Notes on *Bit by bit: social research in the digital age* (Salganik 2017)

Ian Stewart

June 26, 2017

1 Introduction

- A digital social science experiment, such as the Rwanda call study that linked call records to wealth, can achieve results faster and at a lower cost than traditional experiments. It may also provoke different responses from different stakeholders. Data scientists may view it as a machine learning problem, social scientists as an economics problem, privacy researchers as an example of surveillance, and business people as a potential for valuable, inaccessible information.

- The digital era presents new possibilities for social science with the ubiquity of technology, an increased scale and variety of data available, and increased computing power. Rather than continuing to do social science as always, researchers should consider ways to integrate traditional methods with new digital-specific methods.

- Typically, social scientists and data scientists have worked with separate methods and goals, but the digital era allows for collaborations between the two disciplines, including a focus on *research design* as the link between questions and answers.

- Digital social science often involves choosing ready-made models (e.g., repurposing existing data), custom-made models (e.g., collecting new data), or a hybrid model.

- Despite new possibilities for complex models, simplicity is often more valuable than complexity in designing research methods, because it often reflects a strong fit between the theory and data and it allows for unexpected results.

- The lower cost of digital experiments provides a wider range of possible data collection but also raises ethical issues, such as obtaining participant consent from internet users.

2 Observing Behavior

- When we talk about *big data* we usually refer to (1) government records, which often include demographic data, or (2) business records, which often include digital traces.

- Big data that is suited for research should be (1) big, (2) always-on, and (3) non-reactive.

    1. *Big* can mean many units, significant amount of information per unit, or many observations over time. A large size can help make estimates for specific subgroups of units, uncover rare events, detect small differences, and make causal estimates. However, researchers must account for the process of generating such big data to account for potential collection errors.
    2. *Always-on* means a constant collection of longitudinal data. This type of data allows researchers to study the same units at different points in time, to estimate causal effect, and it also allows researchers to generate real-time estimates, which can be useful for policy makers.
    3. *Non-reactive* means that study participants are unaware of the data collection process and are less likely to modify their behavior in response to it (e.g. racism in search queries). However, this doesn't preclude the possibility of positive bias in self-presentation in online social situations, or algorithmic confounds from the platform providing data.

- Big data that is not suited for research is (1) incomplete, (2) inaccessible, (3) non-representative, (4) algorithmically confounded, (5) dirty, or (6) sensitive.

    1. *Incomplete* data often misses relevant demographic information, data on other platforms, and data relevant to theory. The lack of demographic and external platform data can be addressed through imputation, or extending information on one group of units to another group, and through record linkage, or combining records from different sources to understand a larger phenomenon.
    2. *Inaccessible* data is generally protected by government or businesses to preserve stakeholder privacy. Researchers can partner with government or business organizations if their interests overlap and both parties can provide something for the other side, but the data may not be shareable with other researchers and the research results may be biased by the organization (e.g., no bad press).
    3. *Non-representative* data can give a biased impression of the underlying population, since the data may reflect different unit characteristics and behavior than the "true" population.
    4. *Drifting* data result from change in the underlying population, behavior, and system and present a distorted view of social activity.

5. Data subject to *algorithmic confounds* reflect the influence of an online platform rather than social activity (e.g. Facebook's friend "minimum" of 20). This can be complicated by performativity, or the influence of a theory on the world, which includes platform-specific algorithms such as Facebook's "suggested friends" feature that encourages triadic closure.
6. *Dirty* data contains spam and bot activity irrelevant to most research, but in order to clean the data researchers need to understand how the data was generated.
7. *Sensitive* data may be inaccessible or may need to be anonymized to protect the stakeholders from emotional or economic harm.

- The research strategies that apply particularly well to big data include (1) counting, (2) forecasting and nowcasting, and (3) causal inference.

- A large proportion of social science is based on *counting* phenomena, and this kind of study should be motivated by interesting or important research questions rather than "filling holes." For example, a study of taxi drivers was motivated by the question of whether the drivers would work more or fewer hours on days with higher wages, i.e. maximize profit or minimize effort.

- Predicting the future through *forecasting* is often conducted with causal models that explain the underlying causes of the dependent variable or with discriminative models that focus on accuracy. *Nowcasting* predicts the present using past data and can be an effective replacement for counting in live-streaming situations such as the Google Flu Trends project, which continuously predicted flu outbreaks by cross-checking search queries with offline flu monitoring. In either kind of prediction it is important to compare results with a baseline to ensure that the model tells researchers something useful.

- Although social science is often too complicated to allow randomized experiments (maximum causal estimate), researchers can still estimate *causal inference* with natural experiments and matching.

  1. In a natural experiment, the control and treatment groups are selected by a (semi-)random event, such as a draft lottery. Big data can facilitate natural experiments if it is always-on and large enough to provide representative control and treatment populations.
  2. In matching, the treatment units are individually matched to control units based on similar values of control variables. Researchers should be careful to draw conclusions in the context of the variables used for matching (e.g. acknowledging unobserved variables) and should be aware that the often limited subpopulation used for matching may differ from the overall population.

- Big data can provide three main benefits for social science research: (1) judging the value of competing hypotheses, (2) improving measurement of policy impact, and (3) measuring causal effects with natural experiments and matching. It can also allow researchers to derive new theory from the data, rather than relying on theory to point research toward relevant data.

3 Asking questions

- Researchers often use surveys to provide insight on social behavior, such as individual intentions (*internal states*), in a way that complements observation. The digital age has provided a wealth of opportunity for new computer-administrated survey methods.

- In the total survey error framework, survey error results from representation error and measurement error.

  1. *Representation* error is a mismatch between the survey respondents and target population, in which the frame population (from which survey sample is drawn) is systematically different from the target population. Researchers can account for this error by understanding how the data was collected to avoid skewing toward the wrong population.
  2. *Measurement* error is a mismatch between the survey responses and the actual thoughts and behaviors of the survey takers. The way that researchers ask survey questions often shapes the response and can bias the inferences made. Running pilot studies with respondents from the target population can reduce the risk of measurement error.

- Researchers must balance reducing error with reducing cost (e.g. sampling a subpopulation rather than the entire population), and digital surveys often produce more data at a lower cost, but with more subtle errors.

- Sampling a population for survey may rely on probabilistic or non-probabilistic methods.

  1. *Probability* sampling weights samples according to how they were collected, such that units with higher probability of collection have more weight. In a sample of unemployment across states, a researcher might have to increase the probability of people from smaller states to collect an equal number of units per state, and then generate a weighted mean to account for different population sizes.
  2. *Non-probability* sampling assumes no knowledge of how the sample was collected (e.g., random ad clicks online) and can produce biased samples. The errors from non-probability sampling can be corrected by post-stratification, which weights

units according to self-reported data such as state of residence. However, this assumes homogeneity response propensity within groups: for example, a New York resident has the same probability of participation as an Alaska resident. Researchers can support this assumption by more rigorous post-stratification that increases the number of groups used, such as gender plus age plus state of residence, without excessive sparsity. They can also support the assumption by weakening it to a non-correlation condition, i.e. that there is no correlation between response propensity and the dependent variable.

3. In non-probability sampling, researchers can produce better samples through more careful data collection that includes quota sampling, in which researchers fill fixed counts of specific demographic quotas, and sample matching, in which researchers match units from a diverse volunteer "panel" to the known population distribution (e.g. matching 25-year old female from population to volunteer).
4. In general, non-probability sampling when done correctly can produce accurate estimates at a lower cost and with less error than probability sampling.

- The digital age has enabled a rethinking of the traditional survey to make it more embedded, open-ended and enjoyable. A digital survey lacks an interviewer and therefore permits increased flexibility but potentially less incentive to stay engaged.

  1. Researchers can embed digital surveys into typical internet browsing activity by decomposing them into *ecological momentary assessments* (EMA), to be conducted at disparate times and places. An EMA (1) collects data in a real-world environment; (2) focuses on an individual's current state or behavior; (3) may be based on an event, time or at random; and (4) requires multiple assessments over time.
  2. An *open* survey permits all written responses to questions, rather than a multiple-choice response. A hybrid *wiki* approach presents fixed responses as well as allowing respondents to contribute their own response, which then become part of the fixed response pool.
  3. *Gamified* surveys encourage participation through play that is often social, such as a survey that compares a respondent's social perception of a friend with an actual response from their friend.

- Researchers can augment the estimates produced by surveys by linking the surveys to external data sources, and the main methods for doing so are amplified asking and enriched asking.

  1. *Amplified asking* connects digital trace data with a sample survey of the units that produced the digital traces, then estimates survey responses over a larger

sample of digital traces. A study of wealth in Rwanda (Blumenstock, Cadamuro, On 2015) used this technique to estimate composite wealth over a large population based on a complete knowledge of phone calls (digital traces) and a survey of wealth for a subpopulation.

2. *Enriched asking* connects digital trace data with survey data to produce a richer dataset, which may require finding a corporate partner who has already aggregated the necessary data. Researchers should be careful to ensure the quality of the digital trace data and survey data, which may not be "ground truth" but can still enrich the digital data.

- Surveys should complement observational studies, and the digital era provides resources for surveys to implement non-probability sampling, computer-driven surveys, and survey linking.

4 Running experiments

- Social scientists undertake experiments to test causal relationships that cannot be inferred from observations alone due to potential confounding factors.

- *Randomized controlled experiments* control for confounding factors by randomly assigning units to control and treatment groups.

- The four components to a successful randomized controlled experiment are (1) participant recruitment, (2) randomized treatment, (3) consistent treatment delivery, and (4) measurement of outcomes.

- Experiment design can vary along two dimensions, (1) lab to field and (2) analog to digital.

  1. *Lab* experiments use highly controlled but artificial conditions to test a causal effect, while *field* experiments use more natural settings to elicit realistic responses from participants. Hybrid experiments can combine these two extremes to test a causal effect in both controlled and realistic scenarios, such as studying the likelihood of hiring mothers based on rated resume responses (lab) and actual hiring responses (field).
  2. *Digital* experiments use computer infrastructure to run experiments, while *analog* experiments use non-computer infrastructure to run experiments. Digital experiments benefit from (1) a larger scale of participant recruitment, (2) more knowledge of participant background information (pre-treatment information), and (3) a larger longitudinal scale. Researchers should be aware of the shortcomings of digital experiments such as limitations on treatment manipulation and ethical concerns for participant treatment.

- Moving past simple experiments requires researchers to consider more nuanced tests of the causal effect under inquiry. For instance, a within-subject study does not have a control group but instead compares subject behavior before and after treatment, thereby improving estimate precision as compared to an aggregate estimate.

- Richer experiments are characterized by (1) validity, (2) heterogeneity of treatment effect, and (3) mechanisms.

  1. A *valid* experiment supports a more general conclusion. Experiments are assessed for (1) statistical conclusion validity, (2) internal validity (correctness of experimental procedures), (3) construct validity (match between data and theory), and (4) external validity (generalizability). Although external validity is the hardest to assess because of the extra experiments required, assessing it can be helped if the system under study is always-on and has a low cost of experimentation (e.g., assessing effect of power bills on electricity consumption).
  2. Measuring the *heterogeneity of treatment* requires accounting for systematic differences between participants, such as the difference between heavy versus light power consumers (Allcott 2011). Merging survey data with digital trace data can produce richer data and reveal unforeseen heterogeneity.
  3. A causal *mechanism* explains how and why the experiment's treatment is connected with the effect. Digital experiments can help reveal mechanisms through (1) collecting process data (e.g., data on possible mechanisms for increased power consumption) and (2) testing different but related treatments (e.g., different combinations of interventions to reduce power consumption). Researchers can test the effects of different treatments by layering the treatment conditions or by trying every possible combination

- Researchers can conduct digital experiments on their own by (1) using existing systems, (2) building an experiment, or (3) building a product.

  1. *Using existing systems* allows the researcher to run a digital field experiment with low cost, low control, medium realism and high ethical risk. An example experiment is an intervention on Craigslist to test the effect of race on seller success (Doleac and Stein 2013). These experiments may only reveal system-specific effects unless generalized appropriately.
  2. *Building an experiment* allows the researcher to run the experiment with medium cost, high control, medium realism and low ethical risk. Like lab experiments, this strategy requires participant recruitment (e.g. MTurk) and can help the researcher isolate a highly specific effect that would otherwise be hard to elicit in field experiments, although the test to isolate the effect might be artificial. An example experiment is testing voter behavior through participation in a game (Huber, Hill and Lenz 2012).

3. *Building a product* allows the researcher to run an experiment with high cost, high control, high realism, and low ethical risk. This strategy recruits participants through their willing engagement with the product and, when successful, generates a positive feedback loop between researchers, product and participants. An example experiment is the MovieLens project that allows users to rate movies and has yielded a wide range of successful field tests.

- Researchers may also *partner with a company* to run the experiment with low cost, medium control, high realism, and high ethical risk. This strategy provides a larger scale as compared with the "do-it-yourself" strategies and significantly more restraint, such as a company's unwillingness to fund research that harms their reputation. By partnering with a company, researchers often perform experiments motivated more by use than by inherent desire for knowledge, but these two dimensions of science can interact productively (e.g., Pasteur's fermentation experiments). An example experiment is Facebook's A/B testing of peer voting information to determine the effect of peer influence on self-reported voting (Bond et al. 2012).

- Before running an experiment, researchers should think over all possible ramifications of the study design and formalize the analysis plan by following reporting guidelines, to avoid needlessly repeating an experiment to collect missing data.

- Rather than design a single perfect experiment, researchers should consider running multiple smaller but self-reinforcing experiments.

- Underemployed digital experiment strategies include creating zero variable cost data, i.e. data whose cost does not increase with additional participants. Researchers can reach this goal by collecting low-cost data (e.g., automating human research labor with scraping) and by making their experiments enjoyable, thereby reducing the cost of participant recruitment. An example experiment is the MusicLab rating system that provided free music listening as compensation for participants (Salganik, Dodds, Watts 2006).

- When possible, researchers should consider lowering their studies' ethical risks by *replacing* experiments with less invasive methods, *refining* treatment to reduce the potential for harm, and *reducing* the number of participants. An example experiment that could have been improved is Facebook's emotional contagion experiment (Kramer, Gillroy, Hancock 2014) in which the experimental manipulation could have been replaced with an observational approach.

- Working with digital experiments can open a variety of possibilities for researchers, including the potential to improve the validity of their study, estimate the heterogeneity of treatment effects, and isolate causal mechanisms.

5 Mass Collaboration

- Massive systems such as Wikipedia would not be possible without collaboration between a large and diverse set of people.

- Researchers can achieve similarly successful results through collaboration by relying on (1) human computation, (2) open calls, and (3) distributed data collection.

  1. *Human computation* projects break a large task into small, easy tasks for regular internet users to perform. An example project is the Galaxy Zoo that solicited basic annotations of galaxies. Many human computation projects rely on the split-apply-combine strategy that allows workers to independently address small problems chunks that can be later combined to produce a consensus solution. Redundant annotations can ensure higher quality data and can make up for lack of formal training among volunteers. To reach a high quality consensus, researchers should clean, debias and reweight the data to account for variation in classification quality among volunteers. Human computation can be augmented with supervised machine learning techniques that use gold-standard annotations to label unseen data at a larger scale than what humans can label.
  2. *Open call* projects seek novel solutions to clearly formulated problems, often prediction tasks, from an open crowd of colleagues, in such a way that it is easier to check solutions than generate them. An example project is the Netflix prize that invited researchers of all kinds to develop new rating prediction algorithms. A clearly formulated problem with a simple evaluation criterion provides a well-scoped space in which researchers can experiment and submit solutions. Open calls may augment expert solutions rather than replacing them entirely, as in the case of the Peer-to-Patent system that provided patent reviewers with crowdsourced information. In addition to improving algorithm performance, open call projects may also identify outliers and discover unexpected aspects that make them hard to predict.
  3. *Distributed data collection* projects solicit data from a wide variety of sources that were previously inaccessible to researchers. An example project is the eBird study that requested bird-watching reports from volunteers. The data collected may be biased as a result of volunteer activities and perspectives, such as the abundance of bird reports that occur near roads or a tendency toward “interesting” birds. This bias can be handled by encouraging the collection of novel rather than redundant data, potentially through gamified scoring systems.

- Researchers interested in designing mass collaboration projects should consider five principles for a successful project: (1) participant motivation, (2) heterogeneity, (3) attention focus, (4) preparing for surprise, and (5) ethics.

1. *Motivation* may come from money, personal interest or collective interest (e.g. wanting to contribute to public knowledge).
2. Data *heterogeneity* is inevitable due to variation in participant skill and effort. Rather than excluding the “bad” participants, researchers should leverage the long-tail of contributions to ensure redundancy and consider guiding participants to better contributions.
3. *Focusing attention* helps maximize participant contributions and may be helped by an explicit contribution score.
4. *Surprising* discoveries can result from participants noticing an unexpected pattern in the data and building consensus among one another.
5. An *ethical* experiment provides sufficient compensation and credit to participants, in addition to ensuring privacy of the data presented to participants (e.g. anonymizing Netflix users).

- Researchers should be aware of the risk of insufficient participation (i.e. wasted effort) and may rely on pilot testing to assess participant interest before potential failure.
- The digital age has provided new methods of mass collaboration in the form of human computation, open calls, or distributed data collection, which allow researchers to democratize science at an unprecedented scale and speed.

6 Ethics

- The large-scale, distributed nature of digital experiments has raised new ethical questions that often do not fit into the typical rules-based approach of social scientists (e.g. IRBs) or the ad-hoc approach of computer scientists.
- Examples of ethically challenging research include the Taste, Ties and Time study (Wimmer and Lewis 2010), which scraped Facebook profiles without informing participants and shared the data with other researchers who were able to de-anonymize the data.
- Digital experiments afford an unprecedented amount of power to the researchers over participants, such as the ability to observe and perturb a participant’s environment without their knowledge (similar to the panopticon). The persistence of digital data means that data collected for research may later be appropriated for unethical use to influence participants without their consent, potentially leading to a “database of ruin.”
- Ethical decisions that researchers make are often complicated by a lack of agreement over rules and norms to govern research, such as the intersection between academic and industry norms in the Facebook emotional contagion study.

- To assess the ethics of their work, researchers should consider a *principles-based* approach to make ethical choices and communicate their reasoning to a wider audience.

- Researchers can rely on four ethical principles to address uncertainty: (1) respect for persons, (2) beneficence, (3) justice and (4) respect for law and public interest.

  1. *Respect for persons* requires researchers to treat participants as autonomous and to honor their wishes. This can be addressed through informed consent.
  2. *Beneficence* requires researchers to not harm participants and to maximize potential benefits while minimizing potential harm. This can be addressed through a risk/benefit analysis that seeks to understand and improve a study's risks and benefits, not just for participants but, when necessary, the broader social world.
  3. *Justice* requires researchers to distribute the burdens and gains of research equally across social groups, rather than having one group reap the gains of another group's efforts. This may entail either extending protection to exploited minorities to reduce their burden or extending access of data on such minorities to allow them to benefit. Justice can be addressed through reconsidering participant compensation.
  4. *Respect for law and public interest* requires researchers (1) to *comply* with relevant laws and terms of service and (2) to be *transparent* about their research goals, methods and results in a way that allows them to take responsibility for their decisions. This can be addressed through a third-party ethical review and an assessment of the potential public impact of a research finding.

- The two frameworks of *consequentialism* and *deontology* can guide ethical thinking by focusing on either the ends or the means of research.

  1. *Consequentialism* requires researchers to make actions that make the world better, such as balancing risks and benefits.
  2. *Deontology* requires researchers to make actions that are inherently good, such as honoring participant autonomy.

- A realistic researcher will blend both consequentialism and deontology to balance the ends and means of their research ethics.

- Digital researchers often face ethical issues from four difficult open problems: (1) informed consent, (2) understanding and managing informational risk, (3) privacy, and (4) making decisions in the face of uncertainty.

  1. Most research should involve some form of *informed consent*, rather than maximizing consent at all times. For example, experiments that study hiring discrimination may need to deceive employers without their consent, which is justified

by the potential benefit to society and the minimal harm to employers. Asking for informed consent may increase participant risks (e.g. exposure to internet censorship), may compromise the study's scientific integrity, or may be logistically impossible to reach all relevant participants.

2. *Informational risk* is common in digital research because of the larger scale and wider variety of data available to researchers. The risk may be reduced through data anonymization provided that the anonymization makes it impossible to identify the participants, which is hard when the "anonymized" data can be merged with other sources to provide more identification power. Since all data is potentially identifiable and sensitive, researchers should consider a data protection plan that maximizes the safety of the project, people involved (trust), data (anonymized, aggregated), data environment/setting (computer security), and output (preventing breaches). Although the decision to not share data ensures maximum security, it also limits the possible benefits from sharing data, and researchers should consider an intermediate approach such as a "walled garden" that allows limited data sharing with trusted people.
3. Data *privacy* is the right to appropriate information, based on the context of the data sender and recipient (actors), the type of information (attributes), and data constraints (transmission principles). Different combinations of these three factor groups lead to different informational norms, such as the norm of allowing researchers to share limited data for clearly scientific purposes. Privacy decisions often require resolving a conflict between different ethical principles such as Beneficence (maximum benefit to society) versus Respect for Persons (minimum harm to participants) in the case of a study that involved spying.
4. Researchers often must resolve *uncertainty* when making ethical decisions. The safest approach is to maximize precaution at the risk of slowing scientific progress, which includes the restriction or cancellation of hazardous experiments. Researchers should consider a minimal risk standard to compare the risks of their experiment with the risks faced by participants in everyday life. Researchers should also use power analysis to minimize the study's sample size, which maintains statistical significance without endangering participants. To gauge participant reaction to the experiment, researchers may conduct pilot surveys and staged trials for potential participants.

- Although an institutional group such as an IRB may seem like the final arbiter of ethical choices, the researcher should consider an IRB as a floor rather than a ceiling, i.e. the establishment of minimal standards that can be improved. Including an ethical appendix in a study provides an opportunity to explain the choices made and issues raised by the work.

- To preempt outsider perception of an experiment, researchers should imagine how participants, stakeholders and non-experts will react to the study.

- Research ethics are not binary (ethical versus non-ethical) but instead continuous, such that researchers can always improve their ethical approach.
- Developing ethical standards that suit both public and scientific opinion will allow researchers to make the most out of their digital experiments.

7 Looking forward

- The future of social science will be guided by digital age affordances, including (1) blending ready-made and custom-made experiments, (2) participant-centered data collection, and (3) research ethics.
  1. While previous social science research has either relied on ready-made or custom-made models, future work will rely on hybrids that combine the scale of ready-made solutions with the theoretical benefits of custom-made models. Combining ready-made (governmental) and custom-made (survey) models presents a richer perspective on the same phenomenon that would otherwise be inaccessible.
  2. Unlike previous research that recruited participants in a "captive" setting, future experiments will have to compete for participants' attention and to motivate their participants to contribute data. Large-scale studies such as the Galaxy Zoo would not have been possible without motivating participants in a game-like fashion.
  3. Due to the wider possibilities of digital research in terms of data collection, future researchers will devote more energy to developing methods that are more ethically sound. Studying the impact of methods such as differential privacy, which allows for aggregated queries on sensitive data to avoid de-anonymization, will help researchers build systems that allow data sharing without harming participants.
- The future of digital social science will combine social science and data science to provide new possibilities for researchers and organizations alike.