The 17 Year Old Who Correlated Domain Names to High Search Rankings - With Mark Collier

Watch the full video at:
http://www.domainsherpa.com/mark-collier-theopenalgorithm-interview/

Three messages before today's interview educates and motivates you.

First, if you’re a domain name investor, don’t you have unique legal needs that require domain name technical know-how and industry experience? That’s why you need David Weslow of Wiley Rein. Go search for David Weslow on DomainSherpa, watch his interview and you can see for yourself that he can clearly explain issues, can help you with buy/sell agreements, deal with website content issues and UDRP actions, and even help you write your website terms and conditions. David Weslow is the lawyer to call for Internet legal issues. See for yourself at NewMediaIP.com.

Second, managing multiple domain name marketplace and auction site accounts is a pain. Inevitably, you forget to sign into one and lose a great domain…or worse. Now imagine using a single, simple-to-use and comprehensive control panel to manage all your accounts. That’s Protrada. You can setup search filters, analyze domains, automate bidding, list domains for sale, and buy domains across all major marketplaces. Protrada also has a new semantic engine that builds Google-friendly websites with rich content and network feeds. Sign up at Protrada.com to get 20 free credits and start building and monetizing your domains today.

Finally, if you have questions about domain names, where should you go to ask them? The answer is DNForum.com. Not only is DN Forum the largest domain name forum in the world, but it's the best. You can learn about domain names and the industry, buy and sell domain names, talk about domain name news, and meet other domainers just like yourself. Register for a free DN Forum account and begin advancing your skills and knowledge today. And when you do signup, send me a friend request so we can connect.

Here's your program.

Michael Cyger: Hey everyone. My name is Michael Cyger, and I'm the Publisher of DomainSherpa.com - the website where you come to learn how to become a successful domain name entrepreneur and investor directly from the experts.

How important is a generic keyword rich domain name to ranking high in the search engines? It's a question that I've asked Danny Sullivan, a journal and worldwide-recognized expert in search engines in a past DomainSherpa.com interview. His response: "If you can come up with the good content to back up the good domain name, that's very, very golden". But I'm a data guy. I want to know that there's real data behind what everyone believes to be true. So, when today's guest reached out to me, I jumped at the opportunity to learn more about what he's doing.

Today we're speaking to Mark Collier, blogger and Chief Data Cruncher at TheOpenAlgorithm.com. Mark, welcome to the show.

Mark Collier: Hi. How are you doing?

Michael: Great. Let's start the show with the bottom line upfront for the audience, Mark. What is the correlation between and exact match domain name and search engine rankings at a high level?

Mark: Okay. So, at the highest level, I'll just let your users know what I did. So I did an analysis of over ten thousand search results. I downloaded the top one hundred results for all those ten thousand and then, I tested a whole bunch of things using Spearman's correlation coefficient. So, for your users, that basically means the results I get back will a number between -1 and 1. And the closer to either one means the stronger the correlation; and usually, something about 0.1 is significant enough correlation. So, over zero - between zero and one - is a positive correlation. Two things move together. And between 0 and -1 is a negative correlation. They don't move together at all. So, for exact match domains, I got a correlation of 0.14, which is rather significant. I tested a hundred and sixty factors. I don't have the list right here, but I'd say it would probably be around number forty. And that forty would be the thirty or thirty-five above that would all be link-based. So, essentially, beyond links, it was probably in the top one, two, or three, so it's fairly significant.

Michael: Wow.

Mark: And then I went a little bit deeper for your users. Exact match .COMs were 0.18 and then, .ORGs were 0.95, .NETs were 0.08, and .INFO and .US had pretty much no correlation. They were just randomly correlated at 0.02. So, obviously, for your users, it is pretty important getting exact match domain. It will help your search engine ranking. And obviously .COMs or .ORGs are definitely the highest correlated. Yeah, but we can dig a bit deeper into that if you want.

Michael: Yeah. So, you just said a whole bunch of information that is really interesting to me and probably to a lot of people who are watching the show that want to know that there is data behind this. So, basically, you said that anything greater than 0.1 is statistically significant. So, you are using the Spearman's coefficient algorithm to actually determine statistical validity of an exact match domain name. So I'm going to dig into exactly how that works and how you measure it so that people just get a better idea of how the analysis is being done. But, first, let's take a step back and learn about why you are doing all of this analysis. When somebody goes to TheOpenAlgorithm.com and looks at The Open Algorithm project, what is it? Can you describe, at a high level, what that is for other people and why you are doing this?

Mark: Sure. So, in essence, I just started a project, and we can get into why a bit later. I just started a project basically to bring Science data-driven knowledge to search engine optimization and then, started (Unclear 4:25.8) marketing and online marketing in general; and basically, that is what the project does. I mean our first study - my first study - is a correlation study. It's pretty easy to run, You just gather data and get the correlations. And then, in the future, we are going to run more studies maybe, hopefully, proving causation and then, obviously, just a few more interesting tests that I sort of have in my mind to run. And we'll probably be running a second correlation study in the coming month. But essentially, it is just a project to bring more data and more Science to search engine optimization because I basically found that a lot of what everybody thought was truths, or sure things, or whatever were not really true and there is a whole lot of bad advice going out there. So, I just thought, 'we'll bring data to it and data usually does not lie'.

Michael: Right. Exactly. What is that famous quote? The biggest lies of statistic. I can't remember it. I'm going to have to post in the comments. Darn it. All right. So you chose TheOpenAlgorithm.com as your domain name. Why did you use the keyword "algorithm"? What is that significant?

Mark: Well, for anybody who knows anything about search engines or how they work, basically, they work on an algorithm. They all have some basic algorithm. And algorithm is just a math formula. So, obviously, you don't have sort of elves and Google figuring out what to supply you with; it is a computer based on a math formula. And the math formula works on trillions, billions, or whatever of URLs and then it picks out the top ten URLs for you to see when you put in a search. So, for search engine marketers, what they want is, they want to understand the algorithm. How to utilize their knowledge of the algorithm to get their sites, pages, and whatever to rank higher in Google and get more traffic, and all that kind of stuff. So, obviously, the Algorithm makes sense. TheOpen part is obviously because we are trying to open it. We are trying to provide more information to the public on how it actually works; not how people think it works. And also, then, hopefully, find out things that people did not know about. So that's really why I chose the domain name; because it is about an algorithm. It is about opening it up.

Michael: Yeah. And so, you supply all of your analysis and then, from many of the posts that I read on the website, you actually supply the data set as well so people can go out and do their own analysis on that; and you are actually encouraging them to come back and share their results in the same fashion that you do.

Mark: Yes. So, in the first study, I did not actually provide that much data to people. I provided some data. So, for example, I provided all the domains we analyzed and other data like how many times they showed up and stuff like that. And then, obviously because it's a free project and it's there for everybody to see, pretty much all of the data is totally transparent; but then, in the next study, we are actually going to go a lot further in that. I did not use very robust database software in my first study. I have upgraded to SQL now and I am going to make the whole database with whatever millions of data points open to everybody to see. And then also, I have only been programming for seven to eight months and I was not a great programmer when I first wrote the software to do the first correlation study, so I have totally rewritten all that. That is what I have spent my summer holidays doing and I am going to make it totally open source. So, hopefully in the second study, we are going to provide a lot more data to people so they can actually go out and run the tests. The first one, we did provide some data and, obviously, all the results were totally transparent and how we came to the results; but in the next one, we are definitely going to provide more data and also, allow people to dig into the software. If they want to use the software that I used to extract the domain from a URL or some other tool, let them use it. That's fine.

Michael: Wow. All right. So why did you start The Open Algorithm project?

Mark: Well, basically, I was interested in search engines and I knew quite a bit about them, or I thought I knew quite a bit about them, and I have done a lot of research on them and tested out a few of my own websites; and I just saw that there was a need for data-driven and Science-driven knowledge about search engines. And I saw forums with absolutely terrible, terrible advice on them, and I saw blogs and so-called experts making assertions that might be true, but they weren't backing it up. They did not have any evidence. And any evidence there was out there was sort of anecdotal evidence; like, 'oh this happened with one or two pages I built', but that is not really Scientifically significant and what if it is not happening for everybody else? So, I came across SEOmoz. SEOmoz were the first to do a correlation study and they have been really helpful with my correlation study; and they did a correlation study slightly smaller scale to mine, but still Scientifically significant and still testing over a hundred factors. Very good correlation study. And people seemed to like it. People seemed to have a need for that kind of blog post; that kind of information, so I decided to go and do it. And that is one of the core things about Science; is that it can be backed up time and time again. So, it's not good enough that SEOmoz is doing a correlation study once and maybe they do it again themselves. If other people can come along and back it up further, better, provide more data, and all sorts of stuff like that, I mean that is good for the industry as a whole.

Michael: Definitely. So I want to step back and understand a little bit more about your background, Mark. Just a moment ago you mentioned summer holiday. When you first sent me an e-mail, I went to your website and looked at your picture on the About Page and thought to myself, "What's the deal here? This thirty-something year old that understand statistical analysis, and search engine rankings, and search engine optimization looks like he is a teenager, or he is using a picture from grade school". How old are you, Mark?

Mark: I am seventeen. I guess that picture you saw of me on the About Page was probably a couple years ago.

Michael: It was probably like last week. So you are seventeen years old. I cannot believe that a person with your maturity and your understanding of all these topics - and people can get the sense from it just in the past five minutes that we have spent together - are able to discuss these topics in such detail. It's unbelievable. I think, when I was seventeen, I was focused on my Atari 400, figuring out how to ask a girl to a date, and - I don't know - maybe writing my bike to the local convenient store to buy some bubble gum. You, on the other hand, have got a fully functional blog. You have got phenomenal technical content. You have got video shows better than I have read from people that are twice or three times your age. And you have got a mission. Have you always been this way?

Mark: I don't know. It is a pain to talk about yourself and what way you've been, but yeah. I have always been interested in sort of doing other stuff like I played football, and cricket, and stuff like that, but I also always used to do quirky stuff, so I was always interested in the business and I read loads of books. When I was eight, nine, ten, I was always interested in business, going out, trying things and little projects. When I was thirteen, I started my first website. Absolutely dreadful website, but I was just interested in it. Interested in code and that kind of stuff. So, it is a pain to talk about yourself and what way you've been and all that kind of stuff, but I have always been interested in quirky stuff and things like that.

Michael: Have you taken computer programming in school or is everything that you have done so far on your websites, in the past and TheOpenAlgorithm.com, just self-taught programming?

Mark: All self-taught. Actually, that is one of the things that sort of infuriate me about the education system in Ireland. There is no option for Computer Science. The furthest you can go is sometimes schools offer an elective of computers, but that goes as far as teaching you Microsoft Word and Microsoft Excel. So we have no Computer Science in any school level in Ireland. I don't know if you guys have heard about CoderDojo. It's an Irish program. It only started up last year. These guys, James Whelton and Bill Liao were pretty annoyed that there was no Computer Science. And I think Jeames Welton is nineteen or twenty, or something like that, and he just set up a Computer Programming Class. It is just a meet up really. And it just grew with no money, no funding, or anything like that. No Government backing. They have literally grown to, I think, thirteen countries and hundreds of clubs. And you just have like seven, eight, nine, and ten year olds teaching themselves how to code; and after like two months in the program, they are teaching seven, eight, nine, and year olds. So there definitely is a need for it, but no, all my Computer Science Programming and that kind of stuff is self-taught. Saying that, I am not a brilliant programmer. I only know Python. I have only been programming for seven or eight months. I can get quite a bit done and I don't know anything about achievement levels or anything like that, or programming quality. I've never programmed in a group, but I think I am an okay programmer; but in saying that, I'm not a technical genius or anything like that.

Michael: Well, my differentiation between a great programmer and an okay programmer is the code works. I'm not one of these guys that gets caught up in how beautiful the code is, or how simple the code is. You know, it takes somebody a month to write the most elegant, simplified code possible that I could write in ten times as many lines in a day; I'll take my ten times as many lines any day, so you don't need to worry about that with me and I do not think the with Domain Sherpa audience. But I should point out to the audience that clearly, you have an accent. You have mentioned Ireland. You are actually calling in from Dublin right now and it is what time in Dublin?

Mark: It's just about eight o' clock right now.

Michael: 8PM. So I appreciate you calling in and making yourself available in the evening time period.

Mark: As you might have guessed, there are no office hours for me. This is just my bedroom behind me, so no, there is no sort of nine to five for me. It is just whenever I feel like doing stuff.

Michael: Excellent. So, how many domain names do you own besides TheOpenAlgorithm.com?

Mark: I probably own around eighty or something like that. I sort of dabbled a bit for a while, so I own about eighty. Most of them are dreadful. Most of them are just end user things like, for me, I sort of have some plans that I want to go build a website here or there. And then some of them I bought because I thought they were good buys like I think I own HowToMakeChocolateCake.com or HowToMakeScones.com. And they were all bought because they are exact match domain based on how many searches there were on them per month. But I definitely would not call myself a domainer or a domain name expert. I've literally only been interested in buy domains for investment purposes maybe in the last month. So, mostly, my experience is in how search engines work and then, obviously, a part of that is how domains work in relation to search engines; how URLs work in relation to search engines. But as an investor, not really many.

Michael: Excellent. All right. And I watched one of your videos on TheOpenAlgorithm.com where you stated you had some clients that you perform SEO for. Do you have a consulting business or do you just work with a certain amount of clients to get pocket money for buying domain names and other things that seventeen year olds want? How does that work?

Mark: Yeah. Essentially, I do not do much consulting and I do not have a fully limited company or anything like that. I do dabble the odd bit for clients. I used to do a bit more of it, but at the minute, I am sort of focused on more of the data sides of thing. The high level thing. Providing people with data. So, I do have a few clients and I work for them, and it goes fine; but I am not big into the consulting side. I prefer the more scalable stuff like, hopefully building software; but at the minute, it's obviously the more scalable stuff of getting data, providing data, and providing it on a scalable platform like blogging where anybody can see it and as many people as they want can see it.

Michael: Smart. All right. So let's get back in the factors. How many factors do the major search engines - let's just say Google and Bing - look at when evaluating whether a website or a specific page ranks at the top of the search engine results?

Mark: So, the actual number is private. Loads of people will tell you the answer. Some people tell you two hundred. Some people tell you ten thousand. So, the major theory is essentially that there is two hundred sort of high level factors. So, for example, if you take page rank. A lot of your users have probably heard about page rank. Page rank is essentially just about the quality and the number of links pointing to your site, and Google will have a whole really complicated algorithm deciding how many links are pointing to your site; the quality of them; the relevance of them; all sorts of stuff like that. So they would essentially take that as one high level factor and then, basically, there are a whole bunch of lower level factors that all feed into a higher-level factor. So the general theory, or based on sort of anecdotal evidence and people saying it just in passing, is there is around ten thousand. Now, there could be a hundred thousand; there could be thousand. Nobody actually really knows. I'd say most of the engineers in Google don't know because they are all sort of marginalized and kept to their own devices. But, essentially, there is a load of factors; but there are key high level factors that are really important and there are some ones that aren't important and then, obviously, there's personalized search. And the search engine industry, as a whole, has just gone so much further than when it started when it was just providing generalized searches irregardless of where you were, irregardless of who you are, and irregardless of all sorts of other factors. So they have impacted things like personalization, localization, geographical targeting and all that that kind of stuff, so there are thousands of factors; but you can say around ten thousand to be safe, but the actual number probably does not matter to anybody.

Michael: Yeah. So, Danny Sullivan of SearchEngineLand.com came up with a graphic entitled The Periodic Table of SEO Ranking Factors. And in that period table, which looks like a periodic table, there are only thirty factors. How do you correlate those thirty factors with, say, the two hundred stated factors or implied factors from the search engines?

Mark: Yeah. So, I have actually seen that period table, and that would definitely be based on anecdotal evidence. I would have said that some of the factors you've got in there are probably a bit dubious, or certainly not worthy of being the top thirty or something like that; but essentially, it has got, in there, all the sort of commonly held theories and some of them are myths. The commonly held idea of what is important to search engine ranking. So, the idea there, or the top process areas that all those ones are heavily correlated and then there are sort of smaller ones that we do not really need to know about. But obviously he is in the news game as apposed to the big data game, and he does it really well, and they have a fantastic site. And that is probably a bit of marketing there. Having a period table that people can reference to, and it is an info-graphic, and all that kind of stuff.

Michael: Right.

Mark: So I wouldn't say that all of them correlate that well. In saying that, some of them will correlate well. So, for example, he has social factors on there. I tested whether Google+ links correlate well and I think they correlated at something like 0.26, so that was a fairly significant correlation. And then, obviously, he has all the link stuff on there, so links correlated at the absolute highest. I think one of SEOmoz's -- I think it was page thirty. They have an algorithm that analyzes links similar to page rank or sort of modeling page rank. Now, I think that correlated at 0.38, and that was, by far, the largest correlation in all of around thirty link-related factors, like internal links, external links, no-follow links, followed links. Every type of link factor I tested all correlated really well, so links definitely are important. But in saying that as thirty factors, I don't have the list of them here. I haven't seen it in a while, but I would say a lot of the on-page stuff actually did not correlate well at all. So, I do have a post on my site. The URL is probably something like TheOpenAlgorithm.com/Correlation-Data/On-Page-Factors, but you can just look up on the site if you wanted. That basically showed that every single commonly held theory about on-page factors that have been spouted for the last fifteen years, according to the correlation data, is pretty much rubbish. And then I tested in-page content, which is like actually when you go and request a webpage as a programmer, you get back the HTML; and then extracting the actual content from that is actually quite a difficult task as a programmer. As a user, it is really for you to separate what is a heading, what is done the side to side bar, what is advertising and all that kind of stuff.

Michael: Like visually, you can easily see because it is the center area - the biggest area.

Mark: Yes.

Michael: But in HTML, it is hard to determine that.

Mark: Exactly. In HTML, it is actually a really difficult task. If you go on to my Google+ Page, you can actually see where Google actually fails at doing that because what they do is, you know when you link to something in Google+ and you put in the link and then, it extracts like the first twenty things or the first hundred characters, or something like that on the page. Well, on a lot of my posts, it puts in Tweet Tweet as the first two words in my posts, but they are not the first two words in my post. That is, I think, from a Twitter Plugin I use on my blog.

Michael: Right.

Mark: So, it is a really, really difficult problem to solve as a programmer, so I used (Unclear 22:28.1) technology. They are a new technology company. They come out of Stanford. Mike (Unclear 22:33.3) is the CEO. And I think they just raised like two million funding round, and he gave me free access to their API. And I used that to analyze the in-content factors, so you think, 'okay. Well, all that bold stuff. Headings; title tags; descriptions; keyword tags. Okay. That is all HTML. That is all manipulatable'. That might have been a commonly held theory in 1998, or 2000/2001. We all know it has moved on from then. Google have really complicated algorithms that know what is content, and what is in content, and what is good content, and all that kind of stuff. So, I tested the in-content stuff, and that did not correlate either. A lot of the in-content, on-page, and that kind of stuff did not correlate at all even though the commonly held belief was that they would correlate. Now, in saying that correlations aren't perfect, and I think you have a question coming up about how correlations work and all that kind of stuff, and whether it is a worthwhile analysis, but certainly, from the correlation data, those things do not matter.

Michael: Yeah. Okay. So let me understand how you put all these things together to come up with the correlations like 0.38 for the SEOmoz rank authority, which models Google Page Rank. So, if I back up and say, let's start with the data. What kind of data and where do you get it from?

Mark: Okay. So, I'm just analyzing Google. So, in future studies, I might analyze Bing, and Ask, or Blekko, and that kind of stuff, but I am just analyzing Google. So, what I did was a fairly basic thing. I went to Google Adwords Keyword Tool. They have - whatever it is - like twenty-seven categories of keywords. Went there. Got the top eight hundred keywords for each category. Maybe sixteen categories or something like that. Anyway. And worked out of about twelve thousand unique keywords.

Michael: Are those individual keywords, Mark, or are they keyword phrases?

Mark: Some of them were individual keywords and some of them were keyword phrases.

Michael: Okay. So these are the most frequently searched phrases using the Google search engine.

Mark: No. No. It was actually designed, essentially, to get a fairly even spread between frequently searched stuff and not frequently searched stuff. So, I'm not sure if it is on the About Page or somewhere on my blog; I have a little graphic that breaks it down into like four categories, like above two hundred thousand and below a thousand searches per month. So I tried to get sort of an even distribution between frequently searched stuff and not frequently searched stuff because, essentially, what could be one of those data points that could throw off the analysis.

Michael: Right.

Mark: For example, above two hundred thousand searches. The chances are that a lot of them are going to be brands that are global brands, and that is going to be impacted. That is probably going to impact the exact match stuff because Disney are going to show up for Disney; but if you search, let's go back to, how to make chocolate cake, that is probably going to be relatively lower than the number of searches for Disney, but it is not going to be a brand name whereas you still could have the exact match domain.

Michael: So you wanted to make sure that your data set was representative of the entire universe of people searching on Google, so you wanted the stuff that was highly searched, and you wanted the stuff that was long-tail, five searches per month. You wanted a representative data sample from the entire spectrum.

Mark: Exactly.

Michael: Okay.

Mark: I mean, in future studies, where I will probably do is I will break them down. So I'll test the correlation of exact matches for the different brands and all sorts of stuff like that. But for this correlation study, that is what I did. I tried to get an even distribution and that kind of stuff. And then, I built a Python script that used US-based web proxy, so that just basically means that it hid my Irish IP address so that Google wouldn't serve me Irish localized web results. And I went and got the top one hundred results for each of those - whatever it was - eleven or twelve thousand keywords and then sorted that locally. So, whatever that is - ten thousand by ten -, that's a hundred thousand. So, a million URLs.

Michael: Yeah. And so, what did you actually extract from the Google result pages?

Mark: I literally just extracted the URL.

Michael: So, the URL from the top ten results for each of the keywords?

Mark: The top two hundred.

Michael: Top two hundred.

Mark: The top two hundred for each keyword. And I did that and actually, one of those points that could change, I'm not actually sure; I haven't tested it. But let's say if you went down through the different ten pages in Google as apposed to requesting a hundred at a time; that could potentially slightly change like a personalized thing in Google. I couldn't tell you if that is a factor or not. If it is, it's minor. So, I just tested. I took a hundred at a time because it saves computer resources because, obviously, using a web proxy that can fool Google costs a little bit of money, so I made sure I only used it eleven thousand times or whatever it is as apposed to using it ten times that number, so a hundred and ten thousand times.

Michael: Definitely.

Mark: So, potentially, a very, very minor flaw, but probably not. So, I had my hundred URLs. I had my ten thousand keywords. I had how many times they were searched a month. And then I went through basically a massive list of factors that I wanted to test. And I test as many as I could using the resources I had, so literally I was just using my laptop, and a data drive, and just got all the data I could. So, for example, I wanted to test things like thing in the URLs using the keyword in the page name matter, so that is really easy to test. You have the URL. You have the keyword. You just test them. And then, other things like: does the number of links pointing to it matter? That is a lot harder to test because, obviously, I could not create a database of all the links in the web and analyze them all. So, there is just stuff like that, that costs a couple million to build that kind of an index. So I went to SEOmoz and they gave me access to their API for free, which was great. And I got the data from them and then, I already mentioned, I went to (Unclear 28:40.1) and got some data from them. I went to Link Research Tools and got some data from them and then, obviously, I gathered some of the stuff myself. So, for example, I got the HTML of all the - whatever it was - million URLs and downloaded it to my data drive and then I analyzed different pieces out of that.

Michael: Wow. Very nice of SEOmoz to give you API access to their database so that you could continue to do your analysis. I love Seattle-based companies that are helping out entrepreneurs. So let me ask you this. What is the statistical analysis tool that you used to analyze whether something was significant or not?

Mark: Okay. Yeah. Perfect. So, essentially, I was comparing two data points - the ranking and the actual data point for that ranking. So if you take an example; let's say I want to test page rank. There are clear numbers there. The ranking is number one, two, three all the way down to a hundred and I have the page rank number for each of those results. So I have, let's say, the first result was page rank of seven, the second result was page rank of seven as well, and the fourth result was page rank of three, and everything like that. And then, it's basically a really, really simple statistical tool. It is just a little algorithm - a little formula - called Spearman's Rank Correlation Coefficient. You can Google it and you will get the Wikipedia page. It was created by Charles Spearman, and is essentially just a response to Pearson's Correlation Coefficient, which is slightly more popular; but it essentially assumes that there is a linear correlation, but the truth is, with Google, sometimes it doesn't work that way, so I used the Rank Correlation Coefficient and that allowed for non-linear correlation. And then you just feed the two data points into the algorithm for each keyword and you get the correlation for each keyword and then you just get the average over the eleven thousand keywords.

Michael: Okay. So you got a coefficient for every single keyword and then you looked at that over all eleven thousand data points that you had, and then you just took an average.

Mark: Yeah. Just took an average. So, essentially, you compute the correlation coefficient for each of the ten thousand based on a hundred connected data points, so essentially two hundred data points. And then, the two hundred data point by ten thousand, and got the average.

Michael: Okay. So, easy enough, and that is something you can just calculate in Excel. You do not even need a statistical analysis software, do you?

Mark: No. You do not need statistical analysis software. I actually did it in Python because I was programming a Python, so I actually downloaded the (Unclear 31:11.1) module and it already has the function built in to do the correlation coefficient. But I could have got the algorithm from Wikipedia and just programmed in the algorithm, but they had a function that worked so might as well use it.

Michael: Sure. Makes sense. So the output variable you are measuring is actually the ranking in results and then, you are gathering other pieces of information to go along with that such as the page rank value or the number of letters in the domain name, or the type of TLD it is, and you feed that as a factor to the Spearman's Rank Correlation Coefficient and you get the coefficient as the output.

Mark: Exactly.

Michael: And we want to know how close to one that coefficient is because, if it is closer to one, then there is a larger correlation.

Mark: Yeah. Exactly. So, maybe some of your users are also wondering about stuff like how would you get the correlation for whether the keyword is in the URL or in the title, or whatever like that. So, you just convert it into an (Unclear 32:16.6) value of 1 or 0.

Michael: 1 is in; 0 it is not in.

Mark: Yeah. Exactly. So that is just a sort of common question people ask; but exactly. You are getting a result out then, in the mean, and of all the keywords that you test the correlation for a between -1 and 1. And, sort of 0 is like the breaker in between that number line. And if it goes between 0 and -1, it is a negative correlation. The two things do not move together. And if it is between 0 and 1, then the two things move together. And then the strength of the relationship is determined by how close it is to either of those ones. So, if it was -0.9, that is a really strong negative relationship; and if it is 0.9, then it is a really strong positive relationship. And people get into whether correlation is causation and stuff like that, and there are sort of flaws in the model. It doesn't prove causation. So, for example, I would always put a disclaimer in my data to say correlation doesn't prove causation. It is just a really good marker of whether it is a causal relationship. So, for example, they use it a lot in Science, especially like Bio-Molecular and all that kind of stuff because they can't necessarily get out causation, or often, it's just a lot easier and cheaper to test for correlation. So, I'll just put that disclaimer in there for anybody who sort of jumps into the (Unclear 33:39.2) and says, 'oh, this correlation is that, so I'll just throw it whenever (Unclear 33:42.1)'.

Michael: No. Exactly. There is a clear difference in causation. A correlation does not indicate causation. I've studied a little bit of statistics in the back. If I smoke a pack a day for the rest of my life and I look at a pool of people that are just like me, you can say that smoking causing lung cancer, but you can only say that there is a correlation between smoking and alcoholism because it turns out that people who smoke tend to drink, or people who drink tend to smoke. There is a correlation, but you can't say that one necessarily causes the other. It hasn't been proved in a scientific way.

Mark: Exactly. So that is a perfect case where correlation would actually be used quite successfully; and there are also cases where it can be misused. So, for example, in my data set, like I said, all the link-related factors I tested turned out to correlate really highly, including the number of no-follow links pointing to a URL. Now, that is a classic case where a bit of common sense has to kick in. And Google have said that they don't treat no-follow links as passing juice and don't pass any search engine ranking benefit.

Michael: Right.

Mark: So what you have to say there is you have to think, is there a correlation between no-follow links and followed links? Are websites that have a lot of no-followed links also getting a lot of followed links? And the truth is there is a really good correlation between that. Websites that get shared a lot on Twitter, which use no-follow links or link to on Wikipedia, or Facebook, or anything like that are likely to get more traffic and more people share the URL on their blog with a followed link. So, that is a classic case of, if there is incorrect analysis of the correlation, you get another myth coming out of what should be data providing actually good advice. So, it is partially a subjective thing. It is partially not just looking at the correlation. So that was a classic case where I just said, "Oh, clearly, there is a correlation between no-follow and followed links, and it is not indicative of the fact that there is a causal relationship between and no-follow and ranking. It's actually really indicate of the fact that there is a relationship between followed links and rankings because, if no-follow links are showing such a high correlation that must prove if there is a significant correlation between followed and no-followed links that followed links has an even higher correlation". So, yeah, it does take a little bit of analysis.

Michael: Yeah. Definitely. So, again, I am just amazed at the comprehension and your ability to explain it to others of these topics. Statistical analysis. When I was seventeen, I didn't know anything about statistical analysis. I did not learn that in high school. I don't think I even took any classes in college, if I can remember. And it wasn't until I began work at GE - General Electric - that I started to get involved in statistical analysis for appropriate business decisions being made. Yet, at seventeen, you are speaking with such authority on the topic and I understand that you actually know the topic. I'm just amazed. How did you come up with your background in statistics?

Mark: I just sort of took an interest in it. It is just something that you can learn if you go and read about it or whatever.

Michael: You just read about it on the web and you are able to internalize it and then apply it?

Mark: Yeah. Exactly. Get a few books. Look it up online. Read some articles or whatever.

Michael: Do you get your parents to help you out? Brothers or sisters? Anybody else or are you just analyzing it yourself and figuring it out?

Mark: Well, I like to talk it out with other people, but the majority of it is myself. And then you can look up forums online if you have a question. I had a guy who helped me out with my programming early on. Steven helped me out with some programming early on and some questions that I had. My dad actually did agricultural economics when he was in college, so he helped out a little bit with the correlation stuff early on. But once you sort of get a grasp on it, you can sort of infer things for yourself and learn things for yourself, and then you can get onto a bit more higher level stuff like learning a bit more about statistical modeling and statistics in general, and all that kind of stuff. So, mostly, myself running it, but I do seek out help whenever I have a problem.

Michael: Yeah. Well, it is very impressive. So you are evaluating two factors at a time. You are going through and you are saying, "Here is my output, here is the factor that I have to evaluate, and I'm getting a coefficient". And you are going one, by one, by one, by one, by one. Now, you know just as well as I do that there is another statistical evaluation that would allow you to then analyze multiple factors at one time. The difficult part to doing the analysis is not necessarily figuring out how to use the formula and figuring out how to program it; it's getting the data. You have accomplished a monumental task by gathering the data, connecting the SEOmoz and asking them for access to their API to analyze their in-bound links across thousands of tens of thousands and hundreds of thousands that they have analyzed. So, why would you not start doing multi-variate testing to look at how the correlation shakes out for multiple factors at a time?

Mark: Yeah. So, basically, what you're talking about is trying to get to more causal relationship and proving that. And there are a bunch of things, like I think you suggested Regression analysis, and there are a whole lot of other things. The main reason that I went for correlation is because somebody had already done a correlation study, so I had a benchmark for a methodology. I had a benchmark for potential results that I was going to get out, so that was my first significant study.

Michael: So you could do your analysis and say how it connected to theirs and then you could make some direct comparisons.

Mark: Exactly; for similar factors. Then I basically had a benchmark to say that I was credible. Like a lot of people aren't going to believe some seventeen year old that rocks up and does a regression analysis first off. So, that was the main reason I did.

Michael: And I've got to say, even though I've been talking to you for like thirty minutes or so, I'm still hesitant to believe that you are seventeen years old. So I completely agree.

Mark: Exactly. So some people are a bit skeptical, I kind of think. The other thing is people haven't heard of me before, so maybe they are not going to believe it. But there are reasons that I did a correlation. The first one is, is that it is a lot easier to do. Like I said, you just have to collect the two data points, compare them with a really simple function, and it's really easy to get whereas, regression analysis or other more advanced stuff are harder to do. And I don't know if you've ever heard of Freakonomics or Super Freakonomics. They are sort of interesting books. I took an interested in them. And there was an interesting quote in the Super Freakonomics one from the Economist, Steven Levitt. He does all the data stuff behind the Freakonomics books. I think I'm paraphrasing a bit. It said something like regression analysis is more art than science; and this is a guy who is totally data driven and he is an Economist. He is a lecturer at a Chicago college, or University of Chicago, Illinois, or something like that. He is a completely data driven guy and he said it was more art than science. So there are always going to flaws with models you pick out; and I mean the beauty of it is that, if you get a big database and you hold it in time, you can always go back and do a second one. You can do another correlation study. You can do a regression analysis later. So, essentially, I did the correlation study because it is the easiest to do, because there was a benchmark there, people already somewhat believed in correlation studies, and because it was easiest to conduct and easiest for the reputation. But absolutely. I agree that it has to go a lot further than a correlation studies. It has to go into causal studies. And I mean, even beyond regression analysis, things like actually physically making changes to a website over a large enough scale to be able to measure whether those changes make any difference in search engine ranking is the ideal situation, but that is really hard to get down to. Can you imagine trying to run a test on a million websites and a million webpages, making a change to each of those million webpages and having control groups and all sorts of stuff like that? That is a lot more difficult just from a financial point of view, from a computer resources point of view, from a software point of view, and from pretty much every point of view. So it's all about trying to find models that suit the goal you are trying to achieve and getting it there using the least amount of resources as possible. So, absolutely regression analysis. Absolutely causal results. Absolutely other tests. So, for example, I am thinking about testing a more subjective thing. So, for example, a lot of people will say content - the quality of the content - is really important to your search engine ranking. It is good common sense. It is one of those things that, when you're making that statement what you are saying is, either that good quality content reads other factors that correlate well or have a causal relationship. For example, links as in quality content means more links. Or, what you're saying is, Google have so advanced algorithms that they can figure out what is quality content or not. Now, testing that is quite difficult. So, for example, a test I want to run is the same thing as if you see in the social network. When Mark Zuckerberg - I don't know whether he actually did it or not - built Facemash.com and she scraped all those pictures of all the girls and put them side by side, and made users vote on which one was better. I would do the exact same thing. Essentially, in my test, I would get all those hundred articles, put them side by side because you can't ask a user to go on and say rank those hundred articles one to a hundred. You would have a brain freeze after you read ten articles. How could you say, 'Oh that's ninety-three; oh that's ninety-four. That's one. That's two'? Like making those minute differences. Whereas, if you compare one and ninety-four, it could be really obvious what ones should be better quality or not. So, essentially, you do something like the (Unclear 43:53.4) rating system to compare to and you would actually test based on the subjective judgment of so many users whether quality content matters. So it's all about the stuff about testing all sorts of different things. Testing in different ways and, hopefully, getting the same result across all sorts of different platforms and stuff like that. But it remains to be seen whether I will get the same result or similar result across all sorts of platforms, but absolutely testing for more causal relationships is definitely where I want to go.

Michael: What factors have you evaluated to date and what is remaining to evaluate in this evaluation portion of your project?

Mark: So I think I've tested a hundred and sixty or a hundred and seventy factors. So, that is pretty much all the main factors that anybody really talks about. Like what I did was, when I came up with that list, I literally went through forums and blogs, and Google searched as much as I could and tried to find any single muted factor or partial factor that people think, theorize, or talk about and see if I can test it. Now, some of them I didn't test. So, for example, I didn't really have the computer resources, which was just my laptop to go and fetch all the number of Twitter shares for a URL. I mean the Twitter API was just a bit of a disaster in terms of getting things out of it, and limits, and stuff like that. I could've went for the Facebook one, but seeing as I wasn't doing Twitter, I decided to group them together and not do either of them. So, for example, those two would probably correlate quite highly based on what people say and based on Google+, I think. So, for example, I have got server resources off a company in the UK that also some more results. They gave me access to their servers and Amazon's elastic compute, so I'll be able to run my next correlation study from up in the cloud, which will obviously give me access to much more powerful computing resources and will allow me to test some things that I did not test. But essentially, I tested pretty much every single, or most of the, hard line factors, like what you would say is a factor. Some of the more touchy-feely stuff, like I was saying about quality content; that probably doesn't get tested the way I got it, which was, basically, get a data point and compare it with another data point. You sort of have to come up with some more interesting tests and experiments for them. But essentially I tested pretty much most of the muted factors. So, you were saying about Danny Sullivan's thirty main factors. I mean I probably would've tested ninety percent of them. So, a hundred and seventy is about the extent of any list of factors you will find. There were some others ones I wanted to test as well, like whether having HTML is correlated well. But, again, when I was using the w3 validation API, they have some API limits and also, it was quite slow API. So a lot of it is about trying to get the data and some of the stuff was just trash because I knew I was going to do another correlation study and so, I could always just leave it until I developed a bit more as a programmer, developed more in terms of getting further resources, and all that kind of stuff.

Michael: Definitely. All right. So when I look at the factors and you say that you've evaluated a hundred or sixty or so; that hundred and sixty actually includes the magnitudes higher that fall underneath those. So, one of the factors is link building, for example. When I look through your posts on your website and I look at the correlation of in-bound links to a website and how that affects the page rank, I can see that I think you tapped into SEOmoz's API and you can look at the number of in-bound links with no-follows, the number of in-bound links with follows, and there are probably like twenty different factors or ten different factors of those in-bound links that you measured and determined the correlation for. So, when you say a hundred and sixty factors, you actually evaluated tens of thousands of factors across the entire study so far.

Mark: Yeah. Some of them probably would boil down to like further sub factors. So, for example, I don't think I tested page rank; but if I had tested page rank, that would have actually boiled down to a lot of sub factors.

Michael: Right. Sub factors or elements. However the search engines are referring to them.

Mark: Exactly. They sort of use different language. But I would have tested some sub factors and then I would have tested some higher-level factors that would probably come down to more sub factors. But I think there will always be sub factors of any sort of factor you test, and sub factors, and sub factors, and sub factors, and all sorts of stuff like that. So I tested sort a wide array of stuff. How many factors are higher-level sub factors and stuff like that? It doesn't really matter that much because if the sub factor is really important to the higher-level factor, then it's worth testing that. And as in SEO, what you really want to do is boil it down to really right down to the very last sub factor because that is the stuff that you can actually take action on. Increasing your page rank. Okay. How do you do that? Oh, well, you got to look at the sub factors of what makes up page rank. And then you got to look at in-bound links versus outbound links, versus no-follow versus follow. So it is all about boiling it down to the sub factors and I would have tested sort of a mix of both. So, some of them probably would boil down to more, but I mean my goal is really to test as many of them as anybody can ever think of to test, whether that is sub factors or high level factors.

Michael: Okay. So, at the high level, what are the five biggest factors that influence a higher correlation?

Mark: The common ones you'll hear in the industry is links, which is correct; and then you'll hear things like title, description, on-page stuff, and the correlation that I did, I did not see that at all. So what I'll probably say is, I'll probably say links, links, links, and then maybe social and then probably links again probably. I'm sort of joking a bit, but it was literally in the correlation results that I saw. It was all links. I test something like thirty factors of links and I think pretty much all of thirty would've been the top thirty. And then I tested stuff like in-URL, on-page, in-page, and all sorts of factors, and hardly any of them correlated as highly. There were a few more. So, for example, exact match .COM would have come out at a fairly significant level that would've probably been just below links. So, exact match; links; social. You'll hear a lot about them in the industry, but it's just sort of quantifying them and ranking them, but definitely links are the big ones.

Michael: So get your in-bound links, the exact match domain, and then some social aspect correlates as well. So, just looking at the domains portion, Mark, if I am buying a domain name like UsedCars, it's got an enormous search volume I'm sure and I own UsedCars.com. That is going to have a high correlation to ranking well as long as there is good content on there just like Danny Sullivan was saying, you've proved with statistical analysis. So we're saying that exact match domains are good. They will drive a higher result. What if it's UsedCarsDenver or UsedCarsForLess? What if the search phrase is part of the exact match domain, but the domain is actually larger? The domain extends beyond the search phrase. Have you done a correlation study on that to see how well those rank?

Mark: Yeah. So I tested partial match and I tested other things like hyphenated match and all sorts of stuff. And they're posted in my blog and all that kind of stuff. But yeah, I tested partial match and I forget exactly what the number was. You probably researched it before you interviewed me, but I think it was something like 0.03 or something like that. So I remember the conclusion was, essentially, that partial match didn't correlate well. It had no impact. Now, you're obviously not just reaching out to search engine optimizers and webmasters; you're reaching out to domain investors. Now, analyzing it from an investment point of view versus a search engine point of view, some of it is connected, but some of it is not connected. So, a lot of the stuff you'll see in the domain name market will represent what happens in search engines. So, for example, everybody knows that .COMs are better in the domain name market than .ORGs. There is a hierarchy - a clear hierarchy - there. And that pretty much is representative. Another thing that is representative is I tested whether hyphenated exact match correlated well. It turns out they correlated much worse than non-hyphenated stuff. And that's represented in domain name market. But some of the stuff then that isn't represented well in the domain name market is the relative values of partial matches versus exactly exact matches might not be represented in domain name market as it is in the search engine benefit. And that's because domain names aren't only bought for their search engine value. They're also bought for their brandability. They are also bought for their length. They are also bought for all sorts of other factors. People believe they will increase in value essentially. They are a different security than search engine ranking, which isn't a security at all. So, they are an asset and they have to be treated an asset, but from a search engine point of view, you are right. That is what the correlation came out as.

Michael: All right. What can people do to help your project, Mark? If people want to help out, is there anything they can do?

Mark: Not really. If anybody has data, that is basically the oxygen of the project. If anybody has data and I mean data on anything. Anything. Because I'm interested in testing anything to do with online marketing, whether it is social data, domain name data. Like I have nearly procured a database of historical domain name sales going back to 2008 from NameBio.com. So, they were really good. I e-mailed them and I wanted to do a domain name market analysis. What affects domain name prices and all that kind of stuff? And part of that was needing historical domain name data, so I just e-mailed them and they go back to me. So, anybody who has data that they can help out; but if you don't have data, all you can do really is just follow along if you are interested in following along. You don't have to help out. If you are interested in following along, just follow along in the blog, or on Facebook, or on Twitter. I'm not really great at the social media stuff, but any time there is a new blog post or something important, I'll share it up there. But just follow along.

Michael: All right. Well, thanks to NameBio for getting you that data. I know that a lot of people would be interested to see your correlation results after the analysis is complete, and I'd love to have you back on the show to talk specifically about those results when you have completed that, Mark.

Mark: Cool.

Michael: So the final question is this, Mark. For over a decade, people have focused on keyword density within the articles they publish. You know, I've been on multiple article outsourcing systems. They specifically have a section on the specification sheet for saying how many times would you like the author to specify the keyword phrase in the article. Is this worthwhile from an SEO perspective based on your correlation data?

Mark: No. That will probably be the biggest myth that ever came out of SEO professionals' mouths. I think I read an article a while back that said that even when Google started, back in 1998, they never used keyword density as a ranking factor. It was a ranking factor with like AltaVista; but if you look at all the results from AltaVista, they were an absolute disaster. So, definitely the biggest myth that ever came out of search engine optimization.

Michael: All right. So it does not correlate. You have looked at the number of times the keyword phrase is mentioned on a page, you've take the rank within the results, and you have calculated the coefficient to be below zero or not statistically significant.

Mark: What I test was: as the keyword density increases, does that correlate? Now, a better test would actually be to break it down in the bands. Does the keyword density between zero and one percent correlate? Does a keyword density between one or two percent correlate? But you could take it as, I think I did a post on this, I used a bit of anecdotal evidence, I used hard evidence from Matt (Unclear 56:23.9) and stuff like that, and then I used some statistical evidence that was slightly dubious in that I only tested if it correlated as it increased, so essentially you could have a cancelling out effect. But yeah, you can take it that keyword density doesn't matter.

Michael: Yeah. All right. Well, it's amazing how many analyses can be done on data. You can slice and dice it a number of different ways and you won't know until you strike a gold nugget of data until you have actually done the analysis. And so, I encourage you to do more. If other people want to do analysis also, I'm sure you'd love to have their analysis on your blog as well, Mark?

Mark: Absolutely. Yes. I'm going to publish my data and my code probably in the next month. And that will be open to everybody. Well, my code will be in the next month and then I'll publish the data after I've done my own analysis of the data, so that will be in a couple months time. And then, anybody who wants to can just have access to it. So if you are interested in getting access to my code or my data, then literally just subscribe by e-mail to the blog and then whenever I'm releasing it, I'll let you know about it.

Michael: Excellent. If you have a follow-up question, please post it in the comments below this video and we'll ask Mark to come back and answer as many as he can. If people want to follow you, Mark, they can do it at @OpenAlgorithms or on Facebook. There is a link. We'll put it below this video as well. And then, of course, go to this website TheOpenAlgorithm.com and sign up for his newsletter so you can get notified of his updated posts.

Mark Collier, Founder of TheOpenAlgorithm.com. Thank you for coming on the show, sharing your knowledge of the search engine ranking factors and your statistical analysis, and thank you for being a Domain Sherpa.

Mark: Thanks, Mike.

Michael: Thank you all for watching. We'll see you next time.

Watch the full video at:
http://www.domainsherpa.com/mark-collier-theopenalgorithm-interview/