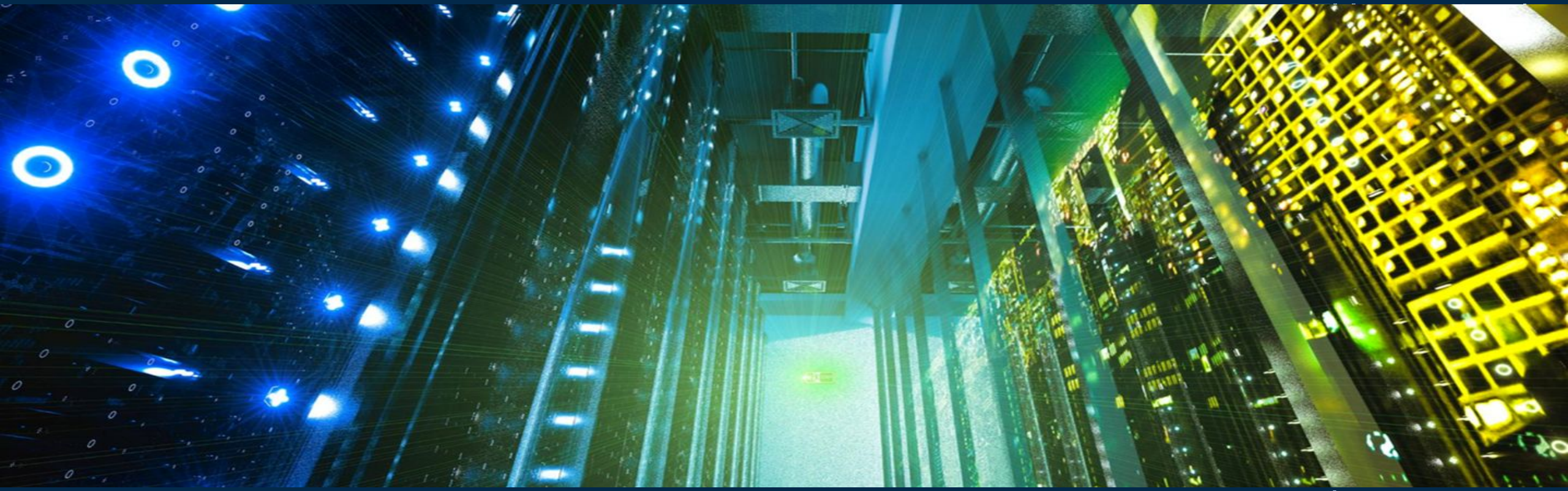
Better Science through Better Data 2018

The Rise of the Data Generalist

OR

Why Research Data Needs Renaissance Men and Women

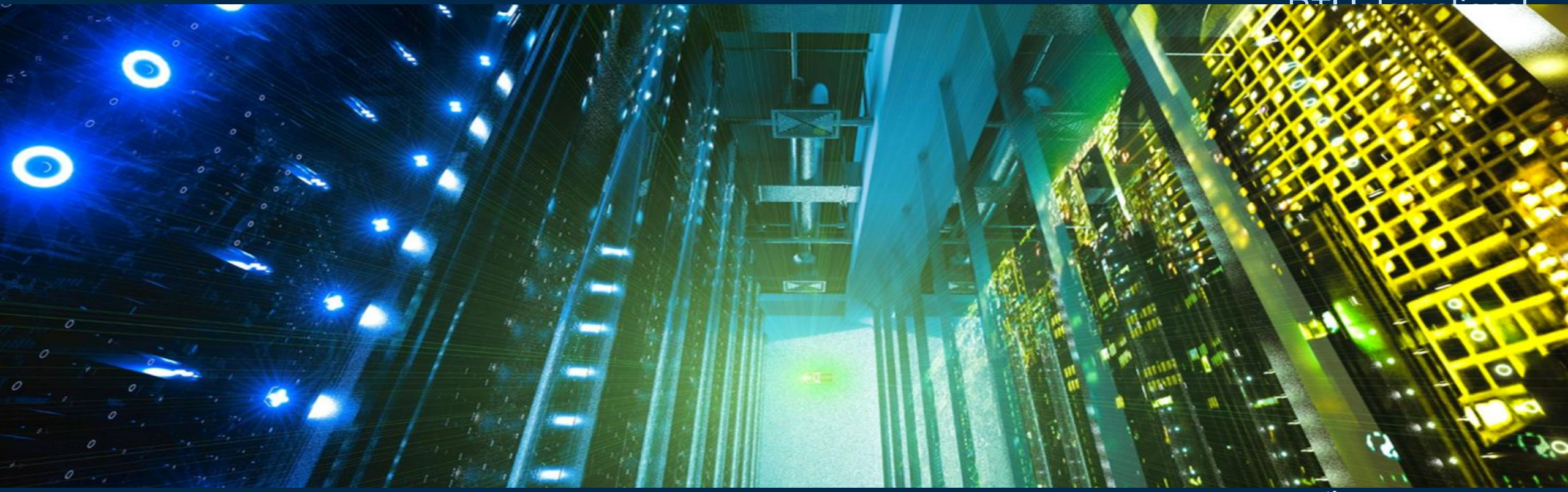
Rebecca Boyles
Oxford, UK
Bioinformatician and Data Scientist, RTI International
@becky_boyles



Better Science through Better Data 2018

www.slido.com
#scidata18

Rebecca Boyles
Oxford, UK
Bioinformatician and Data Scientist, RTI International



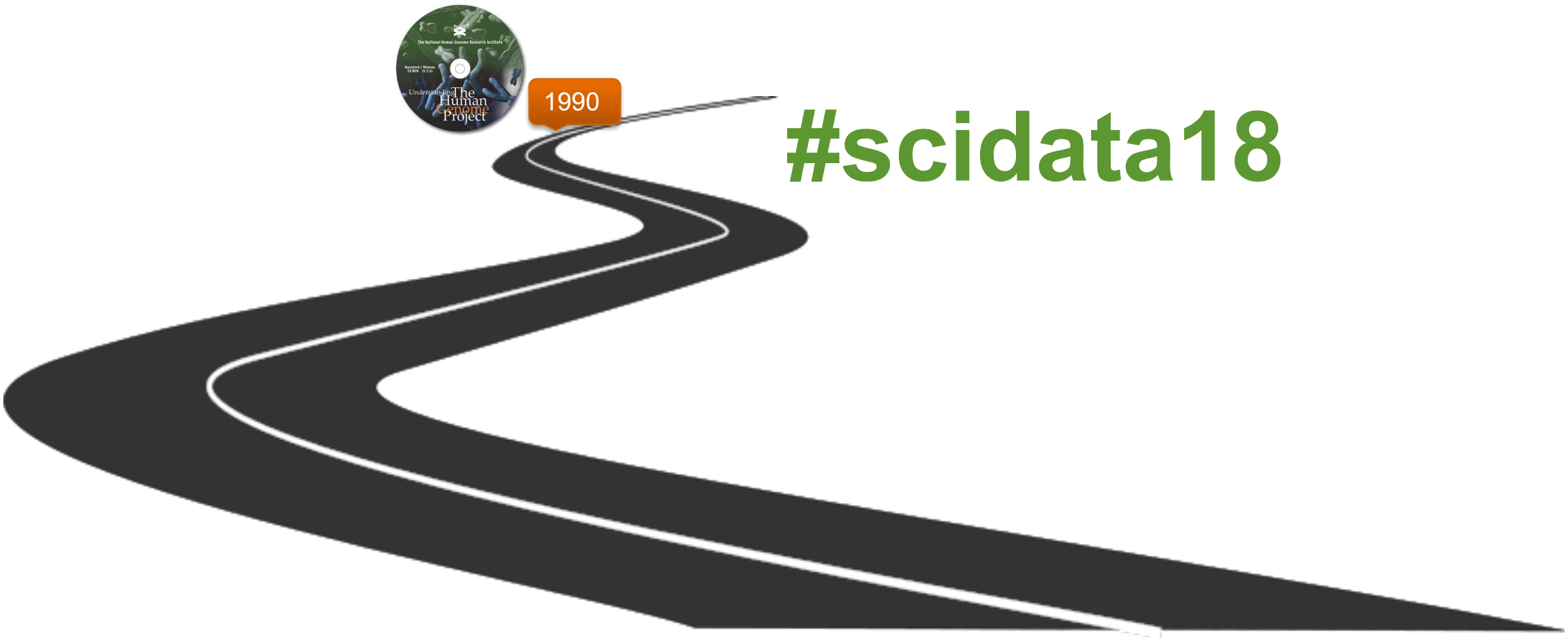
Overview

- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

Overview

- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

Biomedicine as a data discipline



Biomedicine as a data discipline

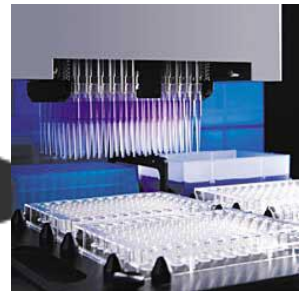
The Human Genome Project

- Compared to landing a man on the moon
- International collaborative program to map and understand the genes of humans... "genome".
- First draft published in *Nature* in February 2001 (~$2.7 billion)
- Francis Collins, then director of NHGRI, *"It is hard to overstate the importance of reading our own instruction book..."*
- The information is only as good as the ability to use it.

https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/
https://www.nature.com/articles/35057062

Biomedicine as a data discipline

1990

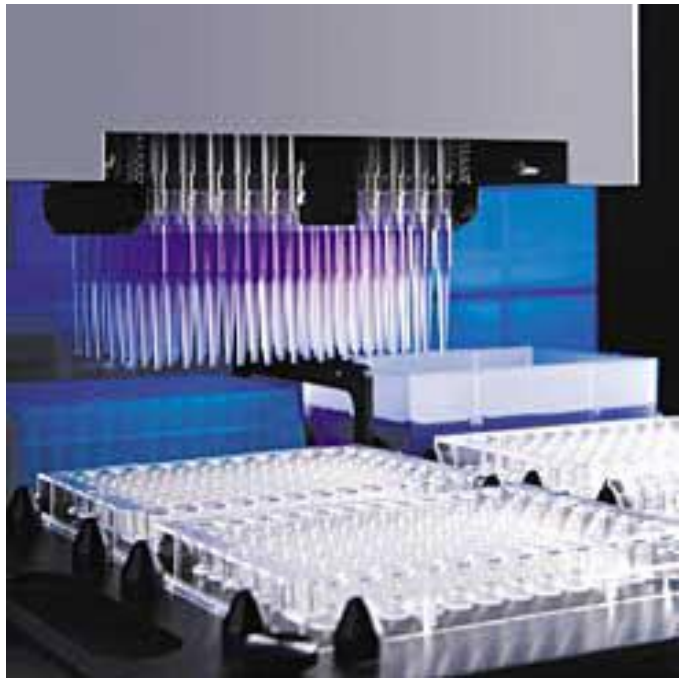


1999

Biomedicine as a data discipline

Robotics and High Throughput Screening

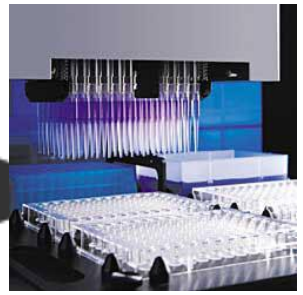
- Developed by Pfizer ~ 1986, but fully integrated in discovery 1999
- Coubled with robotics enabled rapid and repeatable in vitro experiments in 96, 384, 1536 or 3456 wells
- Explosion of data. Challege is identifying **biological significance** among plate effects and noise
- "*Soon, you're probably not going to be able to say that you're a molecular biologist if you don't understand some statistics or rudimentary data-handling technologies*," says [John] Blume. "*You're simply going to be a dinosaur if you don't.*"



https://www.nature.com/articles/4421067a
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1978279/

Biomedicine as a data discipline

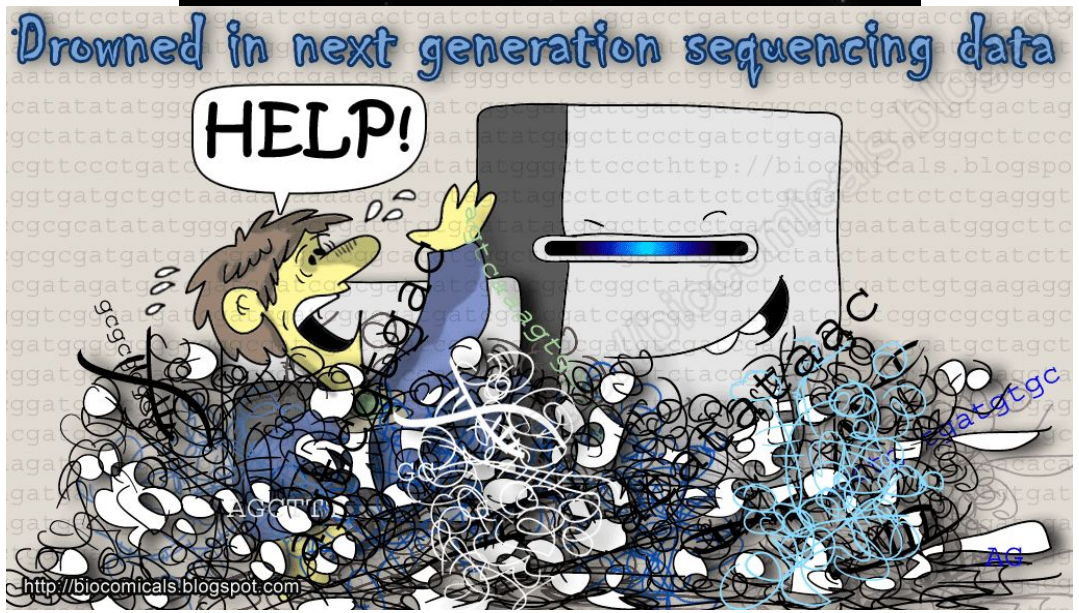
1990



1999

2008

Biomedicine as a data discipline

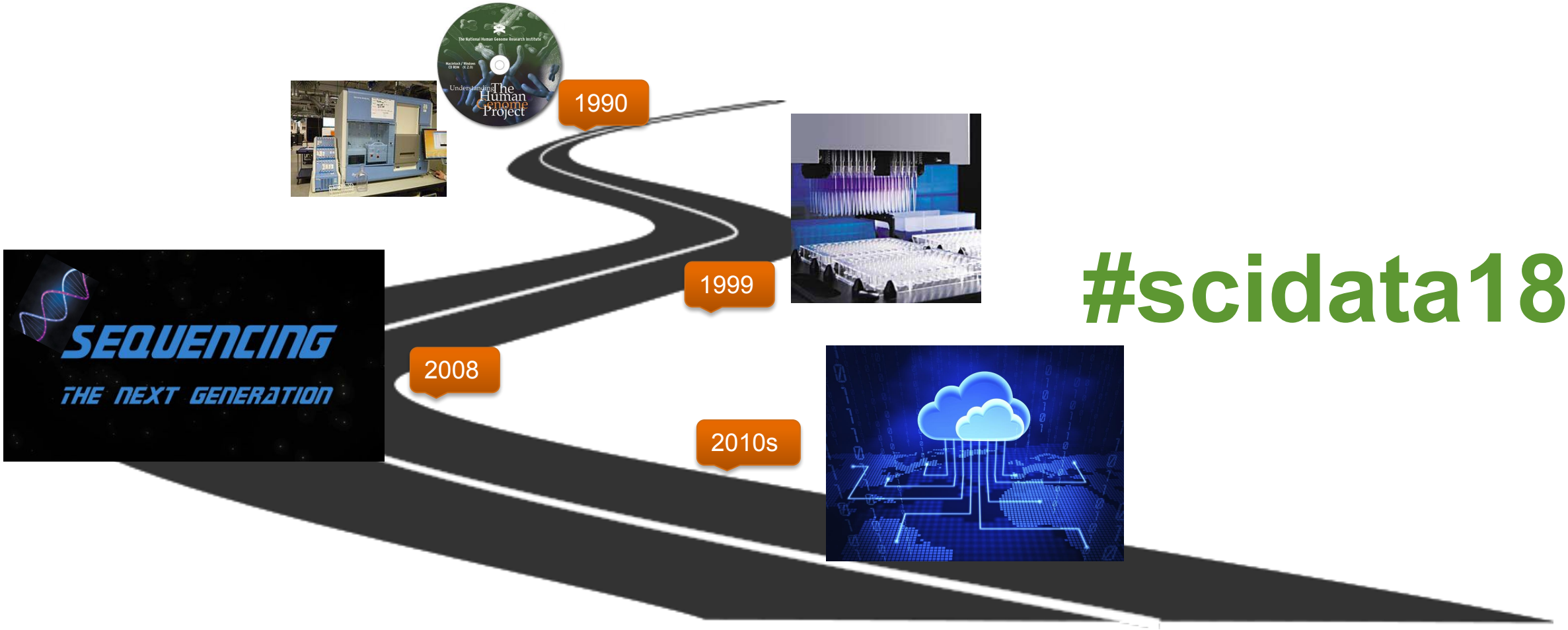


Next Generation Sequencing

- Entire genome in 1 day
- Sequencing millions of small DNA fragments in unison
- First draft published in *Nature* in 2008 by James Watson
- In recent years, coupling to cloud computing and bioinformatics tools has driven down cost (~$1,500 for draft sequence- 2015 and falling)
- Suddenly, the cost of **data storage, compute, and expertise** is a bigger cost than data production

https://www.nature.com/articles/nature06884
https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

Biomedicine as a data discipline



#scidata18

Biomedicine as a data discipline

Cloud Computing & Infrastructure as a Service

- Based on 1960s mainframe sharing, but largely commercially available around 2010s (EC2 launched by Amazon in 2006)
- 50% of all IT will be in Cloud in next 5-10 years
- Solves "geography" of data and tool sharing and democratizes compute access.
- NIH STRIDES to "*establish additional innovative partnerships to broaden access to services and tools, including training for researchers to learn about the latest cloud tools and technologies.*"
- European Open Science Cloud (EOSC) pilot to identify how to support a metadata ecosystem in the cloud

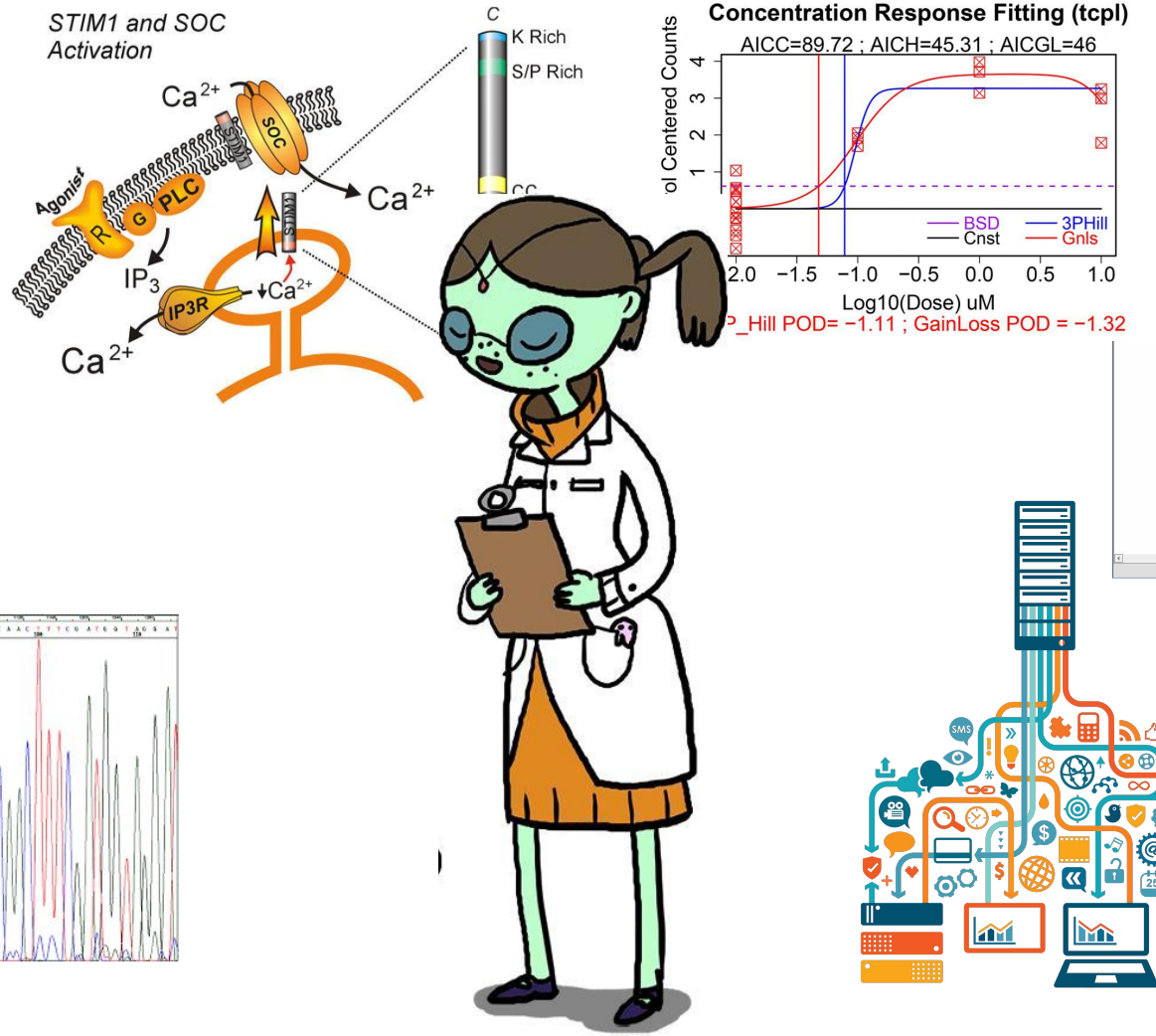
https://timesofcloud.com/cloud-tutorial/history-and-vision-of-cloud-computing/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1978279/
https://www.nih.gov/news-events/news-releases/nih-makes-strides-accelerate-discoveries-cloud
https://www.eoscpilot.eu/

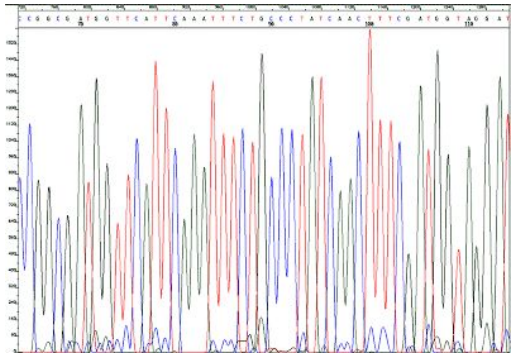
Overview

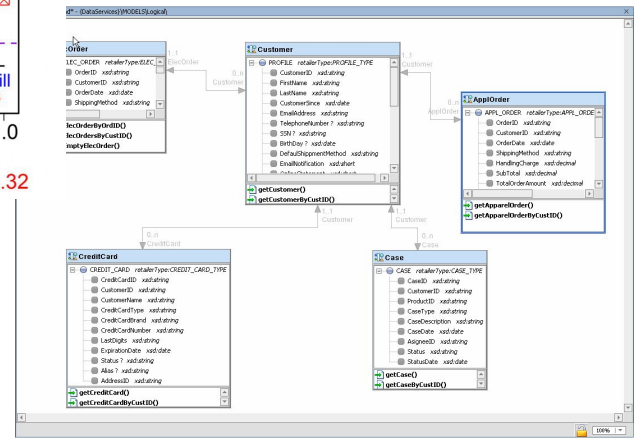
- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

#scidata18

My journey – in brief



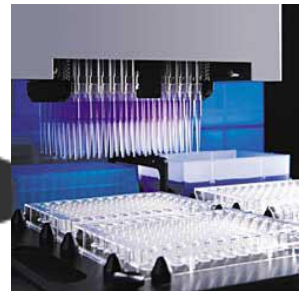




http://jcb.rupress.org/content/jcb/169/3/381.full.pdf

Biomedicine as a data discipline

1990



1999

2008

2010s

What does this mean???

- The information for human biology as written in the genome took 13 years to decipher
- DNA sequence reads used to be rate limiting, now we are talking about DNA as a future data storage device that can be read on demand
- Big data does not necessarily answer big questions, it needs to be analyzed and possibly shared and combined
- New scientific knowledge requires interpreting results in the context of relevant prior knowledge
- Scientists must store large data sets, integrate them, analyze, compare and share them—NOT EASY

AND increasingly they must understand how to work in **teams**, **communicate** data, how to build and use **compute infrastructure**, how to **document/ reproduce** experiments, and how to evaluate **new technologies**.

Overview

- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

Data as a resource

Click to add text

#scidata18

The world's most valuable resource is no longer oil, but data

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Is data a natural resource?

natural resources

noun
noun: **natural resource**

materials or substances occurring in nature which can be exploited for economic gain.
"the sustainable use of natural resources"

Data used to be viewed as a **by product** of research
but now it is as likely a **starting point**.

Data as a resource

£54 million funding to transform health through data science

7 February 2018

https://bit.ly/2z52dT0

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

The data economy is predicted to be worth £94.6 billion by 2025.

https://www.dataiq.co.uk/article/uks-data-economy-worth-ps73-billion-potential-greater

Maximize the Value of Your Data Science Efforts by Empowering Citizen Data Scientists

Published: 12 June 2018 **ID:** G00343732

Analyst(s): Carlie Idoine | Erick Brethenoux

https://www.gartner.com/doc/3878963?ref=mrktg-srch

Data as a resource

Closed Shared Open

Even "open science" projects are not open

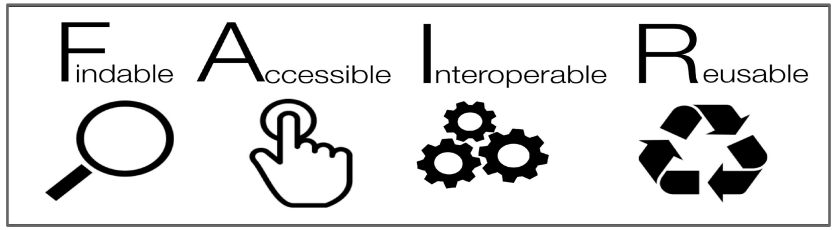


Image by SangyaPundir

The Open Science Prize

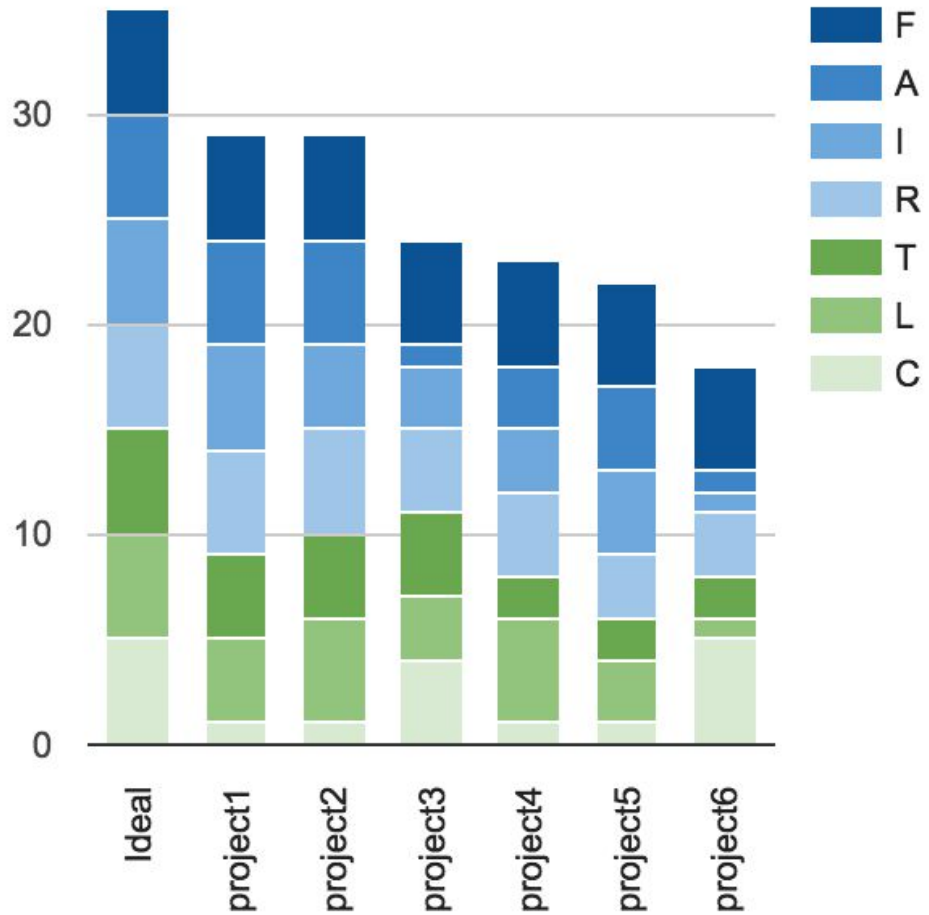
MyGene2

Open Imaging

nextstrain

OpenTrialsFDA
An OpenTrials App

openaq



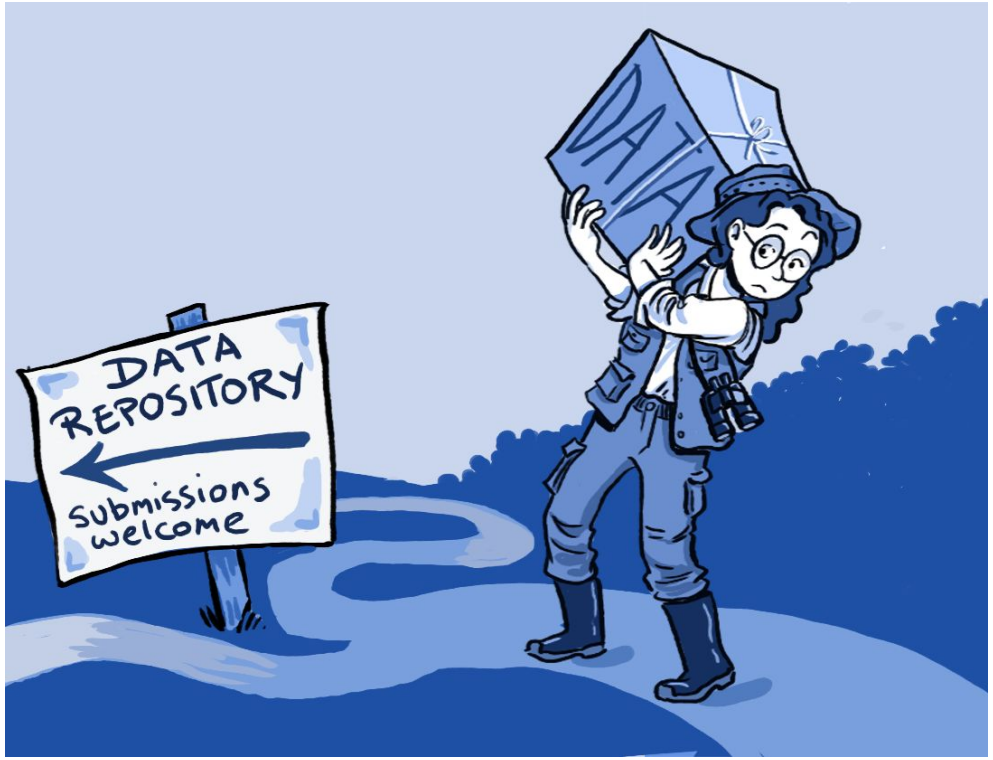
https://doi.org/10.5281/zenodo.253046

Adapted from Melissa Haendel

Open science is work

Data management planning is:

- Expensive
- Time consuming
- Requires expertise



#scidata18

doi:10.1371/journal.pbio.1001779

Overview

- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

Biomedical data commons

The Data Commons is a platform that fosters the development of a digital ecosystem.
- Vivien Bonazzi, NIH

A platform is a plug and play model that allows multiple participants (producers and consumers) to connect to it, interact with each other and create value.
- *Sangeet Paul Choundary*

Changing the conversation in data sharing

Data Commons

How do we find data, software and standards?

How can we make data, annotations, software and metadata accessible?

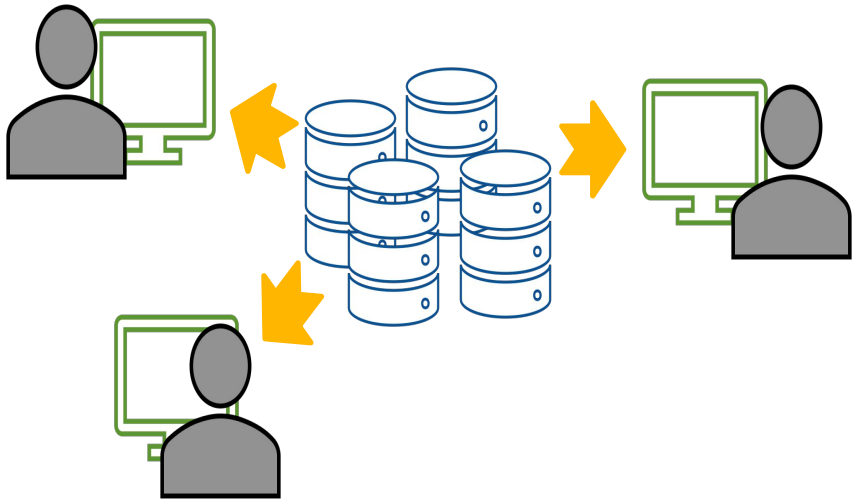
How do we accommodate closed, shared, and open data?

How do we reuse data standards?

How do we make more data machine readable?

A new model for data sharing

Current state

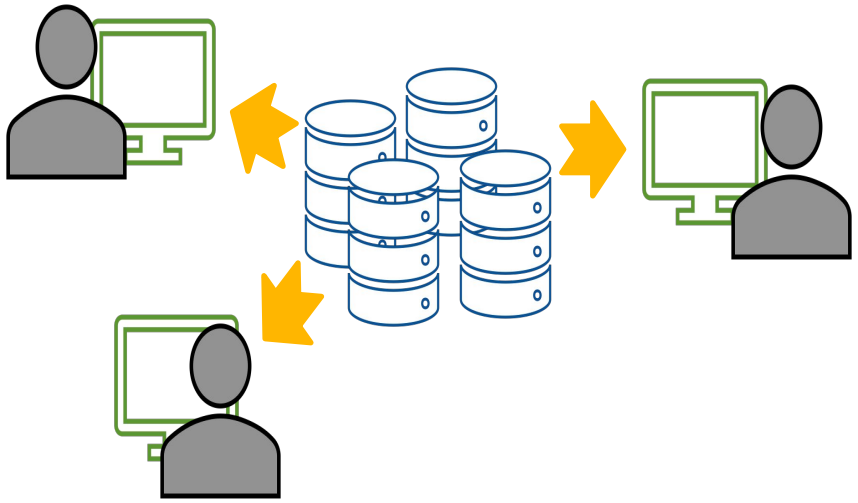


- Data sharing = Data Copying
- Security risks (data handoffs)
- Does not support Team Science
- Siloes compute

Adapted from Moran Caballi

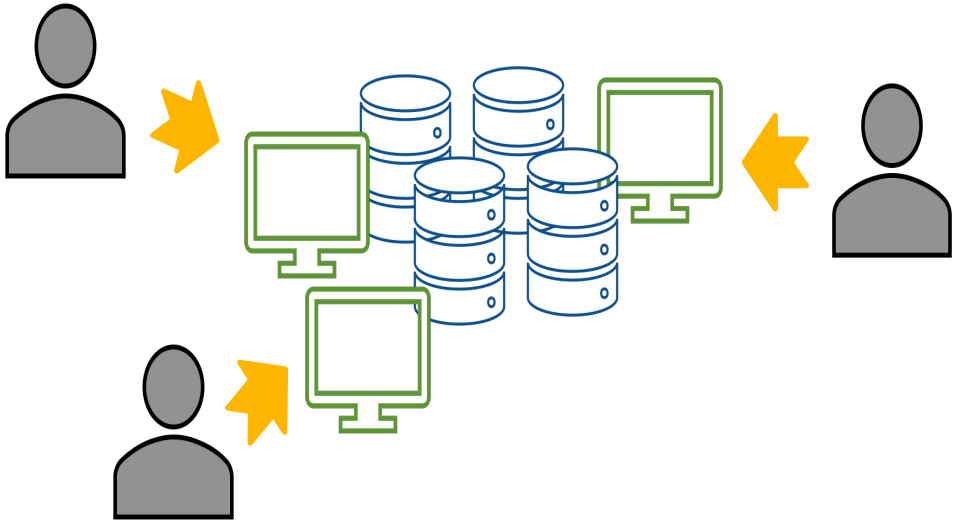
A new model for data sharing

Current state



- Data sharing = Data Copying
- Security risks (data handoffs)
- Does not support Team Science
- Siloes compute

Future state



- Data management
- Enhanced security and controls
- Collaboration space
- Access to compute, tools, and expertise

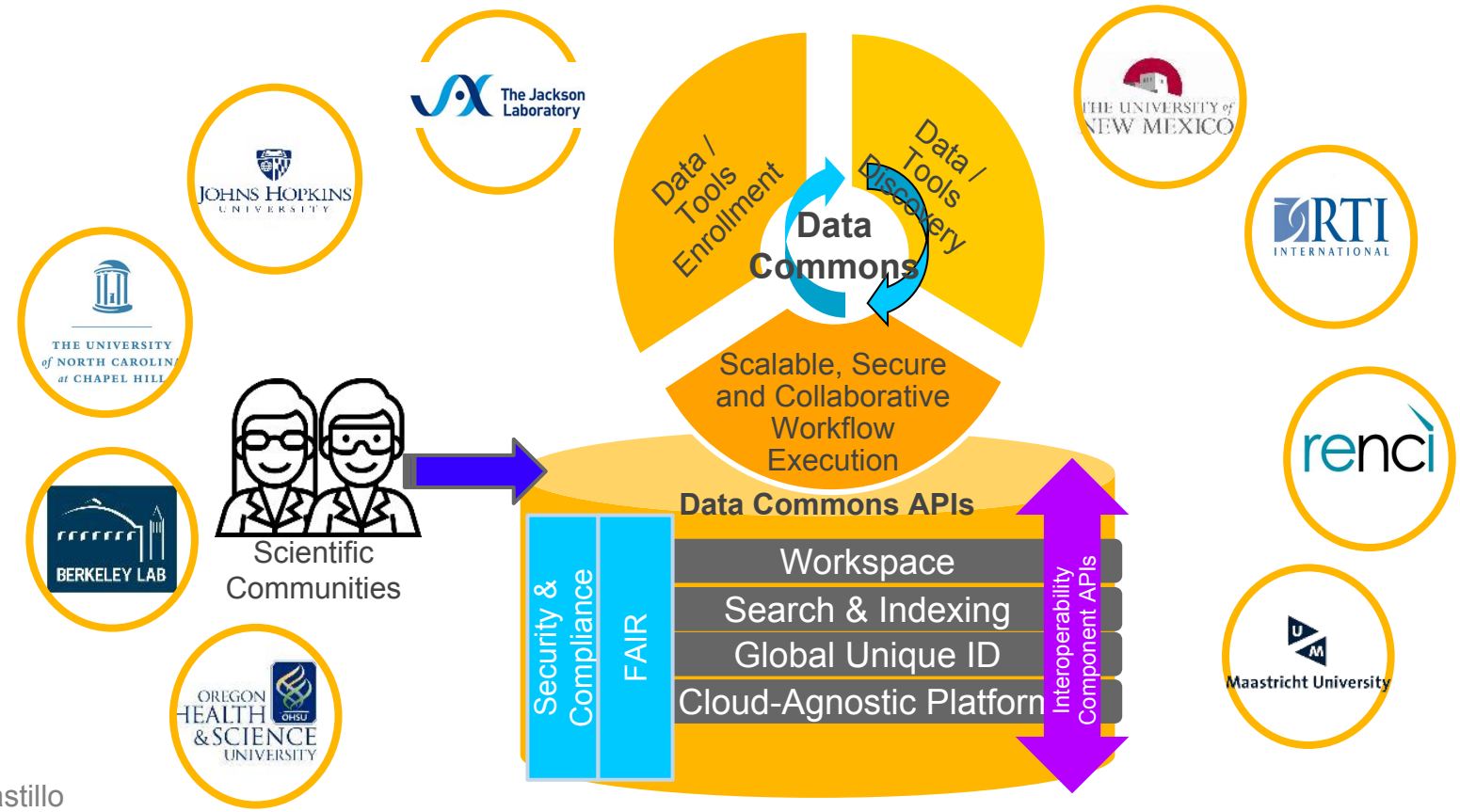
Adapted from Moran Caballi

NIH Data Commons Pilots

Enable users to be both producers and consumers of data and capabilities.

The Commons will provide the platform for a science marketplace and the building blocks to enable forward-thinking capabilities.

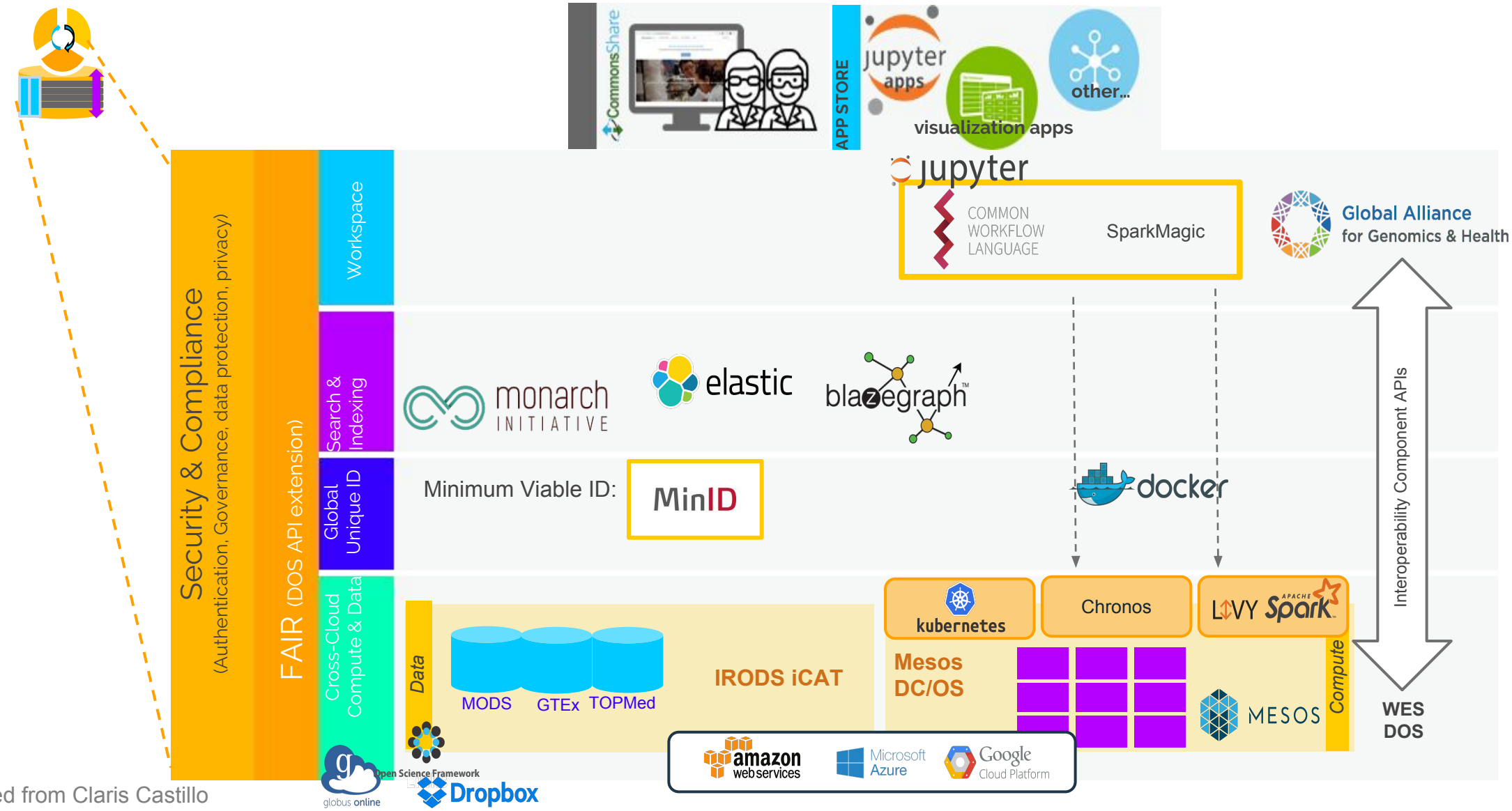
As such the primary asset of the Commons resides and the data and interactions that it enables.



Adapted from Claris Castillo

NIH Data Commons Pilots

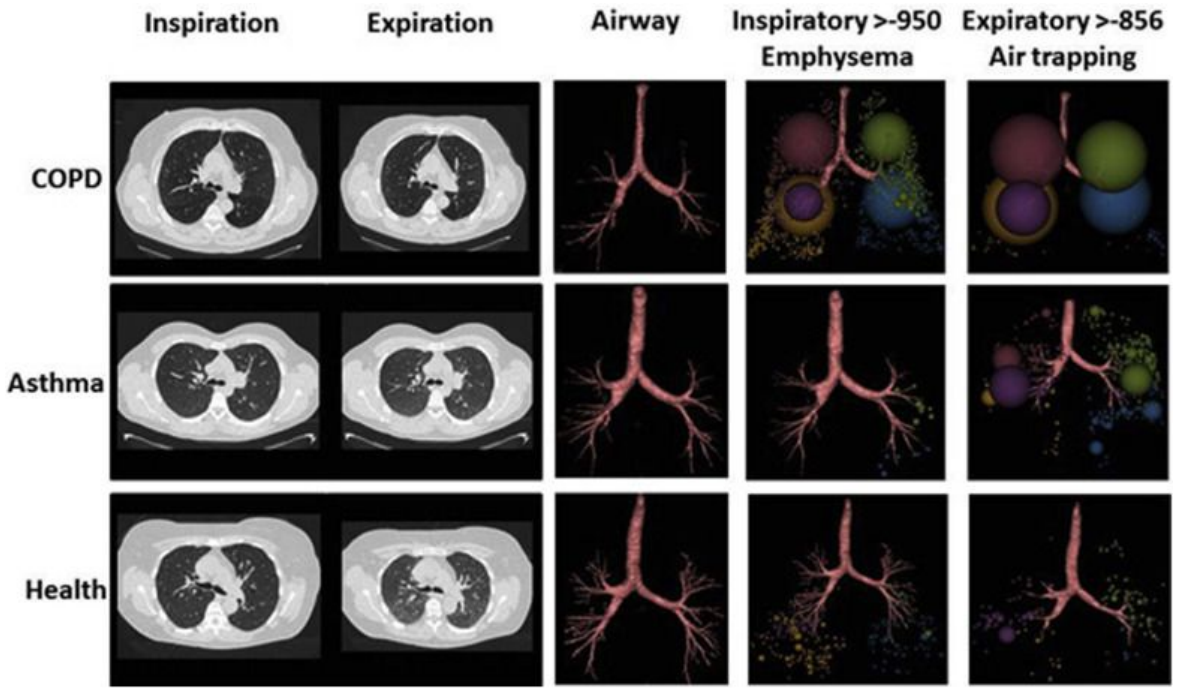
CommonsShare: Cloud-Agnostic Architecture



Adapted from Claris Castillo

Commons use case

Deep Learning on Chest CT Images



A machine learning method on neural networks to learn by training to recognize patterns.

- Improve image segmentation or feature identification
- Predict rates of disease progression
- Classify disease into subtypes

Hartley RA et al. Relationship between lung function and quantitative computed tomographic parameters of airway remodeling, air trapping, and emphysema in patients with asthma and chronic obstructive pulmonary disease: a single-center study. J Allergy Clin Immunol 2016;137(5):1413–1422.e12. doi: 10.1016/j.jaci.2016.02.001.https://www.ncbi.nlm.nih.gov/pubmed/27006248.

Overview

- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

The tragedy and governance of commons

tragedy of the commons

describes a situation in a shared-resource systems where individual users act independently according to their own self-interest, and contrary to the common good- depleting the resource

- **Garrett Hardin**

paraphrased from "The Tragedy of the Commons," Science 1968;162(3859):1243–1248. doi: 10.1126/science.162.3859.1243. http://science.sciencemag.org/content/162/3859/1243

governance of the commons

a general framework for successful self-organization to sustain a community system that includes: size, productivity, mobility, number of users, leadership, social norms and ethics, knowledge, importance, and collective choice

- **Elinor Ostrom**

paraphrased from "A General Framework for Analyzing Sustainability of Social-Ecological Systems," Science 2009; 325(5939):419–422. doi: 10.1126/science.1172133. http://science.sciencemag.org/content/325/5939/419

Tackling the Commons challenges

Data Commons

How do we engage a community and build productivity?

How can we establish social norms around FAIR and sharing?

How do we educate the scientific community and recruit minds?

How do we communicate the importance of the effort and sustain funding for infrastructure?

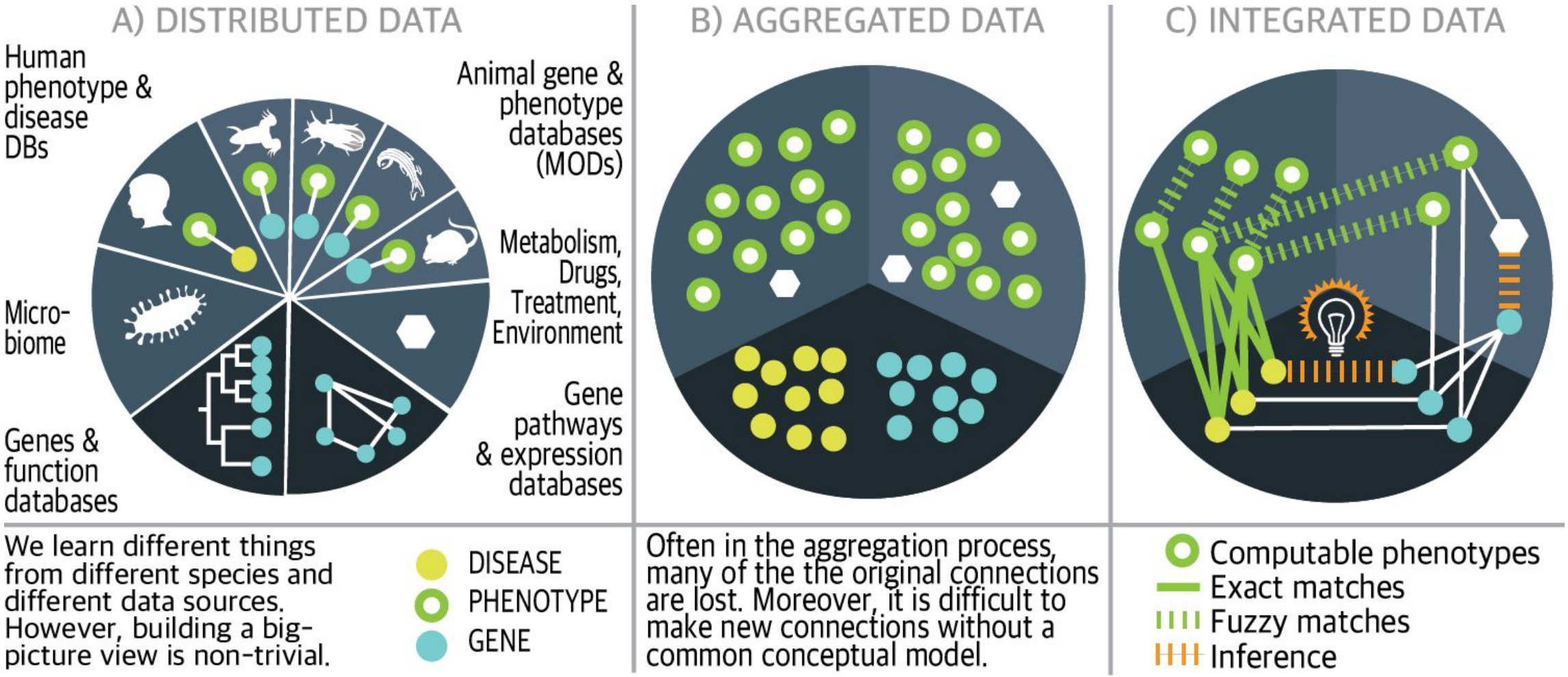
How do we collectively govern the Commons?

a National Cancer Institute program

Data **C**ommons **P**ilot **P**hase **C**onsortium

Data **S**torage, **T**oolspace, **A**ccess and analytics for bi**G** data **E**mpowerment

What about the data?

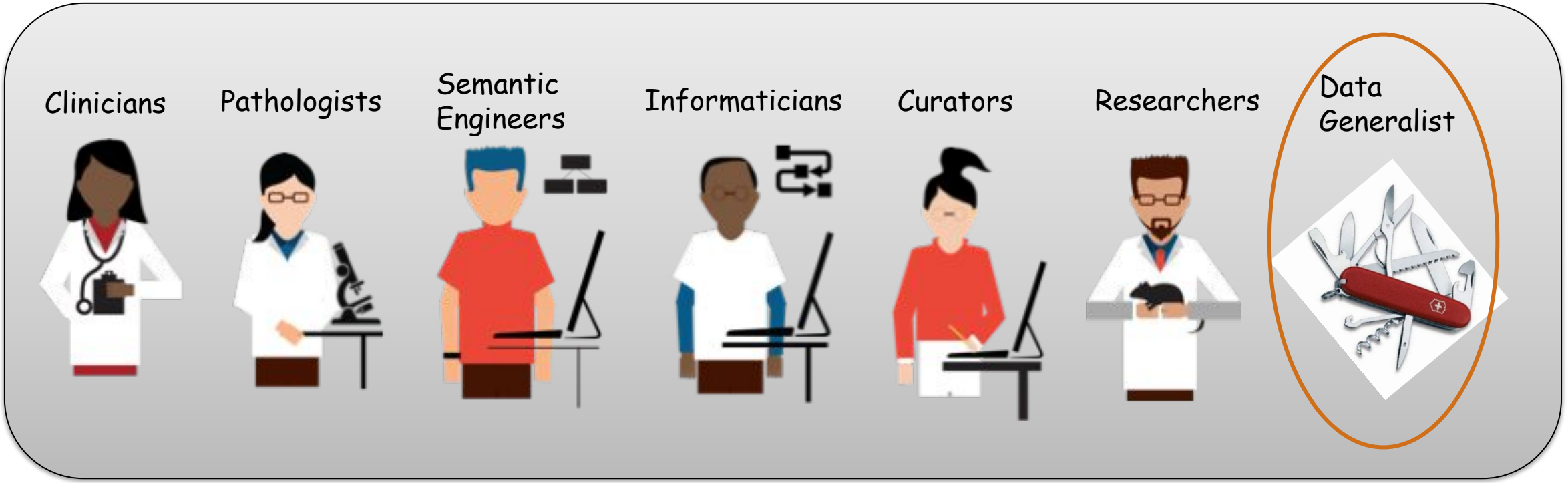


McMurry et al. Navigating the phenotype frontier: the Monarch Initiative. Genetics 2016;203(4):1491–1495. doi: 10.1534/genetics.116.188870.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981258/figure/fig1/

Overview

- Brief history of biomedicine as a data discipline
- My professional journey
- The realization of data as an asset or resource
- Development of "data commons"
- Challenges and approaches to realizing a commons
- The rise of the data generalist

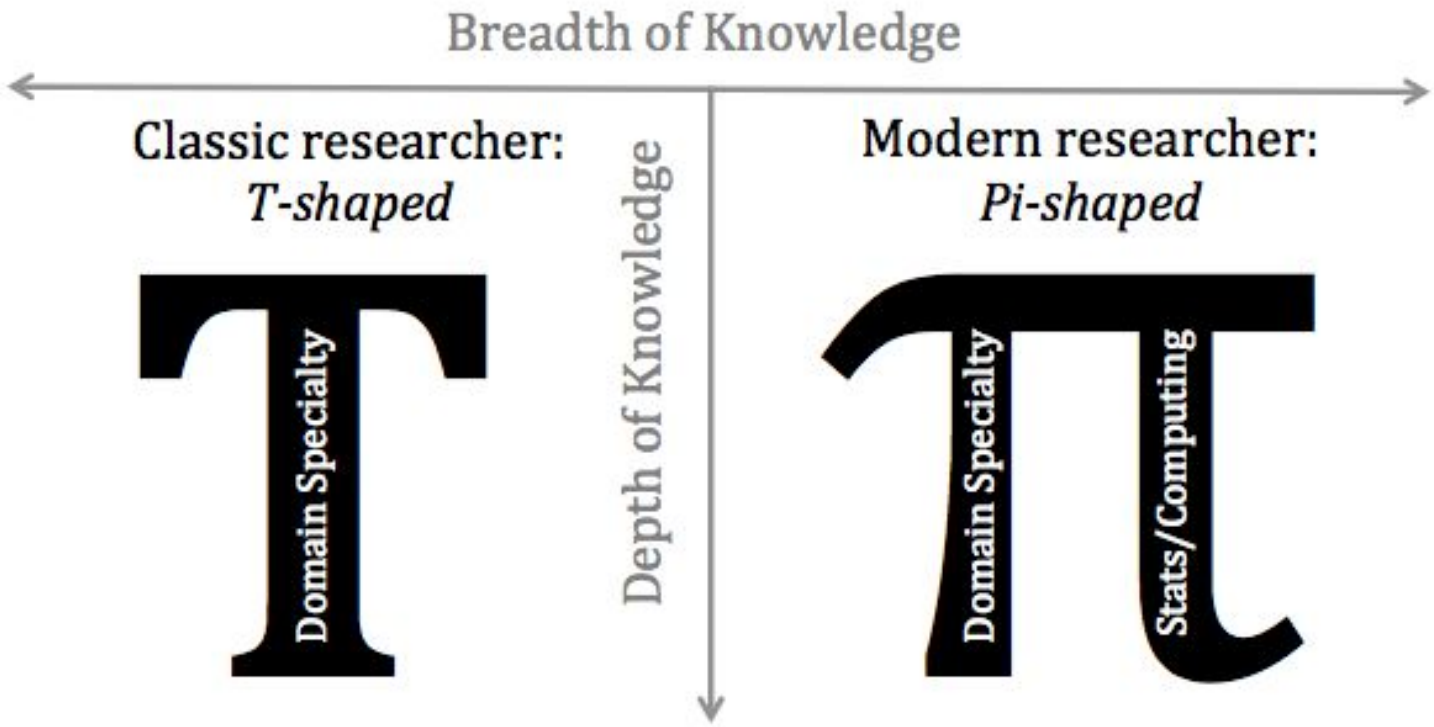
Data Science is Team Science



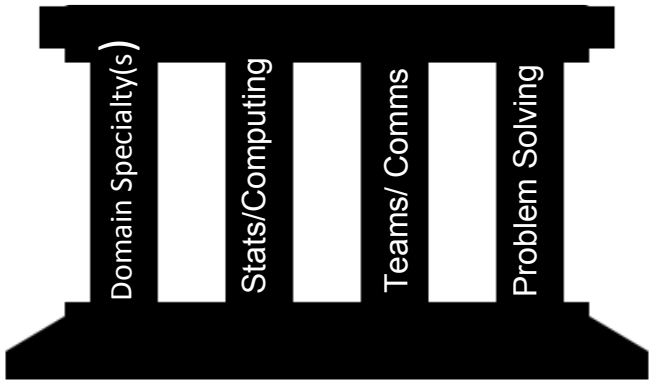
#scidata18

Rise of the Data Generalist

- Understand teams, communications, cost/benefit
- Understanding of clinical, informatics, computational and basic research processes and data
- Problem solving to deploy the expertise: AI, blockchain, biochips, 3D printing, cloud computing



Data generalist



The Future of Biomedicine

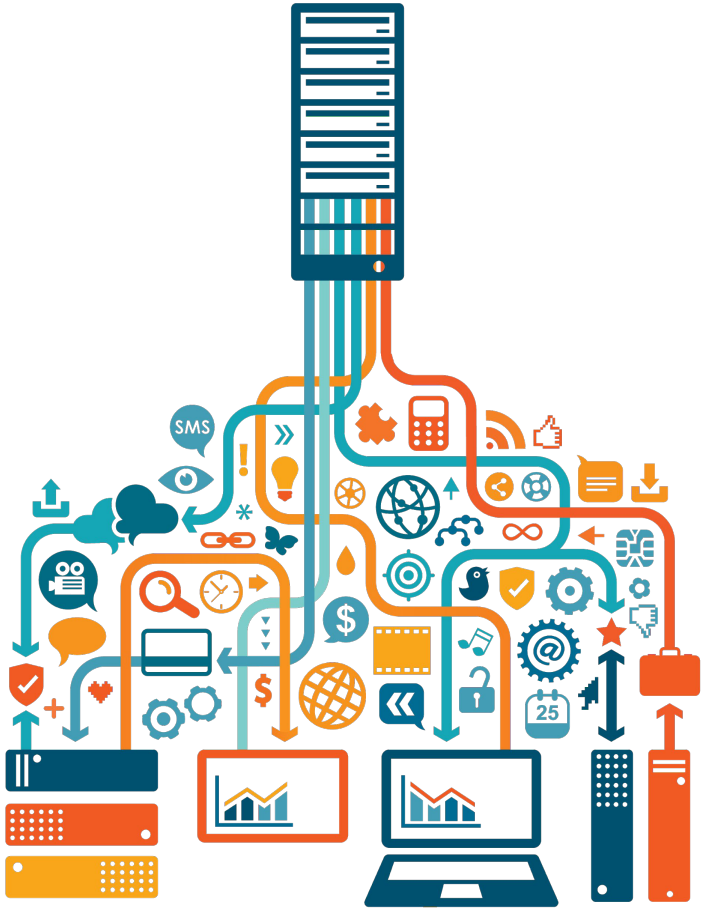
Imagine…

Teams can spontaneously form around commons interests and research questions (and data)

Infrastructure and advanced analysis is offered as a service

Digital objects are FAIR, and interoperability standards let a user navigate from resource to resource seamlessly (and publication)

Scientific advances demonstrate the validity of the open science in research by treating data as a first class scientific contribution



The Future of Biomedicine

Data generalists: a catalyst for change

- Translate across domains
- Understand inherent limitations to data/ experiments
- Identify the right tool for the problem
- Communicate value
- Assess cost/benefit of an approach

Acknowledgements

Data Partners	Chronic Obstructive Pulmonary Disease (COPD) Gene
	Trans-Omics for Precision Medicine (TOPMed)
	Alliance of Genome Resources (AGR)

Teams	Institution
Calcium	Broad Institute
	The University of California, Santa Cruz
	The University of Chicago
Carbon	Harvard Medical School
Helium	Lawrence Berkeley National Laboratory
	Oregon State University
	Renaissance Computing Institute: RENCI
	RTI International
	The Jackson Laboratory
	University of New Mexico Health Sciences Center
Xenon	Elsevier
	Repositive
	Seven Bridges Genomics Inc
	US Department of Veterans Affairs

NIH NHLBI, 1 OT3 HL147154-01

NIH OD, 1 OT3 OD025464-01 S2

RTI International

delivering **the promise of science**
for global good

Rebecca Boyles
Senior Scientist, Bioinformatics and Data Science
Oxford, UK

rboyles@rti.org
@becky_boyles