

Evaluación

¿Cómo saber si un sistema de IR funciona bien?


¿Cómo decidir las múltiples opciones en el diseño y configuración de un motor de búsqueda, optimizarlas y ponerlas a punto?

¿Cómo cuantificar y comparar qué sistema funciona mejor?

Necesidad de evaluar

- ♦ IR es una disciplina altamente empírica
- ♦ Guiar el desarrollo, validación, selección y optimización de modelos, algoritmos y sistemas de IR
 - Elegir modelos IR, variantes (p.e. en *tf-idf*), normalizaciones, stemming, stopwords, etc.
 - Ajuste de parámetros
 - Optimización y puesta a punto general del sistema
- ♦ Comparativas y diagnóstico
 - Evaluación de una respuesta, evaluación global del sistema
 - Identificación de fallos y puntos débiles
 - Comparación de sistemas

Calidad de un sistema de IR

- ◆ Satisfacción de la necesidad de información del usuario (utilidad)
 - **Relevancia** a la consulta 
 - Calidad del contenido, autoridad de la fuente...
 - Actualidad, entretenimiento, idioma, ayudas...
 - ...
- ◆ Requisitos técnicos
 - Efectividad y escalabilidad: tiempo de respuesta, query throughput, rapidez de indexado, cobertura...
 - Flexibilidad: actualización incremental, configurabilidad, extensibilidad...
 - Usabilidad, calidad de la UI, efectividad de los snippets, etc.
 - ...



Metodología



Cyril W. Cleverdon
(1914-1997)

- ♦ Offline vs. online
 - Orientado al sistema vs. estudios con usuarios
- ♦ Paradigma Cranfield
 - Colección, conjunto de consultas, juicios de relevancia
- ♦ Sistemas, métricas, comparación
 - Comparativas entre varios sistemas sobre diferentes métricas
 - Optimización: usar consultas de entrenamiento y consultas de test
- ♦ Simplificaciones
 - La relevancia depende sólo de la consulta y el doc
(nos abstraemos de otras cualidades del documento)
 - La relevancia no depende del usuario
 - Ni del tiempo, ni del contexto
 - La relevancia de un documento es independiente de la relevancia de los demás

} Por la dificultad
de obtener juicios
relativos (si se tuvieran
se podrían usar)

Relevancia (recordatorio)

- ♦ Un concepto central en IR
 - La base de la evaluación de sistemas
 - La base de algunos modelos formales probabilísticos
- ♦ Una propiedad de un par consulta / documento
 - El documento es relevante si **satisface la necesidad de información** que motiva la consulta
 - Comúnmente considerada binaria (relevante / no relevante)
 - Pero métricas más recientes consideran grados de relevancia

Métricas

- ♦ Aplican a una lista ordenada (ránking) de documentos + una consulta
- ♦ Devuelven un número real
 - Generalmente positivo, y típicamente (pero no forzosamente) en $[0,1]$
- ♦ Versión $@k$ aplica a los top k documentos del ránking
- ♦ Diferentes métricas captan diferentes matices
- ♦ Clásicas
 - Precisión, recall, media armónica, curva precisión/recall, R-precision, MAP, MRR, fallout...
- ♦ Más recientes
 - nDCG, ERR, RBP...
- ♦ Un campo muy abierto y activo de innovación e investigación
 - Formalización, unificación y generalización de métricas, modelos de usuario
 - Nuevas dimensiones: diversidad, novedad...

Métricas clásicas

$$P = \frac{|\text{Relevant} \cap \text{Returned}|}{|\text{Returned}|}$$

P.e. típica búsqueda Web



Representa una tarea donde el usuario quiere encontrar un nº razonable de docs relevantes y valora el ratio de docs relevantes por unidad de esfuerzo

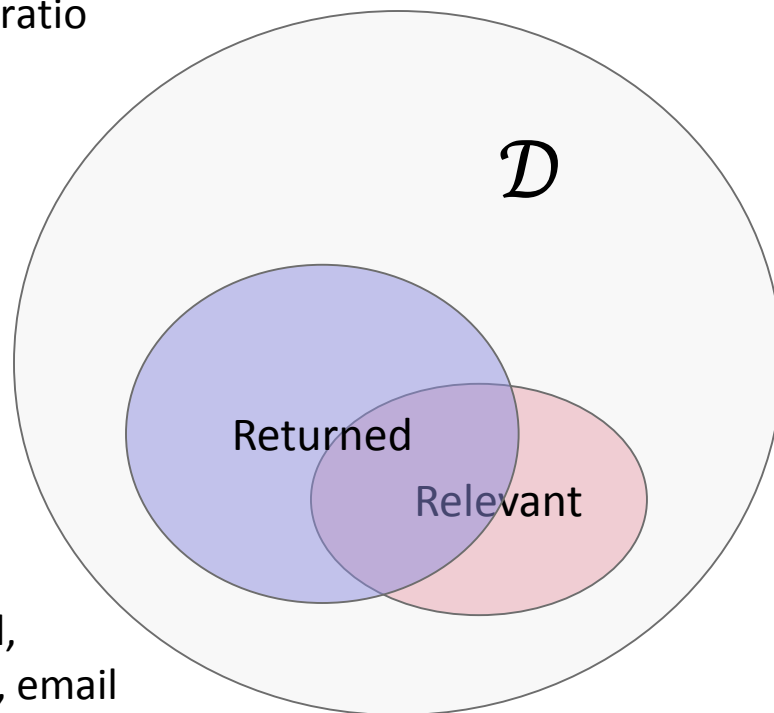
$$R = \frac{|\text{Relevant} \cap \text{Returned}|}{|\text{Relevant}|}$$

Medida de exhaustividad, representa una tarea donde el usuario quiere encontrar todos los docs relevantes y no repara en el esfuerzo de examinar docs irrelevantes

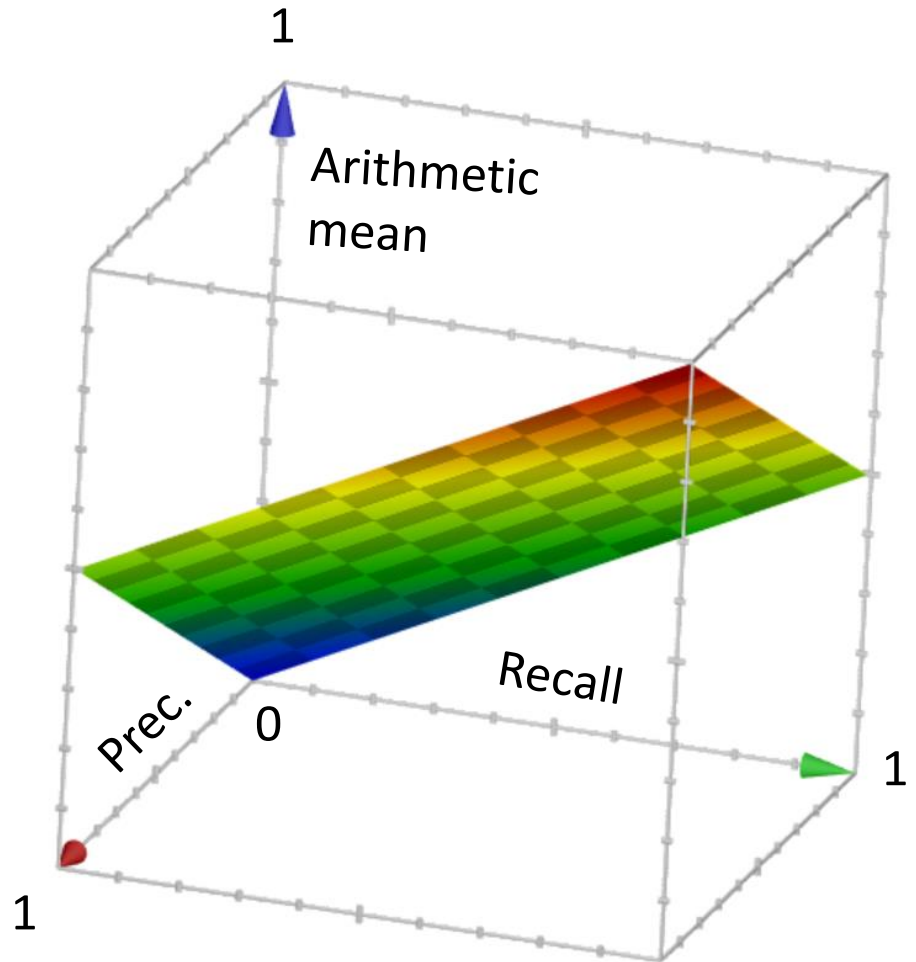
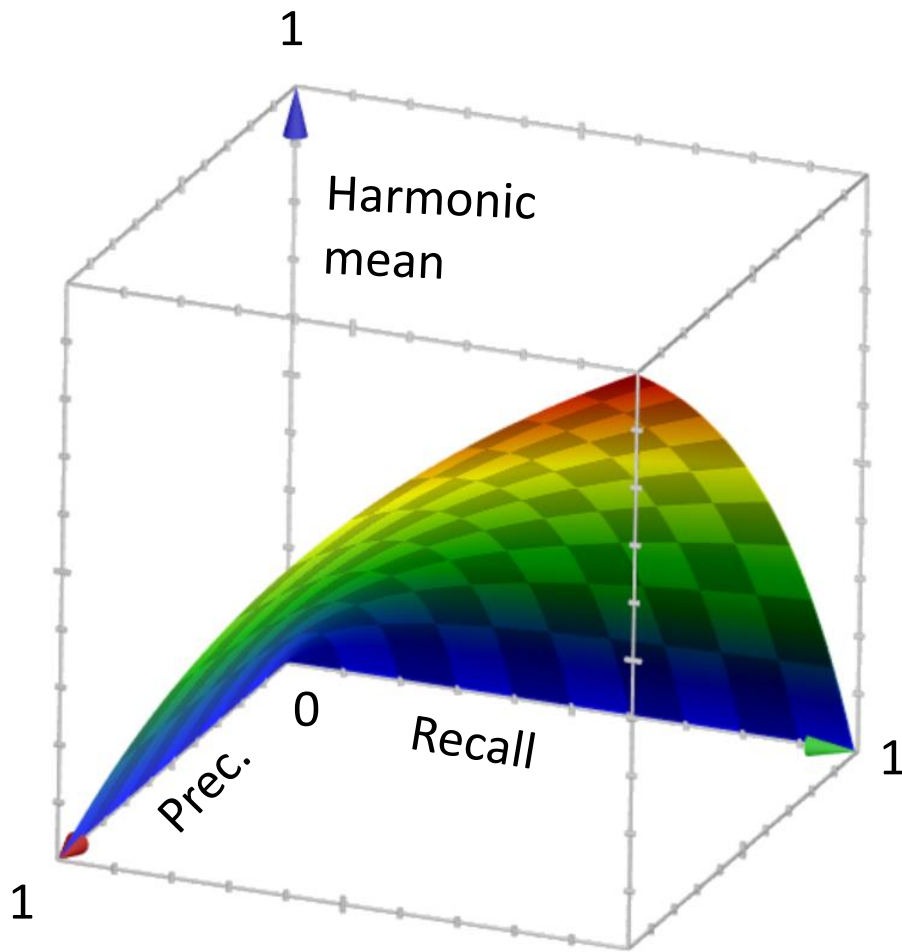
$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

P.e. dominio judicial,
patentes, escritorio, email

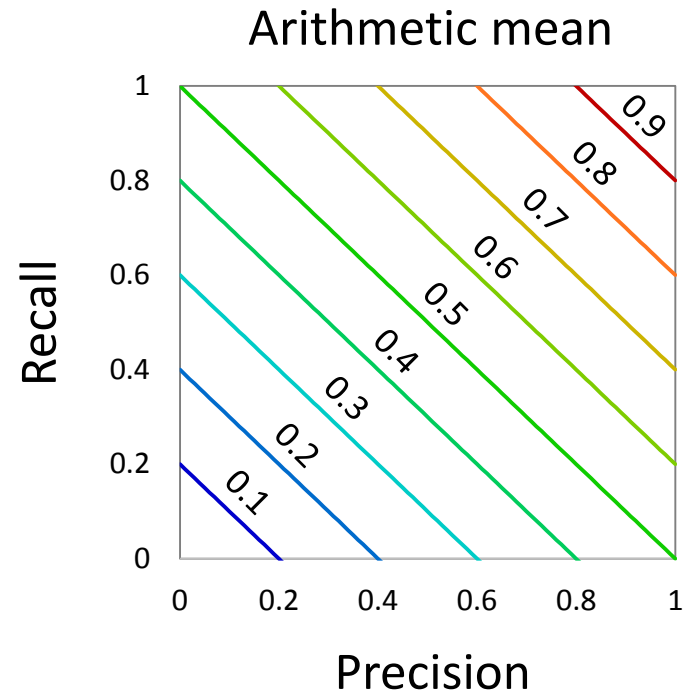
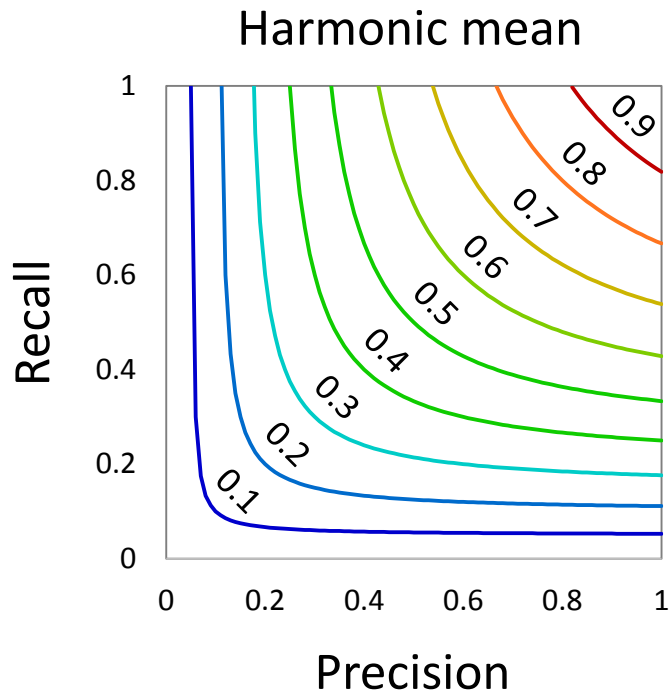
Media armónica, combinación de precisión y recall que penaliza valores muy bajos en una u otra métrica



Efecto de la media armónica



Efecto de la media armónica



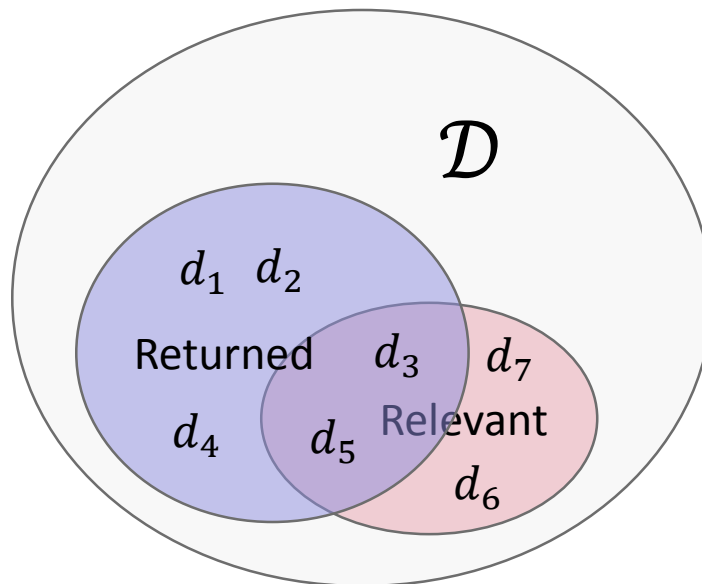
Ejemplo

Para una consulta q los documentos relevantes son:

$d_3 \ d_5 \ d_6 \ d_7$

Un sistema devuelve, por este orden, los documentos:

$d_1 \ d_2 \ d_3 \ d_4 \ d_5$



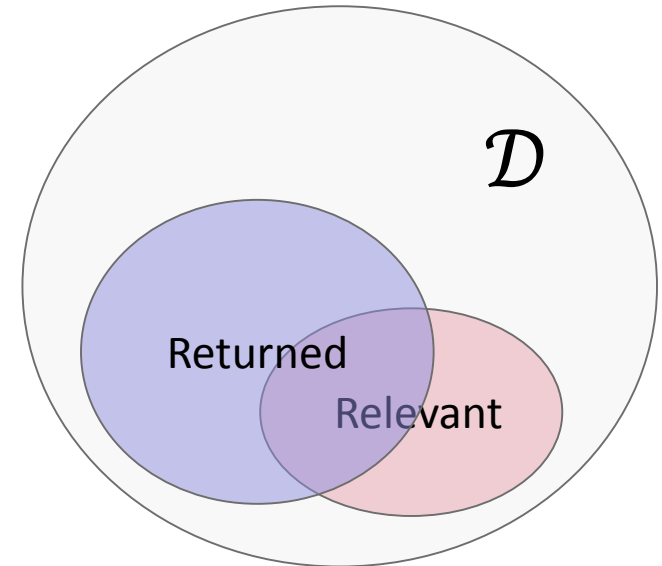
$$P = 2/5$$

$$R = 2/4$$

Otras formas de entender P y R

Matriz de confusión

	P	N
\hat{P}	TP	FP
\hat{N}	FN	TN



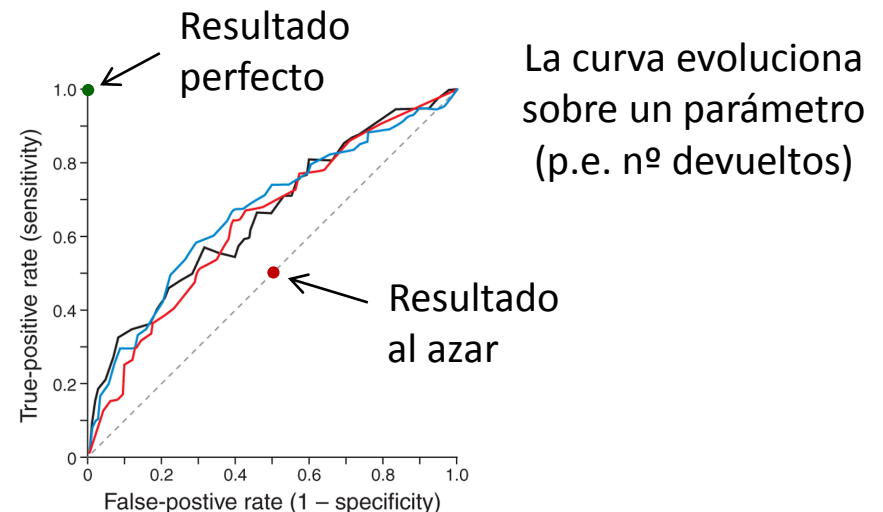
$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$Fallout = \frac{FP}{FP + TN}$$

¿Probabilidad de...?

Curva ROC: recall vs. fallout →
AUC: Area Under the ROC Curve



Métricas

- ♦ Para evaluar un sistema s , las métricas m se promedian sobre una batería de consultas \mathcal{Q} (p.e. mínimo 50)

$$m(s) = \text{avg}_{q \in \mathcal{Q}} \underbrace{m(s(\mathcal{D}, q), q)}_{\downarrow}$$

ránking devuelto por s en respuesta a q

- ♦ Es común también tomar las métricas “at k ” (top k rank cutoff)

$$m@k(R, q) = m(\text{top}k(R), q)$$

$$\text{P.e. } P@k = \frac{|\text{Relevant in top } k|}{k}$$

¿Por qué?

Evaluaciones comparativas

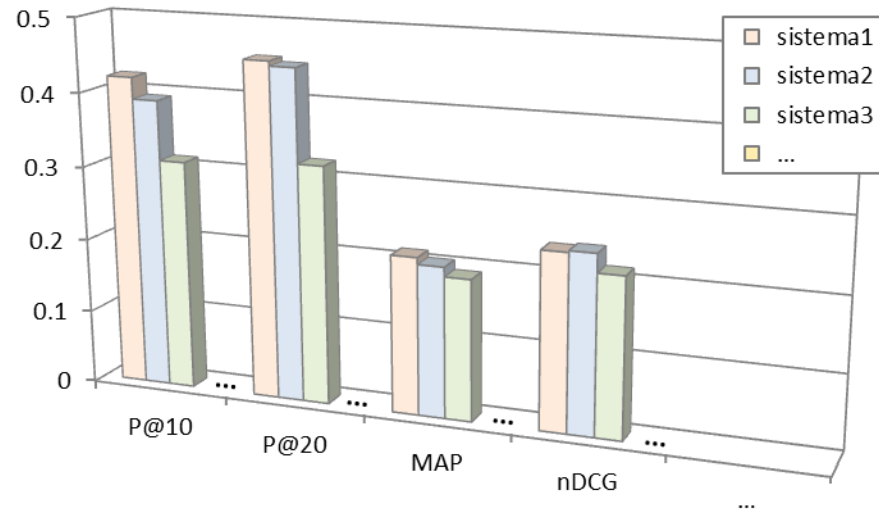
- ◆ Generalmente el objetivo es comparar varios sistemas

	$P@10$	$P@20$	MAP	$nDCG$...
sistema ₁	0.4201	0.4532	0.2124	0.2381	...
sistema ₂	0.3912	0.4467	0.2033	0.2405	...
sistema ₃	0.3109	0.3215	0.1906	0.2153	...
...

- ◆ Conviene medir la *significatividad estadística* de las comparaciones
 - Aunque lo obviaremos en esta asignatura

Evaluaciones comparativas

- ◆ Generalmente el objetivo es comparar varios sistemas



- ◆ Conviene medir la *significatividad estadística* de las comparaciones
 - Aunque lo obviaremos en esta asignatura

Juicios de relevancia

- ♦ ¿Cómo sabemos qué documentos son relevantes al evaluar?
- ♦ Se necesita “etiquetado” manual de los pares consulta / doc (a.k.a. *ground truth*, *gold standard*)
 - Campañas de evaluación: TREC, NTCIR, CLEF...
 - El proveedor del motor de búsqueda recluta evaluadores (*assessors*, a pequeña o gran escala, ver p.e. Google quality rating program)
 - Se infiere una probabilidad de relevancia por los clicks de los usuarios
- ♦ En general difícilmente sabremos *todos* los relevantes
 - Se necesitaría anotar la relevancia de $Q \times \mathcal{D}!!$
 - Se suele hacer un muestreo (*pooling*)
 - Dificultad para calcular el valor exacto de métricas como recall o IDCG

Más métricas clásicas

$$AP = \frac{1}{|d_k \text{ relevant}|} \sum P@k \rightarrow MAP$$

Los relevantes no devueltos
se consideran posición ∞

Algunos autores toman $\min(|Relevant|, n)$
en el denominador de $AP@n$

$$R\text{-Precision} = P@|Relevant|$$

Más “justa” que $P@k$ p.e. cuando
hay menos de k docs relevantes

Es el punto donde $P@k = R@k$

Implicítamente combinan
precisión y recall, pues
dividen por $|Relevant|$

$$RR = \frac{1}{\min\{k | d_k \text{ relevant}\}} \rightarrow MRR$$

Representa una tarea
donde al usuario le basta
un doc relevante

Nuevas métricas

- ♦ No todos los documentos relevantes son igual de relevantes
 - No relevante, poco relevante, relevante, muy relevante...
 - **Grados de relevancia**
- ♦ No todas las posiciones en el ránking son iguales
 - Cuanto más alta la posición, más influye la relevancia del documento en la efectividad real para el usuario
 - Cuanto más baja la posición, menos probable que el usuario llegue a ver el documento
 - **Descuento por posición**
- ♦ No todas las consultas son igual de difíciles
 - **Normalización** para tener en cuenta la dificultad de la consulta

Nuevas métricas – nDCG

Normalized Discounted Cumulated Gain

$$nDCG = \frac{DCG}{IDCG}$$

\uparrow
Normalización
 $nDCG \in [0,1]$

$$DCG = \sum_k \frac{g(d_k)}{\log_2(k+1)}$$

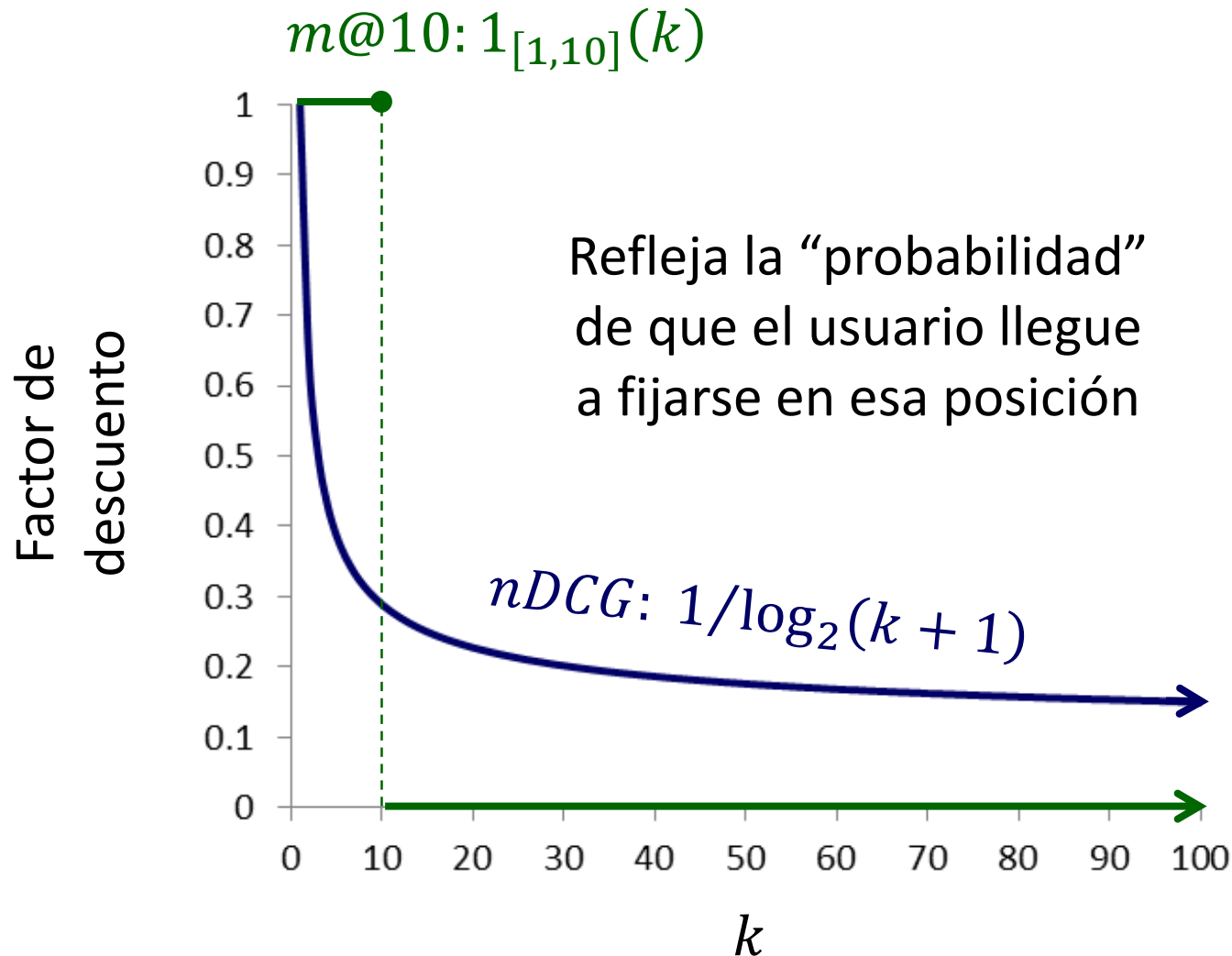
\leftarrow *Grado de relevancia*
 \leftarrow *Descuento por posición*

$\longrightarrow IDCG = \max_{R \in \sigma(\mathcal{D})} DCG(R, q)$

Con cutoff: $nDCG@k = \frac{DCG@k}{IDCG@k}$

K. Jarvelin, J. Kekalainen. Cumulated gain-based evaluation of IR techniques.
ACM Transactions on Information Systems 20(4), 2001, pp. 422–446

Descuento por posición



Ver <https://www.advancedwebranking.com/cloud/ctrstudy>

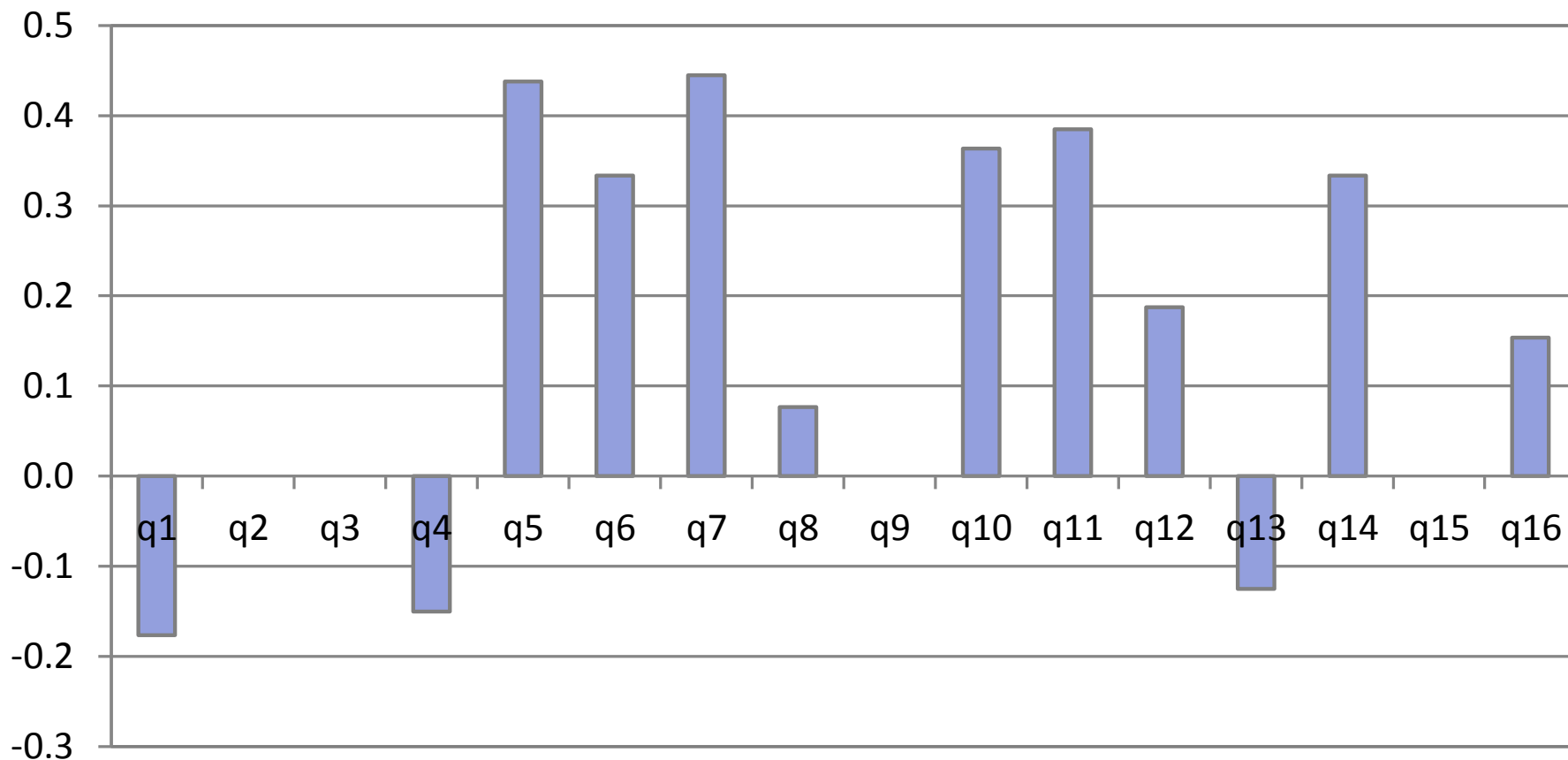
Examinar consultas individuales

Histograma p.e. R-precision

- ◆ Permite comparar dos sistemas en detalle consulta a consulta
- ◆ Permite observar diferencias que se difuminan en el promedio sobre las consultas: varianza, outliers, tipos de consulta, etc.
- ◆ Mostrar $R\text{-precision}(s_1, q_i) - R\text{-precision}(s_2, q_i)$ para varias consultas q_i como diagrama de barras

Ejemplo

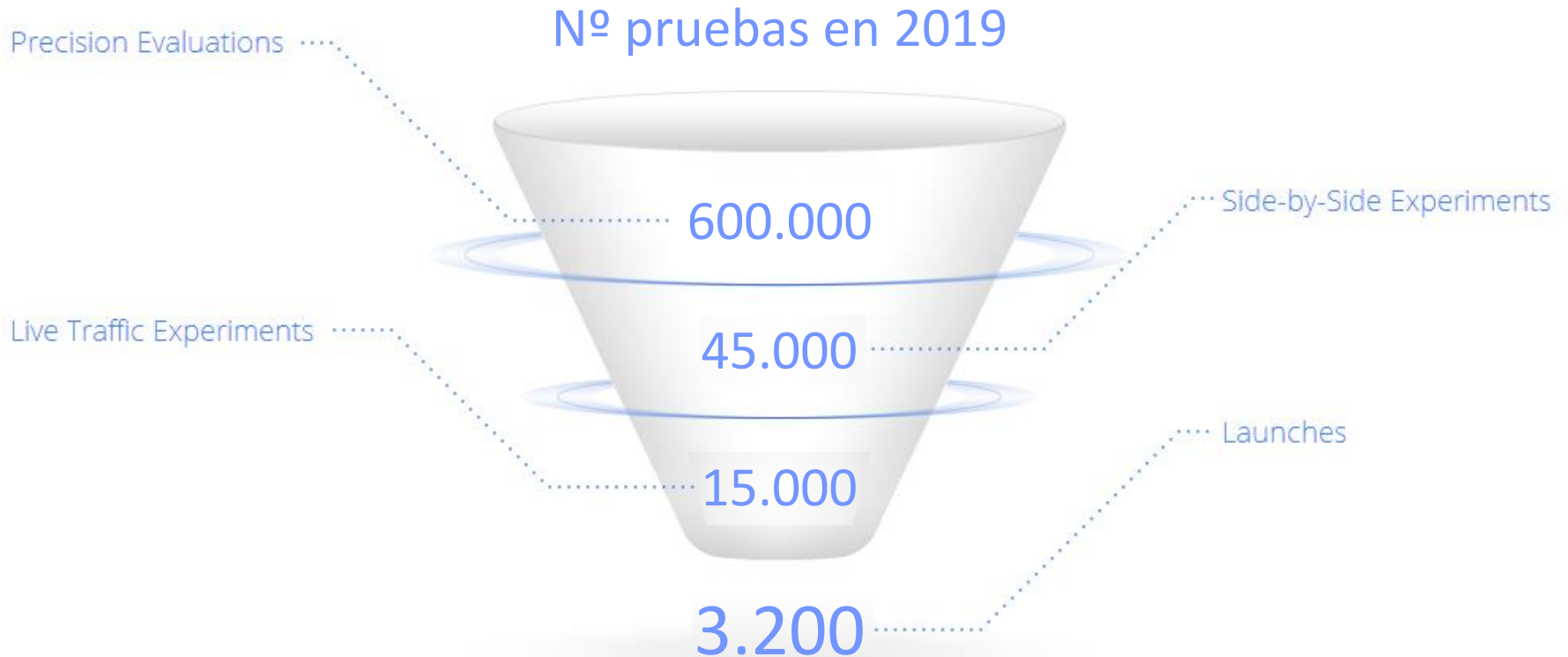
R-Precision histogram Google vs. Bing



Evaluación online

- ♦ Experimentos en laboratorio
- ♦ Crowdsourcing
- ♦ Tests A/B
- ♦ Evaluación con logs de búsqueda y clickthrough

Ejemplo protocolos evaluación en Google



<https://www.google.com/search/howsearchworks/mission/users>

Ver también...

- <http://www.youtube.com/watch?v=J5RZOU6vK4Q>
- <http://googleblog.blogspot.com.es/2008/09/search-evaluation-at-google.html>

Sobre Google raters...

- <http://searchenginewatch.com/article/2172154/How-Google-Uses-Human-Raters-in-Organic-Search>
- <https://www.youtube.com/watch?v=nmo3z8pHX1E>
- <http://searchengineland.com/interview-google-search-quality-rater-108702>