



What is Big Data?

And Why it Matters to You!

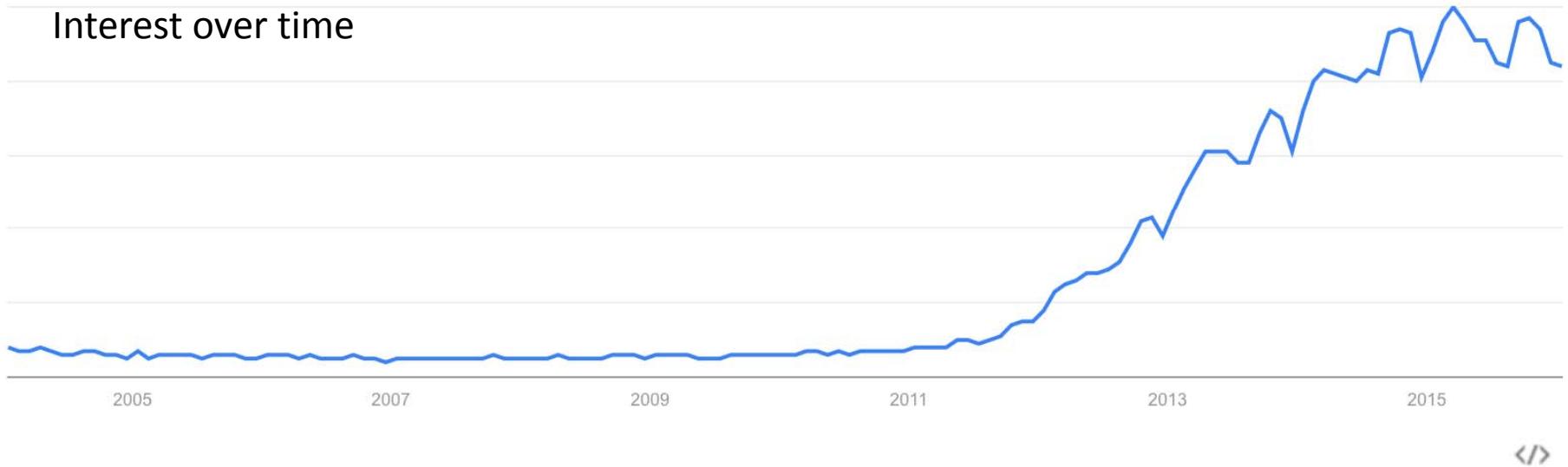


Data After Dark
OHSU BD2K Data Science Workshop

Shannon McWeeney, PhD
13th January 2016

Google Trends: “Big Data”

Interest over time



</>

Chart created 13th January 2016

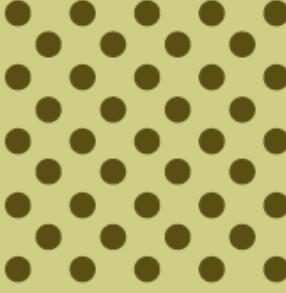
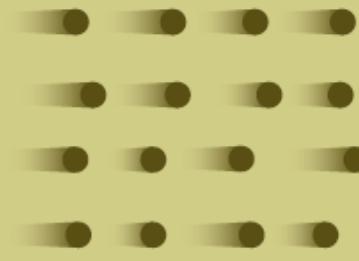
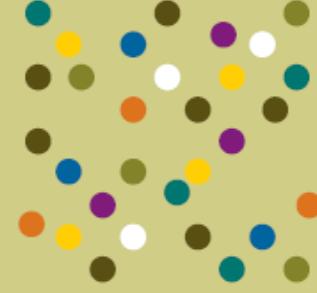
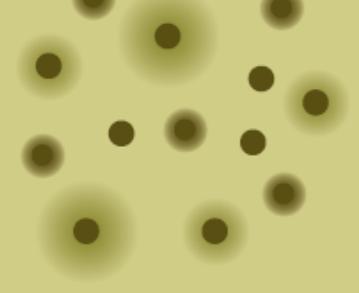


2

What is Big Data?

Interactive Q&A

The 4 V's

Volume	Velocity	Variety	Veracity
			
Data at rest <p>Terabytes to exabytes of existing data to process</p>	Data in motion <p>Streaming data, milliseconds to seconds to respond</p>	Data in many forms <p>Structured, unstructured, text and multimedia</p>	Data in doubt <p>Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations</p>



What does this mean?

Why does this matter to me?

The Fourth Paradigm*



First

DESCRIPTIVE



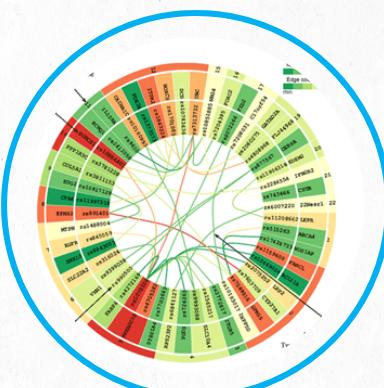
Second

THEORETICAL



Third

COMPUTATIONAL



Fourth

EXPLORATION

*Jim Gray, Microsoft



Career Prospects

Harvard
Business
Review



**SEXIEST
CAREER
ALIVE!**

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE



Career Prospects

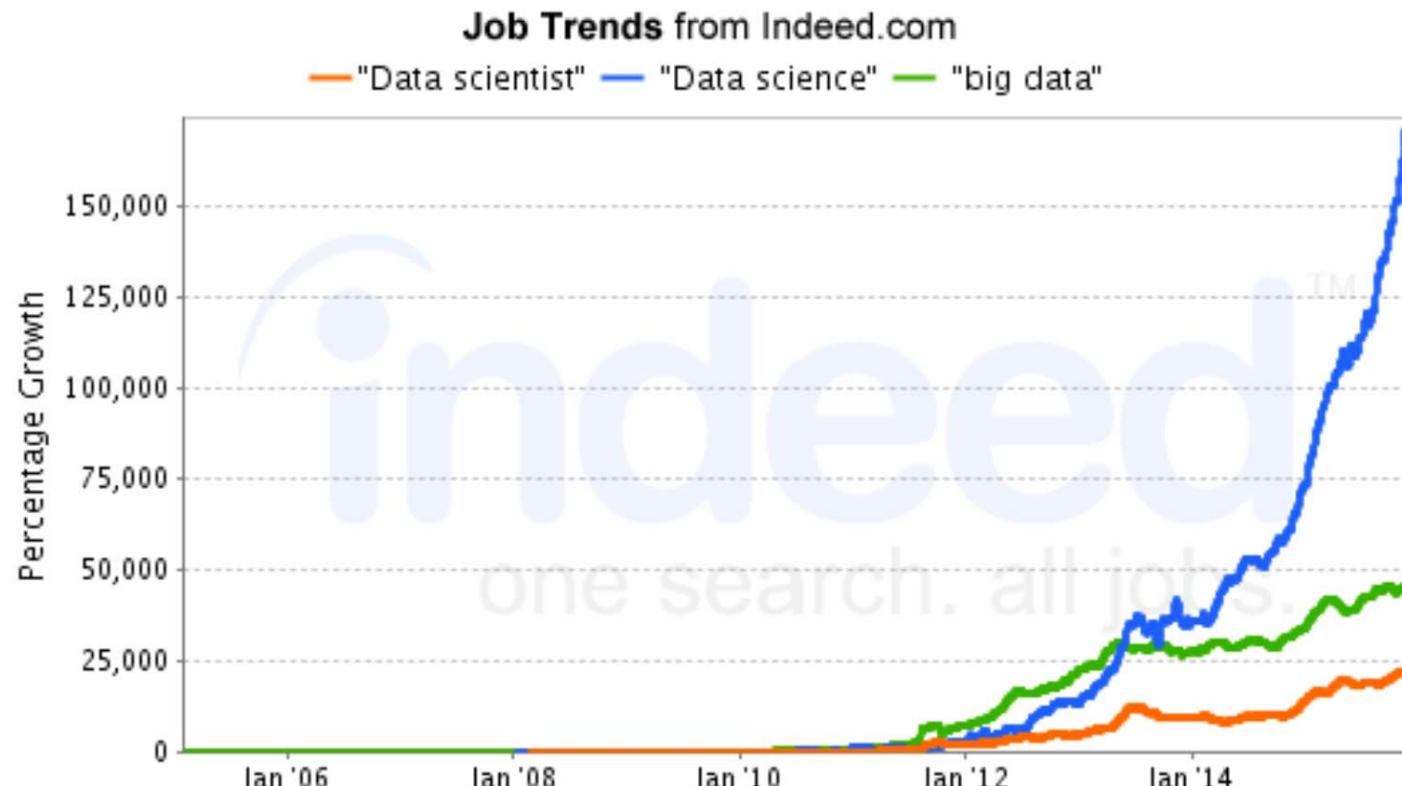


Chart created 13th January 2016



Scientific Considerations

Retraction Watch

Tracking retractions as a window into the scientific process

Court denies appeal of HIV fraudster's 57-month prison sentence

without comments

An appeals court has affirmed the stiff prison sentence for [Dong-Pyou Han](#), the former Iowa State University researcher who faked the results of an HIV vaccine experiment in rabbits. [Read the rest of this entry »](#)



Share this:

[Email](#) [Facebook](#) [Twitter](#)

Written by Ivan Oransky
January 13th, 2016 at 9:30 am

Posted in [dc](#)

Data dispute forces journal to valuable land

without comments

The authors of a paper about the density of an e disputing the journal's decision to pull the paper its contents.

Subscribe to Blog via Email

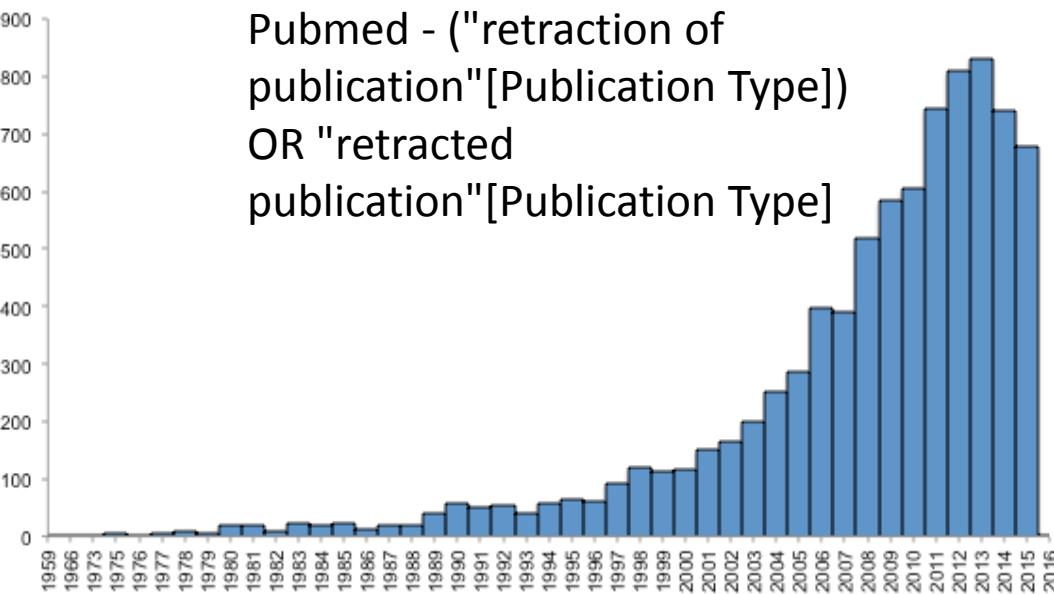
Join 11,609 other subscribers

Email Address

Subscribe

Pages

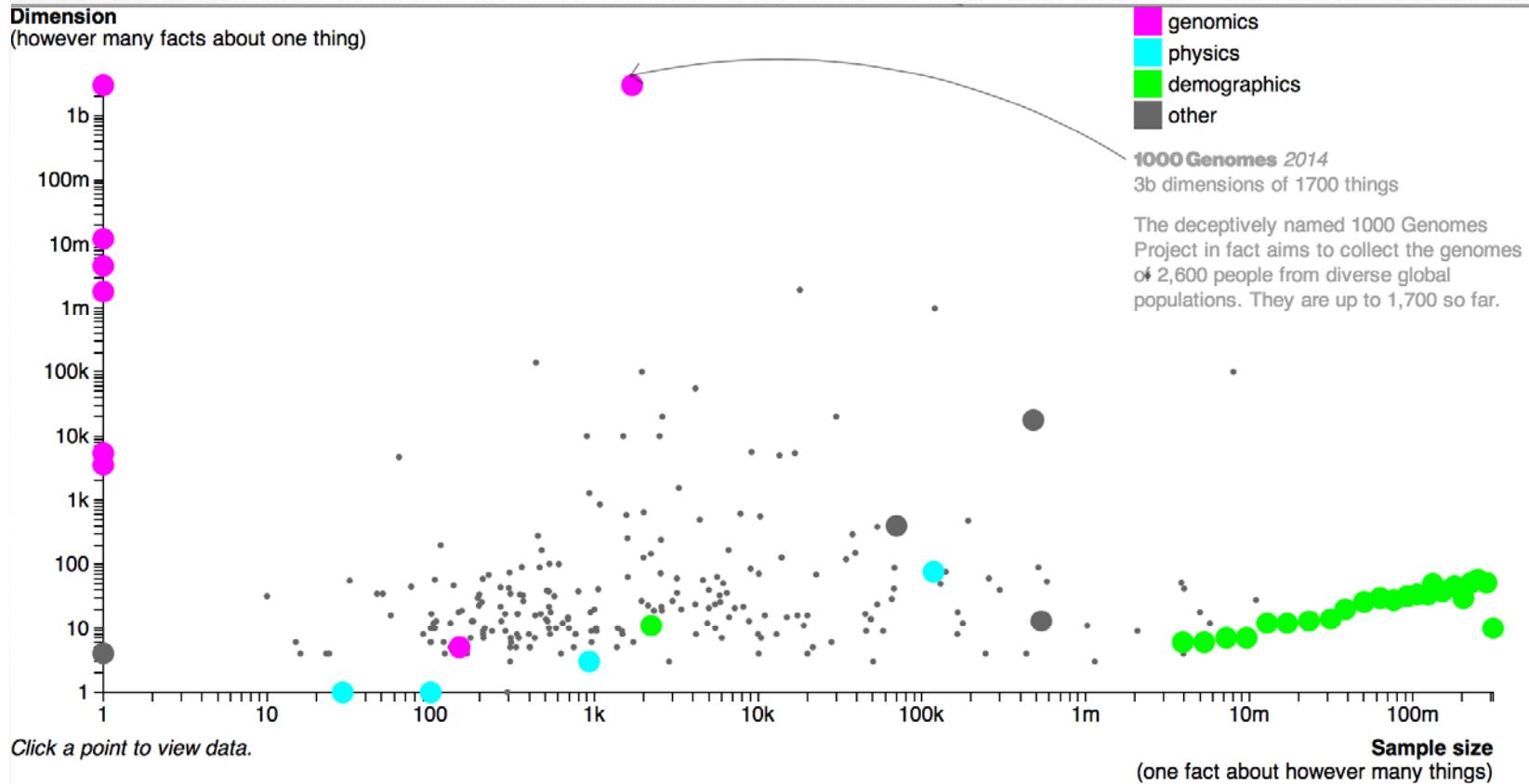
Pubmed - ("retraction of publication"[Publication Type])
OR "retracted publication"[Publication Type]



Quantifying Data

Quantified Self and Your Personal Data

Quantifying Data

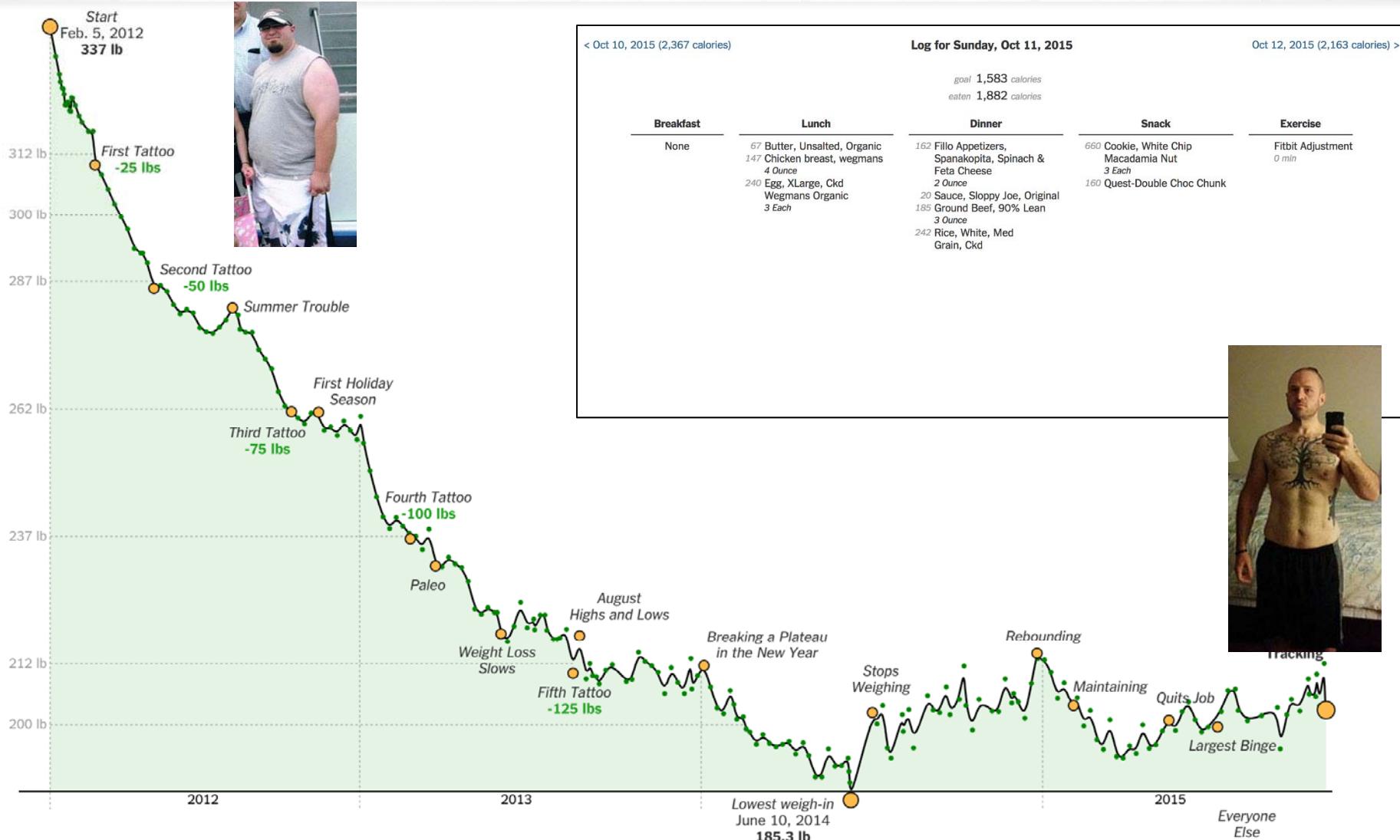


GRAPHIC BY BLOOMBERG BUSINESSWEEK. DATA: UNIVERSITY OF CALIFORNIA IRVINE; AMAZON; KAGGLE; U.S. CENSUS BUREAU; STEPHEN STIGLER; ESA; PLOS; WIKIPEDIA ([CSV](#))



How much data do you generate?

Quantifying Personal Self



Data: Steve Lochner; Graph: New York Times



Quantifying YOUR Data

Exogenous data

(Behavior, Socio-economic, Environmental, ...)

60% of determinants of health

Volume, Variety, Velocity, Veracity

Genomics data

30% of determinants of health

Volume

Clinical data

10% of determinants of health

Variety



1100 Terabytes

Generated per lifetime

6 TB

Per lifetime

0.4 TB

Per lifetime

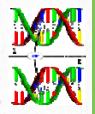
Source: "The Relative Contribution of Multiple Determinants to Health Outcomes", Lauren McGover et al., *Health Affairs*, 33, no.2 (2014)

Image: IBM



Personal Information Streams

'Omics'

	Genome: SNP mutations Structural variation Epigenetics	✓
	Microbiome	✓
	Transcriptome	
	Metabolome	
	Proteome	
	Diseasome	✓
	Environmentome	✓

Traditional

Personal and Family Health History	✓
Prescription History	✓
Lab Tests: History and Current	✓
Demographic Data	✓
Standardized Questionnaires	✓

Quantified Self

	Self-reported data: health, exercise, food, mood journals, etc.	✓
	Mobile App Data	✓
	Quantified Self Device Data	✓
	Biosensor Data Objective Metrics	

Internet of-Things

	Smart Home	✓
	Smart Car	✓
	Personal Robotics	✓
	Environmental Sensors	✓
	Community Data	✓

Legend: Consumer-available

Data Ecosystem

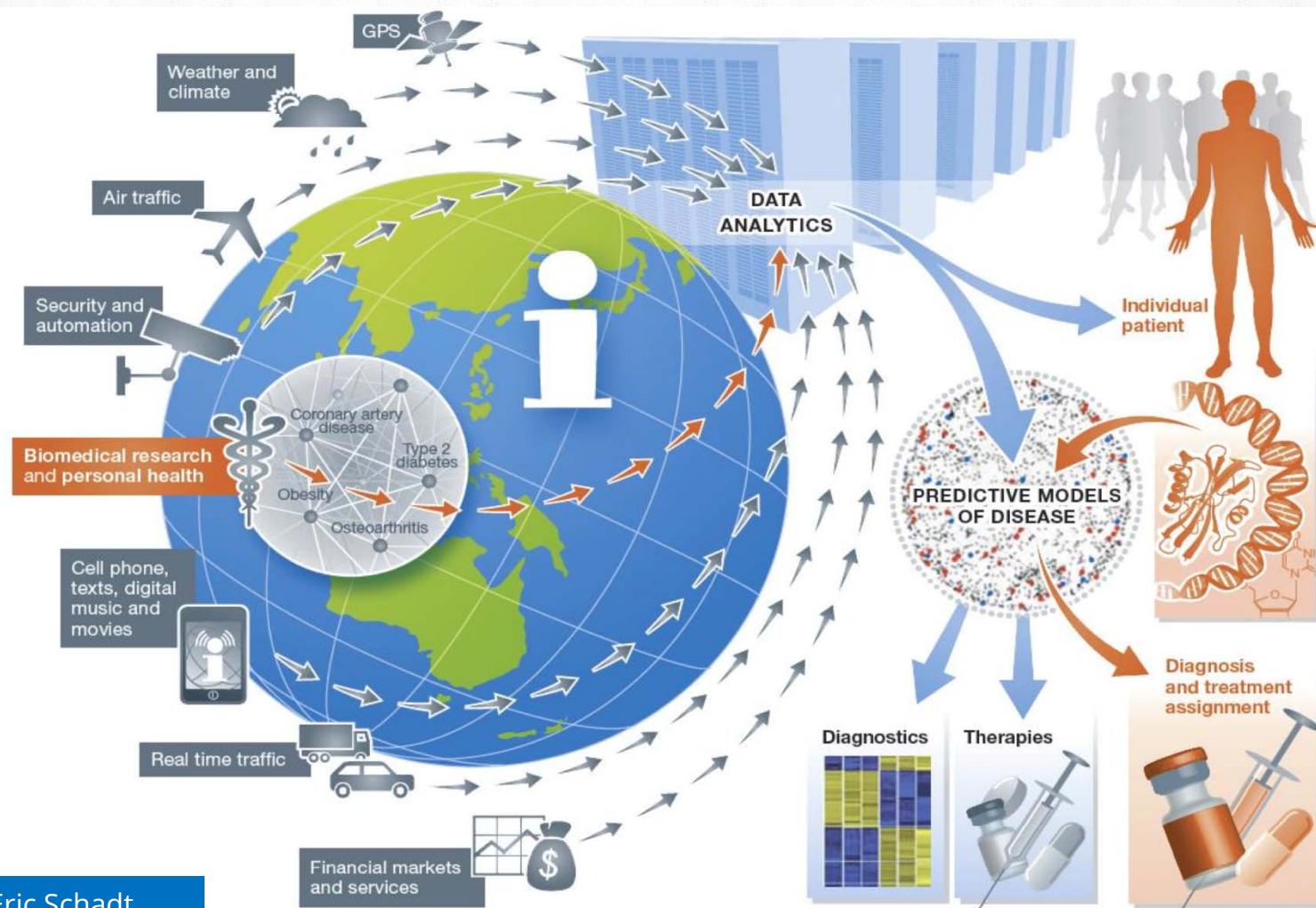


Image: Eric Schadt

From Individual to Population

Key Use Cases

"We are each, in effect, one-person clinical trials"

Op-Ed As I lay dying

By LAURIE BECKLUND

SHARELINES

Do patients fail therapies, or do the therapies fail them? #breastcancer

Breast cancer is not one disease; there is no one 'cure'

FEBRUARY 20, 2015, 7:45 PM

I am dying, literally, at my home in Hollywood, of metastatic breast cancer, the only kind of breast cancer that kills. For six years I've known I was going to die. I just didn't know when.

Then, a couple of weeks before Christmas, a new, deadly diagnosis gave me a deadline. No doctor would promise me I'd make it to 2015.



"Yet the knowledge generated from those trials will die with us because there is no comprehensive database of metastatic breast cancer patients.... **In the Big Data-era, this void is criminal.**"



**“The unfolding calamity
in genomics is that a
great deal of life-saving
information, though
already collected, is
inaccessible.”**

Antonio Regalado

Internet of DNA



Internet of DNA

A global network of millions of genomes could be medicine's next great advance.

Availability: 1-2 years

Breakthrough

Technical standards that let DNA databases communicate.

Why It Matters

Your medical treatment could benefit from the experiences of millions of others.

Key Players

+ Global Alliance for Genomics and Health
+ Google
+ Personal Genome Project

10 Breakthrough Technologies 2015

Introduction

[Magic Leap](#) >

[Nano-Architecture](#) >

[Car-to-Car Communication](#) >

[Project Loon](#) >

[Liquid Biopsy](#) >

[Megascale Desalination](#) >

[Apple Pay](#) >

[Brain Organoids](#) >

[Supercharged Photosynthesis](#) >

[Internet of DNA](#) >

Source: Harvard Business Review



GA4GH

<http://genomicsandhealth.org>

The screenshot shows the homepage of the Global Alliance for Genomics & Health. At the top left is the logo, which consists of a circular emblem made of colored lines forming a grid-like pattern, with the text "Global Alliance" and "for Genomics & Health" to its right. To the right of the logo is a green button labeled "Become a Member". Below the logo is a search bar with a magnifying glass icon. A blue navigation bar at the top has five items: "ABOUT GLOBAL ALLIANCE", "OUR WORK", "MEMBERS", "NEWS & EVENTS", and "CONTACT US". The main content area features a large image of a stethoscope on a clipboard. Overlaid on this image is a dark blue box containing the text "Framework for Responsible Sharing of Genomic and Health-Related Data" and a link "Read Framework here". Below this box are four small white circles. At the bottom of the page are three columns with titles: "What is the Global Alliance?", "What is the Global Alliance doing?", and "Who is involved?". Each column contains a brief description of the Global Alliance's mission and activities.

Global Alliance
for Genomics & Health

Become a Member

ABOUT GLOBAL ALLIANCE OUR WORK MEMBERS NEWS & EVENTS CONTACT US

Framework for Responsible Sharing of Genomic and Health-Related Data

Read the new Framework guided by human rights that offers foundational principles and core elements to facilitate responsible research conduct.

→ [Read Framework here](#)

○ ○ ● ○

What is the Global Alliance?

The Global Alliance for Genomics and Health (Global Alliance) is an international coalition, dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing. The promise of genomic data to revolutionize biology and medicine depends critically on our ability to make comparisons

What is the Global Alliance doing?

Since its formation in 2013, the Global Alliance for Genomics and Health is leading the way to enable genomic and clinical data sharing. The Alliance's Working Groups are producing high-impact deliverables to ensure such responsible sharing is possible, such as developing a [Framework for Data Sharing](#) to guide governance and research and a

Who is involved?

The Global Alliance for Genomics and Health is an independent, non-governmental alliance, made up of hundreds of world-leading organizations and individuals from across the world. The Global Alliance is focused on bringing together a diverse set of key stakeholders across regions and sectors, including leaders in healthcare and research

21

Matchmaker Exchange

<http://www.matchmakerexchange.org>



Matchmaker Exchange



Genomic discovery through the exchange of phenotypic & genotypic profiles



22

Consumer Genetic Testing

The image displays two side-by-side screenshots of consumer genetic testing websites.

Left Website (23andMe):

- Header:** 23andMe logo, Welcome message.
- Section:** "Here's a preview of what you'll receive when you receive your report".
- Content:** "Hereditary Breast and Ovarian Cancer Syndrome (BRCA1- and BRCA2-Related, Selected Mutations)".
- Text:** "Established Research report on 3 reported markers."
- Buttons:** Example Data, Resources, Technical Report.
- Warning:** A blue box with an exclamation mark stating: "If this condition runs in your family or you think you might have this condition, consult with a healthcare provider about appropriate testing." Below it says: "Other factors can also influence your risk for this condition even if you don't have the variant(s) covered by this report."
- Footer:** About Hereditary Breast and Ovarian Cancer Syndrome.

Right Website (Color):

- Header:** color logo, navigation links: BUY, LEARN, SUPPORT, PROVIDERS, ACTIVATE KIT, SIGN IN.
- Image:** A photograph of a Color DNA kit box on a kitchen counter.
- Text:** "Understand your genetic risk for breast and ovarian cancer".
- Text:** "Color analyzes 19 genes—including BRCA1 and BRCA2—to help you understand your risk of developing breast and ovarian cancer. Purchase your Color Kit for \$249."
- Button:** "Purchase Color".

ClinGen

<http://clinicalgenome.org>

The screenshot shows the ClinGen website homepage. At the top left is the ClinGen logo with a stylized DNA helix icon. The top right features a navigation bar with links for "For Patients", "Search", and "Contact". Below the navigation is a horizontal menu with links for "About", "Data Sharing", "Knowledge Curation", "Machine Learning", "Events & News", and "Tools & Resources". The main content area has a blue banner at the bottom with the text "ClinGen: Sharing Data. Building Knowledge. Improving Care." and a paragraph about the need for coordinated effort between clinical and research communities. A "Learn more »" link is also present. The background of the page shows a blurred image of a laboratory or office environment with a computer monitor and test tubes.

For Patients Search Contact

About Data Sharing Knowledge Curation Machine Learning Events & News Tools & Resources

ClinGen
Clinical Genome Resource

ClinGen: Sharing Data. Building Knowledge. Improving Care.

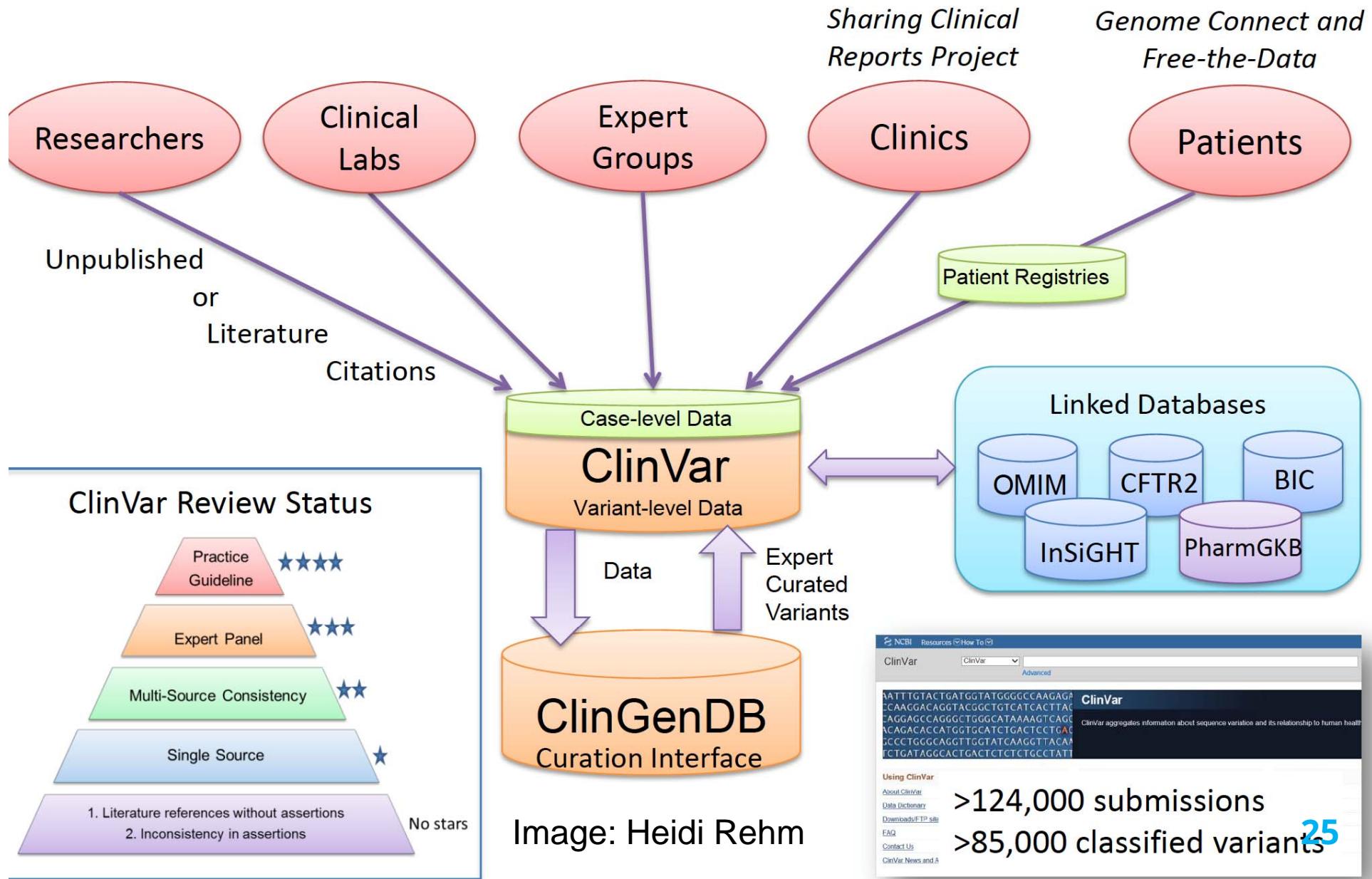
Technological advances are quickly allowing genome-wide analysis to become commonplace in the care of patients. However, the ability to detect DNA variants has greatly surpassed the ability to interpret their clinical impact, limiting patient benefit. Improving genomic interpretation will require a coordinated effort from both the clinical and research communities. [Learn more »](#)



24

Data Flows in ClinGen

(>200 ClinVar submitters)



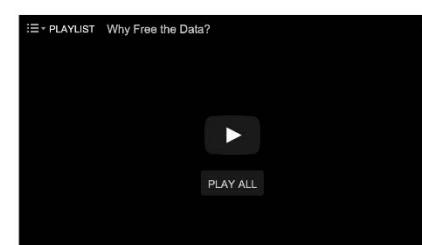
Patient Provided Data



The screenshot shows the homepage of patientslikeme.com. At the top, there's a search bar with placeholder text "conditions, symptoms, treatments...". To the right of the search bar is a link "Already a member? Sign in.". Below the search bar is a large banner with the text "Live better, together!" and "Making healthcare better for everyone through sharing, support, and research". A green button labeled "Join now" with the subtext "(it's free!)" is prominently displayed. To the right of the banner is a logo for "FREE THE DATA" featuring a grid of colored dots. Below the banner, there are three sections: "Learn from others" (Compare treatments, symptoms and experiences with people like you and take control of your health), "Connect with people like you" (Share your experience, give and get support to improve your life and the lives of others), and "Track your health" (Chart your health to research the all). At the bottom of the page are links for "Home", "About CONNECT", "Free My Data", "Join the Movement", "Learn More", "For Clinicians", and "About Us". Social media icons for Twitter, Facebook, and LinkedIn are located at the top right.



The screenshot shows the homepage of PatientCrossroads™. At the top left is the "PatientCrossroads™" logo. In the center is a large image of a family (a man, a woman, and a child) sitting in a field, looking at a tablet together. To the left of this image is the "GenomeConnect The ClinGen Patient Portal" logo, which features a stylized DNA double helix icon. Below the logo is a login form with fields for "User Name" and "Password", a "Remember me" checkbox, and a "Login" button. Below the login form are links for "Forgot username / password?" and "Create account". At the bottom left is a "About GenomeConnect" link. In the center of the page is a call-to-action button "Click here to register now!". At the bottom of the page is a section titled "GenomeConnect – Partnering with YOU to advance genomic health" with a descriptive paragraph about the portal's purpose and how it advances genomic health.

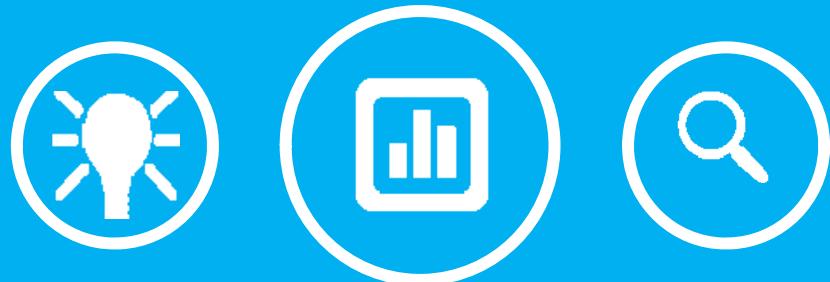


Genetic information is more valuable when shared.
Genes contain important information about your health and disease. Changes, called mutations, in BRCA1 and BRCA2 greatly increase the risk of hereditary breast and ovarian cancer. Sharing these mutations helps clinicians improve patient care and helps researchers advance our understanding of hereditary breast and ovarian cancer. Mutations should not be 'trade secrets' - join us and Free the Data!



This section contains four call-to-action boxes: "Free My Data", "Join the Movement", and "Learn More".
Free My Data: Your mutation and health information are important in the search for better health. Share your information safely and securely.
Join the Movement: Your story, photos, and/or videos are important for others to experience. You can help encourage more people to participate in Free The Data!
Learn More: Learn why sharing your mutation is important and how the privacy and sharing system works.





Clinical data – not as big, but

It may matter even more to you!

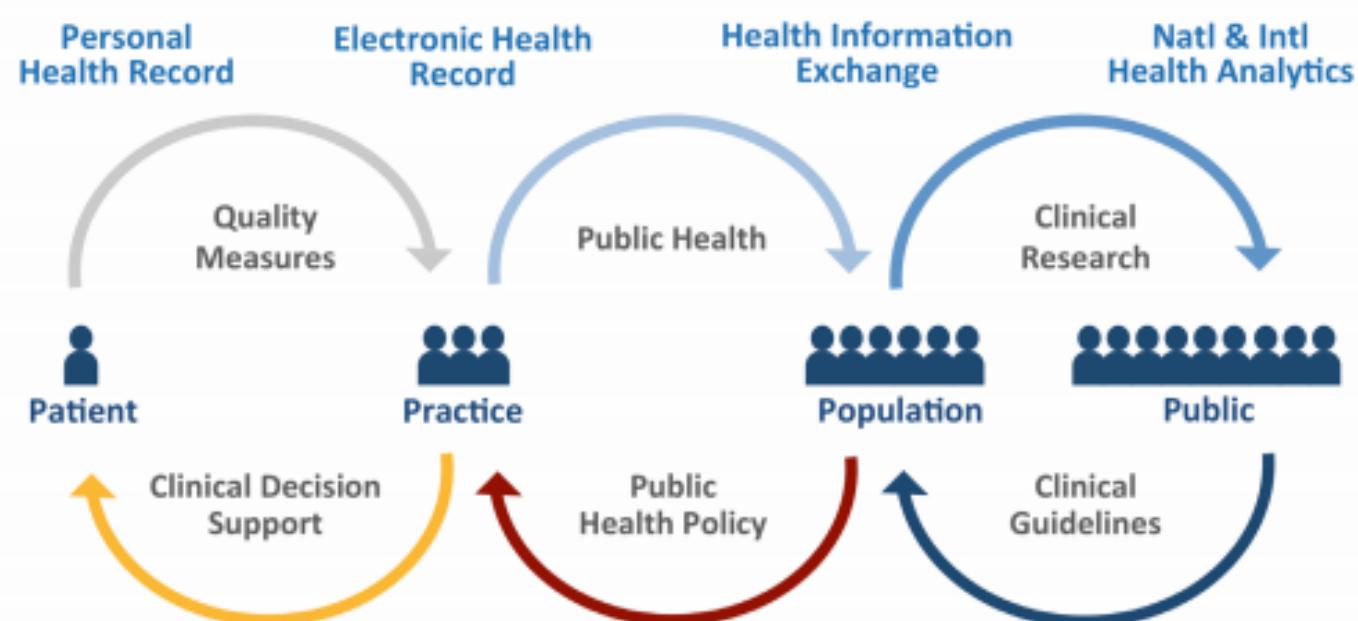


Data After Dark
OHSU BD2K Data Science Workshop

David Dorr, MD, MS
13th January 2016

Clinical Data across the health ecosystem

Figure 1. Vision of the Health IT Ecosystem



Source: Office of the National Coordinator



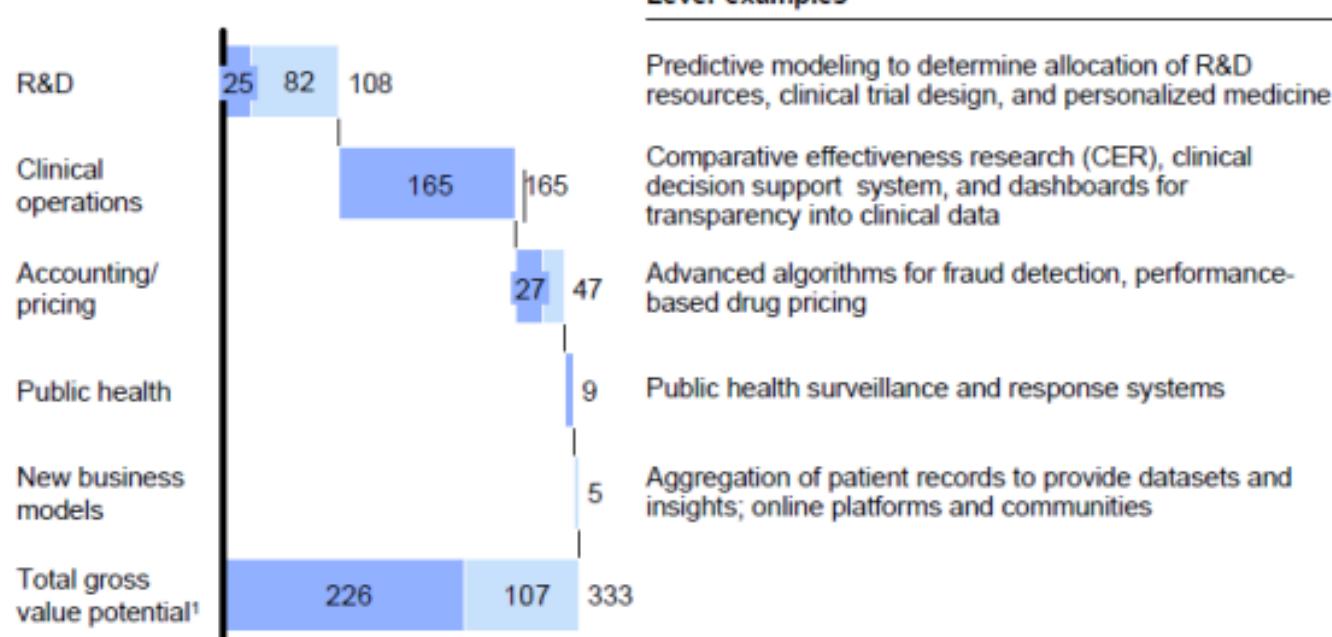
Value of clinical data

The estimated long-term value of identified levers is more than \$300 billion, with potentially more than \$200 billion savings on national health care spending

Value potential from use of big data

\$ billion per year

- Direct reduction on national health care expenditure
- Unclear impact on national health care expenditure

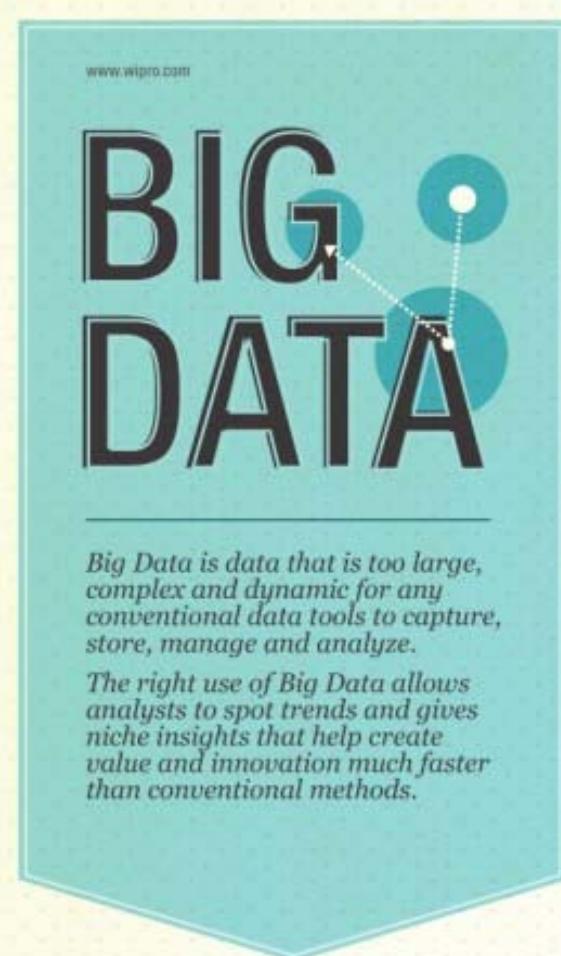


1 Excluding initial IT investments (~\$120 billion–\$200 billion) and annual operating costs (~\$20 billion per annum).

SOURCE: Expert interviews; press and literature search; McKinsey Global Institute analysis



Big Data



CASE STUDY - Healthcare

\$300 billion is the potential annual value to Healthcare



TRANSPARENCY IN CLINICAL DATA AND CLINICAL DECISION SUPPORT

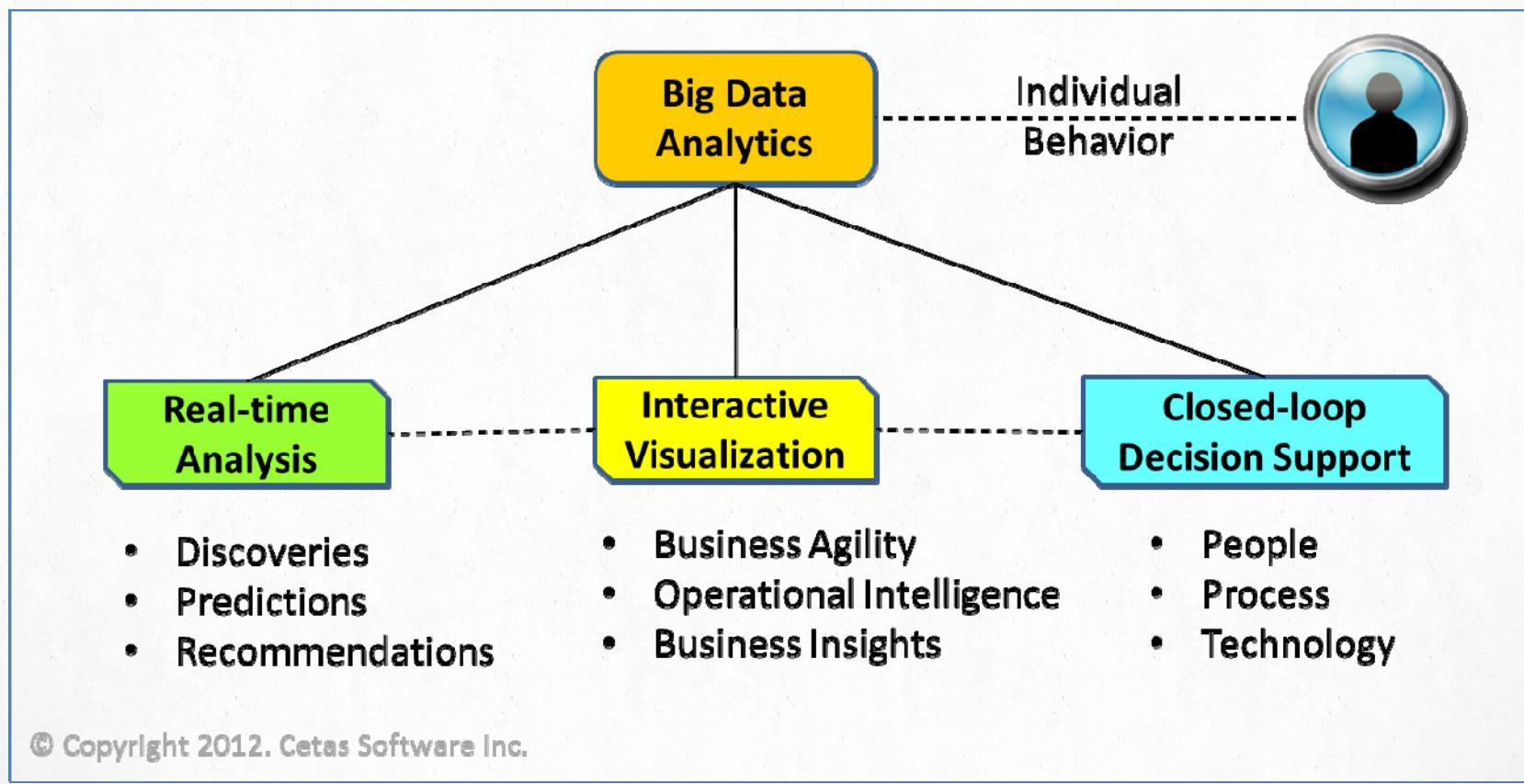
AGGREGATION OF PATIENT RECORDS, ONLINE PLATFORMS AND COMMUNITIES

RESEARCH AND DEVELOPMENT; PERSONALIZED MEDICINE; CLINICAL TRIAL DESIGN

PUBLIC HEALTH SURVEILLANCE AND RESPONSE SYSTEMS

ADVANCED FRAUD DETECTION; PERFORMANCE BASED DRUG PRICING

The future? (or the present, just not evenly distributed*)

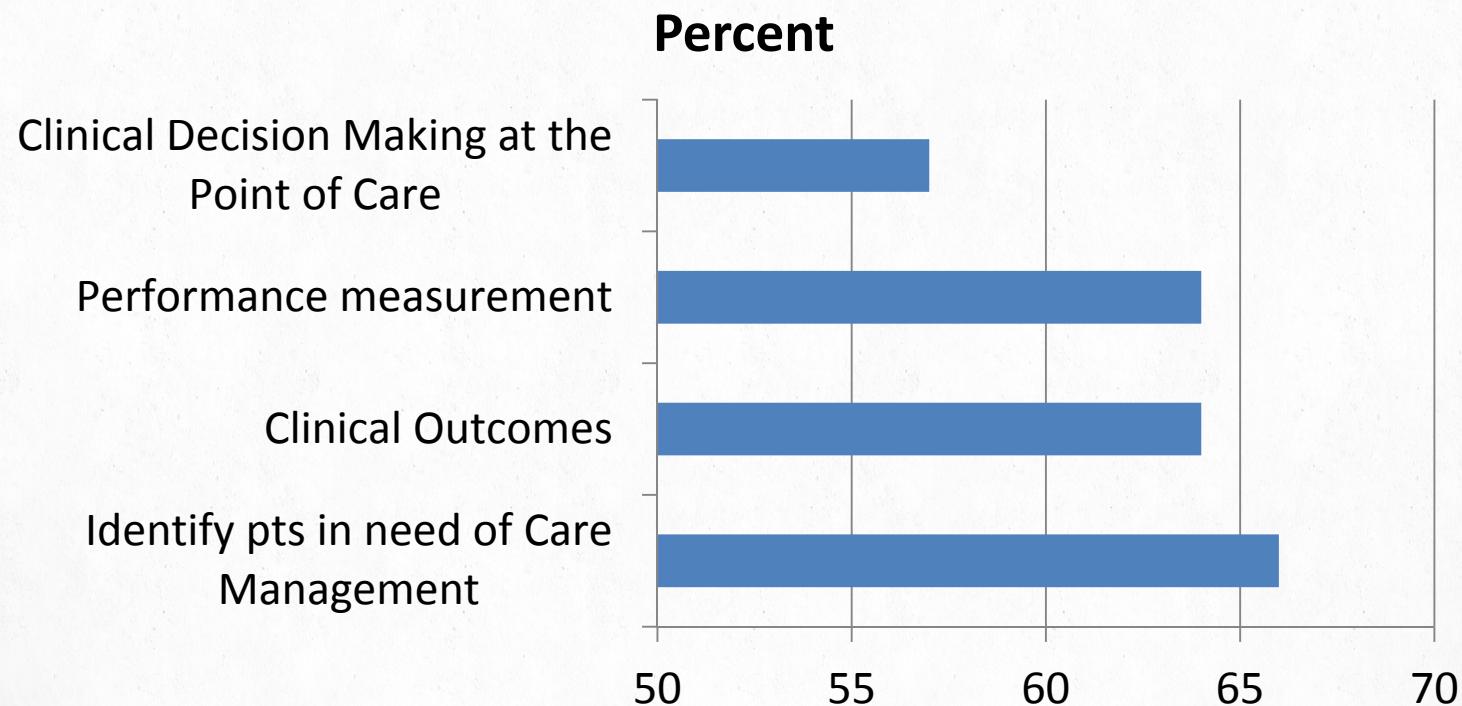


© Copyright 2012. Cetas Software Inc.

<http://www.cetas.net/blog/wp-content/uploads/2012/02/Big-Data-Analytics-Frame.png>; * quote is from William Gibson, author.

What do healthcare institutions plan to use 'analytics' for?

According to IDC, the top four reported capabilities for which healthcare organizations intend to use analytics are:



What do visionaries and innovators want to use analytics for?

HealthData
Management

Elizabeth Gardner
MAR 1, 2013

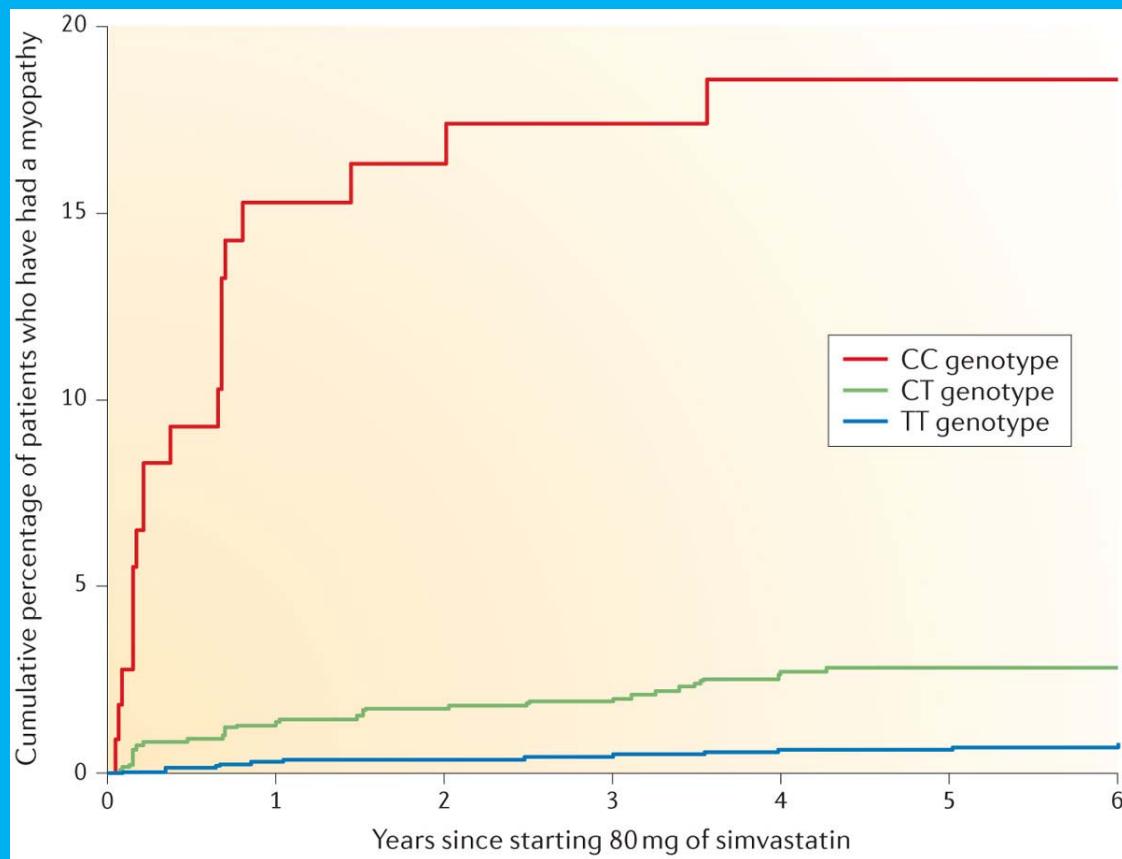
The HIT Approach to Big Data

Brand new insights ... from

1. Harnessing clinical free text
2. Combining data to discover new treatments or approaches
3. Mining device data



We need TRANSLATION now of big data genomic results into clinical care



Nature Reviews | Genetics

Estimated cumulative risk of myopathy associated with high-dose simvastatin by solute carrier organic anion transporter family member 1B1 (*SLCO1B1*) rs4149056 genotype. The figure is modified from Ref. 71 © (2008) Massachusetts Medical Society.

Clinical data major issues

Point	Example
Clinical data has <i>potential</i> to transform research across the translational continuum	Multiple opportunities to transform knowledge discovery and generate value
Clinical data is increasingly <i>available</i> but requires <i>increased ethical protections</i>	Increasing EHR adoption + interest -> concerns about privacy, confidentiality, use, and security
Clinical data is <i>incomplete, inaccurate, and messy</i>	Computing phenotypes is challenging
You can be <i>part of the solution by sharing data, recording metadata, and being responsible</i>	Up to you ... just do it!



Where we need to focus: Human-Data Interaction



