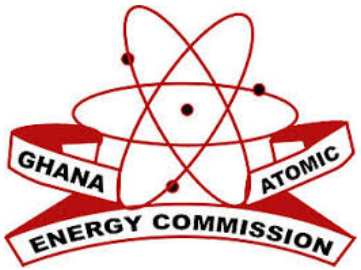
Big Data Techniques in Astronomy

T. Ansah-Narh (Ph.D)

t.narh@gaecgh.org



February 19, 2025

Overview

Introduction

Introduction

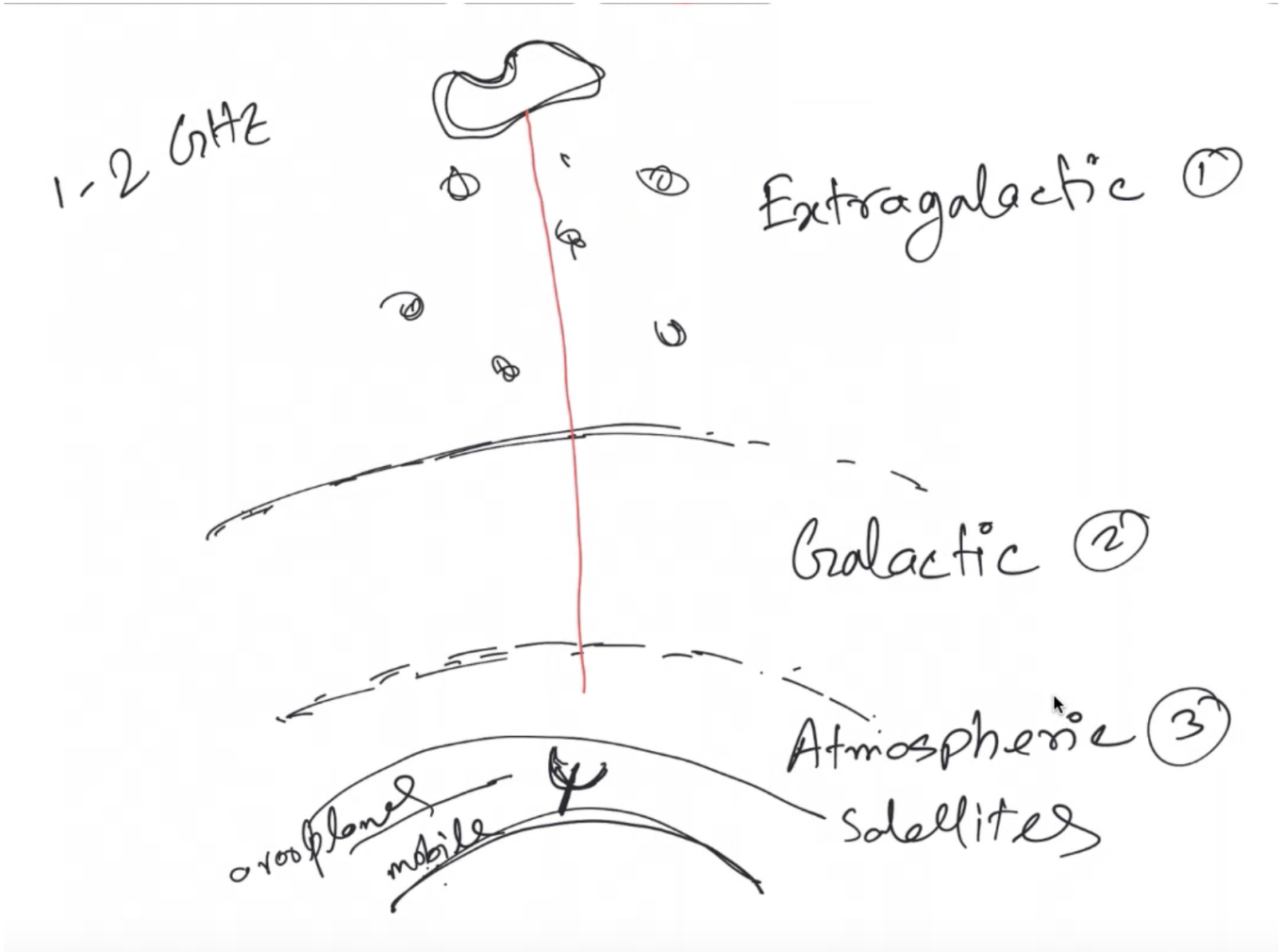


Figure 2: Red line showing the region astronomers observe.

Introduction (cont.)

First light image of SDSS in May 27-28, 1998
Image credit: SDSS Collaboration

- Large digital sky surveys are becoming the dominant source of data in astronomy.
- There are more than 100 terabytes of data in major archives, and that amount is growing rapidly.
- A typical sky survey archive has approximately 10 terabytes of image data and a billion detected sources (stars, galaxies, quasars, etc.)

Introduction (cont.)

SKA project
Image credit: SKA newsletter

- The SKA is currently building the largest and most sensitive radio telescope in the world.
- It will be $\approx 50\times$ more sensitive and $\approx 10\,000\times$ faster w.r.t survey speed than the current instruments.
- The telescope will be able to detect radio signals billions of light years[a] away from us.
- This will help researchers to understand how the universe evolved, and how stars and galaxies form and change.

[a] It's the distance light travels in a year, at a speed of 3×10^5 km/s.

Introduction (cont.)

(a) GRAO

(b) GRAO HPC system

Figure 3: How Ghana Radio Astronomy Observatory (GRAO) works.

Introduction (cont.)

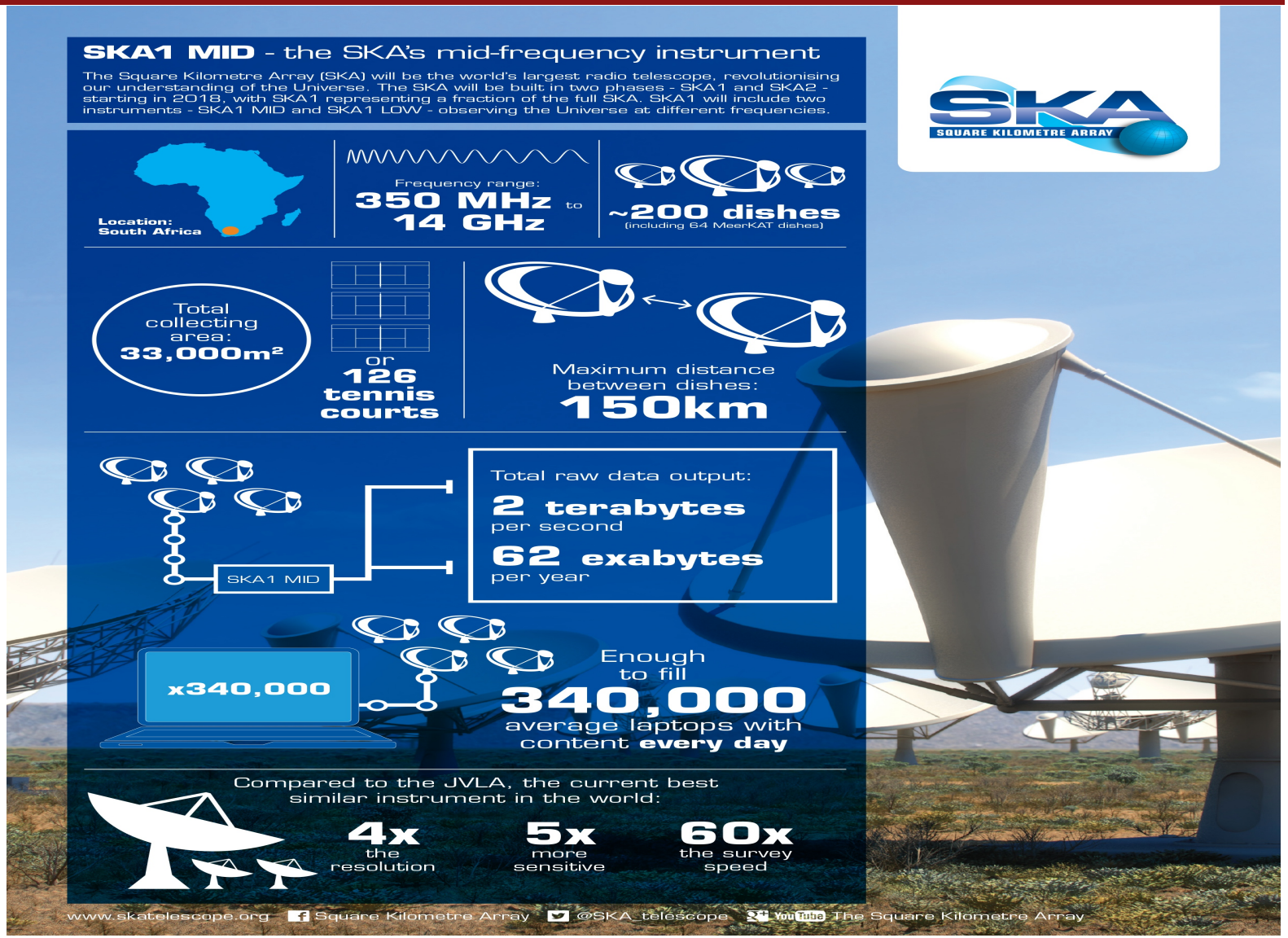


Image source: SKA Precursors – MeerKAT and KAT-7.

Introduction (cont.)

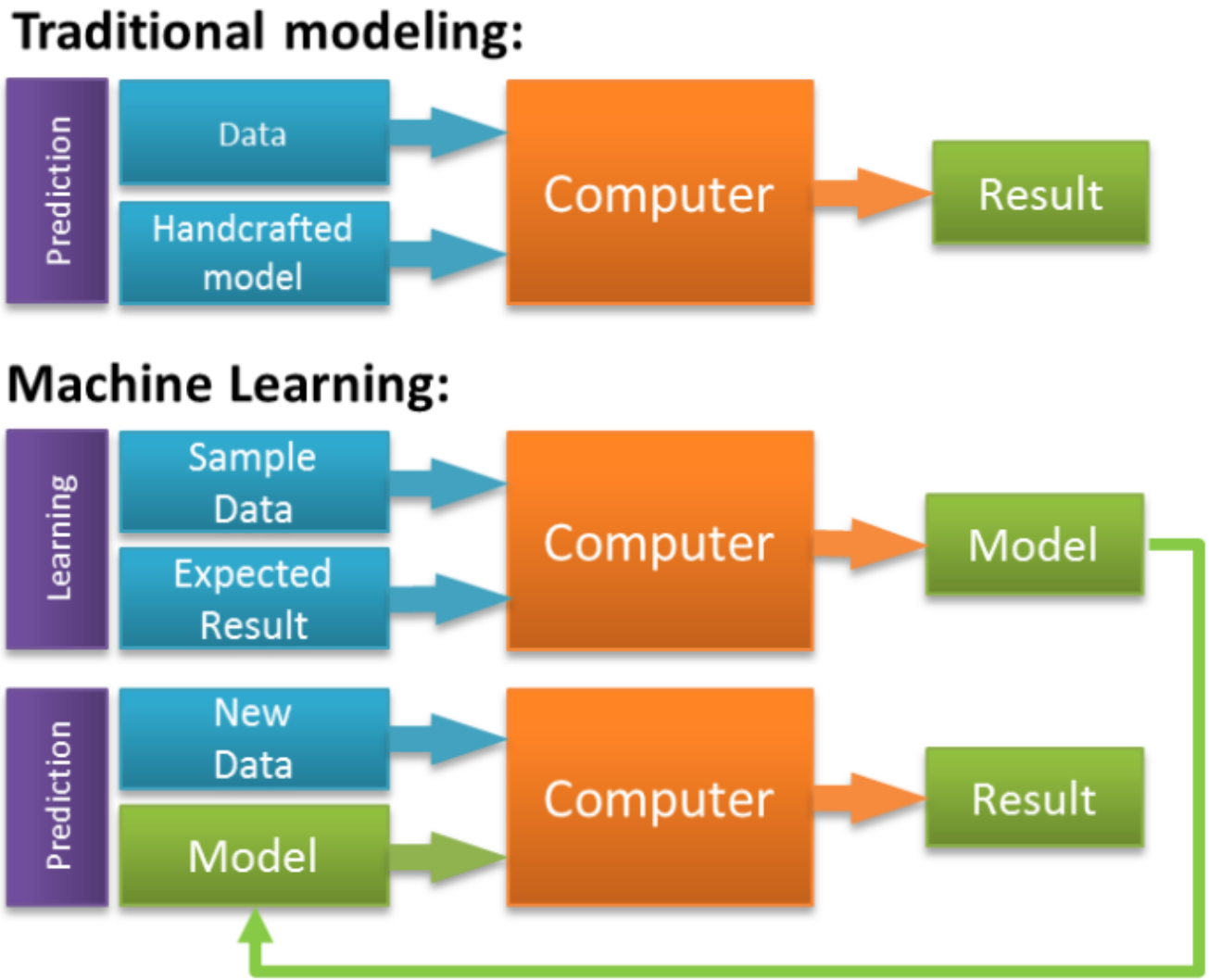
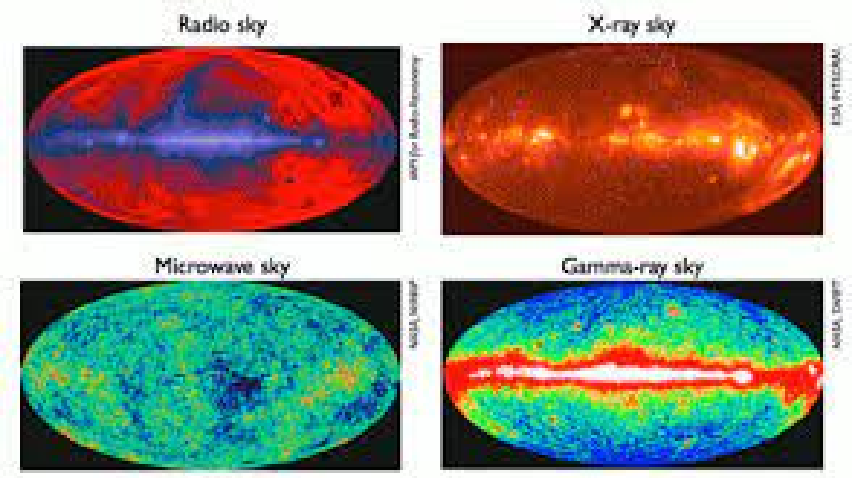


Figure 5: Model learning difference. Image source: Kassel, 2017

Big Data Projects in GSSTI

Intensity mapping techniques



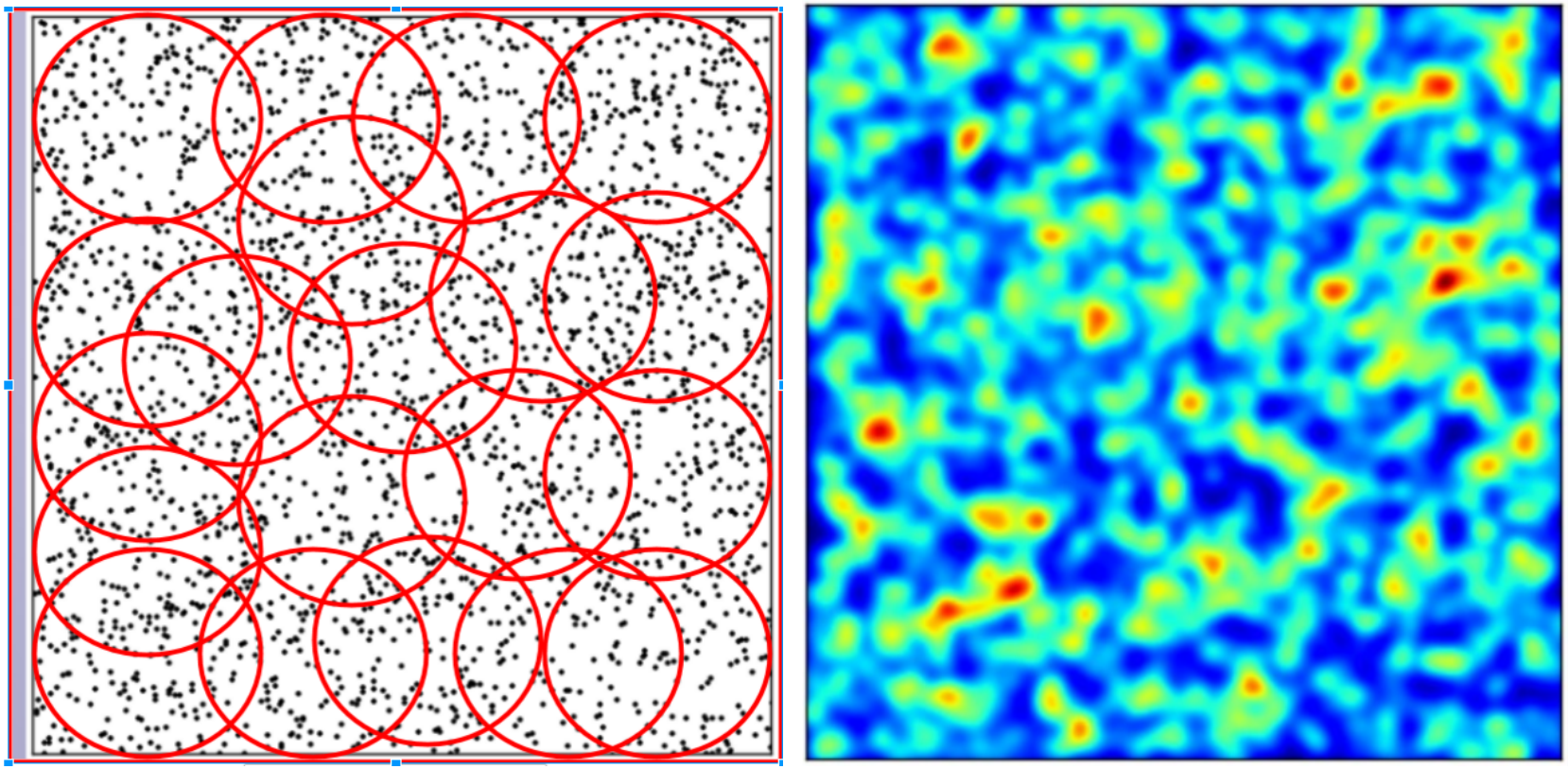
Different windows to the Universe
Image credit: Indigo website

- Motivation behind observational cosmology:
  - understand the large- scale structure of the Universe.
  - its evolution with time.
- This involves observing the entire sky using massive instruments.
- In the past, astronomers use optical telescopes to observe millions of individual galaxies, determine their redshifts and use the information to estimate the energy distribution for each.

Intensity mapping techniques (cont.)

- This approach takes a lot of time and resources, as it is very difficult to detect sufficient numbers of faint sources.
- These optical surveys have been able to map only approximately a **1%** spatio-temporal region of the Universe.

Intensity mapping techniques (cont.)



(a) (b)

Figure 6: Simulated variations in the 21 cm emission brightness temperature. The red rings in (a) show how to perform IM experiment by observing multiple patches in the sky with the radio telescope in order to measure the 21 cm emission to produce (b).

Feature Extraction

- Consider a wavefront denoted by $\Phi_w(\rho, \theta)$, in polar coordinates (ρ, θ), to be a linear combination of Zernike polynomials over a circular unit, then this phenomenon can mathematically be expressed as

$$\Phi_w(\rho, \theta) = \sum_{\beta,\alpha}^{M} C_\beta^\alpha Z_\beta^\alpha(\rho, \theta) \tag{1}$$

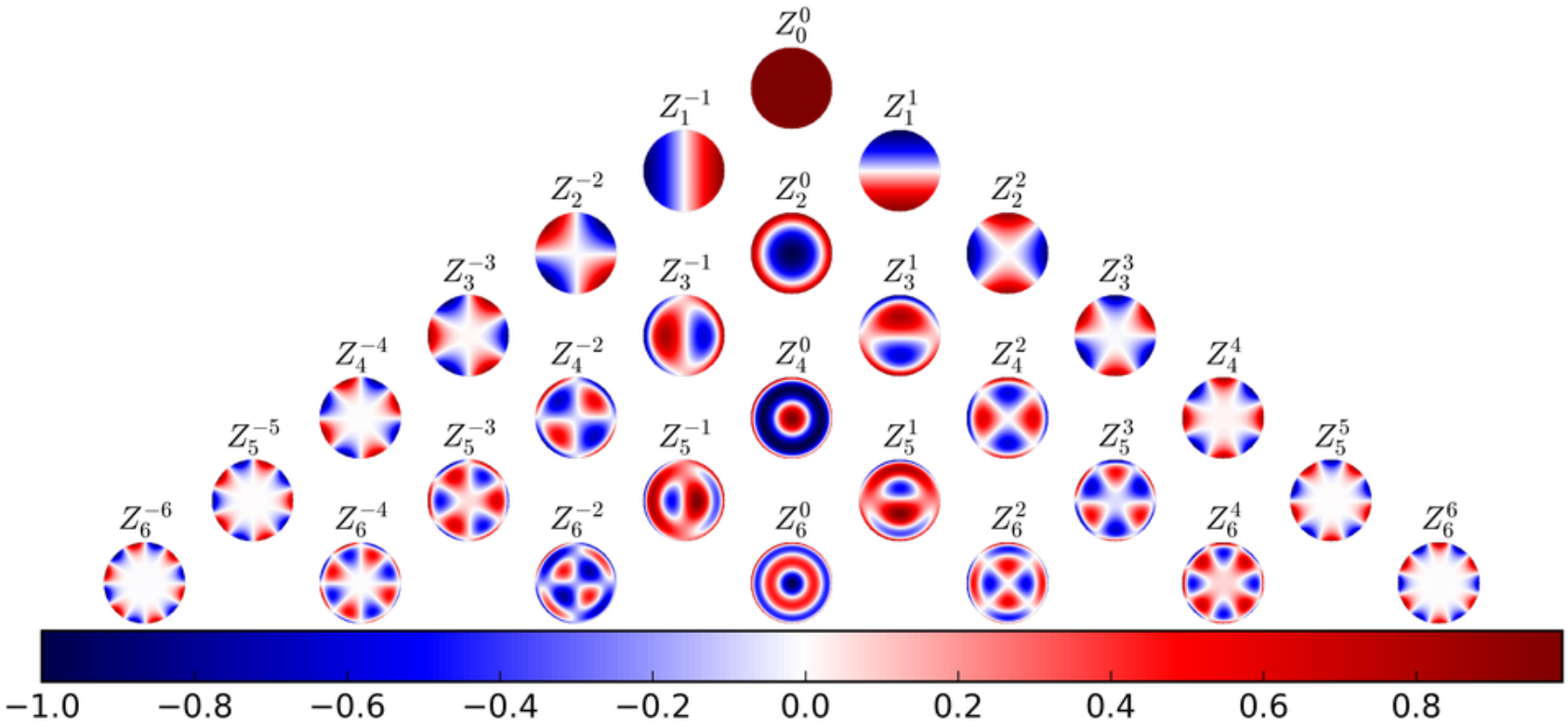
- where the basis function or ZM $Z_\beta^\alpha(\rho, \theta)$ is defined as

$$Z_\beta^\alpha(\rho, \theta) = \begin{cases} \Lambda_\beta^\alpha R_\beta^{|\alpha|}(\rho)\cos(\rho\alpha\theta), & \alpha \geq 0 \\ -\Lambda_\beta^\alpha R_\beta^{|\alpha|}(\rho)\sin(\rho\alpha\theta), & \alpha < 0 \end{cases} \tag{2}$$

Feature Extraction (cont.)

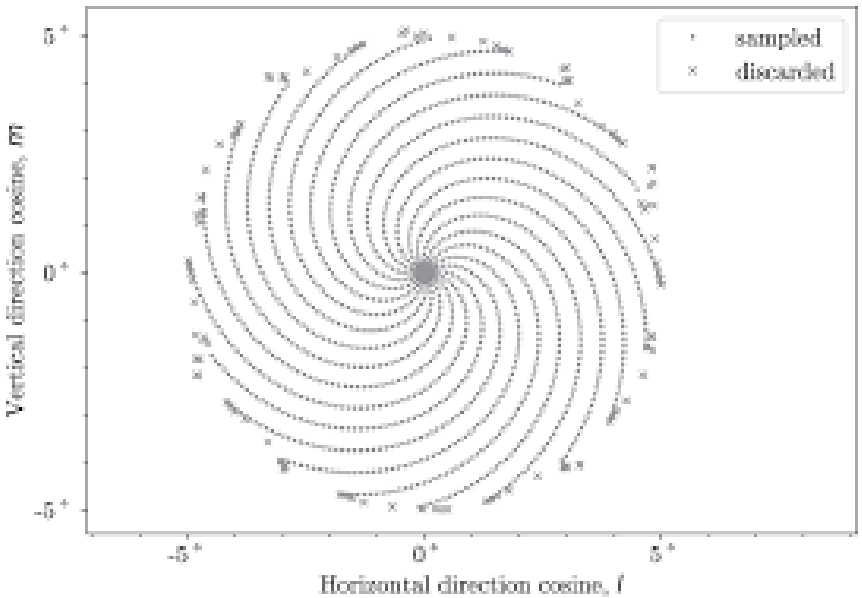
$$\begin{pmatrix} Z_1(\rho_1,\theta_1) & Z_2(\rho_1,\theta_1) & \cdots & Z_P(\rho_1,\theta_1) \\ Z_1(\rho_2,\theta_2) & Z_2(\rho_2,\theta_2) & \cdots & Z_P(\rho_2,\theta_2) \\ \vdots & \vdots & \ddots & \vdots \\ Z_1(\rho_L,\theta_L) & Z_2(\rho_L,\theta_L) & \cdots & Z_P(\rho_L,\theta_L) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_P \end{pmatrix} = \begin{pmatrix} \Phi(\rho_1,\theta_1) \\ \Phi(\rho_2,\theta_2) \\ \vdots \\ \Phi(\rho_L,\theta_L) \end{pmatrix} \tag{3}$$

Feature Extraction (cont.)



Representation of basis patterns of Zernike moments $Z_\beta^\alpha(\rho, \theta)$ of order 6, plotted on a unit circle.

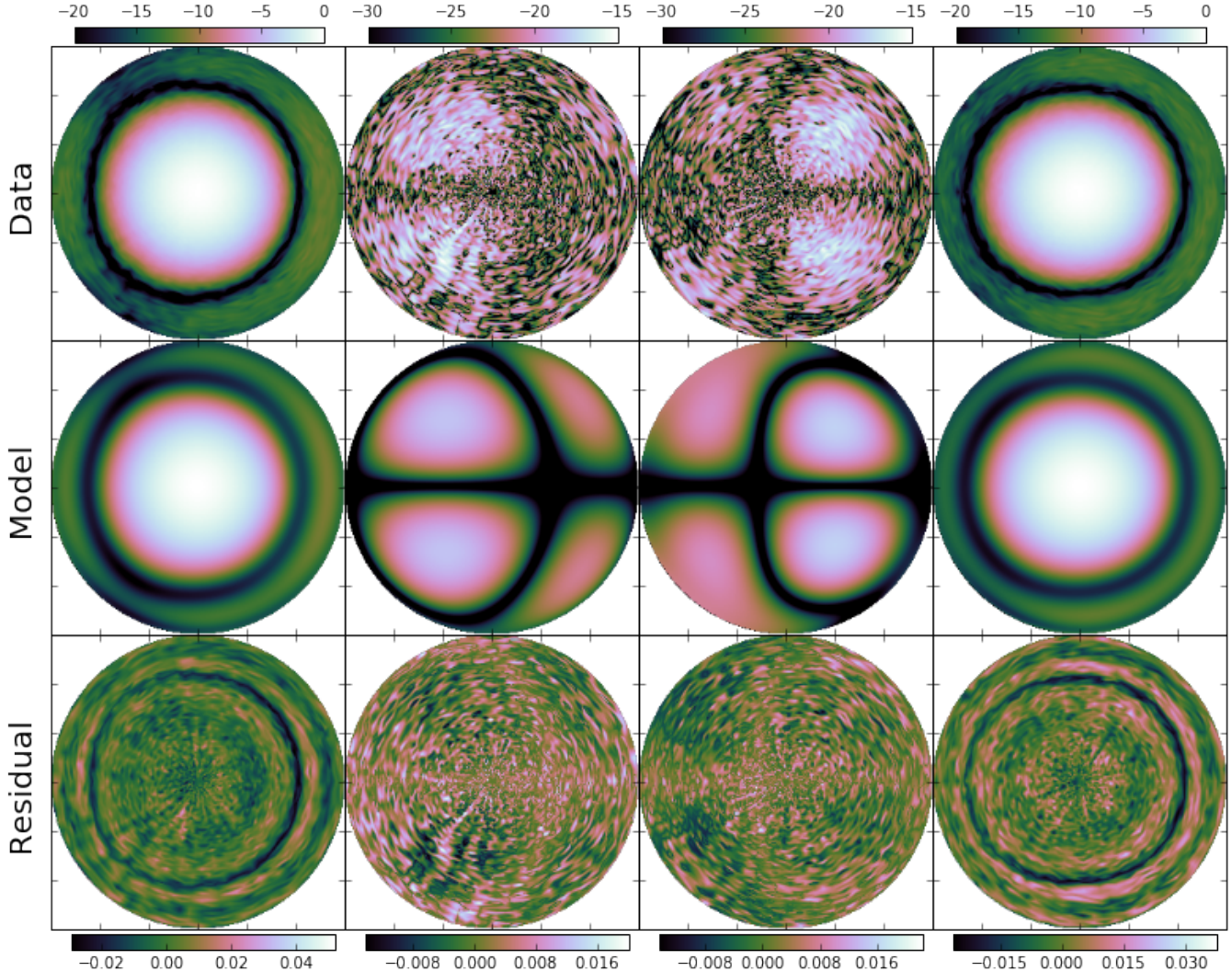
Denoising Holographic Measurements Using ZP



A typical spiral scan pattern observed in the L band showing sampled and discarded slew points.
Image credit: de Villiers & Cotton, 2022

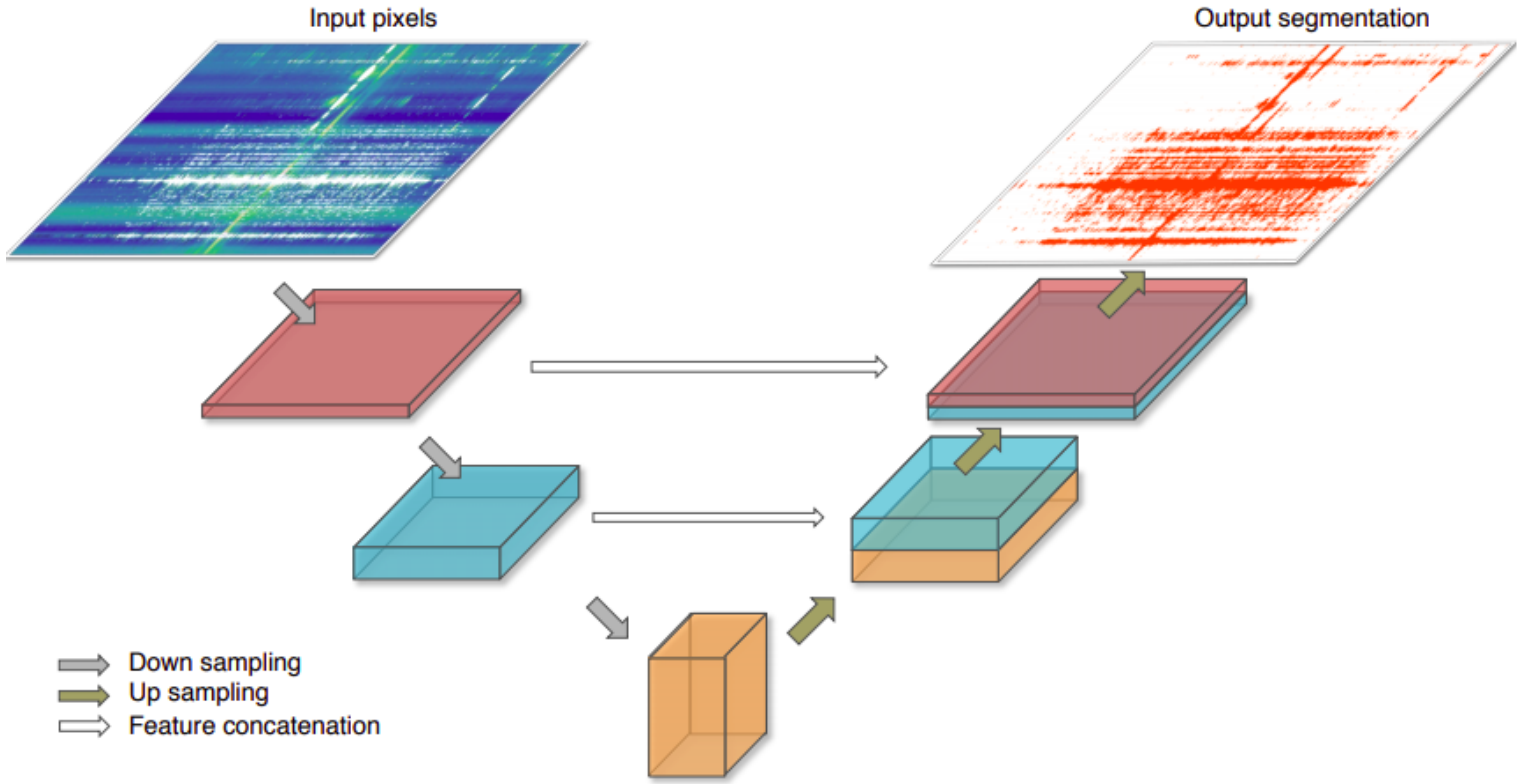
- For holography observation, one antenna constantly tracks a distant radio target while another scans across the target, in a raster pattern.
- The scanning antenna frequently slews back to the target to capture data used to calibrate out gain and phase fluctuations
- A spiral scan pattern is preferred because it reduces the required observation time, has a circular rather than rectangular footprint, and offers variable control over the degree of sampling uniformity across the beam.

Denoising Holographic Measurements Using ZP (cont.)



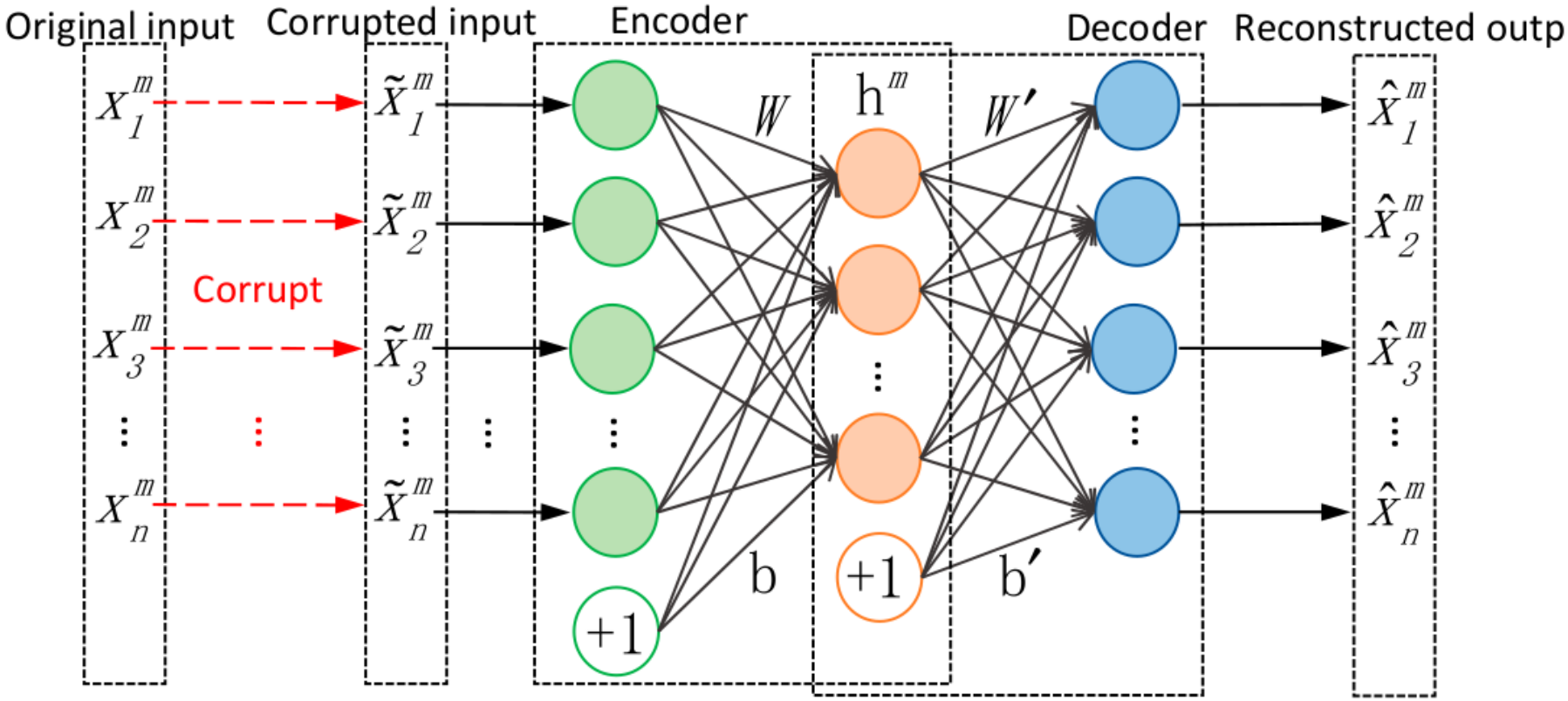
Reconstructing MeerKAT holography beams

RFI Mitigation



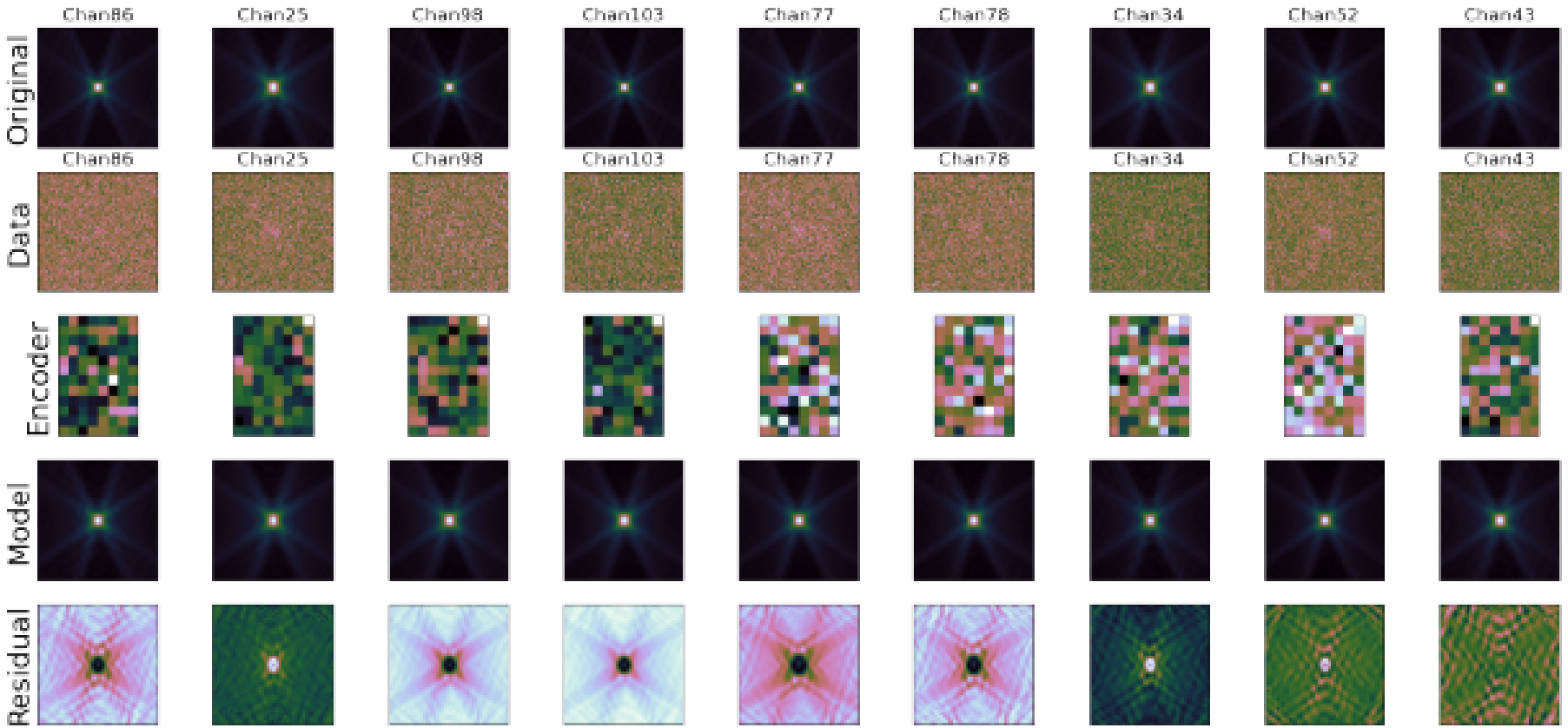
Conceptional architecture of the U-Net with a layer depth of 3.
Source: Akeret et al. 2016.

RFI Mitigation (cont.)



An illustration of the architecture of a de-noising Auto-encoder.
Source: X Liu et al. 2019

RFI Mitigation (cont.)



An illustration of the architecture of a de-noising Auto-encoder. *Source*: X Liu et al. 2019

Recurrent Neural Network (RNN)

1. RNN is a type of ANN that has a recurring connection to itself.
2. Consider a periodic formula: $S_t = R_w(S_{t-1}, X_t)$
   - X_t – Input at time t
   - S_t – Current state at time t
   - S_{t-1} – Initial state at time $t-1$
   - R_w – Periodic function
3. Simple case: $S_t = \tanh(W_s S_{t-1} + W_x X_t)$

Recurrent Neural Network (RNN) (cont.)

- Fig. 12 shows how recurring connection helps RNN to learn the effect of previous input $x(t-1)$ along with the current input $x(t)$ while predicting the output at time t, $S(t)$.

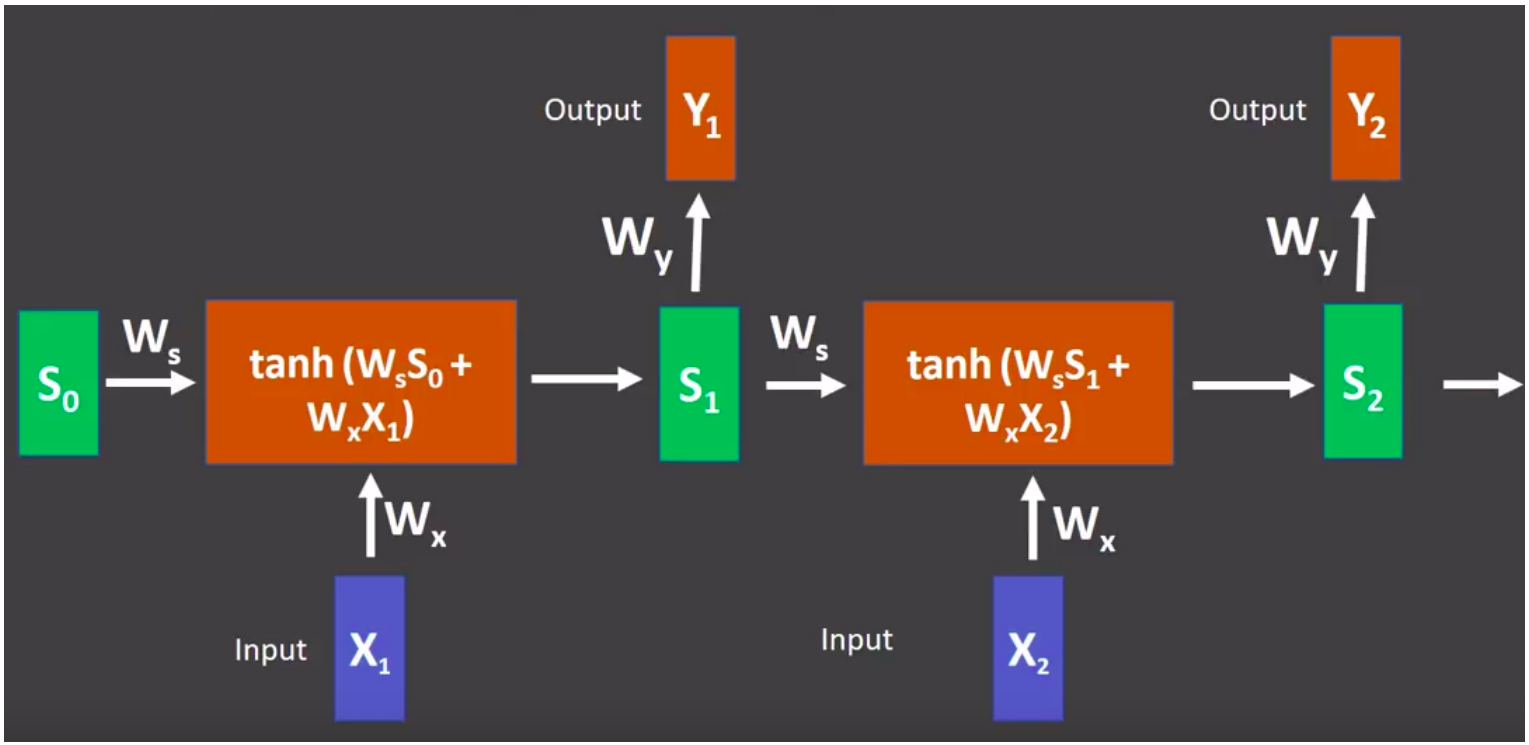


Figure 12: Basic RNN (unfolded). *Credit*: The Semicolon RNN lecture, 2018.

Recurrent Neural Network (RNN) (cont.)

- Fig. 13 shows one of the problems with RNN networks which is vanishing gradients, i.e. the gradients vanish to 0 during backpropagation. It arises because the derivative of the activation functions such as *sigmoid*(σ) or tanh are less than 0.25 and 1 respectively.

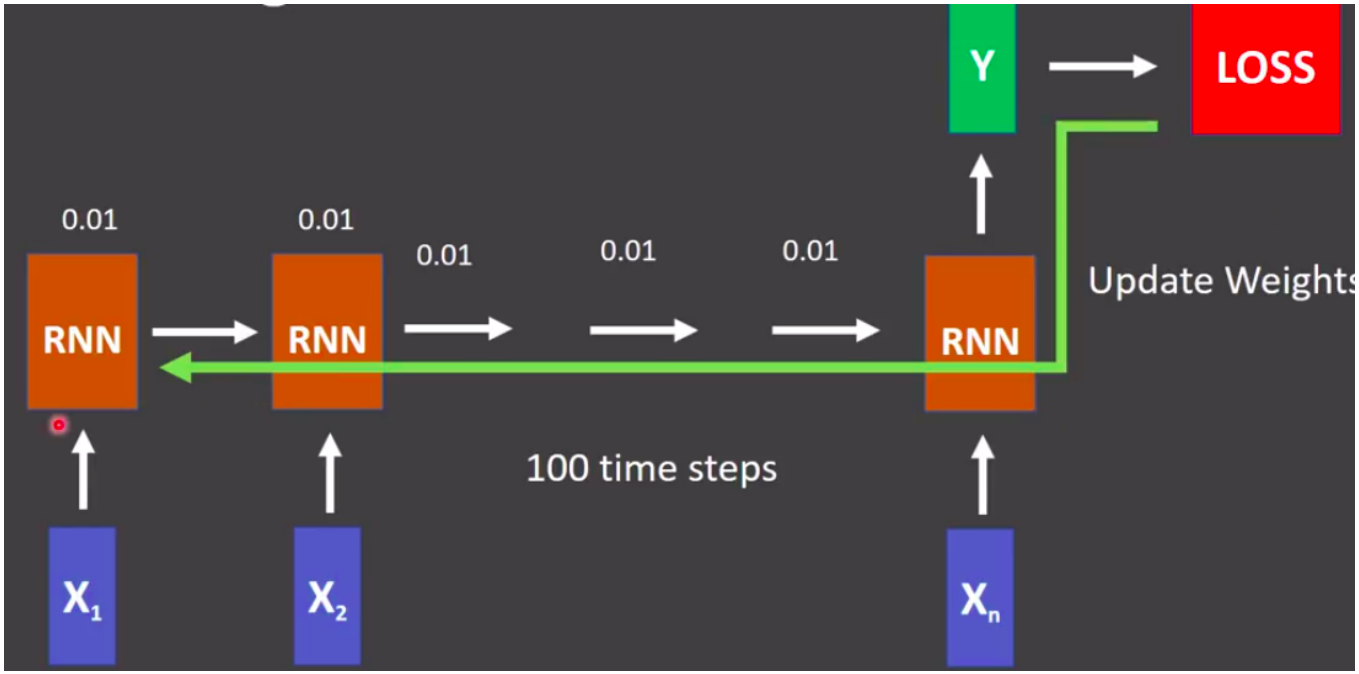
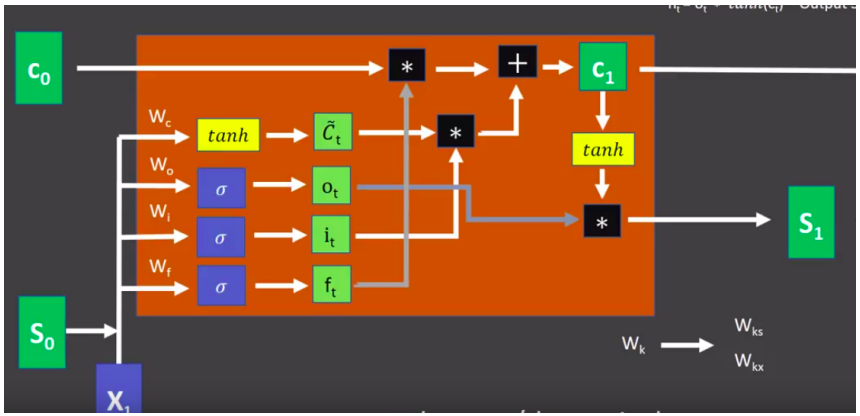


Figure 13: Problem of vanishing gradient. *Credit*: The Semicolon RNN lecture, 2018.

Recurrent Neural Network (RNN) (cont.)

- When many of these derivatives are multiplied together while applying chain rule, the gradients vanish to 0. This causes earlier layers to learn very slowly compared to later layers
- Long Short Term Memory (LSTM) model solves the problem of vanishing gradients by introducing a new state called cell state.

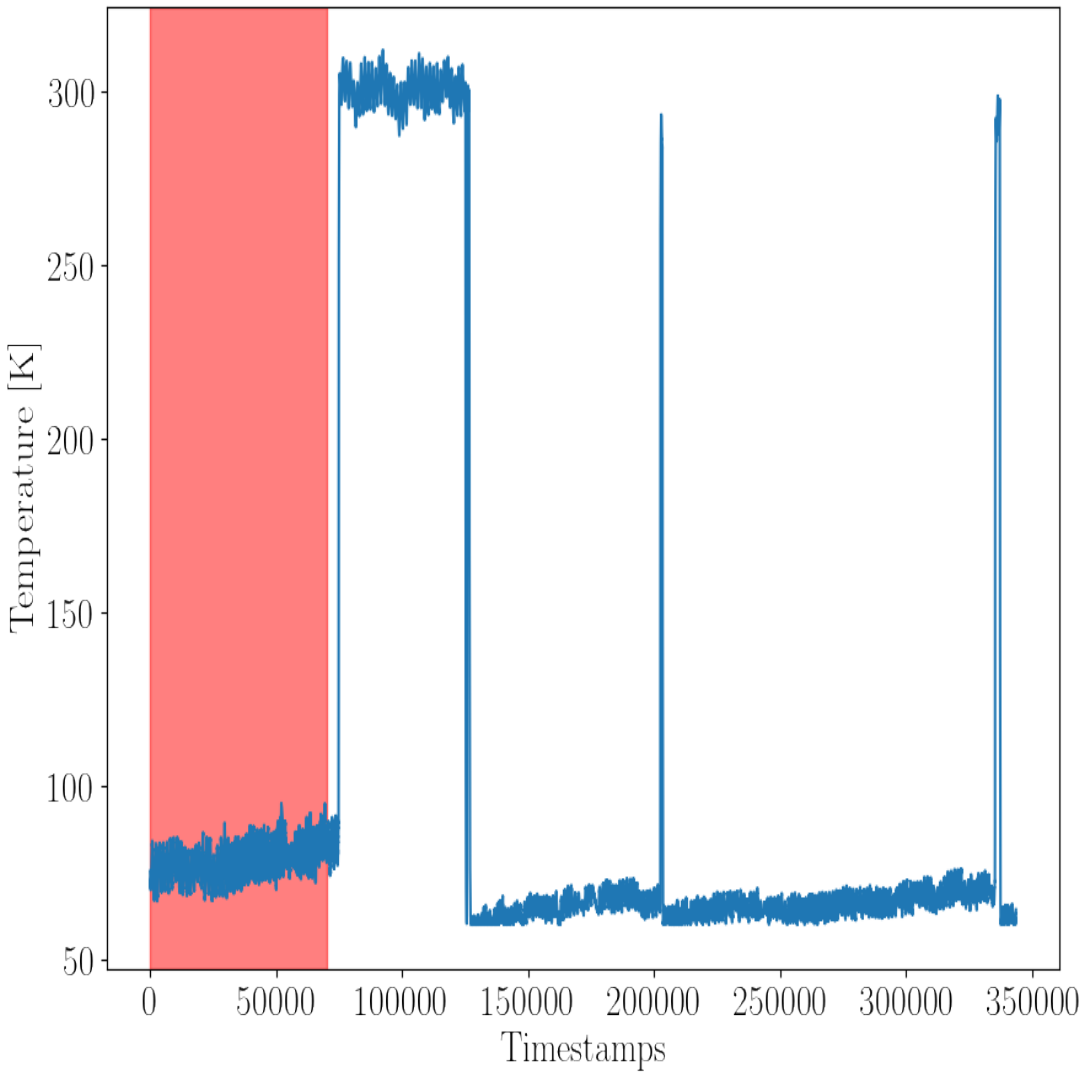
Long Short Term Memory (LSTM)



Credit: The Semicolon RNN lecture, 2018.

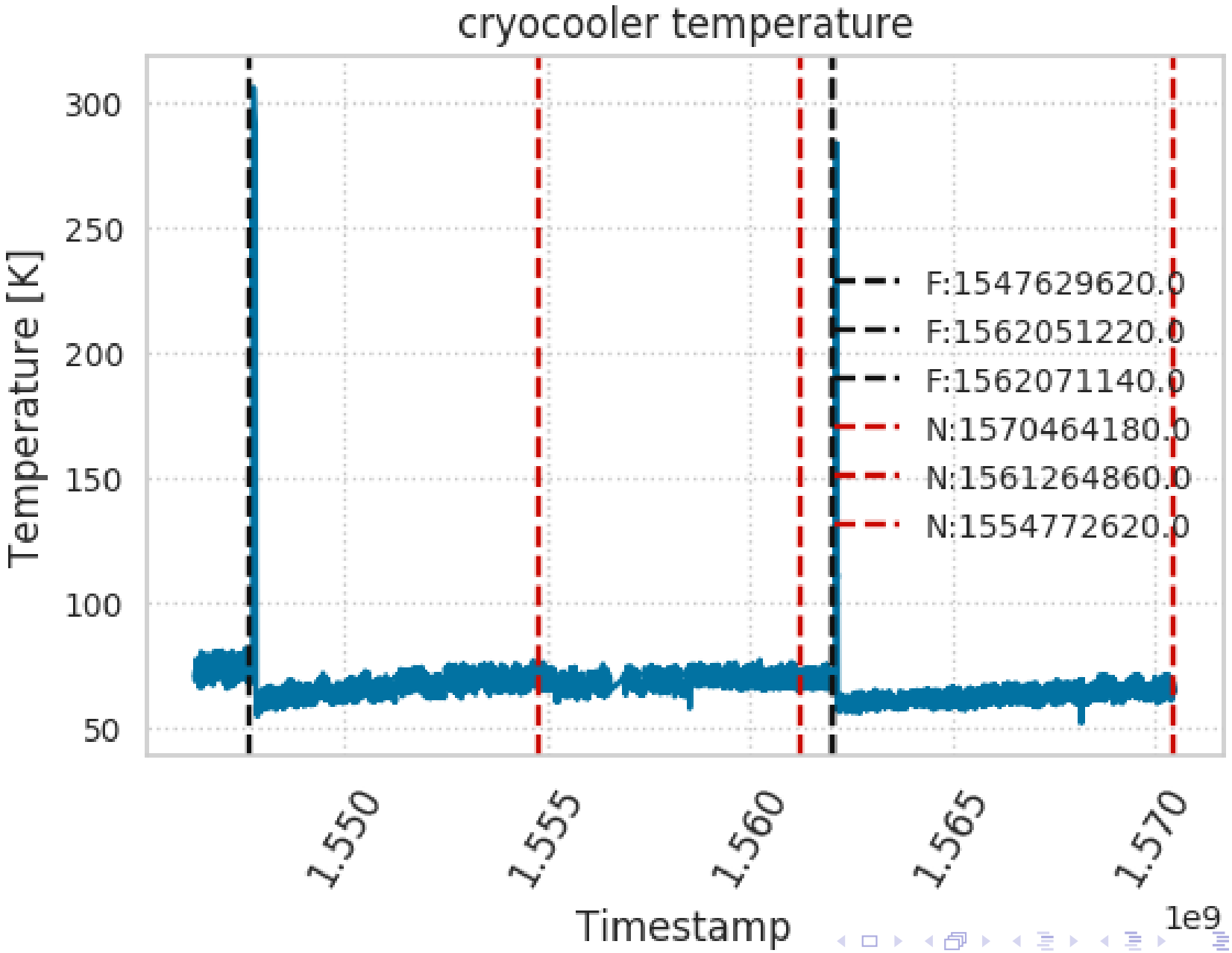
- Input Gate:
  $i_t = \sigma(W_i S_{t-1} + W_i X_i)$
- Forget Gate:
  $f_t = \sigma(W_f S_{t-1} + W_f X_f)$
- Output Gate:
  $o_t = \sigma(W_o S_{t-1} + W_o X_o)$
- Cell State:
  $c_t = (i_t * \tilde{C}_t) + (f_t * c_{t-1})$
  - Intermediate cell state:
    $\tilde{C}_t = \tanh(W_c S_{t-1} + W_c X_c)$

Anomaly detection in Sensor Data



Early failure in temperature reading

Anomaly detection in Sensor Data (cont.)

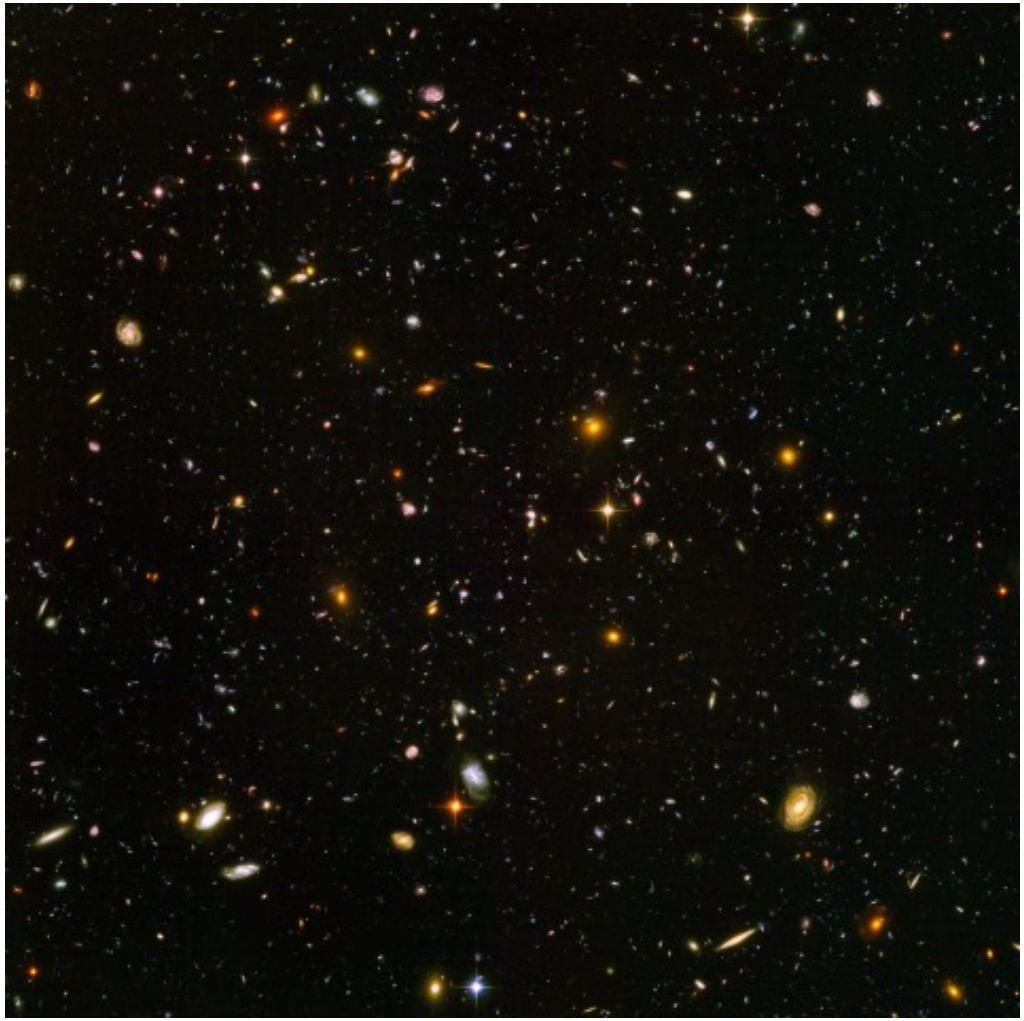


Galaxy Classification

Hubble Space Telescope
Image credit: Hubble site

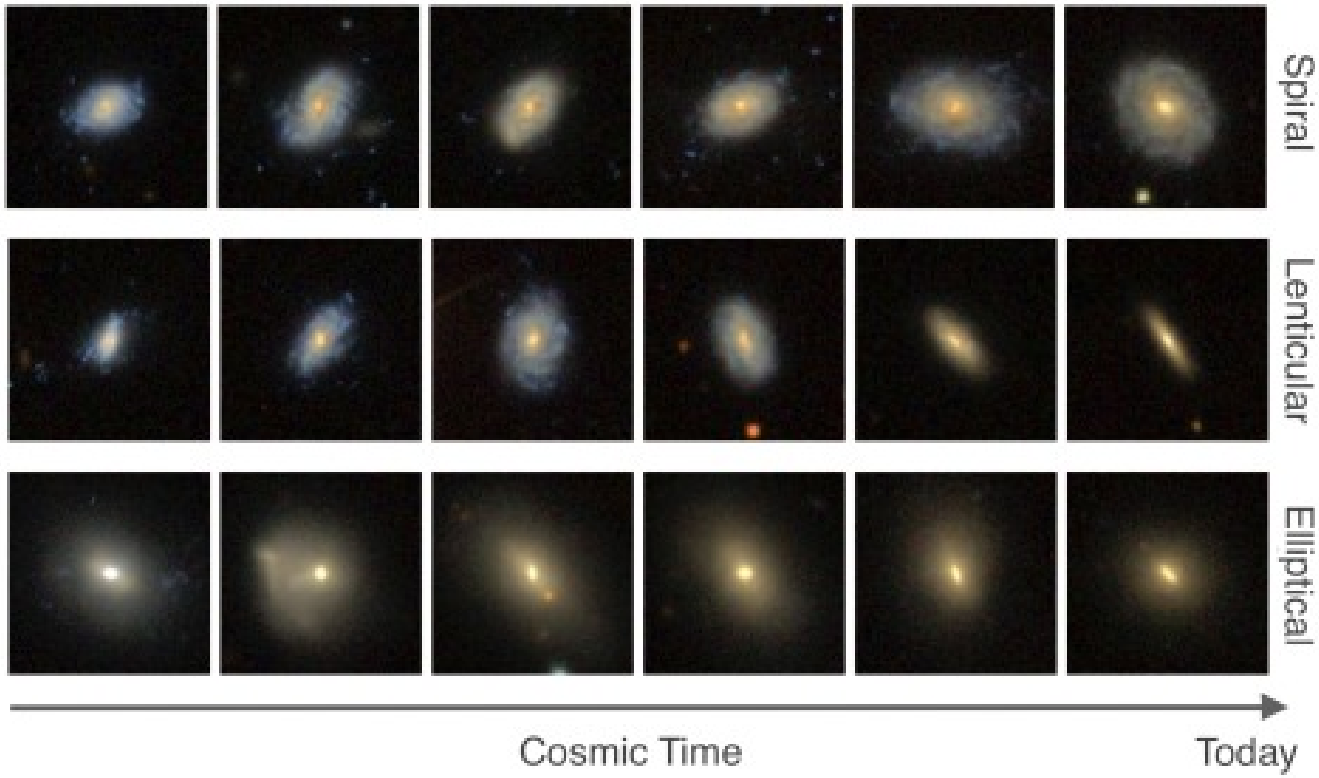
- It is a large space observatory launched in April, 1990.
- It has made more than 1.4 million observations over the course of its lifetime.
- Hubble has a crystal-clear view of the universe. Scientists have used Hubble to observe some of the most distant stars and galaxies yet seen, as well as the planets in our solar system.
- It is helping us better understand the history of the expanding universe.

Galaxy Classification (cont.)



Source: Galaxy Zoo

Galaxy Classification (cont.)



Source: Galaxy Zoo

Galaxy Classification (cont.)

Physics Letters B 795 (2019) 248–258

Contents lists available at ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb

Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey

Asad Khan [a,b,*], E.A. Huerta [a,c], Sibo Wang [a], Robert Gruendl [a,c], Elise Jennings [d], Huihuo Zheng [d]

[a] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[b] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[c] Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[d] Argonne National Laboratory, Leadership Computing Facility, Lemont, IL 60439, USA

ARTICLE INFO

Article history:
Received 6 December 2018
Received in revised form 7 June 2019
Accepted 7 June 2019
Available online 12 June 2019
Editor: H. Peiris

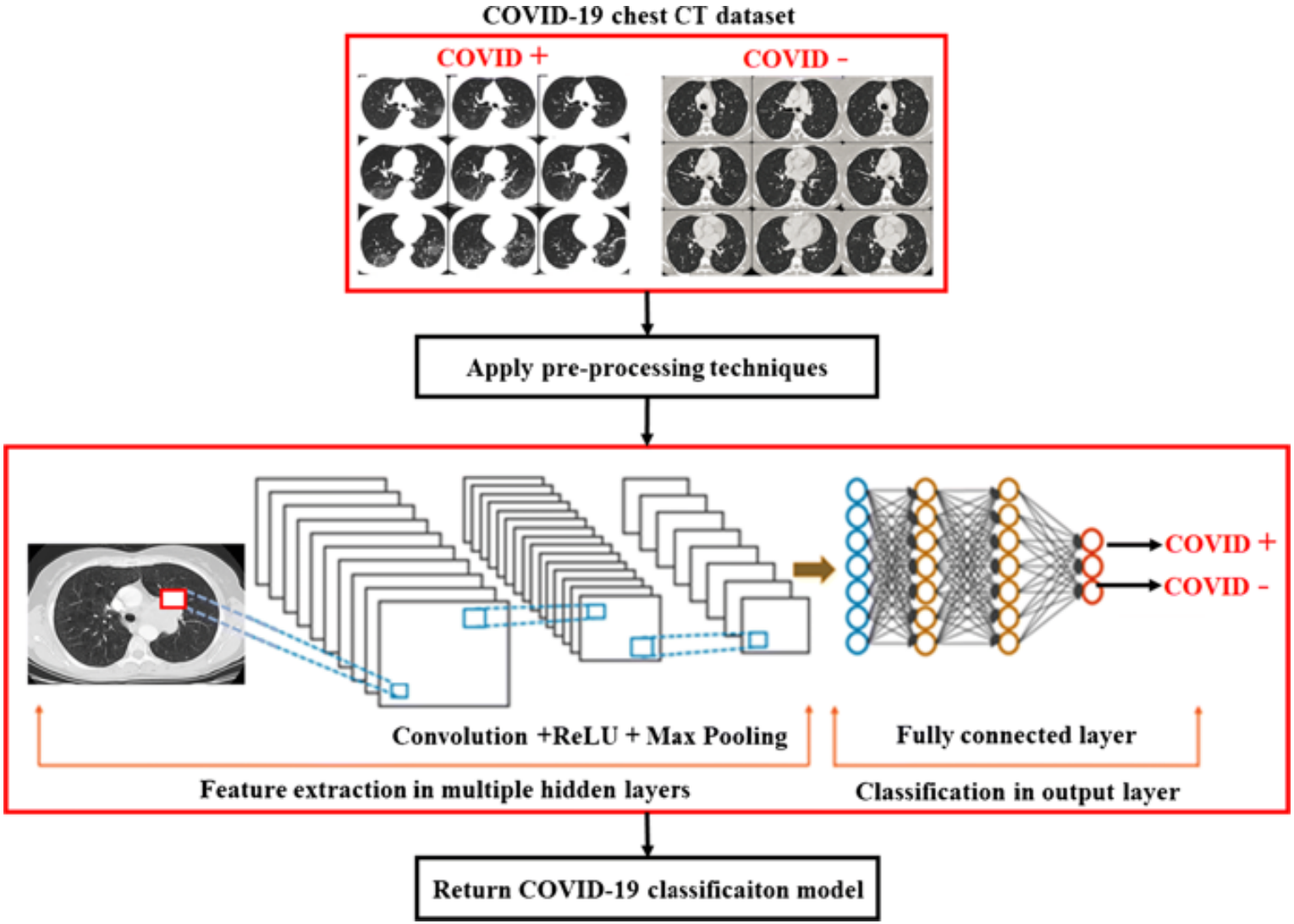
Keywords:
Deep learning
Convolutional neural networks
Sloan Digital Sky Survey
Dark Energy Survey
Large Synoptic Survey Telescope
Unsupervised learning

ABSTRACT

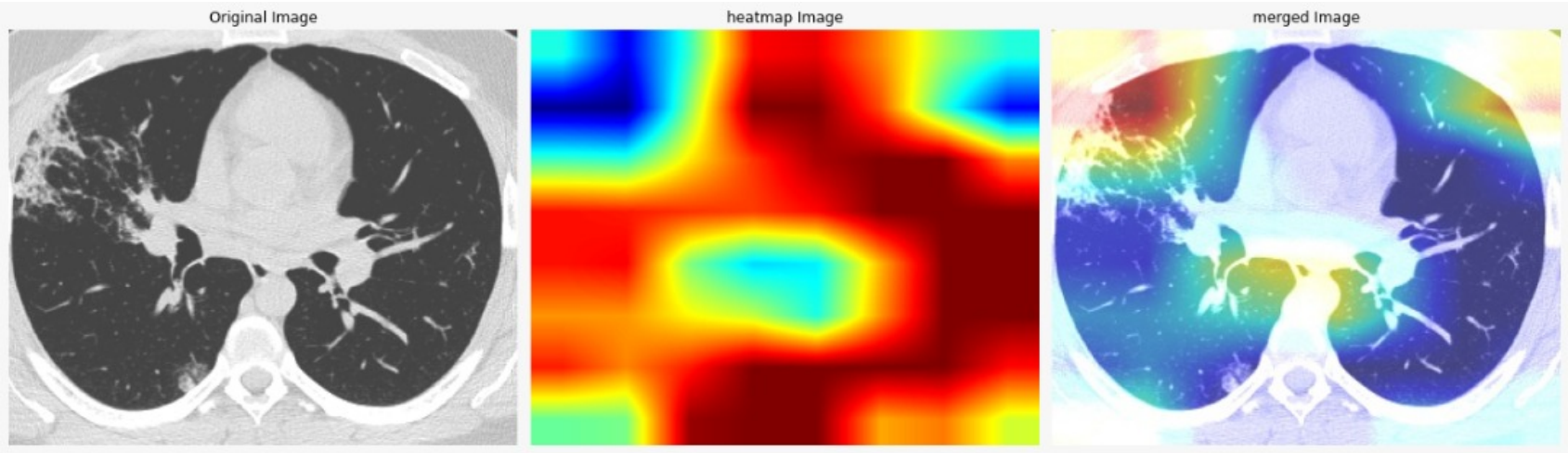
The scale of ongoing and future electromagnetic surveys pose formidable challenges to classify astronomical objects. Pioneering efforts on this front include citizen science campaigns adopted by the Sloan Digital Sky Survey (SDSS). SDSS datasets have been recently used to train neural network models to classify galaxies in the Dark Energy Survey (DES) that overlap the footprint of both surveys. Herein, we demonstrate that knowledge from deep learning algorithms, pre-trained with real-object images, can be transferred to classify galaxies that overlap both SDSS and DES surveys, achieving state-of-the-art accuracy $\gtrsim 99.6\%$. We demonstrate that this process can be completed within just eight minutes using distributed training. While this represents a significant step towards the classification of DES galaxies that overlap previous surveys, we need to initiate the characterization of unlabelled DES galaxies in new regions of parameter space. To accelerate this program, we use our neural network classifier to label over ten thousand unlabelled DES galaxies, which do not overlap previous surveys. Furthermore, we use our neural network model as a feature extractor for unsupervised clustering and find that unlabelled DES images can be grouped together in two distinct galaxy classes based on their morphology, which provides a heuristic check that the learning is successfully transferred to the classification of unlabelled DES images. We conclude by showing that these newly labelled datasets can be combined with unsupervised recursive training to create large-scale DES galaxy catalogs in preparation for the Large Synoptic Survey Telescope era.

Technology Transfer

COVID-19 Detection in CT Scans



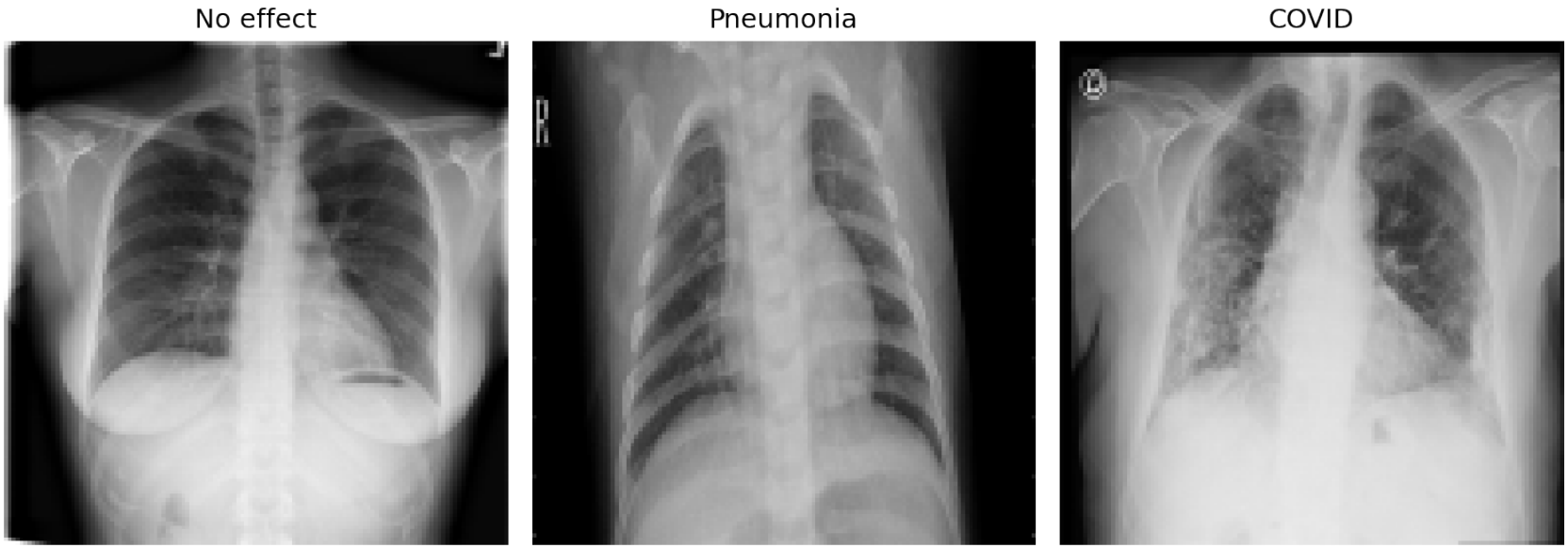
COVID-19 Detection in CT Scans (cont.)



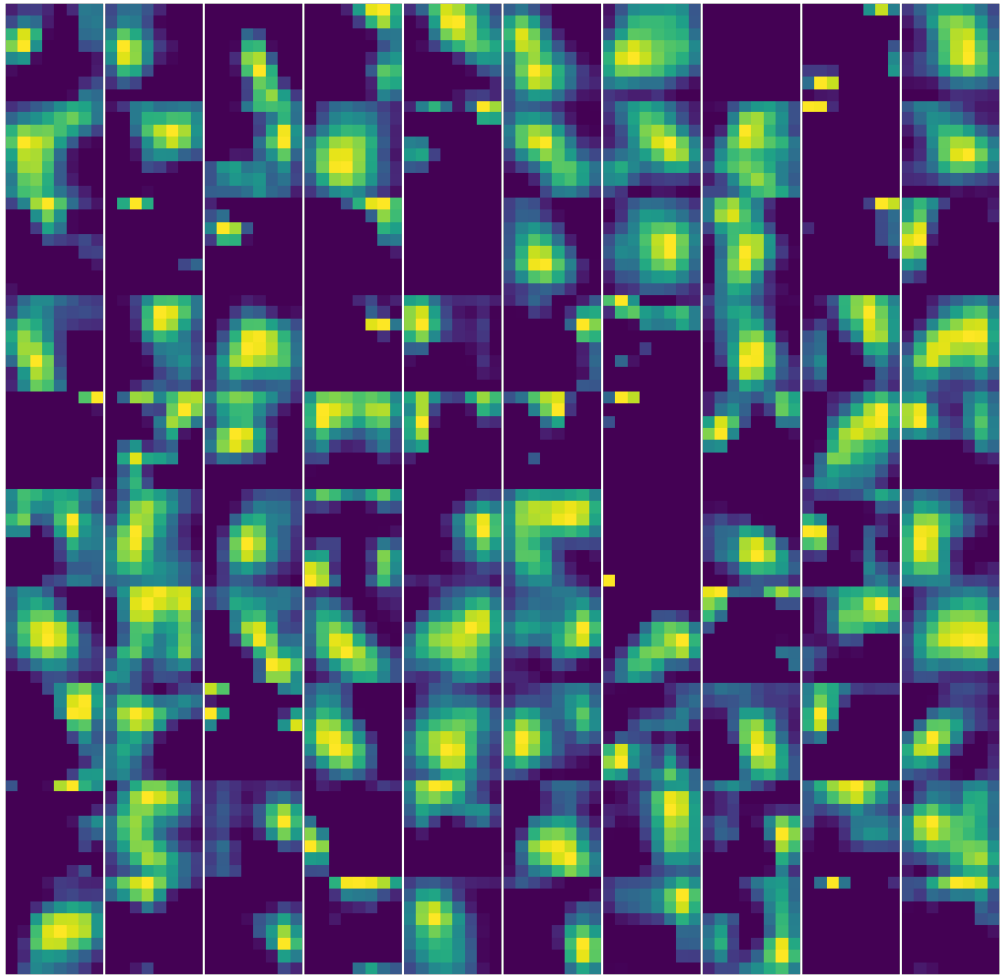
Deep learning approach to detect COVID

Active Learning technique

Active Learning technique (cont.)

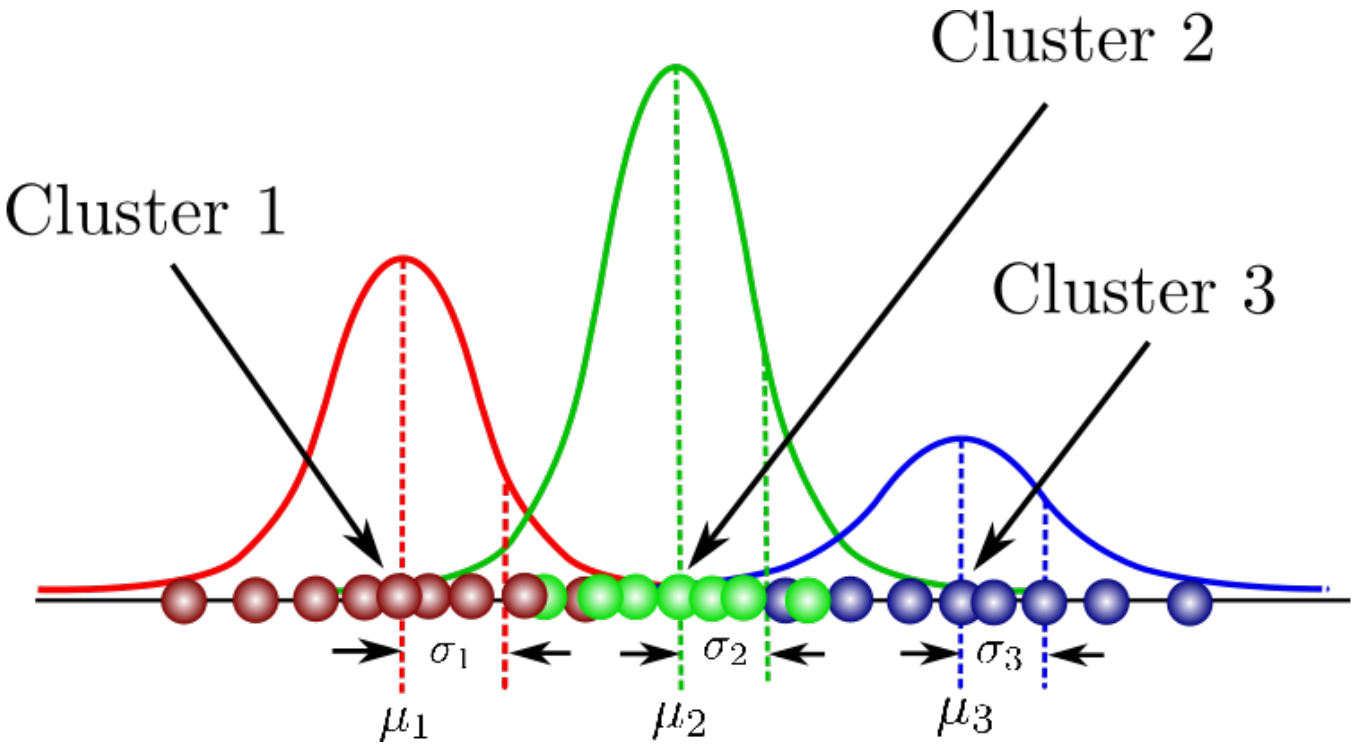


Active Learning technique (cont.)



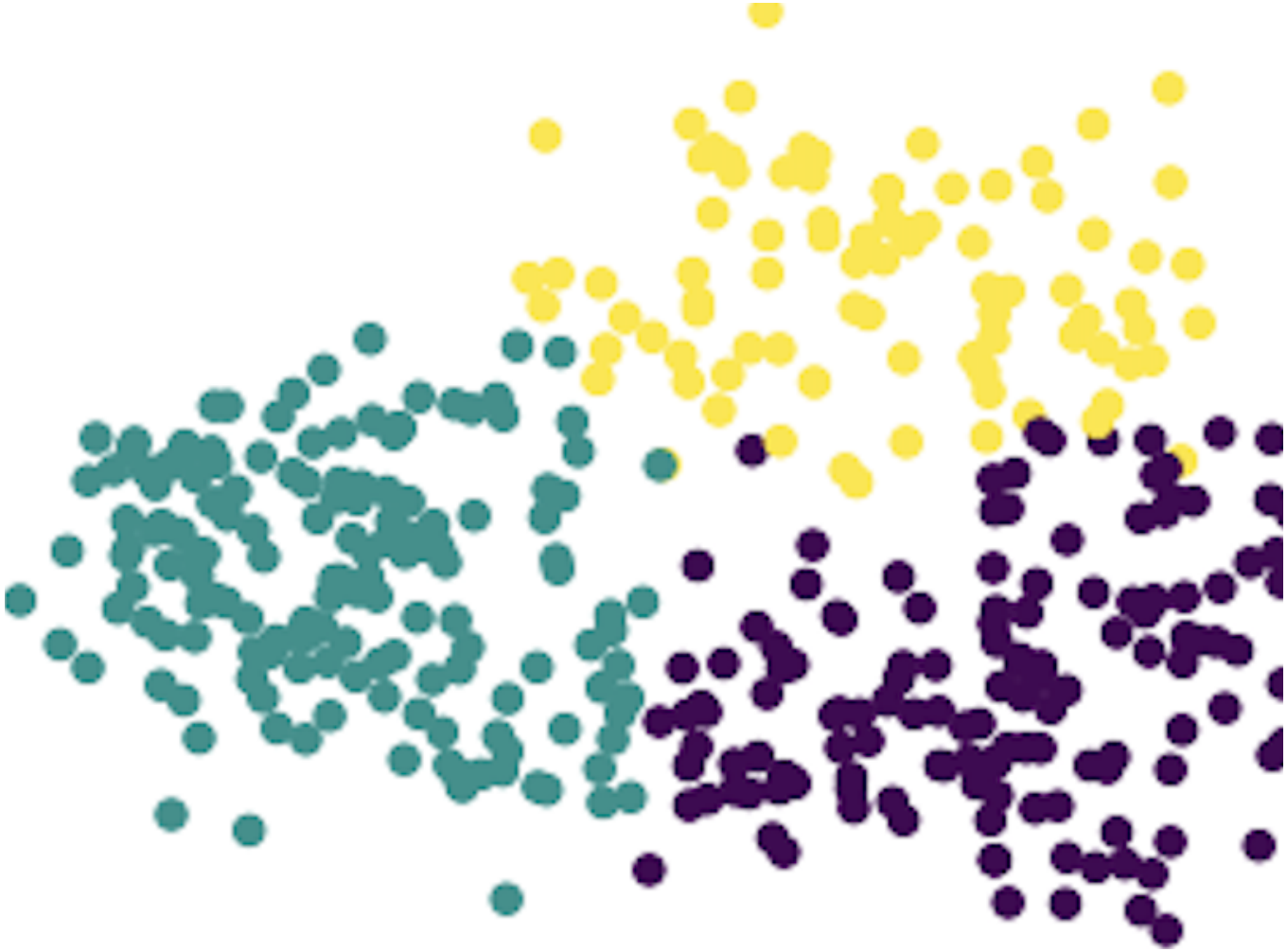
Activation layers

Active Learning technique (cont.)



Gaussian mixture model

Active Learning technique (cont.)

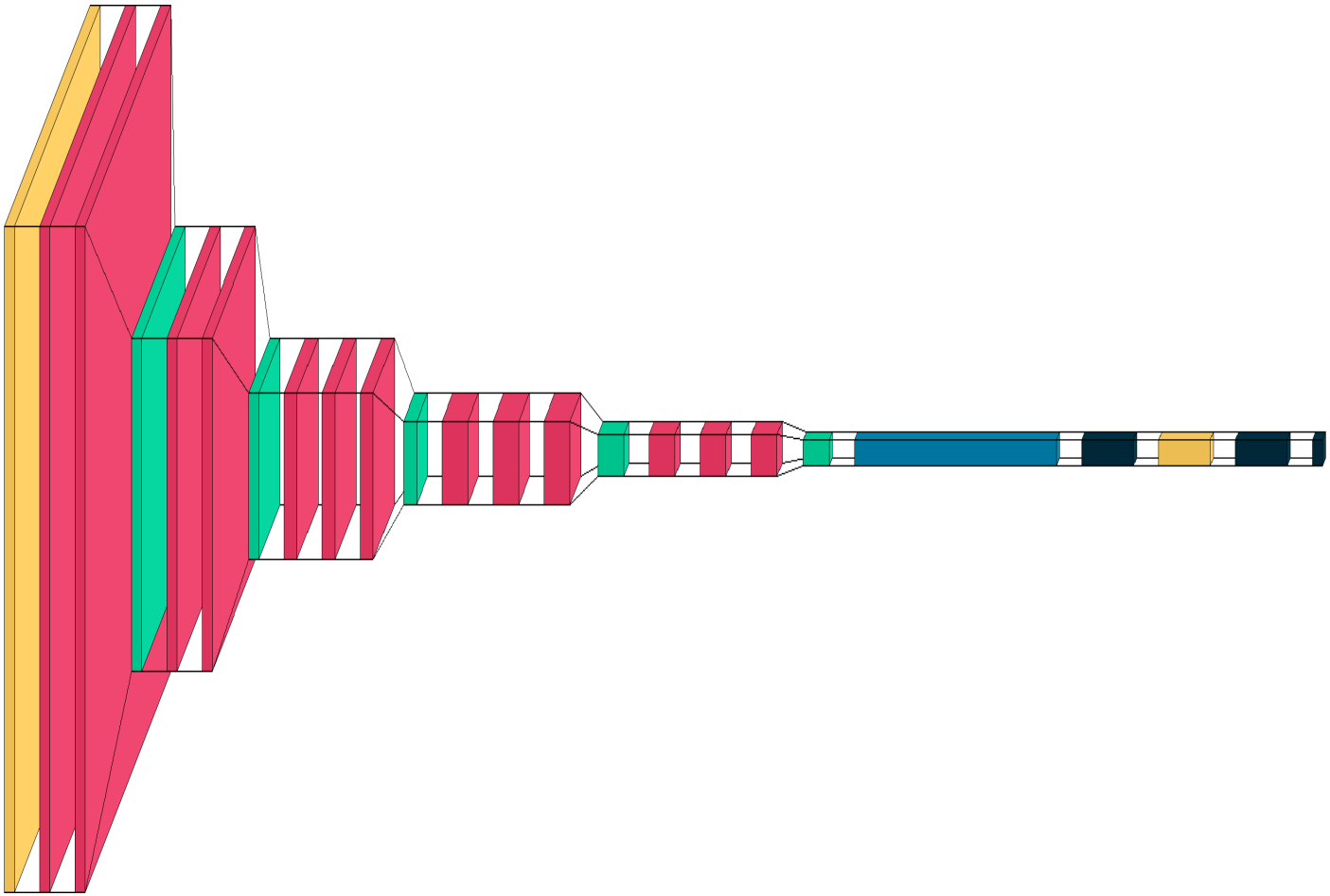


Different components using GMM

Maize leaf disease detection

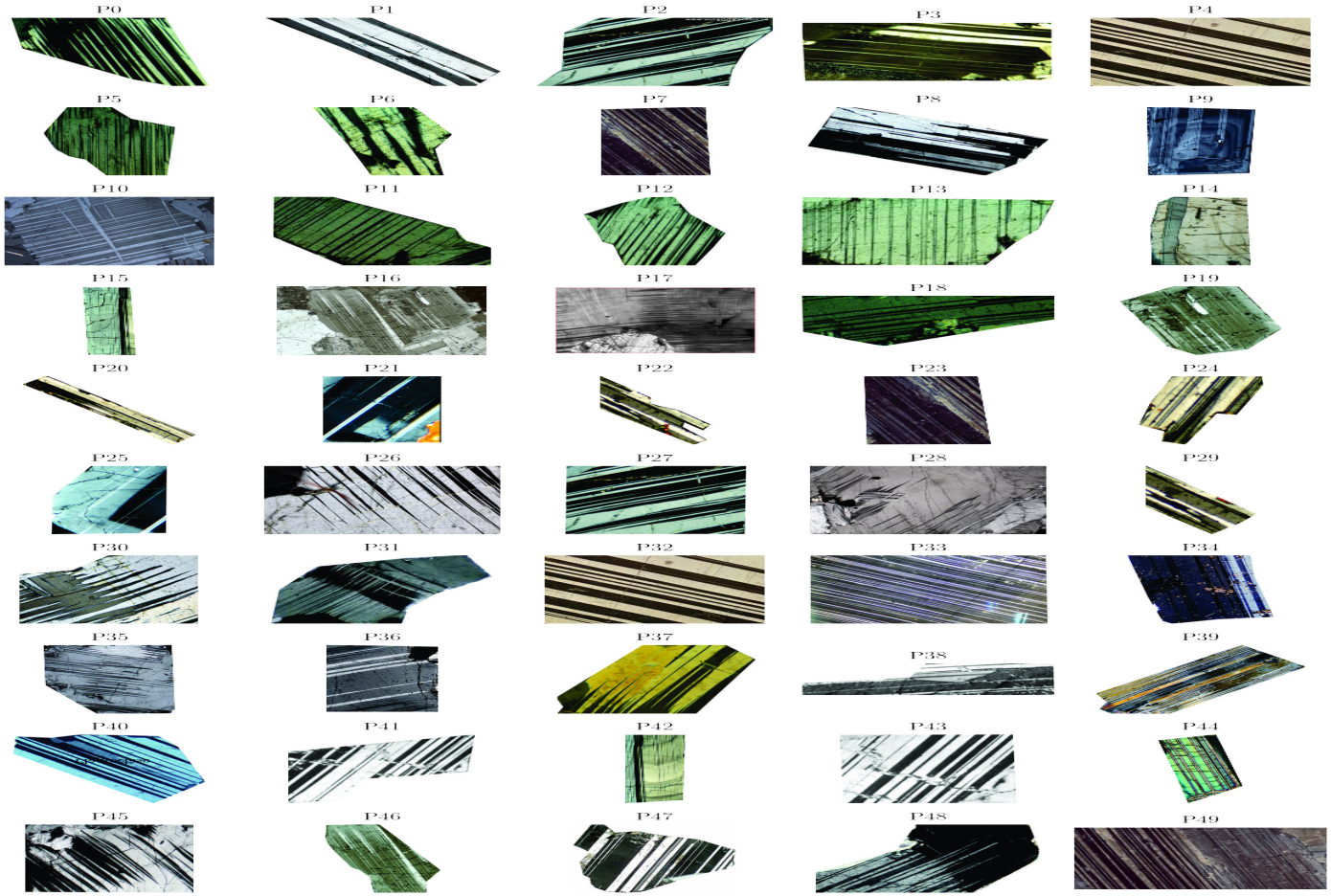
Deep learning approach to detect maize leaf diseases.

Mineral Classification in Petrographic images



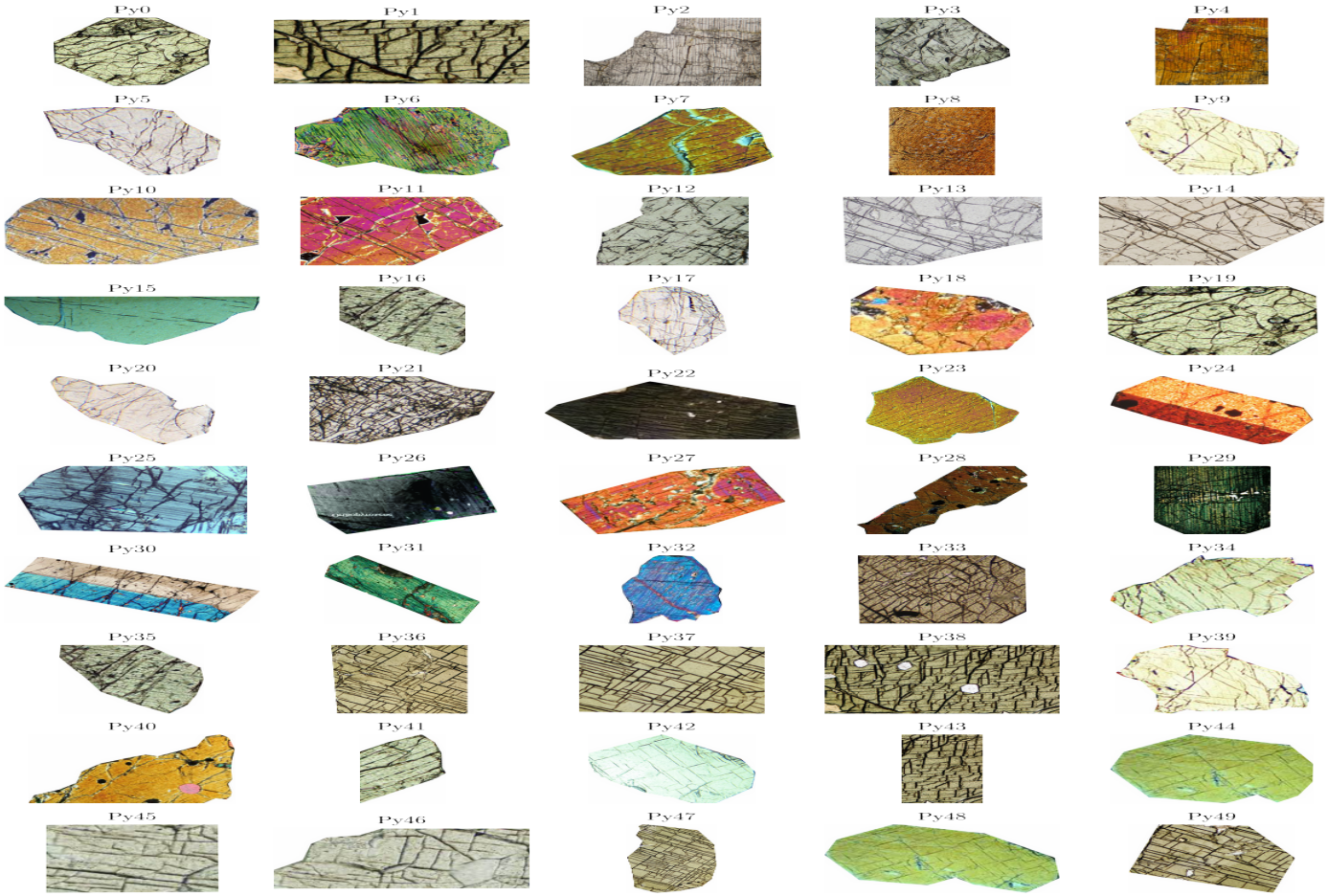
VGG16-Convolutional Neural Network

Mineral Classification in Petrographic images (cont.)



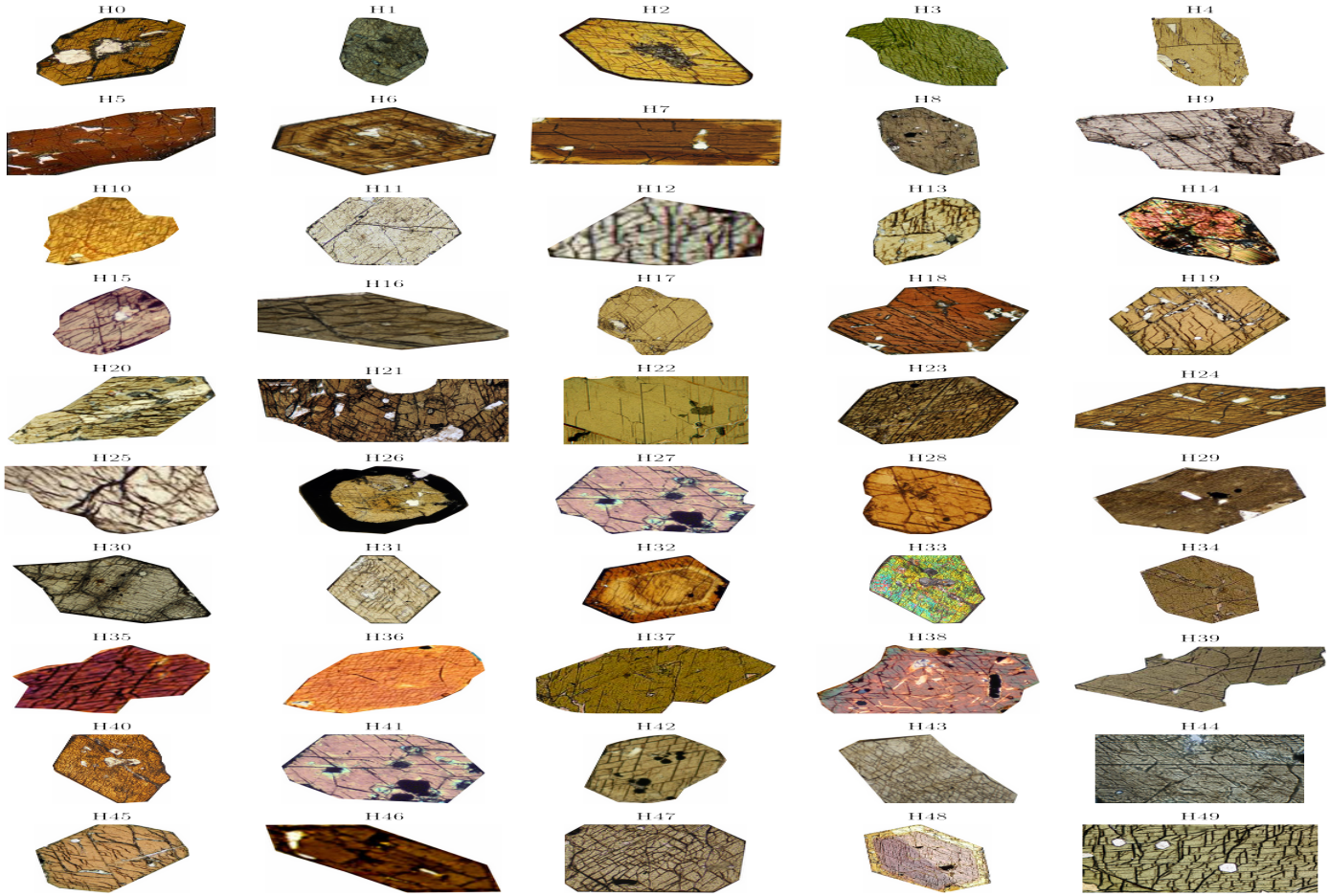
Plagioclase sample images

Mineral Classification in Petrographic images (cont.)



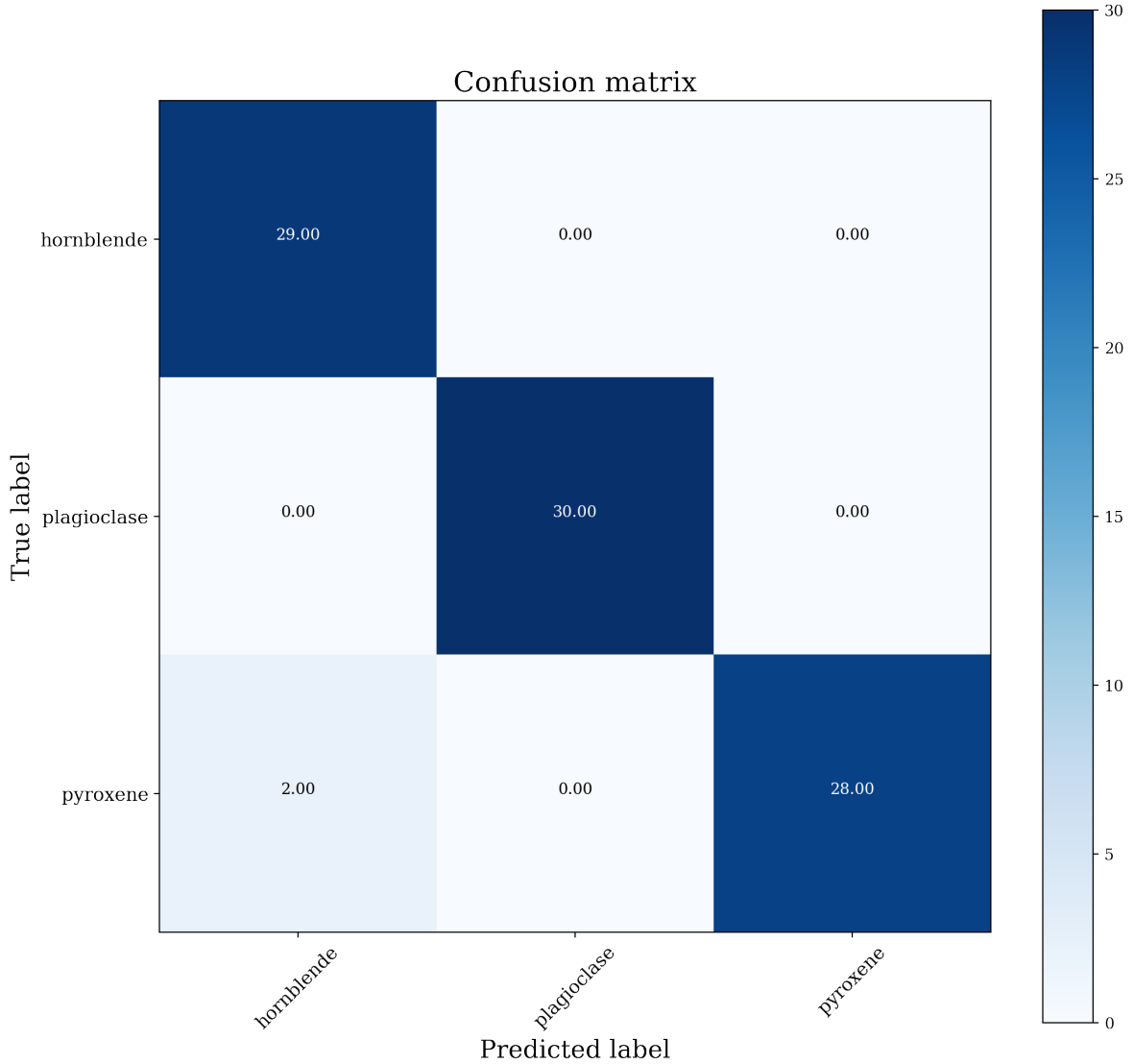
Pyroxene sample images

Mineral Classification in Petrographic images (cont.)



Hornblende sample images

Mineral Classification in Petrographic images (cont.)



Confusion matrix

THANK YOU!!!!