# Domain Shift Problems in Astrophysics:

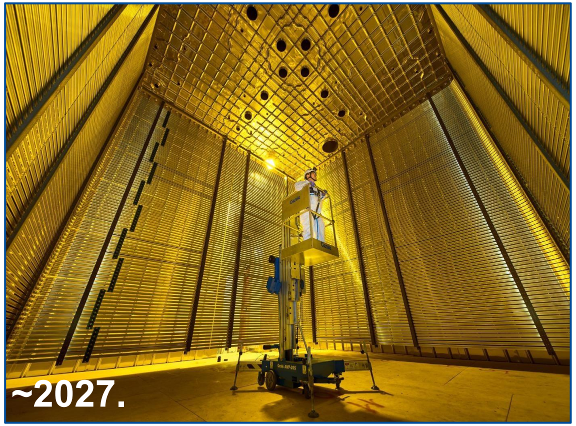
## Bridging the gap between simulated and real data with AI


Aleksandra Ćiprijanović
(she/her/hers)

Fermilab, DSSL
UChicago, Astron. & Astrophys.
aleksand@fnal.gov


MATF Seminar
April, 2023

# Vision of the Future



## Rubin LSST

~ 20 TB / day
~ 100 PB total by DR11

## DUNE

~ 30-60 PB / year (raw)
~ 114x4 TB / month (raw) for Supernovae detection (speed need for follow ups)

## HL-LHC

~ order of magnitude more data
~ 650 PB / year

# 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

**DUNE**

~ 30-60 PB / year (raw)
~ 114x4 TB / month (raw) for Supernovae detection (speed need for follow ups)

**HL-LHC**

~ order of magnitude more data
~ 650 PB / year



Fermilab

# Vision of the Future




**Rubin LSST**

~ 20 TB / day
~ 100 PB total by DR11

**DUNE**

~ 30-60 PB / year (raw)
~ 114x4 TB / month (raw) for Supernovae detection (speed need for follow ups)

**HL-LHC**

~ order of magnitude more data
~ 650 PB / year

# Talk Outline

Domain Shift Problems

Domain Adaptation

Model Robustness

Universal Domain Adaptation

# Talk Outline

Domain Shift Problems

Domain Adaptation

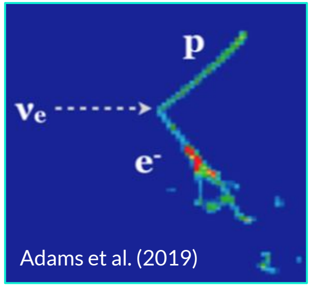
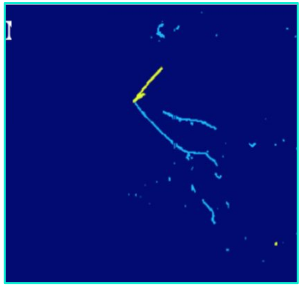
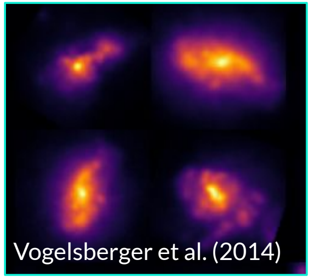
Model Robustness

Universal Domain Adaptation

# 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) |  |  |
| Illustris / Hubble (merging galaxies) |  |  |









Fermilab

# Combining Datasets

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

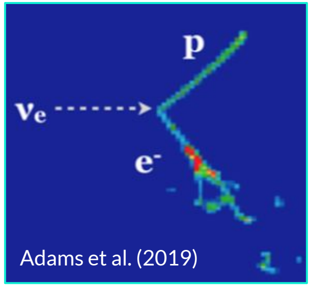
**DATASET SHIFT**

Missing and unknown physics, wrong geometry, background levels

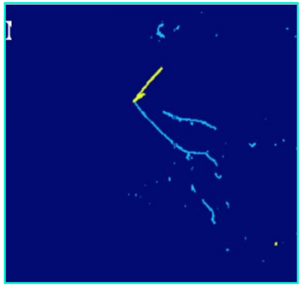
Computational constraints for simulations

SIMULATED | REAL

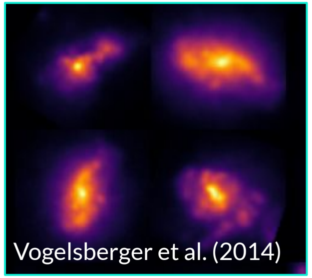
MicroBooNE (neutrinos)



Adams et al. (2019)



Illustris / Hubble (merging galaxies)


Vogelsberger et al. (2014)

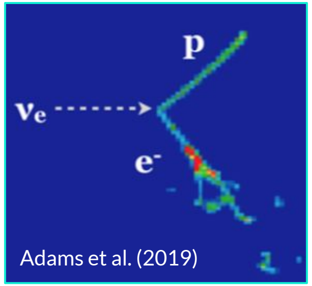
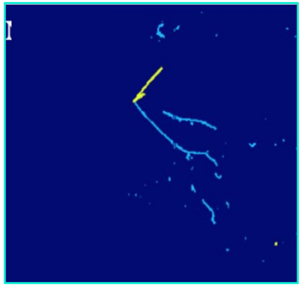
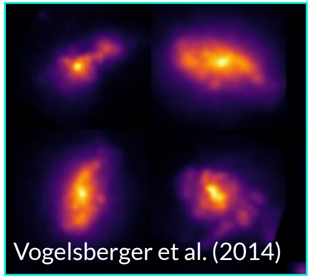

Hubble

# 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) |  |  |
| Illustris / Hubble (merging galaxies) |  |  |



Adams et al. (2019)

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**!

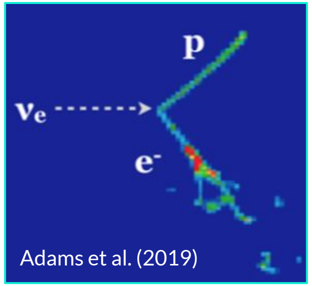
**DATASET SHIFT**

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

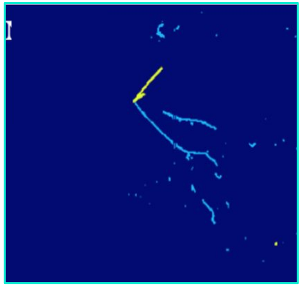
SIMULATED

REAL

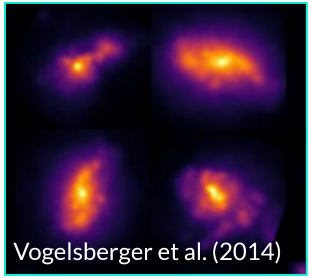
MicroBooNE (neutrinos)



Adams et al. (2019)



Illustris / Hubble (merging galaxies)

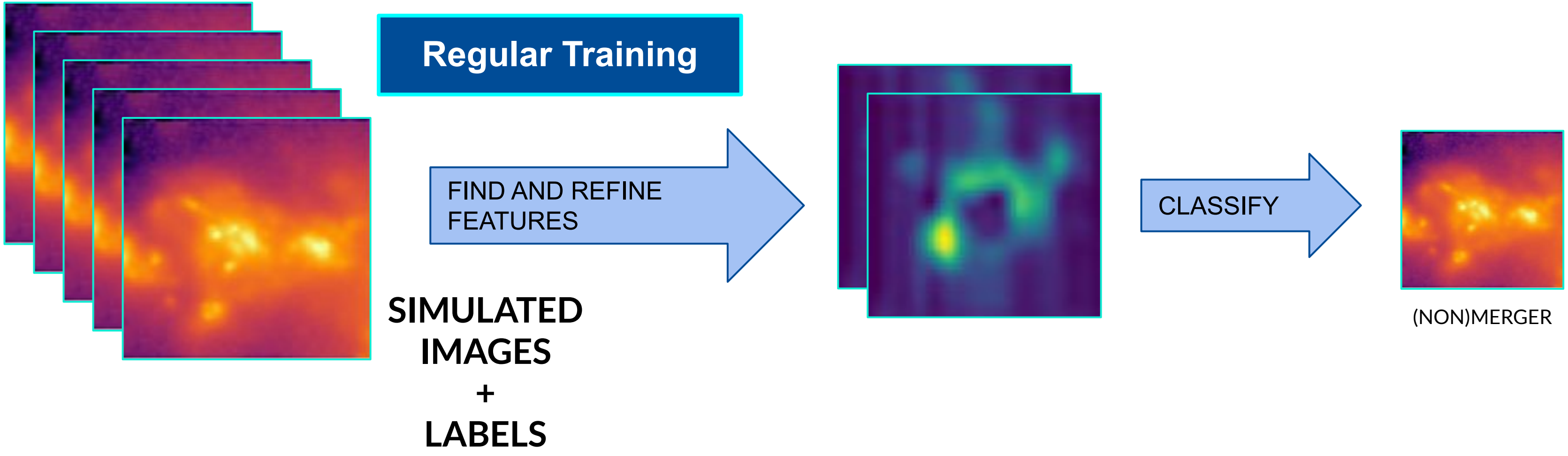

Vogelsberger et al. (2014)


Hubble



Fermilab

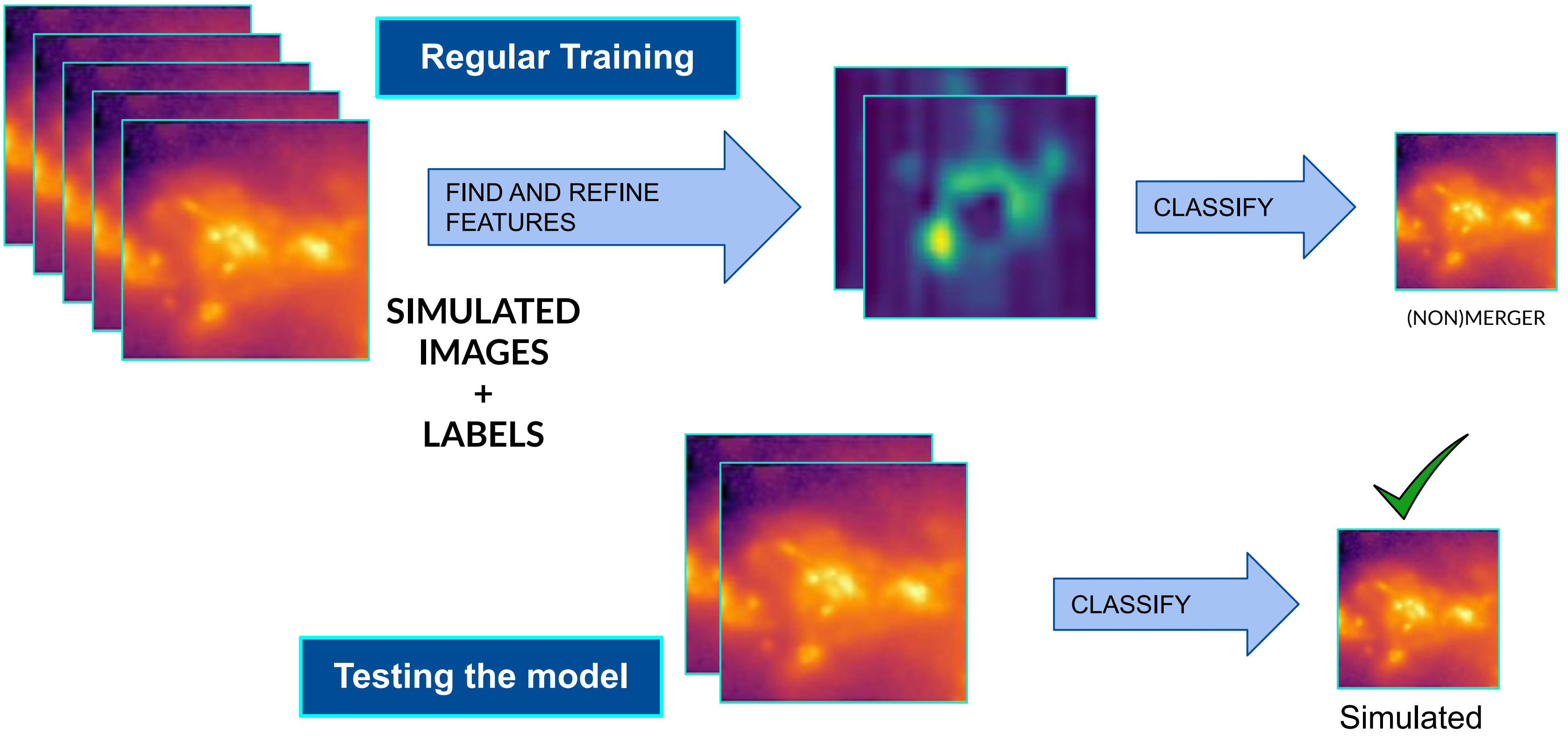
# Combining Datasets



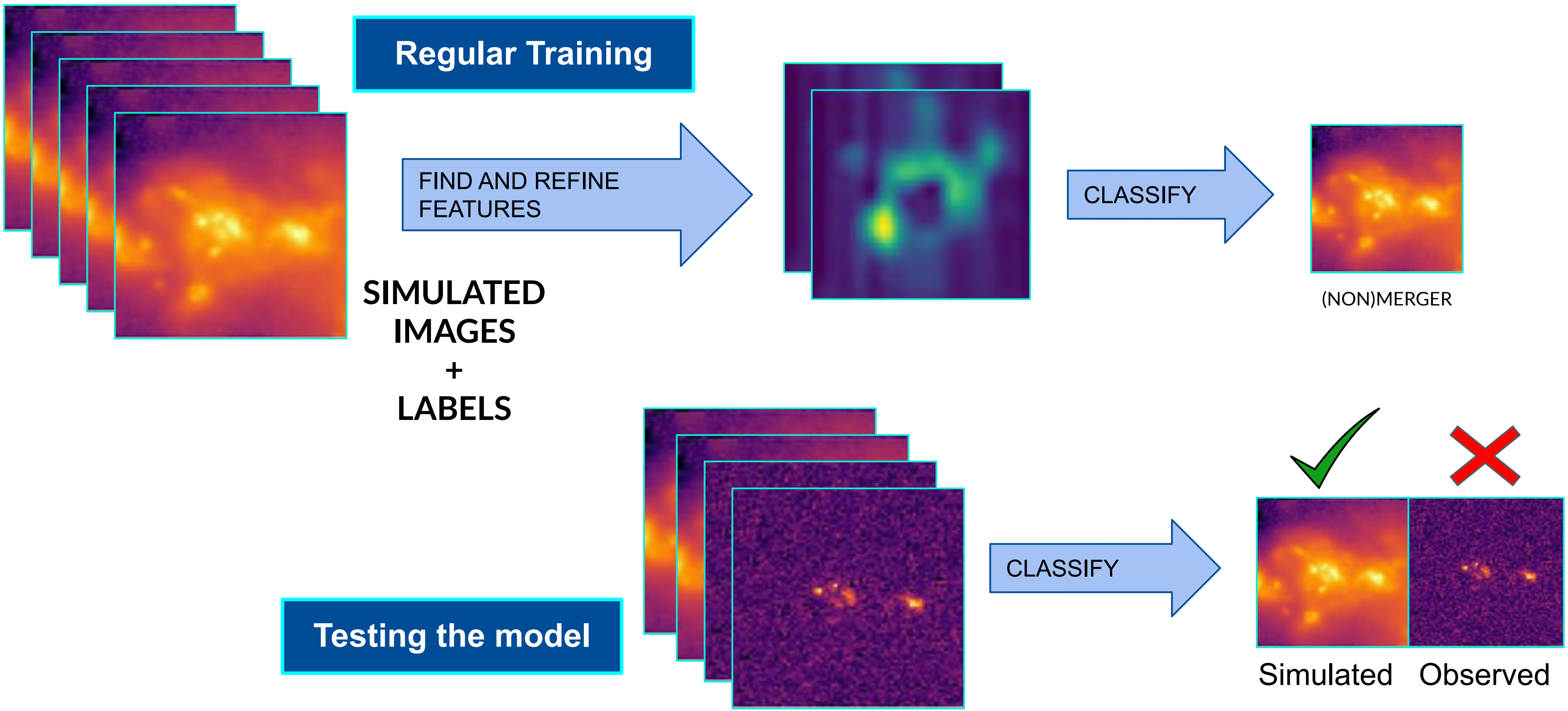


Fermilab

# Combining Datasets



# Combining Datasets



# Combining Datasets

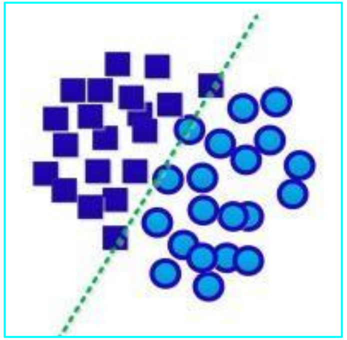
**Why does this happen?**

# Combining Datasets

## Why does this happen?

Source Domain

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





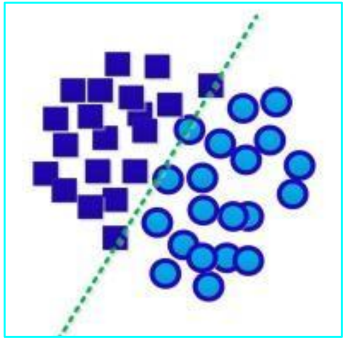
Fermilab

# Combining Datasets

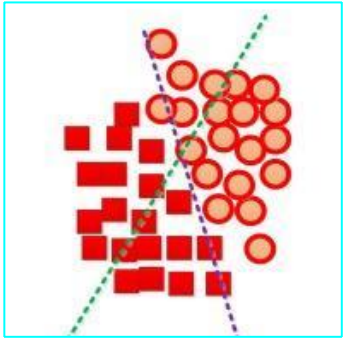
## Why does this happen?

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

Source Domain



Target Domain

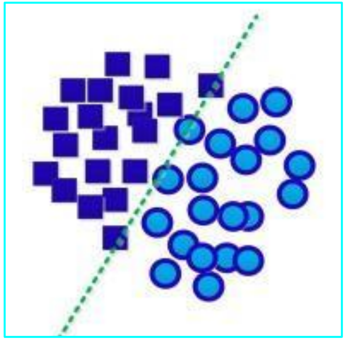


# Combining Datasets

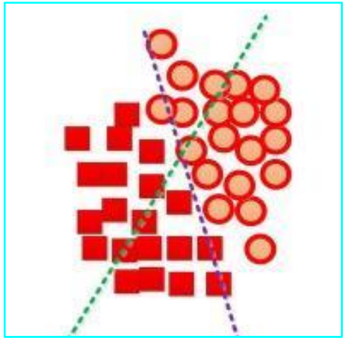
## Why does this happen?

We need to align the data during training!

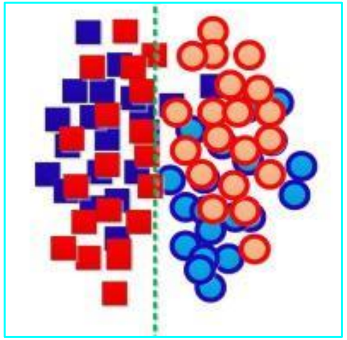
Source Domain



Target Domain



Domain Alignment



# Talk Outline

1. Domain Shift Problems
2. Domain Adaptation
3. Model Robustness
4. Universal Domain Adaptation

# 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.**

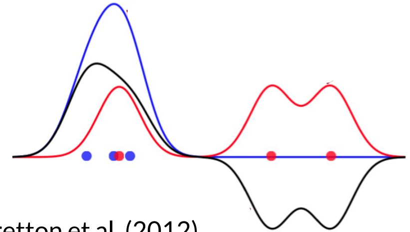
# Combining Datasets

## DOMAIN ADAPTATION

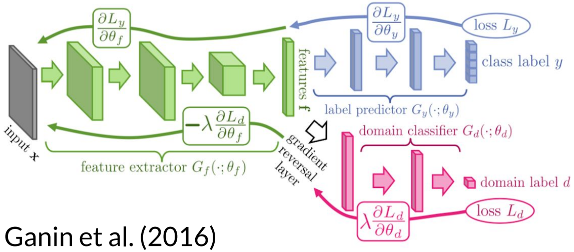
Align data distributions in the latent space of the network by forcing the network to **find more robust domain-invariant features.**

**Distance-based methods**

**Adversarial methods**



Gretton et al. (2012)




Ganin et al. (2016)

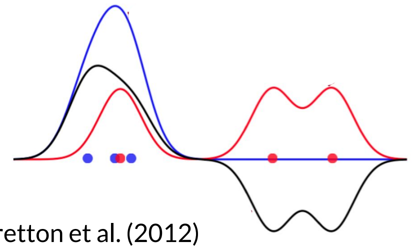
# Combining Datasets

**DOMAIN ADAPTATION**

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

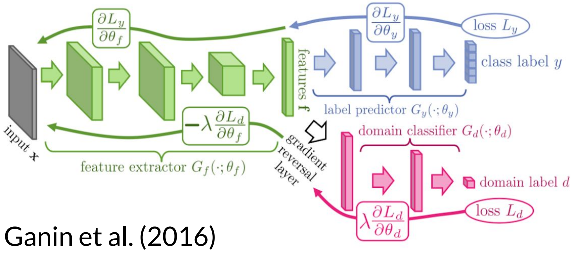
**Distance-based methods**

**Adversarial methods**



Gretton et al. (2012)

**Training = Task Loss + DA Loss**




Ganin et al. (2016)



Fermilab

# 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.**

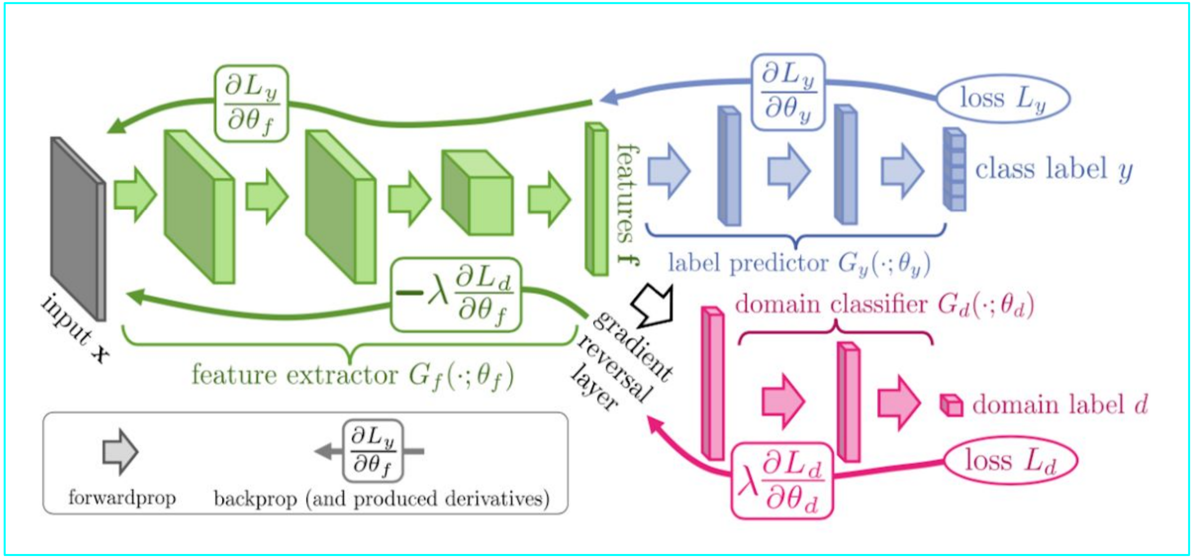
- Distance-based methods
- Adversarial methods

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

# Domain Adversarial Neural Networks - DANNs

DANN - feature extractor + label predictor + domain classifier

- **Gradient reversal layer** - multiplies the gradient by a negative constant during the backpropagation.
- Results in the extraction of **domain-invariant features**.
- Only source domain images are labeled during training.

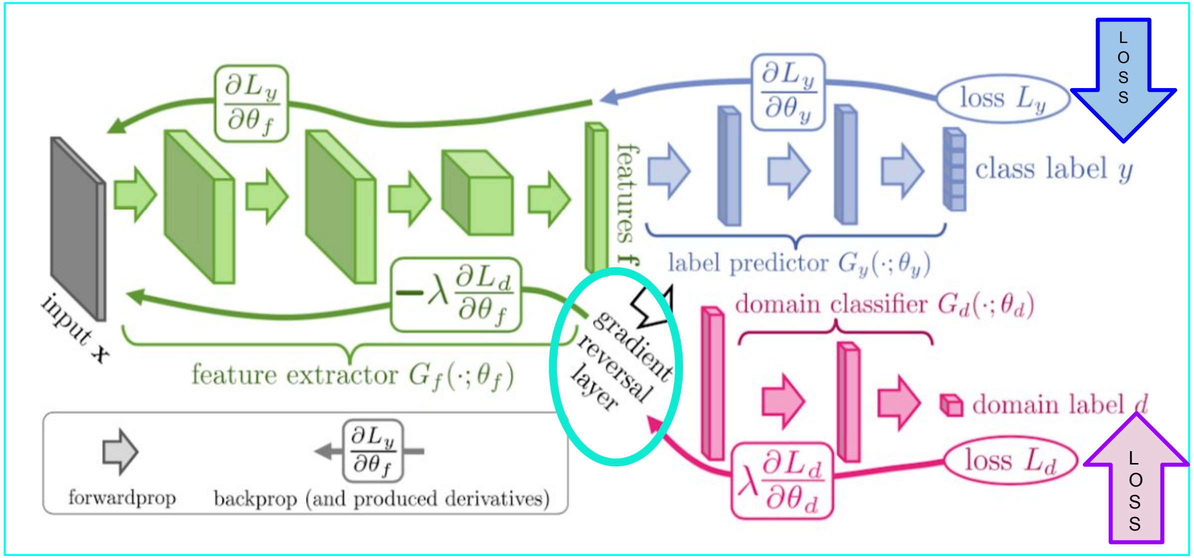



Ganin et al. (2016)

# Domain Adversarial Neural Networks - DANNs

DANN - feature extractor + label predictor + domain classifier

- **Gradient reversal layer** - multiplies the gradient by a negative constant during the backpropagation.
- Results in the extraction of **domain-invariant features**.
- Only source domain images are labeled during training.



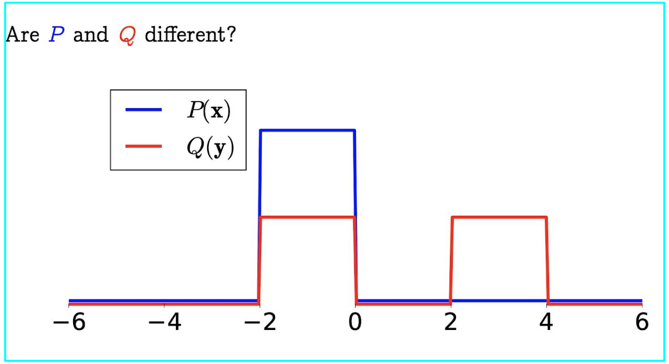

Ganin et al. (2016)

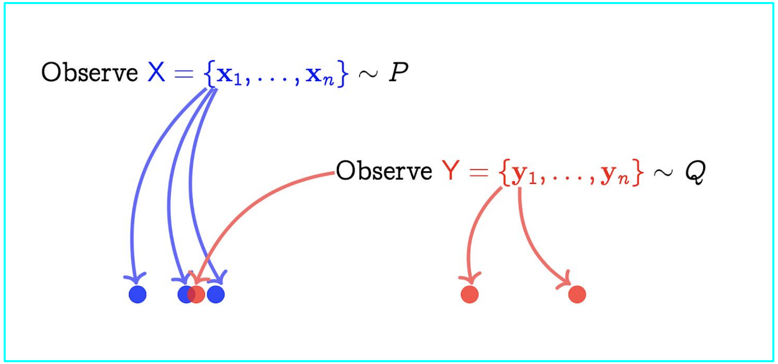


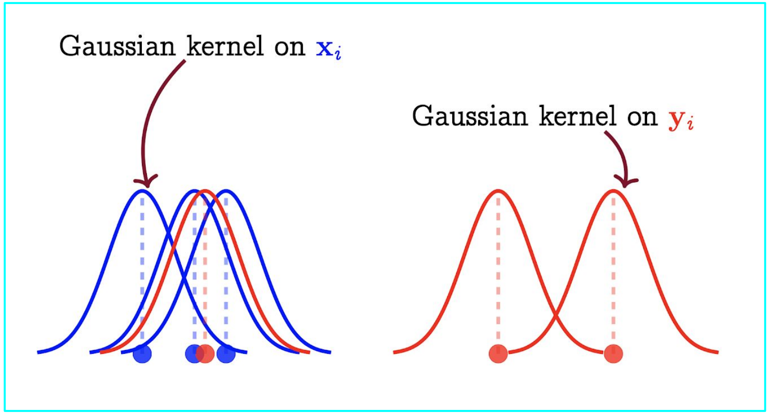
Fermilab

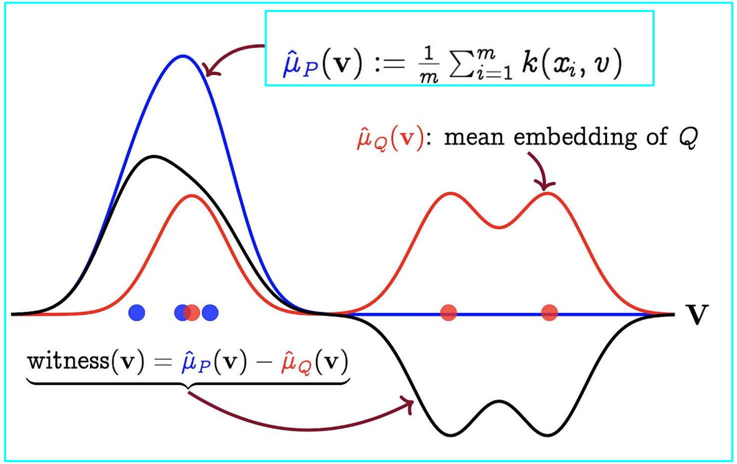
# Maximum Mean Discrepancy - MMD

Smola et al. (2007)
Gretton et al. (2012)



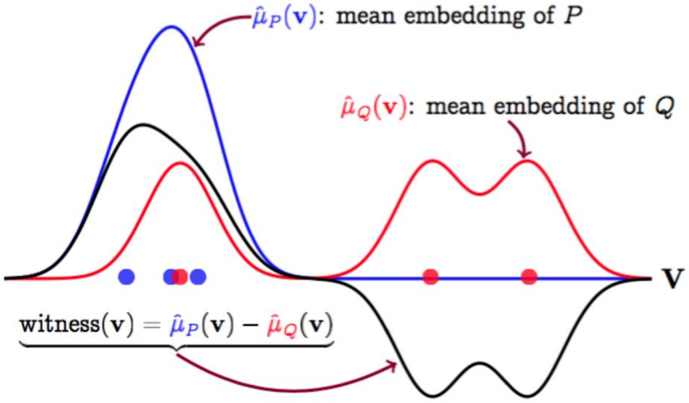



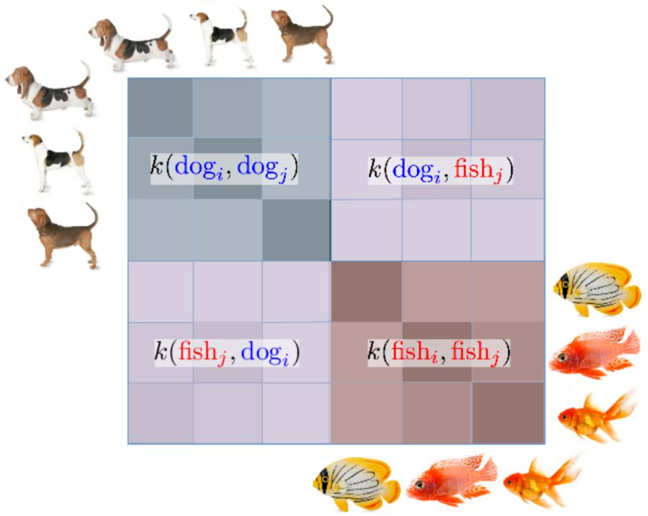
# Maximum Mean Discrepancy - MMD

Smola et al. (2007)
Gretton et al. (2012)




$$\widehat{MMD}^2 = ||\text{witness}(\mathbf{v})||^2_{\mathcal{F}}$$
$$= \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{dog}_i, \text{dog}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\text{fish}_i, \text{fish}_j) - \frac{2}{n^2} \sum_{i,j} k(\text{dog}_i, \text{fish}_j)$$

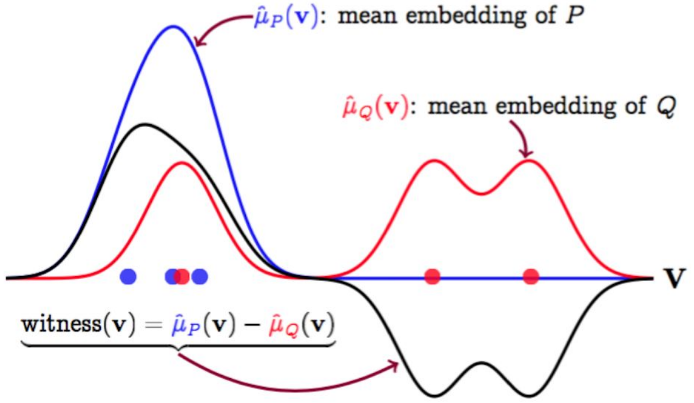





Fermilab

# Maximum Mean Discrepancy - MMD

Smola et al. (2007)
Gretton et al. (2012)

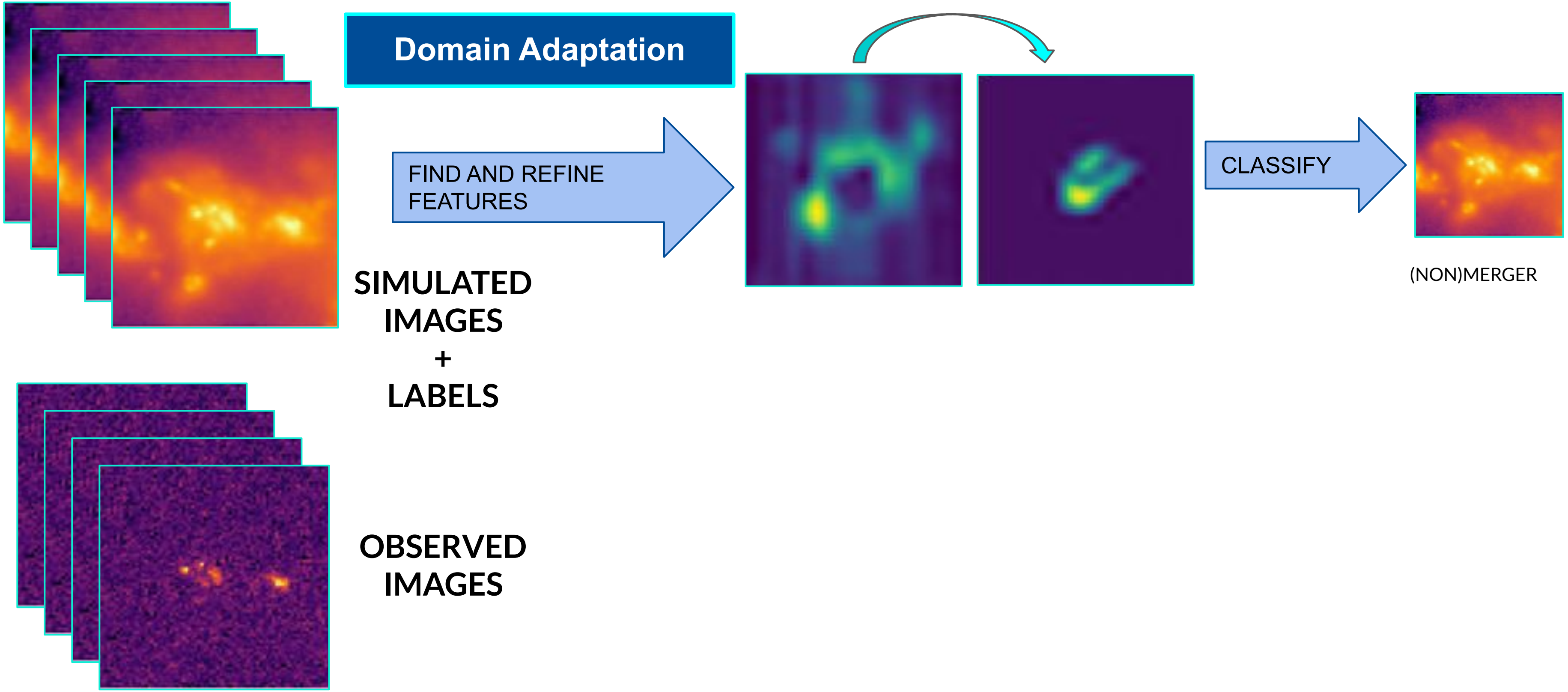



$$\widehat{MMD}^2 = ||\text{witness}(\mathbf{v})||^2_{\mathcal{F}}$$
$$= \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j)$$
$$- \frac{2}{n^2} \sum_{i,j} k(x_i, y_j)$$

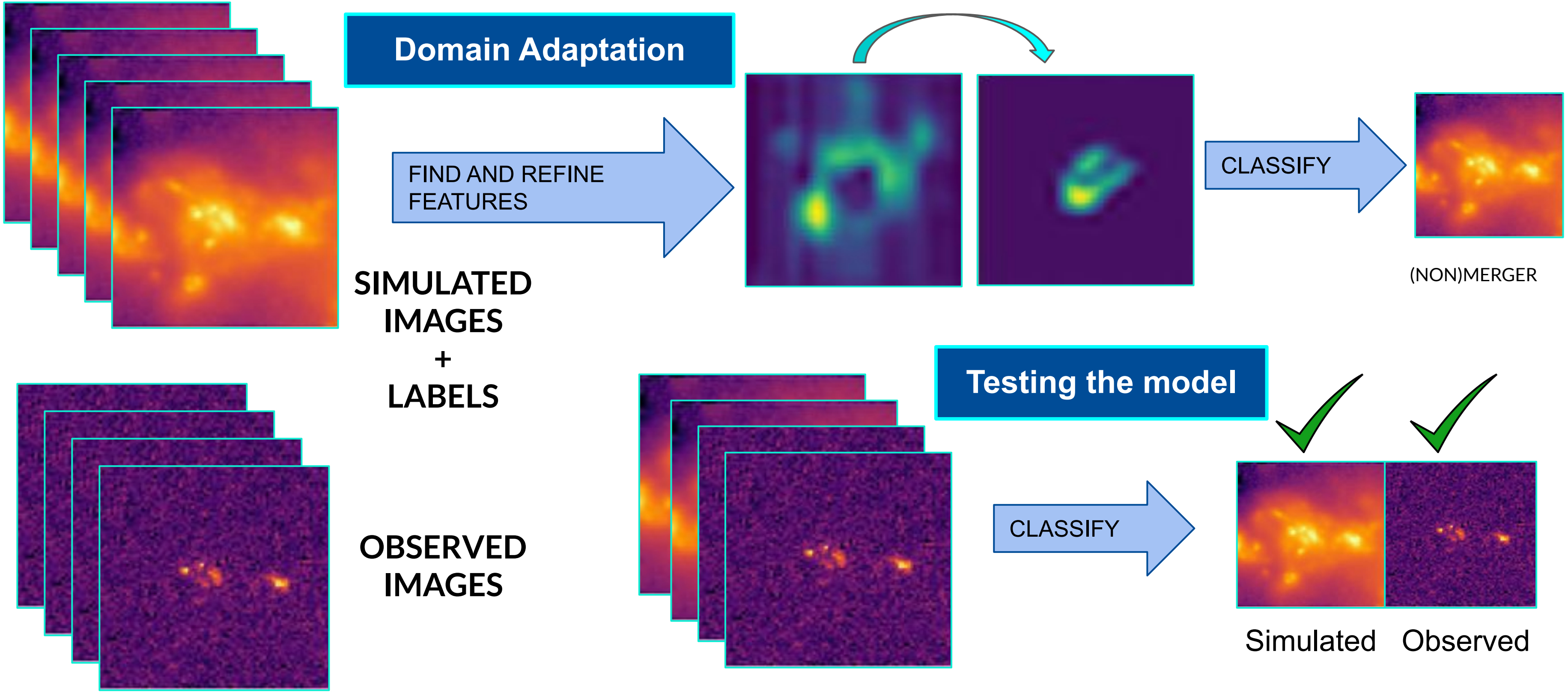


Fermilab

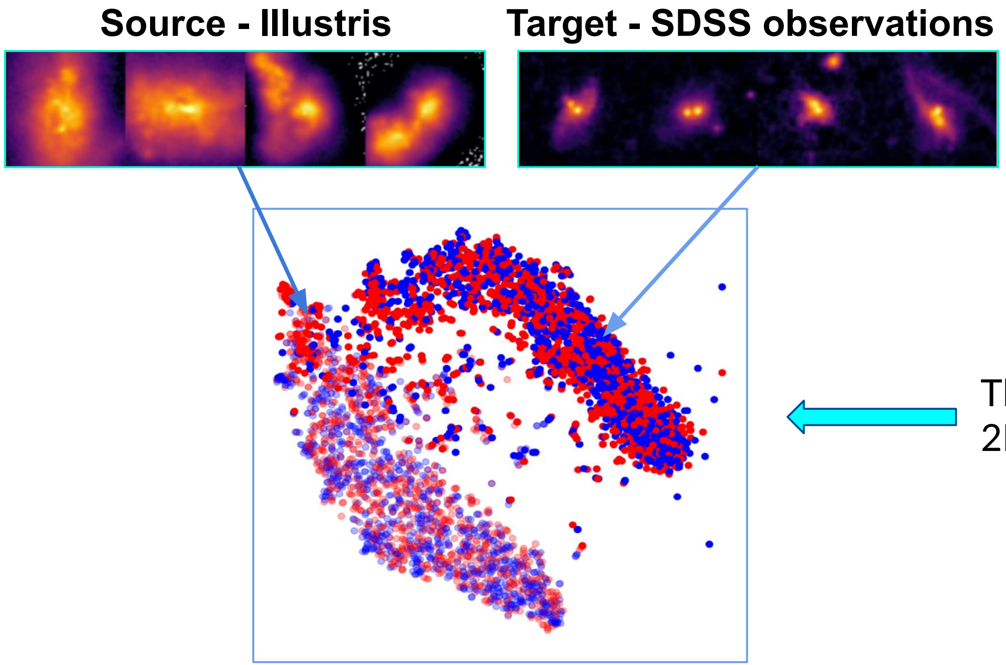
# Combining Datasets



# Combining Datasets



# Combining Datasets



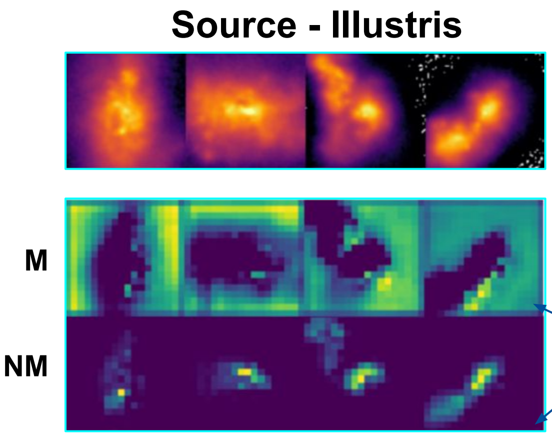

This is how the network sees the data.
2D representation of network's latent space.

Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.



Fermilab

# Combining Datasets



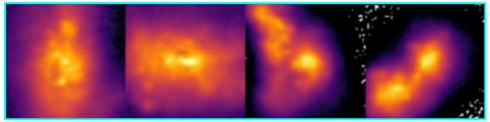

Important regions are highlighted!

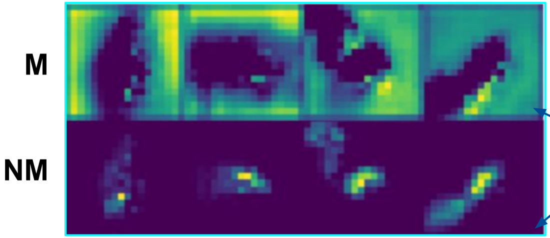
Regular Training

Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

# Combining Datasets

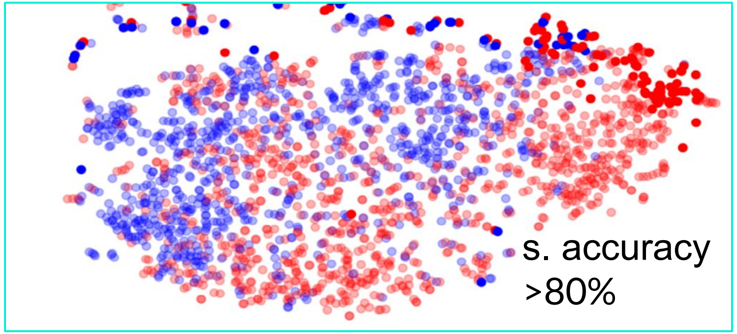
**Source - Illustris**






Important regions are highlighted!

Regular Training



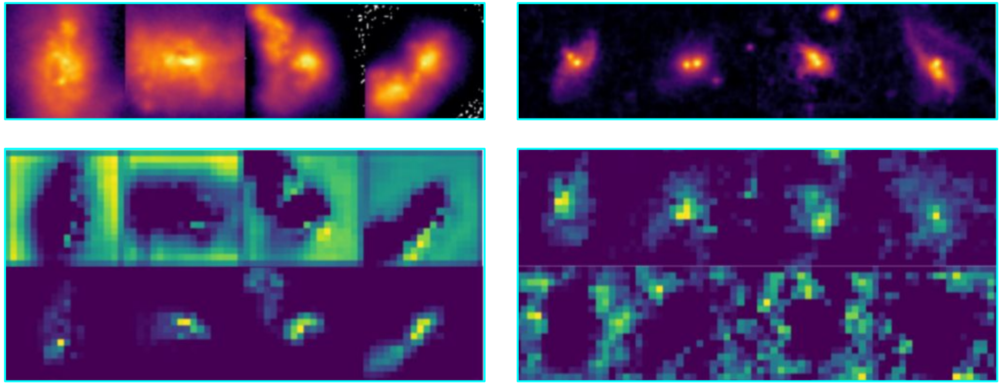

Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

# Combining Datasets

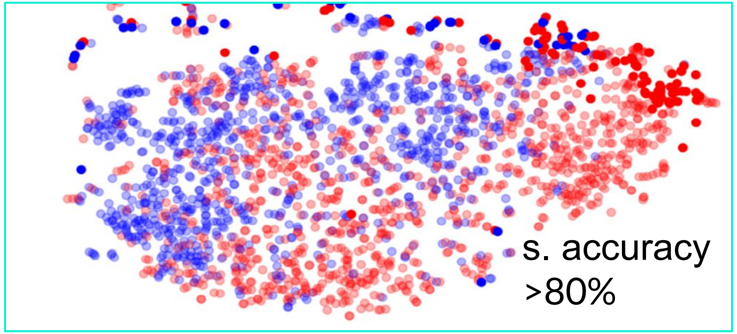
**Source - Illustris** **Target - SDSS observations**

M

NM

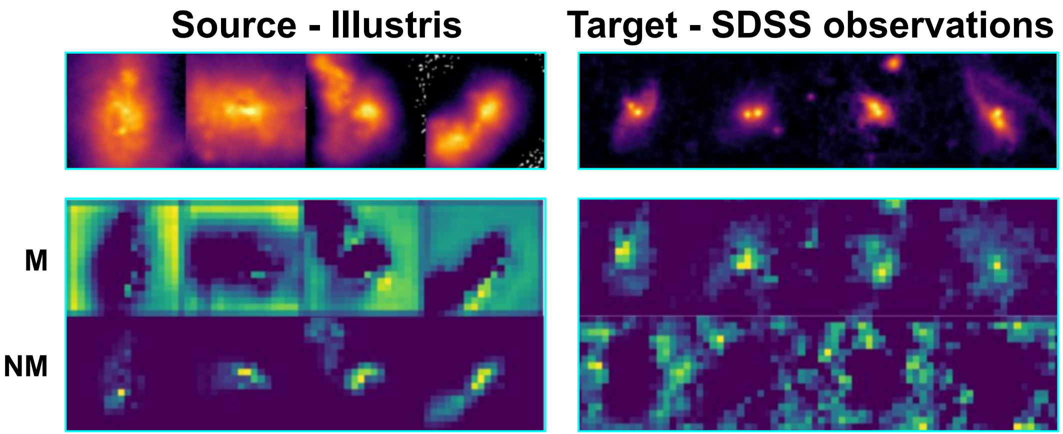


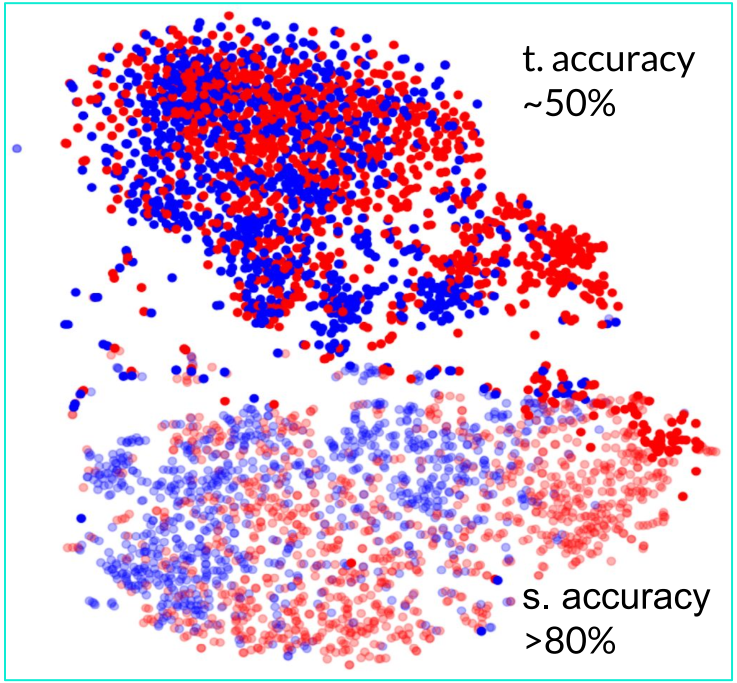
Regular Training




Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

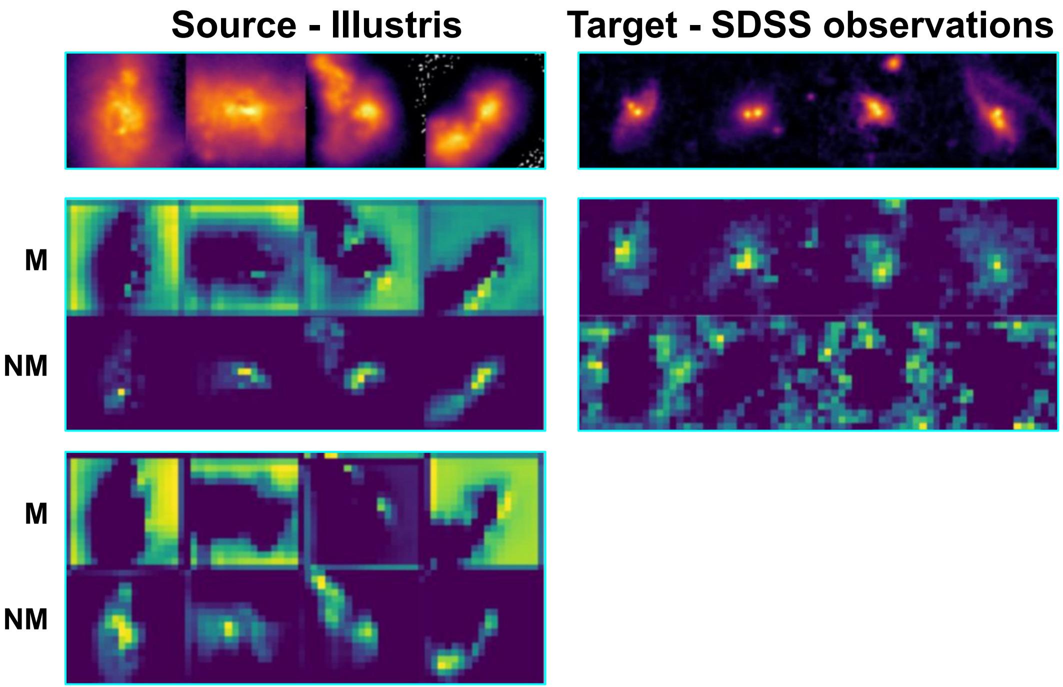
# Combining Datasets




Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

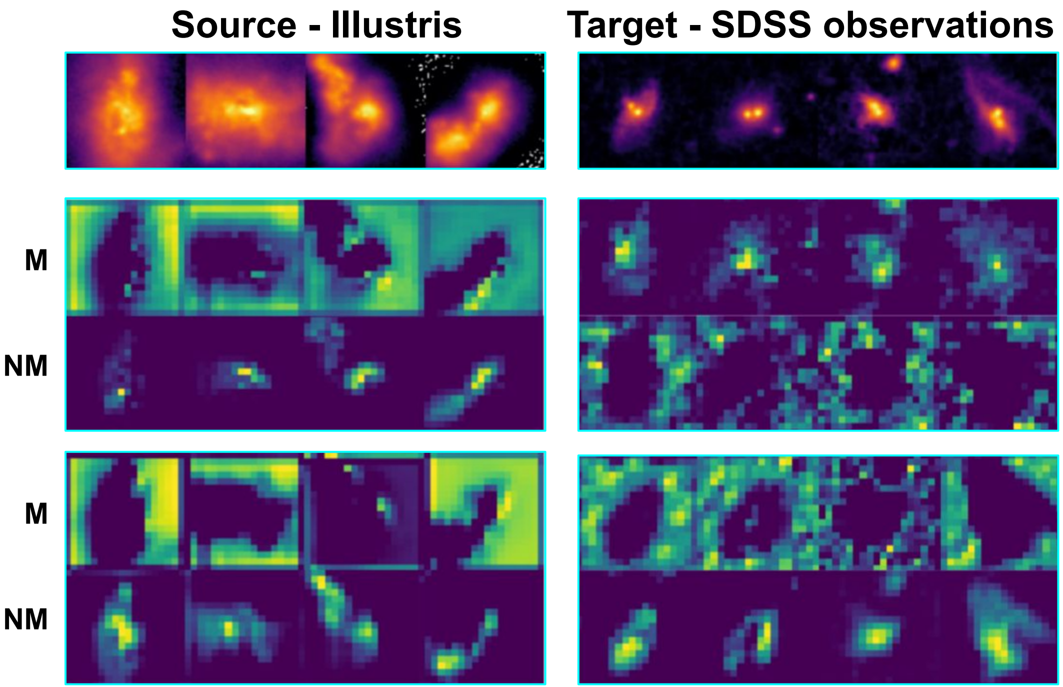
# Combining Datasets




Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

Domain Adaptation

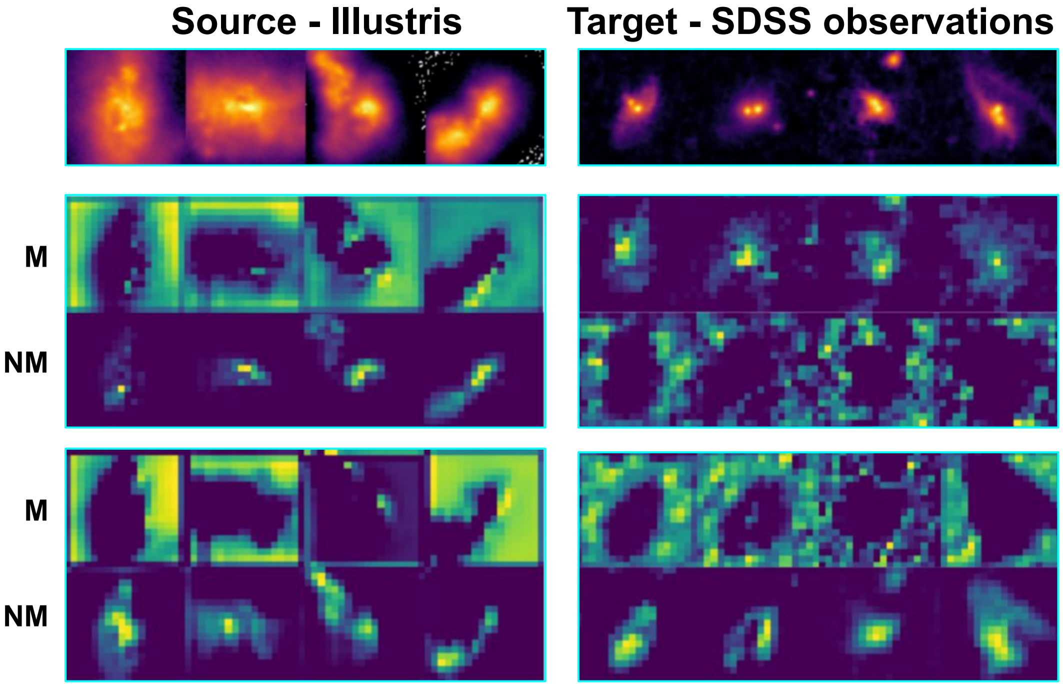
# Combining Datasets




Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

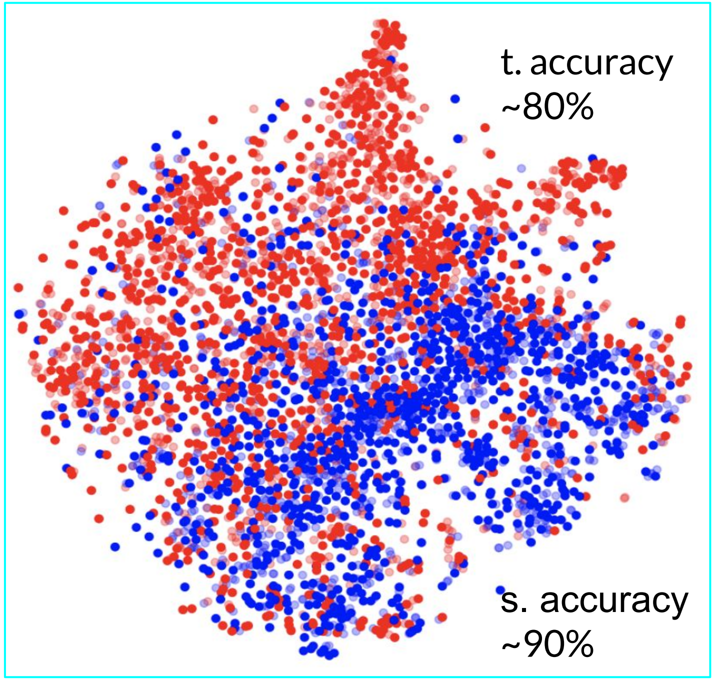
**Domain Adaptation**

# Combining Datasets




Ćiprijanović et al. 2020.
Ćiprijanović et al. 2021.

Up to 30% increase!






Fermilab

# Talk Outline

Domain Shift Problems

Domain Adaptation

**Model Robustness**

Universal Domain Adaptation

# Model Robustness

Scientific data pipelines will introduce **inadvertent data perturbations:**

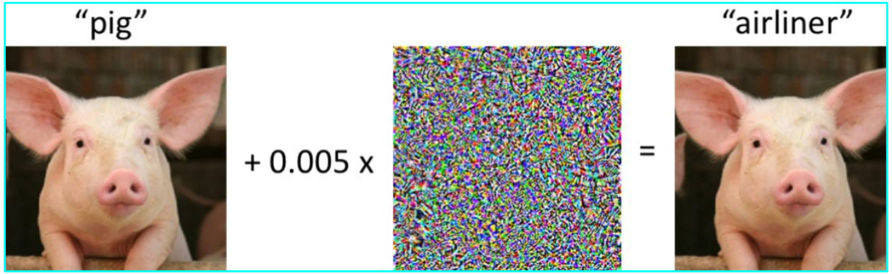
- image compression or blurring
- noise
- data pre-processing
- detector errors
- transient phenomena …

# Model Robustness

Scientific data pipelines will introduce **inadvertent data perturbations:**

- image compression or blurring
- noise
- data pre-processing
- detector errors
- transient phenomena …

**Model performance drops (sometimes catastrophically)**




Targeted attack!

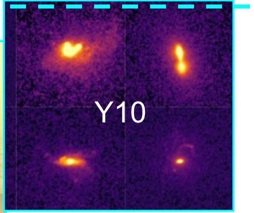
# Model Robustness

Scientific data pipelines will introduce **inadvertent data perturbations:**

- image compression or blurring
- noise
- data pre-processing
- detector errors
- transient phenomena …

**Model performance drops (sometimes catastrophically)**

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.


If we perturb **a single pixel,** model will classify the object **incorrectly**!

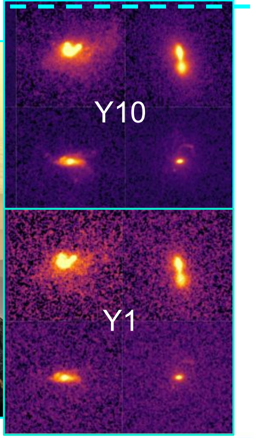
# Model Robustness

Scientific data pipelines will introduce **inadvertent data perturbations:**

- image compression or blurring
- noise
- data pre-processing
- detector errors
- transient phenomena …

**Model performance drops (sometimes catastrophically)**

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

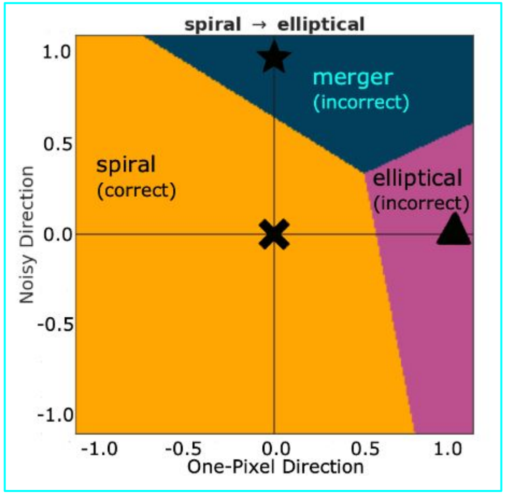

If we perturb **a single pixel**, model will classify the object **incorrectly**!

**Old data can help!**



Fermilab

# Model Robustness



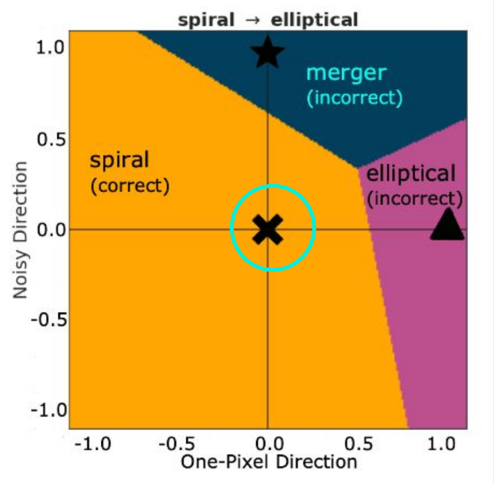

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

Regular Training on Y10 data



Fermilab

# Model Robustness



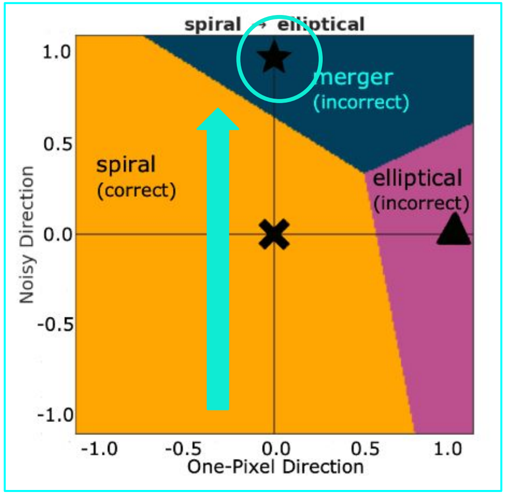

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

Regular Training on Y10 data



Fermilab

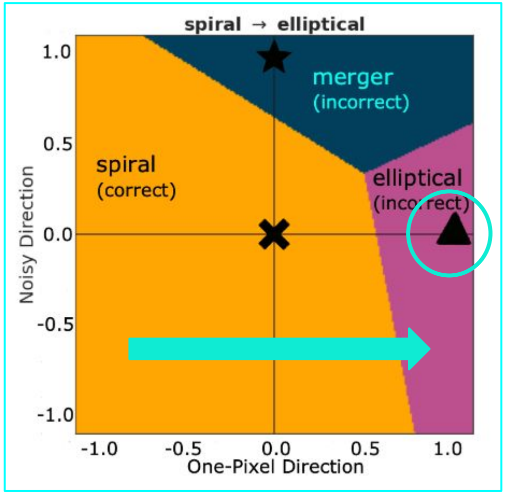
# Model Robustness




Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

Regular Training on Y10 data

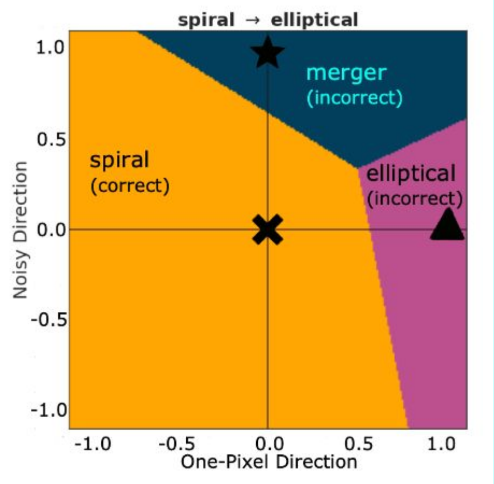
# Model Robustness

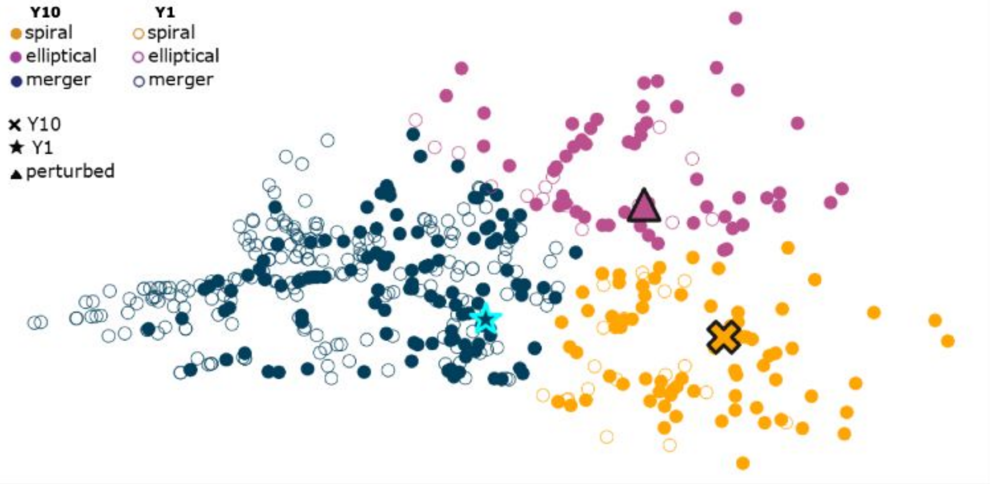



Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

Regular Training on Y10 data

# Model Robustness



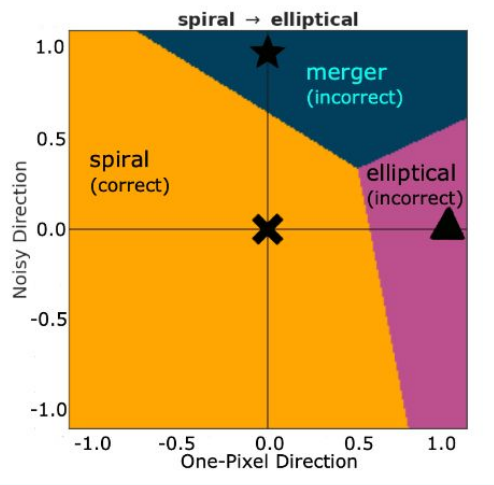

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

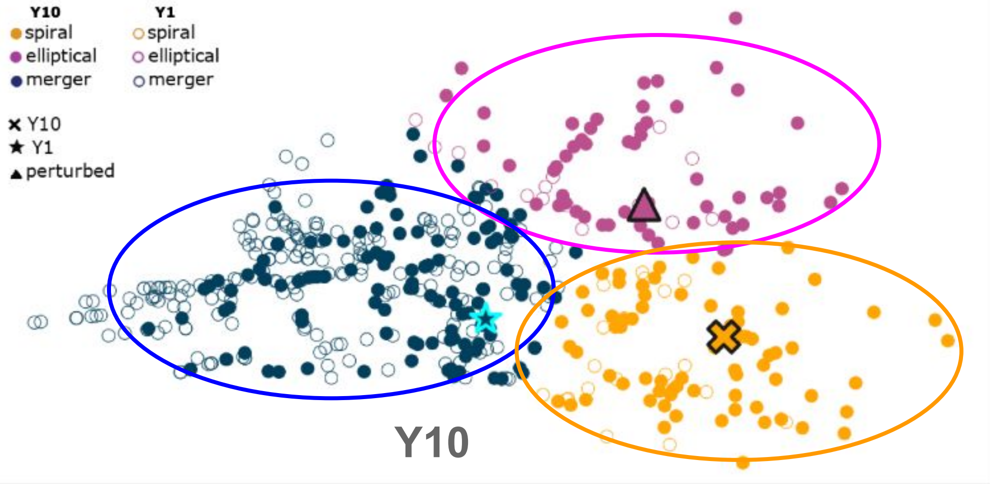
Regular Training on Y10 data



Fermilab

# Model Robustness



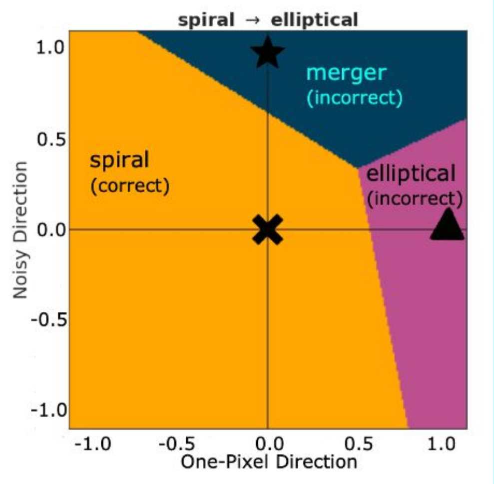

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

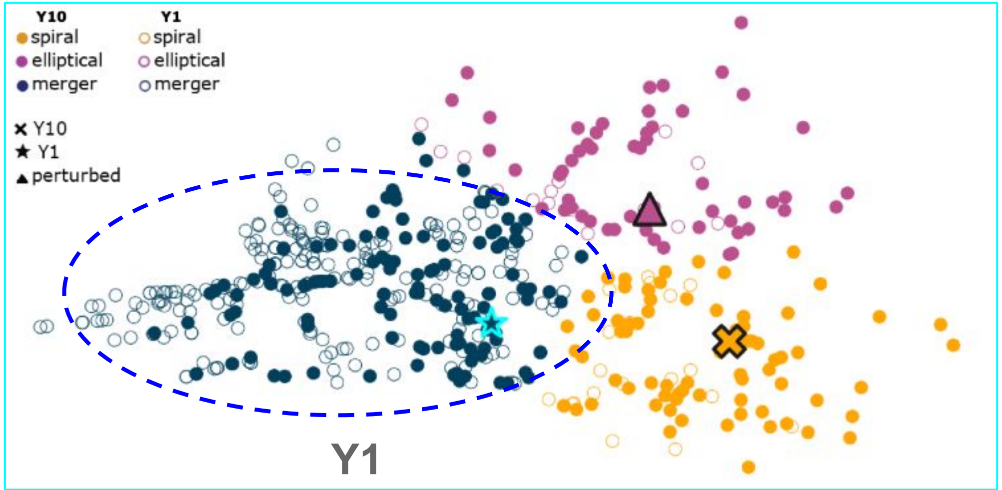
Regular Training on Y10 data



Fermilab

# Model Robustness



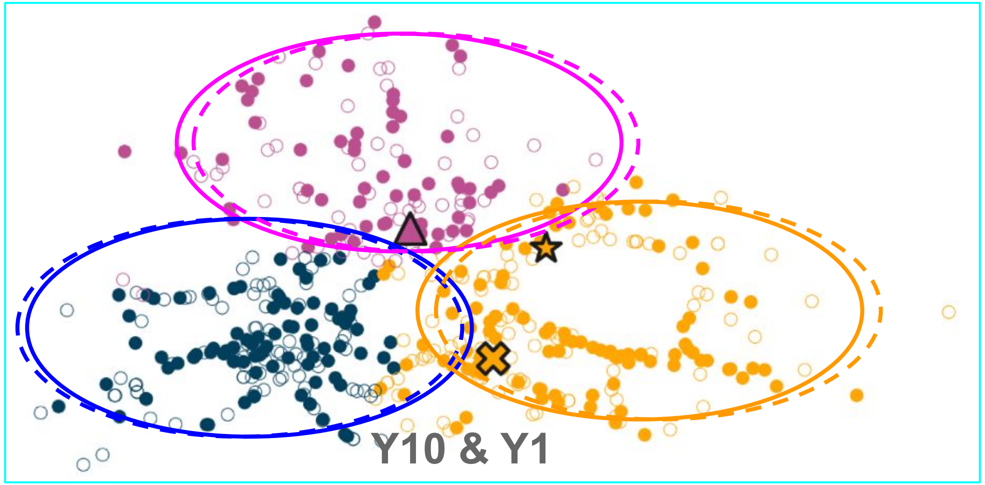

Ćiprijanović et al. 2021.
Ćiprijanović et al. 2022.

Regular Training on Y10 data

Domain Adaptation using Y1 data

# Model Robustness

- Accuracy on both datasets increases (up to 23%)!

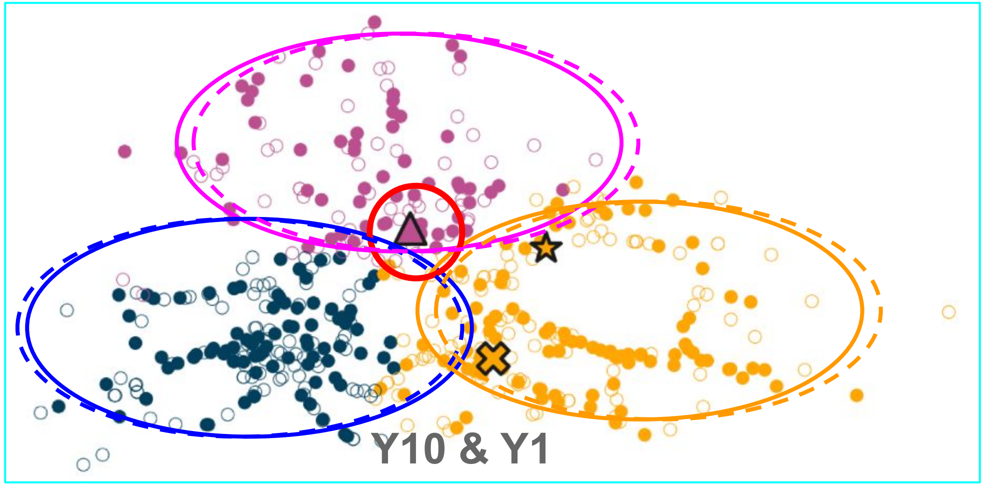



Regular Training on Y10 data → Domain Adaptation using Y1 data

Fermilab

# Model Robustness

- Accuracy on both datasets increases (up to 23%)!
- Distance to the wrong class increases ~2.3!
- **Robustness to inadvertent perturbations increases!**




Regular Training on Y10 data → Domain Adaptation using Y1 data

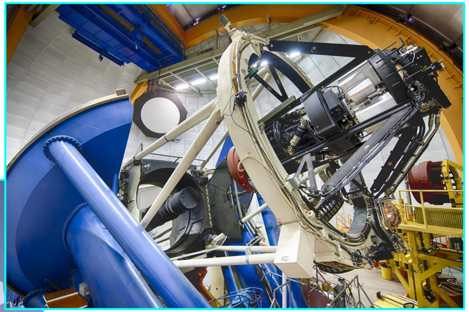
# Talk Outline

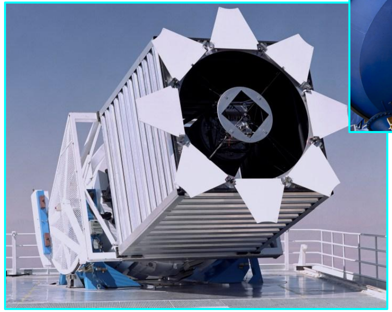
- Domain Shift Problems
- Domain Adaptation
- Model Robustness
- **Universal Domain Adaptation**

# Bridging between observations - Much Harder!

The gap between observational datasets is much larger:

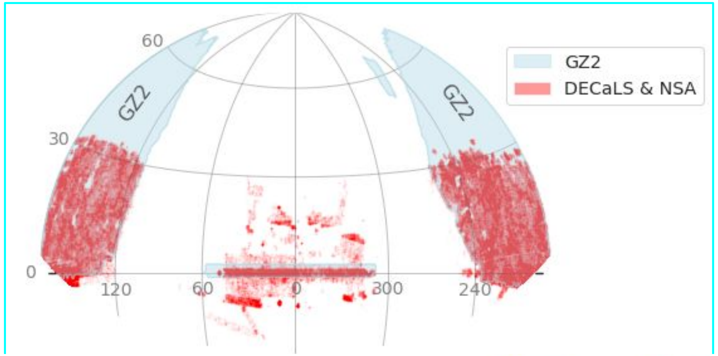
- Noise, PSF
- Pixel scale
- Depth of the survey
- Magnitude limit
- Perhaps different filters
- Different data distributions….





SDSS to DECaLS?

How do we build something flexible enough to handle any kind of data distribution shifts?

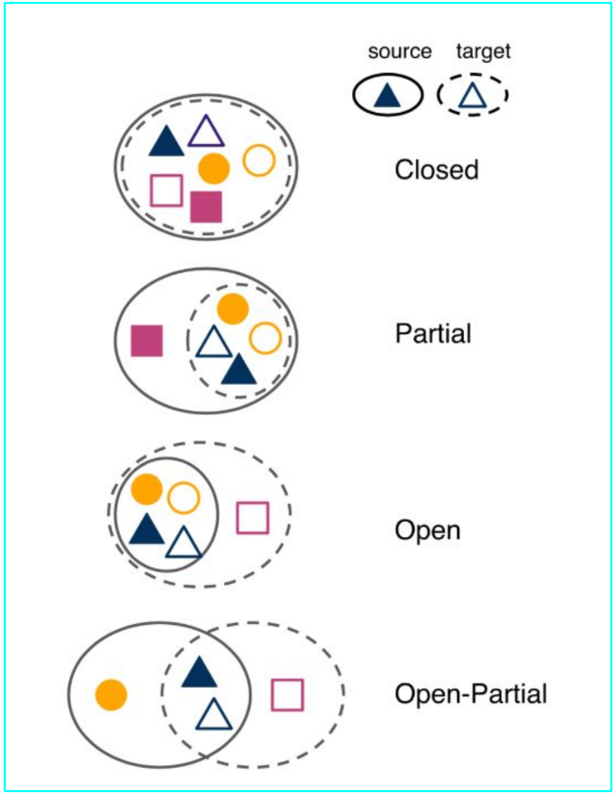





Fermilab

# Types of Dataset Shift Problems

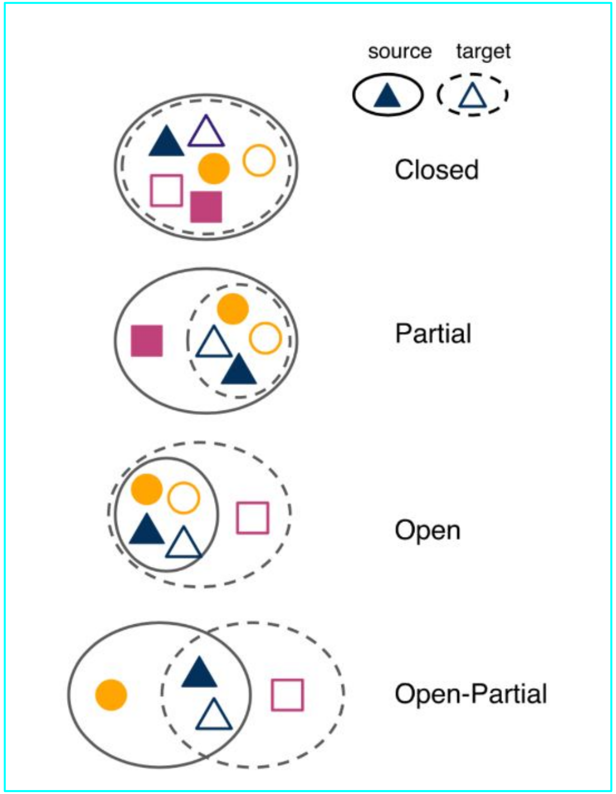
- Overall distribution per class can be different between datasets.
  - Overlapping classes should be aligned independently instead of aligning the entire data distribution.
- We can even have classes present in only one of the datasets - old labeled data or even new unlabeled data (so we won't even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



# 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.

**Use self-supervision and allow model to decide on its own!**



# 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


Fermilab

# 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

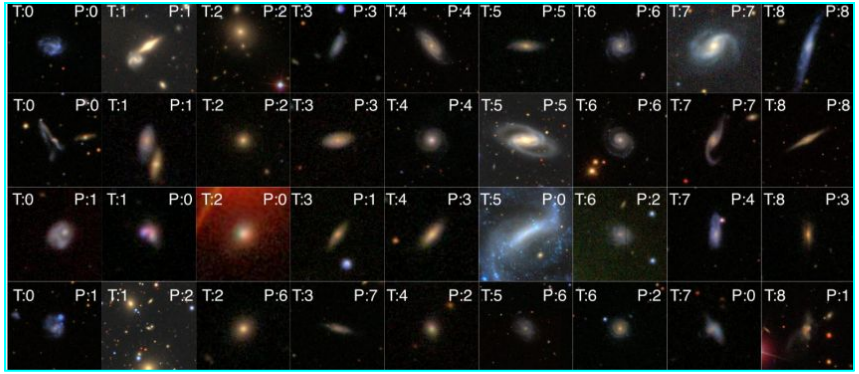
Class labels are from Galaxy Zoo 2 & 3 (crowdsourcing labels ~10^5 volunteers).

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),

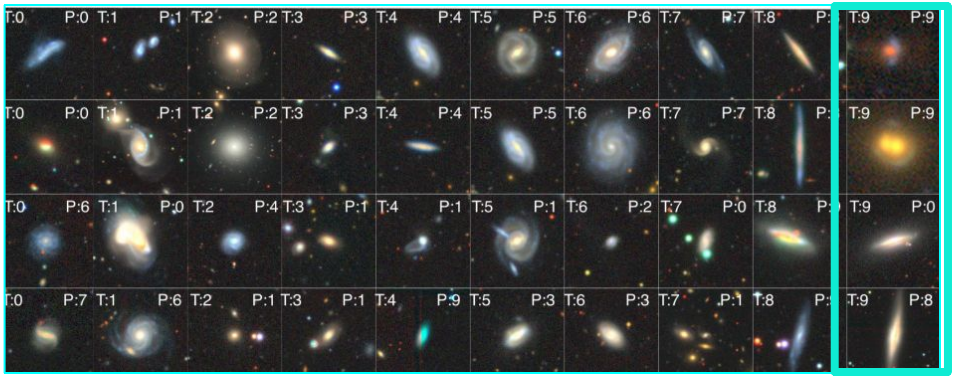
Unknown anomaly class (only in DECaLS):
Strong gravitational lens (9)

# 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)


Fermilab

# Universal Domain Adaptation (DeepAstroUDA)

Classification of known classes

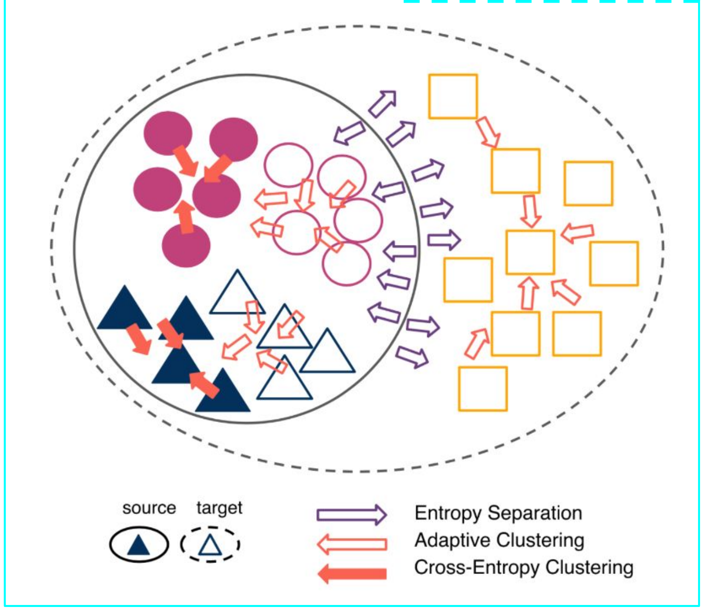
+

Clustering of similar known and unknown samples

+

Separation of different (anomalous) unknown samples

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



# Universal Domain Adaptation (DeepAstroUDA)

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

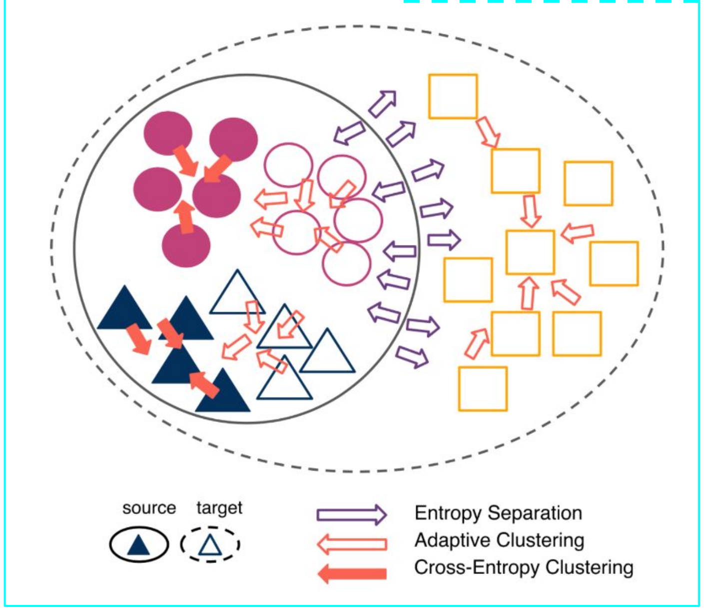
Classification of known classes

+

Clustering of similar known and unknown samples

+

Separation of different (anomalous) unknown samples




0 1 2 3 4 5 6 7 8 9

Output vector **p**

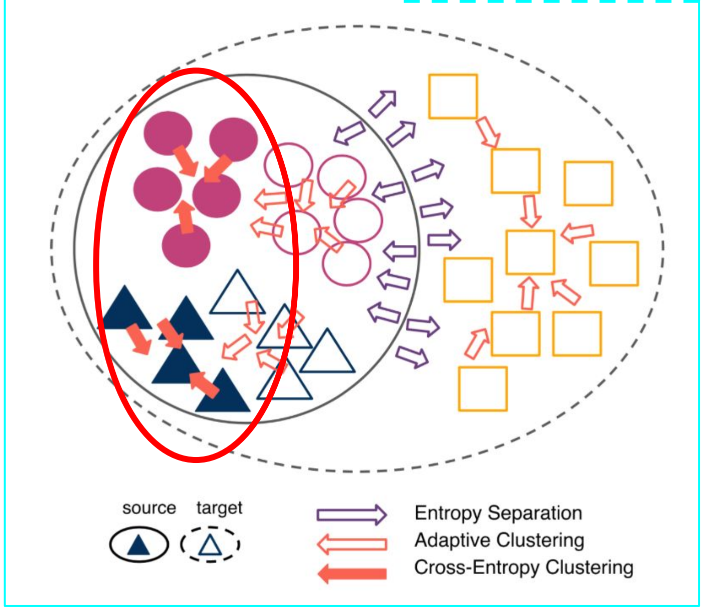
# Universal Domain Adaptation (DeepAstroUDA)

Classification of known classes

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

Using true and predicted labels

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




0 1 2 3 4 5 6 7 8 9

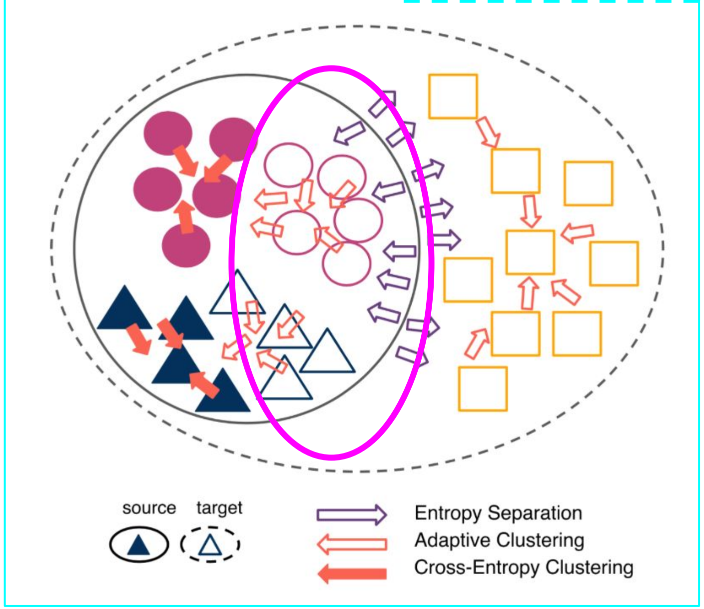
Output vector **p** → compare predicted y’ with true label y

# Universal Domain Adaptation (DeepAstroUDA)

Clustering of similar known and unknown samples

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

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





Output vector **p**

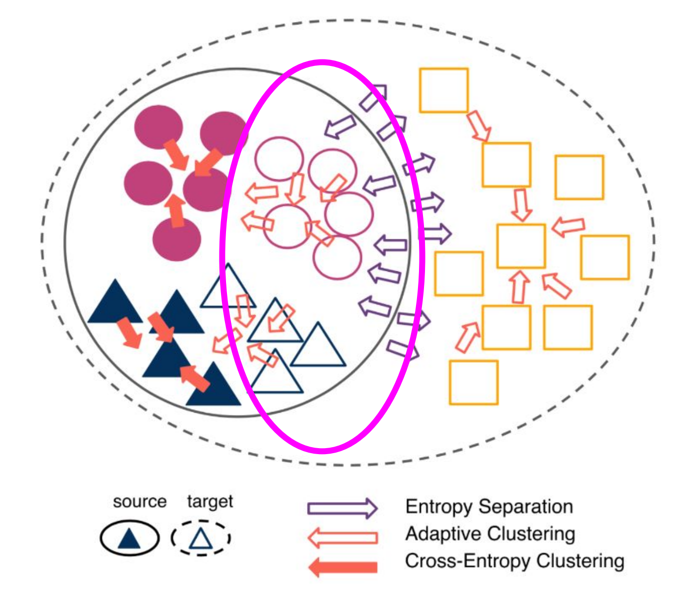
# Universal Domain Adaptation (DeepAstroUDA)

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

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





Output vector p → rank order to create similarity labels



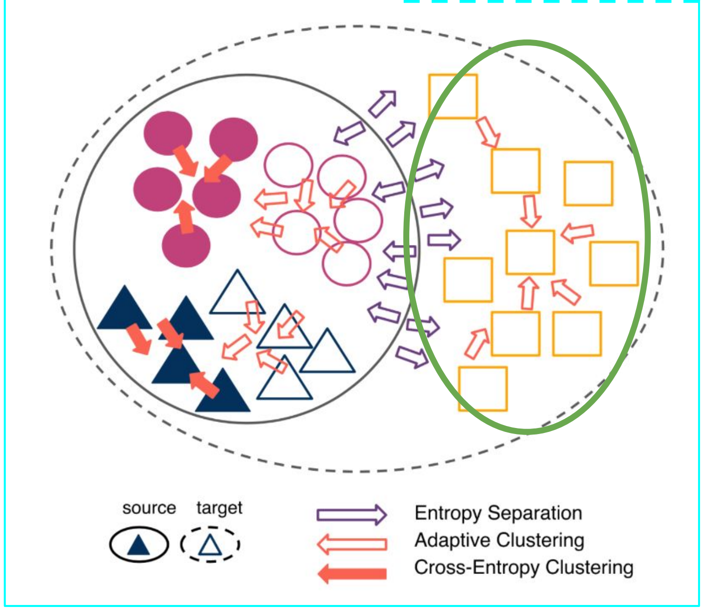
Fermilab

# Universal Domain Adaptation (DeepAstroUDA)

Separation of different (anomalous) unknown samples

Pushing away samples with high entropy of outputs features

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




| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |

Output vector p

# Universal Domain Adaptation (DeepAstroUDA)

Separation of different (anomalous) unknown samples

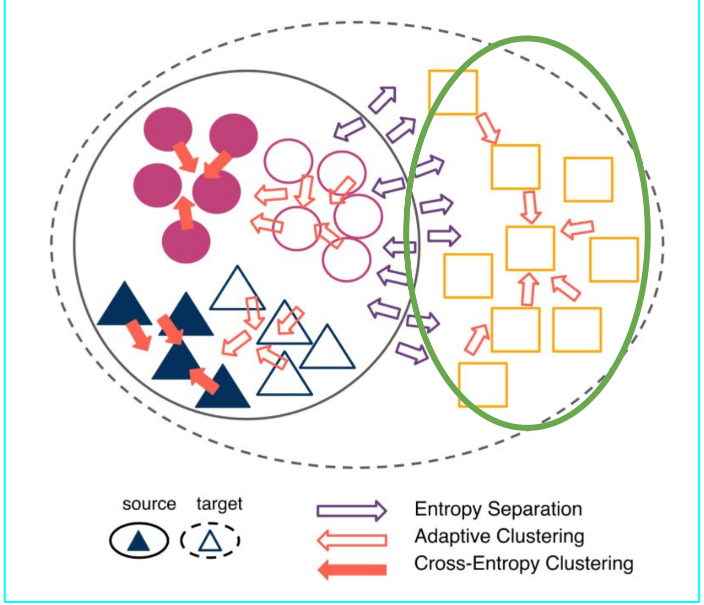
Pushing away samples with high entropy of outputs features

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

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

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

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





Output vector p → calculate entropy of each output



Fermilab

# Universal Domain Adaptation (DeepAstroUDA)

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

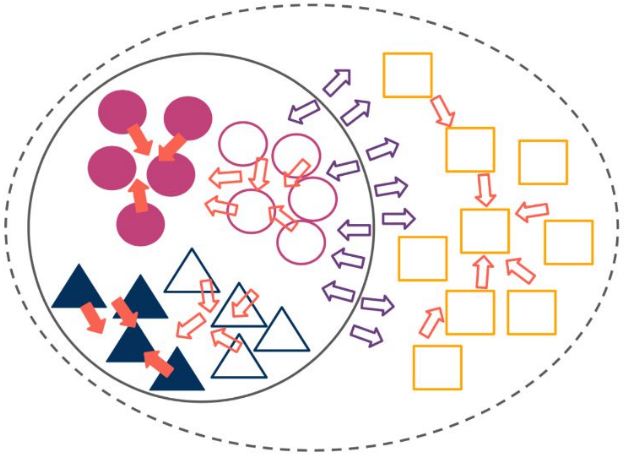
Separation of different (anomalous) unknown samples

Pushing away samples with high entropy of outputs features

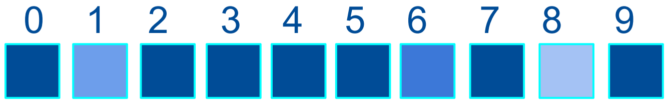
$$\mathcal{L}_{\mathrm{ES}}(\mathbf{p}_i) = \begin{cases} -|H(\mathbf{p}_i) - \rho| & |H(\mathbf{p}_i) - \rho| > m, \\ 0 & \text{otherwise.} \end{cases}$$

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

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



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

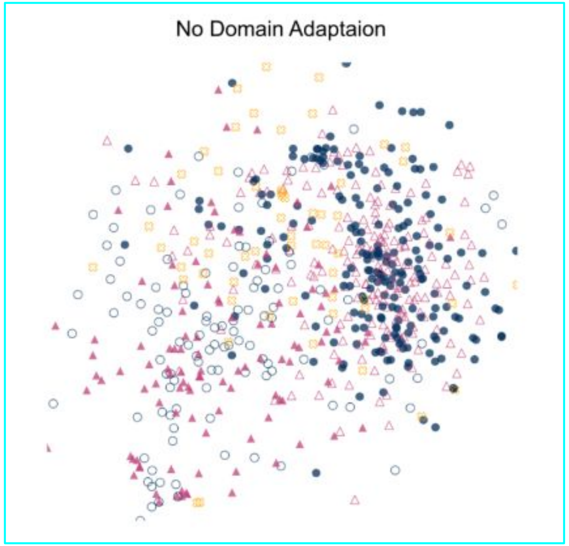



Output vector p → calculate entropy of each output

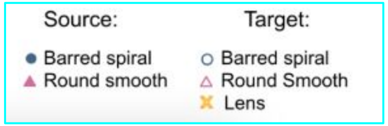


Fermilab

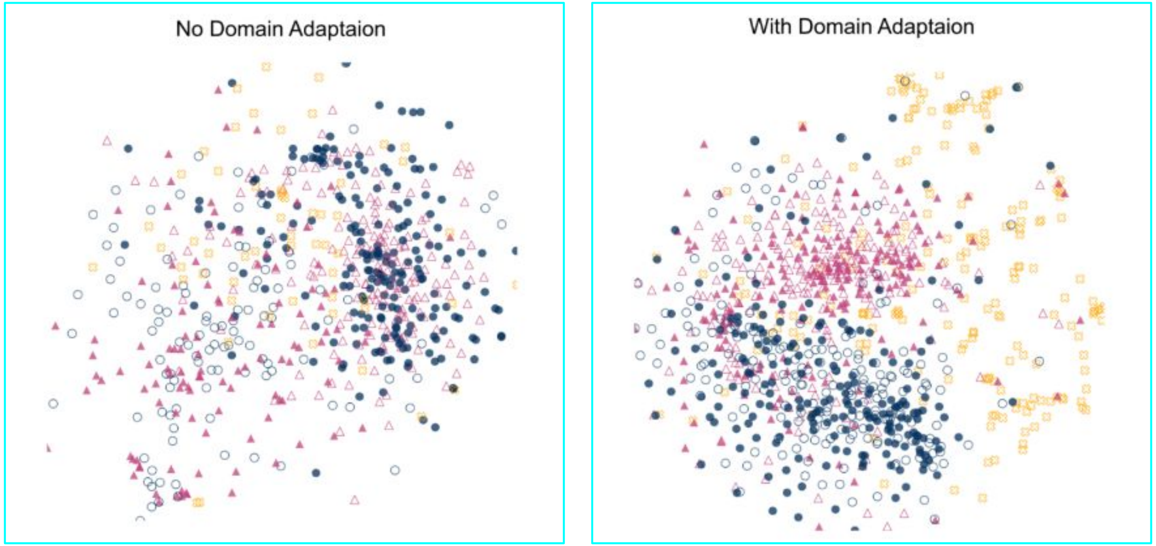
# Universal Domain Adaptation (DeepAstroUDA)



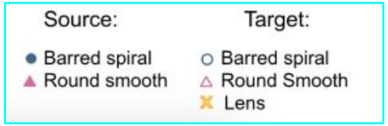

Classes are mixed!



# Universal Domain Adaptation (DeepAstroUDA)

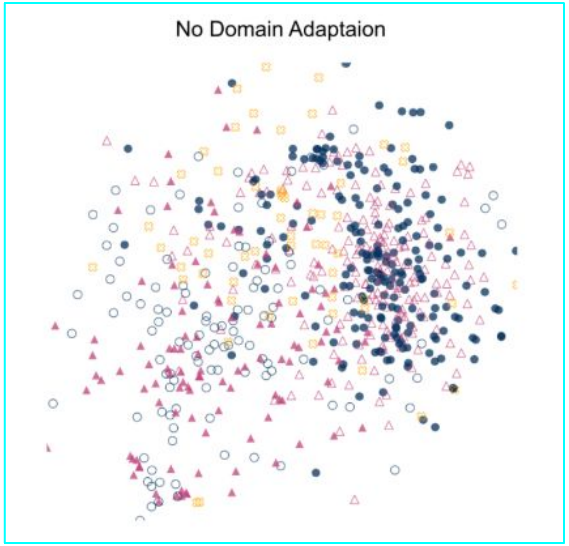


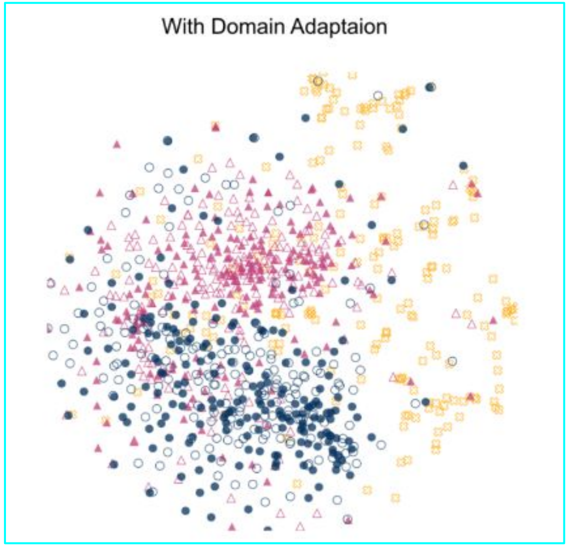

Classes are mixed!

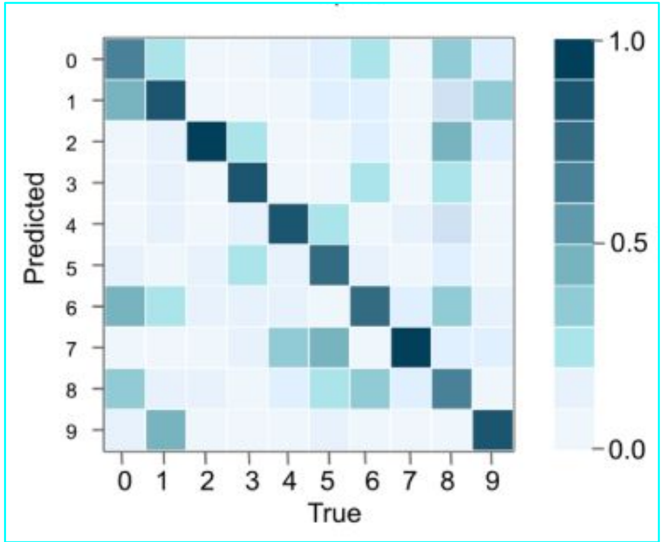



Known classes overlap,
unknown is pushed to the side.

# Universal Domain Adaptation (DeepAstroUDA)




Classes are mixed!

Source:
- Barred spiral
- Round smooth

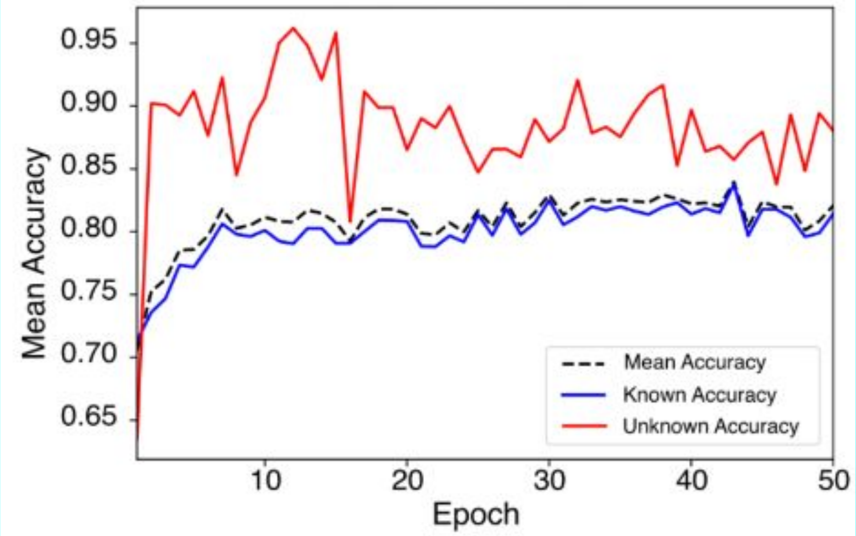
Target:
- Barred spiral
- Round Smooth
- Lens

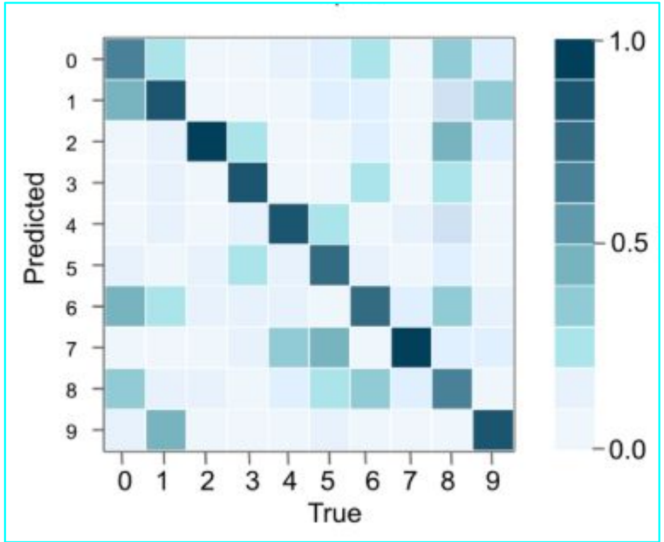
Known classes overlap,
unknown is pushed to the side.



Fermilab

# Universal Domain Adaptation (DeepAstroUDA)



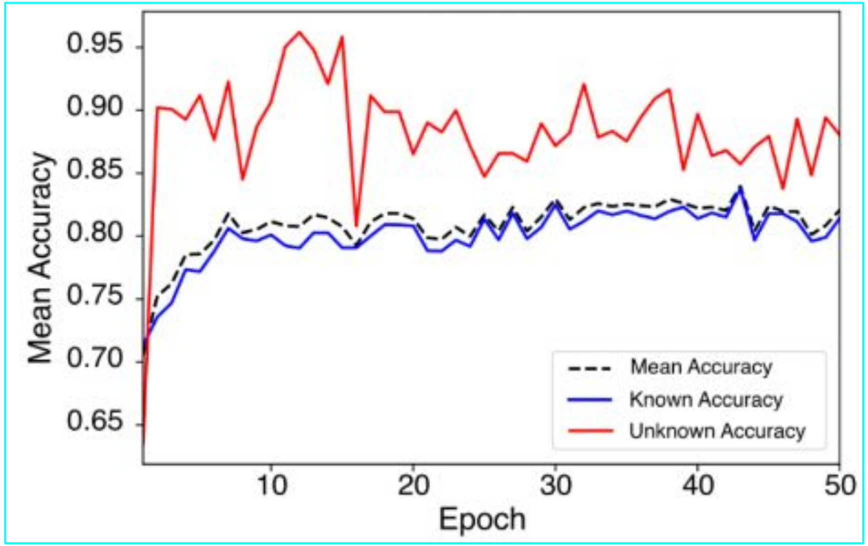




- Most confusion between classes is for truly morphologically similar classes, like disturbed and merging.
- Model is very sure about the unknown lens class - it can recognize these object look different than all other known classes.

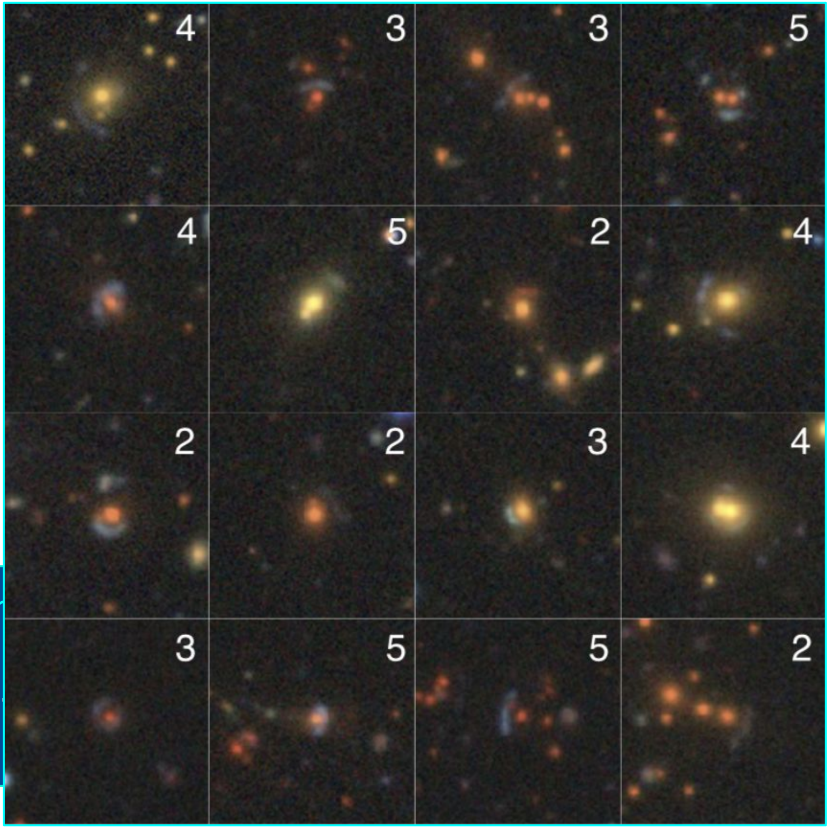


Fermilab

# 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.



# Domain Adaptation to the rescue!

Ćiprijanović et al. 2021.

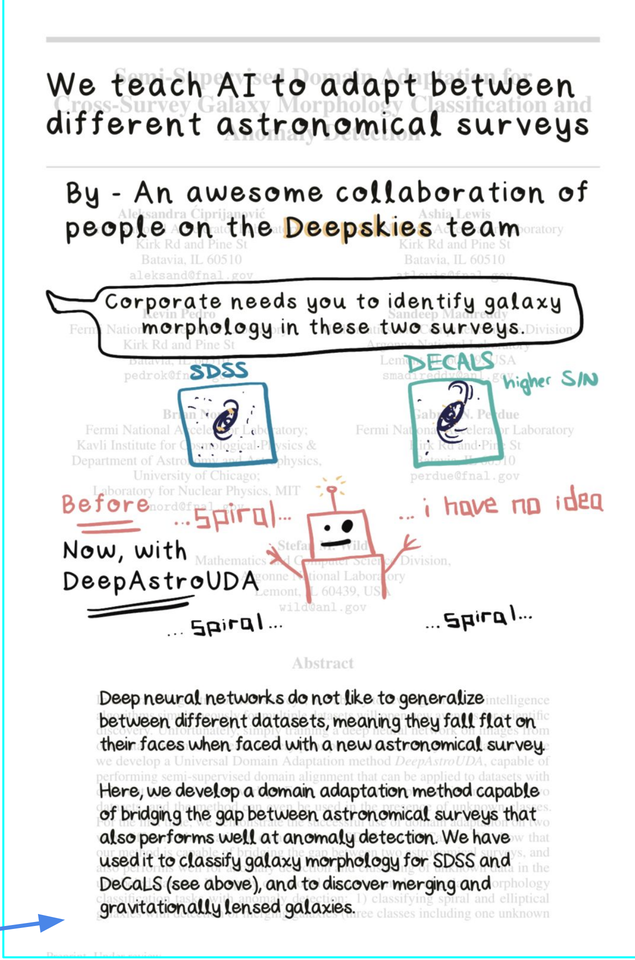
Ćiprijanović et al. 2022.

- Simulation and observations
- Increase robustness to data perturbations
- 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.

Stay tuned for applications to galaxy properties, more galaxy mergers, strong gravitational lensing and cosmology!

By Becky N.






Fermilab

# Big thanks to all my amazing collaborators

**Fermilab**


**University of Chicago**


**Argonne, Oakridge,Berkeley**


**Space Telescope Science Institute**


and many more!

Argonne NATIONAL LABORATORY

STScI

BERKELEY LAB

OAK RIDGE National Laboratory

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science


# THANK YOU!

**Aleksandra Ćiprijanović**
(she/her/hers)

**Fermilab, DSSL**
**aleksand@fnal.gov**

Interested in AI/ML for Astrophysics?
Join the **Deep Skies Lab**!
https://deepskieslab.com/


MATF Seminar
April, 2023