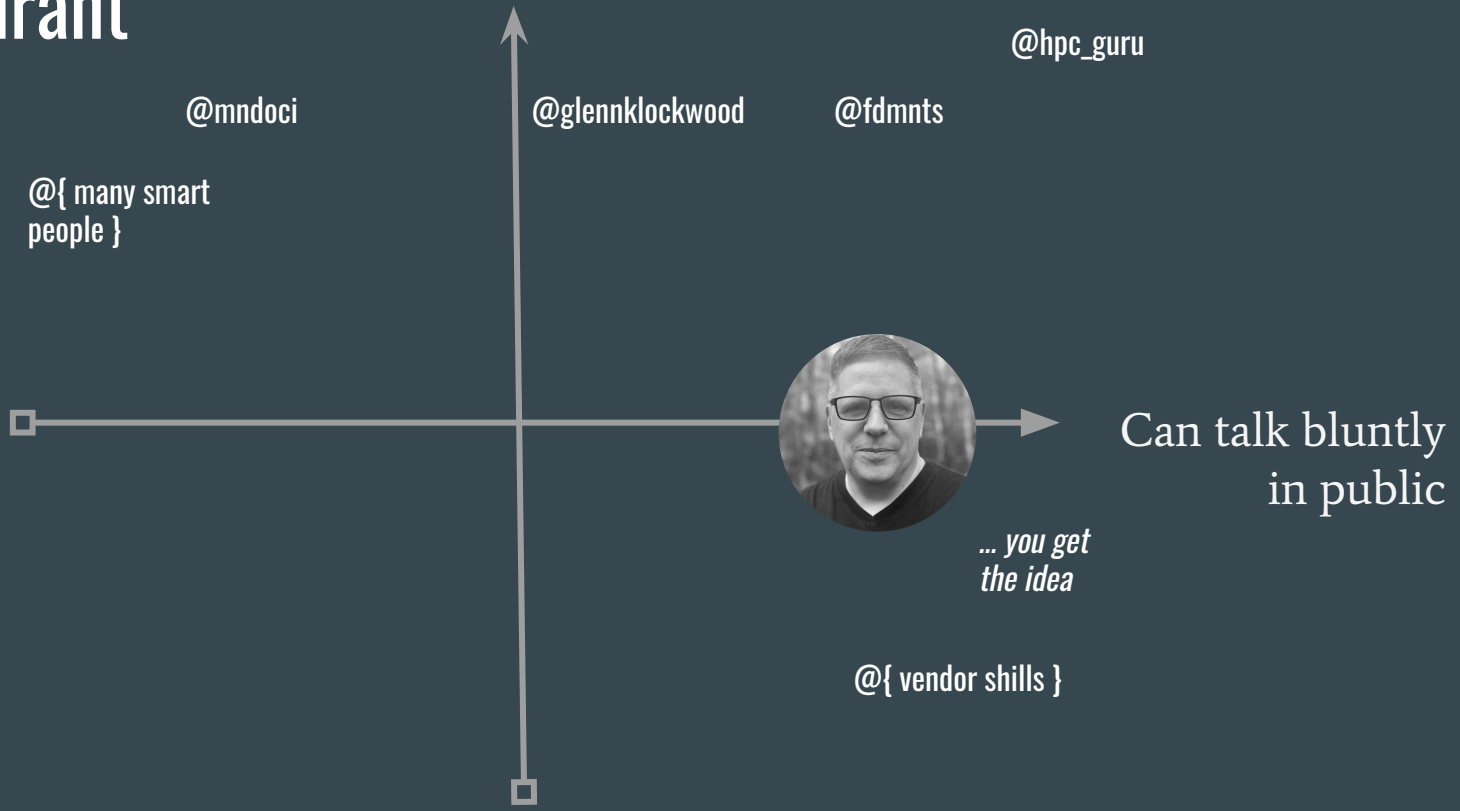
~~Trends~~ Friends From The Trenches




Bio-IT World Conference & Expo 2021

Thought Excretor Magic Quadrant



12 years ago (!)

Trends to Friends: How we got here …

Data Drift: Real Example

- Non-scalable storage islands add complexity
- Example:
  1. Volume "Caspian" hosted on server "Odin"
  2. "Odin" replaced by "Thor"
  3. "Caspian" migrated to "Asgard"
  4. Relocated to "/massive/"
- Resulted in file paths that look like this:

```
/massive/Asgard/Caspian/blastdb
/massive/Asgard/old_stuff/Caspian/blastdb
/massive/Asgard/can-be-deleted/do-not-delete…
```

Flops, Failures & Freakouts

Learning from past mistakes.

#1 - Unchecked Enterprise Architects

- Scientist: "*My work is priceless. I must be able to access it at all times*"
- Storage Guru: "*Hmmm…you want high availibility, huh?"*
- *System delivered:*
  - 40TB Enterprise SAN
  - Asynchronous replication to remote site
  - Can't scale, can't do NFS easily
  - $~500K/year in maintenance costs

#2 - Unchecked User Requirements

- Scientist: "*I do bioinformatics, I am rate limited by the speed of file IO operations. Faster disk means faster science. "*
- System delivered:
  - Budget blown on top tier 'Cadillac' system
  - Fast *everything*
- Outcome:
  - System fills to capacity in 9 months

Data Drift: Real Example

- Non-scalable storage islands add complexity
- Example:
  1. Volume "Caspian" hosted on server "Odin"
  2. "Odin" replaced by "Thor"
  3. "Caspian" migrated to "Asgard"
  4. Relocated to "/massive/"
- Resulted in file paths that look like this:

```
/massive/Asgard/Caspian/blastdb
/massive/Asgard/old_stuff/Caspian/blastdb
```

Lesson Learned:

People appreciated conference talks containing blunt words and real-world stories from hands-on practitioners

- Scientist: "*My work is priceless. I must be able to access it at all times*"
- Storage Guru: "*Hmmm…you want high availibility, huh?*"

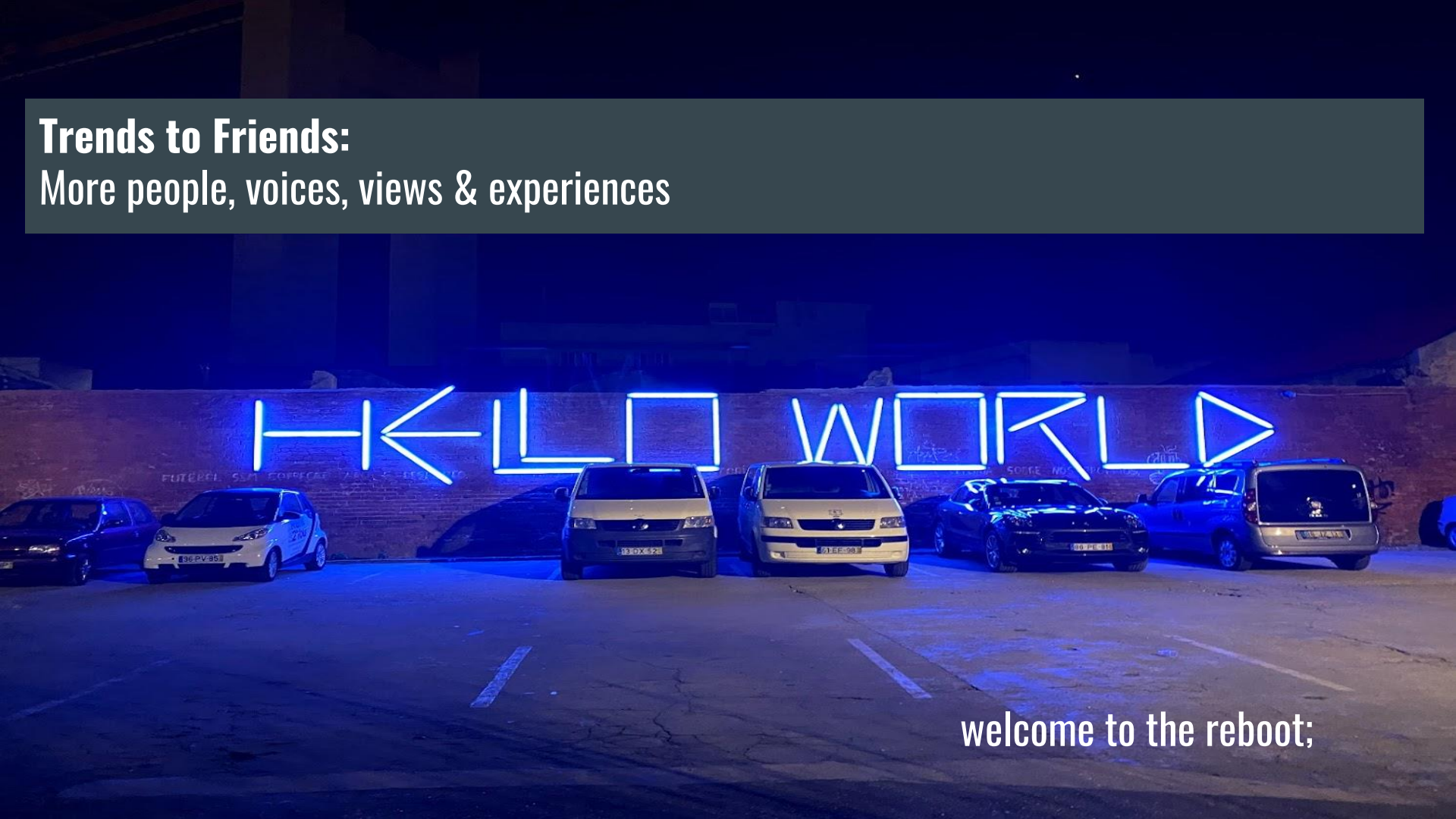
- *System delivered:*
  - 40TB Enterprise SAN
  - Asynchronous replication to remote site
  - Can't scale, can't do NFS easily
  - $~500K/year in maintenance costs

faster science. "

- System delivered:
  - Budget blown on top tier 'Cadillac' system
  - Fast *everything*

- Outcome:
  - System fills to capacity in 9 months

Trends to Friends:
More people, voices, views & experiences



welcome to the reboot;

Adam Kraut: Digital Transformation

BioTeam

Karl Gutwin: Data Transformation

BioTeam

Fernanda Foertter: AI Outside The box

Secret hardware startup still in stealth ... SHHHHH!

Coming Soon:

Stan Gloss: The OG Friend

- Spent 2020-2021 talking to life science leaders and experts via 1:1 interviews and conversations
    - Eric Dishman, Bryn Roberts, Lihua Yu, Matthew Trunnell, Michelle Bennett, Glenn Lockwood, Eli Dart and more …
- Collected stories and experiences into a series of interviews; Thanks to CHI, now available as a PDF E-Book

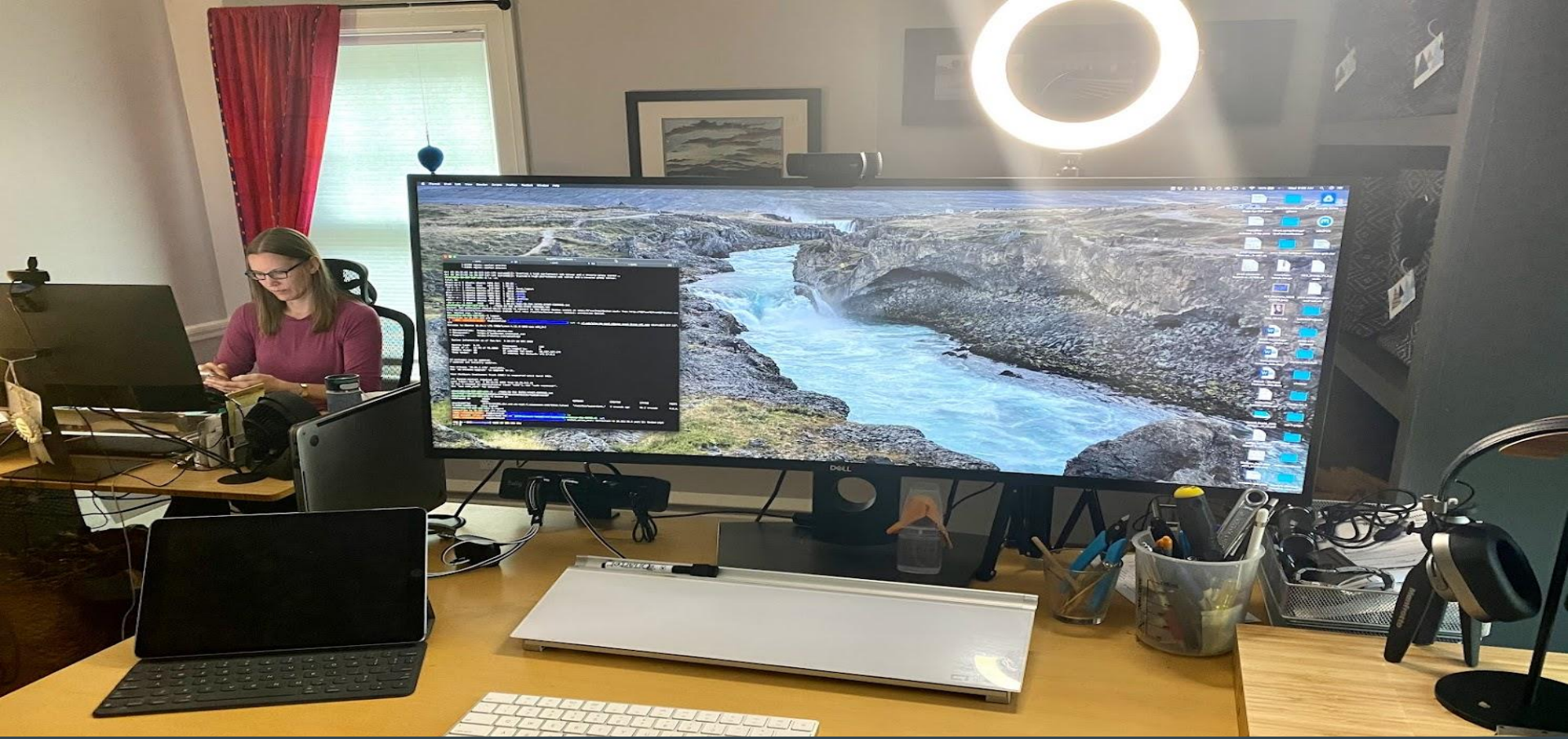

Friend: Chris Dagdigian
Trend: Themes of 2021
Email: dag@bioteam.net
Twitter: @chris_dag

Personae & Places

POSIX

Policy

Petabytes



Personae and Places

Personae and Places

We are all experts on the personal, parental, professional and Org-culture changes that COVID has forced. Let's talk around the edges of some of the new Bio-IT challenges

- Vastly larger and highly distributed remote workforce, including wet lab folk
- New startups selling “API-controlled wet-lab operations as a Service”
- Onsite people may not have fixed desks/offices any more
- “Hard Shell” security, firewall & infosec postures no longer sufficient
- Edge computing to support distributed people, instruments & process
- Effect of Ransomware attacks, Supply Chain SW Attacks & IOT device invasion

End Result:

- Edge Firewalls and “Perimeter Hard Shell” Security Postures Are Obsolete
- Security, identity, authentication, authorization & device recognition requirements now fully extend throughout our entire hybrid premise/remote/cloud/SaaS IT stack
- Security tooling, observability and capability must fully diffuse throughout the organization
- New capabilities that may be required:
  - Pervasive NAC (network access control)
  - Automated device discovery, inventory, profiling and policy application
  - Behavior profiling of unknown or un-classified devices & endpoints
  - Blocking unrestricted internet access within CI/CD workflows and forcing download of patches, containers, libraries, modules, assets and artifacts via controlled repositories
  - Viewing “smart” devices and IOT as high-security internal threats and likely source of intrusion, persistence and lateral movement by threat actors
  - ZeroTrust (people, ports & devices)

Personae and Places

The standard “trends” warning applies here

- Hot and rapid growing IT segments always draw in the marketers, charlatans and snake-oil salesfolk
- Be cynical, be cautious and independently test or verify vendor capability claims
- ML/AI is getting over this hurdle in life sciences but the Infosec-for-LifeSci nonsense is at Peak Marketing

https://twitter.com/cyb3rops/status/1438949275162091524

POSIX

POSIX

Dag for the last few years ...

- “Everybody needs to be peta-capable; not all will require peta-scale”
- “Object Storage is the future of scientific data at rest but it will take years to get there”
- “Purchasing petascale storage is now cheaper than the human cost of managing data @ scale”
- “It is easier to generate/acquire vast piles of data than it is to sensibly manage it over its full lifecycle”
- “Humans are no longer the dominant storage consumer - the “files and folders” paradigm is over”

Dag in 2020

- “ML and AI have fundamentally changed storage architecture and procurement calculus”
  - Can’t bias procurement for size over performance any more
  - Tiering is now much harder and archive discussions are more complex

- IT can no longer coddle scientists who built careers off of data-intensive science refusing to take an active role in managing scientific data. Stop whining and deflecting data curation tasks to IT and TAKE SOME FREAKING RESPONSIBILITY FOR THE STUFF THAT YOUR CAREER IS BUILT OFF OF

POSIX

Dag in 2021:

- Egnyte.com and box.com, huh?
- I guess it's back to "*files and folders*" again ...

Active Trend in 2021

- Almost every single small stealth-mode or startup company I'm working with in 2021 has adopted Egnyte
- Egnyte is embedded in the discovery workflow, starting with direct instrument data capture
- Those that are not using Egnyte are using Box.com for sensitive/regulated scientific data & info

My work in 2022 ...

- Suspect I'm going to be doing a lot of glue/integration/replication work stitching these GUI-biased, human-scale solutions into the large scale data lake and analytic environments that modern data-intensive science organizations require

POSIX

Egnyte or Box style secure storage and file synchronization solutions ...

- Fantastic at local scale
- IT does not have to manage storage, replication, security, DR or backup, what's not to love?
- Feature checklist to make a C-suite IT exec swoon

But from an Enterprise / Data Ecosystem / Data Operations View:

- 100GB max file size. What about petabytes, very large images or time-series/streaming data?
- MacOS/Windows-based GUIs that present "Files and Folders" view of data is quaint @ Enterprise Scale
- We have trillions of files and petabytes of data - a human-centric "files and folders" concept of data organization and presentation should not be the architectural centerpiece and encourages data silo creation
- These products seem to have atrocious Linux support or treat it as as a 2nd class client citizen making it non-trivial to integrate in larger data analytic environments
- Do you know what constant external data sync traffic is gonna do to your cloud network egress fees if not actively managed?

POSIX

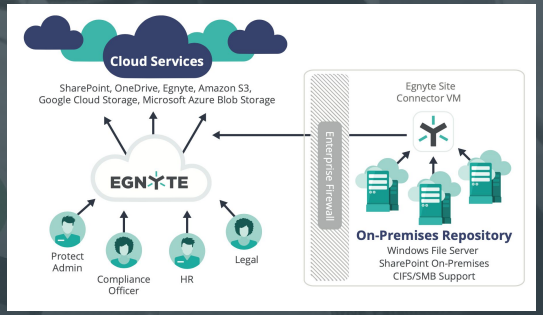
- When marketing owns website copy without supervision you get claims like this:

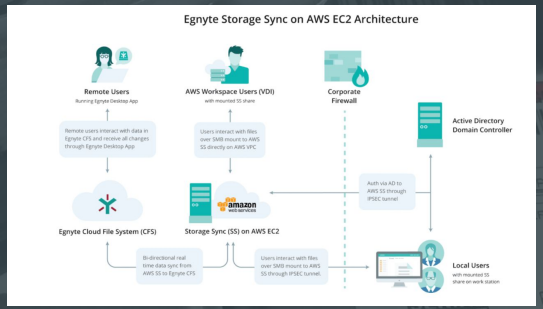
- Downloaded the “File Server Replacement ebook” and got a governance pitch that also assumes Humans are the only consumers:

- The graphical use case for using Storage Sync in AWS is to service Humans consuming storage via Workspaces VDI clients (unless of course your scientists use Linux Workspaces …)

Technology Designed for Enterprises

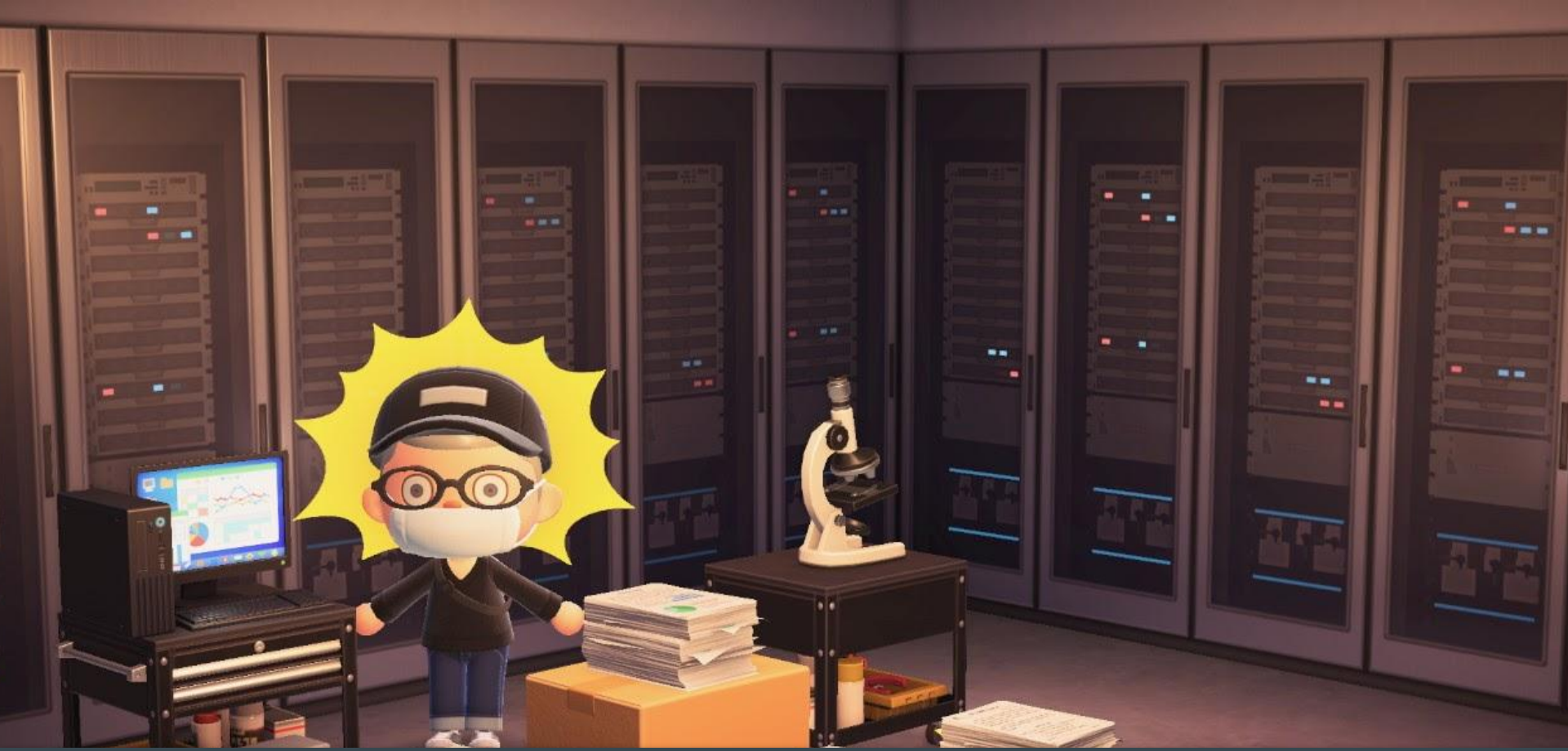
Egnyte was purpose built to address the file sharing needs of businesses. That means you'll be working with a platform capable of handling 100% of enterprise files and use cases.





POSIX

- Honestly not knocking Egnyte. My clients are (largely) happy
- However in terms of “Scientific Storage Trends 2021” the rise of products like this, at least in the startup, stealth-mode and small-but-growing-fast biotech world has been significant to observe
- We (*Research IT, Scientific Computing, HPC, Data Engineers, Data Ops etc.*) are gonna have to figure out how to access, interoperate and integrate with products like this
- Contempt for Linux endpoints/clients and human-centric design bias is a concern for Big Data operations
- My current best guess for small AWS environment (?)
    - Egnyte Selective Storage Sync -> AWS FSx/Windows -> HPC clients using cifs// mounts



POLICY

POLICY

- Significant action in this space in 2020-2021
- We've been hammering the "*it is harder and more expensive to manage data than it is to acquire/store it*" message for a while now. For the first time our consulting workload is starting to measurably change
- "Data Awareness" is a competitive advantage. Failure could be existential threat. Your competitors know this.
- Organizations that used to throw money at expanding storage rather than understanding or curating it have started to hit limits
- Year 2021 so far:
  - ~2 active projects to design/replace a large scale petascale storage environment
  - MULTIPLE active Data Commons projects underway @ BioTeam
  - MULTIPLE active Data Governance / Data Silo Breaking projects underway @ BioTeam
  - MULTIPLE active Assessment projects looking at Org-wide data practices
  - MULTIPLE network refresh/redesign efforts where big data access & movement is key requirement
  - A few of our clients are going for internal "moonshot" efforts to get funding to massively grow out petascale analytic capabilities + budgets for licensing/generating/acquiring new data sources

POLICY

- Ties back to the POSIX section ...
- Starting to see some tension in this space with respect to design patterns
    - Enterprise IT and “human scale” use cases are driving data patterns to POSIX “files and folders” platforms
    - On the R&D, Discovery, Clinical and Commercial Org side the gravitational pull is towards data commons, data lakes and high-scale search, query, aggregate and analysis capabilities that are built for automation and API-driven workloads and pipelines
- This is a fun/interesting space right now. No “one size fits all” solutions
- Not a pure technical problem. Addressing this properly requires internal honest assessments of
    - Technical capability, products, platforms & skills
    - Org Culture (*data sharing vs data hoarding and the career/publication/political drivers that incentivise hoarding*)
    - Org Charts (*what functions sit where*)
    - Org Communication/Collaborations practices (*cross-functional work is hard*)

PETABYTES

PETABYTES

- Gratifying to see the growing trend of people interested in data awareness & data management
- Looking back over the last 12-18 months we still have the same clear Bio-IT drivers and operational challenges
- Buying petascale systems and storing petascale data is pretty straightforward - premise & cloud
- Lets talk
    - Drivers
    - Culture
    - Challenges

PETABYTES - Drivers

Science Driver	Context	IT Impact
Genomics and Bioinformatics	Historically dominant consumer of both storage and compute resources. This will continue as sequencing becomes less expensive and more widely used in both the lab and clinic settings.	• Storage capacity • Non-GPU computing • Large Memory computing • Data Ingest & movement
Image-based data acquisition and analysis	The fastest-growing IT driver BioTeam observes "in the trenches" continues to be image capture and image-based storage driven by the increasing importance of both light (*confocal and lattice light-sheet*), 3D microscopy, **CryoEM,** MRI and fMRI image analysis.	• Storage capacity • Storage performance • GPU computing • Large scale data ingest & movement

PETABYTES - Drivers, continued:

Science Driver	Context	IT Impact
ML and AI	ML and AI techniques are expected to make a significant future contributions to Bio-IT requirements and platforms. These approaches may need hardware beyond the general purpose GPU and may require advanced GPUs, FPGAs, and neural processors.	• Storage performance • Storage capacity • GPU computing • *Cloud workload migration*
Chemistry and Molecular Dynamics	Computational chemistry and MD simulations requirements differ significantly from bioinformatics/genomics requirements. It is worth noting that chemists are capable of consuming nearly infinite amounts of compute capacity -- if more power is available they simply run longer or more complex simulations.	• GPU computing • Scratch storage performance • *Cloud workload migration*

PETABYTES - Drivers, continued:

Emerging	Context	IT Impact
Wearables & Sensors *Time-series data* *Streaming data* *IOT data*	New data types particularly time-series and streaming can be difficult to ingest, store & exploit on existing systems and platforms. This is an area where new products, tech and platforms may be required. IaaS cloud vendors fighting hard to make their platforms the default (AWS Kinesis etc) for these data types	• Data integration • Data storage • Operations • Cloud

PETABYTES - Culture

Event speaker protip: Shrilly hector your audience about the way YOU think it should be ...

Trends that I'm trying to encourage or will into more widespread use:

- DATA AS CURRENCY
- VIEW STORAGE AS A CONSUMABLE RESOURCE JUST LIKE LAB REAGENTS
- THE NEW SOCIAL CONTRACT BETWEEN IT & RESEARCH
- KEEP A TIDY $HOME

PETABYTES - Culture - Data As Currency

It's about the data, not the storage platform

- Hearing leadership say *"all our data is important" is far worse than "... we don't know how to figure out what is important ..."*
- Your data has value; treat it as currency. For too long we've focused on "spend wisely" not "manage effectively"
- Not understanding the true value of data leads to hoarding, massive inefficiencies and inability to properly leverage the data at hand
- Data management, scientifically-relevant metadata, tracking the use, derivative uses, and amount of repeated uses of data could totally change how we approach scientific data storage

PETABYTES - Culture - Storage Is a Consumable

The no-quota, endless expansion, no-justification era is clearly over. We can't afford to operate storage safely and reliably with unchecked growth.

- Management, budgeting & procurement of lab consumables is a well understood thing
  - We should start to view scientific storage the same way
  - Storage is an expensive consumable, users should be expected to have a plan for what they need, how they plan to use it and what efforts it will aid, answer or accelerate
  - Not demanding chargeback accounting or even showback. We just need a slight culture shift away from the belief that IT will always simply keep expanding the underlying storage environment
  - Ties in well if the 2021 "data management" trend is real. Storage analytics and reporting tools will help

PETABYTES - Culture - The New Social Contract

Let's reset the working assumptions between Research IT and end-users

- IT-managed, policy driven storage tiering and management has generally failed us utterly
- Scientific data lifecycles work on a human-controlled "project", "publication" or "project" basis
  - Storage management / tiering based on "file creation date" or "last access time" NOT A GREAT FIT
- IT can't automate around data lifecycle management. Scientist and end-users must take on this role
  - *Dag in 2010: "IT can't make data deletion decisions, that has to come from research"*
  - *Dag in 2019: "Done with coddling scientists who built careers off of big data but refuse to help manage it"*
  - *Dag in 2021: "IT can't automate storage ops with simple policy rules; stakeholder inputs are required"*

PETABYTES - Culture - The New Social Contract

My ideal social contract between IT and scientific end users:

IT Responsibilities

- Storage that meets business and scientific requirements
  - Including scratch, active, nearline, object and archive services
  - Durable, available and reliable
- Metrics, monitoring, reporting and tools that end-users can use to report, manage and understand their data

End User Responsibilities

- View storage as a consumable that requires planning, forecasting and at least minimal justification
- Take ownership and responsibility for data management through full lifecycle
  - Classifying, curating, organizing, retention & archive

Joint Responsibilities

- Large scale data ingest, export and movement

PETABYTES - Culture - Keep A Tidy $HOME

No more large storage allocations to individuals.

- We've seen the problems that unrestricted $HOME filesystems have caused
  - Data provenance issues
  - Replication, duplication and wastage
  - Data hoarding & silos
  - Ownership issues when original person departs

- NERSC has been doing this for years. I want to see more of it in my world
  - https://docs.nersc.gov/ -- great resource to see high quality HPC/storage policy & docs
  - Heavy quota enforcement on "personal" storage allocations and $HOME folders
  - Large storage is allocated to Projects, PIs, Labs or Programs (not individuals)

PETABYTES - Challenges

The hard stuff in 2021-2022

- Friction: POSIX vs Object and having to support both for a very long time
- Friction: Design/build for human “files and folders” browsing or unattended API-driven analysis?
- Friction: Cloud economics and data egress fees @ petascale
- Turning the new interest in data insights, observability & governance into real, actionable long-lasting practices
- Data silo-breaking within an Org
- Extending storage and data analytics consistently & supportably across an enterprise
    - Discovery + Research + Preclinical + Manufacturing + Clinical + Commercial

End; begin->{rusticate};

Thanks for your kind words, attention & feedback all these years!


Friend: Adam Kraut
Trend: Digital Transformation
Email: kraut@bioteam.net
Twitter: @adamkraut

Cloud Computing

Big Data

Internet of Things

AI

Cloud Computing

- Dag in 2009:

Utility Computing

In 2009 I will try hard never to use the word “cloud” in any serious technical conversation …

Digital Transformation as a Leadership Framework

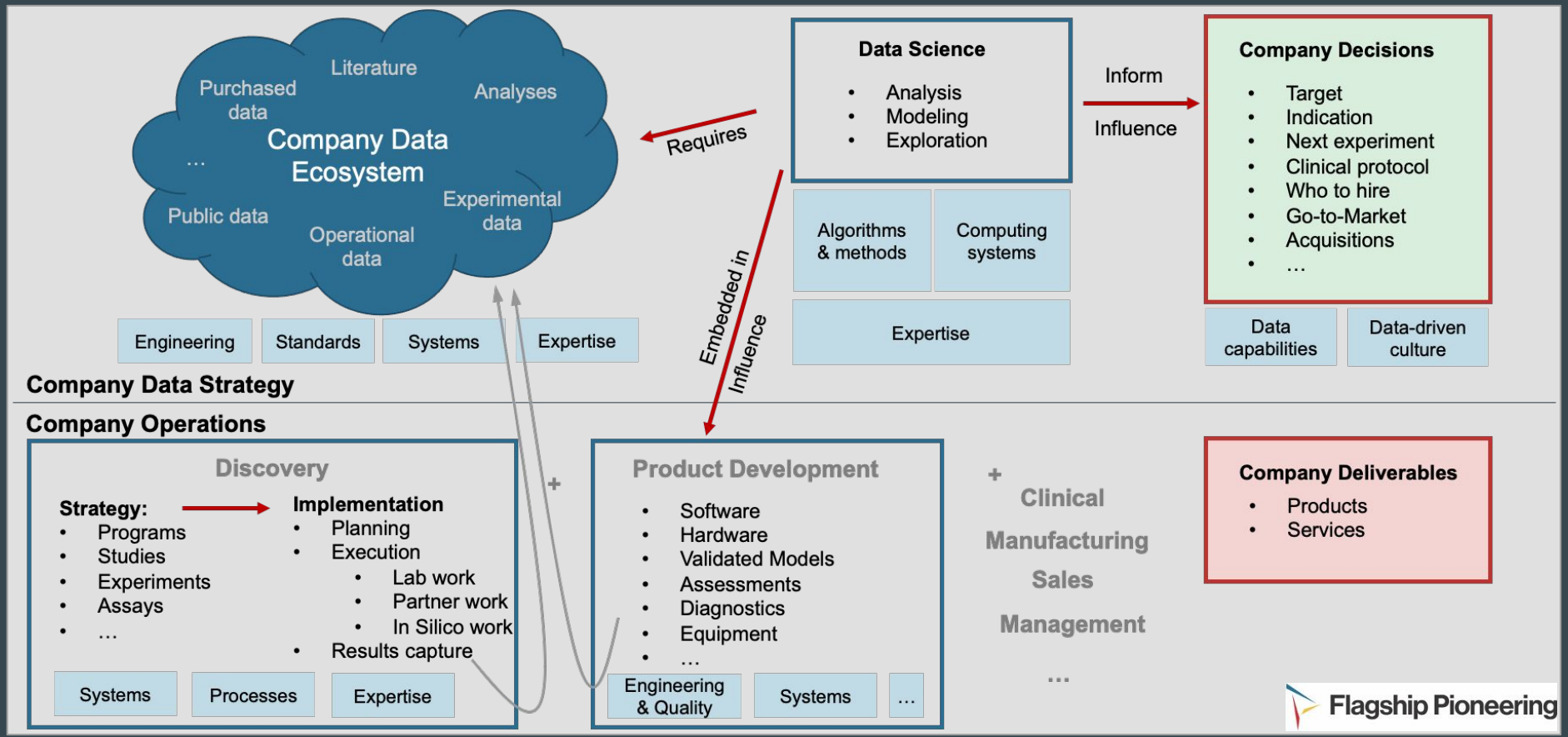
- How can our data inform and influence Company Decisions?
  - Targets, Indications, Next Experiment, Who to Hire, Acquisitions, Go-to-Market
  - Build up our Data Capabilities and foster a Data Driven Culture

- How can we have Data Science embedded in Product Development?
  - Influence Company Deliverables (Productization and Services)

- Performing Data Science (Analysis, Modeling, Exploration) requires solid fundamentals
  - Algorithms and methods
  - Computing Systems
  - Human Expertise

- High-Performance Data Science requires a healthy and robust Data Ecosystem
  - Bring together Experimental Data, Operational Data, Public Data, Literature, ...
  - Data Engineering, Data Standards, Systems, Platforms, Expertise

Digital Transformation as a Leadership Framework



Digital Transformation is a healthy Data Ecosystem

- Your Data Ecosystem is the set of infrastructure and services that empowers a community of data scientists and engineers to make decisions and influence business outcomes

- Key characteristics of a healthy Data Ecosystem:

    - Data Discoverability
    - Data Integrity at the Origin
    - Data Citizenship
    - Common Languages
    - Pipelines and Infrastructure as Code
    - Experiment tracking and shared Workspaces
    - Continuous Delivery mindset for ML and Discovery (DataOps and MLOps)

Data Discoverability

- The primary goal of a data scientist is to locate data, make sense of it, and determine if it is trustworthy or not
- Best case scenario you have clean authoritative data to use for problem solving and decision making

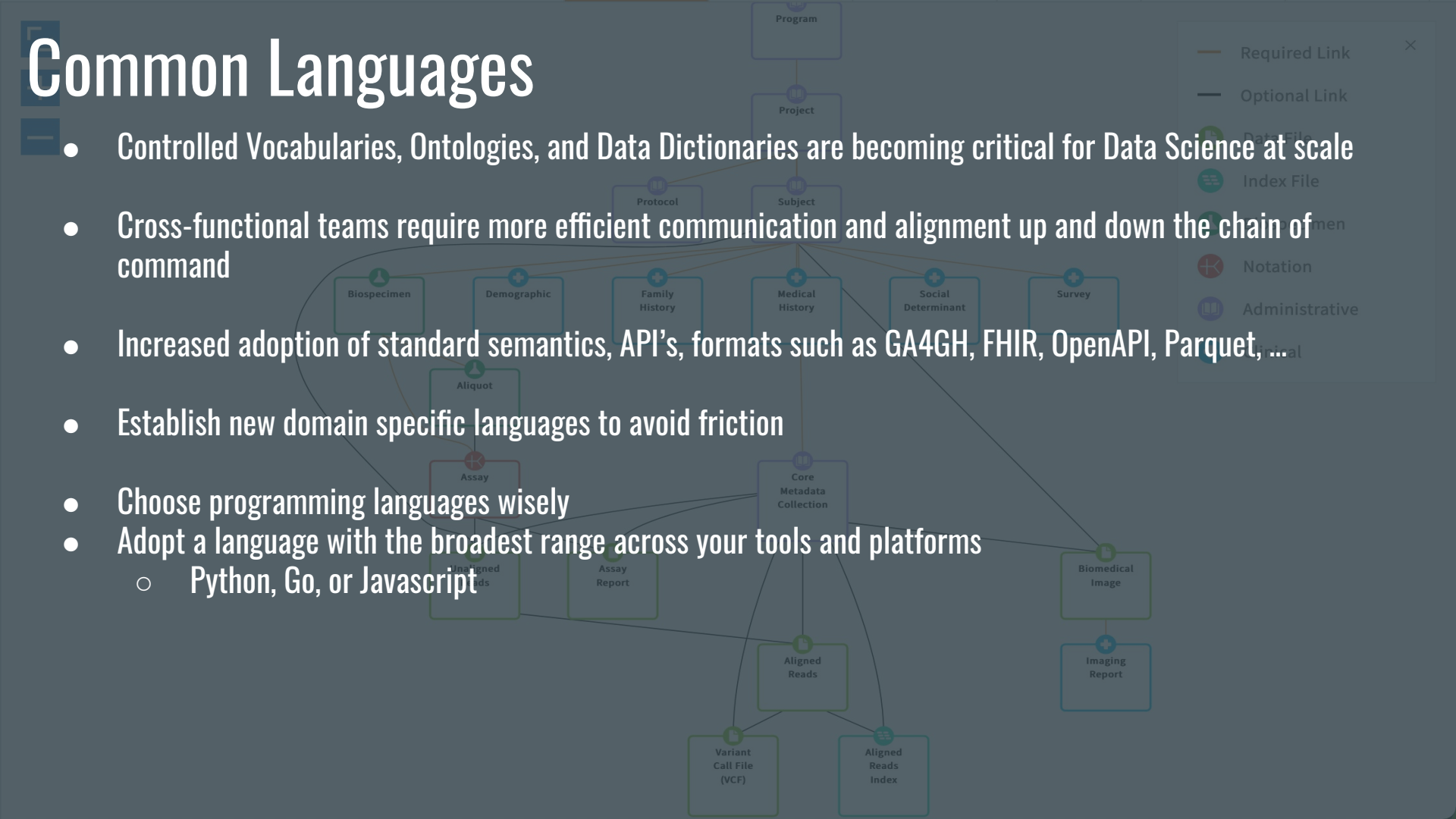
- Datasets often diverge into silos which become problematic
  - Human nature creates silos
  - Applications and databases create silos
  - Businesses and geography creates silos

- Searching and finding data is the primary objective
- Assessing the quality is a supporting objective

- We need well defined metadata and curation at the point of data instantiation
- We need data registration with experiment design, analysis-ready data, and abstract data from storage services

Common Languages

- Controlled Vocabularies, Ontologies, and Data Dictionaries are becoming critical for Data Science at scale
- Cross-functional teams require more efficient communication and alignment up and down the chain of command
- Increased adoption of standard semantics, API's, formats such as GA4GH, FHIR, OpenAPI, Parquet, ...
- Establish new domain specific languages to avoid friction
- Choose programming languages wisely
- Adopt a language with the broadest range across your tools and platforms
    - Python, Go, or Javascript



Pipelines and Infrastructure as Code

- Informatics has benefitted from advancements in software development and Infrastructure as Code
- Puppet > Chef > Ansible > Terraform > CloudFormation > CDK > Helm > ...
- Pipelines as Code: Nextflow, Airflow, Cromwell, CWL
- HPC Clusters as Code: AWS Parallel Cluster (v3.0)
- Infrastructure as Code should make our lives easier not harder
  - Infrastructure is not as static or easily testable like software (1:1 code:resource)
  - Tradeoffs between efficiency gains versus the ROI for creating and maintaining the automation
  - One minor configuration change requires pull requests, code review, build pipelines for CI/CD
- Just because FAANG does it that way doesn't mean your company should too
- Advanced techniques and processes increase complexity and are harder to train and communicate

Continuous Delivery for Data Science and ML

- Discipline of bringing DevOps principles and practices to to ML workflows

- DevOps teams should bridge the gap between ML training environments and deploying models
  - Accelerate the build-test-deploy cycle and shorten feedback loops
  - Eliminate manual handoffs between teams
- Automate the end to end process: Versioning, Testing, Deployments of components (data, model, code)

- Trend towards explainability of models as selection criteria
  - Explainable model allows us to say how a decision is made
  - Avoid biases in data and bad mental models
- May be critical to understanding the fundamental biology and chemistry

Skill Development and Leadership - Go further faster

- High Performance problem solving requires both a framework for skill development and platforms for training
- Skills are perishable... must have a continuous learning mindset to stay sharp (Co-opetition)
  - “Problem Solving is the ultimate skillset”
- Searching for “unicorn” AI or ML expert or 10X engineers should be avoided
  - *“Data Science is a Team Sport!”*
- Diversity is key in high-performance teams
  - Recruit people with mixed talent and experience
  - Include clinicians, lawyers, and other outside expertise

- Every member of the team has an opportunity to lead
- Requires discipline at first and strong communication

- *“Everybody on the team has leadership. You don't have to be a captain to be a leader...Rookies give us leadership, Veterans give us leadership...If they have a good attitude and they work hard and they put the team first, then that's what leadership is all about.”* - Bill Belichick


Friend: Karl Gutwin
Trend: Data Transformation
Email: kgutwin@bioteam.net
Twitter: @kgutwin

Data Flow Automation

Data Commons

FAIR Data

Data Flow Automation

Data Flow Automation

The tedium of data entry and/or transfer by your scientists is real!

If you don't believe me, go into the lab and ask them how they get data from here to there

Why isn't it easy to automate this? Where did we go wrong?

There are just too many proprietary, custom, or constantly-changing data formats out there

Project Review: Mid-size Pharma

Design and build a system capable of taking instrument-generated data from a wide range of sources and make it available to scientists in searchable, raw, and processed formats

Core component of this system was the data processor - receive, reformat, and push to destination

Because of the large number of custom data formats, one key concern was balancing customization versus complexity

Takeaways:

1. Use off-the-shelf tooling when possible - Airflow (or many others...)

2. Consider the delicate balance between custom components, off-the-shelf tools, and your organization's capacity

Data Commons

Data Commons

Purpose: Bring multiple user communities together into one place
- More data sharing, less data siloing

So You Want To Build A Data Commons --- how?
- Use an existing platform such as Gen3?
- Build it as a "data lake" with market-standard components?
- Scratch-build for maximum flexibility?

The choice here is non-linear and multi-factor, so consider carefully

Project Review: BMS

Design a system that can consolidate data from silos across the organization, with a standard SDK and cloud workspace environment

Original design was based on Gen3, but evolved to a custom architecture more in line with their organizational goals

Takeaways:

1. The commons will change its focus over time - embrace it, and anticipate flexibility

2. You will have multiple audiences - be prepared to consider custom portals or workflows for specific user communities

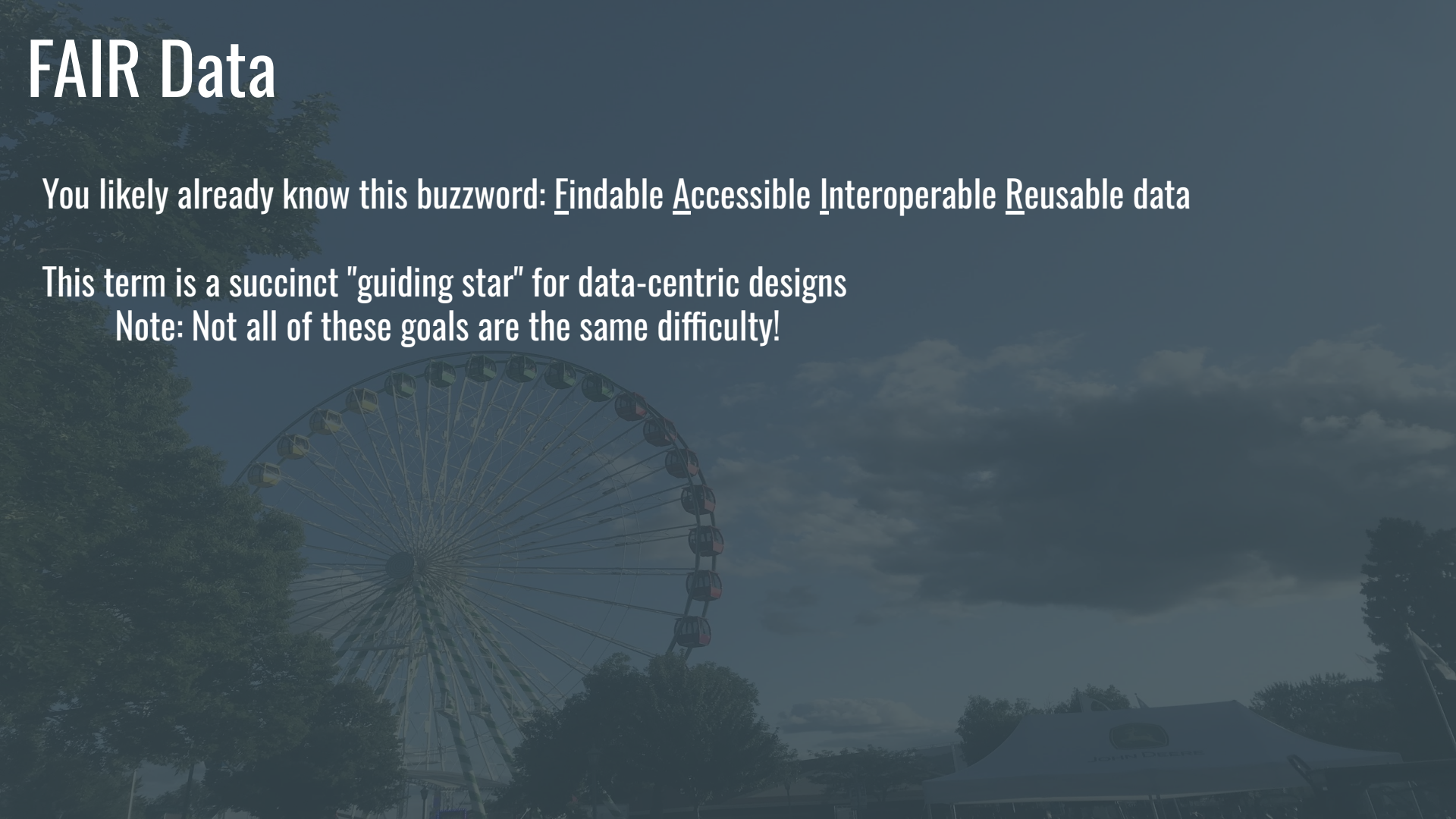
FAIR Data

FAIR Data

You likely already know this buzzword: Findable Accessible Interoperable Reusable data

This term is a succinct "guiding star" for data-centric designs

Note: Not all of these goals are the same difficulty!



Project Review: NHLBI BioData Catalyst

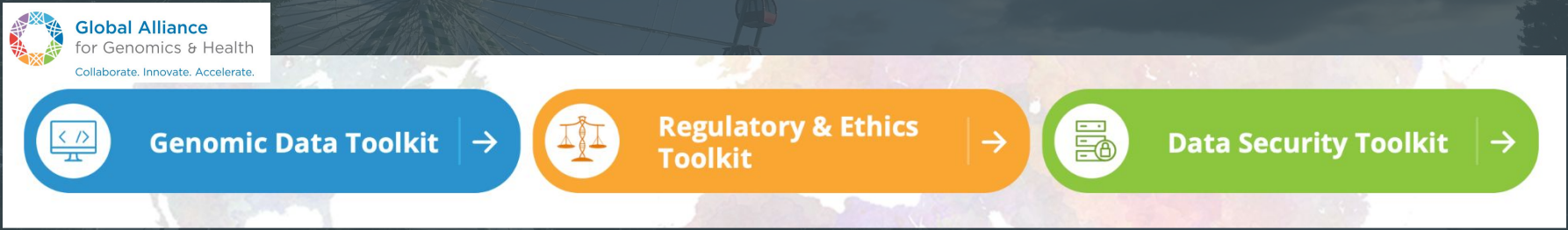
Over the past several years, NHLBI at NIH has built a data commons to share and encourage the use of data from TOPMed and other heart/lung/blood research studies.

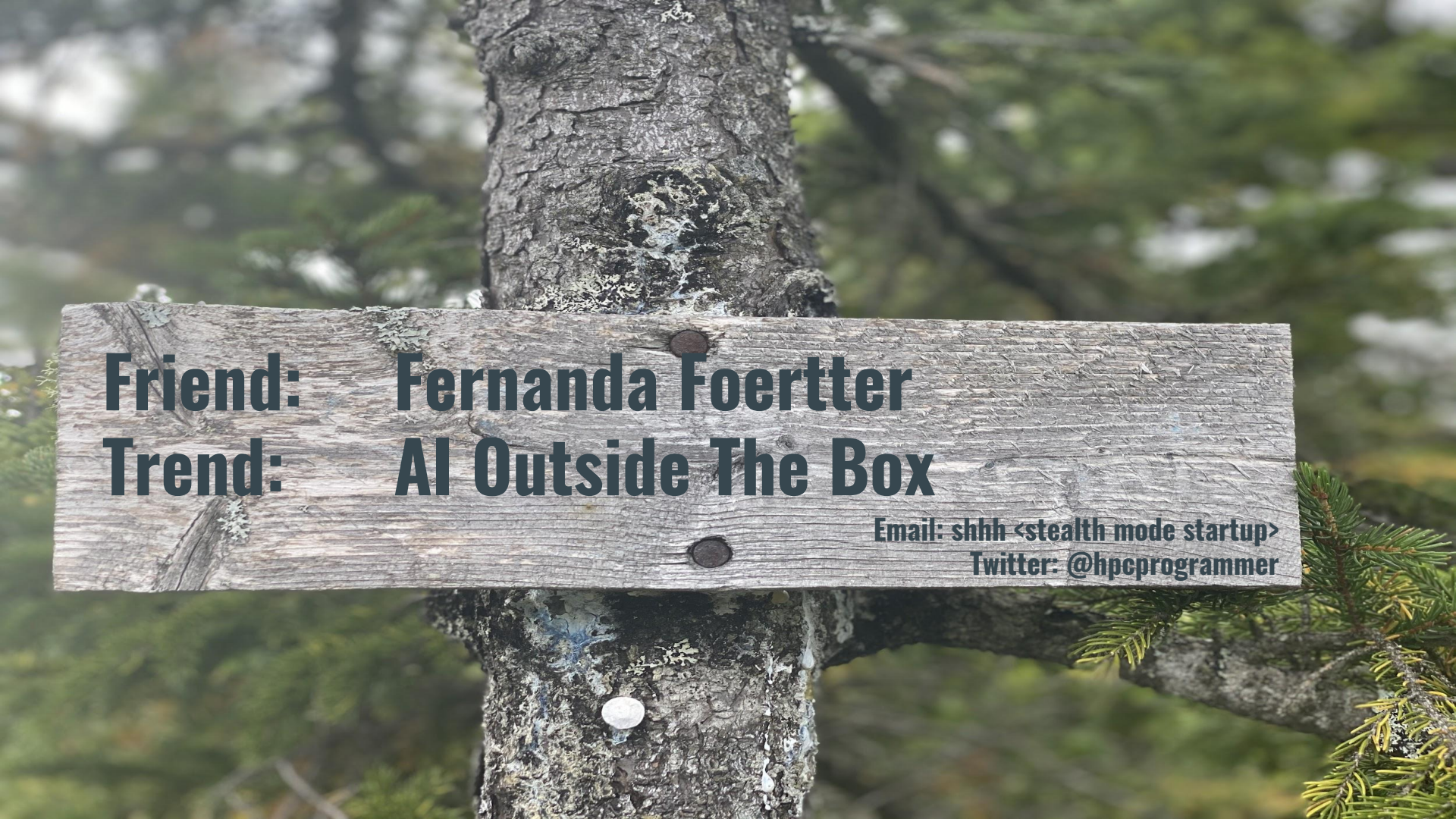
One of its guiding principles was interoperability between components, particularly as part of the NCPI (NIH Cloud Platform Interoperability) effort.

Takeaways:

1. Use standards if you can - look for them first before embarking on a design

2. Ask (even demand!) for standards to be implemented in your preferred platform




Friend: Fernanda Foertter
Trend: AI Outside The Box
Email: shhh <stealth mode startup>
Twitter: @hpcprogrammer

Google research: [Music Conditioned 3D Dance Generation with AIST++]

DanceNet

Li *et. al.*

Ours

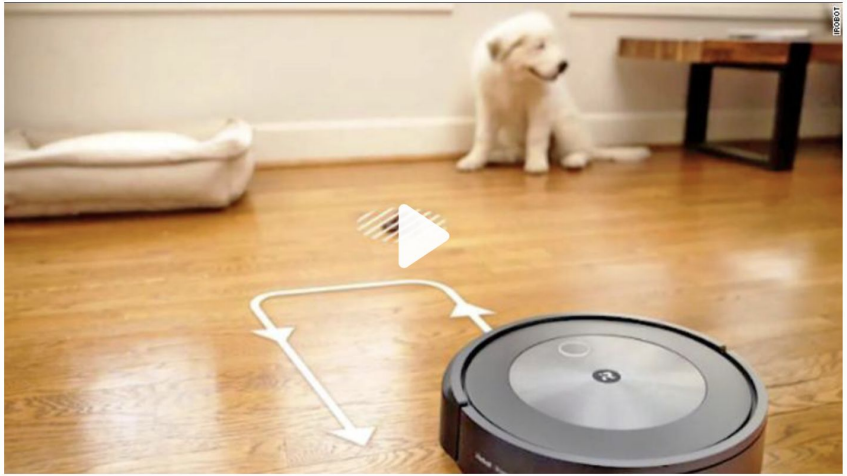
erated motion

The dream...

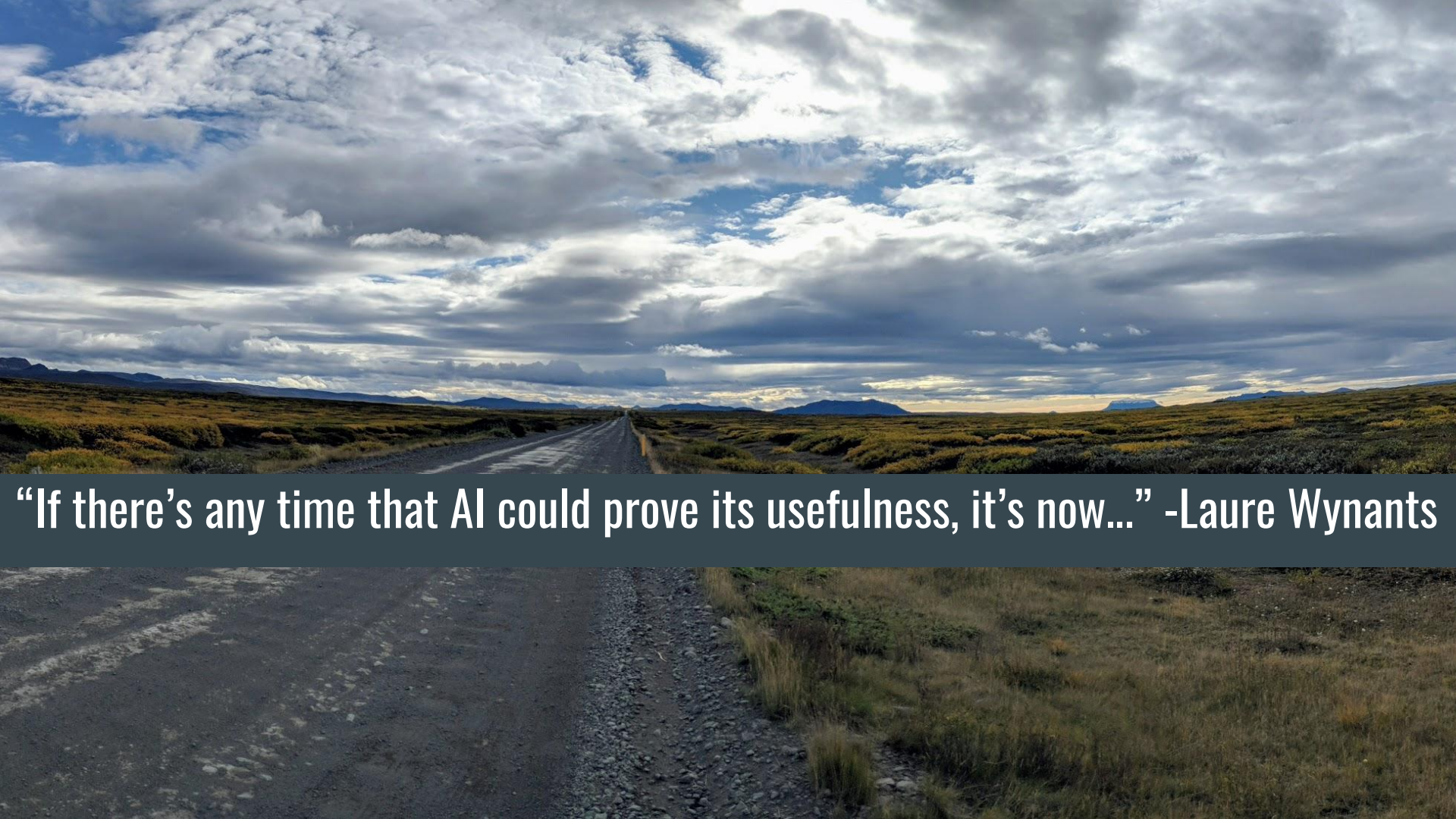
The new Roomba uses AI to avoid smearing dog poop all over your house

By Rachel Metz, CNN Business

Updated 0159 GMT (0959 HKT) September 10, 2021



The reality...



“If there’s any time that AI could prove its usefulness, it’s now…” -Laure Wynants

Artificial intelligence / Machine learning

Doctors are using AI to triage covid-19 patients. The tools may be here to stay

Faced with staff shortages and overwhelming patient loads, a growing number of hospitals are turning to automated tools to help them manage the pandemic.

by **Karen Hao** April 23, 2020

GETTY IMAGES

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

by **Will Douglas Heaven** July 30, 2021

AP

"I thought, 'If there's any time that AI could prove its usefulness, it's now,'" says Wynants. "I had my hopes up."

hundreds of predictive tools were developed. None of them made a real difference, and some were potentially harmful.

The Turing Institute said AI tools had made little, if any, impact in the fight against covid.

232 algorithms for diagnosing patients or predicting how sick those with the disease might get.. none of them were fit for clinical use.

They looked at 415 published tools and concluded that none were fit for clinical use.

Reality check: AI is primarily a research tool for now

Your organization is likely swimming in data

All of it is disconnected

Nothing is actually usable today

The person that understood any of it, left

Marketing Dept has already published the piece announcing it

On data...

“Just because you can search it doesn’t mean you will find it.
Just because you find it doesn’t make it usable.”

-- Fernanda Foertter @BioIT Workshop 09/21/21

What companies are doing right

- Hiring Data Curators
- Hiring Ethicists
- Baking in ways to tag data into already existing processes
- Data commons/lakes
- Starting with fresh new data
- Targeting AI to improve internal processes
- Finding ways to share or buy data
- Reusing mature algos from Google/FB etc.
- Solving narrowly focused problems
- Hiring consultants

What companies are doing wrong

- Buying AI startups
- Ignoring ethics and bias
- Adding complexity to processes
- Proprietary data lakes
- Starting with historical data
- Targeting AI for customer facing apps
- Believing their data is enough
- Building their own AI models
- Trying to solve broad problems
- Ignoring AI altogether (wait and see approach)

If you give a mouse AI...

- It’s going to want more data
- More data will need better data infra
- Better infra will need long term thinking


Make 2022 the year of building
good infrastructure for data

How to learn to stop worrying and love AI:

- Start with a fresh project
- Generate new data
- Keep the problem narrow
- Use it to improve internal process
- Reuse models that already exist