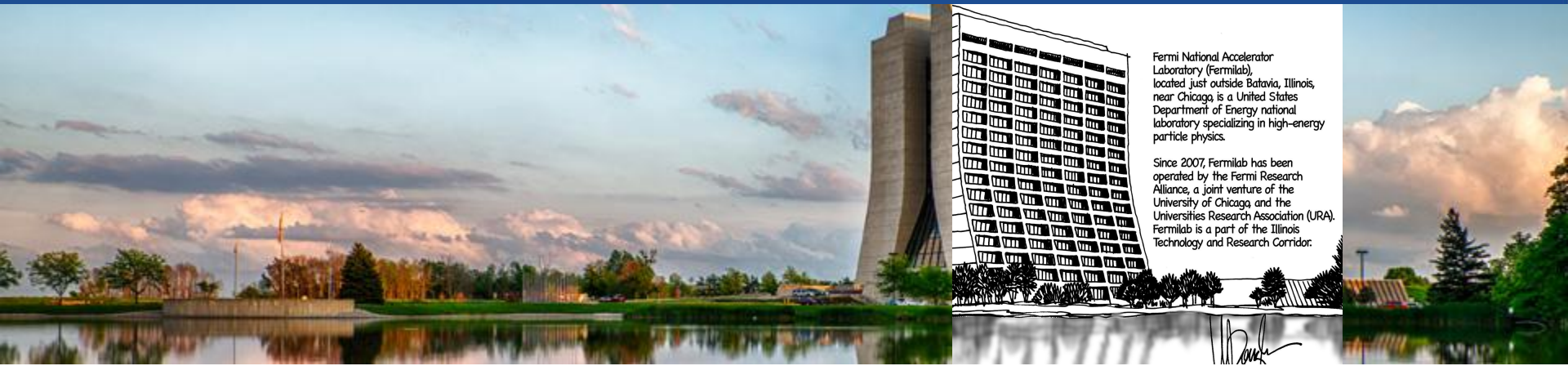


Domain adaptation for cross-domain studies in astronomy

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

HEP - ML Seminar
Imperial College London
January 2021

Which one is real?

Which one is real?

Fermilab

Talk outline

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

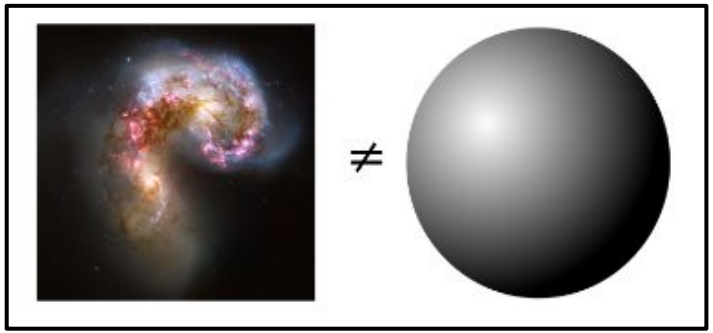
Why should we care?

Accelerates research

Models based on the data

Unexpected discoveries

Parameter space reduction

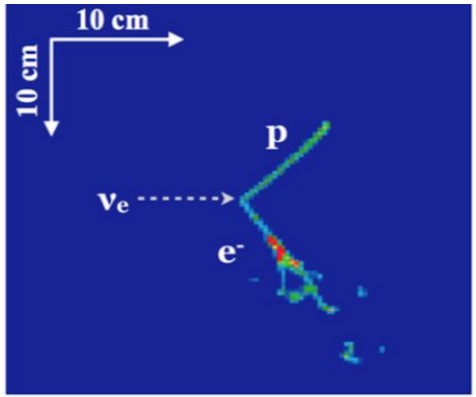


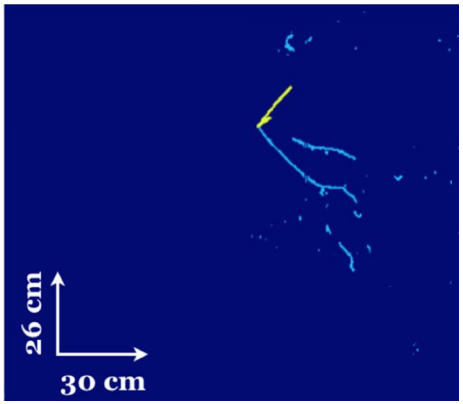
Fermilab

What about physics?

MicroBooNE
Neutrinos

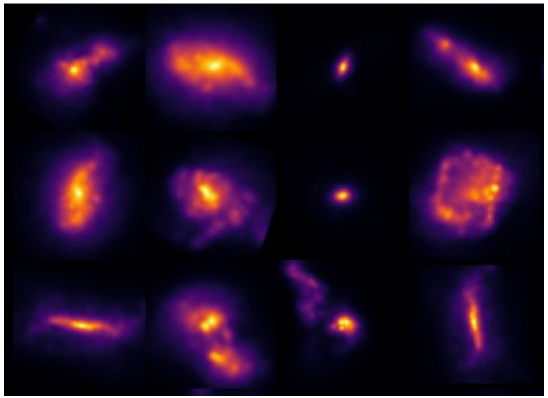
Adams et al. (2019)

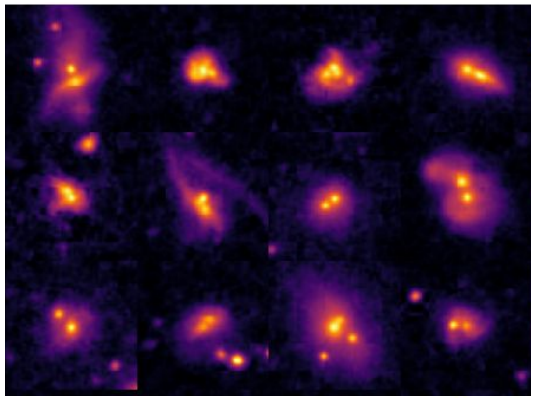




Illustris and SDSS
Merging galaxies

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





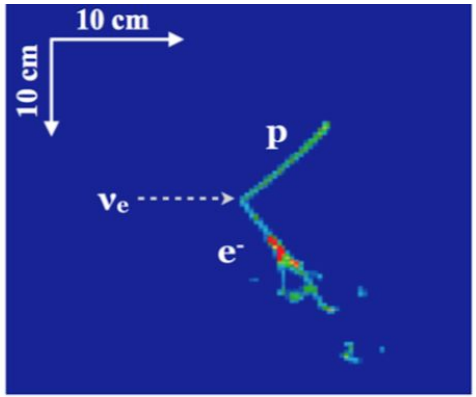
Fermilab

What about physics?

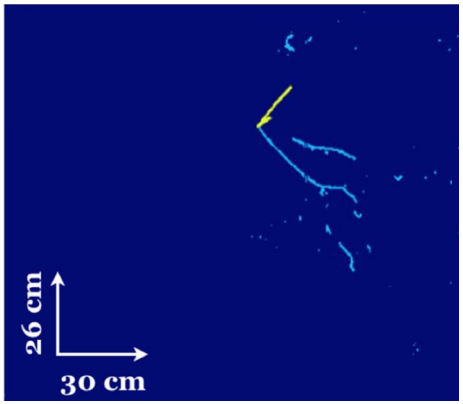
MicroBooNE
Neutrinos

Adams et al. (2019)

Simulation

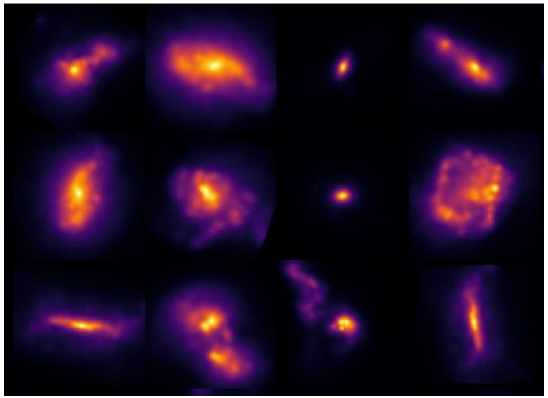


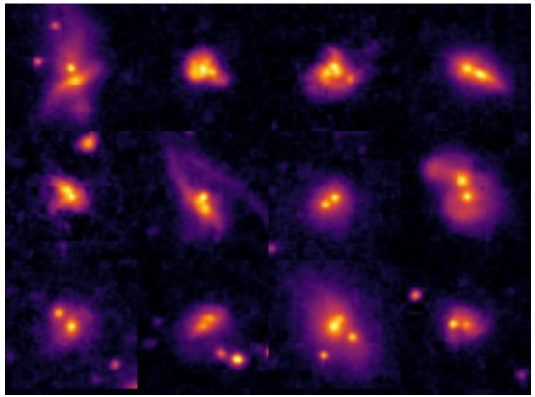
Real



Illustris and SDSS
Merging galaxies

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

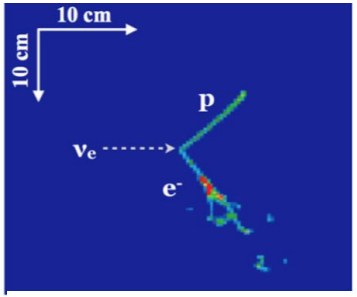




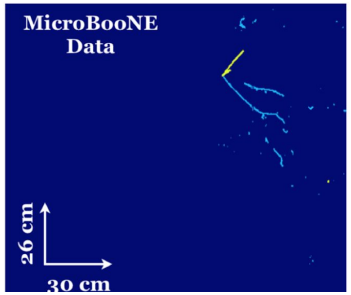
Fermilab

Where are differences coming from?

Simulation (source)



Real (target)



Wrong estimates of the total rate of the background or data

Dataset shift in physics

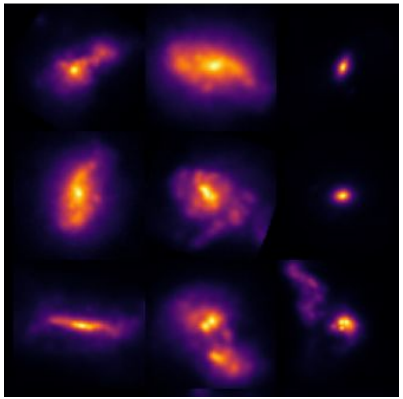
Approximate process geometry, imperfect detector

Not considering all the categories of physics processes

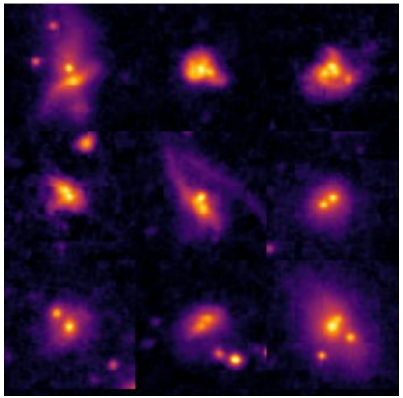
Fermilab

Where are differences coming from?

Simulation (source)



Real (target)



Simulations are not perfect - physics missing, computational resources

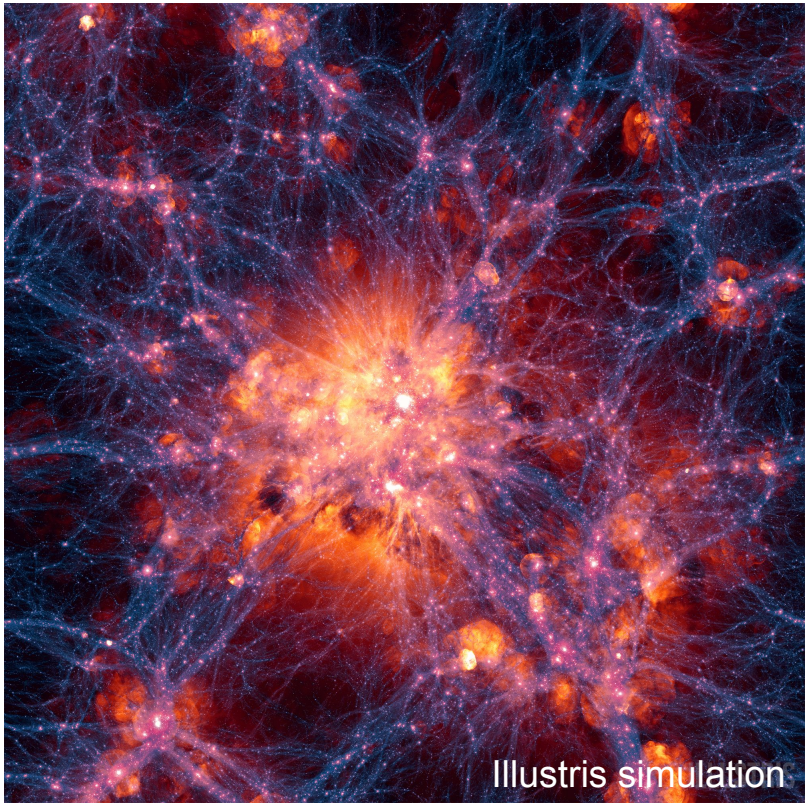
Dataset shift in astronomy

Use data from different telescopes

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

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



Merging galaxies

Standard Classification Methods

visual, morphology parameters
(concentration, asymmetry, clumpiness)

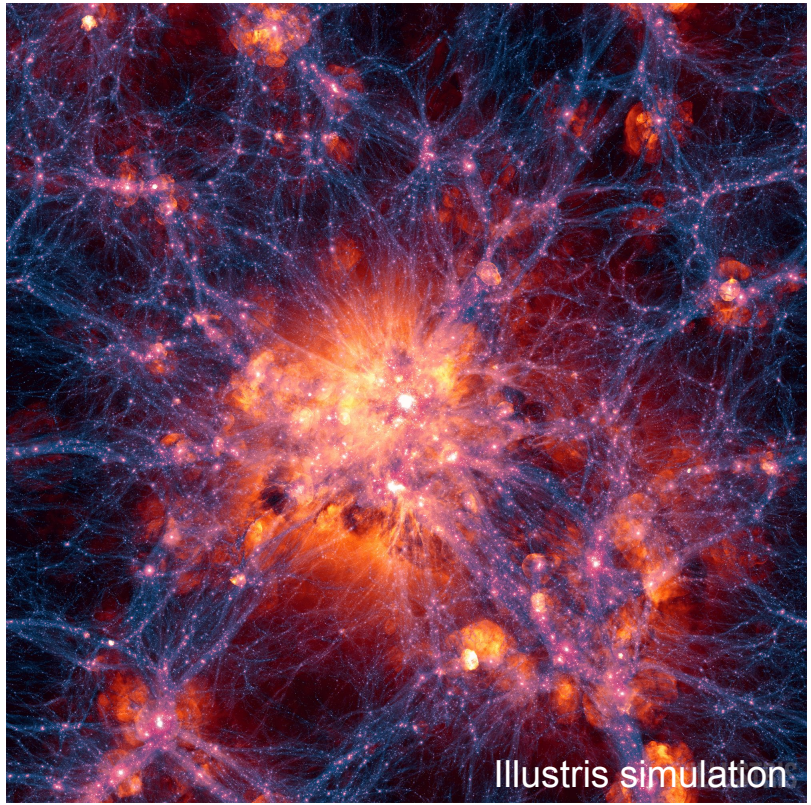
Machine Learning

Random Forests

Snyder at al. (2019)

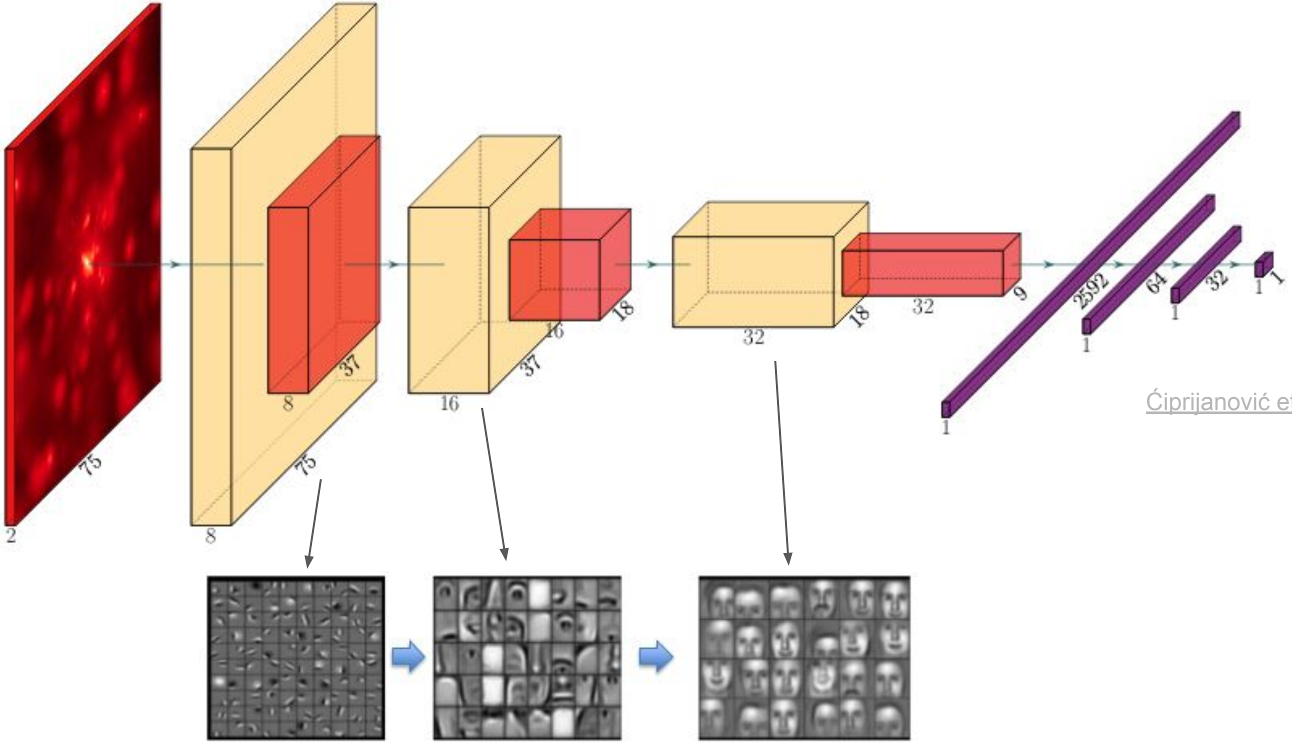
Deep Learning

Ćiprijanović et al. (2020)



Fermilab

Convolutional Neural Networks



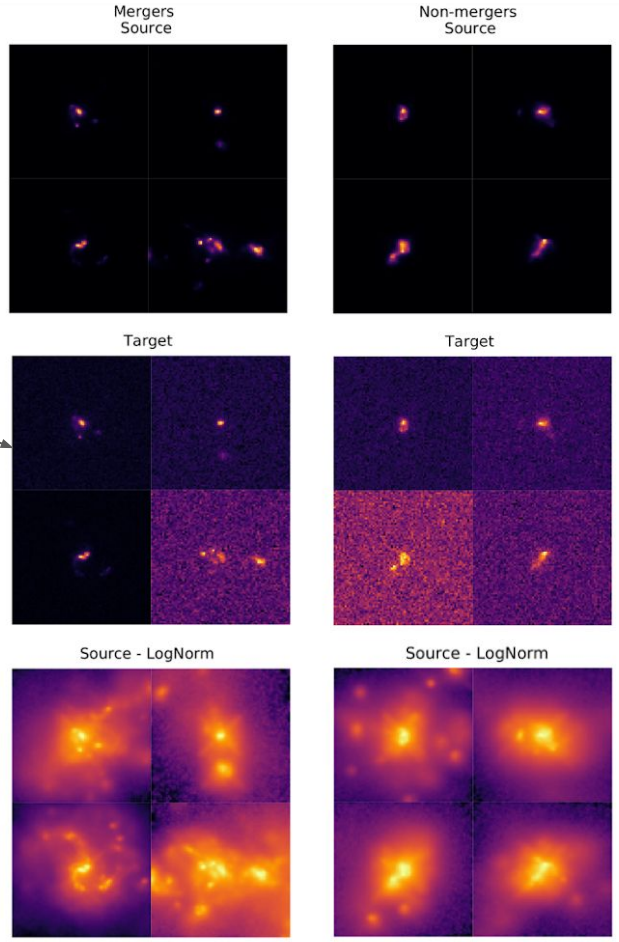
Ćiprijanović et al. (2020)

Fermilab

Distant Merging Galaxies

Metric \ Train / Test	Pristine / Pristine	Noisy / Noisy
AUC	0.86 ± 0.01	0.82 ± 0.01
Accuracy	0.79 ± 0.01	0.76 ± 0.01
Precision	0.81 ± 0.02	0.77 ± 0.02
Recall	0.80 ± 0.02	0.78 ± 0.02
F1 score	0.81 ± 0.02	0.77 ± 0.02
Brier score	0.15 ± 0.007	0.17 ± 0.007

Ćiprijanović et al. (2020)

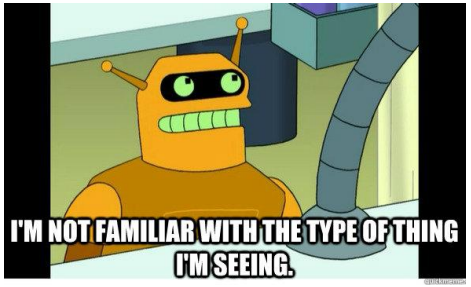


- Illustris simulation - source (simulation + Hubble PSF); target (simulation + PSF + random sky shot noise)
- 2 filter images - one for bluer features in galaxies (star formation, clumps, and asymmetries), the other redder features (reveal stellar mass and mergers)
- z=2, 2233 galaxies, around 15 000 images

New Discoveries will leverage all available data

Train on simulated data

Use the model on observed data



	Trained on Pristine, Tested on Noisy	Trained on Noisy, Tested on Pristine
AUC	0.53 ± 0.02	0.79 ± 0.02
Accuracy	0.47 ± 0.02	0.56 ± 0.02
Precision	0.74 ± 0.01	0.55 ± 0.02
Recall	0.03 ± 0.009	0.98 ± 0.007
F1 score	0.06 ± 0.02	0.71 ± 0.01
Brier score	0.42 ± 0.01	0.30 ± 0.01

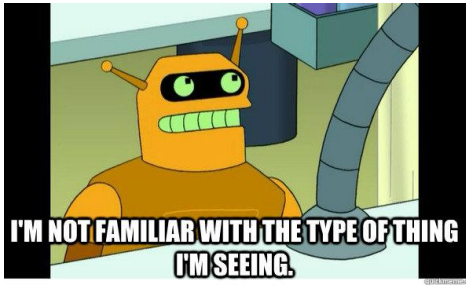
Ćiprijanović et al. (2020)

Fermilab

New Discoveries leverage all available data

Train on simulated data

Use the model on observed data

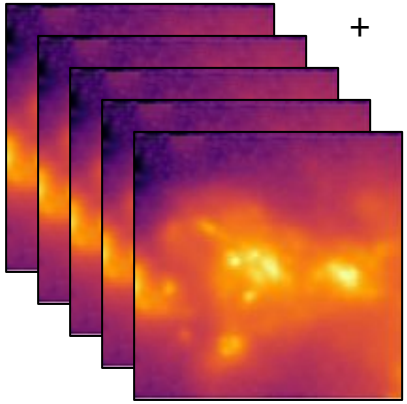


	Trained on Pristine, Tested on Noisy	Trained on Noisy, Tested on Pristine
AUC	0.53 ± 0.02	0.79 ± 0.02
Accuracy	0.47 ± 0.02	0.56 ± 0.02
Precision	0.74 ± 0.01	0.55 ± 0.02
Recall	0.03 ± 0.009	0.98 ± 0.007
F1 score	0.06 ± 0.02	0.71 ± 0.01
Brier score	0.42 ± 0.01	0.30 ± 0.01

Ćiprijanović et al. (2020)

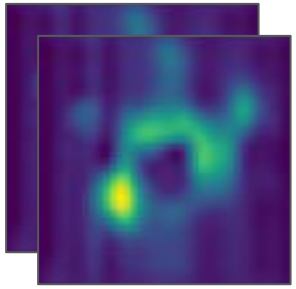
Fermilab

TRAINING



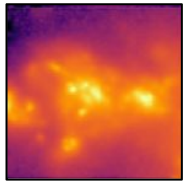
\+ Labels

FIND AND REFINE FEATURES



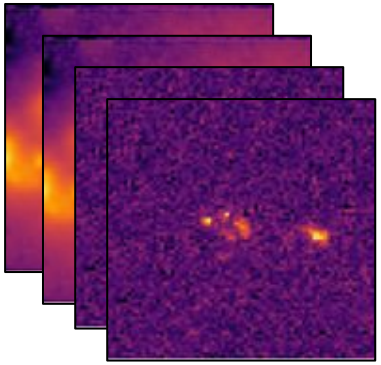
grad-CAM

CLASSIFY

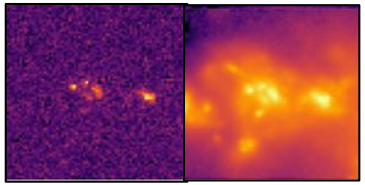


(NON)MERGER

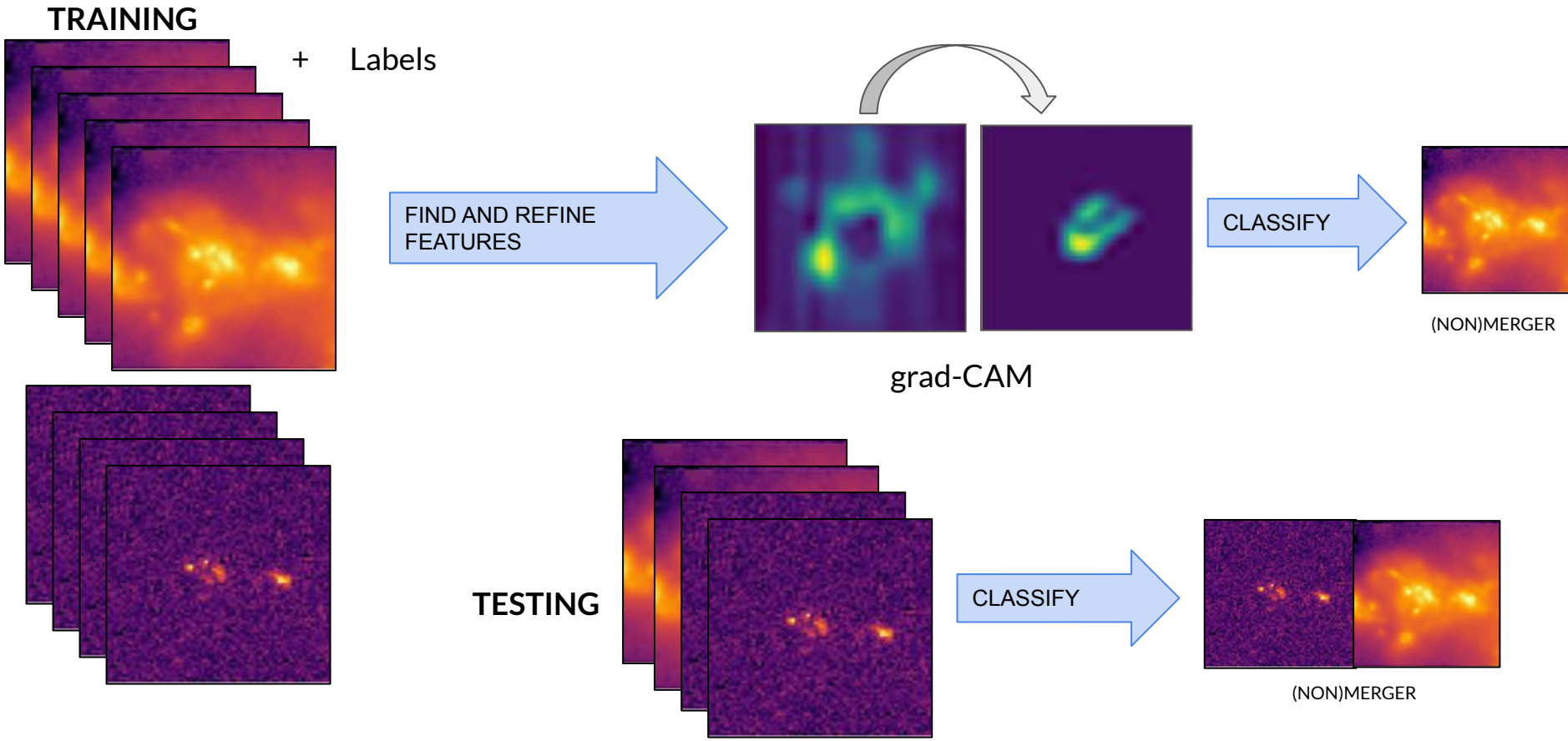
TESTING



CLASSIFY



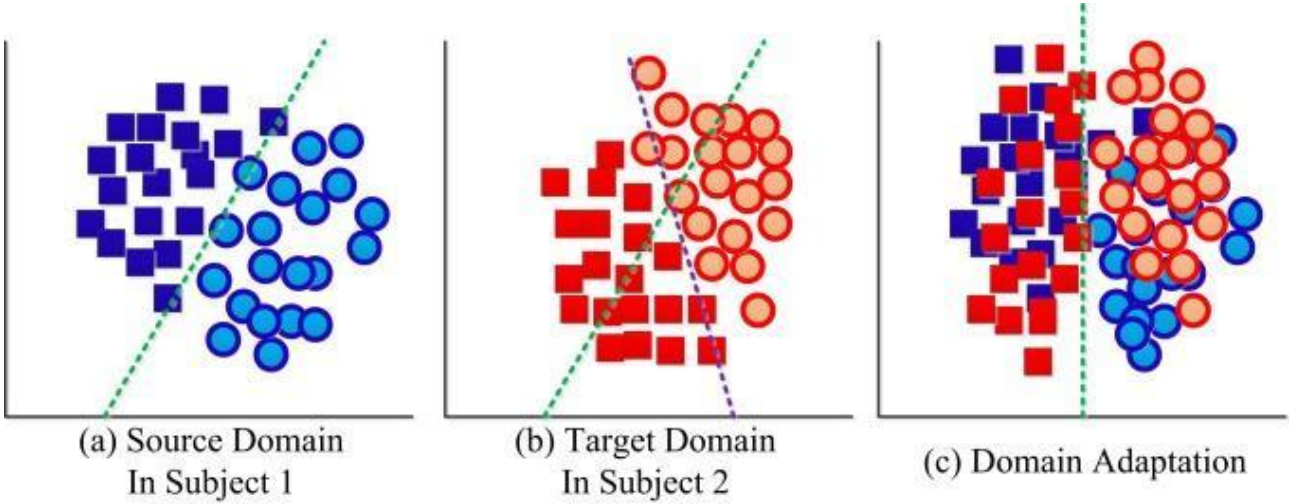
(NON)MERGER


TRAINING
+ Labels
FIND AND REFINE FEATURES
CLASSIFY
(NON)MERGER
grad-CAM
TESTING
CLASSIFY
(NON)MERGER

Fermilab

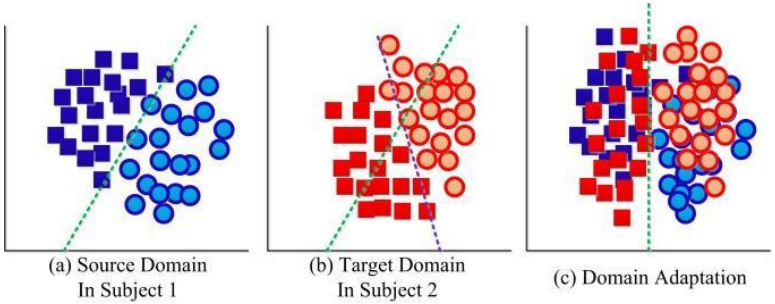
Domain adaptation

Neural networks extract features from images and use them for classification.

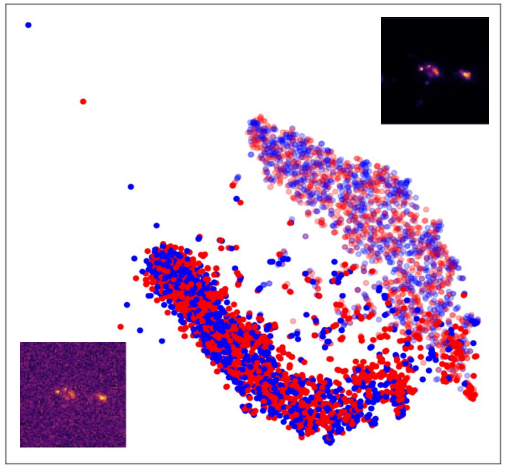


- Neural networks try to find a decision boundary between clusters in the feature space which represent different classes.
- Will the distribution of features for images from a different domain match the one NN was trained on?

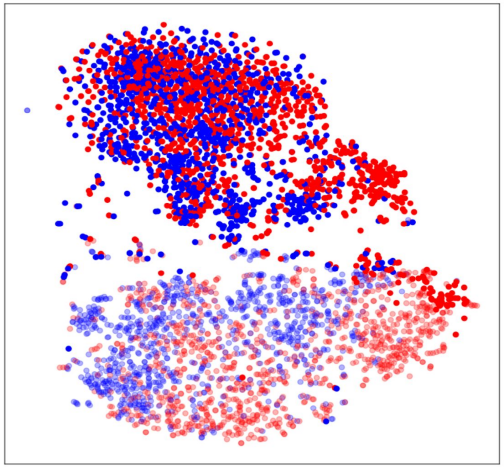
Domain adaptation



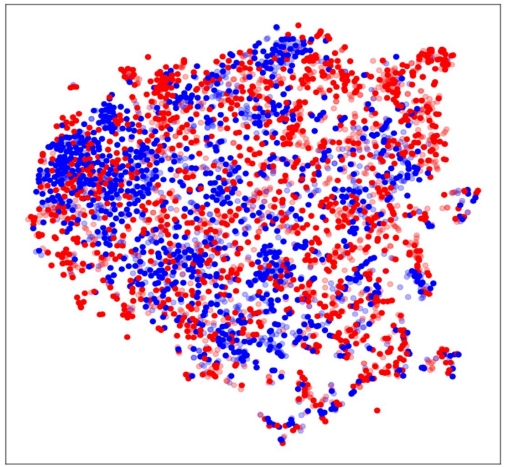
Start of the training



No domain adaptation

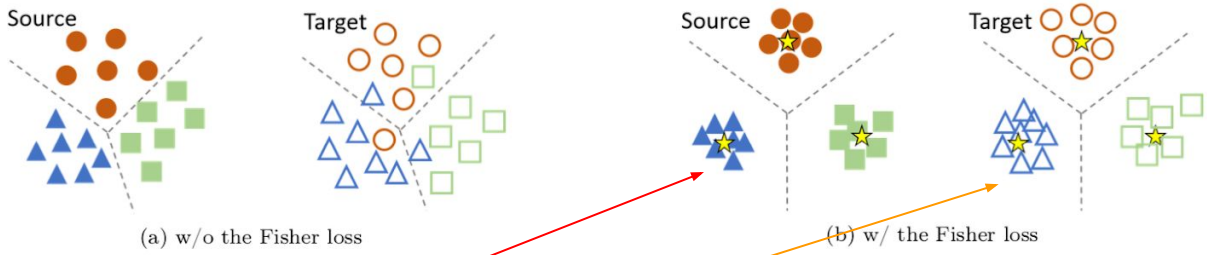


With domain adaptation



Fermilab

Domain adaptation



Zhang et al. (2020)

Total loss:

$$\min_{G_f, G_y} \mathcal{L}_y + \lambda_0 \cdot \mathcal{L}_{Fisher} + \lambda_1 \cdot \Omega - \lambda_2 \cdot \mathcal{L}_{TL},$$

Task loss | Fisher loss | Entropy minimization | Transfer loss

- Task loss - whatever we need for our problem
- Fisher loss - enforces between class separation and within class compactness
- Entropy - pushes each of the target domain samples towards one of the centers
- Transfer loss - empirical estimate of the domain discrepancy

Fermilab

Task loss - very often categorical cross-entropy loss

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

Fisher loss - Enforces within class compactness and between class separability.

$$\mathcal{L}_{Fisher} = \alpha(\text{tr}(\mathbf{S}_w), \text{tr}(\mathbf{S}_b))$$

$$\mathbf{S}_w = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\mathbf{h}_{k,j} - \mathbf{c}_k)(\mathbf{h}_{k,j} - \mathbf{c}_k)^T$$

$$\mathbf{S}_b = \sum_{k=1}^{K} (\mathbf{c}_k - \mathbf{c})(\mathbf{c}_k - \mathbf{c})^T$$

Trace ratio $\mathcal{L}_{Fisher-TR} = \text{tr}(\mathbf{S}_w)/\text{tr}(\mathbf{S}_b).$

Trace difference $\mathcal{L}_{Fisher-TD} = \text{tr}(\mathbf{S}_w) - \lambda_b \cdot \text{tr}(\mathbf{S}_b),$

Entropy minimization - Discriminability in the target domain is not guaranteed - centers of classes are unknowns, so the Fisher loss cannot be applied.
Help - entropy minimization, which pushes each sample to one of the classes centers in the target domain - estimated by label predictor $p(y|\mathbf{h})$.

$$\Omega = -\sum_j \sum_{k=1}^{K} p(y_j = k|\mathbf{h}_j) \log p(y_j = k|\mathbf{h}_j)$$

Fermilab

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.

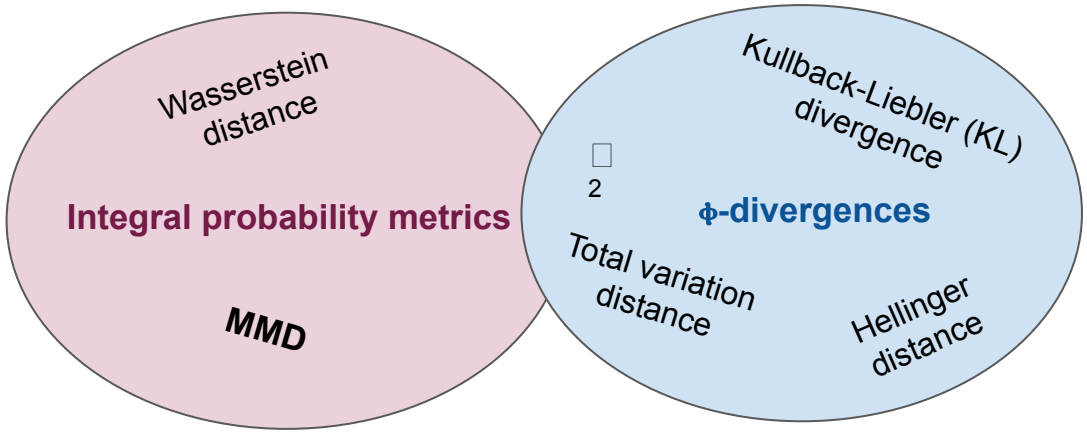
Maximum Mean Discrepancy - MMD

Want to learn more about MMD? Watch these two great videos:

- Kernel Methods, part 1 - Arthur Gretton - MLSS 2020, Tübingen
- Kernel Methods, part 2 - Arthur Gretton - MLSS 2020, Tübingen

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

Distance measures between probability distributions



$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int_M f \, d\mathbb{P} - \int_M f \, d\mathbb{Q} \right|$$

$$D_f(P \parallel Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ$$

Sriperumbudur et al. (2009)

Fermilab

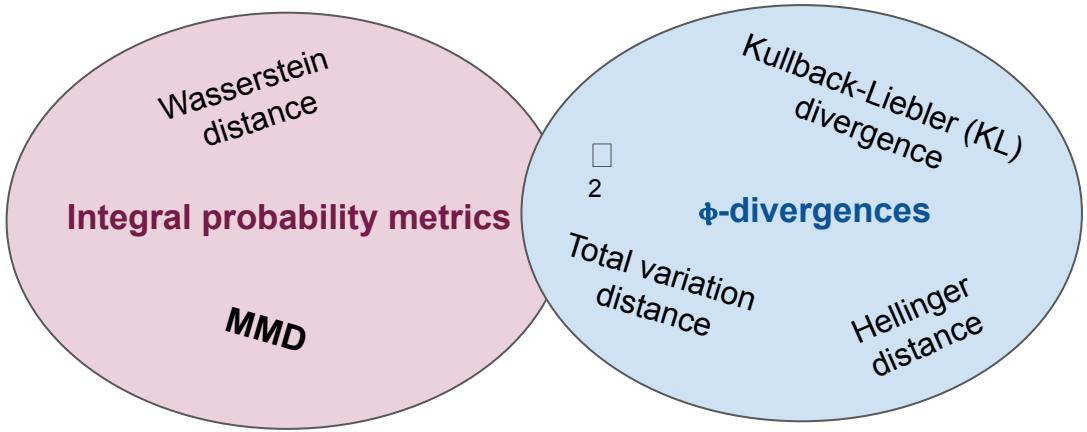
Maximum Mean Discrepancy - MMD

Want to learn more about MMD? Watch these two great videos:

- Kernel … Arthur Gretton … übingen
- Kernel … Arthur Gretton … Tübingen

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

Distance measures between probability distributions



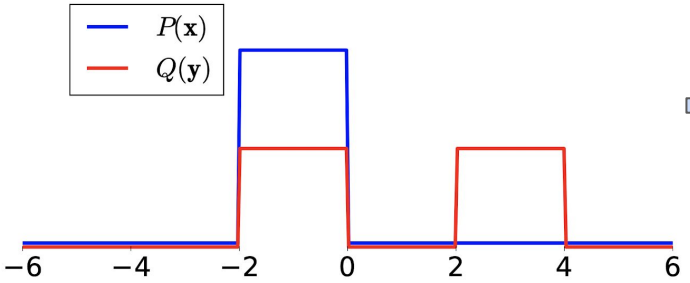
$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_M f\, d\mathbb{P} - \int_M f\, d\mathbb{Q} \right|$$

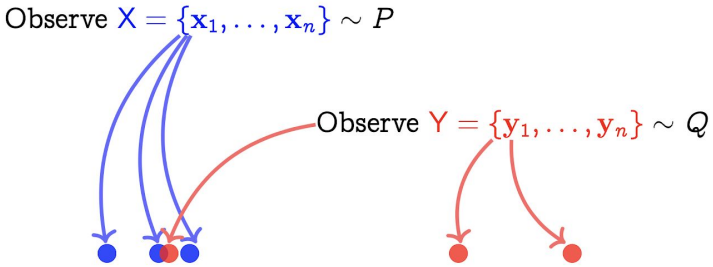
$$D_f(P \parallel Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ$$

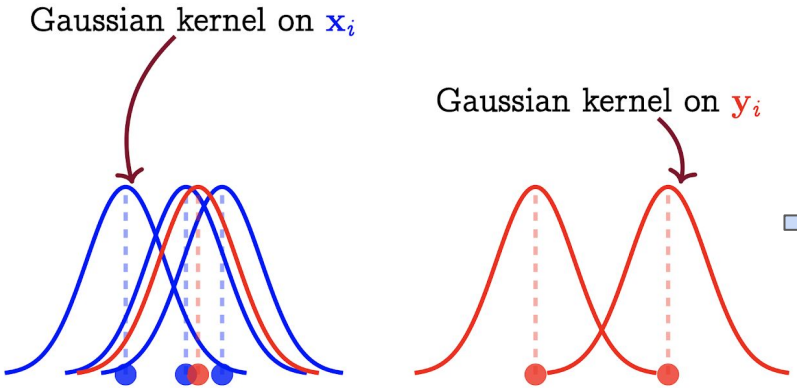
Sriperumbudur et al. (2009)

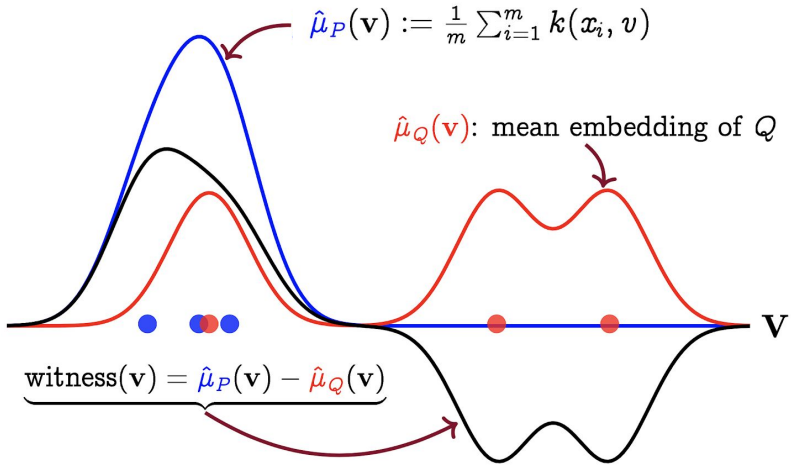
Fermilab

Are P and Q different?



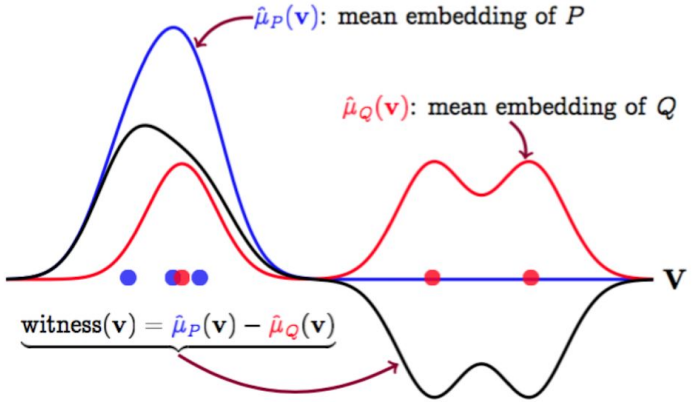






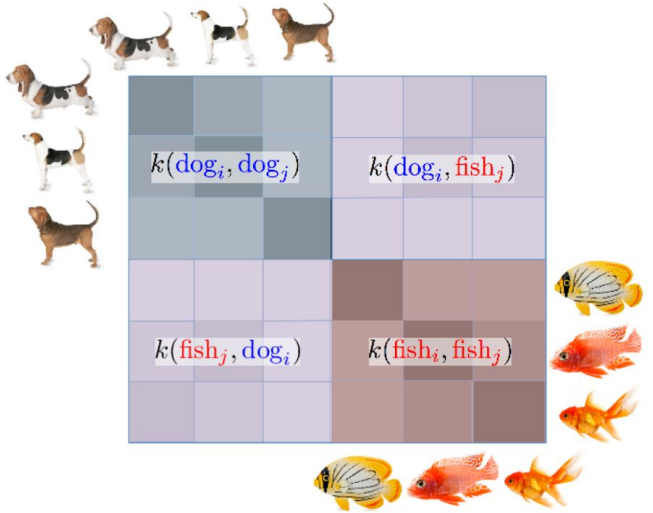
From Arthur Gretton (NIPS 2016 Workshop on Adversarial Learning, Barcelona Spain)

Fermilab



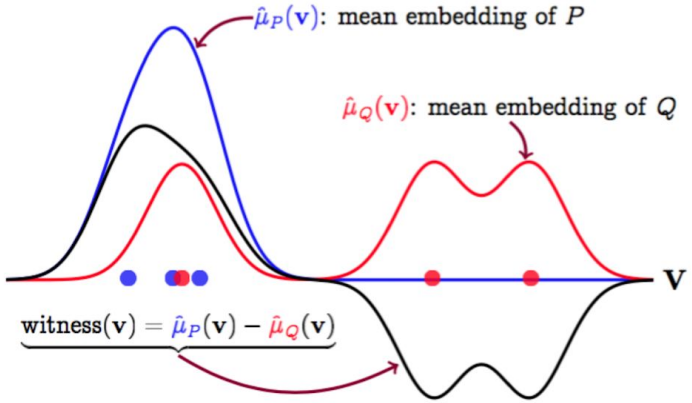
$$\widehat{MMD}^2 = \|\text{witness}(\mathbf{v})\|^2_{\mathcal{F}}$$
$$= \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{y}_i, \mathrm{y}_j)$$
$$- \frac{2}{n^2} \sum_{i,j} k(x_i, \mathrm{y}_j)$$

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



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

Fermilab



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

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

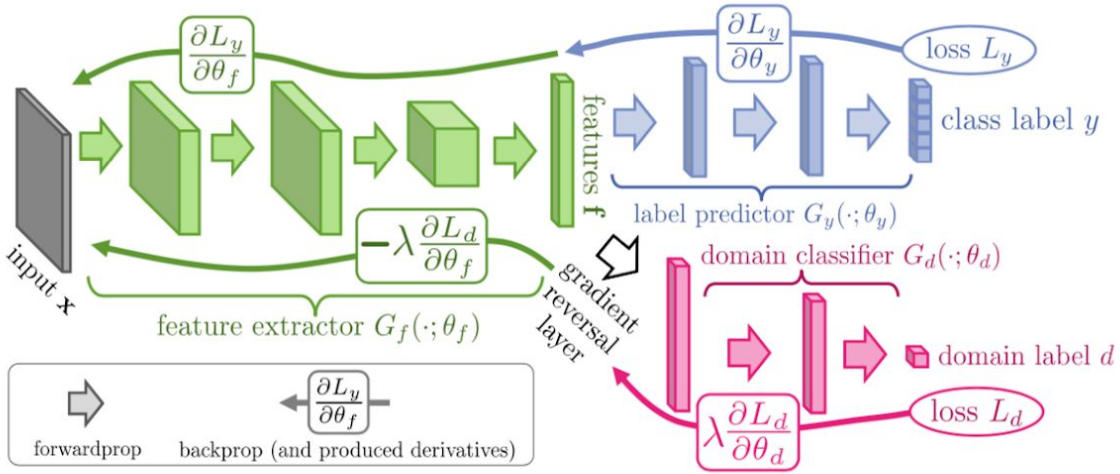
From Arthur Gretton (NIPS 2016 Workshop on Adversarial Learning, Barcelona Spain)

Fermilab

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)

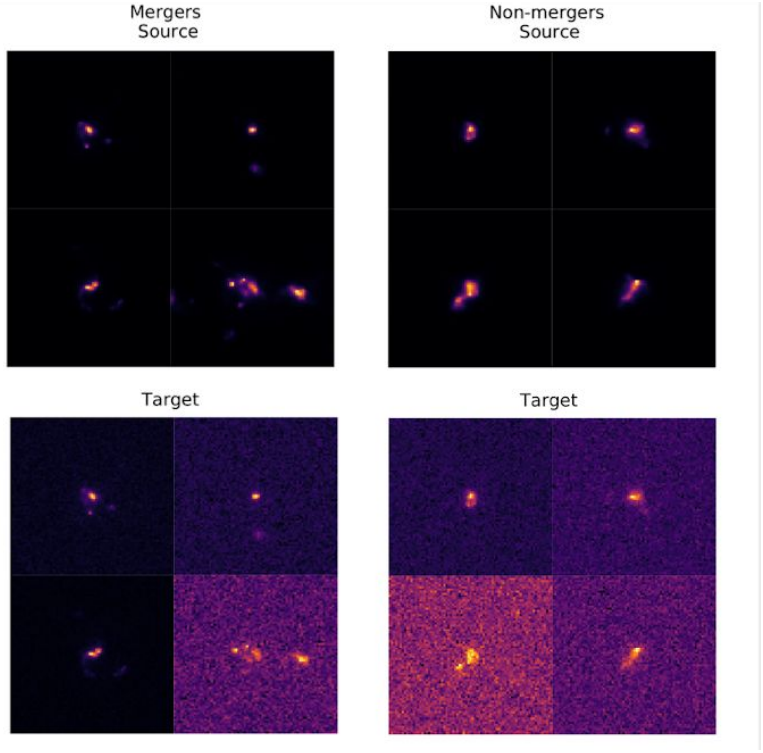
Fermilab

Experiments

1.
Domain adaptation
MMD
&
DANN

2.
Additional
Fisher Loss
&
Entropy min.

3.
Use two networks
DeepMerge
&
ResNet18

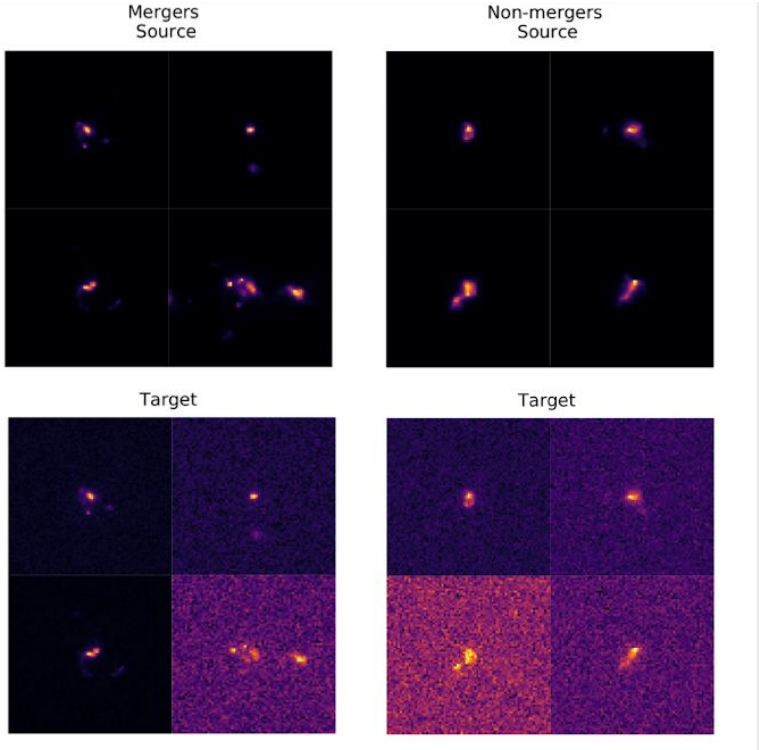


Fermilab

Results

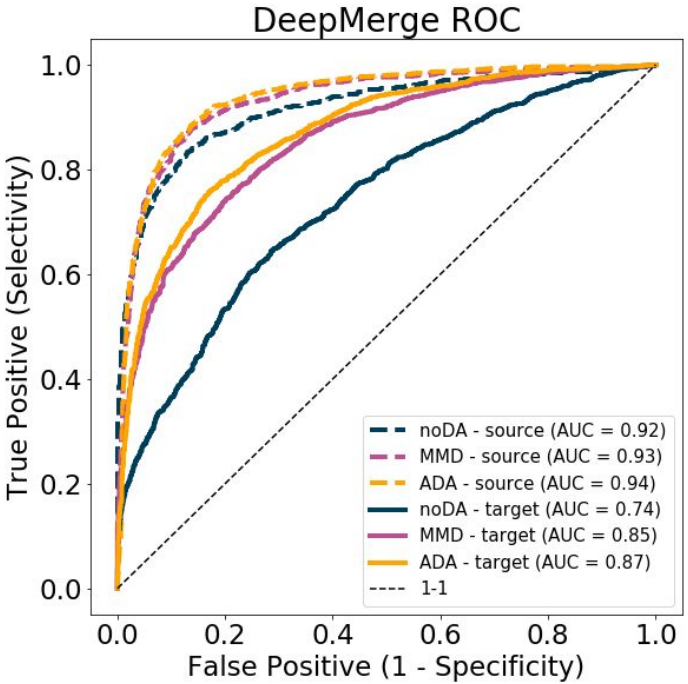
PRELIMINARY

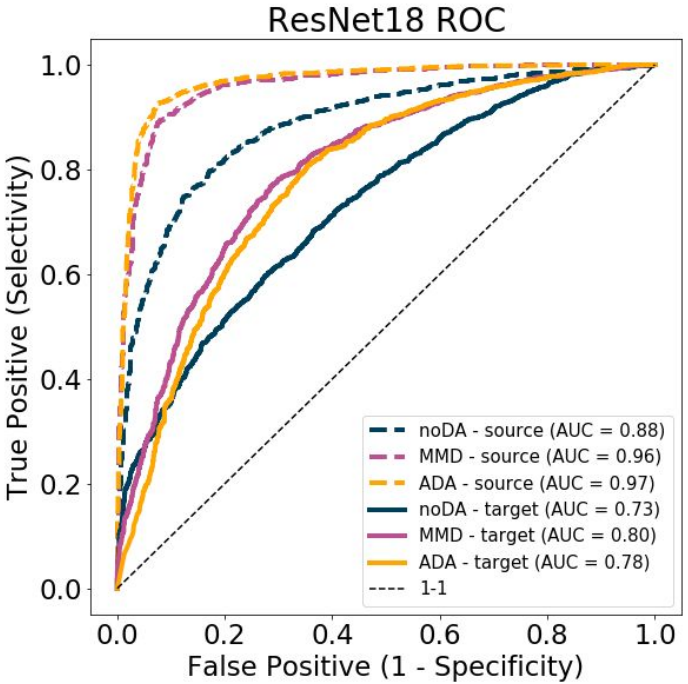
	Source Domain		Target Domain	
	DM	RN18	DM	RN18
noDA	85%	82%	58%	60%
MMD	87%	90%	77%	74%
DANN	87%	92%	**79%**	72%
MMD +F+E	84%	89%	77%	**75%**
DANN +F+E	87%	87%	74%	70%



Fermilab

Performance





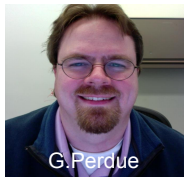
Ćiprijanović et al. (2020) @ NeurIPS

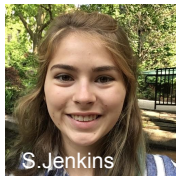
Fermilab

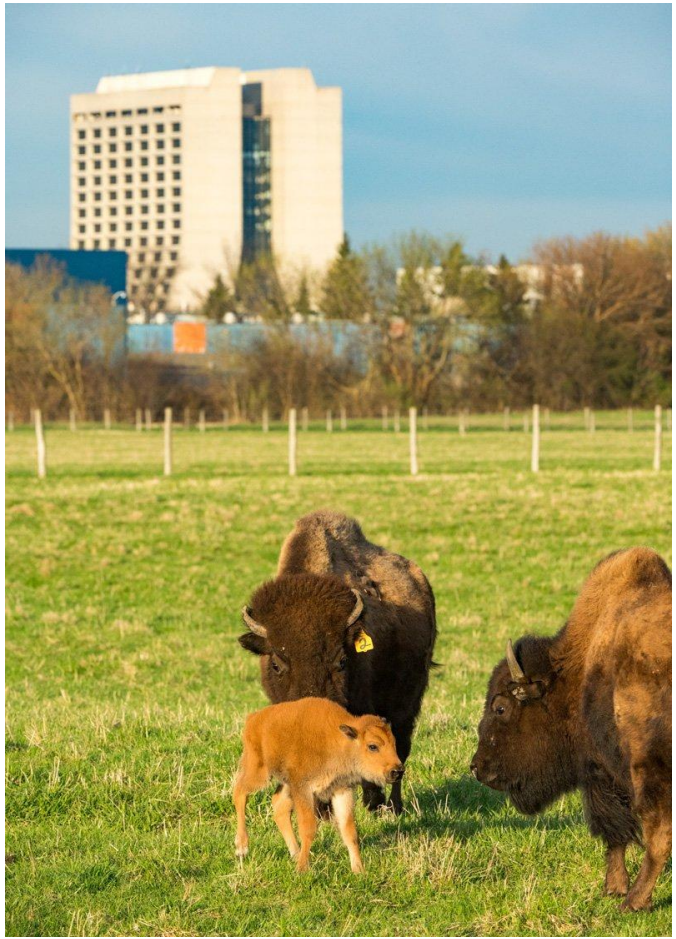
Summary & What's next?

- **Merging galaxies** are important for the study of galaxy morphology, but also evolution of structure in the Universe.
- We talked more about **MMD and DANNs** and introduced **Fisher Loss and Entropy minimization**.
- **Domain adaptation (DA) is crucial** for successful bridging between different data sets and full utilisation of ML in science.

- Working on domain transfer problems between **simulated and real telescope** images (Illustris to SDSS).
- 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?



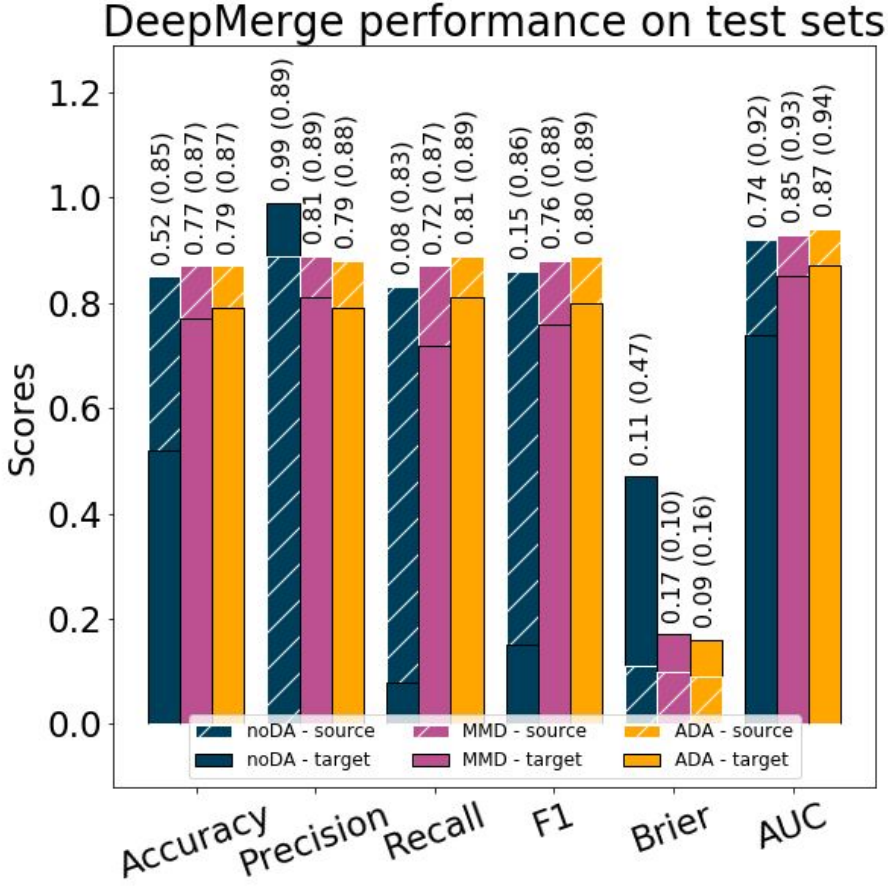


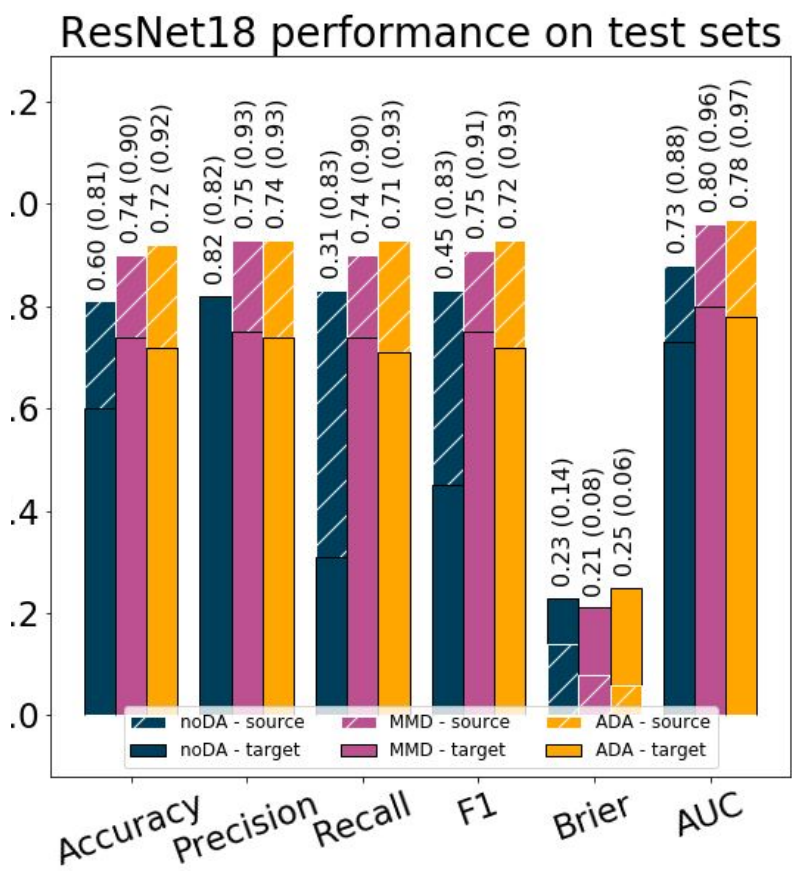


Thank you!

aleksand@fnal.gov

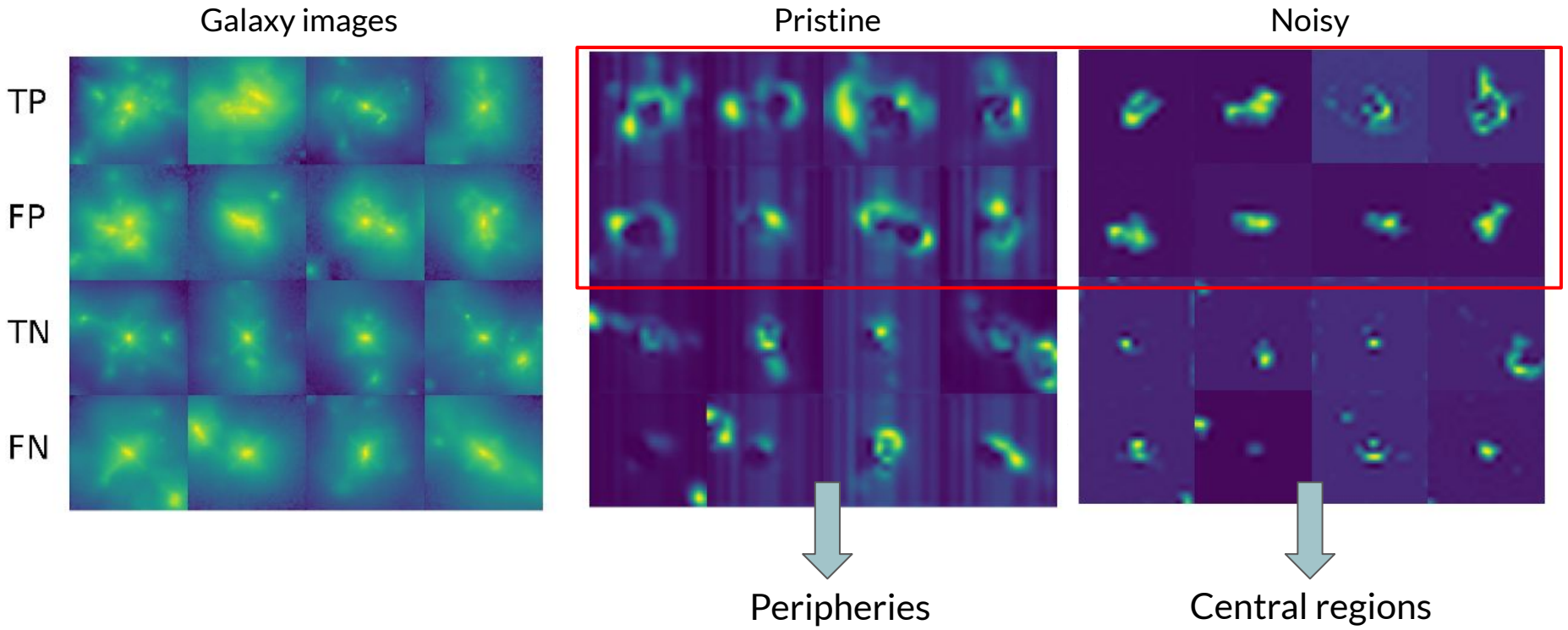
Appendix


DeepMerge performance on test sets
Scores
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.52 (0.85)
0.77 (0.87)
0.79 (0.87)
0.99 (0.89)
0.81 (0.89)
0.79 (0.88)
0.08 (0.83)
0.72 (0.87)
0.81 (0.89)
0.15 (0.86)
0.76 (0.88)
0.80 (0.89)
0.11 (0.47)
0.17 (0.10)
0.09 (0.16)
0.74 (0.92)
0.85 (0.93)
0.87 (0.94)
noDA - source
noDA - target
MMD - source
MMD - target
ADA - source
ADA - target
Accuracy
Precision
Recall
F1
Brier
AUC


ResNet18 performance on test sets
0.60 (0.81)
0.74 (0.90)
0.72 (0.92)
0.82 (0.82)
0.75 (0.93)
0.74 (0.93)
0.31 (0.83)
0.74 (0.90)
0.71 (0.93)
0.45 (0.83)
0.75 (0.91)
0.72 (0.93)
0.23 (0.14)
0.21 (0.08)
0.25 (0.06)
0.73 (0.88)
0.80 (0.96)
0.78 (0.97)
noDA - source
noDA - target
MMD - source
MMD - target
ADA - source
ADA - target
Accuracy
Precision
Recall
F1
Brier
AUC

Fermilab

Visualizing important regions for classification with Gradient-weighted Class Activation Mapping (Grad-CAM)



Fermilab

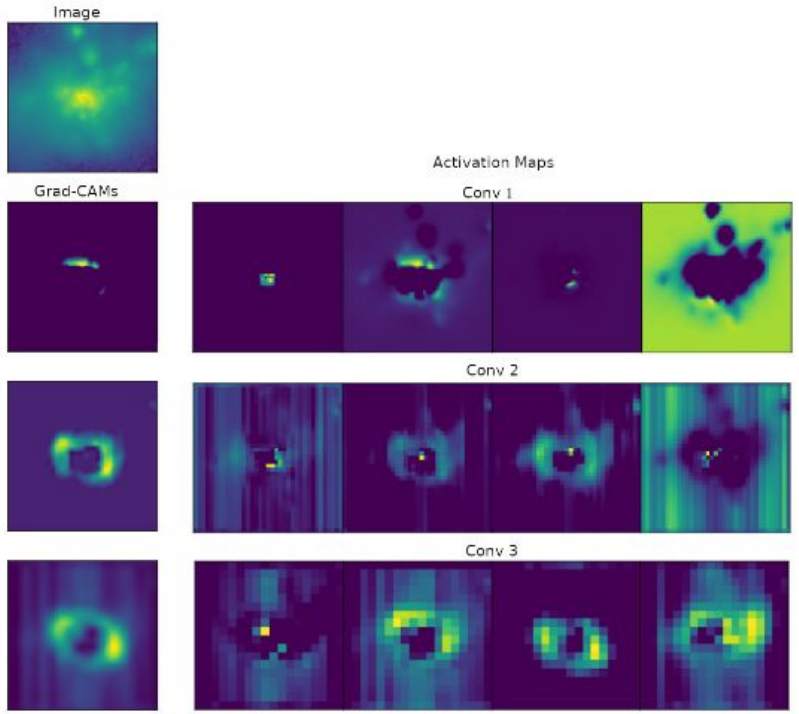
Visualizing important regions for classification with Gradient-weighted Class Activation Mapping (Grad-CAM)

Derived by calculating gradients of the score for a given class (before the activation function, with respect to feature map of a convolutional layer (usually the last one).

Salvaraju et al. (2016)

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$



Map shows which pixels were the most important for classification.

Fermilab