Utilizing the Seven Bridges Platform for Training of a New Generation of Health Data Scientists

Yuriy Gusev, PhD
Associate Professor of Bioinformatics,
Director of Health Informatics & Data Science Program
Innovation Center for Biomedical Informatics (ICBI),
Georgetown University
Profile: https://icbi.georgetown.edu/gusev/
Publications: orcid.org/0000-0001-7371-4715

Krithika Bhuvaneshwar
Research Instructor Faculty / Senior Bioinformatician,
Curriculum coordinator, Health Informatics & Data Science,
Innovation Center for Biomedical Informatics (ICBI),
Georgetown University
Profile: https://icbi.georgetown.edu/Bhuvaneshwar
Publications: bit.ly/krithikab
Outline:

Part I:
- Our background in bioinformatics and data analytics
- Our approach, teaching philosophy and methodologies for teaching and training of a new crop of biomedical data scientists

Part II:
- Examples of graduate-level courses that leverage the Cancer Genome Cloud as a teaching platform in the Masters in Health Informatics and Data Science Program at Georgetown University.
Part 1
Background: who we are

The Georgetown Innovation Center for Biomedical Informatics (ICBI) is a hub for cutting-edge research in biomedical informatics.
Background: who we are

• Innovation Center for Biomedical Informatics at Georgetown (since 2012)
  • https://icbi.georgetown.edu/

• Biomedical Informatics Shared Resource for Lombardi Comprehensive Cancer Center (for the last 10 years)
  https://lombardi.georgetown.edu/research/sharedresources/bbsr/

• Health Informatics and Data Science Program (MS HIDS) since 2019
  • https://healthinformatics.georgetown.edu/
Main areas of research activities @ ICBI

- ICBI is an academic hub for Biomedical Data Science with the primary mission of conducting investigator initiated research in biomedical data science & Informatics, education and training; and collaborative projects with biomedical investigators.

- ICBI is home to the Georgetown Data Base of Cancer (G-DOC) – open access on-line platform for Translational Cancer Research.

- New Directions: Data Science and Bioinformatics for Immuno-Oncology (IO);
- Big Genomic Data Analytics for Non-cancer applications:
  - Neuro Degenerative Diseases,
  - Transplant Research,
  - Viroinformatics, Microbiome & Metagenomics

- Growing number of collaborations in Immuno-Oncology:
  - Colorectal, Pancreatic and other GI cancers
  - Liver Cancer
  - Bladder Cancer
  - Breast Cancer and
  - Ovarian Cancers
  - Lung Cancer
  - Melanoma
  - Brain Cancer
Overarching Themes in all of our areas of research:

• Big Biomedical Data:
  How Big is it?

• Cloud computing:
  Can not avoid it – better embrace it

• Data Analytics:
  Mining the Big Biomedical Data
Big Biomedical Data: Genomics and Beyond

- Massive amount of Genomics data is being generated as we speak - Petabyte scale
- Big Genomics Data transform Biomedical Research and Clinical Practice
- Emerging Trend: Utilizing Massive Genomics Data for Precision Medicine/Precision Health
- Genomics Data Science, Bioinformatics and Computational Biomedicine – bridging the gap and connecting the dots
Big Data Mining – moving forward fast
Teaching and Training Philosophy

• These transformative times in Biomedical Research and Medicine require adequate workforce trained to deal with Big Data:

• Increasing Demand on interdisciplinary professionals who are proficient in both computational methods and biomedical domain knowledge

• Two possible scenario:
  • We can train biomedical researchers in computational methods
  • We can train STEM trained professionals to learn about biomedical technologies and data types

• The ideal new generation of data scientists could be fluent in both domains: computational and biomedical
A very brief history of curriculum development for Biomedical Data Scientists Training at ICBI

- 2010-12 – Training workshops utilizing our on-line platform G-DOC
  - gdoc.georgetown.edu
- 2012  New Graduate Course in Translational Bioinformatics utilizing on-line platform G-DOC to train bioinformatics graduate students
- 2014 New Curriculum development for undergraduate biology course in Systems Biology, utilizing on-line knowledge base Pathway Studio (Elsevier)
- 2016-17 EdX MOOC development: combining various on-line resources including G-DOC, TCGA, Xena, and others to provide students anywhere with uniform, reproducible, hands-on experience with navigating and exploring big biomedical data.
- 2018-2019  New graduate program development (HIDS) with various courses incorporating approaches, methods, and on-line resources for hands-on training on how to handle Big Biomedical Data on a cloud.
G-DOC: The Georgetown Database of Cancer

Important features:

- Study based
- Combines omics data and clinical attributes
- Provides analytical tools on a fly (bringing tools to the data)
- User accounts with analysis sessions history
- Sharable findings – user groups on-line could access the same data, use the same tools and view the same results

https://gdoc.georgetown.edu/
EdX MOOC: Demystifying Biomedical Big Data: A User’s Guide

https://www.edx.org/course/demystifying-biomedical-big-data-a-users-guide

Demystifying Biomedical Big Data: A User’s Guide

Whether you are a student, basic scientist, researcher, clinician, or librarian, this course is designed to help you understand, analyze, and interpret biomedical big data.

Estimated 8 weeks
3–6 hours per week

Self-paced
Progress at your own speed

Free
Optional upgrade available

This course is archived
View course materials

Funded by NIH/NLM R25
# Masters in Health Informatics & Data Science (HIDS) at Georgetown University

[https://healthinformatics.georgetown.edu/](https://healthinformatics.georgetown.edu/)

<table>
<thead>
<tr>
<th><strong>What</strong></th>
<th>In-Person, Full-time or Part-time Masters program</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Who</strong></td>
<td>Undergraduate/Advance degrees in Computer Science, Statistics, Bioinformatics, Math, Medicine, Nursing &amp; Life Sciences and similar fields</td>
</tr>
</tbody>
</table>
| **Length (Full-time)** | 1 year (3 semesters)  
Fall-Spring-Summer, 32 credits total |
| **Competencies** | health informatics, health data science, big data analytics, digital ethics, artificial intelligence and machine learning, imaging informatics, precision health informatics applications to inform health outcomes and reduce costs |
| **Target Industries for employment** | Healthcare systems, Pharma/Biotech, Software companies, Academia, Device companies, Diagnostic companies, Healthcare consulting firms, and more. |
| **Ideal Location** | Located in Nation’s Capital Washington DC, which is home to numerous federal agencies (NIH, NCI, FDA, NIST, etc.) and over 2000 Life Sciences and Healthcare companies |
Unique Program Features

**HANDS-ON EXPERIENCE**
- Real world project experience in the classroom & through capstone projects with industry and government agencies

**EXPERTS**
- Courses taught by well-respected faculty who are practitioners and leaders in the field of health informatics

**FLEXIBILITY**
- 1-year accelerated, career-ready program with the flexibility to enroll part-time or full-time

**IDEAL LOCATION**
- Located in the nation’s capital Washington D.C., which is home to numerous federal agencies (NIH, NCI, FDA, NIST) and 2,000 biotechnology companies

**MULTI-DISCIPLINARY COURSEWORK**
- Students exposed to and interact with field experts, and faculty from top-ranked Georgetown University depts including Massive Data Institute, The McDonough School of Business, The Kennedy Institute of Ethics & Innovation Center for Biomedical Informatics (ICBI)

https://healthinformatics.georgetown.edu/
HIDS Curriculum Overview

**FALL**
- Introduction to Health Data Science & Analytics (HIDS-501)
- Utilizing Data in Electronic Medical Records (HIDS-502)
- Precision Health Informatics (HIDS-503)
- Evidence based data analysis in population health (HIDS-504)

**SPRING**
- Massive Data Fundamentals (ANLY-502)
- Advanced topics in Health Informatics (HIDS-505)
- AI for health applications (HIDS-506)
- Digital Health Applications (HIDS-507)

**SUMMER**
- Human Factor Engineering, Usability and Safety (HIDS-508)
- Imaging Informatics (HIDS-509)
- Mandatory Capstone (HIDS-510)

[https://healthinformatics.georgetown.edu/curriculum/course-schedule/](https://healthinformatics.georgetown.edu/curriculum/course-schedule/)
Industry and Government Advisory Panel for HIDS Capstone

- Sage Bionetworks
- ESAC
- FDA
- Deloitte
- Flatiron Health
- COTA
- Syapse
- TrinetX
- INOVA
- Meds
- Indivumed Group
- Argentys
- CVS Health
- PwC
- Open Source Electronic Health Record Alliance (OSEHRA)
- LexisNexis
- Assymetrik
- RadAmerica
- Cerner Corporation
- Medstar Institute for Innovation (MI2)
Part II: Using the Seven Bridges (SB)  
Cancer Genomics Cloud (CGC) in a classroom  

Krithika Bhuvaneshwar  
Research Instructor Faculty / Senior Bioinformatician,  
Health Informatics & Data Science (HIDS),  
Innovation Center for Biomedical Informatics (ICBI),  
Georgetown University  

Profile: https://icbi.georgetown.edu/Bhuvaneshwar  
Publications: bit.ly/krithikab
About the HIDS Masters Program

- **Health Informatics and Data Science (HIDS) Masters Program at Georgetown University**
  - 1 year MS program
  - Focus on hands-on workshops and real data examples
    - Python
    - R
    - SQL
    - Machine Learning and AI
    - Modern Cloud based tools
    - More.....

[https://healthinformatics.georgetown.edu](https://healthinformatics.georgetown.edu)
Why the Seven Bridges CGC platform?
Advantages of using a platform like Seven Bridges for data analysis

- Multiple tools “stitched” into one workflow – called an “app”.
- Easy to run this “app” – don’t need knowledge of command line
- Easy to run ”batch jobs” – multiple samples at once
- Easy to scale i.e. More computational resources as needed
- Access results on the cloud / online – through web browser
- Easy to review, and share results
- Colleagues can easily reproduce results
- No need to babysit data/workflow/local computer
SB-CGC pipelines used in the HIDS program

- **Course - Imaging Informatics**
  - Medical image segmentation pipelines

- **Course - Precision Health Informatics**
  - Next Generation Sequencing (NGS) pipelines

- Seven Bridges partnering with HIDS, providing free cloud credits for all students for these 2 courses
Advantage of using CGC platforms for teaching

- Simple visual representation of the steps in the pipeline
- Easy for students to understand basic concepts, before diving into command line details
- Easy to run pipeline
- Easy for instructor to check in-class work and homework
- Easy to teach students about input files, input files, workflows, tasks, cloud costs
SB-CGC environment set up for teaching/training
Summary of the SB-CGC set up

- One project for each student
- Instructor(s) added to each project as admins
- Use public/open source datasets and workflows for teaching
  - Data
  - Workflow (apps)
- How to run workflows in batch mode (parallel processing)
- Run in-class hands-on demo
Example screenshots

Project set up in SB-CGC

One project for each student
Inside a student project

Welcome to your new project!

Projects are the core building blocks of the CGC Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

Within your project, you can:
- Start exploring public datasets straight away
- Install your tools on the CGC and create workflows
- Upload your own private data and analyze it along with public datasets
- Collaborate securely with other researchers

Please record the details of your project here, such as its aims, experimental context, and any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the CGC are logged on the task page. This notepad is just for your own notes.

You can also use markdown here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the Knowledge Center

The Seven Bridges CGC Team

Citation

Project participants agree to acknowledge the funding for the CGC in all publications and external presentations, as follows:

*The Seven Bridges Cancer Research Data Commons Cloud Resource has been funded in whole or in part with..."
Students learn FAIR principles through SB-CGC

- The FAIR Guiding Principles for scientific data management
  - Findable (F)
    - Easily findable, searchable, unique ID
  - Accessible (A)
    - Metadata retrievable, open, free
  - Interoperable (I)
    - Easily integrate with other data
  - Reusable (R)
    - Easily reuse the data and metadata

https://www.nature.com/articles/sdata201618
Example Use case 1 – Pipeline for analysis of Medical Imaging
How to run a workflow on Seven Bridges CGC

Create/choose project

Select or Create a workflow ("app")

Select input files

Run the "app"

- Indicates the type of analysis, input and output etc.
- Takes in Input files
- Runs the workflow
- Produces the appropriate Outputs
Workflows in SB-CGC

**FSL Brain Segmentation**

FSL 5.0

This workflow is used to analyze brain MRI data. It enables image preprocessing (e.g. skull stripping, intensity inhom...)

**SBG TCIA Archive Viewer**

SBG 1.0

The TCIA Archive Viewer consists of two steps:

- The first step involves unpacking of the archive file using the SBGD...
Copy the “app” to the project

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Type</th>
<th>Modified by</th>
<th>Modified on</th>
<th>Toolkit (version)</th>
<th>CWL Version</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSL Brain Segmentation</td>
<td>This workflow is used to analyze brain MRI data. It enables image preprocessing (e.g., automatic intensity normalization)</td>
<td>Workflow</td>
<td>kb472</td>
<td>May 26, 2021 21:08</td>
<td>FSL (5.0)</td>
<td>sbg:draft-2</td>
<td>Run</td>
</tr>
<tr>
<td>SBG TCIA Archive Viewer</td>
<td>The TCIA Archive Viewer consists of two steps: * The first step involves unpacking ...</td>
<td>Workflow</td>
<td>kb472</td>
<td>May 26, 2021 21:08</td>
<td>SBG (1.0)</td>
<td>sbg:draft-2</td>
<td>Run</td>
</tr>
</tbody>
</table>
Query Public Datasets => Get Input files

In this example, we queried:
- Medical imaging MRI files
- One patient
- Brain Cancer in the
- TCIA collection
### Input Data/Files

<table>
<thead>
<tr>
<th>Name</th>
<th>Experimental strategy</th>
<th>Platform</th>
<th>Paired-end</th>
<th>Created on</th>
<th>Type</th>
<th>Size</th>
<th>Case ID</th>
<th>Gender</th>
<th>Race</th>
<th>Disease type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.148794664711827901715849860167.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>1.6 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.150346522548097750188956570502.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>1.6 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.240283449041382781659568683216.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>5.8 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.445635019400433103248745468431.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>1.4 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.466873485067873546718804101423.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>1.7 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.507509430627489084746877138456.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>1.5 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.68597855331495334440262662908.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>197.6 KIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
<tr>
<td>1.3.6.1.4.1.14519.5.2.1.1706.4001.779013211760175351323192631290.zip</td>
<td>-</td>
<td>-</td>
<td></td>
<td>May 26, 2021 21:22</td>
<td>ZIP</td>
<td>1.5 MIB</td>
<td>TCGA-02-0009</td>
<td>female</td>
<td>white</td>
<td>Glioblastoma Multiforme</td>
</tr>
</tbody>
</table>

**Patient clinical data/meta data**
Run workflow

Batch mode (parallel processing)

Workflow to uncompress and view MRI images
Check Status of a task

Shows 4 child tasks running (blue color).
4 child tasks completed (green color)
Go to tab TASKS – workflow completed

Cost = 0
Run time < 1 min
Workflow complete

- Cost <10 cent.
- Run time = 1 - 2 min

Concepts learnt at end of class
- Workflows (app) for medical imaging and genomics
- Open source workflows
- Public datasets
- Running a workflow in batch mode
- Expected outputs
- Cloud costs
Example use case 2 – Run python notebook on SB-CGC
Use case 2

• We worked with David Roberson from the SB-CGC Team
• David demoed a real life example
  • COVID-19 Lung segmentation using Deep Learning
• CGC-SB environment
  • Jupyter Lab notebook
  • SB Machine learning environment with Tensor Flow
  • Instance type (GPUs)
    • p3.2xlarge (1024 GB EBS, 8 vCPUs, 61 GB RAM, 1 GPU)
    • Price: $3.2 per hour
Files inside student project

All relevant information
• Data
• Python notebooks
• Models

Slides
Python notebook

**Concepts learnt at end of class**
- Experience with running command line code on the cloud
- Experience with using GPU and Tensor flow in SB-CGC
- Expected outputs
- Cloud costs
- Data / Algorithm
- All aspects of a data science project
Example use case 3 – Genomics pipelines
NGS pipelines - RNA-seq and DNA-seq

How to process raw NGS files (FASTQ files)
RNA-seq pipeline on SB

Input:
- FASTQ file
- Genome Reference file
- GTF file (info about gene and CDS location)

Output files
- From alignment
- Gene and Isoform Quantification
Run the workflow

- Set app to run for a batch of samples as input

1. Batching ON
2. GTF file
3. Input files
4. Reference file
5. Run workflow
Output files from workflow

Input:
• FASTQ file
• Genome Reference file
• GTF file (info about gene and CDS location)

Cost = 40 cents
Duration = About 30 minutes
(using spot instances)

Run time depends on size of VM chosen

Gene expression results
Transcript BAM file after alignment
Quality check of BAM files after alignment
Quality check of input FASTQ files
Log files
Trimmed FASTQ files
DNA-seq pipeline on SB

Input:
- FASTQ file
- Genome Reference file

Output:
- Variants – in a Variant Call File (VCF file)
- VCF file with and without annotation
- Filter/tag low quality variants
Concepts learnt at end of class
- Basic concepts about NGS data analysis
- Running RNA-seq, WGS and WXS pipeline
- What the input and output files are
- Cloud costs
- File sizes for genomics data
- Run times based on size of VM
Student feedback

- Very positive feedback from students

- Able to focus on concepts instead of worrying about
  - Environment set up
  - Tool installations
  - Version related challenges
  - Cloud costs
Summary

Great tool to teach concepts of cloud computing

- Data - already present in the cloud
- Tools/Pipeline - already present in the cloud
- Scalable computing options
- Learn about cloud costs
- Learn about parallel processing

Easy management

- Monitor student work easily
- Students easily reproduce in-class work
- Students can use SB-CGC for their capstone projects
- Students can keep copies of their work after end of the course