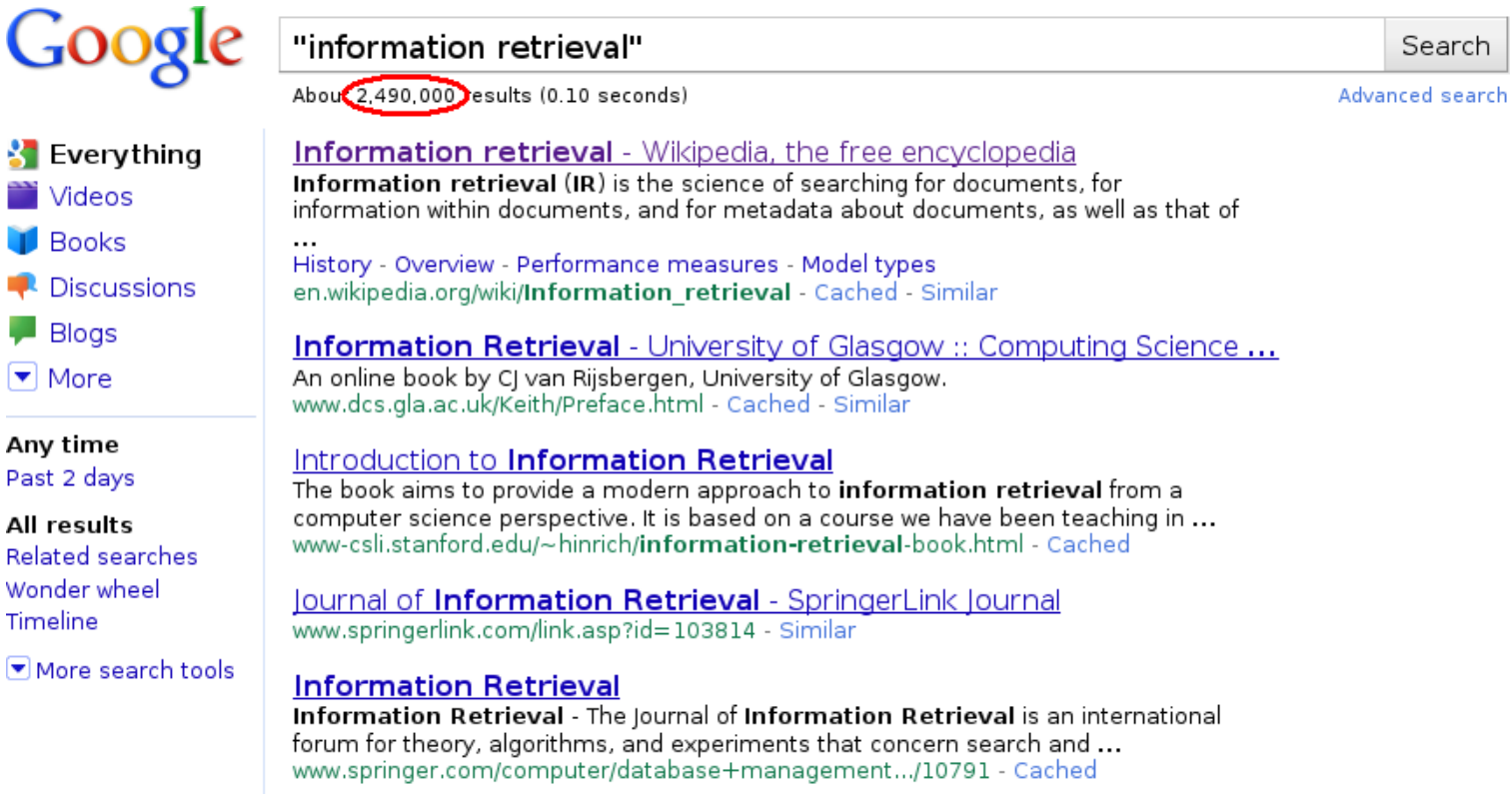
Making the most of a Web search session

Benno Stein Matthias Hagen

Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de

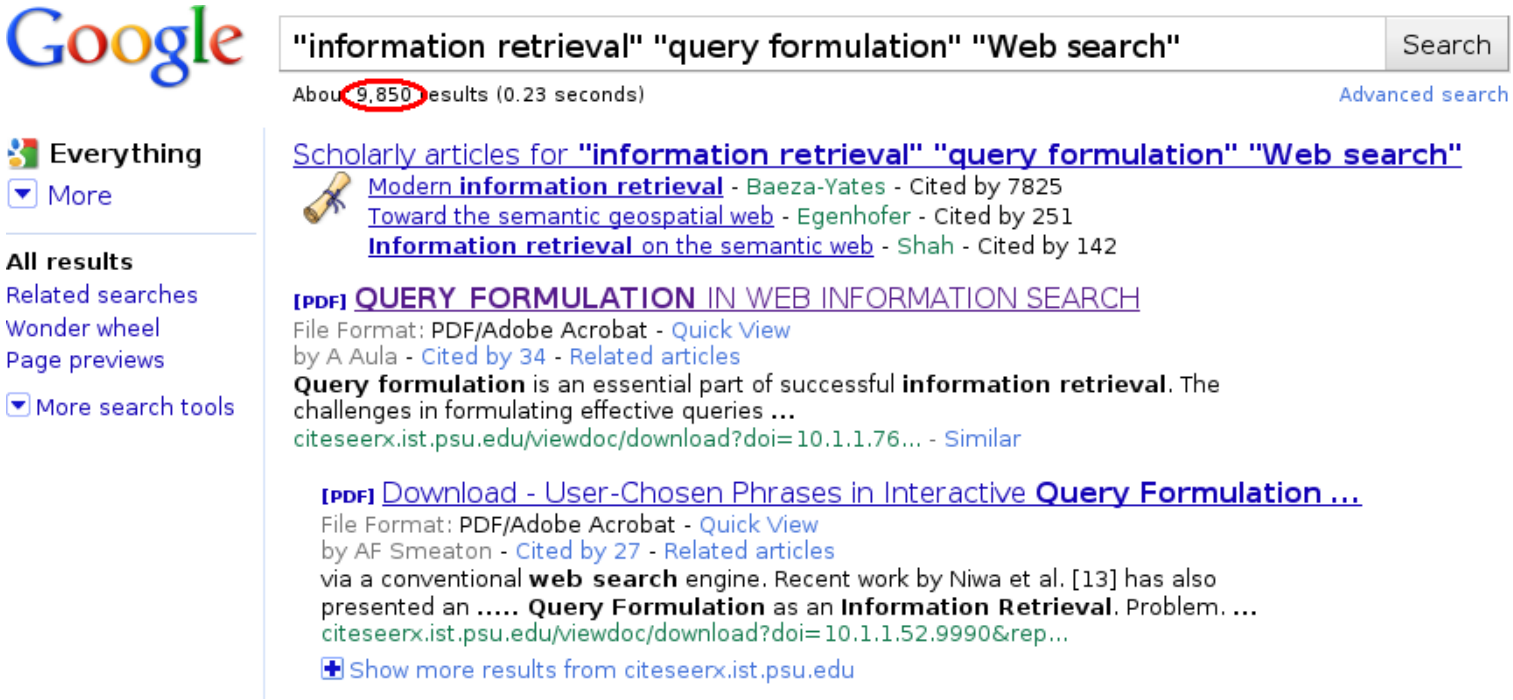
Web Information and Quality Evaluation
Valencia
September 13, 2010

Observing a sample user

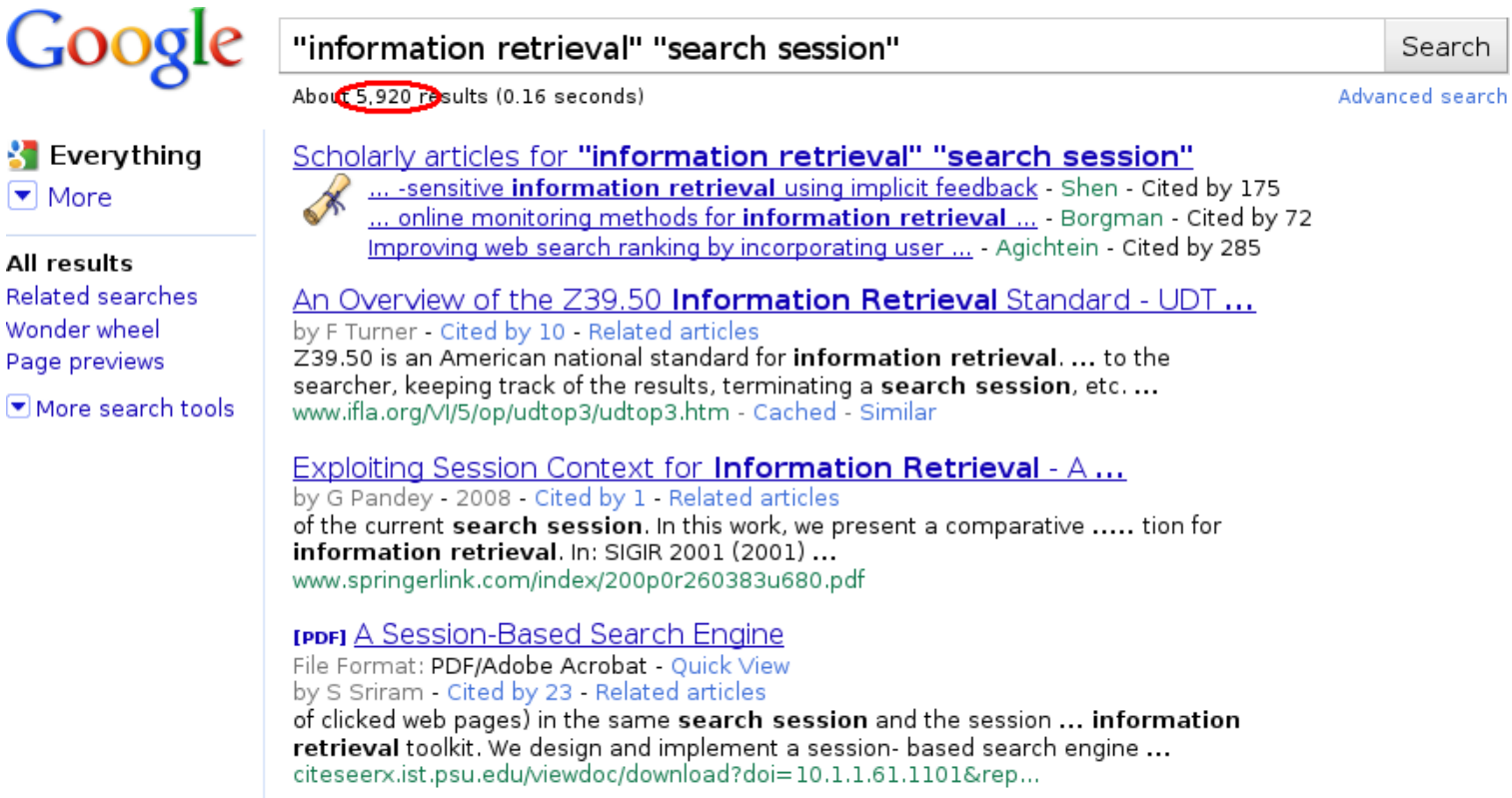


Observing a sample user – query 2

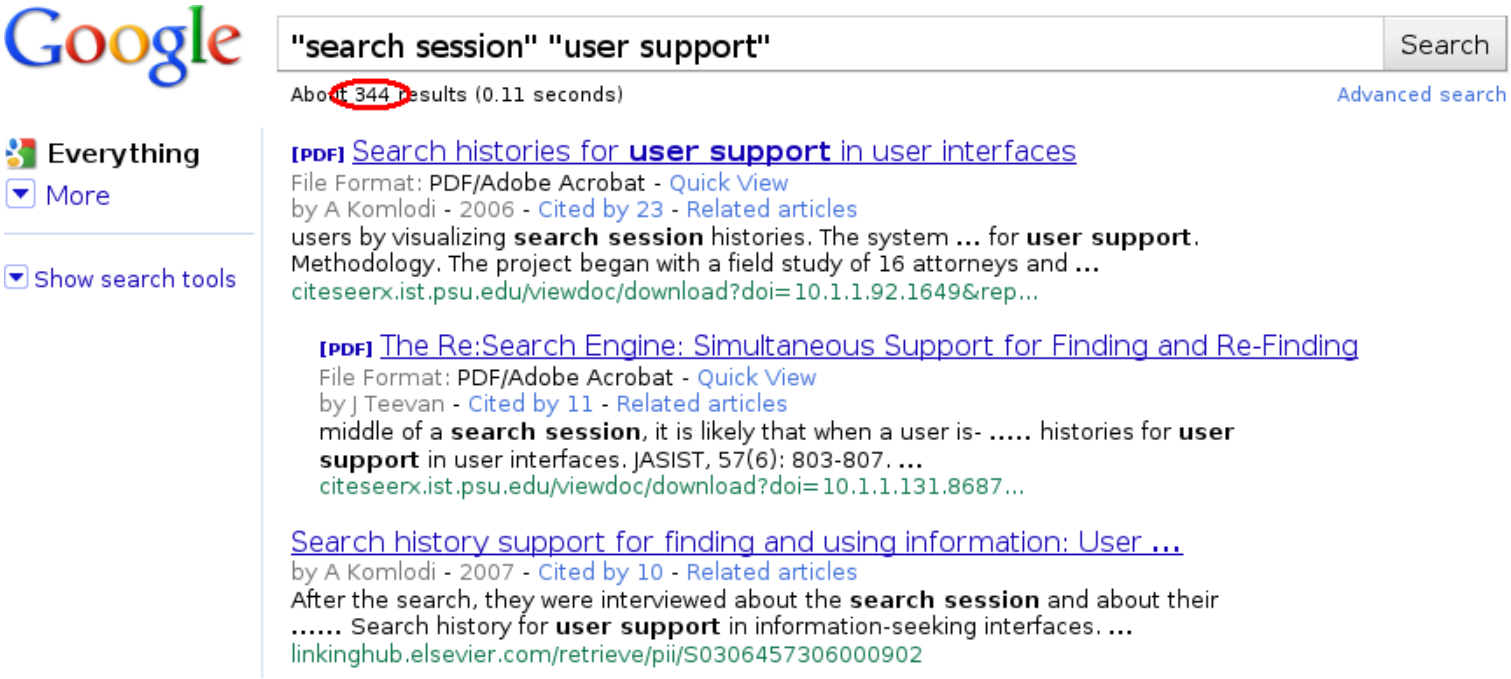
Observing a sample user – query 3



Observing a sample user – query 4



Observing a sample user – query 5



Observing a sample user – query 6

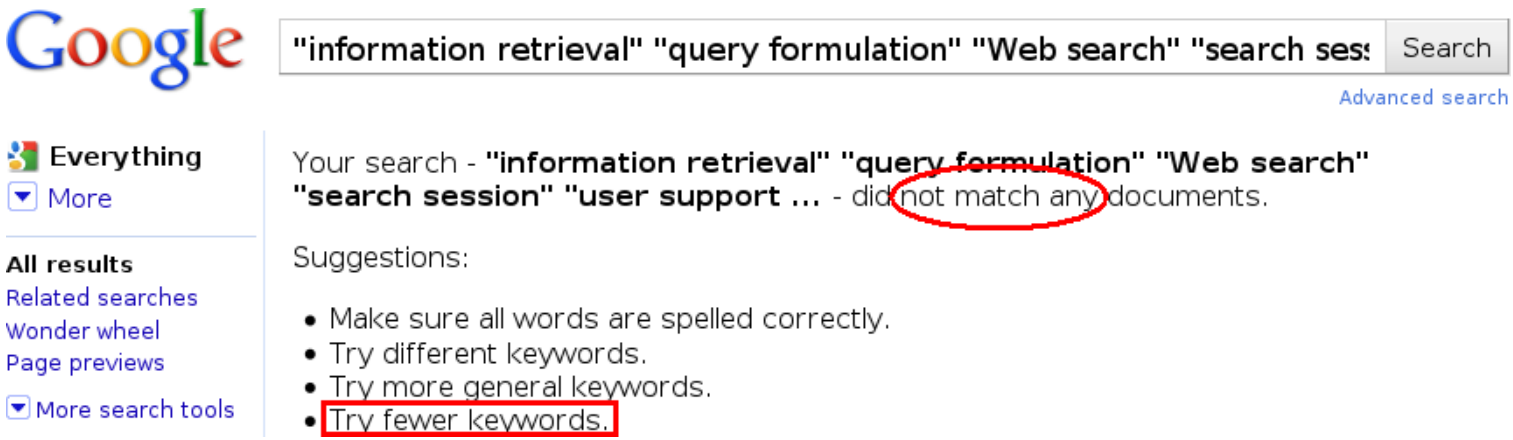
The complete search session

1. “information retrieval”
2. “information retrieval” “query formulation”
3. “information retrieval” “query formulation” “Web search”
4. “information retrieval” “search session”
5. “search session” “user support”
6. “search engine” “cost optimization”

How to get better results from the session?

All keywords as one query?

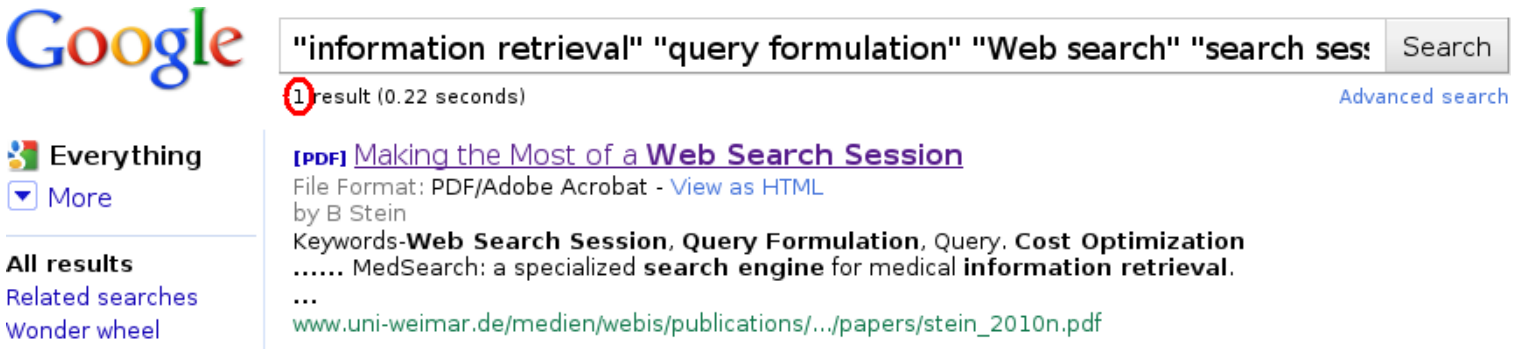
"information retrieval" "query formulation"
"Web search" "search session" "user support"
"search engine" "cost optimization"



How to get better results from the session?

Use **as many keywords as possible**!

"information retrieval" "query formulation"
"Web search" "search session" ~~"user support"~~
"search engine" "cost optimization"



"As many keywords as possible"-Query

- "Best" single query to capture user's articulated information need
- Ideally not too many results $\rightarrow$ user can check complete list
- Potential of improved user experience in search sessions

"As many keywords as possible"-Query

- "Best" single query to capture user's articulated information need
- Ideally not too many results $\rightarrow$ user can check complete list
- Potential of improved user experience in search sessions

"Problem"

Current engines do not suggest "as many keywords as possible"-queries.
So: How to find them?

"As many keywords as possible"-Query

- "Best" single query to capture user's articulated information need
- Ideally not too many results $\rightarrow$ user can check complete list
- Potential of improved user experience in search sessions

"Problem"

Current engines do not suggest "as many keywords as possible"-queries.
So: How to find them?

Idea:

Externally compute them at client site!

As a Formal Problem Statement

Maximum Query

- Given:
  1. A set W of keywords.
  2. A query interface for a Web search engine $\mathcal{S}$.
  3. An upper bound k on the result list length.
- Find a maximum subset $Q \subseteq W$ yielding at most k Web results.

Optimization Problem!

Minimize the number of submitted Web queries to find Q.

As a Formal Problem Statement

Maximum Query

- Given:
  1. A set W of keywords.
  2. A query interface for a Web search engine $\mathcal{S}$.
  3. An upper bound k on the result list length.
- Find a maximum subset $Q \subseteq W$ yielding at most k Web results.

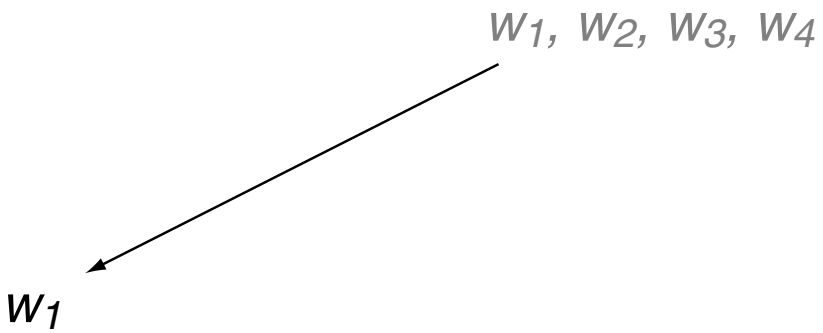
Optimization Problem!

Minimize the number of submitted Web queries to find Q.

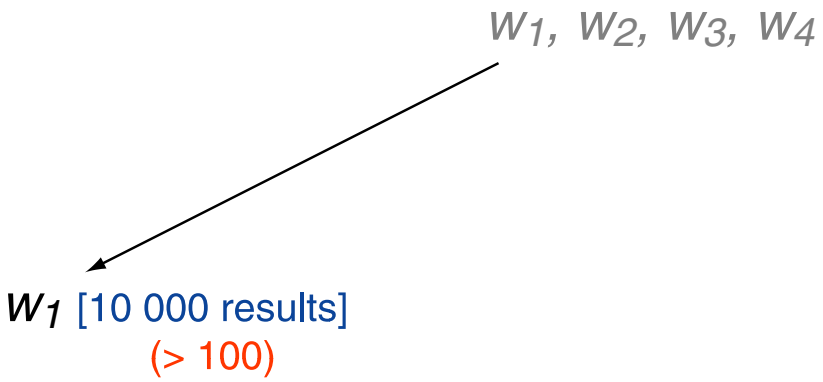
Simple Attack: Depth-First Search

w_1, w_2, w_3, w_4

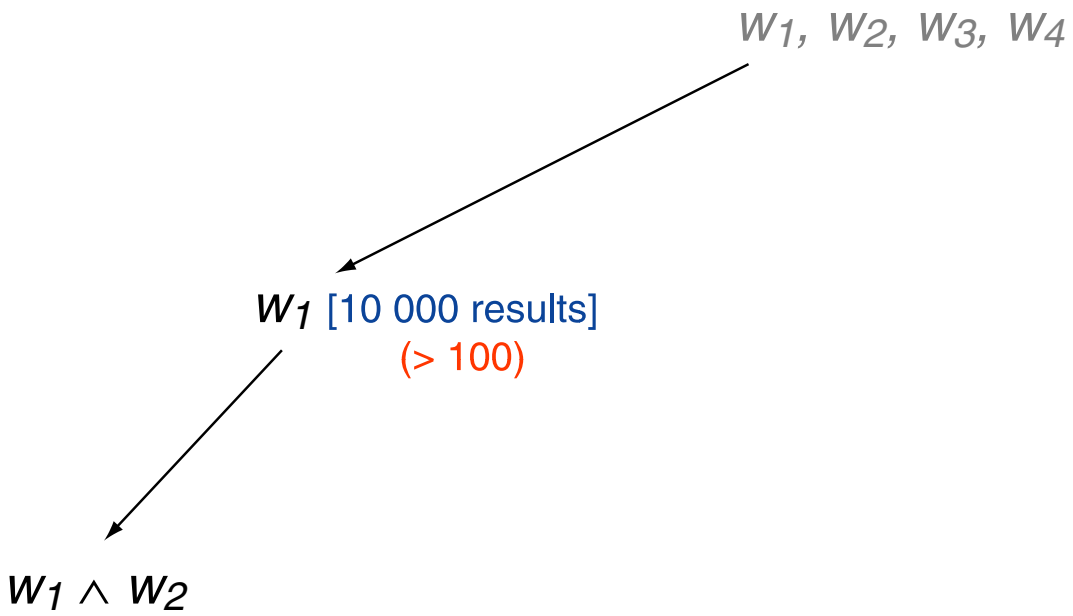
Simple Attack: Depth-First Search



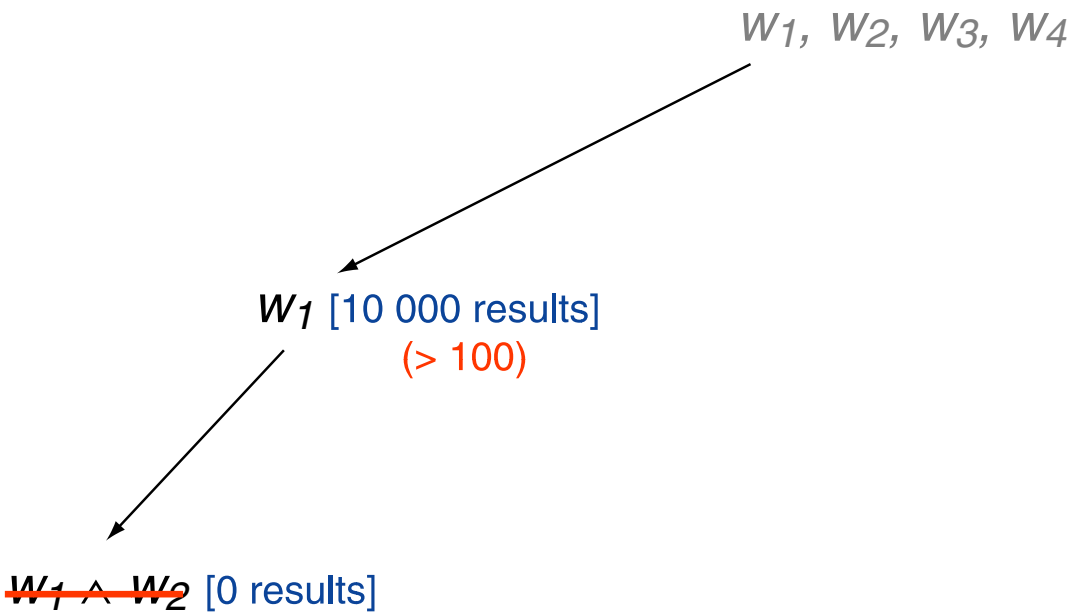
Simple Attack: Depth-First Search



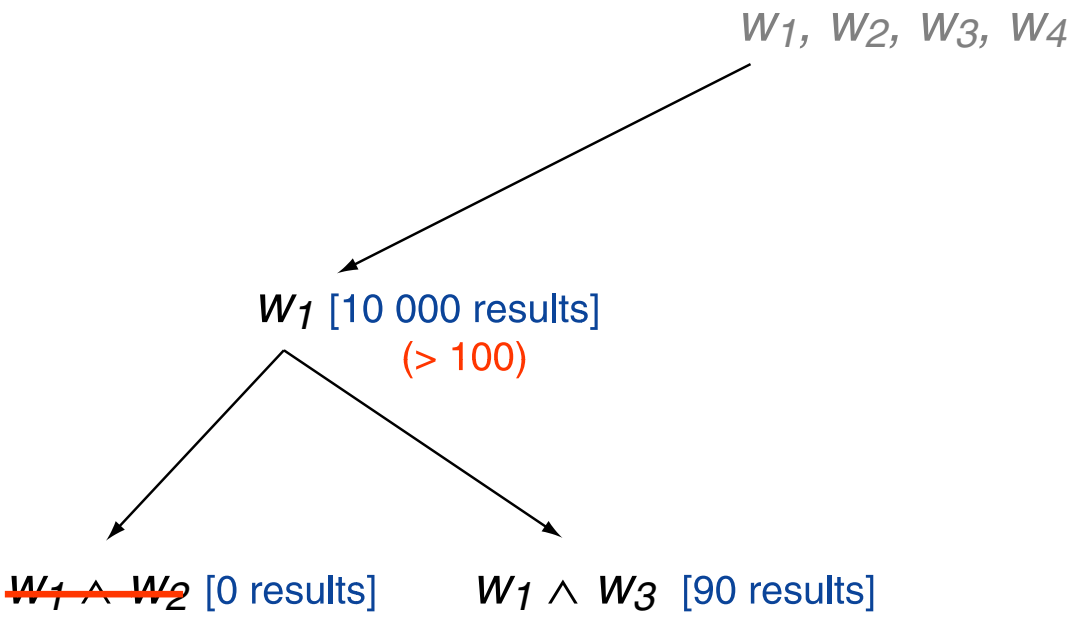
Simple Attack: Depth-First Search



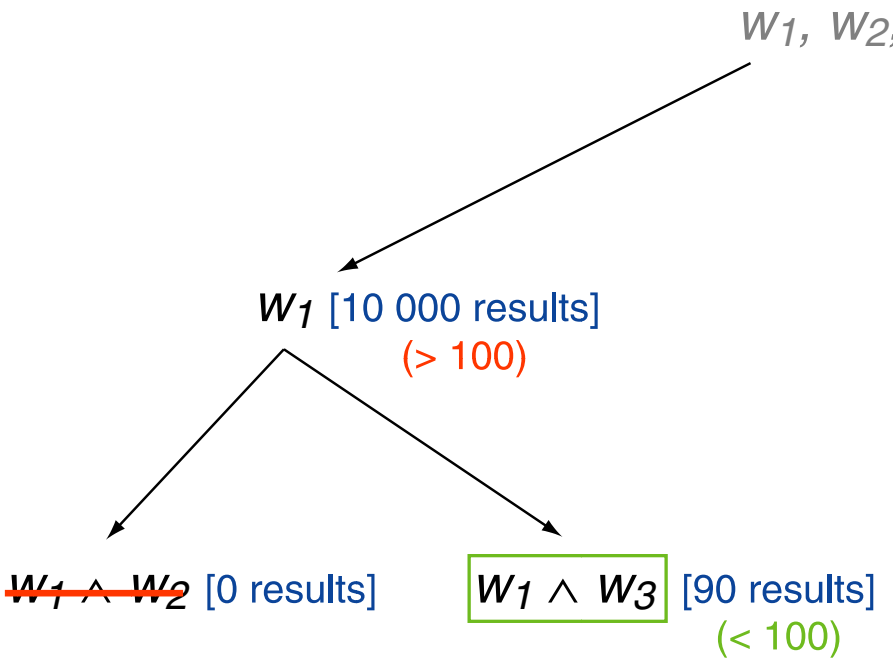
Simple Attack: Depth-First Search



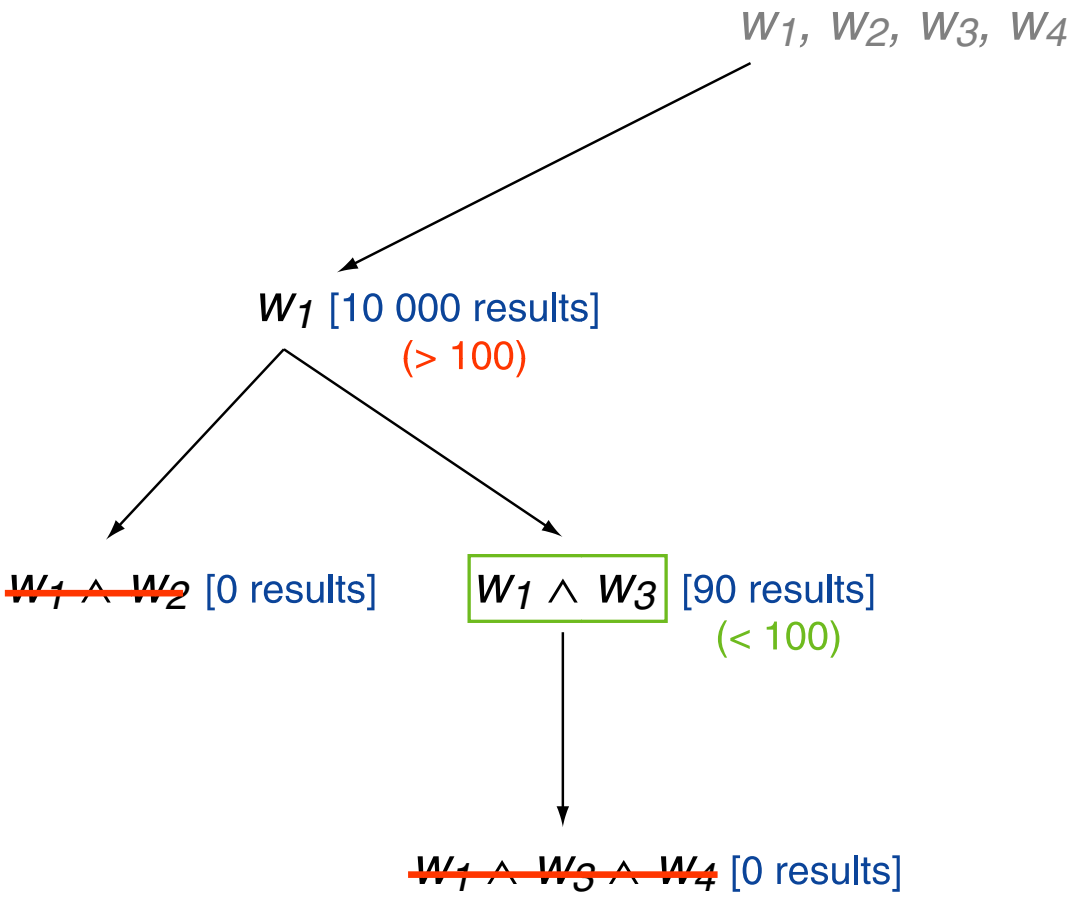
Simple Attack: Depth-First Search



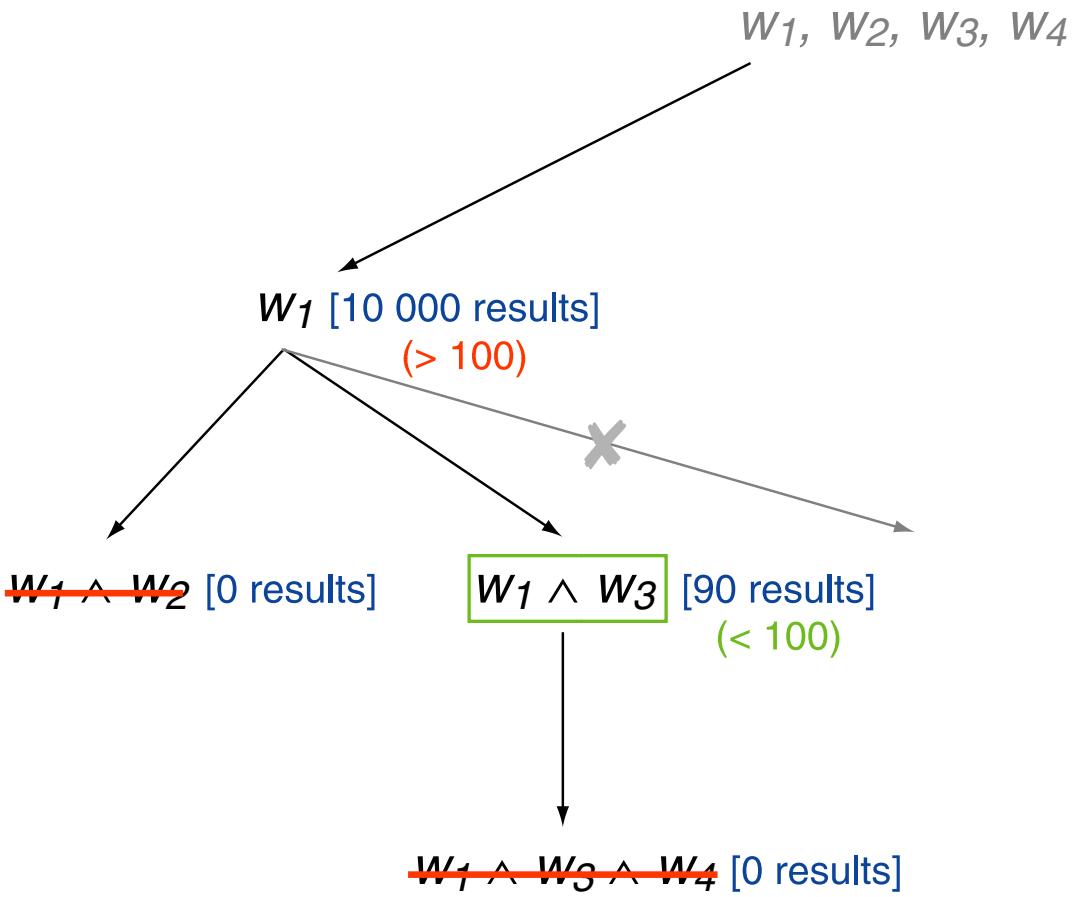
Simple Attack: Depth-First Search



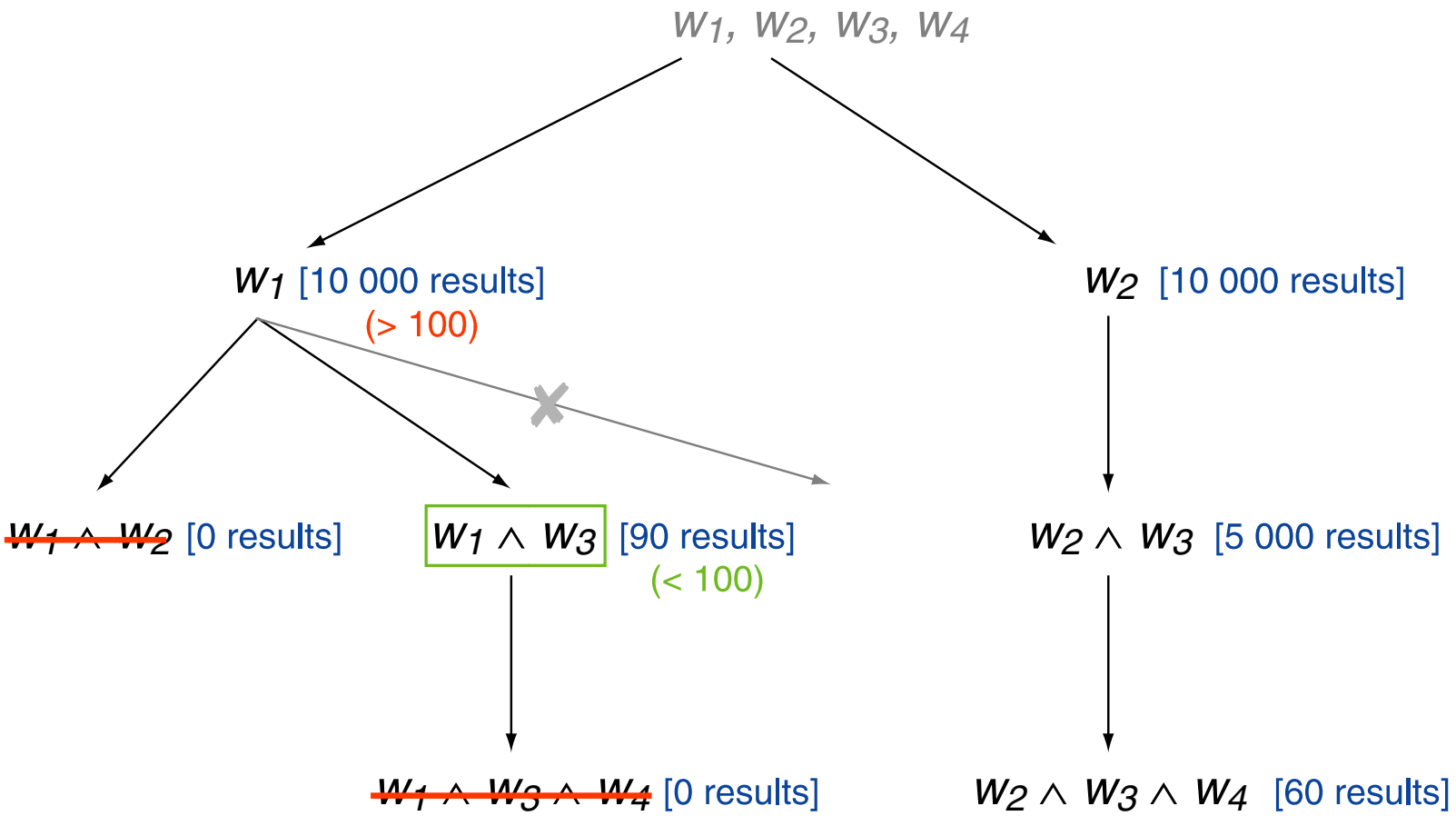
Simple Attack: Depth-First Search



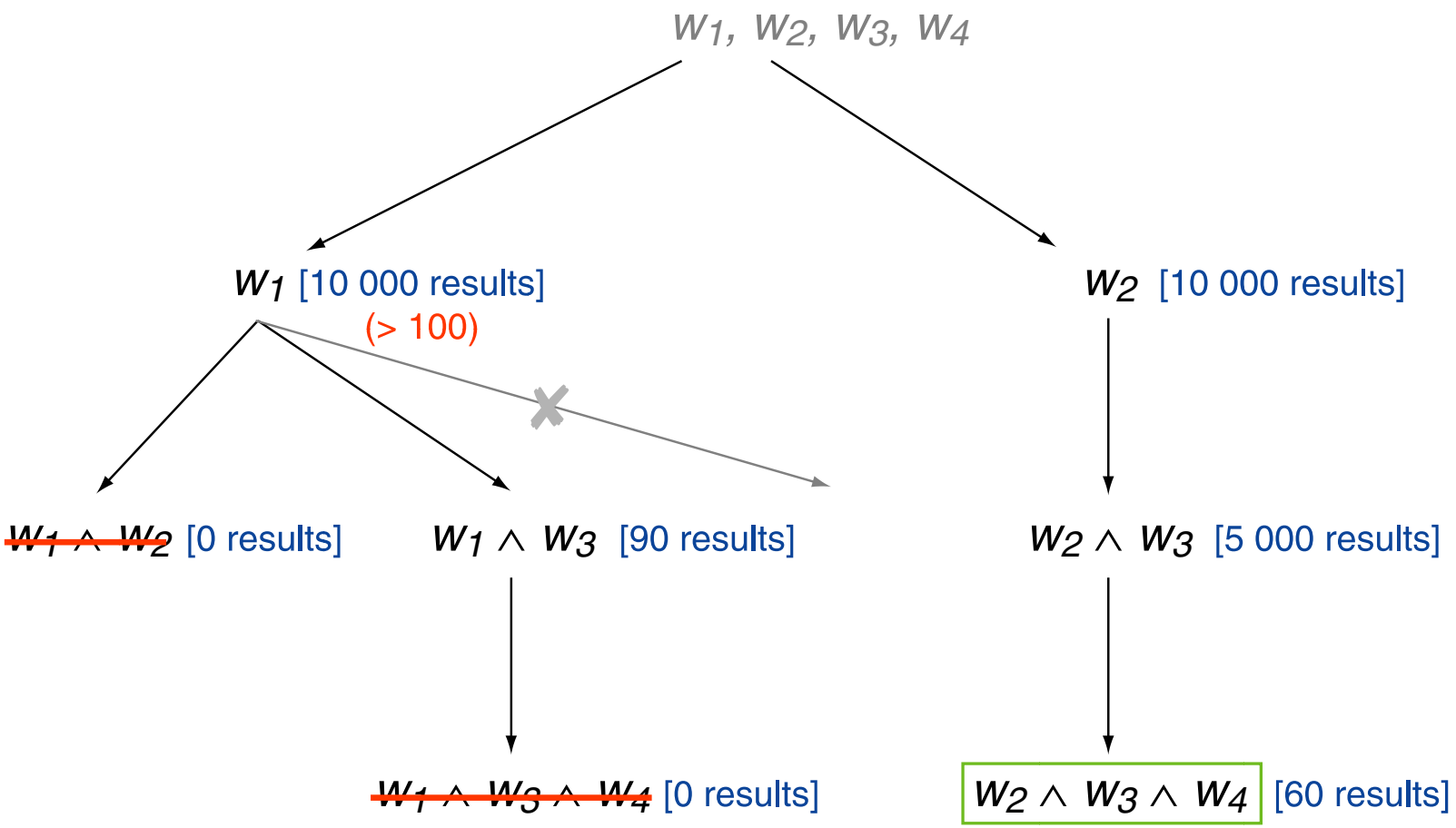
Simple Attack: Depth-First Search



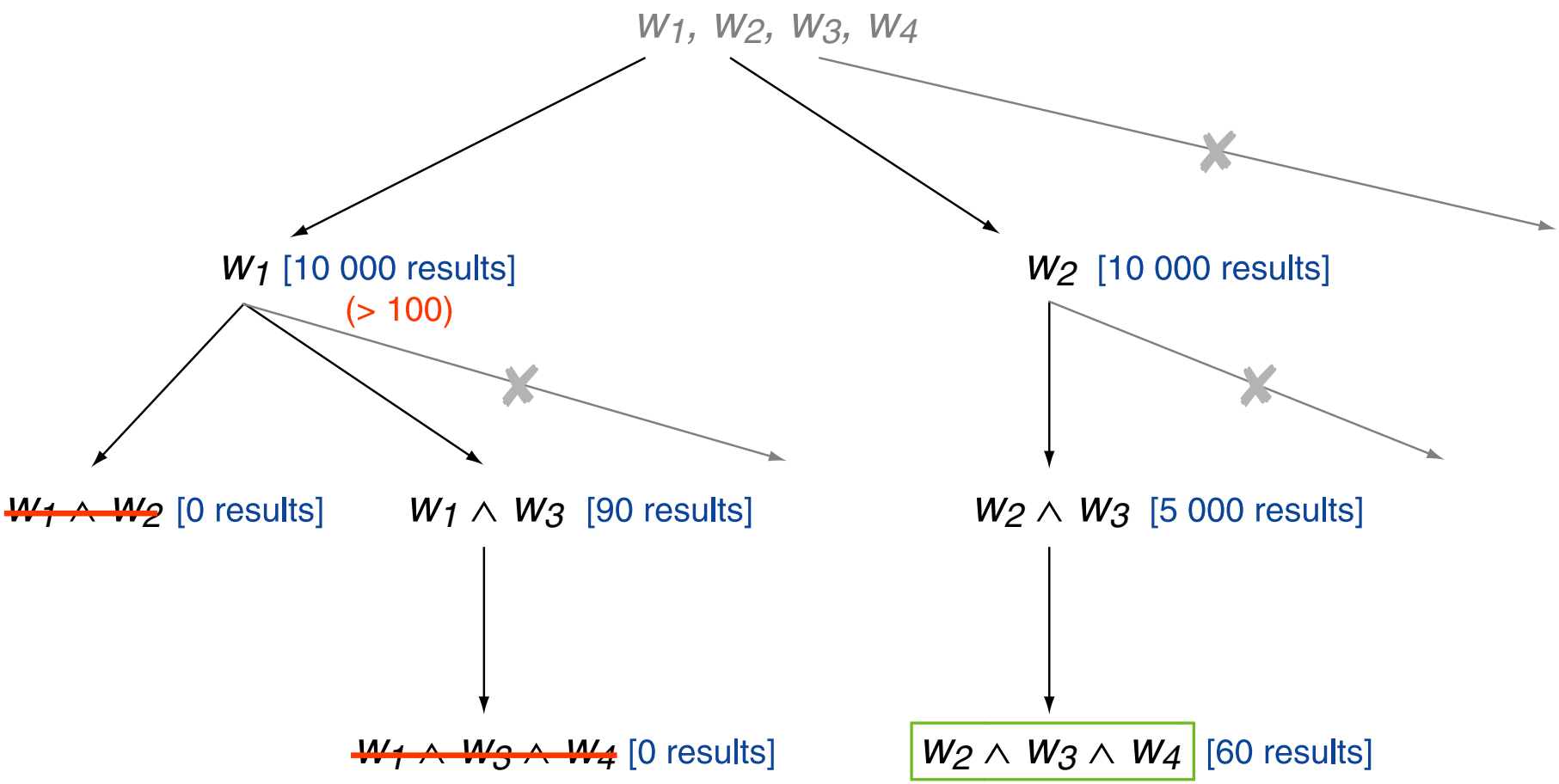
Simple Attack: Depth-First Search



Simple Attack: Depth-First Search



Simple Attack: Depth-First Search



Baseline's Analysis

Major Drawback

All intermediate queries submitted to the search engine. $\Rightarrow$ Bad run time!

Baseline's Analysis

Major Drawback

All intermediate queries submitted to the search engine. $\Rightarrow$ Bad run time!

Solution Idea:

Estimate a query candidate's Web count before submission.

Co-occurrences

Google	"information retrieval"	2,500,000 results
Google	"information retrieval" "query formulation"	25,000 results

gives yield-factor: $\gamma(\mathrm{IR} + \mathrm{QF}) = 0.01$

Co-occurrence based Web Count Estimation

Estimate: “information retrieval” “query formulation” “Web search”

Co-occurrence based Web Count Estimation

Estimate: "information retrieval" "query formulation" "Web search"

Known numbers:

$25,000$	results for "information retrieval" "query formulation"
0.06	$\gamma(\mathrm{IR} + \mathrm{WS})$
0.16	$\gamma(\mathrm{QF} + \mathrm{WS})$

Co-occurrence based Web Count Estimation

Estimate: "information retrieval" "query formulation" "Web search"

Known numbers:

$25,000$	results for "information retrieval" "query formulation"
0.06	$\gamma(\mathrm{IR} + \mathrm{WS})$
0.16	$\gamma(\mathrm{QF} + \mathrm{WS})$

Our scheme:

$$25,000 \cdot \mathrm{avg}(0.06, 0.16) \quad = \quad 2,750 \text{ results}$$

Co-occurrence based Web Count Estimation

Estimate: "information retrieval" "query formulation" "Web search"

Known numbers:

$25,000$	results for "information retrieval" "query formulation"
0.06	$\gamma(\mathrm{IR} + \mathrm{WS})$
0.16	$\gamma(\mathrm{QF} + \mathrm{WS})$

Our scheme:

$$25,000 \cdot \mathrm{avg}(0.06, 0.16) = 2,750 \text{ results}$$

Control: Google $10,000$ results

Co-occurrence based Web Count Estimation

Estimate: "information retrieval" "query formulation" "Web search"

Known numbers:

$25,000$	results for "information retrieval" "query formulation"
0.06	$\gamma(\text{IR} + \text{WS})$
0.16	$\gamma(\text{QF} + \text{WS})$

Our scheme:

$$25,000 \cdot \text{avg}(0.06, 0.16) = 2,750 \text{ results}$$

Control: Google $10,000$ results

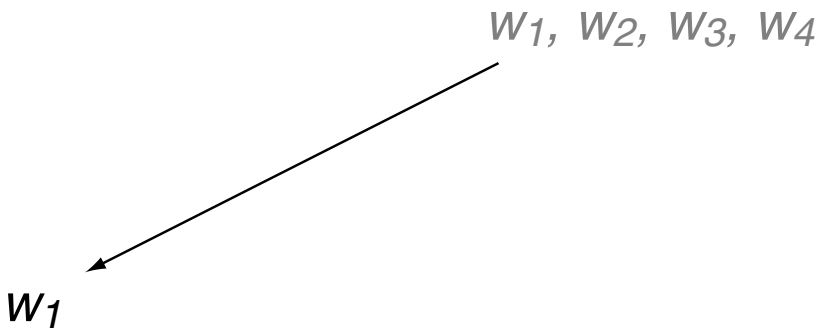
Observation:

Our co-occurrence based estimations usually underestimate real Web count.

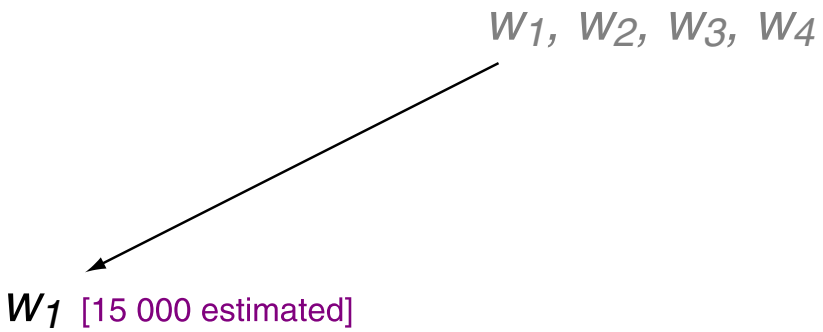
"Informed" Depth-First Search

w_1, w_2, w_3, w_4

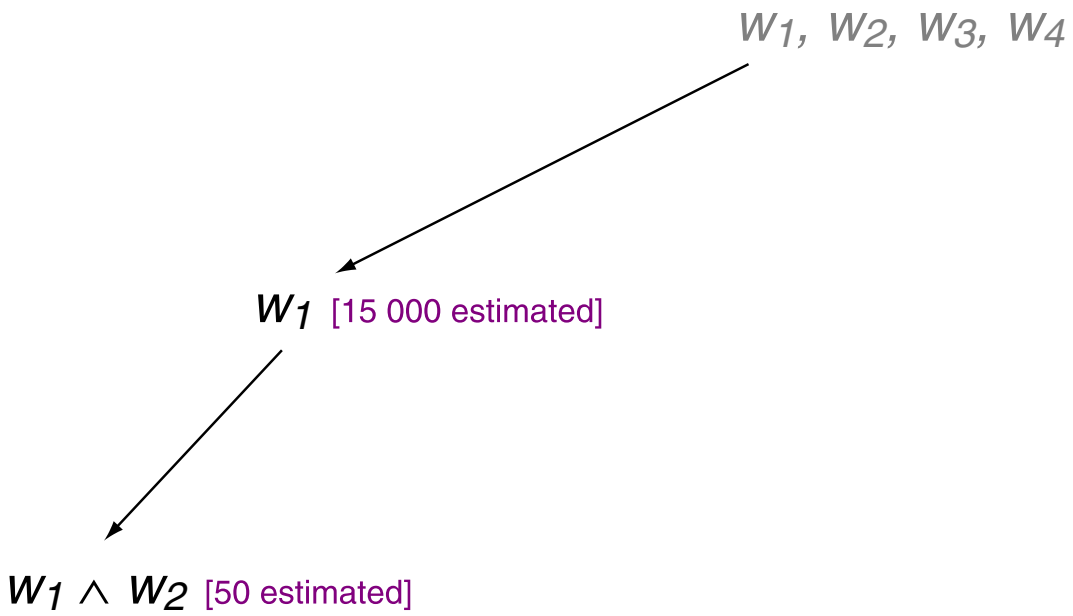
"Informed" Depth-First Search



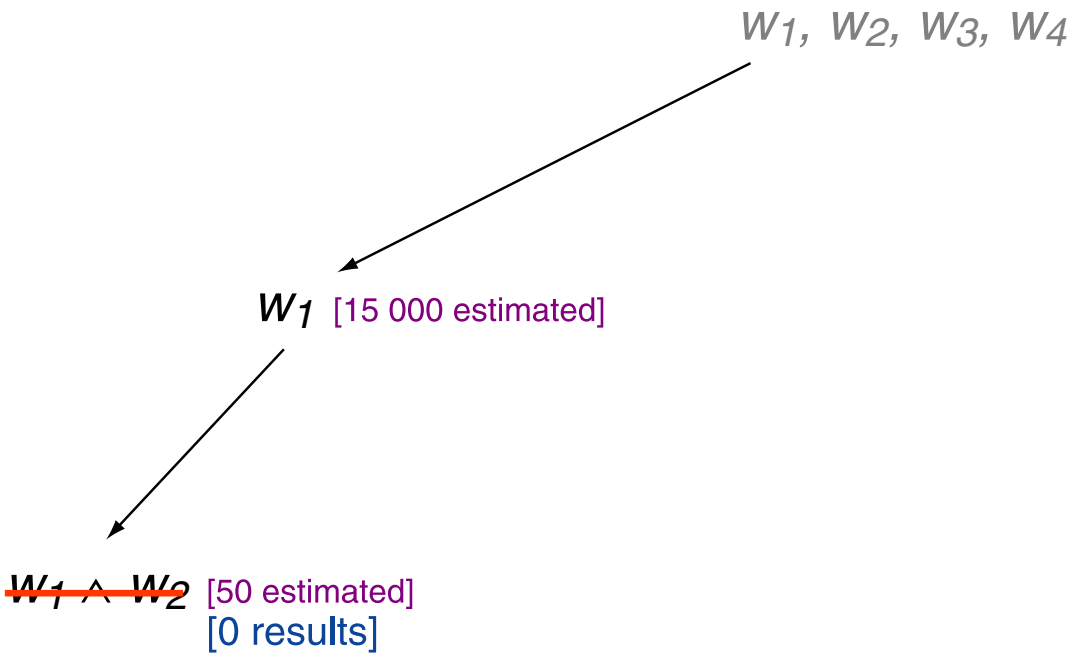
"Informed" Depth-First Search



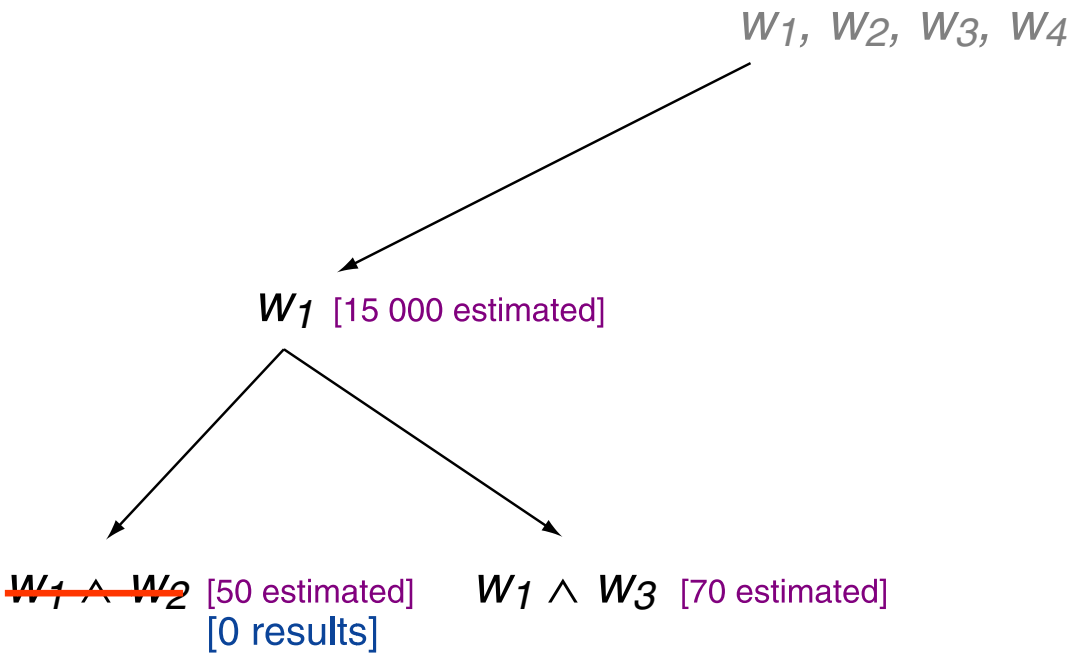
"Informed" Depth-First Search



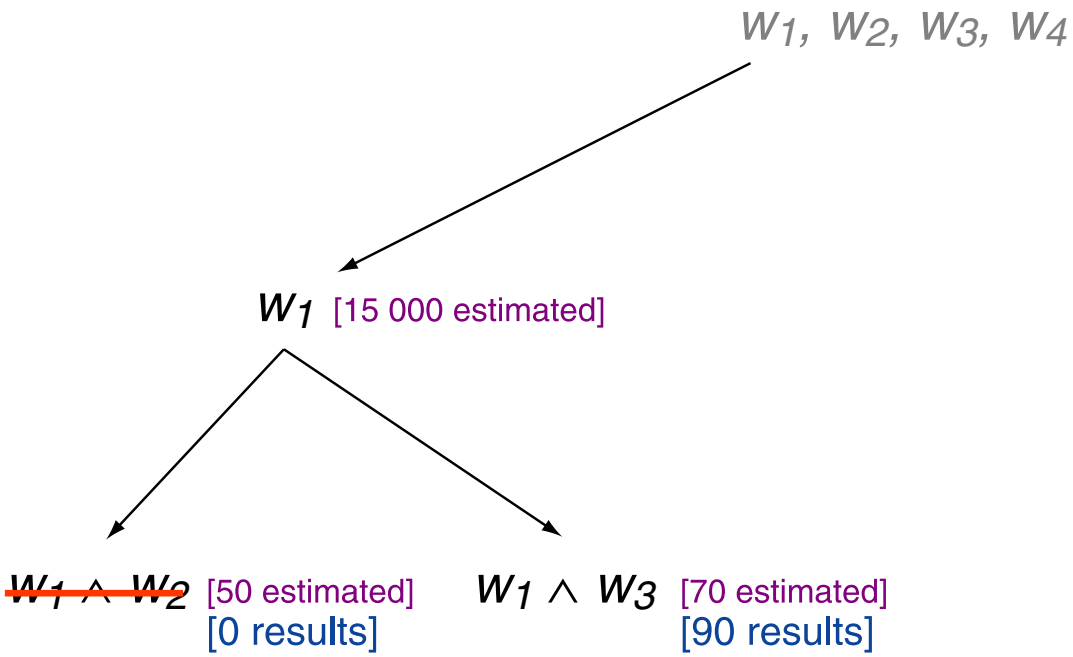
"Informed" Depth-First Search



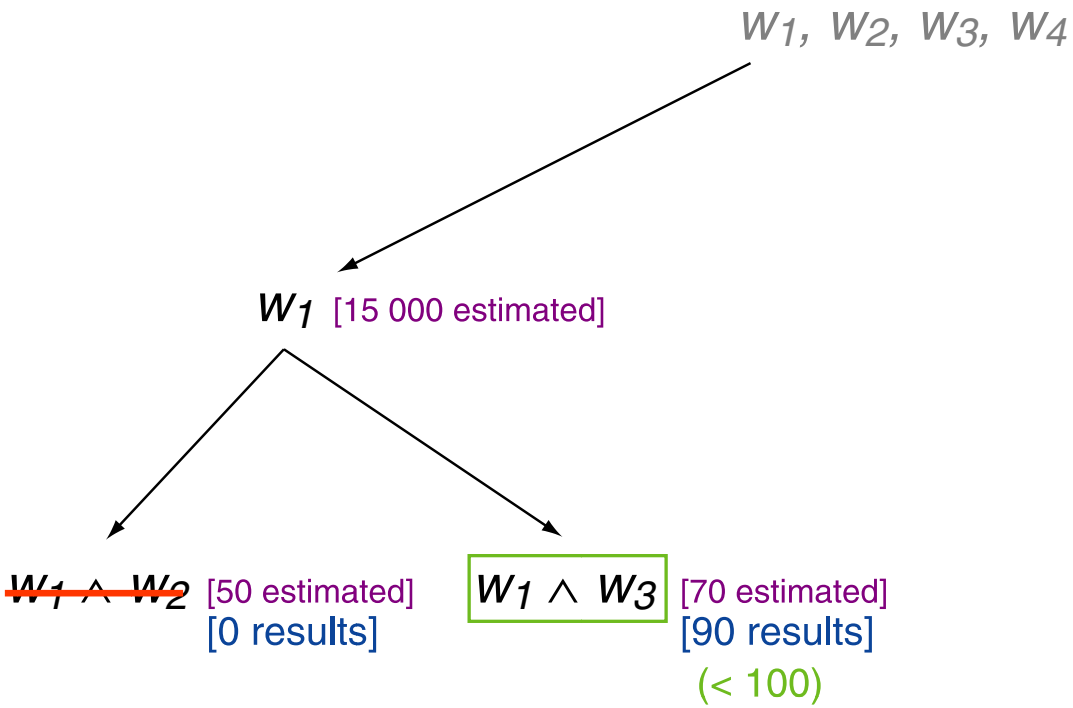
"Informed" Depth-First Search



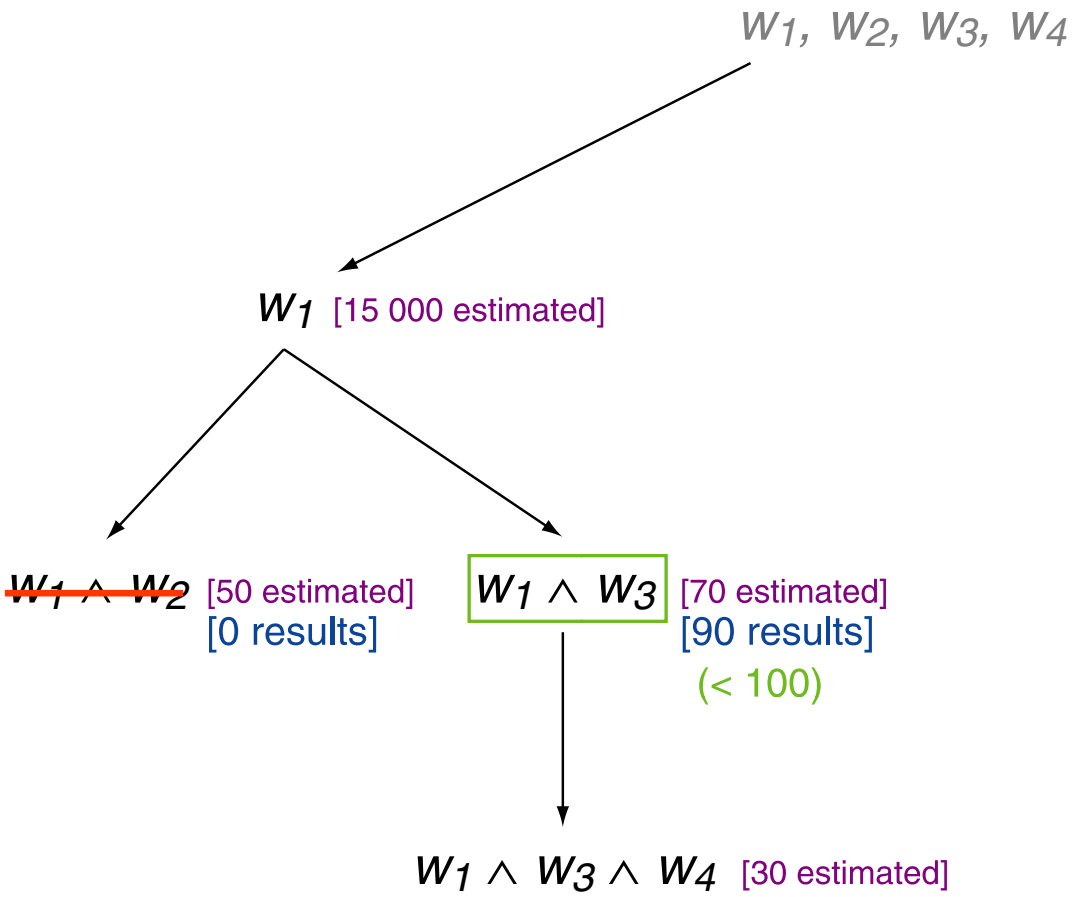
"Informed" Depth-First Search



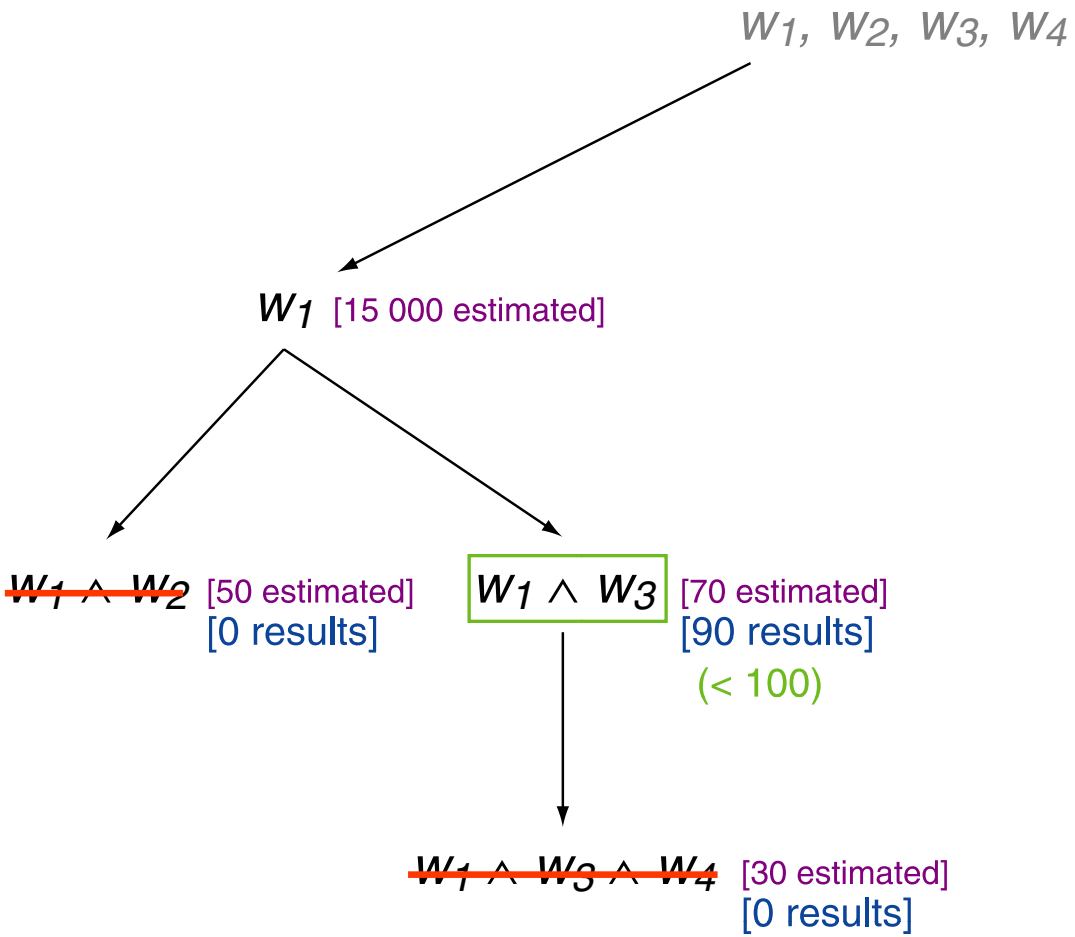
"Informed" Depth-First Search



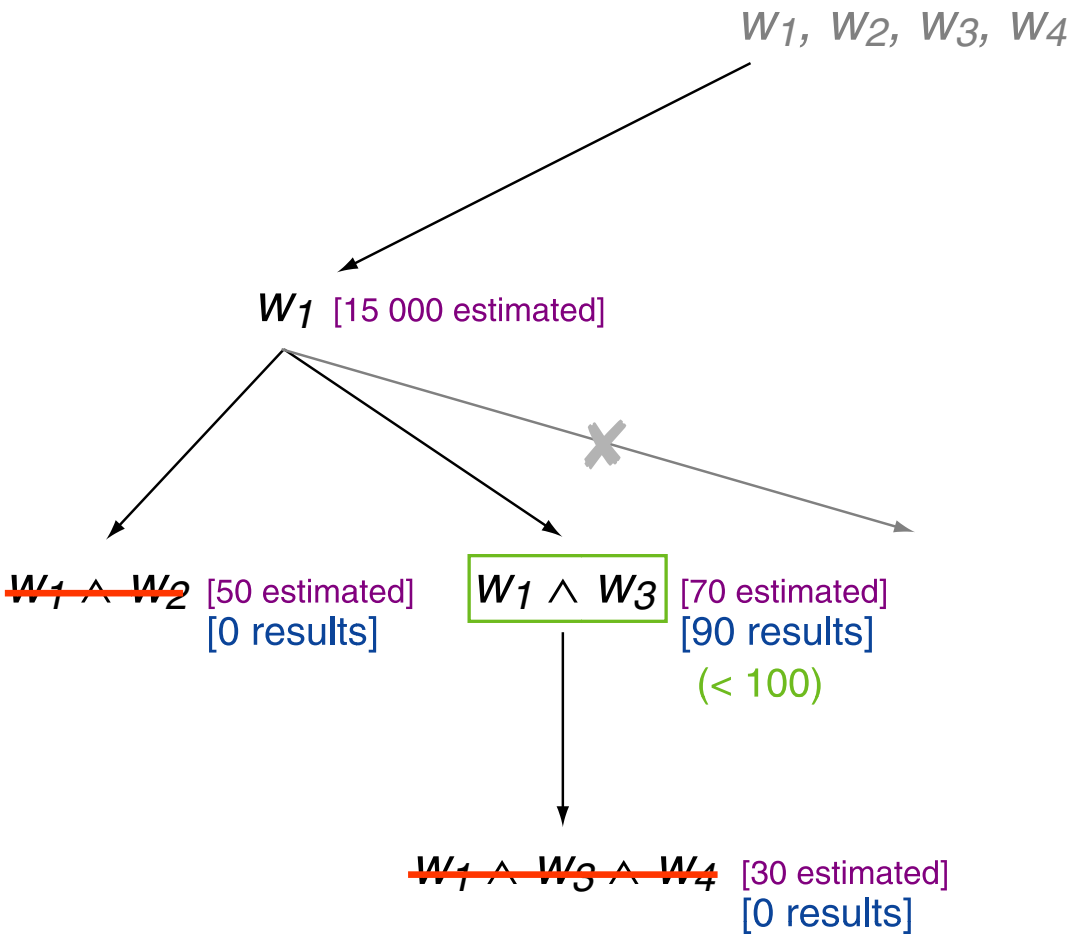
"Informed" Depth-First Search



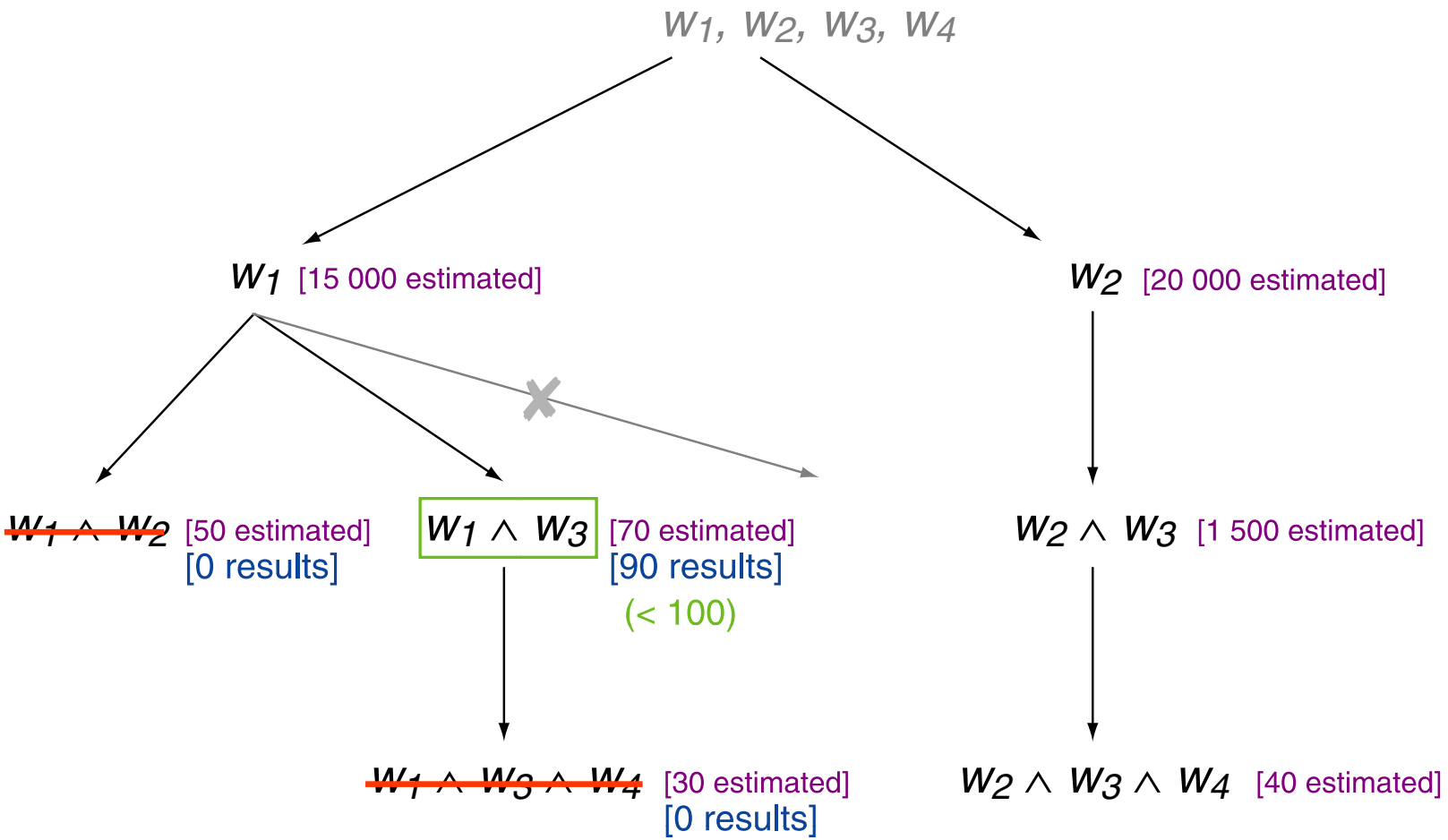
"Informed" Depth-First Search



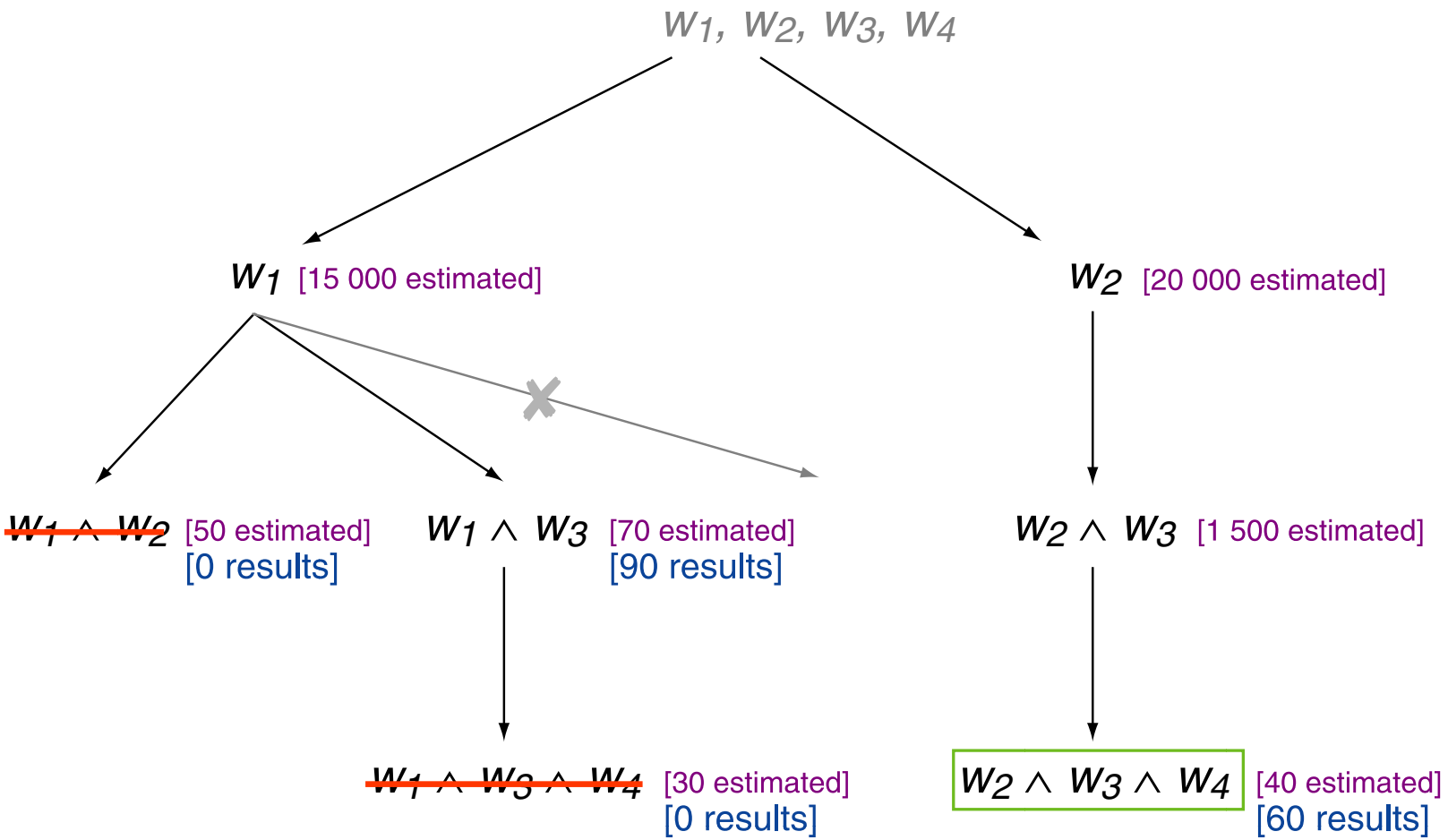
"Informed" Depth-First Search



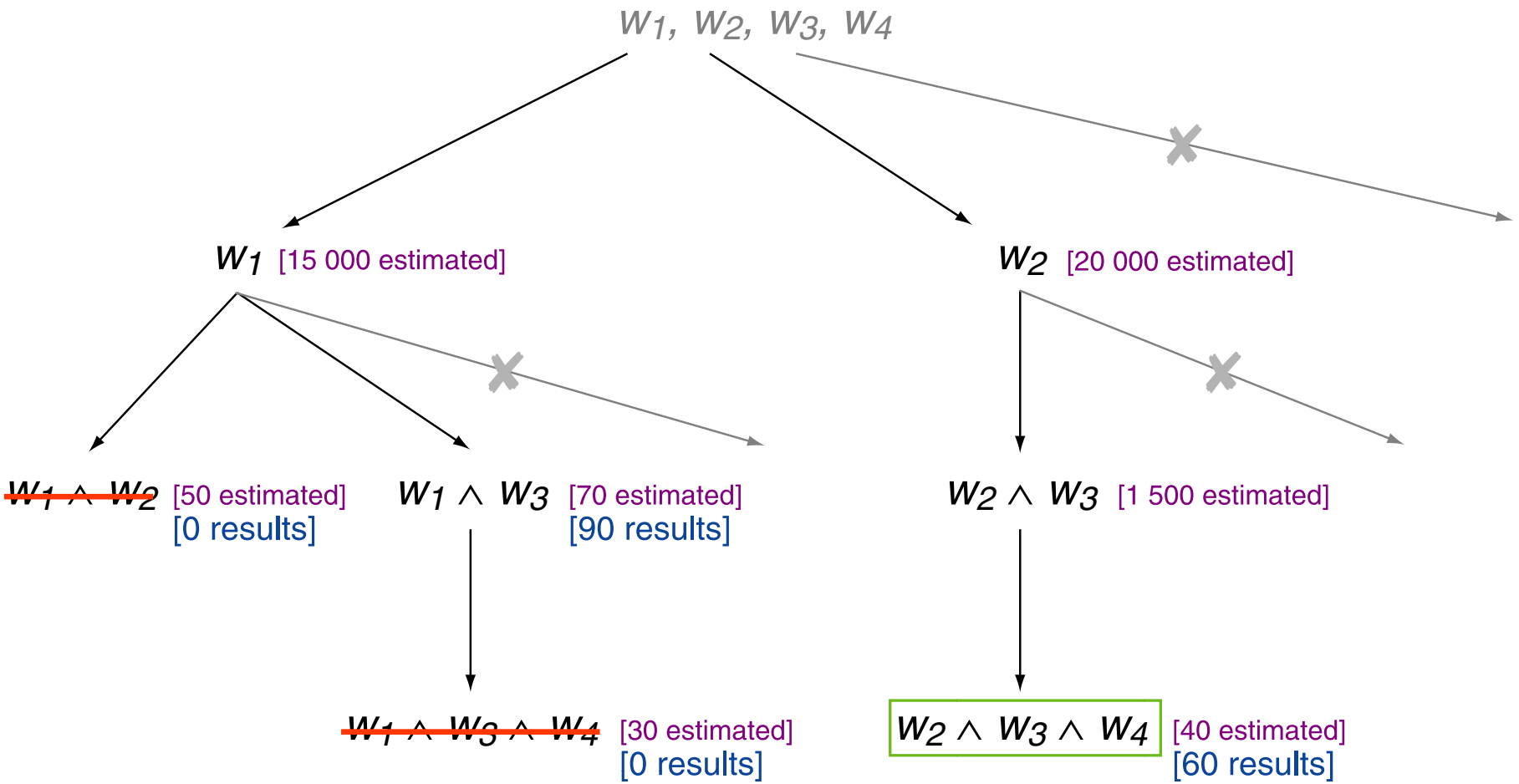
"Informed" Depth-First Search



"Informed" Depth-First Search



"Informed" Depth-First Search



Experimental Setup

Corpus:

Extracted sessions (≥ 2 queries) from AOL log with two methods:

- 10 minute time cut-off,
- Geometric method [Gayo-Avello 2009].

Removed stopwords.
For $i \in \{3, \ldots, 15\}$ sampled 1000 i-keyword-sessions from each method.

System:

Bing API as search engine.
Set $k = 100$.
Measured number of submitted Web queries.

Experimental Setup

Corpus:

Extracted sessions (≥ 2 queries) from AOL log with two methods:

- 10 minute time cut-off,
- Geometric method [Gayo-Avello 2009].

Removed stopwords.
For $i \in \{3, \ldots, 15\}$ sampled 1000 i-keyword-sessions from each method.

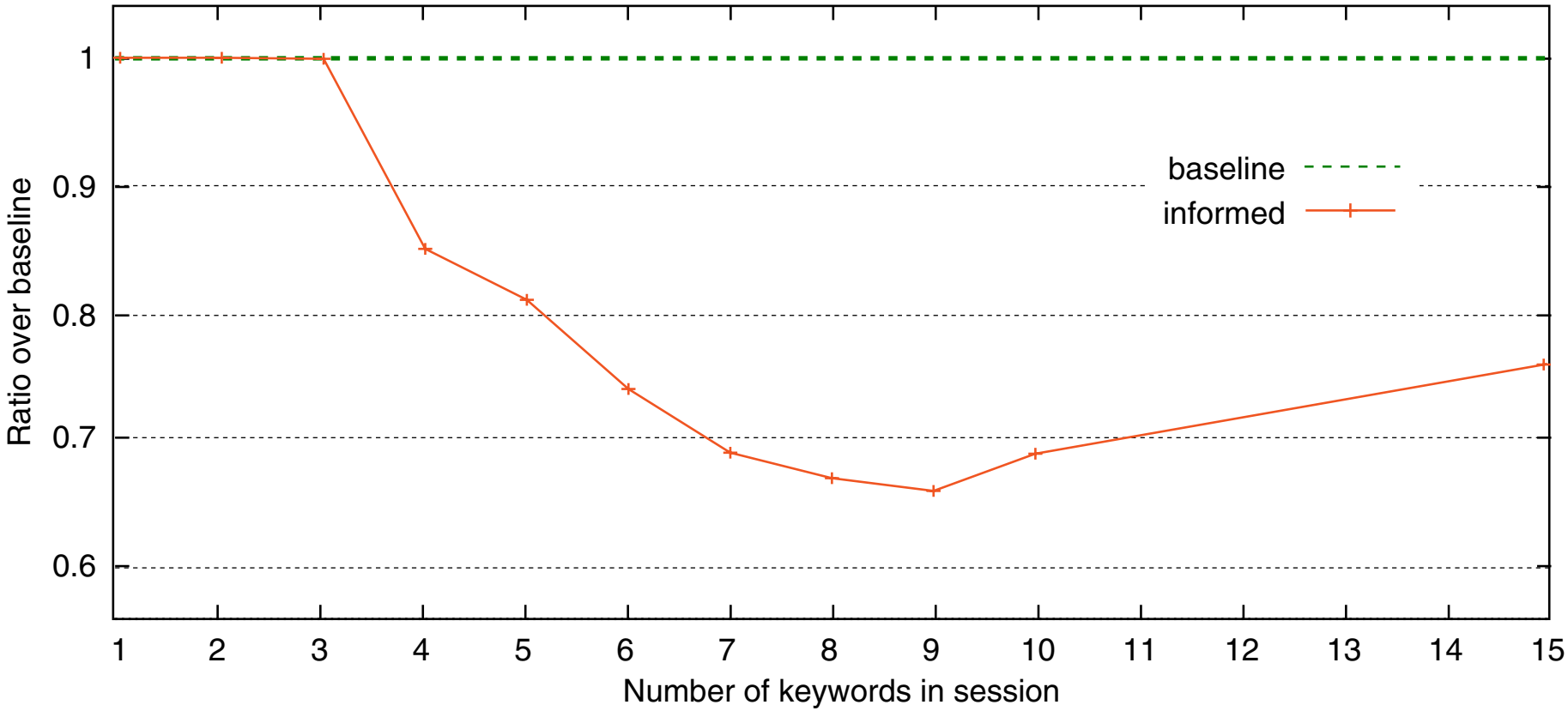
System:

Bing API as search engine.
Set $k = 100$.
Measured number of submitted Web queries.

Experimental Results

Number of keywords	5	10	15
No maximum query possible	1 913	1 599	953
Maximum query found	87	401	1 047
Avg. queries submitted informed	10.90	48.15	394.41
Avg. queries submitted baseline	13.38	70.29	516.46
Avg. Web query time (ms)	359.05	367.94	336.45
Avg. size maximum query informed	3.09	7.47	11.93
Avg. size maximum query baseline	3.19	7.71	12.34

Experimental Results



Almost the end: The take-away messages!

What we have done

Results

- Maximum Query
- Query Cover (in the paper)
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines

Open Problems

- Improved heuristics
- Co-occurrence source
- User study

What we have (not) done

Results

- Maximum Query
- Query Cover (in the paper)
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines

Open Problems

- Improved heuristics
- Co-occurrence source
- User study

What we have (not) done

Results

- Maximum Query
- Query Cover (in the paper)
- External (client site) algorithms
- Co-occurrence based heuristics
- Heuristics outperform baselines

Open Problems

- Improved heuristics
- Co-occurrence source
- User study

Thank you

☺