

BioIT World, April 17th, 2024

Best Practices to Solve Data Science Problems that Slow Down Scientific Research

Ari E. Berman, Ph.D. & Simon Twigger, Ph.D.

BioTeam
Accelerate Science

Introductions and outline

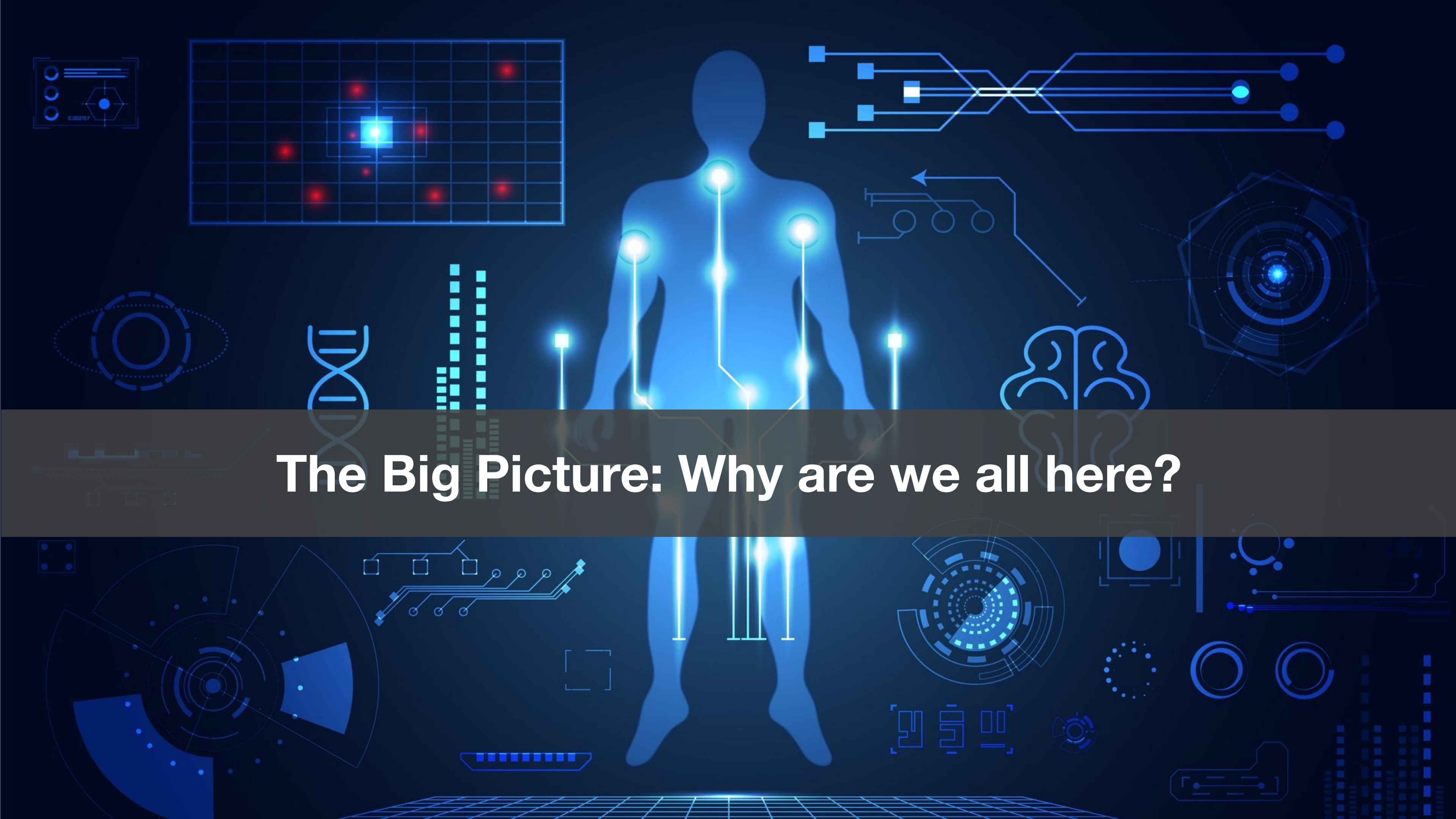
Ari E. Berman, Ph.D.,
CEO,
BioTeam, Inc.

- Neuroscientist/Molecular Biologist/Computational Biologist
- 30 years building HPC for science
- BioTeam for 12 years - Bio-IT World for 12 years

Simon Twigger, Ph.D.,
Principal Consultant,
BioTeam, Inc.

- Biochemist then Bioinformatics, Genomics, Proteomics, Clinical informatics.
- 25+ years building software and systems for the life sciences
- BioTeam for 11 years

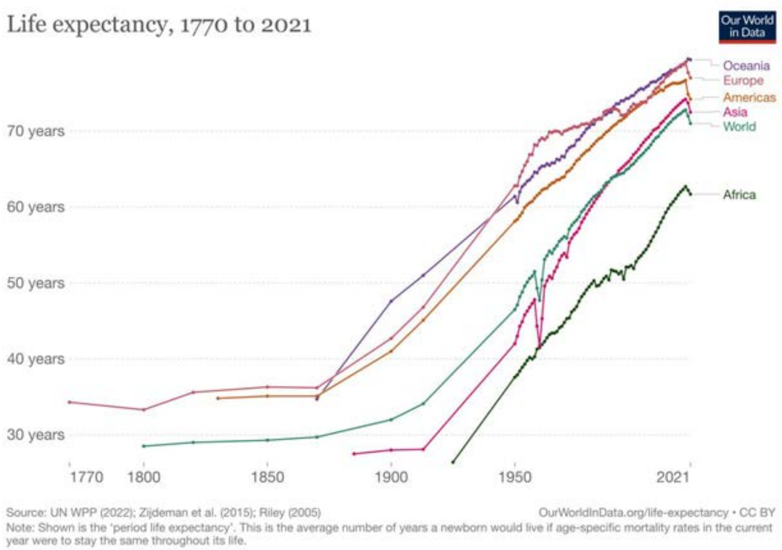
The Big Picture: Why are we all here?



Scientific Discovery Drives Modern Medicine

Science: *the systematic study of the structure and behavior of the physical and natural world through observation and experiment*

- Human lifespan hit a maximum in 2019 (76.7 years US)
- First major inflection in 1865 – Germ Theory, antiseptics in surgery, washing hands (doubled lifespans)
- Vaccinations, epidemiology, anesthesia, antibiotics are others
- All driven by basic life sciences research and applied to human health



Long Lifespan and Healthspan Now Out of Sync

- Healthspan – the percentage of life that one is considered healthy
- "Healthy" means different things to different people
- Average age of health decline in US is 63yo vs. 76.7-year lifespans
- Live nearly 20% of our lives unhealthy – with lower quality of life
- Resulted in soaring healthcare costs and increased burden on medical and economic systems as the population ages poorly

Disease	Deaths per year	Age of 1st occurrence
Heart Disease	610,000	65
Lung cancer	158,060	60
Chronic obstructive pulmonary disease (COPD)	147,101	45
Stroke	140,000	65
Lower respiratory infections	131,800	75
Alzheimer's disease	93,541	65
Type 2 diabetes	69,071	54
Colorectal cancers	50,260	70
Breast cancer	40,000	62
Prostate cancer	25,000	66

Table 1: Top ten causes of death in the US and their average or median age of first occurrence.

Medical Science Needs a New Focus

- Historically, medicine has focused on reducing risk of death as an outcome
- Medical science has advanced to the point where that focus isn't as relevant anymore
- Need new breakthroughs and renewed focus on being healthy for longer, vs being alive longer (likely linked)
- Modern diagnostic and computational methods have already started driving towards this reality

Drive Towards Precision Medicine is Data Intensive

- Move away from the “one-size-fits-all” approach to health care delivery and to instead tailor treatment and prevention strategies to people's unique characteristics, including environment, lifestyle, and biology
- Medical decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease
- Goal of increased Healthspan
- Foundational data initiatives like All of Us, CANDLE, TCGA, InsightRX, and many more
- Involves a large amount of data collection, modeling, and classification to pull off

Biomedical Research Grand Challenges

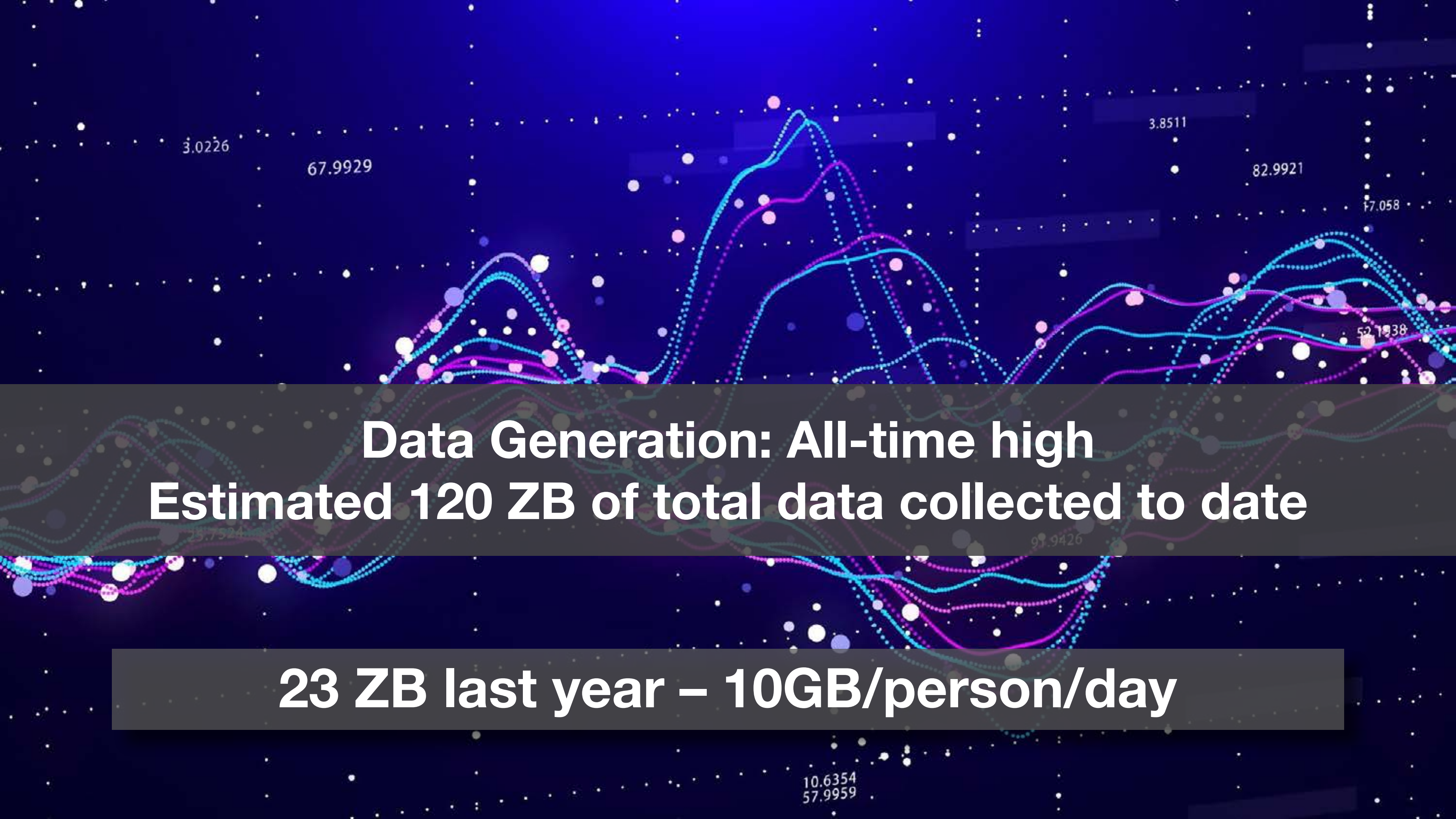
Context for this workshop, driven by what we're trying to help the community do better

Grand Challenges

Current state	IT support for science inconsistent, not aligned with science	Fractured data ecosystem - minimal standards/FAIR	Advanced analytics are built by data scientists for experts
Effect on science	Infrastructure is a barrier, not an asset, scientists do it themselves	Enormous productivity sink just getting data ready for analysis - 80% of time	Data scientists become a bottleneck, lab scientists can't do analysis on their own
Grand challenge	Create advanced infrastructure for science - abstracted from users	Unify Biomedical Data Ecosystem	Make advanced analytics accessible to the long-tail of scientists

Science goes faster!

Infrastructure in Life Sciences and Healthcare: Scoping the problem



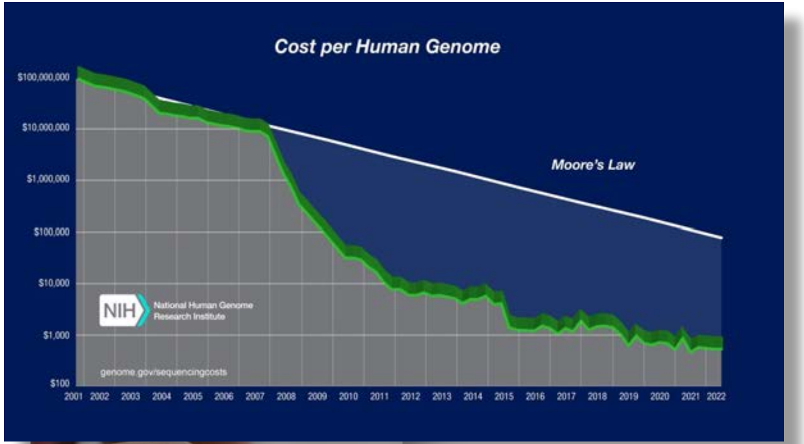
Data Generation: All-time high
Estimated 120 ZB of total data collected to date

23 ZB last year – 10GB/person/day

Laboratory and Health Diagnostics Innovation Accelerating

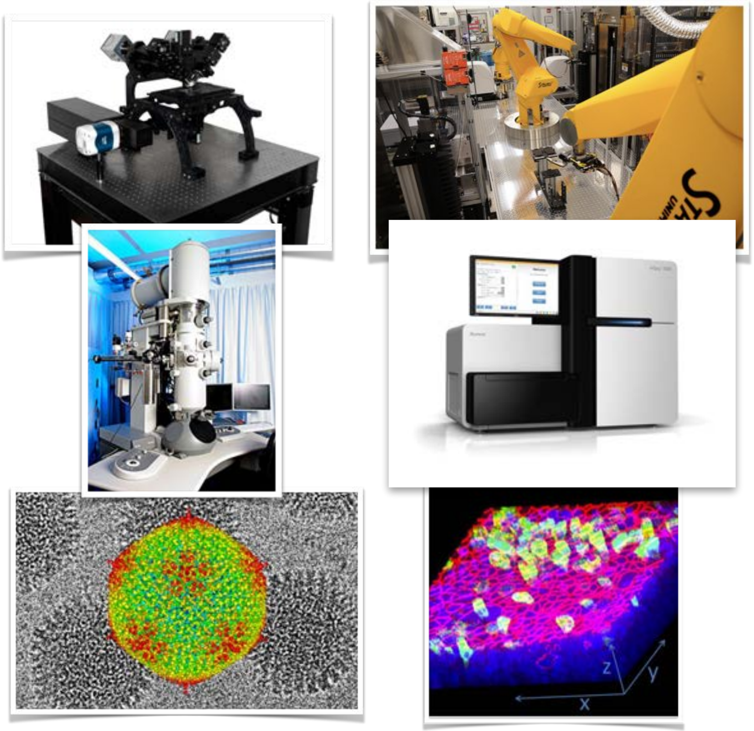
Rate of innovation in data generating equipment in life sciences far outpaces Moore's Law

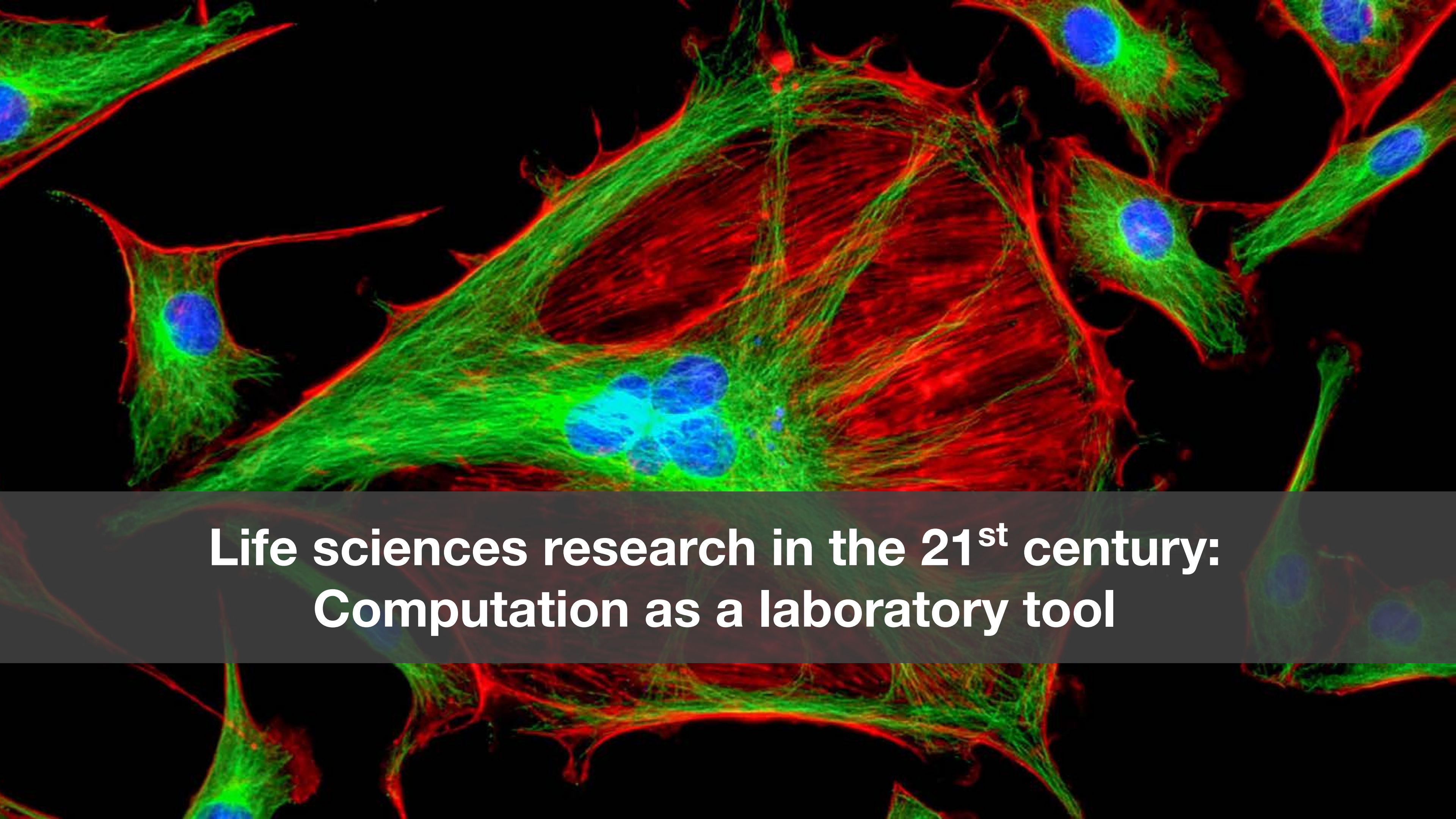
- Laboratory technology changing on the order of months
- HPC and IT tends to move on the order years – equipment lifecycles are 3-6 years
- Cloud has enabled more rapid cycling, but at increased cost and complexity
- Health diagnostics and IoT sensor data pushes amount and complexity of data



Laboratory Data Generation at an All-time High

- Sequencers were just the beginning
- Life sciences and Healthcare among the top data generators of all the sciences
- Most labs have equipment that can generate 10s TB of data/week
- Bioinformatics/Data Science have taken over biomedical analytics space
- Computational sophistication required to compete
- Clinical informatics becoming a major health diagnostics tool





Life sciences research in the 21st century: Computation as a laboratory tool

Infrastructure for Biomedical Research

Infrastructure a Critical Issue in 2024

- Big Data era led to huge data generation and now AI/ML
- Sophistication needed for data intensive science is high
- Backlog in data to be analyzed, slow to interpret due to state of data
- Stress on existing resources
- Led to a decrease in discovery and productivity

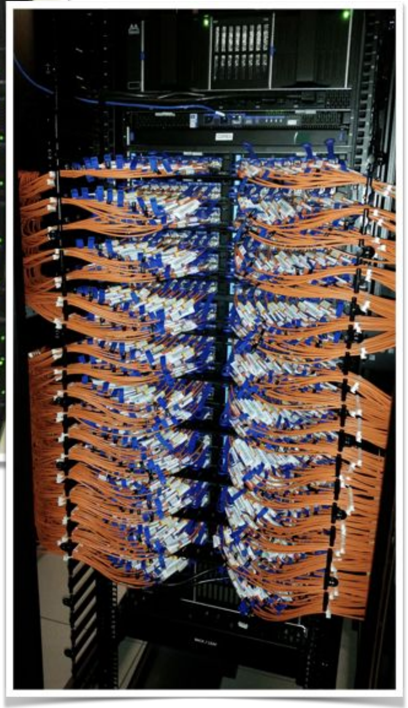
The pandemic changed the game

- Everyone went home into isolation
- Labs shut down – researchers started remotely analyzing data
- On-premises HPC and compute systems ramped up usage a lot – users that wouldn't use it that often were now clamoring for access and hours on systems
- On-premises infrastructure became expensive, hard to manage remotely
- Supply chain drove up prices, slowed delivery – local IT barely maintaining status quo (still)

On-prem Compute: Increasing in Complexity

Modern Scientific Computing

- Used to be “figured out”, not so simple anymore
- Explosion of Data Science, huge datasets, accelerators and co-processors (GPUs!!!!!), cloud
- AI/ML in particular
- Caused a previously simple set of decisions to become extremely complex
- Also, diversification of CPU, back-end network options, matched to use cases

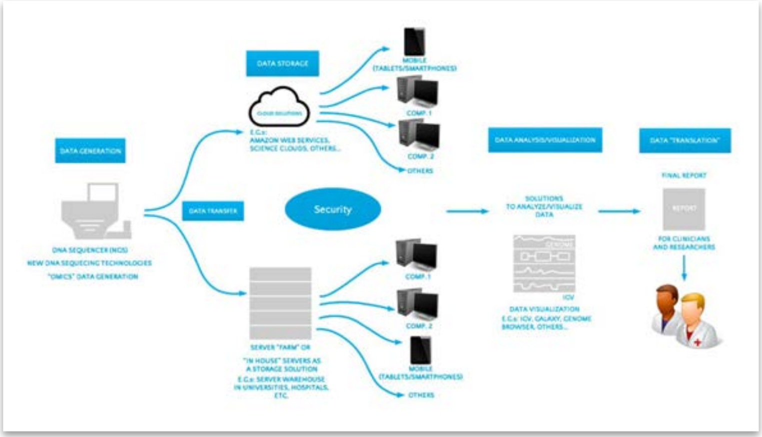


Networking Challenges

You have to move the data!

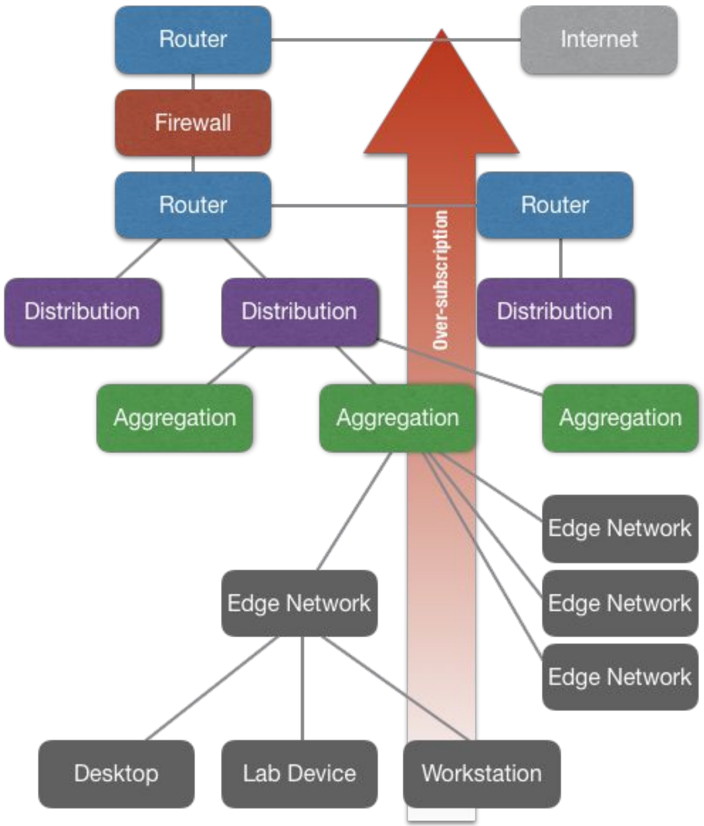
A lot of talk about moving compute to data, but you always have to move it at least once

- Laboratory equipment only able to store last few experiments
- Not designed for any analysis past initial data reduction
- Have to move it to compute storage for processing
- Many labs can generate 1PB/year or more – Non-trivial network requirements
- Data is valuable, needs to be protected, replicated, or backed up
- Data sharing is a standard requirement



Enterprise Networks

- Viewed as a cost to be controlled
- IT in general has flat budget, tight fiscal control
- Networks optimized for web/email traffic, not for large sustained data transfers
- Results in highly under- provisioned and oversubscribed network deployments
- Expertise split and siloed - networking org isolated
- Really need at least 10Gb from local data storage to Cloud for data intensive science
- Security designed for risk, not science



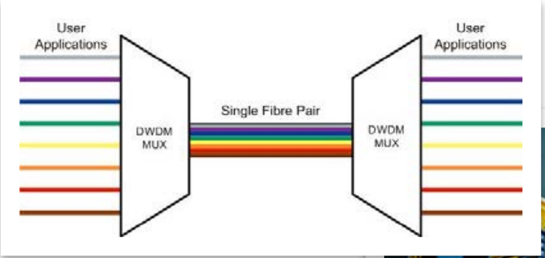
More on Security…

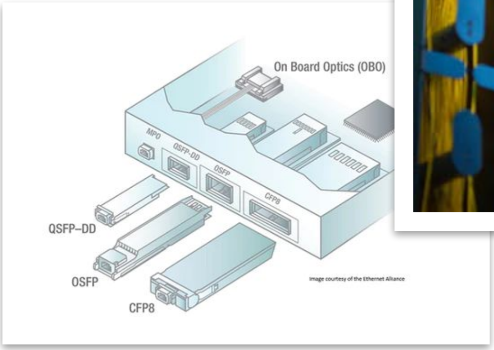
- Enterprise security usually universally applied to all use cases
  - Security designed for highest risk data in network
  - Usually smallest percentage of data
- Malware in a genome file is extremely unlikely – packet scanners don't know what a BAM file is
- Slows down transfer, scans every packet for 300GB files with no useful outcome
- Takes the sledgehammer approach, then all is covered
- Seriously hampers data sharing, movement, reference datasets, etc.

Most common high-speed network - STILL

Networking: key for modern biomedical science

- Modern science requires 100Gb data speeds (1TB transfer in 1.5min*)
- Next-gen filesystems require 100Gb networking to function
- 400Gb current standard
- 600-800Gb Optical Transport Networks are out now
- 1 Tb networks in early release
- If you aren't thinking of high-speed science for your network, you are hampering the science of your org





Cloud Challenges

Cloud providers launched a BRILLIANT marketing campaign: convinced decision makers to go cloud.

People Tend to Think in Absolutes

- A solution can only be this or that
- If cloud, then ALL cloud
- Abandon things that work well, because this will be better!
- Reasons:
  - Everyone's doing it
  - It's cool so people will invest in it, support it
  - People like to talk about it a lot
- Reality:
  - Very nuanced situation, there are no absolutes, and it isn't simple

Aggressive Cloud Migration Programs were Started

- Most science organizations started planning for cloud-first or all-cloud transitions away from local infrastructure
- Cloud providers gave deep discounts and a lot of direct support during the planning and the migrations to cloud
- Stopped planning on local infrastructure: storage, HPC, even networking
- All future planning was for cloud-based architectures
- We've seen this movie before…

A Look Back: The Last Wave of Cloud Migrations

- 2008 – 2014: we helped several orgs migrate to the cloud and close their datacenters (AWS)
- The draw: cheap, easier to manage, endless supply of compute power, less staff needed
  - Better shared access for external collaborations
  - Better access to public datasets
- The reality:
  - 10-50x the cost of operating datacenters
  - Cloud couldn't replace all local infrastructure
  - Required specialized skillsets in IT, harder to use
  - Didn't meet scientists' requirements
- The Result: Cloud sobriety – massive pullback

Ah, what short memories we have

Now it’s Happening Again…

- …for slightly different reasons this time:
  - There are more cloud providers, way more sophisticated, clouds are largely designed to handle enterprise needs now
  - Competition has forced huge innovation in cloud services
    - There are aspects of cloud that you can’t reproduce locally
    - Deep learning applications and specialized hardware/services are huge draws
  - There’s so much data now, that storing it locally is non-trivial
  - Promise of it being somehow cheaper to operate
- Flexibility of cloud architectures combined with supply chain issues for local hardware made moving that direction more attractive

Also: Cloud business model locks your data/operations into their platforms

- The most successful ransom scheme in history
- Free to put data in, charged to get it out
- Save money by using proprietary (and very useful) services like:
  - Serverless
  - Bulk operations services
  - Specific analytics services
- Locks you into using their services – can't reproduce them yourself (e.g. – Parler)

Advantages of the Cloud

- Large capacity, automatic upgrades, reliability
- Virtual orchestration and serverless technologies
  - Can architect very sophisticated environments more easily
- Containerization and portability of workloads
- Share data more easily, and in more sophisticated ways
- Shared standards in architecture, great way to prototype
- Availability of accelerators, like GPUs
  - Though, they aren't so available, really
- Specialized services that can speed up deployments

Relevant Questions: To Cloud or not to Cloud?

- Can you do real HPC in the cloud?
  - Yes, now you can, but depends on what kind you need
- Can you create a secure environment?
  - Absolutely, but it's yours to mess up. The models are different than on-prem
- Can scientists use it out of the box?
  - Absolutely not: requires set up of services, user interfaces, platforms
  - Even the sophisticated users need to architect their environments
  - Some things are there out of the box, but they're basic
- Can we use it for storage?
  - Definitely: but storage in the cloud is complex, hard to price, and easy to overdo
  - Once your data is in the cloud, it costs money to get it out again

So, Why not All Cloud?

- Fundamental mismatch between cloud business model and long-term research goals
- These are private, for-profit companies, not public utilities. They offer services that they can change at any point (and do!)
- Shortest full scientific study, 5-10 years. Some 100s of years
- Clouds don't even disclose time period if they shut down to get your data back – this should concern you
- Hybrid computing model preserves data ownership – source of truth kept locally
  - Copy data to the cloud for analysis (ingress – no charge)
  - Only copy results of analyses back – delete source data (minimize data egress fees)
  - Still not solved!!

Let's Talk About Data


How are we doing with FAIR so far?
Not great—limited pockets of excellence

What's blocking FAIR?

- Short answer: people, not technology
- Long answer—it's complicated:
  1. It's really hard to do, extremely complex—lots of historical data—but it can be done
  2. Lack of unified data standards in the field
  3. Lots of distraction, lack of sophistication, lack of tools and training, lack of accessibility of solutions
  4. No incentive to make the effort—all incentives are individual (i.e., NIH grants), this is a community effort (some EU efforts towards open data)
  5. It pays to do your own thing—your funding gets diluted if you join with groups of collaborators
  6. No clear vision, leadership, or directives from funding organizations to get there

Lack of unified data standards

- Biomedicine and life sciences research lack unified data standards
  - Very difficult to combine data from multiple sources for deeper analysis
- Clinical information largely contained in electronic health records (EHR) systems
  - Not designed for analytics, designed for compliance
  - Different systems store and make data available differently—far from FAIR
  - Deidentifying data is difficult and loses key information, while improving access
- Historical data locked in publications
- Each domain and each project has its own format/standards

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

HL7 FHIR

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

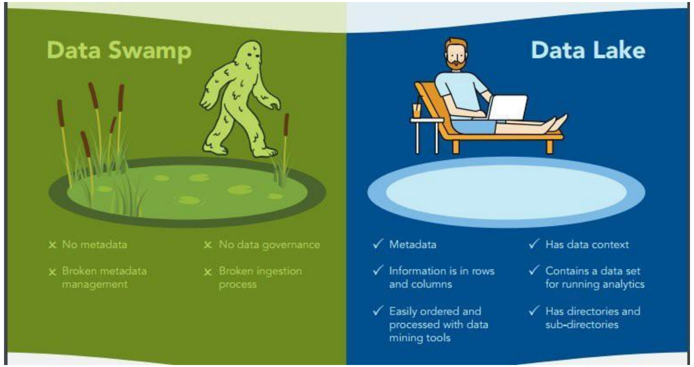
MIAME and ArrayExpress
- a standard for microarray data annotation and a database to store it

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

RDA
RESEARCH DATA ALLIANCE

Data platforms—they're everywhere!

- Another buzz-loaded space:
  - Data Lakes, Oceans, Fogs, Swamps, Islands, Universe, Warehouse, Commons, Ecosystem, Mesh, Fabric…
- Truth: just building more silos of excellence—and not accessible to long tail of scientists (laboratory)
- Need a diversity of tools and approaches, but they aren't cross-compatible
- Lack of data standards and interoperability lose the power of the data in collaboration



GEN3 DATA COMMONS

Flywheel

Terra

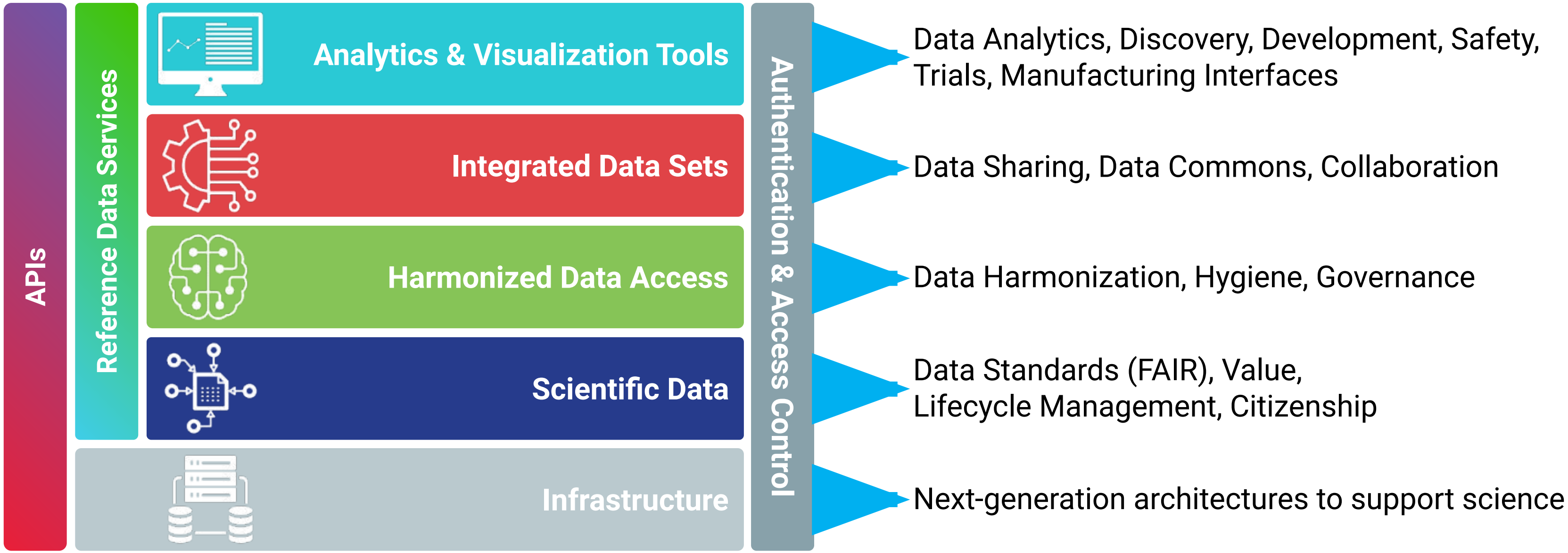
Palantir

VELSERA

Data value in life sciences: key for digital transformation

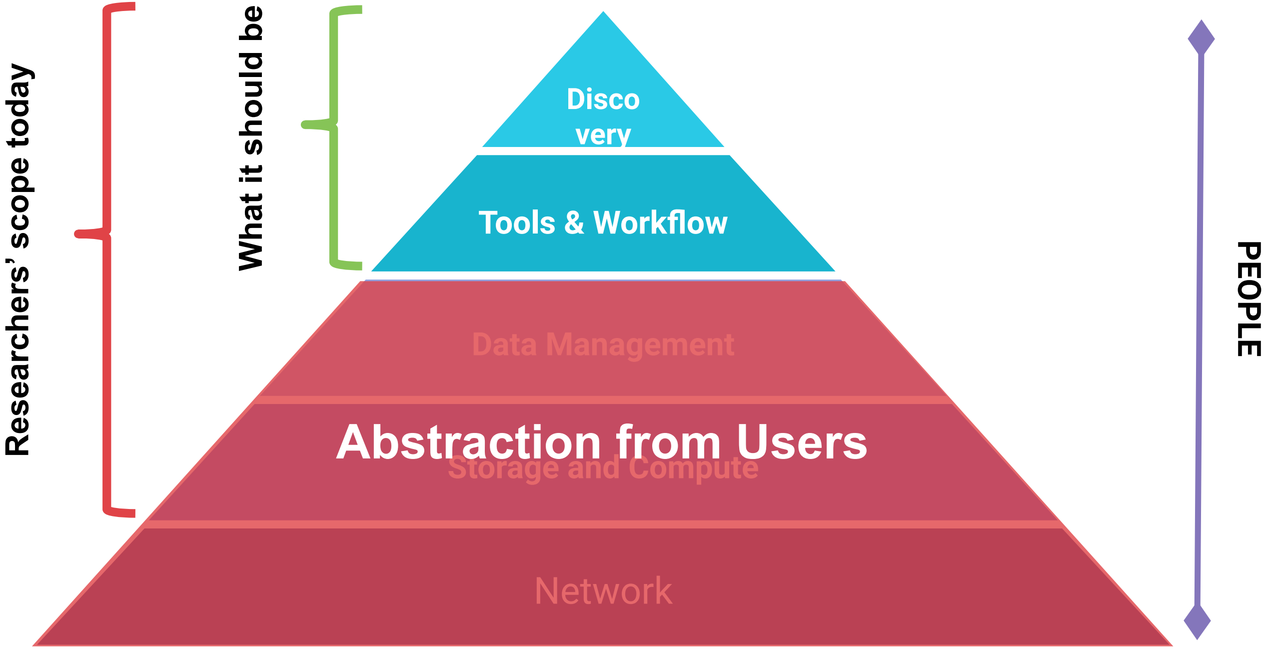
- In general—no understanding of scientific data value
  - Investment—how much is data worth, what did it cost to generate
  - Scientific value—will this data ever be reused, how long do we keep it, what's the impact of each unit of data?
- Without this understanding, everything is high value so we just keep everything
- Infinitely expand storage—hoarding
- Can't prioritize investments in data that are aligned with strategy
- FAIR also doesn't contemplate data quality

Scalable data structure strategy



Making Science Accessible

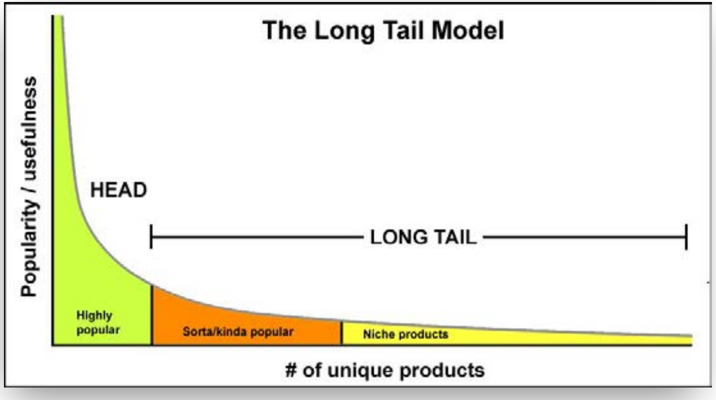
Maslow's hierarchy of IT needs



Research IT: Integrated spectrum of infrastructure, software, services and support, focused on science

Accessible Advanced Analytics

- Most life sciences researchers and clinicians lack computational sophistication (can barely print and use email). Systems need to be designed to be accessible by that population.
- Most infrastructure and analytics platforms target people who have a lot of expertise (i.e., know APIs, how to code, understand how to use HPC).
- This creates a barrier for the long tail of science, and a bottleneck for efficient discovery.
- Funding organizations need to incentivize and require data standards.
- Design systems that non-sophisticated users can use – (e.g. – Jupyter Notebooks = not accessible)



Practical Tools and Approaches

aka ‘unleashing the possibilities’

Grand Challenges

Current state

IT support for science inconsistent, not aligned with science

Fractured data ecosystem - minimal standards/FAIR

Advanced analytics are built by data scientists for experts

Effect on science

Infrastructure is a barrier, not an asset, scientists do it themselves

Enormous productivity sink just getting data ready for analysis - 80% of time

Data scientists become a bottleneck, lab scientists can't do analysis on their own

Grand challenge

Create advanced infrastructure for science - abstracted from users

Unify Biomedical Data Ecosystem

Make advanced analytics accessible to the long-tail of scientists

Science goes faster!

To address the challenges, something needs to change!

Our storage is full/broken/old/ not backed up, etc.

We want to move on/off the Cloud!

The scientists aren't happy with IT.
(and probably vice versa!)

We don't have access to enough compute.

Moving our data is too slow!

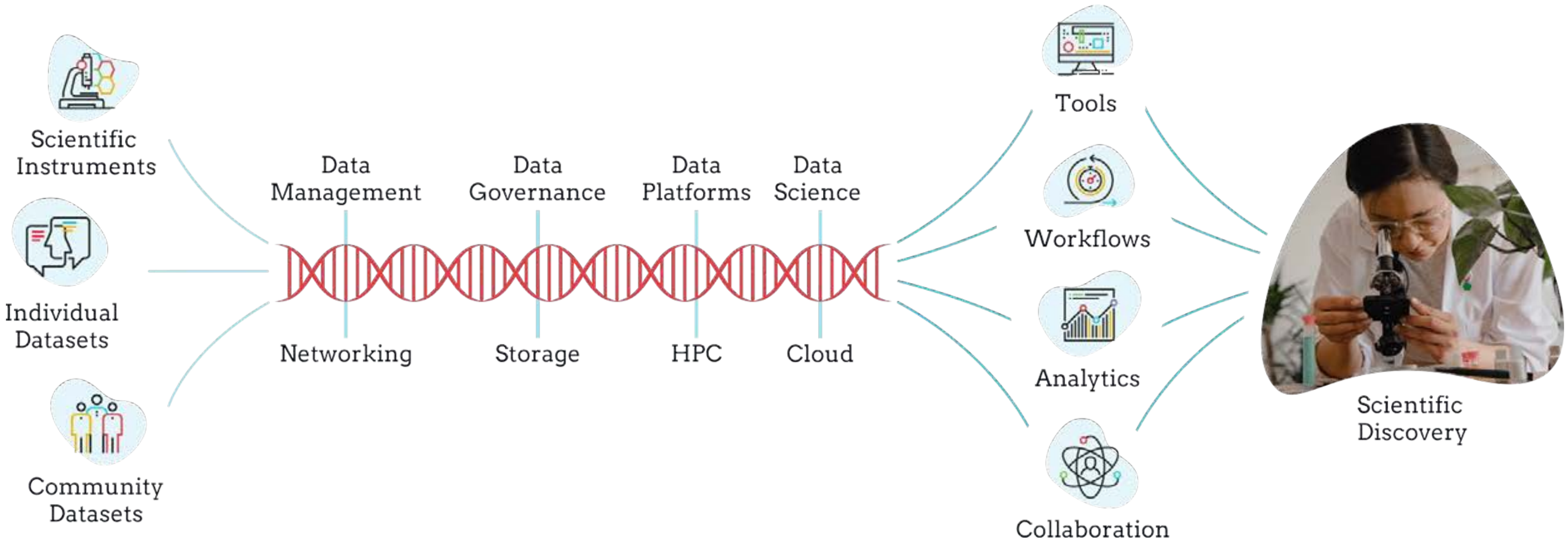
How long do we want to keep the data? Forever.

The data is a mess, its in different places, different formats!

All the data needs to be FAIR!

The *[insert critical infrastructure here]* is failing and the person who built it just left…

It gets quite complicated...



There are a lot of interconnected pieces, each of which is complex in its own right, all of which are created, managed, and used by people...

Given the complexities of tech, people, etc.

How do you:

- Find out about common goals?
- Identify common ground?
- Understand the complexities?
- Help people work together?
- Communicate effectively?
- Figure out what's really going on?
- Know what to change?

Design

Milwaukee Art Museum,
Santiago Calatrava

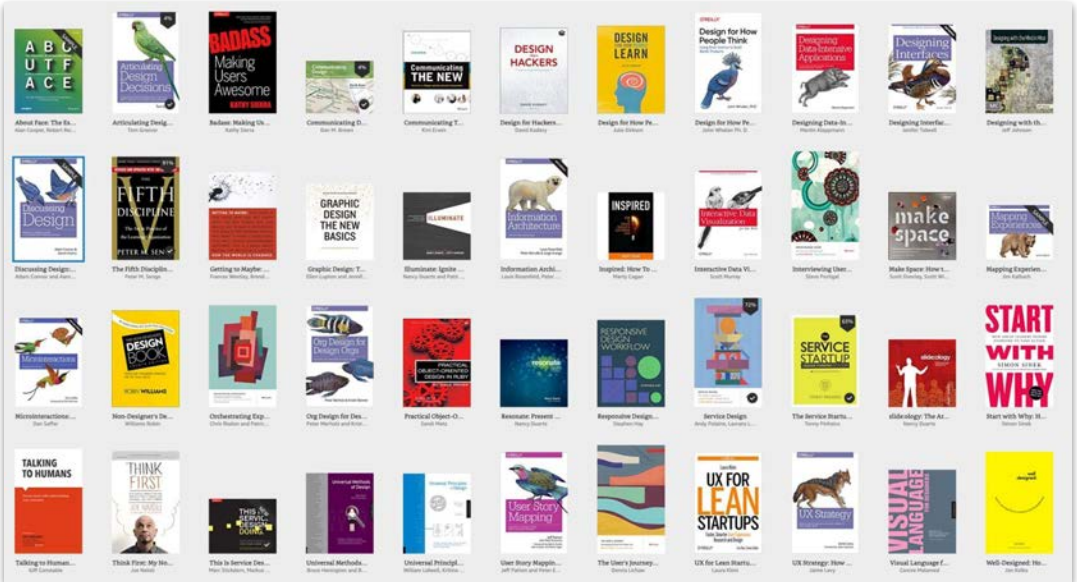
"Everyone designs who devises courses of action aimed at changing existing situations into desired ones."

Herbert A. Simon

Nobel prize winner for economics, 1978

What sort of ‘design’ is needed in BioIT?

- **Software design**
    - **Interface design** - for web sites and tools
    - **Responsive design** - for web sites and tools
    - **Object-oriented design** - for software
- **Information design** (Information architecture) - documents, websites, etc.
- **Graphic design** - for ppts, documents, posters, graphics, etc.
- **System design** - networks, compute, storage, etc.
- **Organizational design** - hiring, org charts, reporting structures
- **Service design** - for ourselves, for our clients
- **Learning design** - for training, educating
- **Human-centered design** - as an overall theme
    - User experience, UX
- **Articulating/communicating design** - to explain why *this* rather than *that*…?

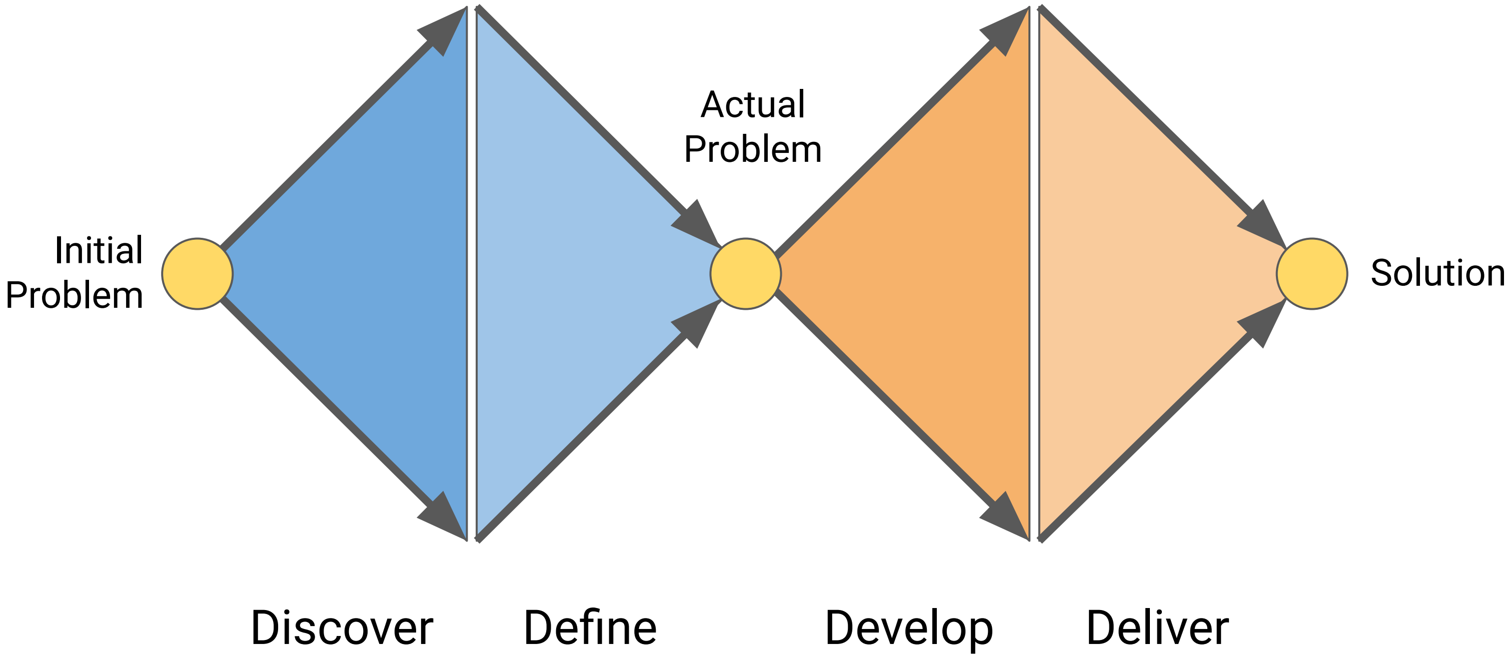


How can design help us?

Design has a rich body of knowledge, established tools to help us understand what is going on, and established processes that we can follow to figure out how to solve the problems we find.

Given that we're going to have to create all these things anyway, we might as well design them intentionally and leverage the tools and approaches that the design community has created.

Specific phases, specific activities



Double Diamond design process - popularized by the British Design Council in 2005

What happens at each stage?

	Discover	Define	Develop	Deliver
Goal	Expand options, gather as much information about the problem and situation	Group, aggregate, narrow down to a solid definition of the key problem at hand	Explore and expand the potential approaches to the problem	Implement the most effective option
Activities	Interviews, reading and research, workshops and group discussions, use cases, personas, user journeys	Synthesize information, create systems maps, identify potential intervention points	Whiteboarding, discussions with stakeholders, prototypes, mockups, etc.	Creation and delivery of the selected solution

Interviews (discussions, chats, etc)

Never underestimate the power of talking to people!

- Do your research - who are they, what do they do, why are they relevant to the problem at hand
- Have a plan for what you want to ask
- Be prepared!
  - It may be the first time anyone has actually sat down and listened to them…
- Take good notes, record the discussion if everyone agrees.
- Afterwards, review your notes for
  - Things that are working
  - Things that are not working
  - Existing workarounds
  - Suggestions and ideas

Tell us about data sharing?

Why bother sharing data, I'm not going to use it!

That data is old now, I don't have time to clean it up!

There's no incentive for me to share my data…

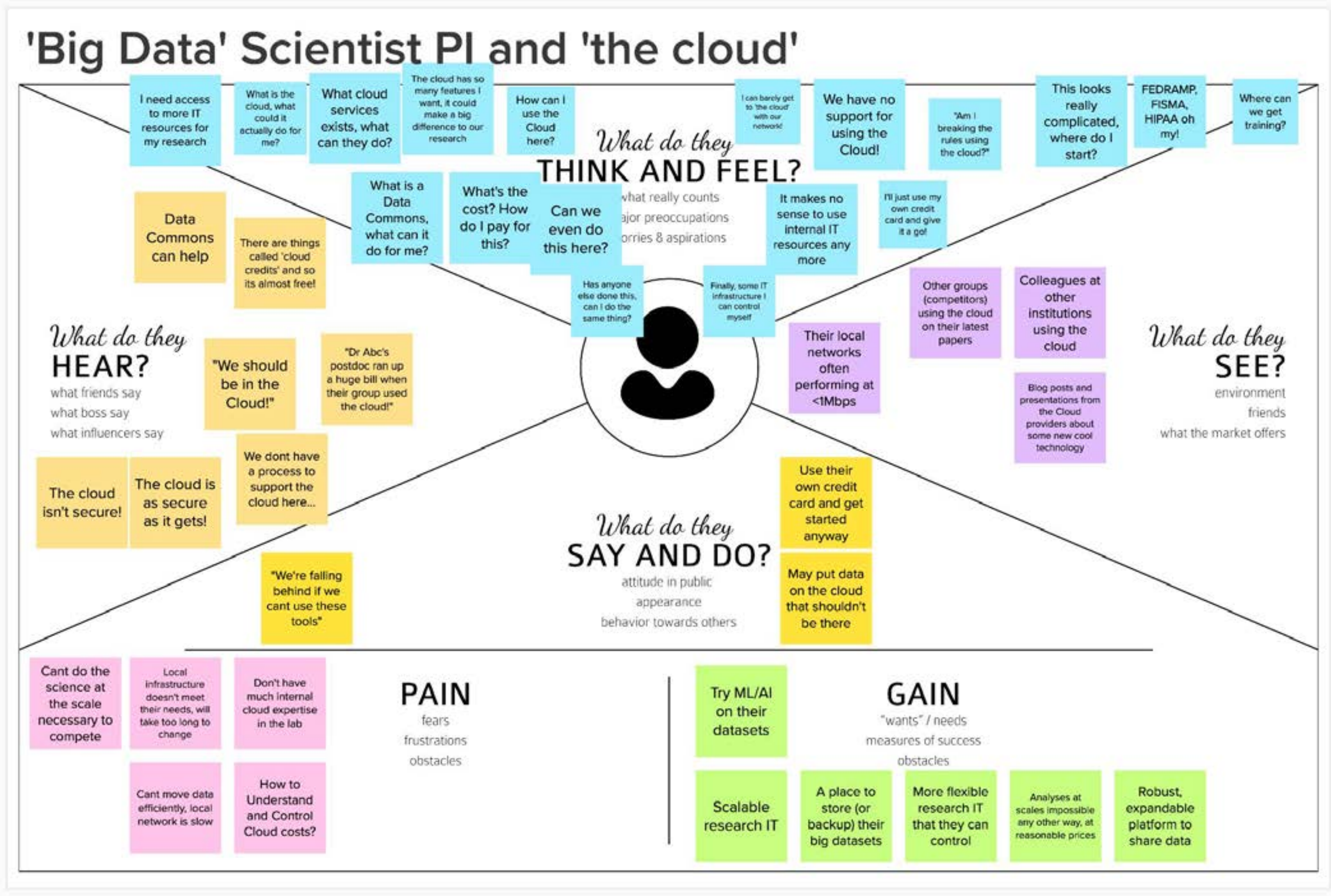
We try but have no idea what the data needs to look like

Who's going to pay for the time it takes to prepare the data and submit it!?

We have our own data standards…

We use a shared format and data dictionary, that really helps

Get inside their head with Empathy Maps



Document key Use Cases

Staff Researcher

Support to use the cloud effectively

As a **PI**, I want **training and guidance to move to the cloud** in order to **be able to effectively utilize the cloud to support my research.**

Icon by Round Icons from flaticon.com

Core Motivations

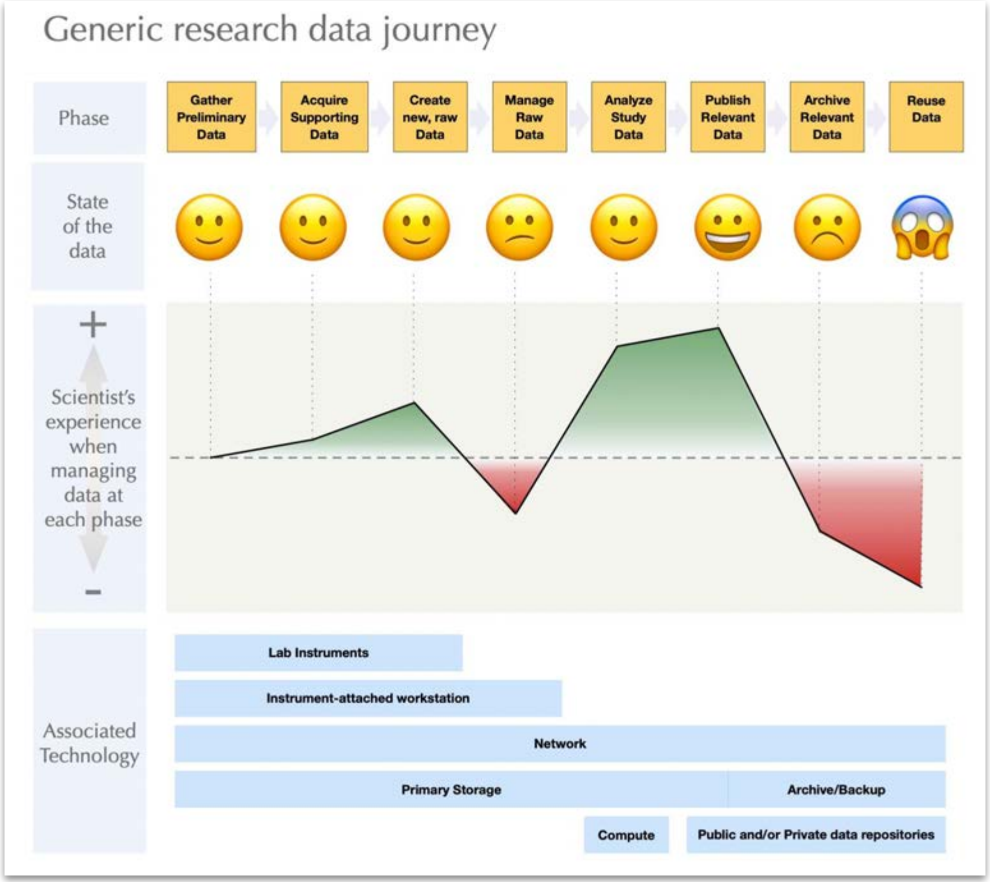
- Many PIs are interested in using the cloud in their work
- However, there are many unknowns: how to design a cloud system, how to move existing workflows, the budgetary implications, how to move data to and from the cloud, how the Org can help, etc.
- These are all barriers to more widespread adoption of the cloud

Potential Data Sources

- Amazon Web Services
- Microsoft Azure
- Google Cloud Platform
- Internal cloud team

Create Journey Maps of processes

- Map out key processes
- Identify pain points
- Communicate to stakeholders



In-depth journey / service maps

Empathy map

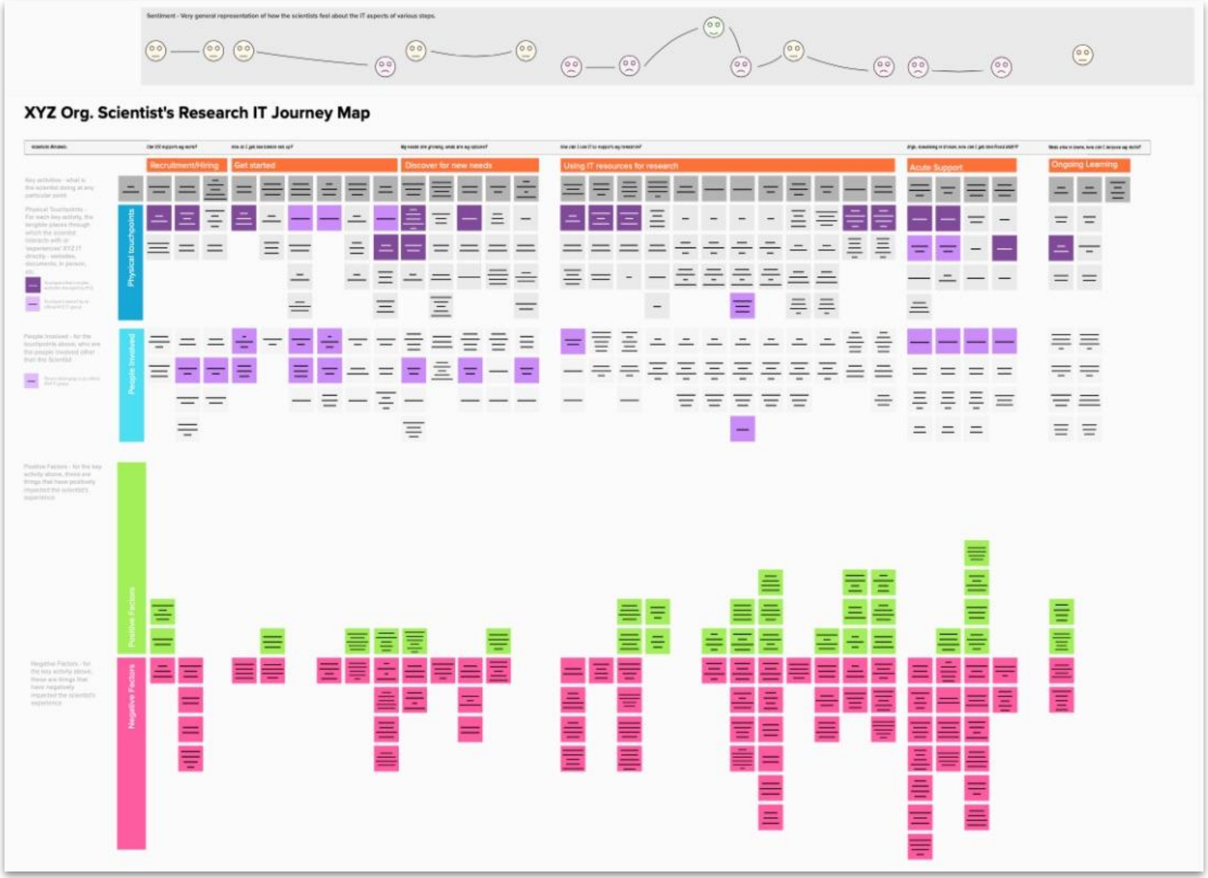
Key phases

Physical touchpoints

People involved

Positive feedback

Negative feedback



In-depth journey maps

Key activities - what is the scientist doing at any particular point

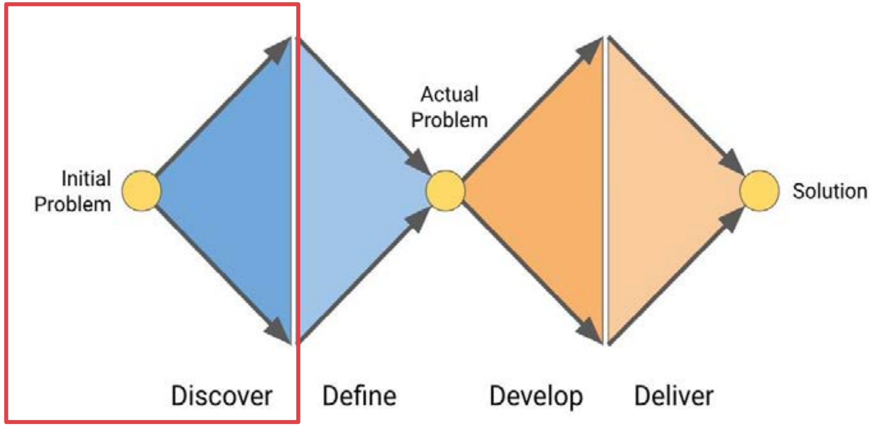
Physical Touchpoints - For each key activity, the tangible places through which the scientist interacts with or 'experiences' XYZ IT directly - websites, documents, in person, etc.

website — Touchpoint that includes websites managed by XYZ

Website — Touchpoint owned by an official XYZ IT group

Scientist Mindset:		Can XYZ support my work?			How do I get the basics set up?						My needs are growing, what are my options?				
		Recruitment/Hiring			Get started						Discover for new needs				
	Key Activities	Hear about XYZ IT	Look for more info	Ask questions about specific lab requirements	Account creation	Purchase computer & software	Software Installation and config	Get Network Access	Advice on Hardware config	Set up website	Learn about available options	Learn about other (not so available) approaches	Evaluate services	Cloud or HPC	Get assistance to implement
Physical touchpoints		XYZ websites	Various XYZ websites	Phone call with relevant staff	XYZ Website and form	XYZ Bookshop	IT support	IT Support	XYZ Bookshop	IT Support	XYZ resource websites (HPC center, etc)	Papers and journal articles	Website	Cloud Vendor Website	Email
		Meetings and conference presentations	Email to contacts	Email to contacts		External Vendor website	Department IT support		Email External Vendor	Dept. website	XYZ web Search	Personal contact	In person discussion	In person discussion	In person discussion
							Ask Colleague		Ask Colleague	External vendor website	Ask Colleague	Seminar or presentation	Documentation	Informatics core or other service that uses cloud	Online documentation or examples
							Online documentation or examples			'The Web' for DIY approach		Websites inc. YouTube, Stack overflow, BioStars, etc			Phone/chat support

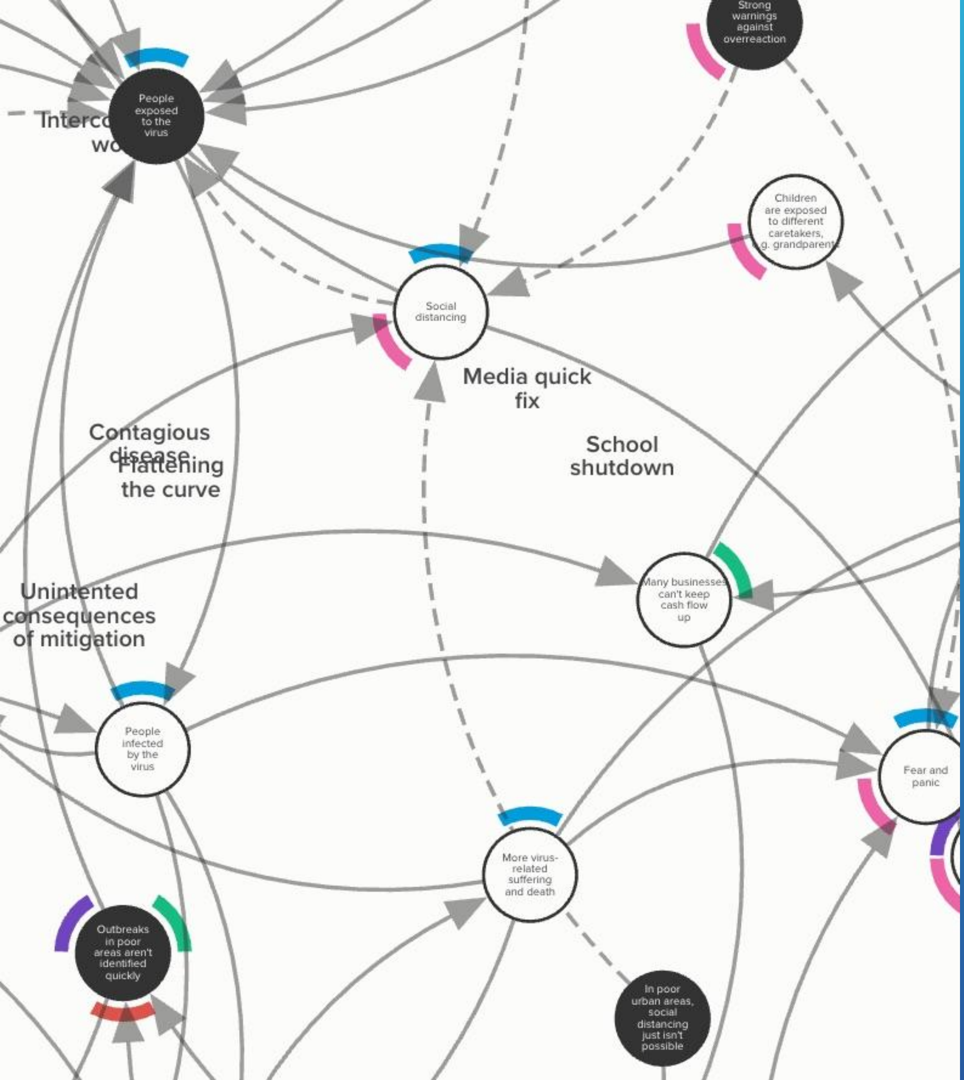
Design methods help us collect and communicate information



We now know about

- The People
- Their Use Cases
- Their Processes
- Their pain points

But what's really going on?

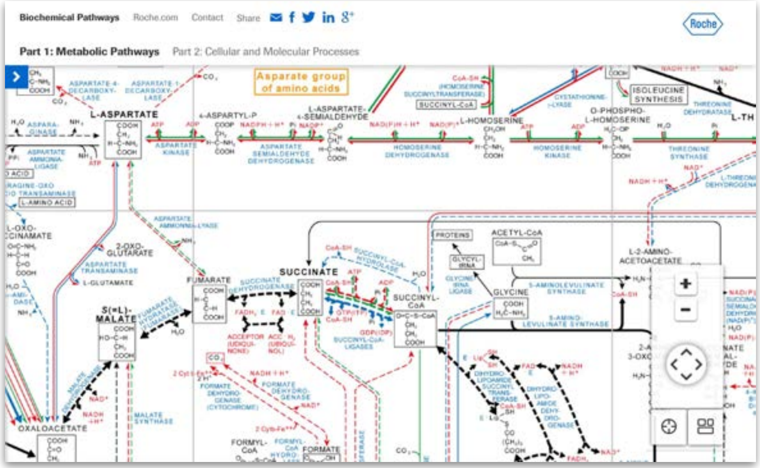


Systems

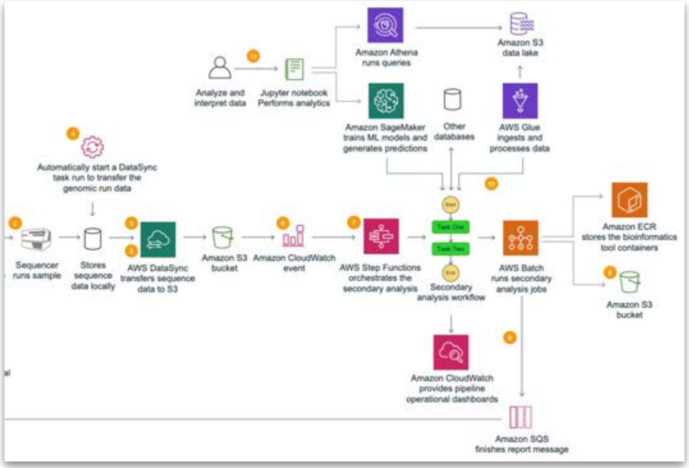
COVID19 systems map by Alex Vipond
https://kumu.kumu.io/covid-19

What do we mean by a system?

Below are examples of systems familiar to the Bio and Bio-IT community: they have components, relationships, dependencies, feedback loops, but we're focusing on ones where people are in the mix…

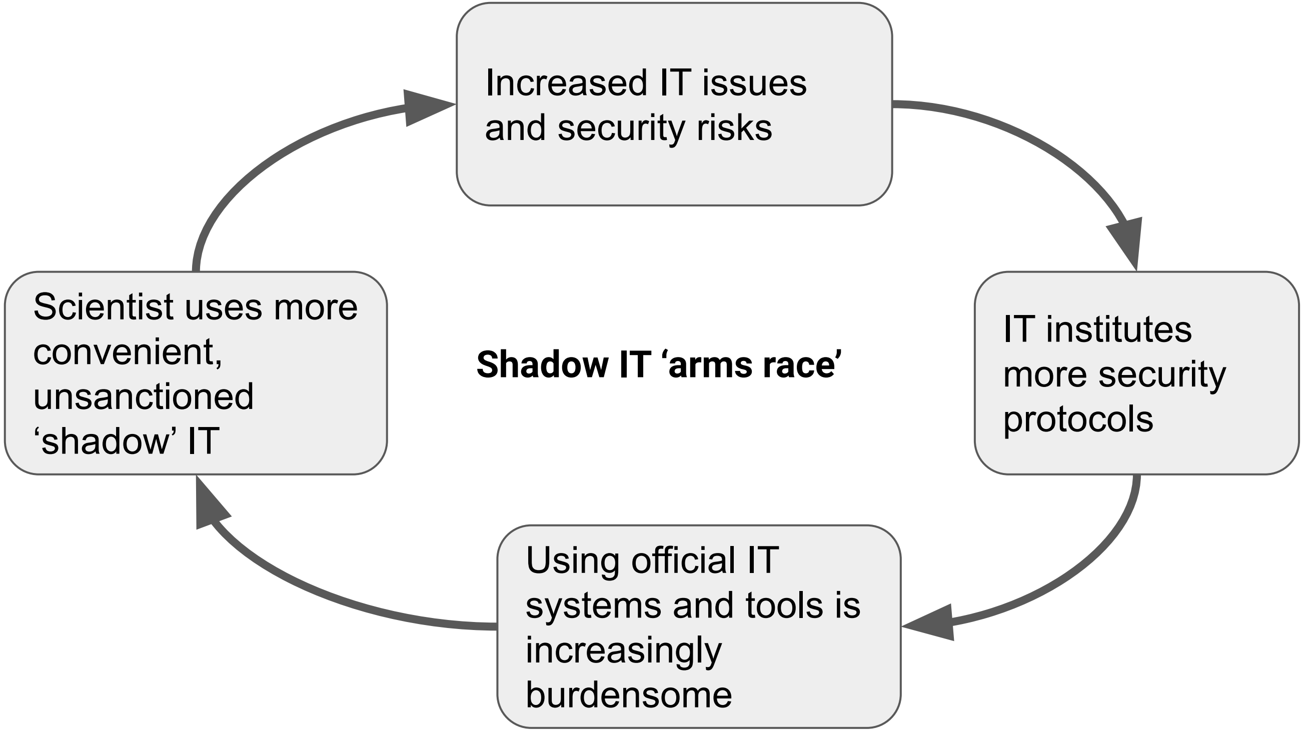


Roche Metabolic Pathways (biochemical-pathways.com)

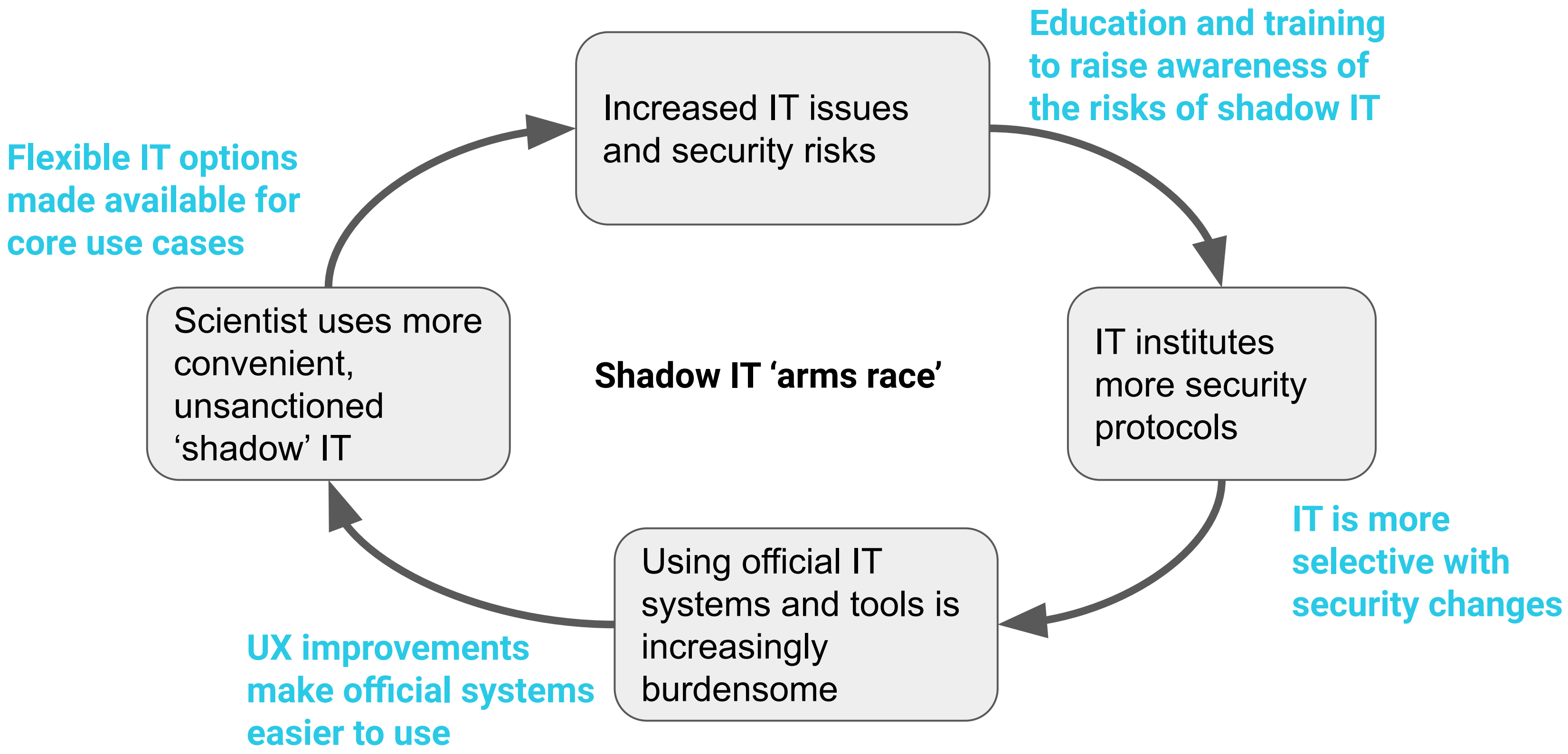


AWS Genomics Reporting Architecture

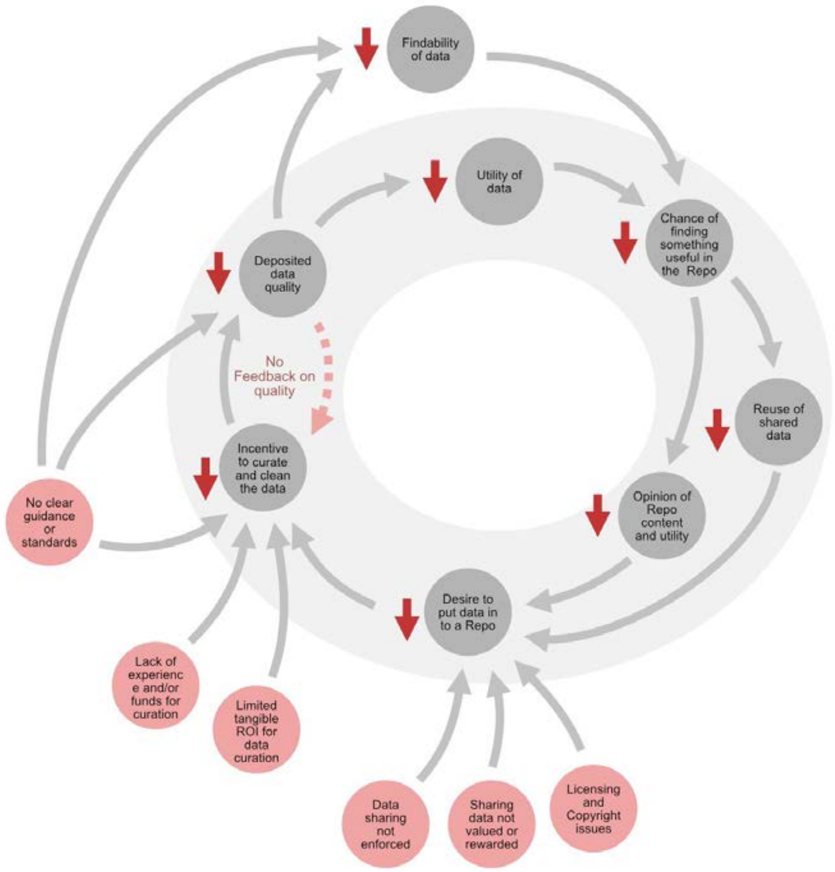
Systems maps - A leads to B, which leads to C



Identification of intervention points



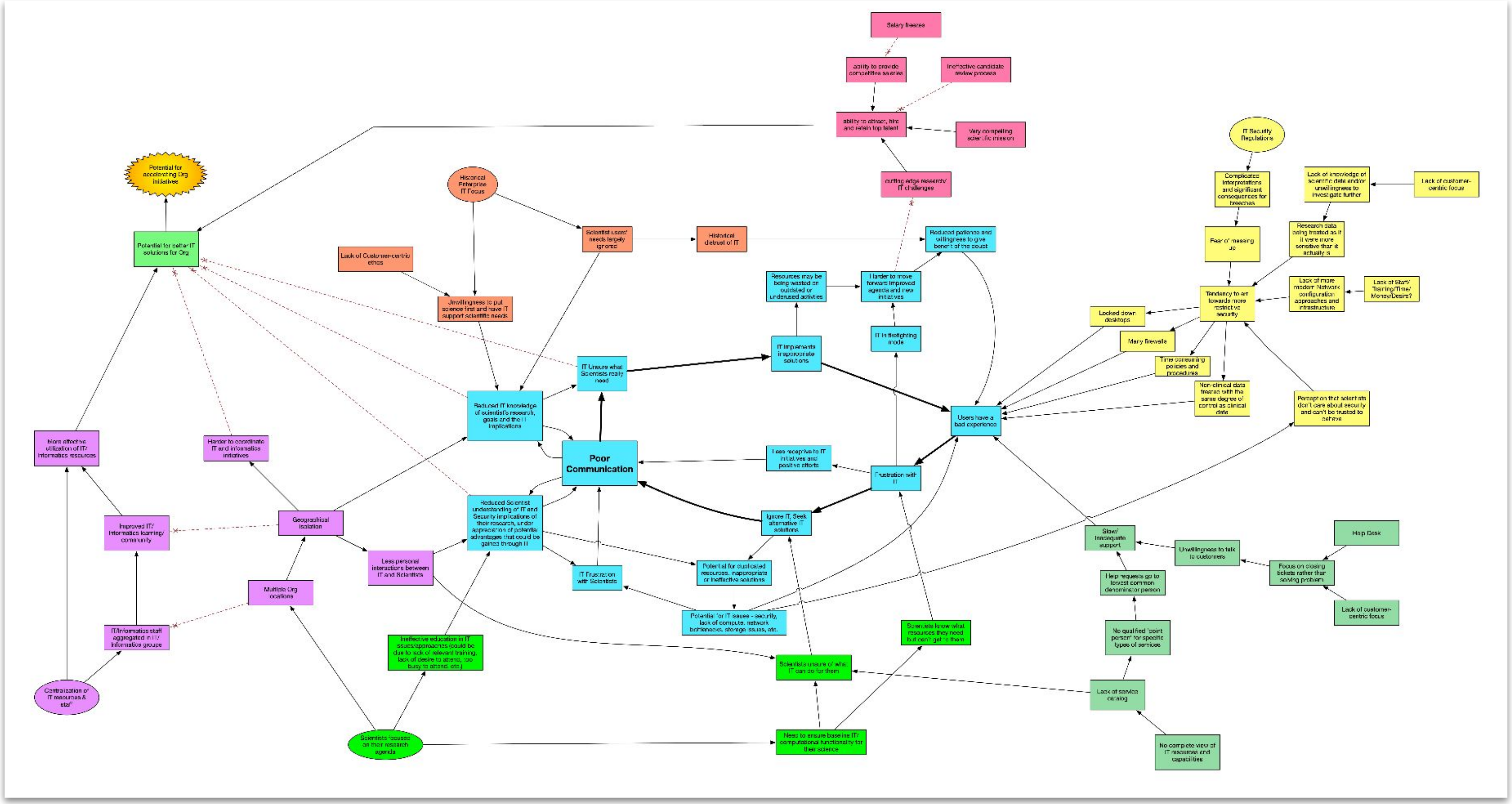
Generic FAIR Data Systems Map



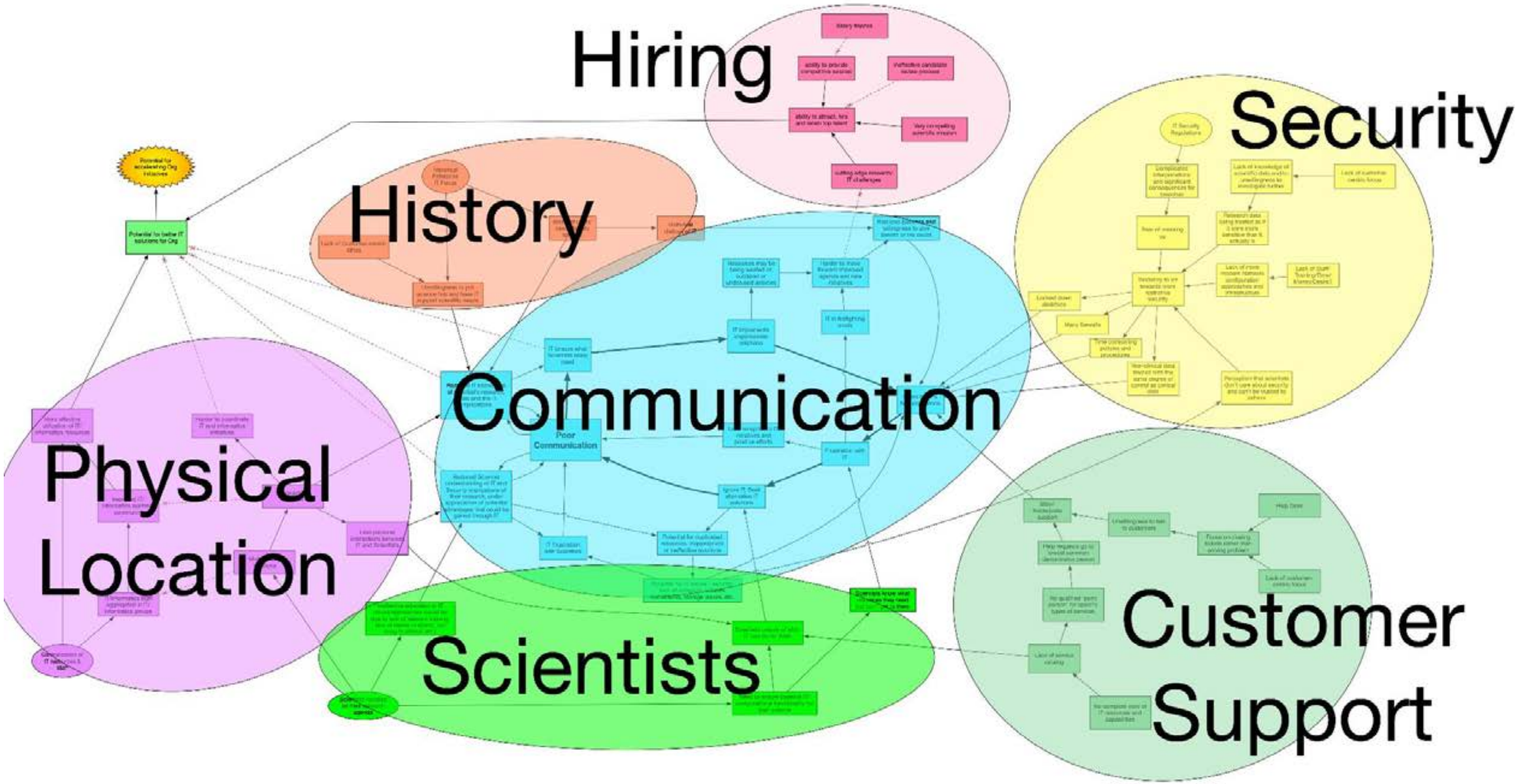
- Making data FAIR is a common desire expressed by labs, organizations, and the scientific community as a whole
- However the system often has various factors that get in the way of these initiatives
- A systems map can help identify some of these factors and how they relate to each other.

- These relationships often form loops that can progressively help (or hinder) the system.
- Identifying these loops can be vital in changing the behavior of the system.

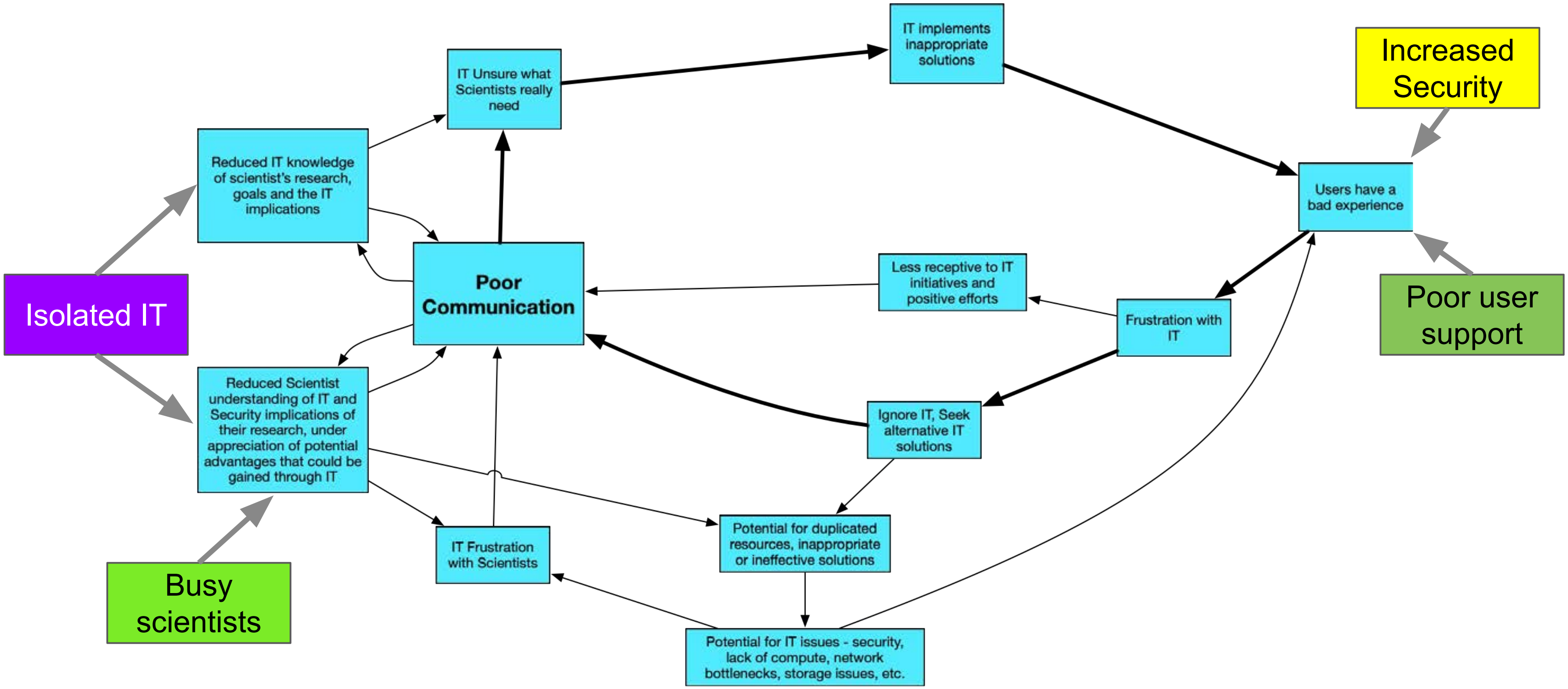
Large Bio organization IT systems map



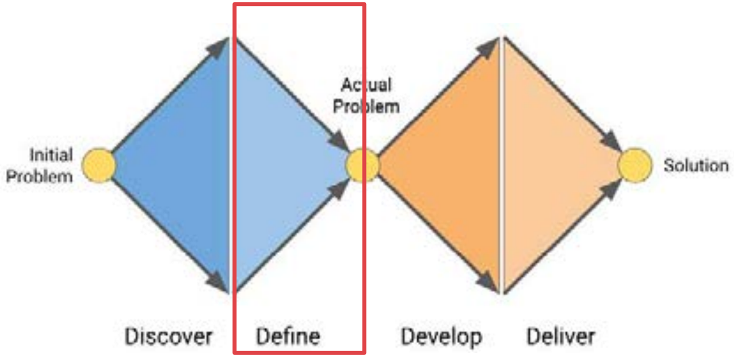
Large organization Bio-IT systems map

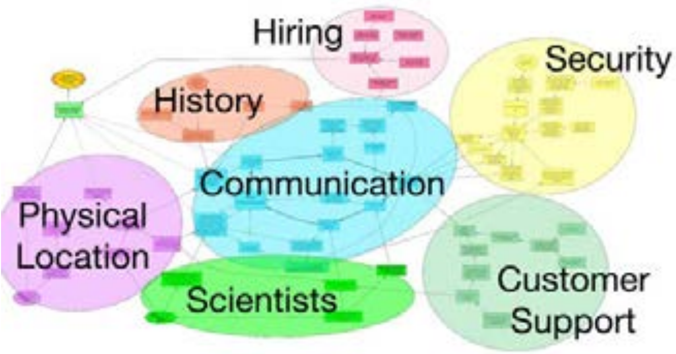


Core loop - communication



Systems maps help us put the information in context

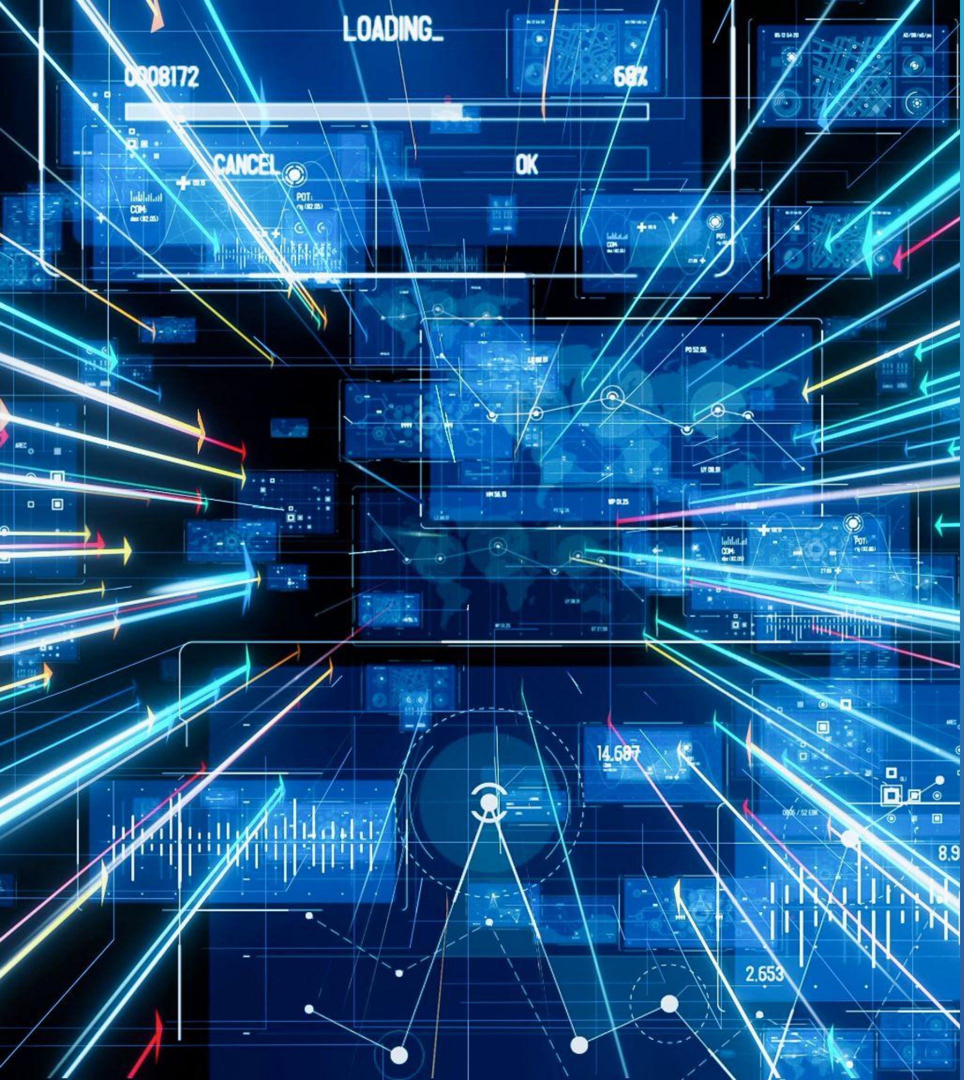




We now know about

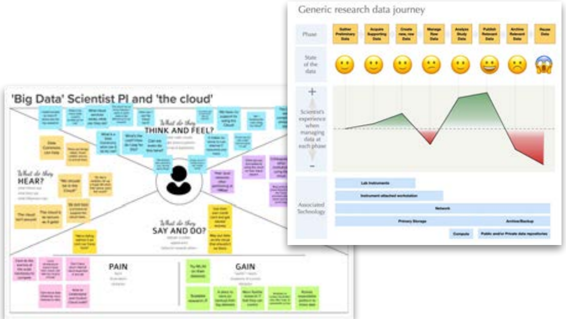
- The components
- The relationships
- The feedback and regulatory systems
- The actual problem

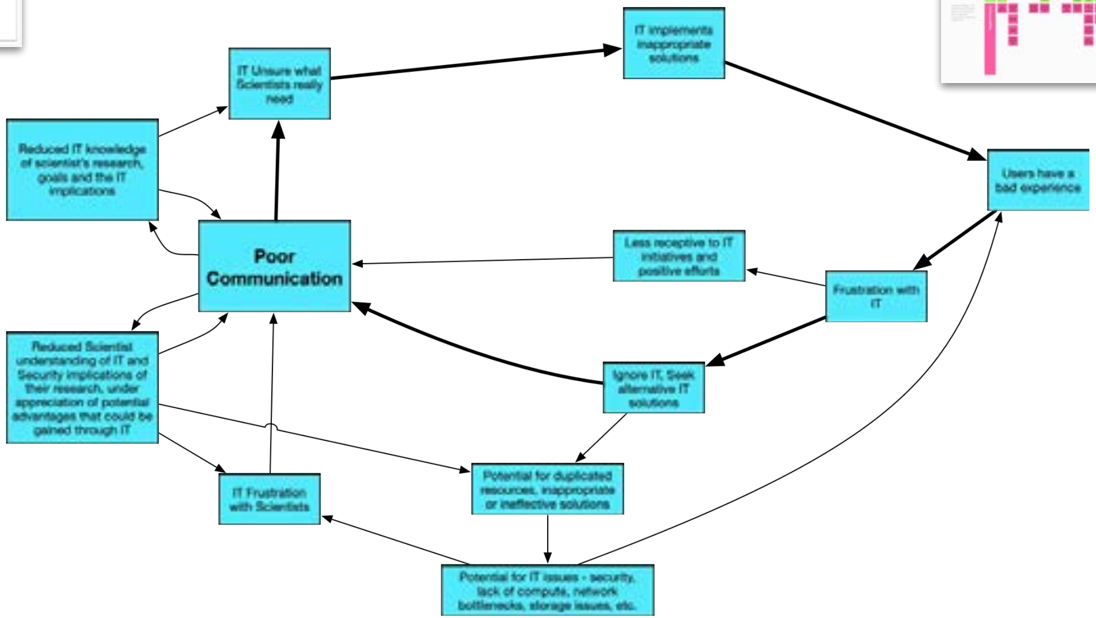
But how to change it?



Change

We now have a lot of information





Identify other places for change

Isolated IT	Busy scientists	Increased Security	Poor user support

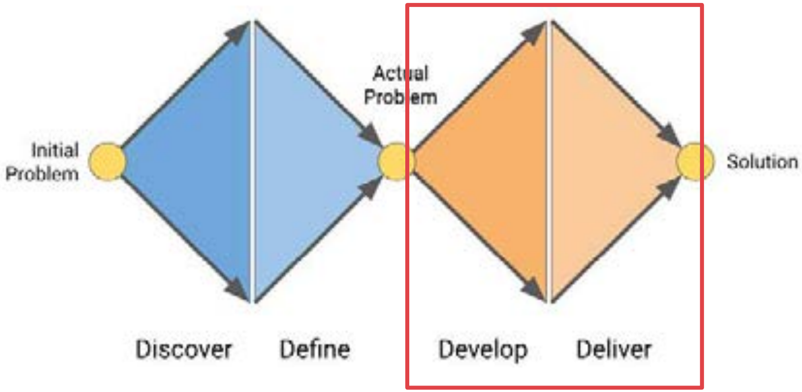
In this example, various factors in the organization are contributing to poor communication and this is impacting IT's ability to support and enable the research organization - fixing this and establishing common ground are key:

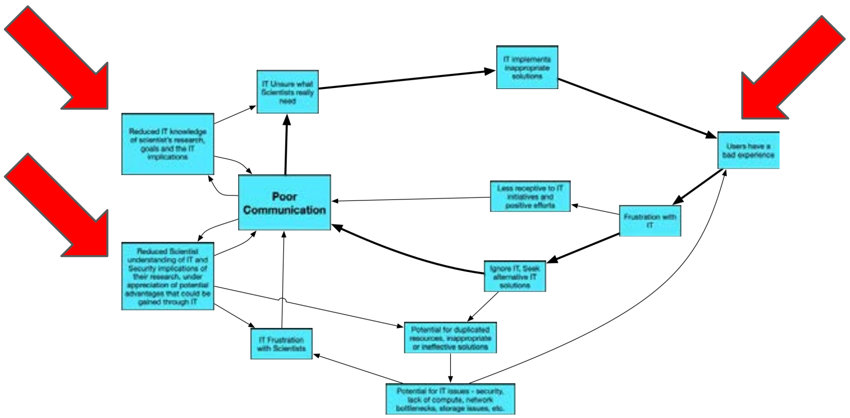
How would you do that?

Potential Ideas might include:

- *Town Hall meetings to discuss the issues more widely*
- *Collaborative projects to get IT and researchers working together, aligned around a common goal*
- *IT office hours in the lab space to make IT more accessible*
- *Lunch and Learn on Data Management, using the tiered storage, getting the most out of the network, large-scale data transfer methods, getting started on the cloud, etc.*

Systems maps help identify how to change

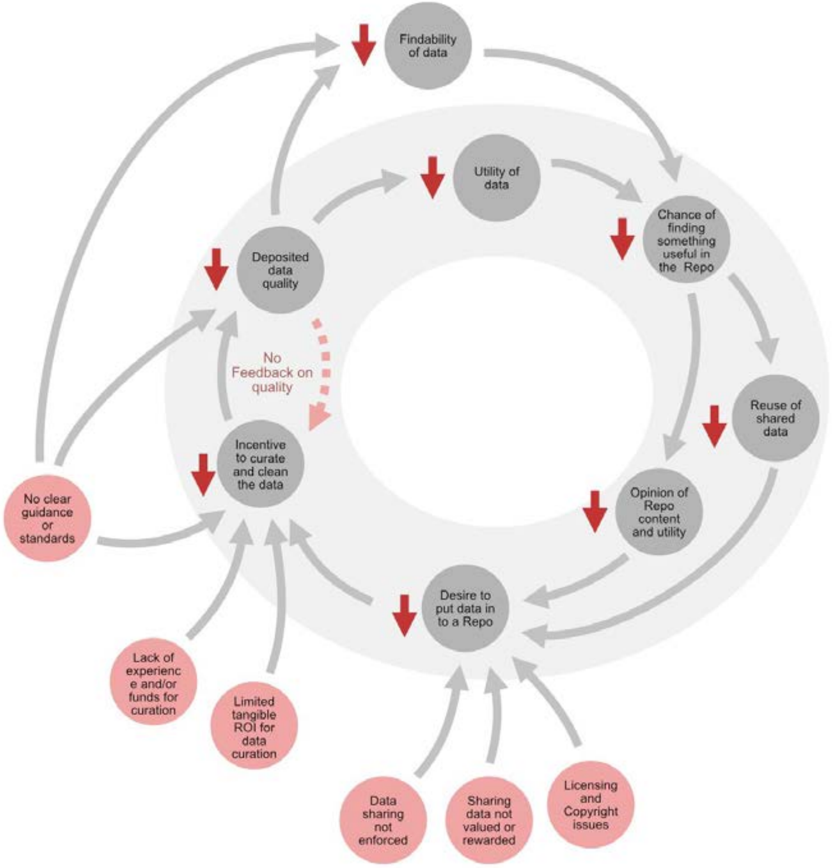


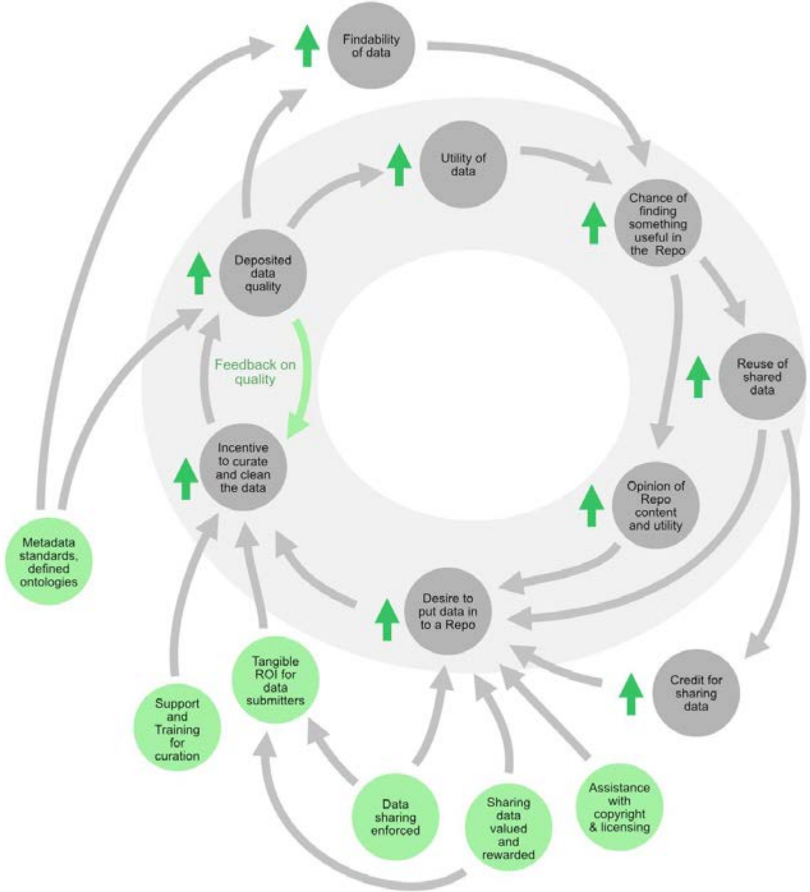


We have a model to help:

- Flag areas to intervene
- Identify possible solutions
- Identify potential unintended consequences
- Suggest small proof of concept activities

Generic FAIR Data System - Ways to improve





Something (always) needs to change - now we have some new tools to make it happen

Our storage is full/broken/old/ not backed up, etc.

We want to move on/off the Cloud!

The scientists aren't happy with IT.
(and probably vice versa!)

We don't have access to enough compute.

Moving our data is too slow!

We need our own network to connect all our teams across the entire country!

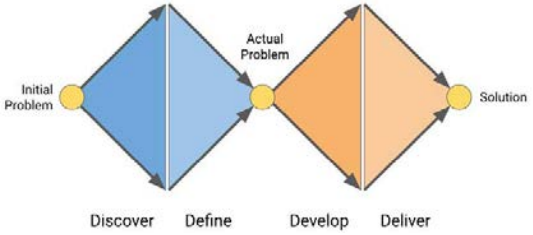
We want a custom cloud SaaS platform so people can easily use our product!

All the data needs to be FAIR!

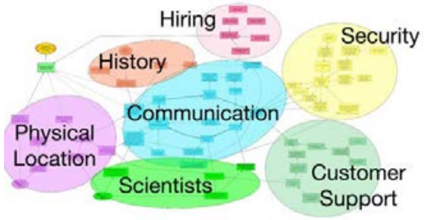
The *[insert critical infrastructure here]* is failing and the person who built it just left...

Summary

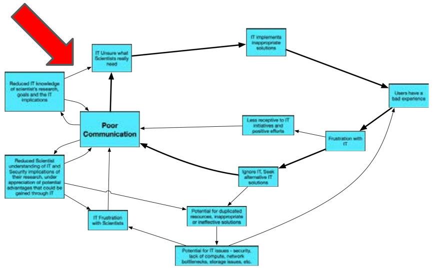
"Everyone designs who devises courses of action aimed at changing existing situations into desired ones."



Design tools and processes



Systems Maps



Opportunities for Change

BioTeam
Accelerate Science

Thank you!

Useful links

Design Thinking

- interaction-design.org/literature/article/what-is-design-thinking-and-why-is-it-so-popular
- **Mapping Experiences**, Jim Kalbach, O'Reilly publishing, Journey mapping and other visualizations
- **Double Diamond** - fulcrum.rocks/blog/double-diamond-design

Systems Thinking, Systems Mapping

- "Thinking in Systems" - Donella Meadows
- Systems Mapping Training
  - acumenacademy.org/course/systems-practice/

Systems thinking and Design, Systemic Design

- https://medium.com/@miekevanderbijl/systems-thinking-design-72209d534c4c