Big Data and the Promise and Pitfalls when Applied to Disease Prevention and Promoting Better Health

Philip E. Bourne Ph.D., FACMI
Associate Director for Data Science
National Institutes of Health
philip.bourne@nih.gov

http://www.slideshare.net/pebourne

Agenda

- What are Big Data anyway?
- What are the implications for healthcare generally?
- What are the implications for NIH specifically?
- Examples of big data applied to disease prevention & promoting better health

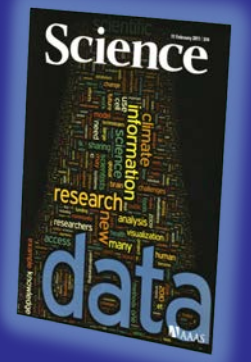
What are Big Data: Quantifying the Problem

- Big Data
  - Total data from NIH-funded research currently estimated at 650 PB*
  - 20 PB of that is in NCBI/NLM (3%) and it is expected to grow by 10 PB this year
- Dark Data
  - Only 12% of data described in published papers is in recognized archives – 88% is dark data^
- Cost
  - 2007-2014: NIH spent ~$1.2Bn extramurally on maintaining data archives

* In 2012 Library of Congress was 3 PB

^ http://www.ncbi.nlm.nih.gov/pubmed/26207759

Big Data in Biomedicine…



This speaks to something more fundamental that more data …

It speaks to new methodologies, new skills, new emphasis, new cultures, new modes of discovery …

NIH

Agenda

- What are Big Data anyway?
- What are the implications for healthcare generally?
- What are the implications for NIH specifically?
- Examples of big data applied to disease prevention & promoting better health

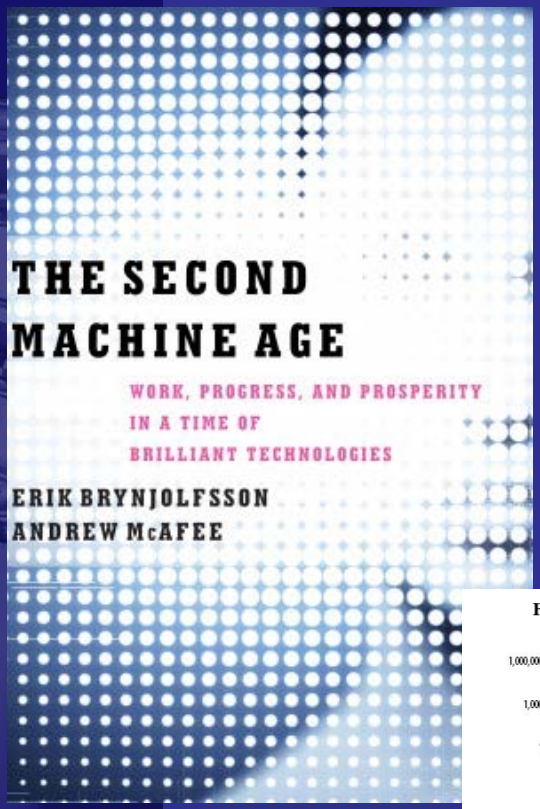
It Follows …

We are entering a period of disruption in biomedical research and we should all be thinking about what this means

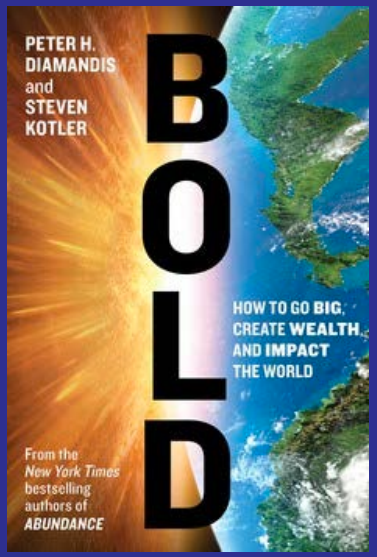
http://i1.wp.com/chisconsult.com/wp-content/uploads/2013/05/disruption-is-a-process.jpg

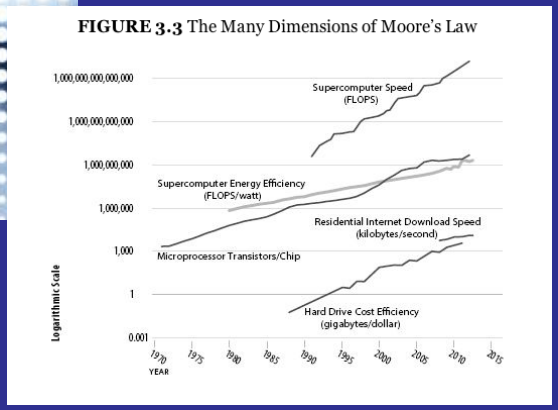
http://cdn2.hubspot.net/hubfs/418817/disruption1.jpg

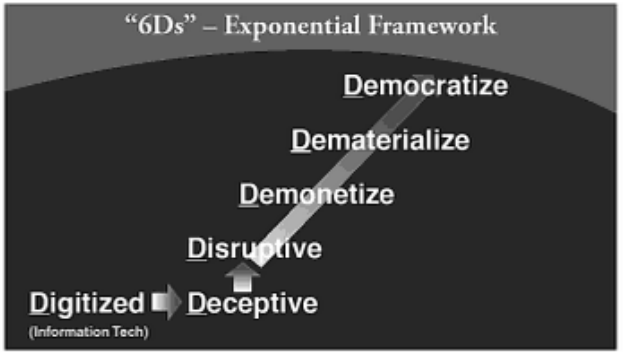
We are at a Point of Deception …



- Evidence:
  - Google car
  - 3D printers
  - Waze
  - Robotics
  - Sensors







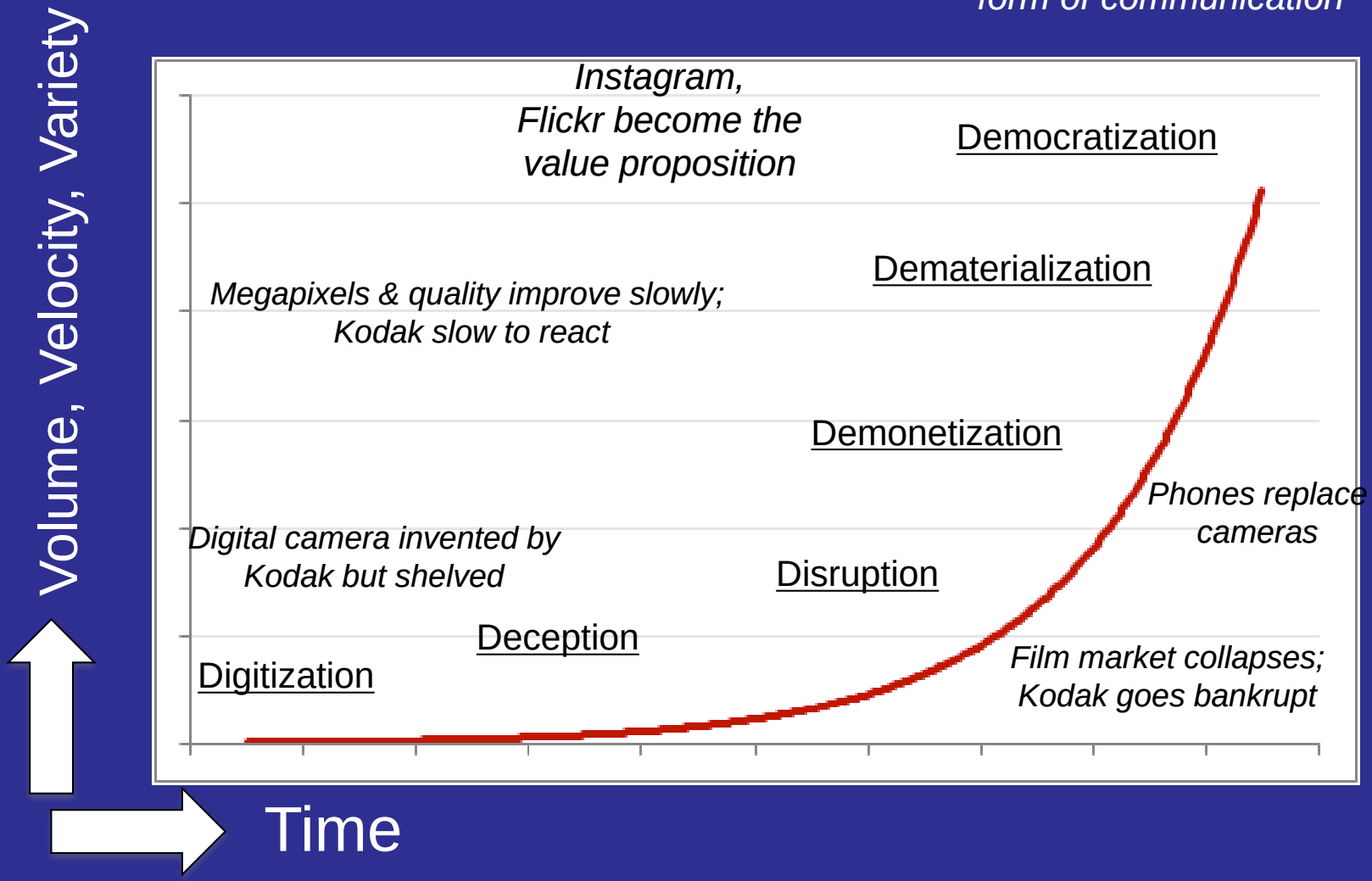
The 6 Ds of Exponentials: Digitalization, Deception, Disruption, Demonetization, Dematerialization, and Democratization

Source: Peter H. Diamandis, www.abundancehub.com

From: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies by Erik Brynjolfsson & Andrew McAfee

Disruption: Example - Photography

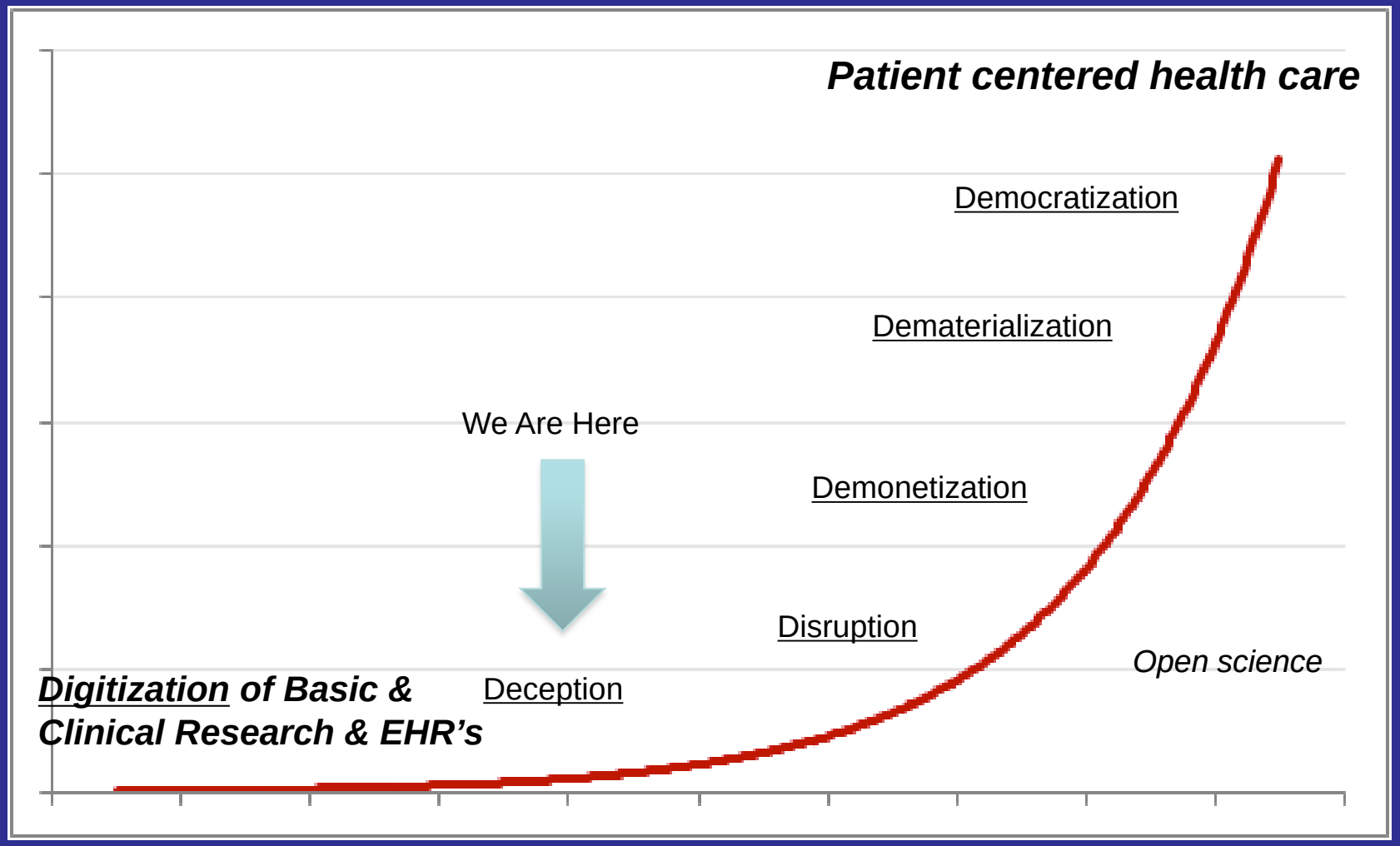
Digital media becomes bona fide form of communication



Agenda

- What are Big Data anyway?
- What are the implications for healthcare generally?
- What are the implications for NIH specifically?
- Examples of big data applied to disease prevention & promoting better health

Disruption: Biomedical Research



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

Archive | Volume 527 | Issue 7576 | Outlook | Article

NATURE | OUTLOOK

Perspective: Sustaining the big-data ecosystem

Philip E. Bourne, Jon R. Lorsch & Eric D. Green

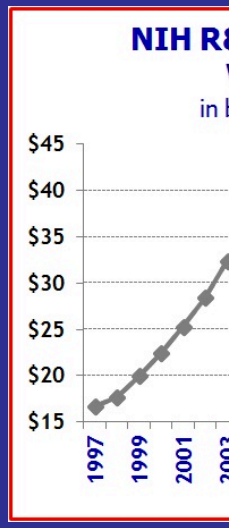
Affiliations | Corresponding author

Nature **527**, S16–S17 (05 November 2015) | doi:10.1038/527S16a
Published online 04 November 2015

Organizing and accessing biomedical big data will require quite different business models, say Philip E. Bourne, Jon R. Lorsch and Eric D. Green.

Subject terms: Genomics · Computational biology and bioinformatics

Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries



.bell1/blog/?p=830

Implications: Reproducibility Changing Value of Scholarship (?)

COMMENT

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 502 > Issue 7470 > Editorial > Article

NATURE | EDITORIAL

Announcement: Launch of an online data journal

09 October 2013

PLOS | BLOGS

Monday, July 05, 2014 | The PLOS ONE Community Blog

Home | STAFF BLOGS | BLOGS NETWORK | COMMUNITY

EveryONE

PLOS | ONE community blog

Home | About EveryONE | Author FAQ | Media | ONE World | Open Access | Resources | Why ONE?

← It's a Mad, Mad, Mad, Mad, but Predictable World: Scaling the Patterns of Ancient Urban Growth

Impending Flood? Hold Onto Your Family! →

PLOS' New Data Policy: Public Access to Data

By Liz Silva

Posted: February 24, 2014

Search EveryONE

Search

Categories

- Aggregators
- Apps
- article-level metrics

PRESS RELEASE 18-Sep-2013

PRESS RELEASE

FOR IMMEDIATE RELEASE

J. Craig Venter Institute Receives $2.4M Grant from National Science Foundation to Develop Arabidopsis Information Portal (AIP)

Implications – New Science

"And that's why we're here today. Because something called precision medicine … gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen."

President Barack Obama

January 30, 2015

Precision Medicine Initiative

- National Research Cohort
  - >1 million U.S. volunteers
  - Numerous existing cohorts (many funded by NIH)
  - New volunteers
- Participants will be centrally involved in design and implementation of the cohort
- They will be able to share genomic data, lifestyle information, biological samples – all linked to their electronic health records

What Are Some General Implications of Such a Future?

- **Open collaborative science** becomes of increasing importance nationally and internationally
- Global **cooperation between funders** will be needed to sustain the emergent digital enterprise
- The **value of data** and associated analytics becomes of increasing value to scholarship
- Opportunities exist to improve the **efficiency** of the research enterprise and hence fund more research
- Current **training** content and modalities will not match supply to demand
- Balancing **accessibility vs security** becomes more important yet more complex

What are the implications of not acting?

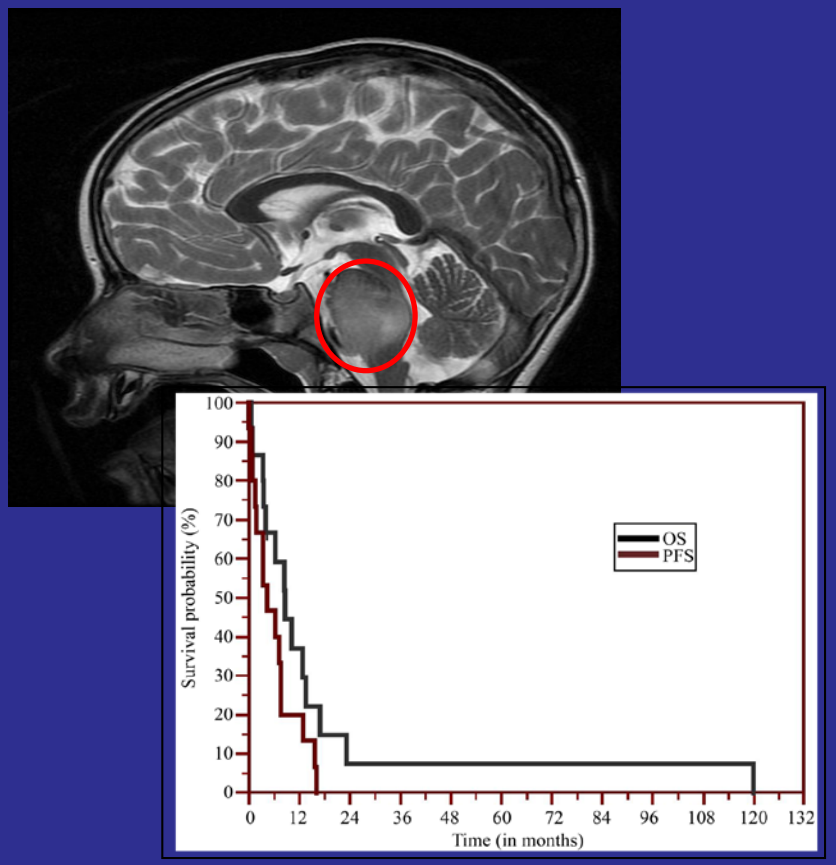
Use Case:

Aggregate integrated data offers the potential for new insights into rare diseases …

As we get more precise every disease becomes a rare disease

Diffuse Intrinsic Pontine Gliomas (DIPG): In need of a new data-driven approach

- Occur 1:100,000 individuals
- Peak incidence 6-8 years of age
- Median survival 9-12 months
- Surgery is not an option
- Chemotherapy ineffective and radiotherapy only transitive

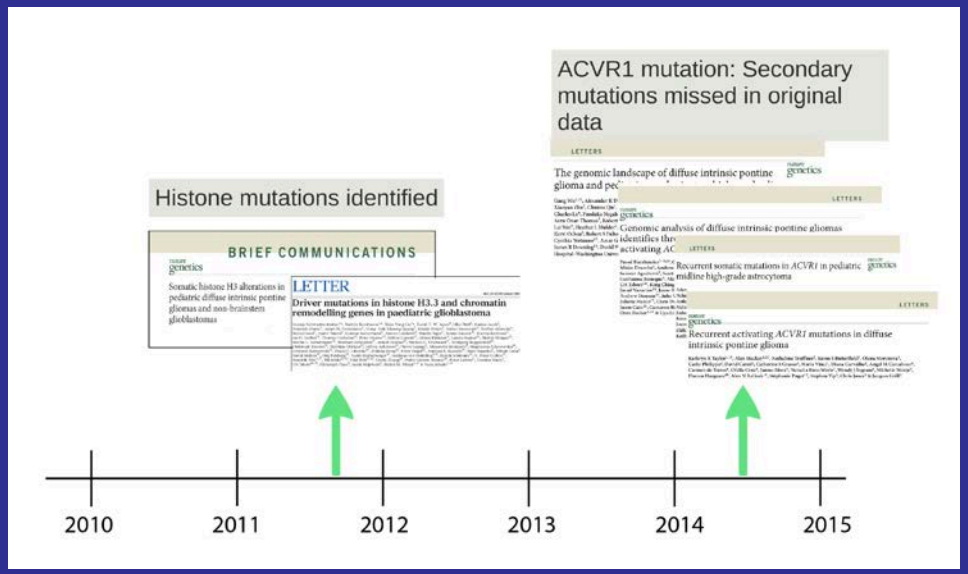


From Adam Resnick

Timeline of Genomic Studies in DIPG

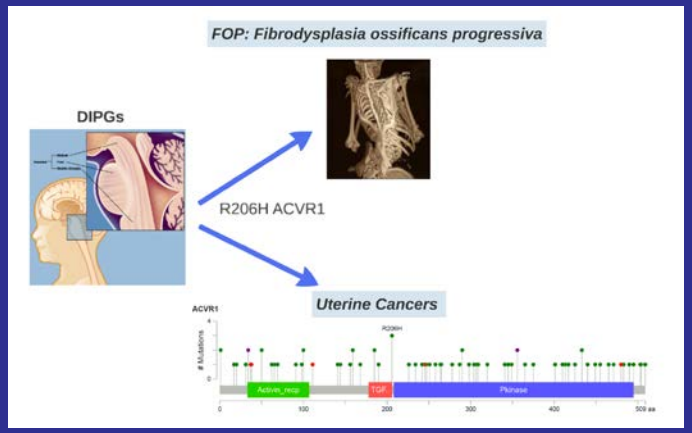
- Landmark studies identify histone mutations as recurrent driver mutations in DIPG ~2012

- Almost 3 years later, in largely the same datasets, but partially expanded, the same two groups and 2 others identify ACVR1 mutations as a secondary, co-ocurring mutation



From Adam Resnick

Hypothesis: The Commons would have revealed ACVR1



- ACVR1 is a targetable kinase
- Inhibition of ACVR1 inhibited tumor progression in vitro
- ~300 DIPG patients a year
- ~60 are predicted to have ACVR1
- If large scale data sets were only integrated with TCGA and/or rare disease data in 2012, ACVR1 mutations would have been identified
- 60 patients/year X 3 years = 180 children's lives (who likely succumbed to the disease during that time) could have been impacted if only data were FAIR

From Adam Resnick

The Commons – The Internet of Data

The Commons offers a path forward to integrate discreet cloud-based initiatives using BD2K developments to make data FAIR*

- Findable
- Accessible
- Interoperable
- Reusable

The internet started as discreet networks that merged - the same could happen with data

* http://www.ncbi.nlm.nih.gov/pubmed/26978244

Examples of Commons Based Initiatives

Program Snapshot

40TB AWS

The Common Fund's **Human Microbiome Project (HMP)** is developing research resources to enable the study of the microbial communities that live in and on our bodies and the roles they [illegible] human health and disease.

5 PB

The NCI Genomic Data Commons

The NCI Genomic Data Commons (GDC) is a unified knowledge base that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.

Cancer is fundamentally a disease of the genome, caused by mutations and other harmful genomic changes that alter its function and contribute to the malignant behavior of cancer cells. Genomic aberrations can influence the aggressiveness of tumors and the response of tumors to particular drugs.

The NCI Genomic Data Commons is housed at the University of Chicago Kenwood Data Center
Credit: University of Chicago

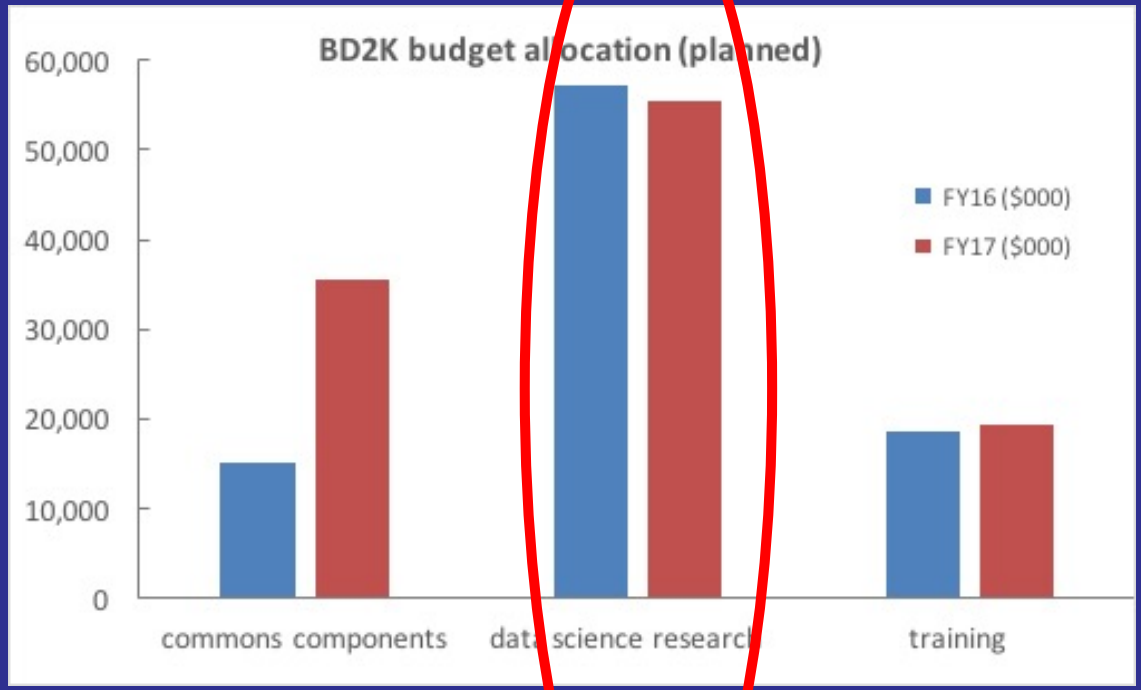
The Role of BD2K

1. Commons

- Resource Indexing
- Standards
- Cloud & HPC
- Sustainability

2. Data Science Research

- Centers
- Software Analysis & Methods

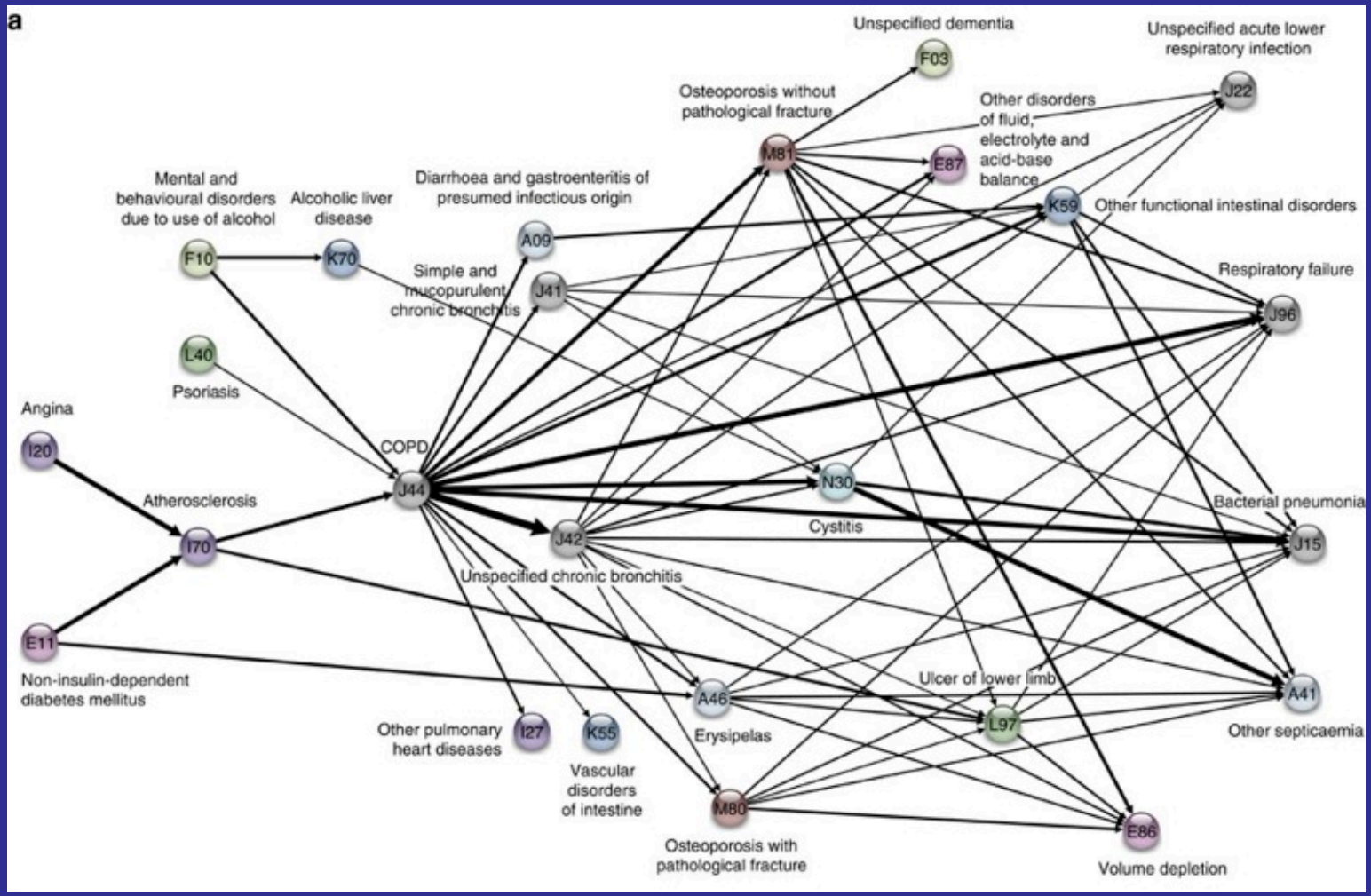


3. Training & Workforce Development

Agenda

- What are Big Data anyway?
- What are the implications for healthcare generally?
- What are the implications for NIH specifically?
- Examples of big data applied to disease prevention & promoting better health

An Example of That Promise: Comorbidity Network for 6.2M Danes Over 14.9 Years



Jensen et al 2014 Nat Comm 5:4022

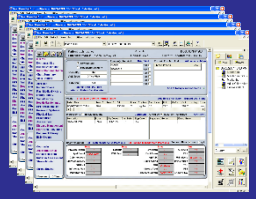
The Center for Predictive Computational Phenotyping

THE UNIVERSITY of WISCONSIN MADISON

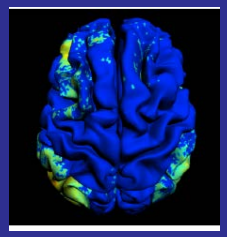
Marshfield Clinic

MORGRIDGE INSTITUTE FOR RESEARCH AT THE UNIVERSITY OF WISCONSIN-MADISON

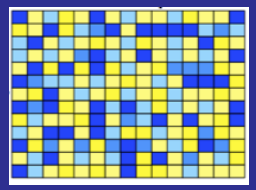
Projects



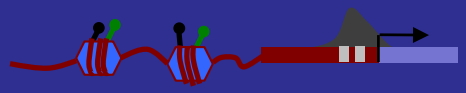
EHR-based phenotyping



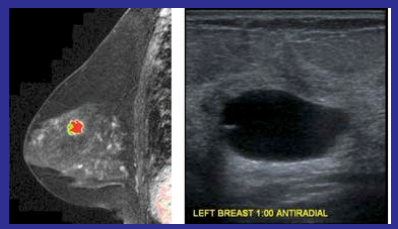
neuroimage-based phenotyping



transcriptome-based phenotyping



epigenome-based phenotyping



phenotype models for breast cancer screening

Labs

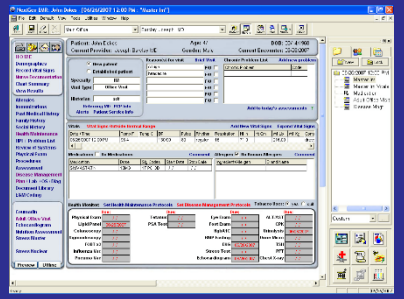
stochastic modeling

low-dimensional representations

value of information

data management

EHR-based phenotyping



genotype
demographics

events in EHR (diagnoses, procedures, medications, labs, etc.)

?

time

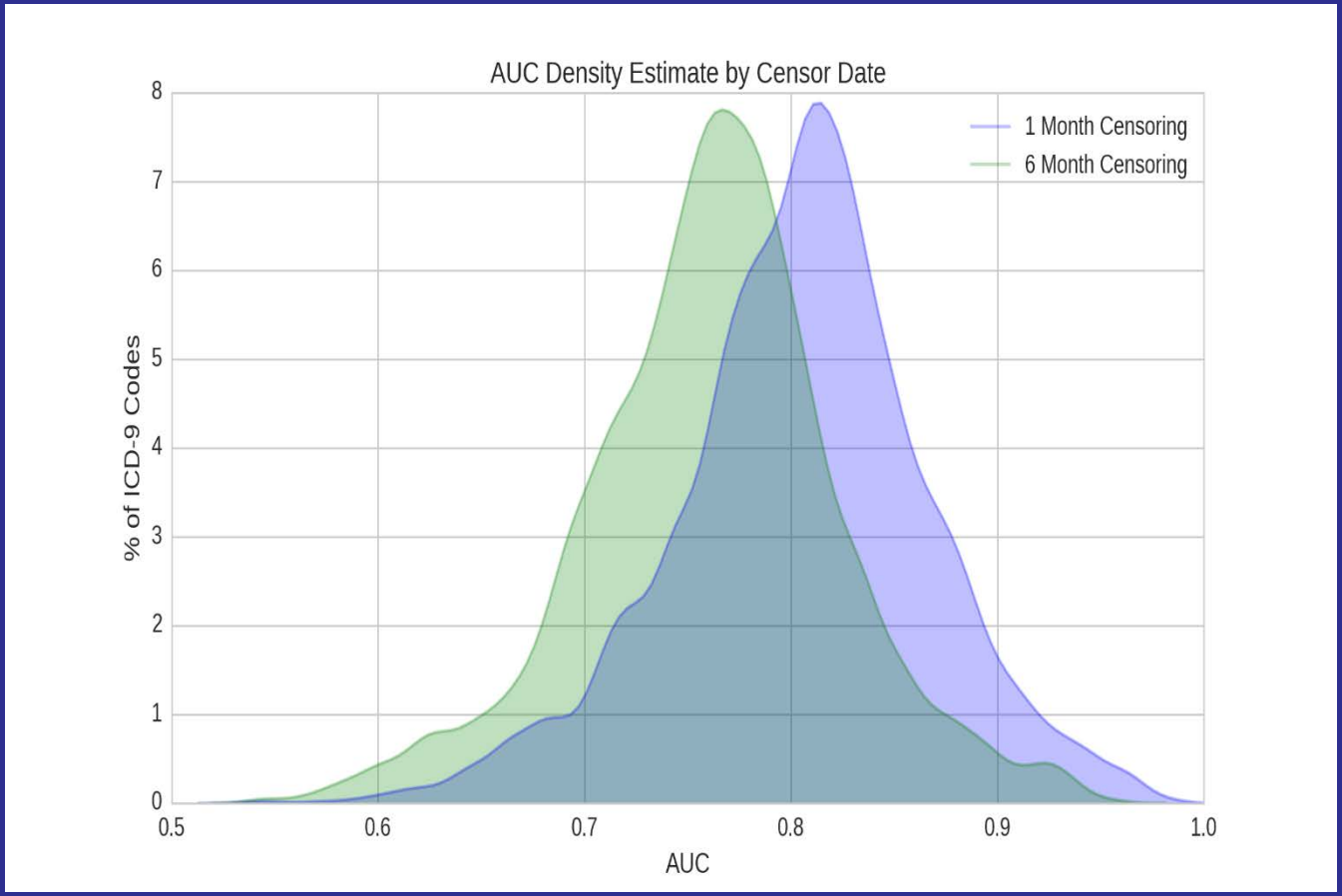
now

retrospective phenotyping: identify subjects who have exhibited a phenotype of interest (i.e. identify cases and controls)

prospective phenotyping: predict a phenotype of interest before it is exhibited

We can predict thousands of diagnoses months in advance of being recorded in an EHR

- ~ 1.5 million subjects from Marshfield Clinic
- models learned for all ICD-9 codes (~3500) for which 500 cases and controls identified



mobilize

Center for Mobility Data Integration to Insight

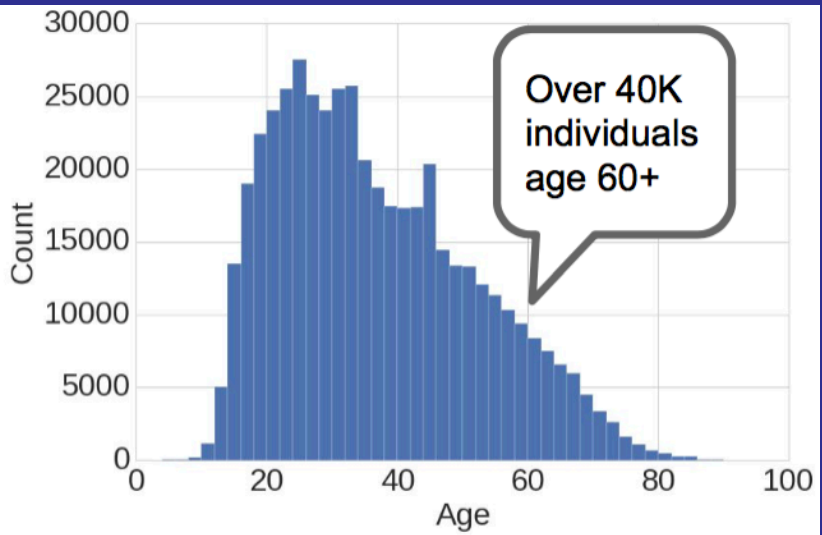
Scott Delp, PhD
Department of Bioengineering
Stanford University

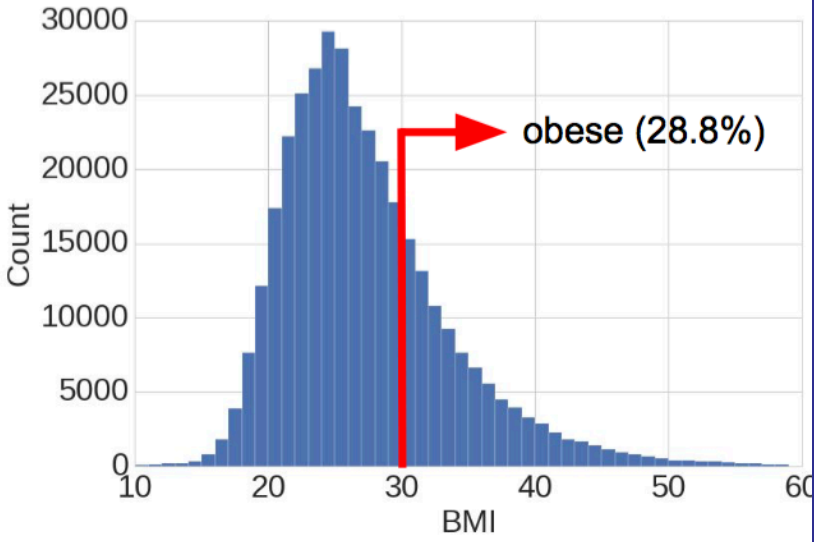
Physical Activity from Personal to Planetary Scale

Physical activity helps prevent heart disease, stroke, diabetes, and weight gain, but inactivity remains a worldwide public health issue.

Azumio
Smartphone App

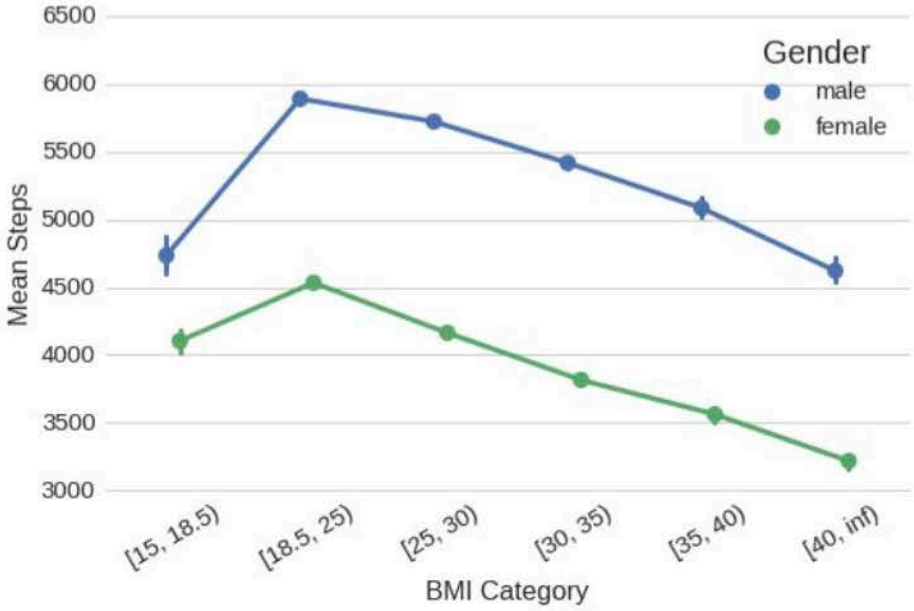
- 2M subjects worldwide and 74M days of activity
- 100B data points, which is ~1000X more than NHANES





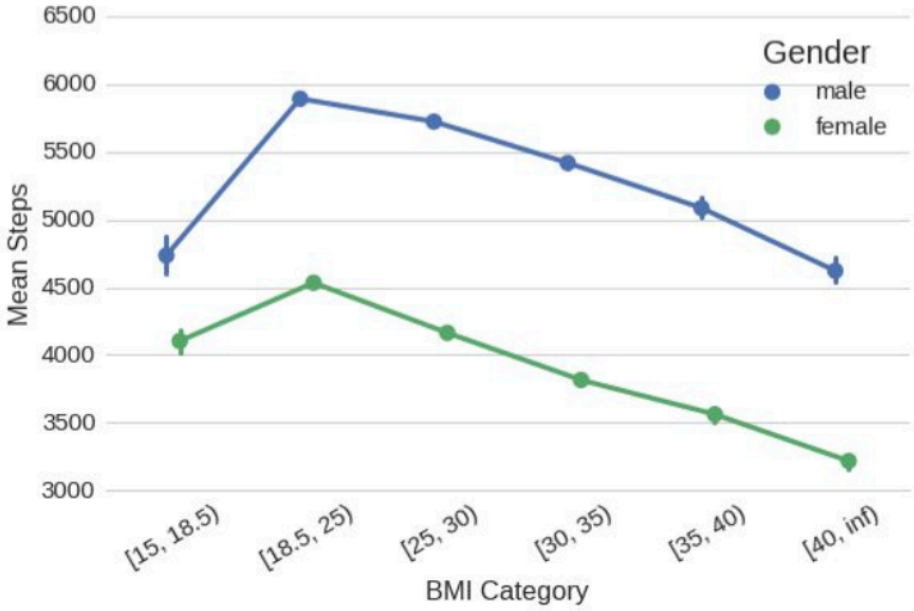
Personal Scale: Physical Activity & BMI

Activity decreases with increasing **BMI** and activity is lower in females (e.g., Tudor-Locke et al., 2010).

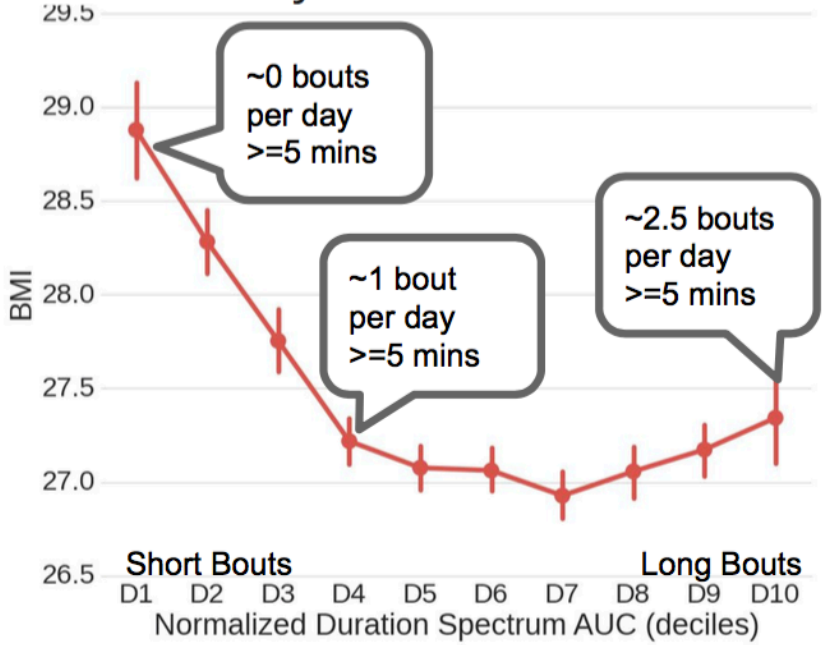


Personal Scale: Physical Activity & BMI

Activity decreases with increasing **BMI** and activity is lower in females (e.g., Tudor-Locke et al., 2010).

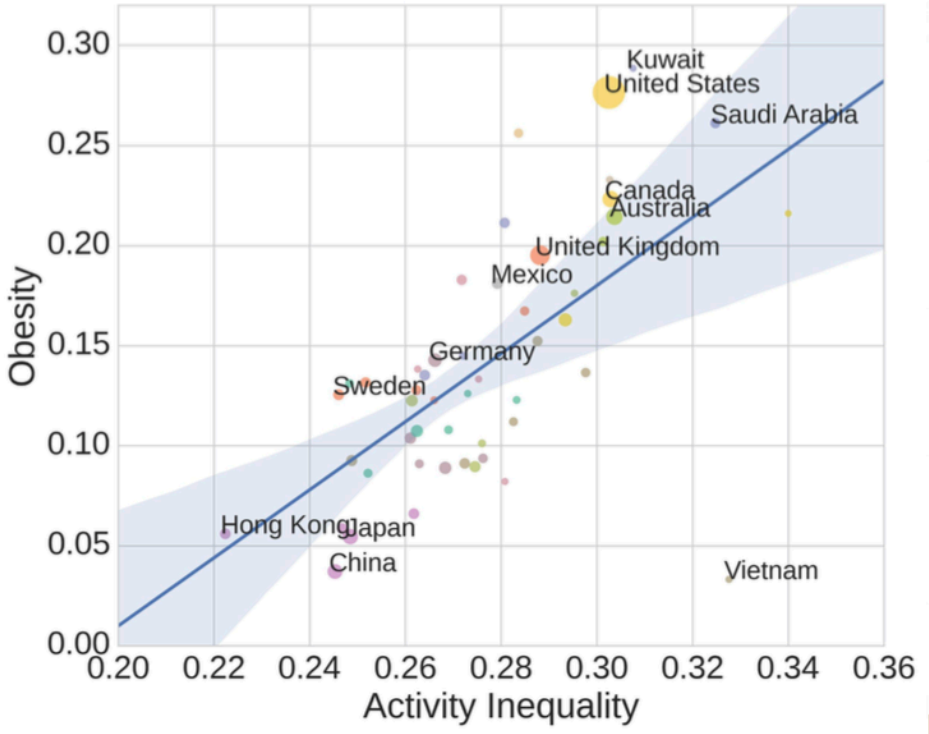


How is **bout duration** related to **BMI**? Control for gender, age, daily wear time, and volume of activity.



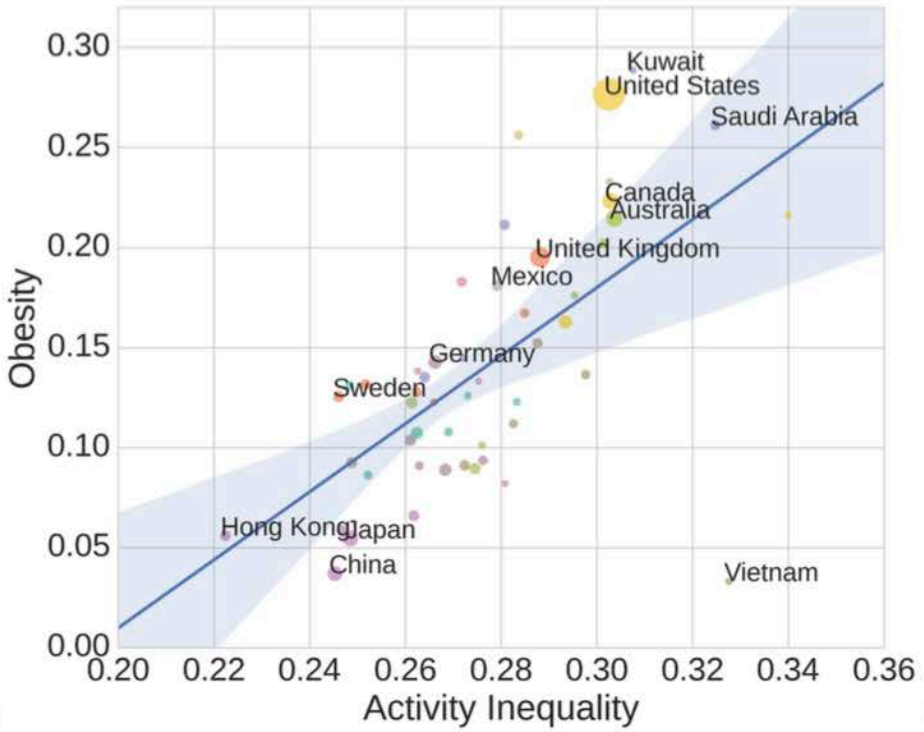
Planetary Scale: Environment, Activity, & Obesity

How is obesity related to **activity inequality** (Gini Coefficient)?

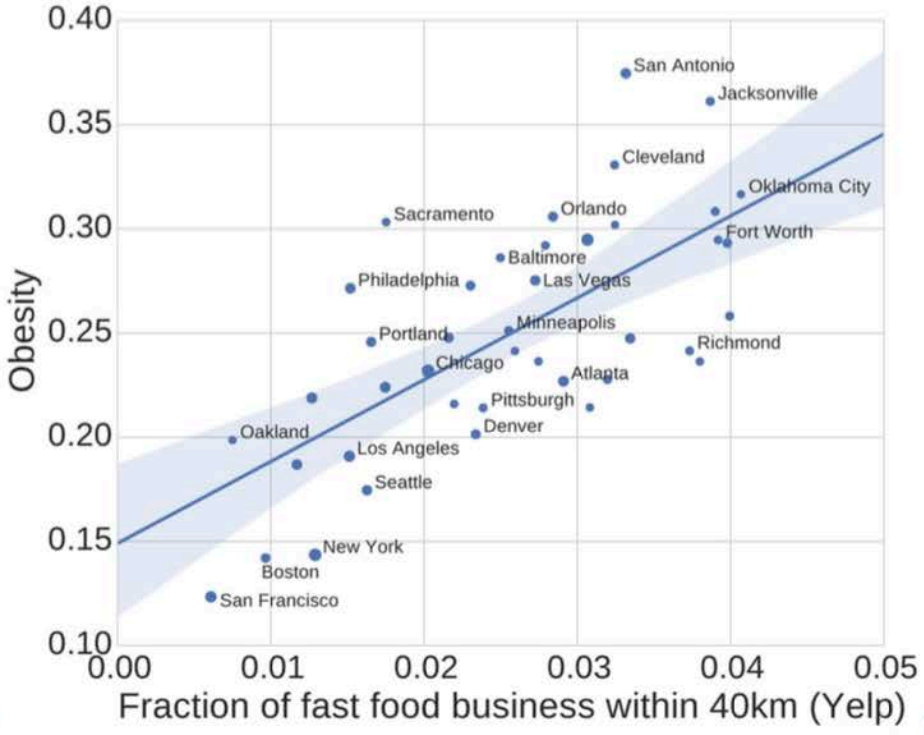


Planetary Scale: Environment, Activity, & Obesity

How is obesity related to **activity inequality** (Gini Coefficient)?



How is obesity related to **fast food access**?



Mobile Sensor Data-to-Knowledge (MD2K)

Mobile Sensors

Smartwatch

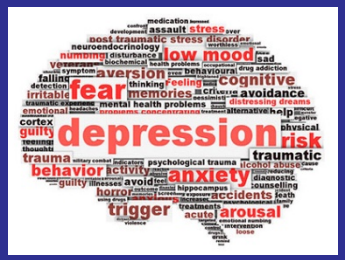
Chestbands

Smart Eyeglasses

Exposures

Behaviors

Outcomes



NIH

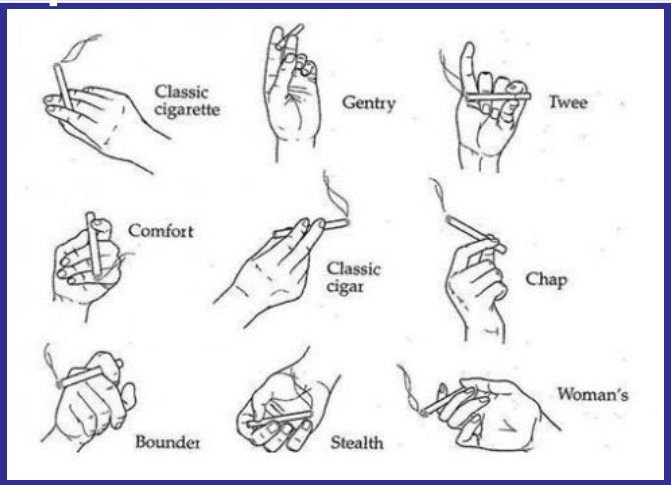
Detecting First Lapses in Smoking Cessation

Saleheen, et. al., ACM UbiComp 2015

Modeling Challenges

1. **Ephemeral (very short duration)**
   - **3~4 sec** for each puff
   - 10,000 breaths in 10 hours
   - 2,000 hand to mouth gestures
   - But, only 6~7 positive instances
   - **Need high recall & low false alarm**
2. **Numerous confounders**
   - Eating, drinking, yawning

Wide person & situation variability



https://www.pinterest.com/pin/567101188907120 75/

Main Results

- Applied on smoking cessation data from 61 smokers
- Detected 28 (out of 32) first lapses
- False alarm rate of 1/6 per day

Key Observations

- First lapse consists of 7 (vs. 15) puffs
- Only 20 (out of 28) reported lapse
- Inaccuracy of self-reported lapse
  - 12 min before to 41 min after lapse
  - Recall inaccuracy even higher

Summary

- Digital Big Data offers unprecedented opportunities
- Those opportunities require a cultural shift – small for some communities large for others – *never easy*
- We are implementing an environment to encourage change
- We would very much like to hear from you opportunities for disease prevention and promoting better health

I not only use all the brains I have, but all I can borrow.

– Woodrow Wilson

ADDS Team

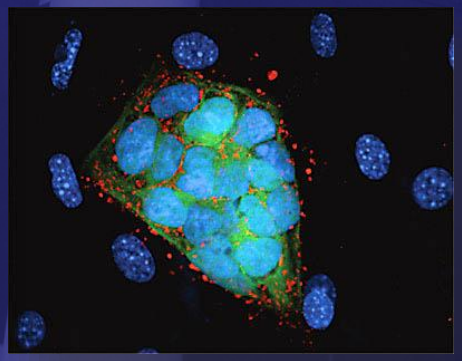
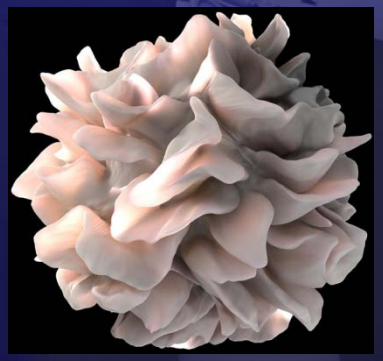
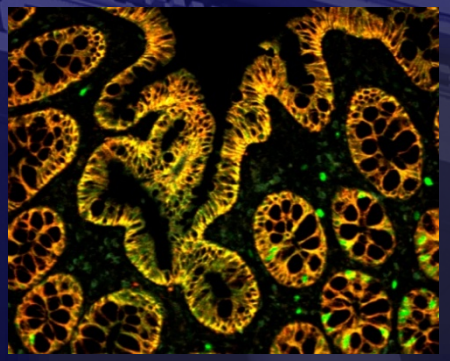
BD2K Representatives




philip.bourne@nih.gov
https://datascience.nih.gov/
http://www.ncbi.nlm.nih.gov/research/staff/bourne/

NIH…

Turning Discovery Into Health

Goal: To strengthen the ability of a diverse biomedical workforce to develop and benefit from data science

Strengthening a diverse biomedical workforce to utilize data science

BD2K funding of Short Courses and Open Educational Resources

Building a diverse workforce in biomedical data science

BD2K Training programs and Individual Career Awards

Discovery of Educational Resources

BD2K Training Coordination Center

Fostering Collaborations

BD2K Training Coordination Center, NSF/NIH IDEAs Lab

Expanding NIH Data Science Workforce Development Center

Local courses, e.g. Software Carpentry