Visualizing big science projects

Katy Börner, Filipi Nascimento Silva and Staša Milojević

Abstract | The number, size and complexity of 'big science' projects are growing — as are the size, complexity and value of the data sets and software services they produce. In this context, big data gives a new way to analyse, understand, manage and communicate the inner workings of collaborations that often involve thousands of experts, thousands of scholarly publications, hundreds of new instruments and petabytes of data. We compare the evolving geospatial and topical impact of big science projects in physics, astronomy and biomedical sciences. A total of 13,893 publications and 1,139 grants by 21,945 authors cited more than 333,722 times are analysed and visualized to help characterize the distinct phases of big science projects, document increasing internationalization and densification of collaboration networks, and reveal the increase in interdisciplinary impact over time. All data sets and visual analytics workflows are freely available on GitHub in support of future big science studies.

'Big science' today is international, interdisciplinary and inter-institutional. Big science projects are anchored around expensive, large and complex instruments, they can run for several decades and they involve thousands of experts. Big science projects make breakthroughs not only in basic research but also in innovation that impacts economy and solves challenging societal needs. As more science fields move towards the big science model of knowledge creation, the lessons learned from previous successful endeavours become essential. This is because big science projects are not just larger and more expensive than other projects but they require specific organizational and management structures. Different knowledge production processes also bring new research roles, changes in the division of labour and adjustment in formal and informal scholarly communication. One way to communicate these aspects of big science, on which this Perspective focuses, is to use various visualizations. Visualizations in this Perspective — and interactive online ones — show that big science projects go through phases with different input needs, expected outputs and impacts. As big science projects mature, their collaborations densify and internationalize; at the same time, scholarly impact increases in terms of citation counts and interdisciplinary reach.

Big science as a phenomenon can be traced all the way back to fifteenth-century cartography and astronomy[1–3] or to eighteenth-century natural history expeditions[2,4]. Nineteenth-century extensive archival projects (the Corpus Inscriptionum Latinarum and the Carte du Ciel) had many characteristics of present-day big science in terms of funding (state backing by Prussia and France), workforce and timescale (requiring more than a lifetime of effort), and were associated with the initial coinage of the term 'big science' (or, originally, Gorswissenschaft) by classical philologist and Prussian Academy of Sciences member Theodor Mommsen[5]. The better known and more immediate precursors of what became known as big science are the establishment of the University of California cyclotron by Ernest Lawrence in the 1930s for energy research[6] and the World War II Manhattan Project[7]. The term 'big science', however, was introduced in the 1960s by Alvin M. Weinberg[8,9] and Derek J. De Solla Price[1] to describe post-World War II developments in physics that built large and very expensive instruments (reactors and accelerators), accompanied by the growth in scientific team sizes working on nuclear-related research[7]. Making advances in nuclear and, later, particle physics became part of the competition among superpowers, with the expectation that breakthroughs would lead to both scientific and technological superiority[10,11]. In addition, big science has been propelled into the general public's awareness by the founding of the National Aeronautics and Space Administration (NASA) and its active and publicly visible space programme[2]. Although most of the early focus regarding big science was on physics, as early as 1965, Weinberg[12] proposed that biomedical science and biomedical technology were ready to enter the 'big biology' era. This entry was made only in the 1990s with the Human Genome Project (HGP), the first big science project in biology[13]. The expansion of the big science mode of knowledge production to other areas of science, such as big biology, brought with it new organizational and collaborative forms, such as 'networked' science enabled by information and communication technologies[14] and some debates as to whether such coordinated efforts can be called big science[15,16].

Big science accentuated the central role instruments play in the development of science as "engines of discovery"[17]. Historically, instruments such as the telescope, the microscope and the air pump opened new vistas and led to scientific revolution, fundamentally changing the nature of scholarship[18–21]. The quest for increased sensitivity and accuracy of instruments led to their constant evolution, making these ever more expensive tools[19,22] obsolete fairly quickly[19]. This process has been described[23] as 'tinkering', in which 'lineages of technology' are adapted and combined, leading to networks, or 'genealogies' of technologies. However, the power of instruments, such as a scanning tunnelling microscope, can be realized only when they engage a community of researchers in what has been called 'an instrumental community,' eventually leading to the formation of new scientific fields, such as nanotechnology[24]. Furthermore, the relationship between science and technology is complex and interdependent, with science also contributing to technology development[25–27].

Early scientists, such as Galileo Galilei and Isaac Newton, engaged in instrument building as well as theoretical and experimental work[28,29]. While not without precedent, instrument building

is detached from performing research in modern science. But ever larger and more complex instruments are only one component of the larger "technological systems"[2,30] that emerged in the 1940s and 1950s out of war labs as facilities (such as particle accelerators) in which scientists and engineers worked together in new organizational structures and where scientist themselves often engaged in engineering tasks[31–33], often making instrument construction and experiment design inseparable[34]. In high-energy physics (HEP), advances in big instruments are both essential for research advancement and time consuming, and "a HEP experiment looks much more like a technological project than basic science"[15], resulting in new professional roles[10].

The most obvious and least disputed characteristics of big science have been the need for large and diverse human resources — including scientists, engineers and technicians — together with large investments[35] to build, service and run very large and expensive instruments[2,36]. Additional features of big science projects have been identified, such as increase in duration to decade(s)[37], increase in multidisciplinary and international teams[6], as well as "the industrialization of research"[14], as exemplified by "an increasingly differentiated and hierarchical division of labour"[2] and profound changes in "political and organizational forms"[10].

Three types of big science have been distinguished[13]: centralized (examples of which are the Manhattan Project and the Apollo programme), which entail centralized effort to build and operate "a major technological system"; federal (such as the HGP and catalogues of stars and galaxies), which entail decentralized efforts to acquire information or knowledge "concerning big subjects" and to integrate it into databases; and mixed (research carried out in big facilities such as accelerators).

In this Perspective, we demonstrate how publication, funding and other data can be used to analyse, visualize, understand and communicate the evolving dynamics of big science projects. High-quality and high-coverage databases combined with open code make it possible to examine the evolution of author teams, institutional collaborations, impact and reach in a scalable and reproducible manner. We start by introducing six big science projects, two each from physics, astrophysics and biomedical sciences, and the data used for the analyses and visualization. For each of the projects, we survey productivity (measured by publications) and impact (measured by citations) over different project phases and milestones. We then discuss the evolution of collaborative networks — demonstrating growth and internationalization of team sizes. We conclude by surveying the issues regarding the big data that these projects generate, cyberinfrastructure needs and initiatives, and the ways of measuring and communicating success of these projects.

Comparing six big science projects

To exemplify key characteristics of big science projects, we identified six prototypical projects. ATLAS[38] and BaBar[39] are physics projects with 30 years of history. The Laser Interferometer Gravitational-Wave Observatory (LIGO)+Virgo[33,40–43] and the IceCube Neutrino Observatory[44] have opened new windows in astrophysics and are enabling multi-messenger astronomy[45]. In the biomedical sciences, we selected the completed HGP that resulted in the human genome[46–50] and two rather young efforts that aim to map the human body at the single-cell level: the Human Cell Atlas (HCA)[51] and the Human BioMolecular Atlas Program (HuBMAP)[52]. Note that three of these projects generated Nobel-worthy results: BaBar was mentioned in a Nobel award in 2008, ATLAS research won a Nobel in 2013 and LIGO in 2017, providing another indicator of the scientific value and impact of big science projects.

To study and compare these six projects, we compiled a data set comprising 13,893 publications and their 333,722 citations, 1,139 funded grants and the 21,945 experts that authored these publications and received the grants (FIG. 1). Publication data for the physics and astronomy projects were downloaded on 8 January 2021 from INSPIRE, an open access library widely used in the field of HEP. The data set includes data output, documents (such as articles, conference abstracts and reports), journals and disambiguated authors with disambiguated institutions. We employed the internal experiment and collaboration categories of INSPIRE and extracted only the records associated with each of the four physics and astronomy projects. Funding data for the US-based projects were collected from the NSF Award Search.

Biomedical publications were retrieved from different sources depending on the project. For the HGP, we manually extracted all publications from the project landmarks webpage on 16 January 2021. For HuBMAP, we collected all the funding information and related publications from the NIH Research Portfolio Online Reporting Tools (RePORT) on 18 January 2021. For the HCA, publications were collected from the official project website on 16 January 2021. Publications for all three biomedical projects were matched against the 5 January 2021 Microsoft Academic Graph (MAG)[53] data dump using digital object identifiers (DOIs), PubMed IDs, PubMed Central (PMC) IDs, year and titles (depending on the available information), to retrieve citation counts for each publication. The granularity of institutions collected from INSPIRE differs from the MAG data used for the biomedical projects. For instance, all the Max Planck Institutions in MAG are listed as a single affiliation, Max Planck Society, whereas in INSPIRE, each campus is considered as a separate institution. Data details and code are available at https://bigscience.github.io.

Phases of big science projects

Big science projects typically go through the following phases and milestones: the project initiation, which includes understanding the goals, priorities, deadlines and risks of the project; securing funding; project planning, including outlining the tasks and timeline required to execute the project; instrument building, validation and calibration; running of experiments and data gathering; data analysis; dissemination of findings; and the project closure. The road from the original idea to initial funding is likely to take years, as is the building of the instrument.

Understanding the duration, needs, expected outputs and impacts of project phases is important for project funding, management and communication of results. An analysis of 53 physics projects indicated that it takes, on average, 2.25 years from "the initial formulation of the project to the point of funding" and, on average, 3.5 years "after the initial idea before publication"[54]. However, the range of time frames for the phases is wide, owing to the diverse nature of the projects studied. In addition, big science projects are taking longer and longer to complete, from some of the earlier experiments lasting ten or more years to some projects now lasting 20 years or longer[15]. As they move through different phases, projects grow and shift in personnel composition. A large number of physics and astrophysics projects start 'small', with individuals playing important roles, only for those roles to later be played by institutions[10]. When it comes to HEP, time spent on designing and building instruments is much longer than that spent on acquiring data. Additionally, building some of these instruments happens over multiple

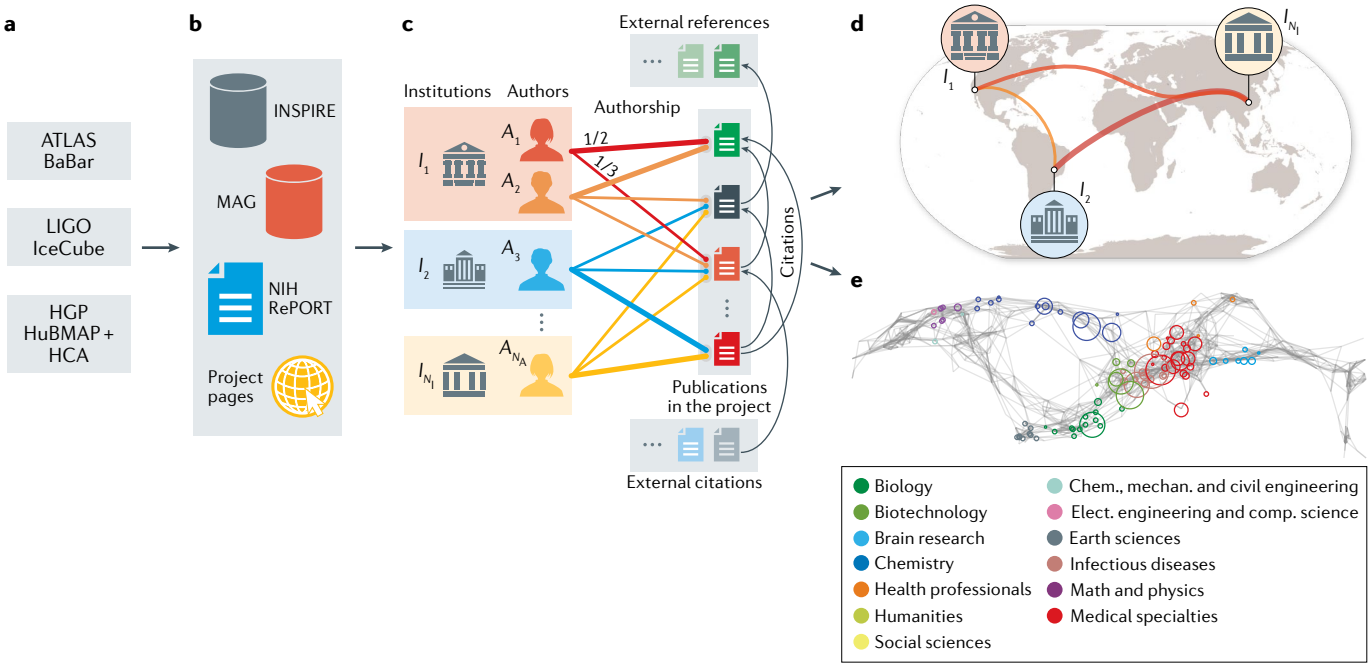


Fig. 1 | **The data compilation, analysis and visualization workflow. a** | Six big science projects are identified. **b** | INSPIRE data were retrieved for physics and astronomy projects based on the INSPIRE collaboration and experiment categorization. The Microsoft Academic Graph (MAG) was used for the biomedical projects, using data from NIH RePORT to obtain Human BioMolecular Atlas Program (HuBMAP) publication metadata. For the Human Genome Project (HGP) and the Human Cell Atlas (HCA), publication information was collected from their respective websites, and records were matched against the MAG to retrieve citation data. **c** | From the publication metadata, information on authors ($A_1, A_2, \ldots, A_{N_A}$, where N_A is the number of authors) and their affiliations ($I_1, I_2, \ldots, I_{N_I}$, where N_I is the number of institutions), publication venue and citation counts were extracted and stored as a multilayer knowledge graph. Publications that cite or are cited by the papers in the six data sets were added to the knowledge graph. **d** | For the physics and astronomy projects, institution address data were used to geolocate authors and to draw the collaboration network of authors aggregated by their respective affiliation. Each edge is weighted by the inverse of the number of co-authors of the publication. **e** | For all six projects, resulting publications were mapped topically based on keywords and overlaid on the 554 subdisciplines of the University of California San Diego (UCSD) classification system; citation linkages to references and from citing papers were overlaid and can be accessed as interactive visualizations on GitHub.

'generations' of projects. For example, ATLAS began as a concrete experimental effort called Future UA2 (REF.[15]), IceCube is the successor of AMANDA and LIGO was preceded by iLIGO.

Using the data set introduced above, we identified the different phases for the six projects (FIG. 2). All physics projects have a construction phase, in which big instruments are developed and constructed. During this phase, there are only a few publications and citations (see, for example, BaBar and LIGO). Not surprisingly, a significant jump in publication counts can be observed during the research phase. Citation counts follow publication trends, including the relative jumps during the research phase. For ATLAS and BaBar, the number of citations increases rapidly during the research phase. Among the considered physics projects, only BaBar is in the closure phase; however, data analysis is still ongoing, resulting in a smaller number of publications per year compared with when the project was active. Of the biomedical projects, HGP publications started in the research phase, and the last paper was published at the end of the project; all papers are still acquiring citations — 20 years after the official completion of the HGP. The HCA project officially started in 2017 and HuBMAP in 2018, and much focus is on technology development and data collection. However, it is important to note that these two rather young projects combined are comparable with the other five projects not only in terms of the number of publications and citations (FIG. 3) but also in terms of team size, discussed below. All projects have several hundred expert authors from hundreds of institutions.

Big teams and their management

Contemporary science has witnessed the growth of collaborative work, both in terms of the increasing prevalence of team-authored papers and the growth of team sizes[55,56]. Nowhere is this as visible as in big science, where author lists have reached thousands — for instance, the paper Combination of the W boson polarization measurements in top quark decays using ATLAS and CMS data at $\sqrt{s} = 8$ TeV has 5,239 authors, and the number of authors on HGP papers can reach hundreds[57,58]. There has also been an increase in the number of institutions per paper[59], especially for big science projects, which are characterized by multi-organizational teams[54]. In HEP, it is institutions and not individuals who enter into these collaborations — which has significant implications for governance and leads to greater equality of members[15]. Big science teams are also international[11,60], interdisciplinary[19,54] and cross-sectoral[54]. In the research areas we examine, members from existing teams may form the core of future collaborations[15], expanding the team based on available budgets and required human resources, and adjusting team expertise as needed to meet the needs of a new big science project. At times, big science collaborations are the only places for performing certain types of science; in physics and astronomy, we now have situations in which most of the researchers in a given scientific field are in one collaboration, effectively creating a monopoly[15].

The evolution of these large-scale collaborations can be visualized via networks overlaid on geospatial maps.

For example, FIG. 3 shows the international collaboration network for the ATLAS project. Each node represents an institution, geolocated using the Nominatim service from OpenStreetMap. We manually checked a sample of 50 records (about 0.15% of the total) and found the procedure to be 96% accurate. Institution disambiguation was performed by geolocation (for example, there are 28 Max Planck institutions in 28 cities in Germany), with the majority of the research centres involved in the projects being located in Europe. Linkages denote collaborations among institutions and the weight of each link equals the number of publications two institutions have co-authored over the project lifespan. To reduce edge clutter, we applied the multiscale backbone extraction method[61], which employs a disparity filter to assign a significance level (in terms of a *P* value) for each of the edges based on the local weight distributions of nodes; edges with a *P* value less than 0.05 were removed. The number of participating countries/territories over time is shown in FIG. 3b. Note the significant jump in the number of participating countries/territories that occurred during the recent upgrade phase. ATLAS started with a large number of participating countries/territories, hence, most collaborations between institutions were created earlier in the project lifespan. By contrast, IceCube (see maps at https://bigscience.github.io) started with less than ten countries/territories but added many more institutions from other countries during the research phase.

Sociologists of science have been captivated by the sort of challenges involved in the management of hundreds of institutions with thousands of participants scattered all over the world and practising different 'cultures' of science[15]. It has been argued that managing large interdisciplinary transnational teams and large budgets requires "industrialization of research"[2]. Multiple factors leading to successful management of such projects have been identified. One of these can be called object-oriented management, or management by content: namely, scientists convene in fluid working groups for a limited time period in order to solve specific problems. These groups report the status of the task regularly, both orally and in status reports. In the biomedical sciences, quarterly demo days and annual all-hands-meetings bring the entire team together and are key for coordinating research and development work by hundreds of experts. Monthly meetings by steering committee members and regular formal feedback by external advisory teams with much participation from industry help to adjust strategy in light of rapid science and technology progress and novel collaboration opportunities. Weekly technical and research meetings bring together lead principal investigators and their extended team to discuss recent developments, get updates by working group leads and disseminate information on publication, funding and training opportunities. Technical experts from diverse, geospatially distributed institutions are involved in the design and provisioning of core data and computing resources. They run weekly stand-up meetings to prioritize data ingest, workflow development and testing, user interface and experience design, and to coordinate regular software releases.

Four major facets are needed for collaborations to be functional: "(1) the extent to which they employ formal rules and documents, (2) their use of specialized division of labor to carry out research, (3) their decision-making hierarchy, and (4) the degree to which a scientific leader sets research directions"[54]. Interestingly, there is more than one way that collaborations handle these requirements. One example of different approaches can be seen in publishing. Big science publication efforts require planning and coordination. For instance, most biomedical projects publish marker papers at the start of the project that lay out project plans and ambitions. Annually, publication packages (sets of interlinked publications that are often published in one issue that report major results achieved that year) are compiled and coordinated via a publication committee. In physics and astronomy, heterogeneous teams[54] with a highly specialized division of labour[2,14] have authorship committees to develop and enforce author protocols that specify types of scholarly communication, their audience and the authorship criteria[62]. However, this approach is not universal. Out of 53 physics projects analysed in REF.[54], only two-thirds used formal authorship contracts. In the area of project leadership, our literature review revealed that bringing scientists with significant experience in big science to leadership positions made a

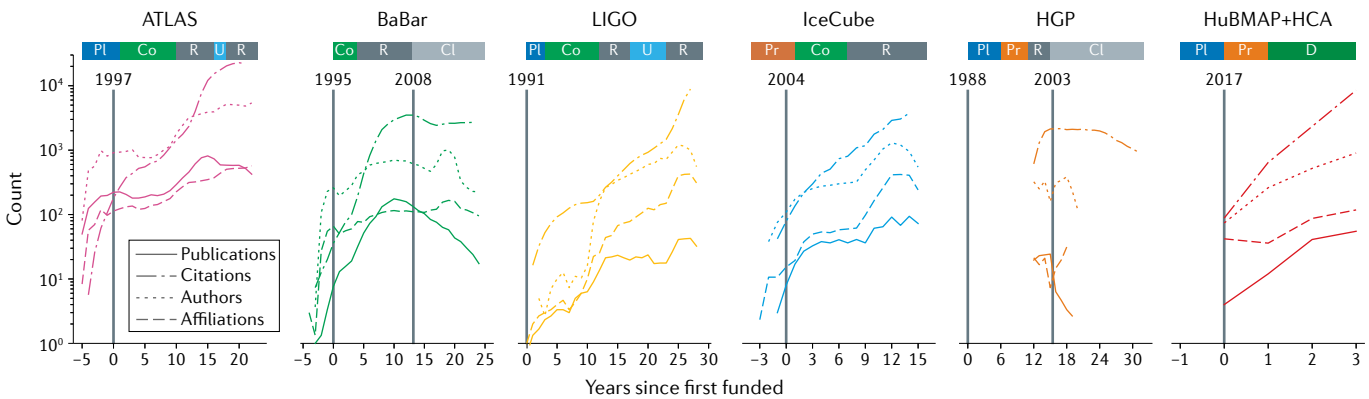


Fig. 2 | **Number of publications, citations, authors and author affiliations for six big science projects.** A 3-year moving average window was applied to improve legibility. To ease comparison across projects, the horizontal axis plots years before (negative values) and after (positive values) a project officially receives funding. Projects go through distinct phases — shown above the graphs — such as planning (Pl), construction (Co), data acquisition (D), prototyping (Pr), research (R), upgrade of instruments (U) and closure (Cl). These phases differ in terms of required inputs (team size and composition, leadership, funding) and output (publications, citations, data) as they mature. Projects start with 10–100 authors but quickly increase in the number of collaborators. Initially, the number of publications is low while big instruments are built or biomedical data generation workflows are developed, validated and productized, but increase substantially during the research phase. Citation counts take time to accumulate but soon reach a factor of 10–300 times the number of publications per year; they continue to amass long after projects close (see BaBar and the Human Genome Project (HGP)). HCA, Human Cell Atlas; HuBMAP, Human BioMolecular Atlas Program; LIGO, Laser Interferometer Gravitational-Wave Observatory.

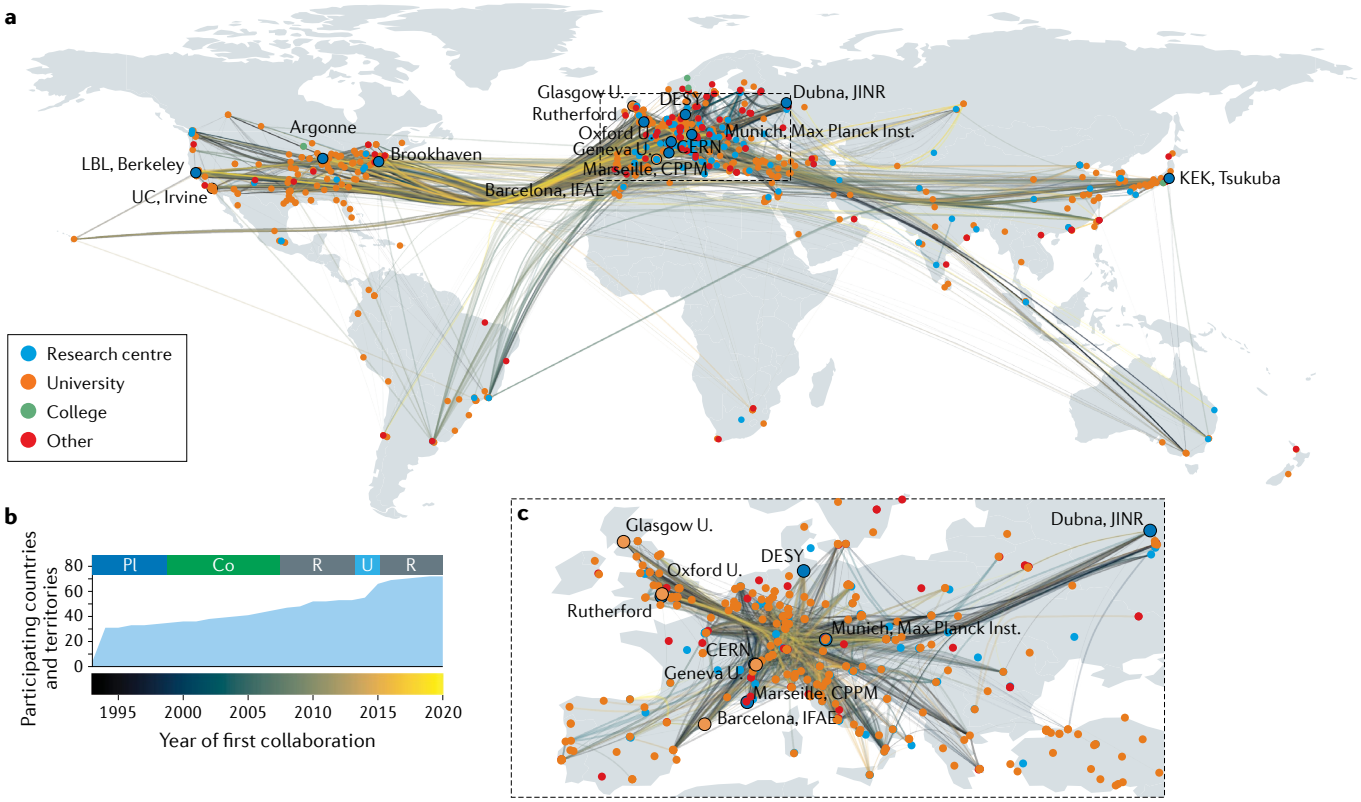


Fig. 3 | **Geodistributed collaboration network of institutions for ATLAS.** Each node represents an institution, with connections indicating the existence of significant collaborations between them (part **a**). Institutions are colour-coded by their type. Edge colour indicates the first time a pair of institutions started collaborating; see colour scale in part **b**, which also plots the total number of countries/territories comprising the participating institutions for each year. As can be seen, the number of collaborations steadily increases over the three decades both in number of centres, universities, colleges and other sites, as well as geographically, engaging scientists from around the globe. A zoomed-in view (part **c**) shows the European institutions. Interactive data visualizations can be explored at https://bigscience.github.io/institutionsmaps. Co, construction; Pl, planning; R, research; U, upgrade of instruments.

substantial difference in transitioning the gravitational wave detection project LIGO from small to big science[63]. Additionally, hiring a project manager with big science experience proved essential for executing the project on time and within budget in the case of IceCube[44].

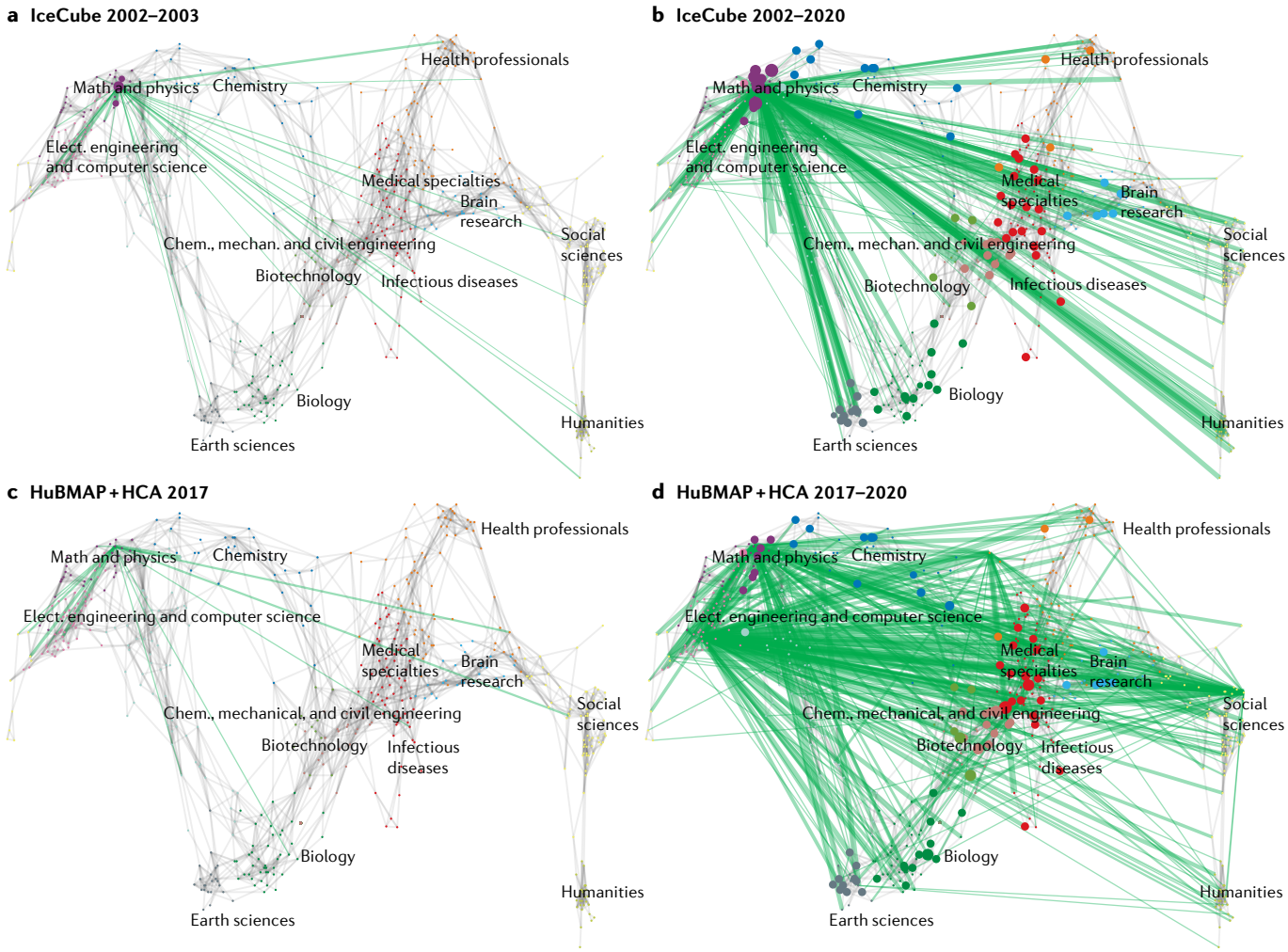
Big data and cyberinfrastructure

Increases in team size are driven by the need for resources that exceed the capabilities of single scientists, labs and even organizations to obtain data and cyberinfrastructures that can move research forward[54]. Data are essential for making "claims of empirical knowledge"[54] and for building "research careers"[54], and can be described as "inputs, outputs, and assets of scholarship"[64]. However, huge data collection efforts are not new. One fascinating example is the Carte du Ciel project: established in 1887, the project mobilized 18 observatories around the world and spent decades compiling a sky map of millions of stars into a catalogue completed only in 1964 (REF.[5]).

Big science particle physics projects from the 1950s and 1960s brought significant changes in "the kind and quantity of data that had to be analysed"[31]; data analysis assumed the central role, leading to ever-increasing demands for computing power[65]. It also brought the need to develop protocols for data acquisition and data sharing[54], as well developing 'pipelines' for the detection of events, such as gravitational waves[42]. However, decisions on who gets access to data and when are not easy. There are many new questions that arise in connection with big data projects that the funders, and policy-makers, need to consider, such as the question of the ownership of the data (does it belong to individual scientists who produced it, the team or the country that funded the project?). Making the data public or open access, though a worthy goal, might open up the possibility of an 'unfair' competition by outside teams[33]. The HGP has been one of the first big projects to have a prominent open access component[64], which is why the project has been described as "an experiment in sharing"[66]. The need for openness goes beyond the data. For example, in recent recommendations regarding multi-messenger astronomy, experts mentioned the need for sharing computer codes and software and learning from communities that have been successful in this[45].

Some of the big science projects we examined are leading to the shifting of traditional roles that experimenters and theorists have played in projects. Data processing often requires specialized knowledge of the instrument; therefore, calibration and software development has been commonly performed by experimentalists themselves, who have the most intimate knowledge of the instruments used. In preparing for LIGO data, however, theorists were heavily involved in data analysis because the project required significant advances in numerical relativity simulations. This need spurred a related collaboration of approximately 50 computational physicists called Simulating eXtreme Spacetimes (SXS)[43].

Many current astrophysics projects require quick detection of potentially interesting events. For example, the promise of multi-messenger astrophysics lies in "joint real-time observation campaigns"[45],


a IceCube 2002–2003
Health professionals
Math and physics
Chemistry
Elect. engineering and computer science
Medical specialties
Brain research
Social sciences
Chem., mechan. and civil engineering
Biotechnology
Infectious diseases
Biology
Humanities
Earth sciences
b IceCube 2002–2020
Health professionals
Math and physics
Chemistry
Elect. engineering and computer science
Medical specialties
Brain research
Social sciences
Chem., mechan. and civil engineering
Biotechnology
Infectious diseases
Biology
Humanities
Earth sciences
c HuBMAP + HCA 2017
Health professionals
Math and physics
Chemistry
Elect. engineering and computer science
Medical specialties
Brain research
Social sciences
Chem., mechanical, and civil engineering
Biotechnology
Infectious diseases
Biology
Humanities
Earth sciences
d HuBMAP + HCA 2017–2020
Health professionals
Math and physics
Chemistry
Elect. engineering and computer science
Medical specialties
Brain research
Social sciences
Chem., mechanical, and civil engineering
Biotechnology
Infectious diseases
Biology
Humanities
Earth sciences
Subfields colours
Biology
Biotechnology
Brain research
Chem., mechan. and civil engineering
Chemistry
Earth sciences
Elect. engineering and comp. science
Health professionals
Humanities
Infectious diseases
Math and physics
Medical specialties
Social sciences
Node and edge size (citations)
1
100
10,000
1
100
10,000
e
ATLAS
Pl Co R U R
1997
BaBar
Co R Cl
1995 2008
LIGO
Pl Co R U R
1991
IceCube
Pr Co R
2004
HGP
Pl Pr R Cl
1988 2003
HuBMAP+HCA
Pl Pr D
2017
Portion of citations
100%
80%
60%
40%
20%
0%
f
Number of subdisciplines
Unique
Effective
Years since first funded

◀ Fig. 4 | **Citations breakdown by disciplines and subdisciplines. a**,**b** | Science maps showing the number of citations for IceCube (about 20 years old), with circle size indicating the number of citations and green links showcasing the number of received citations in the early years of 2002–2003 (part **a**) and for the total time frame 2002–2020 (part **b**). **c**,**d** | Science map for the Human Cell Atlas (HCA) and the Human BioMolecular Atlas Program (HuBMAP) combined (about 3 years old) using the same colour and size coding; 2017 citations (part **c**) and citations for 2017–2020 (part **d**). Interactive data visualizations of reference and citation links to and from these six core data sets can be explored at https://bigscience.github.io/sciencemaps. **e** | Distribution of disciplines citing the projects. **f** | Unique and effective number of subdisciplines citing publications from the six projects over time. Initially, papers from only a few subdisciplines cite the work published by the six projects. Over time, the number of citing subdisciplines increases until reaching a max of 386 (out of 554) for the HuBMAP and the HCA combined. Cl, closure; Co, construction; D, data acquisition; HGP, Human Genome Project; LIGO, Laser Interferometer Gravitational-Wave Observatory; Pl, planning; Pr, prototyping; R, research; U, upgrade of instruments.

which require the application of deep learning techniques to enable real-time alerts necessary to carry those out[67]. Deep learning is also required for more efficient matching and filtering solutions, especially given that new efforts such as the Vera C. Rubin Observatory and Legacy Survey of Space and Time (LSST) expect to have 10 million alerts per night. Also, thanks to their efficiency, the solutions developed for IceCube (and its predecessor AMANDA) to enable on-the-fly data processing with limited resources[44] might also be valuable in regular circumstances.

Today's science in general — and big science in particular — has been described as "distributed, data-intensive, and computation-intensive"[64]. Such science requires significant investments in cyberinfrastructure. The particle physics community has been developing complex, distributed, high-throughput grids, such as the Large Hadron Collider Computing Grid (LCG) with 170 computing centres in 34 countries[68], and the UK's Grid for Particle Physics (GridPP), which connects 19 UK universities and Rutherford Appleton Laboratory[26]. Whereas grid solutions have worked well for the physics community, others have been turning to cloud computing[26]. The fast-evolving area of multi-messenger astrophysics[45] called for the developments in cyberinfrastructure (including supercomputing centres[69], data science and high-performance computing). In the biomedical sciences, funders such as NIH are partnering with commercial providers to support rich data sets and advanced computational services, such as the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. However, with sciences generating data at an unprecedented scale, calls for improved national and international cyberinfrastructure are only expected to grow.

Measuring and communicating success

Publicly funded big science projects are closely scrutinized, especially by "funding agencies and the politicians to whom they answer"[33]. Such scrutiny leads to an increasing need to define and measure success. However, it is not clear "how best to assess the impact of technology and instrumentation"[70], because impact might unfold over decades and secondary impacts might outweigh primary impact[70]. After World War II, investment in big science, primarily in the big science facilities, was seen as a way to enhance national security[2]. Capital investments made big science highly visible, which "elevated political stakes"[10] and made it necessary for scientists to seek public support via the popular press[8]. Post-World-War-II high-energy physicists turned the knowledge and ties they forged into positions that could shape science, both as government advisers and as university administrators (deans, provosts and presidents)[71]. Big science projects require big resources, often over multiple funding cycles, making their leaders tireless fundraisers, who seek support from government agencies, home-universities and, increasingly, philanthropic foundations. However, obtaining funding for big science projects is not easy, especially when funding is sought from the National Science Foundation, where getting funding is seen as a zero-sum game, and where other science specialties are afraid that the big science project will deplete the resources they need[16]. Therefore, convincing the larger scientific community of the benefits of a project also becomes very important.

Compelling and insightful visualizations are powerful tools that can be used to present short-term and long-term impact and to support data-driven decision-making. For example, project timelines help to communicate the chronological order of events, and other data visualizations help to communicate the quantitative impact of research and development (R&D) or the impact of federal funding for fundamental research on industry sectors and society at large; examples include Human Genome Project Timeline[72], Chemical R&D Powers the US Innovation Engine and IT Sectors with Large Economic Impact[73,74]. Publication and funding data make it possible to compute and compare funding intake with publication and citation output, as well as delays between input and output. Maps of science help to reveal the number of publications and citations that individuals, institutions, countries or projects contribute to different areas of science. They can be used to show publication and citation activity over time: for instance, disciplinary impact during the project and interdisciplinary impact after the project ends.

The citation networks for IceCube (FIG. 4a,b) and for the HCA and the HuBMAP combined (FIG. 4c,d) can be visualized using the University of California San Diego (UCSD) map of science and classification system computed using 2006–2008 data from Scopus and 2005–2010 data from Web of Science[75]. The map organizes more than 25,000 journals and conference venues into 554 subdisciplines, which are further aggregated into 13 main scientific disciplines. In order to allocate a number of citing publications to each subdiscipline, a new publication data set is 'science-coded' using the keywords associated with each of the subdisciplines. Using keyword-based matching, 89% of the total of 127,972 publications (including the core and cited publications) can be mapped to at least one of the 554 subdisciplines, reaching a maximum of 386 unique subdisciplines (out of 554) for the HuBMAP and the HCA. IceCube has a very focused impact in the 'Math and physics' discipline, with 1,054 publications over 21 years (2000–2020) and a 89% matching rate. The HCA and the HuBMAP have a much more interdisciplinary impact, with a total of 92 publications published over the initial four years (2017–2020), with a matching rate based on keywords of 94%. Online interactive visualizations make it possible to explore publication counts and citation impact over time and to compare metrics over decades.

The physics and astronomy project citations are dominated by just a few disciplines, such as 'math and physics' and 'chemical, mechanical and civil engineering' (FIG. 4e), in agreement with the science map for IceCube. The engineering discipline seems to play an important role at the beginning of the physics and astronomy projects but 'Math and physics' grows in size during the research phase. By contrast,

biomedical projects are considerably more diverse in terms of their citing disciplines. Initially, the HuBMAP and HCA projects acquire a substantial number of citations from the 'social sciences'. Later, 'electrical engineering and computer science' is more dominant.

The number of unique subdisciplines of the papers that cited the considered projects and the effective number of subdisciplines can also be calculated (FIG. 4f). The effective number of subdisciplines corresponds to the diversity index[76] (also known as true diversity) calculated for the distribution of subdisciplines according to the number of citations in a given year. For each (sub) discipline d, the probability p_d of the project receiving a citation from d is $p_d = f_d/\Sigma_k f_k$, where f_d is the number of times a citation from d is received by a paper in the corresponding project. The entropy of this distribution is $H = -\Sigma_d p_d \log p_d$. The effective number of (sub)disciplines is defined as $N_{\text{eff}} = e^H$, and corresponds to the number of dominating (sub)disciplines in terms of citations. For instance, if a certain project receives thousands of citations from only two (sub)disciplines and just a few dozen citations from others, the effective number of (sub) disciplines will be approximately 2. Although all six projects manage to attract citations from papers in hundreds of subdisciplines, physics and astronomy projects are limited to about 10–20 dominating subdisciplines, whereas biomedical projects receive citations effectively from up to 100 disciplines. These quantitatively confirm previous observations.

Outlook

High-quality and high-coverage databases such as INSPIRE and the MAG combined with open code make it possible to examine different aspects of the big science projects, such as the evolution of author teams, institutional collaborations, impact and reach, in a scalable and reproducible manner. Here, we used big science visualizations to communicate scientific progress to relevant stakeholders — including the general public. At the same time, these visualizations (especially interactive ones) can be used to inform data-driven decision-making by project leaders and funding agencies. They can also provide the context for students and project members, so they understand their place within these complex enterprises. We encourage the readers to also explore the interactive visualizations that we provided at https://bigscience.github.io.

The analysis results discussed here can be further enriched by complementary data, such as interviews and surveys involving project members and leaders. Such surveys could help to understand the challenges and the opportunities associated with planning, managing, evaluating and communicating research collaborations within and across big science projects.

Code availability

Data details and code[77,78] are available at https://bigscience.github.io.

Katy Börner[1] ✉, *Filipi Nascimento Silva*[2] *and Staša Milojević*[1]

[1] *Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA.*

[2] *Indiana University Network Science Institute, Indiana University, Bloomington, IN, USA.*

✉*e-mail: katy@indiana.edu*

https://doi.org/10.1038/s42254-021-00374-7

Published online: 28 September 2021

1. Price, D. J. D. S. *Little Science, Big Science* (Columbia Univ. Press, 1963).
2. Capshew, J. H. & Rader, K. A. Big science: price to the present. *Osiris* **7**, 3–25 (1992).
3. Smith, R. W. in *Big Science: The Growth of Large-Scale Research* (eds Galison, P. & Hevley, B.) 184–211 (Stanford Univ. Press, 1992).
4. Knight, D. M. *The Nature of Science: The History of Science in Western Culture Since 1600* (A. Deutsch, 1976).
5. Daston, L. in *Sciences in the Archives: Pasts, Presents, Futures* (ed. Daston, L.) 159–182 (Univ. Chicago Press, 2017).
6. Galison, P. in *Big Science: The Growth of Large-Scale Research* (eds Galison, P. & Hevley, B.) (Stanford Univ. Press, 1992).
7. Hiltzik, M. *Big Science: Ernest Lawrence and the Invention That Launched the Military-Industrial Complex* (Simon & Schuster, 2016).
8. Weinberg, A. M. Impact of large-scale science on the United States. *Science* **134**, 161–164 (1961).
9. Weinberg, A. M. *Reflections on Big Science* (MIT Press, 1967).
10. Hallonsten, O. *Big Science Transformed: Science, Politics and Organization in Europe and the United States* (Springer, 2016).
11. Wagner, C. S. *The New Invisible College: Science for Development* (Brookings Institution Press, 2009).
12. Weinberg, A. M. Scientific choice and biomedical science. *Minerva* **4**, 3–14 (1965).
13. Kevles, D. & Hood, L. in *The Code of Codes: Scientific and Social Issues in the Human Genome Project* (eds Kevles, D. & Hood, L.) 300–331 (Harvard Univ. Press, 1992).
14. Vermeulen, N. *Supersizing Science: On the Building of Large-Scale Research Projects in Biology* (Maastricht Univ. Press, 2009).
15. Cetina, K. K. *Epistemic Cultures: How the Sciences Make Knowledge* (Harvard Univ. Press, 2009).
16. No authors listed. No final frontier. *Nat. Rev. Phys.* **1**, 231 (2019).
17. Smith, R. W. Engines of discovery: scientific instruments and the history of astronomy and planetary science in the United States in the twentieth century. *J. Hist. Astron.* **28**, 49–77 (1997).
18. Price, D. J. D. Of sealing wax and string. *Nat. Hist.* **93**, 48–56 (1984).
19. Ziman, J. M. *Prometheus Bound* (Cambridge Univ. Press, 1994).
20. Helden, A. V. & Hankins, T. L. Introduction: instruments in the history of science. *Osiris* **9**, 1–6 (1994).
21. Shapin, S. *The Scientific Life: A Moral History of a Late Modern Vocation* (Univ. Chicago Press, 2008).
22. Hoddeson, L. & Kolb, A. W. The Superconducting Super Collider's Frontier Outpost, 1983–1988. *Minerva* **38**, 271–310 (2000).
23. Collins, R. *The Sociology of Philosophies: A Global Theory of Intellectual Change* (Harvard Univ. Press, 1998).
24. Mody, C. C. M. *Instrumental Community: Probe Microscopy and the Path to Nanotechnology* (MIT Press, 2011).
25. Brooks, H. The relationship between science and technology. *Res. Policy* **23**, 477–486 (1994).
26. Meyer, E. T. & Schroeder, R. *Knowledge Machines: Digital Transformations of the Sciences and Humanities* (MIT Press, 2015).
27. Schroeder, R. *Rethinking Science, Technology, and Social Change* (Stanford Univ. Press, 2007).
28. Biagioli, M. *Galileo's Instruments of Credit: Telescopes, Images, Secrecy* (Univ. Chicago Press, 2007).
29. Gleick, J. *Isaac Newton* (Vintage, 2004).
30. Hughes, T. P. in *The Social Construction of Technological Systems* (eds Bijker, W. E., Hughes, T P. & Pinch, T.) 51–82 (MIT Press, 1989).
31. Galison, P. L. *Image & Logic: A Material Culture of Microphysics* (Univ. Chicago Press, 1997).
32. Pickering, A. *Constructing Quarks: A Sociological History of Particle Physics* (Univ. Chicago Press, 1984).
33. Collins, H. *Gravity's Shadow: The Search for Gravitational Waves* (Univ. Chicago Press, 2004).
34. Smith, R. W. & Tatarewicz, J. N. Counting on invention: devices and black boxes in very big science. *Osiris* **9**, 101–123 (1994).
35. Sklair, L. *Organized Knowledge: A Sociological View of Science and Technology* (Hart-Davis MacGibbon, 1973).
36. Galison, P. & Hevley, B. *Big science: The Growth of Large-Scale Research* (Stanford Univ. Press, 1992).
37. Lambright, W. H. Downsizing big science: Strategic choices. *Public Adm. Rev.* **58**, 259–268 (1998).
38. The ATLAS Collaboration. *ATLAS: A 25-Year Insider Story of the LHC Experiment* (World Scientific, 2019).
39. Quinn, H. R. & Harrison, P. F. *The BaBar Physics Book: Physics at an Asymmetric B Factor* (SLAC, 1998).
40. Barish, B. C. in *Einstein Was Right: The Science and History of Gravitational Waves* (ed. Buchwald, J. Z.) 6–18 (Princeton Univ. Press, 2020).
41. Collins, H. *Gravity's Ghost: Scientific Discovery in the 21st Century* (Univ. Chicago Press, 2010).
42. Collins, H. *Gravity's Kiss: The Detection of Gravitational Waves* (MIT Press, 2017).
43. Thorne, K. S. in *Einstein Was Right: The Science and History of Gravitational Waves* (ed. Buchwald, J. Z.) 19–46 (Princeton Univ. Press, 2020).
44. Bowen, M. *The Telescope in the Ice: Inventing a New Astronomy at the South Pole* Vol. 212 (St. Martin's Press, 2017).
45. Huerta, E. A. et al. Enabling real-time multi-messenger astrophysics discoveries with deep learning. *Nat. Rev. Phys.* **1**, 600–608 (2019).
46. Bodmer, W. & McKie, R. *The Book of Man: The Human Genome Project and the Quest to Discover Our Genetic Heritage* (Oxford Univ. Press, 1997).
47. Kevles, D. & Hood, L. *The Code of Codes* (Harvard Univ. Press, 1992).
48. Hilgartner, S. in *Handbook of Science and Technology Studies* (eds Jasanoff, S., Markle, G. E., Petersen, J. C. & Pinch, T.) 302–315 (SAGE Publications, 1995).
49. Watson, J. D. The Human Genome Project: Past, present, and future. *Science* **248**, 44–49 (1990).
50. Gates, A. J., Gysi, D. M., Kellis, M. & Barabási, A. L. A wealth of discovery built on the human genome project — by the numbers. *Nature* **590**, 212–215 (2021).
51. Rosen-Rozenblatt, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
52. Snyder, M. et al. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
53. Sinha, A. et al. in *Proceedings of the 24th International Conference on World Wide Web* 243–246 (ACM, 2015).
54. Shrum, W., Genuth, J. & Chompalov, I. *Structures of Scientific Collaboration* (MIT Press, 2007).
55. Milojević, S. Principles of scientific research team formation and evolution. *Proc. Natl Acad. Sci. USA* **111**, 3984–3989 (2014).
56. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
57. Dennis, C., Gallagher, R. & Campbell, P. The human genome. *Nature* **409**, 814–816 (2001).
58. Jasny, B. R. & Kennedy, D. (eds) The human genome. *Science* **291**, 1177–1180 (2001).
59. Jones, B. F., Wuchty, S. & Uzzi, B. Multi-university research teams: shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
60. Wagner, C. S. *The Collaborative Era in Science: Governing the Network* (Palgrave Macmillan, 2018).
61. Serrano, M. Á., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl Acad. Sci. USA* **106**, 6483–6488 (2009).

62. Galison, P. in *Scientific Authorship: Credit and Intellectual Property in Science* (eds Biagioli, M. & Galison, P.) 325–355 (Routledge, 2003).
63. Collins, H. in *Einstein Was Right: The Science and History of Gravitational Waves* (ed. Buchwald, J. Z.) 111–128 (Princeton Univ. Press, 2020).
64. Borgman, C. L. *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, 2015).
65. Borgman, C. L. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (MIT Press, 2007).
66. Roberts, L. Genome project: an experiment in sharing. *Science* **248**, 953 (1990).
67. No authors listed. The big three. *Nat. Rev. Phys.* **1**, 579 (2019).
68. Zheng, Y., Venters, W. & Cornford, T. Collective agility, paradox and organizational improvisation: the development of a particle physics grid. *Inf. Syst.* **21**, 303–333 (2010).
69. National Academies of Sciences, Engineering and Medicine. *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017–2020* (National Academies, 2016).
70. Kurczynski, P. & Milojević, S. Enabling discoveries: a review of 30 years of advanced technologies and instrumentation at the National Science Foundation. *J. Astron. Telesc. Instum. Syst.* **6**, 030901 (2020).
71. Traweek, S. *Beamtimes and Lifetimes: The World of High Energy Physicists* (Harvard Univ. Press, 1988).
72. Leja, D. *Human Genome Project Timeline* (Department of Energy, 2003).
73. National Academies of Sciences, Engineering and Medicine. *Continuing Innovation in Information Technology: Workshop Report* (National Academies, 2016).
74. National Research Council. *Innovation in Information Technology* (National Academies, 2003).
75. Börner, K. et al. Design and update of a classification system: the UCSD map of science. *PLoS ONE* **7**, e39464 (2012).
76. Chao, A., Chu, C.-H. & Jost, L. Phylogenetic diversity measures and their decomposition: a framework based on Hill numbers. *Biodivers. Conserv. Phylogenet. Syst.* **14**, 141–172 (2016).
77. Börner, K., Silva, F. N. & Milojević, S. Visualizing big science projects — Institution collaboration maps. *zenodo* https://doi.org/10.5281/zenodo.4835034 (2021).
78. Herr, B. W. II et al. Visualizing big science projects — Science maps. *zenodo* https://doi.org/10.5281/zenodo.4884741 (2021).

Acknowledgements
The authors thank the interviewed experts for their time and expert input. T. Schwander gave guidance for compiling the INSPIRE data sets. B. W. Herr II implemented the interactive science maps. T. N. Theriault compiled references and provided professional copy-editing support. This work is funded by the NSF under grants NRT-1735095, AISL-1713567 and DMS-1839167 and the Precision Health Initiative as part of Indiana University's Grand Challenges programme. In addition, this material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0391. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Author contributions
S.M. led the literature review, K.B. led the expert survey and science mapping effort and F.N.S. led the data analysis and visualization. All authors contributed equally to the write-up of other article parts.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Physics thanks Junming Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

RELATED LINKS
Council for Chemical Research. Chemical R&D Powers the US Innovation Engine: https://scimaps.org/map/5/6
Human Cell Atlas. Publications: https://www.humancellatlas.org/publications
Human Genome Project Information Archive, 1990–2003: landmark HGP papers: https://web.ornl.gov/sci/techresources/Human_Genome/project/journals.shtml
INSPIRE: https://inspirehep.net
National Institutes of Health. NIH Research Portfolio Online Reporting Tools (RePORT): https://reporter.nih.gov/
National Science Foundation. BaBar award search results: https://www.nsf.gov/awardsearch/simpleSearchResult?queryText=babar
National Science Foundation. IceCube award search results: https://www.nsf.gov/awardsearch/simpleSearchResult?queryText=icecube
National Science Foundation. LIGO award search results: https://www.nsf.gov/awardsearch/simpleSearchResult?queryText=ligo
National Institutes of Health. NIH RePORTER search results (I): https://reporter.nih.gov/search/bC3_awAf4U6Hl7zF9rQEZQ/projects?shared=true
National Institutes of Health. NIH RePORTER search results (II): https://reporter.nih.gov/search/BwnasVXfbUiGwaac353HGw/projects?shared=true
NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative: https://datascience.nih.gov/strides
Nominatim. Home page: https://nominatim.org
Vera C. Rubin Observatory. Rubin Observatory System & LSST survey key numbers: https://www.lsst.org/scientists/keynumbers