SEXY SCIENTISTS WRANGLING DATA AND BEGETTING NEW INDUSTRIES

Chris Wiggins
(The New York Times)

Caitlin Smallwood
(Netflix)

Amy Heineike
(Quid)

Jonathan Lenaghan
(PlaceIQ)

DATA SCIENTISTS AT WORK

Roger Ehrenberg
(IA Ventures)

Erin Shellman
(Nordstrom)

Victor Hu
(Next Big Sound)

John Foreman
(MailChimp)

Claudia Perlich
(Dstillery)

Daniel Tunkelang
(LinkedIn)

Kira Radinsky
(SalesPredict)

Eric Jonas
(Independent Scientist)

Yann LeCun
(Facebook)

Anna Smith
(Rent the Runway)

Jake Porway
(DataKind)

André Karpištšenko
(Planet OS)

SEBASTIAN GUTIERREZ

FOREWORD BY PETER NORVIG (GOOGLE)

Contents

CHAPTER

1

Chris Wiggins

The New York Times

Chris Wiggins *is the Chief Data Scientist at* The New York Times *(NYT) and Associate Professor of Applied Mathematics at Columbia University. He applies machine learning techniques in both roles, albeit to answer very different questions.*

In his role at the NYT, Wiggins is creating a machine learning group to analyze both the content produced by reporters and the data generated by readers consuming articles, as well as data from broader reader navigational patterns—with the overarching goal of better listening to NYT consumers as well as rethinking what journalism is going to look like over the next 100 years.

At Columbia University, Wiggins focuses on the application of machine learning techniques to biological research with large data sets. This includes analysis of naturally occurring networks, statistical inference applied to biological time-series data, and large-scale sequence informatics in computational biology. As part of his work at Columbia, he is a founding member of the university's Institute for Data Sciences and Engineering (IDSE) and Department of Systems Biology.

Wiggins is also active in the broader New York tech community, as co-founder and co-organizer of hackNY—a nonprofit organization that guides and mentors the next generation of hackers and technologists in the New York innovation community.

Wiggins has held appointments as a Courant Instructor at the New York University Courant Institute of Mathematical Sciences and as a Visiting Research Scientist at the Institut Curie (Paris), Hahn-Meitner Institut (Berlin), and the Kavli Institute for Theoretical Physics (Santa Barbara). He holds a PhD in Physics from Princeton University and a BA in Physics from Columbia, minoring as an undergraduate in religion and in mathematics.

Wiggins's diverse accomplishments demonstrate how world-class data science skills wedded to extraordinarily strong values can enable an individual data scien-

tist to make tremendous impacts in very different environments, from startups to centuries-old institutions. This combination of versatility and morality comes through as he describes his belief in a functioning press and his role inside of it, why he values "people, ideas, and things in that order," and why caring and creativity are what he looks for in other people's work. Wiggins's passion for mentoring and advising future scientists and citizens across all of his roles is a leitmotif of his interview.

Gutierrez: Tell me about where you work.

Wiggins: I split my time between Columbia University, where I am an associate professor of applied mathematics, and The New York Times, where I am the chief data scientist. I could talk about each institution for a long time. As background, I have a long love for New York City. I came to New York to go to Columbia as an undergraduate in the 1980s. I think of Columbia University itself as this great experiment to see if you can foster an Ivy League education and a strong scientific and research community within the experiment of New York City, which is full of excitement and distraction and change and, most of all, full of humanity. Columbia University is a very exciting and dynamic place, full of very disruptive students and alumni, myself included, and has been for centuries.

The New York Times is also centuries old. It's a 163-year-old company, and I think it also stands for a set of values that I strongly believe in and is also very strongly associated with New York, which I like very much. When I think of The New York Times, I think of the sentiment expressed by Thomas Jefferson that if you could choose between a functioning democracy and a dysfunctional press, or a functioning press and a dysfunctional democracy, he would rather have the functioning press. You need a functioning press and a functioning journalistic culture to foster and ensure the survival of democracy.

I get the joy of working with three different companies whose missions I strongly value. The third company where I spend my time is a nonprofit that I cofounded, called hackNY,[1] many years ago. I remain very active as the co-organizer. In fact, tonight, we're going to have another hackNY lecture, and I'll have a meeting today with the hackNY general manager to deal with operations. So I really split my time among three companies, all of whose mission I value: The New York Times and the two nonprofits—Columbia University and hackNY.

Gutierrez: How does data science fit into your work?

[1] `http://hackNY.org`

Wiggins: I would say it's an exciting time to be working in data science, both in academia and at The New York Times. Data science is really being birthed as an academic field right now. You can find the intellectual roots of it in a proposal by the computational statistician Bill Cleveland in 2001. Clearly, you can also find roots for data scientists as such in job descriptions, the most celebrated examples being DJ Patil's at LinkedIn and Jeff Hammerbacher's at Facebook. However, in some ways, the intellectual roots go back to writings by the heretical statistician John Tukey in 1962.

There's been something brewing in academia for half a century, a disconnect between statistics as an ever more and more mathematical field, and the practical fact that the world is producing more and more data all the time, and computational power is exponentiating over time. More and more fields are interested in trying to learn from data.

My research over the last decade or more at Columbia has been in what we would now call "data science"—what I used to call "machine learning applied to biology" but now might call "data science in the natural sciences." There the goal was to collaborate with people who have domain expertise—not even necessarily quantitative or mathematical domain expertise—that's been built over decades of engagement with real questions from problems in the workings of biology that are complex but certainly not random. The community grappling with these questions found itself increasingly overwhelmed with data.

So there's an intellectual challenge there that is not exactly the intellectual challenge of machine learning. It's more the intellectual challenge of trying to use machine learning to answer questions from a real-world domain. And that's been exciting to work through in biology for a long time.

It's also exciting to be at The New York Times because The New York Times is one of the larger and more economically stable publishers, while defending democracy and historically setting a very high bar for journalistic integrity. They do that through decades and centuries of very strong vocal self-introspection. They're not afraid to question the principles, choices, or even the leadership within the organization, which I think creates a very healthy intellectual culture.

At the same time, though, although it's economically strong as a publisher, the business model of publishing for the last two centuries or so has completely evaporated just over the last 10 years; over 70 percent of print advertising revenue simply evaporated, most precipitously starting around 2004.[2] So although this building is full of very smart people, it's undergoing a clear sea change in terms of how it will define the future of sustainable journalism.

[2] www.aei-ideas.org/2013/08/creative-destruction-newspaper-ad-revenue-has-gone-into-a-precipitous-free-fall-and-its-probably-not-over-yet/

The current leadership, all the way down to the reporters, who are the reason for existence of the company, is very curious about "the digital," broadly construed. And that means: How does journalism look when you divorce it from the medium of communication? Even the word "newspaper" presumes that there's going to be paper involved. And paper remains very important to The New York Times not only in the way things are organized—the way even the daily schedule is organized here— but also conceptually. At the same time, I think there are a lot of very forward-looking people here, both journalists and technologists, who are starting to diversify the way that The New York Times communicates the news.

To do that, you are constantly doing experiments. And if you're doing experiments, you need to measure something. And the way you measure things right now, in 2014, is via the way people engage with their products. So from web logs to every event when somebody interacts with the mobile app, there are copious, copious data available to this company to figure out: What is it that the readers want? What is it that they value? And, of course, that answer could be dynamic. It could be that what readers want in 2014 is very different than what they wanted in 2013 or 2004. So what we're trying to do in the Data Science group is to learn from and make sense of the abundant data that The New York Times gathers.

Gutierrez: When did you realize that you wanted to work with data as a career?

Wiggins: That happened one day at graduate school while having lunch with some other graduate students, mostly physicists working in biology. Another graduate student walked in brandishing the cover of *Science* magazine,[3] which had an image of the genome of *Haemophilus influenzae*. *Haemophilus influenzae* is the first sequenced freely living organism. This is a pathogen that had been identified on the order of 100 years earlier. But to sequence something means that you go from having pictures of it and maybe experiments where you pour something on it and maybe it turns blue, to having a phonebook's worth of information. That information unfortunately is written in a language that we did not choose, just a four-letter alphabet, imagine ACGT ACGT, over and over again. You can just picture a phonebook's worth of that.

And there begins the question, which is both statistical and scientific: How do you make sense of this abundant information? We have this organism. We've studied it for 100 years. We know what it does, and now we're presented with this entirely different way of understanding this organism. In some ways, it's the entire manual for the pathogen, but it's written in a language that we didn't choose. That was a real turning point in biology.

[3] www.sciencemag.org/content/269/5223/496.abstract

When I started my PhD work in the early 1990s, I was working on the style of modeling that a physicist does, which is to look for simple problems where simple models can reveal insight. The relationship between physics and biology was growing but limited in character, because really the style of modeling of a physicist is usually about trying to identify a problem that is the key element, the key simplified description, which allows fundamental modeling. Suddenly dropping a phonebook on the table and saying, "Make sense of this," is a completely different way of understanding it. In some ways, it's the opposite of the kind of fundamental modeling that physicists revered. And that is when I started learning about learning.

Fortunately, physicists are also very good at moving into other fields. I had many culture brokers that I could go to in the form of other physicists who had bravely gone into, say, computational neuroscience or other fields where there was already a well-established relationship between the scientific domain and how to make sense of data. In fact, one of the preeminent conferences in machine learning is called NIPS,[4] and the *N* is for "neuroscience." That was a community which even before genomics was already trying to do what we would now call "data science," which is to use data to answer scientific questions.

By the time I finished my PhD, in the late 1990s, I was really very interested in this growing literature of people asking statistical questions of biology. It's maddening to me not to be able to separate wheat from chaff. When I read these papers, the only way to really separate wheat from chaff is to start writing papers like that yourself and to try to figure out what's doable and what's not doable. Academia is sometimes slow to reveal what is wheat and what is chaff, but eventually it does a very good job. There's a proliferation of papers and, after a couple of years, people realize which things were gold and which things were fool's gold. I think that now you have a very strong tradition of people using machine learning to answer scientific questions.

Gutierrez: What in your career are you most proud of?

Wiggins: I'm actually most proud of the mentoring component of what I do. I think I, and many other people who grow up in the guild system of academia, acquire a strong appreciation for the benefits of the way we've all benefited from good mentoring. Also, I know what it's like both to be on the receiving end and the giving end of really bad and shallow mentoring. I think the things I'm most proud of are the mentoring aspects of everything I've done.

[4] http://nips.cc

Here at the data science team at The New York Times, I'm building a group, and I assure you that I spend as much time thinking hard about the place and people as I do on things and ideas. Similarly, hackNY is all about mentoring. The whole point of hackNY is to create a network of very talented young people who believe in themselves and believe in each other and bring out the best in themselves and bring out the best in each other. And certainly at Columbia, the reason I'm still in academia is that I really value the teaching and mentoring and the quest to better yourself and better your community that you get from an in-person brick-and-mortar university as opposed to a MOOC.

Gutierrez: What does a typical day at work look like for you?

Wiggins: There are very few typical days right now, though I look forward to having one in the future. I try to make my days at The New York Times typical because this is a company. What I mean by that is that it is a place of interdependent people, and so people rely on you. So I try throughout the day to make sure I meet with everyone in my group in the morning, meet with everyone in my group in the afternoon, and meet with stakeholders who have either data issues or who I think have data issues but don't know it yet. Really, at this point, I would say that at none of my three jobs is there such a thing as a "typical day."

Gutierrez: Where do you get ideas for things to study or analyze?

Wiggins: Over the past 20 years, I would say the main driver of my ideas has been seeing people doing it "wrong". That is, I see people I respect working on problems that I think are important, and I think they're not answering those questions the right way. This is particularly true in my early career in machine learning applied to biology, where I was looking at papers written by statistical physicists who I respected greatly, but I didn't think that they were using, or let's say stealing, the appropriate tools for answering the questions they had.

And to me, in the same way that Einstein stole Riemannian geometry from Riemann and showed that it was the right tool for differential geometry, there are many problems of interest to theoretical physicists where the right tools are coming from applied computational statistics, and so they should use those tools. So a lot of my ideas come from paying attention to communities that I value, and not being able to brush it off when I see people whom I respect who I think are not answering a question the right way.

Gutierrez: What specific tools or techniques do you use?

Wiggins: My group here at The New York Times uses only open source statistical software, so everything is either in R or Python, leaning heavily on scikit-learn and occasionally IPython notebooks. We rely heavily on Git as version control. I mostly tend to favor methods of supervised learning rather than unsupervised learning, because usually when I do an act of clustering, which is generically what one does as unsupervised learning, I never know if I've done it the best. I always worry that there is some other clustering that I could do, and I won't even know which of the two clusterings is the better.

But with supervised learning, I usually can start by asking: How predictive is this model that we've built? And once I understand how predictive it is, then I can start taking it apart and ask: How does it work? What does it learn? What are the features that it rendered important?

That's completely true both at The New York Times and at Columbia. One of the driving themes of my work has been taking domain questions and asking: How can I reframe this as a prediction task?

Gutierrez: How do you think about whether you're solving the right problem?

Wiggins: The key is usually to just keep asking, "So what?" You've predicted something to this accuracy? So what? Okay, well, these features turned out to be important. So what? Well, this feature may be related to something that you could make a change to in your product decisions or your marketing decisions. So what?

Well, then I could sit down with this person and we could suggest a different marketing mechanism. Now you've started to refine and think all the way through the value chain to the point at which it's going to become an insight or a paper or product—some sort of way that it's going to move the world.

I think that's also really important for working with junior people, because I want junior people always to be able to keep their eyes on the prize, and you can't do that if you don't have the prize in mind. I can remember when I was much younger—a postdoc—I went to see a great mathematician and I talked to him for maybe 20 minutes about a calculation I was working on, as well as all of the techniques that I was learning. He sat silently for about 10 minutes and then he finally said, "What are you trying to calculate? What is the goal of this mathematical manipulation you're doing?" He was right, meaning you need to be able to think through toward "So what?" If you could calculate this, if you could compute this correlation function, or whatever else it is that you're trying to compute, how would that benefit anything? And that's a thought experiment or a chain of thinking that you can do in the shower or in the subway. It's not something that even requires you to boot up a computer. It's just something that you need to think through clearly before you ever pick up a pencil or touch a keyboard.

John Archibald Wheeler, the theoretical physicist, said you should never do a calculation until you know the answer. That's an important way of thinking about doing mathematics. Should I bother doing this mathematics? Well, I think I know what the answer's going to be. Let me go see if I can show that answer. If you're actually trying to do something in engineering, and you're trying to apply something, then it's worse than that, because you shouldn't bother doing a computation or collecting a data set or even pencil-and-paper work until you have some sense for "So what?" If you show that this correlation function scales to $T^{7/8}$, so what? If you show that you can predict something to 80-percent accuracy on held-out data, so what? You need to think through how it will impact something that you value.

Gutierrez: What's an interesting project that you've worked on?

Wiggins: One example comes from 2001 when I was talking to a mathematician whom I respect very much about what he saw as the future of our field, the intersection of statistics and biology, and he said, "Networks. It's all going to be networks." I said, "What are you talking about? Dynamical systems on networks?" He said, "Sure, that and statistics of networks. Everything on networks."

At the time, the phrase "statistics of networks" didn't even parse for me. I couldn't even understand what he was saying. He was right. I saw him again at a conference on networks two years later.[5] Many people that I really respected spoke at that conference about their theories of the way real-world networks came to evolve.

I remember stepping off the street corner one day while talking to another biophysicist, somebody who was coming from the same intellectual tradition that I had with my PhD. And I was saying, "People look at real-world networks, and they plot this one statistical attribute, and then they make up different models—all of which can reproduce this one statistical attribute." And they're basically just looking at a handful of predefined statistics and saying, 'Well, I can reproduce that statistical behavior.' That attribute is over-universal. There are too many theories and therefore too many theorists saying that they could make models that looked like real-world graphs. You know what we should do? We should totally flip this problem on its head and build a machine learning algorithm that, presented with a new network, can tell which of a few competing theorists wins. And if that works, then we're allowed to look at a real-world network and see which theorist has the best model for some network that they're all claiming to describe."

That notion of an algorithm for model testing led to a series of papers that I think were genuinely orthogonal to what anybody else was doing. And I think it was a good example of seeing people whom I respect and think are very smart people but who were not using the right tool for the right job, and then trying to reframe a question being asked by a community of smart people as a prediction problem. The great thing about predictions is that you can be wrong, which I think is hugely important. I can't sleep at night if I'm involved in a scientific field where you can't be wrong. And that's the great thing about predictions: It could turn out that you can build a predictive model that actually is just complete crap at making predictions, and you've learned something.

[5] http://cnls.lanl.gov/networks

Gutierrez: How have you been able to join that point of view with working at a newspaper?

Wiggins: It's actually completely the same. Here we have things that we're interested in, such as what sorts of behaviors engender a loyal relationship with our subscribers and what sorts of behaviors do our subscribers' evidence that tends to indicate they're likely to leave us and are not having a fulfilling relationship with *The New York Times*. The thing about subscribers online is that there are really an unbounded number of attributes you can attempt to compute. And by "compute," I really mean that in the big data sense. You have abundant logs of interactions on the web or with products.

Reducing those big data to a small set of features is a very creative and domain-specific act of computational social science. You have to think through what it is that we think might be a relevant behavior. What are the behaviors that count? And then what are the data we have? What are the things that can be counted? And, of course, it's always worth remembering Einstein's advice that not everything that can be counted counts, and not everything that counts can be counted. So you have to think very creatively about what's technically possible and what's important in terms of the domain to reduce the big data in the form of logs of events to something as small as a data table, where you can start thinking of it as a machine learning problem.

There's a column I wish to predict: Who's going to stick around and who's going to leave us? There are many, many attributes: all of the things that computational social science, my own creativity, and very careful conversations with experts in the community tell me might be of interest. And then I try to ask: Can I really predict the thing that I value from the things that the experts believe to be sacred? And sometimes those attributes could be a hundred things and sometimes that could be hundreds of thousands of things, like every possible sequence element you could generate from seven letters in a four-letter alphabet. Those are the particular things that you could look at.

That is very much the same here as it is in biology. You wish to build models that are both predictive and interpretable. What I tell my students at Columbia is that as applied mathematicians, what we do is we use mathematics as a tool for thinking clearly about the world. We do that through models. The two attributes of a model that make a model good are that it is predictive and interpretable, and different styles of modeling strike different balances between predictive power and interpretability.

A few Decembers ago, I had a coffee with a deep learning expert, and we were talking about interpretability, and he said, "I am anti-interpretability. I think it's a distraction. If you're really interested in predictive power, then just focus on predictive power." I understand this point of view. However, if you're interested in helping a biologist, or helping a businessperson, or helping a product person, or helping a journalist, then they're not going to be so interested in .08

error on held-out data. They're going to be interested in the insights and identification of the interesting covariates, or the interesting interactions among the covariates revealed to you.

I come from a tradition in physics that has a long relationship with predictive interpretability. We strive to build models that are as simple as possible but not simpler, and the real breakthroughs, the real news-generating events, in the history of physics have been when people made predictions that were borne out by experiment. Those were times that people felt they really understood a problem.

Gutierrez: Whose work is currently inspiring you?

Wiggins: It's always my students. For example, I have a former student, Jake Hofman, who's working with Duncan Watts at Microsoft Research. Jake was really one of the first people to point out to me how social science was birthing this new field of computational social science, where social science was being done at scale. So that's an example of a student who has introduced me to all these new things.

I would also say that all of the kids who go through hackNY are constantly introducing me to things that I've never heard of and explaining things to me from the world that I just don't understand. We had a hackNY reunion two Friday nights ago in San Francisco. I was out there to give a talk. We organized a reunion, and the Yo app had just launched. So a lot of the evening was me asking the kids to explain Yo to me, which meant explaining the security flaws in their API and not just how the app worked. So that's the benefit of working with great students. Students are constantly telling you the future of technology, data science, and media amongst other things, if you just listen to them. Former students and postdocs of mine have gone on to work at BuzzFeed, betaworks, Bitly, and all these other companies that are at the intersection of data and media.

I have also benefited greatly from really good colleagues whom I find inspiring. The way I ended up here at The New York Times, for example, was that, when I finally took a sabbatical, I asked all my faculty colleagues what they did with their sabbaticals, because I had never taken one. My friend and colleague Mark Hansen did the "Moveable Type" lobby art here in the New York Times Building. So if you go look at the art in the lobby, Mark Hansen wrote the Python to make the lobby art "go", and he did that in 2007 when they moved into this building. So he knew many people at The New York Times, and he introduced me to a lot of people here and was somebody who explained to me—though he didn't use these words—that The New York Times is now in a similar state to the state that biology was in 1998. That is, that it's a place where they have abundant data, and it's still up for grabs what the right way is to use machine learning to make sense of those data.

Mark Hansen is a good example of somebody who's done great work. In fact, although he won't admit it, he was using the phrase "data science" throughout the last 12 years. He's been writing for years about what he often called "the science of data." He's been somebody who's been really thinking about data science as a field much longer than most people. Actually, he worked with Bill Cleveland at AT&T. Bill Cleveland, in turn, had worked with Tukey, so there is a nice intellectual tradition there. There's a reason why data science resonates so much with academics. I feel it's because there's been an academic foundation there in the applied computational statistics community for half a century.

David Madigan, who's the former chair of stats at Columbia, is also inspiring. He is somebody who's done a great job showing the real impact of statistics—good and bad—on people's lives. All the people I respect are people who share my value for community. Mark Hansen is trying to build a community of data journalists at the journalism school. His PhD was in statistics, but now he's a professor of journalism who is trying to build a community of data journalists. David Madigan similarly—he was the chair of statistics and now he's the Executive Vice President for Arts and Sciences at Columbia.

The people I find the most inspiring are the people who think about things in this order: people—in terms of how you build a strong community; ideas—which is how you unite people in that community; and things that you use to build the community that embodies those ideas.

But mostly, I would say my students—broadly construed, at Columbia and at hackNY—who inspire me.

Gutierrez: What was it that convinced you to join The New York Times and try to make a difference when you did your sabbatical?

Wiggins: It was clear to me by the end of my first day here that we should build a predictive model for looking at subscriber behavior. I spent some time interviewing or meeting everyone around here in the company who I felt was likeminded. I found some good collaborators, worked on this project, and it was clear from the way people reacted to it that no one had done that before. I did that without a real clear sense of whether or not I was reinventing a common wheel.

I got the impression from the way people reacted that people had been sort of too busy feeding the goat, meaning doing their daily obligations of running a company, even worse in journalism. In journalism, you have constant deadlines, but even on the business side, there's a business to run. Nobody has time to do a two-month research project. I think that's what convinced me that there really was a lot to be learned from the data that this company is gathering and curating.

Gutierrez: What do you look for in other people's work?

Wiggins: Creativity and caring. You have to really like something to be willing to think about it hard for a long time. Also, some level of skepticism. So that's one thing I like about PhD students—five years is enough time for you to have a discovery, and then for you to realize all of the things that you did wrong along the way. It's great for you intellectually to go back and forth from thinking "cold fusion" to realizing, "Oh, I actually screwed this up entirely," and thus making a series of mistakes and fixing them. I do think that the process of going through a PhD is useful for giving you that skepticism about what looks like a sure thing, particularly in research. I think that's useful because, otherwise, you could easily too quickly go down a wrong path—just because your first encounter with the path looked so promising.

And although it's a boring answer, the truth is you need to actually have technical depth. Data science is not yet a field, so there are no credentials in it yet. It's very easy to get a Wikipedia-level understanding of, say, machine learning. For actually doing it, though, you really need to know what the right tool is for the right job, and you need to have a good understanding of all the limitations of each tool. There's no shortcut for that sort of experience. You have to make many mistakes. You have to find yourself shoehorning a classification problem into a clustering problem, or a clustering problem into a hypothesis-testing problem.

Once you find yourself trying something out, confident that it's the right thing, then finally realizing you were totally dead wrong, and experiencing that many times over—that's really a level of experience that unfortunately there's not a shortcut for. You just have to do it and keep making mistakes at it, which is another thing I like about people who have been working in the field for several years. It takes a long time to become an expert in something. It takes years of mistakes. This has been true for centuries. There's a quote from the famous physicist Niels Bohr, who posits that the way you become an expert in a field is to make every mistake possible in that field.

Gutierrez: What's been the biggest thing you've changed your mind about?

Wiggins: That's a tough choice. There are so many things that I've changed my mind about. I think probably the biggest thing I've changed my mind about is the phrase that you can't teach an old dog new tricks. I think if you really care about something, you'll find a way. You'll find a way to learn new tricks if you really want to.

The other thing that I've changed my mind about is that I grew up, like most academics, with the sense that scientists somehow functioned with some orthogonal value system that was different than the world. I think one thing that I've changed my mind about in this area—but this is over like a 20-year period since I was not yet a PhD, is that scientists are human beings too, whether they know it or not. And that science is done by scientists, and

scientists are human beings. And so all the good and the bad about humans, and how they make their choices, and what they value, carries over to the scientific and academic enterprise. It's not different. It's a lovely guild and it's functioned fantastically for centuries, and I hope it continues to function for a long time because I think it has been very good for the species, but we shouldn't believe that scientists are somehow not subject to the same joys and distractions as every other human being.

So that's part of what I learned—that science is somehow not a qualitatively different enterprise than, let's say, technology or any other difficult human endeavor. These are very difficult human endeavors, and they take planning, and attention, and care, and execution, and they take a community of people to support it. Everything I just said is completely true of academic science, writing papers, winning grants, training students, teaching students, as well as forming a new company, doing research, or using technology in a big corporation that's already been established. All of those things are difficult and require a community of people to make it happen. As they say: "people, ideas, and then things, in that order" That's true in any science and that's also true in the real world.

Gutierrez: What does the future of data science look like?

Wiggins: I don't see any reason for data science not to follow the same course as many other fields, which is that it finds a home in academia, which means that there becomes a credentialing function, particularly around professional subjects. You'll get master's degrees and you'll get PhDs. The field will take on meaning, but it will also take on specialization. You see this already with people using the phrases "data engineering" and "data science" as separate things. My group here at The New York Times is the Data Science group, which is part of the Data Science and Engineering larger group. People are starting to appreciate how a data science team involves data science, data engineering, data visualization, and data architecture.

Data Product is not sort of a thing yet, but certainly, if you look at how, say, data science happened at LinkedIn—data science reported up through the product hierarchy. At other companies, data science reports through business; or it reports through engineering. Right now I'm located within in the engineering function of The New York Times, separate from the product, separate from marketing, and separate from advertising. Different companies are locating data science in different arms.

So I think there'll be credentialing. I think there will be specialization. New fields are born—I wouldn't say all the time, because by real-world standards, nothing ever happens in academia—but there are new departments born at universities every few years. It happens, and the way that it happens is part of the creation of new fields. I'm old enough that I had the benefit of watching, say, systems biology be born as a field, synthetic biology be born as a field,

and even nanoscience be born as a field in the time that I've been a practicing academic. My first research project in the 1980s was in chaos, which at that time was being born as a new field. There's a famous book on this by James Gleick, at that time writing for The New York Times, called *Chaos: Making a New Science*.[6] It's not that new fields aren't created in academia. It's just that it's so damn slow compared to the pace of the real world, which I think is really for the best. There are young people's futures at stake, so I think it's actually not so bad.

So I think the future of data science is for it to become part of academia, which means a vigorous, contentious dialog among different universities about what is really data science. You're already starting to see work in this direction. For instance, at Columbia, a colleague of mine, named Matt Jones, who's an historian, is writing a book about the history of machine learning and data science. So you're already starting to see people appreciate that data science wasn't actually created from a vacuum in 2008. Intellectually, the things that we call data science had already been sort of realized—that is, that there was a gap between statistics and machine learning, that there was sort of something else there. So I think there will be a greater appreciation for history.

Part of what happens when a field becomes an academic field is that three main things occur—an academic canon is set, a credentialing process is initiated, and historical study provides the context of the field. An academic canon is the set of classes that we believe are the core intellectual elements of the field. The credentialing process, which is another separate function from academia, which can be unbundled, is initiated so you can get master's and PhD degrees. Lastly, historical study occurs to appreciate the context: Where did these ideas come from?

As the names and phrases people use become more meaningful, then you get the possibility of specialization, because what we have now is that when people say "data science" they could mean many things. They could mean data visualization, data engineering, data science, machine learning, or something else. As the phrases themselves become used more carefully, then I think you'll get to see much more productive specialization of teams. You can't have a football team where everybody says, "I'm the placekicker." Somebody needs to be the placekicker, somebody needs to be the holder, and somebody needs to be the linebacker. And as people start to specialize, then you can pass. You can have meaningful collaborations with people because people know their roles and know what "mission accomplished" looks like. Right now, I think it's still up for grabs what a win in data science really looks like.

[6]James Gleick, *Chaos: Making a New Science* (Viking, 1987).

Again, I come from a very old field. Physics is a field where the undergraduate curriculum was basically canonized by 1926. Years ago I picked up a book at the Book Scientific bookstore called *Compendium of Theoretical Physics.*[7] It had four chapters: classical mechanics, statistical mechanics, quantum mechanics, and E&M—electricity and magnetism. Those are the four pillars on which all of physics stands. And physics has a pretty rich intellectual tradition, with some strong clear wins behind it, but it's really built on those four pillars. You can see that it has a strong canon. Most fields don't enjoy that. I think you really need to have a well of a mature field for you to be able to say, "Here are the four classes that you really need to take as an undergraduate."

Gutierrez: What does the academic canon at the Institute for Data Sciences and Engineering at Columbia cover?

Wiggins: I'm on the education committee for the Data Science Institute at Columbia, so we've created a canon of four classes: Probability and Statistics, Algorithms for Data Science, Machine Learning for Data Science, and EDAV, which is short for Exploratory Data Analysis and Visualization. The three letters, *EDA*, are taken directly from John Tukey.

Tukey had a book in the 1970s called *Exploratory Data Analysis*—which was basically a description of what Tukey did without a computer, probably on the train between Princeton and Bell Labs, whenever somebody gave him a new data set.[8] The book is basically a description of all the ways he would plot out the data, histograms, Tukey boxplots, Tukey stem-and-flower plots—all these things that he would do with data. If you read the book now, it looks like, "man, this guy was kooky. He should have just opened up R. He should have just opened up matplotlib."

Around the same time, he was co-teaching a class at Princeton with Edward Tufte. If you pick up the book *Visual Display of Quantitative Information,* look at whom it's dedicated to.[9] It's dedicated to Tukey. Again, there's a very old academic tradition on which many of the data science ideas lie. People have been thinking in academia for a long time about what the visual display of quantitative information is. How do we meaningfully "do" data visualization? What do we do when someone hands us data and we just have no distribution? The world doesn't hand you distributions. It hands you observations.

---

[7] `www.wachter-hoeber.com/Books.html?bid=002`

[8] John W. Tukey, *Exploratory Data Analysis* (Pearson, 1977).

[9] Edward R. Tufte, *The Visual Display of Quantitative Information* (2nd ed.) (Graphics Press, 1983).

Much of what we do in physics or mathematical statistics organizes our worldview around what the appropriate model is. Is this the time when I should treat it as statistical mechanics and, if so, what terms do I put in my Hamiltonian? Is it the case that this is a quantum mechanical problem? If so, what terms do I put in my Hamiltonian? Is this a classical mechanics problem? If so, what terms should I put in my Hamiltonian?

The world's like that. The world doesn't hand you models. It doesn't come to you with a model and say, "Diagonalize this Hamiltonian."[10] It comes to you with observations and a question usually being asked by the person who gathered those data. So that's the tradition that I thought was important enough that we make one of our four pillars of data science at Columbia. We want students to think about how we explore data before we decide that we're going to model it using some particular distribution or some particular graphical model. How do you explore a data set that you've been handed?

Gutierrez: What are the most exciting things in data science for you?

Wiggins: The things that are most exciting to me are not new things. The most exciting thing to me is realizing that something everybody thinks is new is actually really damn old. That's why I like Tukey so much. There's a lot of excitement about this new thing called "data science." I think it's really fun to go see really old papers in statistics that are even older than Tukey. For instance, Sewall Wright was using graphical models for genetics in the 1920s.[11] The things that really capture my excitement are not the newfangled things. It's particularly around the ideas, not so much things, because, again—people, ideas, and things in that order. The things change. It's fun when we think we have a new idea, but usually we then realize the idea is actually very old. When you have an understanding of that, it's a really frickin good idea.

Stochastic optimization and stochastic gradient descent, for example, has been a huge, huge hit in the last five years, but they descend from a paper written by Robbins and Monro in 1951.[12] It is a good idea, but the fact that I think it's a good idea means somebody really thought through it very carefully with pencil on paper a long time back. Trying to understand the world through data and your computer is a very good idea. That's why Tukey was writing about it in 1962 when he was ordering everybody to reorient statistics as a professional discipline and a funding line for the NSF organized around computation and data and data analysis. He wrote an article in 1962 called "The Future of Data Analysis."[13] And he wasn't the last, right?

---

[10] `http://vserver1.cscs.lsa.umich.edu/~crshalizi/reviews/fragile-objects/`
[11] Wright, Sewall. "Correlation and causation." *Journal of Agricultural Research* 20.7 (1921), 557-585.
[12] Herbert Robbins and Sutton Monro, "A Stochastic Approximation Method": *Ann. Math. Statist.*, Volume 22, Number 3 (1951), 400-407.
[13] John W. Tukey, "The Future of Data Analysis": *Ann. Math. Statist.*, Volume 33, Number 1 (1962), 1-67.

Leo Breiman all throughout the 1990s was writing to his community of statisticians, "Let us get with data, statistics community!" He was writing papers in the late 1990s telling all his colleagues to start going to NIPS.[14] It was like he had gone into the wilderness and come back and said to everybody at Berkeley, which was one of the first mathematical statistics departments, "You guys need to wake up because it's on fire. You guys are proving theorems. It's on fire out there. Wake up!"

So I think there was a strong tradition of people understanding how powerful and how different it was to understand the world through data. The "primacy of the data" was a phrase that one of the mathematical statisticians at Berkeley used a long time back for Tukey's emphasis.[15] This strong tradition carried on through this sort of heretical strain of thought from John Tukey through Leo Breiman to Bill Cleveland in 2001. All of them saw themselves as orthodox statisticians, though they were people who were sufficiently heretical. It's just that as statistics kept doubling down on mathematics every five years because of their origin from math that made statistics a bona fide field, you found this strain of heretics who were saying, "No, you should really try to get with data." That's what I think is most exciting in terms of people, ideas, and things—don't be distracted by today's things but find the people and their ideas that are actually much older.

[14]https://www.stat.berkeley.edu/~breiman/wald2002-1.pdf
[15]Erich L. Lehmann, *Reminiscences of a Statistician: The Company I Kept* (NY: Springer Science+Business Media, 2008), 198.