FERMILAB-SLIDES-21-043-SCD
This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Domain Adaptation for Cross-Domain Studies in Astronomy:

Merging Galaxies Identification

Aleksandra Ćiprijanović
Research Associate
Scientific Computing Division
aleksand@fnal.gov

Where the Earth Meets the Sky
27-28 May 2021
Cosmic Dawn Center at DTU

Talk outline

1. Astro example and what I work on
2. What is domain discrepancy?
3. Domain adaptation - two methods
4. How does domain adaptation help?

Merging galaxies

WHY

To understand the evolution of our Universe (galaxy mergers lead to hierarchical formation of structures).

HOW

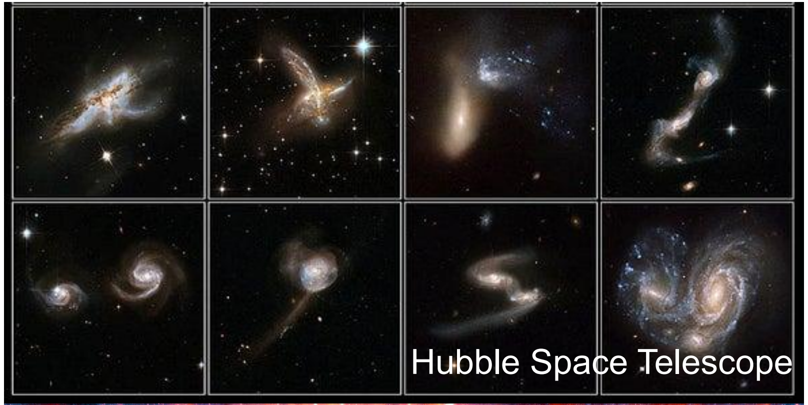
Leverage a large sample of merging galaxies to study.

PROBLEMS

Standard methods require knowledge about the morphology (we need for precise observations). Visual classification is very time consuming and prone to errors.

SOLUTION

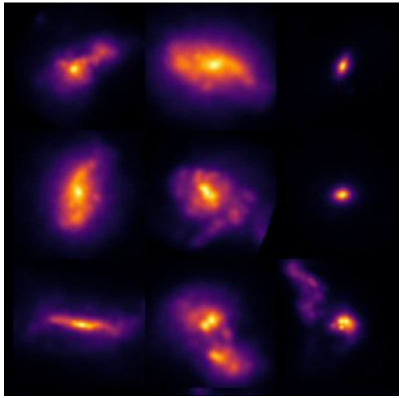
Large simulations (we know the ground truth) + machine learning



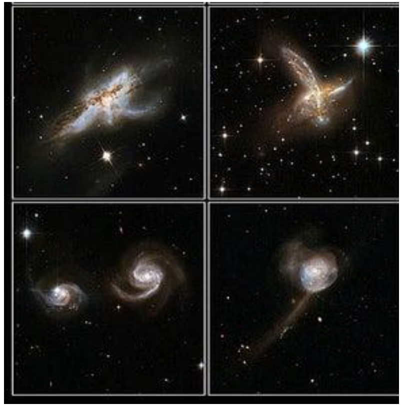
Fermilab

Where are differences coming from?

Simulation (source) LABELED!



Real (target) UNLABELED!



Simulations are not perfect - physics missing, computational resources

Use data from multiple telescopes with different specs

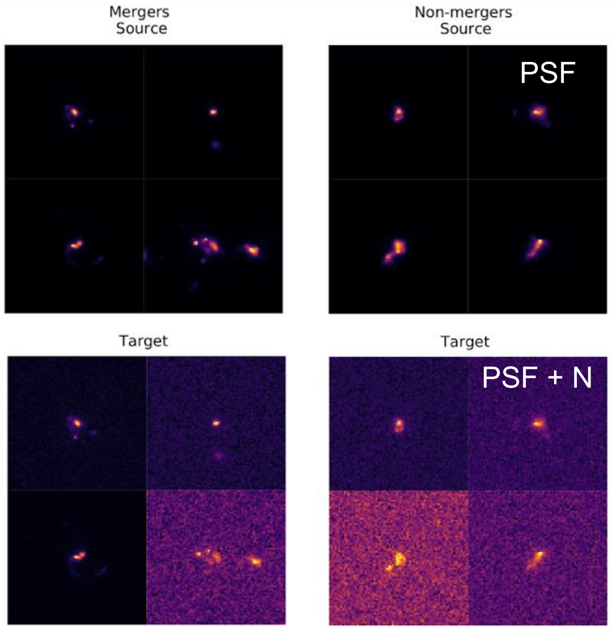
Dataset shift in astronomy

Making mock images is hard - adding noise, PSF, telescope imperfections

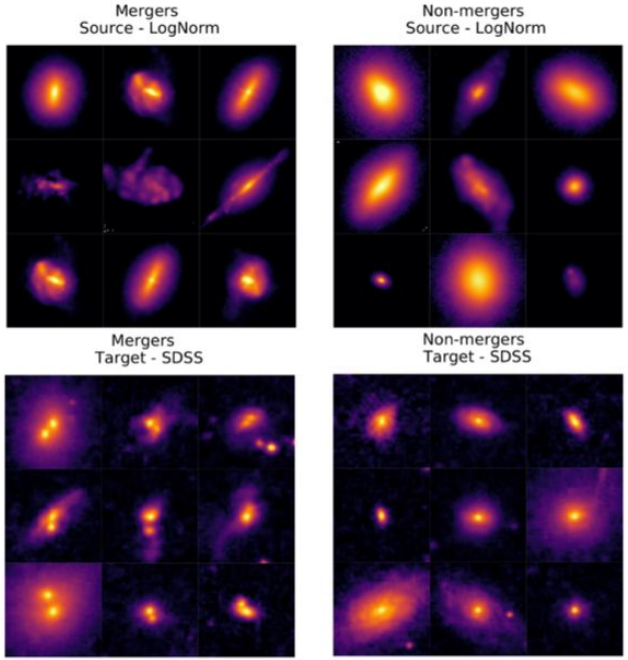
Fermilab

Two experiments

Simulation → Simulation + Noise

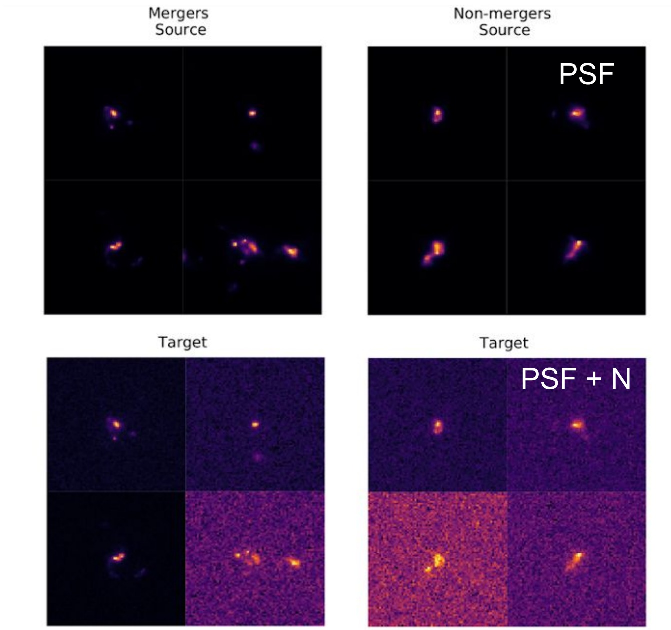


Simulation → Observations



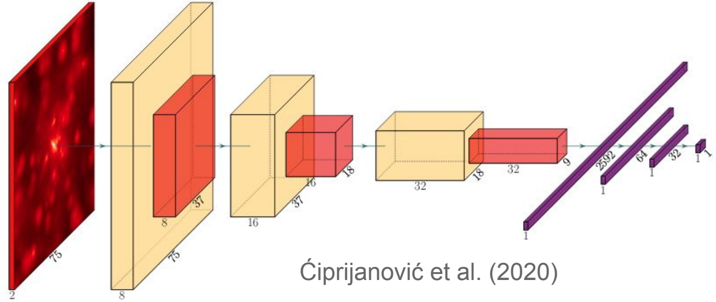
Two experiments

Simulation → Simulation + Noise



Vogelsberger et al. (2014)

- **Illustris simulation**
  - source (sim. + Hubble PSF)
  - target (sim. + PSF + random sky shot noise)
- Distant mergers at z=2
- 2233 individual galaxies
- **~15 000 images**



Ćiprijanović et al. (2020)

Total Loss = Task Loss + Transfer Loss

Task loss - very often categorical cross-entropy loss

$$L_{\text{cross-entropy}}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_i y_i \log(\hat{y}_i)$$

Transfer loss - domain alignment

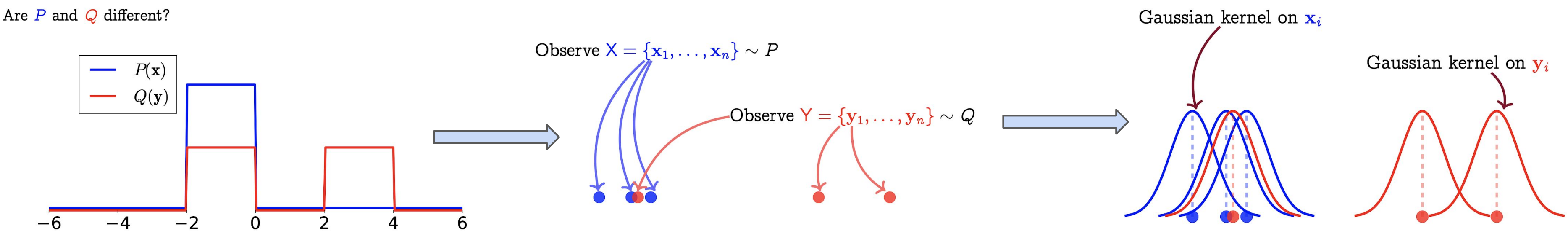
Maximum Mean Discrepancy

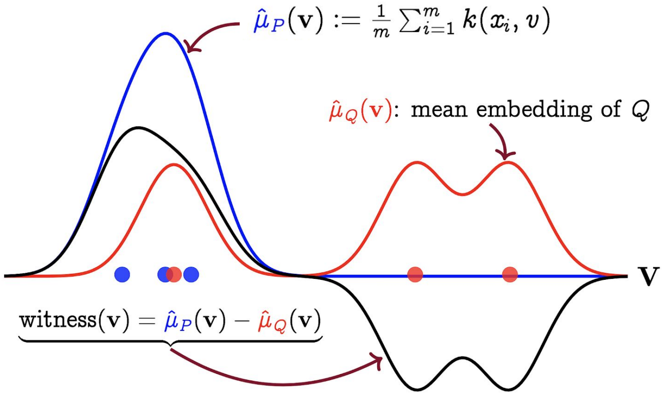
Non-parametric distance between two probability distributions (distance of the mean embeddings of the samples in the kernel space).

Adversarial training on domain labels

Using Domain Adversarial Neural Network (DANN) to force domain-invariant feature extraction.

Maximum Mean Discrepancy - MMD





$$\widehat{MMD}^2 = \|\mathrm{witness}(\mathbf{v})\|_{\mathcal{F}}^2$$
$$= \frac{1}{n(n-1)}\sum_{i\neq j} k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i\neq j} k(\mathbf{y}_i, \mathbf{y}_j)$$
$$\boxed{- \frac{2}{n^2}\sum_{i,j} k(x_i, \mathbf{y}_j)}$$

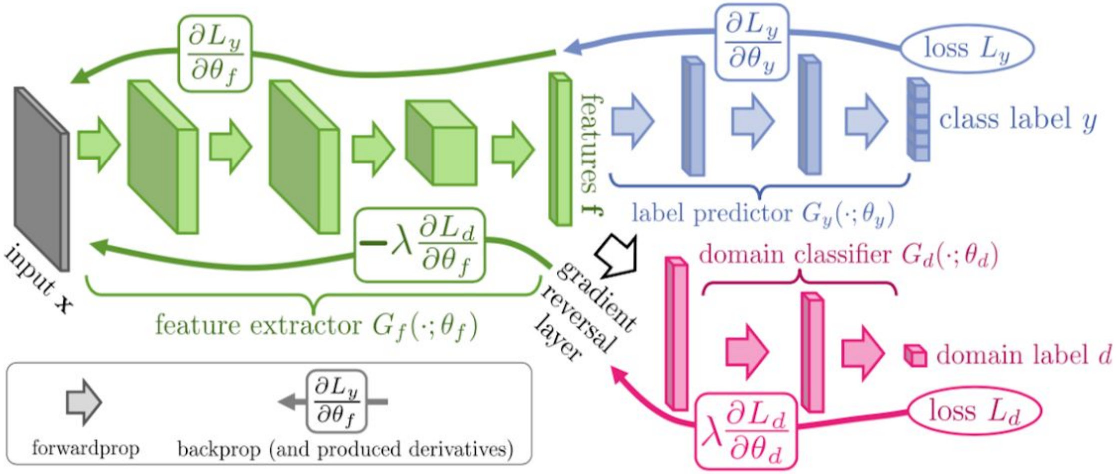
Minimize MMD loss by maximizing cross-similarities ⇒ we find **domain invariant features**!

From Arthur Gretton (NIPS 2016 Workshop on Adversarial Learning, Barcelona Spain)

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.

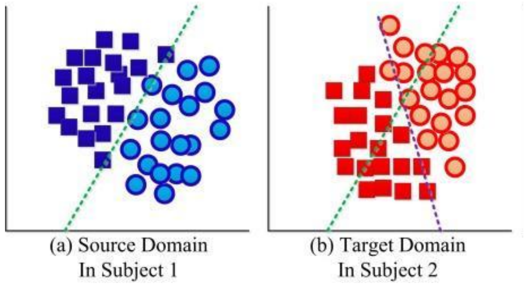


Ganin et al. (2016)

Results

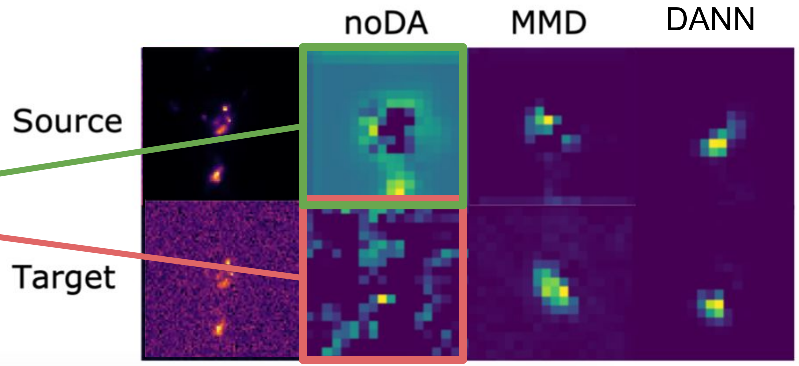
DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains

arxiv:2103.01373



	Source Domain	Target Domain
noDA	85%	58%
MMD	87%	77%
DANN	87%	**79%**

Grad-CAM (merger class)

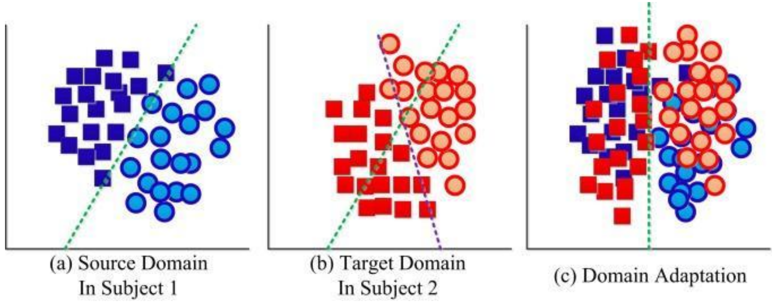


What is the network focusing on?

Results

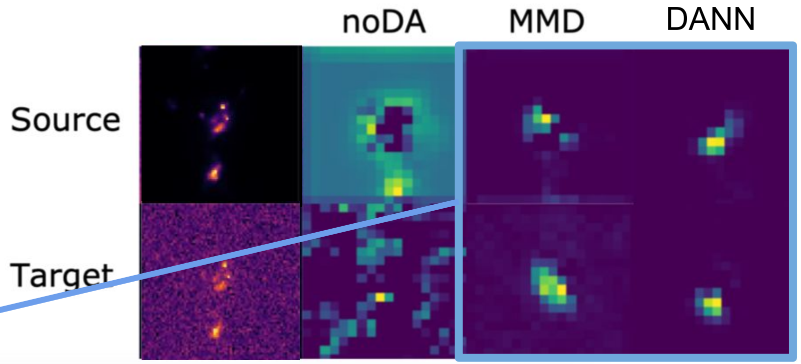
DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains

arxiv:2103.01373



	Source Domain	Target Domain
noDA	85%	58%
MMD	87%	**77%**
DANN	87%	**79%**

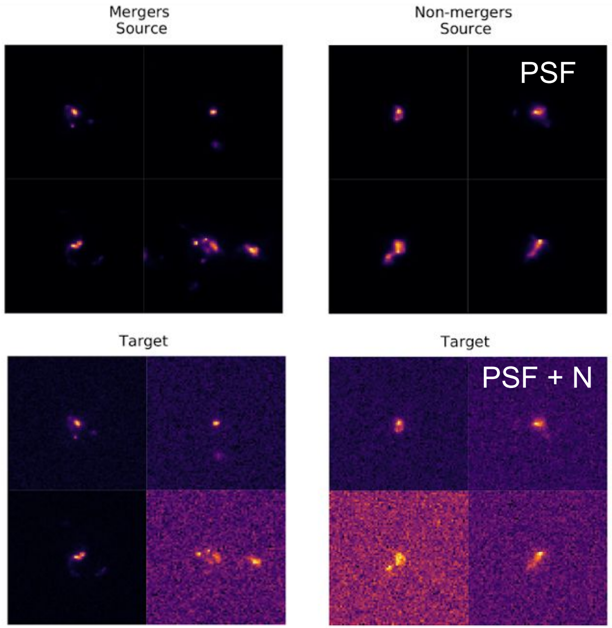
Grad-CAM (merger class)



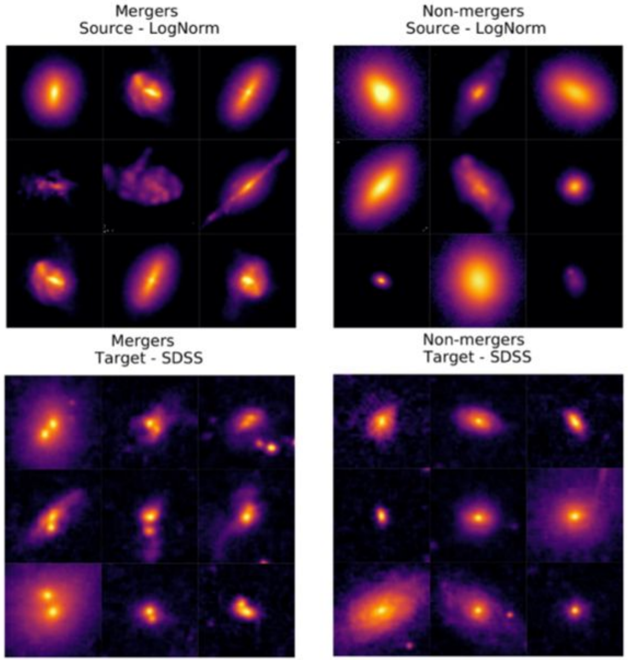
What is the network focusing on?

Two experiments

Simulation → Simulation + Noise



Simulation → Observations



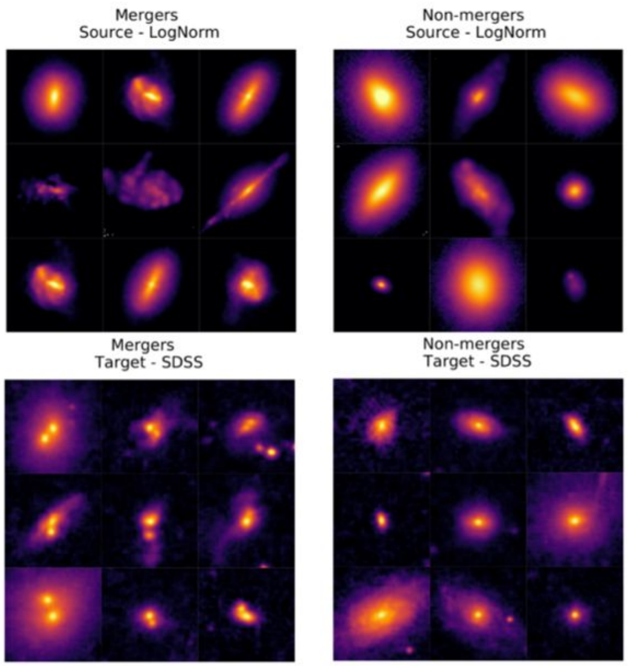
Two experiments

- **Source:** Illustris nearby galaxies
  - z=0
  - small dataset (44 mergers) !

- **Target:** Real galaxies - SDSS:
  - small dataset (310 mergers)!
  - z<0.1
  - very different, only simple examples!
  - labeled by humans !
- ~6000 images

Vogelsberger et al. (2014)
Darg et al. (2010)
Lintott et al. (2008)

Simulation → Observations

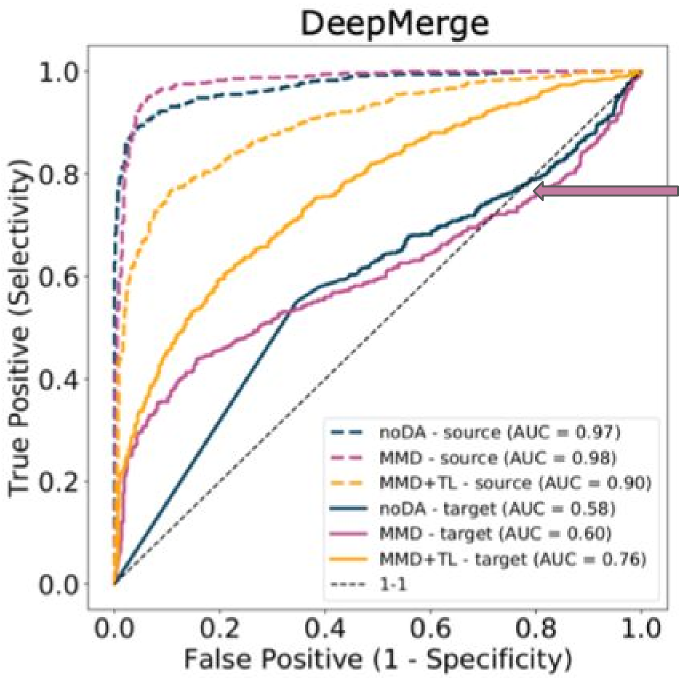


Fermilab

Results

DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains

arxiv:2103.01373



noDA and MMD

Accuracy ~ 50%

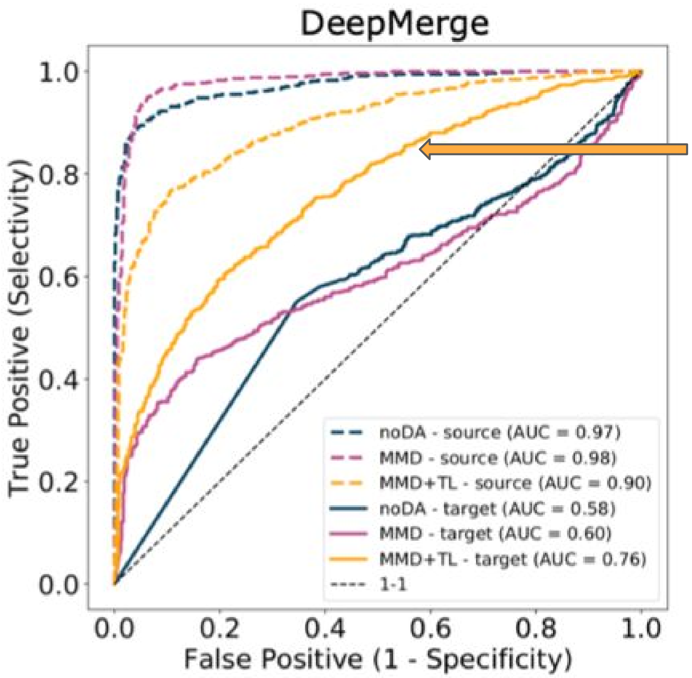
	Source Domain	Target Domain
noDA	91%	50%
MMD	94%	53%
MMD+T	83%	**69%**

Fermilab

Results

DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains

arxiv:2103.01373



MMD with TL form previous model - works!

Accuracy ~ 70%

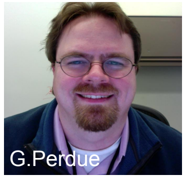
	Source Domain	Target Domain
noDA	91%	50%
MMD	94%	53%
MMD+T	83%	**69%**

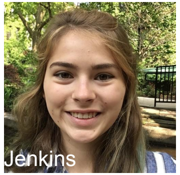
Fermilab

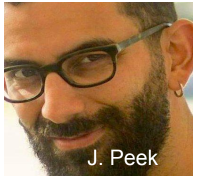
Summary & What's next?

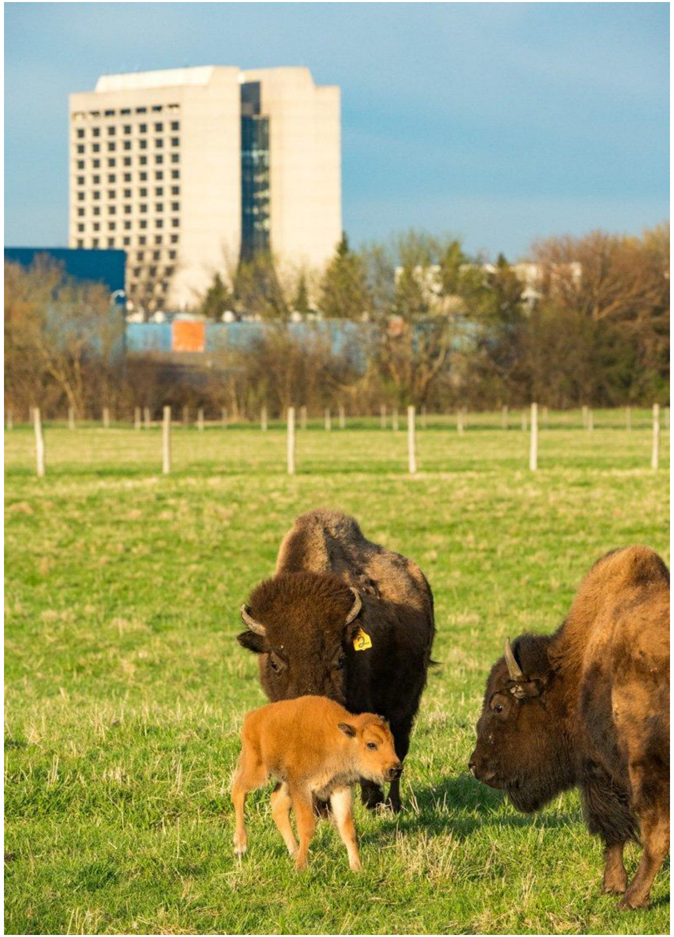
- **Merging galaxies** are important for the study of galaxy morphology, but also evolution of structure in the Universe.
- **Domain adaptation (DA) is crucial** for successful bridging between different data sets and full utilisation of ML in science.

- Harder problems will need more sophisticated **methods that try to align classes** (MMD aligns the entire distribution).
- Discrepant domains can lead to **negative transfer** and impact the performance.
- Can DA help us **make more robust algorithms**, understand decision boundaries and uncertainties of our ML algorithms?









Thank you!

aleksand@fnal.gov