Introduction to Biomedical Data Science

Aik Choon Tan

8/31/2018

aikchoon.tan@ucdenver.edu

http://tanlab.ucdenver.edu/labHomePage/teaching/BSBT6111/

Outline

- Course outline
- https://www.youtube.com/watch?v=F6CI7jXHGWg
- Introduction to Data Science
- Case Study

Learning Objectives

At the completion of this course, students will be able to understand:

- The characteristics and challenges of big data science
- Understand the characteristics of various biomedical sources
- Understand data analytics and their usages in biomedical data science
- Understand data repurposing concepts and techniques
- Understand the concept of data sharing and data reproducibility concepts
- Understand of Data Visualization and visual analytics for data representation, presentation, exploration, and manipulation

Grading

1. Participation and presentation (50% of course grade)

   – This is a high level didactic course with minimal student participation. However, student understanding of the materials will be measured occasionally from in-class interaction.

2. Quizzes (50% of course grade)

   – There are multiple quizzes for the student to be graded for the final grade for the course

Classes

1. INTRODUCTION TO BIOMEDICAL DATA SCIENCE – 8/31/18
2. BIOMEDICAL DATA SOURCES AND INTEGRATION - 9/7/18
3. DATA MINING AND ANALYTICS - 9/14/18
4. DATA MINING AND ANALYTICS II - 9/21/18
5. GUEST LECTURE - Clinical Informatics - MINING ELECTRONIC HEALTH RECORDS - Michael Ames, Associate Director, COMPASS - 9/28/18
6. GUEST LECTURE - MINING CLINICAL DATA IN VA HEALTH SYSTEMS - Thomas Glorioso, Senior Data Scientist, VA - 10/5/18
7. GUEST LECTURE - DATA & METHODS REPRODUCIBILITY - Wladimir Labeikovsky, Bioinformationist, HSC Library - 10/12/18
8. DATA VISUALIZATION - 10/19/18
9. GUEST LECTURE - INTRODUCTION TO BIOMEDICAL TEXT MINING - Dr. Kevin Cohen, Group Leader, Dept. Pharmacology - 11/2/18
10. DATA REPURPOSING AND CONCLUSIONS - 11/9/18
11. FINAL - 11/16/18

Buzzwords: Data Science

DATA ANALYTICS

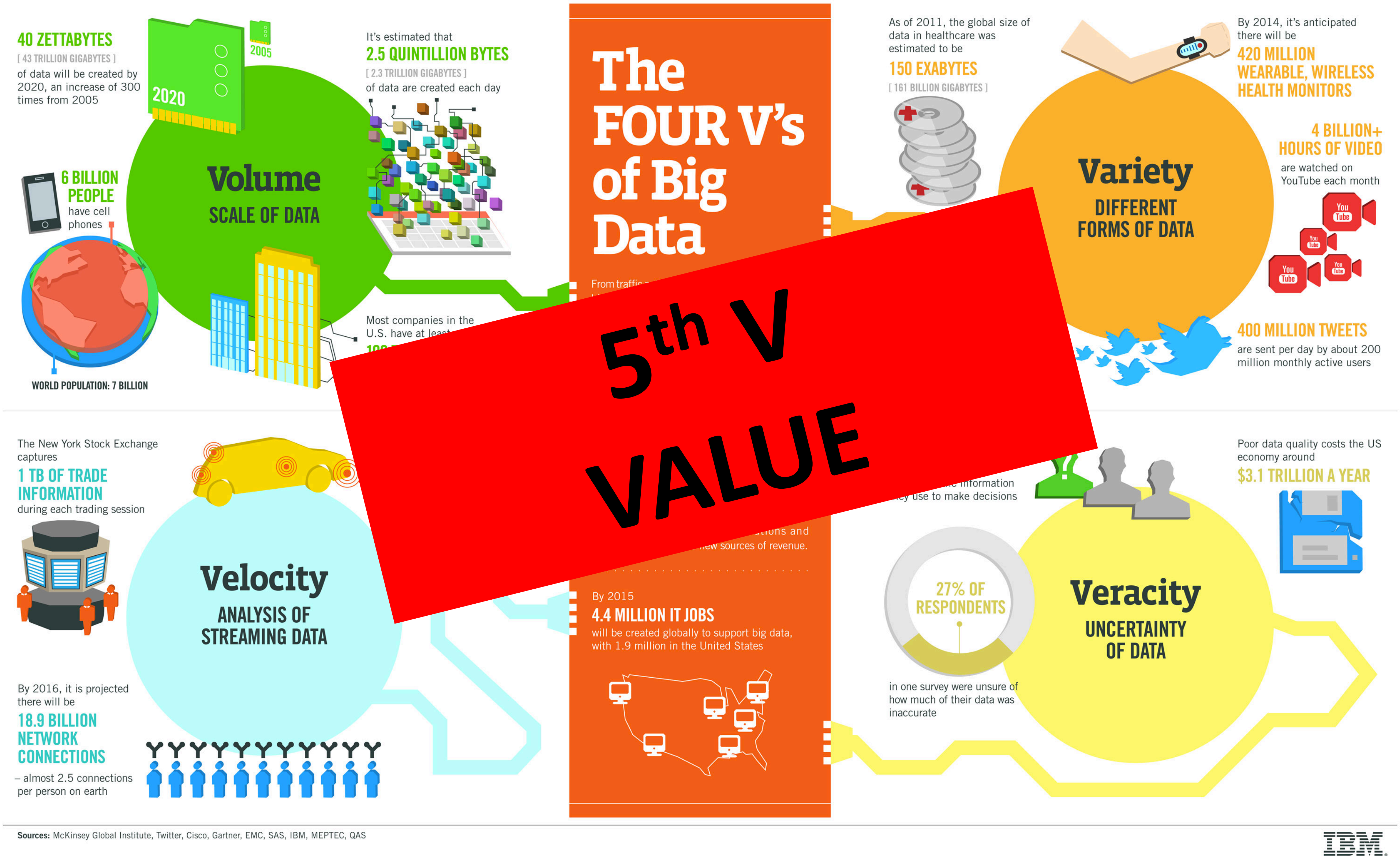
BIG DATA

THE CLOUD

What do they mean?

And how can they help our “Business”

BIG DATA



5th V
VALUE

The FOUR V's of Big Data

From traffic

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that 2.5 QUINTILLION BYTES [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session

By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth

tions and new sources of revenue.

By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States

Veracity

UNCERTAINTY OF DATA

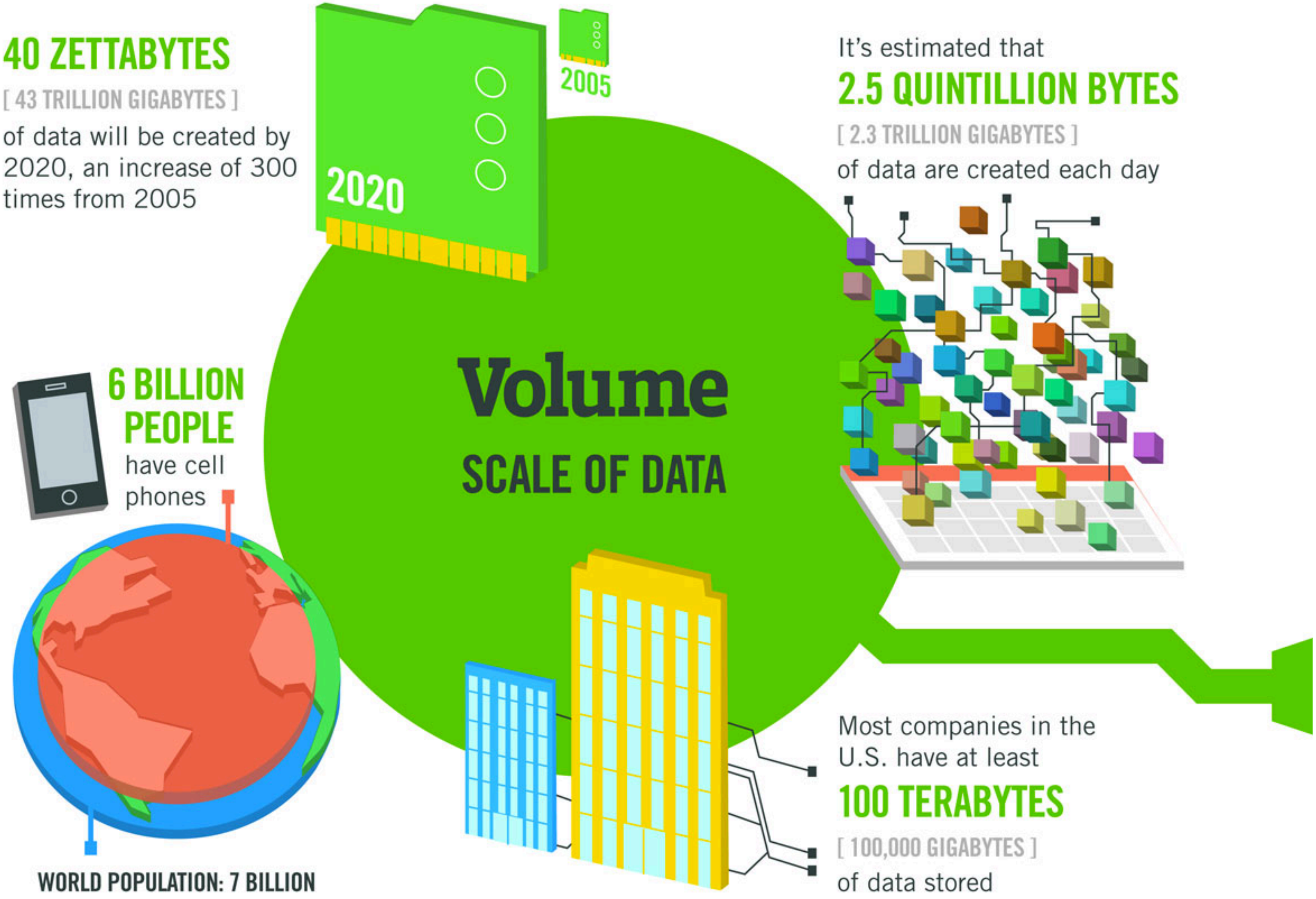
e information ey use to make decisions

Poor data quality costs the US economy around $3.1 TRILLION A YEAR

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

V = Volume



40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

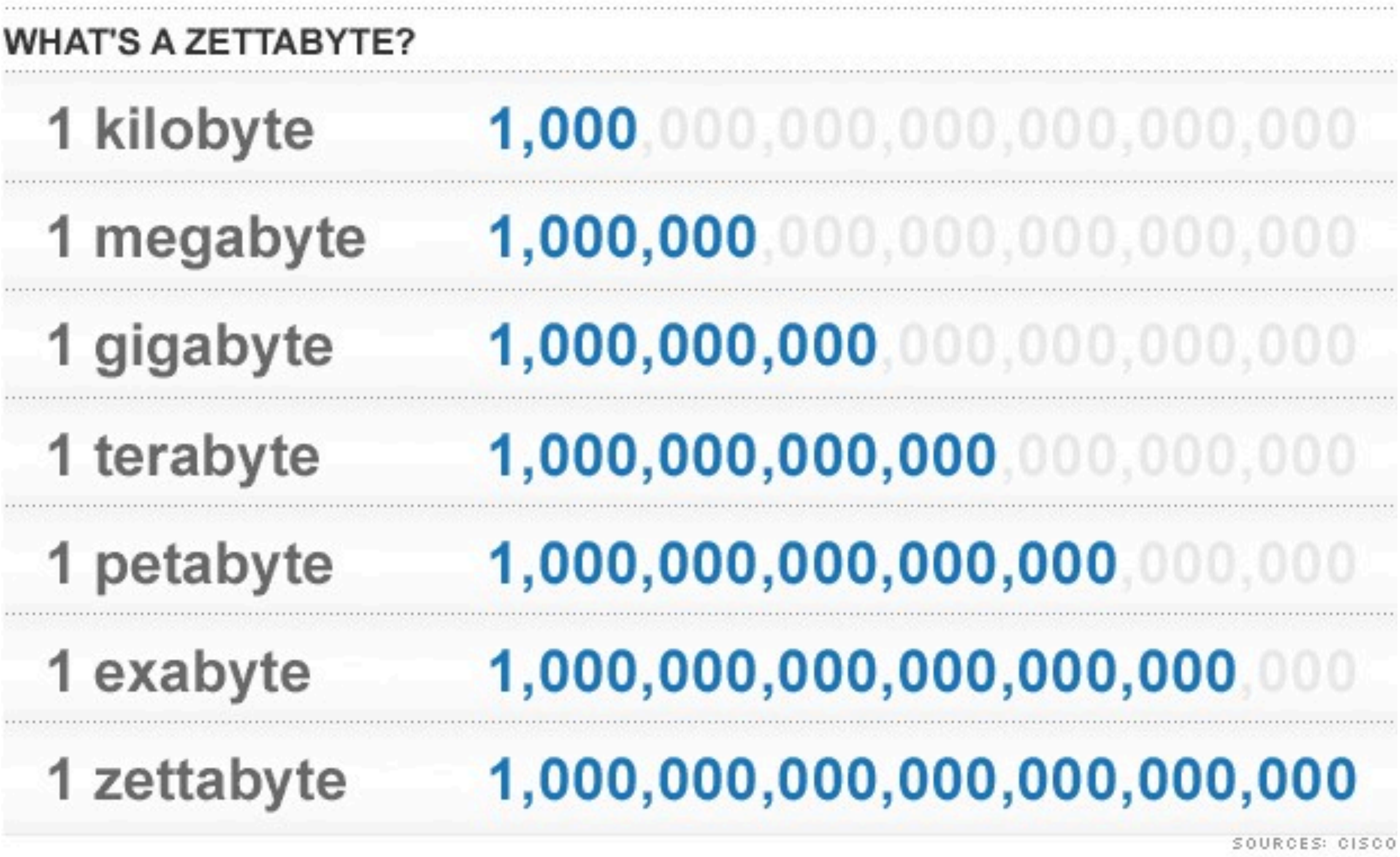
Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

V = Volume

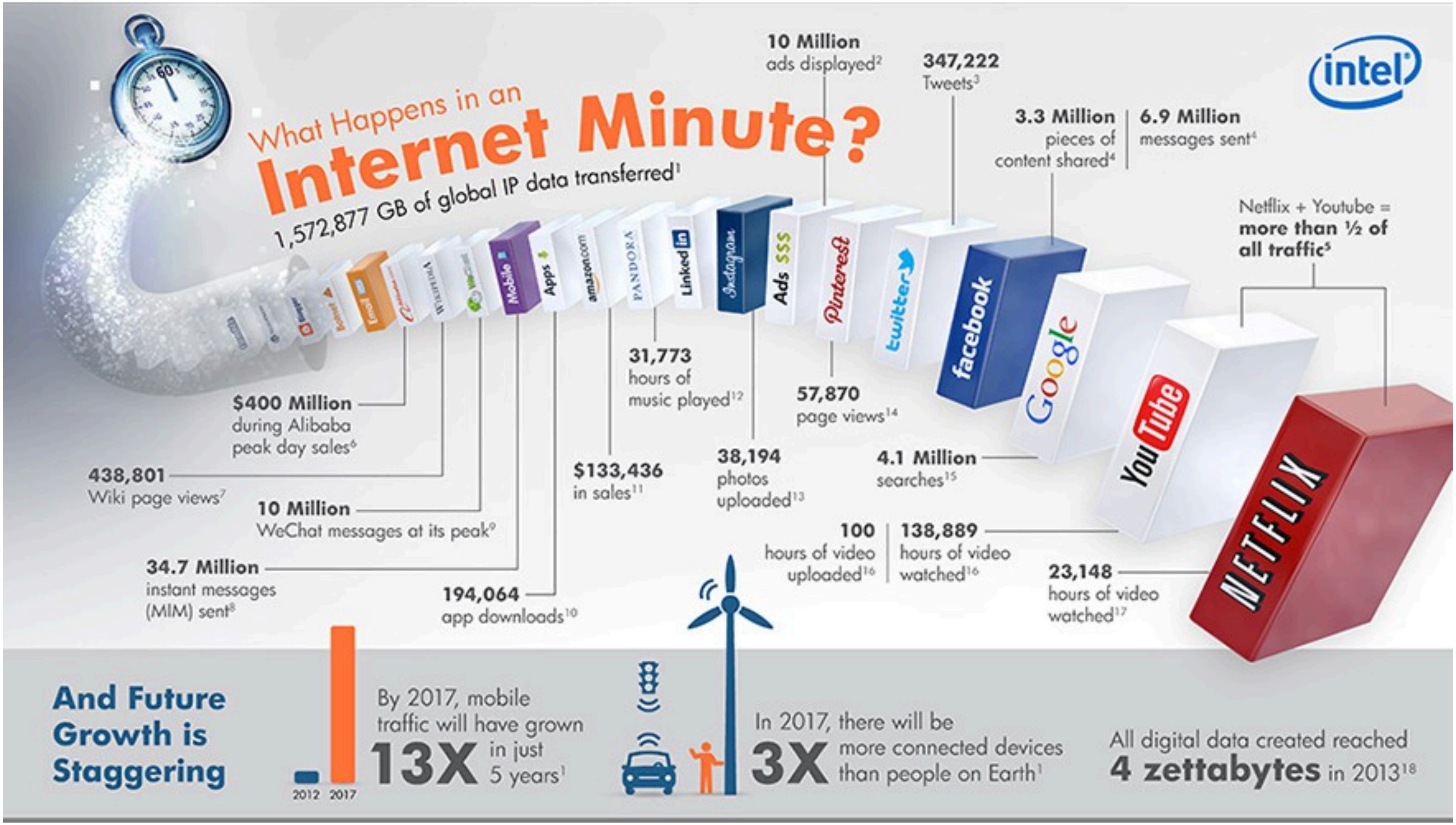


WHAT'S A ZETTABYTE?

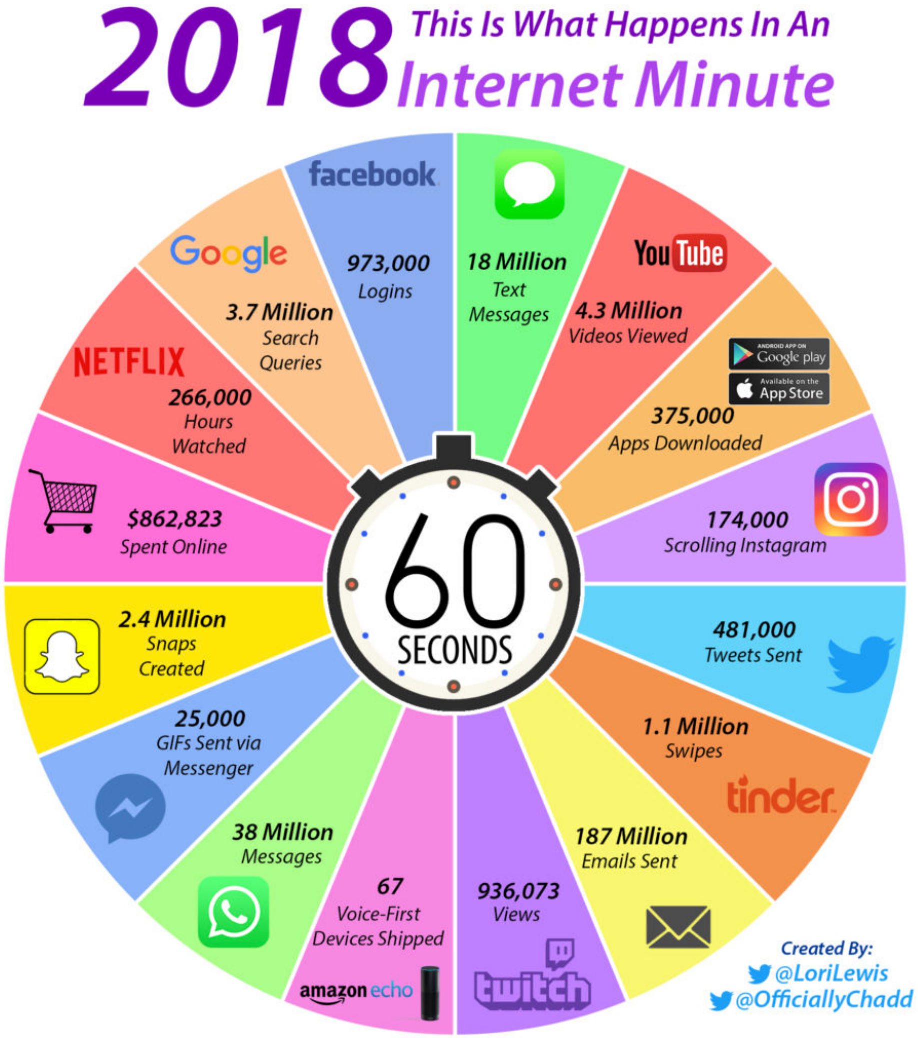
1 kilobyte	**1,000**,000,000,000,000,000,000
1 megabyte	**1,000,000**,000,000,000,000,000
1 gigabyte	**1,000,000,000**,000,000,000,000
1 terabyte	**1,000,000,000,000**,000,000,000
1 petabyte	**1,000,000,000,000,000**,000,000
1 exabyte	**1,000,000,000,000,000,000**,000
1 zettabyte	**1,000,000,000,000,000,000,000**

SOURCES: CISCO

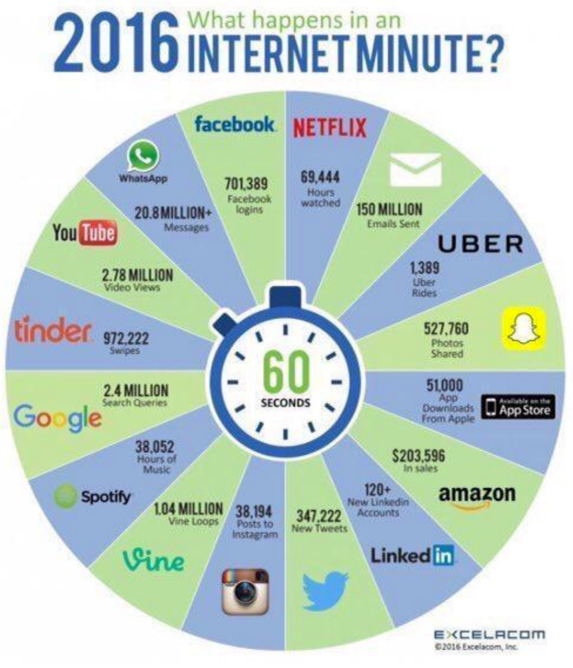
V = Volume

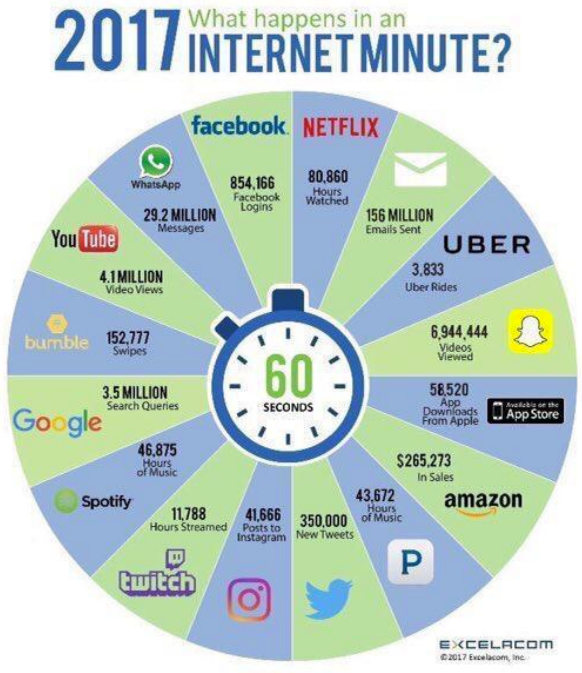

What Happens in an
Internet Minute?
1,572,877 GB of global IP data transferred[1]
10 Million ads displayed[2]
347,222 Tweets[3]
3.3 Million pieces of content shared[4]
6.9 Million messages sent[4]
Netflix + Youtube = more than ½ of all traffic[5]
$400 Million during Alibaba peak day sales[6]
438,801 Wiki page views[7]
10 Million WeChat messages at its peak[9]
34.7 Million instant messages (MIM) sent[8]
194,064 app downloads[10]
$133,436 in sales[11]
31,773 hours of music played[12]
38,194 photos uploaded[13]
57,870 page views[14]
4.1 Million searches[15]
100 hours of video uploaded[16]
138,889 hours of video watched[16]
23,148 hours of video watched[17]
Email
Wikipedia
WeChat
Mobile
Apps
amazon.com
PANDORA
LinkedIn
Instagram
Ads $$$
Pinterest
twitter
facebook
Google
YouTube
NETFLIX
intel
And Future Growth is Staggering
2012
2017
By 2017, mobile traffic will have grown 13X in just 5 years[1]
In 2017, there will be 3X more connected devices than people on Earth[1]
All digital data created reached 4 zettabytes in 2013[18]

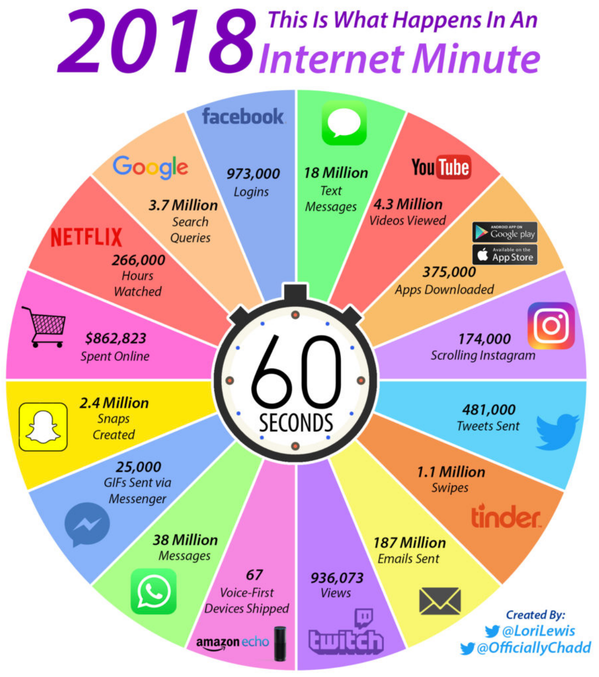
V = Volume


2018 This Is What Happens In An Internet Minute
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
Google play
App Store
375,000 Apps Downloaded
174,000 Scrolling Instagram
481,000 Tweets Sent
1.1 Million Swipes
tinder
187 Million Emails Sent
936,073 Views
twitch
67 Voice-First Devices Shipped
amazon echo
38 Million Messages
25,000 GIFs Sent via Messenger
2.4 Million Snaps Created
$862,823 Spent Online
NETFLIX
266,000 Hours Watched
Google
3.7 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

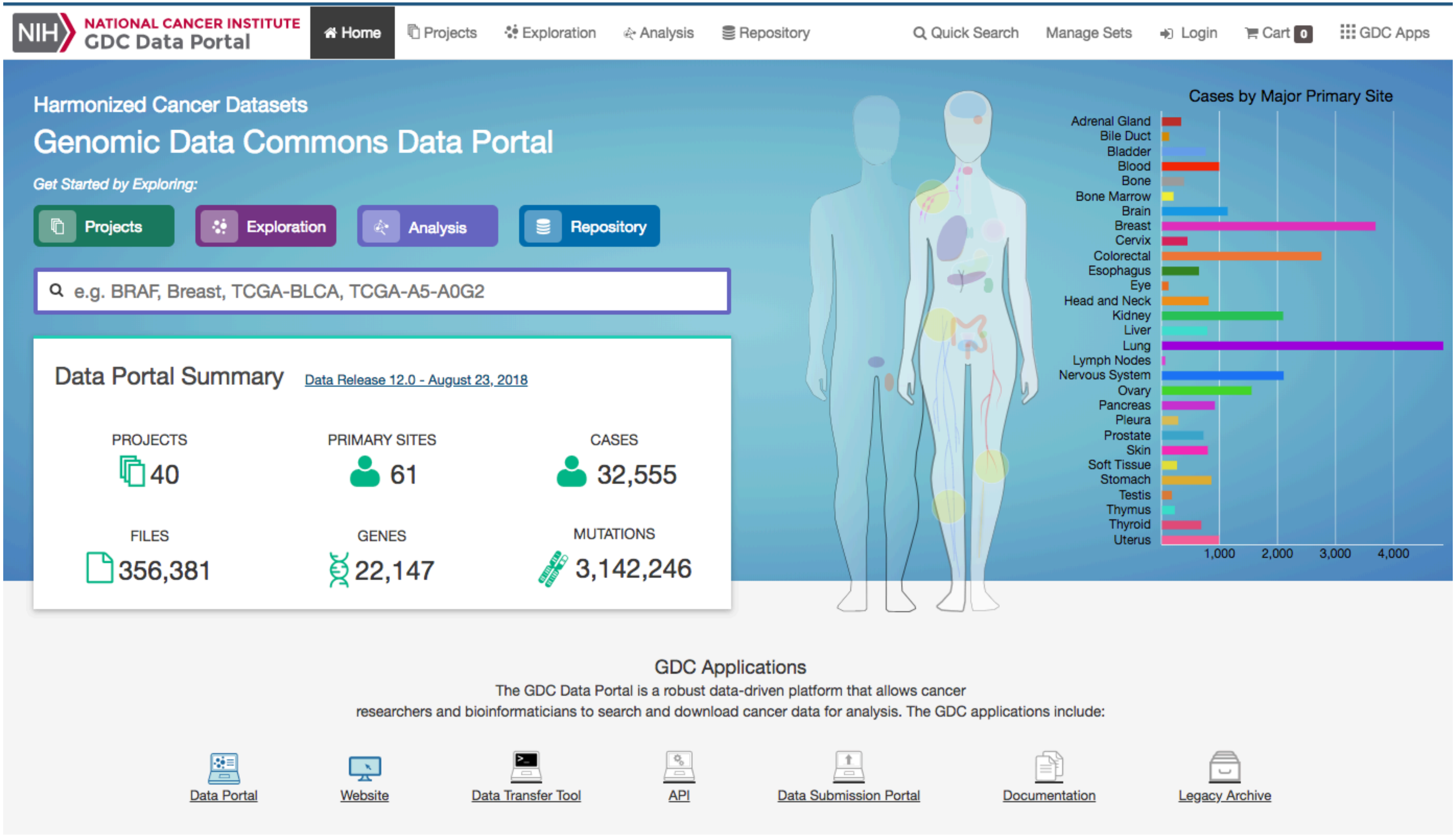
V = Volume


2016 What happens in an INTERNET MINUTE?
facebook
701,389 Facebook logins
NETFLIX
69,444 Hours watched
150 MILLION Emails Sent
UBER
1,389 Uber Rides
527,760 Photos Shared
51,000 App Downloads From Apple
Available on the App Store
$203,596 In sales
amazon
120+ New LinkedIn Accounts
LinkedIn
347,222 New Tweets
38,194 Posts to Instagram
1.04 MILLION Vine Loops
Vine
Spotify
38,052 Hours of Music
2.4 MILLION Search Queries
Google
tinder
972,222 Swipes
2.78 MILLION Video Views
YouTube
20.8 MILLION+ Messages
WhatsApp
60 SECONDS
EXCELACOM
©2016 Excelacom, Inc.


2017 What happens in an INTERNET MINUTE?
facebook
854,166 Facebook Logins
NETFLIX
80,860 Hours Watched
156 MILLION Emails Sent
UBER
3,833 Uber Rides
6,944,444 Videos Viewed
58,520 App Downloads From Apple
Available on the App Store
$265,273 In Sales
amazon
43,672 Hours of Music
P
350,000 New Tweets
41,666 Posts to Instagram
11,788 Hours Streamed
twitch
Spotify
46,875 Hours of Music
3.5 MILLION Search Queries
Google
bumble
152,777 Swipes
4.1 MILLION Video Views
YouTube
29.2 MILLION Messages
WhatsApp
60 SECONDS
EXCELACOM
©2017 Excelacom, Inc.


2018 This Is What Happens In An Internet Minute
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
Google play
App Store
375,000 Apps Downloaded
174,000 Scrolling Instagram
481,000 Tweets Sent
1.1 Million Swipes
tinder
187 Million Emails Sent
936,073 Views
twitch
67 Voice-First Devices Shipped
amazon echo
38 Million Messages
25,000 GIFs Sent via Messenger
2.4 Million Snaps Created
$862,823 Spent Online
NETFLIX
266,000 Hours Watched
Google
3.7 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

V = Volume



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Analysis | Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

Data Release 12.0 - August 23, 2018

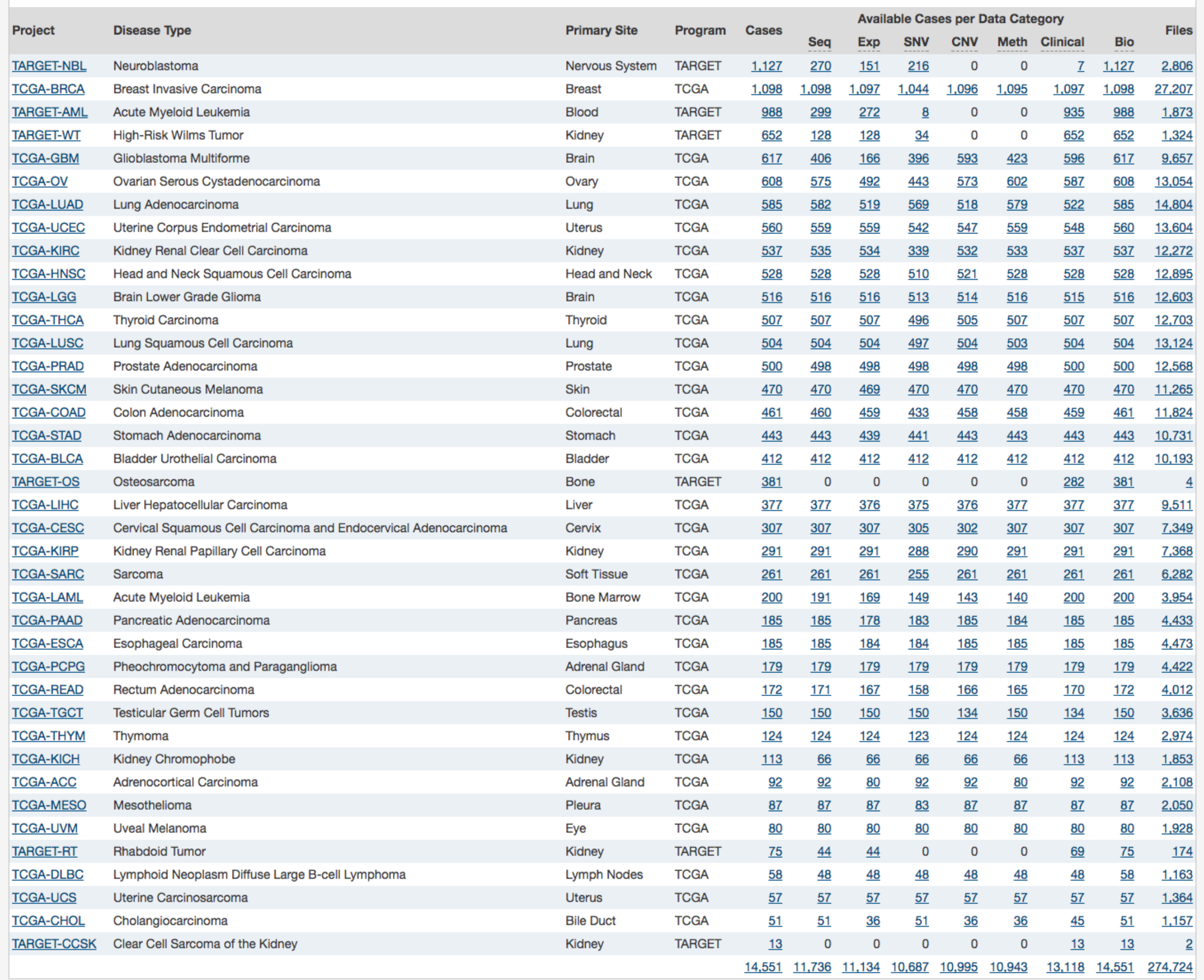
PROJECTS	PRIMARY SITES	CASES
40	61	32,555
FILES	**GENES**	**MUTATIONS**
356,381	22,147	3,142,246

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:

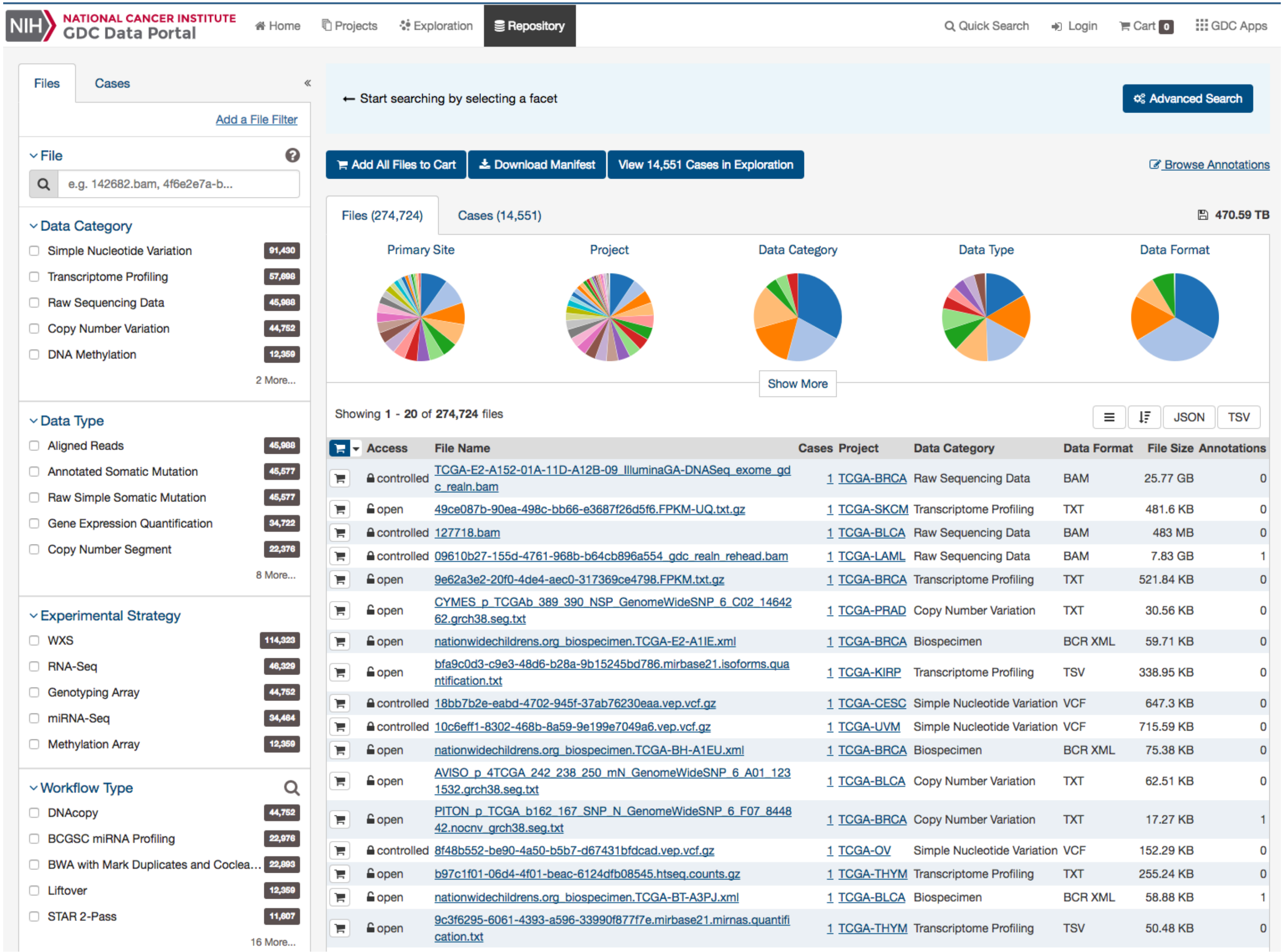
Data Portal | Website | Data Transfer Tool | API | Data Submission Portal | Documentation | Legacy Archive

V = Volume



Project	Disease Type	Primary Site	Program	Cases	Available Cases per Data Category							Files
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
TARGET-NBL	Neuroblastoma	Nervous System	TARGET	1,127	270	151	216	0	0	7	1,127	2,806
TCGA-BRCA	Breast Invasive Carcinoma	Breast	TCGA	1,098	1,098	1,097	1,044	1,096	1,095	1,097	1,098	27,207
TARGET-AML	Acute Myeloid Leukemia	Blood	TARGET	988	299	272	8	0	0	935	988	1,873
TARGET-WT	High-Risk Wilms Tumor	Kidney	TARGET	652	128	128	34	0	0	652	652	1,324
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617	406	166	396	593	423	596	617	9,657
TCGA-OV	Ovarian Serous Cystadenocarcinoma	Ovary	TCGA	608	575	492	443	573	602	587	608	13,054
TCGA-LUAD	Lung Adenocarcinoma	Lung	TCGA	585	582	519	569	518	579	522	585	14,804
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	Uterus	TCGA	560	559	559	542	547	559	548	560	13,604
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	TCGA	537	535	534	339	532	533	537	537	12,272
TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	Head and Neck	TCGA	528	528	528	510	521	528	528	528	12,895
TCGA-LGG	Brain Lower Grade Glioma	Brain	TCGA	516	516	516	513	514	516	515	516	12,603
TCGA-THCA	Thyroid Carcinoma	Thyroid	TCGA	507	507	507	496	505	507	507	507	12,703
TCGA-LUSC	Lung Squamous Cell Carcinoma	Lung	TCGA	504	504	504	497	504	503	504	504	13,124
TCGA-PRAD	Prostate Adenocarcinoma	Prostate	TCGA	500	498	498	498	498	498	500	500	12,568
TCGA-SKCM	Skin Cutaneous Melanoma	Skin	TCGA	470	470	469	470	470	470	470	470	11,265
TCGA-COAD	Colon Adenocarcinoma	Colorectal	TCGA	461	460	459	433	458	458	459	461	11,824
TCGA-STAD	Stomach Adenocarcinoma	Stomach	TCGA	443	443	439	441	443	443	443	443	10,731
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	TCGA	412	412	412	412	412	412	412	412	10,193
TARGET-OS	Osteosarcoma	Bone	TARGET	381	0	0	0	0	0	282	381	4
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	TCGA	377	377	376	375	376	377	377	377	9,511
TCGA-CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	Cervix	TCGA	307	307	307	305	302	307	307	307	7,349
TCGA-KIRP	Kidney Renal Papillary Cell Carcinoma	Kidney	TCGA	291	291	291	288	290	291	291	291	7,368
TCGA-SARC	Sarcoma	Soft Tissue	TCGA	261	261	261	255	261	261	261	261	6,282
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow	TCGA	200	191	169	149	143	140	200	200	3,954
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas	TCGA	185	185	178	183	185	184	185	185	4,433
TCGA-ESCA	Esophageal Carcinoma	Esophagus	TCGA	185	185	184	184	185	185	185	185	4,473
TCGA-PCPG	Pheochromocytoma and Paraganglioma	Adrenal Gland	TCGA	179	179	179	179	179	179	179	179	4,422
TCGA-READ	Rectum Adenocarcinoma	Colorectal	TCGA	172	171	167	158	166	165	170	172	4,012
TCGA-TGCT	Testicular Germ Cell Tumors	Testis	TCGA	150	150	150	150	134	150	134	150	3,636
TCGA-THYM	Thymoma	Thymus	TCGA	124	124	124	123	124	124	124	124	2,974
TCGA-KICH	Kidney Chromophobe	Kidney	TCGA	113	66	66	66	66	66	113	113	1,853
TCGA-ACC	Adrenocortical Carcinoma	Adrenal Gland	TCGA	92	92	80	92	92	80	92	92	2,108
TCGA-MESO	Mesothelioma	Pleura	TCGA	87	87	87	83	87	87	87	87	2,050
TCGA-UVM	Uveal Melanoma	Eye	TCGA	80	80	80	80	80	80	80	80	1,928
TARGET-RT	Rhabdoid Tumor	Kidney	TARGET	75	44	44	0	0	0	69	75	174
TCGA-DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Lymph Nodes	TCGA	58	48	48	48	48	48	48	58	1,163
TCGA-UCS	Uterine Carcinosarcoma	Uterus	TCGA	57	57	57	57	57	57	57	57	1,364
TCGA-CHOL	Cholangiocarcinoma	Bile Duct	TCGA	51	51	36	51	36	36	45	51	1,157
TARGET-CCSK	Clear Cell Sarcoma of the Kidney	Kidney	TARGET	13	0	0	0	0	0	13	13	2
				14,551	11,736	11,134	10,687	10,995	10,943	13,118	14,551	274,724

V = Volume



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Repository | Quick Search | Login | Cart 0 | GDC Apps

Files | Cases «

Add a File Filter

File

e.g. 142682.bam, 4f6e2e7a-b...

Data Category

□ Simple Nucleotide Variation	91,430
□ Transcriptome Profiling	57,698
□ Raw Sequencing Data	45,988
□ Copy Number Variation	44,752
□ DNA Methylation	12,359

2 More...

Data Type

□ Aligned Reads	45,988
□ Annotated Somatic Mutation	45,577
□ Raw Simple Somatic Mutation	45,577
□ Gene Expression Quantification	34,722
□ Copy Number Segment	22,376

8 More...

Experimental Strategy

□ WXS	114,323
□ RNA-Seq	46,329
□ Genotyping Array	44,752
□ miRNA-Seq	34,464
□ Methylation Array	12,359

Workflow Type

□ DNAcopy	44,752
□ BCGSC miRNA Profiling	22,976
□ BWA with Mark Duplicates and Coclea...	22,893
□ Liftover	12,359
□ STAR 2-Pass	11,607

16 More...

← Start searching by selecting a facet

Advanced Search

Add All Files to Cart | Download Manifest | View 14,551 Cases in Exploration

Browse Annotations

Files (274,724) | Cases (14,551) 470.59 TB

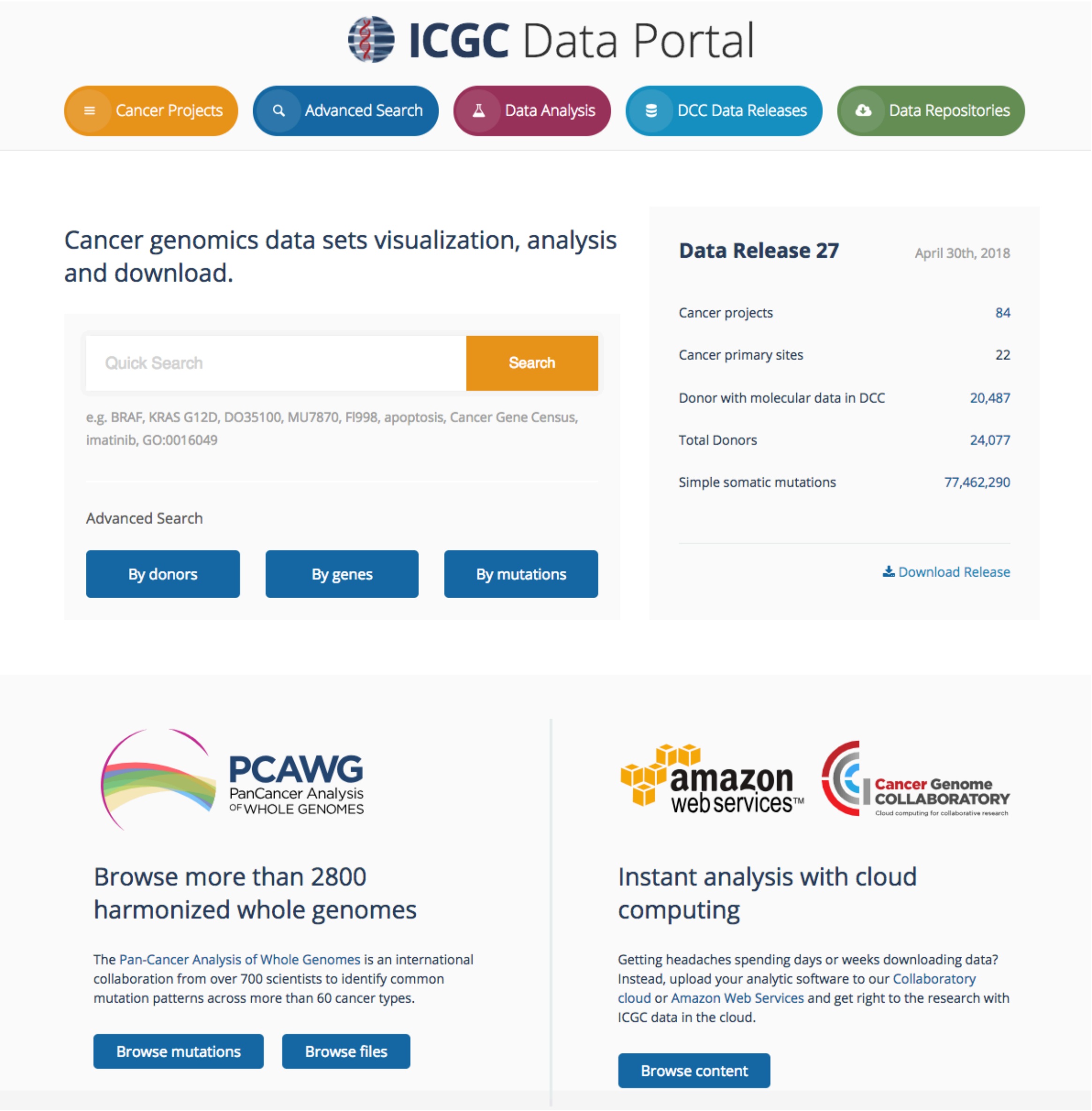
Show More

Showing **1** - **20** of **274,724** files

JSON | TSV

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
controlled	TCGA-E2-A152-01A-11D-A12B-09_IlluminaGA-DNASeq_exome_gdc_realn.bam	1	TCGA-BRCA	Raw Sequencing Data	BAM	25.77 GB	0
open	49ce087b-90ea-498c-bb66-e3687f26d5f6.FPKM-UQ.txt.gz	1	TCGA-SKCM	Transcriptome Profiling	TXT	481.6 KB	0
controlled	127718.bam	1	TCGA-BLCA	Raw Sequencing Data	BAM	483 MB	0
controlled	09610b27-155d-4761-968b-b64cb896a554_gdc_realn_rehead.bam	1	TCGA-LAML	Raw Sequencing Data	BAM	7.83 GB	1
open	9e62a3e2-20f0-4de4-aec0-317369ce4798.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	521.84 KB	0
open	CYMES_p_TCGAb_389_390_NSP_GenomeWideSNP_6_C02_1464262.grch38.seg.txt	1	TCGA-PRAD	Copy Number Variation	TXT	30.56 KB	0
open	nationwidechildrens.org_biospecimen.TCGA-E2-A1IE.xml	1	TCGA-BRCA	Biospecimen	BCR XML	59.71 KB	0
open	bfa9c0d3-c9e3-48d6-b28a-9b15245bd786.mirbase21.isoforms.quantification.txt	1	TCGA-KIRP	Transcriptome Profiling	TSV	338.95 KB	0
controlled	18bb7b2e-eabd-4702-945f-37ab76230eaa.vep.vcf.gz	1	TCGA-CESC	Simple Nucleotide Variation	VCF	647.3 KB	0
controlled	10c6eff1-8302-468b-8a59-9e199e7049a6.vep.vcf.gz	1	TCGA-UVM	Simple Nucleotide Variation	VCF	715.59 KB	0
open	nationwidechildrens.org_biospecimen.TCGA-BH-A1EU.xml	1	TCGA-BRCA	Biospecimen	BCR XML	75.38 KB	0
open	AVISO_p_4TCGA_242_238_250_mN_GenomeWideSNP_6_A01_1231532.grch38.seg.txt	1	TCGA-BLCA	Copy Number Variation	TXT	62.51 KB	0
open	PITON_p_TCGA_b162_167_SNP_N_GenomeWideSNP_6_F07_844842.nocnv_grch38.seg.txt	1	TCGA-BRCA	Copy Number Variation	TXT	17.27 KB	1
controlled	8f48b552-be90-4a50-b5b7-d67431bfdcad.vep.vcf.gz	1	TCGA-OV	Simple Nucleotide Variation	VCF	152.29 KB	0
open	b97c1f01-06d4-4f01-beac-6124dfb08545.htseq.counts.gz	1	TCGA-THYM	Transcriptome Profiling	TXT	255.24 KB	0
open	nationwidechildrens.org_biospecimen.TCGA-BT-A3PJ.xml	1	TCGA-BLCA	Biospecimen	BCR XML	58.88 KB	1
open	9c3f6295-6061-4393-a596-33990f877f7e.mirbase21.mirnas.quantification.txt	1	TCGA-THYM	Transcriptome Profiling	TSV	50.48 KB	0

V = Volume



Cancer genomics data sets visualization, analysis and download.

e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

Advanced Search

Data Release 27 April 30th, 2018

Cancer projects	84
Cancer primary sites	22
Donor with molecular data in DCC	20,487
Total Donors	24,077
Simple somatic mutations	77,462,290

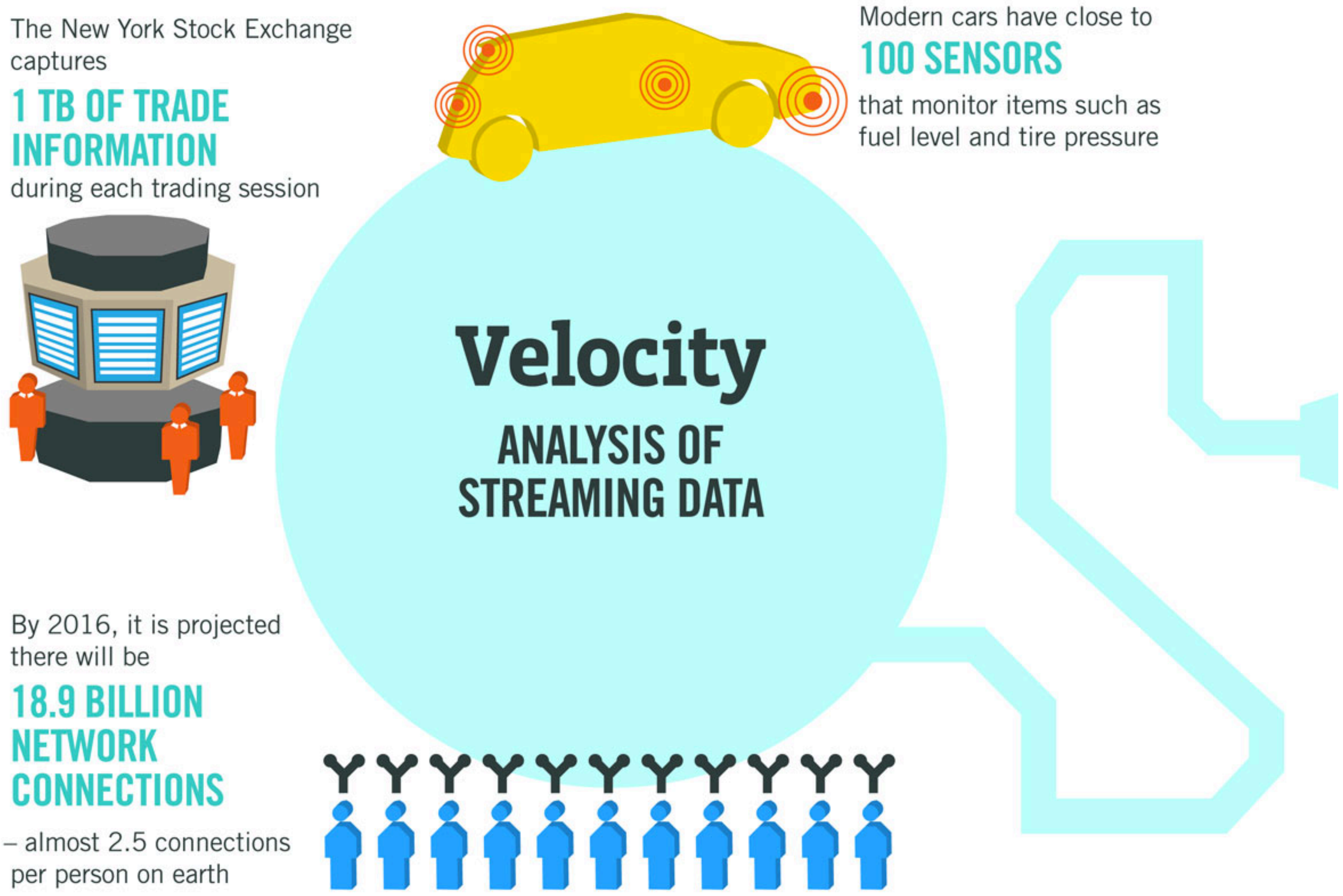
Browse more than 2800 harmonized whole genomes

The Pan-Cancer Analysis of Whole Genomes is an international collaboration from over 700 scientists to identify common mutation patterns across more than 60 cancer types.

Instant analysis with cloud computing

Getting headaches spending days or weeks downloading data? Instead, upload your analytic software to our Collaboratory cloud or Amazon Web Services and get right to the research with ICGC data in the cloud.

V = Velocity



The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

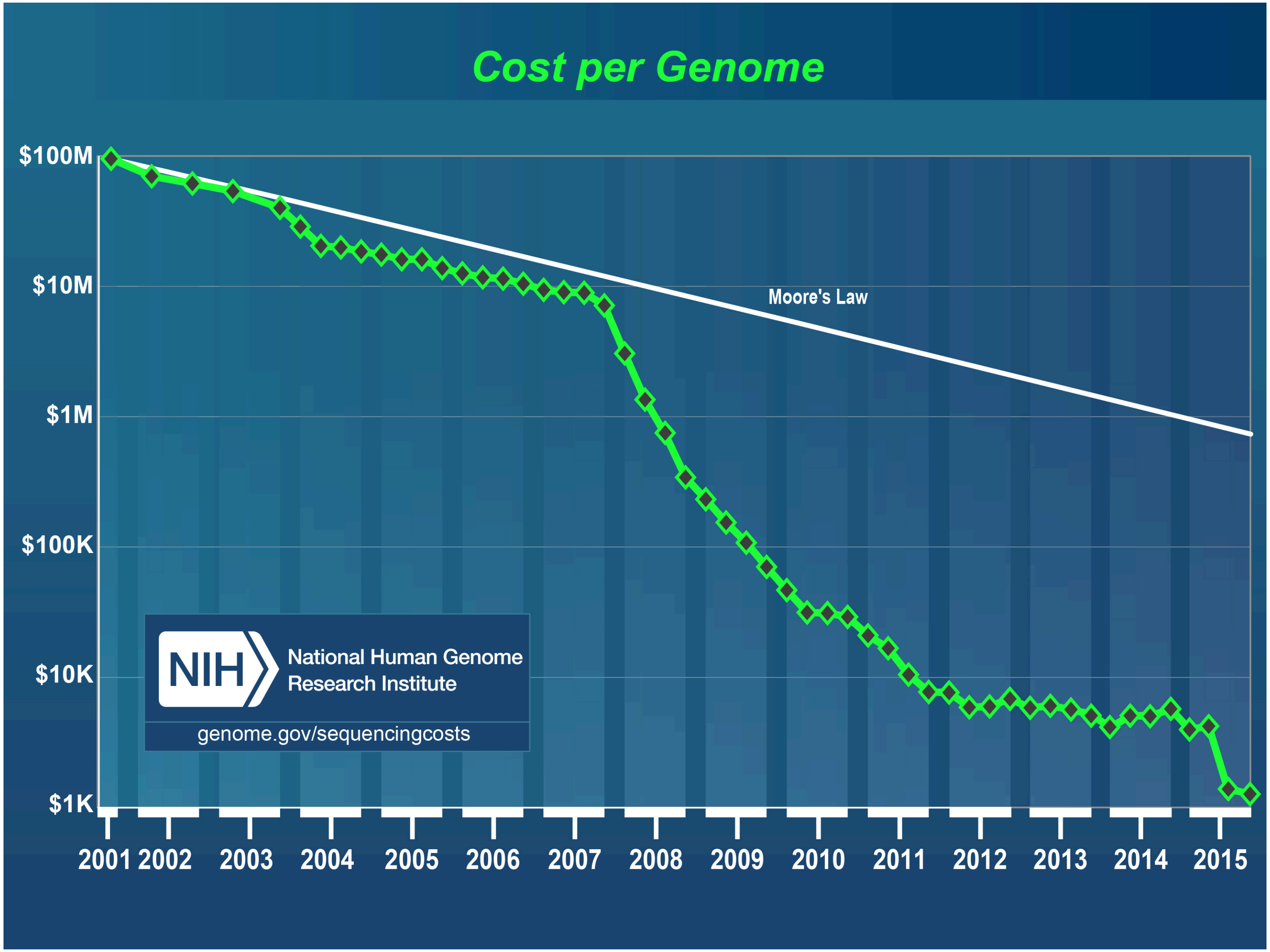
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

V = Velocity


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

V = Velocity

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
SPECIAL ISSUE
Polymers and plastics
THE HUMAN
EXOME
The ExAC catalogue of human protein–coding genetic variation, spanning more than 60,000 samples
PAGES 277 & 285

V = Velocity

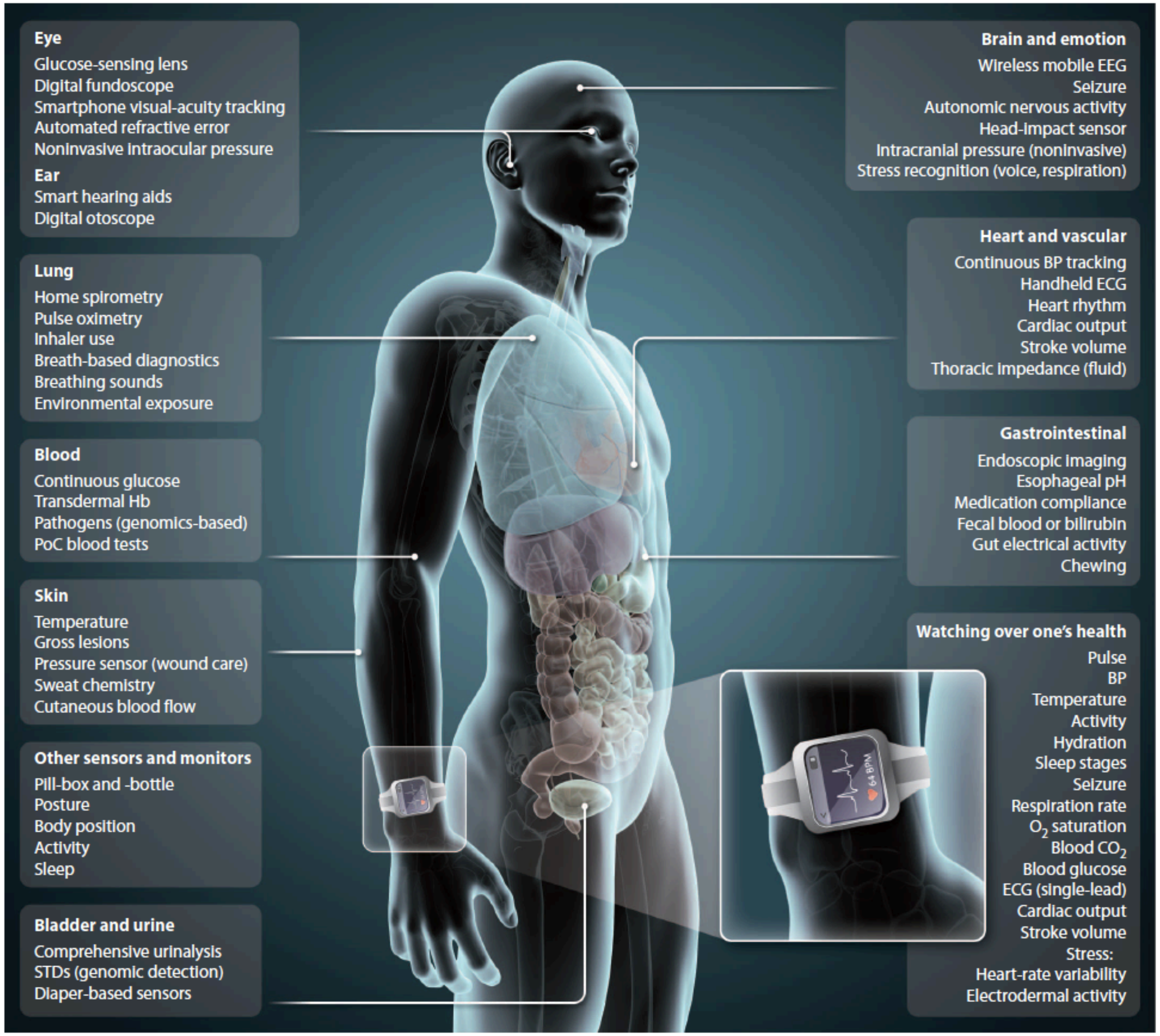
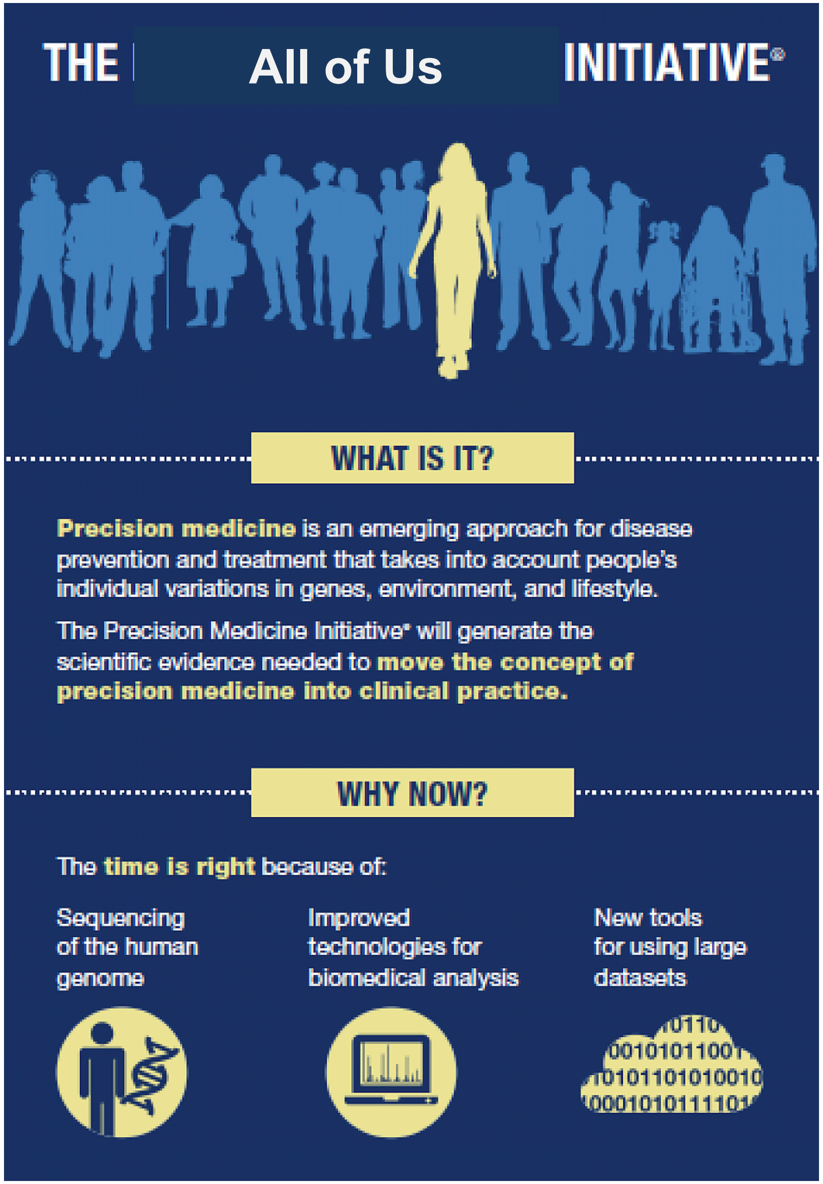


Fig. 1. Sensing a shift in health care. Shown are bodywide measurements by mHealth technologies that are available to health care providers and patients to aid in the tracking, diagnosis, or management of various physiological processes and disease conditions. **(Inset) Watching over one's health.** Multiple developers have reported that the listed physiological parameters are measurable with sensors in a wrist-worn device. BP, blood pressure; Hb, hemoglobin; STDs, sexually transmitted diseases.

(Steinhubl et al Science Translational Medicine 2015)

V = Velocity



THE All of Us INITIATIVE®

WHAT IS IT?

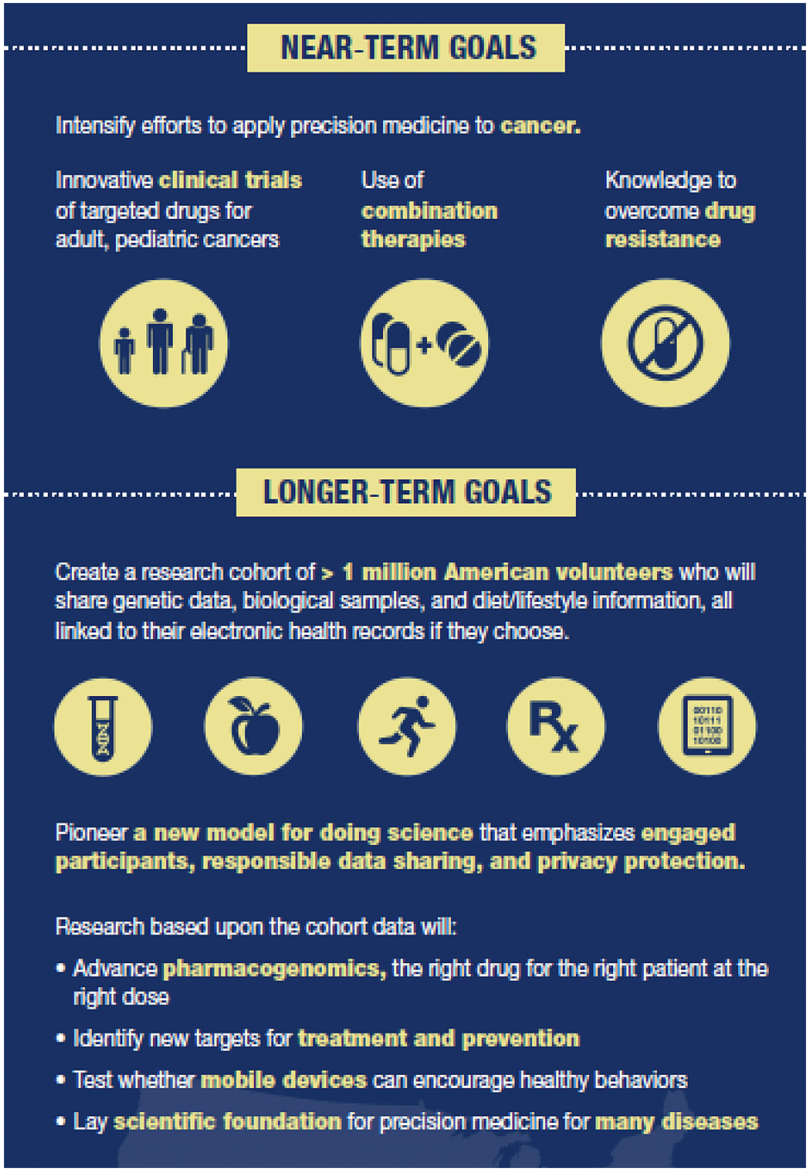
Precision medicine is an emerging approach for disease prevention and treatment that takes into account people's individual variations in genes, environment, and lifestyle.

The Precision Medicine Initiative® will generate the scientific evidence needed to **move the concept of precision medicine into clinical practice.**

WHY NOW?

The **time is right** because of:

- Sequencing of the human genome
- Improved technologies for biomedical analysis
- New tools for using large datasets



NEAR-TERM GOALS

Intensify efforts to apply precision medicine to **cancer.**

- Innovative **clinical trials** of targeted drugs for adult, pediatric cancers
- Use of **combination therapies**
- Knowledge to overcome **drug resistance**

LONGER-TERM GOALS

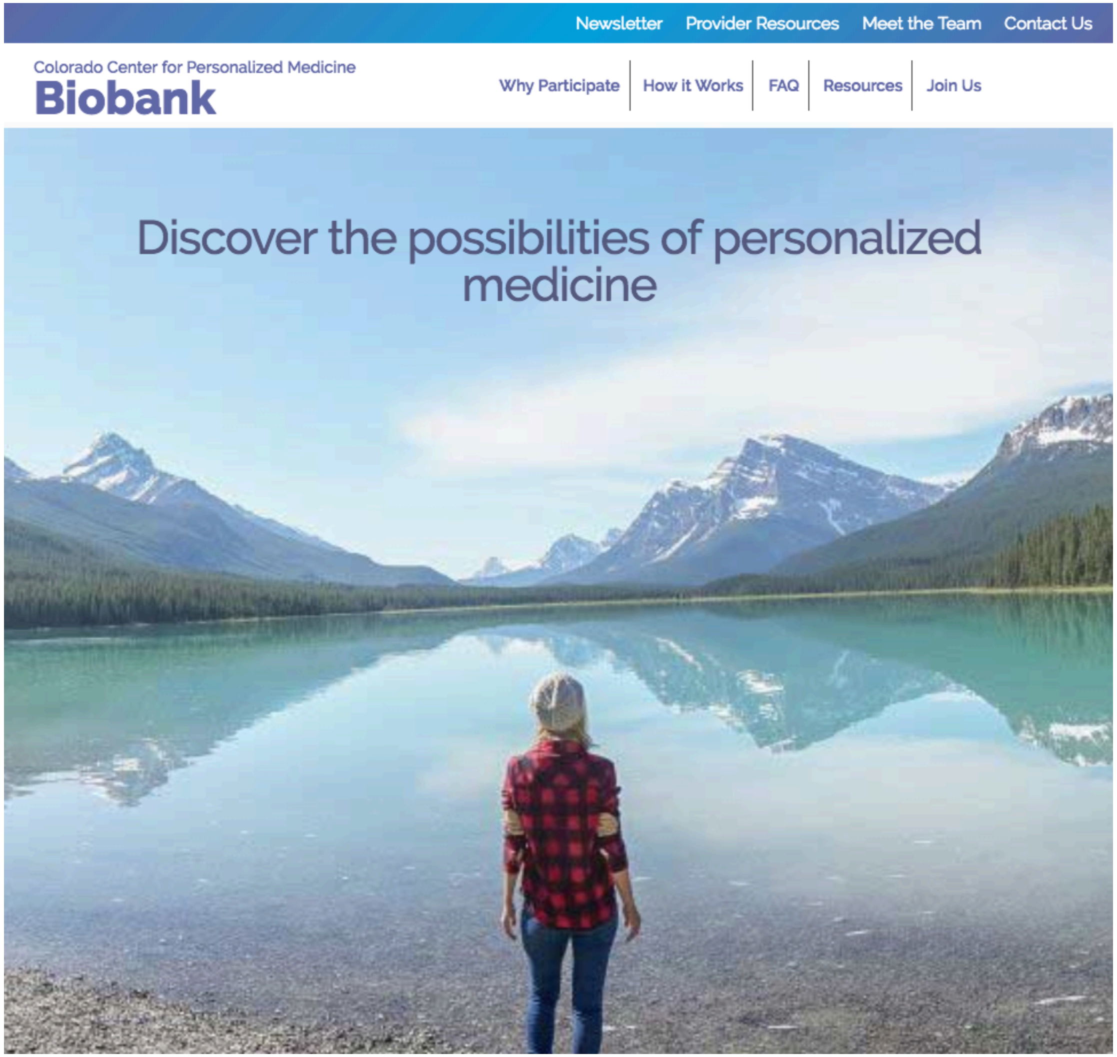
Create a research cohort of > **1 million American volunteers** who will share genetic data, biological samples, and diet/lifestyle information, all linked to their electronic health records if they choose.

Pioneer **a new model for doing science** that emphasizes **engaged participants, responsible data sharing, and privacy protection.**

Research based upon the cohort data will:

- Advance **pharmacogenomics,** the right drug for the right patient at the right dose
- Identify new targets for **treatment and prevention**
- Test whether **mobile devices** can encourage healthy behaviors
- Lay **scientific foundation** for precision medicine for **many diseases**

V = Velocity



Newsletter Provider Resources Meet the Team Contact Us

Colorado Center for Personalized Medicine

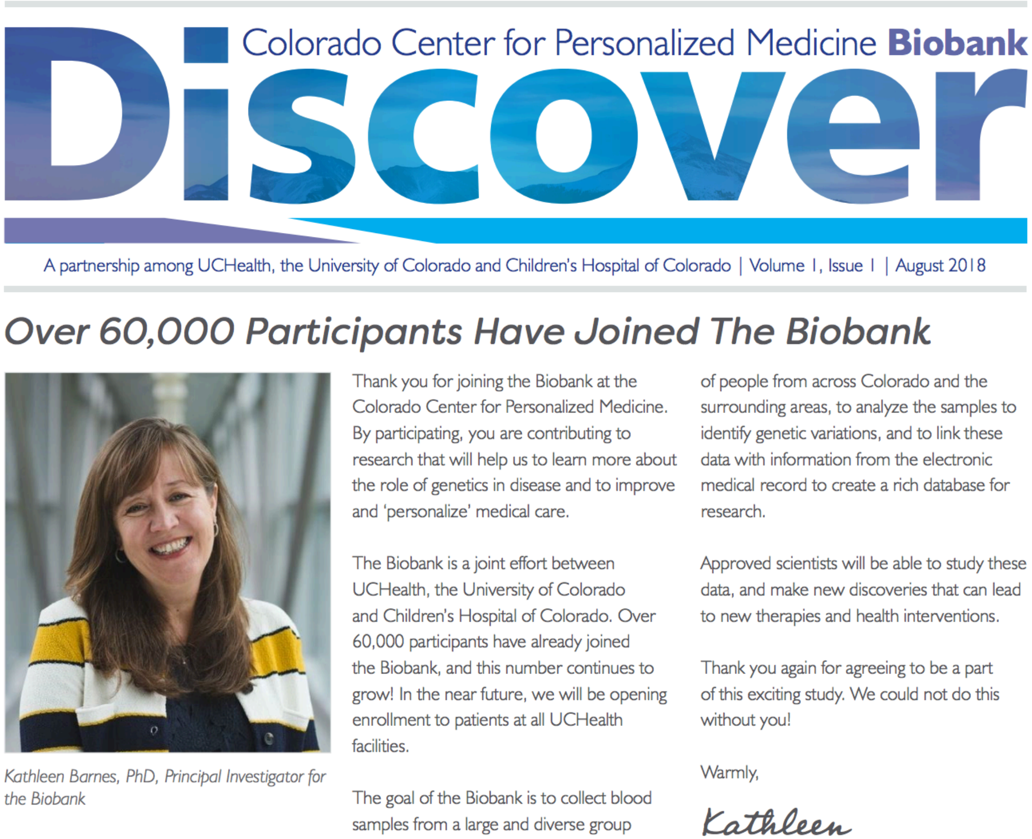
Biobank

Why Participate | How it Works | FAQ | Resources | Join Us

Discover the possibilities of personalized medicine

https://www.cobiobank.org/

V = Velocity



Colorado Center for Personalized Medicine **Biobank**

Discover

A partnership among UCHealth, the University of Colorado and Children's Hospital of Colorado | Volume 1, Issue 1 | August 2018

Over 60,000 Participants Have Joined The Biobank

Kathleen Barnes, PhD, Principal Investigator for the Biobank

Thank you for joining the Biobank at the Colorado Center for Personalized Medicine. By participating, you are contributing to research that will help us to learn more about the role of genetics in disease and to improve and 'personalize' medical care.

The Biobank is a joint effort between UCHealth, the University of Colorado and Children's Hospital of Colorado. Over 60,000 participants have already joined the Biobank, and this number continues to grow! In the near future, we will be opening enrollment to patients at all UCHealth facilities.

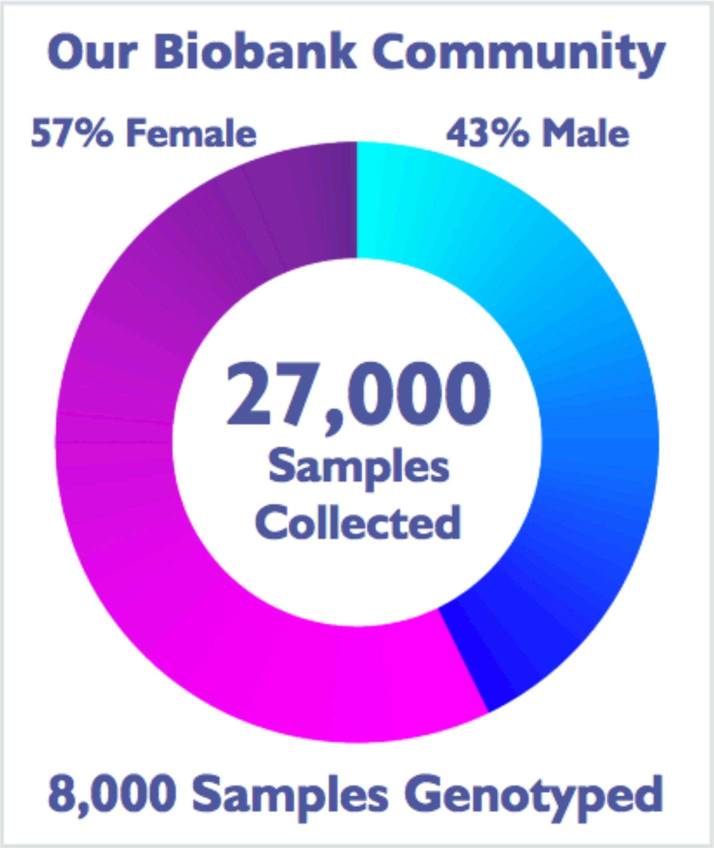
The goal of the Biobank is to collect blood samples from a large and diverse group of people from across Colorado and the surrounding areas, to analyze the samples to identify genetic variations, and to link these data with information from the electronic medical record to create a rich database for research.

Approved scientists will be able to study these data, and make new discoveries that can lead to new therapies and health interventions.

Thank you again for agreeing to be a part of this exciting study. We could not do this without you!

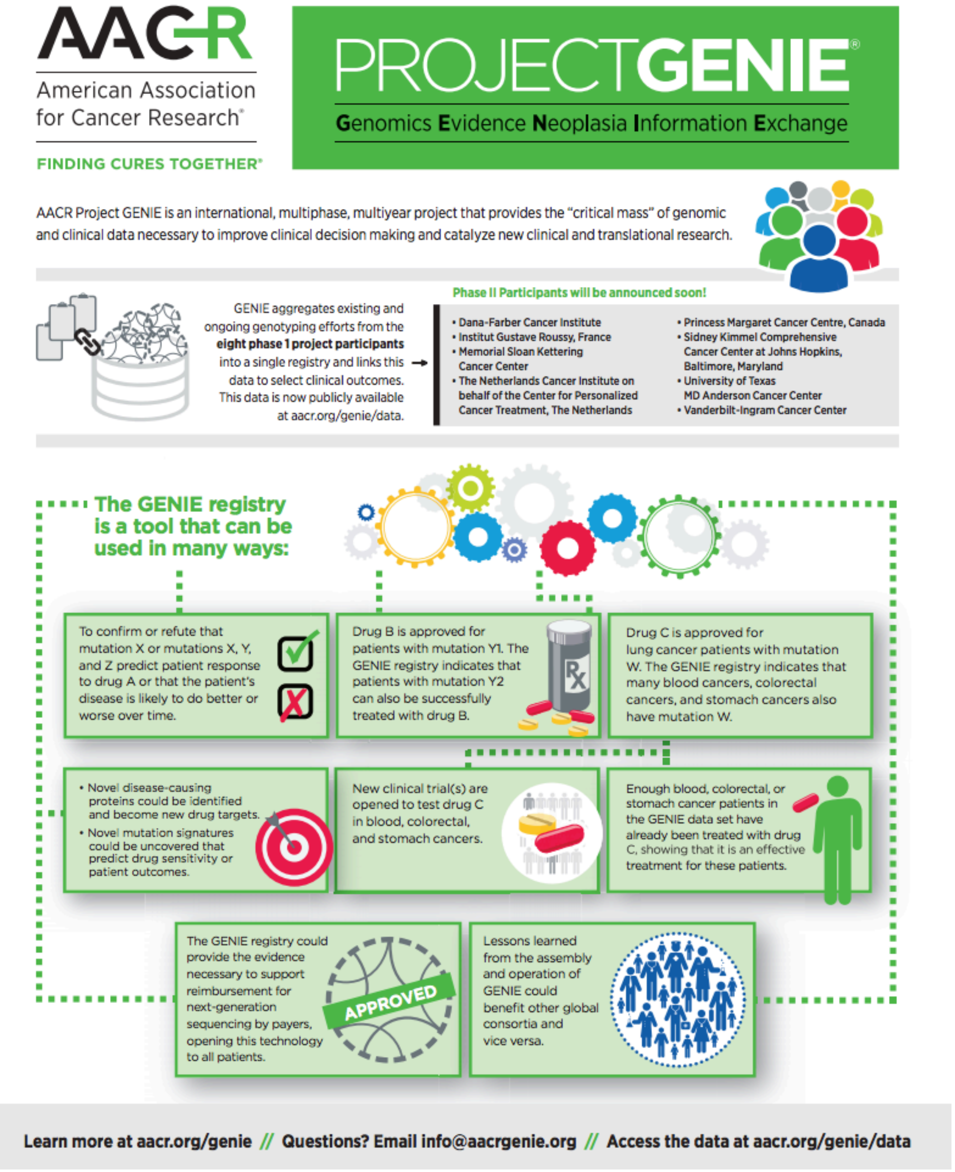
Warmly,

Kathleen



https://www.cobiobank.org/

V = Velocity



AACR

American Association for Cancer Research®

FINDING CURES TOGETHER®

PROJECT GENIE®

Genomics Evidence Neoplasia Information Exchange

AACR Project GENIE is an international, multiphase, multiyear project that provides the "critical mass" of genomic and clinical data necessary to improve clinical decision making and catalyze new clinical and translational research.

GENIE aggregates existing and ongoing genotyping efforts from the **eight phase 1 project participants** into a single registry and links this data to select clinical outcomes. This data is now publicly available at aacr.org/genie/data. →

Phase II Participants will be announced soon!

- Dana-Farber Cancer Institute
- Institut Gustave Roussy, France
- Memorial Sloan Kettering Cancer Center
- The Netherlands Cancer Institute on behalf of the Center for Personalized Cancer Treatment, The Netherlands
- Princess Margaret Cancer Centre, Canada
- Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland
- University of Texas MD Anderson Cancer Center
- Vanderbilt-Ingram Cancer Center

The GENIE registry is a tool that can be used in many ways:

To confirm or refute that mutation X or mutations X, Y, and Z predict patient response to drug A or that the patient's disease is likely to do better or worse over time. ☑ ☒

Drug B is approved for patients with mutation Y1. The GENIE registry indicates that patients with mutation Y2 can also be successfully treated with drug B.

Drug C is approved for lung cancer patients with mutation W. The GENIE registry indicates that many blood cancers, colorectal cancers, and stomach cancers also have mutation W.

- Novel disease-causing proteins could be identified and become new drug targets.
- Novel mutation signatures could be uncovered that predict drug sensitivity or patient outcomes.

New clinical trial(s) are opened to test drug C in blood, colorectal, and stomach cancers.

Enough blood, colorectal, or stomach cancer patients in the GENIE data set have already been treated with drug C, showing that it is an effective treatment for these patients.

The GENIE registry could provide the evidence necessary to support reimbursement for next-generation sequencing by payers, opening this technology to all patients.

Lessons learned from the assembly and operation of GENIE could benefit other global consortia and vice versa.

Learn more at aacr.org/genie // Questions? Email info@aacrgenie.org // Access the data at aacr.org/genie/data

GENIE is Unique

The registry contains the existing CLIA-/ISO-certified genomic data obtained during the course of routine practice at multiple national and international institutions, and will continue to grow as more patients are treated at the participating centers and as new centers join the project. As a result, the registry is derived from a variety of cancer types, including rare cancers, and is enriched in examples of late-stage disease; thus it approximates more of a "real world" dataset.

One Registry, Many Uses

* Powering clinical and translational research

- The database can be used to generate many research hypotheses spanning translational to clinical studies, including those that would inform new or ongoing clinical trials.

* Validating biomarkers

* Drug repositioning/repurposing*

* Adding new mutations to existing drug labels*

* Identifying new drug targets

* Could provide the evidence base necessary to support reimbursement for next-generation sequence-based testing by payers.

* The AACR will be working closely with the FDA to ensure that the registry contains data that could be accepted as evidence supporting regulatory approval.

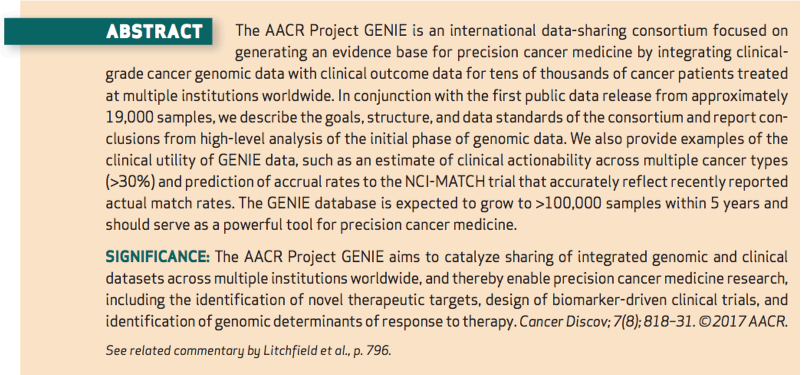
https://www.youtube.com/watch?v=DUc00BjfpMc

V = Velocity

RESEARCH ARTICLE

AACR Project GENIE: Powering Precision Medicine through an International Consortium

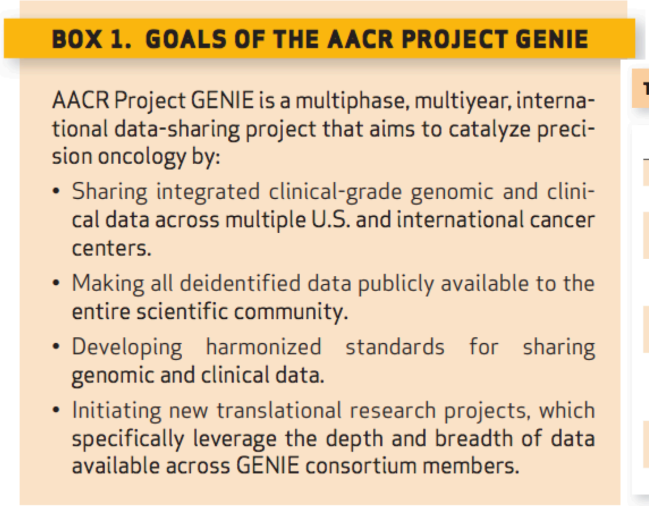
The AACR Project GENIE Consortium



ABSTRACT The AACR Project GENIE is an international data-sharing consortium focused on generating an evidence base for precision cancer medicine by integrating clinical-grade cancer genomic data with clinical outcome data for tens of thousands of cancer patients treated at multiple institutions worldwide. In conjunction with the first public data release from approximately 19,000 samples, we describe the goals, structure, and data standards of the consortium and report conclusions from high-level analysis of the initial phase of genomic data. We also provide examples of the clinical utility of GENIE data, such as an estimate of clinical actionability across multiple cancer types (>30%) and prediction of accrual rates to the NCI-MATCH trial that accurately reflect recently reported actual match rates. The GENIE database is expected to grow to >100,000 samples within 5 years and should serve as a powerful tool for precision cancer medicine.

SIGNIFICANCE: The AACR Project GENIE aims to catalyze sharing of integrated genomic and clinical datasets across multiple institutions worldwide, and thereby enable precision cancer medicine research, including the identification of novel therapeutic targets, design of biomarker-driven clinical trials, and identification of genomic determinants of response to therapy. *Cancer Discov; 7(8); 818–31. ©2017 AACR.*

See related commentary by Litchfield et al., p. 796.



BOX 1. GOALS OF THE AACR PROJECT GENIE

AACR Project GENIE is a multiphase, multiyear, international data-sharing project that aims to catalyze precision oncology by:

- Sharing integrated clinical-grade genomic and clinical data across multiple U.S. and international cancer centers.
- Making all deidentified data publicly available to the entire scientific community.
- Developing harmonized standards for sharing genomic and clinical data.
- Initiating new translational research projects, which specifically leverage the depth and breadth of data available across GENIE consortium members.

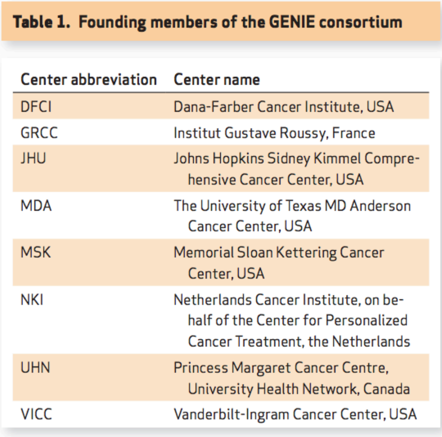
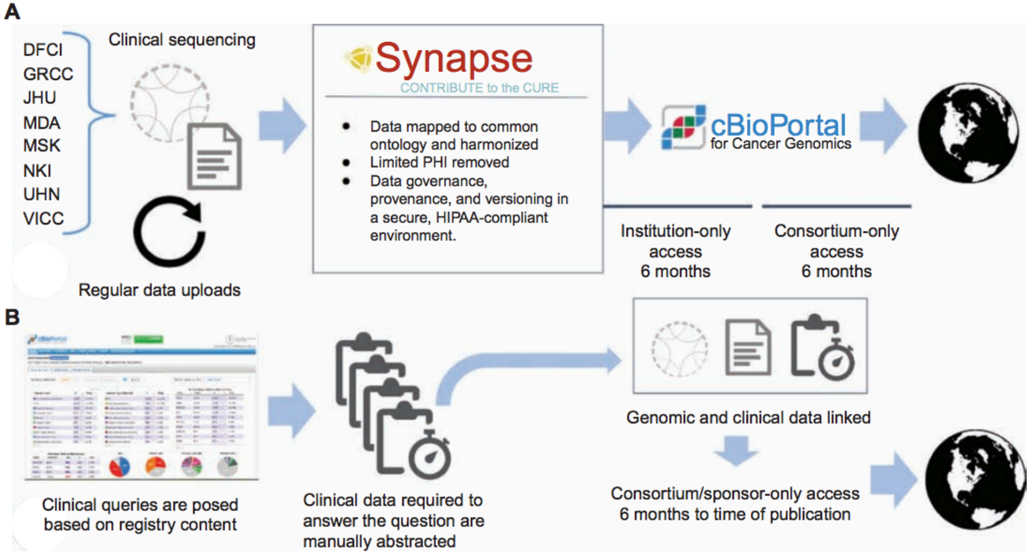
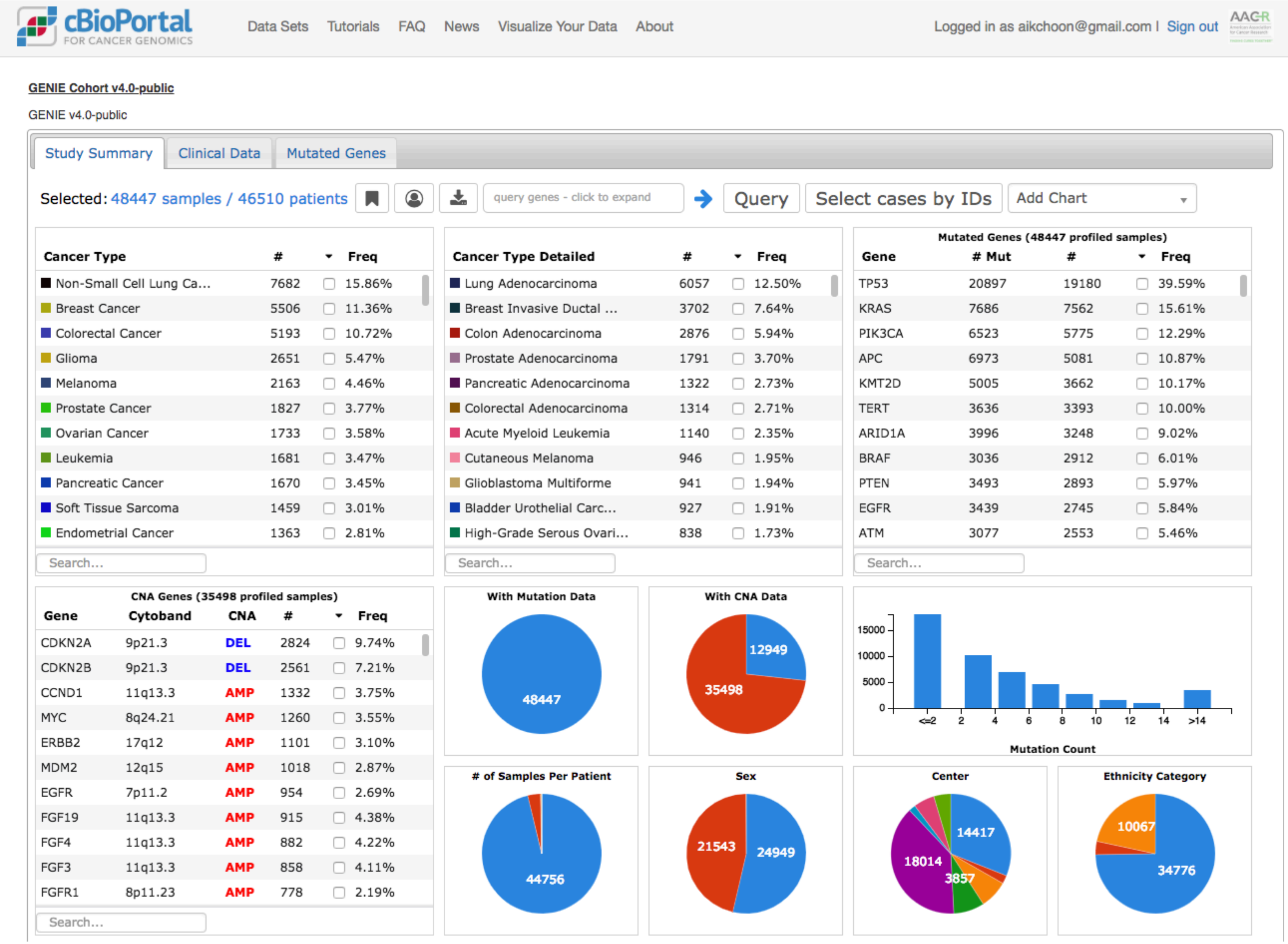


Table 1. Founding members of the GENIE consortium

Center abbreviation	Center name
DFCI	Dana-Farber Cancer Institute, USA
GRCC	Institut Gustave Roussy, France
JHU	Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, USA
MDA	The University of Texas MD Anderson Cancer Center, USA
MSK	Memorial Sloan Kettering Cancer Center, USA
NKI	Netherlands Cancer Institute, on behalf of the Center for Personalized Cancer Treatment, the Netherlands
UHN	Princess Margaret Cancer Centre, University Health Network, Canada
VICC	Vanderbilt-Ingram Cancer Center, USA


A
DFCI
GRCC
JHU
MDA
MSK
NKI
UHN
VICC
Clinical sequencing
Regular data uploads
Synapse
CONTRIBUTE to the CURE
• Data mapped to common ontology and harmonized
• Limited PHI removed
• Data governance, provenance, and versioning in a secure, HIPAA-compliant environment.
cBioPortal for Cancer Genomics
Institution-only access 6 months
Consortium-only access 6 months
B
Clinical queries are posed based on registry content
Clinical data required to answer the question are manually abstracted
Genomic and clinical data linked
Consortium/sponsor-only access 6 months to time of publication

V = Velocity



cBioPortal FOR CANCER GENOMICS
Data Sets
Tutorials
FAQ
News
Visualize Your Data
About
Logged in as aikchoon@gmail.com I
Sign out
AACR

GENIE Cohort v4.0-public

GENIE v4.0-public

Selected: 48447 samples / 46510 patients

Cancer Type	#		Freq
■ Non-Small Cell Lung Ca...	7682	□	15.86%
■ Breast Cancer	5506	□	11.36%
■ Colorectal Cancer	5193	□	10.72%
■ Glioma	2651	□	5.47%
■ Melanoma	2163	□	4.46%
■ Prostate Cancer	1827	□	3.77%
■ Ovarian Cancer	1733	□	3.58%
■ Leukemia	1681	□	3.47%
■ Pancreatic Cancer	1670	□	3.45%
■ Soft Tissue Sarcoma	1459	□	3.01%
■ Endometrial Cancer	1363	□	2.81%

Search...

Cancer Type Detailed	#		Freq
■ Lung Adenocarcinoma	6057	□	12.50%
■ Breast Invasive Ductal ...	3702	□	7.64%
■ Colon Adenocarcinoma	2876	□	5.94%
■ Prostate Adenocarcinoma	1791	□	3.70%
■ Pancreatic Adenocarcinoma	1322	□	2.73%
■ Colorectal Adenocarcinoma	1314	□	2.71%
■ Acute Myeloid Leukemia	1140	□	2.35%
■ Cutaneous Melanoma	946	□	1.95%
■ Glioblastoma Multiforme	941	□	1.94%
■ Bladder Urothelial Carc...	927	□	1.91%
■ High-Grade Serous Ovari...	838	□	1.73%

Search...

Mutated Genes (48447 profiled samples)

Gene	# Mut	#		Freq
TP53	20897	19180	□	39.59%
KRAS	7686	7562	□	15.61%
PIK3CA	6523	5775	□	12.29%
APC	6973	5081	□	10.87%
KMT2D	5005	3662	□	10.17%
TERT	3636	3393	□	10.00%
ARID1A	3996	3248	□	9.02%
BRAF	3036	2912	□	6.01%
PTEN	3493	2893	□	5.97%
EGFR	3439	2745	□	5.84%
ATM	3077	2553	□	5.46%

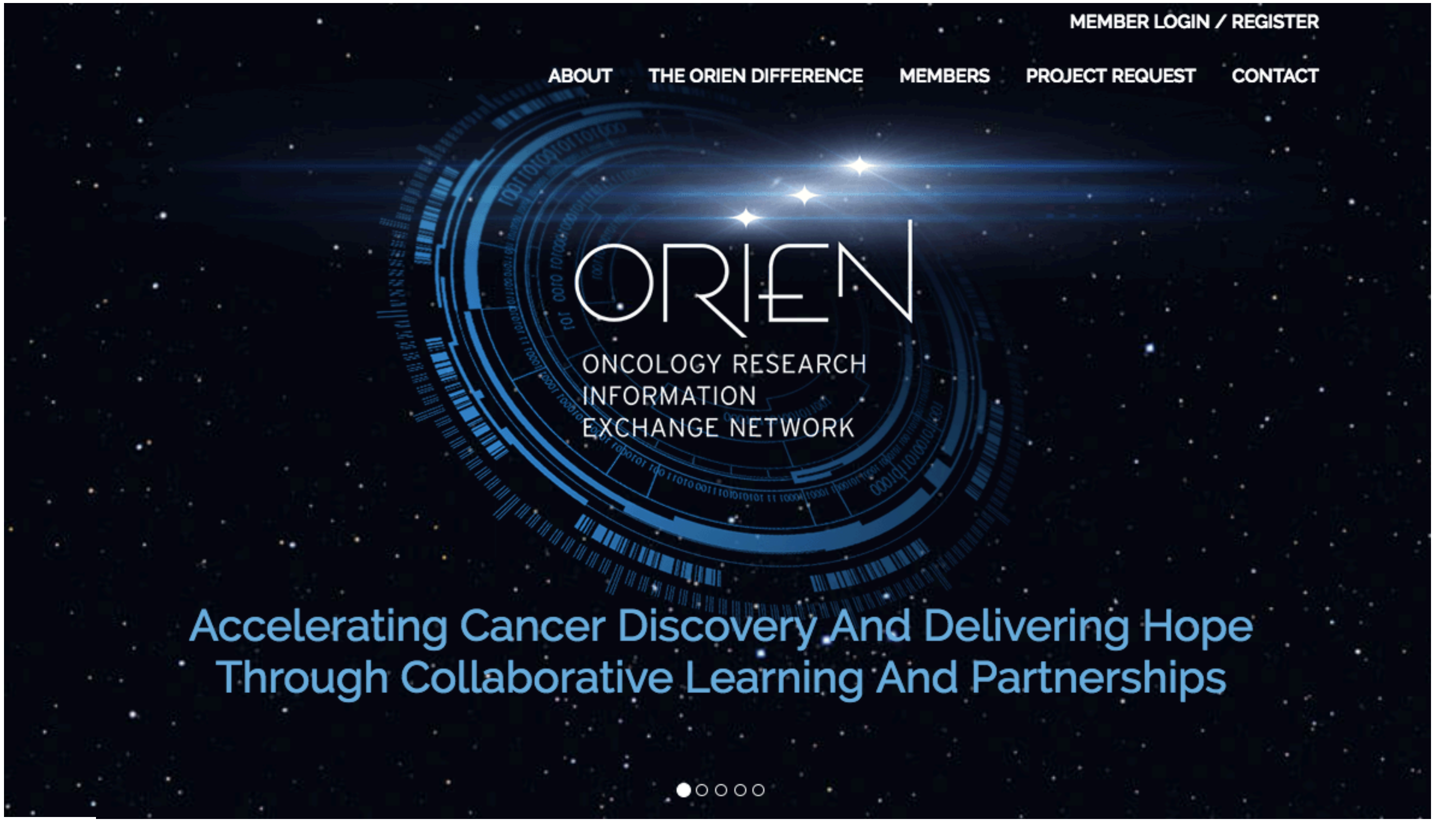
Search...

CNA Genes (35498 profiled samples)

Gene	Cytoband	CNA	#		Freq
CDKN2A	9p21.3	DEL	2824	□	9.74%
CDKN2B	9p21.3	DEL	2561	□	7.21%
CCND1	11q13.3	AMP	1332	□	3.75%
MYC	8q24.21	AMP	1260	□	3.55%
ERBB2	17q12	AMP	1101	□	3.10%
MDM2	12q15	AMP	1018	□	2.87%
EGFR	7p11.2	AMP	954	□	2.69%
FGF19	11q13.3	AMP	915	□	4.38%
FGF4	11q13.3	AMP	882	□	4.22%
FGF3	11q13.3	AMP	858	□	4.11%
FGFR1	8p11.23	AMP	778	□	2.19%

Search...

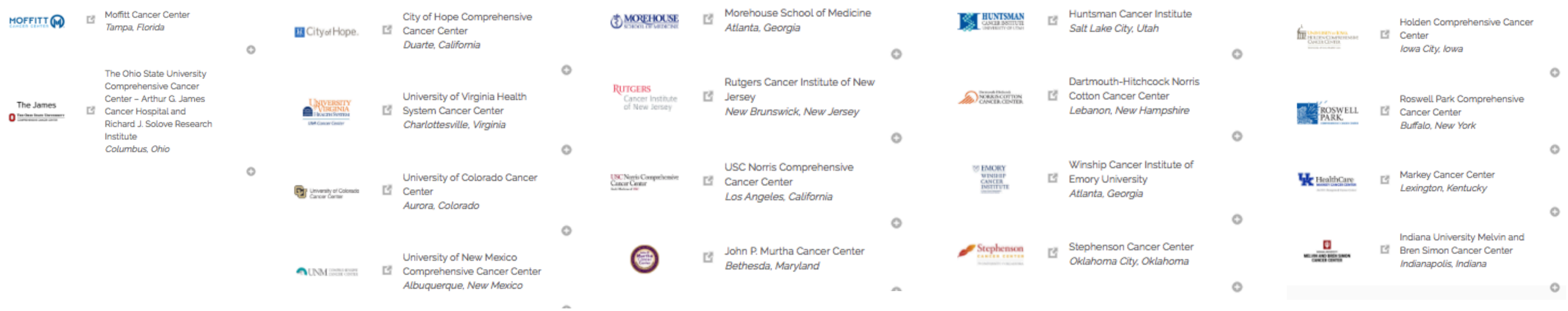
V = Velocity



ORIEN

ONCOLOGY RESEARCH
INFORMATION
EXCHANGE NETWORK

Accelerating Cancer Discovery And Delivering Hope Through Collaborative Learning And Partnerships



Moffitt Cancer Center
Tampa, Florida

The Ohio State University Comprehensive Cancer Center – Arthur G. James Cancer Hospital and Richard J. Solove Research Institute
Columbus, Ohio

City of Hope Comprehensive Cancer Center
Duarte, California

University of Virginia Health System Cancer Center
Charlottesville, Virginia

University of Colorado Cancer Center
Aurora, Colorado

University of New Mexico Comprehensive Cancer Center
Albuquerque, New Mexico

Morehouse School of Medicine
Atlanta, Georgia

Rutgers Cancer Institute of New Jersey
New Brunswick, New Jersey

USC Norris Comprehensive Cancer Center
Los Angeles, California

John P. Murtha Cancer Center
Bethesda, Maryland

Huntsman Cancer Institute
Salt Lake City, Utah

Dartmouth-Hitchcock Norris Cotton Cancer Center
Lebanon, New Hampshire

Winship Cancer Institute of Emory University
Atlanta, Georgia

Stephenson Cancer Center
Oklahoma City, Oklahoma

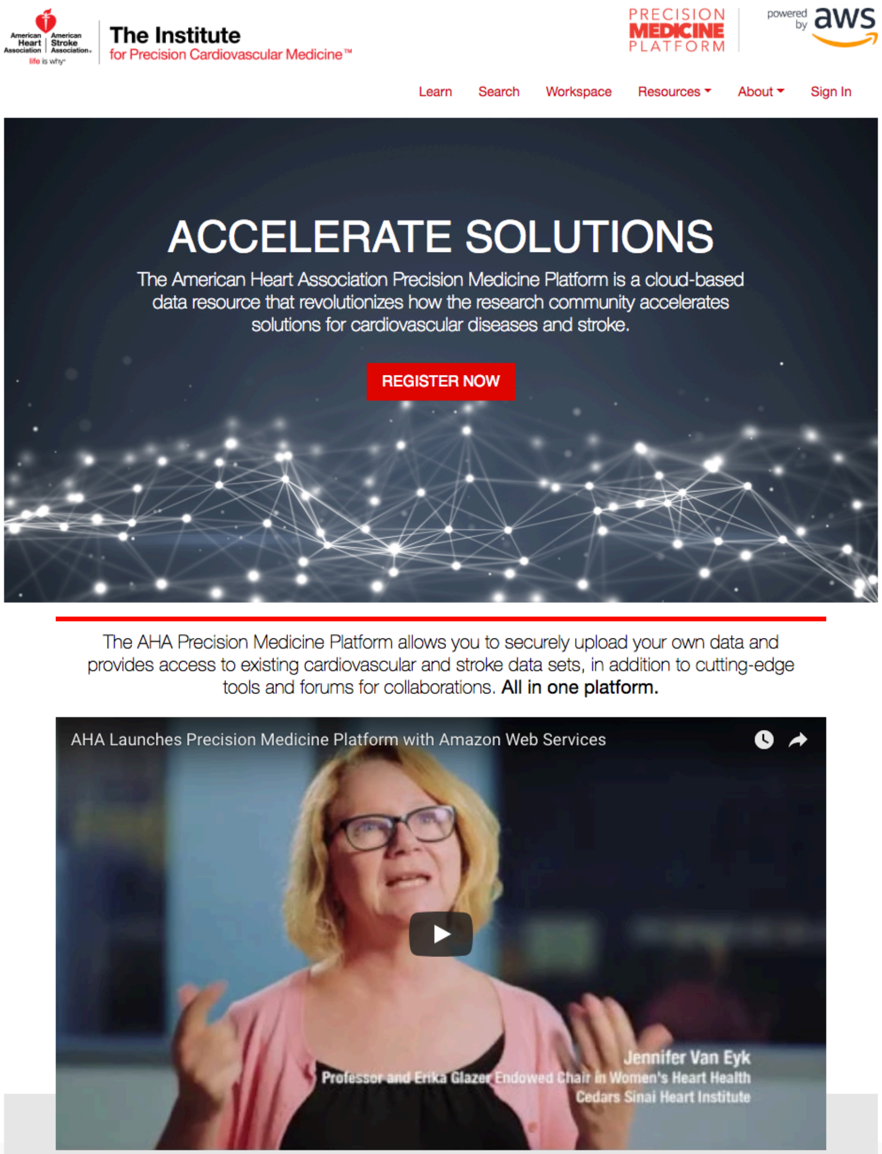
Holden Comprehensive Cancer Center
Iowa City, Iowa

Roswell Park Comprehensive Cancer Center
Buffalo, New York

Markey Cancer Center
Lexington, Kentucky

Indiana University Melvin and Bren Simon Cancer Center
Indianapolis, Indiana

V = Velocity


American Heart Association | American Stroke Association
life is why
The Institute
for Precision Cardiovascular Medicine™
PRECISION MEDICINE PLATFORM
powered by aws
Learn
Search
Workspace
Resources
About
Sign In
REGISTER NOW
Jennifer Van Eyk
Professor and Erika Glazer Endowed Chair in Women's Heart Health
Cedars Sinai Heart Institute

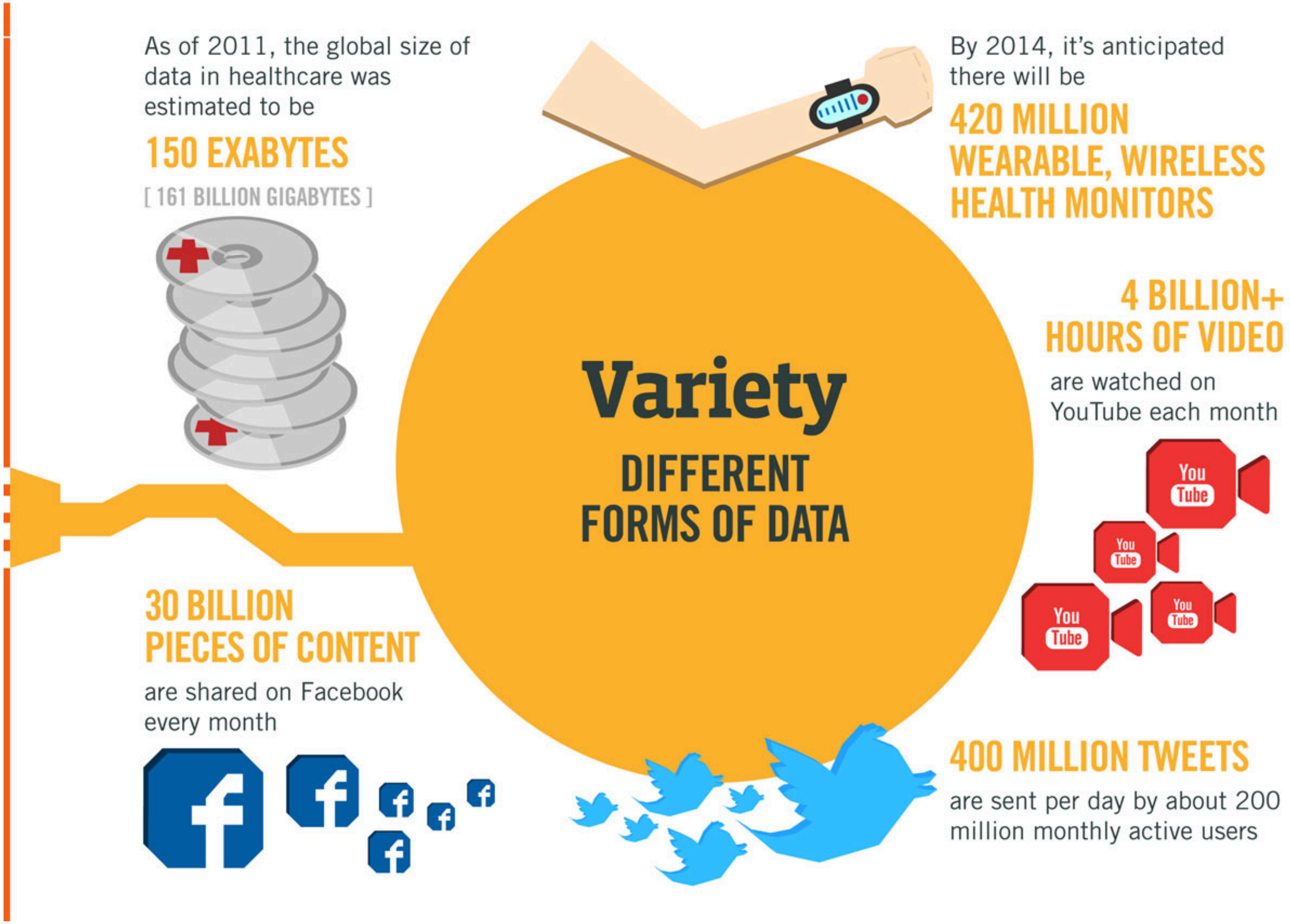
ACCELERATE SOLUTIONS

The American Heart Association Precision Medicine Platform is a cloud-based data resource that revolutionizes how the research community accelerates solutions for cardiovascular diseases and stroke.

The AHA Precision Medicine Platform allows you to securely upload your own data and provides access to existing cardiovascular and stroke data sets, in addition to cutting-edge tools and forums for collaborations. **All in one platform.**

AHA Launches Precision Medicine Platform with Amazon Web Services

V = Variety



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Variety

DIFFERENT FORMS OF DATA

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

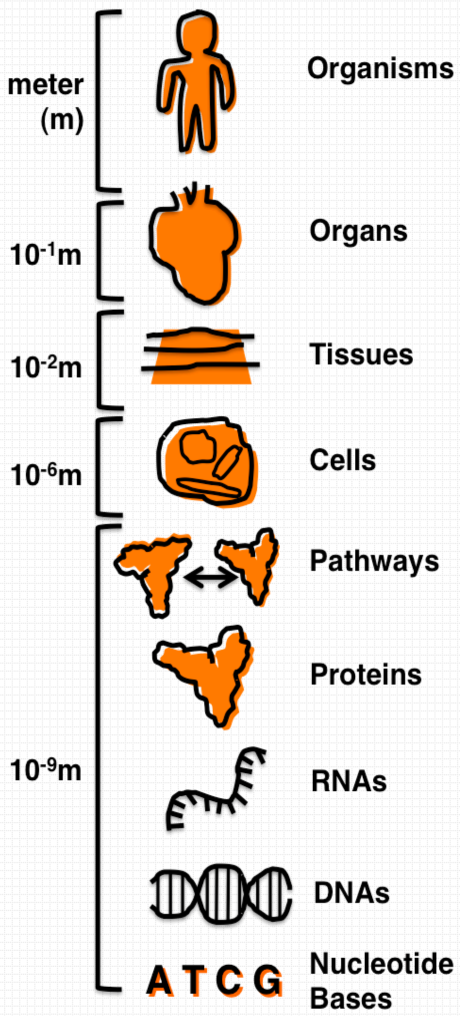
are sent per day by about 200 million monthly active users

Biomedical Big Data Challenges:

Multi-scale, Complex, Heterogeneous and Distributed

Complex & Multi-scale Biological System

Biological Scale



meter (m) — Organisms

10^{-1}m — Organs

10^{-2}m — Tissues

10^{-6}m — Cells

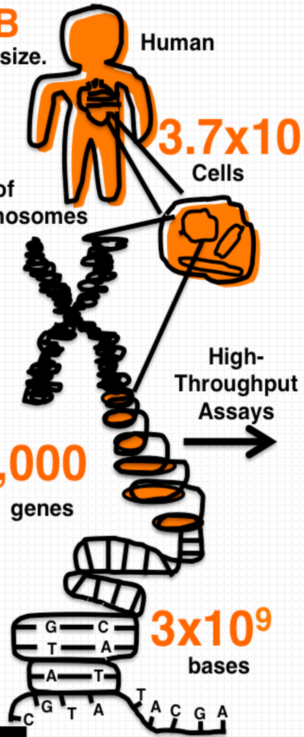
10^{-9}m — Pathways, Proteins, RNAs, DNAs, Nucleotide Bases

Examples of Large-Scale, Complex & Comprehensive Biomedical Genomics Projects

Human Genome Project

Whole genome sequencing of an individual:

4TB data size.



23 pairs of Chromosomes

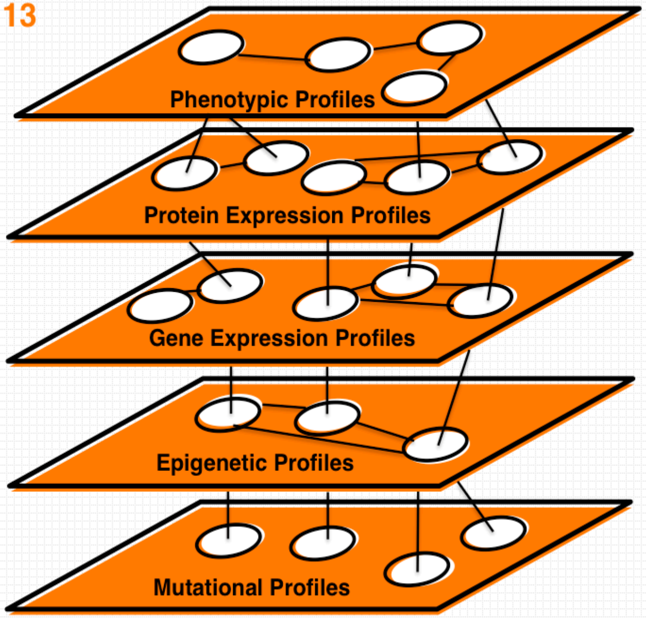
3.7x10^{13} Cells

20,000 genes

3x10^{9} bases

The Cancer Genome Atlas Project (TCGA)

Comprehensive Catalog of Molecular Profiling, Clinical Information and Imaging data of 35 cancer types. Total files currently available from CGhub: 59,998, total data file size: 766TB



ENCODE Project

99%

Comprehensive Catalog of 1% of the coding human genome.

1,649 experiments generating 15TB data.

1000 Genomes Project

A deep catalog of human genetic variation. Phase 1 genomics data of 1,092 individuals: 200TB data size.

CMap Project

A comprehensive catalog of compound-gene expression profiles:

1.3 million experiments.

Heterogeneous & Distributed Data Sources

PubMed

Comprehensive Collection of Biomedical Literature:

23 million abstracts.

Protein Data Bank

Comprehensive Collection of Protein Structures:

95,644 structures.

EBI ArrayExpress

Microarray Gene Expression Repository:

43,787 experiments

1,242,503 assays

18.5TB data size.

Sequence Traces

Comprehensive Collection of next-generation sequences:

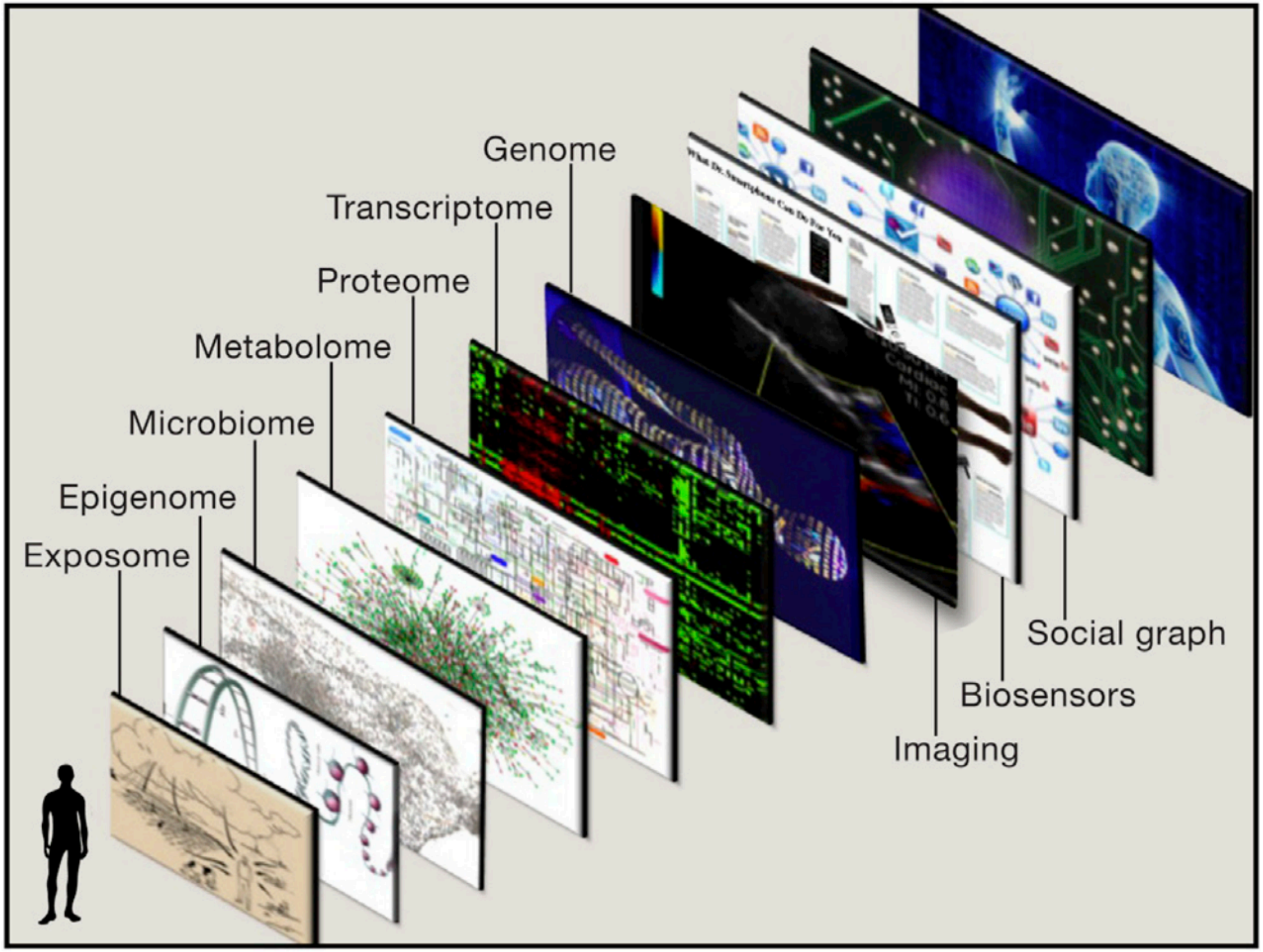
2.02 x 10^{15} bases.

PubChem

Comprehensive Collection of chemical compounds and their bioactivities:

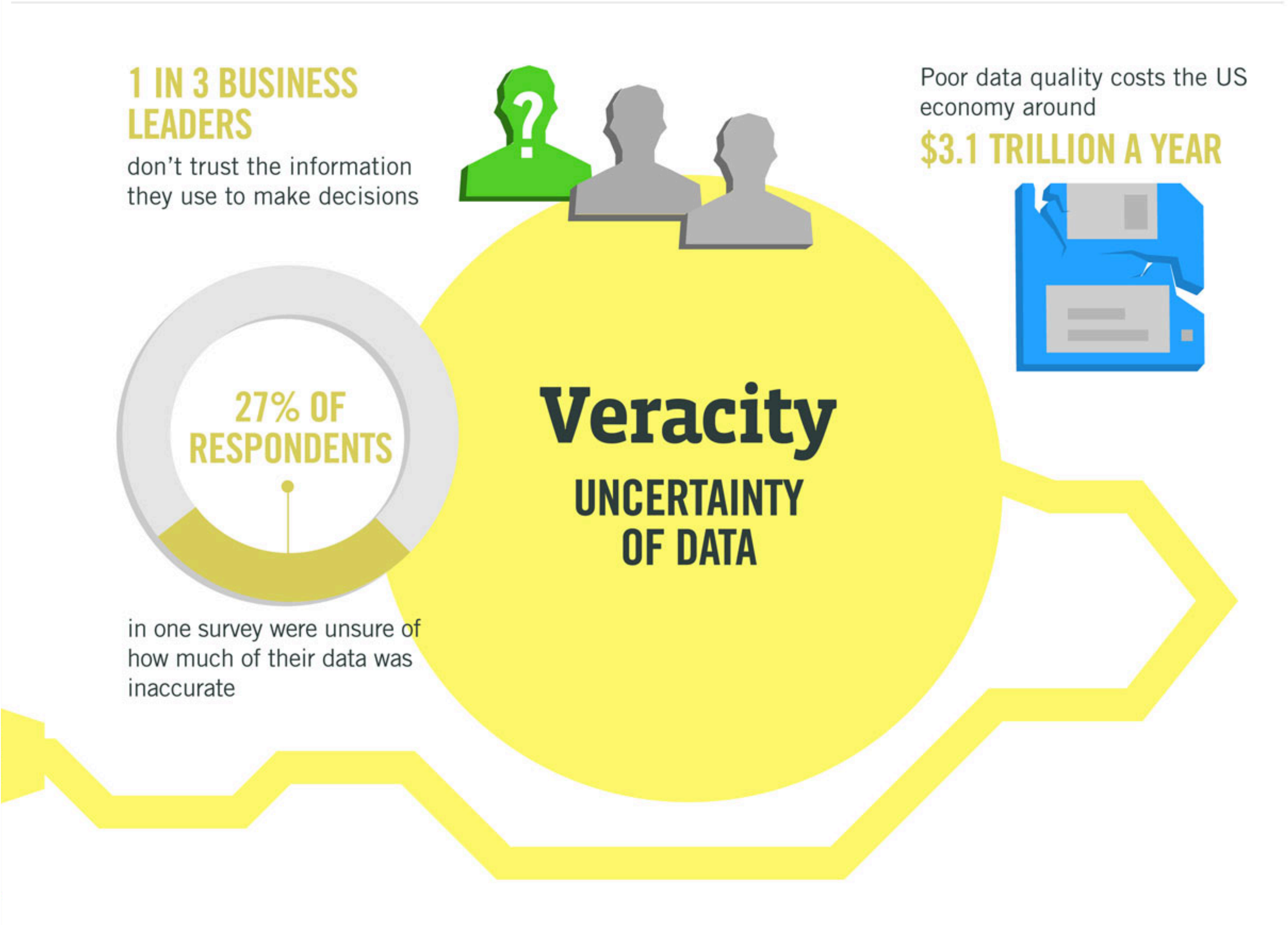
47,725,890 compounds.

V = Variety



(Eric Topol, CELL 2014)

V = Veracity



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

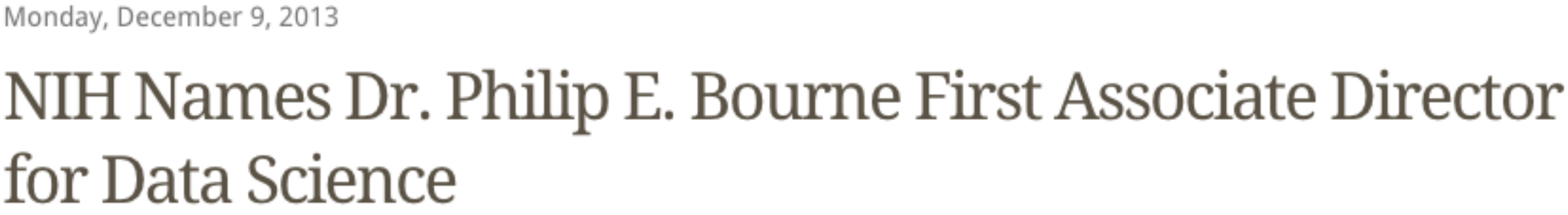
27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Veracity

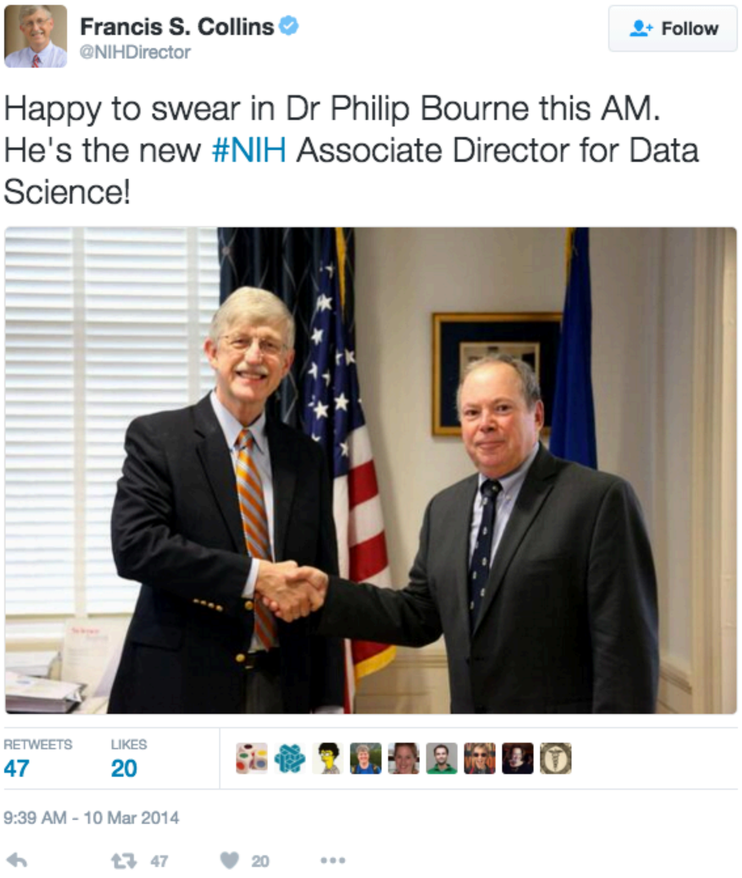
UNCERTAINTY OF DATA

V = Value



Monday, December 9, 2013

NIH Names Dr. Philip E. Bourne First Associate Director for Data Science



Francis S. Collins
@NIHDirector

Happy to swear in Dr Philip Bourne this AM. He's the new #NIH Associate Director for Data Science!

RETWEETS	LIKES
47	20

9:39 AM - 10 Mar 2014

V = Value

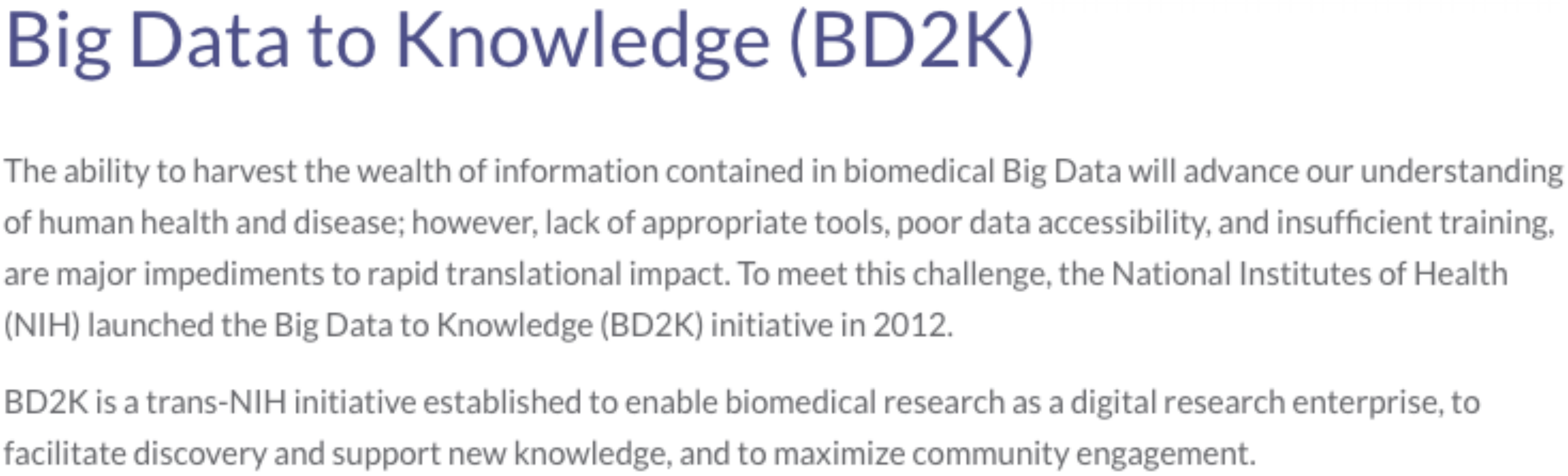
NIH Data Science at NIH

Enabling Biomedical Scientists to capitalize more fully on the Big Data being generated by the research communities

https://www.youtube.com/channel/UCKIDQOa0JcUd3K9C1TS7FLQ/videos

NIH

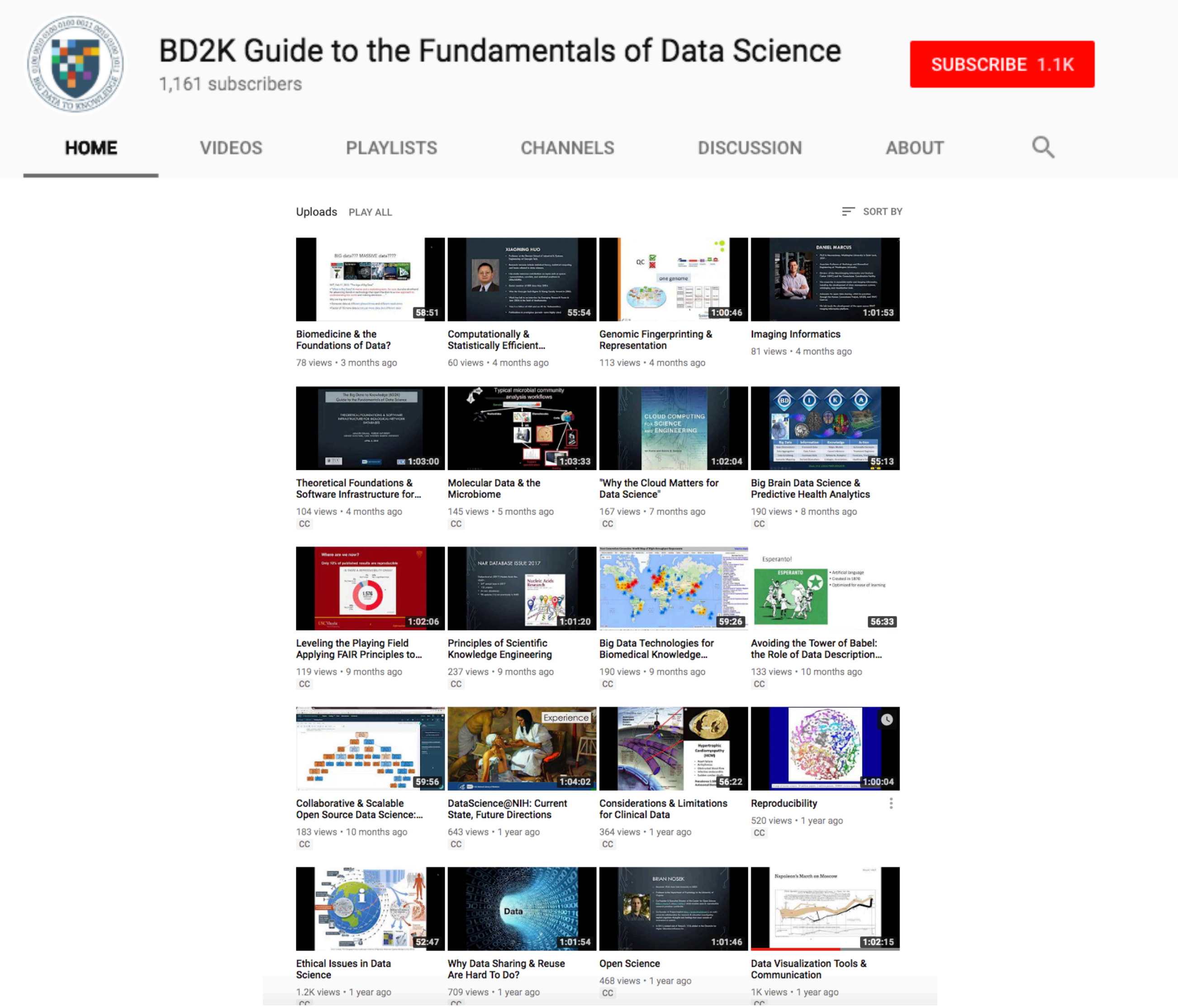
Big Data to Knowledge(BD2K)


Big Data to Knowledge (BD2K)

The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.

BD2K is a trans-NIH initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge, and to maximize community engagement.

V = Value



BD2K Guide to the Fundamentals of Data Science

1,161 subscribers

58:51 Biomedicine & the Foundations of Data? 78 views • 3 months ago

55:54 Computationally & Statistically Efficient... 60 views • 4 months ago

1:00:46 Genomic Fingerprinting & Representation 113 views • 4 months ago

1:01:53 Imaging Informatics 81 views • 4 months ago

1:03:00 Theoretical Foundations & Software Infrastructure for... 104 views • 4 months ago CC

1:03:33 Molecular Data & the Microbiome 145 views • 5 months ago CC

1:02:04 "Why the Cloud Matters for Data Science" 167 views • 7 months ago CC

55:13 Big Brain Data Science & Predictive Health Analytics 190 views • 8 months ago CC

1:02:06 Leveling the Playing Field Applying FAIR Principles to... 119 views • 9 months ago CC

1:01:20 Principles of Scientific Knowledge Engineering 237 views • 9 months ago CC

59:26 Big Data Technologies for Biomedical Knowledge... 190 views • 9 months ago CC

56:33 Avoiding the Tower of Babel: the Role of Data Description... 133 views • 10 months ago CC

59:56 Collaborative & Scalable Open Source Data Science:... 183 views • 10 months ago CC

1:04:02 DataScience@NIH: Current State, Future Directions 643 views • 1 year ago CC

56:22 Considerations & Limitations for Clinical Data 364 views • 1 year ago CC

1:00:04 Reproducibility 520 views • 1 year ago CC

52:47 Ethical Issues in Data Science 1.2K views • 1 year ago CC

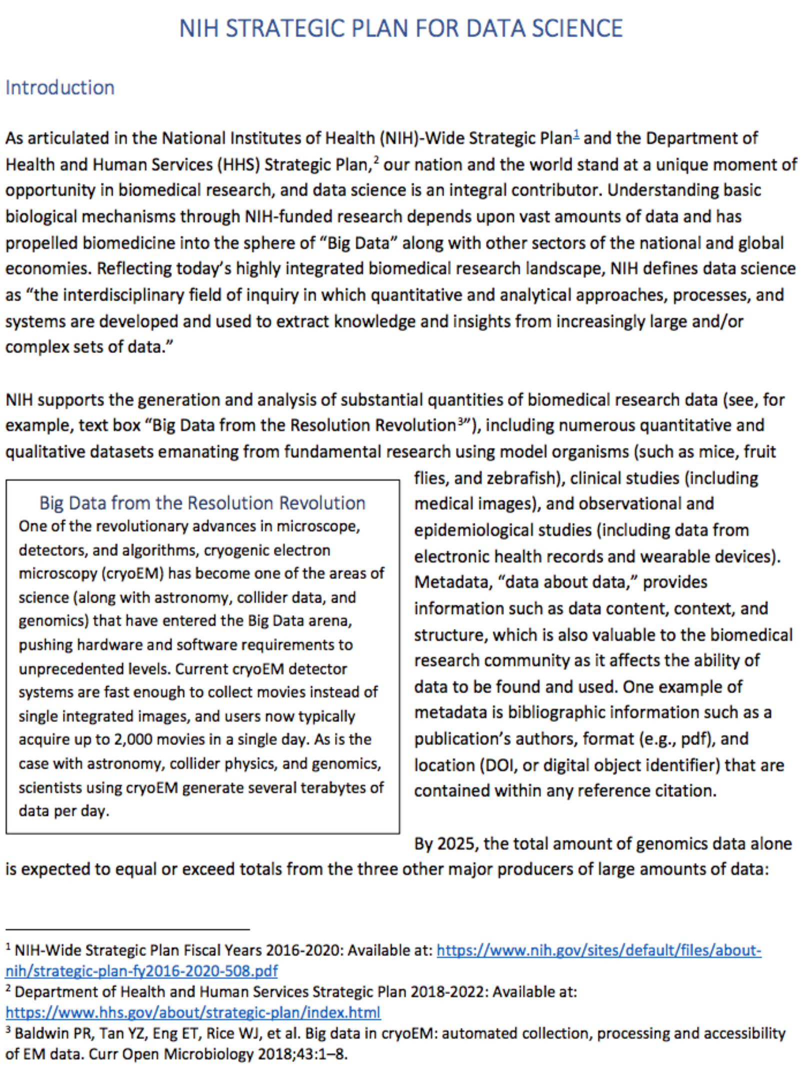
1:01:54 Why Data Sharing & Reuse Are Hard To Do? 709 views • 1 year ago CC

1:01:46 Open Science 468 views • 1 year ago CC

1:02:15 Data Visualization Tools & Communication 1K views • 1 year ago CC

https://www.youtube.com/channel/UCKIDQOa0JcUd3K9C1TS7FLQ/videos

V = Value



NIH STRATEGIC PLAN FOR DATA SCIENCE

Introduction

As articulated in the National Institutes of Health (NIH)-Wide Strategic Plan[1] and the Department of Health and Human Services (HHS) Strategic Plan,[2] our nation and the world stand at a unique moment of opportunity in biomedical research, and data science is an integral contributor. Understanding basic biological mechanisms through NIH-funded research depends upon vast amounts of data and has propelled biomedicine into the sphere of "Big Data" along with other sectors of the national and global economies. Reflecting today's highly integrated biomedical research landscape, NIH defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

NIH supports the generation and analysis of substantial quantities of biomedical research data (see, for example, text box "Big Data from the Resolution Revolution[3]"), including numerous quantitative and qualitative datasets emanating from fundamental research using model organisms (such as mice, fruit flies, and zebrafish), clinical studies (including medical images), and observational and epidemiological studies (including data from electronic health records and wearable devices). Metadata, "data about data," provides information such as data content, context, and structure, which is also valuable to the biomedical research community as it affects the ability of data to be found and used. One example of metadata is bibliographic information such as a publication's authors, format (e.g., pdf), and location (DOI, or digital object identifier) that are contained within any reference citation.

Big Data from the Resolution Revolution

One of the revolutionary advances in microscope, detectors, and algorithms, cryogenic electron microscopy (cryoEM) has become one of the areas of science (along with astronomy, collider data, and genomics) that have entered the Big Data arena, pushing hardware and software requirements to unprecedented levels. Current cryoEM detector systems are fast enough to collect movies instead of single integrated images, and users now typically acquire up to 2,000 movies in a single day. As is the case with astronomy, collider physics, and genomics, scientists using cryoEM generate several terabytes of data per day.

By 2025, the total amount of genomics data alone is expected to equal or exceed totals from the three other major producers of large amounts of data:

[1] NIH-Wide Strategic Plan Fiscal Years 2016-2020: Available at: https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2016-2020-508.pdf

[2] Department of Health and Human Services Strategic Plan 2018-2022: Available at: https://www.hhs.gov/about/strategic-plan/index.html

[3] Baldwin PR, Tan YZ, Eng ET, Rice WJ, et al. Big data in cryoEM: automated collection, processing and accessibility of EM data. Curr Open Microbiology 2018;43:1–8.

U.S. Department of Health and Human Services | National Institutes of Hea

NIH DataScience@NIH

Search

Data Science Community | BD2K | Blog | News & Events | NIH Strategic Plan for Data Science | About

For the latest information on plans for data science at NIH,

VISIT THE NEWS AND EVENTS PAGE

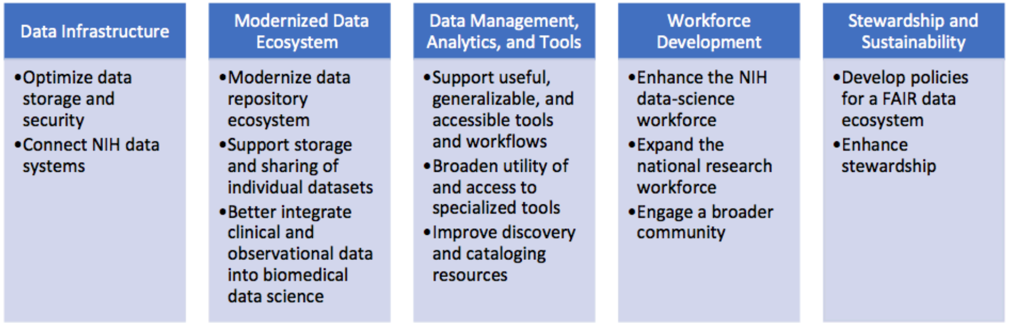
○ ○ ○ ●

Data Science Community

Building a digital ecosystem for biomedicine.

Go to Data Science Community Page

https://datascience.nih.gov/

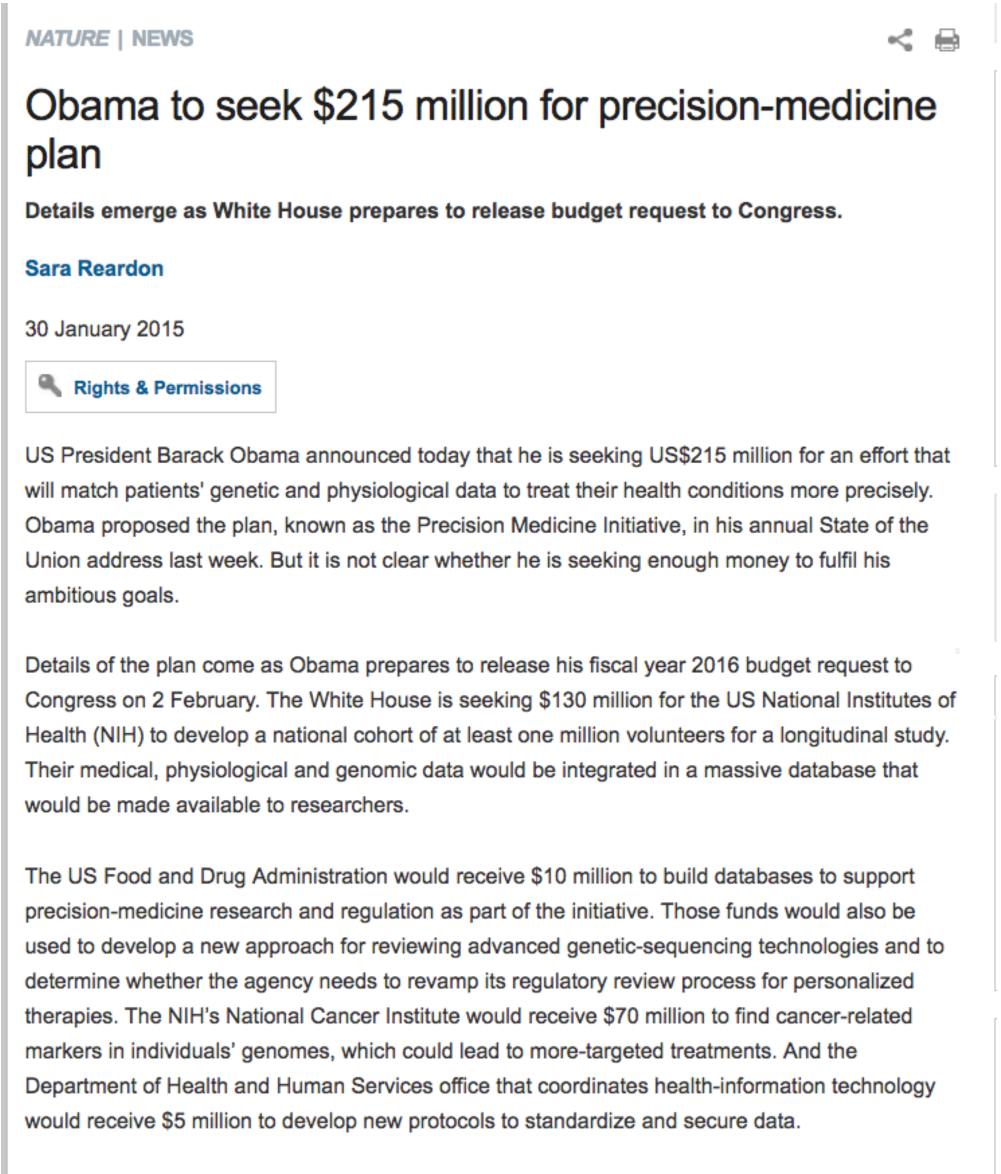


Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
•Optimize data storage and security •Connect NIH data systems	•Modernize data repository ecosystem •Support storage and sharing of individual datasets •Better integrate clinical and observational data into biomedical data science	•Support useful, generalizable, and accessible tools and workflows •Broaden utility of and access to specialized tools •Improve discovery and cataloging resources	•Enhance the NIH data-science workforce •Expand the national research workforce •Engage a broader community	•Develop policies for a FAIR data ecosystem •Enhance stewardship

Figure 2. NIH Strategic Plan for Data Science: Overview of Goals and Objectives

https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

V = Value



NATURE | NEWS

Obama to seek $215 million for precision-medicine plan

Details emerge as White House prepares to release budget request to Congress.

Sara Reardon

30 January 2015

Rights & Permissions

US President Barack Obama announced today that he is seeking US$215 million for an effort that will match patients' genetic and physiological data to treat their health conditions more precisely. Obama proposed the plan, known as the Precision Medicine Initiative, in his annual State of the Union address last week. But it is not clear whether he is seeking enough money to fulfil his ambitious goals.

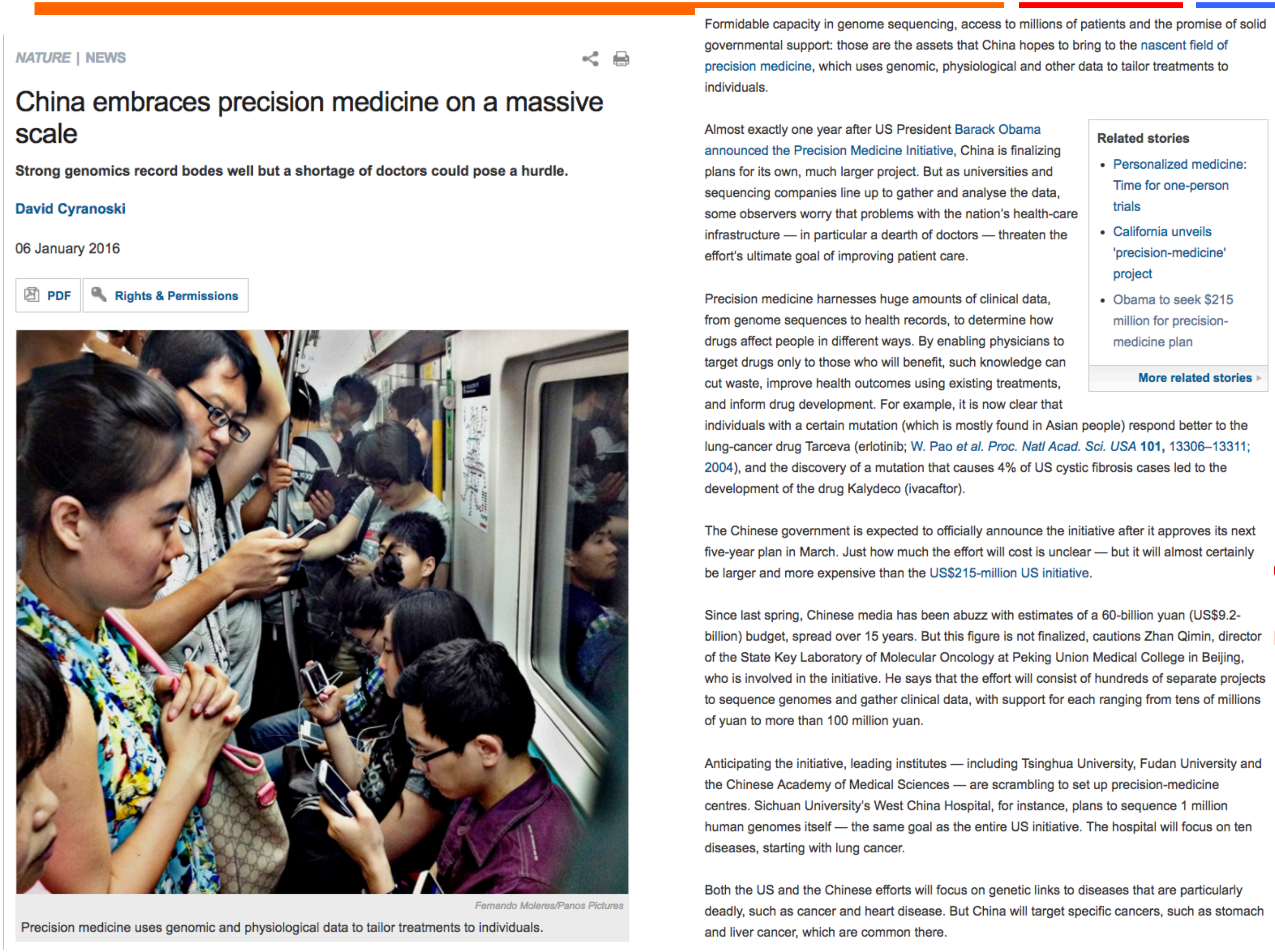
Details of the plan come as Obama prepares to release his fiscal year 2016 budget request to Congress on 2 February. The White House is seeking $130 million for the US National Institutes of Health (NIH) to develop a national cohort of at least one million volunteers for a longitudinal study. Their medical, physiological and genomic data would be integrated in a massive database that would be made available to researchers.

The US Food and Drug Administration would receive $10 million to build databases to support precision-medicine research and regulation as part of the initiative. Those funds would also be used to develop a new approach for reviewing advanced genetic-sequencing technologies and to determine whether the agency needs to revamp its regulatory review process for personalized therapies. The NIH's National Cancer Institute would receive $70 million to find cancer-related markers in individuals' genomes, which could lead to more-targeted treatments. And the Department of Health and Human Services office that coordinates health-information technology would receive $5 million to develop new protocols to standardize and secure data.

USA Budget:

US$ 215 million

V = Value



NATURE | NEWS

China embraces precision medicine on a massive scale

Strong genomics record bodes well but a shortage of doctors could pose a hurdle.

David Cyranoski

06 January 2016

Fernando Moleres/Panos Pictures

Precision medicine uses genomic and physiological data to tailor treatments to individuals.

Formidable capacity in genome sequencing, access to millions of patients and the promise of solid governmental support: those are the assets that China hopes to bring to the nascent field of precision medicine, which uses genomic, physiological and other data to tailor treatments to individuals.

Almost exactly one year after US President Barack Obama announced the Precision Medicine Initiative, China is finalizing plans for its own, much larger project. But as universities and sequencing companies line up to gather and analyse the data, some observers worry that problems with the nation's health-care infrastructure — in particular a dearth of doctors — threaten the effort's ultimate goal of improving patient care.

Related stories

- Personalized medicine: Time for one-person trials
- California unveils 'precision-medicine' project
- Obama to seek $215 million for precision-medicine plan

More related stories

Precision medicine harnesses huge amounts of clinical data, from genome sequences to health records, to determine how drugs affect people in different ways. By enabling physicians to target drugs only to those who will benefit, such knowledge can cut waste, improve health outcomes using existing treatments, and inform drug development. For example, it is now clear that individuals with a certain mutation (which is mostly found in Asian people) respond better to the lung-cancer drug Tarceva (erlotinib; W. Pao *et al. Proc. Natl Acad. Sci. USA* **101,** 13306–13311; 2004), and the discovery of a mutation that causes 4% of US cystic fibrosis cases led to the development of the drug Kalydeco (ivacaftor).

The Chinese government is expected to officially announce the initiative after it approves its next five-year plan in March. Just how much the effort will cost is unclear — but it will almost certainly be larger and more expensive than the US$215-million US initiative.

Since last spring, Chinese media has been abuzz with estimates of a 60-billion yuan (US$9.2-billion) budget, spread over 15 years. But this figure is not finalized, cautions Zhan Qimin, director of the State Key Laboratory of Molecular Oncology at Peking Union Medical College in Beijing, who is involved in the initiative. He says that the effort will consist of hundreds of separate projects to sequence genomes and gather clinical data, with support for each ranging from tens of millions of yuan to more than 100 million yuan.

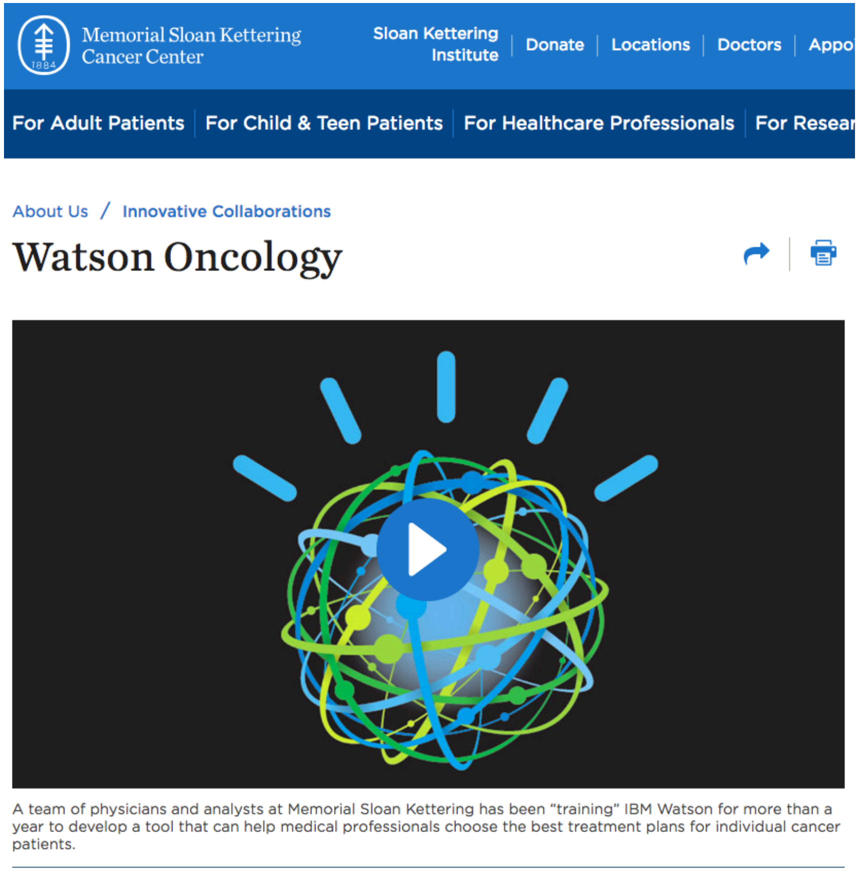
Anticipating the initiative, leading institutes — including Tsinghua University, Fudan University and the Chinese Academy of Medical Sciences — are scrambling to set up precision-medicine centres. Sichuan University's West China Hospital, for instance, plans to sequence 1 million human genomes itself — the same goal as the entire US initiative. The hospital will focus on ten diseases, starting with lung cancer.

Both the US and the Chinese efforts will focus on genetic links to diseases that are particularly deadly, such as cancer and heart disease. But China will target specific cancers, such as stomach and liver cancer, which are common there.

Chinese Budget:

US$ 9.2 billion!!!

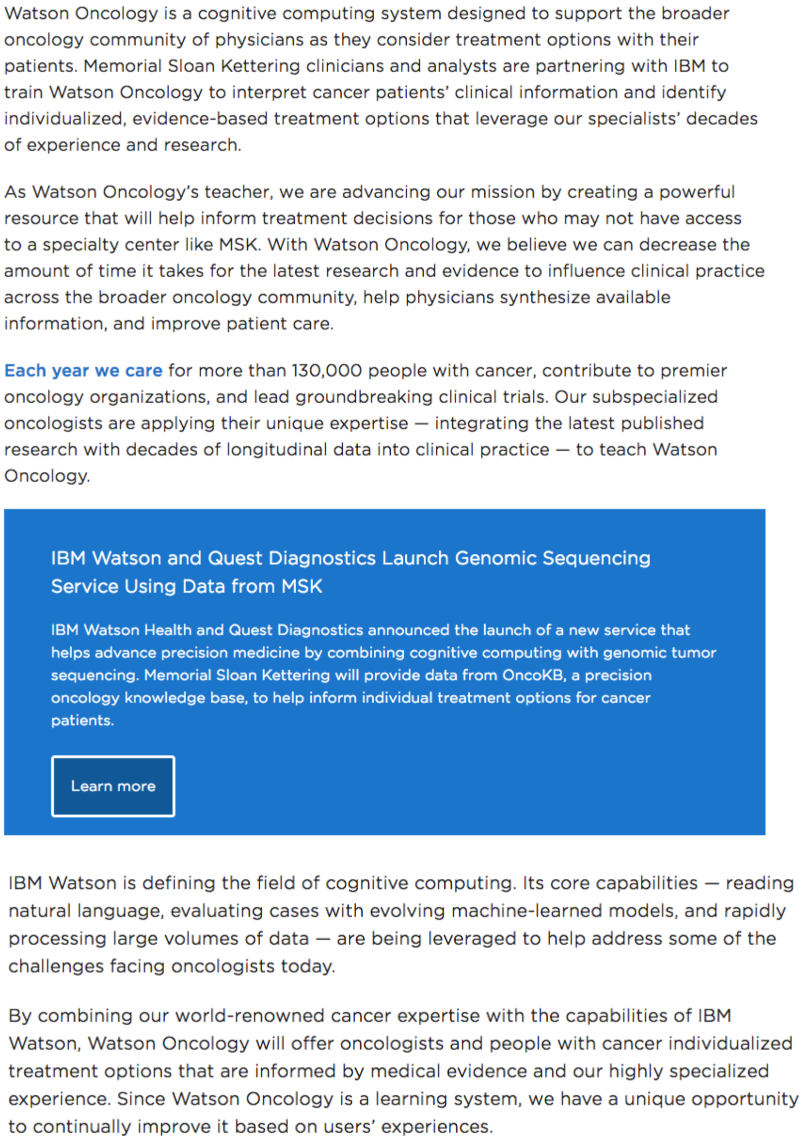
V = Value


Memorial Sloan Kettering Cancer Center
1884
Sloan Kettering Institute
Donate
Locations
Doctors
Appo
For Adult Patients
For Child & Teen Patients
For Healthcare Professionals
For Resear

About Us / Innovative Collaborations

Watson Oncology

A team of physicians and analysts at Memorial Sloan Kettering has been “training” IBM Watson for more than a year to develop a tool that can help medical professionals choose the best treatment plans for individual cancer patients.



Watson Oncology is a cognitive computing system designed to support the broader oncology community of physicians as they consider treatment options with their patients. Memorial Sloan Kettering clinicians and analysts are partnering with IBM to train Watson Oncology to interpret cancer patients’ clinical information and identify individualized, evidence-based treatment options that leverage our specialists’ decades of experience and research.

As Watson Oncology’s teacher, we are advancing our mission by creating a powerful resource that will help inform treatment decisions for those who may not have access to a specialty center like MSK. With Watson Oncology, we believe we can decrease the amount of time it takes for the latest research and evidence to influence clinical practice across the broader oncology community, help physicians synthesize available information, and improve patient care.

Each year we care for more than 130,000 people with cancer, contribute to premier oncology organizations, and lead groundbreaking clinical trials. Our subspecialized oncologists are applying their unique expertise — integrating the latest published research with decades of longitudinal data into clinical practice — to teach Watson Oncology.

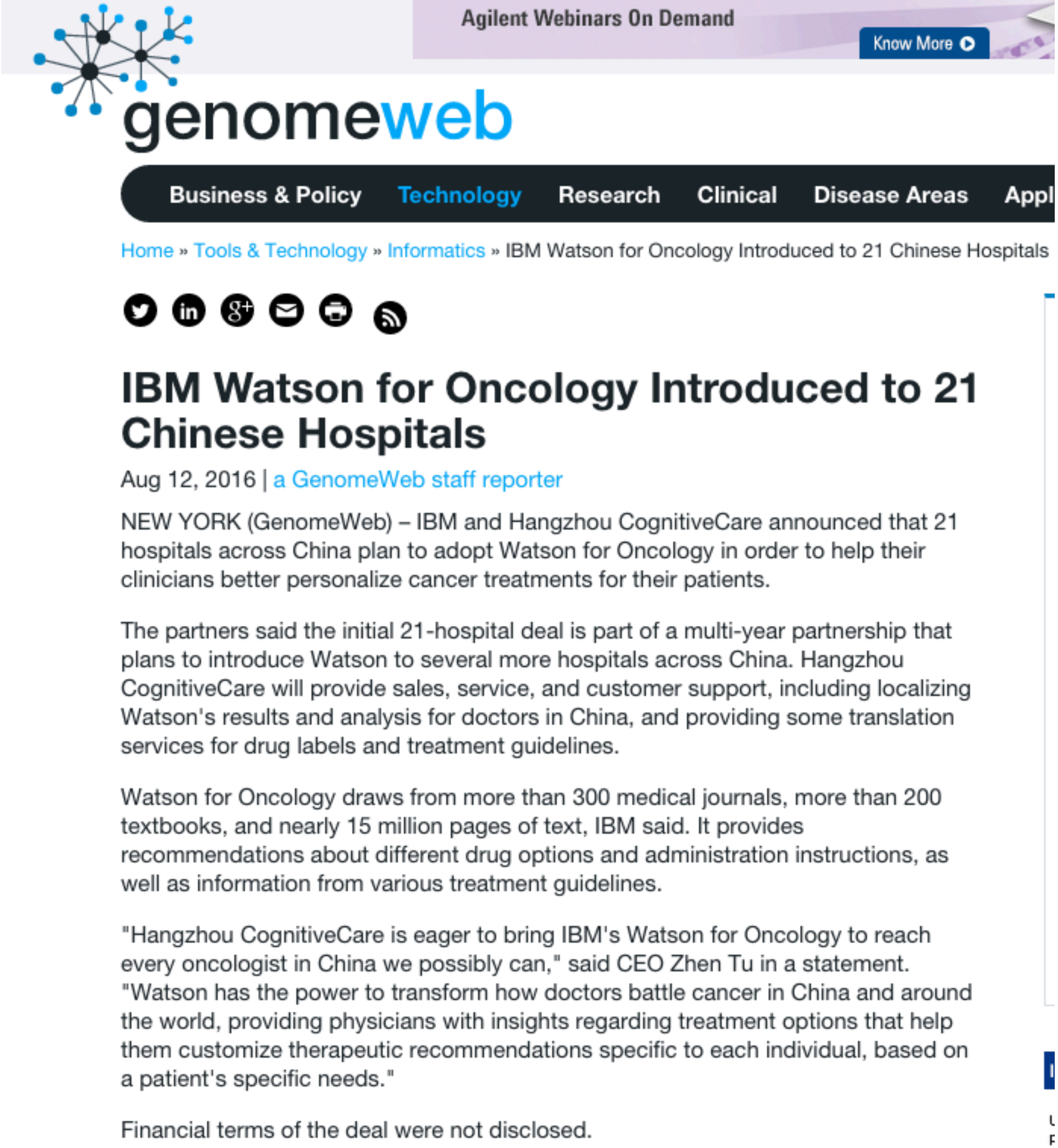
IBM Watson and Quest Diagnostics Launch Genomic Sequencing Service Using Data from MSK

IBM Watson Health and Quest Diagnostics announced the launch of a new service that helps advance precision medicine by combining cognitive computing with genomic tumor sequencing. Memorial Sloan Kettering will provide data from OncoKB, a precision oncology knowledge base, to help inform individual treatment options for cancer patients.

IBM Watson is defining the field of cognitive computing. Its core capabilities — reading natural language, evaluating cases with evolving machine-learned models, and rapidly processing large volumes of data — are being leveraged to help address some of the challenges facing oncologists today.

By combining our world-renowned cancer expertise with the capabilities of IBM Watson, Watson Oncology will offer oncologists and people with cancer individualized treatment options that are informed by medical evidence and our highly specialized experience. Since Watson Oncology is a learning system, we have a unique opportunity to continually improve it based on users’ experiences.

https://www.mskcc.org/about/innovative-collaborations/watson-oncology



IBM Watson for Oncology Introduced to 21 Chinese Hospitals

Aug 12, 2016 | a GenomeWeb staff reporter

NEW YORK (GenomeWeb) – IBM and Hangzhou CognitiveCare announced that 21 hospitals across China plan to adopt Watson for Oncology in order to help their clinicians better personalize cancer treatments for their patients.

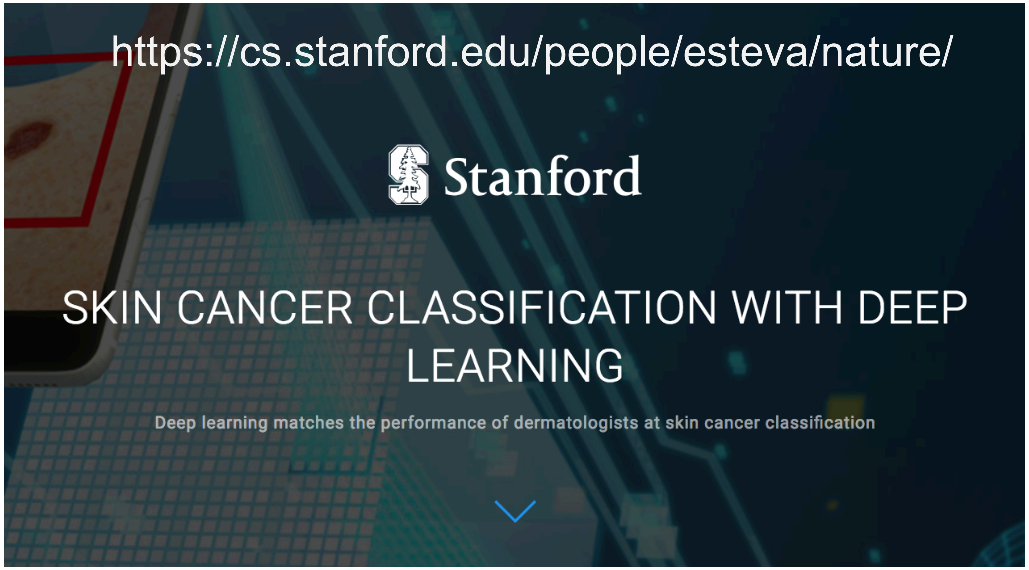
The partners said the initial 21-hospital deal is part of a multi-year partnership that plans to introduce Watson to several more hospitals across China. Hangzhou CognitiveCare will provide sales, service, and customer support, including localizing Watson's results and analysis for doctors in China, and providing some translation services for drug labels and treatment guidelines.

Watson for Oncology draws from more than 300 medical journals, more than 200 textbooks, and nearly 15 million pages of text, IBM said. It provides recommendations about different drug options and administration instructions, as well as information from various treatment guidelines.

"Hangzhou CognitiveCare is eager to bring IBM's Watson for Oncology to reach every oncologist in China we possibly can," said CEO Zhen Tu in a statement. "Watson has the power to transform how doctors battle cancer in China and around the world, providing physicians with insights regarding treatment options that help them customize therapeutic recommendations specific to each individual, based on a patient's specific needs."

Financial terms of the deal were not disclosed.

V = Value

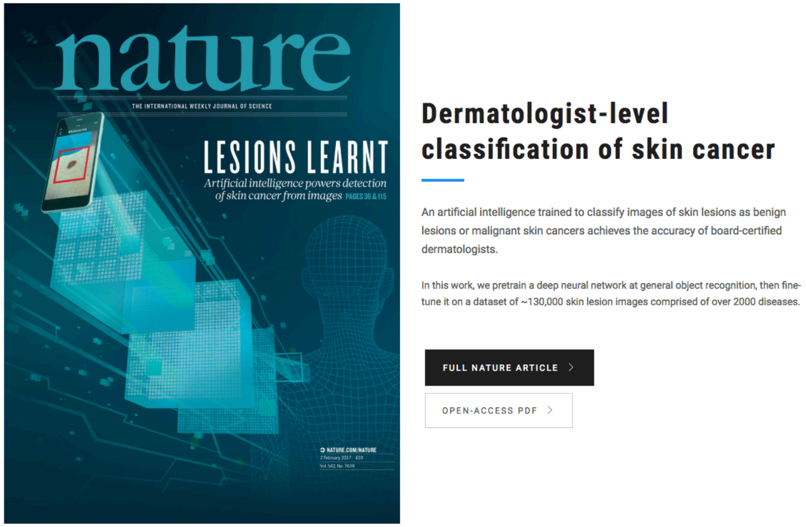


https://cs.stanford.edu/people/esteva/nature/

Stanford

SKIN CANCER CLASSIFICATION WITH DEEP LEARNING

Deep learning matches the performance of dermatologists at skin cancer classification


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
LESIONS LEARNT
Artificial intelligence powers detection of skin cancer from images

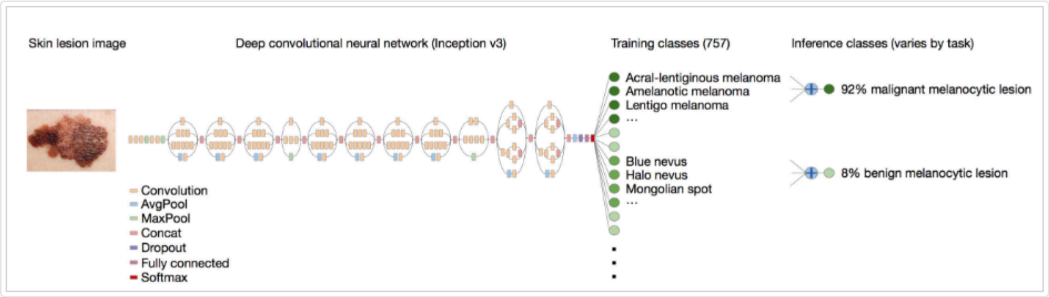
Dermatologist-level classification of skin cancer

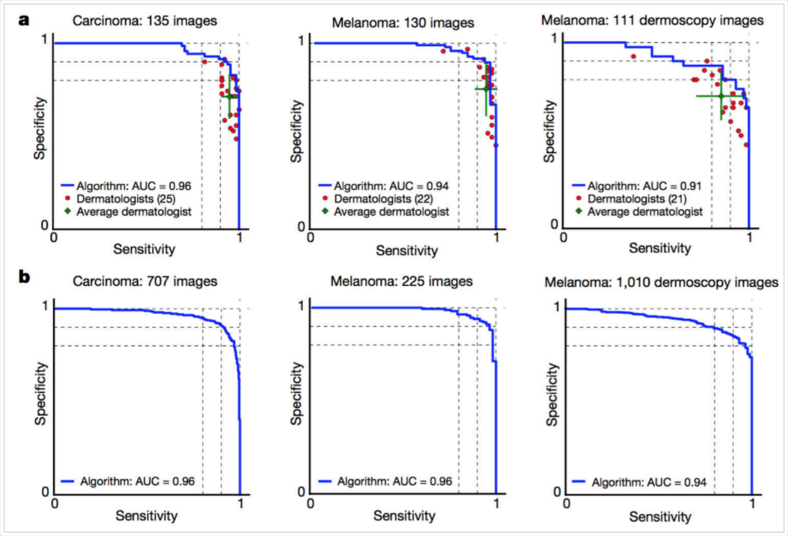
An artificial intelligence trained to classify images of skin lesions as benign lesions or malignant skin cancers achieves the accuracy of board-certified dermatologists.

In this work, we pretrain a deep neural network at general object recognition, then fine-tune it on a dataset of ~130,000 skin lesion images comprised of over 2000 diseases.

FULL NATURE ARTICLE >

OPEN-ACCESS PDF >

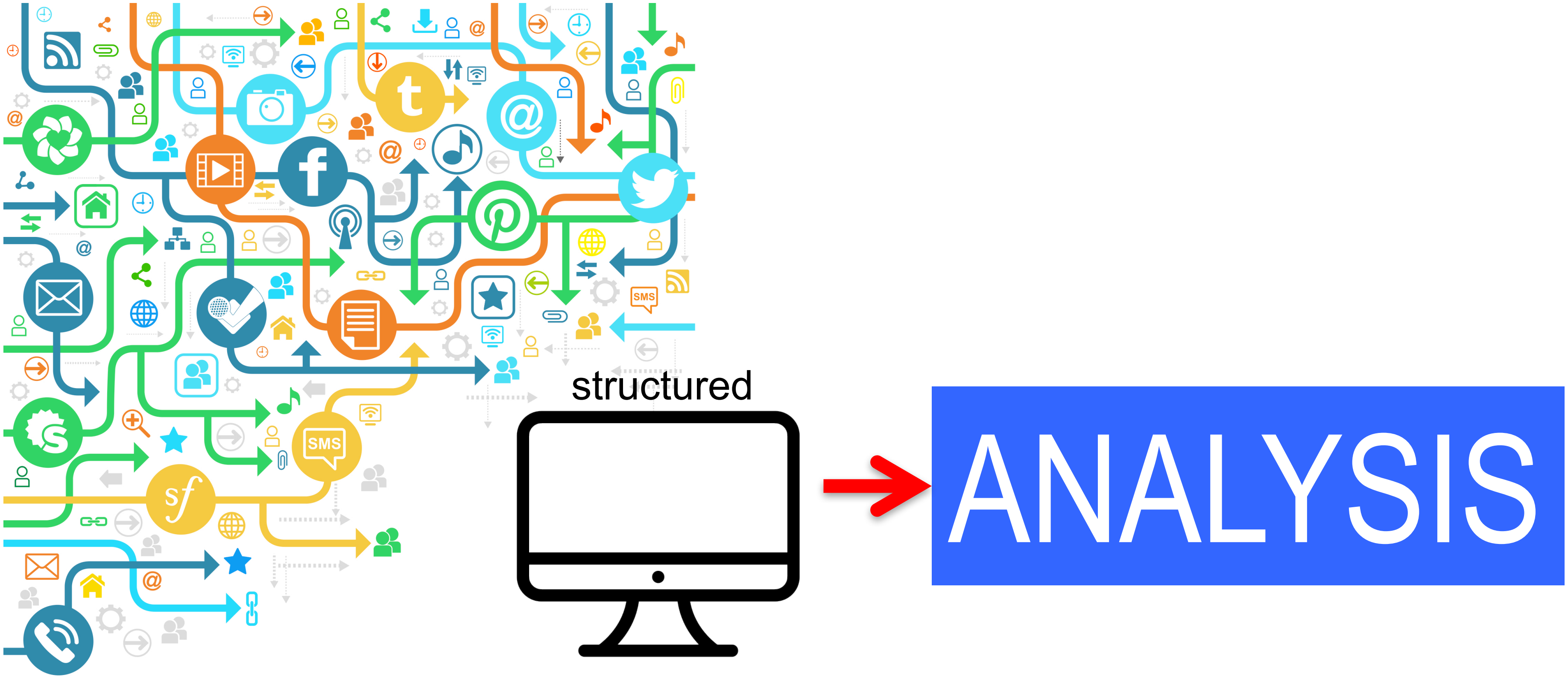




https://www.youtube.com/watch?v=IvmLEq9piJ4

Data

- Structured – transactions
- Unstructured – text



structured

ANALYSIS

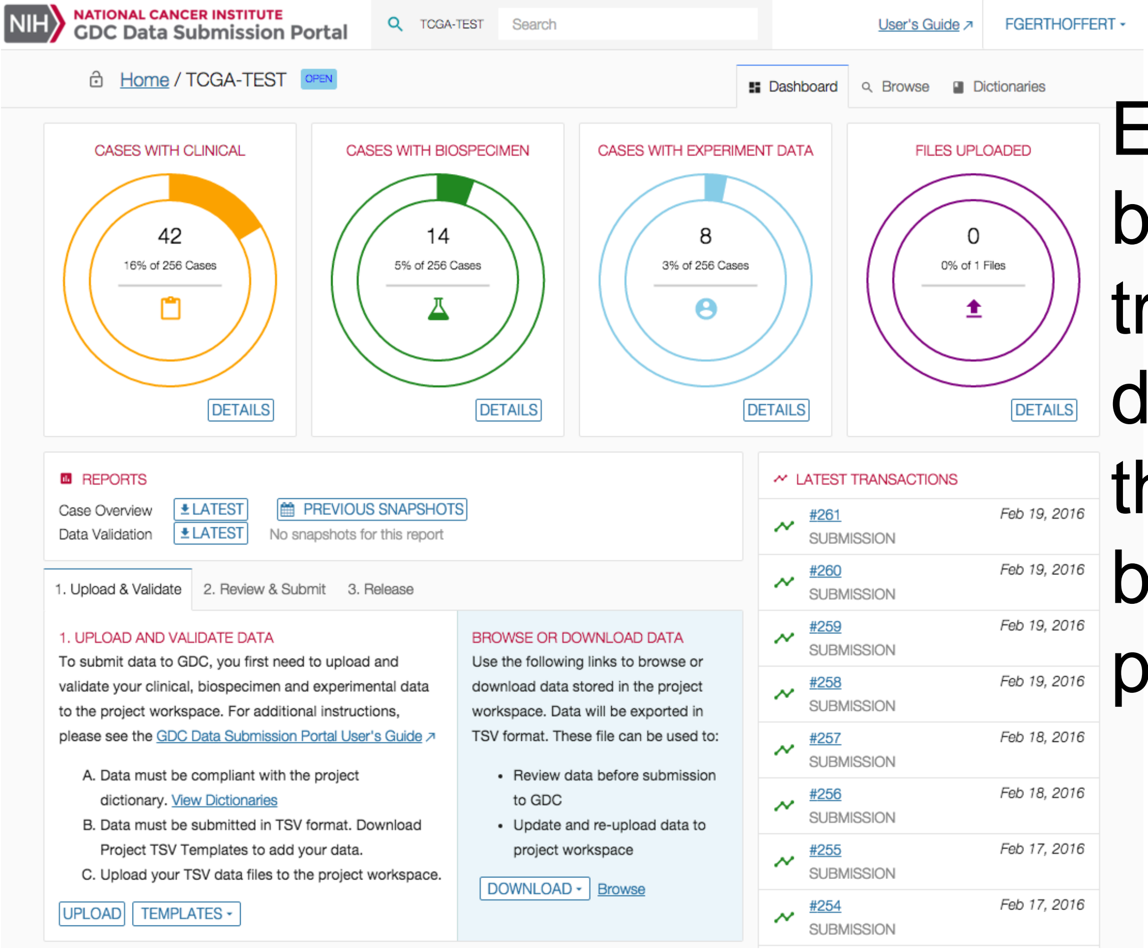
Data Analytics

The use of machine learning and statistics to derive meaning from data in order to make better decisions (*translating big data to knowledge*)

Three Types of Analytics

1. Descriptive
2. Predictive
3. Prescription

Descriptive Analytics



NIH NATIONAL CANCER INSTITUTE GDC Data Submission Portal

TCGA-TEST Search User's Guide ↗ FGERTHOFFERT ▾

Home / TCGA-TEST OPEN

Dashboard Browse Dictionaries

CASES WITH CLINICAL	CASES WITH BIOSPECIMEN	CASES WITH EXPERIMENT DATA	FILES UPLOADED
42 16% of 256 Cases	14 5% of 256 Cases	8 3% of 256 Cases	0 0% of 1 Files
DETAILS	DETAILS	DETAILS	DETAILS

REPORTS

Case Overview ⬇LATEST PREVIOUS SNAPSHOTS

Data Validation ⬇LATEST No snapshots for this report

1. Upload & Validate 2. Review & Submit 3. Release

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 dictionary. View Dictionaries
B. Data must be submitted in TSV format. Download Project TSV Templates to add your data.
C. Upload your TSV data files to the project workspace.

UPLOAD TEMPLATES ▾

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 submission to GDC
- Update and re-upload data to project workspace

DOWNLOAD ▾ Browse

LATEST TRANSACTIONS

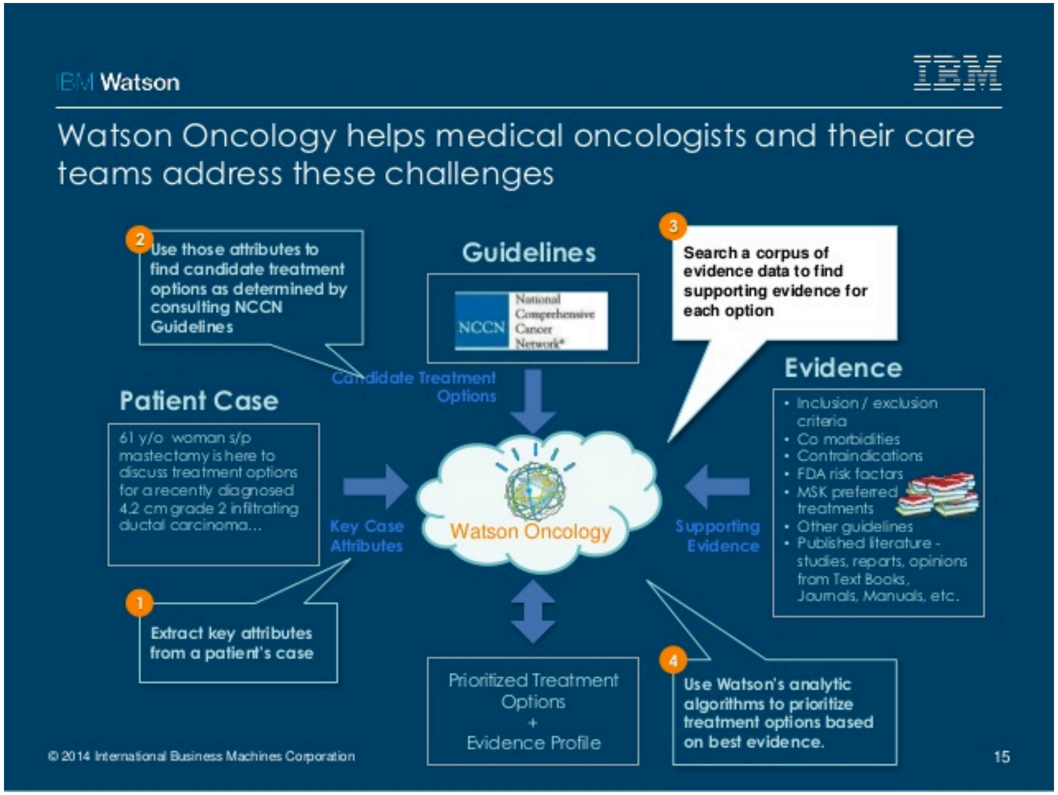
#261 SUBMISSION	Feb 19, 2016
#260 SUBMISSION	Feb 19, 2016
#259 SUBMISSION	Feb 19, 2016
#258 SUBMISSION	Feb 19, 2016
#257 SUBMISSION	Feb 18, 2016
#256 SUBMISSION	Feb 18, 2016
#255 SUBMISSION	Feb 17, 2016
#254 SUBMISSION	Feb 17, 2016

Example: Dash boards, portals, trends, alerts – displaying data from the past (history), but no predictive power

Predictive Analytics

Example: machine learning and statistical tools - Use data to build models that can predict future “unseen” situation.

Prescriptive Analytics



Watson Oncology helps medical oncologists and their care teams address these challenges

1. Extract key attributes from a patient's case

2. Use those attributes to find candidate treatment options as determined by consulting NCCN Guidelines

3. Search a corpus of evidence data to find supporting evidence for each option

4. Use Watson's analytic algorithms to prioritize treatment options based on best evidence.

Patient Case

61 y/o woman s/p mastectomy is here to discuss treatment options for a recently diagnosed 4.2 cm grade 2 infiltrating ductal carcinoma…

Guidelines

NCCN
National Comprehensive Cancer Network®
Candidate Treatment Options
Key Case Attributes
Watson Oncology
Supporting Evidence

Evidence

- Inclusion / exclusion criteria
- Co morbidities
- Contraindications
- FDA risk factors
- MSK preferred treatments
- Other guidelines
- Published literature - studies, reports, opinions from Text Books, Journals, Manuals, etc.

Prioritized Treatment Options + Evidence Profile

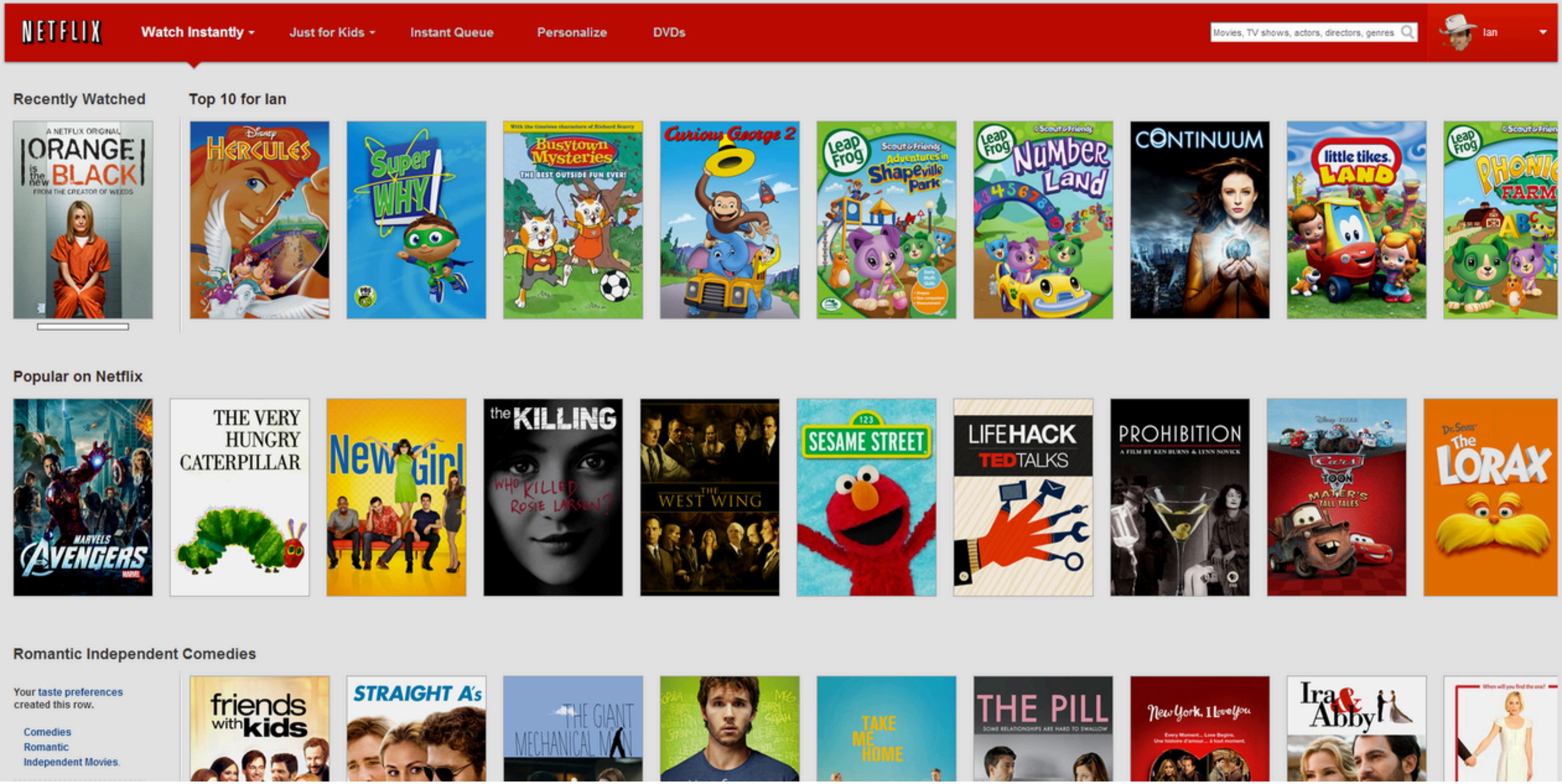
© 2014 International Business Machines Corporation

15

Example: optimization algorithms to suggest the best solution

https://www.mskcc.org/videos/mskcc-and-ibm-collaborate-applying-watson-technology-help-oncologists

Example: Netflix Recommendation System



Recently Watched

Top 10 for Ian

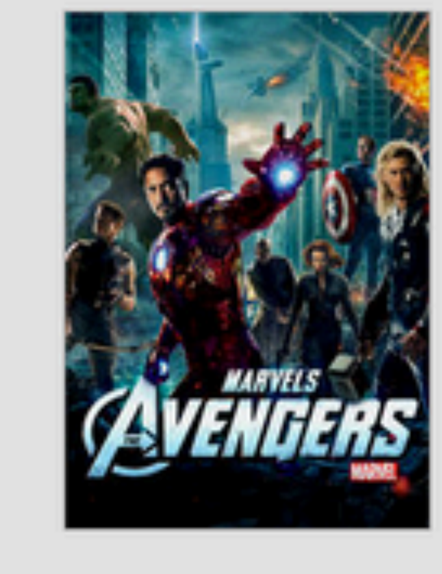
Popular on Netflix

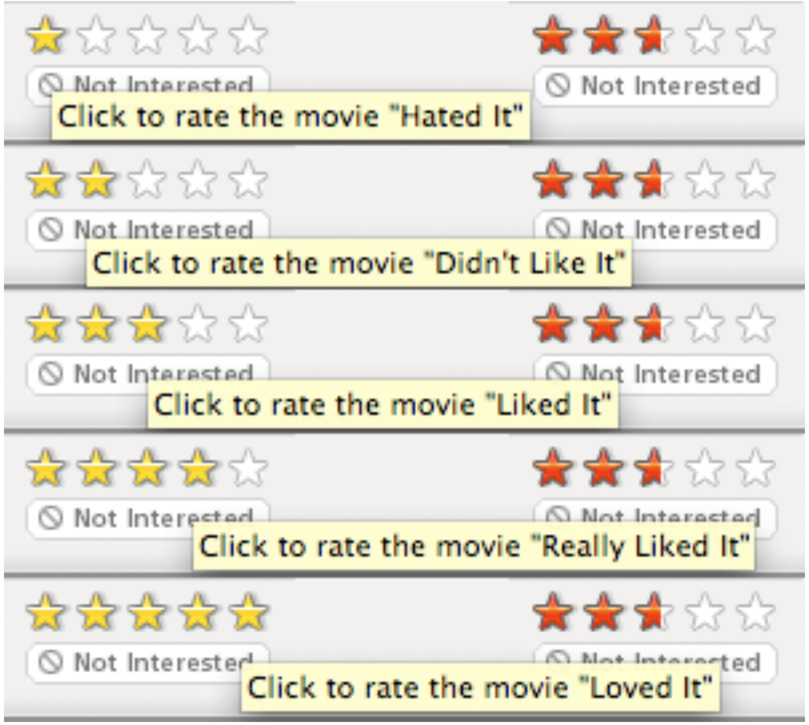
Romantic Independent Comedies

Your taste preferences created this row.

Comedies
Romantic
Independent Movies.

Example: Netflix Recommendation System




Not Interested
Click to rate the movie "Hated It"
Not Interested

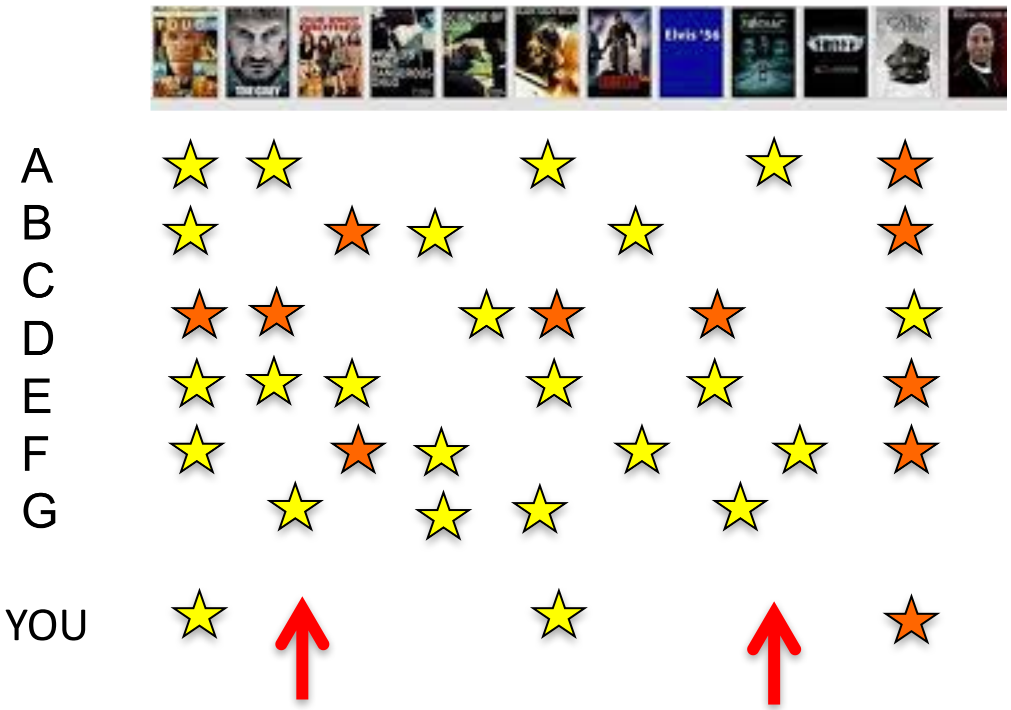
Not Interested
Click to rate the movie "Didn't Like It"
Not Interested

Not Interested
Click to rate the movie "Liked It"
Not Interested

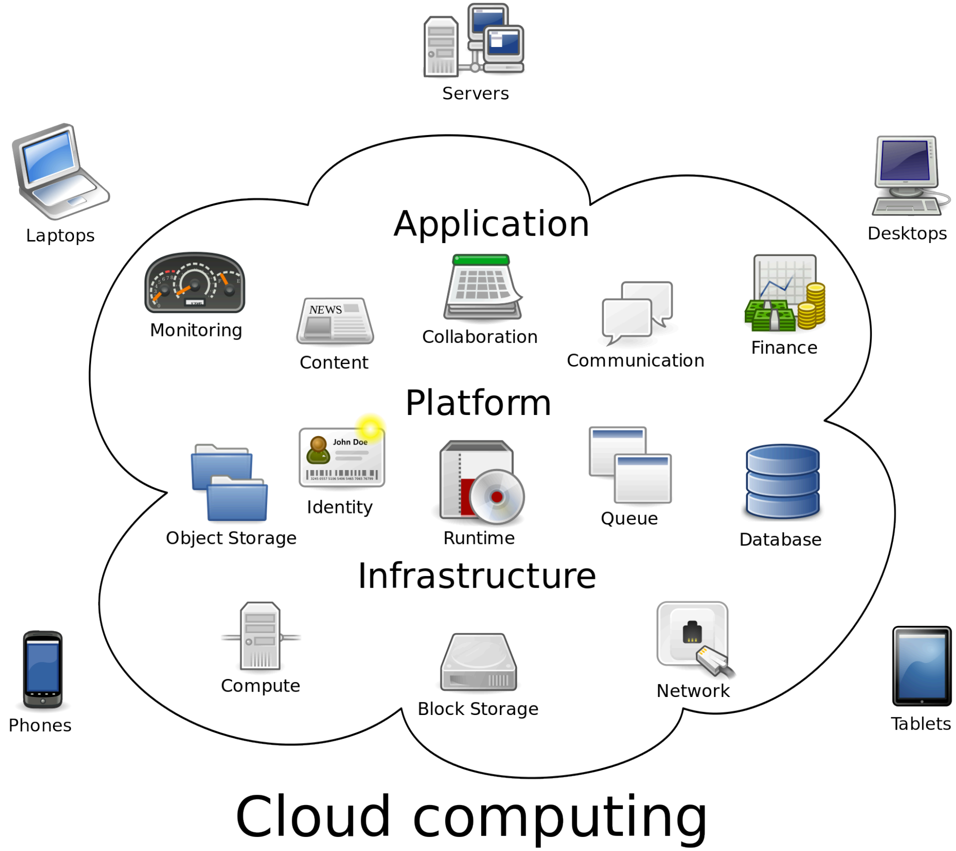
Not Interested
Click to rate the movie "Really Liked It"
Not Interested

Not Interested
Click to rate the movie "Loved It"
Not Interested

Example: Netflix Recommendation System



Cloud Computing

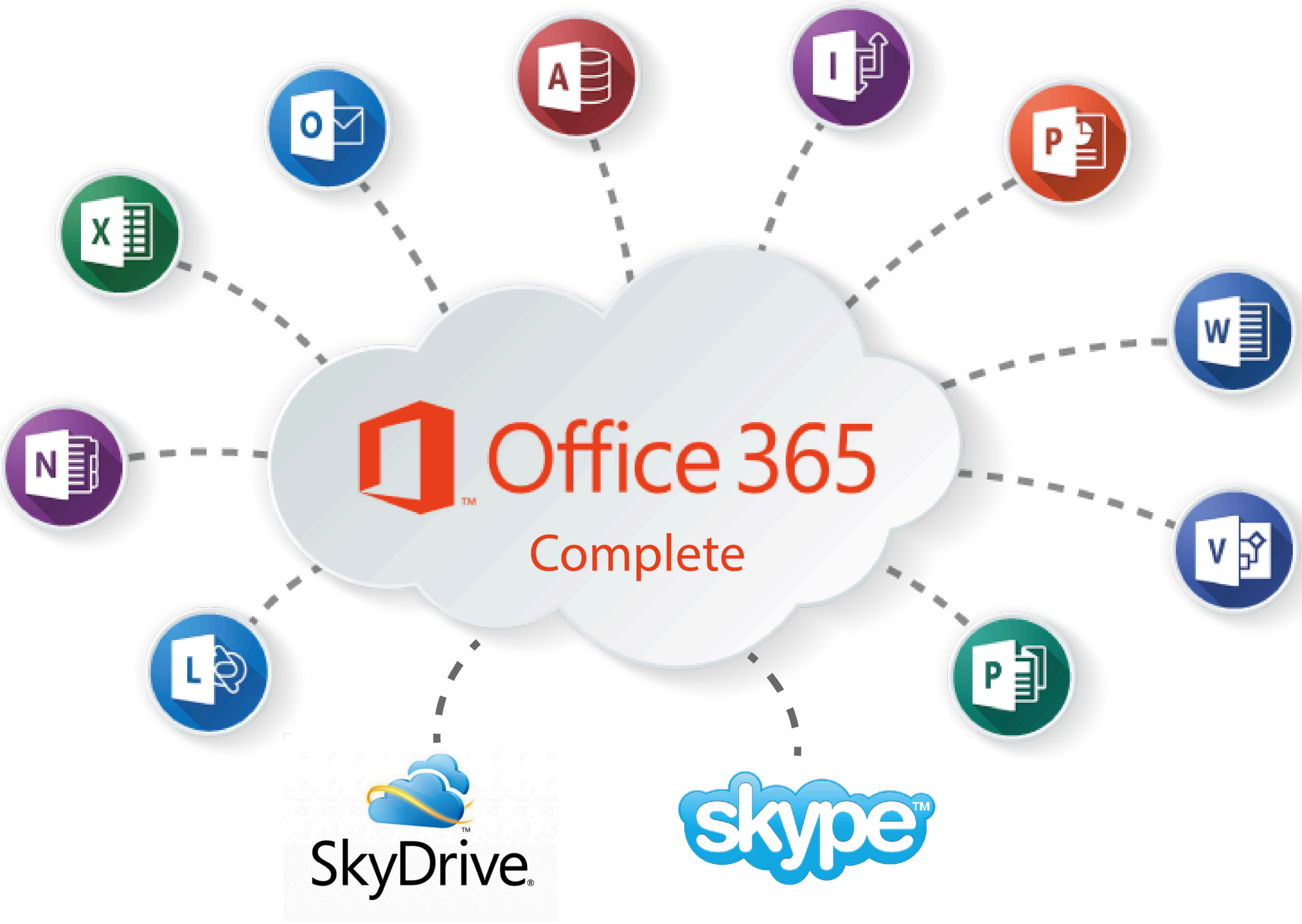


The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer. (From Wikipedia)

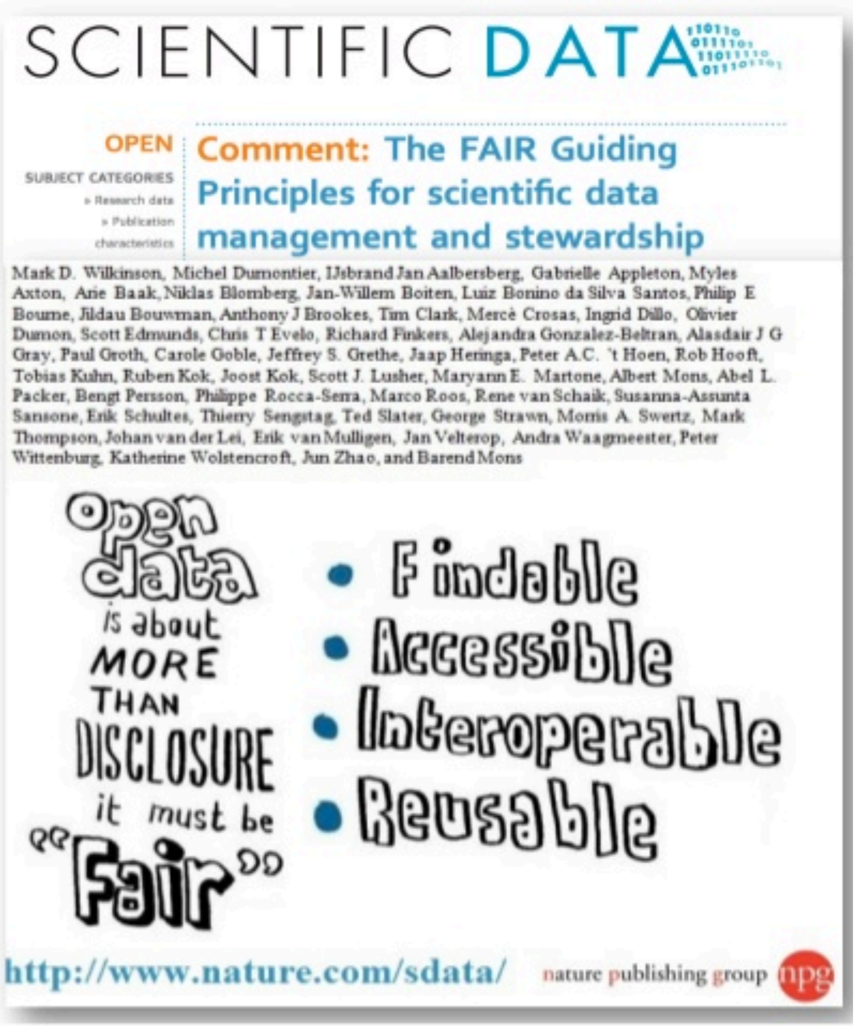
amazon web services™

Google Cloud Platform

Cloud Computing



Open Science



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons

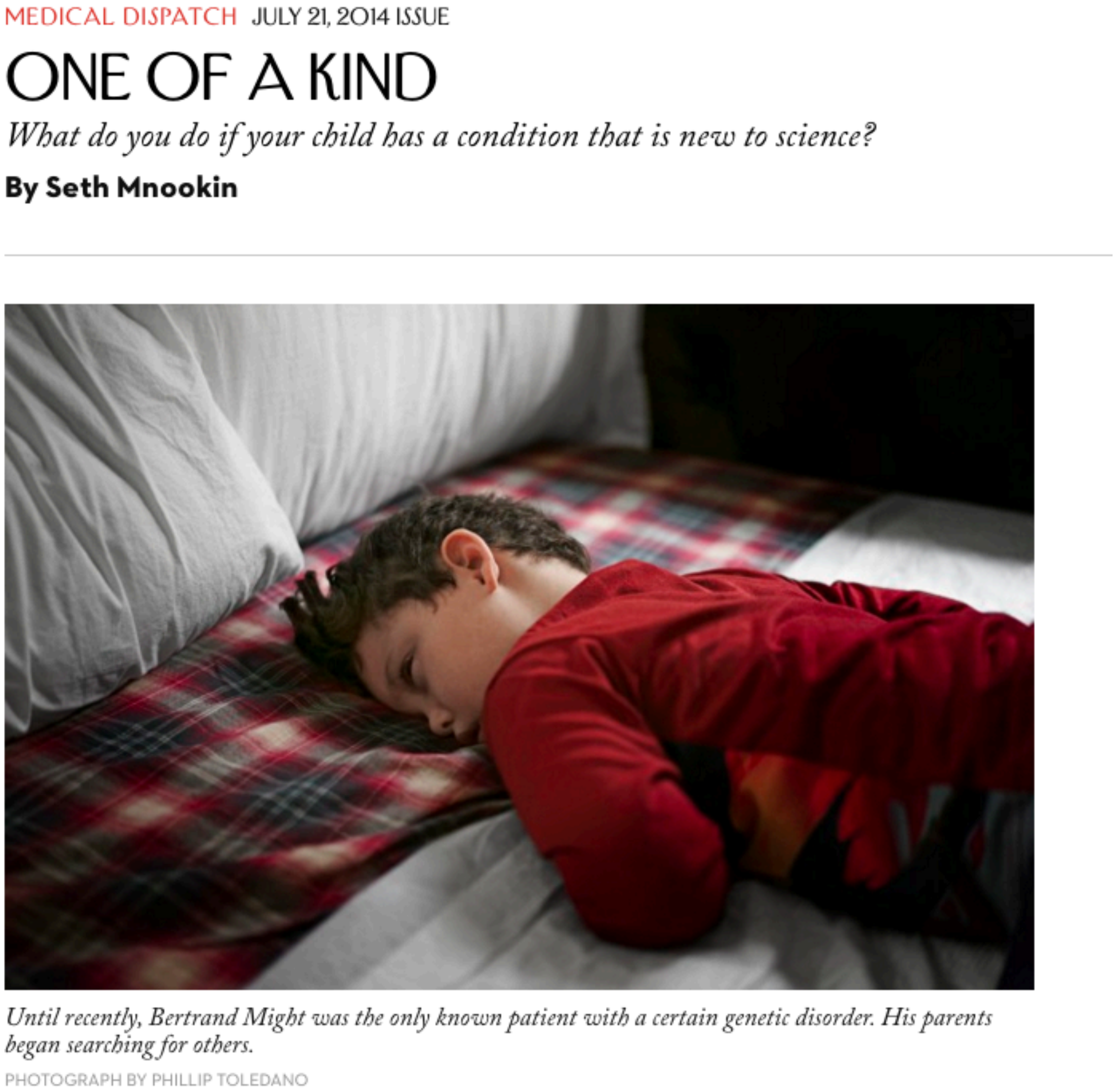
Open data is about MORE THAN DISCLOSURE it must be "Fair"

- Findable
- Accessible
- Interoperable
- Reusable

http://www.nature.com/sdata/

nature publishing group npg

A CASE STUDY



MEDICAL DISPATCH JULY 21, 2014 ISSUE

ONE OF A KIND

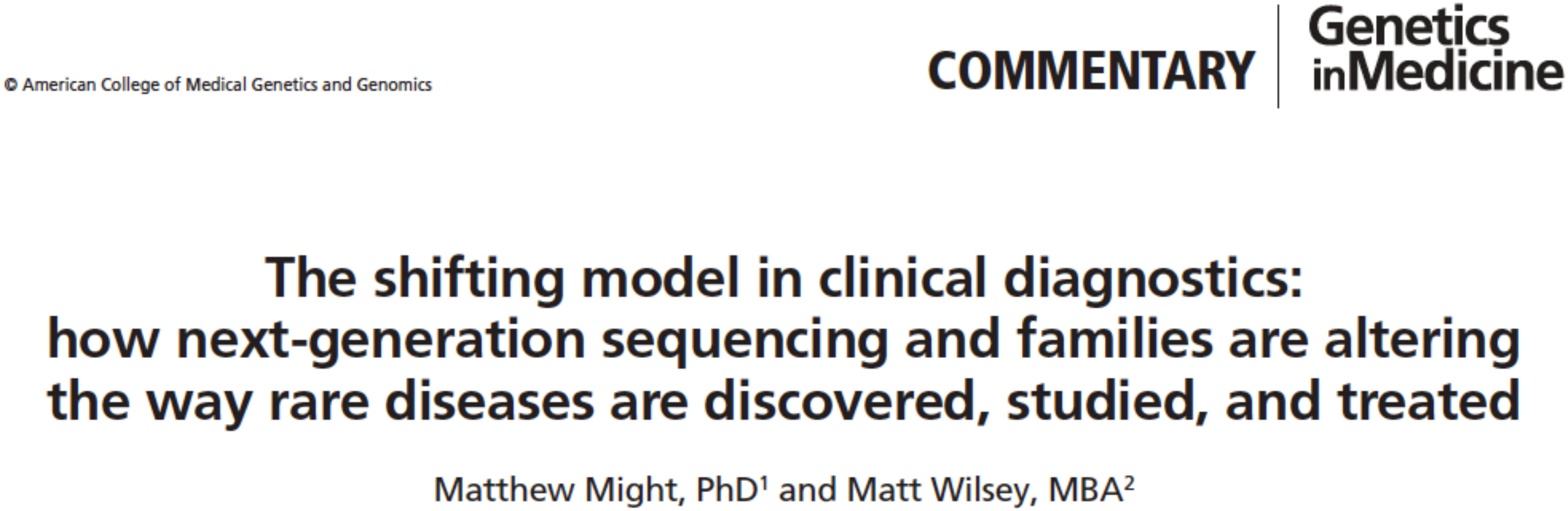
What do you do if your child has a condition that is new to science?

By Seth Mnookin

Until recently, Bertrand Might was the only known patient with a certain genetic disorder. His parents began searching for others.

PHOTOGRAPH BY PHILLIP TOLEDANO

A CASE STUDY



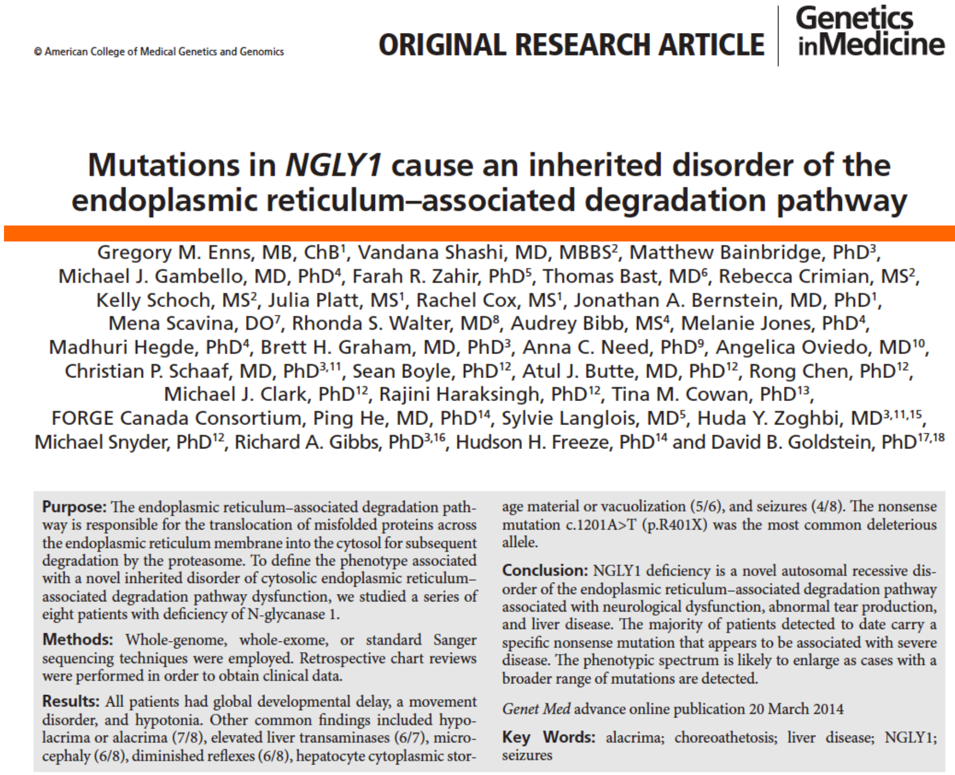
© American College of Medical Genetics and Genomics

COMMENTARY | Genetics inMedicine

The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated

Matthew Might, PhD[1] and Matt Wilsey, MBA[2]

NGLY1



© American College of Medical Genetics and Genomics

ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Mutations in *NGLY1* cause an inherited disorder of the endoplasmic reticulum–associated degradation pathway

Gregory M. Enns, MB, ChB[1], Vandana Shashi, MD, MBBS[2], Matthew Bainbridge, PhD[3], Michael J. Gambello, MD, PhD[4], Farah R. Zahir, PhD[5], Thomas Bast, MD[6], Rebecca Crimian, MS[2], Kelly Schoch, MS[2], Julia Platt, MS[1], Rachel Cox, MS[1], Jonathan A. Bernstein, MD, PhD[1], Mena Scavina, DO[7], Rhonda S. Walter, MD[8], Audrey Bibb, MS[4], Melanie Jones, PhD[4], Madhuri Hegde, PhD[4], Brett H. Graham, MD, PhD[3], Anna C. Need, PhD[9], Angelica Oviedo, MD[10], Christian P. Schaaf, MD, PhD[3,11], Sean Boyle, PhD[12], Atul J. Butte, MD, PhD[12], Rong Chen, PhD[12], Michael J. Clark, PhD[12], Rajini Haraksingh, PhD[12], Tina M. Cowan, PhD[13], FORGE Canada Consortium, Ping He, MD, PhD[14], Sylvie Langlois, MD[5], Huda Y. Zoghbi, MD[3,11,15], Michael Snyder, PhD[12], Richard A. Gibbs, PhD[3,16], Hudson H. Freeze, PhD[14] and David B. Goldstein, PhD[17,18]

Purpose: The endoplasmic reticulum–associated degradation pathway is responsible for the translocation of misfolded proteins across the endoplasmic reticulum membrane into the cytosol for subsequent degradation by the proteasome. To define the phenotype associated with a novel inherited disorder of cytosolic endoplasmic reticulum–associated degradation pathway dysfunction, we studied a series of eight patients with deficiency of N-glycanase 1.

Methods: Whole-genome, whole-exome, or standard Sanger sequencing techniques were employed. Retrospective chart reviews were performed in order to obtain clinical data.

Results: All patients had global developmental delay, a movement disorder, and hypotonia. Other common findings included hypolacrima or alacrima (7/8), elevated liver transaminases (6/7), microcephaly (6/8), diminished reflexes (6/8), hepatocyte cytoplasmic storage material or vacuolization (5/6), and seizures (4/8). The nonsense mutation c.1201A>T (p.R401X) was the most common deleterious allele.

Conclusion: NGLY1 deficiency is a novel autosomal recessive disorder of the endoplasmic reticulum–associated degradation pathway associated with neurological dysfunction, abnormal tear production, and liver disease. The majority of patients detected to date carry a specific nonsense mutation that appears to be associated with severe disease. The phenotypic spectrum is likely to enlarge as cases with a broader range of mutations are detected.

Genet Med advance online publication 20 March 2014

Key Words: alacrima; choreoathetosis; liver disease; NGLY1; seizures

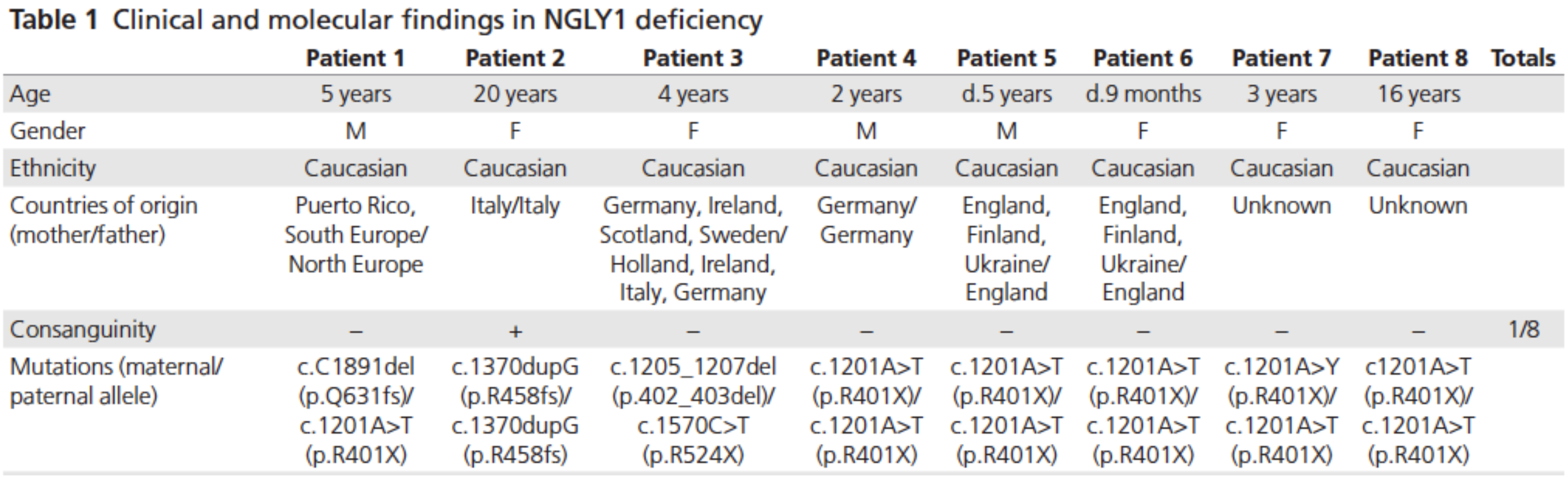


Table 1 Clinical and molecular findings in NGLY1 deficiency

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Totals
Age	5 years	20 years	4 years	2 years	d.5 years	d.9 months	3 years	16 years	
Gender	M	F	F	M	M	F	F	F	
Ethnicity	Caucasian	Caucasian	Caucasian	Caucasian	Caucasian	Caucasian	Caucasian	Caucasian	
Countries of origin (mother/father)	Puerto Rico, South Europe/ North Europe	Italy/Italy	Germany, Ireland, Scotland, Sweden/ Holland, Ireland, Italy, Germany	Germany/ Germany	England, Finland, Ukraine/ England	England, Finland, Ukraine/ England	Unknown	Unknown	
Consanguinity	–	+	–	–	–	–	–	–	1/8
Mutations (maternal/ paternal allele)	c.C1891del (p.Q631fs)/ c.1201A>T (p.R401X)	c.1370dupG (p.R458fs)/ c.1370dupG (p.R458fs)	c.1205_1207del (p.402_403del)/ c.1570C>T (p.R524X)	c.1201A>T (p.R401X)/ c.1201A>T (p.R401X)	c.1201A>T (p.R401X)/ c.1201A>T (p.R401X)	c.1201A>T (p.R401X)/ c.1201A>T (p.R401X)	c.1201A>Y (p.R401X)/ c.1201A>T (p.R401X)	c1201A>T (p.R401X)/ c.1201A>T (p.R401X)	

Conclusion

- Biomedical research is in the center of digital revolution.
- Every biomedical problem is a data problem. In this
- Harnessing the power of big data in understanding disease mechanisms (basic) and enabling precision medicine (clinical).

So, do you want to learn data science?

HBR.ORG

Harvard Business Review

OCTOBER 2012

46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin

84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz

93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks

GETTING CONTROL OF BIG DATA

How vast new streams of information are changing the art of management

PAGE 59

Data Scientist:

The Sexiest Job of the 21st Century