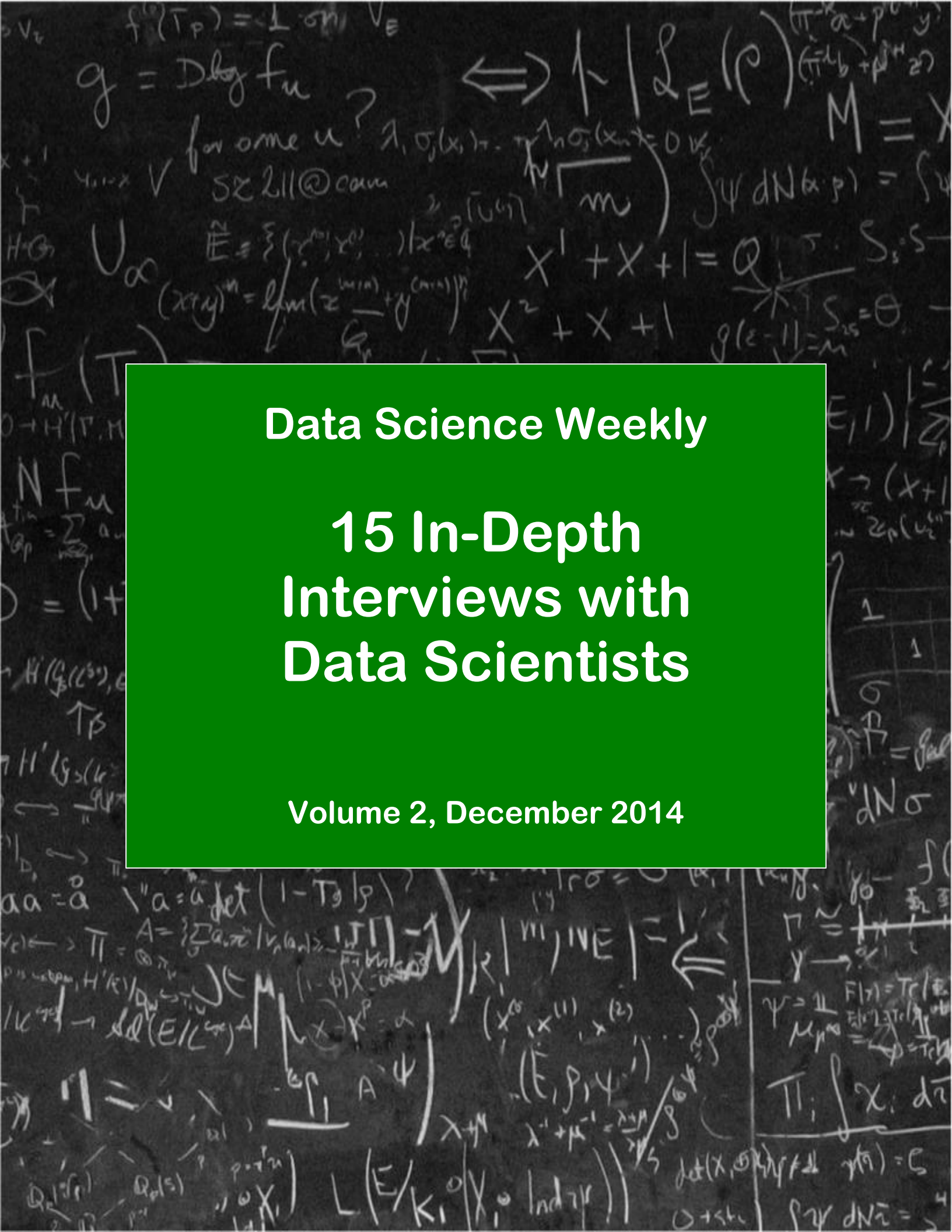


Data Science Weekly

15 In-Depth Interviews with Data Scientists

Volume 2, December 2014

FOREWORD

Given the success of our first Interview Series, we kept going!

Over the past few months we have been lucky enough to conduct in-depth interviews with another 15 different Data Scientists.

The 15 interviewees have varied roles and focus areas: from start-up founders to academics to those working at more established companies; working across content discovery, gaming, genomics, health and fitness, entertainment, mobile/communications and more...

We wanted to make the interviews more easily accessible to the community, so have put them together in this pdf.

We hope you enjoy the read and thanks again to all the interviewees!

Hannah Brooks, Co-Editor
DataScienceWeekly.org

P.S. – If you enjoy these interviews, and want to learn more about

- what it takes to become a data scientist
- what skills do I need
- what type of work is currently being done in the field

Check out ***"Data Scientists at Work"*** – a collection of 16 interviews with some the world's most influential and innovative data scientists, who each address all the above and more! :)

[Full disclosure: author is my co-editor at DSW, Sebastian Gutierrez!]

CONTENTS

Jarno Koponen: "Random" Predictive Content Discovery.....Page 5
Co-Founder of Random

Emmett Shear: Data Science at Twitch: CEO Perspective.....Page 16
CEO of Twitch

Dag Lohmann: Machine Learning in Catastrophe Modeling ..... Page 23
Co-Founder of KatRisk

Chul Lee: Data Science to improve Health & Fitness ..... Page 32
Director of Data Engineering & Science at MyFitnessPal

Michael Watson: Applying & Teaching Data Science ..... Page 42
Partner at Opex Analytics; Adjunct Professor at Northwestern University

Chris Checco: Cloud-Based Analytics for Communications .....Page 50
President and Chief Analytics Officer of Razorsight

Reid Robison: Creating the "Dropbox of your Genome" .....Page 64
MD, MBA and CEO at Tute Genomics

Kirk Borne: Data Mining at NASA to Teaching at GMU ..... Page 75
Data Scientist and Professor of Astrophysics & CS at GMU

Dmytro Karamshuk: Optimal Retail Store Location..... Page 88
Researcher in Computer Science & Engineering at King's College London

Nathan Kallus: Big Public Data to Predict Crowd Behavior.....Page 98
PhD Candidate at the Operations Research Center at MIT

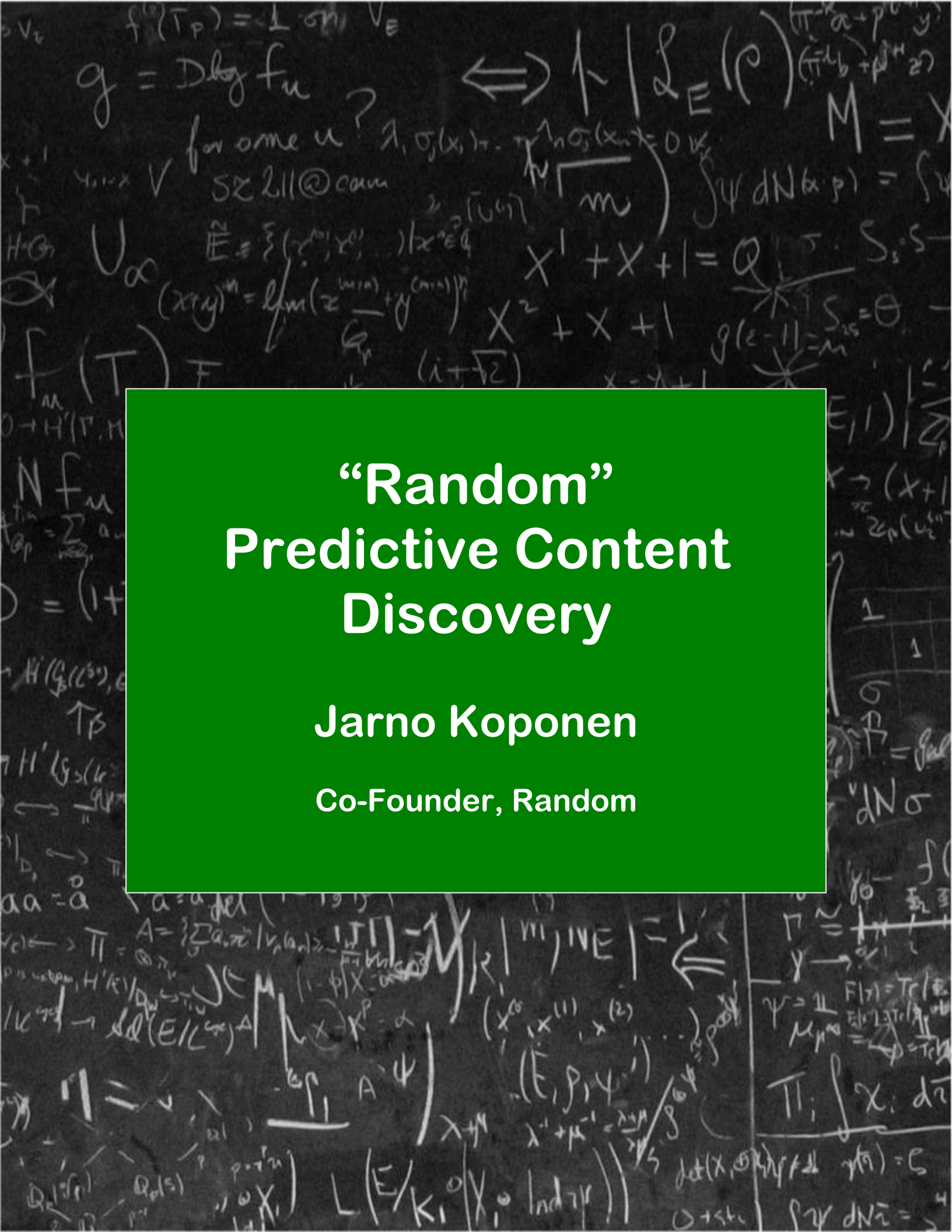
Joseph Misiti: Reducing Friendly Fire to Analyzing Social Data....Page 108
Co-founder of Math & Pencil, SocialQ and more

Jeroen Janssens: Data Science at the Command Line..... Page 117
Author of Data Science at the Command Line

Sudeep Das: Transforming OpenTable -> Local Dining Expert.... Page 128
Astrophysicist and Data Scientist at OpenTable

Nick Elprin: Building a Data Science "Experiment Platform"..... Page 137
Founder of Domino Data Lab

Szilard Pafka: Building the LA Data Science Community.....Page 145
Chief Scientist at Epoch and Founder of Data Science LA



“Random” Predictive Content Discovery

Jarno Koponen

Co-Founder, Random

“Random” Predictive Content Discovery

We recently caught up with Jarno Koponen, co-founder of Random. We were keen to learn more about his background, his perspective on predictive content discovery and what he is working on now at Random...

Hi Jarno, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - I am the co-founder of Random, a designer and humanist passionate about augmenting our personal and collective future thinking. First step (with Random): a system that enables you to explore the unexpected and discover new interesting things...

Q - How did you get interested in working with data?
A - I believe we make sense of ourselves, other people and the world around us through stories that consist of bits and pieces of information...

that is, data. Data is everywhere. Yet data in itself is nothing. It needs to be processed and refined to become meaningful and valuable to anyone. Data needs a use case, it comes alive through a story or a product.

I've always been passionate about personal stories and our digital existence. At some point I started thinking, how could personal data be used to open up an individual's future horizon. What kind of an application would help him / her to see the existence of different alternatives around themselves and to discover new interesting things. And that led me to start working around the topic of personal future simulations and predictive discovery.

Q - What was the first data set you remember working with? What did you do with it?
A - This is a bit more of an old-school thing, not a standard answer you'd get from a hard-core computer scientist... In a sense, we as humans are made of stories. I'm a deep-rooted humanist and I started working with history books. Going to the very source by looking at old texts and then starting to construct their meaning through related sources and relevant research literature. That to me was a unique data-set.

Q - What makes you believe in the power of data?
A - It's said that very little can be learned from history. This thought reflects a more fundamental belief: humanity itself doesn't change - basic things, like motivations and intentions affecting our behavior remain constant.

However, I believe change can and does happen through individuals. If

learning and change can happen on a micro-scale, it could happen on a macro-scale too over time. If our personal data can be used to open up our personal future horizon, it might mean that such an application of data could potentially have more far-reaching consequences... even for humanity itself.

In the end, everything depends on finding the use case and building products that matter. In a digital product both qualitative and quantitative data come together in a unique way. Both macro and micro events do count. Humanists, designers and data scientists are all needed when building adaptive and predictive systems - and together can unleash the power of data.

Interesting to get such a different, humanist perspective - thanks for sharing! Let's talk in more detail about predictive content discovery...

Q - What changes are we seeing in the world around us that are enabling Predictive Content Discovery?
A - The amount of personal data has increased dramatically. Widely adopted use cases - such as Facebook or Netflix - make it possible to utilize data to understand an individual in a specific context. Every use case serves a certain purpose. Every user interface directly affects what kind of data can be meaningful and how the system can learn and adapt. Simultaneously, any specific use case does not provide a holistic understanding of an individual. New methods are needed to make it possible for an individual to benefit from his / her own data as much as

possible.

The amount of information around us is increasing. Our current tools are inadequate for making sense of our needs and intention in relation to the information around us. New methods and tools are needed to access information that matters to us as unique individuals. We'll be moving from the age of search and social feed to a new territory, to a new paradigm in which the right information comes to us. We will go beyond linear to truly non-linear experiences, from feed scrolling or input-driven to something proactive, adaptive and personalized. Search and social will remain for the time being, yet a new user experience paradigm will be born around predictive content discovery.

Q - What are the main types of problems now being addressed in the Predictive Content Discovery space?
A - I don't think it's about problems, rather it's the new opportunities that are born with the evolving digital ecosystem. I think the opportunities can be put into two buckets:
i) How can we get more of what we know we want (without asking for it)
ii) How can we get (more) of what we don't know we want (without asking for it)
For example, how to get better search results when you know what you want. Or how to get new movie recommendations when you don't quite know what you'd like to watch next. Both i) and ii) have content specific and more universal applications.

It's said that there's more and more information and new content out there. However, we don't have proper tools to access them. To generalize,

with our current tools and services our everyday Internet has grown smaller instead of bigger. We end up in the same places far too often. The great opportunity lies in making tools and applications that widen our chances to find the things that matter to us personally.

Q - Who are the big thought leaders?
A - I think there are many interesting people exploring this field both in practice and theory. In this space, an interdisciplinary approach is vital. I would concentrate more on use cases and products rather than individuals. For example, Netflix is a great example of bringing together algorithms and human curation to power content discovery. Also, I've always been fascinated by Wolfram Alpha and the ideas and inspiration behind their work. I'm also following the development in the field of digital humanism with great interest. Pierre Lévy's (@plevy) work is full of inspiring insights. Alex Soojungkim Pang (@askpang) has explored how digital rhythms affect human behavior. We need a more human-centered approach when building digital products and especially systems that will augment integral parts of our existence.

Q - What excites you most about bringing Machine Learning and Content Discovery together?
A - To understand an individual we need to create a concrete connection between physical and digital worlds. We need to be able to bring the human mind closer to the digital realm - at some point even inside it - in a human-centered way. That's why the interaction between Machine Learning and human behavior is so exciting to me.

To be more specific, to build a truly adaptive predictive system, both the

human and the system need to learn from each other continuously. To build such a system, both the user experience and system mechanics are equally important. Searching and discovering are different modes of thinking - we think and act differently when we explore to discover new things. And thus the system should act and learn differently too. In curiosity-driven exploration the unconscious and irrational parts of our thought-processes surface more explicitly. Using Machine Learning to capture our irrationality and associative thinking enables a loop to be created such that humans and machines are learning from each other continuously in a unique way.

Q - What are the biggest areas of opportunity / questions you would like to tackle?
A - It goes back to the previous themes we've discussed. How can we help an individual explore the unknown and unexpected for positive surprises? How can we better understand the needs and intentions of an individual? How can we augment our personal future thinking by creating a truly adaptive and predictive interface to the information around us?

Really interesting questions - and fascinating to think about how the world of content discovery could evolve going forward. On that note, let's talk more about what you are building at Random...

Q - Firstly, how did you come to found Futureful/Random?
A - In a nutshell: a great bunch of talented people with an aligned vision came together and Futureful/Random was born. We used to be Futureful and now we're Random. The name change is related to the evolution of

our system and application as well as the world around us - the view of the future contained in Futureful was actualizing in the present so we updated our name.

Q - What specific problem does Random solve? How would you describe it to someone not familiar with it?
A - Random helps you to explore the unexpected and discover new interesting things. The app lets you go beyond what you know and find positive surprises in the Internet. Random is built to feed your curiosity: you can start with the topic "design" and end up learning new things about "algae". If it matches your personal interests that is. All this happens in a seamless personalized flow. Random never gives you the same content twice.

Q - Why is it interesting to you? What is the most surprising insight you have found?
A - Making an adaptive and predictive system that finds the right balance of serendipity and relevance is a great puzzle. And to do that one needs to understand both rational and irrational sides of a human being.

Q - How do users find relevant content? How is this different from / better than traditional web-browsing?
A - As I just mentioned, Random balances relevance and serendipity. It starts by suggesting topics that might be of interest to you. You then choose and get new content in front of you - be it a a blog post, video, photo or a news article. And then you choose again to find new things, at your own pace. You do not need to type in anything. You do not need to follow anyone. You do not need to sign in. The system learns from you

and also brings up stuff from the periphery of your interests. Everyone has a unique journey when exploring new things.

Q - So, you don't have a sign-in process or link to social feeds to suggest content - could you tell us a more about the technology behind it, and how Machine Learning is helping...?
A - From the very start, Random learns from you and lets you define what might be interesting for you. It learns from every interaction or non-action and starts to map the way you see the world.

Everyone's reality consists of unique connections between things. To someone "art" is more strongly connected to "abstract art" than "photorealistic art". For someone else "art" refers most strongly to "oil paintings". Our system tries to capture and understand what kind of unique connections an individual has. Random tries to understand what "art" means to you personally and what kind of related things might be interesting to you.

The app also allows you to connect things freely thus letting you express both your rational and irrational self. There're no universal categories or connections between different things - rather it's about an individual's own "ontology" that's created through usage. The "associative ontology" evolves continuously both through the actions of the individual and other people using the system.

Random is out now in the App Store. Feel free to give it a spin and let me know what you think!

Q - That's great - look forward to trying it out :) ... Last thing on Random ... You mention on your website that you see Random becoming much bigger than an app - could you talk more about your future vision?
A - Recommendation systems will play a key role in shaping our thinking / processes going forward. The technologies powering Random could be used to build an adaptive and predictive interface to all sorts of different kinds of information - hence driving much greater impact than just personal content discovery. For example, the next generation of operating systems will have predictive recommender systems built in their very core. A capability to learn, adapt and predict will not be a feature, but the core of the operating system itself.

Jarno, thanks so much for all the insights and details behind Random - we wish you the best with the new product! Finally, it is time to look to the future and share some advice...

Q - What does the future of Predictive Content Discovery look like?
A - It will be more human(-centered). More mobile and ubiquitous. A new language - consisting of gestures, natural language and metadata - will be used to power the new interface to the information around us.

Q - Any words of wisdom for Data Science / Machine Learning students or practitioners starting out?
A - Follow the stories that matter to you personally. That's how it started with Random.

Jarno - Thank you so much for your time! Really enjoyed learning more

about your background, your perspective on how predictive content discovery is evolving and what you are working on now at Random. Random can be found online at http://www.random.co/ and Jarno is on twitter @ilparone.

Data Science at Twitch: CEO Perspective

Emmett Shear

CEO of Twitch

Data Science at Twitch: CEO Perspective

We recently caught up with Emmett Shear, CEO of Twitch. We were keen to learn more about his background, how data and Data Science have influenced Twitch's growth to this point, and what role they have to play going forward...

Hi Emmett, firstly thank you for the interview. Let's start with your background and how you became interested in data and the early days of Twitch/Justin.tv...

Q - What is your 30 second bio?
A - I've been doing startups my whole life. I started a now-forgotten calendar startup called Kiko Calendar right out of school with my friend Justin Kan, which we sold on eBay a year later. Then we started Justin.tv as a 24-7 experiment as a new kind of reality TV that we called "lifecasting", along with our two new co-founders Michael Seibel and Kyle Vogt. Justin.tv grew to be pretty big, but eventually our growth plateaued and we had to figure out what to do next.

I had the idea to pivot the company entirely and go after live video game streaming and became the CEO of Twitch, which we built out of Justin.tv gaming. I still run Twitch today, and also spend some time as a Y-Combinator part-time partner mentoring startups.

Q - How did you get interested in data?
A - I've always been interested in data. Even before starting companies, I loved "personal analytics" where I'd graph things happening in my own life. The hard part isn't having fun looking through data for patterns – it's figuring out how to find data that can actually drive decisions.

Q - Can you talk about how data was used in the early days at Twitch/Justin.tv?
A - When we worked on Kiko and Justin.tv, we mostly just shot from the hip. While we occasionally referenced data in our decisions they were driven primarily by a product vision.

With Twitch, I did things very differently. From day 1 we've focused on data to guide our decision-making process. Some of the time, this means classic data science and crunching the numbers on user behavior in aggregate across millions of actions. Sometimes it means interviewing 10 key example broadcasters to understand their views of the world. Sometimes it means market research on how competitor's features are working.

This was a magical, transformative experience for me. Instead of guessing what users wanted, I actually knew the answer. We tended to jump to

actual solutions for their problems instead of adding features that no one wanted or used.

Q - Was there a specific "aha" moment when you realized the power of data within Twitch/Justin.tv?
A - When we were first getting started working on gaming, I wanted to figure out what would be responsible for driving growth. I had the idea to go and see which streamers on the site were pulling in the most new users -- getting the word out about us. When we focused down on delivering value for those streamers, it caused a huge bump in our growth. And it was thanks to having real data and understanding that we knew who to talk to.

Q - What role did data have in deciding to create Twitch from Justin.tv?
A - Honestly, less than you'd think. I started Twitch because I was passionate about video game content and I thought there would be a big market for it, based on how many gamers were out looking for interesting things about video games on IGN and YouTube / Machinima. It wasn't a particularly data driven decision.

Very interesting background and context for where both you and Twitch are at today - thanks for sharing! Let's talk in more detail about the role Data Science can play at Twitch going forward...

Q - Which companies do you think Twitch should try to emulate with respect to their use of Data Science? Why?
A - I think mobile gaming companies are some of the sharpest in the

world in their Data Science usage. I always learn something when I talk to a product manager or Data Scientist from one of those companies.

Q - Where can Data Science create most value in Online Gaming?
A - I'm not really an expert on gaming per-se. I know a lot about video and a little about gaming!

Q - Fair enough! Let's focus on Twitch ... How is Data Science used at Twitch now? Is this different from Justin.tv?
A - It's very different from Justin.tv. We've invested a huge amount more into Data Science now than we ever had before. The team at Twitch is 5 full-time members now and growing. We had zero full-time on Justin.tv.

It's also much more useful because we have dashboards internally which allow everyone in the company to get insights without having to do all the scripting and analysis themselves. That's been a huge shift.

Q - What are the biggest areas of opportunity/questions you want Data Science to tackle at Twitch?
A - It's easy to calculate things like Lifetime Value (LTV) in a retrospective way. Figuring out your predictive LTV for new users as they come in and understanding how changes impact that LTV is the holy grail.

Q - What projects has the Data Science team been working on this year, and why/how are they interesting?
A - We've started digging in on much more sophisticated questions recently. The newest one is working on clustering viewers to understand

the different types of viewing habits in aggregate. Hopefully we'll have fun results to share on that front eventually!

Q - What has been the most surprising insight/development they have found so far?
A - I was surprised to learn how common it was for people to subscribe to a single channel one month, and then jump to another channel the next month. That accounts for a huge number of our subscriptions!

Q - Interesting! So, how does the Data Science team work with the rest of the organization? How would you describe their role?
A - The Data Science team owns and builds the backends and dashboards and notifications that allow the entire organization to understand the impact of their actions and make good predictions.

Q - And where/how are they integrated into decision making?
A - They're brought in on almost every product decision that gets made, because you can't make good decisions without understanding the data behind it. Usually very early in the process, as we're figuring out if we need more instrumentation and analysis, and how we're going to test for success.

Q - Which other groups/teams do they work with most closely? Why?
A - Product management for sure. That's where the data gets integrated with the rest of the business realities.

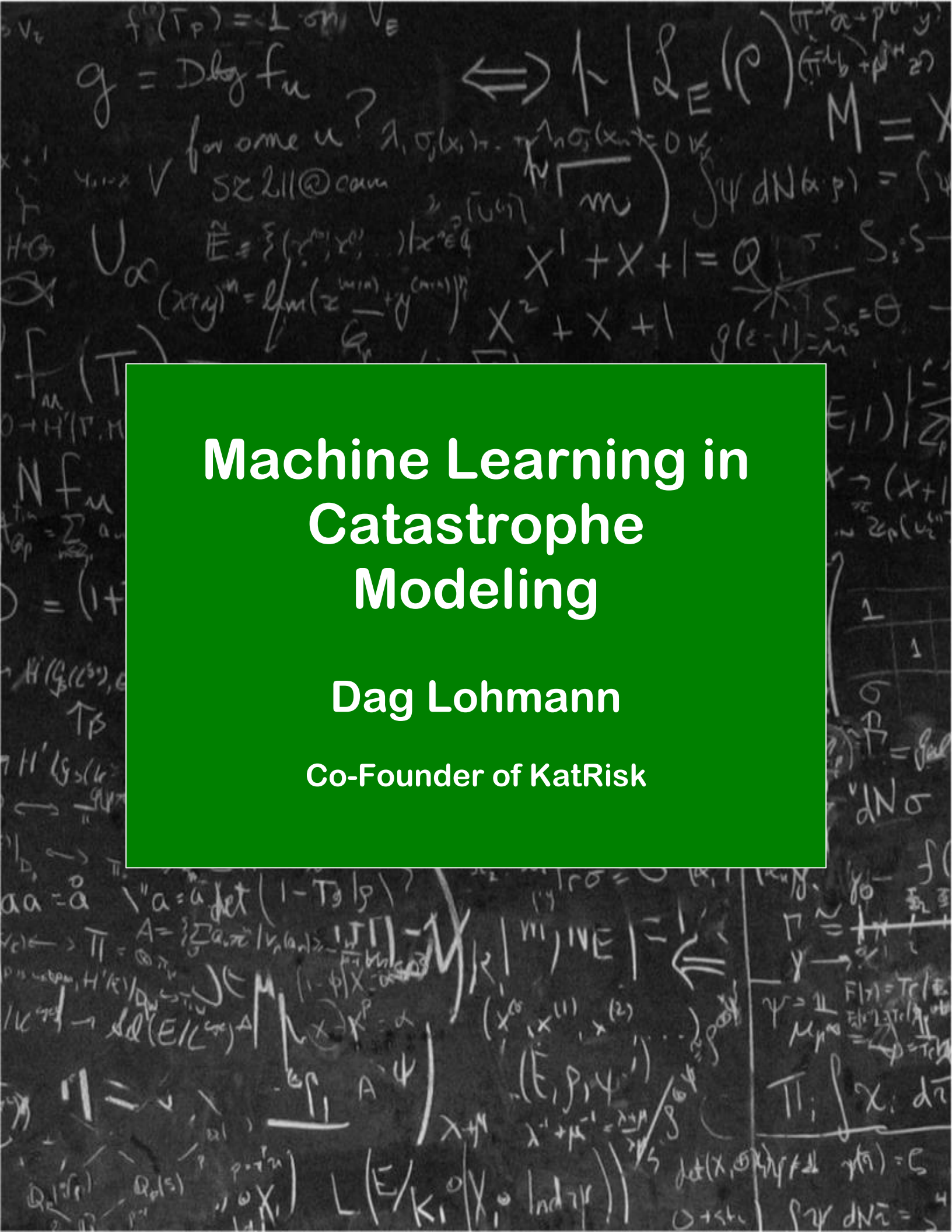
Q - Makes sense. You are currently hiring/building out the Data Science team at Twitch... What are you most looking for in candidates?
A - We need great engineers to build analytics systems, and we need great data scientists to make use of them. Both pieces are crucial.

Q - What is the most compelling reason why they should join Twitch?
A - It's incredibly impactful and challenging work. We deal with tens of billions of events per month, and the output of that analysis directly drives all our most important decisions.

Emmett, thanks so much for all the insights and details behind Twitch - sounds like a great time to be part of the Data Science team! Finally, one quick question on your role at YC...

Q - You've advised several batches of YC startup companies - how has their use of data/Data Science been evolving? What excites you most about these developments? What advice do you give in general and related to data?
A - More and more startups are actually actively tackling these problems as a service. I've seen several companies starting to produce automated tools to attack problems we've had to solve one-off at Twitch, which is really exciting for new companies that don't have the advantage of their own in-house Data Science team.

Emmett - Thank you so much for your time! Really enjoyed learning more about your background, how data and Data Science have influenced Twitch's growth to this point, and what role they have to play going forward. Twitch can be found online at http://www.twitch.tv and on twitter at @twitch.



Machine Learning in Catastrophe Modeling

Dag Lohmann

Co-Founder of KatRisk

Big Data & Machine Learning in Catastrophe Modeling

We recently caught up with Dag Lohmann, Co-Founder of KatRisk. We were keen to learn more about his background, how the world of Catastrophe Modeling is evolving (and the influence of Big Data and Machine Learning) and what he is working on at KatRisk...

Hi Dag, firstly thank you for the interview. Let's start with your background and how you became interested in data and Catastrophe Modeling..

Q - What is your 30 second bio?
A - Before co-founding the risk modeling company KatRisk LLC, I was Vice President of Model Development at Risk Management Solutions in Newark, CA leading a team of modelers and software engineers building and implementing catastrophe models. I worked for 7.5 years on the development of continental scale flood risk models in RMS and RMS' Next Generation risk modeling methodology. Before that, from 1999 to

2004, I was with the National Weather Service, NOAA/NCEP/EMC in Camp Springs, MD, where my main interest was data assimilation with real-time data, forecasting and hydrological modeling.

In terms of education, I received a Physics Diploma (Masters) from the Georg-August University in Goettingen (Germany) and a Ph.D. from Hamburg University (Germany) before working for 2 years as a postdoc at Princeton University. I received the 1999 Tison Award of the IAHS and have published numerous papers on risk modeling, hydrological modeling, model uncertainty, forecasting, data assimilation, and climate change.

Q - How did you get interested in Catastrophe Modeling?
A - I remember having a conversation with my brother (Gerrit Lohmann, Professor at AWI Bremerhaven in Germany) in 1998 about climate and extreme events. We thought about looking into this as a commercial enterprise. I then looked up what the marketplace was and found out pretty quickly (not surprisingly) that are companies working on this. I then applied to one of these companies, but it wasn't until 2004 that I started working for RMS, the current market leader in catastrophe modeling.

Q - What was the first data set you remember working with? What did you do with it?
A - The first real data set I worked with was in 1992 when I was still a "real physicist". I did my Masters degree in high energy physics then and worked on Coherent Bremsstrahlung. It was a very interesting problem and I had a very good supervisor (Prof. Schumacher, Goettingen). I had

written some code (in C) that would do quantum electrodynamic computations of high energy photons that were created by fast electrons hitting a diamond. After the results from the experiment at MAMI B in Mainz (Germany) came back and I saw that the computations matched almost exactly the predictions I was excited. We had run the experiment all night and in the morning I drove 250 miles back home (on an old Suzuki 400 motorcycle) to show these results to my professor. I still remember riding that bike for 4 hours -- tired, but happy!

Q - Was there a specific "aha" moment when you realized the power of data?
A - That came much later, and in a way that "aha moment" is still happening to me quite a lot today. Data was never too special for me, I always liked simple concepts and models that are able to reflect reality. I am now quite amazed by what is available out there and I always want to do more... It is quite an exciting time for people that like data and models. I feel we are only scratching the surface right now.

Very interesting background - thanks for sharing the personal story - sounds like a memorable bike ride :) Let's talk in more detail about Catastrophe Modeling and how Big Data and Machine Learning are influencing the field...

Q - How is the rise of Big Data and Machine Learning changing the world of Catastrophe Modeling? What is possible now that wasn't 5-10 years ago?
A - I always find it interesting what "Big Data" really means in Catastrophe Modeling. I think many people largely think about Big Data

as unstructured data that tracks behavior on the internet. Meteorologists and climate scientists have dealt with very similar problems and tools (EOF/PCA analysis is a good example) for a long time. I think BD and ML will change how we structure code and data, but the math behind will stay the same (or evolve slowly). There is a lot of well organized meteorological data and many smart people are using them for many different purposes.

Big Data for me also means large computations. We are in a world where I can now build a computer for $50K that has the same compute power as the world's fastest computer in 2002. Quite a change - and cloud computing hasn't even really started yet. Catastrophe models will use more and more information in the future. I can easily imagine how all the different and divergent data sources might be used (in aggregate) for decision making in the future.

Q - What are the biggest areas of opportunity / questions you want to tackle?
A - Scientifically I would like to do risk forecasting. But we first have to be better in modeling the current weather and climate risk before we do future risk. There is still so much work to be done to understand climate and weather. I sometimes think that the scientific outlook on future risk is changing too quickly and therefore loosing credibility. But I find the combination of climate models and risk models very interesting and we are already thinking about the best way to do this soon.

Q -What Machine Learning methods have you found most helpful?
A - Everything that simplifies and classifies. For a long time my favorite algorithms have been PCA/EOF based.

Q - What are your favorite tools / applications to work with?
A - My toolset is rather simple: R, Fortran, csh, bash, emacs and QGIS on Ubuntu Linux. I am still yet to find problems that I can't solve using these. I for some reason never really liked C# and Java, or any of the languages that are popular now (JS, etc.). Too many lines of code to write before something happens. R for me is unbelievable, and one must applaud the people behind RStudio for providing great tools to the community!

Q - And what statistical models do you typically use?
A - We've recently started digging in on much more sophisticated questions. As such, we're deploying the newest most models that an actuary would be using, plus models that reflect nature at a very basic level (Poisson distributions, etc.).

Makes sense - interesting to hear how the field is evolving! Let's talk more about what you're working on at KatRisk...

Q - Firstly, how did you come to found KatRisk?
A - I was very fortunate to start KatRisk with two of the smartest people I know (Stefan and Guy). I always wanted to see what I can do with others (without the limits a large company sets) - plus the excitement is hard to beat when you're out by yourself. The timing was great when we started 1.8 years ago. We are getting good feedback from the marketplace and hope to be a nimble, smart and agile player.

Q - What different types of catastrophes does KatRisk assess?
A - Right now we're doing flood, storm surge and tropical cyclones. We

just released our US and Asia Flood maps, also we have online versions of our tropical cyclone models at http://www.katalyser.com

Q - What have you been working on this year, and why / how is it interesting to you?
A - We are working on our US and Asia models. Overall it's an interesting scientific problem - but we also think that we can have commercial success with these offerings. Nature is quite complicated and to describe natural phenomena with a mix of deterministic and statistical methods is quite challenging.

Q - What has been the most surprising insight you have found?
A - I have gained much more insight into the relationship between climate and extreme weather. Another surprise - on a more personal level - was how much one can work when you have to. The pressure when you are really dependent on yourself is quite different than what you experience in a large company.

Q - Sounds like an interesting time to be working in the field! Can you talk about a specific question or problem you're trying to solve?
A - We're solving the problem of adequately pricing insurance based on each individual building's characteristics. That's why we created 10m resolution flood maps for the US. The basic underlying principle we follow is that risk adverse behavior should be rewarded with lower premiums. We like to believe that we can contribute to a more risk resilient society that way.

Q - What is the analytical process you use? (i.e., what analyses, tools, models etc.)

A - We do everything from finite element modeling on graphics cards written in C, Fortran, CUDA and shell scripts to data analysis with R and QGIS. I like to keep things simple but powerful. As much as I would like to learn new languages and concepts, I can still do everything I need with these tools. We have written all tools and models from scratch by ourselves!

Q - How / where does Machine Learning help?
A - Not as much as I would like to claim. In principle one could apply machine learning to many more problems for insurance pricing. We are more focused on the climate and modeling problem currently. Once we have all these models up and running I would like to go back and take a deep look at ML again. I believe there is a lot of untapped potential.

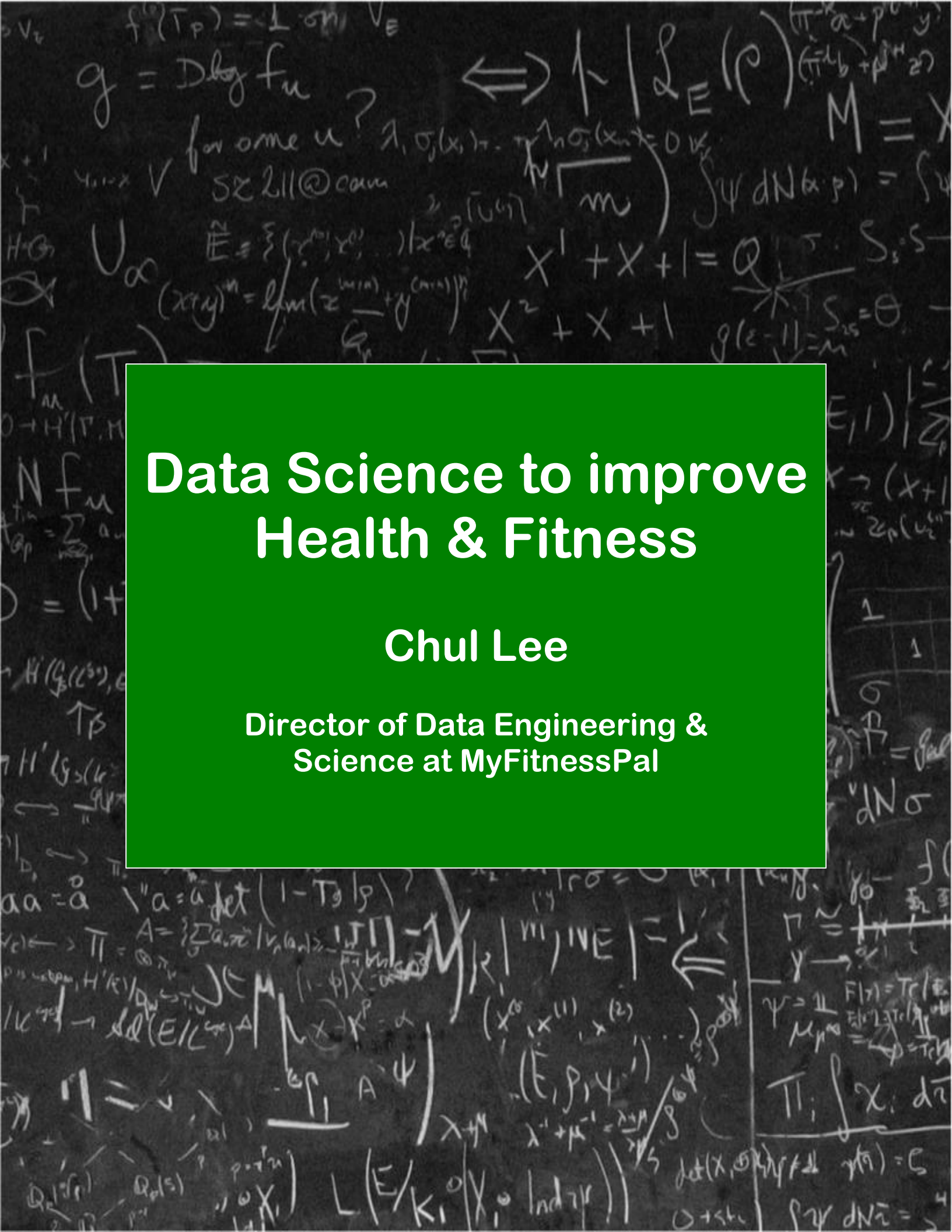
Q - What are the next steps for this product?
A - We have to start marketing it now. We are essentially just done with our first model and data release. We would like to make these products available to a large audience, but that requires that people understand the value for their business. The development on these models never stops. I also think that we are just at the beginning of the open data revolution and that this whole field will look quite different in 5 years.

Dag, thanks so much for all the insights and details behind KatRisk - sounds like you are building some very valuable tools/products! Finally, its time to look to the future and share some advice...

Q - What does the future of Catastrophe Modeling look like?
A - In the future we'll debate complicated issues, such as the cost of climate change, more through data and less through opinions. I really like that science seems to be becoming more open and data more accessible. Catastrophe models and their principles are really at the heart of many discussions when it comes to global change and how we can think about it. I really look forward to that!

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - Learn statistics and R - and find an interesting problem. There is a whole world out there that needs these skills, not just analytics about ads and what people will buy next.

Dag - Thank you so much for your time! Really enjoyed learning more about your background, how the world of Catastrophe Modeling is evolving and what you are working on at KatRisk. KatRisk can be found online at http://www.katrisk.com/.



Data Science to improve Health & Fitness

Chul Lee

Director of Data Engineering & Science at MyFitnessPal

Data Science to improve Health & Fitness

We recently caught up with Chul Lee, Director of Data Engineering & Science at MyFitnessPal. We were keen to learn more about his background, how Data Science is shaping the Health and Fitness industry and what he is working on at MyFitnessPal...

Hi Chul, firstly thank you for the interview. Let's start with your background and how you became interested in working with data.

Q - What is your 30 second bio?
A - I am highly multi-cultural: born in South Korea, grew up in Mexico, and received my high education in Canada. I obtained my Ph.D. in CS from the University of Toronto specializing in Web Algorithms. After my Ph.D, I co-founded a social news aggregation startup, Thoora. I joined LinkedIn with other Thoora members in early 2012. At LinkedIn, I led a team that operated several relevance engines for LinkedIn's content products. I became the head of data engineering & science at MyFitnessPal just 3 months ago.

Q - How did you get interested in working with data?
A - I have been always amazed by the power of computing and data in general. At the same time, I have always appreciated the beauty of mathematics. I originally studied mathematics at college but eventually switched to computer science once I realized that computer science would allow me to pursue all my passions. At graduate school, I studied various web algorithms including PageRank and I further developed my interest in working with data.

Q - What was the first data set you remember working with? What did you do with it?
A - A bunch of text files in MS-DOS. When I was a child, I wrote a simple BASIC program that was able to count the number of lines, words, characters, spaces, etc in these files. Using this simple program, I wrote another program that constructed uni-grams and bi-grams and did a very rudimentary plotting of their distributions.

Q - Was there a specific "aha" moment when you realized the power of data?
A - The core essence of Thoora, the first startup I worked for, was to rank and present each news article by the volume of social buzz around it. At Thoora, I was amazed by how fast and accurately social media can break interesting news stories. For instance, when Michael Jackson passed away, the social media space immediately became full of eulogies for Michael Jackson in the matter of few seconds and it was an incredible experience to observe that phenomenon in actual numbers and stats.

Very interesting background - thanks for sharing! Let's talk in

more detail about the Health & Fitness space and how data/data science are shaping it...

Q - What attracted you to the intersection of data/data science and Health & Fitness?
A - I learned a valuable lesson at LinkedIn in terms of how data products and data science could create tremendous value for users. I had the feeling that similar success could be replicated in the intersection of data/data science and Health & Fitness. Interestingly enough, the nature of many data problems in health & fitness is very similar to that of problems I have worked on previously. The overall idea of being a pioneer in the new field was exciting to me.

Q - Which Health & Fitness companies do you consider pace-setters in terms of their use of data science? What in particular are they doing to distinguish themselves from peers?
A - I think health & fitness big data innovation is in its nascent stage and therefore it is not clear which companies are pace-setters in terms of their use of data science. I think companies like Jawbone, Zephyr Health, Zynopsis, explorys, HealthTap, etc. are attempting interesting and novel ways of applying data science in their product offerings. I personally find the possibility of using IBM's Watson in healthcare very interesting. I would like to say that our endeavor at MyFitnessPal of using data science in some of our product offerings is also interesting.

Q - I think you're right to say so :) ... Given the current nascent stage, which companies/industries do you think Health & Fitness should emulate with respect to their use of Data Science? Why?
A - I think big internet and social media companies are the true pioneers

of data science without calling their approach as "data science". The success of data products developed by companies like Google, ebay, Yahoo, Facebook, LinkedIn, Twitter, etc clearly demonstrates the power of data science. For instance, the Google Translate Service is one of the most interesting and powerful statistical engines based on big data techniques. LinkedIn's well-known PYMK (People You Might Know) is another big success story. Also note that these companies pioneered several tools (e.g. Hadoop, Pig) that eventually became essential in data science. Thus, I think health & fitness companies should try to emulate the success of these companies in the use of data science to tackle their own data problems especially in the development of new products.

Q - Makes a lot of sense! ... On that note, where do you think Data Science can create most value in Health & Fitness?
A - Almost everywhere! I might be slightly over-optimistic here but I think data science can create value in almost every spectra of health & fitness because the right usage of data science will increase the overall information processing power of health & fitness data while providing insights about our health habits, consumption, treatments and medication. Thus, the room for growth of data science in health & fitness for the next few years is big.

That's great to here - and makes for interesting times to be working at MyFitnessPal! Let's talk more about that ...

Q - What are the biggest areas of opportunity/questions you want to tackle at MyFitnessPal?
A - I think offering Amazon like recommendations on what you should

eat and what exercise you should do, based on your dietary preferences would be the biggest areas of opportunity since it will have a big product impact and will trigger technology innovation as well. To achieve that goal, discovering hidden patterns for diets that people are following, how and why their behaviors change over time, why certain diets work while others don't, etc would be very important.

Q - What learnings from your time at LinkedIn will be most applicable in your new role?
A - At LinkedIn I learnt that feature engineering and reducing noise in data are crucial for the successful development of data products. I also learnt the importance of building and operating large-scale data processing infrastructure. Thus, I am currently paying special attention to the development of a scalable data processing infrastructure while making sure that all data cleaning, standardization and feature extraction tasks are well supported.

Q - What projects are you currently working on, and why/how are they interesting to you?
A - There are several feature extraction and data processing projects that I am working on with other team members. One project that I am particularly interested in is the food categorization project which is an attempt to categorize a given food item into one or many food categories. This is particularly interesting because it sits in an intersection of data science and engineering involving different large-scale machine learning, natural language processing, and data analysis techniques. In addition, it will have direct impact on other data science problems that we are trying to tackle.

Q - That sounds really interesting! What has been the most surprising insight you have found so far in your work?
A - The overall variety and complexity of different data problems that have to be tackled in health & fitness. Note that health & fitness touches pretty much every aspect of human activity. Different companies in health & fitness have been accumulating different types of semi-structured and structured data related to health & fitness. The overall variety and complexity of problems that need to be solved in order to better understand health & fitness data is way bigger that I originally envisioned.

Q - And to help you do so you have recently recruited a top notch team from Google, Ebay, LinkedIn, etc - how did you find the recruiting process?
A - As you know, there is a fierce competition to attract top engineering & data science talent in the startup world. Thus, it was not easy to find and secure top talent from Google, Ebay & LinkedIn. We were specifically looking for candidates who had previous experience at companies that had a strong presence in data science since we wanted to emulate their success with respect to their use of data science. We were also looking for candidates that were passionate about building great data products that could have real impact on users. I think our current team members joined MyFitnessPal because they were impressed by the nature and scale of data problems that need to be tackled and MyFitnessPal's overall vision for data products also helped a lot. [Editor Note: see for example this example on Growing the World's Largest Nutrition Database!]

Q - How does your team interact with the rest of the organization?
A - Since we are a startup, my team works with a wide range of functional teams. More specifically, we mainly collaborate on different data related projects with product managers, platform engineers and app engineers. In terms of the decision making process, I would say that we participate in almost every decision making process as long as the given project has some "data" component. It is sometimes challenging to communicate our work/findings with the rest of the organization since not everyone has a strong quantitative mindset. I do not think this is necessarily bad since many times different perspectives can lead to creative solutions. Thus, for a data scientist, having story-telling skills is important (as it has been pointed out by other data scientists). Yet, we have to admit to ourselves that data science, as any other scientific discipline, involves a certain level of complexity and therefore it is important to make sure we pursue scientific rigor while maintaining the accessibility of our work/findings.

Chul, thanks so much for all the insights and details behind what you are working on at MyFitnessPal - sounds like you have a lot of data to play with and are building some very valuable tools/products! Finally, its time to look to the future and share some advice...

Q - What excites you most about recent developments in Data Science?
A - I am most excited about the re-surgance of traditional optimization and machine learning techniques at large-scale in a wide range of application scenarios. Due to this new trend, some traditional algorithms are being re-evaluated and revised to accommodate different computation

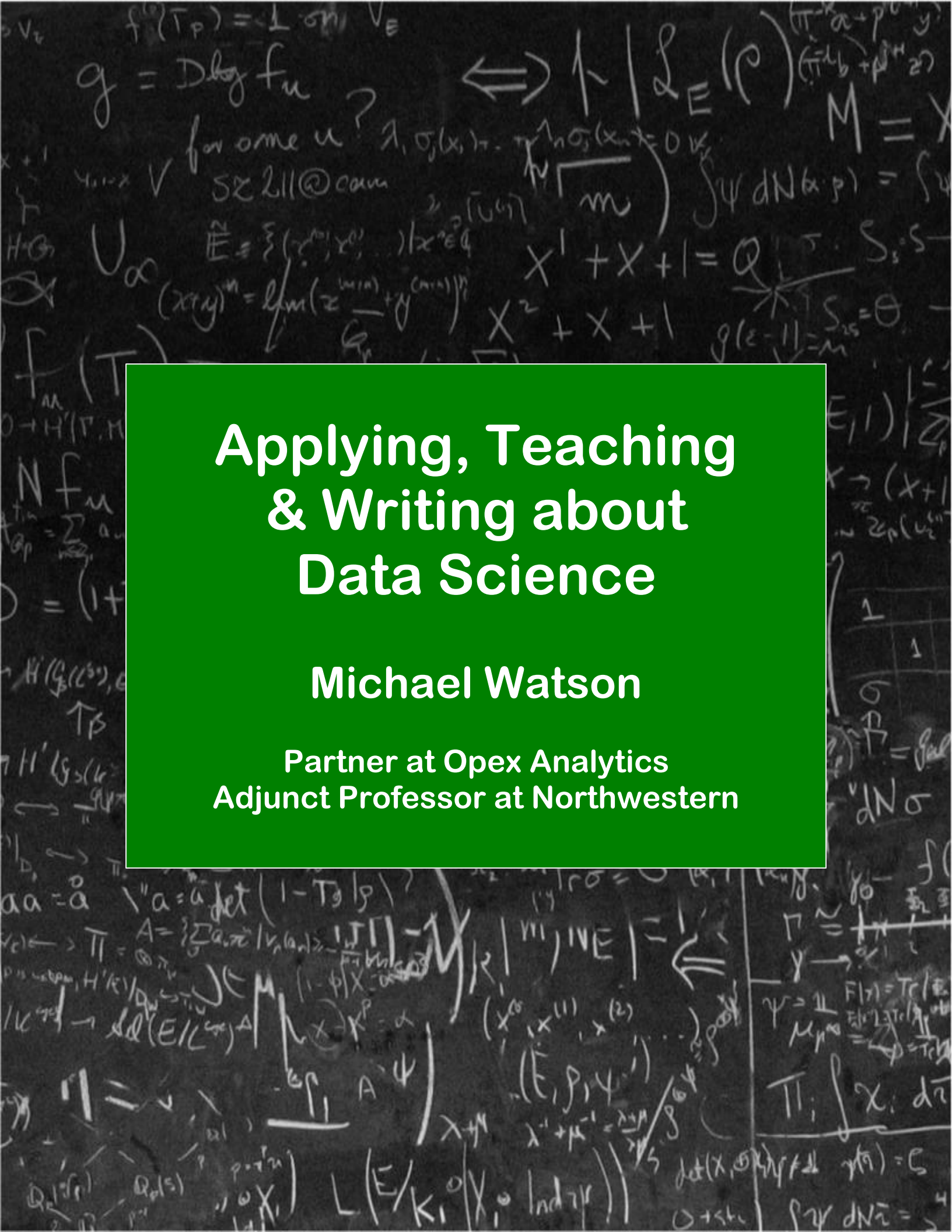
models at large-scale, especially in distributed and parallelized environments. Meanwhile, new computation infrastructures are being proposed to support these new algorithms.

Q - What does the future of Data Science look like?
A - Very promising and rosy! There is no doubt that data is becoming the definitive currency of today's digital economy. I believe that data science will continue finding new applications in many domains. Thus, new problems and challenges in these applications will continue stimulating innovation in data science itself.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Don’t be afraid to explore new exciting and emerging areas in data science like healthcare, ecology, agriculture, green tech, etc. I believe that these new areas will be booming for the next few years and you shouldn't let these opportunities go, especially when you are starting out as a data scientist. As such, I think it is important to have a good preparation in the fundamentals of data science as many of the techniques are somewhat universal and transferrable from one domain to another - and these new areas are not exceptions.

Chul - Thank you so much for your time! Really enjoyed learning more about your background, how data and data science are influencing the Health & Fitness space and what you are working on at MyFitnessPal.

MyFitnessPal can be found online at http://www.MyFitnessPal.com/ and Chul is on twitter @Chul_Lee.



Applying, Teaching & Writing about Data Science

Michael Watson

Partner at Opex Analytics
Adjunct Professor at Northwestern

Applying, Teaching & Writing about Data Science

We recently caught up with Michael Watson, Partner at Opex Analytics and Adjunct Professor at Northwestern University, where he teaches Analytics in two of their masters programs. We were keen to learn more about his background and his various current roles and projects...

Hi Michael, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - I am a partner with a new company called Opex Analytics. We are helping bring new data science techniques and products to operational companies (retailers, manufacturers, etc). I'm also an adjunct professor at

Northwestern University, teaching in two of their masters programs. Previously I was at IBM.

Q - How did you get interested in data / analytics?
A - I have always been interested. I used to track statistics of our sandlot baseball and basketball games as kid — probably much to the annoyance of the other kids. And, this carried with me through schooling — I loved my math classes and I loved the idea that mathematical models could help you understand the real world.

Q - What was the first data set you remember working with? What did you do with it?
A - Back in 1989 (or so) when I was a sophomore in college, I got our school to enter in the Mathematical Contest in Modeling. This gave us access to a realistic data set that we had to analyze and come up with a solution over a weekend. I can't remember the tools we used to analyze the data, but we calculated basic statistics and built simple models to analyze the problem. This was a great way to see that data could have an impact on real problems.

Q - Was there a specific "aha" moment when you realized the power of data ?
A - In the summer of 1991, I was north of Cairns, Australia, in a small camp in the middle of the rainforest. After dinner, I had the most interesting conversation with someone who was convinced that mathematical modeling was the way to solve complex and real business problems. It was during this conversation that I decided to take the career path that I did.

Very interesting background - thanks for sharing! Let's talk in more detail about the work you're doing now at Opex Analytics...

Q - What attracted you to the intersection of advanced analytics and operations?
A - At IBM (and ILOG and LogicTools through the successive acquisitions), I was focused on helping companies apply optimization to make better supply chain decisions. So, I knew the challenges faced by operational companies. I wanted to be in a position to bring the new data science ideas to these firms. Combining techniques like optimization and data science can be a powerful combination.

Q - What has been the most interesting project you have worked on? Why?
A - The most interesting project I worked on was a merger of two large manufacturing companies. We had to help them figure out which facilities to keep, which ones to remove, and how to structure their supply chain. It was interesting because the savings were huge (more than $100 million) and the project received the attention of the CEO. And, the project was mentioned several times in the CEO's quarterly earnings announcements.

Q - That's great! And how did advanced analytics help?
A - There is no way to solve these types of problems without advanced analytics. You have an enormous amount of data to sort through, and you need to build models that help you analyze the alternatives.

Q - Where do you think advanced analytics can create most value going forward? (i.e., what industries, what types of problems etc.)
A - I think that a lot of web companies have been taking advantage of analytics for a long time. There is a lot of potential to apply some of these ideas to traditional manufacturing companies, traditional retailers, or traditional service providers. These firms have a lot of data that they are not leveraging.

Makes sense! Let's switch gears and talk in more detail about your teaching - you've been an Adjunct Professor at Northwestern for a long time (since 1999) ...

Q - What are you teaching currently?
A - I've recently started teaching in Northwestern's new Masters in Analytics program (a great program for readers of this newsletter to check out) and have started teaching "Managerial Analtyics" in Northwestern's Masters in Engineering Management (a program for engineers who are working full time and want a more technical business masters degree).

Students are hearing a lot about analytics, but aren't quite sure what it is. It is exciting to be able to teach a new and fast-moving field to students who are eager to learn more about it. And, I get to learn a lot by teaching analytics.

Q - How has the subject matter changed since you first started teaching? What has driven the evolution?
A - Back in 1999 when I started teaching, no one was talking about analytics. And, my course was focused on operational excellence — how to

run a better factory, a better supply chain, or a better services operation. Now, with the rise of analytics, managers realize the importance of data in many different parts of the business. So, there is a desire to learn the basics of data science so managers can apply it to whatever area they happen to be working in.

Let's switch gears and talk about your new book - Managerial Analytics: An Applied Guide to Principles, Methods, Tools, and Best Practices - which has been receiving terrific reviews...

Q - What does the book cover?
A - The first thing the book does is to help people understand what the field of analytics and Big Data is all about—what do these terms mean, what do they include. When we first started researching analytics (and Big Data), we found that a lot of people were using the terms, but not defining them. And, based on how people were using the terms, they were not talking about the same thing. Second, we discuss what it means to have an analytics mindset—how do you need to think about data. And, then we devote the third section of the book talking about the science behind analytics. The science section is meant to show the manager that these techniques are accessible and show how they might be used.

Q - Who is the book best suited for? What can readers hope to learn?
A - The book is best suited for managers who need to understand analytics or for people who are just getting into the field of analytics. The book paints a broad picture of the field and helps people understand how the different pieces fit together and what the different terms mean.

Q - What was your favorite part of writing the book?
A - Since the field is moving so fast, a lot of people are coming out with important ideas or new ways to look at things. My favorite part was learning about new developments and incorporating these into our book. Readers have told me that they have found our bibliography and references in the endnotes very helpful.

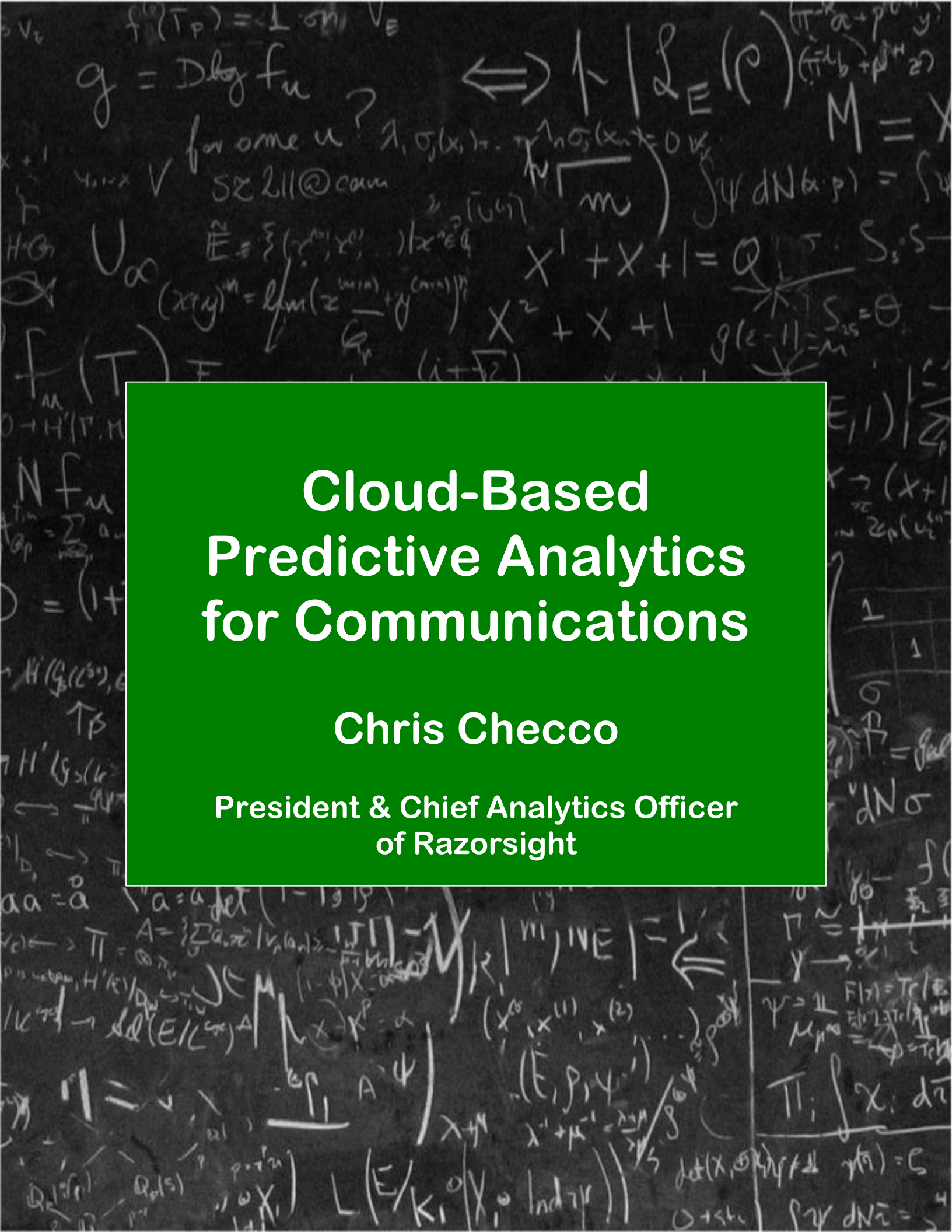
That's great - congratulations! Finally, let's talk a little about the future and share some advice...

Q - What excites you most about recent developments and the future of Advanced Analytics?
A - I like the fact that the technology of advanced analytics continues to push into new areas. But, just as important, it is great to see that companies are starting to embrace the idea that they should be taking advantage of the data they have. This combination should allow us to see many new developments that have a big impact in the marketplace.

Q - Any words of wisdom for Advanced Analytics students or practitioners starting out?
A - For analytics students, make sure you also understand business — to make analytics stick in a company or organization, you need more than just the technology. And, for companies and organizations just starting out, my suggestion is to start with a small team, pick a few projects and grow from there.

Michael - Thank you so much for your time! Really enjoyed learning

more about your background and your various current roles and projects! Opex Analytics can be found online at http://OpexAnalytics.com and Michael on his faculty page.



Cloud-Based Predictive Analytics for Communications

Chris Checco

President & Chief Analytics Officer of Razorsight

Cloud-Based Predictive Analytics for Communications

We recently caught up with Chris Checco, President and Chief Analytics Officer, Razorsight. We were keen to learn more about his background and perspectives on the evolution of cloud analytics, how data science can impact mobile/communication companies and what he is working on now at Razorsight...

Hi Chris, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - At the heart of it all, I'm just a data geek with some business skills who thrives to help others make informed decisions. I've spent my entire career delivering insights to businesses to help them run their business more effectively, whether it be for selling more products, identifying fraud, or finding hidden cost leakages. I went from database programmer to project manager to management consultant to executive. That path has

helped me understand the world of analytics from the ground up – I feel very lucky in that sense.

Q - How did you get interested in working with data?
A - It started when I was a kid eight or nine years old, collecting comic books. While I never read a page of the comic books, I was interested in the economics of them; I would track the price changes of my comics from month-to-month on graph paper to see where the value was, and then try to invest the bits of money I made from cleaning pools, gardening, and delivering newspapers in the comic books with the best future value. Sad, isn't it?

Q - Very smart! :) ... What was the first data set you remember working with? What did you do with it?
A - From a professional perspective, the first dataset I worked with was a mobile phone customer database back in the late 1980s. I learned how to program on that dataset, tracked and reconciled commissions and made a marketing database for myself to call customers for upgrades and promotions. It was a crude green-screen system, but it ultimately helped drive sales and reduce commissioning leakage.

Q - Was there a specific "aha" moment when you realized the power of data ?
A - While I spent many years futzing with data in many ways, I really learned the power of data when I was working as a programmer at a wireless provider as a contractor. I was working with a statistician for the first time, trying to create some funky dataset they could use for their analysis on a repeatable basis – it was new ground for me. That statistician was thoughtful enough to show me how she was going to use

the data, and shared some of her results with me. I was absolutely blown away. While the reality of advanced analytics is complex, the concept was so simple and natural that I immediately started thinking how else we could apply this magic to other parts of the operation. That moment was a turning point in my professional life.

Very interesting background - thanks for sharing! Given your current role, let's talk in more detail about the world of Cloud Analytics...

Q - In research published recently, Gartner analyst Dan Sommer says, "we're at the cusp of a series of tipping points which will facilitate unprecedented interest and adoption of analytics." Do you agree, and if so, what's driving this uptick?
A - I think Dan is absolutely right, however, I would say that statement has a long tail. Here's what I mean. While advanced analytics has been around forever, and really hasn't fundamentally changed since the early days of agricultural use, each industry has it's own adoption curve. If you went to any major credit card provider, insurance provider or oil company, they would have been using these advanced techniques for decades, at least in certain areas of their businesses. Other industries, such as telecommunications, are in the midst of this adoption phase. Others have yet to adopt those technologies. Dan's statement is profound in another way - there is a proliferation of advanced analytics across industries, with less mature industries borrowing the lessons learned from the more mature industries.

Q - What have been some of the main advances that have fueled the move toward cloud based analytics? How do you see this playing out going forward?
A - I think there are three main drivers in the march towards cloud analytics, and believe the pro-cloud trend will continue. Cost – the investment in the infrastructure, maintenance, and migration has plagued companies since we started using computers. The cloud removes much of that pain.

Second is convenience – the ability to stand up a system in minutes, hours, or days, versus months or years, means business can move at the speed of business rather than at the speed of technology integration.

And finally, core competency. Companies are realizing that cloud applications level the playing field in terms of the technology advantage. Companies need to put their energy into delivering the best products within their industry their core competency, not into the latest and greatest super computer, tool, or programming language.

Makes sense! Let's switch gears and talk in more detail about your the Communications and Mobile industry, given that's your current focus...

Q - For years, the communications and media industries focused on Business Intelligence platforms, but with only marginal performance improvements. Now, with the move to Big Data analytics, data science is all the rage. What precipitated this shift?

A - The realization that BI, while still a necessary and beneficial tool, has limitations if used alone. Imagine the first explorers to the New World trekking across the ocean with a rudimentary map, compass, and a vague idea of where to go. That is BI alone – the people asking the questions aren't armed with solid facts. The explorers, similarly, were limited by their vision, assumptions, and knowledge. Now imagine if those same explorers had todays GPS technologies and knew precisely the route they were taking and the route to get there. It drives intelligence far beyond because it can see the big picture (which couldn't be seen by the explorers) and the details all at once. It provides an answer to a question – what precise direction do I need to go to get to India, and what is the best way to get there. Big Data Analytics provides the latter … and that shift, as has happened with many technologies, is increasing in utility and benefits as it gets used more broadly.

Q - What are the biggest areas of opportunity/questions you want to tackle in the $2 trillion global communications market? Do you see one sub-area such as cable or mobile as the most ripe for disruption?
A - While advanced analytics has great potential in almost every aspect of the global communications industry, there are 2 specific areas that are most exciting.

First is the addressable advertising space – this is an area that has had slow traction since its inception a couple of years ago, but it promises to deliver up to 1000% improvement on a significant portion of ad revenues for cable and satellite TV operators. The historical gap was being able to deliver these targeted ads, but that gap has been filled by the latest ad delivery platforms and IP-based set top box technologies. The gap today is

a sales challenge – accurately estimating the impressions at a granular level and optimizing sales to that volume.

The second area that provides significant growth potential is data monetization. Many communications providers are taking big bets on this space, but are in uncharted territory in relation to their traditional businesses. The first gap is the business model – how will they actually drive significant revenue streams. Second is the differentiation, since all of the major players are working toward the same goals with similar data availability. Analytics in my mind will both provide vast differentiation and enable them to have more options in terms of their business models. For instance, selling data versus enriching data versus selling insights warrant three different levels of compensation. Analytics enables a higher multiple.

Q - You've said that customer retention continues to be a major issue for cable companies as well as mobile operators, where recent research shows that every quarter some 16% of postpaid customers say they'd like to switch service providers. What's behind these high levels of "churn" and how can Data Science help?
A - The desire to switch carriers is driven by two key factors. First are the reduced barriers to exit and entry. I can now cancel my service and have my termination fee paid by the competition. Second are the alternative options for consumers. While cable TV is a must in most households, the younger generation is using IPTV as their television medium.

Data science can aid organizations in managing retention in several ways. Historically, these organizations wanted to know who was going to churn,

and then provide them an offer. Today, data science can help build a better picture of the customer, telling the operators why they are dissatisfied, how much they will be worth, how to best address the customers' needs, and the best communication options. It's the difference between throwing a steak at someone who's hungry or giving them a seven-course, well-balanced meal that meets their palate. The latter provides a richer experience and a more satisfied customer.

Q - Which companies or industries do you think communications companies should emulate with respect to their use of data science – and why?
A - Casinos. The best-in-class casinos have the customer experience down to a science – literally. They have created systems to gather information on their clients and turn those into actionable insights for each and every one. In the pre-data-science days of casinos, only the wealthy players got a personal experience – the 1% of the 1%. Today, at the analytically driven casinos, customers are treated based on their individual profiles. This drives more repeat business, more time in the casinos, more money changing hands, and a better overall customer experience.

Now, 18 months or so ago you joined Razorsight, let's talk more about that…

Q - What guided your decision to join Razorsight in late 2012?
A - The decision was one of the easiest of my careers. After spending many years consulting for companies and government agencies, and designing/building one-off analytic platforms, I knew there had to be a better way. With these customized solutions, I watched customers either

struggle to support it themselves or pay huge consulting fees to keep them alive. I also witnessed them struggle to hire and keep the scarce resources required to fully leverage the platforms.

Razorsight provided the ideal opportunity to create a leading edge, cloud-based solution that avoided the one-off approach, avoided the need for these scarce resources, but still provided the power and results of a customized solution. And that is precisely what we've built and delivered on top of the mature cloud-platform that Razorsight has perfected over the last dozen years. The ideas that have been culminating and evolving for years have come together in a repeatable, business-user application that doesn't skimp on anything ... and now you can see why the decision was easy.

Q - Razorsight has attracted some big name clients -- AT&T, Verizon, Comcast, T-Mobile, Dish, Facebook and some 80 other leading brands. What does Razorsight do in the areas of data science and advanced analytics, and why does your work matter to these companies?
A - It's really simple – we help these organizations make better fact-based decisions by providing them insights that they can't create at scale or through brute force. We do it on a repeatable basis at scale delivered through a business-user-friendly interface. They ask the tough questions, and our RazorInsights system provides the answers and insights. While there's no silver bullet to addressing our clients' business needs, our goal is to help them continuously and to cost-effectively "turn the dial" in areas like customer acquisition & retention, cross-selling, operations, and advertising.

At the core, our client base struggles with one thing in relation to decision-making – Time Management. The first part of the Time Management is making timely decisions – business is moving faster and faster, and traditional analysis can't keep pace. The second part of Time Management has to do with where resources are spending their energy – some customers have estimated that they spend roughly 80% of the time munging data, 15% analyzing the data and 5% making decisions. Hence the analysis and decision-making suffer because of the long data preparation time. If they can eliminate all or most of the data munging time, and have a stronger starting point for their analysis, they can put more effort into the advanced analyses and decision-making processes. It only stands to reason that better decisions will be made and at the speed their business requires those decisions.

Q - What makes Razorsight's approach new or unique, and how does it differ from prior attempts to deliver insights on the customers' mindsets "in the moment"?

A - The uniqueness of our approach is that we don't just provide more data or answer a one-dimensional question – rather we surround the core answer with additional critical decision-making facts. If I'm lost and dehydrated, I don't just need to know where water is located, I also need to know if the water is drinkable, the best way to get to the water, the risks in getting there, and how much water there is available. Our solution provides these types of critical answers, but for the communications industry. If we just tell them a customer is likely to churn, there are still many unknowns. So we also tell them why the customer is likely to churn, what they should spend on that customer to retain them, how they should entice the customer to stay, and where they should interact with that

customer. We can even tell them the likelihood that the customer will accept a specific offer, and how long they will likely stay after they accept.

Q - What projects have you been working on this year, and why are they interesting to you? What has been the most surprising insight you have found?
A - While I can't give provide specificity on discrete customer insights, as they are highly confidential, I can tell you that there are definitely surprises. Typically, 60% of the insights were known or perceived – while one might state we are simply replicating existing insights (which is partially true) we are also invalidating some of the previous insights and assumptions. This is the truth serum for perceptions and assumptions, as myths and broad statements are dispelled. One of my favorite examples, which I've seen across multiple clients, is the "usage trap", that is, the perception that heavy users are highly satisfied and lighter users, less satisfied. Hence they spend enormous marketing dollars trying to drive higher usage rates. What they don't realize is that the pure usage, in many cases, has little correlation to satisfaction level.

Roughly 20% of the insights are things they had an idea about, but couldn't quite put their finger on it with confidence or accuracy. The remaining 20% of the insights are the true "aha" moments – these are the ones that, many times, go against conventional wisdom and drive positive changes in the way the business thinks, acts, and communicates with customers.

Q - Very interesting!... It's also been said that IT departments are the place data science and analytics go to die because traditional IT is just too

slow on the draw. How does Razorsight deal with this challenge -- do you have greater success approaching different departments within a company? What, if anything, would make this easier?
A - We do interact with IT, but our buyers are generally business buyers: Marketing, Sales, Finance, Operations, and Media teams. It's because we are not selling a technology or tool, per se, but rather a solution that directly addresses business issues. IT typically has a set of enormous challenges, namely to keep everything running smoothly while there are a thousand changes in place and external forces driving new behaviors. In this sense, I see IT as the factory – they have to be great at operationalizing massive processes, putting in fail-safe measures, and dealing (calmly) with the daily fires that arise from the unforeseen. he core of advanced analytics simply doesn't fit this mold – it's not rigid system, which once defined, runs forever. It adapts as the business changes.

Think for a moment about a triage unit. General medics handle the easiest cases. This is equivalent in the business world to the group-specific analysts – they need to be smart and very efficient at what they do, but they can do it with the tools they have on their desks on a small scale. The moderate cases require surgeons to perform more complex but routine operations to address the moderately wounded. These take longer, require more training and experience, and have larger impacts if not done correctly. This is your IT group – harder issues, large scale, and repeatable.

Severe cases require ER doctors who can handle a myriad of issues, get

them to a point of stabilization, and then pass them to the surgeons. These are your data scientists. Give them a complex question with a little direction and let them go to work. Once they find an answer, the outputs can be operationalized in the IT systems. level.

That makes a lot of sense! Thanks for all the insights. Finally, let's talk a little about the future and share some advice…

Q - What excites you most about rrecent developments in Data Science?
A - The best development in recent Data Science history, in my humble opinion, is the realization that Data Science is a requirement. Companies can no longer afford to live without this staple and realize continued success. Things are getting better all around due to the collateral impacts – companies become more efficient at their core business, they find new ways to expand into new businesses, and the customers get new and better products at more competitive prices.

Q - What does the future of Data Science look like?
A - Well, I don't have a predictive model for that question (!). However, I hope to see two things: First, I would like to see some revolutionary advances in the scientific aspects – new algorithms and techniques which I believe will emerge in the next decade. Second, I'd hope to see the creative side prevail as much as the scientific aspects – that is, creating new innovative uses for the science, just as folks like Pandora and Match.com have over the last decade.

Q - Any words of wisdom for Data Science students or practitioners starting out?

A - These students and early-life practitioners have such a leg up – I'm a bit jealous. This is because they are getting into Data Science at an extremely exciting time, and will be adding value in ways that some of us more aged folks didn't have the opportunity to contribute. With that preamble, here's my advice.

Experience is King – get as much experience as possible in the shortest amount of time, across a broad spectrum of applications.

There is no silver bullet – analytics is a vast science, and it must be realized that an approach that provides an answer today may not work on the same question tomorrow. Use all of the tools at your disposal and please do not rely on one approach. And finally,

Innovate – the uses of data science are still emerging, so take the non-traditional route sometimes and experiment. Taking a chance may pay off in spades.

Chris - Thank you ever so much for your time! Really enjoyed learning more about your background, your perspectives on cloud analytics and what you're working on at Razorsight. Razorsight can be found online athttp://www.razorsight.com.

Creating the "Dropbox of your Genome"

Reid Robison

MD, MBA
CEO at Tute Genomics

Creating the "Dropbox of your Genome"

We recently caught up with Reid Robison, MD, MBA and CEO at Tute Genomics. We were keen to learn more about his background, his perspectives on the evolution of genomics, what he's working on now at Tute - and how machine learning is helping...

Hi Reid, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - Physician & genetics researcher turned data scientist. Serial entprereneur. Studied neuroscience as undergrad, while doing mouse genetics studies. Then med school by day, MBA school by night.

Completed psychiatry residency then jumped into a genetics fellowship focused on autism gene finding. Did a post-doc in Bioinformatics to try and merge computer science & genetics. Joined the faculty at the University of Utah doing research for a while, running a genetics lab, conducting clinical trials. Then left to start my first company, Anolinx, a health-care data analytics firm, in the early stages of the big data hype. We mined health data from big hospital systems, and queried it to answer big reserach questions for pharma. Pharmacoepidemiology, pharmacovigilence, virtual registry studies. Cool stuff. Then it was acquired pretty quickly and is still doing quite well. I had started and sold another company in the life sciences space, and was looking for my next thing. That's when Kai Wang, computational biologist at USC, and I started brainstorming about a genetics venture, and Tute Genomics was born.

Q - What drove your transition from being a Physician to a Data Scientist?

A - Late last year, Vinod Khosla said "By 2025, 80 percent of the functions doctors do will be done much better and much more cheaply by machines and machine learned algorithms". Treating patients one by one was personally satisfying, and I got to help a number of people. But it was hard work, and there was never enough time to help everyone. Long waiting lists, people without insurance, imperfect treatments. Our healthcare system is a bit of a mess... a huge mess in fact... I wanted to move to more of a macro level in terms of conducting research, solving big problems with data science, changing the way we practice medicine.

Q - Very inspiring! ... So, how did you get interested in working with data?

A - I was about 11 or 12 years old when we got our first computer. An early Texas Instruments, I think. It didn't come with any games. Instead, I had to check out a book of code from the library, type in hundreds and hundreds of lines, hit the go button and voila! A game appeared. It was magical, and I've had a bit of an obsession with computers and intelligent machines ever since... And then as time went by I started to become more and more fascinated with all the unanswered questions out there, especially when it came to the human brain and the human genomes - both are massive unexplored frontiers that are very ripe for discovery.

Q - What was the first data set you remember working with? What did you do with it?

A - During my genetics fellowship in 2009, I was doing 'linkage analysis' on large extended pedigrees. Linkage is basically a statistical approach to finding 'peaks', or regions of interest, in the genomes of large families (and groups of families) with a certain disease or phenotype. Back then, we didn't really sequence 'genomes' per se, but we had panels of markers scattered across genomes, instead of information from every letter in the genome like we can do now. I had 6000 markers on each person from a linkage panel chip, and I had this for hundreds of people across dozens of large families. I set it up using some command-line, open-source linkage analysis software called MCLINK on a linux box under my desk at the University of Utah, and it sat there processing for, I kid you not, over a month. I would stare at it every day saying "please don't crash". Eventually, it worked, and we got some decent results from it: http://www.molecularautism.com/content/1/1/3

Q - That's amazing - great ingenuity and patience! :) Final background question ... Was there a specific "aha" moment when you realized the power of data?
A - This study I just mentioned was a bit of an 'aha' moment for me. It was impressive what we could do with just 6000 genetic markers... I couldn't help but wonder what we could find if we had access to all 6 billion letters in the human genome. My research focus shifted from these 6000 marker panels to microarrays with 250,000 genetic markers, then 500k, then 1 million... By then, next-generation sequencing was becoming available so I jumped right into trying to figure out how to use whole exome and whole genome sequencing for gene discovery in autism and other neurodevelopmental conditions, as well as helping to develop the tools to make this possible.

Very compelling background - thanks ever so much for sharing! Let's switch gears and talk in more detail about the "genome revolution"...

Q - What have been some of the main advances that have fueled the "genome revolution" in recent years?
A - The cost of sequencing the human genome has dropped over one million fold and this is literally transforming healthcare as we know it. The $1000 genome was announced this year, and it is now cheaper to sequence the entire human genome than it is to order a bunch of single gene tests. Instead of paying thousands of dollars for a few specific genetics test, why not pay a fraction of that amount to sequence your entire genome and get information on all of your 25,000 genes at once.

The problem is that no-one, until now, could handle this massive amount of data.

The advancements in sequencing technology are definitely making it more accessible to use genomics to solve medical problems. This combined with the research insight to use genomic science to design treatments and prevention strategies for major diseases is pushing society to become more accepting of genomics and promote putting resources into this industry...

Q - What are the main types of problems now being addressed in the Genomics space?
A - Our biggest problem in this space is how to quickly translate enormous amounts of sequencing data into information that can be used to fuel discovery. The industry is really advancing with sequencing technologies, and so many researchers and labs have the data, but they don't have the time and resources to make sense of it at the pace that patients and society would like to see it get done. We are basically delayed in making major strides toward understanding and treating disease.
See my post here about the size of the human genome and the problems this causes in terms of bottlenecks in data transfer and processing.

Genomics also faces an issue within the general knowledge base. We need more participation in the collection and distribution of human genomic data to identify disease causing variants, variants responsible for drug response, and so on - and this information needs to be located in a central database which is easily accessed from anywhere, by any researcher.

Q - Who are the big thought leaders?
A - My co-founder, Dr. Kai Wang, is a well-known computational biologist in this space and wrote software called ANNOVAR that quickly became the gold standard in genome analysis and has been cited by over 750 scientific papers now

Another researcher I admire is Dr. Gholson Lyon (physician scientist at Cold Spring Harbor Laboratory in New York) who led the discovery of Ogden Syndrome, a rare, previously undiagnosed genetic disorder that he named after the families he worked with in Ogden, Utah. You can read his account of the discovery here.

Q - What excites you most about bringing Genomics and Data Science / Machine learning together?
A - It's all about getting things done quickly and accurately. Training our machines to identify novel disease causing variants is priceless, this alone can eliminate months or more of work from a research project.

Q - What are the biggest areas of opportunity / questions you would like to tackle?
A - Before long, everyone will get his or her genome sequenced. Your genetic blueprint can and should service as a reference for you and your doctors to query at every important medical event and decision throughout your life. Someone needs to be the keeper of that data, in a secure, accessible and meaningful way. That's what we're working on at Tute Genomics.

On that note, let's talk more about what you're working on at Tute...

Q - What specific problem does Tute Genomics solve? How would you describe it to someone not familiar with it?
A - Tute is kind of like the dropbox of your genome - we are a big data cloud-based platform that lets researchers & healthcare organizations analyze entire human genomes. By doing so, Tute is opening a new door for personalized medicine by helping researchers and clinicians interpret genetic variants and find disease-related genes.

Q - That sounds great! Could you tell us a little more about the technology - firstly, how does it work?
A - Tute Genomics is a clinical genome interpretation platform that assists researchers in identifying disease genes and biomarkers, and assists clinicians/labs in performing genetic diagnosis. Given sequencing data on a genome or a panel of genes, Tute can return over 125 annotations on variants and genes, perform family-based, case/control or tumor sample analyses to identify causal disease genes, and generate clinical reports for clinicians to focus on clinically relevant and actionable findings.

Q - How is Machine Learning helping?
A - Machine learning enables our software to quickly go from DNA to diagnosis. The Tute platform uses machine-learning algorithms to score and rank all genes and genetic variants in a given genome by their likelihood of causing disease. We call this the Tute Score, and it's used to predict whether a genetic variant is likely to be damaging or disease-

causing. This machine learning approach shows much improved predictive power compared to traditional approaches, based on cross-validation of a number of genetic data sets. We have acquired multiple public and proprietary databases, along with commonly utilized genetic scoring algorithms, and we utilized Support Vector Machine (SVM)) to build & train the predictive models. SVM is a supervised classifier in the field of machine intelligence. The classification is formulated as the optimization problem to identify the optimal hyperplane that creates the biggest margin between the training points for neutral and deleterious variants/genes. More importantly, linear separability can be obtained in an expanded input feature space by using kernel functions. First we identified a set of functional prediction scores for which coding and non-coding variants can be assigned into. We then built and tested SVM prediction models using a variety of kernel functions and other parameters. The SVM models were optimized using known disease causing mutations from our test data sets.

To comprehensively evaluate the false positive and negative rates of this approach, we've been validating the Tute score on both synthetic and real-world data sets... So far so good, and we've been able to crack undiagnosed genetic diseases in a matter of minutes when you combine our annotation engine and these machine learning algorithms.

Q - Very impressive! What further advances could the Tute approach / technology enable going forward?
A - We are excited about the opportunity we have to make a meaningful dent in the universe accelerating precision medicine by unlocking your genome, personalizing treatment, and powering discovery. This is a

massive amount of complex data, and we are making it accessible and useful so that we can all query our genomes at every important medical question throughout our lives.

In terms of next steps, we are already starting to integrate with patient health records, so that genomic data can be accessible where it is most useful and actionable. We are basically sick of our messed up healthcare system and are on a mission to accelerate progress towards patient-centric, precision medicine!

That's a great goal - good luck with next stage of the journey! Finally, let's talk a little about the future...

Q - What does the future of Genomics & Data Science look like?
A - Our healthcare is not yet personalized to each of us as individuals. When you receive a prescription for blood pressure medicine, or cholesterol, or even for cancer, there is a very real chance that it may be the wrong medicine for you, or even the wrong diagnosis. Fast forward a few years to a world where your medical treatment can be 100% unique. Every diagnosis, every treatment, every drug and every dietary change is tailored to you and you alone. Every treatment works in a predictable way. When you walk into the hospital, instead of feeling like a car on an assembly line, you can be treated like the unique human being you are. Instead of seeing a specialist in a certain field of medicine, the information in your genome can turn any doctor into a specialist in YOU. All of this, thanks to one test: whole genome sequencing, and Tute

Genomics software technology - an app to unlock your genetic blueprint and enable genome-guided medicine.

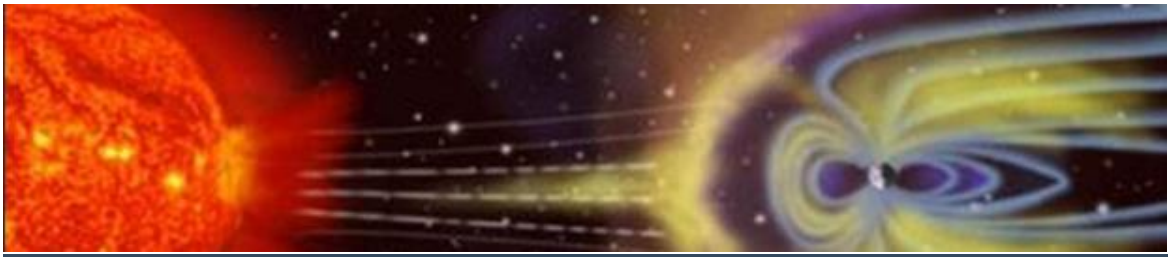
Reid - Thank you ever so much for your time! Really enjoyed learning more about your background, your perspectives on the evolution of genomics, what you're working on at Tute - and how machine learning is helping. Tute can be found online at http://tutegenomics.com.

Data Mining at NASA to Teaching Data Science at GMU

Kirk Borne

Trans-Disciplinary Data Scientist
Professor of Astrophysics & CS at GMU

Data Mining at NASA to Teaching Data Science at GMU



We recently caught up with Kirk Borne, trans-disciplinary Data Scientist and Professor of Astrophysics and Computational Science at George Mason University. We were keen to learn more about his background, his ground-breaking work in data mining and how it was applied at NASA, as well as his perspectives on teaching data science and how he is contributing to the education of future generations...

Hi Kirk, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - I am a trans-disciplinary Data Scientist and Professor of Astrophysics and Computational Science at George Mason University. My professional career was primarily astrophysics for two decades, but I focused on data systems for large space astronomy projects at NASA during the years following my graduate and postgraduate work. That focus on data led me into the field of data science starting in 1998. I left my NASA position and moved to GMU in 2003 to pursue two things: data science research and the creation of the first data science undergraduate degree program in the world.

Q - What appealed to you about getting a doctorate in astronomy from CalTech?
A - Caltech was (and still is) the world's leading university for astronomy and astrophysics graduate education. Ever since high school, my goal was to go to Caltech and use the big telescopes at Mount Palomar Observatory, whose astronomical images appeared in all of the astronomy books that I read during my youth. In order to pursue a career in astronomical research, a PhD is required, and Caltech is second to none for that.

Q - What was the transition from Academics to working at NASA like?
A - The transition was mostly seamless for me since I used large telescopes at Caltech and in my subsequent postdoctoral research positions at University of Michigan and the Carnegie Institution of Washington. Therefore, the fact that my first job was as a supporting scientist for NASA's Hubble Space Telescope (HST) was a natural step for me. The HST Science Institute in Baltimore was growing into becoming the absolute best astronomy research institute in the world at that time (e.g., three of its associated scientists have won a Nobel Prize in the past dozen years). I wanted to be part of that growth and that telescope. My research on colliding galaxies continued throughout and beyond that transition. There was really no transition, other than the normal one that occurs when going into your first real job.

Q - Makes sense ... So what was the first data set you remember working with? What did you do with it?
A - Well, if you want to talk "small data", I worked with professors on two small astronomy data research projects when I was an undergraduate at

LSU in the 1970's. For one of those, I analyzed data on the hottest and bluest stars in the Milky Way Galaxy, which led to the discovery of some very unusual stars (which we now call cataclysmic variable stars). For the other project, I helped create the discovery star charts for some "stars" that didn't seem to be stars at all – many of these turned out to be quasars. That was very exciting. My first independent data project as a graduate student was to analyze the shapes and distortions of colliding galaxies, as observed through astronomical images obtained at Palomar Observatory. I was able to use those distortions to infer the masses and orbits of the colliding galaxies – I was one of the first astronomers in the world to do that.

Q - That's amazing - very impressive :) Final background question ... Was there a specific "aha" moment when you realized the power of data?
A - As an astronomer, I have used data my whole life. There was never really an "aha" moment with that. But there was a huge "aha" moment when I realized that the volumes of data that we are creating in science were reaching astronomical proportions (pun intended!). That occurred in 1998, when the astrophysics data center at NASA (where I was working) was offered a two-terabyte data set from a single experiment. That data set was larger than the cumulative volume of all of the previous 15,000 space science experiments that NASA had flown during the previous 40 years of NASA history, combined! I knew at that moment that things were drastically changing, and the power of data for new discoveries was now growing beyond our wildest dreams.

Very compelling background - thanks ever so much for sharing!

Let's switch gears and talk in more detail about data mining, and your time at Raytheon onward...

Q - Tell us about what you learned from founding and becoming the Co-Director For Space Science in which you and your team carried out research into scientific data mining...
A - I was working as a Raytheon contract department manager in NASA's Astrophysics Data Facility from 1995 through 2003. In 1998, I realized that that the huge increase in data volumes were leading to huge potential for new discoveries. To achieve those discoveries, we needed the special machine learning algorithms that are used in data mining. I began devoting all of my research time to data mining research, initially on the very same colliding galaxies that I had previously studied "one at a time" but now "many at a time."

By 2001, I had developed something of a reputation as a leading data mining researcher at NASA. I didn't realize that was happening until October 2001 (about one month after the horrible events on September 11, 2001) – in October, I was asked to brief the President of the United States on data mining initiatives at NASA. I didn't actually do that, for various logistical reasons, but that event convinced me that we needed to step up our game in data mining. So, I worked hard to convince my Raytheon bosses that the company needed to develop expertise and a corporate capability in information science and data mining (which we now call Data Science, but we didn't use that phrase in 2001). My efforts led to the creation of IST@R (the Institute for Science and Technology at Raytheon), and I became its first Co-Director for Space Science. I was able to obtain a few small NASA research grants to continue my data mining

research within IST@R, which carried over into 2003 when I moved from NASA to GMU.

Q - Very interesting! What were the successes?
A - We secured grants to discover unusual super-starbursting galaxies in large astronomy data sets. We had a grant to build a neural network model to identify wildfires in remote sensing satellite images of the Earth. My colleagues at UMBC and I collaborated on a grant to develop data mining algorithms on distributed data – the algorithms were designed to work on the data in their distributed locations – it was one of the first successful examples of "ship the code, not the data", which now everyone is trying to accomplish.

Q - What would you do differently now given all the new technologies and techniques that have been developed since then?
A - If I could do it again, I would have focused more on the Hadoop and MapReduce technologies, which are still not part of my own skill set. But, I have enjoyed developing and testing new algorithms for discovery and inference. So, I guess I won't give that up – there is such great pleasure in that discovery process.

Q - Makes sense :) Final question on the data mining front ... How has the work you've done in consulting, managing and developing technologies for data mining changed since you first started working in it?
A - The biggest change is that everyone now wants to do it. In those days, I could not convince most companies that I consulted with that they needed data mining technologies. I would get one or two consulting gigs per year, at most, and those businesses were almost entirely focused on data management, not on data science, or data mining, or discovery. Now,

a lot of people have forgotten the importance of data management, data modeling, metadata, data organization, clever data indexing schemes, data quality, etc. So, the pendulum needs to swing back to an equilibrium place where all of the data technologies and processes are in play. Consequently, I no longer need to convince people of the importance of data science – my phone and email are now flooded with dozens of requests for assistance from companies everywhere!

That's a nice problem to have! On that note, let's switch gears and talk about another area that keeps you busy - teaching Data Mining…

Q - What compelled you to start teaching Data Mining?
A - I always loved to teach – it was a natural gift for me. When I started discovering the power of data mining and experiencing the joy of knowledge discovery from big datasets, I was like a kid in a candy store. I wanted everyone to know about it. I was giving talks on data mining in many places. I gave such a talk in the Baltimore area in 2002, and one of the database program directors from the UMUC graduate school was there – he said that they were planning to start a new data mining course in early 2003, and he asked me to teach it. I jumped at the opportunity. I couldn't imagine anyone getting a job in the modern world where they didn't have data skills – it became my mission in life to teach data mining to anyone and everyone who would listen to me. I frequently said (then and now) that we need to teach data mining to school kids (starting in the elementary grades), and I still believe that. Of course, the examples and methods that are taught must be tuned to the appropriate grade level, but

the concepts of classification, clustering, association, and novelty discovery are all part of our human cognitive abilities from birth.

Q - That's a bold goal - how would you approach that? Approach high schools and then march down the age groups?
A - I am thinking a lot about that these days. So, the answer might be yes. The goal would be to establish professional development workshops for teachers, who might receive continuing education credits as they learn data science and create curricular materials related to it. Watch us and see what happens…

Q - Will do! Back to your current teaching for now though … As you now teach onsite and online how do they compare and contrast?
A - When I was at UMUC, I taught both online and face-to-face. I still do the same at GMU, so it is not really a change for me. However, the two learning environments are vastly different for me – I can interact more freely and tell my "war stories" from my NASA days more fluidly in the face-to-face class. Also, the one-on-one interactions in the online class are very time-consuming for me, compared to the one-on-many interactions in the face-to-face class, which I find to be much more manageable.

Q - Makes sense. So what about the experience at UMUC excited you enough to help you eventually become a full professor of Astrophysics and Computational Science?
A - I loved teaching the one graduate course in data mining at UMUC, but I really wanted to create a whole data science curriculum, including data visualization, databases, computational methods, data ethics, and more. That was one of my main motivations for going to GMU. I could never do that at UMUC, since my one course was part of a bigger database

technologies program. However, I was pleasantly surprised this past year to learn that UMUC now has a graduate degree in Big Data Analytics. I am a member of that program's advisory board – that is a very gratifying experience for me, the fact that they remembered me and asked me to join their board.

Q - What are your views on teaching Data Mining / Data Science / Machine Learning now? How have they changed since you first started teaching?
A - I think every student in every discipline needs at least one such course. I also believe that every student in a science or engineering discipline needs much more than one course. That hasn't changed. But what has changed for me is this: I used to think that undergraduates needed to major in Data Science, and so we created a BS degree program for that, but I am now more convinced that students should take Data Science electives or take a Minor in it, to accompany their own choice of Major. That's because we need a data-savvy workforce in all disciplines in our data-rich world.

That would certainly help general workforce data-literacy! Also, a good opportunity to talk a little about the GMU Data Science BS degree program...

Q - What was the impetus to starting a Data Science BS degree program?
A - I was convinced that data would govern the world, and would change everything in business, government, social, academia, etc. So, I was driven to start a program that taught students all of the skills that make up data science, to prepare them for the data-rich world that they would

be entering after college. I knew that the job opportunities would be large, but I never imagined that the number of jobs would be as huge as they have now become! So, the impetus was both my love of data science and my belief that it was absolutely essential. We started offering courses in 2007. At that same time, my GMU colleagues and I wrote a proposal to the NSF undergraduate education division to develop the program further – we called it "CUPIDS = Curriculum for an Undergraduate Program In Data Science." It was funded, and we were on our way.

Q - How did you approach it? What tools/methodologies does the program use?
A - We began as (and remain) a science degree program. So, the focus is on science problem-solving, applied math, statistics, machine learning and data mining algorithms, basic programming skills, simulation and modeling skills, and data visualization techniques. We are gradually moving toward a more general data science focus, but we are still keeping our students focused on the core problem-solving, modeling, and data mining skills of a data scientist.

Q - What has been the most surprising insight in creating the program?
A - The most surprising thing is that most students coming out of high school have never heard of data science – almost no students have received any guidance counseling about the importance of data and information in the modern world. Also, most students think of computational and data science as “information technology” – i.e., as word processing, or internet security, or system administration. They aren’t particularly interested in making a career out of that – neither am I. They don’t realize that it is all about discovery, discovery, discovery!

And they don’t realize that being a data scientist is the sexiest job of the 21st century. When they do finally realize these things, they then become our best ambassadors and evangelists to their fellow students – most of our recruitment comes from "word of mouth" from student’s own peers. The other big insight for us is that we thought that students could jump from the introductory courses into the advanced courses – it is now obvious that we needed intermediate-level bridge courses, which we subsequently developed – the most successful of these courses has been our "Computing for Scientists" course, which is packed with students every semester.

Q - That's great to hear! How would you describe the goals of the program??
A - I list the goals very simply in this way: Students are trained

- to access large distributed data repositories,
- to conduct meaningful inquiries into the data,
- to mine, visualize, and analyze the data, and
- to make objective data-driven inferences, discoveries, and decisions.

The single overarching goal since the beginning of our CUPIDS program has been to increase student's understanding of the role that data plays across the sciences as well as to increase the student's ability to use the technologies associated with data acquisition, mining, analysis, and visualization.

That's a great goal - good luck with the ongoing journey! Finally, as one of the most often cited "Big Data Influencers", would love to get your thoughts on the future of Data Science...

Q - What excites you most about recent developments and the future of Big Data / Data Science?
A - The opportunity to work with many different businesses and disciplines is truly the most exciting aspect of the work. Data scientists can work in many areas – for example, I work with people in astrophysics, aerospace engineering, transportation safety, banking, finance, retail, medical research, text analytics, climate modeling, remote sensing, and more. I now call myself a trans-disciplinary data scientist because my work in data science transcends discipline boundaries – I do not need to become an expert in those fields (i.e., not multidisciplinary) in order to work with people with a different domain expertise than mine. I see a very bright future as more and more organizations get on board the Big Data / Data Science train – there are many new technologies, algorithms, problems to solve, and things to do. It is almost overwhelming, but it is definitely exhilarating. My favorite catch phrase these days is "Never a dull moment!" That sums it all up.

Q - Couldn't agree more! Last one! ... Any words of wisdom for Big Data / Data Science students or practitioners starting out?
A - Start early and often in doing, in learning, and in absorbing data science. It takes some serious education and training, but it is worth it. Be true to yourself – know your aptitudes, your skills, your interests – don't force something that isn't there. There is a place for everyone in this data-rich world. Don't underestimate the power of curiosity, communication,

and collaboration skills. They will take you further than just about anything else in life. Above all else, be enthusiastic and passionate about it. If you can see the power, discovery potential, and wonder of big data, then the passion and enthusiasm will follow. The future is very bright for those who are able to derive insights from big data, in any discipline or any job. Find the right opportunity and pursue it. There will never be a dull moment after that.

Kirk - Thank you ever so much for your time! Really enjoyed learning more about your background, your ground-breaking work in data mining and how it was applied at NASA, as well as your perspectives on teaching data science and how you are contributing to the education of future generations. Kirk can be found online at http://kirkborne.net or on twitter @KirkDBorne

Optimal Retail Store Location

Dmytro Karamshuk

Researcher in Computer Science
& Engineering at
King's College London

Social Media & Machine Learning tell Stores where to locate

We recently caught up with Dmytro Karamshuk, Researcher in Computer Science and Engineering at King's College London - investigating data mining, complex networks, human mobility and mobile networks. We were keen to learn more about his background, how human mobility modeling has evolved and what his research has uncovered in terms of applying machine learning to social media data to determine optimal retail store placement...

Hi Dmytro, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - I'm a computer scientist with startup experience working on understanding user behavior in social media and mobile networks to make the world slightly better... Currently at King's College London; previously at University of Cambridge; Italian national research council; Institute of Market and Technologies in Lucca; and as a managing partner at a software engineering startup (http://stanfy.com).

Q - How did you get interested in Machine Learning?
A - It was a commercial project we did with my classmates (co-founders of stanfy.com) during our university years. We were creating a web 2.0 system for our client, Banks.com Inc., when the idea of recommending interesting stuff to users emerged. Although we didn't know much about recsys at that time, we implemented something very similar to item-to-item collaborative filtering. It was the year 2005. Everyone was happy with the outcome.

Q - That's great! ... So, what was the first data set you remember working with? What did you do with it?
A - The first dataset was at the age of 15 when I coded a web chat on Perl. The data was stored in a MySql database and a bunch of my friends would use it for fun. Didn't do much with it to be honest, but it was interesting to play with the logs and run some simple stats.

Q - Was there a specific "aha" moment when you realized the power of data?
A - When there is a lack of data, researchers would usually start making assumptions and build synthetic models. An “aha” moment is when at last you get your hands on the real data and see how all your assumptions crash one by one. Had it during my PhD :)

I guess that's good and bad! Thanks for sharing your background - let's switch gears and talk in more detail about your research field - Machine Learning Applied to Human Mobility & Urban Dynamics...

Q - How has human mobility modeling been approached in the past?
A - Interest in urban mobility was raised many decades (or even centuries) before we (very recently) got the first large scale trace of human movements. With the lack of data researchers had to rely on either probabilistic synthetic models (the most prominent ones we have discussed in our review) or coarse-grained statistics usually collected with user surveys. There would also be more creative ways of gathering data such the one dollar bill experiment one dollar bill experiment where users were collectively reporting locations of one dollar bills across the US and used that as a proxy of human movements. The scale of the dataset was phenomenal for that time.

Another groundbreaking experiment, called Reality Mining, has been conducted by guys from MIT when a number of volunteers among MIT students agreed to carry a mobile device in their pockets with a piece of

software which would record all bluetooth communications with other similar devices. This was probably the first dataset on human contact obtained in an automated way.

Frankly, and related, a large scale data set of human movements had already existed for a few decades: mobile phone operators have been collecting logs of mobile users' locations (i.e., base stations from which they access the network) from the earliest days of mobile phones. However, this information was under lock for a long time given operators were afraid of leaking commercially sensitive information. The first one to break this taboo was a group of physicists from Northeastern University in Boston who published a large scale study from mobile phone data in their prominent 2008 Nature paper.

More recently, with the emergence of location-based social networks (such as Foursquare, Gowalla or Altergeo) where users voluntarily share their very-abouts with the world (via Twitter for example) we have finally got public access to a massive trace of human mobility.

Q - Really interesting context ... What then excites you most about bringing Machine Learning and Human Mobility & Urban Dynamics together? How is this approach different from traditional mobility models?

A - Machine-learning is useful in two ways in this context: as a tool to build something very practical, very useful, such as recommender systems; and as an explorational tool for theoretical research to understand complex dependencies and correlations between various variables in a given physical system.

Q - What are the biggest areas of opportunity / questions you want to tackle?
A - One interesting challenge lies in disambiguating information across various sources of data. There are dozens of different signals which we employ in urban computing: from social networks, census data to signals collected from sensors installed in personal devices, cars and embedded in the streets (CCTV cameras for example). So far all these data sources have been mostly considered separately because it is very difficult if not impossible to link a user of, say, an Oyster Card [used on the London subway / tube system] with a Twitter user, or his account in a fitness app or sensor installed in his car. But if we could draw these links at least on some coarse-grain level, among social groups or users with similar demographic profiles for example, it could skyrocket our understanding of user behavior in the cyber-physical world and open up huge space for exploration in various aspects of urban studies. I believe we will see a number of statistical methods as well as technological solutions emerging in this field in the near future.

Q - That would be very exciting! ... Couple of more practical questions. First, what Machine Learning methods have you found most helpful?
A - I play with various supervised-learning models. I find them more suitable for "exploratory" research where one wants to test some hypothesis or check some dependencies in his data. As long as a problem can be formalized in a supervised learning task, the results can be directly validated over the data which makes it more convenient than unsupervised learning where manual validation is required.

Q - And what are your favorite tools / applications to work with?
A - I have recently switched to Python (after ten years with Java) and found it very suitable for data analysis. I would name scipy, scikit-learn and graphlab as my new favorites. In Java my data analysis bundle would consist of Weka, Ranklib, Lenskit, Gephi, apache common math and other statistical libs. When I have inspiration I play with Processing to draw some fancy visualizations.

That is fancy! :) On that note, let's talk more about your recent work on Mining Online Location-based Services for Optimal Retail Store Placement, which has caught retailers attention - it is a great example of how your research can be applied ...

Q - Could you tell us a little more about this work?
A - In this work we solve an old good problem of finding the best place to open a business in the city but with a new data coming from location-based social networks. Here we wear the hat of a retail chain manager (say, a Starbucks manager) to solve the question: given a number of locations in the city can we predict the one where a new restaurant will thrive? See, big retailers spend a fortune for sophisticated user surveys and expensive market analysis, so, we wondered how user-generated data which was already available online could be used to solve the very same problem in a cost-effective way.

Q - That sounds great! How did you approach it?
A - We collected a dataset of public tweets where users were reporting their location with the Foursquare app. A basic record in this dataset would say a venue V in New York (e.g., a restaurant, railway station or a

bank) where a user U was at time T. We used this data to measure how popular each venue is and to build various indicators of popularity. This would include things like the number of competitors in the area, presence of transportation hubs or other place-attractors nearby, intensity of human mobility in the area etc.

Q - What were the major steps?
A - Initially we tried to solve this problem for a general business. It took us a month of try-and-fail trials to understand that the popularity indicators may vary significantly across different types of venues: placing a chinese restaurant in China Town might be a good idea but not as good for an italian restaurant. We then decided to focus our efforts on 3 chains McDonalds, Starbucks and Dunkin' Donuts.

The next important insight came from the work of a physicist, Pablo Jensen, who a decade ago proposed to use a complex network toolkit to analyze geographic interactions between various business activities. We used some of his ideas to build a popularity indicator that assesses the importance of co-locating a restaurant with other business activities.

Q - So what unlock did the social media data provide?
A - The most crucial difference between the social networking data we had and traditional data sources is the fact that we have fine-grained data of individual users movements. So, we can ask questions like what is a place A which users who have just visited B would also usually visit? Like a coffee shop on one's way from a train station to his office. We learned these patterns from the dataset and used them as a yet another source to predict popularity.

Q - How did Machine Learning help?
A - Once we devised these various indicators of human movements and co-location of places we used machine learning to build a popularity prediction algorithm. Our goal was to use all various indicators and the popularity data to train a supervised model which would rank a given list of locations according to their prospective success. We tried various different models and found that a regression-to-rank model based on support vector regressions performed the best in this case.

Q - What answers / insights did you uncover?
A - Our main finding was that the model which combined user mobility information built out of users' checkins and data about geographic location of users was performing better than geographic data alone. In other words, we can indeed achieve incredibly valuable insights on user retail preferences from the social media data. You can see more details on the exact results of the different models we tried in this recent presentation.

Q - That's really interesting! What are the next steps / where else could this be applied?
A - Not only is this very relevant for retailers, we believe our work can inspire research in various other areas where urban mobility is important. For example, the same way of thinking can be applied to study land prices, the optimal location for opening an office, understanding public use of common spaces, etc.

That would be great! Finally, let's talk a little about the future...

Q - What does the future of Machine Learning look like?
A - As a machine learning practitioner I would be really happy to see a basic machine learning course a in high school program. We have already realized that coding is the must know tool for a 21st century person and I believe data science is next in line.

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - Solve the problems that are worth solving. A quote from a wise-man which I strongly second. One would rather start as an amateur in something that is really important with a potentially groundbreaking impact rather than an expert in something that no one cares about.

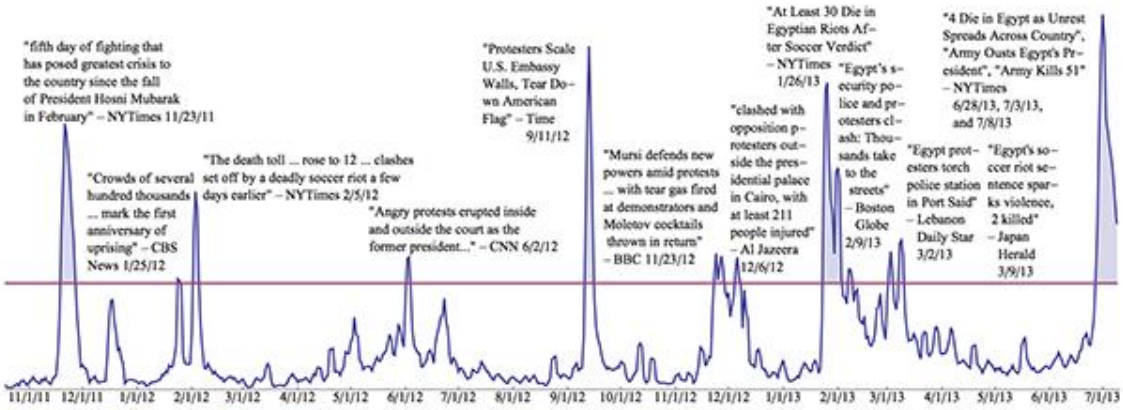
Dmytro - Thank you ever so much for your time! Really enjoyed learning more about your background, how human mobility modeling has evolved and what your research has uncovered in terms of applying machine learning to social media data to determine optimal retail store placement. Dmytro can be found online here and on twitter @karamshuk.

Big Public Data to Predict Crowd Behavior

Nathan Kallus

PhD Candidate at the
Operations Research Center at MIT

Big Public Data to Predict Crowd Behavior



We recently caught up with Nathan Kallus, PhD Candidate at the Operations Research Center at MIT. We were keen to learn more about his background, his research into data-driven decision making and the recent work he's done using big public data to predict crowd behavior - especially as relates to social unrest...

Hi Nathan, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - I grew up in Israel and went to college at UC Berkeley where I first discovered my passion for statistics and optimization. Today I am a PhD Candidate at the Operations Research Center at MIT and my research revolves around the combination of statistics/data sci with mathematical optimization. I am really interested in the theory and practice of data-driven decision-making and in general the analytical capacities and challenges of unstructured and large-scale data.

Q - How did you get interested in working with data?
A - Most things in life cannot be known with certainty. In my own life, this is one of the reasons why I always like to keep an open mind toward new things and never judge others. Statistics and related data science is the most important tool to understand things that are not absolute, as most things are. It allows us, first, to describe uncertainty in the real world and then, second, to investigate it using real data. This, to me, is very stimulating and makes working with data quite exciting. Optimization is the mathematics of making the best decision possible but what that decision is depends on the settings. An optimal decision in unrealistic or misspecified settings may end up being a very bad one in practice so it is critical to recognize uncertainty when optimizing, including uncertainty in one's model.

The field of operations research has, historically, been primarily model-driven -- necessarily so due to a past dearth of data. Many quantitative methods for decision making were based on modeling and on distributional assumptions with little to no deference to data. At most it was estimate, then optimize. Nonetheless, the field has transformed whole industries: airlines, advertising, retail, finance, and more. The explosion in the availability and accessibility of data is ushering forth a shift toward a data-driven paradigm for decision making. New theory, methods, and applications are necessary to realize this and the combination of statistics and data science with optimization provides the right toolset. I find it critical and fascinating to work on the methodological advances in data-driven decision making that must come

hand-in-hand with the technological advances of the era of information, and on practical applications that combine these effectively.

Q - Its definitely an exciting time to be in this field! ... So, what was the first data set you remember working with? What did you do with it?
A - Perhaps my first endeavor into predictive analytics was a final project in a course I took in undergrad. I was wondering if I could uncover certain innate personal characteristics of users such as sex (which is at least for the most part innate) based on petty aspects of their Facebook profile such as group memberships (this was before Likes etc). I wrote up a dummy Facebook app and got some friends to install it so I can scrape their friend network. Soon enough I had a few thousand profiles. The prediction was reasonably accurate, if I recall.

Q - Was there a specific "aha" moment when you realized the power of data?
A - I don't know about a first "aha," but the biggest "aha" was definitely when I saw how accurately I could predict real-world events like mass protests and violent social unrest accurately using data from the "virtual" world of social media. There the scale of the data was really critical for success.

We'll dive into that in more detail shortly! First, let's talk more broadly about your field of research - data-driven decision making...

Q - What have been some of the main advances in recent years?
A - The most important advances have been technological and resulting

in an increase in the availability and accessibility of useful data. Enterprise resource planning (ERP), supply chain management (SCM), and customer relations management (CRM) software platforms are becoming more ubiquitous and collecting more raw data by default as they simply operate. A lot of things that used to be offline like the news, government data, etc are now online and machine-readible and therefore can be used for analysis, often with the help of natural language processing. New modes of communication such as Facebook and Twitter have taken hold online and data from these provide a digital window into sentiments, consumer behavior, the behavior of crowds, and more. In 2012 about 2.5 exabytes of data were created each day and this number has increased by some 25% each year since (fun fact: Walmart consumer transactions alone make up approximately 0.01% of daily data collection/creation).

This explosion of data is changing how we think about a lot of things but most importantly it needs to change the way we make decisions or its collection is for naught. The lack of real impact on decisions was (or still is) one of the biggest criticisms of the "Big Data" buzzword frenzy. I was reading the Technology Review's Business Report recently and liked the tagline they had: "What's the point of all that data, anyway? It's to make decisions" (Jan 2014).

Q - That's a great quote! ... So what are the biggest areas of opportunity / questions you would like to tackle?
A - How to use unstructured (e.g. text, video, health records) and irregular (e.g. non-IID) data -- properties that characterize (or at least

should characterize) the kind of data referred to as "Big" -- for decision-making in a theoretically principled manner with practical impact.

Q - A great goal! :) Now, two of your more theoretical papers recently won awards in the Operations Research community (congratulations!) - can you share a high level summary of one or both? (e.g., what problem you were tackling, how you approached it, what you found etc.)
A - In both papers we were addressing the fundamental question of how to go from data to decisions and in both the idea of combining statistical tools (both theoretical and practical) with optimization tools (same) was the key. On both I worked with Dimitris Bertsimas and Vishal Gupta.

In one, we developed a new framework for data-driven optimization under uncertainty that is notable for combining generality, tractability, convergence, and finite-sample performance guarantees. The key was in making theoretical connections between statistical properties of hypothesis tests and optimization properties of decision problems. This yielded a new theoretical framework that unified some existing work and resulted in new tools that are well suited for practical use.

In the other, we worked with an already widely popular method for optimization under uncertainty, robust optimization (RO), and showed how it can be made data-driven and tailored for data-rich environments. The key ingredient in RO is an uncertainty set. These are traditionally designed ad-hoc and in a modelling-driven manner. The paper proposes several procedures for designing these based directly on data and shows that data-driven variants of existing RO approaches consistently outperform their data-poor analogues.

Q - Thanks for sharing - and congrats again on the recognition! ... Let's switch gears and talks about your recent work on Predicting Crowd Behavior with Big Public Data, which is a great example of how some of your research can be applied - and has been featured in the news ... could you tell us a little more about this work? First, what specific problem were you trying to solve?

A - I had become quite interested in the predictive power of social media. Never before has this amount of communication been publicly available and so accessible. When it comes to predicting the actions of people, it makes a lot of sense –– it's crowds talking about themselves. For example, the manifestation of mass demonstrations often involves collective reinforcement of shared ideas and public calls to action to gather at a specific place and time. Both of these now take place to some extent online, perhaps providing a sufficiently wide window into the real world and its future trajectory. So I set out to both verify and quantify this idea that social media data, and other online data, can predict significant future events such as mass protests, violent social unrest, and cyber hacktivism campaigns.

Q - Makes sense. How did you approach it?

A - I teamed up with a company called Recorded Future. They collect a ton of data from various open-content online sources like news, government publications, blogs, social media, etc. This is something that's sometimes called web intelligence or open source intelligence. Importantly, this data included lots of Twitter activity of the sort I talked about above.

I first looked for signs of potential predictive signals in the data. A

descriptive analysis of the data suggested a potential signal in the Tweets that were posted before this day but seemingly talked about a protest to occur on that day (for example, social-media calls to arms are often like this: posted before the day in question, about the day in question). The key here was to analyze the unstructured text for reported/discussed events in it and the time frame for the event's purported occurrence. Other signals emerged from analyzing, for example, the news. In tandem, these were helpful in sifting out the social media trends that fizzled out before materializing on the ground.

Then it was a matter of employing machine learning to train a predictive classification model to uncover the right predictive patterns in the set of signals I thus extracted. Transforming and normalizing the data in the right way was also critical to making it work. There was also simply a ton of data to handle so of course there was also an ample amount of engineering necessary in implementing this to be able to handle this -- that, and a rather big computer.

Q - I can imagine! What answers / insights did you uncover?
A - While the issue of whether mobilization, and how much of it, occurs online is highly controversial, it did become quite clear to me that the permeation of connective technologies was sufficient so that the dialogues on platforms like Twitter provided a sufficiently clear window into these processes to allow pretty accurate prediction. I really could verify that social media data had the power to predict some futures and that it does so quite accurately. In the paper I looked at historical data in order to validate this, but right now the system is running live in real-time and it's exhilarating to follow it and see it actually predicting these events before

they occur and before I see them reported in the news. We have documented many cases including in Pakistan, Bahrain, and Egypt where we saw a clear prediction and then two to three days later headlines everywhere.

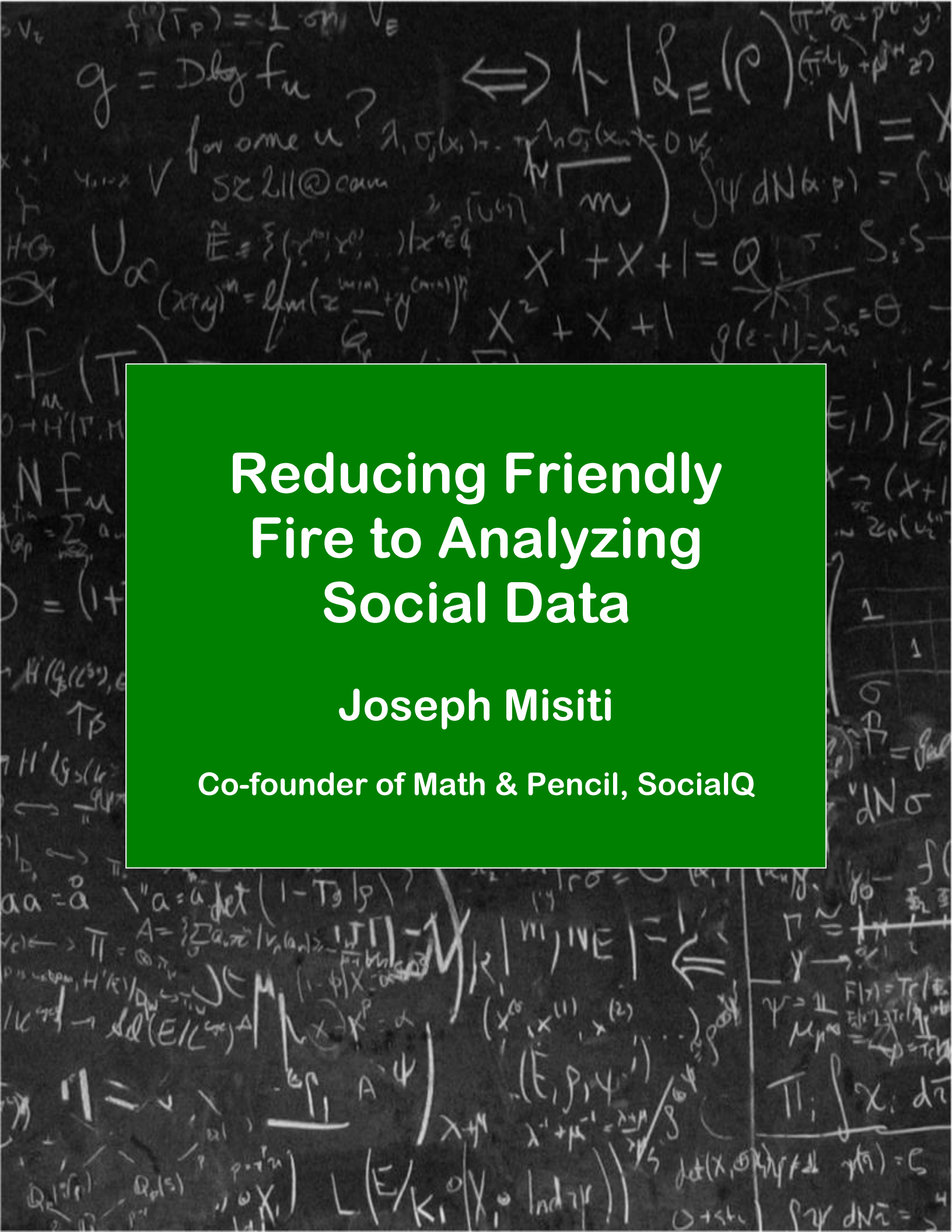
Q - What are further potential applications of this approach? Which get you most excited?
A - The ability to forecast these things has clear benefits. Countries and authorities faced with a high likelihood of significant protest can prepare themselves and their citizens to avoid any unnecessary violence and damages (sadly a lot of unrest in certain regions ends up with clashes and deaths). Companies with personnel and supply chain operations in an affected region can ask their employees to stay at home and to remain apolitical and can attempt to safeguard their facilities in advance. Countries, companies, and organizations faced with possible cyber campaigns against them can beef up their cyber security in anticipation of attacks or even preemptively address the cause for the anger directed at them.

Besides providing predictions, the system can be useful to pundits and decision makers because it allows one to follow in particular those trends in social media that will actually lead to a major event. Then one can better understand what is going on now, where it will lead, as well as why. Let me give an example. The system had predicted in advance the unrest in Egypt that sadly led to 5 deaths on March 28. If you looked at the tweets that supported that prediction, you could see calls to demonstrate from both the pro-Brotherhood and pro-El-Sisi sides, raising fears of street clashes, and that the group Tamarod was behind the latter calls.

These were actually two facts that were at a later time specifically mentioned in an AP release.

More broadly I am very excited about the predictive power of such web intelligence in more business operations, as it has for example been shown to predict product demand and other consumer behavior. With this sort of prediction, the decision theory actually becomes more complicated and is the subject of a current research project of mine.

Nathan - Thank you ever so much for your time! Really enjoyed learning more about your background, your research areas and the work you've done using big public data to predict crowd behavior - especially as relates to social unrest. Good luck with your ongoing research!



Reducing Friendly Fire to Analyzing Social Data

Joseph Misiti

Co-founder of Math & Pencil, SocialQ

From Reducing Friendly Fire to Analyzing Social Data

We recently caught up with Joseph Misiti, co-founder of Math & Pencil, SocialQ and more! We were keen to learn more about his background, his work at SocialQ, and thoughts on how Data Science is evolving. Also, given his thought-provoking article "Why becoming a data scientist is NOT actually easier than you think", we were keen to garner his advice on how best to enter the field...

Hi Joseph, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - I'm 31 years old and live in New York City. I hold a BS in Electrical Engineering which focused on signal processing and numerical analysis, and an MS in Applied Math with a focus on computer vision, data mining and wavelet analysis. I started out in DOD, building SATCOM radios at Harris Corporation, moved on to missile defense algorithms at Lockheed Martin, and capped my work in that sector with building a lie detector

(Thin Slice Detectorif you have read the book Blink) using computer vision and wavelet analysis. I moved to New York City three years ago and started a consultancy called Math & Pencil which is behind start ups including:

- SocialQ
- Fetcher
- ReWire
- Employii

Q - How did you get interested in working with data?
A - I have always loved math and computer science, so analyzing data was a natural next step. I suppose I really got excited after studying numerical solutions to partial differential equations while an undergrad, because that was the first time I really saw the power of computer modeling/applied mathematics.

Q - So, what was the first data set you remember working with? What did you do with it?
A - The first data set I can recall playing with (I think) was the Hair-Eye Color data set. I was using it in an introduction to statistics course in undergrad to learn about linear regressions, coefficients, p-values, etc

Q - Was there a specific "aha" moment when you realized the power of data?
A - When I was working for Lockheed Martin, we used Kalman filters to model errors in missile measurements. The actual algorithms are fairly simple from a math perspective, but when applied to missile data, were actually really good at predicting errors in the measurements. Removing

these errors would theoretically reduce friendly fire, so I would say this was the first time I saw a simple algorithm applied to a real-life data set that could literally save lives. It was pretty amazing at the time.

Q - Wow, that's very powerful! ... On that note, what excites you most about recent developments in Data Science?
A - The most exciting thing about data science in my opinion is the open source movement combined with Amazon EC2 prices dropping. For the first time, I do not need to have access to or purchase a cluster of computers to run an experiment. With a lot of recent developments in deep learning being run on GPUs rather than CPUs, I can very easily rent a GPU instance on EC2, install the software, and use an open source library written in Python like pylearn2 to test out an hypothesis.

The open source movement in general is really amazing. I would say mostly because of the rise in popularity of Github, it's very easy to contribute to projects now. For instance, I created an open source project last month calledawesome-machine-learning which is basically a list of all the machine learning resources on the web. Within a few weeks, over 2.9K people had starred it and I have had 48 contributors help me out. If you step back and think about it, this is really amazing (and most of us just take it for granted).

It is amazing - and that is a terrific resource you've put together - thanks! Let's switch gears and talk about your current work at SocialQ...

Q - What attracted you to the intersection of Data Science and Social Media? Where can Data Science create most value?
A - I was originally attracted to the idea of building a company with the potential of using my machine learning skill set, but also realized the company would have to have some level of success to get there. SocialQ started as a SaaS based tool to help marketing researchers dig into their social data via a few dashboards. After a few years, we have built up a rather larger data set and we are now able to offer statistical tools. Also, we work directly with customers/marketing researchers to figure out what type of questions they want answered, and then come up with statistical solutions to these problems. It has been a really interesting learning experience.

Q - So what specifically led you to found SocialQ? What problem does SocialQ solve?
A - One of the problems SocialQ solves is what marketing researchers can do with their social data, when they don't necessarily have comprehensive tools or the math background to make sense of what they've collected. We have created a platform that not only helps them answer those questions, but also makes the collection of the data easier. It is bundled into a set of SasS based tools so the researcher can initiate a study and then login in the next day and see the results.

Q - What are the biggest areas of opportunity/questions you want to tackle?
A - I am interested in helping companies improve their brands on social media using mathematics. There are a lot of different ways to do that, but that is the problem area I am trying to tackle currently.

Q - What learnings/skills from your time building data models for the US intelligence community are most applicable in your current role?
A - The skill set I was using previously is still being applied today in my day-to-day, the only difference is the features I'm extracting. Computer vision features are very specific to the field, (SIFT, wavelets, etc). My new job requires more NLP techniques, but a lot of the algorithms I am using are the same (SVMs,PCA, kmeans, etc). I have more freedom now, because the code I am using is not owned by the US government and its contractors. I can download a package of Github and start playing around without having to go through six different people for approval.

Q - Makes sense! So what projects are you currently working on, and why/how are they interesting to you?
A - Recently I have accepted roles as a consultant for a few interesting companies. One was using NLP to built classifiers/recommendations around RSS feed data set for a start-up, another is a computer vision problem involving human hand writing. In my spare time, I have been studying a lot of Bayesian Inference and reading papers and listening to lectures on Deep Learning. I find almost all aspects of statistics/math interesting so any chance I get where someone will compensate me to solve such a problem, pending I have the time and interest in the query, I'm in.

Q - And how is data science helping? What techniques, models, software etc are you using?
A - For NLP I have been using a lot of Latent Dirichlet allocation to build features. For computer vision, it has been a lot of OpenCV and almost all classifiers are trained in Scikit-Learn. All data analysis, pre-processing,

and exploratory statistical analysis is using iPython notebooks, Pandas,Statsmodels, and either Matplotlib or ggplot2. Basically the same toolset everyone else is using that has moved on from R.

Q - How do you/your group work with the rest of the team?
A - My technical team consists of two other engineers and one designer. The designer is Tim Hungerford, and the two engineers are Andrew Misiti and Scott Stafford. I have worked with them so long at this point that everything just seems to work, but interestingly enough, all of us work remotely, with Andrew and Tim based in Buffalo, Scott based in DC, and myself based in Manhattan. We do all our work using a combination of Github issues, campfire, Gmail/hangouts, and sometimes (although rarely these days) Asana.

Thanks for sharing all that detail - very interesting! Finally, let's talk a bit about the future and share some advice ... ...

Q - What does the future of Data Science look like?
A - It's exciting. I think the open source movement around data science will continue to be the driving force. I am excited about new languages like Julia and frameworks like Scikit-learn/Pandas/iPython. I think more and more, we are going to see researchers moving from R/MATLAB to Python. I am not an MATLAB/R hater at all, but I do think Python is much easier to work with, document, and read. In the end, the reason for not using R/MATLAB anymore is simply because you cannot integrate them into a web service - unless it's some simple exploratory analysis, you're going to need to eventually convert them into another language

anyways, so why not just avoid that step (also MATLAB isn't always cost effective).

I think because the demand for data science is increasing, more smart people are going to go into it. I also think long-term, it will solve a lot of larger problems like bringing down heath care costs, reducing government waste, detecting government/private company fraud etc.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Never stop learning. Take as many math and statistics courses as you can, and implement all the basic algorithms on your own before you use open source versions. I am only advocating "reinventing the wheel" because I truly believe the only way to understand something is to build it from scratch, even though it is extremely time consuming.

Also, write blog posts and open source code, because that is what potential employers in the future are going to want to see.

Finally, do not be afraid to learn this stuff on your own rather than going back to school for it. If you already have an MS, I would save the money and invest in yourself, the job is more about getting meaningful results, not the school or degree you have, especially to me as an employer. You can see this by looking at the wide range of backgrounds employed data scientists have.

Joseph - Thank you ever so much for your time! Really enjoyed learning

more about your background, your work at SocialQ, and your thoughts on how Data Science is evolving. Good luck with all your ongoing projects!

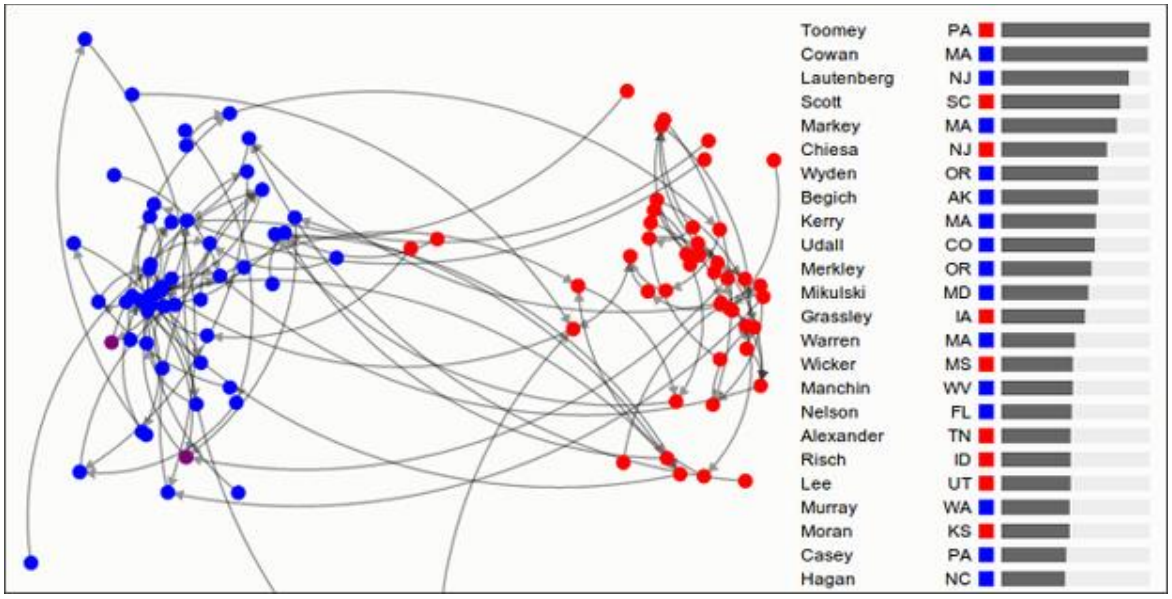
Readers, thanks for joining us! If you want to read more from Joseph he can be found online here and on twitter @josephmisiti.

Data Science at the Command Line

Jeroen Janssens

Author of Data Science
at the Command Line

Anomalies, Concerts & Data Science at the Command Line



We recently caught up with Jeroen Janssens, author of Data Science at the Command Line. We were keen to learn more about his background, his recent work at YPlan and his work creating both the book and the (related) Data Science Toolbox project...

Hi Jeroen, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - Howdy! My name is Jeroen and I'm a data scientist. At least I like to think that I am. As a Brooklynite who tries to turn dirty data into pretty plots and meaningful models using a MacBook, I do believe I match at least one of the many definitions of data scientist. Jokes aside, the first time I was given the title of data scientist was in January 2012, when I joined Visual Revenue in New York City. At the time, I was still finishing my Ph.D. in Machine Learning at Tilburg University in the Netherlands.

In March 2013, Visual Revenue got acquired by Outbrain, where I stayed for eight months. The third and final startup in New York City where I was allowed to call myself data scientist was YPlan. And now, after a year of developing a recommendation engine for last-minute concerts, sporting events, and wine tastings, I'm excited to tell you that I'll soon be moving back to the Netherlands.

Q - How did you get interested in working with data?
A - During my undergraduate at University College Maastricht, which is a liberal arts college in the Netherlands, I took a course in Machine Learning. The idea of teaching computers by feeding it data fascinated me. Once I graduated, I wanted to learn more about this excited field, so I continued with an M.Sc. in Artificial Intelligence at Maastricht University, which has a strong focus on Machine Learning.

Q - So, what was the first data set you remember working with? What did you do with it?
A - The very first data set was actually one that I created myself, albeit in quite a naughty way. In high school--I must have been fifteen--I managed to create a program in Visual Basic that imitated the lab computers' login screen. When a student tried to log in, an error message would pop up and the username and password would be saved to a file. So, by the end of the day, I had a "data set" of dozens of username/password combinations. Don't worry, I didn't use that data at all; this whole thing was really about the challenge of fooling fellow students. Of course I couldn't keep my mouth shut about this feat, which quickly led to the punishment I deserved: vacuum cleaning all the classrooms for a month. Yes, I'll never forget that data set.

Q - I can imagine! Maybe it was that moment, though was there a specific "aha" moment when you realized the power of data?
A - Towards the end of my Ph.D., which focused on anomaly detection, I was looking into meta learning for one-class classifiers. In other words, I wanted to know whether it was possible to predict which one-class classifier would perform best on a new, unseen data set. Besides that, I also wanted to know which characteristics of that data set would be most important.

To achieve this, I constructed a so-called meta data set, where its 36 features were characteristics of 255 "regular" data sets (for example, number of data points, dimensionality). I evaluated 19 different one-class classifiers on those 255 data sets. The challenge was then to train a meta classifier on that meta data set, with 19 AUC performance values as the labels.

Long story short, because I tried to do too many things at once, I ended up with way too much data to examine. For weeks, I was getting lost in my own data. Eventually I managed to succeed. The lesson I learned was that there's also a thing as too much data; not in the sense of space, but in density, if that makes sense. And more importantly, I also learned to think harder before simply starting a huge computational experiment!

Makes sense! Thanks for sharing all that background. Let's switch gears and talk about this past year, where you've been the Senior Data Scientist at YPlan...

Q - Firstly, what is YPlan? How would you describe it to someone not familiar with it?
A - Here's the pitch I've been using for the past year. YPlan is for people who want to go out either tonight or tomorrow, but don't yet know what to do. It's an app for your iPhone or Android phone that shows you a curated list of last-minute events: anything ranging from Broadway shows to bottomless brunches in Brooklyn. If you see something you like you can book it in two taps. You don't no need to go to a different website, fill out a form, and print out the tickets. Instead, you just show your phone at the door and have a great time!

Q - That's great! What do you find most exciting about working at the intersection of Data Science and entertainment?
A - YPlan is essentially a market place between people and events. It's interesting to tinker with our data because a lot of it comes from people (which events do they look at and which one do they eventually book?). Plus, it's motivating trying to solve a (luxury) problem you have yourself, and then to get feedback from your customers. Another reason why YPlan was so great to work at, was that everybody has the same goal: making sure that our customers would find the perfect event and have a great time. You can improve on your recommendation system as much as you want (which I tried to do), but without great content and great customer support, you won't achieve this goal. I guess what I'm trying to say is that the best thing about YPlan were my colleagues, and that's what made it exciting.

Q - So what have you been working on this year? What has been the most surprising insight you've found?

A - At YPlan I've mostly been working on a content-based recommendation system, where the goal is essentially to predict the probability a customer would book a certain event. The reason the recommendation system is a content-based one rather than a collaborative one, is that our events have a very short shelf life, which is very different from say, the movies available on Netflix.

We've also created a backtesting system, which allows us to quickly evaluate the performance of the recommendation system to historical data whenever we make a change. Of course, such an evaluation does not give a definitive answer, so we always A/B test a new version with the current version. Still, being able to quickly make changes and evaluate has proved to be very useful.

The most surprising insight is, I think, how wrong our instincts and assumptions can be. A recommendation system, or any machine learning algorithm in production for that matter, is not just the math you would find in textbooks. As soon as you apply it to the real world, a lot of (hidden) assumptions will be made. For example, the initial feature weighting I came up with, has recently been greatly improved using an Evolutionary Algorithm on top of the backtesting system.

Thanks for sharing all that detail - very interesting! Let's switch gears and talk about the book you've been working on that came out recently...

Q - You just finished writing a book titled Data Science at the Command Line. What does the book cover?

A - Well, the main goal of the book is to teach why, how, and when the command line could be employed for data science. The book starts with explaining what the command line is and why it's such a powerful approach for working with data. At the end of the first chapter, we demonstrate the flexibility of the command line through an amusing example where we use The New York Times' API to infer when New York Fashion Week is happening. Then, after an introduction to the most important Unix concepts and tools, we demonstrate how to obtain data from sources such as relational databases, APIs, and Excel. Obtaining data is actually the first step of the OSEMN model, which is a very practical definition of data science by Hilary Mason and Chris Wiggins that forms the backbone of the book. The steps scrubbing, exploring, and modeling data are also covered in separate chapters. For the final step, interpreting data, a computer is of little use, let alone the command line. Besides those step chapters we also cover more general topics such as parallelizing pipelines and managing data workflows.

Q - Who is the book best suited for?

A - I'd say everybody who has an affinity with data! The command line can be intimidating at first, it was for me at least, so I made sure the book makes very little assumptions. I created a virtual machine that contains all the necessary software and data, so it doesn't matter whether readers are on Windows, OS X, or Linux. Some programming experience helps, because in Chapter 4 we look at how to create reusable command-line tools from existing Python and R code.

Q - What can readers hope to learn?
A - The goal of the book is make the reader a more efficient and productive data scientist. It may surprise people that quite a few data science tasks, especially those related to obtaining and scrubbing, can be done much quicker on the command line than in a programming language. Of course, the command line has its limits, which means that you'd need to resort to a different approach. I don't use the command line for everything myself. It all depends on the task at hand whether I use the command line, IPython notebook, R, Go, D3 & CoffeeScript, or simply pen & paper. Knowing when to use which approach is important, and I'm convinced that there's a place for the command line.

One advantage of the command line is that it can easily be integrated with your existing data science workflow. On the one hand, you can often employ the command line from your own environment. IPython and R, for instance, allow you to run command-line tools and capture their output. On the other hand, you can turn your existing code into a reusable command-line tool. I'm convinced that being able to build up your own set of tools can make you a more efficient and productive data scientist.

Q - What has been your favorite part of writing the book?
A - Because the book discusses more than 80 command-line tools, many of which have very particular installation instructions, it would take the reader the better part of the day to get all set up. To prevent that, I wanted to create a virtual machine that would contain all the tools and data pre-installed, much like Matthew Russell had done for his book Mining the Social Web. I figured that many authors would want to do something like that for their readers. The same holds for teachers and workshop

instructors. They want their students up and running as quickly as possible. So, while I was writing my book, I started a project called the Data Science Toolbox, which was, and continues to be, a very interesting and educational experience.

Q - Got it! Let's talk more about the Data Science Toolbox. What is your objective for this project?
A - On the one hand the goal of the Data Science Toolbox is to enable everybody to get started doing data science quickly. The base version contains both R and the Python scientific stack, currently the two most popular environments to do data science. (I still find it amazing that you can download a complete operating system with this software and have it up and running in a matter of minutes.) On the other hand, authors and teachers should be able to easily create custom software and data bundles for their readers and students. It's a shame to waste time on getting all the required software and date installed. When everybody's running the Data Science Toolbox, you know that you all have exactly the same environment and you can get straight to the good stuff: doing data science.

Q - What have you developed so far? And what is coming soon?
A - Because the Data Science Toolbox stands on the shoulders of many giants: Ubuntu, Vagrant, VirtualBox, Ansible, Packer, and Amazon Web Services, not too much needed to be developed, honestly. Most work went into combining these technologies, creating a command-line tool for installing bundles, and making sure the Vagrant box and AWS AMIs stay up-to-date.

The success of the Data Science Toolbox is going to depend much more on the quantity and quality of bundles. In that sense it's really a community effort. Currently, there are a handful of bundles available. The most recent bundle is by Rob Doherty for his Introduction to Data Science class at General Assemblyin New York. There are a few interesting collaborations going on at the moment, which should result in more bundles soon.

Thanks for sharing all the projects you've been working on - super interesting! Good luck with all your ongoing endeavors! Finally, let's talk a bit about the future and share some advice...

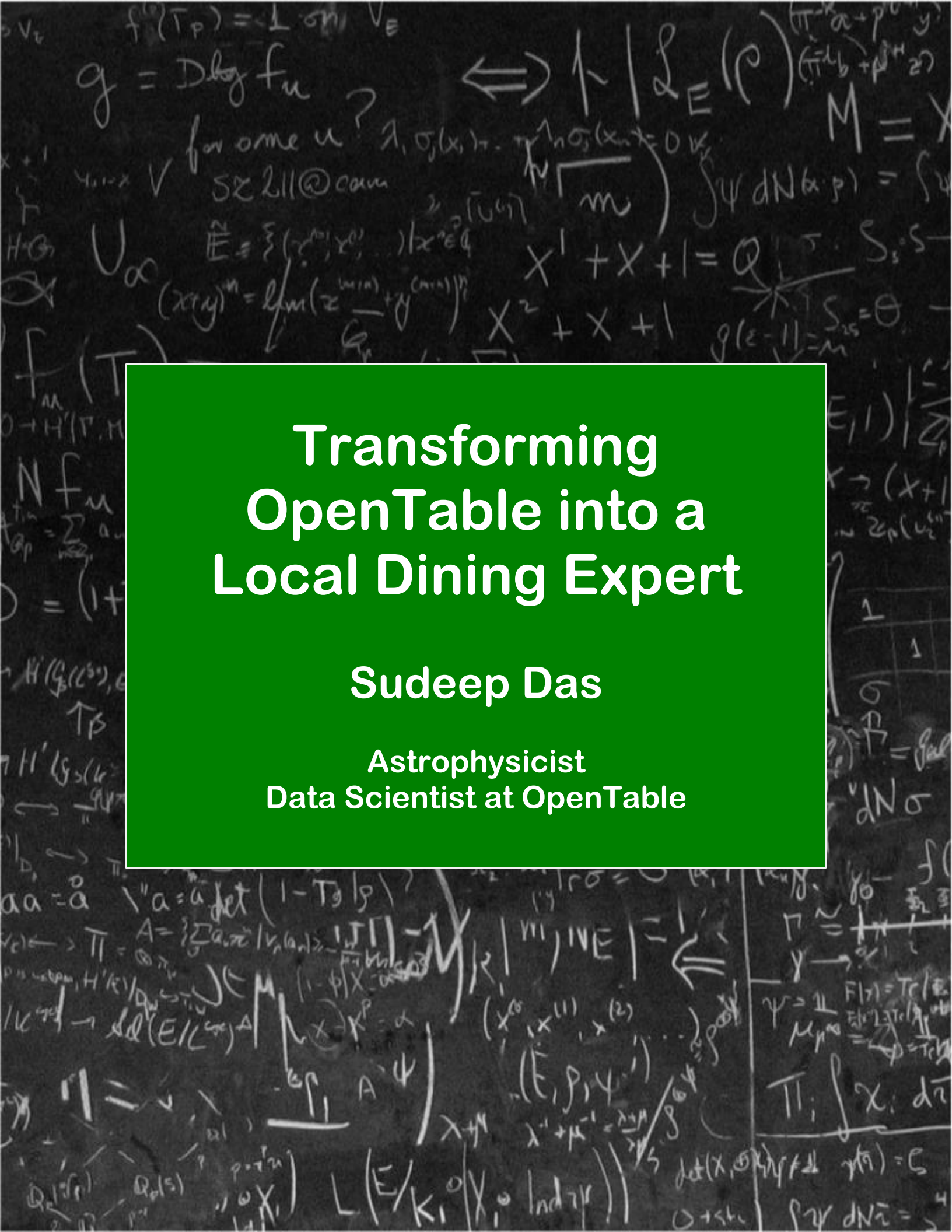
Q - What does the future of Data Science look like?
A - For me, and I hope for many others, data science will have a dark background and bright fixed-width characters. Seriously, the command line has been around for four decades and isn't going anywhere soon. Two concepts that make the command line so powerful are: working with streams of data and chaining computational blocks. Because the amount of data, and the demand to quickly extract value from it, will only increase, so will the importance of these two concepts. For example, only recently does R, thanks to magrittr and dplyr, support the piping of functions. Also streamtools, a very promising project from the New York Times R&D lab, embeds these two concepts.

Q - One last question, you said you're going back to the Netherlands? What are your plans?
A - That's right, back to the land of tulips, windmills, bikes, hagelslag, and hopefully, some data science! About three years ago, when I was

convincing my wife to come with me to New York City, the role of data scientist practically didn't exist in the Netherlands. While it still doesn't come close to say, London, San Francisco, or New York City, it's good to see that it's catching up. More and more startups are looking for data scientists. Also, as far as I'm aware, three data science research centers have been formed: one in Amsterdam, one in Leiden, and one in Eindhoven. These developments open up many possibilities. Joining a startup, forming a startup, teaching a class, consulting, training, research; I'm currently considering many things. Exciting times ahead!

Jeroen - Thank you ever so much for your time! Really enjoyed learning more about your background, your work at YPlan and both your book and toolbox projects. Good luck with the move home!

Readers, thanks for joining us! If you want to read more from Jeroen he can be found on twitter @jeroenhjanssens



Transforming OpenTable into a Local Dining Expert

Sudeep Das

Astrophysicist
Data Scientist at OpenTable

Transforming OpenTable into a Local Dining Expert

We recently caught up with Sudeep Das, Astrophysicist and Data Scientist at OpenTable. We were keen to learn more about his background, his work in academia and how he is applying data science in his new role - transforming OpenTable into a local dining expert...

Hi Sudeep, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - In no particular order, I am a coffee aficionado, a foodie, and a scientist. For most of my professional life, I have been an astrophysicist. After finishing my Ph.D. from Princeton in 2008, I moved to the Berkeley Center for Cosmological Physics as Prize Fellow, and then to Argonne National Laboratory as a David Schramm Fellow working on minute fluctuations in the afterglow of the Big Bang called the cosmic microwave

background. During the past year, I became increasingly interested in the booming field of data science and decided to switch fields to start an adventure in this new area. Currently, I am a data scientist at OpenTable using dining related data to help personalize the user experience, discover cultural and regional nuances in dining preferences, as well as help provide insights to restaurateurs. I am also an avid blogger, and write about science and data science on my blog datamusing.info.

Q - How did you get interested in working with data?
A - Well, much of my thesis work involved dealing with vast amounts of data from the extreme edges of the observable universe. In my every day work, I would perform a significant amount of munging, reduction, and analysis of the raw and noisy data collected by our telescope stationed in Chile. The last stage of the analysis would be applying machine learning and Bayesian techniques to make inference about fundamental parameters of the universe! This was a long and tortuous process with all kinds of nasty data problems one could imagine, but the results were rewarding! You could not do it if you did not love working with data.

Q - I can imagine! :) So, what was the first data set you remember working with? What did you do with it?
A - The first significant data set I worked on was the cosmic microwave background data on a patch of sky where there was supposed to be a big cluster of galaxies called the Bullet Cluster, and this cluster was supposed to leave an impression in the data. For several weeks, all we saw was noise. I was involved in making a map of that patch of sky, and tried various filters for suppressing the noise and various forms of

visualizations. Finally, I was able to find the tiny dark dot at the position where the Bullet Cluster was supposed to be.

Q - That must have been very rewarding! Maybe it was that moment, though was there a specific "aha" moment when you realized the power of data?
A - Undoubtedly, this was when I first saw the signs of an extremely faint signal called the gravitational lensing of the cosmic microwave background (CMB) in the data from our telescope. These are tiny distortions to the patterns in the CMB due to gravitational pull of massive structures in the universe. It took excellent observations, a large arsenal of statistical tools, excellent team work and careful analysis of data to come to this point. It was the first ever detection of this effect, it was amazing, and definitely an "aha" moment for me.

Q - Wow, that's very powerful! ... On that note, what excites you most about recent developments in Data Science?
A - While machine learning has been around for a very long time, and the basic methods are well established, what is really new is the enormous scale of data sets that is prompting both new ways of implementing established algorithms, as well as novel approaches to solving familiar problems at scale. Academic data sets used in traditional machine learning used to be small. Now, the game has changed with data sets becoming so large and also live (as in streaming). For example, even the apparently simple task of computing similarity between users has warranted new algorithms when the user base is in billions. Along with algorithmic developments, new ways of introspecting data have also come into play. Visualizations play a huge role in all stages of data science from

initial introspection to the interpretation of results. Modern day data science demands a multifaceted skill set that ranges from the ability to efficiently clean huge data sets, solid understanding of basic algorithms, excellent visualization skills, to creative ways of solving problems and in many cases, just seeing through the haze and applying common sense. All of this has made data science a very dynamic and colorful space to work in, which is what I like most about my current role.

I also believe that data science can play an important role in solving social problems. I am a mentor in the non-profit Bayes Impact program, and currently I am mentoring the fellows on two projects based in India and the US.

That must be fascinating! Thanks for sharing all that background. Let's switch gears and talk about your current role at OpenTable...

Q - What attracted you to the intersection of data science and Restaurants/Dining? Where can Data Science create most value?
A - I have always been a die-hard foodie, and even before joining OpenTable, I used the service frequently to make restaurant reservations. While doing so, I always wondered how nice it would be if the app somehow knew my dining habits and preferences and suggested restaurants to my liking. Specially, as an academic, I was traveling frequently, and I wanted the app to be my local foodie expert, rather than just a tool to book tables. Also, I felt that there should be a way to distill the reviews at a restaurant into a set of succinct insights that would tell

me what this restaurant is all about at a glance, without having to read through all the reviews. If I were a restaurateur I would like to have my restaurant's data analyzed to inform myself of how my business is doing, in general, and in comparison to others. Now, as a data scientist at OpenTable I am helping build many of these data driven features. The idea is to transform OpenTable from a transactional to an experiential company, and that is where I think Data Science is going to create the most value.

Q - That's great! So what specifically led you to join OpenTable?
A - I have always wanted a to work in a space that will marry my data science skills with my passion and domain knowledge in the food and dining space. OpenTable is the world leader in dining reservations and has an extensive and rich data set to work with, so it was an obvious choice.

Q - What are the biggest areas of opportunity/questions you want to tackle?
A - Using data science to help transform OpenTable into a local dining expert who knows me very well, and can help me and others find the best dining experience wherever we travel is incredibly exciting. This entails a whole slew of tools from natural language processing, recommendation system engineering, predictions based on internal and external signals that have to work in synch to make that magical experience happen. We also want to use data science to create unprecedented value and tools for the restaurateurs who use our service.

Q - What learnings/skills from your time in academia will be most applicable in your new role?

A - Coming from academia, and especially with a background in math, computation, and data intensive field, I feel at home with munging through large data sets and have a strong footing in statistical methods and algorithms. Academia has also trained me to pick up a fresh research paper, and quickly implement its algorithm and adapt it to our use cases. Astrophysics is also very visually driven, so I have a knack for visualizing OpenTable data in novel and non-standard ways to extract insight. Another thing that research has taught me is the importance of experimentation. I'd love to play an important role in designing experiments to field test various flavors of our data science solutions.

Q - Makes sense! So what projects are you currently working on, and why/how are they interesting to you?
A - Broadly speaking, I work on extracting insights from reviews and past dining habits of diners using a whole suite of machine learning tools that include Natural Language Processing, Sentiment Analysis, Recommendation Systems, Clustering and classification algorithms, just to name a few.

Q - And how is data science helping? What techniques, models, software etc are you using?
A - I use Python a lot, relying heavily on Pandas, scikit-learn and gensim. For visualizations I use d3.js, Matplotlib, Bokeh. Recently, I have also been using the scala-based package Spark to implement machine learning solutions at scale.

Q - What has been the most surprising insight/development you have found?

A - There are many, but nothing we're ready to share quite yet. Stay tuned.

Q - Will do! Final question on OpenTable … How do you/your group work with the rest of the organization?
A - We sit at the heart of various projects that transcend the boundaries of several teams here. From marketing, to mobile and web in the front end, to architecture, engineering and product, we work in close collaboration with a large number of teams across the organization and around the world.

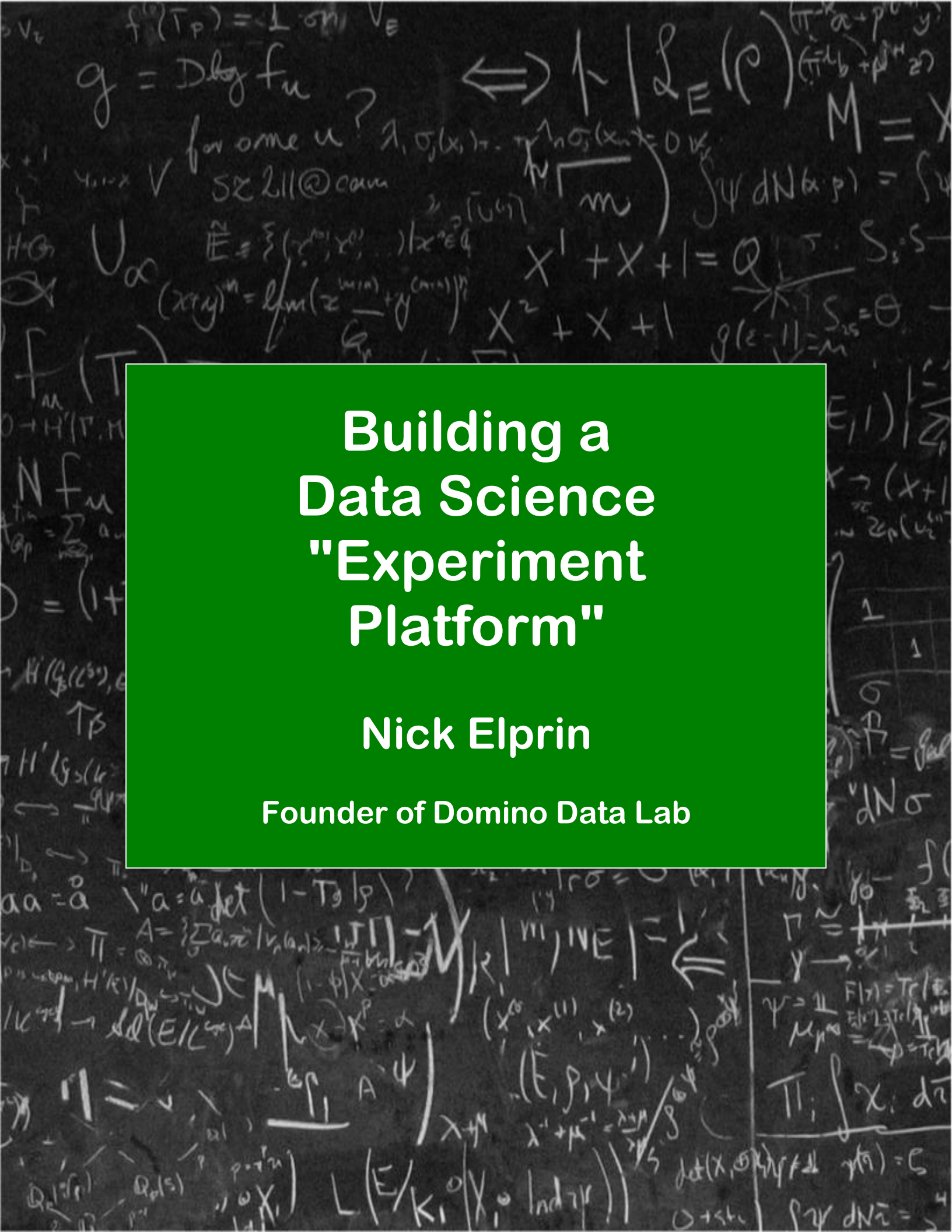
Thanks for sharing all that detail - very interesting! Good luck with all your endeavors - sounds like a great time to be part of the OpenTable team! Finally, let's talk a bit about the future and share some advice … …

Q - What does the future of Data Science look like?
A - It looks very promising, and I feel like we are really scratching the surface now. With time, this role will take a more defined shape, and will be able to tackle bigger and broader problems. We will have great power in harnessing the information in data to really impact society on various fronts. The internet of things will also bring data science into every day appliances and how they communicate with each other and the environment. As with any great power, comes great responsibility! I think applying data science in a responsible way will be the key to continued success in this new field.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Don't be afraid to attack a problem from non-standard angles, and be always in the know of new advances in the field. Pick problems that really matter and can have impact. Never stop learning. Share your learnings through open source code and blog posts.

Sudeep - Thank you ever so much for your time! Really enjoyed learning more about your background, your work in academia and your remit and objectives in applying data science at OpenTable. Good luck with all your ongoing projects!

Readers, thanks for joining us! If you want to read more from Sudeep he can be found online here and on twitter @datamusing.



Building a Data Science "Experiment Platform"

Nick Elprin

Founder of Domino Data Lab

Building a Data Science "Experiment Platform"

We recently caught up with Nick Elprin, Founder of Domino Data Lab. We were keen to learn more about his background, his thoughts on Data Science as a Service and the functionality he has built at Domino...

Hi Nick, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - Before I founded Domino Data Lab, I spent about seven years building analytical software tools at a large hedge fund. As you can imagine, economic modeling at that scale requires a lot of sophisticated data analysis, so we built some interesting technology. Before that, I studied computer science at Harvard.

Q - How did you get interested in working with data?
A - Well, I was interested in working with software, and building software to solve interesting problems. Then it turned out that there are a lot of interesting problems around data and data science that demand help from software.

Q - Was there a specific "aha" moment when you realized the power of data?
A - My first job out of college was at an algorithmic hedge fund. We would process hundreds of thousands of data series to predict market movements. That was certainly a really powerful example of how you can use data.

Q - I can imagine! ... Let's talk a little about the evolving field of Data Science - how have things changed over the past 5 years?
A - It has scaled up, along every dimension: bigger data sets, more sophisticated analytical techniques, large teams working together, and a wider range of problems that now seem like good opportunities for applying data science. As companies transitioned from “we should collect lots of data” to “we should _do something_ with all our data,” the job of data scientist became much more demanding.

Q - How has the life of a Data Scientist changed as a result?
A - The variety of skills needed is really overwhelming, resulting in “unicorn-like” job descriptions for data scientists. One thing we see a lot is that data scientists have to do software engineering to build tools for themselves. This happens even within companies with strong engineering teams, because the engineers are all dedicated to working on the product (or something else) rather than providing support to data scientists.

Q - What range of tools / platforms are being developed to support this evolution?
A - There are lots of tools to address specific parts of a data science workflow. For example, there are tools that make it easier to do data cleaning; tools that make it easier to manage and explore big data sets;

lots of great libraries in Python and R for specific data science techniques; lots of great tools for visualization and reporting. But nobody is really stepping back and saying, “you know, it doesn’t make sense that we’re asking our data scientists to do so many different things.” With Domino, we’re trying to cut a lot of the engineering “schlep” out of the entire analytical lifecycle, from model development all the way to deployment.

Q - Got it. So what is the key enabler of Data Science as a Service?
A - One of the things we think is really important is supporting the data scientist in the way they want to work rather than trying to change it. You see products that expect users to change the language they use or the workflow or something like that. At Domino, we really try to minimize any impact on how the data scientist wants to work. So for instance we support R, Python, Matlab, Julia, etc; and our users work in the IDEs and tools they already use, not some new editor that we’ve built.

On that note, let's switch gears and talk about Domino in more detail...

Q - What specific problem does Domino Data Lab solve? How would you describe it to someone not familiar with it?
A - I like to describe Domino as an "experiment platform": it lets data scientists improve their analyses faster by making it easy to run, track/reproduce, share, and deploy analytical models. Normally these capabilities would require a lot of engineering work and hassle to build and maintain, but Domino gives you these “power tools” out of the box.

That's the short version. For a longer version, it's easiest to just describe Domino's main areas of functionality:

1. Domino lets you move your long-running or resource-intensive compute tasks off your machine onto powerful hardware with "one click" (either in the cloud, or on a cluster behind your company's firewall). And you can run as many compute tasks as you want in parallel. So instead of being limited by your desktop or laptop, you can run more experiments in parallel across an unlimited number of machines. It's basically the simplest way to get access to an industrial strength compute cluster.
2. Every time you run your code, Domino automatically keeps a snapshot of your work — including your data sets, and the results you produce — so you can reproduce past work and always have a record of how your analysis has evolved. This is critical to analytical workflows, which tend to be highly iterative and exploratory.
3. Because Domino tracks and organizes your work centrally, it's easy to work with collaborators. It's like Github for data science. Domino will keep your team updated as changes happen, and let you merge your work with other people's.
4. Finally, Domino lets you package up your analytical models for future use. You can put a UI around your model, so non-technical users can run your analysis without interacting with your code -- or bothering you -- at all. Or you can put a RESTful API interface on top of your model, so existing software systems can interact with it. Domino provides all the plumbing to let you "deploy" your model without any setup or hassle.

Q - That's a lot of functionality! What would you say are the main benefits / attributes of your platform?
A - I think the key attribute of the platform is the idea of centralizing your analysis. Moving analysis off of analysts' desktops onto a central server unlocks a lot of power. For example, because Domino is a central hub for analysis, it scales out the hardware behind the scenes easily; it synchronizes and shares work across multiple people; and it can automatically track changes as you work. And at a higher level what all this means is that the data scientist gets to experiment faster … and therefore learn faster.

Q - That's great! What are some of the most interesting analysis/projects you have hosted?
A - Unfortunately I can't say much about the most interesting ones, because most of our clients are doing proprietary work (some of which we don't even know about). One of our customers uses machine learning to do spam detection in social media, which is interesting because the "vocabulary" changes rapidly: as new people and terminology enter the social media domain, models need to be updated and re-trained rapidly. Another one of our big customers is a car manufacturer that uses Domino to process data it collects from various sensors. More specifically, a reliability engineering team uses Domino to run analysis to improve the reliability of different parts of the car. It's very rewarding to know we're helping — if indirectly — with something as serious as that.

Thanks for sharing all that detail - very interesting! Good luck

with everything you're doing at Domino! Finally, let's talk a bit about the future and share some advice … …

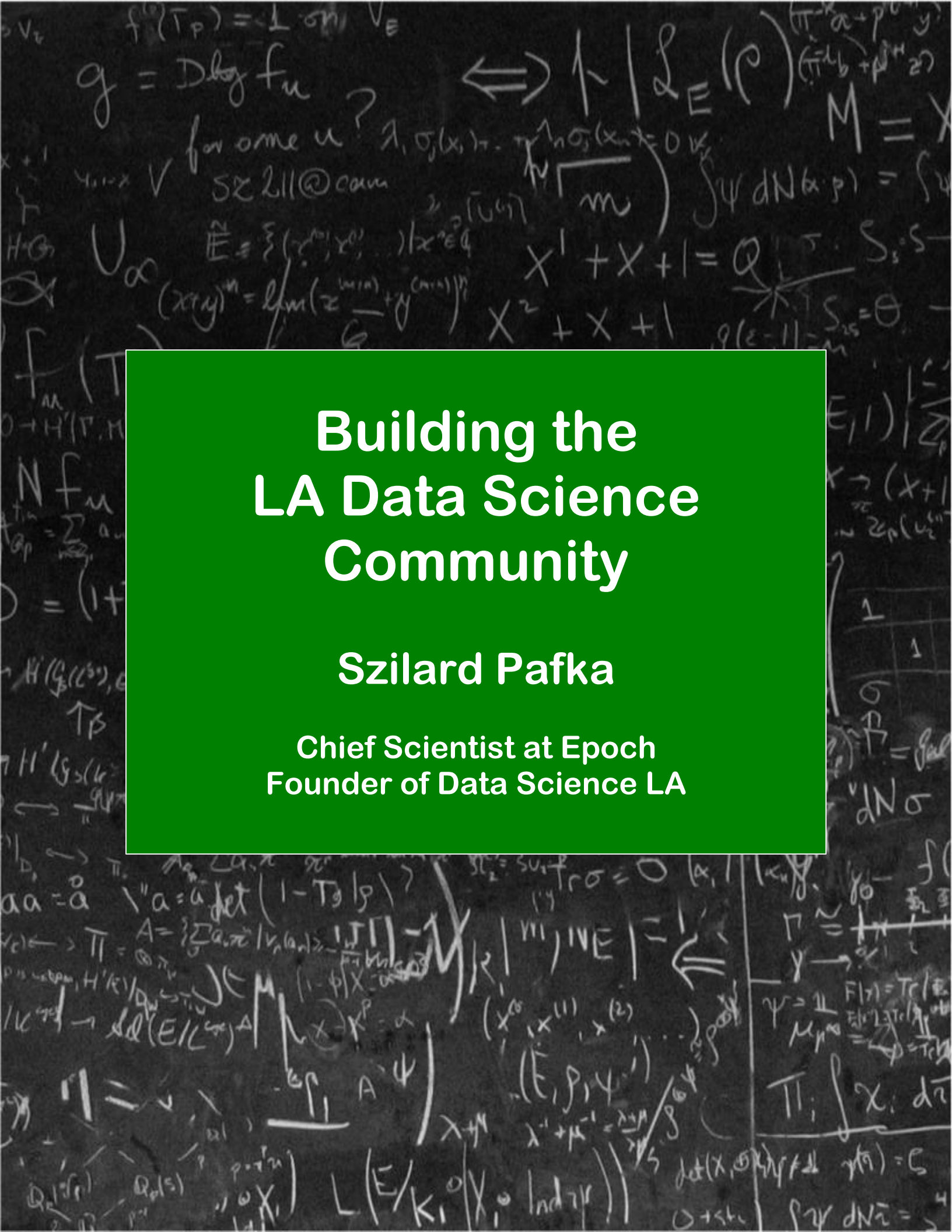
Q - What does the future of Data Science look like?
A - I think one interesting question is, "how much can be automated?" I've seen several products that seem to promise something like, "upload your data and we'll find the insights for you." My personal view is that this is misguided; that there is a critical "human" element of this work that won't be automated for a long time (until we develop a real artificial intelligence, I suppose). "Can we automate data science" is a bit like asking, "could we automate science." That's because, like any scientific or truth-seeking activity, the questions you ask, and the places you look, matter as much if not more than the techniques you use. And I don't see software being able to ask insightful questions anytime soon.

So… my bet is that we will make tools better and better at augmenting — rather than replacing — human intelligence, understanding, inquisitiveness, and domain expertise. I think the key metric for measuring the progress of our data science tools is: what percentage of time are analysts spending on their core problem, rather than on distractions (e.g., connecting to data sources, waiting for code to run, configuring infrastructure). I don't think that number will ever get to 100%, but it should get much higher than it is today.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Get your hands dirty as much as you can. Trying things is the best way to learn.

Nick - Thank you ever so much for your time! Really enjoyed learning more about your background, your thoughts on Data Science as a Service and the functionality you have built at Domino. Good luck with all your ongoing projects!

Readers, thanks for joining us! If you want to know more, Domino can be found online here. and on twitter @DominoDataLab .



Building the LA Data Science Community

Szilard Pafka

Chief Scientist at Epoch
Founder of Data Science LA

Building the LA Data Science Community

We recently caught up with Szilard Pafka, Chief Scientist at Epoch and Founder of Data Science LA. We were keen to learn more about his background, his role building the LA Data Science community and his work at Epoch...

Hi Szilard, firstly thank you for the interview. Let's start with your background and how you became interested in working with data...

Q - What is your 30 second bio?
A - My primary field of study at university was Physics (BSc/MSc/PhD) with adventures in Computer Science (BSc) and Finance (MA). Like many people in Physics at that time (late 90s) I was working with data, models, computational approaches, and I ended up working in risk management in a bank while still working on my PhD research involving statistical modeling of financial prices. In 2006 I came to California to be the Chief Scientist for Epoch, essentially doing data science (data analysis, modeling/machine learning, data visualization etc.) way before the "data science" term has been used to describe this. In 2009 I started organizing an R meetup in Los Angeles (which retrospectively was the very first data science meetup in LA) with the goal of bringing together data

professionals to learn from each other. More recently I started other meetup groups that are focused on my other professional interests (DataVis, Data Science), and finally a few weeks ago with the involvement of a couple of other volunteers, we started datascience.la, a website serving the growing LA data science community.

Q - How did you get interested in working with data?
A - I was always interested in math, physics and later in computers (which, for me, the first was C64 in the late 80s). Later on I got involved in data, modeling and computing, my Monte Carlo simulations in the field of materials science (dislocation systems) generated lots of data that needed to be analyzed, I think that's how I started more seriously to use tools for data munging/analysis/visualization.

Q - So, what was the first data set you remember working with? What did you do with it?
A - There were a couple of datasets like my running times that I used paper and pencil to graph them (late 80s). Later on I played with inputing the data on the C64 and graphing it that way. Without reading any formal literature at that time, I became fascinated by the power of visualization (e.g. to see trends or detect outliers).

Q - Was there a specific "aha" moment when you realized the power of data?
A - I cannot pinpoint a specific time, but there is also a trend: the more sources we have for data (collected/generated etc) the more useful it can become. But there are also dangers especially to privacy and security as it becomes more and more clear.

Q - Makes sense ... On that note, what excites you most about recent developments in Data Science?
A - First, with all the hype let's recognize that data science is many decades old (we could go back even to John Tukey). Many of the machine learning algos have been developed in the 90s or 2000s. The basic software tools used by most data scientists are over 10 years old. One the other hand, besides this solid foundation, there is an extraordinary pace of new developments. Many of the new add-on tools have increased hugely my productivity (e.g. Rstudio, knitr, shiny, and more recently dplyr). Many others make it possible to do things that was not possible before (increasing computing capacity also helps). We also have now open source tools to tackle larger and larger datasets (Hadoop, but more excitingly for data scientists tools that support interactive analysis such as Impala or Spark).

Q - What industries do you think will benefit most?
A - I think it's all over the place. We are getting/collecting data from more and more sources, more and more industries, from sensors, from humans, from crowdsourcing and the list goes on. Next this data is processed, analyzed, used to improve processes. It's hard to imagine any industry that will not benefit.

Very true! Thanks for sharing all that background. Let's switch gears and talk about your role promoting the data science community in LA...

Q - How did you come to found / organize Data Science LA?
A - DataScience.LA has its roots in the LA R meetup that I started with Prof. Jan de Leeuw in 2009. While it was an R meetup, my goal from the beginning was to put everything in a more general context of data analysis / modeling. With the raise of "data science" as a term for our essentially old craft, we started to have events on more general topics and ultimately I started new meetup groups to focus on specific parts of data science (DataVis) or the overall process of combining tools in various companies. DataScience.LA takes this to a new level, by preserving the content of the meetups (slides, code, video recording etc.) and involving the community in new ways (such as blogging). It is also a way to scale up the community leadership by involving top-notch data scientists from LA in serving the needs of the growing data science community.

Q - That's great! So what are your primary goals?
A - I touched a bit on that in the answer to the previous question, but in one phrase it would be building a world-top data science community in LA.

Q - On that note - What has been the most memorable meet-up presentation(s) you've seen?
A - We had many-many excellent talks from professionals in LA and outside LA, for example at the R meetup we had Hadley Wickham, Dirk Eddelbuettel, Michael Driscoll, Ryan Rosario (to name just a few of the better known names from outside LA). At the DataVis meetup we had a fascinating talk by the LA Times' Data Desk, while at the Data Science/ Machine Learning meetup we had talks e.g. by Netflix, Activision (Call of Duty) and Factual.

Q - Wow - that's a terrific bunch of speakers! ... What advice would you give to others looking to organize a Data Science group/meet-up in their own city?
A - I would encourage everyone, it's such a rewarding endeavor. It ultimately boils down in getting speakers, a venue (and a sponsor for food) and most importantly members. Use meetup.com, it takes care nicely of all administration for you. For venue, talk to companies willing to host (and provide food), it is much easier now than say 5 years ago. Don't get discouraged by low attendance, we had about 30 people attending the R meetup in the first 2 years (well, the number of R users in general exploded only after that).

Got it - good advice! Now, we'd also love to talk a bit about your role at Epoch...

Q - How are you using Data Science at Epoch? What types of questions does it help you solve?
A - Epoch is an online credit card transaction processor, so obviously the main problem is fraud detection, but there are many other areas for example in sales tracking, marketing or consumer satisfaction that can be improved by models or insights from data. Epoch was wise enough to hire a data scientist way before “data science” got hot and we have developed several sophisticated tools starting many years ago.

Q - What has been the most surprising insight you've found?
A - Unfortunately I'm not allowed to share details about Epoch, but my general philosophy is to start with a business problem a company needs to solve (usually improving the bottom line), understand the domain, look

at the data and come up with solutions that are best suited for the problem – the outcome can be an advice for an action or a model that can be deployed. Sometimes simple things such as a real-time dashboard can provide a lot of value (monetarily), in other cases you might need a fancy machine learning algorithm.

Q - Makes sense! How does your team interact with the rest of the organization?
A - In any data-driven organization, data science should have a central role. It has to interact with (advise) top management and it has to connect to all parts of the organization where data can drive decisions or optimize processes. This is fairly easy to do in a small organization but in a larger one it has its challenges. Ideally, data scientists learn the domain knowledge in the various parts of the organization, explore the data, give strategic advice, develop models that can operationalize micro-decisions, but they also disseminate a data-centric view across the organization and mentor key personnel in other departments so that they can use increasingly data and results of data analysis and models in their day-to-day job.

Thanks for sharing all that detail - very interesting! Good luck with all your endeavors - both at Epoch and in the broader LA Data Science community! Finally, let's talk a bit about the future and share some advice … …

Q - What does the future of Data Science look like?
A - Let me quote Niels Bohr: “Prediction is very difficult, especially if it's about the future.” Data science is good at predicting micro-events where

we have data about lots of past micro-events, we fit a distribution (implicitly most of the time e.g. in some non-parametric model or by applying some learning algorithm) and we assume our world is stationary. Predicting macro-events in society or technology is a completely different thing.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Sure. Get a balance of theory and hands on experience. For theory there are numerous books, free classes etc. For hands-on spend time with "looking" at data. This involves mostly tedious data munging, but besides preparing, cleaning the data you gain understanding about the data and the domain. If you do modeling, make sure you understand how it works, what are the assumptions, limits, pitfalls, and spend enough time with evaluating your models.

Szilard - Thank you ever so much for your time! Really enjoyed learning more about your background, your role building the LA Data Science community and your work at Epoch. Good luck with all your ongoing projects!

Readers, thanks for joining us! If you want to find out more about Data Science LA, they can be found online here and on twitter @wwwDSLA.

CLOSING WORDS

We very much hope you enjoyed this interview series.

If so, we would ask 2 things – please

1. Share it with friends and colleagues you think would enjoy it

2. Consider signing up for our newsletter – to keep up with the next interview series as well as receive weekly curated news, articles and jobs related to Data Science (http://www.DataScienceWeekly.org)

Thanks again to all the interviewees – looking forward to the next 15!

Hannah Brooks, Co-Editor
DataScienceWeekly.org