What Our Google Searches Reveal About Who We Really Are

May 1, 2017 · 9:01 PM ET

SHANKAR VEDANTAM, HOST:

A note before we get started - this episode includes a racial epithet and discussions about pornography. If you have small kids with you, please save this for later.

This is HIDDEN BRAIN. I'm Shankar Vedantam. We start today's show with a personal question. Have you ever Googled something that you would never dream of saying out loud to another human being? When we have a question about something embarrassing or deeply personal, many of us today don't turn to a parent or to a friend but to our computers.

UNIDENTIFIED WOMAN #1: Because there's just some things you just can't ask a real person in real life, and you need to ask Google. (Laughter).

UNIDENTIFIED WOMAN #2: Because it's completely anonymous, and there are no judgements attached.

UNIDENTIFIED MAN: Well, Google knows everything.

UNIDENTIFIED WOMAN #1: I agree to that.

VEDANTAM: Every time we type into a search box, we reveal something about ourselves. As millions of us look for answers to questions or things to buy or places to

meet friends, our searches produce a map of our collective hopes, fears and desires.

SETH STEPHENS-DAVIDOWITZ: You do learn a lot about people that's very, very different from what they say and kind of the weirdness at the heart of the human psyche that doesn't really reveal itself in everyday life or at lunch tables. But it does reveal itself at 2 a.m. on PornHub.

VEDANTAM: Today, on HIDDEN BRAIN, what big data knows about our deepest thoughts and secrets.

(SOUNDBITE OF MUSIC)

VEDANTAM: My guest today is Seth Stephens-Davidowitz. He used to be a data scientist at Google. And he's the author of the book, "Everybody Lies: Big Data, New Data And What The Internet Can Tell Us About Who We Really Are." Seth, welcome to HIDDEN BRAIN.

STEPHENS-DAVIDOWITZ: Thanks so much for having me, Shankar.

VEDANTAM: So Seth, we all know that Google handles billions of searches every day. But one of the insights you've had is that the reason Google knows a lot about us is not just because of the volume of search terms but because people turn to Google as they might turn to a friend or a confidant.

STEPHENS-DAVIDOWITZ: That's exactly right. I think there's something very comforting about that little white box that people feel very comfortable telling things that they may not tell anybody else about their sexual interests, their health problems, their insecurities. And using this anonymous aggregate data, we can learn a lot more about people than we've really ever known.

VEDANTAM: And one of the ways we can learn a lot more about people is through these very strange correlations. You find, for example, there's a relationship between the unemployment rate and the kinds of searches people make online.

STEPHENS-DAVIDOWITZ: Yeah, I was looking at what searches correlate most with the unemployment rate. And I was expecting something like new jobs or unemployment benefits. But during the time period I looked at, the single search that was most highly correlated with the unemployment rate was Slutload, which is a pornography site. And you can imagine that if a lot of people are out of work and they have nothing to do during the day, they may be more likely to look at porn sites.

Another search that was high on the list was solitaire. So again, when people are out of work, they're kind of bored. They do leisure activities. And potentially, this measure of how much leisure there is on the Internet may help us know how many people are out of work on a given day.

VEDANTAM: And, of course, this sort of helps us reconsider what we think of as data. So when we think about the unemployment rate, as you say, you know, our normal approach is to say, how many people are still in jobs? Let's track down all the jobs. This is coming at the question entirely differently.

STEPHENS-DAVIDOWITZ: Yeah, I think the traditional way to collect data was to send a survey out to people and have them answer questions, check boxes. There are lots of problems with this approach. Many people don't answer surveys, and many people lie to surveys. So the new era of data is kind of looking through all the clues that we leave, many of them not as part of questions or as part of surveys but just clues we leave as we go through our lives.

VEDANTAM: One of the important differences between mining this kind of data and the responses we get on surveys has to do with how people report their sexual orientation. I understand that the kind of queries that you see on Google might reveal something quite different than if you ask people if they're gay.

STEPHENS-DAVIDOWITZ: That's right. If you ask people in surveys today in the United States, about 2.5 or 3 percent of men say that they're primarily attracted to men. And this number is far higher in certain states where tolerance to homosexuality is greater. So there are a lot more gay men, according to surveys, in California than in

Mississippi.

But if you look at search data for gay male pornography, it's a tiny bit higher in California but not that much higher. And overall, about 5 percent of male pornography search is for gay porn, so almost twice as high as the numbers you get in surveys.

VEDANTAM: Your research has important implications for a topic that we've looked at a lot on HIDDEN BRAIN - the topic of implicit bias. People aren't always aware of the biases they hold. And so scientists have had to find clever ways to unearth these biases. You think that Google searches can reveal some forms of implicit bias?

STEPHENS-DAVIDOWITZ: That's right. So one I look at is the questions that parents have about their children. If you ask many parents today, they would say that they treat their sons and daughters equally, that they're equally excited about their intellectual potential, equally concerned about maybe their weight problems.

But if you aggregate everybody's Google searches, you see large differences in gender - that when parents in the United States ask questions starting, is my son, they're much more likely to use words such as gifted or a genius than they would in a search starting, is my daughter. When parents in the United States search is my daughter, they're much more likely to complete it with is my daughter overweight or is my daughter ugly. So parents are much more excited about the intellectual potential of their sons and much more concerned about the physical appearance of their daughters.

VEDANTAM: You report that in some states, after Barack Obama was elected president, there were more Google searches for a certain racist term than searches for first black president.

STEPHENS-DAVIDOWITZ: I think there is a disturbing element to some of this search data, where in the United States today, many people - and maybe this is a good thing - don't feel comfortable sharing that they have racist thoughts or racist feelings. But on Google, they do make these searches in strikingly high frequency. I need to use

sordid language to this. The measure is the percent of Google searches that include the word nigger. And these searches are predominately searches looking for jokes mocking African-Americans.

I should clarify. This is not searches for rap lyrics, which tend to use the word nigga (ph), ending in A. But if you look at the racist search volumes, I think if you had asked me based on everything I had read about racism in the United States, I would have thought that racism in the United States predominately concentrated in the South, that really the big divide of the United States, when it comes to racism, is South versus North.

But the Google data reveals that's not really the case, that racism is actually very, very high in many places in the North, places like western Pennsylvania or eastern Ohio or industrial Michigan or rural Illinois or upstate New York. The real divide these days when it comes to racism is not north versus south. It's east versus west. There's much higher racism east of the Mississippi than west of the Mississippi.

content removed here (i.e., this documented was edited here) for use in PSY 532 course

VEDANTAM: You spend a lot of time in the book talking about sex. It turns out to be an area where marketers and companies know that what we say about ourselves is nowhere close to the truth. Most people report being not interested in pornography. But the website PornHub reports that in 2015 alone, viewers watched 2 and a half billion hours of porn, which is apparently longer than the entire amount of time that humans have been on Earth. What does this say about us - the fact that we either have very little insight about ourselves or we're actually lying through our teeth?

STEPHENS-DAVIDOWITZ: Yeah, I'd say we're probably lying through your teeth. Yeah, I'd say that - I do talk a lot about sex in this book. One thing I like to say is that big data is so powerful, it turned me into a sex expert because it wasn't a natural area of expertise for me. But I do talk a lot about sexuality, and I think you do learn a lot about people that's very, very different from what they say and kind of the weirdness

at the heart of the human psyche that doesn't really reveal itself in everyday life or at lunch tables. But it does reveal itself at 2 a.m. on PornHub.

VEDANTAM: One of the things that I was wondering about as I read your book was how much search terms tell us about what people are actually thinking or actually feeling and how much they might just tell us about things that people are curious about. So, you know, certainly people search for a lot of things related to sex that would indicate that there is a large amount of interest in, you know, sadomasochism and fetishes and so forth. But could some of it just be that people are curious? People hear a lot about this in the news or on social media, and they Google something because they're just curious about it, not necessarily because they themselves want to, you know, be part of the BDSM culture.

STEPHENS-DAVIDOWITZ: I think it depends on the particular question you're looking at. So the reasons we can trust the racism data as meaningful is because it correlates with voting patterns. With the sex data, there's not really necessarily something to check it against. On the internet, we do see the videos that people watch, and I think that is pretty telling about some people's fantasies, even if it's not definitive because some people may just be curious.

VEDANTAM: Pornography sites aren't the only ones gathering information about our sexual and romantic preferences. We now have apps like Tinder and sites like OKCupid that gather tons of data about us. As a result, these apps and sites know a lot about our romantic preferences. But for a long time, we've had a human version of big data for romance, grandma.

Seth has some personal experience with this big data source. A couple of years ago, he was having Thanksgiving dinner with his family. He was 33, didn't have a date with him, and his family was trying to figure out the qualities Seth needed in a romantic partner.

STEPHENS-DAVIDOWITZ: My family was going back and forth. My sister was saying that I need a wacky girl because I'm wacky. My brother was saying that my sister was

wacky, that I need a normal girl to balance me out. And my mom was screaming at my brother and sister that I'm not wacky, and my dad was then screaming at my mom that, of course, Seth is wacky.

VEDANTAM: (Laughter).

STEPHENS-DAVIDOWITZ: So it's kind of a classic Stephens-Davidowitz. It's a family Thanksgiving where everyone's just yelling at each other for being wacky, and we're not really getting any progress in learning about what I need in my love life. And then my soft spoken 88-year-old grandma started to speak, and everyone went quiet. And she explained to me that I need a nice girl, not too pretty, very smart, good with people, social so you would do things, sense of humor because you have a good sense of humor.

And I describe why was her advice so much better than everybody else's? I think one of the reasons that she's big data, right? So grandmas and grandpas throughout history have had access to more data points than anybody else, and they've been able to correlate larger patterns than anybody else has because they've been around longer. And that's why they've been such an important source of wisdom historically.

VEDANTAM: The problem, of course, as you also point out is that it's very hard to disentangle your personal experiences from what actually happens in the world. And in your grandmother's case, she actually had a very specific piece of relationship advice about the kind of person you should want, and some of that might not actually be backed up by the empirical evidence.

STEPHENS-DAVIDOWITZ: Yeah. Well, my grandma told me - has told me on multiple occasions that it's important to have a common set of friends in a partner, so she lived in a small apartment in Queens, N.Y. with my grandfather. And every evening they'd go outside and gossip with their neighbors, and she thought that was a big part in why their relationship worked. But actually recently computer scientists have analyzed data from Facebook, and they can actually look when people are in relationships and when they're out of relationships and try to predict what factors a

relationship make it more likely to last.

One of the things they tested was having a common group of friends. Some partners on Facebook share pretty much the same friend group and some people have totally isolate friend groups and they found contrary to my grandmother's advice that having a separate social circle is actually a positive predictor of a relationship lasting.

VEDANTAM: And so, of course, the risk of trusting the individual is that the individual's intuition about what worked for his or her life might not work for everyone else.

STEPHENS-DAVIDOWITZ: That's right. I think we tend to get biased by our own situation. Data scientists have a phrase called weighting data. Some data points get extra weight in our models, and our intuition gives too much weight to our own experience. And we tend to assume that what worked for us will work for others as well, and that's frequently not the case.

VEDANTAM: Many companies know that we don't really understand ourselves. When we come back, we look at how companies are using big data to predict what we're going to do before we know it ourselves. We'll also ask if sites like Google can use data to forecast whether you're going to get a serious illness. Should they give you that information? Stay with us.

(SOUNDBITE OF MUSIC)

VEDANTAM: This is HIDDEN BRAIN. I'm Shankar Vedantam. Netflix used to ask users what kind of movies they wanted to watch. Seth Stephens-Davidowitz says eventually the company realized that asking this kind of question was a complete waste of time.

STEPHENS-DAVIDOWITZ: Yeah. Initially, Netflix would ask people what they want to view in the future, so they queue up the movies that they said, and if you ask people what are you going to want to watch tomorrow or this weekend, people are very aspirational. They want to watch documentaries or about World War II or avant-garde

French films. But then when Saturday or Sunday comes around, they want to watch the same lowbrow comedies that they've always watched. So Netflix realized they had to just ignore what people told them and use their algorithms to figure out what they'd actually want to watch.

VEDANTAM: So one of the things that's intriguing about what you just said is it - I don't think it's actually the case that people were lying to Netflix when they said they wanted to watch the avant-garde film. They actually genuinely probably aspire to do that. It might actually be that big data understands people better than they understand themselves.

STEPHENS-DAVIDOWITZ: Yeah. Probably even more common than lying to other people is lying to ourselves, particularly when we're trying to predict what we're going to do in two or three days. We tend to assume that we're going to go to the gym more than we go to the gym or eat better than we actually will eat or watch more intellectual stuff than we actually will watch. So the algorithms can correct for this overoptimism that we all tend to share.

VEDANTAM: When you look at a company like Facebook which has access to these, you know, huge amounts of data about us and what we like and whom we like and our relationships, you have to wonder how the company is using this data in all kinds of different ways. I remember Facebook got into some hot water a couple of years ago because they ran an experiment that seemed to be manipulating how people feel. And, of course, there was a huge outcry about the experiment at the time. And since then, there hasn't been very much reported about what Facebook is doing, but I suspect that it might just be because Facebook is no longer telling us what it's doing, but it's still doing it anyway.

STEPHENS-DAVIDOWITZ: Every major tech company now runs lots and lots of what are called A/B tests, which are little experiments where you put people into two different groups - treatment and control group, and you show one group one version of your site and the other group another version of the site. And you see which version gets the most clicks or the most views. This has really exploded in the tech industry.

VEDANTAM: It's not just the tech industry that uses A/B testing. Newspapers do, too, newspapers like The Boston Globe. A few years ago, The Globe tried out two different headlines for the same story, and then measured which headline got the most clicks. The newspaper then used the more effective headline for the rest of the day. I've been a journalist for about 25 years and spent most of that time working at newspapers. Seth wanted to test my headline-writing expertise. He read out two versions of a headline for a Boston Globe story, and he asked me to guess which one worked better.

STEPHENS-DAVIDOWITZ: So let's see, Shankar, if you can guess some of these winners. The first headline test - I'll give you headline A first and then headline B second - headline A - "When The First Subway Opened In Boston," headline B - "Cartoons From When The First Subway Opened In Boston."

VEDANTAM: All right. That's going to be easy, "Cartoons From When The First Subway Opened In Boston."

STEPHENS-DAVIDOWITZ: No, it's headline A. It got 33 percent more clicks for "When The First Subway Opened In Boston"

VEDANTAM: Oh, no.

STEPHENS-DAVIDOWITZ: You want another one?

VEDANTAM: Yeah. Let's try it. I know where this is going, but let's try it.

STEPHENS-DAVIDOWITZ: (Laughter) OK. Headline A is "Woman Makes Bank Off Rare Baseball Card" and headline B is "Woman Makes $179,000 Off Rare Baseball Card."

VEDANTAM: I'm going to go with the specific dollar amount so B.

STEPHENS-DAVIDOWITZ: No, it's headline A, 38 percent more clicks for headline A. You're zero for two.

VEDANTAM: Is there a third one? Can I redeem myself?

STEPHENS-DAVIDOWITZ: Yeah. OK. Let me - let's do another one. All right. Headline A - "Hook Up Contest At Heart Of St. Paul Rape Trial," headline B "No Charges In Prep School Sex Scandal."

VEDANTAM: All right. So I'm going to follow a completely different strategy than I did the last two times which is I'm just going to pick a number, and I picked the number before you even read it - read that (laughter)...

STEPHENS-DAVIDOWITZ: OK.

VEDANTAM: ...To prevent myself from being biased. I'm going to go with B again just on the off chance that you couldn't tell me three answers where all the answers were A.

STEPHENS-DAVIDOWITZ: That's right. I didn't even realize I was doing that, but headline B is correct - 108 percent more clicks for headline B. So good job. You got it one for three, not so bad.

VEDANTAM: And the interesting thing, of course, is I used sort of an algorithmic solution...

STEPHENS-DAVIDOWITZ: Yeah. You outsmarted me, I guess. Right? Yeah. So I think what this shows is that the reason that A/B testing is so important is because our intuition can trick us, that you've been around journalism for many, many years, and you have your own ideas of what makes a successful headline. But even someone like you is frequently wrong. And we can use A/B testing to correct our faulty intuition, find what actually works, not what we think works.

VEDANTAM: It's one thing when companies use big data to serve us better. You could argue that a newspaper that delivers the catchier headline is serving its audience better, but there are many, many instances where companies are now using big data against us. Banks and other financial institutions are using clues from big data to decide who should get a loan.

STEPHENS-DAVIDOWITZ: I think it's an area of big concern. So I talk about a study

in the book where they cite a peer-to-peer lending site, and they cite the text that people used in their requests for loans. And you can figure out just from what people say in their loans how likely they are to pay back. And there are some strange correlations. For example, if you mentioned the word God, your 2.2 times less likely to pay back, 2.2 times more likely to default.

And this does get eerie. Are you really supposed to be penalized if you mention God in a loan application? That would seem to be really wrong, even evil - right? - to penalize somebody for religious preference. Basically everything is correlated with everything, right? So just about anything anybody does is going to have some predictive power for other things they do. And the legal system is really not set up for a world in which companies potentially can mine correlations over just about everything anybody does in their life.

VEDANTAM: I was thinking about an ethical issue. I'm not sure if necessarily this is a legal issue. But you mention in the book that, you know, if someone is Googling I've been diagnosed with pancreatic cancer, what should I do? It's reasonable to assume that this person has been diagnosed with pancreatic cancer. But if you collect all of the people who are Googling what to do about that diagnosis with pancreatic cancer and then work backwards to see what they have been searching for in the weeks and months prior to their diagnosis, you can discover some pretty amazing things.

STEPHENS-DAVIDOWITZ: Yeah. This is a study that - researchers used Microsoft Bing data. They looked at people who searched for just diagnosed with pancreatic cancer, and then similar people who never made such a search, and then they looked at all the health symptoms they had made in the lead-up to either a diagnosis or no diagnosis. And they found that there were very, very clear patterns of symptoms that were far more likely to suggest a future diagnosis of pancreatic cancer.

For example, they found that searching for indigestion and then abdominal pain was evidence of pancreatic cancer, while searching for just indigestion without abdominal pain meant a person was much more unlikely to have pancreatic cancer. And that's a really, really subtle pattern in symptoms, right? Like, a time series of one symptom

followed by another symptom is evidence of a potential disease. It really shows, I think, the power of this data, where you can really tease out very subtle patterns in symptoms and figure out which ones are potentially threatening and which ones are benign.

VEDANTAM: So here's the ethical question. Once you established that there is this correlation that you sort of say I have a universe of people who clearly have pancreatic cancer, and I work backwards through their search history. And I detect these patterns that no one had thought to look at before that say these particular kinds of search terms seem to be correlated with people who go on to have the diagnosis versus these search terms that do not go on to predict a diagnosis.

So does a company like Microsoft now have an obligation to tell people who are Googling for these combinations of search terms, look, you might actually need to get checked out? You might actually need to go see a doctor? Because, of course, if you can be diagnosed with pancreatic cancer, you know, four weeks earlier, you have a much better chance of survival than if you have to wait for a month.

STEPHENS-DAVIDOWITZ: I lean in the direction of, yes, some people would not lean that direction. It could be a little creepy if Google - right below the button I Feel Lucky, you know, I have - you may have pancreatic cancer. It's not exactly the most friendly thing to see on a website, but, personally, if I had some sort of symptom pattern that suggests that I may have a disease, and there was a chance of curing it if I was told, I'd want to know that. It's just another example that really the ethical and legal framework that we've set up is not necessarily prepared for big data.

VEDANTAM: Seth Stephens-Davidowitz is a former data scientist at Google and the author of the book "Everybody Lies: Big Data, New Data, And What The Internet Can Tell Us About Who We Really Are." Seth, thank you for joining me today on HIDDEN BRAIN.

STEPHENS-DAVIDOWITZ: Thanks so much for having me, Shankar.

VEDANTAM: This week's episode was produced by Rhaina Cohen and edited by Tara Boyle. Our staff includes Jenny Schmidt, Maggie Penman and Renee Klahr. Our unsung hero this week is Hugo Rojo. Hugo works on NPR's media relations team, and he's one of those people who's always willing to be helpful. Hugo helps us with social media for the show. He's also our in-house professional photographer. When we need a producer to record a line of narration in Spanish, Hugo puts up his hand. He's had some terrific ideas on how to reach new listeners, and he's always willing to share those ideas with us. Thanks, Hugo.

Copyright © 2017 NPR. All rights reserved. Visit our website terms of use and permissions pages at www.npr.org for further information.

NPR transcripts are created on a rush deadline by Verb8tm, Inc., an NPR contractor, and produced using a proprietary transcription process developed with NPR. This text may not be in its final form and may be updated or revised in the future. Accuracy and availability may vary. The authoritative record of NPR's programming is the audio record.