



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

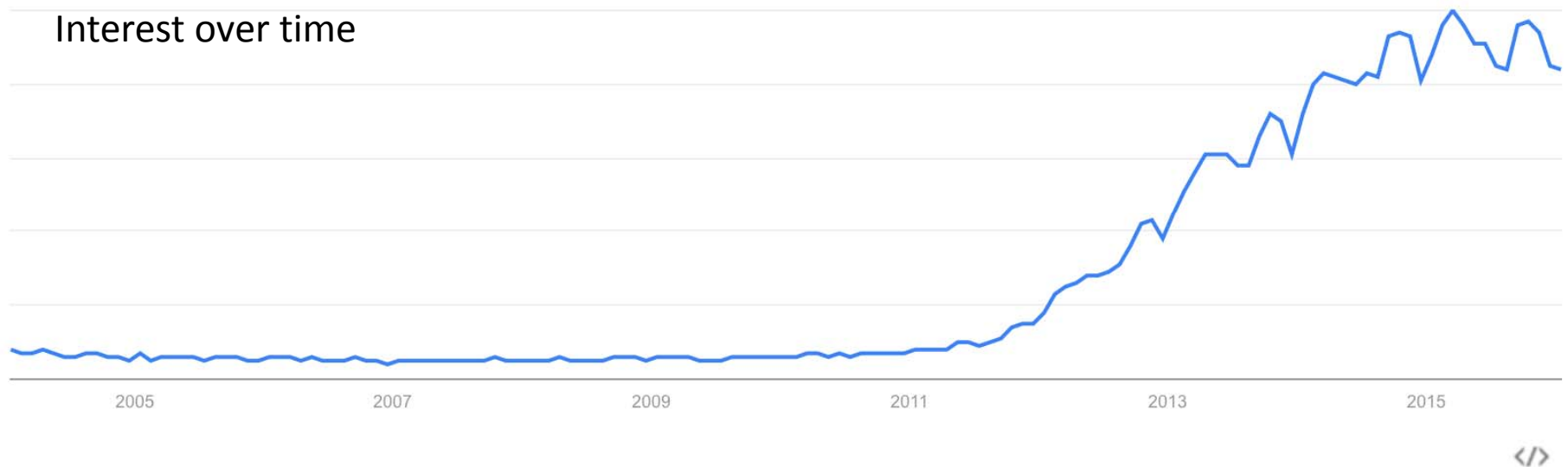


Chart created 13th January 2016



What is Big Data?

Interactive Q&A

The 4 V's

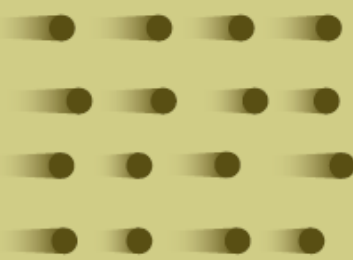
Volume



Data at rest

Terabytes to exabytes
of existing data
to process

Velocity



Data in motion

Streaming data,
milliseconds to
seconds to respond

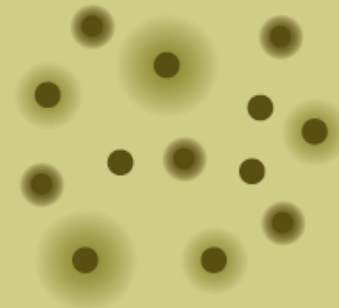
Variety



Data in many forms

Structured, unstructured,
text and multimedia

Veracity



Data in doubt

Uncertainty due to data
inconsistency and
incompleteness,
ambiguities, latency,
deception and model
approximations



What does this mean?

Why does this matter to me?

The **Fourth** Paradigm*



First

DESCRIPTIVE



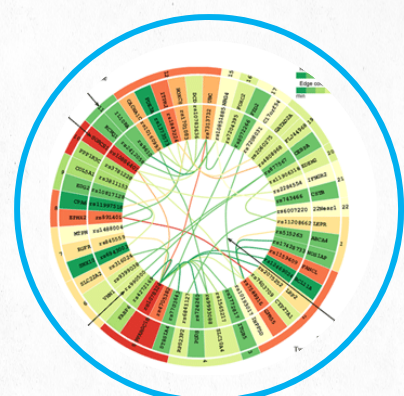
Second

THEORETICAL



Third

COMPUTATIONAL



Fourth

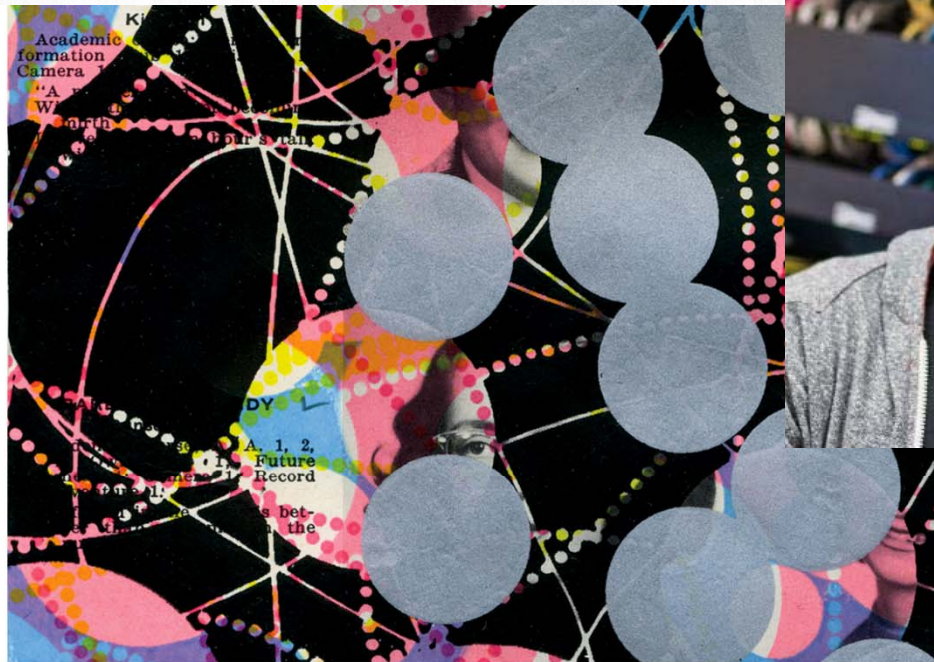
EXPLORATION

*Jim Gray, Microsoft



Career Prospects

Harvard
Business
Review



**SEXIEST
MAN
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

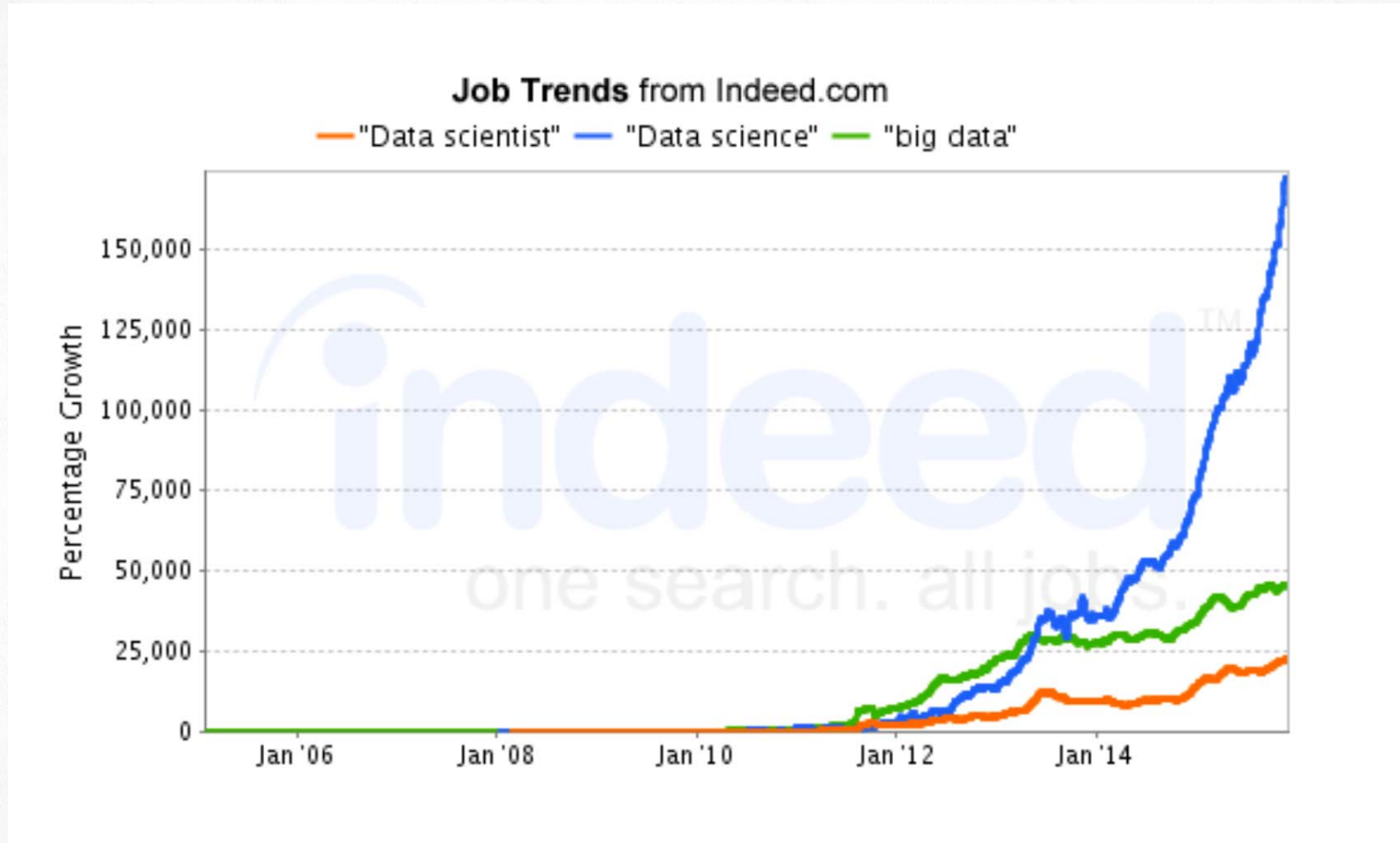


Chart created 13th January 2016



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 »](#)



Subscribe to Blog via Email

Join 11,609 other subscribers

Subscribe

Pages

Share this:

Email Facebook 1 Twitter

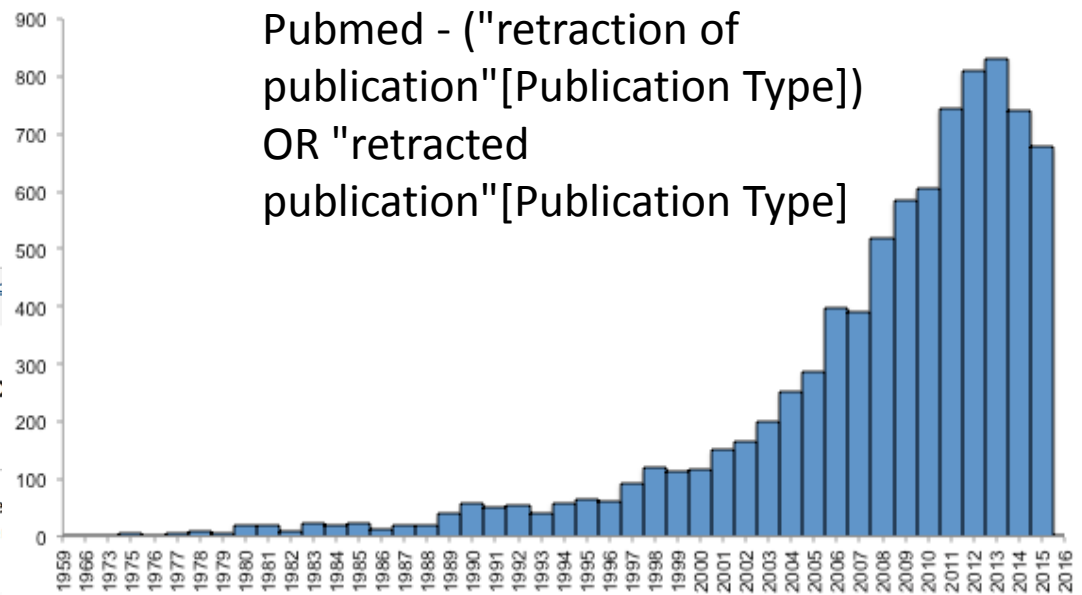
Written by Ivan Oransky
January 13th, 2016 at 9:30 am

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 pape its contents.



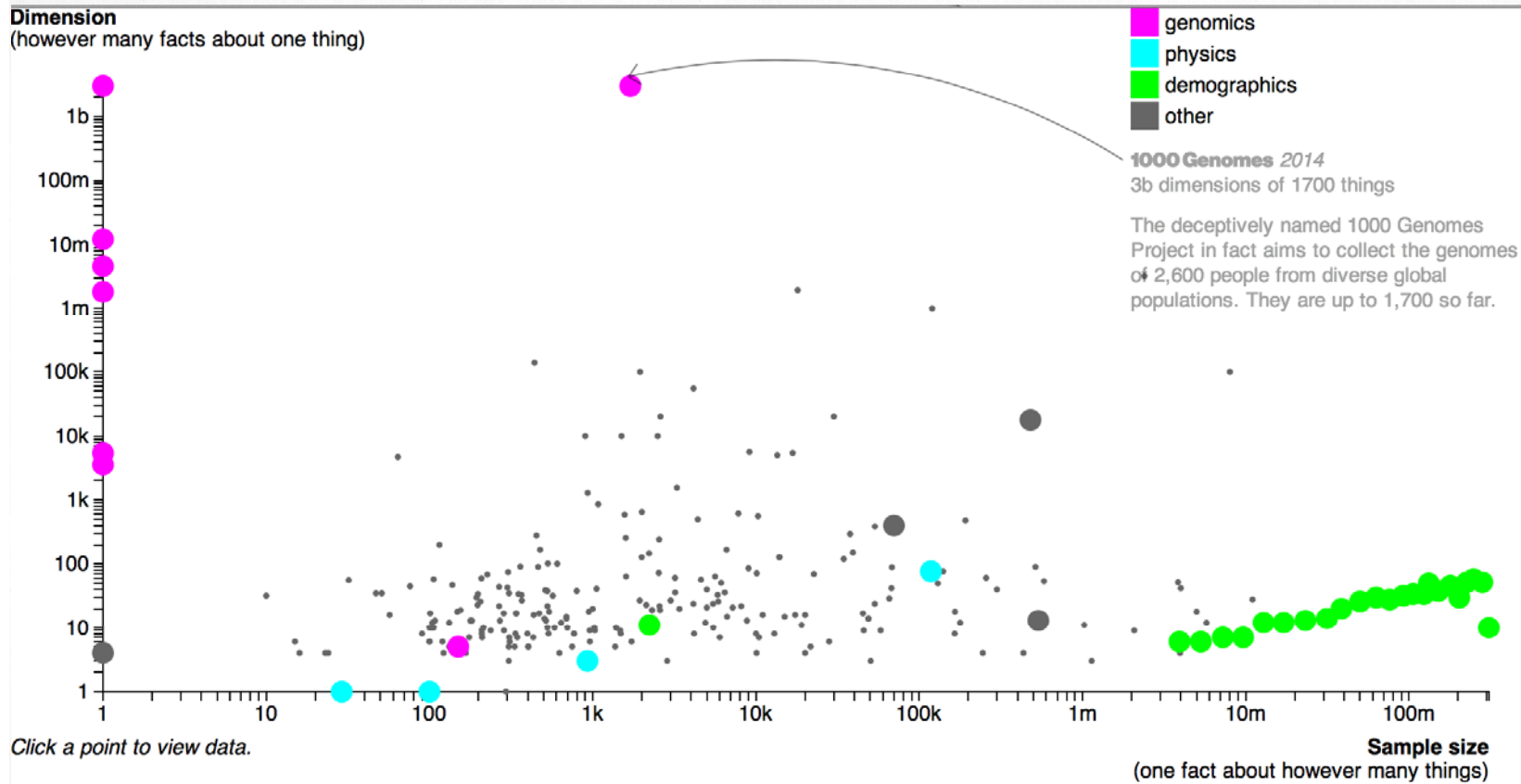
Quantifying Data

Quantified Self and Your Personal Data

Quantifying Data

Dimension

(however many facts about one thing)

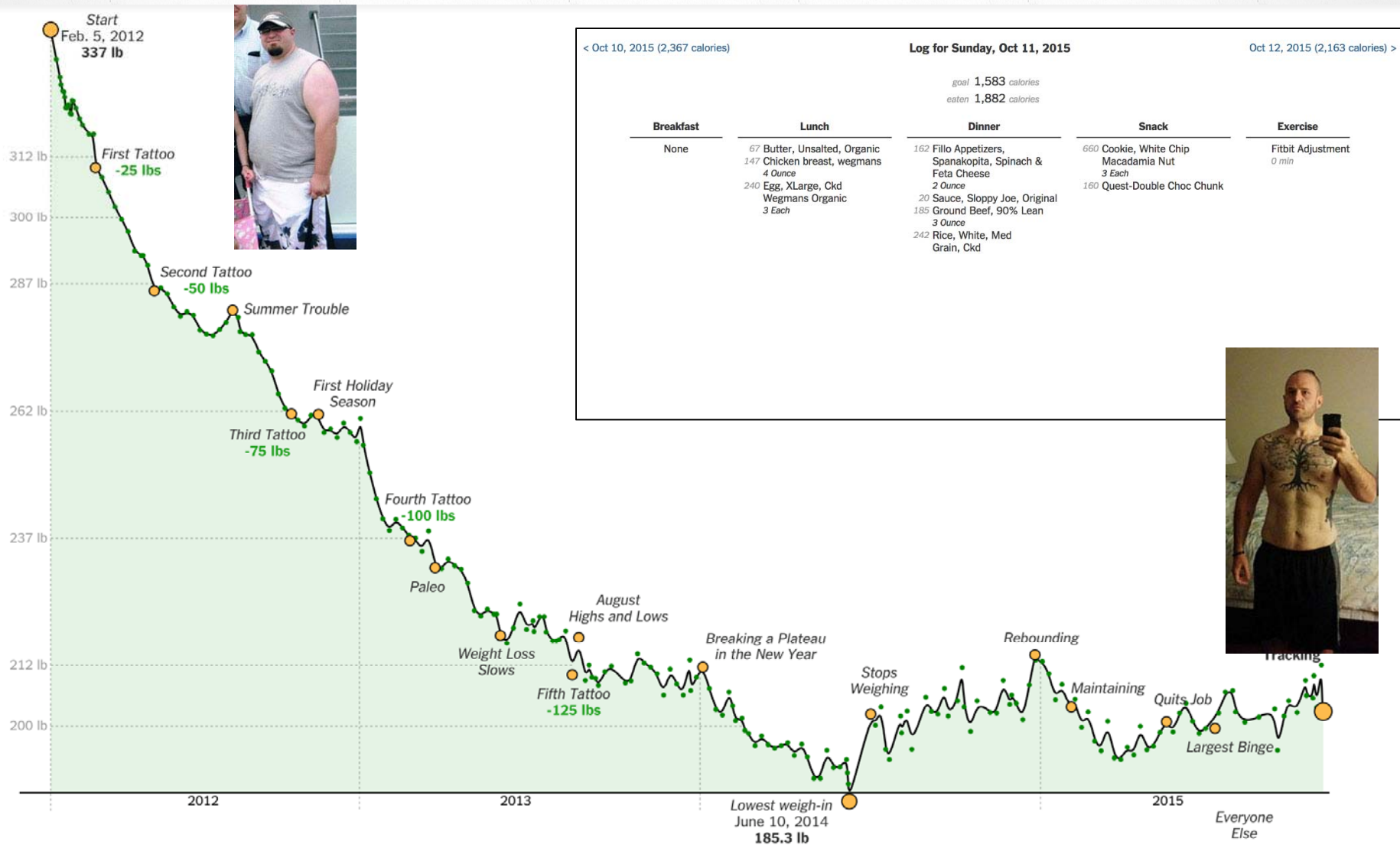


GRAPHIC BY BLOOMBERG BUSINESSWEEK. DATA: UNIVERSITY OF CALIFORNIA IRVINE; AMAZON; KAGGLE; U.S. CENSUS BUREAU; STEPHEN STIGLER; ESA; PLOS; WIKIPEDIA ([CSV](#))



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

1100 Terabytes
Generated per lifetime

Genomics data

30% of determinants of health
Volume

6 TB
Per lifetime

Clinical data

10% of determinants of health
Variety

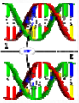















0.4 TB
Per lifetime

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'	Traditional	Quantified Self	Internet of-Things
 <p>Genome: ✓ SNP mutations ✓ Structural variation ✓ Epigenetics ✓</p>	<p>Personal and Family Health History ✓</p>	 <p>Self-reported data: health, exercise, food, mood journals, etc. ✓</p>	 <p>Smart Home ✓</p>
 <p>Microbiome ✓</p>	<p>Prescription History ✓</p>	 <p>Mobile App Data ✓</p>	 <p>Smart Car ✓</p>
 <p>Transcriptome</p>	<p>Lab Tests: History and Current ✓</p>	 <p>Quantified Self Device Data ✓</p>	 <p>Personal Robotics ✓</p>
 <p>Metabolome</p>	<p>Demographic Data ✓</p>	 <p>Biosensor Data Objective Metrics</p>	 <p>Environmental Sensors ✓</p>
 <p>Proteome</p>	<p>Standardized Questionnaires ✓</p>	 <p>Community Data ✓</p>	 <p>Community Data ✓</p>
 <p>Environmentome ✓</p>			

Legend: Consumer-available ✓

Data Ecosystem

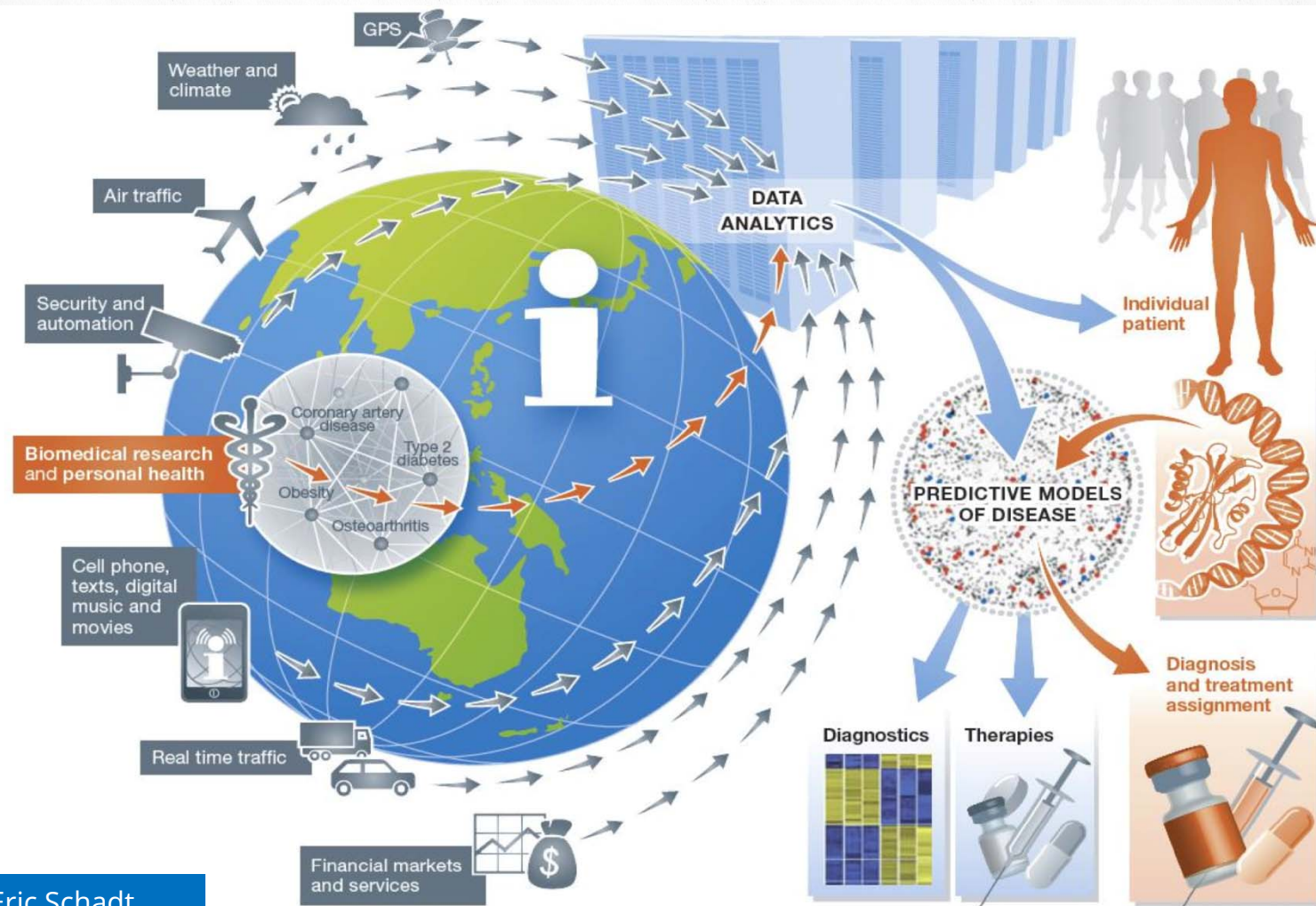


Image: Eric Schadt

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>



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

Matchmaker Exchange

<http://www.matchmakerexchange.org>



Consumer Genetic Testing



23andMe

welcome

Here's a preview of
when you receive yo

Hereditary Breast and Ovarian Cancer Syndrome (BRCA1- and BRCA2-Related, Selected Mutations)

Established Research report on 3 reported markers.

Example Data

Resources

Technical Report



If this condition runs in your family or you think you might have this condition, consult with a healthcare provider about appropriate testing.

Other factors can also influence your risk for this condition even if you don't have the variant(s) covered by this report.

About Hereditary Breast and Ovarian Cancer Syndrome

color

BUY

LEARN

SUPPORT

PROVIDERS

ACTIVATE KIT

SIGN IN

Understand your genetic risk for breast and ovarian cancer

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.

Purchase Color



23

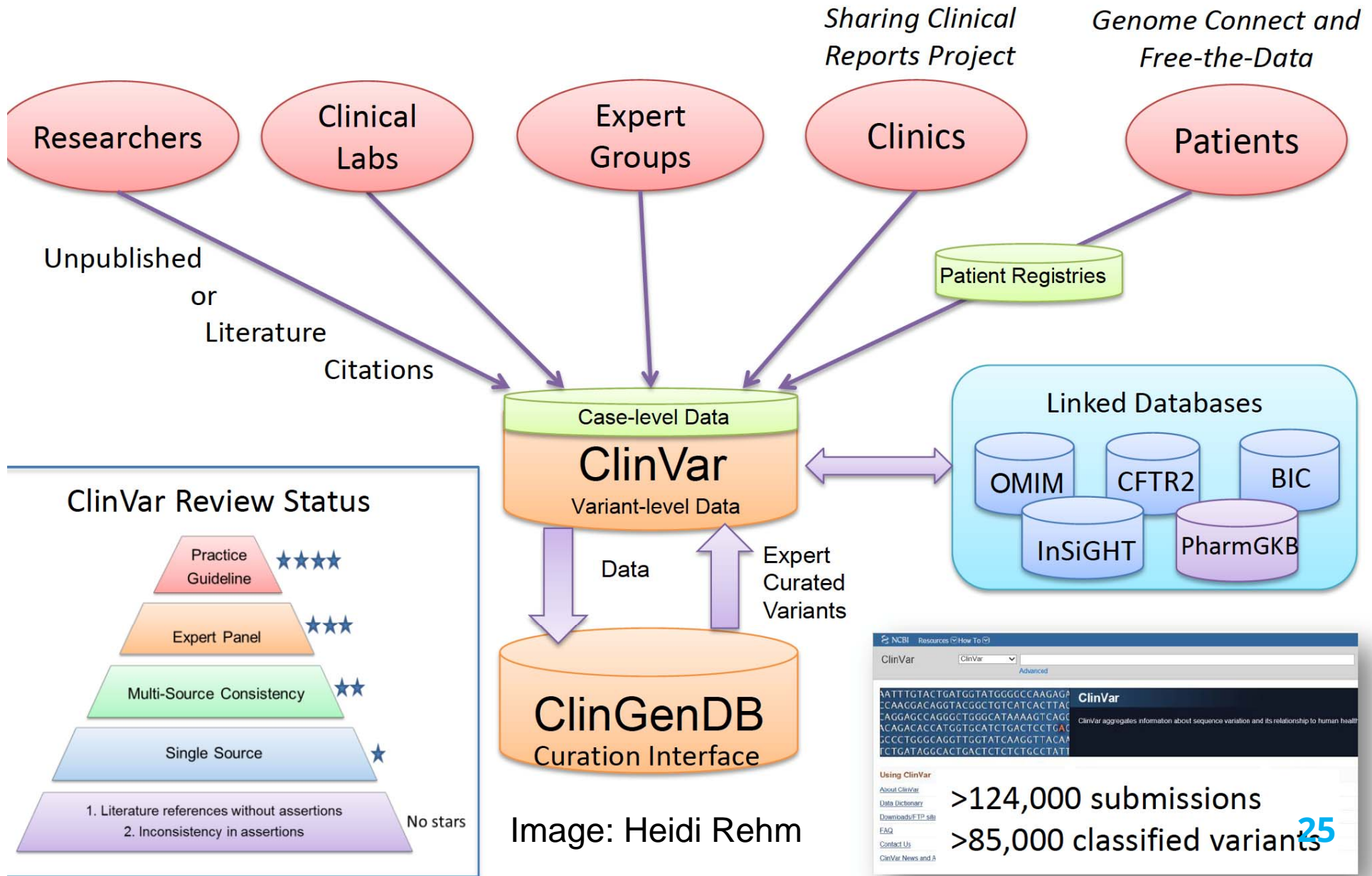
ClinGen

<http://clinicalgenome.org>

The banner features a background image of a laboratory with a person in a lab coat and various test tubes. Overlaid on this is a large, semi-transparent blue rectangle containing text. In the top left corner of the banner is the ClinGen logo, which consists of a stylized DNA double helix and the text 'ClinGen Clinical Genome Resource'. A horizontal navigation bar is positioned above the main text area, containing links for 'About', 'Data Sharing', 'Knowledge Curation', 'Machine Learning', 'Events & News', and 'Tools & Resources'. To the right of this bar are three additional links: 'For Patients', 'Search', and 'Contact'. The main text area contains the headline 'ClinGen: Sharing Data. Building Knowledge. Improving Care.' followed by a paragraph about technological advances in genome-wide analysis and the need for coordinated effort. A 'Learn more »' link is provided at the end of the paragraph. Below the text is a small navigation indicator consisting of four circles, with the first one filled in white and the others empty.

Data Flows in ClinGen

(>200 ClinVar submitters)



Patient Provided Data

patientslikeme®

Already a member? Sign in.



Join now

(it's free!)



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



Track your health

Chart your health to research the all

PatientCrossroads™

Home

About CONNECT ▾



Click here to register now!

GenomeConnect – Partnering with YOU to advance genomic health

GenomeConnect is a patient portal, or registry, that is working to build the knowledge base about genetics and health that will allow researchers and doctors to study the impact of genetic variation on health conditions. This knowledge is key to the development of new treatments and therapies.

Registries like GenomeConnect make medical discoveries possible by bringing together information from a large number of patients. YOUR participation in GenomeConnect will help bring the future of genomic medicine one step closer!

FREE
THE DATA

Free My Data

Join the Movement

Learn More

For Clinicians

About Us

FOLLOW US:

PLAYLIST Why Free the Data?



PLAY ALL

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!

Free My Data

Your mutation and health information are important in the search for better health. Share your information safely and securely.

Free My Data

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!

Join the Movement

Learn More

Learn why sharing your mutation is important and how the privacy and sharing system works.

Learn More





Clinical data – not as big, but

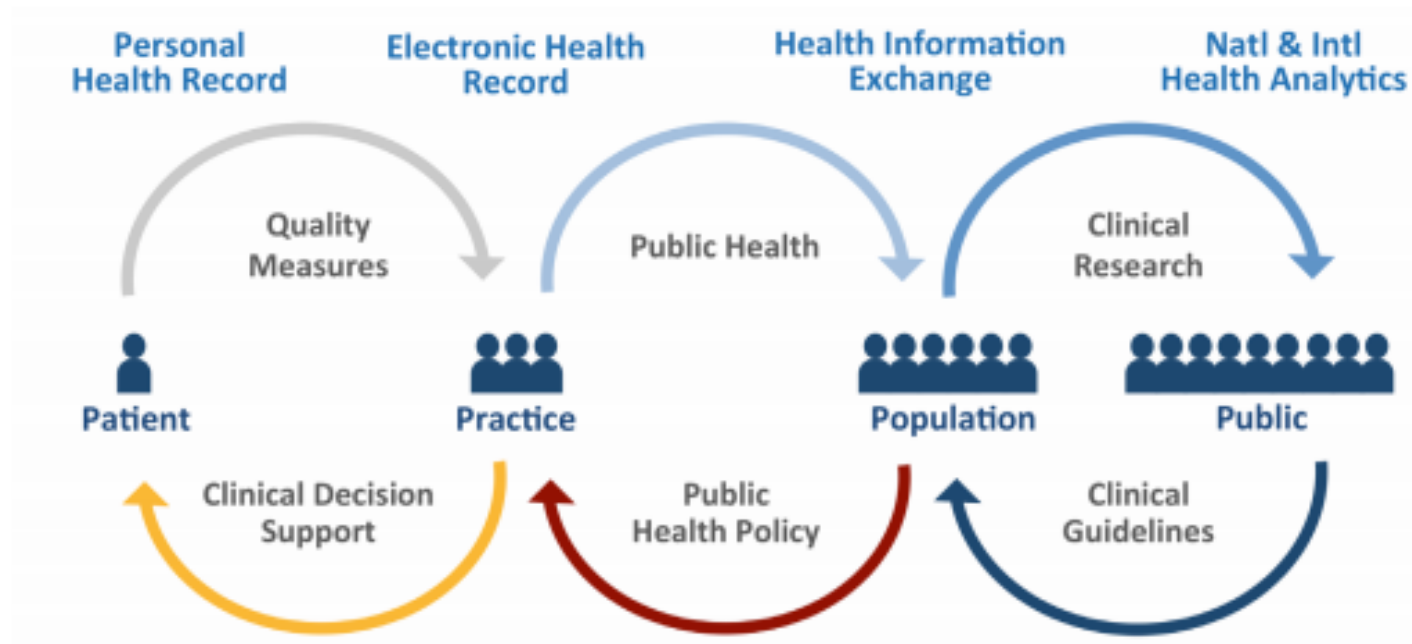
It may matter even more to you!



David Dorr, MD, MS
13th January 2016

Clinical Data across the health ecosystem

Figure 1. Vision of the Health IT Ecosystem



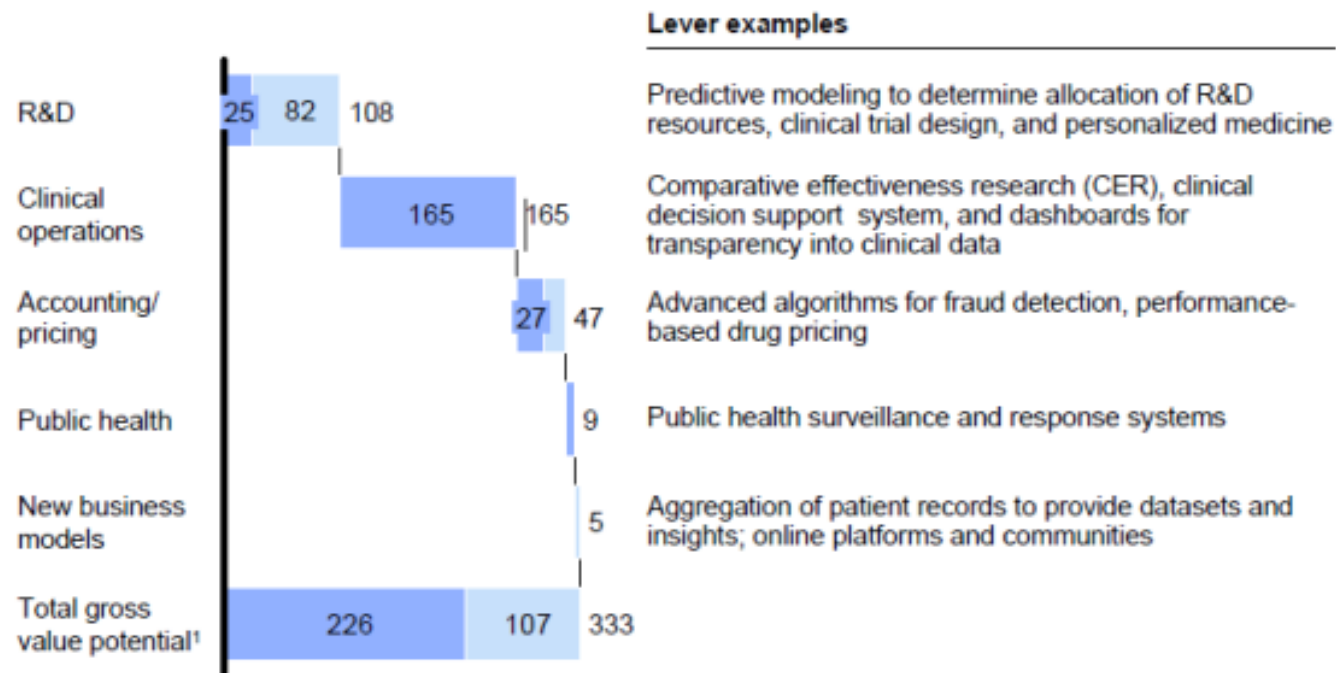
Source: Office of the National Coordinator

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

www.wipro.com

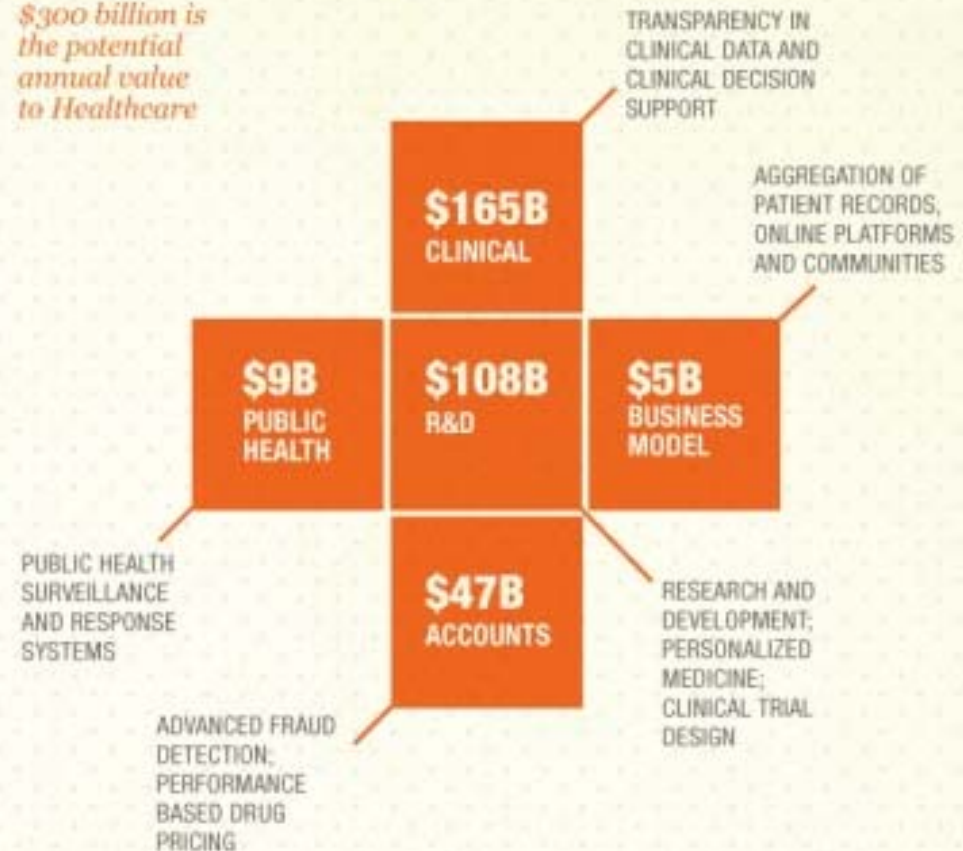
BIG DATA

Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze.

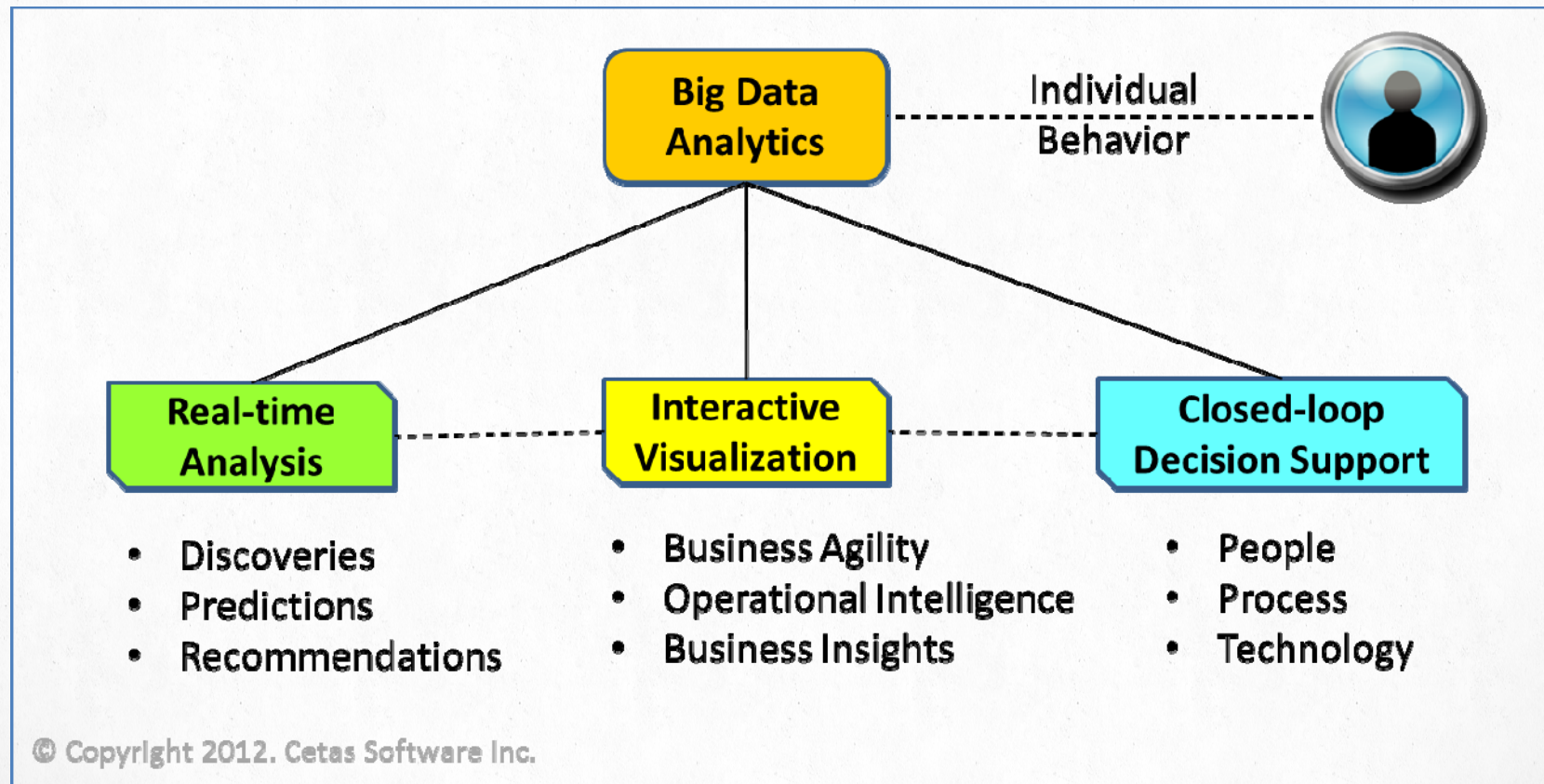
The right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods.

CASE STUDY - Healthcare

\$300 billion is the potential annual value to Healthcare

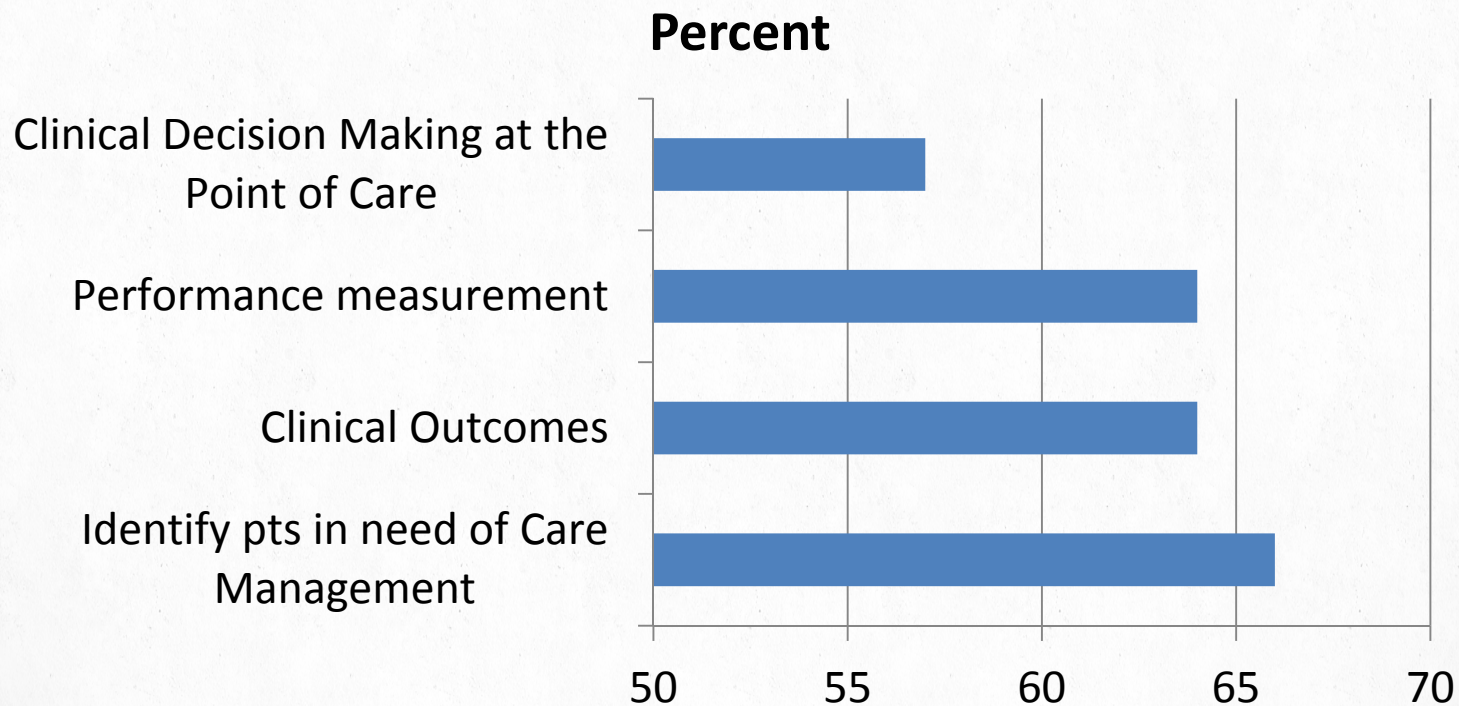


The future? (or the present, just not evenly distributed*)



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

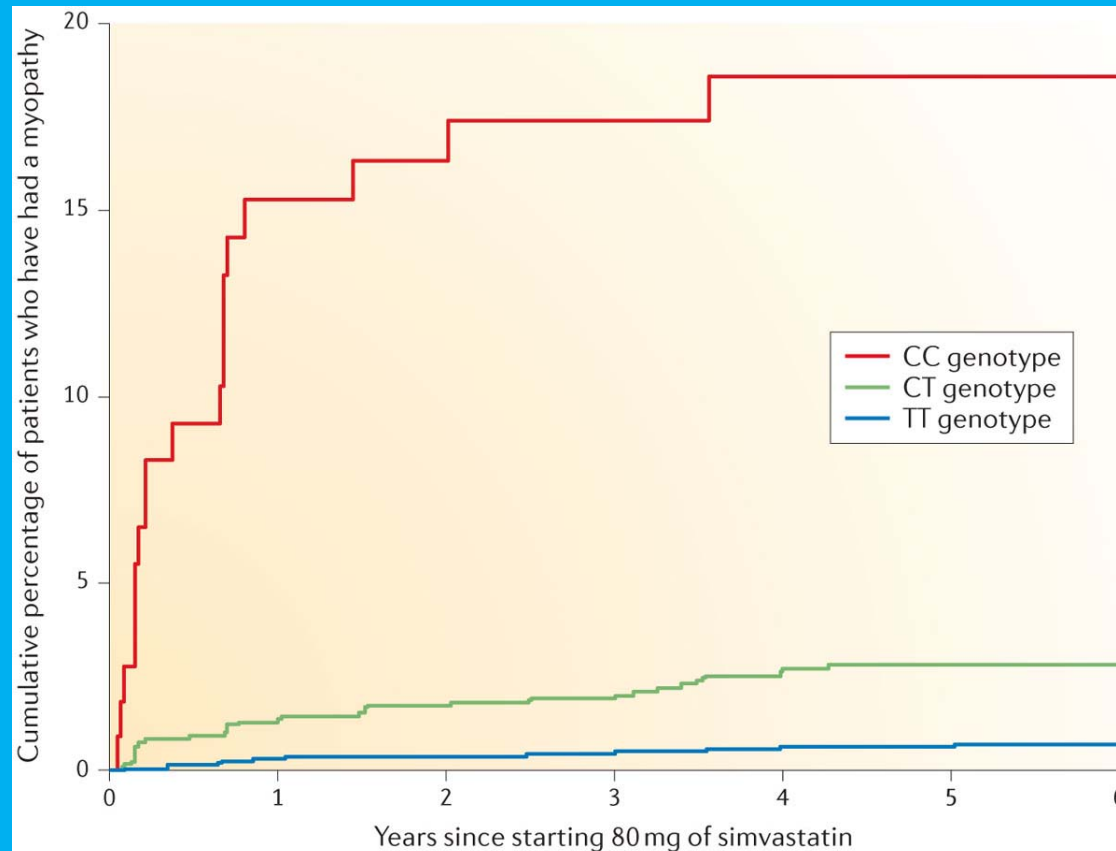
The HIT Approach to Big Data

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</i> by <i>sharing data, recording metadata, and being responsible</i>	Up to you ... just do it!



Where we need to focus: Human-Data Interaction

