Pegasus Users Group

MEETING

Machine Learning Workflows using Pegasus

Patrycja Krawczuk

Research Assistant

25th February 2021

Pegasus is funded by the National Science Foundation under grant #1664162

Outline

1. Steps of Machine Learning (ML) Workflows
2. Pegasus for ML Workflows
3. Work of ML Workflows team at Scitech lab

Machine Learning in Science

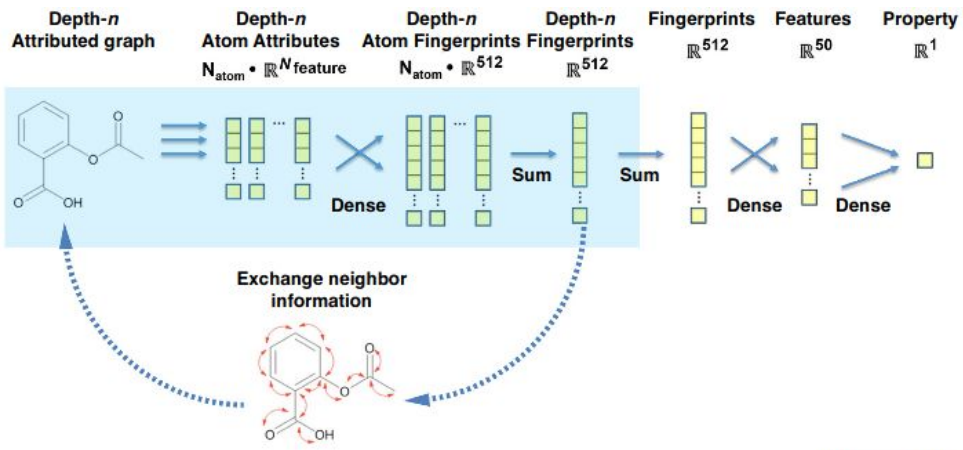


Image **A** Source: A.Lavecchia, "Deep Learning in Drug Discovery: opportunities, challenges, and future prospects."

Image **B** Source: A. Khan et al. "Deep Learning at Scale for the Construction of Galaxy Catalogs in the Dark Energy Survey."

Image **C** Source: DeepMind

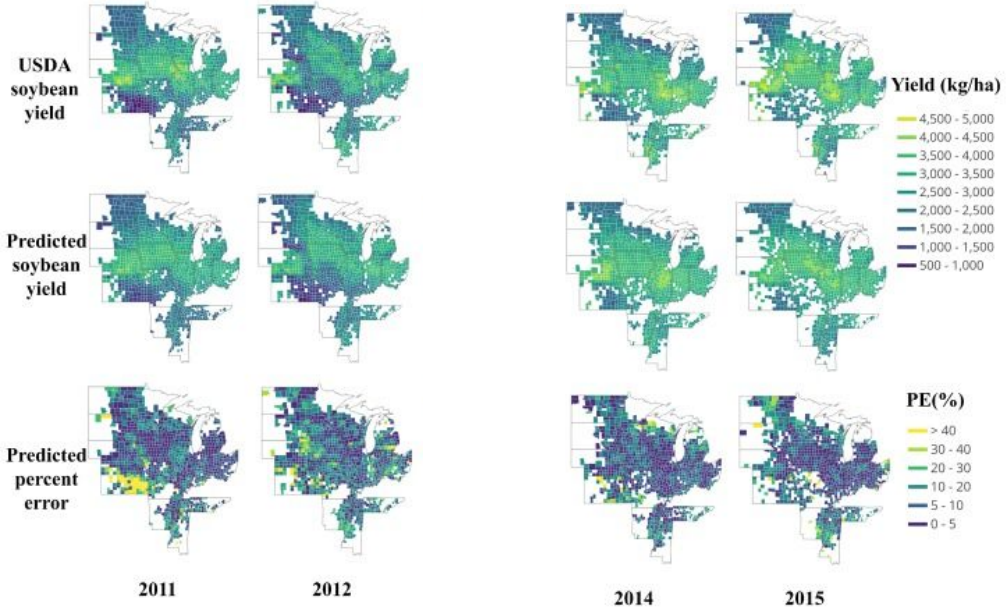
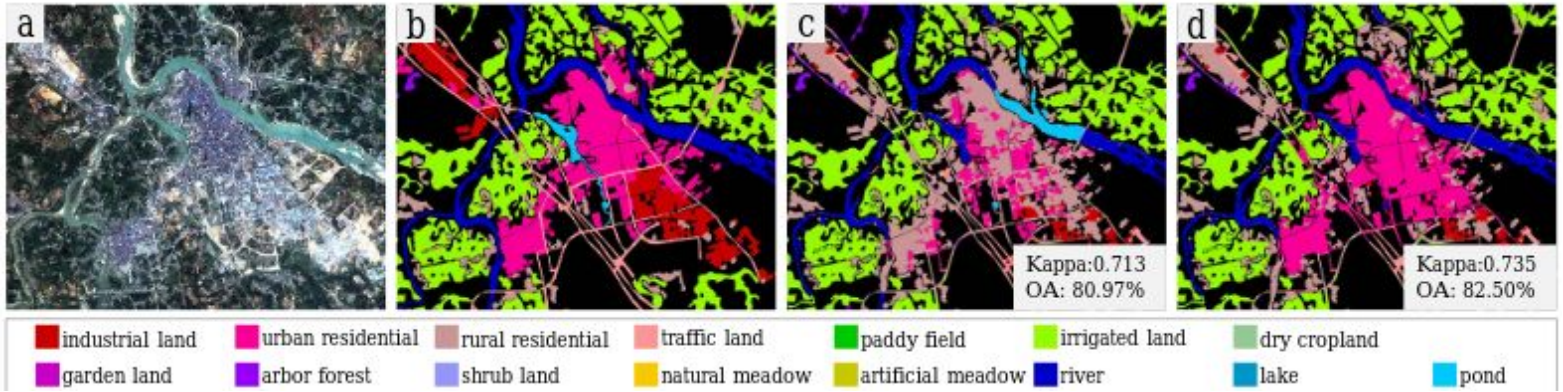
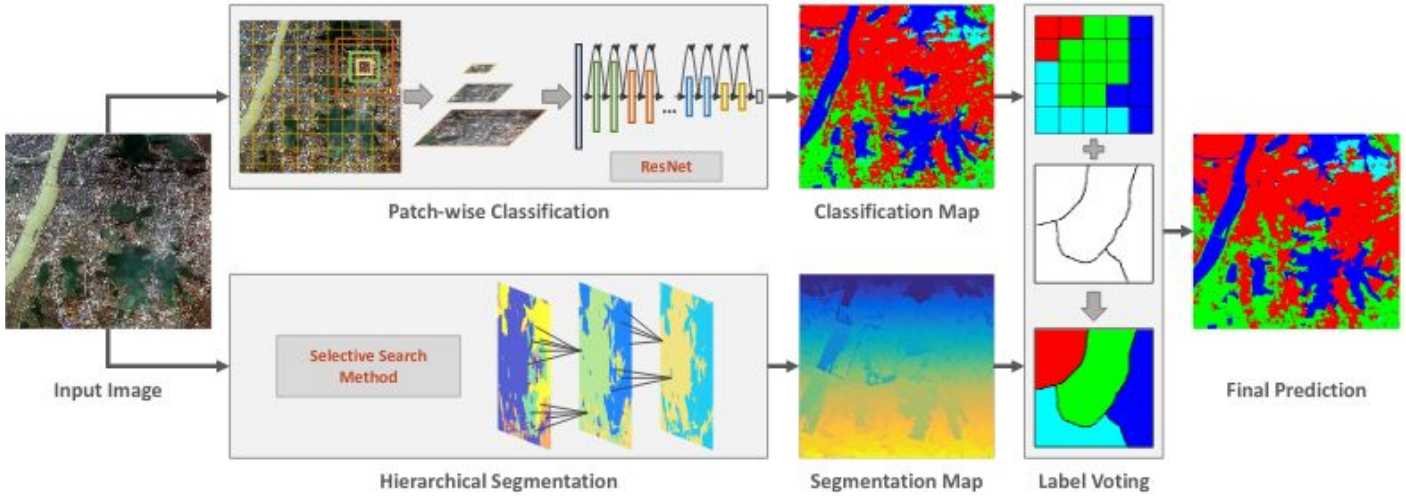


Image **D** Source: J.Sun et al. "County-Level Soybean Yield Prediction Using CNN-LSTM Model".

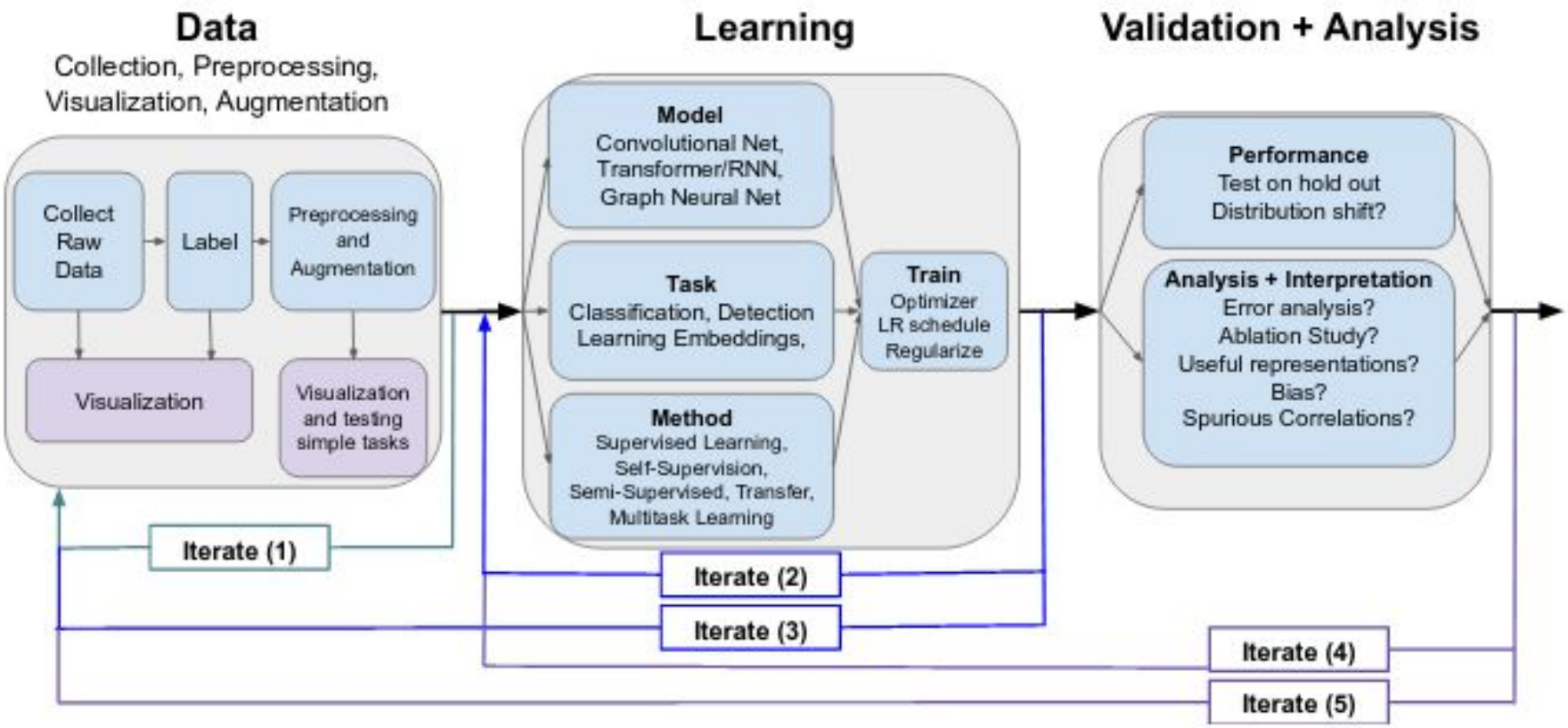
Machine Learning in Science: Land-Cover Classification





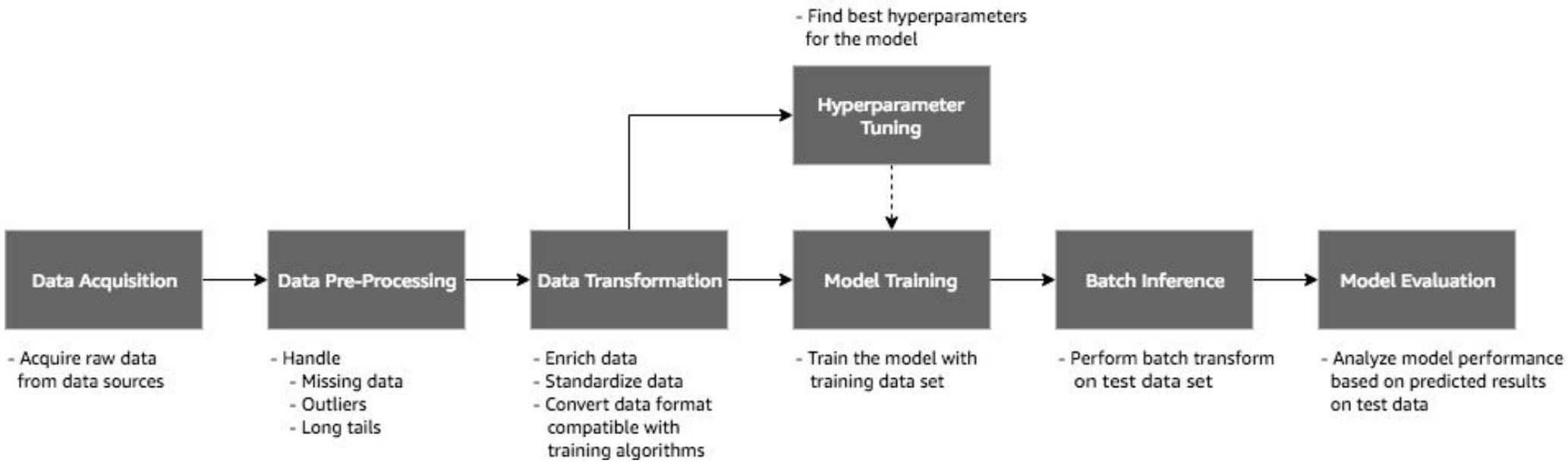
Source: S.Li et al. ,Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models"

Machine Learning Workflows: General View I



Source: M. Raghu et al. "A Survey on Deep Learning for Scientific Discovery"

Machine Learning Workflows: General View II

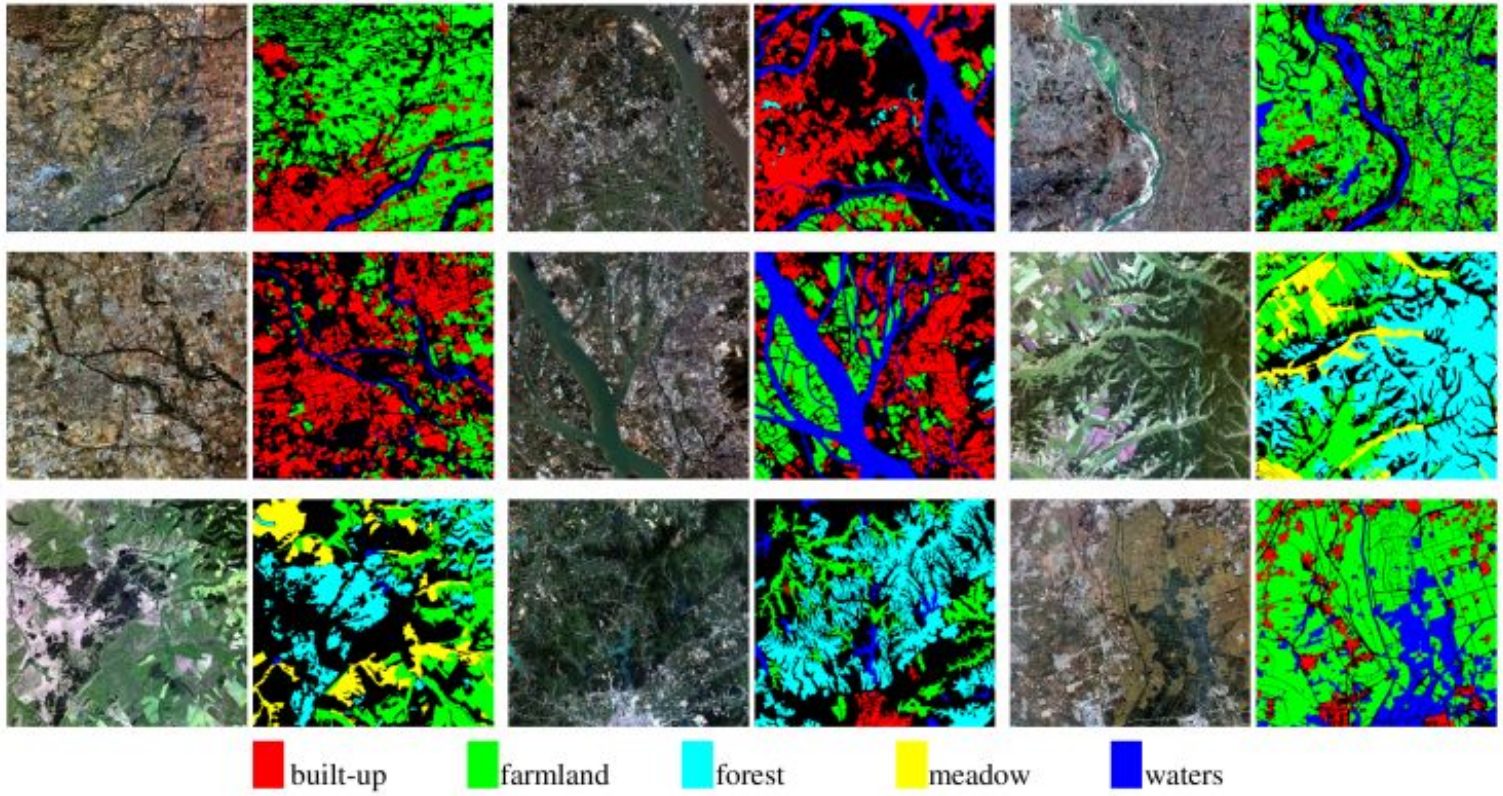


Source: H. Li et al. "Build end-to-end machine learning workflows with Amazon SageMaker and Apache Airflow"

Data Acquisition

Gaofen Image Dataset (GID) contains images from Gaofen-2 (GF-2) well-annotated dataset
High-Resolution Remote Sensing (HRRS) images up to 4m, covers more than 50,000 km

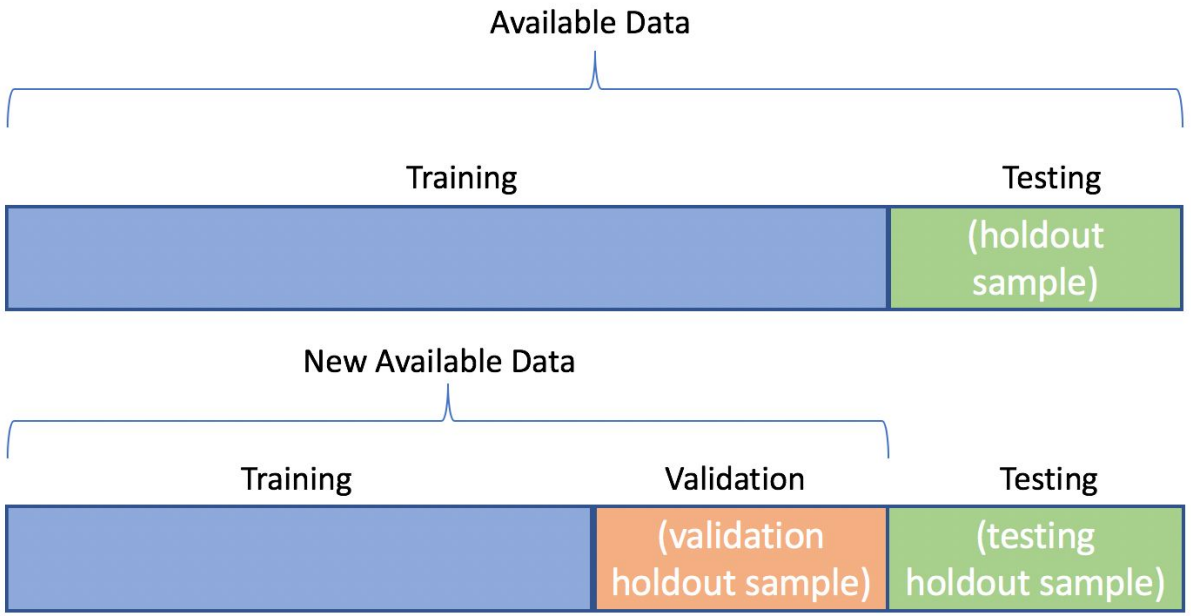
Gaofen-1 (GF-1), Jilin-1 (JL-1), Ziyuan-3 (ZY-3), Sentinel-2A (ST-2A)
and Google Earth images of Wuhan, Hubei (GE-WH)

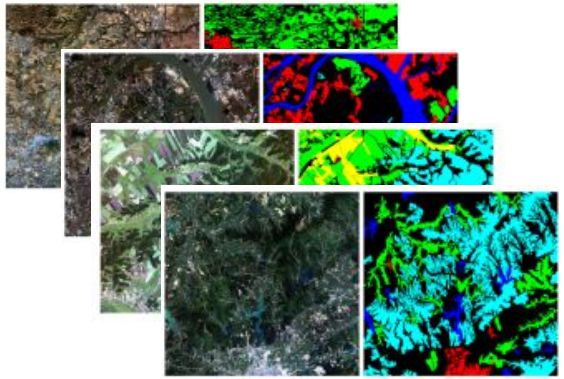


Data Management:
Integrity Checking

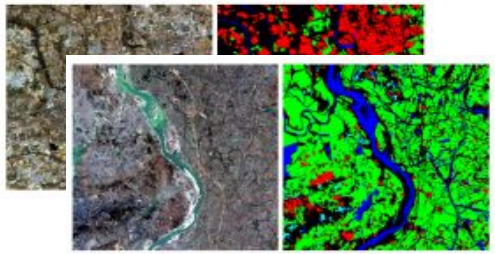
Source: H. Li et al. "Differentially Private Synthesization of Multi-Dimensional Data Using Copula Functions."

Data Split

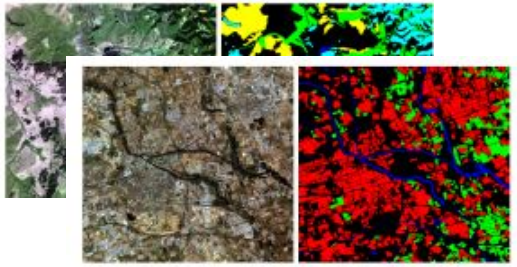




Train Dataset



Validation Dataset



Test Dataset

Data Management: Integrity Checking

Source: S.Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.”

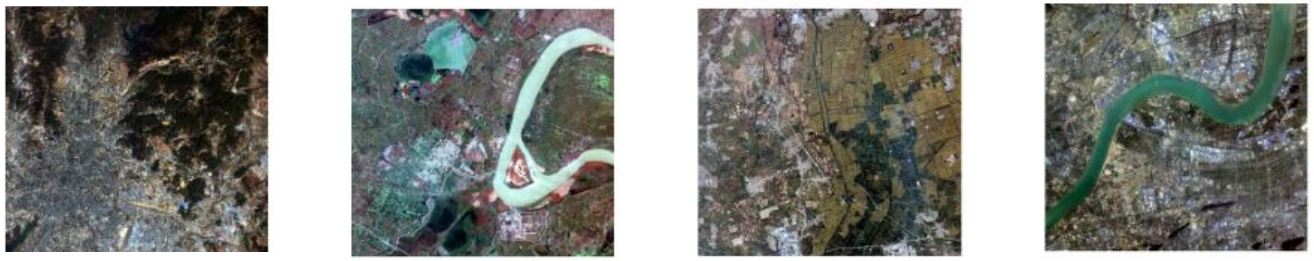
Data Preprocessing

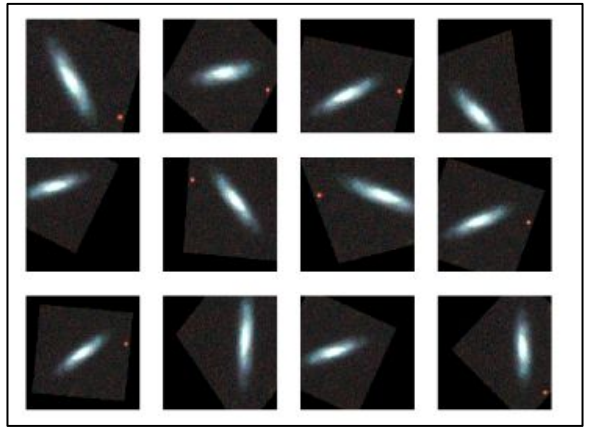
Train Dataset Preprocessing:

- Requantizes the images to 8-bit using optimized linear search (ENVI)
- Sample patches of size 56x56, 112x112, 224x224
- Resize the images to 224x224x4

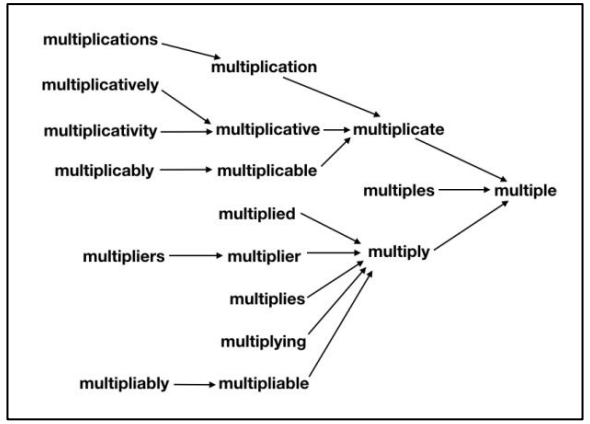
Test Data Preprocessing:

- Requantizes the images to 8-bit using optimized linear search (ENVI)
- Partition into patches with multi-scale siding window based on image resolution
- Resize the images to 224x224x4





Data Augmentation includes random rotations, flips, zooms, height and width shifts.



Lemmatization is the process of grouping together the inflected forms of a word so they can be analysed as a single item.

Distributed Execution:
Data Parallelism

Source: A. Khan et al. "Deep Learning at Scale for the Construction of Galaxy Catalogs in the Dark Energy Survey."

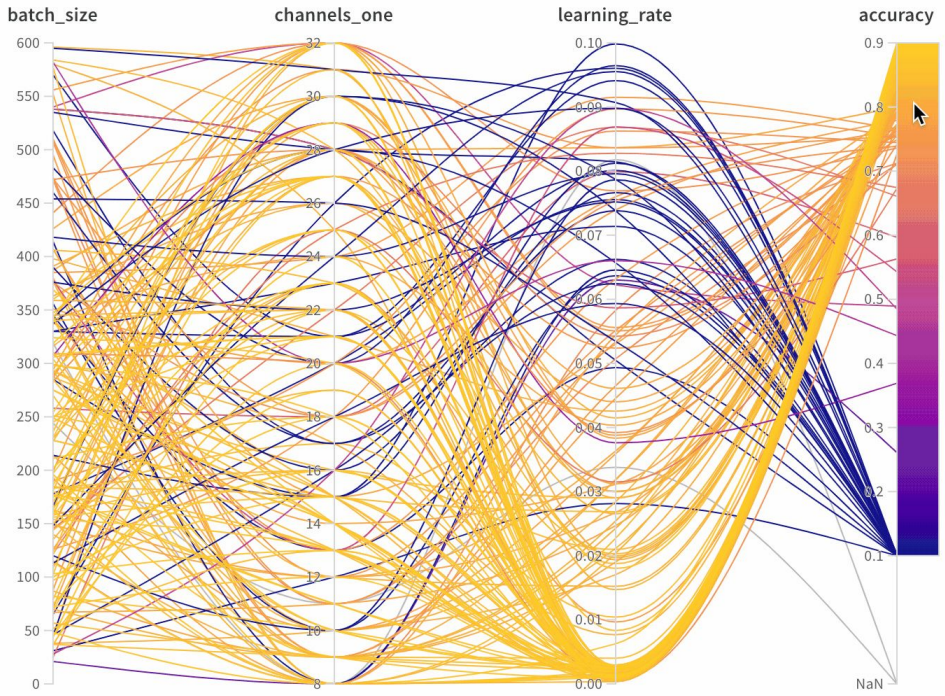
Hyper-parameter Optimization (HPO)

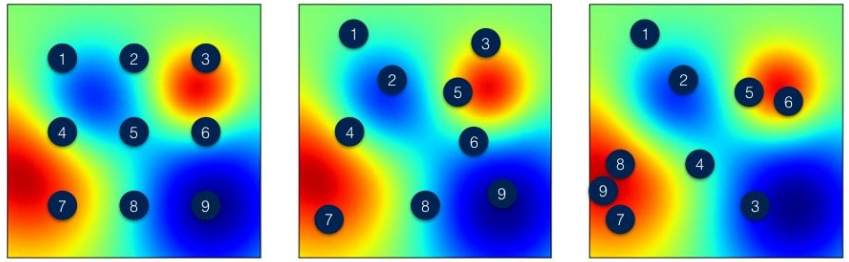
Hyperparameters:

- Learning Rate
- Number of Epochs
- Batch Size
- Momentum Regularization Constant
- And many more depending on model and its architecture

Hyperparameters Search Techniques:

- Grid Search
- Random Search
- Bayesian Optimization.
- Gradient-Based Optimization
- Evolutionary Optimization





Container Execution

Fault Tolerance: Checkpointing

Model Training

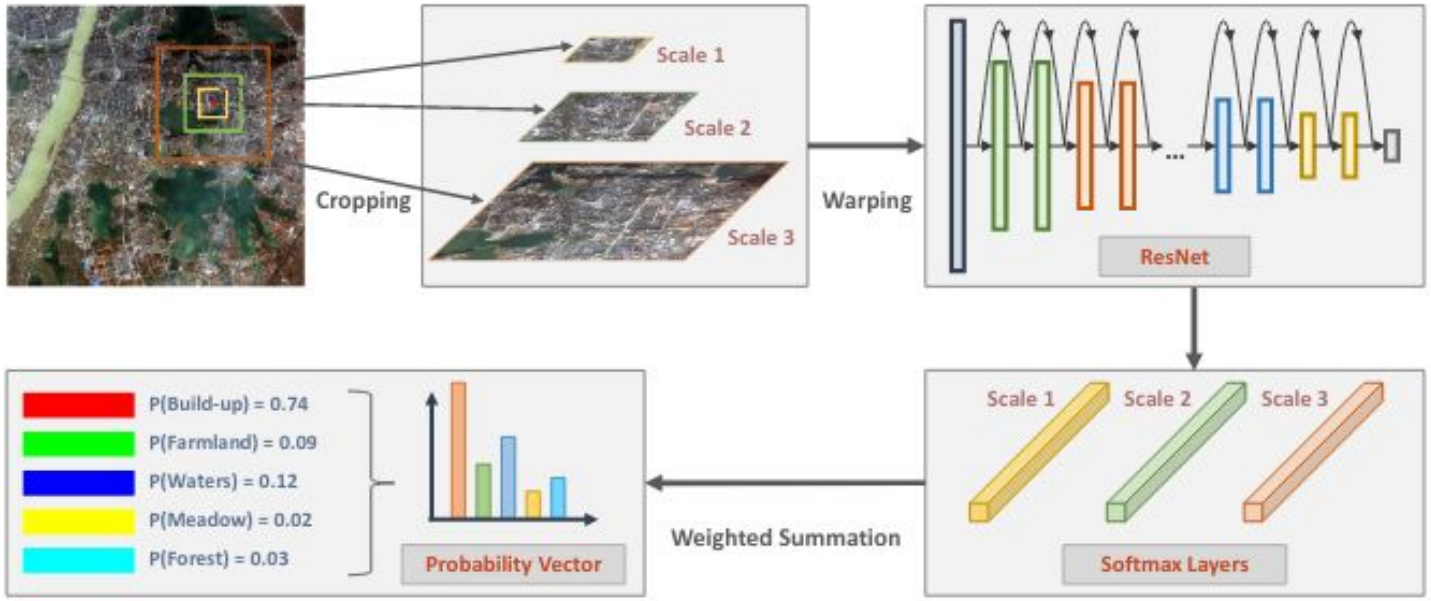


Figure 4: Multi-scale contextual information aggregation.

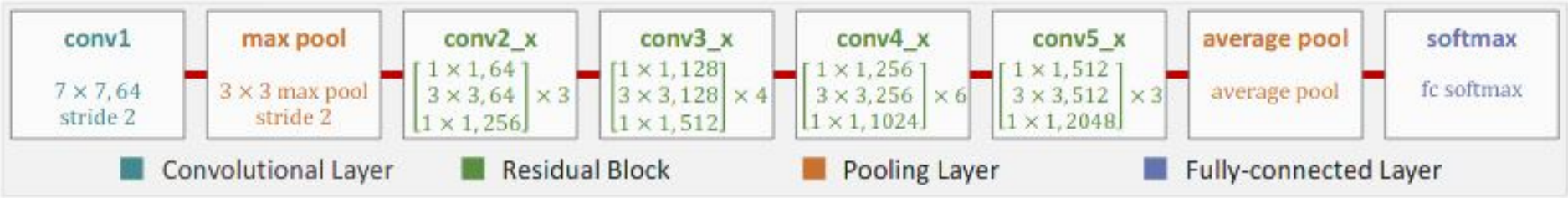
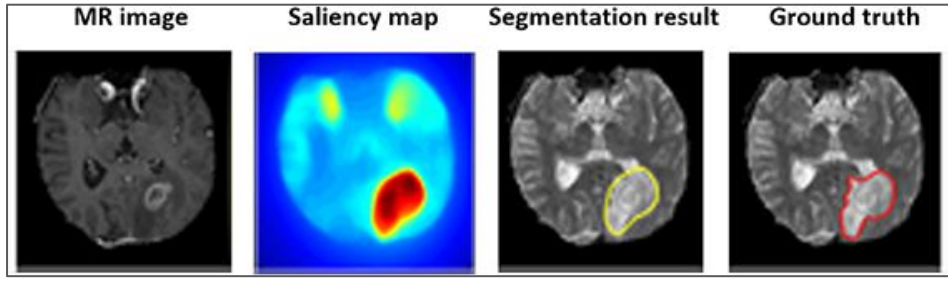


Figure 1: The structure of ResNet-50. Different structures are represented by different colors.

Fault Tolerance: Checkpointing

Model Evaluation



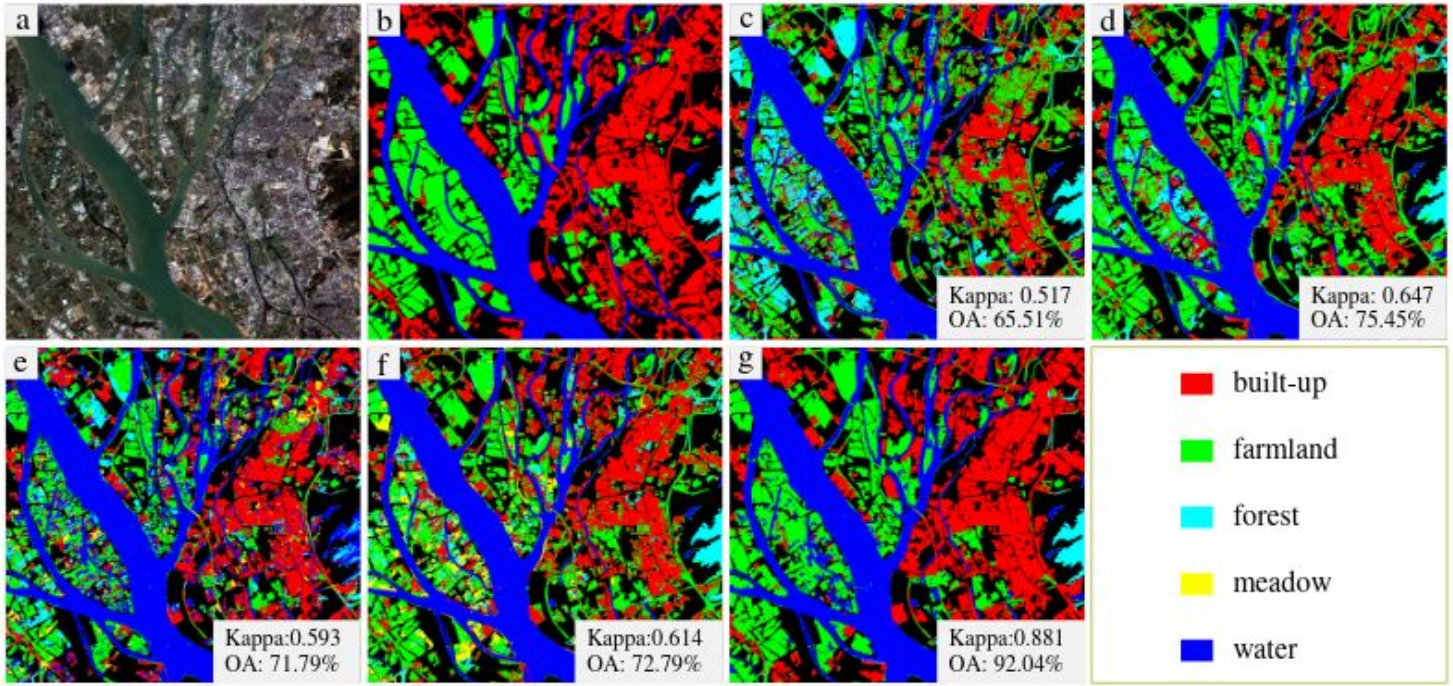
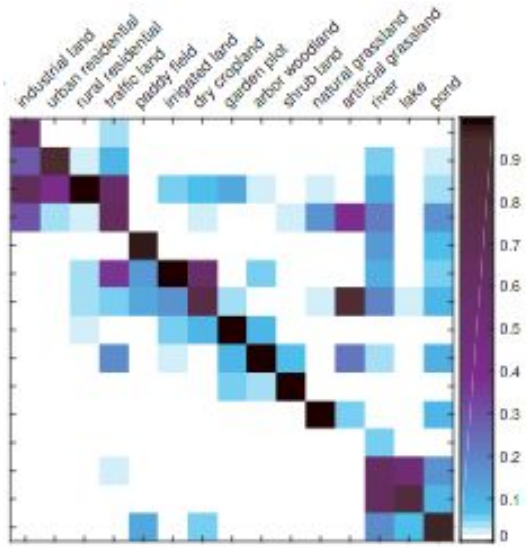


Figure 12: Land-cover classification maps of a GF-2 image obtained in Dongguan, Guangdong Province on January 23, 2015. (a) The original image. (b) Ground truth. (c)-(g) Results of MLC+Fusion, RF+Fusion, SVM+Fusion, MLP+Fusion, and PT-GID.

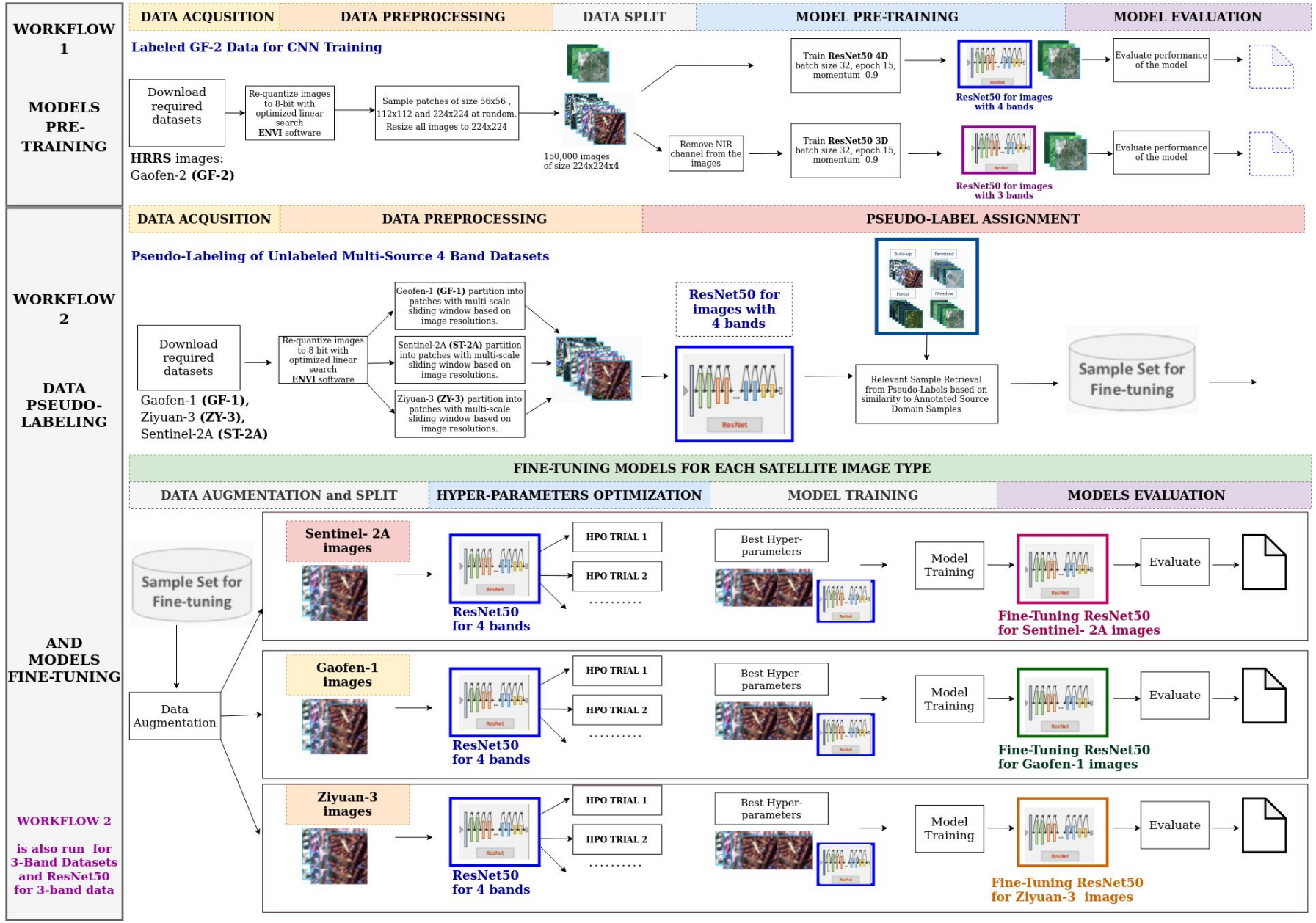


(c) FT-$\mathbf{U}_{tg}$ (multi-source)

Data Management:
Integrity Checking

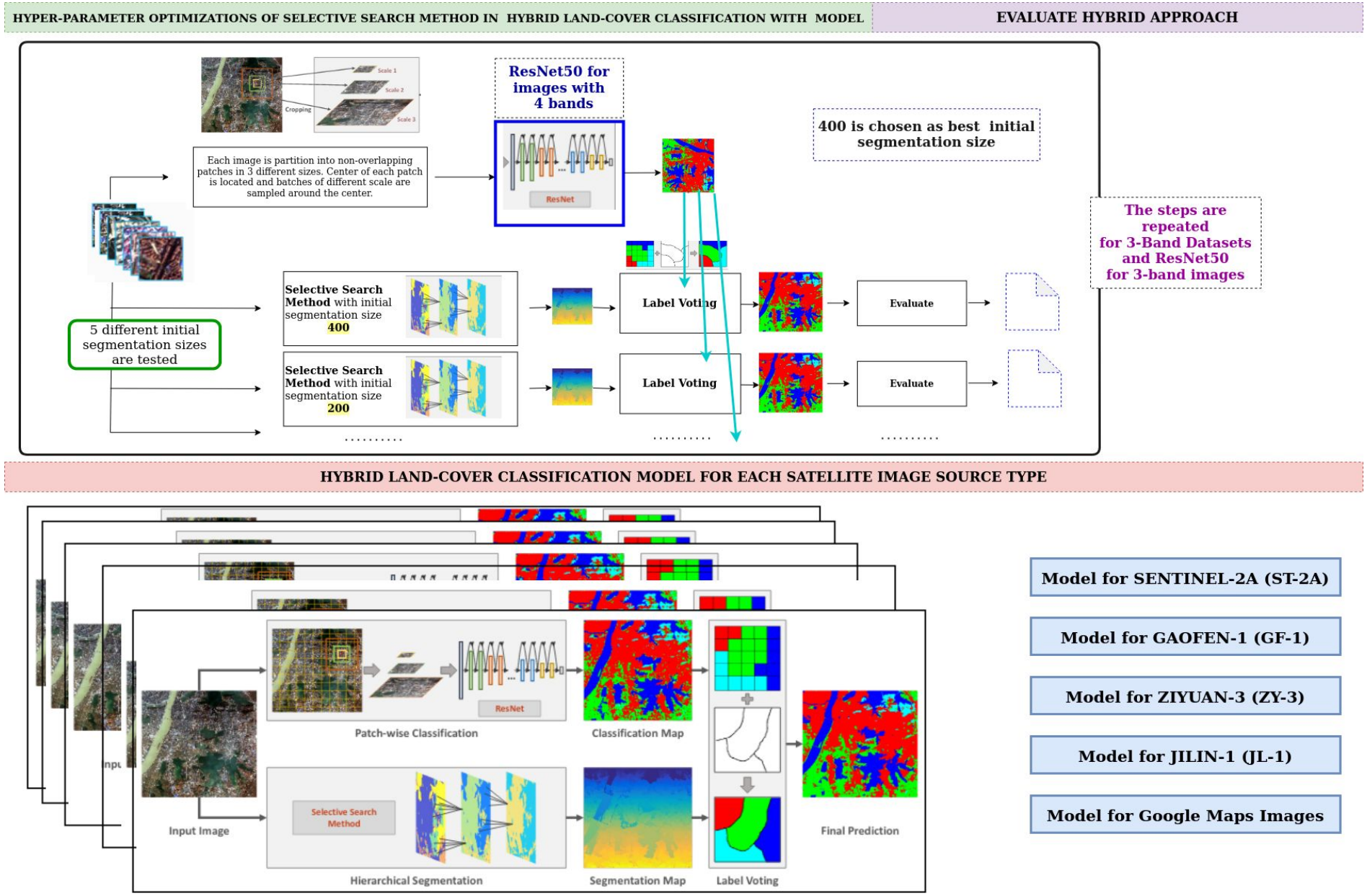
Source: Ben-Ahmet et al. “Visual saliency for medical imaging and computer-aided diagnosis”

Land-Cover Classification Workflow I



Source: S.Li et al. ,Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models

Land-Cover Classification Workflow II



Source: H. Li et al. "Build end-to-end machine learning workflows with Amazon SageMaker and Apache Airflow"

ML Workflows in Pegasus:

Data Management

Parallelization

Checkpointing

Container Execution

ML Workflows Group

Ryan Tanaka M.S.
Programmer Analyst II

Patrycja Krawczuk
Graduate Student (Ph.D.)

Srujana Subramanya
Graduate Student (MS)
Galaxy Morphology Classification

Aditi Jain
Graduate Student (MS)
Lung Segmentation (X-ray images)

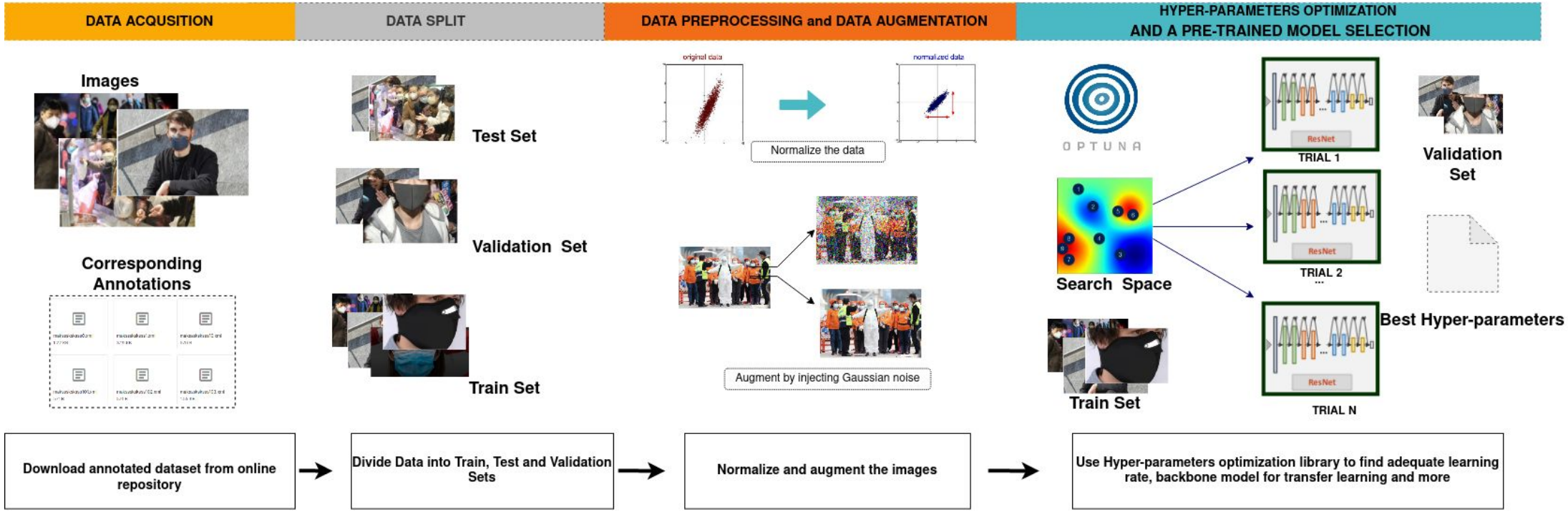
Shubham Nagarkar
Graduate Student (MS)
Crisis Computing

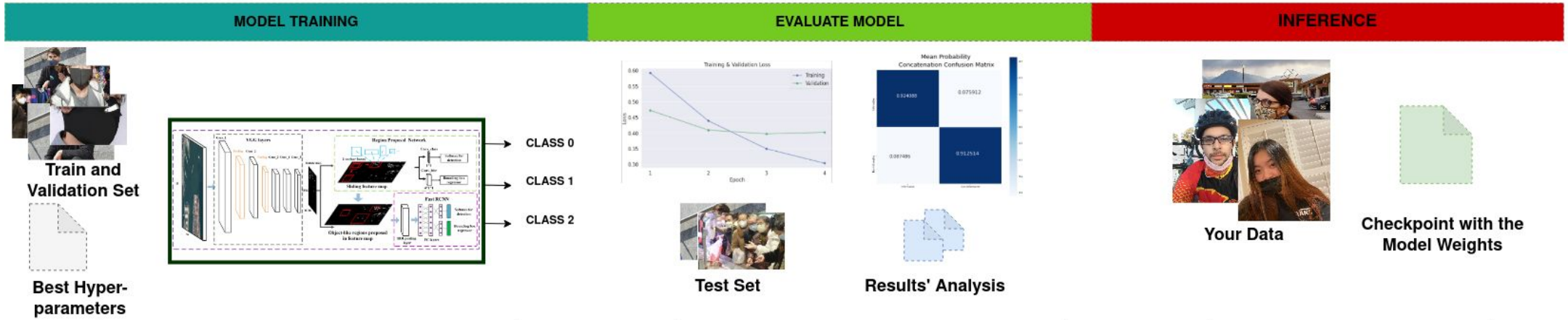
Kelsie Lam
Intern (HS)
Face Mask Detection and Classification

Rebecca White
Undergraduate Student
Classification of noise transients in LIGO data

Mask Detection and Classification Workflow

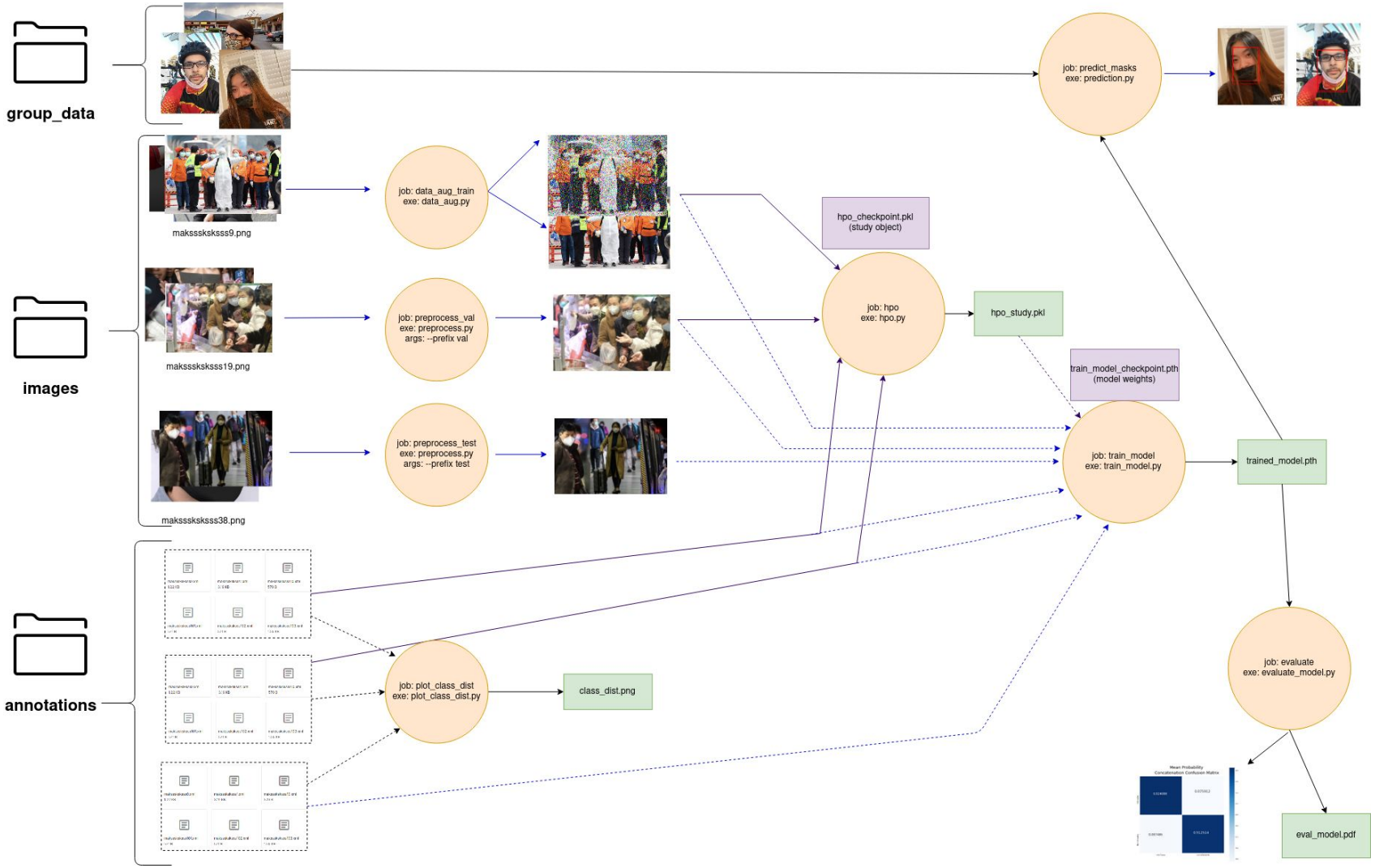
work by
Kelsie Lam





Mask Detection and Classification Workflow

work by
Kelsie Lam

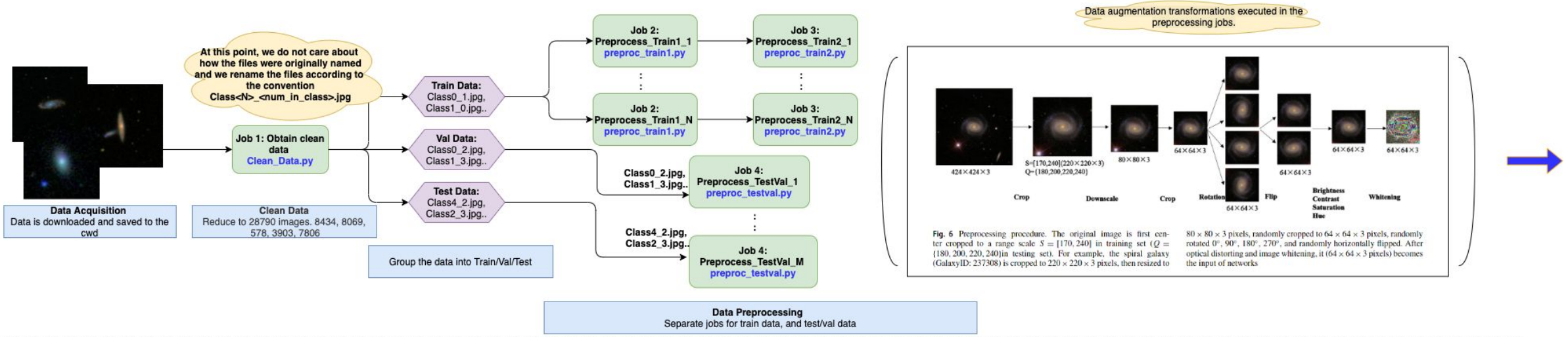


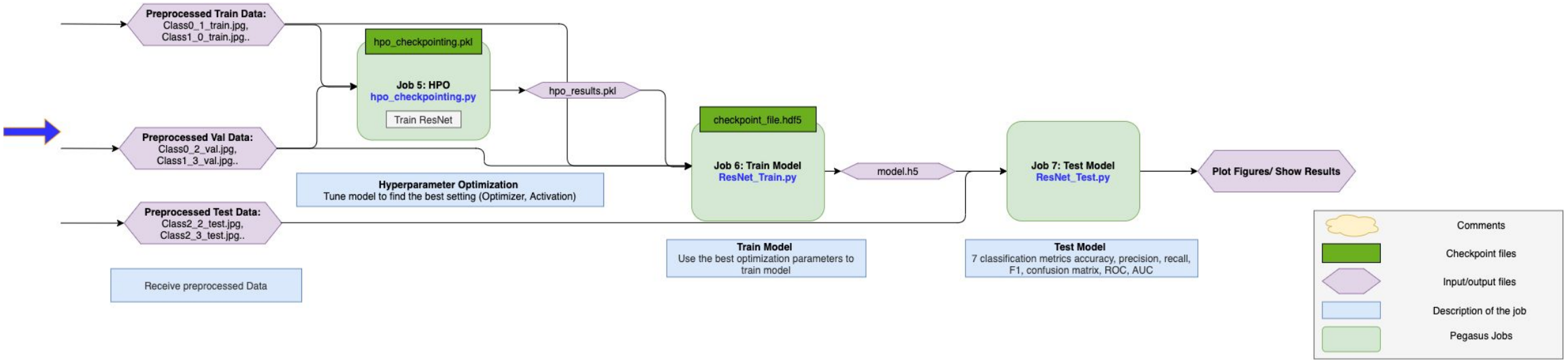
Galaxy Classification Workflow

work by
Srujana Subramanya

Galaxy morphology classification with deep convolutional neural networks

Xiao-Pan Zhu[1,2] · Jia-Ming Dai[1,2] · Chun-Jiang Bian[1] · Yu Chen[1] · Shi Chen[1] · Chen Hu[1,2]



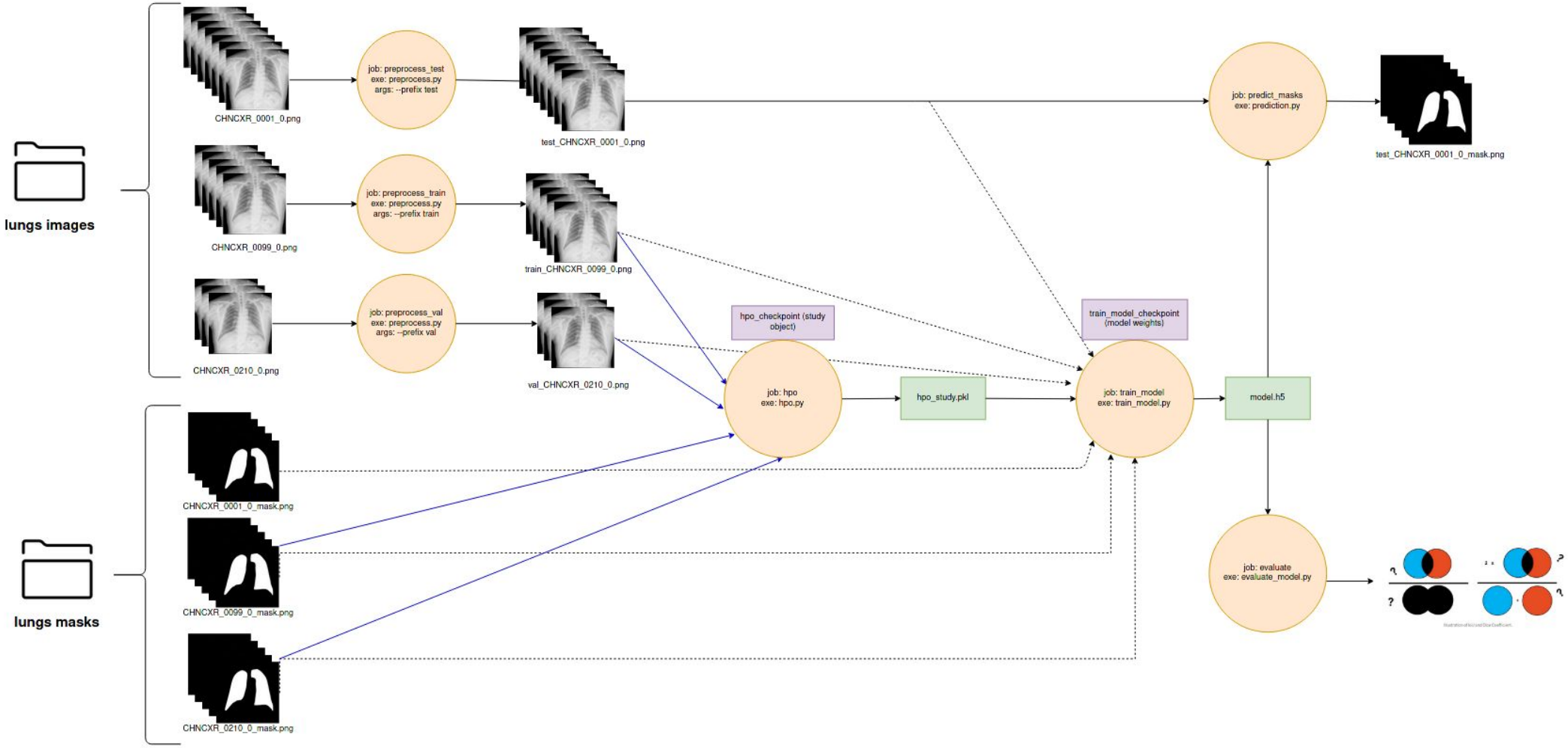


Lung Segmentation Workflow

work continue
by
Aditi Jain

U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox

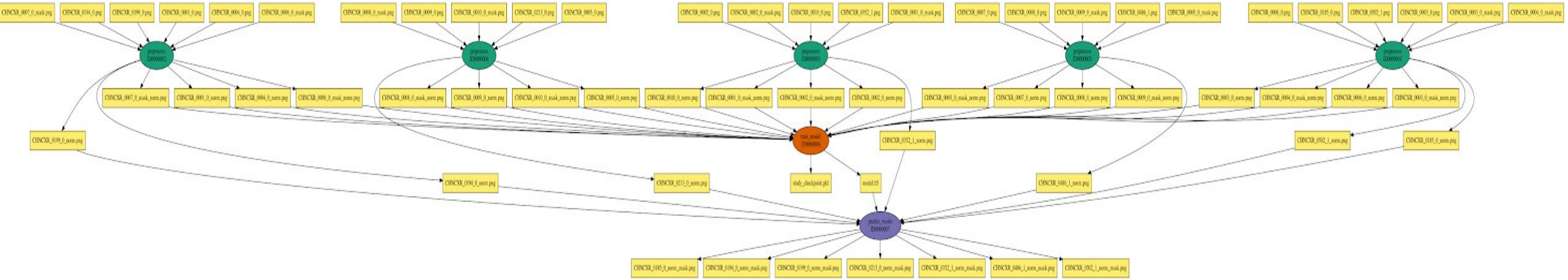


Lung Segmentation Workflow

U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox

work continue by **Aditi Jain**



Thank You!

Pegasus is funded by the National Science Foundation under grant #1664162