The NIH Data Commons

NHGRI Council – February 6, 2017

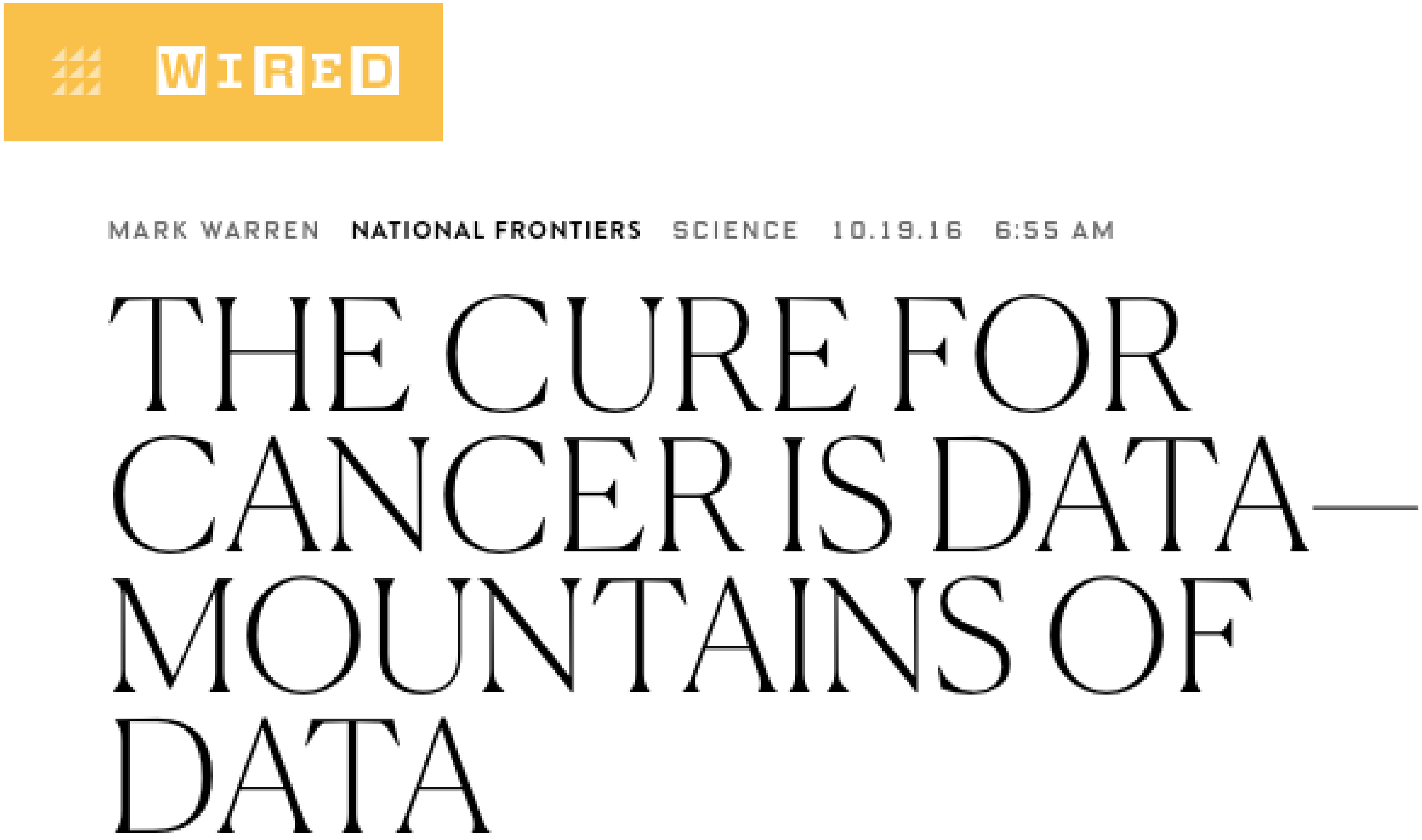
Vivien Bonazzi Ph.D.

Senior Advisor for Data Science & Data Commons
National Institutes of Health, Bethesda

What’s the driving the need for a Data Commons?

Convergence of factors

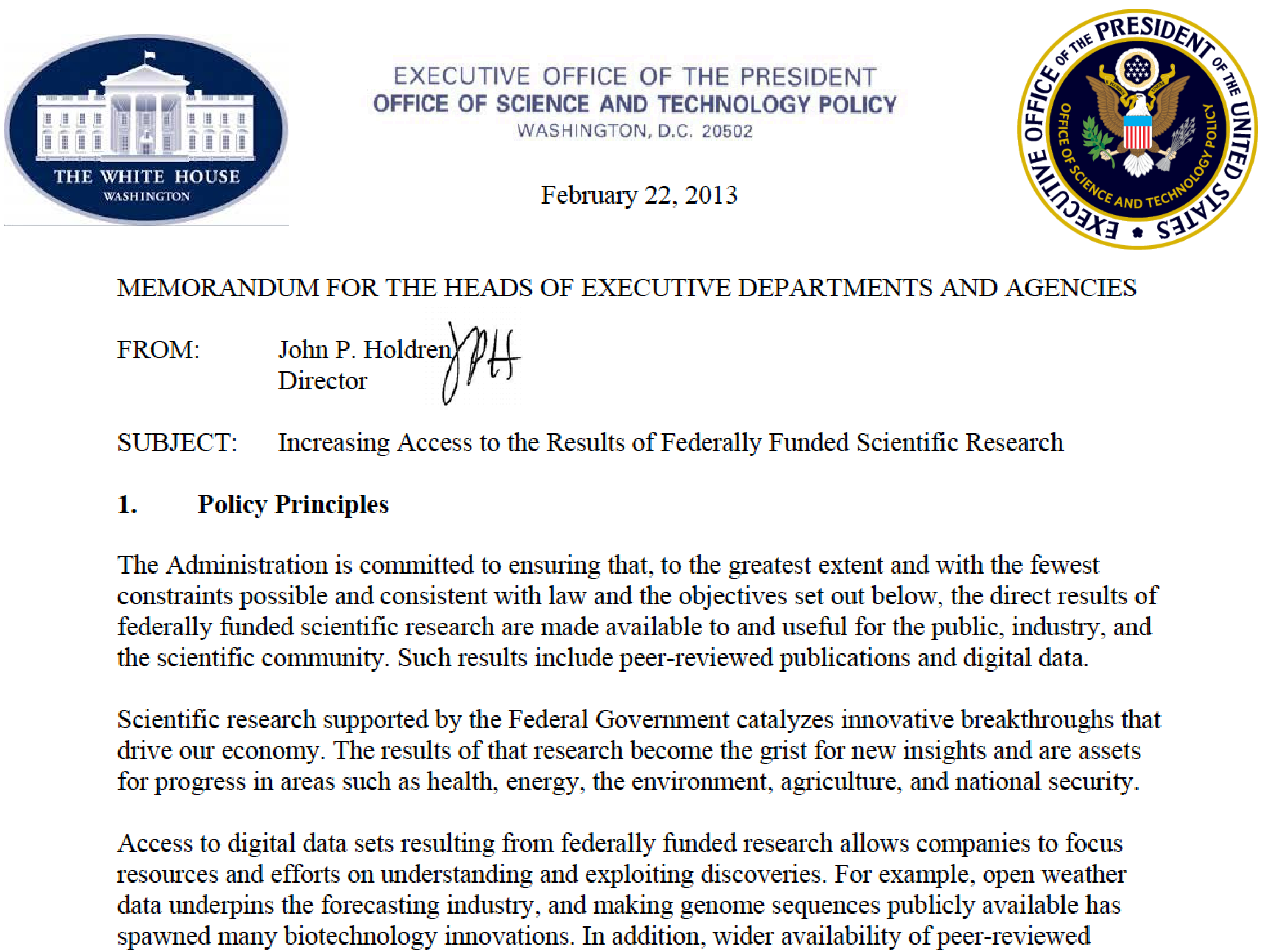
- **Mountains of Data**
- **Increasing need and support for Data sharing**
- **FAIR – *Findable Accessible Interoperable Reproducible***
- **Availability of digital technologies and infrastructures that support Data at scale**



WIRED

MARK WARREN NATIONAL FRONTIERS SCIENCE 10.19.16 6:55 AM

THE CURE FOR CANCER IS DATA—MOUNTAINS OF DATA



THE WHITE HOUSE
WASHINGTON

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren JPH
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed

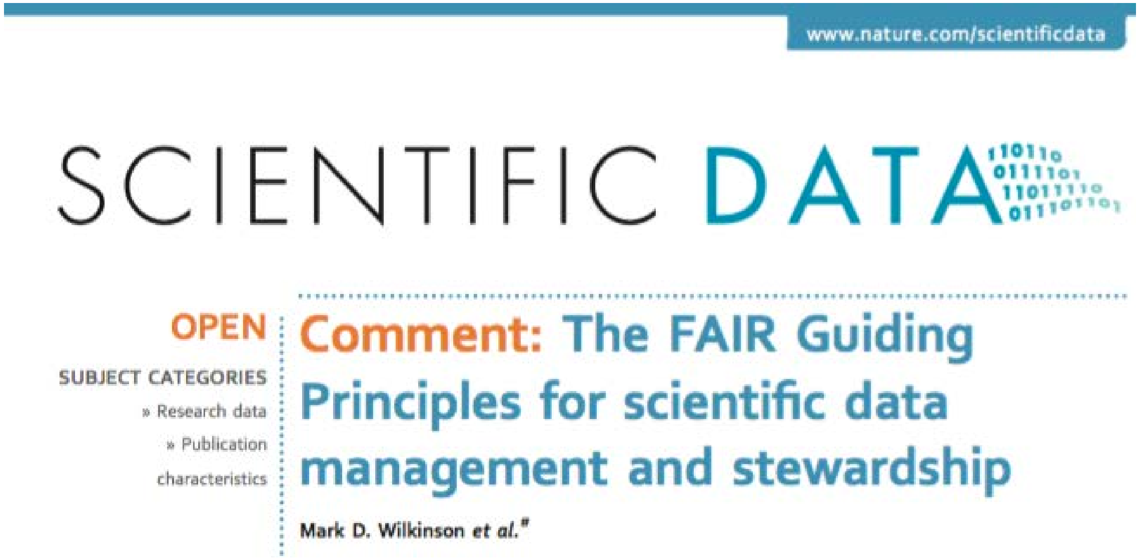
U.S.Department of Health & Human Services

NIH Genomic Data Sharing (GDS)

https://gds.nih.gov/
Went into effect January 25, 2015

NCI guidance:
http://www.cancer.gov/grants-training/grants-management/nci-policies/genomic-data

Requires public sharing of genomic data sets



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

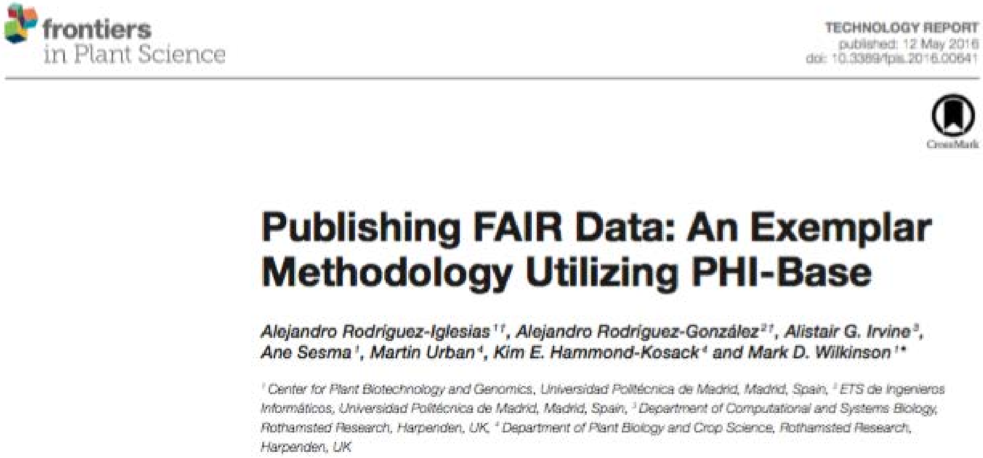
SUBJECT CATEGORIES

» Research data

» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#



frontiers in Plant Science

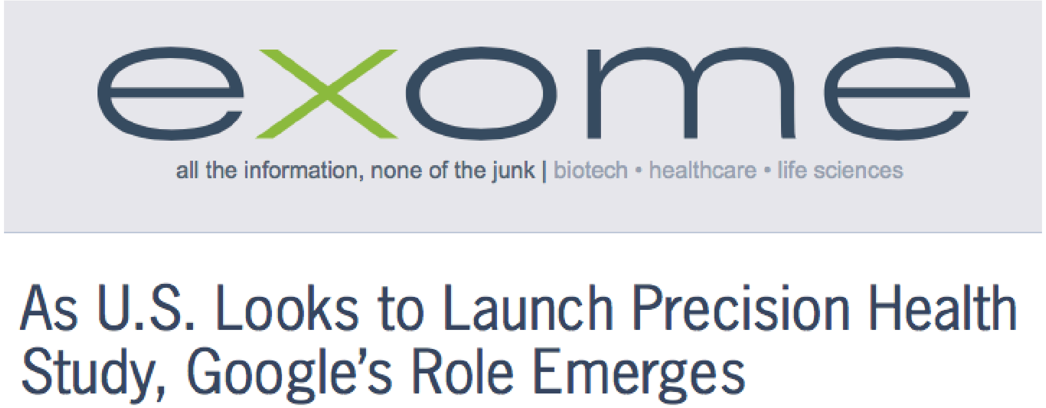
TECHNOLOGY REPORT
published: 12 May 2016
doi: 10.3389/fpls.2016.00641

CrossMark

Publishing FAIR Data: An Exemplar Methodology Utilizing PHI-Base

Alejandro Rodríguez-Iglesias[1†], *Alejandro Rodríguez-González*[2†], *Alistair G. Irvine*[3], *Ane Sesma*[1], *Martin Urban*[4], *Kim E. Hammond-Kosack*[4] *and Mark D. Wilkinson*[1*]

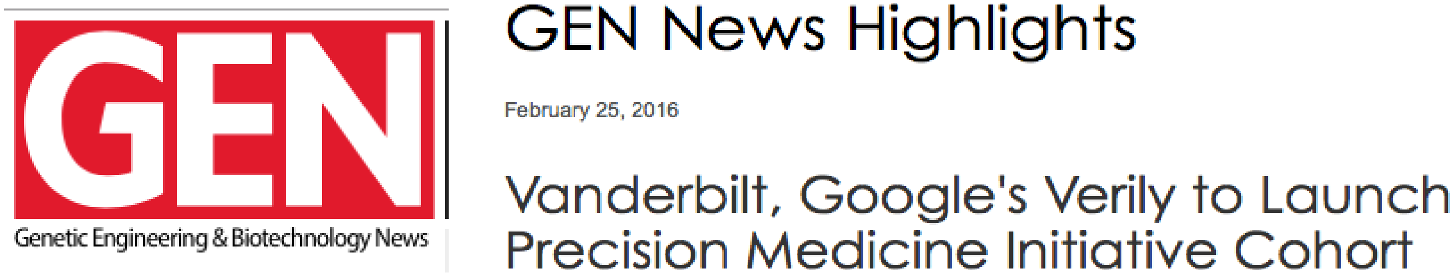
[1] Center for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid, Madrid, Spain, [2] ETS de Ingenieros Informáticos, Universidad Politécnica de Madrid, Madrid, Spain, [3] Department of Computational and Systems Biology, Rothamsted Research, Harpenden, UK, [4] Department of Plant Biology and Crop Science, Rothamsted Research, Harpenden, UK



exome

all the information, none of the junk | biotech • healthcare • life sciences

As U.S. Looks to Launch Precision Health Study, Google's Role Emerges



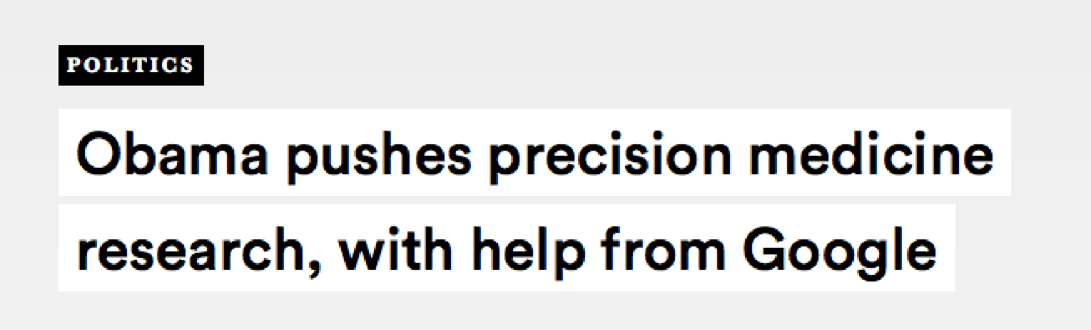
GEN

Genetic Engineering & Biotechnology News

GEN News Highlights

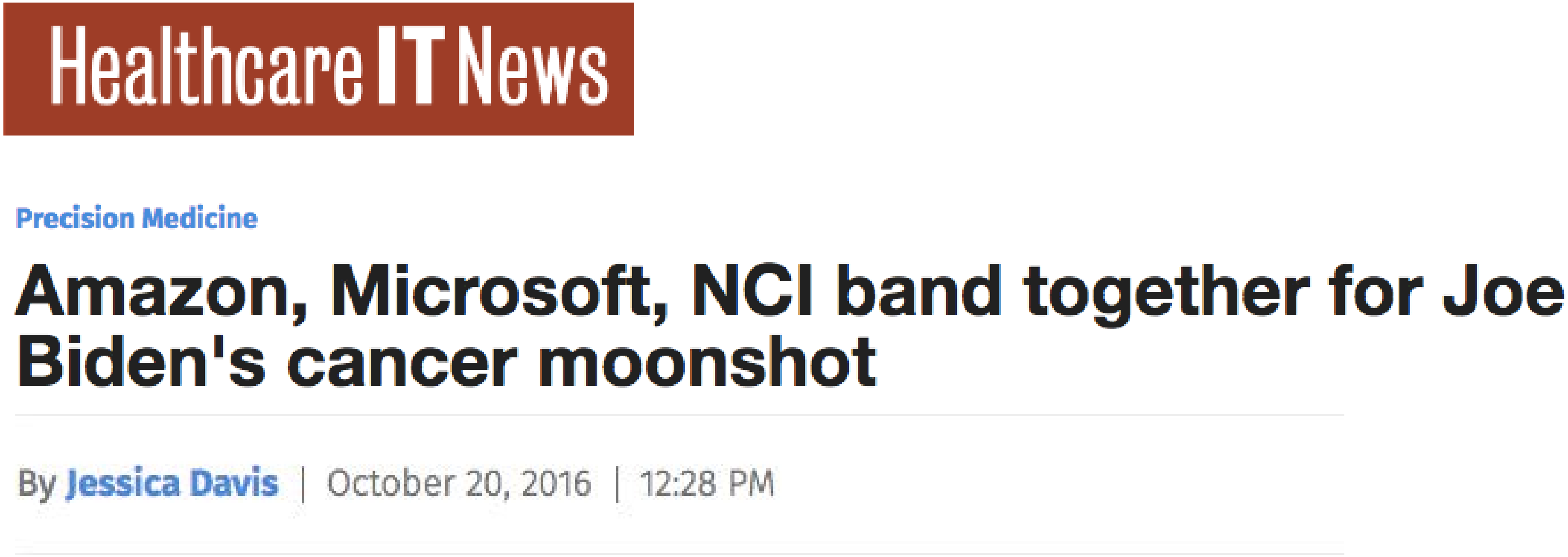
February 25, 2016

Vanderbilt, Google's Verily to Launch Precision Medicine Initiative Cohort



POLITICS

Obama pushes precision medicine research, with help from Google



Healthcare IT News

Precision Medicine

Amazon, Microsoft, NCI band together for Joe Biden's cancer moonshot

By Jessica Davis | October 20, 2016 | 12:28 PM

Data Commons
enabling data driven science

Enable investigators to leverage **all possible** data and tools in the effort to accelerate biomedical discoveries, therapies and cures

by

driving the development of **data infrastructure** and **data science capabilities** through collaborative research and robust engineering

Developing a *Data Commons*

- Treats products of research – data, methods, papers etc. as *digital objects*
- These digital objects exist in a shared virtual space
  - Find, Deposit, Manage, Share, and Reuse data, software, metadata and workflows
- Digital object compliance through FAIR principles:
  - **F**indable
  - **A**ccessible (*and usable)*
  - **I**nteroperable
  - **R**eusable

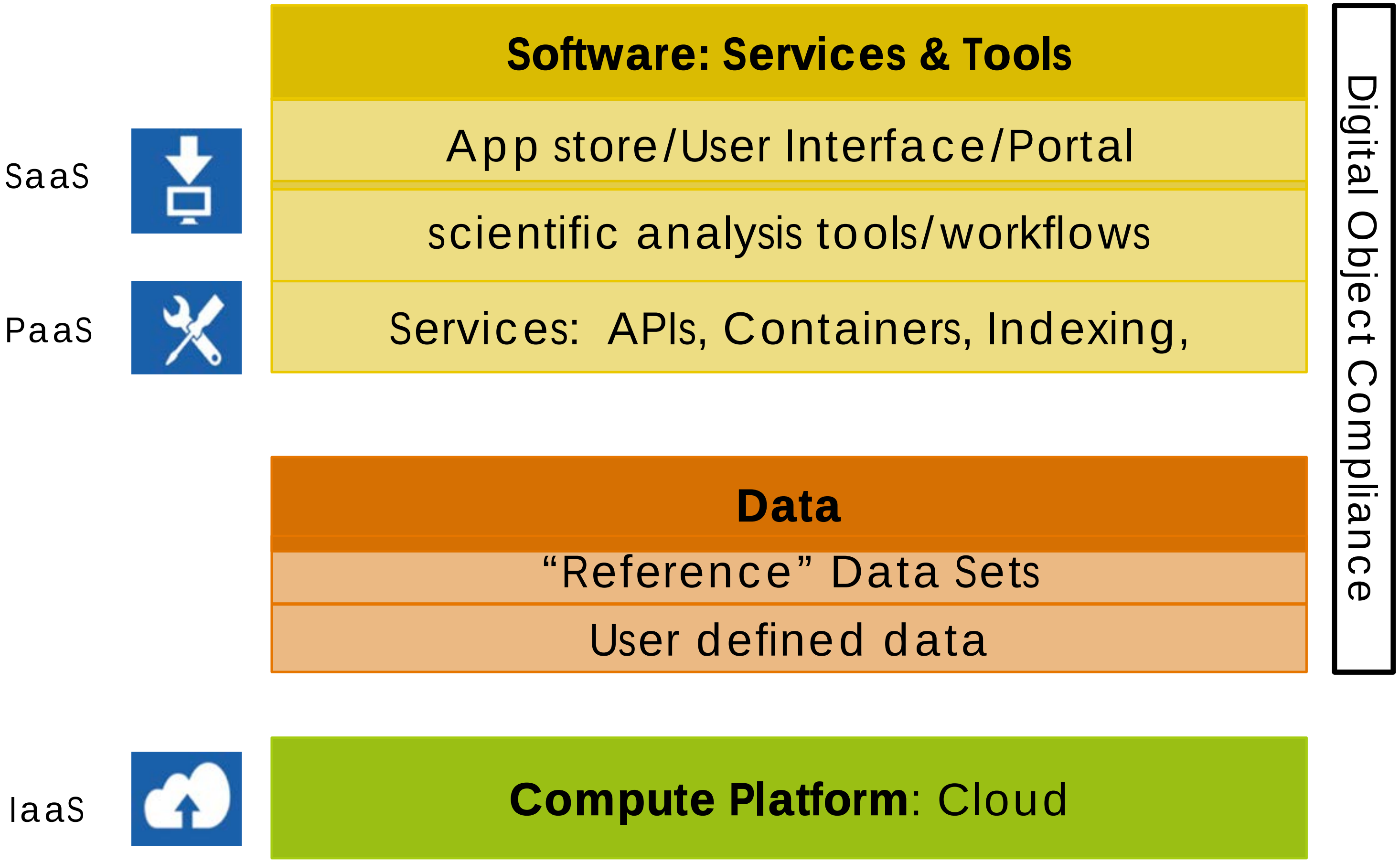
The Data Commons

is a platform

that allows transactions to

occur on FAIR data at scale

The Data Commons Platform



Software: Services & Tools

App store/User Interface/Portal

scientific analysis tools/workflows

Services: APIs, Containers, Indexing,

Data

“Reference” Data Sets

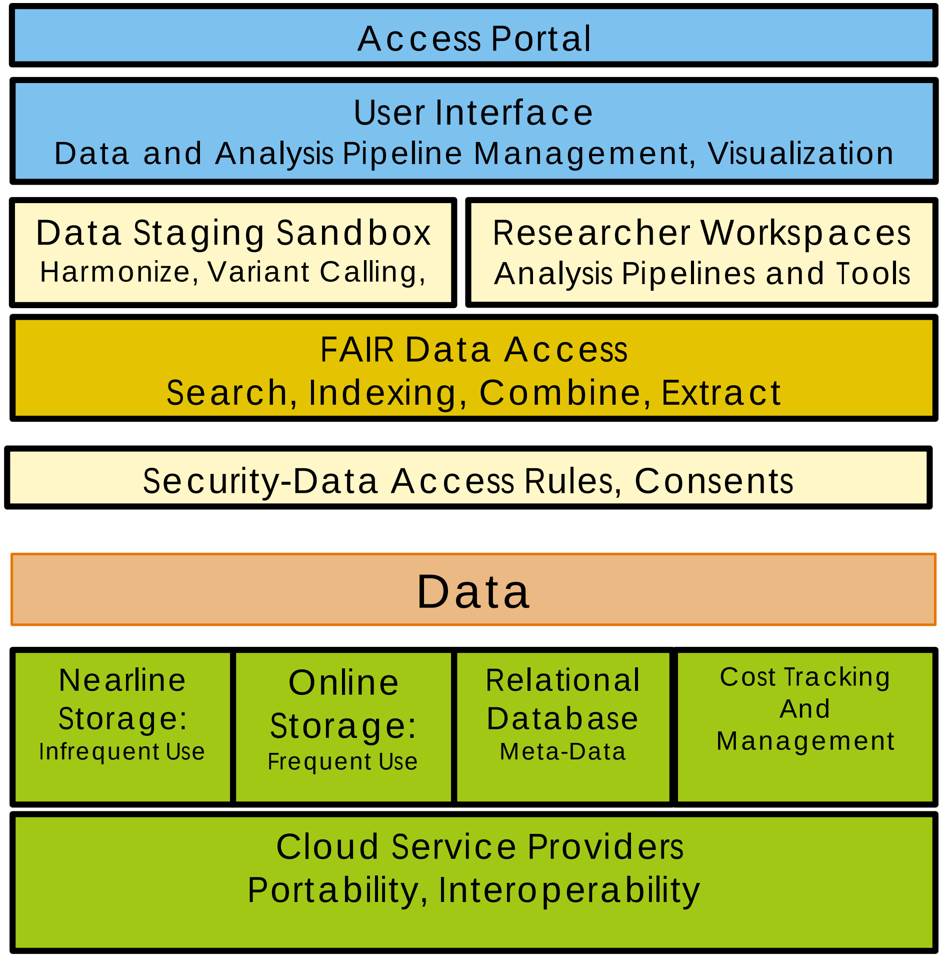
User defined data

Compute Platform: Cloud

Digital Object Compliance

https://datascience.nih.gov/commons

Commons Architecture



Access Portal

User Interface
Data and Analysis Pipeline Management, Visualization

Data Staging Sandbox
Harmonize, Variant Calling,

Researcher Workspaces
Analysis Pipelines and Tools

FAIR Data Access
Search, Indexing, Combine, Extract

Security-Data Access Rules, Consents

Data

Nearline Storage:
Infrequent Use

Online Storage:
Frequent Use

Relational Database
Meta-Data

Cost Tracking And Management

Cloud Service Providers
Portability, Interoperability

Other Data Commons’

GDC NCI Genomic Data Commons

NEW YORK GENOME CENTER®
commons

FRED HUTCH
CURES START HERE®

HUTCH DATA COMMONWEALTH

NIH NEPHELE
Microbiome analysis without boundaries.

UC SANTA CRUZ
UCSC Cloud Commons Infrastructure

Other Data Commons'

GD NCI Genomic Data Commons

FRED HUTCH
CURES START HERE®

HUTCH DATA COMMONWEALTH

NIH NEPHELE
Microbiome analysis without boundaries.

NEW YORK
GENOME CENTER®
commons

THE UNIVERSITY OF CALIFORNIA 1868
UC SANTA CRUZ

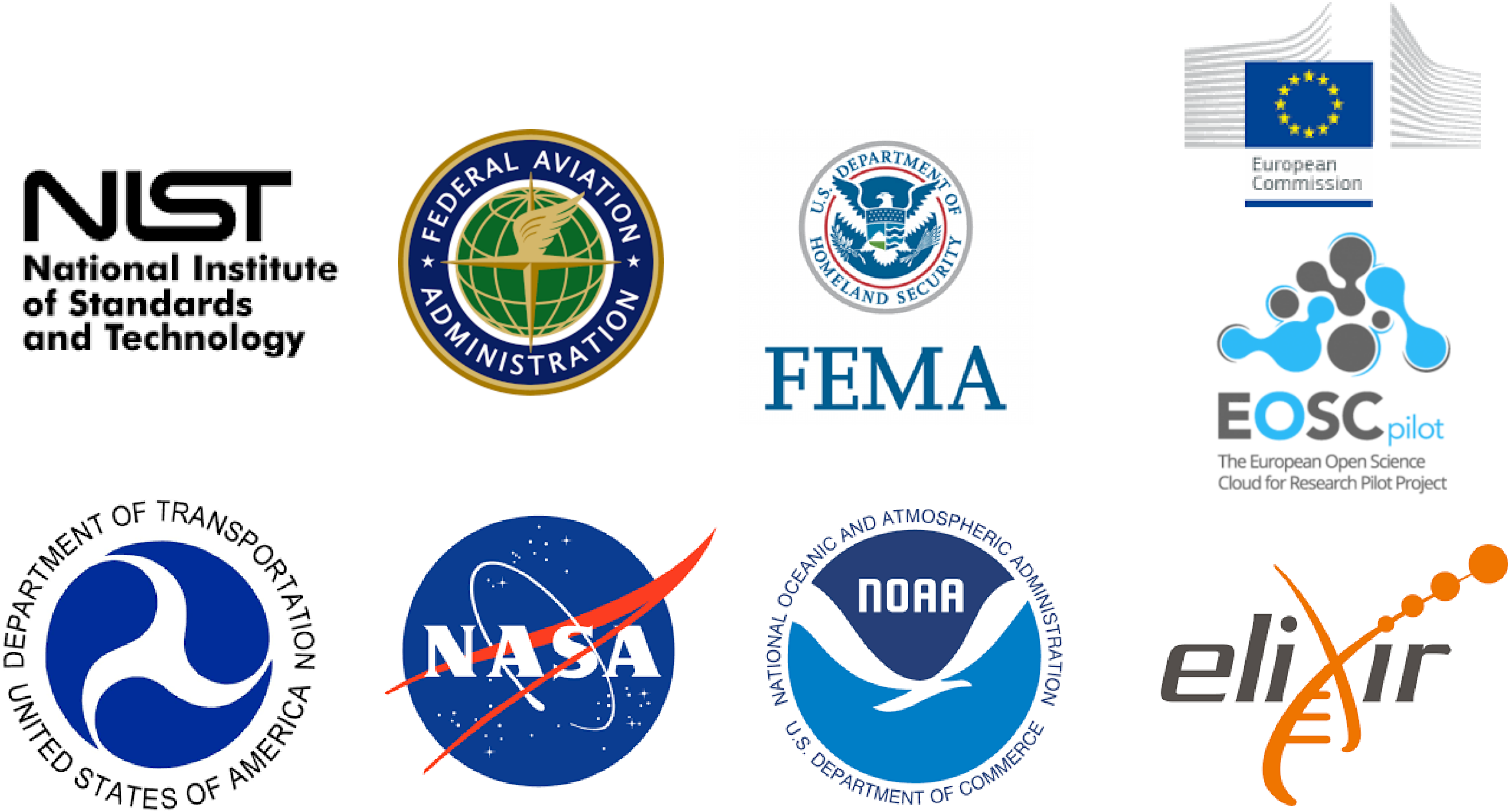
NIH National Heart, Lung, and Blood Institute

NIH
The Common Fund

NIH National Human Genome Research Institute

Commons Engagement

US Government Agencies & EU groups


NIST
National Institute of Standards and Technology
FEDERAL AVIATION ADMINISTRATION
U.S. DEPARTMENT OF HOMELAND SECURITY
FEMA
European Commission
EOSCpilot
The European Open Science Cloud for Research Pilot Project
DEPARTMENT OF TRANSPORTATION
UNITED STATES OF AMERICA
NASA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE
elixir

Interoperability with other Commons'

- **Common goals – democratizing, collaborating & sharing data**
- **Reuse of currently available open source tools which support interoperability**
  - **GA4GH, UCSC, GDC, NYGC**
  - **Planned meeting for current major Commons developers/NIH Staff**
  - **BioIT Commons Session?**
- **Shared open standard APIs for data access and computing**
- **Ability to deploy and compute across multiple cloud environments**
- **Docker containers – Dockerstore/Docker registry**
- **Workflows management, sharing and deployment**
- **Discoverability (indexing) objects across cloud commons**
- **Global Unique identifiers**
- **NIH Commons Working Groups: BD2K, ELIXR members & broader community**
  - **Commons FAIRness metrics WG:**
  - **Interoperable APIs**
  - **Docker registry /workflow sharing**
  - **Data Object registries**
- **Common user authentication system**

Acknowledgments

ADDS Office: Jennie Larkin, Phil Bourne, Michelle Dunn, Mark Guyer, Allen Dearry, Sonynka Ngosso, Tonya Scott, Lisa Dunneback, Vivek Navale (CIT/ADDS), Ron Margolis

NCBI: George Komatsoulis

NHGRI: Valentina di Francesco, Ajay Pillai, Ken Wiley

NIGMS: Susan Gregurick

CIT: Andrea Norris, Debbie Sinmao

NIH Common Fund: Jim Anderson , Betsy Wilder, Leslie Derr

NCI: Ian Fore, Sean Davis, Warren Kibbe, Tony Kerlavage, Tanja Davidsen

NIAID: Maria Giovanni, Alison Yao, Eric Choi, Claire Schulkey

NHLBI: Weiniu Gan, Alastair Thomson

NIH Clinical Centre: Elaine Ayres, (BITRIS),

NIBIB: Vinay Pai (DK),

OSP: Dina Paltoo, Kris Langlais, Erin Luetkemeier, Agnes Rooke,

Research and Industry: Mathew Trunnell (FHC), Bob Grossman (Chicago), Toby Bloom (NYGC)

Vivien Bonazzi

bonazziv@mail.nih.gov

Stay in Touch

QR Business Card

LinkedIn

Slideshare

@Vivien.Bonazzi

Blog
(Coming soon!)