Data Science Weekly

Interviews with Data Scientists

Volume 1, April 2014

FOREWORD

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

The 15 interviewees have varied roles and focus areas: from start-up founders to academics to those working at more established companies; working across healthcare, energy, retail, agriculture, travel, dating, SaaS 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

CONTENTS

Visual Image Extraction

Parham Aarabi

CEO of ModiFace &
University of Toronto Professor

Visual Image Extraction

We recently caught up with Parham Aarabi - PhD in Electrical Engineering from Stanford and Professor of Electrical and Computer Engineering at the University of Toronto since 2001; and co-founder of ModiFace Inc. He and his team have been leading fascinating work leveraging Big Data to bolster social image search...

Hi Parham, firstly thank you for the interview. Let's start with your background.

Q - How did you get interested in working with data?
A - I was always interested in making computers more intelligent, and the best way to do that was to enable them to hear and see better. This motivated me to focus on signal and image processing, as well as data processing algorithms in general.

Q - What was the first data set you remember working with? What did you do with it?
A - The first data I remember working with was a set of analog readings from an exercise bike that I connected to my computer in order to create a very crude augmented reality exercise bike. I was 13 at the time.

Q - Was there a specific "aha" moment when you realized the power of data?
A - The true "aha" moment came a few years ago. This was the moment that my team and I realized the power of data and that it could actually solve relevant problems. For years, many companies and researchers had been exploring facial skin-tone detection systems that would tell you your true skin pigment. The hard part with this is that your skin will look vastly different in photos (and sometimes, in real life) due to different lighting and facial tanning conditions. We tried a simple idea. What if you analyzed hundreds of photos of a person over many years instead of just a few photos? Could you remove the lighting noise or errors and find the true tone of a person's face?
The answer turned out to be an astounding yes. By scanning hundreds of tagged Facebook photos of a person, we could accurately detect their true skin tone. We were surprised how well this worked.

Parham, very compelling background. Thank you for sharing. Next, let's talk about your work in Visual Information Extraction.

Q - What are the main "types" of problems being tackled in the visual information extraction space? Who are the big thought leaders?

A - The biggest problem is to identify objects in images and categorize them (i.e. find out a cat vs. a person vs. a tree). There are many researchers working on this including our research group at the University of Toronto.

Q - What have you been working on this year, and why/how is it interesting to you?
A - This year my team has focused on a number of projects, but the most interesting has been "understanding" relationships based on location of tags in images. It is a simple idea, that whoever you stand beside when you take photos is more closely affiliated with you. In practice, it results in very interesting relationship graphs which often mimic the real-life family or friend structure. There are other projects that we are working on, including detecting heart signal shapes from color changes on a person's face, or detecting cancerous moles by visually analyzing skin. However, the latter projects are still ongoing.

Very interesting - look forward to hearing more about the ongoing projects. Let's move on to talk about what you are building at ModiFace.

Q - Firstly, how did you come to found ModiFace?
A - We had been working on lip reading and lip tracking systems for years, when interest from a cosmetics brand resulted in us applying our lip detection to lip enhancement and visual alteration. This was in 2006 and from there ModiFace was founded.

Q - What specific problem is ModiFace trying to solve? How would you describe it to someone who is not familiar with it?
A - ModiFace technology simulates skin-care and cosmetics products on user photos. So, a skin care product that reduces dark spots, or a shiny lipstick, or a glittery eyeshadow ... we specialize in making custom simulation effects for all facial products. This is us as a core. From this technology we have built a variety of apps for popular beauty brands (like L'Oreal or Vogue magazine), as well as ModiFace-branded mobile apps for consumers which have been extremely popular (with over 27 million downloads to date).

Q - What do you find most exciting about the intersection of technology and beauty/fashion?
A - First, it is an open space. There have not been too many signal processing scientists tacking problems in this area, which makes it interesting. Finally, consumers who are deeply engaged with beauty/fashion are very tech friendly, and open to new technologies. All of these elements combine to make an area ripe for disruption.

Q - Which ModiFace applications/technologies have been most successful?
A - Our core technology which consists of simulating makeup and skin-care/anti-aging effects is at the heart of what we do.

Q - What else should we know about ModiFace?
A - We are the largest mobile beauty company on the planet with 27+ million mobile downloads. Not a lot of people know/realize this about us.

Sounds like things are going very well - long may it continue! Parham, next, let's talk about you a little more. .

Q - What in your career are you most proud of so far?
A - By far, ModiFace. It is an interesting company at the intersection of mobile, beauty, fashion and technology. But the team that has come together over the last 7 years is what makes it truly special.

Q - What mistakes have you made (if any!)?
A - Many. There have been projects that we have pursued that were dead-ends. For example, I created a multi-camera image searching company ten years ago which was technically neat, but offered a solution that no one really wanted. The company didn't work. There are many examples like this. But as long as you learn from your mistake, you look at them as steps towards a positive outcome.

Q - What distinguishes your work from that of your contemporaries?
A - That is a hard question to answer. Everyone pursues directions based on what motivates and excites them. My career path has been perhaps a bit different in that it has this blend of entrepreneurship and academia. Most eventually settle in one of these, for good reason. I have tried to make the two fit together in synergy. Time will tell if this was yet another mistake, or an unorthodox but synergistic equilibrium.

Q - What publications, websites, blogs, conferences and/or books do you read/attend that are helpful to your work?
A - There are many signal processing conferences/journals. I also read

the popular tech blogs on a daily basis, and of course daily news like CNN. That is an essential part of my morning routine.

Q - How important is data in your personal life?
A - It is hard to answer this one. It is perhaps fair to say that personally, I am driven more by gut feelings than data.

Very insightful. Really appreciate your honesty. Finally let's talk about the future and where you think your field is headed...

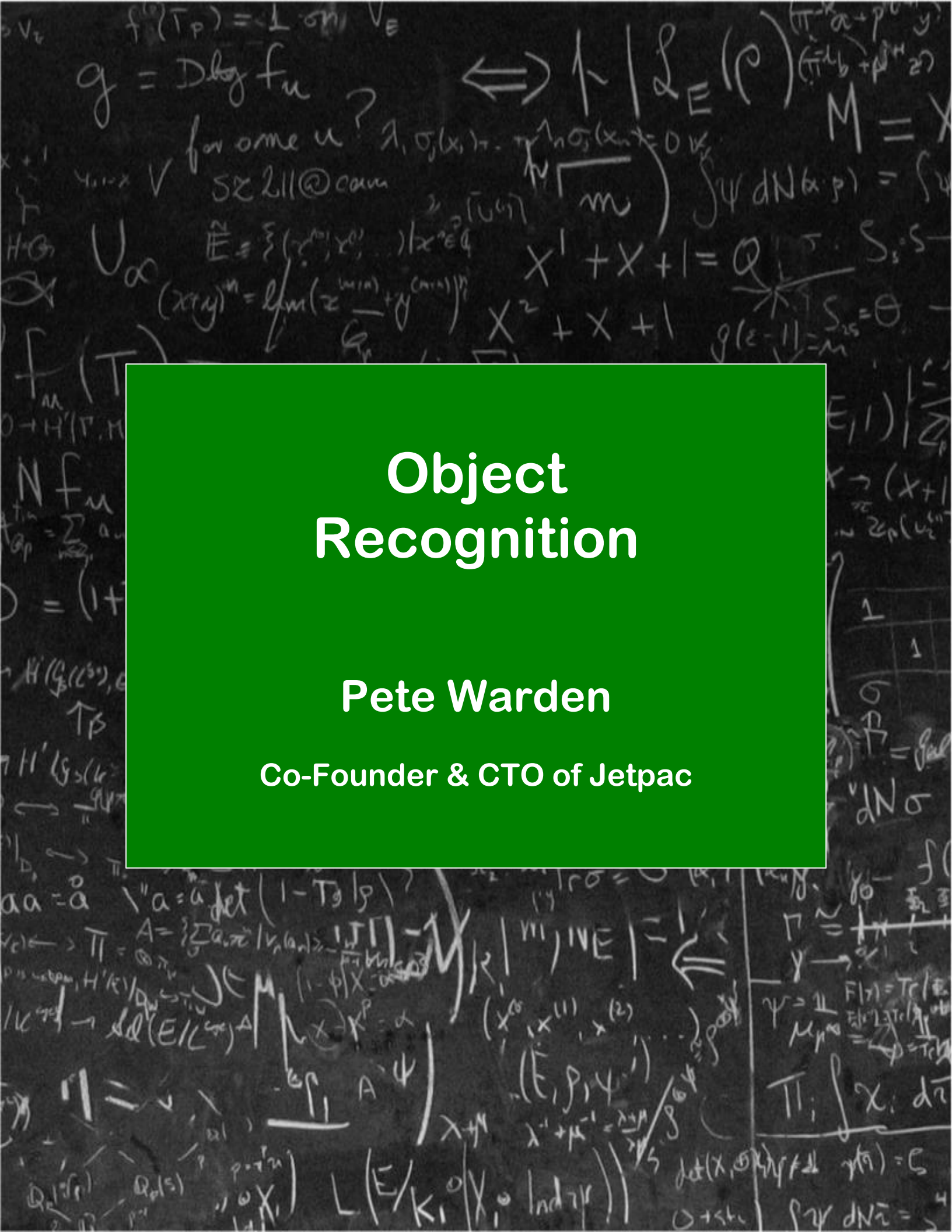
Q - What does the future of Big Data / Visual Information Extraction look like?
A - It will be both exciting and scary. Computational image understanding will get better, and sooner or later Facebook and other social networks will know everything about us, from our relationships to what we like to do. This much information at the hands of a few companies will eventually be cause for concern, but so far our loss of privacy has been inch by inch which has been hard to notice.

Q - What is something a smallish number of people know about that you think will be huge in the future?
A - The impact that mobile apps, especially data processing and intelligent apps, will have on our society. Everything from how we study, to how we treat diseases, to how we shop, will change. But we are just beginning this process ... much more awaits us in the next few years.

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

A - Pick problems that in your view truly matter. Too often, we find ourselves pursuing goals that deep down we don't believe in, and this will only lead to failure or unappreciated success.

Parham – Thank you so much for your time! Really enjoyed learning more about your research and what you are building at ModiFace, as well as your career to-date and your views on the future of Data Science. ModiFace can be found online at http://modiface.com and Parham Aarabi @parhamaarabi.



Object Recognition

Pete Warden

Co-Founder & CTO of Jetpac

Object Recognition

We recently caught up with Pete Warden, Co-Founder and CTO of Jetpac, which is using Big Data and Object Recognition to build a modern day Yelp...

Hi Pete, firstly thank you for the interview. Let's start with your background and the work going on in Object Recognition right now...

Q - What is your 30 second bio?
A - As you mentioned, I'm the CTO of Jetpac. I was born in Britain and am now living in San Francisco, I used to work for Apple, I've written some books on data for O'Reilly, and I blog at petewarden.com.

Q - What are the main "types" of problems being tackled in the Object Recognition space? Who are the big thought leaders?
A - There's an amazing amount of great research out there around recognizing objects in images, but there have been surprisingly few commercial applications. The biggest successes have been specialized facial recognition security applications, bar-code scanners like Occipital's Red Laser, and Google's image search. Everybody knows

object recognition is a crucial foundational technology for the future, but because it's currently so unreliable it's been hard to build any consumer applications around it.

The problem is that object recognition is incredibly hard, and even the best algorithms make a lot of mistakes. If you're doing a search application, these mistakes mean a lot of bogus images showing up in the search results. The fortunate thing about Jetpac is that we have hundreds or thousands of photos of each place we feature, so we're able to derive data from applying our algorithms to all these samples. An algorithm that only spots a mustache 25% of the time would give a terrible experience if you were relying on it to deliver search results, but applying to a lot of photos at the same place gives you a reliable estimate of how many mustaches are present. Even if individual photos might be mis-identified the errors cancel out.

Q - What are the biggest areas of opportunity / questions you want to tackle?
A - Photos are data! That's the most exciting thing about what we're doing, once you're able to extract useful information about a place from a collection of photos taken there, all those billions of photos gathering digital dust on hard drives around the world turn into an incredible source of data. We'll be able to answer questions about pollution by analyzing the intensity of sunsets, spot smog in photos, build a much better picture of how people move around neighborhoods to help plan urban regeneration, there's an endless number of pressing problems this data can help with.

Q - What Data Science methods have you found most helpful?
A - My friend Monica Rogati likes to say that division is her favorite algorithm. I specialize in uncovering new information from discarded sources, mining neglected data exhaust, so most of the work I do is the initial extraction of useful features from apparently useless noise. Once I have the data, most of the analysis is fairly primitive database joins, sums, and division. We use machine learning, neural networks, and a lot of other fancy approaches to analyze the images, but Excel formulas are key too. A lot of people underestimate the usefulness of old-school data tools like spreadsheets.

Q - What are your favorite tools / applications to work with?
A - I have to give a plug to the Data Science Toolkit here. It's a custom virtual machine, available as a Vagrant box and an Amazon EC2 image, and it comes pre-installed with my favorite open source tools and data sets. It's focused on taking messy, unstructured data and turning it into something useful, so it has everything from geocoders, sentiment analysis, and document conversion, to entity extraction from text. There are a lot of amazing open-source tools out there, but they're often hard to install and interface with, so I wanted to make my personal favorites available in a turn-key way.

Pete, very interesting background and context - thank you for sharing! Next, let's talk more about what you are working on at Jetpac.

Q - How did you come to found Jetpac?
A - My co-founder Julian was using some of my open-source tools, and

he was peppering me with questions. As soon as I talked with him, I realized how fantastic a source of data he was looking at in the hundreds of billions of social photos we're sharing.

Q - What specific problem is Jetpac trying to solve? How would you describe it to someone who is not familiar with it?
A - We help you discover fun places to go, both locally and when you're traveling. We aim to offer the kind of insights you'd get from a knowledgeable local friend about the best bars, hotels and restaurants. The information we get from the mass of pictures, and the pictures we present in the guide, combine to give you a much better idea of what a place is like than any review-based service.

Editor Note - If you are interested in more detail behind how Jetpac's technology works, Pete's recent blog article is very insightful. Here are a few highlights:

Image-based measurements *- The most important information we pull out is from the image pixels. These tell us a lot about the places and people who are in the photos, especially since we have hundreds or thousands of pictures for most locations.*

One very important difference between what we're doing with Big Data and traditional computer vision applications is that we can tolerate a lot more noise in our recognition tests. We're trying to analyze the properties of one object (a bar for example) based on hundreds of pictures taken there. That means we can afford to have some errors in whether we think an individual photo is a match, as

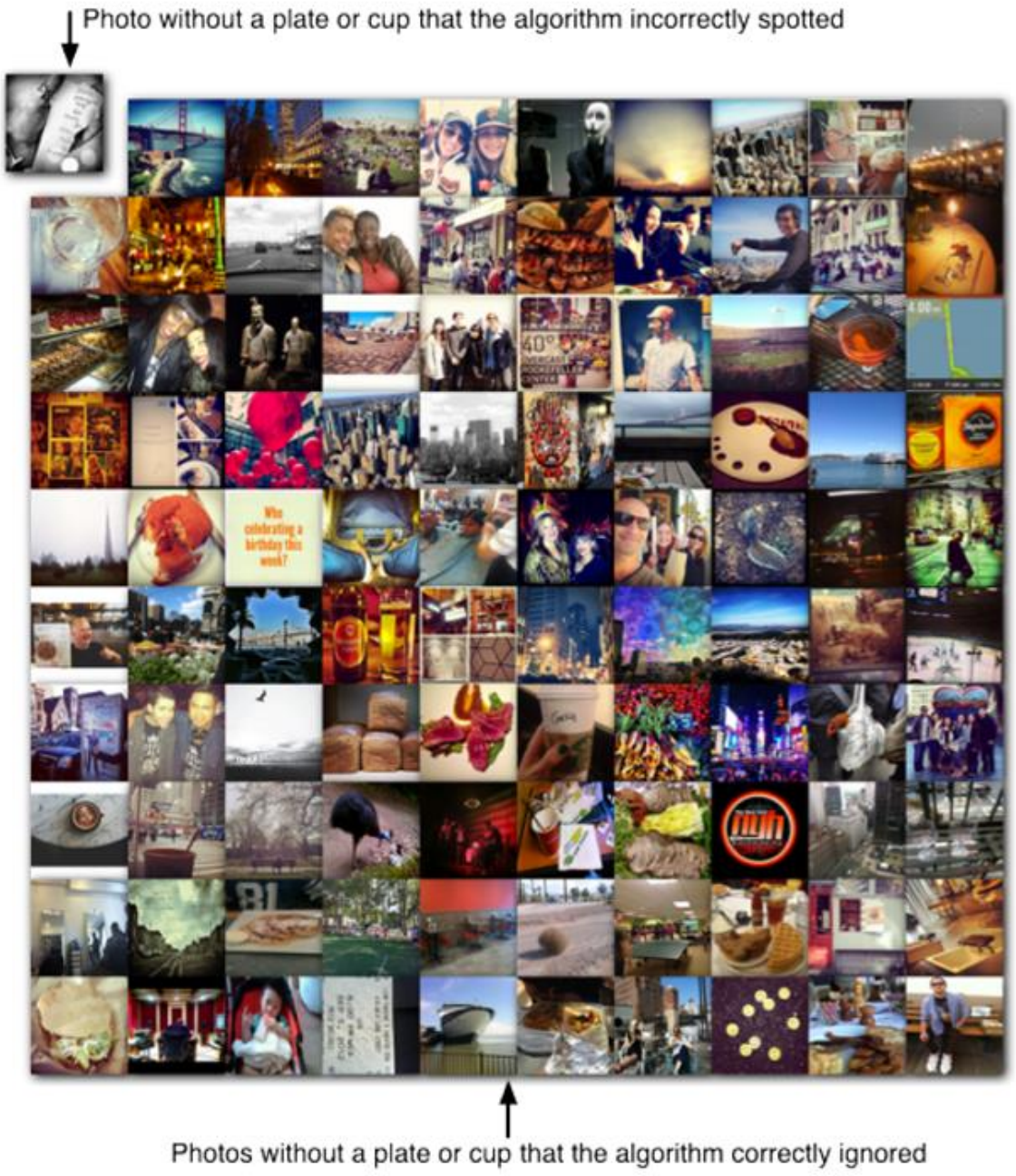
long as the errors are random enough to cancel themselves out over those sort of sample sizes

Testing *- Internally, we use a library of several thousand images that we've manually labeled with the attributes we care about as a development set to help us build our algorithms, and then a different set of a thousand or so to validate our results. All of the numbers are based on that training set, and I've included grids of one hundred random images to demonstrate the results visually.*

We're interested in how well our algorithms correlate with the underlying property they're trying to measure, so we've been using the Matthews Correlation Coefficient (MCC) to evaluate how well they're performing. I considered using precision and recall, but these ignore all the negative results that are correctly rejected, which is the right approach for evaluating search results you're presenting to users, but isn't as useful as a correlation measurement for a binary classifier.

Example: Pictures of Plates = Foodies *- We run an algorithm that looks for plates or cups taking up most of the photo. It's fairly picky, with a precision of 0.78, but a recall of just 0.15, and an MCC of 0.32. If a lot of people are taking photos of their meals or coffee, we assume that there's something remarkable about what's being served, and that it's popular with foodies.*


Photos with cups or plates that the algorithm correctly spotted
Photos with cups or plates that the algorithm missed



Editor Note - Back to the interview!...

Q - What publications, websites, blogs, conferences and/or books are helpful to your work?

A - O'Reilly have been true pioneers in the data world, I recommend

following their blog at http://radar.oreilly.com, and the Strata conference has always been a blast.

Very interesting - look forward to following Jetpac's progress! Finally, it is advice time!...

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Don't listen to old farts like me. Figure out how we're all doing it wrong, and show us! I'm looking forward to being rendered obsolete by a whole new generation with tools and insights that leave us in the dust. We really have only scratched the surface in what we can do with all the data we're generating, so be ambitious and attack problems everyone else is ignoring as too hard.

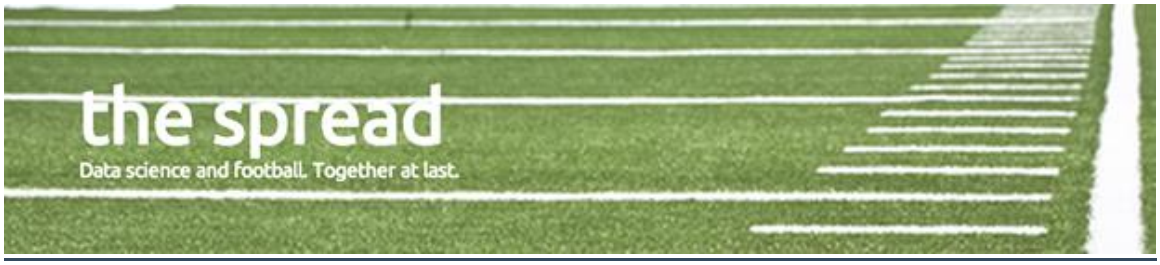
Pete - Thank you so much for your time! Really enjoyed learning more about Object Recognition and what you are building at Jetpac.
Jetpac can be found online at https://www.jetpac.com and Pete Warden @petewarden.

Data Science & Football

Trey Causey

Founder of *the spread*,
Data Scientist at zulily

Data Science A Football: Together at Last



We recently caught up with Trey Causey - Data Scientist at zulily and Founder of the spread - bringing Data Science and Football together at last...

Hi Trey, firstly thank you for the interview. Let's start with your background.

Q - What is your 30 second bio?
A - As you mentioned, I'm a Data Scientist at zulily, a site offering daily deals for moms, babies, and kids. I've spent most of my adult life as a quantitative and computational social scientist. I'm also a huge sports fan and really want to advance the state of sports analytics and statistics.

Q - How did you get interested in working with data?
A - It's hard to say, though I was hooked after my first statistics class as an undergrad, and I've always loved computers and hacking. My first computer was a Commodore VIC-20.

Q - Was there a specific "aha" moment when you realized the power of data?
A - Unsupervised learning techniques have always seemed kind of magical to me, whether a simple clustering algorithm or more

complicated methods like latent Dirichlet allocation. The idea that you can discover structure in a pile of data without telling the algorithm what you're looking for is pretty amazing. I think once I started working with more unstructured text I realized there was a whole other level of power available here

Trey, very interesting background. Thank you for sharing. Next, let's talk more about Football Analytics and what you are working on with the spread.

Q - Why are you excited about bringing Data Science and Football together?
A - A large part of Football Analytics is conducted by self-taught hobbyists - which is amazing and really makes the community lively and passionate. The down side is that I see a lot of wheel-reinventing and a lot of ad hoc, arbitrary decisions in work. It's not uncommon to see things like "I've included all players who started more than 10 games, have more than 4 years in the league, didn't miss time due to injury, and stayed with one team the entire time." This often amounts to selecting on the dependent variable and biases your results. I think a lot of people don't realize that many of the problems in sports analytics are just specific substantive examples of commonly occurring modeling problems. I'm hoping to change this.

Q - What are the biggest areas of opportunity / questions you want to tackle?
A - A big, low-hanging fruit that I see is the explicit incorporation of uncertainty into estimates of things like win probabilities. Data

Scientists encounter this problem all the time - we need to provide decision-makers with succinct, often single-number summaries that can be used to take action. But we also want to express how confident we are about those summaries and estimates.

Q - What project(s) are you working on at the moment?
A - Right now I'm working on two projects, one using ensemble models (random forests, gradient boosted classifiers, etc.) to build a win probability model and then building a Bayesian model of so-called 'field goal range' that gives us better estimates of kicking success.

Q - Tell us a little more about the spread - what are your goals for the site?
A - First and foremost, I want it to be fun for myself and for readers -- it's a hobby. Besides that, my goals are to a) improve the state of Football Analytics by offering a different perspective on some commonly explored questions and b) to teach some people some basic data science methods. Sports provide lots of great teaching cases for explaining the reasoning behind some common modeling problems. So, I hope it's educational and causes people to think.

Very interesting - look forward to learning from both those projects - and having some fun along the way! Let's talk briefly about your work at zulily and helpful resources...

Q - What does a typical day at zulily look like for you?
A - We're a fast-paced organization and my role covers a lot of different areas. I get the opportunity work on a diverse set of ever-changing

projects. A given day could range from tackling more traditional business statistical problems to sketching out the math behind an algorithm with the engineers to teaching seminars on statistics to non-experts in the company.

Q - What publications, websites, blogs, conferences and/or books do you read/attend that are helpful to your work?
A - On the data science side, the value of the connections I've made via Twitter really can't be understated. I've made professional connections, personal friends, and have an always-on network of frighteningly smart people who are always willing to help answer a question. I'd say that John Myles White and Drew Conway deserve special mention here. When I started getting to know them, they were both grad students in the social and behavioral sciences like myself. Their book, Machine Learning for Hackers, explains a lot of complicated topics in machine learning while being fun and conversational.

Interesting that Twitter has been proven such a valuable connector - good to keep in mind! Finally let's talk about the future and where you think your field is headed...

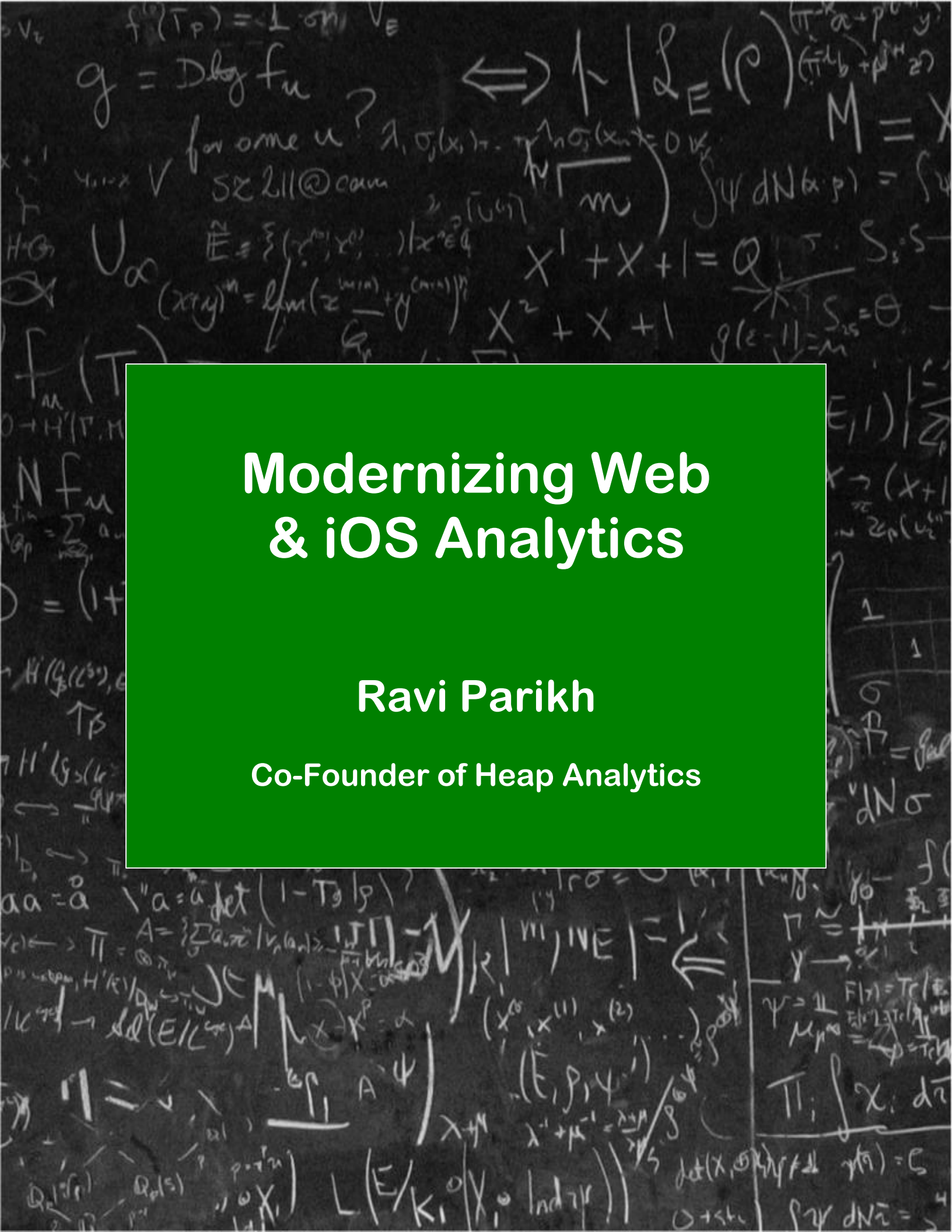
Q - What does the future of the spread and/or Football Analytics look like?
A - This is a great question. I don't know, but I hope it's a more transparent, peer-reviewed future with lots of collaboration. I'm a firm believe that we all improve when we make our methods transparent and open to critique. That being said, sport is a business with extremely high stakes and there's a tension there. I think that as analyses become more

complicated, the role of data visualization will become much more important in conveying lots of information in an easy-to-understand fashion. Have you seen the laminated play sheets that coaches have on the sidelines? They're not nicknamed "Denny's menus" for nothing.

Q - That's funny! ... Finally, how about any words of wisdom for Data Science students or practitioners starting out?
A - I'd say to pick a data set or sets you know really well and explore it like crazy. It's really helpful to be able to apply a new method to a dataset and have the ability to assess the face validity of your findings. It's fun to get counter-intuitive findings, but you should really stop and check your work if somehow you find that Ryan Leaf is actually a better quarterback than Peyton Manning. Examples that use uninteresting data (iris anyone?) are a lot less likely to result in you going the extra mile to learn more and exploring after the lesson is over.

I'd also say not to get too discouraged. This stuff is hard and it takes a lot of practice and a lot of willingness to make mistakes and be wrong before you get it right. And, if I had one single piece of advice -- take more matrix algebra.

Trey - Thank you so much for your time! Really enjoyed learning more about the convergence of Data Science and Football and what you are building at the spread. the spread can be found online at http://thespread.us and Trey Causey @TreyCausey.



Modernizing Web & iOS Analytics

Ravi Parikh

Co-Founder of Heap Analytics

Modernizing Web and iOS Analytics

We recently caught up with Ravi Parikh, Co-Founder of Heap, (YC W13) which is harnessing the power of Big Data to modernize web and iOS analytics...

Hi Ravi, firstly thank you for the interview. Let's start with your background...

Q - What is your 30 second bio?
A - I studied computer science at Stanford, where I did research with Professor Jeff Heer on data visualization. In 2012 I co-founded Heap, a user analytics company, and I've been working on that since. I also do quite a bit of data visualization work independently.

Q - What was the first data set you remember working with? What did you do with it?
A - When I was young I was really enjoyed filling out March Madness brackets. I loved poring over statistics and historical trends in order to "engineer" a perfect bracket. Ironically though the only bracket I ever filled out that did well was one where I didn't do any analysis and instead put my hometown team in the finals.

One of the lessons I learned from all the analysis I did was the

importance of avoiding "data dredging" - the practice of blindly mining data to find relationships. If you look long enough and hard enough at a large set of data you'll find plenty of seemingly interesting relationships that are just products of random chance. It's important to be disciplined and use methods like multiple hypothesis testing correction to avoid being misled.

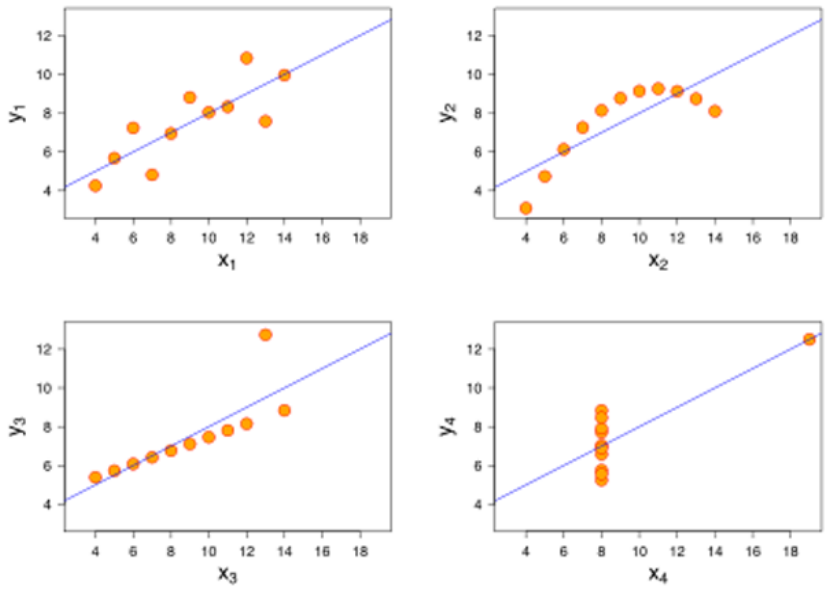
Q - Was there a specific "aha" moment when you realized the power of data?

A - For me that "aha" moment was when I learned about Anscombe's quartet. It's a group of four datasets each of which consist of several (x,y) pairs. Each of these datasets has the same mean of x, mean of y, variance of x, variance of y, x/y correlation, and the same linear regression line. Basically many of the "standard" summary statistics we might use to characterize these datasets are identical for all four. However, when visualized, each of the four datasets yield significantly different results. This was when I truly understood that asking deeper questions about data and visualizing data is incredibly important and powerful.

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)



Ravi, very interesting background and context - thank you for sharing! Next, let's talk more about what you are working on at Heap.

Q - What specific problem is Heap trying to solve? How would you describe it to someone who is not familiar with it?
A - Heap is web and iOS analytics tool that captures every user interaction on a website or mobile app: every click, form submission, pageview, tap, swipe, etc. Instead of having to write tracking code, Heap captures everything upfront and lets you analyze it later. When you want to answer a question with data, you can do it immediately, instead of writing code, deploying it, and waiting for metrics to trickle in.

Q - How did you come to found Heap?
A - Matin Movassate, my co-founder at Heap, had the initial idea. He used to work at Facebook as a product manager. To make any data-driven decision, he was forced to figure out what he wanted to track, ask a developer to write event tracking code, wait for the next product release cycle, wait for data to trickle in, and then finally have an answer. This is a process that could take weeks or months just to answer simple questions like "How many people are using the messages feature?" We decided to build Heap to eliminate that entire cycle.

Q - What have you been working on this year, and why/how is it interesting to you?
A - One of the coolest things I've built at Heap is the iOS tracking library, which automatically grabs touch and gesture events on mobile apps. Figuring out how to automatically capture event data from iOS

apps while taking into account performance and network overhead was a fun challenge.

Q - What has been the most surprising insight you have found?
A - We built Heap because we, as developers, were frustrated with the current state of the art in analytics. However we've found that our approach to tracking data without writing code has enabled product managers, marketers, and other non-technical folks to conduct end-to-end analysis on their data. We're looking forward to a future where anyone can be a data scientist.

Q - What technology are you using?
A - Our stack is Node + Redis + Postgres + Backbone + D3. Some things we're working on:

- Data capture. We're integrating with more clients and frameworks, including Android, AngularJS, and Backbone.js, all with virtually no performance overhead or integration cost.
- Real-time infrastructure. We support an expressive set of queries that allow our users to slice and dice the data in arbitrary ways. The results need to come back with sub-second latencies and reflect up-to-the-minute data.
- Data visualization. Simple pre-generated graphs just don't cut it. There's an enormous number of ways to organize the data. Existing tools only scratch the surface.

Q - What else should we know about Heap?
A - Heap is a small, engineering-focused company with a growing user base. We collect orders of magnitude more data than other analytics

products, and it's a complex technical problem to store and analyze that volume of data.

Editor Note - If you are interested in more detail on some of the neat properties of Heap's approach, this Quora discussion is very insightful. Here are a few highlights:

Automatically retroactive *- Heap captures all raw interactions since install time, so your analysis isn't constrained by events you remembered to log upfront.*

Super granular *- You can drill down into a cohort of users (or a specific user) and visualize their precise path through your app... You can define cohorts (without shipping code) as things like "users who added items to their shopping cart but never checked out".*

Untouched application code *- As the surface areas of your application increases, sprinkling tracking/logging calls across your app can be error-prone and difficult to manage. Heap entirely decouples analytics from development.*

Editor Note - Back to the interview!...

Finally let's talk a little about the YC experience, helpful resources and the future of Web/Mobile Analytics...

Q - How did you find the YC experience? What was most surprising?
A - YC was an amazing experience. There's not much I can say about it that hasn't already been said more eloquently by someone smarter than me, but I will reiterate that anyone in the early stages of building a technology company should consider it very strongly. I was most surprised by the incredibly high caliber of everyone else in my batch. If nothing else, YC puts you in proximity with other talented and unique people.

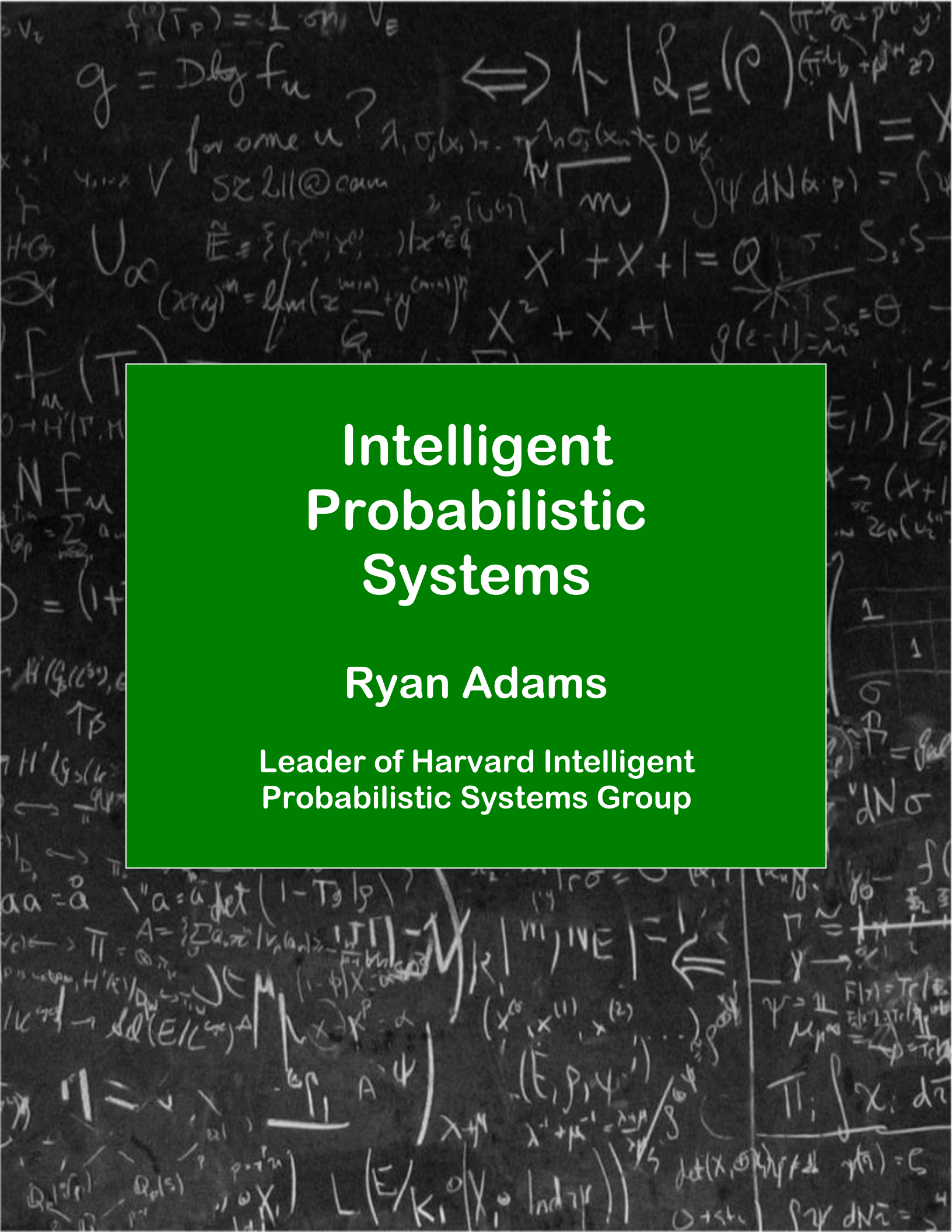
Q - What publications, websites, blogs, conferences and/or books do you read/attend that are helpful to your work?
A - I follow a number of blogs, websites, and people who are always teaching me new things about data science and visualization. The NYTimes graphics department is incredibly high-quality and staffed with some very impressive people. The Economist also puts out a daily chart which, while simple, are well done and insightful. Visualizing.org is a great website that hosts challenges for any visualization designer to hone their skills. One website that I'm looking forward to following once it's up and running again is Nate Silver's FiveThirtyEight.com, which is currently in the process of relaunching.

Q - What does the future of Web/Mobile Analytics look like?
A - We're moving towards a more integrated future. Currently the landscape is fragmented. A large, modern organization takes advantage of hundreds of disparate data sources, but the real power comes from integrating these and finding deeper insights that way.

Q - What is something a smallish number of people know about that you think will be huge in the future?
A - It's probably not fair to say a small number of people know about this, but I'm incredibly excited about the future of bioinformatics. The cost of genome sequencing and other technologies is dropping rapidly, and we're on the verge of an explosion in the amount of data that researchers will have access to.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Get your hands dirty. There's no faster way to learn than finding an interesting data set and playing around with it.

Ravi - Thank you so much for your time! Really enjoyed learning more about your background and what you are building at Heap.
Heap can be found online at https://heapanalytics.com and Ravi Parikh @ravisparikh.

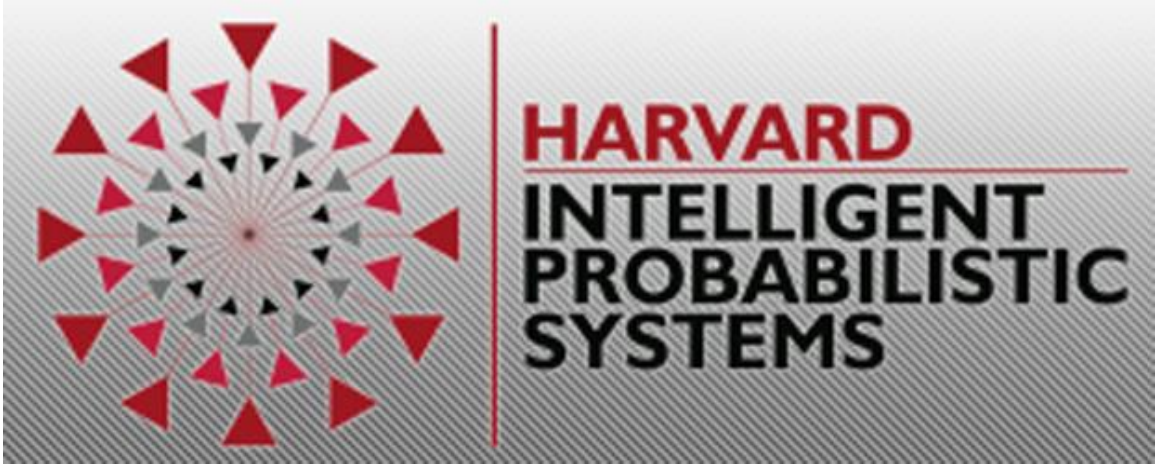


Intelligent Probabilistic Systems

Ryan Adams

Leader of Harvard Intelligent Probabilistic Systems Group

Intelligent Probabilistic Systems



We recently caught up with Ryan Adams - Assistant Professor of Computer Science at the Harvard School of Engineering and Applied Sciences and leader of the HIPS (Harvard Intelligent Probabilistic Systems) group - to learn more about the research underway at HIPS and his recent work putting powerful probabilistic reasoning algorithms in the hands of bioengineers...

Hi Ryan, firstly thank you for the interview. Let's start with your background.

Q - What is your 30 second bio?
A - I grew up in Texas, where my family has a ranch. I went to MIT for EECS and spent some time at NASA and in industry. I got my PhD in Physics at Cambridge University as a Gates Cambridge Scholar. I spent two years as a CIFAR Junior Research Fellow at the University of Toronto. I joined Harvard in the School of Engineering and Applied Sciences two and a half years ago as an assistant professor.

Q - How did you get interested in Machine Learning?
A - I was deeply interested in Artificial Intelligence, and as an undergrad received the excellent advice to work with Leslie Kaelbling, a premier researcher in the field and professor at MIT.

Q - What was the first data set you remember working with? What did you do with it?
A - As an undergrad, I spent some time doing financial modeling with neural networks.

Ryan, an impressive and interesting background - thank you for sharing. Next, let's talk more about Intelligent Probabilistic Systems and your work at HIPS.

Q - What excites you most about your work at HIPS?
A - I'm most excited about my fantastic students and collaborators, and the range of science that we can all do together. We pursue a lot of different research interests in my group. I'm excited about our new theoretical and methodological developments in areas such as Markov chain Monte Carlo, Bayesian optimization, Bayesian non-parametrics, and deep learning. I'm also excited about our collaborations in astronomy, chemistry, neuroscience, and genetics.

Q - What are the biggest areas of opportunity / questions you want to tackle?
A - There are several big questions that I'd like to make progress on in the near future:

- How do we scale up computations for Bayesian inference, to reason under uncertainty even when data sets become large?
- How do we perform optimizations over complicated structured discrete objects?
- How can we automatically discover hierarchical modularity in data?

Q - What is the most interesting model / tool / computational structure you have developed thus far?
A - Of the work we've done recently, our stuff on practical Bayesian optimization is the hottest, I think. We're actually working on a startup based on this technology as well. (Keep an eye on whetlab.com for developments.)

Q - What problem does it solve?
A - It optimizes difficult functions, but in particular, it can automatically tune other machine learning algorithms. It's had big success in tuning deep learning procedures without human intervention. Our open source software tool (called "Spearmint") has enabled non-experts to apply machine learning to novel domains.

Q - How does it work?
A - It uses relatively sophisticated Bayesian modeling and inference for Gaussian process function models to make recommendations on what function evaluations to try. The idea is to use information theory to make good decisions about optimization.

Q - What has been the most surprising insight it has generated?
A - It's a case where marginalizing over uncertainty in a probabilistic

model really gives a huge win. It turns out that humans are pretty bad at these problems in more than a couple of dimensions, but that machine learning algorithms can often do a great job.

Very interesting - look forward to hearing more about Whetlab in the near future! Let's talk about your recent work with Wyss Institute for Biologically Inspired Engineering, which has shown how AI algorithms could be implemented using chemical reactions...

Q - What question / problem were you trying to solve?
A - We were initially working on distributed inference algorithms for robotics, but we realized that chemical reactions mapped much better onto inference problems. We then focused on figuring out how chemical reaction networks could be used to implement the belief propagation algorithm.

Q - How is AI/Machine Learning helping?
A - It's not that AI/ML are helping, it's that in the longer term we're hoping these algorithms will be useful for synthetic biology.

Q - Got it, so what answers/insights did you uncover?
A - We showed that these important classes of computations can be performed without needing a digital computer. In particular, chemical reactions turn out to be a very natural substrate for graphical model computation.

Q - What are the next steps?
A - In addition to working with experimentalists to try to implement

these ideas in vitro, we have several theoretical directions we want to pursue. For example, these algorithms should lead to improved error correction in synthetic biology implementations.

Editor Note - If you are interested in more details on this research, Ryan's paper on Message Passing Inference with Chemical Reaction Networks is very insightful; and recent press coverage, such as this Phys.org report, provides a little more color ... Now, back to the interview!

Finally let's talk a little about helpful resources and where you think your field is headed...

Q - What publications, websites, blogs, conferences and/or books are helpful to your work?
A - The main ML publication venues that I read and contribute to are:

- Conferences: NIPS, ICML, UAI, AISTATS
- ML Journals: JMLR, IEEE TPAMI, Neural Computation, Machine Learning
- Stats Journals: JASA, Annals of Statistics, Journal of the Royal Statistical Society, Biometrika, Bayesian Analysis, Statistical Science, etc.
- Blogs: Andrew Gelman, Radford Neal, Larry Wasserman, Yisong Yue, John Langford, Paul Mineiro, Yaroslav Bulatov, Il

'Memming' Park, Hal Daume, Danny Tarlow, Christian Robert, and others I'm sure I've forgotten.

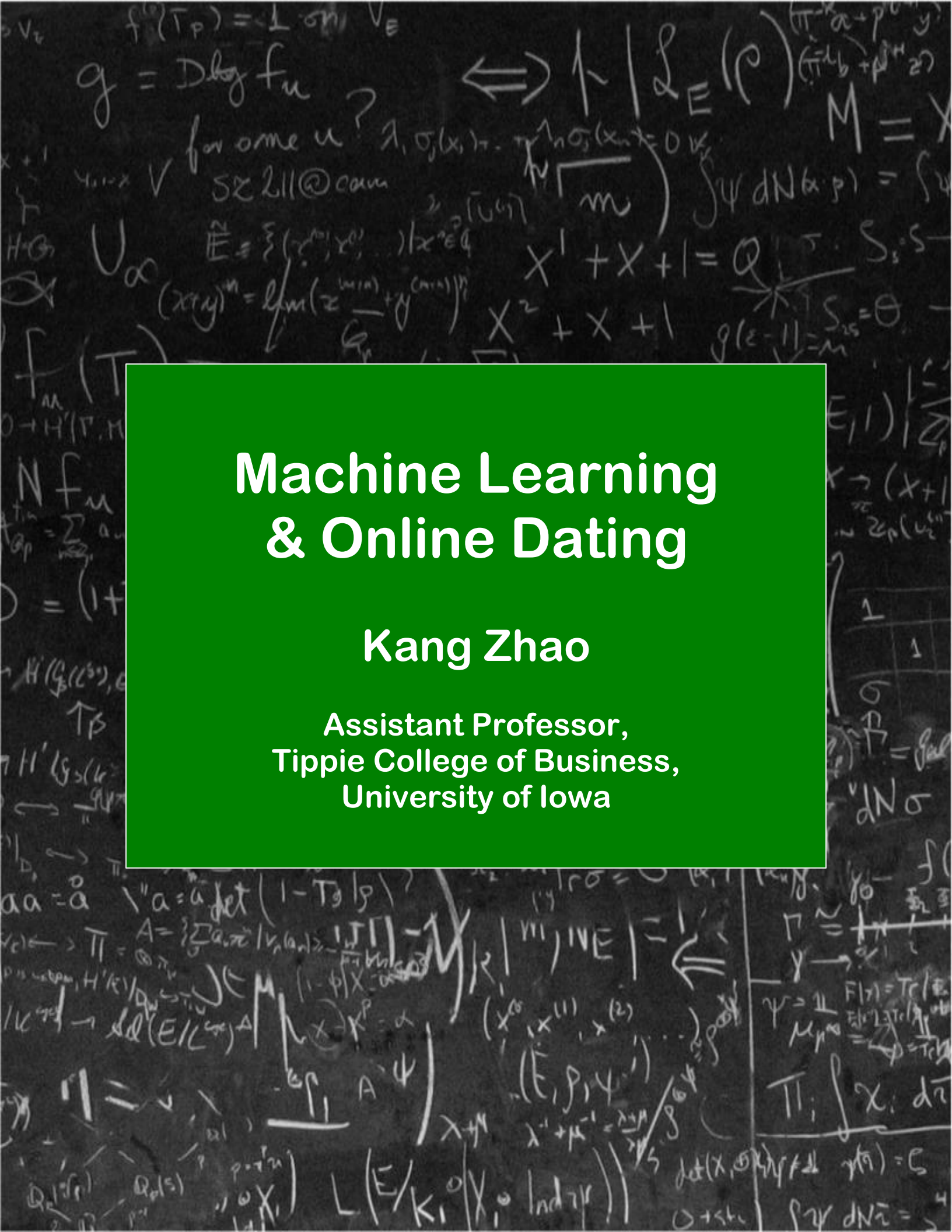
Q - What does the future of Machine Learning look like?
A - I think machine learning will continue to merge with statistics, as ML researchers come to appreciate the statistical approach to these problems, and statisticians realize that they need to have a greater focus on algorithms and computation.

Q - What is something a smallish number of people know about that you think will be huge in the future?
A - I think some of the recent work on the generalized method of moments (and tensor decomposition) is very interesting. I also think that the area of Bayesian optimization is going to get bigger, as we figure out how to tackle harder and harder problems. People are also beginning to understand better the behavior of approximate inference algorithms, which will become a bigger deal I expect.

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - Go deep. Learn all the math you can. Ignore artificial boundaries between institutions and disciplines. Work with the best people you can. Be wary of training in "data science" where you just learn to use other people's tools. To innovate, you have to learn how to build this stuff yourself.

Ryan - Thank you ever so much for your time! Really enjoyed learning more about the work you are doing at HIPS and where it could go next.

HIPS can be found online at http://hips.seas.harvard.edu and Ryan Adams at http://www.seas.harvard.edu/directory/rpa.

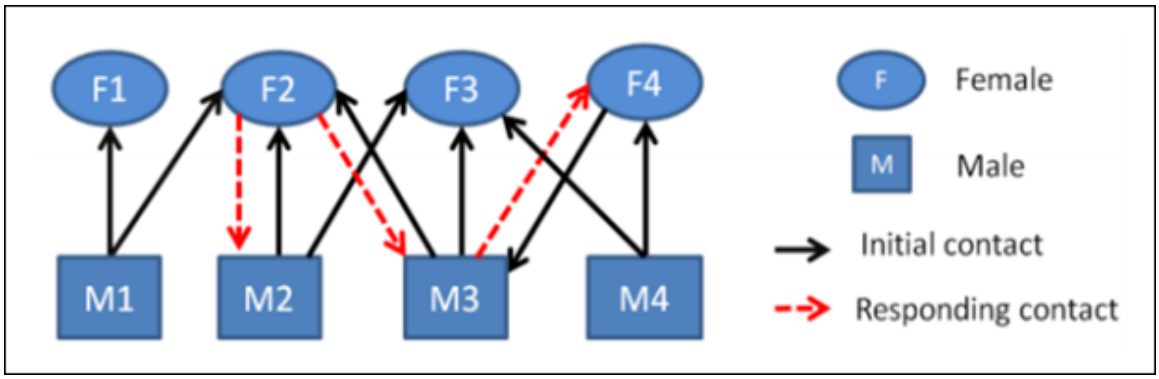


Machine Learning & Online Dating

Kang Zhao

Assistant Professor,
Tippie College of Business,
University of Iowa

How Machine Learning Can Transform Online Dating



We recently caught up with Kang Zhao, Assistant Professor at the Management Sciences department, Tippie College of Business, the University of Iowa. His work applying Machine Learning to the world of online dating has generated significant coverage (Forbes, MIT Technology Review, UPI, among others), so we wanted to know more!...

Hi Kang, firstly thank you for the interview. Let's start with your background...

Q - What is your 30 second bio?
A - As you mentioned, I am an Assistant Professor at the Management Sciences department, Tippie College of Business, University of Iowa. My research focuses on business analytics and social computing, especially in the context of social networks and social media. I also hold a PhD in Information Sciences and Technology from Penn State University.

Q - How did you get interested in Data Science / Machine Learning?
A - That dates back to my grad school days. I was involved in research projects that leveraged data from online social networks and social media. It is amazing that nowadays all the large-scale and distributed

interactions among people are available online thanks to the advances of online social networking/social media sites. Such data not only reveals who is talking to whom (i.e., helps us build a regular social network based on "knowing" or simple interaction), but also the time and the content of their online communication, which enable us build other social networks based on the nature of interactions (such as support network, information spread network). All these made me believe that the availability of such data will bring a brand new perspective to the study of people's social behaviors and interactions.

Q - What was the first data set you remember working with? What did you do with it?
A - My first research project using a real-world dataset was about collecting and analyzing data about humanitarian agencies and their networks. The scale of the data was actually "tiny" (several mega bytes) but the data did show us some interesting patterns on the topological similarities between different networks among these organizations (e.g., communication and collaboration networks), which inspired us to develop a simulations to model the co-evolution of multi-relational networks.

Kang, very interesting background and context - thank you for sharing! Next, let's talk more about Machine Learning in Social Networks and Social Media.

Q - What excites you most about bringing Machine Learning and Social Networks / Social Media together?
A - It is about the opportunity to do better prediction. With larger-scale

data from more sources on how people behave in a network context becoming available, there are a lot of opportunities to apply ML algorithms to discover patterns on how people behave and predict what will happen next. Such prediction can help to validate/test existing theories about people's social behaviors at an unprecedented scale. It is also possible to derive new social science theories from dynamic data through computational studies. Besides, the education component is also exciting as industry needs a workforce with data analytics skills. That's also why we, at the University of Iowa, have started a bachelor's program in Business Analytics and plan to roll out a Master's program in this area as well.

Q - What are the biggest areas of opportunity / questions you want to tackle?
A - I want to better understand and predict social networks dynamics at different scales. For example, dyadic link formation at the microscopic level, the flow of information and influence at the mesoscopic level, as well as how network topologies affect network performance at the macroscopic level.

Q - What Machine Learning methods have you found most helpful?
A - It really depends on the context and it is hard to find a silver bullet for all situations. I usually try several methods and settle with the one with the best performance.

Q - What are your favorite tools / applications to work with?
A - I use JUNG, a Java framework for graph analysis, Mallet for topic modeling, lingpipe for text analysis, and Weka for data mining jobs.

Q - What publications, websites, blogs, conferences and/or books are helpful to your work?
A - I usually keep an eye on journals such as IEEE Intelligent Systems, numerous IEEE and ACM Transactions, Decision Support Systems, among many others. As for conferences, I found the following helpful for my own research: ICWSM, WWW, KDD, and Workshop on Information Technologies and Systems. I also enjoy several conferences related to social computing, such as SocialCom and SBP.

Improving our ability to make predictions is definitely very compelling! Now, let's discuss how this applies in some of your research...

Q - Your recent work on developing a "Netflix style" algorithm for dating sites has received a lot of press coverage ... what question / problem were you trying to solve?
A - We try to address user recommendation for the unique situation of reciprocal and bipartite social networks (e.g., dating, job seeking). The idea is to recommend dating partners who a user will like and will like the user back. In other words, a recommended partner should match a user's taste, as well as attractiveness.

Q - How did Machine Learning help?
A - In short, we extended the classic collaborative filtering technique (commonly used in item recommendation for Amazon.com or Netflix) to accommodate the match of both taste and attractiveness.

Q - What answers / insights did you uncover?
A - People's behaviors in approaching and responding to others can provide valuable information about their taste, attractiveness, and unattractiveness. Our method can capture these characteristics in selecting dating partners and make better recommendations.

Editor Note - If you are interested in more detail behind the approach, both Forbes' recent article and a feature in the MIT Technology Review are very insightful. Here are a few highlights:

Recommendation Engine (from MIT Tech Review) *- These guys have built a recommendation engine that not only assesses your tastes but also measures your attractiveness. It then uses this information to recommend potential dates most likely to reply, should you initiate contact. The dating equivalent [of the Netflix model] is to analyze the partners you have chosen to send messages to, then to find other boys or girls with a similar taste and recommend potential dates that they've contacted but who you haven't. In other words, the recommendations are of the form: "boys who liked this girl also like these girls" and "girls who liked this boy also liked these boys".*

The problem with this approach is that it takes no account of your attractiveness. If the people you contact never reply, then these recommendations are of little use. So Zhao and co add another dimension to their recommendation engine. They also analyze the replies you receive and use this to evaluate your attractiveness (or unattractiveness). Obviously boys and girls who receive more replies

are more attractive. When it takes this into account, it can recommend potential dates who not only match your taste but ones who are more likely to think you attractive and therefore to reply. "The model considers a user's "taste" in picking others and "attractiveness" in being picked by others," they say.

Machine Learning (from Forbes) *- "Your actions reflect your taste and attractiveness in a way that could be more accurate than what you include in your profile," Zhao says. The research team's algorithm will eventually "learn" that while a man says he likes tall women, he keeps contacting short women, and will unilaterally change its dating recommendations to him without notice, much in the same way that Netflix's algorithm learns that you're really a closet drama devotee even though you claim to love action and sci-fi.*

"In our model, users with similar taste and (un) attractiveness will have higher similarity scores than those who only share common taste or attractiveness," Zhao says. "The model also considers the match of both taste and attractiveness when recommending dating partners" ... After the research team's algorithm is used, the reciprocation rate improves to about 44% - a better than 50% jump.

Finally, for more technical details, the full paper can be found here.

Editor Note - Back to the interview!...

Q - What are the next steps / where else could this be applied?
A - We want to further improve the method with different datasets from

either dating or other reciprocal and bipartite social networks, such as job seeking and college admission. How to effectively integrate users' personal profiles into recommendation to avoid cold start problems without hurting the method's generalizability is also an interesting question we want to address in future research.

That all sounds great - good luck with the next steps!... You are also working on other things - your work on sentiment influence in online social networks (developing a "Good Samaritan Index" for cancer survivor communities) has been well documented ... could you tell us a little more about this work?

Q - What question / problem were you trying to solve?
A - We tried to find who are the influential users in an OHC (Online Health Community). Here we directly measure one's influence, i.e., one's capabilities to alter others' sentiment in threaded discussions.

Q - How did Machine Learning help?
A - Sentiment analysis is the basis for our new metric. We developed a sentiment classifier (using Adaboost) specifically for OHCs among cancer survivors. We did not use off-the-shelf word list because sentiment analysis should be specific to the context. Some words may have different sentiment in this context than usual. For example, the word "positive" may be a bad thing for a cancer survivor if the diagnosis is positive. The accuracy rate of our classifier is close to 80%.

Q - What answers / insights did you uncover?
A - When finding influential users, the amount of contributions one has made matters, but how others react to one's contributions is also extremely valuable, because it is through such reactions inter-personal influence is reflected and thus measured.

Q - What are the next steps / where else could this be applied?
A - We would like to further investigate the nature of support in OHCs, so that we can build users' behavioral profiles and better design such communities to help their members.

Very interesting - look forward to following all of your different research paths in the future! Finally, it is advice time!...

Q - What does the future of Machine Learning look like?
A - This is a tough question. I don't know the exact answer but I guess ML will develop along two directions. The first would be on the algorithm side--better and more efficient algorithms for big data, as well as machine learning that mimics human intelligence at a deeper level. The second would be on the application side - how to make ML understandable and available to the general public? How to make ML algorithms as easy to use as MS Word and Excel?

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - I am not sure whether my words are of real wisdom, but I'd say for a beginner, it is certainly important to understand ML algorithms.

Meanwhile, it is equally important to develop the right mindset--a data scientist needs to be able to come up with interesting and important ideas/questions when given some data. In other words, one must learn how to answer the question-- "Now we have the data, what can we do with it?". This is very valuable in the era of big data.

Kang - Thank you so much for your time! Really enjoyed learning more about your research and its application to real-world problems.
Kang can be found online at his research home page and on twitter.

Future of Neural Networks & MLaaS

Dave Sullivan

Founder & CEO of Blackcloud BSG
- The company behind Ersatz

The Future of Neural Networks and MLaaS

We recently caught up with Dave Sullivan, Founder and CEO of Blackcloud BSG - the company behind Ersatz - and host of the San Francisco Neural Network Aficionados group. We were keen to learn more about his background, recent developments in Neural Networks/Deep Learning and how Machine Learning as a Service (MLaaS) is evolving...

Hi Dave, firstly thank you for the interview. Let's start with your background...

Q - What is your 30 second bio?
A - Sure, so I was born in '85 in the SF bay area. I got a chance to play with computers when I was 10 and by 12 had started learning to program - first with BASIC, then visual basic, then java, c/c++. Although I stuck with it over the years, I always viewed programming and technology in general as a hobby rather than a potential career opportunity. Because work is supposed to suck, right?

I graduated in '09, right in the middle of the Great Recession with a degree in history. I moved back to the bay area and started going to tech

meetups where, for the first time, I realized that there was a huge industry in tech and entrepreneurship was actually a valid adult career choice.

So in November 2010, I started Blackcloud BSG as a software consultancy, basically software development for hire. Deep learning is something that I've been working with for the past 3 years - it started as a hobby, then became somewhat of an obsession, and now it's a product. But I definitely took the unconventional route to get there...

Q - How did you get interested in working with data?
A - Well, I was researching poker bots. Not super seriously, it was just one of those random wikipedia searches, but it kind of introduced me to this whole world of machine learning, intelligent agents, etc. I kept reading and even though I'd have *tons* of *stuff* to learn before I could do *anything* with machine learning, it was enough to spark my interest and I was able to motivate myself to really study and develop these skills. And once you're working with it, once you start solving practical problems, you start to appreciate what the term "data science" really means. It's really about being able to visualize in your mind how data all interacts together, starting to think of data as its own entity with characteristics - it's about developing that intuition.

Q - What was the first data set you remember working with? What did you do with it?
A - Email data. I started learning about NLP with NLTK, a python NLP library. I wanted to cluster emails according to similarity, mostly by looking at the content. I started doing cool things like using it to surface

emails that required follow up, etc. That actually led to me learning about deep learning through some of the NLP research that was coming out back in ~2008 with word embeddings and all that.

Q - Was there a specific "aha" moment when you realized the power of data?
A - Honestly, not really. It's been more of a slow creep. I think people realize that data is going to be a big deal, that data is all around us, etc. etc. But even though everyone is saying it, I think most people don't quite understand how important it's going to be. You've got to think through all the ramifications of it, and the more I do, the more I become convinced "data" and what we do with it is going be as transformative to our society during the next 20 years as the Internet has been in the past 20. But it's taken me a while to come to that conclusion.

Dave, very interesting background and context - thank you for sharing! Next, let's talk more about Neural Networks...

Q - What have been some of the main advances that have fueled the "deep learning" renaissance in recent years?
A - Well, it started in 2006 when people started using "unsupervised pre-training" as a way to sort of "seed" a neural network with a better solution. Traditionally, one of the problems with neural networks had been that they were very difficult to train - particularly with lots of layers and/or lots of neurons. This was a way around that and it helped renew interest in a field that was all but dead (neural networks I mean).

So more research started coming out, there was a lot of looking into this

unsupervised pre-training idea, trying to figure out why it was working. In ~2008, 2009 people started using GPU implementations of neural networks and got massive performance boosts, as high as 40x in many cases. Suddenly that makes neural networks a lot more attractive - a 40x speedup makes something that used to take 40 days take 1 day - that's huge. So with that came much bigger models.

Since then, there's been a lot of really interesting new research. Google, Facebook, et al. have been getting into it, companies like mine are trying to build products around it - deep learning has come a long way in a short time and a lot of problems from the past have been revisited and solved. For instance, recurrent neural nets used to not be particularly practical - but now they hold state of the art in audio recognition and they are a really powerful tool for time series analysis. All of this is very recent though, just a few years old.

So now you have this situation where there's a ton of money getting pumped into this area, and thus there's a ton of people working on these problems, many more than there used to be. The models are pretty well defined at this point in the sense that much of the research is ready to be applied to industry. Meanwhile, the pace of new breakthroughs seems to be increasing (with "dropout" and word compositionality being two recent major developments)

Now, it could turn out that there's some other deal killer or gotcha with neural nets and they turn out to not be as useful as everyone thought (this would be the third time...) But personally, I think there's

something there, and I think it's the area where we're going to see the biggest machine learning breakthroughs in the near term. But at the same time, just because deep learning is gaining momentum doesn't mean that everything that came before it should be written off. Different tools should be used for different jobs...

Q - What are the main types of problems now being addressed in the Neural Network space?
A - The really big wins for deep learning have been in vision and audio problems. Those are the types of algorithms that are already in use by several really big companies. So you're seeing gains in just about any industry that could benefit from being able to interpret images and sounds better. But that whole process is really just getting started.

The next big area for deep learning is going to be natural language processing. Our demo (wordcloud.ersatz1.com) is a good example of some of the work that's being done there. Google in particular has been leading some really interesting research. For instance, it turns out that they can train models that can learn to organize data and words in such a way that different words become linearly combinable - like king + chair might be very close to the vector for throne. Everyone's sort of scratching their heads on why that happens/works, but the answer to that could be pretty interesting. If you solve vision, audio, and text, you've got a pretty robust set of "inputs" with which you can then start building more complex agents. You can start thinking about this idea of "higher level reasoning" and what that even means exactly. But even before all that, our sorta-smart devices are all going to get upgrades

thanks to deep learning and software is going to be making a lot more decisions that humans used to make.

Q - Who are the big thought leaders?
A - Haha, other than me? j/k... But there are basically 3 big guys in deep learning: Hinton (at google), LeCunn (at Facebook), and Bengio (I don't think anyone's snagged him yet?). But each of those guys have a lot of students, and those are really where the new ideas are going to come from. The big thought leaders - no one has really heard of them yet, but they're definitely there, they're the guys publishing at NIPS and a whole bunch of others that are self taught and tinkering with this stuff in their spare time in some yet unknown corner of the world.

Q - What excites you most about working with Neural Networks?
A - Well, in the near term, I think the most fundamental win from neural networks is this idea of automating the feature engineering process. I saw some cool research at NIPS this year that basically used these concepts to build a system like Pandora automatically (in a day, perhaps). But in order to do the same thing, Pandora spent years and probably a lot of money building a database of features - this was all feature engineering. You cut down on that part of the pipeline, and huge value is created.

In the longer term, I'm excited to be working with neural networks and machine learning more generally because, like I say, I really do think the impact on the world is going to be as important as the Internet has been. I mean, theoretically we could end up in a place where the idea of "work" as we know it just becomes relatively unnecessary - perhaps even

economically inefficient. That poses all kinds of really fundamental questions for society, just like the Internet has already started doing in a major way. And it's really cool to think about taking part in that conversation and really exciting to think about having an opportunity to shape it.

Q - What industries could benefit most from deploying Neural Network algorithms and techniques?
A - Any industry where the accuracy of their predictions can make a significant financial impact to their business. For a company like Netflix, increasing the accuracy of movie recommendations from what they were doing before by 10% might not be a huge deal. But for a company involved in any kind of algorithmic trading (be it options, commodities, or comic books), an extra 10% increase in the quality of certain decisions in their pipeline can make a really big difference to their bottom line. Oil exploration is another one that fits this. But those are the obvious ones - these kinds of techniques can also be applied to robotics (self driving cars, housekeeping robots, car driving robots), game design (Minecraft is algorithmically generated, so imagine something like that but way more original/complex/varied every time you play - and tailored to your unique gamer tastes), blogging (there will definitely be companies that crawl the internet and generate pretty readable articles with linkbait headlines with minimal human involvement), our phones (Siri will get better), there's a bunch more. Business X "with machine learning" will probably be a semi-valid business strategy soon enough. But really, every industry is going to benefit from better tools and an expanding pool of people that know how to use those tools.

Really compelling and inspiring stuff - thanks for all the insights! Now let's talk more about Ersatz...

Q - How did you come to found Ersatz?
A - Well, the 30 second bio kind of gives the broad strokes, but I really built Ersatz because I was frustrated by the existing tools available. I mean, you can download Pylearn2 (and we use Pylearn2 for certain pieces of ersatz, actually) and get started with it. But there's a lot of ground to cover in just getting something up and running. Then you also have to worry about the hardware component (GPU done right gives 40x speedups, which makes neural nets practical, so you want that). Then once you're training models, you want to learn things about them, but you're not really sure how. And that's to say nothing of the subtle kinds of bugs that can creep into this type of software - it's hard enough troubleshooting data issues, having to debug algorithmic issues too doesn't really help. This is all the kind of stuff we're trying to make easy with Ersatz. Making it so you can become a neural network practitioner instead of having to learn how to build them.

This process has played out with other product categories already - people used to build their own operating systems, databases, etc. Some people still do I guess, but there's a lot more that choose to buy software. And I think neural nets are kind of a good example of this, where many companies could benefit from them, and even more products can benefit from them. So people will have a database storing their data and a neural net back-end making it smarter. And we want Ersatz to be that neural net back-end.

Q - What specific problem does Ersatz solve? How would you describe it to someone not familiar with it?
A - Sure, so the hard part about machine learning is learning to think about your problem in terms of machine learning. Knowing how to frame your questions to get the answers you're looking for... So, assuming you're at least familiar with machine learning basics... Ersatz can do a few things: dimensionality reduction, data visualization, supervised learning, unsupervised learning, sample generation. We use neural networks to do all the heavy lifting on these tasks, and that's the part that we hide away from the user. Basically, if you provide the data and you understand how to ask the right questions, Ersatz takes care of all the learning algorithms, setting of parameters, the GPU hardware, etc. etc.

Q - Makes sense. Could you tell us a little more about the technology - how does it work, what models do you typically use - and why?
A - Sure, so basically, you've got 2 basic units: a worker and a job server. When you upload your data, it gets uploaded to S3. When you create a model, it creates a job for a worker. An available worker sees the job and pulls the data down from S3. These workers are GPU servers, and that's where the actual neural network stuff all happens. As it trains, the worker reports information back to the job servers which update statistics and dashboards in Ersatz. The stack is pretty much entirely Python, with a good bit of C/C++ in there too. And of course, quite a bit of JS on the frontend - we use D3js on our charts. Pretty standard fare really, we try to be relatively conservative about technology we use without going overboard.

In terms of models, well, we've got a few that we support, depending on the type of problem/data you have. We have standard deep nets (with different types of non-linearities, dropout, all the bells and whistles), autoencoders (for dimensionality reduction or feature learning), convolutional nets (for image problems), and recurrent nets (for time series problems).

Great, that makes it very straightforward to understand - thanks! Now, a couple of more operational questions...

Q - What has it been like boot-strapping the company throughout?
A - Really tough! And humbling, I think. You're kind of forced to learn to work with limited resources, which turns out to be a good skill to have. But it's also very frustrating, and often times it can feel like bootstrapping is slowing you down. Sometimes you actually can accelerate growth by throwing more money at a problem. The problem is, if you throw it at the wrong stuff, you just start losing money faster, and you lose momentum going down the wrong path. I think companies that raise money too early really do themselves a disservice. Once you do that, a clock starts ticking. In the beginning, being bootstrapped gives you a bit more time and flexibility to look at various options, try different ideas, and it also gets you in the mindset of needing to conserve resources. But I also don't believe in the "bootstrapped 4life!" mantra - I think that's just masochistic.

Q - You mention on your website that you manage an entirely remote group of developers (in 14 different countries!) - how do you make that work?
A - The actual number varies depending on what client projects we're

working on (right now it's 11 people in 8 countries, for instance). But in order to get it to work - there are a few things I've learned... First, I think you kind of either need to be remote or have an office - mixing the two isn't great and I think companies that have tried strapping on a remote team but had it not work out basically have this problem - everyone in the office communicates fine, they neglect communication with the remote team, and when things break as a result of that communication breakdown it's the fault of "remote work". So assuming you want to do remote... You've got to instill a culture of DIY - everyone can make, and is responsible for, their own decisions. Technically, you're required to be online on Skype for 3 hours a day usually, 8am PST - 11am. This is relatively loosely enforced, depending on what's going on. So you really have to be self-sufficient here sometimes.

Using Skype as your office is really cool because meetings can occur while you're not there and you can just read back what happened while you were out. Meetings happen asynchronously, which simply doesn't happen at an office. People put thought into their communications. Also, fewer interruptions - you can just sit and meditate on certain issues, not be interrupted. We're pretty good about using our various project management systems - that part is really important. It is nice to be able to hire anywhere in the world. I just think it requires a generally different management style, but it's a viable organizational model and it allows my team to get a lot done. But I won't lie, it introduces its own problems - it's hardly some holy grail that magically solves all your operational issues. And it's also still very new - even just 10 years ago, it would have been very difficult to start this company the way that I have.

But people live their lives on the Internet now, geography matters less and less and everyone anywhere in the world is just a click away.

Very interesting - look forward to hearing more about your and Ersatz' successes going forward! Finally, it is advice time!...

Q - What does the future of Neural Networks / Machine Learning look like?
A - That is the gazillion dollar question isn't it?

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - Don't be intimidated about getting into it! The basics aren't that complicated - with enough banging your head against the wall, anyone can get it. This is a field that is wide open - there is no "theory of relativity" for AI yet, but there probably will be, and I think it's actually pretty likely that we'll see that in our lifetimes. It's a really unique time in history right now, and this is a revolution that pretty much anyone in the world with an Internet connection can take part in. While many aspects of the worldwide economy are really messed up and will continue to be, I don't think there's ever been a time where economic mobility has been more decentralized. No matter where you are or who you are, you can take part if you're smart enough. So yeah, my advice: jump in, before it gets crowded!

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

about the work going on in neural networks and what you are building at Ersatz. Ersatz can be found online at http://www.ersatz1.com and Dave is on twitter @_DaveSullivan.

Big Data & Agriculture: Next Green Revolution

Wolfgang van Loeper

Founder & CEO of MySmartFarm

Big Data & Agriculture: The Next Green Revolution

We recently caught up with Wolfgang van Loeper, Founder and CEO of MySmartFarm. Once a wine farmer himself, he is now using Big Data to transform agriculture. We were keen to learn more about his background and what he is building at MySmartFarm (MSF) ...

Hi Wolfgang, firstly thank you for the interview. Let's start with your background...

Q - What is your 30 second bio?
A - I am South African born, though I finished my last three years of schooling in Germany and then went on to study business economics in Germany as well. Coming back to South Africa I started up a family wine business / farming operation, converted it to organic and won a few wine awards. This saw me using and recording high volumes of technical data over many years. As I was harvesting not only grapes but data (!), I began to transition from farmer to Big Data Entrepreneur.

Q - How did you get interested in working with data?
A - As a farmer you are forced to work with data. I respect the odd organic small holder farmer who doesn't work with data, but for myself, it was the organic farming operational requirements that really got me into collecting farming data. I couldn't imagine doing without. All I now

do with the development of MySmartFarm, is to make this whole 'data thing' much easier and quicker for all other farming colleagues.

Q - What was the first data set you remember working with? What did you do with it?
A - We used to receive pages and pages of faxes from the labs, with the analyses of all our soil and leaf samples. Being the structured person I am, I re-typed all the data sets into excel sheets. This helped me understand and structure the data better. We then used my analysis/insights to balance our soils and fertilize appropriately. Very soon thereafter we also started working with soil moisture data, hosted in an independent software package, to further refine our farming techniques.

Q - Was there a specific "aha" moment when you realized the power of data?
A - We were preparing for last pre-harvest irrigations in our vineyards and observed a coming heat wave in combination with already high levels of plant stress. With harvest being eminent, we were looking at obtaining optimum phenolic constitutions within the grapes. It was a major manual exercise combining and understanding all the data - not made for the everyday farmer - but I am glad we did and we were very happy with the results. And I, for one, realized the power of combining all that data.

Wolfgang, very interesting background and context - thank you for sharing! Next, let's talk more about Data Science and Agriculture...

Q - What excites you most about bringing Big Data and Agriculture together?
A - The opportunity to create a move in farming practices - literally being part of a new Green Revolution, which, unlike the last one, potentially has the ability to fix the problems the previous one has bought us.

Q - Which areas of agriculture do you think will be most impacted by Data Science?
A - The application of chemicals and water and the effective use of natural fertilization.

Q - What are the biggest areas of opportunity / questions you would like to tackle?
A - Creating an environment for farmers to farm crops more in tune with nature. And making more use of nature's tricks, to harvest crops that cost less to produce and contain more of the all-important natural phytonutrients, which conventional, heavily chemically treated, farm produce has very little of.

Q - What was your reaction to Monsanto acquiring Climate Corporation last year? How does that deal change the Data Science-Agriculture landscape?
A - Although the Climate Corporation does not cover as many varied sources of data as MySmartFarm (MSF) does, it was inevitable that one

of the big players was going to move into agricultural data science. But only recent developments in SaaS systems make it possible to collect data directly from the farmer. So it would have been hard to develop a MSF-like system any earlier, or similarly, for an acquisition of such a scope to be possible. Subsequent to Climate Corporation's acquisition, Du Pont and Deere have also partnered to drive their own move into agricultural data science. Sitting in the middle of this development, I'm first glad to be part of it and second, to have my patents in place!

Definitely sounds like an exciting time to be developing technology in this space! On that note, let's talk more about MySmartFarm...

Q - How did you come to found MySmartFarm?
A - After years of using and fine-tuning my excel sheets - where I manually collected/entered/analyzed all the data - farmers, agronomists and scientists said I should think about solving the problem in a way that I could commercialize it so other farmers could benefit. So creating a SaaS Cloud based platform that automatically collected all the data just made so much sense.

Q - Got it. So what specific problem does MySmartFarm solve? How would you describe it to someone not familiar with it?
A - With MySmartFarm a farmer has all his data (harvesting, fertilization, laboratories, weather, disease and sensor data - such as from local soil or leaf moisture and satellite sensors) alongside his important mapping and GIS data. MSF then combines all that data with climate data and from there generates new intelligence. Added to the

secure storage of a farmer's complete set of data, he has the added benefit, by getting a very convenient management dashboard, illustrating what is important to him to make fast and efficient decisions.

Q - Could you tell us a little more about the technology?
A - MSF makes use of a whole host of available high tech services to farmers - basically farmers make use of sensors and laboratory information from these service providers and MSF collects the data from them all onto one platform. All the data that we collect for the farmer on our SaaS systems is hugely valuable in that we can combine that data with forecasted data and help the farmer act on predictions or tendencies we pick up over the years. With these insights the farmer can act in a much more timely manner on an enormous set of parameters, which otherwise would be impossible. In terms of technology stack, IBM is supporting the development of MySmartFarm and we're using business intelligence stacks from their portfolio, this saves us a lot of costs and development money.

Q - What have you been working on most recently?
A - We're busy with beta testing and developing the first version. In the last four weeks we've been busy with the dashboard, for me a very important aspect, as it has very unique patented features that make it literally visually fun for the farmer to interact with all his data - and he will be able to specifically select what is important for him. The feedback to-date from farmers is that they like being able to interact with the data from multiple sources on one dashboard, without being required to change to different platforms or software.

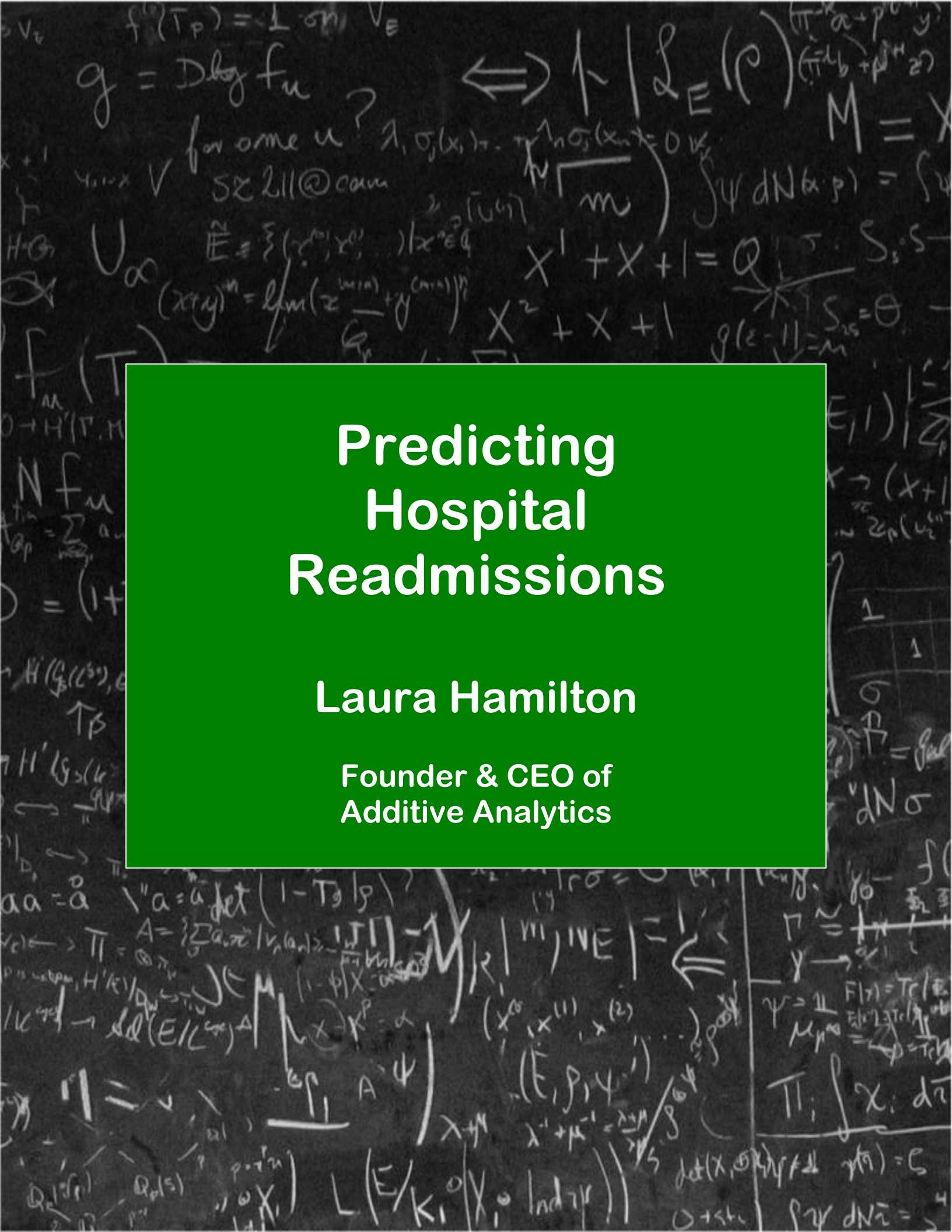
Q - What else should we know about MySmartFarm?
A - MSF will drive farmers to more sustainable farming practices, not only saving water and chemicals, but assisting them on the move to more agro-ecological practices through knowledge transfer of successful and profitable, more ecological practices; especially if it is linked to high tech and data.

Very interesting - look forward to hearing more about MySmartFarm going forward! Finally, it is advice time!...

Q - What does the future of Data Science & Agriculture look like?
A - Rosy!

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Studies and data collection are nice but of limited use without the wisdom of practical solutions that actually help people achieve new ways of doing things, such that we can still live on this planet in a 100 years time.

Wolfgang - Thank you so much for your time! Really enjoyed learning more about the evolving Data Science - Agriculture landscape and what you are building at MySmartFarm. MySmartFarm can be found online at http://mysmartfarm.info and on twitter @MySmartFarm.



Predicting Hospital Readmissions

Laura Hamilton

Founder & CEO of
Additive Analytics

Predicting Hospital Readmissions

We recently caught up with Laura Hamilton, Founder and CEO of Additive Analytics. We were keen to learn more about the evolution of the health analytics space, how data science / machine learning is helping, and what she is building at Additive Analytics ...

Hi Laura, firstly thank you for the interview. Let's start with your background...

Q - What is your 30 second bio?
A - I graduated from the University of Chicago with a mathematics degree. From there I joined Enova International, where I launched 3 businesses within a 3 year span; 2 of those were profitable within 18 months of launch. After Enova, I joined ecommerce startup BayRu as Head of Analytics. At BayRu, I built a proprietary analytics engine that compares the company's performance to benchmarks. In September 2013, I launched a healthtech startup called Additive Analytics. We provide analytics for hospitals.

Q - How did you get interested in working with data? Was there a specific "aha" moment when you realized the power of data?
A - It was during my econometrics class at the University of Chicago. For the final project, we did a linear regression on some labor data using Stata. I just liked the idea that I could get real, actionable, objective results with a computer program and a few commands. Then at Enova, I brought on an additional data vendor, which improved our underwriting and reduced our default losses. It's often more effective (and easier) to go out and obtain additional data sources than it is to keep trying to make your algorithm more sophisticated!

Laura, interesting background and context - thank you for sharing! Next, let's first talk about Healthcare Analytics..

Q - What have been some of the main advances that have fueled the rise of Healthcare Analytics in recent years?
A - We've seen a dramatic increase in the number of providers using Electronic Health Records (EHRs) in recent years as a result of government incentives. There's all this clinical data sitting in machine-readable form right now. It used to be all paper charts in boxes in the basement. Now that there's all this data, people are really excited to harness the data and use it to provide better care at a lower cost. In the future, I think that we will see a lot more analytics geared towards patient engagement. I am really excited about the potential of Blue Button+, which is an initiative by the federal government to enable patients to view their own personal health data online or via mobile devices.

Q - What are the main types of problems now being addressed in the Healthcare Analytics space?
A - One of the key priorities of the Centers for Medicare and Medicaid Services (CMS) is moving from a fee-for-service payment model to a pay-per-episode-of-care model. Currently, 60 hospitals are participating in CMS' advanced bundled payments model. These hospitals need to start taking full financial risk by October 2014. Under the new model, a participating hospital will receive a flat payment up front (depending on the patient's condition). The hospital will not receive any additional payments if the patient is readmitted within 30 days of discharge. That type of financial risk is new to hospitals. Hospitals are used to fee-for-service. As a result, there's a lot of demand for analytics solutions that will reduce 30-day readmissions. For example, the Henry Ford Innovation Institute has issued a challenge to find innovative, technology-driven solutions to reduce 30-day hospital readmissions.

Q - Who are the big thought leaders?
A - There are several...

1. Healint is using data from cell phone sensors to identify neurological conditions such as stroke.
2. AliveCor built a $199 ECG sensor that attaches to a cell phone.
3. Google is using search data to uncover flu trends/li>
4. 23andme analyzes patients' DNA to identify genetic disease risks.
5. IBM researchers have found a way to extract heart failure diagnostic criteria from free-text physicians notes.

6. Additive Analytics predicts which patients are likely to get readmitted to the hospital within 30 days of discharge and gives providers actionable steps to take to reduce readmissions.
7. AllScripts is building an ecosystem of technology and analytics apps that integrate with its electronic medical record system.

Q - What excites you most about bringing Healthcare and Machine learning together?
A - Healthcare in the United States is so broken right now. The United States spends $2.8 trillion per year on healthcare. That's $8,915 per person. The United States spends 17.2% of its GDP on healthcare - far more than its peers spend. Take a look at this graph:

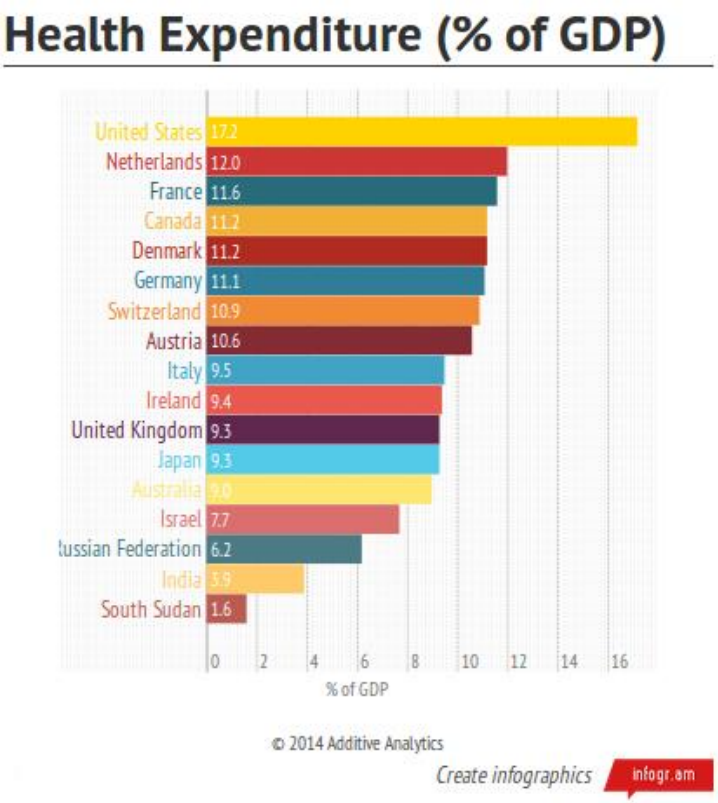


Image Credit: Additive Analytics

By using the right analytics, we can reduce costs while increasing the quality of care.

Q - What are the biggest areas of opportunity / questions you would like to tackle?
A - One of the key areas of focus for us is reducing 30-day hospital readmissions. Sixty hospitals have joined CMS' most advanced payments model. They've agreed to receive a single payment covering the whole episode of care, including all hospital readmissions up to 30 days after discharge. We're offering a solution to enable hospitals to understand their data and reduce their readmissions.

Definitely sounds like an exciting time to be developing technology in this space! On that note, let's talk more about Additive Analytics...

Q - How did you come to found Additive Analytics?
A - My background is in technology and analytics, and I wanted a way to leverage that. And I think now is the right time to be working on technology and analytics for healthcare. We are just now getting electronic access to patients' medical records. Also, there are lots of payment changes coming in the near future as a result of the Affordable Care Act. With those two changes happening right now, I think there is a ton of opportunity in the healthtech and healthcare analytics space.

Q - Got it. So what specific problem does Additive Analytics solve? How would you describe it to someone not familiar with it?

A - We provide analytics for hospitals. In the past few years, most healthcare providers have moved from paper charts to electronic charts. Now there is a huge amount of clinical data. Additive Analytics takes that clinical data and generates useful insights from it. For example, our model can identify what patients are likely to get readmitted to the hospital within 30 days of discharge. We suggest actionable steps that providers can take to reduce 30-day hospital readmissions. By reducing readmissions, we can save money and save lives.

Q - Could you tell us a little more about the technology? Firstly, how does it work?
A - Our software integrates directly with hospitals Electronic Medical Record (EMR) systems. Our software takes clinical data from the EMR and runs various analyses on it. From there, we generate intuitive graphs that give hospitals insights into their performance, both on clinical measures as well as financial measures. We also provide concrete steps that hospitals can take in order to improve their quality measures. For example, if a hospital wants to reduce 30-day hospital readmissions, we can provide specific steps on a per-patient basis that the hospital can take to prevent excess readmissions.

Q - What tools / applications are you using?
A - I like to use Octave, Python, and Vowpal-Wabbit. Sometimes I find it's helpful to do some initial summary and graphing with Excel. The Additive Analytics web application is built with Ruby on Rails. It sits on top of a Postgres database. For data visualization, I like D3 and DataTables. If I need a quick chart for the Additive Analytics blog, sometimes I will use Infogr.am.

Q - How is Machine Learning helping?
A - Machine learning helps us make sense of the huge amounts of clinical data in hospitals' EMRs. For example, we can use Natural Language Processing to extract meaning from free-text physicians' notes. Also, we can use techniques such as logistic regression and neural networks to predict which patients are likely to get readmitted to the hospital within 30 days.

Q - What is the most surprising insight you have found?
A - Simply giving patients a phone call after they are discharged from the hospital reduces the risk of 60-day readmission by 22%. It's so simple, but it's so powerful.

Q - What is your favorite example of how Additive Analytics is having real-world impact?
A - Last week we launched an online tool that expectant parents can use to compare maternity wards at different hospitals. It was written up in TechCrunch. Our goal is to give patients tools to evaluate providers based on objective, quantitative quality metrics. We hope to provide increased transparency to hospitals' performance. Currently, you can go on Yelp and find the best restaurant or you can go on Angie's List and find the best plumber. You can go to US News & World Report to find the best college. We think you should be able to go on the Internet and find the best healthcare provider, too.

Q - What advances could your approach / technology enable going forward?
A - We can run analytics on clinical EMR data to figure out which treatments are working better than others. For example, we could

analyze the outcomes of patients who were treated with proton therapy via a $150 million cyclotron. We could compare how those patients fared versus patients treated with traditional (much cheaper) methods. Perhaps we would find that proton therapy didn't improve outcomes at all; that could provide significant cost savings to hospitals as well as payers.

Very interesting - look forward to hearing more about Additive Analytics going forward! Finally, it is advice time!...

Q - What does the future of Healthcare & Machine Learning look like?
A - A few thoughts ...

1. Researchers have found a way to extract Framingham heart failure diagnosis criteria from free-text physicians' notes using Natural Language Processing. In the future, I think we'll see many more applications of Natural Language Processing for diagnosis, for anomaly detection, and for billing.
2. I think that we're going to see a much tighter integration of the clinical and financial sides of things in the future. It will be much easier for physicians and hospital administrators to understand which treatments are the most cost effective.
3. We have a huge number of data sources now. I have a Withings scale. My husband has a Fitbit. We've got all these new sources of data from wearables and even from our cell phones. In the future, I think a lot more of the data will be connected. Your physician

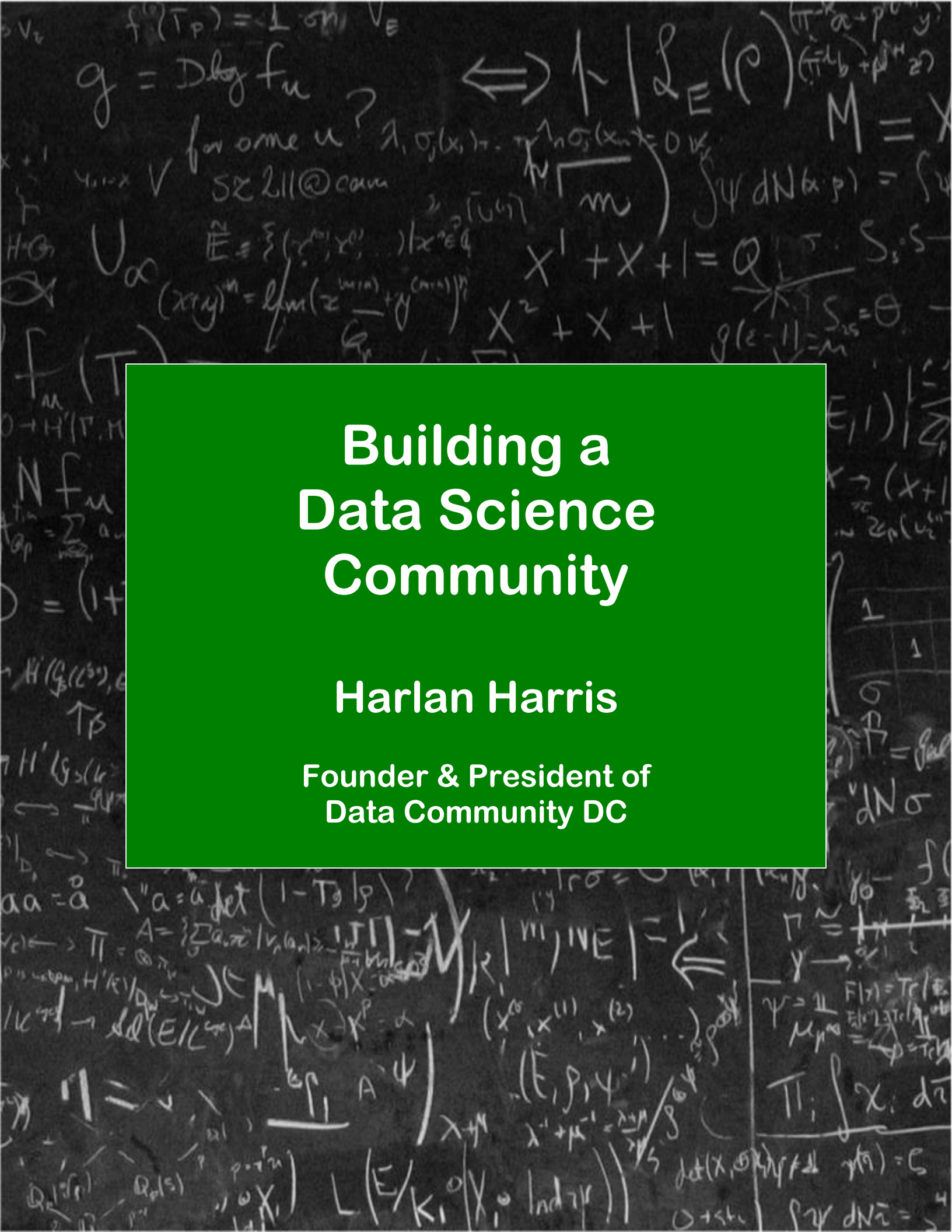
will be able to see a chart of your Withings data, your Fitbit data, your cell phone data.

4. I think that we're going to get better at finding adverse drug events. Now that we have electronic medical records for millions of patients, we can mine that data to find drug interactions and problematic side effects—even ones that only affect a small subset of patients. Problems such as those with Vioxx and Thalidomide will be found more quickly, and fewer patients will be affected.
5. We're going to have a better understanding of disease transmission. Already, we can use Google search terms to understand flu trends. If we combine social media data with electronic medical record data and perform aggregate analyses, we can predict epidemics and take steps to halt disease transmission in its tracks.

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - 5 things:

1. Take Dr. Andrew Ng's Machine Learning course on Coursera.
2. Take Dr. Abu-Mostafa's Learning from Data course on edX
3. Get as many features as you can. Think about where you can get additional data, and add as many new data sources as you can.
4. Data visualization is as important as the model. You can have the most sophisticated model in the world, but if you don't present it in a way that's intuitive to the user it will be useless. All analyses should be actionable
5. Beware overfitting!

Laura - Thank you so much for your time! Really enjoyed learning more about the evolving Health-tech landscape and what you are building at Additive Analytics. Additive Analytics can be found online at http://www.additiveanalytics.com and Laura on twitter @LauraDHamilton



Building a Data Science Community

Harlan Harris

Founder & President of
Data Community DC

Building a Data Science Community

We recently caught up with Harlan Harris, Co-Founder and current President of Data Community DC (DC2). After multiple years (and degrees!) in academia he transitioned to industry as a Data Scientist in 2009. We were keen to learn more about his background, the vision for DC2 and his views on how Data Science is evolving ...

Hi Harlan, firstly thank you for the interview. Let's start with your background...

Q - What is your 30 second bio?
A - I'm from Madison, WI. Undergrad from UW-Madison in Computer Science with most of a second major in Linguistics. Grad school at Illinois - Urbana/Champaign, where I wrote a dissertation in Computer Science (Machine Learning), while doing a lot of Psycholinguistic and Cognitive Science coursework and research on the side. Cognitive Psychology post-docs at Columbia University/UConn and at NYU. I was good at the pieces, but not at the whole, and couldn't make an academic career work. Switched to industry as a data scientist in 2009, first at Kaplan Test Prep and now at Sentrana, Inc in Washington, DC. I was pretty involved in the professional data Meetup scene in NYC before I

moved, and even more so here in DC. ... Also: married, foodie, lapsed fencer.

Q - How did you get interested in working with data?
A - As an undergrad I took several AI classes, including a Machine Learning class taught by Jude Shavlik. That was my favorite undergrad class in my major, and led me to continue with ML as a graduate student. I then got further hooked by the statistics side of things when I started doing psychological research.

Q - What are you favorite tools / applications to work with at home &/or at the office?
A - I almost entirely use R and Julia. I was an initial developer of some of the statistical data representation and manipulation functions for Julia, but I haven't had much time to work on that recently, and people much smarter than me have taken the reins. I also do a little programming in Javascript, but mostly side projects - nothing fancy.

Q - What was the first data set you remember working with? What did you do with it?
A - Ever? I think there was a 4th grade science project that involved talking to plants. I'm pretty sure I proved (with n=3) that houseplants do better when you yell at them.

That's a pretty powerful learning at such a young age :) Thanks for sharing your background. Let's talk more about Data Science and how the landscape is evolving...

Q - What excites you most about recent developments in Data Science?
A - I'm fascinated by the idea that we're watching a new kind of professionalization of a discipline. Existing academic and professional boundaries are being redrawn. But unlike the creation of new disciplines in the last century, such as the formation of Computer Science out of Electrical Engineering and Mathematics, we now have a bottom-up, peer-driven community, supported by on-line tools such as Meetup and StackOverflow. Being part of a professional society seems less important now than ever before. But having visibility and credibility - and of course skills - are as important as ever. (See my recent Ignite talk about this...).

On a technical level, it's interesting that Neural Nets are back in fashion in the form of Deep Learning. I'm also interested to see what happens with probabilistic programming and the maturation of Monte Carlo modeling techniques.

Q - What industries do you think will benefit most?
A - Basically anything with repeated processes, lots of data exhaust, and a well-defined success criterion. The relative cheapness of data science techniques these days means that stuff that used to be limited to just governments and enormous businesses can be applied by small teams to things like healthcare analytics and journalism, which is drastically changing those fields. On the other hand, there are a lot of really interesting domains where there's no relevant data, or where you can't usefully define success, or where every situation is basically unique. For example, you can't use predictive analytics to tell you how to write a healthcare law.

Q - What are the biggest areas of opportunity / questions you would like to tackle?
A - Drew Conway famously put Domain Knowledge as a key part of the Data Science Venn Diagram. I'm interested to see whether simple AI systems that have simple domain knowledge capabilities can supplement the statistical tools in a useful way in a broader set of applications. Right now, the domain knowledge is in our heads - is it possible to extract just enough domain knowledge into software so that more people can more efficiently focus on the questions rather than the tools? IBM's Watson is one approach to this, but I think there will be a lot more systems that try different approaches in coming years.

Very thought-provoking - that would definitely transform a lot of professions! Let's change gears and talk more about your involvement in the Data Science Community...

Q - How did you come to found / organize Data Science DC??
A - I was an occasional presenter at the R and other Meetups in NYC before I moved in 2011. When I came to DC, there was an R Meetup, run by Marck Vaisman, but nothing else. Along with a data scientist at WaPo Labs, Matt Bryan, we formed Data Science DC that summer. It was a bit ahead of the times to call the Meetup "Data Science" - everything else was Predictive Analytics or Machine Learning or something. The Meetup's been very successful, and in 2012, we decided we wanted the capability to do bigger and better things, so, along with several others, we created an umbrella organization called Data Community DC, or DC2. DC2 now has six Meetup groups with over 5000 unique members,

a board of 12 people, a blog, occasional workshops, and plans for bigger events in the future. I'm the current President of DC2.

Q - What are the primary goals of the organization?
A - Here's DC2's current mission statement: *Data Community DC is an organization committed to connecting and promoting the work of data professionals in the National Capital Region by fostering education, opportunity, and professional development through high-quality, community-driven events, resources, products and services.*

Within DC2, Data Science DC, which I'm still the primary organizer of, focuses on the "algorithmic" or problem-solving level. Basically, we want to give people an opportunity to share what they're working on and what approaches they're excited about and to meet other people in their professional community, even those who work on wildly different problems and domains.

Q - What have been 3 of the most memorable Meetup presentations?
A - Wow, DSDC alone has had 30 events... Let's see, of those... I really liked the Recommendation Systems event, where two great presenters, from WaPo Labs and LivingSocial, talked about real-life applications of the technology. We had a presentation by a team at the Sunlight Foundation that involved everything from problem formulation to data collection to graph analysis to data visualization. Another great one was a panel discussion about Data Science in political campaigns - entertaining and fascinating. In all three cases, our presenters had real problems, in retail, or journalism, or marketing, and used a very wide variety of tools and techniques to do things that would have been flat

impossible, or taken orders of magnitude more resources, just a decade ago. It's really inspiring... The other DC2 Meetup groups have all had amazing events too!

Q - What has been the most surprising insight / learning from organizing the group?
A - Hhmmm. One thing is that almost everybody who gathers up the courage to give a presentation to scores or hundreds of their peers knocks it out of the park. It gives me amazing faith in humanity that everyone seems to be so good at their jobs!

Q - What advice would you give to others looking to organize a Data Science group/Meetup in their own city?
A - Get sponsorship, and minimize support from your employer. Astroturf Meetups don't last. But there are many, many great companies that would love to chip in some money for potential customers and employees to get pizza and soda before presentations. Don't be afraid to ask individual people who you think do interesting work to speak - most will, and do a great job. Steal ideas from Meetups in NYC. :)

Makes sense :) Harlan, what you have managed to build for the Data Science community in DC is really impressive - look forward to hearing more about the various groups going forward! Finally, it is advice time...

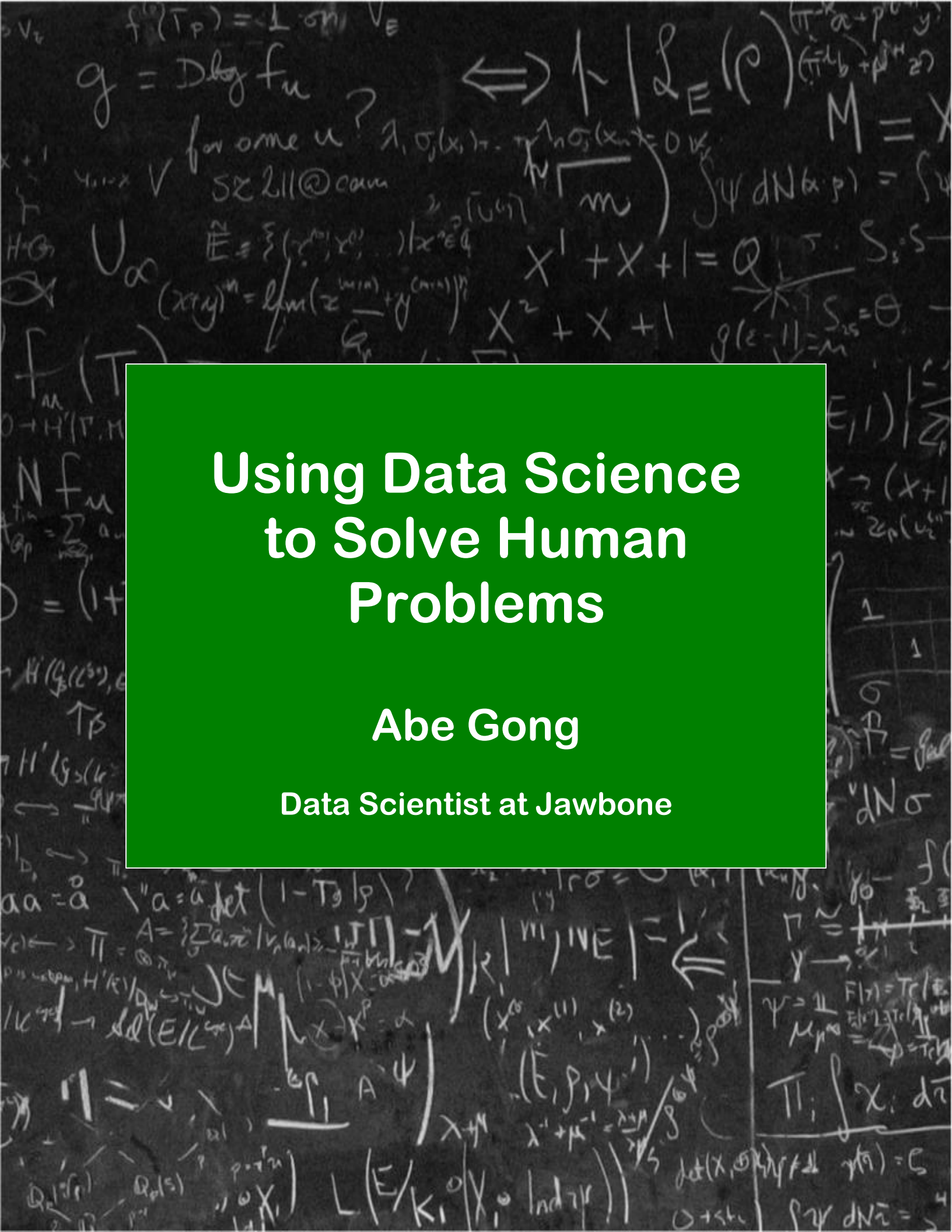
Q - What does the future of Data Science look like?
A - There will be people coming out of academic programs with Masters degrees in Data Science very soon. It'll be very interesting to see how

those people interact with people who pivoted professionally. There'll be more certification and more coherence in terms of what people know and are expected to be able to do.

I suspect techniques for Big Data analysis will continue to be important, but perhaps relatively less so over time as those tools mature. Medium Data, where you have to think about the scale of the problem to solve it, but where you can move the data around without too much problem, will be where most of the action is. ... I'm also personally interested in the impact of Open Data and Civic Analytics on people's lives around the world.

Q - Any words of wisdom for Data Science students or practitioners starting out?
A - Get involved in your professional community, whether it's attending Meetups (and meeting people at the bar afterwards), or answering questions on StackOverflow or CrossValidated, or trying your hand at a Kaggle competition or a hackathon. Learn about the many different points of view of people doing work related to your interests.

Harlan - Thank you so much for your time! Really enjoyed learning more about the evolving Data Science landscape and what you are building at Data Community DC. DC2 can be found online at http://datacommunitydc.org/blog and Harlan Harris online at http://www.harlan.harris.name or on twitter @HarlanH.



Using Data Science to Solve Human Problems

Abe Gong

Data Scientist at Jawbone

Using Data Science to Solve Human Problems

We recently caught up with Abe Gong, Data Scientist at Jawbone and thought-leader in the Data Science community. We were keen to learn more about his background, his work at Jawbone and his latest side projects - including thought-provoking insights on how the ROI on Science is evolving ...

Hi Abe, firstly thank you for the interview. Let's start with your background and some of your early public policy related work...

Q - What is your 30 second bio?
A - I'm a hybrid social/computer scientist - interested in human problems, and how the right computational systems can sometimes solve them. I studied communications at BYU, then public policy, political science, and complex systems at the University of Michigan. I'm currently a data scientist at Jawbone, working on the UP fitness tracker.

Practically speaking, that means I get to spend my time building data systems to nudge people to form good habits and live healthier.

Q - How did you go from Communications to Public Policy to Data Science?
A - I feel like I've always done data science - we just didn't call it that until recently. I've been writing code since I was 10, and my side jobs and internships were always data-related. Comms taught me formal statistics, with application to marketing and PR. Public policy extended that training to government and policymaking. My PhD followed up with a stiff dose of web scraping, natural language processing, and research design. The subject domains sound different, but the core skills are very similar. When data science became a thing a couple years ago, I said, "Great! Now there's a name for this kind of work!"

Q - Was there a specific "aha" moment when you realized the power of data?
A - My senior capstone project in PR and market research was a statewide survey of military families in Utah. This was 2005 and 2006, so the first deployments to Iraq and Afghanistan were just ending. For the first time in 30 years, lots of soldiers were coming home with PTSD and other combat-related issues. No one really knew how to cope with it, and the strain was tearing up a lot of people and families.

I worked in the call center and did most of the data analysis for the project. As I talked to the officials in the military and the Veteran's Affairs office, I realized that our little amateur research team had the clearest picture of how deployments and PTSD were affecting the state.

By collecting the right data (i.e. really listening to people), we had become the representatives for a suffering constituency with no other voice.

That was an "aha" moment for me. I think I had previously assumed that big institutions, like the VA, were basically well-informed and rational. In that project, I realized how often the right information is missing from important conversations.

Q - What attracted you to the intersection of data, politics and blogs?
A - Science always follows data. Blogs struck me as a fantastic source of readily available data, and I was sure it must be good for something. From there I worked backwards to political theories - because I was in a PoliSci program - that could be enriched by bringing in data from the blogosphere: political participation, new media, and civil discourse. That's not the way you're supposed to do it - theory and research questions are supposed to drive data collection. But it worked in this case because blog data is just so rich. (Also, my committee was very supportive. I think they were curious whether I could actually pull off the research design I'd pitched.)

Q - How do you think the future of Public Policy will look as people like you, Nate Silver and others apply a very data heavy approach?
A - Ha - that's probably the first and only time that Nate Silver and I are mentioned in the same sentence. I've moved away from direct policy work, so others (Jake Porway, Drew Conway, Matt Gee) can probably answer this question better. My sense is that open data will improve policymaking, but that progress will be slow and uneven: two steps

forward, one step back. Many opportunities to improve governance through data science are going to open up in the coming years, but I don't think I have the patience to wait for them.

Very interesting background and insights - thanks for sharing! Let's change gears and talk more about Data Science and Machine Learning...

Q - What excites you most about using Machine Learning and Data Science in your professional life?
A - I want to push back on the question a little, because in my experience, machine learning is only a small fraction of data science. Case in point: I came to Jawbone a little more than a year ago as the company's first data scientist. Since then, at least half my time has been spent building infrastructure: ETL, scheduling, making sure systems scale, making sure we have the right instrumentation, making sure that other groups know to tell us before changing their data structures.

It's not the sexy part of data science, but when you get it right, everything else falls into place. Your analysis is faster and more conclusive. Your data products are more fun to build and ten times more reliable in production. A lot of data scientists miss the importance of the infrastructure layer, and that ends up seriously constraining the speed, scope, and quality of their work. Now and then my work calls for statistics/machine learning, but it's usually the last step in a long data pipeline - the icing on the cake, really. You can have a cake without icing, but not the other way around.

To your question about what's most exciting: one of my big projects right now is developing Jawbone's system for AB testing on the UP band. It's a great business intelligence asset for the company, and it's also a fantastic platform for nudging and improving user behavior. In other words, it's a great place for doing Science. We have all the same levers and tools as a growth hacking team at a typical SaaS company (content changes, UI changes, timing changes, in-app messaging, email, etc.), but our dependent variables are a lot more interesting. Instead of trying to convert/retain/upsell customers, we get to optimize for things like miles walked/run per user, sleep quality, and habit formation.

In other words, we're building infrastructure to tackle some of the big, unsolved problems in psychology and behavioral economics. I love working on these problems from a vantage point with such awesome data and reach. I also love that our relationship with users is fully collaborative - instead of trying to grab more eyeballs or induce more clicks ("Find out how this Mountain View mom makes over $6,000 a month with this one weird trick!") - we're trying to help users achieve their own lifestyle goals. There's nothing wrong with ad targeting, but I feel blessed to work on data problems with more direct human impact.

Q - That sounds fantastic! Now, while you're doing all this - what are your favorite tools/applications to work with?
A - I'm a python guy. I love ipython, pandas, scikit-learn, and matplotlib. Probably two-thirds of my workflow revolves around those tools. I used R a lot in grad school, but gave it up as I started working more closely with production systems -it's just so much easier to debug,

ship, and scale python code. For backend systems, I'm agnostic. I tend to use the AWS stack for my own projects, but the right combination of streaming/logging/messaging/query/scheduling/map-reduce/etc. systems really depends on the problem you're trying to solve. In my opinion, a full-stack data scientist should be comfortable learning the bindings to whatever data systems he/she has to work with. You don't want to be the carpenter who only knows hammers.

Q - What are the biggest areas of opportunity/questions you want to tackle?
A - Habit change at scale. Habits are an awfully important part of what it means to be human, but we really don't know that much about how they work. That is, our theories of motivation, psychology, incentives, etc. don't yet explain why some habits stick and others don't. The science hasn't developed that far. That's changing, though. I'm convinced that this field is ripe for an explosion. The data is there, the commercial incentives are right, and there's enough existing social/psychological theory to prime the pump. In the next few years, I expect to see theories of habit change improve by leaps and bounds - we're talking about a minor revolution in the science of human behavior - and I'm really looking forward to being part of it.

Q - What personal/professional projects have you been working on this year, and why/how are they interesting to you?
A - I've already mentioned the stuff I'm doing at work, so let me tell you about a couple of side projects ... First, storytelling: after watching D.J. Patil's talk about how storytelling is an important skill for data scientists, I put a lot of my spare cycles into reading about, thinking

about, and practicing storytelling. I learned to look for story elements in data: plot, characters, scenes, conflict, mood, etc. Often, our first instinct is to reduce data to numbers and hypothesis tests. Looking for the stories in data is another good way to make data meaningful, especially when you want users to get personally involved with the meaning-making.

I've really enjoyed exploring the craft of storytelling. It's a tradition at least as old as the scientific method, and sometimes much more powerful: you may be able to persuade individual humans without telling stories, but it is almost impossible to persuade a whole group without good storytelling - stories are the API to human culture change. I'm not sure that this is unique to data science, but it's definitely worth knowing. If others want to read up on the subject, I highly recommend Story, by Robert McKee, Save the Cat, by Blake Snyder, and Campbell's classic The Hero with a Thousand Faces - in that order.

More recently, I've been exploring a topic I call "the ROI for science." This started with a blog post speculating about how data science might evolve as a profession, branched out into a search for root causes ("Why is data science getting big now?"), and led to a fascinating thesis. Here's the gist: cheap and ubiquitous data are driving up the return on investment for many kinds of research, causing a boom in the use of scientific methods in business and day-to-day life.

Once you spot the trend, you'll start to see examples all over the place: the recent J-curve in patent filings, the growth of the hacker/maker

movement, the big data infrastructure supporting scientific efforts like CERN. If we stopped with the simple trend - more Science!" - this would be a very optimistic story.

But the same premise leads to a counterintuitive corollary: as more research is driven by private investment, the benefits of science are increasingly being captured by private interests. Think of all the investment that goes into in business intelligence and operations research (and data science): many person-years and millions of dollars to develop the equivalent of a whole scientific discipline - devoted entirely to the success of a single business model. Other examples of the scientific method serving narrow interests: a growing body of industrial trade secrets that never pass into the public domain; secret surveillance technologies developed by governments; the increasing dependence of many academic researchers on datasets owned by corporations.

We're used to thinking of science as a public good - open, democratic, and freely shared - but as the ROI on science increases, we should expect far more science to be privatized. That's not necessarily bad, but it brings new risks, power relationships, and thorny ethical questions. I'm very interested in starting a conversation around these issues, including the role that data scientists can play in nudging the system in constructive directions.

Okay, stepping back. These are fun things to think and talk about - blue sky, big picture stuff. I also have some technical side projects in the

works (mostly quantified self projects about goal-setting, mental acuity, and productivity) but they're not ready for prime time yet.

Very thought-provoking - will definitely be interesting to see how data availability and the data science profession influence / impact the ROI on science - and to see who gains. Look forward to following your thoughts / the conversation on this topic! Finally, it is advice time...

Q - What does the future of Data Science look like?
A - Exciting! Like I said earlier, I'm convinced we're living through a renaissance of the scientific method. There's a scary side to the new power of data, but on the whole I'm optimistic about where we're headed. Science always thrives in a data-rich environment, and the information revolution ("software eating the world") is generating a wealth of data. More and more, science is going to be something that everyone can - and to some extent, needs - to do. That's the common thread behind the appeal of data science, the quantified self movement, and the emphasis on "big data." They're all about capturing data, applying the scientific method, and making life better by making it smarter.

Abe - Thank you so much for your time! Really enjoyed learning more about your background and what you are working on now - both at Jawbone and personally. Abe's blog can be found online at http://blog.abegong.com and Abe himself is on twitter @AbeGong.

Machine Learning => Energy Efficiency

Kari Hensien,
Cameron Turner

Optimum Energy & The Data Guild

Machine Learning => Energy Efficiency

We recently caught up with Kari Hensien - Sr Director Product Development at Optimum Energy and Cameron Turner - Data Scientist at The Data Guild. We were keen to learn more about their recent collaborations, bringing Data Science and Machine Learning to the world of energy efficiency...

Hi Kari and Cameron, firstly thank you for the interview. Let's start with your respective backgrounds and your first experiences with data...

Q - What is your 30 second bio?
A - Cameron - Architecture student becomes software engineer becomes data analyst becomes data scientist - still learning about all of the above.
[Editor note - Cameron co-founded ClickStream Technologies in 2003, which was acquired by Microsoft in 2009. Cameron holds a BA in Architecture from Dartmouth College and an MBA from Oxford University.]
A - Kari - Long-time product planner who made the jump from large

software company to startup - currently building products that leverage data-science and machine-learning disciplines.
[Editor note - Kari leads product management strategy at Optimum Energy. Prior to Optimum, Kari spent 15 years at Microsoft, most recently as Senior Product Planner in the Windows Product Group, where she directed product planning for the Windows hardware application development platform.]

Q - How did you get interested in working with data?
A - Cameron - I believe that statistics and creative data science can create answers to some of the world's toughest problems. Sometimes solutions can be finessed by correlation and analysis, rather than brute force approaches that attempt to answer a question directly. Data (at times) can make sense of the world and unlock the universe's secrets.
A - Kari - I have a natural instinct to question and dig deeper to gather the data needed to make an informed decision. It's incredible to now be looking at how a system is changing over time based on what it is learning.

Q - What was the first data set you remember working with? What did you do with it?
A - Cameron - In high school my friends and I created a tiny FM transmitter and hid it in the teachers' lounge. We were hoping to develop a database of article ideas for the student newspaper. Incidentally, we didn't get that far before our bug was found in the lunch table napkin holder. They weren't too pleased with us when we fessed up (to get our bug back).
A - Kari - Honestly, I was a girl scout and the cookie season was upon

us. I found myself trying to figure out how many houses I would need to stop at in order to sell enough boxes to get the Rubik's Cube.

Q - Was there a specific "aha" moment when you realized the power of data?
A - Cameron - While working at Microsoft in the 90s we developed an opt-in program for enthusiasts to share their software usage with us through nightly uploads. No one knew if people would agree to do this, but after the first night there were hundreds of new uploads to parse and analyze. I got goose bumps and remember thinking: "This is going to completely change how software is made." Of course now, the data is real time and the analysis can be done through learning algorithms. At the Data Guild we see opportunity everywhere for machine learning to massively disrupt industries under human control. We're humbled daily to be a part of this transformation.
A - Kari - At Microsoft we were trying to prioritize feature requests. There was no shortage of feedback from customers telling us what they do. Cam's work allowed us, for the first time, to put together a picture that compares what people say they do with what they actually do. We ended up prioritizing very differently as a result of data of actual use.

Very interesting background and insights - thanks for sharing! Let's change gears and talk more about Optimum Energy...

Q - What specific problem is Optimum Energy trying to solve? How would you describe it to someone who is not familiar with it?
A - Kari - Optimum Energy is focused on energy optimization in

enterprise facilities with a solution that provides automated, continuous commissioning through dynamic adaptation of complex HVAC systems. Essentially, Optimum uses technology that manages HVAC systems directly and reduces the amount of energy that they consume.

Q - Which Optimum Energy technologies/solutions have been most successful?
A - Kari - Optimum Energy is best known for its OptimumLOOP technology, which provides continuous, system-level energy optimization of centrifugal chilled water plants. The technology continuously and dynamically adapts to fluctuating load, weather and occupancy conditions to yield the lowest possible energy draw while maintaining occupant comfort.

Q - Tell us a little more about the partnership with The Data Guild - how did it come about? And what projects will you be working on together?
A - Kari - As a startup, having the funds and ability to invest in the R&D required to implement machine learning is challenging. When we recognized that this was an investment we needed to make, we connected with Cam and the Data Guild team to help us with the expertise needed to begin our efforts building the discipline. We are focused on equipment-level and building-system level projects that enable autonomous optimization of an HVAC System.

Q - What excites you most about bringing machine learning and energy issues together?
A - Cameron - There is an immediate opportunity here to substantially reduce carbon emissions through machine learning. I love the fact that I

can draw on data science best practices that are working in other verticals and apply them to improving energy efficiency.

Q - What are the biggest areas of opportunity or questions you want to tackle?
A - Cameron: Three things

1. Product line expansion and enhancement: equipment-and system-level HVAC efficiency
2. Ecosystem enablement: real-world equipment operating specifications
3. Customer targeting and opportunity analysis: initial plant assessment

Q - What machine learning methods have you found or do you envision being most helpful? What are your favorite tools/applications to work with?
A - Cameron - We use correlation/covariance analysis along with regressions to do basic modeling and build out our view of the landscape. We use both supervised and unsupervised learning to build clustering and identify untold structure in plant performance. We use recursive partitioning to identify custom rules for local set points based on global algorithm development. In terms of favorites: R/R-Studio, Python, Java, SQL, Tableau, Hadoop, AWS

Q - What publications, websites, blogs, conferences and/or books are helpful to your work?
A - Cameron - We feel indebted to our network of affiliates (http://www.thedataguild.com/people) for ongoing support and review

of ideas and approaches. Specifically, Paco Nathan and Dennis Lucarelli for their continuing support of our work. We love O'Reilly and look forward to their Strata conference here in the Silicon Valley (on now!), as well as the Data Visualization Summit where Cameron Turner will be speaking this winter.

Q - What project have you been working on this year, and why/how is it interesting to you?
A - Cameron - We're moving into pilot-test phase with a project that focuses on equipment recommendations in HVAC systems. This expands on the energy optimization that Optimum Energy currently provides and sends additional information to a facility about the most efficient combinations of equipment to run at a given time. Initial tests have been promising, and we're excited for the next test stage.

Q - What has been the most surprising insight or development you have found?
A - Cameron - We wanted to better understand what was happening within a chilled water system around a chiller surge. Engineers know a chiller is going to surge instinctively. They just need to be at the plant and they can see it, and feel it. We are trying to create vibrational and acoustic classifications around a surge to be able to better understand and predict them.

Very interesting - look forward to following future progress as these projects reach completion! Finally, it is time to talk a little more about each of your accomplishments and what you think the future of data science/machine learning and energy efficiency looks like...

Q - What in your career are you most proud of so far?
A - Cameron - Developing high-performing teams of great data scientists with diverse backgrounds and skills.
A - Kari - Seeing the products I've helped to develop in use and valuable to customers. Most recently, I planned, designed, built and shipped my first mobile app: OptiCx Trend.

Q - What does the future of machine learning and energy look like?
A - Cameron - Big question. It is inevitable that near-to-real-time cloud-based decision support systems will come to the energy sector. In fact, energy and related fields will be the first to embrace and extend the concepts of machine learning and true big data opportunity, due to: 1) the closed form of some aspects of the problem (for example, lower kWh consumption, higher savings), 2) the enormous upside of successful implementation, 3) the critical impact CO_2 emissions have on the earth's future.

Q - What's next for Optimum Energy?
A - Kari - We are compiling a lot of valuable real-world operating data: we currently have 250 cumulative years of data, and we add approximately 8 years each month. Our long-term plans involve continuously improving energy efficiency for its customers by leveraging

this data using machine-learning and predictive-maintenance algorithms. It's an honor to be playing a part in this transformation.

Kari and Cameron - Thank you so much for your time! Really enjoyed learning more about your backgrounds and what you are working on together. Optimum Energy can be found online at http://optimumenergyco.com and The Data Guild at http://thedataguild.com.

Training Deep Learning Models in a Browser

Andrej Karpathy

Machine Learning PhD, Stanford
Creator of ConvNetJS

Training Deep Learning Models in a Browser

We recently caught up with Andrej Karpathy, Machine Learning PhD student at Stanford and the man behind the innovative ConvNetJS - a JS library for training Deep Learning models (mainly Neural Networks) entirely in your browser. We were keen to learn more about his background, the motivation and potential applications for ConvNetJS, and his research agenda ...

Hi Andrej, firstly thank you for the interview. Let's start with your background and how you became interested in Machine Learning and AI...

Q - What is your 30 second bio?
A - I was born in Slovakia and my family moved to Toronto when I was 15. I finished a Computer Science / Physics undergraduate degree at University of Toronto, went on to do Master's degree at University of British Columbia working on physically-simulated animation (think simulated robots), and finally ended up as a Computer Vision / Machine Learning PhD student at Stanford where I am currently a 3rd year student. Along the way I squeezed in two wonderful internships at Google Research working on neural nets (Google Brain) for video classification.

Q - How did you get interested in Machine Learning?
A - As an undergraduate I studied Computer Science/Physics and Math with intentions of working on Quantum Computing. I saw that computing applications were going to completely transform the world and I wanted to help create the most efficient computing devices. This meant going down as far as possible to quantum level and utilizing the most elementary physical laws and particles to perform computation. However, as I took my Quantum Mechanics classes it became apparent that I was not having fun. It was too distant, too limiting. I couldn't get my hands dirty.

At the same time, I felt myself consistently gravitating towards topics in Artificial Intelligence. As one of my influential turning points, I remember walking around in a library realizing that there were zillions of amazing books around me and that I wanted to learn everything in all the books and know everything there is to know. Unfortunately, I also realized that this would be, for all practical purposes, a hopeless endeavor: I'm a blob of soft tissue with finite, leaky memory, a slow CPU, and damnit I felt hungry. However, it also occurred to me that if I can't learn everything there is to know myself, maybe I could build something that could. I refocused on Artificial Intelligence, and later (after a period of confusion when I was asked to code up graph search algorithms and minimax trees in my AI class), narrowed in on the branch I felt was closest to AI: Machine Learning.

Compared to my Quantum Computing escapade, I finally felt that the only thing that stood between me and my goal was entirely my own

ingenuity, not some expensive equipment or other externalities. Additionally, I realized that working on AI is arguably the most interesting problem because it's the ultimate meta problem: if I was successful in my quest, the AI could in principle learn all about anything, with Quantum Mechanics merely as a relatively insignificant special case.

Q - What was the first data set you remember working with? What did you do with it?
A - It was probably MNIST digit classification, hacking on Restricted Boltzmann Machines while auditing Geoff Hinton's Neural Nets class at the University of Toronto somewhere around 2007. But strangely, I don't remember the class having much impact on me at the time and in fact I remember being dismissive of it. I considered the digits classification to be a cute but useless toy problem and I couldn't understand why Geoff got so excited about digits. At the time I wasn't ready to extrapolate what I was seeing to my own motivating examples, including for example, reading and understanding the content of all books/websites/videos in the entire world.

Q - Was there a specific "aha" moment when you realized the power of data?
A - I think it was a gradual process. After I learned the mathematical formulation of regression and some of the related approaches I started to realize that many real-world problems reduced to it. Armed with logistic regression (but also more generally Machine Learning mindset) as a hammer, everything started to look like a nail.

Very interesting background and insights - thanks for sharing! Let's change gears and talk more about Neural Networks...

Q - What are the main types of problems now being addressed in the Neural Network space?
A - Well, first what we do know is that neural nets seem to work very well in fully-supervised regression problems (i.e. learning fixed mappings from input to output) when you have a Lot of data and sufficient computational budget. In fact, I would argue that we've learned something more profound: neural nets are in their basic form just non-linear regression and the more general lesson that has emerged is that you can, in fact, get away with formulating scary-looking non-convex objective functions and that it is seemingly possible to "solve" them in real-world scenarios, even with simple first order methods. This is not an obvious conclusion and for a long time many people were worried about local minima, lack of theoretical guarantees, etc. However, the field has recently seen an explosion of renewed interest, in large part due to dramatic improvements obtained on important, large-scale problems in the industry (speech and vision being among first few).

There are plenty of open questions and exciting directions. I'll list just a few: We haven't really figured out very good ways of taking advantage of unsupervised data (most serious industry applications right now are large, fully supervised neural nets with a huge amount of labeled training data). We still don't have a very satisfying principled approach

for optimizing neural nets that is in practice consistently better than a few tricks and rules of thumb that have been developed over the years largely by the "guess-and-check" method (and it is common to rely on vigorous hand motions rather than mathematical theorems to justify these tricks). Neural nets are also in their basic form fixed, feed-forward functions of data and many people are interested in composing them (for example as seen in recursive neural networks), using them in structured prediction tasks, reinforcement learning tasks, or perhaps most interestingly formulating loopy models and training neural nets as dynamical systems (recurrent neural networks). I also think we'll see interesting multi-task learning approaches for embedding different types of data into common "semantic" vector spaces (especially words, n-grams or entire sentences, which is currently a very active and relatively new area of research), techniques for mapping between modalities, etc. It is an exciting time to be in the field!

Q - Who are the big thought leaders? [you can say yourself :)]
A - Haha, maybe if you ask me in 20 years I could list myself, but for now I'll go with a largely uncontroversial answer: Geoff Hinton, Yann LeCun and Yoshua Bengio.

Q - What excites you most about working with Neural Networks?
A - Neural Networks enjoy many desirable properties! Just to list a few: They can be trained online, they are efficient at test time (in both space and time), they are modular (the gradient is possible to derive locally and decomposes very simply through chain rule), they are simple, and they work. There are also plenty of interesting connections (pun not intended) to neuroscience and the human brain.

Q - What are the biggest areas of opportunity / questions you would like to tackle?
A - On an application-agnostic side of Neural Networks, I'm always more interested in simplifying than complicating, which is likely attributable to my physics background where all fundamental equations are sweet, beautiful and short. If the core of a (well-designed) Neural Networks library is longer than few hundred lines of code, it means there's still more unnecessary complexity and options to get rid of. The models we use are already fairly simple (nets with Rectified Linear Unit activation functions really just come down to repeated matrix multiplication and thresholding at zero), but the training protocol, hyper-parameter choices, regularization and data preprocessing/augmentation tricks are infamous for being a messy dark art that requires expensive cross-validations.

Sounds like a very interesting time to be in the field, with lots of areas to explore - look forward to keeping up with your work! On that note, let's talk more about your recent project creating a JS library for Deep Learning models entirely in a browser...

Q - Firstly, what motivated you to create it?
A - I noticed that there were many people interested in Deep Learning but there was no easy way to explore the topic, get your feet wet, or obtain high-level intuitions about how the models work. Normally you have to install all kinds of software, packages, libraries, compile them for your system, and once (and if) you get it all running you can usually

look forward to a console window that spams the current average loss across the screen and you get to watch the loss decrease over time. The whole experience from user's point is dreadful. I wanted to make the models more accessible, transparent, and easy to play with and I figured that browser is the perfect environment.

Q - How did you build it? How does it work?
A - It started off as a fun hack over my christmas break in Toronto. I've already implemented Support Vector Machines and Random Forests in Javascript and as I sat in a Tim Hortons one day sipping delicious Canadian coffee I decided it was a good time to write up a Deep Learning library. My initial intuition was that the (convolutional) networks would be too slow but I ran some quick benchmarks and was blown away by the efficiency I was observing in Chrome with the V8 engine. That encouraged me to continue the development. Along the way I was also labeled as "insane" for working on this project by a good friend of mine, which only made me further step up my efforts - when people call you insane you know you've hit on something interesting!

Q - What Machine Learning processes, models, techniques etc. does it enable?
A - The library allows you to specify and train neural networks. It also supports convolutional neural networks which are designed specifically for images. Images are technically very high-dimensional input spaces but also possess certain simplifying properties that are explicitly taken advantage of in regularizing the network through restricted connectivity (local filters), parameter sharing (convolutions), and special local invariance-building neurons (max pooling).

Q - What is your favorite demo application, and why?
A - I haven't fully finished my favorite demo yet. It is a 60-million parameter convolutional network that is trained on ImageNet (the largest image recognition dataset) using GPUs and then ported to ConvNetJS through JSON. The issue right now is that the model is about 200MB in raw number of bytes and several gigs in JSON format so it's a little hard to work with. I'm working on quantizing and compressing the network in various ways so that it fits into a few tens of megabytes and still delivers state of the art image recognition performance, but in your browser. I don't consider this to be exceptionally useful but I think it would go a long way in demonstrating what's possible in Javascript, today.

Q - What was the most surprising result / learning?
A - Javascript's efficiency was the most surprising result and I only expect the trend to continue. We're starting to see a lot of low-level support built into browsers (for example with WebGL) and once there is support for efficient matrix multiplication (a necessary, core building block of not only Neural Networks but almost all Machine Learning algorithms), it will enable a myriad of highly-accessible applications. I'm keeping my fingers crossed - if that happens I'll be able to change a few lines of code and expect to see at least an order of magnitude increase in efficiency.

Q - Where could this functionality be applied? What potential application are you most excited about?
A - First, I think there are obvious educational applications since it is so easy to set up, train, visualize and tune networks. You can quickly build

intuitions for different variables (such as the learning rate, or weight decay) if you fiddle with a slider and see the effect on the accuracy or the features right away.

There are some potentially interesting uses for training Neural Networks in browser extensions (modeling some aspects of user behavior, or content of sites they visit), uses in web-based games, or from a site owner's perspective immediate user modeling based on site interaction to deliver a customized experience. Since it's entirely pure Javascript, it's also instantly available on node.js and available for use on server side.

The library could also be used in larger and more serious applications in a hybrid setup where the expensive training is done on an efficient backend and the network weights are loaded through JSON for a reasonably quick test-time prediction in Javascript. Alternatively, it could serve as nice browser-based visualization frontend that interacts with a C++ library that potentially trains a network on GPUs. This is how I'm currently using a fork of the library in my own research, as the C++ code I'm working with only scrolls numbers in a console and I have a preference for looking at pretty pictures of learned weights, neat d3js loss/accuracy curves and example predictions to monitor the progress of my network while it trains. It's also easier to click a button to anneal the learning rate without having to enter commands in the console.

Lastly, here's a crazy idea: massively distributed Neural Network training (think FoldIt, or SETI@Home), except every client merely visits

a URL and right away starts to contribute Javascript compute time by sending gradient updates to a central server. A few issues have to be addressed first in terms of the modeling: vanilla Neural Networks have dense interactions so they are difficult to parallelize and naive use of distributed optimization techniques is likely to pose problems with stale gradients.

It really is a terrific resource, with many applications - the ML community should be very grateful that you developed it!

Q - Now, in terms of your more formal research, what are you working on day-to-day??
A - As I alluded to above, my preferred application domain for Machine Learning is Computer Vision. The vast majority of the content on the Internet (by both size in bytes and attention) is visual media content, yet we have almost no idea what's in it! In a sense, images and videos are the dark matter of the Internet. I also like that analogy because that makes me an Internet Astronomer.

In terms of my general approach, my motto has always been "I like my data large, my algorithms simple and my labels weak". I'm seeking to develop algorithms that gobble up all the images/videos on the Internet and learn about the visual world automatically with very few human-provided annotations. With the current state of the art methods if you want your algorithm to recognize a sheep in an image you have to first provide it hundreds, thousands (the more the better) of examples before it can do so reliably. This low-hanging-fruit approach turns out to work

well and is essentially how all current industrial applications work, but philosophically it is revolting: your parents didn't have to show you thousands of images of sheep in all possible angles, poses, occlusions and illumination conditions before you could recognize one. They pointed it out once as a "sheep" and you associated the sound to that already-familiar visual concept. The learning was instantaneous and did not require thousands of examples. I work on developing learning algorithms that can do the same, provided that they have seen and processed millions/billions of unlabeled images first.

So far I've only run at most the first mile of the marathon that I expect the above task to be. A representative paper that is relevant to my current interests at Stanford is my NIPS paper with Adam Coates on training an unsupervised deep learning model on millions of images from YouTube and automatically discovering frequently recurring visual concepts. The paper is very similar in spirit to the famous Google network discovered cats paper from Le et al [ICML 2012]. In retrospect I think there were many things wrong with the modeling and the approach but I'm hoping to address all that in the coming years!

Definitely a fascinating area to explore - good luck with the next few miles of the marathon! Finally, it is advice time...

Q - What does the future of Machine Learning look like?
A - Broad question! In terms of research I can confidently say that we are going to see a lot of rapid progress in Neural Networks areas I outlined above. There are other areas which I consider promising but I

know relatively little about, such as Bayesian Optimization and Probabilistic Programming. In terms of applications, I'm convinced that the future of Machine Learning looks very bright and that we will see it become ubiquitous. The skill to manipulate/analyze and visualize data is a superpower today, but it will be a necessity tomorrow.

Lastly, I expect that Machine Learning in Javascript will also become ubiquitous due to its wide availability, accessibility, the benefits of the browser as a wonderful, powerful, efficient and interactive UI framework and my hunch that the vast majority of machine learning applications are actually only of medium size and easily and immediately crushed with Javascript (on desktop/mobile browser, or node.js server) without a need for complicated backends, pipelines or communication protocols. I expect we should see a very successful and widely used Machine Learning library for Javascript within a few years. Feel free to take this with a grain of salt though, since I have a history as a notorious fan of web-based technologies.

Q - Any words of wisdom for Machine Learning students or practitioners starting out?
A - You learn the most by reinventing the wheel. Don't just read about Machine Learning algorithms and fall into trap of thinking you understand the concepts because everything you read sounds reasonable. Read it once and then re-implement it from scratch, yourself. And while you're at it, do it in Javascript ;)

Andrej - Thank you so much for your time! Really enjoyed learning more about your background and what you are working on now - both your personal projects and more formal research agenda. Andrej's blog can be found online at http://karpathy.ca/myblog and he is on twitter @karpathy.

Predictive Policing

George Mohler

Chief Scientist at PredPol
Asst. Professor Mathematics & CS,
Santa Clara University

Predictive Policing

We recently caught up with George Mohler, Chief Scientist at PredPol, Inc and Assistant Professor of Mathematics and Computer Science at Santa Clara University. We were keen to learn more about his background, the theory and technology behind predictive policing and the impact PredPol is achieving ...

Hi George, firstly thank you for the interview. Let's start with your background and how you became interested in predicting crime hotspots...

Q - What is your 30 second bio?
A - I am Chief Scientist at PredPol, Inc and Assistant Professor of Mathematics and Computer Science at Santa Clara University. Prior to joining the faculty at Santa Clara University I was CAM Assistant Adjunct Professor of Mathematics at UCLA from 2008 to 2010. I received a B.S. in Mathematics from Indiana University and my Ph.D. in Mathematics from the University of California Santa Barbara.

Q - How did you get interested in Data Science and Machine Learning?
A - I became interested in data science rather late, during my postdoc at UCLA. Prior to that I was working on computational methods for

variational models of polymers in graduate school. When I joined the crime modeling group at UCLA, I started to work on similar types of optimization problems, but applied to spatio-temporal crime patterns. We had a large dataset provided by the Los Angeles Police Department and I was interested in understanding the statistics of crime hotspots and how they could be predicted.

Q - What excites you most about bringing Data Science and Policing together?
A - Some of the models we use at PredPol are self-exciting point processes that were originally developed for modeling earthquake aftershock distributions [Marsan and Lenglin, 2008]. The fact that these point process models fit earthquake and crime event data quite well is, by itself, a cool result. However, in the context of policing we can actually send police into the hotspots that we predict in order to prevent crime. So not only does predictive policing present an interesting modeling problem, but the models then have a societal impact that can reduce the risk that one's car is broken into or that they are a victim of gun violence.

Editor Note - If you are interested in more details on the research underlying the models, the original academic paper is very insightful. Here are a few highlights:

- Criminological research has shown that crime can spread through local environments via a contagion like process. For example, burglars will repeatedly attack clusters of nearby targets because local vulnerabilities are well known to the offenders

- Self-excitation is also found in gang violence data as a gang shooting may incite waves of retaliatory violence in the local set space (territory) of the rival gang. The local, contagious spread of crime leads to the formation of crime clusters in space and time. Similarly the occurrence of an earthquake is well known to increase the likelihood of another earthquake nearby in space and time
- Mohler and his fellow authors propose (and demonstrate!) that self-exciting point processes can be adapted to capture the spatial-temporal clustering patterns observed in crime data. More specifically, spatial heterogeneity in crime rates can be treated using background intensity estimation and the self-exciting effects detected in crime data can be modeled with a variety of kernels developed for seismological applications or using nonparametric methods

Editor Note - Back to the interview!...

Very interesting - fascinating that the earthquake models can be applied to crime event data! Let's talk more about the technology that you have built at PredPol...

Q - Firstly, what specific problem does it solve?
A - Police departments have limited resources. Officers have large beats to cover and limited time when they are not on a call to service. So during a given shift in a given beat, an officer might only be able to patrol k hotspots. From a prediction perspective, we would like to flag

for patrol the k hotspots that are most likely to have crime in the absence of police. This means that hotspot policing is actually a learning to rank problem.

Q - How does the technology work?
A - PredPol is a SaaS company and officers access the software on a computer or smart phone with an Internet connection during their shift. They then pull up a UI that includes a map with the hotspots (150m x 150m) displayed. The idea is that the officers then make extra patrols in those areas when they are not on a call to service. The key with this sort of technology is to make it as simple and easy to use as possible, because police have a very difficult, dangerous job and they don't have time to mess around with complicated software in the field.

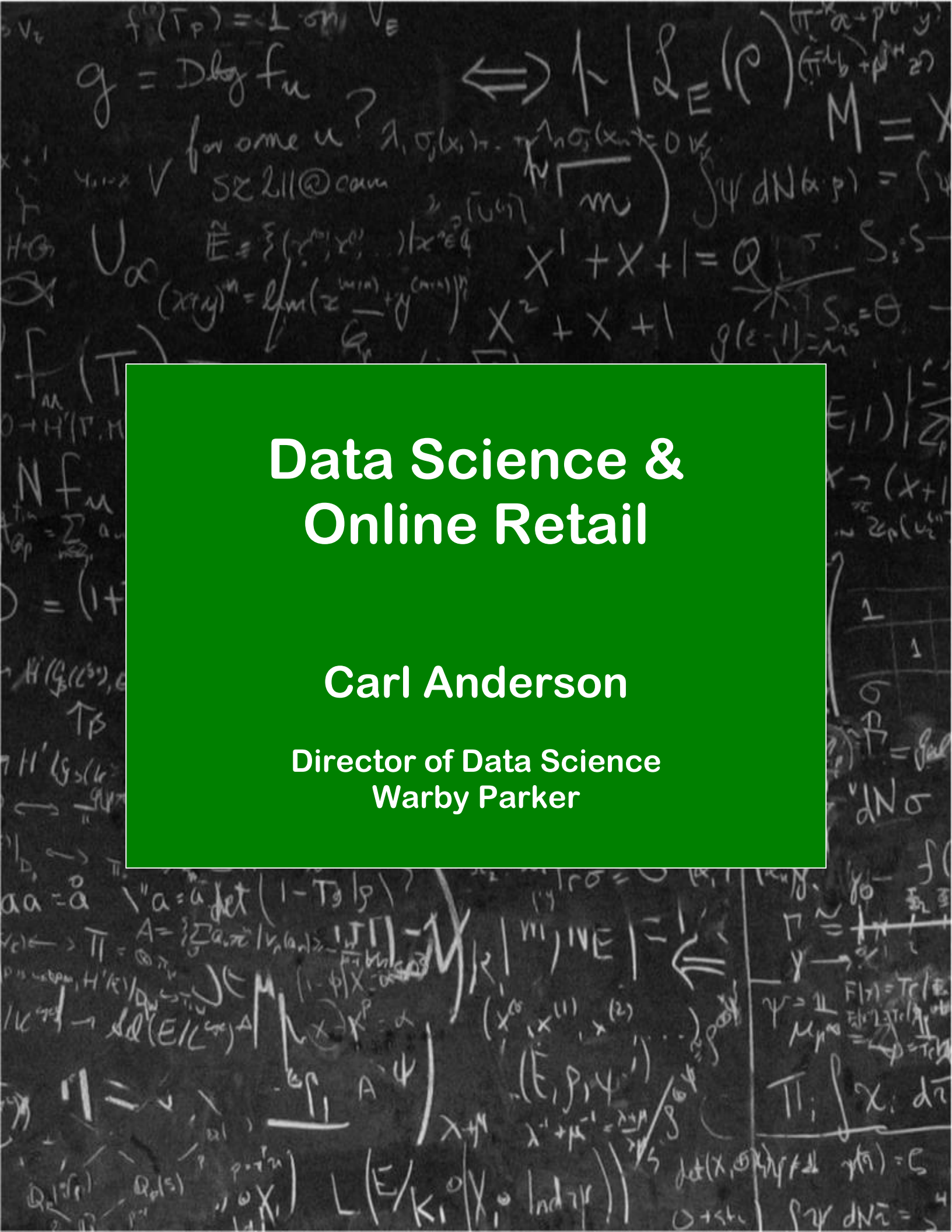
Q - That makes sense... And what is your favorite example of how PredPol is having real-world impact?
A - We have run randomized controlled trials to measure accuracy of the PredPol algorithms and impact on crime rates. These are necessary, because without them it is impossible to determine whether a crime rate increase/decrease is due to the technology and its use or because of some exogenous factor. But my favorite examples are at the scale of individual hotspots. For example one agency had a guy stealing cars with a tow truck. So the police put a decoy car with a GPS tracker in one of the PredPol hotspots and sure enough he came and towed it away (and the police were able to catch him).

That's a great example :) It sounds like you've already had significant impact with PredPol - let's talk a little about the future...

Q - What research areas would you like to explore more going forward?
A - We still do a lot of work on improving our algorithms at PredPol, both in terms of bringing in new modeling approaches and also exploring what loss functions make sense for policing (and how to optimize them). However, predictive policing is not just about designing accurate algorithms. Ultimately the software has to be used by police in the field and so human-computer interaction is really important. We are exploring ways in which the software can increase officer patrol time in hotspots while still fitting seamlessly within their existing practices.

Q - And finally, any words of wisdom for Data Science / Machine Learning students or practitioners starting out?
A - Many universities have courses in machine learning and some are starting to have degrees specifically in data science. But there are many ways to learn data science on your own. I think Kaggle is a great way to start out and there are some entry level competitions that walk you through some of the basics of data science. Coursera has free courses in data science and machine learning; I took Bill Howe's "Introduction to data science" class over the summer and thought it was really well put together. I recommend to my students that they try to do an internship or REU in data science if they are interested in pursuing a career in the area.

George - Thank you so much for your time! Really enjoyed learning more about what you are achieving with PredPol. George can be found online here and PredPol at predpol.com.



Data Science & Online Retail

Carl Anderson

Director of Data Science
Warby Parker

Data Science & Online Retail – at Warby Parker and Beyond

We recently caught up with Carl Anderson, Director of Data Science at Warby Parker (and previously at One Kings Lane). We were keen to learn more about his background, his perspective on how data science is shaping the online retail landscape and what he is working on now at Warby Parker...

Hi Carl, 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 have a fairly diverse background. I grew up in the UK. I have a B.Sc. in Biology, M.Sc. in Biological Computation, and a Ph.D. in Mathematical Biology from a Probability and Statistics department. I spent a few years doing postdocs at Duke, Georgia Tech, and in Europe before becoming a faculty member at Georgia Tech. I left academia to do consulting in complex systems and agent-based modeling. After that

there were some short stints doing lossless data compression algorithms and working for the Sunlight Foundation. I spent 4 years as a senior scientist building analytical systems for our large-scale models at Archimedes in San Francisco, before a stint as a data scientist at One Kings Lane. I'm currently the Director of Data Science at Warby Parker in New York.

Q - How did you get interested in working with data?
A - Like many data scientists, I have a classical pure science background and data are the lifeblood of the empirical scientific method. If you want to be a scientist, you have to care about data.

Q - Was there a specific "aha" moment when you realized the power of data?
A - As a biological sciences undergrad, I remember being fascinated by some classes and labs in paleoclimatology. I loved the idea that using a simple non-specialized microscope one could count and measure tree ring growth over hundreds, and with Bristlecone Pines, thousands, of years. With similar evidence from ice cores (which spans longer time periods) and shells of tiny Foraminifera animals, a person can combine these data proxies to provide a compelling reconstruction of ancient climates. In a sense, all of this evidence is just sitting around us waiting to be used (just like the current big data hype today). I loved that idea.

The other key moment for me was in my Master's degree in Biological Computation. This unique masters degree took biologists and taught them to do three things: computer programming, mathematical modeling, and statistics. This gave me a whole new suite of tools in my

toolbox and helped me realize that I could come up with hypotheses and run computer simulations and generate data *de novo* and not be limited by sample size in the way that the biologist empiricists were. This led me to do a lot of agent-based models on badgers (Masters) and with ants, bees, wasps (Ph.D.), later with humans, information packets and so on (consulting).

Very interesting background and insights - thanks for sharing! Let's change gears and talk more about Data Science in Online Retail...

Q - What attracted you to the intersection of data/data science and online retail?
A - When I left scientific consulting to join One Kings Lane (a successful home décor flash sales site), I was interested in working somewhere that was consumer facing. I had always been interested in style and design and the fact that the company dealt with 4000 new products, 60% of which had never been sold on the site, each and every morning was a real draw from a data perspective. Likewise, Warby Parker designs and makes beautiful products that customers love. Both of these companies represented interesting challenges.

To be honest, I am fairly agnostic about the context and generator of data (which in my background has been as diverse as ants, robots, and battleships) and am drawn to interesting intellectual problems. How do we deal with the cold start problem of 4000 new products? How do we generate a holistic view of our customers as they interact across our

channels of mobile, web, school buses, and physical retail stores? These are all fun challenges.

Q - Which online retail 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 - To state the obvious, Amazon has been a leader for many years now. Their site is littered with different recommenders and they are sitting on a mountain of data. Their UI and experience can be kind of clunky though. That is, the site can expose weird recommendations that can be kind of jarring: people who bought X happened to buy something completely random and seemingly unrelated Y. The best data science, though, will be invisible. It will just work (more on that below).

Birchbox (with Liz Crawford as CTO) is doing some great work. Nordstrom's data lab has done an impressive job putting together a new team, selling the power of data science internally, and getting proven better-than-human recommenders in front of customers.

Q - Which companies/industries do you think online retail should emulate with respect to their use of data science? Why?
A - Netflix drove the field forward with the Netflix prize. However, they continue to innovate with A/B testing and measuring everything, (see http://www.slideshare.net/justinbasilico/learning-a-personalized-homepage). Interestingly, they have a very strong focus on driving the re-subscription rate (i.e. getting people to keep paying a monthly fee), akin to maximizing the lifetime value of the customer. This is a very downstream metric that rolls up all of the user's experiences but it also

means that they know how much they can spend to acquire and keep customers, (see https://blog.kissmetrics.com/how-netflix-measures-you/).

I'm also a really big fan of LinkedIn's data science team. One thing that can happen is that data scientists can try to be too clever and develop an unnecessarily complex algorithm to infer customer's intent, taste etc. They forget that one can simply ask the customer directly. Does Joe Doe know about project management? We don't know for sure so let's ask their contacts. While the system can be gamed (same as Yelp reviews and any other user generated content), it provides a simple --- and in LinkedIn's case, a very large --- direct source of data. Online retailers are a little too reticent to ask customers what they want. What would you like to see in our next flash sale? If you could choose the color of the next widget, what would you choose? Get the data and let product managers and data scientists sift through it.

Q - Where can data science create most value in online retail?
A - There is the old cliché about the right thing at the right place at the right time....One Kings Lane was always trying to find the right balance between getting products that the data science team and OKL buying team believed individual customers would love versus serendipity, letting them explore and chancing upon something new, different and unexpected. This is a really delicate balance and the oft-cited example of Target knowing a teenage girl was pregnant before the girl's father, fell to the wrong side of that line. Great data science will help companies understand their customers better, not just from a historical context but grasp the current context --- what do they really want or need right

now? --- so that the data products will be viewed as a boon, a digital assistant that will aid the customer.

For instance, Google Now is viewed by many as helpful and not creepy as it provides you the train or bus times or weather at your current location when you need them. This, however, is not true of most online retail experiences where the cross sell and upsell is crude and in your face. Data science needs to disappear from the customer's perspective. Jack Dorsey said it better:

"I think the best technologies -- and Twitter is included in this – disappear... They fade into the background, and they're relevant when you want to use them, and they get out of the way when you don't.".

Data science in online retail is not restricted to customer facing websites and services though. My team works with just about all teams in the company and has potential to improve operations and decision making across the board from supply chain, finance, merchandising, and marketing. I reviewed the range of these contributions in a recent blog post and in the past ten days have been working with business owners in all of these different areas. Particular examples we are working on right now include performing unsupervised clustering and decision-tree based classifiers of customers to identify signals early on in the relationship that might identify high lifetime value customers. We are also iterating on our tools and models for predictive sales forecasting at the sku level.

Really interesting ... fascinating to see how data science can have an impact across retail functions; not just the consumer facing parts of the business model. On that note, let's talk more about your work at Warby Parker...

Q - Firstly, what are the major similarities and/or differences between how data science is being applied at Warby Parker vs. One Kings Lane?
A - When I arrived at One Kings Lane, the data engineering, warehousing, and analytics were pretty much all in place. The data were, for the most part, there ready to use; of course, it is never as clean as you might think or want. At Warby Parker, in contrast, all of that was missing and it has been my team's responsibility to create that, to get the data in place before we start to make use of it. We have just turned that corner and with a core set of data in place, we are now shifting our focus to building models and data products, as well as provide a robust set of reporting and business intelligence tools to support our analysts..

Q - What are the biggest areas of opportunity/questions you want to tackle at Warby Parker?
A - Much of the last year has been spent putting our data infrastructure in place. Getting data into databases, creating a single source of truth, and creating an accurate catalog. Now that we have that, this year is going to be very different. We are going to focus a lot more on understanding the customer. Who are they and what makes them tick? What's the relationship between online and offline (we plan to open more physical retail stores in the coming years) experiences? For this, we will marry sources such as clickstream, transactional history, in-store analytics and social media.

Q - What projects have you been working on this year, and why/how are they interesting to you?
A - Two weeks ago, we sent our Home Try-On recommender out for A/B testing (our Home Try-On program allows a customer to order five frames, have them shipped home to try-on and can then send the box back to us, all free of charge.) Unlike most e-Commerce sites our basket size is very small. Customers wouldn't normally purchase a pair of glasses frequently, as they would groceries. However, our HTO program ships boxes of five frames to customers. This is a really great dataset because it is reasonably large and you can look at the covariances among the five frames plus what they subsequently purchased and build a recommender based on basket analysis. We hope that the tool will help customers better choose frames that they'll want to purchase.

Q - What has been the most surprising insight you have found?
A - Warby Parker sells a monocle and it has an extremely high conversion rate. Most people who order this in their Home Try-On boxes end up purchasing it. Conversion is so high that we had to tweak our basket analysis algorithm specifically to account for it.

Q - That's very surprising! :) ... Last thing on Warby Parker ... About a year ago you wrote a very interesting blog piece on "How to create a data-driven organization" and how you planned to do so at Warby Parker ... a year on, what do you think is working well?
A - Great question. I have been working on a follow up "one year on" post that will appear in Warby Parker's tech blog very soon. (As such, I'll keep these responses short.) ... We have got much better at evaluating what the tech team should be working on. Different business owners

essentially have to compete for software developer (agile team) time and have to quantify the return on investment. The underlying assumptions for both the return and investment have to be clear, justifiable, and the metrics must be comparable across proposals. That way, all managers (who vote on all the competing initiatives) are able to view what different teams are interested in and what will drive the business forward.

Q - Has it proven harder than you imagined?
A - Getting analysts across the company to knuckle down and learn statistics. With a Ph.D. from a statistics department, I am very biased, but a sound basis in statistics seems to be an essential tool of any analyst. Like many skills, statistics may not feel useful and relevant until a specific project comes along and the analyst needs it --- for instance, when we need to A/B a change in the website or optimize email subject lines and send times.

Q - Anything you wouldn't repeat if you could start over?
A - There is nothing that we've done that I've completely regretted and wouldn't do. There are many things, however, that I would do better the second time around. These range from getting business intelligence tools in place, running our analyst guild, and productionizing our systems (according to my boss, things don't exist until they are in production).

Carl, thanks so much for all the insights - really interesting to get such a detailed feel for the work and culture you are developing at Warby Parker!

Finally, it is time to look to the future and share advice...

Q - What excites you most about recent developments in Data Science?
A - Without doubt, deep learning. With my biologist's background, I have always had an interest in cognition and Artificial Intelligence. However, it has never delivered. When I read Jeff Hawkins' *On Intelligence*, a few years back, it blew my mind. Here was a biologically plausible hierarchical model of the neo-cortex and which represented a generic learning model. Numenta's (Hawkins' company around this) model at the time was in its infancy. It has since improved significantly but I think has been leapfrogged by the exciting deep-learning work by Hinton et al. Here we have a relatively simple hierarchical model (stacked auto-encoders, which if you squint, arguably look similar to Hawkins' model) that is proven to work: it really can classify cats autonomously, it can win kaggle competitions without any domain knowledge etc.

Q - What does the future of Data Science look like?
A - I think that it is very bright indeed. Computational power is only going to get bigger and faster, algorithms (such as deep learning) will become easier to train and use, and hopefully tooling will be easier for the average non-technical user to harness machine learning.

I do wonder about the term data science though. It may well disappear over time as sub disciplines become more established. If you look at the history of science, rich gentlemen were not scientists but "natural philosophers" and members of the Royal Society would attend readings of papers that spanned the gamut of what is today known as biology,

physics, chemistry, and mathematics. Differentiation into those disciplines only happened relatively recently (1800s) and as those fields grew and became more established, we defined sub-disciplines such as oncology and later pediatric oncology. Data science as this overly broad umbrella term is still akin to natural philosophy.

Q - That makes sense - in that context, any words of wisdom for Data Science students or practitioners starting out?
A - Just do it. There is no substitute for getting your feet wet and working on things. These days, there is no shortage of data (even the government has finally caught on: https://www.data.gov), free online courses, books, meetups of like-minded individuals and open source projects. Reading a book is one thing but getting real data and having to deal with missing data, outliers, encoding issues, reformatting, i.e. general data munging, which really can constitute 80% of time of a data science project are the kind of dues that you must pay to build the suite of skills to be a data scientist. Kaggle has some great introductory 101 competitions and http://scikit-learn.org/ has some great documentation that can help get you started.

Carl - Thank you so much for your time! Really enjoyed learning more about your background, your perspective on how data science is shaping the online retail landscape and what you are working on now at Warby Parker. Carl's blog can be found online at http://www.p-value.info and he is on twitter@LeapingLlamas.

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