機械学習を用いた近年の宇宙物理研究

森脇可奈（東京大学UTAP）

Contents:
- ML in astrophysics
- ML methods + our recent results
- General issues

2021/11/17-19 第10回 観測的宇宙論ワークショップ

Machine learning in astrophysics

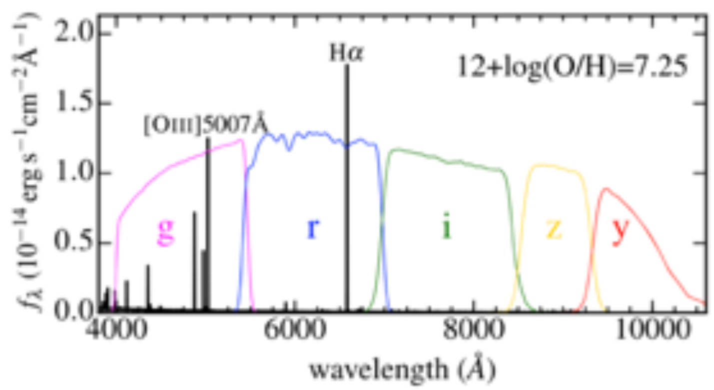
cf. review by Fluke & Jacobs 2020

- **Machine learning**: automated processes that learn by examples (training data)

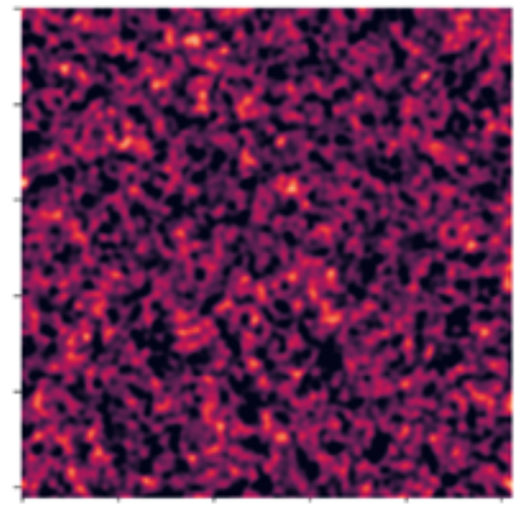
Images (Dieleman+15)

Spiral? Merger? (classification)

Photometry (Kojima+20)



Redshift, metallicity, etc. (regression)



Weak lensing map (Shirasaki+19)

Cosmological parameter (regression)

De-noising (reconstruction)

Random noise

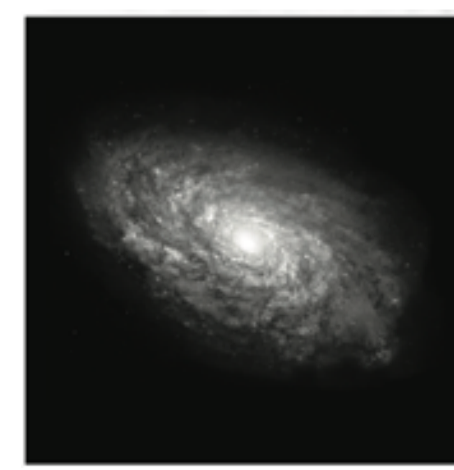


Image of galaxy (Ravanbakhsh+16)

(Generation)

Machine learning in astrophysics

We review the current state of data mining and machine learning in astronomy. *Data Mining* can have a somewhat mixed connotation from the point of view of a researcher in this field. If used correctly, it can be a powerful approach, holding the potential to fully exploit the exponentially increasing amount of available data, promising great scientific advance. However, if misused, it can be little more than the black box application of complex computing algorithms that may give little physical insight, and provide questionable results. Here, we give an overview of the entire data mining process, from data collection through to the interpretation of results. We cover common machine learning algorithms, such as artificial neural networks and support vector machines, applications from a broad range of astronomy, emphasizing those in which data mining techniques directly contributed to improving science, and important current and future directions, including probability density functions, parallel algorithms, Peta-Scale computing, and the time domain. We conclude that, so long as one carefully selects an appropriate algorithm and is guided by the astronomical problem at hand, data mining can be very much the powerful tool, and not the questionable black box.

Ball & Brunner (2010)

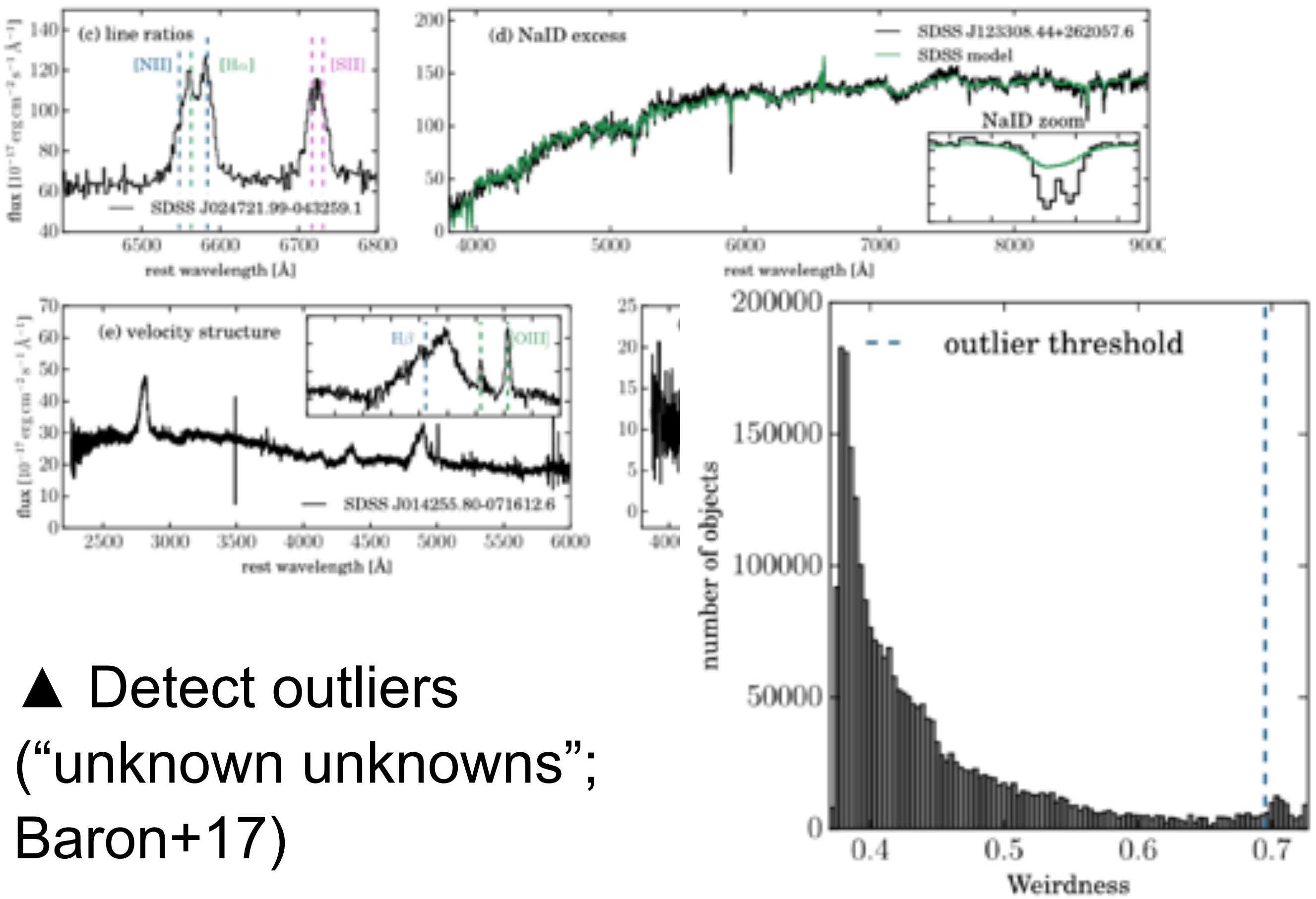
Abstract: In recent years, machine learning (ML) methods have remarkably improved how cosmologists can interpret data. The next decade will bring new opportunities for data-driven cosmological discovery, but will also present new challenges for adopting ML methodologies and understanding the results. ML could transform our field, but this transformation will require the astronomy community to both foster and promote interdisciplinary research endeavors.

Ntampaka et al. (2020)

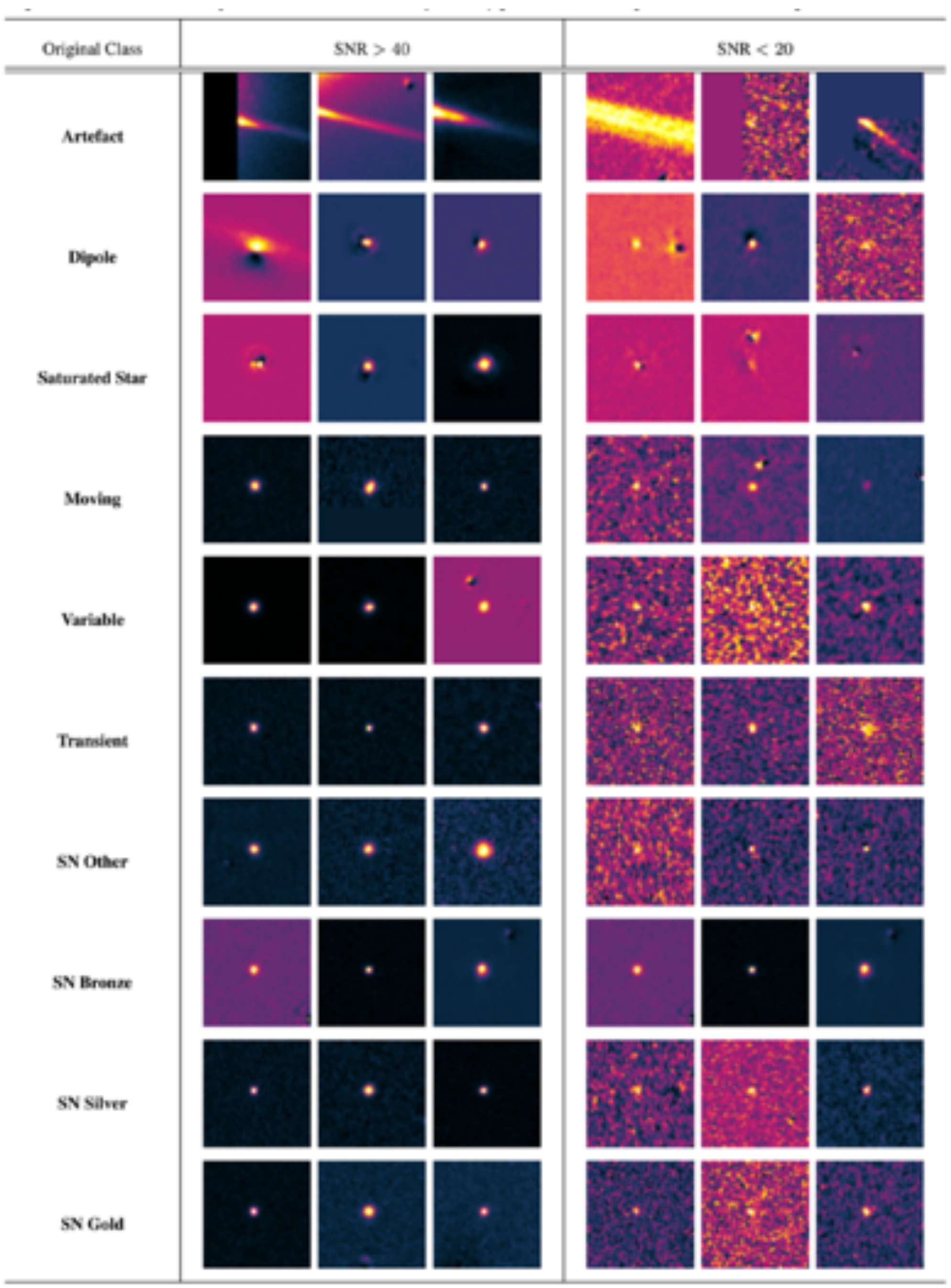
When is ML useful?

- **High-speed processing of large amounts of observational/simulation data** — Classification, search for rare objects, analysis of e.g., LSST, SKA, etc. in the future, generate realistic images of galaxies, emulator, etc.

Talks by Nishimichi-san, Tanaka-san



▲ Detect outliers (“unknown unknowns”; Baron+17)

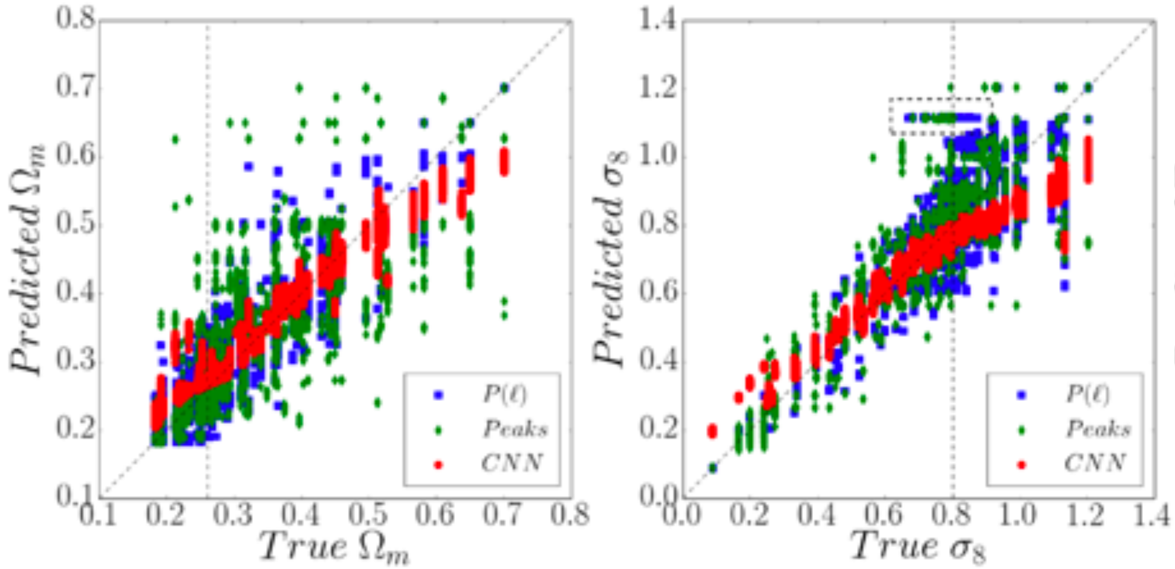


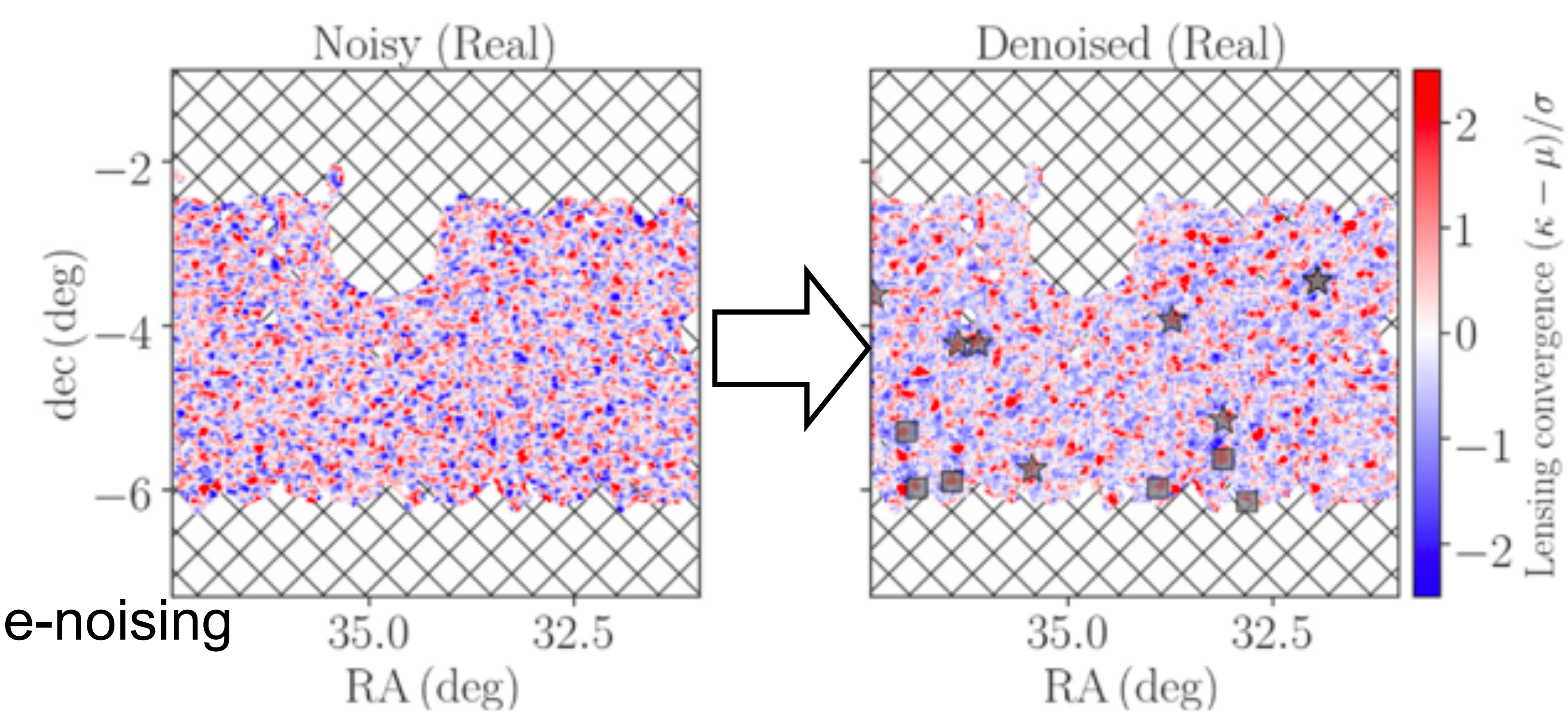
▲ Classification of SDSS transients (duBuisson+15)

When is ML useful?

- **High-speed processing of large amounts of observational/simulation data** — Classification, search for rare objects, analysis of e.g., LSST, SKA, etc. in the future, generate realistic images of galaxies, emulator, etc.
- **Capture complex structures/relationships** — Non-Gaussianity in WL maps, halo-galaxy relation, etc.

▼ Parameter estimation from noiseless WL maps (Gupta+18)



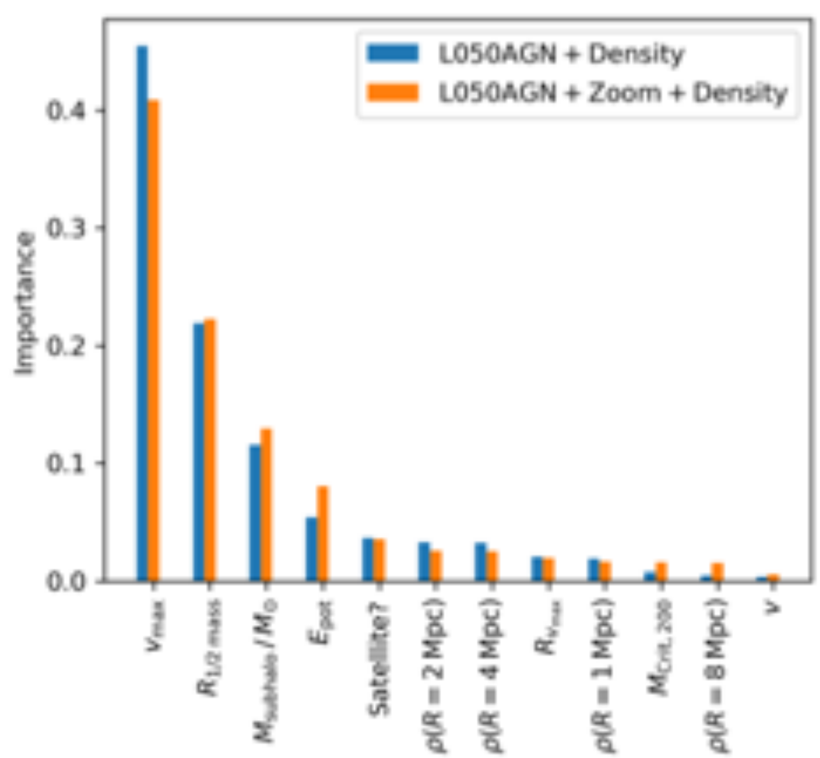


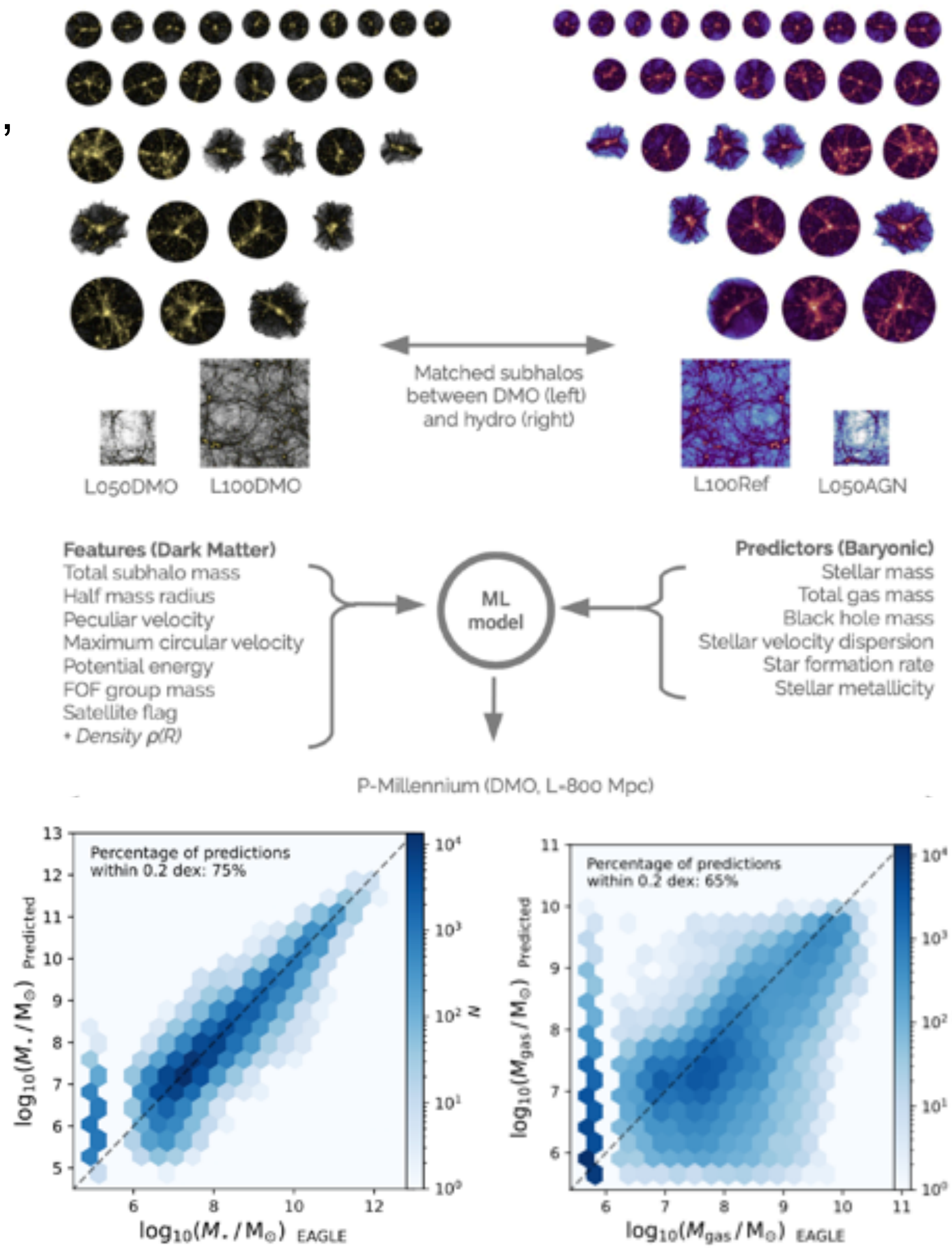
▶ Subaru HSC WL maps de-noising with cGAN (Shirasaki+21)

When is ML useful?

- **High-speed processing of large amounts of observational/simulation data** — Classification, search for rare objects, analysis of e.g., LSST, SKA, etc. in the future, generate realistic images of galaxies, emulator, etc.
- **Capture complex structures/relationships** — Non-Gaussianity in WL maps, halo-galaxy relation, etc.
- **(New) insights**

Importance of parameters



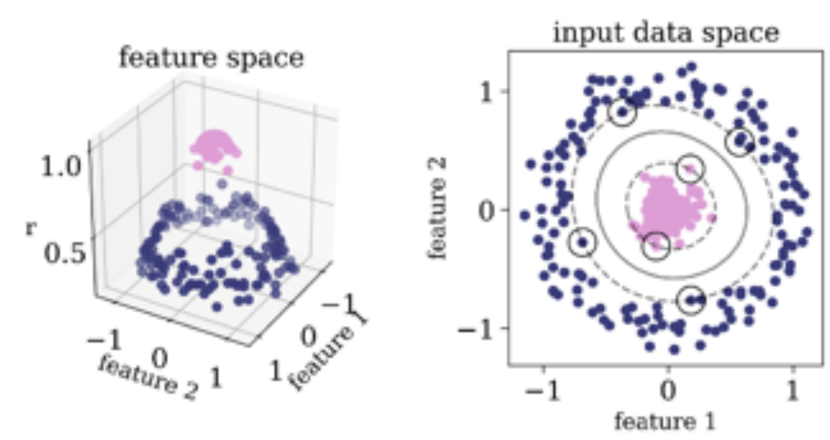


Mapping baryons onto DM haloes (Lovell+21)

Machine learning methods

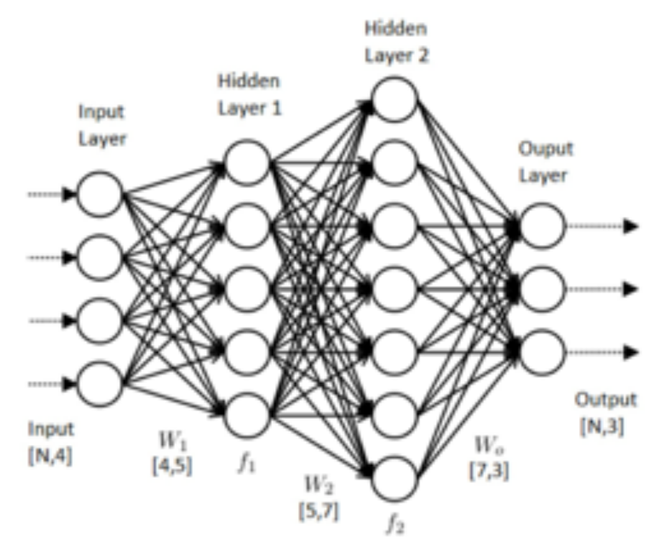
- Principal component analysis (PCA; ~1980s-)
- Artificial neural network (ANN; ~1990s-)
- Decision tree (DT; ~1990s-)
- Support vector machine (SVM; ~2000s-)

Support vector machine



Decision tree

Artificial neural network

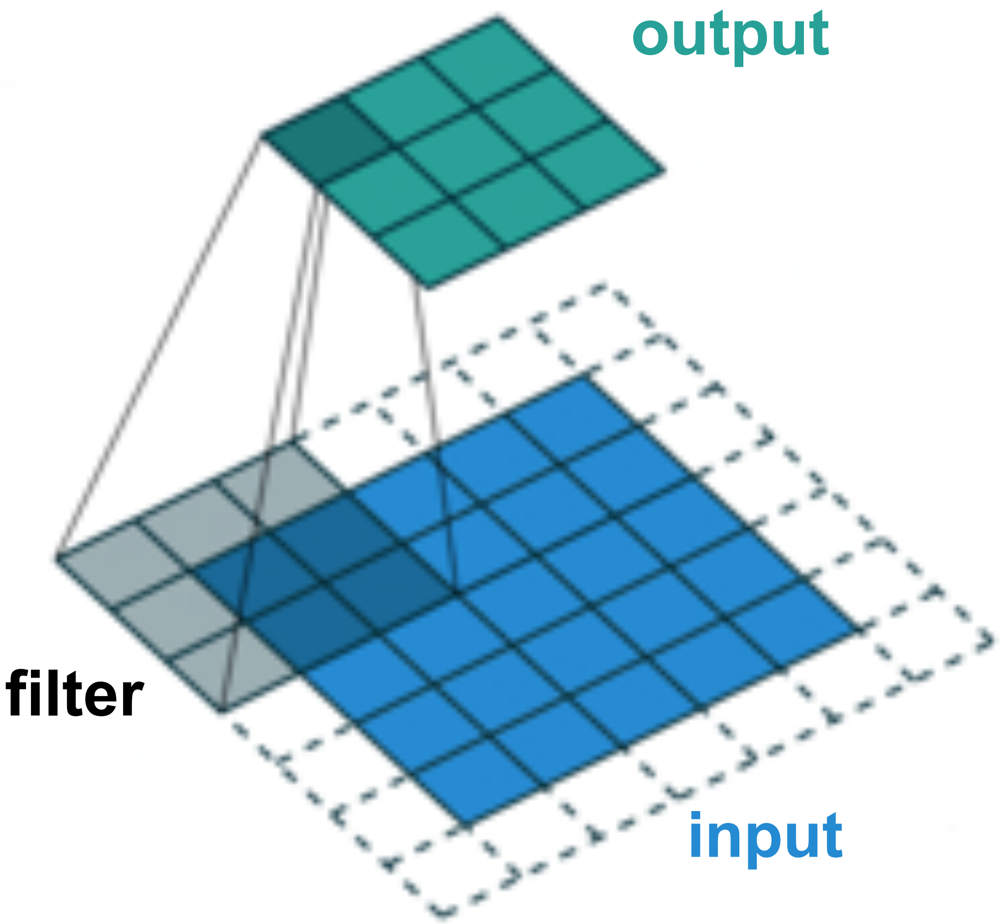


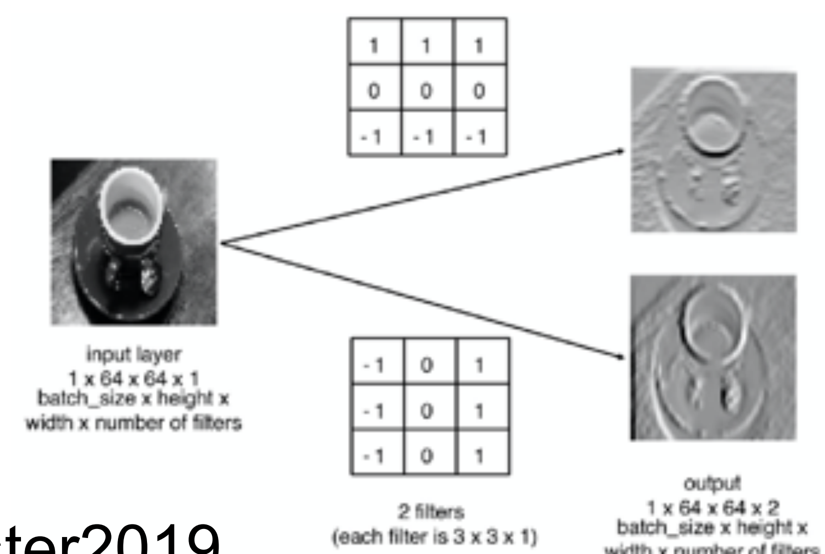
After appearance of GPU…

- Convolutional neural network (CNN)
- Recurrent neural network (RNN)
- Graph neural network (GNN)
- Transformer

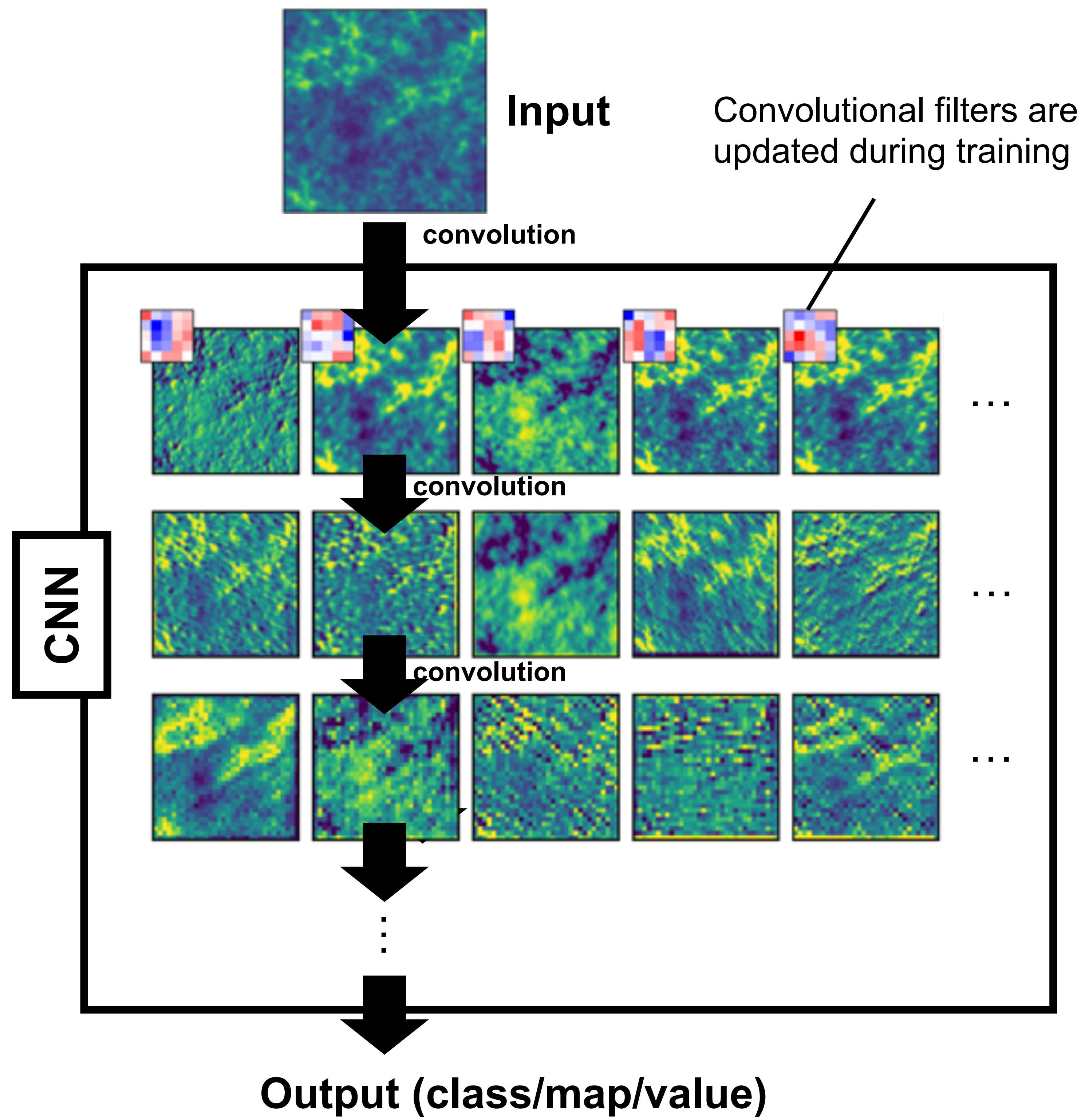
Figures from Baron (2019)

Convolutional Neural Network (CNN)





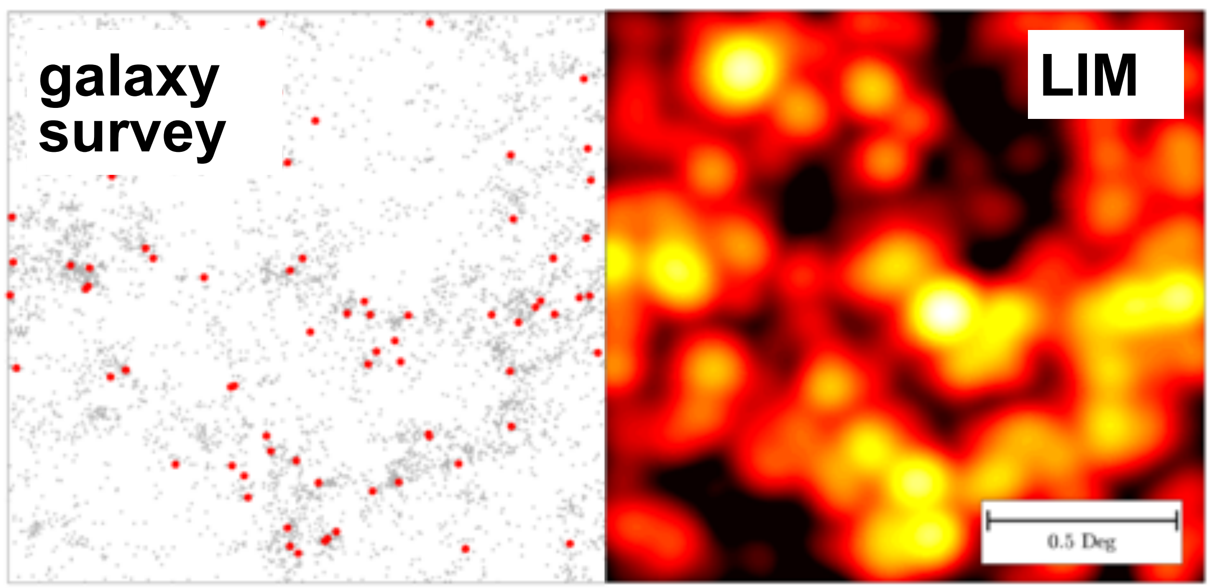
Foster2019



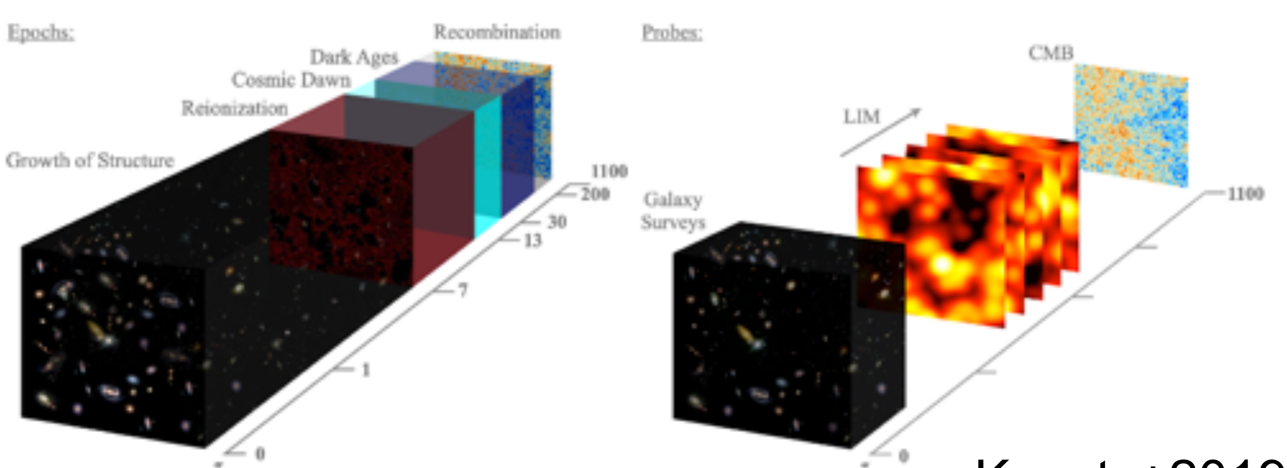
Our recent work: application of CNNs to LIM data

KM & Yoshida (2021)

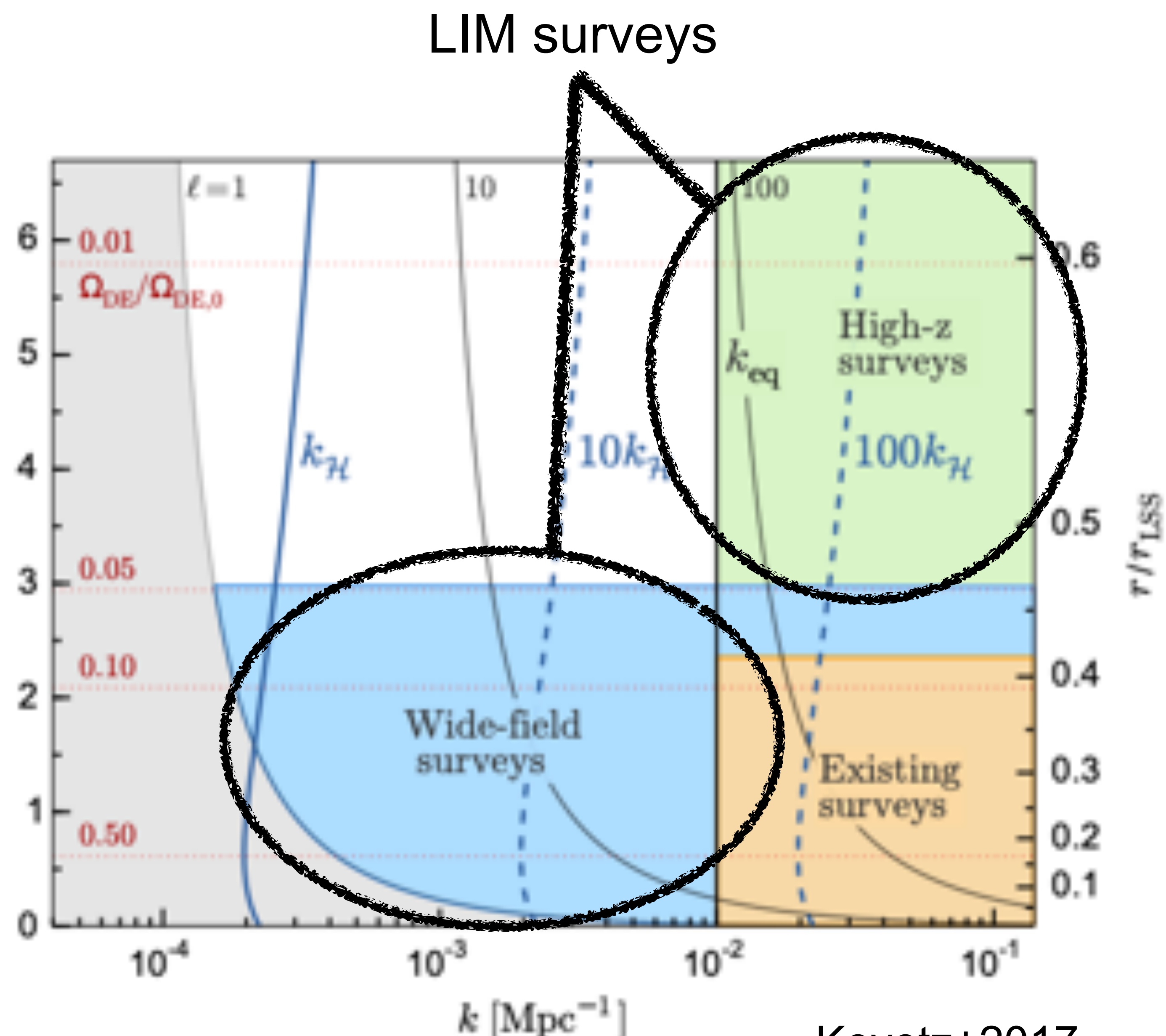
Line intensity mapping (LIM) observations provide large-scale 3D distributions of line intensities



Breysse et al. (2016)



Kovetz+2019



Kovetz+2017

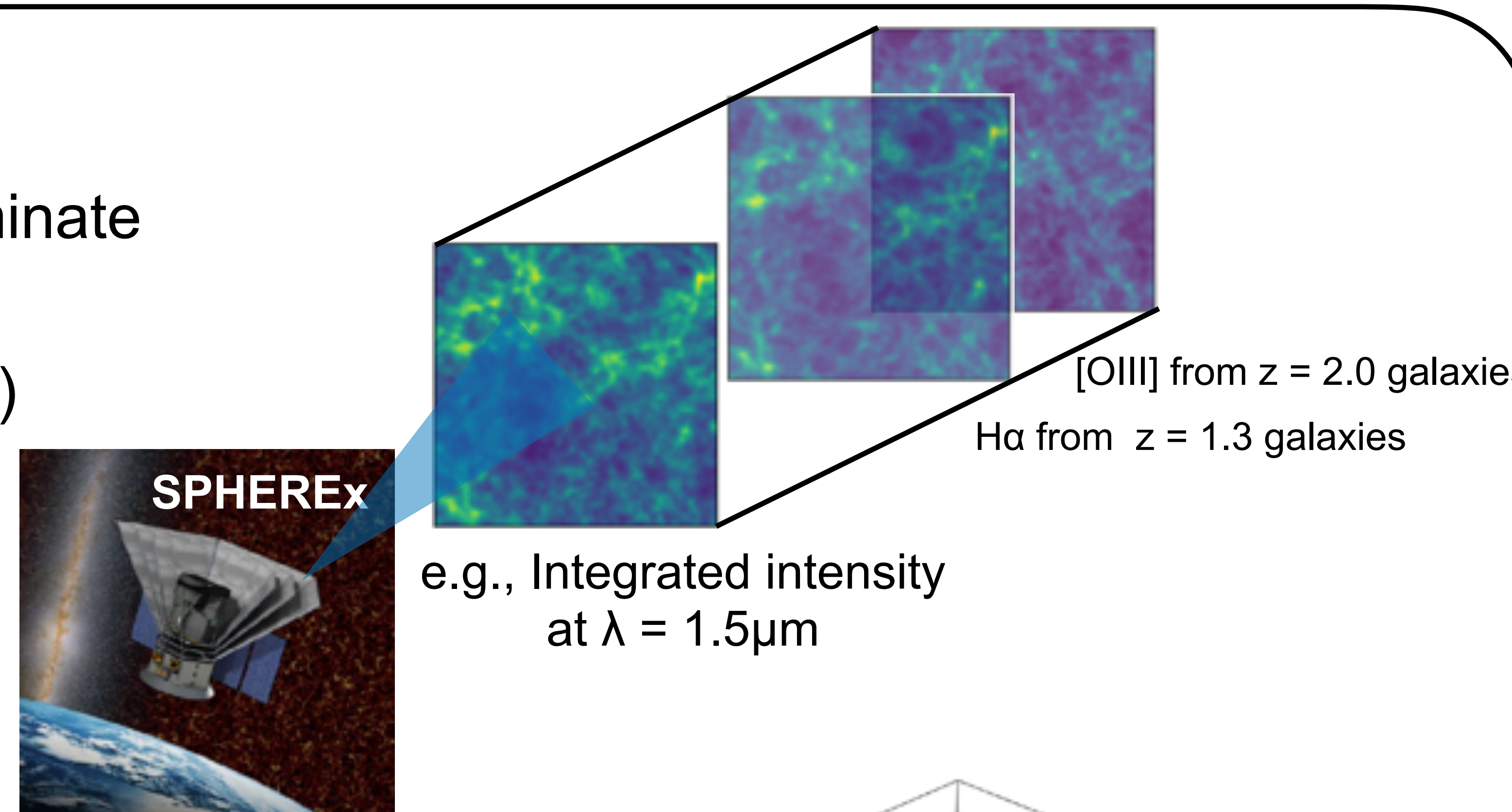
De-confusing interloper lines with CNNs

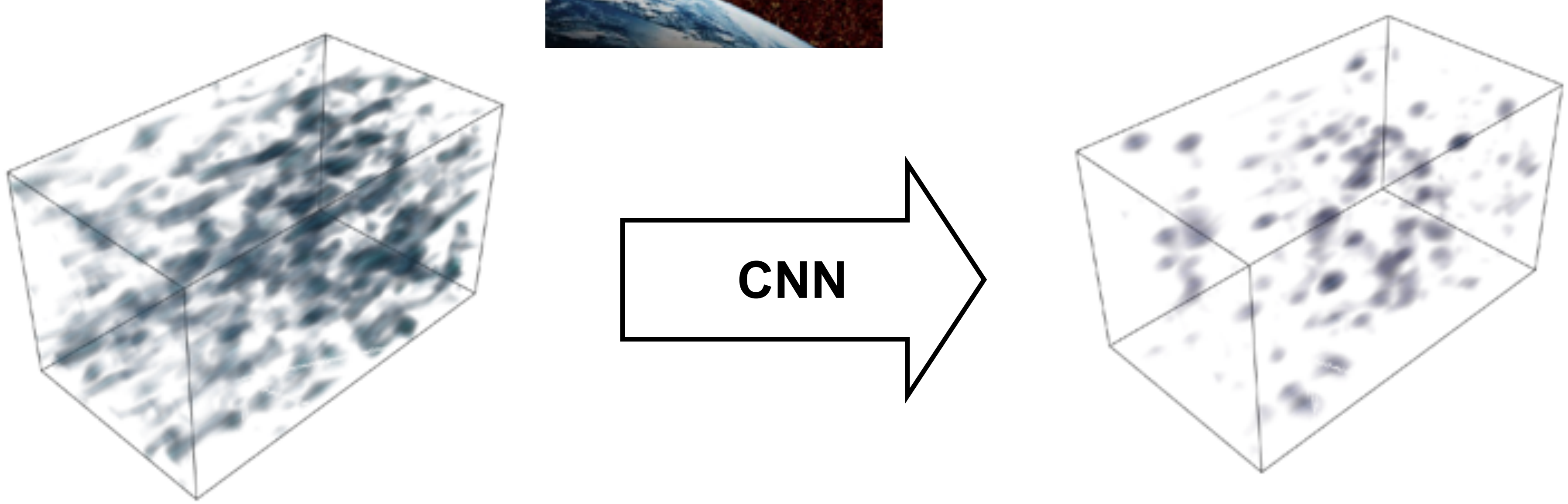
KM & Yoshida (2021)

- **Line confusion problem**

Multiple emission lines contaminate each other when they satisfy

$\lambda_{obs} = \lambda_1 (1 + z_1) = \lambda_2 (1 + z_2)$



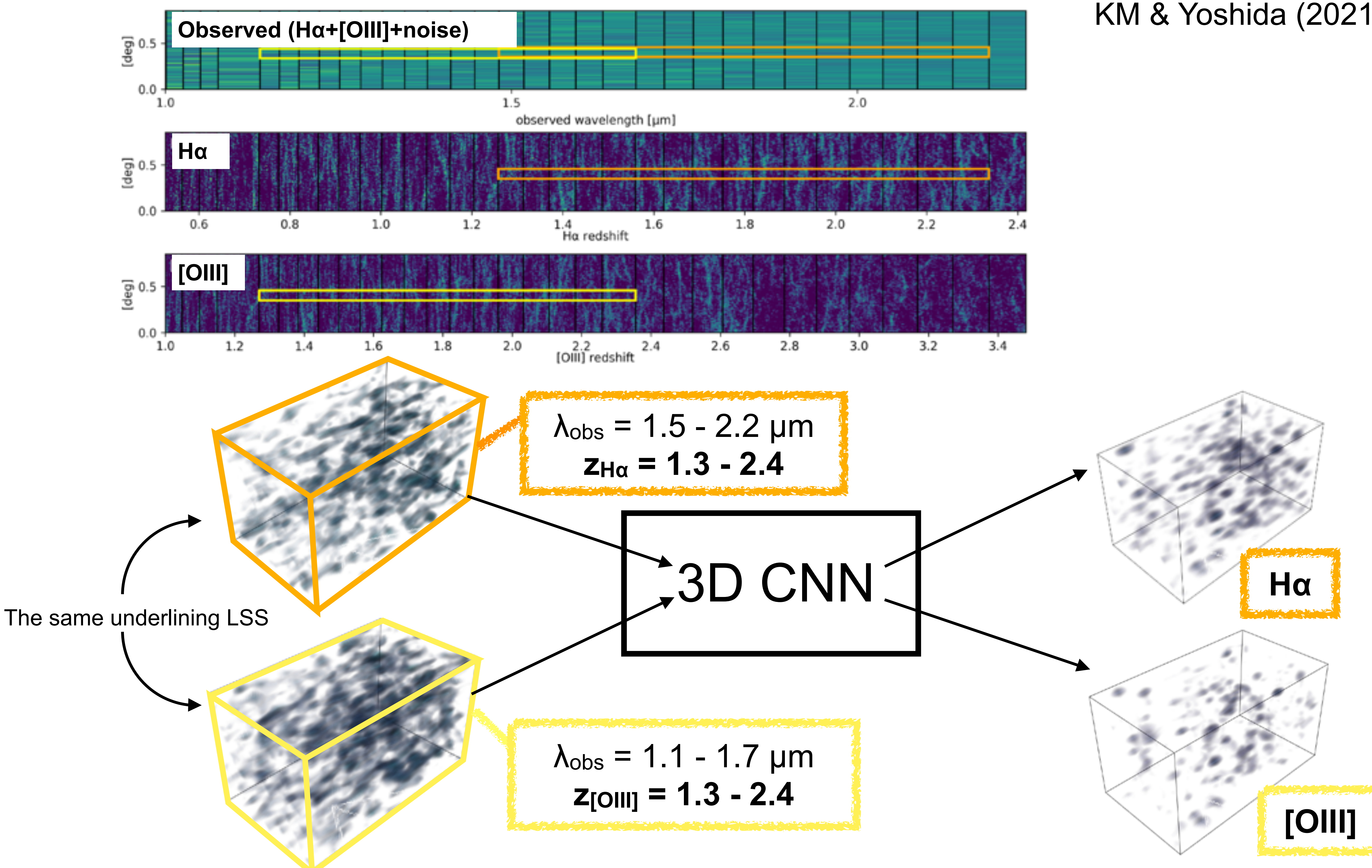


Observed data (3D data cube)

[OIII] distribution at z = 2.0

De-confusing interloper lines with CNNs

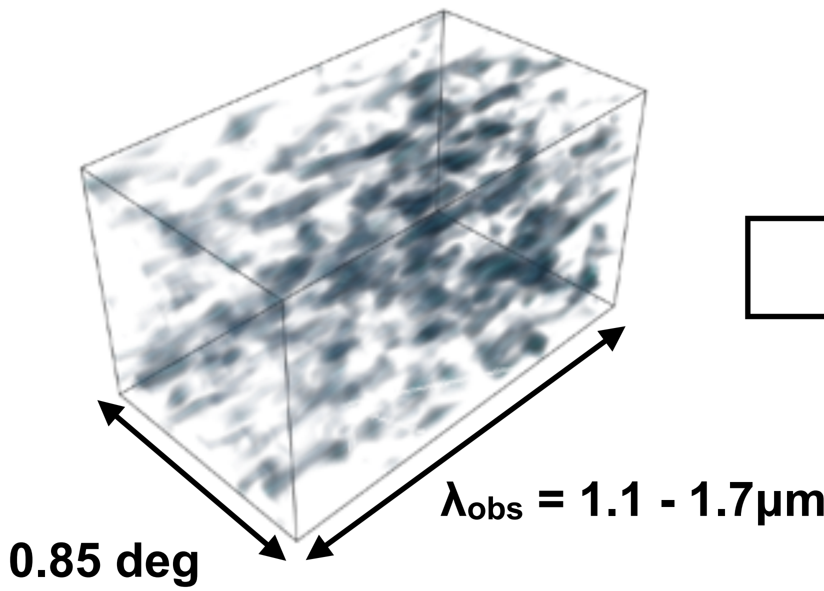
KM & Yoshida (2021)



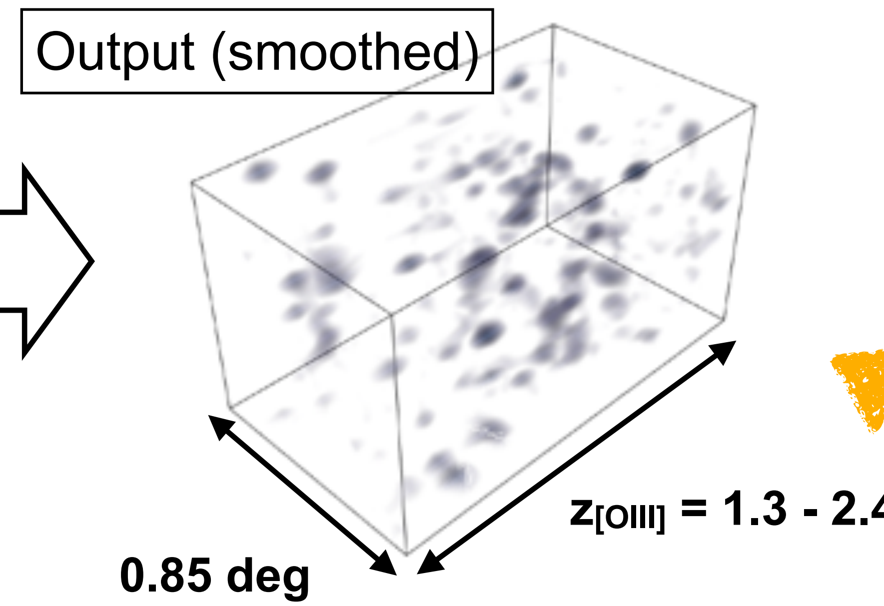
Reconstruction result

KM & Yoshida (2021)

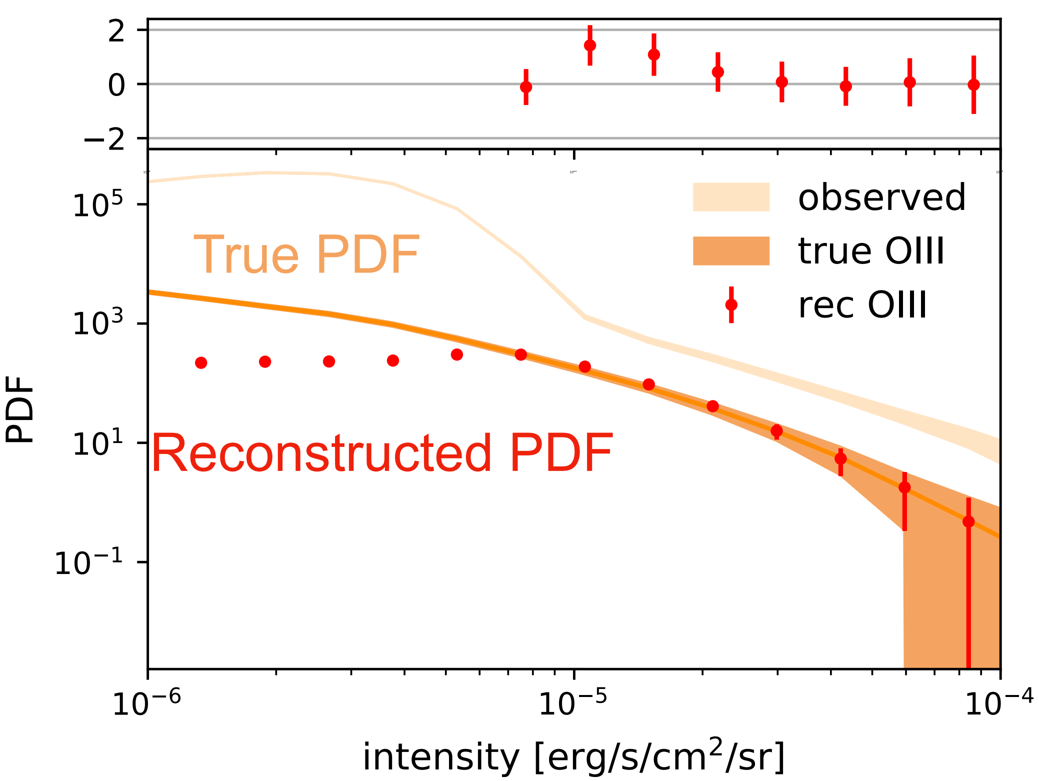
Observed (Hα+[OIII]+noise)



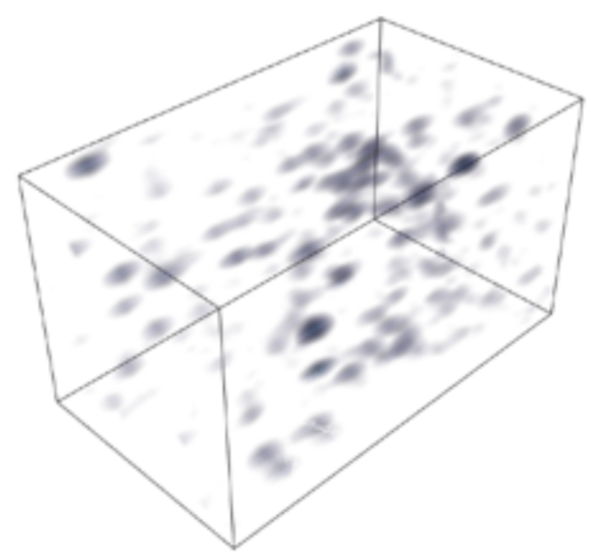
Reconstruct [OIII]



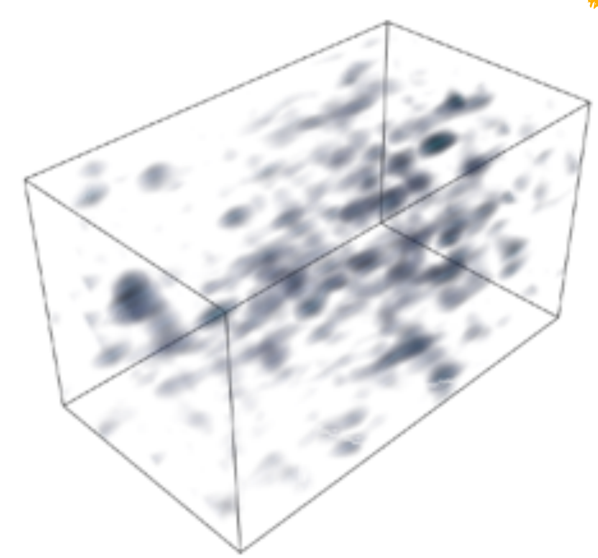
Foreground interlopers are properly removed



True [OIII]

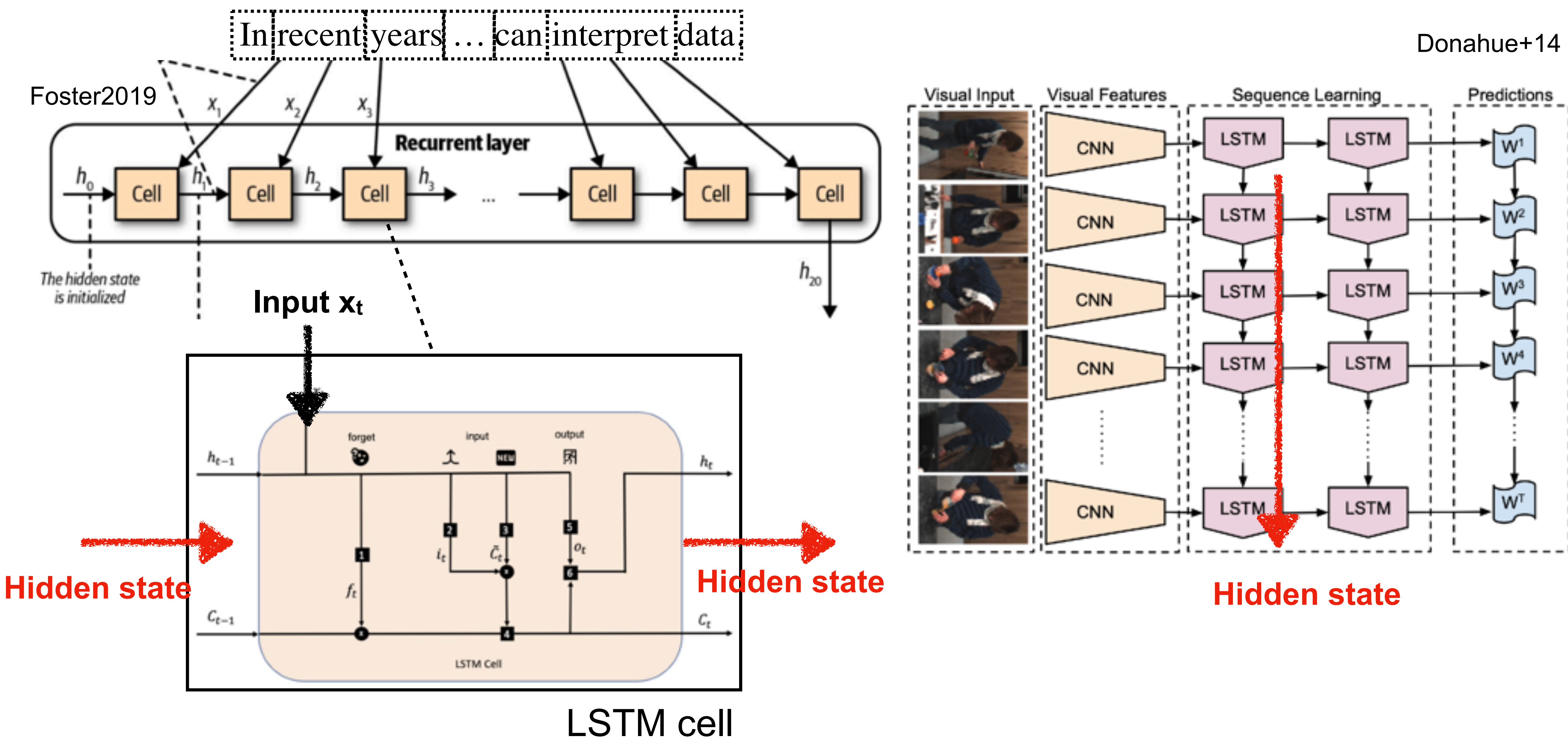


cf) True Hα+[OIII]



Recurrent Neural Network (RNN)

- Process sequential data (sentences, videos, etc.)
- Hidden state is propagated to downstream (e.g., long short-time memory, LSTM)

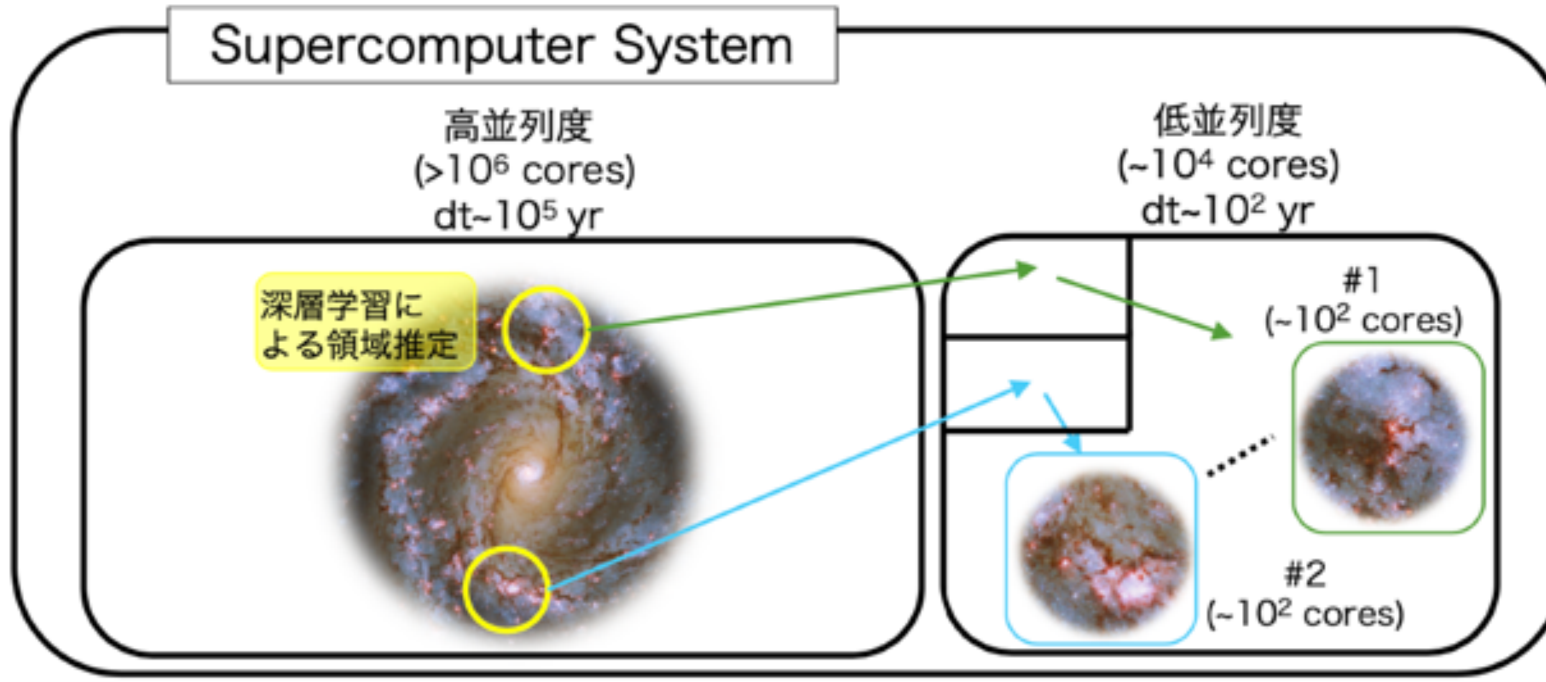


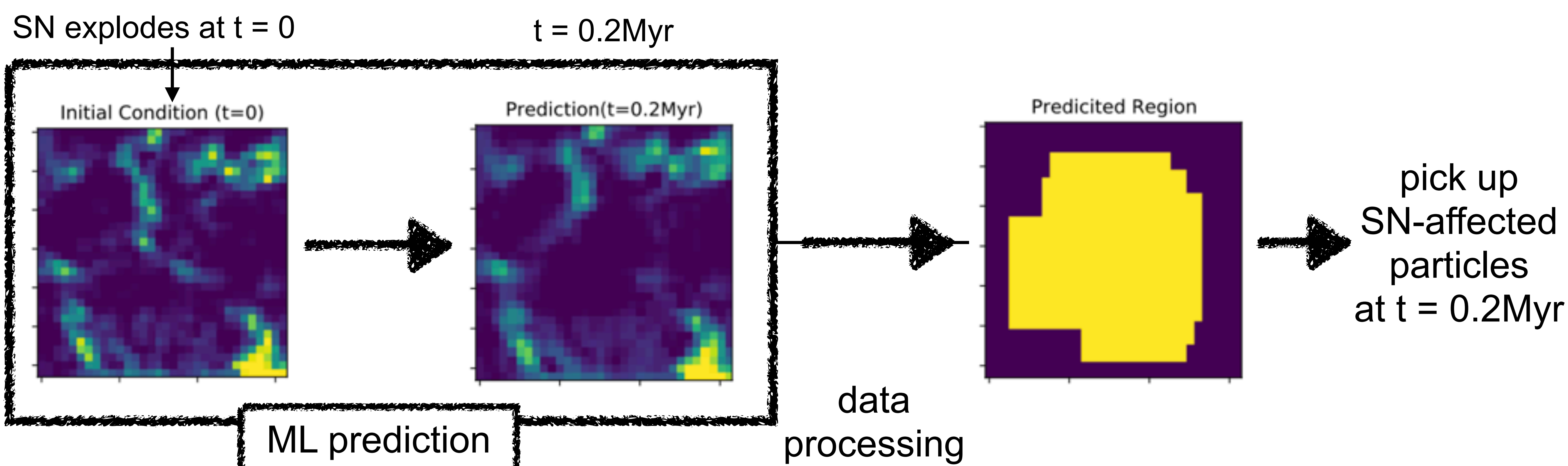
Processing simulation data with CNN + RNN

Hirashima, KM, et al. (in prep.)

ASURA-FDPS

- Purpose: galaxy formation simulation with ~1$M_\odot$ resolution.
- Bottleneck: very short time steps required to compute SN-influenced particles.
- We want to send such particles selectively to the low DOP server.



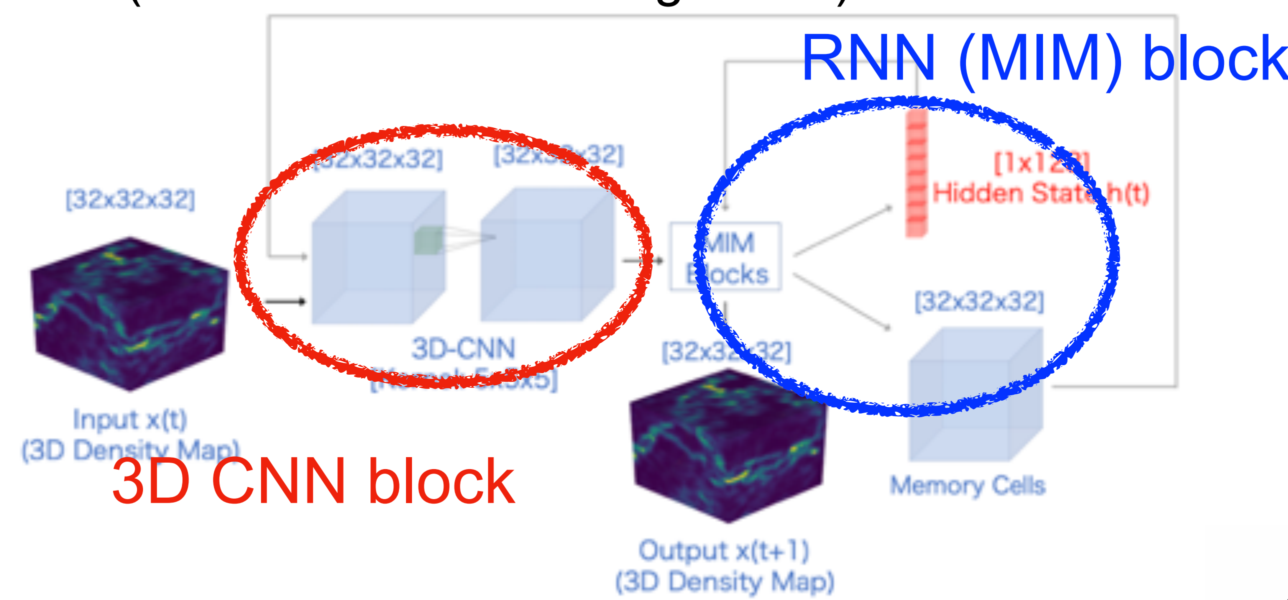


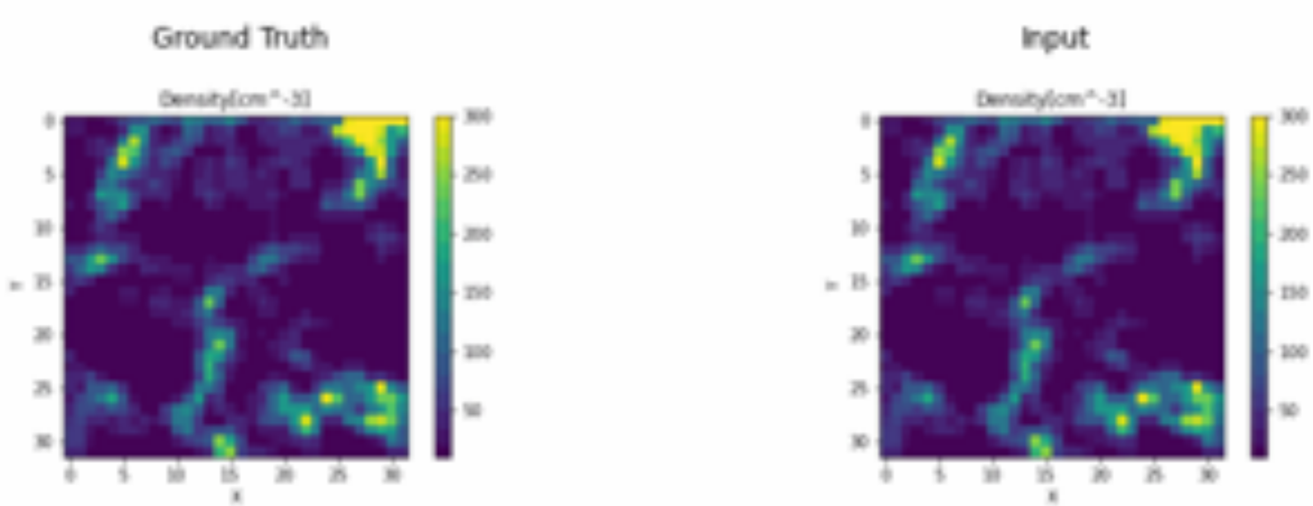
Processing simulation data with CNN + RNN

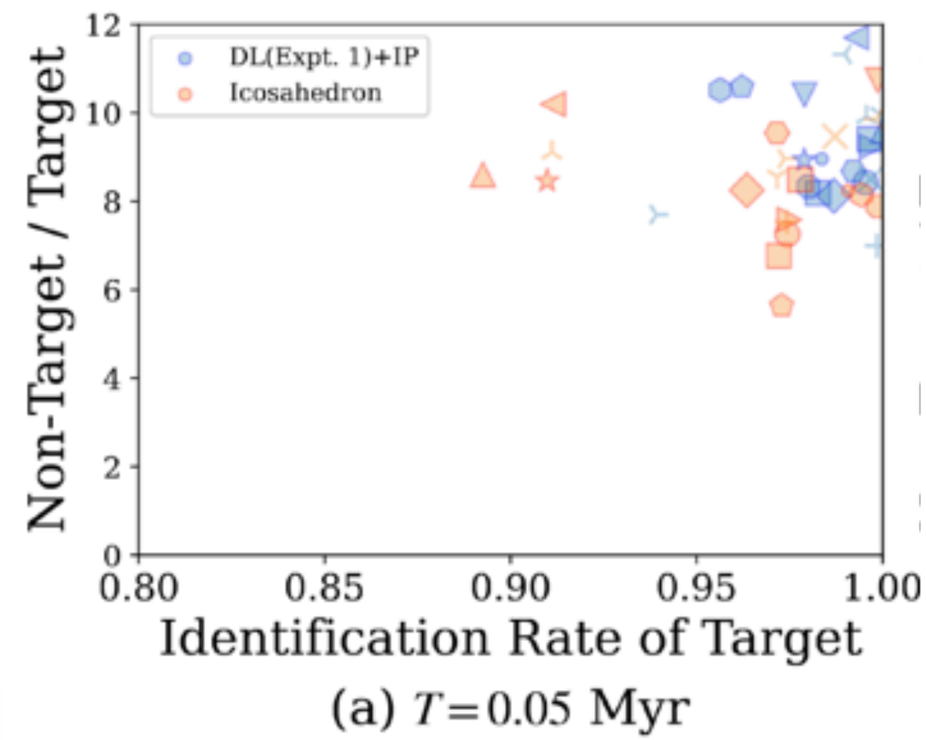
Hirashima, KM, et al. (in prep.)

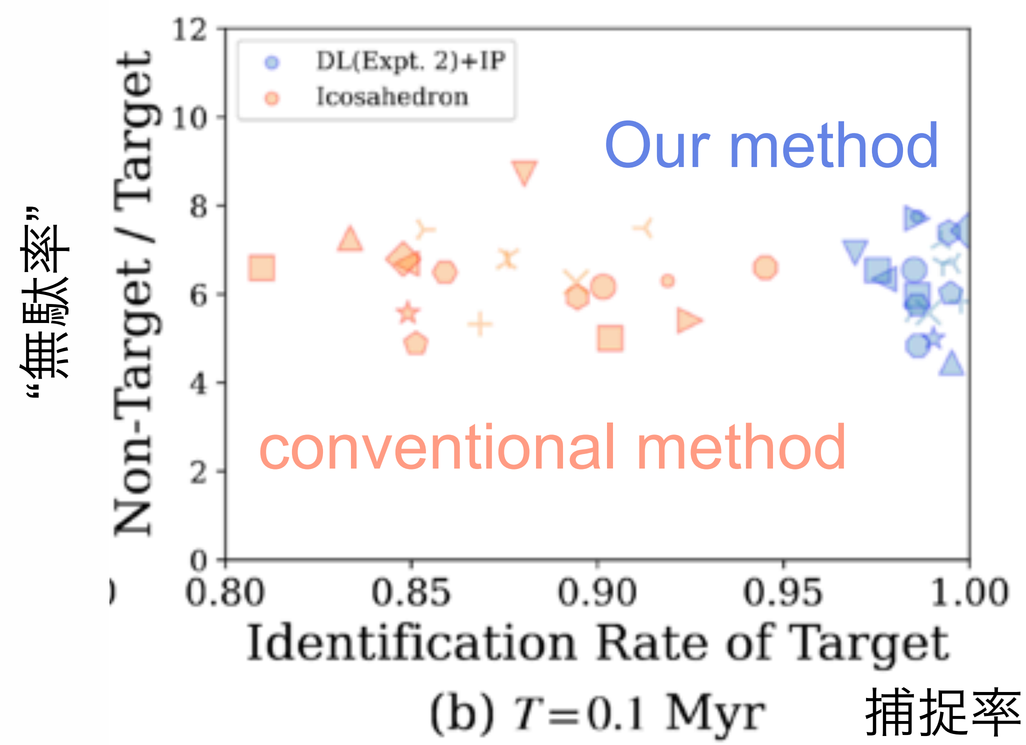
Extended MIM model

cf. MIM (2D CNN × RNN: Wang+2018)







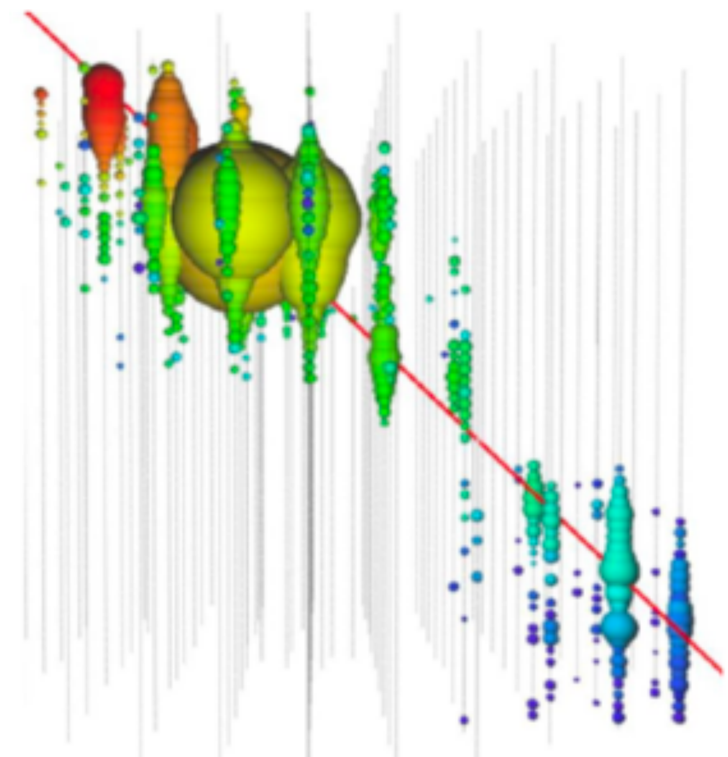


Graph Neural Network (GNN)

- Graph = a collection of nodes and edges
- Applications:
  - particle physics (e.g., Shlomi+2020),
  - neutrino detector (Choma+2018)
  - SPH simulation (Sanchez-Gonzalez+2020)

Figure 17: A visualisation of the dataflow for the three flavours of GNN layers, g. We use the neighbourhood of node b from Figure 10 to illustrate this. Left-to-right: **convolutional**, where sender node features are multiplied with a constant, c_{uv}; **attentional**, where this multiplier is *implicitly* computed via an attention mechanism of the receiver over the sender: $\alpha_{uv} = a(\mathbf{x}_u, \mathbf{x}_v)$; and **message-passing**, where vector-based messages are computed based on both the sender and receiver: $\mathbf{m}_{uv} = \psi(\mathbf{x}_u, \mathbf{x}_v)$.

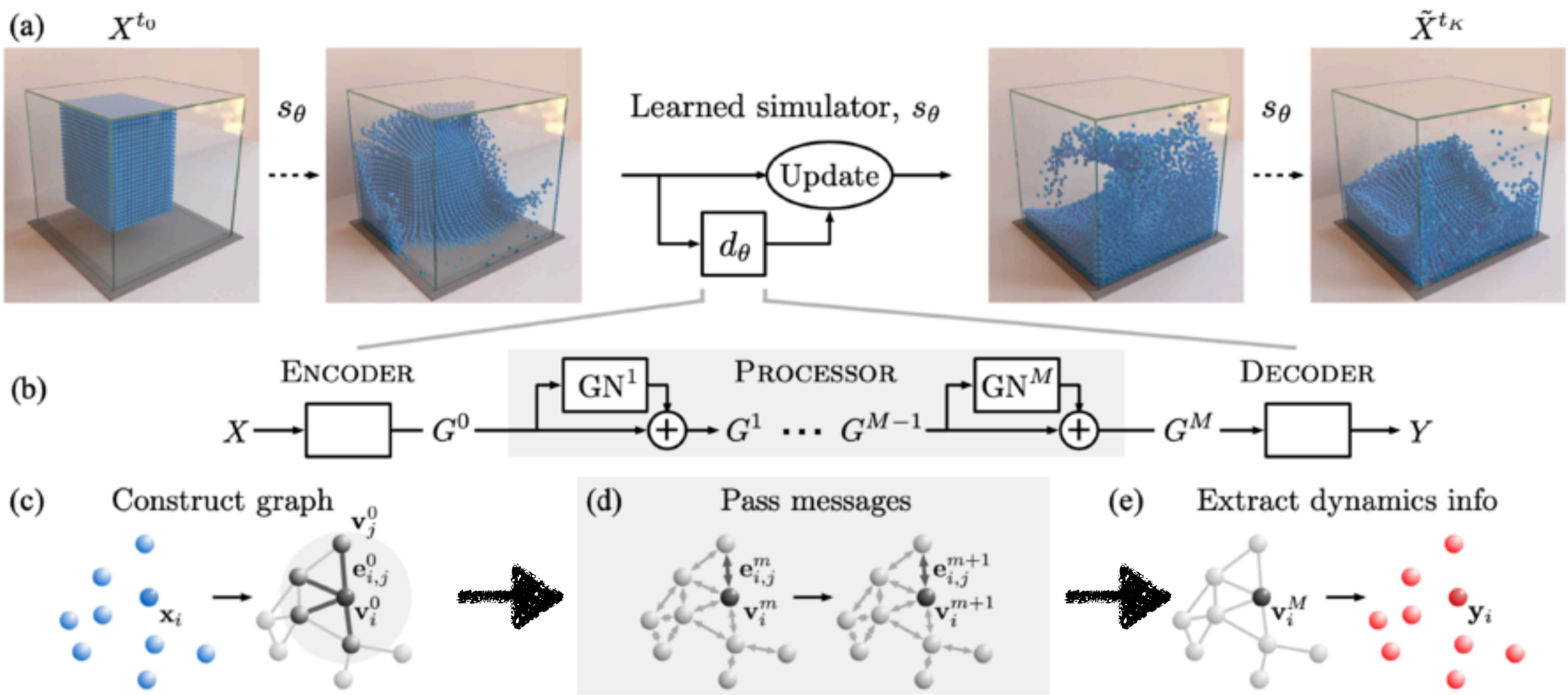
Bronstein et al. (2021)



Neutrino detector (Choma+2018)

SPH simulation with GNN

Sanchez-Gonzalez+2020 (DeepMind)



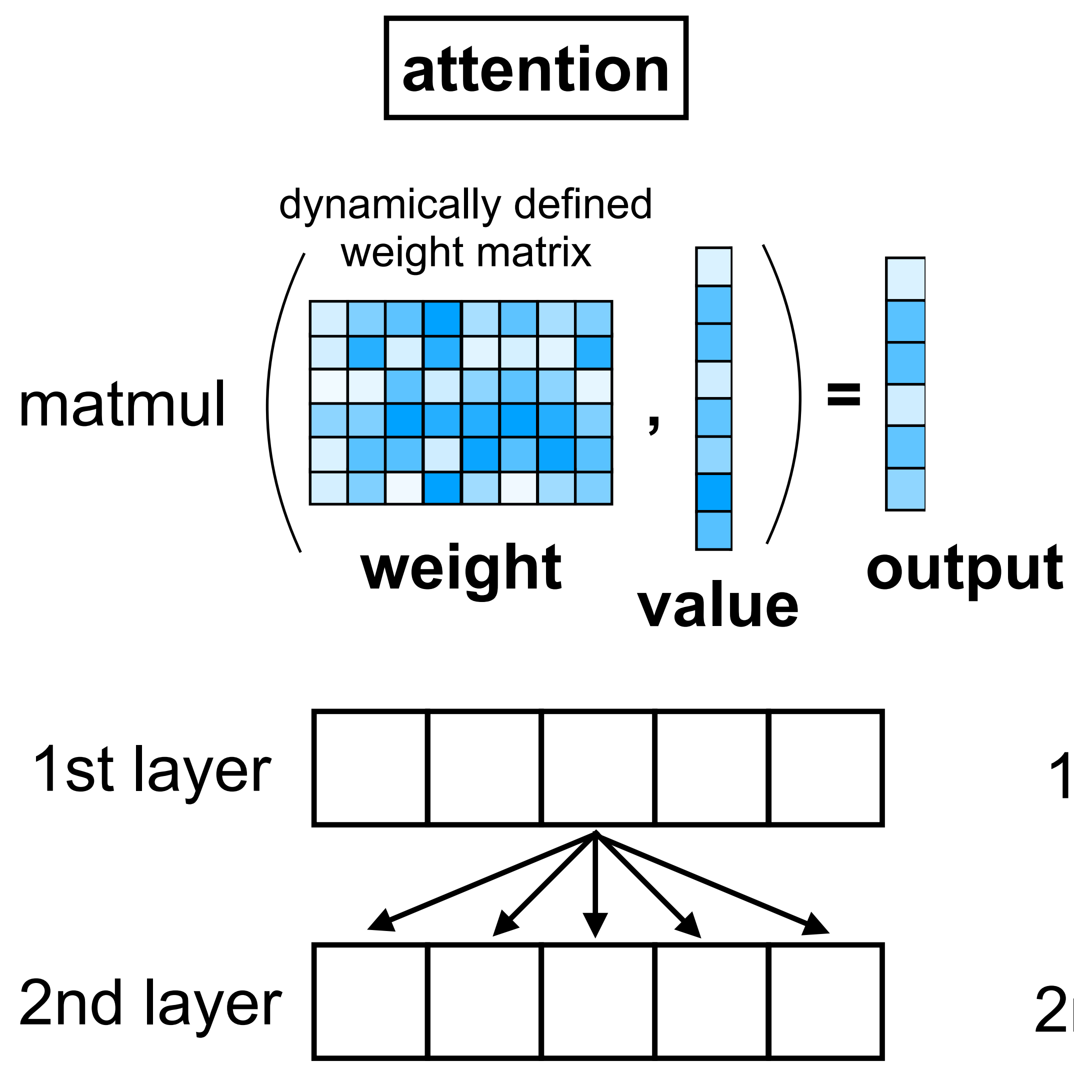
Possible advantages of the machine learning simulators

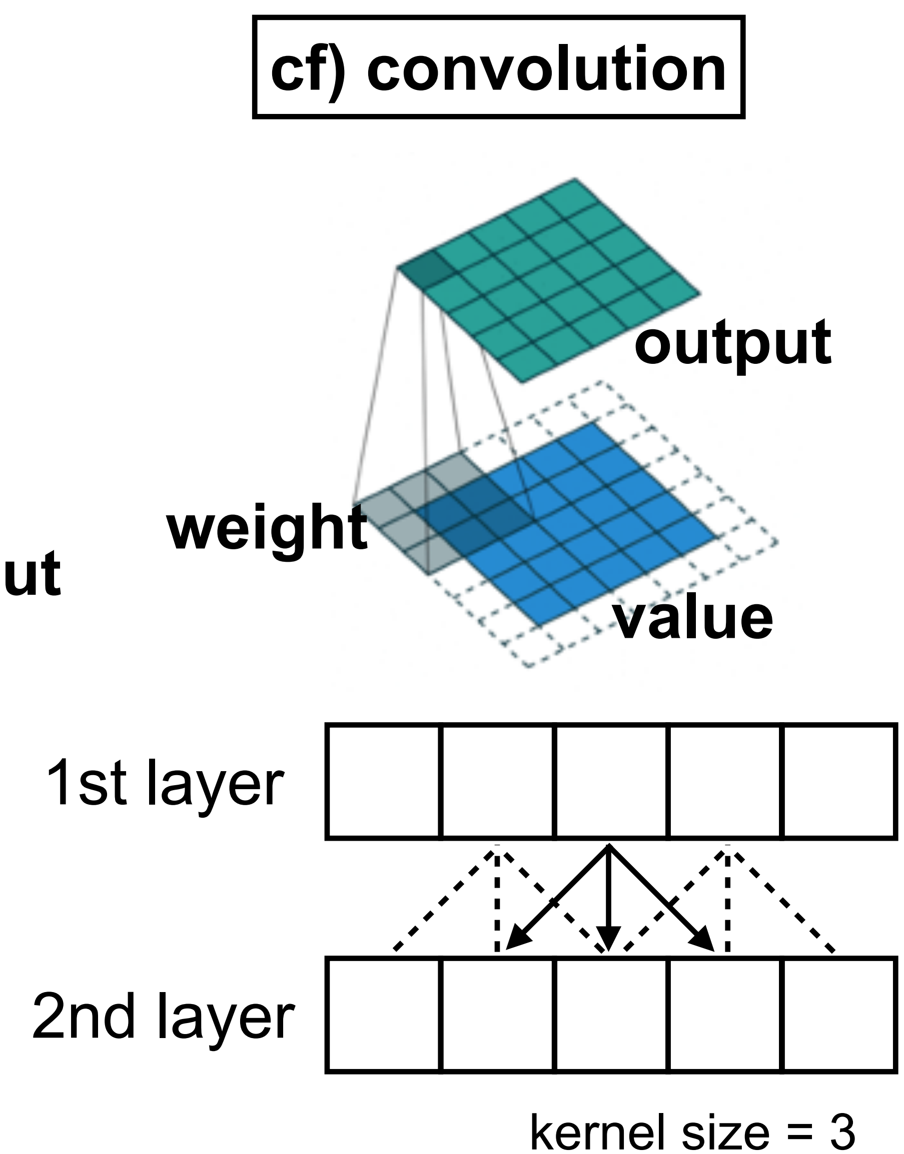
- Could be faster than numerical simulation (future work: more efficient, parallelizable networks).
- Trained directly from observed data.
- If it is optimized for inverse objectives it would be valuable for solving inverse problems.

Transformer

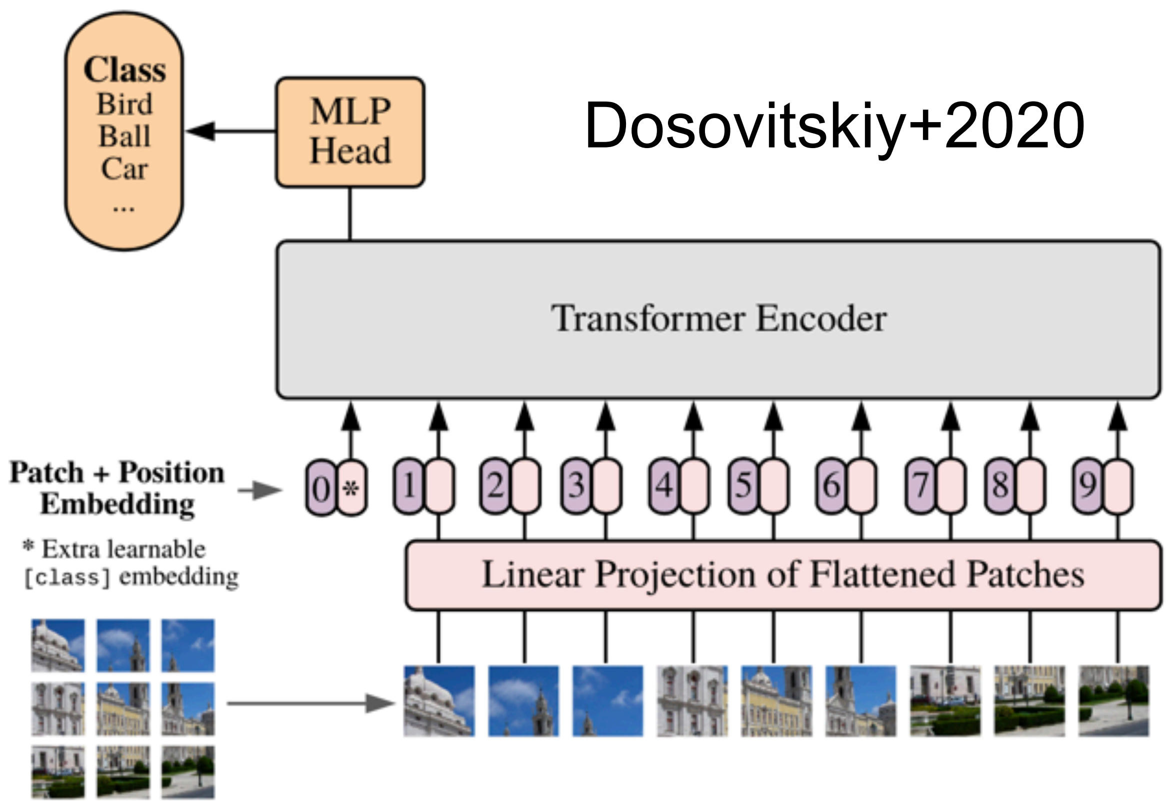
- Transformer: ML model based on **attention mechanisms** (Vaswani+2017)
  = a GNN with every node connected to the every other node
- SoTA NLP models: BERT (Devlin+2018; Google), GPT-3 (Brown+2020; OpenAI)
- **Long-range dependencies** are more easily computed.

Vaswani+17





Vision Transformer (ViT)

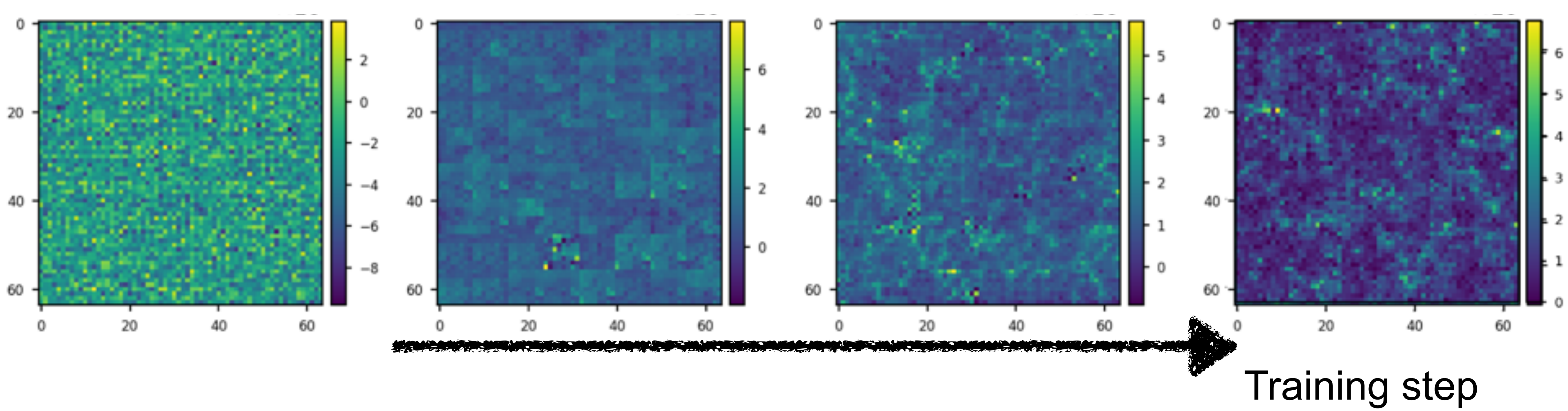


Dosovitskiy+2020

See also

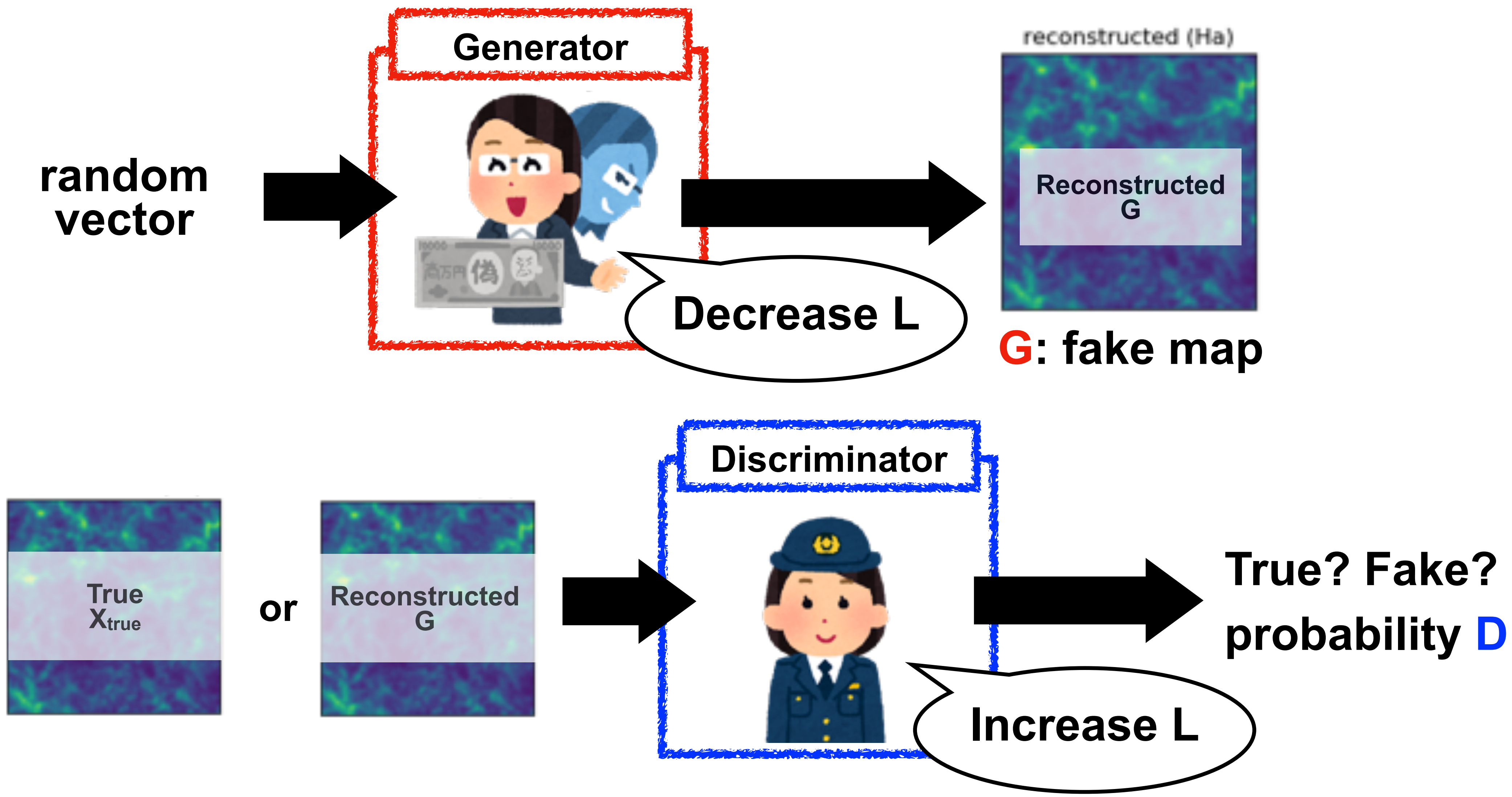
- Thuruthipilly et al. (2021): application of ViT for lensed images
- Allam Jr. & McEwen (2021): classification of supernovae

e.g., generation of LSS



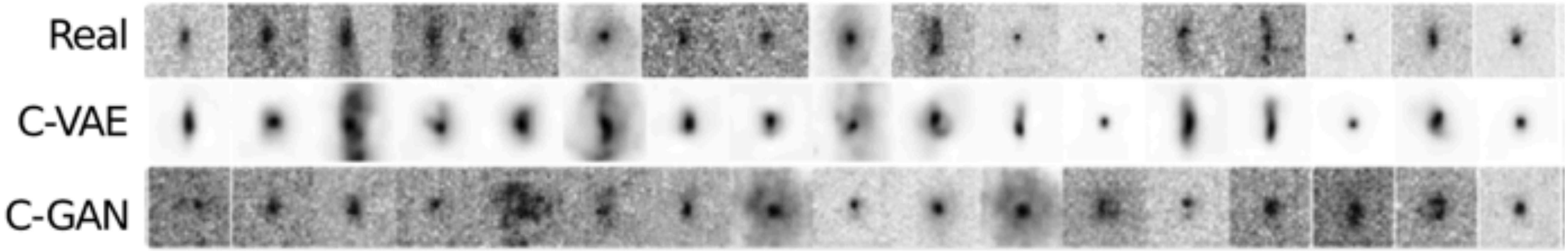
Generative Adversarial Network (GAN)

- GAN (Goodfellow+2014), conditional GAN (Isola+2016)
- Two networks generator and discriminator are updated in an adversarial way.

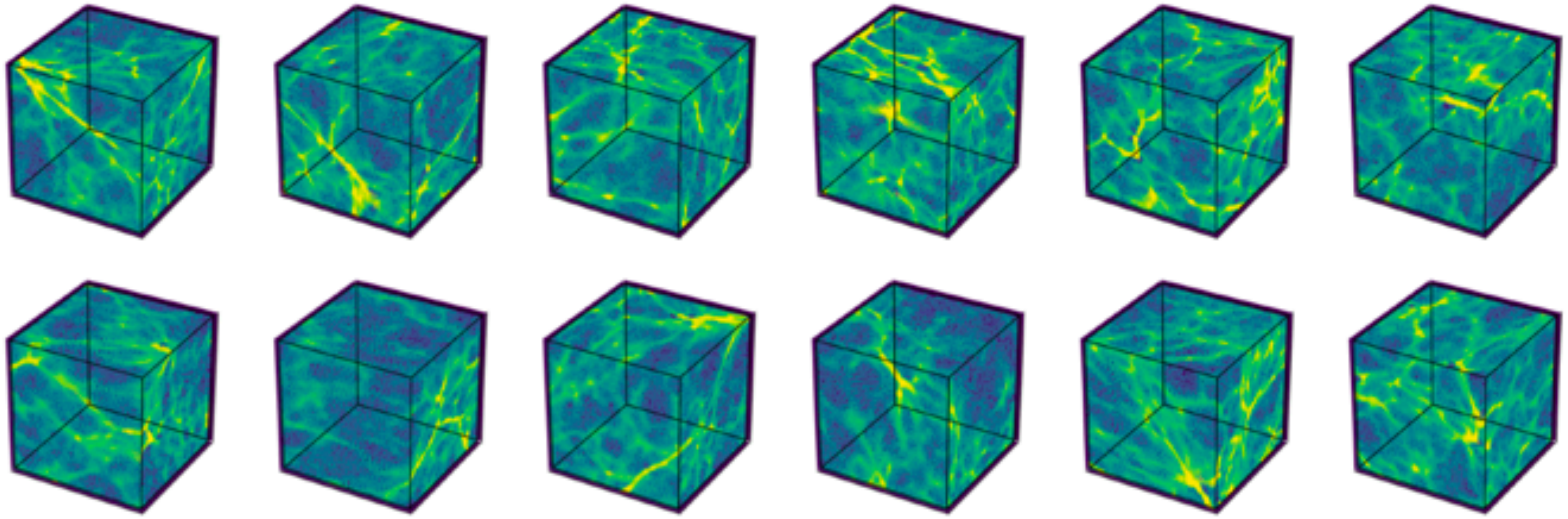


loss function: $$L[G, D] = \log D(X_{\rm true}) + \log[1 - D(G)]$$

Generative Adversarial Network (GAN)



Generate realistic images of galaxies for lensing study (Ravanbakhsh+16)



Generate mock neutral hydrogen map (HIGAN; Zamudio-Fernandez+19)

General Issues

- Comparison between models
- Accuracy and validity of ML outputs

Comparison between models

- Which model is the best? Pros/cons of each method?

→ Need for publicly accessible reference datasets for systematic comparison (cf. MNIST, ImageNet, etc.)

Photo-z estimation contest with PHAT-1 sample (Hildebrandt+10, Cavuoti+12)

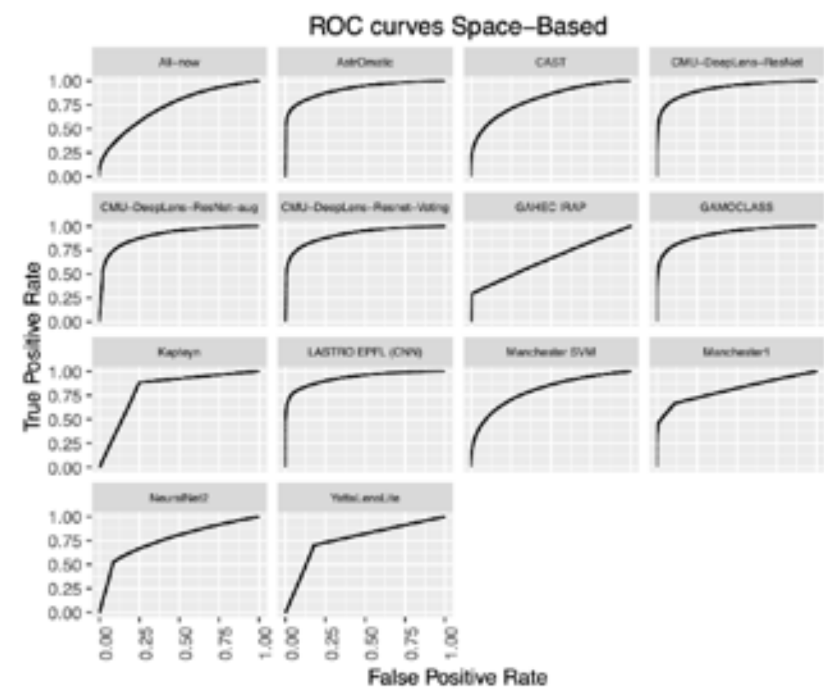
A	18-band; $\vert\Delta z\vert \le 0.15$			14-band; $\vert\Delta z\vert \le 0.15$			18-band; $R < 24$; $\vert\Delta z\vert \le 0.15$			14-band; $R < 24$; $\vert\Delta z\vert \le 0.15$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	0.0006	0.056	16.3	0.0028	0.063	19.3	0.0002	0.053	11.7	0.0016	0.060	13.7
AN-e	–0.010	0.074	31.0	–0.006	0.078	38.5	–0.013	0.071	24.4	–0.007	0.076	32.8
EC-e	–0.001	0.067	18.4	0.002	0.066	16.7	–0.006	0.064	14.5	–0.003	0.064	13.5
PO-e	–0.009	0.052	18.0	–0.007	0.051	13.7	–0.009	0.047	10.7	–0.008	0.046	7.1
RT-e	–0.009	0.066	21.4	–0.008	0.067	24.2	–0.012	0.063	16.4	–0.012	0.064	18.4
B	18-band; $\vert\Delta z\vert \le 0.5$			14-band; $\vert\Delta z\vert \le 0.5$			18-band; $R < 24$; $\vert\Delta z\vert \le 0.5$			14-band; $R < 24$; $\vert\Delta z\vert \le 0.5$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	–0.0028	0.114	3.8	–0.0046	0.125	3.8	–0.0039	0.101	1.7	–0.0039	0.101	1.7
AN-e	–0.036	0.151	3.1	–0.035	0.173	4.2	–0.047	0.130	1.4	–0.047	0.130	1.4
EC-e	–0.007	0.120	3.6	–0.003	0.114	3.6	–0.015	0.106	1.9	–0.015	0.106	1.9
PO-e	–0.013	0.124	3.1	0.001	0.107	2.3	–0.020	0.098	1.2	–0.020	0.098	1.2
RT-e	–0.031	0.126	3.2	–0.028	0.137	3.6	–0.034	0.111	1.4	–0.034	0.111	1.4

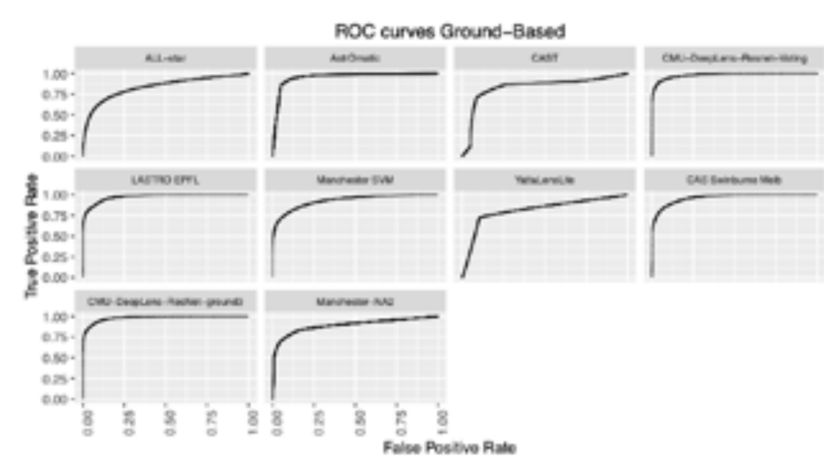
Comparison between models

“The strong gravitational lens finding challenge” (Metcalf+2019)

- A training set was provided to participants
- The participants are requested to upload the results within 48hrs after given test data set

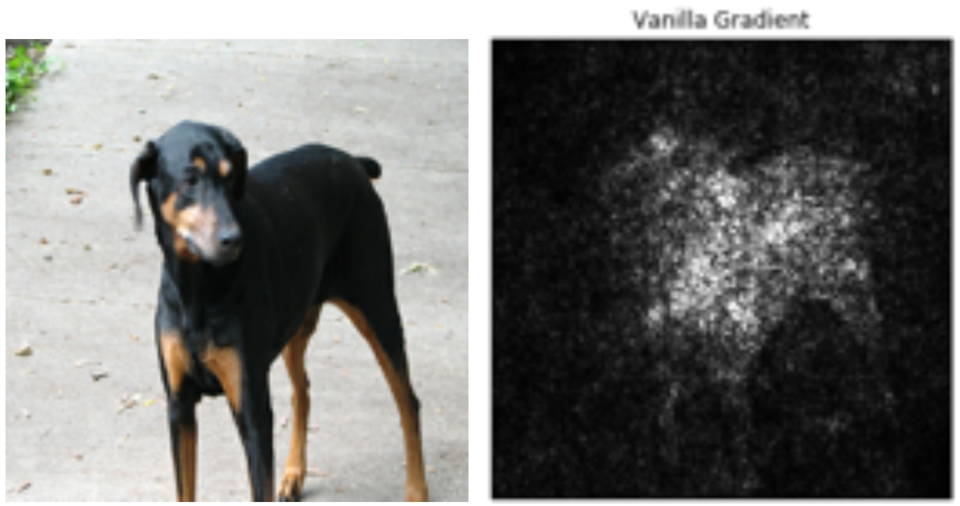
Name	Type	AUROC	TPR_0	TPR_{10}	Short description
Manchester SVM	Ground-based	0.93	0.22	0.35	SVM/Gabor
CMU-DeepLens-Resnet-ground3	Ground-based	0.98	0.09	0.45	CNN
LASTRO EPFL	Ground-based	0.97	0.07	0.11	CNN
CMU-DeepLens-Resnet-Voting	Ground-based	0.98	0.02	0.10	CNN
CAS Swinburne Melb	Ground-based	0.96	0.02	0.08	CNN
ALL-star	Ground-based	0.84	0.01	0.02	Edges/gradiants and Logistic Reg.
Manchester2	Ground-based	0.89	0.00	0.01	Human Inspection
YattaLensLite	Ground-based	0.82	0.00	0.00	SExtractor
CAST	Ground-based	0.83	0.00	0.00	CNN/SVM
AstrOmatic	Ground-based	0.96	0.00	0.01	CNN
CMU-DeepLens-Resnet	Space-based	0.92	0.22	0.29	CNN
GAMOCLASS	Space-based	0.92	0.07	0.36	CNN
CAST	Space-based	0.81	0.07	0.12	CNN
All-now	Space-based	0.73	0.05	0.07	Edges/gradiants and Logistic Reg.
Manchester SVM	Space-based	0.80	0.03	0.07	SVM/Gabor
Manchester1	Space-based	0.81	0.01	0.17	Human Inspection
LASTRO EPFL	Space-based	0.93	0.00	0.08	CNN
GAHEC IRAP	Space-based	0.66	0.00	0.01	Arc finder
AstrOmatic	Space-based	0.91	0.00	0.01	CNN
Kapteyn Resnet	Space-based	0.82	0.00	0.00	CNN
CMU-DeepLens-Resnet-aug	Space-based	0.91	0.00	0.00	CNN
CMU-DeepLens-Resnet-Voting	Space-based	0.91	0.00	0.01	CNN
NeuralNet2	Space-based	0.76	0.00	0.00	CNN/wavelets
YattaLensLite	Space-based	0.76	0.00	0.00	Arcs/SExtractor



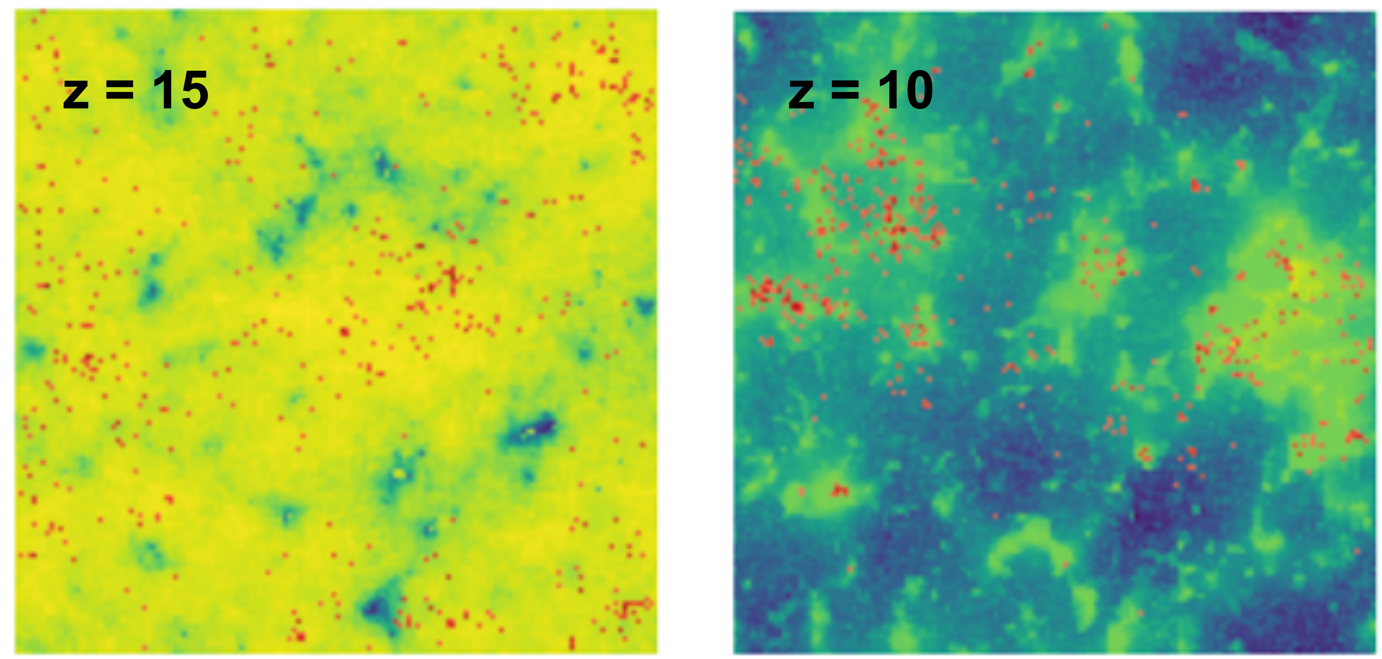


Accuracy and validity of ML outputs: Can we explain the machine's strategies?

Saliency analysis (SA)



$$\text{vanilla gradient} = \frac{\mathrm{d}y_{\text{class}}}{\mathrm{d}x_{ij}}$$



Villanueva-Domingo & Villaescusa-Navarro (2020)

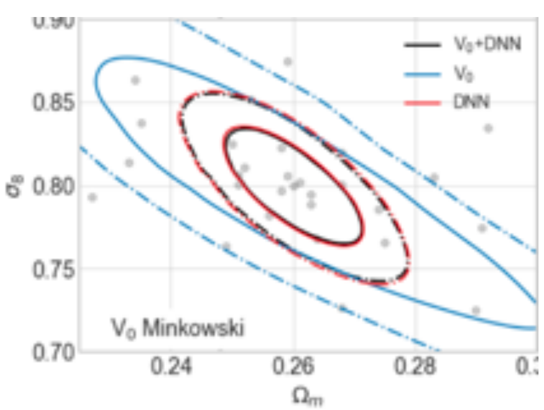
- 21-cm map → properties of ionizing/heating sources (e.g., M_{turn}, L_X, N_γ)
- Red points: large saliency for L_X
- Insight from SA: the machine focus on bright 21cm regions when estimating L_X

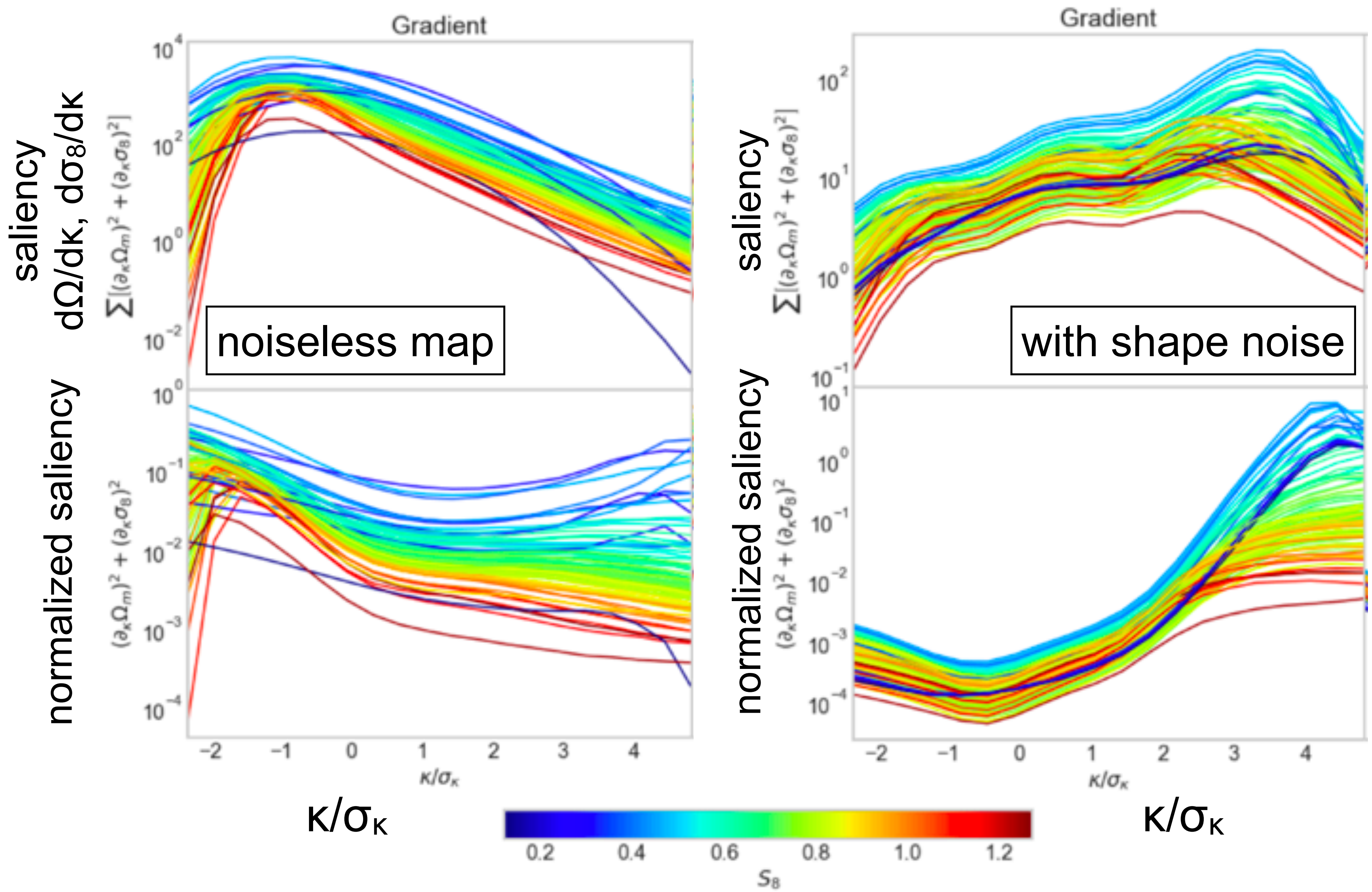
Saliency analysis

Matilla+2020

- Parameter estimation from weak lensing maps
- The machine put more focus on the low-κ regions, consistent with previously known results

WL map (κ)

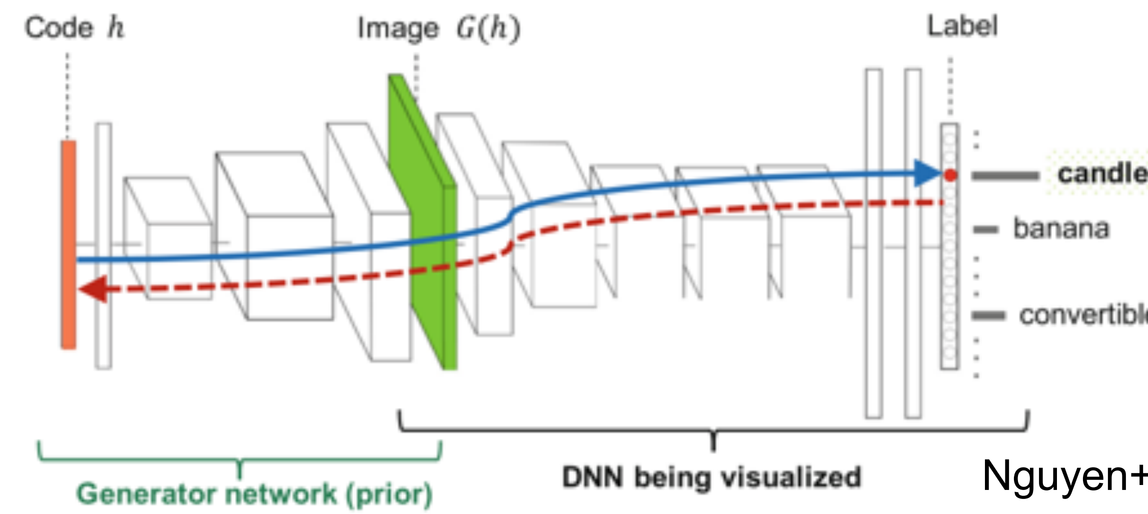




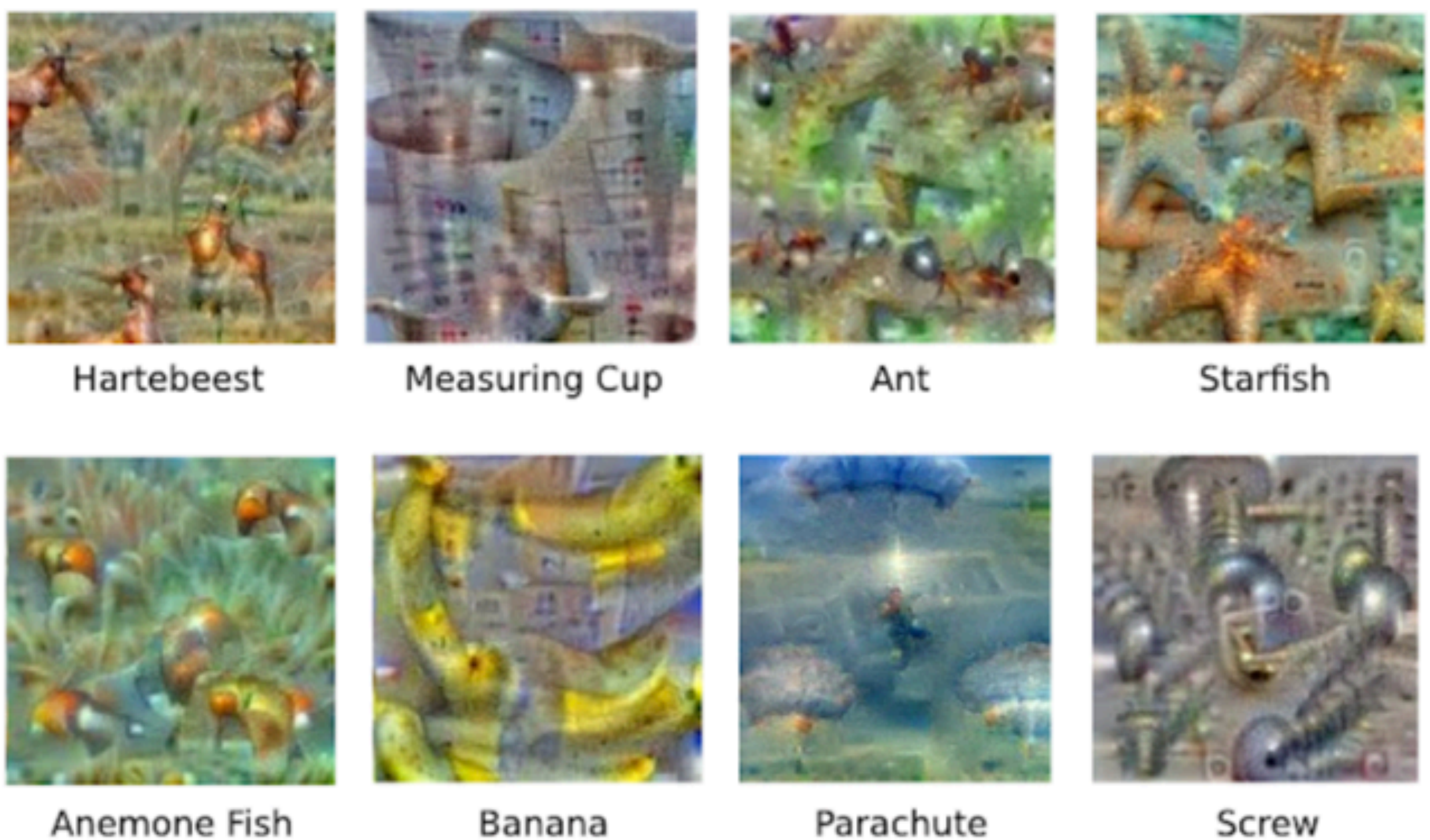
Activation maximization method

DeepDream (Mordvintsev+15)

- Activate all the units across a given layer simultaneously
- Visualization of “higher-order” information
- Could also be used for explaining ML classifications of astrophysical data



Nguyen+19

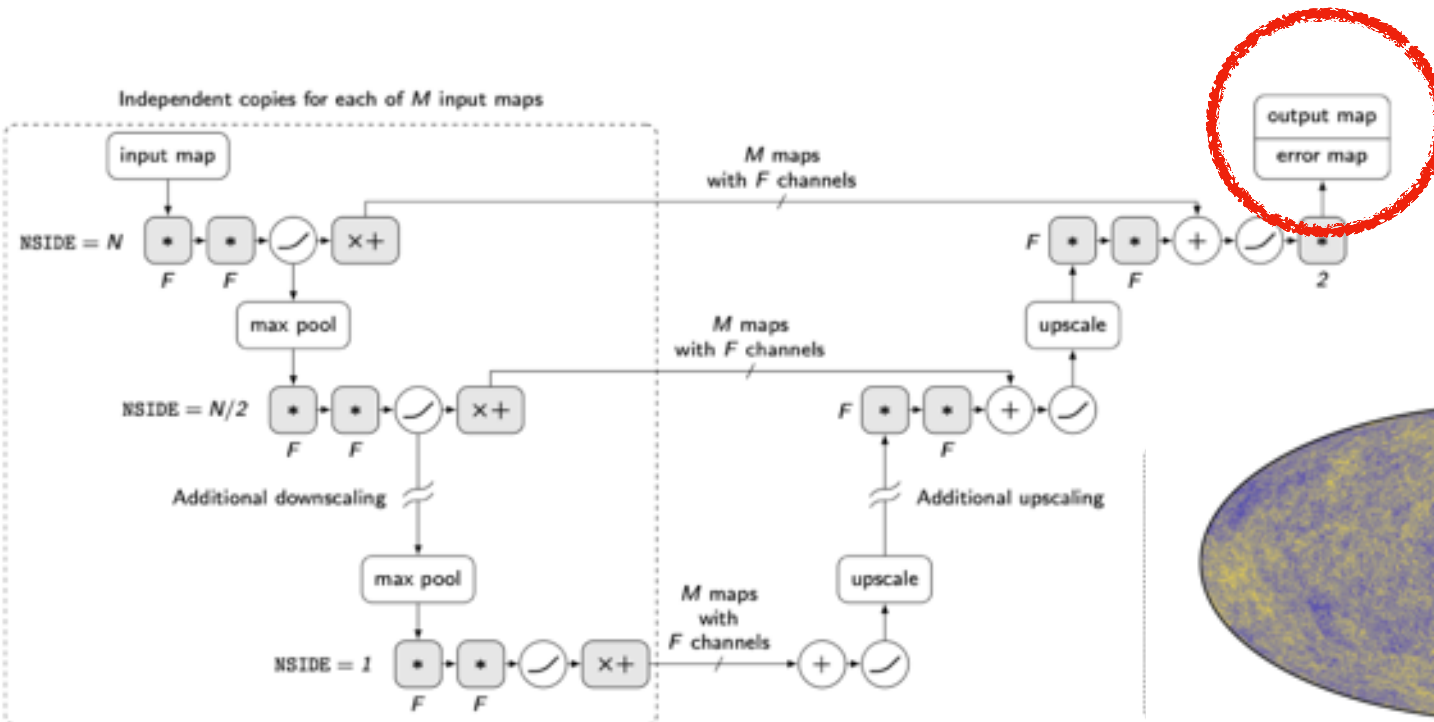


https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

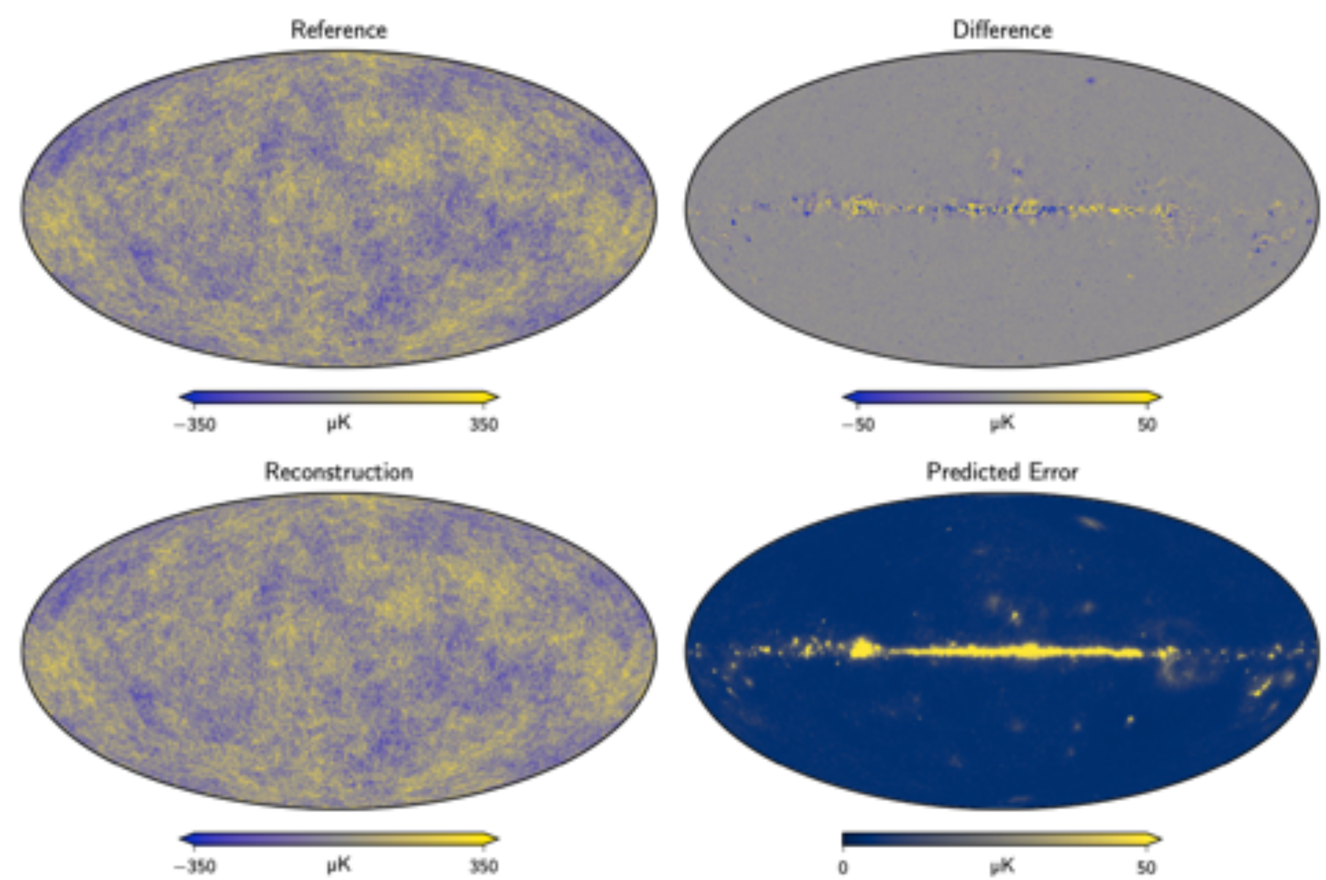
Accuracy and validity of ML outputs: Uncertainty of the machine's output

- How much uncertainty is there in the machine's output?

→ Train the network to estimate uncertainties as well



CMB foreground cleaning (Patrof+20)

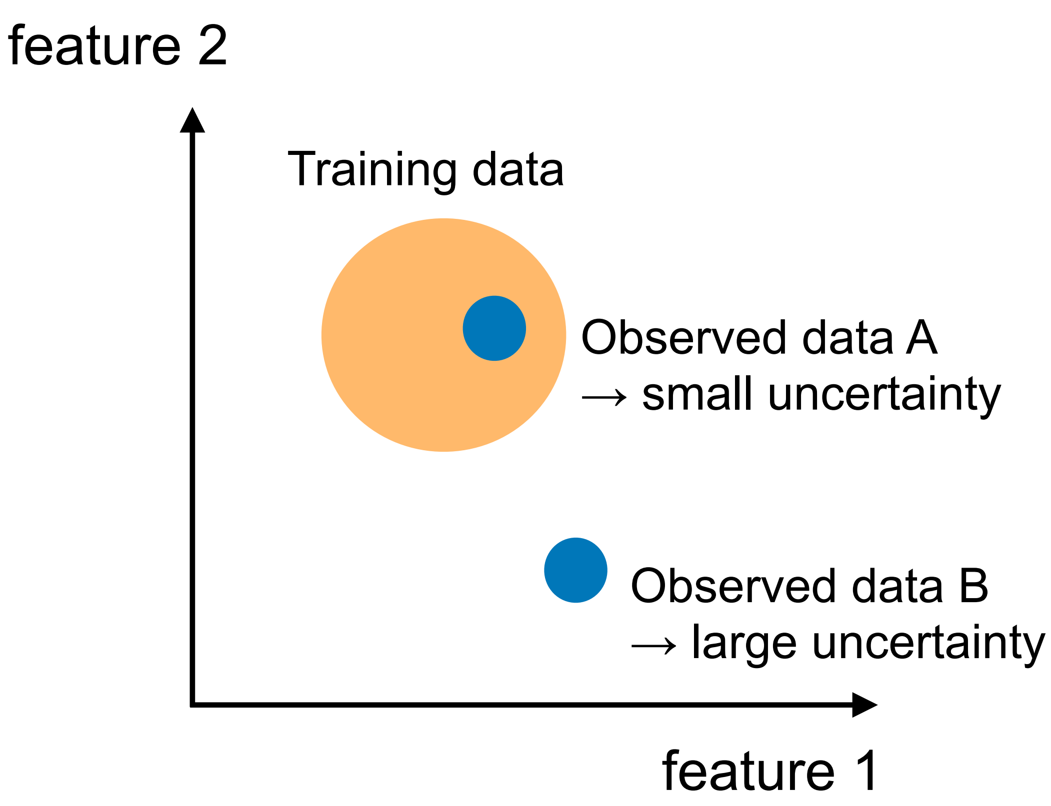


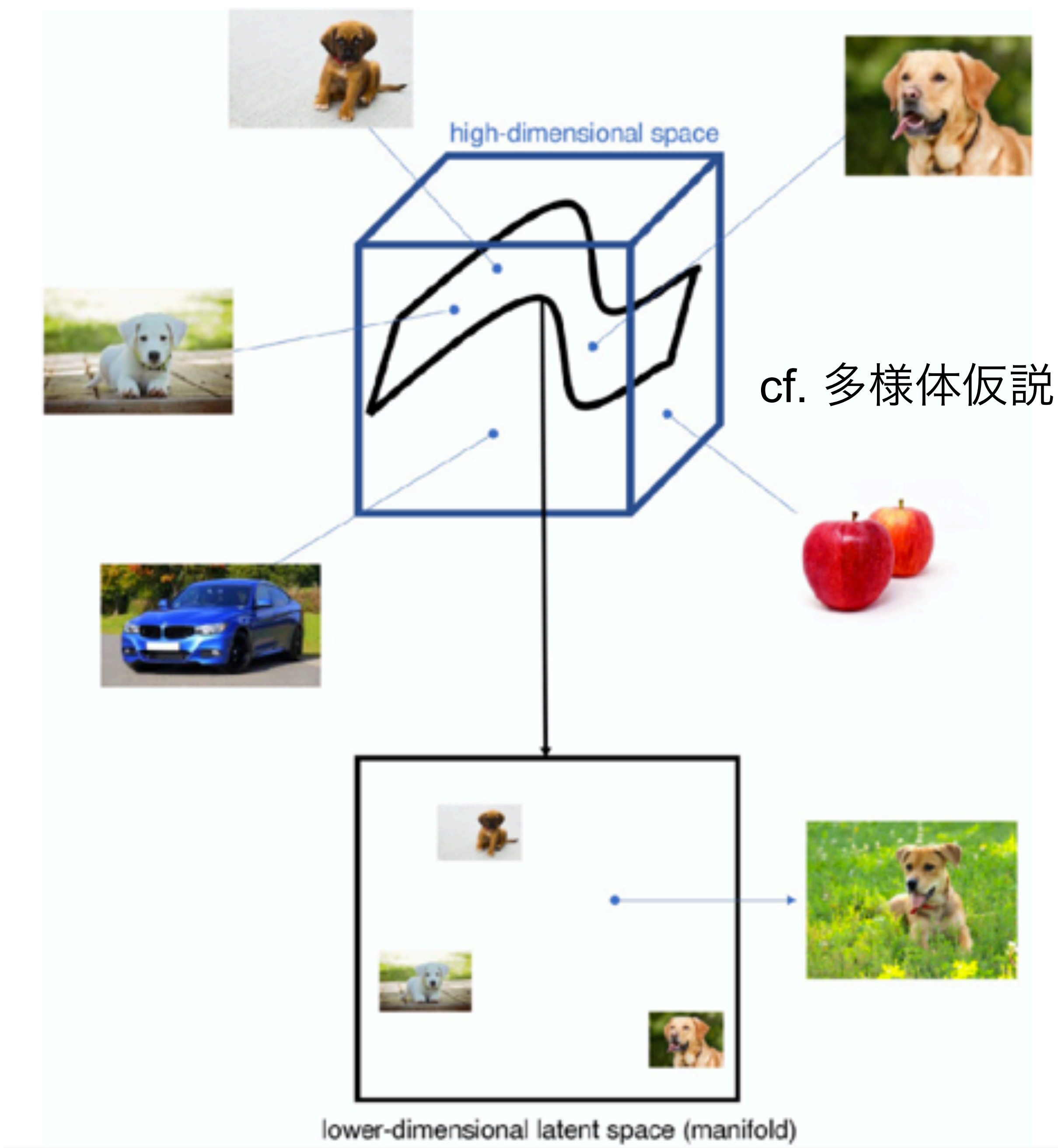
Aleatoric uncertainty（観測データに内在するノイズ）に対してはうまく予測できるが、学習中に見たことがないようなデータに対してはうまく予測できない（cf. Kendall & Gal 2017)

Distribution in high-dimensional space

- What to do with data that has never been seen during training?

→ Let's think about the distribution in high-dimensional space





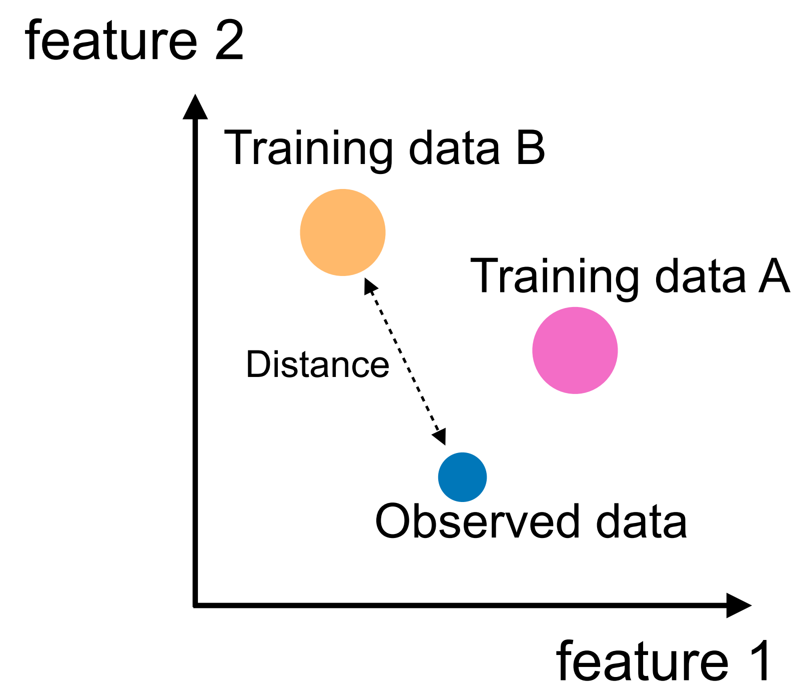
cf. 多様体仮説

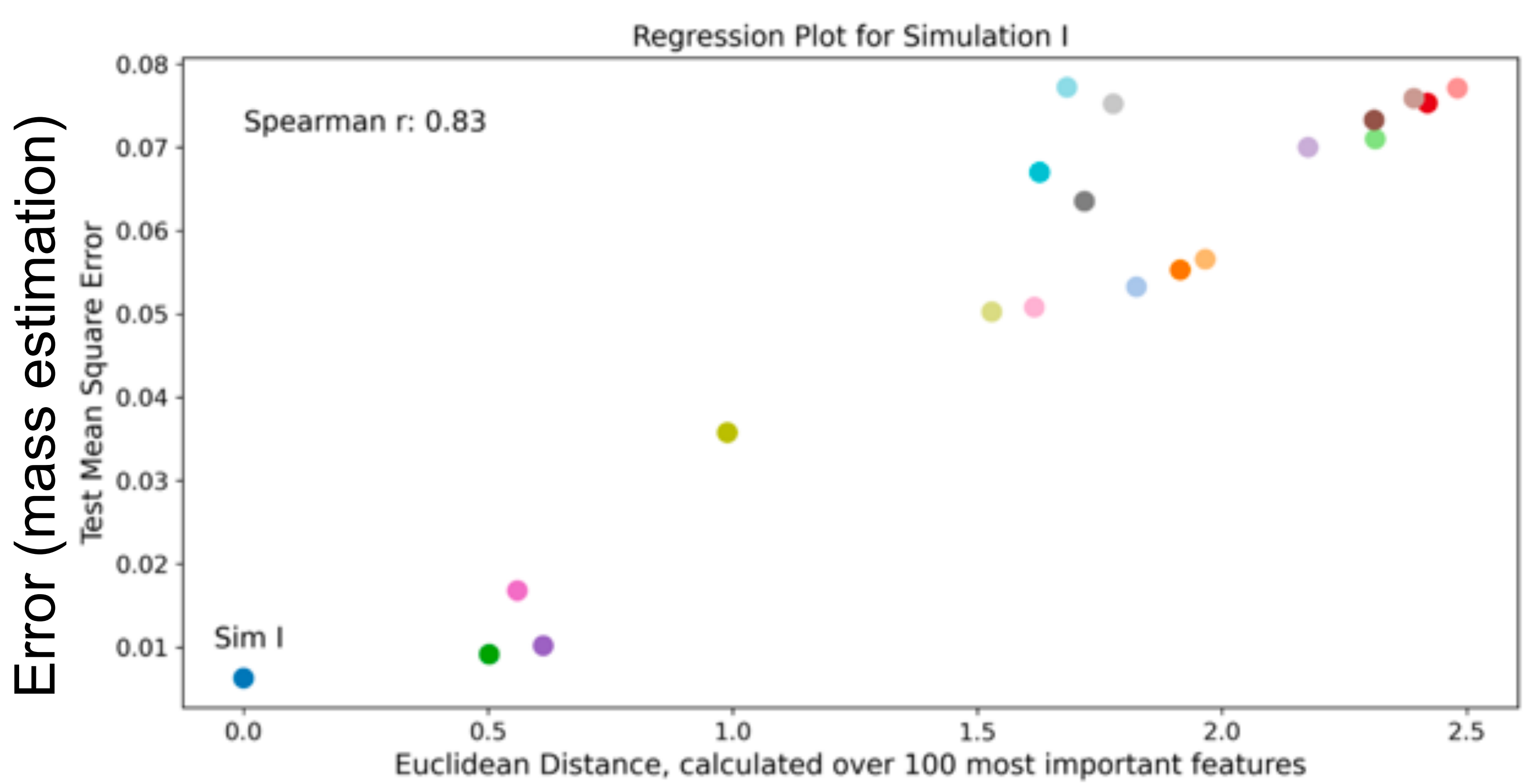
cf. Deep k-nearest neighbors (Papernot & McDaniel 2018)

Use distributions in high-dimensional space to estimate uncertainties

Acquaviva+2020

- Task: galaxy spectrum → stellar mass (regression)
- Train multiple models with different training datasets generated with different physical models
- Assumption: the generalization error depends on the “distance” between observed and training data.
- Train another machine to learn the distance metric.





Distance between input data and the training datasets

Summary

- Various ML methods are used for various kinds of astronomical data for various purposes.
  - ML could deal with a lot of data at high speed
  - ML could capture complex structures/relations
  - We may be able to obtain new insights from ML outputs
- Relatively new methods such as CNNs, RNNs, and GANs are also proving to be useful.
- A lot of collaboration with people in the statistics/AI fields, especially in emerging phases.

- General issues
  - Sharing reference datasets is crucial for systematic comparison between conventional methods and ML models.
  - Explainable AI techniques are getting applied for astrophysical studies — mostly just making sure that the machine is working as expected rather than finding new physics.
  - Several methods to estimate the uncertainties of the ML outputs are proposed
  - Still, it is difficult to deal with completely unpredictable data (but this could also be true for the other methods besides ML!)

References

- Ball & Brunner (2010) “Data Mining and Machine Learning in Astronomy”, International Journal of Modern Physics D, 19, 1049
- Baron (2019) “MACHINE LEARNING IN ASTRONOMY: A PRACTICAL OVERVIEW”, arXiv:1904.07248
- Ntampaka et al. (2019) “The Role of Machine Learning in the Next Decade of Cosmology”, BAAS, 51, 14 (Astro2020 Science White Paper)
- Fluke & Jacobs (2020) “Surveying the reach and maturity of machine learning and artificial intelligence in astronomy”, WIREs Data Mining and Knowledge Discovery, 10, e1349

- Samek et al. eds. (2019) “Explainable AI: Interpreting, Explaining and Visualizing Deep Learning”, Springer International Publishing
- Bronstein et al. (2021) “Geometric Deep Learning Grids, Groups, Graphs, Geodesics, and Gauges”, arXiv:2104.13478