UNIVERSITY OF BELGRADE

FACULTY OF PHILOSOPHY

Vlasta J. Sikimić

OPTIMIZATION OF SCIENTIFIC REASONING: A DATA-DRIVEN APPROACH

Doctoral Dissertation

Belgrade, 2019

UNIVERZITET U BEOGRADU

FILOZOFSKI FAKULTET

Vlasta J. Sikimić

OPTIMIZACIJA ZAKLJUČIVANJA U NAUCI: PRISTUP ZASNOVAN NA PODACIMA

doktorska disertacija

Beograd, 2019.

Supervisor:

Prof Dr Slobodan Perović
Associate Professor, Department of Philosophy, Faculty of Philosophy, University of Belgrade, Serbia

Members of the Committee:

Prof Dr Slobodan Perović
Associate Professor, Department of Philosophy, Faculty of Philosophy, University of Belgrade, Serbia

Prof Dr Kevin J.S. Zollman
Associate Professor, Department of Philosophy, Dietrich College of Humanities and Social Sciences, Carnegie Mellon University, USA

Dr Miloš Adžić
Assistant Professor, Department of Philosophy, Faculty of Philosophy, University of Belgrade, Serbia

Defense date:___May 14, 2019___

Mentor:

prof. dr Slobodan Perović
vanredni profesor, Odeljenje za filozofiju, Filozofski fakultet, Univerzitet u Beogradu, Srbija

Članovi komisije:

prof. dr Slobodan Perović
vanredni profesor, Odeljenje za filozofiju, Filozofski fakultet, Univerzitet u Beogradu, Srbija

prof. dr Kevin J.S. Zolman [Kevin J.S. Zollman]
vanredni profesor, Odeljenje za filozofiju, Ditrik koledž za humanistiku i društvene nauke, Karnegi Melon univerzitet, SAD

doc. dr Miloš Adžić
docent, Odeljenje za filozofiju, Filozofski fakultet, Univerzitet u Beogradu, Srbija

Datum odbrane: __14.5.2019.__

Abstract

Scientific reasoning represents complex argumentation patterns that eventually lead to scientific discoveries. Social epistemology of science provides a perspective on the scientific community as a whole and on its collective knowledge acquisition. Different techniques have been employed with the goal of maximization of scientific knowledge on the group level. These techniques include formal models and computer simulations of scientific reasoning and interaction. Still, these models have tested mainly abstract hypothetical scenarios. The present thesis instead presents data-driven approaches in social epistemology of science. A data-driven approach requires data collection and curation for its further usage, which can include creating empirically calibrated models and simulations of scientific inquiry, performing statistical analyses, or employing data-mining techniques and other procedures.

We present and analyze in detail three co-authored research projects on which the thesis' author was engaged during her PhD. The first project sought to identify optimal team composition in high energy physics laboratories using data-mining techniques. The results of this project are published in (Perović et al. 2016), and indicate that projects with smaller numbers of teams and team members outperform bigger ones. In the second project, we attempted to determine whether there is an epistemic saturation point in experimentation in high energy physics. The initial results from this project are published in (Sikimić et al. 2018). In the thesis, we expand on this topic by using computer simulations to test for biases that could induce scientists to invest in projects

beyond their epistemic saturation point. Finally, in previous examples of data-driven analyses, citations are used as a measure of epistemic efficiency of projects in high energy physics. In order to additionally justify and analyze the usage of this parameter in their data-driven research, in the third project Perović & Sikimić (under revision) analyzed and compared inductive patterns in experimental physics and biology with the reliability of citation records in these fields. They conclude that while citations are a relatively reliable measure of efficiency in high energy physics research, the same does not hold for the majority of research in experimental biology.

Additionally, contributions of the author that are for the first time published in this theses are: (a) an empirically calibrated model of scientific interaction of research groups in biology, (b) a case study of irregular argumentation patterns in some pathogen discoveries, and (c) an introductory discussion of the benefits and limitations of data-driven approaches to the social epistemology of science. Using computer simulations of an empirically calibrated model, we demonstrate that having several levels of hierarchy and division into smaller research sub-teams is epistemically beneficial for researchers in experimental biology. We also show that argumentation analysis in biology represents a good starting point for further data-driven analyses in the field. Finally, we conclude that a data-driven approach is informative and useful for science policy, but requires careful considerations about data collection, curation, and interpretation.

Keywords: *data-driven approach, optimization, scientific reasoning, social epistemology of science, formal models, empirical calibrations, high energy physics, experimental biology, inductive patterns*

Scientific discipline: **Philosophy**

Narrow scientific discipline: **Philosophy of Science**

UDC: 101.1

Sažetak

Zaključivanje u nauci ogleda se u složenim argumentativnim strukturama koje u krajnjoj instanci dovode do naučnih otkrića. Socijalna epistemologija nauke posmatra nauku iz perspektive celokupne naučne zajednice i bavi se kolektivnim sticanjem znanja. Različite tehnike su se primenjivale u cilju maksimizacije naučnog znanja na nivou grupe. Ove tehnike uključuju formalne modele i kompijuterske simulacije naučnog zaključivanja i interakcije. Ipak, ovi modeli su uglavnom testirali hipotetičke scenarije. Sa druge strane, ova disertacija predstavlja pristupe u socijalnoj epistemologiji nauke koji se zasnivaju na podacima. Pristup zasnovan na podacima podrazumeva prikupljanje podataka i njihovo sistematizovanje za dalju upotrebu. Ova upotreba podrazumeva empirijski kalibrirane modele i simulacije naučnog procesa, statističke analize, algoritme za obradu velikog broja podataka itd.

U tekstu predstavljamo i detaljno analiziramo tri koautorska istraživanja u kojima je autorka disertacije učestvovala tokom doktorskih studija. Prvo istraživanje imalo je za cilj da odredi optimalnu strukturu timova u laboratorijama fizike visokih energija koristeći algoritme za obradu velikog broja podataka. Rezultati ovog istraživanja su objavljeni u (Perović et al. 2016) i ukazuju na to da su projekti u koje je uključen manji broj timova i istraživača efikasniji od većih. U drugom istraživanju smo pokušali da utvrdimo da li postoji tačka epistemičkog zasićenja, kada su u pitanju eksperimenti u fizici visokih energija. Inicijalni rezultati ovog istraživanja objavljeni su u (Sikimić et al. 2018). U disertaciji produbljujemo ovu temu korišćenjem kompjuterskih simulacija da

bismo testirali mehanizme pristrasnosti koji navode naučnike da ulažu u projekte iznad tačke epistemičkog zasićenja. Konačno, u prethodnim primerima analiza zasnovanih na podacima, citiranost je korišćena kao mera epistemičke efikasnosti pojekata u fizici visokih energija. Da bi dodatno opravdali upotrebu ovog parametra u svojim analizama, u trećem istraživanju Perović & Sikimić (under revision) su razmatrali i upoređivali induktivne šematizme u eksperimentalnoj fizici i biologiji sa pouzdanošću mere citiranosti u ovim oblastima. Zaključili su da, iako su citati relativno pouzdana mera efikasnosti u fizici visokih energija, to nije slučaj u najvećem delu istraživanja u oblasti eksperimentalne biologije.

Povrh toga, doprinosi autorke koji su prvi put objavljeni u ovoj disertaciji jesu: (a) empirijski kalibrirani model naučne komunikacije unutar istraživačkih grupa u biologiji, (b) analiza neočekivanih argumentativnih struktura u otkrićima nekih patogena i (c) uvodna diskusija u pogledu prednosti i ograničenja pristupa zasnovanih na podacima u socijalnoj epistemologiji nauke. Korišćenjem kompjuterskih simulacija na empirijski kalibriranim modelima, pokazujemo da je raslojavanje i podela na manje istraživačke timove epistemički korisno za istraživače u eksperimentalnoj biologiji. Takođe, pokazujemo da je analiza argumenata u biologiji dobra osnova za dalje analize zasnovane na podacima u ovoj oblasti. Na kraju, zaključujemo da je pristup zasnovan na podacima informativan i koristan za kreiranje naučne politike, ali da zahteva pažljiva razmatranja u pogledu prikupljanja podataka, njihovog sortiranja i interpretiranja.

Ključne reči: *pristup zasnovan na podacima, optimizacija, zaključivanje u nauci, socijalna epistemologija nauke, formalni modeli, empirijske kalibracije, fizika visokih energija, eksperimentalna biologija, induktivni šematizam*

Naučna oblast: **filozofija**

Uža naučna oblast: **filozofija nauke**

UDK: 101.1

Acknowledgments

First and foremost, I wish to thank my supervisor, Slobodan Perović, for his guidance during my PhD and for the time, knowledge and energy spent on the successful research projects in which we participated together.

I am grateful to Kevin Zollman for sharing his insights during the excellent intensive course held in Belgrade in spring 2018. I am especially thankful to him and Miloš Adžić for agreeing to be members of the dissertation committee and finding time to read and comment on my thesis. During my PhD I was lucky to collaborate with Sandro Radovanović, Kaja Damnjanović and Andrea Berber, and to exchange ideas with Milan Z. Jovanović. I would also like to thank Dunja Šešelja and Christian Straßer for opening their doors for my research endeavor, supporting my academic growth, and for the current collaboration. I am thankful to Živan Lazović, head of the project "Dynamic Systems in Nature and Society: Philosophical and Empirical Aspects", which I am part of.

I wish to thank Sonja Smets and Alexandru Baltag for their constant encouragement in my scientific aspirations. I am grateful to Peter Schroeder-Heister for all his kindness and valuable advice he has provided me during my research stay in Tübingen in 2015. I am indebted to Kosta Došen, who first welcomed me in the academic world and from whom I learned what it means to be a scientist.

I have to thank all my close friends who kept cheering for me during this period. Finally, I owe most to my father, Jovan, and to my husband, Ole, who supported me in every imaginable way during the turbulent PhD period: you mean the world to me.

To my grandparents and heroes, Danica and Čedo, for everything they did for me

Contents

Chapter 1

Introduction

Social epistemology of science studies group knowledge acquisition and aims to understand how this process can be optimized. This contrasts with traditional epistemology, which focuses on individuals' knowledge acquisition (Goldman & Blanchard 2016). This new focus on group knowledge acquisition is accompanied by a shift in the philosophical perspective. New topics arise from it, including the analysis of information exchange during deliberation and other democratic processes or on the Internet and social media.

The central hypothesis in social epistemology of science is that members of a group all together can know more than any individual member of it. This assumption might be counterintuitive, as we can see that a group of runners will always be slower than the best runner (Surowiecki 2004), but it becomes clearer when we examine scenarios in which more complex goals are involved. For example, at Tour de France, a group exploiting "drafting" will be substantially faster than an individual cyclist, because of the superior aerodynamic properties of this group formation (McCole et al. 1990). Similarly, when it comes to knowledge acquisition, groups can profit from different backgrounds, viewpoints, approaches, and skills of their members. The distributed knowledge of a group – i.e., the potential knowledge that a group can reach when all members share their private knowledge in the sense of collecting different pieces of the puzzle – is

therefore usually greater than the knowledge of the most knowledgeable member (Baltag et al. 2013). Another example of this effect is the phenomenon called *the wisdom of crowds* (Surowiecki 2004): aggregated knowledge of a large group of agents can outperform expert knowledge. However, the wisdom of crowds is a phenomenon that occurs only when certain conditions are met, such as decentralization and independence of the agents (Surowiecki 2004). Each agent should have prior knowledge that she acquired independently and that she was able to expand on individually. After an adequate aggregation procedure, a large group composed of such agents will outperform expert knowledge (Surowiecki 2004).

The shift from individual to group knowledge acquisition also requires a different approach than the "armchair" one associated with traditional epistemology. To study the process of group knowledge acquisition and its optimization, we argue for an interdisciplinary approach, combining the expertise of philosophers, computer scientists, and psychologists, among others. In the next section (1.1), we will first discuss the topics studied in social epistemology in general and then introduce social epistemology of science. We will present the contemporary approaches to social epistemology of science, their benefits and limitations, and future directions in the field. These new trends in philosophy require interdisciplinary approaches, enriching philosophical reasoning with ideas from logic, computer science, and psychology. They have the potential to increase the scope of philosophically relevant topics, and thus to increase the reach (i.e., applicability) of the field of philosophy (Sikimić 2017).

1.1 Topics in social epistemology

Some of the aspects of collective knowledge production that social epistemology studies are the evaluation of evidence from the perspective of a group and the analysis of epistemic systems. For instance, it is concerned with questions of how to evaluate

testimonies of others and how complex epistemic systems such as the citizens in a democracy or a scientific community reason (Goldman & Blanchard 2016). While traditional epistemology focuses on knowledge acquired by individual senses, inferences, or introspection, social epistemology is concerned with questions of how individuals interact and evaluate each other when acquiring knowledge (Sikimić 2017). For instance, *how do we recognize some people as experts?* Or, *how do situations arise in which different experts are forming different opinions based on the same evidence?*

Social epistemology also studies group knowledge acquisition and its maximization: how group beliefs and knowledge are formed in an optimal way. As mentioned above, the underlying hypothesis of social epistemology is that the knowledge of a group is usually greater than the knowledge of any individual member, as different group members have different private information (Surowiecki 2004). This reasoning is supported by another observation in both academia and industry: diverse groups tend to outperform more uniform ones (Powell 2018, Hunt et al. 2018).

Theoretical considerations and experience show that group knowledge often does not reach its maximum potential in practice; handbooks on encouraging participation in meetings and sharing viewpoints can be found in countless workplaces. One example of suboptimal information aggregation is the phenomenon of an information cascade: individuals ignore their private knowledge because of the beliefs of others. When we choose one of two restaurants only because it is popular, we might be affected by an information cascade. This effect can be frequently observed in tourist destinations. Even in disciplines guided by science, such as medicine, we can observe information cascades. One frequently used example is unnecessary tonsillectomy – the removal of the tonsils often performed in the middle of the 20^{th} century in the United Kingdom (Bikhchandani et al. 1992), and elsewhere. Although the benefit for patients had not been scientifically confirmed, tonsillectomies were very common. Still, already at the time, experts in the field considered the procedure to be unnecessary. Thus, the high frequency of opera-

tions indicates that they were performed based on an erroneous belief set. Moreover, the frequency of tonsillectomies varied between regions, further supporting the view that many procedures were performed because of information cascades among doctors in certain regions (Bikhchandani et al. 1992).

To develop strategies against such inference mistakes, one first has to understand the communication patterns between group members that resulted in several groups forming opinions which contradicted the view of the medical authorities. We must consider both the private evidence and the evidence from medical authorities presented to the group, with the understanding that the group's prediction (i.e., judgment) would then be made after all the evidence had been presented and discussed within the group (Baltag et al. 2013).

With the advent of instant messaging and social media, the effect of information cascades has become a major topic of public debate, and developing ways to avoid them is now an important topic in interdisciplinary research (Hendricks & Hansen 2014, 2016). For instance, in an attempt to prevent the spread of wrong beliefs among subgroups, the popular instant messenger WhatsApp restricted the forwarding of messages in India after the viral spread of fake news resulted in several cases of lynching (Ryan 2018).

A similar fallacy in group belief aggregation has been analyzed by Hartmann & Rafiee Rad (2018). They observe that a specific type of anchoring occurs in groups formed of individually rational agents. The standard anchoring phenomenon occurs when an individual sets all her beliefs close to her initial belief (the anchoring belief). However, Hartmann & Rafiee Rad (2018) argue that in the context of group deliberation, the first speaker has the highest impact on the beliefs of all other participants in the decision-making process.

Another topic studied by contemporary social epistemology is the dynamics of knowledge exchange within complex systems. This topic concerns the fluctuations of beliefs in groups such as Internet users exchanging information, voters in a democracy,

or scientists forming dominant research hypotheses. The last is mainly studied by social epistemology of science.

The central topics of social epistemology of science include (among others) the division of labor among scientists, their communication practices, and their mechanisms of knowledge aggregation. One of the field's central findings concerns scientific communication: interestingly, limiting communication among scientists has been shown to be epistemically beneficial (Zollman 2007, 2010). However, funding incentives and other rewards systems in scientific research strongly discourage limiting the information flow between scientists. Scientific communities sometimes form erroneous beliefs, and, because those beliefs easily spread through a well-connected scientific network, the scientists entrench themselves. In order to overcome this epistemic dead end, outsiders with rogue ideas are helpful (Kitcher 1990).

As an attempt to address topics in social epistemology of science using a data-driven method, Perović et al. (2016) used data envelopment analysis to determine which tendencies influence epistemic efficiency in high energy physics. Their results show that smaller teams outperform big ones and that the total number of researchers in each team should be relatively small. We analyze in detail the results of this research in chapter 3. This interdisciplinary research paves the way for further studies that will target philosophically-flavored questions with data collected in the field. Such an endeavor can include joint efforts of data scientists, philosophers, computer scientists, psychologists, and others.

1.2 Motivation

There are different approaches in social epistemology of science. Some involve agent-based models, e.g., (Borg et al. 2017, Grim 2009, Zollman 2007, 2010), or more traditional philosophical analysis based on case studies, e.g., (Šešelja & Straßer 2014,

Vickers 2014). However, questions about optimal division of labor or communication structures in science are also situation-dependent questions and can be answered using a data-driven analysis. As a broad term "science of science" stands for the transdisciplinary research that treats science as its topic (Ossowska & Ossowski 1964). This research is largely based on data. For instance, in sociology of science, data-driven analyses are frequent, e.g., (McLevey & McIlroy-Young 2017, Mutz et al. 2017). We refer to abstract models as *hypothesis-driven* and to the ones based on actual data from a scientific field as *data-driven*.

Abstract models can be enriched with data and empirically calibrated to give answers to more precise questions, while data-mining techniques should be applied carefully and with previous theoretical considerations. For instance, citation metrics are not equally reliable measures of scientific performance in every discipline. This thesis represents an attempt to bridge the gap between these two approaches. In particular, in chapter 6 we present an empirically calibrated model for comparing efficiencies of different team structures in experimental biology. The data-driven research in social epistemology of science comes with a smaller degree of abstraction, which in turn provides a clear and straightforward domain for the application of the results. The interpretation of the research is defined by the data that are used for calibration. In this case, we used data from qualitative interviews with biologists and analysis of their existing team structures to model realistic division of labor in biological laboratories. When we consider that professors have limited time for communication, the results after 1000 simulations show that groups with additional levels of hierarchy perform much better than centralized groups. The performance further improves when group leaders communicate with each other. The results also indicate that group leaders should not have too many students, and that large groups should be decentralized. These findings match the assumptions interviewees brought up.

Apart from empirically calibrated models, another important data-driven technique

that can be used in social epistemology of science is data-mining based on different machine learning algorithms and statistical analyses. In chapter 4, we present results from the statistical analysis of 49 projects from Fermilab conducted by Sikimić et al. (2018). They analyzed whether there is an epistemic saturation point after which further investments in experiments in high energy physics will not deliver further important results. This analysis, along with the analysis from (Perović et al. 2016), is based on citation metrics. Citations cannot always and unreservedly be used as a measure of project impact. For instance, in social sciences the replication problem is correlated with the fluctuation of the impact of publications over time (Baker 2015). Perović & Sikimić (under revision) argue that fields such as high energy physics and phylogenetics exhibit relatively regular inductive behaviors, which is later reflected in their citation patterns. Thus, citation metrics represent a good proxy for measuring scientific performance in these fields. However, they also point out that other areas of research such as plant biology and the study of pathogenesis do not exhibit these features and are not the best candidates for citation-based analyses. We address this problem in chapters 3 and 5.

1.3 Structure of the thesis and key points

The thesis is structured as follows.

Introduction

In the introduction, the research subject is situated in the field of social epistemology of science. Also, we discuss the existing approaches to the optimization of scientific reasoning. This chapter is partially based on (Sikimić 2017).

Optimization of scientific reasoning

In this chapter, we analyze different formal approaches to the optimization of resources

in science. The goal of this chapter is the evaluation and classification of existing formal approaches in social epistemology of science. The main classification criterion is the demarcation between hypothesis- and data-driven approaches. Moreover, we discuss general topics related to the application of data in science, such as data journeys and ways of collecting data.

Optimization of resources within a scientific project: the case of high energy physics

The goal of this chapter is to show different methods of data-driven analysis, such as Data Envelopment Analysis (DEA), applied to specific cases in high energy physics. In particular, we analyze results that demonstrated that smaller teams are more efficient than large ones. Also, we discuss reasons why contemporary experimental physics is particularly suitable for operational analysis, while other subjects might not be. In this chapter results from the co-authored research from (Perović et al. 2016) and (Perović & Sikimić under revision) are presented and discussed.

Investments in HEP and the halting problem

We explore whether there is the epistemic saturation point in science, after which further research on the same project will most likely not be fruitful. In this chapter, we will discuss further the results from the co-authored research published in (Sikimić et al. 2018).

Optimization of resources within a scientific project: the case of experimental biology

In this chapter, we analyze whether biology can be suitable for citation-based analysis and which other types of data-driven analysis can be used in order to maximize epistemic efficiency in the field. We argue that in the sub-fields that exhibit relatively regular behaviors, such as phylogenetics, the use of citation metrics can be justified,

while this is not the case for the majority of research in experimental biology. This chapter is partially based on the results from (Perović & Sikimić under revision).

Empirically calibrated agent-based models

This chapter discusses the potential of empirically calibrated agent-based models and presents a data-driven model of optimal divisions of labor in experimental biology. Team structures in science are field-dependent. In biology, laboratories are typically structured hierarchically. We developed models simulating three different management styles, ranging from groups with one leader controlling everybody to groups with two levels of hierarchy. In qualitative interviews performed with biologists, participants brought up and discussed these structures and their effects on group performance. The groups with several layers of hierarchy outperformed centralized ones. In this chapter, we present unpublished results of the author.

Argumentation patterns in life science: a study of pathogen discoveries

Scientific argumentation, as a fine-grained type of argumentation, is a fruitful ground for formal analyses of different reasoning strategies. In this chapter, the analysis of argumentation that led to discoveries of disease-causing mechanisms in life science is examined. This chapter is based on unpublished research by the author.

Benefits and limitations of data-driven approaches

Data-driven approaches are a powerful tool for the investigation of the process of scientific inquiry. There are different ways of collecting data about scientific interaction, values in science, scientific practice, and so forth. In this chapter, we compare data collecting using qualitative, quantitative, and mixed methods, where the mixed one is clearly most informative.

The other big challenge for data-driven approaches is the systematization of the

data. For this purpose, a field-specific approach has to be employed. After addressing the problem of data collection and their systematization, we turn to the normative aspect of public availability of data from scientific projects. Public data availability is necessary for the ideal of open science, but one must take care to protect scientists' privacy.

Conclusions and further research

In the final chapter, we summarize the results of this thesis and point towards further research topics. In particular, we address the Optimist platform, which is the result of a collaboration between scientists from the University of Belgrade, Serbia and the Ruhr-University Bochum, Germany. The acronym stands for "Optimization Methods in Science and Technology" (Optimist 2018).[1] The first effort of the collaboration was gathering data about job satisfaction in physics laboratories via a survey.

[1]More information is available on the following link: <http://www.ruhr-uni-bochum.de/optimist-survey/>.

Chapter 2

Optimization of scientific reasoning

Formal approaches to the optimization of scientific reasoning include 1) hypothesis-driven analyses, such as agent-based simulations, e.g. (Übler & Hartmann 2016, Šešelja & Straßer 2013, Zollman 2007, 2010), 2) logical models, e.g. (Baltag & Smets 2011), and 3) data-driven modeling, e.g. (Irvine & Martin 1984b, Martin & Irvine 1984,b, Perović et al. 2016). Hypothesis-driven models make predictions about the consequences of different epistemic behaviors of scientists; although they do not use empirical datasets, they are meant to model relevant phenomena of scientific interaction. Logical models also have explanatory power and make reasoning patterns of the scientific community explicit; they describe optimal and suboptimal processes for updating and aggregating beliefs. Data-driven techniques have been developed in sociology of science. This approach makes use of real-life data to identify patterns and make predictions, e.g., (McLevey & McIlroy-Young 2017, Mutz et al. 2017). Finally, hypothesis-driven models can benefit from empirical calibrations, in the sense that it becomes clearer whether they model a specific scientific environment. In this way, a data-driven and hypothesis-driven approach are combined.

2.1 Hypothesis-driven approaches

The social epistemological perspective about which research hypothesis to pursue does not always agree with the perspective of the individual researcher. For the individual's research career, it is often best to explore the most probable hypothesis because this gives one the highest chance to publish one's results, even considering that one might have to publish them in lower-impact journals (Sikimić 2017). It is important to note that some people are excited to pursue bold hypotheses for intrinsic reasons, or because they desire a more influential publication, etc. However, for the scientific community as a whole, it is desirable to have a fraction of researchers pursuing less likely and thus riskier hypotheses, as demonstrated with examples from history of science and probability models (Kitcher 1990, 1993) and computer simulations (Strevens 2003, Weisberg & Muldoon 2009). The listed authors claimed that novel, less probable ideas can result in very influential research; they argued that it could be beneficial for the formation of the correct belief that scientists work independently of the ideas of their peers. It should be noted that the methods of Weisberg & Muldoon (2009) are questioned by other research groups. Alexander et al. (2015) argued that their conclusions were based on a poor implementation of the rules governing the agents in the simulations.

The assumption that groups of scientists perusing diverse hypotheses outperform those working in accordance with existing tendencies stands in contrast to one of the most relevant reward systems in science. Journals follow the same trends as the scientific community, they mainly publish research within the most cited subfields. This strongly motivates scientists to direct their research in the same direction, reducing the diversity of approaches and harming science as a whole. On the other hand, from the community perspective, it can be beneficial if the scientific workforce is divided into two fractions: a larger group working on the most probable scenarios, and a smaller group exploring new areas and less likely hypotheses. Interestingly, some funding agencies do follow this approach – for example, the European Research Council requires project proposals to

contain both a relatively safe part and a highly ambitious one. The safe part should most likely be answered within the project duration, and should also result in some publication for the involved scientists. The second, more novel and risky part is not necessarily expected to be successful. Although these novel parts do not produce the desired results as frequently, they tend to be much more influential and much more widely published, justifying the invested resources.

In addition to the research direction that the members of the scientific community take, their communication is a very important part of the scientific process. This communication affects not only the speed at which scientists reach a consensus, but also group knowledge acquisition. Zollman (2007, 2010) used computer simulations to study the relationship between the time needed to reach a scientific consensus and the accuracy of the conclusions. As expected, a well-connected group reaches the consensus faster than a less connected one, but this comes with a trade-off: strongly connected networks marginalize minority positions and thus do not necessarily reach the optimal decision. In contrast, groups with a smaller degree of connectivity are more successful at exploring all possible hypotheses and are more likely to reach the optimal group knowledge. These results favor the proposal to limit the communication between scientists.[1] Moreover, Kummerfeld & Zollman (2015) claim that scientists should sometimes get external incentives to explore risky hypotheses.

All these and many more studies on the social epistemology of science are based on computer simulations and logical models, e.g., see (Baltag et al. 2013, Kelly & Mayo-Wilson 2010). However, they are mostly not data-driven but rely on idealized assumptions. In this thesis, we will follow a new tendency and use real-world data to optimize scientific resource distribution and maximize group knowledge acquisition. While general and theoretical approaches result in good basic conclusions that should be

[1] It should be noted that Rosenstock et al. (2017) argued that a decreased connectivity is only beneficial under very specific circumstances.

considered when research groups are structured (Kitcher 1990, Kozlowski & Bell 2003, Zollman 2007), these general arguments can be adapted for the specific requirements of a scientific field. For instance, in chapter 6 we make an attempt in this direction and present models based on data about realistic group structures in HEP and experimental biology. Data-driven approaches usually use data from a specific field and account for its particularities. While theoretical models can only provide evidence for or against the hypotheses considered when the model was designed, real-world data might not only point out that these hypotheses are faulty or insufficient but can also highlight less understood properties. One further advantage concerns the applicability of the results: because data-driven models use data from a specific field, they can be immediately and unambiguously applied to this field. This is important if we want to improve the efficiency of scientific knowledge acquisition and promote excellent research in times of budget constraints and beyond (Sikimić 2017).

2.2 Data-driven approaches

We use data-driven approaches to identify efficient projects and make predictions about the worthiness of project proposals. In general, when we think about scientific resources, we mainly consider financial, time, and human investments in a research project.

One prominent example of a data-driven analysis that prompted a major shift in funding schemes was the one conducted by Michael Lauer, the National Institutes of Health's (NIH)[2] deputy director for extramural research, and his colleagues (Lauer et al. 2017). They used data about their grants, awarded to more than 70 000 principal investigators over almost 20 years, and showed that the efficiency dropped when a laboratory was holding three or more standard grants. Considering that scientific productivity fluctuates during one's career (Way et al. 2017), the NIH decided to redirect

[2]The NIH, with its about 30 billion dollar budget, is one of the largest biomedical funding organization in the world.

3% of its budget, about 1 billion dollars, from the biggest laboratories to early-career researchers in order to increase the total number of researchers being funded (Collins 2017). Similarly, new research policies in South Korea redirect some money from large governmental research projects to individual researchers (Yeom 2018). Even though such changes might face powerful opposition from the established scientists leading those large-scale projects, policy makers have begun to understand that supporting individual researchers and small labs and projects is usually more efficient (Cook et al. 2015, Yeom 2018).

To analyze the social epistemology of science, a variety of approaches have been employed, including different statistical tools and machine-learning algorithms (Kelly 2004, Schulte 2000, Thagard 1988). Recently, Perović et al. (2016) published their results of a Data Envelopment Analysis (DEA) based on data from the high energy physics laboratory Fermilab. DEA was developed to enable theoreticians and policy makers to measure production efficiencies without needing to know all possible input and output combinations. It was first used to assess agricultural productivity (Farrell 1957) and has since been applied to many different industries, including banking, health care, science, and education. However, to our best knowledge, the analysis of project efficiencies in high energy physics by Perović et al. (2016) is the first application of DEA to social epistemology of science. We believe that this approach can be applied to many more scientific fields and institutions. This highlights the need for philosophers to collaborate in interdisciplinary research teams to analyze scientific projects with tools from computer science and psychology.

When we try to understand the output of research projects, we can consider different measures such as the number of trained researchers, granted patents or publications. For instance, Perović et al. (2016) focused on the number of citations per paper. As input parameters Perović et al. (2016) used basic properties of a research collaboration: the number of researchers, the number of research teams involved, and the project

duration. While the funding of a project is mirrored in the number of researchers and the invested time, the number of research teams is relevant for the interactions of the researchers with each other (i.e., the network structure of the collaboration). While researchers in one laboratory typically interact closely with each other, the interactions between researchers from different groups and locations are less frequent. In addition, Sikimić et al. (2018) studied the project duration in HEP to determine when to stop an unsuccessful experiment. This external analysis should be particularly useful, as researchers frequently find it hard to stop projects in which they have invested much already (Arkes & Blumer 1985). Sikimić et al. (2018) used the accessible data of the HEP laboratory Fermilab to show how the chance of achieving the desired goal (e.g., publishing one influential paper) decreases over time. Furthermore, in section 3.2 we will discuss why the specific context of HEP laboratories is particularly suitable for this assessment.

2.3 Data journeys

Rich information about all accepted and declined proposals requesting the use of the accelerators hosted by Fermilab is available on the high energy physics repository INSPIRE-HEP (https://inspirehep.net/). Perović et al. (2016) used this data source to identify efficient and inefficient experiments conducted in Fermilab. They showed that smaller teams tend to be more efficient, providing further support for the policy of funding more junior researchers and their small laboratories. With this data-driven analysis of projects in high energy physics, Perović et al. (2016) provided a framework that can be considered in funding decisions to increase the scientific output. The results from (Perović et al. 2016) are also supported by studies in other disciplines. As we mentioned above, Cook et al. (2015) showed that small life science laboratories outperform large ones, unless the budget for financing postdocs and PhD students is too limited.

Real-world data are never complete or error-free. To use them in a DEA, we must first analyze them carefully. For example, when we obtain data from the INSPIRE-HEP repository we must consider its historical context, because the data was collected over several decades. Therefore, some standards, such as the number of authors on a publication list, have changed over time (Birnholtz 2008). In addition, the value of academic titles has changed – in the past, seniority was not always associated with a PhD title.

These data are not only valuable for researchers working on similar topics, but also in ways that were not originally intended – for instance, for the research in "science of science". Unfortunately, in many cases only a very small data portion is made available, typically in the form of a publication containing the key findings. Moreover, these publications are frequently protected by paywalls, preventing other researchers from accessing the findings. How much data is published in databases, or whether only the conclusions are made available, varies strongly across fields (Gentil-Beccot et al. 2009).

The second big problem concerns the curation of the data. Only when the data are sorted and searchable they can be useful for other researchers. For example, genomic data about sequenced organisms are usually made available on specialized websites; there researchers can search for genes of many species, analyze their relationships, and study their functions. These websites function as hubs, collecting data from many different institutions, curating them, and making them freely available. However, in many other fields data are not made accessible for various reasons. The most important ones preventing researchers from making their data available include considerations about competition, and a lack of motivation and resources to curate and store the data. To improve this situation, Borgman (2015) argues that all stakeholders must work together and promote open databases as the standard. This can only improve if funding agencies encourage and support the open databases and if the publication of databases is

associated with career benefits. Long-term funding schemes are necessary to provide support for the IT infrastructure and positions for data curators. Furthermore, researchers sometimes prefer to keep large datasets for themselves as long as they hope to find useful information in them. This would only change if they can be sure that their hard work in generating this data is properly acknowledged. One additional problem concerns data privacy. For example, the Chinese government enforces very strict rules on the publication and sharing of genomic data. Research institutions and companies sharing data without the agreement of the government are punished and temporarily banned from international collaborations that use human genetic resources (Cyranoski 2018). This policy, while understandable from the perspective of privacy advocates, limits the accessibility of valuable databases generated by Chinese institutions.

Even though the availability of all research proposals, protocols, and positive and negative results is desirable from the perspective of social epistemology of science, it also requires a big shift in information exchange and publishing policies. The first step could be the publication of articles on open access platforms. In Europe, major funding agencies have created an initiative to enforce open access. Partners include the European Research Council and the national funding agencies of Austria, Finland, Ireland, the Netherlands, Norway, Poland, and several other countries. They will demand that from 2020 onwards, all research that they support gets published in open access journals or on open access platforms. They will also prohibit embargo periods (open access after a certain time, e.g., one year) and hybrid models (publishing in journals that publish only certain articles open access).[3] This initiative has the potential to change the business models of scientific publishers and make open access the standard. However, right now, different aspects prevent the publication of articles in open access journals, including publication fees, quality problems, and psychological aspects, such

[3]More information about the funders and their principles can be found on https://www.coalition-s.org.

as peer pressure and individual motivation. Weckowska et al. (2017) analyzed the psychological aspects and suggested a targeted approach to tackle the problems preventing researchers from publishing in open access journals. They argue that only scientists with strong motivation will go against the stream and publish in open access journals unless publication policies change.

In physics, open access is very common, and almost all papers are made available on open access servers, such as arXiv (Gentil-Beccot et al. 2009). This approach also manifests itself in online repositories like INSPIRE-HEP, which make external data on experiments in high energy physics available. Collecting these data is an interdisciplinary endeavor: data is provided by physicists, curated and archived by humanities scholars, and made available for many different questions, including science policy studies. It is a great example of how open data repositories can help answer questions not originally intended by the researchers who generated the data.

2.4 Understanding the optimal team structure

How research projects should be structured to perform optimally is a question that funding bodies should consider when making decisions about science policy. Also, research institutions and department heads should examine the advantages and disadvantages associated with the team structure and internal communication among scientists. Some aspects, like the value of interdisciplinary research groups and a healthy proportion of junior to senior researchers, are supported by diverse funding organizations and institute bodies, e.g., (Rylance 2015). The question of how researchers within projects should be structured has also been investigated within social epistemology of science, psychology of science, and science policy studies (Kozlowski & Bell 2003, Milojević 2014, Olson et al. 2007). Two contradictory effects about epistemic efficiencies in science have been singled out. On the one hand, as discussed in section 2.1, a high degree of independence

among researchers and a high number of researchers is considered beneficial (Jackson 1996). When a high diversity of ideas is encouraged and supported by weaker interaction between researchers, innovative ideas have the chance to flourish and challenge the mainstream. On the other hand, large and less connected teams are difficult to manage; they easily suffer from communication problems and the absence of a common goal.

How much independence is necessary to support the exploration of new ideas, and how much control is needed to ensure the proper function of a laboratory, mainly depends on the question the specific scientific project seeks to answer. A project that tests a set of hypotheses requires a vastly different organization than a project designed to thoroughly explore a certain natural phenomenon. Therefore, we can employ various techniques and approaches, and different levels of abstraction, to assess the epistemic efficiency of various projects and the factors affecting the efficiency (Seijts & Latham 2000).

Many modern scientific studies require diverse competencies and thereby a large number of scientists working on them. This results in large laboratories or collaborations. In high energy physics, laboratories often consist of hundreds of members with thousands of collaboration partners. In the most extreme case, the 2015 paper estimating the mass of the Higgs Boson, this resulted in a paper stating more than 5 000 authors – the author list and their affiliations is longer than the rest of the paper (Aad et al. 2015). How teams of increasing size can be managed and structured efficiently is a very important and challenging question in science policy studies. Carillo et al. (2013), Katz (1982), and Von Tunzelmann et al. (2003) studied how the increased size of research institutions affects scientific performance.

The efficient structures of collaborative teams are also frequently discussed in management studies. In fact, many consulting companies sell the development and application of new structures to their customers. To quantify efficiency, the industry has

developed various measures, typically focusing on productivity. Increasing productivity is one of the main goals of industrial companies, especially in saturated markets. By contrast, in science- and research-driven businesses (like parts of the pharmaceutical industry), the main goal is the acquisition of knowledge, which cannot be measured as easily as the output of a production line.

To study the effect of different sizes and structures of research groups, one can use not only techniques like computer simulations and logical models, but also data-driven analyses. Data-driven approaches assess the epistemic efficiency of group knowledge acquisition at various levels of abstraction and are characterized by specific advantages and disadvantages as outlined in chapter 3. If the necessary data is available and the field satisfies the conditions, DEA can be used to produce immediately applicable recommendations.

2.5 Psychology and social epistemology

Several psychological mechanisms influence the efficiency of research projects. These include factors affecting the individual, such as the sunk cost bias which influences the time spent on futile research projects (Arkes & Blumer 1985), and group phenomena, such as the principle of compliance with authority which makes scientists vulnerable to incorrect beliefs (Cialdini 2001). We will focus on how individuals react to resource allocation, when to stop unsuccessful experiments, and what the optimal structure of research teams is.

The first problem, which relates to the number of researchers involved in a project, is the optimization of resources invested in a project. Even though new expensive technologies might be needed to answer certain questions, they are never sufficient, nor does the use of expensive technology necessarily correlate with the quality of the research. Still, judgmental heuristics sometimes link expensive to superior and inexpensive to

inferior. This psychological effect is frequently abused in marketing, when items are priced higher because consumers evaluate more expensive products more positively. A similar effect also influences our judgment when we trust statements more because we hear them from an expert (Cialdini 2001). In science, we observe a similar effect: judgmental heuristics produce a better evaluation of expensive equipment, as opposed to smarter work-arounds. Good examples of this judgment can be observed in life sciences, when expensive technologies (like sequencing) are used to answer questions accessible to technologies for a fraction of the cost because they can be published in higher-impact journals. To protect us from such misjudgments, a comparative analysis of similar projects is very useful. When we include the invested resources in our consideration, we can establish when a cheaper approach is sufficient and presents a superior epistemic approach.

The second problem, and the one that strongly affects the involved scientist, is when to stop an unsuccessful experiment (Sikimić 2017). To stop an experiment, or any investment, before the desired goal is reached violates the psychological principle of commitment and consistency. After we have made a commitment (for example, starting out in a particular research direction), we tend to continue in the same direction. Even when we acquire novel evidence that would have prevented us from going in this direction in the first place, we tend to continue in the direction of our original beliefs (Cialdini 2001). Another psychological fallacy is the already mentioned sunk cost bias (Arkes & Blumer 1985). For instance, there is no reason to believe that the prior value of any stock will influence its future value (the prior value is public knowledge, and thus is included in the value at the stock market); despite this fact, we tend to keep stocks that have lost value after we bought them because we do not want to write off our prior investment. Instead, we consider this irrecoverable investment as a part of our decision. In science, these psychological principles lead to similarly biased investments of financial and human resources. With the help of statistical tools, we can identify efficient and

inefficient experiments and define a point of diminishing returns, e.g., (Sikimić et al. 2018). This method provides us with an objective tool with which we can decide to stop an unsuccessful experiment and identify the best strategy to reach our scientific goal.

The last problem, the optimal structure of a research group, is especially interesting, because every principal investigator and every institution can influence it rather easily (Sikimić 2017). From a global perspective, the group structure is mainly expressed as the number of groups, the communication structure within these groups, and the communication between groups. An optimal structure can help to reduce powerful social effects such as the one imposed by the principle of compliance with authority (i.e., the observation that group members tend to comply with the incorrect beliefs of an authority, even when they have evidence disagreeing with this authority (Cialdini 2001)). In addition, information cascades can be avoided by a well-structured group. When we incorporate all accessible data, such as the number of researchers involved, the ratio of senior and junior group members, and the number of groups, we can use a network DEA (a DEA analysis that accounts for the structure of the parameters) to identify optimal values. Alternatively, we can develop an empirically calibrated model to simulate realistic team structures in science and evaluate their efficiencies. We present such models in chapter 6.

2.6 Summary

We discussed how social epistemology could help us make the scientific endeavor more efficient. We discussed hypothesis- and data-driven models used in social epistemology of science and their advantages and disadvantages. We highlighted the benefits of formal models and computer simulations for group knowledge maximization: they can help to clarify group dynamics and identify phenomena influencing group decisions. Finally,

we showed how data-driven approaches could help us identify factors associated with efficient spending and thereby shape how funding agencies such as the NIH distribute their resources.

One of the main obstacles for data-driven approaches is the availability of the necessary data. We analyzed factors promoting and impeding their availability and discussed open data as an aspect of open science, which requires the free availability of scientific data, protocols, and publications. Furthermore, we showed how psychological aspects influence scientists' decisions and communication, as well as how optimal research team structures can help to overcome obstacles caused by psychological principles.

All aspects of a data-sensitive analysis require philosophical considerations: the significance of variables and data points must be established; appropriate hypotheses and models have to be developed and tested; and data sources should be identified and evaluated. However, data-driven analyses also require input from additional disciplines: for example, the algorithms for analyzing large-scale datasets are borrowed from computer science and, when we talk about group structures and scientific reasoning, we need to include results from psychology. We see that interdisciplinary work provides a great opportunity for contemporary philosophy to enrich its scope and increase its reach.

Chapter 3

Optimization of resources within a scientific project: the case of high energy physics

In the present chapter, we discuss methods that can be applied to improve the allocation of resources in high energy physics. First, we show that high energy physics is a fruitful ground for operational analysis based on citations since the field itself expresses an inductive behavior that can be captured by machine learning algorithms (Perović & Sikimić under revision). This inductive behavior guarantees a relatively quick, reliable and stable consensus about the results in HEP (Schulte 2000).

Once we established that the usage of citation metrics is meaningful in this field, we discuss a variety of different operational analyses that are and could be applied to external data in HEP, such as the data-mining techniques or citation-based statistical analyses from (Perović et al. 2016, Sikimić et al. 2018).

3.1 Operational approach

An operational analysis is a special type of a data-driven analysis that provides computable data-driven models. Operational approach (OA) studies are mainly applied in sociology of science, e.g., (McLevey & McIlroy-Young 2017). However, some philosophically flavored studies have recently been published (Perović et al. 2016, Sikimić et al. 2018). An operational approach investigates numerically expressed relations between properties of scientific projects such as the number of researchers, the project duration, the publication outputs, etc. An inductive approach (IA), on the other hand, is an analysis of scientific learning within a field, e.g., (Kelly et al. 2016, Kelly & Genin 2018, Schulte 2000, 2018).

Studies presenting OA and IA results are published by different communities. While studies using OA are usually found in journals publishing about science and research policy or social epistemology, e.g., (McLevey & McIlroy-Young 2017, Mutz et al. 2017), IA based on Formal Learning Theory is mainly used in philosophy of science, e.g., (Baltag et al. 2016, Kelly et al. 2016, Kelly & Genin 2018, Schulte 2018). Both employ methods and concepts from computer science.

3.2 Inductive behavior as a constraint to operational research

In order to identify optimal conditions for generating scientific knowledge, data-driven analyses can be successfully applied under specific terms. As the citation impact is one of the major output parameters, the main question is whether the citation record can serve as a reliable measure of the success of a project. Perović & Sikimić (under revision) claim that, though this is not always the case, in certain circumstances, citation metrics represent a reliable measure of impact and can be safely used in a data-driven analysis.

The criterion used is the degree of regular behavior in a scientific field. In the fields in which results are found following regular inductive patterns, with a small percentage of exceptions, the citation record will accordingly be a reliable measure of relevance of the experiment. This is a necessary, though not the sufficient, condition a field has to fulfill to be suitable for an OA.

In (Perović & Sikimić under revision), two examples of scientific disciplines with regular inductive behavior were discussed: high energy physics and phylogenetics. On the other hand, the authors argue that experimental plant biology does not express the same regular behavior.

The conservation laws (e.g., conservation of momentum, energy, charge) are the baseline principles of inductive reasoning in high energy physics. In this area of experimental physics, the consensus about results is relatively quick, stable, and relevant over long periods of time (decades). The reason for that is its relatively regular inductive behavior which postulates the conservation principles as the core one. Moreover, the Formal Learning Theory approach demonstrates that this state of affairs is a result of a reliable pursuit (Schulte 2000). Thus, based on the inductive analysis, in this case an operational data-mining technique is favorably qualified. Another example of a subdiscipline with a relatively regular inductive behavior is phylogenetics, which we will discuss in chapter 5.

3.3 Inductive behavior in HEP

High energy physics or particle physics studies the components of matter and radiation. The core principles to identify new particles such as the Higgs boson are conservation laws. Every observation is explained with a model that takes into account the conservation laws.

Formal Learning Theory (FLT) and Machine Learning Theory (MLT) are used to

describe discovery patterns in the field. FLT is used to identify algorithms and rules of inference that constitute reliable inductive methods. Parsimonious algorithms that require a minimal set of rules to explain the results have been shown to be more reliable than more complex solutions (Kelly et al. 1997). General MLT applies the theoretical approach of FLT. It is used by physicists to identify formal rules and identify patterns in scientific reasoning. Thereby, FLT and MLT can be used to reconstruct the scientific pursuit. This regular behavior that can be detected in HEP is advantageous when it comes to its suitability for citation-based analyses, because it guarantees a relatively high degree of reliability of the discoveries.

The second advantage of HEP for quantitative studies is that the consensus is reached relatively quickly and reliably. The experimental results are rather unambiguous and stable over long periods of time. This allows us to measure the quality of a project with the help of citation counts. Influential projects result in more cited publications. Therefore, we are able to track the judgment of the peers by using citation rates as a proxy. One has to keep in mind that author lists with hundreds of publications prohibits us to properly attribute the contribution of the individual researcher to a specific project, but it does not prevent us from evaluating the projects. Finally, the costs of experiments in high energy physics are immense. Because many experiments require equipment only available at a few or even only one single site, experiments are usually unique.

Perović & Sikimić (under revision) proposed the inductive test for the applicability of analyses based on citations. The scientific pursuit has to satisfy the following three conditions:

1. The pursuit follows inductive rules;

2. The scientific community reaches the consensus about the results relatively quickly and reliably;

3. The citation count reflects this consensus.

Several algorithms which reconstruct the inductive scientific pursuit in high energy physics have been developed (Kocabas 1991, Schulte & Drew 2010, Valdés-Pérez & Erdmann 1994, Valdés-Pérez 1996). They can be employed to model the inductive process on an experimental dataset. From the OI perspective, the relatively regular pattern of discovery in HEP together with the publishing and citation tendencies favorably characterize the use of citation metrics in the field as an informative project output.

3.4 A case study: application of DEA to HEP

Since the research in HEP follows the rules of the IA, one has reason to expect that an OA will give insightful results. Indeed, there have been several operational analyses conducted on data from HEP. A three-part assessment of projects in HEP focused on CERN based on the number of published papers and citations was given in (Martin & Irvine 1984, Irvine & Martin 1984b, Martin & Irvine 1984b).

Martin and Irvine assessed the performance of the individual accelerators within CERN, as well as, the performance of CERN in comparison to other HEP laboratories (Martin & Irvine 1984, Irvine & Martin 1984b, Martin & Irvine 1984b). They gathered extensive quantitative data to help improving HEP experiments (Figure 3.1). As a key metrics to compare the laboratory performance, they used the number of published papers and the number of citations. In addition, they presented qualitative data gathered by interviews with physicists. Both qualitative interviews and citation counts show a bigger scientific impact of the American laboratories. Especially the Stanford Linear Accelerator Center outperformed other laboratories in the early phases of particle physics, which was until the late 1970s (Martin & Irvine 1984). In (Irvine & Martin 1984b) major obstacles, such as political questions and the four-year lead of Fermilab, were discussed as relevant factors preventing CERN to utilize its full potential. On the

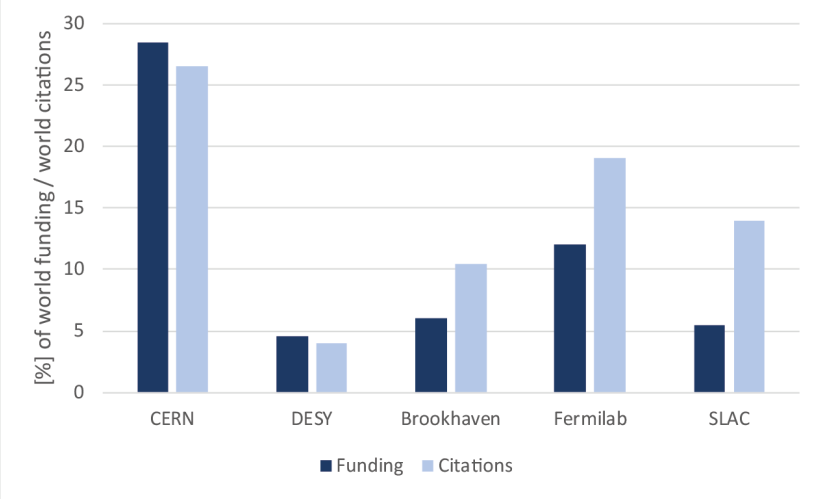


Figure 3.1: **Comparison between the input and output of the major laboratories in the period of 1969-1978.** The American HEP laboratories Brookhaven, Fermilab, and SLAC only received about one quarter of the worldwide funding but their publications got almost half the citations. The figure is based on (Martin and Irvine 1984a).

other hand, they argue, that CERN, being experiences in managing large collaborations, had an advantage over the American HEP laboratories when experiments began to include hundreds of researchers (Martin & Irvine 1984b). In addition, they foresaw already in 1984 that experiments in HEP would become too expensive to be financed by a single country or region and that a cooperative approach between the American laboratories and CERN would be beneficial. Nowadays, the United States has more researchers affiliated with CERN than any other country (Cho 2008).

Following (Perović et al. 2016, Sikimić et al. 2018), we will focus our analysis on the HEP laboratory Fermilab. In (Perović et al. 2016) the history of Fermilab was summarized. Founded in 1967, Fermilab was the first physics mega-laboratory for non-military purposes established in the US. Researchers there have made several breakthrough discoveries, including the bottom and top quarks, key elements of the standard model

of particle physics. This made Fermilab one of the most successful HEP laboratories (Hoddeson et al. 1997, 2009). Only in 2009 the Large Hadron Collider, at the European Organization for Nuclear Research (CERN) in Geneva, broke the energy record for collision energy, taking over the scientific supremacy of Fermilab (Reich 2011). Nowadays, Fermilab has a budget of almost 300 million euros and employs more than two thousand people.

Its development can be separated into four phases with different foci. In the first phase that lasted until the late 1970s, it housed multiple teams performing hundreds of experiments. In contrast to other HEP laboratories like CERN or the Lawrence Berkley Lab, it was not centralized but functioned as a host institution were experimental teams could assemble and perform their experiment within a specified time period (Galison et al. 1992, Hermann et al. 1987, Hoddeson et al. 2009, Nieva 1985, Perović et al. 2016).

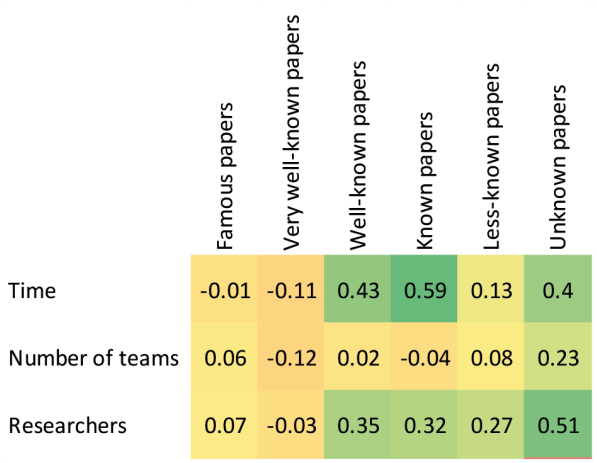
Between the late 1970s to the mid 1990s, the focus shifted and the number of experiments was reduced in favor of longer projects lasting several years. Later, in the third phase, Fermilab became home of the Tevatron collider, the most powerful collider at the time. It was used for colliding particle beams, resulting in a further reduction of the number of experiments and increase in their duration. The timeframe between design, commission and performance of an experiment increased to up to a decade and the research teams became even larger (Perović et al. 2016). These trends continued in contemporary HEP at the Large Hadron Collider in CERN. This shift to projects employing thousands of researchers is much more substantial than the difference we can observe between the first two phases of particle physics in Fermilab (Boisot et al. 2011). At Fermilab itself the projects became smaller again when the Tevatron collider was switched off in 2011. The direction of Fermilab shifted from the "energy frontier" to the "intensity frontier". Researchers now aim at observing rare interactions of known particles rather than discovering new particles using higher energies. This shift was accompanied by a decrease in the team size. While the collaborations working on the

Tevatron experiments had about 600 members, now the research teams comprise of about 100 to 200 members (Reich 2011).

The shift between the first three phases of research at Fermilab was gradual. Perović et al. (2016) and Sikimić et al. (2018) used experiments performed during the second phase to capture fairly homogenous organizational structures. This means they studied projects conducted by teams of moderate size. Their analysis is particularly useful because, despite the very famous examples of research conducted at CERN, the vast majority of experiments conducted in particle physics is still performed by groups of moderate size. The analysis of older research projects allowed Perović et al. (2016) to assess the impact of Fermilab projects quantitatively, based on citations metrics.

Perović et al. (2016) performed a DEA on 27 Fermilab experiments and computed the efficiencies of the individual performances in relation to the size of the research groups involved (Figure 3.2). The number of times the publications from the projects were cited served to establish the performance of the project. This quantitative study revealed that the efficient research groups were smaller than the groups responsible for the inefficient studies. The results of the analysis from (Perović et al. 2016) passed the statistical sensitivity test. They are robust and not influenced by outliers, which is a common concern when using DEA (Cooper et al. 2004). Also, after qualitatively analyzing the results, it was confirmed that the algorithm successfully categorized experiments into efficient and inefficient ones. The efficient projects were smaller than that the inefficient ones both with respect to the number of researchers and the number of teams. These results agree with the work from Lauer et al. (2017) and Cook et al. (2015) in biomedical research. As we discussed in chapter 2, both supported more funding for smaller laboratories at the expense of larger ones.

The impact of the team size on the performance observed by Perović et al. (2016) was high. Even though the projects addressed very different questions, ranging from the establishment of a new experimental technique to projects focusing on the discovery



	Famous papers	Very well-known papers	Well-known papers	Known papers	Less-known papers	Unknown papers
Time	-0.01	-0.11	0.43	0.59	0.13	0.4
Number of teams	0.06	-0.12	0.02	-0.04	0.08	0.23
Researchers	0.07	-0.03	0.35	0.32	0.27	0.51

Figure 3.2: **Correlations between variables in the DEA model.** The table gives the correlation coefficient between the properties. The number of unknown papers correlates with the number of researchers. The figure is based on (Perović et al. 2016).

of new facts, the six efficient experiments were performed by very small teams. In half of the cases the researchers were grouped in only two teams. In contrast, all inefficient experiments involved a bigger number of researchers – sometimes several dozed primary scientists – which were divided into six to eleven teams.

Furthermore, the influence of the project duration on the efficiency became clear: while efficient projects were usually short and lasted between one and seven years, inefficient ones toke up to nine years. This indicates that the prolongation of inefficient experiments usually does not help to reach ambitious goals, but rather wastes additional resources without big impact. This result can be explained by the psychological principle of commitment and consistency with previous beliefs: researchers, like all other humans, have the tendency to prolong unfruitful projects, hoping to justify the past investments, instead of quitting them to cut the losses (Cialdini 2001). Every prolongation of a research project should be judged based on future resources that will be spent during the prolongation, and not based on already invested resources (Sikimić

et al. 2018).

Despite the fact that politicians sometimes promote large institutions and agglomerations, Bonaccorsi & Daraio (2005) showed that the size measured by the number of personal negatively correlates with the productivity. They argued that science does not profit in the same way from the division of labor as industrial production units. As publications are shared international, scientists all over the world can profit from the division of labor – this is not limited to organizational boundaries. This is both true for researchers' productivity in terms of published studies, and when we include teaching and other services provided: researchers in smaller laboratories are in general more productive (Carayol & Matt 2006).

Projects in HEP are too complex to be performed by individual researchers. However, the trend pointing at the highest efficiency of small research groups also agrees with results by Nieva et al. (1985), who found a curvilinear relationship between the number of researchers involved in a project and its efficiency. This means that efficacy increases with adding new research members only up to a certain point, after which it decreases. Even though, one might expect small experiments to be more focused and more efficiently managed, bigger research teams benefit from a larger variety of skills and approaches assembled together. Since larger teams are harder to coordinate, it is necessary to find the saturation point after which adding new team members becomes inefficient. With the results of Perović et al. (2016), we can point to a specific threshold, which limits the group size of efficient projects. However, the question of how this curvilinear relationship develops when we consider a wider organizational structure and measure the output of more parameters than publications and citations, remains. In this case it would be interesting to see, whether the general curvilinear relationship stays intact, or it is sensitive towards the scientific profile and the epistemic diversity of the involved researchers.

Perović et al. (2016) used a relatively homogenous group of experiments, all per-

formed at the same institute within the same time period to understand the relationship between the group size and the efficiency of the project. The numerical results of the study highlight the value of quantitative studies on specifically defined projects. We can see how the group size determines the efficiency of the project. This does not mean that every research goal can be achieved with a small group of scientists; some projects simply require large interdisciplinary collaborations. However, often investing resources into smaller projects increases the performance of the institution and thereby benefit the scientific community. In Fermilab, this approach was introduced in the beginning of the 1970s during the directorship of R.R. Wilson (Hoddeson et al. 2009). Instead of aiming at long-term projects performing few experiments with eventually huge impact, the Fermilab became a turn-around site for many small experiments. Only later, the policy shifted to long-running strings of experiments.

Similarly to the effect of large groups on the efficiency of individual projects, the agglomeration of scientific research at central locations leads to a decrease in the productivity of the researchers. Van der Wal et al. (2009) showed that in environmental research the publication rate and the quality of the publications are negatively impacted by the centralization. This is particularly interesting to see, because one would expect reduced costs from centralized facilities providing services to the researchers and a positive impact of the expertise available at these facilities Bonaccorsi & Daraio (2005). These arguments are often used to justify investments into large research clusters or institutes. However, it seems that productivity decreases with centralization, and that the researchers from smaller sites are more likely to publish impactful research. Van der Wal et al. (2009) argue that two factors mainly contribute to the decline in productivity in large institutions: reduced commitment and communication problems. In large institutions, it becomes harder for employees to understand the purpose and impact of an individual contribution. Because scientists are mainly intrinsically motivated they might be especially vulnerable towards this effect. Furthermore, the time spent

on communication grows disproportionally with the size of the institution. (Perović et al. 2016) results indicate that similar effects might apply to HEP. Because upscaling provides positive aspects as well, the optimal team size depends on the specific requirements of the field. With the help of DEA Perović et al. (2016) were also able to suggest the optimum for experiments in HEP.

In addition to the agglomeration of research in centralized locations, we can observe the impact that the location of the institution has on scientific performance. Despite being more attractive for the researchers as employees and other economic advantages, research institutions in highly developed urban regions do not provide a positive effect on the productivity of the researchers in biomedical research (Bonaccorsi & Daraio 2005). A particular example of successfully decentralized research is the Max Planck Society. Its more than 80 institutes distributed over Germany have an output of high impact publications only surpassed by the Chinese Academy of Sciences and the Harvard University, institutions with vastly higher budgets.[1]

The presented results highlight the impact of the structure of the research groups and institutions on the epistemic performance. We used the example of HEP research in Fermilab and embedded the results from DEA by Perović et al. (2016) into the historical context. In HEP, small collaborations with approximately two teams were singled-out as optimal by the research in (Perović et al. 2016). Furthermore, we discussed the effect of the group structure in other disciplines and presented possible reasons for the worse performance of larger groups. In larger groups the increase in scientific competences is frequently overshadowed by communication problems.

[1]For more details, please consult: https://www.natureindex.com/news-blog/twenty-eighteen-annual-tables-ten-institutions-that-dominated-sciences.

Chapter 4

Investments in HEP and the halting problem

In computer science the halting problem stands for the question of whether an algorithm together with specific input values will terminate or not terminate – *halt*, at any time point. Turing proved that no general algorithm could determine this for every given computer program and input (Turing 1937). The time needed for a program to terminate or not terminate is the basis for determining the complexity of a procedure. Inspired by this fundamental question in computability theory, we refer to the problem of whether a scientific project will be fruitful or futile after a certain period of time, i.e. whether there is an epistemic saturation point in experimentation, analogous to the halting problem in science. In the literature on stopping rules, researchers investigate at which point it is optimal to stop gathering data and start analyzing them, e.g. (Stanev 2012, Steele 2013). In our research, we are interested in understanding at which point scientists working on large projects in HEP should stop investing in a project that is failing to produce results. Since this question is data- and field-dependent, the appropriate way of answering it is via data-driven analyses. How reliable the results of the analysis will be, depends on whether there might be a pattern that governs discoveries

in the field. Often, such a pattern cannot be established. Still, an approach based on citation metrics is meaningful and informative in the field of experimental physics because of its inductive nature (section 3.3).

In order to tackle the question of an epistemic saturation point in HEP, Sikimić et al. (2018) employed a data-driven statistical analysis. The investments in HEP projects are extensive: they include the constructions of costly equipment, its maintenance, employment of numerous academic staff members and technicians over long periods of time, etc. Thus, Sikimić et al. (2018) considered project duration and team members as important parameters for establishing projects' efficiency in HEP.

When it comes to the time invested in an experiment, one should distinguish between two important parts of each project: the time spent on gathering data, and the time spent on analyzing them. For instance, the INSPIRE HEP database provides very rich information about the time spent on each experiment conducted in Fermilab. From the perspective of the scientific community, the time spent using the expensive Fermilab technology is costly. For each day of the experiment, there is a tough competition between research groups and ideas, because many scientists from different institutions want to use the technology simultaneously. In contrast, the time spent on analyzing the results might not be as expensive because it does not involve the usage of excessively costly equipment, technical staff or other scarce resources. Furthermore, most research at HEP laboratories is performed by guest scientists from various institutions. They are not paid by the HEP laboratory while they are working on the site, nor when they are analyzing the data. Therefore, they do not cost the laboratory. However, while they are working on the experiment, they use very valuable equipment owned by the laboratory, and require support from the technical and scientific personal on the site.

The time between the start of a project and the publication can be influenced by many external factors, such as the delay between the acceptance of an article and the final publication of it. For example, some experiments took only a few months, but the

related publications only appeared in the next years.

When Sikimić et al. (2018) evaluated the efficiency of scientific projects, they focused on the time spent on performing the experiment at the site. However, when they analyzed the longitudinal data on the performance of individual experiments, they followed the project from the start day until the last publication. This allowed them to show how fast one can expect a publication from a project and how the output developed over time.

4.1 Method

In (Sikimić et al. 2018) the correlations between these investments and scientific outputs were investigated for Fermilab projects that belong to the middle period of HEP, i.e., the period between 1975 and 2003. In the middle period, experiments in Fermilab belonged to the largest physics experiments both in terms of the accelerators used and the size of the collaborations. During this period, Fermilab constructed the Tevatron particle accelerator, which was at the time the collider operating with the highest energies on earth. It was, for example, used to discover the top quark, the last missing particle of the standard model. The period is also particularly instructive for today's science because already then experiments started to last longer than scientists typically stay in one position. Sikimić et al. (2018) opted for the middle period of the development of HEP because the time distance allowed them to access the impact of the projects. They analyzed the impact of the time between the project start and the moment of publication.

The authors employed a data mining technique on external data about 49 experiments extracted from the INSPIRE HEP database to analyze the correlation between the quality and quantity of the publications and the resources invested in a project (Sikimić et al. 2018). They focused on the number of researchers and research teams

involved, and the time invested in each project and analyzed how the publication output of the projects developed over time. Papers were classified into the following six categories according to the classification provided by physicists [1]:

1. famous papers (250 + citations);
2. very well-known papers (100–249 citations);
3. well-known papers (50–99 citations);
4. known papers (10–49 citations);
5. less known papers (1–9 citations);
6. unknown papers (0 citations).

To evaluate the impact of the paper as objectively as possible, self-citations were excluded. As we described in section 3.3, the consensus on results in high energy physics is reached relatively quickly and it remains stable over long time periods (Schulte 2000). This allows us to use the number of citations as a rather good proxy for the impact and value of each experiment. As a last selection criterion, Sikimić et al. (2018) removed linked experiments but instead analyzed experiments that started from scratch.

4.2 Results

Sikimić et al. (2018) first analyzed the time while experiments were running. This analysis revealed an epistemic saturation point and a clear cut between fruitful and futile projects for research in HEP. Interestingly, all but one project resulting in more than one very well known or famous publication were finished after less than three years, resulting in a negative correlation between the time spent on a project and the output of highly regarded papers (Figure 4.1).

[1] This classification is provided on HEP Inspire platform.

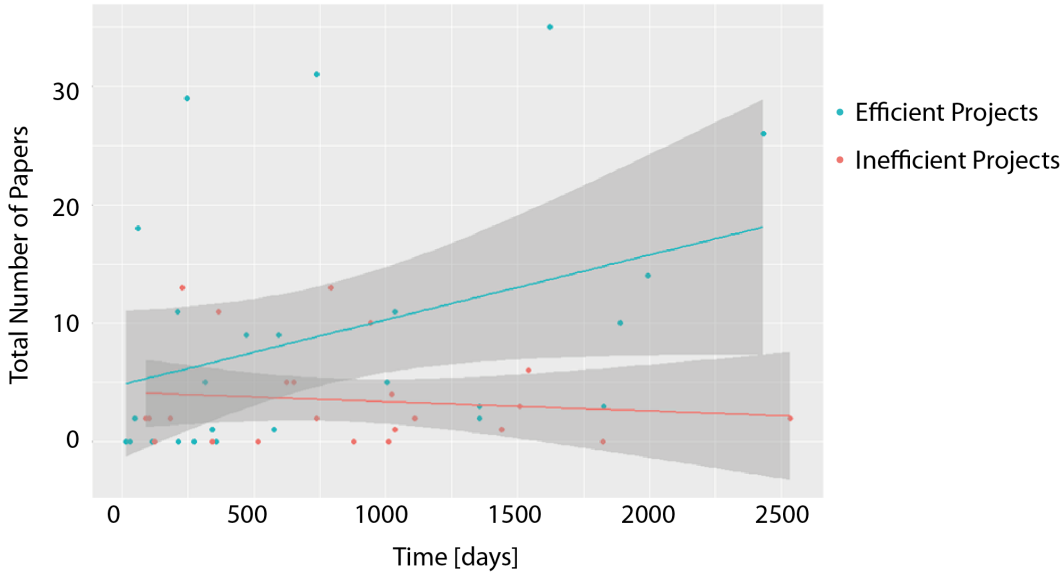


Figure 4.1: **Publications per time.** The figure is based on the unpublished results by Sikimić, Radovanović, and Perović.

In contrast, the number of less cited papers increases over time, highlighting that researches continued to invest valuable resources into their projects far beyond the epistemic saturation point. A further important result concerns the separation between fruitful and futile projects: papers with at least 50 citations correlate strongly with each other while papers which are cited less than ten times or are not cited at all cluster together (Figure 4.2). As the number of those less cited papers increases over time these results further highlight that the epistemic saturation point has been passed before.

However, this result is also valuable from another perspective: the clear separation between fruitful and futile projects and the epistemic saturation point is usually supported by the citation frequency of several papers, supporting the view that the number of citations of individual papers is a good proxy for the value of the research conducted within a project. Another interesting point is that the number of less cited papers correlates with the number of researchers and the time spent on the project (Figure 4.3).

	Famous papers	Very well-known papers	Well-known papers	known papers	less-known papers	unknown papers
Famous papers	1.00	0.68	0.41	0.16	-0.06	0.15
Very well-known papers	0.68	1.00	0.38	0.38	-0.07	0.16
Well-known papers	0.41	0.38	1.00	0.64	0.13	0.64
known papers	0.16	0.38	0.64	1.00	0.38	0.60
less-known papers	-0.06	-0.07	0.13	0.38	1.00	0.27
unknown papers	0.15	0.16	0.64	0.60	0.27	1.00

Figure 4.2: **Table of correlations between the number of publications in each category.** Note that highly cited papers (top left) and not well-cited papers (bottom right) clustered together. The figure is based on the results from (Sikimić et al. 2018).

	Famous papers	Very well-known papers	Well-known papers	Known papers	Less-known papers	Unknown papers
Time	0.06	-0.08	0.16	0.28	0.22	0.27
Number of teams	0.1	-0.05	0.06	0.14	0.19	0.25
Researchers	0.07	-0.04	0.24	0.36	0.49	0.47

Figure 4.3: **Correlations between the time spent on a project and the team structure.** The time spent on a project and the number of researchers correlate with the number of less cited papers but not with the number of highly cited ones (based on (Sikimić et al. 2018)).

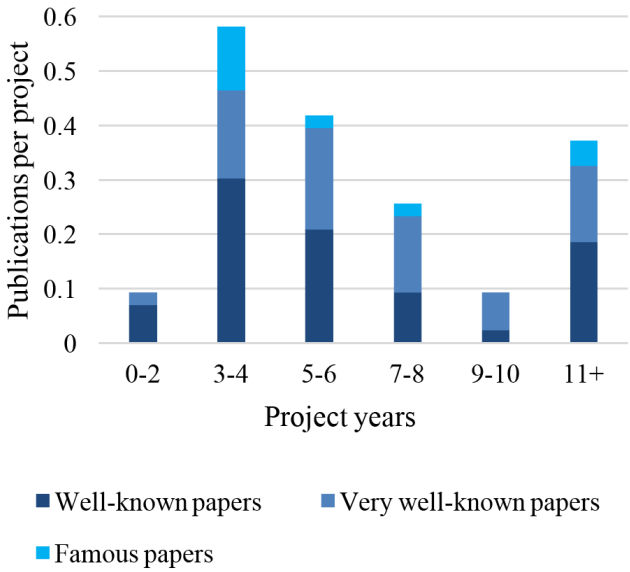


Figure 4.4: **Publications per project from the starting date.** The figure is based on the results from (Sikimić et al. 2018).

In addition, Sikimić et al. (2018) analyzed in which points of time the project results get published and how influential these publications are. They were able to identify an epistemic saturation point that occurs relatively early – three to four years after the project start. After this timepoint, the chance to publish well-known or famous papers decreases by one half every two years (Figure 4.4). These results are in accordance with the trend observed for the project duration: longer projects usually do not result in influential publications (Sikimić et al. 2018). There are occasional outliers, i.e., papers that are published even one decade after the project start. However, the low frequency of such publications should be considered when decisions about the prolongation of projects in high energy physics are due.

It is important to note that the project duration was not dependent on funding schemes, since each project applied for the time researchers found fit. In the analyzed data, the maximal duration of projects was seven years, and 16% of them took more

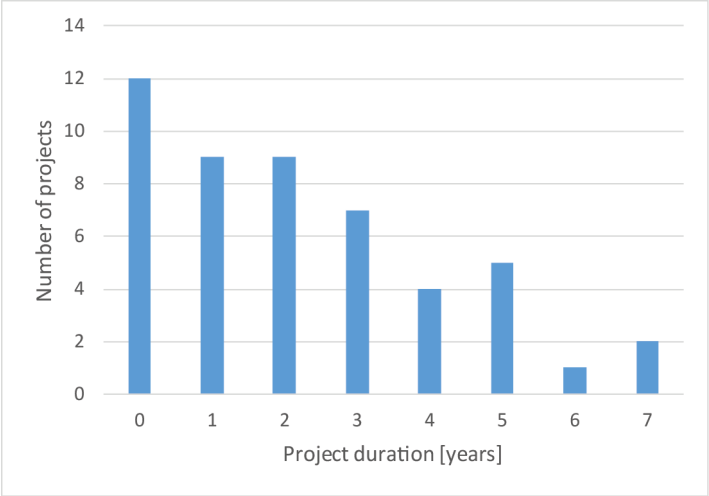


Figure 4.5: **Duration of the analyzed projects.**

than four years, which we checked for the purpose of this thesis (Figure 4.5).

4.3 The sunk cost bias and the halting problem in science

We assume that scientists are in general aware of the negative correlation between project duration and publication output such as the one observed by Sikimić et al. (2018). However, some scientists continue to invest in their own projects long after they crossed the epistemic saturation point. This can be explained by psychological mechanisms such as the principal of commitment and consistency. This psychological principal is responsible for the pressure of the scientists to remain in agreement with their previous beliefs, despite novel evidence (Cialdini 2001). While some degree of persistency is expected from a good scientist, as discussed above, the principal of commitment and consistency can motivate scientists to continue investing in unsuccessful projects, because they previously committed to them. The results by Sikimić et al.

(2018) could be explained by this effect. The authors did show that the total number of papers correlates positively with the duration of the project. However, the increase in the number of publications was not mirrored by citations of the results. In fact, only the number of low-cited papers increased with time (Sikimić et al. 2018).

A second psychological mechanism associated with the prolongation of futile projects is the sunk cost bias. While the principal of commitment and consistency prevents us from changing the directions because we feel committed to our prior beliefs, the sunk cost bias motivates us to continue investing in futile projects because of our prior investments (Arkes & Blumer 1985). Humans have a tendency to keep pursuing their initial projects, because they do not want their investments to fail. The bigger the initial investment in a project is, the stronger is the pressure to deliver the desired results.

The sunk cost bias motivates a consumer to drive to an event through a snowstorm because she already invested money in the tickets. On the other hand, if she had gotten the tickets for free, she would not have gone (Thaler 1980). Sweis et al. (2018) showed that this psychological effect is a very basic psychological principle that both human and non-human animals follow. They analyzed how long humans, rats and mice are willing to wait for a reward. They showed that the sunk cost bias exists in all these species and that it depends on the previously invested time.

In order to understand better the effect that the sunk cost bias might have on scientists, we turn to computer simulations which we discuss in more detail in chapter 6. With our agent-based model we can show that the sunk cost bias is an individual bias and not a group effect. We simulated the communication between scientists for 200 rounds. In each round, the agents interacted with all connected agents and updated their beliefs by a fraction of the beliefs of their interaction partners. We considered completely connected, centralized, hierarchical, and weakly-hierarchical group structures (Figure 4.6). As outcome, we calculated how fast the agents reach a consensus

about a hypothesis. The outcome of the individual simulations depends on the initial beliefs of the agents, which is normally distributed around an undecided state. To account for the sunk cost bias, we introduced a threshold: agents only change their opinion when the evidence for the opposite view is 10% higher than the evidence for their prior belief. The results show that the sunk cost bias is independent of the group structure. It affects scientists organized in completely connected groups in the same way as scientists in centralized or hierarchically structured groups (Figure 4.6). Therefore, we cannot rely on optimal team structures to fight the sunk cost bias. Instead we have to focus on warning scientists about the risks of suboptimal choices. The analysis of Sikimić et al. (2018) is an example of such an attempt. It provides background information based on which scientists could reconsider their decisions about experiment prolongation. For instance, after initial investments in a project, scientists might be incentivized to continue applying for the continuation of their project because they do not want that their investments are in vain. However, evaluation committees can consider that there is a tendency of reaching an epistemic saturation point and decide accordingly.

4.4 Conclusions

In summary, the analysis of project duration highlights two important observations: HEP project outputs have an epistemic saturation point which should be considered when decisions about the prolongation of the project are due. Longer work on the project usually only results in a higher number of less influential publications. The first analysis showed that the total output of impactful papers of a research project is larger when the time conducting the experiments is relatively short. The second one showed that this also holds true for every single experiment: the most famous papers are published shortly after the project start. Furthermore, publications with many

citations as well as publications with few or no citations cluster together, supporting the hypothesis that there is a clear cut between fruitful and futile projects (Figure 4.1).

Still, physicists kept investing in inefficient projects after the saturation point. The reasons behind these decisions most likely involve the sunk cost bias and the principle of commitment and consistency with previous beliefs. Using computer simulations, we demonstrate that the sunk cost bias is not influenced by the group structure. Thus, in order to overcome this fallacy, researchers need to be aware of the epistemic saturation point in their field. A data-driven approach has the potential to deliver an external guideline to fight these psychological fallacies and thereby benefit science.

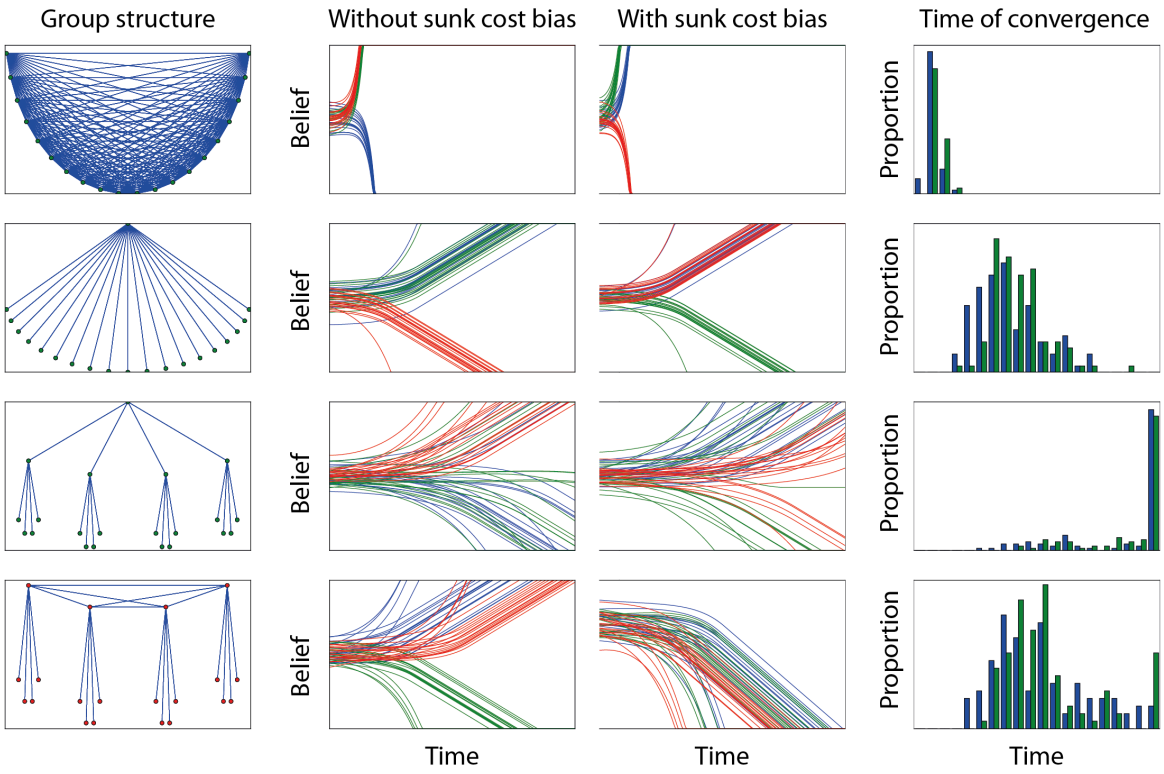


Figure 4.6: **Agent-based models of group dynamics.** In the first column, the analyzed group structures are visualized. Each agent is represented as a red or green dot; the color indicates the final belief in one simulation. The connections between agents are represented as lines. In the second and third column, the individual beliefs of all agents in three simulations are plotted, each simulation is shown in a different color (second column: standard model; third column: model with sunk cost bias). The last column shows the speed of convergence to a consensus. The standard model is represented with blue, while the model accounting for the sunk cost bias with green color. The sunk cost bias is responsible for a small delay, but it does not affect the overall outcome of the simulations.

Chapter 5

Optimization of resources within a scientific project: the case of experimental biology

The consensus on results in contemporary experimental biology is, in general, less quick and stable in comparison to high energy physics. The reason for this is that the nature of research in biology often allows for exceptions of the strict inductive rules that would ideally govern scientific discovery (Perović & Sikimić under revision). The answers in biology cannot be easily computed because many research parameters are open for different interpretations. Even when a current view is prevailing, other views frequently remain supported by at least some researchers and schools. In addition, the consensus in biology is often unstable over time and citations do not necessarily reflect the relevance of the publication. Contopoulos-Ioannidis et al. (2008) analyzed the impact of highly cited publications on medical practice. They focused on publications cited more than 1000 times. Only one from 101 analyzed claims had an extensive application in medical practice. In addition, the median time lag between the first formulation of the hypothesis and the first highly cited publication based on this hypothesis was 24 years.

This shows that very long time periods pass between the discovery of the hypothesis and its translation into medical practice. Furthermore, the majority of highly cited claims did not deliver to the expectations.

Prevailing views in biology are often influenced by external aspects in addition to the presented facts, such as the reputation of the researchers advocating them. As we will discuss in chapter 7, the discovery that the human papillomavirus causes cervical cancer was hampered by the language barrier. When Harald zur Hausen first presented the findings of his group showing the correlation between the virus and cancer, his presentation was dismissed in part because of the language barrier (Cornwall 2013).

Still, we can find some subfields of biology which exhibit stricter inductive behavior and their discoveries can be captured by machine learning algorithms. One of these subfields is phylogenetics, the study of evolutionary relationships.

5.1 Inductive behavior and phylogenetics

Phylogenetics is particularly suitable to be analyzed in a similar way to HEP, because its basic principle – the principle of parsimony – functions analogously to the conservation principle in physics (Perović & Sikimić under revision). The principle of parsimony is based on the idea that a change between a common ancestor and the analyzed species is unlikely. Therefore, the best hypothesis describing the evolutionary relationships between any number of species, is the one that requires the fewest evolutionary changes (Yang & Rannala 2012). As virtually all researchers agree on this basic principle, the consensus is reached efficiently.

Even though this general principle is guiding the construction of all phylogenetic trees, the prevailing views about the relationships between organisms are not always constant. It can change when new technology supersedes older approaches. In the pre-genetic era scientist first compared morphological differences between species and –

after they understood the importance of the embryogenic pattern formation processes – the embryogenesis (Perović & Sikimić under revision). Now, as sequencing got cheaper, genetic sequence comparisons became the gold standard of phylogeny, resulting in the term phylogenetics. While in the beginning of this new era researchers focused on short conserved fragments, they nowadays focus more and more on whole genome comparisons. All these shifts resulted in changes in the phylogenetic trees which are considered to represent the current view of the evolutionary relationships. The shifts in technology are associated with changes in the consensus. It is important to note that these changes, which are every time based on increased knowledge, follow an inductive pattern and the change only results from including more data into the model. The inductive principle of parsimony remains the criterion of selection of hypotheses.

As we have seen in chapter 3 we can use machine learning to infer the efficiency of HEP projects independent of the underlying experimental process. Machine learning uses a learning dataset to construct a model that is then applied to new cases. The reasons for the decisions of the algorithm are frequently not accessible but still highly reliable. They are not only able to beat humans in chess and go but also assess immensely complicated medical databases to find patterns inaccessible for the human brain. In phylogenetics however, biologists use simple models, which are based on the principle of parsimony. To assess them with an OA, we can therefore study the relevant algorithms.

As mentioned earlier, phylogenetic evolutionary relationships are based on sequence comparisons between different species. Usually, scientists focus on the comparison of conserved genes, genes which are present in the whole clade analyzed, these genes are aligned and then the differences are evaluated. How these genes are aligned and how changes are evaluated might differ. For example, in some analyses, only conserved regions will be considered, while others might use the whole gene. Furthermore, not every change in the encoded protein has the same result, as some of the 20 amino acids,

the building blocks of the proteins, are more similar to each other because of their charge or size. This results in unequal exchange rates between the amino acids. As these similarities are based on the chemical properties of the amino acids, the exchange rates are fairly constant over most proteins. Scientists have included this observation into their models and calculated numerical scores of the likelihood of any amino acid change into any other amino acid, based on the observed exchange rates in a large number of homologues proteins. These are proteins that share the same origin. When a tree is constructed not only the number of changes is considered, but also their likelihood. Algorithms consider these likelihoods and reconstruct the trees containing the minimal number of unlikely mutations (Perović & Sikimić under revision).

To illustrate this, we can construct a tree that requires the minimal number of changes starting from the three sequences AAA, AAB, and ABB. When we assign the expected frequency 1 to a change between A and B we can construct different trees 5.1. To reach the smallest number of changes, AAA and AAB should be grouped together, and ABB closer to AAB than to AAA. This requires one change between AAA and the common ancestor of AAB and ABB, and another change between AAB and ABB 5.1 b). Any other tree requires additional mutations, see for example, 5.1 c).

While this can be easily seen and calculated in the case of three sequences of length three, it becomes increasingly hard if there are more sequences of greater length. In fact, the construction of the optimal tree is considered an np-hard problem, meaning that it can only be computed by comparing all possible combinations, a task which becomes impossible for large datasets. To resolve this issue, scientists developed algorithms that can approximate the optimal tree in more efficient ways.[1].

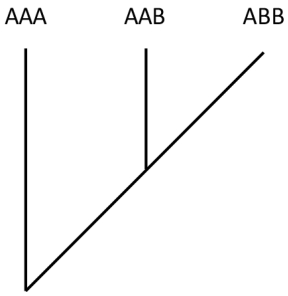
This approach in general results in a reliable tree, but of course also a rather objective algorithm has to be fed with adequate data. The main difference arises from the used sequence information (Perović & Sikimić under revision). In the beginning of

[1]For more on this topic, the reader is referred to (Yang & Rannala 2012)

a)

	AAA	AAB	ABB
AAA		1	2
AAB	1		1
ABB	2	1	

b)



c)

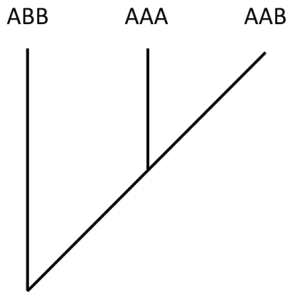


Figure 5.1: **Optimal and suboptimal tree of the sequences AAA, AAB, and ABB.** a) Table of necessary exchanges between the three sequences; b) optimal tree, when we define AAA as last common ancestor only two changes are necessary; c) suboptimal tree with ABB as last common ancestor. This tree requires three sequence changes because the mutation B→A has to occur both between ABB and AAA, and between AAA and AAB.

the genomics era scientists focused on particular stretches of DNA to establish relationships: for bacteria they focused on the 16S rRNA, an essential component of the protein biosynthesis machinery. However, when sequencing became cheaper, it was discovered that, even though closely related species share similar 16S rRNA, the reverse is not true. Species with extremely similar 16S rRNA can strongly differ when we look at the whole genome (Stackebrandt & Goebel 1994). This observation provides another example of how a shift in technology resulted in a change of the perceived view on the correct reconstruction of phylogenetic trees. With next-generation sequencing technology becoming cheaper than ever we are now able to sequence the whole genome of even complex organisms. For example, the sequencing of the genome of a human which took more than ten years when it was first conducted by the human genome project now takes only a few days and a single researcher. Therefore, scientists nowadays usually use whole genome data for phylogenetics.

One of the most common problems answered by phylogenetics is the identification

of proteins with common origin. This is important when we need to identify proteins related to the ones implicated in any human disease. Their relationships are usually being evaluated with the help of specific databases. They contain information on how frequent we observe any amino acid exchange in other homologues proteins. For example, BLOSUM62 contains the frequency of every possible amino acid exchange between proteins with a similarity of 62% (Henikoff & Henikoff 1992). To every exchange between the two proteins of interest a score is assigned. Also, insertion and deletions are scored and a final similarity is calculated. Since every researcher can choose the database and the penalties for insertions and deletions, the final tree may vary at least in details (Perović & Sikimić under revision).

A much bigger effect comes from the decision which data is used for the tree reconstruction. While we could easily just compare the whole sequence to get an "objective" results, this would not necessarily result in a meaningful tree: what matters for the function, and therefore also for the evolution, are the conserved domains in the proteins. Some domains are very conserved, because every mutation in it will destroy the function, others are hypervariable. Which part we focus on depends mainly on the question, if we want to find a receptor binding the same hormone in another organism we want to focus on the binding domain, when we want to understand the relationship of highly similar receptors in an evolutionary context we would focus on the hypervariable regions, because they provide the highest amount of information.

What we discussed above mainly applies to the binary trees that are commonly used. They are very valuable for the phylogenetic reconstruction of the relationships of plants and animals, but they are based on an important assumption that is often overlooked. In most species similar genes are derived from common ancestors and mutate over time ("molecular clock"). However, there are other ways to acquire a sequence. For example, bacteria are able to acquire DNA from other bacteria. Because of this mechanism they can easily gain resistance against an antibiotic if there are other bacteria in the

surrounding are already resistant. The presence of extremely similar DNA fragments in bacteria is therefore not conclusive evidence of a close relationship.

This short introduction into the constructions of phylogenetic trees highlights one of the core problems in the study of evolution. We can only access the currently existing species and we have limited data about extinct species. All conclusions about the likely properties of the last common ancestor are therefore based on models and probabilities. How these properties are weighted, and which possibilities are included, changes over time as new data sometimes challenge established ideas. Nevertheless, the phylogenetic trees are usually robust, and changes in their reconstruction require breakthroughs in technology or in our understanding of the inheritance of genetic material.

The results are relatively robust because the principle of parsimony, which is the basis for all kinds of phylogenetic analyses, is an efficient method to generate rules. The scientific reasoning mainly follows an inductive process (Perović & Sikimić under revision). Because the models which are used for phylogenetic reconstruction are based on parsimony, the scientific pursuit passes the machine learning test.

However, this only shows that the results follow inductive rules, but it does not prove that the output, the publications and their impact, follow them as well. To be suitable for an operational analysis based on citations, the citation metrics in the field has to correlate with the relevance of the research project. This condition is more difficult to satisfy, as the citation counts are influenced by many additional factors. While any proposed change in the phylogenetic history of mankind is heavily discussed and cited, the reconstruction of the phylogenetic tree of any other genus raises far less attention. And even though databases with the phylogenetic information are frequently used by researchers, they are rarely cited. This contrasts with HEP, where the correlation between citations and scientific relevance is investigated and argued for (Perović & Sikimić under revision). Hence, to assess the citation patterns objectively, further investigation is needed.

5.2 Inductive analysis in other areas of biology

Phylogenetics is only one of many subfields of experimental biology. In other research areas many different principles and approaches are applied, hence IA and OA cannot easily be applied in the whole field. One of the main problems is that the consensus is not reached in a straightforward manner. Frequently new studies challenging established views and promoting unorthodox hypotheses are published in high-impact journals. They are heavily discussed and cited only to disappear over time (Perović & Sikimić under revision).

There are many reasons why this is the case. The results discussed in scientific studies in biology are usually not hard facts, but rather interpretations of the available data: authors might use this data either to develop a new model or to support a model they had previously developed. Other researchers with opposing theories in mind might interpret the facts very differently (i.e., in the way which supports their own working hypotheses).

Furthermore, the career and funding incentives in science promote a culture of publishing exciting and highly-cited articles fast, instead of careful work with more than the absolute minimum of controls. Heesen (2018) claims that the scientific reward structure motivates rational agents to publish non-reproducible research. While publishing non-reproducible research has a negative impact on the performance of science as a whole, it can provide career advantages for those researchers. Specifically, Heesen (2018) argues that the harm of publishing non-reproducible research for the community is greater than the penalty for the individual scientist. This results in a high number of papers being published in high-impact journals that cannot be replicated (Pusztai et al. 2013). Reasons for their non-reproducibility can include deliberately vague descriptions of methods, which is a common way to prevent competitors from making fast progress on the newly published results. Even deliberately faked results are sometimes published. Because of the overproduction of publications, researchers can usually as-

sume that they will not be caught. Moreover, even when caught, their work is rarely retracted. In most countries the involved institutions are the ones that would have to pursue an investigation. One last big problem is the expectancy bias in published work. In almost every case, authors only mention experiments that support their explanation (Perović & Sikimić under revision). Fire's early work on RNA interference is an exception: even though it challenged his main conclusion, he published results stating that the proposed orientation dependence was not as absolute as he claimed (Fire & Moerman 1991). Although he did not understand this observation back then, he and Mello received the Nobel Prize in Physiology or Medicine in 2006 for their discovery and later explanation of the phenomenon. However, in general, all these factors slow down consensus because they interfere with the replication of experiments (Goodman et al. 2016).

When it comes to the interpretation of the results, the sheer quantity of phylogenetic studies makes the field accessible to machine learning algorithms. In contrast, many fields in experimental biology rely heavily on images as main results – but how these pictures are taken and analyzed, and which particular picture is chosen for publication, depends on the scientists analyzing them, and on their prior knowledge and beliefs (Perović & Sikimić under revision). A good example of this problem is the pathological assessment of the cancer stage. In this case we can assume that every pathologist has the best intentions, but (Vestjens et al. 2012) still found that even when the same samples were used only 83% of the diagnoses of local pathologists agreed with an independent review. This reveals how difficult it can be to interpret complex pictures. While these issues in routine medicine might become accessible to a more sensitive and objective artificial intelligence because of the wealth of training data (Ehteshami Bejnordi et al. 2017), experimental science will continue to be vulnerable to biased interpretation, and double-blind experiments are still very uncommon in academic research.

Another problem affecting reproducibility concerns the experimental conditions. Ex-

perimental conditions in phylogenetics are usually not as clearly defined as in experimental particle physics, and even when described as clearly as possible they leave significant room for interpretation. Some experimental conditions are difficult to control, such as the quality of light, room temperature, the humidity, the quality of the soil in plant biology, etc. By contrast, researchers in particle physics frequently use the exact same accelerators and detectors for different experiments. The same technical personnel might even be involved in the execution of different experiments, making it substantially easier to understand and replicate the experimental procedures (Perović & Sikimić under revision).

5.3 Non-parsimonious results

The scientific community and reviewers of scientific studies adjust the amount of supporting evidence necessary to accept a hypothesis to account for the perceived likelihood of the hypothesis being correct (Perović & Sikimić under revision). Hypotheses contradicting common beliefs or teachings (i.e., non-parsimonious ones), which require other views to be corrected, require more time to be accepted – if they ever are. This bias against novelty has been studied recently. Wang et al. (2017) showed that despite having a bigger impact on the field, novel results have a delayed recognition and are more frequently published in low-impact journals. Of course, reviewers might have very different views on which explanations are parsimonious, depending on their prior knowledge and their own opinions. This can result in one reviewer rejecting a manuscript for lacking proof for its exceptional claims, while another reviewer rejects the same manuscript for providing no new insights. The effect becomes obvious if we examine the impact of star scientists: after a premature death, non-collaborators publish more, while former collaborators publish less. The papers of the non-collaborators are also disproportionately likely to be cited highly (Azoulay et al. 2015). This shows that some

scientists block the promotion of novel and important results; new ideas only get the credit they deserve if the established scientists clear the way.

The discovery of prions as the cause for scrapie disease is a good example of how long it can take for non-parsimonious explanations to be widely accepted. The protein hypothesis, originally proposed in 1967, challenged Koch's second postulate requiring all infectious diseases to be caused by a self-propagating organism. The prion hypothesis is now widely accepted, and Prusiner was awarded the Nobel Prize 1997, but between his now-famous study in 1982 (Prusiner 1982) and the final proof when mice infected with prions developed scrapie, many years went by before this view became the consensus (Soto 2011, Perović & Sikimić under revision).

Zur Hausen faced similar opposition when he originally proposed that the human papillomavirus is the main cause of cervical cancer (zur Hausen 2009). While it was understood that viruses could integrate in the host genome and cause cancer, it was not considered possible that the main reason for any kind of cancer could be an infectious disease (Perović & Sikimić under revision). Today, health officials recommend HPV vaccination for all girls, and it has recently become understood that the vaccination can also benefit men. However, between the publication of zur Hausen's hypothesis in 1976 and the development of a vaccine in 2006, many years were wasted before this life-saving discovery was accepted and applied. Only after a substantial number of argumentative steps and extensive correlative studies between virus infections and cervical cancer was the ubiquitous hypothesis that cancers are not caused by infectious diseases defeated. This stands in stark contrast to the most famous cause of cancers: mutations in oncogenes. For those cancers, almost every gene mutated in a significant proportion of patients is considered a potential drug target, often raising the attention of multiple pharmaceutical companies that develop their pipelines in parallel. This results in several treatment options for many putative targets being studied in clinical trials in parallel, and therefore before it has been proven that any particular approach

is suitable to prolong the patient's life or improve her quality of life.

In contrast to the above non-parsimonious explanation, the scientific community has fewer acceptance requirements for hypotheses that match their preexisting views. When Koch first proved that *Bacillus anthrax* causes anthrax, he only needed two argumentative steps: first, he established that the microorganism was present in all patients but not in healthy individuals; second, he showed that he could cause the disease with the pure, propagated microorganism (Perović & Sikimić under revision).

In conclusion, we can identify some general criteria for evaluating hypotheses about disease-causing agents, but we cannot find regular principles like the conservation principles guiding discoveries in physics. Additionally, scientific progress in the study of disease-causing agents does not reach a fast and reliable consensus, because the acceptance of an unexpected or non-parsimonious hypothesis is usually slower, despite the bigger impact of that hypothesis in the long run. The problem is not the absence of data, but the disagreement in the research community on the relevance of individual studies. The citation data might partially reflect this division, but we lack an objective output parameter with which we could judge the efficiency or inefficiency of projects. Outsider ideas that are not cited might seed the development of life-saving approaches, while well-connected researchers promoting established but outdated ideas publish their papers and reviews in the highly cited journals.

Chapter 6

Empirically calibrated agent-based models

In the previous chapters we focused on the optimization of scientific inquiry from an external perspective, by looking at the properties of different research groups and analyzing several specific discoveries in life science. Now we will turn to a third approach to analyze the scientific endeavor: agent-based models of group dynamics. Agent-based models rely on computer simulations and are used to identify the properties of a network. They can be used to identify solutions by simulating all possible scenarios (Ormerod & Rosewell 2009). In every agent-based model individuals or groups are called agents; they can have different properties (e.g., knowledge, beliefs) and can be connected with the other members of the model. The simulation is typically run for a finite number of rounds – phases in which the properties of the agents change, depending on the properties of the other agents. This process is repeated until a clear picture emerges. Because certain parameters (e.g., the initial beliefs or the updating process) contain a stochastic element, the simulations can provide general arguments over a large variety of conditions.

Agent-based models are relatively novel, because they require sufficient computa-

tional power to model complex phenomena. They are successfully used in many different scientific fields to test hypotheses or discover novel properties. They can be used to model the spreading of diseases in epidemiology (Kumar et al. 2013), predict the development of a market in an economy (Farmer & Foley 2009), or study efficient communication structures of researchers in social epistemology of science (Grim 2009, Zollman 2007, 2010).

Zollman (2010) modeled the effect of cognitive diversity on the scientific process and found that transient diversity is most beneficial. Transient diversity means that scientists explore diverse hypotheses for a while but not too long before they reach a consensus. Transient diversity can be reached if scientists start with strong individual beliefs, or if the communication between scientists is limited (Zollman 2010). Grim (2009) concluded that, for some scientific questions, it is fruitful to structure the researchers in loosely connected networks, like the ones dominating the 17^{th} century. Those observations challenge the common opinion that scientific exchange between researchers supporting rivaling theories is beneficial (Longino 2002). Therefore, these models have led to intensive discussions among scientists, e.g., (Borg et al. 2017, Rosenstock et al. 2017).

Agent-based models can also be used to detect novel phenomena. For instance, as mentioned in section 1.1, a group anchoring effect was discovered by Hartmann & Rafiee Rad (2018). They observed that the first speaker in a deliberation process disproportionately influences the beliefs of all group members: because the first speaker talks before anyone else has voiced their opinions, she influences them. Moreover, Frey & Šešelja (2018*a*,*b*) argue that it is desirable for abstract agent-based models to pass a robustness test under parameter and assumption changes; according to the authors, robust models can have explanatory value.

Abstract agent-based models can provide good general arguments in favor of a hypothesis, but in order to apply them to specific questions, empirical calibration is

beneficial. With the help of data we can test our model, adapt it to fit typical output data, and understand how big the impact of these parameters is. In physics, we can turn to the extensive data comparing CERN with other laboratories as well as comparing the CERN accelerators with each other by Martin and Irvine (Martin & Irvine 1984, Irvine & Martin 1984b, Martin & Irvine 1984b) and to the analysis of Fermilab projects by Perović et al. (2016). Both studies used statistical analysis and data-driven methods to identify efficient projects and the parameters associated with them.

With the help of similar datasets, we can calibrate agent-based models to specific questions and use them to optimize scientific approaches. Furthermore, we can extend the reach of the data-driven analysis to hypothetical scenarios. Finally, methods to empirically validate agent-based models are increasingly discussed in economics, e.g., (Fagiolo et al. 2006).

There are various data sources for empirical calibrations: data can be gathered via qualitative interviews, quantitative studies, mixed methods, etc. For example, qualitative interviews can be used to gain a better understanding of the communication structures within individual research groups as well as in the specific fields. Furthermore, external data repositories can be very useful. Citation patterns can inform us about the impact of individual projects and about the flow of knowledge within a community. Structures of scientific groups can be made available in repositories. For instance, Harnagel (2018) offered an agent-based model enriched with the data on citations and peer review and argued in favor of randomness in science funding.

Empirically calibrated models can also be used to simulate the internal processes of research groups and provide arguments to support the robustness of data-mining results and qualitative interviews. We will provide two examples to illustrate this. First, we will model the scientific exchange of scientists in high-energy physics. With the group structures analyzed by Perović et al. (2016) we show that agent-based models reach the same conclusion, supporting the observation that a smaller number of researchers as

well as the separation of the researchers into as few groups as possible is beneficial. Such groups can agree on an interpretation much faster than researchers separated into many different groups; thus, this structure promotes efficient work and fast publications.

Next, we will turn to experimental biology. We use qualitative interviews to extract typical team structures and the effect of group performance as assessed by the interviewees. We will model similar structures and can show that our models reproduce the observations.

6.1 Empirically calibrated agent-based models

Models of scientific inquiry often rely on idealized scenarios. Clearly, the only complete model would be the real world. However, when we want to understand the impact of any specific parameter, we need a minimal model that accounts for it. A more complete model might only camouflage the effect of this parameter. On the other hand, network structures commonly used for modeling, such as wheels and circles, do not necessarily resemble the typical communication between scientists in different fields. Thus, we have to turn towards a data-driven approach. The data on group structures, from the high energy physics laboratory Fermilab, show that scientists are usually organized in several teams and members of the same teams communicate more with each other than with members of other teams. See Figure 6.1 for an illustration of typical group structures. This illustrates that we should model scientists as groups. Furthermore, links are usually not equally strong and not necessary bidirectional; for example, the belief of a professor tends to influence the student more than the belief of the student influences the professor.

We developed an empirically calibrated agent-based model that can account for these effects and simulated different aspects of scientific inquiry based on observed team structures. Firstly, from the data on team structures in the HEP laboratory Fermilab,

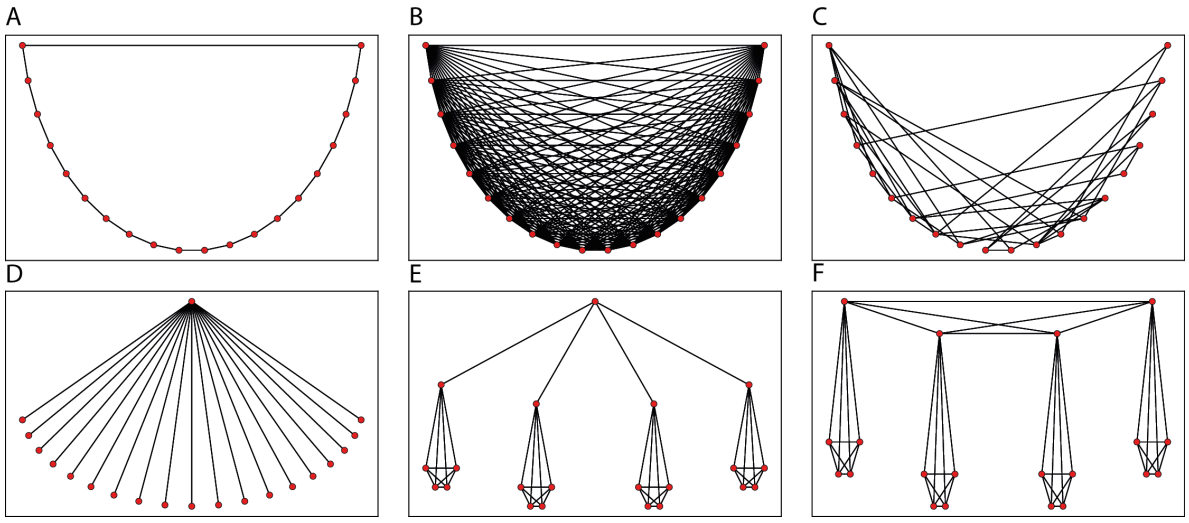


Figure 6.1: **Typical and data-motivated group structures.** Red circles represent agents, black lines communication. A-C) Group structures typically seen in the literature. A) Circle: all agents are connected only with neighbors. B) Completely connected network: every agent is connected with all other agents. C) Random network: every agent is connected with two other agents at random. D-F) Group structures inspired by data. D) Centered group: all agents communicate with one head. E) Hierarchical group: the leader is communicating with several team leaders who in turn communicate with their students. The students in addition communicate with each other. F) Interacting group leaders: four group leaders communicate with each other and with their students; in addition the students communicate with each other. To account for hierarchy levels in the model, agents of every group and hierarchy level are always oriented in a half-circle.

we chose an approximation of how many members research groups had and we also noted that the scientists were often divided into smaller teams. For this purpose, we used the group of projects analyzed by Perović et al. (2016), who showed that smaller teams perform better than bigger ones.[1] Just as in Perović et al. (2016), as team members

[1]These data are available in the INSPIRE-HEP repository (https://inspirehep.net/).

we considered scientists working on a project, without their helping stuff. Thus, we are able to check whether the results of the simulations concur with the previous results by Perović et al. (2016).

For these simulations every scientist or group of scientists is modeled as an agent. This agent has different properties and beliefs and is connected with the rest of the group and the scientific community. The communication network, their prior beliefs, and the updating mechanism define the outcome of the simulations. In every case, we will observe the changes in the beliefs of the individual agents. As starting condition, every agent has a belief normally distributed around the undecided state. They update their beliefs after talking to every connected scientist. During each round of information exchange, the receiving agent updates her belief by adding a fraction of the belief of the communicating agent. We opted for small changes, because scientists most likely do not communicate all their beliefs in group meetings, but only fractions. For example, one strongly convinced agent with the belief 0.9 increases the conviction of the receiving agent by 0.009. Thereby, the belief of the receiving agent can exceed the belief of the speaker. The communication does not have to be two-directional, because we want to be able to simulate asynchronous communication, e.g., weekly presentations. This process repeats itself, and after several rounds these opinions become fixed. In the figure, we can follow the change in the individual beliefs from a value around 0 to 1 or -1. We simulated this process 1000 times; summaries of how fast the groups reached a consensus are provided in the histogram in Figure 6.2.

Two types of group structures dominate the data from Perović et al. (2016). Efficient project groups were usually small, involving about two teams with few members in each team. By contrast, inefficient groups were much larger, involving several institutions (about eight) with a high number of members each (about seven). We illustrate these two group structures in Figure 6.2. Circles represent the individual researchers (agents) that are communicating with the other members in the group as well as with the head

of the group. The interactions are indicated by lines. How the communication and reasoning work within these groups is illustrated in Figure 6.2. We consider two different models: a standard model where every agent is updating her belief rationally based on the beliefs of her peers, and an improved model where the sunk cost bias is considered. Here agents only change their belief if the evidence for the opposing view is significantly better than for their original belief. In the model, this is simulated by a threshold of 0.1. An agent previously disagreeing with a given hypothesis will not immediately change her opinion when the evidence for the hypothesis is objectively larger than against the hypothesis, but only when the evidence is at least 10% larger. In Figure 6.2, we show the beliefs of every agent in three simulations for each condition over 300 rounds. Every simulation is color-coded. The final belief of all agents in the first simulation (blue) in the sunk cost bias model is also included in the group structure. Agents believing in the hypothesis are colored green, while agents who disagree are colored red.

The visualization of the models highlights interesting aspects. In Figure 6.2, we can observe subgroups that are forming an information bubble that opposes the majority opinion, similarly to the tonsillectomy example discussed in section 1.1. In the bottom-left chart in Figure 6.2, we see three subgroups, colored in red, which reached the opposite consensus from the rest of the scientists. This is possible because scientists within subgroups, i.e., research teams communicate more with each other than with the rest of the group members.

The new results agree with the data-driven study by Perović et al. (2016). Indeed, the results of the calibrated simulations show that small research groups, which are divided into a few teams, reach the consensus faster than bigger research groups with a larger number of teams (Figure 6.2).

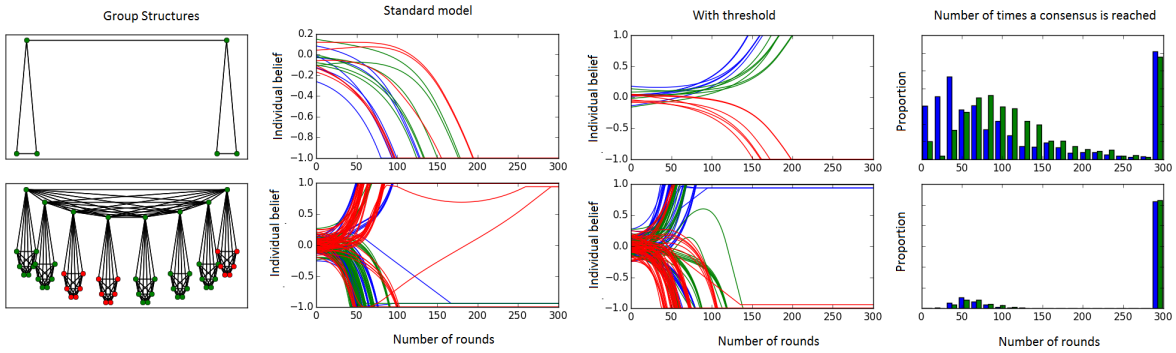


Figure 6.2: **Empirically calibrated models of communication in HEP.** Left: analyzed group structures; agents are represented as red or green dots, connected with lines. The color of the agent represents her final belief in the first simulation of the model accounting for the sunk cost bias. Second column: the beliefs of the individual agents in three simulations (each simulation in its own color) in the standard model. Third column: the beliefs of the agents in a model accounting for the sunk cost bias. Right: histogram showing how fast the agents reach a consensus (green with and blue without sunk cost bias). The results are robust under both modeling conditions.

6.2 Empirically calibrated agent-based models of communication in biology

The number of researchers in a scientific project varies in different fields. In experimental biology, research groups are generally smaller than in HEP. They are also usually hierarchically structured and can contain several layers of hierarchy. We analyzed three management styles: centralized, groups with two levels of hierarchy and decentralized groups (Figure 6.3).[2]

In centralized groups only one professor communicates and influences all junior group members. On the other hand, junior researchers do not have the opportunity to

[2]The inspiration for these structures came from qualitative interviews and laboratory investigations.

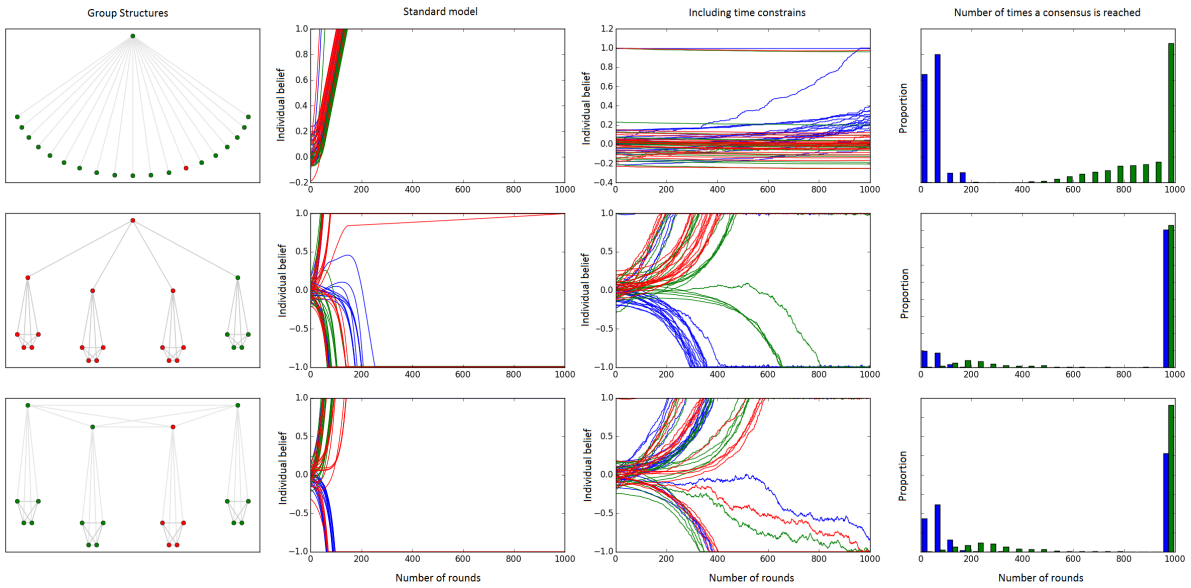


Figure 6.3: **Empirically calibrated models of communication in experimental biology.** On the left side, the analyzed group structures are displayed (the color of the agent indicates her final belief in one simulation). Second column: individual beliefs of the agents in three color-coded simulations of the standard model. Third column: individual beliefs in the model with the constrained communication time. Fourth column: speed of reaching a consensus. The agents reach the consensus much faster when they all communicated in each round (blue) than in the model with time constrains (green). This is particularly evident for the centralized groups.

work together. Such a situation was brought up by one of the interviewed biologists. The second analyzed structure is the one with two levels of hierarchy. In such a group professor communicates with several experienced researchers, e.g., postdoctoral research fellows, who in turn supervise PhD students. This is another typical management style in larger biological groups.

The last network represents a group of four group leaders with small groups. The group leaders communicate closely with each other in a complete network. In addition,

in the last two networks the PhD students are communicating with their peers. In these simulations, we also included time constraints: agents with many connections communicate less frequently with their individual connections. For example, a professor who directly supervises twenty PhD students, will not have time to talk to every single one of them each week. We simulate this time constrains with a probability function based on the number of connections of the agents. In each round, two connected agents only communicate with each other with a probability of one divided by the number of connected agents. We use the highest number of connected agents, because we consider that the agent with most connections, e.g., the group leader, is least likely to find the time to communicate with every individual group member. In our example, a student has only a 5 % chance to communicate with the professor in each round, even when she would like to interact with her supervisor more frequently.

As expected, we observe the strongest effect of the introduction of time constrains in centralized networks. In the standard model, groups converge on a consensus very fast, because all communication happens over the central node. However, this advantage disappears fast when we limit communication with strongly connected nodes. In this scenario, groups stay undecided for long time periods and individual group members stray in the wrong direction. The impact of this scenario becomes particularly interesting when we compare the efficiency of all group structures in the first 300 rounds. When we do not consider time constraints the centralized group converges fastest, but otherwise both hierarchical ones perform better. Furthermore, the interconnected groups perform better than the four groups connected with a single hub. This supports the policy of separating groups into smaller teams and ensuring the best possible communication between the members. Further work will focus on the effect of external evidence from unconnected members (e.g., via publications) and the impact of better communication between junior members from different groups (e.g., via conferences).

These examples show that our agent-based models can reproduce effects observed in

the data of (Perović et al. 2016) and our qualitative interviews. Our examples highlight the impact of the choice of group structure. They can be used to assess the likely effect of changes in management styles.

Chapter 7

Argumentation patterns in life science: a study of pathogen discoveries

We discussed and analyzed optimization methods in HEP and experimental biology based on external data. We pointed out that most subfields of experimental biology are not perfectly suitable for analyses based on citation metrics, because of its non-parsimonious nature (section 5.2). Thus, instead of analyzing experimental biology based on external data, we investigate argumentation patterns by analyzing internal data. As we illustrated in section 6.2, empirically calibrated models can be applied for the optimization of scientific inquiry in the field. Analyzing the scientific process, qualitative and quantitative interviews, as well as citation patterns, are valuable sources of data for calibrating such models. In the following chapter, we switch to the analysis of the internal data on pathogen discoveries. We focus on examples of pathogen discoveries that are relevant for social epistemology of science because of their great impact in the field of life science, and because of the complex argumentative exchange among experts that lead to them.

We are mirroring parsimony with the number of argumentative steps needed for accepting a hypothesis. The greater the number of argumentative steps, the less par-

simonious is the hypothesis. A simpler hypothesis requires fewer steps to be justified than a more complex one, provided that a simpler hypothesis needs to take into account fewer parameters. For instance, in the 17th century, the heliocentric hypothesis required fewer parameters than the geocentric one, which used eccentrics, epicycles, deferents, and equants (Weinert 2008).

We will discuss three breakthrough results in pathogenesis research that were awarded a Nobel Prize. The first is the discovery of protease-resistant protein (PrP); the second is the discovery of the relation between peptic ulcer and *Helicobacter pylori*; and the third one is the finding that viruses can participate in causing cancer, specifically, we will address the case of the Human papillomavirus (HPV). PrP causes several infections degenerative diseases of the nervous system, such as Creutzfeldt-Jakob, bovine spongiform encephalopathy, and Scrapie disease. *Helicobacter* and HPV are each responsible for about 5% of cancer associated deaths worldwide. All cause life-threatening diseases, but studies supporting the claim that they are disease-causing agents were neglected by the scientific community for long periods of time, resulting in unnecessary delays in the understanding and treatment of the diseases. We will analyze why the acceptance of the correct hypothesis was delayed and how a plurality of approaches can help to elucidate complex mechanisms.

We focus on these cases because findings of disease-causing agents are usually only accepted if the discovery follows Koch's postulates – even though it was necessary to adapt them to account for the novel mechanisms discovered by Stanley Prusiner (prions), Barry Marshall and Robin Warren (*Helicobacter*), and Harald zur Hausen (HPV).

7.1 Koch's postulates

Koch's postulates have been very valuable for establishing the causality between an infectious agent and a disease, especially in the 19th century after the role of bacteria in the development of diseases was understood. We follow the standard formulation of the postulates, as presented in (Tabrah 2011):

"1. The organism must be shown to be invariably present in characteristic form and arrangement in the diseased tissue.

2. The organism, which from its relationship to the diseased tissue appears to be responsible for the disease, must be isolated and grown in pure culture.

3. The pure culture must be shown to induce the disease experimentally.

4. The organism should be re-isolated from the experimentally infected subject." (Tabrah 2011) p. 144.

The postulates are illustrated in Figure 7.1. The pathogen is first isolated from sick individuals (1), grown in pure culture (2), used to infect an individual (3), and re-isolated from this new individual (4).

Even though the postulates were widely accepted and a helpful guideline for establishing causality, scientists had severe problems to show every single point for most diseases and modified the guidelines over time. From a contemporary perspective, each postulate can be challenged. The discovery of viruses, which require a living host and therefore, cannot grow in pure culture, led to the modifications of Koch's postulates. Furthermore, already Koch himself realized that the third postulate might be too strict, because he and many others failed to fulfill it for pathogens which were strongly associated with diseases, e.g., like the *Vibrio cholerae* bacterium with cholera. The first postulate was also challenged by new findings. Firstly, not every infection leads to disease symptoms: in a study in 1955 with infants, Huebner found viruses in infants, who

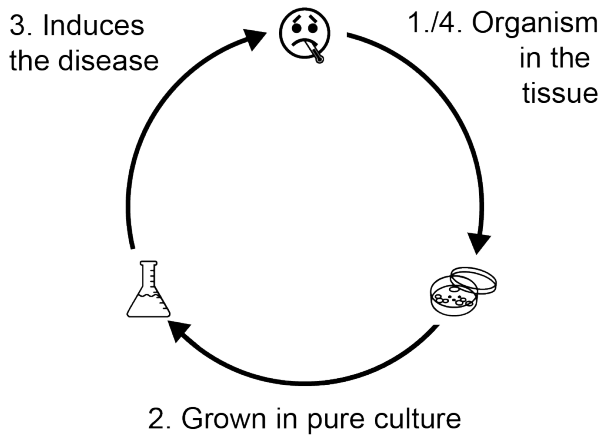


Figure 7.1: **Illustration of Koch's principles.**

were not showing any symptoms (Evans 1976). Furthermore, it was recognized that some diseases such as acute respiratory syndromes, could be caused by several different agents, while other diseases, such as swine influenza, require the synergistic action of two agents (a virus and a bacterium). Thus, in the sixties it was recognized that the nature of the agent, as well as the status of the host, play a role in the development of diseases (Evans 1976).

The fourth postulate, requiring the re-isolation of the organism was added later on (Tabrah 2011). Interestingly, this was more likely fulfilled by the famous French microbiologist Louis Pasteur. He had a dispute with Koch who first demonstrated causality in the case of anthrax. While Koch argued that his isolation in the year 1876 was sufficient and that the demand to isolate bacilli from every contaminant is impossible, Pasteur argued that the only conclusive evidence of causality was the passing of the organisms to successive animals and cultures (Carter 2003). Because the passing through successive animals requires the re-isolation, Pasteur's evidence would more closely follow what later was called Koch's fourth postulate. Koch's postulates became received views on pathogenesis. Even though Koch himself noticed the limitations of his methodology and admitted that the third postulate cannot always be fulfilled, many

researchers followed his work blindly (Rivers 1937). Thus, Koch's postulates can be considered as dogmas which even hindered the research on viruses (Rivers 1937, Keyes 1999).

Based on Koch's first postulate, detecting a high percentage of microorganisms in sick organisms in opposition to healthy ones, is sufficient for establishing the correlation between an infectious disease and a microorganism. For instance, in the case of anthrax, Koch detected bacilli in sick patients, which he used as evidence for the correlation between the bacilli and the disease. More importantly, according to Koch's second postulate, in order to show that there is a causal connection, i.e., to show that an organism causes the infectious disease, the organism has to be grown in pure culture and subjects should be infected with it. He successfully performed this task for anthrax, tuberculosis, and many more. From the external structural perspective, we are talking about four argumentative steps that show the correlation and the causal connection between a microorganism and an infectious disease.

7.2 Misfolded proteins as infectious agents

Prion diseases are neurodegenerative infectious diseases affecting different mammals. The infamous bovine spongiform encephalopathy, commonly known as mad cow disease, belongs to this group. It had an outbreak during the '90s in the United Kingdom and it is responsible for a variant of the Creutzfeldt-Jakob disease – a human version of a prion caused neurodegenerative disease (Will et al. 1996). Prions as disease-causing agents violate Koch's second postulate that all infectious diseases can be grown in pure culture.

Prions do not propagate by an internal mechanism, they cannot grow or divide on their own, they are misfolded proteins inducing the misfolding of the same protein in the brain of the host. This triggers a chain reaction causing a range of diseases,

most famously the mad cow disease and the Creutzfeldt-Jakob disease. Koch did only know bacteria as infectious agents and was fully aware that it might be difficult or even impossible to grow an agent in pure culture as we have seen above.

Even though the need of viruses for a host cell was well understood in the eighties when Prusiner published his first discoveries, the scientific community was reluctant to accept his findings. Prusiner found it hard to convince the scientific community that a protein can be responsible for an infectious disease and in took in total 38 years for the protein hypothesis to get accepted (Soto 2011). The hypothesis was initially published in 1967 and received strong support in 1997 when Prusiner won the Nobel Prize. However, only in 2005, when infected mice with in vitro generated prions showed the symptoms of the disease (Castilla et al. 2005), the scientific community was convinced about a novel disease-causing agent and refuted Koch's second postulate.

Koch's dogma is more general and simpler than the hypothesis that infectious diseases can also be caused by something else than a self-propagating organism. Therefore, in order to accept a simple correlation between the protein and an infectious disease, Prusiner and his team made six argumentative steps (Prusiner 1982, Soto 2011). First, they showed that the disease-causing agent does not react to treatments that destroy nucleic acids. This was very surprising, as all infectious diseases known so far were caused by viruses or bacteria which require intact nucleic acids. Second, Prusiner and his team showed that the disease-causing agent could be killed by a protein destroying treatment. With filtrations, they further demonstrated that the disease-causing agent is as small as a protein. Another observation concerned the first postulate. The problem was that the prion protein was also present in healthy organisms. Therefore, they assumed that the disease-causing protein had some special feature – a different folding.

Finally, scientists expressed the protein with a different folding and were able to infect mice with it. These mice showed symptoms of the disease (Castilla et al. 2005, Soto 2011). The dogma that all infectious diseases are caused by bacteria or viruses was

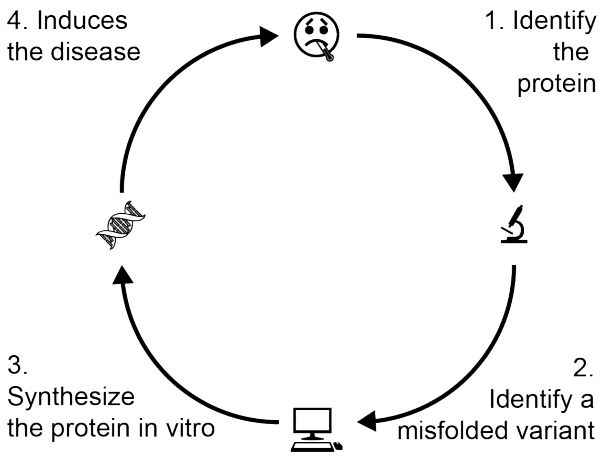


Figure 7.2: **Illustration of the discovery of prions.**

refuted because of these experimental results. The adjusted scheme of Koch's postulates for prions summarizes this (Figure 7.2). Prions were detected in sick individuals (1), a genetically encoded variant which misfolds autonomously was identified (2), this variant was synthesized in pure culture (3), and used to infect mice (4).

These argumentative steps, together with the previous four steps that showed the correlation, are widely accepted by the life science community as evidence that PrP causes an infectious disease. Also, the fifth result (expression of the protein) is analogous to the requirement of the second Kochs principle, i.e., growing an organism in pure culture, while the sixth result corresponds to Koch's third postulate of infecting the subjects. Thus, the difference between parsimonious explanation and the non-parsimonious one is in showing the correlation between two factors.

While in the case of an expected and uniform solution, i.e., Koch's solution, the correlation step is immediate, an unexpected discovery requires more testing before the hypothesis becomes reasonable from the perspective of a scientific jury. This might be a justifiable acceptance requirement. However, the epistemic worthiness of pursuing a hypothesis should be evaluated in a different way. We turn to this question later.

From a formal point of view, we can identify six argumentation points. On the

other hand, discoveries that follow Koch's postulates (e.g., the observation that a specific organism is responsible for a disease) are much quicker accepted by the scientific community. For example, the scientific community accepted Koch's evidence in favor of a bacillus causing anthrax after just two arguments. Firstly, he demonstrated the correlation between the presence of *Bacillus anthracis* and the disease, i.e., the presence of bacillus in sick tissue. Secondly, in order to argue for the causal relationship, he was able to induce the disease in healthy organisms.

7.3 The discovery of *Helicobacter pylori*

Helicobacter pylori causes gastric cancer and is the second most common cause of cancer-related deaths, killing about 700 000 people worldwide per year. Its manifestation usually follows the development of peptic ulcer. For decades scientists and medical practitioners focused on acid disbalances in the stomach as the cause for ulcer, treating it with drugs reducing the acidity to reduce the symptoms. The discovery that *Helicobacter pylori* is responsible for peptic ulcer and gastric cancer was not hampered by problems to grow the agent in pure culture (Koch's second postulate) but by difficulties showing *Helicobacter* in the diseased tissue (Koch's first postulate).

After the long battle between two hypotheses, one stating that the peptic ulcer disease is caused by stomach acidity and the other stating that the main cause of the disease is a bacterium, Warren and Marshall received the Nobel Prize 2005 for their results in favor of the latter. During the course of years the bacterial hypothesis had several results in its favor. However, the possibility that bacteria could survive in the stomach was largely disregarded by scientists supporting the acid hypothesis.

Already in the 19th century, scientists observed bacteria-like organisms in the stomach, e.g., Klebs detected bacteria in gastric glands (Fukuda 2002), while Bizzozero registered the presence of spiral organisms in the stomachs of dogs (Figura & Bianciardi

2002), etc. Clearly, these results were not conclusive, but were showing the possibility of bacteria surviving in the stomach, and thus that they may be responsible for the peptic ulcer disease.

At the same time, there were results that supported the acid hypothesis. For instance, Schwarz observed the high acidity in patients diagnosed with ulcer (Fatović-Ferenčić & Banić 2011). And after performing over 1000 biopsies, Palmer published that he did not observe any colonizing bacteria (Palmer 1954). He concluded that bacteria cannot live in the acidic stomach environment. In his investigation Palmer used hematoxylin and eosin (H&E) staining, which was insufficient for noticing bacteria in the stomach. Šešelja & Straßer (2014) claim that Palmer used the suboptimal staining method, even though the research at the time already indicated that this method would not be the most promising one. At the time, there were already publications available showing that silver staining is superior when it comes to the detection of spirochetes in the gastric mucosa (Šešelja & Straßer 2014). In addition, he did not show a positive control, that would be the successful identification of artificially introduced bacteria, nowadays an absolute requirement for a publication showing any negative result. Still, Palmer's research was so influential that the acid hypothesis prevailed for years.

We talk about the primary and most relevant factor for causing peptic ulcer. However, it should be noted that ulcers can have different causes, not only bacterial, e.g., they can be caused by anti-inflammatory treatments, or can be induced by acid (Šešelja & Straßer 2014). Moreover, gastric acid hypersecretion promotes the colonization of ulcer (Malfertheiner et al. 2009). Thereby, acid and *H. pylori* promote the development of the disease synergistically. Also, the healing is accelerated when acid suppressants are given in addition to antibiotics. However, acids are not considered to be an independent cause for ulcer in humans anymore (Malfertheiner et al. 2009). Palmer's research resulted in a dogma that no bacteria can live in the stomach. Because of his results, further investigations in favor of the second hypothesis stopped (Zollman

2010). Still, the previous results showing the presence of bacteria in the stomach contradicted Palmer's dominant results. Moreover, the dogma was built on a paper, which methodology could have been questioned already at the time of publication (Šešelja & Straßer 2014). On the other hand, Palmer tested his staining method in rhesus monkeys and concluded that spirochetes were easily detectable (Palmer 1954). In order to show a correlation between the bacteria and the disease, Warren and Marshall first had to refute Palmer's result. They used different methods in order to mark bacteria in the stomach. Finally, Warren and Marshall were able to grow *Helicobacter pylori* in pure culture and infect subjects, including Marshall himself, who afterwards showed early symptoms of peptic ulcer. This demonstration was in accordance with Koch's second postulate. However, because developing ulcer lasts for years and it is very dangerous, the causal relation has only been demonstrated between the bacterium and the early stage of the disease (Marshall 2001). Thagard (1988) claims that, though Marshall infecting himself with the disease had a large impact in the public eye, the scientific community was convinced only after curing the ulcer by antibiotics.

In this discovery no clear law of parsimony was violated. Instead, we have the case of reasoning that neglected some of the evidence. If all results were given the same weight, Palmer's dogma would have been questioned earlier. In this example, we see the potential benefits of argumentation analysis in life science. Argumentation analysis using the known results gives a different insight than the prevailing scientific opinion at the time. The arguments are presented and evaluated neutrally, e.g., without a field-specific bias or research bias towards simpler explanations. This enables us to make a more realistic evaluation of their strength. Interestingly, when Marshall submitted his report to the Australian gastroenterology society in 1983, the report was rejected, because this scientific community was skeptical about his results. However, he had the chance to present the results at a workshop in the field of microbiology. Here, his results raised substantial interest among bacteriologists (Thagard 1988). This raises

a question which reasons, e.g., epistemic or non-epistemic, shaped the opinion of the gastroenterological community, to rule out Marshall's research as not worth pursuing.

The question of whether a hypothesis is worth pursuing is relevant not only for new hypotheses, but also for already accepted or overthrown ones (Nickles 2006). Nickles (2006) argues that heuristic appraisal considers the potential fertility of a theory, research funding, availability of technical equipment, etc. Assessing how promising a theory is, differs from the criteria for its acceptance, since the prospective values of a theory do not coincide with the ones relevant for the acceptance (Whitt 1992). Examples of the acceptance values are adequacy of the explanation, explanatory power, problem-solving effectiveness, etc., while establishing the prospective values is about assessing how promising a theory is, i.e., whether it is worth pursuing (Whitt 1992). Šešelja et al. (2012) discuss the distinction between practical and epistemic aspects of pursuit worthiness, as well as the distinction between individual and communal reasons for a scientific pursuit. In (Whitt 1992) the claim of asymmetry is supported by the following example. Even though, several theories can be evaluated as worth pursuing, usually only one theory gets accepted as adequate. Practical worthiness concerns factors such as the technical realizability of a research, and social pressure in a given work context. In the presented examples of pathogen discoveries, we abstracted from technical restrictions. However, because forming the novel hypotheses took a substantial time (years), it is plausible to assume that the social pressure played a role in the assessment of their pursuit worthiness. This could be a reason why the gastroenterology community was closed for Marshall's results. (Šešelja et al. 2012) explain the individual pursuit worthiness in terms of a research directive, while the communal pursuit worthiness can be seen as an evaluative stance. The individual assessment of the pursuit worthiness will usually include both epistemic and non-epistemic goals, while the communal assessment should abstract from the individual goals. Instead, the communal assessment should ideally focus on epistemic values (Šešelja et al. 2012). From

the communal perspective, having in mind the complex nature of pathogenesis, it is epistemically beneficial to pursue diverse paths.

7.4 A cancer causing virus

The last discovery of a disease-causing agent we will discuss is the human papillomavirus (HPV). It causes almost all cervical cancer and the vast majority of anal and oropharyngeal cancers, making it responsible for about 5% of all cancers worldwide. Because cancers are mainly caused by a few aggressive subtypes, the knowledge of the mechanism resulted in the development of potent vaccines which have the potential to rescue hundreds of thousands of lives every year.

When zur Hausen's team discovered that the human papillomavirus (HPV) is the primary cause of cervix cancer they also had significant difficulties to fulfill Koch's first postulate, i.e. to identify the pathogen in diseased tissue (zur Hausen 2009). Even though the possibility that a virus can cause cancer in chicken was understood already (and Francis Rous was awarded the Nobel Prize for the research he conducted on this matter in the twenties), the community stayed skeptic towards the idea that it could have such an impact on human health. It was still not considered that human cancers can be triggered by infections. Therefore, showing a simple correlation between the disease-causing agent and the disease took a substantial number of argumentative steps. The uniform hypothesis that cancers are not infectious was defeated by a non-parsimonious one, which accounts for exceptions. Zur Hausen found that 70% of cervical cancers contained either the HPV subtypes HPV 16 or HPV 18. Unfortunately, the scientific community was not convinced when his group presented the work in 1983 on the Second International Conference on Papilloma Viruses, in part because of the language barrier during the questions (Cornwall 2013). Later an important convincing factor was the replication of the results by Peter Howley with whom they shared their

samples early on (Cornwall 2013). Zur Hausen's final triumph was the development of a life-saving vaccine.

Especially when new claims attack received views, the community values the independent replication of the results very highly. Still, it should be noted that before the development of the vaccine, zur Hausen's research was based on correlation.

We understand the principle of parsimony as a useful method for finding adequate theories (Kelly 2004, 2007). For instance, the parsimony principle is a base-line principle in phylogenetics, which states that 'if all other parameters are the same, the best explanation is the one that makes the smallest number of evolutionary changes necessary'. This method is efficient for the reconstruction of phylogenetic trees as discussed in chapter 5. Also, in many cases Koch's principles were efficient methods for identifying disease-causing agents. Yet, the discoveries of some disease-causing agents, did not concur with the simple and general principles, and thus required testing diverse hypotheses. Though it is reasonable that unexpected discoveries require more testing before being accepted, the simplicity of a hypothesis is not equally epistemically beneficial in every research field. For instance, the application of the parsimony criterion to philosophical questions was criticized by Huemer (2009), because it is often unclear which philosophical view is more complex, using as examples the nominalism/realism and physicalism/dualism debate. Longino (1996) argues in favor of heterogeneity in economics, because it is beneficial for a feminist epistemology. For instance, a heterogeneous ontology of a household makes gender relations more visible. Longino goes further and argues that the simplicity of a theory cannot be a purely cognitive value, unless we assume the simplicity of the universe. While phylogenetics can be regarded as a theory with a high degree of uniformity, pathogenesis does not experience the same degree of regular behavior. The presented examples support this claim. Moreover, disease causes can often be cofactorial (several interconnected factors at the same time) and multifactorial (several independent factors). From the perspective of the scientific

community working on pathogenesis it is epistemically beneficial to pursue diverse and complex hypotheses.

7.5 Parkinson's disease and bacteria

Parkinson's disease is a neurological disorder that was initially described by James Parkinson in the 19th century. The disease is caused by a poorly understood complicated interplay of genetic and environmental factors (Kalia & Lang 2015). Moreover, the symptoms vary and at least two subtypes of the disease can be defined: tremor-dominant and non-tremor-dominant Parkinson's disease. The clinical diagnostics is based on motor performance tests but reaches only a sensitivity of 90%. The gold standard for diagnosing the disease is still the observation of Lewy bodies in the brain (which is done post-mortem). It is still unclear if the Lewy bodies are only correlated with the disease, or if they are part of its cause.

The risk of developing Parkinson's disease is considered to be multifactorial. Environmental risk factors include pesticide exposure and prior head injury; furthermore, some dozen gene loci have been implicated in Parkinson's disease (Kalia & Lang 2015). Moreover, some environmental and genetic factors act synergistically. Some novel results indicate a correlation between Parkinson's disease and bacteria in the intestines (Sampson et al. 2016). These results demonstrated a role for gut bacteria in a complex neurodegenerative disease. The research was conducted on mice that were genetically modified to be more prone to the disease. Sampson and his team showed that mice growing without contact with microbiota rarely develop the disease. Moreover, the chance that mice treated with antibiotics develop the disease is low. Finally, the reintroduction of microbiota or their metabolites triggers the disease.

These results represent only initial steps in discovering which pathogen factors influence Parkinson's disease. For instance, the results cannot pinpoint any particular

gut bacterium. The remaining logical possibilities of the disease causes are numerous, e.g., a combination of gut bacteria might be responsible for the disease, or combined genetic factors and bacteria might be responsible, etc. As in the example of oncoviruses and *Helicobacter pylori*, the prevailing hypothesis was that Parkinson's disease is not infectious. However, a diverse approach in hypotheses that includes testing the bacterial hypothesis is a promising path for exploring complex causal patterns responsible for neurodegenerative diseases, which remains enigmatic until today. Undoubtedly, the diverse hypotheses in this domain are epistemically worth pursuing. Moreover, the graphical representation of the results, both the ones based on the genetic and the ones based on the bacterial hypothesis, might be helpful for examining the plurality of factors that neurodegenerative diseases most likely have, since it can help in understanding their interplay, as well as avoiding potential reasoning mistakes that might occur in the complex argument exchange. Here, we only indicate this possibility.

7.6 Summary

As we have discussed in chapter 5, citation metrics of projects in experimental biology can be misleading. Therefore, we focused on analyzing the internal data of particular projects to study the factors promoting or impeding scientific discoveries. Internal data on important results can help us understand scientific reasoning and belief formation. A further use of analyzing case studies in experimental biology can be gathering data for an empirical calibration of formal models, e.g., for representing epistemic landscapes.

The presented Nobel Prize winning discoveries had a great impact on science and medicine. Yet, these discoveries required novel approaches and followed a unique path. By analyzing them we can point out similarities and differences in the scientific approach to pathogen discoveries that were defeating received views.

In the case of prion diseases, the general received view has been directly refuted by a

single-case hypothesis, demonstrating that the simplicity and generality of the received view did not lead to its accuracy. In the case of the discovery of *Helicobacter pylori*, using argument reconstruction, we have demonstrated that some evidence was neglected. Furthermore, it is important to question received views and the interpretation of old results like the ones from Palmer, which were not performed as stringent as it would be expected in contemporary biology. When considering the epistemic goal of finding the most adequate disease cause, the investigation of diverse hypotheses, even if they appeared fringe at first, is beneficial. Thus, the novel approach, which explores the connections between microorganisms and Parkinson's disease seems worth pursuing, and it might be a part of a more complex explanation that will require complex scientific argument exchange.

Chapter 8

Benefits and limitations of data-driven analyses

For the purpose of social epistemology of science data can be collected both from the scientists themselves discussing their experiences, and from external observations about their scientific practice and outputs. These data need to be curated and interpreted carefully. Finally, adequate algorithms should be applied for their processing.

8.1 Data collection

We are witnessing a shift in philosophy of science towards naturalization in the context of experimental philosophy (Daly 2010, Knobe & Nichols 2017). This naturalization means that data are relevant and informative for the philosophical endeavor. Quantitative and qualitative research represents a valuable data source when it comes to the optimization of scientific reasoning. Quantitative research involves different types of surveys and questionnaires. The main benefit of quantitative analyses is that the obtained results should be statistically justified. With qualitative interviews, however, one can detect topics that scientists find most problematic and bring up themselves (Hangel

& Schickore 2017, Wagenknecht et al. 2015, Wagenknecht 2016). Qualitative research is thus a great source for detecting values in science, challenges that scientists face, or assessing their work conditions. Moreover, when interviewees are limited to filling in an online questionnaire they cannot elaborate further on their answers. We cannot track their reactions or ask them additional explanatory questions. The most comprehensive way of gathering results is using the mixed-methods which combine quantitative and qualitative techniques in various ways. For instance, after the analysis of survey results, qualitative interviews can be conducted to get deeper insights into the answers obtained in the survey. Comprehensive studies, which make use of mixed-methods, represent a rich data source for further optimizations of the scientific environment. After understanding what scientists need and complain about, we are in the position to improve their work conditions.

A different approach to data collection is the creation of online repositories with external data such as projects' structures and resulting papers. All these data require curation. Moreover, there is a question of how much data should be made public without invading the privacy of individual researchers.

An additional problem represents the fact that private companies are not necessarily motivated to share their research data. Yet, this would be very valuable and it is important to encourage data availability. For this purpose, intermediate solutions can be found. For example, in order to motivate pharmaceutic companies to share their data, the European Medical Agency (EMA), the institution responsible for the approval of drugs in the European Union, invited drug developers to discuss confidential information about their Alzheimer's disease programs. Under this protection, the companies were willing to share private knowledge and helped the EMA to improve their regulation and help in the design of clinical studies. This approach is far away from an open science approach but it at least allowed a limited exchange of ideas which already helped the participants and will hopefully lead to faster development of effective

treatments (Alteri & Guizzaro 2018). A similar approach could be applied to facilitate the gathering of the external data about research in industries.

8.2 Practical and theoretical consequences of data availability

When we work with real world data we have to be careful. First, we must evaluate the quality of the data. Second, we must protect the privacy if we work with data which can be associated with individual people.

The quality of the data might be compromised for several reasons. For example, when data is collected only from a few sources, they might be biased towards special institutions or people. Furthermore, they can be influenced by other factors such as the historical context. When the data is collected over longer time periods the meaning of some properties might change (e.g., the value of PhD titles or author list standards) (Sikimić 2017).

Data privacy is another concern. To make research transparent and reproducible, the sharing of data is indispensable. However, when we use real world data, also from public resources, we have to consider negative consequences for the people involved. For example, when we analyze research projects and highlight some as inefficient, people involved in those projects might fear negative consequences. It is our task as researchers to exclude this possibility, for example, by encoding personal data (Sikimić 2017).

Also, examining projects after a time distance has smaller practical implications on the academic careers of the involved scientists. Perović et al. (2016) performed the analysis on data from the eighties and nineties, to be able to estimate their long-term impact; they did not influence the career perspectives of the researchers involved. In addition, they were able to draw general conclusions, which could allow a more efficient distribution of resources and can improve the knowledge acquisition in science. A

responsible and transparent science policy increases both the epistemic efficiency and support by the scientific community.

8.3 Limitations of data-driven models

The empirical calibration of models of scientific inquiry and interaction is beneficial in order to understand the scope of the model, i.e., to understand in which cases this model is applicable. A complete model would be too complex to be useful. However, a model has to be able to reproduce the most important aspects of the process it is supposed to represent. For example, when the observations show that large groups are less efficient, the model has to show the same trend. Hypothesis-driven models can sometimes be questioned from the methodogical side. Alexander et al. (2015) questioned the results of Weisberg & Muldoon (2009) and pointed out some problems in the algorithms. Rosenstock et al. (2017) questioned the results from Zollman (2007, 2010), arguing that the results hold only under certain assumptions.

An advantage of data-driven models is the clear reach of the findings. While hypothesis-driven research can be used to support general conclusions about optimal team structures, e.g., (Kitcher 1990, Zollman 2007), data-driven models unambiguously apply to certain scenarios. However, one has to keep in mind that a data-driven model will most likely only make meaningful predictions about very similar scenarios. It can only be applied to the scenarios for which data are available. In fact, the unavailability of a sufficient amount of high-quality data is one of the main factors currently limiting the wider introduction of data-driven models.

Further problems hindering the development of data-driven models include different biases. These included sampling biases, cultural biases, measurement biases and algorithm biases. A sampling bias occurs when the training data is not balanced (e.g., in terms of ethnicity) (Zliobaite 2015). Similarly, the cultural bias is observed when

we train an algorithm on biased data. In this case, automatically derived semantics will reflect the same biases as humans. For example, machine learning algorithms can associate certain occupations with a specific gender, if the training data was biased (Caliskan et al. 2017). A measurement bias occurs when the device or method used for the generation of the data is suboptimal. In order to overcome it, one has to consider the sources of the training data. For example, the tumor staging methods in radiology and pathology do not always reach the same conclusion (Anderson et al. 2017), thus using the results in a machine learning algorithms without considering the differences could lead to a measurement bias. The last bias we want to mention is the algorithm bias. It is independent of the training data, rather it is encoded in the algorithm. A typical example is an algorithm which considers one option before the other. Because of this procedure, the algorithm would more likely classify novel datasets into the first category and would only compare them with the second category if no match is previously found (Dietterich & Kong 1995).

Another problem that requires philosophical consideration is the comparability of different populations. For example, Mutz et al. (2017) used Stochastic Frontier Analysis and measured the efficiency of research projects in different fields, both in natural and social sciences. However, if we use data from very different fields, we have to account for different styles, traditions, and necessities. Citation metrics are field dependent, publications in experimental biology or cancer research are typically cited much more frequently than publication in, e.g., philosophy. Moreover, as explained in section 5.2, citation metrics are not an equally good measure of efficiency in every field. Perović & Sikimić (under revision) argued that disciplines such as plant biology do not exhibit the same reliable inductive pattern as experimental physics or phylogenetics, which in turn indicates that the citation metrics have a weaker predictive power of efficiency in the field. In life science, the reproducibility crisis causes that consensus on the results is not long-lasting and very reliable, e.g., Pusztai et al. (2013).

Though a very powerful tool, a data-driven approach should be applied carefully, eliminating biases from the data and algorithms. Also, before the application of data-mining algorithms, it is important to check whether the available data is comparable and informative for the desired research goal. Data-driven approaches are useful for the questions in social epistemology of science. Once initial criteria are met, using data-driven analysis one can get predictions about optimal team structure, the optimal number of researchers, project duration, communication frequency among researchers, etc.

8.4 Summary

Data-driven analyses are useful for studying "science of science" because they have a clear interpretation. Still, there are several important steps that need to be addressed when using external data. Firstly, it is important to gather representative data, carefully curate them and make them widely accessible. If this is not possible, also a limited accessibility can be helpful. In this process, it is important to take the necessary measures to protect the privacy of the scientists. Secondly, biases from the datasets should be removed as much as possible. Thirdly, algorithms need to be applied in a meaningful way, e.g., only on comparable datasets.

Chapter 9

Conclusions and further research

9.1 Conclusions

We have examined data-driven approaches in the field of social epistemology of science. Apart from being used internally within scientific disciplines, data are a valuable resource for optimizing scientific inquiry. In this sense, data contribute to making informed judgments about science policy.

There are several ways to incorporate data into the philosophical analysis of scientific inquiry. These include, among others:

- Data-mining techniques,
- Qualitative and quantitative studies,
- Empirically calibrated models, and
- Analyses of argumentation patterns.

We argue that data-driven analyses require a field-specific approach, since different rules and metrics apply in different scientific disciplines. The focus of the present research was on high energy physics and experimental biology. The research in high energy physics exhibits a relatively parsimonious nature, which is mirrored by the machine

learning algorithms used to describe the pattern of discoveries in this field. Citation rates are a relatively reliable metric to keep track of research in the field, because consensus is reached relatively quickly and remains stable over the years. Additionally, the use of expensive equipment and the application process for its use guarantees that experiments in high energy physics are almost unique. Finally, experts mainly cite positive results from their field, which all led us to the conclusion that citation metrics can be used as an efficiency parameter in the field. However, the same does not hold in the majority of areas of biology. Generally, it is not advisable to use citation metrics as the efficiency parameter for research in biology (Perović & Sikimić under revision). On the other hand, case studies of scenarios in experimental biology are a valuable source of data about the epistemic exchange in the field.

In chapter 7 we analyzed cases in which non-parsimonious hypotheses challenged received views on pathogenesis, and noted that cognitive diversity is advantageous for making discoveries in circumstances when the epistemic landscape requires outside-the-box thinking.

In chapter 3 we discussed data-driven analyses of the efficiency of projects in HEP. Since citation metrics in HEP are a relatively reliable measure of project efficiency, data-driven approaches based on them can be fruitful. Perović et al. (2016) and Sikimić et al. (2018) used data from Fermilab to establish optimal team structures in HEP laboratories and determine an epistemic saturation point in the duration of experiments. They showed that efficient projects are usually relatively small, both in the number of researchers and in the number of research teams. The results from Perović et al. (2016) that smaller teams outperform large ones agree with the conclusions of the empirically calibrated models of scientific interaction in HEP that we presented in chapter 6. With the help of our agent-based model, we showed that collaborations involving several medium-sized teams have problems reaching a consensus. Individual teams easily form beliefs that differ from the beliefs of the other teams. The reason for

this becomes obvious in the agent-based models: scientists generally communicate more with members from their team than with members from other teams, so the influence of their close peers outweighs the influence of the rest of the group.

Empirically calibrated models of scientific inquiry and data-driven analyses such as DEA require data collection. In chapter 2, we highlighted how these data could be generated or acquired and briefly discussed the potential of open data to transform scientific disciplines and social epistemology of science. Well-curated datasets are a valuable resource for all scientists, but even the best data sources require a thorough philosophical evaluation to account for biases and variations in data collection.

Finally, the results presented and discussed in this thesis could be relevant for science policy. The data-driven studies by Perović et al. (2016) and Sikimić et al. (2018) can promote an improved allocation of resources. They highlight an epistemic saturation point for the duration of experiments in HEP and optimal team structures. These results might help reviewers to evaluate the potential of projects in HEP and guide researchers to adopt better team structures for their laboratories. Similarly, we have elaborated on the methods that can be used as adequate tools for the optimization of scientific reasoning in experimental biology. Therefore, we hope to contribute with this thesis to the discussion about communication processes in science and promote data-driven approaches as a useful tool in philosophy of science.

9.2 Further directions

The data-driven approach to social epistemology of science requires interdisciplinary research that combines philosophy with psychology, computer science, data science, and other disciplines. In order to advance the field of data-driven social epistemology of science, we suggest a project with a wide and interdisciplinary scope.

For further research, we plan to develop more sophisticated, empirically calibrated

models that will capture fine-grained updating procedures and epistemic landscapes. Moreover, in order to collect data about scientific interaction, researchers from the University of Belgrade and the Ruhr University of Bochum have created the Optimist platform (Optimist 2018).[1] The aim of the collaboration is to make the research process both more human and more efficient. As the first step, a survey about work conditions in HEP was launched. Qualitative interviews and mixed methods combining qualitative interviews and quantitative surveys are also a valuable source of data for social epistemology of science; therefore, information from interviews with experimental biologists about the epistemic practices in their field, conducted by some of the Optimist group members are used for empirically calibrated models.

Finally, processing a larger dataset of Fermilab projects would enable us to use predictive machine learning algorithms. The idea of predictive analysis is to reveal whether an experiment will be efficient exclusively based on data from the project proposal (e.g., the number of researchers and teams involved, expected duration, or necessary funding). Moreover, a predictive analysis should also indicate which way each parameter of an inefficient project should be modified. In this way, both the researchers themselves and funding agencies can optimize their projects. Philosophical considerations represent an important part of the listed approaches, since the motivation for this research and the research questions arise from the philosophy of science.

[1] For additional information please consult: <http://www.ruhr-uni-bochum.de/optimist-survey/>.

Bibliography

Aad, G., Abbott, B., Abdallah, J., Abdinov, O., Aben, R., Abolins, M., AbouZeid, O., Abramowicz, H., Abreu, H., Abreu, R. et al. (2015), 'Combined measurement of the higgs boson mass in p p collisions at s= 7 and 8 tev with the atlas and cms experiments', *Physical review letters* **114**(19), 191803.

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.

Alteri, E. & Guizzaro, L. (2018), 'Be open about drug failures to speed up research'.

Anderson, K. R., Heidinger, B. H., Chen, Y., Bankier, A. A. & VanderLaan, P. A. (2017), 'Measurement Bias of Gross Pathologic Compared With Radiologic Tumor Size of Resected Lung Adenocarcinomas: Implications for the T-Stage Revisions in the Eighth Edition of the American Joint Committee on Cancer Staging Manual', *Am. J. Clin. Pathol.* **147**(6), 641–648.

Arkes, H. R. & Blumer, C. (1985), 'The psychology of sunk cost', *Organizational behavior and human decision processes* **35**(1), 124–140.

Azoulay, P., Fons-Rosen, C. & Zivin, J. S. G. (2015), Does science advance one funeral at a time?, Working Paper 21788, National Bureau of Economic Research.

Baker, M. (2015), 'Over half of psychology studies fail reproducibility test', *Nature News* .

Baltag, A., Christoff, Z., Hansen, J. U. & Smets, S. (2013), 'Logical models of informational cascades', *Studies in Logic* **47**, 405–432.

Baltag, A., Gierasimczuk, N. & Smets, S. (2016), 'On the solvability of inductive problems: A study in epistemic topology'.

Baltag, A. & Smets, S. (2011), 'Keep changing your beliefs, aiming for the truth', *Erkenntnis* **75**(2), 255.

Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992), 'A theory of fads, fashion, custom, and cultural change as informational cascades', *Journal of Political Economy* **100**(5), 992–1026.

Birnholtz, J. (2008), 'When authorship isn't enough: lessons from cern on the implications of formal and informal credit attribution mechanisms in collaborative research', *Journal of Electronic Publishing* **11**(1).

Boisot, M., Nordberg, M., Yami, S. & Nicquevert, B. (2011), *Collisions and Collaboration: The Organization of Learning in the ATLAS Experiment at the LHC*, Oxford University Press.

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

Borg, A., Frey, D., Šešelja, D. & Straßer, C. (2017), An argumentative agent-based model of scientific inquiry, *in* S. Benferhat, K. Tabia & M. Ali, eds, 'Advances in Artificial Intelligence: From Theory to Practice', Springer International Publishing, Cham, pp. 507–510.

Borgman, C. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World*, The MIT Press.

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017), 'Semantics derived automatically from language corpora contain human-like biases', *Science* **356**(6334), 183–186.

Carayol, N. & Matt, M. (2006), 'Individual and collective determinants of academic scientists' productivity', *Information Economics and Policy* **18**(1), 55–72.

Carillo, M. R., Papagni, E. & 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**, 25–36.

Carter, K. C. (2003), *The rise of causal concepts of disease: Case histories*, Ashgate Publishing, Ltd.

Castilla, J., Saa, P., Hetz, C. & Soto, C. (2005), 'In vitro generation of infectious scrapie prions', *Cell* **121**(2), 195–206.

Cho, A. (2008), 'Particle physics. Does Fermilab have a future?', *Science* **320**(5880), 1148–1151.

Cialdini, R. (2001), *Influence: science and practice*, 4 edn, Allyn & Bacon, Boston.

Collins, F. S. (2017), 'New NIH approach to grant funding aimed at optimizing stewardship of taxpayer dollars'.
URL: *https://www.nih.gov/about-nih/who-we-are/nih-director/statements/new-nih-approach-grant-funding-aimed-optimizing-stewardship-taxpayer-dollars*

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, T. C., Ioannidis, J. P. et al. (2008), 'Life cycle of translational research for medical interventions', *Science* **321**(5894), 1298–1299.

Cook, I., Grange, S. & Eyre-Walker, A. (2015), 'Research groups: How big should they be?', *PeerJ* **3**, e989.

Cooper, W. W., Li, S., Seiford, L. M. & Zhu, J. (2004), *Sensitivity Analysis in DEA*, Springer US, Boston, MA, pp. 75–97.

Cornwall, C. (2013), *Catching cancer: the quest for its viral and bacterial causes*, Rowman & Littlefield.

Cyranoski, D. (2018), 'China's crackdown on genetics breaches could deter data sharing', *Nature* **563**(7731), 301–302.

Daly, C. (2010), *An Introduction to Philosophical Methods*, Broadview Press.

Dietterich, T. G. & Kong, E. B. (1995), Machine learning bias, statistical bias, and statistical variance of decision tree algorithms, Technical report, Technical report, Department of Computer Science, Oregon State University.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H. J., Heng, P. A., Hass, C., Bruni, E., Wong, Q., Halici, U., Oner, M. U., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y. W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M. M., Serrano, I., Deniz, O., Racoceanu, D. & Venancio, R. (2017), 'Diagnostic

Assessment of Deep Learning Algorithms for Detection of Lymph N ode Metastases in Women With Breast Cancer', *JAMA* **318**(22), 2199–2210.

Evans, A. (1976), 'Causation and disease: The henle-koch postulates revisited', *J Bacteriol* **49**, 175–195.

Fagiolo, G., Windrum, P. & Moneta, A. (2006), Empirical validation of agent-based models: A critical survey, Technical report, LEM Working Paper Series.

Farmer, J. D. & Foley, D. (2009), 'The economy needs agent-based modelling', *Nature* **460**(7256), 685.

Farrell, M. J. (1957), 'The measurement of productive efficiency', *Journal of the Royal Statistical Society. Series A (General)* **120**(3), 253–290.

Fatović-Ferenčić, S. & Banić, M. (2011), 'No acid, no ulcer: Dragutin (Carl) Schwarz (1868–1917), the man ahead of his time', *Digestive Diseases* **29**, 507–510.

Figura, N. & Bianciardi, L. (2002), Helicobacters were discovered in Italy in 1892: An episode in thescientific life of an eclectic pathologist, Giulio Bizzozero, *in* B. Marshall, ed., 'Helicobacter pioneers: accounts from the scientists who discovered Helicobacters 1892-1982', pp. 1–13.

Fire, A., D. A. S. W. H. & Moerman, D. G. (1991), 'Production of antisense rna leads to effective and specific inhibition of gene expression in c. elegans muscle', *Development* **113**(2), 503–514.

Frey, D. & Šešelja, D. (2018*a*), 'Robustness and idealizations in agent-based models of scientific interaction', *The British Journal for the Philosophy of Science* p. axy039.

Frey, D. & Šešelja, D. (2018*b*), 'What is the epistemic function of highly idealized agent-based models of scientific inquiry?', *Philosophy of the Social Sciences* **48**(4), 407–433.

Fukuda, Y., S. T. S. T. . M. B. J. (2002), Kasai, kobayashi and koch's postulates in the history of helicobacter pylori, *in* B. Marshall, ed., 'Helicobacter pioneers: accounts from the scientists who discovered Helicobacters 1892 - 1982', pp. 15–24.

Galison, P., Hevly, B. & Weinberg, A. M. (1992), 'Big science: The growth of large-scale research', *Physics Today* **45**, 89.

Gentil-Beccot, A., Mele, S. & Brooks, T. (2009), 'Citing and reading behaviours in high-energy physics', *Scientometrics* **84**(2), 345–355.

Goldman, A. & Blanchard, T. (2016), Social epistemology, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2016 edn, Metaphysics Research Lab, Stanford University.

Goodman, S. N., Fanelli, D. & Ioannidis, J. P. (2016), 'What does research reproducibility mean?', *Science translational medicine* **8**(341), 341ps12–341ps12.

Grim, P. (2009), Threshold phenomena in epistemic networks., *in* 'AAAI Fall Symposium: Complex Adaptive Systems and the Threshold Effect', pp. 53–60.

Hangel, N. & Schickore, J. (2017), 'Scientists' conceptions of good research practice', *Perspectives on Science* **25**(6), 766–791.

Harnagel, A. (2018), 'A mid-level approach to modeling scientific communities', *Studies in History and Philosophy of Science Part A* .
URL: *http://www.sciencedirect.com/science/article/pii/S0039368118300049*

Hartmann, S. & Rafiee Rad, S. (2018), 'Voting, deliberation and truth', *Synthese* **195**(3), 1273–1293.

Heesen, R. (2018), 'Why the reward structure of science makes reproducibility problems inevitable', *The Journal of Philosophy* **115**(12), 661–674.

Hendricks, V. & Hansen, P. (2014), *Infostorms: how to take information punches and save democracy.*, Copernicus Books, Springer.

Hendricks, V. & Hansen, P. (2016), *Infostorms: Why do we 'like'? Explaining individual behavior on the social net.*, Copernicus books, Springer.

Henikoff, S. & Henikoff, J. G. (1992), 'Amino acid substitution matrices from protein blocks', *Proceedings of the National Academy of Sciences* **89**(22), 10915–10919.

Hermann, A., Pestre, D., Krige, J. & Mersits, U. (1987), 'History of cern. vol. 1: Launching the european organization for nuclear research'.

Hoddeson, L., Brown, L., Riordan, M. & Dresden, M. (1997), *The rise of the standard model: A history of particle physics from 1964 to 1979*, Cambridge University Press.

Hoddeson, L., Kolb, A. W. & Westfall, C. (2009), *Fermilab: Physics, the frontier, and megascience*, University of Chicago Press.

Huemer, M. (2009), 'When is parsimony a virtue?', *The Philosophical Quarterly* **59**, 216–236.

Hunt, V., Prince, S., Dixon-Fyle, S. & Yee, L. (2018), 'Delivering through diversity'.

Irvine, J. & Martin, B. R. (1984b), 'Cern: Past performance and future prospects: Ii. the scientific performance of the cern accelerators', *Research Policy* **13**, 247–284.

Jackson, S. E. (1996), 'The consequences of diversity in multidisciplinary work teams', *Handbook of work group psychology* pp. 53–75.

Kalia, L. & Lang, A. (2015), 'Parkinson's disease.', *The Lancet* **386**, 896–912.

Katz, R. (1982), 'The effects of group longevity on project communication and performance', *Administrative science quarterly* pp. 81–104.

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.

Kelly, K. T. & Genin, K. (2018), 'Learning, theory choice, and belief revision', *Studia Logica* .

Kelly, K. T., Genin, K. & Lin, H. (2016), 'Realism, rhetoric, and reliability', *Synthese* **193**(4), 1191–1223.

Kelly, K. T. & Mayo-Wilson, C. (2010), 'Ockham efficiency theorem for stochastic empirical methods', *Journal of Philosophical Logic* **39**(6), 679–712.

Kelly, K. T., Schulte, O. & Juhl, C. (1997), 'Learning theory and the philosophy of science', *Philosophy of Science* **64**(2), 245–267.

Keyes, M. E. (1999), 'The prion challenge to the 'central dogma' of molecular biology, 1965–1991', *Studies in History and Philosophy of Science Part C* **30**(2), 181–218.

Kitcher, P. (1990), 'The division of cognitive labor', *Journal of Philosophy* **87**(1), 5–22.

Kitcher, P. (1993), *The advancement of science*, Oxford University Press.

Knobe, J. & Nichols, S. (2017), Experimental philosophy, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2017 edn, Metaphysics Research Lab, Stanford University.

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

Kozlowski, S. W. J. & Bell, B. S. (2003), *Work Groups and Teams in Organizations*, John Wiley & Sons, Inc.

Kumar, S., Grefenstette, J. J., Galloway, D., Albert, S. M. & Burke, D. S. (2013), 'Policies to reduce influenza in the workplace: impact assessments using an agent-based model', *American journal of public health* **103**(8), 1406–1411.

Kummerfeld, E. & Zollman, K. J. S. (2015), 'Conservatism and the Scientific State of Nature', *The British Journal for the Philosophy of Science* **67**(4), 1057–1076.

Lauer, M. S., Roychowdhury, D., Patel, K., Walsh, R. & Pearson, K. (2017), 'Marginal returns and levels of research grant support among scientists supported by the national institutes of health', *bioRxiv* pp. 1–30.
URL: *https://www.biorxiv.org/content/early/2017/05/29/142554*

Longino, H. E. (1996), *Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy*, Springer Netherlands, Dordrecht, pp. 39–58.

Longino, H. E. (2002), *The fate of knowledge*, Princeton University Press.

Malfertheiner, P., Chan, F. K. & McColl, K. E. (2009), 'Peptic ulcer disease', *The Lancet* **374**(9699), 1449–1461.

Marshall, B. (2001), 'One hundred years of discovery and rediscovery of helicobacter pylori and its association with peptic ulcer disease'.

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

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

McCole, S. D., Claney, K., Conte, J. C., Anderson, R. & Hagberg, J. M. (1990), 'Energy expenditure during bicycling', *J. Appl. Physiol.* **68**(2), 748–753.

McLevey, J. & McIlroy-Young, R. (2017), 'Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science', *Journal of Informetrics* **11**(1), 176 – 197.

Milojević, S. (2014), 'Principles of scientific research team formation and evolution', *Proceedings of the National Academy of Sciences* **111**(11), 3984–3989.

Mutz, R., Bornmann, L. & Daniel, H.-D. (2017), 'Are there any frontiers of research performance? efficiency measurement of funded research projects with the bayesian stochastic frontier analysis for count data', *Journal of Informetrics* **11**(3), 613–628.

Nickles, T. (2006), *Heuristic appraisal: Context of discovery or justification?*, Springer Netherlands, Dordrecht, pp. 159–182.

Nieva, V. F., Fleishman, E. A. & Rieck, A. (1985), Team dimensions: Their identity, their measurement and their relationships, Technical report, Advanced Research Resources Organization BETHESDA MD.

Nieva, Veronica F ; Fleishman, E. A. . R. A. (1985), *Team Dimensions: Their Identity, Their Measurement and Their Relationships*, Washington, DC: U. S. Army, Research Institute for the Behavioral and Social Sciences.

Olson, B. J., Parayitam, S. & Bao, Y. (2007), 'Strategic decision making: The effects of cognitive diversity, conflict, and trust on decision outcomes', *Journal of management* **33**(2), 196–222.

Optimist (2018), 'Research Platform'.
URL: *http://www.ruhr-uni-bochum.de/optimist-survey/*

Ormerod, P. & Rosewell, B. (2009), Validation and verification of agent-based models in the social sciences, *in* 'Epistemological aspects of computer simulation in the social sciences', Springer, pp. 130–140.

Ossowska, M. & Ossowski, S. (1964), 'The science of science', *Minerva* **3**(1), 72–82.

Palmer, E. (1954), 'Investigations of the gastric mucosa spirochetes of the human.', *Gastroenterology* **27**, 218–220.

Perović, S., Radovanović, S., Sikimić, V. & Berber, A. (2016), 'Optimal research team composition: data envelopment analysis of fermilab experiments', *Scientometrics* **108**(1), 83–111.

Perović, S. & Sikimić, V. (under revision), 'How theories of induction can streamline measurements of scientific performance'.

Powell, K. (2018), 'These labs are remarkably diverse - here's why they're winning at science', *Nature* **558**(7708), 19–22.

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.

Reich, E. S. (2011), 'Fermilab faces life after the Tevatron', *Nature* **477**(7365), 379.

Rivers, T. (1937), 'Viruses and koch's postulates', *J Bacteriol* **33**, 1–12.

Rosenstock, S., Bruner, J. & O'Connor, C. (2017), 'In epistemic networks, is less really more?', *Philosophy of Science* **84**(2), 234–252.

Ryan, J. (2018), 'WhatsApp fights mob violence in India by limiting message forwarding'.
URL: *https://www.cnet.com/news/whatsapp-limits-message-forwarding-in-ongoing-effort-to-fight-fake-news/*

Rylance, R. (2015), 'Grant giving: Global funders to focus on interdisciplinarity', *Nature News* **525**(7569), 313.

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V., Chesselet, M.-F., Keshavarzian, A., Shannon, K. M., Krajmalnik-Brown, R., Wittung-Stafshede, P., Knight, R. & Mazmanian, S. K. (2016), 'Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease', *Cell* **167**(6), 1469–1480.

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. (2018), 'Causal learning with Occam's razor', *Studia Logica* pp. 1–33.

Schulte, O. & Drew, M. S. (2010), Discovery of conservation laws via matrix search, *in* 'International Conference on Discovery Science', Springer, pp. 236–250.

Seijts, G. H. & Latham, G. P. (2000), 'The effects of goal setting and group size on performance in a social dilemma.', *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* **32**(2), 104.

Šešelja, D., Kosolosky, L. & Straßer, C. (2012), 'Rationality of scientific reasoning in the context of pursuit: drawing appropriate distinctions.', *Philosophica* **86**, 51–82.

Šešelja, D. & Straßer, C. (2013), 'Abstract argumentation and explanation applied to scientific debates', *Synthese* **190**(12), 2195–2217.

Šešelja, D. & Straßer, C. (2014), 'Heuristic reevaluation of the bacterial hypothesis of peptic ulcer disease in the 1950s', *Acta Biotheoretica* **62**(4), 429–454.

Sikimić, V. (2017), Interdisciplinarity in contemporary philosophy: the case of social epistemology, *in* 'Društvene nauke pred izazovima savremenog društva', pp. 19–26.

Sikimić, V., Sandro, R. & Perović, S. (2018), 'When should we stop investing in a scientific project? the halting problem in experimental physics', *Proceedings of the XXIV Conference "Empirical Studies in Psychology"* pp. 105–107.

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

Stackebrandt, E. & Goebel, B. M. (1994), 'Taxonomic note: A place for dna-dna reassociation and 16s rrna sequence analysis in the present species definition in bacteriology', *International Journal of Systematic and Evolutionary Microbiology* **44**(4), 846–849.

Stanev, R. (2012), Stopping rules and data monitoring in clinical trials, *in* S. Hartmann & S. Okasha, eds, 'EPSA philosophy of science: Amsterdam 2009', Springer, pp. 375–386.

Steele, K. (2013), 'Persistent experimenters, stopping rules, and statistical inference', *Erkenntnis* **78**(4), 937–961.

Strevens, M. (2003), 'The role of the priority rule in science', *The Journal of Philosophy* **100**(2), 55–79.

Surowiecki, J. (2004), *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Anchor Books: Random House LLC.

Sweis, B. M., Abram, S. V., Schmidt, B. J., Seeland, K. D., MacDonald, A. W., Thomas, M. J. & Redish, A. D. (2018), 'Sensitivity to "sunk costs" in mice, rats, and humans', *Science* **361**(6398), 178–181.

Tabrah, F. (2011), 'Koch's postulates, carnivorous cows, and tuberculosis today.', *Hawaii medical journal* **70**, 144–148.

Thagard, P. (1988), *Computational Philosophy of Science*, Castle Rock, Pittsford, USA.

Thaler, R. (1980), 'Toward a positive theory of consumer choice', *Journal of Economic Behavior & Organization* **1**(1), 39–60.

Turing, A. M. (1937), 'On computable numbers, with an application to the entscheidungsproblem', *Proceedings of the London mathematical society* **2**(1), 230–265.

Übler, H. & Hartmann, S. (2016), 'Simulating trends in artificial influence networks', *Journal of Artificial Societies and Social Simulation* **19**(1).

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

Valdés-Pérez, R. E. & Erdmann, M. (1994), 'Systematic induction and parsimony of phenomenological conservation laws', *Computer Physics Communications* **83**(2-3), 171–180.

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.

Vestjens, J. H., Pepels, M. J., de Boer, M., Borm, G. F., van Deurzen, C. H., van Diest, P. J., van Dijck, J. A., Adang, E. M., Nortier, J. W., Rutgers, E. J., Seynaeve, C., Menke-Pluymers, M. B., Bult, P. & Tjan-Heijnen, V. C. (2012), 'Relevant impact of central pathology review on nodal classification in individual breast cancer patients', *Ann. Oncol.* **23**(10), 2561–2566.

Vickers, P. (2014), 'Theory flexibility and inconsistency in science', *Synthese* **191**(13), 2891–2906.

Von Tunzelmann, N., Ranga, M., Martin, B. & Geuna, A. (2003), 'The effects of size on research performance: A spru review', *Report prepared for the Office of Science and Technology, Department of Trade and Industry* .

Wagenknecht, S. (2016), *A Social Epistemology of Research Groups*, Palgrave Macmillan UK.

Wagenknecht, S., Nersessian, N. J. & Andersen, H. (2015), Empirical philosophy of science: Introducing qualitative methods into philosophy of science, *in* 'Empirical Philosophy of Science', Springer, pp. 1–10.

Wang, J., Veugelers, R. & Stephan, P. (2017), 'Bias against novelty in science: A cautionary tale for users of bibliometric indicators', *Research Policy* **46**(8), 1416–1436.

Way, S. F., Morgan, A. C., Clauset, A. & Larremore, D. B. (2017), 'The misleading narrative of the canonical faculty productivity trajectory', *Proceedings of the National Academy of Sciences* **114**(44), E9216–E9223.

Weckowska, D. M., Levin, N., Leonelli, S., Dupré, J. & Castle, D. (2017), 'Managing the transition to open access publishing: a psychological perspective', *Prometheus* pp. 1–25.

Weinert, F. (2008), *Copernicus, Darwin, and Freud: Revolutions in the History and Philosophy of Science.*, Oxford: Wiley-Blackwell.

Weisberg, M. & Muldoon, R. (2009), 'Epistemic landscapes and the division of cognitive labor', *Philosophy of Science* **76**(2), 225–252.

Whitt, L. A. (1992), 'Indices of theory promise', *Philosophy of Science* **59**(4), 612–634.

Will, R. G., Ironside, J. W., Zeidler, M., Estibeiro, K., Cousens, S. N., Smith, P. G., Alperovitch, A., Poser, S., Pocchiari, M. & Hofman, A. (1996), 'A new variant of Creutzfeldt-Jakob disease in the uk', *The Lancet* **347**(9006), 921–925.

Yang, Z. & Rannala, B. (2012), 'Molecular phylogenetics: principles and practice', *Nature reviews genetics* **13**(5), 303.

Yeom, H. W. (2018), 'South korean science needs restructuring', *Nature* **558**(7711), 511–513.

Zliobaite, I. (2015), 'A survey on measuring indirect discrimination in machine learning', *arXiv preprint arXiv:1511.00148* .

Zollman, K. J. S. (2007), 'The communication structure of epistemic communities', *Philosophy of Science* **74**(5), 574–587.

Zollman, K. J. S. (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* **392**(1), 1 – 10.