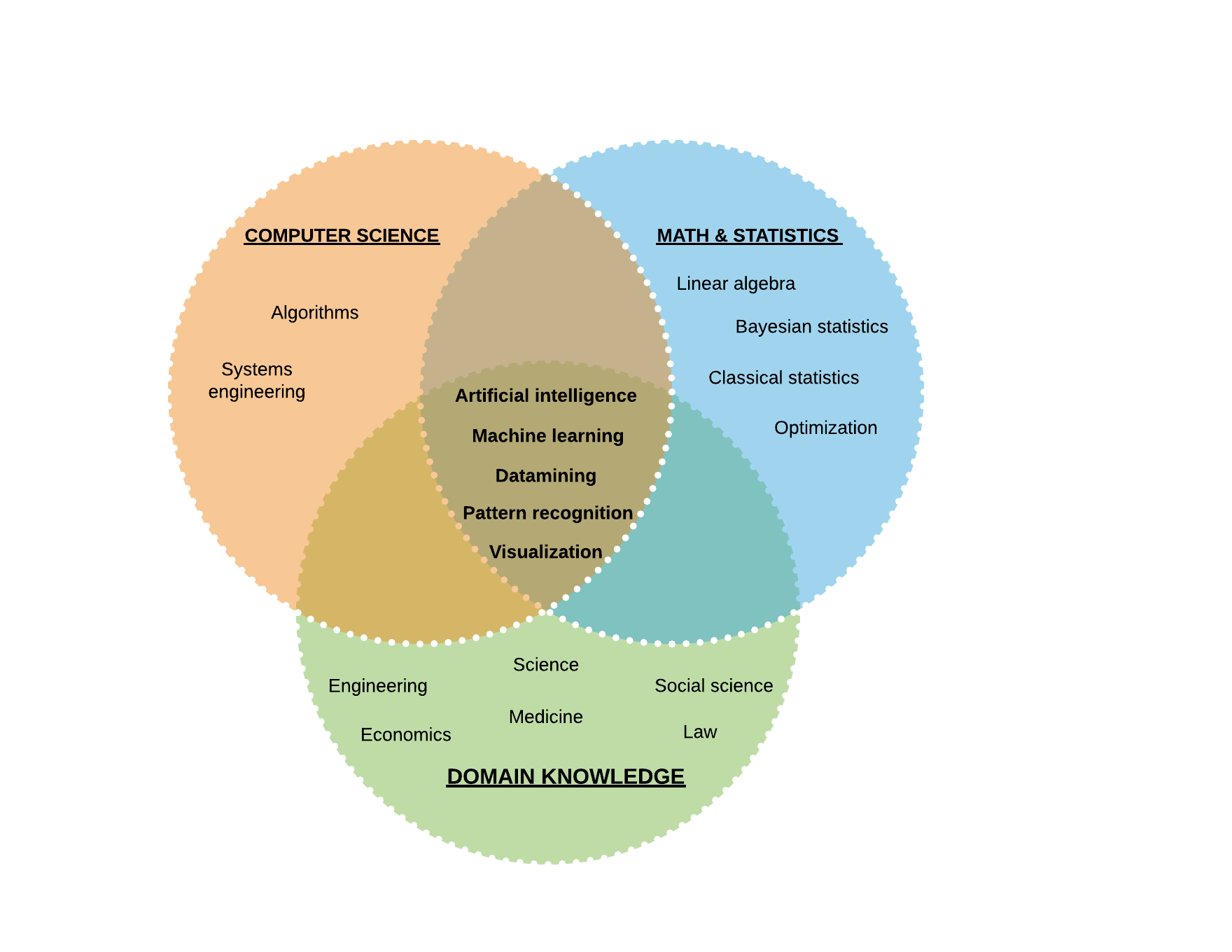
**NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY (NIST)  
BIG DATA INTEROPERABILITY FRAMEWORK**

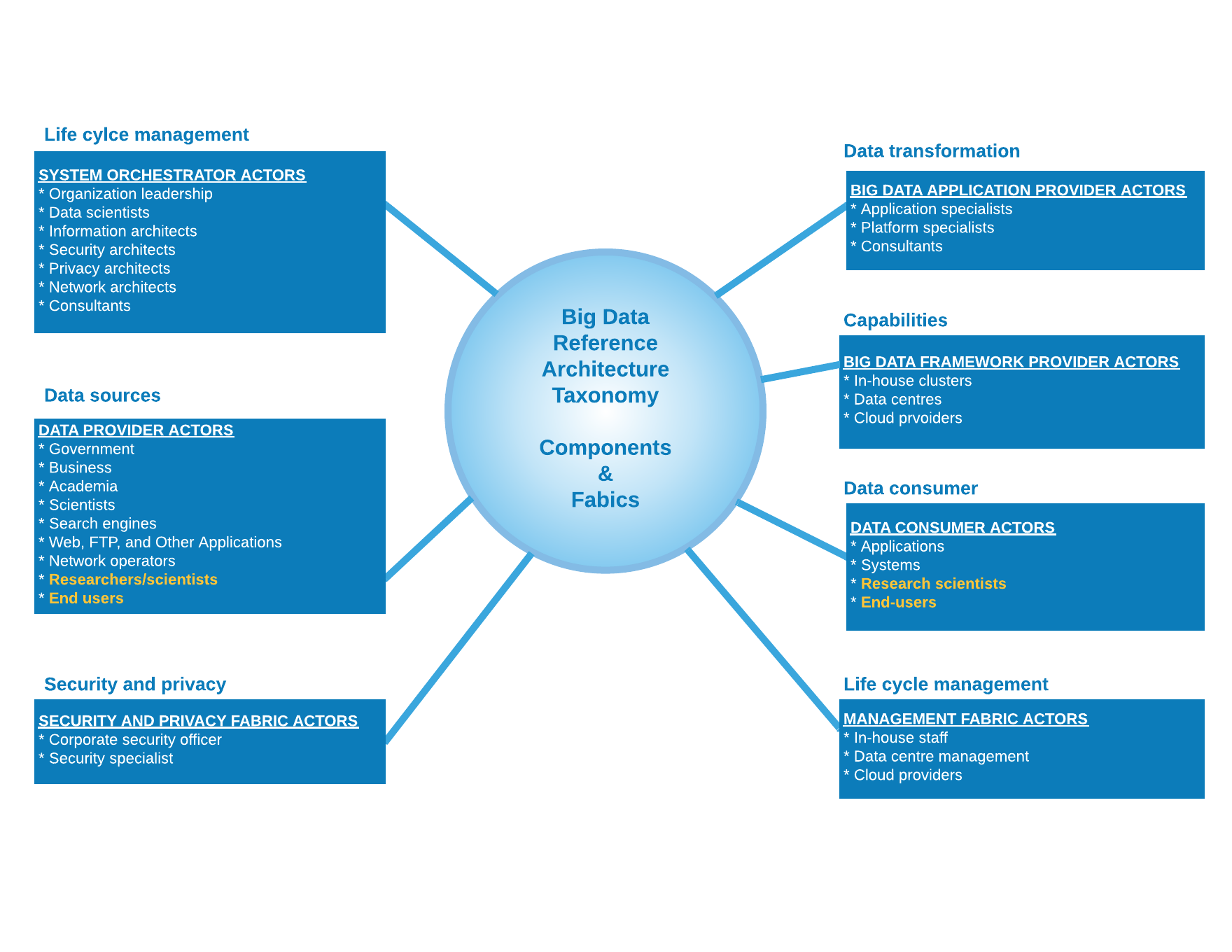
**Version 3**

**SUBMISSIONS  
2018-06-01**

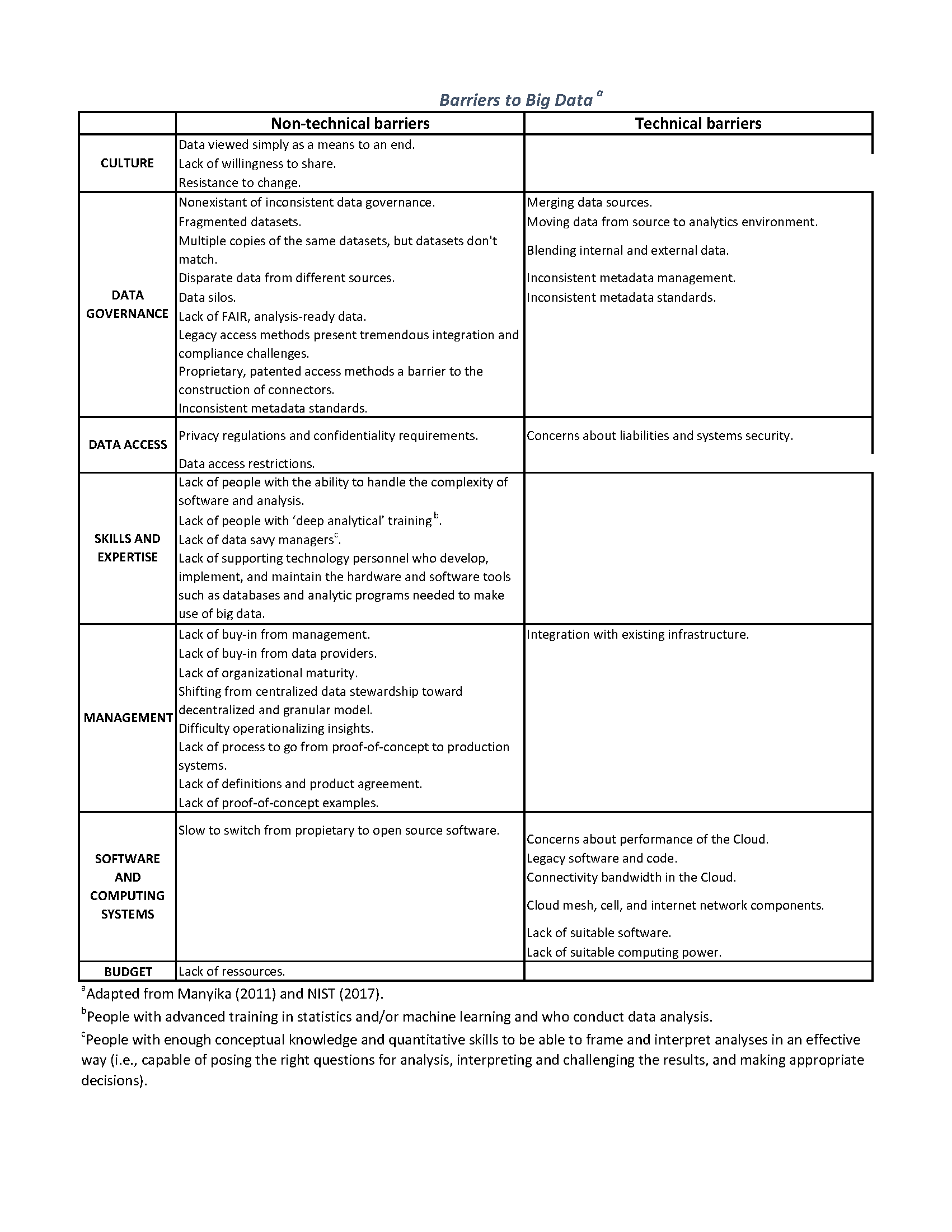
**REPLACEMENT FIGURE  
Volume 1 - Figure 2 – Data science  
Justification**: To better contextualize overlapping roles



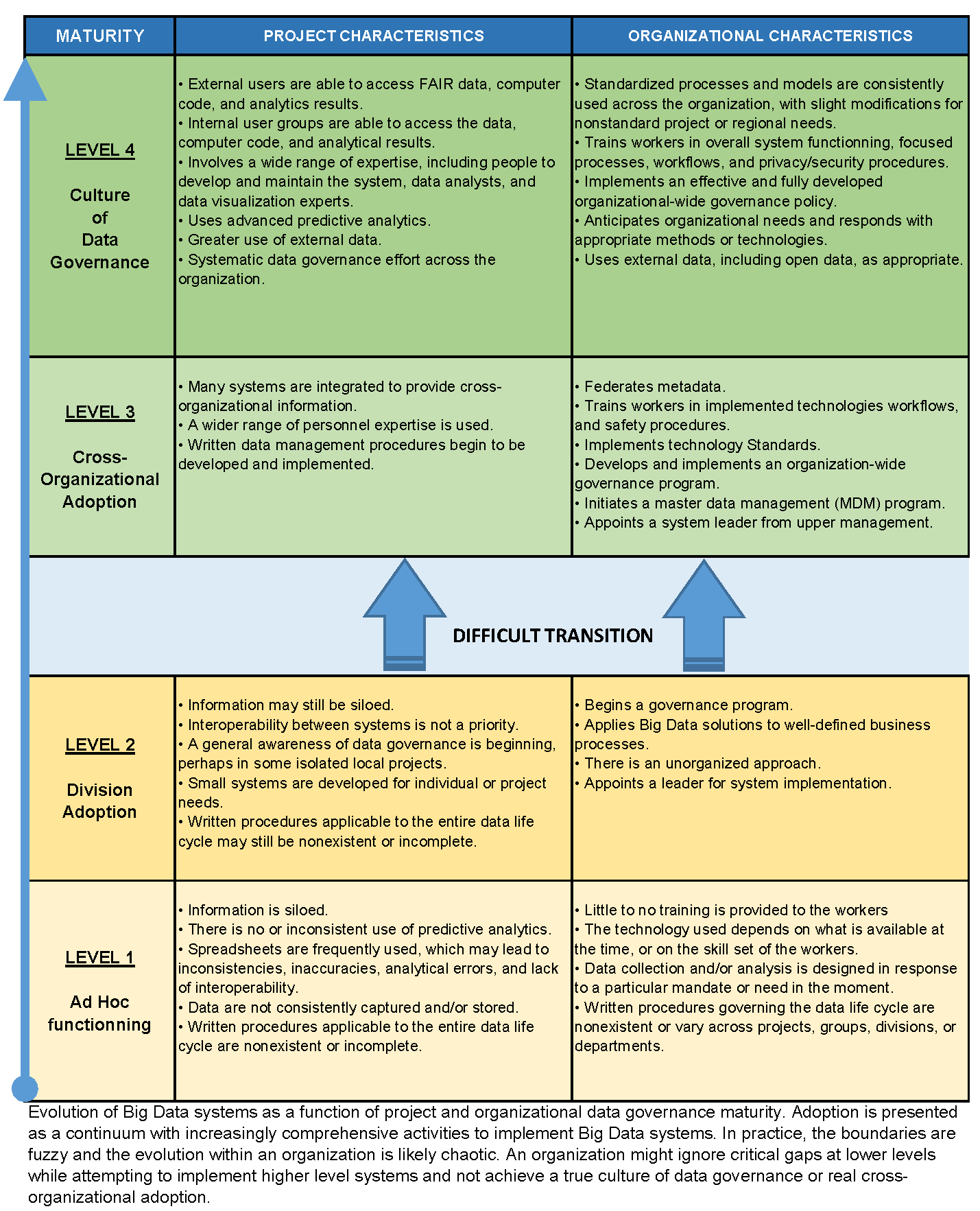
**REPLACEMENT FIGURE  
Volume 7 - Figure 1 – NIST Big Data reference architecture taxonomy  
Justification**: To make the link with Volume 7, Table 2 (Mapping use case characterization categories to reference architecture components and fabrics).

****

**REPLACEMENT TABLE  
Volume 9 – Table 4 – Nontechnical and technical barriers to adoption  
Justification**: Organize examples by major groups and provide more examples.



**REPLACEMENT FIGURE  
Volume 9 – Figures 1 & 2 – Governance gap and organizational maturity  
Justification**: Merge Figures 1 & 2 and add text in a single Figure for clarity.



**VOLME 9 – NEW SECTION - BIG DATA READINESS**

Big Data has the potential to answer questions, provide new insights previously out of reach, and strengthen evidence-informed decision making. However, the harnessing of Big Data can also very easily overwhelm existing resources and approaches, keeping those answers and insights out of reach. As the success of Big Data system adoption relies heavily on organizational maturity (**Vol 9, Fig 1 & 2**), this section offers suggestions for a PATH TO BIG DATA READINESS for the data provider (**Vol 6, Fig 2, top left corner**). An organization does not need to wait for the development of a Big Data Framework (**Vol 6, Fig 2, 4, 6-8**) to take action and help accelerate the implementation of *Big Data*. Tactical actions directed at the working level can help enable *Big Data* without overwhelming workers, managers, or stakeholders and increase the chances of success of a Big Data framework:

1. Create awareness of “*Big Data readiness*” from the bottom up in operations and research contexts via communications such as newsletters, bulletins, and a dedicated website or wiki.
2. Provide online training modules to increase digital literacy across the organization.
3. Deploy “It’s good enough” checklists for data Findability, Accessibility, Interoperability, and Re-usability (FAIR data) to help data providers produce data that are ready for Big Data workflows.
4. Implement “*user-centric,*” approaches to data preparation to replace project- and client-centric approaches. [[1]](#footnote-1)
5. Create “*linear data pathways*” to authoritative data sources to eliminate data fragmentation, duplication, and to preserve data lineage.
6. Develop and pilot test models of data-intensive scientific workflows for the preparation of FAIR, tidy, and analysis ready data and “*reproducible science*” in line with national and international best practices.
7. Implement semi-automated data verification and feedback loops to ensure that data are ready for integration into Big Data workflows.
8. Maximize chances of success of Actions 1-7 by including data providers in the development of solutions.

Generic and universal solutionsPutting the initial focus on structured digital scientific data and the identification of a pathway from Small Data to Big Data for a stepwise, will provide a rational approach to harnessing Big Data. Implement actions that are generic and independent of systems currently in place. This means that they can be implemented “*now*.”

“Lock-in”  
Best practices in data management have not kept up with changes in technology that resulted in a rapid increase in the speed of generation, quantity, variety, complexity, variability and new uses for the data collected. There is, an addition, uncertainty regarding data accuracy, inconsistency in vocabulary and confusion over the meaning of Big Data. Meanwhile, organizations are still struggling to emerge from a paper-based world governed in silos to a digitally interconnected world. This is a difficult transition. It requires the transformation of longstanding, well-adapted thinking processes that no longer work well, to new thinking processes adapted to a new world.

# Change in thinking

*Big Data* is being propelled from an emerging area to the fore of Open Data and Open Science. However, data that may be “locked in” traditional approaches are largely inaccessible to Big Data. This limits an organization’s ability to use Big Data approaches for knowledge acquisition and innovation. A change in thinking across organizations is needed to achieve a coordinated and harmonized system that is simple, effective and geared to meet organizational needs.

# Culture change

Operational and research programs have developed data management processes that work for them internally. They tend to be project- or client-centric to meet their specific mandate and needs, but not necessarily user-centric in the context of Open Science and Big Data. A paradigm shift in thinking and culture is needed across organizations to achieve agile delivery of “*analysis-ready*” data that can be incorporated seamlessly into a *Big Data* workflow. The underlying principle for success is a “*Big Data readiness*” approach from the bottom up at the working level in operations and research. Targeted generic actions will help create the necessary conditions on the ground – culture change will follow.

Big data readiness  
Data providers in the field, laboratory, and other organizational levels need to recognize at the outset that the data users, how the data will be used, and for what purpose are unknown. Data transmitted from one person or group to the next must be FAIR and tidy. FAIR data include all related metadata and documentation so that an unknown end-user can completely understand the data and the data quality without having to contact the data provider. FAIR data have been verified by the data provider to be “fit for use” by any unknown user who is then in a position to assess whether or not the data are “fit for purpose” in some specific context. FAIR, tidy, analysis ready data can be easily integrated into a Big Data workflow. A *Big Data* readiness approach at the working level will concomitantly help solve existing data flow and data quality issues irrespective of whether or not the data will eventually enter a *Big Data* workflow. A *Big Data* readiness approach will improve an organization’s overall data stewardship and governance, help make Open Data and Open Science a reality, and improve the chances of success of future corporate solutions related to Big Data and analytics.

Data governance gaps  
There is a need for common data Standards for the preparation and updating of FAIR data. Previous approaches to Data Governance may have led to data fragmentation (**Vol 9 Fig 8 & 9**), variation in data quality, and incomplete information concerning the data. Where this may be satisfactory within specific mandates, it is problematic for *Big Data*. In order to use such data, each user inherits the task of reassembling the data before being able to use them, yet lacks all the information needed to perform the task reliably. This is an error-prone, costly, time consuming, and inefficient use of resources. Furthermore, it is unlikely that data reassembled by different end-users will result in matching datasets. The problem compounds exponentially when trying to integrate these data into *Big Data.* Targeted actions address gaps in Data Governance to improve the ability to integrate data from multiple sources and to extract reliably new knowledge and insights from large and complex collections of digital data. Adopting a Big Data readiness approach in an organization can help enable Big Data analytics, machine learning and Artificial Intelligence (AI).

# Harnessing Big Data

Harnessing Big Data means extracting more knowledge from existing data. A major hurdle is data preparation which can take up to 70% or more of the total time (**Vol 9 Fig 9**), essentially performing tasks left undone when data are not FAIR (**Vol 9 Fig 10**). The solution to extensive data preparation time is improved data governance. Time is thus freed for the harnessing of Big Data in the continuum of reproducible science (**Vol 9 Fig 11**).

# Disrupting the status quo

Implementation of a *Big Data readiness* approach at the working level may be easier to implement than imagined. The person best equipped to prepare “*analysis-ready*” data is the data provider – the person at the data source who knows the data best. Success requires inclusion of data providers – especially those who are experiencing the greatest challenges – in developing solutions. Inclusion means going beyond providing support. It means saying not only, “*What can we do for you?*” but also, “*This is what we need from you*.” It means disrupting the status quo. *Big Data readiness* requires a paradigm shift in thinking at the working levels that is revolutionary, not evolutionary.

It’s good enough  
Overwhelming people can be avoided by developing well thought out, “*It’s good enough*” modular checklists that will result in what is needed – now – to move forward on the pathway to *Big Data*. It is unrealistic to expect that people at the working level, in the field and in the laboratories, have or can acquire the necessary skills and tools to design and maintain databases or to output their data in unfamiliar formats. However, it is realistic and necessary to expect that they can output their data in a form that can be easily understood and used by other people and systems. If this is achieved, it will be good enough (**Vol 9 Table 13**).

# Cost savings

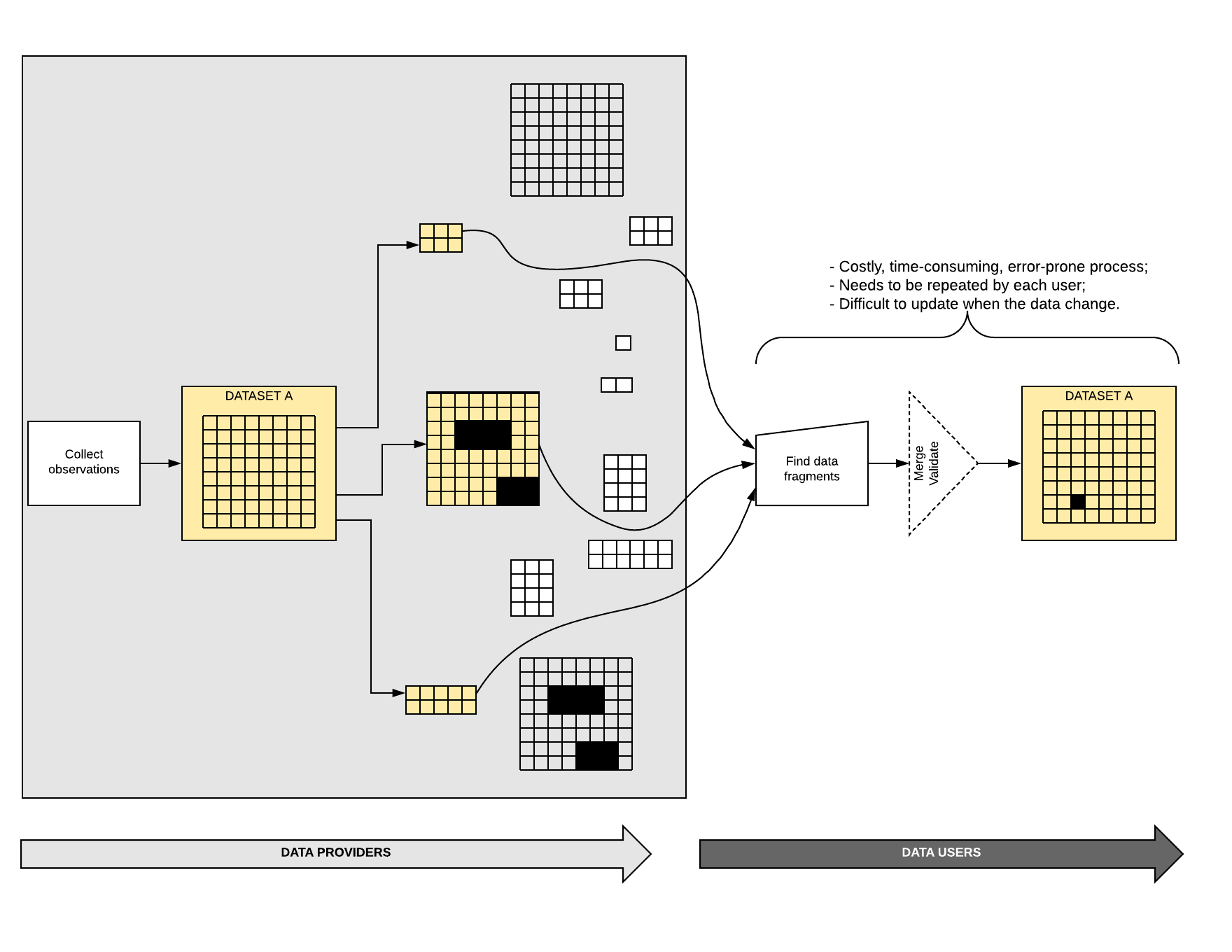
Big Data and Data Governance are also tools to reduce costs. Big Data reduces costs by using existing data instead of collecting more data unnecessarily. Big Data may also reduce costs getting better answers quicker. However, Big Data will not improve data quality, solve data management problems, or obliviate the need for good quality, well managed data. Good data governance and FAIR data will result in the reduction or elimination of inefficiencies and costly errors. Improved data quality, usability and discoverability will increase the value of data products thereby providing a bigger return on investment.

# The bigger picture

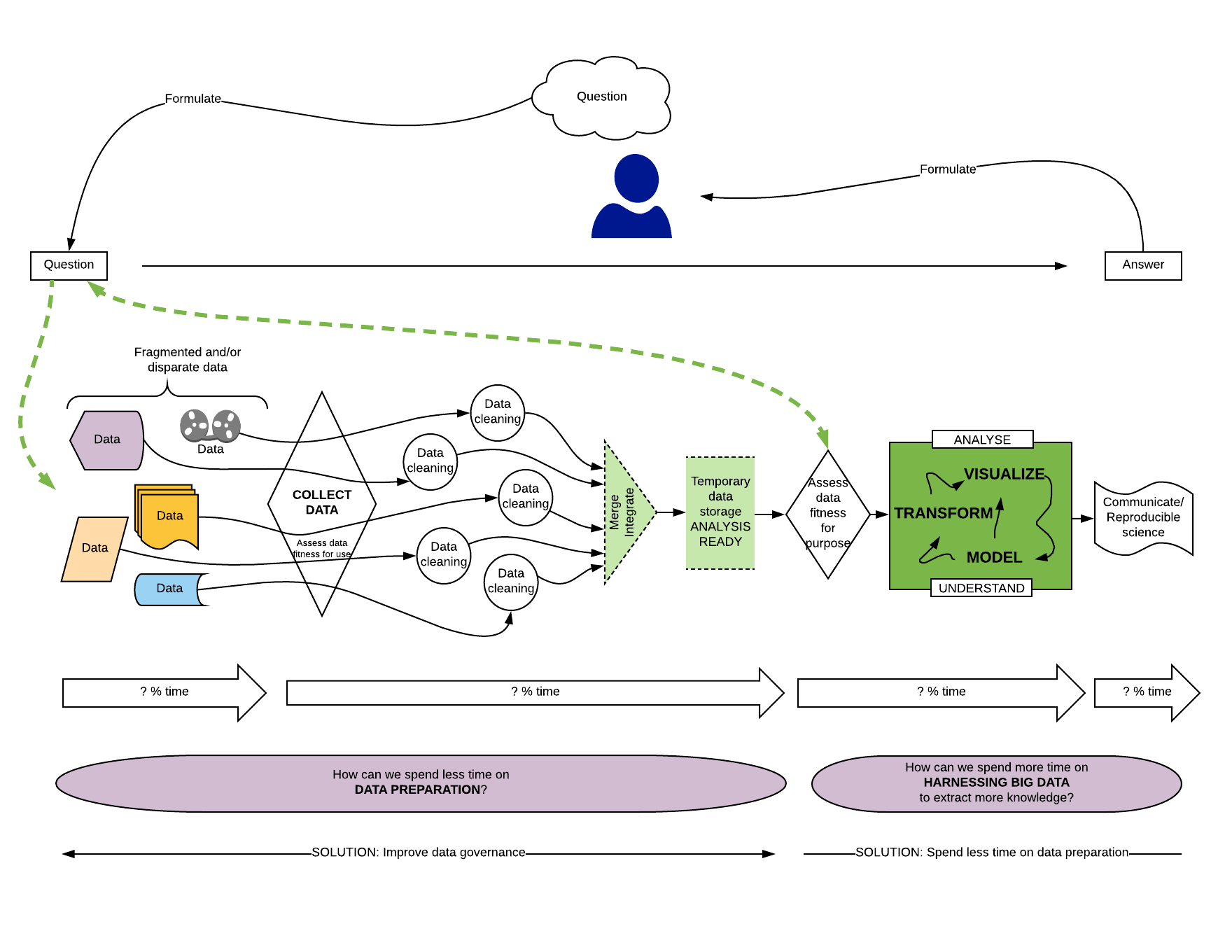
Data management gaps, at both the working and corporate levels, reflect the state of affairs in the private and public sectors in developed countries that are dealing with decades old legacy systems and ways of doing things. *Big Data* is a rapidly evolving area as evidenced by current efforts to develop new international Standards to provide guidance as we collectively move forward with *Big Data*. These will inevitably and necessarily profoundly impact overall data management practices at all levels and for all types of data.

It is important that an organization identify the technical and non-technical barriers to Big Data (**Vol 9 Fig 4**). Contextualization of a path to Big Data readiness within a framework that describes Big Data reference architecture and Big Data governance and metadata management is also important. However, an effective first step will emphasize what can be done now in the present taking into account current realities (**Vol 9 Fig 1 &2**) to position the organization to meet opportunities provided by the Big Data revolution. As the organization matures it will be able to implement linear pathways to authoritative data (**Vol 9 Fig 12**).

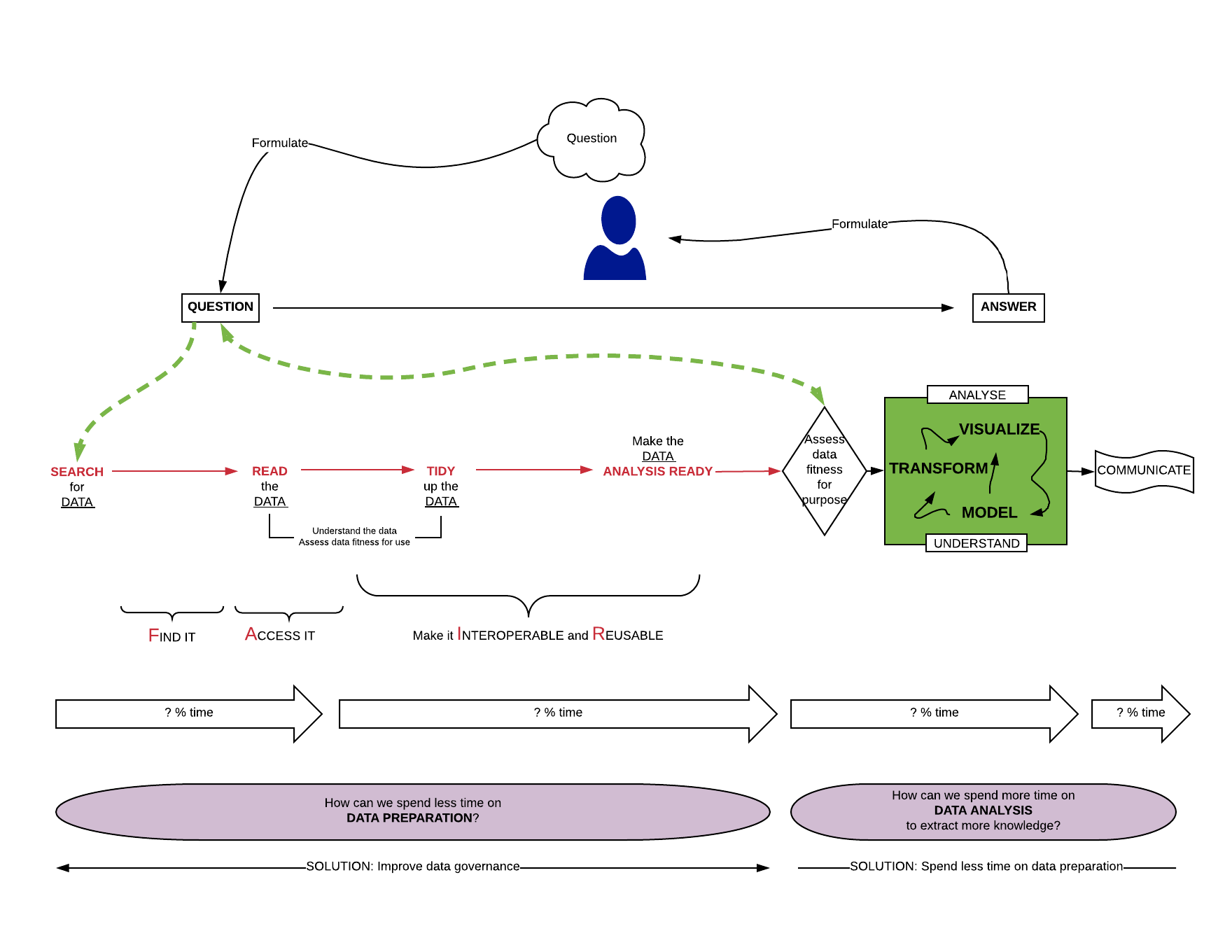
**NEW FIGURE  
Volume 9 Fig 8 – Dataset fragmentation  
Justification**: To accompany Big Data readiness text



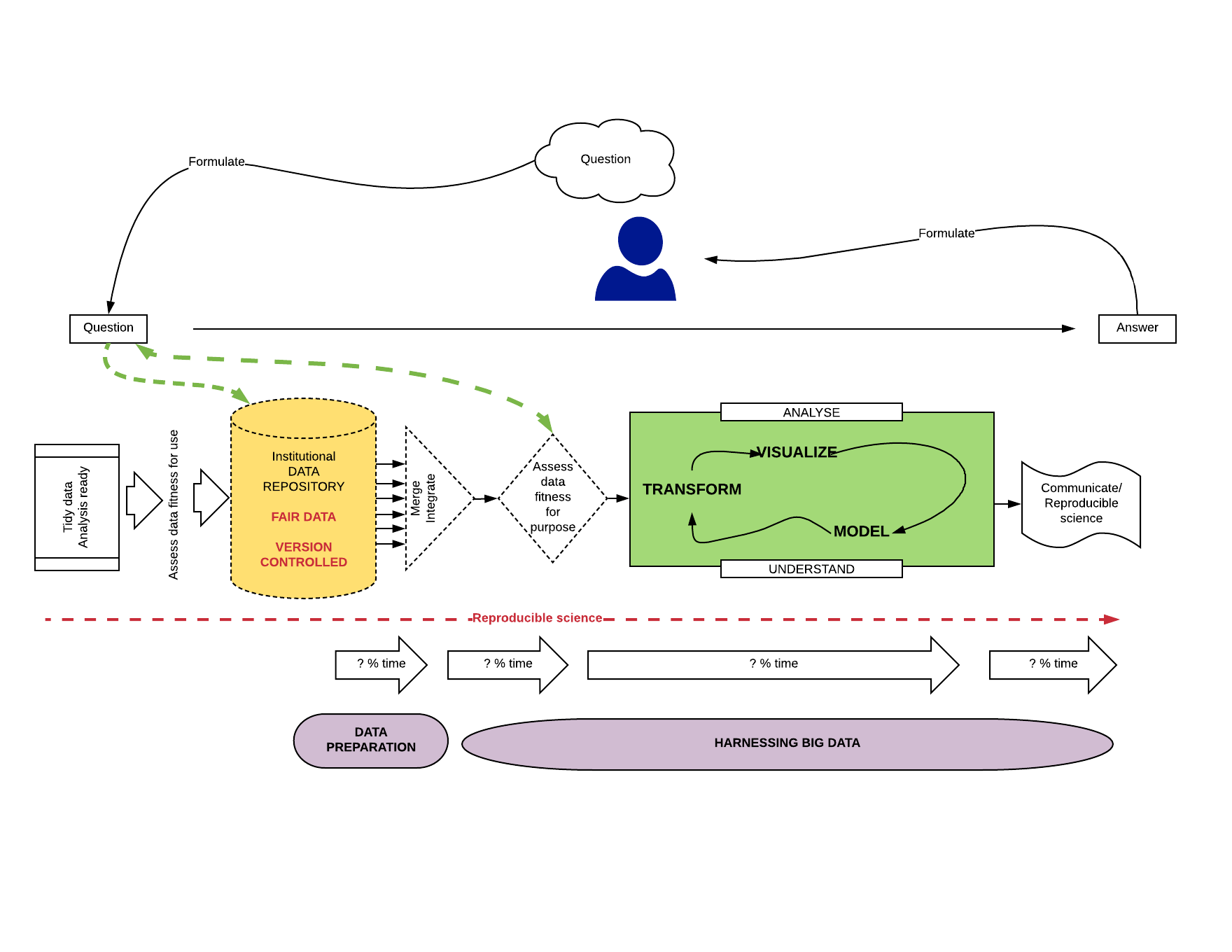
**NEW FIGURE  
Volume 9 Fig 9 – Data preparation  
Justification**: To accompany Big Data readiness text



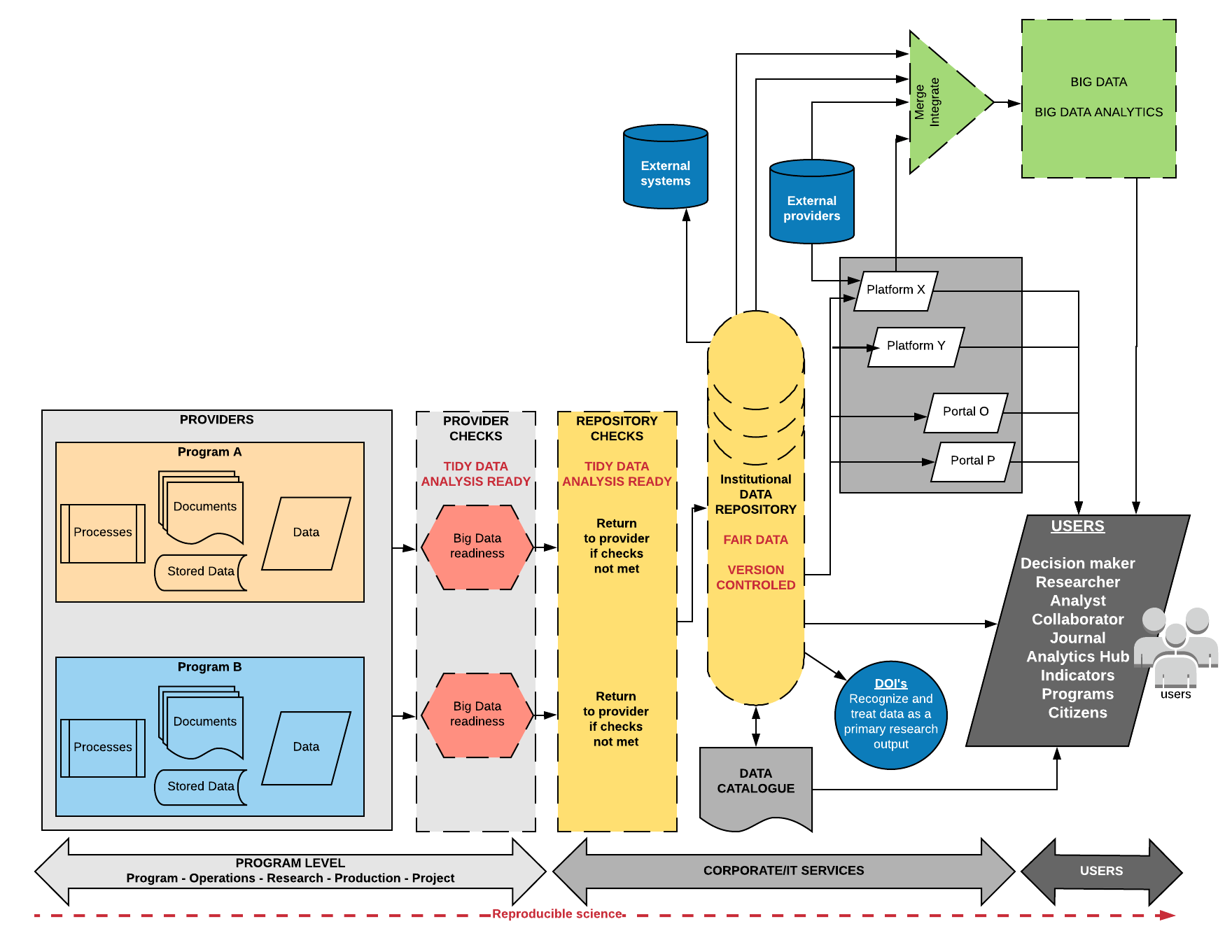
**NEW FIGURE  
Volume 9 Fig 10 – Making data analysis ready  
Justification**: To accompany Big Data readiness text



**NEW FIGURE  
Volume 9 Fig 11 – Harnessing Big Data  
Justification**: To accompany Big Data readiness text



**NEW FIGURE  
Volume 9 Fig 12 – Linear data flows for authoritative data  
Justification**: To accompany Big Data readiness text



**NEW TABLE  
Volume 9 Table 13 – Big Data readiness checklist  
Justification**: To accompany Big Data readiness text

* Checklist questions should be formulated such that the “correct” answer is ‘yes’.
* Allowable answers are ‘*yes*’, ‘*no*’, ‘*I don’t know*’, ‘*not applicable*’.
* Checklists can be used as an auto-evaluation tool.
* Checklist results can be submitted to management for data approval.
* Checklist results can be used by a repository to accept or reject datasets.
* Management can easily merge results received from across an organization.
* Management can quickly scan the results to identify areas in need of improvement.

|  |  |  |  |
| --- | --- | --- | --- |
| line | **MODULES Scientific computing category** | **CHECKLISTS Big Data readiness** | **Does your dataset comply with the items in the checklist?** |
| 1 | Metadata management | Do the metadata include a description of the dataset? | yes |
| 2 | Metadata management | Does the dataset have a persistent identifier? | yes |
| 3 | Metadata management | Do the metadata include a dataset creation date? | yes |
| 4 | Metadata management | Do the metadata include a dataset update date? | yes |
| 5 | Metadata management | Do the metadata include a description of the temporal coverage? | yes |
| 6 | Metadata management | Do the metadata include a description of the geospatial coverage? | yes |
| 7 | Metadata management | Do the metadata identify the creator of the dataset? | yes |
| 8 | Metadata management | Do the metadata identify the contributors to the dataset? | yes |
| 9 | Metadata management | Do the metadata include a link to related publications? | yes |
| 10 | Metadata management | Do the metadata include a link to related data products? | yes |
| 11 | Metadata management | Do the metadata include keywords to improve dataset discoverability? | yes |
| 12 | Metadata management | Are all metadata provided in a machine-readable format? | yes |
| 13 | Metadata management | Are the terms used in the metadata compliant with relevant metadata standards or ontologies? | yes |
| 14 | Metadata management | Do the metadata include a citation that is compliant with JDDCP? | yes |
| 15 | Metadata management | Do the metadata include a description of the methods used for data collection? | yes |
| 16 | Metadata management | If the dataset comes from model output, do the metadata include a description of the model that was used? | yes |
| 17 | Metadata management | Do the metadata include a description of the experimental set-up? | yes |
| 18 | Metadata management | Is this dataset part of a data collection? | yes |
| 19 | Metadata management | Do the metadata include a description of the data collection, if applicable? | yes |
| 20 | Metadata management | Is there a data dictionary that describes the contents, format, and structure of the tables in the data collection, and the relationship between the tables? | yes |
| 21 | Data collection | Was a quality control technique such as" Statistical Process Control" used to ensure that collected data are accurate? | yes |
| 22 | Data collection | If the dataset includes data from a testing or calibration laboratory, was the laboratory method accredited? e.g., ISO/IEC 17025:2017 standard (originally known as ISO/IEC Guide 25). | yes |
| 23 | Data preparation | Were check-digits used on known unique identifiers to ensure valid values? | yes |
| 24 | Data preparation | Were drop-down menus, look-up tables or reference lists used for variables that should have a fixed code set? | yes |
| 25 | Data preparation | Are dates formatted according to the ISO 8601 Standard (e.g., YYYY-MM-DD)? | yes |
| 26 | Data preparation | Are times formatted according to the ISO 8601 Standard (e.g., HH:MM)? | yes |
| 27 | Data preparation | Where the dataset contains measured observations, are the units provided in a separate column? | yes |
| 28 | Data preparation | If the dataset contains latitude/longitude, is the date d a t u m provided? | yes |
| 29 | Data preparation | Are the data files tabular? i.e. There is one rectangular table per file, systematically arranged in rows and columns with the headers (column names) in the 1st row. Every record (row) has the same set of column names. Every column contains the same type of data, and only one type of data. | yes |
| 30 | Data management | Are the raw data available online? | yes |
| 31 | Data management | Are the raw data backed up in more than one location? | yes |
| 32 | Data management | Are all the steps used to process the data recorded and available online? | yes |
| 33 | Data management | Does each record (row) have a unique identifier? | yes |
| 34 | Data management | Have you anticipated the need to use multiple tables? | yes |
| 35 | Data management | Can the tables in a data collection be linked via common fields (columns)? | yes |
| 36 | Data management | Have the data been submitted to a reputable DOI repository? | yes |
| 37 | Data management | Do the files have names that are meaningful to humans? | yes |
| 38 | Data management | Do the variables (column) have names that are meaningful to humans? | yes |
| 39 | Data management | Have the data been deduplicated? | yes |
| 40 | Data management | Are the data FAIR (Findable, Accessible, Interoperable, Re-usable)? | yes |
| 41 | Data management | Was a logical, documented naming convention used for variables (column names)? | yes |
| 42 | Data management | Was a logical, documented naming convention used for file names? | yes |
| 43 | Data management | Were the data documented, "*as-you-go*" rather than at end the end of the process? | yes |
| 44 | Data management | Is a description of the quality control and quality assurance (QA/QC) procedures available online? | yes |
| 45 | Data management | Were measures taken to protect security of data in all holdings and all transmissions through encryption or other techniques? | yes |
| 46 | Data management | Were measures taken to protect against disclosure or theft of confidential information? | yes |
| 47 | Data management | Is a description of the measures taken to protect against disclosure or theft of confidential information available online? | yes |
| 48 | Data management | Were measures taken to ensure a "*single source of truth*" to minimize duplication of information and effort? | yes |
| 49 | Data management | Were standard formats used for names? | yes |
| 50 | Data management | Were standard formats used for civic addresses? | yes |
| 51 | Data management | Are the datasets prepared at the lowest possible level of granularity? (i.e. the data are not summary statistics or aggregated data) | yes |
| 52 | Data management | Are new datasets output at regular, predictable intervals (e.g., the last day of every month, the last day of the year)? | yes |
| 53 | Data management | Is the dataset located in a repository meeting CoreTrustSeal standards? | yes |
| 54 | Data management | Is there a description of the steps performed during data preparation? | yes |
| 55 | Data fitness for use | Are the data tidy? i.e. the data can be read by statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code to extract information to put it in a machine useable form. | yes |
| 56 | Data fitness for use | Are the data analysis ready? | yes |
| 57 | Data fitness for use | Are the data machine readable? | yes |
| 58 | Data fitness for use | Can the data be ingested directly into statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code? | yes |
| 59 | Data fitness for use | Are the data in CSV (i.e. comma separated, or character separated) format? | yes |
| 60 | Data fitness for use | Was a "user-centric" (i.e. the end-user is unknown), rather than a project- or client-centric approach used for data preparation? | yes |
| 61 | Data fitness for use | Can the data be incorporated seamlessly into a Big Data workflow? | yes |
| 62 | Data fitness for use | Are the data files in a non-proprietary format? | yes |
| 63 | Data fitness for use | Are new data appended to existing data files? | yes |
| 64 | Data fitness for use | Did you follow specified data quality assurance practices in the production of these data? | yes |
| 65 | Data fitness for use | Do the metadata include all concepts, definitions and descriptions of all of the variables? | yes |
| 66 | Data fitness for use | Do the metadata include descriptions of methods, procedures and quality assurance practices followed during production of the data? | yes |
| 67 | Data fitness for use | Are the metadata accurate, complete, up to date, and free of contradictions? | yes |
| 68 | Data fitness for use | Are accuracy indicators provided for all of the measured variables? | yes |
| 69 | Data fitness for use | Are there matching variables such as age, sex, address, industry, occupation? | yes |
| 70 | Data fitness for use | Is a description available online of any exceptions or limitations in these data? | yes |
| 71 | Data fitness for use | Do the data meet domain specific standards or requirements? | yes |
| 72 | Data fitness for use | Are the data fit-for-use? | yes |
| 73 | Computer code | Is there a brief explanatory comment at the start of the code? | yes |
| 74 | Computer code | Has the code been decomposed into functions? | yes |
| 75 | Computer code | Has duplication been eliminated? | yes |
| 76 | Computer code | Does the code include well researched libraries or packages to perform needed tasks? | yes |
| 77 | Computer code | Have you tested the libraries or packages before relying on them? | yes |
| 78 | Computer code | Do the functions and variables have meaningful names? | yes |
| 79 | Computer code | Have dependencies and requirements been made explicit? | yes |
| 80 | Computer code | Have you avoided using comment/uncomment for sections of code to control the program's behavior? | yes |
| 81 | Computer code | Have you provided a simple example or test dataset? | yes |
| 82 | Computer code | Has the code been submitted to a reputable DOI-issuing repository? | yes |
| 83 | Computer code | Is an overview of the project available online? | yes |
| 84 | Computer code | Is a shared "to-do" list for the project available online? | yes |
| 85 | Computer code | Is a description of the communication strategy available online? | yes |
| 86 | Computer code | Is there an explicit license? | yes |
| 87 | Computer code | Is the project citable? | yes |
| 88 | Project organization | Is each project in its own directory which is named after the project? | yes |
| 89 | Project organization | Are text documents associated with the project in a documents directory? | yes |
| 90 | Project organization | Are the raw data and metadata in a data directory? | yes |
| 91 | Project organization | Are the files generated during cleanup and analysis in a results directory? | yes |
| 92 | Project organization | Is the project source code in a ‘source’ directory? | yes |
| 93 | Project organization | Are external scripts or compiled programs in a bin directory? | yes |
| 94 | Project organization | Do all filenames reflect their content or function? | yes |
| 95 | Keeping track of changes | Is (almost) everything created by a human being backed up as soon as it is created? | yes |
| 96 | Keeping track of changes | Are changes kept small? | yes |
| 97 | Keeping track of changes | Are changes shared frequently? | yes |
| 98 | Keeping track of changes | Is a checklist created, maintained, and used for saving and sharing changes to the project? | yes |
| 99 | Keeping track of changes | Is each project stored in a folder that is mirrored off the researcher's working machine? | yes |
| 100 | Keeping track of changes | Is there a file called CHANGELOG.txt in the project's docs subfolder? | yes |
| 101 | Keeping track of changes | Is the entire project copied whenever a significant change has been made? | yes |
| 102 | Keeping track of changes | Is a version control system used? | yes |
| 103 | Keeping track of changes | Are changes conveyed to all users in a timely fashion? | yes |
| 104 | Reproducibility | Are the data the result of a reproducible workflow? | yes |
| 105 | Reproducibility | Are all methods documented in detail such that a 3rd party could reproduce the workflow and obtain the same results without needing to consult with the data provider? | yes |
| 106 | Reproducibility | Given the data and information provided, are the data and the limitations of the data completely understandable by a 3rd party without needing to consult with the data provider? | yes |
| 107 | Manuscripts | Are manuscripts written using reference management software? | yes |
| 108 | Manuscripts | Are manuscripts written in a plain text format? | yes |
| 109 | Manuscripts | Are manuscripts deposited in a pre-print repository? | yes |
| 110 | Manuscripts | Are manuscripts submitted to an open source, peer reviewed journal? | yes |
| 111 | Manuscripts | Do manuscripts identify individual authors and co-authors? | yes |
| 112 | Manuscripts | Are manuscripts version controlled? | yes |

**Checklist references**

Broman KW, Woo KH (2017). Data organization in spreadsheets. The American Statistician, 72(1): Special Issue on Data Science <https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989>

Kitzes J (2016). Reproducible workflows. <http://datasci.kitzes.com/lessons/python/reproducible_workflow.html>

Wickham H (2014). Tidy data. Journal of Statistical Software, 59(40), 1-23. <https://www.jstatsoft.org/article/view/v059i10>

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Tea TK (2017). Good enough practices in scientific computing. PLOS Computational Biology | <https://doi.org/10.1371/journal.pcbi.1005510http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745>

1. A pivotal turning point is the release of all data in human readable and machine-readable format. For example, CSV files in tabular form can be understood by humans and can be read by statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code to extract information and put it in a machine useable form. [↑](#footnote-ref-1)