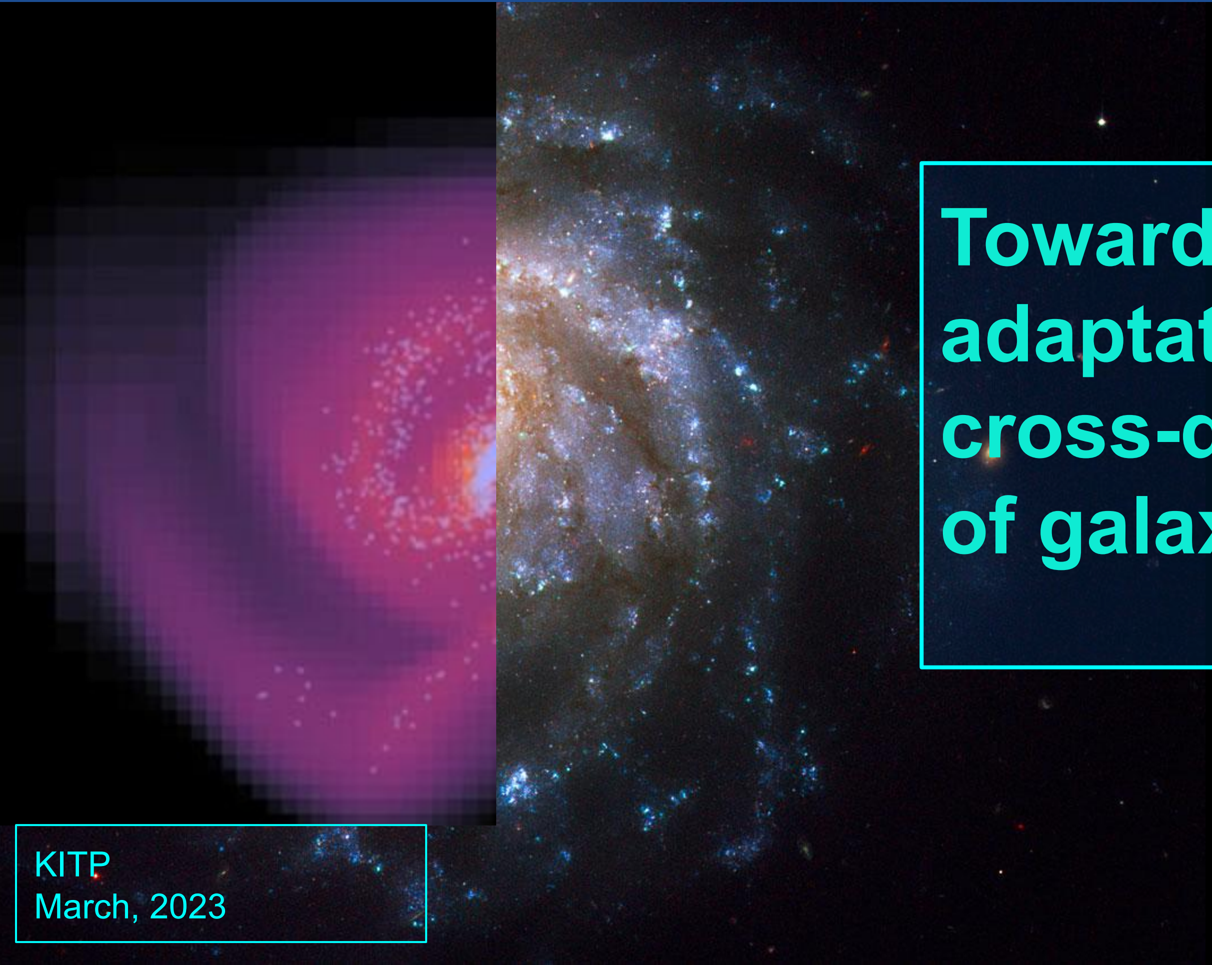
Fermilab U.S. DEPARTMENT OF ENERGY Office of Science



Towards flexible domain adaptation methods for cross-datasets studies of galaxies

Aleksandra Ćiprijanović
(she/her/hers)

Fermilab, DSSL
aleksand@fnal.gov

KITP
March, 2023

Vision of the Future

- **Real-time:**
  - data handling,
  - decision making
  - detection of interesting events
  - inference
- **Automated experiments**
- **Working with big data** later in the process

Rubin LSST

~ 20 TB / day
~ 100 PB total by DR11

Often research starts from simulations…
…but when AI gets involved…
…we face some challenges.

Talk Outline

1. Domain Shift Problems

2. Domain Adaptation

3. Universal Domain Adaptation

4. Future Challenges

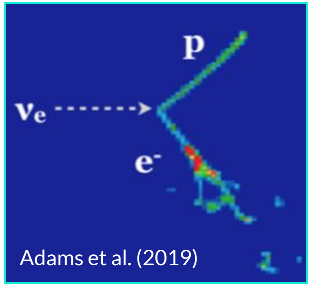
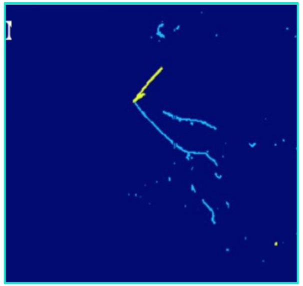
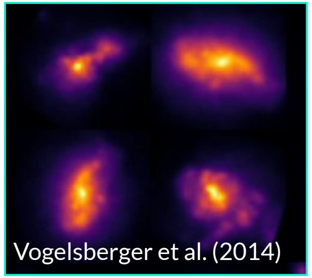
Talk Outline

1. Domain Shift Problems
2. Domain Adaptation
3. Universal Domain Adaptation
4. Future Challenges

Combining Datasets

All areas of Fermilab science often need to create **model trained on simulated data, that also work on real detector data**!

DATASET SHIFT

	SIMULATED	REAL
MicroBooNE (neutrinos)	 Adams et al. (2019)	
Illustris / Hubble (merging galaxies)	 Vogelsberger et al. (2014)	 Hubble

Fermilab

Combining Datasets

All areas of Fermilab science often need to create **model trained on simulated data, that also work on real detector data**!

DATASET SHIFT

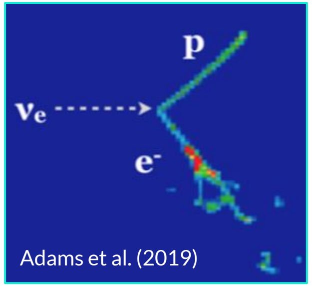
Missing and unknown physics, wrong geometry, background levels

Computational constraints for simulations

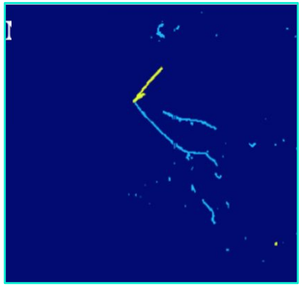
SIMULATED

REAL

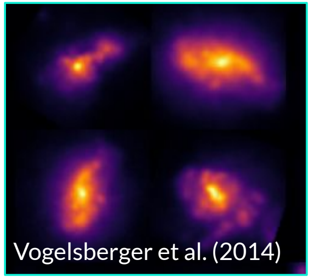
MicroBooNE (neutrinos)



Adams et al. (2019)



Illustris / Hubble (merging galaxies)


Vogelsberger et al. (2014)

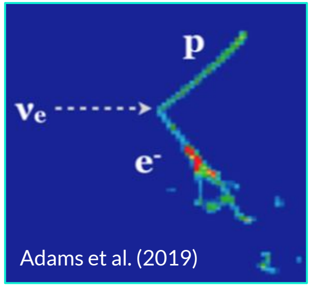
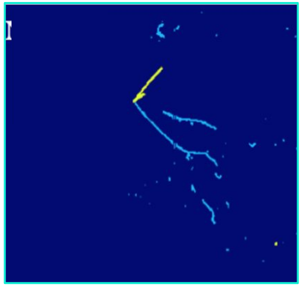
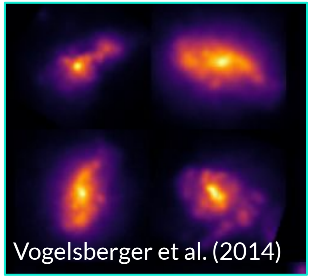

Hubble

Combining Datasets

All areas of Fermilab science often need to create **model trained on simulated data, that also work on real detector data**!

DATASET SHIFT

- Missing and unknown physics, wrong geometry, background levels
- Computational constraints for simulations
- Detector problems, transients, errors, data compression
- Imperfect addition of observational effects

	SIMULATED	REAL
MicroBooNE (neutrinos)	 Adams et al. (2019)	
Illustris / Hubble (merging galaxies)	 Vogelsberger et al. (2014)	 Hubble

Combining Datasets

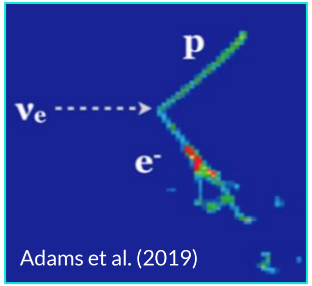
All areas of Fermilab science often need to create **model trained on simulated data, that also work on real detector data**!

DATASET SHIFT

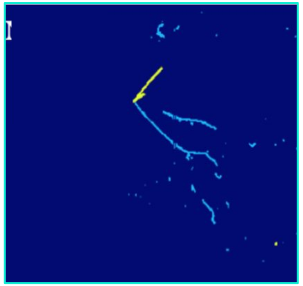
- Missing and unknown physics, wrong geometry, background levels
- Computational constraints for simulations
- Detector problems, transients, errors, data compression
- Imperfect addition of observational effects
- Different detectors or telescopes

SIMULATED | REAL

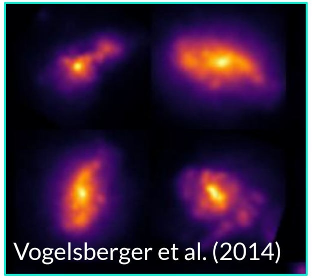
MicroBooNE (neutrinos)



Adams et al. (2019)

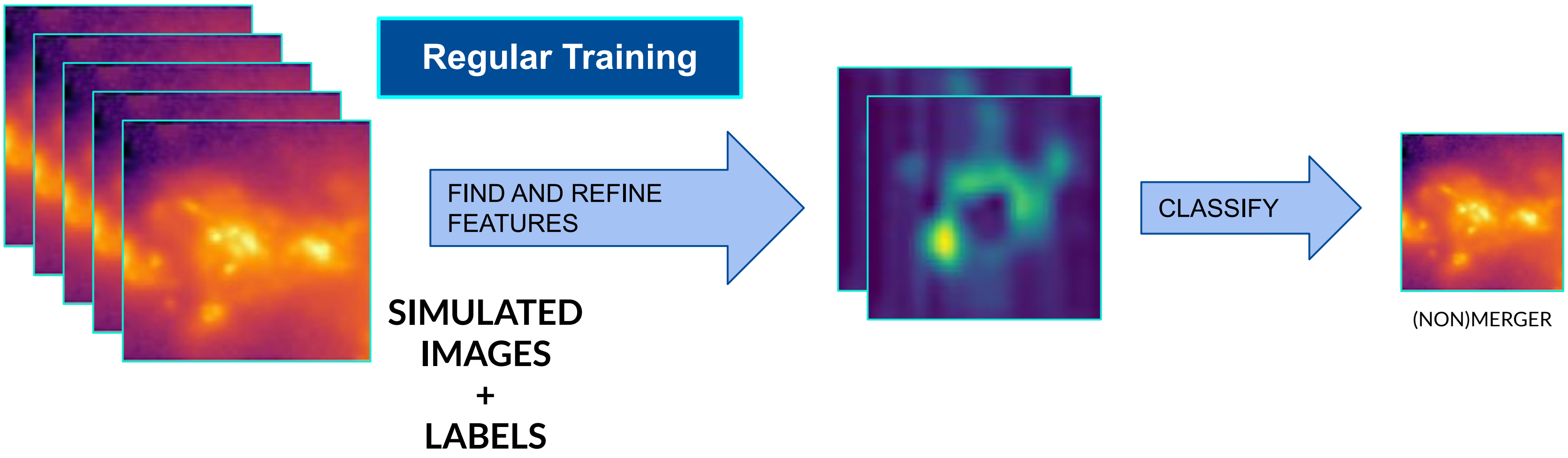


Illustris / Hubble (merging galaxies)

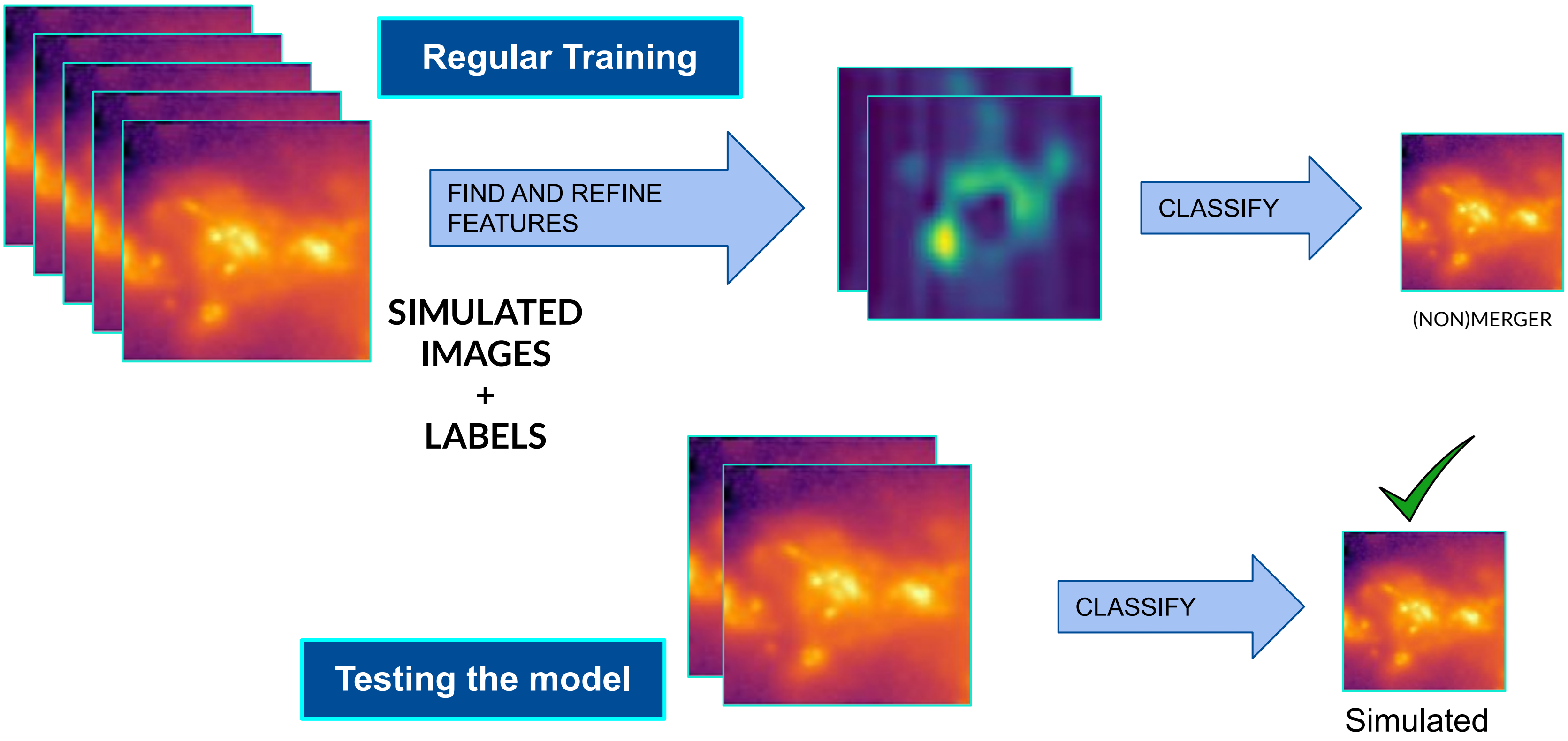

Vogelsberger et al. (2014)


Hubble

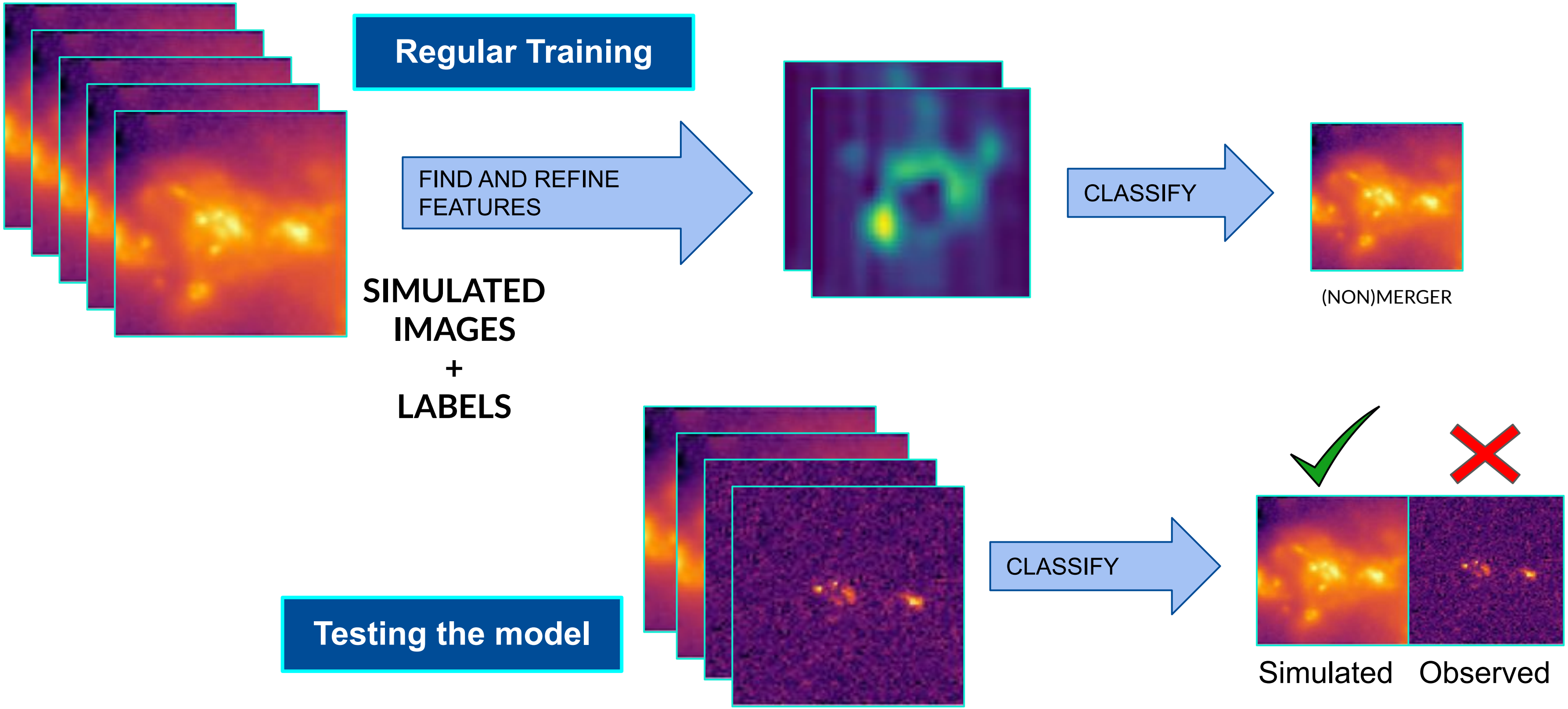
Combining Datasets



Combining Datasets



Combining Datasets



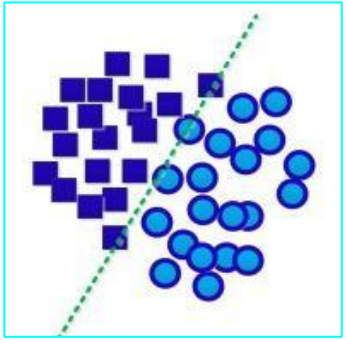
Combining Datasets

Why does this happen?

Combining Datasets

Why does this happen?

Source Domain



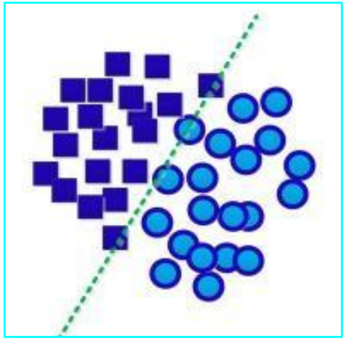
Train the model on source dataset and find the decision boundary.

Fermilab

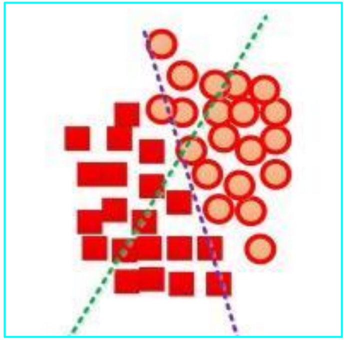
Combining Datasets

Why does this happen?

Source Domain



Target Domain



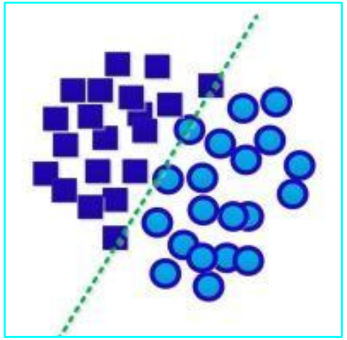
New domain is shifted, learned decision boundary doesn't work.

Combining Datasets

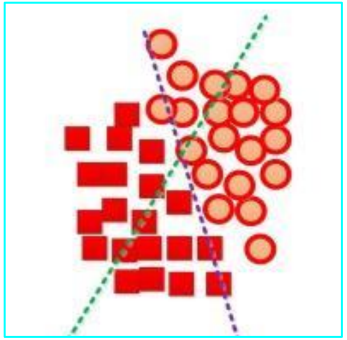
Why does this happen?

We need to align the data during training!

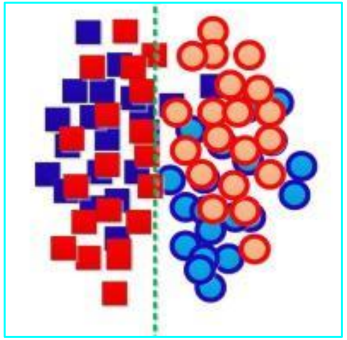
Source Domain



Target Domain



Domain Alignment



Talk Outline

1. Domain Shift Problems
2. Domain Adaptation
3. Universal Domain Adaptation
4. Future Challenges

Combining Datasets

DOMAIN ADAPTATION

Align data distributions in the latent space of the network by forcing the network to **find more robust domain-invariant features.**

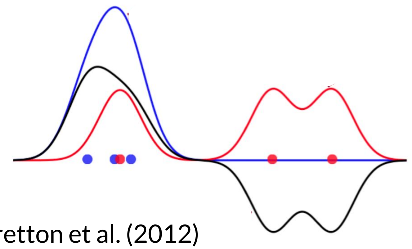
Combining Datasets

DOMAIN ADAPTATION

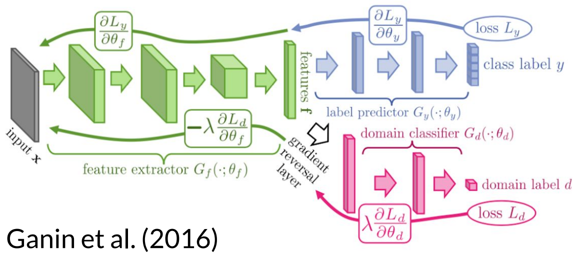
Align data distributions in the latent space of the network by forcing the network to **find more robust domain-invariant features.**

Distance-based methods

Adversarial methods



Gretton et al. (2012)



Ganin et al. (2016)

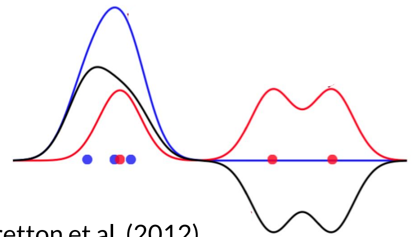
Combining Datasets

DOMAIN ADAPTATION

Align data distributions in the latent space of the network by forcing the network to **find more robust domain-invariant features.**

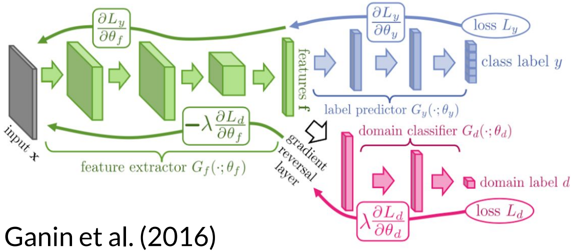
Distance-based methods

Adversarial methods



Gretton et al. (2012)

Training = Task Loss + DA Loss



Ganin et al. (2016)

Combining Datasets

DOMAIN ADAPTATION

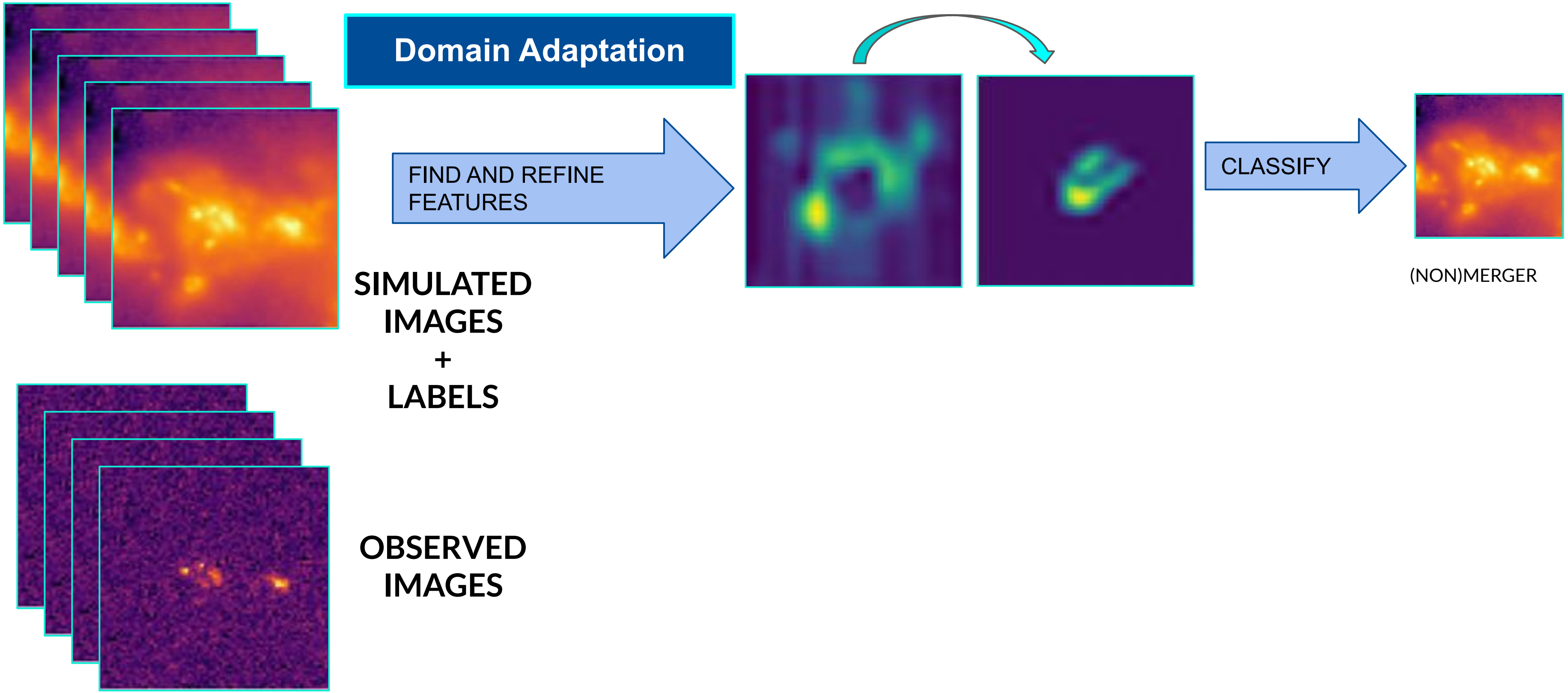
Align data distributions in the latent space of the network by forcing the network to **find more robust domain-invariant features.**

Distance-based methods

Adversarial methods

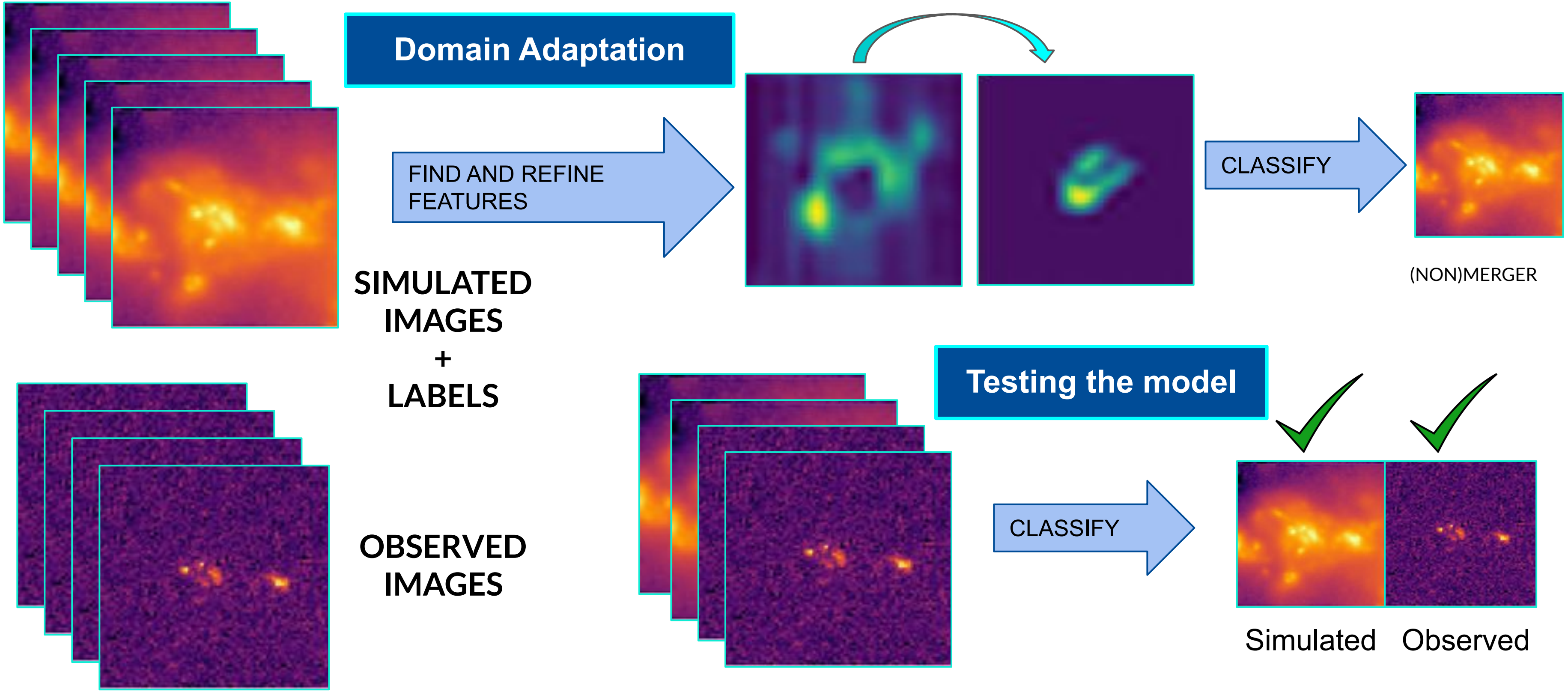
Works on **unlabeled target domain**!
Can be applied to **new data**, no need for scientists to label anythin

Combining Datasets



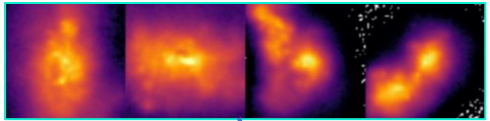
Fermilab

Combining Datasets

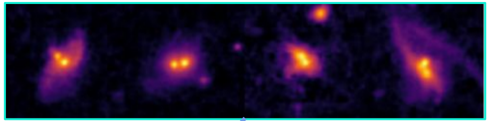


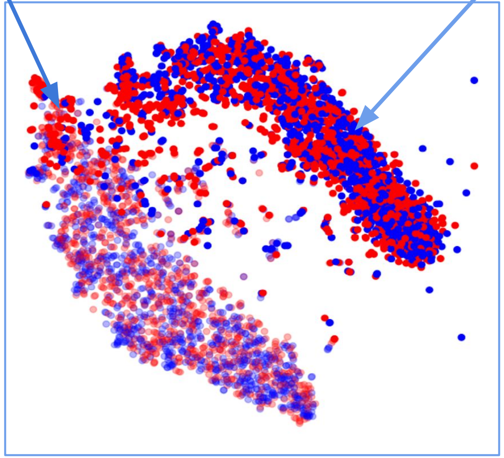
Combining Datasets

Source - Illustris



Target - SDSS observations



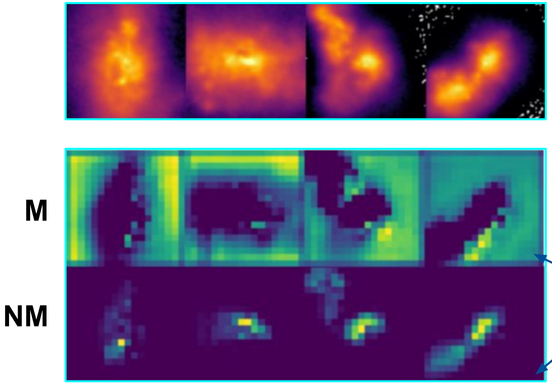


This is how the network sees the data.
2D representation of network's latent space.

Ćiprijanović et al. 2020b.
Ćiprijanović et al. 2021a.

Combining Datasets

Source - Illustris



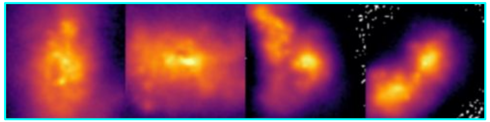
Important regions are highlighted!

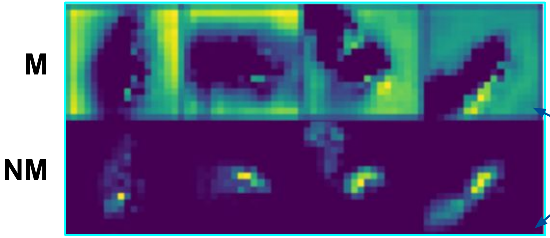
Regular Training

Ćiprijanović et al. 2020b.
Ćiprijanović et al. 202

Combining Datasets

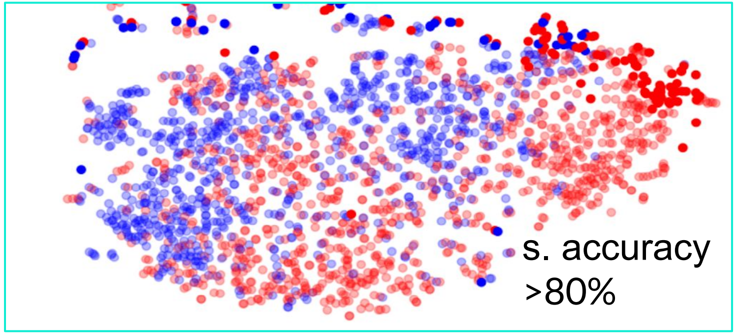
Source - Illustris





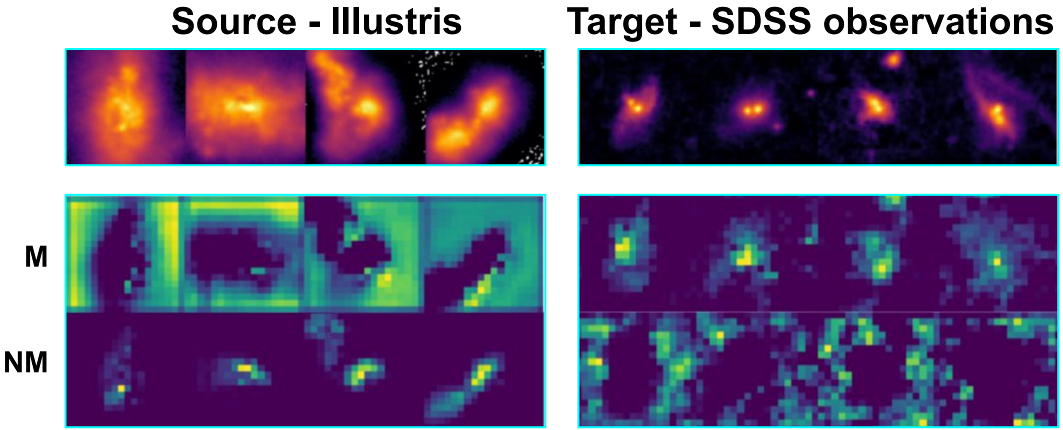
Important regions are highlighted!

Regular Training

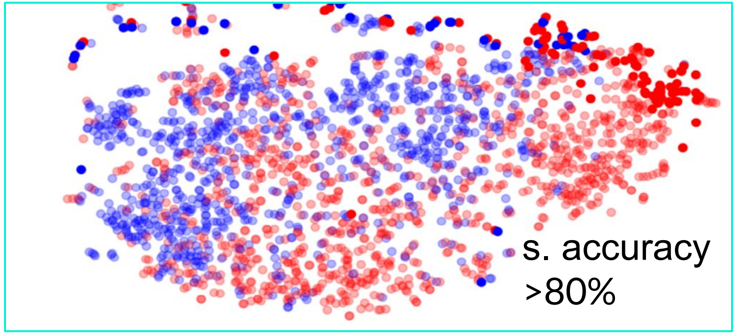


Ćiprijanović et al. 2020b.
Ćiprijanović et al. 202

Combining Datasets

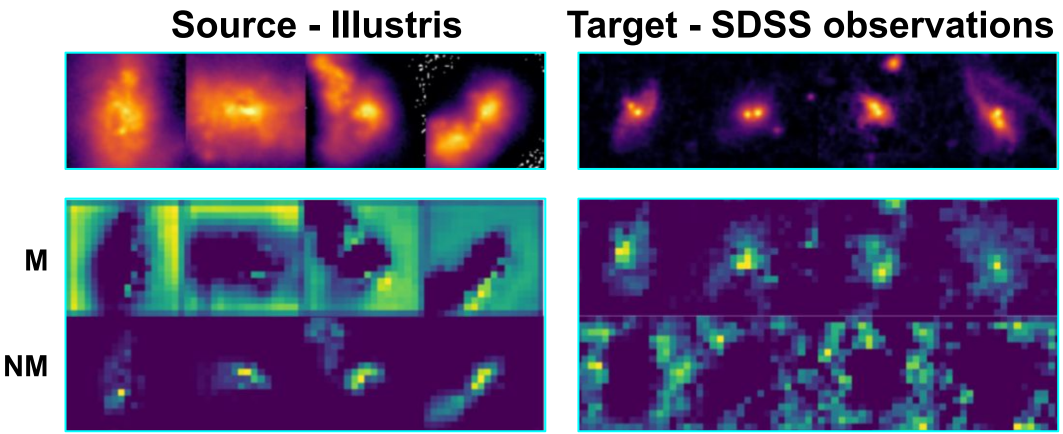


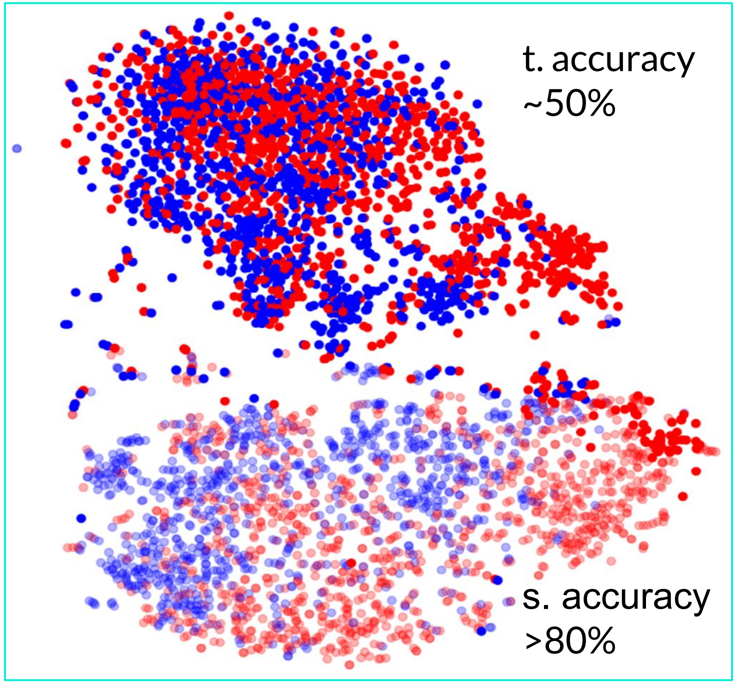
Regular Training



Ćiprijanović et al. 2020b.
Ćiprijanović et al. 202

Combining Datasets

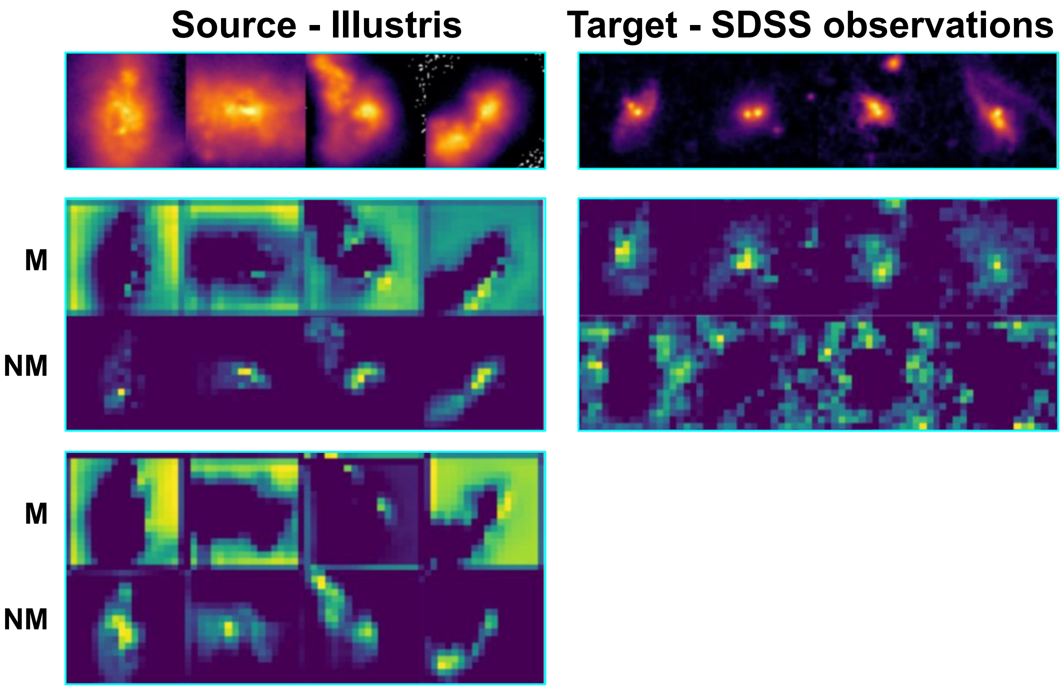




Ćiprijanović et al. 2020b.
Ćiprijanović et al. 202

Fermilab

Combining Datasets

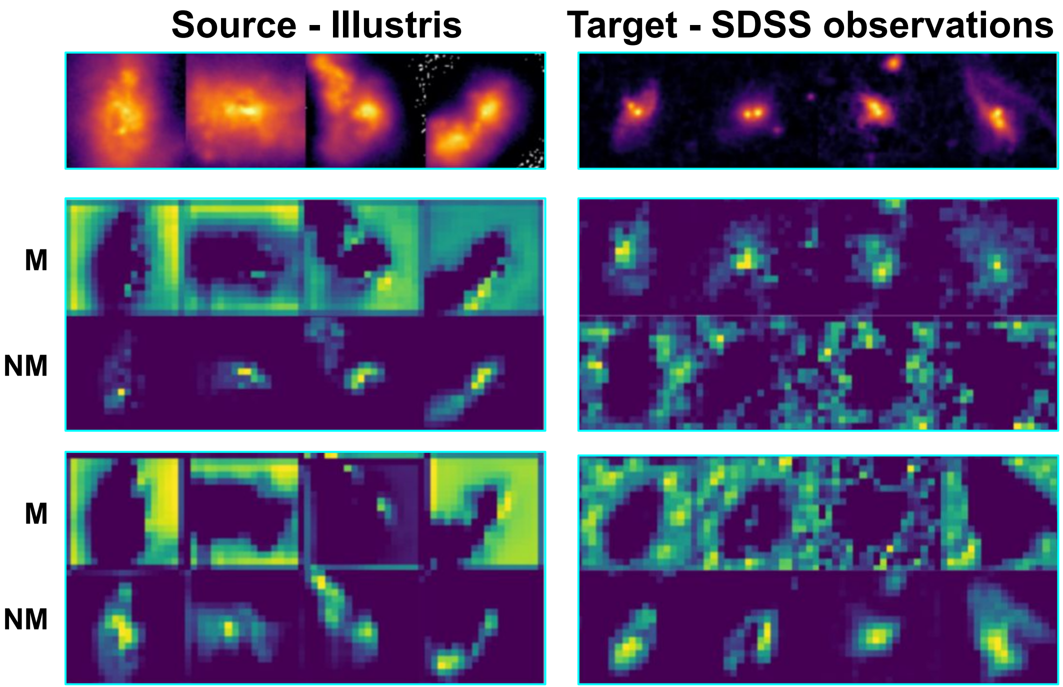


Domain Adaptation

Ćiprijanović et al. 2020b.
Ćiprijanović et al. 202

Fermilab

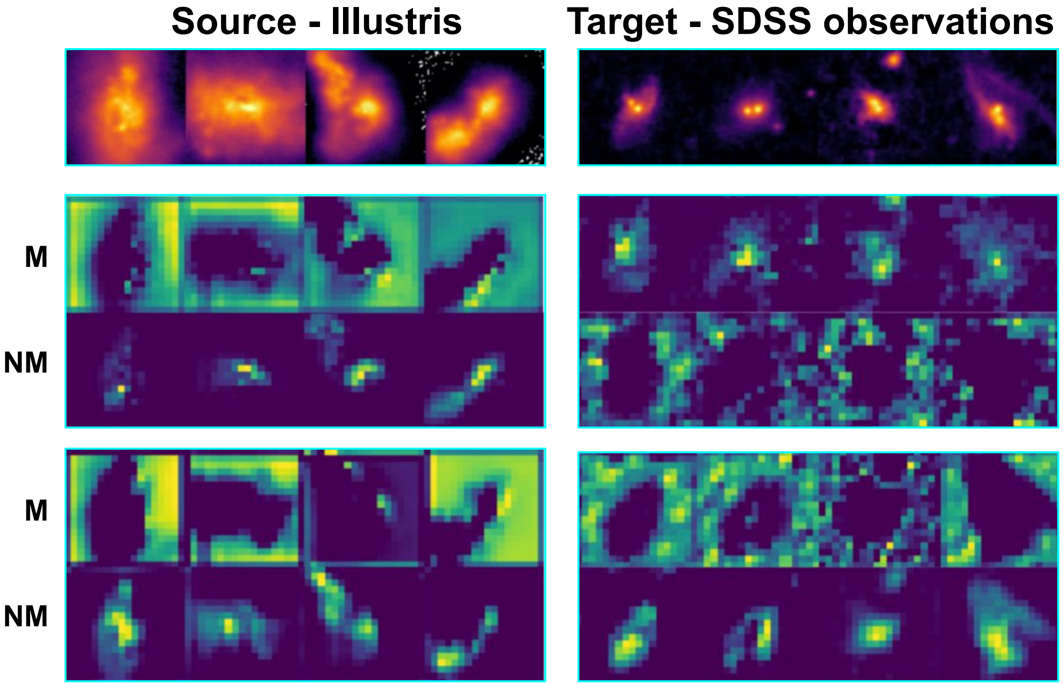
Combining Datasets



Domain Adaptation

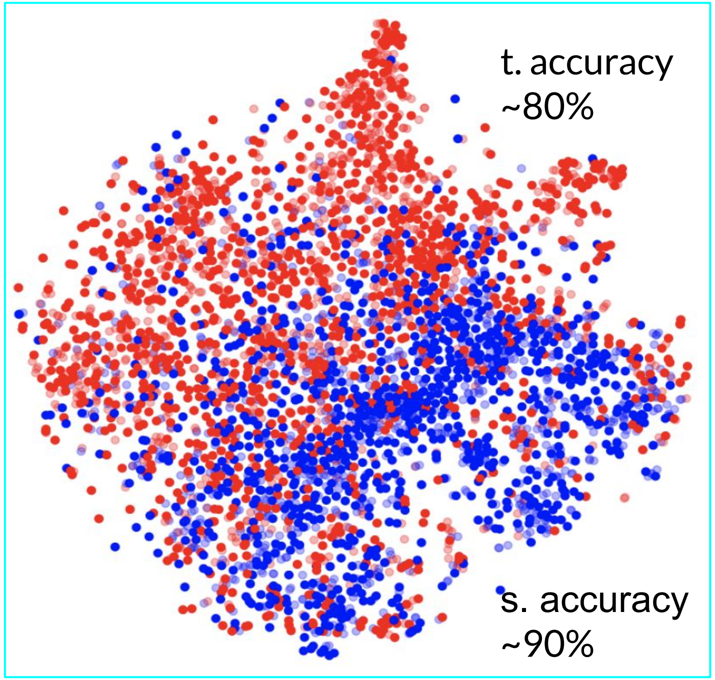
Ćiprijanović et al. 2020b.
Ćiprijanović et al. 202

Combining Datasets



Ćiprijanović et al. 2020b.
Ćiprijanović et al. 2021a

Up to 30% increase!



Talk Outline

1. Domain Shift Problems

2. Domain Adaptation

3. Universal Domain Adaptation

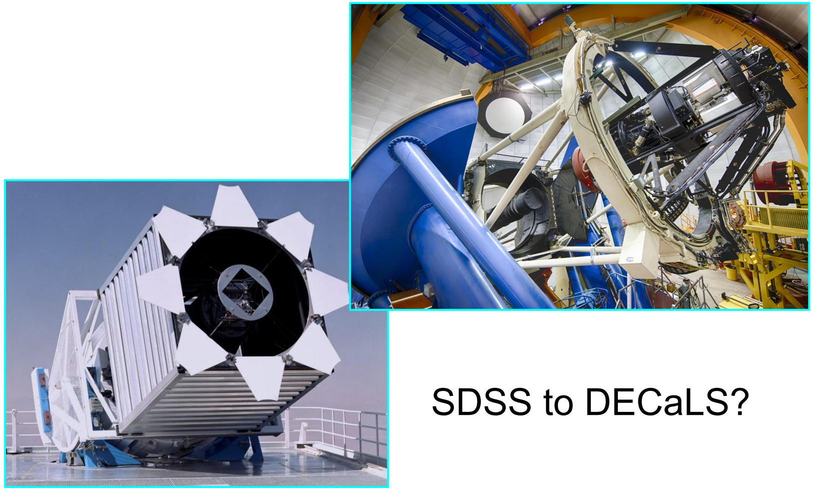
4. Future Challenges

Bridging between observations - Much Harder!

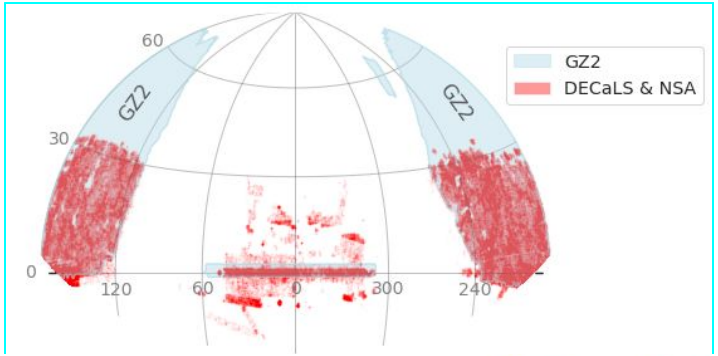
The gap between observational datasets is much larger:

- Noise, PSF
- Pixel scale
- Depth of the survey
- Magnitude limit
- Perhaps different filters
- Different data distributions….

How do we build something flexible enough to handle any kind of data distributions and distribution overlaps?



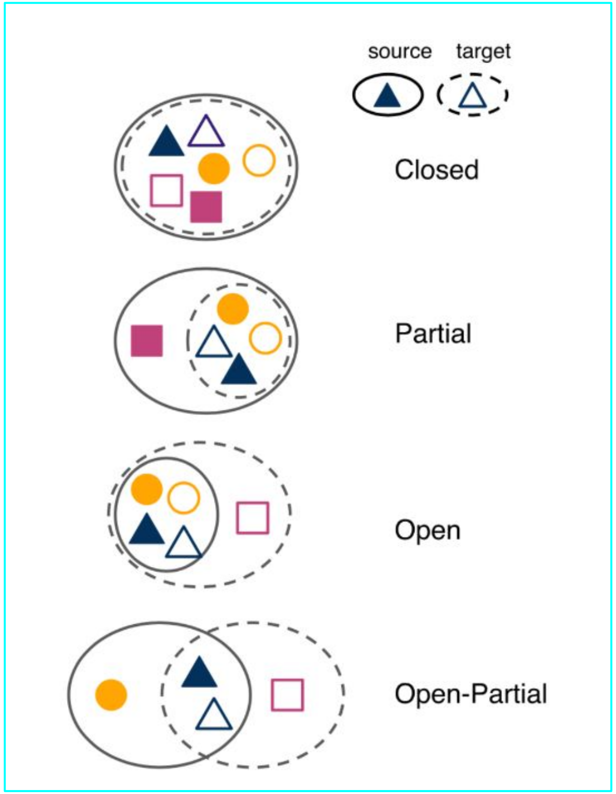
SDSS to DECaLS?



Fermilab

Types of Dataset Shift Problems

- Overall distribution per class can be different between datasets.
    - Overlapping classes should be aligned independently instead of aligning the entire data distribution.
- We can even have classes present in only one of the datasets - old labeled data or even new unlabeled data (so we won't even know it's there!)
    - Non-overlapping classes should not be aligned with anything.



Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

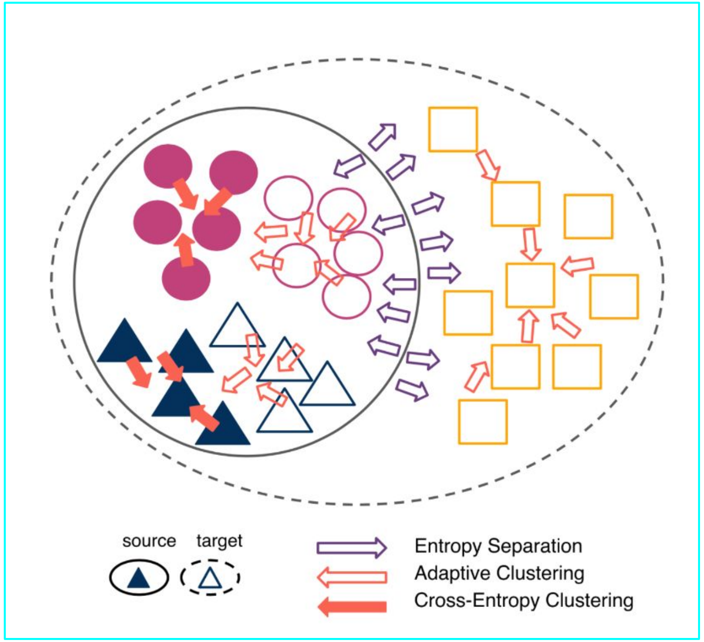
Classification of known classes

+

Clustering of similar known and unknown samples

+

Separation of different (anomalous) unknown samples



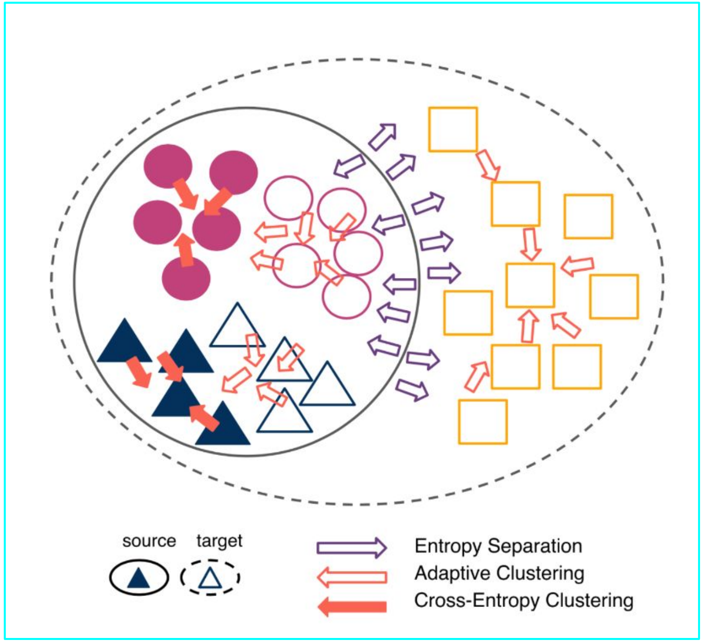
Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

Classification of known classes

Clustering of similar known and unknown samples

Separation of different (anomalous) unknown samples



Output vector **p**

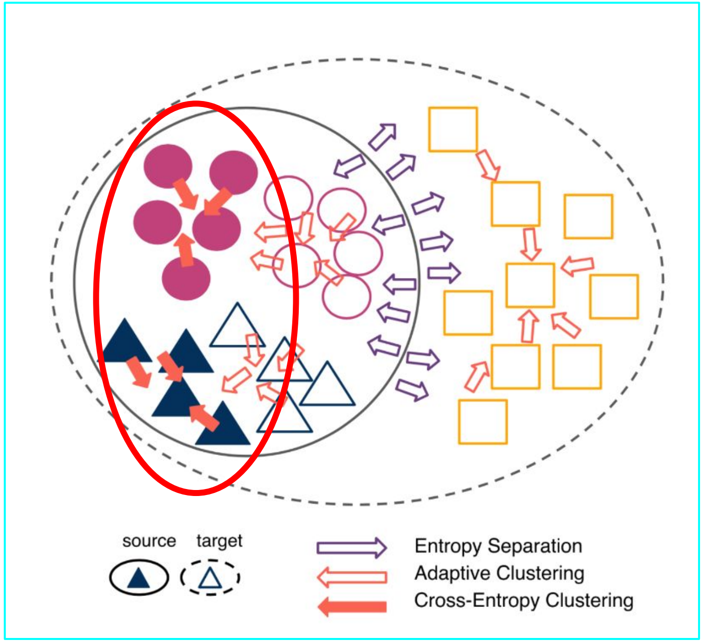
Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

Classification of known classes

$$\mathcal{L}_{\mathrm{CE}} = \frac{-\sum_{k=1}^{\mathrm{K}} w_k y_k \log \hat{y}_k}{\sum_{k=1}^{\mathrm{K}} w_k},$$

Using true and predicted labels



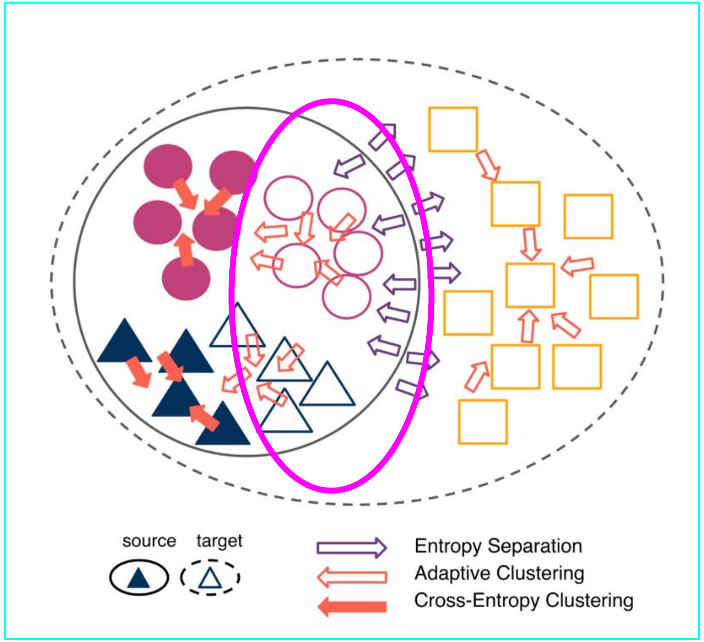
Output vector **p** → compare predicted y’ with true label y

Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

Clustering of similar known and unknown samples

Via self-supervision:
comparing pairs of output features
between all samples from both domains



0 1 2 3 4 5 6 7 8 9

Output vector **p**

Fermilab

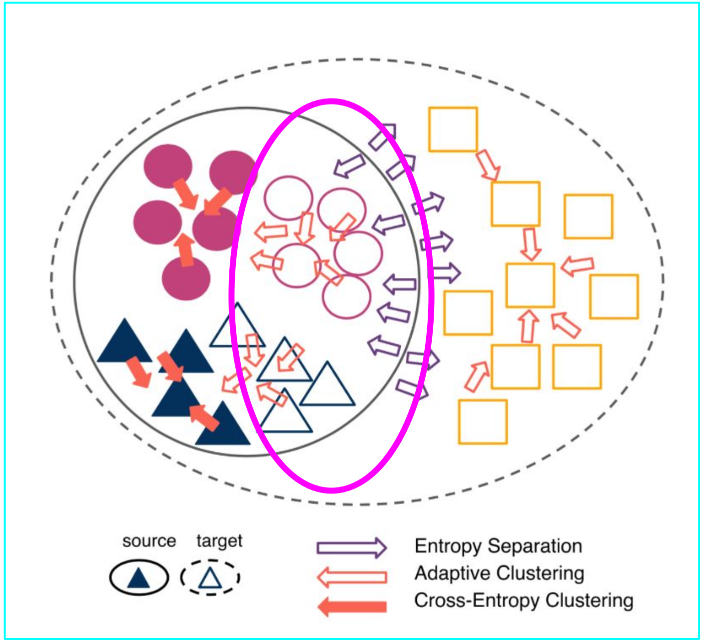
Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

Clustering of similar known and unknown samples

Via self-supervision:
comparing pairs of output features
between all samples from both domains

$$\mathcal{L}_{\mathrm{AC}} = -\sum_{i \in B}\sum_{j \in b_t} s_{ij}\log(\mathbf{p}_i^{\top}\mathbf{p}_j) + (1-s_{ij})\log(1-\mathbf{p}_i^{\top}\mathbf{p}_j), \tag{1}$$



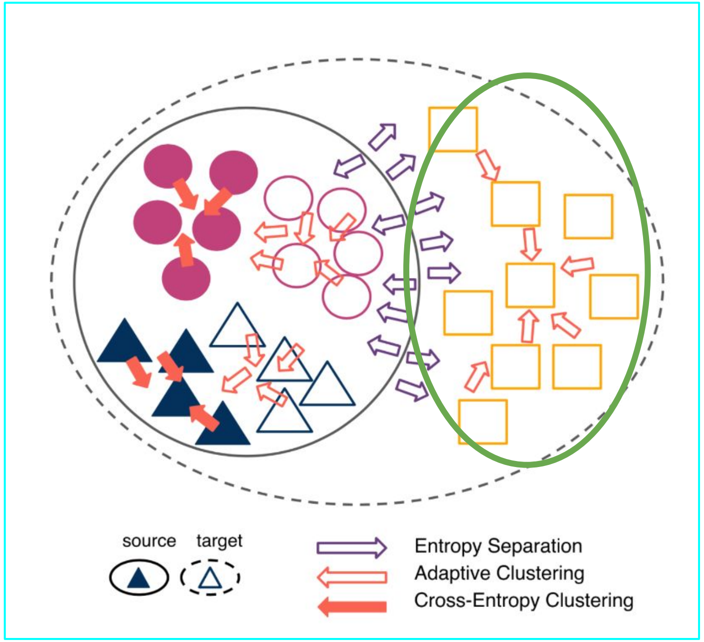
Output vector p → rank order to create similarity labels

Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

Separation of different (anomalous) unknown samples

Pushing away samples with high entropy of outputs features



0 1 2 3 4 5 6 7 8 9

Output vector p

Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

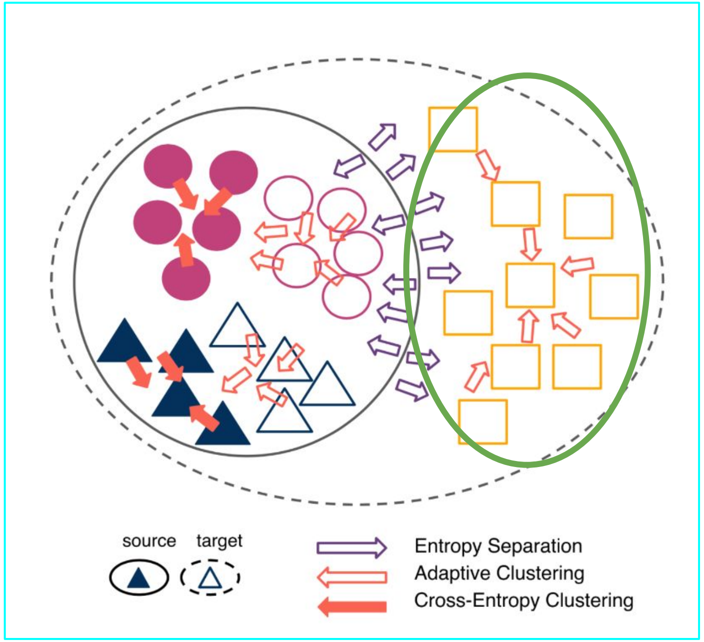
Separation of different (anomalous) unknown samples

Pushing away samples with high entropy of outputs features

$$\mathcal{L}_{\mathrm{ES}}(\mathbf{p}_i) = \begin{cases} -|H(\mathbf{p}_i) - \rho| & |H(\mathbf{p}_i) - \rho| > m, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathcal{L}_{\mathrm{ES}} = \frac{1}{|b_t|} \sum_{i \in b_t} \mathcal{L}_{\mathrm{ES}}(\mathbf{p}_i).$$

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$



0 1 2 3 4 5 6 7 8 9

Output vector p → calculate entropy of each output

Universal Domain Adaptation (DeepAstroUDA)

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

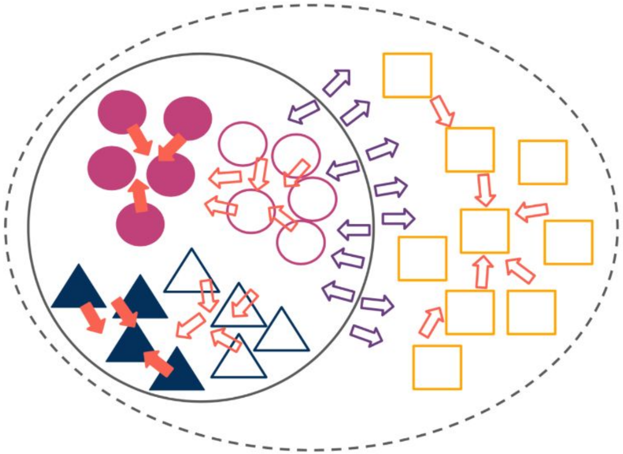
Separation of different (anomalous) unknown samples

Pushing away samples with high entropy of outputs features

$$\mathcal{L}_{\mathrm{ES}}(\mathbf{p}_i) = \begin{cases} -|H(\mathbf{p}_i) - \rho| & |H(\mathbf{p}_i) - \rho| > m, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathcal{L}_{\mathrm{ES}} = \frac{1}{|b_t|} \sum_{i \in b_t} \mathcal{L}_{\mathrm{ES}}(\mathbf{p}_i).$$

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$



$$\mathcal{L} = \mathcal{L}_{CE} + \lambda(\mathcal{L}_{\mathrm{AC}} + \mathcal{L}_{\mathrm{ES}}).$$

0 1 2 3 4 5 6 7 8 9

Output vector p → calculate entropy of each output

Universal Domain Adaptation (DeepAstroUDA)

DA tests we ran:

- Two data releases from the same telescope
  - LSST mocks Y1 and Y10
- Different surveys
  - SDSS and DECaLS
- Wide and deep fields in the same survey
  - SDSS wide and Stripe 82 deep field

Universal Domain Adaptation (DeepAstroUDA)

DA tests we ran:

- Two data releases from the same telescope
  - LSST mocks Y1 and Y10
- Different surveys
  - SDSS and DECaLS
- Wide and deep fields in the same survey
  - SDSS wide and Stripe 82 deep field

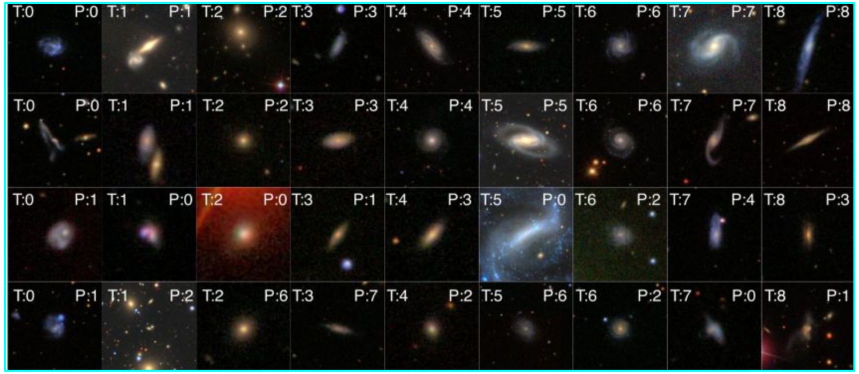
Class labels are from Galaxy Zoo 2 & 3 (crowdsourcing labels ~10^5 volunteers).

Known classes:
Disturbed (0)
Merging (1)
Round smooth (2)
Cigar shaped smooth (3)
Barred spiral (4)
Unbarred tight spiral (5),
Unbarred loose spiral (6)
Edge-on without bulge (7),
Edge-on with bulge (8),

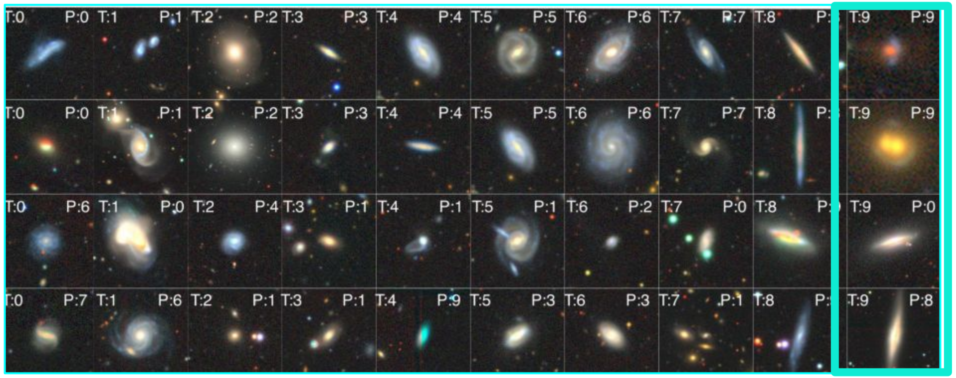
Unknown anomaly class (only in DECaLS):
Strong gravitational lens (9)

Universal Domain Adaptation (DeepAstroUDA)

SDSS



DECaLS

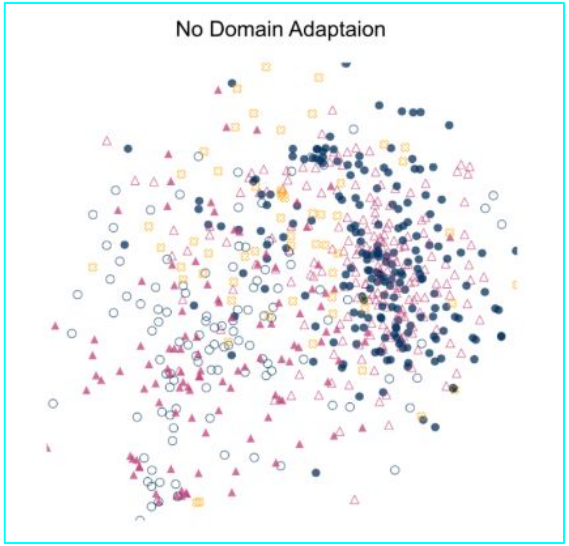


Class labels are from Galaxy Zoo 2 & 3 (crowdsourcing labels ~10^5 volunteers).

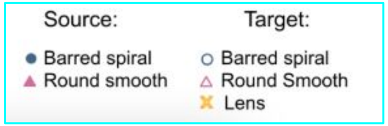
Known classes:
Disturbed (0)
Merging (1)
Round smooth (2)
Cigar shaped smooth (3)
Barred spiral (4)
Unbarred tight spiral (5),
Unbarred loose spiral (6)
Edge-on without bulge (7),
Edge-on with bulge (8),

Unknown anomaly class (only in DECaLS):
Strong gravitational lens (9)

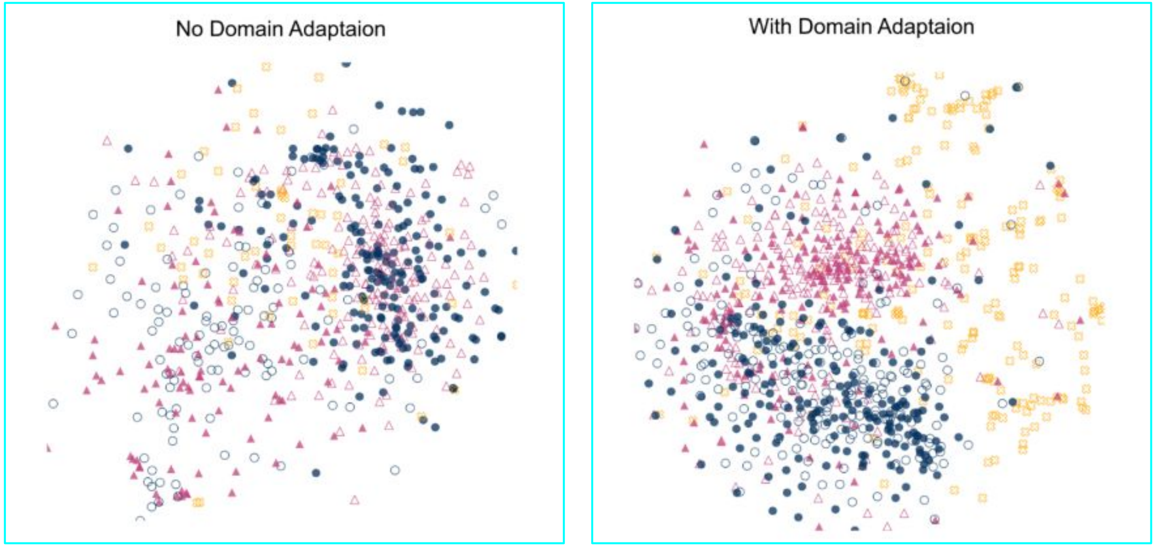
Universal Domain Adaptation (DeepAstroUDA)



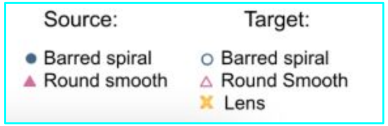
Classes are mixed!



Universal Domain Adaptation (DeepAstroUDA)

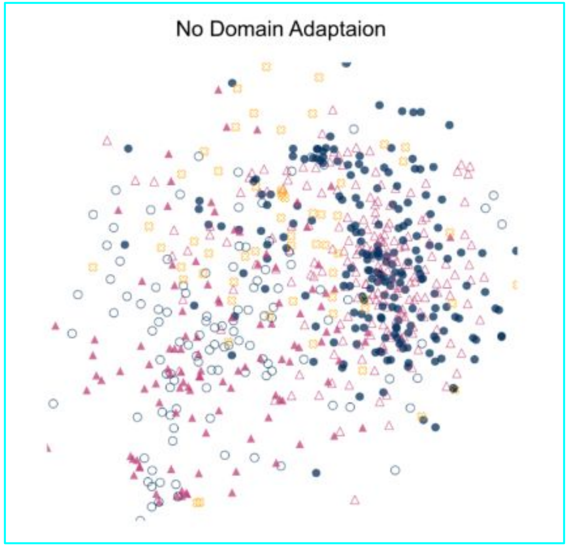


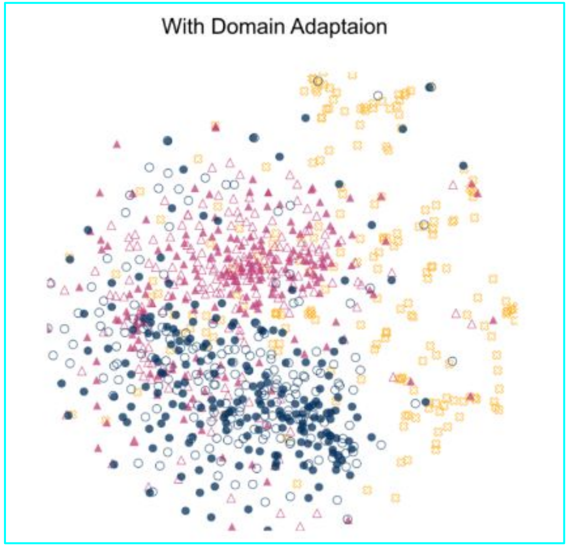
Classes are mixed!

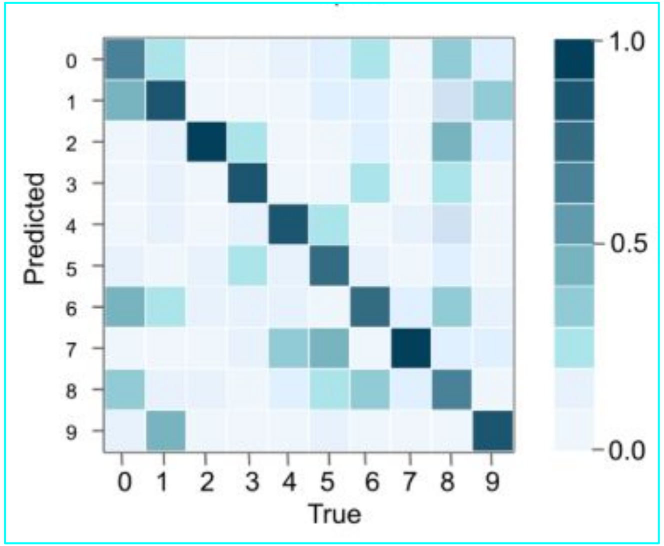


Known classes overlap,
unknown is pushed to the side.

Universal Domain Adaptation (DeepAstroUDA)





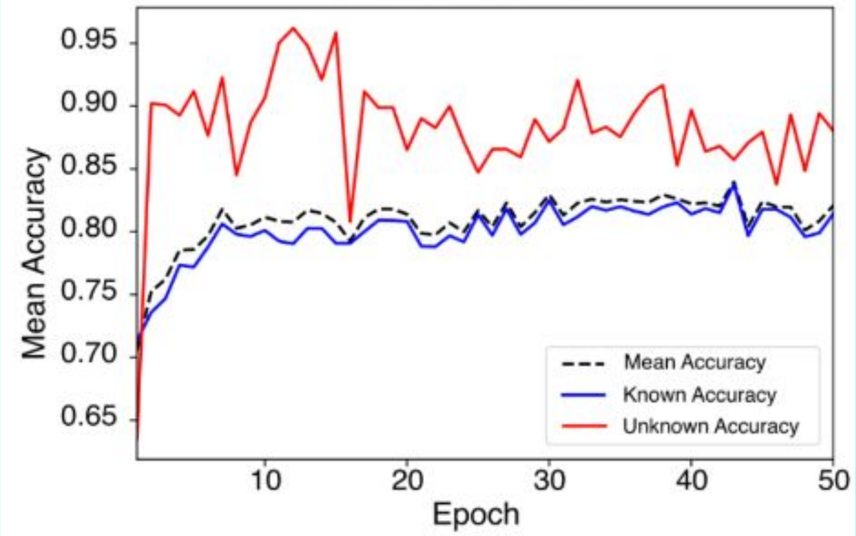


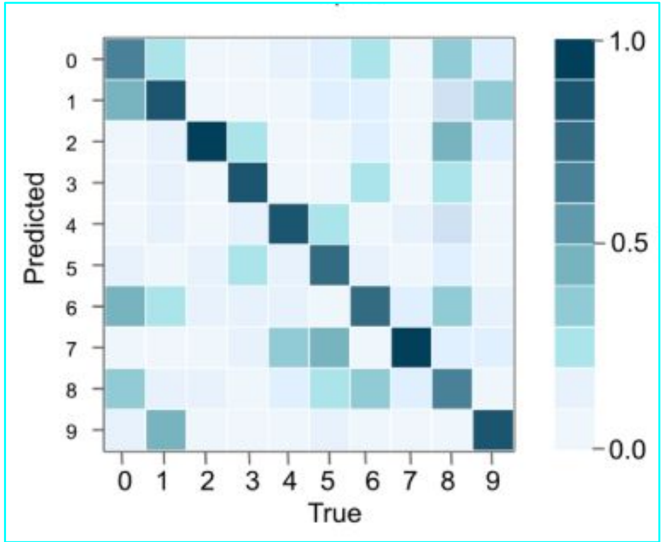
Classes are mixed!

Source:	Target:
● Barred spiral	○ Barred spiral
▲ Round smooth	△ Round Smooth
	✕ Lens

Known classes overlap,
unknown is pushed to the side.

Universal Domain Adaptation (DeepAstroUDA)

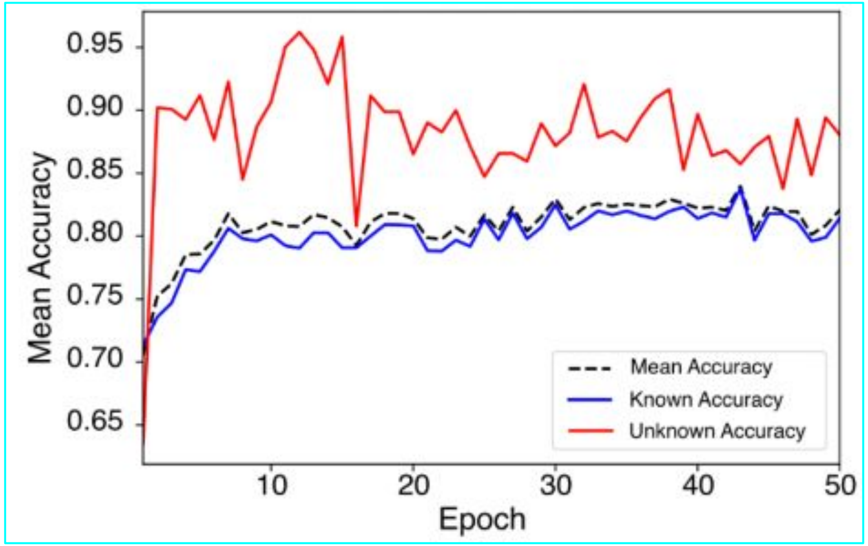


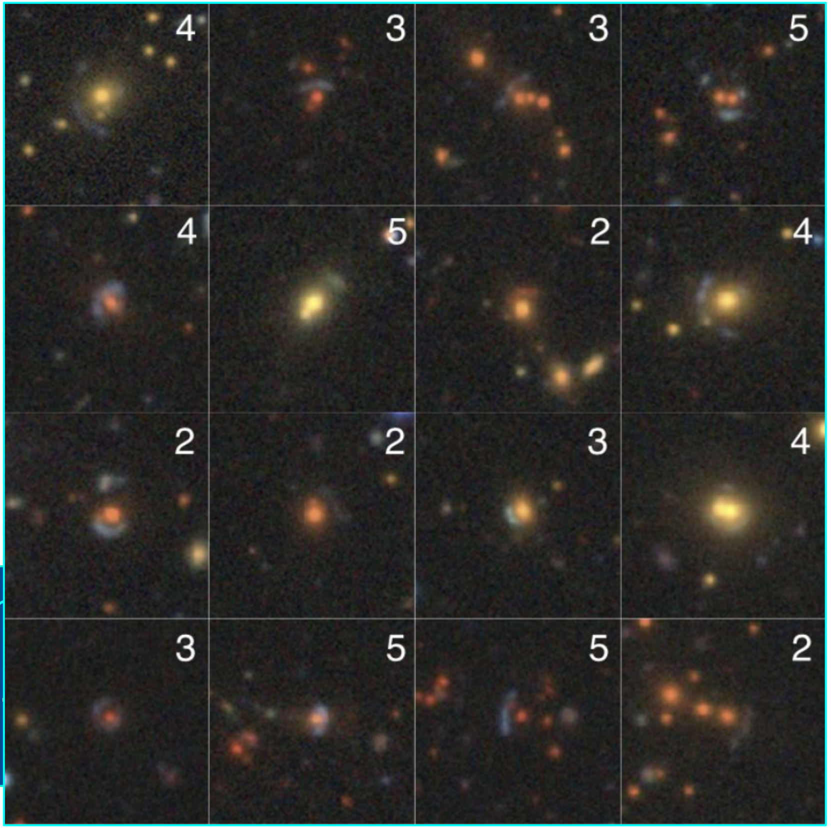


- Most confusion between classes is for truly morphologically similar classes, like disturbed and merging.
- Model is very sure about the unknown lens class - it can recognize these object look different than all other known classes.

Fermilab

Universal Domain Adaptation (DeepAstroUDA)





- Most confusion between classes is for truly morph
  and merging.
- Model is very sure about the unknown lens class
  than all other known classes.

Fermilab

Talk Outline

1. Domain Shift Problems

2. Domain Adaptation

3. Universal Domain Adaptation

4. Future Challenges

- Simulation and observations

Ćiprijanović et al. 2021.

- Increase robustness to data perturbations

Ćiprijanović et al. 2022.

- Different data releases from the same survey
- Different surveys
- Wide and deep fields of the same survey

Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

- Simulation and observations

Ćiprijanović et al. 2021.

- Increase robustness to data perturbations

Ćiprijanović et al. 2022.

- Different data releases from the same survey
- Different surveys
- Wide and deep fields of the same survey

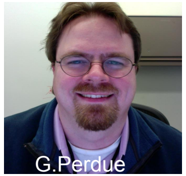
Ćiprijanović et al. 2022.
Ćiprijanović et al. 2023.

FUTURE CHALLENGES

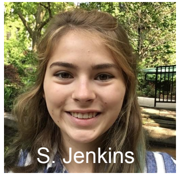
- Connecting extracted features to physical properties.
- Guiding the model to use some preferred physical features and discover the rest.
- Understanding and exploring the latent space.
- Can we get any new insights from AI?
- What if the domain shift is physical not instrumental/computational?

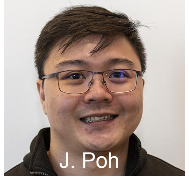
Big thanks to all my amazing collaborators

Fermilab

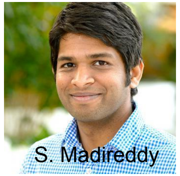


University of Chicago



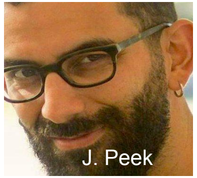


Argonne, Oakridge



and many more!

Space Telescope Science Institute



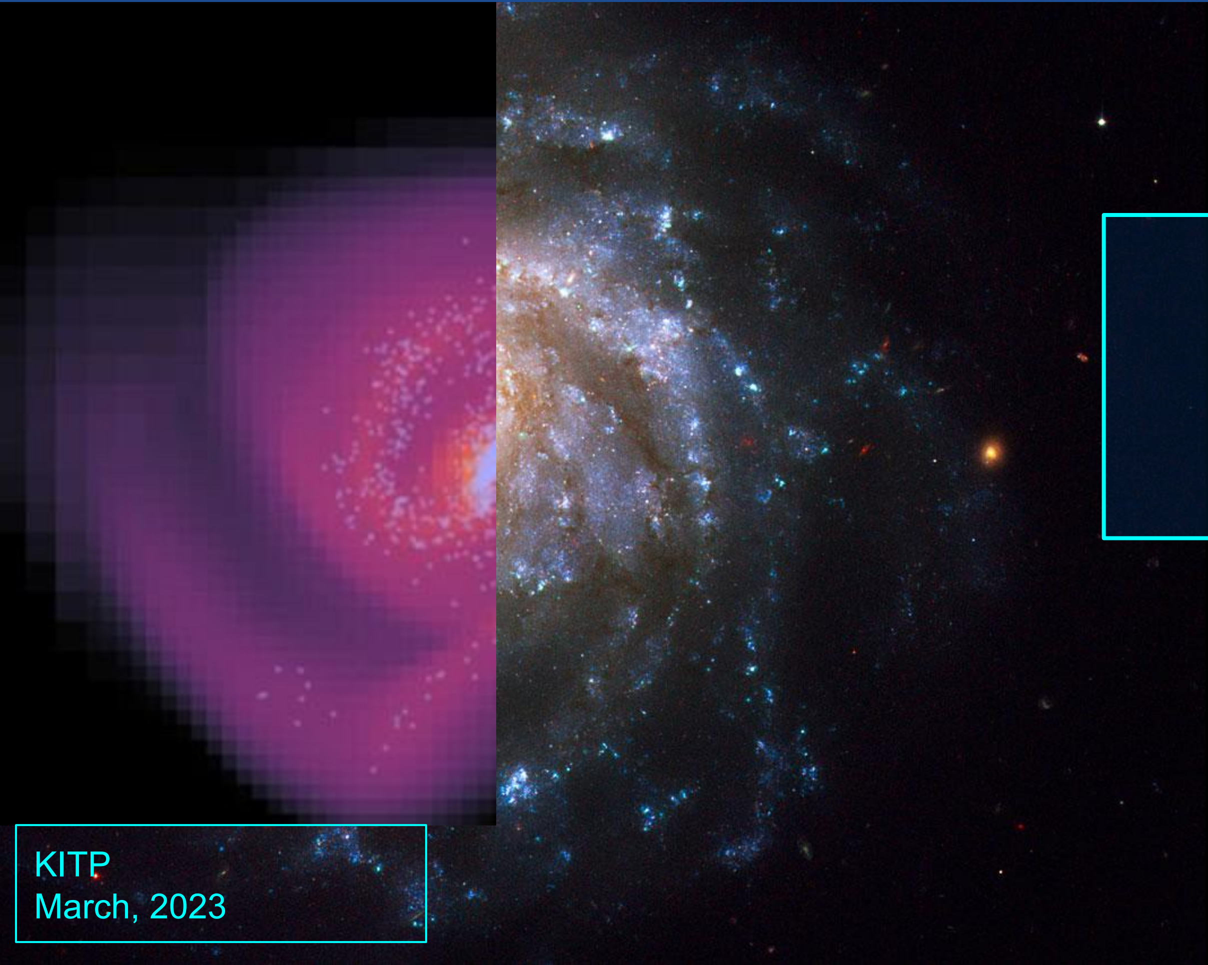
THE UNIVERSITY OF CHICAGO 1890

Argonne NATIONAL LABORATORY

STScI

OAK RIDGE National Laboratory

Fermilab



THANK YOU!

KITP
March, 2023

Aleksandra Ćiprijanović
(she/her/hers)

Fermilab, DSSL
aleksand@fnal.gov