



Fulfilling the Promise of Information Technology

Taha Kass-Hout, MD, MS

FDA Chief Health Informatics Officer (CHIO)

FDA Chief Technology Officer (CTO), *Acting*

Director, FDA Office of Informatics and Technology Innovation (OITI)

PhUSE Computational Science Symposium | March 17, 2014/ 8:30 am- 9:30 am ET

Any views or opinions expressed here do not necessarily represent the views of the FDA, HHS, or any other entity of the United States government. Furthermore, the use of any product names, trade names, images, or commercial sources is for identification purposes only, and does not imply endorsement or government sanction by the U.S. Department of Health and Human Services.



Presidential Executive Order for Open Data

On May 9, 2013, President Obama enacted the executive order surrounding Open Data within the Federal Government



The screenshot shows the White House website with the following content:

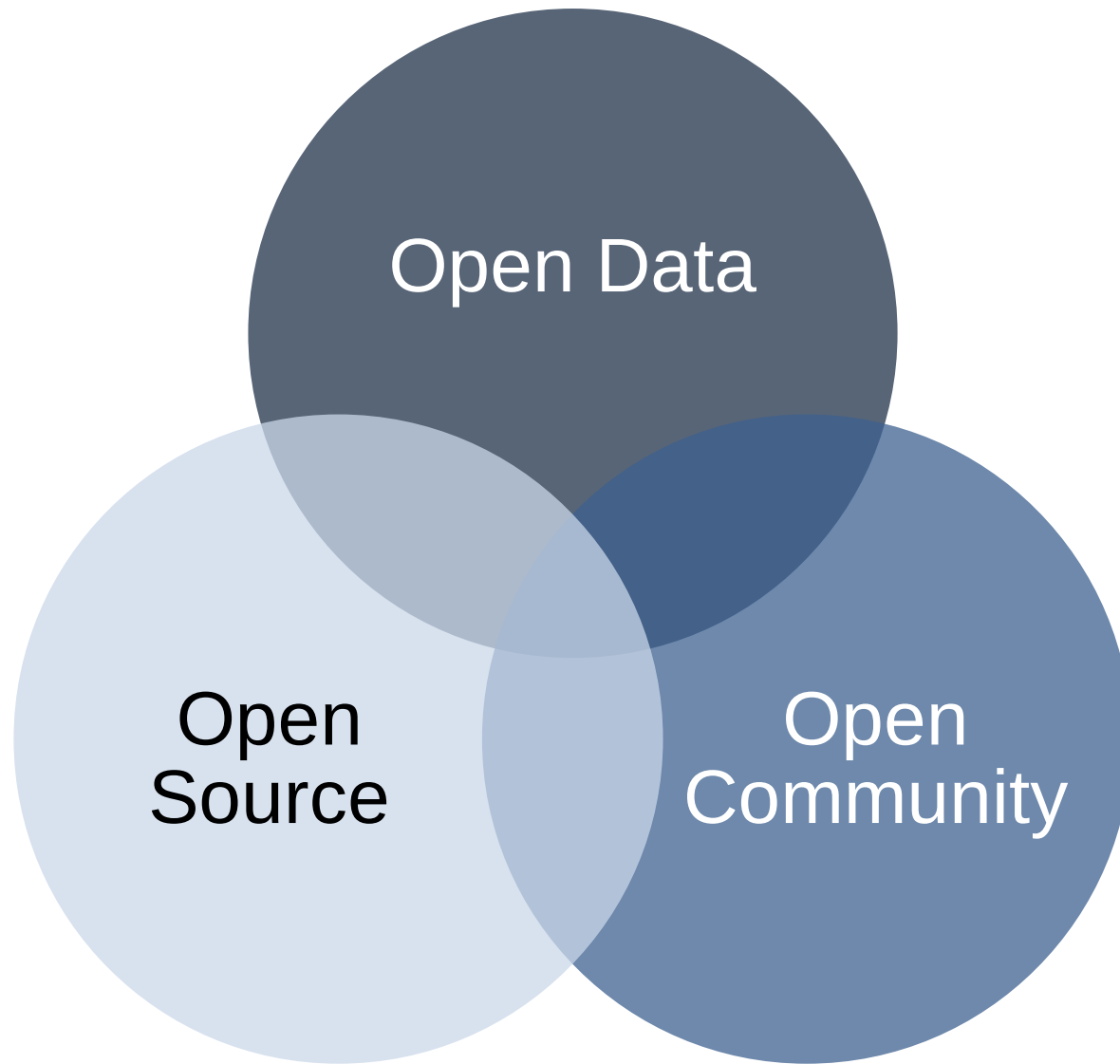
- Header:** the WHITE HOUSE PRESIDENT BARACK OBAMA, with navigation links for BLOG, PHOTOS & VIDEO, BRIEFING ROOM, ISSUES, the ADMINISTRATION, the WHITE HOUSE, and our GOVERNMENT.
- Breadcrumb:** Home • Briefing Room • Presidential Actions • Executive Orders
- Search:** Search WhiteHouse.gov
- Metadata:** The White House, Office of the Press Secretary, For Immediate Release, May 09, 2013.
- Title:** Executive Order -- Making Open and Machine Readable the New Default for Government Information
- Section:** EXECUTIVE ORDER, MAKING OPEN AND MACHINE READABLE THE NEW DEFAULT FOR GOVERNMENT INFORMATION
- Text:** By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:
 - Section 1. General Principles.** Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation.
- Additional Text:** Decades ago, the U.S. Government made both weather data and the Global Positioning System (GPS) data available to the public. American entrepreneurs and innovators have utilized these resources to create navigation systems, location-based applications, precision farming tools, and many other products and services, leading to economic growth and job creation. In recent years, the U.S. Government has posted in machine-readable form for free public use on Data.gov. Entrepreneurs and innovators have continued to develop a vast range of useful new products and businesses using these public information resources, creating good jobs in the process.
- Conclusion:** To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable. Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. In making this the new default state, executive departments and agencies (agencies) shall ensure that they safeguard individual privacy, confidentiality, and national security.

"By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. General Principles. Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation."

February 24, 2014 4:45 PM EST
A White House Call To Action to Advance the BRAIN Initiative
 President Obama is calling for BRAIN Initiative to be an "all hands on deck" effort, involving not only the federal government but also companies, health systems, patient

Executive Order -- Making Open and Machine Readable the New Default for Government Information, May 9, 2013. Available at: <http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government> Accessed on February 25, 2014



>80 public-access resources or data are currently indexed, many updated daily, including adverse drug events, reports involving medical devices, searchable listings of over-the-counter tests cleared or approved by the FDA, and a database of accredited mammography facilities.



FDA Data Inventory

An inventory of all FDA Datasets

Download the Data: [API](#)

Agency and Program Information

Agency:	U.S. Food and Drug Administration
Federated Catalog:	open.fda.gov
Data Series Location:	http://open.fda.gov/data.json
Keywords:	data-json, inventory
Date Updated:	Dec 20, 2013

Data Documentation

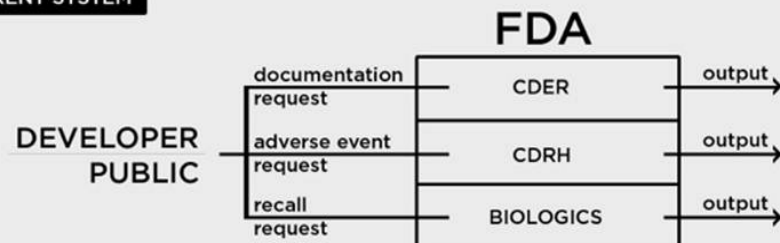
License Agreement:	Public Domain
--------------------	---------------

For Developers

CKAN Page:	http://hub.healthdata.gov/dataset/fda-data-inventory
Metadata API:	JSON RDF/XML N3

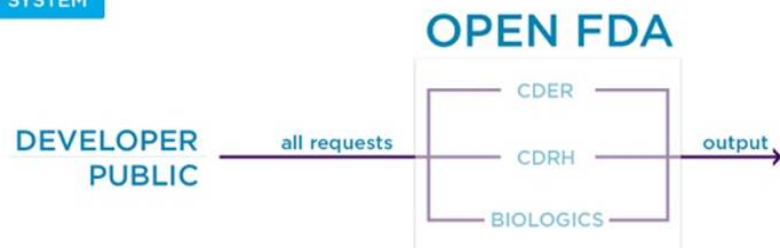
OpenFDA Vision

CURRENT SYSTEM



OpenFDA aims to offer high-value public datasets via developer-friendly means (**APIs or downloads in an elastic Cloud environment**) to further the regulatory and scientific missions, educate the public, spur innovation, and save lives while protecting the privacy and security of the information

NEW SYSTEM



openFDA

» About openFDA
» @openFDA

Open data

Easier and better access to structured FDA data.

Developer-focused APIs, raw data, and documentation for a curated set of high value public FDA datasets.

Open source

Code and documentation.

Open source API code and documentation, shared on [GitHub](#).

Open community

Share examples, apps, and stories.

Developers, researchers, the public, and FDA in one shared community on GitHub, StackExchange, and Twitter to best utilize these datasets.

Sign up for updates

Email address

FDA on GitHub

We're hosting this blog on GitHub, and we'll use it for openFDA API code too.

@openFDA on Twitter

Use hashtag #openFDA to join the openFDA community!

A research project to provide open APIs, raw data downloads, documentation and examples, and a developer community for an important collection of FDA public datasets.

[More about openFDA »](#)

Introducing openFDA

By Taha Kass-Hout | 04 March 2014



Welcome to the new home of openFDA! We are incredibly excited to see so much interest in our work and hope that this site can be a valuable resource to those wishing to use public FDA data in

OpenFDA will **lower the barrier of entry for developers of consumer and enterprise applications** who want to use FDA's public-access data

<http://open.fda.gov> | open@fda.hhs.gov | @openFDA

FDA Open Data - Applications and Benefits

- Adverse Events
 - Drug-Drug Interactions
 - Drug-Food Interactions
 - Medical Devices
- Recalls & Risk Maps
- Label changes & Updates
- Inspections and citations
- More datasets to be announced as development continues



FDA's Impact on Food Pathogen Research

Existing methods for classifying and tracking pathogens in our food and environment are not meeting the demands of a federated, globalized food chain

The objective of this program is to sequence the genetic codes of 100,000 strains of important food pathogens, such as *Salmonella*, and making them available in a free and public database at NIH's National Center for Biotechnology Information.

- FDA's Field Labs will be contributing more than 5000 whole pathogen genomes annually
- FDA's Food Safety Network of State and Federal labs will be contributing all of their whole genome sequences of pathogens associated to food illness
- FDA's Center for Veterinary Medicine will be contributing to the 100K genome project as well

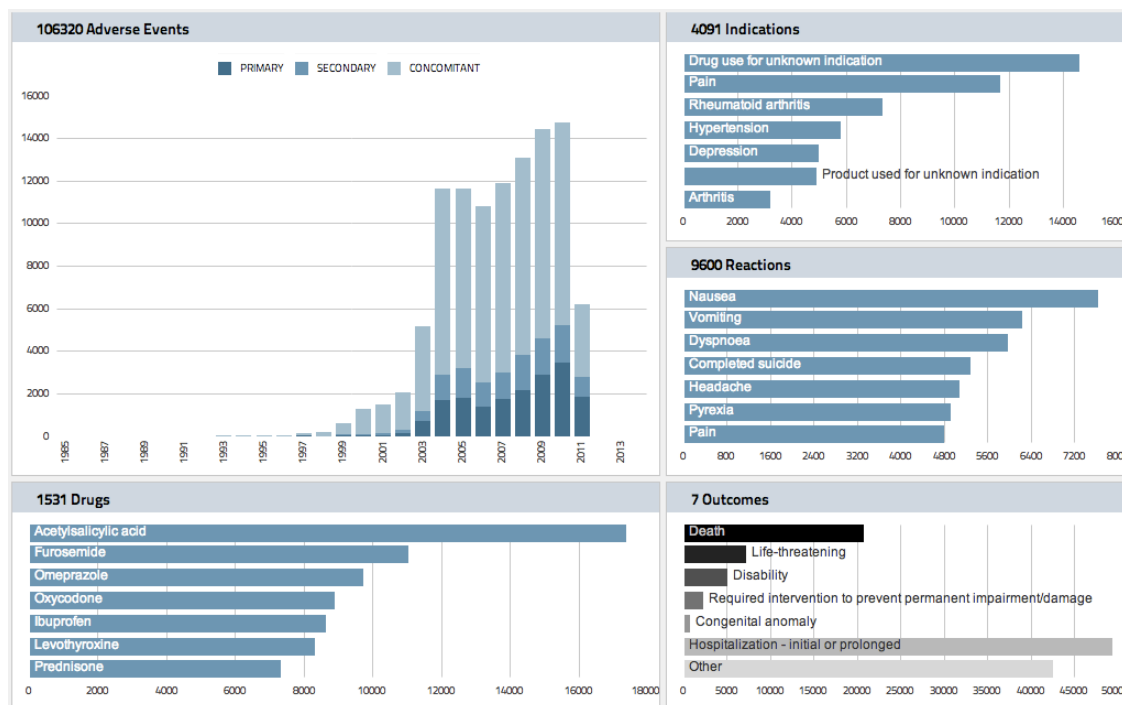


<http://www.hhs.gov/open/initiatives/hhsinnovates/round5/fda-100k-genome.html>

Publically Available Data – Adverse Events via openFDA

Using publically accessible data we can look at the incidence of drug-drug interactions associated with specified outcomes

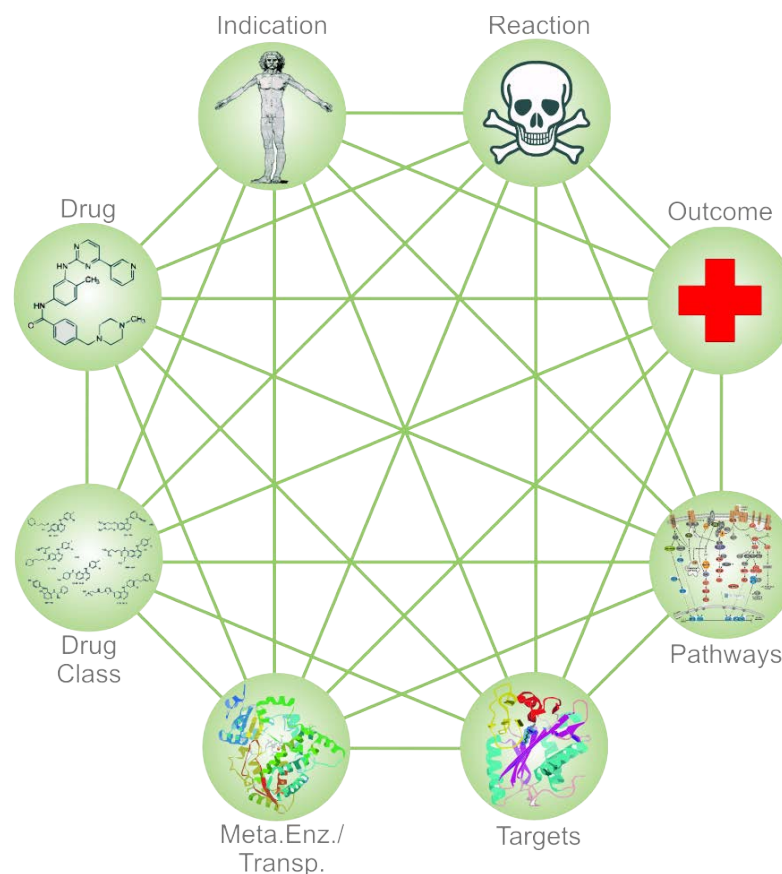
In a simple example, we looked at over **80,000** Acetaminophen adverse events where nausea was an adverse event and death was an outcome



Publically Available Data – Adverse Events via openFDA

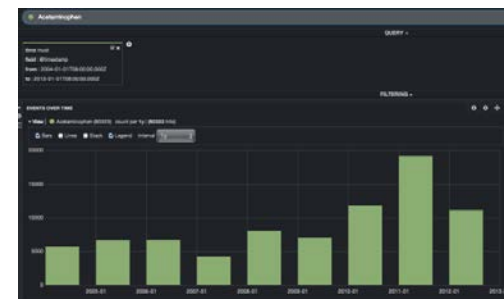
We have the ability to look at adverse event data against a wide array of accessible data:

- Indications
- Drug Data
- Drug Class
- Enzyme Associations
- Biological Targets
- Biological Pathways
- Outcomes
- And DDI Reactions

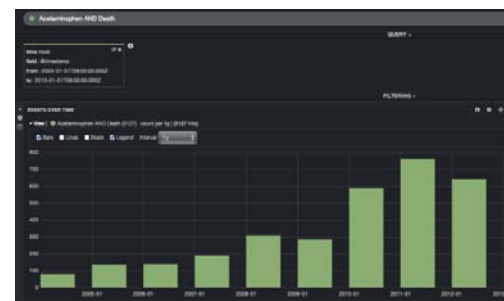


Acetaminophen adverse events where Diphenhydramine was concomitant medication and had death as the outcome

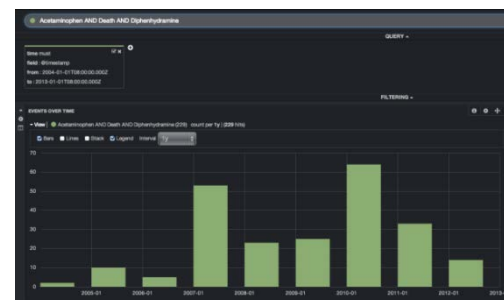
Acetaminophen 80,323 AEs



Acetaminophen and Deaths: 3,127 AEs



Acetaminophen and Deaths and Diphenhydramine: 229

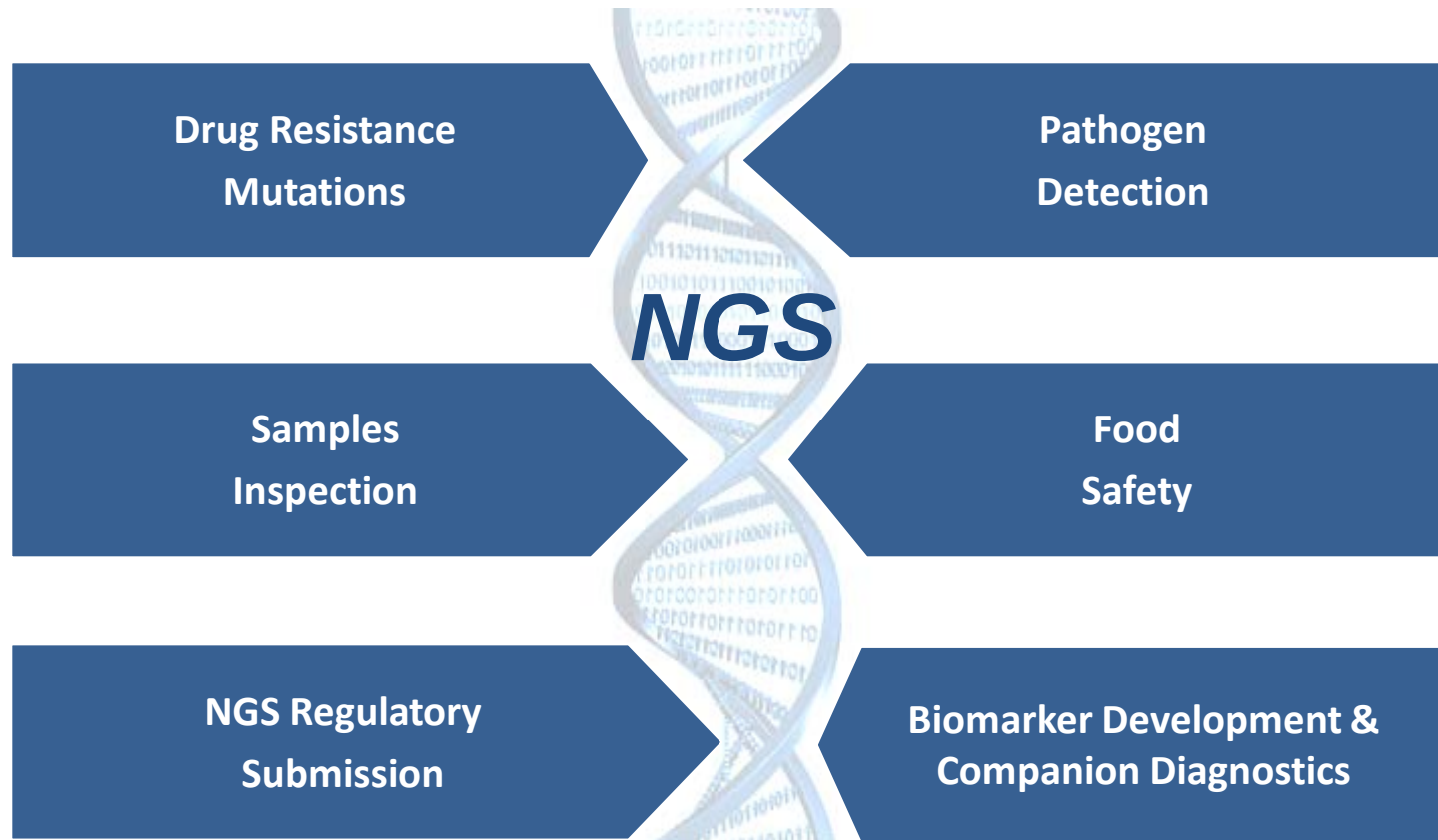


[illegible]

FDA's Need For Next Generation Sequencing Capabilities



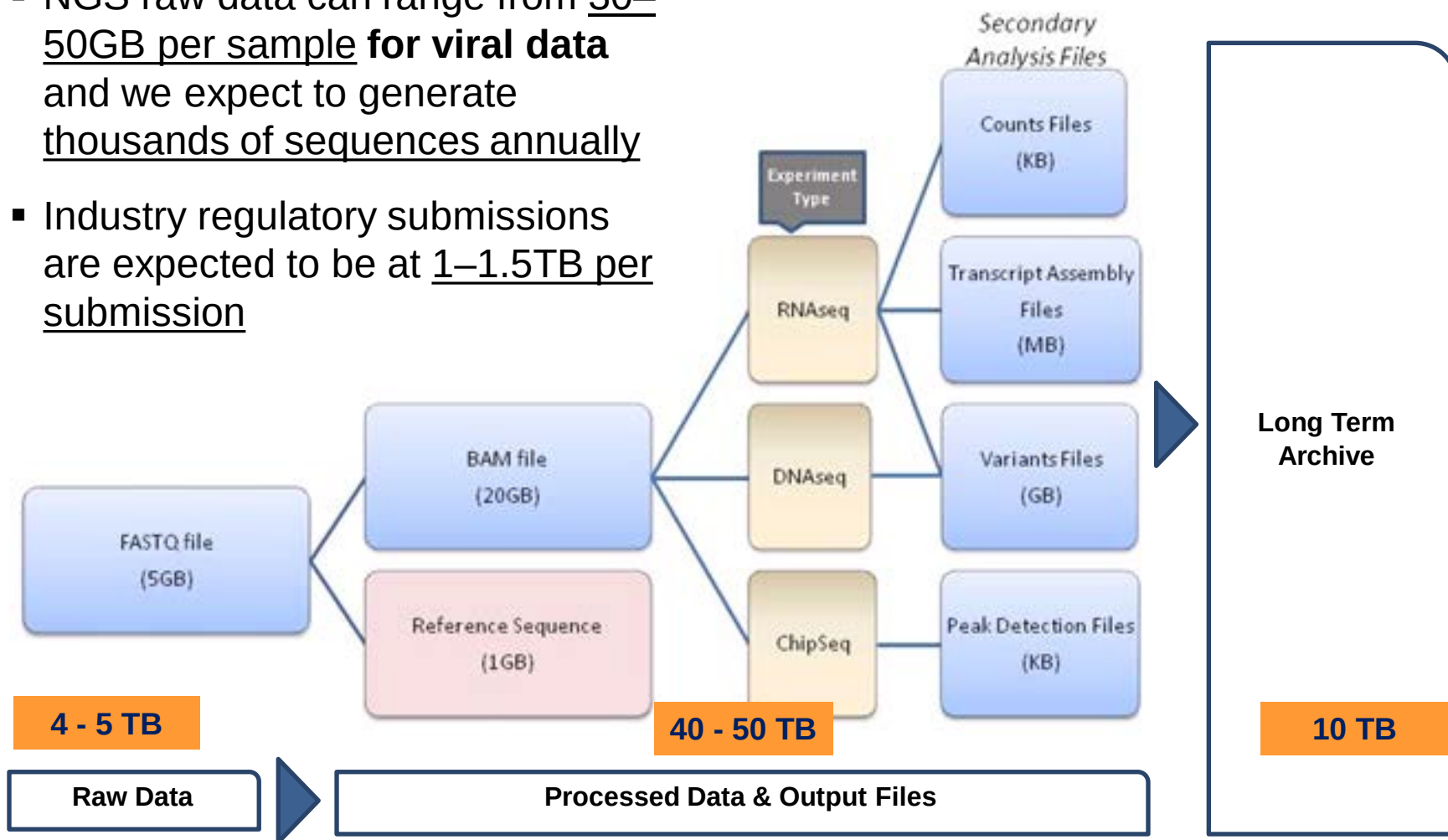
FDA's growing demand for Next Generation Sequencing (NGS) requires the Agency to develop related business and IT capabilities



It is imperative for FDA to effectively generate, analyze, and share NGS data

NGS is a Big Data Initiative

- NGS raw data can range from 30–50GB per sample **for viral data** and we expect to generate thousands of sequences annually
- Industry regulatory submissions are expected to be at 1–1.5TB per submission



Immediate Use Cases For NGS at FDA

Use Case I

Data Sharing

FDA is rapidly expanding the development of Whole Genome Sequencing (WGS) capabilities to more accurately conduct and share epidemiologic outbreak investigations, better understand pathogenic bacteria virulence traits, and the factors that influence their adaptability to food manufacturing environments



Use Case II

Open Data

The Food Safety Genomics team is coordinating efforts with State, Local and Federal public health labs to sequence pathogens collected from foodborne outbreaks, contaminated food products and environmental sources and make available to the public



Use Case III

Regulatory Review

The FDA's Computational Science Center has started to receive NGS data with regulatory submissions and must store and disseminate this data to the review teams for validation and analysis



NGS is a Critical Tool For Modern Food Safety

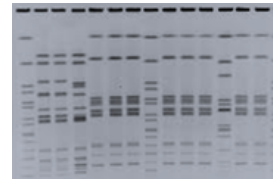


Why Genomic Approach to Pathogen Identification?

Traditional Approach To Pathogen Identification is only effective in a limited scope: Pulse Field Electrophoresis combined with Serology has been the ‘gold standard’ for FDA pathogen identification for years



Bacterial cultures are isolated from a potential contamination source and grown on an agar plate to incubate the bacteria.

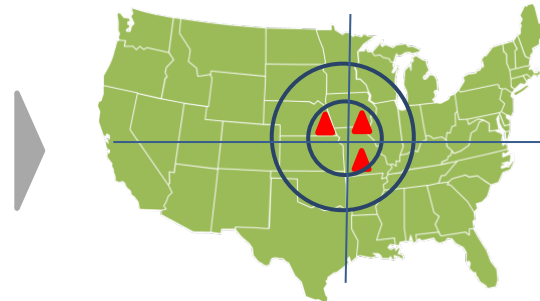
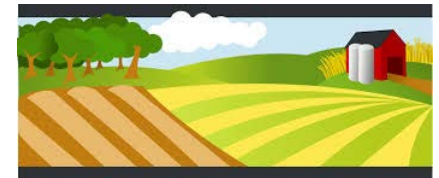


The DNA is extracted, cleaved, and exposed to electric currents that separate the DNA fragments by size – and the resulting displacement is recorded as the “fingerprint” **These bands are stable – and reproducible**

Many isolates show patterns that are distinct to a geographic region that can help investigators determine potential sources. However, this approach has faced major challenges or limitations.

- This is a measure of “Relatedness” and not a true phylogenetic measure
- Many strains cannot be sub-typed in this approach and belong to more uniform clonal groups

Raw Sequencing Data

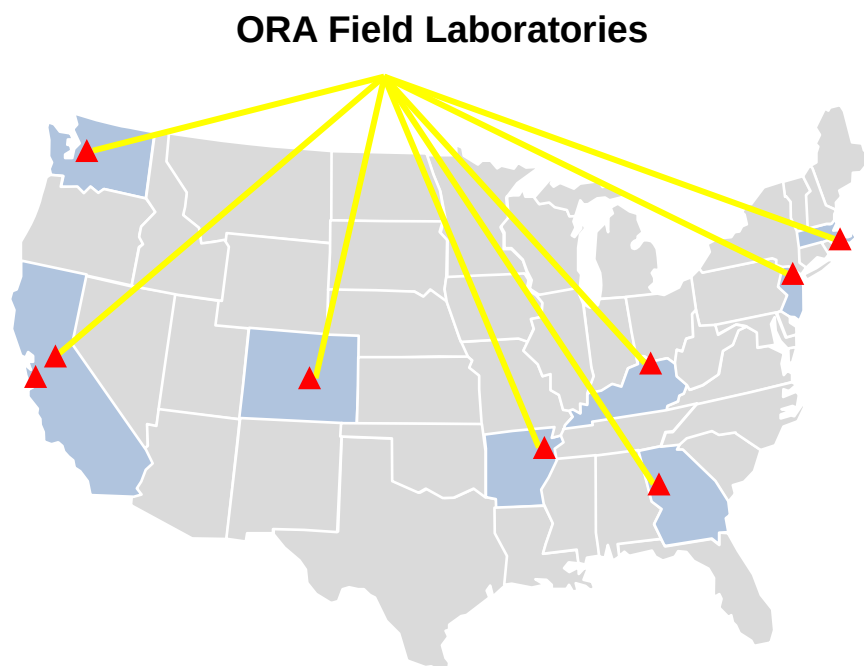


Identify Contamination Source

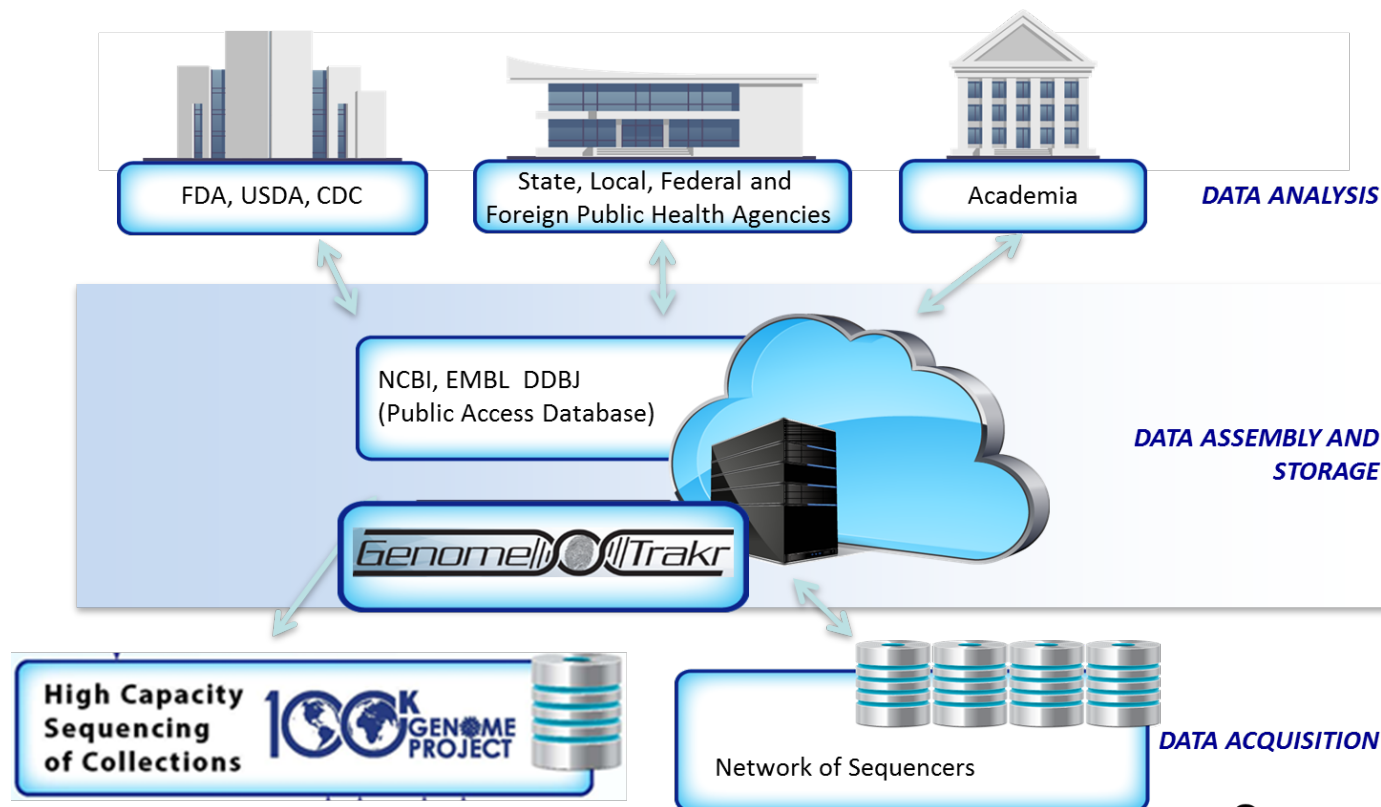
NGS Initiatives at FDA Field Labs

FDA is participating the Next Generation Network for the Food Pathogen Traceability (NGN-FPT) – a network of Federal, State, and global partners to populate a publically accessible Salmonella genome database in collaboration with the 100K Genome project

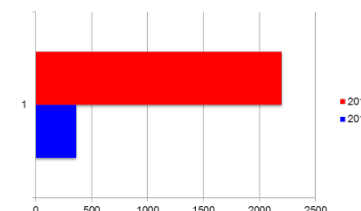
- FDA labs began to sequence the vast archives of bacterial isolates from hundreds of outbreak investigations and surveillance assignments
- 9 field laboratories have been equipped with sequencing platforms and began contributing to Salmonella strain inventories
 - Each lab will be sequencing 200 – 400 Salmonella isolates per year, which will expand the public database by 5200 sequences per year
- FDA is evaluating NGS platforms and comparing effectiveness for use in the field labs



FDA Food Safety Use of Genomic Data

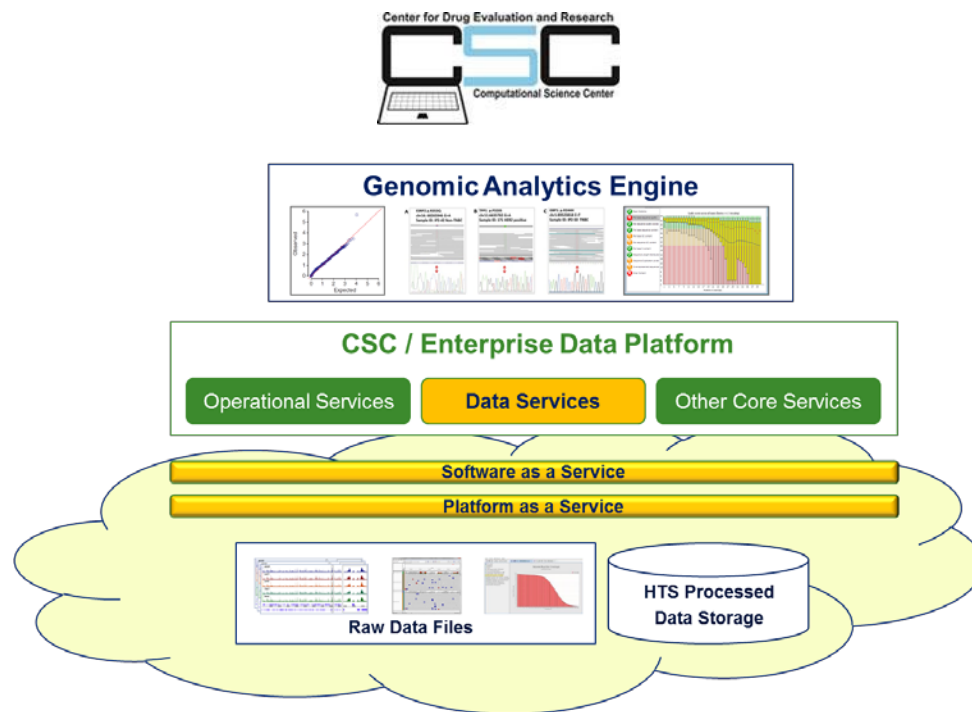


Sequences generated by GenomeTrakr and FDA



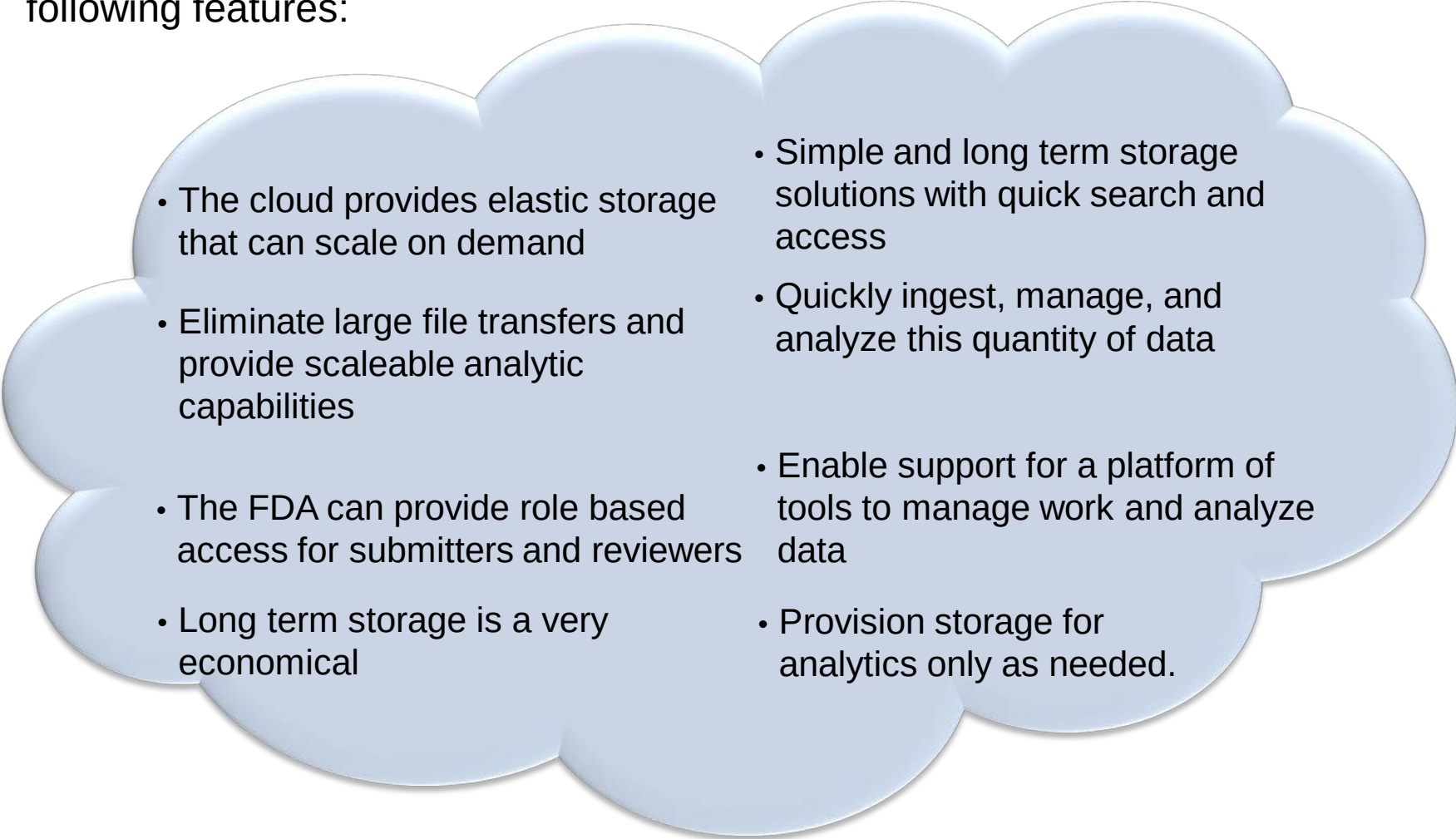
In May of 2013, CDER received its first full product review with large NGS datasets. It became known as the “**Terabyte Submission**” and presented a series of challenges for the review, analytics, and high performance computing environments required to analyze the information.

- The NGS data was put on hard drives and manually moved to workstations and HPC environments throughout the review process
- OITI is working with CDER and the Computational Science Center to define an environment that can manage large data sets, provide controlled access, and integrate with FDA high performance computing and elastic computing environments



The Cloud as an Enterprise Genomic Platform for FDA

A cloud-based genomics platform as an enterprise solution that has the following features:

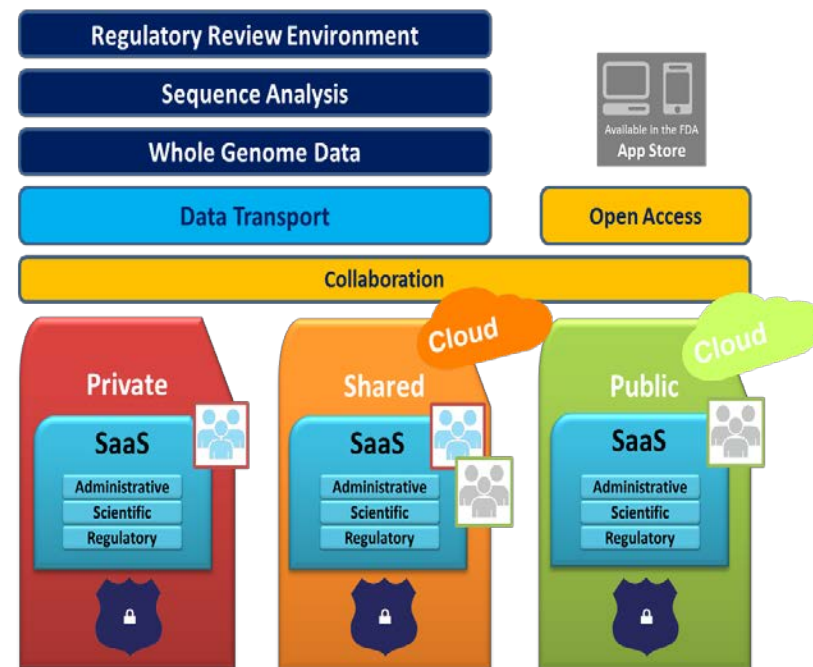
- 
- The cloud provides elastic storage that can scale on demand
 - Eliminate large file transfers and provide scaleable analytic capabilities
 - The FDA can provide role based access for submitters and reviewers
 - Long term storage is a very economical
 - Simple and long term storage solutions with quick search and access
 - Quickly ingest, manage, and analyze this quantity of data
 - Enable support for a platform of tools to manage work and analyze data
 - Provision storage for analytics only as needed.

NGS Platform As A Service

FDA Centers and OITI are moving towards a unified environment for managing NGS data – not only meeting the objectives for the 3 immediate use cases, but providing the platform to support emerging needs as well.

OITI is providing a secure NGS Platform

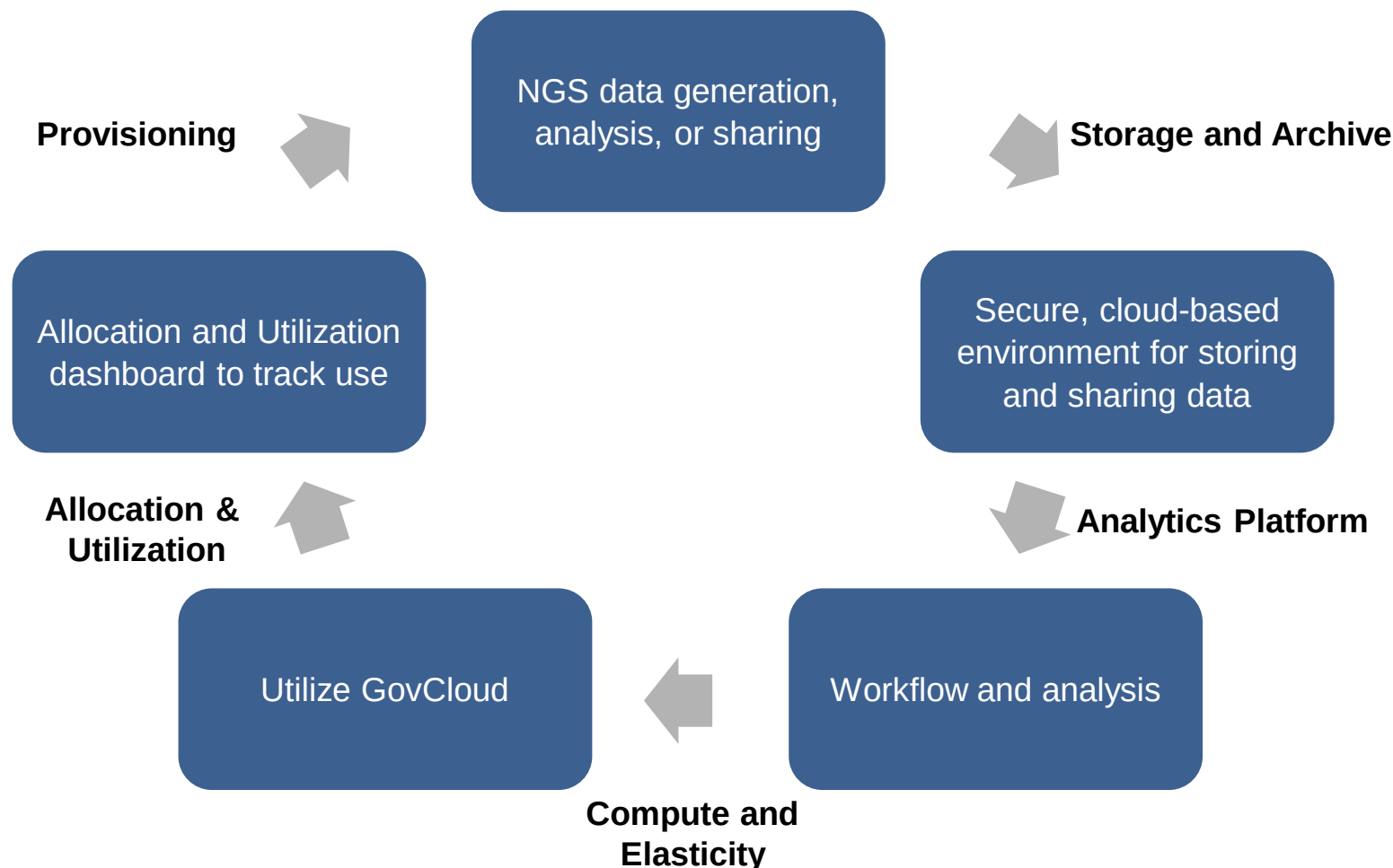
- FDA users can easily generate, analyze and securely share NGS data
- Opens the future to other utility such as; biomarker development, drug safety, companion diagnostics, etc. in collaboration with FDA partners
- NGS data can potentially provide tremendous value to the public when combined with openFDA



Next Generation Sequencing (NGS) is an immediate priority for FDA

NGS Platform As A Service

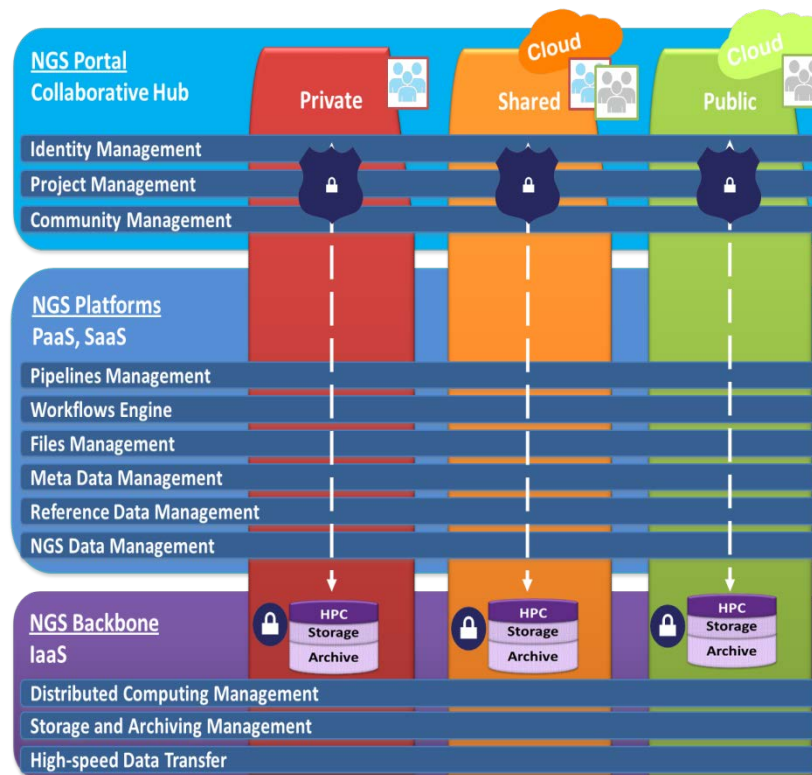
We are working on a cloud-based genomic platform that offers **NGS services** including: storage, analysis and collaboration environments for FDA's Centers and Offices and their partners



NGS is Part of the Big Picture at FDA

Collaborative NGS environments can be set up through the use of public clouds (GovCloud), collaborative NGS platforms and hubs

- **Public Clouds:** FedRAMP certified vendors can provide IaaS as the NGS backbone for FDA.
- **NGS Platforms:** NGS workflow platforms such as Galaxy and Arvados can enable collaboration through analysis pipelines and audit trails
- **Collaborative Hubs:** Scientific collaboration platforms like HUBzero can help create NGS portals for supporting the sharing of NGS data, pipelines and results by FDA with our partners or the public



The NGS space has been maturing in a fast pace with extensive tools and platforms both open source and commercial available for driving the research and collaboration



Thank You!

Taha Kass-Hout, MD, MS

FDA Chief Health Informatics Officer (CHIO)

FDA Chief Technology Officer (CTO), *Acting*

Director, FDA Office of Informatics and Technology Innovation (OITI)



[@DrTaha_FDA](https://twitter.com/DrTaha_FDA)

Any views or opinions expressed here do not necessarily represent the views of the FDA, HHS, or any other entity of the United States government. Furthermore, the use of any product names, trade names, images, or commercial sources is for identification purposes only, and does not imply endorsement or government sanction by the U.S. Department of Health and Human Services.