

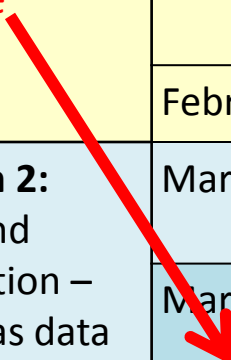
Data and Society

March 11 Meeting

3/11/16

Announcements

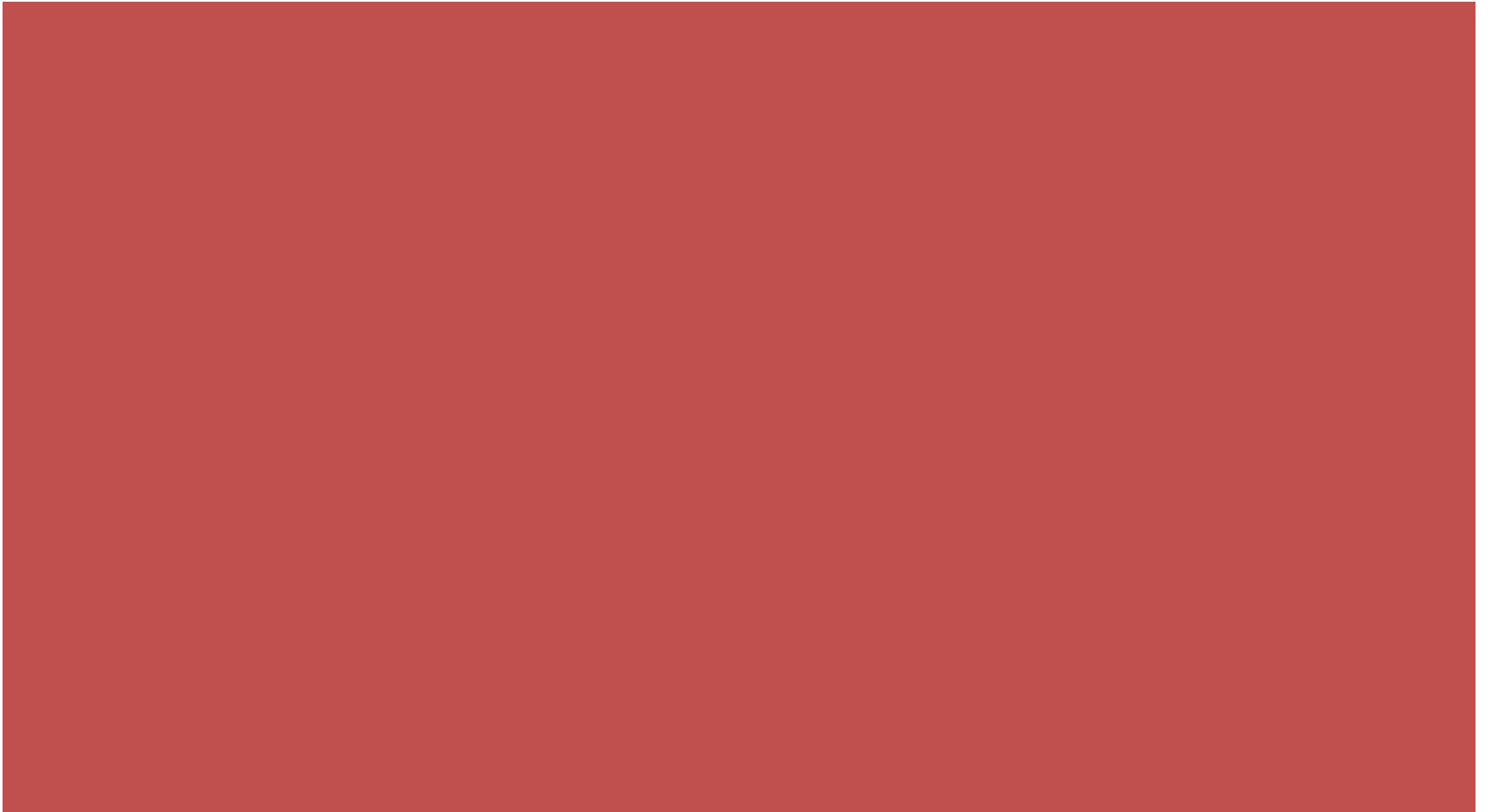
- **Office hours today: 1:30-2:30 (30 minutes later).
Come by if you want to verify scores for your grade so far.**
- **Data Roundtables, etc. announced after break**
- **Paper due April 8 before class**
- **Next week is Spring Break – no class**

Section 1: The Data Ecosystem -- Fundamentals We are here 	January 29	Class introduction; Digital data in the 21 st Century (L1)	Data Roundtable / Fran
	February 5	Data Stewardship and Preservation (L2)	L1 Data Roundtable / 5 students
	February 12	Data-driven Science (L3)	L2 Data Roundtable / 5 students
	February 19	Future infrastructure – Internet of Things (L4)	L3 Data Roundtable / 5 students
	February 26	Section 1 Exam	L4 Data Roundtable / 5 students
Section 2: Data and Innovation – How has data transformed science and society?	March 4	Exams back	Section 1 Data Roundtable / 5 students
	March 11	Data and Health: Phil Bourne guest lecture (L5)	Section 1+2 Data Roundtable / 3 students
	March 18	Spring Break / no class	
	March 25	Data and Entertainment (L6)	Section 2 Data Roundtable / 5 students
	April 1	Big Data Applications (L7)	Privacy Panel / 6 students
Section 3: Data and Community – Social infrastructure for a data-driven world	April 8	Data in the Global Landscape (L8) Section 2 paper due	L7 Data Roundtable / 5 students
	April 15	Digital Rights (L9)	L8 Data Roundtable / 5 students
	April 22	Bulent Yener Guest Lecture, Data Security (L10)	L9 Data Roundtable / 5 students
	April 29	Digital Governance and Ethics (L11)	L10 Data Roundtable / 5 students

Phil Bourne is awesome

- Phil is currently the first Associate Director for Data Science at NIH and manages the Big Data to Knowledge Initiative.
- **Previous positions:**
 - Associate Director of the RCSB Protein Data Bank
 - Founding Editor-in-Chief of PLoS Computational Biology
 - Associate Vice Chancellor of Innovation and Industry Alliances at UCSD
 - President of the International Society for Computational Biology
 - Professor in UCSD Skaggs School of Pharmacology
 - Head of Biology at San Diego Supercomputer Center
 - Writer and Entrepreneur, co-founder of SciVee
 - Etc. etc. etc.

Break



Data Roundtable – March 25

- **“Management Secrets of the Grateful Dead”**, The Atlantic
<http://www.theatlantic.com/magazine/archive/2010/03/management-secrets-of-the-grateful-dead/307918/> (Courtney T.)
- **“How pro teams are using data analytics to draft better players,”** Financial Post,
http://business.financialpost.com/2013/09/03/pro-sports-teams-turning-to-data-analytics-to-fill-seats/?_lsa=88c9-3dab (Sarah S.)
- **“The Real Reason why Google Flu Trends Got Big Data Analytics so Wrong”**,
Forbes, <http://www.forbes.com/sites/teradata/2016/03/04/the-real-reason-why-google-flu-trends-got-big-data-analytics-so-wrong/#509383e31cb1> (Aima M.)
- **“Superman memory crystal lets you store 360TB worth of data”**, CNBC Technology,
February 20, 2016, <http://www.cnbc.com/2016/02/20/superman-memory-crystal-lets-you-store-360tb-worth-of-data.html> (Jordan D.)
- **“Is Keck’s Law Coming to an End?”**, IEEE Spectrum, January 26, 2016,
<http://spectrum.ieee.org/semiconductors/optoelectronics/is-kecks-law-coming-to-an-end> (Evan F.)

April 1 Privacy Panel

- **Side A: Everything is private (complete control of your generated data and approval to use/share data about you) [Ian B., Craig D.]**
 - **Side B: Some things are private (you control your generated data and others may control data about you) [Amreen A., Amelia G-B]**
 - **Side C: Fewer things are private (you control some of your generated data and others may control some of your generated data and data about you) [Caitlin C., Rob R.]**
 - Each team must prepare clear argument that supports it's side's view and describe the repercussions to: **applications, infrastructure and policy/regulations**
 - **Each team member must present and answer questions from the class.** Joint team score of up to 5 given to each team member.
 - **Each team member must write an *individual* 3-4 page summary of their side's perspective** on applications, infrastructure, policy/regulations. Score of up to 5 given. Reviews are expected to be individual.
- Format:
 - 10 minute presentation from Side A team
 - 10 minute Q&A/rebuttal
 - 10 minute presentation from Side B team
 - 10 minute Q&A/rebuttal
 - 10 minute presentation from Side C team
 - 10 minute Q&A/rebuttal
 - Closing statements from each team: 5 minutes per team
 - Class "vote" on most popular perspective

Data Roundtable – April 8

- **“The Shazam Effect”**, The Atlantic,
http://www.theatlantic.com/magazine/archive/2014/12/the-shazam-effect/382237/?single_page=true (Arun V.)
- **“Six Provocations for Big Data”**, Oxford Internet Institute
Network Conference,
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=192643
[1](#) (Kit H.)
- Will provide 3 more articles on March 25

Today: **Data Roundtable**

- **“The Spider that crawls the Dark Web looking for stolen data”**, The Atlantic, February 22, 2016,
<http://www.theatlantic.com/technology/archive/2016/02/the-spider-that-crawls-the-dark-web-looking-for-stolen-data/470220/> (Ethan B.)
- **“M-health: Health and appiness”**, The Economist, February 1, 2014,
<http://www.economist.com/news/business/21595461-those-pouring-money-health-related-mobile-gadgets-and-apps-believe-they-can-work> (Ian B.)
- **“I had my DNA picture taken, with varying results”**, The New York Times, December 30, 2013,
http://www.nytimes.com/2013/12/31/science/i-had-my-dna-picture-taken-with-varying-results.html?pagewanted=all&_r=0 (James B.)

Data & 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>

RPI March 11, 2016



A Few Vignettes to Get the Conversation Going



PROTEIN

DECK NAMES FOR PROTEINS

DECK
NAME

CHYM01

CPASE01

CYTB501

HSDEH01

HSMEH01

INSUL01

IDH01

LDH02

LAMP1

MY0GL01

PAPAIN01

RNASES01

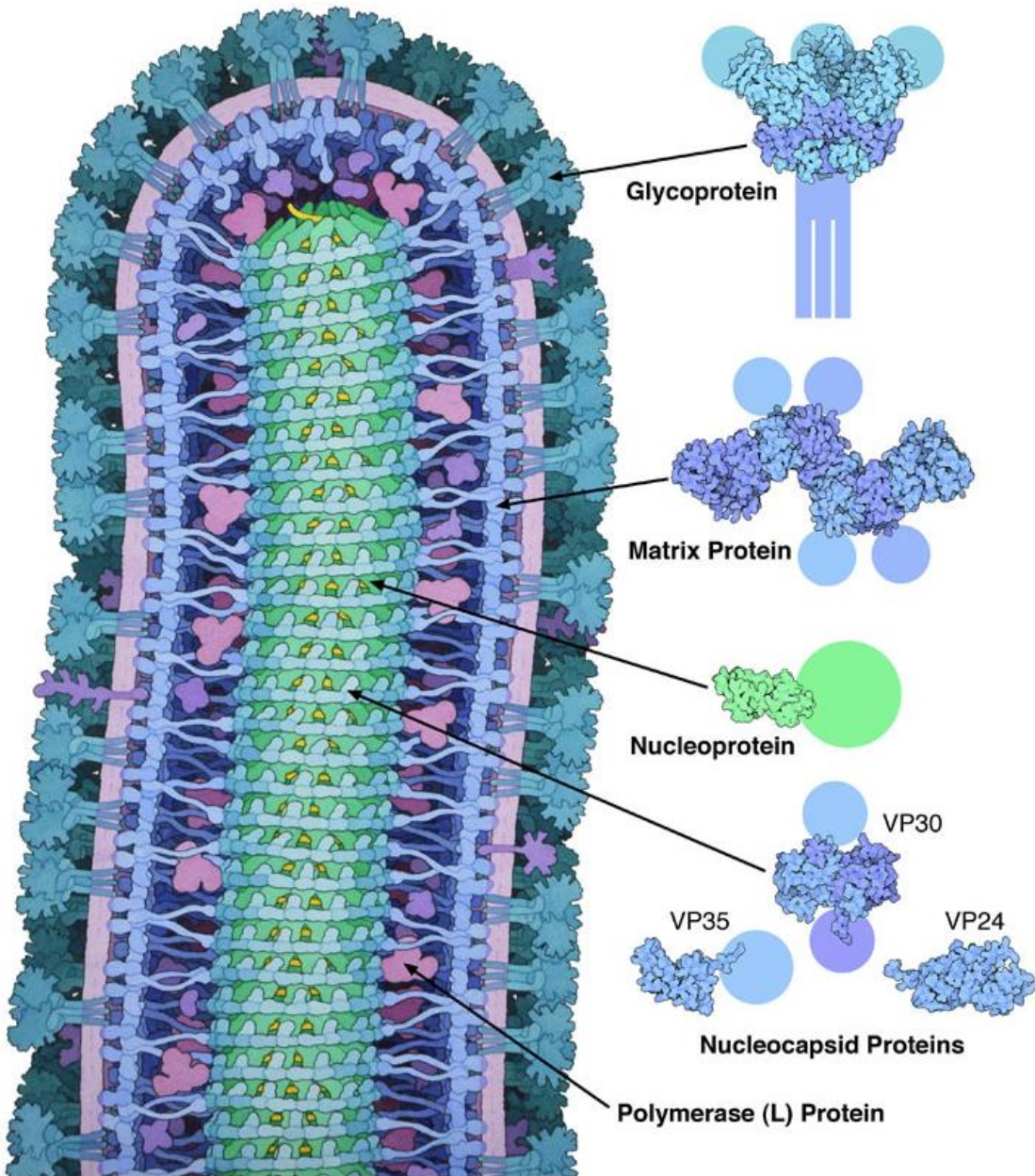
RUBY01

STAPHN01

SUBTL

TRINO

TRINO



Download Image

CLOSE X

#1 Data Sharing

2003
Finished
version of
human
genome
sequence
completed

HGP ends with
all goals achieved

the Human Genome

The Landscape of the Human Genome

1. **Map** Scale in megabases. 2. **Gaps** Red bars indicate gaps for which no large insert clone amenable to sequencing and assembly exists for screening genomic libraries to at least 30X depth (see text of accompanying paper). 3. **Cytosine bands** Grey bars show approximate positions of Gamma irradiated chromosome bands at 500 band resolution. Centromeres are in dark red. Marking heterochromatin, 20, 15, 20, 10, and the heterochromatin block on chromosome 18 is shown. It is expected to exist that centromeric features are only fixed or approximately repeat to sequence-based features. 4. **Systems** Conserved synteny with the mouse genome. 1000 conserved segments of 100,000 bases are color-coded according to the corresponding mouse chromosome. Relative orientation in mouse is indicated by + (matches the assembled mouse sequence) or - (inverted in comparison). Centromeres also indicate the strand of the syntenic match. Based on the Feb. 2003 version of the mouse genome. 5. **Repeats** Long-range repeats of 50,000 bp or greater are clustered in grey. 60-100% in red. Centromeric repeats are green. 6. **GC%** Percentages of bases in a 20,000 base window that are G or C. Scale ranges from 22% (62N) to 28%. 7. **Mouse align** Alignment to mouse. The blue line shows the percentage (0% to 85%) of human sequence within 20 kb blocks that can be aligned to mouse sequence, using anchoring by conserved syntenic regions and gap penalties. The red line shows the base identity (0% to 70%) within the regions and the alignment score in 20 kb blocks. The alignment was done with BLAST2 and a base-to-genome file. (Schwarz et al. Genome Res. 2003 Jan 13;13(1):77-87). 8. **SNP density** SNP density and mutation rate. The purple line shows density of single nucleotide polymorphisms based on 100Kb discovered from random sequence reads, divided by the number of bases from random reads that have had sufficiently high neighborhood quality scores to assess. This measure of heterozygosity is calculated in 1 Mb windows that overlap by 0.5 Mb, and is plotted on a range from 5.0000 to 0.0010. The green line shows mutation rate estimated by the 85% of reads from aligned human reads at sites in haploids that provide the primary-order divergence, calculated over 1 Mb windows. Units are substitutions per site; the scale ranges from 0.2 to 0. 9. **SNP islands** The most common dispersed repetitive elements in the human genome. SNPs are in red. L1s are in blue, both calculated in 100 kb windows. 10. **exDNA** Exons that do not code for protein - e.g. 5'UTRs, 3'UTRs and introns are in green. RFLP positions are in orange. 11. **CpG islands** Each grey bar represents a segment of 200 bases or more with CpG dinucleotide density significantly higher than the genome as a whole. The 20% percentage of CpG dinucleotides in the genome is used as a threshold. 12. **Y-chromosome** The location of Y-chromosome bands. 13. **ESTs** Black ticks show the location of expressed sequence tags (ESTs) with at least one read, aligned against genomic DNA. 14. **Gene tracks** The gene models for RefSeq are indicated in blue. The 2000 reference gene predictions from Ensembl are in dark red (Ensembl: Acc. Nos. 20,1448-70 (2004)). 15. **Gene names** Genes are named in blue, using HUGO nomenclature committee abbreviations. Red indicates a known disease gene. 16. **Indels** Indels are indicated in blue. The 2000 reference gene predictions from Ensembl are in dark red (Ensembl: Acc. Nos. 20,1448-70 (2004)).

Credits: Data: Ewan Birney, Jim Kent, Krishna M. Raskin, Jim Mullikin, Darryl Thomas, Robert Baerbach, Derrick Darryl Lapp



nature

1996

First human gene map established

Pilot projects for human genome sequencing begin in U.S.

First archaeal genome sequenced

Yeast (*S. cerevisiae*) genome sequenced



HGP's mouse genetic mapping goal achieved



Bermuda principles for rapid and open data release established



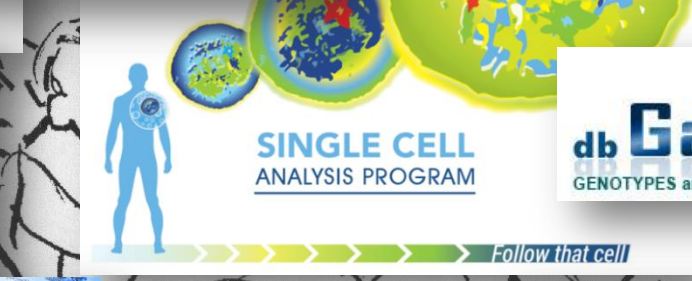
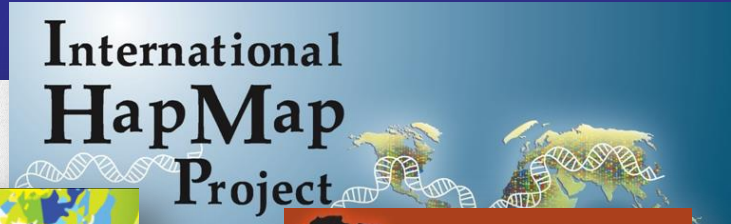
HUMAN GENOMIC SEQUENCE GENERATED BY LARGE SCALE CENTRES

RELEASE

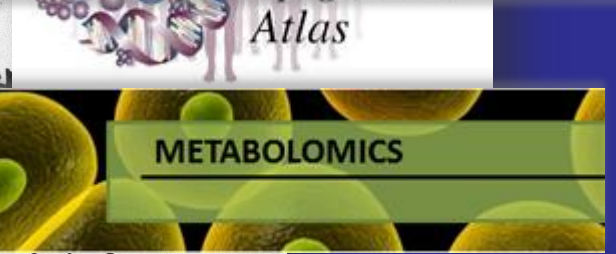
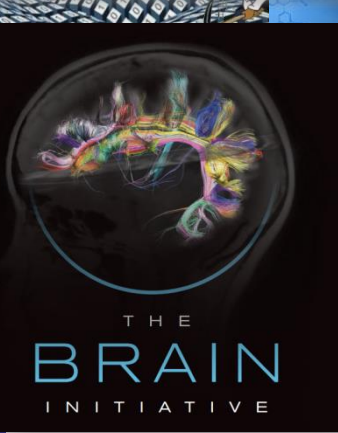
- Automatic release of sequence assemblies >1kb (preferably daily)
- Immediate submission of finished annotated sequence
- ~~_____~~ and in the public domain
- Aim to have all sequence freely available for both research and development, in order to maximise its benefit to society.

POLICY

- The funding agencies are urged to foster these policies



Data Sharing: An Essential Component



The Story of Meredith

STREAM YOUR EVENT | PARTNERS | SPEAKERS CART ITEMS: 0 | CHECKOUT

FORA.tv
CONFERENCE AND EVENT VIDEO

[Join Now](#) or [Log In](#) [Q](#) [Twitter](#) [Facebook](#)

PAY-PER-VIEW BUSINESS ENVIRONMENT POLITICS SCIENCE TECHNOLOGY CULTURE


SPACE | EVOLUTION | PHYSICS | SOCIAL SCIENCES | NATURAL SCIENCES | DNA | PSYCHOLOGY | BIOTECH | MEDICINE | ANTHROPOLOGY | ASTRONOMY

WATCH LIVE
117 hrs 43 mins 13 secs

THE U.S. HAS NO DOG IN THE FIGHT IN SYRIA presented by *intelligence²* DEBATES >>

Congress Unplugged - Phil Bourne
WATCH FULL VIDEO
More from this conference: **Sage Bionetworks Commons Congress 2012**
More videos from this partner: **Sage Bionetworks**

3rd Sage Bionetworks Commons Congress
April 20-21, 2012



sas | THE POWER TO KNOW.

ANALYTICS

Turn what they say into why they stay.

CLICK FOR PAPER ON BUILDING A MARKETING ANALYTICS FRAMEWORK

http://fora.tv/2012/04/20/Congress_Unplugged_Phil_Bourne



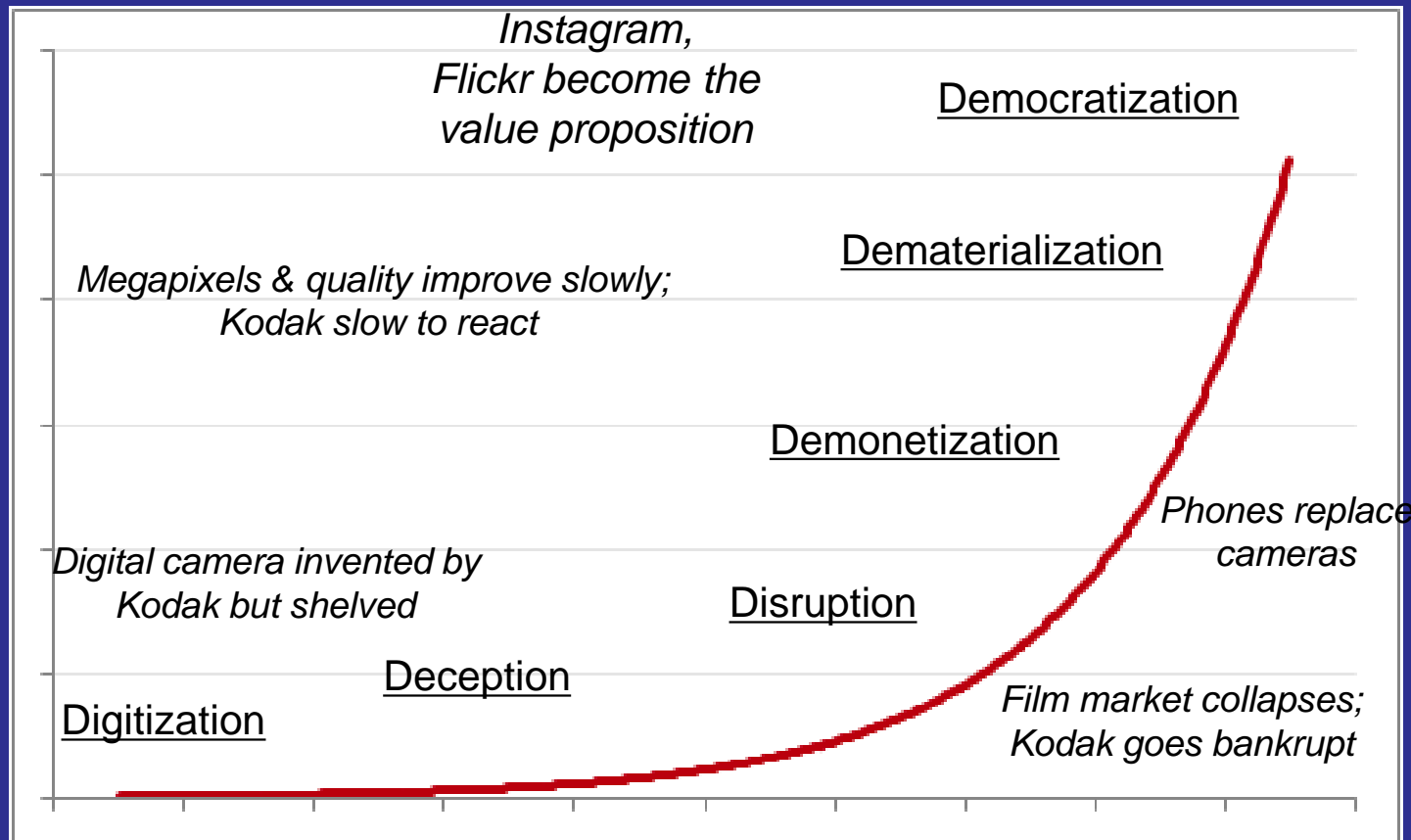
#2

**Will Data Disrupt the Health Care
Market?**

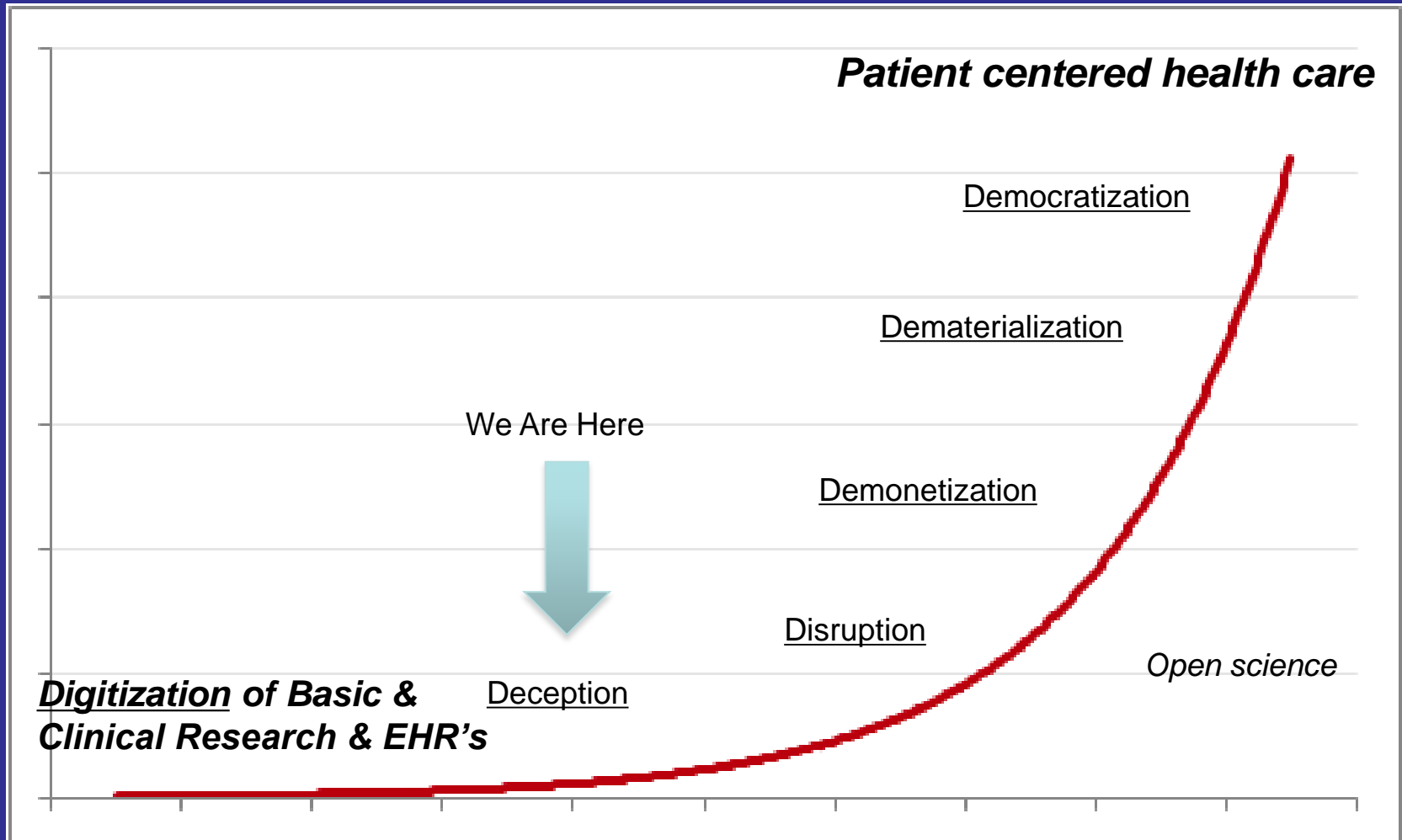
True Free Market - Photography

Digital media becomes bona fide form of communication

Volume, Velocity, Variety
↑
Time →



False Market - Biomedical Research?



#3 Sustaining the System is a Problem

NIH R&D Under BCA Caps With and Without Sequestration

\$45
\$40
\$35
\$30
\$25
\$20
\$15
1997

nature

International weekly journal of science

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

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



PDF



Citation



Reprints



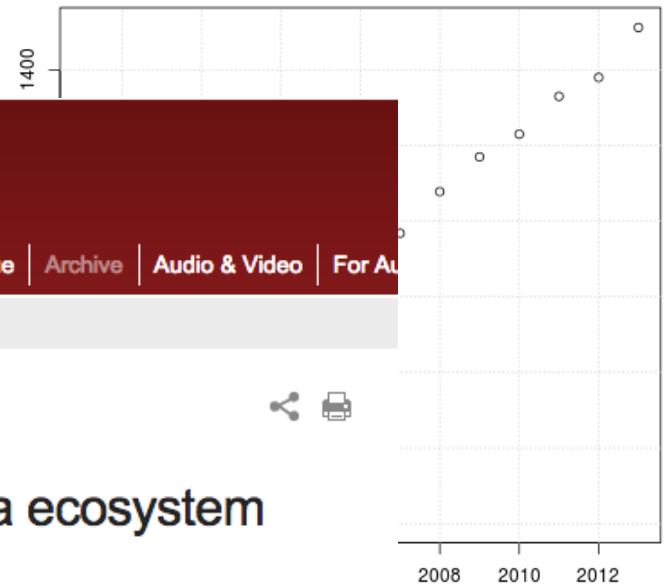
Rights & permissions



Article metrics

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

Growth of Biological Databases



cl.ac.uk/m.j.bell1/blog/?p=830

#4 Reproducibility & Changing Value of Scholarship

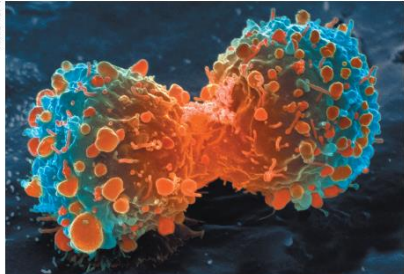
COMMENT

SHIM SHIMURA Shift expertise to track mutations where they emerge p.104

LARIJUNYAN Past climates give valuable clues to future warming p.107

WERNER Past climates? Just letter tracked using Google p.108

WYLLIE Why? Why? and an elusive stress hormone p.142



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

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.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex of diseases. Although we in the cancer field hoped that this would lead to more effective drug, historically our ability to translate this research to clinical trials in oncology have the highest failure rates compared with other therapeutic areas. Given the high stakes need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will reach the bedside. However, the low success rate of drugs that reach the bedside in oncology must mean that our approach to translating discovery research into practical success and impact. Many factors are responsible for this failure rate, notwithstanding the inherently difficult nature of this disease. Clearly, the limitations of preclinical to

nature International weekly journal of science

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

Archive | Volume 502 | Issue 7470 | Editorial | Article

NATURE | EDITORIAL

Announcement: Launch of an online data journal

09 October 2013

PLOS | BLOGS

day, 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! →

Search EveryONE

Categories

- Aggregators
- Apps
- article-level metrics

PLOS' New Data Policy: Public Access to Data

By Liz Silva
Posted: February 24, 2014

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)

#5 New Science



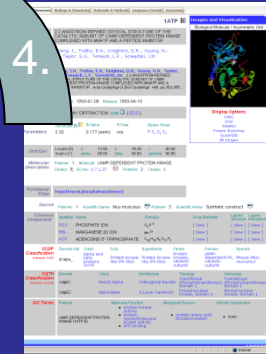
“And that’s why we’re here today. Because something called precision medicine ... gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen.”

President Barack Obama
January 30, 2015

#6 What Communication Should Be

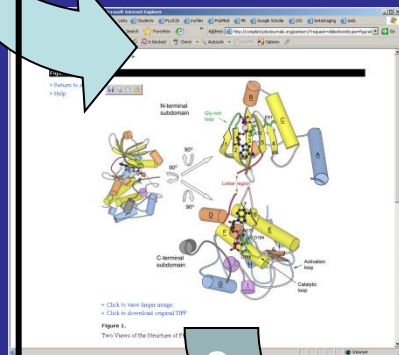
0. Full text of PLoS papers stored in a database

4. The composite view has links to pertinent blocks of literature text and back to the PDB

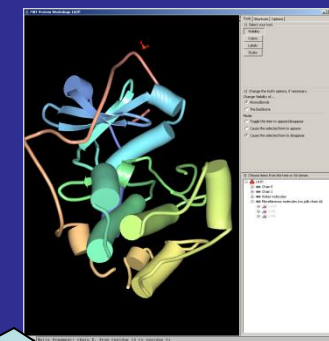


1.

1. A link brings up figures from the paper



2.



2. Clicking the paper figure retrieves data from the PDB which is analyzed

3.

3. A composite view of journal and database content results

Is a database really different than a biological journal?

PloS Comp Biol
2005 1(3) e34

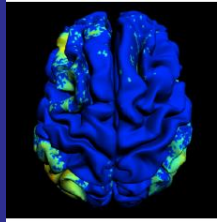
#7 Some Big Data Examples



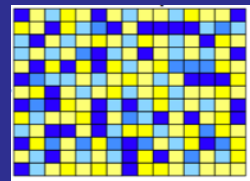
The Center for Predictive Computational Phenotyping



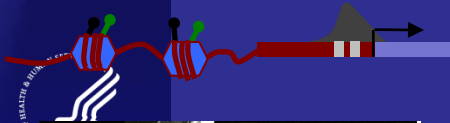
EHR-based phenotyping



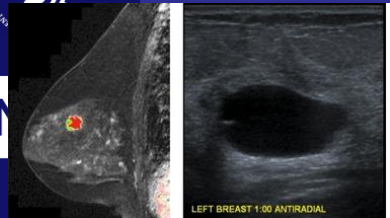
neuroimage-based phenotyping



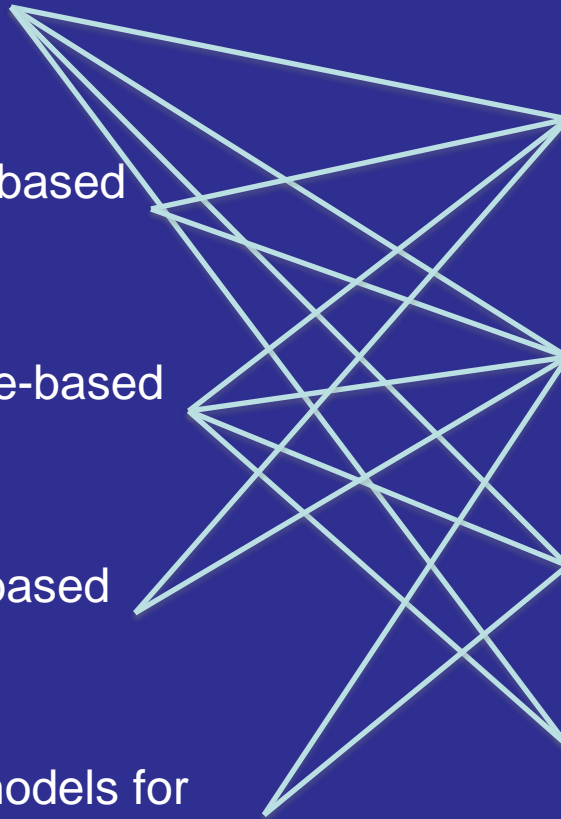
transcriptome-based phenotyping



epigenome-based phenotyping



phenotype models for breast cancer screening



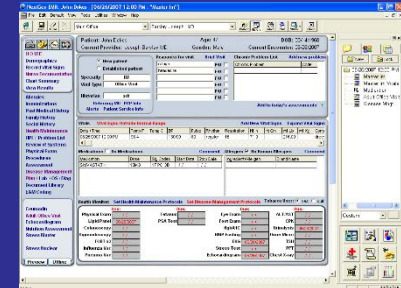
Projects

Labs

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

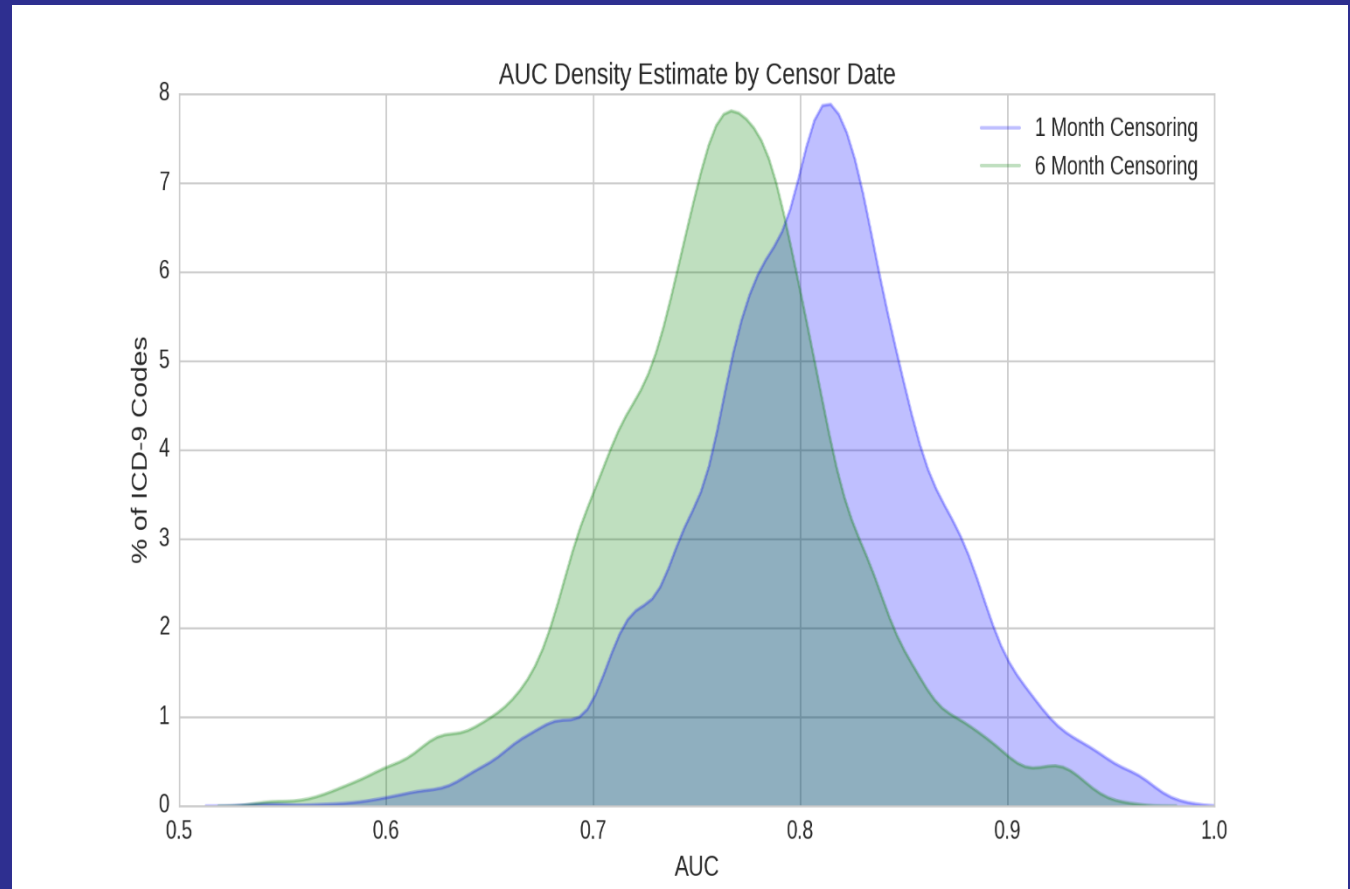
interest (i.e. identify cases and
controls)

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



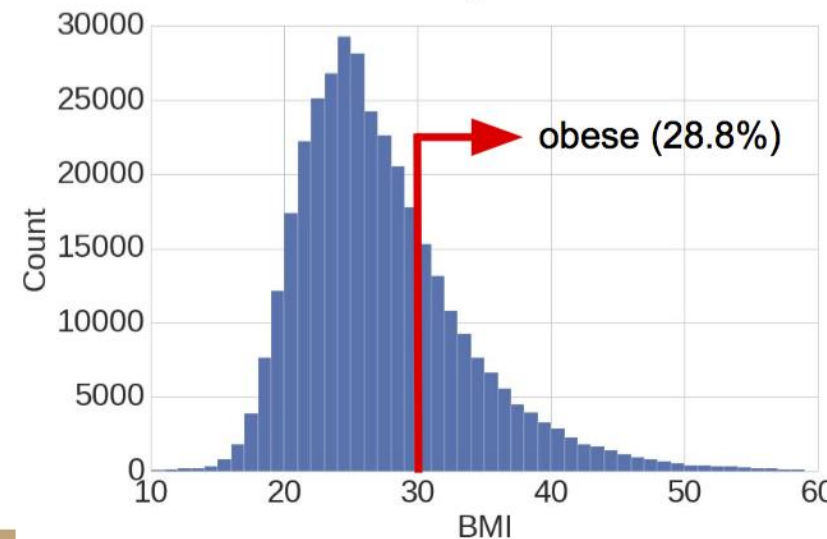
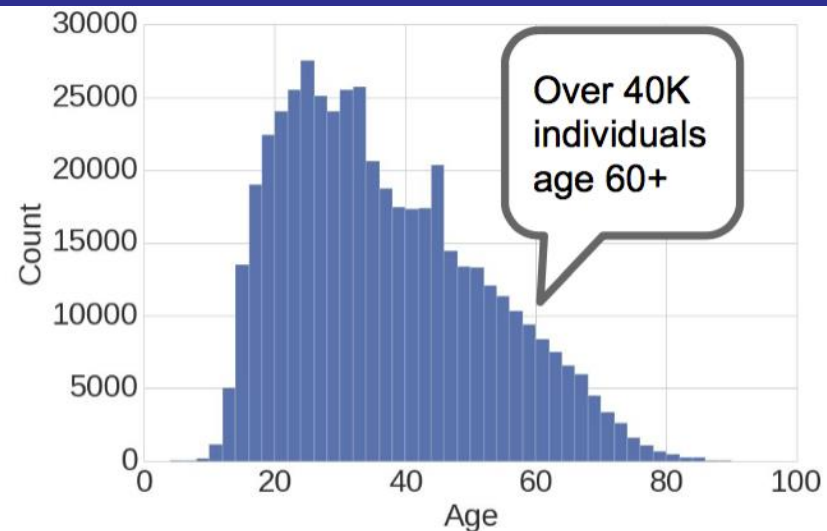
Physical Activity from Personal to Planetary Scale

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



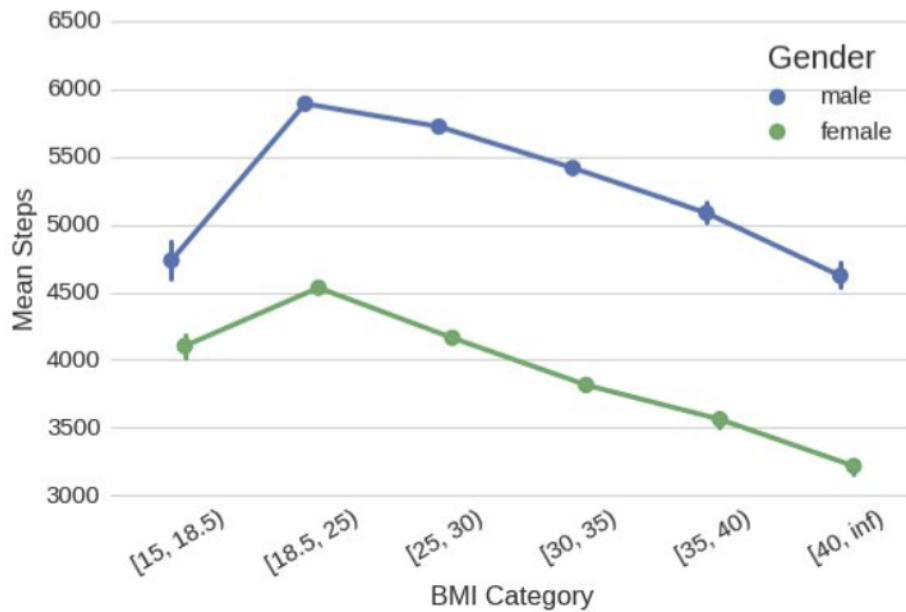
Azumio Smartphone App

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



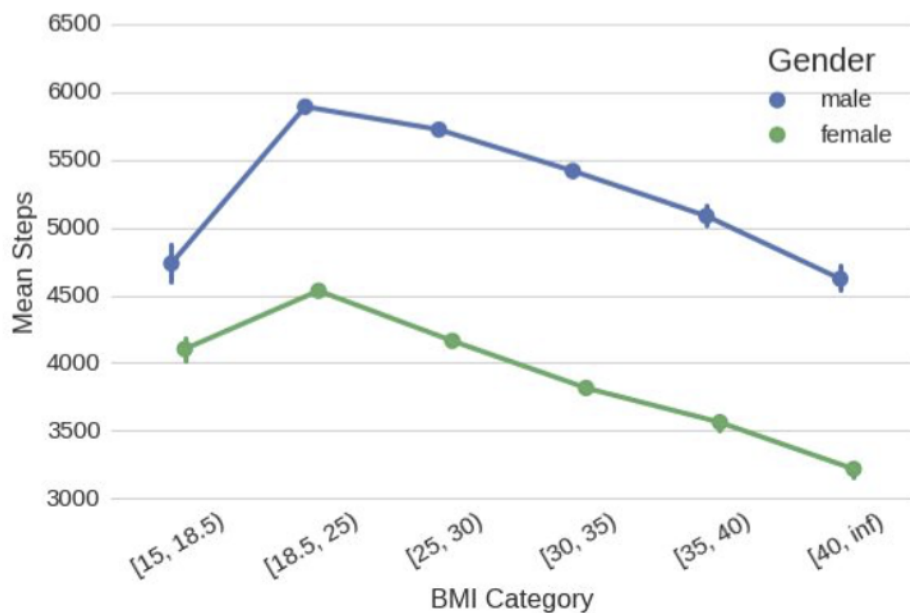
Personal Scale: Physical Activity & BMI

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

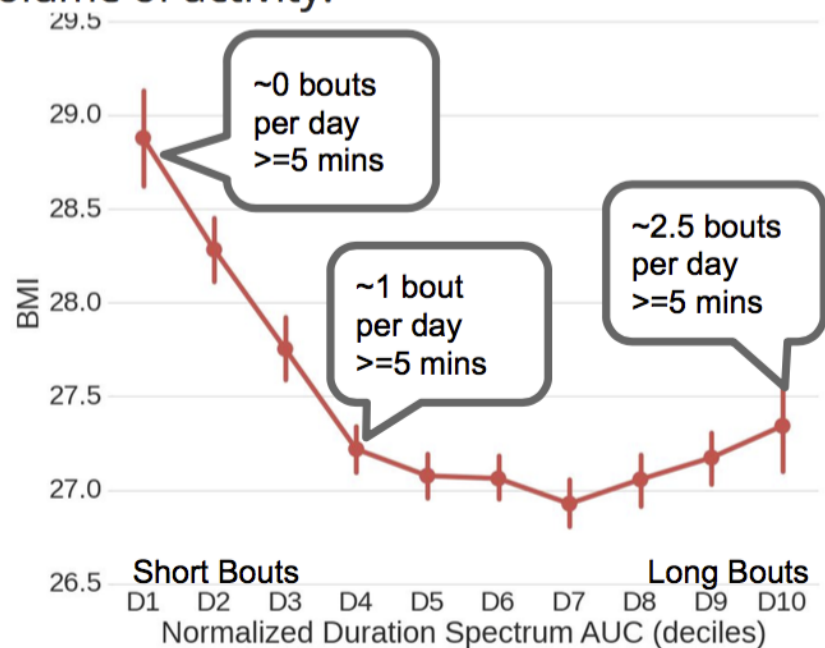


Personal Scale: Physical Activity & BMI

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

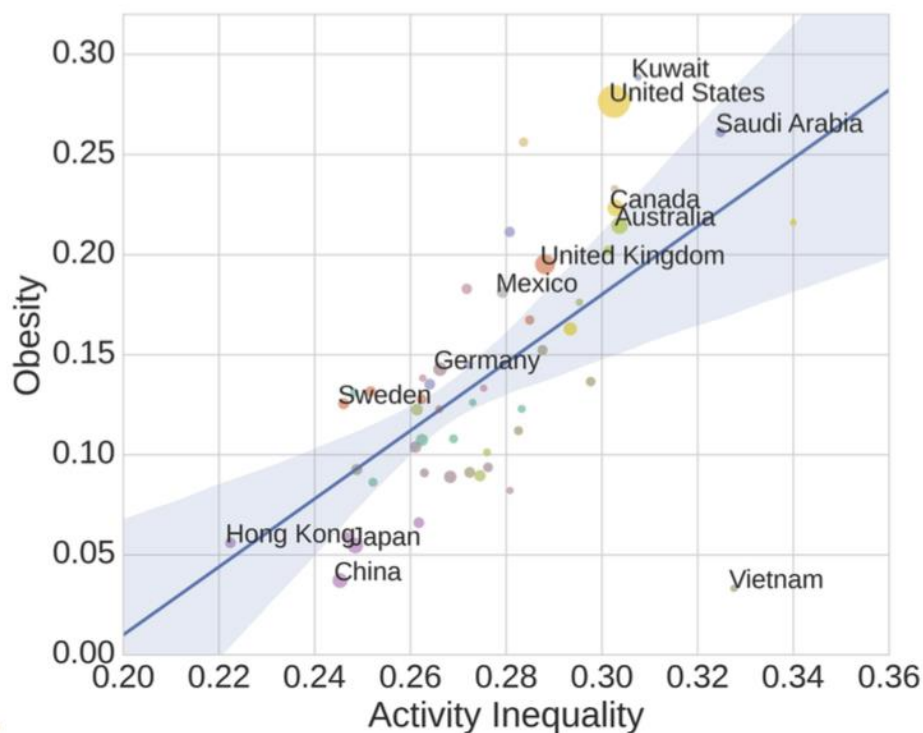


How is **bout duration** related to **BMI**?
Control for gender, age, daily wear time, and volume of activity.



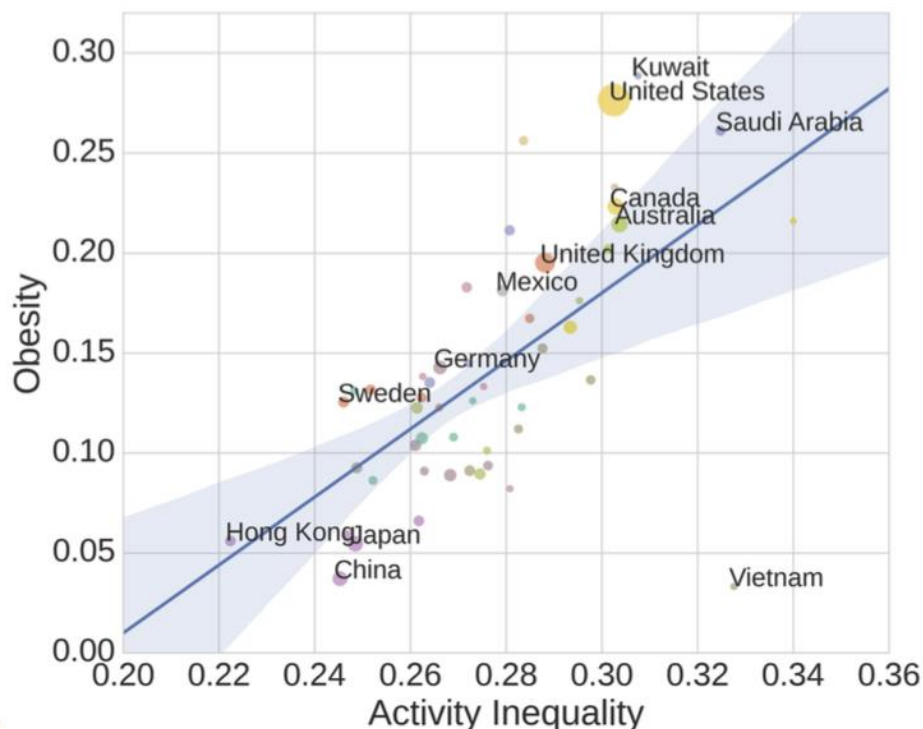
Planetary Scale: Environment, Activity, & Obesity

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

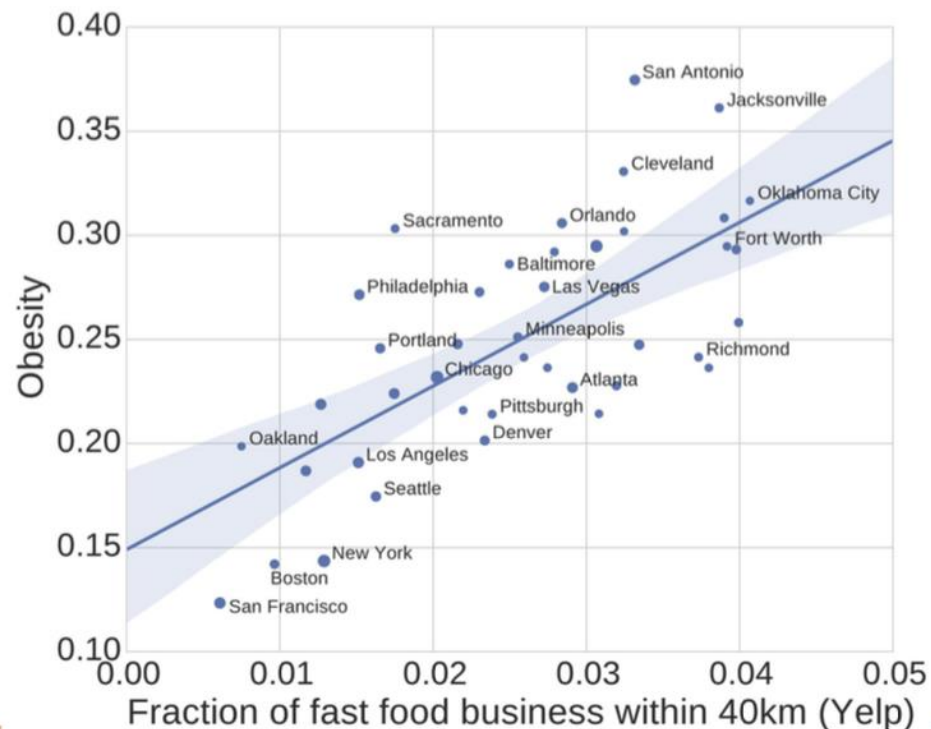


Planetary Scale: Environment, Activity, & Obesity

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



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

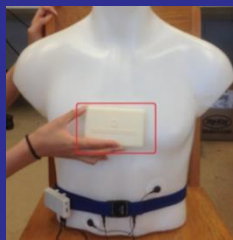


Mobile Sensor Data-to-Knowledge (MD2K)

Mobile Sensors



Smartwatch



Chestbands



Smart Eyeglasses

Exposures



Behaviors



Outcomes



Detecting First Lapses in Smoking Cessation

Saleheen, et. al., ACM UbiComp 2015

Modeling Challenges

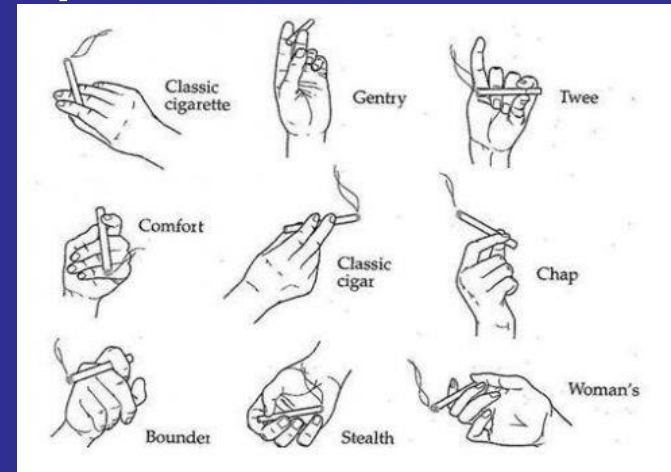
1. Ephemeral (very short duration)

- 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/56710118890712075/>

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

#8 The Commons

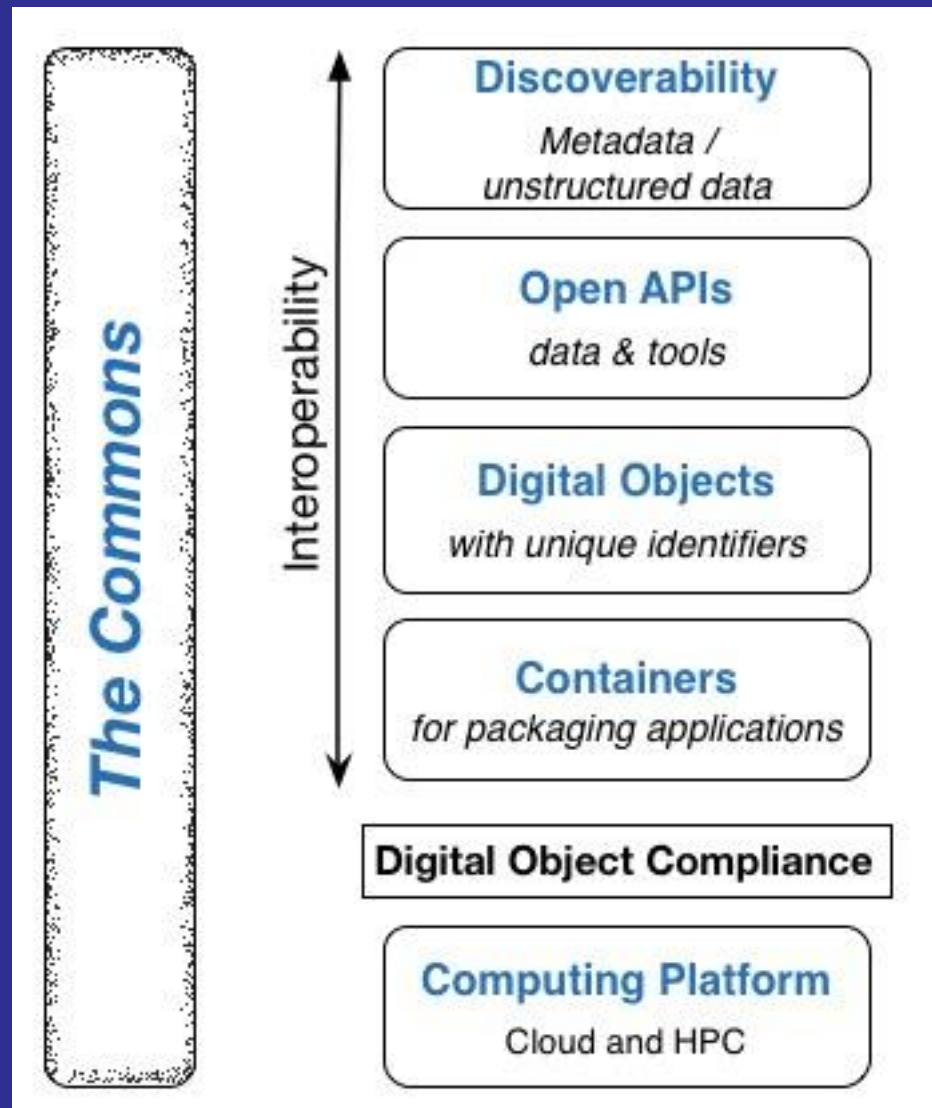
The Commons

Components

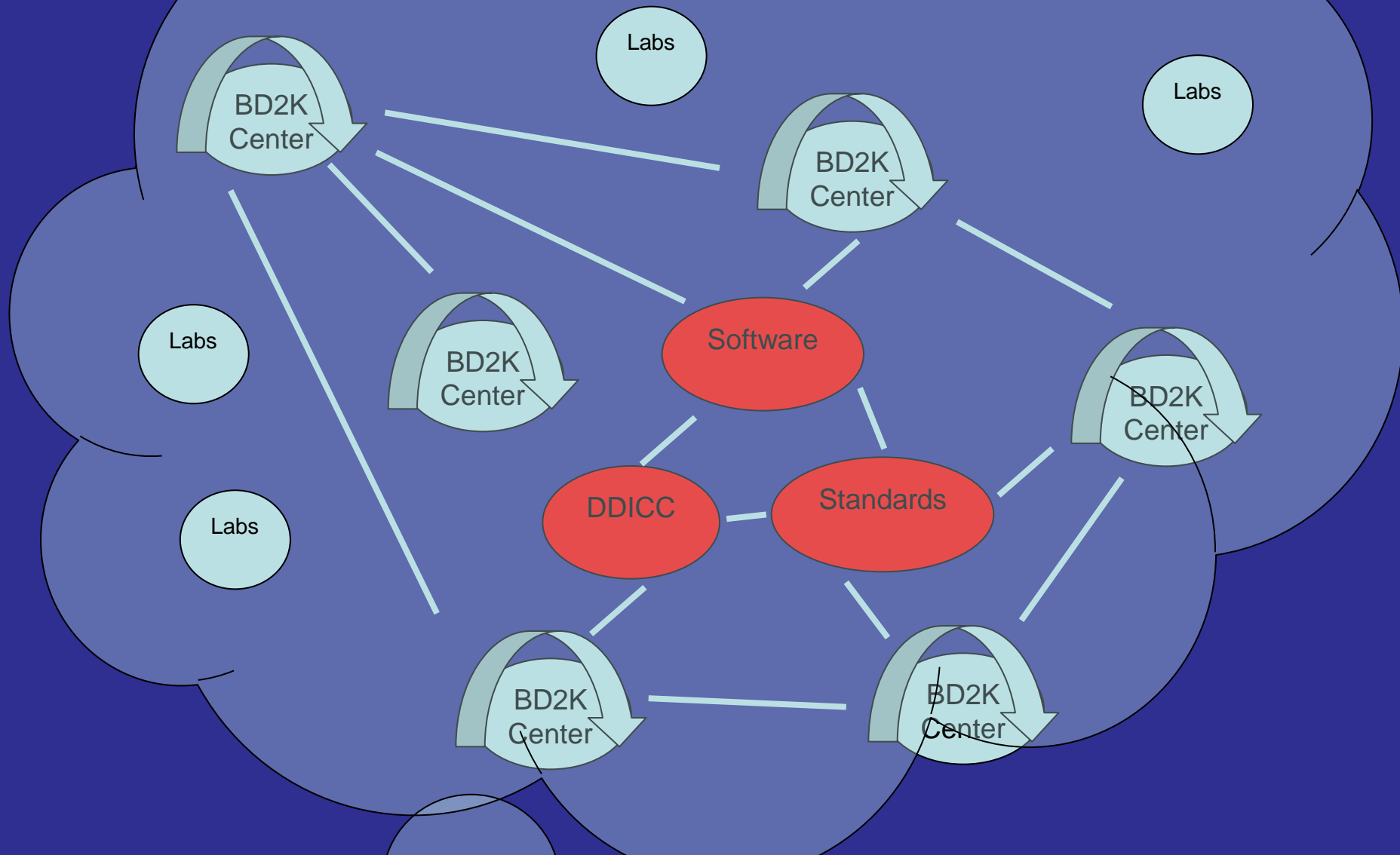
- **Computing environment**
 - cloud or HPC (High Performance Computing)
 - supports access, utilization, sharing and storage of digital objects.
- **Methods for Interoperability**
 - enables connectivity, shareability and interoperability between digital objects.
- **Digital object compliance model**
 - describes the properties of digital objects that enables them to be discoverable and shareable.

The Commons

Components



Infrastructure - The Commons



Commons - Pilots

- The Cloud Credits - business model
- BD2K Centers
- MODs (Model Organism Databases)
- HMP Data and tools available in the cloud
- NCI Cloud Pilots & Genomic Data

Commons

Acknowledgements

**The 133 Folks who have passed
through my lab over the years**

https://docs.google.com/spreadsheets/d/1QZ48UaKcwDI_iFCvBmJsT03FK-bMchdfuHe9Oxc-rw/edit#gid=0



NIH...

philip.bourne@nih.gov

Turning Discovery Into Health

