6. Science Forecasts: Modeling and Communicating Developments in Science, Technology, and Innovation

Katy Börner, Staša Milojević

In a knowledge-based economy, science and technology are omnipresent, and their importance is undisputed. Equally evident is the need to allocate resources, both monetary and human, in an effective way to foster innovation [6.1, 2]. In the preceding decades, science policy has embraced data mining and metrics to gain insights into the structure and evolution of science and to devise metrics and indicators [6.3], but it has not invested significant efforts into mathematical, statistical, and computational models that can predict future developments in science, technology, and innovation (STI) in support of data-driven decision making.

Recent advances in computational power combined with the unprecedented volume and variety of data concerning science and technology developments (e. g., publications, patents, funding, clinical trials, and stock market and social media data) yielded ideal conditions for the advancement of computational modeling approaches that can be not only empirically validated, but used to simulate and understand the structure and dynamics of STI in support of improved human decision making.

In this chapter, we review and demonstrate the power of computational models for simulating and predicting possible STI developments and futures. In addition, we discuss novel means to visualize and broadcast STI forecasts to make them more accessible to general audiences.

6.1 Models and Visualizations

Science, technology, and innovation are crucial for the prosperity of nations, and are a driving force of human civilization. After World War II, science entered a phase of accelerated growth, reflected in the exponential rise in the number of active scientists and an increase of scientific output [6.4]. Science itself is undergoing a transformation, with most researchers engaging in collaborative or team work [6.5, 6]. In order to create effective science policies and maximize the returns on our society's investments in STI, we must understand STI as a complex and dynamic system that emerges from interdependences and interactions of different actors at different levels of aggregation.

Models of STI aim to inform policy decision making in many fields, including education, energy, healthcare, security, and others [6.7, 8]. These models do not replace—but rather empower—experts to make better informed decisions when selecting reviewers, picking the best proposals for funding, or when making resource allocation decisions. They are a new kind of

W. Glänzel et al. (Eds.), *Springer Handbook of Science and Technology Indicators*, Springer Handbooks,
https://doi.org/10.1007/978-3-030-02511-3_6

'macroscope tool' [6.9] that help derive key insights from big data in support of evidence-based policy.

Some existing models of STI are optimized to make recommendations. IBM's *Watson*, for example, can suggest reviewers for a set of proposals without much information on the type of match or the matching process. Other models aim to capture the true structure and dynamics of complex STI systems, simulating the diffusion of ideas and experts, estimating the impact of population explosion and aging, or communicating the probable outcomes of different policy decisions. Still others help answer either resource allocation or multifaceted strategic questions, the latter of which are often used in a team setting where small multidisciplinary groups investigate and debate alternative futures together.

Computational models are well established in many fields: meteorology, where they are used to predict weather and storms; epidemiology, to predict and prevent pandemics; and climate, to predict future scenarios and set carbon prices. In industry (hereafter used as a general term to indicate the various industrial sectors, such as retail, IT, car manufacturing, etc.), computational models are used to optimize operations, management, production, distribution, and marketing. Early adopters of data-driven decision making (most notably Target, Walmart, and Amazon) now dominate their sectors. Those who were slow to invest and then did so in isolated aspects of the organization (most notably Sears and Kmart) are headed towards bankruptcy.

Interactive data visualizations that show probable futures in response to different policy decisions or external events can help stakeholders discuss model assumptions, designs, and outputs. Ideally, stakeholders get to "drive the future before it is implemented" [6.10, 11]; they can quickly explore different policy options and discard those that lead to undesired consequences [6.2, 12]. However, designing effective interfaces that let different stakeholders communicate and explore different scenarios is a nontrivial endeavor.

6.2 Models and Modeling

While our world is infinitely complex, our ability to sense, understand, and act within that world is finite. To capture and interpret the structures and dynamics of a complex system such as STI, scientists build models, which are simplified representations of a system [6.13]. Models bring conceptual unity to what is otherwise too complex to understand and manage. Regardless of the approach, the goal of any model is to simplify thinking "while still retaining some ability to illuminate reality" [6.14, p. 11]. In order to understand and predict different aspects of the world, models reduce the world to a subset of elements and laws that govern the behavior of those elements. Such simplification allows researchers to focus on and elucidate only the specific elements of a system that concern them. While every model is bounded by its initial framework, this does not mean that it cannot increase our understanding of the phenomenon at hand.

Complex systems research, however, challenges the notion that by perfectly understanding the behavior of each component of a system, we will understand the system as a whole. While there is no agreed upon definition of complex systems, the combination of various definitions leads to the following characteristics that a system needs to have in order to be considered complex [6.15, 16]. Two major components of a system are its entities and the interactions among those entities, with a much heavier emphasis on the interactions than on the entities.

A complex system usually has a large number of entities that mainly respond to local information (i. e., each element of the system is ignorant of the behavior of the system as a whole). The interactions these entities have can be: nonlinear (small changes in system variables that can have disproportionate outcomes); dynamic (changes over time); rife with feedback loops (both positive and negative, which can lead to distributions such as a power law); fairly rich (any element in the system influences and is influenced by quite a few others); and fairly short range. So far as the system as a whole is concerned, it is open (i. e., interacts with its environment); requires nonequilibrium conditions (there needs to be a constant flow of energy to ensure the survival of the system); and has a history (not only does it evolve through time, but its past is coresponsible for the present behavior).

Computational models consist of input (theories translated first into mathematical equations, and then into algorithms with different parameter values) and output (structures, or the behavior of the model over time). A dynamic system is one which by its very nature changes its state in ways that can be modeled by an application of an evolution law (a set of rules that describe what phase space configuration the system will occupy in the next moment). In the case of complex systems, these rules—though often very simple—can lead to so-called *emergent behavior*, a phenomenon in which "individual, localized behavior aggregates into global

behavior" [6.14, p. 44]. When modeling the evolution of dynamic complex systems, researchers must remember that the resultant model represents one configuration in a phase space, given at time t [6.17], and as the models aim to capture system dynamics, output at time t often serves as input for computing time $t + 1$.

Typically, the accuracy of simulations increases with both the ease of repetition and the number of simulations run (i. e., the number of possible futures obtained). Running simulations multiple times allows for better estimates concerning the sensitivity of outcomes to initial conditions, as well as the probabilities associated with those outcomes.

Computer modeling is gaining traction as an acceptable approach to doing science in a wide range of fields, from astronomy to economics [6.18]. Computational models use simulations to study the behavior of a system, and these simulations pose new questions regarding the scientific method, the nature of evidence, theory and theory building, and the role of data [6.19]. Outside pressures have forced researchers in climatology—a field heavily dependent on computer models—to be at the forefront of deeply critical thinking regarding the capabilities and limitations of computer modeling [6.20]. In this way, climatology exemplifies how community-lead, large-scale endeavors to gather, model, and visualize data can lead to significant infrastructure building.

Using simulations in the social sciences is a more recent phenomenon; however, a number of excellent resources showcase the benefits of broad usage cases [6.21] and provide practical guidance [6.22, 23]. Computational models can be used both to advance theory via conceptual models, and as tools to enhance decision making via predictions. In both cases, modeling is an iterative process that includes both induction and deduction [6.14] to revise and improve an initial set of assumptions, often leading to better results.

Recent developments in machine learning have dramatically improved researchers' capabilities to identify structures and patterns in data to aid with decision making [6.24, 25]. *Jon Kleinberg* and colleagues [6.26] provide a great overview of what they call "prediction policy problems" ranging from the medical field (predicting which surgeries will be futile) to criminal justice (deciding on whether to detain or release an arrestee).

6.3 Modeling Science

The book *Models of Science Dynamics* [6.8] provides a unique review of major model classes—from population dynamics models to complex network models—accessible to science policy researchers and practitioners. Two special issues in *Scientometrics* entitled *Modeling Science: Studying the Structure and Dynamics of Science* [6.27] and *Simulating the Processes of Science, Technology, and Innovation* [6.28] feature research, applications, and validations of exemplary STI models.

Models capturing the structure and evolution of scientific endeavor fall into one of two categories: descriptive and predictive [6.29]. Descriptive models include maps, and aim to describe the major features of static datasets. Predictive (or process) models aim to capture the mechanisms and temporal dynamics by which real-world systems evolve, focusing on the identification of elementary mechanisms that lead to the emergence of specific structures or dynamics. Ultimately, process models seek to simulate, statistically describe, or formally reproduce statistical characteristics of interest.

Computational models have been developed to enhance our knowledge of fundamental generating processes regarding citing, publishing, careers, rewards, funding, team formation, problem selection, and research areas dynamics. They gained traction in recent years because of their power to simulate processes leading to particular outcomes. Particularly important were findings that identified universal patterns, that is, patterns holding across majority of scientific fields, such as citations dynamics and timing of major discoveries.

A number of scientific models draw from complexity theory. Modern notions of complexity have their roots in theories of chaos, complex systems, fractal geometry, nonlinear dynamics, and self-organizing criticality. Complexity theory has been influential in physical science, technology, and mathematics for some time. This influence is newer and less developed in social science [6.30]. This is unsurprising, since models of complex systems work best with homogeneous elements where there is little or no difference between the individual elements of the system (e. g., atoms and molecules). Using the tools of complexity theory to cover the richness of both entities and their relationships within social systems proves to be a harder task.

The *natural* and *self-organizing* development of science towards more interdisciplinary activities is comparable with ecological systems that exhibit growth and emergent behavior [6.31, 32]. *Anthony van Raan* [6.33] expanded on this idea, portraying science not only as an interdisciplinary, complex, and self-organizing system, but as an amalgam of "cognitive regions" derived from the parts of pre-established disciplines. Such disciplines represent research fields that originated from earlier interdisciplinary developments.

In this model, science itself is a living, complex and dynamic system consisting of several ever-growing subsystems, each of which unfolds further into a myriad of different fields and subfields. A variation of the so-called "epidemics model" [6.34] was used in the 1960s to develop an epidemic theory on the diffusion of ideas and the growth of scientific specialties [6.35]. By using mast cell research as a case study, William Goffman demonstrated that it was possible to see growth and development as sequences of overlapping epidemics.

A wide range of studies used network-based models of citations to understand different aspects of science. A number of studies focused on understanding the dynamics of citation accumulation, starting from the identification of cumulative advantage/preferential attachment as the driving mechanism behind the power-law distribution of citations [6.36, 37]. *Filippo Radicchi* et al. [6.38] found that citation distributions are universal across fields by replacing the raw number of citations with relative ones. More advanced models of citation dynamics have focused on features such as the obsolescence of knowledge, which leads to a decrease in the number of citations as a function of time [6.39, 40].

Dashun Wang et al. [6.41] have developed this idea the furthest, developing a generative model that takes into account three parameters (the number of previous citations, obsolescence, and fitness) to predict citation dynamics of individual papers. Such network approaches are also used to identify communities of research papers that frequently cite one another [6.42], a task of enormous importance for policy and evaluation. These citation networks were also used to trace the usage of words and phrases to determine whether the usage of such words corresponds to the emergence of new paradigms [6.43].

While most network models focus on a single type of node at a time, some combine a number of different types of nodes. For example, work by *Feng Shi* et al. [6.44] aims to understand how choices at the microlevel (e. g., an individual scientist's choice of topics) may constrain the development and advancement of knowledge at the macrolevel, making incremental advances/improvements to the things that are already known, rather than huge leaps to unconnected—and therefore still unimagined—futures.

A number of models focus on scientific careers. For example, *Alexander Petersen* et al. [6.45] developed a generative model showing the detrimental effect policy decisions related to the increased availability of short-term positions have on researchers' productivity levels. Work by *Albert-László Barabási* and his team uses a stochastic model to show that the timing of one's most important contribution is not the result of (academic) age, but can occur at any stage of one's career, and is a function of productivity [6.46].

Agent-based models can reveal the microprocesses of individuals that lead to particular macrolevel patterns. These models focus on the relations and interactions among entities rather than the characteristics of the entities themselves. *Nicholas Payette* [6.47] provides an excellent overview of agent-based models in the context of studying science. *Nigel Gilbert* introduced the first agent-based model to study science focused on papers (rather than authors) as agents and managed to reproduce Lotka's law of productivity, as well as the rise and decline of specializations [6.48].

Katy Börner et al. [6.49] developed a more nuanced model called TARL (topics, aging, and recursive linking), in which they simulate the simultaneous coevolution of networks of papers and authors driven by the "rich-get-richer" phenomenon [6.37, 50, 51]. *Xiaoling Sun* et al. [6.52] revisited the topic of the growth and decline of disciplines, proposing an agent-based model in which the evolution of disciplines is guided mainly by social interactions of scientists writing papers together. The disciplines thus rise and fall through the splitting and merging of communities of collaborators.

A key insight to be drawn from existing model results is that science is complex, and therefore the study of science is also complex. This complexity comes from the fact that not only are communication processes under study multilayered, but that both the data and the latent structures within the data are evolving over time. Furthermore, it is now obvious that the intricate relationships between social, conceptual, cognitive, and institutional forces need to be taken into account to fully understand the organization of science.

Despite great advances towards expanding our understanding of science, there are definite limitations in terms of predicting the emergence of a new field [6.53]. These deficiencies are due to the fact that 'normal science' is much easier to predict than 'radical innovations'. External forces (new policies, wars, etc.) have a major impact on the development of science, while data access and model development are both limited.

Current work focuses on the expansion of data sources from the traditional output of research in the forms of bibliographic data on publications, grants, and patents to the analysis of full text of those resources, grant applications (both successful and unsuccessful), mentorship (formal and informal), conferences, employment data, social media, etc. [6.53]. In parallel, algorithm development aims to capture not only strong associations but also causation [6.53, 54]. One step in that direction is using counterfactual scenarios to assess how well models perform [6.26, 55–57].

6.4 Exemplary Models of Science

This section discusses two science models in more detail. The first example describes how teams in various fields have evolved over time, and what it is they contribute to contemporary science. The second example proposes radical changes to the current funding system. Both of these models have been empirically validated, and reveal a high correlation between the simulated datasets and the structures/dynamics found in publication and funding data.

6.4.1 The Importance of Small Teams in the Big Science Era

Contemporary science is a collaborative effort within an intricate network of people, institutions, concepts, and technology. Many projects are of such complexity and scope that they require the joint efforts of many individuals with diverse expertise, culminating in teams of hundreds. Furthermore, studies suggest that large, interdisciplinary teams are more likely to produce high-impact work.

Yet only 50 years ago, the situation was very different. Most papers were written by single authors, and the largest coauthor teams did not exceed ten members. How did this change in the production of knowledge occur? How do science teams form, and what processes lead to their expansion? Perhaps most importantly, what makes a successful team?

Considering these questions, team size distribution lies at the heart of our understanding of collaborative practices and research productivity. As Fig. 6.1 shows, knowledge production today is qualitatively different from that of earlier times: *little science* performed by individuals or small groups of researchers is largely superseded by *big science* efforts conducted by large teams that span disciplinary, institutional, and national boundaries.

In Fig. 6.1, we see a change in the distribution of research team sizes in physics from a Poisson distribution to one dominated by a fat tail (a power law). In 1941–1945, for each paper with five authors, there were one thousand single-authored papers (blue). In 2006–2009, there were as many papers with five authors as there were single authored papers (red), and very large teams were not uncommon. Such a distribution (Fig. 6.1) can be reproduced using the model developed by *Staša Milojević* [6.5], which demonstrated how teams emerge, grow, and would evolve in the future.

Vitally, Milojević's model shows that team formation was, and remains, a Poisson process resulting in relatively small core teams (including single-investigator teams) carrying out certain types of research. The model also simulates the emergence of larger teams over the last 50 years in all fields of science, albeit with varying pace and magnitude of change.

According to the Milojević model, every big team originates from a small team; while some small teams do not change in size, others quickly accumulate additional members proportionally to the past productivity

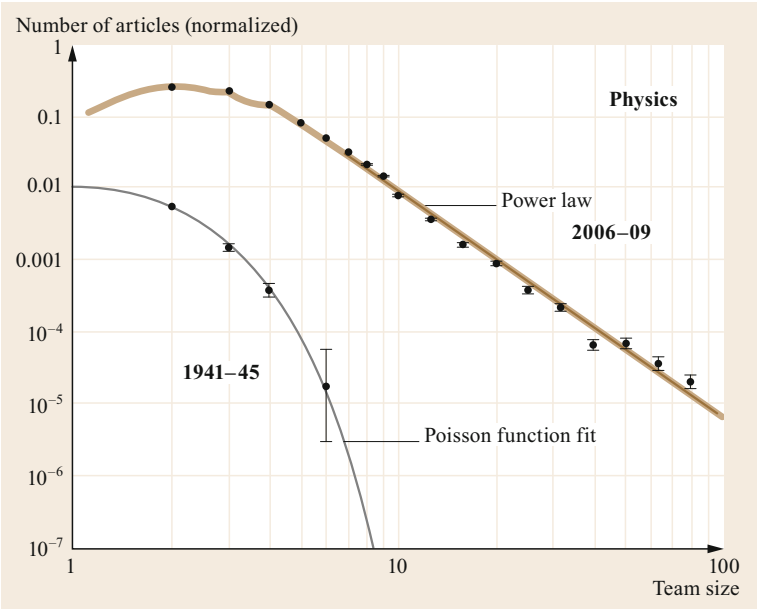

Fig. 6.1 Change in the distribution of team sizes

of preexisting team members, eventually allowing small teams to grow into big teams.

Furthermore, Milojević's model shows that relatively small teams dominate knowledge production in most fields; cumulatively, small teams still contribute more new knowledge than large teams. These findings are of key importance to policy, because they show that increased funding emphasis on large teams may undermine the very process by which large, successful teams emerge.

6.4.2 Crowdsourcing Funding Allocation

As funding agencies consume resources that could be more productively used to conduct and finance research, *Johan Bollen* et al. [6.58] argue that scholars "invest an extraordinary amount of time, energy and effort into the writing and reviewing of research proposals". In their 2014 paper, they used National Science Foundation (NSF) and Taulbee Survey data to calculate the return on investment for scholars in computer science. This calculation reveals a negative return on investment.

Given a computer and information science and engineering (CISE) funding rate of 21%, four professors working full-time for four weeks on a proposal submission with labor costs of about $35 000, five submission-review cycles may be required, resulting in a total expected labor cost of $175 000.

The average NSF grant is $164 526 per year, to which US universities charge about 50% of their overhead, leaving roughly $109 684 for research. Consequently, the four professors in question lose $65 316 of paid research time by obtaining a grant. US universities might even forbid professors to apply for grants—if they can afford to forgo the indirect dollars. Note that this simple calculation does not cover any time spent by scholars to review proposals. In 2015 alone, NSF conducted 231 000 proposal reviews to evaluate 49 600 proposals.

Bollen et al. [6.58] then go on to propose a *FundRank* model to (partially) replace the current process of government research funding allocation by expert-based crowdsourcing. In this new FundRank system, each eligible scholar (e. g., all eligible to submit NSF and National Institutes of Health (NIH) grants today) receives a certain dollar amount each year—let's say $100 000. She then needs to give a certain fraction (e. g., 50%) to colleagues that are most deserving by logging into a centralized website and entering names and amounts. In this way, scholars collectively assess each other's merit and *fund-rank* one another, with high ranking scholars receiving the most funding.

Instead of spending weeks writing and reviewing proposals, scholars are now incentivized to spend time communicating the value and impact of their past, current, and planned work so that others can judge their contributions. Using a fully digital system, conflicts of interest could be easily identified and honored; networks of mutual favors could be detected automatically, and results shared publicly.

FundRank was implemented using the recursive PageRank algorithm pioneered by *Page* and *Brin* [6.59]. Using PageRank, the "importance" (here consisting of reputation, value, and impact) of a scholar depends not only on the number of scholars that vote for her, but also their importance. The more important the scholars that link to a person, the more important the person must be. The FundRank model was validated using citation data from 37 million papers over 20 years as a proxy for how each scientist might distribute funds within the proposed system. Simulation results show funding patterns that have a similar distribution compared to NSF and NIH funding for the past decade—at a fraction of the cost required by the current system.

6.5 Challenges

Using mathematical, statistical, and computational models of STI in decision making poses a new and diverse set of challenges, many of which can be viewed as opportunities. Such challenges/opportunities are related to fundamental research, applied research, cyberinfrastructure, education, and outreach.

6.5.1 Fundamental Research

Research concerning STI is conducted across a wide range of disciplines, including (but not limited to): economics, social sciences, information sciences, science policy, scientometrics/bibliometrics, and physics. Researchers in these disciplines develop mathematical, statistical, and computational models of different types (stochastic, agent-based, epidemics, game-theoretic, network, etc.) to address the questions they are interested in.

One of the factors impeding the advancement of fundamental research is lack of free access to high-quality data. Such access would significantly reduce data curation efforts currently being done by each individual team, and as a consequence, would enable reproducibility. An additional challenge facing this type

of research is the lack of obvious sources offering continuous funding.

Furthermore, researchers exploring STI modeling tend to publish in a wide range of venues, often addressing vastly different audiences. Current research efforts and the results of said efforts are not universally known to the researchers, let alone policy makers. Such a widespread state of ignorance slows scientific progress and can even lead to unnecessary reinventions of the wheel. Scientific events that foster interactions among intellectually diverse communities and shed new light on problems by forcing researchers and practitioners to think and talk about their own research in new ways would help address this issue.

In the meantime, to arrive at policy-relevant solutions, researchers and analysts must pose good questions rather than focus solely on outcomes. Moving from descriptive to normative theories seems desirable. One major research challenge concerns the development of multiscale models—covering the micro (individual) to macro (population) levels—and understanding the appropriateness of particular models for particular scales. Ultimately, STI modeling experts should keep an open mind, and aim to learn from other branches of science (e. g., physics, economics, medicine) that are actively working on systems-science approaches.

6.5.2 Applied Research

One of the main reasons for the relatively low adoption rates of STI models is that these models are developed within different government institutions/agencies, and as a consequence often lack wider exposure. Relationships between model builders and users/stakeholders are often strained by poor communication at all stages of development, from the initial design (what question is being asked, what assumptions are being made, what measures and metrics are being used, etc.) to the interpretation and application of the results to real-world problems. This strain is further exasperated by an inherently opaque modeling process that neither creates nor maintains a sense of *buy-in* from the very beginning of a project.

For those interested in further reading, there are a few case studies that provide insights into the possibilities and challenges of carrying out applied research using modeling [6.60]: *Charles Phelps* et al., for example, implemented a tool for the measurement of the importance of vaccines—*SMART Vaccines*—which was then used by decision makers rather than model builders [6.61].

6.5.3 Cyberinfrastructure

As with many other disciplines, a robust cyberinfrastructure (e. g., data and model repositories, computing and visualization infrastructures) will greatly benefit STI modeling efforts. Many of the sciences have already setup billion-dollar international data infrastructures, and distributed computing systems in close collaboration with their government and industry partners, with impressive effect. Such synergy can be seen in the fields of meteorology (e. g., weather forecasts and hurricane and tornado prediction), epidemiology (e. g., predicting the next pandemic and identifying the best intervention strategies), climate research (e. g., predicting future scenarios and setting carbon prices), and financial engineering (e. g., stock trading and pricing predictions).

Sadly, no such universal infrastructure yet exists for the study and management of STI modeling, leading to up to 80% of project efforts being commonly spent on the acquisition, cleaning, interlinkage, and preprocessing of relevant data. Despite great benefits of building common infrastructure that we've witnessed in the natural sciences—where building a general infrastructure of commonly used data available to all has led to major advances (e. g., climate studies, astronomy, etc.)—STI modeling resources have been largely spent on individual project levels. Such a model of funding is unconducive to quick advancement of this area; successful STI modeling requires validation, iterative improvement, and a community of users, all of which could be provided via appropriate cyberinfrastructure. However, building such an infrastructure will require active partnerships among academia, government, and industry.

6.5.4 Education and Outreach

Advancing science, technology, and innovation requires extensive education and training. Recent studies show that data visualization literacy—the ability to read and write data visualizations—is relatively low [6.62]. Going forward, introducing computational modeling into formal and informal education will prove vital. More proactive and involved partnerships between stakeholders and modelers will allow for simpler models that can be understood and validated more easily. Such active partnerships will in turn help modelers deliver a timely and effective product, while also helping stakeholders determine their usefulness. At the same time, there is an urgent need for researchers, model builders and other users to enhance their communication and visualization skills.

Modeling results also need to be communicated effectively to different types of stakeholders. Storytelling and the art of communicating major results and recommendations in a clear and simple message is vital. Recent reports by the US National Academy of Sciences [6.63] and the National Academies of Sciences, Engineering, and Medicine [6.64] emphasize the importance of communication with nonscientists, and provide excellent examples of how such communication can be achieved.

6.6 Insights and Opportunities

As we have made clear, computational models of STI are deeply complex, and special effort is required to communicate not only their inner workings, but the implications of their results to relevant stakeholders. With this in mind, visualizations of data quality, data analysis, model parameter effects, and near real-time forecasts of STI developments can substantially increase the adoption rate and utility of modeling efforts.

6.6.1 Modeling Needs and Implementation

Modeling research and development strongly depend on understanding the problem at hand, as well as the range of actions a decision maker can take in response to that problem. If the wrong problem is modeled, or if suggested actions are infeasible (e. g., doubling the US R&D funding budget), then model utility as a consequence will be low.

Furthermore, there is a major difference between statistical significance and *business relevance*. For example, models used by PayPal need to avoid causing substantial costs via *false positives* (unidentified malicious users that cost PayPal money) but also via *false negatives* (valued customers with blocked accounts that cost PayPal reputation and might lead to bad press).

Model validation is of paramount importance. Ideally, different types of models can be applied to capture the structure and dynamics of that very same complex system and only if multiple models predict the same results should these results be used to make informed decisions.

Experts will need to work across disciplinary and institutional boundaries to exploit synergies, and to arrive at modeling results that are greater than the sum of their parts. There is a need for—and advantage to be gained from—combining basic and applied contract work [6.65]. Model developers (e. g., in academia and industry) should aim to *room in* with model users (policy and other decision makers) in an effort to foster active relationships.

Computational models also need to be vetted by experts, and as a consequence earn the trust of the scientific policymaking community before many start using them in practice. Key to building trust is the dogged pursuit of transparency, and also engaging stakeholders in the design and application of STI models. Easy-to-use, simple models that answer real-world questions are more readily adopted by decision makers than complex models with many parameter values.

Different policy offices have different abilities to absorb and implement models. Resistance to the adoption of new tools and approaches in general is unavoidable; the United States Federal Government, perhaps most notoriously, is the largest and most complex organization in the world, yet remains poorly understood and continues to use outdated decision support tools and processes. Models could be extremely useful when making resource allocation decisions, whether promoting agency missions, or managing international crises. Systems dynamic modeling is considered the best option, and yet not much has changed over the last decade since these approaches were first suggested. This can be attributed—simply, and frustratingly—to the human unwillingness to adapt and change.

6.6.2 Data Infrastructure

High-quality and high-coverage data is an imperative ingredient in high-quality modeling results. Currently, multiple teams are overlapping their efforts in cleaning, interlinking, and processing the same data (e. g., publication or patent data), and as a consequence are reducing the total amount of resources that could be spent on model research or validation. What's worse is that such data is preprocessed in slightly different ways across teams, making it hard or impossible to replicate results across sites.

While having so-called *big data* regarding science and technology dynamics is important to answer certain questions, having *more data* is not—and should not be—the answer to modeling questions. Though a large number of modelers use unstructured data, structured data boasts unique values, and as it becomes increasingly available [6.66] should also be explored. Given that many high-quality datasets are held by various sectors of industry (e. g., Web of Science and Scopus publication data, LinkedIn expertise profile data, Twitter or Instagram data, etc.) it appears highly desirable to work closely with them. Going forward, data sharing,

data repositories, and joint data curation efforts should be explored as universal practices.

6.6.3 Code Repository and Standards

Efficient means by which to share STI model code are essential, not only to ensure replicability and reproducibility of model results, but also to support model comparisons and effective teaching. While some teams actively use the repository GitHub.com to share code and documentation, STI models remain difficult to locate among the millions of open-source code projects stored there.

The time is ripe to focus the energies and resources of researchers on building a cyberinfrastructure and a research community to support both systematic research and development efforts. Instead of creating a new repository, it would be most efficient to build upon and extend/interlink existing model repositories. Existing data repositories can be broken into three categories: academic, government, and industry.

Academic Repositories

Academic repositories are typically associated with a tool. For example:

- *Agent Modeling Platform* (AMP) project provides "extensible frameworks and exemplary tools for representing, editing, generating, executing and visualizing agent-based models (ABMs) and any other domain requiring spatial, behavioral and functional features." (http://www.eclipse.org/amp)
- *GAMA* is a "modeling and simulation development environment for building spatially explicit agent-based simulations." (https://github.com/gama-platform)
- *NetLogo* is a "multi-agent programmable modeling environment. It is used by tens of thousands of students, teachers and researchers worldwide. It also powers HubNet participatory simulations." (http://ccl.northwestern.edu/netlogo)
- *MASON* is a "fast discrete-event multi-agent simulation library core in Java, designed to be the foundation for large custom-purpose Java simulations, and also to provide more than enough functionality for many lightweight simulation needs. MASON contains both a model library and an optional suite of visualization tools in 2D and 3D." (http://cs.gmu.edu/~eclab/projects/mason)
- The *Repast Suite* is a "family of advanced, free, and open source agent-based modeling and simulation platforms that have collectively been under continuous development for over 15 years." (http://repast.sourceforge.net)

Repositories might also be created for specific research projects. For example, SIMIAN (http://www.simian.ac.uk) funded by the Economic and Social Research Council to promote and develop social simulation in the UK, uses the SKIN model (https://github.com/InnovationNetworks/skin). Another example is OpenABM (https://www.openabm.org) that provides a growing collection of tutorials and FAQs on agent-based modeling as part of the CoMSES Network.

Government Institutions

Government institutions aim to support sharing of datasets or tools. NSF's *SciSIP* program maintains a listing of "Datasets, Graphics & Tools" pertinent to the *Science of Science Policy* (SOSP) community at http://www.scienceofsciencepolicy.net/datasets_tools.

The *Interagency Modeling and Analysis Group* (IMAG) (https://www.imagwiki.nibib.nih.gov) and the Multiscale Modeling Consortium aim to grow the field of multiscale modeling in biomedical, biological and behavioral systems, to promote model sharing and the development of reusable multiscale models, and disseminate the models and insights gained from the models to the larger biomedical, biological, and behavioral research community, among others.

The *Predictive Model Index* lists over 100 reusable, sharable models in support of reproducible science (https://www.imagwiki.nibib.nih.gov/model-indexing). The *Centers for Disease Control and Prevention* (CDC) made the "H1N1 Flu (Swine Flu): Preparedness Tools for Professionals" software available at http://www.cdc.gov/h1n1flu/tools. The page was developed during the 2009–2010 H1N1 pandemic, but it has not been updated, and is being archived for historic and reference purposes only.

Publishers typically aim to ensure replicability of work by asking authors to submit datasets and models. Examples are The *Journal of Artificial Societies and Social Simulation* (JASSS, http://jasss.soc.surrey.ac.uk/JASSS.html), an interdisciplinary journal for the exploration and understanding of social processes by means of computer simulation; published since 1998, *JASSS* recommends authors upload model code and associated documentation to the CoMSES Net Computational Model Library (https://www.comses.net/codebases/). As of June 2016, the CoMSES library featured 352 agent-based models.

Industry

Industry has long embraced big data and advanced data mining, modeling, and visualization algorithms. Computational models are widely used in online recommendation services (e. g., those provided by Amazon

or Netflix), and by financial and insurance companies (e. g., to detect credit card fraud and estimate fees). Many companies use models internally to support strategic decision making, and to guide investment decisions. While code is typically proprietary, close industry-academia-government collaborations are likely beneficial for all parties involved.

6.6.4 Visualization and Communication of Modeling Results

Global operation rooms that provide visualizations of current data and predictions of possible futures (already commonplace in the fields of meteorology, finance, epidemiology, and defense) might soon be commonplace in support of funding, strategic intelligence, or policy decision making.

William Rouse has been pioneering "policy flight simulators" that let decision makers fly the future before they write the check [6.10]. His team uses a combination of commercial off-the-shelf tools (e. g., AnyLogic, D-3, Excel, R, Simio, Tableau, and Vensim) rather than writing software from scratch. This practice can enable creation of a prototype interactive environment within a week or two, which in turn allows rapid user feedback and easy midcourse corrections.

Meanwhile, *Ben Shneiderman* and his team developed EventFlow, a novel tool for event sequence analytics that includes a timeline display showing all individual records, their point and interval events, as well as an aggregated view of all the sequences in the dataset (http://hcil.umd.edu/eventflow) [6.67]. Among others, the tool supports the examination of data quality before any type of data analysis is conducted or visualizations are rendered—blind usage of data is dangerous.

Storytelling in particular provides a powerful means to communicate data analysis and modeling results [6.63, 68]. Merging data with narrative, especially when communicating the value of research, is a primary way to connect to policymakers.

Katy Börner and her team are developing and prototyping *Science Forecasts*; a news show that communicates local and global developments in science, technology, and innovation to a general audience. In Spring 2015, a pilot episode was recorded featuring a moderator that explained trends using an animated map of science, analogous to a weather forecast. Zeroing in on specific research results using Twitter for detecting episodes of depression, the information was presented by Johan Bollen and Fred Cate, both faculty at Indiana University. A still image of the news can be seen in Fig. 6.2.

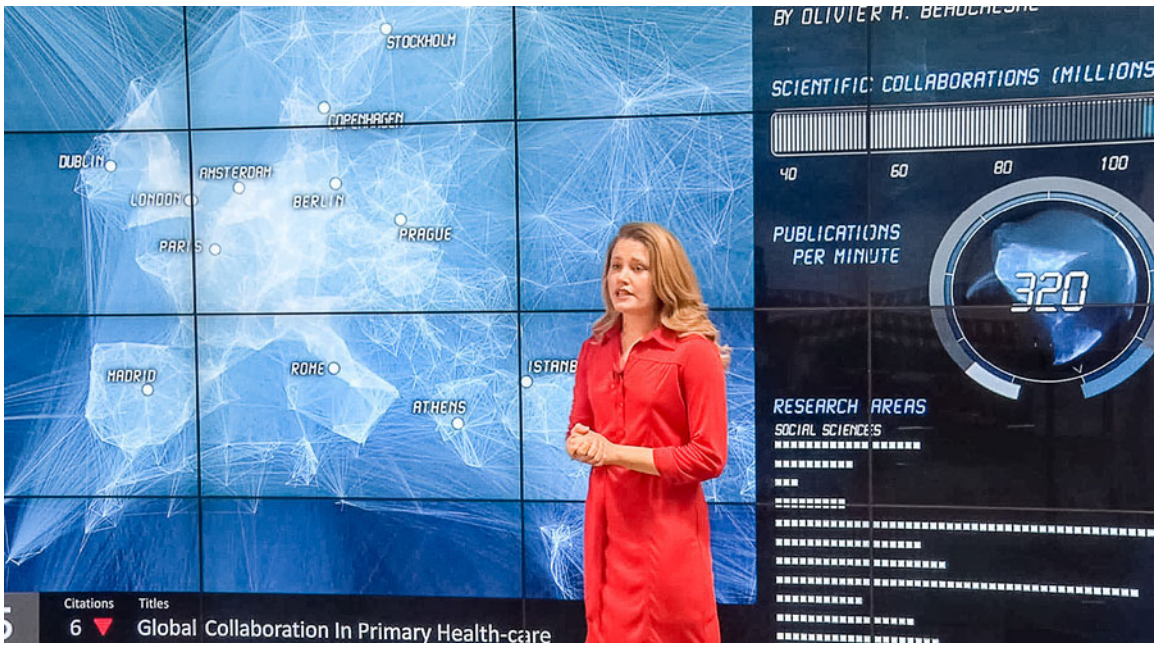

Fig. 6.2 *Science Forecast*, recorded at Indiana University, presents interviews and animated maps of scientific activity in a manner similar to weather forecasts. The program demonstrates the power of data and visual analytics to provide up-to-date stories on science trends and developments

6.7 Outlook

In 2007, *Issues in Science and Technology* published *The Promise of Data-Driven Policymaking* by *Daniel Esty* and *Reece Rushing* [6.69]. In 2016, the same magazine published "Data-Driven Science Policy" [6.7]. The articles both point out that in the corporate sector, a wide variety of data-driven approaches are used to boost profits, including systems that improve performance and reliability, evaluate the success of advertising campaigns, and determine optimal pricing. Both articles argue for the need for—and discuss the premise of—data-driven decision making and policy making in STI using large-scale, high-quality datasets, and computational means to inform human decision makers.

Today, in 2019, a wide range of mathematical, statistical, and computational models exist that were developed and implemented in a variety of settings to increase our collective understanding of the structure and dynamics of STI and to support human decision making. While academic researchers typically focus on work that can be published, there is a growing emphasis by researchers and practitioners on the power of models to advance future decision making, and to communicate the usefulness of models to simulate, explain, and communicate the past, present, and future.

Acknowledgments. The work was supported in part by the National Institutes of Health under awards U01CA198934, P01AG0393, and OT2OD026671 and National Science Foundation awards 1546824, 1713567, 1735095, and 1839167, NETE Federal IT, Thomson Reuters, Indiana University Network Science Institute, and the Cyberinfrastructure for Network Science Center at Indiana University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

6.1 P. Ahrweiler, N. Gilbert, A. Pyka (Eds.): *Joining Complexity Science and Social Simulation for Innovation Policy: Agent-based Modelling using the SKIN Platform* (Cambridge Scholars, Newcastle upon Tyne 2015)
6.2 C. Watts, N. Gilbert: *Simulating Innovation. Computer-Based Tools for Re-Thinking Innovation* (Edward Elgar, London 2014)
6.3 D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, I. Rafols: The Leiden Manifesto for research metrics, Nature **520**, 430–431 (2015)
6.4 D.J. de Solla Price: *Little Science, Big Science* (Columbia Univ. Press, New York 1963)
6.5 S. Milojević: Principles of scientific research team formation and evolution, Proc. Natl. Acad. Sci. USA **111**(11), 3984–3989 (2014)
6.6 S. Wuchty, B.F. Jones, B. Uzzi: The increasing dominance of teams in production of knowledge, Science **316**(5827), 1036–1039 (2007)
6.7 K. Börner: Data-driven science policy, Issues Sci. Technol. **32**(3), 26–28 (2016)
6.8 A. Scharnhorst, K. Börner, P. van den Besselaar (Eds.): *Models of Science Dynamics: Encounters between Complexity Theory and Information Science* (Springer, Berlin 2012)
6.9 J. de Rosnay: *The Macroscope: A New World Scientific System* (Harper Row, New York 1979)
6.10 W.B. Rouse: Human interaction with policy flight simulators, J. Appl. Ergon. **45**(1), 72–77 (2014)
6.11 W.B. Rouse: *Modeling and Visualization of Complex Systems and Enterprises: Explorations of Physical, Human, Economic, and Social Phenomena* (Wiley, Hoboken 2015)
6.12 P. Ahrweiler, M. Schilperoord, A. Pyka, N. Gilbert: Modelling research policy: Ex-ante evaluation of complex policy instruments, J. Artif. Soc. Soc. Simul. **18**(4), 5 (2015)
6.13 C.A. Lave, J.G. March: *An Introduction to Models in the Social Sciences* (Univ. Press of America, Lanham 1993)
6.14 J.H. Miller, S.E. Page: *Complex Adaptive Systems: An Introduction to Computational Models of Social Life* (Princeton Univ. Press, Princeton 2007)
6.15 P. Cilliers: *Complexity and Postmodernism: Understanding Complex Systems* (Routledge, London 1998)
6.16 G. Nicolis, I. Prigogine: *Exploring Complexity: An Introduction* (W.H. Freeman Company, New York 1989)
6.17 J.H. Holland, K.J. Holyoak, R.E. Nisbett, P.R. Thagard: *Induction: Processes of Inference, Learning, and Discovery* (MIT Press, Cambridge 1986)
6.18 E. Winsberg: *Science in the Age of Computer Simulation* (Univ. Chicago Press, Chicago 2010)
6.19 M. Morrison: *Reconstructing Reality: Models, Mathematics, and Simulations* (Oxford Univ. Press, Oxford 2015)
6.20 P.N. Edwards: *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, Cambridge 2010)
6.21 J.M. Epstein: *Generative Social Sciences: Studies in Agent-Based Computational Modeling* (Princeton Univ. Press, Princeton 2006)
6.22 N. Gilbert: *Agent-based Models* (SAGE, Los Angeles 2008)
6.23 N. Gilbert, K.G. Troitzsch: *Simulation for the Social Scientist*, 2nd edn. (Open Univ. Press, Maidenhead 2009)

6.24 T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer, New York 2017)
6.25 K.P. Murphy: *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge 2012)
6.26 J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer: Prediction policy problems, Am. Econ. Rev. Papers Proc. **105**(5), 491–495 (2015)
6.27 K. Börner, W. Glänzel, A. Scharnhorst, P. van den Besselaar: Modeling science: Studying the structure and dynamics of science, Scientometrics **89**(1), 346–463 (2011)
6.28 K. Börner, B. Edmonds, S. Milojević, A. Scharnhorst: Simulating the processes of science, technology, and innovation, Scientometrics **110**(1), 385 (2016)
6.29 K. Börner, K.W. Boyack, S. Milojević, S.A. Morris: An introduction to modeling science: Basic model types, key definitions, and a general framework for comparison of process models. In: *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*, ed. by A. Scharnhorst, K. Börner, P. van den Besselaar (Springer, Berlin 2012) pp. 3–22
6.30 J. Smith, C. Jenks: *Qualitative Complexity: Ecology, Cognitive Processes and the Re-Emergence of Structures in Post-Humanist Social Theory* (Routledge, London 2006)
6.31 E.C.M. Noyons, A.F.J. van Raan: Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research, J. Am. Soc. Inf. Sci. **49**(1), 68–81 (1998)
6.32 A.F.J. van Raan: Fractal dimension of co-citations, Nature **347**, 626 (1990)
6.33 A.F.J. van Raan: On growth, ageing, and fractal differentiation of science, Scientometrics **47**(2), 347–362 (2000)
6.34 W.O. Kermack, A.G. McKendrick: A contribution to the mathematical theory of epidemics, Proc. R. Soc. A **115**, 700–721 (1927)
6.35 W. Goffman: Mathematical approach to the spread of scientific ideas – The history of mast cell research, Nature **212**(5061), 449–452 (1966)
6.36 D.J. de Solla Price: Networks of scientific papers, Science **149**, 510–515 (1965)
6.37 D.J. de Solla Price: A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inf. Sci. **27**(5), 292–306 (1976)
6.38 F. Radicchi, S. Fortunato, C. Castellano: Universality of citation distributions: Toward an objective measure of scientific impact, Proc. Natl. Acad. Sci. USA **105**, 17268–17272 (2008)
6.39 Y.-H. Eom, S. Fortunato: Characterizing and modeling citation dynamics, PLoS ONE **6**(9), e24926 (2011)
6.40 P.D.B. Parolo, R.K. Pan, R. Ghosh, B.A. Huberman, K. Kaski, S. Fortunato: Attention decay in science, J. Informetrics **9**(4), 734–745 (2015)
6.41 D. Wang, C. Song, A.-L. Barabási: Quantifying long-term scientific impact, Science **342**(6154), 127–132 (2013)
6.42 R. Klavans, K.W. Boyack: Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?, J. Assoc. Inf. Sci. Technol. **68**, 984–998 (2016)
6.43 T. Kuhn, M. Perc, D. Helbing: Inheritance patterns in citation networks reveal scientific memes, Phys. Rev. X **4**, 041036 (2014)
6.44 F. Shi, J.G. Foster, J.A. Evans: Weaving the fabric of science: Dynamic network models of science's unfolding structure, Soc. Netw. **43**, 73–85 (2015)
6.45 A.M. Petersen, M. Riccaboni, H.E. Stenley, F. Pammolli: Persistence and uncertainty in the academic career, Proc. Natl. Acad. Sci. USA **109**, 5213–5218 (2012)
6.46 R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási: Quantifying the evolution of individual scientific impact, Science **354**(6312), aaf5239 (2016)
6.47 N. Payette: Agent-based models of science. In: *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*, ed. by A. Scharnhorst, K. Börner, P. van den Besselaar (Springer, Berlin 2012) pp. 127–157
6.48 N. Gilbert: A simulation of the structure of academic science, Sociol. Res. Online (1997), https://doi.org/10.5153/sro.85
6.49 K. Börner, J. Maru, R. Goldstone: The simultaneous evolution of author and paper networks, Proc. Natl. Acad. Sci. USA **101**, 5266–5273 (2004)
6.50 A.-L. Barabási, R. Albert: Emergence of scaling in random networks, Science **286**, 509–512 (1999)
6.51 R.K. Merton: The Matthew effect in science, Science **159**(3810), 56–63 (1968)
6.52 X. Sun, J. Kaur, S. Milojević, A. Flammini, F. Menczer: Social dynamics of science, Sci. Rep. **3**, 1069 (2013)
6.53 A. Clauset, D.B. Larremore, R. Sinatra: Data-driven predictions in the science of science, Science **355**, 477–480 (2017)
6.54 P. Azoulay: Turn the scientific method on ourselves, Nature **483**, 31–32 (2012)
6.55 P. Azoulay, J.G. Zivin, G. Manso: Incentives and creativity: Evidence from academic life sciences, RAND J. Econ. **42**(3), 527–554 (2011)
6.56 O.A.D. Arrieta, F. Pammolli, A.M. Petersen: Quantifying the negative impact of brain drain on the integration of European science, Sci. Adv. **3**, e1602232 (2017)
6.57 J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan: Human decisions and machine predictions, NBER Working Paper No. 23180 (2017)
6.58 J. Bollen, D. Crandall, D. Junk, Y. Ding, K. Börner: From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review, EMBO Reports **15**(2), 131–133 (2014)
6.59 L. Page, S. Brin: The anatomy of a large-scale hypertextual web search engine, Comput. Netw. and ISDN Syst. **30**(1–7), 107–117 (1998)
6.60 K. Börner, S. Milojević (Eds.): *Modeling Science, Technology and Innovation. NSF Conference Report, Indiana University*, https://modsti.cns.iu.edu/report and presenter slides at https://modsti.cns.iu.edu (2016)
6.61 C. Phelps, G. Madhavan, K. Sangha, R. Rappuoli, R.R. Colwell, R.M. Martinez, L. King: A priority-setting aid for new vaccine candidates, Proc. Natl. Acad. Sci. USA **111**(9), 3199–3200 (2014)

6.62 K. Börner, J.E. Heimlich, R. Balliet, A.V. Maltese: Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors, Inf. Vis. **15**(3), 198–213 (2016)
6.63 US National Academy of Sciences: *The Science of Science Communication II: Summary of a Colloquium* Retrieved from Washington (2014)
6.64 National Academies of Sciences, Engineering, and Medicine: *Communicating Science Effectively: A Research Agenda* Retrieved from Washington (2017)
6.65 B. Shneiderman: *The New ABCs of Research: Achieving Breakthrough Collaborations* (Oxford Univ. Press, Oxford 2016)
6.66 J. Hendler, W. Hall: Science of the World Wide Web, Science **354**(6313), 703–704 (2016)
6.67 M. Monroe, R. Lan, C. Plaisant, B. Shneiderman: Temporal event sequence simplification, IEEE Trans. Vis. Comput. Graph. **19**(12), 2227–2236 (2013)
6.68 E. Segel, J. Heer: Narrative visualization: Telling stories with data, IEEE Trans. Vis. Comput. Graph. **16**(6), 1139–1148 (2010)
6.69 D. Esty, R. Rushing: The promise of data-driven policymaking, Issues Sci. Technol. **23**(4), 67–72 (2007)

Katy Börner

School of Informatics, Computing, and Engineering
Indiana University
Bloomington, IN, USA
katy@indiana.edu

Katy Börner is Victor H. Yngve Distinguished Professor of Engineering and Information Science in the School of Informatics, Computing, and Engineering, Adjunct Professor at the Department of Statistics in the College of Arts and Sciences, Core Faculty of Cognitive Science at Indiana University. She holds an MS in Electrical Engineering from the University of Technology in Leipzig (1991) and a PhD in Computer Science from the University of Kaiserslautern (1997).

Staša Milojević

School of Informatics, Computing, and Engineering
Indiana University
Bloomington, IN, USA
smilojev@indiana.edu

Staša Milojević is an Associate Professor of Informatics in the Center for Complex Network and Systems Research in the School of Informatics, Computing, and Engineering at Indiana University. Her work covers a range of topics within the "science of science": dynamics of research teams, collaborative networks, formation and evolution of scientific fields, and research metrics. She received a PhD in Information Studies from the University of California, Los Angeles.