IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

UNITED STATES OF AMERICA, ET AL.,	)	
	)	
Plaintiffs,	)	
	)	CV No. 20-3010
vs.	)	Washington, D.C.
	)	November 15, 2023
GOOGLE LLC,	)	9:30 a.m.
	)	
Defendant.	)	Day 41
	)	Morning Session

TRANSCRIPT OF BENCH TRIAL PROCEEDINGS
BEFORE THE HONORABLE AMIT P. MEHTA
UNITED STATES DISTRICT JUDGE

APPEARANCES:

For DOJ Plaintiffs:

Kenneth M. Dintzer
U.S. DEPARTMENT OF JUSTICE
1100 L Street, NW
Washington, D.C.
(202) 307-0340
Email:
kenneth.dintzer2@usdoj.gov

Diana Arlen Aguilar Aldape
U.S. DEPARTMENT OF JUSTICE
450 Golden Gate Avenue
Room 10-101
San Francisco, CA 94102
(415) 934-5300
Email: diana.aguilar@usdoj.gov

For Plaintiff
State of Colorado:

Jonathan Bruce Sallet
COLORADO DEPARTMENT OF LAW
Consumer Protection Section,
Antitrust Unit
Ralph L. Carr
Colorado Judicial Center
1300 Broadway
Suite 7th Floor
Denver, CO 80203
(720) 508-6000
Email: jon.sallet@coag.gov

William F. Cavanaugh, Jr.
PATTERSON BELKNAP
WEBB & TYLER LLP
1133 Avenue of the Americas
Suite 2200
New York, NY 10036-6710
(212) 335-2793
Email: wfcavanaugh@pbwt.com

APPEARANCES CONTINUED:

For Defendant Google: John E. Schmidtlein
Kenneth Charles Smurzynski
WILLIAMS & CONNOLLY LLP
680 Maine Avenue, SW
Washington, D.C. 20024
(202) 434-5000
Email: jschmidtlein@wc.com
Email: ksmurzynski@wc.com

Court Reporter: William P. Zaremba
Registered Merit Reporter
Certified Realtime Reporter
Official Court Reporter
E. Barrett Prettyman CH
333 Constitution Avenue, NW
Washington, D.C. 20001
(202) 354-3249

Proceedings recorded by mechanical stenography; transcript produced by computer-aided transcription

- - -

WITNESS INDEX

- - -

P R O C E E D I N G S

COURTROOM DEPUTY: All rise. The Honorable Amit P. Mehta presiding.

THE COURT: Good morning. Please be seated, everyone.

COURTROOM DEPUTY: Good morning, Your Honor. This is civil action 20-3010, United States of America, et al., versus Google LLC.

Kenneth Dintzer for the DOJ.

Jonathan Sallet on behalf of Plaintiff States.

John Schmidtlein on behalf of Google.

THE COURT: All right. Good morning, everyone.

So are the plaintiffs, DOJ plaintiffs ready to begin with their rebuttal case.

MS. AGUILAR: Good morning, Your Honor. Diana Aguilar on behalf of the United States.

The United States calls Professor Douglas Oard as their first rebuttal witness.

THE COURT: All right. Thank you, Ms. Aguilar.

COURTROOM DEPUTY: Please raise your right hand.

(Witness is placed under oath.)

COURTROOM DEPUTY: Thank you.

THE COURT: All right. Professor Oard, welcome.

THE WITNESS: Thank you, Your Honor.

THE COURT: Ms. Aguilar, whenever you're ready.

- - -

DOUGLAS WILLIAM OARD, WITNESS FOR THE PLAINTIFFS' REBUTTAL, SWORN

DIRECT EXAMINATION

BY MS. AGUILAR:

Q Good morning, Professor Oard.

Will you please state your full name for the record.

A Yes. My name is Douglas William Oard, O-a-r-d.

Q And, Professor Oard, can you please tell us a little bit about your educational background?

A Certainly.

I have a Ph.D. in electrical engineering from the University of Maryland at College Park, and I have bachelor's and master's degrees in electrical engineering from Rice University in Houston, Texas.

The bachelor's degree is a double major in mathematical sciences and all the degrees have a computer science concentration.

Q And what is your professional background?

A After I completed my master's degree, I flew for the U.S. Navy. I was on active duty for nearly seven years. Then I worked as a defense contractor here in the Washington, D.C., area, and then returned to the university for my Ph.D.

Q And where do you currently work?

A I'm a full professor at the University of Maryland in College Park with joint appointments in the College of Information Studies and the University of Maryland Institute for Advanced Computer Studies.

Q And what is your field of academic research?

A I do research on information retrieval systems.

Q And broadly speaking, what is an information retrieval system?

A So an information retrieval system is a system that is designed to search for unstructured content. It's distinguished from a database system by the lack of structure or relative lack of structure of the content for which it searches.

Q And can you give us an example of an information retrieval system?

A Well, a general search engine, such as Google or Bing, is an example of an information retrieval system.

Q And do you teach classes on information retrieval?

A I do.

Q And have you published peer-reviewed research on information retrieval?

A Yes, I have.

Q And have you served on the editorial boards of any peer-reviewed journals in your field?

A Yes.

Q And have you ever built an information retrieval system?

A Oh, yes.

Q How many?

A It's a little bit hard to count individual systems because we build systems and then modify them many times in research.

So in research, we're trying to get to the right question, and it's often easy to get the answer once you have the right question but finding the right question is an iterative process.

Typically I might say we do two completely different things a year, maybe more than that, and each one would go through dozens of iterations.

Q So two a year over the course of your career?

A Over the course of 25 years, yeah.

MS. AGUILAR: Your Honor, at this point, I would like to offer Professor Oard as an expert in information retrieval and computer science.

MR. SMURZYNSKI: No objection as to qualifications.

THE COURT: We will recognize Professor Oard as an expert in information retrieval and computer science.

MS. AGUILAR: Your Honor, we've prepared slides to

assist Professor Oard with his testimony. May we publish?

THE COURT: You may.

MS. AGUILAR: May I approach?

THE WITNESS: Thank you.

BY MS. AGUILAR:

Q Professor, I'd like to ask you about your assignment in this case. Who hired you as an expert in this case?

A I was hired by the United States Department of Justice.

Q And what assignment did the United States give you?

A On the next slide, it was to provide my expert opinion on the analysis and opinions offered by Google's expert, Professor Fox, in his June 2022 report.

Q And what did you rely on to complete that assignment?

A Well, I read Professor Fox's report carefully. I read other materials in the case, and I relied on my experience of 30 years of information retrieval research.

Q And can you please summarize your overall opinion.

A If we could go to the next slide, I think that Professor Fox substantially understates the beneficial effects of user-side data on search quality.

Q And in his opening report, did Professor Fox state

what his assignment had been for this case?

A He did.

If we could go to the next slide, please.

So he states that he was "asked by Google's 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."

Q And as part of this assignment, did Professor Fox rely on a set of tests carried out by Google?

A Yes, he did.

Q And were those tests reported in Appendix A of his opening report?

A Yes.

Q I will refer to those as a 2022 DRE today. Does that make sense?

A Yes.

Q Professor Oard, were you able to recreate or replicate the 2022 DRE?

A I was not.

Professor Fox produced a description of the report in Appendix A, but I did not have access to the systems that -- on which the experiments were run, and I did not have access to all of the data that was used in running those experiments.

Q And, Professor Oard, did you review

Professor Fox's in-court testimony?

A I did.

MS. AGUILAR: And, Ms. Kons, if we could turn to the next slide.

BY MS. AGUILAR:

Q Were these the conclusions --

THE COURT: I'm sorry, Counsel. Could I ask a question?

MS. AGUILAR: Sure. Yes, Your Honor.

THE COURT: Could you help me understand what peer review means in the context of your work?

And specifically what I mean is, if you have published an article in a journal and you have used systems and data in order to prepare the article, how is that peer reviewed, and if it does -- and does peer review involve getting access typically to the underlying systems and data that are the subject of the article?

THE WITNESS: So, Your Honor, practice varies in the field. It is becoming much more common for people to register their code. For example, we use a system called GitHub so that someone else can download the code and run it and to release the data on which they ran their experiments.

There are cases in which that can't be done for privacy reasons. And so there are cases of papers that are published for which the code is not released, but the

availability of the code and the availability of the data to permit replication of the experiment is a factor that is now routinely considered and weighed against other factors in the peer-review process.

THE COURT: Okay.

And so in an experiment such as that like the one Professor Fox performed, what would a peer -- say he had sought to publish that study, what would a peer review look like of that effort?

THE WITNESS: So normally the peer review would be performed on the manuscript itself, so the final written description of the experiment.

It's not uncommon, it's becoming more common now, for manuscripts to contain supplemental information that's not a part of the main narrative and they can extend the length of the paper quite a bit because we no longer are paying per-page charges to print.

So in the case of an experiment using a live search engine, some of the data is -- implicates privacy so people sometimes will type their own name into the search engine and it's difficult to identify and redact all of those. So it might not be feasible in some experiments to release the data. And when the code base is proprietary, then it might not be possible to release the code base.

There would be a higher standard for then the

review of the document for the details that had been disclosed about how things were done so that someone else with access to similar data would be able to run a similar experiment and get similar results. So that's ultimately the goal of the process, is to ensure that the study could be replicated in some way even if it couldn't be reproduced exactly.

Of course, the peer-review process looks at many other things, such as whether the conclusions are supported by the data, whether the methods are properly applied, things like that.

THE COURT: And in your estimation, could Professor Fox's experiment be replicated, if not duplicated, based upon either what you've received or what you could have received?

THE WITNESS: So I could not replicate the experiment because I do not have access to the systems on which the experiment was run, nor the data on which it was run.

And, in fact, I don't know if those systems and data have been escrowed in a way that would make it possible to gain access to that.

But from the description in his report, it's possible to understand some of what he did and not all of what he did. And I think that it would be difficult for

someone with access to comparable systems and comparable data to replicate the study in a way that closely approximates what Professor Fox, his experiment did, because we are missing important details on the experiment.

THE COURT: Okay.

And what details?

THE WITNESS: Well, so, for example, a number of components generated the signals that were used to lay out the full web page, and the full web page was evaluated, but we're not told how those components were retrained, we're only told how the components for the ten blue links, the -- not the full search engine result page that might have images or video or spelling correction or any number of other things on it, just the ten blue links. Those were the focus of the experimental intervention, that which was changed, the components that contribute to that ranking.

But yet the measure that was used was then looking at the full layout of the search engine result page.

THE COURT: Okay. Thank you.

BY MS. AGUILAR:

Q Professor Oard, did you have an opportunity to review Professor Fox's in-court testimony?

A I read the transcript. I was not in court for his testimony.

Q And is what's shown on the screen, Slide 5, the

conclusions Professor Fox offered in this case?

A Yes, I reviewed his slides as well, and these were the conclusions on his slides.

Q And do you agree with Professor Fox's conclusions?

A I don't agree with the first three of them.

I think that the fourth statement, if considered separate from the experiment itself, isn't really open to question. So I do agree with the statement of the fourth conclusion, but I don't agree that the fourth conclusion is supported by the experiment results.

Q And did Professor Fox emphasize any of these conclusions?

A He stated that the first component -- the first conclusion was his principal conclusion, I think he said the big one, and that the next two followed from that.

Q And I want to focus on the first two conclusions.

Did Professor Fox offer a basis for his conclusions?

A He did.

If you could go to the next slide, please.

This is a bit of an eye chart, but it sort of captures the entire story, so I thought it would be useful to show from the outset.

I'll tell the story from right to left.

And so Google created a snapshot of its system, a

frozen system shown here at Google with an ice cube.

(Laughter)

THE COURT: I'm sorry, it's been a long trial.

THE WITNESS: And then measured the quality of that system using a metric that Google uses in the ordinary course called IS4@5.

Google then did an intervention to that system and reduced the amount of user side data that was used for training six components, and that intervention reduced the measured search quality of the system. And you can see this purple bar with the yellow bar at the top. So the yellow bar is the difference between the purple bar and the gray bar.

It's really hard to see the yellow bar, it's a very small difference.

And so over here you can see a blown-up version of that yellow bar. So frozen versus retrained Google.

So this is one-half of the story. So this is the experimental intervention. And this reduction in search quality was designed, as I understand it according to Professor Fox, to reflect the reduced amount of user-side data that Bing would have compared with Google in Professor Murphy's opinion as I'm told.

Now, that's the right side, that's the experiment. And -- and we could quibble with the details of the

experiment, but that's what was done.

On the left, starting on the far left, Google measures its own search quality in the ordinary course. And they have a measurement here using the same measure, IS4@5. And if you compare the green bar, which is Google, to the gray bar, which is frozen Google, you can see the gray bar is not quite as high as the green bar because when you take a snapshot of the system, some of the components are no longer effective. So, for example, components that are trying to keep up with what's happening today. And so this one on the left is the actual full Google.

And Google also compares its search quality using this measure, in the ordinary course, to Bing. And so this is Google's measurement to Bing.

And Google measures Bing search quality as being somewhat lower than that of Google, the difference being the red bar. And that red bar is blown up in the center to the same scale as the yellow bar.

So Professor Fox then says, Well, the yellow bar is smaller than the red bar, so that which was done in the experiments can't explain the difference between Google and Bing.

Now, of course, the measurements over here are being made by Google. If Bing had made the measurements, they might have come out differently, but this is Google's

view of the difference between Google and Bing, and Google's view of the difference between frozen and retrained Google.

And this is essentially the basis for Professor Fox's conclusions, is that the yellow bar is smaller than the red bar, quite a lot smaller than the red bar, and thus does not explain the difference between Google and Bing.

THE COURT: And just, you would agree that -- let me ask you: Do you agree that this IS difference is a significant difference? Leaving aside whether it's correct or not, that delta is a large --

THE WITNESS: Oh, indeed, that's quite a large difference, would certainly be noticeable to users and would make a difference to users.

THE COURT: Okay.

THE WITNESS: If it were correct, it would indeed be.

BY MS. AGUILAR:

Q And, Professor Oard, do you agree that this analysis supports Professor Fox's two first conclusions?

A Well, I don't.

The -- I think the analysis overlooks three important factors, which I've summarized on the next slide.

Speaking colloquially, I think the yellow bar is

too small, the red bar is too big, and there are things that the experiment could not measure that also make use of user-side data. So there are unmeasured benefits, there are measurement errors, and there are important other benefits.

I want to be clear that by "measurement errors," I don't mean mistakes, I mean something like if I had gone to the scale to weigh myself this morning and I know my scale's always two pounds off, then I would adjust my weight by two pounds when I was trying to understand what my true weight was, because I know this about my scale. So this is just a fact of the instrument.

Q I'd like to break this down a bit, Professor Oard.

You note that there are unmeasured benefits of user-side data in this experiment. What did you mean by that?

A Well, if we could go to the next slide.

In the upper right, I've summarized what Professor Fox says in Appendix A he did. He worked with Google to identify components to retrain, and he chose six components to retrain, and these six components were motivated by a desire to see the effect on web ranking.

So web ranking uses a number of signals, and there are 11 signals, if I recall correctly, generated by these 6 components that were retrained, and there are another 11 signals that were not retrained, or generated by other

components.

And so by retraining those components, Professor Fox estimated that perhaps 82 percent of effect of user-side data on web ranking would have been explained.

Now, I have quibbled in my report with that calculation of 82 percent in a number of ways, but I think if you think about it as a back-of-the-envelope calculation, it might be 70 percent, it might be 85 percent. But it's a way of estimating which 6 components would have the greatest effect.

And he didn't retrain the six components with the greatest effect, but he looked at the effect sizes and chose six components with large effects, large estimated effects. There are a couple of other components that he skipped over for one reason or another.

He then measured the effect not on web ranking, but he measured the effect on whole-page ranking. And so in web ranking, you get ten blue links, and then in whole-page ranking, what you try to do is you try to intersperse other things into those ten blue links.

So you sort of take those ten blue links as a grid, and you say, Well, I have images, and this is a query where I think an image would probably be what the user's looking for, so I'll show images first and then I'll show the blue links after that.

Or videos, they might be looking for videos for this kind of a query. I'll show that ahead of blue links. Or I might show a couple of blue links and then some images if I wasn't quite so sure that they were going to want to see images but I thought they might.

And not just images and videos, there are answer panels, best guess snippets. There are just all kinds of -- of things.

Google calls these search features.

And so essentially you're trying to get the search features in there.

So the things that get the search features in were not retrained. The search feature -- the things that generate the search features, they weren't the basis for choosing these components to be retrained, and many other factors.

THE COURT: I'm sorry.

Just in other words, the whole-page ranking is essentially the ranking of the SERP in its entirety?

THE WITNESS: It is, with the exception of such advertising.

I think the search advertising might be sort of put in separately than this whole-page ranking process.

BY MS. AGUILAR:

Q And, Professor Oard, you list five items here.

Do each of these benefit from user-side data?

A They do.

Q Okay.

And can you briefly explain how indexing benefits from user-side data?

A Certainly.

So when Google crawls the web, they -- they're trying to find not just every page on the web, which isn't possible, they're finding as many as they can, but they're particularly interested, when they build the index, in indexing pages that people might want to see.

But there are people creating content for the web that have an interest contrary to Google's and contrary to Google's users that are creating spam pages, and Google would like to get those spam pages out of there.

So Google uses user-side data to see what kinds of pages people click on and stay on, what kind of pages people click on and come back from right away, and then they look at the characteristics of those pages and use that to train systems to decide whether to keep those pages -- or whether to keep future pages in the index or not.

That's not the only thing that they use user-side data for. The index has a complex structure, it's a hierarchy. There's a small, fast index and a large, slow index, and many pieces in between. And figuring out where

to put things can also benefit from user-side data.

But as I understand it in Professor Fox's experiments, he's not clear on this, but -- but my understanding is that the web was not reindexed for each of these versions. A snapshot of Google was taken, and it had Google's index, and so the process of getting things into the index wasn't changed.

This could be quite a large effect. I've seen documents that indicate that Google estimates that it's about a 50 percent savings in index size, you need twice as large an index to hold the good content, plus all the spam, as they presently are able to do by -- by doing this index selection process.

Q And, Professor Oard, did you review documents in the record that are consistent with your description of how user-side data is used in indexing?

A I did.

MS. AGUILAR: And, Your Honor, may I approach?

THE COURT: You may.

BY MS. AGUILAR:

Q Professor Oard, I'm showing you what's been marked as UPX2149. This document contains privileged information, so -- confidential information, so please don't read it into the record?

A Okay.

Q Can I ask you to turn to the second page of the document.

And I want to focus you on the first paragraph under "Overviewing Motivation."

Do you see that?

MR. SMURZYNSKI: Your Honor --

THE WITNESS: I do.

MR. SMURZYNSKI: -- this is an exhibit that is neither on Professor Oard's materials relied-upon list, nor on the government's exhibit list. We received it the first time on November 13th, which was two days ago.

So I object to Professor Oard testifying to its contents, the supporting opinion, when it was not on, for both of those bases.

MS. AGUILAR: Your Honor, we've asked that it be -- it's a document that is from Google's files, and we've asked that it be admitted into the record. It supports Professor Oard's opinion, and I believe Professor Fox used a document in the same way during his testimony.

THE COURT: Was it this document or a different document?

MS. AGUILAR: He used a Microsoft document in the same manner.

MR. SMURZYNSKI: Your Honor, I don't believe Professor Fox used anything in his examination that was not

on his materials relied-upon list, but we can -- if there's something counsel wants to direct me to, I'm happy to look at it.

So two points stand with this, it's not on his materials relied-upon list, it was not on the government's exhibit list until two days ago.

Rebuttal is not the time for us to get new documents, new materials relied upon, when our expert can't come and address any of these things.

THE COURT: Ms. Aguilar, what's the explanation for why this wasn't either among the materials that Professor Fox considered or included in an exhibit list?

MS. AGUILAR: This is an opinion that Professor Oard did offer so it was --

THE COURT: I'm sorry, Professor Oard.

MS. AGUILAR: Yeah, it is an opinion that Professor Oard did offer. We think this document explains his opinion or is consistent with his opinion so it's not a surprise to Google that this opinion is something he would offer.

THE COURT: Well, I don't think it's about the opinion as much as what the material is that supports the opinion and whether they're on notice as to whether this would be a document that you would seek to put in front of him during testimony.

MS. AGUILAR: We notified Google of this three or four days ago. I think it might have been Sunday or Monday, so they had notice and they didn't object.

MR. SMURZYNSKI: Well, Your Honor, could I just address that? We're objecting now. We don't have to object when they send us something on Monday to them sending it to us on Monday.

THE COURT: Let's do it. Let's move past this for a moment. I don't want to spend a lot of time with Professor Oard here on the stand arguing about this.

So why don't we continue, we'll move past this and then we can revisit it.

MR. SMURZYNSKI: Your Honor, I just note, I think this is not going to be the only document of this type based on other materials that we received, today's Wednesday, on Monday, on Sunday, and the like. So this is not a one-off, and so I just take your position, but I -- just to be clear, we're going to see this again.

THE COURT: Okay.

BY MS. AGUILAR:

Q Professor Oard, the next item on your list is spelling correction.

A Yes.

Q Can you explain to the Court how spelling correction benefits from user-side data.

A Certainly.

When users type queries in, they often don't type them correctly. And if this produces bad results, they might then correct their own spelling, and search engines can observe this sequence of queries and can then infer that the second try was what had been intended in the first try. So that's an example of a case in which user-side data can be used to learn to perform spelling correction.

The spelling correction systems also have to learn how strongly they believe that, and so they see many such instances and they might see different kinds of correction and then have to learn to guess which is the most likely correction and whether to offer that correction and how strongly to offer that correction, and it's based on what it is that users are doing.

It's not only based on the way they change their queries. Google can also see which documents they click on and spend time on and what words are present there and what spelling changes might have helped to better match the query to the document.

Q And, Professor Oard, the third one you list are search features like images. Can you please explain how search features like images benefit from user-side data?

A Yes.

I mean, at its core, these are just other search

engines. These are search engines not for web pages but for images, not for web pages but for video.

There are many search features like this.

And these search engines need to learn how to represent the content in a way that can be matched to the queries.

The way that's done is by seeing examples of queries and desirable content, and one source of that is to look at where do users pay attention.

Now, in web pages, you can look at where users click, but a user who's looking for images might not click on the image, they might just see the image that they wanted to see and be done with it.

But they might move their mouse pointer as they're doing that, that's referred to as hovers, or on a mobile device, they might change the part of the page they're looking at, change the view port of their mobile device because you can only see a little bit on a mobile device. So all of those features are used in similar ways to the ways that clicks are used with web pages.

To train these systems, components specific to these systems, in conjunction with other things, for example, image interpretation or video key frame extraction, or something like that. So these are just search engines -- or search systems in their own right, and they benefit from

user-side data in the same way that web search might. In fact, perhaps to a greater degree because it might be more difficult to interpret features and images than to interpret words on web pages.

THE COURT: Just to be clear, I think you said this, but just to be clear in my own mind, but something like video and images, they are part of the web ranking -- the whole-page ranking.

THE WITNESS: They're part of the whole-page ranking and not part of the web ranking. So the web ranking is looking at traditional web pages.

THE COURT: Okay.

BY MS. AGUILAR:

Q And, Professor Oard, the next item you have here is search advertising. Can you explain briefly how user-side data benefits search advertising?

A Yes.

So in search advertising, one of the risks is that advertisers' interests might not match the interest of the users, and, of course, search engines have to balance multiple interests. And so in order to estimate the interest of the user, the search engines can look at the quality of the page that the advertising would take the user to. So they're going to try to estimate the quality of that page.

Well, other users have gone to that page before by clicking on the ad. And if they ran screaming from the room, then perhaps we shouldn't show this to too many more users.

And so by looking at the user behavior, it's possible to learn about specific so-called landing pages, the pages that the advertising would take you to, and also general things about landing pages.

Q And, Professor Oard, when you talked about the benefits of user-side data to indexing, one of the things you mentioned was the tiering of the index; is that correct?

A Yes.

Q Okay.

A So it's not just a question of whether it's in the index but the question of where in the index it is.

In fact, you can think of the question of whether it's in the index simply as the last tier. We simply don't bother putting it in the index in the very last tier.

THE COURT: Let me ask -- sorry, another question.

Just to go back to something you just -- the search advertising. I thought you said earlier that whole page web ranking, which was used in Professor Fox's study, did not include search advertising.

THE WITNESS: So the measure that's used by Professor Fox does not include search advertising, that's

correct. But search advertising does take space on the page, it does affect actual user behavior.

So if you focus just on the IS4@5 measure, search advertising not measured. But the IS4@5 measure only sees some aspect of search quality. It's like if I'm concerned about my health, then my weight is an important factor and I should certainly keep track of it, but I probably should get a blood test every year as well from my doctor and maybe cancer checks as well.

THE COURT: So your point is that an important aspect of search quality is not even considered in the ranking, and that aspect of search quality is something that is -- to which user data is critical?

THE WITNESS: So user-side data is used. It's important.

And Professor Fox tells us nothing about whether there were any effects from the six components that he retrained on search advertising. So I'm not saying that there are no effects on any of these from the six retrained components. What I'm saying is there's no analysis in Professor Fox's report that tells us what the size of the effects would be.

And there are many more components that make use of user-side data in Google's broader search engine than are used specifically in web ranking.

And so it seems to me quite reasonable to expect that if you choose components based only on the effect on web ranking, you might have some effect on these things or on some of them, but that effect likely would not be nearly as large as it would be for web ranking which was the focus of the experiment.

BY MS. AGUILAR:

Q Professor Oard, I'd like to now turn to the effects of user-side data on whole-page ranking.

Did Professor Fox describe how Google uses user-side data in whole-page ranking?

A So if we could go to Slide 11, Professor Fox talks about a system called Glue, which aggregates these signals that I mentioned earlier, the clicks, hovers, scrolling, which is horizontally across images, for example, and swipes -- I'm sorry, scrolling and swipes are one's horizontal and the other is vertical. And Glue puts all of that together.

There was an earlier system called NavBoost, which only looked at clicks because NavBoost was designed for web ranking, but as I said, people don't always click on some of these newer search features so a broader range of signals is useful.

So Glue was designed to assemble this broader range of signals and to make those signals available across

Google's search stack. So very many components in the search stack make use of Glue as their source or one of their sources of user-side data.

And Professor Fox is clear here that this Glue signal is used by the system that does the layout of the whole-page ranking to determine whether a search feature should appear at all and where that search feature should appear on the page. So the system that does that is called Tentgram, and its predecessor system was called Tetras.

Q And, Professor Oard, have you seen internal Google documents that describe the important of Glue to those systems?

A Yes.

If we could go to Slide 10, please, one slide back.

So this is from a 2016 document, quite dated now, so it's talking about Tetras, not Tentgram.

And this talks about -- this is when Glue was fairly new. And it's talking about Glue already being used in several components in the search stack. And noting that it's particularly important in the whole-page ranking, the layout of the search engine result page because it's used by Tetras.

Q And was Glue one of the components retrained in the 2022 DRE?

A No.

Professor Fox is clear, if we could go to Slide 12, please, that these were the six retrained components.

Glue is not listed on the list.

This is an important issue because the evaluation measure that Professor Fox used evaluated whole-page ranking. And so in response to questions about whether Glue was retrained, Professor Fox's statements did not state that Glue was retrained, but, rather, stated that there might be some effect on Glue from retraining NavBoost.

Q Now, Professor Oard, I think you've described what the evaluation entailed, the human reader evaluation of search results entailed for the 2022 DRE, but did Professor Fox also explain it?

A Yes, Professor Fox is clear on this in a footnote in his report.

So it's Appendix A in his first report which describes the experiment, and in a footnote there, he's clear about the measure that was used, IS4@5, which rates the top five positions on the page, counting both search features, such as video, he gives here the example, one boxes, and blue links, the 10 blue links.

Q And what is the effect of this -- of the things you've just talked about?

A Well, if we can go to the next slide.

This is the second I chart.

I'll start on the left.

The frozen version of Google was evaluated by human raters. These are people that are hired by Google through a contractor, as I understand it, and trained to perform this rating, and they give a quality score to each of the items that appear on the page.

They do this without regard to where they appear on the page, just whether they're good or not.

And then if another experiment is later run when the same thing appears in a different position on a page in response to a query, Google will already have stored the quality score for that.

And so you can see here an example in which the top three elements are search features in this sketch, and the bottom two elements are web pages.

And just imagine that they're all good, so they get -- they get green stars.

Now, when Google retrains the system, so the retrained frozen system, that's the ice cube with the little experiment symbol in it, then it may be that some different information will be shown on the search engine result page that the assessors are assessing. And the question is, is this worse than it would have been with more user-side data.

That's the question that's being asked in the experiment.

Now, since the first three elements are not web results, we wouldn't expect them to be as likely to change, and so imagine they don't change at all, and what changes are the web results down in Positions 4 and 5.

Now, if, in this example, that were to happen, the score that Google is computing, this IS4@5, has a weight, gives a higher weight to the first position, less weight to the second position, and you can see here on the right the weighting. And so the -- the lower positions get lower weight.

Now, if the web pages were above the search features, then the search features, which weren't changing, would get low weight. But if the search features -- if any search feature's in Position 1, it will dominate the total score that's computed for this page.

And we -- since the -- the experiment is focused on measuring what's happening to the web pages, this tends to reduce the measured effect on the web pages without reducing the actual effect on the web pages. The effect on the web pages could be quite large.

Google has another measure, they call it IS4@5 web, that only pays attention to the top five web pages. And if that had been used, then we could have gotten a clean comparison with what was happening to the web-page ranking,

but we still wouldn't have known what was going on in whole-page ranking, which is what Google is producing, what Google is competing with Bing and other providers on, because components that focus on other aspects of the search stack weren't retrained.

So it would have been possible to make a better matched experiment, and by not doing that, we measure less of the effect than actually occurs, and the effect is focused on web ranking, not on the full search engine.

So those are the two main flaws in the experiment as I see it.

THE COURT: And the web ranking, just to be clear, that's the -- your yellow-starred portions of the page?

THE WITNESS: Right.

So on the left, the bottom two elements were intended to indicate web ranking, or web pages, and on the right, two different web pages than the ones on the left, in this example, better than the ones on the right because they came from a system with more user-side data. So just as an example of how the score would be computed.

But the point here is that -- that we're only measuring a fraction of the effect, maybe a moderate fraction. These search features, they fire quite often. They're quite common to have a search feature on the search engine result page. So this is --

THE COURT: So just to be clear, your assessment is that the retrained version, to the extent it is having an impact on the IS4@5 score, it's with respect to less weighted portions of the overall score?

THE WITNESS: In this example.

If you also now issue another query and you get web results ahead of the search features, then there would be less of an effect to wash out the score.

And if you ever got a page that had nothing but 10 blue links on it, then the IS4@5 measure for that query would be exactly the same as the IS4@5 web measure would have been.

So it's just that whenever search features appear, it's reducing our ability to see the differences that were the focus of Professor Fox's experiment.

BY MS. AGUILAR:

Q Now, Professor Oard, the way we've shown this on the slide is sort of a side-by-side comparison. Did Professor Fox conduct a side-by-side comparison of these results?

A No.

There are a number of ways of performing evaluation. This way of performing evaluation is very efficient because you can reuse the scores from results that later show up on another page.

But humans are less good at assigning a score to something. You think about, you know, scoring Olympic divers or something. Humans don't always hold up the same card as other humans. But they're pretty good at making comparisons; you can say that diver was better than that diver.

And so Google also, in addition to doing experiments like this early in their development process, will then do side-by-side experiments to get more fine-grained comparisons, particularly where search features are involved.

And then they'll do live experiments with live users because live users can show you things that human raters aren't as good as measuring. And the interpretation of the entire result set then allows you to draw a conclusion.

But the IS4@5 measure is just looking at this in one way, and this way is designed for efficiency, so it's designed to allow you to work through ideas quickly, and it provides important information that aren't as well provided by other measures, and so you need to look at all the measures together to get as broad a perspective as you possibly can.

Q And, Professor Oard, you mentioned that search features trigger often in search results. Did you review

evidence in the record consistent with your opinion?

A Yes.

If we could go to the next slide.

MR. SMURZYNSKI: Your Honor, this is another document that was not on Professor Oard's materials relied-upon list.

MS. AGUILAR: Your Honor, this document --

MR. SMURZYNSKI: We object to its use.

MS. AGUILAR: This document is in the record, and experts have been using documents in the record.

It's in evidence, sorry. It's in evidence, and experts have been using documents in evidence during their testimony.

MR. SMURZYNSKI: Your Honor, we had a Case Management Order in this case, paragraph 23.

Thank you.

THE COURT: Why don't we go ahead and take a break since this is a recurring issue.

Professor Oard, I'm going to ask you to step outside for a few moments, and we'll call you back in.

THE WITNESS: Certainly.

THE COURT: Thank you, sir.

Look, I think the question -- and I appreciate the issue. I wish I had a little bit of time to think about this a little bit more, but, you know, the question is

whether if a document -- look, Rule 26 requires, as part of the expert disclosures, the facts and data considered by the witness in forming an opinion. That is obviously a Discovery rule.

I think the question is, if a fact or data is not disclosed as part of the Rule 26 process, and it doesn't sound like some of these documents were, whether an expert witness is foreclosed from relying upon it at trial.

I don't know the answer to that off the top of my head, and it would --

MS. AGUILAR: And, Your Honor, if I may, Professor Oard is a rebuttal witness. Yesterday or the day before, Professor Murphy was on the stand, and he had reviewed the trial transcripts and documents and he was testifying as to them.

And so this is a similar situation, the document is in evidence, and he is allowed to look at documents that are in evidence at trial and use them in his testimony as a rebuttal expert.

THE COURT: I guess the question, though, he was already -- well, look, two things.

One, he was already designated, in part, as I understand, as a rebuttal expert, and so his report presumably was rebutting and prepared to rebut what Professor Fox had done. That's one.

And two, it's not clear to me that what is being presented is rebutting something that Professor Fox relied upon; in other words, some piece of evidence that was presented to Professor Fox that perhaps even you didn't have notice of, and that it sort of would be fair game for him to come forward with something that would rebut that issue.

So it just -- I mean, there's an aspect of this that's difficult in that because he is a rebuttal expert, to introduce new evidence through him or to put evidence before him that was not either designated in his report or disclosed in advance of Professor Fox's testimony, sort of deprives Google of the opportunity to question him about it in advance of trial, one; and two, to put it before Professor Fox to see whether they would have had him testify about it in a way that was different.

I mean, what I'm not hearing you say is that these records are coming forward because of something Professor Fox said in his testimony that was unanticipated or unexpected.

MS. AGUILAR: No, but I -- but I also think these records are not coming forward from something that Professor Fox did not have notice of. Professor Oard was very clear about this issue. This is a central issue that makes the 2022 DRE flawed.

Professor Fox addresses it in his reply to

Professor Oard. And if counsel wanted to address it head-on during the direct, it had a choice to do that. It ignored the entire issue and now is not wanting to let our expert talk about the flaws in this experiment.

THE COURT: I don't think it's that.

I mean, I don't think it's quite that broad.

I mean, he obviously has been talking about what he thinks are flaws in the experiment. The question is simply what evidence he can rely on in order to substantiate that issue.

MS. AGUILAR: Your Honor, I'll also note that I don't think there is any sort of -- I don't hear counsel saying that it's not true that search features fire quite often on the SERP. And so this is sort of --

THE COURT: Well, I don't think that's -- that's not in dispute.

MS. AGUILAR: Okay.

THE COURT: I think the question is what -- what's coming up. And we all have been here long enough to know what a SERP looks like and what the features are of a SERP, and it's not clear to me what's being brought forward is going to elucidate that in any meaningful sense.

MS. AGUILAR: Understood, Your Honor.

Would it be good to take a moment to reconvene and just see where this issue will come up again? I don't think

it's going to come up a lot from here on.

THE COURT: Sure.

I mean, I think the question -- well, do you --

MS. AGUILAR: We can --

THE COURT: Well, let me ask --

MS. AGUILAR: Yes.

THE COURT: -- Google's counsel, or do you have a sense of how many more of these objections you have in -- I mean, you've seen the slide deck now so...

MR. SMURZYNSKI: Well, I have been trying to follow along.

I've not gone and reviewed the slide deck in toto. I can do that on the break, and we can identify how many more instances there are. But I've not -- I've been trying to pay attention to the questions and the answers as opposed to reviewing the deck which I got, you know, so I can't -- I just know that there were some other documents.

THE COURT: Were you all -- actually, I just assumed you had gotten the deck in advance, but you're getting the deck just as it's --

MR. SMURZYNSKI: No, I got it --

THE COURT: Okay. All right.

I didn't appreciate that but that's --

MS. AGUILAR: That's been normal practice, Your Honor, for expert witnesses.

THE COURT: Okay.

MS. AGUILAR: There is testimony and documents that have been used with other witnesses that we do think is appropriate for an expert to use, and that's been the practice for all experts.

THE COURT: Yeah, let me look at the issue briefly.

I don't think -- look, I think we all understand what the ground rules are in the following sense. There's some principles that apply here, right. Experts can rely on testimony that an ordinary witness cannot.

At the same time, you're supposed to make disclosures about what an expert is relying upon, and that's in order to give the other side notice about what the opinion is based upon.

And so the question is whether the -- if you haven't done that, and it seems like at least with respect to some of these that hasn't been done, what are the appropriate circumstances in which there's some exception to that, and I just -- sitting here today, since it hasn't come up for me so I'd like to have a few minutes to just think about it. So why don't we come back at 10:45 and we'll resume at that point.

MR. SCHMIDTLEIN: Just to be clear so you understand the ordering, as you recall, there were multiple

rounds of expert reports. Professor Fox issued a report. They did not offer Professor Oard as a witness in the first round.

THE COURT: Right.

MR. SCHMIDTLEIN: So this report that he's testifying to came in in the second round, and then Professor Fox was given an opportunity to respond to that in a third round.

So to the extent that we are hearing for the first time all of this, you are absolutely correct that Professor Fox had no ability to respond to that in his report and to anticipate it, therefore, on the front end of his testimony.

MS. AGUILAR: Your Honor, there's nothing that we've discussed today that wasn't disclosed in Professor Oard's report and that wasn't replied to in Professor Fox's reply report, to be clear.

THE COURT: No, I --

MS. AGUILAR: Yes.

THE COURT: I don't think there's a dispute about the general -- what he's talking about generally in his testimony.

MS. AGUILAR: Yes.

THE COURT: This is really just about the specific documents that you're looking to put in front of him.

And, again, I don't -- without having gone through this, I don't know how terribly useful the documents are, but there's been an objection, so I need to at least consider it.

So just give us a few minutes. We'll see what we can find and go --

MS. AGUILAR: Yes, Your Honor. When would you like to resume?

THE COURT: Let's just take -- why don't we say 10 of, and if somebody would let Professor Oard know we're going to take this break to resolve this issue.

COURTROOM DEPUTY: Thank you. This Court stands in recess.

(Recess from 10:50 a.m. to 10:55 a.m.)

COURTROOM DEPUTY: All rise.

All rise. This Honorable Court is again in session.

THE COURT: Thank you, everyone. Please be seated.

All right. The first question is how much more, how many additional exhibits are we talking about, pages of a slide deck?

MS. AGUILAR: I think there was two, Your Honor.

THE COURT: Okay.

Anybody wish to be heard any further on the issue?

MS. AGUILAR: Your Honor, I just wanted to note one sort of issue we didn't talk about was we do have some slides with documents Professor Fox relied upon in his reply that we would like to use.

MR. SMURZYNSKI: We have no objection to that. That's fair game in our view.

THE COURT: Right. Right. Okay.

Well, look, I think the parameters of this are as follows which is, you know, Rule 26 does require the part of expert disclosures, that the material that an expert relies upon to form the expert opinion has to be disclosed.

Rule 26(e) further provides that there is an ongoing supplemental disclosure obligation.

And what Rule 26(e)(2) says is that "For an expert whose report must be disclosed under Rule 26(a)(2)(B), the parties' duty to supplement extends both to information included in the report and to information given during the expert's deposition. Any additions or changes to this information must be disclosed by the time the parties' pretrial disclosures under Rule 26(a)(3) are due."

So at least the supplement -- the duty to supplement is ongoing and it does have a timing element to it and it provides that any additional material needs to be disclosed prior to the Pretrial Statement, pretrial submissions.

You know, very quick research, there was a case called *English versus The District of Columbia*, this is a D.C. Circuit case, 651 F.3d 1, and it involved expert testimony in a Section 1983 case in which a doctor was called to testify, and in his reply testimony for the first time identified that he had actually, as part of his opinion, his case-specific opinion, relied on a particular interview that he had conducted that had not been disclosed to the other side.

And what the D.C. Circuit held was that was a violation of Rule 26 not to have, in particular, Rule 26(e)(2), not to have made that disclosure as a supplement but that ultimately it was harmless in that case for a whole variety of reasons.

So, look, I think the bottom line is that relying -- his reliance on documents now that were not disclosed as part of his report or as a supplement, I think, is a, at least, presumptive violation of Rule 26. And in the absence of circumstances such as Professor Fox said something on the stand that warrants the introduction of new records, I think it's probably inappropriate, I think it is inappropriate to permit him to rely on that kind of document to state his opinion here today. Okay?

So to the extent we've got one or more of these, I'll sustain the objection to them on those grounds.

All right. Why don't we bring Professor Oard back in and continue his testimony.

All right, Professor Oard, welcome back. We appreciate your patience.

THE WITNESS: Thank you, Your Honor.

THE COURT: All right, Ms. Aguilar, whenever you're ready.

MS. AGUILAR: Ms. Kons, will you please bring up Slide 14.

BY MS. AGUILAR:

Q Professor Oard, just to sort of re-orient ourselves, can you sort of describe in your view the severity of the implications of this mismatch between the metric used and the results evaluated?

A Yes.

I think there are two points.

One is that if the goal were to compare web ranking, the metric that was chosen is not designed specifically for that purpose and, hence, has less sensitivity, less of an ability to measure that effect.

On the other hand, I think the other serious concern that I have is that the comparison between Google and Bing, and as I understand it, the question in the case is whether user-side data is used -- whether to and to what degree user-side data is useful in search engines generally,

not just in web ranking.

And for that, the measure that was used is not inappropriate, but the focus of training just on web ranking is inappropriate, and the failure to use other measures to complement what IS4@5 is able to measure is also a serious deficiency from that perspective.

MS. AGUILAR: Ms. Kons, if we could skip to Slide 16.

BY MS. AGUILAR:

Q Professor Oard, I'd like to return to your conclusion and response to Professor Fox's conclusions.

You note that there are measurement errors in the quality gap. What did you mean by that?

A Well, as I explained earlier using the example of my bathroom scale being off by an amount that I know, any sensor, any measurement has some error. That error can be a random error or it could be a systematic error. We call -- in this case, random errors are often the result of sampling, so we'd call them sampling errors, and then the systematic errors are errors in the sampled measures, and averaging cannot remove the effect of measurement errors, it can only reduce the effect of sampling errors.

Q And what specific examples of measurement errors were present in the 20 -- in the search quality gap measurement?

A Well, I've identified three on the next slide, on Slide 17.

There are cases in which Google would seem to me at risk, be at risk of overstating its own performance, and cases -- I've shown that at the top, and cases in which Google would be at risk of understating Bing's performance.

Q And let's take those one at a time.

What did you mean by the effect of Google teaching to test?

A Well, if you can go to the next slide, these are the six components that were retrained by Professor -- or for Professor Fox's experiments.

And just to focus on the first one, NavBoost, it says that, in describing the data used, that an effort was made to maximize the IS score, for example, this IS4@5 score, some version of the IS score here.

And, in fact, every one of these is trained on the data that is used to compute the IS score, in addition to user-side data.

Now, if I'm trying to maximize a score, I might not actually maximize my search quality to the same degree.

So if you think about a case where I'm running a driving school and I'm teaching Americans to drive and I have a friend in the U.K. running a driving school, teaching Brits to drive, then you would expect my students

would do better on the American test and that their students would do better on the British test. But in the end, they're both going to be able to drive in each other's countries, and they might be equally good drivers but one will do better on one test, the other better on the other because they've been taught to the test.

Now, this is not something that is an intentional obfuscation or anything. This is simply a natural effect for training to the same measure that you're going to use to do evaluation.

Now, Bing doesn't use the IS measure specifically, it can't use the IS measure specifically, it's a proprietary measure, the details of it are proprietary. And so it optimizes to some other measure, which is also not necessarily search quality, but its mistakes might be different. So Google will be learning better search quality and also the peculiarities of the IS measure, and Bing will be learning better search quality and also the peculiarities of that measure.

So Google will tend, I would expect, to have a somewhat inflated opinion of itself. And if Bing had done the measurements, Bing would likely have a somewhat inflated opinion of itself. It's just a matter of who's doing the measurements and what they can see.

Q And did Professor Fox correct for this measurement

error?

A No, he did not.

Q Another measurement error you identified was the choice of evaluating the queries in mobile presentation. What did you mean by that?

THE COURT: I'm sorry, could I ask a question, a follow-up.

How would one correct for such a thing?

THE WITNESS: The only way I can think to do it is to just bear in mind that the effects might be somewhat overstated.

If we, outside of Google, wanted to know the answer, we might ask Google and Bing to make measurements both of the other and see how much of a difference there was, but we don't have that available to us in the record so that's an experiment that could be run by someone outside of both organizations by asking each organization to make these measurements and make the comparison.

BY MS. AGUILAR:

Q And, Professor, turning your attention back to your second identified measurement error, which was the effect of evaluating queries on mobile presentation, what did you mean by that?

A Right.

So this is -- the first one was something that

tends to reduce the height of the bar from the top. This tends to reduce the base of the bar -- or to increase the base of the bar from the bottom.

So what's going on here is that the way Google is making these judgments is very much like what I showed on an earlier slide. They show, in a mobile presentation form, maybe on a laptop computer or a desktop computer, but in the form that it would be shown for mobile presentation, the results from every query, and the top five results are then scored by a human rater.

Now, about a third of the queries that were used in the experiment for the comparison were desktop queries, they were queries that came from a device with a screen like this.

And for those, my understanding is that Bing has made substantial investments in, for example, the richness of search features for desktop.

But when these are viewed on a mobile device, the human rater might not see the full richness that Bing had designed for the desktop device, and as a result, might tend to undervalue the Bing results systematically.

Again, not in a way that's nefarious, this is simply the ordinary course process that Google uses to do evaluation, because Google is focused on, as I understand it, performance on mobile devices.

So if Bing had done the measurements, they might have seen quite different results for desktop queries if their evaluation process were different.

Q Professor Oard, did you review any evidence in the record that supports this quality differential between mobile and desktop on Bing?

A Yes. If we could go to Slide 19.

So the graphic here is redacted, and so I won't read off the numbers.

But the -- this is a comparison between Google and different search engines, and the one on the far left is a comparison between Google and Bing.

There's a color code in the upper right that indicates that the bluish color, bluish green here, is a difference on mobile, and a bar that's above zero indicates Google is somewhat better.

A bar that is below zero would indicate Bing is somewhat better.

And you can see that the red bar is -- the ones to the left of the blue bars, show the performance on desktop.

Now, this is from running a side-by-side experiment of the type we talked about earlier where you show people Bing results and Google's results from the actual live search engine. Nothing is scraped here. This is -- these are the real results, and you just ask people

which one is better.

These are trained raters. And so the trained raters see a little difference. The note here, which is not redacted, is that on desktop, Google is comparable to Bing.

Q Professor Oard, did Professor Fox correct for this measurement error?

A He did not.

There's a statement he has in his report that states that the results that are reported are simply the measures that were taken with no corrections.

Q And did you review Professor Fox's in-court testimony?

A Yes.

If you could go to the next slide.

He was asked about this specific effect, and the question was, What would the IS gap be if the human raters had been looking at desktop presentations, and his response is that -- that Google just doesn't do that, and that's all he was able to measure, that Google's evaluation process is set up the way it's set up.

THE COURT: Can I just ask you.

The prior slide that shows the quality comparison on desktop, I think you said that that was a side-by-side comparison. Do you know whether that was a side-by-side comparison done on desktop?

THE WITNESS: Yes, I believe it was.

THE COURT: It was?

THE WITNESS: Yes.

THE COURT: They were --

THE WITNESS: The desktop would have been done on desktop and the --

THE COURT: I just wanted to make sure.

THE WITNESS: I believe that -- I believe that's correct.

There's also a comparison that Apple has made that has similar results.

BY MS. AGUILAR:

Q And, Professor Oard, the last measurement area you identified was Google's difficulties accurately rating Bing's results. What did you mean by that?

A Right.

Are we able to show Slide 21?

Q Yes, we are. Thank you for asking.

A So the way that the evaluations are done, they're not normally done on -- using a side-by-side evaluation, they're done with the evaluation of individual results. And so the individual results have to be removed from their original context and built into result lists where the user can't guess which came from which condition.

And so this process is called scraping.

And Google has to scrape their own results, and they have to scrape Bing's results.

Now, when Google scrapes their own results, they, of course, have access to the entire system. So they know, for every query, where that query was issued -- well, for almost every query -- and they can inject that into their system and return results that are appropriate to where that query was issued.

They know how they build the search engine result page, what the code is that actually displays things, so they can scrape it reliably. And they know if they make a change to their search engine result page, to update their scraping code.

When they look at Bing, on the other hand, they look at Bing the same way you and I look at Bing. They see what Bing displays. If Bing has changed it since yesterday, well, sometimes you see their results bounce around because they can't scrape Bing accurately today, but they can fix the problem and scrape it more accurately next week.

So the -- the scraping process is difficult.

The location injection process is difficult. It's harder to tell the system where you are, because you can't actually go there and be there, and so Bing will see a query that might have the wrong location or with no location at all, and so that's what this slide is commenting on is a

case in which they had made these side-by-side comparisons on the live systems and had no such problems, and then they made a comparison with the full-page scrape, and Bing was disadvantaged because it didn't have the location of the searcher.

A third problem is that when Google scrapes Bing, Bing undoubtedly has systems that can tell when it's getting scraped by a -- by a bot rather than by a human, like humans type at a different speed, humans don't issue their queries as regularly. Bots are supposed to identify themselves as bots.

And so Bing might try less hard to make a bot happy than it would to make a human user happy. There's actually a document from Bing saying that when they scrape Google, this is one of the factors that they allow for in interpreting the Google results, is that Google may well do that.

That's, again, not nefarious, it's just why should you work so hard to give the best possible result to a machine. That's not the business they're in.

And so if you can tell it's a machine, you might get a less quick result.

Now, we don't know if any of these problems occurred in the Google scrape of Bing.

What we know is the process by which Google

scrapes Bing was not fully described in Fox's -- Professor Fox's report, and so we don't know that these problems don't exist.

But to the extent that any of these problems exist, and they'd be completely natural problems if they did, then they tend to inflate -- they would tend to reduce Google's opinion of Bing and thus make the bar larger than it should be.

So together, these three effects tend to make the bar somewhat smaller than it looks like just from the measurement, and when Google discusses these measurements, they allow for these effects. They know that scraping isn't perfect, that things change over time, and I expect, although I haven't seen a document from the Google side in the record, that they know that Bing might give them less good results just because they're not trying as hard.

Q Now, Professor Oard, do you --

THE COURT: I'm sorry.

Just to be clear, going back to your original Slide 7, so these three effects, in your view, if they operate in a way that you believe they may have or the way it does -- the way they do, essentially would shrink the size of this red bar?

THE WITNESS: That's right.

We don't know by how much, but -- but we can make

reasonable inferences about the direction of the error, of the measurement error, and it would be to shrink the red bar by reducing the top of the red bar in this original computation off on the left and raising the floor.

THE COURT: Got you.

THE WITNESS: To some degree, but not all four points but something.

THE COURT: And you're not able to quantify what that might be even ballpark, right?

THE WITNESS: I really couldn't, without more details about what was done, give you any useful estimate.

I mean, I doubt that it would cut it in half, I mean, but that's -- that's not really a very helpful response, I don't think.

THE COURT: No.

BY MS. AGUILAR:

Q Professor Oard, we've just gone over several measurement errors, unmeasured benefits of user-side data. I'm going to refer to them as measurement deficiencies.

Did Professor Fox correct for any of these measurement deficiencies?

A He did not.

Q And what would the effect of correcting for these measurement deficiencies have on Professor Fox's conclusions?

A Well, I've tried to make an impressionistic sketch on that on Slide 22.

On the far left is Professor Fox's reported measurement of Google's ordinary course difference between Bing and Google on a specific date.

And on the right, the very small yellow bar, in each case the bar on the left, is the difference that Professor Fox actually measured in his experiment in the IS4 and 5 measure.

Accounting for the things that we talked about with the unmeasured benefits, I'd expect that yellow bar to be larger, considerably larger, if everything had been retrained and the same measure had been used. And I'd expect the bar on the left to be smaller, somewhat smaller.

And this is drawn with sort of fuzzy boundaries because I don't know where to draw the boundaries, but I don't believe Professor Fox would know where to draw the boundaries either, but I expect we would agree that one must account for things like this in interpreting the results of the experiment.

MS. AGUILAR: And, Ms. Kons, if we could go to then Slide 23.

BY MS. AGUILAR:

Q I'd like to return to your response to Professor Fox's central conclusions and focus on the

third bullet.

You note that, "There are important benefits of user-side data that this experiment cannot measure."

What did you mean by that?

A Well, if you can go to the next slide.

I see several issues, and I've listed three here. One is that search engines are developed over time, and those effects haven't been accounted for in this experiment, and couldn't be. An experiment of this design simply couldn't account for those effects.

The second is that the IS4 metric has strengths, but it also has limitations, and so there are things that simply haven't been measured because only this human rater individual item assessment has been done.

And then finally, there are some things that a frozen system can't measure.

Q And I'd like to take those one by one starting by the beneficial effects of user-side data on innovation.

What did you mean -- what are some -- what are those benefits?

A Well, if you go to the next slide, people have to get ideas from someplace, the people who build search engines. They might get ideas from reading papers where somebody else has done something, or they might get ideas from looking at what's happening with the search engine. So

there's a live search engine there, and Google, you could just watch the queries scroll by and see what people are asking.

You could do log analysis and find out where people are not reacting in ways that would indicate that they're pleased with the results. And that might give you ideas of things that you could do.

Then you'd need to build systems.

Now, Professor Fox writes his report entirely about the systems that already exist, but those systems had to be built, and to build those, you have to do what we do in research and development, you have to try out ideas. That's on the left.

On the right, you have to find out if those ideas work. An awful lot of the ideas that we try out don't work. These are very complex systems, among the most complex things humans have ever designed.

And so in order to see if your intuition is right and your implementation is good, you have to measure the results. And you can't measure the results just one way, you'll see just one thing. If you want to make a high stakes decision, like whether to deploy a new component in the search engine, you want to look at it many different ways, and then you want to make that decision based on judgment, not based on the results of the measure. Right?

And this is something that Google and other search engine companies have been doing for a long time, and user-side data is used in every step of the process. So the design of the search engine, the implementation of the search engine, gets better over time because of the existence of user-side data, and you simply can't measure that with an experiment in which you say, now let's take a snapshot of the system. I mean, it just --

THE COURT: Do you mean to say that what is not accounted for in this is the fact that the metrics or the signals that Professor Fox used are themselves based upon user-side data that has been used over the years to develop those systems or develop those signals?

THE WITNESS: That's right.

The mere fact that that's the way Google does it is based in part on Google seeing what works and trying out new ideas, and user-side data is just all over that process. And so that if you have access to more and better user-side data, then you have opportunities to do things here you might not otherwise have.

And that's simply not measured in the experiment, right. That experiment of this general design couldn't possibly measure that effect. I mean, you'd have to replay 20 years of search engine development.

BY MS. AGUILAR:

Q Professor Oard, I'd like to focus on the left-hand side of this innovation cycle, ideation and implementation, you've talked about that a little bit.

Can you provide a little more detail about that process and how it benefits from user-side data.

A Sure.

So perhaps if I could give an example, maybe we could go to next slide.

So this is John Giannandrea, who was previously at Google and is now at Apple.

And they're talking about the use of queries and, I presume, the response to those queries from the user, to improve.

And he's pointing out that what you look for is patterns and then try to improve the algorithm.

So this is not a matter of the algorithm learning from the data. It's a matter of the person who designed the algorithm learning from the data what works.

Q And did you review similar testimony from current Google employees?

A Yes.

On the next slide Pandu Nayak is talking about how Google lays these things out, and so they have a process of looking -- they have many queries in a query set, and

they'll look at which queries are things getting better at and which queries are things getting worse at, and then use that to try to figure out how to improve the algorithms. And you can do that, as he says here, with the IS metric, or you can do it with live experiments as well.

Q And, Professor Oard, I'd like to focus now on the right-hand side of the innovation cycle evaluation and decision.

Can you explain a little bit about how user-side data benefits this process.

A Yes.

So the next slide might help us to do that.

This is a case where we're not talking about just the engineers who build the system, but also their managers who make decisions on what to deploy. You don't want to go turning things loose on users without believing that you're at least doing no harm and hopefully improving things.

And so Google has a well-thought-out process for launching new components that looks a bit like a funnel.

And the very first step, the top of the funnel, they call here precision evaluations, this is from a 2018 slide. This is the kind of study that Professor Fox did.

So this is where you take an idea and you're going to try it out and you see what happens with these individual results that are scored by human raters.

And you don't have to show anything to actual live users of the search engine, and you can reuse the results if you make a small change to your system, you might be able to reuse many of the results. So it's an efficient way of doing it, and as a result, Google engineers do a lot of experiments with this design.

Then things that look like good ideas, you can spend a lot more money doing side-by-side experiments. You can't do as many side-by-side experiments as you can afford to do precision evaluation. And the numbers are all much larger now than they were in 2018.

But in a side-by-side experiment, you take the output of the system, you take the system as it existed before your supposed improvement, and then you take advantage of people's ability to do holistic comparisons, right. And as I said, humans are good at that.

But if you change your system again, you have to pay all over again for that. So you don't want to do that on your first try, you want to do that on something that looked good in the precision evaluation.

Then, before you turn it loose on a user, you really want to find out how users react.

And so what search engines do, and what Google in particular does, is they set up an experiment in which, it's a little like when you went to the eye doctor and they're

like, you know, hey, do you like A or do you like B? Some people see A, some people see B. And we look at what they do.

And so if they click on a page and stay there for a long time, that's good. If they -- if they click on a page and come back right away, maybe not so good. If they click on the third page but not the first page, maybe we didn't get the right thing in the first position and we can learn from that.

And so this is a -- somehow this is the acid test, right? These are the real users, they actually know what they're looking for. But on the other hand, there's things that users might want that we might not want to show them. So they might go after clickbait, and we might be after high quality content. And so the precision evaluation, the side-by-side evaluation are better for that. And so you need all of these perspectives.

And so Google lines up all of these perspectives in a launch report of a standard format report. Every time they want to launch something, there's a committee that reviews these things. And then they build a holistic understanding.

And the answer depends on the question you were asking. So if I'm going to show you on the result page the answer to your question, you're not going to click on

anything. And so I reduce the number of clicks.

Well, I'll check to see if actually we did reduce the number of clicks. I might have other things where I want to show you good results and I want to increase the number of clicks that you make on the good results.

So you have to think about what it is you're trying to change and which of these measures can show you what kinds of insights.

And so this is the decision process. This is what engineers and their managers engage in together to make data-driven decisions about which things to launch. This is where they bet the company.

Q And, Professor Oard, did Professor Fox correct for the effects of user-side data on innovation in reaching his conclusions?

A I just don't understand how you could correct for it. Professor Fox's experiment wasn't designed to measure anything about the innovation cycle. It starts with a snapshot of a system, and so now we have the system. We can't -- we can't ask questions in this experiment design about how the system came to be.

He could have accounted for this in drawing his conclusions. He could have made more modest conclusions, but he didn't.

Q And did you review Professor Fox's in-court

testimony on that issue?

A If we can go to the next slide.

So he was asked what explains the part that's not the 97 -- or the 97 percent that's not that little 3 percent bar, right?

So there's -- the 3 percent bar is explained by his experiment, so he says, and there's 97 percent unexplained.

And he says he doesn't know what the other parts are, but then he corrects that and he says, well, one thing he does know is that it's not from user-interaction data. And I agree with him on that first claim and I just don't see how he could make the second claim.

THE COURT: In other words, you agree with the first sentence, that he doesn't know what the other parts are, but...

THE WITNESS: That's right.

THE COURT: Right.

THE WITNESS: That's right.

But you'll -- I don't believe it makes sense to reject user-interaction data. In the first part, there's all that unmeasured piece that the yellow bar is actually larger. That's one part of the story. But then there's all of this unmeasurable piece that is what's been going on in the innovation cycle and what would have been going on if

you'd had less user-side data available.

BY MS. AGUILAR:

Q Now, Professor Oard, returning to the three items you described as not being capable of measurement in this experiment, the second item you listed was the effects of user-side data that IS4@5 can't measure.

Can you tell us what you mean by that?

A Yes.

So if you can go to the next slide.

Google has a problem, as do all companies that rely on machine learning, that the students we graduate from our universities are trying to make the scores of their systems better.

And Google's goal isn't to make its score better. Google's goal is to make its search quality better, its actual quality better.

And so they have to sort of retrain the people we graduate to think that way.

And so in, you know, orientation and indoctrination slides, this is all over those, to make the point that you don't want to just tune for the metric. Tuning for the metric is a fine thing to do, it's a first good step, but just remember that metric, IS4, is only an approximation of the benefit to the user of what we're doing, and you need to both think about it, that's the

analysis side, and look at other metrics to help you to think about it, right?

And it doesn't mean that improving IS4 won't improve the search engine, but at some point you're chasing IS4 and you're not chasing the search engine, and you have to guard against that and you guard against that by looking holistically.

Q And, Professor Oard, can you give us some examples that things that IS4@5 -- aspects of search quality that IS4@5 is not good at capturing?

A Sure.

So at its most basic level, if we can go to the next slide, IS4@5 is based on these ratings, and Google has a rating guide where they specify what it is the human rater should do, and they have -- the subcontractors have training processes that train them and they have ways of checking to see if they're doing it, but that doesn't make them super human, that just makes them well-trained. There are still things that people aren't good at.

So if I typed in a query and I ask you to guess what I meant from a query, there's only a couple of words in the query, you might guess wrong what it was I was looking for.

Well, okay, but maybe a lot of people would guess right on average.

But I have a Ph.D. in electrical engineering. I can probably type in a query you couldn't judge so well. You could probably type in several queries I couldn't judge so well.

So raters may not understand highly technical queries, right?

And raters are judging the queries later, so they're not so good at judging what's popular. Raters are adults, not all the queries come from adults. Adults, my experience, don't always understand what younger people are looking for. So, you know, the raters might not get it right.

And if the search engine is updating what it's doing every few minutes, well, the raters are going to be evaluating this thing two weeks later, they don't remember every detail of what was happening two weeks ago, even if they know that the query was issued two weeks ago, right? You can't unremember things that have happened since then.

And so anyway, raters -- raters are very useful because not only do they guess whether or not the query -- the needs are met, but they guess about page quality, and so that part of it is extremely useful.

So raters are good for some things, very good for some things. You wouldn't want to stop doing this. But raters aren't everything. Raters have weaknesses, and

Google knows raters have weaknesses. These are from Google documents. It's not a secret.

Q And, Professor, I didn't mean to interrupt you but I wanted to caution you that we're not going to be relying on the next two slides, so I'll ask you my questions without reference to them.

And, Professor Oard, you just said that Google is aware of these shortcomings in IS4@5 from the documents you reviewed; is that correct?

A Yes.

Q And how does Google address these shortcomings in the ordinary course documents you've seen?

A In exactly the way we talked about. They look at multiple metrics. They -- they don't look at a single metric. And they exercise judgment, they don't follow metrics blindly.

Q And, Professor Oard, you mentioned earlier that you had reviewed standard launch reports from Google.

A Yes.

Q And in those processes, were live experiments included?

A Yes.

I can't recall having seen a single launch report that didn't include live experiments, but I was focused on launch reports for ranking components. And so it might be

possible that some other things that are focused on efficiency or something might occur without live experiments.

But I think we -- the -- I just can't think of a single instance where I saw a launch report for a ranking component that didn't have live experiments.

Q And based on your understanding of the document, do you think Google would implement this change of reducing this amount of data without running a live experiment?

A I'm sorry, let me see if I understand the question.

You're asking if Google would reduce its user-side data to this degree without running a live experiment?

Q Yes.

A I couldn't imagine that they would.

Q Professor Oard, what is the effect of not having done a live experiment as part of the 2022 DRE have on the conclusions drawn by Professor Fox?

A Well, you can draw conclusions about what's happening with the IS4@5, and you can't draw conclusions about what's happening with search quality holistically because you're not looking at search quality holistically.

For things that IS4@5 is good at measuring, you could draw stronger conclusions, and for things you know it's weak at measuring, you could draw weaker conclusions.

THE COURT: Can I ask you, are you aware whether there are more holistic measures, whether there is a metric that attempts to actually do more than just capture the IS4?

THE WITNESS: Well, you can build arbitrarily complex measures, but -- but people don't do that.

There's no perfect measure.

George Box, a statistician, said that all models are wrong but some models are useful. And by that what he meant is that we're reducing reality to a cartoon version of reality by intentionally removing lots of details so that we can understand it and make inferences about it.

And so that makes the model wrong in the sense that it's not reality.

But the question is, did you reduce it in a way that was useful.

And so the best approach that we have in information retrieval where we're trying to actually understand what's going on in the user's mind, is to look at this from multiple perspectives, and the kinds of metrics that Professor Fox used are one way of doing that, that have the advantage that they're inexpensive relative to the other measures that we know how to build, and they're particularly good at looking at things like page quality, because people can be trained for that specific task.

And they have value, considerable value, for

telling how we're doing on needs met, what we would call more generally relevance or topical relevance.

But the amazing thing about search engines is that they also have all these users, all this behavior from which you can generate all this insight. And so search engines have ways of doing measurements that would be hard to do in -- it would be very expensive in other contexts.

So the experiments that are run with human users run with thousands or maybe even millions of human users because any one user is behaving very eclectically. I mean, the phone rings, and, you know, all of a sudden there -- seems like they're on the page for a long time, but really they weren't.

And so the fact that search engines have such large numbers of users and so much access to user-side data gives them a perspective that would otherwise be hard to generate that's complementary to the perspective you can get with human raters either doing these -- well, I call them point-wise experiments, the -- you know, making a judgment about each element or doing side-by-side experiments, which are also done by human raters.

BY MS. AGUILAR:

Q And, Professor Oard, just to sort of put a finer point on the question, is it right that because no perfect measure exists, you have to look at multiple measures? Is

that what you were getting at?

A Yes.

It is generally the case that you will benefit from looking at multiple measures. And when you're trying to make high-stakes decisions, then the cost of looking at multiple measures would be, it seems to me, well justified.

You might easily do early work with only a single measure and see if that work were promising, and that's what -- what's done in the innovation tunnel -- or in the evaluation tunnel that I showed you.

Q Professor Oard, returning to your listed facts of user-side data the experiment cannot measure, the last item on your list is "The effects a frozen system can't measure." What did you mean by that?

A Well, if you can go to the next slide, the redacted content here is using data from Professor Fox's report.

The picture on the left shows the difference in Google's measurement between the live version of Google and the frozen version of Google. This is before any user-side data is reduced. This is the full frozen version of Google. And it's doing quite a bit worse.

And that's what you would expect. And, in fact, that's what the content on the right says, is that's what you would expect, because there are components in any search

engine that are trying to keep up with what's happening now, and those components don't do you any good once you take a snapshot of the system.

And that's just a necessary characteristic of this kind of an experiment design. If you had done a live experiment, you wouldn't have this limitation, you would have other limitations, and you'd be able to draw other conclusions.

But you can't draw conclusions about some things using frozen systems just because the -- experiments with frozen systems aren't designed to support that. It's a necessary condition that you -- that you operate without that insight when you do this kind of an experiment.

Q And, Professor Oard, can you provide an example of how a frozen system might differ from a live system?

THE COURT: Sorry, can I interrupt just quickly?

From what I recall, Professor Fox used pre-existing datasets or query sets with Google?

THE WITNESS: He used pre-existing queries, and he also used user-side data as it existed on a certain date. So pre-existing user-side data.

THE COURT: And do you have any insight into whether those query sets themselves were tuned to this sort of freshness issue, for example?

In other words, are we comparing query sets that

were developed at a period of time that's closer to what the -- the frozen version looks like such that the absence of newness in the frozen set would not really be reflected with the query set that's run, because the query set doesn't reflect newness either? Does that make sense?

THE WITNESS: It does.

Professor Fox is clear, admirably clear in his description of how these queries sets were created. They're created in the ordinary course. So the process is -- is well documented.

As I understand it, the covert query set, the one that is being used here, it's called covert not because it's top secret, it's because you don't want the engineers to see what the test is going to be. They might tune to the test even without intending to do so. So it's just withheld from the engineering groups and used to make comparisons.

So on this covert query set, Google simply randomly samples queries coming in. So if they're taking every 4-million query, they just -- whatever it is, they take the query.

They occasionally will remove a query for one reason or another. Maybe it's mangled. Maybe it's asking for inappropriate content and they don't want to spend their rater's efforts to see how well they can find an appropriate content or whatever. So a couple of queries might be

manually removed, and so a 5,000-query set might have 4,970 queries under it or something like that, but they're just random selected.

And then when you run on frozen, when you run with those queries on frozen systems, you get what you get. And if there was something asking about something that happened five minutes ago, if you froze the system at the right time, you get it, and if you didn't freeze the system at the right time, you don't.

But then Google allows for this. They look at the queries on which a new system is doing badly and on which a new system is doing well. And if this effect is present, they'll see it when they look at those queries.

So the entire process requires judgment.

THE COURT: Thank you.

BY MS. AGUILAR:

Q And, Professor Oard, can you provide an example of how a frozen system differs from a live system.

A So it might be useful to go to Slide 41 here and just look at the actual behavior.

This is from a -- the effect of the system called Instant Glue.

And what it's trying to do is it's trying to keep up with what's changing in the world.

So Glue keeps track of clicks on web pages and

attention signals on web pages, but web pages have been around for a long time, they get a lot of attention. And things that are brand-new, they don't have any attention.

So there's this system called Instant Glue that compensates for that by giving attention to new things.

And you can see here Instant Glue in operation.

The pictures here, if you ask for nice pictures, you know, you need something to decorate your PowerPoint slides, then Google tries to give you nice pictures.

But if there's a terrorist attack in Nice, France, and you ask for Nice pictures, which, you know, is spelled just exactly the same way, the day before the terrorist attack, you probably wanted nice pictures, and it's a good guess that you wanted Nice pictures, and we can tell because everybody's clicking on these pictures from Nice or paying attention to the pictures from Nice. And so this is the outcome of the process, the goal of the process.

Q And, Professor Oard, you mentioned Instant Glue. Can you provide a little more detail about what Instant Glue is.

A Yes.

If you can go back to Slide 40.

So this is -- Google has had different -- different variants of this process for accounting for this paucity of user interaction data that's available on new

content. And so Instant Glue is only looking at the last 24 hours of logs. And because of that, the processing can be faster. And so that allows them to get updates available to the search engine in something on the order of 10 minutes.

And this isn't the only way that they can recognize fresh content. They can also recognize fresh content in indexing. But -- because they know sources that provide fresh content. But there are -- the indexing systems weren't retrained either. So all of this is on its -- it is just not modeled by the frozen.

THE COURT: I'm sorry, this is user-side data in what sense? Because the Instant Glue is crawling the web, the Internet, and finding new web pages and evaluating how users are interacting with those web pages?

THE WITNESS: No. What's happening is that the crawler is crawling the new web pages, right. And so you don't crawl everything the same way.

My academic web page, Google doesn't check it every five minutes, but CNN, they check every -- God, I hope faster than that. And so that's, in fact, why some of the index tiering is done is to put the freshest content up at the top where it's going to get searched because people will be looking for it.

And then what happens is, somebody comes along and

they type "nice pictures" and you give them the same nice pictures you did yesterday, and they're like, you know, no, no, Nice, France, pictures is their next query and then you give with them the Nice, France, pictures and they start paying attention there, maybe not clicking but hovering or sending them to their friends or doing whatever they're going to do.

And at that point you say, oh, nice pictures was actually Nice pictures, and after you see the third person, you know, go, no, no, Nice, France, pictures, you're like, okay, fine, let's start putting those pictures up.

THE COURT: I see.

THE WITNESS: And the question is, how fast can you do that?

And so you need a separate component the way the system is architected, to give credit to new things because the main system is giving credit to things that are getting a lot of clicks and these haven't had a chance to get a lot of clicks. So if you want to follow people around the corner, you need to be watching them as they turn the corner.

BY MS. AGUILAR:

Q And to be clear, the Glue, the Instant Glue, is that system that's watching the user?

A That's right.

These words get use said in complex ways. So Glue is the set of data put together by the system called Glue. So there's the Glue system making the Glue data which is creating Glue signals, all of which are different things. But because they're in this chain, people will often just say Glue.

Q And, Professor Oard, was Professor Fox aware that the 2022 DRE's design was not suited to measure the effects of freshness?

A He commented on this in particular, yes.

If you could go to Slide 42.

So here in this redacted content, he names a specific signal which, in his computation of 82 percent, got a substantial amount of weight but wasn't retrained. And the question is, well, why wasn't it retrained? Well, it wouldn't have been useful to have retrained it because we wouldn't have expected it to have any effect. So he exercised judgment here and said that that signal wouldn't have been the best investment to retrain in an experiment simply because the experiment is not designed to measure things like this.

Q Now, Professor Oard, we covered a number of measurement deficiencies and a number of items that the 2022 DRE was unable to measure.

Can you explain how accounting for all of these

things affects your opinion?

A Sure.

If you can go to Slide 43, please.

So this is the picture we looked at earlier with just one addition, that there are some beneficial effects, we haven't even discussed them all, of user-side data that the experiments simply wasn't designed to measure.

And then there were things that could have been measured if more components were retrained. That's on the right, the larger yellow bar.

And then on the left, we have the measurement of the difference between Google and Bing, which we can judge might be somewhat smaller than was actually measured.

And then when you want to draw interpretations from this, what you want is this slide in front of you, and you want to say, okay, I've got this fuzzy red bar, I've got this fuzzy yellow bar, I've got these unmeasured things in the middle.

And that's it. That's Appendix A to Professor Fox's report. This is the situation that we need to interpret in order to draw conclusions about the -- what the effect actually is of additional user-side data from scale for the particular cases that he ran, he ran two cases, one he called low mobile, and one he called high mobile, and the 100 percent which represented Google, but as

we've discussed, was a frozen version of Google because of the experiment design.

So this picture allows you to then think through what's not accounted for in the numbers that we're seeing, right. The numbers help us to think, but they don't tell us the answer.

Q And, Professor Oard, is it your opinion that the effects of user-side data could explain the whole difference that Professor Fox modeled?

A They could. I don't know that they do. I don't know that they don't.

As I say, you just have to exercise judgment here once you have this data in front of you. And if the question is, could it be completely explained? Sure, it could.

Q Now, Professor Oard, I'd like to pivot a bit and turn to Professor Fox's third conclusion.

THE COURT: Quick question.

THE WITNESS: Sure.

THE COURT: I hear you saying when you say could, as in, sure, it's possible.

THE WITNESS: Yes, sir.

THE COURT: But can you put a likelihood of that being the possibility, that, in fact, this is all user-data driven, the difference?

THE WITNESS: I can't.

If we could go back to the last slide.

I think the big question is the effect on the innovation cycle.

A search engine with access to more user-side data is able to do things that can't be done with less user-side data.

So, for example, I've seen, in testimony, people have been talking about the long tail. So there are queries that don't individually come up very often, but together they account for -- well, in Professor Fox's experiments, by definition, the bottom one quarter of the query distribution.

Some of these are seen very rarely. I mean, some of them have never been seen before, some of them are seen once or twice a month in Google's entire search string.

If you had 5 percent that amount of data, that amount of user-side data, and you had something that was occurring once or twice a month, it would occur, at most, once a year. So something that looks to Google as something they can model, something they can work on, is invisible when you only have 5 percent of the user-side data, because your long tail gets to zero, whereas somebody that has 20 times as much user-side data will see 10 or 20.

THE COURT: Thank you.

BY MS. AGUILAR:

Q Professor, I'd like to pivot to discussing Professor Fox's third conclusion.

Do you agree with Professor Fox's third conclusion?

A I don't, and, frankly, I'm surprised Professor Fox doesn't disagree with it, too.

There's direct contradiction of this in his results.

Q Can you please explain.

A Well, our discussion of long tail queries was timely.

Professor Fox measured the differences. Now, these were with the experiment we just talked about, with all of its limitations.

He saw an effect.

Can we go to the next slide?

That redacted here on the projected slides, but you can see the size of this difference.

So this is simply taking the training queries. Now, these are not the covert queries we've been talking about. There are only about 5,000 of those. And only a quarter of those are long tails, so about 1250.

But the training query set is half long tail queries.

And there are 15,000 queries in the training query set.

It's not actually used for training but that its name.

And so half of that 7,500.

So there are a lot more queries in the training query set.

And as a result, we can get more accurate measurements.

And so using the long tail queries in the training query set, Professor Fox measures this reduction in IS4@5.

Now, this is the whole page, IS4@5, when he's only manipulating the web. So the real reduction is probably larger than this, but I don't know how much larger than this.

And so this says what I would have expected, and what I think the complaint in the case talks of, which is that user-side data is particularly useful for long tail queries.

To make that comparison, all the other queries, just everything that's not a long tail query, the other half of the training query set, is then getting a much smaller effect.

And so it follows exactly what you would expect, that the long tail queries are where user-side data can be

particularly valuable, because if I have a head query, a query that's occurring very often, maybe, I don't know, when I started working on this, it was Madonna, right, everybody was asking about Madonna, well, I mean, you know, then I don't have to have a whole lot of user-side data before I've seen a lot of Madonna queries. And if I see a lot more Madonna queries I'm not probably going to get a whole lot better. But if I'm seeing zero or 20, there's a big difference.

Q And, Professor Oard, can you help us contextualize what a drop in quality of that size means.

A Well, I can't say what that size is, but if we could go to the next slide.

Q Or --

A I know. If we can go to the next slide.

So this is Pandu Nayak commenting on something that I actually first saw in a different Google document. So Google talks about this quite a lot. In fact, Professor Fox mentions this in his report, that if you want to think about what an IS point is or a half an IS point or something like that, these are on a scale of zero to 100 and it's just hard to get an intuition. So internally at Google to help people talk about this with each other, they tried just taking Wikipedia out of the index. And they did this by just not retrieving any document that mentions Wikipedia.

And the effect is that it never returns a Wikipedia page, nor indeed any page that even mentions the existence of Wikipedia. And that knocks IS by down by about half a point.

Now, the first time I saw this, it wasn't the IS4 measure, it was the IS3 measure, but the IS3 measure and IS4 measure are rather similar in terms of how they score things.

And so this -- and Pandu Nayak's testimony is recent, right.

So anyway, I think half a point is the effect of removing all of Wikipedia.

And so he says, it's a pretty significant difference, right?

And so if you had -- if you measured a difference, it was about half a point. And you measured it with an experiment design that could only see some of the effects. So there's maybe more than one Wikipedia difference if you -- on long tail queries. Right?

Q Now, Professor Fox -- Professor Fox -- Professor Oard, I'd like to turn to discussing Professor Fox's final conclusion.

What does Professor Fox's final conclusion tell us about the benefits of user-side data to search quality?

A Well, I don't think it actually says anything.

If we could go to the next slide.

This is a plot that Professor Fox showed. This was from a Microsoft document that is just a sketch that says, as you get more user-side data, and the original NDCG was on the vertical axis so some measure of search quality gets better.

And when you have very little, then not only do you get better, but you keep getting better at a faster and faster rate up to some point where the rate at which you're getting better starts to slow down. Now, we call that the inflection point, and everything after that is the region of diminishing returns.

And so if you look just beyond the inflection point, you see the curve is still going up; in fact, it's going up pretty fast. And if you're in that region, you're in the region of diminishing returns. But you're not in the region of vanishing returns. You're in the region of pretty damn big returns.

And then if you go further out where it's getting flatter, the reason it's getting flatter is that we plot the amount of training data on a linear scale. So we say, Well, if you gave me a, I don't know, billion queries, here's how I'd do. And if you gave me another billion queries, here's how much better I'd do.

But if we asked instead, If I doubled the data,

how much better would I do, and then doubled it again, and then doubled it again, that's putting it on a logarithmic scale, then this line that's curving down like that would stay quite a lot longer, still going up.

So now what it says is, Well, yeah, you might have to get 10 times the amount of data, and if you're a small company, that might be possible. And if you've already got a -- most of the queries, that might not be possible. But if you had 10 times the amount of data, you'd still have returns.

So the question isn't whether you have diminishing returns, the question is whether you have vanishing returns, right?

And Professor Fox's experiment simply can't show us that. He has a plot where he shows the -- the dots all look like they're in a line, but -- but that's just because he can't measure the differences, right, all these measurement limitations of the experiment. And if he could measure the differences, maybe the -- maybe the line would look something like this. I don't know. He doesn't know.

The other point to make is that there's not really one line. I mean, if I'm trying to do head queries, there's one line. And if I'm trying to do tail queries, there's another line.

If I'm trying to do location at the scale of

United States, there is one line. If I'm trying to do it at D.C. metropolitan area, there's another line. If I'm trying to do this courthouse, there's another line. Fine-grained location gives us less and less data for each of those locations. But fine-grain location has value for serving user needs if I know what other people in that region have looked for.

And so really, it's a suite of lines, and engineers wouldn't think of it as one line. They would think of it as what they were trying to work on. Somebody trying to work on tail, they'll think of it one way. Somebody trying to work on -- I'm from College Park, so College Park is a location, that scale, they would think of it another way.

Q And, Professor Oard, did you review evidence that suggests that Google obtains a positive return to the user-side data it receives beyond Microsoft's scale?

A Yes.

I'm not actually sure you need a document to support that claim, but -- but if you go to the next slide, this is Pandu Nayak, and he's sketching out something we all know, which is that as you get more data, it costs more to keep it, and it costs even more to not just keep it but to be able to use it responsibly because we have to give responses to users very quickly. So the cost of keeping and

using this data goes up with the amount of data that we keep.

The value goes up as well. And at some point, if the value were to decline to the point where it wasn't worth the cost, people would stop doing it, and this is what he points out, that there's a sweet spot where you would stop doing it, and Google hasn't stopped doing it yet.

Q And, Professor Oard, can you provide a comparison of how much data Google uses in some of its ranking compared to Bing?

A Sure.

If we go to the next slide, it's my understanding that this is an unredacted slide, and that it's in the record that there are some systems that use 13 months of user-side data.

If Bing, which, according to Professor Fox's report, has something on the order of 5 percent of Google's user-side data, then we wouldn't be talking about 13 months of data, we'd be talking about more than 13 years of data, two decades of data.

But, really, months aren't the right unit anyway because the amount changes over time. So the amount of user-side data in a month is higher now than it was 20 years ago. So even two decades probably wouldn't be enough.

But even if it was, an awful lot's happened in the

two decades, and user-side data from two decades ago might be more harmful than helpful on some things.

And so having large amounts of user-side data lets you respond to changes in the world with good visibility and insight into what's going on in the long tail, and that's what Google's chosen to do. And it looks to me like a very good engineering decision, and it's been very well-considered.

MS. AGUILAR: Professor Oard, thank you for your testimony.

Your Honor, I have no further questions.

THE COURT: Okay. Thank you.

Do the States have any questions for Professor Oard?

MR. SALLET: No, Your Honor.

THE COURT: Why don't we go ahead and just break for lunch now instead of starting on the cross-examination. So it's 12:15 now. We'll resume at 1:15.

Professor Oard, I'll ask you to please not discuss your testimony with anyone over the break. See you then. Thank you, sir.

COURTROOM DEPUTY: All rise. This Court stands in recess.

(Recess from 12:15 p.m. to 1:15 p.m.)

C E R T I F I C A T E

I, William P. Zaremba, RMR, CRR, certify that the foregoing is a correct transcript from the record of proceedings in the above-titled matter.

Date: November 15, 2023

William P. Zaremba, RMR, CRR