The SBG Image Processing Toolkit
Easy Image Classification with No Coding Needed

Soner Koc
Ana Stankovic
Agenda

1. Overview
2. Power of AI on the CGC
3. Image Processing Toolkit
4. Demo: Image Analysis on the CGC
5. Conclusion and Q/A
Power of AI on the CGC
CGC Support for Machine Learning

- **Seamless integration** of Data, Code, Model, and Compute Environment for **Reproducible Pipelines** with the power of multiple GPUs

- Powerful and **Robust Documentation**
  - docs.cancergenomics.org/docs

- **Bridging the Gap** Between Exploratory Computing Environment and Production Application Easily and User-friendly.

- **Best Security** and Compliance

- **Expert Scientific Support** Team, Trained in Machine Learning
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

Data cleaning
Feature engineering
Model selection
Hyperparameter tuning
Model training
Model predictions
Experiment logging
Analysis & Interpretability
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

- Data cleaning
- Feature engineering
- Model selection
- Hyperparameter tuning
- Model training
- Model predictions
- Experiment logging
- Analysis & Interpretability
CGC Support for Machine Learning

- Data
  - TCGA, TCIA
  - SevenBridges Platforms
  - Any Public/Private Dataset

- ETL, Data Ingestion
  - Raw Data Cleaning
  - Raw Data Transform
  - Data Stats

- Feature Extraction
  - Define Objectives
  - Model Training
  - Hyperparameter Tuning

- Model Engineering
  - Interactive Analysis
    - Setup Environment
    - Clone GitHub Repository
    - Visualize Intermediate Results

- Model Evaluation
  - Monitoring
  - Diagnosis
  - Visualization
  - Train, Validation, Test Evaluation

- Workflow/App
  - Portable, Reproducible, Replicable ML/DL workflows
CGC Support for Machine Learning

- **Data**
  - TCGA, TCIA
  - SevenBridges Platforms
  - Any Public/Private Dataset

- **ETL, Data Ingestion**
  - Raw Data Cleaning
  - Raw Data Transform
  - Data Stats

- **Feature Extraction**
  - Define Objectives
  - Model Training
  - Hyperparameter Tuning

- **Model Engineering**
  - Setup Environment
  - Clone GitHub Repository
  - Visualize Intermediate Results

- **Model Evaluation**
  - Monitoring
  - Diagnosis
  - Visualization
  - Train, Validation, Test Evaluation

- **Workflow/App**
  - Portable, Reproducible, Replicable ML/DL workflows

---

Cancer Genomics Cloud
CGC Support for Machine Learning

Data
- TCGA, TCIA
- SevenBridges Platforms
- Any Public/Private Dataset

ETL, Data Ingestion
- Raw Data Cleaning
- Raw Data Transform
- Data Stats

Feature Extraction
- Define Objectives
- Model Training
- Hyperparameter Tuning

Model Engineering

Model Evaluation
- Monitoring
- Diagnosis
- Visualization
- Train, Validation, Test Evaluation

Workflow/App
- Portable, Reproducible, Replicable ML/DL workflows

Interactive Analysis
- Setup Environment
- Clone GitHub Repository
- Visualize Intermediate Results

Cancer Genomics Cloud
CGC Support for Machine Learning

Data
- TCGA, TCIA
- SevenBridges Platforms
- Any Public/Private Dataset

ETL, Data Ingestion
- Raw Data Cleaning
- Raw Data Transform
- Data Stats

Feature Extraction
- Define Objectives
- Model Training
- Hyperparameter Tuning

Model Engineering

Model Evaluation
- Monitoring
- Diagnosis
- Visualization
- Train, Validation, Test Evaluation

Workflow/App
- Portable, Reproducible, Replicable ML/DL workflows

Interactive Analysis
- Setup Environment
- Clone GitHub Repository
- Visualize Intermediate Results

Cancer Genomics Cloud
CGC Support for Machine Learning

- **Data**
  - TCGA, TCIA
  - SevenBridges Platforms
  - Any Public/Private Dataset

- **ETL, Data Ingestion**
  - Raw Data Cleaning
  - Raw Data Transform
  - Data Stats

- **Feature Extraction**
  - Define Objectives
  - Model Training
  - Hyperparameter Tuning

- **Model Engineering**

- **Model Evaluation**
  - Monitoring
  - Diagnosis
  - Visualization
  - Train, Validation, Test Evaluation

- **Workflow/App**
  - Portable, Reproducible, Replicable ML/DL workflows

---

Cancer Genomics Cloud
CGC Support for Machine Learning

Data
- TCGA, TCIA
- SevenBridges Platforms
- Any Public/Private Dataset

ETL, Data Ingestion
- Raw Data Cleaning
- Raw Data Transform
- Data Stats

Feature Extraction
- Define Objectives
- Model Training
- Hyperparameter Tuning

Model Engineering

Model Evaluation
- Monitoring
- Diagnosis
- Visualization
- Train, Validation, Test Evaluation

Workflow/App
- Portable, Reproducible, Replicable ML/DL workflows

Interactive Analysis
- Setup Environment
- Clone GitHub Repository
- Visualize Intermediate Results

Cancer Genomics Cloud
Fast Prototyping
Interactive Analysis

- **Familiar Environment** (Jupyter Lab and RStudio) Integrated in the Platform for ML Analysis
- **No Need** to Download or Move Data to Analyze them Interactively!
- **Easy Visualizations** and Notebooks Previews
- **Division Predefined Environments** are Available on Data Cruncher
  - No need to install the libraries every time the analysis starts!
  - Docker images with frequently used libraries by the client, so that prerequisites are already installed when Interactive analysis starts
  - Reproducibility
Fast Prototyping

Public/Private Data and Codebase

- Data
  - Brain Cancer Predictive Modeling and Biomarker Discovery Challenge
- Gene Expression, DNA Methylation, Phenotype Data
- Code Repository

Jupiter Notebook

- Clone GitHub Repository into CGC Data Cruncher
- Setup DL Environment

Visualize the Result

Model Engineering and Evaluation

Execute ML/DL Model

Terminal
Fast Prototyping

More Flexibility

- **Data Drift.**
  
  Trained on a certain distribution of inputs, but this distribution changes over time.

- **Concept Drift.**
  
  Trained to learn an x->y mapping, but the statistical relationship between x and y changes.

- **Changing Requirements.**
  
  Built to perform a particular task, but the product team decides to modify its capabilities.
How is Your ML Tool Executed on the Platform?

1. According to the resource description in the CWL, Scheduler assigns an empty instance to this job.

2. Executor mounts the input files to the instance and pulls the image (according to the CWL) from the SB Docker repository, then creates a container.

3. According to the CWL, the Executor generates CMD, and runs the job.

4. After the job is done, the output will be fetched back to storage.
Optimize and Debug ML Tool in an Easier Manner

View Instance Metrics

- Instance type, purchasing type and status
- Instance configuration - i.e. available vCPUs, Memory, Disk space
- CPU, Memory, and Disk usage

Instance Metrics are preserved for 2 weeks after the task completion

Real Time Job Monitoring

- Live Logs
- Standard output/error streams
- Workspace directory
- Workspace files preview

This is available during the duration of the task, while instance is active.
Reproducibility in ML for Healthcare*

• Technical Replicability
  - Privacy Sensitive Data
  - Missing/Unorganized Code
  - Test Environment Setup

• Statistical Replicability
  - Dataset Differences

• Conceptual Replicability
  - Lack of Multi-Institution Data

*Reproducibility in machine learning for health research: Still a ways to go, Matthew et. al. 2021 STM
Image Processing Toolkit
SBG Image Processing Toolkit

• **Goal:** Enable users to create machine learning (ML) image classifiers on the CGC without coding

• **Highlights:**
  - Utilize platform computing
  - Flexible for various types of image data
  - Allows image processing and genomic data integration
Image Processing Tools

- **Deep learning**
  - SBG Deep Learning Image Classification Exploratory Workflow
  - SBG Deep Learning Prediction

- **Quality control**
  - HistoQC tool (open-source tool)

- **Preprocessing**
  - SBG Histology Whole Slide Image Preprocessing
  - SBG Stain Normalization
  - SBG X-Ray Image Preprocessing Workflow

- **Utility**
  - SBG Medical Image Convert
  - SBG Split Folders
Image processing flow
QC & Preprocessing tools

HistoQC

SBG Whole Slide Image Preprocessing

SBG Stain Normalization

SBG X-Ray Preprocessing Workflow
Deep learning apps

- Main features:
  - Deep learning over classic machine learning
  - Keras and Tensorflow
  - Transfer learning - use of pre-trained models
Deep learning apps

- **Exploratory classification**
  - SBG Deep Learning Image Classification Exploratory Workflow

- **Prediction**
  - SBG Deep Learning Prediction

SBG Deep Learning Image Classification Exploratory Workflow is an image classifier pipeline. It relies on the...
Pipeline Steps
1. Create multiple configurations
2. Support multiple image formats
3. Deep learning Image processing
4. Prepare comprehensive reports

Inputs
- Train Directory (Required)
- Test Directory (Required)
- Validation Directory (Optional)

Outputs
- Interactive HTML Report
- Visualisations from best configuration
- Metrics from all configurations
- Visualisations from all configurations
SBG Deep Learning Image Classification Exploratory Workflow

Image Processing Report

Configuration string

User can see the results for each model in the HTML report and use the best model for SBG Deep Learning Prediction tool by using the CONFIGURATION STRING

SBG Deep Learning Prediction tool

Image classifier that trains models on all labeled data in order to be able to classify unlabeled images that weren't included in the training or validation set

- Takes the best model configuration, provided by the Exploratory Workflow
- Trained model is then used to classify unlabeled images

![Image 1](image1.png)

labeled dataset

![Image 2](image2.png)

SBG Deep Learning Prediction

SBG Deep Learning Prediction is an image classifier tool that classifies unlabeled images based on labeled data.

Run

Copy

![Image 3](image3.png)

image label + prediction probability
Live Demonstration - Image Classification

1. Set up an example project
2. Perform exploratory image classification to find the best model
3. Perform image label prediction on an unlabeled dataset
Performance Benchmarking - SBG Deep Learning Image Classification Exploratory Workflow

<table>
<thead>
<tr>
<th>Number of images</th>
<th>Average image size</th>
<th>Number of configurations</th>
<th>Number of parallel instances</th>
<th>Execution time*</th>
<th>Price*</th>
<th>Instance type</th>
</tr>
</thead>
<tbody>
<tr>
<td>940</td>
<td>250 KiB</td>
<td>10</td>
<td>1</td>
<td>3h 57m 22s</td>
<td>$4.10</td>
<td>p2.xlarge</td>
</tr>
<tr>
<td>5856</td>
<td>200 KiB</td>
<td>1</td>
<td>1</td>
<td>42m 43s</td>
<td>$0.77</td>
<td>p2.xlarge</td>
</tr>
<tr>
<td>5856</td>
<td>200 KiB</td>
<td>24</td>
<td>3</td>
<td>8h 56m 45s</td>
<td>$28.02</td>
<td>p2.xlarge</td>
</tr>
<tr>
<td>650</td>
<td>20 MiB</td>
<td>1</td>
<td>1</td>
<td>54m 39s</td>
<td>$0.99</td>
<td>p2.xlarge</td>
</tr>
<tr>
<td>650</td>
<td>20 MiB</td>
<td>24</td>
<td>5</td>
<td>5h 57m 23s</td>
<td>$30.94</td>
<td>p2.xlarge</td>
</tr>
</tbody>
</table>

*Benchmarking was performed on On-Demand AWS instances*
Live Demonstration - Fast Prototyping

1. NCI-DOE-Collab-Pilot1-Tumor_Classifier repository
2. Create a tool
3. Integrate tool into the exploratory workflow
Live Demonstration - Conclusion

1. Easy access to image processing apps
2. Fast and cheap analysis and image prediction by using the newly published apps
3. Easy integration of custom machine learning tools
Conclusion and Future Work
Conclusion

• Summarized How Easy **Scope AI Project and Deploy** on the CGC

• Image Processing **Toolkit is Publicly Available**
Conclusion and Future Work

• Continue to **Extend** Publicly Available Image Processing Toolkit
  - More Pre-processing and Post-processing Options
  - More State-of-the-art Models and Workflows

• **Multi-omics Deep Learning** Solutions

• **Multi-omics with Imaging** Deep Learning Solutions
Acknowledgement

Nevena Nikolić

Jovana Babić

Sai Subramanian

Dennis Dean

Vladimir Kovačević

Ana Popić
Thank you

Questions?

- soner.koc@sevenbridges.com
- ana.stankovic@sevenbridges.com
- cgc@sevenbridges.com