Big Data Transcript

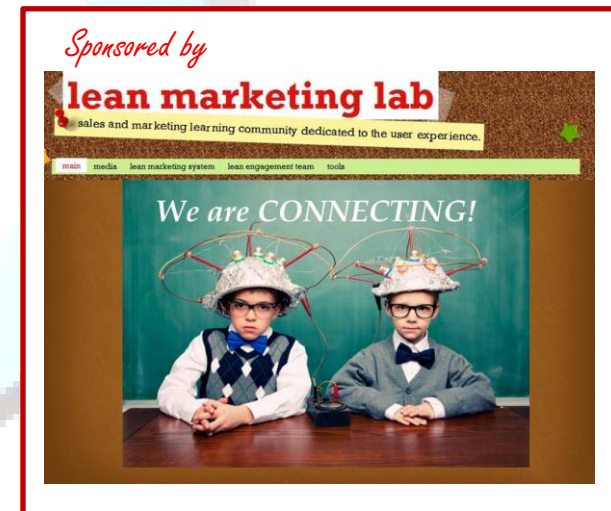## Guest was Kaiser Fung

Related Podcast:

[Framing Big Data, part 1 of 2](#)

[Framing Big Data, part 2 of 2](#)

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

Transcription of Podcast

Joe Dager: Welcome everyone; this is Joe Dager, the host of the Business901 podcast. With me, today is Kaiser Fung. Kaiser is a professional statistician with over a decade of experience applying statistical methods to marketing and advertising businesses. He holds an MBA from Harvard Business School, his acclaimed blog "Junk Charts", pioneered the critical examination of data and graphics in the mass media. I would like to welcome you and congratulate you, Kaiser, on your previous book, "Numbers Rule the World" and your newest book, "NumberSense".

Kaiser Fung: Thanks a lot Joe and I am very happy to be back.

Joe: I enjoyed our last podcast immensely. Is "NumberSense" a continuation of "Numbers Rule the World" or is it an entirely different book?

Kaiser: It is both a continuation and it is a little bit different. The first book, "Numbers Rule the World", is organized around the sort of five important statistical concepts which I tried to explain using plain English, focusing on people who are non-specialists. Then I look at different stories that illustrate how these concepts work, and that framework has sort of its own limitations. This new book is in part a reaction to the whole Big Data movement. It allows more freedom because it is more topical. I take topics from things like Fantasy

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Football, things like Economic Indicators, College Rankings, and Marketing Data. It is large topics and we walk through how statistical thinking can help us understand the types of information that is put out there right now in the media.

There is a lot more freedom for me to explore different aspects of it. Not constrained by specific concepts that I wanted to put out. If you have read my first book, you will find that there are a lot of parallels, a lot of links back to "Okay, here I am really using that concept from Chapter 2 of the book." You will find a little bit different, but you know, generally the same line of thinking.

Joe: Before the podcast, we talked about John Paulos. He of course wrote the book, "A Mathematician Reads the Newspaper." I think his book framed how we should view numbers in everyday life. I thought your book did a great job of reframing how we should be looking at Big Data without saying who is right or wrong, but putting it in our court saying, "we need to frame it."

Kaiser: I think it is important people need to realize before they dig into the book is, I focus on a specific aspect of Big Data, which is not, receiving enough attention right now. We keep hearing about Big Data and it is like volumes of data, there are all kinds of new types of data. We are tracking everybody, every movement and that is true and an important aspect of it. Much of that aspect is much what I would call supply folks – it is all about the people who are doing the data work, and the people that are tracking us. My book is focused

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

on the consumption side, so one consequence of having so much data is that it is going to be that there are tons and tons of people who are going to come to us with all kinds of arguments, and they are going to tell us that their argument is supported by such and such data.

For most of us who do data analysis, it is probably even for people who read data analysis, you will realize that you can pretty much find data to support anything you want to say. So what is going to happen; we are going to have a lot of contradiction and confusion. There will be so much data analysis out there; we do not know what to think. In Number Sense, what I try to do is to give people, as you say, a framework to start thinking about how you would interpret all these things out there.

If you have two sets of researchers who are telling you contradictory things, and they have their own data sets to support it; how do you tell which one is believable and which one is junk. Like you alluded to this is not an exercise in figuring who is a hundred percent correct and who is not. Our problems are so complex, and the data sets, even though they are so luminous will never be complete. We will never be able to know for sure that, you are right, and he is wrong.

I encourage people to take a skeptical attitude, and develop your own framework for interpreting the data analysis out there. I would have expected some people would probably not hundred percent agree with everything that I

have to say in the book either. That is totally acceptable, that is sort of part of the mentality of how you approach the interpretation of data analysis.

Joe: Well that is the whole point to it. Who do we believe, and how do we analyze? Should we embrace Big Data, or should we be somewhat skeptical and scared of it?

Kaiser: It is interesting because I just put up a blog post this morning – It is about Snapchat. I wrote that blog post, and at the very end, I cited Eric Schmidt, the past CEO of Google. He had a quotation for that I used to think "Oh my god, this is so creepy." He said something like "There is no privacy anymore if you do not want anyone to know that you have done something, then you should not have done it in the first place." I used to think that it is like a common trend how creepy this technology companies are, but I think all these revelations are essentially making me rethink what he actually said. No matter, whether we like it or not, the data is out there and somebody will collect it. It is extremely easy for some people to collect it, we just cannot avoid it. I think he is just basically saying "If you take that as your starting point, then you should think about whether you should be doing things that you do not want other people to know." It is a different thing from saying it is creepy.

Joe: I notice after visiting certain sites, the different ads that are coming up and I am seeing on the web how they reference where I have been. Sometimes

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

when I go surfing the web, I am looking for new information. Is my search being more constructed or funneled into my past, more so than I think?

Kaiser: Yes and this actually leads me down this other path. I actually do think that for me, I have been in the marketing business for a long time working with marketing people and helping them use data to help improve their performance. We should have essentially a Code of Conduct to respect people's privacy, more than we do. A lot of that is based on the fact that I believe that a lot of the data that is being collected and being used is being used in a way that actually does not improve the performance. There is a lot of data that is being collected that actually is not necessary. You actually raised this example which is, one of a clear example of how things are not as they seem.

There is a whole sector of online marketing, which is often times called remarketing, or retargeting. You are right that the reason why all these ads chase you around is that, what happens is we have the ability to know where you have been. You went to Amazon or some place and you browse around, and you were looking at something you did not buy. It is not too surprising that if I keep putting it in front of you for the rest of the days or the weeks. Some of these people will actually buy eventually. If they eventually buy because we have had ads chasing them around, I think most likely is not true. People just have a consideration cycle that is longer and they could not make

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

up their mind in the first place but eventually they will. However, the way that the web works, often times the last person that can be tied to your purchase is going to get the credit and would get paid for having supposedly generated that sale.

Now, that goes into the whole conversation, correlation issue that we have one of the topics that I have kind of talked about on and off in the book. The way that the industry is currently working is that the last person who shows you the ad before you make the purchase gets full credit oftentimes, for having generated a sale. So, that is exactly why the ads are chasing you around. In my opinion, those types of things are quite ineffective because oftentimes "You are preaching to the choir" – I mean these people oftentimes had already purchased, and that goes into the less than a hundred percent precision of the sub- tracking. Even if they have not bought yet, most of the time, without the retargeted ads that kind of chases you around, those people will eventually buy, are going to buy anyway. Remember that in one part of the book, I talked about the concept style thinking to illicit, cause and effect, and this is not an application of it. Those things are only effective if you are able to motivate people who would otherwise, not came back to buy. If all you are doing is to capture the people who will all come back to buy no matter what, then you are just spending money for something that you do not need to spend money on.

Joe: Are reward cards and discount cards and all that data becoming intermingled

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

with Big Data?

Kaiser: Yes. You would think those are essentially the same tools. Big Data is just that, we now collect a lot more types of data in a lot more places. Even way back when, when we start having those supermarket loyalty cards or any kind of loyalty cards, just about has always been that if you collect data about your customers, you had better be able to tailor the experience to them. In theory, that actually works. The whole point and even like credit cards, where you have Amazon credit cards, you have Macy credit cards, and all these retailers have joint programs with credit card companies. The whole point is the reason why they give you the points and benefits is that they want us to consolidate our purchases in one place, so that they have a more comprehensive database to understand our purchasing habits. One thing about data analysis is that you will get to the wrong conclusions if you start buying alphabet buyers' data sets. If you spread your purchases over a lot of different not-linked together retailers, then each will only have like a small window on your purchasing experience so it is difficult to understand who you are. Even back in the days of the loyalty cards, they started building profiles of users. Now, it has just got into a much bigger scale because pretty much every company is in the game and there are a lot of like data exchanges and places where you can actually fill in the holes and gaps in your data set.

Joe: I think "NumberSense" starts addressing this with me and raises a few questions. It is not the data, it is the analysis and how you use is that is

important.

Kaiser: Absolutely. It is ultimately, the data can be as big and as massive as it can be – and we might need to build giant data centers to hold them. But ultimately, nobody can use all the data points. The analysis part is how do you extract out the information from the amounts of data, how do you summarize it where in statistics, we like to talk about the summarization of the data – We talk about reduction of the data to its essence. This is the same thing that Nat Silver was talking about in the "*Signal and the Noise,*" how do you pull the signal out from all that noise that it is in. What how is going to impact us consumers is that people are going to look at the data in some way. They may look at it in a good way or bad way leading to either good or bad or somewhere-in-the-middle types of data analysis. It is the analysis that draws out what we call the insights that actually leads to policies and actions that affect us.

Joe: How do I get better at analysis or do I need to have a statistician on call?

Kaiser: That is one thing that I play it out – is this notion of "NumberSense". I found over the years since I manage teams of data analysts and think about who turns out to be good, and who turns out to be not as good. It is ultimately not about your technical ability, it is not about how good a mathematician you are, it is not about how good a coder you are, although those things are some sort of minimum requirements for people on my team. The one thing that distinguishes the good ones from the not-so-good ones is what I call

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

"NumberSense" – it is like an intangible thing. It is this idea of if you are given a data analysis, do you know how to take it apart? Do you know what questions to ask? It turns out that this is not something that you can teach in a classroom, it is almost impossible. It is almost like trying to teach common sense and in a classroom. I find that the best way to learn is by what I call apprenticeship. So it is looking at what other people who have this "NumberSense" have done, with data analysis and by sort of reading broadly and by just reading a lot of discussions about data, that you kind of pick up these things. That is sort of the whole premises of the book, it is like you said kind of like Paulos. It is like me, the statistician – data scientist, reading the newspaper. Reading all these things that people have said in the recent past and I give you a window into how I work. If I need to, want to have an opinion on these particular topics that is sort of how I dissect the analysis and what I do with it. Sometimes, I have to go to other sources. Oftentimes I have to dig to some details as to how the methodology works. What I hope to illustrate to people is that you do not need to learn math or calculus; or, you do not need to take stats classes that are being taught at math classes. You do not need to know computer programming; you actually do not have to pull down terabytes and gigabytes of data. You can get quite far without doing any of that.

Joe: Two questions come to me. One is that you seem to describe someone that is able to step back and look at the big picture but is willing to take the dive in, and dig deep.

Kaiser: Yes, I mean, you kind of describe what you need to become a statistician. You will find what happens is you pretty much have to be in the weeds and in the details. Then there is always the summarization. It is always sort of how do you, after you have understood everything how the data is generated; how to analyze, and all the little minute details of it. Ultimately, we have to remember that our jobs is to take us out one of those details, and build a bridge to – for a way of describing what is going on. Anyone who is not comfortable in being able to come in and out of the details and sort of the big picture is not going to make a good statistician.

Joe: One of things you alluded to earlier, when we were talking about outcomes. How do we stop from having this outcome sitting out there that we want, and just making the numbers work for us?

Kaiser: I think one of the things that I have mentioned; even in the prologue of the book is that you really can't. The reason is that the human brain is actually extremely prolific in creating theories of our world. You sit here and you just generate theories. There is both the opportunity and the risk of this age of Big Data that there is so much data out there. On one hand, it would mean that there are more ingredients that would allow us to prove certain theories, sensible theories. It creates a risk because now people can easily, just by looking at certain parts of the data, support their versions of the story.

Now, not all of this is malicious. Some of it may be that, people just want data

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

to support their own points. If you talk to people who do data analysis, you will find that oftentimes, we fool ourselves. A good example of that, is that you thought that you are looking at a certain type of data, and then you might subsequently found that that data has a lot of missing components that you were not aware of that were not collected. Or that, certain parts of it were erroneously collected. It is labeled as one thing but in fact it is actually if you go down to the computer circle itself, it turns out that maybe somebody made a mistake, and it is actually mentioning something else.

It reminds me of, I do not know if you have read that article or I wrote an article in New York Times, I think last year, just saying that there were a lot of researchers. I think these were cancer researchers and had been working with certain cell lines. They have all these very lax Aussies that kind of, exchanging their research ingredients with each other. There are lots of papers that were published where people thought that they were studying a certain type of cancer, in fact, they were studying other types of cancer. You know that thing happens to data a lot too, and I think the whole point is that you can never avoid making mistakes, but if you have the right framework and if you have this intangible thing that I call "NumberSense". When you see the bad data, when you see the bad analysis, you feel uncomfortable; there is a noise in your head. That is when you start digging around, and you will then be able to find things that will help you either, interpret better or discover that it actually does not make any sense. I think the key is not going to be that we can avoid anyone from doing either maliciously or inadvertently a bad analysis. My

attitude is that, we just have to make ourselves smarter consumers of the data analysis out there.

Joe: All these different services that are selling data, should a smaller company use them and how selective should they be?

Kaiser: For that type of question, I would always start with the objective. I think it is never a good idea, and certainly not for a company of a small size where you have limited budget to go buy a whole bunch of data and then, figure out what you want to do with it. I think you need to start with "Okay, what is the business problem that I am trying to solve?" Then you need to develop some type of theory as to "What is it that will lead to success? You might be solving a problem, "I need customers to walk into my store." Then, you need to solve some theories around like "What are the things that are motivating people to come into my store?", and how the data could help you there, may be that you want to as of, attain which of a bunch of different theories may, in fact, be correct. In figuring out what are either the motivating factors of people coming to your store, or the factors why people are not coming to your store. If you are driven by an over-arching objective, you will find the right pieces of data to address your own questions. It is a lot more productive than just having a tremendous amount of data that you really do not know what you are doing with it.

Joe: I develop this buyer persona and I say "We are going to be entering this

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

market; we are looking to see if this product works." I assign a particular demographic or psychographic to try it out and get some groups around to be able to do that. When I go out there, am I really just getting sold a list, or am I able to fine-tune it that much with these services?

Kaiser: I think one of the unfortunate things about the current hype with Big Data is that there is not enough attention to reporting the accuracy of these procedures. A couple of my chapters in the book are dedicated to this. Actually, both this book and the previous book are dedicated to talking about how we measure the performance of either, predictive modeling or in a Big Data mining. How you would do that, is you would enter a test agreement. Most data places will allow you to do that. You essentially, instead of going for a long term contract, you would sign a short term contract and you sometimes, you can only just rent the information just for the one thing that you are doing. The important thing is to know how you are going to use that data and for your own problem and then measure the value of it. Typically the set-up to sort of branches where you would have one situation in which you actually leverage the data to your work, and then you have a separate group of say, customers or people, for whom you do not use the data. So this goes back to the whole counterfactual thing. What you compare is the difference between optimizing my strategies assuming I have such data, versus, the other situation where you assume that you do not have the data. Even though you have it, you choose not to use it. If you choose not to use it, you would undergo a different strategy and you compare it to the strategies and then,

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

the difference will tell you the value of the data. The nice thing about the data space is that, you can always and most of the time, I would not say always. Most of the times, you can find ways to put a quantifiable value on what having that data gives you. You can actually calculate an ROI of the data that you are buying, and that is nice.

One thing that is missed oftentimes out there when you read the newspaper about all this Prescriptive Modeling and Big Data mining is that they are not telling you enough about the accuracy of these methods. That is another long conversation, and I get into it in multiple parts of this book as well as the last book. It is obviously important for us as data analysts to be honest with ourselves, to look at what analysis is actually buying us. A great example is the NetFlix's price which got a lot of attention. Some of these researchers won a million dollars for helping NetFlix improve their rating algorithms by ten percent. They developed this extremely complex, so called ensemble of over hundred different component models. Later on, I think a year later, the engineers at NetFlix put up a blog post that talks about the aftermath of the Netflix price. In it, they say "Well, we actually did not implement the winner. We took two of the methods out of hundred submitted. Then for the ones submitted, the explanation was that the gains and performance was actually not large enough to merit that kind of complexity, like this gigantically complex network of models. That is the type of stuff that we need to be starting to talk about and not just hyping everything up as saying "Oh, well you would definitely find that one terrorist" or in another case that I talked

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.

Copyright Business901

about in the book, there is this whole conversation around how Target, the retailer, finds pregnant women and send them special brochures. What sort of missed in that conversation is the type of errors that can be made when you deploy such a system, finding women who are not pregnant, or missing the women who are pregnant. We kind of get ourselves into that conversation and are very honest about what these things can do, and what these things cannot do. We really are not consuming this information in a smart way.

Joe: I think you are so right there. Because for years, I got this horse magazine and I am so far removed from needing a horse magazine, that you know, by ZIP code, everything is wrong for it. I always laugh and thought "What data set are these people using?"

Kaiser: Yes, I work with these models every day. I have been for many years. I mean I can tell you most of time they are wrong. We are playing a game of odds, and going to back what I have said before – from the perspective of the person doing the modeling, all we are looking for is, if we did not have this prediction, we would have gotten a certain level of accuracy just based on intuition and guessing or whatever other methods that we have. What we are shooting for is just an incremental benefit by using data to try to be better at targeting. With our accuracy initially started at ten percent, and we improved it and doubled it, it would be twenty percent. But that would still mean that eighty percent of the time we are wrong. From our perspective, it would have been a massive win because doubling something is not easy. It is actually a

very hard thing to do. But from the perspective of the consumer, like you receiving that horse catalogue, it is still a failure.

Joe: Did you write the book for a statistician, or for a non-statistician?

Kaiser: Well, it is primarily written for non-statisticians. My belief is that even non-statisticians can actually do a certain part of the job that statisticians do. A lot of that interpretation of the data analysis is not rocket science. And for most non-statisticians, you are not going to be dealt spread sheets of numbers or databases – you are going to be given somebody's policy advice or somebody's business tip based on some distilled version of the data. And that is sort of where I start with and I try to help people develop the framework to analyze those things – So, analyze the analysis, so to speak. On the other hand, I suspect that the analytics community and statisticians would also – should be interested in reading the book as well. Because one of the things that I am a big believer in is that "NumberSense" is not something that you can learn from the classroom. You learn by, as I said, reading what other people are doing, understanding how the other people reason about data – So, I am always reading what other statisticians are doing. That is really how you grow your own ability to analyze data analysis. So mostly for non-statisticians, a lot of people who are doing this work would also be interested in it.

Joe: So statisticians cannot be replaced by computers, right?

Kaiser: Well, personally I do not really think so. You kind of read through my whole body of work, you will find that it is really all about how we can marry a thing or two – it is, they like to say statistics is both an art and a science. There is a science part – the science part makes our work more efficient, makes us more certain and productive, but there is still a lot of creativity, a lot of detective work that it is really hard for me to imagine how a computer as it is designed today – the types of computers that we have today – can take over a large part of that type of work. In fact, one of those underlying things is that we are really running the risk of producing a lot of bad analysis when we just pump things into a machine, and hope that the machine will come up with the right answers.

Joe: Should numbers really rule the world? Can I go with my gut instinct, or should we really leave the numbers persuade me?

Kaiser: Well, I think the numbers actually serve a few purposes. I think it is always great even if you sort of intuition, to have numbers to back you up. Like sort of, if you are in marketing, marketer, and you think that by spending money on marketing or an advertising campaign, you are going to generate a lot more sales. Even in that case, it would be great if you are able to find some data to support your theory. So if you spend a lot of money on it, you kind of can see your fruit of your action – So even for measurement, I think it is a great thing. And then you have the other way which is for discovery. Let us

say you have Google or even Facebook – you have a billion users. I would actually – Data analysis – it is almost impossible for Google and Facebook to understand who their customers are. Intuition or even just like talking to some customers is going to give you an extremely biased view. You are going to basically base your entire description of one billion people based on interactions with what – a hundred people? Having the data and the numbers are really important. But like we said previously, completely trusting the numbers and the analysis is also very foolish. You need human intelligence to interpret these numbers to really – it is really an interplay of the numbers and your interpretation because ultimately, even though the numbers will never give us cause of information – they can never really tell you with certainty that A causes B, it would tell you that A is related to B. It is really that, human interaction that is needed that kind of tie these things altogether into a credible story and like I have said before, forget the notion that you will find the one story that is correct, and everything else is wrong. All we are trying to look for is a story that is our best story, given our constraints of what we can and what we cannot.

Joe: I would recommend your book to anyone that is trying to learn numbers, trying to learn how to use Big Data. I think "NumberSense" is a great book. It is a book that once you start reading, there are certain pages that I push myself through and then all of a sudden, it was like "Oh, wow, I get it!" I think there were a couple of "Ah-hah!" moments in there like that. I highly recommend it because I think we have to understand Big Data because it is

part of our lives now. I appreciate the book very much and being able to read it and I want to thank you for it. Where is the book available? I am sure it is on Amazon – is there a separate book site for it?

Kaiser: Yes, I am going to be rebranding my blog. I have two blogs. I have the "Junk Charts" that you have mentioned that is about data-graphics and how to interpret those. I have a blog that is associated to my first book, "Numbers Rule Your World" blog. That, I am going to rebrand very soon, because now I have two books and put on that blog, I actually will be continuing these type of exercise where I will look at the big topics of the day. The topics that which people are publishing some numbers to support their theories and now taking those apart and telling people how I would look at such things.

Joe: Well, I would like to thank you very much. What is the best way for someone to contact you? Through the blog?

Kaiser: Yes, I also am on Twitter, @JunkCharts". I have – and through the blog you can find my email address, I am on Gmail. So people can also directly write me, but these days Twitter is probably as good as anything. Guess I am actually on it quite a bit.

Joe: Well, I would like to thank you very much for your time. Again, the new book is "NumberSense". Podcast will be available on Business901 iTunes store and Business901 blog site. So, thanks again, Kaiser.
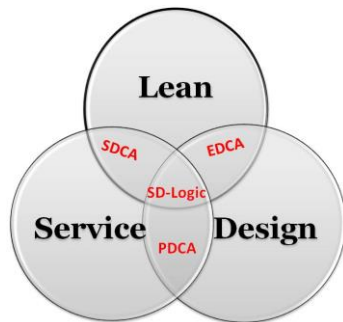
Business901   *Podcast Transcription*

Implementing Lean Marketing Systems

Joseph T. Dager

Business901

Phone: 260-918-0438

Skype: Biz901

Fax: 260-818-2022

Email: jtdager@business901.com

Website: http://www.business901.com

Twitter: @business901

Joe Dager is president of Business901, a firm specializing in bringing the continuous improvement process to the sales and marketing arena. He takes his process thinking of over thirty years in marketing within a wide variety of industries and applies it through Lean Marketing and Lean Service Design.

Visit the Lean Marketing Lab: Being part of this community will allow you to interact with like-minded individuals and organizations, purchase related tools, use some free ones and receive feedback from your peers.

Framing Big Data, part 2 of 2   *Framing Big Data, part 1 of 2*.