Back to the Future: Review of *Bit by Bit* by Matt Salganik

Gabriel Rossman
Socioplogy, UCLA

In one of my undergraduate courses, I show students a photo of Paul Lazarsfeld and Frank Stanton. Of course neither social scientist is familiar to them, but I argue to my students that Lazarsfeld had a bigger impact on the daily practice of sociology than any member of the Marx/Weber/Durkheim triumvirate they study in classical theory. But even those of us who are aware of Lazarsfeld's impact on sociology can forget how much of his work and social science of his era generally either relied on data collected as an industrial byproduct or piggy-backed theory-driven data collection on industrial efforts. After all, in the picture Lazarsfeld is collaborating with Stanton, a psychology PhD who was first the head of research at CBS radio and then the president of CBS as it transitioned to television. Before the establishment of the National Science Foundation's SBE division and the proliferation of regularly collected surveys with about 1500 randomly sampled respondents and about a hundred theory-driven Likert scale questions, sociologists had to hustle. Data collected through industry sometimes feel deficient by contemporary standards, with non-random samples and few covariates. Nonethless, work along these lines done in collaboration with CBS, Pfizer, *Life* Magazine, and the Department of War gave us many of the great works of social science published prior to the first wave of the GSS. We would do well to learn from the mid-twentieth century model as we consider our response to threats to the late-twentieth century model in plummeting response rates and threatened federal funding for social and behavioral science.

In *Bit by Bit*, Matt Salganik provides a book on methods that will prove instrumental in institutionalizing and building up the new approach of computational social science which in some ways is a return to the mid-twentieth century approach to social science I described above. Computational data can be proprietary, it can be missing variables like education, race, and gender that we consider obviously necessary, and it is almost never based on a random sample of the non-institutionalized adult population. Judged by the standards of canonical survey datasets, computational social science deals with convenience samples that don't have very basic control variables. But computational datasets are also convenience samples that don't have very basic control variables that are enormous, have unbelievably detailed interaction-level data, that are constantly collected and so allow us to observe behavior before whatever event attracted our attention to the site, and have many other attractive properties by which standards it is traditional survey data that looks ridiculously inadequate. I am not arguing, and Salganik is certainly not arguing, that computational data is *better* than traditional quantitative (or for that matter, qualitative) data, but only that it is *different*. Salganik correctly urges us to appreciate how computational social science has enormous potential but also notable frustrations and just like any other method we must learn to appreciate its strengths and its weaknesses, and ideally complement its weaknesses with the strengths of other methods.

*Bit by Bit* is a book that is well-suited for course adoption, but also for self-teaching. I wrote this review on the way home from a computational social science conference and excitement about the book came up several times. Having finished the book, I can say the excitement is justified. Salganik

describes how computational social science can both engage in secondary data analysis of big data created as a byproduct of industry and use a wide variety of purpose-built data collection strategies from wiki surveys (e.g., AllOurIdeas) to citizen science supervised learning (i.e., volunteers code scientific data at moderate scale and this corpus is then used to train an AI that can code scientific data at any scale). An important hybrid of industrial data and purpose-built data is randomized field trials, though this approach has been the source of several ethical controversies. Throughout Salganik is explicit about the intellectual and ethical trade-offs compared to traditional methods.

Salganik is a talented writer and carefully crafts the book to fit multiple points on the learning curve from beginner to expert. Intermediate and advanced readers will appreciate the annotated bibliographies that close each chapter. These go well beyond a list of citations but are short review essays that recapitulate the themes of the chapter but giving more emphasis than the chapters to the sources.

The main way that *Bit by Bit* will serve readers across the range of the learning curve is through the exercises that close each chapter. Salganik annotates each exercise with icons denoting both overall difficulty and whether the exercise requires data collection, math, or coding. Note that in practice "data collection" in the exercises generally means about $10-$50 of Amazon Mechanical Turk labor, which is absurdly cheap by research standards but could very quickly become expensive by pedagogical standards if each student were to collect her own data. As such an instructor considering *Bit by Bit* for course adoption should probably choose sparingly from the data collection assignments and declare the Amazon budget in the syllabus alongside the textbook. Alternately, the instructor could collect a single dataset for the entire class (or each discussion section) which each student would then download from the class website and interpret. Note, I am not complaining about Salganik including these exercises since you can't really understand data until you've picked a clump of it off the ground and rubbed it between your fingers to take in its texture and smell. Rather, I think it would be valuable in traditional methods classes for students to conduct their own random-digit dialing telephone surveys, but this is obviously unfeasible (and arguably unethical) so it is unthinkable, but MTurk data collection is cheap enough that it becomes concievable to imagine it as homework, albeit homework that will quickly add up unless the instructor lets students pool resources or otherwise economize.

I should be clear that the book is *not* a manual for how to code. *Bit by Bit* will tell you how to think about computational social science approaches and which are appropriate for what research questions, but it will not teach you how to scrape a website, implement a mass collaboration, or anything of the sort. There are no blocks of code to type into your computer and only a few schematics and equations, which are there to explain data structures and causal inference, respectively. Salganik's aim is more abstract, he is teaching us how to use computational methods, not how to use a computer. Just as a traditional quantitative methods curriculum includes both a methods course supported by a methods textbook (e.g., Singleton and Straits's *Approaches to Social Research*) and a statistics courses supported by a statistics textbook (e.g., Agresti's *Statistical Methods for the Social Sciences*), so too will a computational social science curriculum need to include both *Bit by Bit* and coding instruction. For readers who want to practice and not just

understand computational social science, I recommend reading *Bit by Bit* and then complementing it with tutorials starting with good introductions to R and Python like Grolemund and Wickham's *R for Data Science* and Lubanovic's *Introducing Python*. From there, one can focus in on libraries relevant to a particular research interest. From personal communication, I know Salganik recommends DataCamp: an R and Python training website offering all-you-can-eat video coursework with an integrated cloud-based programming interface. Nonetheless, that *Bit by Bit* abstracts away from the actual .py or .R files needed to implement its advice is not a bug but a feature as it both makes it accessible to readers who just want to comprehend these approaches and lets advanced readers focus on the distinctly methodological issues without getting distracted by geeking out on functions and libraries. The book's abstraction should be supplemented by a good coding tutorial, but if I had to choose between a social scientist who can code but hasn't thought about the implications of computational work as a research method and one who understands computational work as a method but needs to rely on computer science co-authors for the actual code, I would choose the latter.

One way that bracketing methods goes a bit too far is that for all the thoughts on creative data collection, *Bit by Bit* implicitly assumes that once collected, data have a simple structure. Almost everything is presumptively a rectangular dataset (i.e., rows are cases and columns are variables), just like most survey data. In a few places *Bit by Bit* refers to network data, such as phone records, but often as not treats them as simply data with no special properties. For instance, Chapter 3's section on "amplified asking" extensively discusses Blumenstock, Cadamuro, and On's (2015) use of phone records to estimate wealth in Rwanda. We get a great deal of information about the consumption index ("do you have electricity," "do you own a bicycle," etc.) used to train the imputed wealth measure but don't learn if the call records were interpreted just as minutes of talk time or by geolocation or through some sort of network centrality measure or what.[1] Every once in awhile *Bit by Bit* describes data that were originally photos or free-form text and required cleaning to take rectangular form, but this process is treated as simple. In chapter 5 the Galaxy Zoo team coded photos of galaxies for the amount of blue, the variance in brightness, and the proportion of nonwhite pixels, but we don't know how or how difficult it was to do this. It's just the computer summarized key features of the images, Bob's your uncle. Likewise, we don't hear much about processing free form text data. Aside from a few references to sentiment analysis (i.e., coding whether words connote a positive or negative tone) and a homework exercise about the pre-cleaned Google NGrams dataset, there is nothing about natural language processing. No stemming (i.e., recognizing that "cat" and "cats" are the same word), no entity extraction (i.e., flagging proper names, addresses, and the like from free form text), no topic models (i.e., reducing unwieldy numbers of related words down to a small number of categories). Arguably these are

[1] In fairness, *Bit by Bit* mirrors the original article on this point, but in another article Blumenstock says his approach is similar to the bandicoot library in Python, the website (http://bandicoot.mit.edu/) for which says its indicators "fall into three categories: individual (e.g. number of calls, text response rate), spatial (e.g. radius of gyration, entropy of places), and social network (e.g. clustering coefficient)." Blumenstock et. al. reduced these indicators down to a single vector estimating wealth from phone records. Thus the answer to my puzzle of *how* the call records were used to impute wealth is "all of the above."

issues of implementation that Salganik deliberately brackets, but some relatively brief references to the approaches necessary for handling these problems would let the reader have an idea of where to look for more information on implementation. Most of the book is about radical new forms of sampling and collecting data but skips over equally radical approaches to data cleaning. If and when we get a second edition, I hope to see an additional chapter on these issues.

On the plus side, Salganik's emphasis on understanding the trade-offs of methods rather than the syntax of functions means that *Bit by Bit* will be a valuable read even for people who have no interest in using computational work themselves but just wish to be able to understand and knowledgably critique the work of those of us who do. I especially recommend Chapter Two for readers who want to achieve broad familiarity with the trade-offs of these approaches or to assign as a course reading for instructors who want to include a week on computational methods in a traditional methods course. Readers who are critical of computational social science will appreciate Salganik's candor in repeatedly returning to the young field's handful of scandals, even outside the ethics chapter. You will learn something relevant to you about computational social science from *Bit by Bit*, whether or not you reflexively preface the word "algorithm" with the definite article.

When I heard a few years ago that Salganik was writing a textbook, I was surprised and a little disappointed that this would be a distraction from his cutting edge research in areas like information cascades and respondent driven sampling. I was a fool. Just as chapter 5 of the book describes how computational approaches can enable mass collaboration on research projects by spreading the work from credentialed experts to masses of people with low or unkown skill, *Bit by Bit* itself will do more for computational social science by spreading the heretofore tacit knowledge of the field than a top researcher could accomplish directly. I strongly recommend *Bit by Bit* and fully expect it will be the standard methods textbook for computational social science until advances in the field render it dated. If we are lucky, we will benefit from a new edition every five to ten years so the book can keep pace with a rapidly evolving field. However for now it is incredibly current and I highly recommend it to any social scientist who teaches, practices, or aspires to practice or even just understand computational social science.