What is Big Data?

And Why it Matters to You!

Shannon McWeeney, PhD

13th January 2016

Data After Dark

OHSU BD2K Data Science Workshop

Google Trends: “Big Data”

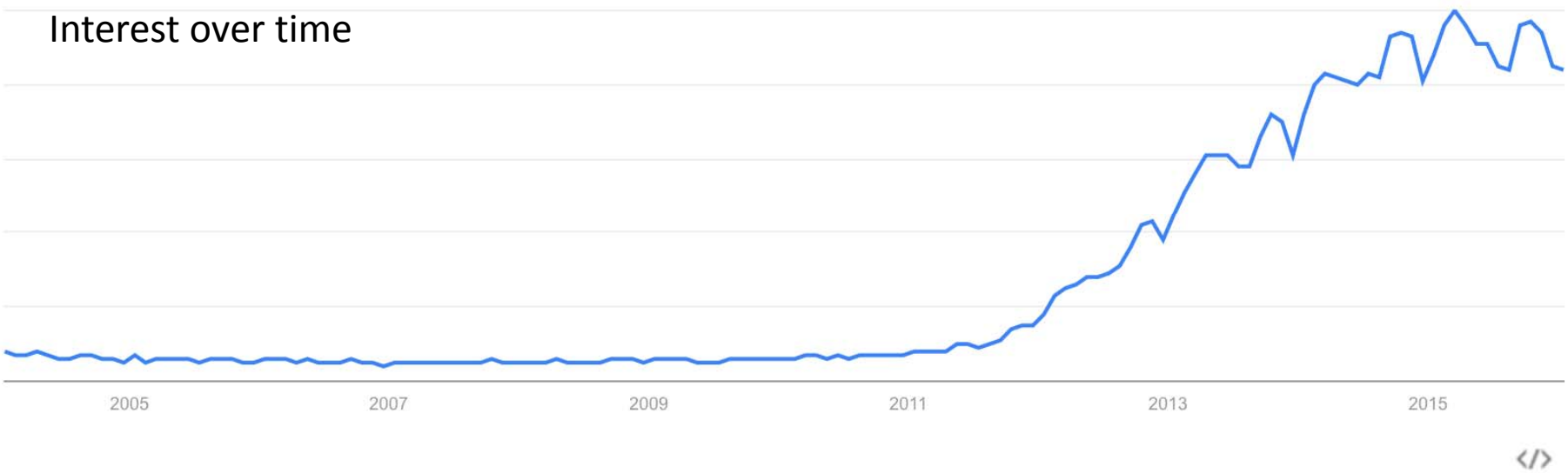
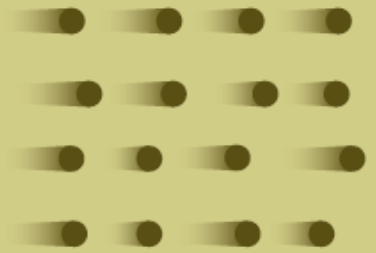


Chart created 13th January 2016

What is Big Data?

Interactive Q&A

The 4 V's

Volume	Velocity	Variety	Veracity
			
Data at rest	**Data in motion**	**Data in many forms**	**Data in doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text and multimedia	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

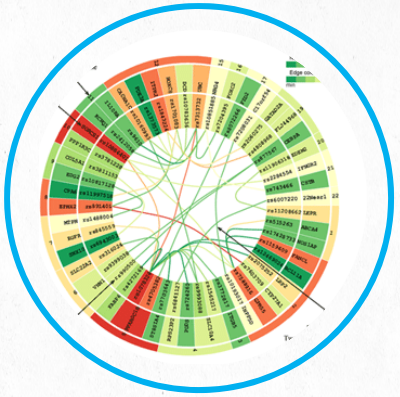
DESCRIPTIVE

Second

THEORETICAL

Third

COMPUTATIONAL



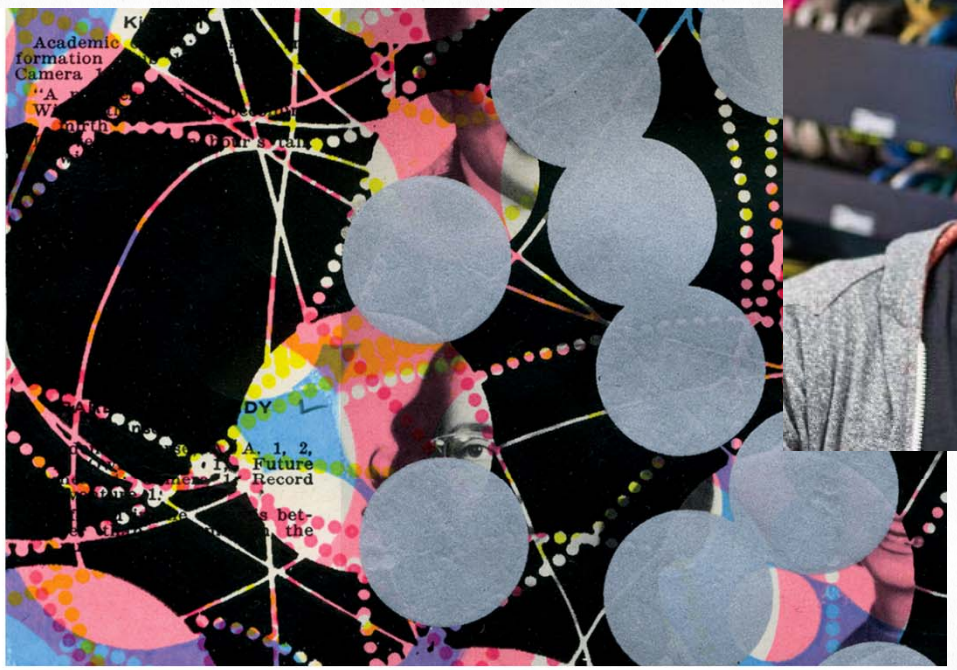
Fourth

EXPLORATION

*Jim Gray, Microsoft

Career Prospects

Harvard Business Review



SEXIEST MAN 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

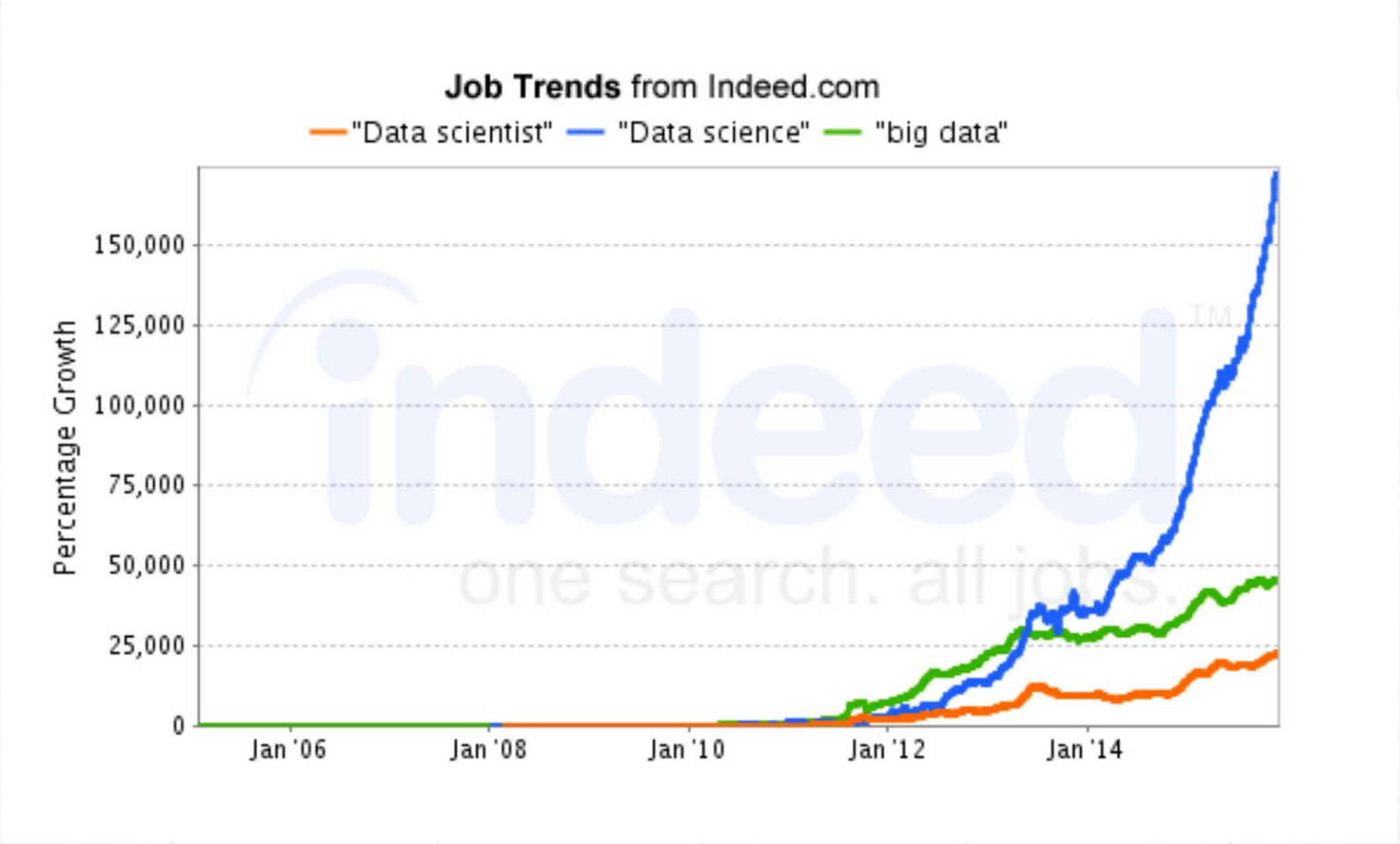


Chart created 13th January 2016

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 »

Subscribe to Blog via Email

Join 11,609 other subscribers

Email Address

Subscribe

Pages

Share this:

Email Facebook 1 Twitter

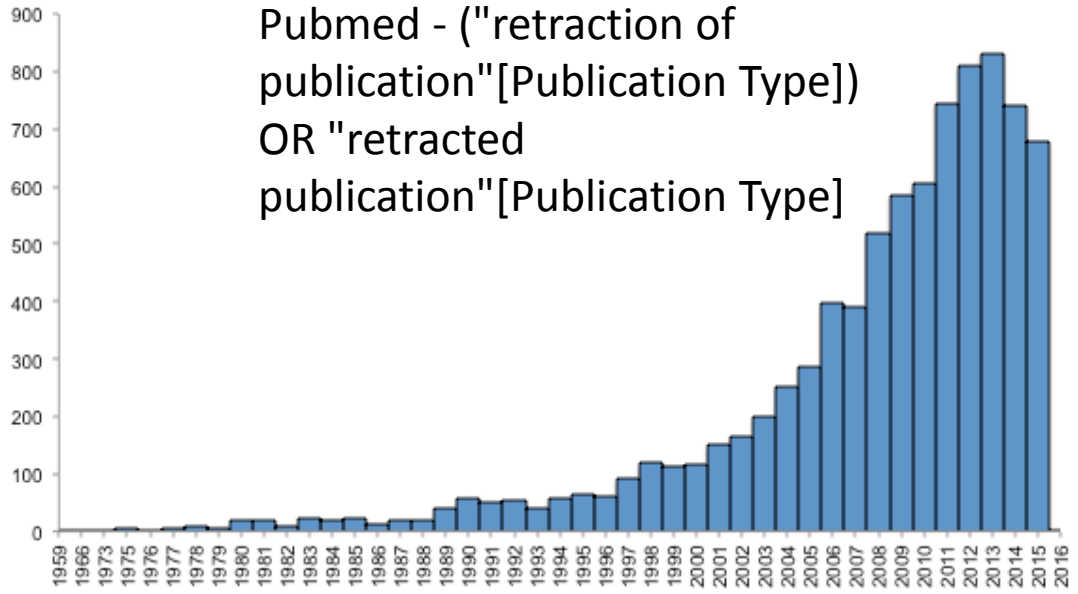
Written by Ivan Oransky
January 13th, 2016 at 9:30 am

Posted in dc

Data dispute forces journal tc 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.

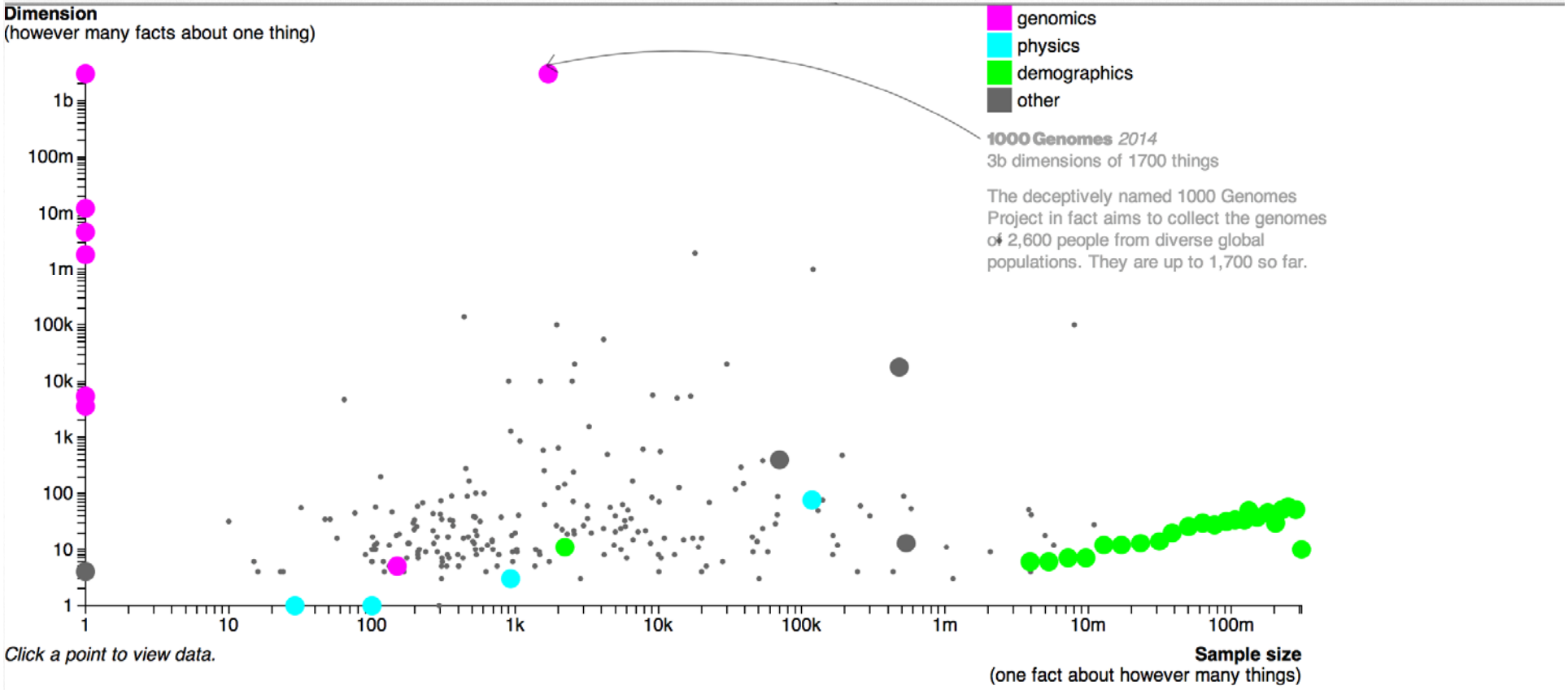


Pubmed - ("retraction of publication"[Publication Type]) OR "retracted publication"[Publication Type]

Quantifying Data

Quantified Self and Your Personal Data

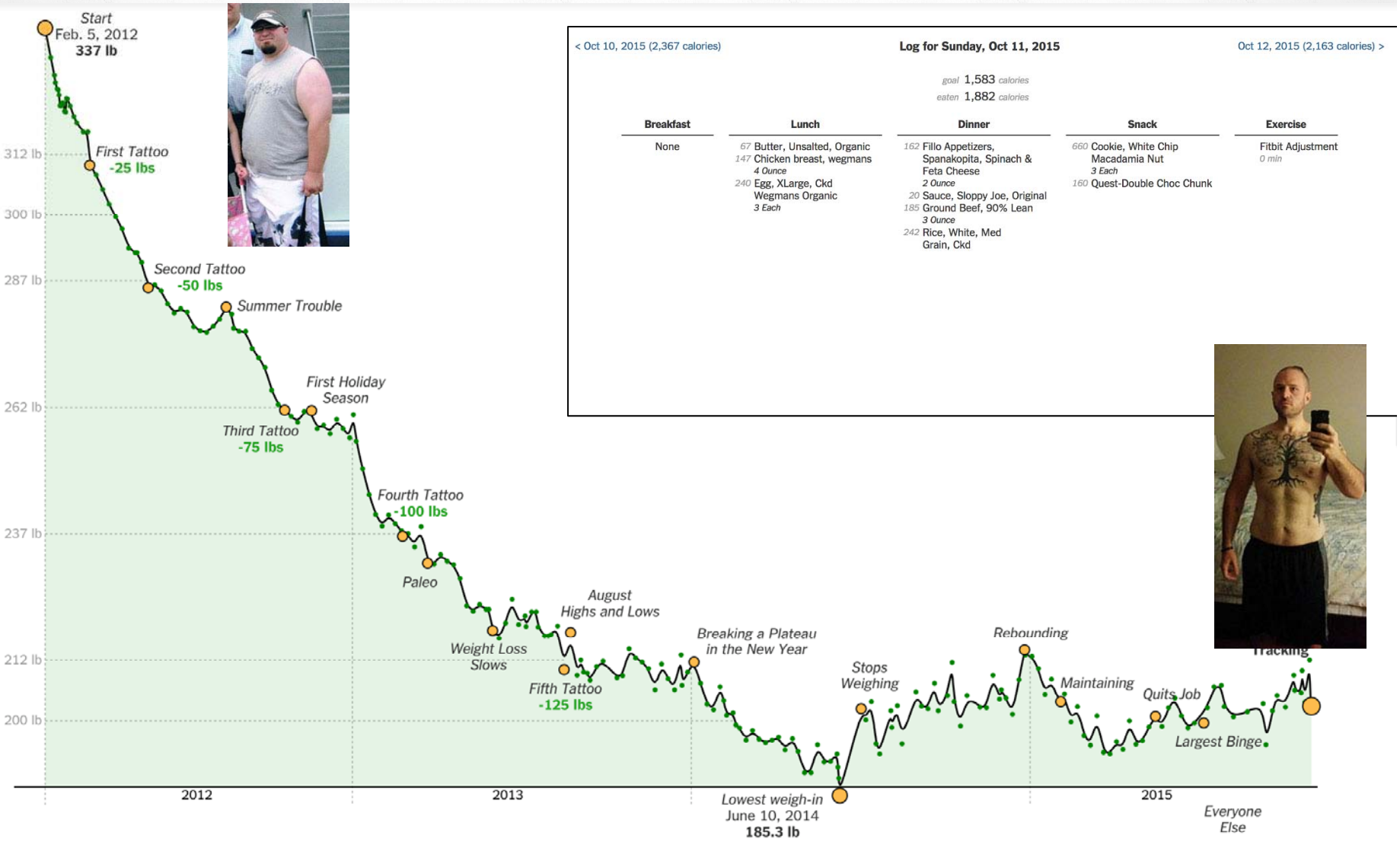
Quantifying Data



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

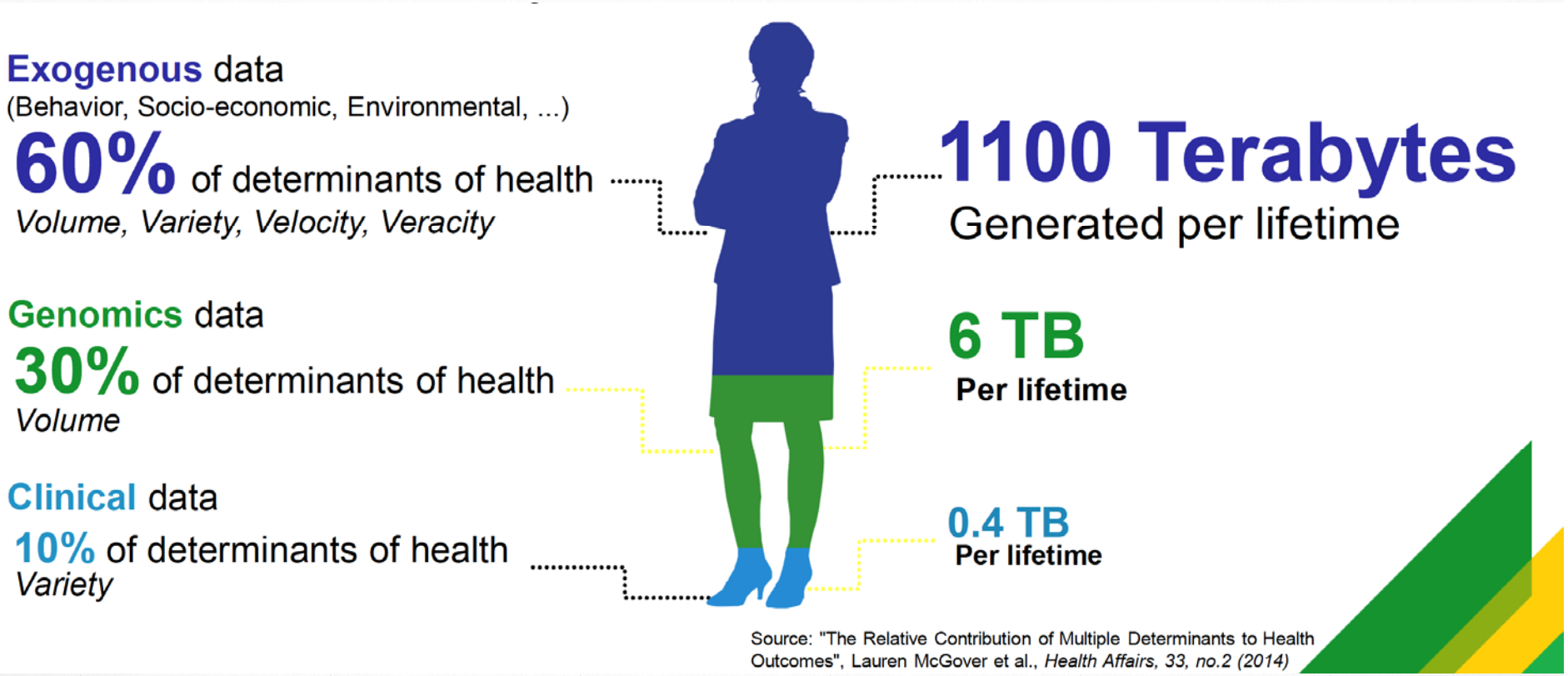
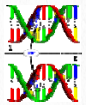


Image: IBM

Personal Information Streams

'Omics'



Genome:
SNP mutations
Structural variation
Epigenetics ✓

Microbiome ✓

Transcriptome

Metabolome

Proteome

Diseasome ✓

Environmentome ✓

Traditional

Personal and Family Health History ✓

Prescription History ✓

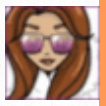
Lab Tests: History and Current ✓

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 ✓

Image: Melanie Swan

Data Ecosystem

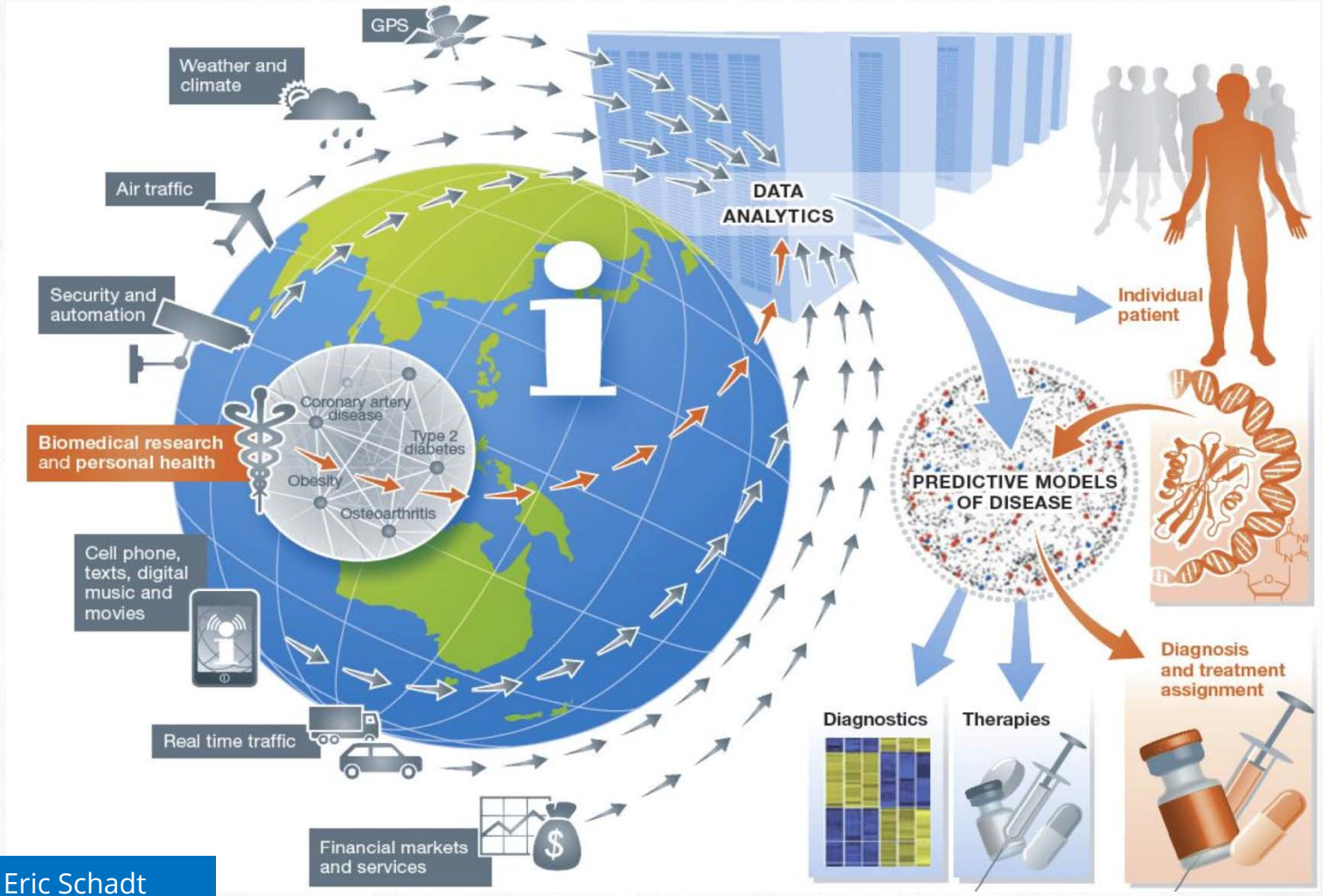


Image: Eric Schadt

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

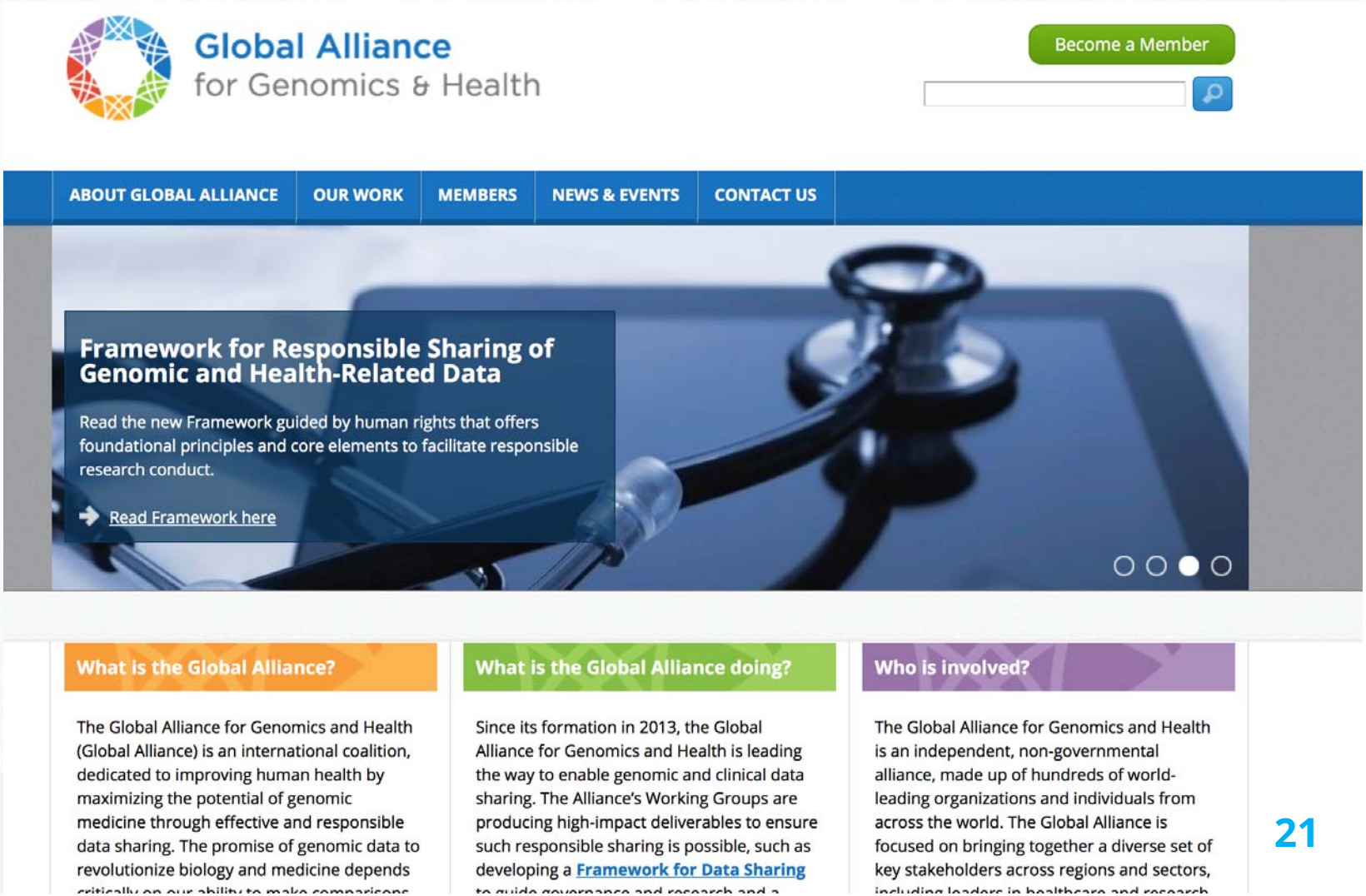
Breakthrough	Why It Matters	Key Players
Technical standards that let DNA databases communicate.	Your medical treatment could benefit from the experiences of millions of others.	+ Global Alliance for Genomics and Health + Google + Personal Genome Project

10 Breakthrough Technologies 2015

Source: Harvard Business Review

GA4GH

http://genomicsandhealth.org

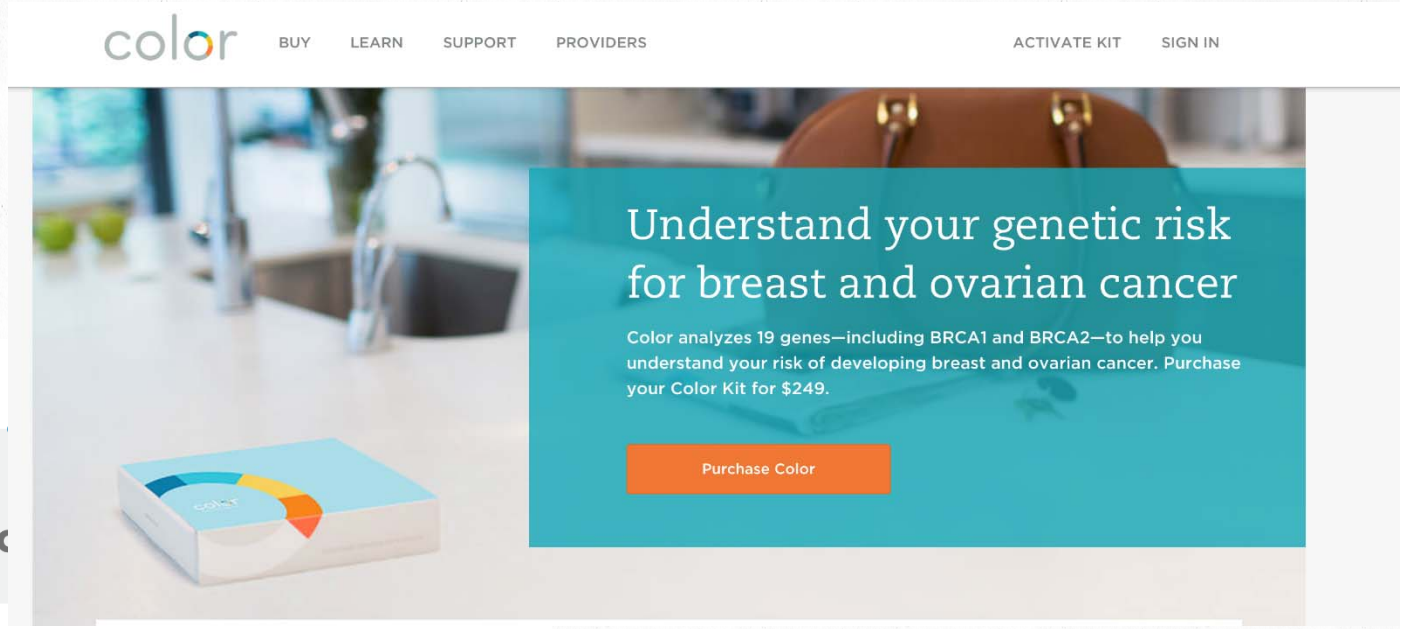


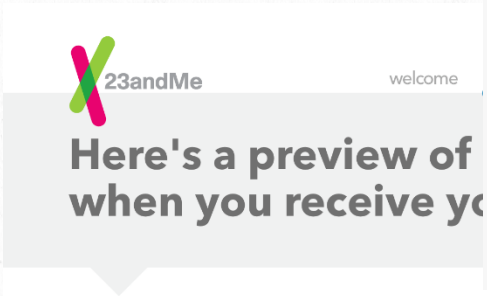
Matchmaker Exchange

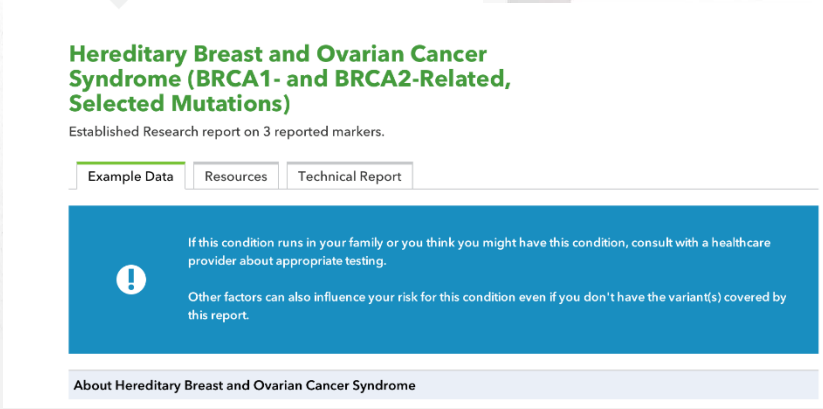
http://www.matchmakerexchange.org

Global Alliance for Genomics & Health

Consumer Genetic Testing







ClinGen

http://clinicalgenome.org

Data Flows in ClinGen

(>200 ClinVar submitters)

Sharing Clinical Reports Project

Genome Connect and Free-the-Data

Researchers

Clinical Labs

Expert Groups

Clinics

Patients

Patient Registries

Unpublished or Literature Citations

Case-level Data

ClinVar

Variant-level Data

Linked Databases

OMIM

CFTR2

BIC

InSiGHT

PharmGKB

Data

Expert Curated Variants

ClinGenDB

Curation Interface

ClinVar Review Status

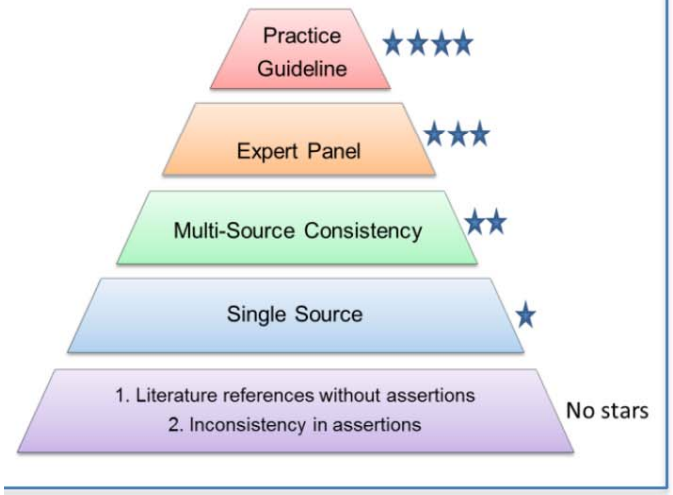
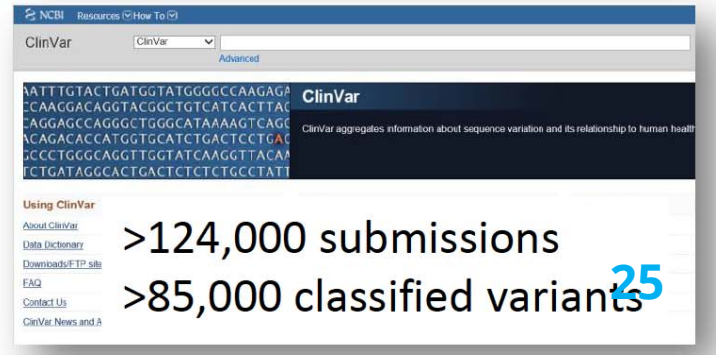


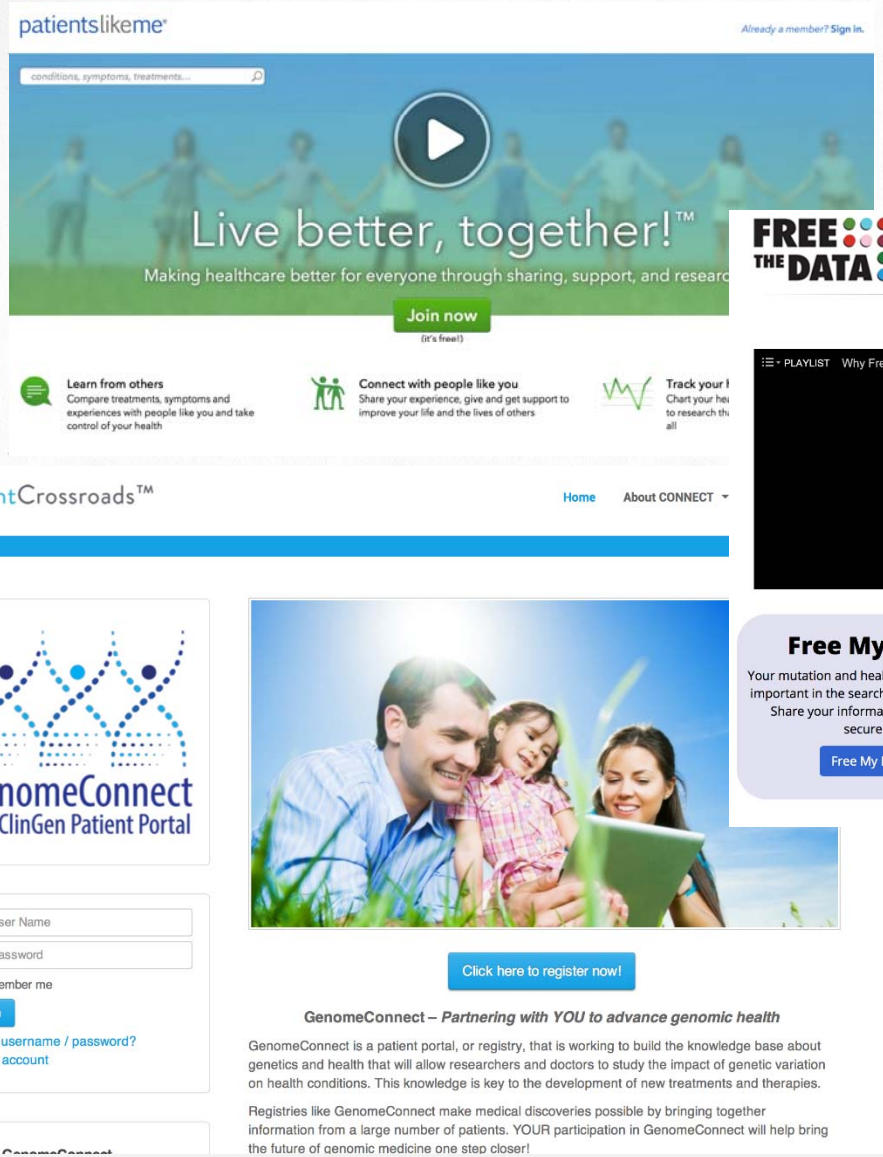
Image: Heidi Rehm

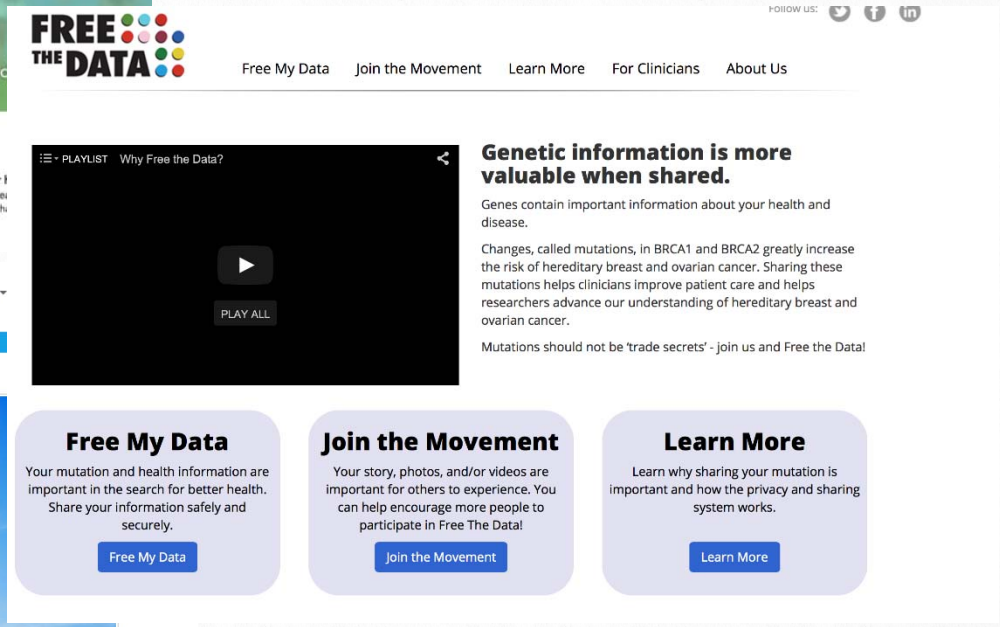


>124,000 submissions

>85,000 classified variants

Patient Provided Data





Clinical data – not as big, but

It may matter even more to you!

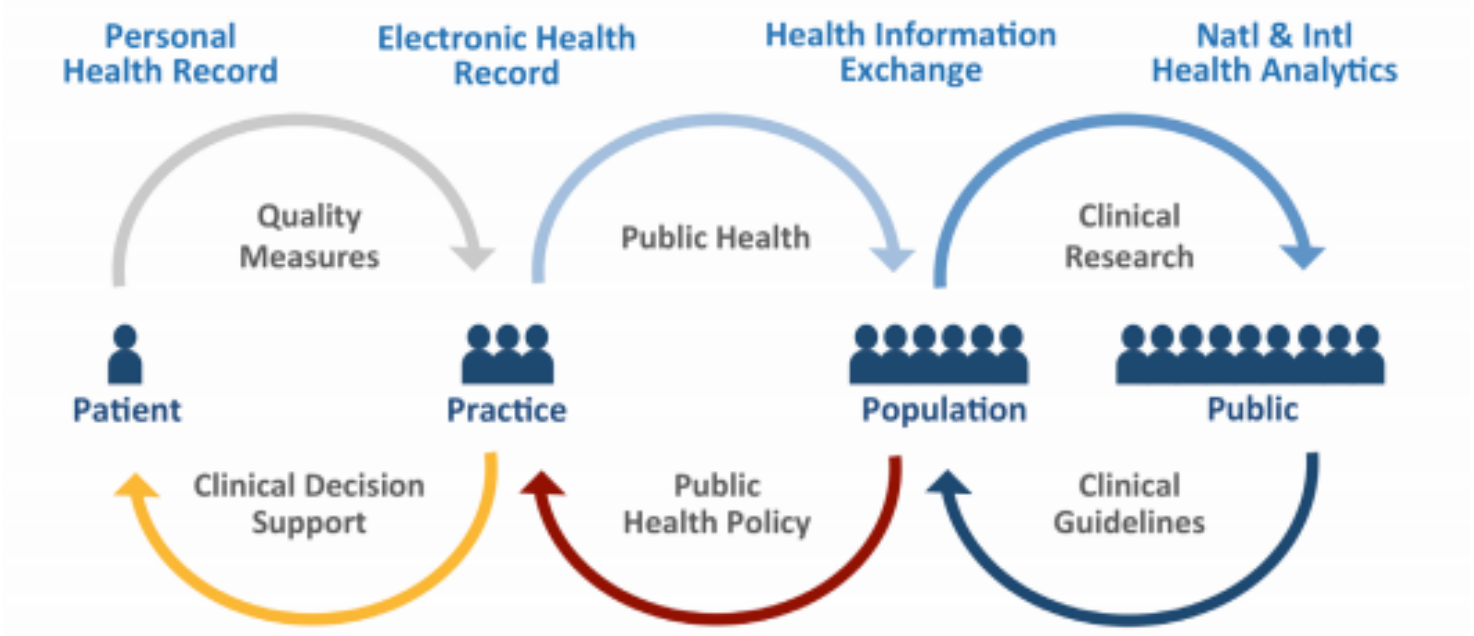
Data After Dark
OHSU BD2K Data Science Workshop

David Dorr, MD, MS

13th January 2016

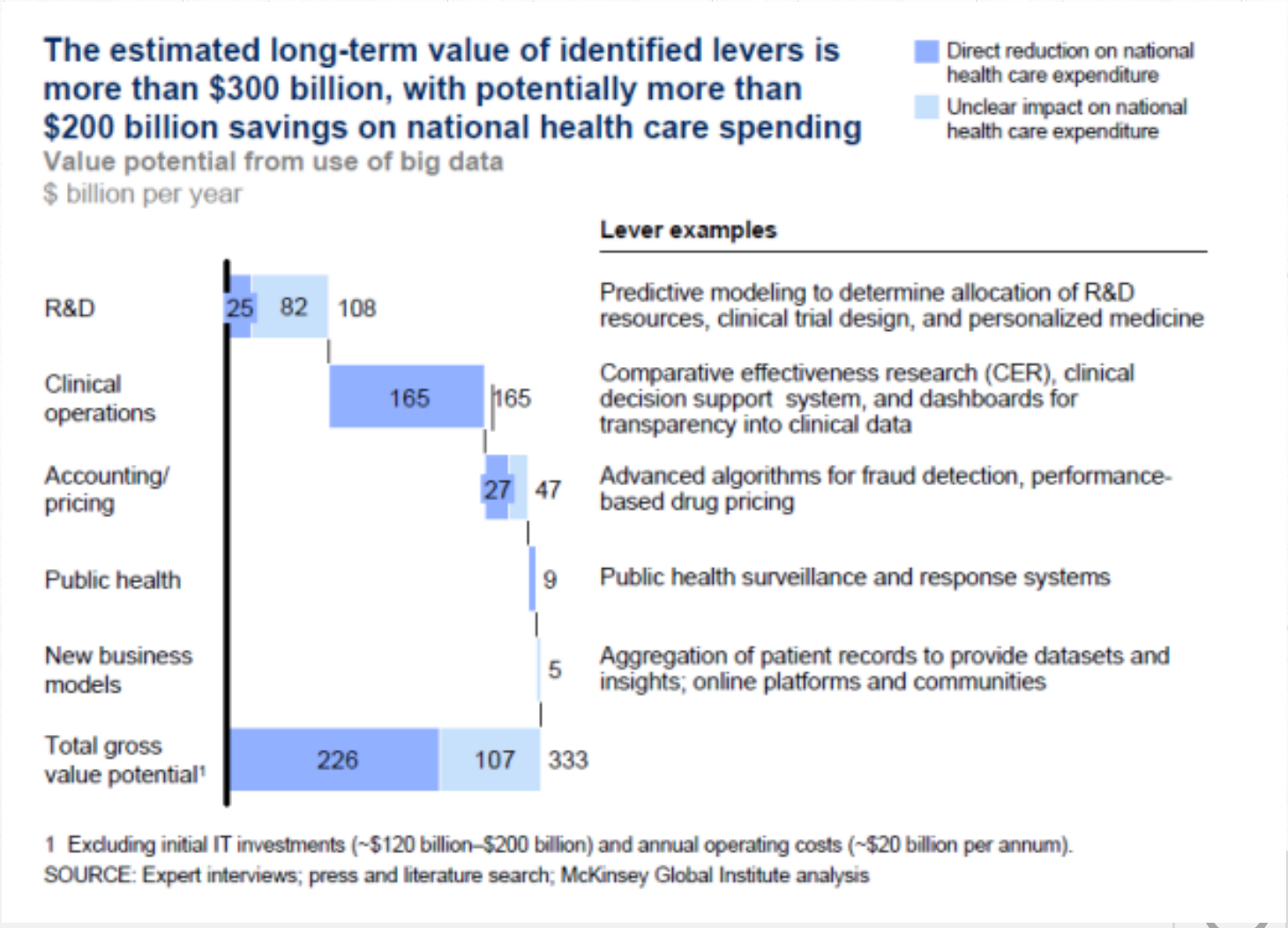
Clinical Data across the health ecosystem

Figure 1. Vision of the Health IT Ecosystem

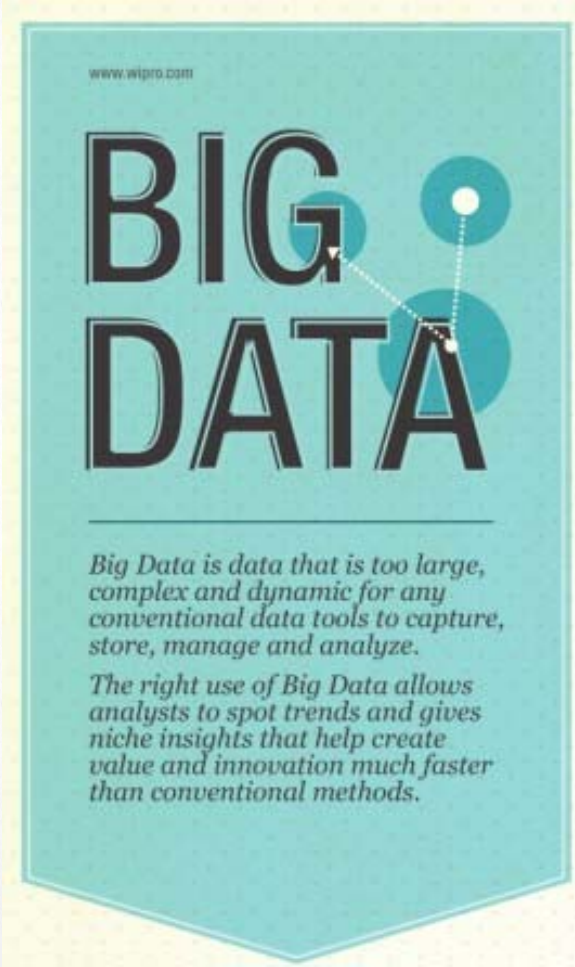


Source: Office of the National Coordinator

Value of clinical data

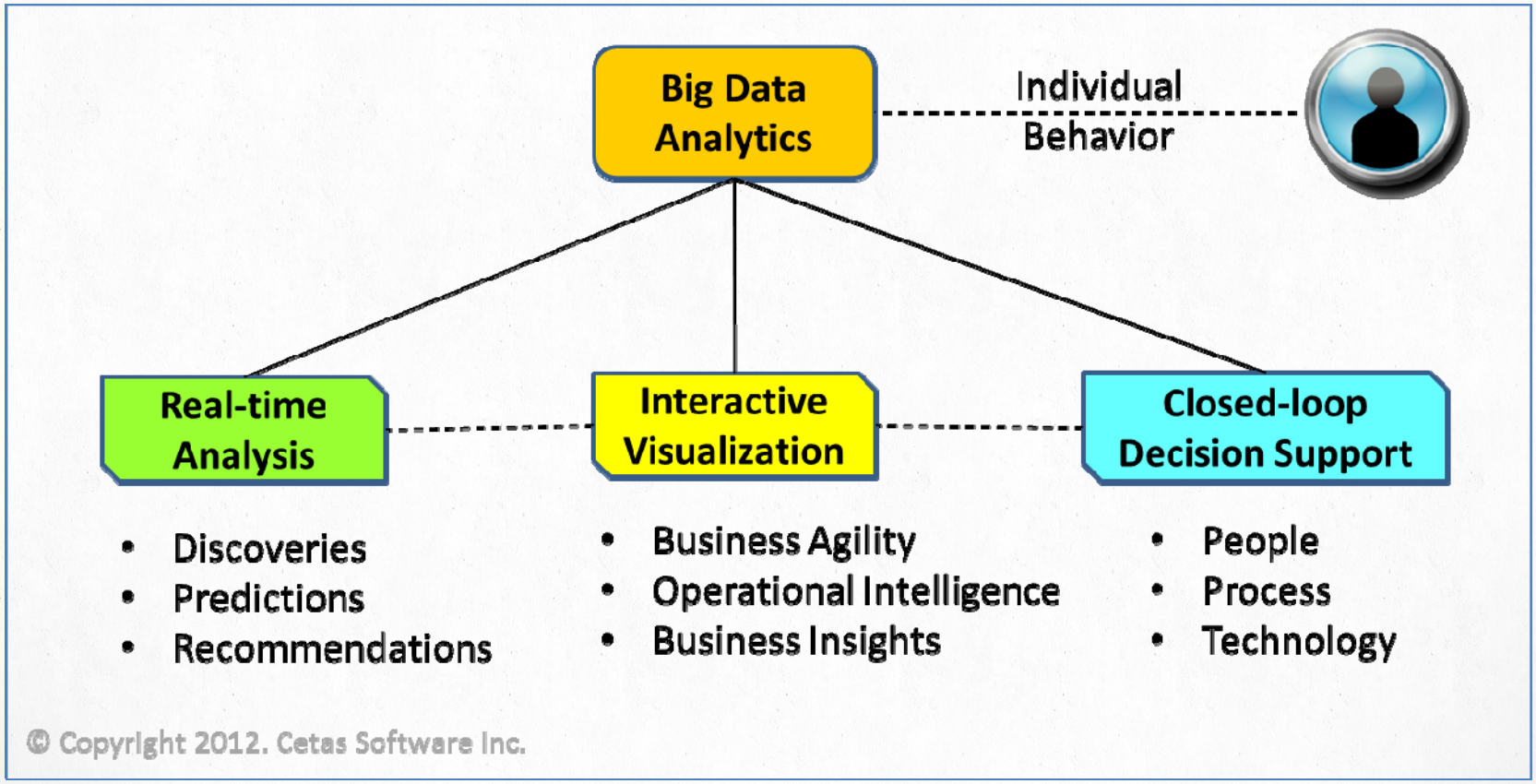


Big Data



http://www.wipro.com/images/infographic.jpg

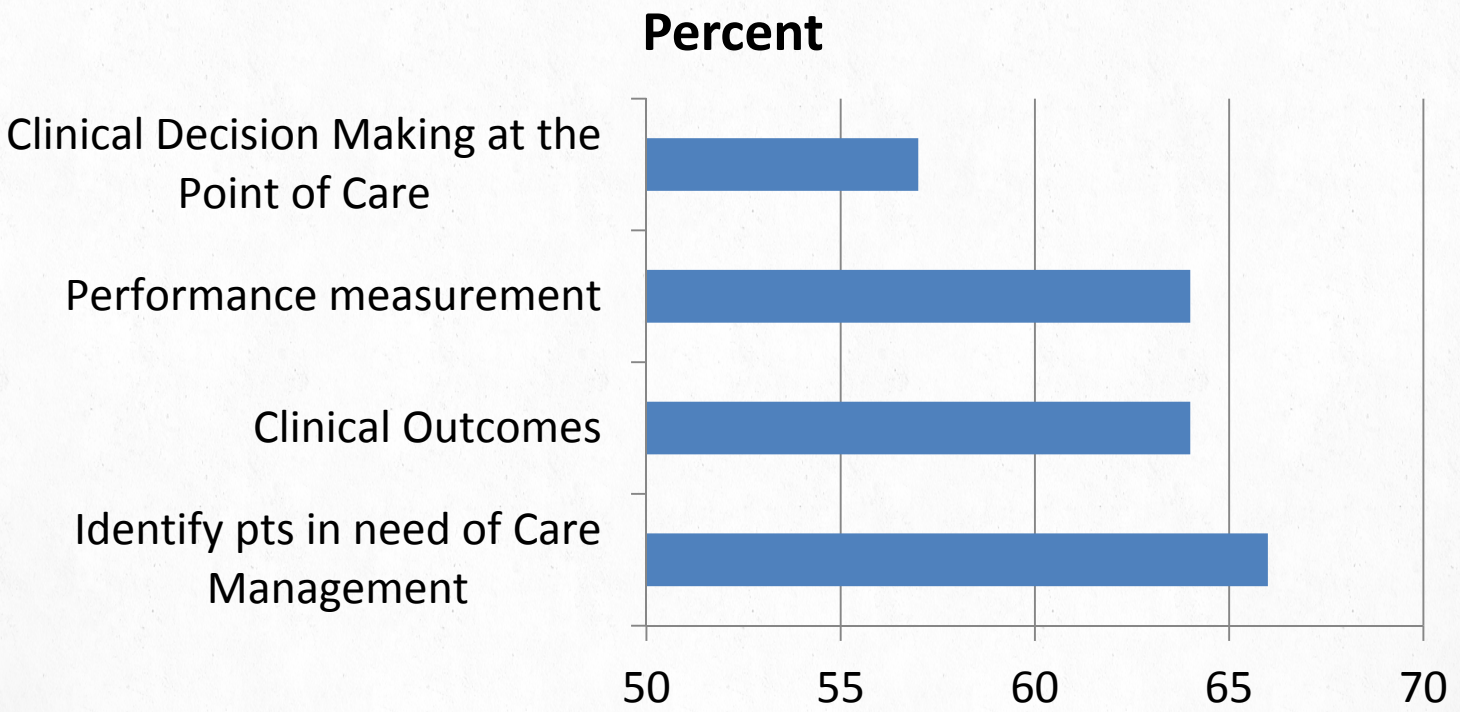
The future? (or the present, just not evenly distributed*)



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:



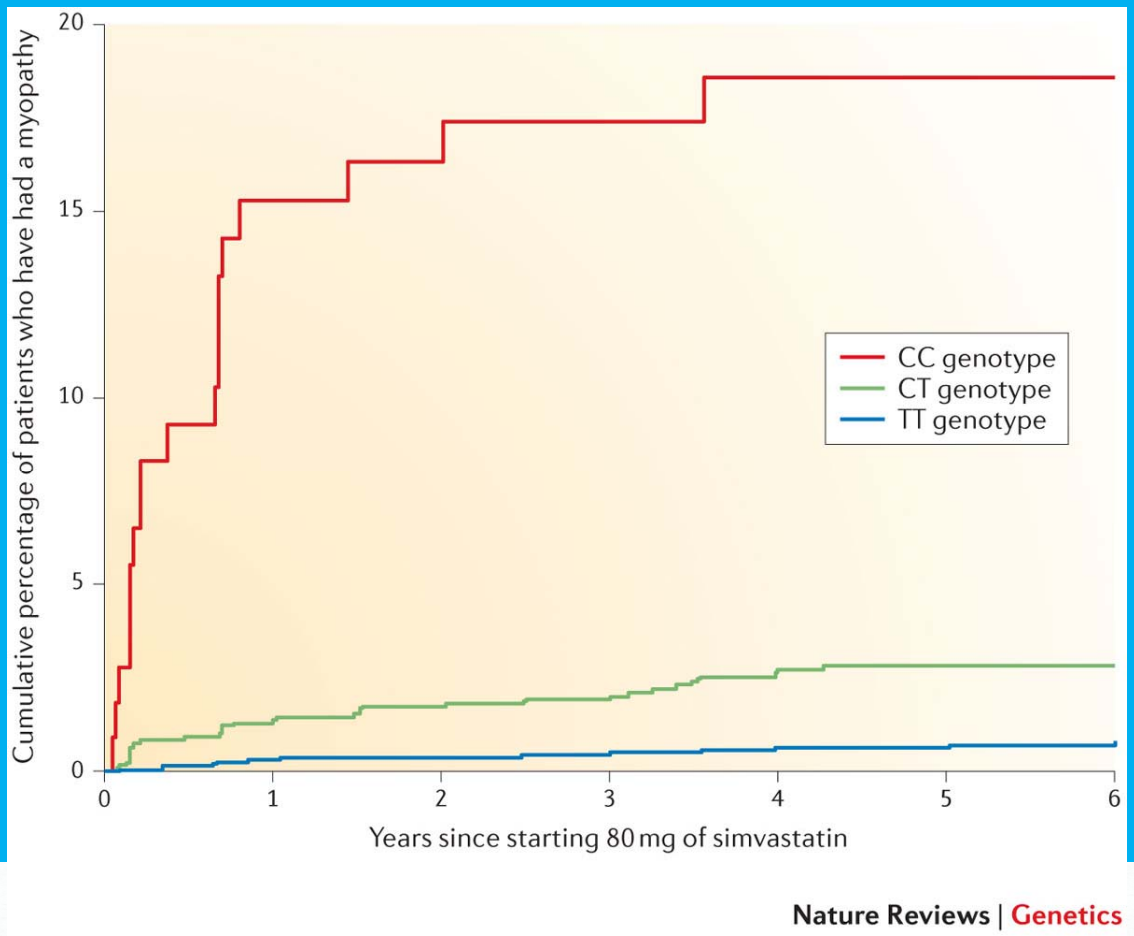
http://www.healthcareitnews.com/news/pop-health-analytics-top-aco-priority

What do visionaries and innovators want to use analytics for?

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

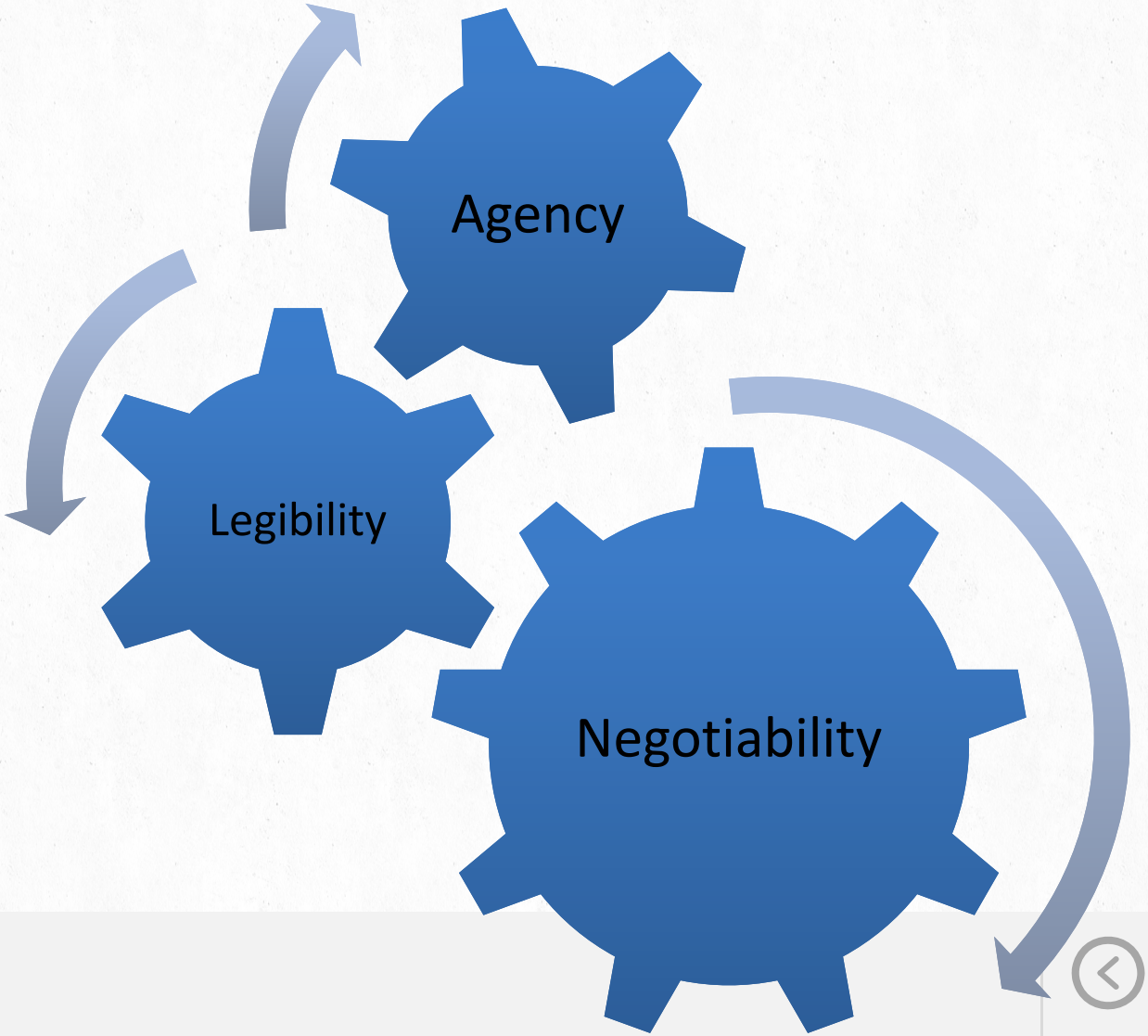


Estimated cumulative risk of myopathy associated with high-dose simvastatin by solute carrier organic anion transporter family member 1B1 (*SLCO1B1*) rs4149056 genotype. The figure is modified from Ref. 71© (2008) Massachusetts Medical Society.

Clinical data major issues

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

Where we need to focus: Human-Data Interaction




Safari File Edit View History Bookmarks Develop Window Stand Help (0:24) Tue 9:25 AM
Tracking Transience v2.0
DL719 JFK SFO
Safari File Edit View History Bookmarks Develop Window Stand Help (0:23) Tue 9:26 AM
Tracking Transience v2.0
Friday, January 9
37.7428, -122.483, ~93.2738
Monday, 07 December 15:27:46 -0500
San Francisco, CA
May 9
Safeway Fuel 0968 Vallejo CA
$14.79