PLENARY KEYNOTE PRESENTATION

Bio·IT World
CONFERENCE & EXPO

TRENDS FROM THE TRENCHES

CHRIS DAGDIGIAN
Senior Director, BioTeam, Inc.

MATTHEW TRUNNELL
Data Commoner

ANNA SOWA, PHD
Senior Scientific Consultant, BioTeam, Inc.

MICHELLE BAYLY, PHD
Senior Scientific Consultant, BioTeam, Inc.

ADAM KRAUT
Director Infrastructure & Cloud Architecture, BioTeam, Inc.

#BioIT22

BIOTEAM
Accelerate Science

Trends From The Trenches




Bio-IT World Conference & Expo 2022

#BioIT22

Want these slides?
https://bioteam.net/present22/

Thank You For Being Here!

- Last few years have been weird and stressful
- We are all excited about the chance to revisit, reengage and reconnect with the larger Bio-IT community
- We've got a lot to catch up on!

Want these slides?
https://bioteam.net/present22/

New? Welcome! Have some background ...

- Started out as a 45min “Track 1 - Infrastructure” talk focusing on HPC and scientific computing topics seen via BioTeam’s consulting work
- Focus has always been on blunt, honest and practical lessons learned from real world projects
- As consultants, we get to see how different groups of smart people tackle similar challenges
- We can (generally) speak in public about what we’ve experienced and learned - no filter, nothing to sell and no marketing BS

@hpc_guru

@{ many smart people not allowed to talk in public }

@mndoci | @boofla | @delagoya

@glennklockwood

@hpcprogrammer

@fdmnts

@DrCuff

@QuinnyPig

... you get the idea

Can talk bluntly in public

Thought Excretor (™) Magic Quadrant

@{ vendor shills & lying marketers }

Competence / Domain Insight

HPC From The Trenches

2010 BioITWorld Conference & Expo

Track 1 – IT Infrastructure & Ha

BIOTEAM

Twitter: chris_dag
Email: <dag@bioteam.net>
LinkedIn: chrisdag
Slides: slideshare.net/chrisdag
Jobs: bioteam.bamboohr.com/jobs/

Trends Newsletter: https://bioteam.net/newsletter/

TRENDS FROM THE SOCIALLY DISTANCED TRENCHES

Trends from the trenches: 2014

slideshare.net/chrisdag/ chris@bioteam.net @chris_dag #BioIT14

slideshare.net/chrisdag/ chris@bioteam.net @chris_dag

Dagdigian had led the project to bui
Pharmaceuticals Inc.'s giant cluster. Van
a big project at Biogen. Athanas, as direc
entific consulting, had his hands in all pr
When Blackstone's business morphed
ly software sales during 2001, the four

, from left: Michael Athanas, Bill Van Etten,

Dagdigian says. They do encounter oc
torage and data integration specialists.
"We know the hardware, the infrastruc

Trends from the Trenches: 2022 Edition

- Every time I talk about retiring from this talk, Stan and others get very sad
- Feedback summary:
  - *“New voices and new focus areas are great but people still want to hear the hardware/HPC/infra/trends bits …”*
- Message received

Don’t make this man sad.

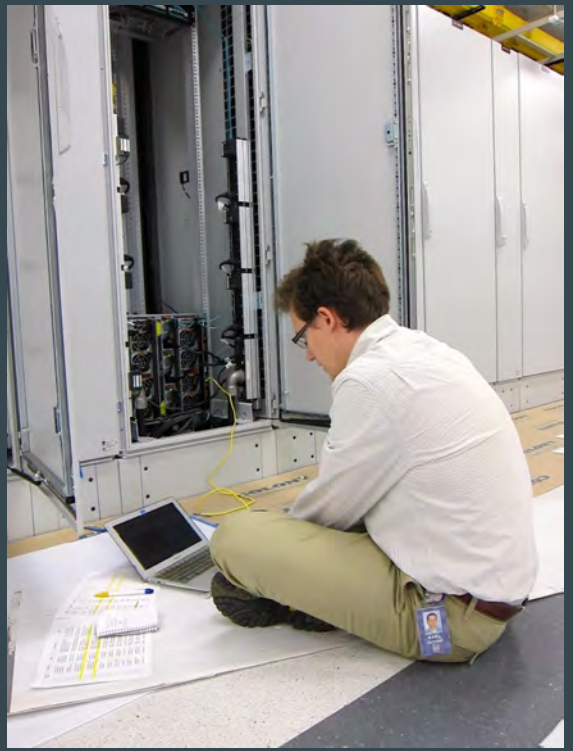
{Mini} Trends 2022

10-Year Recap

Life Science Informatics Mini-Trends 2022

Three Topics Today

- ~10-Year Trends Recap
  (unchanging constants in a world of chaos)
- Squishy Stuff
  (mistakes and things in transition)
- ~1-Year Trends Recap
  (interesting things seen in COVID era}



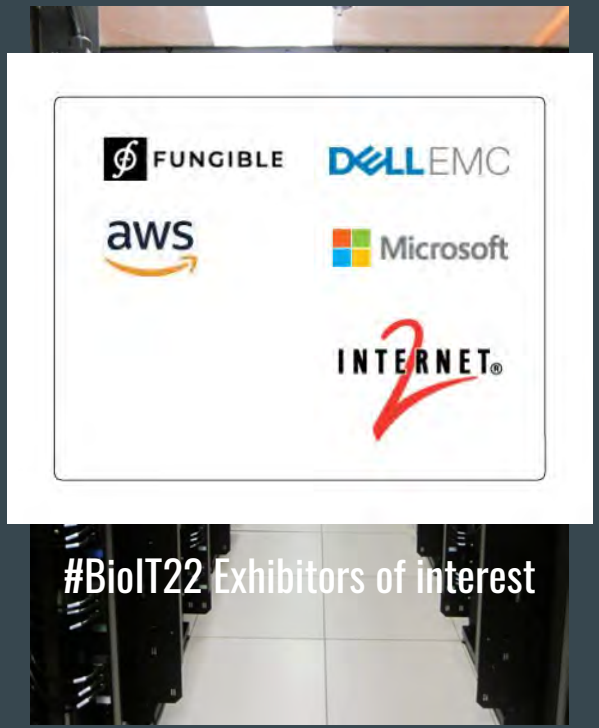
10 Year Recap: What has not changed over time?

"The Existential Dread Problem"

- Science always changes faster than IT can rebuild datacenters, refresh infrastructure or re-architect cloud platforms

- Rate of change in science: Monthly
  IT refresh cycles: 3-4 Years if lucky

- This is THE FUNDAMENTAL problem when supporting life science informatics; Huge risk when IT builds or guesses wrong ...



10 Year Recap: What has not changed over time?

Cloud is for CAPABILITY; not COST

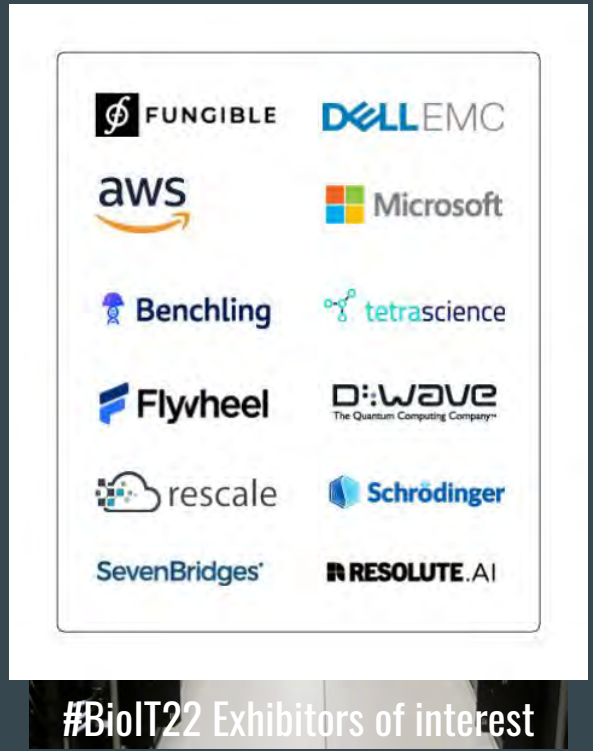
- Moving to the cloud is a play for capability, agility, elasticity and flexibility. It RARELY saves money.
- Despite cost implications the capability gains & flexibility justifies cloud adoption.
- Cost optimization occurs as cloud maturity increases.
- Multi-Cloud is Stupid; Hybrid-Cloud is Fine.

10 Year Recap: What has not changed over time?

Compute generally is a solved problem

- Premise, cloud, colo or hybrid–does not matter
- Bringing compute power to bear on scientific problems has not been particularly difficult for a long time
- Mostly a financial planning exercise with a bit of guidance/guardrails supplied by scientific leadership in order to constrain the "*can consume infinite compute*" people*

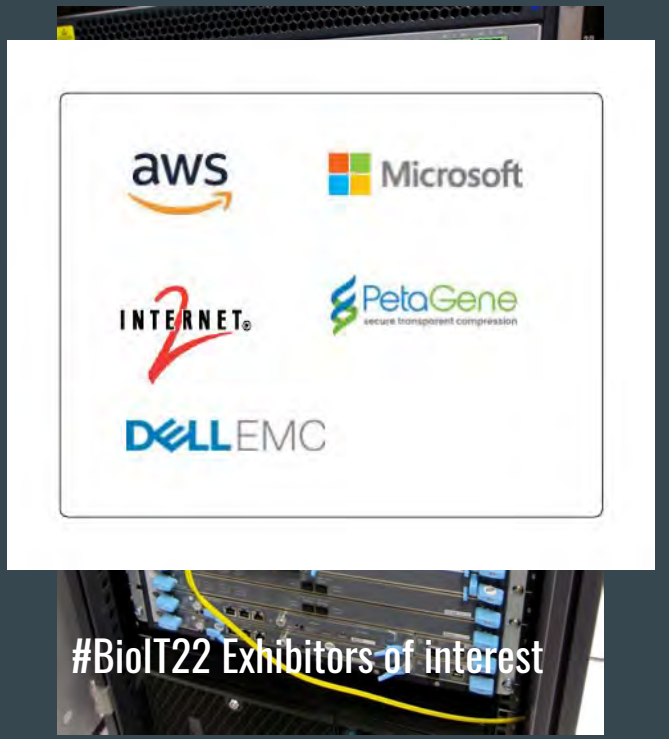
*(*hi chemists, modelers, protein folders and structural folk!)*



10 Year Recap: What has not changed over time?

Networking is a solved problem

- No longer hard, risky or exotic to build high-speed or low-latency networks (if you get off of Cisco ..)
- We regularly see 25Gig, 40Gig, 50Gig, 100Gig networks being deployed by clients; Internet2 has been running 400Gig network links since mid-2021
- The only hassle/mistake: Orgs that build “WiFi only” spaces or don’t account for future need to pull fiber to special wetlab or desktop locations
- ScienceDMZ patterns & out of band traffic mirroring allows for network monitoring and security at scale



10 Year Recap: What has not changed over time?

Large storage is a solved problem

- Petabyte+ capable storage platforms are the norm, not the exception in 2022
- Cloud, Colo or On-premise (does not matter)
- Many vendors, many options, all with track records and referenceable life science customers
- *Not solved: Data movement and Data Management*

#BioIT22 Exhibitors of interest

10 Year Recap: What has not changed over time?

Fundamental storage use cases still apply

- Overwhelming use case for scientific data remains

  “shared read/write access for instruments, automation, pipelines and people”

- SAN or LUN based storage not ideal unless you are being clever and have a plan
- File sync platforms (Box, DropBox, Egnyte, etc.) are solid/proven point solutions but may not be cost effective at scale and are unsuitable for a number of critical use cases; augment these with “something else”
- Object storage or Scale-out NAS are the starting points; Move to something else if business/scientific need requires

10 Year Recap: What has not changed over time?

Storage & data management requires shared responsibility

- IT cannot “own” data management. Period
- Data management and governance has to come from Scientific leadership
- Partnership required:
  - IT provides supported, durable fit-for-purpose ways to store data
  - Science/IT provides tools for data awareness, governance, movement and management
  - End-users must take some responsibility for managing, moving, curating, annotating and handling data through a complete lifecycle

- Scientists who build careers and publication lists via “data intensive science” yet refuse to take responsibility for their own data or engage with IT really make me angry
  - I’m long past coddling those types and will call them out at every opportunity

{Mini} Trends 2022

Squishy stuff (mistakes & transitional trends)

I was so smug (and so wrong) about this 10+ years ago ...

"The future of scientific data at rest is object storage" -- some jerk

Why?

- Erasure coding type methods most cost effective way to store petascale data durably
- Object stores support metadata and tagging of objects–fantastic for scientific data management, querying, search and automated management
- Purpose built for FAIR, data discovery and API/automation heavy future where humans are no longer the dominant consumer of files and data

I was so smug (and so wrong) about this 10 years ago ...

"The future of scientific data at rest is object storage" -- some jerk

What went wrong ...

- We still assume "human browsing a folder" is the dominant use case
- We still treat POSIX folder names and file paths as a reasonable way to capture essential metadata:

```
/data/KatesLab/2019/joe/project22/microscope01/
experiment-01/controls/mouse23A1-brain-03.tiff
```

- Still use Active Directory or POSIX for access permissions and ACLs
- Tons of commercial code not object-aware
- Tons of open-source code not object-aware
- Many scientists who just need to transform data in R or Python are not "object aware", don't have time to learn, don't really see the benefits from changing current methods

Transitional Trend

ML and AI have really messed up long-held storage design patterns in our world

- For a long time my ideal was “single global namespace”
- Then I made peace with different namespaces if they came with compelling capability
  - Tiers: `“/burst”, “/active”, “/online”, “/nearline”, “/archive”` etc.

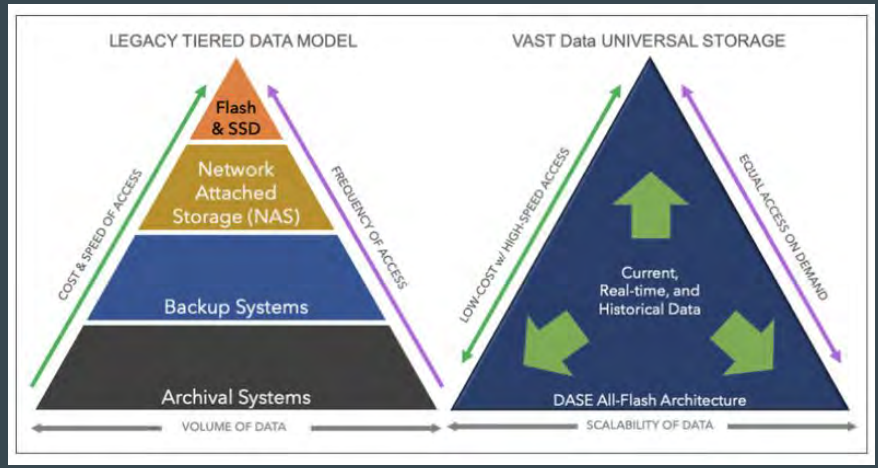
- Machine Learning and AI sort of blow up the tiered storage argument. Why?
  - With ML and AI *all* your data is potentially active and in-use
  - Your storage layer needs to be fast and online for all data (old and new)

- This is a new world for us–making all our data always available via performant methods lest we screw up or slow down the ML/AI folk

Transitional Trend 02

ML and AI have really messed up long-held storage design patterns in our world

- This week VAST Data began sharing links to a Futurum Analyst paper called “Is this the end of tiered storage?”
- This is an area ripe for disruption and displacement of storage incumbents
- VAST, WEKA, HAMMERSPACE & DellEMC are all here at #BioIT22 and I’m sure they’d love to chat about this with you



Source: https://futurumresearch.com/research-reports/is-this-the-end-of-tiered-storage/

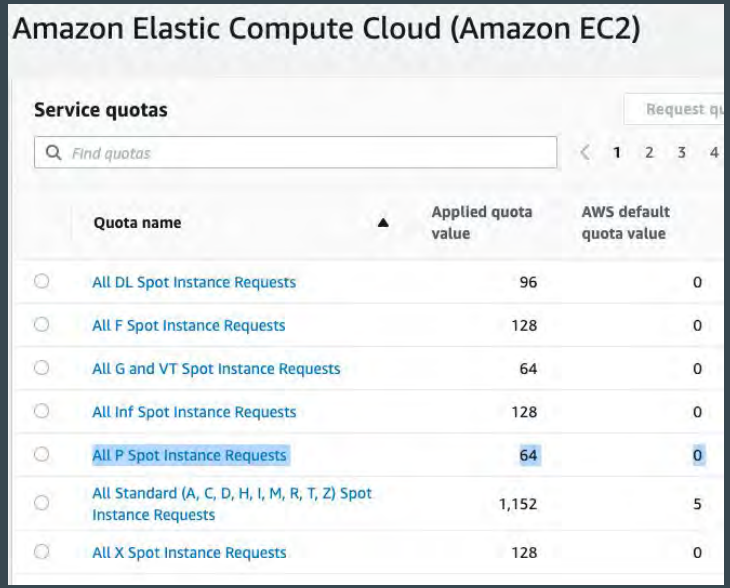
{Mini} Trends 2022

1-Year Recap

Cloud GPU Scarcity Driving Real Change

One of the more annoying BioIT infra challenges of 2021-2022

- GPUs have been scarce for a long time, but ...
- For the first time ever, I had a client with a monthly AWS spend of ~$100K/month on EC2 who was denied a quota increase for a single additional GPU node in US-East-1
- At a Boston-area client, we've moved TWO AWS Parallelcluster HPC grids from US-East-1 to US-West-2 simply to chase more GPUs within our quota limit
- A Boston-based biotech is building their first AWS footprint in US-East-2 because they don't trust US-East-1 to have the GPUs they need for CompChem and ML workloads



Defensive Hedge: Premise/Colo GPU systems

I think we will see more of this in 2022 and beyond. This is practical “Hybrid Cloud”

- This is a Boston-area client 1U CompChem/MDSim box with:
  - 1.5 Terabytes of RAM
  - Dual-Socket / 40 CPU cores without HyperThreading
  - Four (4x) Nvidia Tesla V100 GPUs
  - Ballpark cost: $70K
- Heck of a “fat node” for GPU, SMP and large-memory workloads, all in a 1U form factor
- Great building block to soak up 24x7 workloads cost effectively relative to cloud GPU or cloud large-memory instance pricing
- Racked and connected inside New England’s best connected colo space: https://www.markleygroup.com/ -- ideal for hybrid cloud setup

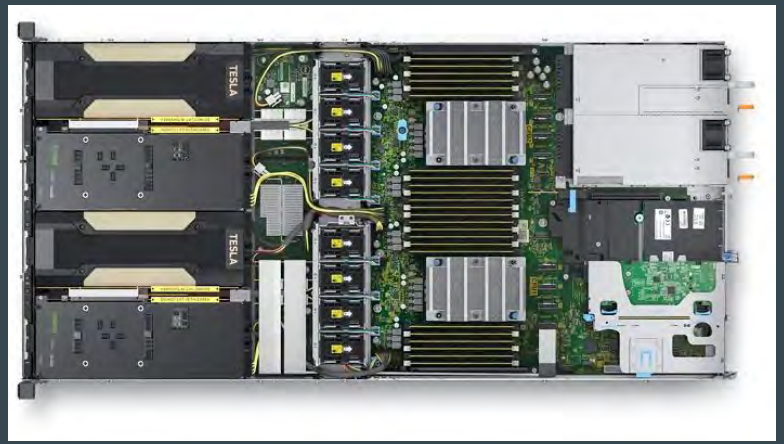
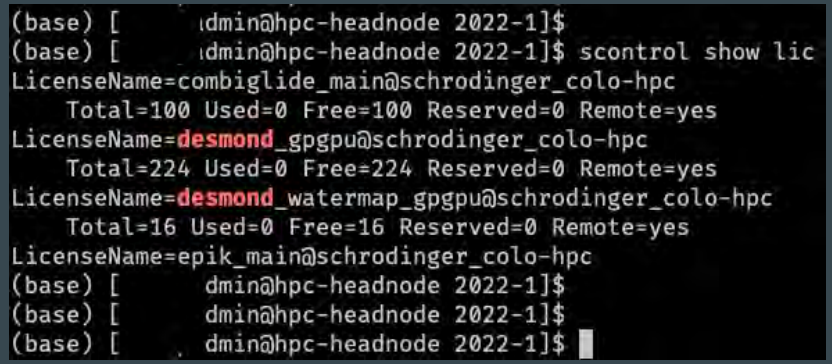


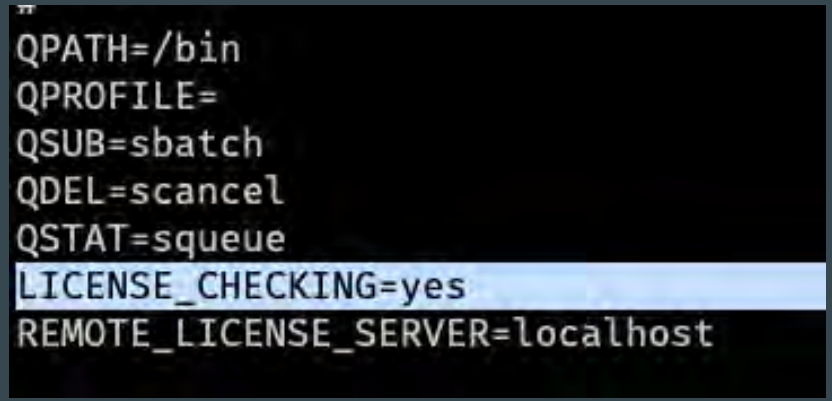
Image source:
https://www.dell.com/en-us/work/shop/productdetailstxn/poweredge-c4140

Biggest Personal Infra/HPC Trend 2022

AWS Parallelcluster + Schrödinger + SLURM “License aware job scheduling”

- This has become “my gig” @ BioTeam. Currently doing this for 5 different clients. Just did this with the ‘early access’ Schrödinger jobServer stack that will replace jobcontrol framework
- Schrödinger computational chemistry suite is popular, capable and very much Not Cheap -- Have clients with 5, 6 & 7-figure Schrödinger investments. Maxing ROI on those investments is critical.
- AWS Parallelcluster 3.x w/ SLURM scheduler allows for flexible compute fleets and deep integration with remote FlexLM license servers and the native Schrödinger jobControl() and jobServer() stacks.
    - End result? Auto-scaling HPC grid as backend for Schrödinger Computational Chemistry workloads, complete with “license aware” job handling by SLURM
- This is not the only way to run Schrödinger of course but we’ve been asked to do it often enough that it is an absolute trend





Mini Trend

Zero-trust, SD-WAN, and SASE networking becoming more real

- Starting to see real world SASE deployments across Orgs of varying sizes, no longer just a buzzword
    - One of our most interesting (and insanely fast growing) clients is using www.catonetworks.com to stitch together multiple clouds, office locations and mobile workers into a seamless cohesive SD-WAN network. It's pretty slick.
    - Also seeing perimeter81.com but only been hands-on with their "VPN alternative"

"SASE capabilities are delivered as a service based upon the identity of the entity, real-time context, enterprise security/compliance policies and continuous assessment of risk/trust throughout the sessions. Identities of entities can be associated with people, groups of people (branch offices), devices, applications, services, IoT systems or edge computing locations."

-- Gartner via https://www.paloaltonetworks.com/cyberpedia/what-is-sase

Mini Trend Of Some Concern

Startups going all-in on “zero infra” with Egnyte

- Egnyte file-share/file-sync platform is pretty popular at Orgs large and small–it’s great for a variety of use-cases, requirements and access patterns ...
- But for the first time I’m seeing startups exit stealth-mode or incubator space who have decided to go all-in on Egnyte as a company-wide storage solution for all scientific data and office/business files:

  “Whee! The only actual IT hardware we have on-premise are the WiFi APs. Cloud-only baby!”

Mini Trend Of Some Concern

Startups going all-in on “zero storage infra onsite” with Egnyte

This is concerning for a few reasons

- Most of these companies are entirely supported by MSPs because they are not yet ready to hire IT staff or fill IT Director / CTO type roles. Would companies with dedicated IT staff and deeper life science domain expertise be making the same decisions?

- Egnyte website & marketing is very compelling, however under the hood:
  - Max file size is 100GB; 50K files per folder max
  - No support for Linux. At all. Best you can do is expose smb:// share to Linux clients

- **Lack of support for Linux and files larger than 100GB is a dealbreaker for a huge swath of the Discovery/Research ecosystem**

  - Leadership at multiple companies growing out of this stage are talking candidly to us about coping strategies or how to sensibly (and selectively) integrate some Egnyte data with petabyte-scale scientific data resources and Linux-heavy infra on-prem and in the cloud

  - I have two active projects where I have to figure out how to expose Egnyte data to AWS Parallelcluster HPC grids running Linux

Figuring this out is going to be interesting. Maybe I can do a 2023 BioIT talk on best practices / lessons-learned.

Emerging Trend?

Data Science PaaS vendors starting to look attractive as shared, supported centralized standard

- Data Science interest and need has diffused org-wide
- Many people, many use cases across many different departments and groups
- We are seeing:
  - Rapid growth in the number and nature of people who want to run R-Studio, Jupyter Notebooks or Shared JupyterHub servers
  - Some are doing this on GPU instances without much cost-control
  - Many are DIY'ing it via Docker containers or self-managed single servers
    - IT starting to see the support, operations, cost & security impact of this

- Data Science Platform as a Service–huge energy and innovation in this space right now
  - Feels like the "cloud spend optimization" market in that many vendors are still trying to figure out the best revenue and pricing model so there is huge diversity in what they specialize in, what you can buy, how you deploy it and how it is paid for.
    Many options to eval and choose from!

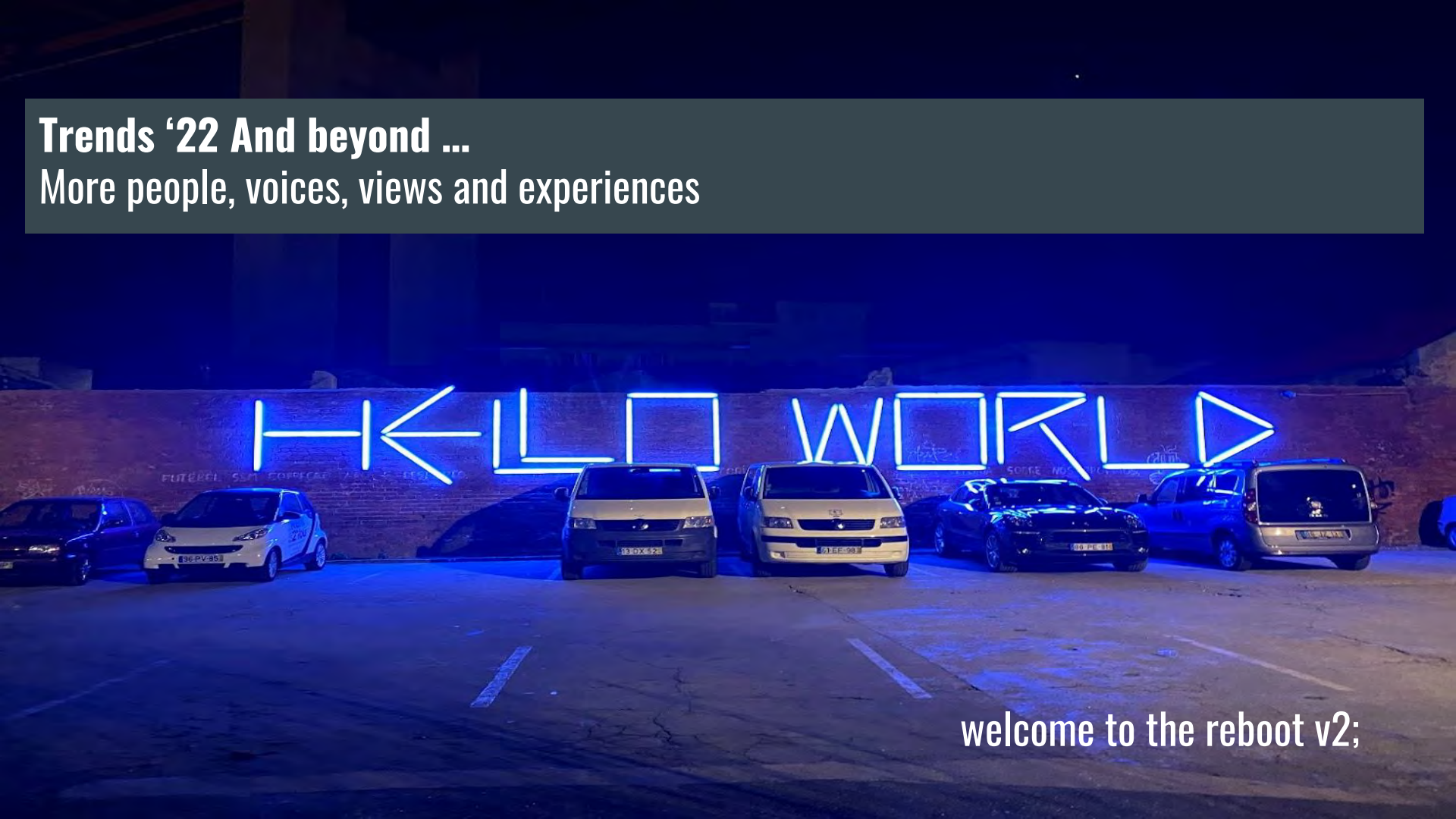
  - My personal and likely highly biased current preference:
    - saturncloud.io largely because of deployment model (in your own VPC) and pricing model (AWS marketplace).
      They make money based on AWS resources consumed so they are highly motivated to provide high-quality / high-touch support to both IT and end-users. I like this model.

{Mini} Trends 2022

end;

Trends ‘22 And beyond ...

More people, voices, views and experiences



welcome to the reboot v2;

Trends '22 and Beyond...

More people, voices, views and experiences

- Been fiddling with format of this session for three years now
- Trying to expand diversity of topics, voices, expertise and knowledge

- We think we've found a viable framework:
  - 3-4 speakers + moderator
  - Speakers own/control their own content, but ...
  - ... a common thread ties all the topics together

Common Thread ‘22

- After 10+ years of shouting “*storing data is easy; managing data is the hard problem …*”
- 20+ years of BioTeam consulting and we have never seen such interest (and active, funded efforts …) aiming straight at the “data insights”, “data access”, “data governance”, and “data management” realms

Common Thread ‘22

“... people tend to back into data from an IT or storage perspective without aligning to culture or a larger strategy ”

-- a ‘Data Commoner’

Why now?

- Downstream effects from many years of poor scientific and IT leadership are now too large to hide or obfuscate
    - Unchecked, unrestricted storage expansion as a means of avoiding difficult conversations and data ownership is no longer viable
- War stories and gossip about efforts that failed due to focus on flashy tech instead of “unexciting” first principles like ontologies and data dictionaries
- Data Science has proven effective and the techniques, tooling and skills have started to permeate org-wide
- Orgs that are effective at exploiting the vast array of diverse data across the Enterprise are out-competing and out-innovating their competition
- C-Suite and Investor Relations need more grist for their “*brah we are serious about ML and AI*” investor briefing materials before the next JPMorgan event

More specifically ...

- Notable since ~2020 but rate/interest increasing fast
- Multiple projects, partners and clients getting serious about:
    - "*humans browsing files in folder structures"* is no longer dominant data access use case
    - Hiring for "Data Product Manager" roles
    - Investing in "data insights" tooling & capabilities
    - Resourcing and hiring humans to work on Governance, Ontologies, Data Dictionaries & FAIR
    - Addressing isolated data "silos" across the organization
    - Building common data platforms that span:
      Research -> Preclinical -> Clinical -> Commercial Ops

Thank You For Being Here!

- Adam Kraut
  Data Product Management, DataOps and MLOps
- Michelle Bayly
  DataOS, Data platforms and Organizational Considerations
- Anna Sowa
  Strategy and Vision for Data and AI

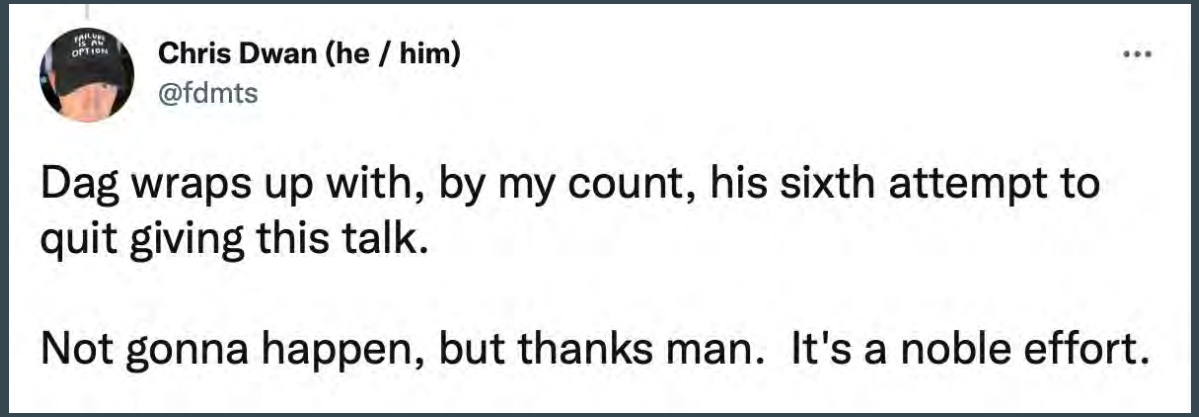
Image source: @datachick

Adam Kraut

Dag Retiring?!? Is it true?

First, let's all acknowledge that Chris announced his retirement from

Trends...Do we really believe it? Or is he pulling a Tom Brady on us?



Leadership Trends: Digitization and Digital Transformation

- Digital Transformation is C-suite-speak for Big Data, Cloud, IoT, and AI...
  - Data Strategy, Org dynamics, and Ops rolled into an ecosystem: BigDataOps, CloudDevOps, MLOps, AIOps
- Digital Transformation is the DevOps mindset applied to your digitized organization
  - Technology, Process, and PEOPLE - Changing the way we work together
- 70% of Digital Transformation Fails–GOOD!
  - As a society, we are too obsessed with winning (“I do is WIN WIN WIN no matter what!”)
  - Accept that failure is inevitable when you are doing something hard, Take the L, and learn from it
  - No transformation actually fails, continuous improvement is the goal, grow by 1%, a never ending journey
  - Building healthier habits and increasing the overall fitness of your Organization
- Digital Transformation requires a high EQ - guiding people through change is hard

Leadership Trends: Digital Drives Business Outcomes

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

- How can we have Data Science and AI embedded in Product Development?
  - Influence Company Deliverables (Productization and Services)
  - Applied across the entire value chain

- Performing Data Science (Analysis, Modeling, Exploration) requires solid fundamentals
  - Algorithms and model development
  - Computing Systems and Infrastructure
  - Applying Human Expertise and using those Skills correctly

- 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 Leads to a Healthier 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:

    - Discoverability–Findability and Exploration
    - Integrity at the Origin–Generating ML-ready or analysis ready data
    - Citizenship and Stewardship–Fostering a sense of community and ownership
    - Common Languages and Standardization–Controlling the Chaos
    - Automation as leverage–Infrastructure as Code, Fast Feedback Loops
    - Experiment tracking, versioning, and shared Workspaces
    - Continuous Delivery mindset (DataOps and MLOps)

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 standards and technologies with the broadest range across your tools and platforms
    - Python, Go, or Javascript

Agility as a differentiating capability

- Research Informatics has benefitted from advancements in software development and Infrastructure as Code

- Puppet > Chef > Ansible > Terraform > CloudFormation > CDK > Helm > ...
- Managing everything as source code
- Pipelines as Code: Nextflow, Airflow, WDL/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-style 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
- In this case, automation is fundamental
- 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
  - AI is an immature misunderstood teenager
  - Avoid biases in data and bad mental models
- May be critical to understanding the fundamental biology and chemistry

From DevOps to MLOps: Applying the Principles to AI

Where will you encounter MLOps?

- Dataset versioning, Model development, Training execution, Continuous Deployment, and Integration
- Data quality control and harmonization–Identifying bias and ethical consequences
- Software development, algorithm development, backend and frontend
- Abstracted Infrastructure

Leadership Trends: Shifting to a Data Product Mindset

- Pharma is shifting to a Data Product mindset (GSK, Moderna)
  - “What’s the difference between Recursion and a traditional Pharma? Traditional Pharma view data as exhaust. We view it as fuel” - Mason Victors, Recursion Pharma
- Product Mindset builds a sense of Ownership and Discipline
- Data as a Product, Capability as a Product, Security/Compliance as a product
- Not all products are valuable and not for everyone
- Understand the producer and consumer dynamics
- Incentivize, organize, prioritize, and execute
- Ownership is the space you will create for future leaders to grow into their power
  - Empowering others through advocacy and ownership is how we get the next generation leaders

AI has arrived whether or not we are skeptical of the hype

- AlphaGo, AlphaFold, Halicin from MIT
- Digital Twins
- Powerful narrow AI (real) versus Artificial General Intelligence (hype)
- Augmented Human Intelligence
- Preparing for AI companions in the lab and in our lives
- Future generations will be living and working side-by-side with some form of AI
- Are we ready for that? Have we discovered all of the consequences?

Core Competencies of Change–Bias Toward Action

- Align data to your organizational strategy
- All companies will become tech companies and will need to find ways to develop their AI/ML in an agile way–building, deploying, and running AI/ML models at scale will become a matter of survival for most organizations
- It will require new roles and new ways of running projects
- New ways of operating data commons and data platforms
- If you move fast, you can try more things
- And if you try more things, you're likely to find something that works for you

Dichotomy of Leadership in Digital Transformations

- Balancing Yourself–You versus You
  - A leader and a follower
  - Plan, but don't Over-plan
  - Humble, not Passive
  - Focused, but Detached (keeping your head up)
- Balancing People
  - Own it all, but Empower others
  - Resolute, but Not Overbearing
  - When to Mentor, When to Fire
- Balancing the Mission
  - Train Hard, but Train Smart
  - Aggressive, Not Reckless
  - Disciplined, Not Rigid
  - Hold People Accountable, But Don't Hold Their Hands

- *"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

Transformation requires true diversity of Leadership

- Balancing masculine and feminine
- Less masculine domination and more understanding
- We need radical compassion and nurturing forces
- Men in service to feminine power
- Strong big hearted men in service of strong powerful women
- Not survival of the fittest
- Survival of the nurtured

Michelle Bayly

About Me

I was in the room where it happens... and here's what I've learned.

Data Platforms - more science, lower costs

The Dream:

- A place where researchers, clinicians, and leadership access clean FAIR data
- Fresh, quality data that can be fed into other systems and used for ML/AI discoveries
- Easy to use and easy to maintain
- Low cost

What are the use cases?

Researcher

- Searchability
- Access to data
- Skills to access data

System Administrator

- Documentation and easy maintenance
- Control access to PHI/PII
- Easy to update CDEs

Institutional Leader

- Control access to PHI/PII
- Data curation that can be used to drive business decisions
- Need to demonstrate compliance with federal regulations and guidelines

Be honest about where you are starting from

- Is the **infrastructure** there?
- Is your **data management** a hot mess?
- What are the **phases** of this initiative?
- **Beta** is okay.
- Plan for **change management/conflict resolution.**
- Don’t forget about the **users!**

Building Data Platforms–Before you Build

Consider:

- Expense
- Authz capabilities (e.g. easily limiting access to PHI/PII)
- The interface (is programmer-friendly, but not user-friendly)
- Often these are complex distributed system, with no “simple implementation” for simple use cases
- Requires developers experienced with cloud computing to stand up and administer (e.g. edit user yaml file and run CI/CD to add users)

Building Data Platforms–wait, there's more

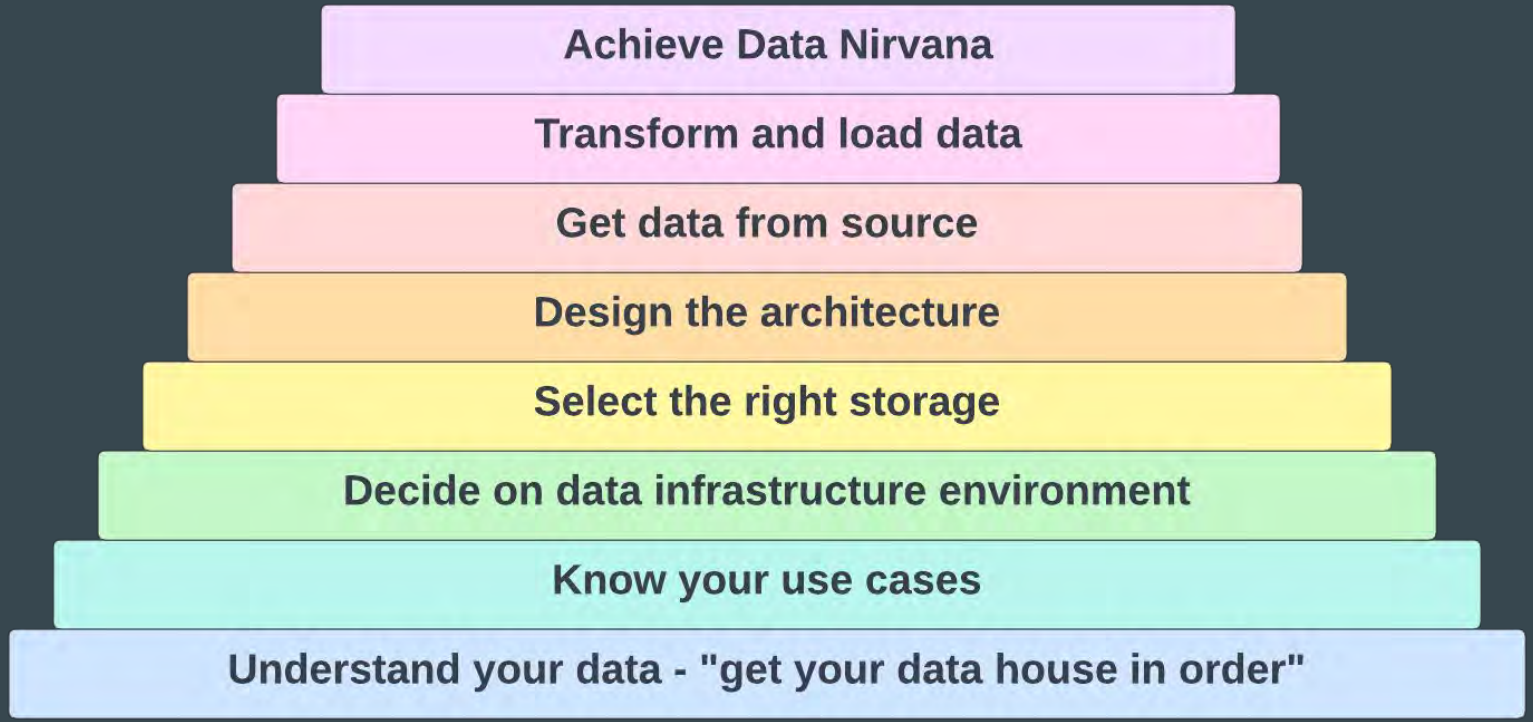
Consider:

- Reliant on the cloud
- Cloud vendor specific (i.e. works best on AWS)
- Sandbox capabilities for local development and experimentation
- Changing code/configuration requires CI/CD
- Data storage may not be flexible
- Database type (may not use a true Graph DB)

Ask before you build

- Cost
  - What is the baseline expense to run this infrastructure?
  - How does it scale with the size and complexity of data?
- Administrator and User UI/UX
  - Is the interface programmer-friendly but not user-friendly?
  - How does an administrator add users and manage permissions?
  - Can non-technical users and administrators use this system?
- Security
  - How will we handle AuthN AuthZ?
  - Can we security handle PHI/PII?
- Development and infrastructure
  - How much technical expertise will standing up and maintaining the system require?
  - Is the system cloud vendor agnostic and on-prem compatible?

Building Data Platforms



Who do you need? Where to they come from?

- **Data scientist/managers**- they have to have domain knowledge
- They translate the science problem into a data science problem
- They should:
  - Lead data standardization efforts
  - Check data suitability
  - Conduct exploratory data analysis
- You will need more and more of these people- lots of these people
- How do we develop these positions?
- “Data product manager” - data is a product

What? We need standards?

- This is the hardest and most important task
- It takes a village
- Standards are iterative

"I have not failed. I've just found ten thousand ways that won't work." - Thomas Edison

Setting up Architecture

- This is the second hardest task
- Who is authorized and who decides if they are are just as important as authn/z
- Data scientists/managers are critical

Lessons Learned–Stay People Focused

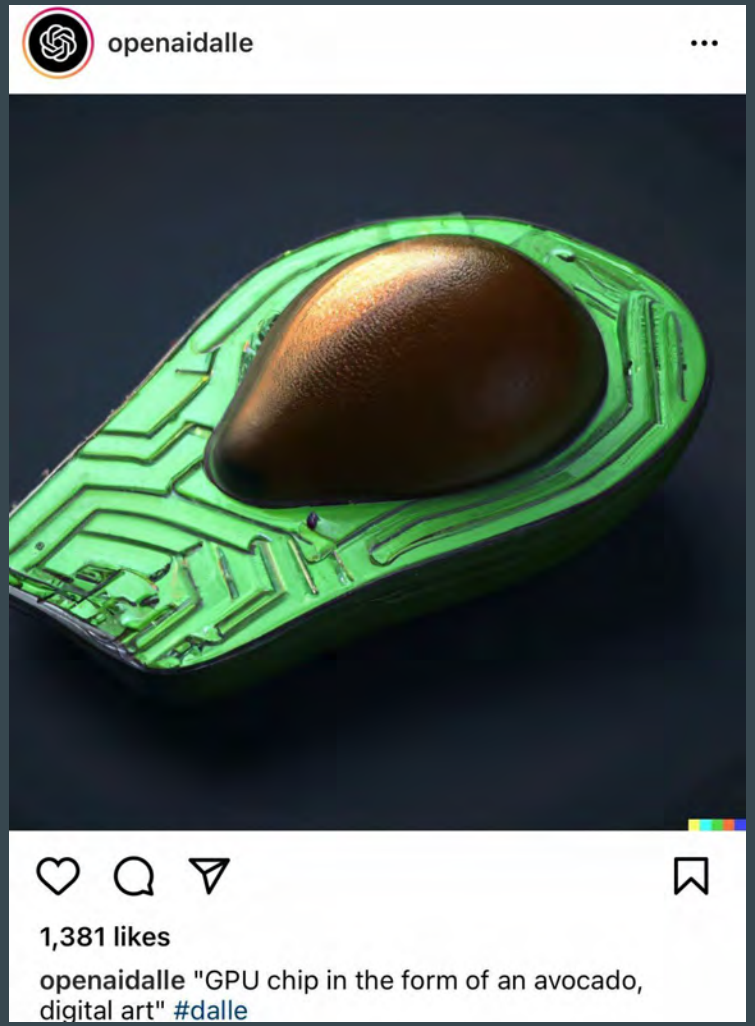
- Software CANNOT replace the need for human data management
- Before you, during and after stay focused on your user
  - If the system is not user friendly or the user does not have enough support, they will go back to using excel.
- Eat the elephant a bite at a time: ID the 4-5 biggest challenges and focus a solution for those first
- Eye on the future: is what you are doing scalable and flexible?

Anna Sowa

Who am I and why talk about strategy?

We are at a time and place where there is a chronic lack of equilibrium between opposing forces.

At its core, that's what strategy is: generating momentum out of polarity.


openaidalle
1,381 likes
openaidalle "GPU chip in the form of an avocado, digital art" #dalle


openaidalle
1,077 likes
openaidalle “Emotional baggage” #dalle

AI is both underrated and overrated

- Fernanda told us last year that most of us are not ready for AI, and it's really in research. She's right (obviously) but there's another side to consider.
- We have to think about the role we actually want to give AI in our biomedical and healthcare systems.
- Is AI curing CANCER? No, but is it doing amazing things? Yes–some of which you can look into with researchers like Matt Might.
- AI is not magic but it is pretty cool and very much appreciated a lot of times. Image search and processing, **do it!** NLP and speech and text mining–**excellent!** AlphaFold–**five stars!**
- We are not taking advantage of the "easy wins" while we overinvest in the unrealistic possibilities.

AI will require more and less human at the same time

- Is AI a tool or social construct? We socially create these tools and transfer our biases on them so we can't ignore the social aspect of it.
- **Less human:** Let the machines do the things that machines are good at–and more of it–we don't need scientists drag and dropping files. There is room for a lot more automation in our scientific workflows. We need more machine readable data.
- **More human:** let the humans own the **mindset**
- AI is going to be a mirror for our biases and ethics. The first question is whose ethics are "our ethics". Ethics of Corporations? Biomedical community? Governments? Industry can be on both sides of this–both develop it and then lobby against it.
- **We have to accept that ethical AI is beneficial to all of us.**

The importance of team dynamics vs. AI Cowboys

- You need AI talent, but searching for someone to save you should be avoided
- I hope if all 4 of us have said this, maybe it will really land, start with internal folks who care about data and processes
- Let your teams form and start working together–it takes time to build a common language -“Data commons is fundamentally a language building exercise”
- Give people time to stretch into the positions–but give them enough support to be effective
- 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

Organizations can be in two dystopias: data hoarding and data fearing at the same time

- General advice “out there” (which I have also given before): Before organizations consider digital transformation they should develop a comprehensive data awareness to ensure transformation is successful and sustainable.
- Data as an asset.
- Data is energy and should be treated and directed as such.
- It needs to be harnessed, not quantified, not feared.
- Direct your data to flow instead of rot.
- If you are trying to manage data traditionally, you are already lost.
    - 25PB of fast GPFS, 16PB is just sitting there rotting

XAI: How are we expecting AI to earn our trust?

- On a very serious note: it is going to cost you a lot if you don't have a source of truth with your data. You have to trust the data that's feeding your algorithm in order to trust and explain your algorithm.
- Opacity is especially problematic in high stakes settings such as clinical care.
- What does it mean to trust an algorithm? Is it any different than trusting and understand ourselves? It's different than our mathematical understanding of trust.
- Trust is a social construct. Trust is just risk disguised as a promise and it's the key to successful AI.
- The differences in what it means to understand and trust AI are not only based on degrees of technical expertise, but also domain-specific norms of decision-making.
- We have to be realistic about the trust we expect.
- Explainability won't save AI.

Your AI question should be meaningful but not a crystal ball

- Your AI question has to be meaningful enough to justify a large investment. With a scope that’s too big, your model will be undone by the complexity.
- Narrow questions that can give us superhuman results versus general questions that put us on the cycle of disappointment.
- Ask meaningful questions, but don’t ask the model to be a crystal ball AND don't just make it a compute engine.
- Find joy in the achievable AND don’t use that as an excuse to do nothing.
- Allow AI to explore the boundaries between what we know and what we don’t know.

Be aware of AI's limitations: "it is my nature"

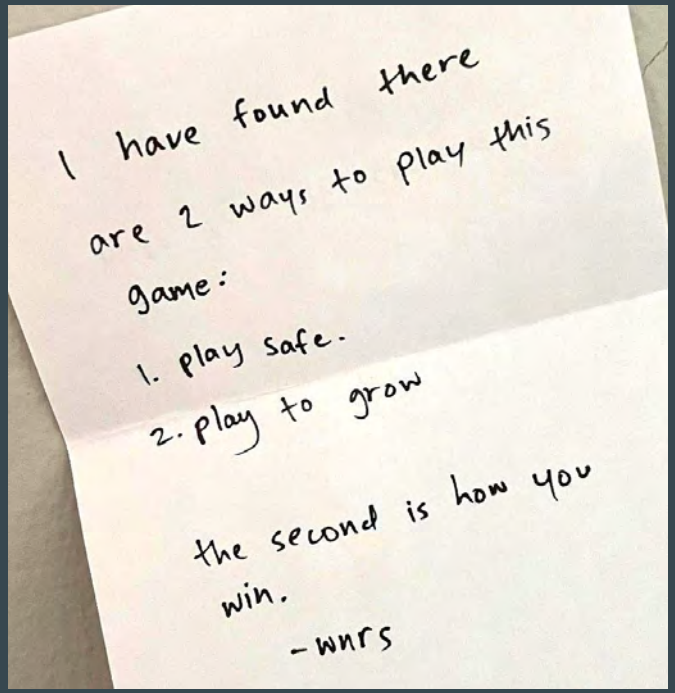
"A scorpion asks a frog to carry him over a river. The frog is afraid of being stung, but the scorpion argues that if it did so, both would sink and the scorpion would drown. The frog then agrees, but midway across the river the scorpion does indeed sting the frog, dooming them both. When asked why, the scorpion shrugs and says "I could not help myself. It is my nature --Fable of the Scorpion and the Frog

- Sample does not equal representation
- Data does not equal reality
- Correlation is not causation
- Discrimination does not equal personalization
- Past does not equal future
- Meaning does not equal meaningful

Leading AI transformation: own it but in your own way

- The CEO has to drive this transformation–set the strategy and turn the whole ship while most likely not understanding what the final outcome is going to look like.
- There are so many layers underneath that.
- It's about tapping into a different way of leadership AND not putting it off–because anyone who says we will deal with data and AI later is in for a rude awakening.
- Follow sound advice AND release the need to run your business like others.
- Don't confuse tactics and strategy.

Preparing for AI/ML is going to be messy and will change you in new ways



- Balance surrender and alignment
- I could tell you a lot about vision and how much it matters.
- I could tell you that AI will expose all of the places where your company is not truly aligned.
- That may or may not land with you.
- The truth is that change will happen–it just depends on if you are in control of driving it or being driven by it.

Being disoriented does not equal being lost

- The new era is emerging out of a small niche. Although it has big promises, it is still very experimental without stringent processes that are considered standard processes.
- Change is hard for people. Make it beneficial.
- Every transformation needs to start somewhere: break up the first silo.
- Infuse even more “human”: digital transformation is a long game and it’s hard. It’s how humans learn to supervise and integrate the data around them.

Moving fast and breaking things versus moving slow and dragging our feet

Consider one of these alternatives:

- Moving to achieve what you believe in
- Move fast and create possibility
- Move fast and learn something
- Move fast and take responsibility
- Move slow and make things

Some closing thoughts

- Be you on purpose.
- Chase the right thing.
- Your energy is the biggest currency you have.
- Play the long game.
- Your organization will follow.
- And then AI (or something even better) will naturally follow

end;

Discussion Time

BIOTEAM
Accelerate Science

Want these slides?
bioteam.net/present22/