STRIDES Initiative: Applications in NHLBI BioData Catalyst

January 20, 2023

Ashok Krishnamurthy and Matthew Satusky

renci

Acknowledgements

- **NHLBI:** Alastair Thomson, Sweta Ladwa, Noble Dzakpasu
- **Seven Bridges:** Tony Patelunas, Jack DiGiovanna
- **RENCI:** Paul Kerr, Julie Hayes, Asia Mieczkowska, Sarah Tyndall
- **RTI:** Kira Bradford

Agenda

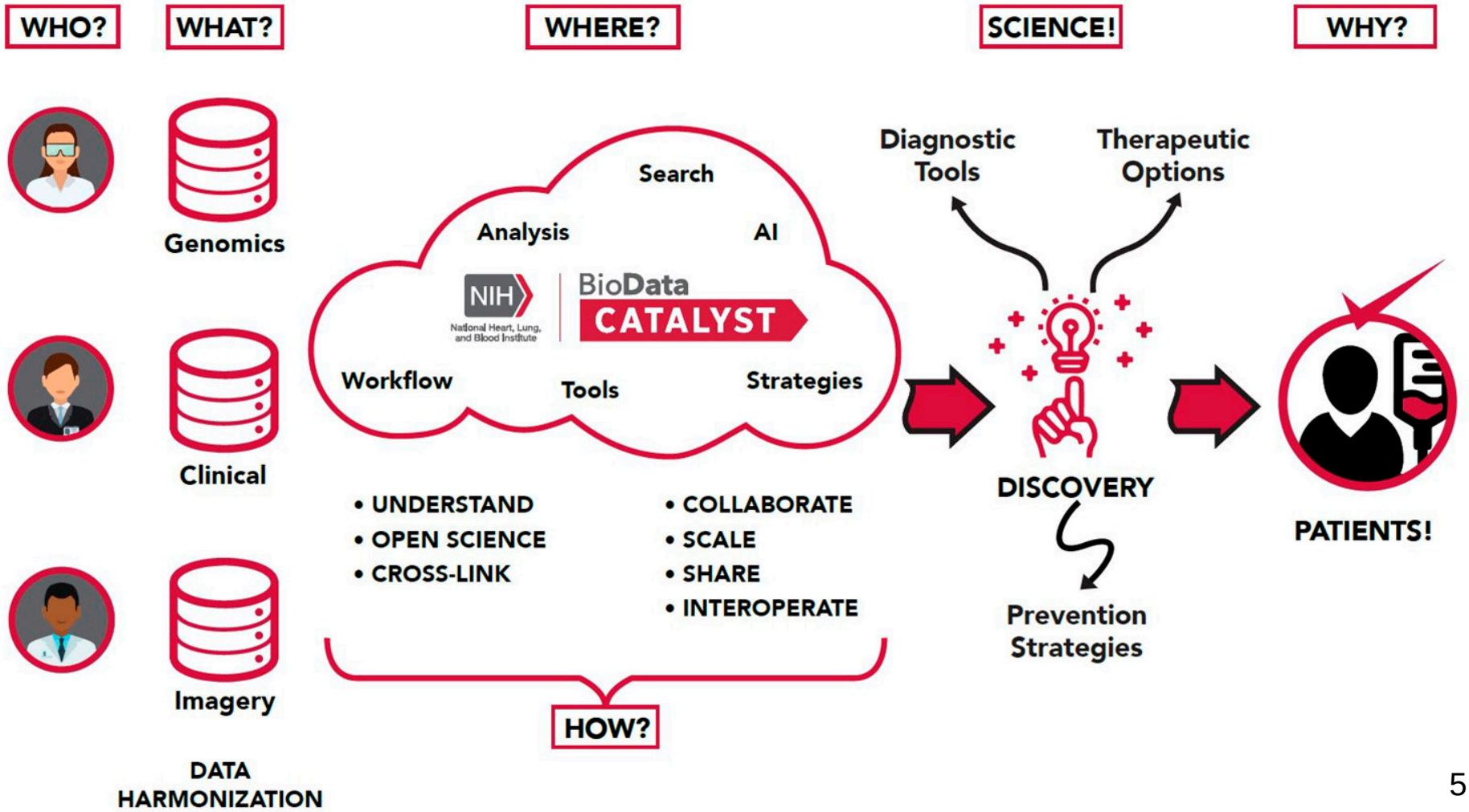
1. Background on BioData Catalyst
2. RENCI's experience: BioData Catalyst before and after STRIDES
3. Across BioData Catalyst: Applications and partnerships enabled by STRIDES
4. Back to RENCI's experience: BioData Catalyst Cloud Cost Modeling Project
5. Upcoming applications of STRIDES in BioData Catalyst
6. Closing thoughts

Background: NHLBI BioData Catalyst

Cloud-based ecosystem developed to advance Heart, Lung, Blood, and Sleep (HLBS) science with a focus on democratizing access to tools and data for researchers.

- Allows HLBS researchers to **find, search, access, share, store, crosslink, and compute** on large-scale data sets
- **Democratizes data and compute access** while ensuring appropriate data security
- **Accelerates efficient biomedical research** and maximizes community engagement, productivity, and discovery

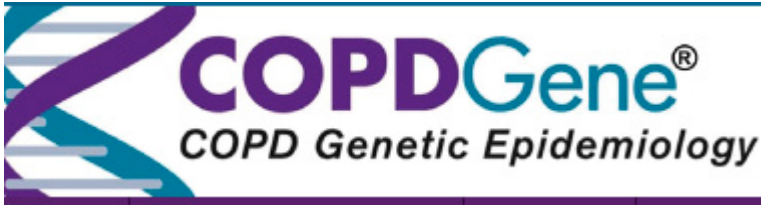
NIH National Heart, Lung, and Blood Institute | BioData CATALYST


WHO?
WHAT?
WHERE?
SCIENCE!
WHY?
Genomics
Clinical
Imagery
DATA HARMONIZATION
Search
Analysis
AI
NIH
National Heart, Lung, and Blood Institute
BioData CATALYST
Workflow
Tools
Strategies
• UNDERSTAND
• OPEN SCIENCE
• CROSS-LINK
• COLLABORATE
• SCALE
• SHARE
• INTEROPERATE
HOW?
Diagnostic Tools
Therapeutic Options
DISCOVERY
Prevention Strategies
PATIENTS!

RENCI's Experience: Before & After STRIDES

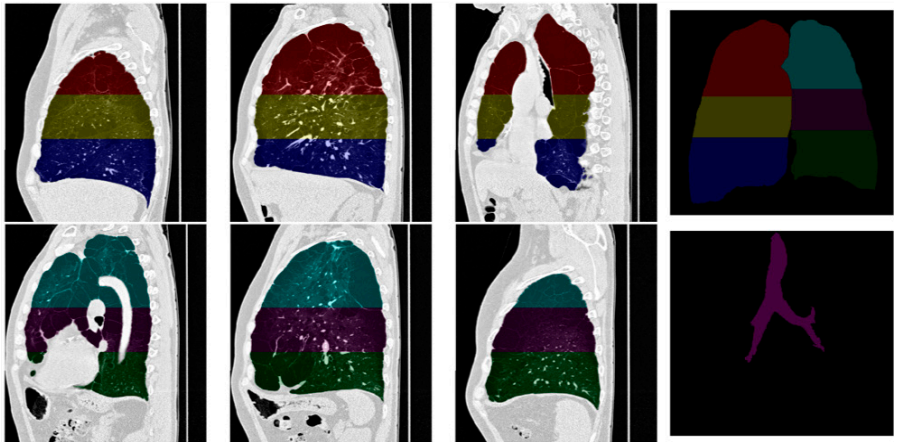
Before STRIDES

- Early on in BioData Catalyst program (formerly DataSTAGE): RENCI was involved in developing **imaging and deep learning solutions for COPDGene researchers**
- Set up our own computational workspace in AWS
- Included a budget for cloud costs
  - No discounts from AWS (UNC did not have a relationship with AWS)
  - Insufficient cost monitoring as a research team



After STRIDES

- Monitoring and reports from STRIDES that helped us **manage cloud costs significantly better**
- Helped with **security monitoring**
- **Two major early users:** Raul San Jose Estepar from Brigham and Women's Hospital and Yonghua Zhuang from University of Colorado (3+ publications)
- **Costs covered by NHLBI**

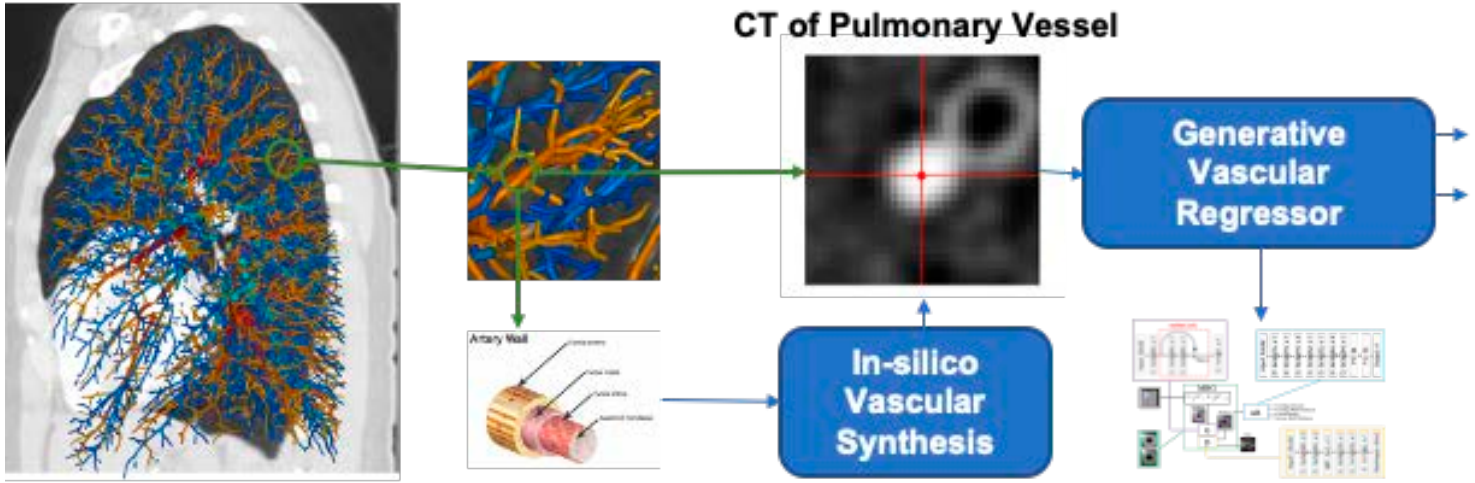


Quantification of Pulmonary Vascular Remodeling

Need: Understand vascular remodeling vs, pruning due to tobacco injury in COPD leading to pulmonary hypertension phenotype.

Goal: Develop novel generative deep learning techniques to quantify vascular muscle layer and lumen size of intraparenchymal vasculature

Progress: Preliminary results show that generative approach can estimate average vascular muscle layer



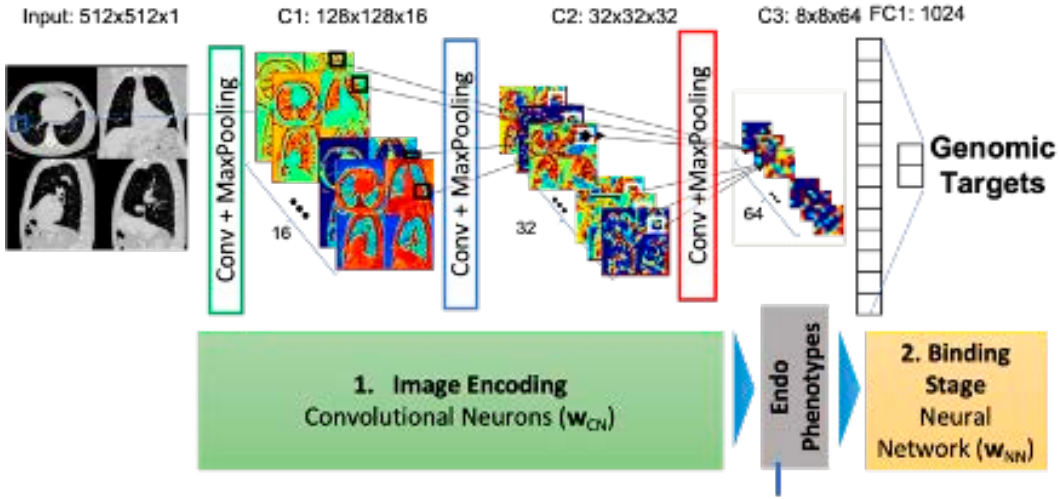
BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

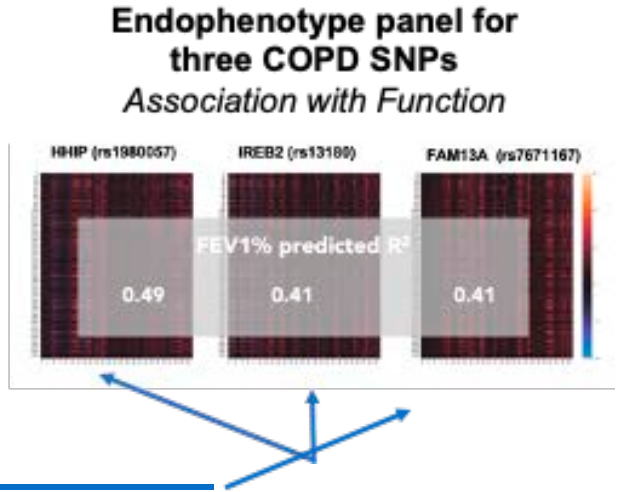
HARVARD MEDICAL SCHOOL

doi:10.1177/20458940211061284

Discovering Novel Endophenotype in COPD with DL

Need: Discovery new endophenotypes that relates to COPD endotypes
Goal: Discover novel relations that link the image phenotype to the genotype in COPD using deep learning and radiogenomic approach
Progress: New panel of endophenotypes discovered with DL has associations with functional decline.





BRIGHAM HEALTH
BRIGHAM AND WOMEN'S HOSPITAL

HARVARD MEDICAL SCHOOL

doi: 10.23866/BRNRev:2019-0014

STRIDES Across BioData Catalyst

renci

Applications of STRIDES in BDCatalyst

- BDCatalyst provides **two computational workspaces:** Seven Bridges (SB) and Broad Institute-Terra
- Investigators can apply for **$500 in cloud credits** on either platform
  - The goal is to enable researchers to get started on the platform quickly and assess utility
- In past year, **83** such requests have been approved, **72** for SB and **11** for Terra
- The goal is that once researchers are comfortable with BDCatalyst, they can **write STRIDES costs into their NIH application**

Benefits of STRIDES in BDCatalyst

- At the start, NHLBI had access to **technical expertise from cloud providers** to get up to speed on cloud technology, storage, compute capabilities of various platforms, etc.
- Significant **cost savings** across teams
- Would have been extremely costly to implement in-house solution at NHLBI (plus hardware and security considerations)
- Significant role of cloud providers in implementing key functions (**De-identification**, **FHIR, Imaging)**
- STRIDES **expedited BDCatalyst's launch**, especially since many teams already had established relationships with cloud providers
- Currently there are 54 active NHLBI STRIDES accounts (billed to awardee or NHLBI)

Seven Bridges (SB) and STRIDES

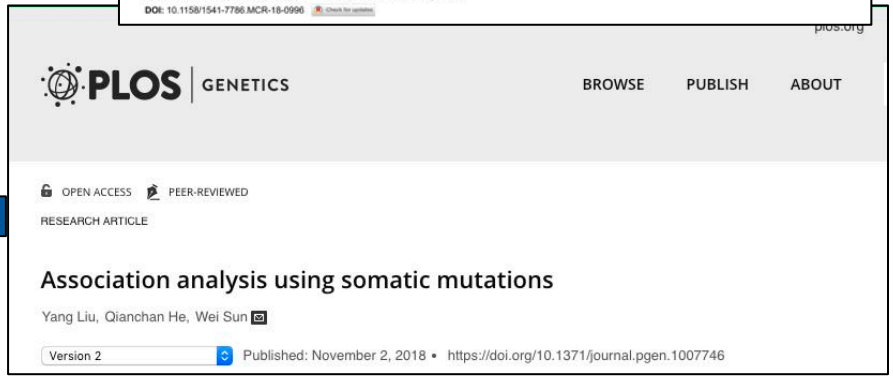
- SB processes make availing of the STRIDES discount completely transparent to the researcher – **researchers have no need to interact directly with STRIDES**
- As cloud costs get written into NIH grant submissions, **a possible suggested model**:
  - Institution sets up STRIDES account that holds STRIDES cloud cost funding
  - As NIH awards come in with STRIDES funding, these go into the Institution account
  - This will further simplify the researcher's administrative load of cloud cost accounting

SB Billing Groups Ease STRIDES spending

Funding agencies can set **appropriate standard funding** (e.g. $100-$500) for *any new users*. Ex. NHLBI Pilot Cloud Credits

Larger amounts (e.g. $5-$20k) can be created for *selected users* with an ambitious research plan

- HeartShare consortium aggregated pilot credits for all in group to use ($2,000)
- Westover sleep consortium aggregated pilot credits ($4,000)
- BDC Fellows program funded at various $1,000+ levels



renci

BDCatalyst-Cloud Provider Partnerships Enabled by STRIDES

- **Amazon Web Services (AWS):** early STRIDES partner
  - Imputation Server from University of Michigan
  - TOPMed data of 4 Pbytes was ingested into AWS
  - Data is now been duplicated on Google Cloud Platform (GCP) and Microsoft Azure
- **Google Cloud Platform (GCP):**
  - Imaging use leveraged the Google Health API for DICOM viewer
  - GCP helped Deloitte de-identify imaging data
- **Azure:**
  - Significant focus on imaging and AI currently
  - C4R imaging data ingestion and use on Azure
  - Planned use by the HeartShare program

Google Cloud

RENCI's Experience: Cloud Cost Modeling

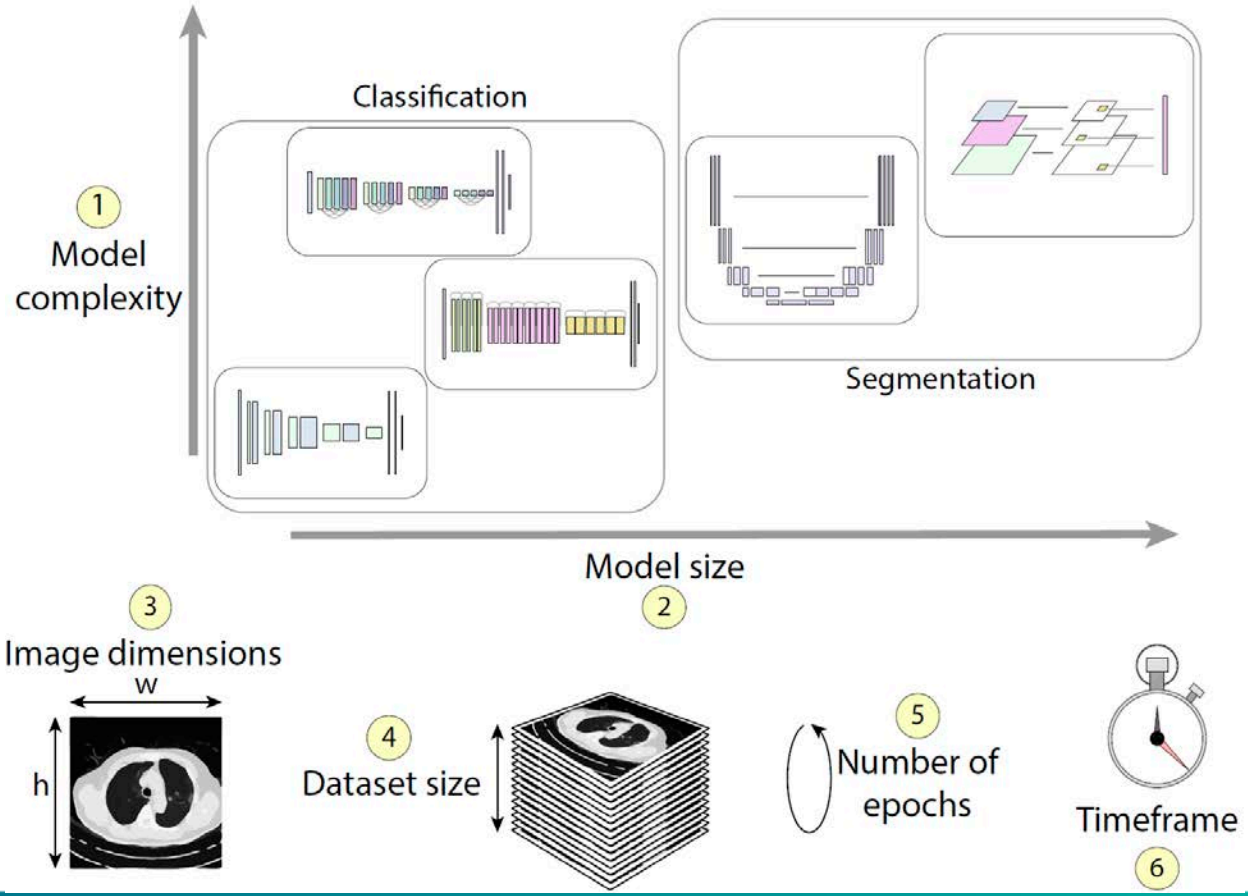
BDCatalyst Cloud Cost Modeling Project

- **Goal:** understand how researchers can model costs in the cloud to better utilize cloud platform resources (AWS and GCP) to more effectively conduct computationally intensive research
- Deloitte worked with NHLBI and BDCatalyst researchers to **identify and run common workflows in the cloud**, elucidating **pricing dynamics** of various cloud compute configurations
- Deloitte conducted **interviews with the principal investigators** to determine specific workflows to run and benchmark across AWS and GCP
- **Scope of work:** two genomic workflows and one **imaging workflow**

Google Cloud

Cloud costs for deep learning projects depend on many factors

Note: 1-6 correspond to factors that can affect cloud costs

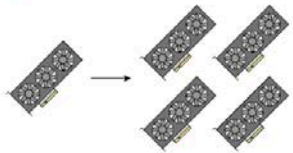


Cloud costs were modeled for broad applicability

Compute

NIH National Heart, Lung, and Blood Institute | BioData CATALYST Powered by Seven Bridges | aws

3 instance classes:
- g4dn
- p2
- p3



Single- and multi-GPU support

Data

CANCER IMAGING ARCHIVE

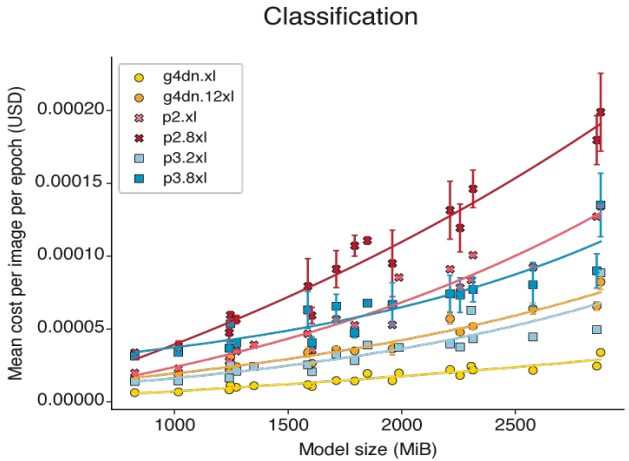
Publicly available lung CT image sets:

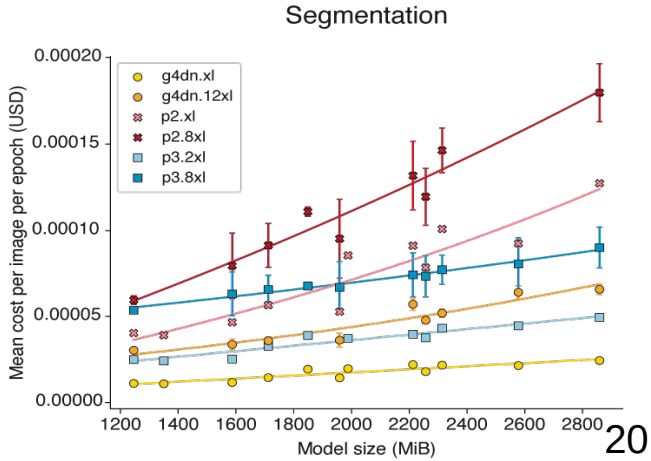
5k images ⟶ 500k images

Model

Classification	Segmentation
✖ brain ✓ lung ✖ heart	
8 model architectures from common families: ResNet DenseNet EfficientNet NASNet	6 UNet variations 6 FPN variations TensorFlow

455 benchmarks
4550 data points
Extrapolatable





Case 1: Research to develop a deep learning model to identify COPD in lung CT scans

Dataset: COPDGene

Subjects: 9,390 subjects with COPD status
390 reserved for validation
9,000 for training

Images: Lung CT stacks (512 x 512)
~500 images per subject
100 central images from each

Total training images: 900,000

Task: Classification

✓ No COPD
✖ Minor
✖ Moderate
✖ Severe

Model: ResNet152 (1793 MiB)
Epochs: 50

Calculated estimates:

	Instance	GPUs	Time	Cost/epoch	Total cost	STRIDES pricing
Least expensive	g4dn.xlarge	1	54 days, 12 hours, 40 min	$13.77	$688.37	
	g4dn.12xlarge	4	17 days, 12 hours, 59 min	$32.94	$1646.93	
	p2.xlarge	1	119 days, 7 hours, 37 min	$51.55	$2577.27	
	p2.8xlarge	4	24 days, 8 hours, 30 min	$84.17	$4208.43	
Recommended	p3.2xlarge	1	19 days, 4 hours, 0 min	$28.15	$1407.63	
Fastest	p3.8xlarge	4	8 days, 20 hours, 16 min	$51.96	$2598.22	

Case 2: Modifying the classification model for semantic segmentation

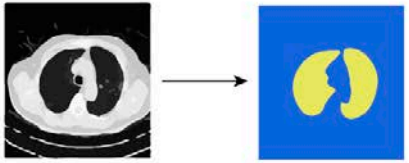
Dataset: COPDGene

Subjects: 9,390 subjects with COPD status
390 reserved for validation
9,000 for training

Images: Lung CT stacks (512 x 512)
~500 images per subject
100 central images from each

Total training images: 900,000

Task: Segmentation



Model: ResNet152-FPN (2859 MiB)
Epochs: 50

Calculated estimates:

	Instance	GPUs	Time	Cost/epoch	Total cost	STRIDES pricing
Least expensive	g4dn.xlarge	1	90 days, 7 hours, 49 min	$22.81	$1140.28	
	g4dn.12xlarge	4	32 days, 22 hours, 49 min	$61.87	$3093.68	
	p2.xlarge	1	257 days, 19 hours, 1 min	$111.37	$5568.32	
	p2.8xlarge	4	47 days, 0 hours, 8 min	$162.45	$8122.63	
Recommended	p3.2xlarge	1	30 days, 16 hours, 17 min	$45.06	$2253.07	
Fastest	p3.8xlarge	4	13 days, 13 hours, 57 min	$79.80	$3989.75	

Future Applications of STRIDES in BDCatalyst

- Upcoming work on BDCatalyst **FHIR EHR** ingestion pipeline
  - Plan to leverage **Azure's extensive experience** with FHIR on Azure
- Upcoming work on **RECOVER BioData Catalyst Data Gateway (RBDG)** project
  - Again we hope to leverage STRIDES Azure AI/Imaging

Closing Thoughts: Big Picture

- Having a **FAIR data collection of health data in the cloud** can be a game changer
  - You are already seeing how AI is revolutionizing things with Large Language Models such as GPT3, ChatGPT, etc.
  - In Imaging, Large models such as DALL-E and Stable Diffusion are creating realistic images from prompts.
  - These leverage the **power of enormous data sets with enormous computational capability**
- Can the biomedical data in STRIDES be leveraged to produce such **Foundation AI models?**
  - Such models then gets specialized for different diseases, for example

Thank you!

Supplemental Slides

Genomics Workflows: Approach

- **Two strategically important genomic workflows were identified:** 1) Genome Wide Association Studies (GWAS), and 2) Structural Variant Callers (SVC)
  - These workflows were **executed on the Seven Bridges Genomic Research Platform**, currently in use within the BDCatalyst Environment
  - The workflows were then comparably engineered and **run on both AWS and GCP** to evaluate costs, assess the relative performance between the platforms, and determine if there were efficiencies that could be engineered into the workflows
  - After analyzing the results, Deloitte identified **best practices and suggested improvements** to the platform to optimize future performance

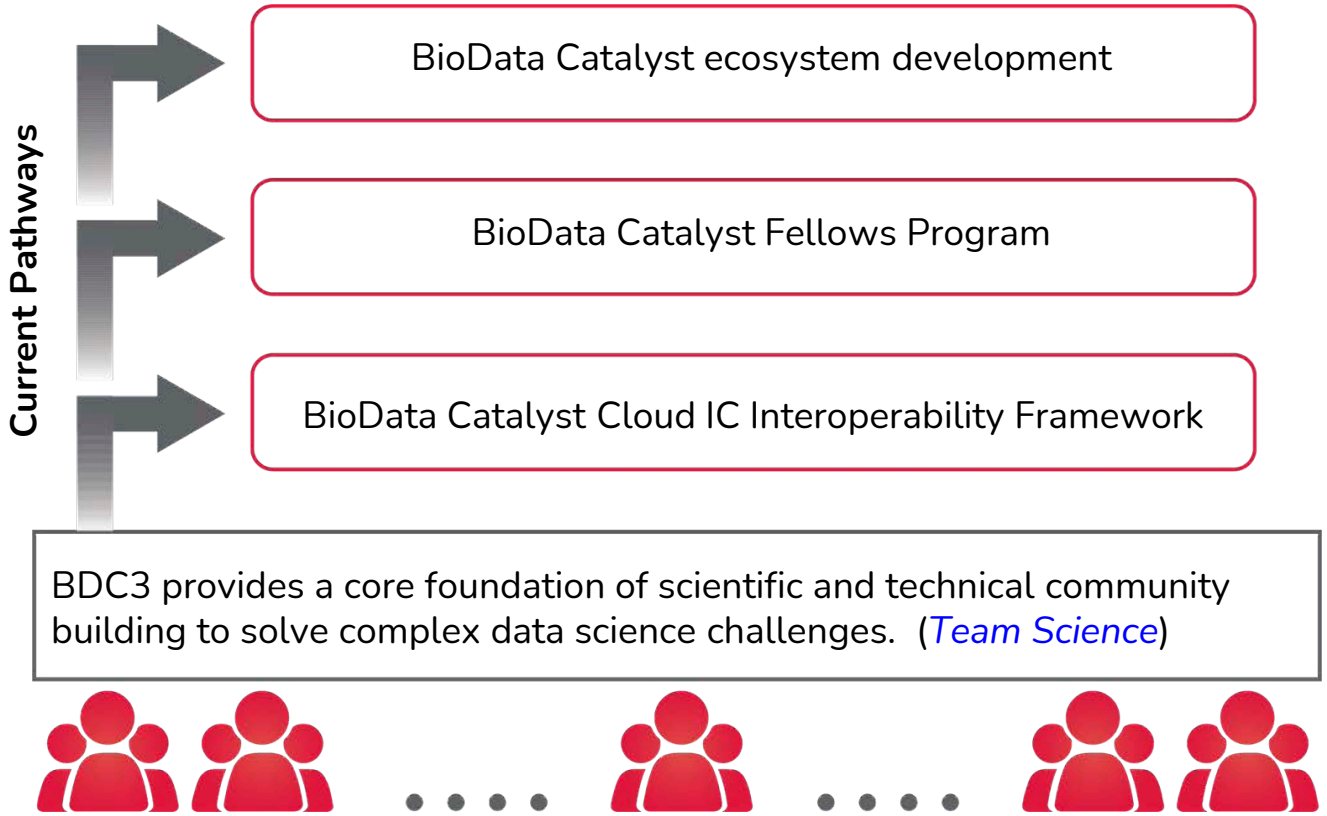
Genomics Workflows: Conclusions

1. **Processor generation** is one of the most important determinants of **price/performance** for large scale compute oriented genomic workflows
2. **Large-scale genomic workflows push the technological edge**, and incorporation of the **latest cloud technologies are critical** to perform these large-scale workflows efficiently and in a price performant way
3. Each genomic workflow **utilizes resources in a different manner** and requires **different compute configurations** to maximize performance/minimize cost
4. Before scaling up a workflow, it is important to **set up a testing protocol** to determine the proper compute configurations for each stage of the workflow, and be aware that these **configurations may change** as you scale the workflow up
5. Since each stage of a workflow can require different resources, **combining numerous stages into a pipeline may not be the optimal way** to structure a workflow
6. **Preemptible/Spot instances are cost-effective for shorter run workflows**, but on the Seven Bridges Platform, the v**alue diminishes the longer the workflow runs beyond 24 hours** and can actually be more expensive and take a longer period to run than on-demand instances
7. **Large-scale workflows may hit technological hurdles** (run out of disk space, RAM, etc.) and may need to be reconfigured or use alternative tools to complete at scale
8. **Support is critical** to being able to effectively perform the analysis/research

Genomics Workflows: Best Practices

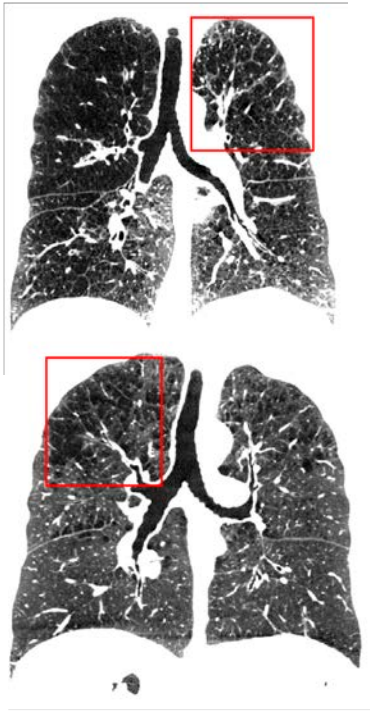
1. Always use the **most recent generation of processor** when running workflows
2. Determine **where your data is stored** (which cloud platform) and choose that platform for **all your work** (to minimize networking costs), unless you **don't have access to the latest generation of processors** on that platform. In this case, it may be better to **move your data,** as the increased networking cost may be less than the increased compute costs
3. Devise a **testing strategy before scaling workflows** to understand the compute requirements and what type of instances to configure for your workflow
    a. **Critical determinants are number of vCPUs, RAM, and disk space requirements**
    b. After running your testing protocol, understand that **as you scale, your workflow you may require different solutions** than your low volume testing indicated (more RAM/disk space/compute cores)
4. **Only combine processing steps that use resources in a similar way**
    a. Stages that require different resources should **not be combined** in pipeline stages that share resources
    b. On Seven Bridges, you can **specify the compute configuration** at the tool level (not workflow) to allow your pipeline to have different configurations at different stages of the pipeline
5. Use **Preemptible/Spot Instances for shorter running workflows** (less than 18-24 hours)
    a. For workflows that run for greater than 24-36 hours, Preemptible/Spot Instances did not consistently demonstrate cost savings and took longer periods of time to run
6. (Seven Bridges-specific) If you are using **Preemptible/Spot Instances** (or even if not), use **Memoization** so that any pre-computed results can be re-used in case of an interruption

BioData Catalyst Coordinating Center



Imaging Analysis: Approach

- Focus on **medical imaging** since the vast quantity, size, and growth of medical images is stressing current on-premise compute and storage resources
- Deloitte conducted interviews to determine relevant imaging workflows for benchmarking
- **Key functionality:** ability to share sensitive imaging files in a secure manner
- **Potential solution:** The Cloudtop Imaging infrastructure was built out on GCP, data imported from the Cancer Imaging Archive (TCIA) public healthcare dataset, and the open-source software installed on a Kubernetes Cluster.
    - Cost of operating was a very manageable **$285 per month** for full-time use by one rotating active user



Imaging Analysis: Key Takeaways

1. Data networking charges were **minimized** in this deployment of the application
    a. If data was imported from a different cloud platform or on-premise computing center, that could materially impact the initial deployment cost, although monthly charges would remain consistent with the costing work done in the study
2. Since **data storage costs** are a significant ongoing expense, reconfiguring the application to **directly access the original repository** would avoid data duplication and reduce costs
3. The Google cloud-native tools and innate VPC and security protocols are utilized and leveraged in this application and demonstrate **Google's effectiveness** in these types of data sharing applications
4. **Open-source tools** were effectively integrated and deployed to leverage Google's cloud native tools
5. While AWS does have third party DICOM imaging tools, the **lack of cloud-native functionality** has appeared to slow its adoption in the area of imaging analysis within the NHLBI environment