Cancer Data Sharing
Genomic Data Commons

Warren Kibbe, PhD

NCI Center for Biomedical Informatics

@wakibbe

NATIONAL CANCER INSTITUTE

The views expressed are my views and not those of NCI

June 21st, 2016

The Genomic Data Commons (GDC) is open!

Vice President Biden visited the GDC and spoke at ASCO on the 6th of June

The New York Times

Biden Unveiling Public Database for Clinical Data on Cancer

FOX NEWS Health

Biden unveiling public database for clinical data on cancer

HealthData Management

NCI launches open access resource to spur cancer research

HOUSTON CHRONICLE

Biden unveils searchable government cancer database

SCIENTIFIC AMERICAN · REUTERS PUBLIC HEALTH

Biden Unveils Major Database to Advance Cancer Research

The Washington Post

Biden unveils launch of major, open-access database to advance cancer research

CHICAGO SUN-TIMES

VP Joe Biden in Chicago to promote Moonshot Initiative vs. cancer

THE CANCER LETTER

Biden Designates NCI's Genomic Data Commons As Foundation of Cancer Moonshot

Daily Mail

New US data system to centralize cancer information

genomeweb

NCI Launches Genomic Data Commons for Cancer Data Sharing

HealthITAnalytics

NIH Launches Genomic Data Commons Supporting Cancer Moonshot

THE HUFFINGTON POST

Biden Announces Crucial Piece Of His Cancer Moonshot Initiative

FORTUNE

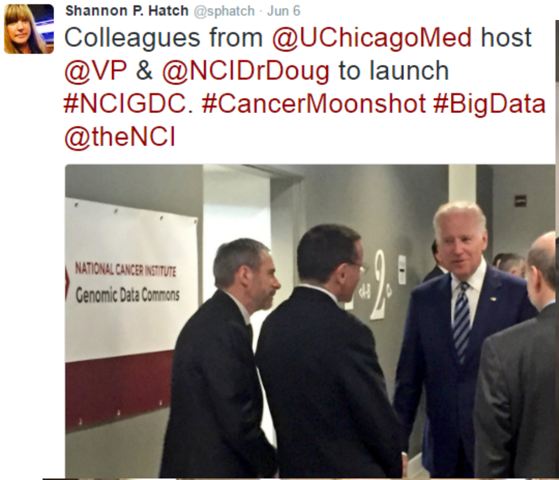
Joe Biden Just Announced a Huge New National Cancer Database

REUTERS

Biden announces U.S. project to promote cancer data sharing

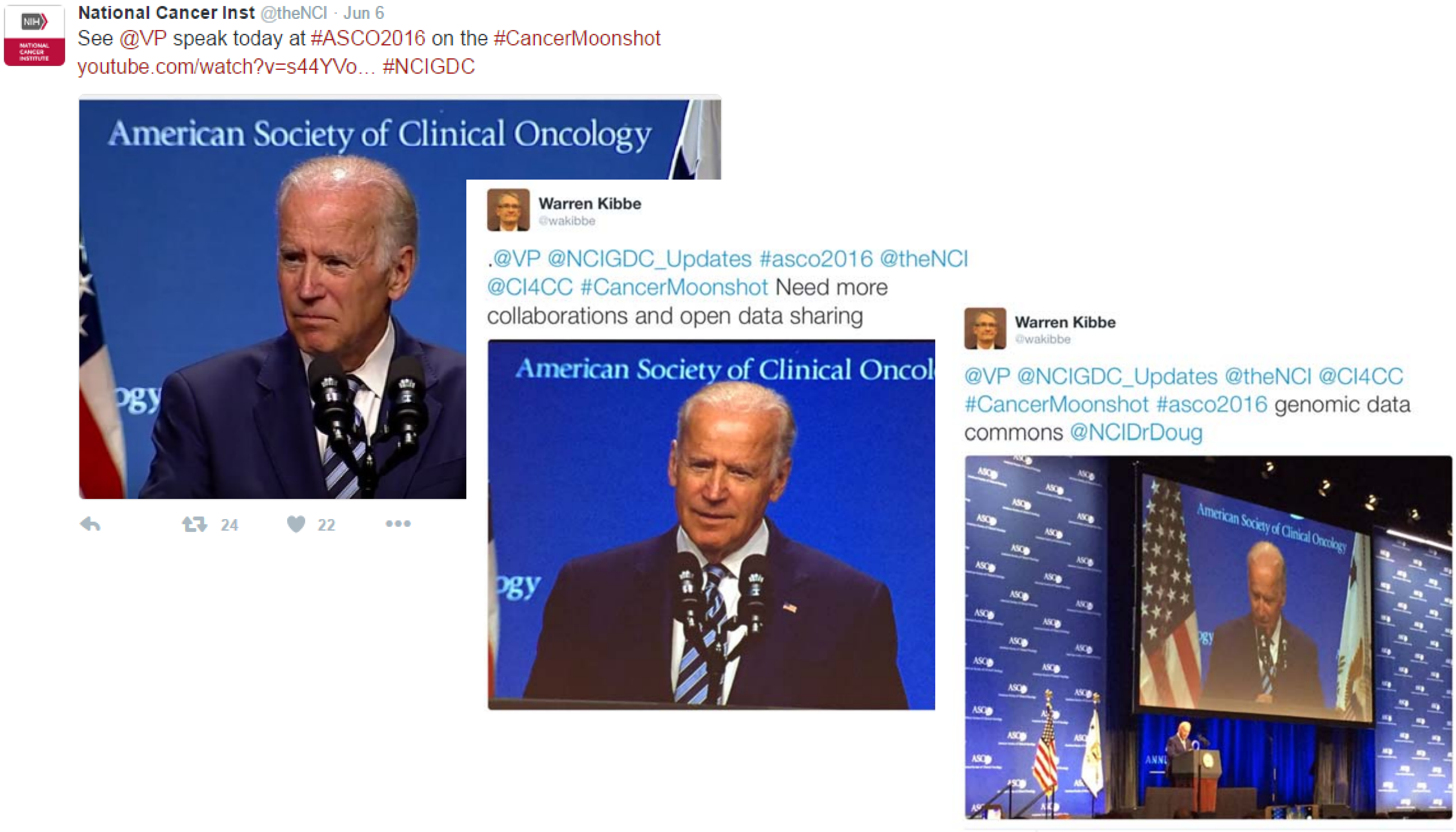
fedscoop

Biden launches data portal to back Cancer Moonshot



Shannon P. Hatch @sphatch · Jun 6

Colleagues from @UChicagoMed host @VP & @NCIDrDoug to launch #NCIGDC. #CancerMoonshot #BigData @theNCI



National Cancer Inst @theNCI · Jun 6

See @VP speak today at #ASCO2016 on the #CancerMoonshot youtube.com/watch?v=s44YVo… #NCIGDC

24 22

Warren Kibbe
@wakibbe

.@VP @NCIGDC_Updates #asco2016 @theNCI @CI4CC #CancerMoonshot Need more collaborations and open data sharing

Warren Kibbe
@wakibbe

@VP @NCIGDC_Updates @theNCI @CI4CC #CancerMoonshot #asco2016 genomic data commons @NCIDrDoug

American Society of Clinical Oncology
American Society of Clinical Oncol
ASCO

Genomic Data Commons

The Cancer Genomic Data Commons (**GDC**) is an existing effort to standardize and simplify submission of genomic data to NCI and follow the principles of **FAIR** – Findable, Accessible, Interoperable, Reusable.

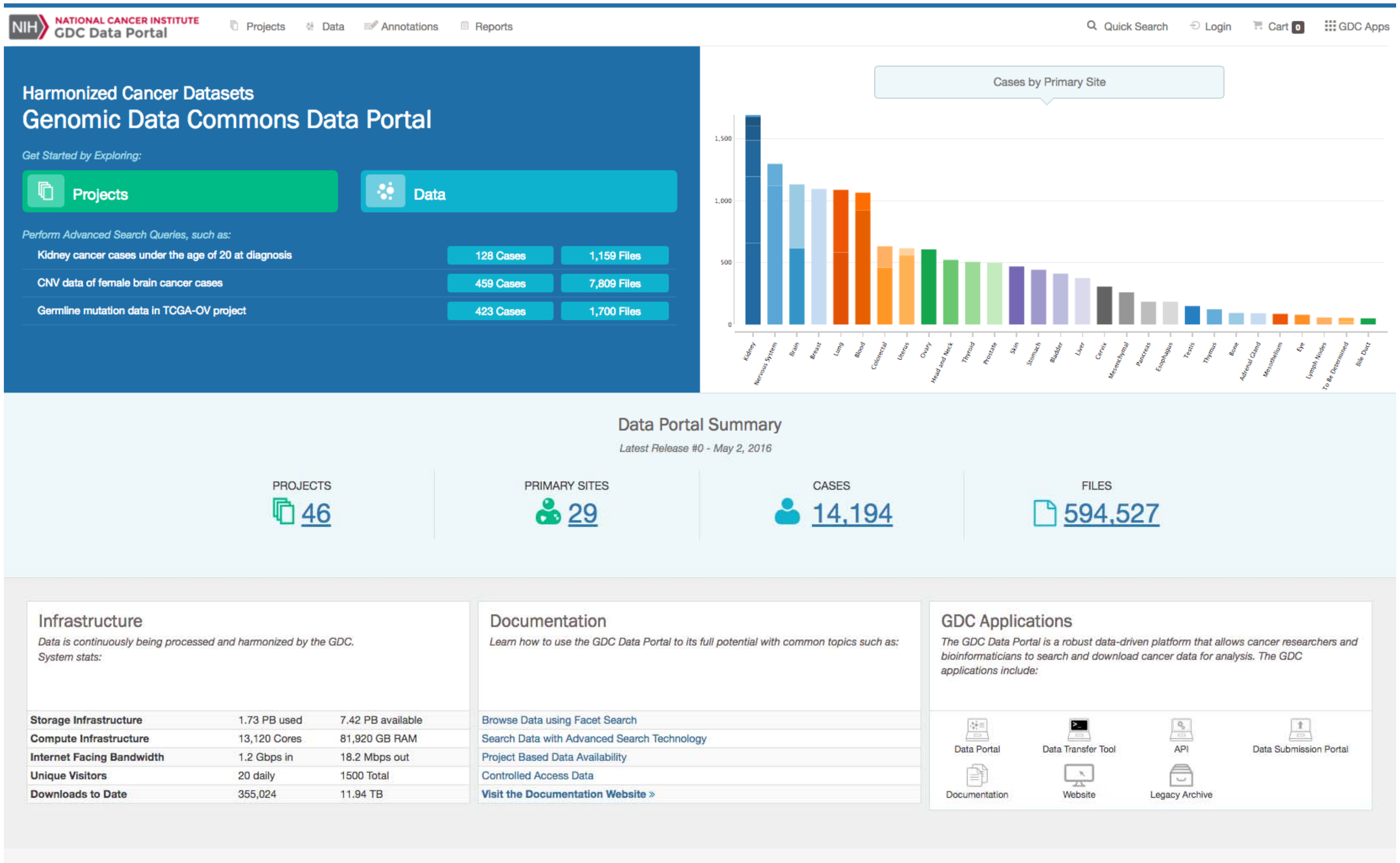
The GDC is part of the NIH Big Data to Knowledge (**BD2K**) initiative and an example of the **NIH Commons**

Microattribution, nanopublications, tracking the use of data, annotation of data, use of algorithms, supports the data /software /metadata life cycle to provide credit and analyze impact of data, software, analytics, algorithm, curation and knowledge sharing

NCI Genomic Data Commons

- The GDC will go live with approximately 4.1 PB of data.
- This includes: 2.6 PB of legacy data;
- and 1.5 PB of “harmonized” data.
- 577,878 files about 14194 cases (patients), in 42 cancer types, across 29 primary sites.
- 10 major data types, ranging from Raw Sequencing Data, Raw Microarray Data, to Copy Number Variation, Simple Nucleotide Variation and Gene Expression.
- Data are derived from 17 different experimental strategies, with the major ones being RNA-Seq, WXS, WGS, miRNA-Seq, Genotyping Array and Expression Array.

Genomic Data Commons Data Portal



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Perform Advanced Search Queries, such as:

Kidney cancer cases under the age of 20 at diagnosis	128 Cases	1,159 Files
CNV data of female brain cancer cases	459 Cases	7,809 Files
Germline mutation data in TCGA-OV project	423 Cases	1,700 Files

Data Portal Summary

Latest Release #0 - May 2, 2016

PROJECTS	PRIMARY SITES	CASES	FILES
46	29	14,194	594,527

Infrastructure

Data is continuously being processed and harmonized by the GDC.
System stats:

Storage Infrastructure	1.73 PB used	7.42 PB available
Compute Infrastructure	13,120 Cores	81,920 GB RAM
Internet Facing Bandwidth	1.2 Gbps in	18.2 Mbps out
Unique Visitors	20 daily	1500 Total
Downloads to Date	355,024	11.94 TB

Documentation

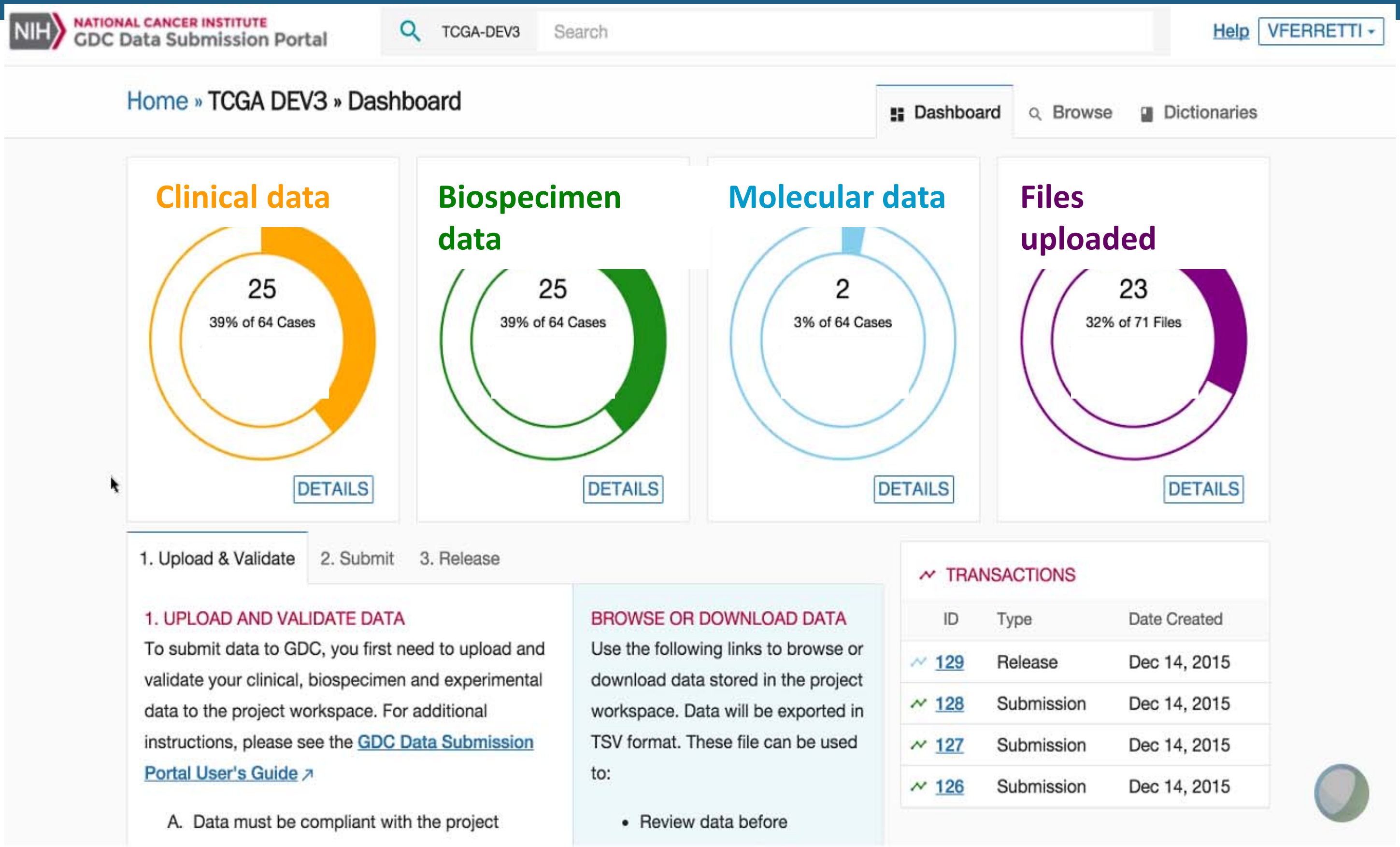
Learn how to use the GDC Data Portal to its full potential with common topics such as:

- Browse Data using Facet Search
- Search Data with Advanced Search Technology
- Project Based Data Availability
- Controlled Access Data
- **Visit the Documentation Website »**

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

The NCI Genomic Data Commons User Interface
Data Submission Dashboard



Clinical data

Biospecimen data

Molecular data

Files uploaded

1. UPLOAD AND VALIDATE DATA

To submit data to GDC, you first need to upload and validate your clinical, biospecimen and experimental data to the project workspace. For additional instructions, please see the GDC Data Submission Portal User's Guide ↗

A. Data must be compliant with the project

BROWSE OR DOWNLOAD DATA

Use the following links to browse or download data stored in the project workspace. Data will be exported in TSV format. These file can be used to:

- Review data before

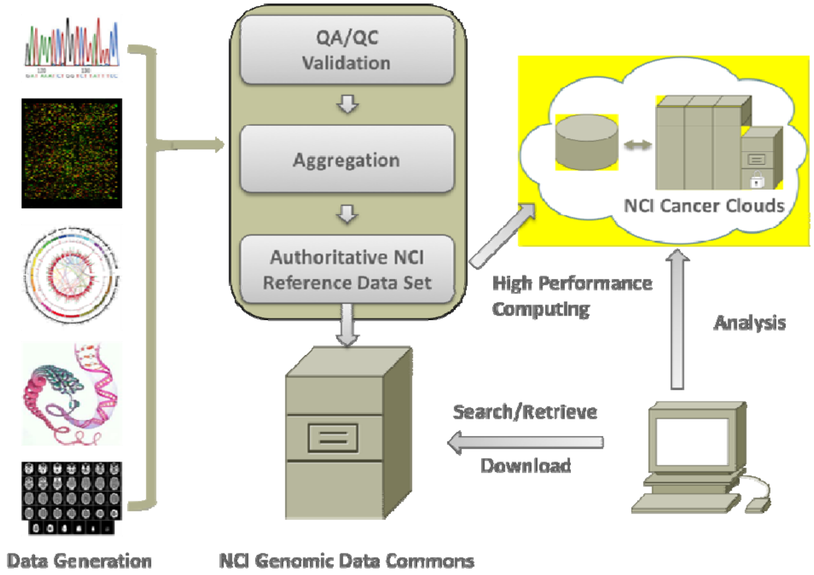
TRANSACTIONS

ID	Type	Date Created
129	Release	Dec 14, 2015
128	Submission	Dec 14, 2015
127	Submission	Dec 14, 2015
126	Submission	Dec 14, 2015

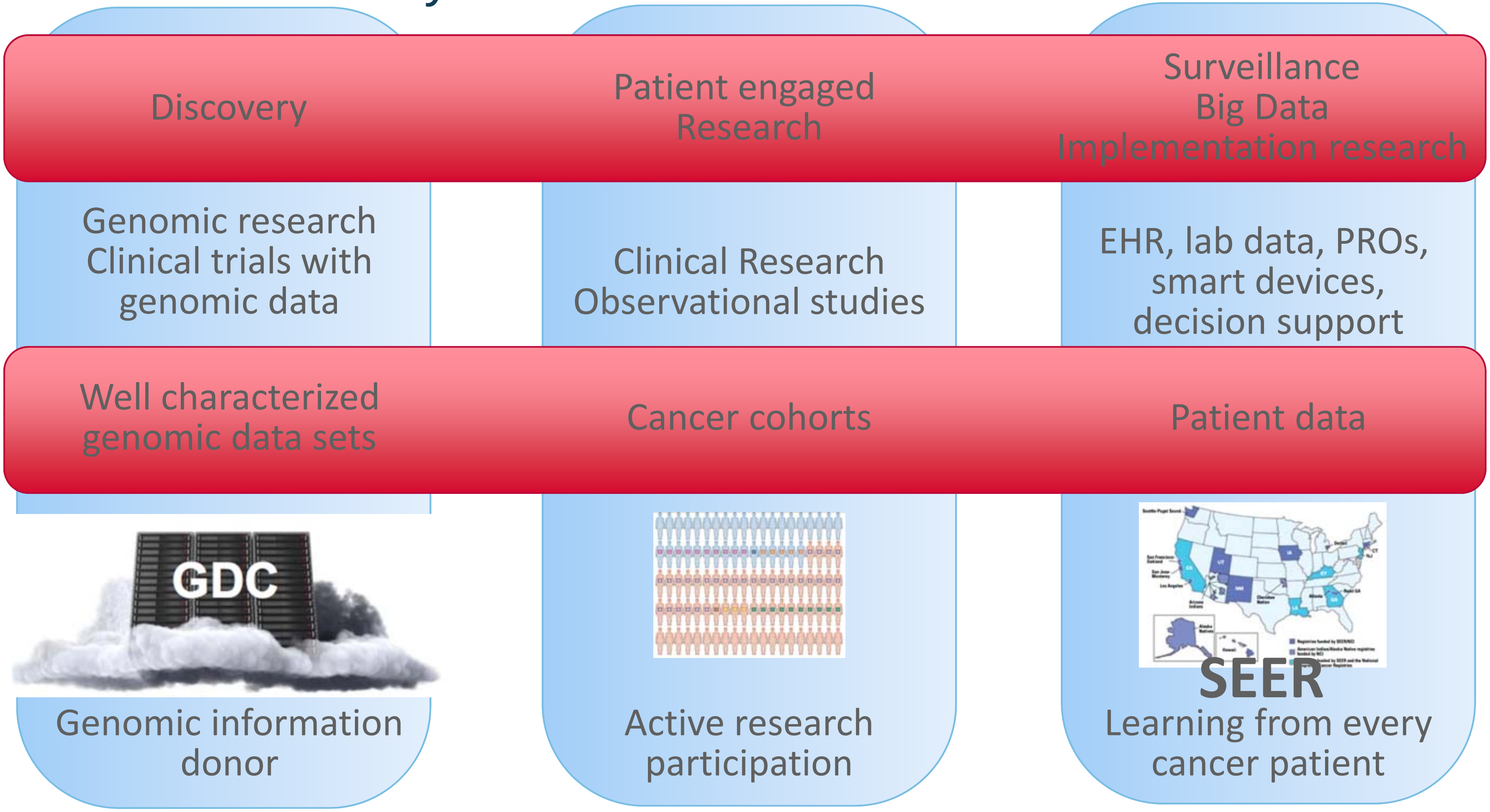
Support the Precision Medicine Initiative

The Genomic Data Commons and Cloud Pilots

- Expand data model to include other data (e.g. imaging and proteomics)
- Allow easy publication of persistent links to data, annotations, algorithms, tools, workflows
- Measure usage and impact
- Change incentives for public contributions



Cancer data ecosystem



Discovery

Patient engaged Research

Surveillance
Big Data
Implementation research

Genomic research
Clinical trials with genomic data

Clinical Research
Observational studies

EHR, lab data, PROs, smart devices, decision support

Well characterized genomic data sets

Cancer cohorts

Patient data

Genomic information donor

Active research participation

Learning from every cancer patient