INCENTIVIZING AN AI-ABLE DATA ECOSYSTEM BETWEEN FEDERAL AND NON-FEDERAL ENTITIES

GIL ALTEROVITZ, PHD

BACKGROUND

As part of the "AI and Open Data for Innovation in Health" event and associated sprint, a four-level tiered interlinked incentivization AI-able data Ecosystem framework was established (bronze, silver, gold, diamond) for qualitatively measuring and incentivizing: Data's Choice for industry perspective and AI's Choice for federal. It works by creating a data linkage between data producers and AI/model creators.

On the federal side, the sprint saw agencies are seeking to leverage industry-based tools that themselves used federal data. So, if any agency sees one company claim an accuracy of 99% and another 90% on a particular AI solution (like matching patients to clinical trials), which would be the better solution for acquisition? It would seem the former is more accurate. But, the key is in the underlying type of data used for training and testing, how the model was built with that data, and how the data was applied to get results on the data. In fact, a high accuracy like 99% may actually suggest an over-fitting solution that may not generalize well to other cancer clinical trials beyond that ones used to train/test the model. What was needed was more than metrics.

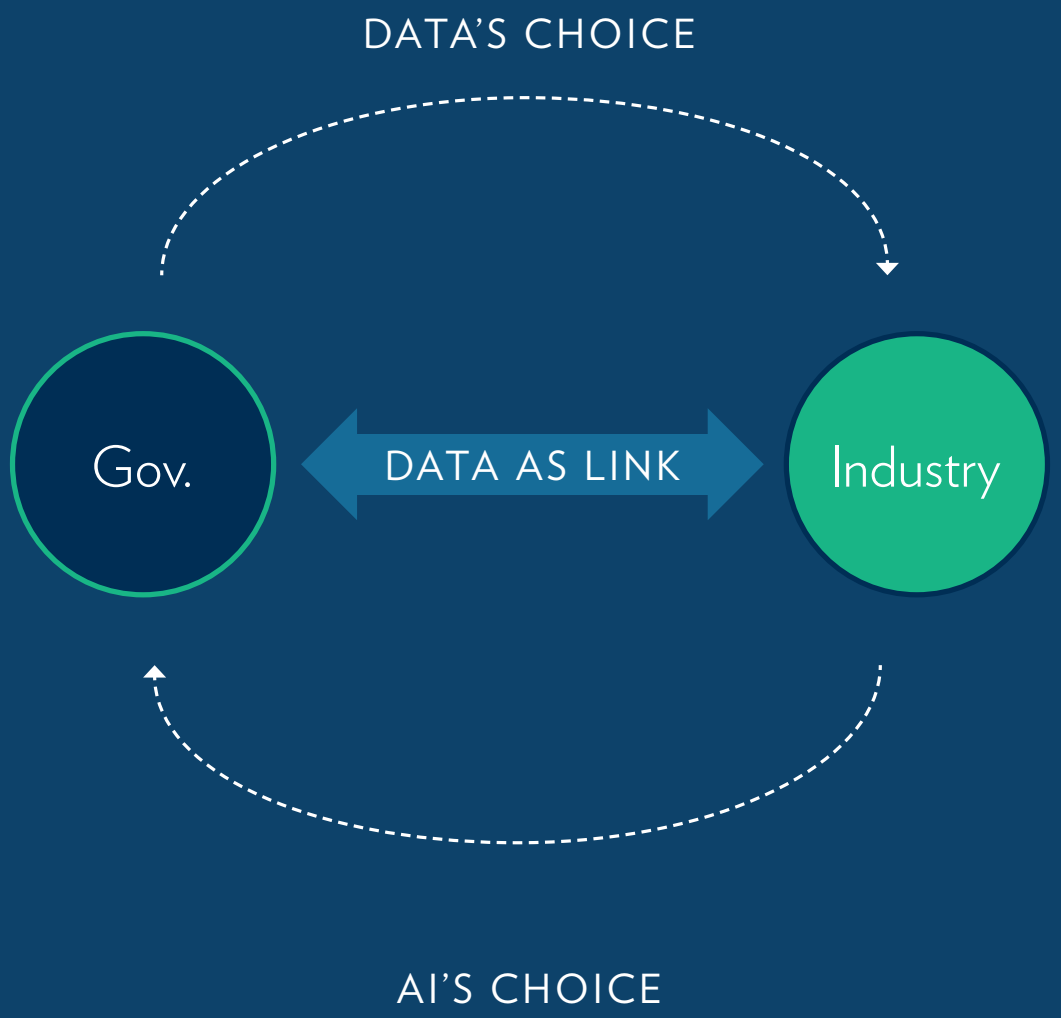


Figure 1. Public-private collaborations via data linkages

INCENTIVIZING AN AI-ABLE DATA ECOSYSTEM BETWEEN FEDERAL AND NON-FEDERAL ENTITIES TEAMS

Thus, the question became: What makes industry AI/data results usable by others, likefor federal agencies, so that they leverage industry tools -– and how could this be potentially measured and incentivized?

Given that the focus of the tech sprint was on creating AI-able data for an AI ecosystem, the project had to reimagine how to generate structured data. To that end, new datasets were created together with NCI data scientists and other federal agency change makers, specifically for the challenge itself to make "AI-able" data as well as enable the tech sprint platform AI ecosystem to serve as an honest broker. Given the new purpose, a novel approach also had to be taken to finding and engaging companies and other players for teams. For instance, this involved specifically seeking out and working with the AI-related technical teams where possible, rather than traditional contacts involved in government interactions and contracting. The end result was the development of a bidirectional AI ecosystem, new industry analysis, and selection of players specifically designed to maximize the platform.

This pioneered a new, nimble approach to create and leverage a data ecosystem that does not require contracts for partnerships, but where we use data as the bidirectional link between government and users of that data across industry, government labs, etc. Thus, public-private links are formed rather than traditional public-private partnerships.

Traditional data ecosystem approaches typically generate and release data via "broadcast" model, mostly one-way communication. It may have some limited input (like radio station caller), but priority is often on timing and/or meeting internal (government) requirements.

The new approach is similar to a group call. It has an open 2-way communication for quick, iterative feedback. It enables data (and perhaps just as importantly, the final AI-based results) to be more usable to all parties. The other key part of the approach is a novel incentivization framework that leads the data link to yield useful results for both sides:

1) government generating data and being able to better use industry AI-based results on that data and

2) industry getting useful data for training AI as well as gaining a better understanding of government use cases.

One of the lessons learned was that the underlying data and tools as well as partnerships and incentives need to be designed for AI from the beginning in order to prevent having to be redone later.

Data's Choice and AI's Choice

From the industry perspective, information that enabled them to judge data utility was paramount. Data's Choice provides for the needs/perspective of data users and AI's Choice for users of AI models/results. For data users, information that enables them to judge data utility is paramooice thus asks:

What makes (e.g. federal) data also useful to nonfederal entities to build tools, and how can this be potentially measured and incentivized? The four levels are designed so that: Data can be analyzed in more efficient manner by industry, Data is de-risked for industry to evaluate overall quality, Data useful for specific industry use cases, and AI-able data to be created that is useful for training/testing AI models in industry.

This is done via levels from bronze to diamond including: machine readable, decimation including provenance and data quality metrics, stakeholder feedback and iterative data release, and AI-centric dataset and tool design. "Data's Choice" is sort of like "People's Choice" – except that the data is the key to selecting the level received.

INCENTIVIZING AN AI-ABLE DATA ECOSYSTEM BETWEEN FEDERAL AND NON-FEDERAL ENTITIES TEAMS

On the other side, AI's Choice asks: What makes AI/data results usable by others (from federalto non-federal), and how could this be potentially measured and incentivized? Here the levels enable the promotion of: leveraging open data, promoting transparency and reproducibility, trust of AI solution through testing, demonstrating usability of AI solution by other parties, and enhancing of the AI ecosystem and re-use of usable solutions. This is done via levels from bronze to diamond including: federal dataset application, test data analysis, independent use/ validation, and giving back to the AI ecosystem/ community.

Details

The "Data's Choice" medal levels are for measuring usefulness, as perceived by industry, of open data generated on the federal side. This was then applied to the federal datasets created during the sprint in order to judge the current state and potential iterate.

For example, for bronze, the data generated should be "machine readable." Having data in JSON, XML or similar format that has elements in easily parsable format enables industry and others to quickly process and deploy the information. Each medal level implicitly builds on previous one.

For silver, documentation is key. Documentation on provenance and data quality metrics can provide a means for industry to evaluate whether or not that dataset is useful for their application and what types of quality control/filtering processes may be needed to handle that data, if they decide to invest in using it. At the gold level, federal data would have gone through stakeholder feedback (e.g. by industry/ users) and iteratively data release to capture ongoing feedback. This not only ensures the data is useful for practical use cases, but also reduces ongoing maintenance costs. At the final level, diamond, datasets are constructed with AI-centric thinking from the beginning. To do so, elements linked to terminologies/ ontologies, as applicable.

Training/ testing datasets designed for AI with testing datasets tests until model is trained. Agencies may also build/leverage tools to serve as honest broker for testing.

The key was not to be have explicit competition between companies through this framework. To that end, none of the levels required achieving certain numeric results. Rather, the framework inspired companies to compete internally (within the company) to increase transparency as well as usefulness of results and obtain the diamond level.

DATA'S CHOICE AND AI'S CHOICE

The 21st Century Cures Act established priorities for initiatives across Federal agencies to reduce roadblocks and enable work toward new therapeutics. One of the chief obstacles to the timely completion of clinical trials is recruitment of participants. To improve the precision of searching for experimental therapeutics, whether they be in clinical trials or under the "Right to Try" Act , this work tested approaches for structuring eligibility criteria to make it be easier to find relevant experimental therapeutics (and clinical trials, where applicable) without having to a read through a large number of trial protocol texts manually.

As part of our sprint, we established a four-level tiered system (bronze, silver, gold, diamond) for qualitatively measuring and incentivizing: Data's Choice for industry perspective and AI's Choice for federal. From the industry perspective, information that enabled them to judge data utility was paramount. As inspired by People's choice awards, it the data and AI approaches themselves selecting the recipient of the recognition and applied to the Health Tech Sprint.

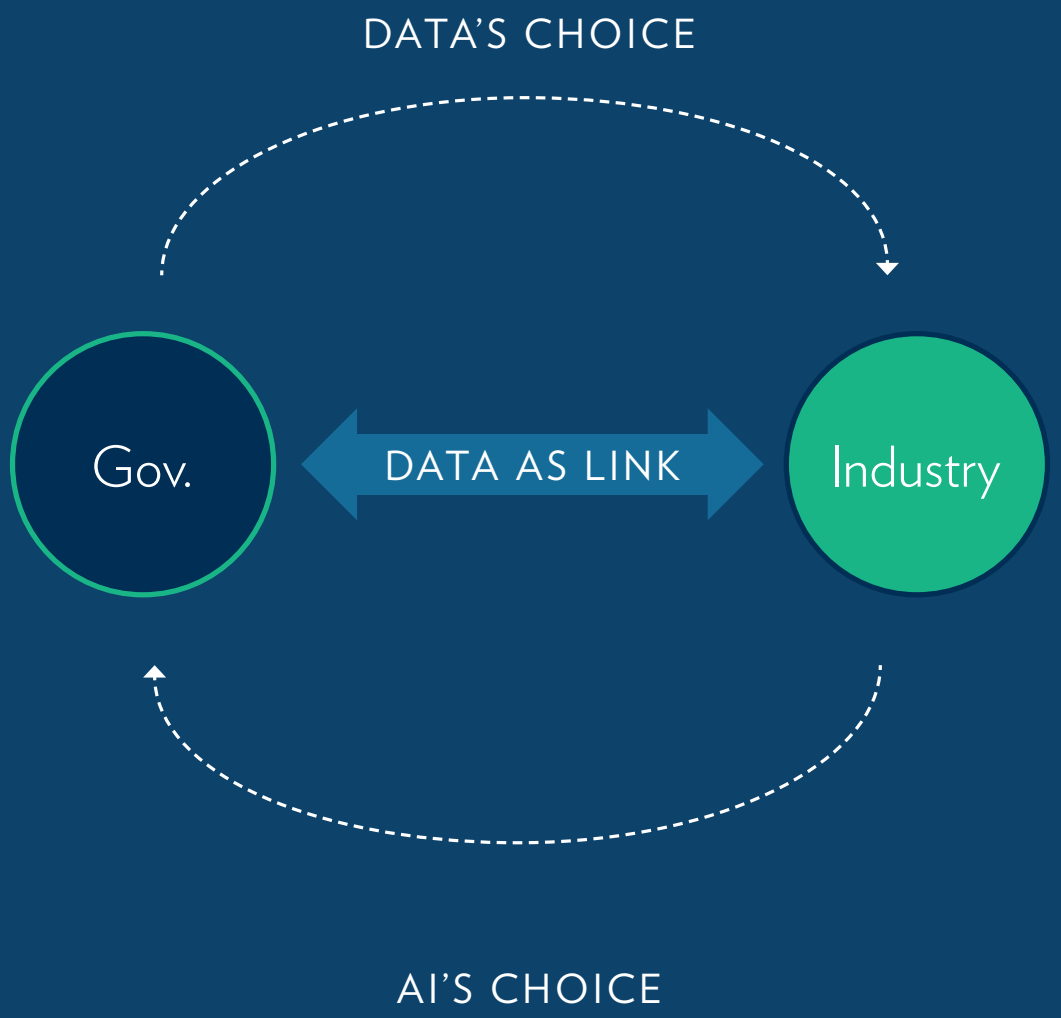


Figure 1. Data as link between Government and other sectors like Industry.

DATA'S CHOICE AND AI'S CHOICE

The "Data's Choice" medal levels are for measuring usefulness, as perceived by industry, of open data generated on the federal side. This was then applied to the federal datasets created during the sprint in order to judge the current state and potential iterate.

For example, for bronze, the data generated should be "machine readable." Having data in JSON, XML or similar format that has elements in easily parsable format enables industry and others to quickly process and deploy the information. Each medal level implicitly builds on previous one.

For silver, documentation is key. Documentation on provenance and data quality metrics can provide a means for industry to evaluate whether or not that dataset is useful for their application and what types of quality control/filtering processes may be needed to handle that data, if they decide to invest in using it. At the gold level, federal data would have gone through stakeholder feedback (e.g. by industry/users) and iteratively data release to capture ongoing feedback. This not only ensures the data is useful for practical use cases, but also reduces ongoing maintenance costs. At the final level, diamond, datasets are constructed with AIcentric thinking from the beginning.

To do so, elements linked to terminologies/ ontologies, as applicable. Training/ testing datasets designed for AI with testing datasets tests until model is trained. Agencies may also build/ leverage tools to serve as honest broker for testing.

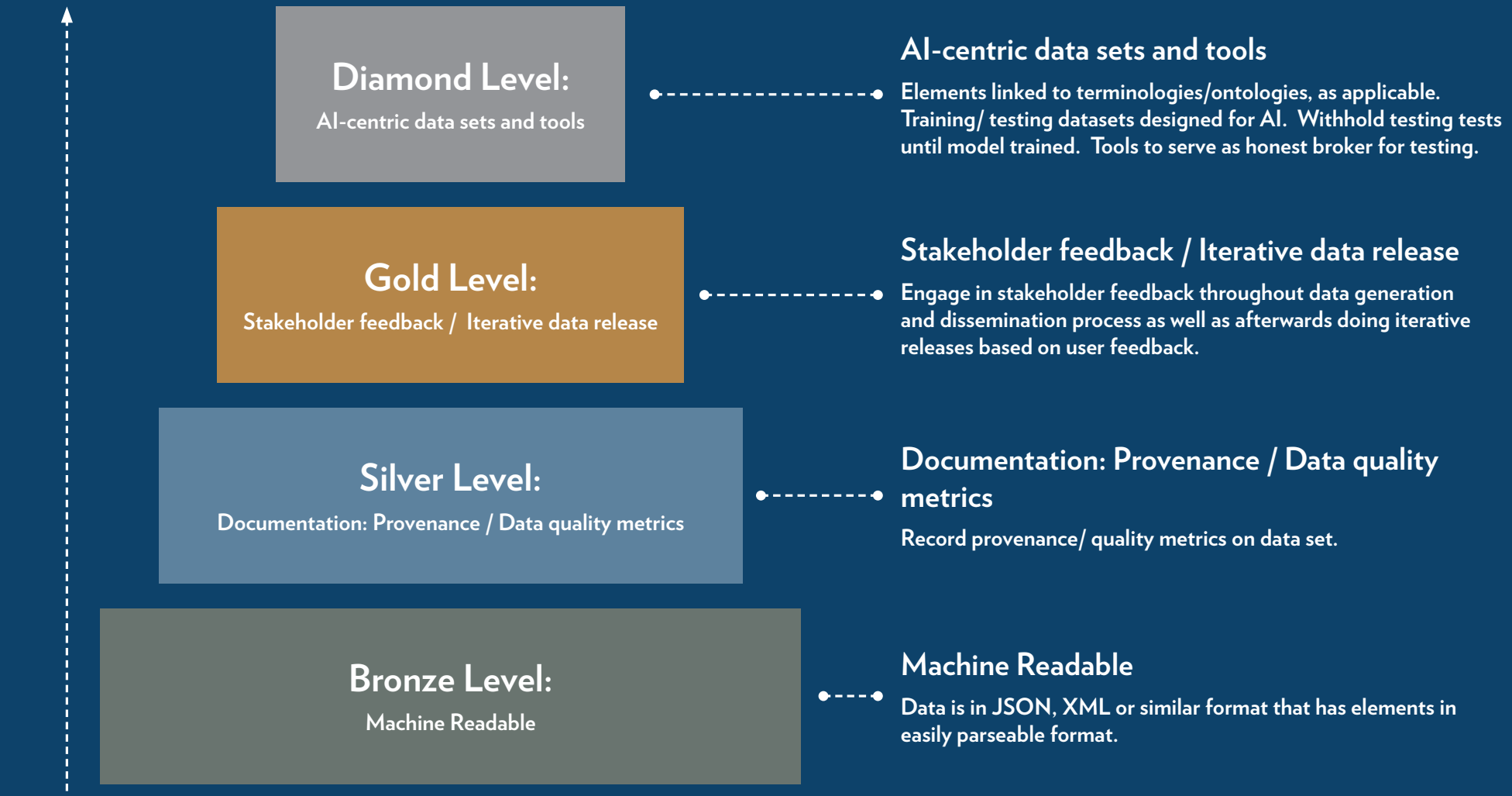
On the other side, "AI's Choice" incentivizing the voluntary release of information on the underlying AI used incrementally, to facilitate agency trust of and evaluate solutions of industry-based AI solutions. The bronze level involves using an AI Ecosystem of linked data and specifying which datasets (e.g. federal) were used in the application. In the Health Tech Sprint, bronze meant that the company /organization used the provided AI ecosystem datasets (e.g. participant data, eligibility criteria data, and health professional matches), among potentially other federal or other data, for new tools.

This lets federal agencies (and others) know if that tool's AI was designed based on the type of data/ use cases that they are interested in.

For silver, the company would use the tool should predict and share results based on independently provided (e.g. federal) test datasets not seen before. In the Health Tech Sprint that meant that teams were asked to agree to that any tool models first be fixed, before test input was shared for AI based prediction. This lets agencies and others see if the AI tool is generalizable to a new test dataset and is fit for purpose. Government agencies (or independent third parties) may also play a role in the future as an honest broker for AI tool data sets and testing to ensure tools are exposed to testing data/predictions evaluated only after the training process is completed.

For gold, there is independent use and validation. In the Health Tech Sprint , patient advocate provided critical feedback and evaluation of the tools. This provided input for iteration of tooling by the companies. It can help build trust with agencies to see independent use and validation. Finally, the diamond level involves giving back to the AI Ecosystem and community. This involves contributing back in some way that helps others in AI ecosystem. In the TOP Health tech sprint, this was designed to be done by, for example, through industry/organizations giving data or open source code, making commitment to hiring workers in AI, and/or patient journey matching actual patients to actual new trial that they enrol in, etc.

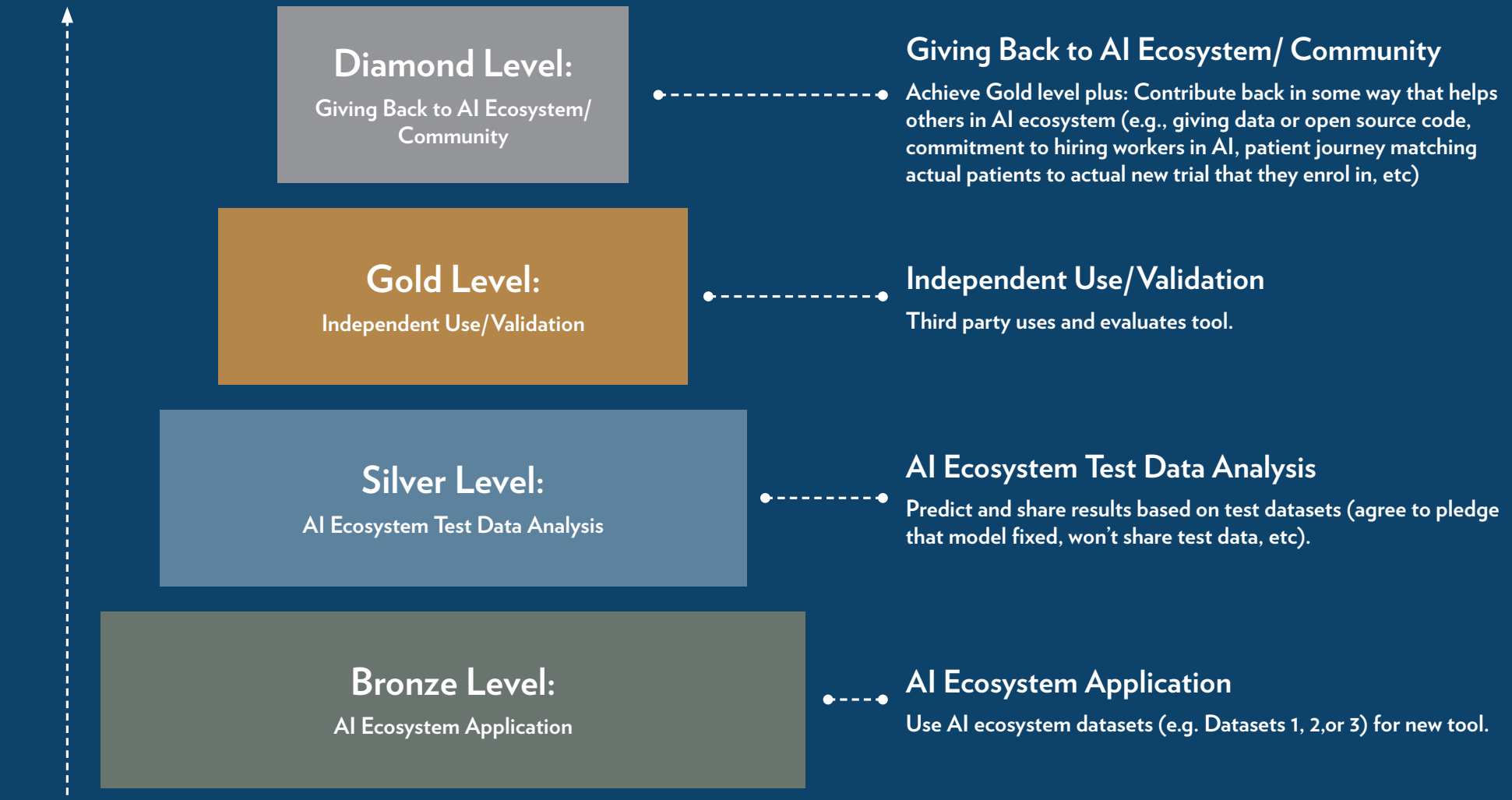
DATA'S CHOICE LEVELS



* Each level implicitly builds on previous one.

Figure 2. Data's Choice Levels: What makes federal data also useful to industry to build tools, and how can this be potentially measured and incentivized?

AI'S CHOICE LEVELS



* Each level implicitly builds on previous one.

Figure 3. AI's Choice Levels: What makes industry AI/data results usable by others, like for federal agencies, so that they leverage industry tools -- and how could this be potentially measured and incentivized?

ADDENDUM: BARRIERS IDENTIFIED OVERCOMING AI BARRIERS VIA SPRINTS

Background

Enclosed are barriers identified during the Health Tech Sprint. As background, the sprint has worked with 11 teams delivering digital tools — built with open federal data and emerging technologies like Artificial Intelligence (AI) — to improve clinical trials, experimental therapies, and data-driven solutions for complex challenges from cancer to Lyme and tick-borne diseases. The teams ranged from two international teams (i.e. Microsoft Healthcare and Philips Research) to one originally at a medical center that officially spun out into a new company specifically based on the AI clinical trial search tool developed during the sprint. It included physicians and patient advocates as well as data stewards and experts in the relevant domain areas from within government. The AIable Data Ecosystem pilot seeks to address these barriers via sprint process and framework. For more information, please see corresponding white paper on Data's Choice and AI's Choice from which the 4-tiered levels below is derived.

Item	Data's Choice	AI's Choice
Lack of machine readability/parsers for dataset/custom format	•	
Ease of access/use of datasets can be challenging X	•	
It is hard to trust/compare models/reproduce results without some transparency on scope/data used.		•
Data is useful for non-federal entitie s (and not sensiti ve), but is not made available for download/access.	•	
Data quality is unpredictable. Few datasets data quality metrics (either qualitative or quantitati ve).	•	
Capturing data's provenance is important. It is used by model builders to understand the potential scope of use of the data for different AI models/applicati ons.	•	
It is not enough to have structured data, it is also key that structured data is complete and properly aligns to the schema.	•	
Poor documentation can lead to datasets/models not being used.	•	•
Matching ID's across datasets can be challenging	•	
Getting/understanding metadata can be challenging.	•	
Getting labeled data can be challenging.	•	
It is hard to connect with original data curators/model builders to understand use/generalizability of dataset	•	
Different systems have dif erent means for access in in terms of authent i cation/authorization- so can be tricky to find/coordinate the accounts/informati on.	•	
It is difficult to trust/compare model results without knowing that models were tested against the same test sets.		•
It is difficult to trust/compare model results without knowing that models were fixed (not altered) after training and before testing.		•
Data designed for archive/release,rather than for use by external parties	•	
It is important to engage end-user of product to know what best data to capture is for doing AI/formats needed, etc.	•	•

Bronze | Silver | Gold | Diamond

Item	Data's Choice	AI's Choice
Criteria prioritization should be impacted also by domain knowledge – domain experts shared their experience and knowledge about the current recruiting approach	•	•
Using layperson'sterms for consumer facing tools is key, since consumers are not necessarily familiar with special language or specialized AI-enabling formats.	•	•
Lack of external users / evaluators of AI models/tools. Evaluation/validation of AI model/tool by third parties can give additional confidence when evaluating model.		•
Data set is obsolete	•	
Data set/form at has not kept up with current needs	•	
Model is superceded by others		•
Model works only suitable for certain users/conditions		•
Annotating and structuring by humans is hard–domain experts/annotators have trouble translating text into logical structured data, which can later be used to build machine learning	•	•
There is often a trade off between granular, technical terms/definitions ideal for AI and specialist or consumer terms/definitions. For products to be useable by different audiences	•	•
There is a lack of testing datasets that can be used to test models after training. Data often all released at once-so hard to have independent set for testing	•	•
There is no 'honest broker' to determine if arbitrarily defined 'testing data' is appropriate, making accuracy metrics hard to evaluate as meaningful.	•	•
There is no 'honest broker' to determine if model is re-trained after 'testing data' is seen, thus making accuracy metrics hard to evaluate as meaningful.	•	•
Data is often not formatted with standardized ontologies/coding systems (so need to do lossy mappings)	•	
For large datasets needed for AI, they can be so large that the computer needs to be brought to the data, rather than downloaded, or accessed via API. This brings access, security, cost	•	•
Dataset are often not generalizable - e.g. only applicable to limited scope/use case or location/time period	•	
Current data is generally structured in away that is suitable for learning associations/ correlations, rather than causation. AI models could use time/invention relationships information encoded	•	
Versioning can be an issue(e.g. use data with coding system in one version vs different dataset with a different version number)	•	•
Models not designed with re-use in mind		•