CA Journal of Cultural Analytics

November 23, 2022

Theory-Driven Statistics for the Digital Humanities: Presenting Pitfalls and a Practical Guide by the Example of the Reformation

Ramona Roller

Ramona Roller ETH Zurich. rroller@ethz.ch.

Peer-Reviewer:

Dataverse DOI:

ABSTRACT

The Digital Humanities face the problem of multiple hypothesis testing: Evermore hypotheses are tested until a desired pattern has been found. This practice is prone to mistaking random patterns for real ones. Instead, we should reduce the number of hypothesis tests to only test meaningful ones. We address this problem by using theory to generate hypotheses for statistical models. We illustrate our approach with the example of the European Reformation, where we test a theory on the role of opinion leaders for the adoption of Protestantism with a logistic regression model. Given our specific setting, including choice of data and operationalisation of variables, we do not find enough evidence to claim that opinion leaders contributed via personal visits and letters to the adoption of Protestantism. To falsify or to support a theory, it has to be tested in different settings. Our presented approach helps the Digital Humanities bridge the gap between the qualitative and quantitative camp, advance understanding of structures resulting from human activity, and increase scientific credibility.

The question of how to incorporate theory into research has been a recurring narrative in the Digital Humanities (DH). Critics at one end of the opinion spectrum disavow theory.[1] They argue that methodological debates are more easily resolved than theoretical ones[2] and that practitioners require guidance based on positivist rather than theoretical ideas.[3] At the other end of the spectrum critics condemn the lack of theory, i.e., humanistic values, such as deconstructionism, relativism, and poststructuralism, in the usage of DH tools, like maps and topic models.[4] In between these two extremes critics argue for a combination of theory and method to improve the interpretability of findings.[5]

Interestingly, such a debate about theory does not exist in the Social Sciences. The sister-discipline of the DH has agreed on a specific usage of theory within the scientific method to test hypotheses in a statistical model. How can the

DH profit from this theory usage and how can they help to address drawbacks of this usage? This article addresses both questions, first, by showing that theory usage with a statistical model improves the quality of data-driven analyses, and second, by encouraging the DH to use narratives to find a balance between theory-driven and exploratory analyses.

The DH and Social Sciences overlap in important areas. They are both interested in structures resulting from human activity, such as societies, cultures, texts, and paintings,[6] and they both use a data-driven approach resulting in several shared tools, such as sentiment analysis, topic modeling, and social networks. However, this data-driven focus makes both disciplines susceptible for data mining[7]: the unscientific practice of looking for patterns in the data until a desired one has been found. In this article, we address a particular but crucial case of data mining: the problem of multiple hypothesis testing (MHT). MHT carries the danger that spurious results are published, which decreases their credibility.[8] We explain how the Social Sciences address MHT by using theory and argue that the DH could not only copy this successful mechanism but also improve it by balancing it with exploratory approaches.

In the Social Sciences, the usage of theory is restricted to the application of the scientific method.[9] Based on an observation capturing a phenomenon of interest, one formulates a hypothesis (induction), tests it empirically with a model, and uses the result to infer new insights about the phenomenon of interest (deduction). Within the scientific method, a theory is used to derive a hypothesis. Formally, a scientific theory is a universal statement to explain, predict, and generalise outcomes resulting from an initial condition beyond the singular case.[10] Scientific theories are falsifiable, meaning that if new experimental observations are incompatible with theoretical expectations the theory is either dismissed or modified.[11]

For example, the theory of opinion leaders states that social change is brought about by important individuals.[12] This theory is falsified if one finds that social changes can occur without the support of opinion leaders. To provide a

use case of how theory-driven analyses can prevent MHT, this article applies the theory of opinion leaders to the example of the European Reformation. By testing to what extent famous reformers (the opinion leaders) affected the adoption of Protestantism in 16th century Europe (the social change), this article exemplifies the benefits of theory-driven statistics.

This article invites the DH and Social Sciences to learn from each other and adopt each other's methods when faced with similar problems. Theory-driven statistical analyses from the Social Sciences can help the DH to prevent unscientific data mining, and exploratory approaches from the DH can help the Social Sciences to account for data that are not readily available as a whole.

Previous Research on Testing Scientific Theories

Previous research in the Social Sciences has used theories to generate hypotheses in various subfields. The aim of the following overview is to provide exemplary cases of theory-driven research that may inspire future analyses of practitioners. We do not intend to provide a complete overview of the literature for theory-driven analyses.

For example, Box-Steffensmeier et al. used structural and interactionist theories of social roles[13] to define roles of interest groups in lobbying coalitions.[14] Iyengar and Westwood used social identity theory[15] to motivate their investigation of polarisation in the electorate along party lines.[16] Matthieß used mandate theory[17] to study the effect of pledge fulfilment of political parties on electoral outcomes.[18] Buggle used a theory of individualism and collectivism[19] to study how differences in societal collaboration have led to divergences in culture and technology.[20] Leal used migration systems theory[21] to study migration flows between countries.[22] Nelson used token theory[23] and racial domination theory[24] to study the use of social capital among ethnographic groups in settings where they represent the minority.[25] Light used a theory of legal decision making[26] to study discrimination in court based on

citizenship.[27] These examples can guide similar research in other disciplines, such as historiography.

With data and computational power now readily available, an opportunity has arrived for the DH to test scientific theories of the classic humanities. For example, in historiography, a subfield of the DH, there is substantial interest in applying statistical methods to existing problems, often trying to test theories concretely.[28] Famous theories include the theory of confessionalisation, describing the impact of the European Reformation on the formation of the modern state;[29] domino theory, relating the fall of the Roman Empire to pressure spreading from peoples outside the empire to those at its borders, which resulted in migration to the empire;[30] and Sonderweg (German for 'special path') theory, arguing that an authoritarian government in Germany was inevitable after the Weimar Republic because of the nation's unique history and development.[31]

The Scientific Theory in Practice

How do we test theories with quantitative methods ideally? We formulate a research question, connect it to an existing theory, and translate aspects of this theory into a testable hypothesis. Suppose our research question states: 'Why did peasant revolts in 16th century Europe occur?', and our specific driving factor of interest is the occurrence of famines. To statistically test how famines affected peasant revolts, we start with a conservative assumption: Famines do not affect the probability of revolts. This assumption of the lack of an effect is called the *null hypothesis* and is tested in a statistical model at a certain confidence level.

Two outcomes are possible: First, if the model provides enough evidence for the effect of famines on peasant revolts, we reject the null hypothesis. If the effect is positive (negative), we infer that famines make peasant revolts more (less) likely. Second, if the model does not provide enough evidence for an

effect of famines on peasant revolts, we fail to reject the null hypothesis. We infer that famines did not affect peasant revolts. Importantly, failing to reject the null hypothesis does not mean that we proved that the effect of interest does not exist. It only means that we did not find enough evidence to claim that the effect exists. With respect to the famine-revolt example, this means that famines may have affected revolts in reality, but our model did not detect this connection.

Both conclusions (famines affect or do not affect peasant revolts) are not absolute because we test the null hypothesis at a certain confidence level. This means that we have a certain probability of drawing the wrong conclusion for each case, i.e., to make an error. In the first case, the error means that we think that famines affected peasant revolts, although they did not (type I error). In the second case, the error means that we think that famines did not affect peasant revolts, although they did (type II error).

The Problem of Multiple Hypothesis Testing

Conducting many hypothesis tests is problematic because it undermines the definition of statistical significance. Statistical significance indicates whether an effect can be attributed to a factor of interest or chance. If we attribute an effect to chance, we fail to reject the null hypothesis. If we attribute an effect to a real pattern, we reject the null hypothesis.

In statistical models, significance is represented by the p-value. It represents the confidence level of the statistical test, i.e., the probability of finding a pattern in the data when in fact, this pattern does not exist. In statistics jargon, the p-value is the probability of conducting a type I error or finding a false positive, which is equivalent to rejecting the null hypothesis when it is true. The counterpart of the type I error is the type II error or false negative, where one fails to reject the null hypothesis when it is false.[32] In statistical models, we try to reduce the probabilities to commit type I and type II errors as much

as possible, e.g., by increasing the sample size. However, the probabilities of type I and type II errors are always larger than zero.[33]

To illustrate the problem of multiple hypothesis testing (MHT), we return to our previous example of peasant revolts. Suppose we have 20 potential driving factors available whose effect on the probability that a peasant revolt occurs we can statistically test, such as the socio-economic situation of peasants, the type of rule of their feudal lord, and the influence of famines. If we test the 20 factors separately, we test 20 hypotheses, one for each driving factor. Suppose that, in reality, each of these factors does not affect peasant revolts (null hypothesis is always true). Of course, this information would be unknown in real-world analyses. The hypothesis test aims to reveal this actual pattern from our data.

For each of the 20 hypothesis tests, we choose an acceptable significance level (α). α is the maximum probability with which we allow ourselves to commit a type I error. For this example, we choose $\alpha = 0.05$ for each test, meaning that we accept a 5% chance to commit a type I error. Assuming all 20 hypothesis tests are independent, the significance level over all hypothesis tests combined (called 'experimentwise significance level') will be given by $1 - (1 - \alpha)^n$.[34] In this equation, α is the acceptable significance level of an individual hypothesis test, i.e., 0.05, and *n* is the number of hypothesis tests, i.e., 20. So, with 20 hypothesis tests being conducted, we have a 64% ($1 - (1 - 0.05)^{20} \approx 0.64$) chance of observing at least one significant result, even if all the individual tests are not significant. The experimentwise significance level under MHT represents a drastic increase from the accepted 5%, which we chose initially. MHP increases the probability of getting a significant result simply due to chance. The large experimentwise significance level falsely indicates to us that some of the significant driving factors affected peasant revolts.

The problem of MHT is not specific to the study of culture and history but a problem in all empirical research relying on statistical inference. One falls

into the trap of thinking that the real pattern can be found with an exhaustive trial and error procedure and that every result of this procedure reveals a valid pattern, i.e., is interpretable. The following three unscientific practices are commonly used and illustrate this trap. First, one looks for patterns in visualisations without statistical testing whether these patterns are random or likely to be real. Second, one optimises the parameters of models without being able to interpret the 'optimal' parameter value, such as accepting a value of zero years for the parameter 'age of a person'. Third, one tests the effect of as many variables as possible on an outcome measure and chooses the variables that explain the outcome best, without correcting for multiple hypothesis testing.

The challenge is to reduce the number of hypothesis tests while testing meaningful hypotheses. That is, if we test a hypothesis, we have to be convinced that it is an interesting one to check. This approach increases the chances that the patterns we find are likely to be real and not random, which boosts science's credibility.

Turning Theories into Hypothesis Tests

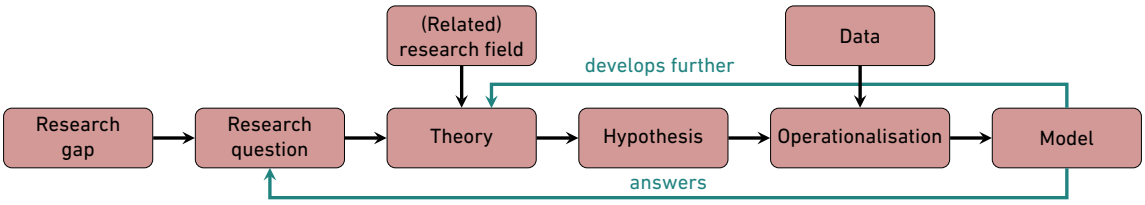

Figure 1: Schematic approach to address multiple hypothesis testing.
Interpretation of model

Figure 1 shows how we can use theory to address the problem of multiple hypothesis testing. Theories that can be tested statistically are conceptual tools that interrelate measurable concepts of interest. Concepts of interest may be the adoption of Protestantism, socio-economic status, hegemonic structures, or the climate. A theory specifies under which circumstances these concepts occur and can be falsified when tested empirically.

Based on a research gap, we formulate a research question and select a theory that deals with this question. This theory can be a historiographical one or one from a related field. With the theory, we formulate specific hypotheses by which we restrict the number of tested variables in the model. So we test the available hypotheses and do not continue testing if we do not find an effect for the formulated hypotheses. Usually, one generates more than one hypothesis from the theory, so we still have to account for multiple hypothesis testing. We can use established statistical correction procedures such as Bonferroni or Benjamini-Hochberg.[35]

Like a theory, a hypothesis connects measurable concepts but defines the connection more precisely. For example, whereas a theory would state that the Reformation drove the formation of the modern state, a related hypothesis would state that if a territorial ruler adopts Protestantism, he is more likley to establish institutions in the territory.

Operationalisation translates abstract concepts into measurable quantities, which can be included as variables in the statistical model. The result of the model indicates whether we reject or fail to reject the null hypothesis. Based on this finding, we can answer our initial research question and provide evidence or counterevidence for the theory. By testing a theory in different settings, we can falsify it, which helps to differentiate the theory further.

Example: Reformation

To illustrate how to integrate theory into a statistical analysis in historiography, we apply the approach from the previous section to a use case of the Reformation. The following analysis is a case study to exemplify the usage of theory-driven statistical models in the DH. It does not provide a comprehensive historiographical study about the Reformation but rather deliberately simplifies historical concepts and the statistical approach to be of interest to a broader DH audience and to offer a template for practitioners.

The Reformation was a socio-transformative movement in 16th century Europe, which overthrew the catholic church, established protestant denominations (e.g. Lutheranism), and initiated political changes. Understanding the Reformation better is relevant because the Reformation is associated with various developments that shaped our modern life, such as the formation of the national state,[36] justification of communist policies in East Germany,[37] and its impact on economic growth, which was first hypothesised by Max Weber[38] is still highly debated among researchers today.[39]

Research gap. The question of why the Reformation took hold in some places but not in others has been addressed by many generations of historians.[40] Often, the adoption of Protestantism was associated with the confessional decision of a territorial ruler. In the 16th century, central Europe was politically divided into many territories, each governed by a prince who decided for his subject which denomination to adopt, i.e., whether his territory should become Protestant or remain Catholic.

Previous research has used this policy to investigate driving factors for the adoption of Protestantism in territories. Qualitative historiographical research has analysed individual territories in isolation.[41] Quantitative historiographical research has examined several driving factors across territories.[42] However, when focusing on the effect of human individuals, quantitative research has often taken a Luther-centric view. For example, studies have analysed the impact of territories' distances to Wittenberg (Luther's place of residence) and the impact of Luther's students on the adoption of Protestantism.[43] However, Luther was not the only person spearheading the Reformation. Thousands of other reformers exchanged ideas via letters and personal visits. We lack a perspective that studies the combined influences of all these reformers on the adoption of Protestantism in the territories.

Research question. How did reformers drive the adoption of Protestantism in territories?

Theory. To address the research question, we rely on a theory of the role of opinion leaders for behaviour change.[44] The theory describes the importance of selected individuals for the adoption of ideas or products among a larger group of people. To our knowledge, the theory has not been applied in a historical context.

The definition of 'opinion leaders' as well as their identification differ based on the research context.[45] Opinion leaders are individuals who leverage their reputation to convince others to adopt an idea. Reformers embody this role. They were primarily theologians, like Martin Luther, but also included noblemen, like Philip of Hesse, and other scholars, who supported the Reformation and had a high social standing because they occupied important offices. For example, Georg Spalatin was the secretary of the Saxon Elector Frederick the Wise, Joachim Vadian was Dean of the University of Vienna, and Martin Bucer was the pastor of the largest gild in Strasbourg. Given these influential positions, we assume that reformers had substantial means to spread their confessional convictions.

However, not all characteristics of reformers match those of opinion leaders from the theory. Reformers varied in their commitment to spread the Reformation. Some were fanatic, even willing to die for their faith (martyrdom was especially common among Baptists, a protestant denomination); others were more moderate. Some were leaders, setting the first steps, others were more reluctant, waiting to spread the Reformation until the political situation was calmer. Due to this variety, reformers fulfilled the role of opinion leaders to different extents. We test the theory of opinion leaders in the context of the Reformation by assuming that, mostly, reformers are opinion leaders. Specifically, we analyse how these reformers affected the adoption of Protestantism in the territories.

Hypothesis. The higher a territory's exposure to reformation ideology through reformers, the larger its chance to become Protestant.

Operationalisation. We consider reformers to represent opinion leaders who spread their ideas in the territories. This spread of ideas can happen via several mechanisms, and we limit our attention to two of those. First, reformers can physically visit the territory to preach, advise the ruler, attend disputations, or inspect whether protestant rules are implemented correctly (so-called 'visitations'). Second, reformers can send letters to individuals living in a territory and convey their ideas via the letter text.

To operationalise the impact of visits and letters, we use a data set of letter correspondences. It consists of the letter editions of nine notable reformers[46] and comprises $3,370$ individuals and $26,663$ letters. We use the sending date and the sending and receiving locations of letters to infer the time reformers have stayed in a territory and to which territory a letter was sent.

We operationalise the impact of physical visits as the number of days a reformer spent in a territory before it became protestant. Since we assume that the impact of visits decreases over time (individuals forget whom they met 10 years ago), we time-weigh each visit. That is, we walk over the years from the foundation year of the territory until it either becomes protestant or ceases to exist. For every year after a reformer had visited a territory, we assume that the impact of the visit decreases. This decrease of impact is modelled with an exponential decay function with a half-life of 15 years. This half-life means that after 15 years (a generation), a visit has become half as influential compared to the day where the visit happened.

As a result, we obtain a series of values for each visit a reformer conducted. Each series runs from the day a reformer visited a territory to the day the territory either became protestant or ceased to exist. The first value in the series is always one because we assume the impact of the visit to be most prominent

on the day the visit occurred. As the series proceeds, the values decrease according to the exponential decay function. So if a reformer visited a territory on January 1st 1520, the impact of the visit is more considerable on that day than a day later on January 2nd 1520. To summarise the impact of physical visits across reformers and time, we merge the values from all series and take the mean. We call the resulting variable `visits`.

To operationalise the impact of letters, we count the number of letters a reformer sent into a territory before that territory became protestant. As for the `visits` variable, we assume that the impact of a letter decreases over time, which is why we apply the same time-weighing as for `visits`. For every day after a letter was sent to a territory, the impact of the letter decreases according to an exponential decay function with a half-life of 15 years. For each letter a reformer sent to a territory, we obtain a series of values, similar to `visits`. To summarise the impact of letters across reformers and time, we merge the values from all series and take the mean. We call the resulting variable `letters`.

To construct the dependent variable, whether or not a territory adopted Protestantism, we manually collected the denominational adoptions of territories in the 16th century using a historiographical book series on territories that were relevant during the Reformation.[47] To simplify the analysis, we only track the first switch from Catholicism to Protestantism of territories. This data set comprises 262 territories, of which 192 became protestant and 70 remained catholic.

We map the conceptual hypothesis from above to the proposed independent variables and generate two variable-related hypotheses. We formulate each of these variable-related hypotheses in two versions: as null hypothesis (H_0) which is tested in the model and as alternative hypothesis (H_a) which specifies the outcome that we try to infer from the model.

$H1_0$ *The number of days reformers spend in a territory does not affect its*

chance to become Protestant.

$H1_a$ *The more days reformers spend in a territory, the larger its chance to become Protestant.*

$H2_0$ *The number of letters reformers send to a territory does not affect its chance to become Protestant.*

$H2_a$ *The more letters reformers send to a territory, the larger its chance to become Protestant.*

Model. For this example, we chose a logistic regression. This model estimates a binary outcome (becoming protestant vs remaining catholic) from a set of independent variables (`visits` and `letters`). Logistic regression is an established model for inferential statistics and has been widely used in many fields.[48]

Note that for the example at hand, logistic regression is not the optimal model. Logistic regression does not take into account temporal changes in the territories before they became protestant. Moreover, the model makes assumptions that do not capture the situation in the 16th century, e.g., the model assumes that a territory's decision to become Protestant happened independently of other territories, which was not the case. Last, our model ignores control variables,[49] such as whether a territory was involved in a military conflict. Since our aim is not to 'proof a theory' but rather to present a testbed for theory-driven statistics, we use logistic regression without control variables for this example. For our example, the formal definition of the model is:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \texttt{visits} + \beta_2 \texttt{letters} + \epsilon \tag{1}$$

p is the probability that a territory becomes protestant. $1-p$ is the probability that a territory remains catholic. $\frac{p}{1-p}$ is the odds ratio and indicates how much more likely a territory is to become Protestant than to remain Catholic. *log*

corresponds to the natural logarithm, and $\log \frac{p}{1-p}$ is called the log odds. By using the logarithm, the right-hand side of the equation is linearised, which facilitates parameter estimation.[50]

The βs are the coefficients that the model estimates. β_1 and β_2 indicate how large the effects of the independent variables on the dependent variable are. Hence, how important `visits` and `letters` are for the adoption of Protestantism. β_0 is called the intercept and indicates, in this particular setting, how likely territories are to become Protestant when all other variables are set to zero, i.e., reformers neither visited territories nor sent letters to them. The βs indicate the average change in the dependent variable if the respective independent variable changes by one unit, and the other independent variables are held constant. We can interpret the βs as odds ratios by exponentiating them (e^{β}). For example, if $\beta_1 = 0.3$, then $e^{\beta_1} = e^{0.3} = 1.35$. This result means that for every additional day a reformer spends in a territory, a territory is 1.35 times more likely to become Protestant than remain Catholic.

ϵ is the model's error term and captures the variance in the data that the model does not explain. The smaller ϵ, the better `visits` and `letters` explain the adoption of Protestantism, i.e., the model is good. Last, for all hypothesis tests in this example, we choose an acceptable significance level of 0.1, meaning that we accept a 10% chance to commit a type I error.

Interpretation. Before we interpret the effect of `visits` and `letters` on the adoption of Protestantism, we test whether the overall model is good. For this, we run a global F-test.[51] The F-test compares the residuals of the tested model, to the residuals of the model where only the intercept term is included, i.e. a horizontal line is used to describe the data. Residuals measure the difference between the adoption of Protestantism predicted by the model and the real-world adoption which is captured in the data. The corresponding null hypothesis of the F-test states: the tested model is no better (in terms of likelihood) than a model fit with only the intercept term. The p-value of the F-

Table 1: Logistic regression results to explain the first switch to Protestantism of territories. `Letters` *represent the mean time-weighted number of letters sent by reformers to a territory before it became protestant or ceased to exist.* `Visits` *represent the mean time-weighted number of days reformers spent in a territory before it became protestant or ceased to exist. Left: no correction for multiple hypothesis testing. The significance level is set to* 0.1*. Right: Bonferroni correction. The significance level is reduced to* 0.05*, i.e., the uncorrected significance level of* 0.1 *is divided by* 2*, the number of tested hypotheses. With the Bonferroni correction,* `letters` *is no longer significant.*

	No correction sig. level = 0.1	Bonferroni sig. level = 0.05
Intercept	$0.8806\ (0.1481)^{***}$	$0.8806\ (0.1481)^{***}$
Letters	$1.8542\ (1.0755)^{*}$	$1.8542\ (1.0755)$
Visits	$-0.0090\ (0.0069)$	$-0.0090\ (0.0069)$
AIC	300.8645	300.8645
BIC	311.5695	311.5695
Log Likelihood	−147.4322	−147.4322
Deviance	294.8645	294.8645
Num. obs.	262	262

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

test is 0.00966. This result means we have a 0.97% chance of falsely thinking that the intercept model is better than the tested model. Since this percentage is smaller than the chosen 10%, we are confident to assume that our tested model is overall useful.

Table 1 shows the results of the logistic regression. The numbers outside the brackets refer to the estimates of the βs for the independent variables `visits` and `letters`. The numbers in brackets correspond to the standard error of the estimates.[52] The columns represent the results of the same model but differ in whether they correct for multiple hypothesis testing. Hence, the estimates are the same (numbers), but their statistical significance changes (stars).

We see that the estimate for `visits` is −0.009. This result corresponds to the average effect a day spent by a reformer in a territory has on the log odds of

becoming protestant vs remaining catholic. Since it is easier to interpret the odds ratio, rather than the log odds, we remove the logarithm by exponentiating the estimate: $e^{-0.009} = 0.99 \approx 1$. An odds ratio of 1 indicates that, on average, a territory is as likely to become Protestant as to remain Catholic for each additional day a reformer spends in a territory. This result means that, according to the model, physical visits of reformers did not affect the adoption of Protestantism in territories.

The estimate for `letters` is 1.8542. This result means that a territory is 6.39 times ($e^{1.8542} = 6.39$) more likely to become Protestant than to remain Catholic if a reformer sends a letter to a territory.[53] Translated into probabilities, this means that if the reformer sends one letter to a territory, the territory has a 93.90% ($\frac{e^{(0.8806+1.8542)}}{(1+e^{(0.8806+1.8542)})}$) probability to become Protestant.[54] This may seem a lot, however, to correctly interpret the effect, we have to compare it to the baseline, i.e., the probability of a territory to become Protestant if reformers send no letters to the territory. The baseline probability is 70.69% ($\frac{e^{0.8806}}{(1+e^{0.8806})}$) and can be computed from the intercept term in Table 1. So for every letter reformers send to a territory the probability of the territory to become Protestant increases by 23.21% (90.93% − 70.69%).

In the first column, we see that `letters` is significant at the 0.1 level (one star), whereas `visits` is not (no stars). The star is missing because the chance to falsely think that `visits` affects the adoption of Protestantism is larger than 10% (probability to commit type I error). Since we consider a probability to commit a type I error above 0.1 too large for a correct analysis, we conclude that insufficient evidence is provided to conclude that `visits` affects the adoption of Protestantism.

Since we test two independent variables in the model, `visits` and `letters`, we test two hypotheses. If we test multiple hypotheses without correction, we increase the probability of finding an effect, even if it does not exist. This is why we have to reduce the significance level, which is what we do with the Bonferroni correction (2nd column). The corrected significance level is the

old one divided by the number of hypotheses.[55] In our case, this would be $0.1/2 = 0.05$. In the second column, 0.05, rather than 0.1, is used as significance level. We see that the star for `letters` disappears, indicating that `letters` is not significant at the 0.05 level. This result means that the observed effect is too small to be considered different from random. At the 0.1 level, the accepted difference between random and real effect was allowed to be smaller.

This result shows that we fail to reject the null hypothesis after applying multiple hypothesis testing with Bonferroni correction. Neither the time reformers spent in territories nor the number of letters they sent to territories affect the adoption of Protestantism in territories. This result does not provide support for the theory of opinion leaders in our specific setting.

To prevent wrong conclusions from these results, we rebut some common misinterpretations of statistical models. Our results do not show that physical visits and letters of reformers were irrelevant for the adoption of Protestantism. In contrast, the results indicate that, given our data set and our chosen operationalisation, the model does not provide sufficient evidence to conclude that physical visits and letters affected the adoption of Protestantism. Had we included different letters in our data or measured the presence of opinion leaders differently, we might have received different results. Moreover, our results do not show that the theory of opinion leaders is wrong. In contrast, we show that our specific model did not find support for this theory. So in our chosen setting, we did not find evidence for this theory, whereas, in other settings, it may still hold. These remarks show that hypotheses and theories are tested in particular settings defined by the model specifications. To support or to falsify a theory, several models should be tested, which allows for generalisation of the results.

We can draw three major conclusions from our results: First, we showed that it makes sense to test the proposed model. As the F-test showed, the proposed model is better than the intercept model, which indicates that the chosen inde-

pendent variables carry value. Second, we see that letters were more important for the adoption of Protestantism than personal visits because the β for `letters` is larger than the one for `visits`. This may indicate that different communication media contribute to the spread of ideas to different extents. Third, our results provide new research questions and testable hypotheses. For example, we may ask whether the large effect of `letters` was driven by all or a subset of reformers. We could hypothesise that representatives of different protestant denominations affected the adoption of Protestantism in territories to different extents. For example, Baptists were more radical in their views than Lutherans because they also wanted to change the worldly order, not just the inner faith.[56] Baptist claims would have decreased the power of territorial rulers who, in order to keep their position, might have rather supported Lutherans. One testable hypothesis could state: Lutherans have a larger impact on the probability that a territory becomes protestant than Baptists.

Addressing Doubts against a Theory-driven Statistics Approach

We address major concerns against theory-driven statistics, which qualitative researchers often put forward. To illustrate these concerns, we refer back to the Reformation example from the previous section.

Imprecision. *Allegation:* Abstract historiographical concepts are too imprecise to be operationalised. For example, the 'adoption of Protestantism' has many different meanings (among rulers, laypeople, and scholars; public vs private behaviour; etc.), which cannot be put into numbers.

Counterargument: The aim of operationalisation is neither to capture *all* meanings of a concept nor its *correct* meaning. It is to capture *one* meaning of a concept and to justify why this specific meaning and the specific translation into a measure are useful for the analysis at hand. If others reject this justification, a new operationalisation of the same concept can be proposed

and compared to the existing one. This comparison is highly valuable because it indicates whether different treatments of the same concept affect the results.

Exceptions. *Allegation:* The testable theories presented in this paper are too general to capture all the historical exceptions. For example, eastern Europe was not affected by the same dynamics of the adoption of Protestantism as core lands of the Holy Roman Empire.[57]

Counterargument: As presented in this paper, the aim of a theory is to generalise findings across multiple settings. Using theories facilitates understanding since individual cases do not need to be studied in isolation but can be interrelated via the theory. If a theory is too general, it falsely claims to generalise a finding from one setting to another. That is, the scope of the theory had not been determined correctly. Statistics measures external validity, which indicates how generalisable the findings are to other settings, namely from the data sample at hand to the population of interest. Whether or not a theory is too general can only be known *after* the theory is tested. In theory-driven statistics, the theory serves as a starting point and is differentiated into subtheories accounting for exceptions in the data that require a different explanation.

Oversimplification. *Allegation:* The presented model cannot consider all relevant explanatory factors; hence, it is oversimplified. For example, in the model of this paper, migration flows and climate (cf. small ice age) also contributed to the adoption of Protestantism of the territories but are not included.

Counterargument: The omission of critical explanatory factors in statistical models is a well-known problem, called 'omitted variable bias'.[58] Since omitted variables violate the assumptions of a model, the induced bias can be detected when these assumptions are tested. For example, simple regression[59] assumes that the explanatory variable and the error terms are unrelated.[60]

Once a systematic relation between error terms and explanatory variable exists, the model is misspecified, indicating that a variable was omitted. However, this check does not indicate *which* variable is missing.

Omitted variable bias addresses the consequences if a model leaves out relevant variables. However, it does not pick up the omission of less relevant variables, i.e., variables which still explain an outcome, but only to a small extent. Statistical modelling aims to deliberately exclude these minor variables. Important explanatory factors rather than all of them have to be captured.

Historical determinism. *Allegation:* By explaining a historical event or process with precisely defined factors, statistical models imply clear cause-effect chains without uncertainties. This practise is historical determinism, which is wrong.

Counterargument: No, statistical models do not support a deterministic view of history. These models test to what extent a selection of factors affects a precisely defined outcome. In addition, other factors, including chance, also affect that outcome. These factors are included in the error term of the model (ϵ). The error term indicates how much variance in the data persists, which cannot be allocated to the tested factors.

Moreover, the existence of uncertainties does not mean that all events are equally likely. We live in a world of tailed, not uniform, probability distributions. For example, it is more likely that a revolt was caused by famine than by choice of shoe colour of some individuals. The model aims to capture factors which correspond to high probabilities because they capture the broad patterns which shape history.

Causation vs effect. *Allegation:* It is impossible to disentangle causation from effect, so statistical models cannot do it either.

Counterargument: Everyday examples show that we can disentangle causation from effect. If a person accidentally breaks a glass, we know that touching the glass caused it to break, rather than the other way round, although the two events seemingly occurred together. In historical contexts, cause and effect are also distinguishable, they are only difficult to identify. Many statistical tools have been developed to investigate causal relations in observational data, such as difference in differences,[61] event history models,[62] propensity score matching,[63] and instrumental variables,[64] as well as comparisons of them.[65]

We can use these tools to enrich simpler correlation-based models. These correlation-based models only claim that there is an association between an explanatory factor and an outcome but do not make claims about the direction of causality between the two. For example, a territorial ruler becoming protestant may have attracted more theologians into the territory. Alternatively, the visits of theologians may have convinced the ruler to become Protestant. To understand why the ruler became protestant in our example, we analysed the visits of theologians over time up until the point where the ruler became protestant, so his confessional switch cannot affect later visits.

Populist. *Allegation:* Researchers only develop theories to become famous. Since a theory distils something complex to something simple, individuals understand it better and like it more. Hence, theories are populist instruments.

Counterargument: Yes, a theory is a tool for simplification, but in the sense of revealing the structure behind complexity in the world. This simplification is subject to bounding assumptions and is only valid in specific situations. Making these assumptions and situations transparent is essential to create complementary theories and develop old theories further once new insights are available. Using any scientific insights as gatekeeping tools to prevent others from accessing that knowledge or to show off to others does not help anyone

and should not be done.

Conciliating theory-driven and exploratory research

The focus on theory-driven research in the Social Sciences has lead to various measures to further reduce the chance of spurious results. The pre-registration of hypotheses has become an accepted standard to prevent MHT.[66] Rigorous justification of tested explanatory variables[67] is supposed to strengthen theoretical foundations. Simpler models are preferred over complex ones (cf. Occam's Razor)[68] which should reduce the ambiguity of model results. Strict criteria for sampling help counteract selection bias. Whereas these measures tend to increase the confidence of social scientists in their results, they constrain research in the DH.

Data in the DH is usually observational rather than collected in experiments making the pre-registration of hypotheses obsolete. Sampling restrictions are rather a question of availability than of theory since data are scarce and the underlying databases grow slowly over time due to resource intensive digitisation and editing steps. The tight rules of theory-driven research in the Social Sciences tend to contradict relativist values of the DH, where one adheres to alternative explanations long into the analysis.

To satisfy the needs of the DH, exploratory research could be given more room within data-driven analyses. However, since the Social Sciences consider the dangers of exploratory research to outweigh its benefits, they tend to stick to a strictly theory-driven approach and are unlikely to address the needs of the DH.[69] The DH are required to decide on their optimal relation between theory-driven and exploratory research that is compatible with their research setting and goals. The established culture of theory narratives within the DH could provide a useful tool to engage in this debate.

Conclusion

The Digital Humanities (DH), like other data-intensive disciplines, face the problem of multiple hypothesis testing: Multiple hypotheses are tested until a desired pattern is found. Without correction, this approach is prone to mistaking random patterns for real ones. By using theory to formulate hypotheses, we restrict the number of hypothesis tests. By using statistical corrections, such as Bonferroni, we account for the remaining hypothesis tests.

As an example, we tested a theory on the role of opinion leaders for the adoption of Protestantism in territories during the Reformation in 16th century Europe. This theory is testable because it interrelates measurable concepts, the impact opinion leaders have on others via their communication and the adoption of Protestantism. Due to this testability, we can falsify this theory with statistics. Based on the assumption that reformers, such as Martin Luther, represent opinion leaders, we have operationalised their presence in territories with their number of days spent there (`visits`) and with the number of letters they sent to the territory (`letters`). After having corrected for the two tested hypotheses (`visits` and `letters`), none of the tested variables was significant. We failed to reject the corresponding null hypotheses and therefore did not find enough support for the theory in our specific setting.

On the one hand, our example illustrates the importance of theories of statistical analyses in the DH. First, theory enables us to test specific hypotheses and to distinguish random from real patterns. Through this process, theories are either supported or falsified which advances our understanding of the phenomenon of interest. Second, theory enables us to compare studies systematically. We could use a different data set, operationalisation and model to test the theory of opinion leaders on the adoption of Protestantism and compare the results with those of this paper. This process increases the robustness of results and hence their credibility. Third, theory guides research through establishing a basis of knowledge that can be taken for granted by future research, which does not have to establish that basis again.

On the one hand, our analysis has shown that a complete theory-driven focus may constrain the DH in their relativist approach. Building on their established narrative culture, we argue that the DH posses a promising tool to modify the theory-driven focus borrowed from the Social Sciences to their needs. Specifically, the DH could look for a new balance between theory-driven and exploratory research to account for characteristics of the data and to give more value to alternative explanations of results.

As a future outlook, we argue that theory can bridge the gap between qualitative and quantitative camps in the DH, which emerged due to the digitisation wave. Representatives of the quantitative camp tend not to believe case studies because the selected case may not mirror the broad lines of the phenomenon of interest, which they consider to be necessary. Representatives of the qualitative camp tend not to believe numbers because they oversimplify cases and only reveal what is already known. This blame in both directions does not advance the DH. The two camps should join efforts and focus on their common aim: understanding human activity and the resulting structures better. We discussed theory-driven statistics as one concrete methodological step towards this aim.

Notes

[1]Chris Anderson, "The end of theory: The data deluge makes the scientific method obsolete," *Wired magazine* 16, no. 7 (2008): 16–7.

[2]Tom Scheinfeldt, "Why Digital Humanities Is 'Nice.'," *Found History* 26 (2010).

[3]Nigel A Raab, "The End of Theory in the Humanities," in *The Humanities in Transition from Postmodernism into the Digital Age* (Routledge, 2020), 70–88.

[4]Johanna Drucker, "Humanistic theory and digital scholarship," *Debates in the digital humanities* 150 (2012): 85–95; Gary Hall, "Toward a postdigital humanities: Cultural analytics and the computational turn to data-driven scholarship," *American Literature* 85, no. 4 (2013): 781–809; Alan Liu, "The state of the digital humanities: A report and a critique," *Arts and Humanities in Higher Education* 11, nos. 1-2 (2012): 8–41.

[5]David M Berrry et al., *No signal without symbol: Decoding the digital humanities*, 2019.

[6]David Budtz Pedersen, "Integrating social sciences and humanities in interdisciplinary research," *Palgrave Communications* 2, no. 1 (2016): 1–7; Humanities Paul Rosenbloom, "Toward a conceptual framework for the digital humanities," in *Defining Digital Humanities* (Routledge, 2016), 235–50.

[7]Data mining is also known under several other names, including data dredging, data fishing, data snooping, data butchery, significance chasing, significance questing, selective inference, p-hacking or data piñata; Ronald L. Wasserstein and Nicole A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician* 70, no. 2 (2016): 129–33, doi:10.1080/00031305.2016.1154108; George D. Smith and Ebrahim Shah, "Data dredging, bias, or confounding," *BMJ (Clinical research ed.* 325, no. 7378 (2002): 1437–38, doi:10.1136/bmj.325.7378.1437; Simon Lindgren, "Beyond Method," chap. 1 in *Data Theory* (Medford: polity, 2020), 23; David Garcia, "data piñata," 2015, accessed November 4, 2021, https://www.urbandictionary.com/define.php?term=data%20pi%C3%B1ata

[8]Cosma Rohilla Shalizi, "Confidence Sets for Multiple Coefficients," chap. 16 in *The Truth about Linear Regression* (unpublished manuscript, 2019), 290; Frank Bretz, Torsten Hothorn, and Peter Westfall, *Multiple Comparisons Using R* (New York: Taylor & Francis, 2011).

[9]Karl Popper, *The logic of scientific discovery*, 2nd (Routledge, 2005 [1959]).

[10]Karl Popper, "Theories," chap. 3 in *The logic of scientific discovery*.

[11]Karl Popper, "Falsifiability," chap. 4 in *The logic of scientific discovery*.

[12]Everett Rogers, *Diffusion of innovations*, 5th ed. (New York: Free Press, 2003); E Katz and P E. Lazarsfeld, *Personal influence: The part played by people in the flow of mass communication* (New York: FreeP Press, 1955).

[13]B. J. Biddle, "Recent Developments in Role Theory," *Annual Review of Sociology* 12, no. 1 (1986): 67–92, doi:10.1146/annurev.so.12.080186.000435.

[14]J.M. Box-Steffensmeier et al., "Role analysis using the ego-ERGM: A look at environmental interest group coalitions," *Social Networks* 52 (2018): 213–27, doi:https://doi.org/10.1016/j.socnet.2017.08.004.

[15]Henri Tajfel and John C. Turner, "An Integrative Theory of Intergroup Conflict," chap. 3 in *The Social Psychology of Inter-group Relations*, ed. William G. Austin and Stephen Worchel (Monterey, CA: Brooks-Cole, 1979), 33–34.

[16]Shanto Iyengar and Sean J. Westwood, "Fear and Loathing across Party Lines: New Evidence on Group Polarization," *American Journal of Political Science* 59, no. 3 (2015): 690–707, doi:https://doi.org/10.1111/ajps.12152.

[17]Jane Mansbridge, "Rethinking Representation," *American Political Science Review* 97, no. 4 (2003): 515–28, doi:10.1017/S0003055403000856.

[18]Theres Matthieß, "Retrospective pledge voting: A comparative study of the electoral consequences of government parties' pledge fulfilment," *European Journal of Political Research* 59, no. 4 (2020): 774–96, doi:https://doi.org/10.1111/1475-6765.12377.

[19]Harry C. Triandis et al., "Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships," *Journal of Personality and Social Psychology* 54, no. 2 (1988): 323–28, doi:10.1037/0022-3514.54.2.323; Harry C. Triandis and Michele J. Gelfand, "A theory of individualism and collectivism.," in *Handbook of theories of social psychology, Vol. 2*, ed. P. A. M. Van Lange, A. W. Kruglanski, and Higgins E. T. (Thousand Oaks: Sage, 2012), 498–520, doi:10.4135/9781446249222.n51.

[20]Johannes C. Buggle, "Growing collectivism: irrigation, group conformity and technological divergence," *Journal of Economic Growth* 25, no. 2 (2020): 1573–7020, doi:10.1007/s10887-020-09178-3.

[21]Akin L. Mabogunje, "Systems Approach to a Theory of Rural-Urban Migration," *Geographical Analysis* 2, no. 1 (1970): 1–18, doi:https://doi.org/10.1111/j.1538-4632.1970.tb00140.x.

[22]Diego F. Leal, "Network Inequalities and International Migration in the Americas," *American Journal of Sociology* 126, no. 5 (2021): 1067–126, doi:10.1086/713877.

[23]Rosabeth M. Kanter, *Men and Women of the Corporation* (New York: Basic, 1977).

[24]Matthew Desmond and Mustafa Emirbayer, "WHAT IS RACIAL DOMINATION?," *Du Bois Review: Social Science Research on Race* 6, no. 2 (2009): 335–55, doi:10.1017/S1742058X09990166.

[25]Jennifer L. Nelson, "How Organizational Minorities Form and Use Social Ties: Evidence from Teachers in Majority-White and Majority-Black Schools," *American Journal of Sociology* 125, no. 2 (2019): 382–430, doi:10.1086/705158.

[26]Donald Black, *The Behavior of Law* (San Diego: Academic Press, 1976).

[27]Michael T. Light, "Punishing the "Others": Citizenship and State Social Control in the United States and Germany," *European Journal of Sociology* 58, no. 1 (2017): 33–71, doi:10.1017/S0003975617000029.

[28]Nagendra Singh Rawat et al., "Networked medieval strongholds in Garhwal Himalaya, India," *Antiquity* 95, no. 381 (2021): 753–72, doi:10.15184/aqy.2021.4; Davide Cantoni, "THE ECONOMIC EFFECTS OF THE PROTESTANT REFORMATION: TESTING THE WEBER HYPOTHESIS IN THE GERMAN LANDS," *Journal of the European Economic Association* 13, no. 4 (2015): 561–98, doi:https://doi.org/10.1111/jeea.12117.

[29]Heinz Schilling, ed., *Die reformierte Konfessionalisierung in Deutschland – Das Problem der "Zweiten Reformation"* (Gütersloh: Gütersloher Verlagshaus, 1986), doi:10.11588/fr.1990.2.54173; Wolfgang Reinhard, "Reformation, Counter-Reformation, and the Early Modern State a Reassessment," *The Catholic Historical Review* 75, no. 3 (1999): 383–404, doi:10.1353/cat.1999.0218.

[30]Peter Heather, *The Fall of the Roman Empire: A New History of Rome and the Barbarians*, 1st ed. (New York: Oxford university Press, 2005).

[31]Jennifer Llewellyn and Steve Thompson, "THE HISTORIOGRAPHY OF THE WEIMAR REPUBLIC," 2019, accessed November 4, 2021, https://alphahistory.com/weimarrepublic/historiography-weimar-republic/; Fritz Fischer, *Griff nach der Weltmacht: Die Kriegzielpolitik des kaiserlichen Deutschland 1914-1918* (Düsseldorf: Droste, 1961); Hans-Ulrich Wehler, "Das Deutsche Kaiserreich 1871-1918: Einleitung," in *Die Bielefelder Sozialgeschichte: Klassische Texte zu einem geschichtswissenschaftlichen Programm und seinen Kon-*

troversen, ed. Bettina Hitzer and Thomas Welskopp (transcript Verlag, 2015), 255–62, doi:doi:10.14361/9783839415214-011; Edmond Vermeil, *L'Allemagne contemporaine: Sociale, politique, et culturelle, 1890-1950* (Paris: Aubier, 1952); Alan J. P. Taylor, *The Course of German History* (London: Hamish Hamilton, 1945); William L. Shirer, *Rise And Fall Of The Third Reic: A History of Nazi Germany* (New York: Simon & Schuster, 1960).

[32]An engraving example to distinguish type I from type II error is to determine whether a person is pregnant. The null hypothesis states that the tested person is not pregnant. If we test a man and claim that he is pregnant, we have falsely rejected the null hypothesis, i.e., conducted a type I error. If we test a pregnant woman and claim that she is not pregnant, we fail to reject the null hypothesis, i.e., conduct a type II error.

[33]Amitav Banerjee et al., "Hypothesis testing, type I and type II errors," *Industrial Psychiatry Journal* 18, no. 2 (2009): 127–31, doi:10.4103/0972-6748.62274.

[34]Thomas H. Ryan, "Significance tests for multiple comparison of proportions, variances, and other statistics," *Psychological Bulletin* 57, no. 4 (1960): 318–28, doi:10.1037/h0044320.

[35]Carlo Emilio Bonferroni, "Teoria Statistica Delle Classi e Calcolo Delle Probabilità," in *Encyclopedia of Research Design*, ed. Neil J. Salkind (Thousand Oaks: Sage, 2010), doi:10.4135/9781412961288.n455; Yoav Benjamini and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)* 57, no. 1 (1995): 289–300.

[36]Schilling, *Die reformierte Konfessionalisierung in Deutschland – Das Problem der "Zweiten Reformation"*; Reinhard, "Reformation, Counter-Reformation, and the Early Modern State a Reassessment."

[37]Robert Walinski-Kiehl, "Reformation History and Political Mythology in the German Democratic Republic, 1949-89," *European History Quarterly* 34, no. 1 (2004): 43–67, doi:10.1177/0265691404040008.

[38]Max Weber, *The Protestant Ethic and the Spirit of Capitalism*, trans. Talcott Parsons (London: Routledge, 1930).

[39]Sascha O. Becker and Ludger Woessmann, "Was Weber Wrong? A Human Capital Theory of Protestant Economic History*," *The Quarterly Journal of Economics* 124, no. 2 (2009): 531–96, doi:10.1162/qjec.2009.124.2.531; Cantoni, "THE ECONOMIC EFFECTS OF THE PROTESTANT REFORMATION: TESTING THE WEBER HYPOTHESIS IN THE GERMAN LANDS"; Gharad Bryan, James J Choi, and Dean Karlan, "Randomizing Religion: the Impact of Protestant Evangelism on Economic Outcomes*," *The Quarterly Journal of Economics* 136, no. 1 (2020): 293–380, doi:10.1093/qje/qjaa023.

[40]"Religious History beyond Confessionalization," *German History* 32, no. 4 (2014): 579–98, doi:10.1093/gerhis/ghu104; Sascha O. Becker, Steven Pfaff, and Jared Rubin, "Causes and consequences of the Protestant Reformation," *Explorations in Economic History* 62 (2016): 1–25, doi:https://doi.org/10.1016/j.eeh.2016.07.007.

[41]Heinz Schilling, "Konfessionskonflikt und Staatsbildung: Eine Fallstudie über das Verhältnis von religiösem und sozialem Wandel in der Frühneuzeit am Beispiel der Grafschaft Lippe," in *Quellen und Forschungen zur Reformationsgeschichte*, ed. Gerd Mohn (Gütersloh: Gütersloher Verlagshaus, 1981), ch. 2.

[42]Becker, Pfaff, and Rubin, "Causes and consequences of the Protestant Reformation."

[43]Davide Cantoni, "ADOPTING A NEW RELIGION: THE CASE OF PROTESTANTISM IN 16TH CENTURY GERMANY," *The Economic Journal* 122, no. 560 (2012): 502–31; Hyojoung Kim and Steven Pfaff, "Structure and Dynamics of Religious Insurgency: Students and the Spread of the Reformation," *American Sociological Review* 77, no. 2 (2012): 188–215, doi:10.1177/0003122411435905; Sascha O. Becker et al., "Multiplex Network Ties and the Spatial Diffusion of Radical Innovations: Martin Luther's Leadership in the Early Reformation," *American Sociological Review* 85, no. 5 (2020): 857–94, doi:10.1177/0003122420948059.

[44]Rogers, *Diffusion of innovations*; Katz and Lazarsfeld, *Personal influence: The part played by people in the flow of mass communication*.

[45]Thomas W. Valente and Patchareeya Pumpuang, "Identifying Opinion Leaders to Promote Behavior Change," PMID: 17602096, *Health Education & Behavior* 34, no. 6 (2007): 881–96, doi:10.1177/1090198106297855; Seyed Mojtaba Hosseini Bamakan, Ildar Nurgaliev, and Qiang Qu, "Opinion leader detection: A methodological review," *Expert Systems with Applications* 115 (2019): 200–22, doi:https://doi.org/10.1016/j.eswa.2018.07.069.

[46]Martin Luther: ProQuest-LLC, "Luthers Werke on the World Wide Web," 2015, accessed May 17, 2019, http://luther.chadwyck.co.uk/, Philipp Melanchthon: Christine Mundhenk, "Melanchthons Briefwechsel – Regesten online," 2019, accessed May 17, 2019, https://www.haw.uni-heidelberg.de/forschung/forschungsstellen/melanchthon/mbw-online.de.html, Martin BucerReinhold Friedrich, "Bucer Briefkorrespondenz," 2018, accessed May 17, 2019, https://www.theologie.fau.de/lehrstuhl-kirchengeschichte-ii-neuere-kirchengeschichte/bucer-forschungsstelle/, Huldrych Zwingli: Christian Moser, "Huldreich Zwinglis sämtliche Werke," 2016, accessed May 17, 2019, http://www.irg.uzh.ch/static/zwingli-briefe/?n=Main.Overview, Heinrich Bullinger: Reinhard Bodenmann, "Heinrich Bullinger's Correspondence," 2016, accessed May 17, 2019, http://www.arpa-docs.ch/SedServer/SedWEB.cgi?fld_41a=&fld_30b=&fld_41c=&fld_30c=&fld_41e=&search=&range=&Alias=Briefe&Lng=0&first=0&session=0&awidth=1440&aheight=769&PrjName=Bullinger+-+Briefwechsel, Andreas Karlstadt: Thomas Kaufmann, "Kritische Gesamtausgabe der Schriften und Briefe Andreas Bodensteins von Karlstadt, Teil I (1507–1518)," 2012, accessed May 17, 2019, http://diglib.hab.de/edoc/ed000216/start.htm, Myconius Oswald: Martin Wallraff, "Erschließung des Briefwechsels von Oswald Myconius," 2016, accessed May 17, 2019, https://myconius.unibas.ch/briefdb.html, Joachim Vadian: Amy Burnett, *Letter correspondence of Oekolampad and Vadian*, Personal communication, 2019, Johann Oekolampad: Burnett, *Letter correspondence of Oekolampad and Vadian*. The letters were crawled from public databases.

[47]Anton Schindling and Walter Ziegler, eds., *Die Territorien des Reichs im Zeitalter der Reformation und Konfessionalisierung: Land und Konfession 1500-1650, Bände 1-5 (Südosten, Nordosten, Nordwesten, Mittleres Deutschland, Südwesten)* (Münster: Aschendorff, 1989–1995).

[48]Pierre-François Verhulst, "Notice sur la loi que la population suit dans son accroissement," *Correspondance Mathématique et Physique* 10 (1838): 113–21; Jan S. Cramer, *The Origins of Logistic Regression*, Tinbergen Institute Working Paper No. 2002-119/4, 2002, doi:10.2139/ssrn.360300; Richard M. Tolman and Arlene Weisz, "Coordinated Community Intervention for Domestic Violence: The Effects of Arrest and Prosecution on Recidivism of Woman Abuse Perpetrators," *Crime and Delinquency* 41, no. 4 (1995): 481–95, doi:10.1177/0011128795041004007; Hwei-Lin Chuang, "High school youths' dropout and re-enrollment behavior," *Economics of Education Review* 16, no. 2 (1997): 171–86, doi:https://doi.org/10.1016/S0272-7757(96)00058-1; James Janik and Howard M. Kravitz, "Linking work and domestic problems with police suicide," *Suicide Life-Threatening Behavior* 24, no.

3 (1994): 267–74; Sanaz Mobasseri, "Race, Place, and Crime: How Violent Crime Events Affect Employment Discrimination," *American Journal of Sociology* 125, no. 1 (2019): 63–104, doi:10.1086/703883; Michael A. Lewis and Kristin M. Ferguson, "Predicting Methamphetamine Use of Homeless Youths Attending High School: Comparison of Decision Rules and Logistic Regression Classification Algorithms," *Journal of the Society for Social Work and Research* 5, no. 2 (2014): 211–31, doi:10.1086/676830; Richard L. Fox and Jennifer L. Lawless, "To Run or Not to Run for Office: Explaining Nascent Political Ambition," *American Journal of Political Science* 49, no. 3 (2005): 642–59, doi:https://doi.org/10.1111/j.1540-5907.2005.00147.x; Romain Ferrali et al., "It Takes a Village: Peer Effects and Externalities in Technology Adoption," *American Journal of Political Science* 64, no. 3 (2020): 536–53, doi:https://doi.org/10.1111/ajps.12471; Jr. Fryer Roland G. and Steven D. Levitt, "Hatred and Profits: Under the Hood of the Ku Klux Klan*," *The Quarterly Journal of Economics* 127, no. 4 (2012): 1883–925, doi:10.1093/qje/qjs028; Ufuk Akcigit et al., "Taxation and Innovation in the Twentieth Century*," qjab022, *The Quarterly Journal of Economics*, June 2021, doi:10.1093/qje/qjab022; Amitabh Chandra and Douglas O Staiger, "Identifying Sources of Inefficiency in Healthcare*," *The Quarterly Journal of Economics* 135, no. 2 (2020): 785–843, doi:10.1093/qje/qjz040; Lawrence F. Katz and Alan B. Krueger, "The Role of Unemployment in the Rise in Alternative Work Arrangements," *American Economic Review* 107, no. 5 (2017): 388–92, doi:10.1257/aer.p20171092; Maximilian Filsinger, Kathrin Ackermann, and Markus Freitag, "Surfing to help? An empirical analysis of Internet use and volunteering in 27 European societies," *European Societies* 22, no. 3 (2020): 368–89, doi:10.1080/14616696.2019.1663895; Catherine E. Harnois, "Race, Ethnicity, Sexuality, and Women's Political Consciousness of Gender," *Social Psychology Quarterly* 78, no. 4 (2015): 365–86; J. Ron Nelson et al., "Which Risk Factors Predict the Basic Reading Skills of Children at Risk for Emotional and Behavioral Disorders?," *Behavioral Disorders* 33, no. 2 (2008): 75–86.

[49]Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricists's Companion* (Princeton: Princeton, 2009), p.64.

[50]Specifically, through linearisation established estimators for parameter estimation can be used, such as ordinary least squares and maximum likelihood estimation.

[51]Cosma Rohilla Shalizi, "F-Tests, R^2, and Other Distractions," chap. 10 in *The Truth about Linear Regression* (unpublished manuscript, 2019), 165–71; Debbi L. Hahs-Vaughn and Richard G. Lomax, "One-Factor Analysis of Variance - Fixed Effects Models," chap. 11 in *An Introduction of statistical Concepts*, 4th (New York: Routledge, 2020), 427, doi:10.4324/9781315624358.

[52]Each standard error indicates how different the real adoption of Protestantism is, on average, from the adoption which was predicted by the corresponding β. The smaller the standard error, the better the corresponding variable predicts the adoption of Protestantism.

[53]A closer analysis could reveal whether this effect is driven by some famous reformers, such as Martin Luther, or whether all reformers contributed to this effect to a similar extent. This could be done with marginal effects.

[54]Using odds ratio to compute probability. The actual computation of the probability is more complicated than presented here, because the other independent variables of the model need to be fixed at a certain value, which is ignored in the calculation below. The calculation below is correct, if only one independent variable was tested in the model. To compute the correct probabilities, 'marginal effects' have to be calculated, which account for the remain-

ing independent variables in the model. Marginal effects are implemented in all major statistics software.

$$\frac{p}{1-p} = e^z \tag{2}$$

$$p = e^z - pe^z \tag{3}$$

$$p(1+e^z) = e^z \tag{4}$$

$$p = \frac{e^z}{(1+e^z)} \tag{5}$$

$$\text{where } z = \beta_0 + \beta_1 * x_1 \tag{6}$$

[55]Shalizi, "F-Tests, R^2, and Other Distractions."

[56]Heinold Fast, *Der linke Flügel der Reformation: Glaubenszeugnisse der Täufer, Spiritualisten, Schwärmer und Antitrinitarier* (Bremen: Carl Schünemann Verlag, 1962).

[57]Bruce Gordon, "Geisteswissenschaftliches Zentrum Geschichte und Kultur Ostmitteleuropas," *German history* 17, no. 1 (1997): 90–94, doi:10.1093/026635599674233424.

[58]Angrist and Pischke, *Mostly Harmless Econometrics*, p.59–64.

[59]Ordinary least squares estimator

[60]That is, the amount of unexplained variance in the data (error) does not systematically differ between low and high values of an explanatory variable (e.g., low and high socio-economic status of inhabitants).

[61]Jeremiah E Dittmar and Ralf R Meisenzahl, "Public Goods Institutions, Human Capital, and Growth: Evidence from German History," *The Review of Economic Studies* 87, no. 2 (2019): 959–96, doi:10.1093/restud/rdz002; Davide Cantoni, Jeremiah Dittmar, and Noam Yuchtman, "Religious Competition and Reallocation: the Political Economy of Secularization in the Protestant Reformation*," *The Quarterly Journal of Economics* 133, no. 4 (2018): 2037–96, doi:10.1093/qje/qjy011; Sascha O. Becker and Luigi Pascali, "Religion, Division of Labor, and Conflict: Anti-semitism in Germany over 600 Years," *American Economic Review* 109, no. 5 (2019): 1764–804, doi:10.1257/aer.20170279.

[62]David Michael Green, Chad Kahl, and Paul F. Diehl, "The Price of Peace: A Predictive Model of UN Peacekeeping Fiscal Costs," *Policy Studies Journal* 26, no. 4 (1998): 620–35, doi:https://doi.org/10.1111/j.1541-0072.1998.tb01936.x; Jonathan Golub, "In the Shadow of the Vote? Decision Making in the European Community," *International Organization* 53, no. 4 (1999): 733–64; Janet Box-Steffensmeier and Bradford S. Jones, *Event History Modeling: A Guide for Social Scientists* (New York: Cambridge University Press, 2004).

[63]James J. Heckman, Hidehiko Ichimura, and Petra Todd, "Matching As An Econometric Evaluation Estimator," *The Review of Economic Studies* 65, no. 2 (1998): 261–94, doi:10.1111/1467-937X.00044; Sebastian Galiani, Paul Gertler, and Ernesto Schargrodsky, "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy* 113, no. 1 (2005): 83–120; Victor Lavy, "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy* 110, no. 6 (2002): 1286–317.

[64]Cantoni, "ADOPTING A NEW RELIGION: THE CASE OF PROTESTANTISM IN 16TH CENTURY GERMANY"; Joshua D. Angrist and Alan B. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics* 106, no. 4 (1991): 979–1014; Daron Acemoglu and Joshua Angrist, *How Large are the Social Returns to Education? Evidence from Compulsory Schooling Laws*, Working Paper, Working Paper Series 7444 (National Bureau of Economic Research, 1999), doi:10.3386/w7444.

[65]D. A. Freedman, "Linear Statistical Models for Causation: A Critical Review," in *Encyclopedia of Statistics in Behavioral Science* (American Cancer Society, 2005), doi:https://doi.org/10.1002/0470013192.bsa598; Austin Nichols, "Causal Inference with Observational Data," *The Stata Journal* 7, no. 4 (2007): 507–41, doi:10.1177/1536867X0800700403; Markus Gangl, "Causal Inference in Sociological Research," *Annual Review of Sociology* 36, no. 1 (2010): 21–47, doi:10.1146/annurev.soc.012809.102702.

[66]Sandy Schumann et al., "Towards Open and Reproducible Terrorism Studies: Current Trends and Next Steps," *Perspectives on Terrorism* 13, no. 5 (2019): 61–73; Robert M. Kaplan and Veronica L. Irvin, "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time," *PLoS ONE* 10, no. 8 (2015), doi:10.1371/journal.pone.0132382; Jakub Prochazka, Yulia Fedoseeva, and Petr Houdek, "A field experiment on dishonesty: A registered replication of Azar et al. (2013)," *Journal of Behavioral and Experimental Economics* 90 (2021): 101617, doi:https://doi.org/10.1016/j.socec.2020.101617.

[67]Cantoni, "ADOPTING A NEW RELIGION: THE CASE OF PROTESTANTISM IN 16TH CENTURY GERMANY"; Box-Steffensmeier et al., "Role analysis using the ego-ERGM: A look at environmental interest group coalitions"; Soumyajit Mazumder, "The Persistent Effect of U.S. Civil Rights Protests on Political Attitudes," *American Journal of Political Science* 62, no. 4 (2018): 922–35, doi:https://doi.org/10.1111/ajps.12384.

[68]Brian Duignan, "Occam's razor," 2021, accessed November 4, 2021, https://www.britannica.com/topic/Occams-razor; Hugh G. Gauch Jr., *Scientific Method in Practice* (Cambridge: Cambridge University Press, 2002), doi:10.1017/CBO9780511815034.

[69]J Scott Armstrong, "How to avoid exploratory research," *Journal of Advertising Research* 10, no. 4 (1970): 27–30.