Temporal Validity in Online Social Science

Kevin Munger

December 11, 2018

Abstract

The centrality of the internet to modern life means that the social and political world is changing faster and less predictably than ever before. At the same time, the "credibility revolution" has forced social scientists to confront the limits of our methods for creating knowledge. The interaction of these two trends is not yet well understood. I argue that the increasing rate of change of the objects of our study makes "knowledge decay" a potentially large source of error. "Temporal validity" is a form of external validity in which the target setting is in the future—which, of course, is always the case.

1 When is Social Science Possible?

Social science was born understanding the tradeoff between generating general knowledge and making accurate predictions about behavior in a given case: "For among statements about conduct those which are general apply more widely, but those which are particular are more true, since conduct has to do with individual cases, and our statements must harmonize with the facts in these cases" (Aristotle, 1954).

Depending on the phenomenon under investigation, this "harmonization" is more or less difficult. This difficulty is related to the *heterogeneity* of the phenomenon: the number of divergent outcomes our theory must explain. Greater heterogeneity means more difficult social science.

Recent methodological innovations have consistently demonstrated that social science is indeed difficult. The rise of randomized control trials (RCTs), regression-discontinuity and natural experimental approaches has increased the credibility of social science research, but it has also increased the relevance of concerns about external validity. Compared to regressions that aim to describe global phenomena, this research generates an internally valid estimate of a causal effect in a given time and place, and on a given subject population (Samii, 2016).

The goal of this research is to accumulate generalizable knowledge; in the words of Dehejia, Pop-Eleches and Samii (2015), "with a large number of internally valid studies across a variety of contexts, it is reasonable to hope that researchers are accumulating generalizable knowledge, i.e., not just learning about the specific time and place in which a study was run but about what would happen if a similar intervention were implemented in another time or place. The success of an empirical research program can be judged by the diversity of settings in which a treatment effect can be reliably predicted."

The "credibility revolution" means that more attention must be paid to causal identification, and thus that researchers must devote more effort developing their identification strategy and less on novel theorizing. In a landmark outline of this paradigm, Samii (2016) argues that this does not mean that "'theory is being lost' but rather that theory is being held constant as we go about the difficult business of trying to do credible causal inference," and that "generalization and theory development are better left to synthesis studies."

This approach to research represents an important innovation, but it highlights a blind spot in the way that social science methods have been adapted to study human

behavior on the internet. Academic research takes place as time advances. As internally valid studies accumulate, the world changes. To the extent that social science has thus far succeeded, it is because the rate of knowledge accumulation has outpaced the rate at which old knowledge becomes obsolete due to the world changing.

The internet is now important for many aspects of our social and political lives, and it changes incredibly quickly. The implicit assumption that accumulation outpaces obsolescence no longer holds, particularly for knowledge about online behavior. This represents a serious challenge to the practice of academic social science, but not an insurmountable one. Industry researchers at powerful technology companies are able to conduct thousands of experiments with millions of subjects, and have developed new statistical techniques to take advantage of this capacity. The rate of internal, industry-secret knowledge production has increased, but academic knowledge production has not kept pace.

This paper conceptualizes this problem as specific form of external validity: *temporal validity*. Recent economics research has outlined conditions by which knowledge from a collection of contexts can be generalized to make predictions about a novel context, but (as I argue below) these conditions cannot hold when the unidirectional nature of time is considered. The extent of this problem varies across different realms of inquiry; for most social science question, there are many more pressing sources of error. However, the baseline rate of temporal validity for research on online behavior is sufficiently low that this source of error is a first-order problem. Unless the rate of change of the internet slows or the rate of academic knowledge production increases, extant social science research paradigms may be fatally inappropriate to the study of online behavior.

2 External Validity/Generalizability

The recent explosion of interest in RCTs among development economists has led to a growing literature on strategies for aggregating locally estimated treatment effects and applying this knowledge to other contexts—the "external validity" or "generalizability" of findings.[1] This turns out to be difficult (Deaton, 2010).

Frequently replicated experiments on a given population are insufficient, even in the presence of large sample sizes and rich individual-level covariate information. Allcott (2015) demonstrates this limitation in a paper on "site selection bias": even with "large

[1]In this paper, I follow the Rubin potential outcomes framework and refer to causal effects as "treatments" and internally causally identified research as "experiments."

samples totaling 508,000 households, 10 replications spread throughout the country, and a useful set of individual-level covariates to adjust for differences between sample and target populations." However, the "extrapolation bias" of the effect of the same intervention applied at other sites is an order of magnitude larger than the estimated standard error of the treatment effect. Similarly, Vivalt (2016) aggregates the results of impact evaluations of international development programs from 635 published papers. Development economics is "one of the first fields...with enough papers on comparable topics to do this analysis," and the results are not promising: "results are much more heterogeneous than in other fields."[2]

Manageably unbiased extrapolation has been shown to be empirically possible. Frequently replicated experiments that span decades and the globe scale can be used to aggregate treatment effects and extrapolate them to novel contexts. Using the Angrist and Evans (1996) natural experiment (that the sex distribution of a household's first two children acts as an as-if random assignment to have additional children), Dehejia, Pop-Eleches and Samii (2015) use 166 country-years of census data (with an aggregate sample size of 12 million) from the Integrated Public Use Microdata Series. Models with over 50 country-years of data can generally produce unbiased extrapolations to other country-years, accounting for both micro- and country-level covariates.[3] Bisbee et al. (2017) extends this approach to the case of instrumental variables. Both of these cases require knowledge of the covariate values in the context being extrapolated to, and cannot account for the creation of novel covariates. For an example of the latter, consider a country which implemented a strictly enforced two-child policy: the fertility treatment effect in this country would be 0, regardless of other covariate values, and the value of the two-child policy variable in the future produces a difficult-to-model source

[2]In a comparable paper from social psychology, Paluck, Green and Green (2018) perform a meta-analysis of the literature on the theory that inter-group contact reduces prejudice (Allport, 1954). This discipline has not fully embraced field experiments, so they are only able to aggregate across 27 randomized field studies. The results are very different from the previous gold standard meta-analysis on the topic: Pettigrew and Tropp (2006) aggregates more than 500 studies and finds strong, context-independent and homogeneous effects of contact reducing prejudice. Restricted to the 27 well-conducted studies, however, Paluck, Green and Green (2018) find that these effects are in fact weaker, context-dependent and more heterogeneous. Even more troublingly, "not one study [of the over 500] assesses the effects of interracial contact on people older than 25." The lack of population sampling leaves open the possibility of far greater heterogeneity; although the results are not conclusive, there effect sizes of the studies conducted on adults over 25 were in general smaller than those on younger people.

[3]The authors admit that they cannot account for site selection into their database; all of the country-years share the property of "have data archived at IPUMS," and it is possible that the model would not extrapolate correctly to country-years which do not have this property.

of extrapolation error. Dehejia, Pop-Eleches and Samii (2015) demonstrate that this "intrinsic variability" swamps prediction error and does not decrease even as sample sizes increase.

Rosenzweig and Udry (2016)'s work on "External Validity in a Stochastic World" is the only attempt to model this form of error—and the only other work to use the term "temporal external validity"—of which I am aware. They first identify several high-profile papers in which either pre- or post-treatment data from a single year randomly had above-average rainfall, a covariate which led to inflated treatment effects but which was not included in the original models. They then explore several contexts in which stochastic shocks moderate treatment effects (eg micro-loans to individuals who fall ill have no effect on their productivity). In order to assess the temporal external validity, they argue, researchers need to be able to estimate the moderating effect of stochastic shocks *and* characterize the distribution of those shocks.

In the framework below, I conceptualize these "macro shocks" as a special case of covariate non-overlap. Depending on their magnitude, frequency, and predictability, "macro shocks" pose a serious to social science. During the mid-20th century, Western political science devoted considerable resources to interpreting the secretive communications of the Soviet Union; after 1989, much of this highly specialized knowledge became largely useless. If world-upending "macro shocks" like the fall of the Soviet Union were more common, we might well rethink the way that we conducted social science.

A pessimistic reading of the argument I present in the current paper is that politics on the internet may be more similar to a world in which Soviet Unions are built and destroyed every five years than we have yet realized. The internet is increasing Knightian uncertainty (Knight, 2012) by creating novel forms of stochastic shocks faster than we can hope to study them and characterize their distributions.

3 The Internet and Politics

The internet (and the way it is used for politics) changes rapidly. The population of people who have used the internet for politics has changed in non-transparent ways over the past twenty years. Perhaps worst of all, a huge percentage of internet politics takes place on privately-held platforms which can change without oversight or notice.

These issues imply that research about the internet and politics is of low temporal

validity.

Although this problem is acute for research about the internet and politics, it is not unique to this field. Many empirical findings change over time, but any findings which are conditional on the cultural/technological/regulatory media environment are less temporally valid as the rate of change of that environment increases and our ability to predict those changes decreases.

A illustrative example is Gelman and Huang (2008), with reference to one of the most robust results in American electoral politics: the incumbency advantage. For a number of reasons, incumbents are likely to win re-election, but the degree of their advantage has varied over time. Gelman and Huang (2008) estimate a substantial increase from 1950 to 1990 and then a gradual decline.[4]

The introduction of new information technologies has rapidly increased the rate of change of this phenomenon. Using the (as-if random) deployment of broadband internet from 2002-2008 as a source of exogenous variation in the media-technology environment, Trussler (2018) demonstrates that the incumbency advantage varied rapidly during this period—from 12% in areas with low internet connectivity to 5.8% in places with the median number of broadband providers.

Fowler (2015) models incumbency advantage from the perspective of an imperfectly informed voter, and argues that incumbency advantage should decrease in the amount of information the voter has about the candidate. Trussler (2018), Jacobson (2015) and Hopkins (2018) emphasize the increasing nationalization of politics and growing party loyalty leave fewer degrees of freedom for incumbency advantage. In either case, the media environment is the conduit by which people acquire this information, and the increased rate of change of that environment in the 21st century has limited the temporal validity of the incumbency advantage in a given year during this time period.

Social science is not documenting empirical regularities for their own sake; the goal of aggregating empirical knowledge is to develop a theory based on understanding the *mechanisms* that produce those outcomes.[5] The research cited above is an excellent example; the authors advance different mechanisms to explain the incumbency advantage, and future research will help adjudicate between them. Understanding the mechanism allows us to make predictions about how the incumbency advantage will vary across

[4]With a different methodology, Fowler (2015) estimates the peak incumbency advantage in the US House of Representatives slightly earlier than 1990, but agrees that rates of change were gradual throughout the 20th century.

[5]At least, this is my understanding of the difference between how social scientists and historians understand the goals of their respective endeavors. This is not a permanent or intrinsic distinction.

contexts.

Descriptive analysis is a necessary first step: no one can have a theory of incumbency advantage before knowledge of such an advantage exists. Descriptive analysis tells us *what is*, allowing us to think about *why*. But the internet—as it exists today, permeating our society and our politics—ensures that *what is* is changing faster than ever before.

This speed poses a number of challenges to the practice of research; it has become more difficult to hold enough of a given context constant in order to appropriately test underlying mechanisms, especially when the context in question is online. Furthermore, this speed affects the real world in a way that makes our research increasingly urgent. In the wake of the 2016 US Presidential Election, the country (and the world) has come to terms with the fact that we do not fully understand the ways that the internet and social media are affecting us. Facebook itself does not know how to fix what it has broken.

This goal of this essay is to begin a discussion of these challenges and propose potential strategies for tackling them.

4 Formalizing Generalizability

4.1 Econometric Approaches

In this section, I will present the model of external validity developed by Hotz, Imbens and Mortimer (2005) to "Predict the efficacy of future training programs using past experiences at other locations." The approach used in this paper relies on a pair of assumptions which are (of course) false, but are "true enough" to be useful in the context of job training programs. However, I argue that they are not "true enough" for the study of online politics.

Their inferential setup is as follows:

"A random sample of size N is drawn from a large population. Each unit i, for i=1,2,...,N, is from one of two locations, indicated by $D_i \in 0, 1$. For each unit there are two potential outcomes, one denoted by $Y_i(0)$, describing the outcome that would be observed if unit i received no training, and one denoted by $Y_i(1)$, describing the outcome given training. Implicit in this notation is the Stable Unit Treatment Value Assumption (SUTVA) of no interference and homogeneous treatments (Rubin 1974, Rubin 1978). In addition, there is, for each unit, an indicator for the treatment received, $T_i \in 0, 1$ (with $T_i = 0$ corresponding to no-training or control, and $T_i = 1$ corresponding to training),

and a set of covariates or pretreatment variables, X_i. The realized (observed) outcome for unit i is $Y_i \equiv Y_i(T_i) = T_i * Y_i(1) + (1 - T_i) * Y_i(0)$.

We are interested in the average training effect for the $D_i = 1$ population:

$$\tau_1 = E[Y_i(1) - Y_1(0)|D_i = 1]$$

We wish to estimate this on the basis of N observations $(X_i, D_i, (1 - D_i) * T_i, (1 - D_i) * Y_i)$. That is, for units in the $D_i = 0$ location we observe the covariates X_i, the program indicator D_i, the treatment T_i and the actual outcome Y_i. For units in the $D_i = 1$ location we observe covariates X_i and the program indicator D_i but neither the treatment status nor the realized outcome."

The first assumption is the "unconfounded location" or "no macro-effects" assumption:

$$D_i \perp (Y_i(0), Y_i(1))|X_i$$

This means that any systematic differences between the locations are only due to the distribution of the covariates X_i at each location.

A related and necessary assumption is the "support condition": for all X

$$\delta < Pr(D_i = 1|X_i = x) < 1 - \delta,$$

for some $\delta > 0$ and for all x in the support of X. This means that, for all values of the covariates, there are some units that take that value in the first location.

The second assumption was mentioned in the quote above: homogeneous treatments. Each "treatment" is in fact a bundle of "treatment components."

Consider a treatment with $K + 1$ treatment components. For each component t, with $t \in \Theta = \{0, 1, ..., K\}$, and each unit i, there is a potential outcome $Y_i(t)$. For unit i, $\tilde{T}_i \in \Theta$ is the treatment component received The researcher only observes the binary treatment assignment $T_i = 1\{\tilde{T}_i \geq 1\}$, where $\tilde{T}_i = 0$ refers to the control condition.

The homogeneous treatment assumption is that:

$$\tilde{T}_i \perp Y_i(1), ..., Y_i(K)$$

If all of these assumptions hold, it follows that we can generalize the results from the first location to the second location by averaging the conditional treatment effects calculated in the former over the covariate distribution in the latter.

Of course, these assumptions are routinely violated in practice, and researchers have taken several different approaches to dealing with this problem. The least satisfying is to restrict the scope of inquiry; Hotz, Imbens and Mortimer (2005) do this by "restricting comparisons to the sub-populations in each location for which we have sufficient overlap."

Even if a single study is necessarily restricted in its scope, the premise of the causal empiricist enterprise is to aggregate findings from across many studies such that the union of their scope covers the entirety of the covariate and treatment component spaces. This knowledge aggregation is not trivial, however, and this subject has been the focus of much recent methodological attention (Athey and Imbens, 2016; Dehejia, Pop-Eleches and Samii, 2015; Egami and Hartman, 2018; Gechter, 2015; Green and Kern, 2012; Hartman et al., 2015; Ho et al., 2007; Imai, Ratkovic et al., 2013; Kern et al., 2016; Nguyen et al., 2017; Stuart et al., 2011; Taddy et al., 2016; Wager and Athey, 2017).

A tractable problem is that both the covariate space and the treatment component space are large; for $\mathbf{C}$ covariate levels and $\mathbf{K}$ treatment components, we need to estimate $\mathbf{C} * \mathbf{K}$ values from the data.[6]

Proposed solutions to this problem involve reducing the treatment component X covariate space. This can be done parametrically; this is most common in fields with well-developed theories about how to collapse variables into a smaller parameter space. A prime example is in economics, the rational choice school assumes that people are motivated to have more money. This allows researchers to map each treatment component (a new version of a given policy implementation, say) onto a single dimension: how much it will change the monetary endowment of treated subjects. Of course, the behavioral economics revolution has falsified the strong version of this assumption (Deaton and Cartwright, 2018). Still, weaker versions of the theory allow researchers to parameterize a portion of the treatment component space, and in general, the goal of theory is to allow researchers to specify parametric relationships that can inform predictions about treatment effects in a given case.

However, the majority of the developments cited above use non-parametric methods. One approach is to use matching and reweighting to find the location where the treatment effect is known that is most similar to the target context in the treatment X covariate space. A complementary approach uses machine learning to discover the

[6]In principle, this space is infinitely large.

covariate values with the largest effect heterogeneity, restricting the space in which matching/reweighting is necessary.

These statistical innovations reduce the costs of precise generalizability by identifying the portion of the covariate X treatment component space for which we need internally valid estimates of treatment effects in order those effects to the entire space.

The fundamental problem of generalizability, then, is not that the covariate X treatment component space is large, but that it *expands over time*. Because all of our knowledge is from the past and all the contexts to which we hope to apply that knowledge are in the future, Hotz, Imbens and Mortimer (2005)'s "no macro-effects assumption"/"support condition" will always fail to obtain.

Let $t \in \mathbb{I}$ denote the time at which a study was/will be conducted, where $t < 0$ denotes the past, $t = 0$ the present, and $t > 0$ the future. Because time is unidirectional,

$$X_{t<0} \subseteq X_{t>0}$$

Of course, time is infinitely divisible, and this process is not instantaneous. Let the *rate of change* of a given phenomenon r be the minimum time difference such that the covariate set expands:[7]

$$X_t \cap X_{t+r} \neq \emptyset$$

At any given time, r varies across different subject areas. In general, though, accelerating technological progress increases r.

For a concrete example, consider the incumbency advantage discussed above. In all of the studies conducted prior to 2002, the value of the covariate *Broadband* was undefined.[8] After the (as-if random) rollout of broadband internet, however, *Broadband* takes the value of 1, violating the support condition.

An analogous quantity is the *rate of knowledge decay*, d. This is the rate at which knowledge of a treatment effect at a given time period improves our ability to estimate that effect in the future, conceptualized as the existence of relevant covariate overlap.

[7]Further assume that the covariate set expands in such a way that the value of the novel covariate cannot be predicted by other covariates:

$$x_{ij} \perp x_{i0}, x_{i1}, ..., x_{iC}$$

for some $x_{ij} \in X_t \cap X_{t+r}$

[8]Alternatively, if we define the covariate space as infinitely large, we can say that there was no variation in *Broadband* in this time period, as it always took the value 0.

$$x_i \notin X_{t+d} \forall x_i \in X_t$$

Knowledge of the incumbency advantage prior to 2002 *decays* when $Broadband_i = 1$ for all units.

If a "macro-effect" occurs, there is perfect separation in the covariate values; *none* of the covariate values taken in the future existed in the past, meaning that all of our knowledge has decayed.[9]

Obviously, this claim is extreme. Some of our knowledge must be transferable from existing covariates to the novel covariate—intuitively, we might select the covariate that is "most similar" to the novel covariate.

This appeal to "similarity" is ultimately unavoidable for research that takes place in the unidirectional flow of time. The philosopher of science Nancy Cartwright has repeatedly criticized RCTs on the grounds that generalizability ultimately requires some appeal to the target context being "similar enough" to known contexts (Cartwright, 2007*a*,*b*; Deaton and Cartwright, 2018).[10]

My critique is related: *non-parametric approaches are insufficient for temporal validity.* Replication is impossible becuase time is unidirectional. This has been known since Hume but is increasingly relevant due to trends in the way social science is conducted.

Our only recourse is to use theory to bridge these gaps; theory is the best guide to extrapolate knowledge of treatment effects to truly novel settings. But the process of theorization itself takes time. Even researchers who prefer to discover treatment heterogeneity with the non-parametric methods described above are not immune to the need for theorization: *novel covariates need to be conceived of before they can be measured and exist in data.*

In practice, the amount of bias in future predictions is related to the rates of r and d, as well as the total variance of effect heterogeneity for a given treatment. For many of the subjects that have been studied with RCTs, these rates have been sufficiently low that their contribution to bias has been small relative to problems of experimental design/implementation.

[9]This is, of course, the fundamental problem of induction, best dramatized by Bertrand Russell (Russell, 2001). Through repeated observation, a chicken estimates the causal effect of the farmer's daily visit to be that he is fed. There is a perfect separation between the past and the future on a crucial covariate: in all of the observations in the past, $ChristmasDay = 0$. When $ChristmasDay = 1$, however, the causal relationship changes, and the farmer's visit causes the chicken to be slaughtered.

[10]But see, among others, Imbens (2018), who argues that Cartwright's understanding is mistaken, or at a minimum that she and the applied statisticians she criticizes are talking past each other.

RCTs are becoming more common in fields in which the rates of r and d are much higher, however—most obviously, in the study of online behavior. The temporal validity of online treatment effects poses a serious epistemic challenge.

4.2 Generalizability in Practice

David Karpf's 2012 article "Social Science Research Methods in Internet Time" makes a series of arguments related to the one in this essay. This article has been relatively influential in the fields of Communication and Information Studies, but has been largely ignored by political scientists (of 170 citations, 3 are found in Political Science journals: *Political Communication*, three times).

Karpf points out that:

"(1) The rate at which the Internet is both diffusing through society and developing new capacities is unprecedented. (2) Many of our most robust research methods are based upon *ceteris paribus* assumptions that do not hold in the online environment. The rate of change online narrows the range of questions that can be answered using traditional tools."[11]

Table 1 formalizes the problem that Karpf describes in his point (2). The assumptions underlying the rows and columns are closely related to treatment homogeneity and support conditions discussed in the econometric literature above, but have been re-written to focus on the issues most relevant to the study of online behavior. The four boxes describe the research designs necessary to estimate treatment effects under different combinations of assumptions. These assumptions are, of course, false—but in certain cases they are useful. Throughout, we assume that treatment effects are heterogeneous (they vary among people with different characteristics).

The two columns differentiate between a world in which we assume that treatment effects are stationary (left) and one in which they are non-stationary (right). The rows denote worlds in which the population of interest has a stable or changing composition:

Assumption 1 *Effect Stability: The treatment effect will not change over time.*

Corrollary 1 *Heterogeneous Stability: The treatment effect on each specified subgroup will not change over time.*

[11] My conception of *temporal validity*—which I developed without knowledge of Rosenzweig and Udry (2016)'s work on the subject—owes an obvious debt to Karpf's piece; my aim in writing this essay to broaden Karpf's insight.

Assumption 2 *Constant Composition: The composition of the population of interest will not vary over time.*

Corrollary 2 *Completely Theorized Composition: All of the relevant covariates have been identified and can be measured.*

In the context of the incumbency advantage, the "treatment effect" of incumbency is the vote share of a given politician in the world in which they are the incumbent compared to the world in which they are not, *ceteris paribus*.

Box A refers to a world in which the causal effect is stable and the population constant. These modeling assumptions are always false when studying human behavior—they only apply to ideal-conditions hard sciences like Chemistry or Physics. In the incumbency example, this would mean that a single study that estimates the heterogenous effect of incumbency on each relevant subgroup in the population would be sufficient to know the true effect of incumbency on vote share, forever.

Box B relaxes the assumption of effect stability, allowing the effect of incumbency to vary over time. Note that the stationarity assumption contains also the realm of *predictable change*. If there were a truly predictable change in treatment effects, we could build this into our estimates. True predictability is generally implausible; consider the shifting incumbency advantage, due either to the nationalization of politics or to the increased information environment provided by the internet. No one could have fully anticipated these developments, and we are today unable to fully anticipate potential technological or institutional changes which could affect the incumbency advantage. This world requires that we perform frequent studies on samples with full support in the relevant covariate space to capture the changing causal effect on each identified subgroup, as this world also relaxes the corollary of heterogeneous stationarity.

Box C assumes effect stability but allows for a dynamic composition. Again, we need to frequently repeat the initial study as the population shifts in order to measure the true effect of incumbency. Covariate weights can allow for some adaptation of previous estimates to the population's new demographics, but this is not possible if a new subgroup enters the population, as is possible when we relax the corollary of completely theorized composition. That is, if in addition to white and black incumbents, Asian incumbents enter the population, our sample needs to reflect this; this subpopulation had 0 support when the initial study was conducted, so our initial estimate of this heterogeneous effect would be undefined.

Table 1: Assumptions and Research Desings

	Causal Effect Stable	Causal Effect Non-Stable
Constant Composition	*(A)* Single Study on Sample With Covariate Support	*(B)* Frequent Studies on Samples With Covariate Support
Dynamic Composition	*(C)* Frequent Studies on Representative Population, Track Demographics	*(D)* Frequent Studies on Representative and Frequently Updated Panel

Box D relaxes both assumptions; this is the real world of social science research. Here is where the lack of both the heterogeneous stationarity and completely theorized composition corollaries bites: the *theories* of effect heterogeneity that allow us to specify and measure the subgroups of interest become invalid. In the incumbency example, the geographic location of an incumbents' constituency was not theorized as relevant—and indeed, it wasn't relevant when the theory was produced. But with the quasi-random rollout of broadband internet access, this previously orthogonal geography subgroup becomes an essential moderator of the effect of incumbency. In order to address this world, we need panel surveys to track within-individual effect changes; we also need theory building (often by conducting studies on theoretically novel sub-populations) to determine how to update the panels to ensure representativeness in both sampling and covariate analysis.

Again: this is the real world of social science research, the enterprise we have been engaged in for many years. The framework above is meant to clarify the role played by these *ceteris paribus* assumptions in how we think about research. The crux of my argument here is that the internet has changed the meaning of the word *frequently*; following the discussion in section 4.1, the internet has increased r.

Academic research is produced along several time cycles. The broadest is the overall production of knowledge and its forgetting, or disappearing from a discipline. The next is in the life span of individual researchers, who accumulate knowledge throughout their lives, and produce knowledge at several stages with different incentive structures (as

graduate students, as untenured faculty, and with tenure). And the shortest is the timespan of a given research project, which can entail a number of steps—acquiring a deep understanding of the literature, gaining necessary skills, applying for grants, conducting a field experiment, data collection and analysis, preparing a manuscript, submission, rejection, revisions—before it results in a peer-reviewed publication.

Before the advent of the internet, the rate of change of the objects under study was generally not high enough to "intersect with" these time cycles.[12] Our society, culture and politics—even elements like the incumbency advantage which are not obviously related to the internet—are changing faster due to the increasing connectedness and decreased costs of communication entailed by mutually reinforcing technologies of the internet, social media and smart phones. The pace of academic knowledge production has increased, but it has not kept up.[13]

Offline, the internet has decreased effect stationarity more rapidly than it has constant composition; the latter is, in many contexts, constrained by human life cycles and other biological/material frictions. For social science research studying offline phenomena, this means more research designs need to be in the realm of Box B that before might have been in Box A. This is a shift that is well within the technological capacities that digital automation and communication have afforded us, at least for many research questions. Previously, a study might estimate incumbency advantage with a dataset from a fixed time period. If all of the data collection, cleaning, and analysis is automated, the results of that study can be updated for minimal cost.[14] This introduces a powerful new form of knowledge production: conditional predictions that clarify the mechanisms at play. Returning to the example of incumbency advantage, we could adjudicate between Fowler (2015)'s proposed mechanism of increased information and Hopkins (2018)'s proposed mechanism of increased nationalization of politics; if nation-

[12]Several high profile exceptions prove the rule: with the unexpected fall of the Soviet Union in 1989, a generation of Sovietologists found that rate of change spike vertically through their career timelines.

[13]A skeptic might counter that previous information technology advances have not forced us to rethink the structure of knowledge production. This is the weakness of induction: barring some rapid increase in the rate of academic knowledge production (or slowdown in the rate of change in the subject), this trend line will only intersect the rate of change of the subject *once*.

[14]There are many research designs for which this approach will not work, of course. Experiments and qualitative studies in the physical world have large fixed costs that automation cannot relieve, and design-based causal inference approaches rely on idiosyncrasies that automation cannot create. As I discuss below, however, software that dramatically lowers the cost of online experimentation could lead to an explosion of such research. An analogous advance in the real world is the Broockman, Kalla and Sekhon (2017), who develop statistical techniques that dramatically reduce the cost of running RCTs.

alization/party loyalty decreases but the incumbency advantage remains unchanged, we should lower our credence in that mechanism.[15]

Even if academic knowledge production keeps pace with the changes wrought by the internet, our subject matter follows unchanging cycles that limit the rate at which data about electoral outcomes and voting behavior can be collected. There is only one US Presidential Election every four years. This is an especially serious problem for political actors and campaign strategists, which I discuss in the final section.

For research that studies online behavior, however, the difficulties are much more severe, and the necessary adjustments to our research strategies and institutional frameworks more dramatic. Returning to Karpf's point (1), recall that "the internet" is in fact a constantly evolving, overlapping set of hardware, software and combinations of users. From our vantage point in 2018, it is obvious that no effect of "the internet" has been constant over time. The changing composition of internet users (to say nothing of a given platform eg Facebook) represents a particularly thorny problem.

It must be noted that the kind of research designs necessary to address this problem are actually already being implemented—by Facebook, and a select few other companies with a sufficient userbase, technological capacity and supply of social scientists. Most of this research is not being published, and is thus of limited use to academic social scientists. However, there a burgeoning academic/industrial subfield developing the methodological tools necessary to conduct this research. Peysakhovich and Eckles (2017) and Athey et al. (2017) are recent papers that develop tools to do research in Box D where "frequent" is in internet time.

To provide a more concrete argument for why this kind of research is necessary, the next two sections explain the relevant dimensions on which internet is unique.

5 Ceteris Paribus Non: Changing Compositions

Figure 1 plots the changing age composition of the internet and social media over time. I take longitudinal survey data from Pew (Pew Research Center, 2018; Smith and Anderson, 2018) to calculate the relative adoption rates of the internet and social media among different age groups. I then use data from the census about the number of American adults in each of these age categories in 2010 to calculate the proportion

[15]Samii (2016) agrees with the necessity for this kind of work: "An experiment or natural experiment is especially interesting if it provides an opportunity to asses the value of competing models of causal mechanisms...credible empirical work clarifies situations where one or another model is useful."

of internet and social media users in each age category.

In 2006, 75% of adult social media users were between the ages of 18 and 29, but 0% of social media users were over 65.

These age categories are crude, but stark. If age is a moderator of the effect of Facebook (as I argue below), then a single study conducted in 2006 could not be used to estimate the overall effect of Facebook use in the future.[16]

A more general concern is that the Facebook users are far from a random sample of Americans of their age; "early adopters" differ from "late adopters" on many previously untheorized and difficult to observe dimensions.[17]

At first, Facebook famously restricted membership to undergraduates at elite colleges, then expanded to include anyone with a university email address. Once Facebook opened up membership, social media adoption rates were higher among people who were wealthy, educated and urban (Smith and Anderson, 2018). In a detailed study of a population of 18-19 year-olds in 2007, Hargittai (2007) finds that Facebook adoption was higher among people who lived with their parents, had internet access at home/at a friend or family member's home, and, mostly prominently, spent more time on the internet.

Even these figures, though, do not get at the issue of *how different people use Facebook*.

Sociologist Eszther Hargittai has written widely about the "second digital divide"—the inequality in internet skills among people with internet access (DiMaggio, Hargittai et al., 2001; Hargittai, 2001). The most consistent finding is that people who use the internet more often are better at using the internet; recent research co-authored by Facebook employees demonstrates that frequency of Facebook use is predictive of Facebook skill, measured both through a survey and an objective frequency of clicking on spam (Redmiles, Chachra and Waismeyer, 2018).

Hargittai has demonstrated inequality in internet skill among both young "digital natives" (Hargittai, 2010) and the elderly (Hargittai, Piper and Morris, 2018). Among both age groups, people who use the internet more often and who have greater autonomy of use have higher levels of internet skill.

There is thus good reason to think that a random sample of elderly Facebook users

[16]Facebook has been the most widely and frequently used social network to date, so I simplify the discussion about social media to focus on Facebook.

[17]A similar phenomenon has been document among subjects recruited from Mechanical Turk (Coppock, 2018; Stewart et al., 2015).

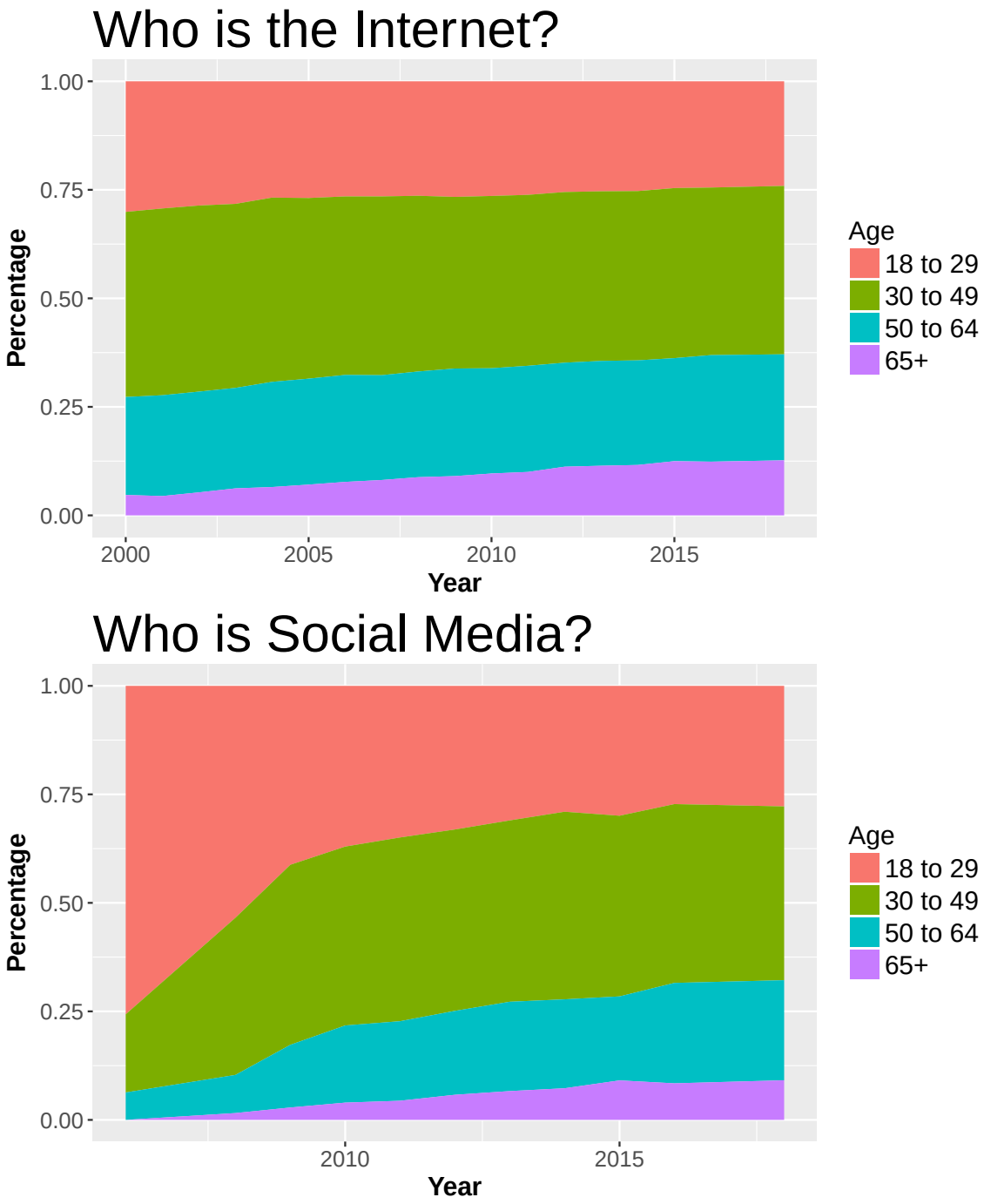

Figure 1: Data from Pew and the CIA World Factbook. The overall % of US adults using the internet grew from 52% in 2000 to 89% in 2018. The overall % of US adults using social media grew from 5% in 2006 to 69% in 2018.

in 2010 would be less technically savvy than one drawn in 2016. Until this theory is produced, however, it is impossible to use it to model heterogeneous effects.

To illustrate how this might be a problem, I will discuss a series of classic papers in the social media and politics literature: the Facebook social pressure and voter turnout experiments. At the time of writing, the original study (Bond et al., 2012) has over 1,300 citations. This (and the following papers) are excellent studies, among the gold standard examples of fruitful collaboration between academics and industry platforms. My goal in discussing the temporal validity limitations of these studies is not to "debunk" this research—indeed, the authors at several points discuss the problem of the shifting Facebook architecture and population, in slightly different terms—but simply to raise the salience of these limitations.

Bond et al. (2012) presents two main findings: (1) social messages (in which the faces of subjects' friends who voted were included along with information about how to vote) caused a significant increase in validated voting; (2) non-social, informational messages had a precisely estimated null effect on validated voting. This experiment was conducted during the 2010 US Congressional election.

Bond et al. (2017) presents a subgroup analysis of the original experiment, exploring how the effect of social pressure varies by subjects' age, gender, and level of education. By far the largest heterogeneity is in the age of the subjects: "The effect size for those 50 years of age and older versus that of those ages 18 to 24 is nearly 4 times as large for self-reported voting and nearly 8 times as large for information seeking"; the effect on validated voting is also far and away the largest of the demographic variables available.

In discussing the limitations of the this finding, the authors say that they "do not assume that these results would hold in other settings. At the time the study was fielded, more than 50% of U.S. adults were Facebook users (Facebook, 2011), but these results may not hold for non-Facebook users."

The latter clause correctly implies that people who joined Facebook after 2010 might respond differently to the treatment than those in the sample, but the fact that "more than 50% of U.S. adults were Facebook users" masks the fact that just 11% of adults over 65 used social media while this study was in the field. The generalizability of one theorized mechanism—that age moderates the effectiveness of social pressure—is thus of limited temporal validity due to the dynamic composition of Facebook users.[18]

[18]There is an optimistic case to be made that we are in a period of disequilibrium: there a huge number of digitally naive people are coming to use the internet and social media for the first and only time. The early days of the internet saw it populated solely with savvy first movers. As the internet

6 Ceteris Paribus Non: Changing "Facebooks" (Non-Stationarity)

The other constantly shifting variable in any equation designed to estimate the effect of Facebook use is the Facebook platform itself.

There have been many changes to the Facebook interface since its inception, but the most relevant changes have come in the form of improved individual targeting made by tweaks to the News Feed algorithm. With the introduction of the now-iconic "like" button, Facebook allowed users to give positive feedback to posts. This enabled Facebook to learn what kinds of posts a given user wanted to see more of, and to fill each users' News Feed with more specifically targeted posts.

The cumulative impact of these changes may also have been responsible for the diminished effect of social pressure between 2010 and 2012. Jones et al. (2017) reports the results of a replication of the previous GOTV experiment on Facebook during the 2012 campaign. The effect of the social recommendation button was still to increase voting, but not nearly as much. The point estimate of the identical treatment on validated voting was .39 in 2010 and .17 in 2012; the latter was not significant without the addition of control variables.

There were a number of modifications to the Facebook interface between the two campaigns, including a complete visual overhaul and increasing the character limit on users' posts from 500 to 63,206. But from the perspective of the effectiveness of the GOTV button, the most significant change was the introduction of advertisements ("Featured Posts") directly into users' News Feeds on January 10, 2012. In 2010, the GOTV button was a very rare instance of a visual stimulus that did not come directly from a users' friend (aside from the display ads on the side of the page, to which any casual internet user has become inured), but in 2012, these kind of visual stimuli were commonplace.

Jones et al. (2017) are forthcoming about this issue: "both the Facebook platform, and how voters and campaigns use the service change over time." They also point to the literature on the increased salience of Presidential elections and the concomitantly decreased effectiveness of GOTV campaigns. The conclusion to the paper indicates

became more financially and technically accessible and more essential for daily life, entire populations got online. This will will never happen again, as new generations of "digital natives" replace older cohorts who will never fully get the hang of the technology. If this story is correct, the sting of dynamic composition might in some contexts return to pre-internet levels.

that the latter is their preferred explanation for the decreased effect size from 2010 to 2012.

How could we know? How can we estimate what portion of this difference is due to unpredictable effect non-stationarity due to changes in the Facebook platform, and how much due to the electoral context?

The general approach of social science would be to accumulate knowledge by conducting additional experiments; by replicating the study in Presidential and Congressional elections, perhaps at subnational levels, we could begin to zero in on the true effect of social pressure in the form of a GOTV button on Facebook in these different electoral contexts.

As with all social science, a research agenda that aims to measure the effect of Facebook-driven social influence on voter turnout must be answered with attention to unpredictable effect non-stationarity and untheorized dimensions of heterogeneity. The solution to this problem, as with all the rest, is the use of representative panel surveys, frequent replication, and novel theorization of effect heterogeneity.

I have chosen this example to illustrate how research studying online behaviors changes the meaning of the word *frequent*. Facebook can be radically and unpredictably changed overnight, in a way that physical institutions cannot; new Facebook users can differ from more experienced users in ways that require theorization. Research on the effect of Facebook-driven social influence on voter turnout is constrained from becoming sufficiently frequent in two ways: first, the academic-hours required to acquire the skills/resources/access necessary to conduct that research (only one research team has successfully conducted these studies) and the publication time cycle (the paper describing results from the 2012 study was published in 2017); and second, the rate at which elections occur (once every two years).

The upshot of this critique is that researchers should strive for designs that are faster and more replicable, and that social science should adopt institutions to adapt to this.

7 Conclusion

As a final illustration of the challenge presented by temporal validity for doing social science on the internet, I will discuss the limitations of my own work. As part of my dissertation, I conducted a pair of experiments to test how social norm enforcement

happens on Twitter. Munger (2017*b*) describes an experiment in which I used Twitter "bots" to send messages discouraging a sample of white men from using a racial slur to harass others. This experiment was conducted in the summer of 2015.

In Munger (2017*a*), I extended this experimental paradigm to study norms of partisan incivility during the 2016 US Presidential election, among a sample of Republicans and Democrats.

The studies differed in many ways, but here I focus on one crucial element: the role of subject anonymity. In preparing the first study, I read through the literature on online anonymity and theorized (in my pre-registration) that the subjects who had opted into creating Twitter profiles that protected their anonymity would be less responsive to the treatment (would reduce their rate of racist harassment less).

I found the opposite; it was the people who tweeted the word "n*****" with an account that contained their real names who did not change their behavior. In the follow-up study, on the other hand, I found results consistent with my original expectations—but I had updated my pre-registered expectations to be in line with the results from the first experiment.

These two studies are thus inconclusive as to the role of subject anonymity in moderating social norm enforcement. There are many possible explanations as to why. My hypothesis is that there are different norms as to the acceptability of the behavior in each case; everyone knows that racist harassment is a norm violation, but the norm about partisan incivility (especially during an election as marred by incivility as 2016) are far from settled. The anonymous racist harassers were doing something they thought was wrong (hence their decision to be anonymous), and were thus susceptible to social pressure. The non-anonymous racist harassers were openly violating the norm, suggesting that they thought it was a bad norm that they wanted to change; if this is true, it is unsurprising that they were not swayed by a single message. This hypothesis is plausible as a post-hoc explanation, but it was hard to conceive as a testable theory (with implications for heterogenous effects on different samples) until after conducting the experiments.

There is no way to theorize the mechanisms that explain the role of anonymity in moderating treatment effects on all relevant subpopulations until the distribution of that heterogeneity is established. This would require either a much larger experiment on a representative sample of all racist harassment or many smaller experiments on different subpopulations.

Each of the experiments represents a contribution to knowledge on its own, but they

cover only a very small amount of the combinatorial space of theoretically interesting variations on the design. Other experiments could vary the identity of the sender, the language of the message, or the number of messages; each of these experiments could be run on a different sub-population of interest.

But this takes time. Initially, I had to conceive of the experiment, develop a codebase, work with the IRB, run the experiment, analyse the results, write a manuscript, submit it for review, wait, revise the manuscript, and wait, until the paper was published. This process took two years (16 months from experiment to publication), beginning November 2014—this was a best case scenario, as *Political Behavior* was the first journal to which I submitted the paper, and the wait time was relatively short.

Tunning these two experiments and writing these two papers by myself was not, in hindsight, the most temporally valid use of my time. A better approach to the study of social norm enforcement online would have been to either a) get a large grant and run an experiment with many treatment arms and a much larger sample size; or b) develop software and tutorials to enable other researchers to implement the experimental design at much lower cost.

Both of these are difficult for graduate students/early career researchers. Grantwriting is costly and uncertain; further, this is not a general solution unless the total amount of available grant money increases. Software requires an ongoing time commitment to maintain, and is insufficiently acknowledged as the kind of contribution that should "count" in hiring and tenure decisions.[19]

Career incentives, the top-heavy distribution of resources and the lethargy of traditional journal publication are all obstacles to conducting more temporally valid research about online behavior.

To close on an optimistic note, I want to highlight two examples of work that sets the standard for temporal validity given the current constraints.

Matias and Mou (2018) proposes and demonstrates the implementation of community-led experimentation in online contexts. Operating as a nonprofit, the project develops software and trains volunteer citizen scientists in research design. Where a traditional researcher-led experiment costs hundreds of social scientist-hours, this paradigm has

[19]Of course, this point may soon become moot: Twitter has dramatically stepped up their bot detection and become much more aggressive about suspending accounts in the wake of public uproar surrounding the 2016 Election. I have spoken with researchers who have had their accounts suspended by Twitter while trying to replicate my experimental design. This completely unpredictable uncertainty is an unavoidable feature of research designs that use proprietary platforms that can change without warning.

the potential to produce hundreds of experiments for the same cost. One drawback of this approach is that many of these experiments may not be ideally designed to provide evidence towards a specific social science theory, but having orders of magnitude more experiments is better on balance.

Though using a more traditional research design, the theory-building effort in Settle (2018) is done with an eye towards generalizability in studying the rapidly changing world of social media and its relationship with polarization. Through an intensive study of the function of the Facebook News Feed, Settle generates a novel theory of how it works to polarize people who consume it more. Book-length theory-building is necessary for social science, but incredibly time intensive; in order to maximize the temporal validity of her insight, Table 9.1 provides a matrix of predictions for specific mechanisms in different technological contexts.

As a community of scholars aiming to produce generalizable knowledge about online behavior, we should embrace methodological approaches like Matias and Mou (2018) and aim to produce theory like Settle (2018).

References

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130(3):1117–1165.

Allport, Gordon Willard. 1954. *The Nature of Prejudice*. Basic Books.

Angrist, Joshua D and William N Evans. 1996. Children and their parents' labor supply: Evidence from exogenous variation in family size. Technical report National bureau of economic research.

Aristotle. 1954. *Nicomachean Ethics*. Translated and Introduced by Sir David Ross, Oxford University Press.

Athey, Susan and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2017. "Matrix completion methods for causal panel data models." *arXiv preprint arXiv:1710.10251* .

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches and Cyrus Samii. 2017. "Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect." *Journal of Labor Economics* 35(S1):S99–S147.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295.

Bond, Robert M, Jaime E Settle, Christopher J Fariss, Jason J Jones and James H Fowler. 2017. "Social endorsement cues and political participation." *Political Communication* 34(2):261–281.

Broockman, David E, Joshua L Kalla and Jasjeet S Sekhon. 2017. "The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs." *Political Analysis* 25(4):435–464.

Cartwright, Nancy. 2007*a*. "Are RCTs the gold standard?" *BioSocieties* 2(1):11–20.

Cartwright, Nancy. 2007*b*. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Coppock, Alexander. 2018. "Generalizing from survey experiments conducted on mechanical Turk: A replication approach." *Political Science Research and Methods* pp. 1–16.

Deaton, Angus. 2010. "Instruments, randomization, and learning about development." *Journal of economic literature* 48(2):424–55.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210:2–21.

Dehejia, Rajeev, Cristian Pop-Eleches and Cyrus Samii. 2015. From local to global: External validity in a fertility natural experiment. Technical report National Bureau of Economic Research.

DiMaggio, Paul, Eszter Hargittai et al. 2001. "From the digital divideto digital inequality: Studying Internet use as penetration increases." *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University* 4(1):4–2.

Egami, Naoki and Erin Hartman. 2018. Covariate Selection for Generalizing Experimental Results. Technical report Working Paper.

Fowler, Anthony. 2015. A Bayesian Explanation for Incumbency Advantage. In *111th APSA Annual Conference, San Francisco*.

Gechter, Michael. 2015. "Generalizing the results from social experiments: Theory and evidence from mexico and india." *manuscript, Pennsylvania State University* .

Gelman, Andrew and Zaiying Huang. 2008. "Estimating incumbency advantage and its variation, as an example of a before–after study." *Journal of the American Statistical Association* 103(482):437–446.

Green, Donald P and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly* 76(3):491–511.

Hargittai, Eszter. 2001. "Second-level digital divide: mapping differences in people's online skills." *arXiv preprint cs/0109068* .

Hargittai, Eszter. 2007. "Whose space? Differences among users and non-users of social network sites." *Journal of computer-mediated communication* 13(1):276–297.

Hargittai, Eszter. 2010. "Digital na (t) ives? Variation in internet skills and uses among members of the net generation." *Sociological inquiry* 80(1):92–113.

Hargittai, Eszter, Anne Marie Piper and Meredith Ringel Morris. 2018. "From internet access to internet skills: digital inequality among older adults." *Universal Access in the Information Society* pp. 1–10.

Hartman, Erin, Richard Grieve, Roland Ramsahai and Jasjeet S Sekhon. 2015. "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3):757–778.

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15(3):199–236.

Hopkins, Daniel J. 2018. *The Increasingly United States: How and Why American Political Behavior Nationalized*. University of Chicago Press.

Hotz, V Joseph, Guido W Imbens and Julie H Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125(1-2):241–270.

Imai, Kosuke, Marc Ratkovic et al. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Imbens, Guido. 2018. "Comments on understanding and misunderstanding randomized controlled trials: A commentary on Cartwright and Deaton." *Social science & medicine (1982)* .

Jacobson, Gary C. 2015. "Its nothing personal: The decline of the incumbency advantage in US House elections." *The Journal of Politics* 77(3):861–873.

Jones, Jason J, Robert M Bond, Eytan Bakshy, Dean Eckles and James H Fowler. 2017. "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US presidential election." *PloS one* 12(4):e0173851.

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill and Donald P Green. 2016. "Assessing methods for generalizing experimental impact estimates to target populations." *Journal of research on educational effectiveness* 9(1):103–127.

Knight, Frank H. 2012. *Risk, uncertainty and profit*. Courier Corporation.

Matias, J Nathan and Merry Mou. 2018. CivilServant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM p. 9.

Munger, Kevin. 2017*a*. "Dont@ Me: Experimentally Reducing Partisan Incivility on Twitter.".

Munger, Kevin. 2017*b*. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39(3):629–649.

Nguyen, Trang Quynh, Cyrus Ebnesajjad, Stephen R Cole, Elizabeth A Stuart et al. 2017. "Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects." *The Annals of Applied Statistics* 11(1):225–247.

Paluck, Elizabeth Levy, Seth A Green and Donald P Green. 2018. "The contact hypothesis re-evaluated." *Behavioural Public Policy* pp. 1–30.

Pettigrew, Thomas F and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90(5):751.

Pew Research Center. 2018. *Internet/Broadband Fact Sheet*. Pew.

Peysakhovich, Alexander and Dean Eckles. 2017. "Learning causal effects from many randomized experiments using regularized instrumental variables." *arXiv preprint arXiv:1701.01140* .

Redmiles, Elissa M, Neha Chachra and Brian Waismeyer. 2018. Examining the Demand for Spam: Who Clicks? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM p. 212.

Rosenzweig, Mark and Christopher Udry. 2016. External validity in a stochastic world. Technical report National Bureau of Economic Research.

Russell, Bertrand. 2001. *The problems of philosophy*. OUP Oxford.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

Settle, Jaime. 2018. *Frenemies: How Social Media Polarizes America*. Cambridge University Press.

Smith, Aaron and Monica Anderson. 2018. *Social Media Use in 2018*. Pew.

Stewart, Neil, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci, Jesse Chandler et al. 2015. "The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers." *Judgment and Decision making* 10(5):479–491.

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw and Philip J Leaf. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2):369–386.

Taddy, Matt, Matt Gardner, Liyun Chen and David Draper. 2016. "A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation." *Journal of Business & Economic Statistics* 34(4):661–672.

Trussler, Marc. 2018. The Effects of High Information Environments on the Incumbency Advantage and Partisan Voting. In *MPSA Annual Conference, Chicago*.

Vivalt, Eva. 2016. "How much can we generalize from impact evaluations?".

Wager, Stefan and Susan Athey. 2017. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* (just-accepted).