Knowledge Decays: Temporal Validity and Social Science in a Changing World

Kevin Munger*

September 2, 2019

Abstract

The "credibility revolution" has forced social scientists to confront the limits of our methods for creating general knowledge. The current approach aims to aggregate valid but local knowledge. At the same time, the increasing centrality of the internet to political and social processes has rendered untenable the implicit *ceteris paribus* assumptions necessary for aggregating knowledge produced at different times. The interaction of these two trends is not yet well understood. I argue that a high 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. A crucial distinction between cross-sectional external validity and temporal validity is that the latter implies a fundamental incompleteness of social science that renders the project of non-parametric knowledge aggregation impossible. I discuss the limitations of extant strategies for knowledge aggregation through the lens of temporal validity, and propose strategies for improving practice.

*kevinmunger@gmail.com

1 What is the Goal of Social Science?

Recent methodological innovations have consistently demonstrated that conducting internally valid empirical research is more difficult than once thought (Angrist and Pischke, 2010). 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 is a giant step forward, but it highlights a blind spot in the way that social science methods have been adapted towards this goal. 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 decays 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 a specific form of external validity: *temporal validity*. Recent methodological research has outlined conditions by which knowledge from a collection of contexts can be generalized to make predictions about a novel context, but these conditions cannot hold when the passage of time is taken seriously. The extent of this problem varies across different realms of inquiry; for most social science questions, 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 dramatically, extant social science research paradigms may be fatally inappropriate to the study of online behavior.

Globalization and an interconnected world imply that the issues I highlight may extend beyond the virtual. Consider the agricultural, medical and cartographic knowledge that has been accumulated over generations. The climate changes, but slowly enough that this knowledge accumulation had not been threatened. However, the recent advent of hyper-accelerated anthropogenic climate change poses a problem for all of these fields. The ideal time and location to plant a given crop—something that may have been more or less constant for hundreds of years—could change dramatically, in unpredictable ways. Using a related logic to the one I develop below, Gail (2016) suggests the possibility that the current rate of knowledge decay is increasing faster than the rate of knowledge production, implying that humanity's moment of "peak knowledge" is already behind us.

A provocative claim, but there's cause for optimism. Modern computation and communication methods have produced an exponential growth in knowledge production that can be harnessed to offset the decreased half-life of knowledge. I wholeheartedly endorse Munger (2019)'s call to raise the status of purely descriptive work in political science, particularly in the study of online politics. Qualitative description is essential for identifying novel research questions and identifying breakdowns in extant models. Quantitative description can be kept up to date at much lower cost than the causal knowledge produced by the research designs that populate top journals.

We should not abandon the goal of improving political decision making by increasing

the diversity of settings in which we can predict outcomes. At present, however, too much of our methodological innovation and too great a percentage of researcher energy is spent reinforcing the internal validity of the single research project. Taking the knowledge produced from each such endeavor and using it towards the goal of predicting the future requires an explicit model of the entire evidentiary chain. At the current margin, *most* of our energies should be devoted to improving our understanding of 1) how to aggregate the knowledge we have generated, and 2) how to take that knowledge and apply it to a given context *in the future*.

2 External Validity and Knowledge Aggregation

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 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]

[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."

[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]

Tolerably unbiased extrapolation has been shown to be empirically possible. Frequently replicated experiments that span both decades and the globe 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 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

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, the 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.

covariate non-overlap. Depending on their magnitude, frequency, and predictability, "macro shocks" pose a serious threat 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.

The interconnectedness of the contemporary world allows money, ideas and warfare to disrupt local systems faster and with fewer frictions than ever before.

Consider the case of the Arab Spring. While avoiding the simplistic argument by techno-fetishists that Twitter was a necessary and sufficient cause of the revolutions in the Middle East in the early 2010s, it is undeniably the case that the spread of information technology changed the nature of the strategic problem facing regimes and protesters (Tufekci, 2017). In particular, government censorship used to cause a favorable media ecosystem. Once enough people had cameraphones and internet access, this causal relationship failed to hold.

It may be the case that the internet is approaching maturity in the West, and that the past decade's disruptions were the result of a once-in-a-century technological advance (Karpf, 2019). But the interconnectedness of the global information sphere is here to stay.[4] Each new innovation in technology or its application can thus rapidly take over the world. The distribution of possible effects of each of these macro shocks on each of the relevant causal political relationships is unknowable ex ante. We need more and better technology for aggregating and applying past knowledge to novel contexts, techniques designed for the interconnected world in which we now reside. To that end, I will now discuss the limitations of the methods currently in use.

3 Generalizability

3.1 Non-Parametric 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

[4]Some of the more tech-savvy authoritarian regimes have taken steps to create their own national internets, most famously in China. Until and unless this happens, though, global interconnectedness and speed will continue to characterize information flows.

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 and many other policy evaluation contexts.

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.[5]

Proposed solutions to this problem involve reducing the treatment component $*$ 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

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

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 $*$ covariate space. A complementary approach uses machine learning to discover the covariate values with the largest effect heterogeneity, restricting the space in which matching or reweighting is necessary.

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

The fundamental problem of generalizability, then, is not that the covariate $*$ 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 is 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:[6]

[6]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 $x_{ij} \in X_t \cap X_{t+r}$

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

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

For a concrete example, consider censorship and the Arab Spring discussed above. In all of the "studies" conducted prior to the mid-2000s, the value of the covariate *Smartphone* was undefined.[7] After exogenous spread of smartphones, however, *Smartphone* takes some non-zero value, 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.

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

Knowledge of censorship prior to the Arab Spring *decays* when $Smartphone_i \neq 0$ 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.[8]

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. 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).[9]

My critique is related: *non-parametric approaches cannot achieve temporal validity.* Because time is unidirectional, the future will contain novel states of the world that

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

[8]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.

[9]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.

a given model *cannot* account for, regardless of how much data from the past it has access to. This has been known since Hume but is only now a consistent and serious issue for the practice of social science.

Even the standard practice of social science is imperiled. Replication is generally considered a key component of this practice. But true replication—of all but the most tightly controlled experiments—is impossible without some recourse to the a non-rigorous "similarity" between contexts. Given the time involved in someone publishing a scientific article and someone else reading it and developing the infrastructure to replicate it, this "similarity" is implausible on its face. As Nancy Cartwright has repeatedly argued, causal chains are only as strong as their weakest link—all of the rigor brought to bear to create internally valid knowledge is wasted without an equally rigorous way to generalize that knowledge.[10]

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.

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.

Our only recourse is to use theory to deal with these contexts; theory is the best guide to extrapolate knowledge of treatment effects to truly novel settings (Lucas, 2003). 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.*

Theory is thus necessary. So the challenge is to "bring it back in" while avoiding the pathologies that motivated the move to agnostic social science in the first place.

[10]Judea Pearl claims to have solved this problem, which he calls "transportability" for a given finding (Pearl and Bareinboim, 2014) and "data fusion" for the general task of aggregating knowledge from different contexts (Bareinboim and Pearl, 2016). Indeed, in a commentary on Deaton and Cartwright (2018), his primary critique is puzzlement that they (and everyone else) has not yet acknowledged that he has solved the problem (Pearl, 2018). I am unconvinced. Perhaps this is simply because I lack the capacity to understand Pearl's novel causal calculus, but I believe that the assumptions he outlines as "sufficient" for transportability are so stringent as to be useless. The framework holds promise, but a compelling empirical demonstration would be welcome.

3.2 Theoretical Approaches

The standard approach for social scientists is to develop, test and apply *theories*. Theories are, in essence, a *dimensionality reduction* technique: a way to take information from different (and potentially incommensurate) sources and encode it into language that can inform predictions in novel contexts.

The idea of using theories to inform decision-making is intuitive; indeed, human cognition seems optimized for discerning patterns and dreaming up causal theories to explain them. The enterprise of social science has in large part been premised on the idea that humans are the locus of knowledge aggregation. Our research produces knowledge that some human can incorporate into their decision-making process.

The above process merits explication because the thrust of research on human cognition, expertise and decision-making over the past decades has been to demonstrate just how *poorly* we tend to do this. Without belaboring a discussion of what is now well-known, research in psychology and behavioral economics have produced a laundry list of cognitive biases. Research on experts consistently finds that their predictions are only slightly better than chance (Tetlock, 2017).[11] As Deaton and Cartwright (2018) argue, the current preference for RCTs stems from the perception that they are "largely independent of 'expert' knowledge that is often regarded as manipulable, politically biased, or otherwise suspect."[12]

Human cognition is a challenging margin along which to improve, and the path that leads beyond these critiques is unclear. One promising avenue is to treat the acquisition of knowledge by social scientists with as much rigor as other subjects we study. Little and Pepinsky (2019) takes up this subject and makes the case that researchers need to think explicitly about their own prior beliefs about the world when learning from a novel piece of research. The present hesitance to put faith in expert judgment discussed above is potentially justified, but there's ultimately no transcending ourselves.

Regardless of the specific form a framework takes, any approach that treats the researcher as the locus of knowledge aggregation will need to incorporate methods

[11]The institutions of academic knowledge production have also been shown to be biased (Franco, Malhotra and Simonovits, 2014). Significant progress has been made on this dimension, and these biases are equally an issue for any kind of knowledge aggregation process, but it bears repeating that everything downstream of these biases is seriously imperiled by them.

[12]Although as Slough (2019) argues, the purported advantage that RCTs produce "agnostic" knowledge is mistaken. The researcher must specify a model of the world in order to precisely define the estimand of RCTs; the failure to do so is not agnostic but rather uses some implicit "shadow model" of the world that is left unevaluated.

for aggregating knowledge *among* researchers. At its simplest, this might entail the institutionalization of a survey of established political scientists that asks their position on disciplinary or policy debates. The Chicago Booth IGM Expert Panel of economists has achieved considerable success with this model, which serves as a useful resource for the general public in addition to helping economists determine the overall level of consensus on a given research question.

A more rigorous approach might involve an explicit "elicitation of priors" among experts in a given subject area, as proposed by Gill and Walker (2005). Again without wading into a larger debate about Bayesian statistics, a rigorous framework for creating synthetic theories from knowledge from disparate contexts would be a major contribution to the use of theoretical knowledge to inform decision making.

Having discussed limitations and paths forward for these approaches, I turn to a discussion of how social science is being done in the meantime.

4 Pragmatic Social Science

David Karpf's 2012 article "Social Science Research Methods in Internet Time" makes a series of arguments related to the one in this essay. 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."

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.*

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.*

The incumbency advantage serves as an accessible case. Here, 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 (Hopkins, 2018) or to the increased information environment provided by the internet (Trussler, 2018). 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

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

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.

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 modern information technology has changed the meaning of the

word *frequently*; following the discussion in section 3.1, r has increased.

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.

For most of the history of political science, the rate of change of the objects under study was generally not high enough to "intersect with" these time cycles. 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 or 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.

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 nationalization/party loyalty decreases but the incumbency advantage re-

[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*.

mains unchanged, we should lower our credence in that mechanism.[14]

Even if academic knowledge production keeps pace with the contemporary rate of change, 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 problem may be intractable, and it is possible that our ability to predict election outcomes peaked sometime in the past.

For research that studies online politics, however, the advantages of continuous, real-time data collection can be used to address the extremely high rate of change. However, taking full advantage of this data will require extensive adjustments to our research strategies and institutional frameworks. 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 2019, it is obvious that no effect of "the internet" has been constant over time. I turn to this point in the conclusion.

5 We Need More Description

The nascent field of data science offers some useful perspective on the topic of temporal validity. Data science is essentially atheoretical; the field is designed to take advantage of the recent explosion in data availability and computing power, so it privileges research questions that can credibly be answered with the data alone. A crucial component of this approach is out-of-sample prediction—absent theory, data science evaluates the validity of a model in its capacity to make accurate predictions on novel data.

This data tends to be temporally granular, and the half-life of a model trained on data from a particular time period can be explicitly measured. The term for this is "concept drift": the relationship between the data and the outcome to be predicted changes over time, in unpredictable ways. This is inevitable; old enough data is simply useless for today's prediction, and the knowledge encoded in the model trained on that data has fully decayed.

This framework—of continuously generating data and directly applying it to the near future—is the only endpoint for the non-parametric approach to temporal validity. It is also technically achievable, and would be the pinnacle of social science. However, it is ruinously expensive; an institution capable of internally valid knowledge of the topics

[14]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."

studied by political scientists (say, the effect of government censorship on protest) would be the most powerful entity in the world.

Given that we cannot hope to create knowledge as quickly, we need to think seriously about the tradeoffs inherent in different approaches to organizing the discipline. The credibility revolution has laid bare the cost of generating causal knowledge, but the necessary next step is a better framework for knowledge aggregation, synthesis and application. Both the non-parametric and theoretical approaches discussed above have serious limitations, and while neither is perfectible, each offers great promise.

This difficult work will take many years to reach any pragmatic consensus. In the meantime, there is a straightforward change to the practice of political science that will provide both an immediate knowledge windfall and will complement the rest of our research designs, increasing their temporal validity in perpetuity.

We must elevate the status of purely descriptive research designs in the discipline. Just like RCTs or other casual research designs, the quality of a given descriptive project can vary considerably. But we need to re-orient the discipline towards asking these kinds of questions, from graduate training to publication standards.

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. It is here that qualitative descriptive analysis is necessary. Before any statistical analysis could determine that the presence of smartphones changes the effect of state censorship, a human with deep understanding of the relevant context needs to observe that smartphones change how people communicate and theorize that these changes might shift the informational equilibrium.

Quantitative description enables us to evaluate the breadth of novel developments, allowing us to gauge their relative importance without resorting to our own biased intuitions. But it is also necessary for transporting knowledge across contexts, including both forward and backwards in time.

The non-parametric research on generalizability emphasizes the importance of fine-grained covariate data at the level of both the unit of analysis and the context in which that unit is situated. By adjusting for these covariate levels, a causal estimate can be credibly transported (albeit never perfectly). Not all covariates need to be adjusted for, only those which are causally relevant. The temporal challenge comes from the fact that we cannot know which covariates are causally relevant *ex ante*; that knowledge

is produced through the process of theorization done after the empirical research has been conducted.

Consider the case of social media adoption. We now think that social media use has an important effect on many political processes; comprehensive data about the pace and distribution of social media use would be a boon to scholars currently in possession of this theoretical insight. But they cannot now go back in time and begin to collect this data.

The need for this data is reflected in the citation counts on polls about social media use published by Pew Research. Pew's mission is simply to document trends of general interest, so they began asking about social media use well before it was of obvious relevance to politics. Duggan et al. (2015), a single survey with N = 3,500 about social media use habits, currently has 1,700 citations on Google Scholar. Lenhart et al. (2010), reporting longitudinal data on social media use among young people from 2006-2010, has over 3,000.

The other advantage of quantitative descriptive research is that it is *itself* more temporally valid. The ANES is among the most impactful research projects in the history of political science because of the power of holding the research instrument constant and simply allowing time to pass.

Unlike RCTs or surveys, some descriptive research can be kept up-to-date at next to zero marginal cost. Plenty of important data is generated by government statistics agencies or can be passively collected from online platforms. Models which transform this data into useful measures are highly temporally valid: they provide knowledge about the past, present and future.[15]

The credibility revolution has successfully steered empirical political science away from producing spurious knowledge. The next step will be to develop better technologies for aggregating and applying the knowledge we are now creating. The unceasing passage of time and accompanying decay of our stores of knowledge about political phenomena is a fundamental limitation to the knowledge aggregation process. For researchers studying social phenomena that change quickly, temporal validity is a first-order consideration.

[15]Many of the arguments for description in this section are laid out in a more general context in Gerring (2012)'s article "Mere Description." I avoid summarizing the entire article, but suffice it to say that temporal concerns are far from the only reason political scientists need to rethink our discipline's relationship with description.

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 Jörn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24(2):3–30.

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.

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

Bareinboim, Elias and Judea Pearl. 2016. "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences* 113(27):7345–7352.

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.

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.

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.

Duggan, Maeve, Nicole B Ellison, Cliff Lampe, Amanda Lenhart and Mary Madden. 2015. "Social media update 2014." *Pew research center* 19.

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*.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Gail, William B. 2016. "A new dark age looms." *New York Times* .

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

Gerring, John. 2012. "Mere description." *British Journal of Political Science* 42(4):721–746.

Gill, Jeff and Lee D Walker. 2005. "Elicited priors for Bayesian model specifications in political science research." *The Journal of Politics* 67(3):841–872.

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.

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)* .

Karpf, David. 2019. "Something I No Longer Believe: Is Internet Time Slowing Down?" *Social Media+ Society* 5(3):2056305119849492.

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.

Lenhart, Amanda, Kristen Purcell, Aaron Smith and Kathryn Zickuhr. 2010. "Social Media & Mobile Internet Use among Teens and Young Adults. Millennials." *Pew internet & American life project* .

Little, Andrew and Tom Pepinsky. 2019. "Learning from Biased Research Designs.".

Lucas, Jeffrey W. 2003. "Theory-testing, generalization, and the problem of external validity." *Sociological Theory* 21(3):236–253.

Munger, Kevin. 2019. "The Limited Value of Non-Replicable Field Experiments in Contexts With Low Temporal Validity." *Social Media+ Society* 5(3):2056305119859294.

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.

Pearl, Judea. 2018. "Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright." *Social Science & Medicine* .

Pearl, Judea and Elias Bareinboim. 2014. "External validity: From do-calculus to transportability across populations." *Statistical Science* pp. 579–595.

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.

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.

Slough, Tara. 2019. "On Theory and Identification: When and Why We Need Theory for Causal Identification.".

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.

Tetlock, Philip E. 2017. *Expert Political Judgment: How Good Is It? How Can We Know?-New Edition*. Princeton University Press.

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

Tufekci, Zeynep. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

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).