CDCS
CENTER FOR DATA AND COMPUTING
IN NATURAL SCIENCES

Janis Kummer (CDCS/UHH)

# Radio galaxy classification with GAN-generated data

**CDL1 — in collaboration with Florian Griese (CDCS/TUHH), Lennart Rustige (CDCS/DESY), Marcus Brüggen (HS,UHH), Frank Gaede (DESY), Gregor Kasieczka (IEXP,UHH), Peter Schleper (IEXP,UHH), Kerstin Borras (DESY/RWTH Aachen)**

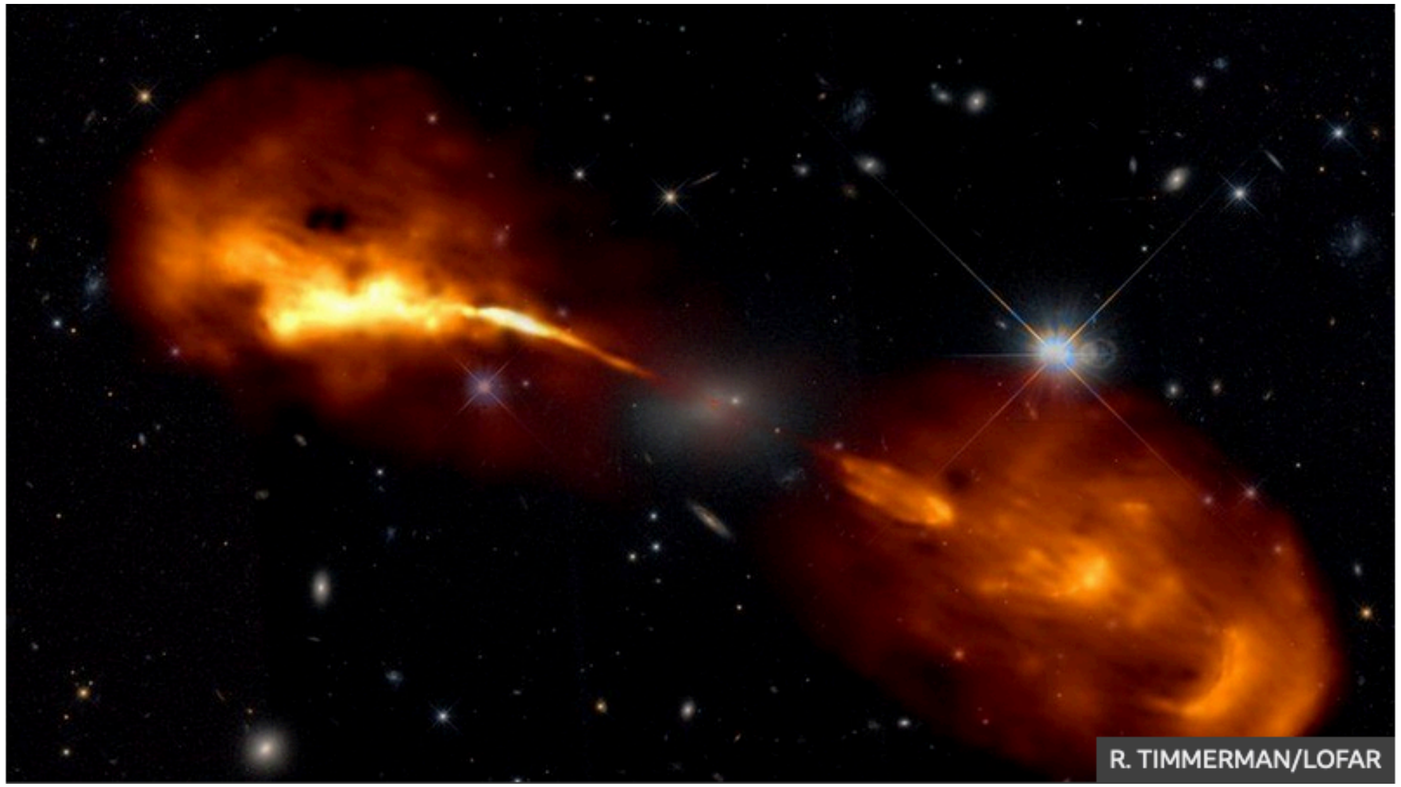
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

DESY.

TUHH
Technische Universität Hamburg-Harburg

# Radio galaxies

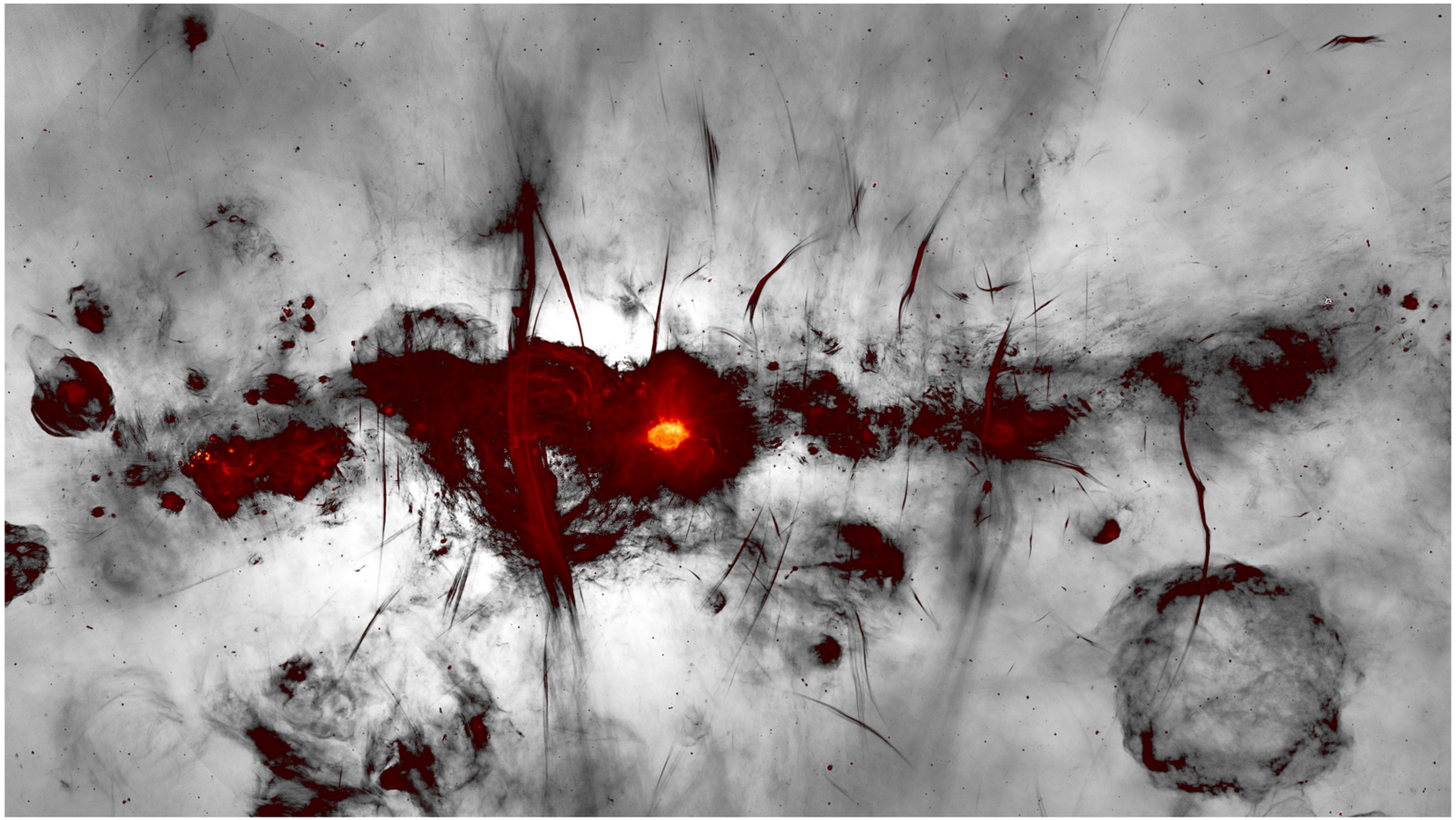





- Radio astronomy reveals processes that cannot be seen with optical telescopes
- A new generation of radio telescopes such as LOFAR, MeerKAT and SKA in the near future will generate an incredible amount of data and will be much more sensitive
- This may lead to a revolution in the field and necessitate a new level of automation for processing the data

# Radio galaxies





Heywood et al. (2022)

- MeerKAT image of the galactic centre

# Radio galaxies



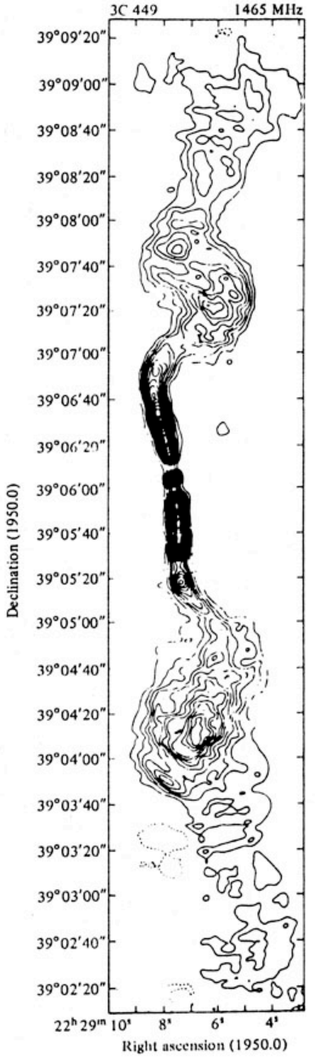
- Accreting black holes in the centre of massive elliptical galaxies power active galactic nuclei (AGN)
- AGNs have jets of charged particles which emit synchrotron radiation
- Studying radio galaxies means understanding massive black holes and their evolution
- Radio-loud sources are highly interesting for observational cosmology as they are observable at very large distances.

ESO/WFI (Optical); MPIfR/ESO/APEX/A.Weiss et al. (Submillimetre); NASA/CXC/CfA/R.Kraft et al. (X-ray)

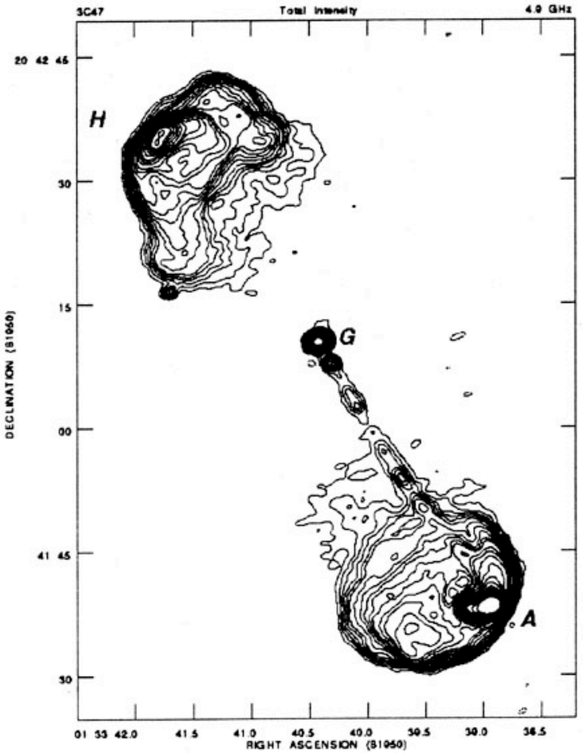
# Radio galaxy classification



- Fanaroff-Riley Classification:
  - Two classes based on the ratio $R$ of the distance between the regions of highest surface brightness on opposite sides of the central galaxy, to the total extent of the source up to the lowest brightness contour in the map.
  - Sources with $R < 0.5$ were placed in Class I (FRI) and sources with $R > 0.5$ in Class II (FRII).




FRI



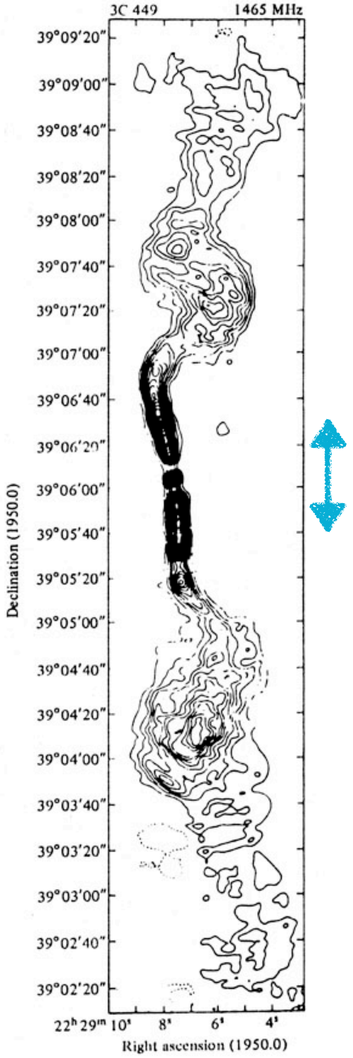

Fanaroff & Riley (1974)

FRII

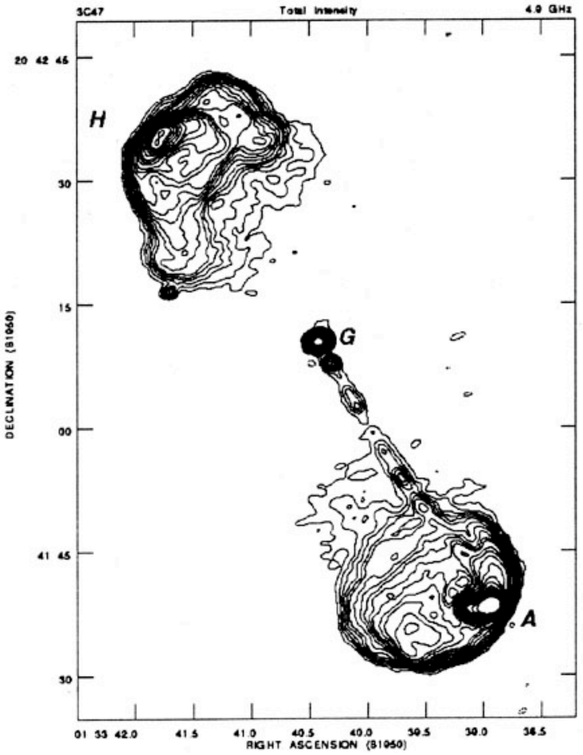
# Radio galaxy classification



- Fanaroff-Riley Classification:
  - Two classes based on the ratio $R$ of the distance between the regions of highest surface brightness on opposite sides of the central galaxy, to the total extent of the source up to the lowest brightness contour in the map.
  - Sources with $R < 0.5$ were placed in Class I (FRI) and sources with $R > 0.5$ in Class II (FRII).




FRI



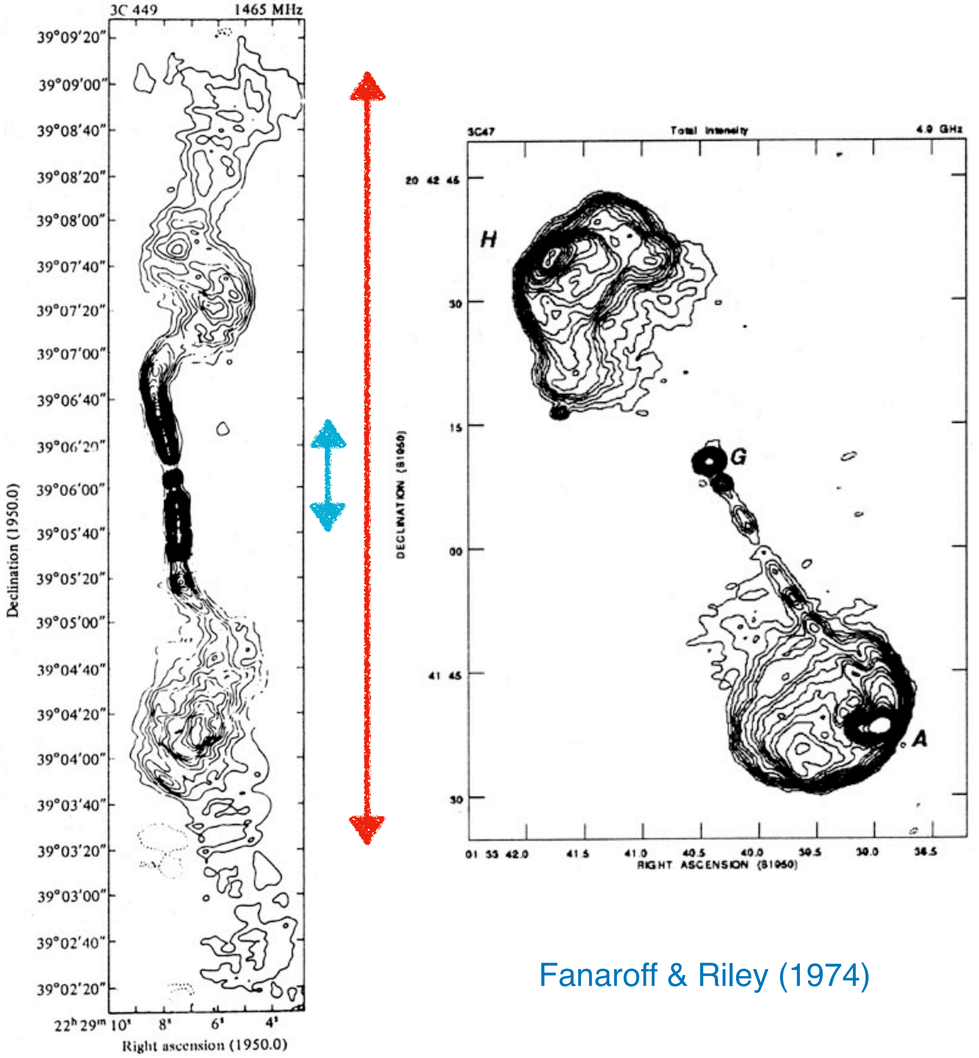

Fanaroff & Riley (1974)

FRII

# Radio galaxy classification



- Fanaroff-Riley Classification:
  - Two classes based on the ratio $R$ of the distance between the regions of highest surface brightness on opposite sides of the central galaxy, to the total extent of the source up to the lowest brightness contour in the map.
  - Sources with $R < 0.5$ were placed in Class I (FRI) and sources with $R > 0.5$ in Class II (FRII).



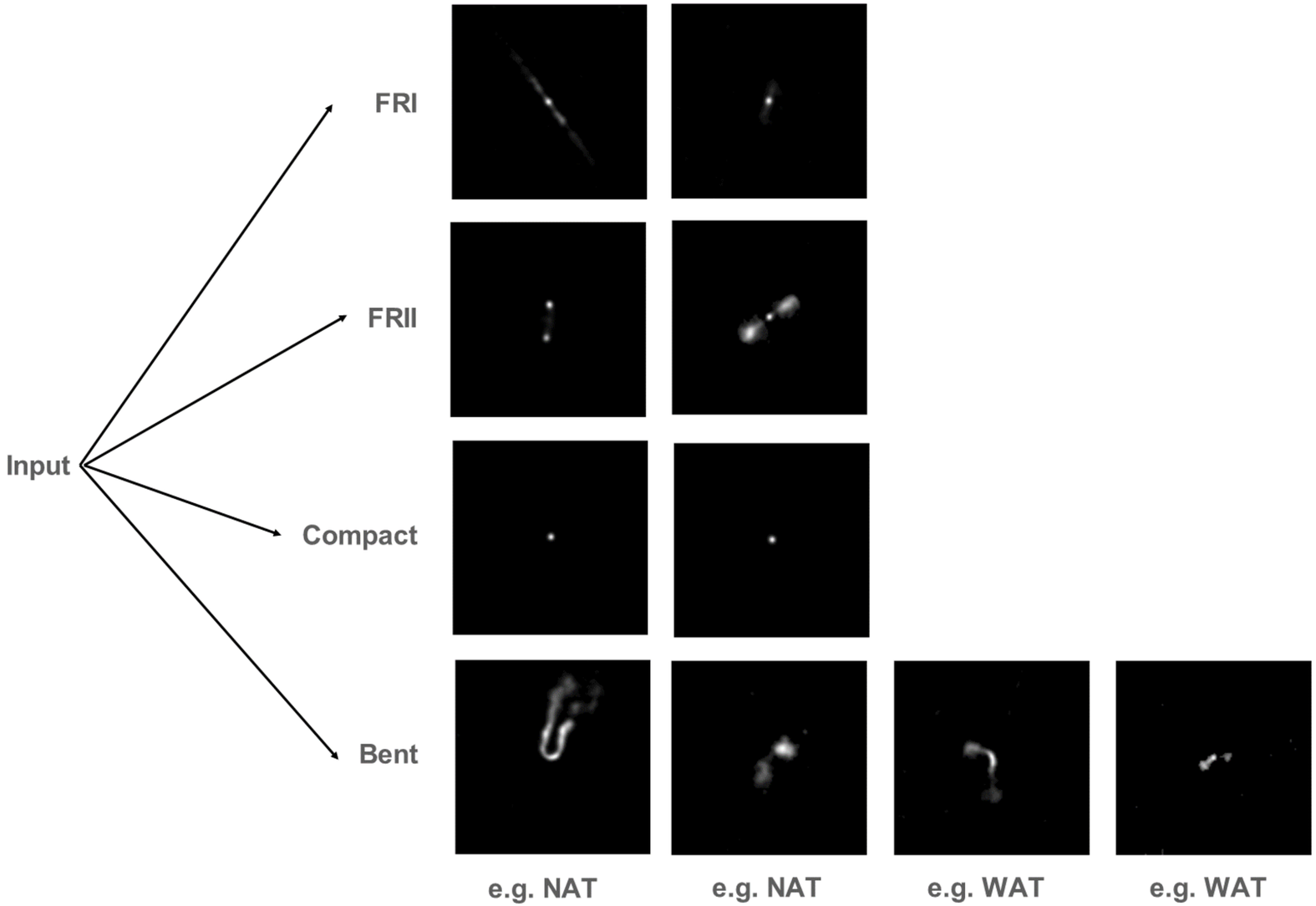

Fanaroff & Riley (1974)

FRI FRII

# Radio galaxy classification



- We consider a 4 class classification problem:
  - FRI, FRII plus compact and bent-tailed sources



# Radio galaxy classification



- Machine learnings models are successful in morphological classification of radio galaxies (see e.g. Aniyan & Thorat (2017), Alhassan et al. (2018) or Tang et al. (2019))
- Large labelled data sets are needed for such deep learning models
- However morphological labels depend on visual inspection of experts from the field → labelled data is limited
- Data sets are enlarged by data augmented (rotated and flipped images)

# Radio galaxy classification



- Machine learnings models are successful in morphological classification of radio galaxies (see e.g. Aniyan & Thorat (2017), Alhassan et al. (2018) or Tang et al. (2019))
- Large labelled data sets are needed for such deep learning models
- However morphological labels depend on visual inspection of experts from the field → labelled data is limited
- Data sets are enlarged by data augmented (rotated and flipped images)
- Our idea: train classifiers on data sets augmented with the help of generative models i.e. we add generated images to the data set

# Data set



- We combined the several catalogues and checked for duplicates
  - E.g. CoNFIG (Gendre & Wall (2008), Gendre et al. (2010)) combines observations from FIRST, NVSS and SDSS to characterise radio source including the morphology i.e. FRI/ FRII/ Compact
- We download images of the FIRST survey (Becker et al. 1995) from the virtual observatory skyview
- Cropped to 128 x 128 pixels and all pixel values below 3 x local RMS noise set to zero

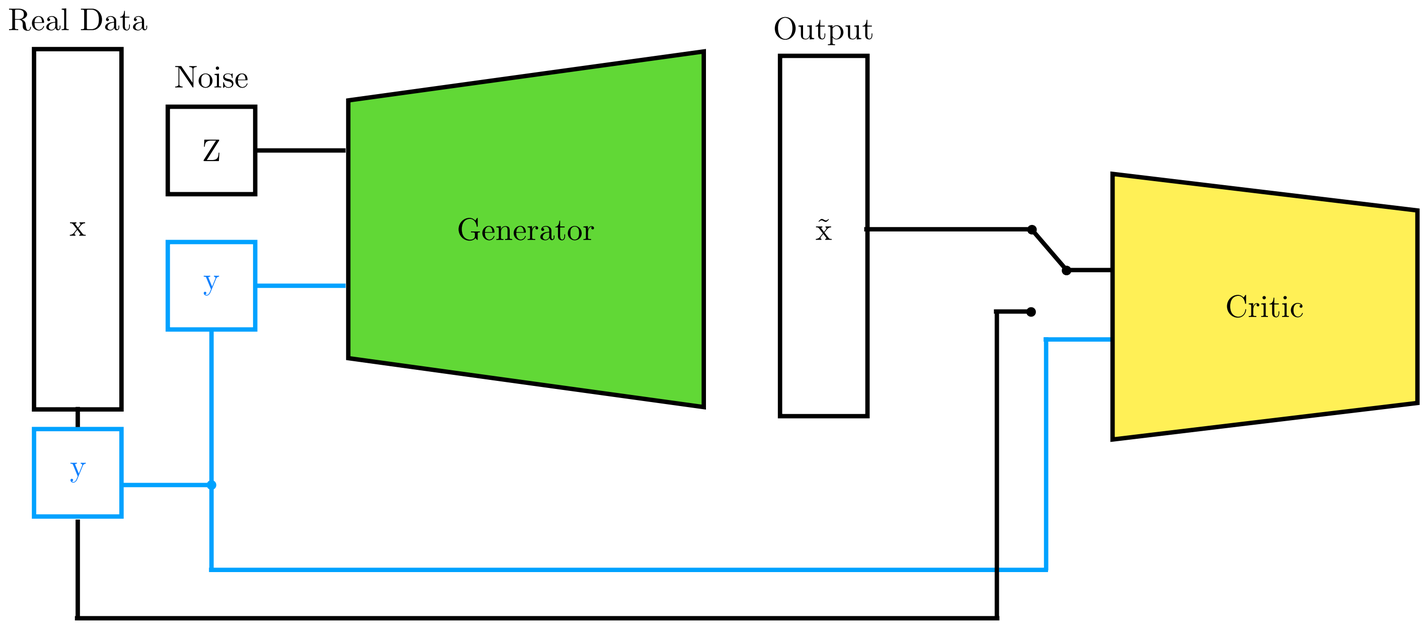
https://skyview.gsfc.nasa.gov

| | FRI | FRII | Compact | Bent | Total |
|---|---|---|---|---|---|
| train | 400 | 823 | 291 | 242 | 1756 |
| validation | 50 | 50 | 50 | 50 | 200 |
| test | 50 | 50 | 50 | 50 | 200 |
| total | 500 | 923 | 391 | 342 | 2156 |

# Generative model



- Wasserstein GAN with gradient penalty

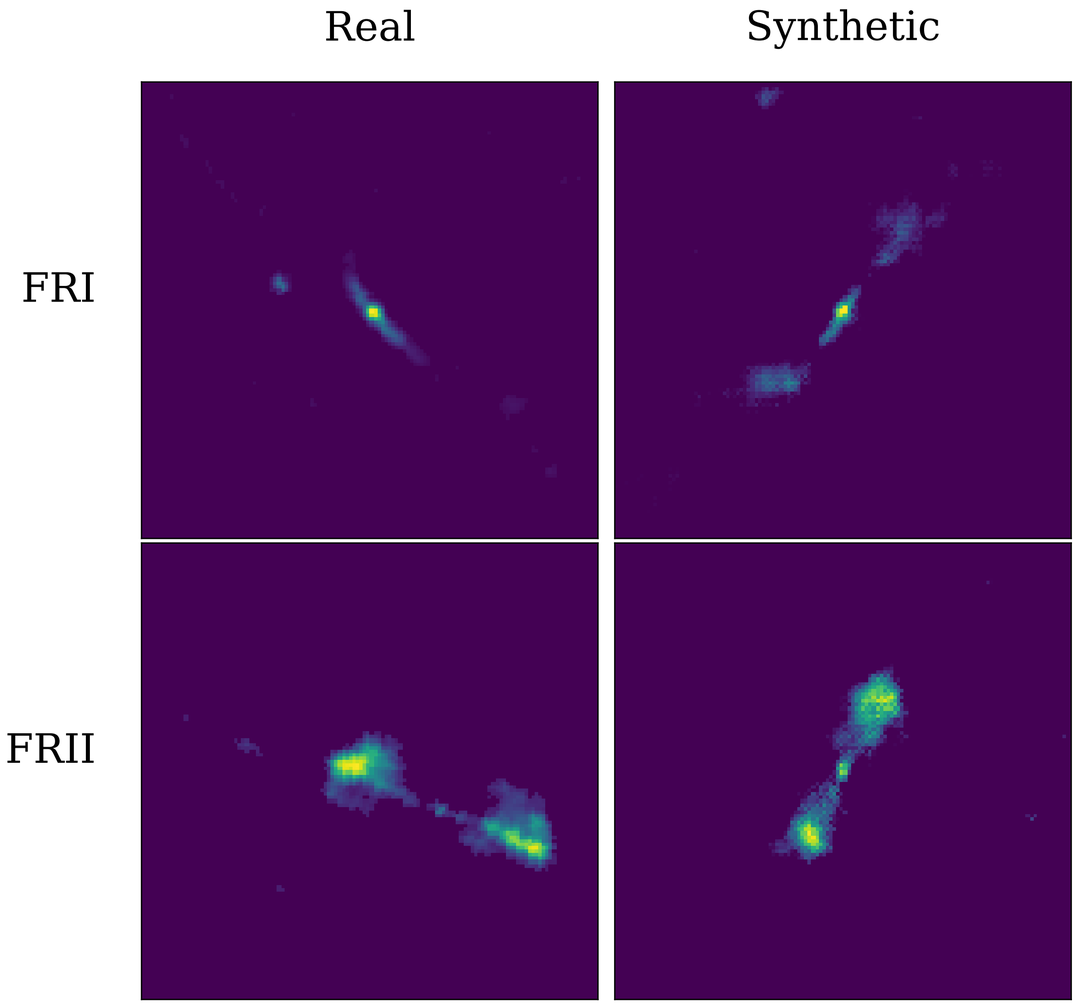



- Generator and Critic based on convolutional layers
- Conditioned on the class labels

# Generated images



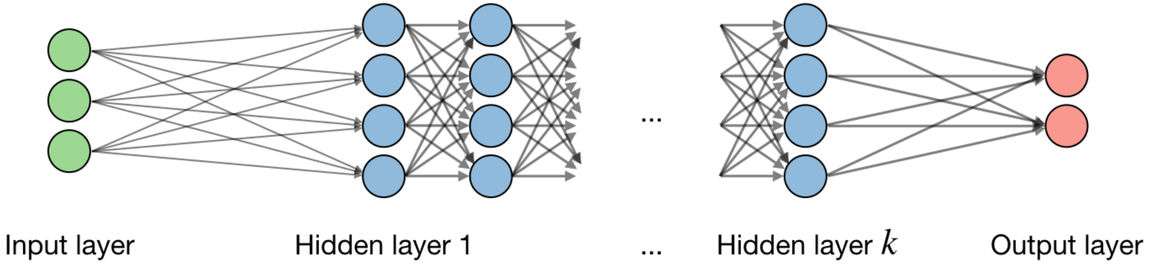
- Best generator model found by a metric based on statistical properties of the images
- In particular, we compare several histograms (e.g. pixel intensities) of real and synthetics images



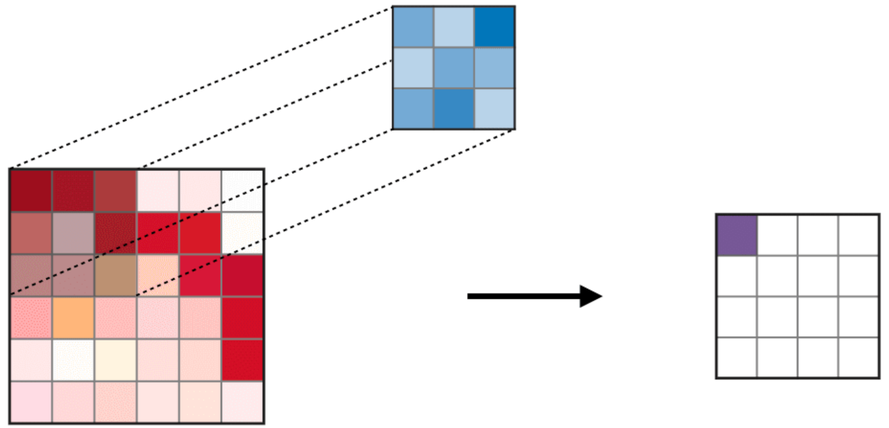
# Classifiers



- Two classifier setups:
  - Fully-connected network (FCN)
  - Convolutional neural network (CNN)



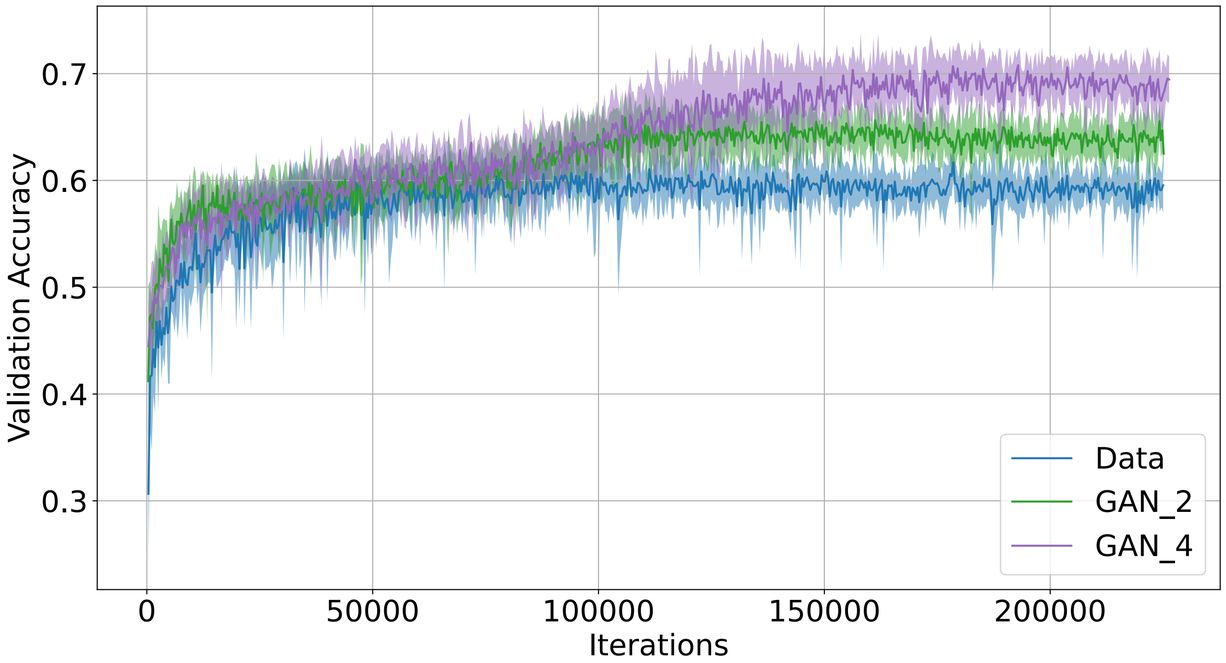

- Two approaches:
  - Generating fixed sets of images before training
  - On the fly generation of images (not covered)
- In both cases the resulting data set is balanced

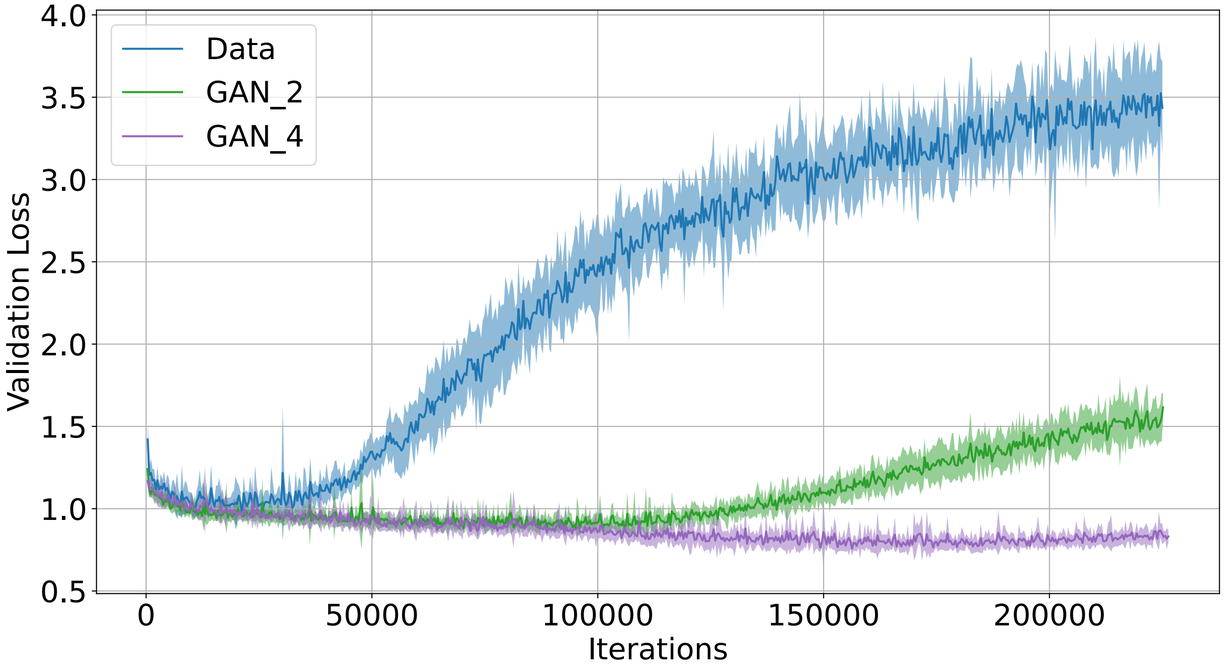


https://stanford.edu/~shervine/teaching/

# Fully-connected network






- GAN_n: size(generated set) = n x size(training set)

# Fully-connected network

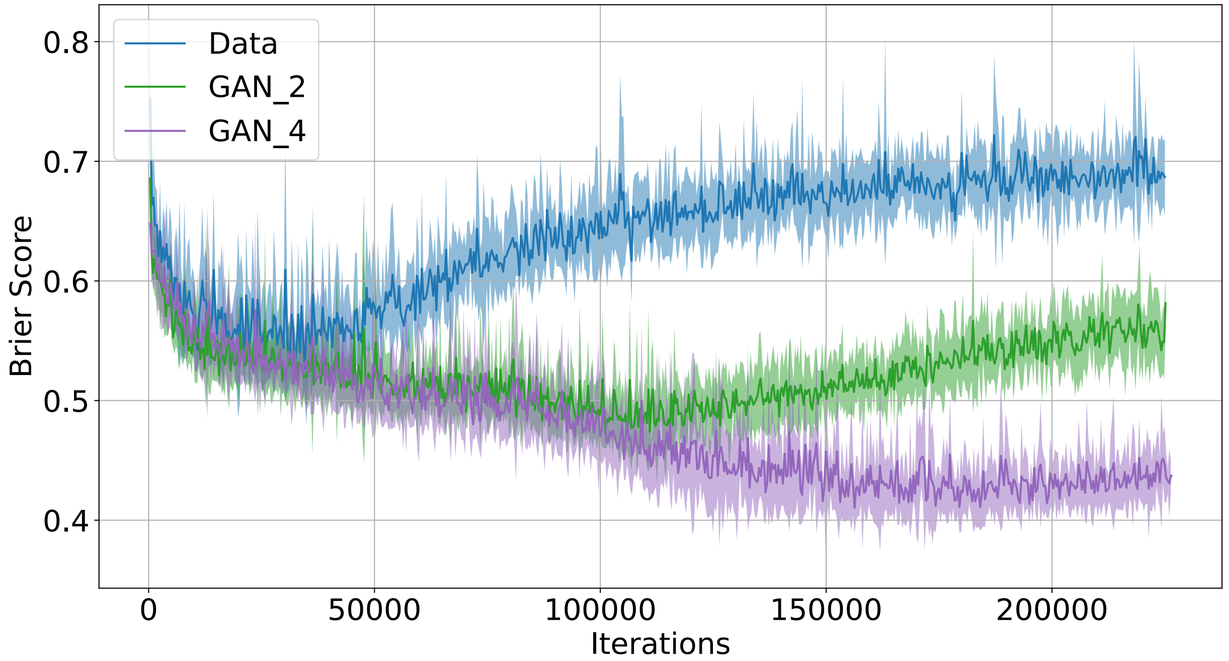


- Best model = minimal Brier Score

$$BS = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{R} (f_{ti} - o_{ti})^2$$

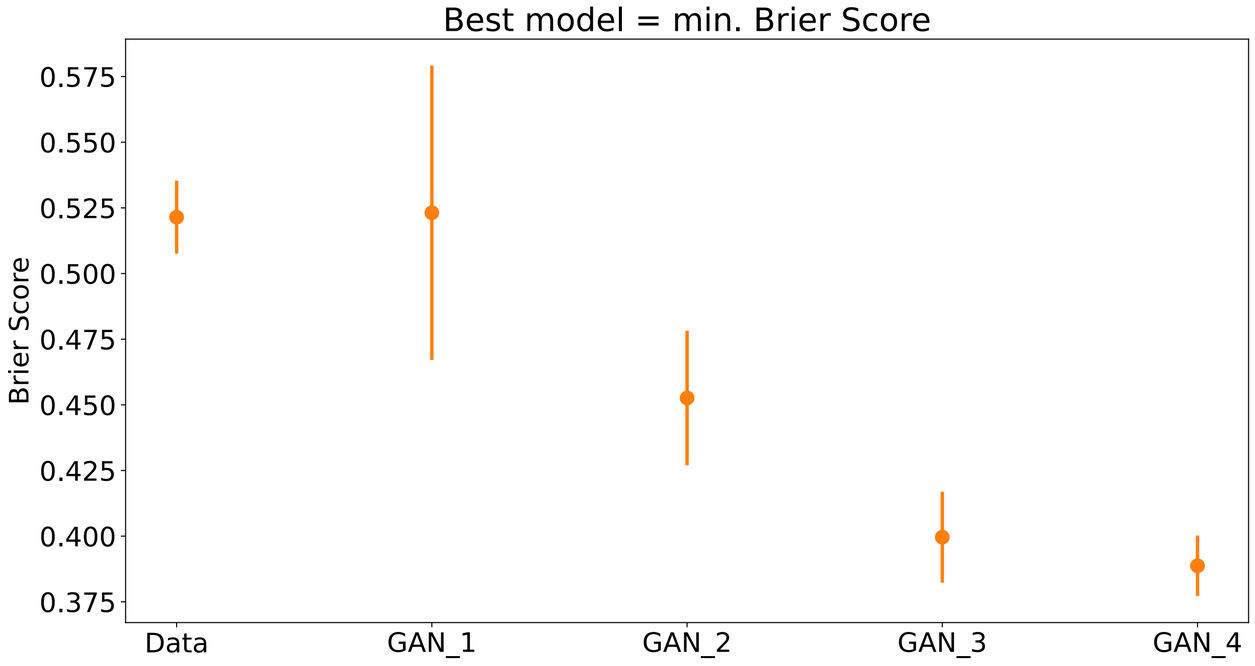
Brier (1950)

- Mean squared error of predicted probabilities



# Fully-connected network on test set



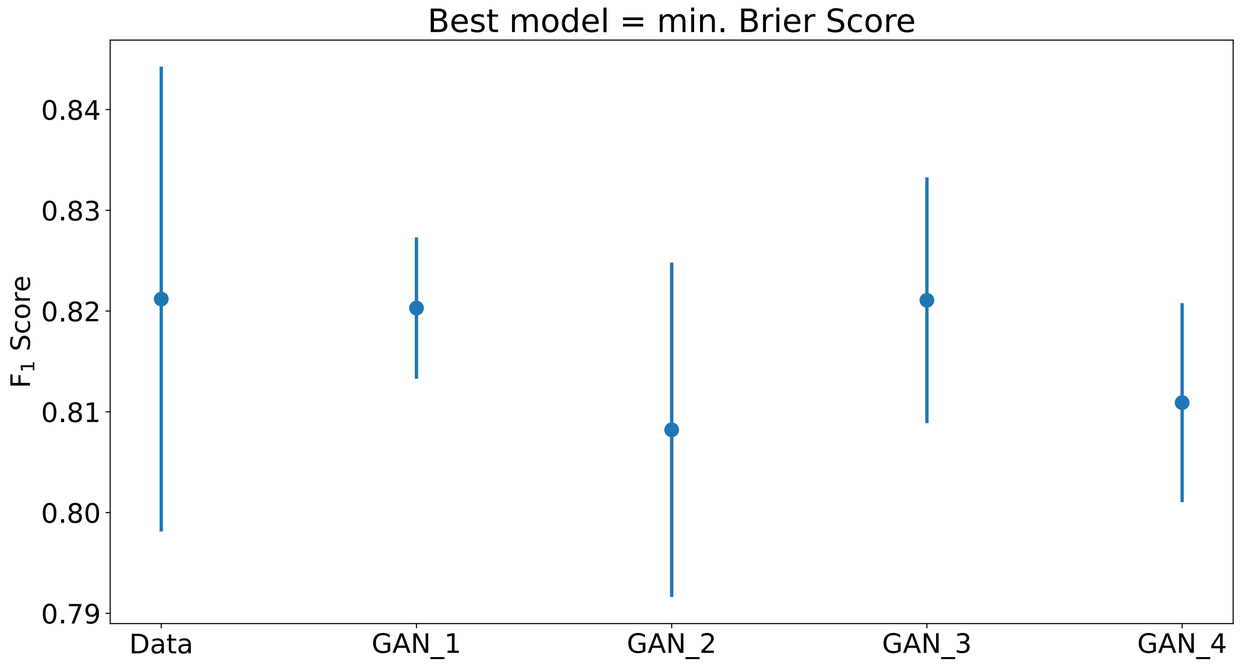

## $F_1$ Score

## Brier Score

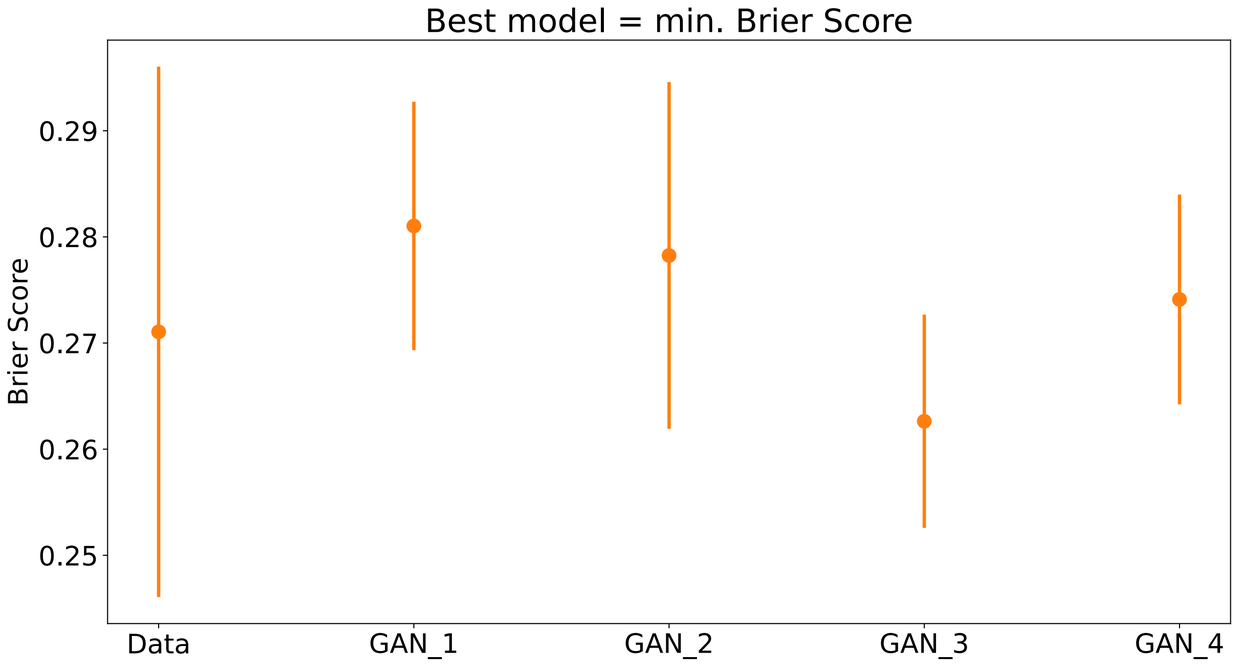
$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \text{Precision} = \frac{tp}{tp + fp}, \text{Recall} = \frac{tp}{tp + fn}$$

# CNN on test set





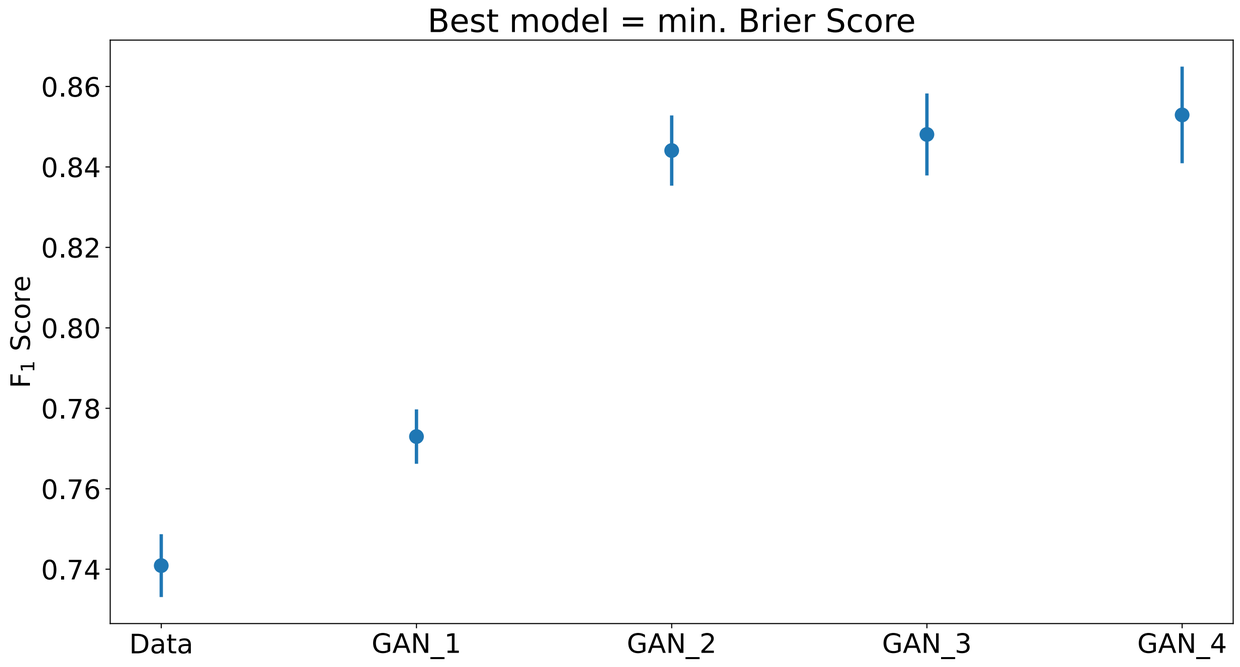

$F_1$ Score



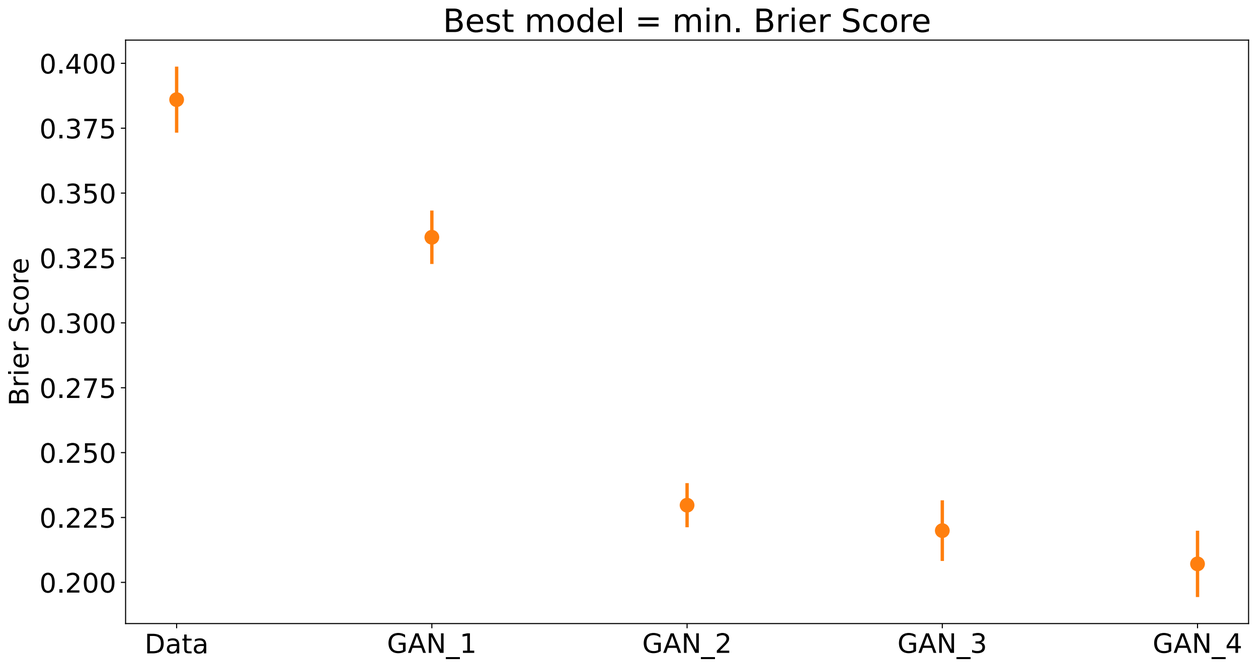

Brier Score

# CNN on GAN-generated test set






$F_1$ Score




Brier Score

# Conclusions



- We managed to simulate radio galaxies with generative models realistic enough for subsequent applications.
- We are able to improve a simple (FCN) classifier significantly compared to the data only baseline by adding GAN-generated images to the training set.
- More about the project → Florian Griese's talk in the CCU session