Ethnography for a data-saturated world

(2018)

Edited by Hannah Knox and Dawn Nafus

Manchester University Press

Contents

11

The other ninety per cent: thinking with data science, creating data studies – an interview with Joseph Dumit

Joseph Dumit and Dawn Nafus

Editor's note: This is a jointly edited transcript of an interview with Joseph Dumit (professor of Science & Technology Studies and Anthropology) about the Data Studies undergraduate minor being designed at University of California at Davis. This programme began in late 2015, and is led jointly by Dumit and Duncan Temple Lang, director of the Data Science Initiative at UCD, professor of Statistics, and formerly of Bell Labs.

DN: How did Data Studies become an interest of yours?

JD: The immediate genesis was meeting with an alumnus, Tim McCarthy, who had been a social science major and went on to work in senior positions at a series of international banks and financial institutions. He was concerned that Liberal arts majors were declining, even though it was the critical thinking skills of the liberal arts that were incredibly valuable in his career. He was concerned that, when he was starting out, companies hired liberal arts majors and then trained them for one to two years. But today, companies can't afford to train anyone for more than a couple of weeks, even though they do want critical thinking skills. This was juxtaposed with my observations that many students in the social sciences and humanities are interested in how technology is changing the world, and have critiques of it, but for various reasons are never getting their hands

on data. Partly that's because they think it's math, or they think the bar to entry is learning how to program.

When we met, we discovered a shared interest in teaching data skills, and critical thinking around data as a means to, at minimum, get them jobs. We wanted to make sure this was doable, however, so a team of us visited senior people at over twenty large and small Silicon Valley companies, banks, a grocery chain company, consulting agencies etc., asking them what they would like to see out of liberal arts majors. They all said that students do need to know Excel – it's a legacy standard that companies can't really get out of – but they also wanted people who, if you asked them a question, could tell you that you were asking the wrong question. They were hiring recent computer science and statistics graduates as data analysts, but one complaint was that, if you gave many young data analysts a problem, they could too easily give you an answer. They would not necessarily take the time to reflect and say, 'Maybe this is not the right question'. The companies were very interested in being able to also get critical thinkers who talk the language of data science, and who know what a data science question is, but who don't necessarily have to know the algorithms in order to engage with them.

Then Duncan Temple Lang and I discussed what it would take to design a set of classes geared toward liberal arts majors (social sciences and humanities students) to learn enough about data handling and data science to be able to apply their critical thinking skills to work in government, industry or NGOs *alongside* data scientists. This would be a minor that would complement, not replace, their major: Data Studies.

DN: DID ANY OF THESE COMPANIES GIVE YOU AN EXAMPLE OF WHY REFRAMING THE QUESTION HAD BUSINESS VALUE?

JD: One company had engineers and data scientists designing front-end interfaces for managing hundreds of databases at a time, and what they ended up with was an interface full of all the information *available*, but not the information that was *needed*. Another example was a bank trying to optimise credit scores, and they needed to take into account the regulatory context (see also O'Neil 2016). Given a particular set of data, there might be an algorithm that would optimise the selection of good credit versus bad credit given that dataset. The regulatory context is constantly changing, so what you can and can't

do is something you have to take into account as you optimise the answer, otherwise you might optimise the answer in an illegal or unfair way. That means you have to ask: why do we have this dataset versus another, and are there biases in that dataset so we can tweak it a little, or seek more data?

DN: YOU PROPOSED TO ME THAT WE SHOULD TALK ABOUT 'THE OTHER NINETY PER CENT". WHAT DO YOU MEAN BY THAT?

JD: In order to design classes for Data Studies, I'd been reviewing textbooks for data science, looking at online courses etc., and at the beginning of each, there was a nice acknowledgement that 'figuring out the question' and 'data cleaning' were the most important things that a data scientist does, and it often takes eighty to ninety per cent of their time. But then the book will only spend ten per cent of the contents on cleaning, ninety per cent about algorithms and programming, and almost nothing on clarifying the question. (There are, however, a couple books (O'Neil and Schutt 2013; Peng and Matsui 2015) that do devote a whole chapter to clarifying and cleaning). 'The other ninety per cent' challenges the irony where textbooks are saying that ninety per cent of your time is going to be spent on one thing, but then focus all of their effort on something else.

Data science is in this funny space. Duncan had started his career in the Statistics and Data Mining group at Bell Labs, where exploratory data analysis (EDA) and figuring out what the real problem was and what data matters was how data scientists spent their time. That was a place where these skills were as important as algorithm-tuning. Duncan was one of core developers of the statistical computing language R, designed precisely to focus attention on thoughtfully exploring data visually as well as numerically (see Mackenzie 2007). He had been writing articles and books for over 17 years about refocusing classroom time on giving students a feel for the data (Nolan and Temple Lang 2015). He had been flipping curriculum using case studies and online tools like Piazza to put students into a collective learning environment. In fact one of the hardest skills to teach was one that all good programmers use: being willing to ask questions online and make use of collective wisdom when wrestling with data and tools.

Occasionally experts do talk about experience – that if you do this long enough you start to get better at asking questions. But the idea that there are actual fields where people are trained to think critically and expansively about interpretation, and to be rigorous about ambiguous situations rather than avoiding them, is absent. These books and courses don't mention, for example, how you should understand the range of your stakeholders – not just those who are invested in your results but those who are potentially affected by the way you analyse and present your data.[1] They might give one example of how a particular stakeholder is important, but the idea that there's a whole discipline devoted to mapping stakeholders and understanding social implications and feedback effects is avoided. Anthropologists and sociologists are trained in how to interview, and how to ask questions that might not just lead stakeholders to the answer they think you want but allow them to reveal to you what they actually care about. Of course, in some amazing world, people would actually just tell you what they want, but sometimes they don't know, or sometimes the right answer is at the intersection of multiple stakeholders. In these situations you do not necessarily need long interviews, but you do need careful interviews with them.

Cleaning has its own sets of mysteries to deal with. We started the introductory class by making students answer a survey of 30 questions by filling in text boxes. We asked them their name, their year in school, how many friends they had on Facebook, sibling rank, favourite movies, shoe size etc. And then we give them their own collective data back, and said: clean this so we can ask questions about possible trends in the class and whether we can use the survey to potentially predict things. They immediately saw that the way most of them answered the questions did not make cleaning the data easy. They learned that they should regularise how people input shoe sizes, for example. Sometimes with sibling rank, students from Asian countries will put their rank among male children if they are male, or among female if female. It didn't occur to them to rank all the children together, just like it didn't occur to other students that rank might be separated by gender. As they tried to clean up the data to be able to quantify it, they came to see that every decision they make about cleaning the data is going to affect them or their friends directly. There's no right answer in most cases, so we have a discussion about

that, and about how each form of cleaning is 'political', meaning that each decision affects the future of how the data will be understood because it will delete some data that others might consider important, collapse some differences and emphasise others.

DN: How do your students encounter this broader notion of 'bias', when societally there are much narrower notions of what bias is?

JD: In one approach, we used Latanya Sweeney's (2013) work on bias in algorithms that pull up Google Adwords that can be racially biased even though the algorithm probably did not have a human behind it except someone who decided that first name was a variable to throw into a giant machine-learning algorithm. Sweeney's discussion of the ethics of algorithms was incredibly helpful for students (Brennan 2015). We also used Cathy O'Neill's presentation to TalksAtGoogle (TalksAtGoogle 2016) and her work on *Weapons of Math Destruction* (O'Neil 2016) describing the work that ProPublica and others had conducted on policing and sentencing to get them to understand that something could be *both objective and unfair.* For instance, if arrests are racially biased, then predictions based on crimes will be racially biased. 'Objective' was always indexed to the dataset as it was gathered. It was indexed to drawing a box around a problem and saying 'Given this data or this criteria, this is an objective answer.' The question was not whether it was biased, because all samples and all collection mechanisms are biased in the sense that they have made choices, but she emphasises that what matters is whether it is unfair in terms of the outcome. If the outcome is unfair, you would reconsider the algorithm and the particular form of objectivity being built in.

We could then extend this to seemingly different workflows. If a company is comparing sales and bonus from different regions, the way in which sales are counted ends up being political and potentially unfair. The people who get urban regions versus rural regions can get penalised in different ways, and there is always fighting over what is going to be the metric. These metrics are also data-formulating and cleaning problems, because there will always be outliers and someone who says 'You can't count that one the same way as this one.' It's important for students to understand that here is where decisions are made, and, once you start putting algorithms on it, you are going to carry forward any choice you made. If you haven't explored the data

and just run with it, you are really building in a type of bias that no one will necessarily notice. To not have people spending serious time exploring data is, from an employer point of view, dangerous. Maybe that person has to gather together the stakeholders to say, what do we do about these differences in how things are counted or with our outliers? It's in our training data, or base set, so it's going to bias our future.

So the first step was getting them to see that there is more than one 'logical' way to do something, and more than one 'objective' answer. Some students end up still thinking that bias is only race, but they do get to the point where they can see that racial bias is made up of a bunch of choices to do with initial data choices, question formulation and cleaning, and doesn't necessarily have to do with how the algorithm was designed. Rather bias often has to do with how data are not independent of the algorithms that they are put into. The next stage is showing that all data is biased and that is an important starting point, which they all got when we worked with the airline data. But you are right that on the test there were students who defaulted to the idea that bias equals race or other key social categories. My hope is that is that, with more classes that emphasise how to think critically, with and about data and unfairness, they can learn to extend these arguments.

When we take it to actual datasets, like an airline flight delay dataset, they start to see how even something that looks very neutral, when you dig into it, you see layers and layers of politics. I was trying to teach them the concept that politics here is just making decisions where it affects others, to the extent that that algorithm gets employed in some way. Politics isn't always 'bias', but it is a change. Or rather, it is bias but it's not evil – it's inevitable. It's a choice. It's more important to be able to explain the results to the public, or to an employer, and say why you made the decisions you did. This means leaving a trail of why I cleaned it this way, and that I did indeed clean it. Otherwise it's just magic; it's just, 'I have an answer' and no one can go back and say 'Why not clean it this way?'

DN: TELL US MORE ABOUT THAT AIRLINE DATASET. WHY WAS IT GOOD TO WORK WITH?

JD: This dataset is a favourite of data scientists because it is large – over five million domestic flights per year for starters. It's a database

maintained by the US Bureau of Transportation. Every flight is one entry in a giant spreadsheet with over a hundred measurements (Excel columns). It includes data about when the flight took off, how long it was on the runway, when they closed the doors, when it was supposed to land, when it actually landed, how much it was delayed or early, and if delayed why. Our whole course was designed around cases with actual data in all of their messiness, and having students imagine having stakeholders, and work through how different stakeholders might want different types of information. It could be someone who is trying to fly back and forth between San Francisco and New York every week, and wants to know which airline or flight has the least chance of being delayed. Or, it could be American Airlines wanting to know which categories it is beating United in, so they can make a campaign around it. It could be someone taking a vacation next year, and that person might be interested in lost luggage, but the data doesn't have that. So if they say they really don't want to lose their luggage, the answer is to find another dataset.

DN: IS THAT WHAT YOU MEAN BY TEACHING THEM HOW DATA IS ALREADY POLITICAL – IT'S DESIGNED FOR THE NEEDS OF CERTAIN PEOPLE, BUT NOT OTHER PEOPLE?

JD: Yes. Furthermore, we, and anyone analysing or presenting data, is making decisions about who our audience is and what they care about. So we 'censor' data, or, specifically, results in order to address what is important to our 'audience'. We hope our choices reflect the audience's goals, as well as social fairness, because there are always multiple audiences. Except in the case of a classroom test!

We started by teaching the students pivot tables in Excel. Most students' computers can handle up to a million rows in Excel, so we did a sub-set of the flights for one year from six California cities and New York cities (approximately three hundred thousand flights). We gave them enough data that they could ask questions about it. We had them first try to sit and examine the data and ask, why are there so many columns? Why is there a column for how many minutes it is delayed or early, but another column identical to that one that had zeroed out all the negative numbers so there is no 'early'? Then there was another column that showed a one or a zero depending on whether it was delayed 15 minutes or more.

As they started posing questions, they realised that each of these columns was differently useful. For example, if you averaged 'arrival time' one airline might appear much better than another one, but then you notice that it includes negative numbers for flights that arrive early. If what you meant to study was which airline had the fewest delays, especially the least long delays, then the column with delays longer than 15 minutes was far more immediately useful. We could see how the bureau started adding columns to this data so others didn't have to create them. If you were creating them on a million or a billion rows, it would take a long time. Here they just give it to you, and so the columns themselves reveal a history about who uses this data and what kinds of questions they pose to it.

Then there is the question of why is one of the columns '15 minutes or more delayed'. We could think about who thresholded it at 15 minutes and why. We wanted to ask whether '30 minutes or more' might be meaningful to someone, but we'd have to make a new column for that. We noticed that there were a lot of early flights, too. Analysts at OAG Aviation Worldwide explored the data and figured out that either the Earth is getting bigger or the projected flight time from San Francisco to New York gets longer, which means that the airlines are listing their arrival time later and later so that more flights are less than 15 minutes late (Morris 2015). Once you put a number like 15 minutes into a data threshold, and turn it into tables that are used to publicise who is on time the most, you see Goodhardt's Law at work, which says that any time an index becomes a target, it no longer functions as an index (see Strathern 1997; Dumit 2016; Griesemer 2016). From a customer point of view, it's nice because I would rather the airline say it takes longer, and know that I have an hour to make my connection. It's good for me, and good for the airline, but it means comparing data across years becomes an interesting challenge, because the airlines are changing the flight times.

Then we get to the real cleaning question. The data is presented apparently quite clean, meaning that every box that is supposed to have data has something in it, unlike when you are at a company or in an environmental group where you get lots of missing data (see Ribes and Jackson 2013). We were trying to ask, if you have data from 2012, could you predict something about 2013? Which airline has the fewest delays for certain routes? We have them do exploratory data analysis (EDA), which is something Duncan Temple Lang often

talks about (Tukey 1977). This starts with actually sitting in front of your data, looking at and thinking about it, and visualising it in many different ways. I think of this as a non-Tufte-style visualisation. Edward Tufte has a powerful approach to visualisations which show how good a map or a chart is as a measure of the distance between what someone who hasn't seen it before gets from it and what the intended meaning was. That's important. If I am the engineer or the social scientist and I know what my data means, and now I want to find the best way to graph it so that somebody else gets my meaning, I measure how much they get out of it by comparing what they understood with what I am trying to convey. Did they get enough, or too much or not the right thing? Exploratory data analysis is totally different. The analyst is the recipient of their own data visualisation and they don't know what the data means except through visualising it multiple times in lots of different ways. They try a bunch of different things to see if there's something that could be a hypothesis, and then they dig in and ask why.

DN: ONE THING I STRUGGLE WITH AS SOMEONE WHO IS STILL LEARNING THIS IS THAT THERE ARE SOME PATTERNS THAT ARE CLEAR IN MY MIND BECAUSE I SEE THEM SO OFTEN, LIKE RECURRENCES BY HOUR, OR BY LOCATION. THEY BECOME MY GO-TO PATTERNS, AND I'M ALWAYS SURPRISED WHEN SOMEBODY COMES UP WITH ONE THAT IS CONCEPTUALLY SIMPLE BUT NOT YET IN MY REPERTOIRE. I WOULDN'T EXPLORE IT IN THOSE NEW WAYS BECAUSE I DIDN'T KNOW THAT WAS EVEN A WAY TO EXPLORE. HOW DO YOU SEED THE POSSIBILITIES WITH YOUR STUDENTS?

JD: Partly we do this through having them sketch on paper what they thought relations would be [editor's note: see also Lupi and Posavec 2016]. This is so they don't get habitual. One of the dangers of most tools like Excel and Tableau is that they come with pull-down menus of charts that make a lot of choices for you, and worse, 'Recommended Charts'. The problems are (1) most of the time these charts *look good* so that, even if they don't make sense of the data, it is easy to stop and hit *print*; (2) when they choose your axes and colour schemes for you, you don't necessarily notice that they made all of these choices; and (3) therefore they don't encourage thinking about what a useful chart might be for yourself. Thus one of our exercises

is always to sketch by hand charts you might find helpful as a way to think out loud about what your data means. Another practice is for our students to post their exploratory charts to one another, so they can be inspired by each other. We try to have them work in teams, so they can bounce ideas off one another. Each person tends to have partial ideas, based on their ongoing hypotheses, and they may think it's obvious and therefore don't think it is worth sharing. But when they do say or show it, someone is often inspired.

This approach is also enhanced by thinking through the lenses of different stakeholders. For example, if you care about a particular person who's flying regularly, certain things are more likely to come to mind as being important. When you think about United versus American Airlines trying to do a marketing campaign, or the ground crew and what they might be worried about, you start thinking about different kinds of relations that you then put together in a graph, and different temporal signatures might come to matter.

For example, as we were exploring and looking at the data for 2012 and delays by month, November looked like this amazing month because it had very few delays relatively speaking. But then we did a graph that showed there were six days in November when there were no delays at all for New York airports. So we looked at it and asked 'What's that? That never happens!' We looked at cancellations, and sure enough every flight was cancelled. So then, what do you do? What's going on? Because it is real data, you can Google '2012 flight delays November New York' and what you find is Hurricane Sandy. And so now you have a cleaning issue: does the question that I am trying to answer with the data in front of me need to be cleaned of Hurricane Sandy? If I include cancellations, then November turns out to be this horrific month and I might end up telling everyone not to fly in November, but if I use delays then November turns out to be a great month. Should I just forget November, and average December and October as a more plausible stand-in for November? Should I look for which month is most like November and put that in instead? So now we are into politics. I have to make a choice here, and each choice will have different consequences for different questions. Not making a choice by just using the actual data I have is going to throw off some recommendations.

There was another occasion where we were exploring the data for airlines day by day, and found there was one period where American

Airlines was super-delayed, and United had almost no delays. This outlier only showed up with one type of graph, it was weird, but you are looking for outliers like this. So then you Google for more about what's going on: 'American Airlines flight delays September 2012'. And you find a labour dispute was going on then, and there were intentional delays. For this period of about 15 days there are tons of delays, and then American and the pilots settled their next ten-year contract. So now we have a dilemma: if I am trying to recommend to my stakeholder which airline to fly next year, do I just use the data as I have it, in which case American is not going to get recommended because it has higher delays, or do I get rid of those two weeks, averaging out that month because the dispute is settled, and it's a ten-year contract, so the odds of something similar happening next year would be almost nil? But United's contract is coming up next year ... so do I take this into account by making the recommendation that says we should take into account labour contract negotiation as putting airlines at higher risk for delays? In that case, I should apply the risk of delays to the other airlines, and not to American.

This is cleaning. It's when you recognise some trend in your data that may or may not be relevant to the question, or hurt the correctness of the recommendation, and you have to make a decision as to whether to leave it in or not. Even this simple flight data turns out to have politics in it – labour disputes, hurricanes etc. Do I take Hurricane Sandy to mean that we are having worse weather and you should stop flying to the East Coast?

DN: I IMAGINE THAT ONE PROBLEM YOUR STUDENTS MIGHT BE FACING IS WHEN TO CALL SOMETHING DONE. HOW DO THEY FIGURE OUT WHEN THE EXPLORATION IS ENOUGH?

JD: Right in the beginning I give them slides that are almost like marketing slides, about the 'Deep Skills' they will be learning: being comfortable with vagueness, willingness to fail, ability to explore, being able and willing to ask for help, understanding and embracing randomness. There are students who are really good *students*, and, because of that, sometimes they run into trouble. I had one student drop the class because she was spending too much time on the assignment. She couldn't figure out how to stop because she wanted to be *right*. I had another student say that 'the word "explore" paralyses me'. Still other students have the opposite problem and just stop at the first

thing that comes to mind (often the first 'recommended chart' by Excel). When I was teaching critical reading and writing at MIT, I had to develop the incompressible assignment, because all the physics students would have these problem sets that took a certain number of hours. They were incompressible. If I gave them a writing assignment next to that, they would squeeze the writing assignment into as few minutes as possible. So I ended up writing these elaborate protocols that tricked them into thinking this could not be compressed.

In Data Studies, I have to make exploratory data analysis as detailed incompressible assignments, because a lot of students have this sense that, if they were going down a route that turned out not to be interesting, they've wasted their time. How do I, as a teacher, communicate that, if you think that way, you are never going to do really good work? You need to try a bunch of things, so one of the things you really need to learn is to try things fast. So the assignment then becomes, 'show me two tries', and then 'show me two more tries'. We can then talk about the value of each try. It's about giving them a sense of progress, but the risk is always that they develop a sense that there is a set of twenty things that they could do to most any dataset, and they can apply it, and then think that they've explored all twenty relations and be done. We all want shortcuts and heuristics, but we first need to really understand that this data in front of us is specific to a time and place, it is situated in historical and cultural and regulatory settings, and therefore may not be 'the same' as previous data we have looked at. We want them to actually find a way to pause and stare at the data, to think about the question they are trying to answer, and see if the data actually answers it, or if they might need to revise the question. That's how data scientists describe it – as you stare at the data, you graph it in a few different ways, you are getting a feel for what this data is telling me. But it is *vague*; it's not anything where if your boss came by, and asked, 'What do you have?' you might not have anything better to say than, well, I've stared at it for two hours. In this sense it is akin to philosophy. That's the interesting dilemma about the pressure for speed, or pressure for results.

In teaching, I've tried various ways to discuss the assignments collectively and in groups. With any group of thirty students, I can have them all say something that is interesting to them about the data. Collectively they cover an amazing number of things, even if at first individually they all are feeling like, 'Oh, I don't really know

anything, I only had one idea.' I'll say, 'But look, together you have set up a lot of interesting questions about the limits of this data for answering this question.' Patience with exploring, comfort with vagueness, willingness to fail.

DN: YOUR USE OF THE TERM 'DATA CLEANING' IS INTERESTINGLY INCLUSIVE. IT'S A LOT MORE THAN WHAT IS SUGGESTED BY 'DATA JANITOR', FOR EXAMPLE. IS THIS FRAMING OF DATA CLEANING PARTICULAR TO DATA STUDIES AS OPPOSED TO 'ORDINARY' DATA SCIENCE, OR DOES IT REFLECT A CONSENSUS ABOUT WHAT DATA CLEANING INVOLVES?

JD: I don't know if there's any consensus at all in data science right now. I don't even know who you would go to, to adjudicate that question. Sure, many people say they know what 'core data science' is, but could you get any sample of data scientists in a room together to agree on it? It's one of those issues where this stuff might not look like 'cleaning', but it is within the area of doing exploratory analysis, looking for outliers and deciding whether to filter them. Is hurricane Sandy an outlier or not? Is a labour dispute an outlier or not? These are things where you can't avoid the fact that they should at least be part of a cleaning conversation, iterating back and forth with analysis.

The cleaning here is very much the iterative process of EDA and modelling that goes beyond data janitor and that is essential. As discussed, every operation on the data is 'political' or a decision that potentially affects different stakeholders. Therefore, the 'data scientist/analyst' has to make them rather than have this done at an earlier stage by somebody else. Often, the cleaning is done by somebody else and we have lost a lot of information about what decisions were made. This is why we always want the raw data. In one case that the students chose, analysing the results of a large-scale US mental health survey, for convenience, the study authors provided the data in different formats.[2] However, they had collapsed several categories into one for at least one variable. This was far from apparent and changed the conclusions drawn by the students. At first they didn't realise this, and were making erroneous conclusions. So we had to go back to the raw data.

There is a fun book called *Guerrilla Analytics* (Ridge 2014). It is full of war stories about how people don't label their files correctly, or

record the process by which they cleaned data, and it creates big problems down the line when someone assumes something different happened. I teach these things to students to say: You might think labelling files is something you do for yourself, but you are really doing it for other people. Whatever you do, odds are that six months from now someone is going to face your files and have to make sense of it. That person might even be you six months from now. If you don't clarify what goes where, what you actually did, and keep all your metadata close to it: that future you is going to be very pissed at the current you.

DN: SO WHY ISN'T THIS DATA SCIENCE?

JD: When I talk with data scientists, they usually say, 'Well, this *is* data science.' I then say, could we have in your curriculum a whole course, or maybe even two, devoted to just this: how to think critically about question formulation, data cleaning and bias, stakeholders mapping, multiple objectivities, unfairness and politics? They will say, 'Well, I don't know if we have time for that, and every course is going to have this in it anyway.' Yes, but that means that every course is going to have the same problem you enacted for me just now, which is that it will get shrunk each time and become secondary to the algorithm types and optimisation – it will be assumed but not taught. That's why I'm calling this Data Studies: a whole minor – it's not a major – but it has its own name precisely so that there will be whole courses devoted to teaching people how to do *the critical social science and humanities work that is an integral part of data science thinking.* Data scientists will always say this is data science, and I say yes, it is, so have your students take the courses!

This is a well-known issue in science, statistics and engineering pedagogy, where the pressure to cover the other core material means squeezing out exploratory data analysis, ethics or other things that are acknowledged as important but, in any particular battle, end up being less important than the material that the faculty in that discipline are primarily trained in. 'Data Studies' remembers that there is an additional set of skills that are required training and are equally necessary for every data scientist, as I am told by data scientists. I try not to tell them, 'You need to be doing this', I just say 'I'm doing this' and then they say 'Oh, that's data science.' I now call these things 'deep versus shallow skills', as opposed to 'soft skills'. I sometimes provocatively

try calling algorithms shallow skills because they do not require long ambiguous conversations and bringing in stakeholders.

Of course, the Data Science programme and major we envision creating at UC Davis will involve a solid mix of EDA, framing questions, interpreting insights, social science and also the usual statistics, machine learning and computing. Already in our pilot classes we have had graduate students take them, and stats students take them. There is a void that needs to be filled. This is why we are experimenting with deep skills pedagogy through cases.

DN: IT SEEMS IMPORTANT TO MAKE THE DEEP SKILLS EXPLICIT – TO NAME WHAT THESE SKILLS ARE, SO THEY DON'T GET MUDDLED AS SOMEHOW 'SOFT' OR IMPONDERABLE. IN ANTHROPOLOGY AND SOCIOLOGY, THEY ARE MADE EXPLICIT.

JD: Yes! In fact, I call this work: 'data archaeology'. Archaeology is a parallel to exploratory data analysis. It involves looking at where the data came from – the provenance or chain of custody, as well as stopping and carefully thinking about what those measurements mean quantitatively. In the airline example, the airlines are sending in their own data to the agency, and it might be necessary to know that airlines are competing with each other in this common data format. An analyst at FiveThirtyEight studied another effect of the '15 minute delay' threshold, and found that for the long flights, e.g. from San Francisco to LA, you could make up the time by going faster (Montet 2014). So if the flight took off within 30 minutes of its departure time, they could make up the time by flying faster while costing the airline a little bit more fuel money, and land within the 15-minute delay window. But if they took off any later, they didn't fly faster at all since they would still be 'late'. So they were behaving in relation to the data collection rule, not in relation to the passengers who would miss a connection if they are 40 minutes late. Data archaeology is figuring out how the data got to you – what each group did, why those columns were added, and that gets you to all the possible places where you can start to ask questions.

There are ways of looking at datasets to make a guess about the culture or historical period in which those questions make sense. In a Foucauldian way, you can start to ask about the genealogy about the dataset itself, and these categories. You can then do a reverse

stakeholder map and ask, why do all these columns exist here? Which people cared that they ended up here? Do I think that there are things that should have been in here but aren't? There's probably a story in that too. These questions provide a more critical sense of why you have this data and not other data in front of you, and they are relevant to what you do to that data. You recognise that the data is one piece of a bunch of possible data, and you might say we should go get other data to answer the question. Even though I could answer the question with the data I have, that might be a defective answer for the people I am trying to help. The data scientist is the last person to see that data before there is an answer floating around, and that answer gets shorn from the data. By the time you get a chart, you are no longer in a position to ask, what choices were combined to produce that?

We hope that the audience to whom the 'answer' is reported would be asking all of these sceptical questions. And a good data scientist should be mentioning any important decisions which may change the conclusion and build them into presentations as well as 'metadata' that travels with the analysis and charts. But STS scholars have long shown that, even with the best intentions, charts and graphics fly free of their qualifications. In our courses, we try to insist that titles, axes, captions and colours have to convey the real, sometimes messy, always limited, story about the data and one's conclusions.

DN: TRADITIONALLY AFTER EXPLORATION AND CLEANING, THERE'S A NOTION OF BUILDING A MODEL. FOR YOUR STUDENTS, WHAT'S NEXT AFTER THESE STAGES?

JD: I'm emphasising the front end because, in the class I taught, students were using Excel for exploratory data analysis, and the plan of the minor is that they learn to work with data scientists and their models. Other classes currently include 'Data Sense and Exploration: Critical Storytelling with Analysis', which introduces the data language R and more visualisation approaches; 'Survey of Data Analytic Concepts & Methods via Case Studies' that introduces machine-learning techniques from a historical and use point of view; electives in areas such as ethics, policy, regulation; and a capstone course that focuses on team approaches to an entire data science pipeline.

But in the introductory course we do work on simple models. For data on the *Titanic* survivors, we built up 'one by one models', where

you could ask, does gender matter? or does gender *and* class matter? and then try gender and class and family size. In this way, you could build up a model that was the equivalent of one type of algorithm that would help you come up with a model of survival. The notion of a 'model', however, gets used in various ways. In statistics, problems are often tied to a sampling question – how do I tell if my sample is good enough (P-values and T-tests and so on). In data science, usually you have something much closer to the population, or found data, rather than an intentional sample. You have all the company's customers, for example, or all the flights for that year, or all the students in the class. So there is no sampling test now, but a need to recognise the data you have and its relation to the inference you want to make. There is a question of what can be predicted, for which I can look at strength of regression, or strength of association of a particular set of features. For that, I can create my own partial universe by taking eighty per cent of my data and seeing what it implies, or predicts, and figure out if there are certain features that can indicate something like survival. Once I have something that works on the eighty per cent of the data, I try to see if it works on the remaining twenty per cent. If it does, it suggests that I may be on to something. There are more sophisticated versions where I try to automate the whole process and repeat ten thousand times: select a random eighty per cent of the data, and figure out what works on that and then test it on the remaining twenty per cent. Our other classes would introduce students to types of modelling, various kinds of regressions, clusterings, machine learnings and so on. The model building, then, is trying to figure out different ways of crunching vast sets of data into better prediction than another one.

In much of data science training, as I mentioned, the model building is part of the ninety per cent of the time spent teaching. There are many different approaches, and each depends on what type of question you are asking: e.g., if you are classifying customers versus trying to predict something. There are a lot of technical choices here, which is the reason why this is ninety per cent of the textbook. But making sure you are asking the right question, clarifying the problem, figuring out what matters in the data so that models can be applied to it, these also matter and are skills that we can deepen in data studies. These hopefully have the effect of helping the students learn what data science can actually do (seeing past the hype) and being able to

work with data scientists. And data scientists, at the same time, training in these data studies skills.

DN: IS IT POSSIBLE THAT SOME OF THE MISRECOGNITION IS HAPPENING HERE, IN THE SENSE THAT PERHAPS SOME FOLKS MIGHT THINK PICKING AN APPROPRIATE MODEL IS THE 'PROBLEM DEFINITION'?

JD: This is a lively debate, in that there are many blogs out there posing questions about this. There is also a debate about Kaggle and other similar data science competitions. The Kaggle approach is to provide eighty per cent of a dataset, and reserve twenty per cent to be predicted by the competing data scientists. Everyone gets the eighty per cent along with a definition of what it is – Netflix user reviews, perhaps, or pictures of fruit. Everyone optimises on that eighty per cent, usually by taking an eighty per cent sub-set of that, and trying to find an optimal model that works on it. You then submit it, and Kaggle runs it on the secret data, and gives you an answer back saying how well you did in the form of a score.

But this is the kind of problem definition where the question of whether the right problem is being asked is off the table. This is the data and you know the question, answer it, full stop. It reinforces the idea that modelling alone can get *the right answer.* However, the people that often win, or some of them, have won partly by what they call 'leakage', which is when the real world leaks into the dataset. This is defined in competition terms as a trace that's not 'supposed' to be in the data. For example, in a hospital data competition, the number system for patient IDs allowed the competitors to figure out groups of patients from different hospitals. By the rules, they 'shouldn't' have been able to do that, but they were able to use it to make a better algorithm. The very term 'leakage' proves that Kaggle poses toy problems, because it defines using the real world as breaking into the 'pureness' of the problem.

There is a side of data science that is trying to become a pure science, the way that statistics went from being applied math to a discipline, and now applied statistics then became data science. Data science is now trying to become a science that defines itself against 'applied data science'. This requires acting as if there is a world where the algorithms can be separated from leakage. But when you are in the real world, leakage is what you want. Any way to figure out a

better answer is supposed to be used. In Kaggle, some people think it is a form of cheating. My son, who enters these competitions, at one point said it should be more 'real data science-y', where you try to think about the world that the data comes from and use that. He used that term in a way that I would like to see everyone use it, where leakage was the whole point, even though at the time he was really obsessed with just tuning his deep learning algorithms to better do prediction.

It might be changing, but so far the emphasis has been on the math, and pushing the limit of computational power. A good Kaggle competition is trying to find the limit of brute force, which is interesting because data science in many ways is the computational brute force approach to the problems that statistics was invented to solve without computational power. The science part of 'pure' data science is about going to the limits of the machine. When we talked to companies about this, some managers would talk about hiring a data science/computer science graduate from an elite programme who only wants to tune cool algorithms. They are almost intentionally ignoring all the leakage. But real data is all leaks, or better put: real data is all context.

DN: DO YOU HAVE A SENSE FOR WHY THERE MIGHT BE THAT DISCONNECT BETWEEN BUSINESSES AND THE WAY DATA SCIENCE IS USUALLY TAUGHT?

JD: I don't want to be misunderstood here. Coming up with a problem that requires all of your programming ingenuity is fascinating, and that joy and fascination is one element that's sustaining the disconnect. Companies and governments and the rest of us eventually have data that currently exceeds our computational power, so coming up with computational tricks to get new insights is amazing. As with applied statistics, there are a lot of people doing the hard work of problem formulation and poring over the data, but that is not what is taught or makes it into the news articles. At the same time, faculty at universities work on the math, that's what they publish on, so the nitty gritty of *teaching how to apply it in context* can take them away from their comfort zones. The default is therefore to teach the 'math' and 'algorithms', which also are easier to test.

But this is also a place where Duncan would remind me that statistical thinking – being able to understand and incorporate randomness

and uncertainty in decisions – is at the heart of data analysis. I'm intrigued by the ability to circumscribe the world, like in a Kaggle competition, and challenge people. Given this or that 'simplified' or 'toy' problem, it's still hard. That toy problem is indexed to a real-world problem, but it turns out you can spend all your time on the toy problem. Maybe a useful analogy is to the way in which chess became a marker of artificial intelligence because it was really hard.[3]

DN: SELF-DRIVING CARS PROVIDE EXCITING ENGINEERING PROBLEMS.

JD: When engineers are imagining self-driving cars, what toy universe are they making in which they can throw all their resources at it and not solve it quickly? They are making incremental progress, but still ignoring the non-toyness of the eventual car; the car that will be making decisions about whether to kill a pedestrian or the driver, whether to get you somewhere on time by taking some risks, or play it safe.

The data disconnect is a historical one, too, where the amount of data being collected and processed has become so large that companies are realising they can still get value out of algorithm-tuning data scientists, even if they could get more value if they added better data scientists. Critical thinking is its own discipline. It's the other half of campus. Doing social science on what's the real problem among this core group of people is a field that includes management consultants. Consultants spend a lot of time learning from people in all parts of companies because companies get so siloed into different departments, and they stop talking. So they hire someone whose job it is to interview everyone. Anthropologists are in companies for the same reason: they don't stay in their box. Of course, Duncan points out: neither do good data scientists. They add value precisely through talking with different groups and putting data in context. Teaching data studies as part of data science is our goal.

DN: IN SOME PREVIOUS CONVERSATIONS, YOU MENTIONED THIS NOTION OF DATA AS A THIRD KIND OF THING. WHAT DO YOU MEAN BY THAT?

JD: I've been thinking about the sense among students and much of the rest of the world that there's a difference between qualitative and quantitative, and that they involve different skills. There is a way in

which that is true, but when you are looking at data like flight delays or surveys, it is amazing how you can apply quantitative skills and interpretive skills to it. At every moment, I find that you can apply both approaches – and you do need to spend time learning to think and reason with uncertainty quantitatively, statistically and socially. I think that there is a third approach. There is a way to talk about data itself as a kind of third skill, where we recognise that there are not two approaches being applied, but you are actually applying a new approach.

My practical definition of data is: anything you can put in an Excel spreadsheet cell. I like the fact that it's in Excel, because Excel itself demonstrates why companies run on Excel. For example, the Excel date format breaks if you give it dates from the nineteenth century (we discovered this when trying to explore early cholera data). It seems like it should be easy enough to fix the date format so that all dates work, but Excel is so full of technocultural legacies that programmers cannot fix this problem without breaking everything. I tell my students that for every headache they experience trying to use Excel, it's just further proof of why companies can't get off of Excel. Excel can't get off of itself. It's proof of the inseparability of history, culture and technology.

Data is Excel with all its broken date conventions. It's not just programming, it's interpretive, cultural, historical world programming. Every piece of data has to be seen as all of those things at the same time, and if we can get to the point where we just say data and mean that, then we'll be in a much better place. Right now we just oscillate. Right now, politically speaking, I'm creating Data Studies as part of this oscillation, but good data scientists are blogging all the time about this stuff – that data can't be treated the way math wants to treat numbers. When you subtract one in Excel, it can break a date if it's from the nineteenth century. That's not just math – the data runs out at a certain point. That's why the Kaggle competitions are effectively 'toy' problems, because they are supposed to be solvable without ever having to ask where are its edges and how do leakages, politics and stakeholders reframe the questions and data. People who win by leakage are good proof that data is this other thing, this third thing, where you have to always use all of these skills, recognising that every time you write an algorithm you are also deciding the cultural and political boundaries of your data. Every time you are

making choices about who to interview to make the analysis, or decide to eliminate data from November, you are making a quantitative and qualitative decision. You are influencing what that algorithm is going to make. You are never separable from these things that appear separate.

I think of data studies as training people right at that intersection. It's training them in Excel and interpretation, in a manner that they can't escape the fact that they are oscillating between the two. It's what I heard the companies saying: we need people who know how to ask the right question about sales data, or engineering data, or public access data. What they were really saying was that we need people who don't think quantitative and qualitative are separate worlds, we need people who can think data critically.

Notes

1 I use the term 'stakeholder' because it is understood across business and government, but I expand it to think through social and environmental justice concerns, drawing on STS and anthropology scholarship. (See for example Callon in Callon and Lacoste 2011; and Fortun 2009.)
2 They were working with the Substance Abuse and Mental Health Services Administration (SAMHSA) population data www.samhsa.gov/data/ [accessed 4 March 2018].
3 See Dumit 2016 for this early history of AI.

References

Brennan, Michael. 2015. 'Can Computers Be Racist? Big Data, Inequality, and Discrimination'. Ford Foundation blog, 18 November. www.fordfoundation.org/ideas/equals-change-blog/posts/can-computers-be-racist-big-data-inequality-and-discrimination [accessed 6 April 2017]. Video available at: https://vimeo.com/146814921.

Callon, M. and Lacoste, A. 2011. 'Defending Responsible Innovation'. *Debating Innovation* 1(1): 19–27.

Dumit, Joseph. 2016. 'Plastic Diagrams: Circuits in the Brain and How They Got There'. In *Plasticity and Pathology: On the Formation of the Neural Subject*. Edited by D.W. Bates and N. Bassiri. Oxford: Oxford University Press, 219–43.

Fortun, K. 2009. *Advocacy after Bhopal: Environmentalism, Disaster, New Global Orders*. Chicago: University of Chicago Press.

Griesemer, James. 2016. 'Taking Goodhart's Law Meta: Gaming, Meta-Gaming, and Hacking Academic Performance Metrics'. Paper given at Gaming Metrics: Innovation & Surveillance in Academic Misconduct, University of California at Davis, 4 February.

Lupi, Giorgia and Posavec, Stephanie. 2016. *Dear Data.* New York: Chronicle Books.

Mackenzie, Adrian. 2007. 'Plying R: A Statistical Programming Language and the Credibility of Data'. Unpublished manuscript.

Montet, Benjamin. 2014. 'Flight Delayed? Your Pilot Really Can Make Up the Time in the Air'. *FiveThirtyEight*, 24 April. https://fivethirtyeight.com/features/flight-delayed-your-pilot-really-can-make-up-the-time-in-the-air [accessed 6 April 2017].

Morris, Hugh. 2015. 'Are Airlines Exaggerating Flight Times so They're Never Late?' *The Telegraph*, 3 December. www.telegraph.co.uk/travel/travel-truths/Are-airlines-exaggerating-flight-times-so-theyre-never-late [accessed 6 May 2017].

Nolan, Deborah and Lang, Duncan Temple. 2015. *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving.* Boca Raton: CRC Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishing Group.

O'Neil, Cathy and Schutt, Rachel. 2013. *Doing Data Science: Straight Talk from the Frontline.* Sebastapol, CA: O'Reilly Media, Inc.

Peng, Roger D. and Matsui, Elizabeth. 2015. *The Art of Data Science: A Guide for Anyone Who Works with Data.* Baltimore, MD: Skybrude Consulting.

Ribes, David and Jackson, Steven J. 2013. 'Data Bite Man: The Work of Sustaining a Long-Term Study'. In *'Raw Data' Is an Oxymoron.* Edited by Lisa Gitelman. Cambridge, MA: MIT Press, 147–66.

Ridge, Enda. 2014. *Guerrilla Analytics: A Practical Approach to Working with Data.* San Francisco: Morgan Kaufmann.

Strathern, Marilyn. 1997. '"Improving ratings": Audit in the British University System'. *European Review* 5: 305–21.

Sweeney, Latanya. 2013. 'Discrimination in Online ad Delivery'. *Communications of the Association of Computing Machinery (CACM)* 56(5): 44–54.

TalksAtGoogle. 2016. 'Cathy O'Neil. Weapons of Math Destruction', 2 November. www.youtube.com/watch?v=TQHs8SA1qpk [accessed 24 February 2018].

Tukey, John. 1977. *Exploratory Data Analysis.* Boston, MA: Addison-Wesley.

Ethnography for a data-saturated world

(2018)

Edited by Hannah Knox and Dawn Nafus

Manchester University Press

Contents