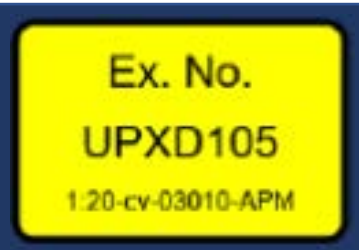
Rebuttal Testimony of Professor Douglas W. Oard

U.S. et al. v. Google LLC
United States District Court for the District of Columbia

November 15, 2023

REDACTED FOR PUBLIC FILING



My Assignment

To provide my expert opinion of the analysis and opinions offered by Google's expert, Prof. Edward A. Fox, in his June 3, 2022 expert report (the "Fox Report").

REDACTED FOR PUBLIC FILING

My Overall Conclusion

Prof. Fox substantially understates the beneficial effects of user-side data on search quality.

REDACTED FOR PUBLIC FILING

Prof. Fox's Assignment

Prof. Fox states he was asked by Google counsel to:

"test the extent to which Google's search quality is affected by the volume of user interaction data available to train its ranking algorithms"

REDACTED FOR PUBLIC FILING

Prof. Fox's Conclusions

Vast majority of Google-Microsoft search quality gap must be explained by factors other than volume of user interaction data

A company as efficient as Google could have search quality similar to Google even at Microsoft's scale

A company as efficient as Google but with Microsoft's scale would not meaningfully benefit from increase in user interaction data

There are diminishing returns to search quality from an increase in the quantity of user interaction data

REDACTED FOR PUBLIC FILING

Prof. Fox's Conclusions

Vast majority of Google-Microsoft search quality gap must be explained by factors other than volume of user interaction data

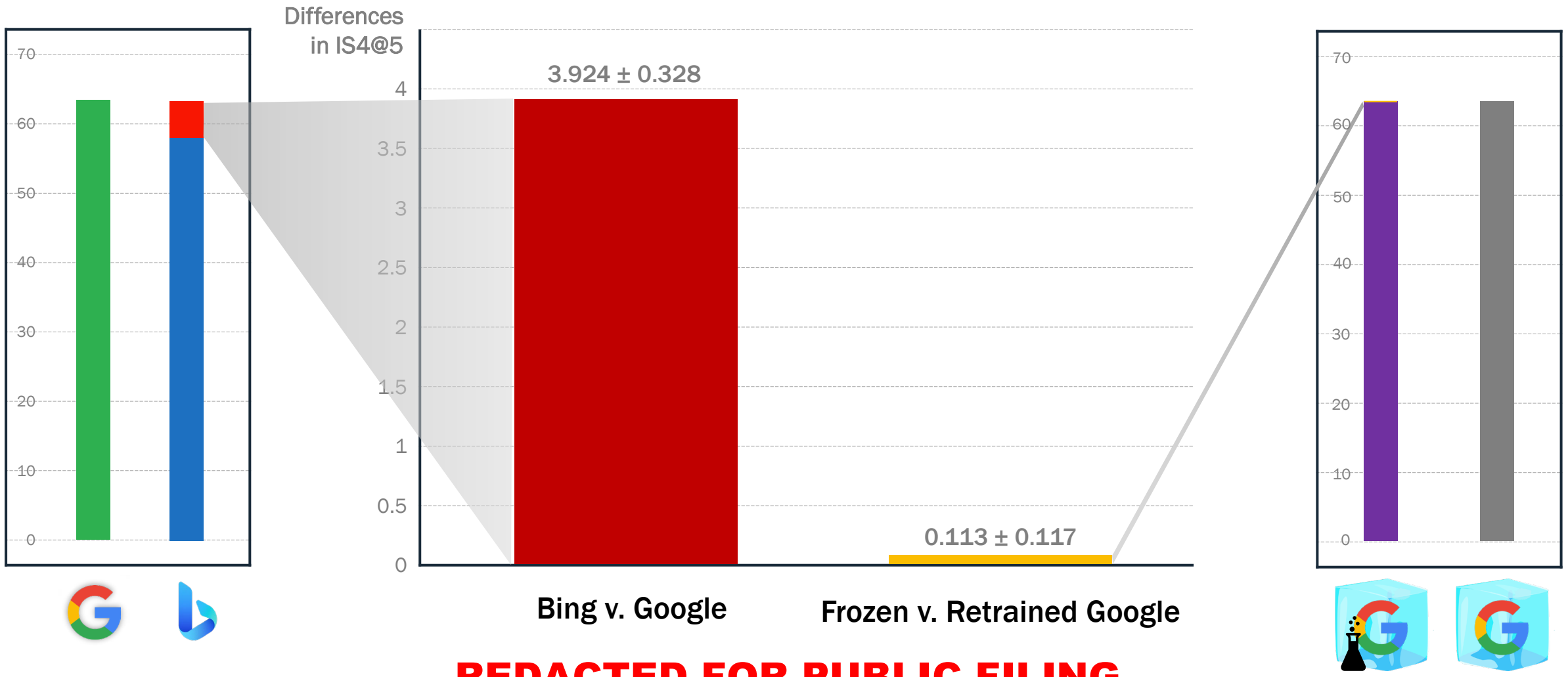
A company as efficient as Google could have search quality similar to Google even at Microsoft's scale

A company as efficient as Google but with Microsoft's scale would not meaningfully benefit from increase in user interaction data

There are diminishing returns to search quality from an increase in the quantity of user interaction data

REDACTED FOR PUBLIC FILING

The Basis for Prof. Fox's Central Conclusion



REDACTED FOR PUBLIC FILING

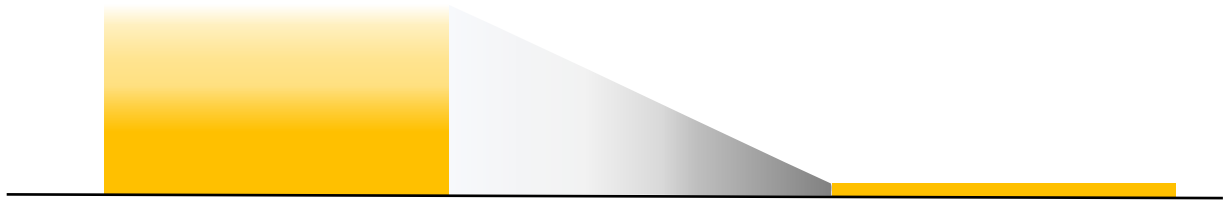
My Response to Prof. Fox's Central Conclusions

Prof. Fox's conclusions are unsupported because of:

- Unmeasured benefits of user-side data in this experiment;
- Measurement errors in the "quality gap"; and
- Important benefits of user-side data that this experiment cannot measure

REDACTED FOR PUBLIC FILING

Many Components Not Retrained



Measurable effect of user-side data

Measured effect of user-side data

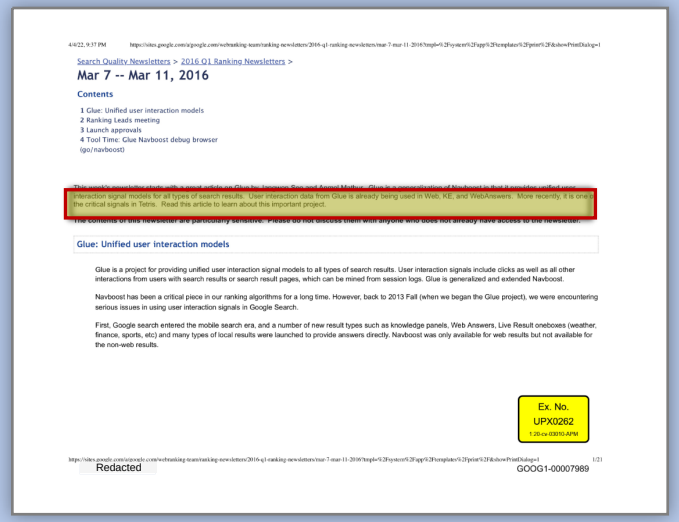
Google only retrained 6 components, chosen based on their expected effect on web ranking (i.e., 10 blue links)

Components were not chosen based on their effect on:

- Indexing
- Spelling Correction
- Search features like images, video…
- Search advertising
- Whole-page ranking

REDACTED FOR PUBLIC FILING

Glue's Importance to Whole-Page Ranking



4/4/22, 9:37 PM https://sites.google.com/a/google.com/webranking-team/ranking-newsletters/2016-q1-ranking-newsletters/mar-7-mar-11-2016?tmpl=%2Fsystem%2Fapp%2Ftemplates%2Fprint%2Fshow&showPrintDialog=1

Search Quality Newsletters > 2016 Q1 Ranking Newsletters >

Mar 7 -- Mar 11, 2016

Contents

1 Glue: Unified user interaction models
2 Ranking Leads meeting
3 Launch approvals
4 Tool Time: Glue Navboost debug browser
(go/navboost)

This week's newsletter starts with a guest article on Glue by Jeongyun Seo and Ammar Mathur. Glue is a generalization of Navboost in that it provides unified user interaction signal models for all types of search results. User interaction data from Glue is already being used in Web, KE, and WebAnswers. More recently, it is one of the critical signals in Tetris. Read this article to learn about this important project.

The contents of this newsletter are particularly sensitive. Please do not discuss them with anyone who does not already have access to the newsletter.

Glue: Unified user interaction models

Glue is a project for providing unified user interaction signal models to all types of search results. User interaction signals include clicks as well as all other interactions from users with search results or search result pages, which can be mined from session logs. Glue is generalized and extended Navboost.

Navboost has been a critical piece in our ranking algorithms for a long time. However, back to 2013 Fall (when we began the Glue project), we were encountering serious issues in using user interaction signals in Google Search.

First, Google search entered the mobile search era, and a number of new result types such as knowledge panels, Web Answers, Live Result oneboxes (weather, finance, sports, etc) and many types of local results were launched to provide answers directly. Navboost was only available for web results but not available for the non-web results.

Ex. No.
UPX0262
1:20-cv-03010-APM

https://sites.google.com/a/google.com/webranking-team/ranking-newsletters/2016-q1-ranking-newsletters/mar-7-mar-11-2016?tmpl=%2Fsystem%2Fapp%2Ftemplates%2Fprint%2Fshow&showPrintDialog=1 1/21

Redacted

GOOG1-00007989

"**User interaction data from Glue is already being used** in Web, KE, and WebAnswers. More recently, **it is one of the critical signals in Tetris**."

2016

REDACTED FOR PUBLIC FILING

UPX0262, at -989 (Mar. 7, 2016) (emphasis added).

Glue Is Used to "Trigger" and Position Search Features

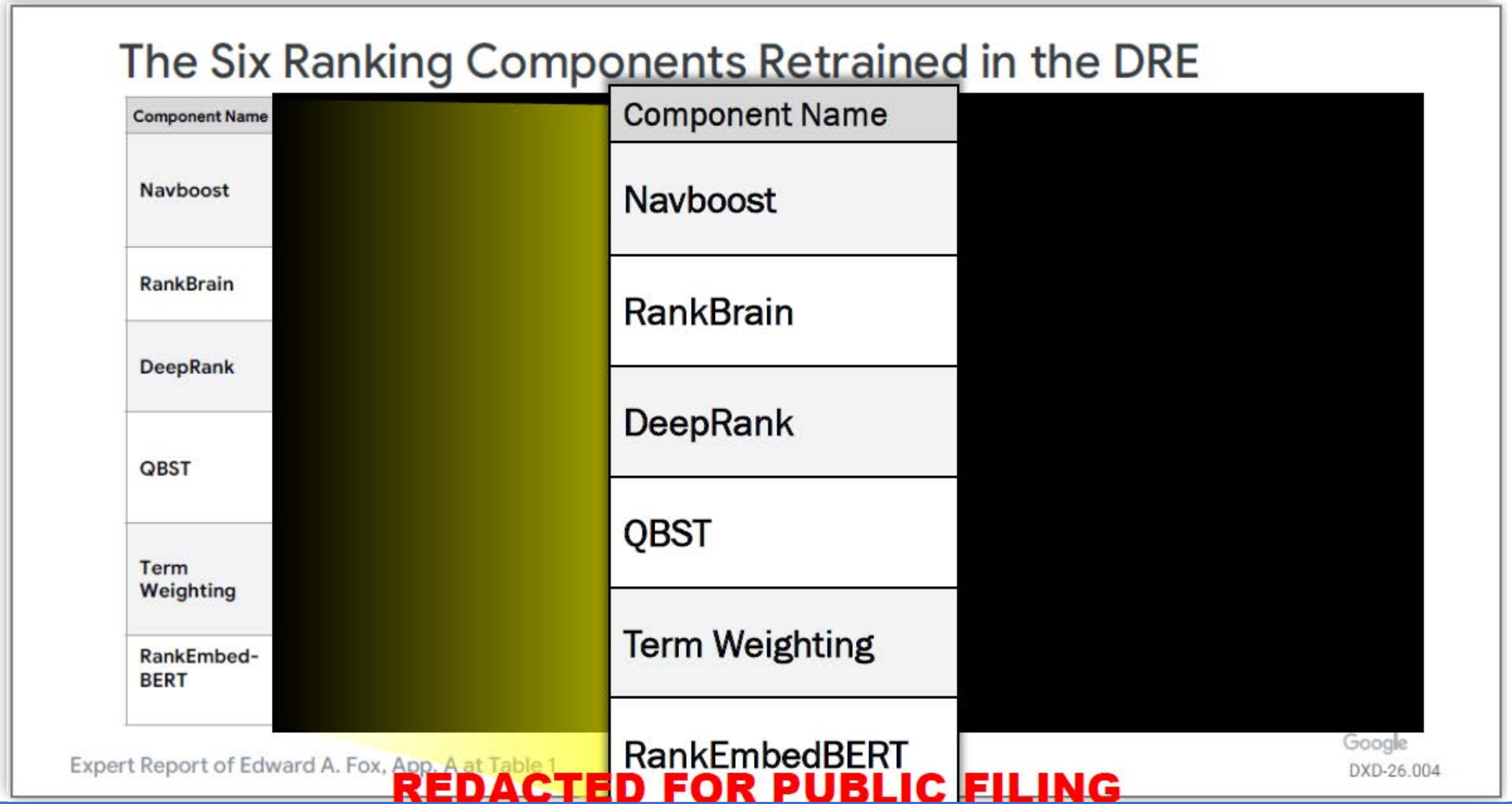
Prof. Edward Fox
Google's Expert Witness

"In simpler terms, **Glue aggregates diverse types of user interactions—such as clicks, hovers, scrolls, and swipes**—and creates a common metric to compare web results and search features. This process determines **both whether a search feature is triggered and where it triggers on the page**."

REDACTED FOR PUBLIC FILING

Expert Reply Report of Edward A. Fox, Oct. 10, 2022, ¶ 46 (emphasis added).

Prof. Fox Has Never Stated that Glue Was Retrained



DXD-26.004 (emphasis added).

The IS4@5 Metric Evaluates Web and Search Features Results

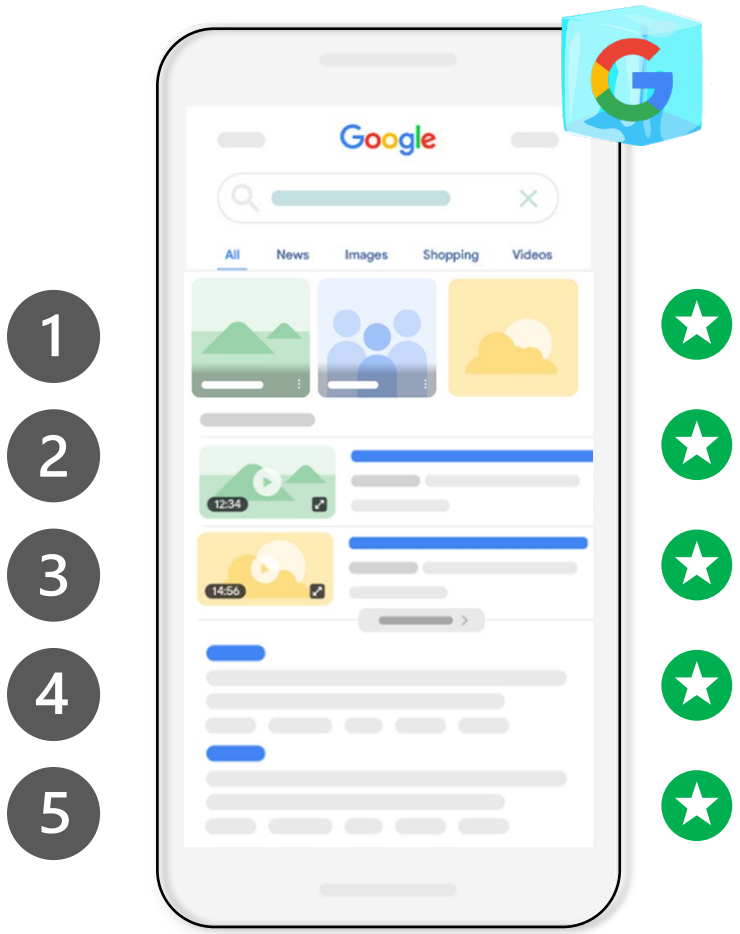
Prof. Edward Fox
Google's Expert Witness

"Google rates the top five positions for IS4@5 **counting both search features** like OneBoxes **and 'blue links.'**"

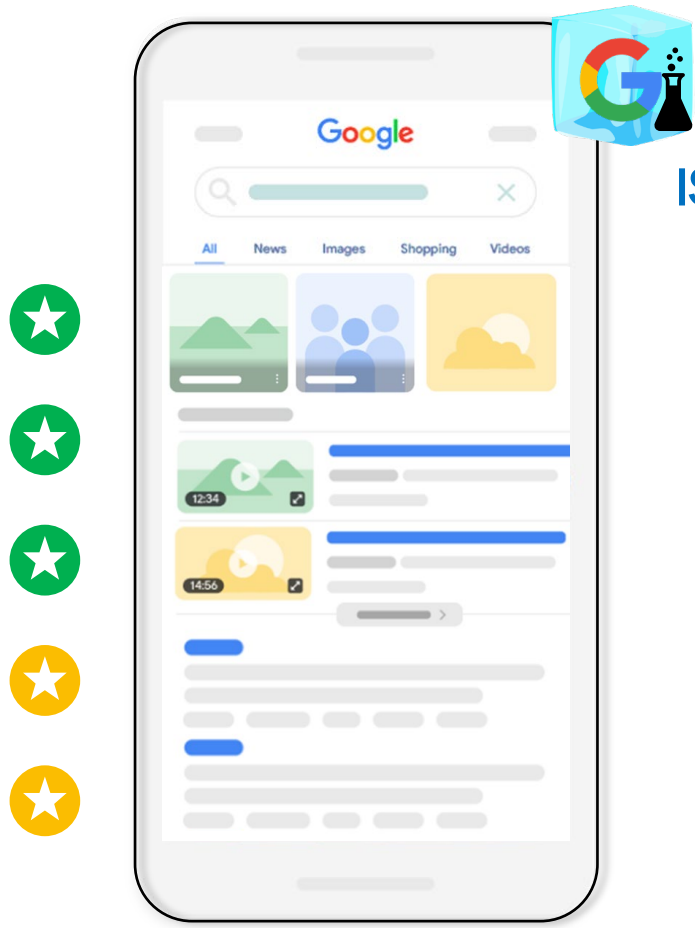
REDACTED FOR PUBLIC FILING

Expert Report of Edward A. Fox, Jun. 3, 2022, App. A, n. 12 (emphasis added).

“10 Blue Links” Ranking’s Effect on IS4@5 Can Be Small



Frozen Google



Retrained Google

IS4@5 Weighting

1.00

0.50

0.33

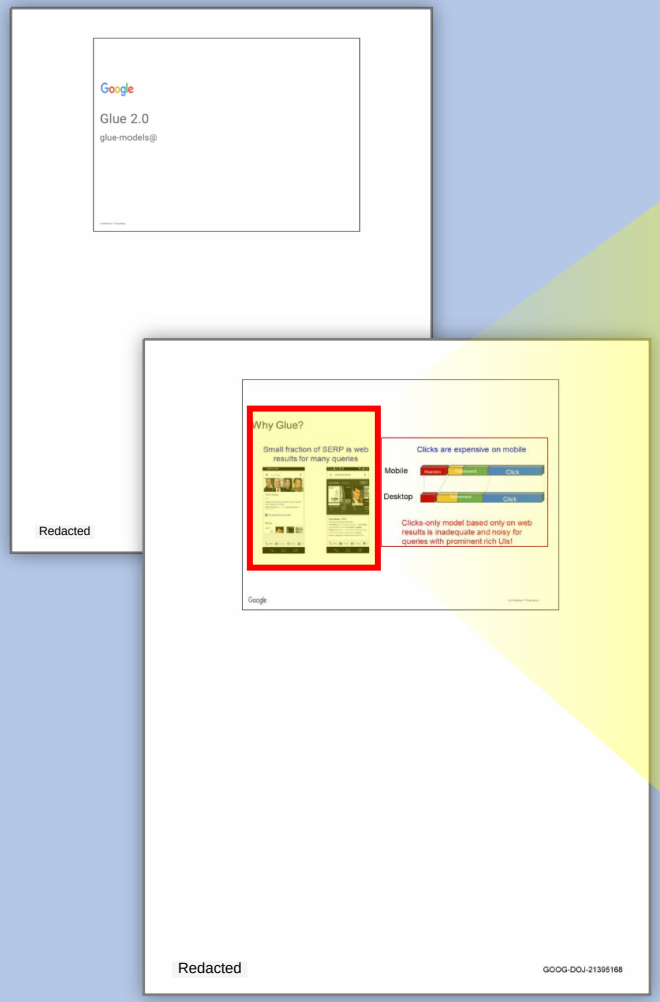
0.25

0.20

REDACTED FOR PUBLIC FILING

Expert Rebuttal Report of Douglas Oard, Aug. 19, 2022, Table 2, at 23; UPX2134, at -077 (Apr. 2019).

Search Features “Trigger” for Many Results



“Small fraction of SERP is web results for many queries”

REDACTED FOR PUBLIC FILING

UPX1114, at -168 (May 3, 2017*) (emphasis added).

My Response to Prof. Fox's Central Conclusions

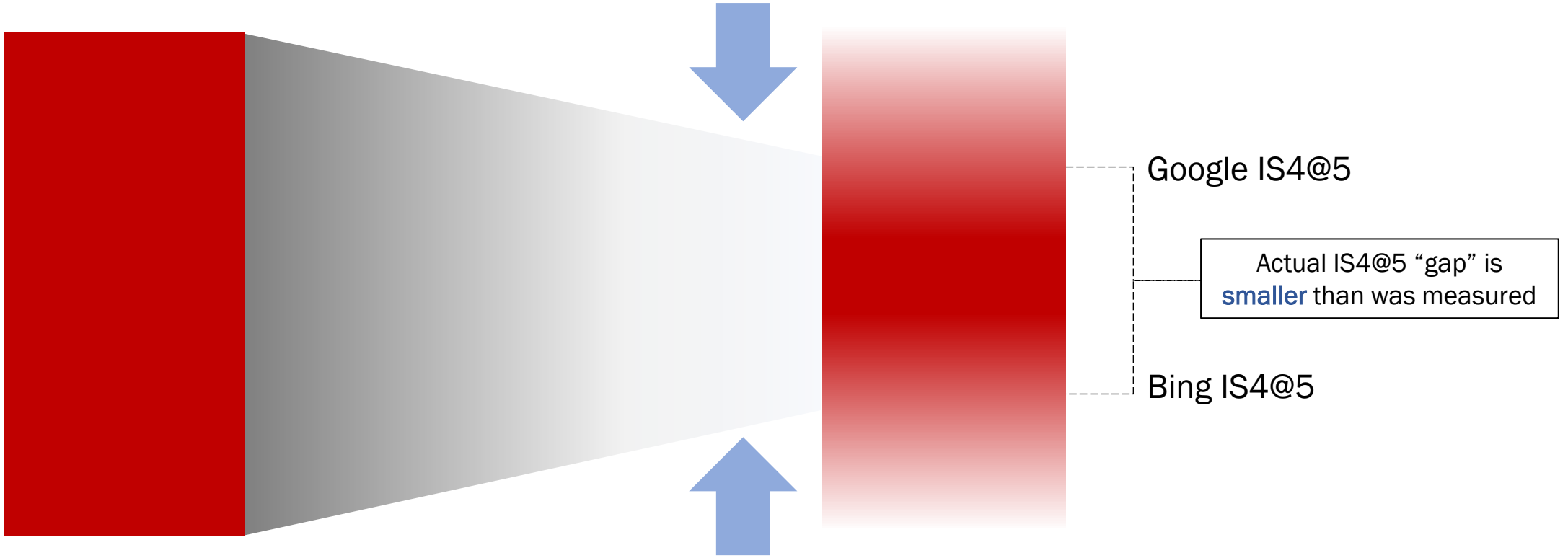
Prof. Fox's conclusions are unsupported because of:

- Unmeasured benefits of user-side data in this experiment;
- Measurement errors in the "quality gap"; and
- Important benefits of user-side data that this experiment cannot measure

REDACTED FOR PUBLIC FILING

Correcting for Measurement Errors

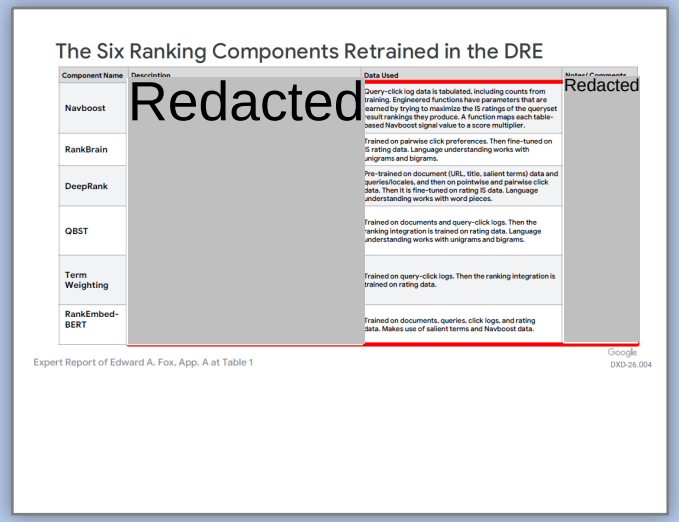
1. Effect of Google "teaching to the test"



2. Google's choice to "rate" all queries using based on mobile presentation
3. Google's difficulties accurately rating Bing's results

REDACTED FOR PUBLIC FILING

Teaching to the Test: Google Trains Using IS, Bing Does Not


The Six Ranking Components Retrained in the DRE

Component Name | Description | Data Used | Notes / Comments

Navboost — Redacted — Query-click log data is tabulated, including counts from training. Engineered functions have parameters that are learned by trying to maximize the IS ratings of the queryset result rankings they produce. A function maps each table-based Navboost signal value to a score multiplier. — Redacted

RankBrain — Trained on pairwise click preferences. Then fine-tuned on IS rating data. Language understanding works with unigrams and bigrams.

DeepRank — Pre-trained on document (URL, title, salient terms) data and queries/locales, and then on pointwise and pairwise click data. Then it is fine-tuned on rating IS data. Language understanding works with word pieces.

QBST — Trained on documents and query-click logs. Then the ranking integration is trained on rating data. Language understanding works with unigrams and bigrams.

Term Weighting — Trained on query-click logs. Then the ranking integration is trained on rating data.

RankEmbed-BERT — Trained on documents, queries, click logs, and rating data. Makes use of salient terms and Navboost data.

Google

Expert Report of Edward A. Fox, App. A at Table 1

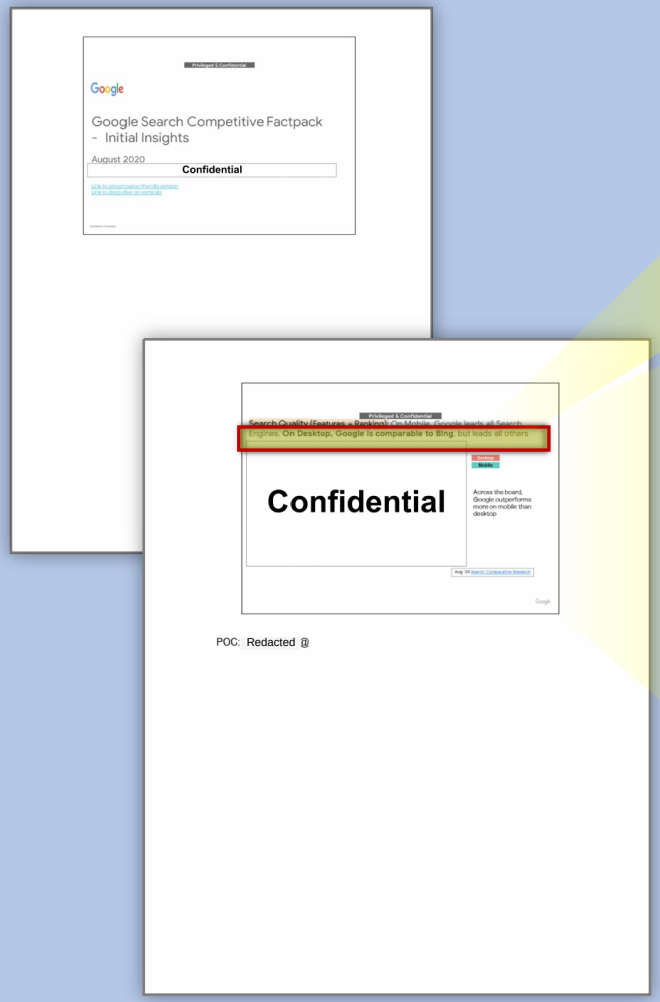
DXD-26.004

Component Name	Data Used
Navboost	…Engineered functions have parameters that are learned by trying to maximize the IS ratings of the queryset result rankings they produce…
RankBrain	…Then fine-tuned on IS rating data…
DeepRank	…Then it is fine-tuned on rating IS data…
QBST	…Then the ranking integration is trained on rating data…
Term Weighting	…Then the ranking integration is trained on rating data.
RankEmbedBERT	Trained on documents, queries, click logs, and rating data…

REDACTED FOR PUBLIC FILING

DXD-26.004 (modified).

Mobile Evaluation Understates Bing's Search Quality



"On Desktop, Google is comparable to Bing"

Privileged & Confidential

Search Quality (Features + Ranking): On Mobile, Google leads all Search Engines. **On Desktop, Google is comparable to Bing**, but leads all others

Redacted

Desktop
Mobile

Across the board, Google outperforms more on mobile than desktop

Aug '20 Search Comparative Research

Google

2020

REDACTED FOR PUBLIC FILING

UPX0268, at -133 (Aug. 2020).

The Measured "Quality Gap" Does Not Account for This

Prof. Edward Fox
Google's Expert Witness

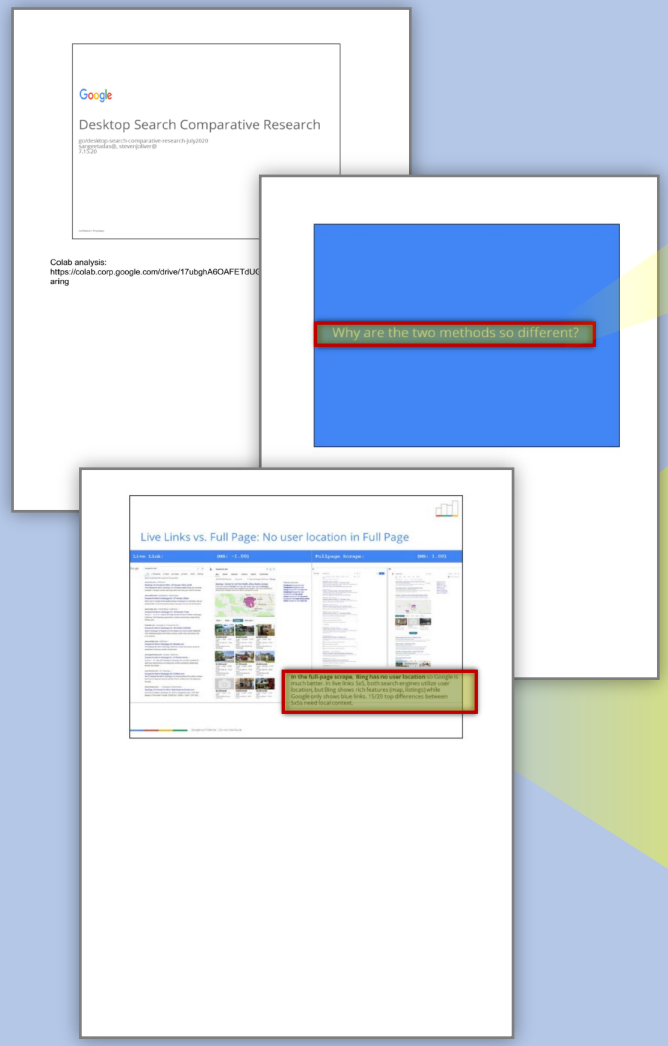
Q. You don't know what the IS gap would be if human raters were looking at desktop presentation; right?

A. Google made a decision some years ago to do all the rater experiments with mobile. So that's all I know.

REDACTED FOR PUBLIC FILING

Testimony of Edward Fox, Oct. 31, 2023, 7977:5–8 (emphasis added).

“Scraped” Results Can Understate Bing’s Search Quality



“Why are the two methods so different?”

“In the full-page scrape, **Bing has no user location so Google is much better.**

In live links SxS, both search engines utilize user location, **but Bing shows rich features** (map, listings) **while Google only shows blue links.**”

2020

REDACTED FOR PUBLIC FILING

UPX0220, at -882–883 (Jul. 15, 2020) (emphasis added).

Correcting for Measurement Errors

Measured difference between Bing and Google

Correcting for measurement errors

Accounting for Unmeasured Benefits

Effect of retraining six components with less user-side data

Effect of retraining all components with less user-side data

REDACTED FOR PUBLIC FILING

My Response to Prof. Fox's Central Conclusions

Prof. Fox's conclusions are unsupported because of:

- Unmeasured benefits of user-side data in this experiment;
- Measurement errors in the "quality gap"; and
- Important benefits of user-side data that this experiment cannot measure

REDACTED FOR PUBLIC FILING

The Experiment Cannot Measure All Effects of User-Side Data

1	Effects on the Innovation Cycle
2	Effects that the IS4@5 Metric Can't Measure
3	Effects that a Frozen System Can't Measure

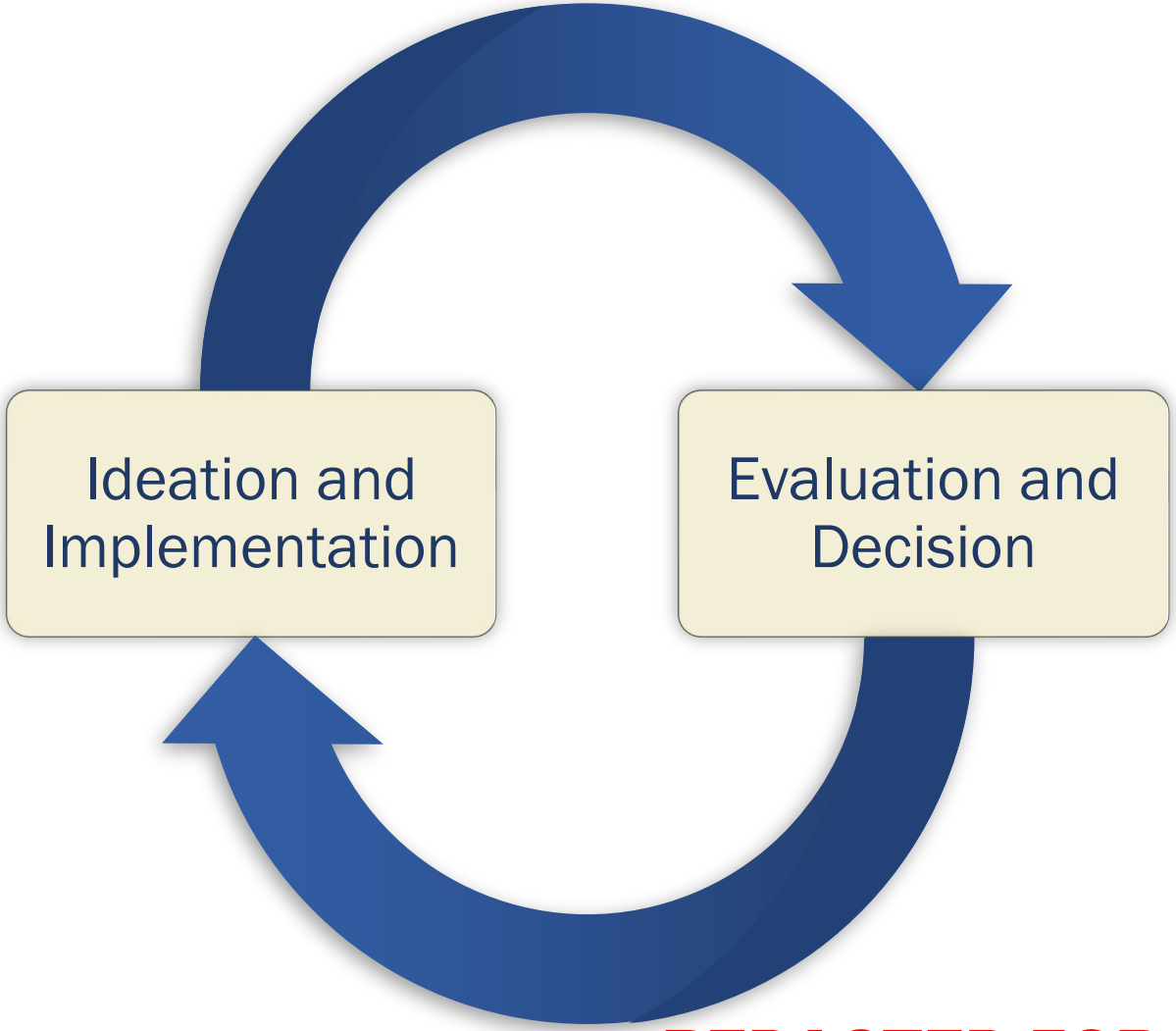
REDACTED FOR PUBLIC FILING

The Experiment Cannot Measure All Effects of User-Side Data

1. Effects on the Innovation Cycle
2. Effects that the IS4@5 Metric Can't Measure
3. Effects that a Frozen System Can't Measure

REDACTED FOR PUBLIC FILING

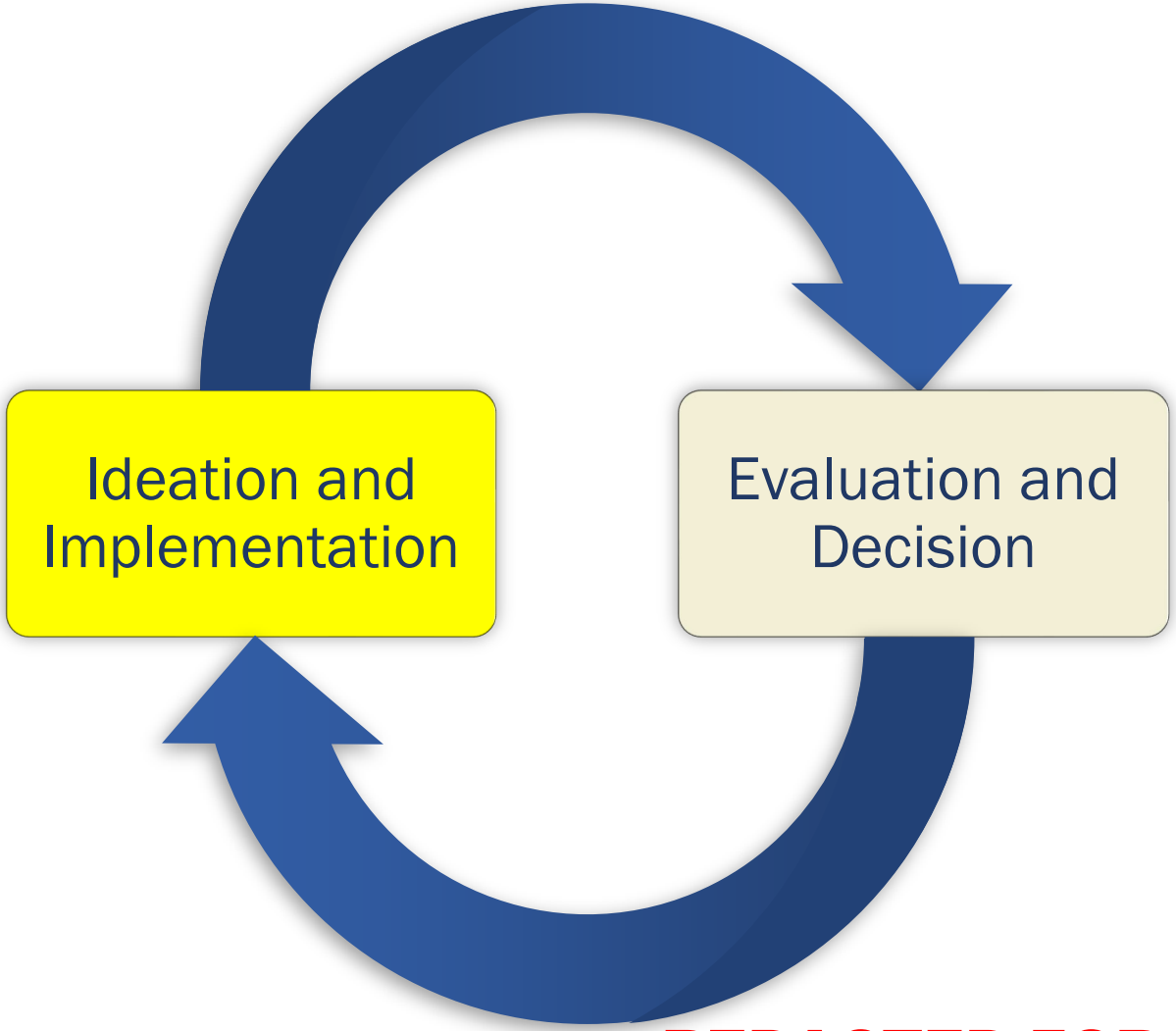
User-Side Data Benefits the Innovation Cycle



People leverage **user-side data**

REDACTED FOR PUBLIC FILING

User-Side Data Benefits the Innovation Cycle



Engineers leverage **user-side data**

REDACTED FOR PUBLIC FILING

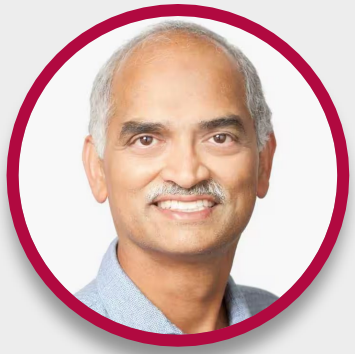
John Giannandrea
Apple SVP of Machine Learning and AI Strategy; Former Google Head of Search and AI

Q. ...So the more queries a search engine sees, the more opportunities it has to improve in this manner?

A. The more opportunities the engineers have to look for patterns and improve the algorithm, yeah.

REDACTED FOR PUBLIC FILING

Testimony of John Giannandrea (Apple), Sept. 22, 2023, 2257:11–2257:15 (emphasis added).



Pandu Nayak
VP, Search

Q. ...[O]ne thing that Google might do is look at queries for inspiration on what it might need to improve on. Does that sound familiar?

A. Yes.

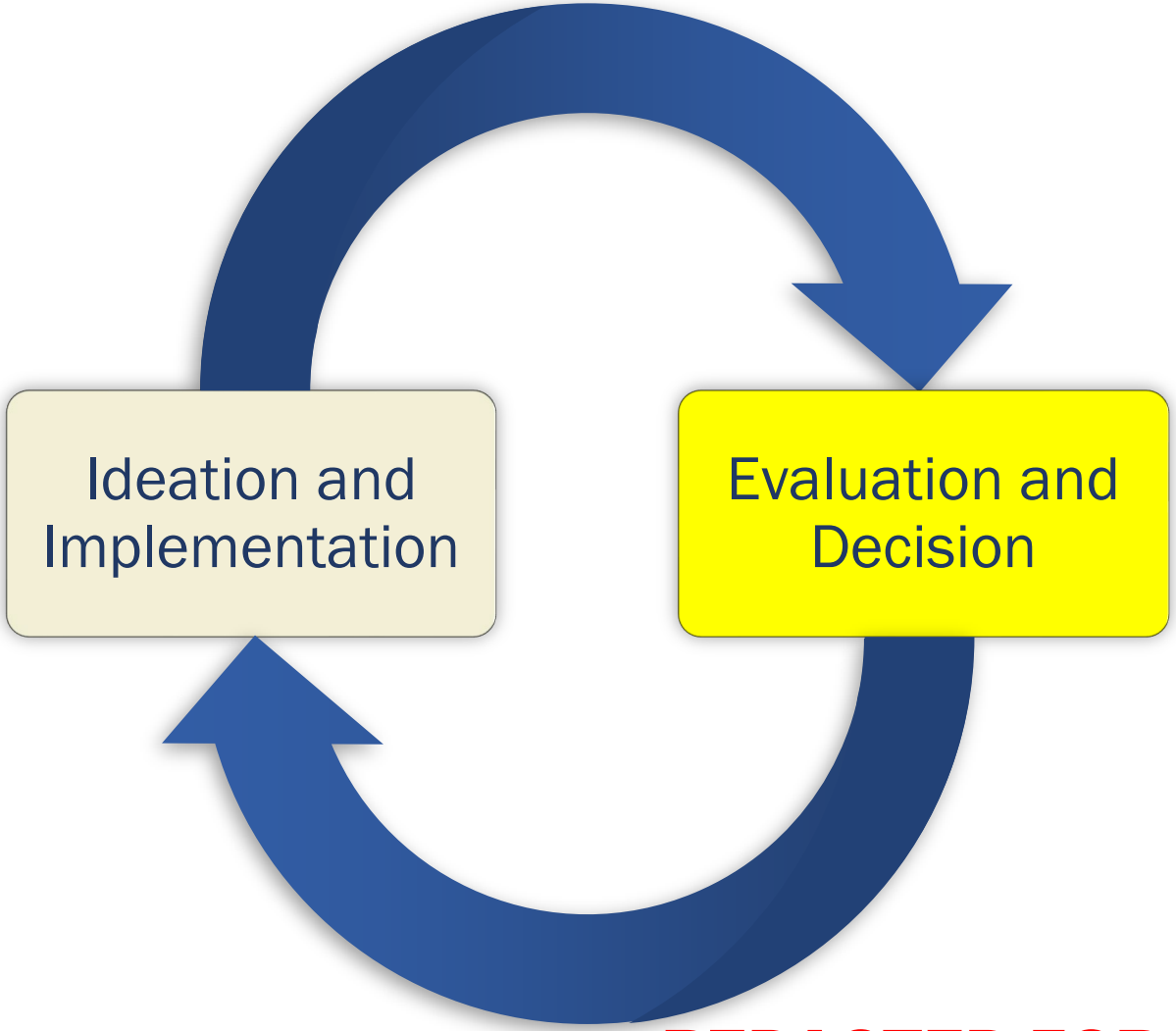
Q. And what does that mean?

A. So we create samples of queries that -- on which we evaluate how well we are doing overall using the IS metric, and we look at -- often we look at queries that have low IS to try and understand what is going on, what are we missing here...So that's a way of figuring out how we can improve our algorithms.

REDACTED FOR PUBLIC FILING

30(b)(6) Deposition of Pandu Nayak (Google), Apr. 7, 2022, 153:4–153:24 (emphasis added).

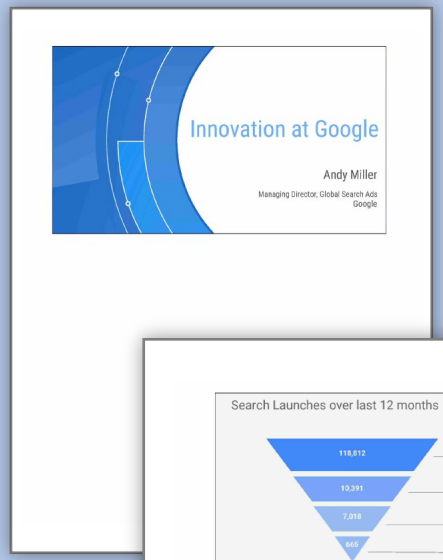
User-Side Data Benefits the Innovation Cycle

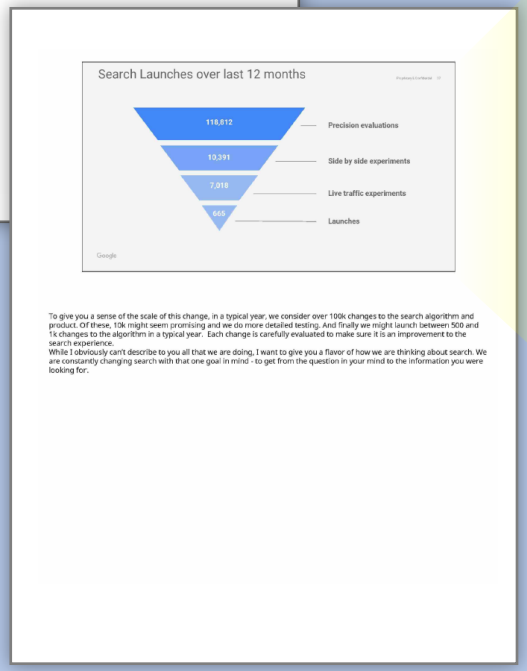


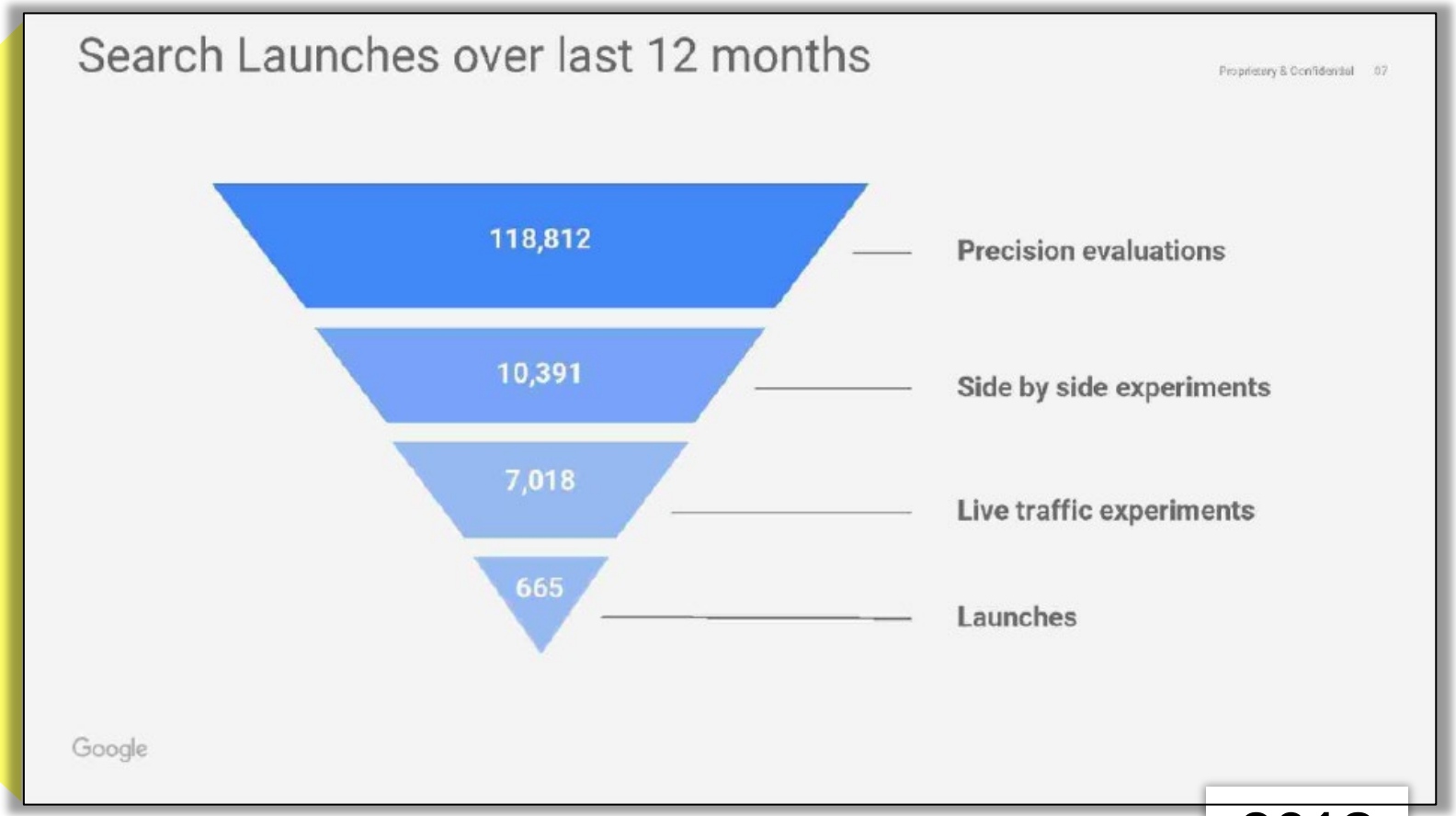
Engineers & Managers leverage **user-side data**

REDACTED FOR PUBLIC FILING

User-Side Data Is Key to Launch Decisions







2018

REDACTED FOR PUBLIC FILING

On the Role of User Interaction Data in Innovation

Prof. Edward Fox
Google's Expert Witness

Q. Whether it's innovation, better algorithms or the like you didn't study, but that's what accounts for the other 97 percent, in your view?

A. So, I don't know what the other parts are. I have guesses because I've worked in the field for a long time, but it's not from user interaction data. That's what I can tell.

REDACTED FOR PUBLIC FILING

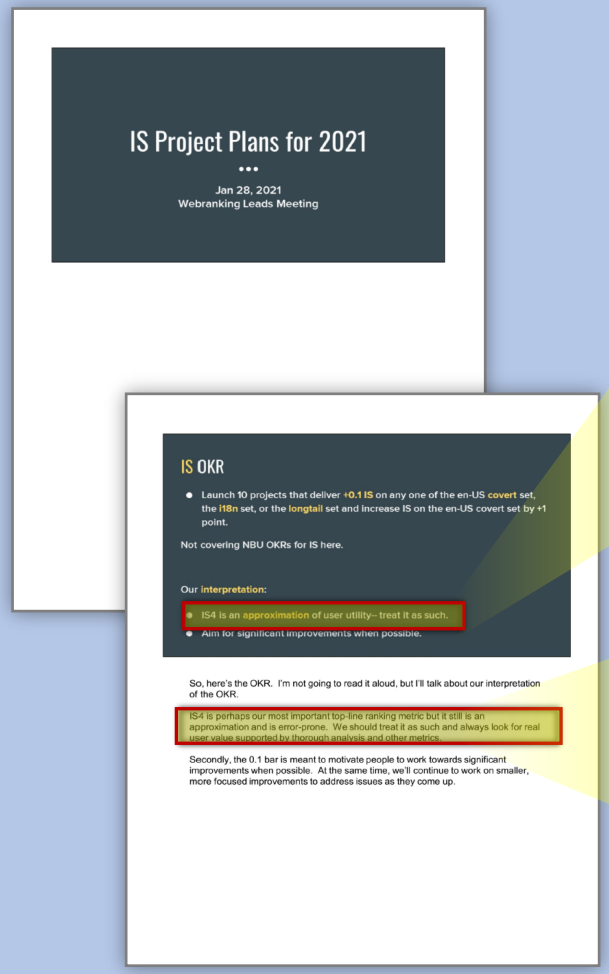
Testimony of Edward Fox, Oct. 30, 2023, 7850:3–7850:8 (emphasis added).

The Experiment Cannot Measure All Effects of User-Side Data

1. Effects on the Innovation Cycle
2. Effects that the IS4@5 Metric Can't Measure
3. Effects that a Frozen System Can't Measure

REDACTED FOR PUBLIC FILING

Metrics Are Not Search Quality



"IS4 is an approximation of user utility-- treat it as such."

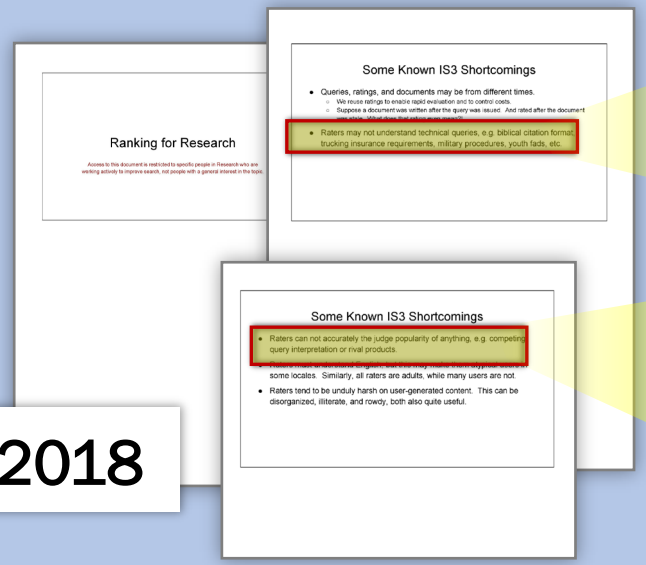
"[A]lways look for real user value supported by thorough analysis and other metrics."

2021

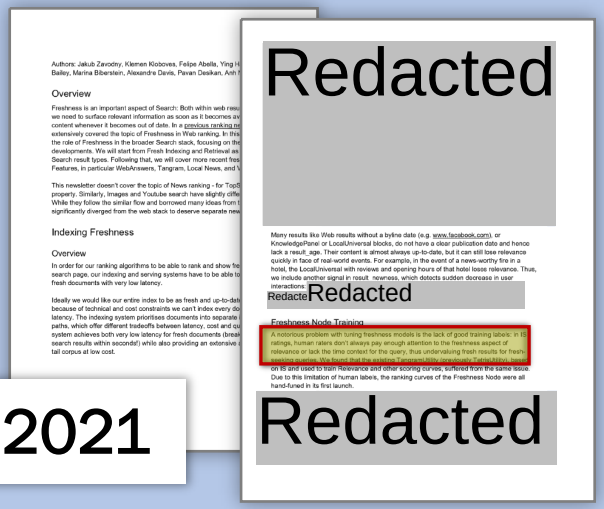
REDACTED FOR PUBLIC FILING

UPX0217, at -792 (Jan. 28, 2021) (emphasis added).

The IS4@5 Metric Is Only a Part of the Story



2018



2021

"Raters may not understand technical queries"

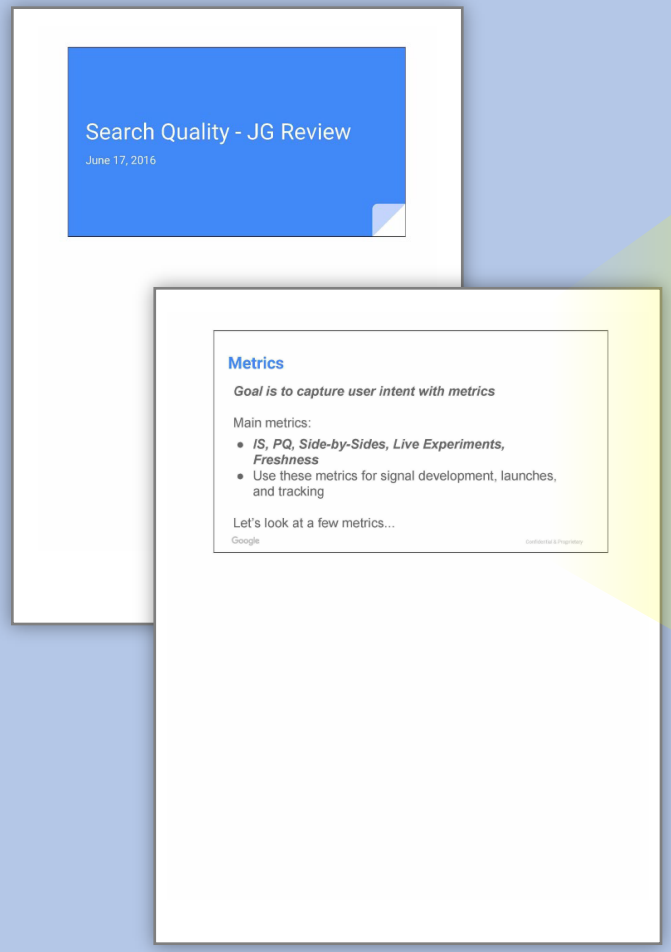
"Raters can not accurately the judge popularity of anything"

"[I]n IS ratings, **human raters don't always pay enough attention to the freshness aspect of relevance** or lack the time context for the query, thus **undervaluing fresh results for fresh-seeking queries**"

REDACTED FOR PUBLIC FILING

UPX0204, at -223, -225 (Nov. 16, 2018*); UPX2133, at -420 (Aug. 11, 2021*) (emphasis added).

Google Uses Many Metrics to Evaluate Search Quality



Metrics

Goal is to capture user intent with metrics

Main metrics:

- ***IS, PQ, Side-by-Sides, Live Experiments, Freshness***
- Use these metrics for signal development, launches, and tracking

Let's look at a few metrics...

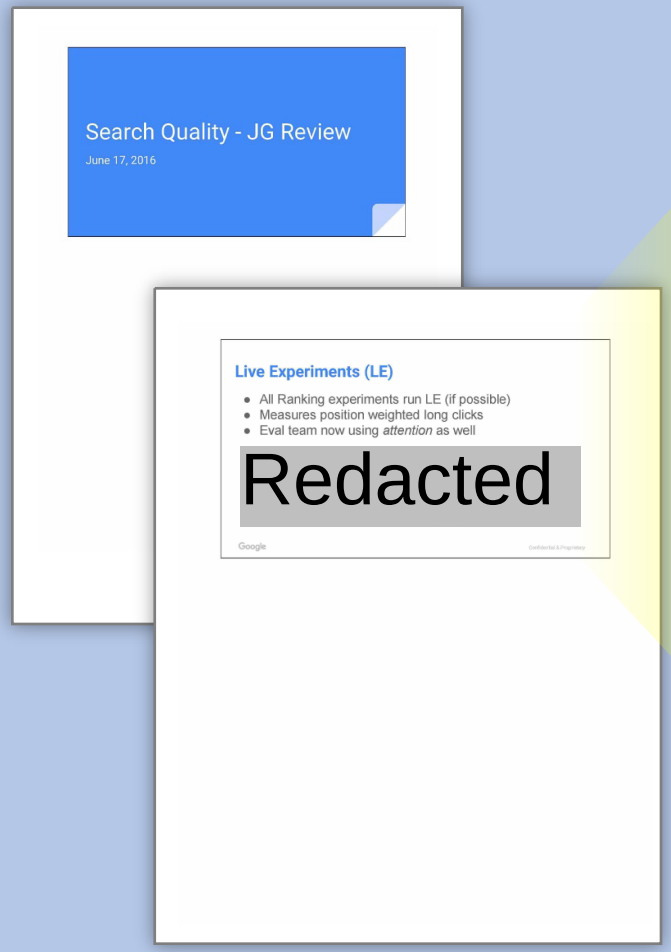
Google

2016

REDACTED FOR PUBLIC FILING

DX0080, at -738 (Jun. 17, 2016).

Live Experiment Metrics Provide Crucial Insights



Live Experiments (LE)

- All Ranking experiments run LE (if possible)
- Measures position weighted long clicks
- Eval team now using *attention* as well

Redacted

Google

2016

REDACTED FOR PUBLIC FILING

DX0080, at -743 (Jun. 17, 2016) (modified).

The Experiment Cannot Measure All Effects of User-Side Data

1. Effects on the Innovation Cycle
2. Effects that the IS4@5 Metric Can't Measure
3. Effects that a Frozen System Can't Measure

REDACTED FOR PUBLIC FILING

Frozen Systems Are Different from Live Systems

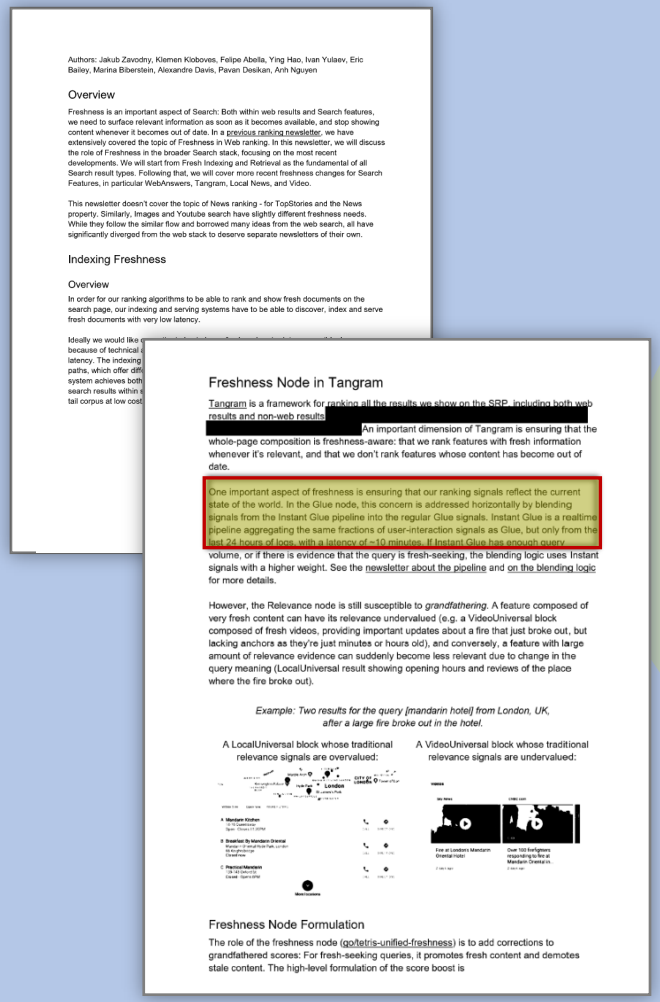
IS4@5 Difference

Redacted

Redacted

REDACTED FOR PUBLIC FILING

Frozen Systems Lack Fresh User-Side Data



Authors: Jakub Zavodny, Klemen Kloboves, Felipe Abella, Ying Hao, Ivan Yulaev, Eric Bailey, Marina Biberstein, Alexandre Davis, Pavan Desikan, Anh Nguyen

Overview

Freshness is an important aspect of Search: Both within web results and Search features, we need to surface relevant information as soon as it becomes available, and stop showing content whenever it becomes out of date. In a previous ranking newsletter, we have extensively covered the topic of Freshness in Web ranking. In this newsletter, we will discuss the role of Freshness in the broader Search stack, focusing on the most recent developments. We will start from Fresh Indexing and Retrieval as the fundamental of all Search result types. Following that, we will cover more recent freshness changes for Search Features, in particular WebAnswers, Tangram, Local News, and Video.

This newsletter doesn't cover the topic of News ranking - for TopStories and the News property. Similarly, Images and Youtube search have slightly different freshness needs. While they follow the similar flow and borrowed many ideas from the web search, all have significantly diverged from the web stack to deserve separate newsletters of their own.

Indexing Freshness

Overview

In order for our ranking algorithms to be able to rank and show fresh documents on the search page, our indexing and serving systems have to be able to discover, index and serve fresh documents with very low latency.

Ideally we would like

because of technical

latency. The indexing

paths, which offer diff

system achieves both

search results within

tail corpus at low cost

Freshness Node in Tangram

Tangram is a framework for ranking all the results we show on the SRP, including both web results and non-web results [redacted] An important dimension of Tangram is ensuring that the whole-page composition is freshness-aware: that we rank features with fresh information whenever it's relevant, and that we don't rank features whose content has become out of date.

One important aspect of freshness is ensuring that our ranking signals reflect the current state of the world. In the Glue node, this concern is addressed horizontally by blending signals from the Instant Glue pipeline into the regular Glue signals. Instant Glue is a realtime pipeline aggregating the same fractions of user-interaction signals as Glue, but only from the last 24 hours of logs, with a latency of ~10 minutes. If Instant Glue has enough query volume, or if there is evidence that the query is fresh-seeking, the blending logic uses Instant signals with a higher weight. See the newsletter about the pipeline and on the blending logic for more details.

However, the Relevance node is still susceptible to *grandfathering*. A feature composed of very fresh content can have its relevance undervalued (e.g. a VideoUniversal block composed of fresh videos, providing important updates about a fire that just broke out, but lacking anchors as they're just minutes or hours old), and conversely, a feature with large amount of relevance evidence can suddenly become less relevant due to change in the query meaning (LocalUniversal result showing opening hours and reviews of the place where the fire broke out).

Example: Two results for the query [mandarin hotel] from London, UK, after a large fire broke out in the hotel.

A LocalUniversal block whose traditional relevance signals are overvalued:

A VideoUniversal block whose traditional relevance signals are undervalued:

Freshness Node Formulation

The role of the freshness node (go/tetris-unified-freshness) is to add corrections to grandfathered scores: For fresh-seeking queries, it promotes fresh content and demotes stale content. The high-level formulation of the score boost is

"One important aspect of freshness is **ensuring that our ranking signals reflect the current state of the world.**

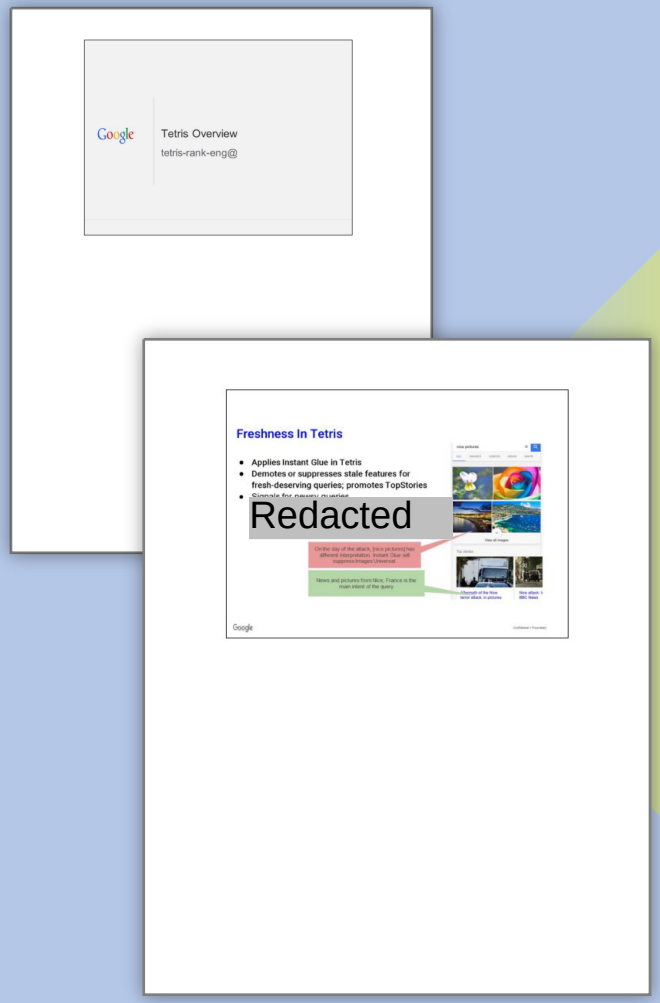
Instant Glue is a realtime pipeline aggregating the same fractions of user-interaction signals as Glue, **but only from the last 24 hours of logs**, with a latency of ~10 minutes."

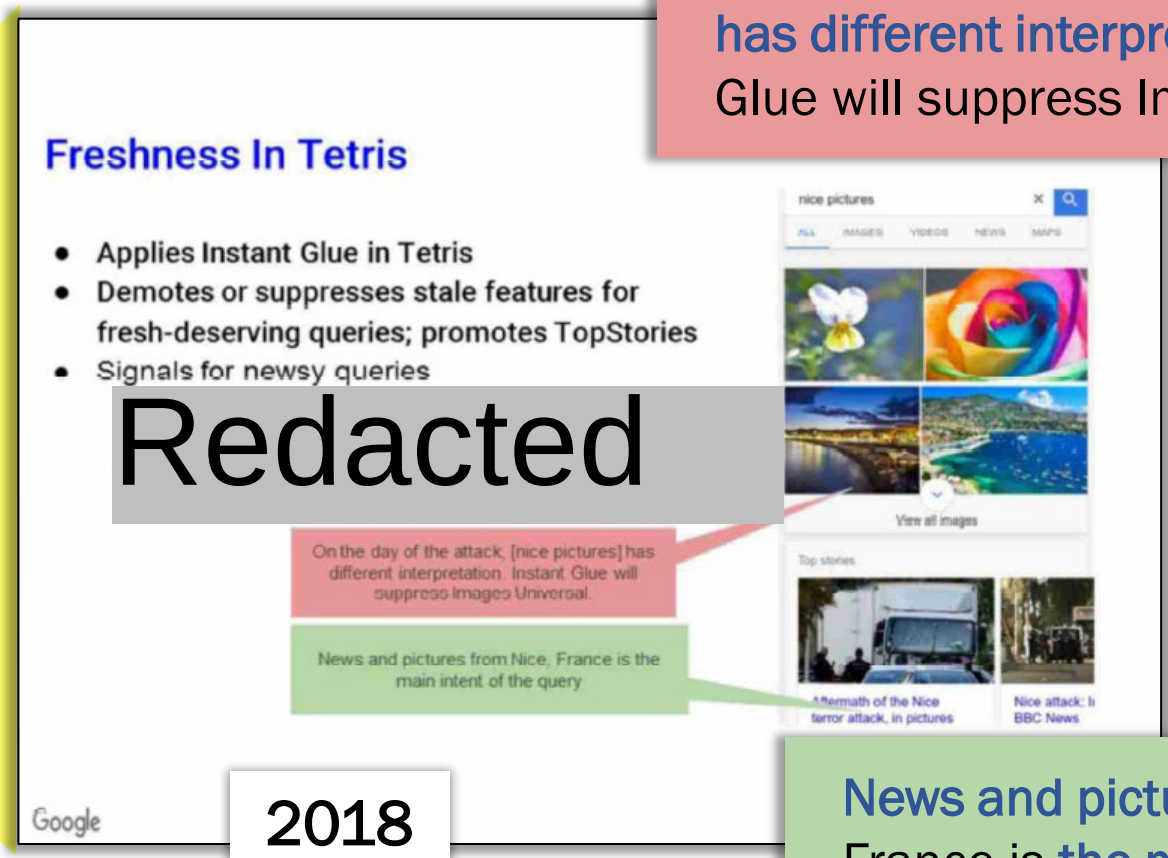
2021

REDACTED FOR PUBLIC FILING

UPX2133, at -419 (Aug. 11, 2021*) (emphasis added).

Freshness Benefits from User-Side Data





On the day of the attack, [nice pictures] has different interpretation. Instant Glue will suppress Image Universal.

News and pictures from Nice, France is the main intent of the query.

REDACTED FOR PUBLIC FILING

DX0116, at -409 (May 24, 2018*) (emphasis added).

This Experiment Can't Test Effects of User-Side Data on Freshness

Prof. Edward Fox
Google's Expert Witness

Redacted

REDACTED FOR PUBLIC FILING

Expert Report of Edward A. Fox, June 3, 2022, App. A, at 10 (emphasis added).

Correcting for Measurement Errors

Measured difference between Bing and Google

Correcting for measurement errors

Beneficial Effects of User-Side Data this Experiment Cannot Measure

Accounting for Unmeasured Benefits

Effect of retraining six components with less user-side data

Effect of retraining all components with less user-side data

REDACTED FOR PUBLIC FILING

Prof. Fox's Third Conclusion

Vast majority of Google-Microsoft search quality gap must be explained by factors other than volume of user interaction data

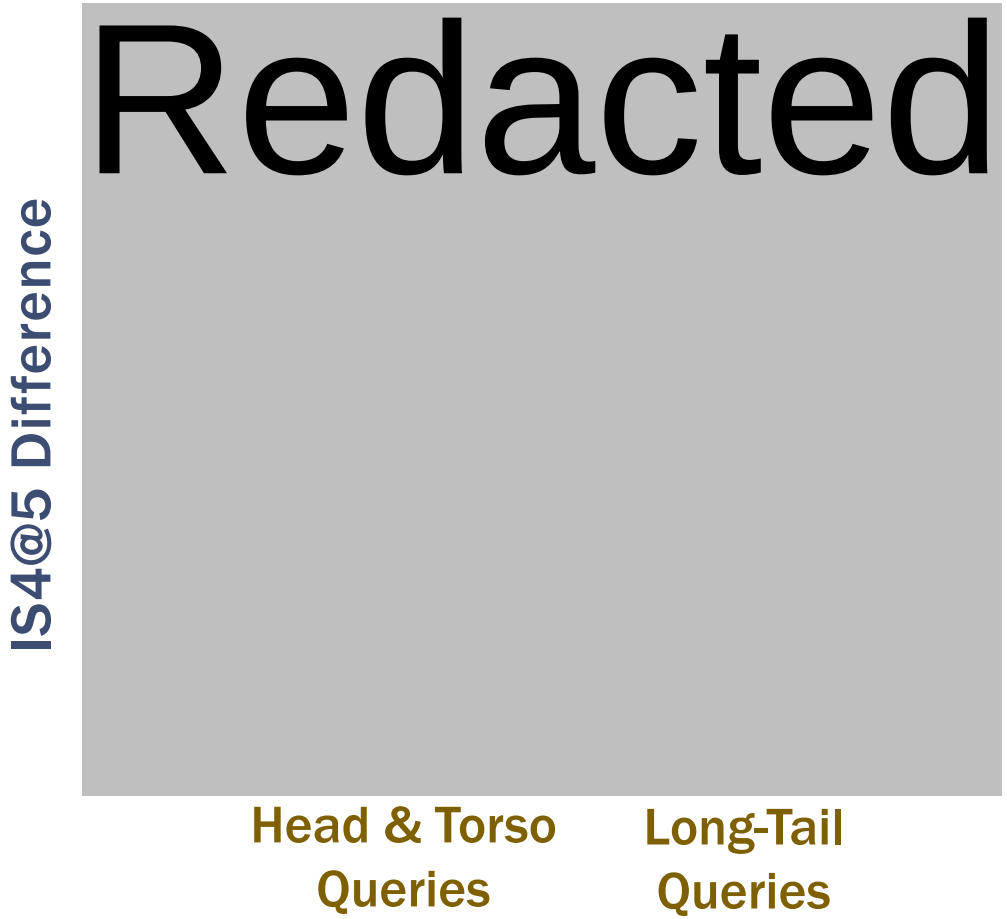
A company as efficient as Google could have search quality similar to Google even at Microsoft's scale

A company as efficient as Google but with Microsoft's scale would not meaningfully benefit from increase in user interaction data

There are diminishing returns to search quality from an increase in the quantity of user interaction data

REDACTED FOR PUBLIC FILING

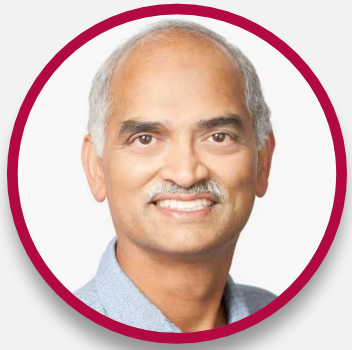
The Results Show a Substantial Effect on Long-Tail Queries



Beneficial effects of user-side data can be different for different queries

REDACTED FOR PUBLIC FILING

Expert Rebuttal Report of Douglas Oard, Aug. 19, 2022, Table 8 (results for “training” query set).



Pandu Nayak
VP, Search

A. So we came up with the following way of thinking about it: **Wikipedia is a really important source on the web, lots of great information.** People like it a lot. **If we took Wikipedia out of our index**, completely out of our index, **then that would lead to an IS loss of roughly about a half point**. So that gives you a sense for what a point of IS is. **A half point is a pretty significant difference** if it represents the whole Wikipedia wealth of information there...

REDACTED FOR PUBLIC FILING

Testimony of Pandu Nayak (Google), Oct. 18, 2023, 6323:8–18 (emphasis added).

Prof. Fox's Final Conclusion

Vast majority of Google-Microsoft search quality gap must be explained by factors other than volume of user interaction data

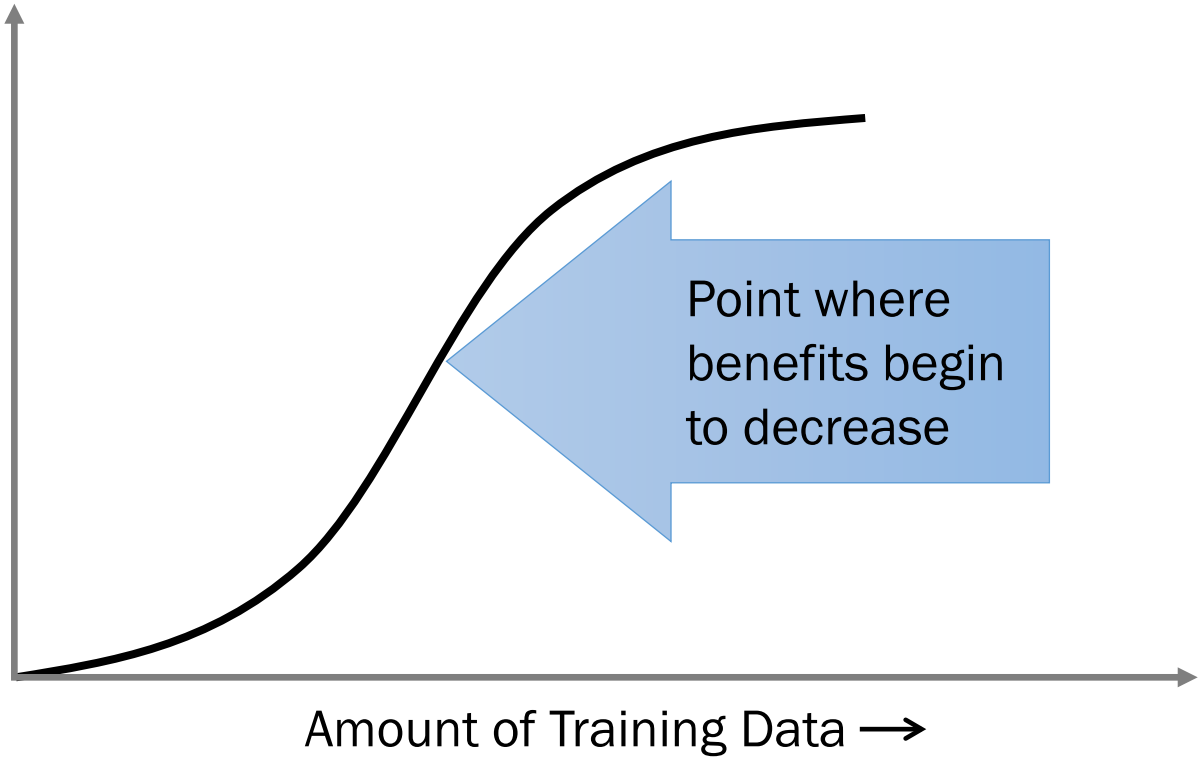
A company as efficient as Google could have search quality similar to Google even at Microsoft's scale

A company as efficient as Google but with Microsoft's scale would not meaningfully benefit from increase in user interaction data

There are diminishing returns to search quality from an increase in the quantity of user interaction data

REDACTED FOR PUBLIC FILING

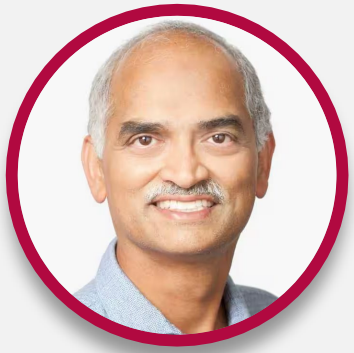
Diminishing Returns Are Not Vanishing Returns



- Benefits continue to accrue
- Benefits would be greater for tail queries, fine-grained location, etc.

REDACTED FOR PUBLIC FILING

Google's Choices Confirm Benefits Continue to Accrue



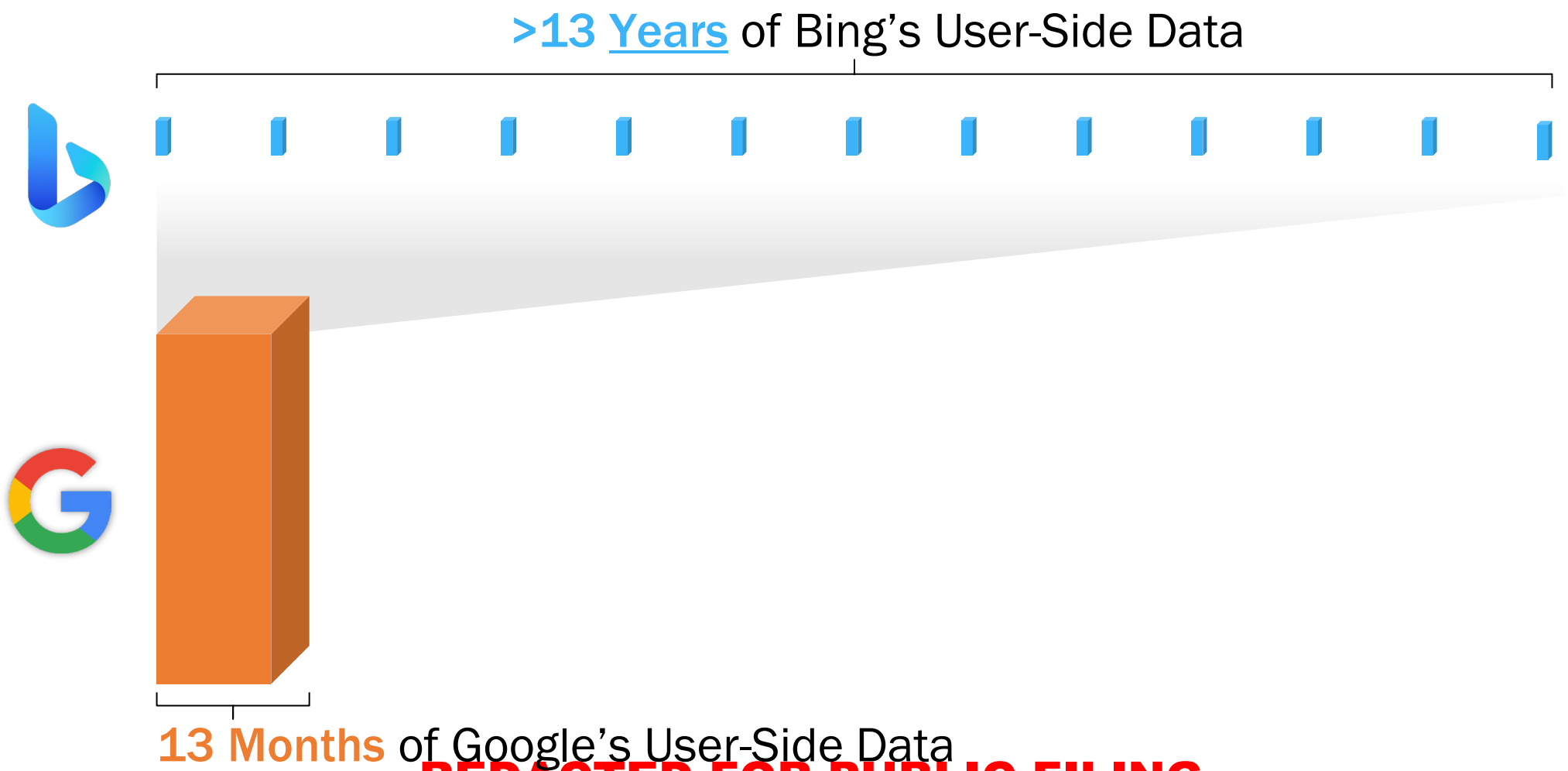
Pandu Nayak
VP, Search

Q. Google has a large collection of sessions logs. Does each click, each piece of data have the same value to Google?

A. ...And so there is this trade-off in terms of amount of data that you use, the diminishing returns of the data, and the cost of processing the data. And so usually there's a sweet spot along the way where the value has started diminishing, the costs have gone up, and that's where you would stop.

REDACTED FOR PUBLIC FILING

Google Clearly Gets "Returns" from User-Side Data



REDACTED FOR PUBLIC FILING

StatCounter GlobalStats (2020 Data).