PLOS ONE

RESEARCH ARTICLE

Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity

Berna Devezer[1,5☯]*, **Luis G. Nardin**[2,5☯], **Bert Baumgaertner**[3,5☯], **Erkan Ozge Buzbas**[4,5☯]

1 Dept. of Business, University of Idaho, Moscow, ID, United States of America, **2** Dept. of Informatics, Brandenburg University of Technology, Cottbus, Germany, **3** Dept. of Politics and Philosophy, University of Idaho, Moscow, ID, United States of America, **4** Dept. of Statistical Science, University of Idaho, Moscow, ID, United States of America, **5** Center for Modeling Complex Interactions, University of Idaho, Moscow, ID, United States of America

☯ These authors contributed equally to this work.
* bdevezer@uidaho.edu

Citation: Devezer B, Nardin LG, Baumgaertner B, Buzbas EO (2019) Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PLoS ONE 14(5): e0216125. https://doi.org/10.1371/journal.pone.0216125

Editor: Daniele Fanelli, London School of Economics and Political Science, UNITED KINGDOM

Received: July 2, 2018

Accepted: April 16, 2019

Published: May 15, 2019

Data Availability Statement: The code to perform the simulations and analyze the data generated in this project, and a summary data set are included as a Git repository at https://github.com/gnardin/CRUST. Refer to the description in the main page of this repository for further instructions and details of the implementation.

Funding: Research reported in this publication was supported by the University of Idaho Vandal Ideas Project grant and the National Institute of General Medical Sciences of the National Institutes of

Abstract

Consistent confirmations obtained independently of each other lend credibility to a scientific result. We refer to results satisfying this consistency as reproducible and assume that reproducibility is a desirable property of scientific discovery. Yet seemingly science also progresses despite irreproducible results, indicating that the relationship between reproducibility and other desirable properties of scientific discovery is not well understood. These properties include early discovery of truth, persistence on truth once it is discovered, and time spent on truth in a long-term scientific inquiry. We build a mathematical model of scientific discovery that presents a viable framework to study its desirable properties including reproducibility. In this framework, we assume that scientists adopt a model-centric approach to discover the true model generating data in a stochastic process of scientific discovery. We analyze the properties of this process using Markov chain theory, Monte Carlo methods, and agent-based modeling. We show that the scientific process may not converge to truth even if scientific results are reproducible and that irreproducible results do not necessarily imply untrue results. The proportion of different research strategies represented in the scientific population, scientists' choice of methodology, the complexity of truth, and the strength of signal contribute to this counter-intuitive finding. Important insights include that innovative research speeds up the discovery of scientific truth by facilitating the exploration of model space and epistemic diversity optimizes across desirable properties of scientific discovery.

Introduction

Consistent confirmations obtained independently of each other lend credibility to a scientific result [1–4]. We refer to this notion of multiple confirmations as *reproducibility of scientific results*. Ioannidis [5] argued that a research claim is more likely to be false than true, partly due

Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research made use of the resources of the High Performance Computing Center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the US DoE under Contract No. DE-AC07-05ID14517. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

to the prevalent use of statistical significance and null hypothesis significance testing as method of inference. Recent theoretical research explored aspects of scientific practice contributing to irreproducibility. McElreath and Smaldino [6] modeled a population of scientists testing a variety of hypotheses and tracking positive and negative published findings to investigate how the evidential value of replication studies changed with the base rate of true hypotheses, statistical power, and false positive rate. Other studies found that current incentive structures may lead to degradation of scientific practice [7, 8]. Publication bias was also proposed to contribute to the transitioning of incorrect findings from claim to fact [9]. These studies focus on how structural incentives and questionable research practices (QRPs) influence reproducibility of scientific results within a hypothesis-centric framework, and how to improve statistical practices and publication norms to increase reproducibility. Under limitations of hypothesis testing [10], however, understanding salient properties of the scientific process is challenging, especially for fields that progress by building, comparing, selecting, and re-building models.

In this work, we make three contributions to the literature on meta-research. First, we present a model-centric mathematical framework modeling scientists' convergence to truth in the process of scientific discovery. We identify, mathematically define and study the relationship between key properties of this process such as early discovery of truth, persistence on truth once it is discovered, time spent on truth in a long-term scientific inquiry, and rate of reproducibility. Second, in a system without QRPs or structural incentives, we study how the diversity of research strategies in the scientific population, the complexity of truth, and the noise-to-signal ratio in the true data generating model affect these properties. Third, we study the scientific process where scientists engage in model comparison instead of statistical hypothesis testing. Model comparison aims to select a useful model that approximates the true model generating the data and it has long been a cornerstone in many scientific disciplines because of its generality. Our model-centric view allows us to study the process of scientific discovery under uncertainty, bypassing the complications inherited from hypothesis testing [10].

A model-centric meta-scientific framework

We adopt a notion of confirmation of results in idealized experiments and build a mathematical framework of scientific discovery based on this notion.

Model, idealized experiment, replication experiment, and reproducibility

We let K be the background knowledge on a natural phenomenon of interest, M be a prediction in the form of a probability model parameterized by $\theta \in \Theta$, that is in principle testable using observables, and D be the data generated by the true model. The degree of confirmation of M by D is assessed by S, a fixed and known method. We define ξ, an *idealized experiment*, as (M, θ, D, S, K).

In an idealized experiment ξ, the data D confirms the model M if $\mathbb{P}(M|D, K) > \mathbb{P}(M|K)$, where $\mathbb{P}(M|D, K)$ and $\mathbb{P}(M|K)$ are probabilities of M after and before observing the data, respectively. By Bayes's Theorem, $\mathbb{P}(M|D, K)/\mathbb{P}(M|K)$ is proportional to the likelihood $\mathbb{P}(D|M, K)$. Large $\mathbb{P}(D|M, K)$ implies high degree of confirmation of M. Complex models, however, have a tendency to inflate $\mathbb{P}(D|M, K)$ and hence $\mathbb{P}(M|D, K)$. As a measure against overfitting, modern model comparison statistics S are not only based on $P(D|M, K)$ but also penalize the complexity of M to prevent inflating the likelihood under complex models. For several well-known S, smaller $S(M)$ means the model M is more favorable in a set of models, and we follow this convention here.

In a scientific inquiry, a novel prediction is often tested against a status quo consensus. We formulate this situation by denoting the novel prediction as a *proposed model* M_P which is

tested against the *global model* M_G, the scientific consensus. Conditional on the data, $S(M_P) < S(M_G)$ means that the proposed model is more favorable than M_G. In this case, M_P becomes the new scientific consensus, otherwise the global model remains as the scientific consensus.

The description of scientific inquiry given in the last paragraph and *reproducibility of results* in a replication experiment as follows. If ξ_1 given by $(M_P, \theta, D_1, S, K_1)$ is tested against M_G, then the experiment ξ_2 immediately following ξ_1 is a replication experiment for ξ_1 if and only if ξ_2 is given by $(M_P, \theta, D_2, S, K_2)$ and it is tested against the same M_G as ξ_1. That is, the replication experiment proposes the same model, uses the same methods, and is tested against the same global model as the original experiment. Of two elements that differ between the original experiment and the replication experiment, the first is D_2, which is the data that is generated in the replication experiment independent of the data D_1 of the original experiment. The second is the background information K_2 which includes all the information necessary to replicate ξ_1. In particular, K_2 includes the knowledge that M_P was the proposed model in ξ_1, it was tested against M_G, and the outcome of this test—whether M_P was updated to consensus or M_G remained as the consensus. We say that the replication experiment ξ_2 *reproduces the results* of ξ_1 if the results of ξ_1 and ξ_2 are the same in terms of updating the consensus. There are two mutually exclusive ways that ξ_2 can reproduce the results of ξ_1: 1) If the proposed model in ξ_1 won against the global model, then this must also be the case in ξ_2, that is $S(M_P) < S(M_G)$ in both experiments. 2) If the proposed model in ξ_1 lost against the global model, then this must also be the case in ξ_2, that is $S(M_P) > S(M_G)$ in both experiments. Otherwise, we say that ξ_2 fails to reproduce the results of ξ_1. These definitions of replication experiment and reproducibility of results formalize necessary open science practices for potential reproducibility of results: Information about the proposed and global models in the original experiment and the results of this experiment which we capture by K_2 must be transferred to ξ_2.

Stochastic process of scientific discovery

We assume an infinite population of scientists who conduct a sequence of idealized experiments to find the true model generating the data (see S1 File for mathematical framework). In the population, we consider various types of scientists, each with a mathematically well-defined research strategy for proposing models. Scientists search for a true model in a set of linear models. Linear models were chosen because they can accommodate a variety of designs with straightforward statistical analysis, and their complexity is mathematically tractable (S2 File). We define model complexity as a function of the number of model parameters and interaction terms, and visualize it by representing each model with a unique geometry on an equilateral hexagon inscribed in its tangent circle (Fig 1A).

We assume a discrete time process with $t = 0, 1, \cdots$, where at each time step an experiment $\xi^{(t)}$ is conducted by a scientist randomly selected from a population of scientists with equal probability. The experiment entails proposing a model $M_P^{(t)}$ as a candidate for the true data generating mechanism. The probability of proposing a particular model is determined by the scientist's research strategy and the global model $M_G^{(t)}$—the current scientific consensus. The scientist compares the global model against the proposed model using new data $D^{(t)}$ generated from the true model and a model comparison statistic S. The model with favorable statistic is set as the global model for the next time step $M_G^{(t+1)}$. Because the probability of proposing a model is independent of the past and the transition from $M_G^{(t)}$ to $M_G^{(t+1)}$ admits the (first order) Markov property, the stochastic process representing the scientific process is a Markov chain. This mechanism represents how scientific consensus is updated in light of new evidence. We study the mathematical properties of this process for different scientist populations representing a variety of research strategies.

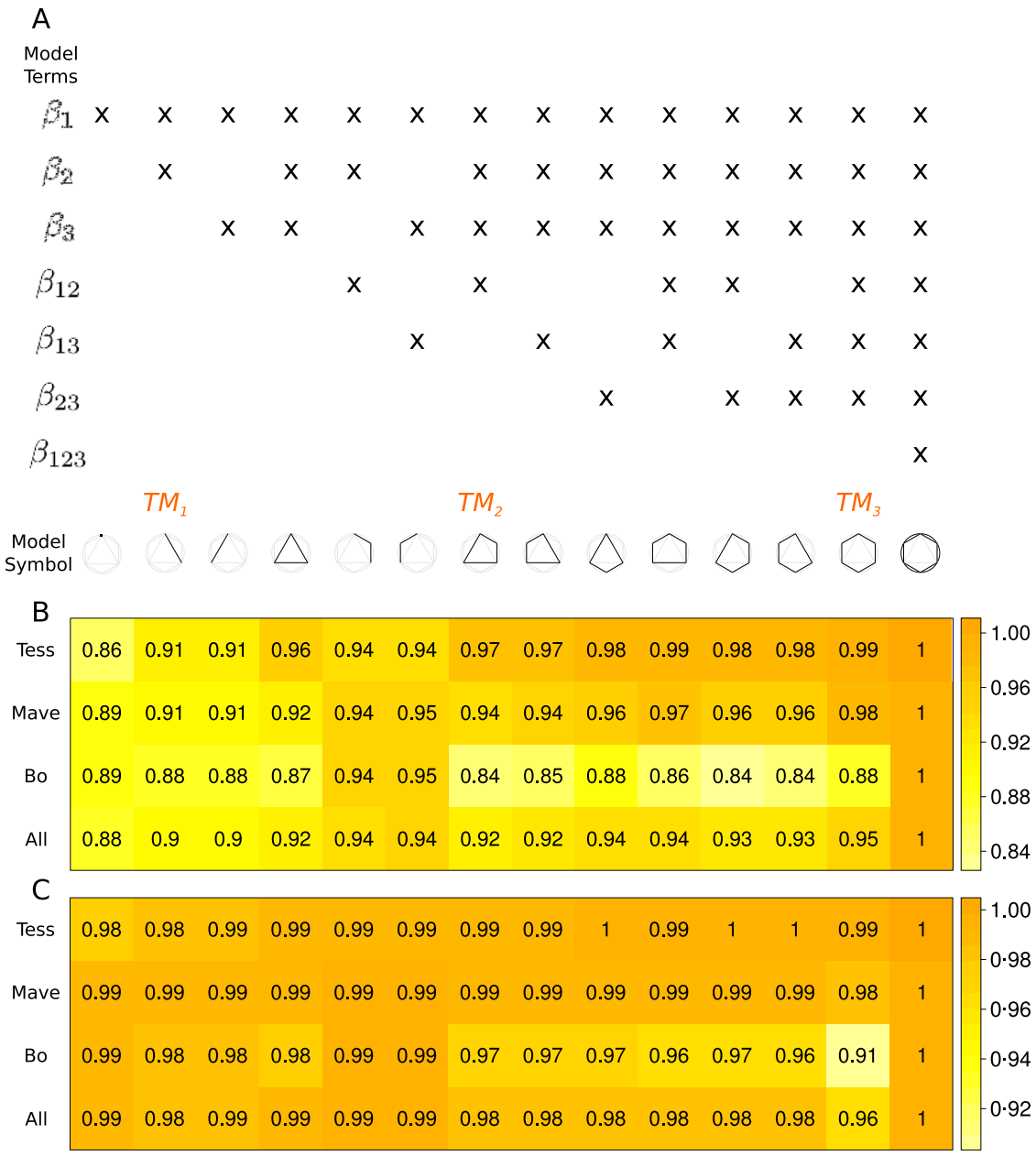


Fig 1. (A) Each column of the matrix indicates the terms included in the model shown by a symbol at the bottom of the column. For example, the fifth column denotes the model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$, and is represented by three corners of hexagon connected with two lines. TM_1, TM_2, TM_3 are the three true models used in our agent-based model simulations. Symbols representing each model are ordered from simple to complex, left to right. Symbols are used as y-axis labels for heat maps in (B) and (C). Stickiness of each true model as a global model for each scientist population under AIC (B) and SC (C).

https://doi.org/10.1371/journal.pone.0216125.g001

Introducing replication experiments to the process fundamentally alters the probability mechanism of updating global models: By definition, a replication experiment depends on the experiment conducted at the previous time step via K_2. Hence the stochastic process is a higher order Markov chain (see [11]) and we turn to an agent-based model (ABM) [12, 13] to analyze the process with replication. Our ABM is a forward-in-time, simulation-based, individual-level implementation of the scientific process where *agents* represent scientists (S1 Algorithm).

We assume that reproducibility is a desirable property of scientific discovery. However, arguably early discovery of truth, persistence on truth once it is discovered, and long time spent on truth in a long-term scientific inquiry are also desirable properties of scientific discovery since they would characterize a resource efficient and epistemically accurate scientific process. We seek insight into the drivers of these properties and the relationship among them, which we motivate with the following questions (see S3 File for mathematical definitions). *How quickly does scientific community discover the true model?* We assess this property by the mean first passage time to the true model and view it as a key indicator of resource efficiency in the process of scientific discovery. *How "sticky" is the true model as global model?* We define the stickiness of the true model as the mean probability of staying in the true model once it becomes global model. *How long does scientific community stay on the true model?* The stationary probability that the true model is global model has the interpretation of the long-term stay of the scientific community on the true model. *How reproducible are the results of experiments?* We track replication experiments to calculate the rate of reproducibility when the true model is global model, as well as when it is not. We study the answers to these questions as a function of the following aspects of model-centric approach to scientific discovery: the proportion of research strategies in the scientific population, the complexity of the true model generating the data, the ratio of error variance to deterministic expectation in the true model (i.e., noise-to-signal ratio), and the model comparison statistic.

Scientist types

We include distinct research strategies to explore the effect of epistemic diversity in the scientific population. We define simple research strategies where our scientists do not have a memory of their past decisions and they do not interact directly with each other, but only via the global model. Nonetheless, the research strategies we include in our model seem reasonably realistic to us in capturing the essence of some well-known research approaches. Fig 2 illustrates our stochastic process of scientific discovery for a specific scientist population.

For the process with no replication, we define three types of scientists: *Tess, Mave, Bo* (S4 File). *Tess, the theory tester*, uniformly randomly selects a proposal model that is only one main effect predictor away from the current global model. We impose a hierarchical constraint

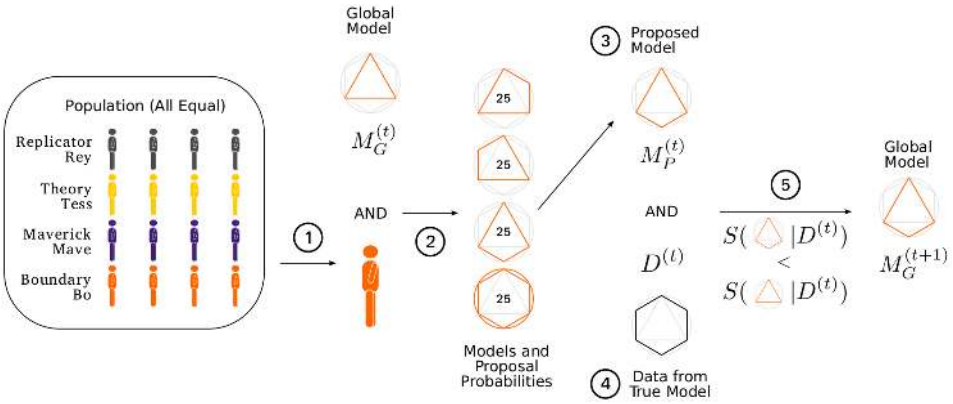


Fig 2. A transition of our process of scientific discovery for an epistemically diverse population with replicator. A scientist (Bo) is chosen uniformly randomly from the population (1). Given the global model, the set of proposal models and their probabilities (given in percentage points inside models) are determined. In this population with no replicator, Bo proposes only models formed by adding an interaction (2). The proposed model selected (3) and the data generated from the true model (4) are used with the model comparison statistic (SC or AIC) to update the global model (5).

https://doi.org/10.1371/journal.pone.0216125.g002

on interaction terms in the sense that when *Tess* proposes to drop a predictor from the current global model, all higher order interactions including this term are dropped too. We think of *Tess*' strategy as a refinement of an existing theory by testing current consensus against models that are close to it. *Mave, the maverick* does not build off of the current consensus but she ignores it. She advances novel ideas and uniformly randomly selects a model from the set of all models. The novelty-seeking aspect of her strategy is similar to a maverick strategy proposed in prior research on epistemic landscapes and the division of cognitive labor [14–16]. However, in contrast to that strategy, *Mave* does not actively aim to avoid previously tested models and she acts independently of the current scientific consensus. *Bo, the boundary tester* systematically tests the boundaries of the current global model. She selects a model that adds interactions to the current global model to explore the conditions under which the current global model holds. After her, if the tested boundary has been confirmed by the data, the global model may earn a new predictor representing a main interaction and the lower order interactions associated with it that are not currently in the model. If the tested boundary has not been confirmed by the data, the global model does not change.

For the process with replication, we introduce *Rey, the replicator* who conducts a replication experiment which is the exact same experiment conducted by her precedent, but with new data. She reproduces the results of the preceding experiment if she confirms as global model the same model confirmed as global model by her precedent. Thus, *Rey* compares the same pair of proposed model (M_P) and global model (M_G) as that of her predecessor, using new data, and the replication is successful if two conditions are satisfied: 1) either the results of her predecessor and the replication experiment are both judged favorable against the global model (M_G), or they both are not, leaving scientific consensus unchanged, 2) sufficient information about the the result of her predecessor's experiment is available through the background knowledge of the replication experiment to assess if the first condition holds. This second condition implies that a replication experiment requires open science practices for transferring sufficient information about the experiment in the previous time step to the replication experiment in order for the latter to assess the reproducibility of the result.

Scientist populations

We assess the effect of each strategy by considering populations of scientists in which *Rey*, *Tess*, *Bo*, and *Mave* are represented at varying proportions (S1 Table). Of particular interest to us are homogeneous populations where the dominant scientist type comprises 99% of the scientist population and epistemically diverse populations where all scientist types are represented in equal proportion.

Accumulation of evidence

The background knowledge that a scientist brings to an experiment consist of the global model, all other models in the system, predictors, and their parameters as well as the results from the previous experiment if the current experiment is a replication experiment. Scientific evidence in our model accumulates through experiments and all the evidence is counted at the end of a long run. Data set in each experiment has a weight of one and the sample proportion of experiments which reproduce a result converges to the true value of reproducibility rate by the Law of Large Numbers.

Model comparison criteria

We adopt two well-known likelihood-based criteria for model comparison and show how these interact with the behavior of scientists in the population: the Schwarz Criterion (SC) [17]

and the Akaike's Information Criterion (AIC) [18, 19]. A smaller value of these statistics indicate a better model performance than a larger value. When the true model generating the data is in the universe of candidate models, SC is statistically consistent and selects the true model with probability 1 as $n \to \infty$ (S4 File).

Maximum number of factors in the model

For computational feasibility, we fix the number of factors in the linear model to 3, which results in 14 models. Each of these 14 models refers to its linear structure. In this sense, there are infinitely many probability distributions that can be fully specified within a model. For the system without replication, we analyze all 14 models as true models. The most complex model has 7 predictors (Fig 1A), including three main effects, three 2–way interactions, and one 3–way interaction. We fix the sample size to 100 and calibrate the ratio of the error variance σ^2 to expected value of the model at the mean value of the predictors $\mathbb{E}(y|\mu_x)$ where $\mu_x = \mathbb{E}(x)$. We fix $\sigma^2 : \mathbb{E}(y|\mu_x)$ to (1: 4) (and include results for (1: 1), (4: 1) in S1–S10 Figs; S4 File).

Design of ABM experiments

For the system with replication, we use three true models representing a gradient of complexity (Fig 1A TM_1, TM_2, TM_3). We set up a completely randomized factorial simulation experiment: 3 true models, 3 $\sigma^2 : \mathbb{E}(y|\mu_x)$ levels at (1: 4), (1: 1), (4: 1), 5 scientist populations (S1 Table), and 2 model comparison statistics (AIC, SC). We run each experimental condition as an ABM simulation for 11000 iterations and replicated 100 times, each using a different random seed. We discard the first 1000 iterations as burn-in, except when analyzing the mean first passage time to true model. Code and data are given in S1 Code and Data.

A brief discussion of our modeling choices, assumptions, and their implications

Our model-centric framework facilitates investigating the consequences of the process of scientific discovery, including reproducibility, as a system-wide phenomenon. System-wide reproducibility and its relationship to scientific discovery are largely unexplored topics. Navigating through numerous potential variables and parameters to create a realistic system rich in behavior whose outcomes are easily interpretable is challenging. Our model aims to create such a system by making design choices and simplifying assumptions. Among many results that we obtain, we report some intuitive results as reasonableness checks. These results connect our idealized system to reality. However, we highlight the results that seem counter-intuitive to us because we find them to be interesting patterns warranting further investigation. The implications and limitations of each specific result are discussed in the Results section.

Here, we qualitatively clarify the implications and limitations of our system and emphasize the assumptions which constitute its salient features for our results to hold. We anchor our system firmly against the backdrop of guarantees provided by statistical theory to avoid overgeneralization.

What statistical theory offers in isolation

A well-known statistical inference mode is comparing a set of hypotheses represented as probabilistic models, with the goal of selecting a model. A statistical method selects the model which fits the stochastic observations best according to some well-defined measure. Consider the following three conditions:

1. There exists a true model generating the observations and it is in the search set.

2. The signal in the observations is detectable.
3. A reliable method whose assumptions are met is used to perform inferences.

If these conditions are met, then the statistical theory guarantees that under repeated testing with independent observations, the true model is selected with highest frequency. This frequency approaches to a constant value determined by conditions (1), (2), and (3). The practical implication of this guarantee is that the results under the true model are reproducible with a constant rate.

We now contemplate on the consequences of violating conditions (1)-(3). If condition (1) is not met, then the true model cannot be selected. In this case, well-established methods select the model that is closest to the true model and in the set with highest frequency. As we discuss in research strategies below, this is a situation where if results are reproducible they are not true.

Conditions (2) and (3) work in conjunction. A method is reliable with respect to the strength of the signal it is designed to detect. There are a variety of ways to evaluate the reliability of statistical methods. Hypothesis tests use specificity and sensitivity. Modern model selection methods often invoke an information-theoretic measure. Intuitively, we expect a statistical test designed to detect the bias in a coin to perform well even with small sample size if the coin is heavy on heads because the data structure is simple and the signal strong. We would be fortunate, however, if a model selection method can discern between two complex models close to each other with the same sample size. If we violate condition (2) or (3), then we have an unreliable method to detect the strength of the signal. In this case, even if the true model generating the observations is in the set, we might not be able to select it with high frequency due to the mismatch between the performance of the method and the strength of the signal (see also [20] for a discussion of how method choice might affect reproducibility).

When conditions (1)-(3) are met, statistical guarantees hold in the absence of external factors that are not part of the data generating mechanism and the inference process. To quote Lindley [21]: "Statisticians tend to study problems in isolation, with the result that combinations of statements are not needed, and it is in the combinations that difficulties can arise [. . .]" Scientific claims often are accompanied by statistical evidence to support them. However, we doubt that in practice scientific discovery is always based on evidence using statistical methods whose assumptions are satisfied. A variety of external factors such as choices made in theory building, design of experiments, data collection, and analyses might affect system-wide properties in scientific discovery. Our work is motivated to develop intuition on how some of these external factors affect the guarantees made by statistical theory. In particular, we introduce external factors which violate conditions (1)-(3), and produce counter-intuitive results. We explicitly discuss two factors that have major effects on our outcomes next.

Research strategies as an external factor and their potential counter-intuitive effect on reproducibility

When scientists aim to discover a true model among a large number of candidate models, reducing the search space is critical. Our system introduces one external factor to statistical theory as *research strategies* which determine the models to be tested at each step of the scientific process thereby serving as a means to reduce the search space. However, by choosing models to reduce the search space, the research strategies also affect the frequency of testing each model. As a consequence of affecting frequencies of tests, these strategies may alter the results guaranteed by statistical theory in many ways. Results depend on how frequently these strategies are employed by the scientific community and how frequently they propose each

model. In this sense, the strategies determine the opportunity given to each model to show its value.

To clarify the effect that strategies can have on reproducibility, we give an extreme example. Consider a search space with only three possible models. We pursue the bizarre research strategy to always test two of these models against each other, neither of which is the true model generating the data. Then:

1. The true model will never be selected because it is never tested.
2. Between the two models tested, the model that is closer to the true model will be selected with higher frequency than the model further.
3. The result stated in item (2) is reproducible.

This toy example shows that we can follow strategies which produce results that are reproducible but not true. In this work, we further show that counter-intuitive results like this can arise under mild research strategies that modify the search space in subtle ways. However, our results *do not mean* that true results are not reproducible. In fact, this is a reasonableness check that we have in our system: provided the three conditions of the previous section are met, *true results are reproducible*.

System updating as an external factor and its effect on reproducibility

A second external factor that our system introduces is the temporal characteristic of the scientific discovery. Probabilistic uncertainty dictates that one instance of statistical inference cannot be conclusive even if the true model is included in the test set and it produces highly reliable data. Thus, repeated testing through time using independent data sets calls for a temporal stochastic process. A state variable defines this process whose outcome is determined as a function of this state variable with respect to a reasonable measure of success.

The natural state variable in our system is the model selected in each test. We think of this model as a pragmatic consensus of the scientific community at any given time in the process of scientific discovery. When another model is proposed, it is tested against this consensus.

There are difficulties in defining a reasonable success measure for models, however. A pragmatic consensus of the scientific community is presumably a model which withstands testing against other models to some degree. The consensus is expected to survive even if it is not selected, say, a few times. A tally of each model against every other model can be kept introducing a system memory. This tally, can be used as prior evidence in the next testing instance of particular models. Introducing memory into the temporal process is technically easy. The real difficulty is how to choose the success measure. A decision rule about when a model should lose confidence and be replaced by another model is needed. Consider the following example: Consensus model A and model B were compared two times each winning once. In the third comparison, model B wins. Should we abandon model A and make model B consensus? If not, how many more times should model B win against model A before we are willing to replace model A?

We find these questions challenging, but they help illustrate our point. One of several well-known rules from decision theory can be implemented to update the consensus. No matter which rule is chosen, however, it will affect system-wide properties including the reproducibility of results. In this sense, a decision rule is another external factor: Precisely because the rule dictates when a model becomes consensus, it has the power to alter the frequency of statistical results in the process, otherwise obtained in isolation.

Even without the complication of a decision rule, scientific strategies make our system complex. They produce a diverse array of results whose implications we do not fully understand. Hence, we left the complication of model memory out of our system by choosing to update the consensus with the selected model at each test. This corresponds to a memoryless 0–1 decision rule. We admit that this memoryless property of our model is unsatisfactory and might not reflect a realistic representation of the scientific process. We caution the reader to interpret our results with care on this aspect. On the other hand, we are interested in system-wide and aggregate results of our model through time. That is, we look at the rate of reproducibility and other properties of scientific discovery in a given process by integrating across many independent iterations of tests and systems. Thus, our reasonableness checks still apply. For example, we expect (and find) the scientific consensus to converge to the true model once it is discovered, and the true model to be sticky and reproducible. While the memorylessness of our system might prevent the scientific process being realistically captured at any given point in the process, on the aggregate we are able to observe certain realistic patterns.

Results

First, we present results in a system with no replicator where properties of our scientific process can be obtained for all true models in our model space using Markov chain theory and computationally efficient Monte Carlo methods (S5 File). We use this computational advantage to gain insight into process properties and to inform ABM experiments for the system with replication, in which exploring all model space is computationally unfeasible. Second, we present results from these ABM experiments (S2 Table).

Results in a system with no replication

We examine stickiness of the true model, time spent at the true model, and mean first passage time to the true model for populations composed of different proportions of *Tess, Mave*, and *Bo*. Our interest is in how the proportion of different research strategies represented in scientist populations influences these desirable properties of scientific discovery. A key feature of the theoretical calculations we present in this section is the implementation of *soft* research strategies where all scientists are allowed to propose a model not consistent with their strategy with a small probability. Technically, this feature guarantees that the transition probability matrix of the Markov chain is well connected. In the system with replicator, we investigate *hard* research strategies—where scientists are allowed to propose only models consistent with their strategy—in addition to soft strategies. We compare results across four scientific populations, two model comparison statistics—Akaike's Information Criterion (AIC) and Schwarz Criterion (SC), and all possible true models in our model space (Fig 1A). We fix a low error variance to model expectation ratio (1: 4).

Stickiness, the probability of the true model staying as global model once it is hit, is high under low error both for AIC (Fig 1B) and SC (Fig 1C). This result serves as a reasonableness check for our theoretical model. Once the true model becomes the consensus, it stays as such most of the time. The stickiness of the true model increases with complexity, except for the *Bo*-dominant population. For *Bo*-dominant population, stickiness decreases with complexity, except for the most complex true model which *Bo* cannot overfit.

Even though the true model is sticky under low error, and hence, tends to stay as global model once it is hit, the system still spends considerable time at models that are not true. For example, *Bo*-dominant population overfits unduly complex models, spending only 25% of the time at the true model under AIC and 48% under SC (S11 and S12 Figs). This population spends most time in models more complex than the true model. Out of 14 true models in our

model space, under AIC 4 are not among *Bo*-dominant population's top 3 most visited models, and under SC 4 are not the most visited model. This is a consequence of *Bo*'s strategy to add only interaction predictors to her proposed models, which regularly pits the global model against more complex models. *Bo*'s poor performance is striking because boundary conditions show whether the relationship between variables holds across the values of other variables and hence, boundary testing is a widely used strategy for theory development in many disciplines [22, 23]. In comparison with *Bo*-dominant population, *Tess*- and *Mave*-dominant populations spend more time in the true model, regardless of its complexity (47% and 41% under AIC and 67% and 72% under SC, respectively). Overall, the theory testing and maverick strategies maximize the probability of spending time at the true model under AIC and SC, respectively.

For the epistemically diverse population, the true model is in top 3 most visited models irrespective of its complexity (S13 Fig). Because boundary testing strategy is ineffective in capturing the true model, the presence of *Bo*s causes the epistemically diverse population to spend less time at the true model (36% under AIC and 62% under SC) than *Tess*- or *Mave*-dominant populations. However, the effect of overfitting complex models by *Bo* is alleviated in this population due to the presence of other research strategies in the population and does not prevent the epistemically diverse population from consistently recovering the true model. In essence, epistemic diversity protects against ineffective research strategies.

We assessed the speed of scientific discovery by the mean first passage time to the true model. In this system without replication, where the transition matrix is well connected due to the implementation of soft research strategies, the true model is hit quickly across all populations (between 3.39 and 6.43 mean number of steps) when noise-to-signal ratio is low. Increasing the proportion of boundary testers in the population, however, slows down the discovery of the true model. Further, the model comparison statistic interacts with scientist populations with respect to the speed of discovery. Under AIC, *Tess*-dominant population is the fastest to find the true model (Fig 3A, *Tess*). In comparison, as shown by red region in Fig 3A, *Bo*-dominant population is slow to discover the true model. *Tess*'s speeding up the discovery of truth is also reflected in the epistemically diverse population. Using SC as opposed to AIC as the model comparison statistic decreases differences across populations (Fig 3B), increasing the speed of discovery for all populations. The fastest population to hit the true model is the epistemically diverse population under SC. We find that the speed of discovery slows down considerably as the noise-to-signal ratio is increased to (4: 1) (S10 Fig).

These results from the system with no replication show that while the true model is sticky and reached quickly under low error in a well connected system, the scientific population still spends considerable time in false models over the long run. Moreover, proportion of research strategies in scientific populations, true model complexity, and model comparison statistic have an effect on all of these properties. Overall, *Bo*-dominant population performs poorly in most scenarios whereas *Tess*- and *Mave*-dominant populations excel in different scenarios. Epistemically diverse population minimizes the risk of worst outcomes. These patterns that we described change substantially as the ratio of error variance to model expectation in the system increases (S1–S10 Figs). We now discuss the implications and limitations of results presented so far for the scientific practice.

Implications and limitations. When the truth exists and is accessible, we find that scientific process indeed discovers and sticks to it in most situations. The exceptions to this result come from 1) research strategies that search the model space in a biased manner and fail to test the true model against alternatives often enough, and 2) large error in the data generating process. In practice, (1) might be realized when there is no overarching theoretical framework guiding the search of model space but instead folk theories or intuitions are used to reduce the possibilities [24]. Further, (2) is a real challenge, especially in disciplines where data do not

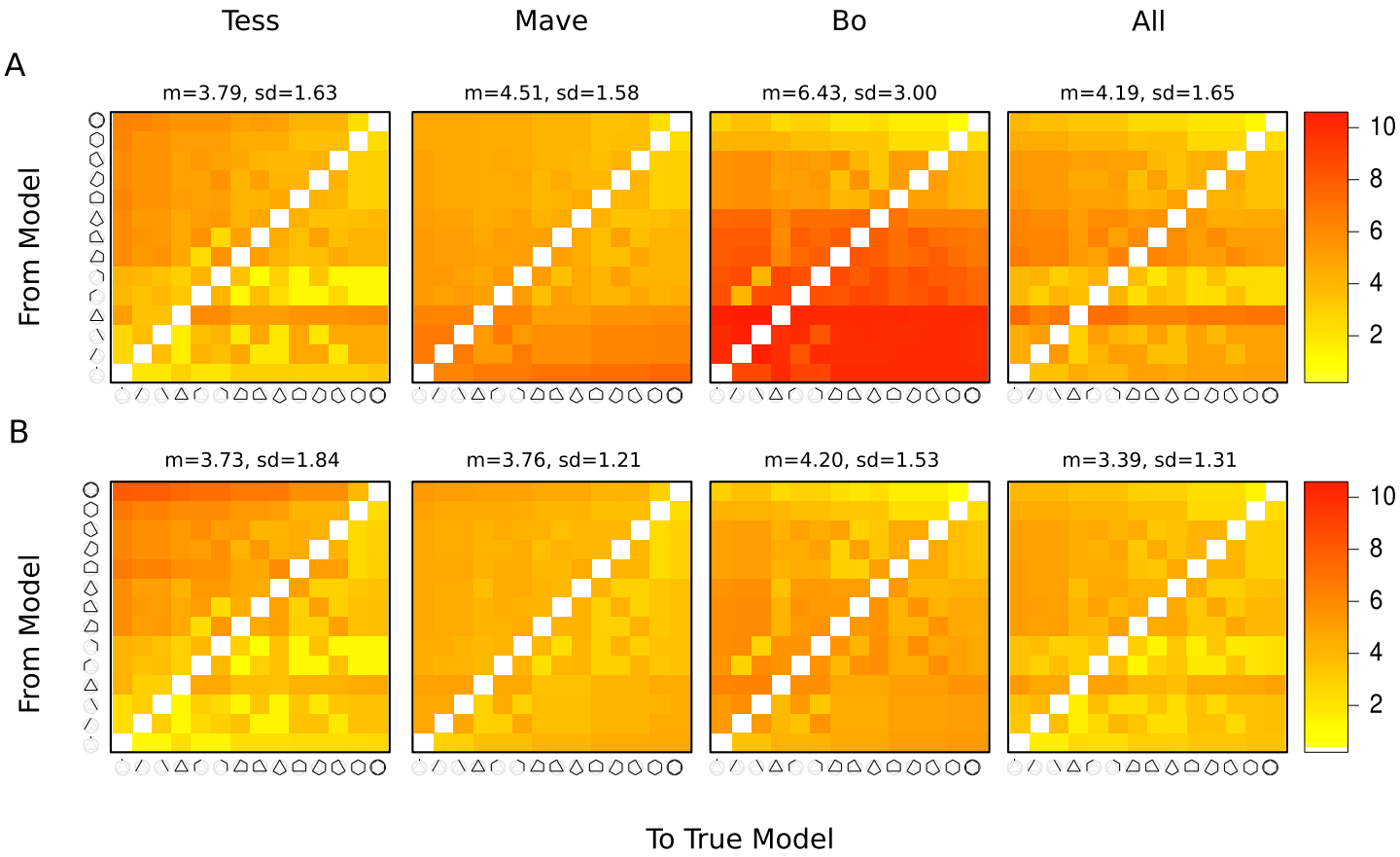


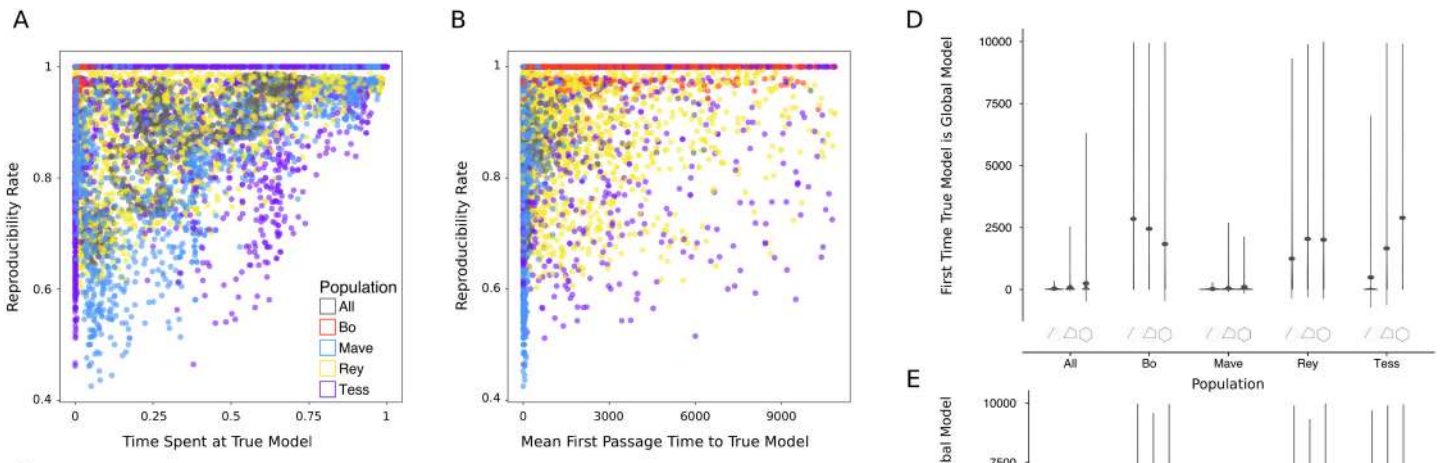
Fig 3. The mean first passage time from each initial model (vertical axis) to each true model (horizontal axis) using AIC (A) and SC (B) as model comparison statistics per scientist populations. *All* stands for epistemically diverse; other populations are dominant in the given type.

https://doi.org/10.1371/journal.pone.0216125.g003

carry a strong signal (e.g., low estimated effect sizes in psychology literature) or whose methods are not sufficiently fine-tuned to detect the signal (e.g., high measurement error). While these are implications that might hold qualitatively for real-life scientific practice, we caution the readers to not over-generalize specific parameters such as the stickiness of a true model and the proportion of time spent at the true model. These quantities depend on the parameters of our system, such as the number of models in the universe, and the linear models framework.

Results in a system with replication

In addition to the properties analyzed in the previous section, in the system with replication we can also analyze the rate of reproducibility since we introduce a replicator in the system. One of our goals is to understand the relationship between reproducibility and other desirable properties of scientific discovery. Informed by the findings reported in the previous section, we run the ABM under three true models of varying complexity, three levels of error variance to model expectation ratio, five scientific populations, and two model comparison statistics (S2 Table). Moreover, ABM allows us to implement hard research strategies where scientists propose only models complying with their strategies and all models incompatible with a given research strategy have zero probability of being proposed by the scientists pursuing that strategy (S4 File). Thus, connectedness among models is restricted in this framework for all scientists with the exception of *Mave* whose research strategy allows her to propose any model at



C

	Median					IQR				
	Rey	Tess	Mave	Bo	All	Rey	Tess	Mave	Bo	All
Time Spent at True Model	23.9	30.6	40.1	0.3	26.0	41.9	60.8	66.5*	0.7	43.7
Stickiness	90.8	97.7	93.0	90.0	91.9	17.2*	16.0	17.0	16.2	14.9
Mean First Passage Time to True Model	820.5	403.5	16.0	1592.5	29.0	2137.0	2267.0	36.0	3384.5*	73.0
Reproducibility Rate	89.9	93.1	90.3	100.0	90.2	12.9	20.0*	17.6	<0.01	12.3
Reproducibility Rate if True Model is Global Model	90.8	100.0	97.5	100.0	91.7	17.2*	8.3	11.1	<0.01	14.0

Fig 4. For *Rey*-, *Tess*-, *Mave*-, *Bo*-dominant, and epistemically diverse populations: (A) The rate of reproducibility against time spent at true model. (B) The rate of reproducibility against mean first passage time to true model. (C) Summary statistics with highest IQRs indicated by *. Mean first passage time to true model in number of time steps; all else in percent points. Violin plots for the mean first passage time to the true model per population type versus (D) complexity of true model and (E) error variance to model expectation ratio. Dots mark the means.

https://doi.org/10.1371/journal.pone.0216125.g004

any given time, and who thereby maintains a soft research strategy. When the transition matrix is highly connected, the discovery of truth is fast, as shown in the previous section. In the current section, we explore how the speed of discovery changes under restricted connectedness for different scientist populations.

Reproducible results do not imply convergence to scientific truth. We first explored the relationship between the rate of reproducibility and other desirable properties of scientific discovery, and found that this relationship must be interpreted with caution. In our framework, we defined the rate of reproducibility as the probability of the global model staying the same after a replication experiment. We show that the rate of reproducibility has no *causal* effect on other desirable properties of scientific discovery including: the probability that a model is selected as the global model in the long run, the mean first time to hit a model, and stickiness of a model (see S6 File for mathematical proof). Thus, although multiple confirmations of a result in a scientific inquiry lend credibility to that result, withstanding the test of multiple confirmations is not sufficient for convergence to scientific truth.

On the other hand, desirable properties of scientific discovery and the rate of reproducibility might be correlated. Whether there is any correlation depends on the research strategies and their frequency in the population (see S6 File for mathematical explanation). We present scatter plots (Fig 4A and 4B) as evidence for the complexity of these correlations across scientist populations. From these scatter plots and Fig 4C, we see, for example, that *Bo*-dominant populations reach perfect rate of reproducibility while spending little time at the true model (as assessed by the very low Spearman rank-order correlation coefficient, $r_{SR} = -0.06$), which confirms that high rate of reproducibility does not imply true results.

Across all simulations, as the rate of reproducibility increases, scientist populations do not necessarily spend more time on the true model, as indicated by a lack of correlation between rate of reproducibility and time spent at the true model ($r_{SR} = -0.02$, Fig 4A). Further, as the rate of reproducibility increases, the discovery of truth slows down rather than speeding up as shown by a positive but small correlation ($r_{SR} = 0.26$, Fig 4B). Crucially, both of these correlations are driven by the research strategy dominant in the population and should only be taken as evidence for the complexity of these relationships between rate of reproducibility and other desirable properties of scientific discovery (see S3 Table for all correlation coefficients per scientist population). For example, *Bo*-dominant population reaches almost perfect reproducibility (Fig 4A, 4B and 4C, red) while taking a long time to hit the true model and spending short time at it. On the other hand, *Mave*-dominant population hits the true model quickly and spends long time there (Fig 4A, 4B and 4C, blue), but it has a much lower rate of reproducibility than *Bo*-dominant population.

Implications and limitations. This counter-intuitive result on reproducibility is due to violating assumptions of statistical methods. Statistical theory guarantees to find the true data generating mechanism as the best fit for the observed data, if a reliable method operates in the absence of external factors. The research strategies implemented in our ABM include critical external factors that determine how the model space is searched. For *Bo*-dominant populations, we get high levels of reproducibility and low level of actual discovery for the same reason: Models proposed by *Bo*s consistently result in fitting overly-complex models to data but lead to reproducible inferences since the comparisons are often between models that are 1) not true and 2) far from each other in the sense of statistical distance. As a result, even though the true model is not proposed, and hence not selected as the global model, the method consistently favors the same untrue model when a specific comparison is repeated with independent data.

Mave-dominant populations search the full model space and consequently discover the true model quickly. Their rate of reproducibility is lower relative to *Bo*-dominant populations. This is because randomly proposed models are typically not true, but also, *Mave*s do not have a biased strategy of proposing models that are far away from the global model. Hence, *Mave*'s comparisons do not always favor a specific model especially when models close to each other are tested.

The effects discussed in this section are marginal main effects of scientist populations over other factors that we vary in our ABM, including levels of noise-to-signal ratio. Due to this marginalization, the mean noise-to-signal ratio that affects the results is higher in our ABM than the system without replication. Nonetheless, our reasonableness checks still capture the salient properties of scientific discovery well. For example, the basic expectation that a successful scientific endeavor will move us closer to truth is captured. Fig 4C confirms that most scientific populations (with the exception of *Bo*-dominant) spend considerable time at the true model. In all populations, the true model is sticky. The rate of reproducibility under the true model is higher than the overall rate of reproducibility across all models that become the global model. The scientific community ultimately discovers the true model with varying speed, depending on dominant strategies represented in the population. We also find that the rate of reproducibility is highly positively correlated with stickiness of the true model in most scientist populations except *Bo*-dominant (S3 Table). The rate of reproducibility is also positively correlated with time spent at true model for most populations, although these correlations are expectedly lower because they are unconditional on the first time to discovery. Speed of discovery has low correlation with the rate of reproducibility. This makes intuitive sense because speed of discovery is largely determined by how the model space is searched (i.e., an external factor) whereas the rate of reproducibility is collectively determined by all variables in the system.

Against this backdrop, we speculate how these counter-intuitive findings might extend to the practice of science beyond our theoretical framework. What we observe appears akin to the tension noted by Shiffrin, Börner, and Stigler [25] regarding the risk of obstructing scientific exploration by imposing restrictions on how science should be conducted. Indeed, we show that exploratory strategies (represented by *Mave*s in our system) are needed to speed up scientific discovery. But then we also need scientists testing theory and running replication studies (e.g, *Tess* and *Rey*s) to establish which discoveries are *true*. If we restrict exploration to allow only research strategies that search the model space in an extremely biased manner (e.g, *Bo*s), we may lock ourselves in a vicious circle of never making a discovery. The reason is that we may be able to obtain high rate of reproducibility as an artifact of this research strategy. Our cases may be extreme and in reality we might expect diverse scientist populations to emerge naturally. However, past research suggests that if incentive structures reward *Bo*-like strategies due to high rates of reproducibility they report, these strategies may be widely selected for in scientific populations [7] thereby resulting in canonization of false results [9].

Innovation speeds up scientific discovery

Mave-dominant population is the fastest to hit the true model (Fig 4C, S14A Fig) regardless of the true model complexity and the error variance to model expectation ratio (Fig 4D and 4E). Further, for epistemically diverse population in which all scientist types are equally represented, the proportion of mavericks is sufficient to garner this desirable result. The reason is that *Mave* provides connectedness in transitioning from model to model via her soft research strategy even when all other scientists represented in the scientific population pursue hard research strategies. All other homogeneous populations take a long time to reach the truth due to pursuing hard research strategies. For example, the estimate for mean first passage time to the true model for *Bo*-dominant population is 1592.5 steps (Fig 4C). We also ran the ABM with soft research strategies and include the results regarding speed of discovery (S7 File) as further confirmation that connectedness among models leads to faster discovery in *Tess*- and *Bo*-dominant populations, besides *Mave*-dominant and epistemically diverse populations.

Implications and limitations. The idea that innovative research plays a significant role in scientific discovery is intuitive and hardly new [14–16]. Our results qualify this idea in a particular way: Innovation leads to fast discovery, which is a property determined by the stochastic process governing the connectedness of models. We should note that the memorylessness property of our system might have exaggerated the role of *Mave*s in making a quick discovery. If all scientists carry a tally of past results and adjust their strategies accordingly, it is possible that the model space could be explored more efficiently by scientist types other than *Mave*. What is needed in essence is not *Mave*s necessarily but a way to guarantee high connectedness among models in the search space and an efficient search algorithm. Arguably the role of innovative, exploratory research is more critical early on in the research cycle and once we are in the vicinity of truth, limited scientific resources might be better spent elsewhere (e.g., confirmatory research or pursuit of other research questions).

Epistemic diversity optimizes the process of scientific discovery

We looked at which scientific population optimizes across all desirable properties of scientific discovery. Fig 4C summarizes the sample median and interquartile range for the time spent at, the stickiness of, and the mean first passage time to the true model, as well as the rate of reproducibility (also see S14 Fig). These statistics show the advantage of an epistemically diverse population of scientists on the efficiency of scientific discovery. Homogeneous populations with one dominant research strategy tend to perform poorly in at least one of these desirable

properties. For example, *Rey*-dominant population has low median rate of reproducibility. *Mave*-dominant population has low median rate of reproducibility and high variability in time spent at the true model. *Tess*-dominant population has high variability in mean first passage time to the true model and the rate of reproducibility. *Bo*-dominant population has low median time spent at the true model, low median stickiness, and high variability in mean first passage time to the true model. In contrast to all these examples, epistemically diverse population *always* performs better than the worst homogeneous population with respect to *all* properties and further, it has low variability. Thus, epistemic diversity serves as a buffer against weaknesses of each research strategy, consistent with results from the system with no replication. We conclude that among the scientist populations we investigate, epistemic diversity optimizes the properties of scientific discovery that we specified as desirable.

Implications and limitations. We believe that the importance of epistemic diversity is intuitive, yet, it cannot be emphasized enough. Our definition of epistemic diversity is limited to the representation of the four research strategies that we included in our system. In reality, there are numerous philosophical (e.g., logical positivist, post-modernist), research methodological (e.g., empirical experimentation, computer simulations, ethnography), and statistical (e.g., frequentist, likelihoodist, Bayesian) approaches to conducting science and our model is agnostic as to what kind or what degree of epistemic diversity would optimize scientific discovery. We merely find that the role of epistemic diversity in scientific population is akin to diversifying an investment portfolio to reduce risk while trying to optimize returns.

Methodological choices affect time spent at scientific truth

The choice of method may appear to be perfunctory if multiple methods perform well. However, violating the assumptions of a method affects the results of an analysis performed with that method. The effects of the model comparison statistic in our system, where a comparison of misspecified models is routinely performed, is not trivial [26]. When true model complexity is low, using SC for model comparison increases the time spent at the true model compared to AIC (S15A Fig). As model complexity increases, however, this difference disappears and further, AIC has lower variability. When the ratio of error variance to model expectation is low, SC leads to a longer time spent at the true model. As the ratio of error variance to model expectation increases, AIC and SC spend comparable amount of time at the true model, but AIC has smaller variability (S15B Fig). Averaged over all other parameters, SC spends longer time at the true model than AIC (*medians* = 27.05% and 19.83%, respectively), but with greater variability (*IQR* = 66.03% and 33.80%, respectively).

Implications and limitations. The finding that methodological tools might affect scientific progress is factually known [27, 28] and being studied extensively by statisticians and meta-scientists alike. Model comparison methods such as AIC and SC as well as all other statistical inference methods work best when their assumptions are met and might lead to invalid inferences under assumption violations. An unsurprising implication of our findings is that statistical theory should inform statistical practice even in the absence of well-known procedural violations such as p-hacking.

Conclusion

We studied the process of scientific discovery and reproducibility in a meta-scientific framework using a model-centric approach. We have chosen a model-centric approach because 1) it translates to scientific models directly, 2) it is a generic mode of inference encompassing hypothesis testing, and 3) model selection methods bypass difficulties associated with classical hypothesis testing.

Our scientists engage in straightforward research strategies and do not commit experimenter bias, learn from their own or others' experiences, engage in hypothesis testing, or commit measurement errors. Further, they are not prone to QRPs or structural incentives. We also assume that there exists a true model that our scientist population attempts to discover and that this true model is within the search space readily available to the scientist population. These factors that we have abstracted away are potential avenues for future research, particularly for complex social dynamics, but our goal here was to explore how the process of scientific discovery works in an idealized framework. We did, however, provide reasonableness checks to make sure that our system behaves in meaningful ways with respect to what we would expect from a well-functioning scientific process.

Our study shows that even in this idealized framework, the link between reproducibility and the convergence to a scientific truth is not straightforward. A dominant research strategy producing highly reproducible results might select untrue models and steer the scientific community away from the truth. Reproducible false results may also arise due to bias in methods and instruments used, as discussed by Baumgaertner and colleagues [20]. While both reproducibility and convergence to a scientific truth are presumably desirable properties of scientific discovery, they are not equivalent concepts. In our system inequivalence of these concepts is explained by a combination of research strategies, statistical methods, noise-to-signal ratio, and the complexity of truth. This finding further indicates that issues regarding reproducibility or validity of scientific results should not be reduced down to QRPs or structural incentives. Considering such methodological and institutional factors, however, would add additional layers of complication, moving us even further away from the guarantees provided by statistical theory.

Not all our results are as counter-intuitive however. On a positive note, we find that the process of scientific discovery is rendered efficient if the transitions between models in the model space are easy. Scientist populations that expedite transitions via promoting innovative research or pursuing flexible strategies will discover the truth quickly. In real life, we surmise that the model space might be much larger and the true model–if it exists–might not necessarily be easily accessible in the search space. Therefore, an outstanding challenge for science appears to be to attain a scientific population that can realize optimum connectedness in the model space to expedite the discovery of truth.

Recently, Shiffrin, Börner, and Stigler [25] have warned against "one size fits all" type of approaches in science and scientific reforms, advising a nuanced approach instead (p.2638). Complementary to their perspective, our results also advise against homogeneity in scientific practice. We find that a diversity of strategies in the scientific population optimizes across desirable properties of scientific discovery—a finding consistent with the cognitive division of labor literature [29]. If populations are largely homogeneous, with one research strategy dominant over others, then the scientific population tends to perform poorly on at least one of the desirable properties which might mean forsaking reproducibility or delaying discovery.

We find that the choice of statistics relative to true model complexity has non-trivial effects on our results. This is corroborated by recent statistical theory [26]. The difficulty is that the complexity of the true model is often unknown to scientists who make not only their statistical inference but also their methodological choices under uncertainty. We believe that model complexity may have differential effects on the desirable properties of scientific discovery depending on the choice of statistic.

Our model, as any other model, is an abstraction of reality and we believe that we have captured salient features of the scientific process of interest to our research questions. Main limitations of our framework are the lack of capacity to learn and memorylessness of scientists. The replicator only provides meta-level information about the scientific process and does not

contribute directly to the accumulation of scientific knowledge. Incorporating past reproducibility of specific results in decision making strategies might allow the replicator to make substantial contributions to scientific discovery. A realistic implementation of this aspect requires our virtual scientists to adopt machine learning algorithms that can heuristically teach them to become intelligent agents.

Our research also raises questions with regard to reproducibility of scientific results. If reproducibility can be uncorrelated with other possibly desirable properties of scientific discovery, optimizing the scientific process for reproducibility might present trade-offs against other desirable properties. How should scientists resolve such trade-offs? What outcomes should scientists aim for to facilitate an efficient and proficient scientific process? We leave such considerations for future work.

Supporting information

S1 Algorithm. Agent-based model algorithm.
(PDF)

S1 Code and Data.
(PDF)

S1 Fig. The three most visited models for scientist populations with one dominant type and the proportion of time spent at each true model, when AIC is the model comparison statistic and noise equals the signal in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1:1)$, proportion of time spent by a model as the global given a true model, assessed by AIC. Three most visited models are shown. Numbers show time spent at each model in percent points. True models are in red. *Tess*- and *Mave*-dominant populations perform more poorly than they do under low error, however, they still spend more time at the true model than any other models. Surprisingly, *Bo*-dominant population captures the true model more often now than under low noise although it still performs relatively poorly as compared to other homogeneous populations.
(EPS)

S2 Fig. The three most visited models for scientist populations with one dominant type and the proportion of time spent at each true model, when SC is the model comparison statistic and noise equals signal in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1:1)$, proportion of time spent by a model as the global given a true model, assessed by SC. Three most visited models are shown. Numbers show time spent at each model in percent points. True models are in red. Under SC and with this level of noise, *Tess*-dominant population performs more poorly than both *Mave*- and *Bo*-dominant populations, spending much less time in the true model.
(EPS)

S3 Fig. The three most visited models by the epistemically diverse population for each true model and when noise equals signal in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1:1)$, proportion of time spent by a model as the global given a true model for an epistemically diverse population. Three most visited models are shown for AIC and SC. Numbers show time spent at each model in percent points. True models are in red. For epistemically diverse population, the true model is the most visited model, for all true models except one under AIC and for all true models under SC. It spends more time in simpler models under SC than under AIC.
(EPS)

S4 Fig. The three most visited models for scientist populations with one dominant type and the proportion of time spent at each true model, when AIC is the model comparison statistic and noise-to-signal ratio is 4: 1 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (4 : 1)$, proportion of time spent by a model as the global given a true model, assessed by AIC. Three most visited models are shown. Numbers show time spent at each model in percent points. True models are in red. When the level of noise in the system is extremely high, all heterogeneous populations but *Bo*-dominant fail to capture the true model for many true models and spend little time at it overall. For *Bo*-dominant population, true model is among top three most visited models across all true models.
(EPS)

S5 Fig. The three most visited models for scientist populations with one dominant type and the proportion of time spent at each true model, when SC is the model comparison statistic and noise-to-signal ratio is 4: 1 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (4 : 1)$, proportion of time spent by a model as the global given a true model, assessed by SC. Three most visited models are shown. Numbers show time spent at each model in percent points. True models are in red. Under SC, the high performance of *Bo*-dominant population is dampened and all homogeneous populations perform very poorly.
(EPS)

S6 Fig. The three most visited models by the epistemically diverse population for each true model and noise-to-signal ratio is 4: 1 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (4 : 1)$, proportion of time spent by a model as the global given a true model for an epistemically diverse population. Three most visited models are shown for AIC and SC. Numbers show time spent at each model in percent points. True models are in red. When the system noise is high, even the epistemically diverse population cannot prevent poor performance. True model is not captured for most models and most of the time, under both model comparison statistics.
(EPS)

S7 Fig. The true model stickiness when noise-to-signal ratio is 1: 1 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1 : 1)$, stickiness of each true model as global model for each scientific population (vertical axis) for AIC (A) and SC (B). The true model is still sticky when noise is set to be equal to the signal in a system with no replication. True model is stickiest for *Tess*-dominant population (increasing with complexity) and least sticky for *Bo*-dominant population under AIC. Under SC, true model stickiness is even higher, and all populations perform comparably well.
(EPS)

S8 Fig. The mean first passage time to true model when noise is set to be equal to the signal in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1 : 1)$, the mean first passage time from each initial model (vertical axis) to each true model (horizontal axis) using AIC (A) and SC (B) as model comparison statistics per scientist populations. All means epistemically diverse; all others dominant in one type. Epistemically diverse population reaches truth fastest under both AIC and SC. Interestingly, under AIC *Bo*-dominant population is the slowest to reach the truth whereas under SC, it is the *Tess*-dominant population, especially when starting from complex initial models.
(EPS)

S9 Fig. The true model stickiness when noise-to-signal ratio is 4: 1 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (4 : 1)$, stickiness of each true model as global model for each

scientific population (vertical axis) for AIC (A) and SC (B). In this scenario, we observe a substantial decrease in true model stickiness, especially for complex models, both under AIC and SC. Level of noise in the system appears to have a large effect on whether true model will stay as global model once it is hit. In such cases, *Bo*-dominant population appears to perform better than other populations but still not as well as the cases with lower noise.
(EPS)

S10 Fig. The mean first passage time to true model when noise-to-signal ratio is 4: 1 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (4 : 1)$, the mean first passage time from each initial model (vertical axis) to each true model (horizontal axis) using AIC (A) and SC (B) as model comparison statistics per scientist populations. All means epistemically diverse; all others dominant in one type. Due to high variability in this scenario, all values greater than 25 are set to 25 for purposes of illustration. Under high noise, the speed with which the true model is hit is much lower, and the slowest when starting from complex initial models. In this scenario, *Bo*-dominant population is the most efficient out of all four populations.
(EPS)

S11 Fig. The three most visited models for scientist populations with one dominant type and the proportion of time spent at each true model, when AIC is the model comparison statistic and noise-to-signal ratio is 1: 4 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1 : 4)$, proportion of time spent by a model as the global given a true model, assessed by AIC. For *Tess*-, *Mave*-, and *Bo*-dominant populations, three most visited models are shown. Numbers show time spent at each model in percent points. True models are in red. *Tess*- and *Mave*-dominant populations capture the true model more consistently than *Bo*-dominant populations.
(EPS)

S12 Fig. The three most visited models for scientist populations with one dominant type and the proportion of time spent at each true model, when SC is the model comparison statistic and noise-to-signal ratio is 1: 4 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1 : 4)$, proportion of time spent by a model as the global given a true model, assessed by SC. For *Tess*-, *Mave*-, and *Bo*-dominant populations, three most visited models are shown. Numbers show time spent at each model in percent points. True models are in red. *Bo*-dominant population spends much more time at the true model under SC than under AIC.
(EPS)

S13 Fig. The three most visited models by the epistemically diverse population for each true model and when noise-to-signal ratio is 1: 4 in a system with no replication. For $\sigma^2 : \mathbb{E}(y|\mu_x) = (1 : 4)$, proportion of time spent by a model as the global given a true model for an epistemically diverse population. Three most visited models are shown for AIC and SC. Numbers show time spent at each model in percent points. True models are in red. For both AIC and SC, all true models are in top three most visited models.
(EPS)

S14 Fig. A comparison of all five scientist populations in ABM on three properties of scientific discovery in a system with replication. Violin plots showing marginal effects of scientist populations on first passage time to true model (A), proportion of times true model is global model (B), and true model stickiness (C). While different homogeneous populations appear to perform better/worse on different properties, epistemically diverse population (indicated by All) appears to have lowest variability across outcomes.
(EPS)

S15 Fig. Interaction of model comparison statistics with true model complexity (A) and with error variance to model expectation ratio (B) on time spent at the true model in a system with replication. (A) Violin plots for proportion of times true model is global model per model comparison statistic and complexity of true model. (B) Violin plots for proportion of times the true model is global model per model comparison statistic and the ratio of error variance to model expectation. Scientist populations spend more time at the true model under SC than AIC when the model is simple or when error variance to model expectation ratio is low (1: 4). Time spent at true model decreases and difference between AIC and SC disappears as model complexity or error variance to model expectation ratio increases.
(EPS)

S16 Fig. All properties of scientific discovery as a function of true model complexity in a system with replication. Violin plots showing marginal effects of true model complexity on first mean passage time to true model (A), proportion of times the true model is global model (B), stickiness (C), and the rate of reproducibility (D). Complexity does not appear to have a substantial direct effect on any property and most of its effect comes through interactions with other model parameters.
(EPS)

S1 File. A stochastic process of scientific discovery.
(PDF)

S2 File. Description of example system of linear models.
(PDF)

S3 File. Properties of scientific discovery.
(PDF)

S4 File. Properties of the model.
(PDF)

S5 File. Monte Carlo estimates of model comparisons.
(PDF)

S6 File. Reproducibility does not imply discovery of truth.
(PDF)

S7 File. ABM with soft research strategies.
(PDF)

S1 Table. Populations of scientists with varying proportions of scientist types.
(PDF)

S2 Table. Parameter values used in ABM experiment.
(PDF)

S3 Table. Correlations per scientist population. Spearman rank-order correlation coefficients between *rate of reproducibility* and other desirable properties of scientific discovery for each scientist population. *Overall* is averaged over all scientist populations.
(PDF)

Author Contributions

Conceptualization: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Data curation: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Formal analysis: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Funding acquisition: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Investigation: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Methodology: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Software: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Visualization: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Writing – original draft: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

Writing – review & editing: Berna Devezer, Luis G. Nardin, Bert Baumgaertner, Erkan Ozge Buzbas.

References

1. Ramsey FP. Truth and probability (1926). In: The Foundations of Mathematics and other Logical Essays. London: Routledge and Kegan Paul Ltd; 1931. p. 156–198.
2. Popper KR. The logic of scientific discovery. London: Hutchinson & Co.; 1959.
3. Kyburg HE, Smokier HE, editors. Studies in subjective probability. New York, NY: Wiley; 1964.
4. Schmidt S. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology. 2009; 13(2):90–100. https://doi.org/10.1037/a0015108
5. Ioannidis JPA. Why most published research findings are false. PLOS Medicine. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
6. McElreath R, Smaldino PE. Replication, communication, and the population dynamics of scientific discovery. PLOS ONE. 2015; 10(8):e0136088. https://doi.org/10.1371/journal.pone.0136088 PMID: 26308448
7. Smaldino PE, McElreath R. The natural selection of bad science. Royal Society Open Science. 2016; 3(9):160384. https://doi.org/10.1098/rsos.160384 PMID: 27703703
8. Higginson AD, Munafò MR. Current incentives for scientists lead to underpowered studies with erroneous conclusions. PLOS Biology. 2016; 14(11):e2000995. https://doi.org/10.1371/journal.pbio.2000995 PMID: 27832072
9. Nissen SB, Magidson T, Gross K, Bergstrom CT. Publication bias and the canonization of false facts. eLife. 2016; 5:e21451. https://doi.org/10.7554/eLife.21451 PMID: 27995896
10. Gelman A, Carlin J. Some natural solutions to the p-value communication problem—and why they won't work. Journal of the American Statistical Association. 2017; 112(519):899–901. https://doi.org/10.1080/01621459.2017.1311263
11. Benson AR, Gleich DF, Lim LH. The spacey random walk: A stochastic process for higher-order data. SIAM Review. 2017; 59(2):321–345. https://doi.org/10.1137/16M1074023
12. Epstein JM. Generative social science: Studies in agent-based computational modeling. Princeton University Press; 2006.
13. Gilbert N. Agent-based models. vol. 153. SAGE Publications, Inc; 2008.
14. Weisberg M, Muldoon R. Epistemic landscapes and the division of cognitive labor. Philosophy of Science. 2009; 76(2):225–252. https://doi.org/10.1086/644786
15. Muldoon R. Diversity and the division of cognitive labor. Philosophy Compass. 2013; 8(2):117–125. https://doi.org/10.1111/phc3.12000
16. Alexander JM, Himmelreich J, Thompson C. Epistemic landscapes, optimal search, and the division of cognitive labor. Philosophy of Science. 2015; 82(3):424–453. https://doi.org/10.1086/681766
17. Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6(2):461–464. https://doi.org/10.1214/aos/1176344136
18. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Proceedings of the 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado; 1973. p. 267–281.

19. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705
20. Baumgaertner B, Devezer B, Buzbas EO, Nardin LG. A model-centric analysis of openness, replication, and reproducibility, ArXiv e-prints. 2018;Nov:arXiv:1811.04525.
21. Lindley DV. Philosophy of Statistics. Journal of the Royal Statistical Society: Series D (The Statistician). 2000; 49(3):293–337.
22. Gonzalez-Mulé E, Aguinis H. Advancing theory by assessing boundary conditions with metaregression: A critical review and best-practice recommendations. Journal of Management. 2017; 44(6):2246–2273. https://doi.org/10.1177/0149206317710723
23. Whetten DA. What constitutes a theoretical contribution? Academy of Management Review. 1989; 14(4):490–495.
24. Muthukrishna M, Henrich J. A problem in theory. Nature Human Behaviour. 2019; https://doi.org/10.1038/s41562-018-0522-1 PMID: 30953018
25. Shiffrin RM, Börner K, Stigler SM. Scientific progress despite irreproducibility: A seeming paradox. Proceedings of the National Academy of Sciences. 2018; 115(11):2632–2639. https://doi.org/10.1073/pnas.1711786114
26. Lv J, Liu JS. Model selection principles in misspecified models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2013; 76(1):141–167. https://doi.org/10.1111/rssb.12023
27. Box GEP. Science and statistics. Journal of the American Statistical Association. 1976; 71(356):791–799. https://doi.org/10.1080/01621459.1976.10480949
28. Nelder JA. Statistics, science and technology. Journal of the Royal Statistical Society: Series A (General). 1986; 149(2):109–121. https://doi.org/10.2307/2981525
29. Zollman KJS. The epistemic benefit of transient diversity. Erkenntnis. 2009; 72(1):17–35. https://doi.org/10.1007/s10670-009-9194-6