Scientific Data Security Concerns and Practices

A survey of the community by the Trustworthy Data Working Group

https://trustedci.org/2020-trustworthy-data

December 15, 2020

Distribution: Public

Andrew Adams, Kay Avila, Jim Basney, Melissa Cragin, Jeannette Dopheide, Terry Fleury, Florence Hudson, Jenna Kim, W. John MacMullen, Gary Motz, Sean Peisert, Mats Rynge, Scott Sakai, Sandra Thompson, Karan Vahi, Wendy Whitcup, John Zage

About the Trustworthy Data Working Group

The Trustworthy Data Working Group is a collaborative effort of Trusted CI[1], the four NSF Big Data Innovation Hubs[2], the NSF CI CoE Pilot[3], the Ostrom Workshop on Data Management and Information Governance[4], the NSF Engagement and Performance Operations Center[5] (EPOC), the Indiana Geological and Water Survey[6], the Open Storage Network[7], and other interested

[1] https://www.trustedci.org
[2] https://www.bigdatahubs.org
[3] https://cicoe-pilot.org
[4] https://ostromworkshop.indiana.edu/research/data-management
[5] https://epoc.global
[6] https://igws.indiana.edu
[7] https://www.openstoragenetwork.org

community members. The goal of the working group is to understand scientific data security concerns and provide guidance on ensuring the trustworthiness of data.

Acknowledgments

The co-authors thank all the other members of the Trustworthy Data Working Group for their help with drafting the survey, advertising the survey, and analyzing the results, including Galen Collier, Douglas Ertz, Ezra Van Everbroeck, Matias Kind, Meredith Lee, Jim Leous, Santiago Nunez-Corrales, and Angie Raymond.

Trusted CI is supported by the National Science Foundation under Grant 1920430. The Cyberinfrastructure Center of Excellence (CI CoE) Pilot project is supported by the National Science Foundation under Grant 1842042. The Engagement and Performance Operations Center (EPOC) is supported by the National Science Foundation under Grant 1826994. The Open Storage Network is supported by Schmidt Futures[8] and the National Science Foundation under Grants 1747483, 1747490, 1747493, 1747507, and 1747552. The NSF Big Data Innovation Hubs are supported in part by the National Science Foundation under awards 1916613, 1916585, 1916589, and 1916573. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Using & Citing this Work

Cite this work using the following information:
Scientific Data Security Concerns and Practices: A survey of the community by the Trustworthy Data Working Group. December 2020. https://doi.org/10.5281/zenodo.3906865

This work is available on the web at the following URL:
https://trustedci.org/2020-trustworthy-data

[8] https://schmidtfutures.com

Table of Contents

Executive Summary

In April and May of 2020, the Trustworthy Data Working Group conducted a survey of scientific data security concerns and practices in the scientific community. This report provides a summary of the survey methodology and an analysis of the results.

111 participants completed the survey from a wide range of positions and roles within their organizations and projects, respectively. The survey questions were designed and developed by the working group with several goals in mind.

- Identify the concerns of researchers and operators with respect to scientific data.
- Understand how the roles of individuals influence their definition of secure and trustworthy data.
- Itemize the attributes that are most relevant when discussing trustworthiness of scientific data.
- Unify disparate terminology used by respondents and identify assumptions made for users from diverse environments.
- Categorize tools and processes currently employed to maintain trustworthiness of scientific data.
- Discern shortcomings, if any, in these tools and processes when used to ensure trustworthy data.

The working group analyzed the survey results with an eye for patterns, themes, correlations, and aggregates. From this analysis, several key findings emerged:

1. Data owners/maintainers/users are concerned with the impact trustworthy data has on the scientific process, especially with regard to the loss of reputation.
2. Data owners/maintainers welcome help in securing trustworthy data workflows with encryption, provenance, and regulatory compliance (e.g., FERPA, HIPAA, FISMA).
3. Trustworthiness is not well-defined.

A more thorough explanation of the survey's methodology and goals can be found in section 1. Sections 2 and 3 contain a more detailed discussion on the analysis of the survey and the participants' responses. Finally, section 4 outlines a plan to derive best practices using the key findings identified during the analysis phase.

This December 2020 (final) version of the report contains minor clarifications regarding the responses to questions #1 and #5. Otherwise, the report is unchanged since the June 2020 (initial) version.

1 Survey Goals and Methodology

The Trustworthy Data Working Group convened for the first time in February 2020 to outline goals for the survey and brainstorm potential questions. Membership consisted of individuals from Trusted CI, the four NSF Big Data Innovation Hubs, the NSF CI CoE Pilot, the Ostrom Workshop on Data Management and Information Governance, the NSF Engagement and Performance Operations Center, the Indiana Geological and Water Survey, the Open Storage Network, and other interested community members. Additional working group members are welcome.

To understand how the scientific community views trustworthy data and investigate current guidance and thought in this area, members of the working group first completed a review of published literature exploring ideas of trustworthiness in scientific data and results. Additionally, existing sources of NIST guidance (e.g., NIST 1800-25[9] and 1800-26[10]) were reviewed. These revealed a wide variety of how the community defines "trustworthiness," although "data integrity" emerged as the most common definition. While definitions of this term vary as well, most generally it is used within the scientific community to mean the data has not been altered. This was particularly true for literature originating from security and computer science, and from operators of high-performance computing infrastructure or cyberinfrastructure. The literature review also uncovered instances of failures in trustworthiness in computational and data science, including an inconsistency detected in popular Python scripts used to analyze nuclear magnetic resonance [1] and previously undiscovered errors in scientific workflows discovered by test runs of the Pegasus Scientific Workflow Integrity tool [2].

After this background research, the group created a list of potential survey questions, then narrowed these down to the final form listed in Appendix A. From a high level, the survey was designed to investigate how scientists and researchers define and think about the trustworthiness of data, as well as to potentially understand data security requirements for different science domains, discover which phases of the scientific process are most concerning from a trustworthiness standpoint, and answer whether current guidance around trustworthiness is sufficient. The questions used a variety of formats to solicit input, namely multiple-choice, Likert scale (agree/disagree), short answer, and long answer forms. All multiple-choice questions allowed respondents to select multiple answers. The survey deliberately avoided defining trustworthiness for the respondent and instead asked which

[9] https://www.nccoe.nist.gov/projects/building-blocks/data-integrity/identify-protect
[10] https://www.nccoe.nist.gov/projects/building-blocks/data-integrity/detect-respond

attributes were considered most important for the concept (*e.g.*, integrity, accuracy, or provenance, among other options).

After approval by the University of Illinois Institutional Review Board,[11] the group made the survey available for public response on April 21, 2020. The group solicited participants using a variety of public forums including the Trusted CI blog,[12] Announce mailing list,[13] and Twitter feed,[14] and the Cyberinfrastructure Center of Excellence Pilot Announcement List,[15] as well as a number of private, community-focused mailing lists including the four Regional Big Data Innovation Hubs and their All-Hubs Data Sharing and CyberInfrastructure Working Group, IEEE-Standards Association Working Group P2733/P2933 for Clinical Internet of Things (IoT) Data and Device Interoperability with TIPPSS (Trust, Identity, Privacy, Protection, Safety, Security), Coalition for Academic Scientific Computation (CASC), Campus Research Computing Consortium (CARC), Campus Champions, Earth Science Information Partners (ESIP), and the National Ecological Observatory Network (NEON). The survey closed on May 31, 2020. During the 40 days the survey was open to the public, 111 responses were received.

The initial version of the survey inadvertently omitted question five, a multiple choice question that asks *Which attributes do you believe scientific data must have in order to be trustworthy?* This was discovered after 40 responses had already been received. Of these, 31 both indicated that they were open to follow up questions and provided their email address, so a link to a second survey including just question five was emailed to these individuals. About half of the participants responded to this follow-up query, and their answers are included in the results and analyses below.

The working group then analyzed the received responses. This effort consisted of a high-level analysis of themes that appeared in the open text responses, which will be covered in section 2, and a per-question analysis, covered in section 3.

[11] University of Illinois IRB Review #20777
[12] https://blog.trustedci.org/2020/04/trustworthy-data-survey.html
[13] https://www.trustedci.org/trustedci-email-lists
[14] https://twitter.com/trustedci
[15] https://cicoe-pilot.org/mailinglists/

2 Working Group Keyword Analysis

We reviewed the open-text responses and recorded common words or themes in a spreadsheet. As we found new themes, we added them to the table and reviewed the responses again to find any themes that may have been missed on a previous review. We tallied the results and listed them in Table 1.

The four questions in the survey with open-text responses (Questions 7, 9, 16, and 17) are quoted below:

7. Please explain why protecting the trustworthiness of scientific data is or is not important to you. Does its importance change during different phases of the research cycle (e.g., data collection, calibration, processing, analysis, sharing, and publication)?
9. In your work, what are potential consequences (if any) to using/producing/curating scientific data that is not trustworthy?
16. If you answered Yes or Maybe [that you would want additional tools or technologies to help maintain trustworthiness], please explain. For example, are there specific tools or technologies you would like to use or specific needs you would like guidance addressing?
17. Is there anything else related to trustworthy research data that you would like us to know?

Table 1. Keywords and Themes identified in Questions 7, 9, 16, and 17.

Keywords and Themes	Total Respondents
Impact on scientific results - "bad conclusions," "wrong conclusions"	44
Reputational risk - reputational harm to scientist or institution	27
Integrity of scientific process - the scientific method	18
Trust in science - combating "misinformation," "politics," "loss of public trust"	15
Provenance of data	15
Loss of funding	15
Concerns about storage, data integrity management, and/or data quality mgmt.	14
Seeking guidance, training, and/or audits	13
Reusability	11
Reproducibility	10
Retraction of publication	6
Encryption	4
Compliance - managing sensitive data, controlled unclassified information (CUI)	4

Difficult to work with restrictions - too cumbersome to follow	3
Wasted funds and/or resources	3
Threats from insiders - theft and/or surveillance	1

Some observations can be drawn from this type of analysis, both from what is stated and what is not. Respondents overwhelmingly stressed the impact of trustworthy data on the scientific results, citing "bad conclusions" or "wrong conclusions" as an effect of untrustworthy data. Reputational harm is also seen as a negative consequence of using untrustworthy data but cited about half as often. Few respondents stated that current security policies are difficult to work with, indicating that there might be a level of comfort or acceptance of the policies in place. Also, while one respondent mentioned that “probably one of the biggest threats are from insiders,” there were no comments about external threats.

3 Survey Data and Discussion

In this section, we provide a question-by-question analysis of survey responses. Questions 1-4 aim to identify which job titles, roles, fields of science, and sectors are represented by survey participants. These questions help to group participants and identify any particular leaning a certain group has, as well as trends in how participants for a group or subgroup answer questions about trustworthy data. Questions 5-12 ask about the respondent's opinions and experiences regarding the trustworthiness of research data. In particular, Question 5 provided defined attributes characterizing trustworthy data for participants to agree or disagree on. Questions 13-16 ask about tools and technologies used for securing research data. Question 17 asks for any other information and Questions 18-20 ask about follow-up communication with the respondents. Please refer to Appendix A for the complete list of survey questions.

Question #1: Primary Job Title

What is your primary job title?

We asked this question to better understand the perspectives of survey respondents related to trustworthy data. Respondents answered this question via a free-text field.

There were 108 responses; 3 individuals skipped the question. The following table lists keywords in the answers:

Table 2. Keywords identified in answers to Question #1.

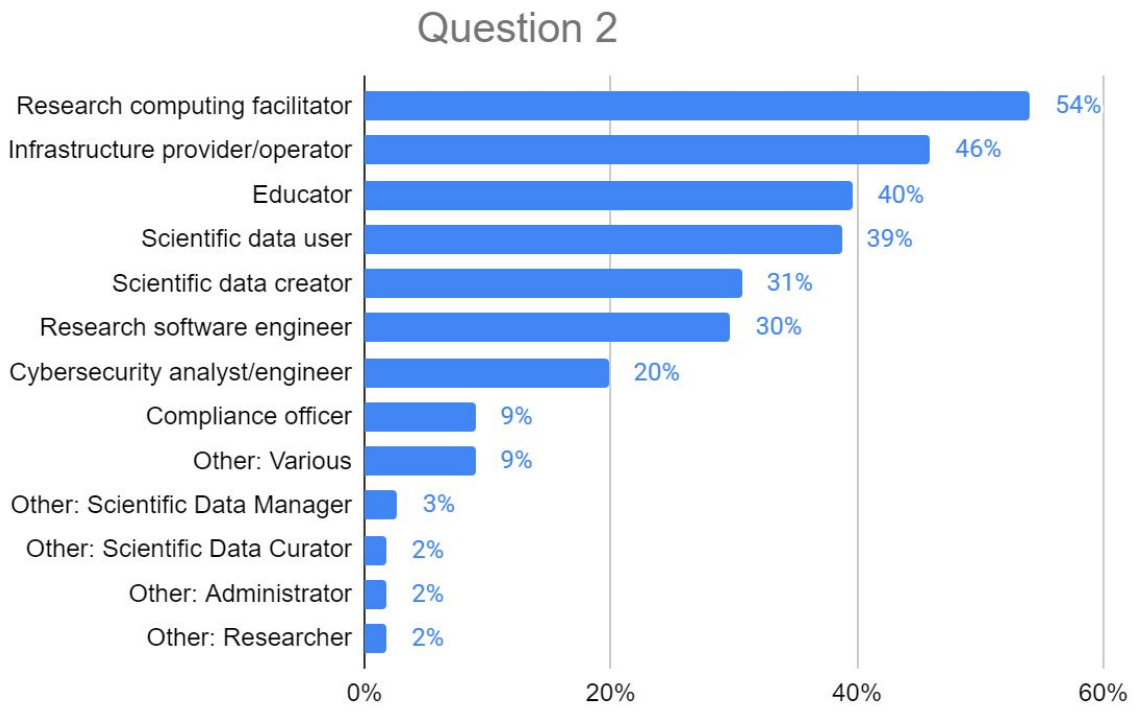
Keywords in job titles	Responses
research	35
director	19
scientist	14
professor	14
manager	13
senior	9
data	7
cyberinfrastructure	4
security / cybersecurity	4
consultant	3
systems administrator	2
chief information officer	2

Other responses included architect, analyst, branch chief, coordinator, curator, engineer, and student.

Question #2: Job Roles

Please select all roles that describe your work, even if they do not correspond to your job title.

A job title might not indicate what tasks one performs, so we asked a follow-up question to discern if, for example, a participant who reported their job title as a professor was also a scientific data user.



There were 110 responses; 1 individual skipped the question. As shown in the graph above, the majority of respondents are in IT support roles (research computing facilitator and/or infrastructure provider/operator), with cybersecurity professionals (analyst/engineer) also represented at 20%. Scientific data users and creators are in the minority at 39% and 31% respectively, further confirming that most respondents are in a support role with respect to scientific data.

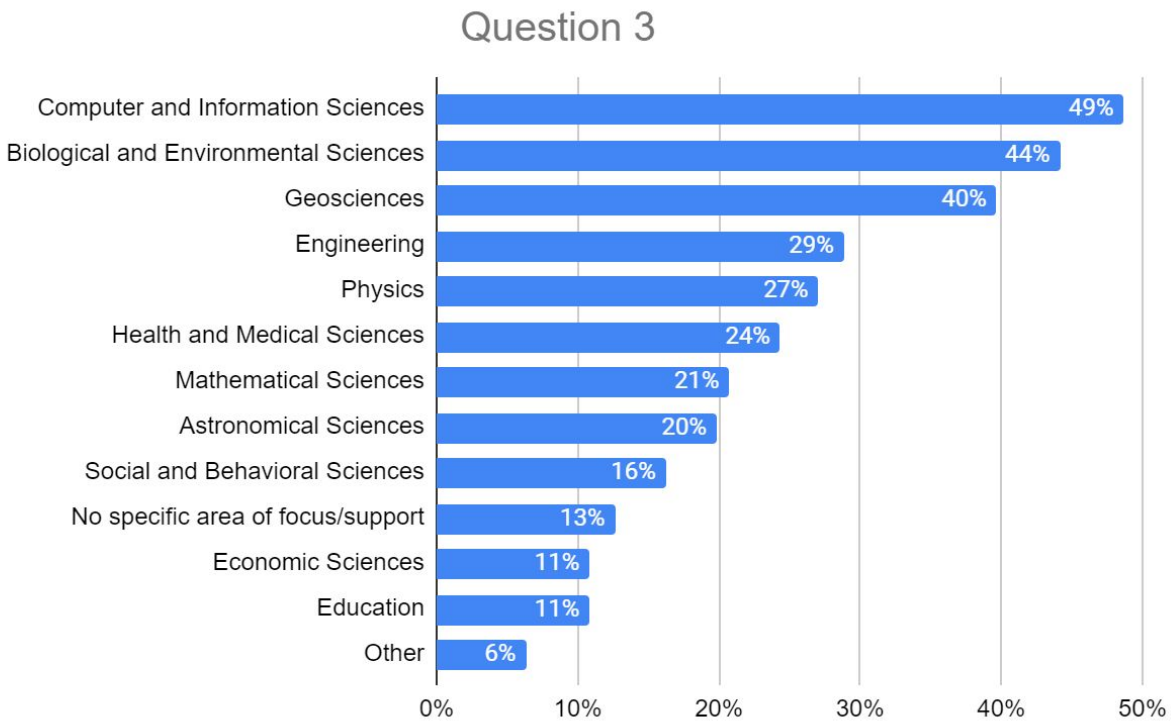
Note that all our multiple-choice questions, including this one, allowed respondents to select multiple answers, so percentages in the above graph sum to more than 100%. For this question, 90 respondents (81%) selected multiple responses (e.g., research computing facilitator and compliance officer).

Question #3: Fields of Science

In the above roles, what field(s) of science do you primarily work in or support?

With the expectation that different fields of science have different perspectives on trustworthy data, we asked about the fields represented by the respondents to better understand the population of responses we received.

There were 110 responses; 1 individual skipped the question. Computer and Information Sciences are well represented, primarily because many of the respondents are in IT roles, as seen in responses to the previous question. Open ended "Other" answers include Humanities, Geospatial and GIS, Molecular sciences, Socio-Environmental Systems, and Fusion experiments.



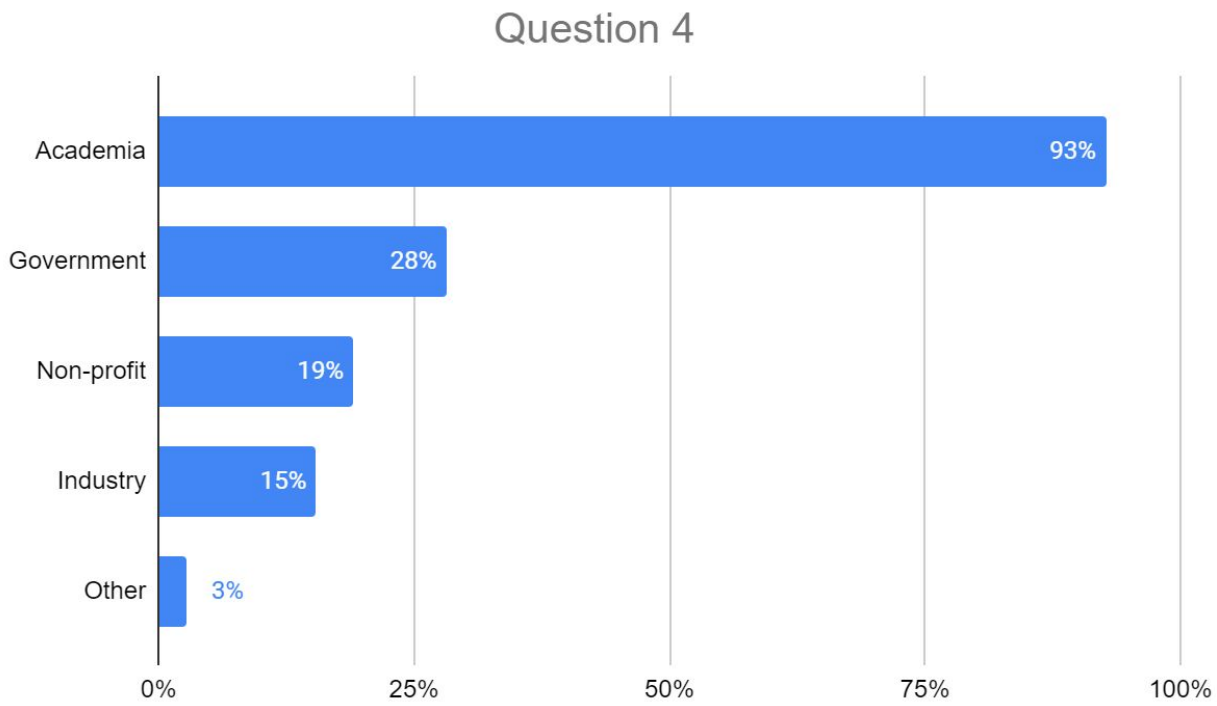
Sixty-three percent of participants reported working in or supporting more than one field of science, which is consistent with research computing facilitators and infrastructure providers/operators who support many fields of science.

Question #4: Work Sectors

What sector(s) do you work with?

By identifying the sectors or industries of participants and then reviewing combined answers, we can better understand the population of survey respondents.

All 111 participants responded, with the majority of them working in the academia sector. Open-ended "Other" responses include NGO, a "not-for-profit federally funded research & development center," and Education.



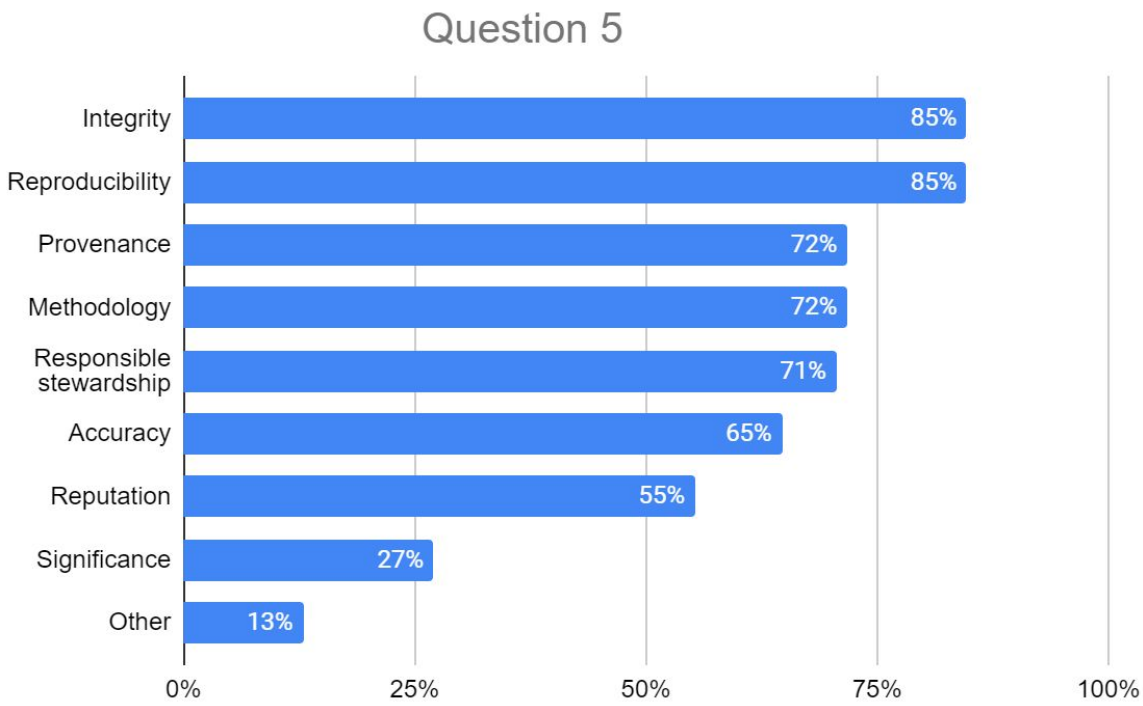
Since the working group members are primarily academics, it is unsurprising that we obtained responses primarily from academics. Additionally, many respondents reported working in multiple sectors. All individuals that selected the choice “Other” selected it in addition to at least one of the provided sectors.

Question #5: Attributes of Scientific Data

Which attributes do you believe scientific data must have in order to be trustworthy?

We asked this question to identify keywords and definitions that participants associate with trustworthiness in the data sciences. See Appendix A for the definitions that we provided for each keyword. The concept of trustworthiness can be complex, and different roles as well as different fields may have different views of what trustworthiness means.

As noted earlier, the first 40 individuals who participated in the survey received surveys where this question was accidentally omitted. A follow up email with a targeted survey link was sent to these 40 participants, resulting in 15 more responses, for a total of 86 responses. Open-ended "Other" responses include transparency, documentation, and methodology.



The choice of what attributes make up trustworthiness by each participant was an open selection, rather than a ranking. This means even though certain attributes were selected more often than others, such as integrity over reputation, it does not necessarily mean participants thought it was more important, only that there was a greater consensus it was an important attribute. Again, note that each of these options included a short definition sentence in the survey questionnaire, which can be found in Appendix A.

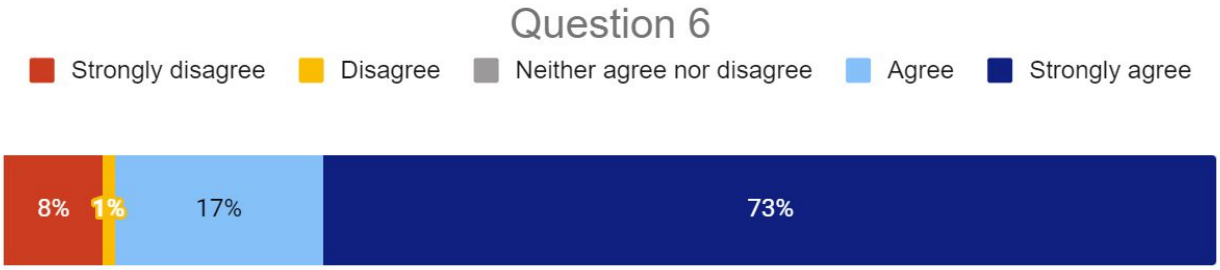
A respondent's field(s) of science may help explain some of the answers from this question, such as accuracy being chosen less frequently than integrity, though we have not (yet) confirmed such correlations in our analysis. We might expect participants focusing on computational approaches to select reproducibility more often, versus observational approaches like in Earth and Life sciences, where reproducibility is more difficult. It should be noted that respondents used the "other" option *in addition to* selecting the available answers and not to the exclusion of the answers. A potential topic of further study could be to understand how participants would rank importance of these attributes.

Question #6: Importance of Protecting Trustworthiness

Do you agree with the following statement? I think that protecting the trustworthiness of scientific data is important.

We asked this question to get an understanding of the participants' valuing of trustworthiness as a concept.

There were 110 participants who responded; 1 individual skipped the question.



There was a 90% consensus of the importance of protecting trustworthiness. Ten participants responded "disagree" or "strongly disagree" to this question. However upon further review of their responses to question 7, which effectively affirmed that protecting trustworthy data was important to them, we have concluded that 9 of those 10 respondents chose the negative side of the Likert scale in error. Thus we conclude that the trustworthiness of scientific data is almost universally valued by our respondents.

Question #7: Importance of Protecting Trustworthiness (Follow-Up)

Please explain why protecting the trustworthiness of scientific data is or is not important to you. Does its importance change during different phases of the research cycle (e.g., data collection, calibration, processing, analysis, sharing, and publication)?

This is a follow up question to the previous one, with open ended answers. It provides details on why trustworthiness is important to different participants as well as how the importance of trustworthy data changes through the different stages of the research cycle.

There were 97 responses to this question; 14 individuals did not provide an answer. The most common theme was that trustworthy data is a cornerstone in the scientific process. Many

participants cited the scientific process/method specifically, while others used language such as *"The very mission of science as a whole is to produce data that is processed, analyzed, and published with integrity,"* or *"results based on improperly collected, maintained, or understood data are inherently untrustworthy*." Some respondents mentioned that trust can come from data quality and thus quality control/assurance is an important element when producing new data sets.

Some respondents indicated that trustworthiness is equally important at each step, while others said it changes (or at least the perceived/measured trustworthiness changes) during the research lifecycle. For example: "*The trustworthiness of data profoundly changes from the collection to the publication. It is not uncommon for researchers to find errors in initial data and have to recollect it or to find errant data during the processing and analysis. It is hoped that by the sharing stage with colleagues, these issues are found and corrected. Trusted colleagues provide great feedback if an issue with a data set continues to exist. By the publication stage, the data should be free from defects.*"

A small number of respondents brought up cost and/or effort in ensuring trustworthy data. For example, one respondent stated, *"If we are going to spend the time, money, and effort to store, analyze, and publish data, it needs to be trustworthy."* However, receiving funding for it might not be as straightforward according to another respondent: *"This is absolutely critical, however, in current funding environments nobody wants to pay for this. It's always looking at the 'new thing', and not ensuring the 'old thing' is properly maintained."*

Yet another response: *"I feel like this has become an overblown/false concern leading to requirements that can become overly burdensome (convert that to time and money). I have never blindly assumed the data I get from elsewhere is 100% accurate. I have always followed an assess and cross-compare model. Just in case you were wondering, there really isn't a cabal of evil scientists out there creating false, untrustworthy data. For me, the question isn't about trust, but quality. The former implies a deliberate attempt to falsify. The later [sic] implies the reality of life: [expletive deleted] sometimes goes wrong (sensor fails, calibration coefficients are incorrect, vendor didn't properly design sensor, etc)."* This response expresses a notably narrow definition of trustworthiness as compared with most other respondents, referring specifically to other scientists deliberately trying to tamper with data. They care about quality as far as whether the information recorded is accurate, but do not seem concerned about whether the data could have changed since it was first recorded.

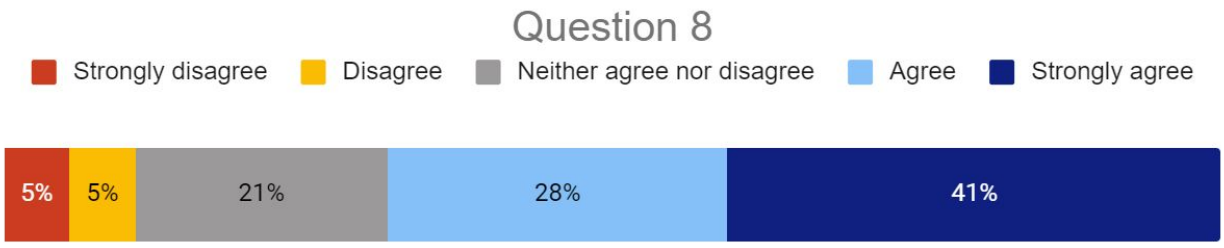
Another small number of respondents felt that trustworthiness is important because of political or anti-science attitudes. Example: *"We live in a society currently affected by notions of mistrust of authority (especially in science), 'alternative facts', and intentional misinformation campaigns. The stakes have never been higher and our need to provide certification and trustworthiness of the provenance of information."* From another respondent: *"In a world that has trended towards partisan division that features a group that is becoming more intensely anti-science, the importance of transparency in data and methods and the appearance of unbiased scientific investigations and discoveries is more important than ever!"*

Lastly, it should be noted that when comparing write-in responses from Question 7 to the responses to Question 6, we have concluded that 9 of the respondents most likely chose the negative side of the Likert scale in error. This represents all but one of the answers selecting the "Strongly Disagree" or "Disagree" options. One respondent selected "Strongly Disagree", and then answered this question with *"Unless you can trust the source of the data, you're not allowed to use it in [a] research paper. Only true data has the ability to illuminate."* Another answered: *"Reliably available, correct data is the most important thing in my job -I literally cannot do my work without it."* Once this discrepancy is taken into account, the consensus among survey takers agreeing that trustworthiness must be protected is substantially higher.

Question #8: Job Responsibilities

Do you agree with the following statement? My job responsibilities include establishing and/or maintaining the trustworthiness of scientific data.

This question was asked to understand if trustworthiness of scientific data is "someone else's problem." For example, scientists may assume that IT support staff are responsible for data protections, while IT support staff may assume that the scientists are responsible.



All survey participants responded to this question, with the majority (69%) self-identifying as people whose job responsibilities include establishing and maintaining trustworthiness of data.

Thus, most of the scientists and IT support staff who responded to the survey feel a shared responsibility for the trustworthiness of scientific data.

Question #9: Potential Consequences

In your work, what are potential consequences (if any) to using/producing/curating scientific data that is not trustworthy?

The question was asked to understand the incentives people have to ensure that data they produce or curate is trustworthy.

There were 93 participants who responded; 18 individuals skipped the question.

A large number of respondents were concerned about reputational risk, both personal (scholarly rebuke, people unwilling to collaborate, loss of position) and to their institutions. Potential consequences mentioned included direct losses of future funding and resulting potential staff cuts. Additionally, people may choose to ignore scholarly work from an organization due to perceived reputation issues, leading to faculty and researchers moving elsewhere. There is also a risk of reputational damage to the associated field of science and methodology where untrustworthy data is generated or used.

One respondent also indicated potential decrease of public trust in science in general as a potential consequence if faulty results in academia are passed to the public sector. Other answers highlighted general public “erosion of trust in science.”

Providers of data repositories also noted the impact of hosting untrustworthy data leading to users not downloading and using data, which they note as their central mission. Also there is a risk of association, where some untrustworthy data makes all the data suspect. As a result all the effort and resources invested in the data go to waste. As one scientific data repository provider explained, “If we provide untrustworthy data as data cutators [sic], that questions all of our work/effort. people are less likely to use our tools or data which may in fact be trustworthy if they have been 'burned' in the past. This affects our reputation, our funding, etc.”

A few respondents also highlighted the impact to the overall scientific process: wrong conclusions can result in retraction of papers, productivity can be reduced as effort is spent handling data security incidents, and using bad data for theoretical modelling can lead to losing

months of researchers' time trying to reconcile known principles with bad observations. Further, it can compound errors in downstream use; for example: “Use of someones [sic] else's data without excellent provenance and an understanding of the data limitations and limited applications results in propagated assumptions through the science process.”

Some respondents identified no consequences of using/producing/curating untrustworthy scientific data as they either work with artificial datasets for teaching, prototype with data without being the primary user of the data, or are solely facilitators. Some CI operators identified themselves as support staff with no direct control of data handling processes and hence no personal consequences to themselves, unless there is something amiss with the underlying cyberinfrastructure. However, one director of research computing services observed that consequences may include “reduced external funding for infrastructure (i.e. NSF MRI or CC* funding for CI)” and that as a result “researchers will use other infrastructure which may not be secure, cause inefficiencies (not centralized).”

A few respondents highlighted potentially bad consequences if untrustworthy data is used in areas such as healthcare, city operations, and others. Another consequence of generating faulty technical and business data is that it may lead to potential loss of millions of dollars for the organization. The issue of government penalties was raised by one respondent, such as when CUI (controlled unclassified information) is mishandled. In clinical research, losing trust can result in loss of study participants which can be detrimental to that research. Also, untrustworthy data can lead to loss of human life (e.g., in health and medical sciences). There is also a risk of untrustworthy data being misinterpreted in larger political context (e.g., in environmental sciences).

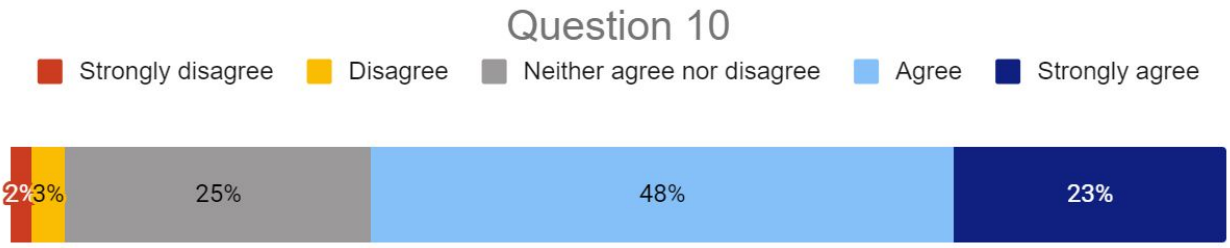
A couple of respondents also highlighted that the issue is not about delivering trustworthy data, but providing the highest quality of data available, that uses well defined and published methodologies, and is reproducible in the sense that steps are clearly listed. They also highlighted the value of adding disclaimers to ensure people using data for analysis understand any assumptions made, or other limitations of the data that could impact reuse: “if I curate the data and share it with the caveat that it should be reused with caution, then the fault may be on the end user for using data that was not trustworthy.”

Question #10: Confidence

Do you agree with the following statement? I am confident in the trustworthiness of the scientific data I use/produce/curate.

This question was asked to gauge people's opinion on their own processes and data they use with respect to trustworthiness.

All participants responded, with the majority (71%) being confident of the trustworthiness of data they use, produce, or curate. A quarter of respondents did not agree or disagree about the trustworthiness of their data. A small percentage (5%) expressed doubt in the trustworthiness of the data.



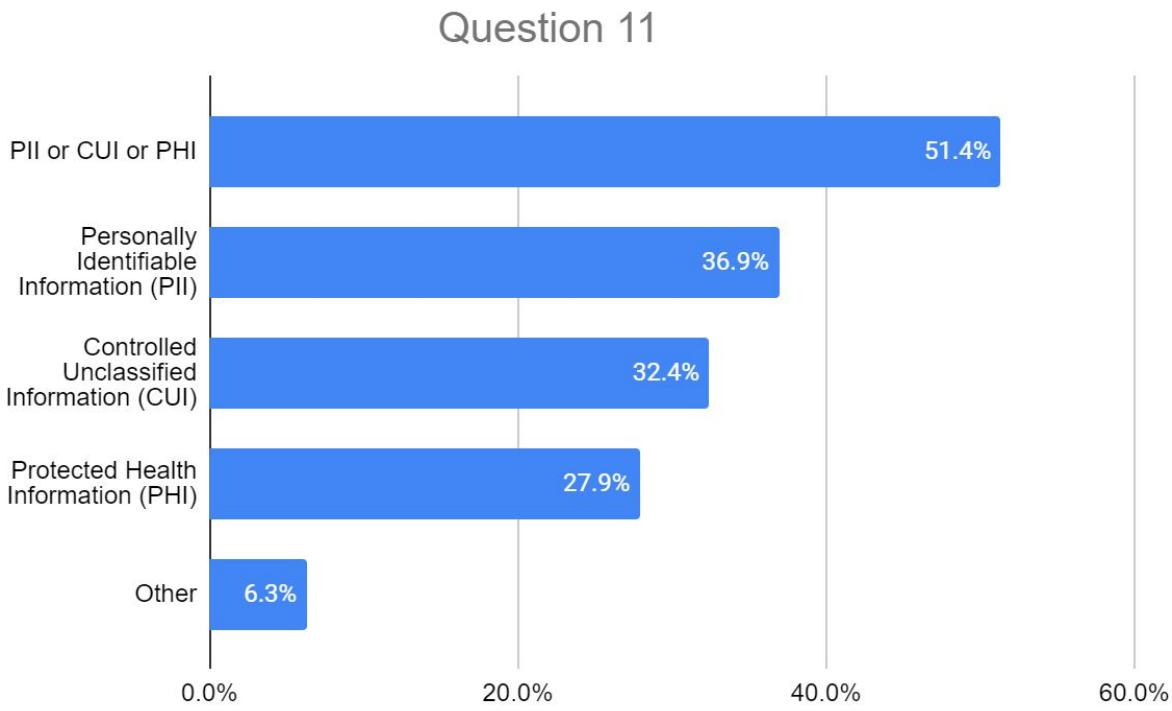
Thus, respondents do not identify a scientific data trustworthiness crisis. They agree that trustworthiness is important for scientific data and that they are achieving the needed attributes of trustworthiness in their experience. While the working group can provide additional support (see below), the respondents feel they are already (mostly) on track.

Question #11: Sensitive Data

Do you perform/support research using sensitive data? Please select any/all that apply.

This question is designed to see if the use and practice of handling sensitive data impacts the individual's thoughts on trustworthy data.

All participants responded, with 57 (51.3%) of the participants indicating use of either PII, CUI, and/or PHI data. Other answers included data from sovereign indigenous peoples, de-identified data, and a small selection of participants using the "other" section to respond that they did not use sensitive data.



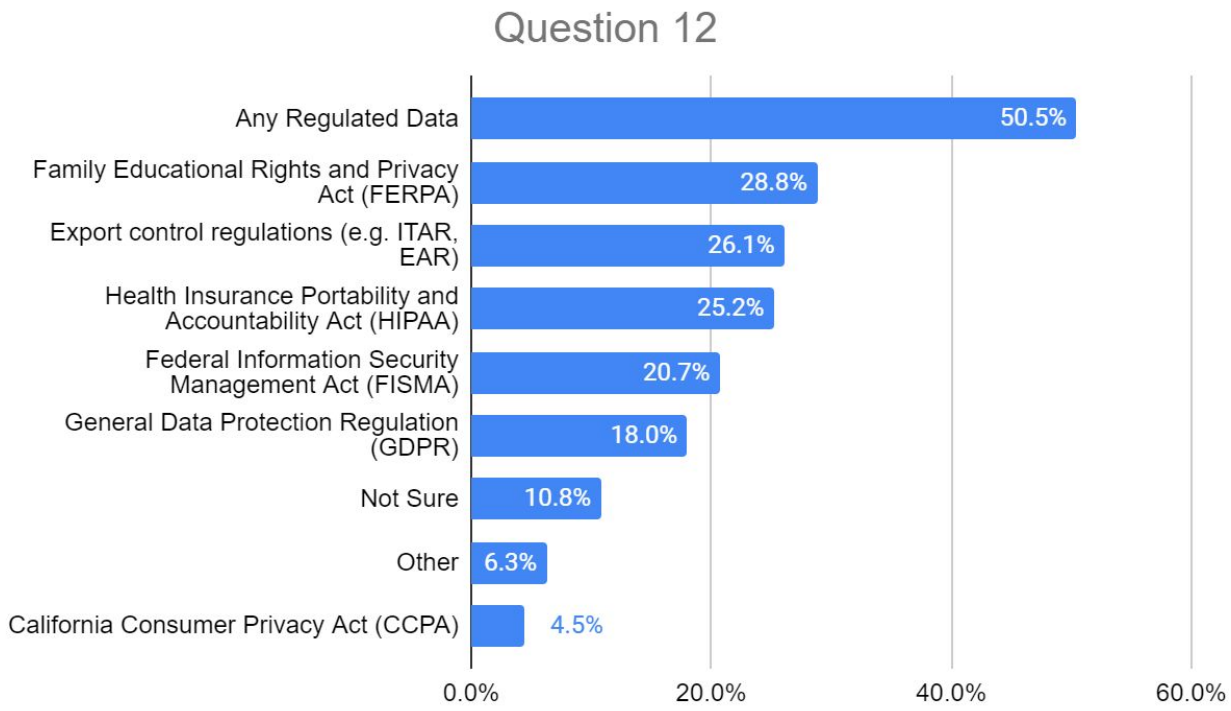
Since over half of our participants work with sensitive data, guidance from the working group should address these types of data.

Question #12: Data Regulations

Do any of the following regulations apply to the research data that you use/produce/curate?

Just like how sensitive data could affect the perception of trustworthiness, practice with handling regulated data could also affect opinions.

Of the 111 participants who responded, 56 (50.4%) are using/producing/curating research data where a regulation applies. In addition to the regulations identified in the question, the respondents included Other responses listing NIST 800, LIMDIS, CMMC, Malaysia National Medicine Research Register and Ethics Approval Body, 45 CFR 46, and RSICC codes.

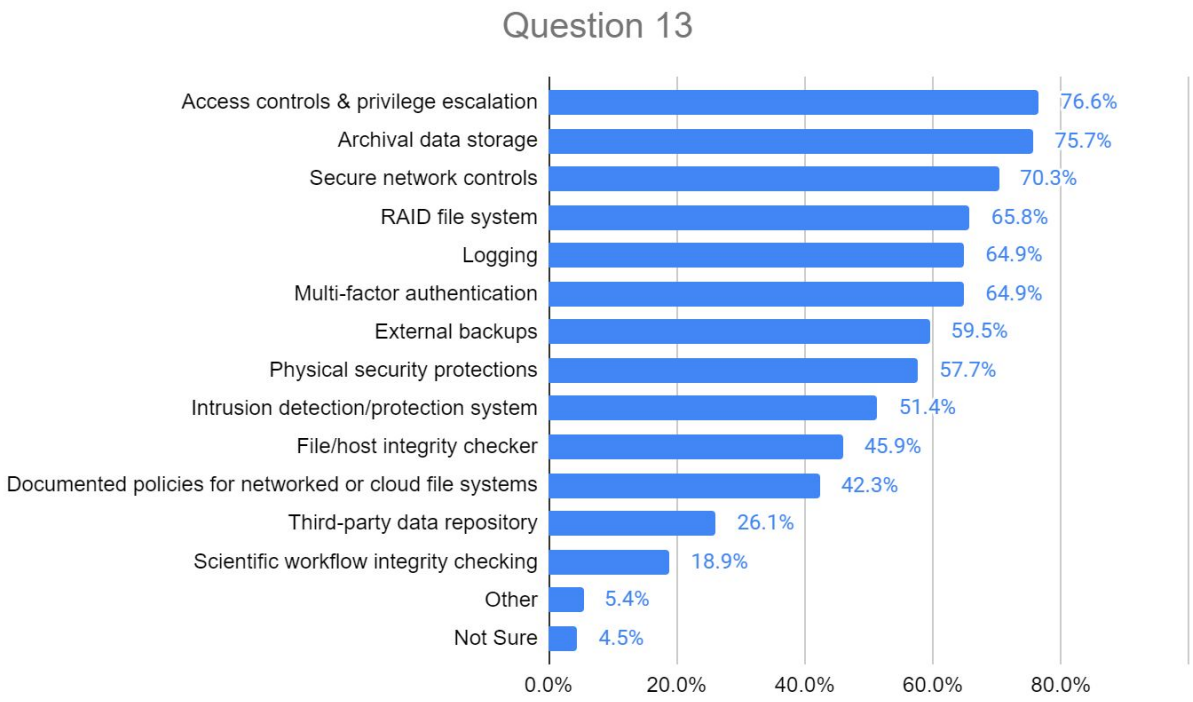


Since half of our participants work with regulated data, this is an important area to cover in future guidance produced by the working group. As expected, there is a strong correlation between work with sensitive data and regulated data: 81% of respondents who work with sensitive data also work with regulated data, and likewise 82% of respondents who work with regulated data also work with sensitive data.

Question #13: Tools and Technologies

Which of the following tools and technologies (if any) help to secure the research data that you use/produce/curate?

Knowing the tools and technologies used by respondents, combined with perspective on how respondents feel about the assurance their tools and technologies give, can give an image of what is perceived to be the most important for trustworthy data.



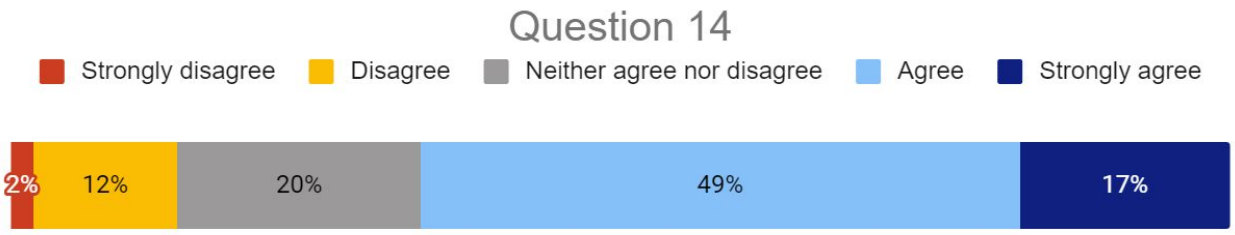
All 111 participants responded to this question. Other responses included “data format compression”, “ curation of data”, and examination under a “certification standard”.

Of interest from the responses is that Multi Factor Authentication (MFA) is not being fully utilized by all participants. Adoption of MFA in the scientific community has been growing (for example, 75% of respondents in the 2019 NSF Community Cybersecurity Benchmarking Survey Report [6] indicated use of MFA). Archival data storage is in second place, but possibly it is being used as a disaster recovery tool and that scientists are encouraged to back up their data. Additional analysis of the popularity of each of these tools by experts in tools and technologies would be beneficial for future work.

Question #14: Sufficiency of Tools and Technologies

Do you agree with the following statement? The tools and technologies that protect the research data that I use/produce/curate provide sufficient assurance against unauthorized changes and/or reputational attacks.

As stated in the previous question, knowing how strongly a survey taker feels about the assurance provided by their tools and technologies reveals the mind set of survey takers.

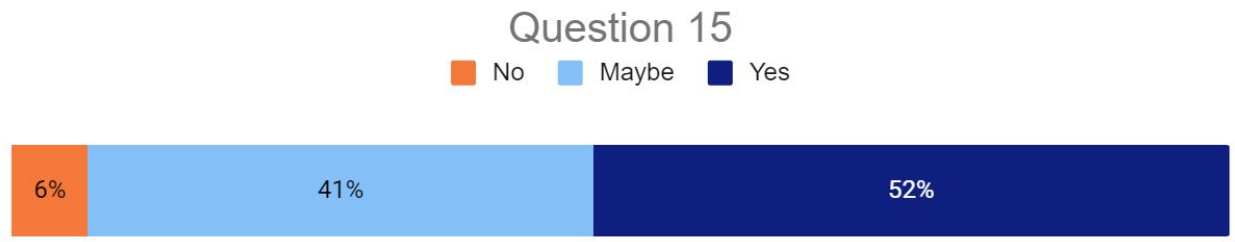


The majority of participants agreed the tools and technologies they use do indeed protect the trustworthiness of their data. Few of the participants did not agree with the statement. Similar to the responses to Question #10 (confidence), our survey data does not indicate a crisis for trustworthy data.

Question #15: Additional Guidance, Tools, and Technologies

If you were provided with additional guidance, resources, or support, would you apply additional tools or technologies to help maintain the trustworthiness of the research data that you use/produce/curate?

This question gives an idea of how receptive the survey takers are to additional help regarding technology and tools. Combined with the previous question of how confident they are about their tools and technologies, it could be possible to see correlation between the amount of confidence and the willingness to consider more technologies and tools.



According to the results in the graph above, respondents are receptive to additional help, even if they generally expressed confidence in the existing tools and technologies (Question 14).

Question #16: Additional Guidance, Tools, and Technologies (Follow-Up)

If you answered Yes or Maybe (Q15), please explain. For example, are there specific tools or technologies you would like to use or specific needs you would like guidance addressing?

There was a lot of interest in additional help and support, and having details on what was needed is vital for future efforts to aid the community.

Out of the 58 "Yes" respondents and the 46 "Maybe" respondents to Question 15, there were 76 responses to this question; 28 participants did not provide an answer, and 7 answered Question 15 as "No" so they were not required to answer.

This was a hard question to answer as some individuals aren't well-versed in what they need to improve or what they haven't thought about because of their lack of knowledge in regards to what trustworthy data needs to be verified. Example responses include: *"Unsure, having brief summaries of each* [tool] *to learn more would be useful." "Not sure, better tools for provenance." "Not sure what those additional tools would look like." "I don't know how to respond about tools and technologies that I am ignorant of."*

Many individuals could not name specific tools or specific needs, but said they were open to new and improved technologies that become available that highlight vulnerabilities. *"Nothing specific; as new guidance is made available, we will continue to adopt and implement best practice." "There are always new and better ways to protect data and so I am open to learning about new tools." "If someone presents me with a better/more effective idea for protecting data and maintaining its trustworthiness, that seems like a good investment for me." "The more guidance/resources/support that is available to everyone the better in my opinion. I am sure some of it would [be] of interest to our community." "Better guidelines with real examples are always welcome when creating and updating security policies."*

While a few participants were very specific: *"The following tools/resources are required to make my data more trustworthy: QA (quality assurance) software for acquisition and processing (provenance generation through manual and automated entry); data management software that ties QA metadata to data streams and points; QC (quality control) software that allows repeatable QC/analysis and generates QA information for those processes; offsite data repositories with metadata indexing." "Encryption." "Tooling to validate integrity of container images. Tooling and methodology to perform 'security posture checking' of end user devices*

that does not limit end users' control over their devices nor their access to the network." "Tools/technologies around CUI. Tools". "We use Globus for file transfers, but are not licensed for more secure features like encryption and handling CUI. Upgrading the license would make secured transfers available to our CUI -based researchers." "We are looking at ERPID to generate PIDs and track the data products that way..But other tools would be of interest as well."

Some participants answered in regards to future concerns revolving around how data will be saved, where it will be saved, maintaining its readability, sustaining it through lack of funding, and updating systems to fulfill new local and federal requirements. "*There is a lot of social media data harvesting that is used in research today. There are no guidelines, it is up to every individual researcher." "Having lived through numerous storage media revolutions, I am mildly concerned with future-proofing the readability of data I save. I have CDs full of data that I can no longer open, and I worry USB drives and even cloud storage will be the same." "I think we do a pretty good job already because it matters to us intrinsically. I'm more worried about new onerous requirements we'll have to implement solely as a CYA…" "Long term preservation on public cloud platforms while avoiding vendor lock-in."*

Some of the participants believed that an internal audit as a first measure of their current implementation would be best before adopting new tools and technologies. One stated: *"An audit of our current practices would be more beneficial, along with suggestions for best practices or inclusion of additional trustworthy tool chains".* And another: *"What would be useful is to have a low or no cost *friendly* "audit" (i.e., one for our use, not to rat us out to funding agencies) to identify low hanging fruit. i.e., opportunities for improvement with positive ROI."*

Many questioned the action of implementing new tools/technologies because of their budget-conscious mindset. *"There are always mechanisms to improve security and trustworthiness, but whether they are feasible depends on the time and money needed to implement them. We operate under strong financial constraints, and generally can only implement changes for the highest priority issues." "I would like to explore external backups (need time for that) and archival storage (need money for that)..." "Depends on the effort required. We already don't have enough staff to implement all the security policies we want."*

A few participants voiced that while they might consider tools to adopt, the ease of use is a main concern. *"There is a lot of complexity in the acquisition of data, and mechanisms to ensure trustworthiness (e.g. provenance capture, integrity checks) often add to that complexity. If tools*

are to be adopted, it needs to be easy to adopt them, or the necessary features need to be added to existing toolchains." "I am not really concerned about unauthorized changes/ reputation attacks. If the tools were easy to use I'd probably use [them]*."*

To summarize, most respondents want to learn about new tools and technologies to improve their data security in terms of encryption, vulnerability points, tracking, and future-proofing. Many don't know what tools they should be considering and most are concerned that implementation will require time and money that is hard to find in addition to an unknown learning curve for these new tools.

Question #17: Request for Additional Information

Is there anything else related to trustworthy research data that you would like us to know?

There were 34 responses to this question. Noteworthy responses include those from individuals who believe data trustworthiness to be important, but they *"can't image [sic] anyone wanting to actively try to corrupt our data or systems"* and feel *"the restrictions now put in place (dual-authentication, VPN's, firewalls, etc.) make it almost impossible to work outside of my office."*

Another respondent said *"I see the efforts around scientific data security focusing on protecting the data as if someone really wants to alter my data instead of using my server to conduct other, more lucrative nefarious acts. This false focus leads to requirements and policies that is [sic] starting to make managing my servers prohibitively expensive."* This respondent did not feel trustworthiness was the most important security concern. Out of this comment came a realization that data trustworthiness could be collateral damage to other threats targeting the system the data is stored on. If a system is compromised, the data on that system becomes subject to tampering from the malicious attacker.

A different respondent pointed out that *"IT in academia generally takes a relaxed view of security."* This may be due to a few factors, such as academia not being as worried about data being stolen because they typically work in open access environments, or not being as concerned about international access to data like in some federal environments.

Concern was shared about the effort to fund trustworthiness initiatives by some respondents: *"These efforts need to be funded at an institutional level. It will not happen at the individual*

researcher level." This is reflected by another respondent, saying *"researchers seem to not worry about this very much."*

Question #18: Contact for Clarification

May we contact you for clarification of your responses if needed?

73 responded Yes, 36 responded No, and 2 skipped this question. As described above, we contacted the initial set of respondents to obtain their answer to Question 5, which was accidentally omitted from the initial published version of the survey.

Question #19: Notification of Draft Report

Would you like to be notified when our draft report is ready for review?

85 responded Yes, 24 responded No, and 2 skipped this question. We will be contacting the 85 respondents who selected Yes to get their feedback on this report and potentially revise the report based on that feedback.

Question #20: Contact Information

If we may contact you and/or you would like to be notified, please enter your name and email address.

The survey team stored names and email addresses separate from the rest of the survey data, to be used only for clarification/notification purposes. All survey analysis was performed using the de-identified data set.

4 Conclusion and Next Steps

The survey provided the working group with a breadth of perspectives from scientists across many scientific disciplines and from cyberinfrastructure professionals in a variety of roles. Respondents overwhelmingly agreed with the importance of protecting the trustworthiness of scientific data, and they are using multiple tools and technologies to accomplish that goal. 52% of respondents indicated that they would apply additional tools or technologies to help maintain the trustworthiness of research data, if they were provided with additional guidance, resources, or support, which motivates the Trustworthy Data Working Group to develop such guidance, with reference to existing resources. These resources include standards such as NIST 1800-25 [3] along with related work from RDA, ESIP, and others on "TRUST Principles for Digital Repositories" [4] and "Risk Assessment for Scientific Data" [5]. The working group will also be providing input into the next revision of the Open Science Cyber Risk Profile (OSCRP).[16]

For more information about the Trustworthy Data Working Group, please visit https://www.trustedci.org/2020-trustworthy-data.

[16] https://trustedci.github.io/OSCRP/OSCRP.html

References

[1] J. Bhandari Neupane, R. P. Neupane, Y. Luo, W. Y. Yoshida, R. Sun, and P. G. Williams, "Characterization of Leptazolines A–D, Polar Oxazolines from the Cyanobacterium Leptolyngbya sp., Reveals a Glitch with the 'Willoughby–Hoye' Scripts for Calculating NMR Chemical Shifts," Org. Lett., vol. 21, no. 20, pp. 8449–8453, Oct. 2019. https://doi.org/10.1021/acs.orglett.9b03216

[2] M. Rynge *et al.*, "Integrity Protection for Scientific Workflow Data: Motivation and Initial Experiences," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines - PEARC '19*, Chicago, IL, USA, 2019, pp. 1–8. https://doi.org/10.1145/3332186.3332222

[3] NIST Special Publication 1800-25 [A-C] - Data Integrity Identifying and Protecting Assets Against Ransomware and Other Destructive Events." https://www.nccoe.nist.gov/sites/default/files/library/sp1800/di-identity-protect-nist-sp1800-25-draft.pdf

[4] Lin et al., 2020. The TRUST Principles for Digital Repositories. Scientific Data. https://doi.org/10.1038/s41597-020-0486-7

[5] Mayernik, Matthew S., Kelsey Breseman, Robert R. Downs, Ruth Duerr, Alexis Garretson, Chung-Yi (Sophie) Hou, and Environmental Data Governance Initiative (EDGI) and Earth Science Information Partners (ESIP) Data Stewardship Committee. 2020. "Risk Assessment for Scientific Data." Data Science Journal 19 (10): 1–15. https://doi.org/10.5334/dsj-2020-010

[6] S. Russel, “2019 NSF Community Cybersecurity Benchmarking Survey Report”, Trusted CI Technical Report, December 2019. https://hdl.handle.net/2022/24912

[7] Reinhard Gentz and Sean Peisert, “An Examination and Survey of Random Bit Flips and ScientificComputing,” Trusted CI Technical Report, December 20, 2019. https://hdl.handle.net/2022/24910

Appendix A: Online Survey Questionnaire

1. What is your primary job title?
[Short answer text field]

2. Please select all roles that describe your work, even if they do not correspond to your job title.

- ❑ Compliance officer
- ❑ Cybersecurity analyst/engineer
- ❑ Educator
- ❑ Infrastructure provider/operator
- ❑ Research computing facilitator
- ❑ Research software engineer
- ❑ Scientific data creator
- ❑ Scientific data user
- ❑ Other: _____

3. In the above roles, what field(s) of science do you primarily work in or support?

- ❑ Astronomical Sciences
- ❑ Biological and Environmental Sciences
- ❑ Computer and Information Sciences
- ❑ Economic Sciences
- ❑ Education
- ❑ Engineering
- ❑ Geosciences
- ❑ Health and Medical Sciences
- ❑ Mathematical Sciences
- ❑ Physics
- ❑ Social and Behavioral Sciences
- ❑ No specific area of focus/support
- ❑ Other: _____

4. What sector(s) do you work with?

- ❑ Academia
- ❑ Government
- ❑ Industry
- ❑ Non-profit
- ❑ Other: _____

5. Which attributes do you believe scientific data must have in order to be trustworthy?

- ❑ Accuracy - The data is free from error.

- ❑ Integrity - The data has not been altered.
- ❑ Methodology - The processes and inputs used to create the data are well-established and accepted by the community.
- ❑ Provenance - The data's origin and lineage can be readily established.
- ❑ Reproducibility - The data can be re-created, or the associated scientific results are replicable.
- ❑ Reputation - The data was generated by a credible or trusted source.
- ❑ Responsible stewardship - The ownership of the data is well managed and can be transferred.
- ❑ Significance - The data enables future research directions (with associated funding/support).
- ❑ Other: _____

6. Do you agree with the following statement?
I think that protecting the trustworthiness of scientific data is important.

- ❑ Strongly disagree
- ❑ Disagree
- ❑ Neither agree nor disagree
- ❑ Agree
- ❑ Strongly agree

7. Please explain why protecting the trustworthiness of scientific data is or is not important to you. Does its importance change during different phases of the research cycle (e.g., data collection, calibration, processing, analysis, sharing, and publication)?
[Long answer text field]

8. Do you agree with the following statement?
My job responsibilities include establishing and/or maintaining the trustworthiness of scientific data.

- ❑ Strongly disagree
- ❑ Disagree
- ❑ Neither agree nor disagree
- ❑ Agree
- ❑ Strongly agree

9. In your work, what are potential consequences (if any) to using/producing/curating scientific data that is not trustworthy?
[Long answer text field]

10. Do you agree with the following statement?
I am confident in the trustworthiness of the scientific data I use/produce/curate.

- ❑ Strongly disagree
- ❑ Disagree
- ❑ Neither agree nor disagree

- ❑ Agree
- ❑ Strongly agree

11. Do you perform/support research using sensitive data? Please select any/all that apply.

- ❑ Controlled Unclassified Information (CUI)
- ❑ Personally Identifiable Information (PII)
- ❑ Protected Health Information (PHI)
- ❑ Other: _____

12. Do any of the following regulations apply to the research data that you use/produce/curate?

- ❑ California Consumer Privacy Act (CCPA)
- ❑ Export control regulations (e.g. ITAR, EAR)
- ❑ Family Educational Rights and Privacy Act (FERPA)
- ❑ Federal Information Security Management Act (FISMA)
- ❑ General Data Protection Regulation (GDPR)
- ❑ Health Insurance Portability and Accountability Act (HIPAA)
- ❑ Other: _____
- ❑ Not Sure

13. Which of the following tools and technologies (if any) help to secure the research data that you use/produce/curate?

- ❑ Access controls & privilege escalation
- ❑ Archival data storage (e.g., long-term data preservation)
- ❑ Documented policies for networked or cloud file systems (e.g., Google Drive)
- ❑ External backups
- ❑ File/host integrity checker
- ❑ Intrusion detection system / intrusion protection system
- ❑ Logging
- ❑ Multi-factor authentication
- ❑ Physical security protections (e.g., locked file cabinets, restricted data center access)
- ❑ RAID file system (or other mechanisms to ensure data-at-rest integrity)
- ❑ Scientific workflow integrity checking (e.g., Scientific Workflow Integrity with Pegasus)
- ❑ Secure network controls (firewall, science DMZ, encrypted remote communication, flow monitoring, etc.)
- ❑ Third-party data repository
- ❑ Other: _____
- ❑ Not Sure

14. Do you agree with the following statement?

The tools and technologies that protect the research data that I use/produce/curate provide sufficient assurance against unauthorized changes and/or reputational attacks.

- ❑ Strongly disagree
- ❑ Disagree
- ❑ Neither agree nor disagree
- ❑ Agree
- ❑ Strongly agree

15. If you were provided with additional guidance, resources, or support, would you apply additional tools or technologies to help maintain the trustworthiness of the research data that you use/produce/curate?

- ❑ No
- ❑ Maybe
- ❑ Yes

16. If you answered Yes or Maybe, please explain. For example, are there specific tools or technologies you would like to use or specific needs you would like guidance addressing?
[Long answer text field]

17. Is there anything else related to trustworthy research data that you would like us to know?
[Long answer text field]

18. May we contact you for clarification of your responses if needed?

- ❑ No
- ❑ Yes

19. Would you like to be notified when our draft report is ready for review?

- ❑ No
- ❑ Yes

20. If we may contact you and/or you would like to be notified, please enter your name and email address.
Name: [Short answer text field]
Email: [Short answer text field]