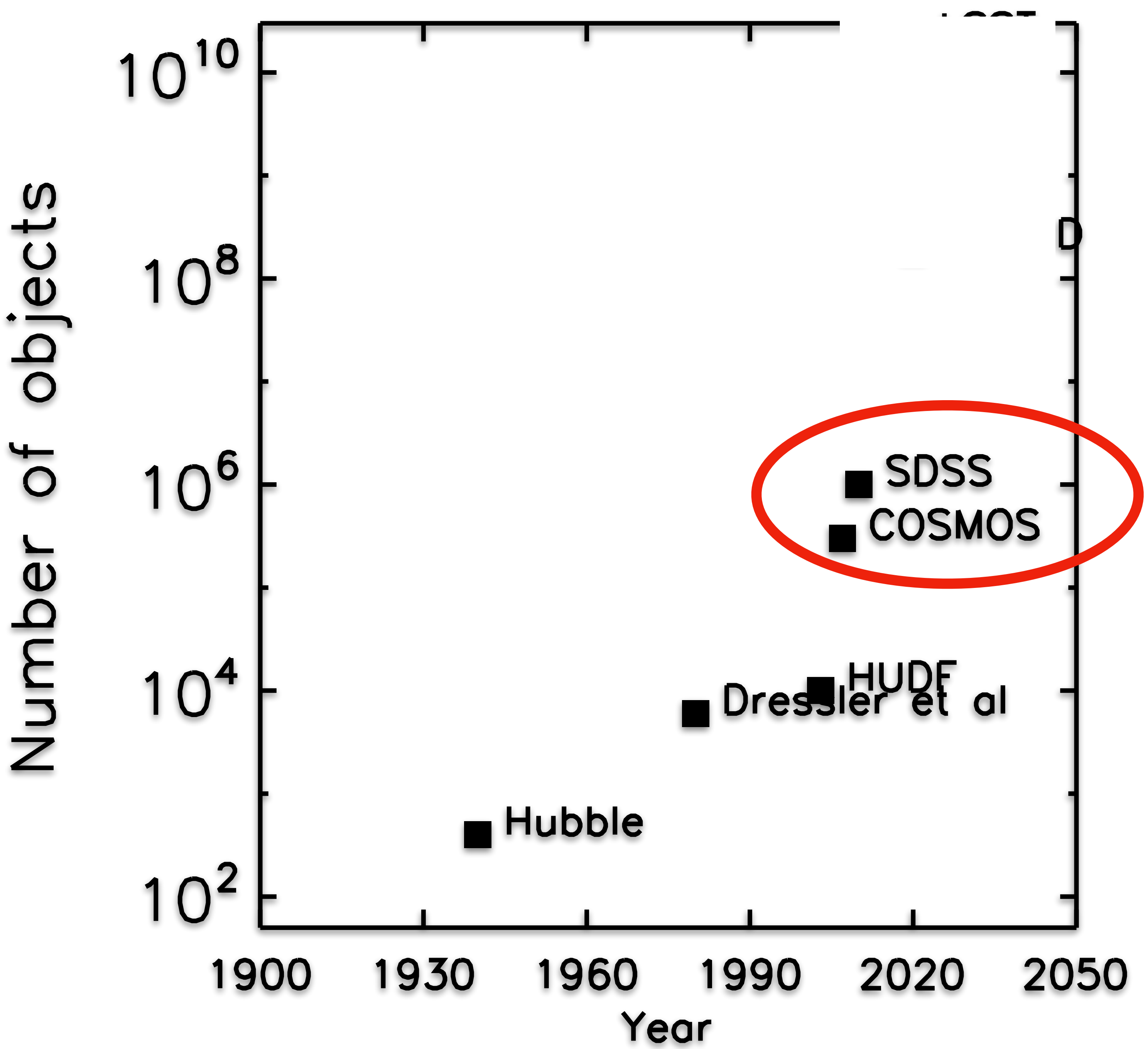
# The impact of deep learning applied to galaxy surveys

Marc Huertas-Company

Cuarta Sinergia SEA Inteligencia Artificial y Astrofísica - 09/02/22

# THE ERA OF STATISTICS

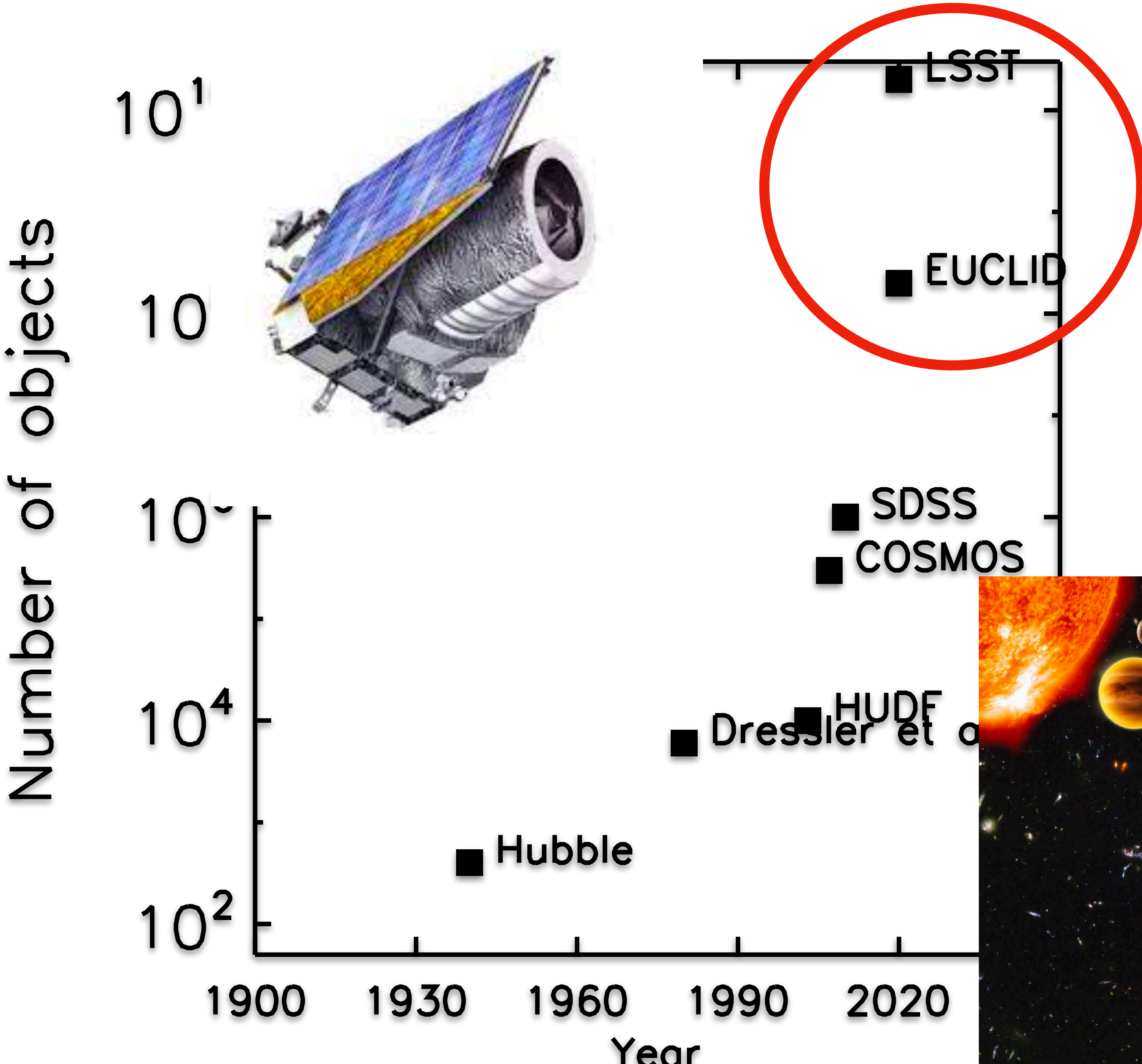



## IN THE LAST ~20 YEARS ….

**THE FIELD OF OBSERVATIONAL GALAXY EVOLUTION HAS EVOLVED INTO A STATISTICAL SCIENCE**

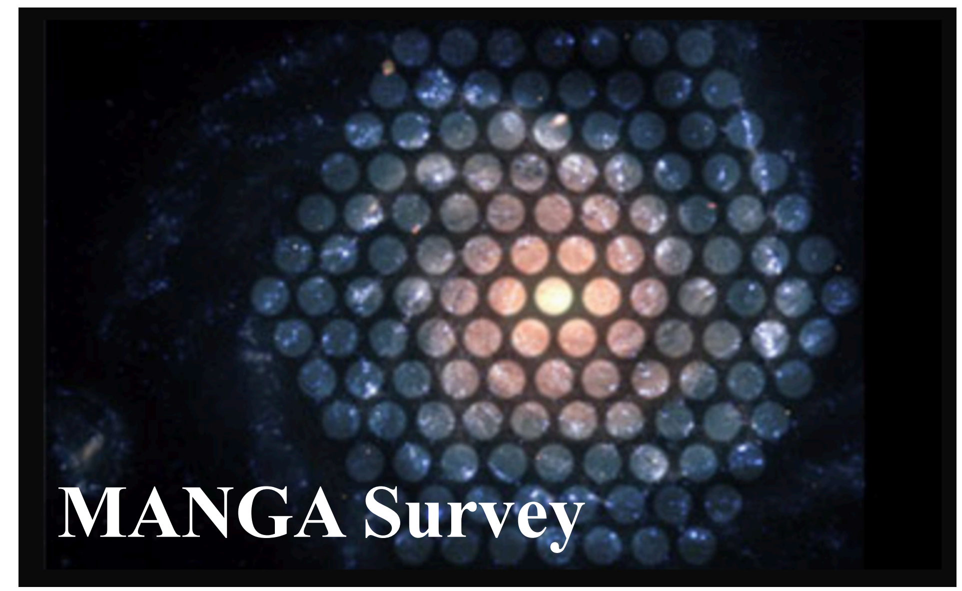
# THE DATA-DRIVEN ASTRONOMY ERA

## BIGGER




**JAMES WEBB SPACE TELESCOPE**

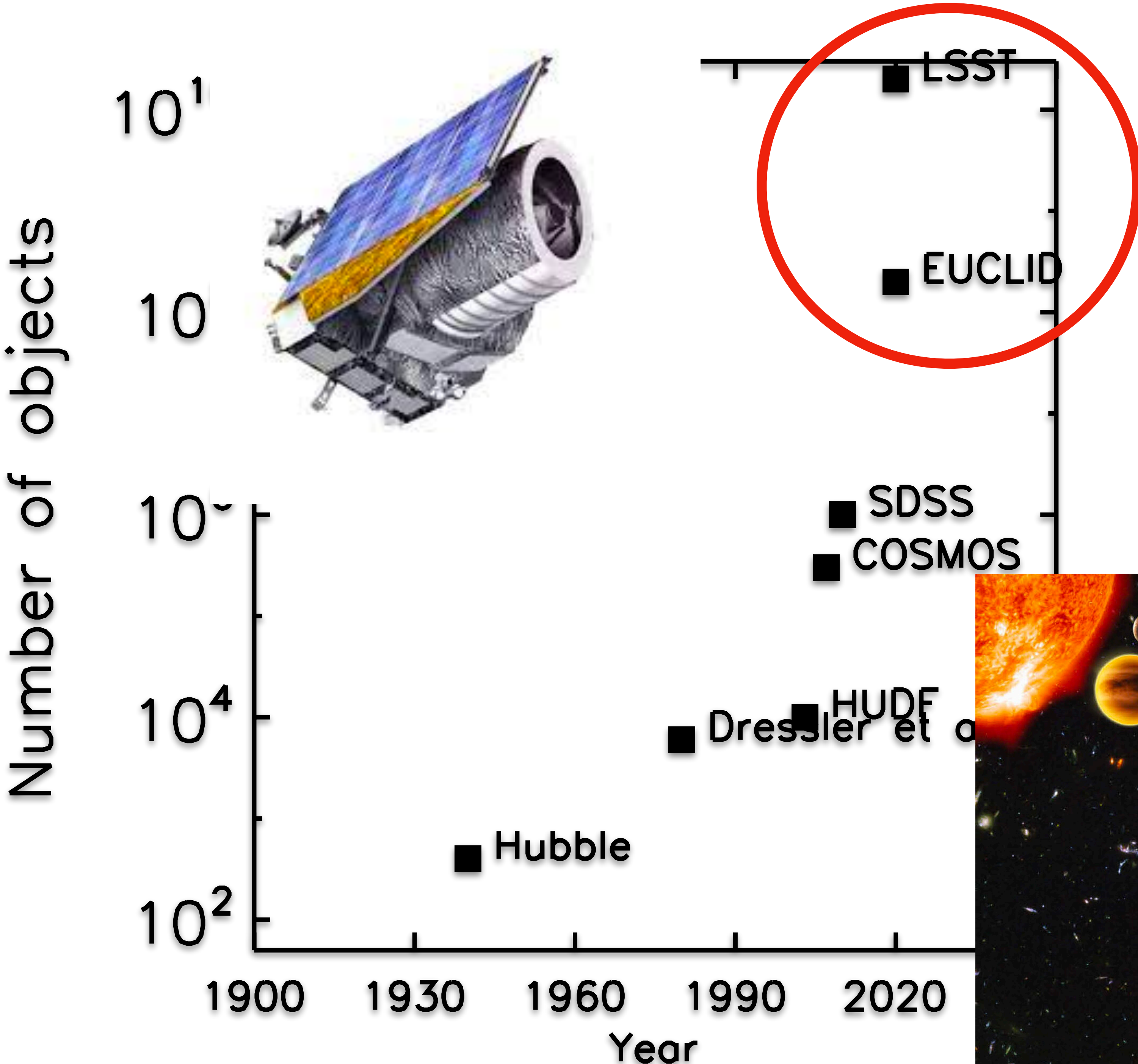
## AND MORE COMPLEX

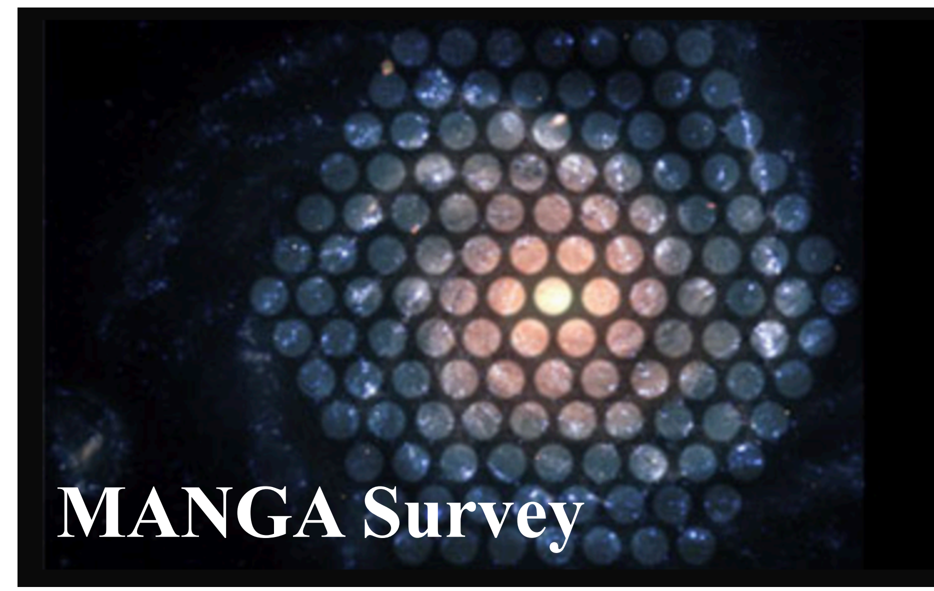



# THE DATA-DRIVEN ASTRONOMY ERA

**BIGGER**

**AND MORE COMPLEX**

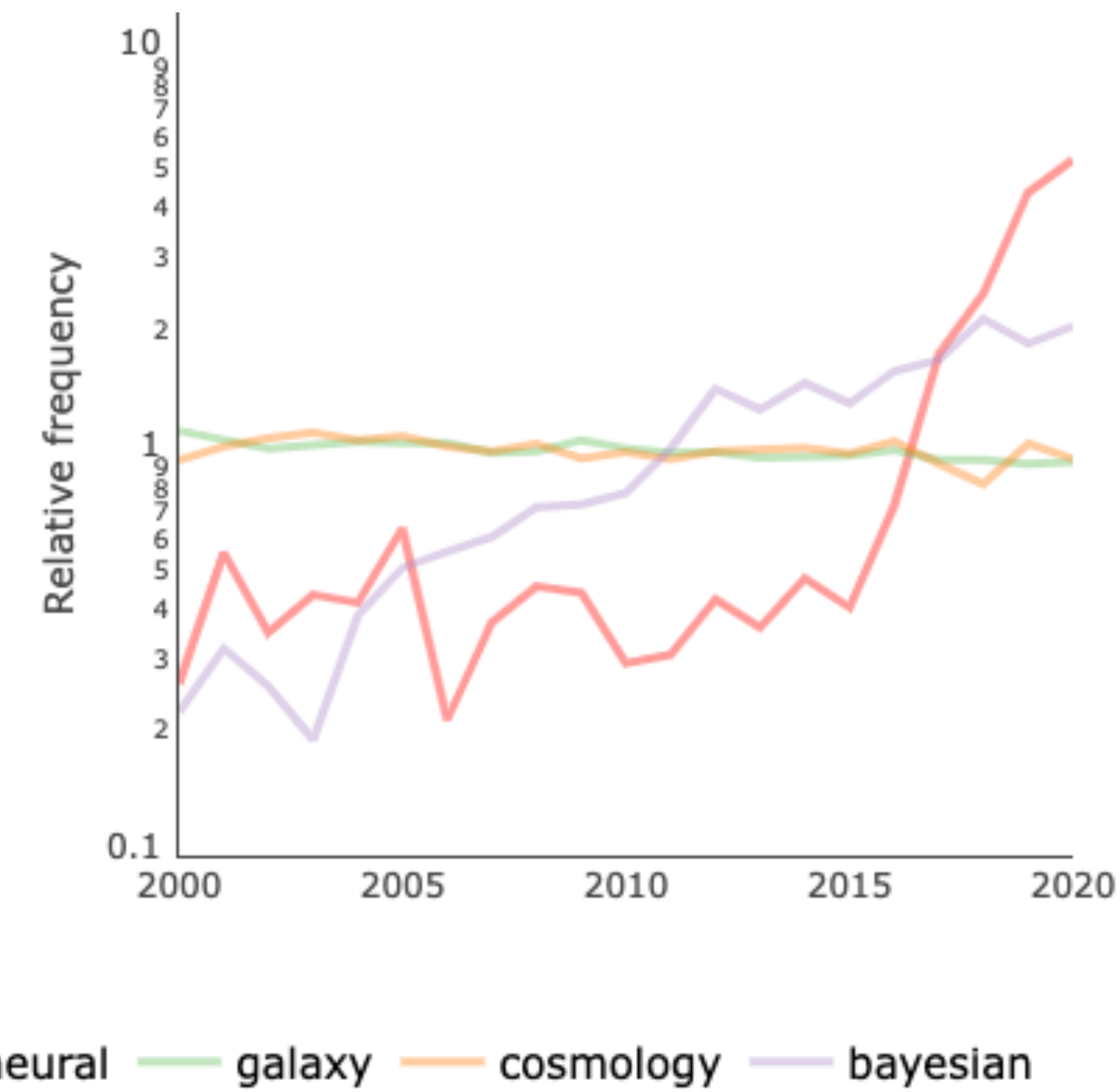




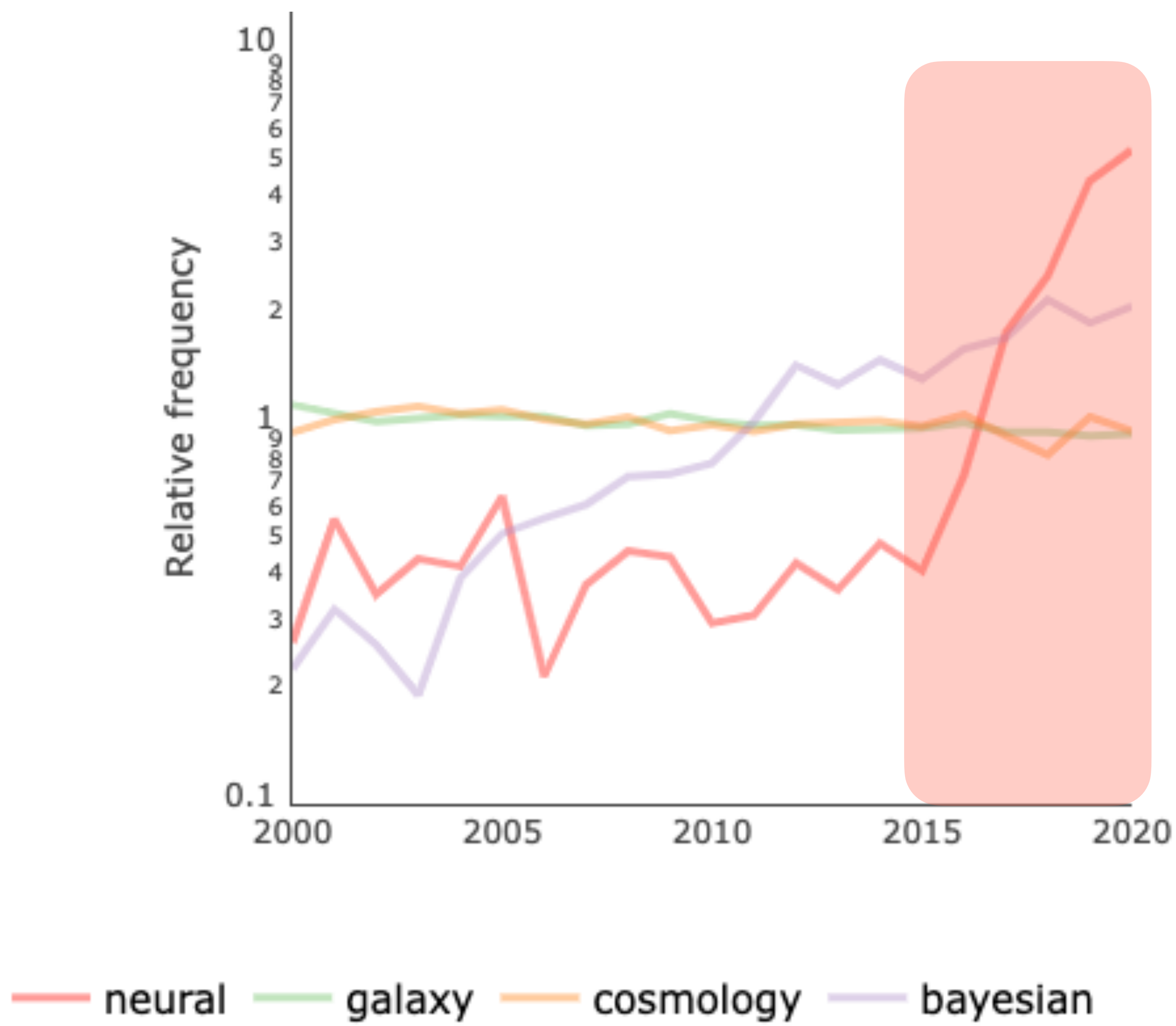
**JAMES WEBB SPACE TELESCOPE**

# Evolution of keywords frequency in astro-ph papers




**Source: ArXivSorter**

# Evolution of keywords frequency in astro-ph papers




**Source: ArXivSorter**

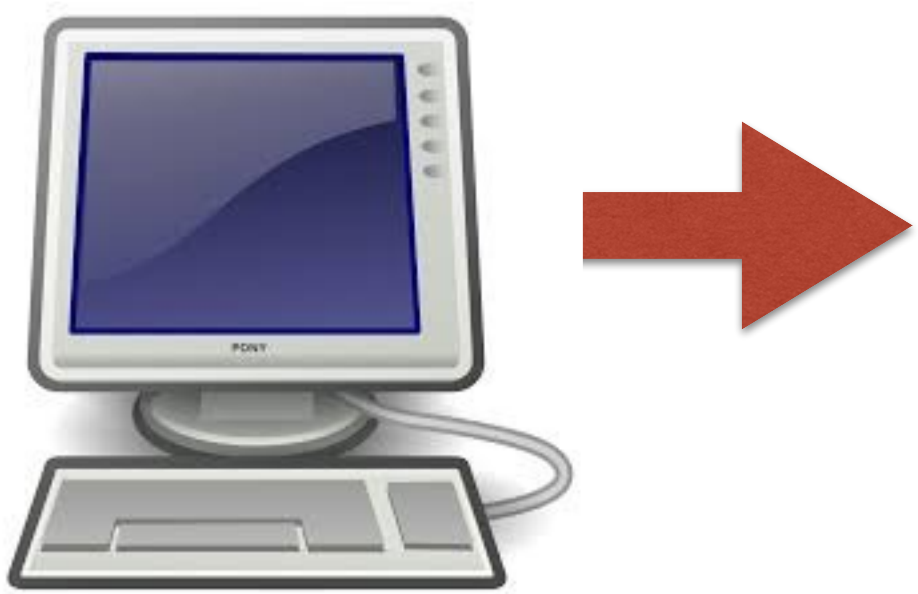
# BEFORE 2012....ML WINTER

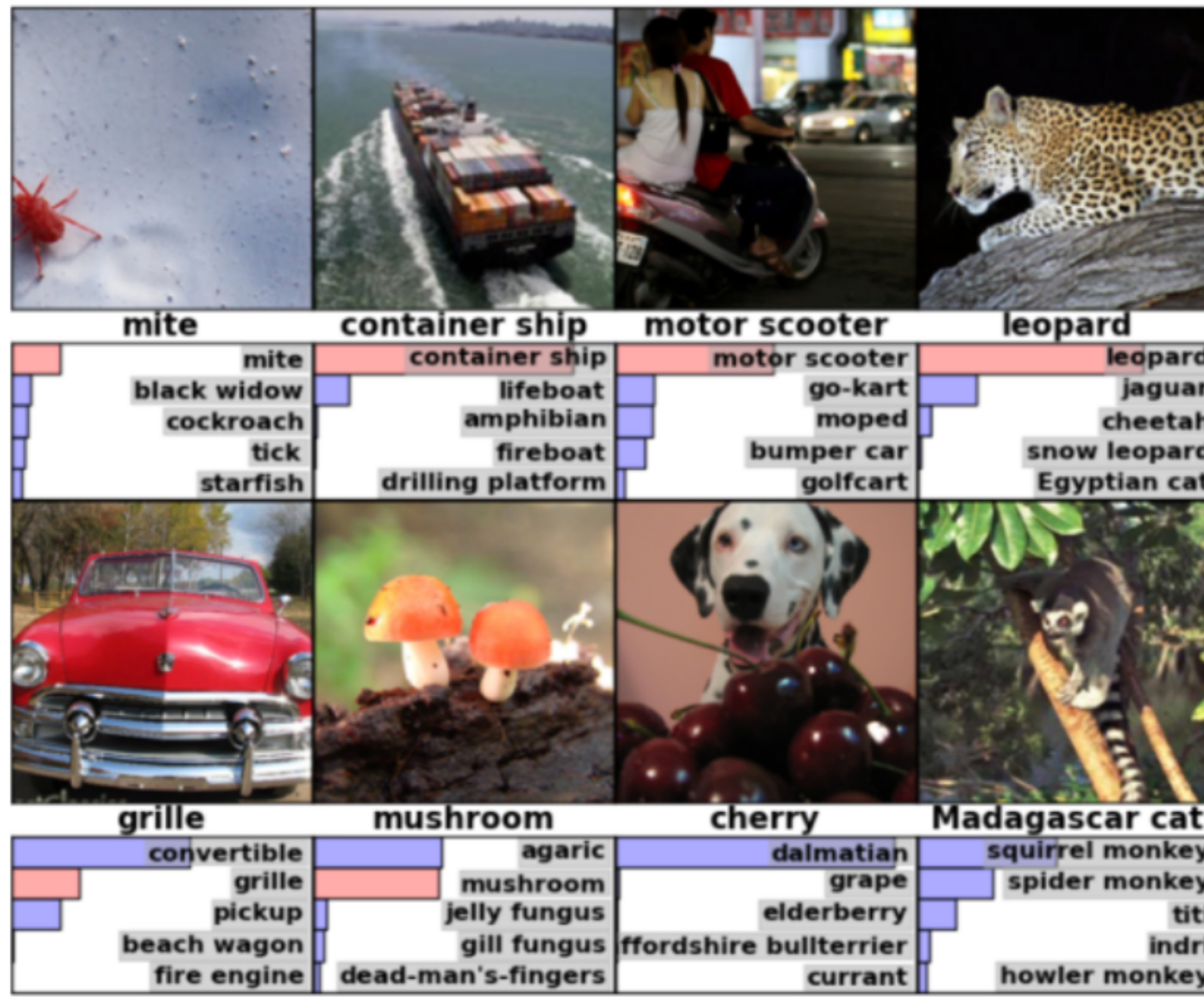
CAT?

DOG?

**TRIVIAL HUMAN TASKS REMAINED CHALLENGING FOR COMPUTERS**

# 2012+: the deep learning era





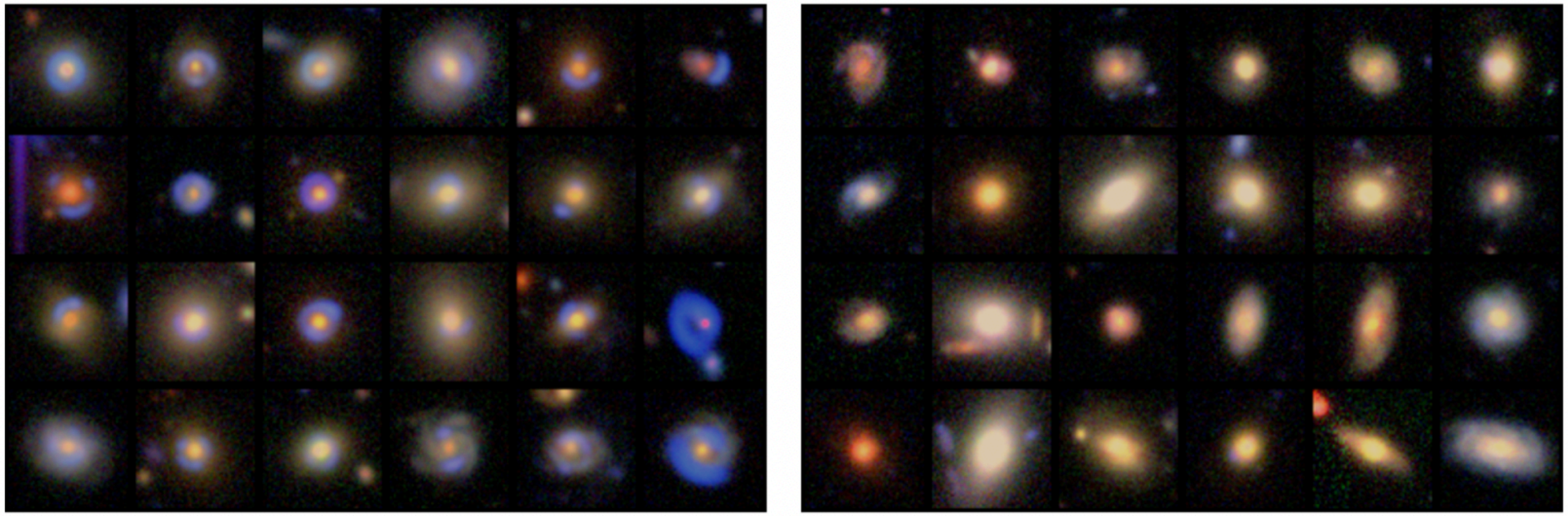

**IT HAS BECOME TRIVIAL....**

# A decade of deep learning, six years in astronomy...

# what type of applications?

| | | |
|---|---|---|
| "Incremental Science" | 1. Deep Learning for low level data processing | close to standard computer vision applications for natural images with no or reduced domain specific content |
| | 2. Deep Learning for galaxy properties | used to replace existing algorithms with a faster and more efficient solution |
| | 3. Deep Learning for emulation of simulations | similar to 2 but forward modelling - learning the galaxy-halo connection |
| "Potentially New" | 4. Deep Learning for assisted discovery | data exploration and visualization of complex datasets, identification of new objects |
| | 5. Deep Learning for hidden correlations | properties of galaxies which are not directly accessible with known observables |
| | 6. Deep Learning to constrain cosmology | bypass summary statistics and constrain models using all available data |

# 1. Deep Learning for low level data processing

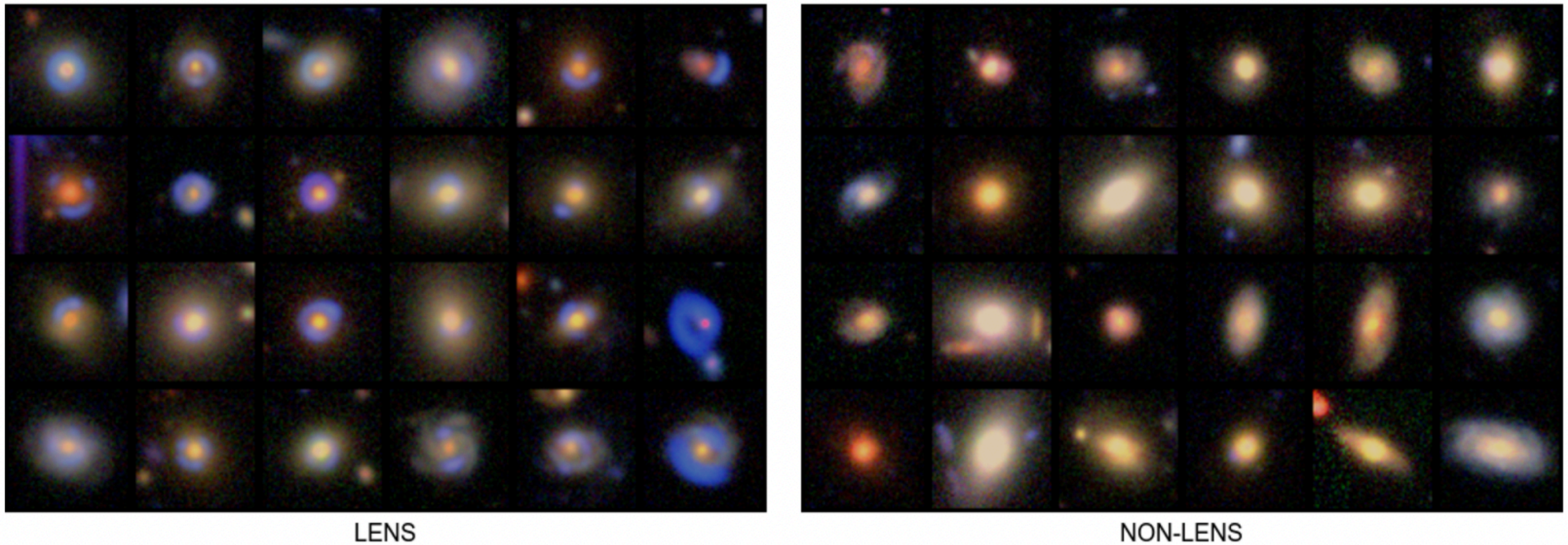


LENS NON-LENS

**Rare events, valuable to constrain cosmology. Future surveys will increase the samples by orders of magnitude.**

Jacobs+17

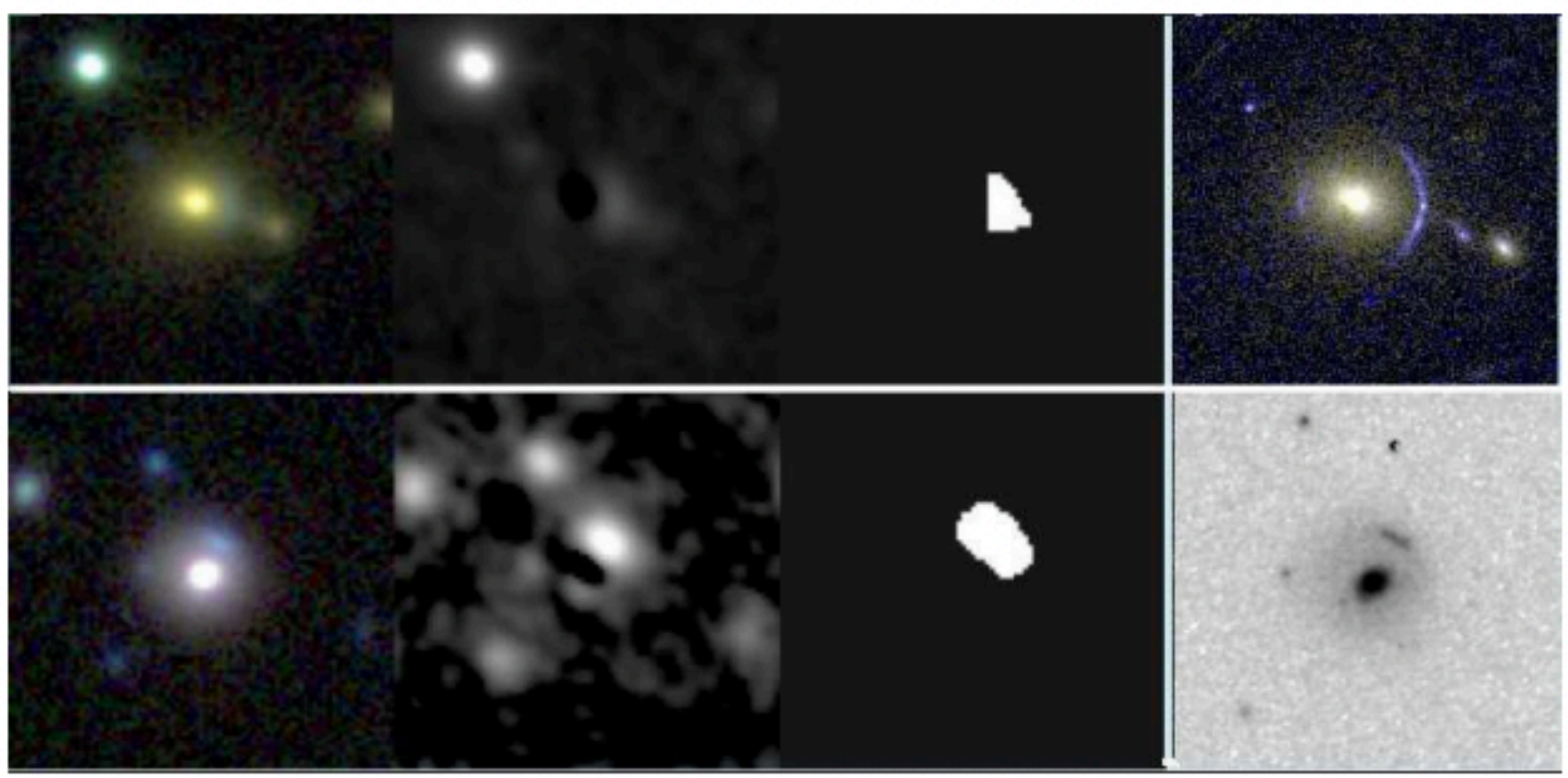
# 1. Deep Learning for low level data processing




Rare events, valuable to constrain cosmology. Future surveys will increase the samples by orders of magnitude.

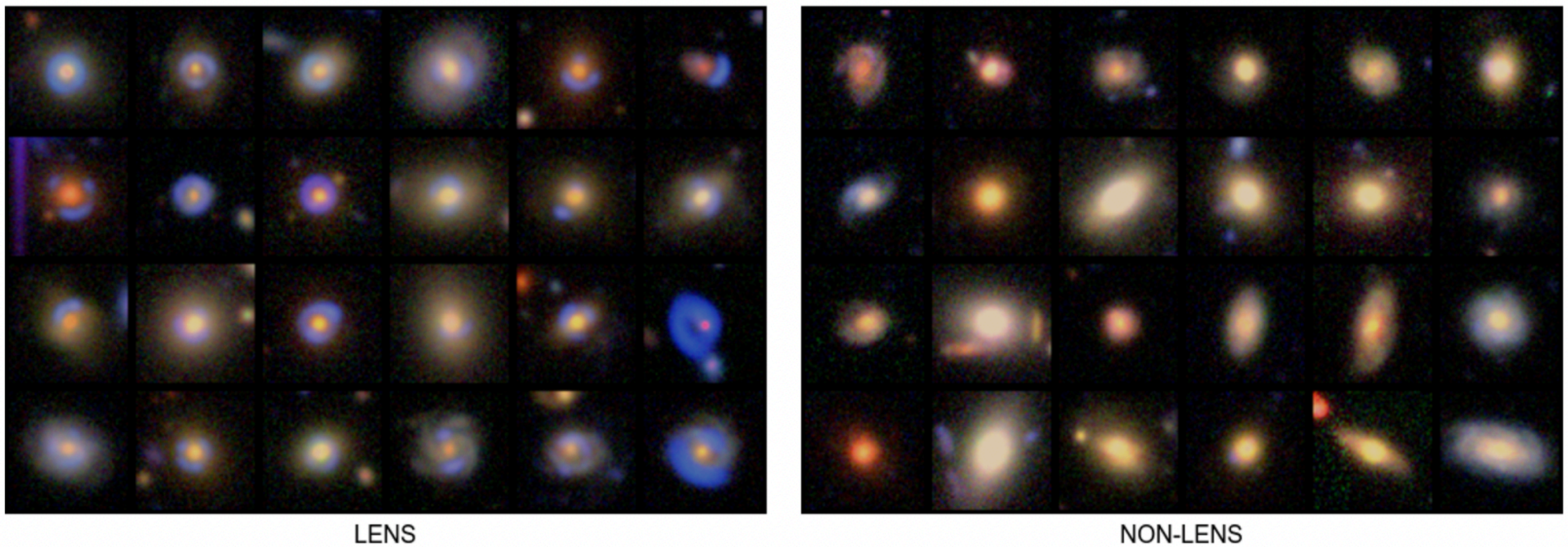
**Jacobs+17**

"Pre Deep Learning" Approach



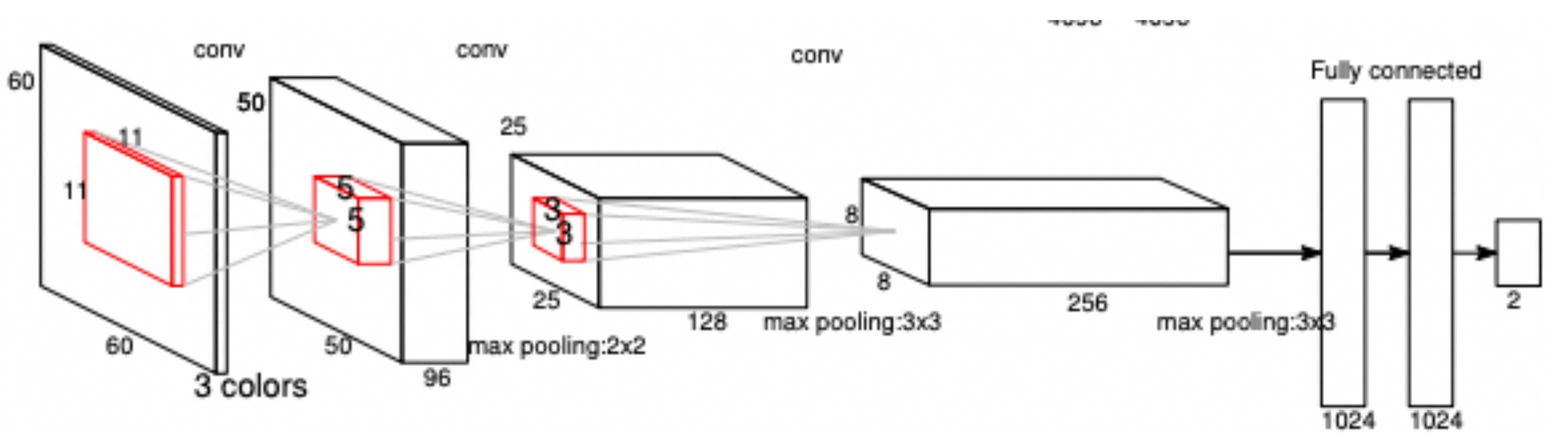
**Gavazzi+17**

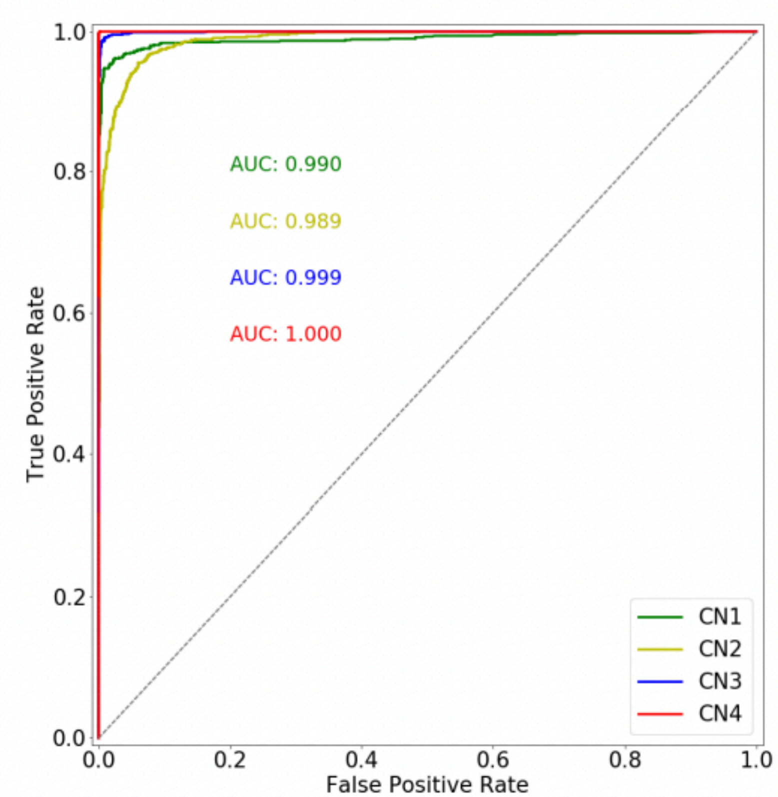
# 1. Deep Learning for low level data processing




Rare events, valuable to constrain cosmology. Future surveys will increase the samples by orders of magnitude.

**Jacobs+17**



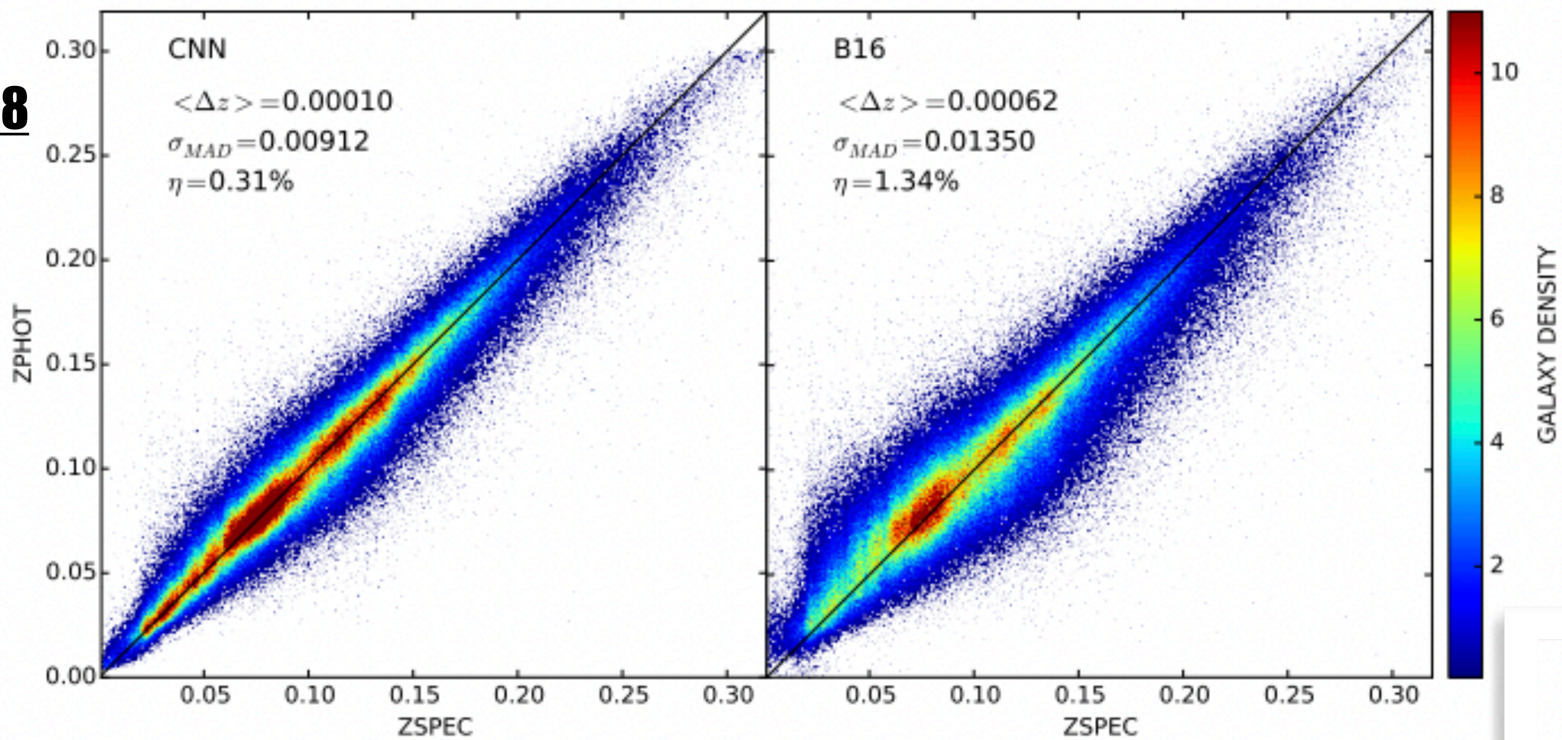




**This is a change of paradigm from an algorithmic centric focus to a purely data driven approach to data**

Dieleman+15, Huertas-Company+15, Aniyan+17, Charnock+17, Gieseke+17, Jacobs+17, Petrillo+17, Schawinski+17, Alhassan+18, Dominguez-Sanchez+18, George+18, Hinners+18, Lukic+18, Moss+18, Razzano+18, Schaefer+18, Allen+19, Burke+19, Carrasco-Davis+19, Chatterjee+19, Davies+19, Dominguez-Sanchez+19, Fusse+19, Glaser+19, Ishida+19, Jacobs+19, Katebi+19, Lanusse+19Liu+19, Lukic+19, Metcalf+19, Muthukrishna+19, Petrillo+19, Reiman+19, Boucaud+20, Chiani+20, Ghosh+20, Gomez+20, Hausen+20, Hlozek+20, Hosseinzadeh+20. Huang+20, Li+20, Moller+20, Paillassa+20, Tadaki+20, Vargas dos Santos+20, Walmsley+20, Wei+20, Allam+21, Arcelin+21, Becker+21, Bretonniere+21, Burhanudin+21, Davison+21, Donoso-Oliva+21, Jia+21, Lauritsen+21, Ono+21, Ruan+21, Sadegho+21, Tang+21, Tanoglidis+21, Vojtekova+21, Dhar+22Hausen+22, Orwat+22, Pimentel+22, Rezaei+22, Samudre+22, Shen+22, Walmsley+22

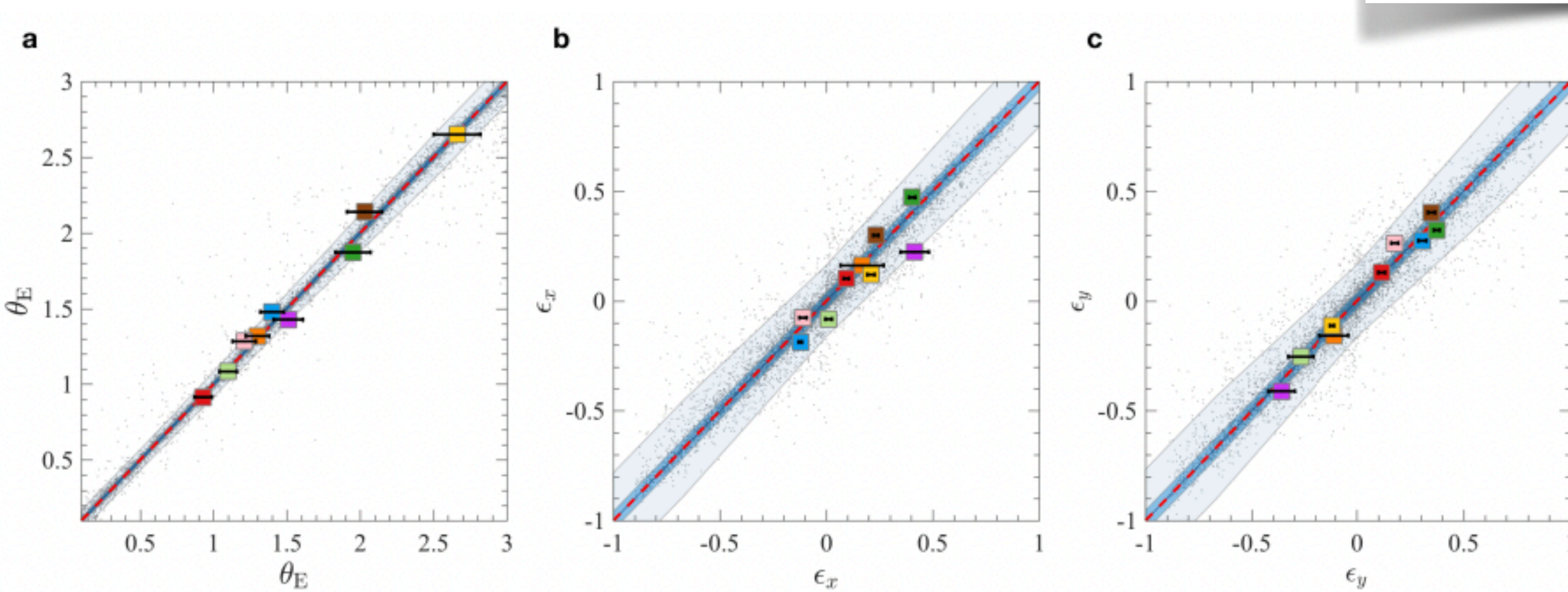
# 2. Deep Learning for galaxy properties (inference)

**Pasquet+18**




## Photometric Redshifts

**Fast, Algorithm-free, no summary statistics**

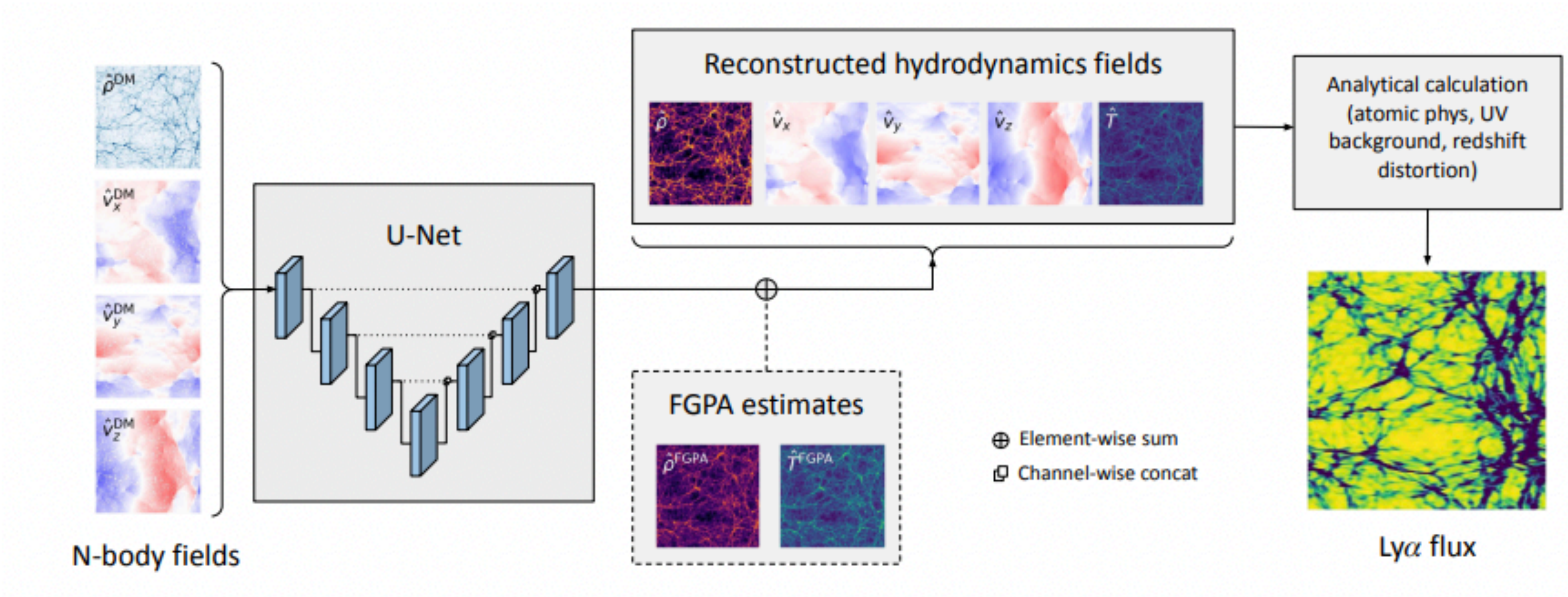



## Strong Lensing

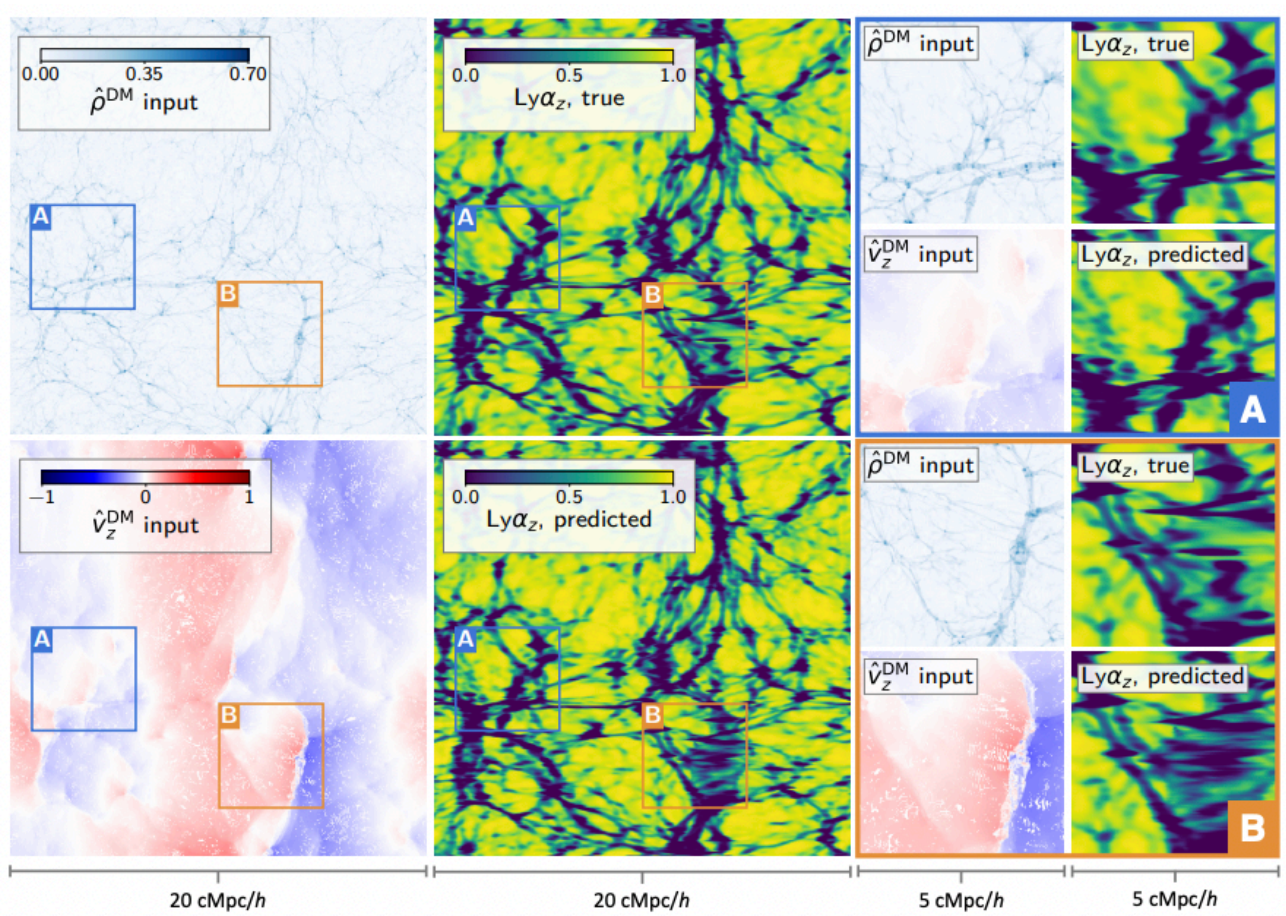
**Hazevah+17**

Hazevah+17, Perreault-Levasseur+17, Morningstar+18, Stark+18, Tuccillo+18, Bom+19, Lovell+19, Madireddy+19, Pearson+19, Simet+19, Pasquet+19, Wu+19, Menou+19, Aragon-Calvo+20, Chianese+20, Stahl+20, Surana+20, Shuntov+20, Cabayol-Garcia+20, Campagne+20, Buck+21, Grover+21, Hayat+21, Li+21, Maresca+21, Qiu+21, Rhea+21, Schuldt+21, Tohill+21, Yao-Yu+21, Dey+21, Lee+21, Zhou+21, Henghes+21, Ansari+21

# 3. Deep Learning for emulating (cosmological) simulations



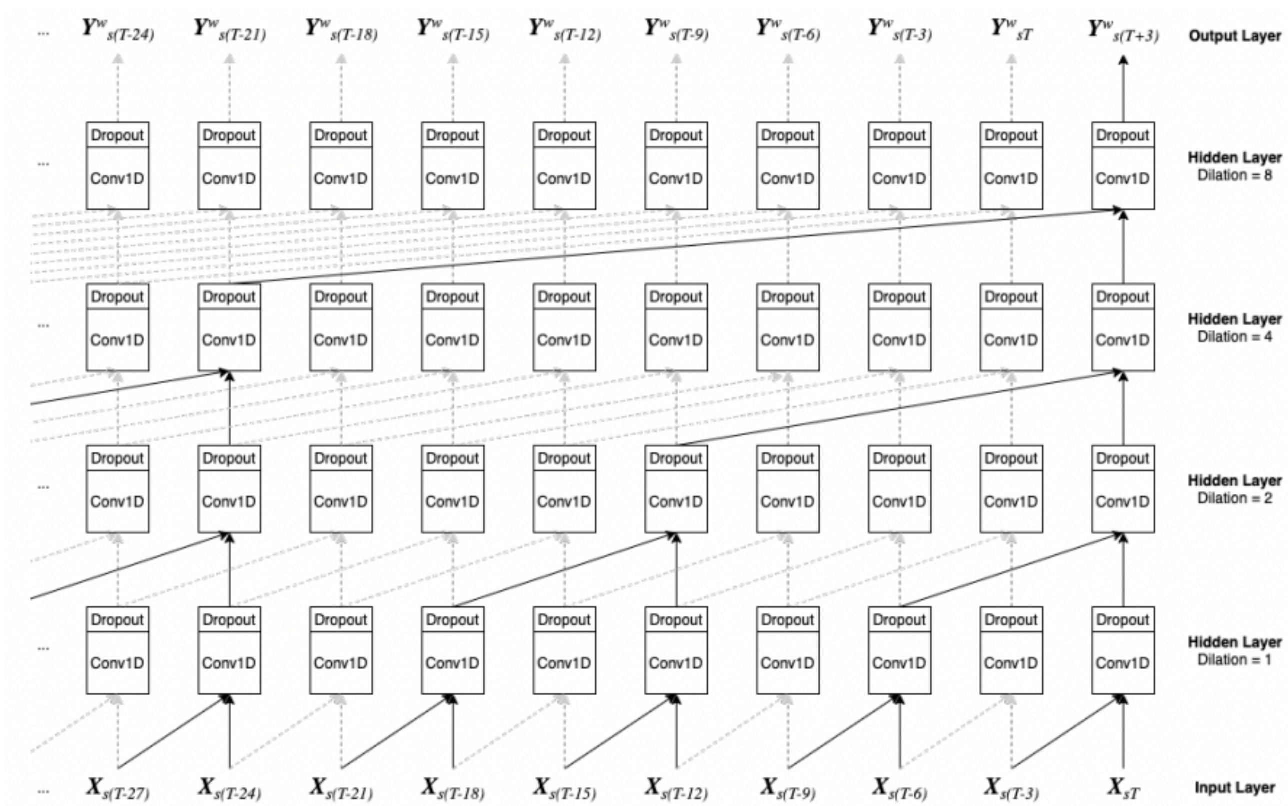

**Harrington+21**

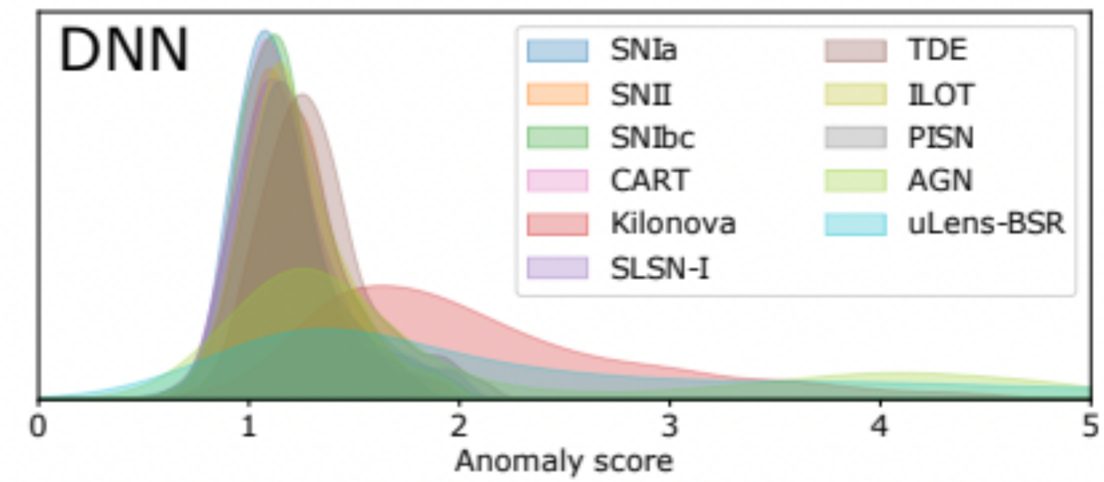


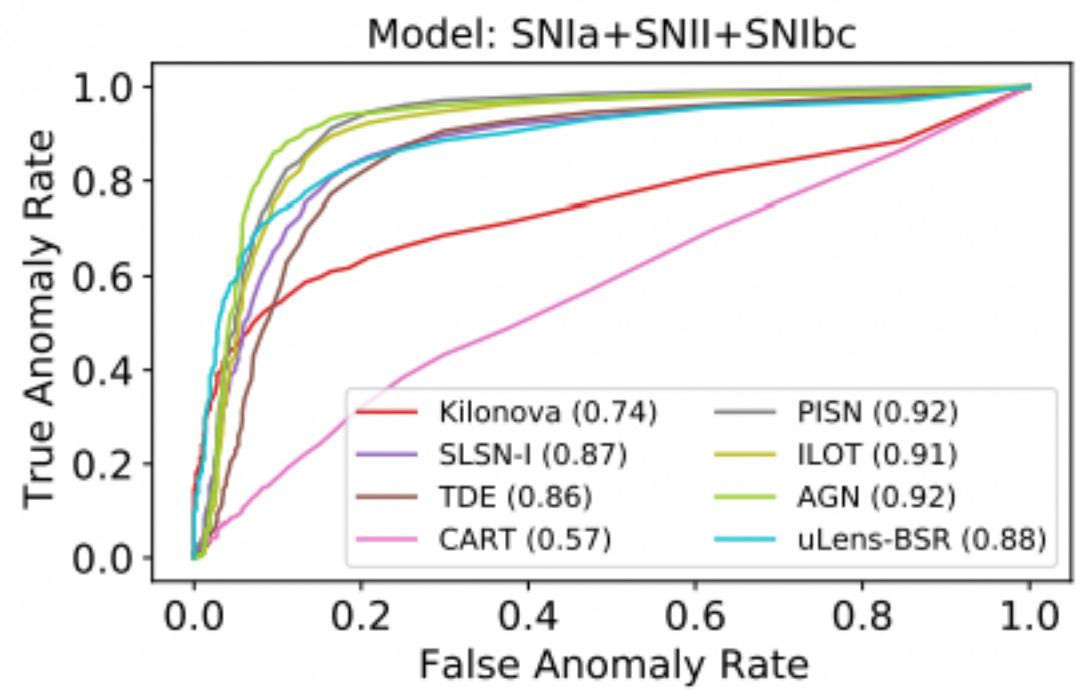

**Neural Networks are used to learn the non-linear mapping between cheap dark matter only simulations to expensive baryonic physics**

Rodriguez+19, Modi+18, Berger+18, He+18, Zhang+19, Troster+19, Zamudio-Fernandez+19, Perraudin+19, Charnock+19, List+19, Giusarma+19, Bernardini+19, Chardin+19, Mustafa+19, Ramanah+20, Tamosiunas+20, Feder+20, Moster+20, Thiele+20, Wadekar+20, Dai+20, Li+20, Lucie-Smith+20, Kasmanoff+20, Ni+21, Rouhiainen+21, Harrington+21, Horowitz+21, Horowitz+21, Bernardini+21, Schaurecker+21, Etezad-Razavi+21, Curtis+21

# 4. Deep Learning for assisted discovery



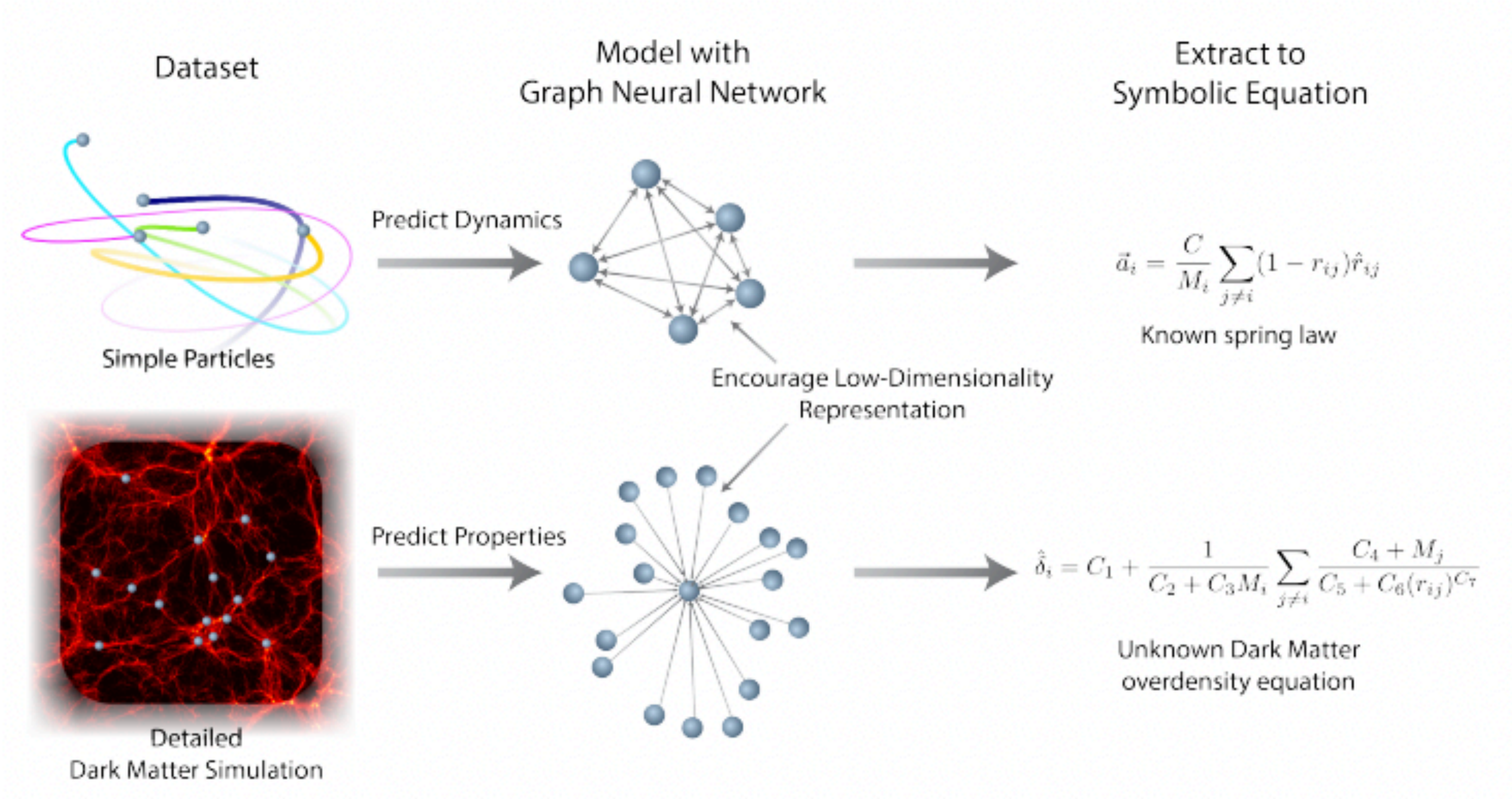




**Something that cannot be modelled is an outlier - Neural Network are used to represent known objects**

**Muthukrishna+21**

Teimoorinia+16, Ma+19, Pruzhinskaya+19, Cheng+20,21, Cranmer+20, Galvin+20, Margalef-Bentabol+20, Mesarcik+20, Portillo+20, Skoda+20, Villar+20, Boone+21, Chan+21, Hayat+21, Lochner+21, Malanchev+21, Muthukrishna+21, Sánchez-Sáez+21, Sarmiento21, Sravan+21, Stein+21ab, Storey-Fisher+21, Tanaka+21, Turner+21, Villar+21, Zanisi+21, Zhou+21, Spindler+21, Li+22

# 4. Deep Learning for assisted discovery

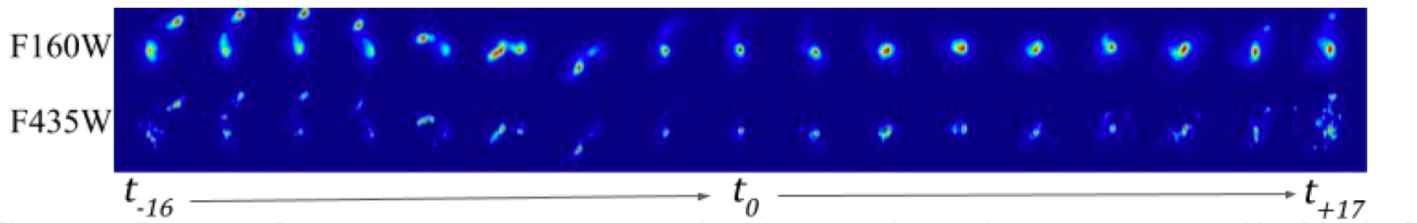



$$\vec{a}_i = \frac{C}{M_i}\sum_{j\neq i}(1 - r_{ij})\hat{r}_{ij}$$

$$\hat{\delta}_i = C_1 + \frac{1}{C_2 + C_3 M_i}\sum_{j\neq i}\frac{C_4 + M_j}{C_5 + C_6 (r_{ij})^{C_7}}$$

**Cranmer+20**

# 5. Deep Learning for hidden correlations

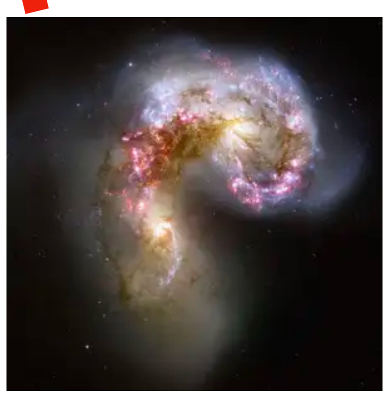



Merger of galaxies sequence
from cosmological simulations

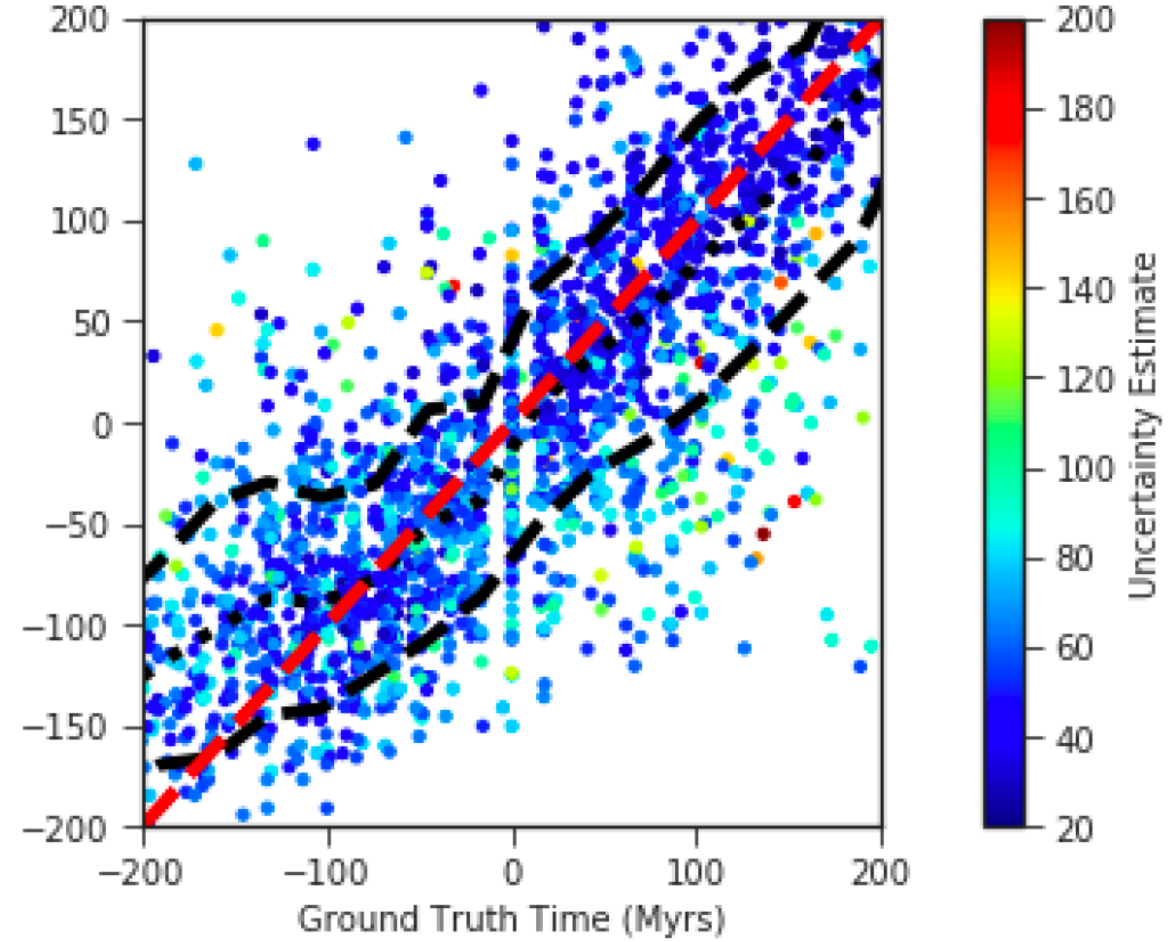
?

?

**Merger stage?**



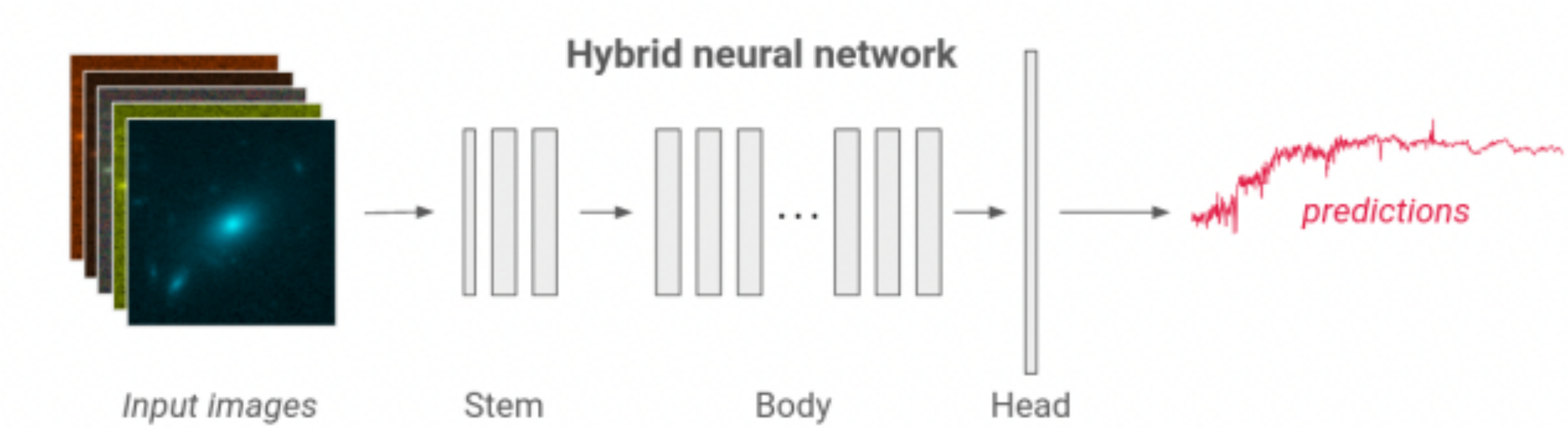
**Neural Networks to find relations between observables and physical processes**

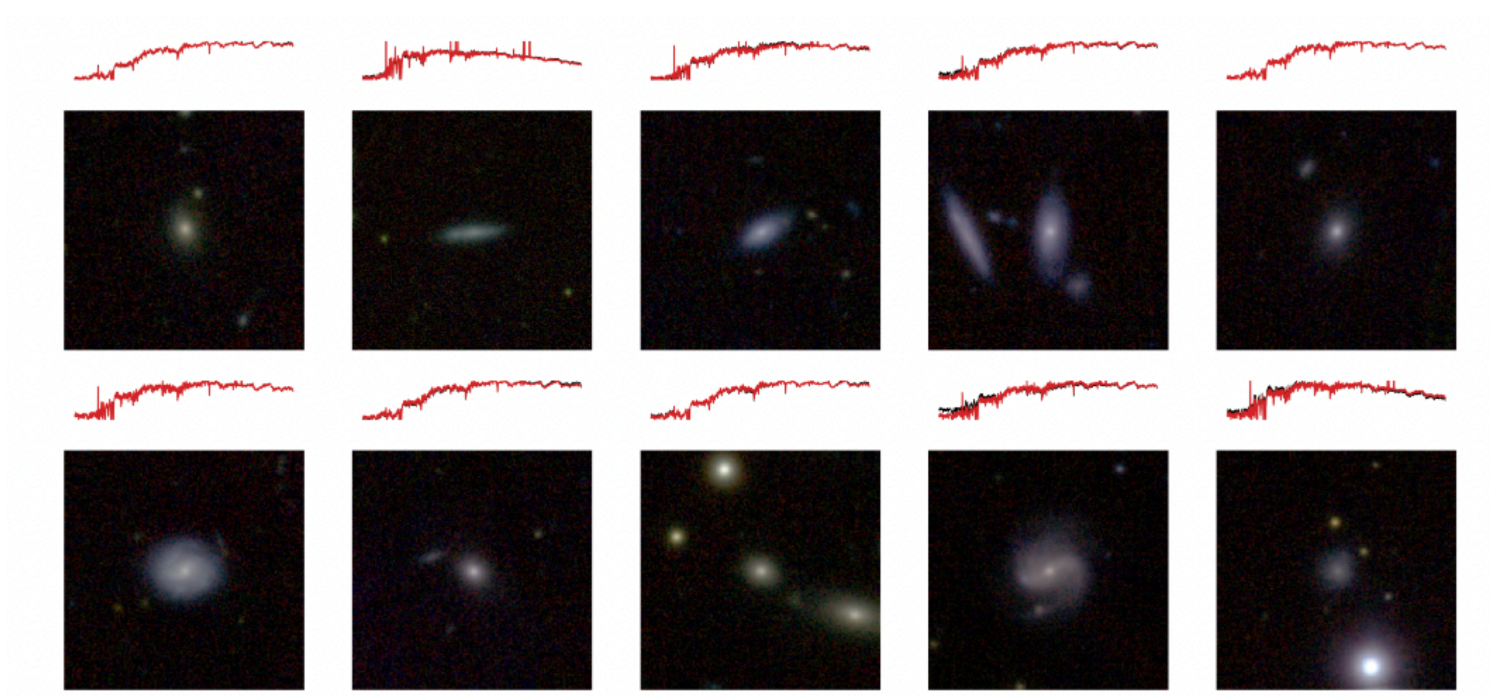



**Koppula+21**

Ntampaka+16, Huertas-Company+18, Schawinski+18, Armitage+19, Bottrell+19, Calderon+19, Diaz+19, Ho+19, Pearson+19, Cai+20, Ferreira+20, Hansen+20, Kodi Ramanah+20, Wu+20, Yan+20, Gupta+20, de Andres+21, Bickley+21, Ciprijanovic+21, Ho+21, Koppula+21, von Marttens+21, Shao+21, Villanueva-Domingo+21, Whithney+21, Holwerda+21

# 5. Deep Learning for hidden correlations



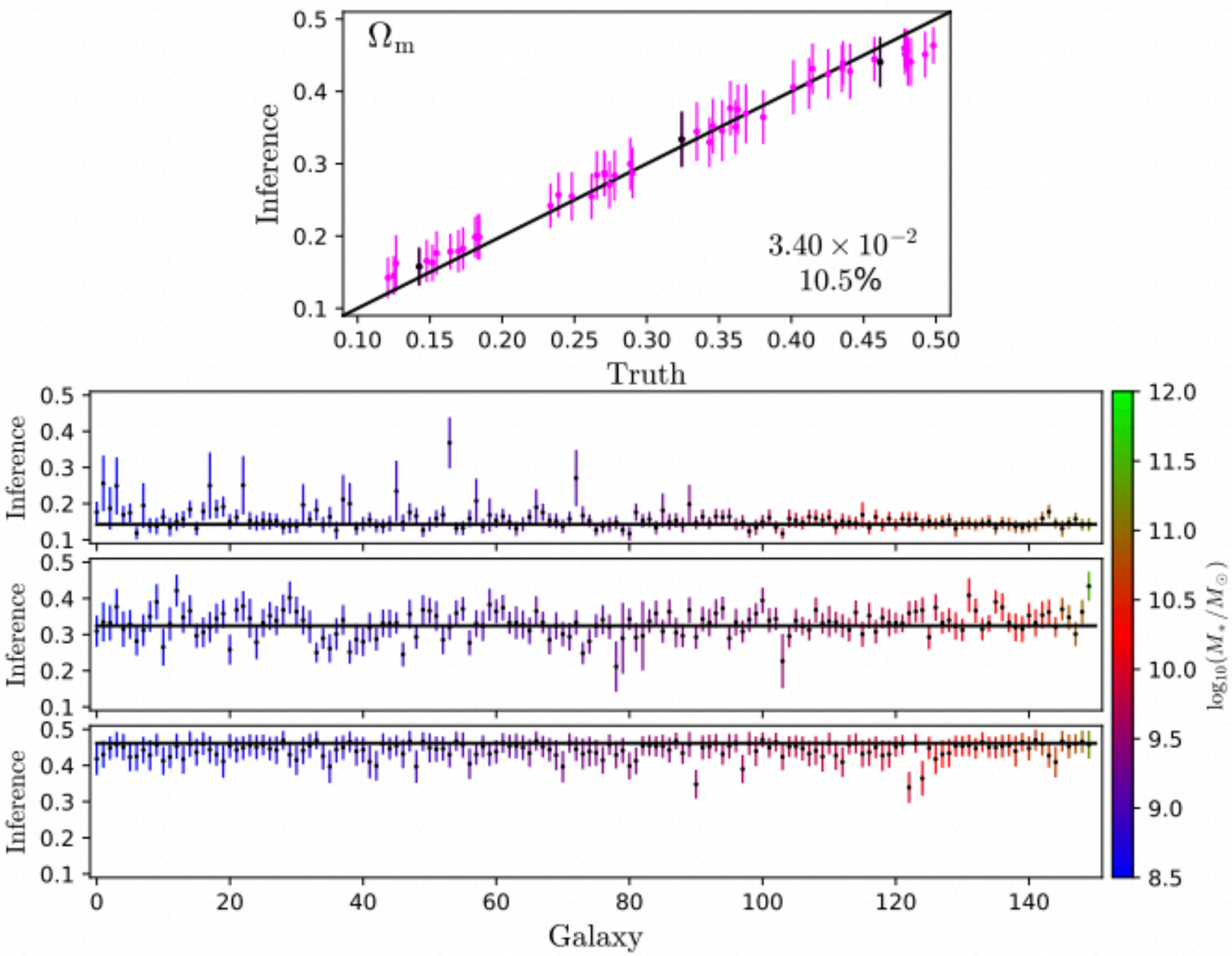



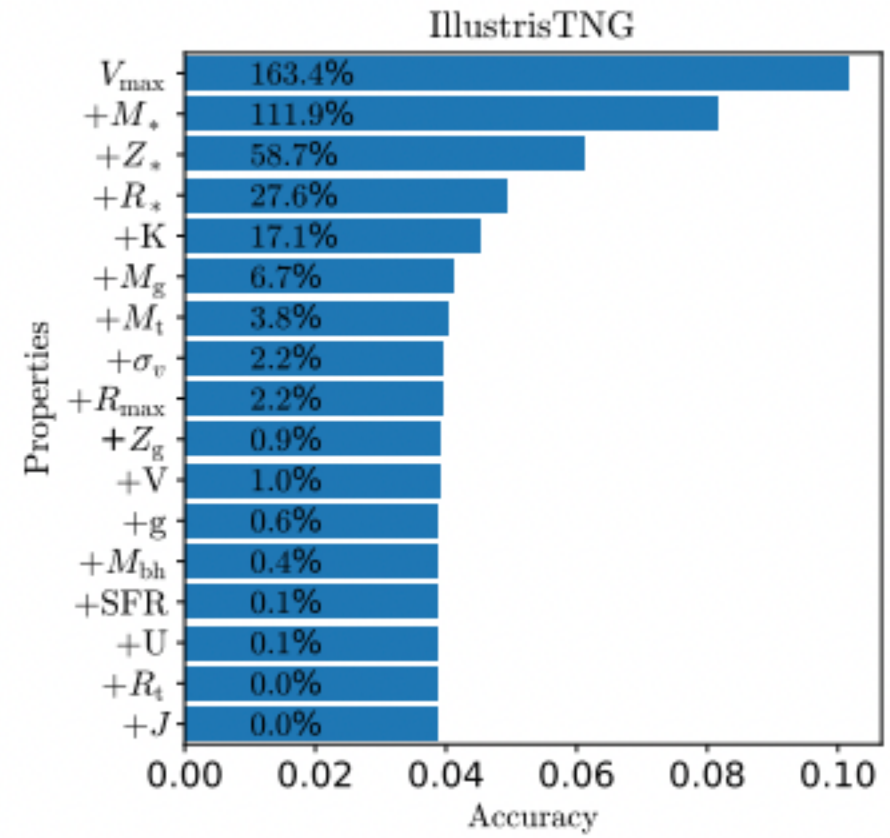
**Neural Networks to learn complex mapping between observables**

Wu+21

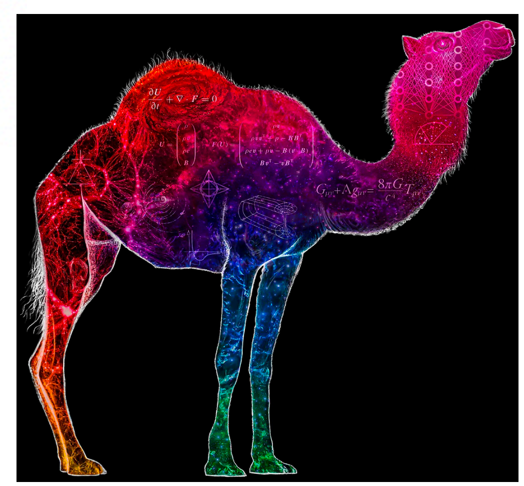
Ntampaka+16, Huertas-Company+18, Schawinski+18, Armitage+19, Bottrell+19, Calderon+19, Diaz+19, Ho+19, Pearson+19, Cai+20, Ferreira+20, Hansen+20, Kodi Ramanah+20, Wu+20, Yan+20, Gupta+20, de Andres+21, Bickley+21, Ciprijanovic+21, Ho+21, Koppula+21, von Marttens+21, Shao+21, Villanueva-Domingo+21, Whithney+21, Holwerda+21

# 6. Deep Learning to constrain cosmology




**Villaescusa-Navarro+22**



CAMELS

**Neural Networks to learn new relations between models and observables**


Ravanbakhsh+17, Brehmer+19, Ribli+19, Pan+19, Ntampaka+19, Alexander+20, Arjona+20, Coogan+20, Escamilla-Rivera+20, Hortua+20, Vama+20, Vernardos+20, Wang+20, Mao+20, Arico+20, Villaescusa_navarro+20, Singh+20, Park+21, Modi+21, Villaescusa-Navarro+21ab, Moriwaki+21, DeRose+21, Makinen+21, Villaescusa-Navaroo+22

# How deep learning techniques penetrate the community ?

| Application \ Model | | | CNNs | Enc. | Gene. | BNNs | RNNs | Trans. | GNNs |
|---|---|---|---|---|---|---|---|---|---|
| Data Processing | Classification | Morphology | ✓ | ✓ | | | | | |
| | | Strong Lenses | ✓* | ✓* | | | | | |
| | | Transients | | | | | ✓* | ✓* | |
| | Segmentation | | | ✓* | ✓* | | | | |
| Galaxy Properties | | Photoz | ✓ | | | ✓ | | | |
| | | Structure | ✓* | | | | | | |
| | | Stellar Populations | ✓* | | | | | | |
| | | Lensing | ✓* | | | ✓* | | | |
| Discovery | | Visualization | ✓ | ✓ | ✓ | | | | |
| | | Outliers | ✓ | ✓ | ✓ | | ✓ | | |
| | | Laws | | | | | | | ✓* |
| Un-observables | | Galaxy Evolution | ✓* | | | | | | |
| | | Dark Matter | ✓* | | | ✓* | | | ✓* |
| Emulation | | | ✓* | ✓* | ✓* | ✓* | | | |
| Cosmological Model | | | ✓* | | ✓* | ✓* | | | |

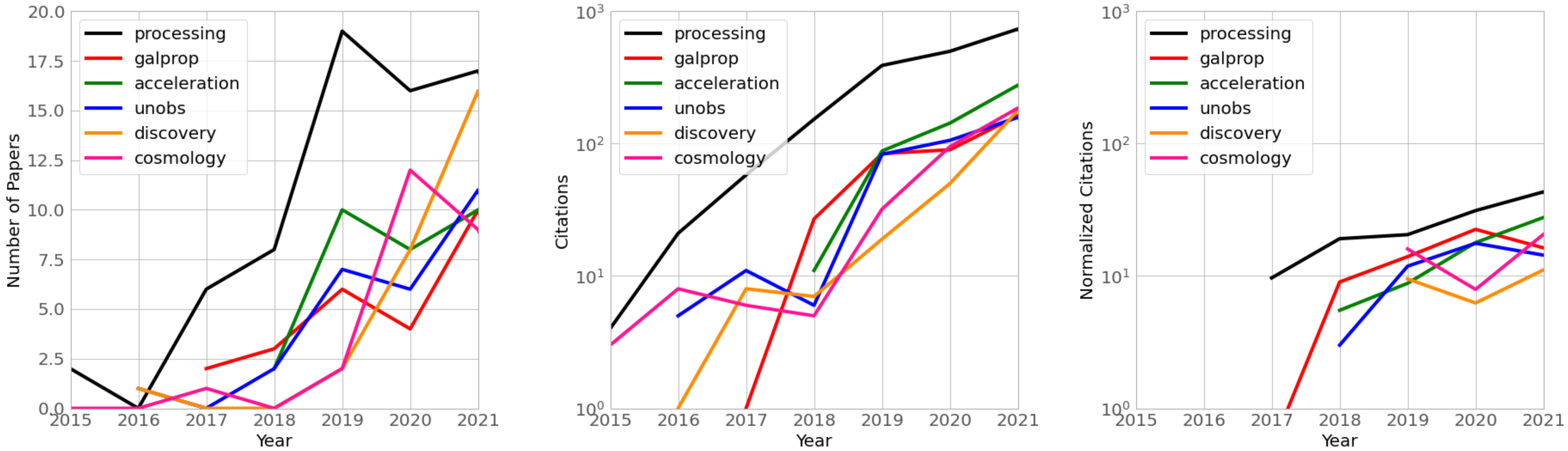
**Supervised** **Unsupervised** * **Training with simulations**

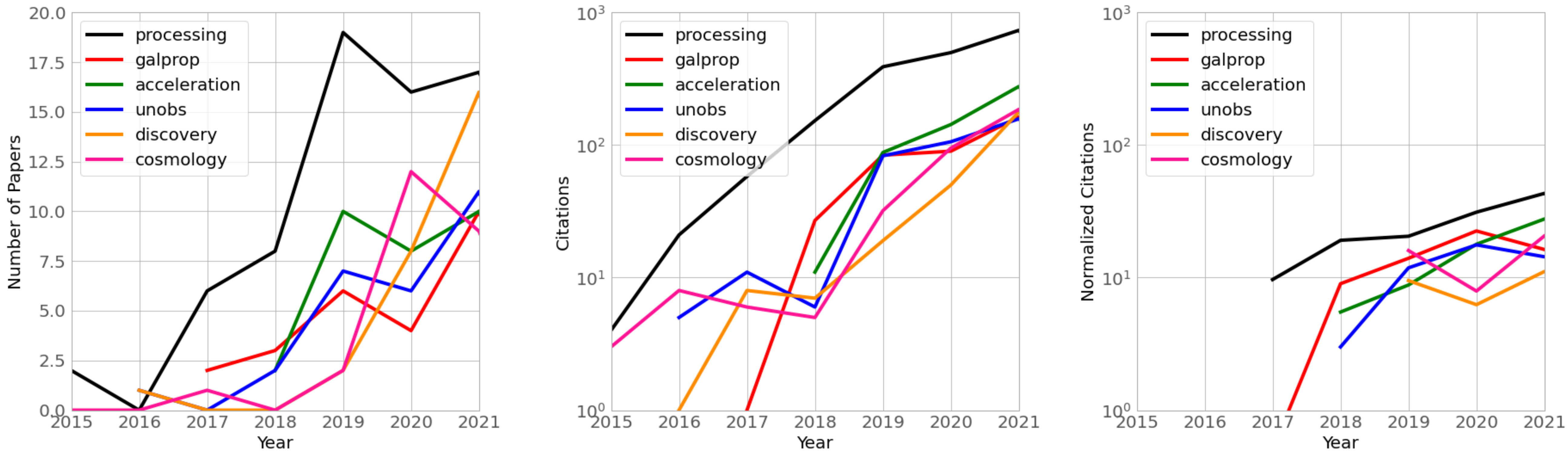
**The community uses in general state-of-the art ML techniques quickly after they become popular in the ML community: <u>Good transfer of knowledge</u>**

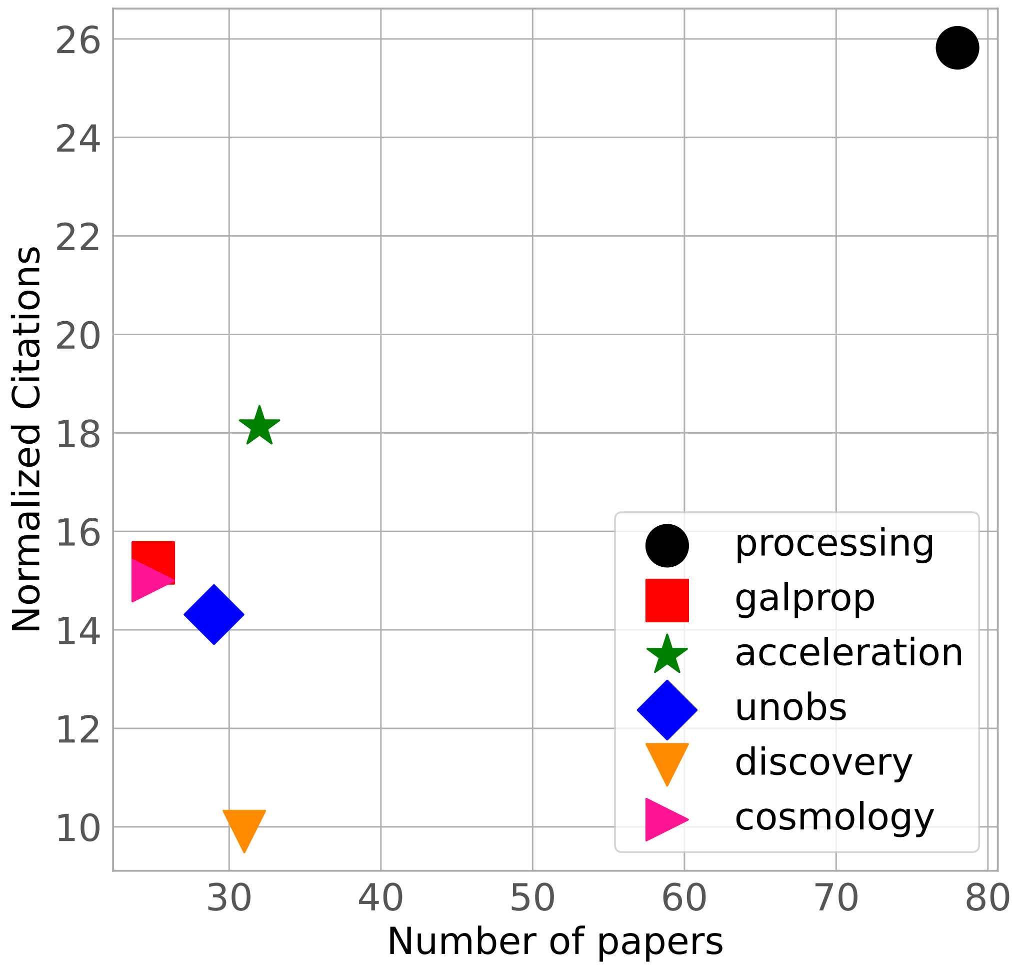
**However, little domain specific adaptation, in general**

**Training on simulations is very common in many applications**

# Measuring the impact of deep learning publications for galaxy surveys

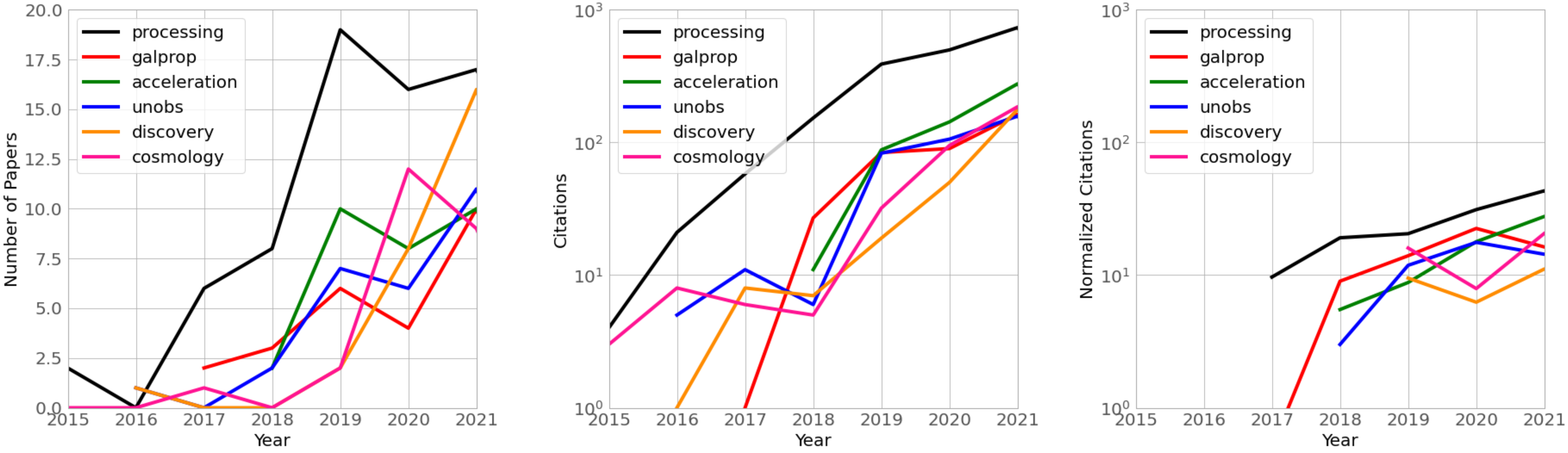


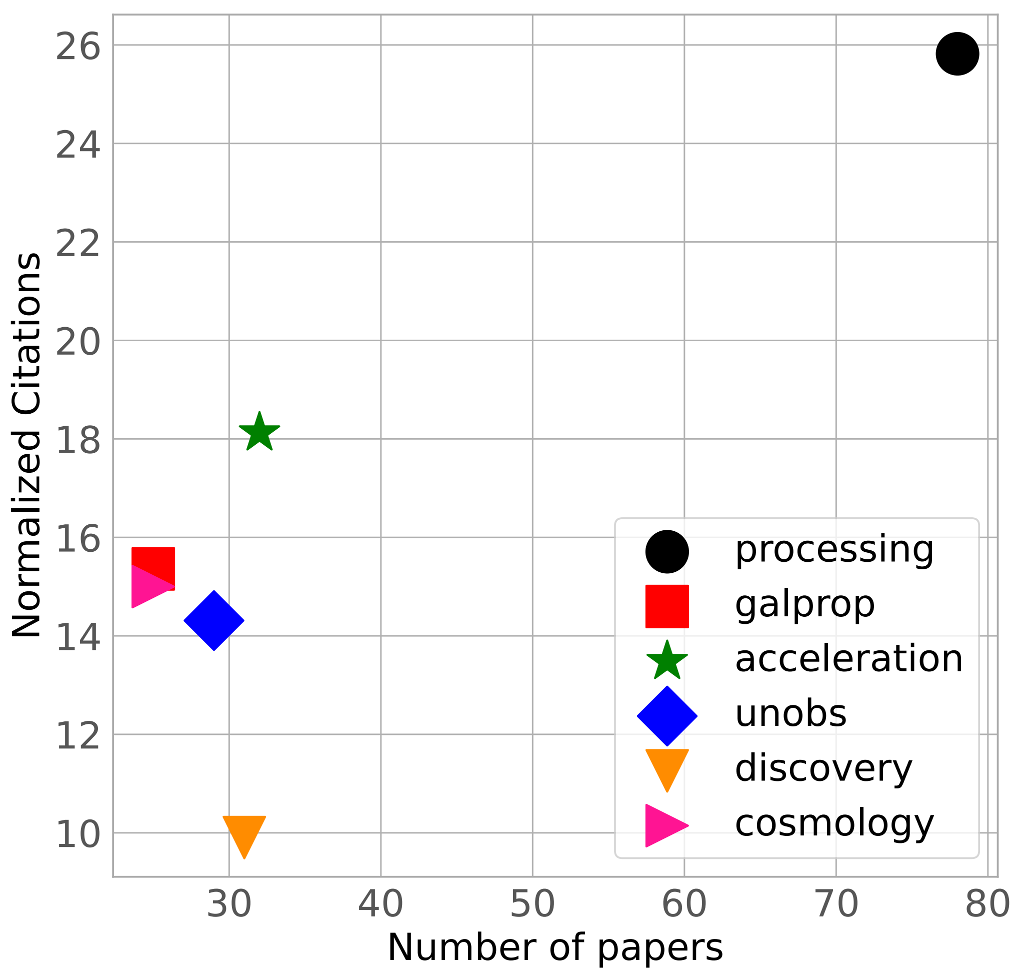



1. Most of the impact of deep learning is still on basic data processing tasks (e.g. mostly incremental science)

2. Despite an increasing trend, and popular interest, potentially novel applications still have a low impact in terms of citations (e.g. discovery)




1. Most of the impact of deep learning is still on basic data processing tasks (e.g. mostly incremental science)

2. Despite an increasing trend, and popular interest, potentially novel applications still have a low impact in terms of citations (e.g. discovery)

**WHY?**

# Possible reasons for limited impact

1. The majority of new datasets have not arrived yet (e.g. Euclid, LSST, Roman, JWST)

2. Too young, still. It is just a matter time. Proof-of-concept stage.

3. Some key issues remain unsolved. Lack of trust.

## Current Challenges (and possible solutions)

| | |
|---|---|
| **Challenge 1 Small labeled datasets** | |
| Solution 1.A Transfer Learning | Domínguez Sánchez et al. (2019) Samudre et al. (2022) Lukic et al. (2019) |
| Solution 1.B Simulated dataset | Jacobs et al. (2017) Vega-Ferrero et al. (2021) |
| Solution 1.C Self-supervised learning | Hayat et al. (2021) |
| **Challenge 2 Uncertainty** | |
| Solution 2.A Bayesian approximations | Walmsley et al. (2020) Perreault Levasseur et al. (2017) |
| Solution 2.B Density Estimators | Kodi Ramanah et al. (2020) |
| **Challenge 3 Interpretability** | |
| Solution 3.A Saliency maps and similar | Huertas-Company et al. (2018) |
| Solution 3.B Symbolic regression | Cranmer et al. (2020) |
| **Challenge 4 Domain Shift** | |
| Solution 4.A Transfer Learning | Tuccillo et al. (2018) |
| Solution 4.A Domain Adaptation | Ćiprijanović et al. (2021) |
| **Challenge 5 Robustness** | |
| Solution 5.A Out of distribution detection | Lee & Shin (2022) |

# Summary: Revolution or Incremental Science?

1. Deep Learning has rapidly penetrated the field of extragalactic astronomy (surveys)

2. State of the art ML techniques are rapidly tested

3. The most impactful applications are low level "out-of-the-box" applications (incremental science)

4. Potentially novel applications have emerged but remain at the proof-of-concept stage with moderate impact