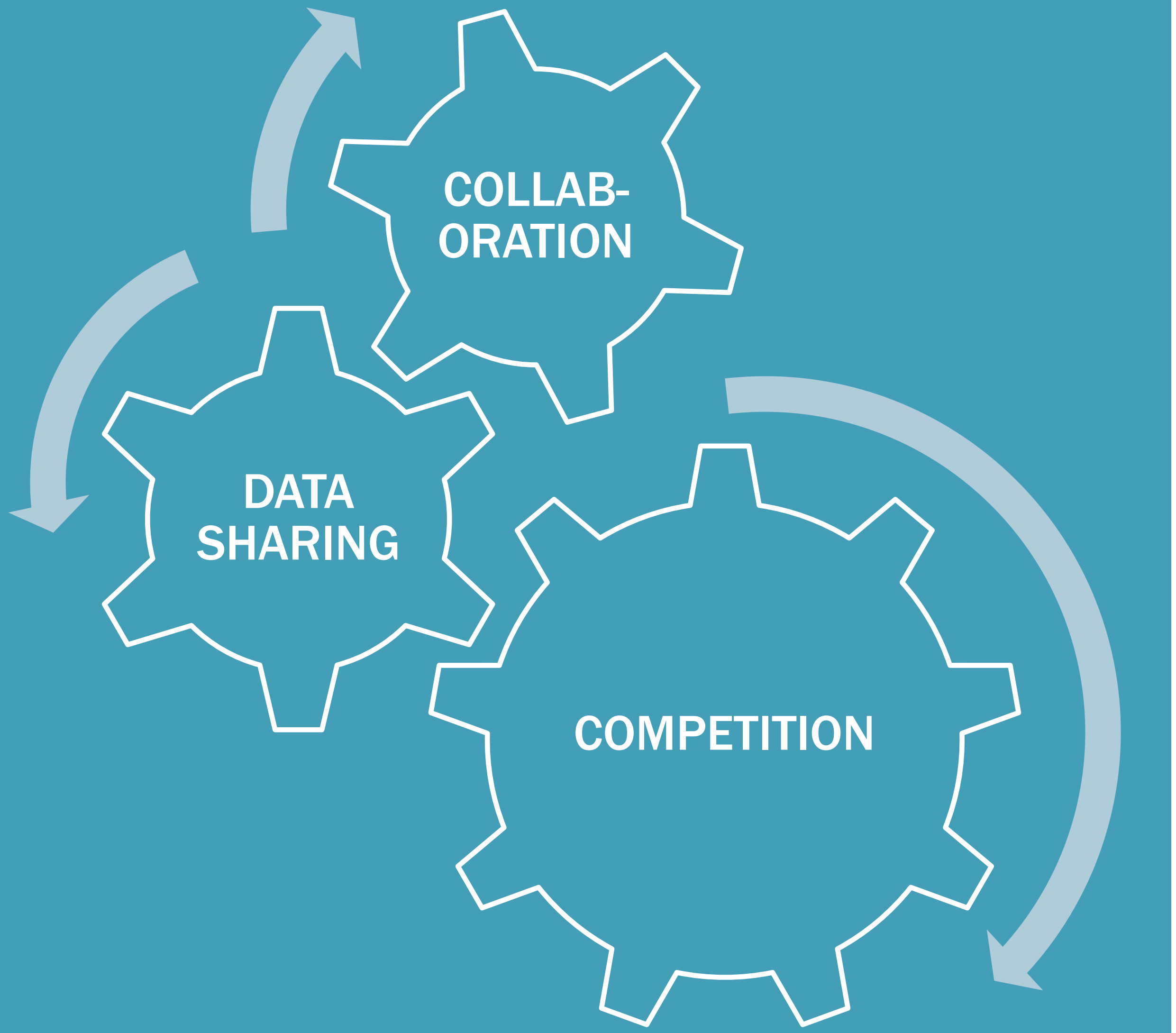
# COMPETITION, COLLABORATION & DATA SHARING IN SCIENCE

*An overview of ongoing challenges*

Glenn Hampson
Program Director
Open Scholarship Initiative
For VI BRISPE Conference
October 29, 2021



# HOW ARE THESE 3 ACTIVITIES RELATED?

1. The gears of science are all connected. Competition is fundamental, but science also requires collaboration and data sharing.
2. Each of these systems has unique practices, outcomes and challenges that end up affecting the other systems.
3. Understanding how this large system operates and is evolving is important for understanding how to improve the science of tomorrow.

# PRACTICES & OUTCOMES

What are the practices and outcomes that define each of these three systems?

# COMPETITION

## PRACTICES

1. The race to discover has always been a key part of science (indeed, it is fundamental to the fabric of science)
2. In academia, competition for funding and tenure is increasingly intense
3. In industry (where most science research happens) competition also means market share and profits

## OUTCOMES

- “Market-like” incentives and intensity
- Secrecy and embargoes (hang on to data until you’re done with it)
- Temptation to commit fraud (p-hacking, fake images, etc.) or take shortcuts (bad science, biased conclusions, etc.)
- Market for publishing solutions that accept bad science (like predatory journals with no peer review)
- Hyper-focus on “winning” the race for citations and attention (at the author level) and measuring/attaining impact (at the publisher, funder, and university levels)
- Hide (or at least don’t publish) negative findings
- Premium on glitz and glamor studies, not replication studies
- “Salami slice” published work: Take one study and make 2-3 papers out of it to get more publishing credit

# EXAMPLE

Rosalind Franklin discovered the double-helix structure of DNA in 1952 and 1953 through her groundbreaking work in x-ray crystallography. Dr. Franklin kept her work tightly guarded, only revealing as much detail as necessary to collaborate with other scientists who were also searching for the true structure of DNA. American researchers James Crick and Francis Watson managed to see photographs of Dr. Franklin's work, though, without permission, and because of this huge "hint," published their findings before Franklin. Crick and Watson were awarded the 1962 Nobel Prize for their discovery while history essentially forgot about Dr. Franklin.

# COLLABORATION

## PRACTICES

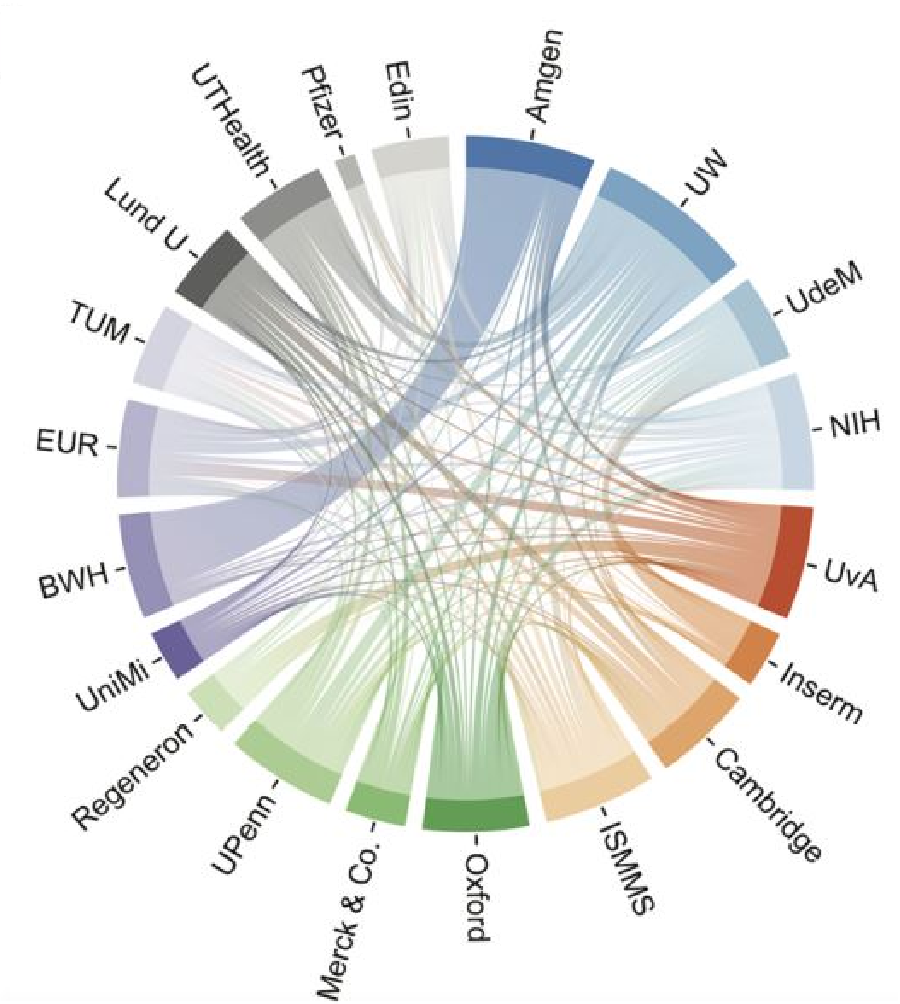
1. Working together with a team of experts is increasingly integral to the success of science.
2. Collaboration can mean distributing resources across institutions (computing, personnel, space, etc.)—resources that a single institution can't afford alone
3. The key to success is striking the right balance between benefit/value and risk/cost

## OUTCOMES

- Scientific collaboration (as measured by research teams and co-authorship) is much more prevalent today than several decades ago, especially in biomedical research.
- The global nature of collaboration is also up sharply (doubling from 10% of published outputs in 2010 to over 20% today).
- Statistically, collaboration may be more like to generate new scientific insights, uncover issues, and lead to additional publications, citations, and collaborations.
- Data Sharing Agreements are key features of these collaborations

# EXAMPLE

The PCSK9 gene provides instructions for making a protein that helps regulate the amount of cholesterol in the bloodstream. Over the last 20 years, the development of the PCSK9 field has involved the collaboration of 9,286 scientists from 4,203 institutions around the world. Interactions between the top 20 institutions are shown in the graphic (Chen 2020).



Source: Chen 2020

# DATA SHARING

## PRACTICES

- This is the most sophisticated and sensitive form of collaboration (due to competition concerns).
- “Open data” is a fairly new practice—only about 10 years old. The Panton Principles (2009) clearly defined what open data should look like in science with regard to licensing
- Data sharing agreements are legal documents that typically define (at minimum) how long data can be used, for what purpose(s), by what means, what constraints will apply, and financial, confidentiality and security requirements.

## OUTCOMES

- Very few researchers (around 15%) currently share their data outside a limited group of colleagues in any comprehensive and meaningful way (with notable exceptions, like astronomy, high-energy physics and genomics)
- Typically, networks allow only “approved” users to access information in order to limit risks of misuse (e.g., patient data)
- Complex and expensive processes are needed to standardize, clean, curate and preserve data and ensure interoperability

# CHALLENGES

What are the main challenges experienced by these three systems considered together (since they all act on each other)?

# RESEARCHER CHALLENGES

**The first and foremost challenge is addressing the needs and concerns of researchers.**

## THESE NEEDS AND CONCERNS FALL INTO SIX MAIN CATEGORIES: IMPACT, CONFUSION, TRUST, ACCESS, EFFORT AND REGIONAL DIFFERENCES.

1. IMPACT: Will my research benefit if I share my data? What benefit will I get from this personally? Will my open data efforts be well received by colleagues and tenure committees?
2. CONFUSION: Where to begin? What kind of license should be used? What data should be shared, in what format, with whom, and in what repository?
3. TRUST: Will my open data be misinterpreted or misused? Will my potential discoveries be scooped?
4. EFFORT: Will complying with data requirements take up too much time? Different publishers and repositories all have different compliance formats and requirements. Will I be responsible for maintaining it long-term?
5. ACCESS: Who needs access to my data and for what reasons? Some datasets are so large that they can't be uploaded via the Internet. For what purpose will my data be uses? Would data summaries suffice?
6. DIFFERENCES: There are wide variations in data needs by field, as well as wide variation in data processing capabilities by institution and region. Less privileged researchers often lack the huge support networks and processing facilities that more privileged researchers can take for granted. Why should they share their hard-earned data? Why should HSS researchers be excited about using a data system design to meet the needs of clinical trials researchers?

# RESEARCHER CHALLENGES

## EFFECTIVELY ADDRESSING RESEARCHER NEEDS WILL GO A LONG WAY TOWARD IMPROVING DATA SHARING PRACTICES AND OUTCOMES AND ALSO TOWARD ANSWERING QUESTION LIKE:

- How can more researchers be persuaded to share more data in usable format, given their concerns about secrecy and the efforts involved in sharing "meaningful" and reusable data?
- How do we broaden the circle of users who can access shared data (considering that sharing full and/or complete datasets is often impractical)
- How do we get researchers more involved in these conversations? This is the most underrepresented stakeholder group (although not for lack of trying). It is also a very heterogenous group with a wide variety of needs and perspectives.

# DATA CHALLENGES

## CHALLENGES REGARDING THE DATA ITSELF ARE ANOTHER MAJOR PRIORITY:

- How can we fund and maintain the infrastructure necessary for data processing, curation, and preservation?
- How do we protect against link rot, and data decay and data obsolescence over time?
- Big data keeps getting bigger. Can we keep pace with sharing tools?
- What happens to data once a research facility is shut down and data needs to be preserved and curated for decades more?
- What happens to long tail data?—the data that sits on laptops or personal websites with minimal or no attached metadata or documentation? Not being able to capture this contributes to issues like irreproducibility, duplicate research, and innovation loss.
- Who pays long term for data care and maintenance?
- How do we ensure the timely sharing of critical data (insofar as rapid sharing impinges on secrecy)
- How do we ensure better data quality, consistency and completeness
- How do we standardize data collection as a necessary approach to ensuring data completeness and comparability?
- How do create internationally agreed-upon minimum standards for metadata (further complicated when metadata are not in English)
- How do establish interoperability and searchability between data platforms (without which researchers need to search and make requests of multiple platforms)
- How do we create internationally agreed-upon standards for Data Availability Statements
- Can we streamline the governance structures used by different platforms
- Often (typically?), data platforms require registration and are only open to “qualified” users. Is this adequate?

# OTHER CHALLENGES

**FUNDERS**

- Very little funding support is available to facilitate data sharing, and to improve data infrastructure systems (and the differences between global regions are stark). What will it take to increase this investment 100-fold?

**CODE**

- How can we better share and preserve code? In many kinds of research, sharing or reanalyzing data without the original code means just sharing and preserving a jumble of numbers.

**POLICY**

- How can we address conflict between data sharing policies in science (e.g., GDPR conflicts have stalled several global research projects)?
- How do we create mandates for sharing that align with the needs and incentives researchers have? The overwhelming majority of researchers want the freedom to submit their work to the journal of their choice.
- How do we improve current policy? At present, NO open access policies require that the official version of record be archived in an open format. The only versions that are made available are the AAM; the VOR can be closed in a subscription journal.
- How will geopolitical tensions (especially with China) affect the future of research collaboration?
- How will infrastructure deficits in much of the world affect the ability of scientists from these regions to participate in the future of science?

**UNIVERITIES**

- Academia still doesn’t generally recognize, reward, or incentivize data sharing or team science in tenure evaluation processes (this is improving but still bad). How can these practices be improved?

**MEASURING SUCCESS**

- Our metrics for measuring the success of data sharing ventures are inadequate at the moment. For example, the number of peer reviewed publications flowing from shared data might be less important than whether this data gets used to inform study design, thereby reducing the need to put patients at risk.

# SOLUTIONS

What solutions and approaches are we using today?

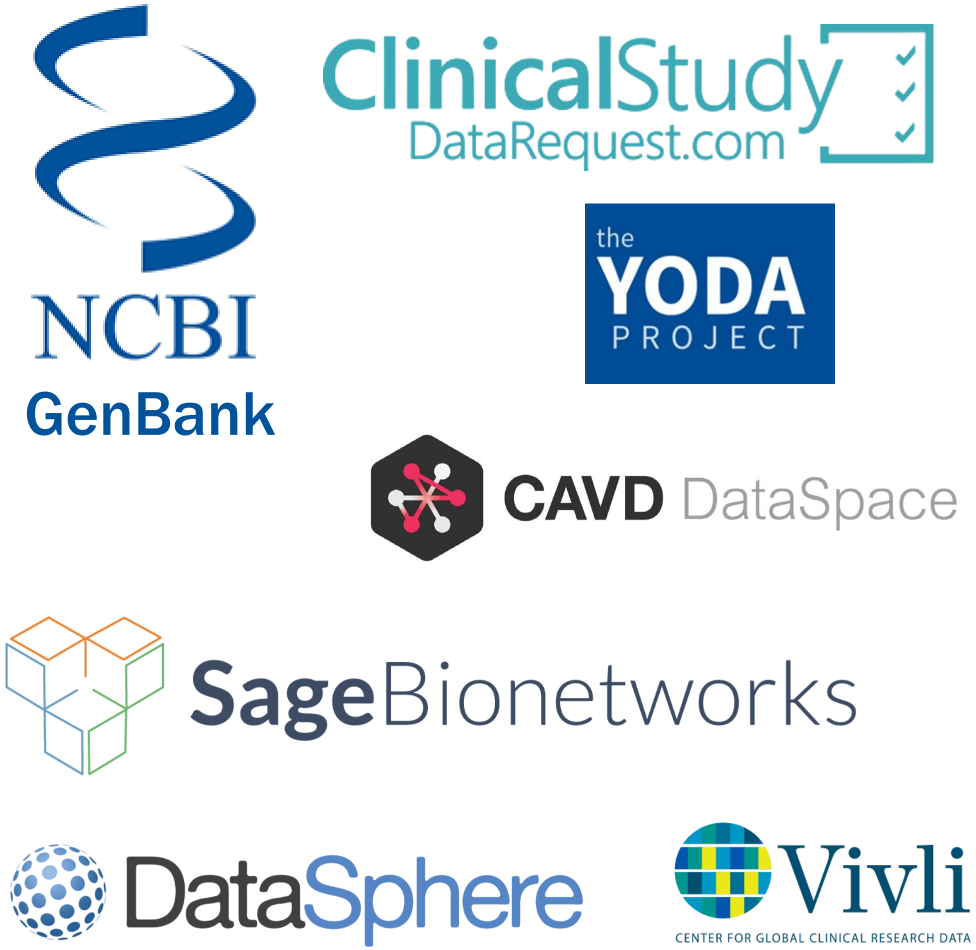
Archaeology Data Service
ArrayExpress
Australian Antarctic Data Centre (AADC)
Australian Ocean Data Network
Beilstein-Institut, STRENDA
Biological General Repository for Interaction Datasets *
Biological Magnetic Resonance Data Bank (BMRB)
BioModels Database
British Geological Survey
caNanoLab
Cell Image Library
ChEMBL
ClinicalTrials.gov
Coherent X-ray Imaging Data Bank (CXIDB)
Crystallography Open Database (COD)
Database of Interacting Proteins (DIP)
dbGAP
dbSNP
dbVar
DNA DataBank of Japan (DDBJ)
Dryad Digital Repository
EarthChem
EBRAINS
EKOS - TERN Ecoinformatics
Electron Microscopy Data Bank (EMDB)
Environmental Data Initiative (formerly LTER Network Information System Data Portal)
Eukaryotic Pathogen Database Resources (EuPathDB)
European Nucleotide Archive (ENA)
European Variation Archive (EVA)
figshare
FlowRepository
FlyBase
GenBank
Gene Expression Omnibus (GEO)
GenomeRNAi
Genomic Expression Archive (GEA)
Global Biodiversity Information Facility (GBIF)
G-Node
Harvard Dataverse
HEPData
HydroShare (CUAHSI)
Image Data Resource
ImmPort
Incorporated Research Institutions for Seismology (IRIS)
Influenza Research Database
IntAct
Integrated Taxonomic Information System (ITIS)
ioChem-BD Computational Chemistry Datasets
Japanese Genotype-phenotype Archive (JGA)
Kinetic Models of Biological Systems (KiMoSys)
KNB: The Knowledge Network for Biocomplexity
Magnetics Information Consortium (MagIC)
Marine Data Archive
Marine Geosciences Data System
MassIVE
Materials Cloud
Mendeley Data
Neuroimaging Informatics Tools and Resources Collaboratory (NITRC)
MetaboLights
MGnify
Morphobank.org
Mouse Genome Informatics (MGI)
Movebank Data Repository
NASA Goddard Earth Sciences Data and Information Services Center
National Addiction & HIV Data Archive Program (NAHDAP)
National Database for Autism Research (NDAR)
National Database for Clinical Trials related to Mental Illness (NDCT)
National Tibetan Plateau/Third Pole Environment Data Center
NCBI Assembly
NCBI PubChem BioAssay
NCBI PubChem Substance
NCBI Sequence Read Archive (SRA)
NCBI Taxonomy*
NCBI Trace Archive
NERC Data Centres
NeuroMorpho.org
NOAA National Centers for Environmental Information
NoMaD Repository
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)
Open Science Framework
openICPSR
OpenNeuro (formerly OpenfMRI)
OpenTopography
PANGAEA
PeptideAtlas
PhysioNet
PRIDE
Protein Circular Dichroism Data Bank (PCDDB)
Qualitative Data Repository
Rat Genome Database (RGD)
Research Domain Criteria Database (RDoCdb)
Science Data Bank
SEANOE
SICAS Medical Image Repository
SICAS Medical Image Repository (formally Virtual Skeleton Database)
SIMBAD Astronomical Database
Structural Biology Data Grid
Synapse
The Cancer Imaging Archive
The European Genome-phenome Archive (EGA)
The Network Data Exchange (NDEx)
UK Data Service
UK Solar System Data Centre
UNAVCO, Inc.
UniProtKB
VectorBase
World Data Center for Climate at DRKZ (WDCC)
Worldwide Protein Data Bank (wwPDB)
Xenbase
Zebrafish Model Organism Database (ZFIN)
Zenodo

Source: https://www.nature.com/sdata/policies/repositories

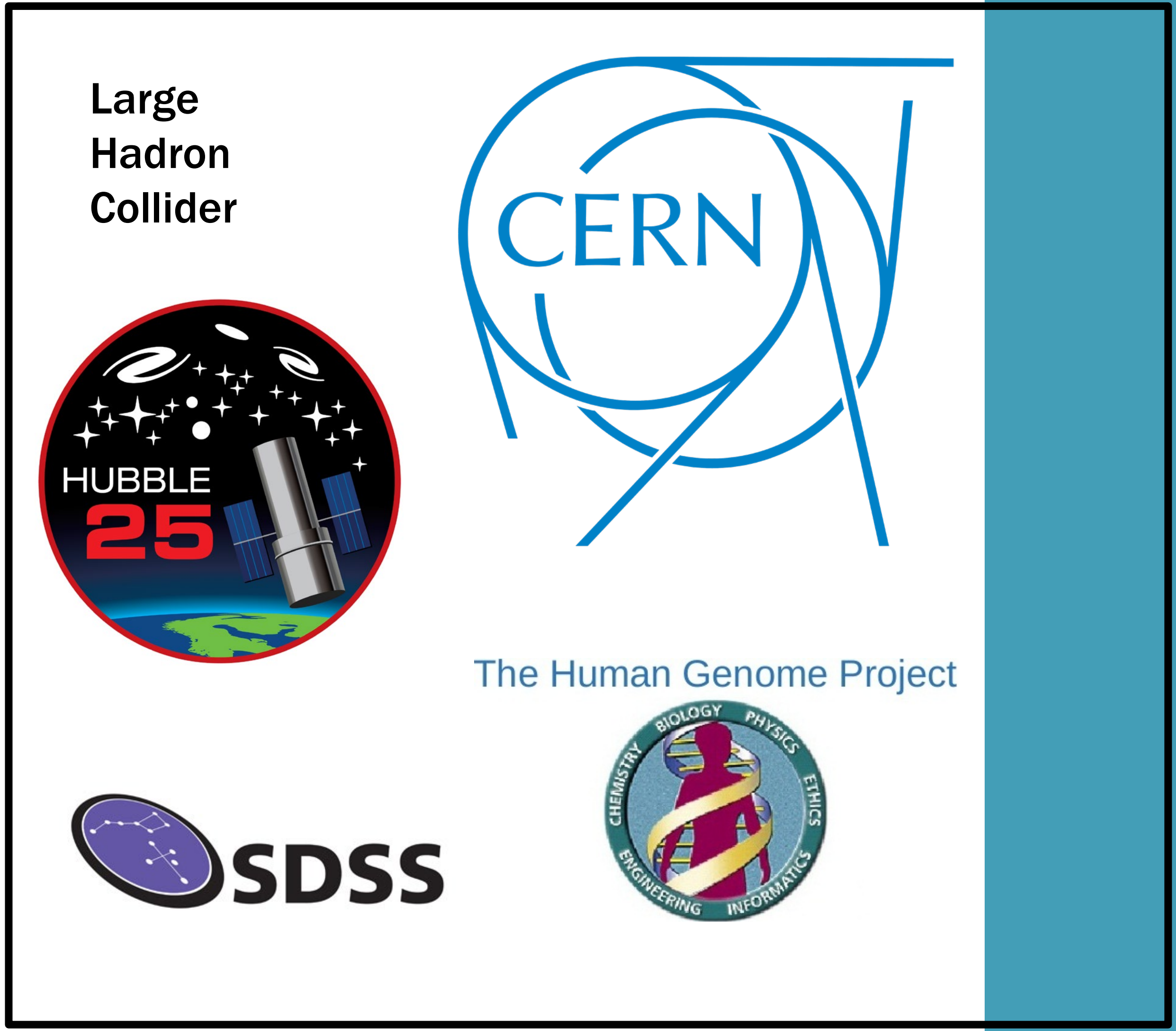
**There are a great many open data repositories currently being used in science (not even including institutional repositories)…**

RDA
RESEARCH DATA ALLIANCE
COAR
Confederation of Open Access Repositories
COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL
Jisc
NISO
OpenAIRE

**...as well as several outstanding organizations working to create best practices in data management and data repository function...**



…several outstanding examples of how real data sharing is working in today’s science environment…



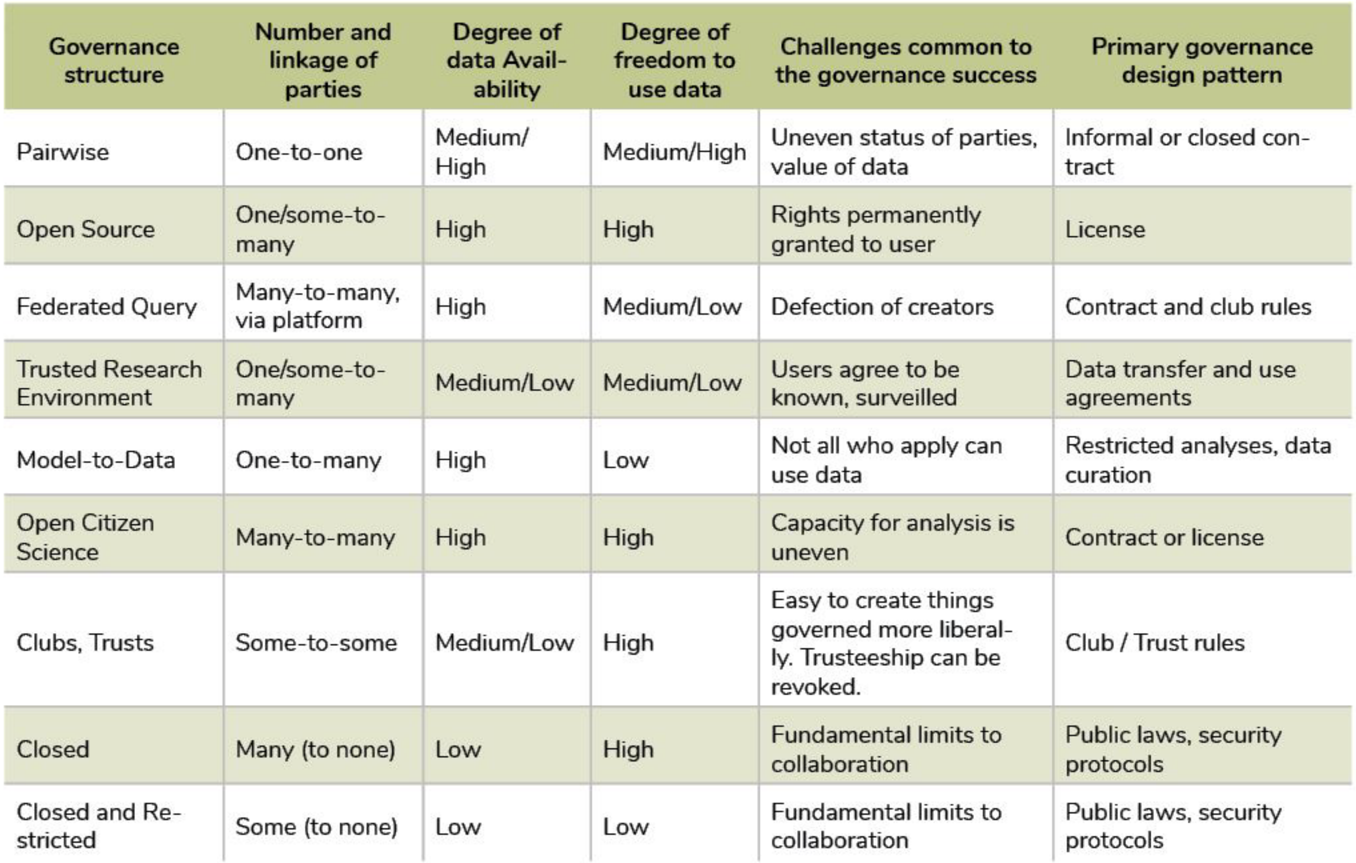
Large Hadron Collider

…a number of high-profile success stories in sharing science data…

# …and a number of widely-used approaches.

- Research sharing principles (like FAIR, DORA and Leiden)
- Open licensing (CC-BY and its variations in publishing, CC0 in data, and various licensing schemes for code)
- Transformative agreements between publishers and university systems for "read and publish" and/or "publish and read" journal publishing arrangements instead of subscriptions
- Growing use of mandates (from governments, universities and funders) for open licensing, limited embargo periods, data inclusion, etc. Data availability statements are now mandated by about a third of publishers and strongly recommended by another third.
- Growing use of tools and systems to catalogue science like Altmetric, Crossref (DOIs), Unpaywall, ORCID, and more. Flowing from these tools and efforts, we are attempting to create systems that give proper attribution to data (data citations).

# We have also learned a lot about the pros and cons of various data governance structures



| Governance structure | Number and linkage of parties | Degree of data Availability | Degree of freedom to use data | Challenges common to the governance success | Primary governance design pattern |
|---|---|---|---|---|---|
| Pairwise | One-to-one | Medium/ High | Medium/High | Uneven status of parties, value of data | Informal or closed contract |
| Open Source | One/some-to-many | High | High | Rights permanently granted to user | License |
| Federated Query | Many-to-many, via platform | High | Medium/Low | Defection of creators | Contract and club rules |
| Trusted Research Environment | One/some-to-many | Medium/Low | Medium/Low | Users agree to be known, surveilled | Data transfer and use agreements |
| Model-to-Data | One-to-many | High | Low | Not all who apply can use data | Restricted analyses, data curation |
| Open Citizen Science | Many-to-many | High | High | Capacity for analysis is uneven | Contract or license |
| Clubs, Trusts | Some-to-some | Medium/Low | High | Easy to create things governed more liberally. Trusteeship can be revoked. | Club / Trust rules |
| Closed | Many (to none) | Low | High | Fundamental limits to collaboration | Public laws, security protocols |
| Closed and Restricted | Some (to none) | Low | Low | Fundamental limits to collaboration | Public laws, security protocols |

Mangravite, L., A Sen, JT Wilbanks. 2020. Mechanisms to Govern Responsible Sharing of Open Data: A Progress Report https://sage-bionetworks.github.io/governanceGreenPaper/v/7ef288619fb46e6d9319433c64fbc5bef6250fe7/

# AND YET...

## Despite this evident activity, need, and potential...OUR GLOBAL DATA SHARING EFFORTS ARE UNDERFUNDED AND FOCUSING ON THE WRONG ISSUES

- There is very little funding or attention in this space. Our data sharing success stories in fields like high-energy physics, astronomy, and genomics, have one primary thing in common: Big money.
- Our approach to data sharing tends to be ideologically-driven — one-size-fits-all approaches that aren't informed by real-world data sharing models and lessons of experience. We pursue open science policies as though they are goals unto themselves, but what are we actually accomplishing without also breaking down the barriers to effective data sharing and building up the necessary capacity for sharing?

# In the meantime, we increasingly recognize an urgent need to work together

- Climate change
- Pandemics
- Food and water security
- Deforestation
- Women’s rights
- Early childhood education
- Global economic development
- Migration

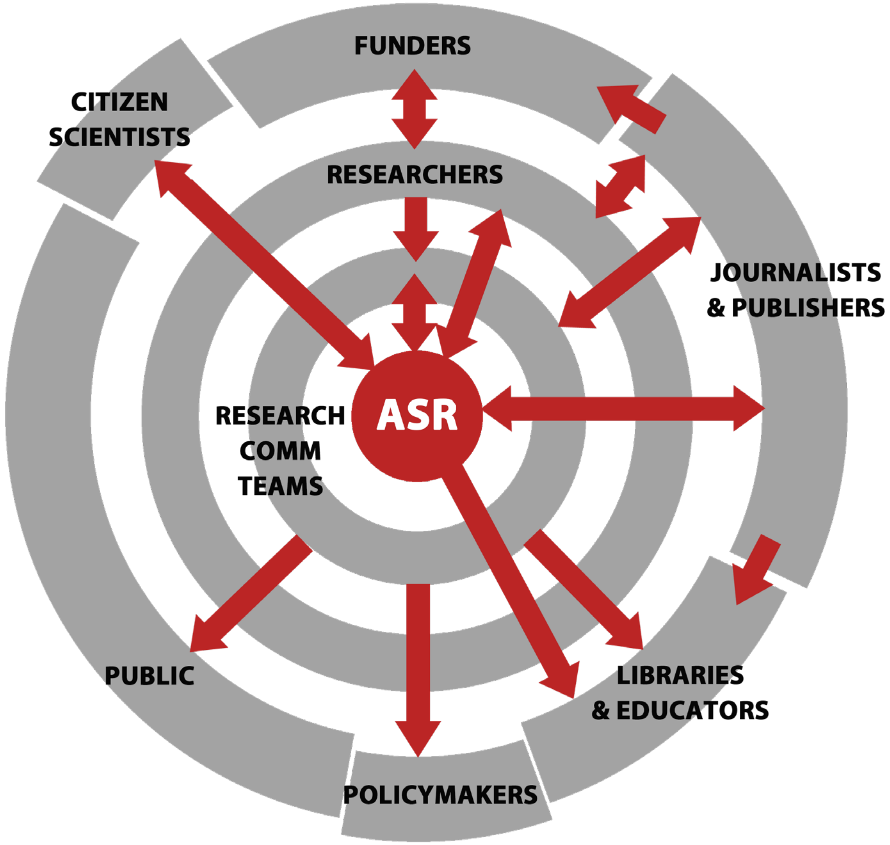
# KEY ACTIONS

How do we get to a better future?

# IS THE ANSWER OUT THERE?

The EU Science Cloud (on the horizon) aims to create a hugely interlinked science repository for the EU. But is this enough? An "All-Scholarship Repository" (ASR) approach would go one step further and instead of linking the thousands of government and institutional repositories currently in use, would instead replace these with one ultra-high functioning repository—a single Amazon for all things research—and would also simplify the flow of research information. All information—not necessarily in research paper format but in all kinds of reporting formats—would simply be deposited into ASR and picked up from there by publishers and other researchers. This information dump would be analogous to simply filing information on a computer, but in an organized and standardized way.

For now, to deal with long tail data (and since doing this for each project in isolation is inefficient and impractical), some universities are starting to build research data repositories to help researchers manage and preserve this data (e.g., Dspace at MIT and Dataverse at Harvard).



Source: Hampson 2020

# WOULD $ HELP?

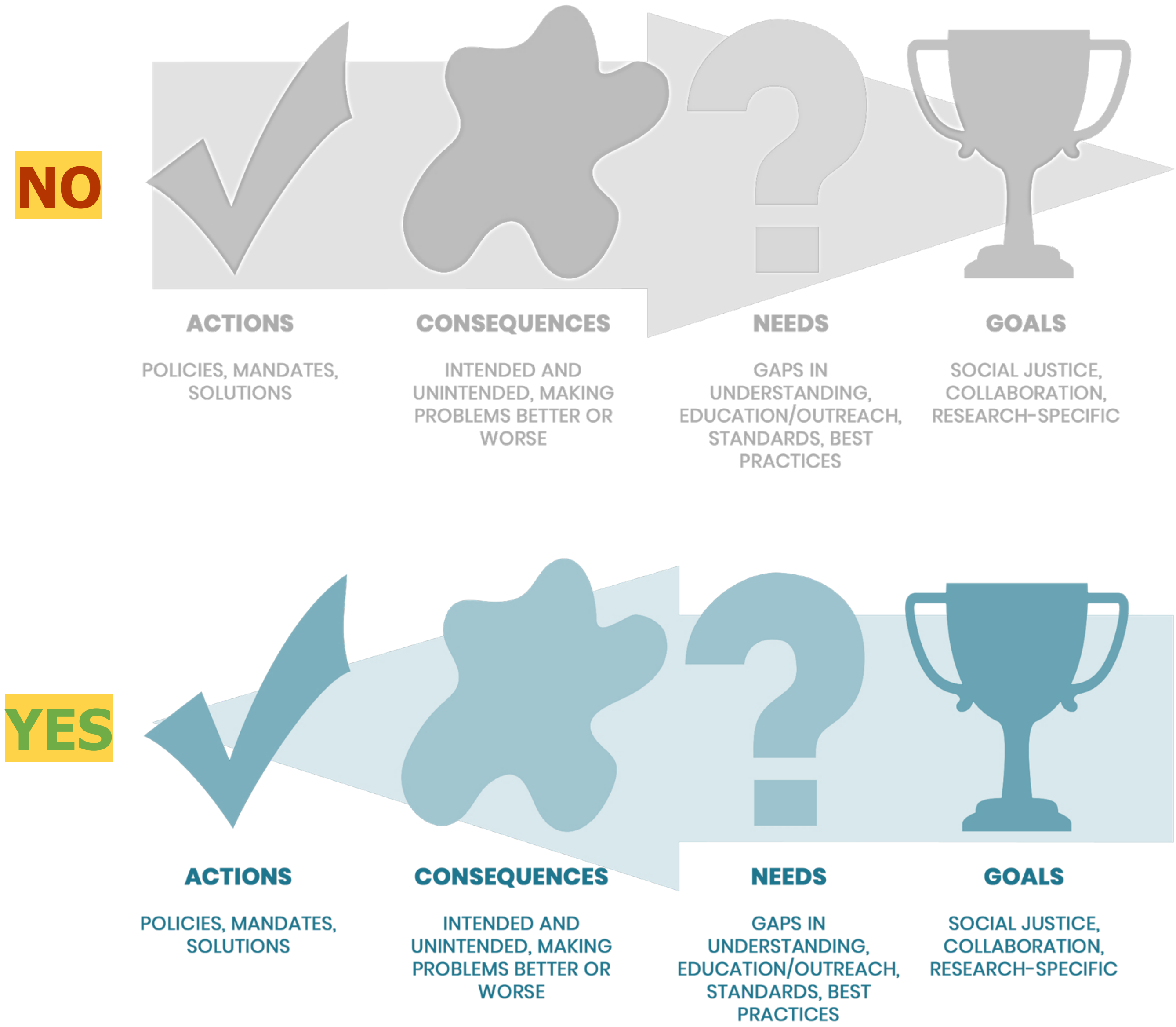
Would a huge influx of money help? If so, from where? And to whom? To a single agency charged with overseeing the future of science data sharing? To one large demonstration project in science data sharing? In just one field?

At present, this is the “winning the lotto” type of concern—there isn’t any serious talk about making these kinds of investments, but they may be what’s needed.

It may be that a private effort like the Chan Zuckerberg Initiative or Google will suddenly come forward with a category killer solution. This may be more likely than expecting governments or international agencies to make the necessary investments.

# UNTIL THEN, IT WOULD HELP TO REVERSE OUR APPROACH TO OPEN SOLUTIONS...

For the past 20 years, our approach to open has been driven by ideology. We have designed our open solutions first, and then tried to sell these solutions to researchers, downplaying unintended consequences, and ignoring the need for a more complete understanding of the open space. Reversing this process is important.



NO

ACTIONS

POLICIES, MANDATES, SOLUTIONS

CONSEQUENCES

INTENDED AND UNINTENDED, MAKING PROBLEMS BETTER OR WORSE

NEEDS

GAPS IN UNDERSTANDING, EDUCATION/OUTREACH, STANDARDS, BEST PRACTICES

GOALS

SOCIAL JUSTICE, COLLABORATION, RESEARCH-SPECIFIC

YES

ACTIONS

POLICIES, MANDATES, SOLUTIONS

CONSEQUENCES

INTENDED AND UNINTENDED, MAKING PROBLEMS BETTER OR WORSE

NEEDS

GAPS IN UNDERSTANDING, EDUCATION/OUTREACH, STANDARDS, BEST PRACTICES

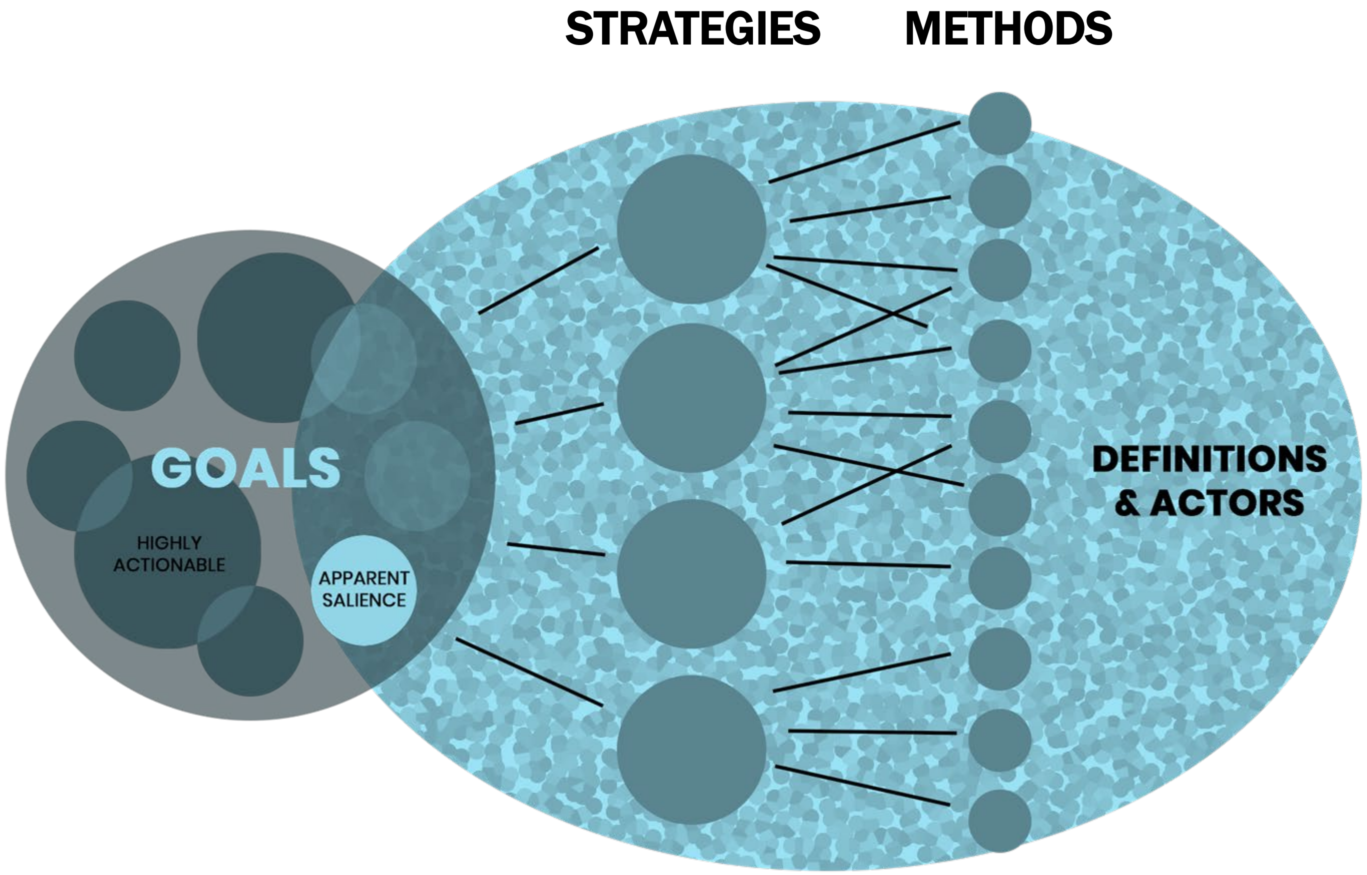
GOALS

SOCIAL JUSTICE, COLLABORATION, RESEARCH-SPECIFIC

Source: Hampson 2020

# …IDEALLY WORKING TOGETHER TOWARD COMMON GOALS

A goals-based approach identifies the long-term changes our broad community desires, and then works backward, together, to map out the actions and policies we need to create this change. By focusing on common goals first, we work together in ways that maximize our mutual benefit across our many differences. The goals-based (Theory of Change) approach is widely used in business, governments, and the United Nations.



Source: Hampson 2020

# IN DOING SO, WE MIGHT SOMEDAY REACH AN“OPEN RENAISSANCE”

## IF WE DO THIS….

- **Clearly define and support** open
- **Make open solutions robust**, inclusive, broad, scalable and sustainable
- **Resolve connected issues** (e.g., impact factors)
- **Align incentives** so scholars embrace open because they want to
- **Make open simple and clear** so scholars know what it means and why they should do it
- **Create clear standards and guidelines**
- **Keep the marketplace competitive** so open products remain cutting edge
- **Integrate open repositories,** not just connect them
- **Standardize data**

## THEN WE GET THIS….

- **The research ecosystem grows more powerful** (with more data, more connections, and more apps),
- **Innovation is catalyzed**
- **Widespread improvements happen in science.**
- **New fields and discoveries emerge** based on “connecting the dots” (thanks to data and repositories)
- **Funding efficiency improves**
- **Discovery accelerates**
- **The social impacts of science surpass today** (including science literacy, public policy, education, more)

# THANK YOU!

Questions? Email Glenn Hampson at ghampson@nationalscience.org

For more information, please visit the OSI website at osiglobal.org

Suggested citation: Hampson, G. 2021. Competition, Collaboration & Data Sharing in Science: An Overview of Ongoing Challenges. VI BRISPE conference.

The opinions in this presentation are the views of the author and are are not an official representation of the views of SCI, OSI, UNESCO, or any individual or institution connected to these organizations.

SCI

OSI

# Sources and additional reading

Cheng, F., Ma, Y., Uzzi, B. et al. 2020. Importance of scientific collaboration in contemporary drug discovery and development: a detailed network analysis. BMC Biol **18,** 138. https://doi.org/10.1186/s12915-020-00868-3

Davies, T, SB Walker, M Rubinstein, and F Perini (eds). 2019. The State of Open Data: Histories and Horizons. African Minds, IDRC. ISBN 9781928331957. Book pdf from https://www.idrc.ca/en/book/state-open-data-histories-and-horizons. HTML rendition at https://www.stateofopendata.od4d.net/

Hampson, G, M DeSart, L Kamerlin, R Johnson, H Hanahoe, A Nurnberger and C Graf. 2021. OSI Policy Perspective 4: Open Solutions: Unifying the meaning of open and designing a new global open solutions policy framework. Open Scholarship Initiative. January 2021 edition. doi: 10.13021/osi2020.2930.

National Academies of Sciences, Engineering, and Medicine 2020. Reflections on Sharing Clinical Trial Data: Challenges and a Way Forward: Proceedings of a Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/25838.

National Academies of Sciences, Engineering, and Medicine. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25116.