Computational History: Challenges and Opportunities of Formal Approaches

JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 05/06 pp232−242
Volume 4, Number 3, September 2023
DOI: 10.23919/JSC.2023.0017

Computational History: Challenges and Opportunities of Formal Approaches

Jürgen Jost*, Roberto Lalli, Manfred D. Laubichler, Eckehard Olbrich, Jürgen Renn, Guillermo Restrepo, Peter F. Stadler, and Dirk Wintergrün

Abstract: We propose a program for a computational analysis, based on large scale datasets, of deep conceptual and formal structures, representing the mechanisms of historical transformations in different domains ranging from biological to social, cultural, and knowledge systems. We conceptualize such systems as consisting of complex multi-layer networks. Structural properties of such networks may explain the spreading of innovations. Temporal relations between the dynamics of interacting networks may help to identify causalities. Complex systems may show path and context dependencies. We illustrate our approach by case studies from all those types of systems.

Key words: computational history; history of science; network analysis; big data

1 Introduction

Historical processes—in biological, social, cultural, and knowledge systems—are governed by mechanisms that transform the structure, dynamics, and function of complex systems. Historians have traditionally described these transformations in form of narratives that suggest implicit causal structures. However, even the most successful narratives, building on cumulative insights of many scholars, are limited by the ability of individuals to manage and manipulate evidence and to deal with complex interdependencies[1, 2]. On the other hand, as in other areas of inquiry, mathematical formalisms and statistical methods reveal deeper conceptual structures that have allowed scientists in various different disciplines to represent and manipulate data and therefore gain insights at larger scales and higher complexity[3]. In the age of big data, including big data about history, the need to further advance these methods and formalisms also for historical analysis has become more important and urgent.

Based on the assumption that an understanding of historical processes also benefits from formal representations and mathematical models, we propose a program for a computational analysis, based on large scale datasets, of deep conceptual and formal, structures, representing the mechanisms of historical transformations in different domains ranging from biological to social, cultural, and knowledge systems. This endeavor is intended to complement traditional case studies based historical scholarship[1, 4]. Furthermore, the development of methods as well as new formal structures representing these complex processes, can only be realized in close collaboration with historians and social scientists as domain knowledge is an essential part of the formal analysis at different levels[5]. This includes questions about data

• Jürgen Jost, Eckehard Olbrich, Guillermo Restrepo, and Peter F. Stadler are with the Max Planck Institute for Mathematics in the Sciences, Leipzig 04103, Germany. E-mail: {jjost, restrepo}@mis.mpg.de.
• Jürgen Jost, Manfred D. Laubichler, and Peter F. Stadler are with the Santa Fe Institute, Santa Fe, NM 87501, USA.
• Roberto Lalli, Manfred D. Laubichler, Jürgen Renn, and Dirk Wintergrün are with the Max Planck Institute for the History of Science, Berlin 14195, Germany.
• Manfred D. Laubichler is also with the School of Complex Adaptive Systems and the Global Biosocial Complexity Initiative, Arizona State University, Tempe, AZ 85287, USA.
• Peter F. Stadler is also with the Department of Computer Science and Interdisciplinary Center for Bioinformatics, Leipzig University, Leipzig 04109, Germany.
* To whom correspondence should be addressed.
Manuscript received: 2023-06-02; accepted: 2023-10-15

structures and their formal representations as well as questions about formal analysis of transformations and inferences about their governing mechanisms[6, 7]. All of these formal issues need to be constrained and parameterized with domain specific knowledge.

Dynamical processes in many fields have a historical aspect insofar as they exhibit strong path dependence and are shaped by internal fluctuations and contingencies and/or external perturbations[3, 8, 9]. Formal representations of such processes have a long history in domains such as evolutionary biology, economics, and the physical sciences. Over the years, this has led to a frequent transfer of such ideas into the historical and social sciences[10]. In many cases, this transfer was both naïve and metaphorical and predicated on a vague notion of similarity between processes in different domains. Here our project is taking a radically different approach. We are not primarily interested in transfer of methods and formalisms between domains, but rather focus on the development of formalisms, mathematical structures, and methods that can adequately represent and analyze complex historical processes[11, 12]. Our goal is to introduce formal mathematical thinking into the domains of history and related fields, which, given the complexity of the subject matter, we fully expect it will also stimulate new developments in the formal and mathematical sciences. Likewise, we aim at presenting the computational approaches to incorporate mathematical formalisms to process, model, and analyse historical data.

For our project, the general formal challenge can be represented as shown in Fig. 1.

We start from two assumptions:

(1) Substantial parts of any historical transformation process are governed by identifiable fundamental mechanisms.

(2) Historical situations can be described as complex and multilayer networks[13, 14].¤

This allows the formalization of historical developments as reconfiguration of these dynamical networks and provides us with a sense of what kind of causal processes can influence these systems, what types of historical processes can exist (for example, genealogical links or division/fusion processes, etc).

¤We use the term *network* here to mean any formal structure that describes objects together with interactions between them, i.e., not just as a synonym of *graphs*.

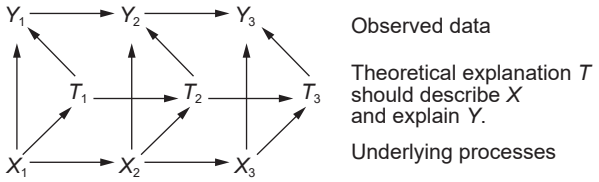


Fig. 1 Relation between data, processes, and explanations.

These are modeled e.g., by transformations, addition or deletion of nodes and/or edges, and topological features in enabling and constraining transformations. In a more general sense, it allows for studying subsets of nodes and relations among subsets in the form of hypergraphs[15]. Examples involve interactions of social structures such as academe-industry, which can be further studied at a higher level of interaction with other network layers[16]. For instance, the semiotic layer, where communication symbols and channels exist, a major goal here is not to just describe patterns, but to gain insights into generative mechanisms that will allow us to explore spaces of possible transformation trajectories and also the patterns and processes of actual historical transformations in a systematic way.

There is a tradition of quantitative history working with data such as demographics, population health, economic, and financial data, and data from empirical social science[17–20]. Sophisticated statistical methods have been developed and applied to analyze such data[21, 22]. Computational history cannot be reduced to quantitative history, although it happily utilizes those data and methods as well. First of all, computational history develops different frames for representing those data. It is interested in interaction patterns, formalized via networks or other more refined mathematical structures, that also represent interactions between more than two entities or hierarchical patterns of interaction. This has not been a traditional focus of quantitative history. Moreover, it draws upon modern tools for the analysis of time series, to find tipping points in historical dynamics and to identify their roots. These roots could intertwine individual innovations and network percolation effects that lead to critical complexity gains or losses. Another aim of computational history is to identify statistical signatures and regularities at aggregate levels across different cultures and periods.

The structure of data and their representation poses a challenge. The data structures for historical processes

in biological, social, cultural, and knowledge systems are intrinsically highly complex and there is a desperate need for systematic analysis. This includes such important questions as coarse graining—i.e., what are the functionally relevant entities for historical processes (not always those that can readily be observed)[23], data representation and compression, and the underlying, often implicit, ontologies that structure databases. Conversely, more readily available data will often correlate or describe epi-phenomena instead of directly representing the entities that are subject to causal processes of historical change. This requires a deeper understanding of the role and properties of "projections" relating intrinsic processes to observable variables[24] (see Fig. 1). Such projections may effectively be suitably defined coarse grainings[25]. In particular, we need new conceptual and formal tools for the integration of facts and knowledge about many different specific subsystems beyond the traditional method of historical narratives.

Our methodology makes it necessary to systematically explore and further advance various methods for the analysis of large complex datasets. This is closely connected to a better theoretical understanding of machine learning and other statistical approaches such as topic modeling[26]. The challenge here is to go beyond the current pragmatic orientation ("it works, it is good enough") and to connect such procedures to the active discovery of underlying mechanisms and formal structures of knowledge and perception.

2 Elements of Computational History

Here we discuss some key concepts of computational history and illustrate each with examples from our research.

2.1 Complexity measure

In order to describe the transformations of complex systems and networks, we need to first be able to measure the degrees of complexity and their changes. Such measures have been developed to quantify the degree of interdependencies between the parts or components of a system or the scope of dependencies in a process[27]. In a pilot project[28], we have applied such complexity measures to literary texts, plays of Shakespeare, and their translations into another language (German).

The linguistic complexity quantified after how many words the information gain from the next word was the highest, and the larger that number, the more complex the text. As expected, Shakespeare's texts achieve particularly high values, but what is perhaps more useful, this allowed for a systematic comparison of the different translations that agreed well with more traditional judgements of literary quality. It is also possible to compare these measures between different historical periods, which allows us to investigate transformations in literary styles. Similarly, evaluating action sequences, the most important characters in a play could be identified via a purely formal analysis. For instance, although Caesar is murdered rather early in *Julius Caesar*, our complexity measure easily spotted him as the key personality of that tragedy. Of course, texts could become overly complex and thereby incomprehensible. Analogously, in interacting dynamical networks, one of them could become complex beyond its own processing capabilities, or provide an input to another system that is too complex for that system.

Complexity measures thus make it possible to investigate whether an increase of complexity beyond a certain threshold in one domain, say the socio-economic one, triggered catastrophic events in another, when for instance, the political system was no longer able to handle that complexity.

2.2 Evolution and historical process

Biology has a tradition of organizing large datasets about historical patterns and processes in a theory driven manner, for instance, in the Linnean classification system or the arrangement of fossils in phylogenetic trees[29–32]. In addition, the field of bioinformatics is highly successful in handling and analyzing datasets produced by molecular biologists through sophisticated chemical and physical measurements[33]. These are annotated, aggregated, and interpreted in terms of biological functions and their consequences, e.g., for human health[34]. This is where similarities between computational history and computational biology emerge.

A closer look at both the history and the current situation of biological data analysis may therefore be useful. The functionalist-structuralist debate, while not

unique to biology (viz. for instance, the contrast between the functionalist approach of Malinowski[35] and the structuralist scheme of Lévi-Strauss[36] in ethnology), has a long history. More than 200 years ago, the functionalist Georges Cuvier, one of the founders of paleontology, was able to use his principles of the correlation of parts and the conditions of existence to reconstruct an entire extinct mammal from a small set of fossil bones. His idea was that all parts of an organism are functionally adapted to its mode of existence, for instance, that of a large herbivore[37–39]. Analogously, the institutions and customs of a nomadic trading society should be characteristically different from that of a sedentary agricultural one. Such differences should be visible in datasets.

Cuvier's antagonist, the structuralist Geoffroy St. Hilaire, in contrast argued that all animal branches, such as the vertebrates or the arthropodes, had their characteristic bauplan, and that knowing that bauplan allowed for inferences about their features[37, 40–42]. From a later evolutionary perspective, both positions are valid: organisms on one hand share ancestral traits, and on the other hand have specific selected functional adaptations.

Similar mixtures of ancestral heritage and path dependency and functional diversification can also be observed in human societies. It was perhaps not a coincidence that biology and linguistics both gained a historical perspective and developed an interest in the reconstruction of ancestors and the construction of corresponding classifications (i.e., phylogenies) at about the same time. Both encountered the difficulty that the discriminatory features depended on the class in questions. In biology, this ranges from the number and arrangement of petals in the Linnean system to the form of the digits for various mammalian branches. In linguistics, the discriminatory grammatical features depend on the language family in question. Likewise, art historians use different features for stylistic analysis in different cultures and historical periods.

Recently, however, biology went through a different phase. It tried to use the same type of data that could be collected in a standardized manner and stored in comprehensive data banks, in particular DNA or protein sequences, across the entire biosphere[43]. For instance, phylogenetic relations are inferred from similarities in genetic sequences[44]. The uniformity and relative simplicity of DNA and protein sequence data were an important factor in the early establishment of centralized data repositories and the emergence of a community consensus requiring data deposition as integral part of the publication process. Methods development in bioinformatics, in turn, was enabled by the availability of extensive data bases.

We do not currently see any such universal data type in the humanities, although there are already efforts in that direction with the International Institute of Social History[45], whose databases offer, for example, information on thousands of occupational titles from countries and languages around the world from the 16th to the 20th century. Likewise, Clio Infra[46] holds interconnected databases containing worldwide data on social, economic, and institutional indicators for the past five centuries, with special attention to the past 200 years. In any case, also in biology, it becomes evident that such data need to be supplemented by semantic interpretations, such as gene ontologies. A key aim is to derive functional interpretations of the genetic and other structures since progress eventually comes from a better understanding of the biophysical mechanisms and biological functions. In computational biology/bioinformatics, this agenda is increasingly supported by statistics and machine learning approaches that encapsulate models of biological systems and turn the "interpretation" of the primary data into a computational problem[47]. At a conceptual level, this does not seem so different from what one would like to do with the data collected in the humanities.

2.3 Dynamics of multi-scale networks

Historical processes transform networks. Reconstructing these transformations is thus a core feature of historical analysis. As these networks are mostly large and complex, involve multiple scales, and need to be reconstructed based on incomplete data, such reconstructions require sophisticated formal representations. For example, paleontologists have provided us with detailed reconstructions of extinct species for centuries. A formal approach similar to Cuvier's logic of functional interdependencies, as described above, has more recently been applied to the reconstruction of ancient ecosystems and food webs[48, 49].

Again, these reconstructions are guided by insights into the topology and multi-scale network structures of

these systems, including a set of possible transformation rules[50]. While the historical reconstruction of biological systems is based on material evidence (fossils) or models reconstructed from present-day DNA sequences, generalized interaction rules (functional network topologies) and patterns of transformation (developmental and evolutionary principles), an additional source of information in human history are text-based archives. These are increasingly available in digital formats. Applying a range of methods from computational linguistics and network analysis to text corpora in the history of science, we could model the growth and differentiation of these discourses[51], distinguish individual discourses through disambiguation techniques, identify sub-clusters of concepts that represent incipient specialization or novel dimensions of discourses, and identify the origin of novelties within these corpora. Time series analyses[52] of these developments give us a dynamic view of change on the semantic or idea level. But we can also link these content-based analyses with the underlying social networks that represent the actual historical actors. Doing this one can, in the case of scientific discourses, identify and measure degrees of interdisciplinarity, the social structures of science, and how these relate to the origin and spread of ideas. This allows us to identify key individuals and institutions as well as socially constructed impediments to the development of science.

These approaches are not restricted to the history of knowledge but can be applied to all textual corpora with structured data and metadata. (For unstructured data one can apply machine leaning techniques.) In our study of evolutionary medicine we could, for instance, demonstrate that this field (as well as many others) exhibits a so-called rich club structure, that innovations tend to originate at the periphery of the social network, rather than within the rich club and that innovative papers have on average higher citation rates[53, 54]. Recent advances[16] also indicate that small teams tend to have more disruptive discoveries. We developed formal criteria and metrics for such observations that can now be applied in other contexts and therefore enable comparisons. The ultimate goal of these studies is to investigate whether there are similar generalities in the reconstructions of these human systems as have been observed in biological systems.

2.4 Socio-epistemic networks

In another example, an analysis of the development of general relativity between 1925 and 1970[55], we further refined the framework of multi-layered network analysis. This framework defines knowledge networks as being composed of three different layers: the social network, the semiotic network, and the semantic network[56]. The social network is defined as the collaboration network of scientists who worked on general relativity in the period under consideration. The co-citation network of papers in research areas related to general relativity is considered a proxy of the semiotic network, which is the collection of the material or formal representations of knowledge. Finally, we explored the network of words in the full-texts of the cited papers. This network is understood as a proxy of the semantic network, which is defined as a collection of knowledge elements and its relations.

This data-driven computational approach was used to uncover the mechanism of the passage between the low-water-mark of general relativity (roughly from the mid-1920s, to the mid-1950s) and so-called renaissance of the theory after the mid-1950s. Based on this multi-layer analysis, we obtained substantial evidence that between the second half of the 1950s and the early 1960s, there was an evident shift in all three layers disproving common explanations of the renaissance process. It shows that this phenomenon was not a consequence of astrophysical discoveries in the 1960s, nor was it a simple by-product of socio-economic transformations in the physics landscape after World War II.

We argue instead that the renaissance has to be understood as a two-phase process both at the social and at the conceptual level. The first occurred between the second half of the 1950s and the early 1960s and was characterized by a return of interest in physical problems in general relativity proper for a growing community of scientists, while the previous period was characterized by a dispersion of research agendas aimed at going beyond the theory. We call this phase the renaissance of the theory. The second period, which we call the astrophysical turn, was instead an experiment-driven process that started with the discovery of quasars and was characterized by the emergence of relativistic astrophysics and physical cosmology as

well as the early phases of gravitational-wave astronomy. Again, we developed a set of formal criteria for the analysis of this multi-layered network dynamics that can now be applied to other cases.

2.5 Formal spaces and their interactions

A space is simply a set of objects endowed with some addition structures describing "nearness"[57]. Organizing sets of facts in such "spaces" not only can draw on our geometric intuition, but also makes a host of mathematical tools available, as soon as we embark to make the structure of space explicit. For instance, the *chemical space*[58, 59] comprises substances, i.e., types of molecules, and organizes them by a measure of similarity between the molecular structures, or a notion of reachability by synthetic steps required to connect any two chemical substances. In the same vein, similarities of words measured in terms of co-occurrences or co-translations organize the lexicon of a language into space in which one can start to argue about semantic shifts in language evolution[60].

Computational history of chemistry[7] offers also interesting results and challenges for computational history in general. Chemical space has grown with a stable rate for more than 200 years, doubling the number of reported substances every 16 years[58]. The analysis of this global statistic revealed that external setbacks such as World Wars have not perturbed the expansion of the chemical space in the long run. A time series analysis of the annual output of compounds showed two major transitions, one around 1860 and the other about 1980[58]. There is strong evidence indicating that the first one had internal causes, namely the crisis of the semiotic system of chemical formulas, which was followed by the incorporation of the molecular structural theory with their molecular structures[2, 61]. This semiotic expansion reduced the internal complexity of chemistry, as molecular formulas expanded the limited combinatorial possibilities of the Berzelian formulas. Structural formulas allowed for accommodating the deluge of new organic compounds and even for estimating new ones[2]. The second major transition in the expansion of the chemical space coincides with a surge in the number of organometallic compounds followed by a rise of bioorganic compounds[58]. It is still an open question to understand the driving forces leading to this transition. Future questions for a computational history of chemistry entail determining the workings of the interaction of the social, semiotic, and material systems of chemistry[2]. This application of computational history methods to chemistry showed a further benefit, namely testing historical narratives. It is traditionally accepted that chemical synthesis began after 1828 Wöhler's famous synthesis of urea[62, 63]. However, our data-driven study showed that chemical synthesis was a major provider of novel substances as early as the dawn of the 19th century[58]. In another study[64], we could analyze how the accumulation of chemical knowledge since the 1840s made the discovery of the periodic system possible and how this knowledge has driven the historical unfolding of the periodic system up to date[65].

Another instance is issue, argument, and conceptual spaces in cultural discussions. In the ODYCCEUS project (https://www.odycceus.eu), we have studied the new cleavage due to economic, political, and cultural globalization that has reconfigured the political space in most of the liberal democracies around the world with a specific emphasis on the impact of the social media platforms on these processes. We can then ask along which lines social and other systems will fracture and where new boundaries will form and perhaps old ones get dissolved between subgroups. Models of the interaction of platform of political parties and the opinions of the electorate and the players therein have a long tradition[66]. Importantly, the structure of the "opinion space" cannot be separated from the way in which the dynamics is modelled. It is important, therefore, to understand the intrinsic structure of these spaces, which—we suspect—will be such that the dynamical processes living become amenable to comprehensible, causal models. As in the case of chemical synthesis, where new molecules are always variations on what has been produced before, we suspect that historical processes are constrained to what one might formalize as a local neighborhood in a suitable constructed formal space. For biological evolution, at least, we have some evidence that this is indeed the case[67]. We shall also need to distinguish between changes, i.e., dynamics within a given space, such a redistribution of economic wealth in a stationary system, and systemic changes of the underlying space itself, for instance through technological innovations

that open up new dimensions or change the topology. This will include a systemic theory of innovations.

3 Formal Structures in Computational History

As in biological evolution, the interplay between functional adaptations and structural constraints is important in human history. Some of the concepts developed in theoretical biology thus may also be relevant and useful here[6]. In particular, because of constraints inherited from biological ancestors or cultural traditions, in either case, the resulting structures are highly path dependent. Also, interactions between networks are important, in hierarchical or interdependent ways[68, 69], such as internal regulation and external niche construction for biological species[11], and social, material, and epistemic or knowledge networks for social structures[56]. In the simplest case, such networks depend on pairwise interactions between elements, and they can then be studied with the well developed tools of graph theory. In other cases, larger groups become causal agents, and since not every subgroup of such a group may assume a causal role, hypergraphs constitute the appropriate mathematical models; a theory which is currently under rapid development[15, 70]. When one understands the dynamics of and in networks, one can, for instance, analyze patterns of innovation and their diffusion. We now possess good tools to evaluate the complexity of structures and processes[27], and we may be able to determine complexity thresholds above which systems collapse. When different network types interact, one can also trace causalities, with classical concepts like Granger causality, transfer entropy, and directed information, or with the more recently developed theory of information decomposition[71]. Building upon the theory of Ref. [72], one can also trace the propagation of disturbances in and between networks to gain insight into causal relations[73]. In evolving networks, one can identify growth patterns and determine scaling relations between different quantities[74], and changes in scaling relations may indicate qualitative transitions or accelerations. More generally, the combination of network analysis and dynamical systems theory should provide useful tools for the analysis of historical processes.

Mathematically, dynamical processes "live on" spaces that have an intrinsic structure determined implicitly by the properties of the objects from which they are composed. The structure of the space in turn influences dynamical systems that are built on them. The importance of the underlying spaces is often glossed over in naïve approaches to data analysis, e.g., by converting everything to feature vectors and distances between them. Dimensionality, measured, e.g., as local degrees of freedom, or as intrinsic asymmetries, however, may have profound effects. Well understood examples can be found in molecular evolution. The dynamics of RNA evolution with its punctuated equilibria and constant rate of innovation[75, 76], for instance, is a consequence of the organization of RNAs in a sequence space rather than, say, a low-dimensional grid. The context-dependence of certain evolutionary transitions in the evolution of development is a consequence of asymmetries in accessibility, which are an intrinsic property of phenotype spaces[67, 77]. Similarly, processes of exploring chemical space depends on the structure of chemical space itself: it imposes constraints on what (known) reactions can be employed and what routes can or cannot be taken to actually synthesize a molecule, even if the desired end-result can be anticipated. In each case, innovation is fundamentally constrained by a notion of reachability that depends on the space on which the dynamics operates. In the same manner, we expect the structure of the underlying spaces to constrain the dynamics of historical processes even if they are steered by purposeful actions. The formal structures just identified also make the application of a wide range of computational tools possible. Methods for the automatic analysis of large corpora are currently rapidly developed. We can link more traditional historical data with other data coming from the natural sciences, as has been practised in archaeology for quite some time, and we can also directly use archaeological data, for instance, for large scale quantitative surveys of trading relations. We can quantify the effects of external perturbations, of catastrophes like earth quakes, volcanic eruptions, floods, plagues, and pandemics as well as of long term climatic trends. Artificial neural networks provide novel methods of data analysis, and autoencoders[78] go beyond the traditional Principal Component Analysis (PCA) method, to name but a few new computational tools.

4 Conclusion

These formal structures also represent opportunities for finding universal relations across different times and cultures. This is a central goal of computational history and also defines its specific focus more traditional historical scholarship. There is, of course, value in detailed studies of individual cases. But the same is true for the question of detecting universal patterns of historical change. These might involve statistical relations between certain aggregate quantities, percolation effects for innovations in intertwined networks, systematic time lags between different subsystems, complexity thresholds, etc. The search for generalities and underlying mechanisms has driven science since its inception. Evolutionary biology is one prime example of what can be gained by searching for regularities behind overwhelming diversity. We see computational history as a related project. And whatever models we come up with, they can be checked by data.

Computational history not only enriches the narratives of the past, but it also allows testing historical hypothesis of the sort "what if this had happened", which by allowing the computational expansion of the model system, leads to possible alternative futures. This has been in particular discussed for the case of chemistry[2]. Therein, the golden challenge for computational history of chemistry lies in estimating the future of the discipline. Chemistry shapes and creates the disposition of the world's resources and exponentially provides new substances for the welfare and hazard of our civilisation[79]. Analysing the historical driving forces of the expansion of the chemical space may lead to detecting the suitable conditions for speeding up the exploration of the space[2], with its associated societal benefits as the tailored discovery of the drug-like space[80].

Acknowledgment

We wish to thank the members of working groups and attendees of workshops at the Santa Fe Institute and the Max Planck Institute for Mathematics in the Sciences for stimulating discussions.

References

[1] M. D. Laubichler, J. Maienschein, and J. Renn, Computational history of knowledge: Challenges and opportunities, *Isis*, vol. 110, no. 3, pp. 502–512, 2019.

[2] J. Jost and G. Restrepo, Modelling the evolution of chemical knowledge, in *The Evolution of Chemical Knowledge*, Cham, Switzerland: Springer, 2022, pp. 23–33.

[3] D. H. Wolpert, M. H. Price, S. A. Crabtree, T. A. Kohler, J. Jost, J. Evans, P. F. Stadler, H. Shimao, and M. D. Laubichler, The past as a stochastic process, arXiv preprint arXiv: 2112.05876, 2021.

[4] M. D. Laubichler, J. Maienschein, and J. Renn, Computational perspectives in the history of science: To the memory of Peter Damerow, *Isis*, vol. 104, no. 1, pp. 119–130, 2013.

[5] J. Maienschein, M. Laubichler, and A. Loettgers, How can history of science matter to scientists? *Isis*, vol. 99, no. 2, pp. 341–349, 2008.

[6] J. Jost, *Biologie und Mathematik*. Wiesbaden, Germany: Springer Spektrum, 2019.

[7] G. Restrepo, Computational history of chemistry, *Bulletin for the History of Chemistry*, vol. 47, no. 1, pp. 91–106, 2022.

[8] J. A. Goldstone, Initial conditions, general laws, path dependence, and explanation in historical sociology, *Am. J. Sociol.*, vol. 104, no. 3, pp. 829–845, 1998.

[9] E. Szathmáry, Path dependence and historical contingency in biology, in *Understanding Change*, A. Wimmer and R. Kössler, eds. London, UK: Palgrave Macmillan London, 2006, pp. 140–157.

[10] A. Mesoudi, *Cultural Evolution*. Chicago, IL, USA: University of Chicago Press, 2021.

[11] M. D. Laubichler and J. Renn, Extended evolution: A conceptual framework for integrating regulatory networks and niche construction, *J. Exp. Zool. B Mol. Dev. Evol.*, vol. 324, no. 7, pp. 565–577, 2015.

[12] J. Renn and M. Laubichler, Extended evolution and the history of knowledge, in *Integrated History and Philosophy of Science*, F. Stadler, ed. Cham, Switzerland: Springer International Publishing, 2017, pp. 109–125.

[13] J. Renn, D. Wintergrün, R. Lalli, M. Laubichler, and M. Valleriani, Netzwerke als wissensspeicher, *Die Zukunft der Wissensspeicher: Forschen, Sammeln und Vermitteln im*, vol. 21, pp. 35–79, 2016.

[14] D. Wintergrün, Netzwerkanalysen und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung, PhD dissertation, Friedrich-Alexander-Universitaet Erlangen-Nuernberg, Erlangen, Germany, 2019.

[15] R. Mulas, D. Horak, and J. Jost, Graphs, simplicial complexes and hypergraphs: Spectral theory and topology, in *Higher-Order Systems*, F. Battiston and G. Petri, eds. Cham, Switzerland: Springer International Publishing, 2022, pp. 1–58.

[16] L. Wu, D. Wang, and J. A. Evans, Large teams develop and small teams disrupt science and technology, *Nature*, vol. 566, no. 7744, pp. 378–382, 2019.

[17] F. Furet, Quantitative history, *Daedalus*, vol. 100, no. 1, pp. 151–167, 1971.

[18] R. W. Fogel, The limits of quantitative methods in history, *Am Hist Rev*, vol. 80, no. 2, pp. 329–350, 1975.

[19] P. Hudson and M. Ishizu, *History by Numbers: An Introduction to Quantitative Approaches*. London, UK:

Bloomsbury Publishing, 2016.
[20] B. M. Roehner and T. Syme, *Pattern and Repertoire in History*. Cambridge, MA, USA: Harvard University Press, 2002.
[21] T. A. Kohler, D. Bird, and D. H. Wolpert, Social scale and collective computation: Does information processing limit rate of growth in scale? *Journal of Social Computing*, vol. 3, no. 1, pp. 1–17, 2022.
[22] J. Shin, M. H. Price, D. H. Wolpert, H. Shimao, B. Tracey, and T. A. Kohler, Scale and information-processing thresholds in holocene social evolution, *Nat. Commun.*, vol. 11, no. 1, p. 2394, 2020.
[23] D. C. Krakauer, J. P. Collins, D. Erwin, J. C. Flack, W. Fontana, M. D. Laubichler, S. J. Prohaska, G. B. West, and P. F. Stadler, The challenges and scope of theoretical biology, *J. Theor. Biol.*, vol. 276, no. 1, pp. 269–276, 2011.
[24] N. Retzlaff and P. F. Stadler, Phylogenetics beyond biology, *Theory Biosci.*, vol. 137, no. 2, pp. 133–143, 2018.
[25] O. Pfante, N. Bertschinger, E. Olbrich, N. Ay, and J. Jost, Comparison between different methods of level identification, *Advs. Complex Syst.*, vol. 17, no. 2, p. 1450007, 2014.
[26] Y. Hu and M. J. Buehler, Deep language models for interpretative and predictive materials science, *APL Mach. Learn.*, vol. 1, no. 1, p. 010901, 2023.
[27] N. Ay, E. Olbrich, N. Bertschinger, and J. Jost, A geometric approach to complexity, *Chaos*, vol. 21, no. 3, p. 037103, 2011.
[28] T. Efer, G. Heyer, and J. Jost, Text Mining am Beispiel der Dramen Shakespeares: Welche neuen Erkenntnisse können moderne formale Methoden liefern? in *Shakespeare unter den Deutschen: Vorträge des Symposiums vom 15. bis 17. Mai 2014 in der Akademie der Wissenschaften und der Literatur, Mainz*, ser. Abhandlungen der Geistes- und Sozialwissenschaftlichen Klasse, C. Jansohn, W. Habicht, D. Mehl, and P. Redl, eds. Stuttgart, Germany: Franz Steiner, 2015, pp. 217–230.
[29] D. A. Baum, S. D. Smith, and S. S. S. Donovan, The tree-thinking challenge, *Science*, vol. 310, no. 5750, pp. 979–980, 2005.
[30] D. A. Baum and S. D. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology*. Greenwood Village, CO, USA: Roberts and Company Publishers, 2013.
[31] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge, UK: Cambridge University Press, 2010.
[32] A. L. Panchen, *Classification, Evolution, and the Nature of Biology*. Cambridge, UK: Cambridge University Press, 1992.
[33] S. Q. Ye, ed., *Big Data Analysis for Bioinformatics and Biomedical Discoveries*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2016.
[34] G. S. Ginsburg and H. F. Willard, *Genomic and Precision Medicine: Foundations, Translation and Implementation*. London, UK: Academic Press, 2017.
[35] B. Malinowski, The group and the individual in functional analysis, *Am. J. Sociol.*, vol. 44, no. 6, pp. 938–964, 1939.
[36] C. Lévi-Strauss, *Structural Anthropology*. New York, NY, USA: Basic Books, 2004.
[37] T. A. Appel, *The Cuvier-Geoffroy Debate: French Biology in the Decades before Darwin*. Oxford, UK: Oxford University Press, 1987.
[38] M. J. S. Rudwick, *Georges Cuvier, Fossil Bones, and Geological Catastrophes: New Translations and Interpretations of the Primary Texts*. Chicago, IL, USA: University of Chicago Press, 2008.
[39] A. L. Panchen, Richard Owen and the concept of homology, in *Homology: The Hierarchical Basis of Comparative Biology*, B. K. Hall, ed. Amsterdam, the Netherlands: Elsevier, 1994. pp. 21–62.
[40] A. L. Panchen, Étienne Geoffroy St.-Hilaire: Father of "evo-devo"? *Evol. Dev.*, vol. 3, no. 1, pp. 41–46, 2001.
[41] E. G. Saint-Hilaire, *Principes de Philosophie Zoologique*. Paris, France: Pichon et Didier, 1830.
[42] E. M. De Robertis, The molecular ancestry of segmentation mechanisms, *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 43, pp. 16411–16412, 2008.
[43] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, et al., Database resources of the national center for biotechnology information, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D10–D17, 2021.
[44] E. O. Wiley and B. S. Lieberman, *Phylogenetics: Theory and Practice of Phylogenetic Systematics, 2nd ed*. Hoboken, NJ, USA: Wiley-Blackwell, 2011.
[45] International Institute of Social History, https://iisg.amsterdam/en, 2020.
[46] Clio Infra, https://clio-infra.eu/, 2020.
[47] M. Cannataro, P. H. Guzzo, G. Agapito, C. Zucco, and M. Milano, *Artificial Intelligence in Bioinformatics—From Omics Analysis to Deep Learning and Network Mining*. Amsterdam, the Netherlands: Elsevier, 2022.
[48] J. A. Dunne, R. J. Williams, N. D. Martinez, R. A. Wood, and D. H. Erwin, Compilation and network analyses of Cambrian food webs, *PLoS Biol.*, vol. 6, no. 4, p. e102, 2008.
[49] J. A. Dunne, C. C. Labandeira, and R. J. Williams, Highly resolved early Eocene food webs show development of modern trophic structure after the end-Cretaceous extinction, *Proc. Biol. Sci.*, vol. 281, no. 1782, p. 20133280, 2014.
[50] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, The structure and dynamics of multilayer networks, *Phys. Rep.*, vol. 544, no. 1, pp. 1–122, 2014.
[51] R. Tripodi, M. Warglien, S. L. Sullam, and D. Paci, Tracing antisemitic language through diachronic embedding projections: France 1789-1914, arXiv preprint arXiv: 1906.01440, 2019.
[52] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*. Cambridge, UK: Cambridge University Press, 2003.
[53] D. T. Painter, F. van der Wouden, M. D. Laubichler, and H. Youn, Quantifying simultaneous innovations in evolutionary medicine, *Theory Biosci.*, vol. 139, no. 4, pp. 319–335, 2020.
[54] D. T. Painter, B. C. Daniels, and M. D. Laubichler, Innovations are disproportionately likely in the periphery

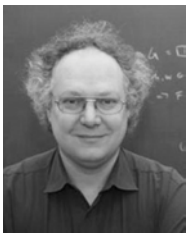
of a scientific network, *Theory Biosci.*, vol. 140, no. 4, pp. 391–399, 2021.
[55] R. Lalli, R. Howey, and D. Wintergrün, The socio-epistemic networks of general relativity, 1925–1970, in *The Renaissance of General Relativity in Context*, A. S. Blum, R. Lalli, and J. Renn, eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 15–84.
[56] J. Renn, *The Evolution of Knowledge*. Princeton, NJ, USA: Princeton University Press, 2020.
[57] J. Jost, *Mathematical Concepts*. Springer, 2015.
[58] E. J. Llanos, W. Leal, D. H. Luu, J. Jost, P. F. Stadler, and G. Restrepo, Exploration of the chemical space and its three historical regimes, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 26, pp. 12660–12665, 2019.
[59] G. Restrepo, Chemical space: Limits, evolution and modelling of an object bigger than our universal library, *Digit. Discov.*, vol. 1, no. 5, pp. 568–585, 2022.
[60] T. Bhattacharya, N. Retzlaff, D. E. Blasi, W. Croft, M. Cysouw, D. Hruschka, I. Maddieson, L. Müller, E. Smith, P. F. Stadler, et al., Studying language evolution in the age of big data, *Journal of Language Evolution*, vol. 3, no. 2, pp. 94–129, 2018.
[61] A. Rocke, What did "theory" mean to nineteenth-century chemists? *Found. Chem.*, vol. 15, no. 2, pp. 145–156, 2013.
[62] J. R. Partington, *A History of Chemistry*. London, UK: Macmillan, 1964.
[63] K. C. Nicolaou, The emergence of the structure of the molecule and the art of its synthesis, *Angew. Chem. Int. Ed. Eng.*, vol. 52, no. 1, pp. 131–146, 2013.
[64] W. Leal, E. J. Llanos, A. Bernal, P. F. Stadler, J. Jost, and G. Restrepo, The expansion of chemical space in 1826 and in the 1840s prompted the convergence to the periodic system, *Proc. Natl. Acad. Sci. USA*, vol. 119, no. 30, p. e2119083119, 2022.
[65] A. M. Bran, P. F. Stadler, J. Jost, and G. Restrepo, The six stages of the convergence of the periodic system to its final structure, *Commun. Chem.*, vol. 6, no. 1, p. 87, 2023.
[66] K. Kollman, J. H. Miller, and S. E. Page, Adaptive parties in spatial elections, *Am. Polit. Sci. Rev.*, vol. 86, no. 4, pp. 929–937, 1992.
[67] B. M. R. Stadler, P. F. Stadler, G. P. Wagner, and W. Fontana, The topology of the possible: Formal spaces underlying patterns of evolutionary change, *J. Theor. Biol.*, vol. 213, no. 2, pp. 241–274, 2001.
[68] J. Jost, Biological information, *Theory Biosci.*, vol. 139, no. 4, pp. 361–370, 2020.
[69] J. Jost, Biology, geometry and information, *Theory Biosci.*, vol. 141, no. 2, pp. 65–71, 2022.
[70] W. Leal, M. Eidi, and J. Jost, Curvature-based analysis of directed hypernetworks, in *Proc. Complex Networks 2019: The 8th International Conference on Complex Networks and Their Applications*, Lisbon, Portugal, 2019, pp. 32–34.
[71] J. T. Lizier, N. Bertschinger, J. Jost, and M. Wibral, Information decomposition of target effects from multi-source interactions: Perspectives on previous. current and future work, *Entropy*, vol. 20, no. 4, p. 307, 2018.
[72] J. Pearl, *Causality*. Cambridge, UK: Cambridge University Press, 2009.
[73] J. Peters, D. Janzing, and B. Scholkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: The MIT Press, 2017.
[74] G. B. West and J. H. Brown, Life's universal scaling laws, *Phys. Today*, vol. 57, no. 9, pp. 36–42, 2004.
[75] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker, From sequences to shapes and back: A case study in RNA secondary structures, *Proc. Biol. Sci.*, vol. 255, no. 1344, pp. 279–284, 1994.
[76] M. A. Huynen, P. F. Stadler, and W. Fontana, Smoothness within ruggedness: The role of neutrality in adaptation, *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 1, pp. 397–401, 1996.
[77] W. Fontana and P. Schuster, Continuity in evolution: On the nature of transitions, *Science*, vol. 280, no. 5368, pp. 1451–1455, 1998.
[78] J. Zhai, S. Zhang, J. Chen, and Q. He, Autoencoder and its various variants, in *Proc. 2018 IEEE Int. Conf. Systems, Man, and Cybernetics* (*SMC*), Miyazaki, Japan, 2018, pp. 415–419.
[79] C. Reinhardt, Introduction, *Isis*, vol. 109, no. 3, pp. 559–564, 2018.
[80] J. Boström, D. G. Brown, R. J. Young, and G. M. Keserü, Expanding the medicinal chemistry synthetic toolbox, *Nat. Rev. Drug Discov.*, vol. 17, no. 12, p. 922, 2018.

Jürgen Jost is a director of the Max Planck Institute for Mathematics in the Sciences, a professor at University of Leipzig, an external member of the Santa Fe Institute, and a principal investigator at ScaDS.AI Dresden/Leipzig. After studying mathematics, physics, economics, and philosophy at Bonn from 1975, he received the PhD degree in mathematics from Bonn University in 1980. He is working in various fields of pure and applied mathematics, theoretical physics, mathematical biology, neuroscience and social science, on a theory of complex systems and network analysis, and in the history of science. He is the (co)author of more than 20 scientific monographs and more than 600 scientific publications. He has received the Leibniz Award of the DFG in 1993 and a European Research Council (ERC) Advanced Grant in 2010.

Manfred D. Laubichler is the global futures professor and president's professor of theoretical biology and history of biology. He is the director of the School of Complex Adaptive Systems and the Global Biosocial Complexity Initiative, Arizona State University, USA. His work focuses on evolutionary novelties from genomes to knowledge systems, the structure of evolutionary theory, and the evolution of knowledge. He is an external professor at the Santa Fe Institute, USA; visiting scholar at the Max Planck Institute for the History of Science, Germany; external faculty member at the Complexity Science Hub Vienna; and vice chair of the Global Climate Forum. He is also an elected fellow of the American Association for the Advancement of Science.

Roberto Lalli is an assistant professor at Politecnico di Torino, Italy. He is a historian of physics who is applying network approaches to the evolution of knowledge in modern physics. Prior to his appointment in Torino, he worked for many years at the Max Planck Institute for the History of Science in Berlin, Germany.

Jürgen Renn is a director at the Max Planck Institute for the History of Science, Berlin, Germany and the Max Planck Institute for Geoanthropology in Jena Germany. He holds an honorary professorship for the history of science at Humboldt University of Berlin and at Free University of Berlin. Since 1998 he has been an adjunct professor for philosophy and physics at Boston University, USA. He is a member of the Academy of Sciences Leopoldina, and the International Academy for the History of Science.



Peter F. Stadler received the PhD degree in chemistry from Univercity of Vienna in 1990 and then worked as an assistant and associate professor for theoretical chemistry at the same school. In 2002 he moved to Leipzig as a full professor for bioinformatics. Since 1994 he has been an external professor at Santa Fe Institute. He has been an external scientific member of the Max Planck Society since 2009 and an corresponding member abroad of the Austrian Academy of Sciences since 2010. His work ranges from the formal structures of evolving systems to the methods development in bioinformatics and cheminformatics as well as applicatioss in molecular medicine, evolutionary, structural, synthetic biology, and cultural evolution.

Dirk Wintergrün is the director of digital transformation at the Klassik Stiftung Weimar. Before that, he worked for many years at the Max Planck Institute for the History of Science, Berlin, where he led many projects in digital and computational history of science.

Eckehard Olbrich is group leader at the Max Planck Institute for Mathematics in the Sciences (MPI MiS), Germany. He received the PhD degree in theoretical physics from TU Dresden, Germany in 1995. From 1995 to 2000, he was a postdoctoral researcher at Max Planck Institute (MPI) for the Physics of Complex Systems. From 2000 to 2004 he worked as a research fellow at University of Zürich mainly on the time series analysis of EEG data. Since 2004 he has been a senior researcher at the MPI MiS. He is working on several aspects of complex systems theory, such as information decomposition, complex networks, game theory, and mathematical modeling of social dynamics and communication with a focus on data analysis.

Guillermo Restrepo received the master degree in chemistry from Universidad Industrial de Santander, Colombia in 2003, and the PhD degree from Universität Bayreuth in 2008. After that, he held the postdoctoral positions at Universität Bayreuth and Texas A&M University. From 2014 to 2017, he was an Alexander von Humboldt foundation fellow at Universität Leipzig. He was a chemistry professor at University of Pamplona, Colombia for a decade and since 2017 he has been a researcher at the Max Planck Institute for Mathematics in the Sciences, Germany. Using large scale chemical and bibliographic databases, he leads projects on the history and evolution of chemistry. He is the 2020 recipient of the Gmelin–Beilstein-Denkmünze of the German Chemical Society.