Slobodan Perović and Vlasta Sikimić
Department of Philosophy
University of Belgrade

How Theories of Induction Can Streamline Measurements of Scientific Performance

Abstract

An inductive analysis of the scientific reasoning process can streamline operational assessments of scientific performance by determining whether the scientific domain at stake is inductively suitable for such assessment. Bringing together inductive analysis (based on Machine/Formal Learning Theory) and operational (citation metrics-based) assessment together, we propose a streamlined hybrid form, whereby the citation metrics used in the operational analysis of a scientific pursuit effectively track its inductive dynamics and measure the efficiency of pursuing the inductive procedures. We demonstrate the use of such inductive streamlining in the cases of high energy physics (HEP) experimentation and phylogenetic research. We find that a general test defining basic internal inductive and external practical conditions can ensure epistemically transparent operational analysis of scientific networks.

Key words: Induction; Formal Learning Theory; Machine Learning Theory; Scientometrics; High Energy Physics; Phylogenetics

1. Introduction

There are two broad approaches to identifying optimal conditions for generating scientific knowledge: the data-driven operational approach (OA) and the hypothesis-driven inductive approach (IA). The former seeks to identify the optimal organizational structure of agents of scientific knowledge-production, such as researchers, research groups, laboratories, etc. It studies relations between their properties (e.g. the size of groups, their social and cognitive diversity, their hierarchy, the relations between researchers and labs, etc.) and the outcomes of their work as these are revealed in specific data (publication rates, citations, patents, educational aims, etc.). The latter approach focuses on identifying optimal and suitably formalized inductive procedures used by the agents or communities to generate reliable knowledge. In other words, OA explores how to optimize scientific groups or networks to make best conclusion acquisition, while IA explores how to optimize the scientific conclusion acquisition process in general. The approaches have developed independently in terms of their focus and methodology. The results of OA are usually published in science and research policy journals, with some recent overlap into social epistemology journals, while IA has developed mainly within the field of philosophy of science in symbiosis with relevant methods and conceptual insights in computer sciences.

The main benefit of OA is its immediate applicability to science policy, something enthusiastically exploited by policy makers. Its scope is limited but precise, making its results immediately applicable. Studies based on citation metrics, one of the main tools of OA, have powerful policy implications for relevant institutions. Consequently, in recent years, both the funding and the organizational structure of scientific institutions have been predicated on them to a large extent.

We propose the inductive streamlining of OA to track the actual inductive reasoning process by means of external operational properties, i.e., citation metrics. This will provide a qualitatively new kind of insight. At the same time, meeting the conditions for achieving it will take care of some of the difficulties OA typically encounters. *Alongside the usual operational insights into the agents' efficiency, our hybrid approach delivers concurrent internal inductive* *(both formal and informal) insights into the domain of application, as it enables citation metrics to track the reasoning dynamics of the scientific pursuit within that domain. It thus makes transparent the core of the target of the operational analysis.*

As a necessary first step, the domain of potential application of citation metrics must be identified as a methodologically coherent pursuit if it is to serve as the domain of the transparent use of OA. As we go on to argue here, the inductive assessment along with the use of citation metrics offers a general way of streamlining OA and enables epistemic transparency, even though particular instances of it, as we demonstrate in our two case studies, high energy physics and phylogenetics, may be successful only in exceptional circumstances. Given possible the use of citation metrics in phylogenetics, we expand the case study to query metrics' use in biology more generally.

In the analysis of the first case, we start from an obvious fact: conservation laws constitute the baseline principle of parsimonious inductive reasoning in High Energy Physics (HEP) and, as such, constitute the baseline of both preliminary informal, as well as various levels of formal inductive analysis of relevant scientific pursuits, including inductive reconstruction based on Formal Learning Theory.[1] In this area of experimental physics, the convergence on results is fast, stable, and relevant over long periods of time (decades), something transparently reflected in the citation metrics. The IA utilizing MLT demonstrates that this state of affairs is a result of a reliable inductive pursuit: the quick convergence on the results in the community, tractable by citation metrics, turns out to be a result of the inductive reasoning of agents as essentially inductive-computing devices, which could be reconstructed via relevant computer models.

[1] As we will see, on the one hand, Formal Learning Theory treats agents as computing devices generating and parsimoniously using rules of inference as a reliable inductive method. On the other, such a process can be reconstructed by modelling algorithms and running computer programs.

In this example, we identify the internal inductive coherence of the pursuit to which OA (citation-based) can be applied and the exact traits the citation metrics will measure. Thus, in this case, *the citation metrics effectively measure the efficiency of pursuing the inductive process. The efficient and inefficient laboratories or teams identified by relevant citation analyses are, in effect, efficient or inefficient at performing the inductive process characterizing the pursuit.*

In biology, the time scales of convergence on the experimental results are typically much longer than in HEP. This, generally speaking, makes the proposed hybrid analysis much harder or in some cases impossible to pursue. Yet analogous to the conservation principle in HEP, the parsimony principle is a baseline principle in phylogeny research. It states that all other things being equal, the best hypothesis is the one that requires the fewest evolutionary changes. As biologists already use a streamlined computing analysis to parse their data, then, OA might fruitfully be applied to this methodologically (inductively) streamlined pursuit.

We suggest two general tests – an MLT/inductive test and a general OA test – to determine whether a scientific pursuit can be justifiably assessed by OA. As a final note, the potential convergence of MLT-based inductive analysis with other inductive approaches, including statistical analysis, would suggest even more strongly that the optimal trade-off between reliability and duration of the experiments is realistic, rendering OA epistemically transparent by tracking the inductive process behind it.

2. **Operational and inductive approaches to epistemically optimal organization of scientific networks**

2. 1 Operational approach

The days of a lone observer who publishes her results after a long solitary process of experimentation and deliberation are mostly gone. In modern science, especially in modern laboratories, the researcher constantly acquires, updates, and revises her beliefs based on her relationship with other researchers in a local or a larger network of researchers. This has motivated fairly recent social-epistemological examinations of science in philosophical literature (Kitcher 1990, Zollman 2010, Weisberg and Muldoon 2009). Much earlier, however, the operational studies of science in science policy research embraced the subject. As we will see shortly, we can fruitfully examine and assess the inductive procedures pursued by scientific agents, understood as a process pursued by a lone researcher or by a team of deliberating researchers. In operational analysis, however, the focus is different: instead of generalizing the patterns of reasoning and inferences themselves, the structure of and relations within the networks of researchers are examined as preconditions for generating knowledge. Thus, as a strand of social-epistemological and science policy analysis, OA focuses on identifying optimal ways to organize scientific networks of agents by studying types of connections between agents, structure of networks, their size, and the extent of their centralization and seeks to identify the operating procedures most likely to yield operational efficiency of networks.

The focus of such analysis is not on a reasoning process and its patterns *per se*, but on the structure of networks and their different properties.[2] This represents a broad quantitative approach to the analysis of a scientific community, rather than an abstract analysis of credence and belief change in the reasoning process. Its upside is the derivation of a quantitative metric of

[2] The results of this sort of research are typically published in science and research policy journals with some recent overlaps with social epistemology. Notable examples, relevant to our argument, include Maruyama et al. (2015), Carillo et al. (2013), Corley et al. (2006), and Martin et al. (1984). All these methods of analysis, including computer simulations, were originally developed in Organization Theory in industrial economics (Peltonen 2016).

efficiency. It is data-driven, and it makes use of the tools of analysis and insights from various quantitative analyses of the organization of scientific institutions.

In citation metrics, perhaps the most powerful and widely used tool of OA, knowledge production and, hence, the optimality of the organization or agent can be measured through relevant metrics of efficiency: the numbers of publications of the results, the citation of the results by others, and their impact in various domains. Yet influential studies often suffer from problems typical of other social sciences research. In fact, the re-examination of the methodology of OA, primarily the use of citation metrics, is ongoing (Bormann 2017, Alexander et al. 2015, Warner 2000, MacRoberts and MacRoberts 1989) and clear methodological guidelines are lacking (Braun 2010).

First, there are problems with transparency. Murky metrics are frequently applied (Van Noorden 2014) or little qualitative analysis is provided. It is sometimes hard to see what justifies the use of citation metrics in a particular domain other than the sheer availability of data, such as citation records and/or desired research goals. The citation context explicated is often little more than a particular operational property suitable for data extraction, and no further internal coherence of the domain selected for the analysis, e.g. methodological coherence or coherent research goals, is required. Generally speaking, there seems to be an assumption in the background of such studies that all researchers within the chosen domain of analysis pursue more or less the same kind of activity (method and goals), e.g. researchers within a particular sub-field working on different tasks or even across diverse scientific institutions, so their production (publication and citation counts) can be justifiably compared.

Second, on a more practical level, the analysed citations can be dispersed across fields without really indicating an expert assessment of the papers or the value of their results. The researchers can take longer periods (years or even decades) to agree on the value of the results, but citations around the time the results were published do not reflect this. In many fields, the convergence on the results does not even occur, and the research remains atomized. On an even more mundane level, the number of published papers and citations can be overwhelming or hard to track, citations themselves can be unreliable for a number of reasons, the relevant papers may have multiple authors who may not equally contribute (Allen et al 2014), and so on.

Now, although we go on to address some of these difficulties using citation metrics, this is not our main goal; instead, we address them indirectly through the conditions we suggest for achieving a new level of epistemic transparency in OA.

We should mention some fairly recent uses of simulations and decision theory which are hypothesis-driven in the same sense as the inductive approach; however, these aim at tangible numerical results. They purport to test reliability and efficiency (time of solving a task) of a scientific pursuit. Results are taken by proponents as informative of the actual properties of scientific networks. Obviously, however, the results are not directly related at concrete target networks, the way they are in citation metrics, so they cannot be as directly used to advance science policy aims (Zollman 2010). The improvement of the simulation-target relationship can rely on ever-more detailed simulations (Rosenstock et al. 2017; Borg et al. 2017) or on empirical calibrations of the simulated model.

We propose a different way to bring together hypothesis-driven (inductive) accounts and the relevant data analysis (citation metrics). It represents a new hybrid approach to the analysis of optimal science processes, one that combines operational and inductive analysis.

2.2 Scientific agents as inductive computing devices

The inductive approach we use here is an agent/belief-centered exploration of epistemic networks. In this approach, it does not matter whether the computing agent is a deliberating collective or a solitary individual, a computer, or a network of computers. The focus is on computing and logical procedures in hypothesis-formation for sorting out data, whatever the structure of the agents. The view of inductive procedures and reasoning optimality in this sort of IA, usually labeled Formal Learning Theory (or simply Learning Theory), is informed by the insights of the Machine Learning Theory (MLT). This approach treats epistemic agents as computing agents rather than as ideal epistemological agents the way traditional epistemology does. It asks whether 'epistemic utilities, … personal probabilities, conformational commitments for how to maintain … [relevant] probabilities, and the rules of hypothetical reasoning' (Kelly, Schulte and Juhl, 248) contain an apparatus for the methods that reliably converge on the truth. Optimality is tied to reliability and convergence on the truth. Thus, '[a]n important learning theoretic project' and certainly the key to our argument 'is therefore to determine whether a proposed methodological norm prevents inquiry from being as reliable as it could have been' (Ibid., 247).

Belief revision is treated as a continuous process dependent on contingencies and, thus, as essentially unpredictable**.** But eventually, in the long run, science corrects itself (Schulte 2000). It is therefore appropriate to search for a rationale for the reliability of the method; although the analysis can never tell us when ideas converge on the truth, it can identify a pursuit as a more or less reliable way to converge on the truth. The key presupposition is that over time a process should lead to convergence, despite short-run errors. The task is to identify procedures that ensure this. *The goal, then, is to identify general principles and inference rules of the pursuit* *and demonstrate their reliability, i.e. that they are better at converging on the truth than the alternatives.*

The agents never know what evidence item is coming at them, and they do not know when they have arrived at the truth, but they can assess the reliability of the pursuit. Following an appropriate method – identified as suitable principles and rules - they will be more likely to arrive at the truth. [3]

A key methodological aspect of this approach is that we never know the final outcomes in concrete cases; only hypothetical outcomes can be justifiably considered in the inductive analysis. However, the approach can compare the epistemic standing of any particular case of pursuit of science to the general rules it generates. In other words, it generates hypothetical, not categorical epistemic norms, but these can be squared with the patterns of reasoning in concrete cases. It is this feature that we propose exploiting to the advantage of the OA. The key to such analysis are the MLT generated reconstructions of the inductive process that can serve as the inductive test of the pursuit in question.

The tools for modelling scientific procedures of reasoning and the search for optimality include causal and neural networks. The cases used in analysis are usually tentative examples or hypothetical tests applied to particular cases (Thagard 1988) or sometimes even to data sets (Chickering 2002) but, generally speaking, this sort of study of the optimality of scientific networks is conducted at an abstract level of analysis. Yet crucial for our aims is that the Learning Theory and MLT are inherently concurrent: the presumed (by Learning Theory) reliable inductive procedures (generated by agents-as-computing devices) composed of parsimonious base-line principles and inference rules, can be verified in concrete cases by generating appropriate algorithms and running computer programs that reconstruct this presumed

[3] IA is a falabilist reliabilism. This is similar to Popper's view, with the addition of various computational tools.

inductive process. Hence, the implementation of inductive procedures as characterized by Learning Theory, can be effectively tested by suitable computer reconstructions.

3. Case 1: Inductive streamlining of operational analysis of experimentation in High Energy Physics (HEP)

3.1 Convergence on actual experimental results and convergence on truth

In HEP, we deal with substantially minimized belief revisions. The convergence on the results is fast, stable, and relevant over long periods (decades, so far). Even the convergence on major discoveries such as J/psi, top quark, or Higgs boson occurred in a matter of days, weeks, or very rarely, months.[4] And retractions are rare and memorable events worthy of media attention in the HEP community.

The HEP experiments are either unique or almost unique – there is either only one or at most only a handful of similar detectors and experimental machines. Peers take into account the results of a handful of experimental centres where research actually occurs, sometimes one or two laboratories, and a limited number of experimental groups. It is thus practically impossible to avoid citing relevant published papers. In addition, the citations of the published experimental results occur almost without exception in journals within the specialized peer group of HEP experimentalists[5], as they are rarely of interest outside this already very streamlined field.[6] This means there is virtually no failure in tracking the impact of the results in publications.

[4] See e.g. historical accounts of the major discovery of J/psi in the 1970s (Ting 1997), or W and Z bosons in the 1980s (Darriulat 2004), or those of a number of other particles and their properties.

[5] The only recent significant exceptions are journals in astroparticle physics where HEP results are relevant and cited by physicists outside HEP laboratories.

[6] In this respect, despite immense resources, the structure of the experimental HEP network may be like that of experimental science in the 17th century which took place in small closed circles.

In other words, the judgement of peers on experimental results is as reliably tracked as it gets by publication and citation rates. The weighted citation metrics straightforwardly indicate which experiments are deemed inadequate, adequate, or fruitful, and, most importantly, without significant danger of divergence or polarization of the produced results in the near or distant future, as is common in other scientific fields, including some other subfields of physics.

The HEP laboratories and their organization have been the target of policy studies based on citation metrics (Perovic et al. 2016; Martin and Irvine 1984a, 1984b), taking advantage of the fact that this fast and stable convergence is reliably reflected in citation counts.[7] These studies explicitly or implicitly rely on the following two aspects: a) time of convergence as the key factor of applicability, and b) convergence as a reliable indicator of agreement on the results.

HEP has extraordinary traits compared to some other scientific fields. We can assess and compare the efficiency of the organization of laboratories and experiments based on citation metrics, as they provide significant assurance that the analysis will not be flawed. Yet we still need an independent argument that the quick and stable convergence on the results, that is, the actual pursuit, is not spurious, accidental, or an artefact of some peculiar traits of the scientific network in HEP. In other words, we want to know whether there is a general indicator that the convergence on the results is the result of a reliable scientific pursuit. If so, the citation metrics provide a new level of insight into the inductive reasoning dynamics involved in the pursuit.

This is where IA based on MLT can help. Is the actual fast and stable convergence on results reflective of the inductively recommended convergence on the truth by a reliable inductive process, given the nature of the pursuit in HEP? If so, and given that the fast and stable convergence on the results actually happens and is suitably reflected in the citation metrics, the

[7] It is also significant that the citations are tracked in the most advanced tracking system of that sort; INSPIRE-HEP categorizes citations into six categories, and has been in place for decades, preceding any currently used citation trackers such as Google or Thomson Reuter’s WoS.

use of operational analysis, citation metrics, in particular, is transparent, as it reveals the inductive process itself and enables the comparison of agents in the efficiency of their inductive pursuit.

3.2 F/MLT-based inductive test of the pursuit in HEP

IA based on MLT aims at identifying whether the method used in a pursuit could have been more reliable. The reliable method – identified as a set of principles and generated rules of inference - is the one that ensures convergence on the truth better than the alternatives, or is perhaps even the *only one* that provides such convergence and, as such, validates the convergence on the actual experimental results. The analysis is concerned with the point at which the method guides a scientist (or a group of scientists) to make a judgement and justifiably stop the sequence of evidence items. The method instructs that she must have enough evidence items and justifiably believe that the theory is adequate whatever future experimental results throw at her – i.e., she can justifiably project her theory in the future (Schulte 2000).

Now, in an infinite inductive process, the global critical-time constraint for making such a judgement is not an issue. Thus, this limit case tells us nothing about the specific flow of a sequence of evidence items. On the one hand, for Bayesian agents, with a long enough period, there is some wiggle room for the convergence on the truth to emerge. On the other hand, in real cases, an actual long-lasting convergence is not simply an empirical fact but may also tell us something about the stable nature of the sequence of the flow of evidence items. Which method is reliable for a particular critical length of pursuit depends on the specific inductive problem. Thus, there is a *critical-time constraint* on any hypothesis testing. Now, if one researcher can show that her method is more reliable than any alternatives, she can (justifiably) project her theory – i.e. *justifiably converge on the truth within the critical time. This requires identifying (e.g. by suitable reconstruction) that the principles and rules governing her inductive inferences are demonstrably more reliable than the alternatives. The key to this demonstration in our case will be to show that the rules of inference based on the core principles are restrictive enough over the data set (actual experimental data); i.e. there are few alternatives or indeed no possible ones whatsoever.*

The practicing experimental particle physicists constrain their derivations from data (i.e. their hypotheses) using conservation principles (conservation of the momentum, energy, spin, charge). They choose the conservation principles which effectively rule out as many unobserved particles as possible, the existence of which would violate them. Thus, in practice, the analysis of particle trajectories is based on the conservation laws; e.g. different potential identities of particles are calculated based on the assumption of the conservation of the momentum of the in-coming and out-going particle tracks. In other words, they opt for the closest fit with the data.[8] The most likely outcomes are selected based on the obtained data and phenomenological (rather than high-level theoretical) models, used for simulation runs, essentially predicated on the conservation principles.

Although it is hard to deny that 'the research program for searching for selection rules [in particle physics] has justified itself by its success so far' (Schulte 2000, 776), the IA analysis should independently reveal the link between the convergence in practice and the inductive reliable convergence on the truth. The discovery of new particles, even very surprising ones, is always possible, but the point is that the stream of actual discoveries based on the sequence of evidence (particle interactions) in the pursuit is the product of an inductively reliable method.

[8] See e.g. (Dissertori et al. 2003).

In fact, without the parsimonious use of conservation principles it would be hard to imagine modern HEP experimentation. The methodology of the field reduces to it in a fairly straightforward way. This is why physicists themselves have been motivated to reconstruct the inductive method driving the field. The conservation principles are the baseline principles of HEP practice and also can be identified as the baseline principles of the inductive process. Thus, both real-world practical derivation procedures and IA/MLT will make recommendations through inference rules based on insights bounded by this same baseline. Now, the baseline principles, in effect, generate suitably applicable selection rules over the experimental data set by providing a restrictive system of inferential rules.

The computable inferential mechanisms that adequately grasp the actual inductive process in particle physics have been investigated (Schulte and Drew 2010; Valdés-Pérez and Żytkow 1996; Valdés-Pérez and Erdmann 1994; Kocabas 1991). Suitable algorithms and models have been constructed and even used for the discovery process, where the inductive process is modelled and computed based on little more than the conservation principles over a data set. Thus, the pursuit has not only been modelled as an inductive process within the MLT framework but also been implemented in the actual pursuit.

These models were either supplied with given constraints or built from scratch. Using the conservation principles parsimoniously if the simplest model does not capture a hypothesis or a set of data, a more complicated one is used.[9] As an example, the standard quark model was reconstructed through such computations, but '[p]robably the most significant result is that an exhaustive search in the space of quark models for baryons followed by the mesons reveals the

[9] Simplicity is defined as the number of constituents and the number of constituents per particle (Valdés-Pérez and Żytkow 1996, 54).

standard quark model stands out nearly uniquely as the simplest, when the constraints of complementary pairs is imposed' (Valdés-Pérez and Żytkow 1996, 2109).

In fact, the key part of this formalization is the proof of the restrictive selection of the rules: 'Under pure induction (i.e. without additional assumptions)' other than those provided by conservation principles and the data set, 'more than one selection rule and quantum property are never needed to distinguish any set of allowed [particle] reactions from any set of prohibited ones' (Valdés-Pérez and Erdmann 1994, 172). This characterization applies to a somewhat simplified model, but computing based on more robust models shows unique determination as well, as demonstrated by Schulte and Drew (2010). In fact, the constraint on the selection rules is strong in all models: in general, assuming conservation laws, the number of selection laws that are not redundant turns out to be small.

As Schulte (2000) points out, it is precisely this restrictiveness that warrants physicists 'projecting the theory': based on it, they are justified in expecting that the theory will be valid for some future expected evidence. Now, since the reconstructed methods of selection based on conservation principles warrant this expectation, the fast and stable convergence on the results we encounter in practice is warranted. The models thus reconstruct an inductive method that generates the procedures and results concurrent with those used by physicists for discovering particular particles (and one could even formally show[10]). It would be indeed hard to imagine a realistic (in terms of base-line principle and inferences) reconstruction of the pursuit that veers far from such models. Thus, given the results of the reconstructions, the pursuit is based on a reliable inductive method, and projecting the theory is justified in the actual pursuit, as we have the same baseline and parameters (conservation rules and evidence items) in both IA and practice.

[10] There is no need to spell out the proofs here; they can be found in (Schulte 2000).

We can, in fact, generalize this case with the IA (MLT) test, i.e., a test of the inductive coherence of the pursuit. The conditions for judging whether a pursuit is MLT-inductively coherent are the following:

1. There are computable models *matching* the actual pursuit (over a relevant set of data – actual experimental data).
2. There is a common, *pursuit-matching* core to these various models: a base-line inductive principle and a set of restrictive rules of inference.
3. By successfully computing (i.e. providing successful retrodictions and predictions) - over the data set via restrictive rules based on the postulated principle - the models *warrant* and *explain* the actual fast and stable convergence of researchers on the results.

Now, the matching models clarify the details of the pursuit but even informal analysis of the inductive process that precedes it reveals the inductive pattern. There are various levels of inductive analysis of this sort, and the informal level is certainly a much more substantial warranty of justified convergence to the results than in many other cases thanks to the nature of the pursuit, which also makes construction of formal models easier.

3.3 <u>Citation metrics and the efficiency of scientific networks in pursuing inductive processes</u>

If a pursuit passes the IA test the OA analyst is justified in treating the fast and stable convergence on the results as the indicator of the use of a reliable method in the pursuit - *the fast and stable convergence on the results is based on the warranted projecting of the theory.* And the OA of the team structure in HEP labs can provide deeper insight beyond the operational traits immediately captured by such analysis. OA applied to the pursuit in HEP can take various forms.

And we can identify a temporal constraint on the applicability of the citation metrics: the long expiry dates of metric analysis are determined by the justifiably long-term convergence on the results in the pursuit, as the revision of beliefs is justifiably minimized.[11]

In perhaps the most comprehensive study of its kind to date, a three-part assessment (Martin and Irvine 1984a, 1984b) of the performance of CERN with respect to other HEP laboratories, as well as the performance of individual accelerators of the laboratory, offered various quantified results with the ambitious normative intention of improving the performance of experimentation in HEP as a whole (Diagram 1). The number of published papers and citations were used as a key metric in this extensive comparative study of the performance (production) of major HEP laboratories.

	n ≥ 15	*n ≥ 30*	*n ≥ 50*	*n ≥ 100*
CERN	111 (26%)[12]	31 (26%)	9 (19.5%)	1 (9%)
DESY	20 (4.5%)	9 (7.5%)	3 (6.5%)	0 (0%)
Brookhaven	37 (8.5%)	6 (5%)	2 (4.5%)	1 (9%)
Fermilab	106 (24.5%)	37 (31%)	17 (37%)	1 (9%)
SLAC	75 (17.5%)	21 (17.5%)	11 (26%)	6 (54.5%)
Others	80 (19%)	15 (13%)	4 (8.5%)	2 (18%)
World total	429	119	46	11

Diagram 1: Numbers of highly cited papers across HEP laboratories within one year in the period 1969-78. (The data based on (Martin and Irvine 1984a, 1984b))

[11] Apart from establishing reliability of the results, IA has the potential to establish the computational properties of a scientific pursuit. For instance, Schulte has investigated the NP hardness of finding a simplest linear causal network from conditional correlations.

[12] All percentages are rounded to the nearest 5%.

A more recent study (Perović et al. 2016) conducted on data from Fermi National Laboratory was based on the actual data from 27 large similar experiments[13] with the goal of computing their efficiencies in relation to the team sizes (Diagram 2). The most inefficient experiments in the quantitative study turned out to be those of the largest teams in the group; they either stalled at the level of realization, or the protocols were so flawed that the data analysis could not be completed. The most efficient teams, those who excelled in weighted citation counts of the publications based on the results in the experiments performed by the teams, were smaller. These results concur with other similar studies across scientific fields (van der Wal et al. 2009; Bonaccorsi and Daraio 2005).

	Experiment	Efficiency
1	FERMILAB-PROPOSAL-0882	1.0000
2	FERMILAB-PROPOSAL-0871	0.4188
3	FERMILAB-PROPOSAL-0868	0.3276
4	FERMILAB-PROPOSAL-0866	0.4019
5	FERMILAB-PROPOSAL-0854	1.0000
6	FERMILAB-PROPOSAL-0802	0.5022
7	FERMILAB-PROPOSAL-0792	1.0000
8	FERMILAB-PROPOSAL-0789	0.3002
9	FERMILAB-PROPOSAL-0774	1.0000
10	FERMILAB-PROPOSAL-0773	0.2814
11	FERMILAB-PROPOSAL-0772	0.9432
12	FERMILAB-PROPOSAL-0770	1.0000
13	FERMILAB-PROPOSAL-0769	0.3066
14	FERMILAB-PROPOSAL-0761	0.2758

[13] Experiments are similar – i.e. homogenous in terms of techniques and other traits of the experimental process – yet varied in terms of their efficiency.

15	FERMILAB-PROPOSAL-0760	0.5000
16	FERMILAB-PROPOSAL-0756	0.4261
17	FERMILAB-PROPOSAL-0747	0.2102
18	FERMILAB-PROPOSAL-0745	0.1447
19	FERMILAB-PROPOSAL-0744	0.3641
20	FERMILAB-PROPOSAL-0743	0.1649
21	FERMILAB-PROPOSAL-0735	0.5000
22	FERMILAB-PROPOSAL-0733	0.2843
23	FERMILAB-PROPOSAL-0713	1.0000
24	FERMILAB-PROPOSAL-0711	0.2009
25	FERMILAB-PROPOSAL-0706	0.1667
26	FERMILAB-PROPOSAL-0705	0.2500
27	FERMILAB-PROPOSAL-0704	0.2087

Diagram 2: Results of the Data Envelopment Analysis comparing efficiency within a series of the experiments performed at Fermi National Laboratory based on weighted citation counts (data based on (Perović et al. 2016))

Regardless of the details and implications of the results of published studies, what really renders the use of weighted citation rates valuable is the fast and stable convergence on the results by real experimental networks and in HEP in general. As noted earlier, the citation metrics indicate fast, strong, and stable peer agreement on the experimental results. The citation metrics, then, can be considered a reliable measure of productivity, i.e. the efficiency of experimental groups (within a laboratory or across laboratories) in producing results that will guide new research. In HEP, the weighted citation counts precisely represent peers' views.

In addition, the field is unusually isolated: researchers publish and cite others in their own journals, they are not cited by external sub-fields, and the experimental centers are few and far between. Thus, the convergence is *accurately reflected* in the citation metrics. (Let us call this *an external condition* of the OA.)

Now, the existing inductive models (reconstructions) we cited are concerned with the abstract level of the theory (QCD). The experimental searches, however, are a matter of phenomenological models constructed in accord with the QCD and the Standard Model (which are again constructed in accord with even higher level of theory, Quantum Field Theory and Quantum Electrodynamics). The existing models reconstructed the process that led to the key properties in Quantum Chromodynamics (QCD), i.e. reconstruction of quarks from the experimental results, while the experiments assessed by Martin and Irvine (1984a, 1984b) account for a wider scope of the experiments. Yet some of the core experiments in the data-set, with most citations, are the key discoveries of the QCD treated by the models. Thus, the results of the existing models are a relevant indication of the kind of the inductive process taking place in the pursuit. In the case of more narrow studies such as Perovic et al. (2015) the experiments explored particle dynamics within the QCD framework rather than the core properties of QCD (i.e. quantum numbers). Thus, the models provide warranty of the core inductive strategy in the pursuit but only indirectly address the actual experiments in the data-set. One could create more custom-made matching models that suit the pursuit within the specific data-set in the studies.

Overall, however, it would be hard to argue that the fast and stable convergence is an accidental outcome unrelated to the inductive pursuit of the outlined sort. Based both on the informal assessment of the inductive process, and on the computable models broadly matching its various aspects, for all practical purposes, the agents in the pursuit act as inductive-computing devices of specific traits. Given this, the context of quick and stable convergence, namely the inductive reasoning dynamics in the network, is *the internal factor* streamlining the agreement. It indicates that *the experimental teams identified as efficient outliers or more productive laboratories - based on weighted citation analysis in the above-mentioned studies, directly reflecting the converging peer view - are significantly better than the other teams at the performance of the inductive process that characterizes the pursuit in question.* Thus, suggesting that other teams should be more like the most efficient teams in terms of the measured operational parameters (team size, number of teams, etc.) is not an operational 'shot in the dark', and the convergence result of spurious or accidental correlation, an artefact of the network, or simply a result of an unknown parameter, since we now know that the pursuit is inherently a specific inductive process and, as such, is effectively tracked by citation metrics.

We can list the following conditions that the pursuit should satisfy to be deemed suitable for the OA test:

1. Internal Condition: The pursuit passes the FLT/MLT test.
2. Empirical condition: Fast and stable convergence on the results in the pursuit.
3. External condition: Convergence is suitably reflected in publication and citation counts.

If a pursuit passes the OA test, the citation metrics effectively measure the efficiency of the inductive process in the scientific network (Diagram 3).

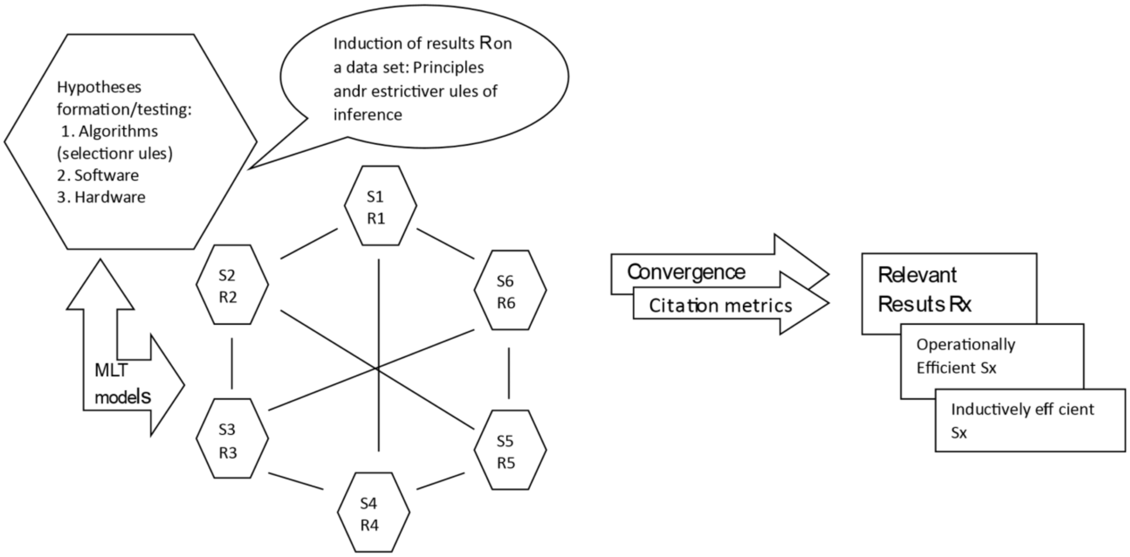

Diagram 3: Inductive/computing units of structure **S** (experimental teams or labs) identified in the pursuit of the network. Operational analysis applied on operational and inductive efficiency.

It should be noted that many experimental outcomes are not reported, as vast numbers of runs of the colliders are based on a great variety of triggers (algorithms that determine the conditions under which the events in the detector will be recorded), the vast majority of which turn out to be of no significance. This is the only reasonable solution given that only particle 'signatures' (decays), not energy scales at which novel particle interactions take place are predicted by the models, so the experimental task is to relentlessly comb vast backgrounds (i.e. known interactions) across the scales.[14] For the results deemed worthy of publication, although the efficiency metric based on weighted citations in HEP is first and foremost a measure of how reliable the results are judged to be by peers, it is also a direct indicator of the fruitfulness of the results. It tells us how excessively the peers relied on and were motivated by the results of the

[14] Most experiments do not purport to establish the existence of new particles; rather, they explore properties of the known ones. The Standard Model is a null hypothesis in vast majority of experiments; it provides the expected background interactions, so the exploratory experiments that do not turn up new particles will be null experiments - but they will also provide important information on their properties (e.g. energy scales) that the model does not deliver. Even if an experiment that does not have any results of significance is reported, it will not result in the number or quality of citations that accompany experiments with confirmatory results.

experiments to further their own work.[15] Fruitfulness gives the edge to efficient outliers over other reliable results. Experiments are fruitful when they confirm or explore a cornerstone of the model so the experiments succeeding them are bound to rely on their results. Thus, they become an indispensable part of the background knowledge of future experiments. In other words, they act as essential data constraints on the event selections within the relevant framework of the conservation laws, both reinforcing and projecting the theory.

Could it be, however, that the production of the results that turn out to be fruitful, or in other words, the measured efficiency, is a result of serendipity rather than a particularly efficient pursuit of the inductive process? Perhaps the successful efficient laboratories or teams simply stumble on fruitful hypotheses, while the inefficient ones are merely unlucky in their choice.

There are, in fact, three different levels of inefficiency tracked by citation metrics. The least inefficient experiments have a problem at the level of 'cables': they do not operate the equipment well and never really take off even though they consume lots of resources (Perović et al. 2016; Martin and Irvine 1984a, 1984b). In other words, the experimental team as an inductive-computing unit has a hardware problem. Then there are those who get stuck at the level of data analysis for various reasons. In such cases, essentially the unit cannot compute well – its deficiency is analogous to software deficiency in performing a computation. Finally, some teams never reach the highest level of efficiency because the hypothesis they test is not fruitful enough.

So it seems only in the last case could we justifiably suppose serendipity is a major factor. Yet perhaps a team makes the initial choice of the hypotheses to be tested as part of its inductive-computing process. The question, then, is to what extent the choice and formation of

[15] This was certainly true of the citation patterns of the experiments from the late 1960s to the mid-1990s – the period analysed by the above-outlined studies; now research has become so centralized that essentially all particle physicists are engaged in one mega-project.

the hypothesis (or rather a set of hypotheses) for testing is: 1) part of the overall inductive-computing process, and 2) shaped by parameters entirely external to the inductive pursuit.

Either way, the difference between stellar and mediocre results is decided at the higher level of the inductive-computing process; in other words, at the level of the choice of algorithms, i.e. the choice of generating rules. The selection rules are generated over an existing data set, so the formation of the list of potential hypotheses for testing is very restricted. We would really need to see how each list was created (along the lines of e.g. Maruyama et al. 2015) to address the possibility of serendipity at this stage of testing; e.g. it is crucial to know how sub-hypotheses are produced from a very general master-hypothesis delivered by the Standard Model or any of the alternative models in particle physics. In any case, smaller teams turn out to produce more fruitful results so it may be that smaller teams are better conditioned for superb computations at the level of picking algorithms/hypothesis, possibly because they are demonstrably better at handling hardware and software.

Finally, when the phenomenon of significance is discovered in the experiments – e.g. a substantial evidence for a new particle (e.g. Higgs-boson), the physicists do not jump to the conclusion what exact particle or property they have discovered (e.g. Higgs-boson of the Standard Model or a Higgs-boson-like particle of Super Symmetry model) as the particle may be accounted by competing models given the evidence. The inductive process leads them to converge on the discovery of a particle or a property of significance, but convergence on its exact nature as it is characterized by a specific model is a more arduous process. One could suggest that the quick and stable convergence we see in the laboratories is a result of the reasoning process more akin to deductive reasoning, or a low-level induction working with simpler data-sets and models, and that the truly inductive process never results in such quick convergence. This is certainly possible especially because the reasoning in these cases concerns phenomenological models. Yet it could be that the inductive process that leads to the convergence on a particular model is of the same sort as the one leading to fast and stable convergence on the phenomena of significance. It is just that it takes longer for new experiments to update the physicists' beliefs and turn their higher-level dilemmas into a search for new phenomena of significance.

3.4 Conclusion

To sum up, the research in HEP follows inductive rules and patterns stemming from the baseline conservation principles. This inductive process, in turn, guarantees a broad and reliable convergence on the results. Based on the inductive convergence on the reliable results, the impact of the results can be measured by weighted citations and taken as representative. In this way, IA justifies the OA identification of optimal organizational structure. If IA cannot guarantee the reliability of the results, then we are not justified in applying OA. In this way, IA based on MLT streamlines OA, although it is possible that there are other internal justifications of this sort. The success of the MLT at reconstructing the inductive method in HEP should not be surprising: the method of data gathering and analysis in HEP is, generally speaking, along the lines of MLT induction. Even a superficial glance at the field, let alone a detailed analysis, suggests this. This is the case in some subfields of contemporary experimental biology as well.

4. Case 2: Phylogenetic research

4.1 Machine learning techniques in phylogeny

In biology, generally speaking, consensus on results is not fast and reliable. The time scale is much longer than in HEP, even if consensus on the results and their relevance eventually occurs. And it is often difficult to find a coherent set of inductive rules governing the research in biology. Yet along the lines of our previous analysis, we can find subfields of biology governed by inductive principles based on the MLT, wherein the pursuit passes the inductive test and gives the green light to operational analysis.

Phylogenetics, a subfield of evolutionary biology identifying trees of evolutionary relations between species (phylogeny), is particularly suitable for such analysis. Analogously to the conservation principle in physics, in phylogenetics, the usual baseline principle is the principle of parsimony. The principle states that all other things being equal, the best hypothesis concerning an evolutionary relationship is the one that requires the fewest evolutionary changes (Yang & Rannala 2012). Analogously to an inductive principle used as a baseline in Learning Theory analysis, it makes the process of reaching a conclusion efficient, even though it does not guarantee its truth.

Over the years, the concept of 'fewest evolutionary changes' has been interpreted in various ways. In the beginning, researchers compared the set of properties of organisms. They gradually moved on to focusing on the common development of species. Finally, they started calculating similarities between sequences of genes. The principle for all three kinds of reconstruction remained the same, however, i.e., the closeness relation.

Now, in the case of HEP, as explained above, we rely on machine learning analysis of the inference procedures independent of the actual experimental process. The MLT application is an afterthought of sorts that produces the models distilling the inductive process behind the pursuit, even though it could be subsequently utilized in it. But in phyolgenetics, based on the guiding principle of parsimony, biologists run much reduced models as the primary tool of analysis, so to assess the suitability of OA, we need to assess the actual application of the relevant algorithms.

Evolutionary relationships are established based on sequence similarities between genes, and the inductive principle (‘fewest changes’) suggests that closely related organisms share a higher degree of sequence similarity. To give an example, some amino-acids are more similar than others; therefore, not every difference in the protein sequence of genes indicates the same evolutionary distance. To account for this, matrices employing observed amino-acid changes between homologous proteins in large data sets have been designed to calculate expected exchange frequencies between genes with similar evolutionary distance. Numerical scores are assigned to differences in the nucleotide, or amino acid sequence, based on the frequencies of the differences. The greater the frequency (in large data sets) the smaller the number assigned. The calculation gives an optimal tree, i.e., the one with the smallest number of differences between branches.

For example, take three sequences with the same length, AAA, AAB, BBA. If we set the expected frequencies to 1 for all differences, the resulting scores are the following: AAA-AAB:1, AAA-BBA:2, and AAB-BBA:3 (Diagram 4).

	AAA	AAB	BBA
AAA		1	2
AAB	1		3
BBA	2	3	

Diagram 4. Differences between sequences in phylogenetic analysis

To create an optimal tree, the algorithm searches for the minimal total score, i.e. the smallest sum. Thus, it will place AAB and AAA together, with BBA as an out-group, resulting in an overall score of 3, with the leaves on the resulting tree grouped as follows: (AAA-AAB)-BBA.

Note that, in general, this approach results in a reliable tree, depending on the adequacy of other assumptions. However, several obstacles can prevent researchers from finding the correct solution[16] (Yang et al. 2016). When sequencing approaches became cheaper, for example, whole genome sequences were suddenly available; the resulting tree depended on which was selected. The problem of homogeneity continues today, albeit to a lesser extent. Now, protein sequences are mainly used to generate trees, as similar homologous proteins can be found in very high amounts. To establish their relationship, researchers use matrices of the frequency of amino acid exchanges. These matrices contain data on how frequently a specific exchange occurs in sequences with a specific similarity; for example, BLOSUM62 holds the observed frequencies for proteins with a similarity of 62% (Henikoff & Henikoff 1992). In addition, inserts and deletions are scored with a specific value. All these scores can be chosen by the researcher, and this affects the result (the tree), at least with respect to some details.

A further question is which exact data are relevant for the tree reconstruction. Do we compare conserved proteins or domains, and how do we weigh exchanges in these conserved positions in contrast to variable regions? All these decisions are based on the researchers' former experience and might therefore vary. Horizontal gene transfer further complicates analysis in some cases (Koonin 2016). Even though researchers tend to construct binary trees, horizontal transfer of DNA between different branches can occur. In such cases, the real tree is not binary but a net. And because scientists can only access information about the current specimen, they

[16] Historically, researchers constructed trees solely based on the 19S RNA, because of the difficulties obtaining sequence information (Yang et al. 2016).

can only infer the sequence of the last common ancestor based on probabilities; they cannot know if they are correct. New information might force recalculation, leading to changes in the tree.

Nonetheless, the principle of parsimony in phylogeny is clearly an efficient method for generating adequate rules of evolutionary relationship.[17] The core tasks and analyses are results of a streamlined inductive-computing process around which the entire scientific process is organized. The reduced models based on parsimony are what the actual pursuit consists of, so the pursuit inherently satisfies the internal condition of the OA test. Whether it satisfies the external one (citation metrics) is less clear: as the results of phylogenetic research are of a wider significance, citation counts will be spread across various fields, much more so than in the case of HEP results. Hence, we need to be able to identify and extract expert-based citations if we are to draw conclusions concerning the inductive efficacy of various elements of the network (teams, sub-teams, labs, individuals, etc.). This requires more research.

4.2. Applicability of inductive analysis in other areas of biology

Phylogeny is one of myriad research topics in contemporary experimental biology. As different principles and approaches are applied in different subfields, the application of a hybrid of IA and OA across biology is a non-trivial task. There are various reasons why results in biology are, in general, not as quickly agreed upon and as reliable as they are in HEP. First, results that cannot be replicated are published in journals with high impact factors and get a high number of citations (Pusztai et al. 2013). Second, there are deliberately faked results because of the inefficient system of paper retraction and individual career benefits from publishing incorrect data. Third, there may be a problem deciding what constitutes sufficient evidence for a

[17] This use accords with an account of parsimony in Kelly (2004, 2007).

hypothesis, especially if the hypothesis is non-parsimonious, i.e. when the hypothesis is not the simplest explanation of the phenomenon. Fourth, an expectancy bias appears in reports on the results. These negatively influence the replicability of biological experiments and slow down consensus (Goodman et al. 2016).

In modern phylogenetics, however, data are numerically expressed, and this makes the field suitable for machine learning analysis. In many other branches of experimental biology, such as cell biology, pictures are the main data. To analyze them, interpretation is crucial. But how these pictures are interpreted is heavily dependent on the prior knowledge and beliefs of the scientist. Another relevant issue is that experimental conditions in biology are not as clearly set as they are in experimental particle physics. Even though efforts are made to provide similar conditions, it is hard to do so when it comes to, for instance, the quality of soil, light, or humidity in plant biology. For example, unless the bulbs in plant growth chambers are simultaneously exchanged and equally used in different laboratories, the light quality will not be exactly the same, and this may affect the results. In particle physics, it is substantially easier to provide equal conditions, especially since the same experimental machines (accelerators and detectors) are often simply recombined to perform different experiments.

4.3 Non-parsimonious results

To understand which evidence is sufficient for the acceptance of a hypothesis by the biological community, we need to consider the expected likelihood of the hypothesis. In the case of non-parsimonious results, acceptance is much slower than for parsimonious ones. For example, consider Koch's second postulate: all infectious diseases are caused by an organism. After showing that protease-resistant proteins, prions, cause Scrapie disease, Koch’s second postulate was abandoned (Soto 2011). It took some time for the argument that a protein can cause an infectious disease and influence the folding of other proteins to be accepted. The first experiments were conducted in 1967, and the protein hypothesis was formulated. Acceptance was bolstered by the famous results of Prusiner in 1982 (Prusiner 1982), but the scientific community was not persuaded until 1990 when mice were infected with the disease in a laboratory (Soto 2011).

The discovery of human papillomavirus as the main cause of cervical cancer took a similar path (zur Hausen 2009). It was already known that viruses could cause cancer, yet it was not accepted that a virus could be the main cause of a specific type of cancer. At the time, the disease was not considered infectious, thus a substantial number of argumentative steps was needed for establishing the correlation between the virus and cervical cancer. In the end, a uniform hypothesis that cancers cannot be caused by infectious diseases was defeated.

When it comes to hypotheses in line with the parsimony principle, the scientific community has fewer acceptance requirements. For instance, results that are in line with Koch's second postulate are accepted by the life science community quicker. Koch's proof that a bacillus causes anthrax required only two argumentative steps. In the first step, the presence of the microorganism in patients was established, while in the other step subjects were infected with the microorganism grown in pure culture.

Generally speaking, in the case of disease-causing agents, we can point to some general criteria, but we cannot find regular principles such as the conservation principle in physics. Yet as illustrated by previous cases, the acceptance of unexpected and/or non-parsimonious hypotheses takes time and requires many argumentative steps, so we cannot talk about concomitantly fast and reliable conclusions. Thus, in these cases, the relevant research line is not predicated on a baseline principle. It isn’t that data aren’t available (e.g. testing the Higgs boson hypothesis waited for three decades because the experimental apparatus was not available) but the community was always divided on the relevance of the existing data. The citation data might reflect this division, but we could not use them to decide which labs were efficient and which ones inefficient, because in this and other cases, those whose results are not cited but denigrated might emerge winners in the end, albeit after a long period of time.

5. Wider inductive convergence and adequacy of operational analyses

An inductive analysis of a scientific pursuit of the sort we discuss here provides at least minimal assurance of the methodological coherence required for operational analysis to yield transparent methodological insights into the pursuit. Yet we need not limit our analysis to the MLT-based inductive account.

Generally speaking, philosophers and theoreticians of induction are selective when choosing cases to illustrate or assess their inductive models. This means that IA is not as open-ended as we may like to think; for example, each scientific pursuit may find a fitting inductive analysis, as several inductive accounts have been developed. In fact, the few, often identical, cases invoked in discussions of IA are simply drops in an ocean of cases and represent those displaying coherence of the pursuit based on at least one inductive model. This provides at least minimal assurance that a selected domain exhibits methodological coherence, as explicated by at least one inductive method. This is much more than operational analysts typically offer in the way of epistemically transparent use of their citation metrics – which is often nothing. Even checking for basic coherence of a pursuit requires a model. And checking a sophisticated pursuit like the one in HEP or research on phylogeny requires sophisticated inductive models and tests.

The hybrid of IA and OA may not be applicable to all pursuits. For instance, exploratory pursuits often do not reflect inductive coherence and are characterized by divergent results. Other research pursuits, as we have demonstrated, have no overarching parsimonious streamlining. The application of OA cannot be inductively justified in such cases, as the inductive process behind the pursuit is not effectively computable - there is no unifying principle or restrictive rules, nor are there computable models of the pursuit. In fact, their inductive logic cannot be identified by one model alone, as background beliefs play a major role in reasoning. Material theory of induction (Norton 2003) that focuses on the factual content may be more appropriate to capture such pursuits. Moreover, it may be problematic, from this standpoint, to apply epistemically transparent OA across pursuits at all. It is not clear what inductive efficiency a citation metric could track in this case. In general, only very streamlined (inductively reduced) or mature pursuits pass the IA test. Perhaps the major challenge is to develop clear criteria for exploratory scientific pursuits and determine the inductive baseline in such cases, if there is one.

Inductive assessment introduces a substantial measure of transparency to operational analysis but, at the same time, puts substantial restrictions to it. Perhaps OA should not be applied prior to identifying methodological coherence of some sort within the domain of citation metrics application. This is a baseline constraint on OA that prevents spurious analysis and undesired side-effects stemming from lack of understanding of the operationally analysed pursuit: insofar as the IA of the pursuit is adequate, such effects are not likely to occur. OA will not suggest anything that will undermine or go against the methodology that made the pursuit successful in the first place.

The inherent inductive process behind the analysed domain guarantees the success of inductive analysis and, thus, ensures the transparency of the operational analysis. Besides a lack of desired inductive streamlining, however, other problems might occur when applying specific types of operational analysis. Not every type of operational analysis can reliably be applied on every data set. But various tests are available to assess how informative an operational analysis has been. For instance, data envelopment analysis, used to find efficiencies in multiple inputs and outputs, evaluates extreme points as efficient. To apply this type of operational analysis, we have to exclude the outliers from the data set. A sensitivity analysis can be conducted for this purpose (Ben-Gal 2005). Another case is the limitation of statistical methods, in particular, types I and II errors. Mayo-Spanos (2006) argues that when hypotheses that pass severe tests are used, these errors are minimized. It is important to note that if a specific OA proves inadequate for a given data set, a different one might apply, providing informative results.

Put otherwise, the approach we suggest does not disqualify other approaches but sets a standard against which they can be developed. For instance, the emerging use of simulations of scientific networks that seeks to empirically calibrate simulated models can introduce a level of inductive coherence in the analysis, improving its justification and transparency. Moreover, a possible cconvergence of different inductive analyses on the reliability of a specific research pursuit might argue in favour of a justified application of the operational analysis of a scientific network in which the pursuit is embedded.

References:

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. Philosophy of Science, 82(3), 424–453.

Allen, L., Brand, A., Scott, J., Altman, M., & Hlava, M. (2014). Credit where credit is due. *Nature*, *508*(7496), 312-313.

Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rockach (Eds.), Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers (pp. 131–146). Kluwer Academic Publishers/Springer.

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, 63(1), 87–120.

Borg, AM., Frey D., Šešelja D., & Straßer C. (2017). An Argumentative Agent-Based Model of Scientific Inquiry. In: *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I*. Ed. by Benfe.

Bornmann, L. (2017). Measuring impact in research evaluations: A thorough discussion of methods for, effects of, and problems with impact measurements. *Higher Education 73*(5), 775-787rhat S., Tabia K., & Ali M. Cham: Springer International Publishing, 507–510.

Braun, T. (2010), How to Improve the Use of Metrics, *Nature* 870-872.

Carillo, M.R., Papagni, E. and Sapio, A., 2013. Do collaborations enhance the high-quality output of scientific institutions? Evidence from the Italian Research Assessment Exercise. *The Journal of Socio-Economics*, *47*, pp.25-36.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, *3*(Nov), 507-554.

Corley, E.A., Boardman, P.C. and Bozeman, B., 2006. Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research policy*, *35*(7), pp.975-993.

Darriulat, P. (2004). The discovery of W and Z, a personal recollection. *European*

Physical Journal C, 34, 22–40.

Dissertori, G., Knowles, I.G, & Schmelling, M. (2003). *Quantum Chromodynamics: High Energy Experimetns and Theory*. Oxford: Clarendon Press.

Goodman S.N., Fanelli D., & Ioannidis J.P.A. (2016). What does research reproducibility mean? Science Translational Medicine, 341ps12.

Kelly, K. T. (2004). Justification as truth-finding efficiency: How Ockham's razor works. Minds and Machines 14(4), 485–505.

Kelly, K. T. (2007). A new solution to the puzzle of simplicity. Philosophy of Science 74(5), 561–573.

Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy*, *87*(1), 5-22.

Kocabas, S. (1991). Conflict resolution as discovery in particle physics. *Machine Learning*, *6*(3), 277-309.

Koonin, E. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res.* **5**, 1805 (2016).

MacRoberts, M. H., & MacRoberts, B. R. (1989) Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40, 342–349.

Martin, B. R., & Irvine, J. (1984a). CERN: Past performance and future prospects: I. CERN's position in world high-energy physics. Research Policy, 13(4), 183–210.

Martin, B. R., & Irvine, J. (1984*b*). CERN: Past performance and future prospects: III. CERN and the future of world high-energy physics. Research Policy, 13(4), 311–342.

Maruyama, K., Shimizu, H., and Nirei, M., 2015. Management of science, serendipity, and research performance: Evidence from scientists' survey in the US and Japan. *Research Policy* (44), pp. 862-873.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science, 70*(4), 647-670.

Rosenstock S., O'Connor C., & Bruner J. (2016). In Epistemic Networks, is Less Really More? Philosophy of Science.

Peltonen, T., 2016. *Organization Theory: Critical and Philosophical Engagements*. Emerald Group Publishing.

Prusiner, S. (1982). Novel proteinaceous infectious particles cause scrapie. Science 216, 136–144.

Pusztai, L., Hatzis, C., & Andre F. (2013). Reproducibility of research and preclinical validation: problems and solutions. Nature Reviews Clinical Oncology 10, 720–724.

Schulte, O. (2000). Inferring Conservation Laws in Particle Physics: A Case Study in the Problem of Induction. The British Journal for the Philosophy of Science, 51(4), 771-806.

Schulte, O., & Drew, M. S. (2010, October). Discovery of Conservation Laws via Matrix Search. In *Discovery Science* (pp. 236-250).

Soto, C. (2011)., Prion hypothesis: the end of the controversy? Trends Biochemical Sciences 36(3), 15–158.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial intelligence*, *46*(3), 259-310.

Ting, Samuel C. C. (1977). The discovery of the J particle: a personal recollection.

Reviews of Modern Physics, Vol. 49(No. 2), 235–249.

Valdés-Pérez R. E. and Żytkow J.M. (1996). A new theorem in particle physics enabled by machine discovery. *Artificial Intelligence*, *82*(1-2), 331-339.

van der Wal, R., Fischer, A., Marquiss, M., Redpath, S., & Wanless, S. (2009). Is bigger necessarily better for environmental research? *Scientometrics*, 78(2), 317–322.

Van Noorden, R. (2014). Transparency promised for vilified impact factor. *Nature News, Jul*, *29*, 2014.

Yang B., Wang Y., & Qian P.Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17: 135.

Yang Z., Rannala B. (2012). Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13, 303—314.

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science*, *26*(6), 453-459.

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science*, *76*(2), 225-252.92(1), 1–10.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. Erkenntnis, 72(1), 17–35.

Zur Hausen, H. (2009). The search for infectious causes of human cancers: Where and why. Virology 3.