20 Years of BioTeam:
Lessons Learned and Future Insights in Scientific Digital Transformation

Welcome

Today's speakers:

- Ari Berman, Ph.D., CEO
- Stan Gloss, Co-Founder and Fellow
- William Van Etten, Ph.D.,Co-Founder and Senior Scientific Consultant
- Chris Dagdigian, Co-Founder and Senior Technical Director of Infrastructure

Chris Dagdigian

Stan Gloss

William Van Etten

Ari Berman

BIOTEAM
Accelerate Science

What we'll cover today

- Take you on a journey through 20 years of BioTeam experience
- Each of us will discuss a different aspect of that experience and provide some insight on what the future may hold for the industry
- Start with a high-level overview of the market
- Stan will talk about the role of data in advancing scientific missions
- Bill will talk about more technical aspects of data management and governance
- Chris will talk about 20 years of lessons learned building infrastructure to support data intensive life sciences

BIOTEAM
Accelerate Science

Scientific Data Ecosystems The Foundation to Accelerate Science

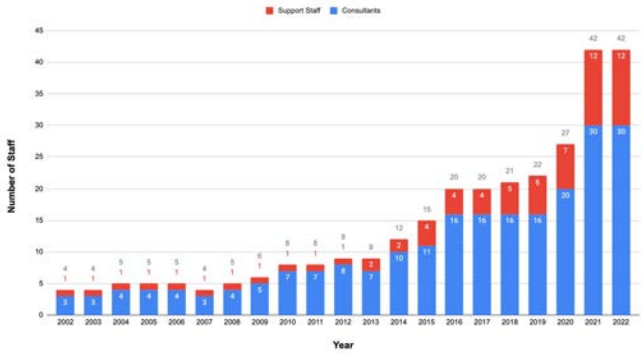
Ari Berman, Ph.D.
CEO

BIOTEAM
20 years
Accelerating science since 2002

BioTeam is 20 years old!

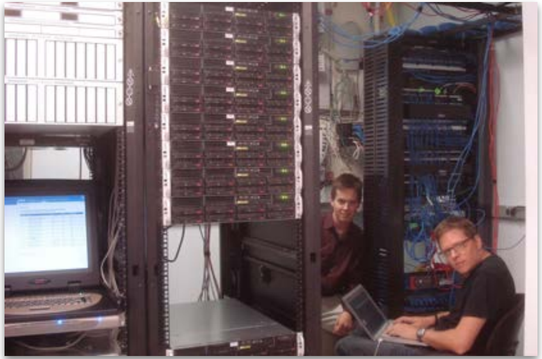
- Started in October 2002
- Four founders from Blackstone Technology Group
- Started with the goal of bridging the gap between IT and science
- Saw a need—computing in life sciences
- First employee in 2004 (Chris Dwan)
- I joined in 2012, #7 at that time (10th overall)
- Grew slowly from there until 2016
- 2x growth through pandemic
- Went from Enabling Science to Accelerating Science

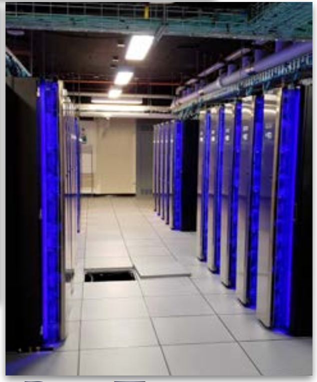
Four-dimensional: *The BioTeam's "accidental" experts, from left: Michael Athanas, Bill Van Etten, Stan Gloss, and Chris Dagdigian.*



Steady goals and values throughout

- Wanted to do cool and impactful work
- Always put supporting science with technology first
- Always took an honest and ethical stance (sometimes too honest)
- Goal was always to do the right thing—it was never about money, always about science
- Those strong principles still guide us today
- Makes BioTeam an awesome place to work, and we get to work with amazing people all over the world





BIOTEAM
Accelerate Science

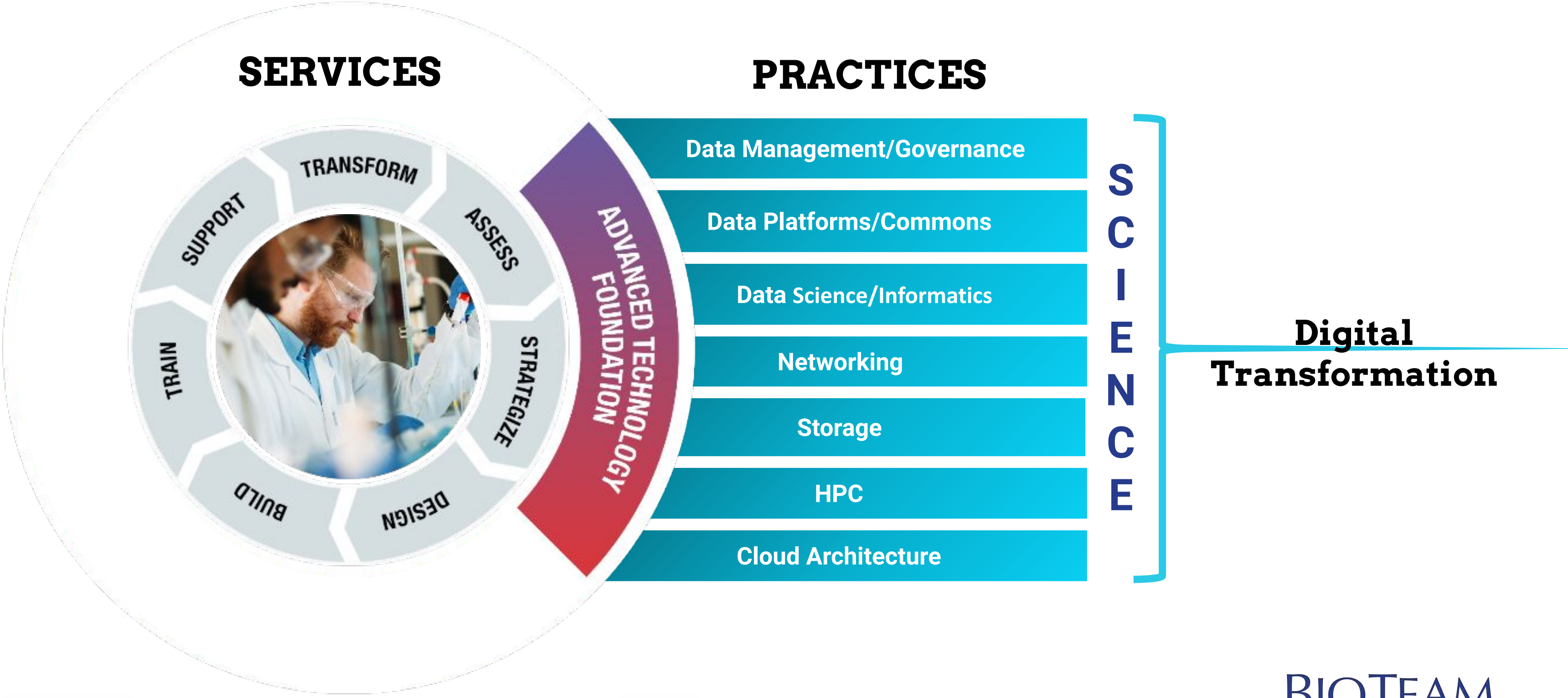
Evolution of BioTeam

First all-hands meeting (2009)

March all-hands meeting (2022)

BIOTEAM
Accelerate Science

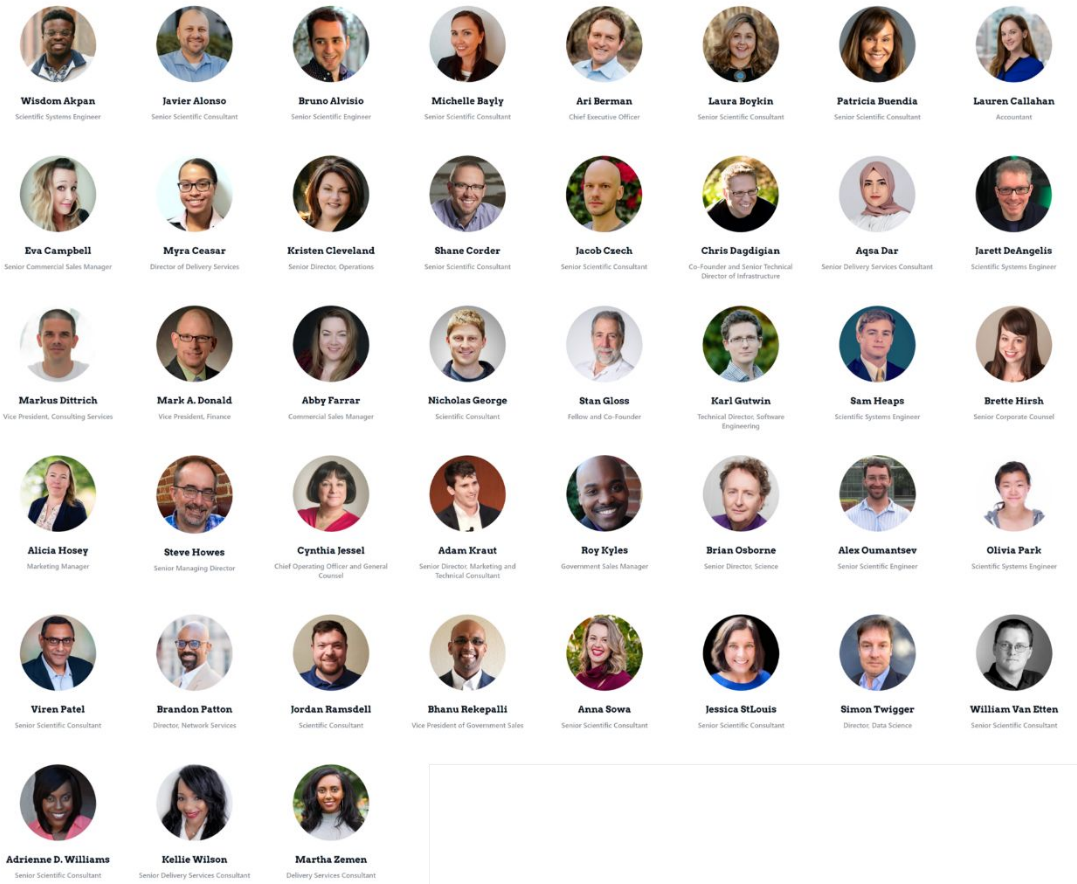
From infrastructure to scientific data ecosystems



BIOTEAM
Accelerate Science

Team of scientists and technologists

- Seamless integration requires interdisciplinary skills
- Our team is diverse, broad, deep, and collaborative
  - Research Scientists:
    - Geneticists, structural biologists and more
    - Data Scientists/Bioinformaticians
  - IT Experts
    - Advanced Infrastructure Experts Cloud Architects and Strategists
    - Software Developers
  - Outstanding Support Staff



BIOTEAM
Accelerate Science

What we've seen through the years

What hasn't changed

- Life sciences data is still hard to work with
- Data generation still at an all time high
- Unified data standards don't exist in the field
- Computing and storage are still at a premium
- Still requires significant computational sophistication to analyze modern research data
- Computing is a laboratory tool, not an IT function

What has changed

- Community movement towards data standards
- More diversity in computational environments (on-prem, cloud, data commons, data services)
- Sophistication of bioinformatics code has improved dramatically
- More willingness to invest in computing from science orgs
- Modern computational methods (i.e., AI/ML) have forced better data habits

Struggle to find storage/compute power

- Most lab scientists spend half their time figuring out where to save their data
  - Without the right support, they make bad decisions
- Most Bioinformatics scientists spend 80% of their time just cleaning up data to be analyzed
- Institutional HPC/Storage is usually an option, but security makes it hard to collaborate
- Large organizations need to store 100s of PBs of data, then need to analyze it
- Intersection of Big Data and AI/ML—forced starvation for storage and compute

Led industry to digital transformation

- What is digital transformation?
  - The unification of digital (data policies, standards, metadata curation, and models) and transformation (people-driven: cultural alignment, funding, and commitment) that allow better use of data assets
- Realization of FAIR
  - Digital drivers—paradigm shifting technologies: Big data -> cloud -> IOT -> now AI
  - Cultural drivers—incentives, strategic alignment, organizational priorities, training

BIOTEAM
Accelerate Science

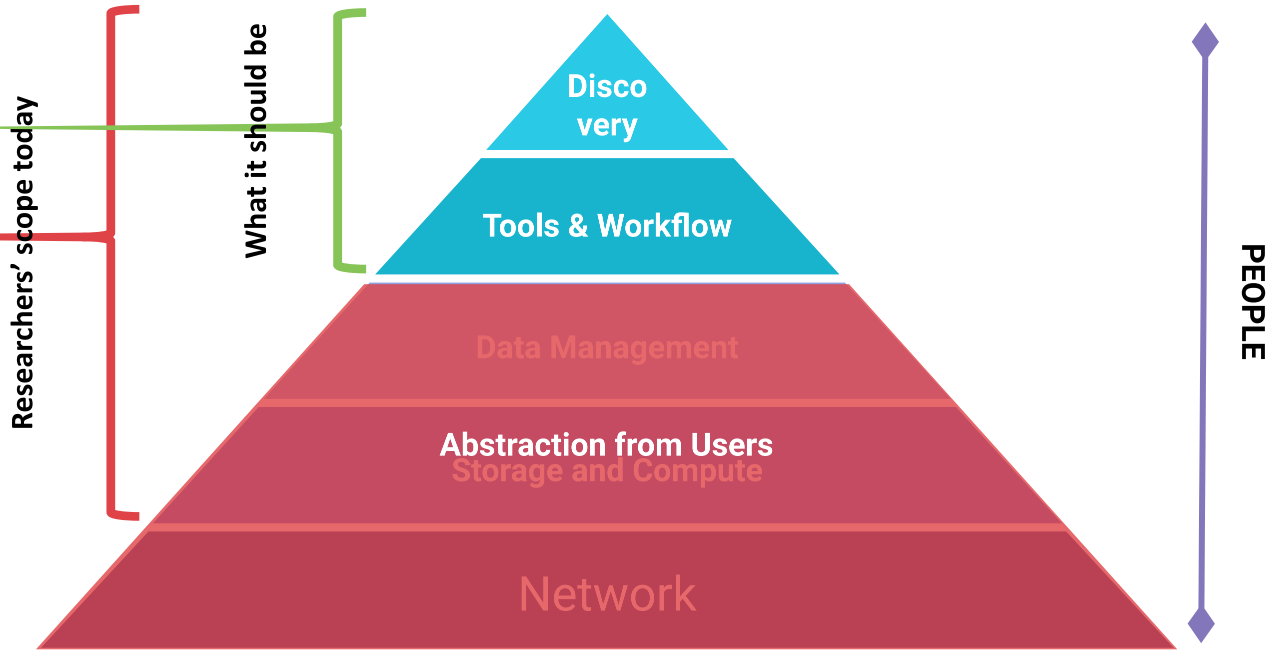
Comparison of industry data capabilities

Capability	Pharma	US Federal Government	Academia
Internal Data Generation	Advancing	Advancing	Advancing
Internal Data Storage Capacity	Sustaining	Lagging/Average	Advancing
Internal Computational Capacity	Lagging/Average	Lagging/Average	Advancing
Cloud Capability	Advancing	Sustaining	Sustaining
Data Transfer	Lagging/Average	Sustaining	Advancing
Data Management	Lagging/Average	Behind/Restricting	Lagging/Average
Data Standards	Lagging/Average	Lagging/Average	Lagging/Average
Data Analytics	Innovating	Sustaining	Advancing
AI/ML	Advancing	Sustaining	Innovating
Data Sharing	Behind/Restricting	Sustaining	Sustaining
Overall Data Strategy	Lagging/Average	Lagging/Average	Sustaining

- Innovating
- Advancing
- Sustaining
- Lagging/Average
- Behind/Restricting

BIOTEAM
Accelerate Science

Maslow's hierarchy of Research Computing needs



Modern Biomedical Research: Integrated spectrum of infrastructure, software, services and support, focused on accessible science

BIOTEAM
Accelerate Science

What does the future of data ecosystems look like?

- **Data unification across the industry**
  - Data scientists and Bioinformaticians spend most of their time getting data into formats and contexts that allow them to be analyzed together (data harmonization)—need to solve that problem universally.
- **Abstracted IT infrastructure for science**
  - Infrastructure abstracted away from the scientists so they can focus on the science, not the technology
  - Scientists spend a lot of time trying to figure out where to save stuff, let alone analyze it. Need systems that they don't have to care where the data are, it just works.
- **Create technology-forward science cultures**
  - 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.

Scientific leaders will be those that manage data as their most valuable asset

Manage Scientific Data as a Product to get the Most Value From Your Data

Stan Gloss
Co-Founder and Fellow

Digitization versus Digital?

Digitized = Operational Excellence

The transformation enhances traditional products and customer service

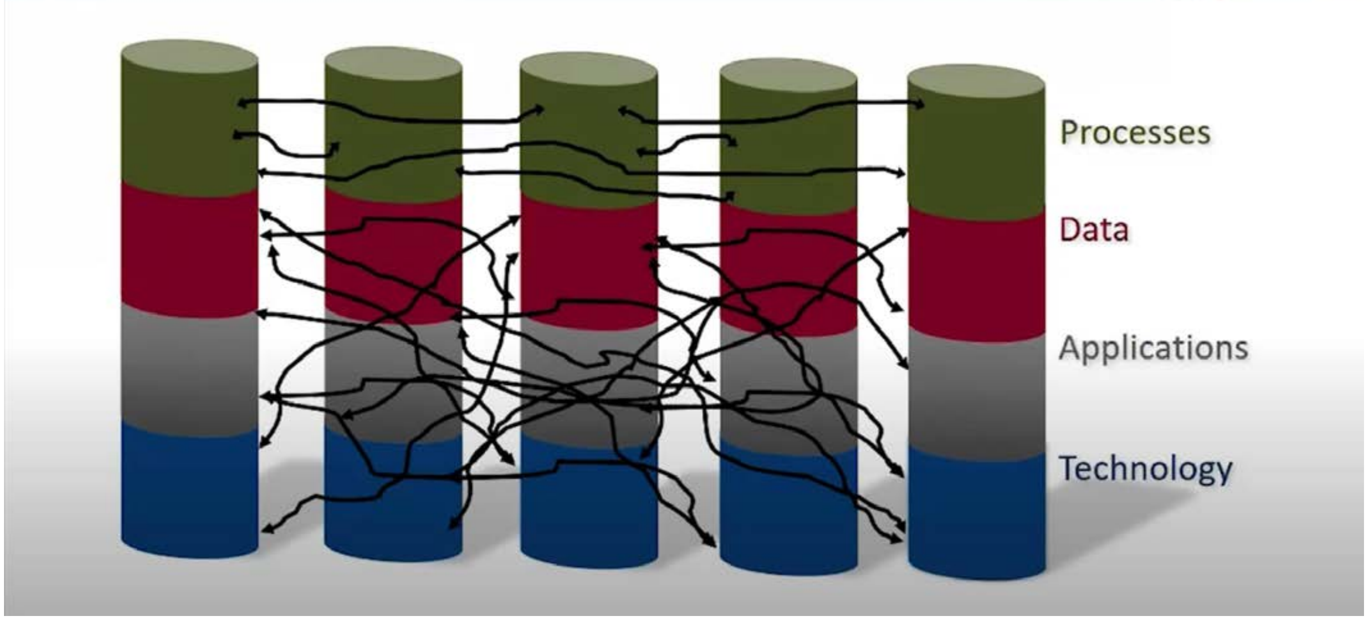
Digital = Rapid Business Innovation

The transformation delivers a new customer value proposition

Source: Jeane Ross - 2020 Digital Transformation youtube.com/watch?v=IrX0UemtWVQ&t=23s

BIOTEAM
Accelerate Science

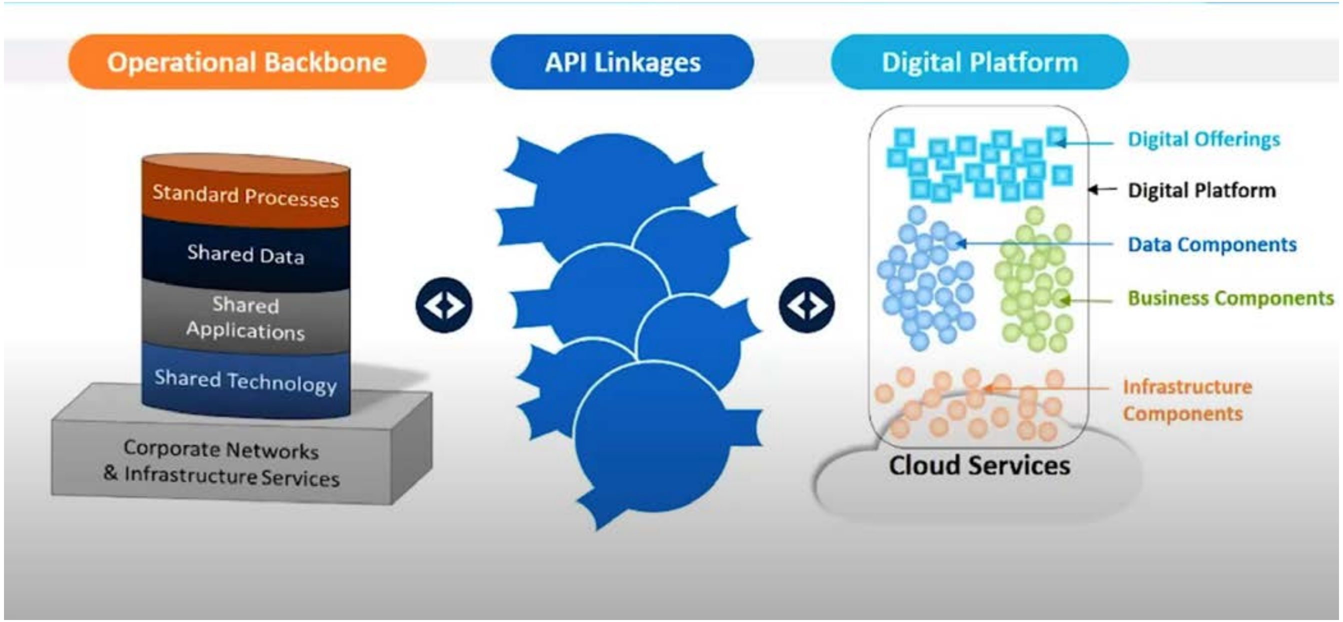
Why is it difficult to get the most value from data?



Source: Jeane Ross - 2020 Digital Transformation https://www.youtube.com/watch?v=IrX0UemtWVQ&t=23s

BIOTEAM
Accelerate Science

Aligning people, processes and technology



Source: Jeane Ross - 2020 Digital Transformation youtube.com/watch?v=IrX0UemtWVQ&t=23s

Approach to data is key to harness its value

To effectively harness digital, data, and analytics to drive better outcomes for patients and customers, scientific organizations are transforming how they approach technology and how they are structured as an organization

Core to this is to switch the organization from a **project-centric operating model to a product-centric model**

Montello, Mike. Transforming from a project to product-centric organization. Mar 23, 2021.
medium.com/gsktech/transforming-from-a-project-to-product-centric-organization-b9af4b58148e

From project teams to product teams

Project teams	**Product teams**
Temporal teams	Stable cross-functional teams
Focus on delivery of outputs	Focus on measurable outcomes and objectives
Long waterfall delivery cycles	**Agile fast feedback loops**
Deliver for the business	Deliver and partner with the business for the customer
Fixed timeframe	Lifecycle oriented

Montello, Mike. Transforming from a project to product-centric organization. Mar 23, 2021.
medium.com/gsktech/transforming-from-a-project-to-product-centric-organization-b9af4b58148e

BIOTEAM
Accelerate Science

Is your data exhaust or renewable fuel?

"Traditional pharmaceutical companies view data as the exhaust of the drug discovery process.

New digital native companies view their data as renewable fuel that powers their discovery."

Mason Victors, Fellow Recursion Pharmaceutical

BIOTEAM
Accelerate Science

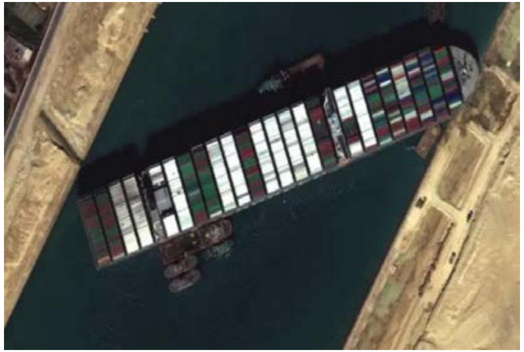
Data = Products in a supply chain

Source: Fernando, Jason "Supply Chain Management (SCM): How It Works and Why It Is Important" July 7, 2022 Investopedia
investopedia.com/terms/s/scm.asp

BIOTEAM
Accelerate Science

Supply chains can be disrupted

Baby Formula



Suez Canal

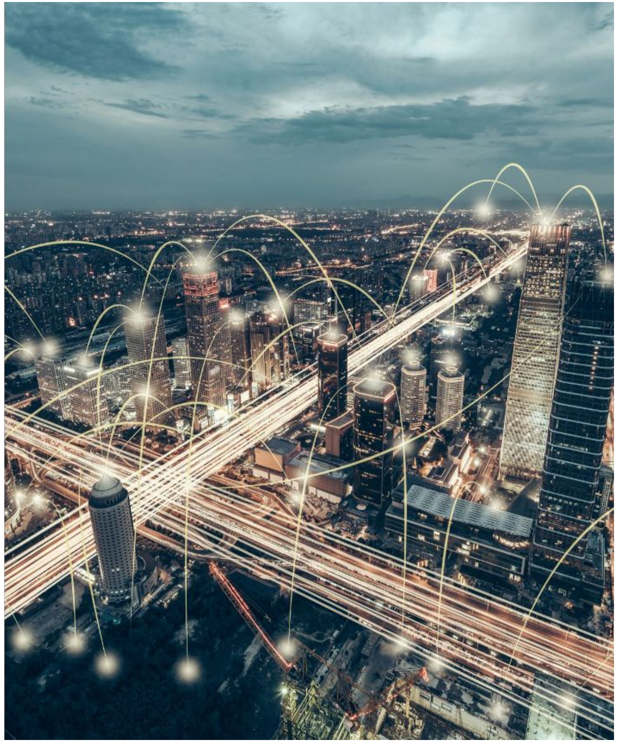
Chip Shortage

War

BIOTEAM
Accelerate Science

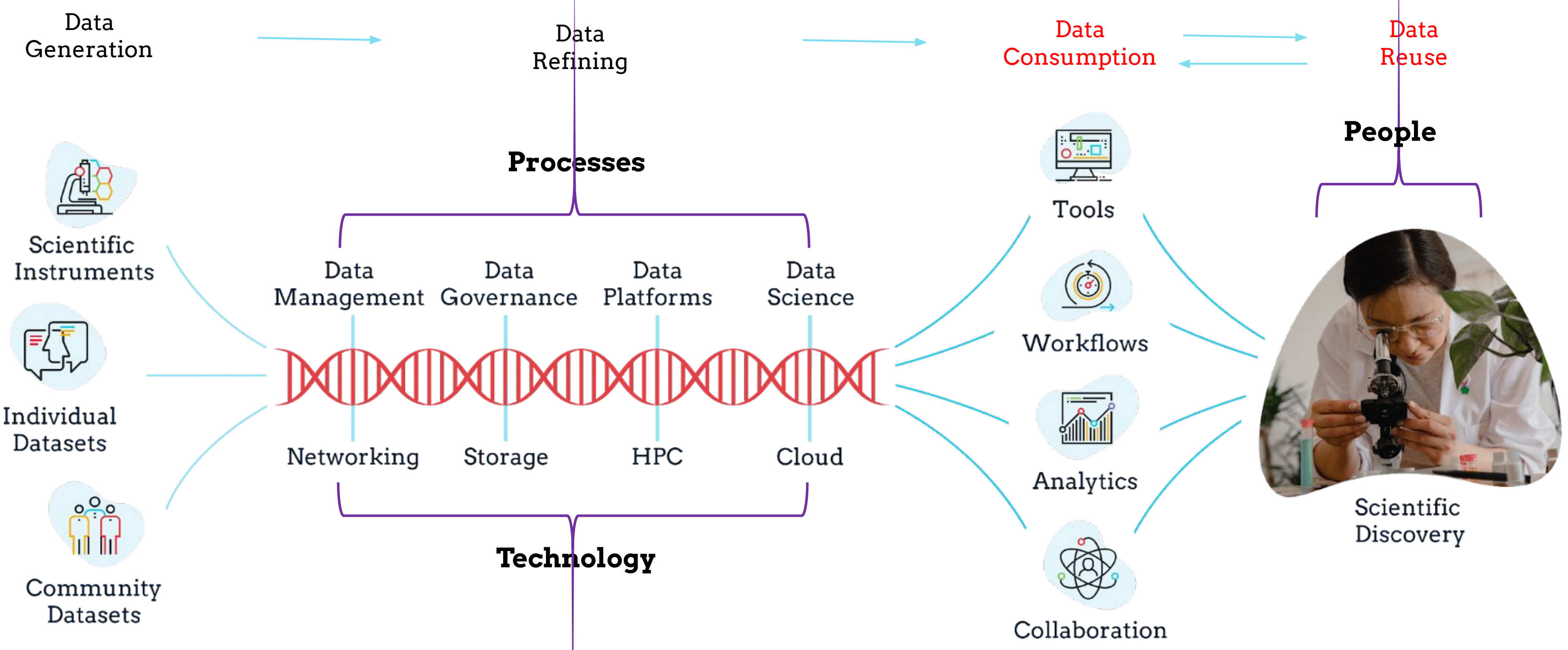
Build frictionless data supply chains for agile data delivery

- Identify barriers in the Data Supply Chain from the data and the infrastructure perspectives
- Identify the right solutions that eliminate bottlenecks and friction given legacy systems
- Seamlessly integrate science, data science and technology capabilities



BIOTEAM
Accelerate Science

Data supply chains harness data and analytics to drive better outcomes

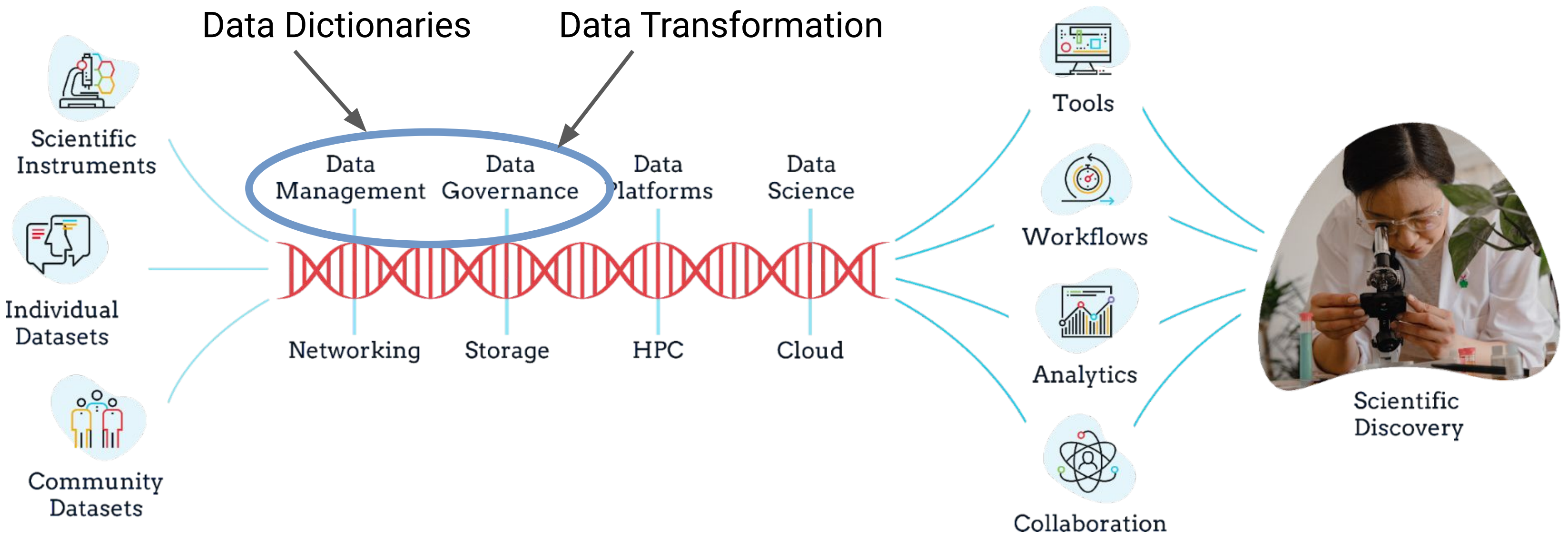


BIOTEAM
Accelerate Science

Making your data analysis-ready
Pitfalls and solutions

William Van Etten, Ph.D.
Co-Founder and Senior Scientific Consultant

Data supply chains: the data lifecycle
My focus



BIOTEAM
Accelerate Science

Data management (FAIR?)

- What data do I have?
- Where does it come from?
  - instruments
  - internal clinical/research silos
  - external public/commercial repositories
  - external collaborators
- How does it all interrelate? (FR)
- What common languages can be used to describe the data? (FR)
- What meta-data can be exposed for search? (FR)
- Who “knows” the data (structure/relationships)? (FR)

BIOTEAM
Accelerate Science

Data governance (FAIR?)

- Who stewards (owns/is responsible for) the data? (A)
- Who should have access to the data? (A)

Data dictionaries make data FAR

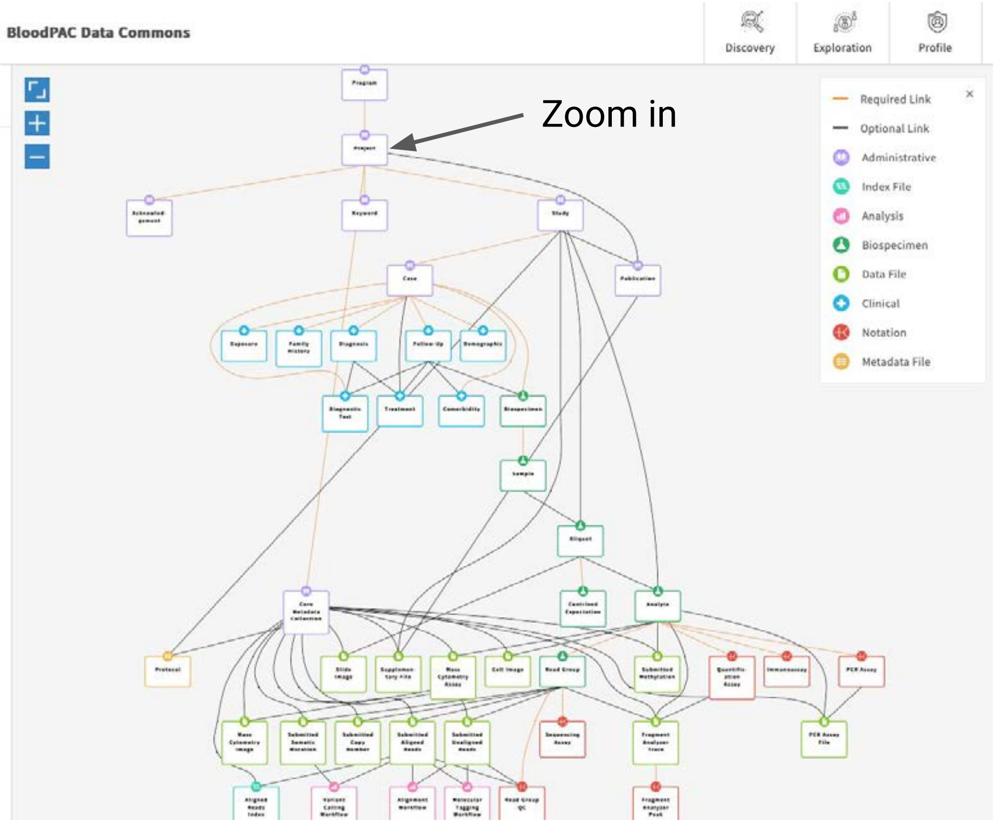
Simple description of what a DD is and why you'd use it

DD: is a validating meta-data schema for Data ecosystems (json) defining the relationships between data elements from various sources, making data FAR.

- Gen3 (gen3.org)
- Terra (terra.bio)
- AWS Omics (aws.amazon.com/omics/)
- Others?
- Or just to stay organized

BIOTEAM
Accelerate Science

Data dictionary example (BloodPAC)



Zoom in

data.bloodpac.org/DD

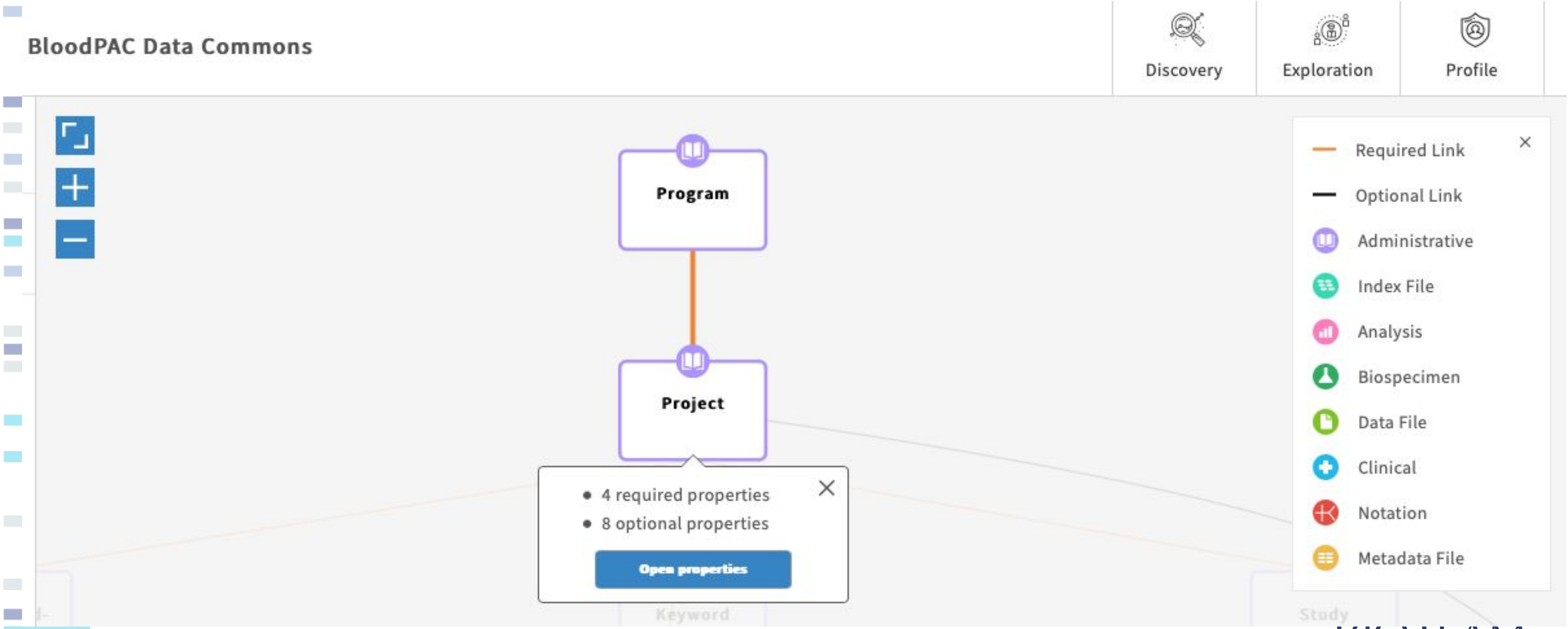
A graph of meta-data nodes

*properties:

- string
- number
- boolean
- array
- object
- enum

BIOTEAM
Accelerate Science

Data dictionary example (BloodPAC)



BIOTEAM
Accelerate Science

Data dictionary example (BloodPAC)

Access

Steward

BloodPAC Data Commons

Discovery Exploration Profile

administrative JSON TSV Close

Project Any specifically defined piece of work that is undertaken or attempted to meet a single requirement. (NCIt C47885)

Property	Type	Required	Description	Term
programs	• array • object	Required	Indicates that the project is logically part of the indicated project.	
name	• string	Required	Display name/brief description for the project.	
dbgap_accession_number	• string	Required	The dbgap accession number provided for the project.	
code	• string	Required	Unique identifier for the project.	
availability_mechanism	• string	No	Mechanism by which the project will be made avilable.	
availability_type	• Open • Restricted	No	Is the project open or restricted?	
date_collected	• string	No	The date or date range in which the project data was collected.	
investigator_affiliation	• string	No	The investigator's affiliation with respect to a research institution.	
investigator_name	• string	No	Name of the principal investigator for the project.	
support_id	• string	No	The ID of the source providing support/grant resources.	
support_source	• string	No	The name of source providing support/grant resources.	
type	• string	No	No Description	

BIOTEAM
Accelerate Science

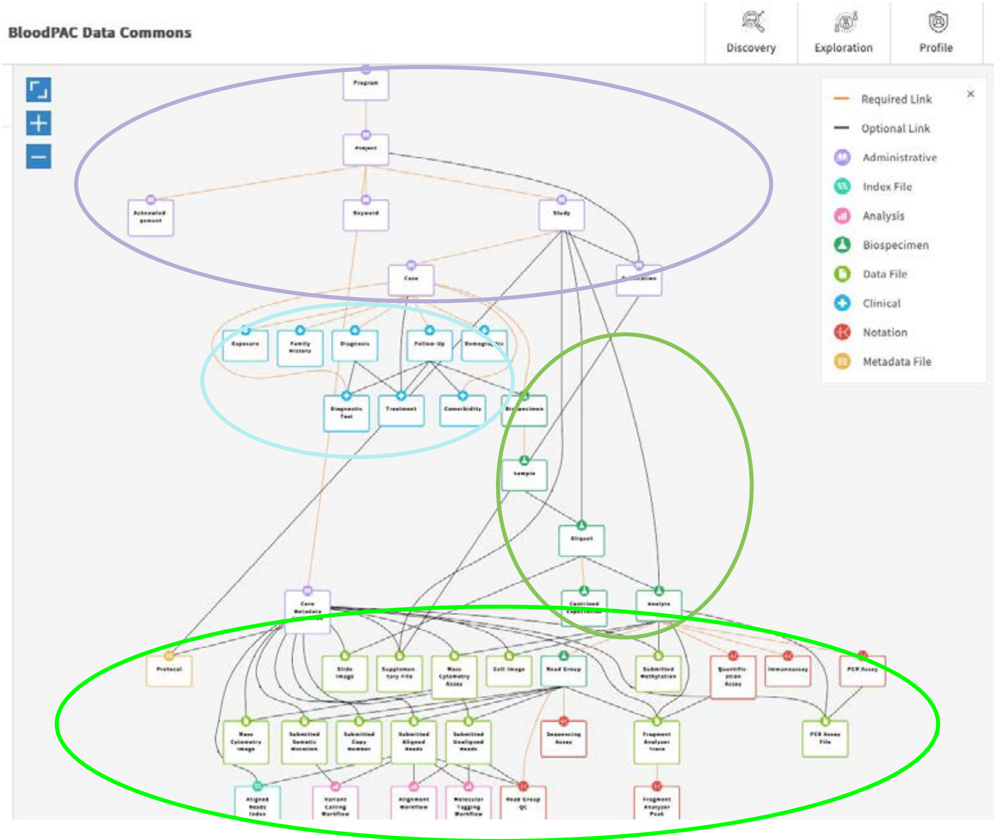
Data dictionary pitfall

Admin

Clinical

Laboratory

Bioinfx



data.bloodpac.org/DD

Who "knows" the data?

BIOTEAM
Accelerate Science

Data dictionary solution

README.md

Data Dictionary

The data dictionary provides the first level of validation for all data stored in and generated by BMS. Written in YAML, JSON schemas define all the individual entities (nodes) in the data model. Moreover, these schemas define all of the relationships (links) between the nodes. Finally, the schemas define the valid key-value pairs that can be used to describe the nodes.

Each **branch** within this repository holds the portion of the data dictionary representing a single BMS data domain. The **root** branch holds the central root of the dictionary that is common to all BMS data domains. The **master** branch holds the current merge of the dictionaries from all participating BMS data domains. The **tagged releases** are **MAJOR.MINOR.PATCH** releases of the master branch.

Visualization

These links below will be automatically updated by a Github Action within a few minutes after creating a branch.

- Travis Build Status BRANCH **master** build passing
- Dictionary Schema BRANCH **master** schema.json
- Dictionary Visualization BRANCH **master** dictionary-visualizer

github.com/bioteam/dictionaryutils

Use standard software development practices to collaborate on data dictionary authoring.

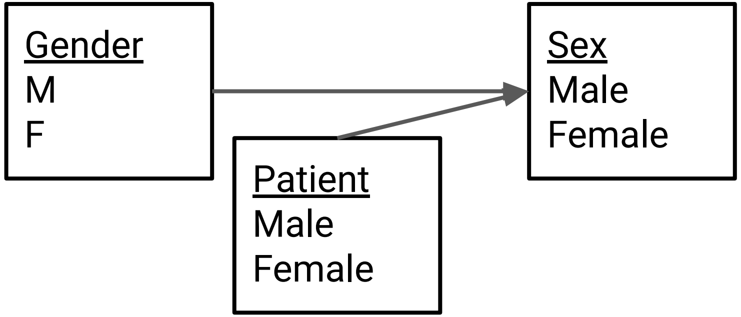
- Groups work on their own branch
- Branches merged upon release
- Build, Test, Deploy upon commit
- Includes serverless DD visualizer

Transformation puts the I in FAIR

Data from multiple independent sources make **I**nteroperability hard.

- instruments
- internal clinical/research silos
- external public/commercial repositories
- external collaborators

Transform before or after load?

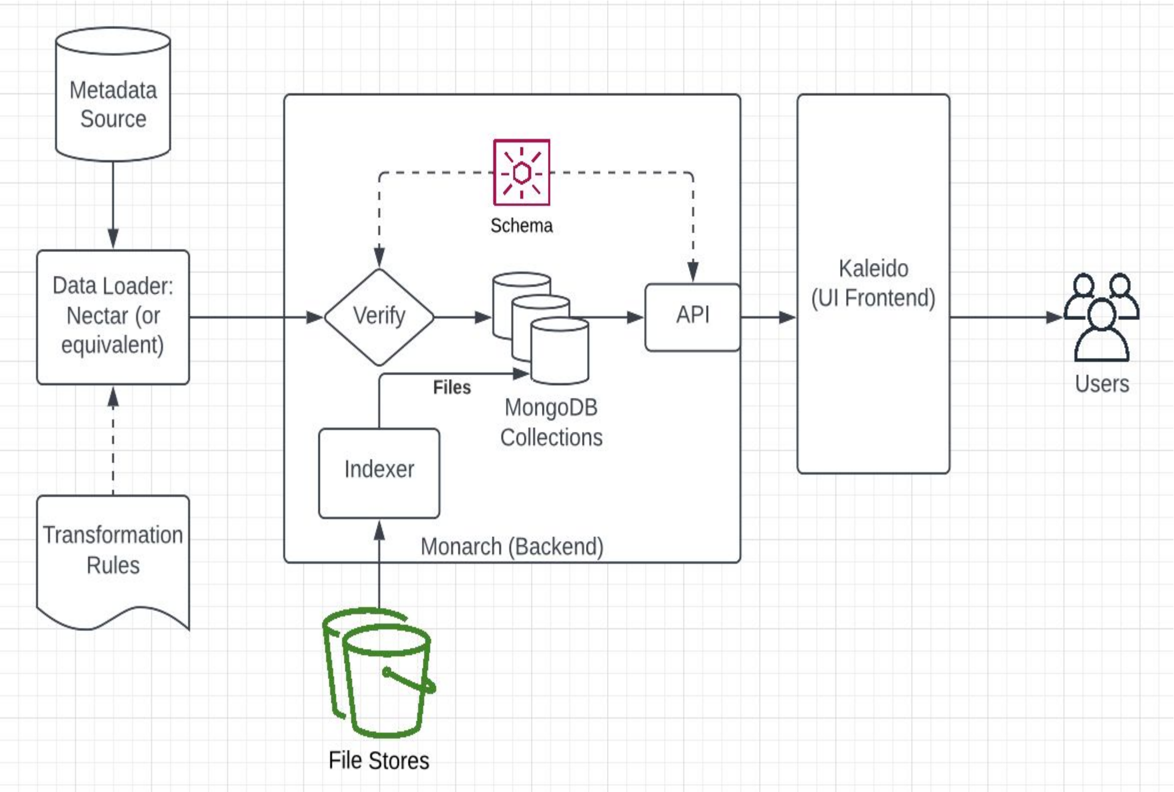


BIOTEAM
Accelerate Science

Transform pitfall (ETL)

ETL (Extract, Transform, Load)

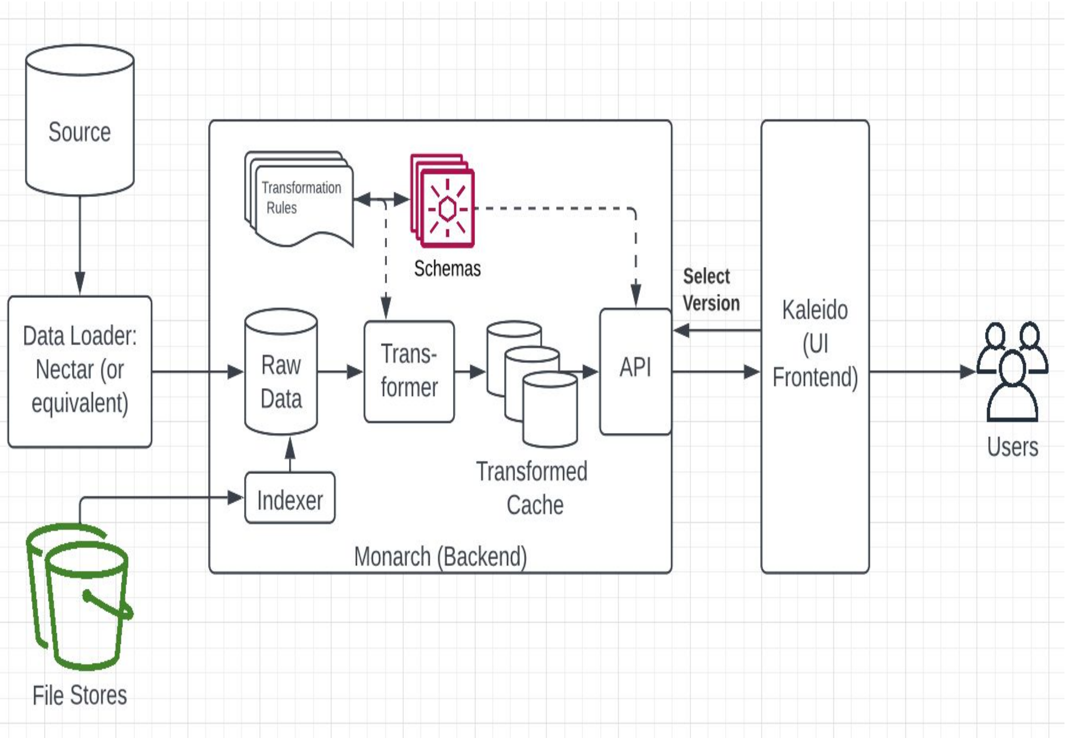
- lose source data
- have only transformed data
- data reload required after every DD change



Transform solution (ELT)

ELT

- Load raw source data
- Transform upon request
- Decouples loading from transformation
- Transformation rules coordinated with DD development
- Permits version controlled and dependable APIs



Webinar coming in February

Join Bill on February 8, 1pm EST for his webinar where he will go into more depth on data dictionaries and data transformation. More details coming soon.

Follow us on social media and sign up for our newsletter to receive the latest updates.

BIOTEAM
Accelerate Science

Choosing the Right Infrastructure to Support Connected Data Ecosystems

Chris Dagdigian
Co-Founder and Senior Director of Technical Infrastructure

Choosing the right infrastructure to support connected data ecosystems

The major transitions in my ~20 year career, part I:

1. **Lab data: Handwritten in lab notebooks → digital storage → ELN**
2. 64bit OS revolution and what it did to memory, storage and compute
3. Refrigerator-sized Unix servers → "Beowulf style" HPC and Linux compute farms on commodity hardware
4. **Research IT storage, networking and compute resources grow larger than infra running the entire org**

BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

The major transitions in my ~20 year career, part II:

5. **Data volumes: Gigabytes → terabytes → petabytes**

6. Networks: Ethernet → fast ethernet → gigabit → 10-Gig → 40-Gig →100-Gig / 400-Gig

7. **Dominant data type by size: Sequence/Genomic → Image-based → (*Future: Time-series ?*)**

8. Virtualization → hyper-converged → composable infrastructure

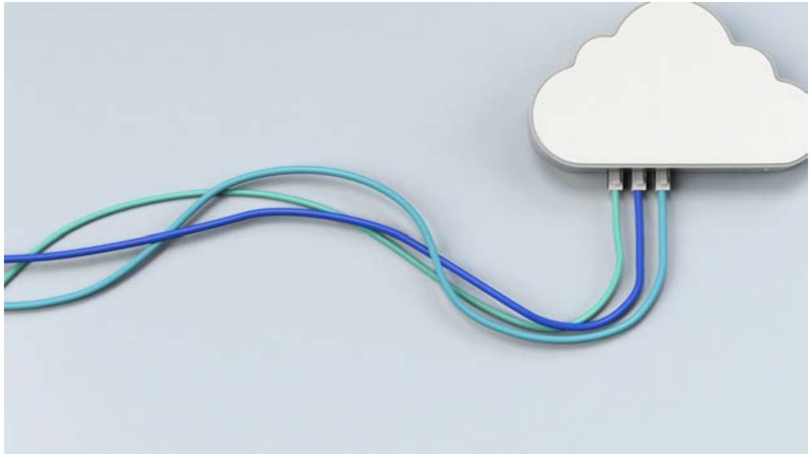
9. Cloud ↔ cloud “retreat/clawback”

BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

Lesson #1: Lean in to the gravitational pull of your data

- Bias your infrastructure to be data-centric with the idea that you will be constantly plugging and unplugging new users, use-cases, workloads and systems into it



BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

Lesson #2: Accept the "*science changes faster than IT*" cadence

- This is your driving philosophy and mission statement
- Accept this cadance and build it into your longer term planning
- The "things" that access your data-centric infrastructure should be modular and support a rapid replace/refresh cadance when business or scientific needs require it

BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

Lesson #3: Infra consolidation may be a lost cause; don't design for it

- Large scale data producers and consumers have diffused "everywhere"
- Complex multi-party/multi-site collaboration is the new normal
- No longer possible to design, assume, or mandate consolidation

The network layer needs to be very fast at the core, inner edge (*labs, IDF, MDF, building-to-building, floor-to-floor*), outer edge (*Internet, Internet2, SDWAN*) and beyond (*direct cloud connectivity*)

Choosing the right infrastructure to support connected data ecosystems

Lesson #4: Consolidate Connectivity

- Can’t consolidate infra, data, or compute but connectivity needs a central hub
  - Large data centers for scientific computing, at least in industry, are becoming harder to operate, connect, and leverage as businesses rapidly change, merge, move, and evolve
  - Yet still need a well-connected “hub” that is “not cloud”

- In ‘22 and beyond this may involve a well-connected colocation facility

- … or a transition to one of the zero-trust/SDWAN-overlay providers

BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

Lesson #5: Storage Design and Operation

- We still design storage assuming "*human browsing files and folders*" is the dominant use case. **We still use filesystem paths to encode simple metadata that are obvious to the eye but perhaps not your tooling**. Foundational storage products (whatever you choose) must treat machine and programmatic access as first class citizens

- Data/storage management still a "grand challenge" and can no longer be considered an IT-only function. Scientists need to be very much involved in storage governance/management from day one. **Reset expectations regarding role of IT in data movement and management decisions to avoid problems**

- ML/AI workloads may require that "all" data be effectively online or nearline

Choosing the right infrastructure to support connected data ecosystems

Lesson #6: Cloud Decisions and Operation

- You need a presence/footprint even if premises-focused for infrastructure
- **Multi-cloud scientific computing is stupid, wasteful, and a sign of poor leadership**
  - ***… at least until we get to the point where K8s runs anything/everything***
  - Pick one cloud to be the primary scientific environment
- Sensible hybrid-cloud patterns are fine
  - Cliché example: AWS for “science”, Azure for “Identity & SSO”

BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

Lesson: #7 Connected ecosystems require multi-architecture support

- Apple moving to silicon based on ARM64 and compelling price/power/performance features of ARM64 in the cloud mean that our scientific computing stacks (*User endpoints, OS, tooling, applications*) can't just assume Intel/AMD x86

- Operationalize support for multi-architecture across the full ecosystem

BIOTEAM
Accelerate Science

Choosing the right infrastructure to support connected data ecosystems

Summary:

- Plan next-gen infra as "data centric" with everything else being a plugin producer or consumer; involve scientists and end-users directly in data management, movement, and governance operations. ***Data can never just be an "IT thing"***

- Plan for disruption similar to the speed at which scientific image data surpassed sequence data by volume in a short period of time. Accept this cadence **and build "reassessment" pauses into your longer term IT roadmap**s

- Build fast networks; **diffuse fast connectivity "everywhere"**

- **Operationalize multi-architecture** (ARM64, X64) support

BIOTEAM
Accelerate Science

Infra trends over 20 years …

"Large scale data producers/ consumers have diffused through the entire organization. IT has lost the data centralization battle."

"Data storage is easy. Data management is still incredibly difficult."

"ML/AI requirements have rendered our long-standing storage tiering/archive strategies obsolete."

"Science increasingly involves complex multi-pa collaboration across phys organizational & network boundaries."

"Future of scientific data at rest is object-based."

"Petascale storage has not been scary for many many years."

"Silicon matters again."

"Cloud is for capability, not cost-saving."

BIOTEAM
Accelerate Science

Wrap Up, Final Thoughts, and Q&A

Ari Berman, Ph.D.
CEO

Summary and Conclusions

- BioTeam has seen a lot of positive change in science and technology over 20 years.
- In life sciences, data is still hard and the industry needs to focus on building functional data ecosystems.
- Thinking about data as a product that brings consumers knowledge may help to unsilo parts of the data supply chain and get us closer to FAIR.
- Collaboratively developed data dictionaries that govern data platforms can help organizations get to FAR, but we might need to rethink ETL to ELT.
- Infrastructure is still both partially worked out and a challenge. The choices (and hype) behind the options are hard to navigate.
- Plan for and strategize around a data-centric model for infrastructure that will allow for data platforms to help manage the supply chain and get us closer to FAIR.

BIOTEAM
Accelerate Science

Thank you for listening: Q&A

THANK YOU! Feel free to reach out to us.

- Ari Berman ari@bioteam.net
- Stan Gloss stan@bioteam.net
- William Van Etten vanetten@bioteam.net
- Chris Dagdigian dag@bioteam.net

Visit our website: bioteam.net

Sign up for our newsletter: bioteam.net/newsletter

BIOTEAM
Accelerate Science