



ARTIFICIAL INTELLIGENCE AND OPEN DATA INNOVATION FOR HEALTH

A 2019 CASE STUDY BY THE PRESIDENTIAL INNOVATION
FELLOWSHIP PROGRAM IN COLLABORATION WITH THE
US DEPARTMENT OF HEALTH AND HUMAN SERVICES

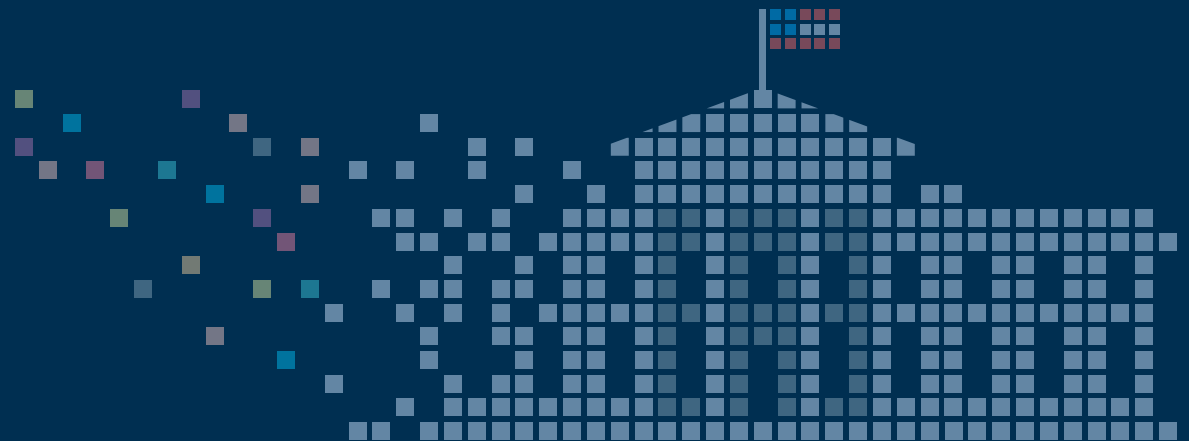
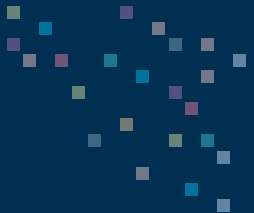




TABLE OF CONTENT

- 
- I. INTRODUCTION
 - II. ARTIFICIAL INTELLIGENCE (AI) ECOSYSTEM TEAMS
 - III. CAPTURING THE PATIENT PERSPECTIVE IN THE
HEALTH TECH SPRINT
 - IV. SHOWCASING RESULTS AT THE WHITE HOUSE
 - V. AI AND OPEN DATA INNOVATION FOR HEALTH AT
THE WHITE HOUSE

HARNESSING OPEN DATA AND AI FOR HEALTH

14
WEEKS

HEALTH TECH SPRINT
BEGAN IN OCTOBER 2018

10
TEAMS

DEVELOPED AND DELIVERED
DIGITAL TOOLS TO ADDRESS
HEALTH CHALLENGES

11
PRODUCTS AND
PROTOTYPES

RESULTED IN DATA-DRIVEN
DIGITAL TOOLS TO BENEFIT
THE AMERICAN PUBLIC

CHALLENGE #1:

ARTIFICIAL INTELLIGENCE (AI) TEAMS

7

TEAMS WORKED TO IMPROVE
HOW PATIENTS ARE MATCHED
TO CLINICAL TRIALS AND
EXPERIMENTAL THERAPIES

The 21st Century Cures Act of 2016 and the
“Right to Try” Act of 2017 helped to inspire and
frame the two challenges.

CHALLENGE #2:

LYME AND TICK-BORNE DISEASE TEAMS

4

TEAMS DEVELOPED CAPABILITIES
TO SUPPORT DATA-DRIVEN
DECISIONS FOR THE PREVENTION,
EDUCATION, AND SCIENCE
TO IMPROVE PUBLIC HEALTH
OUTCOMES RELATED TO
TICK-BORNE DISEASES

This work advanced priorities recently
identified in the American AI Initiative,
established by Executive Order in February
2019.

Our Health Tech Sprint, as known as TOP
Health, in collaboration with the U.S.
Department of Health and Human Services
Office of the Chief Technology Officer and
Presidential Innovation Fellows was modeled
in part after The Opportunity Project
(TOP) at U.S. Department of Commerce. It
was co-led by Presidential Innovation Fellow,
Dr. Gil Alterovitz, and HHS Innovator-in-
Residence, Dr. Kristen Honey.

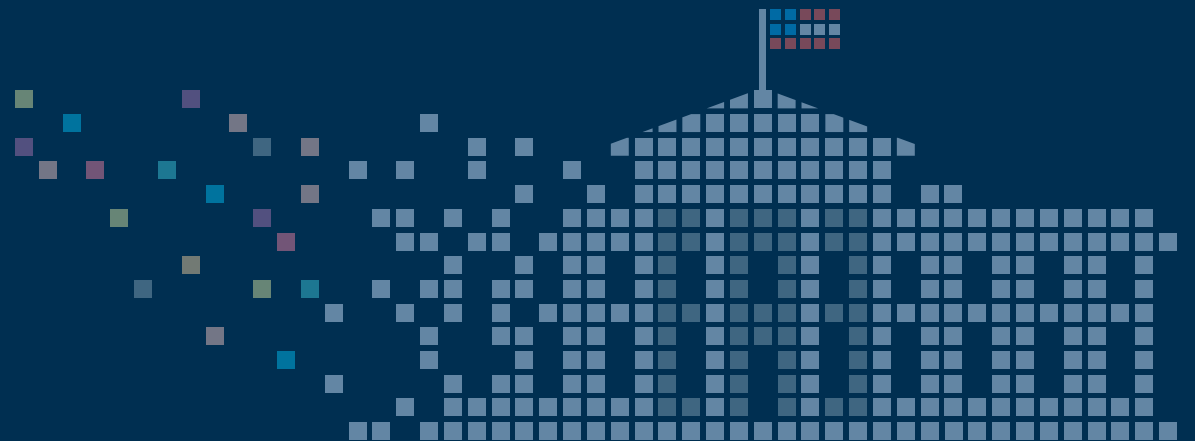
THE HEALTH TECH SPRINT WAS LED BY

Gil Alterovitz,
Presidential Innovation Fellow
at the U.S. General Services
Administration

Kristen Honey,
Innovator in Residence at the
U.S. Department of Human
and Health Services

Federal agencies provided curated datasets on Healthdata.gov and
Data.gov with access to the data stewards

I. INTRODUCTION



HOW A HEALTH TECH SPRINT INSPIRED AN ARTIFICIAL INTELLIGENCE (AI) AND NEW DIGITAL TOOLS

GIL ALTEROVITZ, DAVAR ARDALAN, KRISTEN HONEY

This report highlights a novel sprint and incentivization framework to foster an ecosystem for designing and leveraging data for Artificial Intelligence (AI). The PIF led initiative drew on resources across over a dozen federal agencies and departments. Part of this work was executed as part of the Health Tech Sprint track co-led with the HHS Office of the Chief Technology Officer (CTO). The results of this work were presented/recognized at:

- Artificial Intelligence and Open Data for Health with TOP Health Showcase at the White House Eisenhower Executive Office Building on February 28, 2019, Washington DC.
- AI ecosystem feedback session at the Department of Veterans Affairs, February 28, 2019, Washington DC.
- The Opportunity Project (TOP) Demo Day at the U.S. Census Bureau on March 1, 2019, Suitland, MD.
- “Ecosystem for AI-able Data for Citizen Services and Applications” submitted and selected for funding as part of 10x program via Technology Transformation Service of GSA, March 6, 2019, Washington, DC.
- The National Coordinating Office’s Artificial Intelligence R&D Interagency Working Group, March 2019, Washington, DC. A subset of this group has been writing the updated 2019 National Artificial Intelligence R&D Strategic Plan.
- Newly launched White House site AI.gov cites Health Tech Sprint as example of “AI for American Industry,” March 20, 2019, Washington, DC
- “Getting government ready for AI” Fed100 Award for PIF leading this work specifically cites AI-able data/sprint, March 28, 2019, Washington, DC.
- White House Office of Science Technology and Policy (OSTP) and Office of Management and Budget (OMB) Data/Model Working Group and Machine Learning and Artificial Intelligence Subcommittee of the National Science and Technology Council, April 2019, Washington, DC.
- White House Tech Day at the U.S. Department of Labor as part of the “Industries of the Future” on May 16, 2019, Washington DC.
- U.S. federal agency CIO Council meeting, May 23, 2019, Washington DC.
- “Smoky Mountains Computational Sciences and Engineering Conference Data Challenge,” which will feature DOE/VA/PIF challenge around: May 15- July 31, 2019, virtually and August 23-27, 2019, Kingsport, TN.
- Cited as part of new “Public-Private Partnerships” strategy in White House press briefing and in “The National Artificial Intelligence R&D Strategic Plan” publication, June 20-21, 2019, Washington DC.
- Launched AI-able Data Ecosystem Sprint with Department of Veterans Affairs/PIF as leads, July 2019, Washington DC.



HOW A HEALTH TECH SPRINT INSPIRED AN ARTIFICIAL INTELLIGENCE (AI) AND NEW DIGITAL TOOLS

GIL ALTEROVITZ, DAVAR ARDALAN, KRISTEN HONEY
HEALTH TECH SPRINT

When beginning the 14-week Health Tech Sprint in October 2018, it was hard to realize the profound lessons that would be learned in just a few months. Together with federal agencies and private sector organizations, this work demonstrated the power of applying artificial intelligence to open federal data.

Through this collaborative process, the Health Tech Sprint showed that federal data can be turned into products for real-world health applications with the potential to help millions of Americans have a better life.

The Health Tech Sprint teams developed and delivered digital tools to address challenges. One challenging area is creating data suitable for AI, or AI-able data. As part of the AI-able Data Ecosystem pilot, seven teams created digital tools that help in finding experimental therapies for patients, and vice versa. In a second challenge, four teams developed capabilities to support data-driven decisions for the prevention, education, and science for improving public health outcomes related to tick-borne diseases. This journey report focuses on the AI ecosystem.

As the team's journey mapped this sprint, it asked federal agencies and patient advocates who were part of the AI ecosystem pilot to share their reactions and insights on their collaborations. Many shared that this sprint allowed for an open exchange of ideas and the opportunity for further collaboration between the U.S. government and private sector companies.

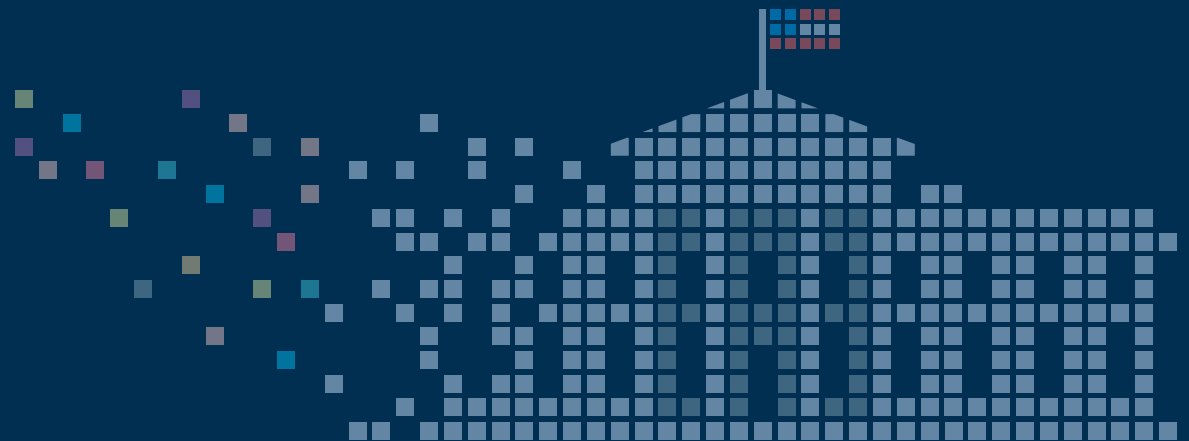
Our Health Tech Sprint, as known as TOP Health, in collaboration with the U.S. Department of Health and Human Services Office of the Chief Technology Officer and Presidential Innovation Fellows was modeled in part after The Opportunity Project (TOP) at U.S. Department of Commerce. It was co-led by Presidential Innovation Fellow, Dr. Gil Alterovitz, and HHS Innovator-in-Residence, Dr. Kristen Honey.

The
Opportunity
Project

II.

ARTIFICIAL INTELLIGENCE (AI)

EMPOWERING PATIENTS WITH THE AI ECOSYSTEM, DATA UTILITY, AND CHOICE



CHALLENGE #1: ARTIFICIAL INTELLIGENCE (AI) TEAMS

Teams from around the globe unveiled their TOP Health results on January 16th
The following digital tools were produced as part of Challenge #1:

Philips Research Trial Explorer

In the Netherlands

Find the most suitable trials for each patient: We derive computable representations of trial criteria and match to harmonized patient data.

Microsoft Clinical Trial Matching Bot

In Israel

Microsoft Healthcare built technology that aims to democratize clinical trials matching by helping patients and doctors find suitable trials quickly and easily, through Conversational AI, advanced Machine Reading on clinical trials eligibility criteria, NLP and smart qualification.

Oak Ridge National Laboratory ORNL SmartClinicalTrials

In Oak Ridge, TN

The ORNL SmartClinicalTrials team is harnessing the power of artificial intelligence (AI) to build a large-scale knowledge network using deep learning and exascale graph analytics. Our scalable AI approach enables discoveries that better match cancer patients with clinical trials.

Oracle Oracle TOP Health Team

In Washington, DC

Oracle built a patient-centric solution for clinical trial matching. The solution matches patients battling cancer with clinical trial programs leveraging modern technologies like artificial intelligence to provide a more simplified and personalized experience for determining treatment.

TrialX iConnect

In New York, NY

iConnect was designed to connect patients with clinical trials and investigative sites. Used at leading organizations such as the University of Pennsylvania and CenterWatch, iConnect is currently being implemented for top clinical trial sponsors in partnership with WCG. Forthcoming releases will include the work done in the TOP Health challenge.

Flatiron Health Flatiron Trials

In New York, NY

Integration of clinical research into everyday cancer care is core to Flatiron Health's mission. Participating in TOP Health helped us accelerate building a framework that improves patient trial matching and gives oncology practices and their patients' access to cutting-edge treatments.

Rush University 3D Innovation Program ACCESS

(Spun out of Rush Medical) In Chicago, IL

Access the medicine of tomorrow, today. If you can find a home or book a vacation online, you should be able to find and share clinical trials just as easily with ACCESS at www.A.CCESS.io

CHALLENGE #1: ARTIFICIAL INTELLIGENCE (AI) TEAMS



Teams leveraged datasets from multiple Federal sources to build new AI tools.

By January 2019, seven Health Tech Sprint teams had delivered digital tools — all built with Federal data and leveraging emerging technology like artificial intelligence (AI). Teams transformed federal data resources into digital products, apps, and inventions to improve clinical trials and data-driven solutions for complex challenges in the specific area of linking patients to experimental therapeutics and clinical trials.

From the onset of the AI ecosystem sprint, Presidential Innovation Fellow, Dr. Gil Alterovitz engaged with Federal and industry thought leaders who emphasized that the key to fostering an Alcentric approach was quality data. To create value — especially with AI data must be in an accessible and in a usable format.

This sprint was unique in a number of aspects. In this sprint, teams (generally non-Federal) gave input on federal data so that it could iteratively be changed in formatting, etc and re-released to suit the use cases/ needs of the participants. This helped lead to Data's Choice, an approach for quantifying usefulness of data. Furthermore, federal entities provided feedback on usefulness of AI results produced and helped define AI's Choice, an approach for quantifying usefulness of AI/model results. See "" in Addendum for more

In a usual sprint, all of the data used is generally open data already available, or released as soon as is possible when unlocked. The Health Tech Sprint was different because it had an AI ecosystem that was designed specifically for AI. Developing accurate AI tools necessitates building a model with one set of data, which acts like its homework, and then testing it on a different one that it has not seen before, which serves as a final exam. The AI ecosystem specifically did not use an open dataset at the beginning of the sprint. Giving all the data out at the beginning would be like giving a final exam with all the answers. So, while part of the dataset developed was released quickly to the companies for training, just as importantly, part was held back by an honest broker for testing. In this case, the honest broker was the organizing sprint personnel, but it could be federal agencies and/or third party industry partners in the future.

Dr. Alterovitz worked across agencies seeking to leverage existing industry-based tools that incorporated Federal data. Health Tech Sprint results included new insight on how sharing Federal data with the private sector could be incentivized and measured. Government agencies were interested in understanding and customizing datasets for various use cases, but were primarily focusing on internal government use. The lightweight collaboration through Health Tech Sprint strengthened the intersection of government and industry.

On the industry side, the private sector was looking to minimize risk when choosing which datasets to leverage. There is a sea of Federal datasets of various quality and usability. Organizing data for a particular business case has costs, both in terms of personnel time to process and map data for a particular use. Through the Health Tech Sprint process, multiple high-value datasets were unlocked from over a dozen collaborating agencies/ departments. For example, the National Cancer Institute developed three datasets specifically for the Health Tech Spring, including: Subset of eligibility criteria translated into machine readable code; records based on callers to the NCI's Cancer Information Service and enhanced with synthetic data, translated into machine readable code; and participant records matched against 50 clinical trials for which the eligibility criteria as well as the participant data had been previously translated into machine readable code.

CHALLENGE #1: ARTIFICIAL INTELLIGENCE (AI) TEAMS

The last data set, produced by oncology professionals, then served as a comparison dataset for the match identified through the application of artificial intelligence.

To inspire prospective inventors, including those with applications in the AI ecosystem and experimental therapeutics space, the U.S. Patent and Trademark Office (USPTO), for instance, assembled a large set of invention topics. USPTO released this data as part of Health Tech Sprint.

The Department of Veterans Affairs (VA) took a different approach. They provided access to de-identified cancer patients for use in matching to trials/therapeutics and analytical tools around AI and machine learning.

These approaches can facilitate the ecosystem but only one which is as strong as its weakest link which is data. The question is whether we can drive a result that is fit for purpose as defined by key stakeholders within the constraints and limitations.”

This first-ever TOP Health tech sprint advanced priorities recently identified in the American AI Initiative, established by Executive Order in

Feb. 2019. It also illustrated real-world value as government data was unlocked for public use as machine-readable “open data,” per the Jan. 2019 signed Foundations for Evidence-based Policymaking Act.

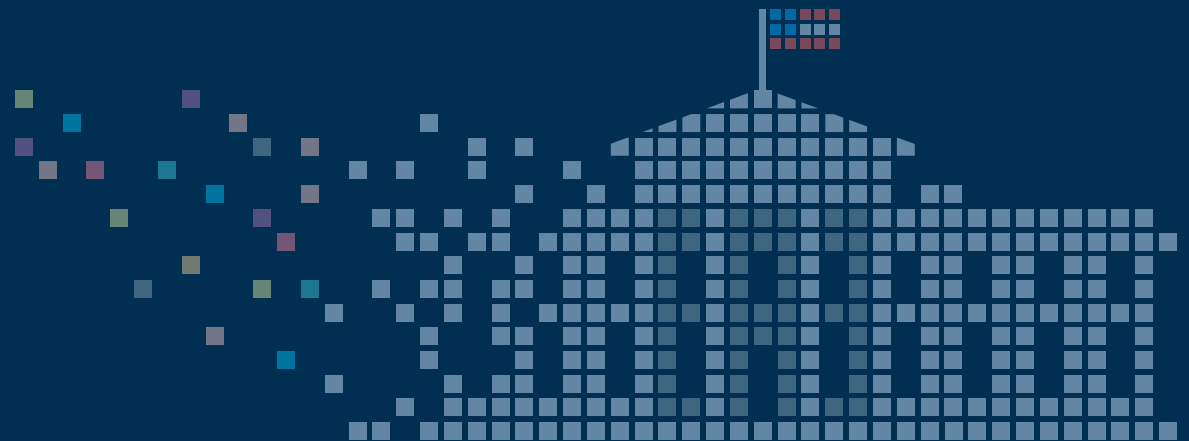


Teams gathered at the White House to present results and discuss next steps. Pictured: Gil Alterovitz, PIF, discussing results with teams.



Microsoft Healthcare team in front of the White House after presenting their work.

III. CAPTURING THE PATIENT
PERSPECTIVE IN THE
HEALTH TECH SPRINT



CHALLENGE #1: CAPTURING THE PATIENT PERSPECTIVE IN THE HEALTH TECH



Stephen Aldrich, ex-founder and CEO of Bio Economic Research Associates, LLC (bio-era) speaks at “AI and Open Data innovation for Health” event at White House Eisenhower Executive Office

Core to the Health Tech Sprint process is people — including patients. User-centered feedback from diverse individuals guided the development of all digital tools. In the AI Ecosystem track, patient advocates like Stephen Aldrich, a 63-year-old ex-founder and CEO of Bio Economic Research Associates, LLC (bio-era) provided valuable feedback and perspective. Aldrich was diagnosed with metastatic Stage IV adenocarcinoma of the esophagus in late March 2017. “My response to my fatal cancer diagnosis served as an inspirational case study for the sprint group,” shares Aldrich. He is hopeful though for the future of experimental therapeutics via AI and Open data.

“I am extremely grateful to live at a time when what used to be a terminal cancer diagnosis can be turned into something much less threatening due, in no small part, to our exploding ability to gather and analyze personal ‘omic’ information. I envision a day when all cancer patients have had their cancers fully sequenced, and enjoy direct control over their fundamental genomic and health data, enabling them to quickly identify the best potential treatment options for their unique cancers. Amazing cures are possible, if we enable them to happen,” Aldrich says.

Rick Bangs, MBA/PMP, is a bladder and prostate cancer survivor and works as a patient advocate, primarily in research and clinical trials. He serves as patient advocate for the National Cancer Institute (NCI), SWOG Cancer Research Network, the National Comprehensive Cancer Network (NCCN), and ASCO and has leadership roles in both the NCI and SWOG. He has also supported the clinicaltrials.gov development team in its efforts to improve its user interface and provide more relevant search results.

Bangs says the AI Ecosystem demos were promising though he believes that no one supplier can make this work without partnering with other teams who bring different capabilities to the table. “The solution here will require vision, and that vision will cross capabilities that no one supplier will individually have. It might help if there was a vision that the suppliers were striving for, an aspirational

North Star if you will. That might result in the original vision being expanded and extended,” Bangs said.

He notes that applied responsibly open data AI approaches “probably” can facilitate a new experimental therapy ecosystem that will benefit the patients but he says data structures and hierarchies must be modernized and matching must account for location as well as disease, “Open data

AI approaches can facilitate the ecosystem but only one which is as strong as its weakest link which is data. The question is whether we can drive a result that is fit for purpose as defined by key stakeholders within the constraints and limitations.”

This first-ever Health Tech Sprint advanced priorities recently identified in the American AI Initiative, established by Executive Order in Feb. 2019. It also illustrated real-world value as government data was unlocked for public use as machine-readable “open data,” per the Jan. 2019 signed Foundations for Evidence-based Policymaking Act.

COLLABORATING LEADING AGENCIES



U.S. General Services Administration (GSA) Presidential Innovation Fellows (PIF) Program

Leadership by PIF Gil Alterovitz
Data gathering coordination
Reporting and administrative support
Web site: tophealth.pif.gov
Storytelling and engagement



U.S. Department of Health and Human Services (HHS) Office of the Chief Technology Officer

Leadership by Kristen Honey, Innovator in Residence
Data gathering coordination
Reporting and administrative support
Communications

FEDERAL AGENCIES



Department of Veterans Affairs (VA)

The Department of Veterans Affairs provided access to de-identified cancer patients for use in matching to trials/therapeutics, expertise in AI, and analytical tools around machine learning.



United States Patent and Trademark Office (USPTO)

To inspire prospective inventors, including those with applications in the AI ecosystem and experimental therapeutics space, the US Patent and Trademark Office assembled a large set of invention topics that have never been combined in a single patent (<https://developer.uspto.gov/visualization/anticlique>).



National Cancer Institute (NCI)

Provided 3 datasets generated at NCI. This included structured eligibility criteria, participants based on call samples to contact center, and medical professional-curated participant/trial match rating. Provided medical professional curation to make datasets, and give guidance.



Department of Commerce (DOC)/Census Bureau

Provided materials and guidance based on TOP. Was working on logistics for presentations (until shutdown).



National Institutes of Health (NIH)/National Library of Medicine (NLM)

Provided tools/api/guidance, curated patient advocate list, etc.



Food and Drug Administration (FDA)

Physician 20% detailee provided dataset curation and guidance

COLLABORATING AGENCIES



Department of Commerce



Department of Defense (DOD)

Health professional 20% detailee provided dataset curation and guidance.



Centers for Medicare and Medicaid Services (CMS)

20% detailee helped on administrative tasks like reports, etc.



Department of Housing and Urban Development (HUD)

20% detailee helped on administrative tasks like reports, etc.



Department of Energy (DOE)

Provided expertise and guidance on Summit (world's fastest computer) and computing facilities.



The White House Office of Management and Budget (OMB)

Provided expertise and collaboration on code.gov



General Services Administration (GSA)



Department of Health and Human Services (HHS)

COLLABORATING ORGANIZATIONS

PHILIPS

OAK RIDGE
National Laboratory

 **Microsoft**

ORACLE



 **Clyme**
HEALTH


CALIFORNIA CENTER for
FUNCTIONAL MEDICINE

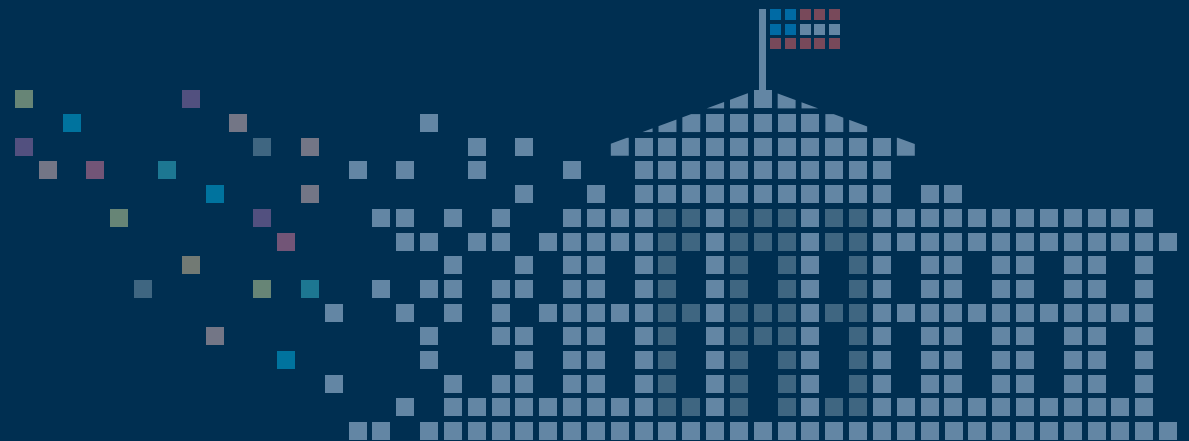

Supporting Children • Curing All

 **GLOBAL**
LYME
ALLIANCE

Trialx

 **flatiron**

IV. SHOWCASING RESULTS
AT THE WHITE HOUSE



FEBRUARY 28, 2019 SHOWCASING RESULTS AT THE WHITE HOUSE AND NEXT STEPS



TrialX/Global Lyme Alliance team in front of the White House/ West Wing

On Feb 28, 2019, the Presidential Innovation Fellows, together with the Department of Health and Human Services, hosted the participants of health sprint and cross-agency leadership in Washington D.C. at the White House Eisenhower Executive Office Building to recognize the incredible work- from data unlocked by government agencies specifically for this sprint to 11 digital tools produced. The event was kicked off by Suzette Kent (Federal Chief Information Officer, Office of Management and Budget, Executive Office of the President) and Sara Brenner (Senior Policy Advisor, White House Office of Science and Technology Policy).

Cross-agency leadership gave talks including Ed Simcox (CTO and Acting CIO at HHS), Charles Worthington (Chief Technology Officer, VA), Thomas Cabbage (Deputy Undersecretary for Science, DOE), and Andrei Iancu (Under Secretary of Commerce for Intellectual Property and Director of the USPTO). Patient perspective by Stephen Aldrich was presented and cited the

Office of the Vice President (OVP)'s invited guests in attendance, including Laura McLinn, whose son Jordan the "S.204 - Trickett Wendler, Frank Mongiello, Jordan McLinn, and Matthew Bellina Right to Try Act of 2017" is named after.

Government contributions and industry participation was recognized at the event, including around the AI-able data ecosystem incentivizing model pioneered in the sprint. For Data's Choice, the National Cancer Institute reached the diamond level. AI's Choice recognized organizations included:

Diamond (Microsoft Healthcare, Flatiron Health, Philips Research, Oracle, TrialX, and Rush Medical). Later that day, a co-working ideation session took place to plan next steps at the Department of Veterans Affairs Central Office.

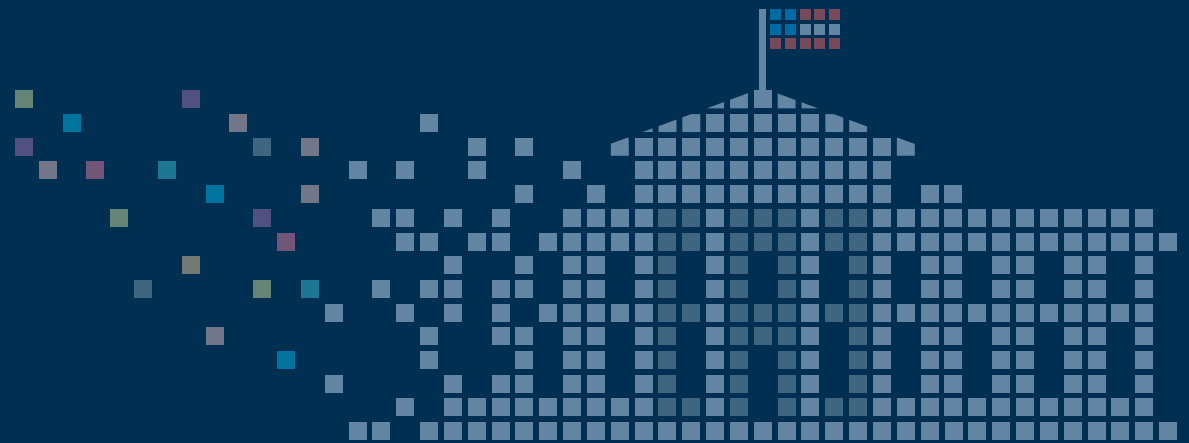


Due to wide applications and public interest, the Health Tech Sprint and its results was covered by media including articles in Fortune, Forbes, Bloomberg, and on television. The sprint was cited on the newly launched White House site [AI.gov](https://ai.gov) as an example of "AI for American Industry." The results are being presented at various other venues and spawning new challenges/sprints. Lessons learned are feeding into national AI strategy and White House-based guidance for agencies.

Are you interested in learning more about using/ extending the AI ecosystem, Data's Choice, AI's Choice, or Health Tech Sprint? If so, please see the Health Tech Sprint site tophealth.pif.gov

V.

AI AND OPEN DATA INNOVATION FOR HEALTH AT THE WHITE HOUSE



“AI AND OPEN DATA INNOVATION FOR HEALTH” AT THE WHITE HOUSE: A COLLABORATIVE EFFORT



“These efforts extract the value locked into data and drive informed decisions,” said Ed Simcox, CTO and Acting CIO, HHS (Pictured second from left)

“Through incentivizing links between government and industry via a bidirectional AI ecosystem, we can help promote usable, actionable data that benefits the American people,” said Joshua Di Frances, PIF Executive Director (Pictured first from left)

Also pictured, from left: Michael Kratsios, US CTO and Deputy Director, White House Office of Science Technology and Policy; Dr. Kristen Honey, Innovator in Residence at HHS; Dr. Gil Alterovitz, PIF.



The Health Tech Sprint “collaboration enables government and industry to collaboratively evolve solutions to meet the needs of patients,” said Rebecca Laborde, Global Healthcare and Precision Medicine Strategist, Oracle. (Pictured fourth from left)

“Collaborative ideation: This was much more open-ended than just a scope of work and provides a different environment to collaborate with government,” said Dan Kuenzig, Director, Public Sector Strategy (Pictured first from left).

Also pictured, from left: Matt Freeman, Oracle; Gil Alterovitz, PIF; Davar Ardalan, PIF Deputy Director.



“AI AND OPEN DATA INNOVATION FOR HEALTH” AT THE WHITE HOUSE: AGENCY PERSPECTIVES



“As demonstrated by the TOP Health initiative, collaboration is a catalyst for realizing innovation faster; it can ensure that the U.S. will always be able to remain competitive,” said Andrei Iancu, Director of the US Patent and Trademark Office.



US Department of Health and Human Services at the White House event.



National Cancer Institute team gets ready to present results. “The collaborations among participants and with the Government initiated by this TOP Health Sprint have been fruitful... This direct interaction helped me think of ways to work with my team to refine the [NCI] datasets to better meet patients’ needs,” said Dr. Gisele A. Sarosy, Associate Director for Informatics and Biomarkers at the NCI’s Coordinating Center for Clinical Trials.



Dr. Nhan Do, Director, Cooperative Studies Program Informatics Center, US Department of Veterans Affairs, presenting at White House event, presented work of the VA toward providing VA data via the Precision Oncology Program.

“AI AND OPEN DATA INNOVATION FOR HEALTH” AT THE WHITE HOUSE: SAMPLING OF PERSPECTIVES FROM THE HEALTH TECH SPRINT



“Health Tech Sprint booths outside Indian Treaty Room at the White House

“Improving recruitment needs to be a community effort. We have the intention to continue the collaboration within the newly created community and to contribute to initiatives focused on facilitating recruitment by standardizing the clinical trial criteria and protocols” - Anca Bucur, Senior Researcher, Philips Research

I was amazed to see the projects developed that leveraged the data. I was very pleased to learn that some of the participants are stepping up their commitment to transparency and will share their machine learning algorithms as well, so the public may understand better on how the data is used.

The biggest takeaway for me was the potential positive impact of open data.” - Irina Muchnik, Backend Web Developer, Consumer Financial Protection Bureau.

“Data quality is critical. It is not enough to have structured data, it is also key that structured data is complete and properly aligns to the schema.”

- Hadas Bitran, Microsoft, Head of Microsoft Healthcare Israel.

“Through the Health Tech Sprint, we found that we can make data designed for AI, namely AI-able data, in a way that can foster an ecosystem for innovation with the private sector and directly impact patient care.”

-Gil Alterovitz, PIF



“AI AND OPEN DATA INNOVATION FOR HEALTH” AT THE WHITE HOUSE: RESULTS FROM THE HEALTH TECH SPRINT



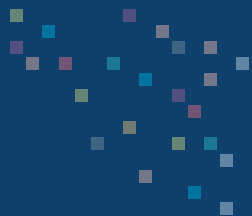
“The logic that we built for [the Health Tech Sprint] will be applied to all patients who go to a clinic that uses Flatiron’s OncoTrials, our oncology-specific clinical trials tracking software. Through TOP, we have developed a mechanism that surfaces patient trial matches directly to the provider. Over 300,000 patients, across 50 clinics in the U.S., will be matched to all actively recruiting trials at their community oncology practice.”



“Our participation in the [Health Tech] Sprint produced a clinical trials (CT) data model that will be the underpinnings of an Exa-scale knowledge graph [for biomedical text/AI],” said Dr. Tourassi, distinguished R&D scientist and the founding director of the Health Data Sciences Institute at the Oak Ridge National Lab.



The information transformation pipeline developed during the sprint to computerize the eligibility criteria of the provided clinical trials (Dataset 3) will be essential for our on-site deployments planned for 2019 at customer sites in North America and Europe... The feedback during the sessions, the user requirements, and the provided datasets led to significant refinements to our trial exploration tool,” said Anca Bucur, Senior Researcher, Philips Research.



“AI AND OPEN DATA INNOVATION FOR HEALTH” AT THE WHITE HOUSE: SHARING PERSPECTIVES



Charles Worthington, CTO of Department of Veterans Affairs speaks about VA efforts around making data available at event at White House.

“[The Health Tech Sprint] collaboration enables government and industry to collaboratively evolve solutions to meet the needs of patients.”

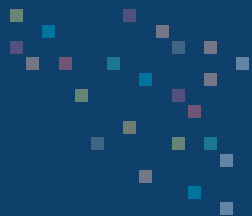
-Rebecca Laborde, Global Healthcare and Precision Medicine Strategist, Oracle.



Getting ready for the White House event. Suzette Kent, Federal Chief Information Officer, Office of Management and Budget, Executive Office of the President; Laura McLinn, patient advocate.



Sharing thoughts on Summit, the fastest computer in the world, at The Opportunity Project Demo Day booths featuring Oak Ridge National Lab team of Health Tech Sprint. Dr. Kelvin Droegemeier, Director, White House Office of Science and Technology Policy; Dr. Gil Alterovitz, PIF



“AI AND OPEN DATA INNOVATION FOR HEALTH” AT THE WHITE HOUSE: NEW INSIGHTS



Dr. Casey Frankenberger of Rush Medical Center talks about clinical trials tool that spun out of Rush Medical into a start-up company during the Health Tech Sprint.



Dorothy Aronson, CIO at the National Science Foundation; Dr. Gil Alterovitz, PIF at “AI and Open Data for Health Innovation” event.



Presenting infographic summarizing the Health Tech Sprint. “Coach” Kathy Kemper, Institute for Education; Dr. Kristen Honey, HHS; Dr. Gil Alterovitz, PIF; Davar Ardalan, PIF..



Dr. Sharib Khan, Chief Executive Officer, TrialX presenting clinical trial application at TrialX booth at event at White House.



INCREASING PUBLIC AWARENESS AROUND AI AND OPEN DATA: APPLICATIONS IN CLINICAL TRIAL MATCHING



Health Tech Sprint DOE Oak Ridge National Lab team participants present at Tech Day 2019 in collaboration with the White House/CIO Council "Industries of Future" booth space hosted at the Dept. of Labor.



Health Tech Sprint cited in newly launched White House site AI.gov as example of "AI for American Industry."



Public display of booths from the Health Tech Sprint teams at the Department of Commerce Census Bureau's The Opportunity project Demo Day, Mar. 1, 2019



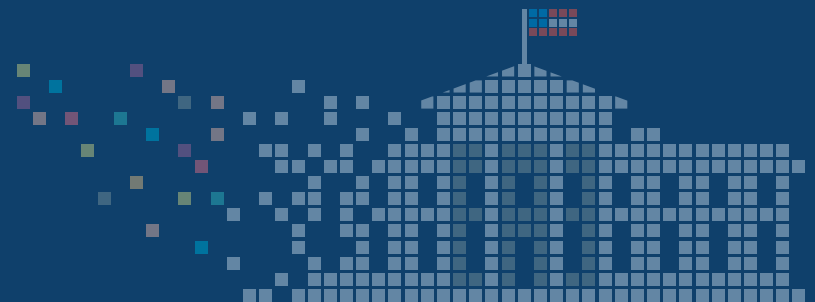
Clinical Trial Bot work from Health Tech Sprint is featured on Bloomberg, Fortune, and other venues.

ADDITIONAL INFORMATION AND THANKS

Presidential Innovation Fellows (PIF), are part of our nation's Technology Transformation Services (TTS) at the U.S. General Services Administration (GSA). TTS's mission is to improve the public's experience with the government by helping agencies build, buy, and share technology that allows them to better serve the public.

The PIF program was established by the White House Office of Science and Technology Policy (OSTP) in 2012 to attract top innovators into government who are capable of tackling issues at the convergence of technology, policy, and process. The program, which in 2013 established a permanent home and program office within GSA, has recruited 135 Fellows working within 35 agencies. PIFs have worked in tandem with federal agencies to solve critical governance issues including accelerating speed to service, developing new methods of procurement, and encouraging risk-taking in public-private partnerships.

Special thanks to: Ed Simcox, Liz Squire, Jacob Ashendorf, and Aspen Kissell at HHS; Drew Zachary and Lorena Molina-Irizarry at the Census Bureau; Joshua Di Frances, Davar Ardalan, Nina Bianchi and Justin Koufopoulos at GSA as well as designer Kate Saker and the following via Open Opportunities: Irina Muchnik, Anniebell Stewart, Abhivyakti Sawarkar, and Lucinda Gish.



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.

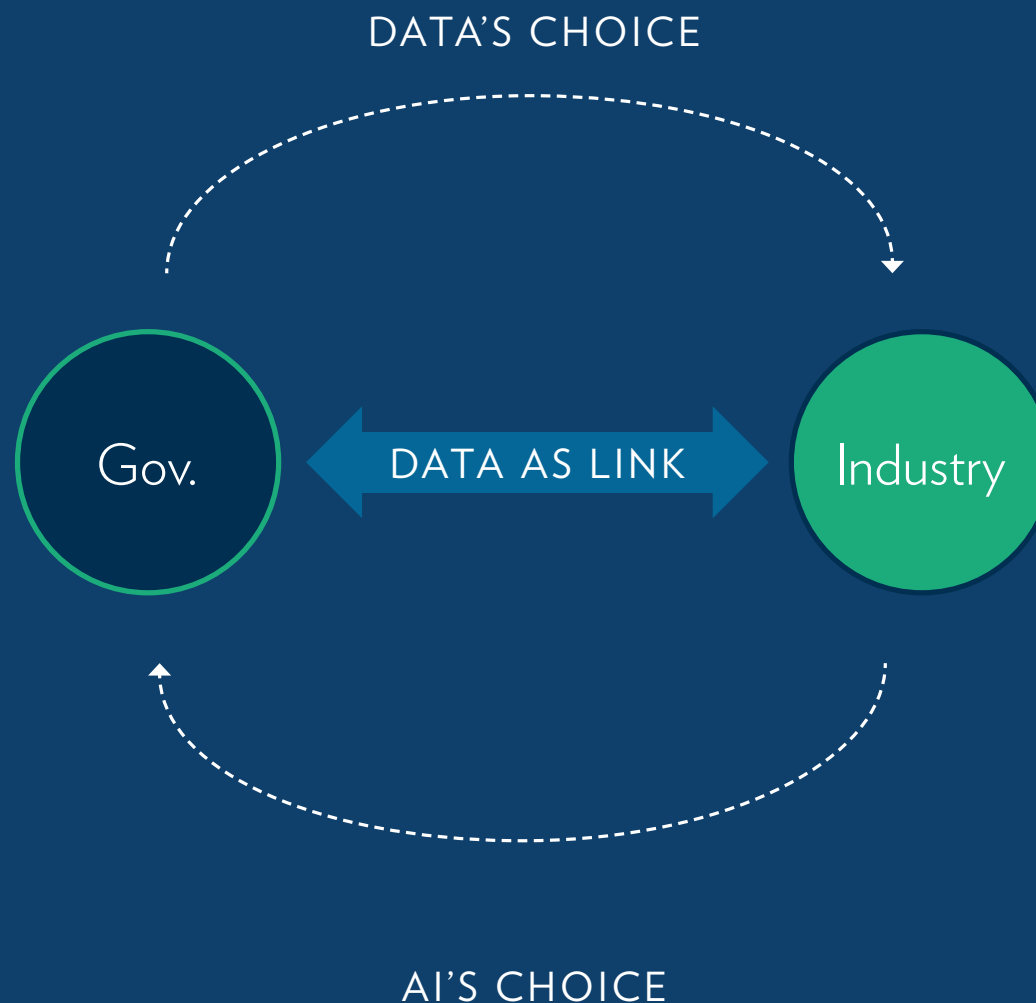


Figure 1. Public-private collaborations via data linkages

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, like for 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 paramount 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 federal to 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 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.

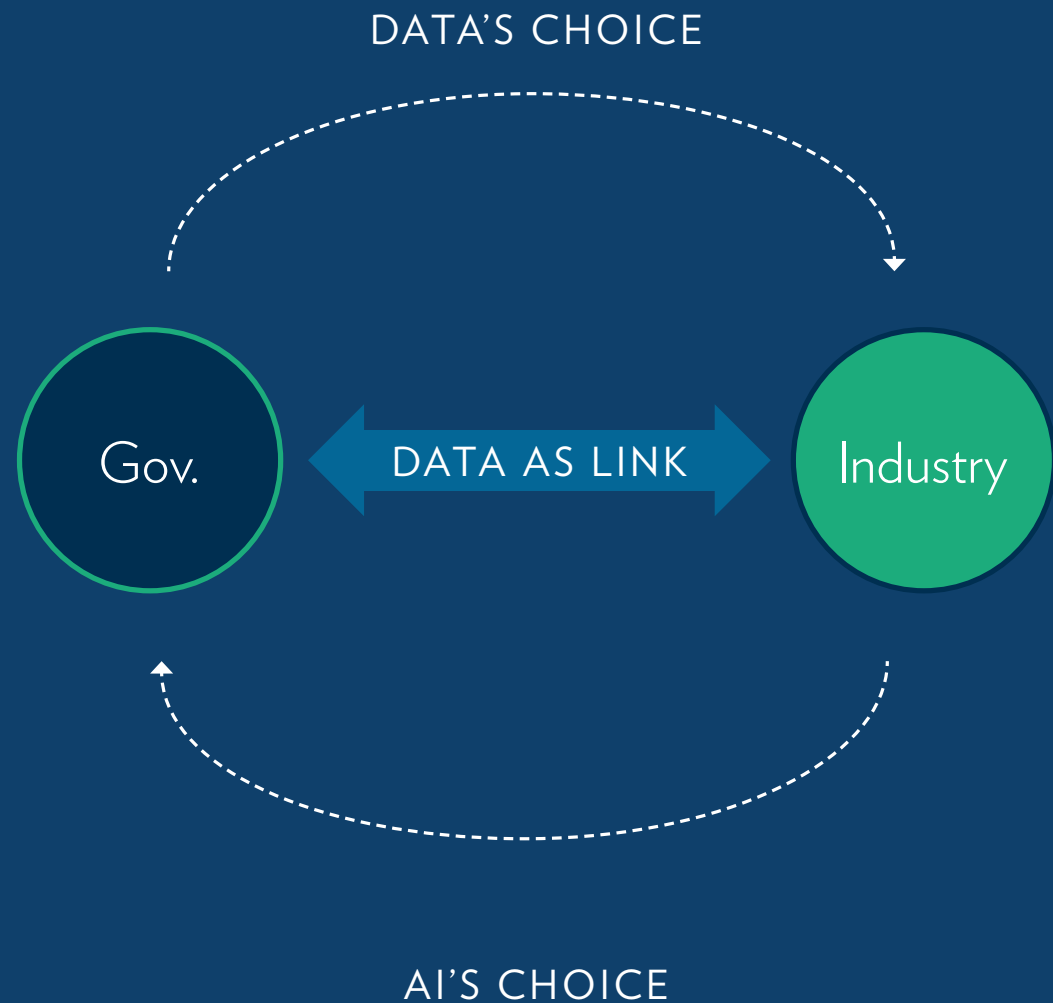


Figure 1. Data as link between Government and other sectors like Industry.

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 Alcentric 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

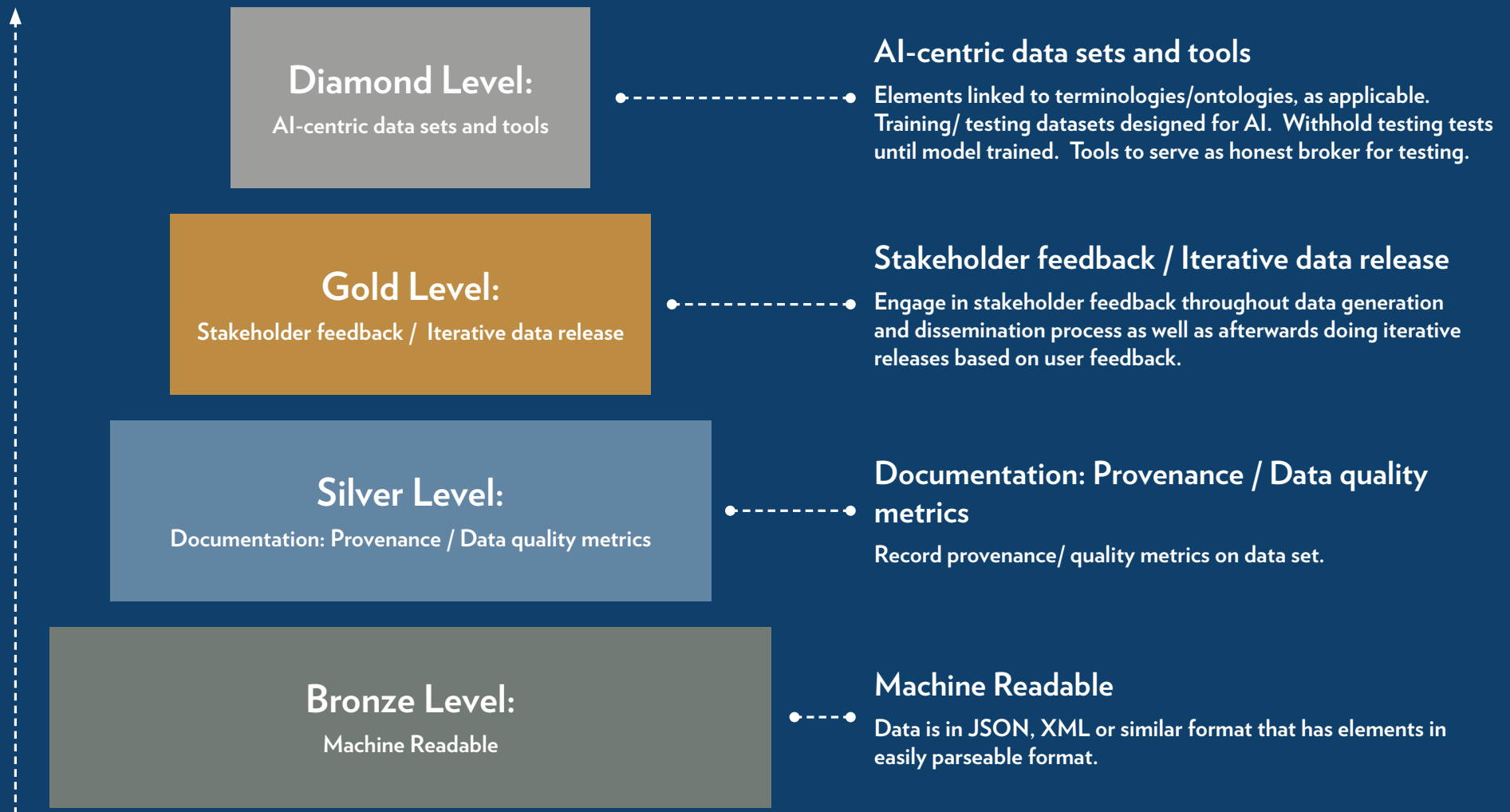


Figure 2. Data's Choice Levels: What makes federal data also useful to industry to build tools, and how can this be potentially measured and incentivized?

* Each level implicitly builds on previous one.

AI'S CHOICE LEVELS

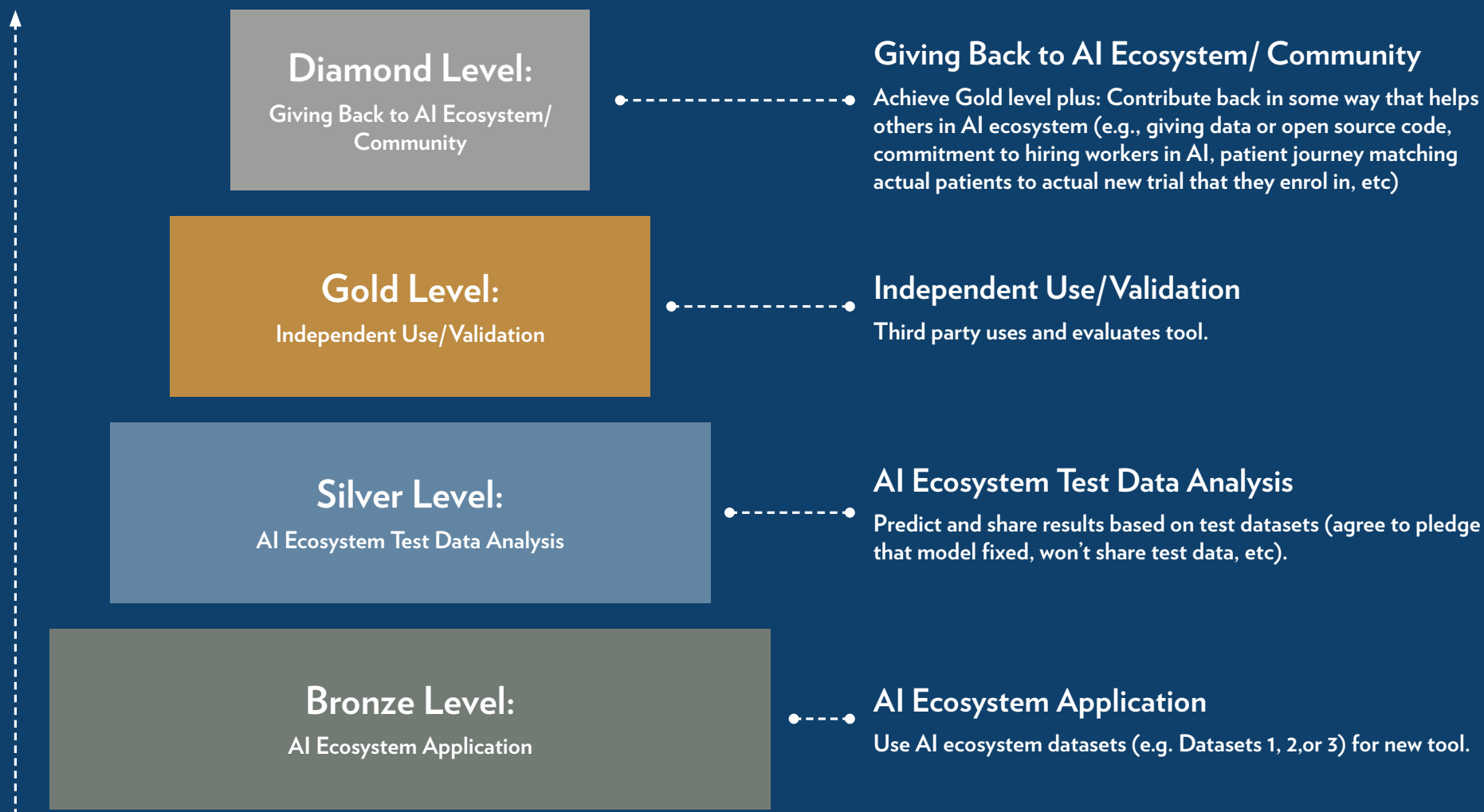


Figure 3. AI's Choice Levels: What makes industry AI/data results usable by others, like for federal agencies, so that they leverage industry tools -- and how could this be potentially measured and incentivized?

* Each level implicitly builds on previous one.

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 Alable 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 entities (and not sensitive), but is not made available for download/access.	•	
Data quality is unpredictable. Few datasets data quality metrics (either qualitative or quantitative).	•	
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/applications.	•	
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 different means for access in terms of authentication/authorization- so can be tricky to find/coordinate the accounts/information.	•	
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's terms 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		•

Bronze

Silver

Gold

Diamond