the features of Internet Explorer 9. ... Internet Explorer 9 ...

What is Windows 7?
See what makes Windows 7 faster, simpler, and more fun ...

Internet Explorer 8 system ...
Find out what your computer needs to run Internet Explorer 8.

Getting started with Internet ...
Learn about the new features in Internet Explorer 9.

More results from microsoft.com »

Internet Explorer - Microsoft Download Center
www.microsoft.com/en-us/download/**Ie**.aspx?q=**Internet+explorer**
Results 1 - 10 of 2090 – The **Internet Explorer** 8 Blocker Toolkit enables users to disable automatic delivery of **Internet Explorer** 8 as an important/ high priority class ...

Internet Explorer - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Internet_Explorer**
Internet Explorer (formerly Windows **Internet Explorer**, commonly abbreviated **IE** or MSIE) is a series of graphical web browsers developed by Microsoft and ...

News for **internet explorer**

Chrome and **Internet Explorer** locked in Web browser desktop battle
ZDNet - by Steven Vaughan-Nichols - 1 day ago
Summary: Chrome has moved ahead of **Internet Explorer** recently, but **IE** is making a comeback.

Internet Explorer needs fresh dev infusion for a full recovery
Register - 2 days ago

New **Internet Explorer** Browser | **Internet Explorer**® 9
www.beautyoftheweb.com/
Discover the new **Internet Explorer**! Redesigned to be the fastest, safest browser with innovative features, **Internet Explorer** will amaze with its speed and ...

COOL…

- But not practical…
  - User experiments are hard to conduct in large scale
  - Would require 2 by 2 comparisons with any system variant

BACK IN THE 1960…

Cyril Cleverdon (1914 - 1997)

- Mostly librarians were interested to find information in a library environment (scientific articles, books)

- Decentralized evaluations were hard to conduct:
  - University A says method A is fantastic in their collection
  - University B says method B is fantastic in their collection

BACK IN THE 1960…

Cyril Cleverdon (1914 - 1997)

- Mostly librarians were interested to find information in a library environment (scientific articles, books)

- Decentralized evaluations were hard to conduct:
  - University A says method A is fantastic in their collection
  - University B says method B is fantastic in their collection
  - University A implements method B and it does not work well
  - University B implements method A and it does not work well either…

BACK IN THE 1960...

Cyril Cleverdon (1914 - 1997)

- Cleverdon proposed that everyone would use the same data in an annual meeting in Cranfield University
- With the same dataset we can fairly compare methods from universities A and B
- Born the Cranfield Experiments / Cranfield Paradigm

CANFIELD EXPERIMENTS

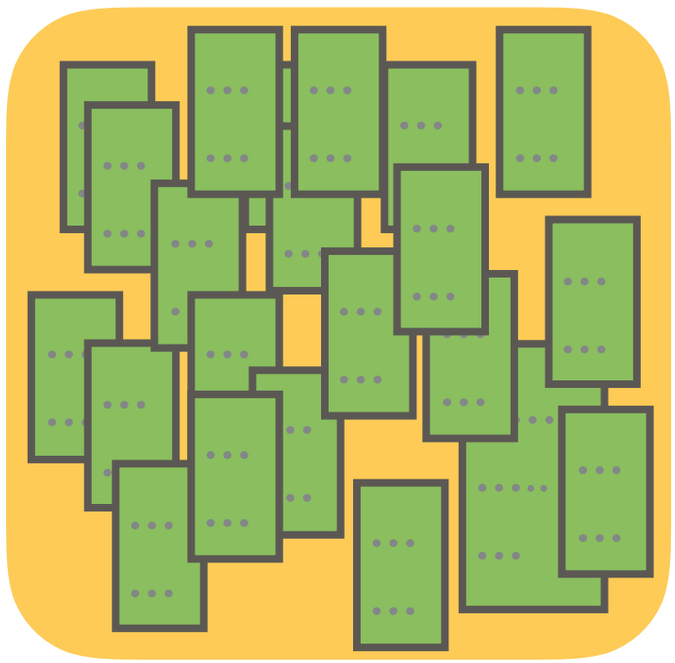
- Key elements:

1. **Collection:** A document collection
2. **Topics:** A test suite of information needs, expressed as queries/topics
3. **Query Relevance Set (QRels):** A set of relevance judgements for each query-document pair

OUR HOMEWORK 2 AND 3 FOLLOW THE SAME PARADIGM

1st Component

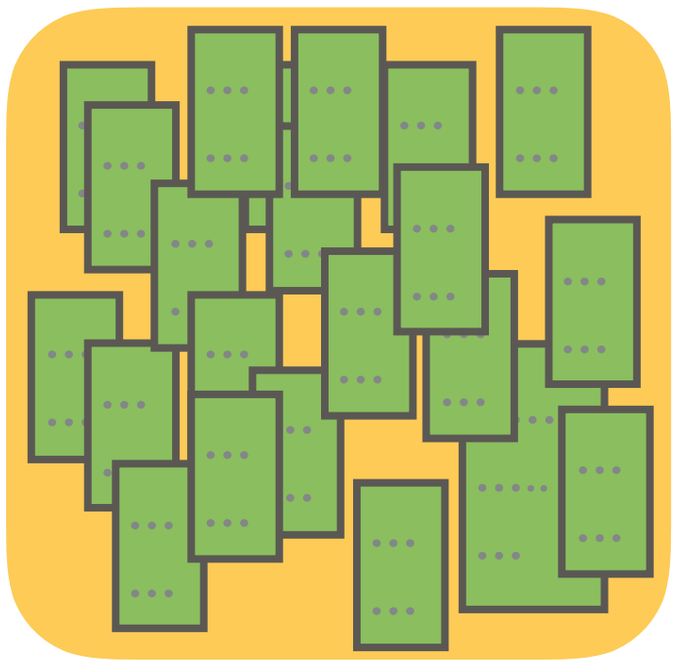
Simple Wikipedia Collection



OUR HOMEWORK 2 AND 3 FOLLOW THE SAME PARADIGM

1st Component

Simple Wikipedia Collection



2nd Component

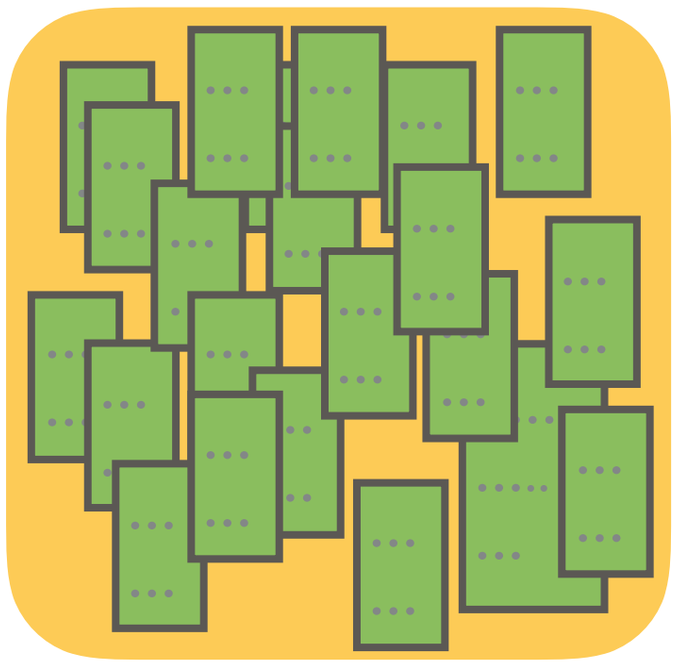
Information needs and queries:

- viking depiction in pop culture
- syrian refugees help
- amazon buying souq
- car breakdown qatar
- post rock bands

OUR HOMEWORK 2 AND 3 FOLLOW THE SAME PARADIGM

1st Component

Simple Wikipedia Collection



2nd Component

Information needs and queries:

- viking depiction in pop culture
- syrian refugees help
- amazon buying souq
- car breakdown qatar
- post rock bands

3rd Component

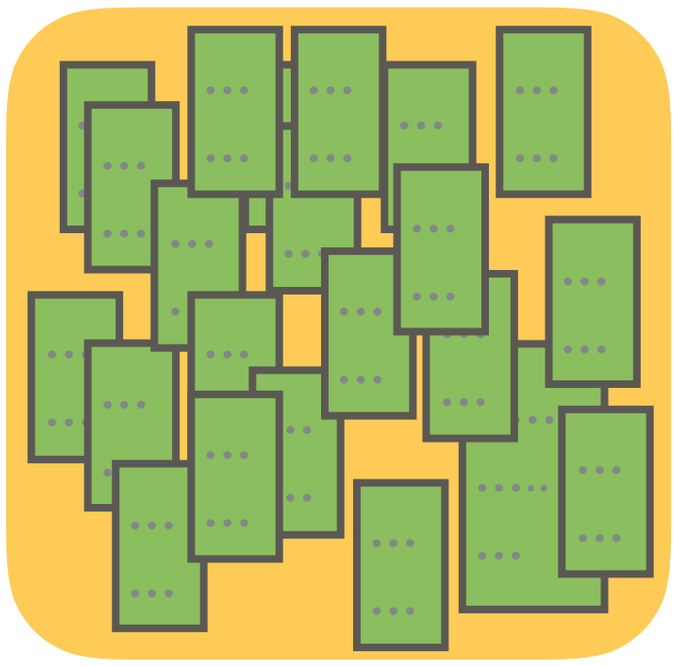
QRels: ground truth / golden standard

- **Query 1 - viking depiction in pop culture:**
  - *simple000001.txt* (Relevant)
  - *simple000002.txt* (Not relevant)
  - *simple000003.txt* (Not relevant)
  - ….-
- **Query 2 - Syrian refugees help:**

OUR HOMEWORK 2 AND 3 FOLLOW THE SAME PARADIGM

1st Component

Simple Wikipedia Collection



2nd Component

Information needs and queries:

- viking depiction in pop culture
- syrian refugees help
- amazon buying souq
- car breakdown qatar
- post rock bands

3rd Component

QRels: ground truth / golden standard

- **Query 1 - viking depiction in pop culture:**
  - *simple000001.txt* (Relevant)
  - *simple000002.txt* (Not relevant)
  - *simple000003.txt* (Not relevant)
  - ….-
- **Query 2 - Syrian refugees help:**

NEXT MONDAY…DO NOT MISS IT!!!

BACK TO CANFIELD EXPERIMENTS…

- Key elements:

    1. **Collection:** 1398 abstracts of aerodynamics journal articles
    2. **Topics:** 225 queries
    3. **Query Relevance Set (QRels):** Exhaustive relevance judgments for all pairs

BACK TO CANFIELD EXPERIMENTS…

- Key elements:

1. **Collection:** 1398 abstracts of aerodynamics journal articles
2. **Topics:** 225 queries
3. **Query Relevance Set (QRels):** Exhaustive relevance judgments for all pairs

Only thing missing: Evaluation metric!

EVALUATION METRICS

- We need numbers that quantifies the quality of a search engine.
- Devising a metric is not a trivial job:
  - metrics should correlate with user perception (many times this is ignored or overlooked)
- We will start metrics for:
  - Unranked retrieval
  - Ranked retrieval

UNRANKED RETRIEVAL EVALUATION

ACCURACY

Confusion Matrix

	Relevant	Non Relevant
Retrieved	true positives (tp)	false positive (fp)
Not Retrieved	false negatives (fn)	true negative (tn)

- Frequently used in Machine Learning
- S.E. classifies each document as either Rel or NRel
- Accuracy - fraction of these classifications that are correct:

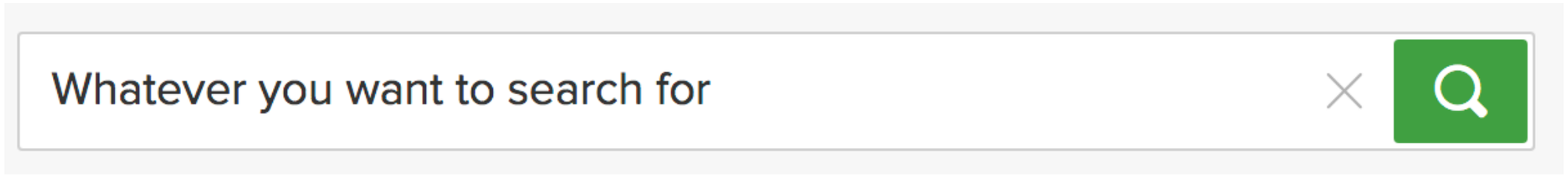
$$(tp + tn) / (tp + fn + fp + tn)$$

Is it a good measure?

ACCURACY

Confusion Matrix

	Relevant	Non Relevant
Retrieved	true positives (tp)	false positive (fp)
Not Retrieved	false negatives (fn)	true negative (tn)



Accuracy of this system is 99.9999999%

We are done! :)

PRECISION AND RECALL

Confusion Matrix

	Relevant	Non Relevant
Retrieved	true positives (tp)	false positive (fp)
Not Retrieved	false negatives (fn)	true negative (tn)

- Precision: Fraction of retrieved docs that are relevant

- Recall: Fraction of relevant docs that are retrieved

Can we deduce their formulas?

PRECISION AND RECALL

Confusion Matrix

	Relevant	Non Relevant
Retrieved	true positives (tp)	false positive (fp)
Not Retrieved	false negatives (fn)	true negative (tn)

- Precision: Fraction of retrieved docs that are relevant

$$(tp) / (tp + fp)$$

- Recall: Fraction of relevant docs that are retrieved

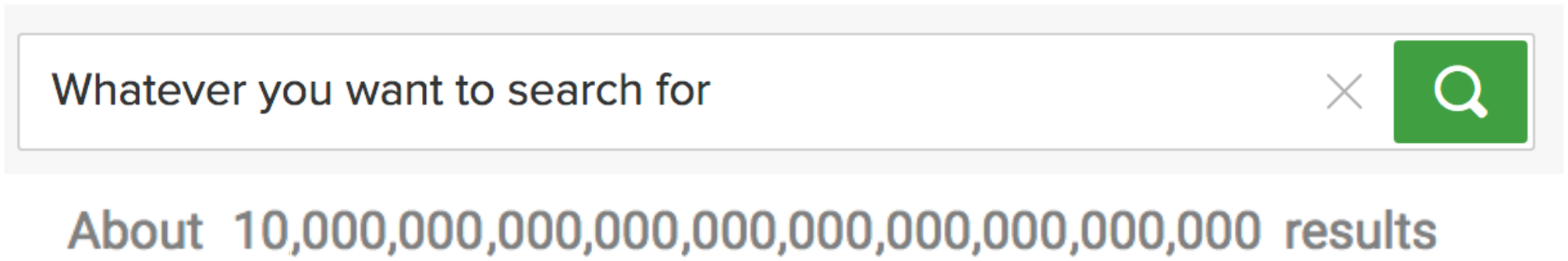
$$(tp) / (tp + fn)$$

PRECISION AND RECALL

- How can we get maximum recall for any query?

PRECISION AND RECALL

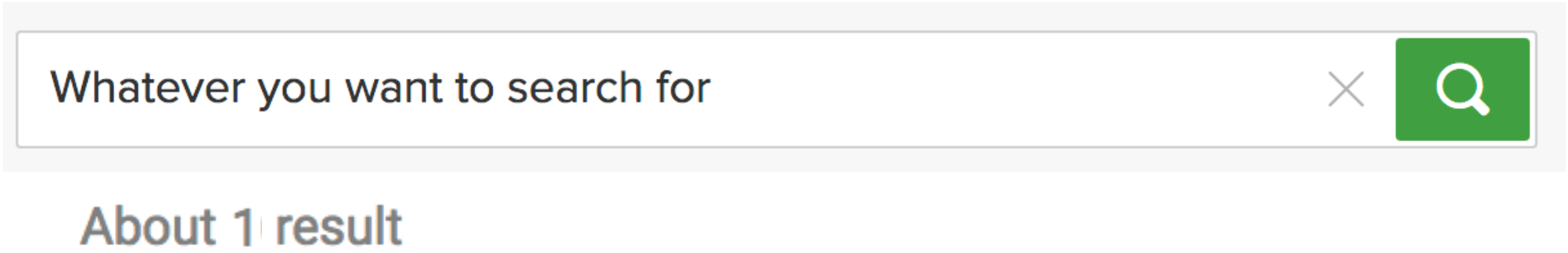
- How can we get maximum recall for any query?



- Recall is a non-decreasing function of the number of docs retrieved

PRECISION AND RECALL

- What is a good strategy to maximize precision?



- Precision often decrease as we retrieve more and more results

PRECISION AND RECALL

- Precision: prefers systems retrieving fewer documents, but highly relevant
- Recall: prefers systems retrieving more documents

HOW TO BALANCE THESE TWO METRICS?

PRECISION AND RECALL

- Precision: prefers systems retrieving fewer documents, but highly relevant
- Recall: prefers systems retrieving more documents

HOW TO BALANCE THESE TWO METRICS?

ARITHMETIC MEAN?

F-SCORE

- F-score combines precision and recall into a single value
- More robust than the arithmetic mean:
  - Harmonic mean!
  - Pessimistic metric – tends to be closer to the minimal value

F-SCORE

- F-score combines precision and recall into a single value
- More robust than the arithmetic mean:
- Harmonic mean!
- Pessimistic metric – tends to be closer to the minimal value

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad == \quad F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

F-SCORE

- F-score combines precision and recall into a single value
- More robust than the arithmetic mean:
- Harmonic mean!
- Pessimistic metric – tends to be closer to the minimal value

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad == \quad F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F-SCORE EXAMPLES

- Precision = 0.60, Recall = 0.70
  - Arithmetic mean: 0.65
  - F1: 0.60
- Precision = 0.10, Recall = 0.90
  - Arithmetic mean: 0.50
  - F1: 0.20

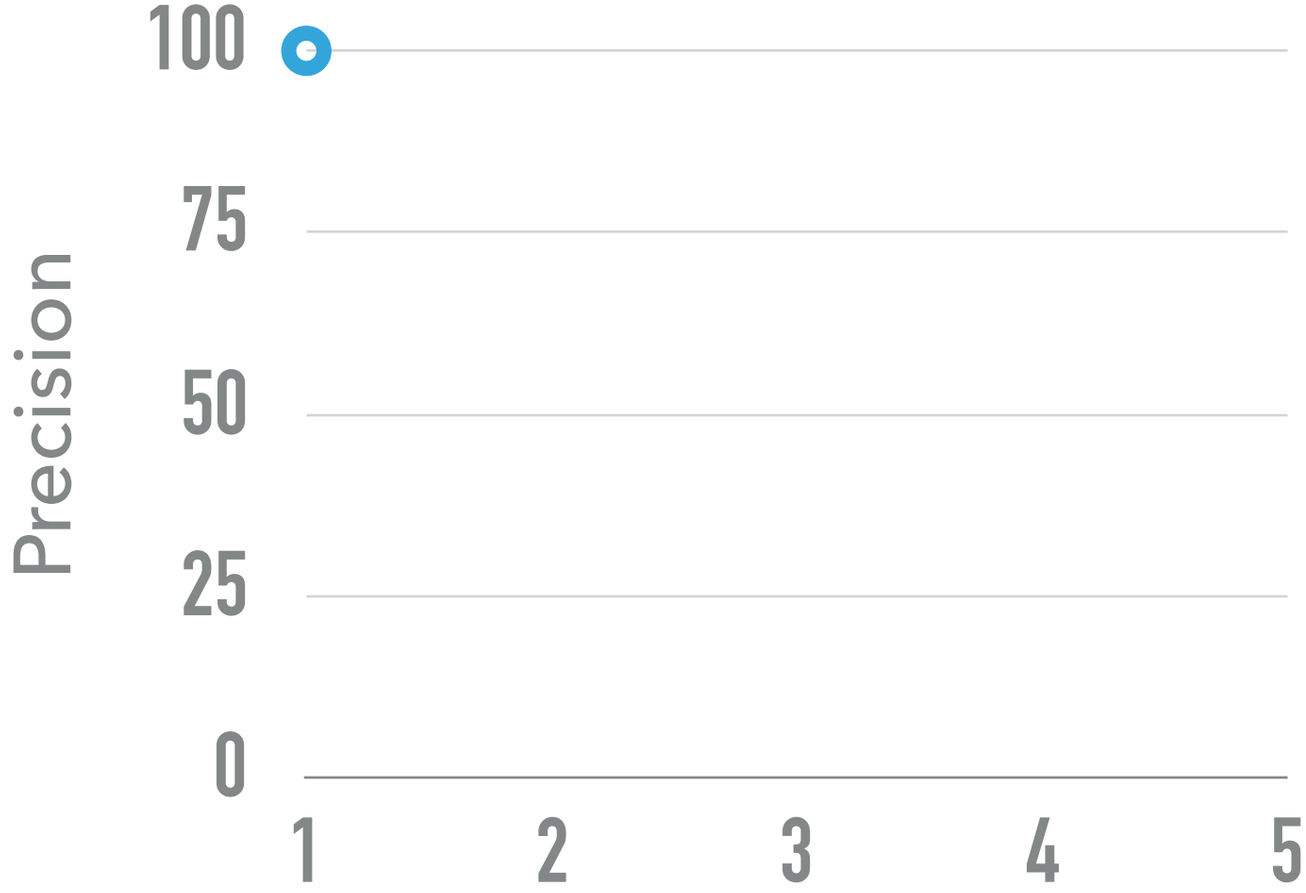
RANKED RETRIEVAL EVALUATION

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- Idea: plot precision as documents are retrieved

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- Idea: plot precision as documents are retrieved

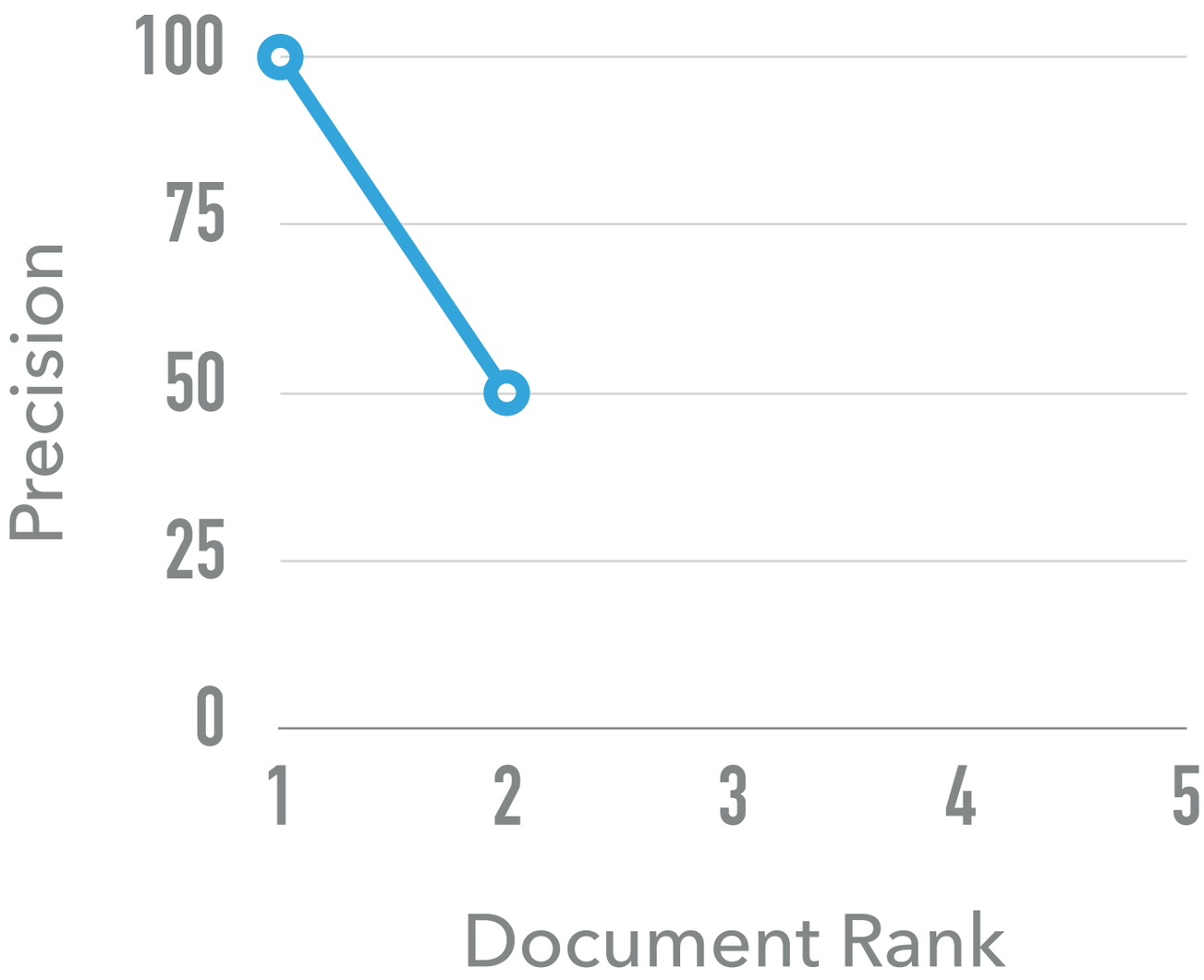


Document 1 is relevant!

Precision Now = 1./1. = 1

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- Idea: plot precision as documents are retrieved

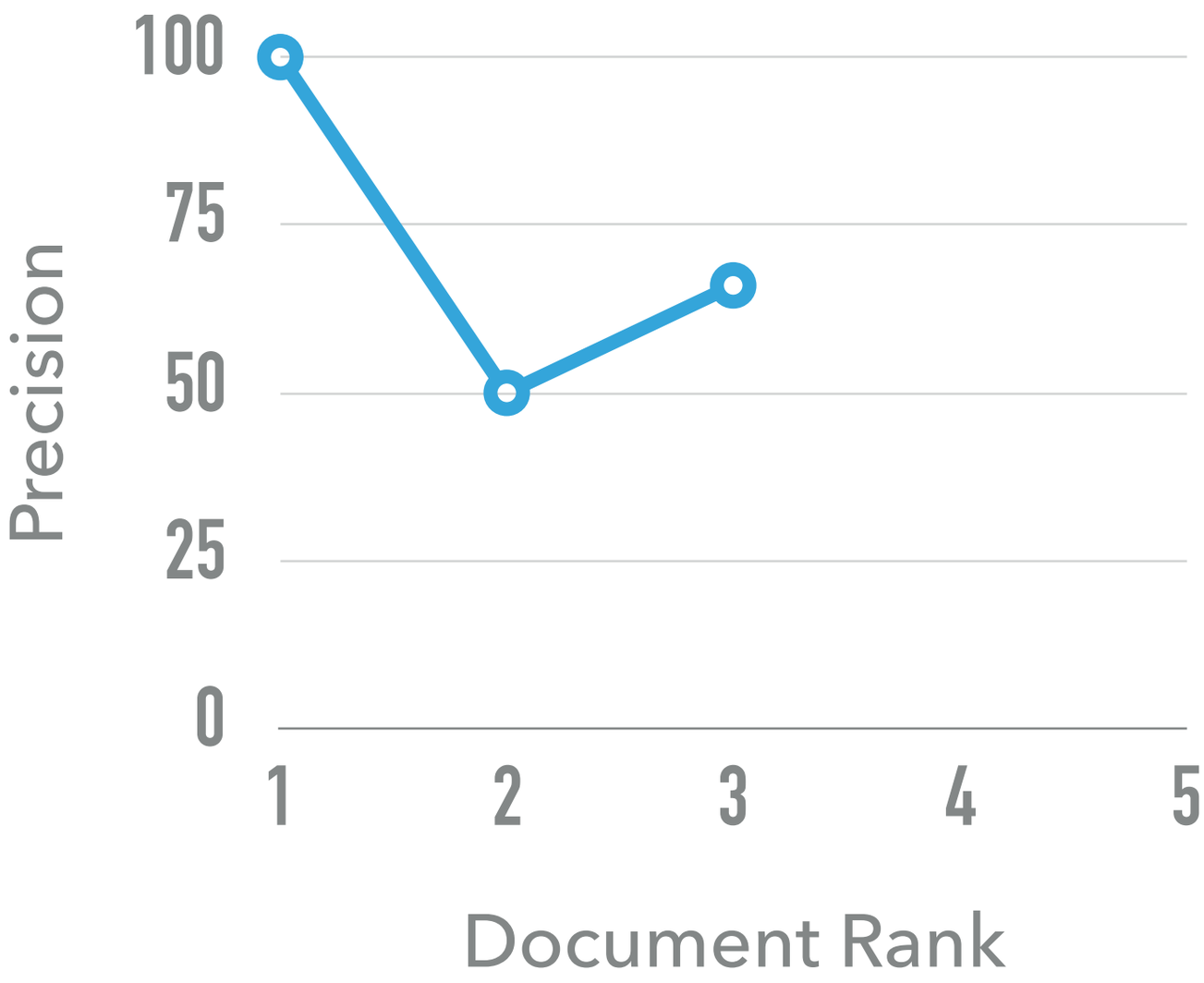


Document 2 is not relevant!

Precision Now = 1./2. = 0.5

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- Idea: plot precision as documents are retrieved

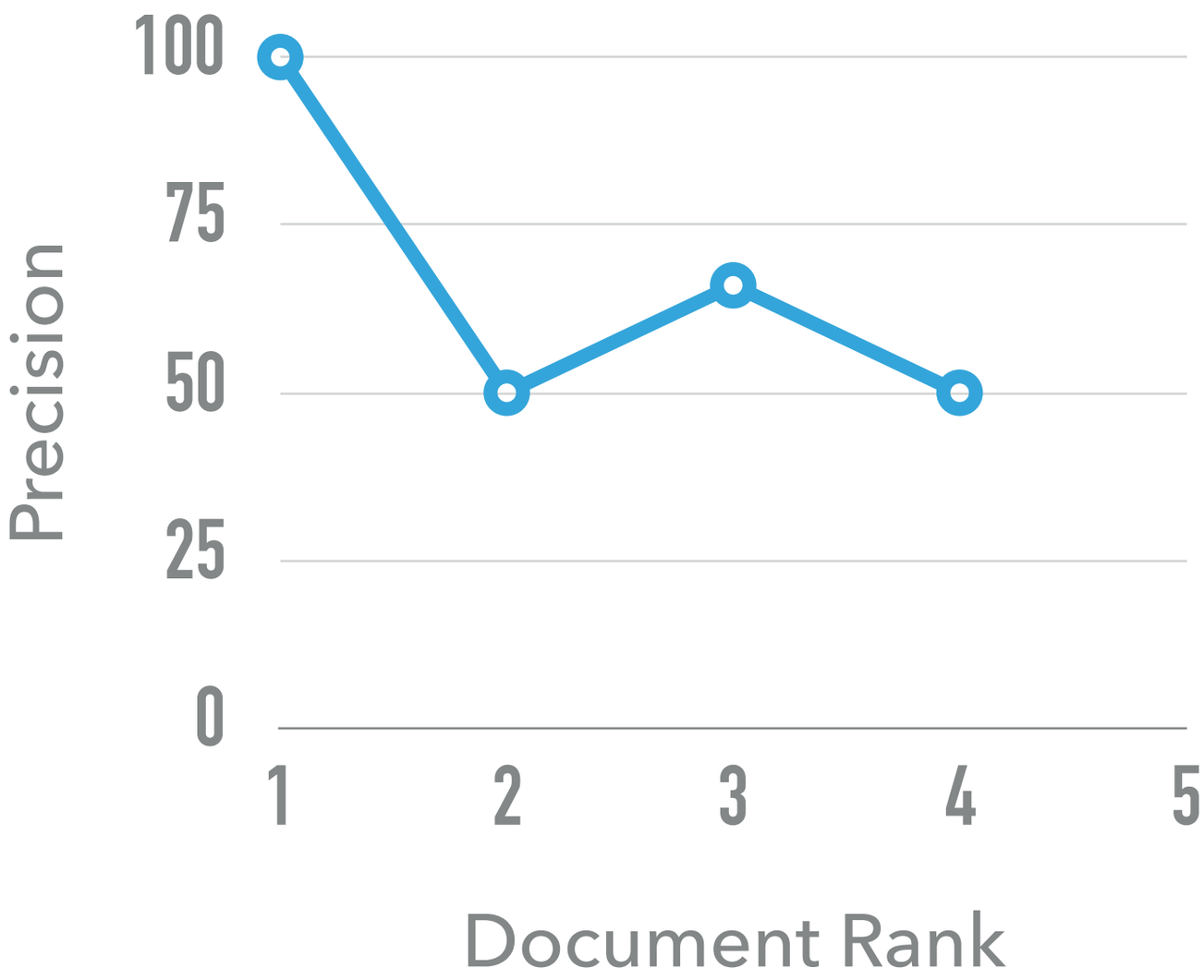


Document 3 is relevant!

Precision Now = 2./3. = 0.66

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- Idea: plot precision as documents are retrieved

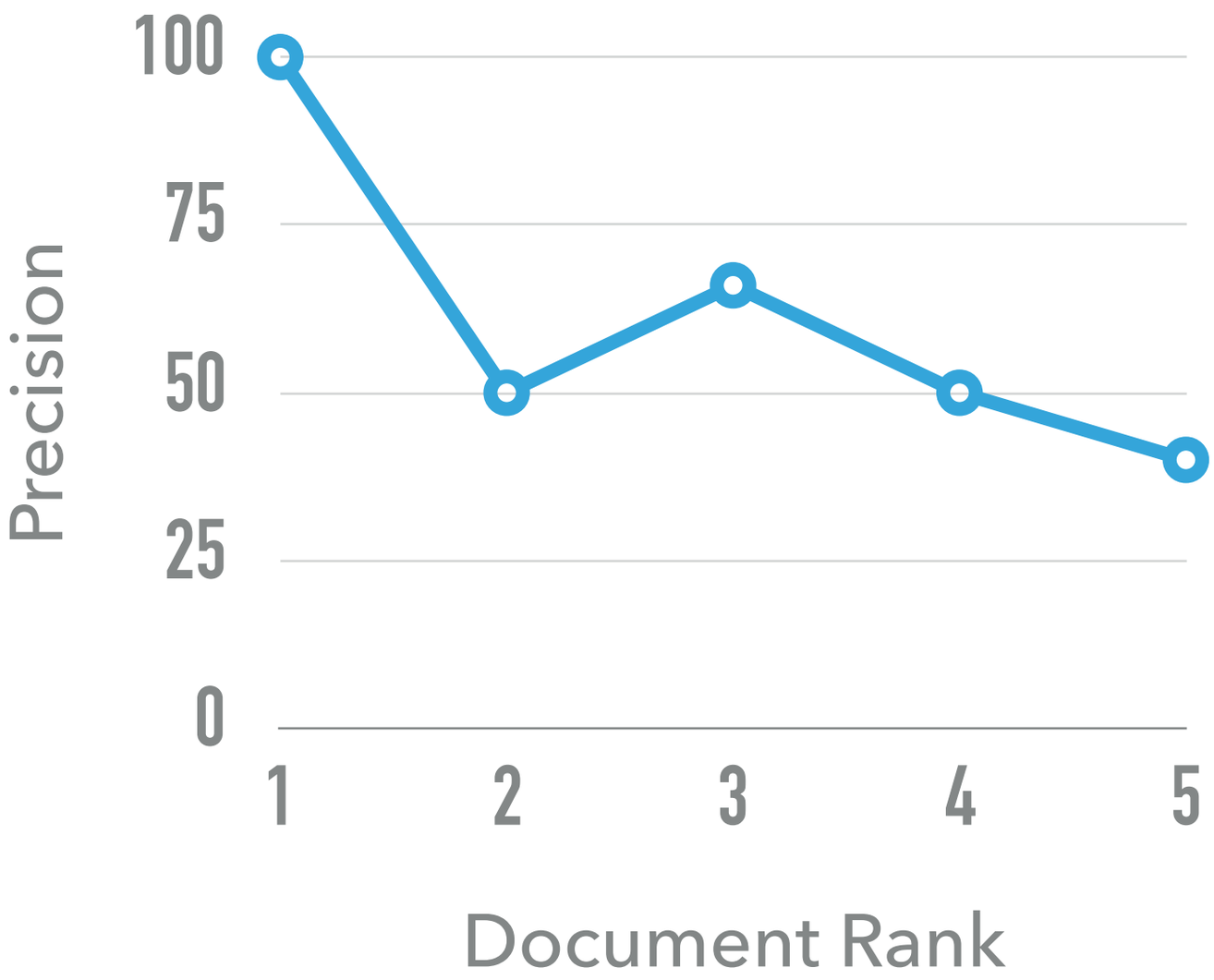


Document 4 is not relevant!

Precision Now = 2./4. = 0.5

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- Idea: plot precision as documents are retrieved



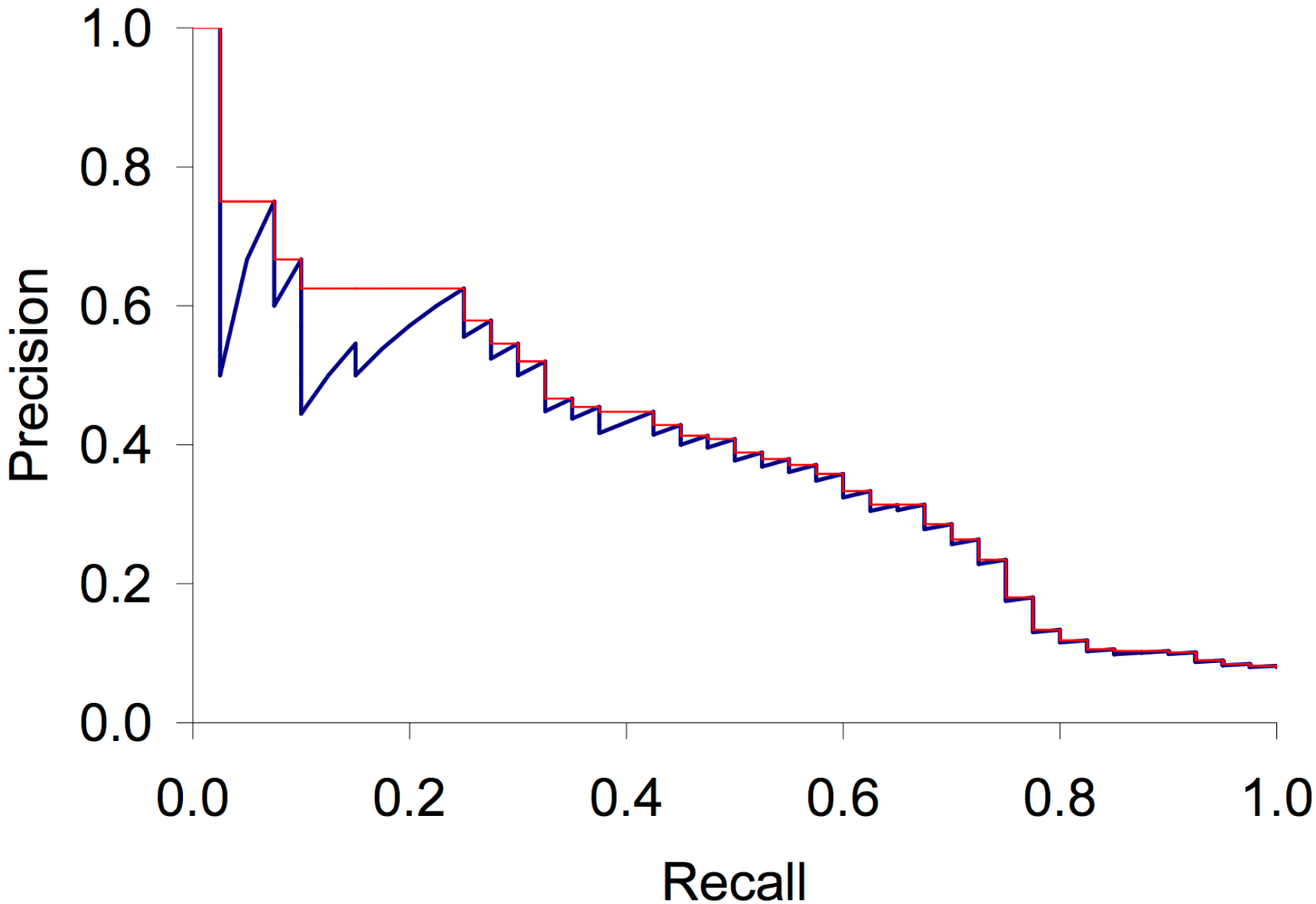
Document 5 is not relevant!

Precision Now = 2./5. = 0.4

Limitation: Recall is **not** considered in this graph!!!

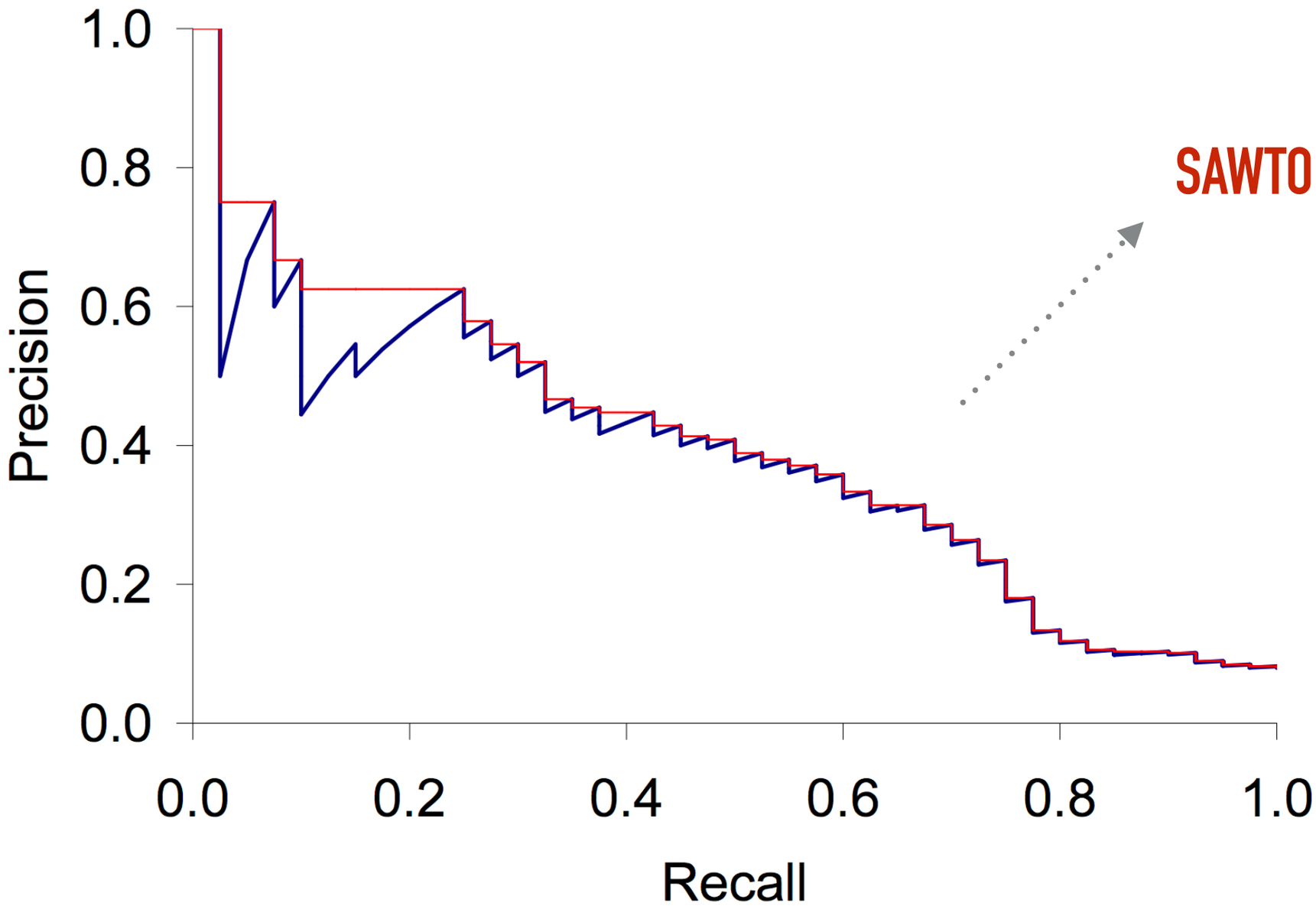
COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- We can do the same at different levels of recall:



COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

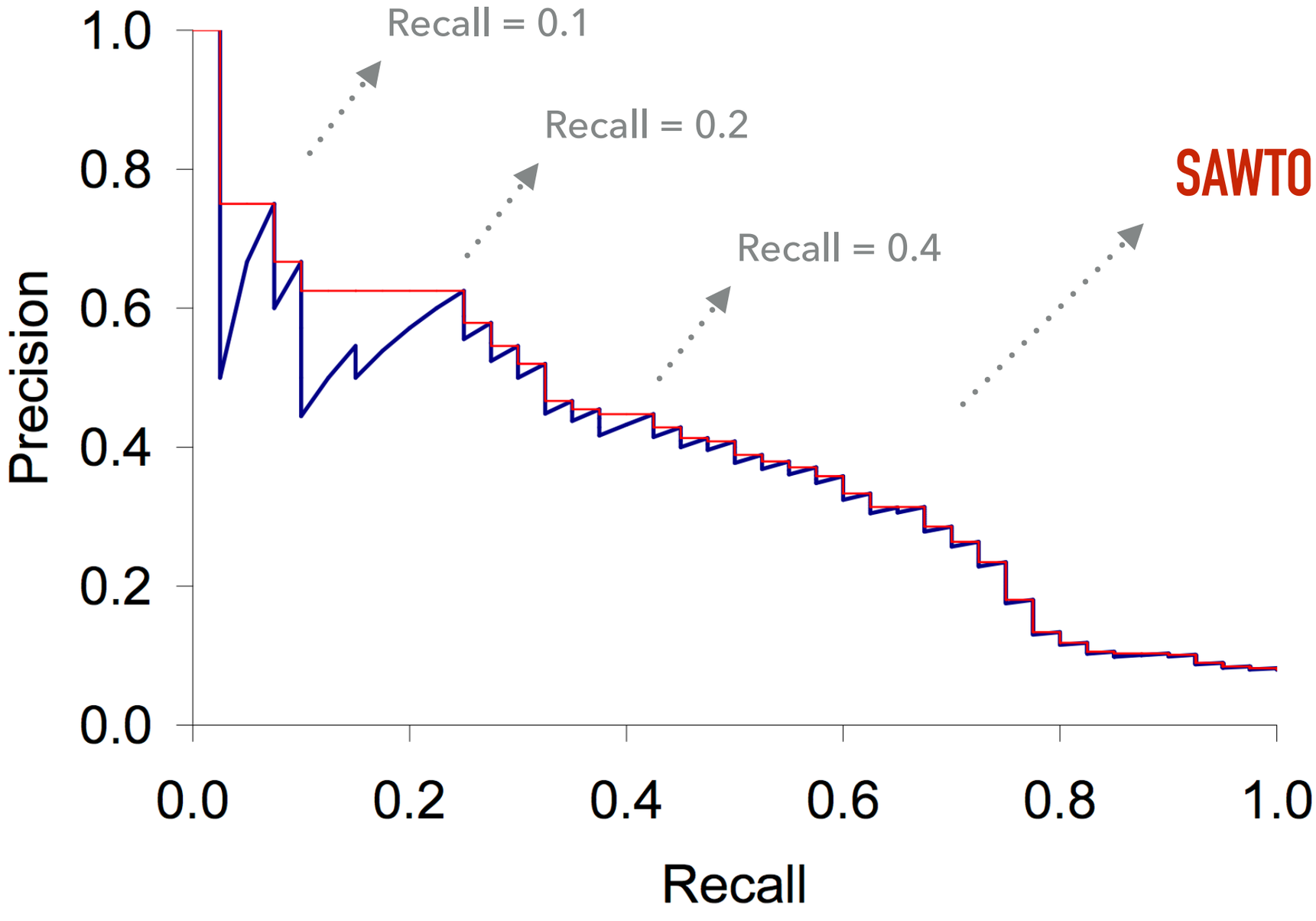
- We can do the same at different levels of recall:



SAWTOOTH SHAPE CURVE

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

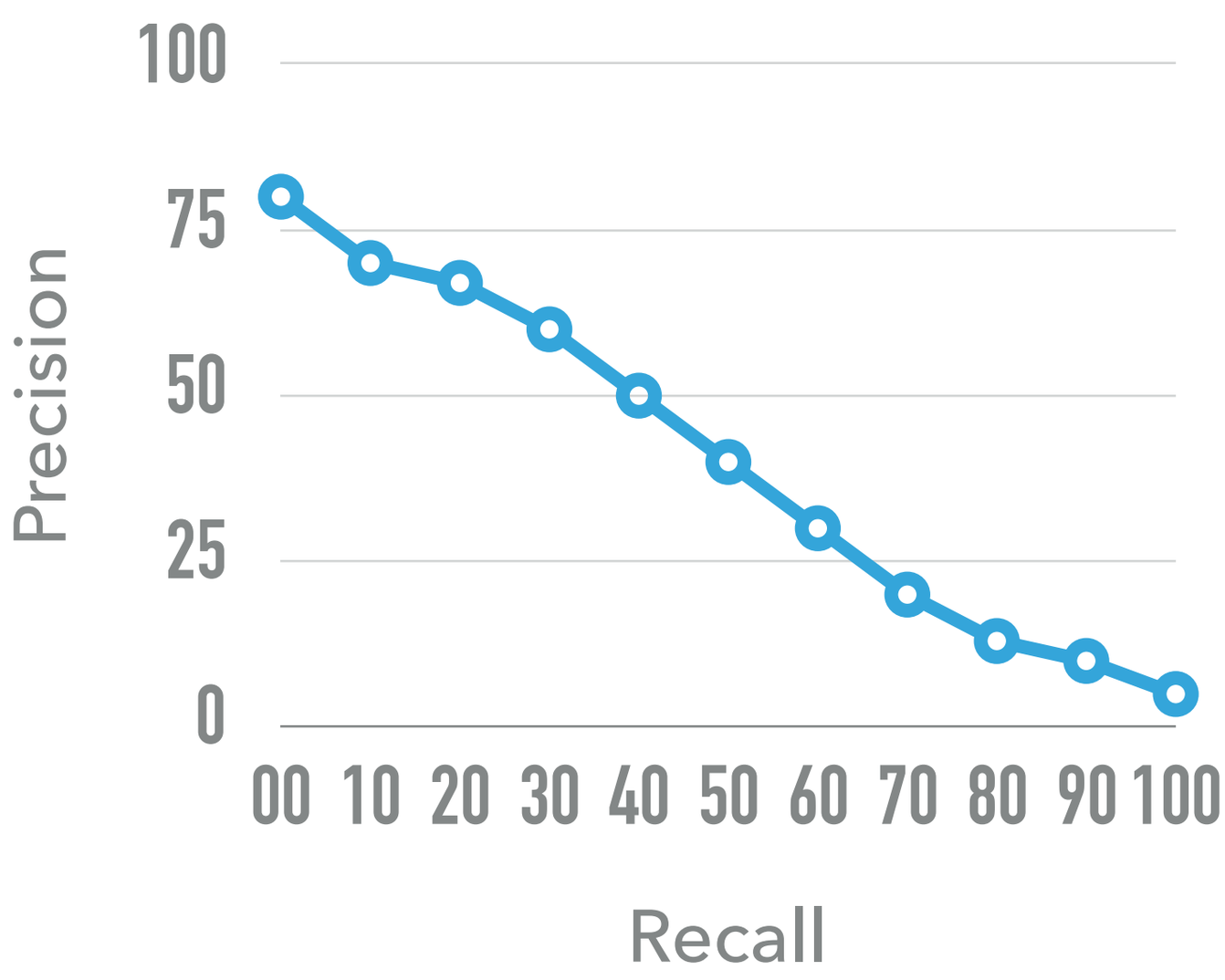
- We can do the same at different levels of recall:



SAWTOOTH SHAPE CURVE

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

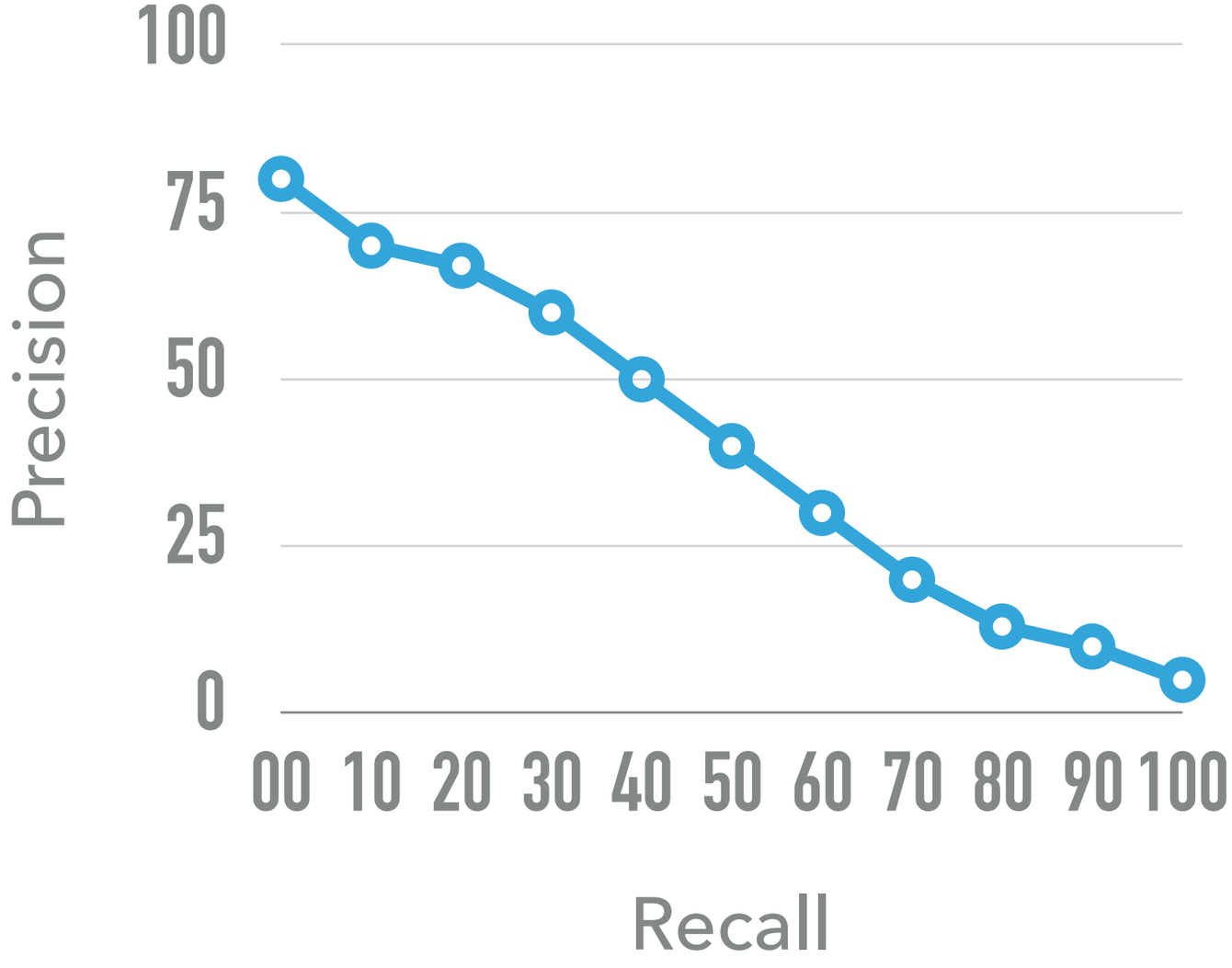
- We can do the same at different levels of recall:



COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- We can do the same at different levels of recall:

A 11-POINT PRECISION-RECALL GRAPH



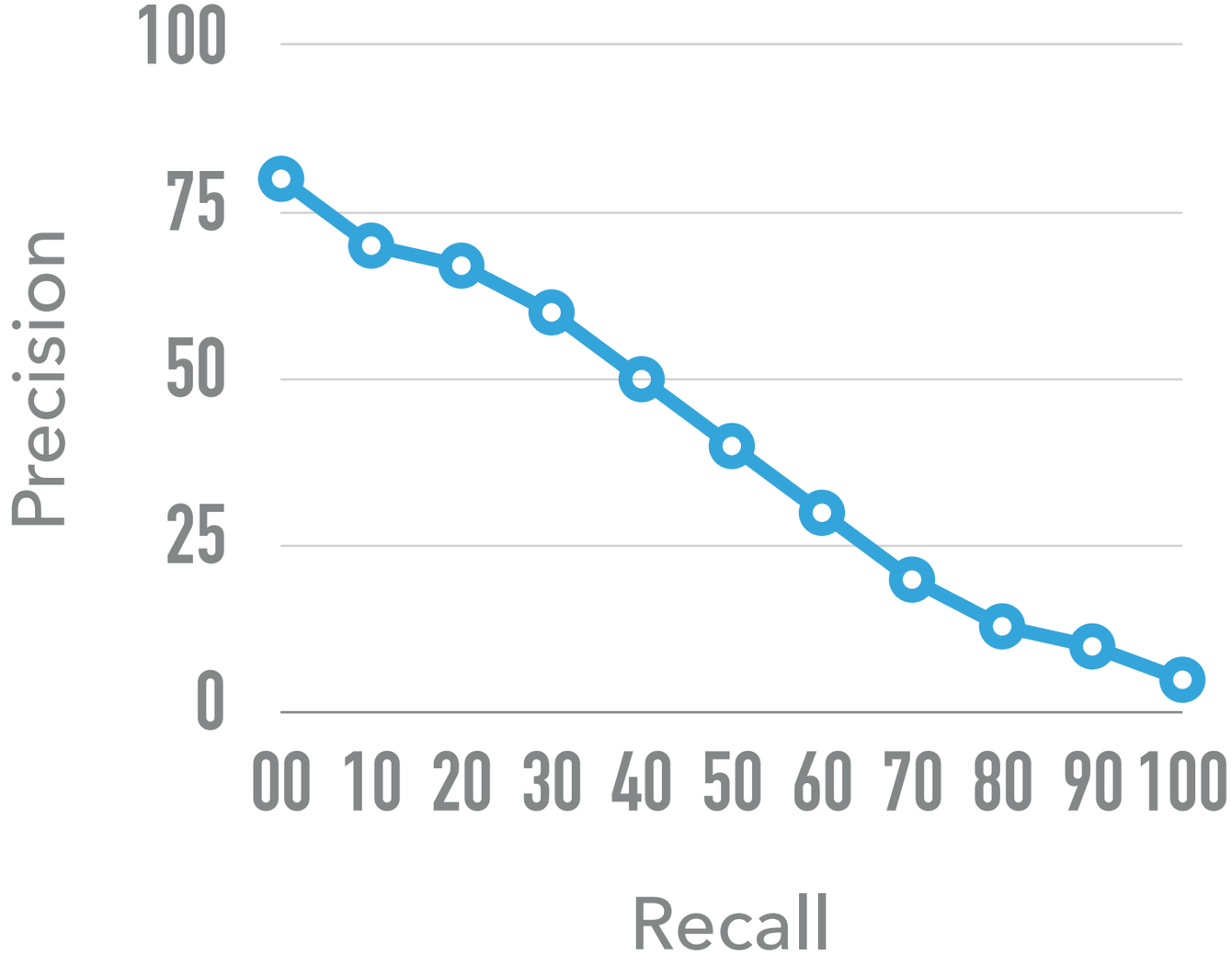
Note

Every one of the points in this graph is the average of all queries in the collection!

COMBINING PRECISION AND RECALL IN RANKED RETRIEVAL SET

- We can do the same at different levels of recall:

A 11-POINT PRECISION-RECALL GRAPH



Note

Every one of the points in this graph is the average of all queries in the collection!

Very descriptive graph...
But we still want one single number to compare systems....

METRICS TO SUMMARIZE PERFORMANCE

- Binary relevance:
  - Average of Eleven-point interpolated average precision
  - Precision@K (P@K)
  - Precision@Recall (P@R)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)
- Multiple grades of relevance
  - Normalized Discounted Cumulative Gain (NDCG)
  - Rank Biased Precision (RBP)

PRECISION AT K (P@K)

- Precision at a pre-determined level, e.g. K = 10; P@10
- Good metric for web retrieval
  - K = 10 – considers only the first page retrieved by a search engine

PRECISION AT K (P@K)

- Precision at a pre-determined level, e.g. K = 10; P@10
- Good metric for web retrieval
  - K = 10 – considers only the first page retrieved by a search engine
- Pro: Easy to understand and explain!

PRECISION AT K (P@K)

- Precision at a pre-determined level, e.g. K = 10; P@10
- Good metric for web retrieval
  - K = 10 – considers only the first page retrieved by a search engine

Collection has 100 docs

Query 1 has 1 relevant doc
Query 2 has 100 relevant doc

- Pro: Easy to understand and explain!
- Con1: Hard to compare across different queries

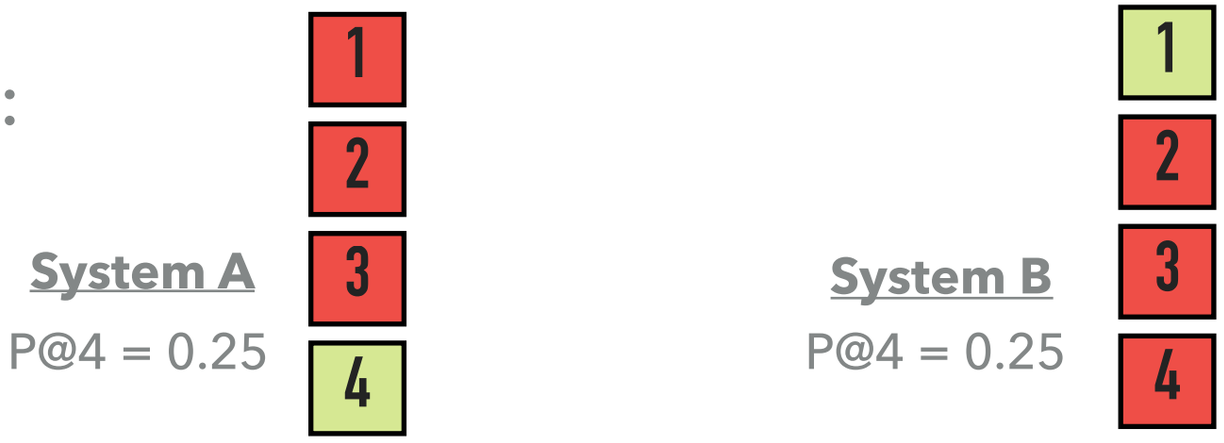
PRECISION AT K (P@K)

- Precision at a pre-determined level, e.g. K = 10; P@10
- Good metric for web retrieval
  - K = 10 – considers only the first page retrieved by a search engine
- Pro: Easy to understand and explain!
- Con1: Hard to compare across different queries
- Con2:

PRECISION AT K (P@K)

- Precision at a pre-determined level, e.g. K = 10; P@10
- Good metric for web retrieval
  - K = 10 – considers only the first page retrieved by a search engine

- Pro: Easy to understand and explain!
- Con1: Hard to compare across different queries
- Con2:



PRECISION AT R

- Adapts the value of K to the number of relevant documents for a given query
- Query with only one relevant document: P@1
- Query with 100 relevant documents: P@100
- Pro: Better to average across different queries
- Con: Harder understand and explain!

MEAN AVERAGE PRECISION (MAP)

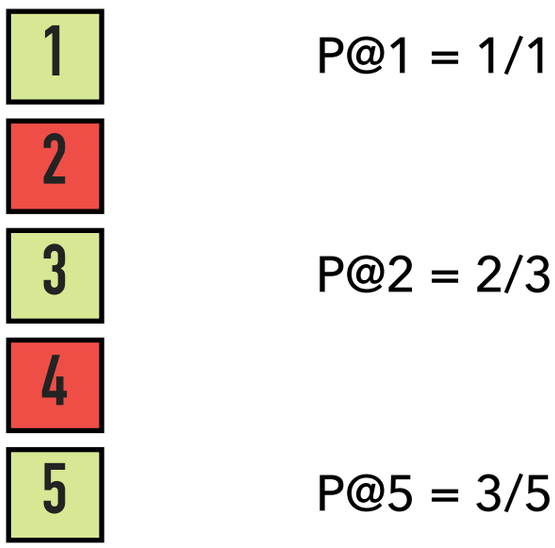
- Compute P@K for each K of relevant documents
- Average those P@Ks

MEAN AVERAGE PRECISION (MAP)

- Compute P@K for each K of relevant documents
- Average those P@Ks

MEAN AVERAGE PRECISION (MAP)

- Compute P@K for each K of relevant documents
- Average those P@Ks



MEAN AVERAGE PRECISION (MAP)

- Compute P@K for each K of relevant documents
- Average those P@Ks

1 P@1 = 1/1

2

3 P@2 = 2/3

4

5 P@5 = 3/5

$$AvePrec = \frac{(\frac{1}{1} + \frac{2}{3} + \frac{3}{5})}{3} = 0.75$$

MEAN AVERAGE PRECISION (MAP)

- Compute P@K for each K of relevant documents
- Average those P@Ks

1 P@1 = 1/1

2

3 P@2 = 2/3

4

5 P@5 = 3/5

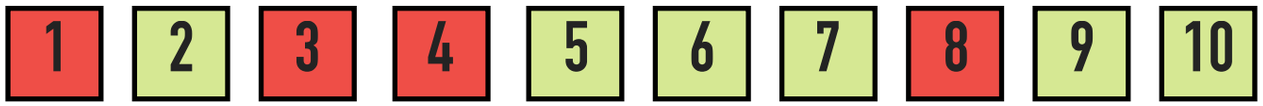
$$AvePrec = \frac{(\frac{1}{1} + \frac{2}{3} + \frac{3}{5})}{3} = 0.75$$

- MAP is the mean of Average Precision across multiple queries

MEAN AVERAGE PRECISION (MAP)

Query 1:

Query 2:



MEAN AVERAGE PRECISION (MAP)

Query 1: 

AVP Query 1 = (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78

Query 2: 

AVP Query 2 = (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6) / 6 = 0.52

MEAN AVERAGE PRECISION (MAP)

Query 1: 

AVP Query 1 = (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78

Query 2: 

AVP Query 2 = (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6) / 6 = 0.52

MAP = (0.78 + 0.52) / 2 = 0.53

MEAN RECIPROCAL RANK (MRR)

- There is only **one single** relevant document for the query OR only consider the **rank of the first** relevant document
- Queries like: “Facebook”, “BBC”, “CNN”, “QNB”
- Widely used in Q&A evaluation

MEAN RECIPROCAL RANK (MRR)

- There is only **one single** relevant document for the query OR only consider the **rank of the first** relevant document
- Queries like: “Facebook”, “BBC”, “CNN”, “QNB”
- Widely used in Q&A evaluation

Reciprocal rank (RR) = 1/4

Reciprocal rank (RR) = 1/1

Reciprocal rank (RR) = 1/2

Mean Reciprocal Rank = 1/4 + 1/1 + 1/2 = 0.58

BEYOND BINARY RELEVANCE

- What if we want to have multiple relevance levels. Example:
  - **Key:** This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine. (relevance grade 2)
  - **Rel:** The content of this page provides information on the topic; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page. (relevance grade 1)
  - **Non**: The content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query (relevance grade 0)
  - **Junk**: This page does not appear to be useful for any reasonable purpose; it may be spam or junk. (relevance grade -2)

BEYOND BINARY RELEVANCE

- We need to metric that takes the relevance level into consideration:
  - A ***'key'*** document should give a greater reward than a ***'relevant'*** document

- Two important ideas to score a ranking list:

  - Idea 1: We will cumulate the individual rewards of documents
  - Idea 2: We will penalize documents that are retrieved in lower ranks

BEYOND BINARY RELEVANCE

- Cumulative gain:
  - Simple idea, document X get reward R_x

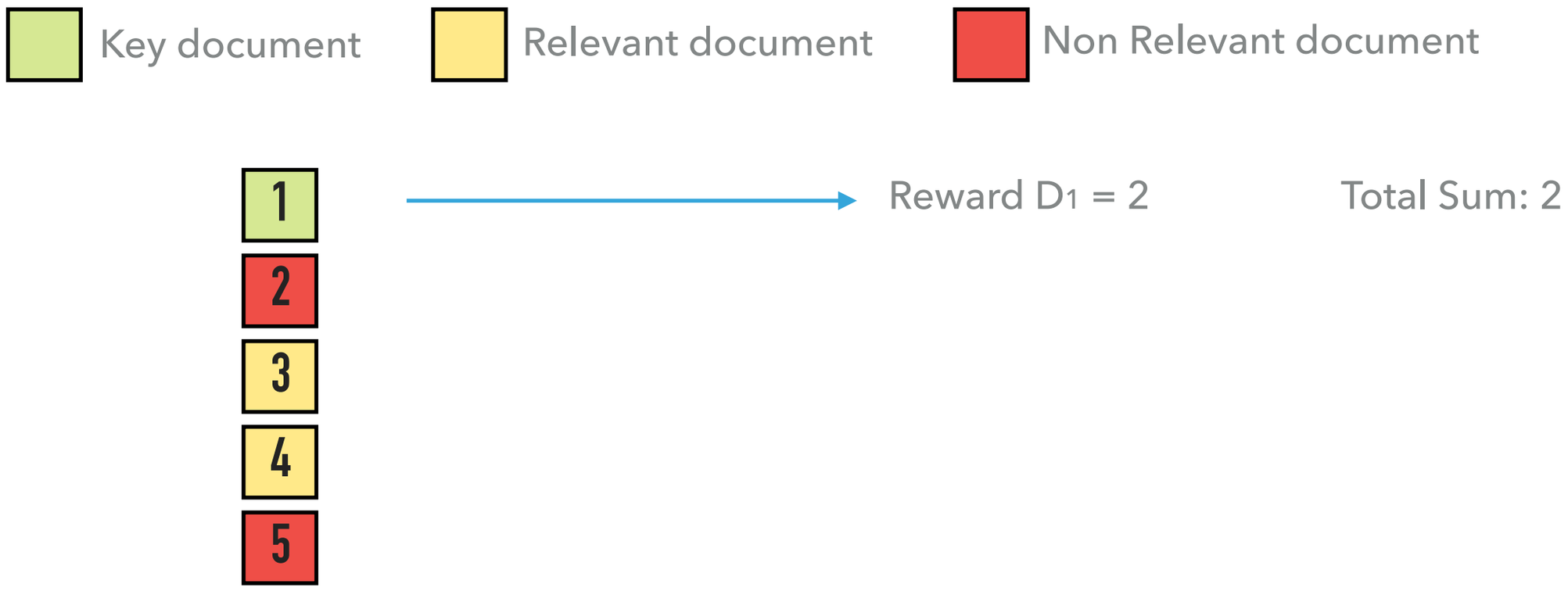
Key document Relevant document Non Relevant document

1
2
3
4
5

Total Sum: -

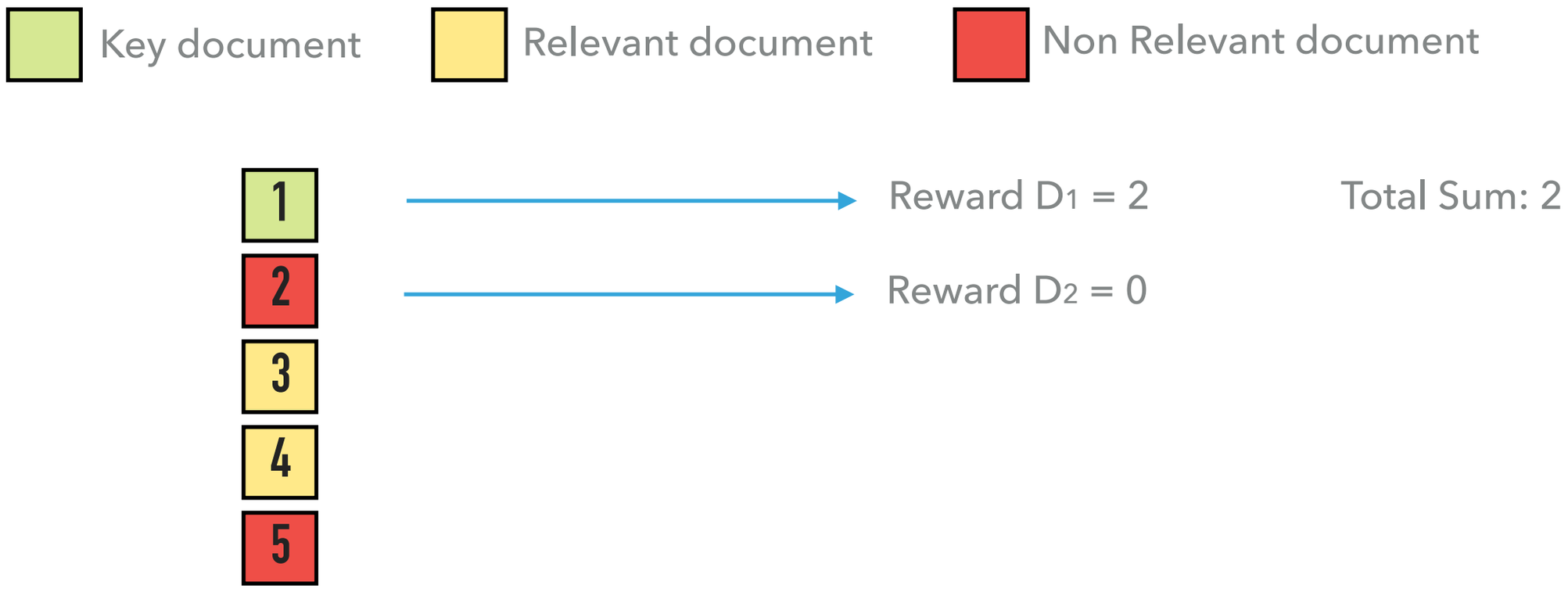
BEYOND BINARY RELEVANCE

- Cumulative gain:
  - Simple idea, document X get reward R_x



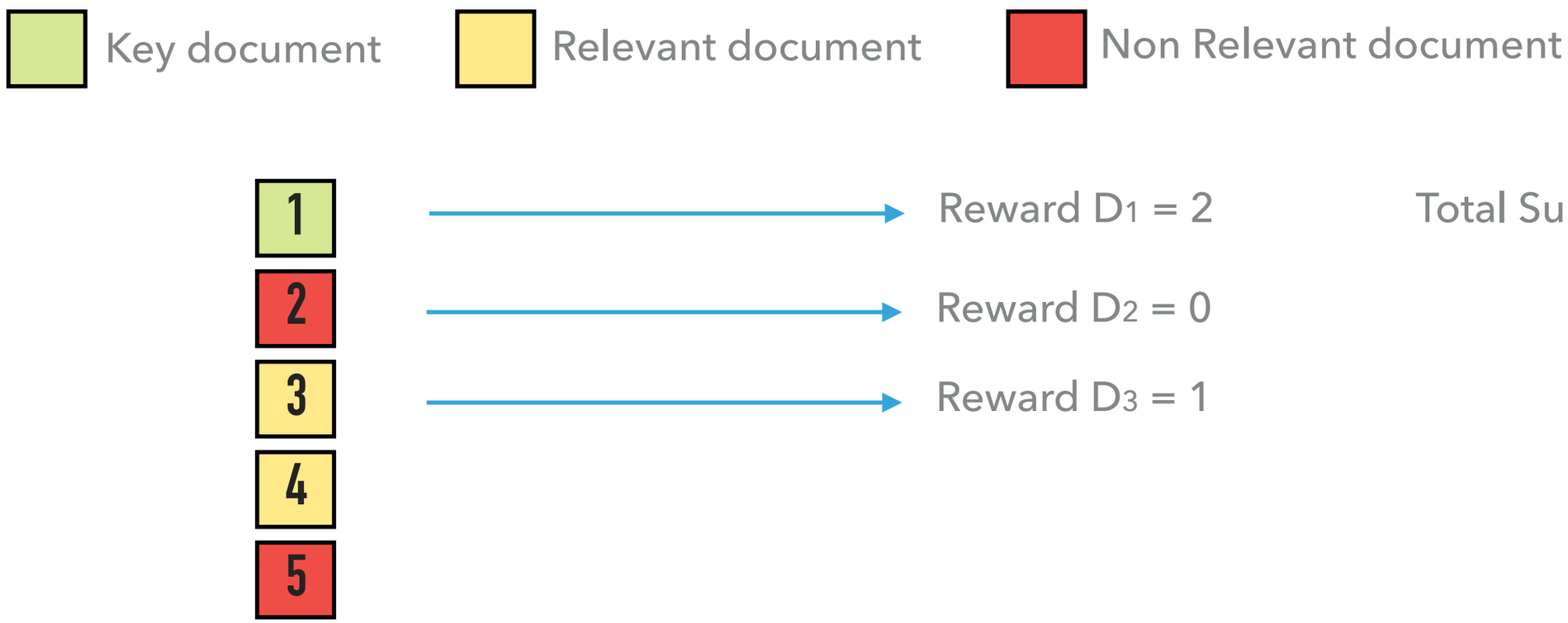
BEYOND BINARY RELEVANCE

- Cumulative gain:
  - Simple idea, document X get reward R_x



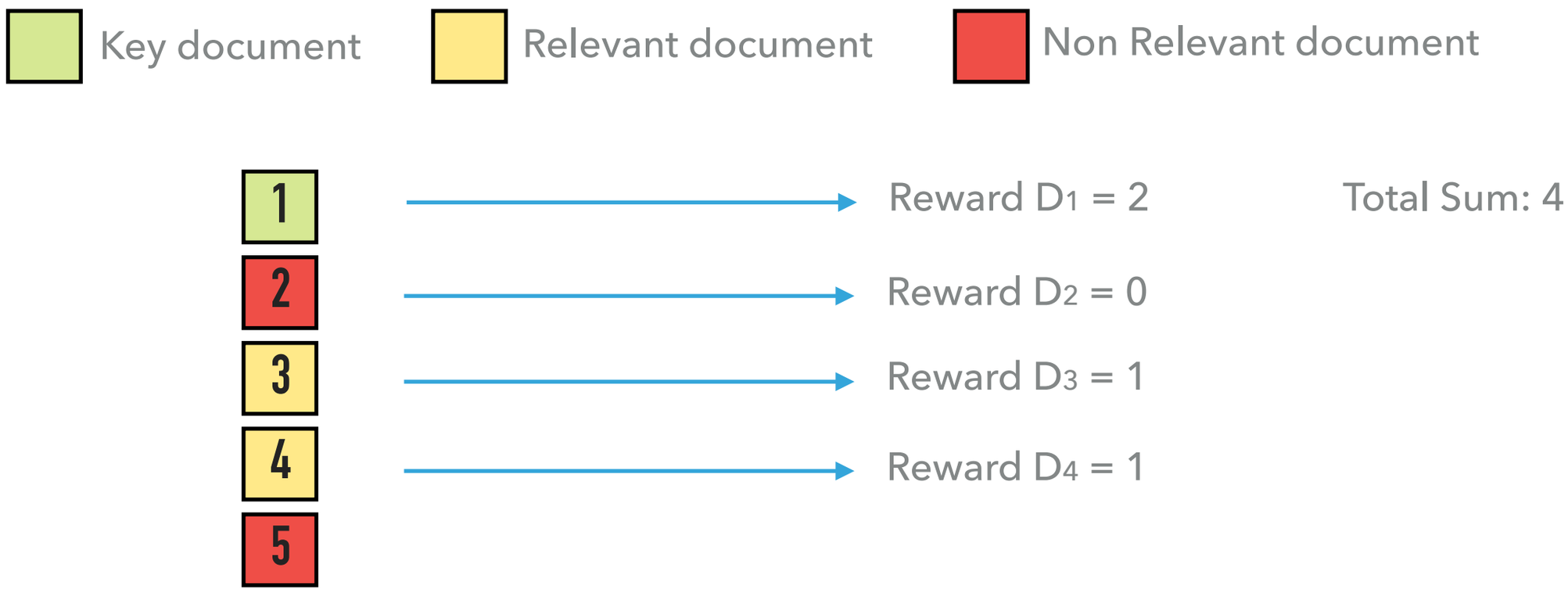
BEYOND BINARY RELEVANCE

- Cumulative gain:
  - Simple idea, document X get reward R_x



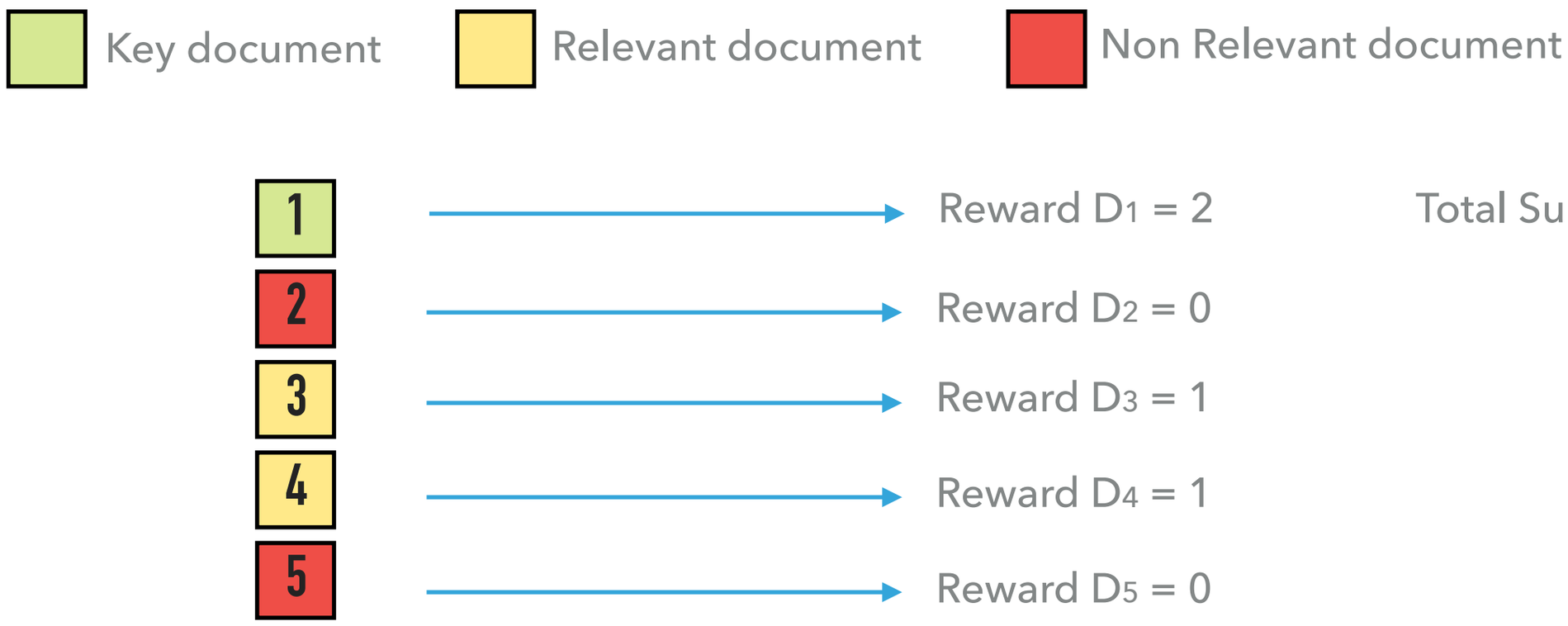
BEYOND BINARY RELEVANCE

- Cumulative gain:
  - Simple idea, document X get reward R_x



BEYOND BINARY RELEVANCE

- Cumulative gain:
  - Simple idea, document X get reward R_x

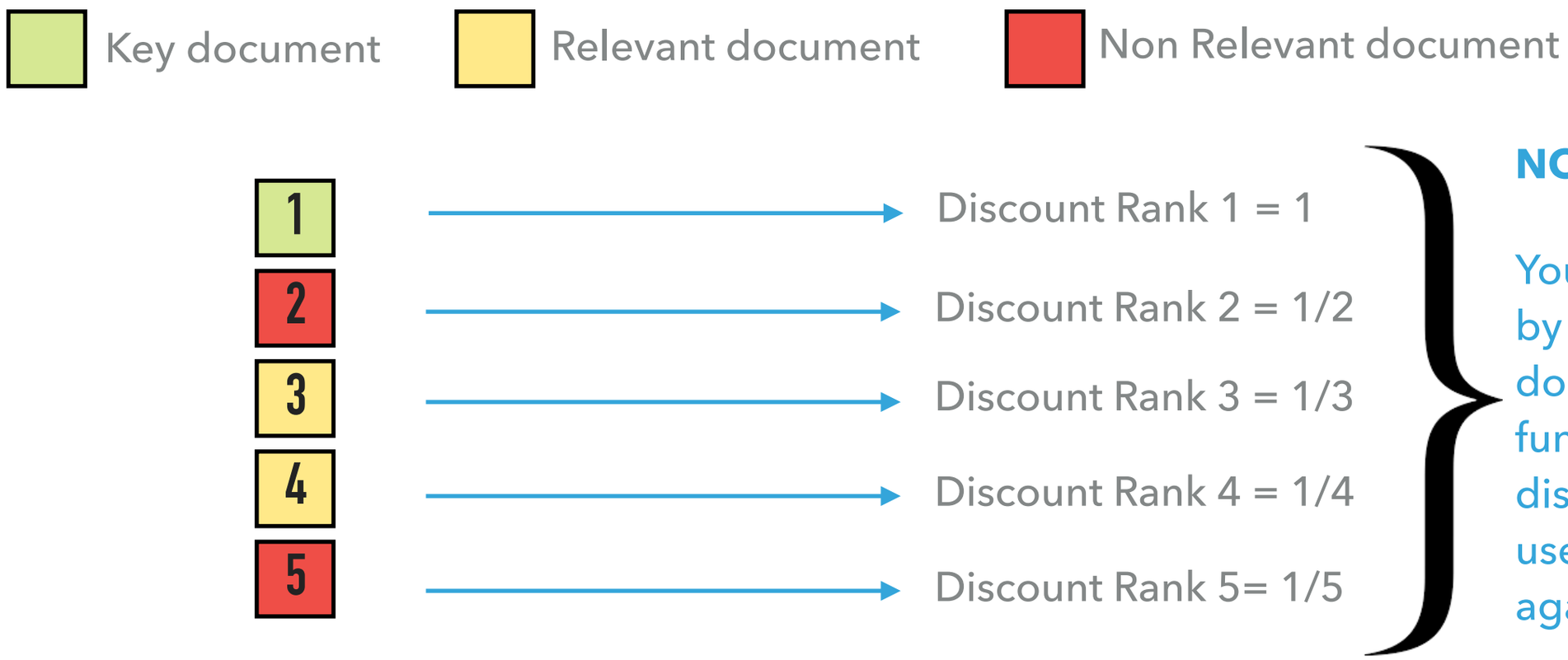


BEYOND BINARY RELEVANCE

- Discounts:
  - Documents at higher ranks are more valuable

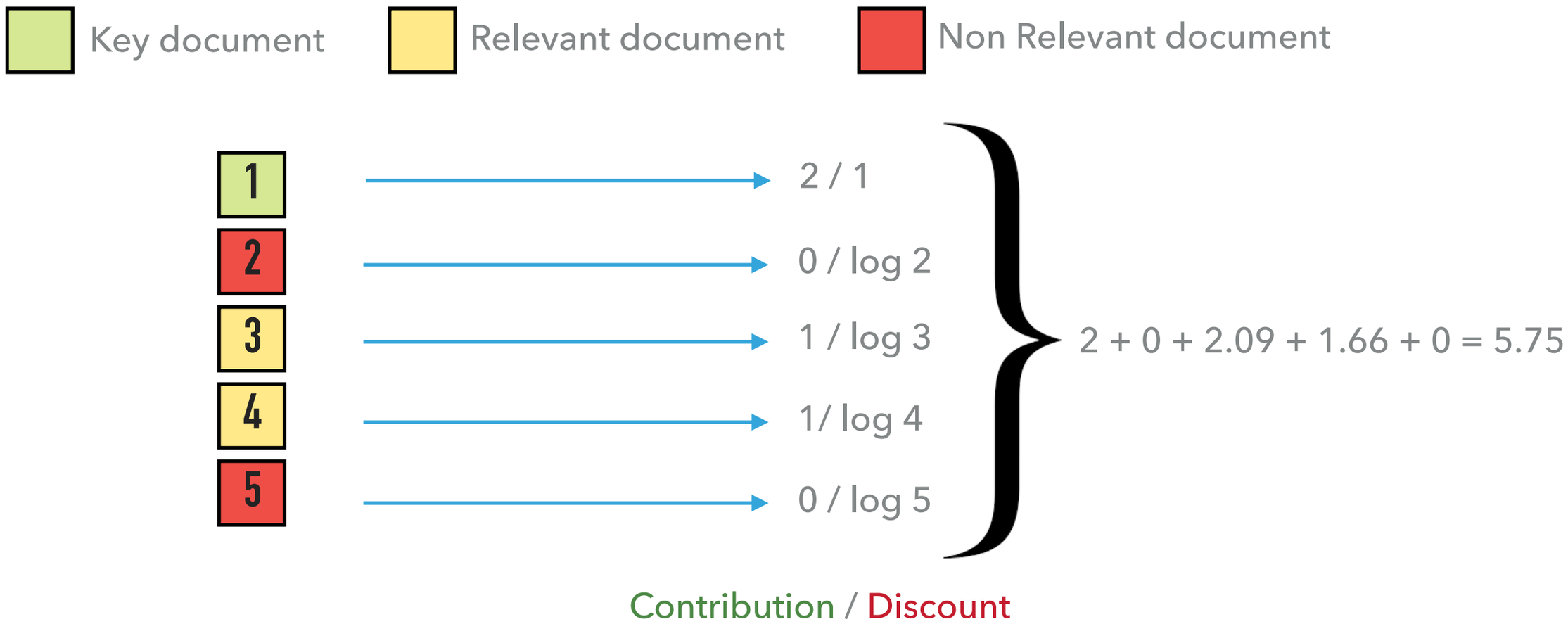
BEYOND BINARY RELEVANCE

- Discounts:
  - Documents at higher ranks are more valuable



BEYOND BINARY RELEVANCE

- Discounted Cumulative Gain (DCG)
  - Sum the contributions up to rank K



BEYOND BINARY RELEVANCE

- Discounted Cumulative Gain (DCG)
  - Formula:

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

- Alternative formula:

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

BEYOND BINARY RELEVANCE

- Nomalized Discounted Cumulative Gain (NDCG)
  - Normalization is made dividing DCG by ideal DCG in which documents are ranked in a optimal/ideal way

$$NDCG_p = \frac{DCG_p}{DCG_{ideal}}$$

- Useful to compare different queries
- Standard metric in many web search companies
- Widely used in related areas (e.g., recommender systems)

METRICS TO SUMMARIZE PERFORMANCE

- Binary relevance:
  - Average of Eleven-point interpolated average precision
  - Precision@K (P@K)
  - Precision@Recall (P@R)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)
- Multiple grades of relevance
  - Normalized Discounted Cumulative Gain (NDCG)
  - ~~Rank Biased Precision (RBP) –~~ ~~Missing, similar idea to NDCG~~

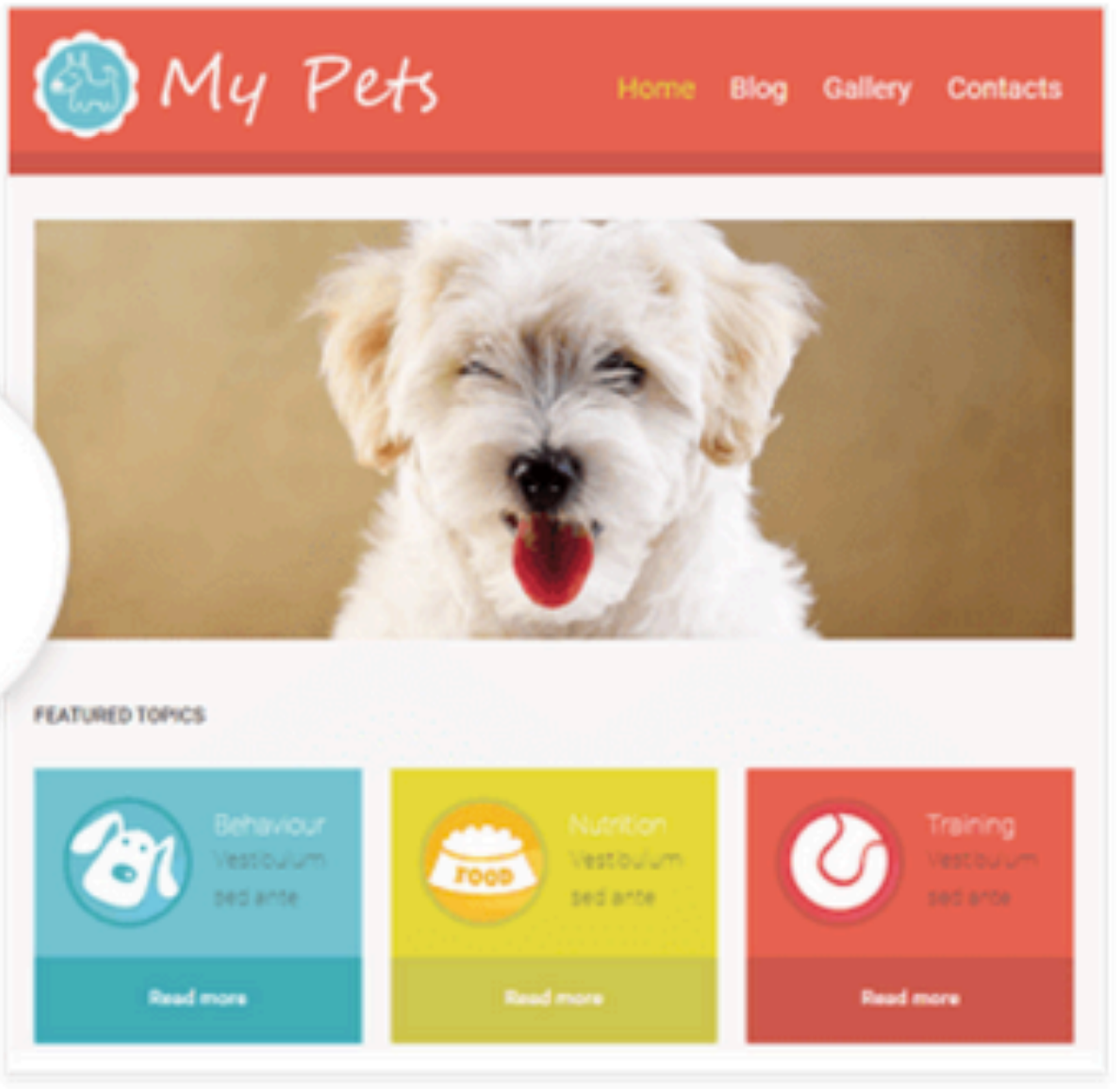
TESTING METHODOLOGY AT LARGE SEARCH ENGINES

- Test collections of queries and hand-ranked results
- Large use of logs to identify problematic queries
- Focus on precision at top 10 or less (mobile phone)
- Often use measures that reward you more for getting more relevant documents right at the top – **NDCG**
- Also:
  - User studies at lab
  - A/B testing
  - Result Interleaving

A/B TESTING METHODOLOGY

- Purpose: Test a single new feature/idea
- Prerequisite: You have a large search engine up and running with a very large number of daily requests
- 99% of users see system A, 1% of users see system B
- Evaluate with an automatic metric (# of clicks? # of purchases, time to perform first click? Rank of the first click?)
- Calculate if user happiness has increased with system B
- We are all constantly being tested...

VS



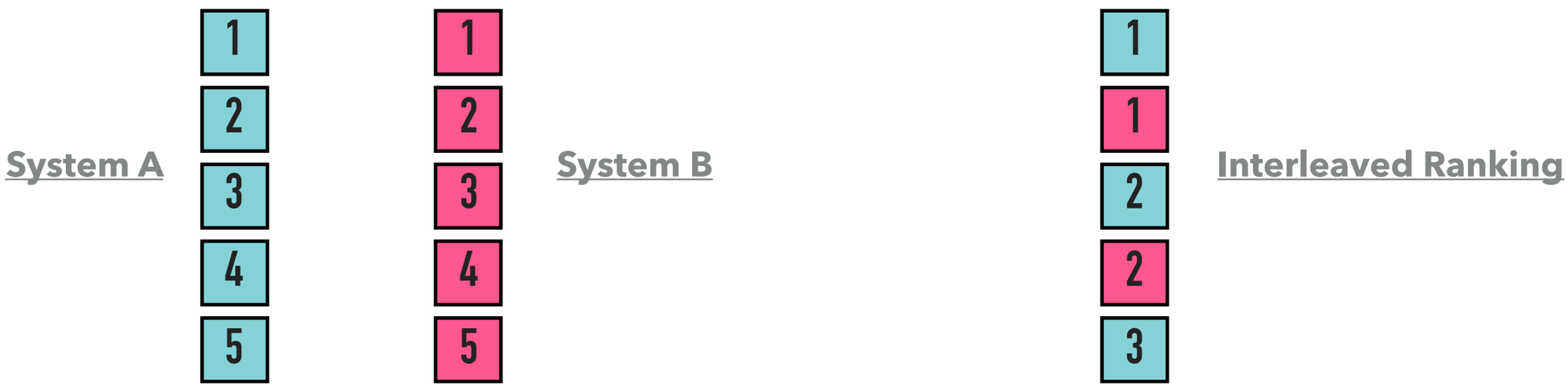
99%

1%

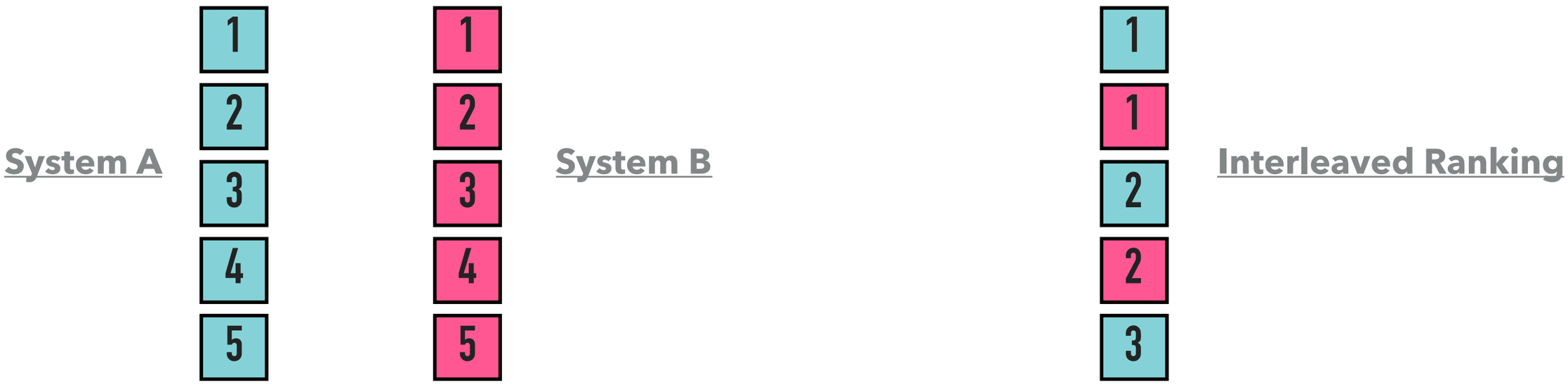
Those numbers are made up, but usually a very large portion of users sees the old system. Why?

INTERLEAVING METHODOLOGY

- Purpose: Direct comparison of two retrieval systems (like AB testing)
- Prerequisite: You have a large search engine up and running with a very large number of daily requests
- How? Main idea is inspired by how sport teams were chosen when we were kids:



INTERLEAVING METHODOLOGY



- How? First element is randomly chosen between both systems
- Evaluate: We count number of times a result from system A, system B and draws. Winner is simply the system with more clicks

EXTRA — STATISTICAL SIGNIFICANCE TESTS

- So far we have talked only on the comparison of averages
- Can we just compare averages?

Queries	System A	System B
1	0.10	0.21
2	0.11	0.22
3	0.09	0.18
4	0.08	0.20
5	0.10	0.20
6	0.12	0.19
Mean	0.10	0.20

Queries	System A	System B
1	0.28	0.21
2	0.05	0.22
3	0.01	0.38
4	0.01	0.15
5	0.07	0.05
6	0.18	0.19
Mean	0.10	0.20

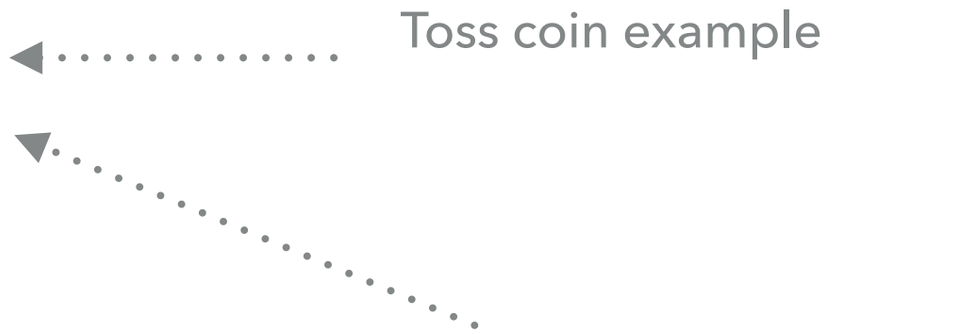
EXTRA — STATISTICAL SIGNIFICANCE TESTS

- Idea: Two groups (control, test)
- A treatment is given to the **test** group only. Control group takes placebo instead.
- Null hypothesis (H_0): No differences between test and control group
- Alternative hypothesis (H_1): the treatment must be affecting the test group
- Measure the p-value: probability of getting data as extreme as was observed

THINK OF COIN EXAMPLE

EXTRA — STATISTICAL SIGNIFICANCE TESTS

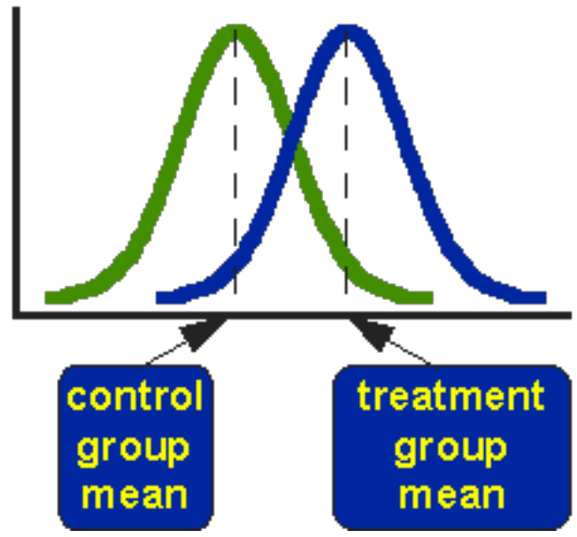
Queries	System A	System B	Sign test
1	0.28	0.21	+
2	0.05	0.22	-
3	0.01	0.38	-
4	0.01	0.15	-
5	0.07	0.05	+
6	0.18	0.19	-
Mean	0.10	0.20	2/6



EXTRA — STATISTICAL SIGNIFICANCE TESTS

Queries	System A	System B	Sign test
1	0.28	0.21	+
2	0.05	0.22	-
3	0.01	0.38	-
4	0.01	0.15	-
5	0.07	0.05	+
6	0.18	0.19	-
Mean	0.10	0.20	2/6

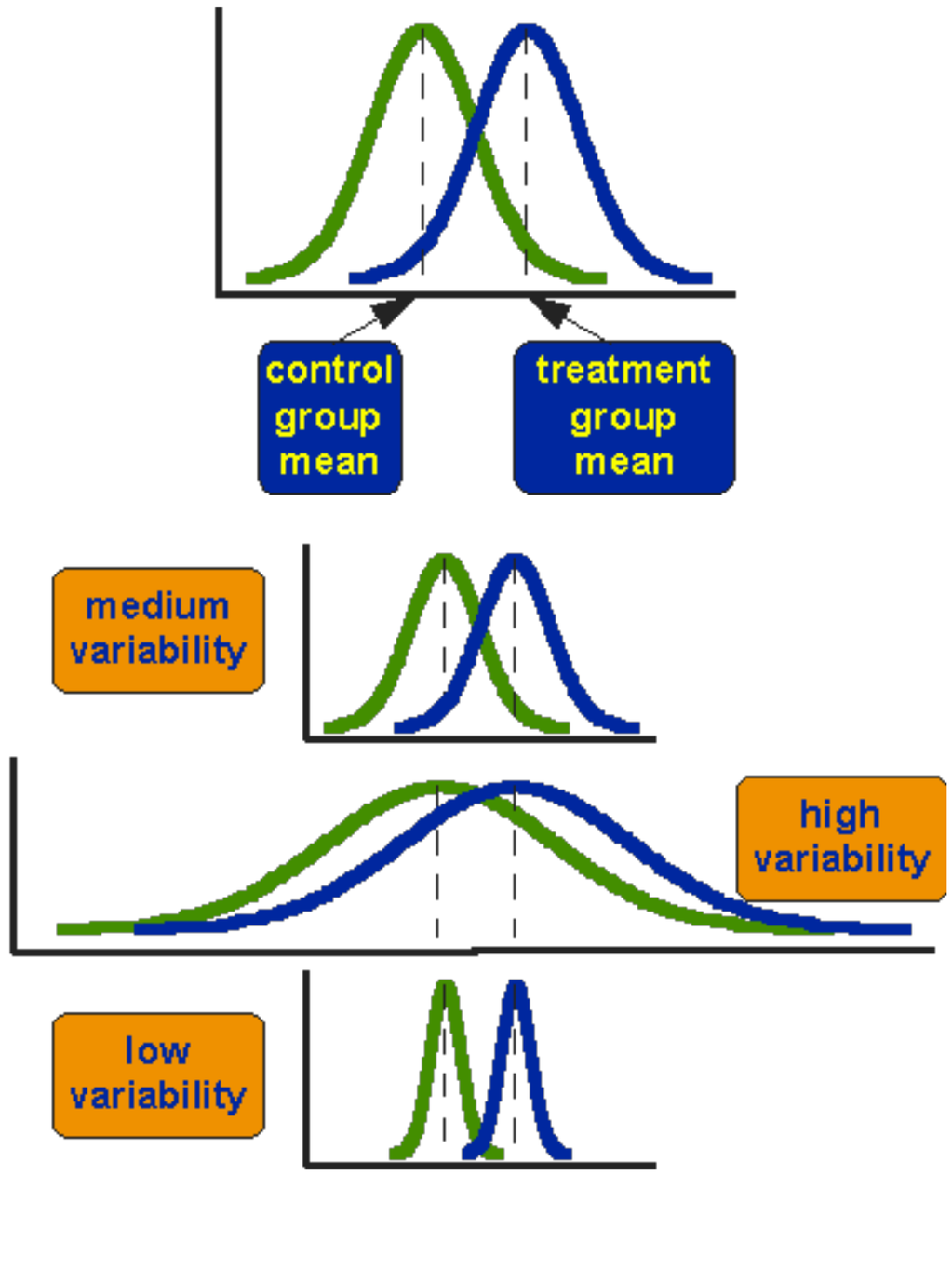
STUDENT T-TEST



EXTRA — STATISTICAL SIGNIFICANCE TESTS

Queries	System A	System B	Sign test
1	0.28	0.21	+
2	0.05	0.22	-
3	0.01	0.38	-
4	0.01	0.15	-
5	0.07	0.05	+
6	0.18	0.19	-
Mean	0.10	0.20	2/6

STUDENT T-TEST



WHAT DID WE SEE? WHAT SHOULD YOU KNOW?

- There is a large number of metrics in IR
- Different metrics should be used in different situations:
  - There is no single ninja metric
- They are all proxies for user happiness in a given context
- Cranfield experiments gave birth to modern IR evaluation (next lecture)
- Large search engine companies use most of the methods academics use + few new methods (AB testing, Interleaving)

TODAY'S LECTURE IN THE STANFORD IR BOOK

- Chapter 8: Evaluation in information retrieval

HOMEWORK 2

- How is it going so far?