機械学習を用いた銀河サーベイデータ解析手法のレビュー

森脇 可奈（東京大学）

2023年度 天文・天体物理若手夏の学校

2023/8/3

自己紹介

森脇可奈

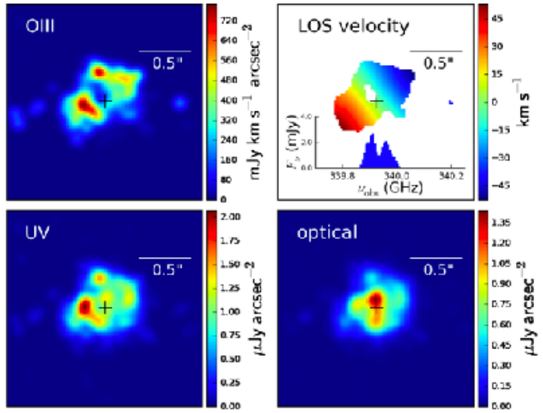
所属：

- 東京大学ビッグバン宇宙国際研究センター（RESCEU）
- 東京大学宇宙理論研究室（UTAP）

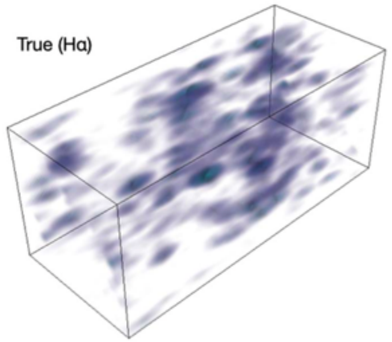
研究内容：

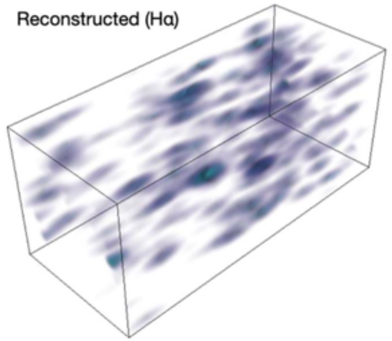
- 高赤方偏移銀河
- 宇宙再電離
- 機械学習

Today’s talk is based on our recent review
Moriwaki et al. (2023), Rep. Prog. Phys. 86 076901

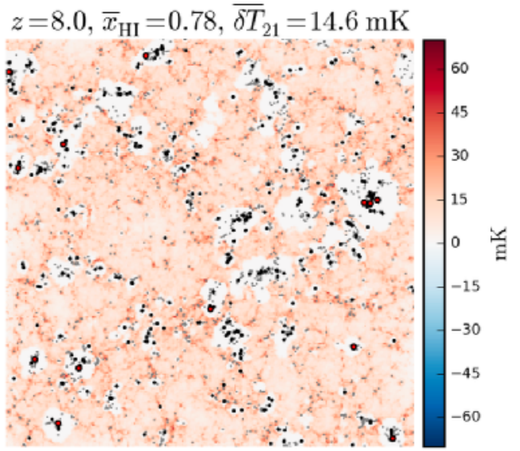


High-z galaxy (Moriwaki+18)



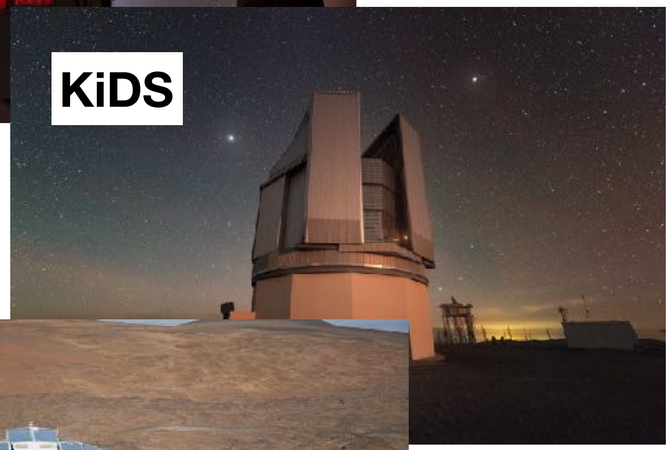


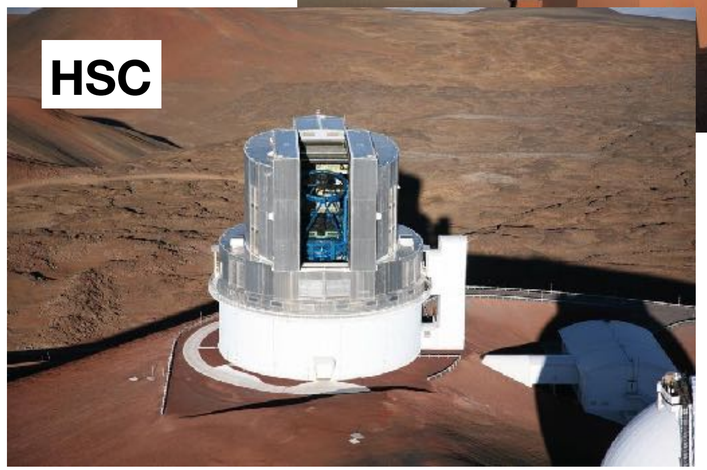
ML for noise removal (Moriwaki+21)

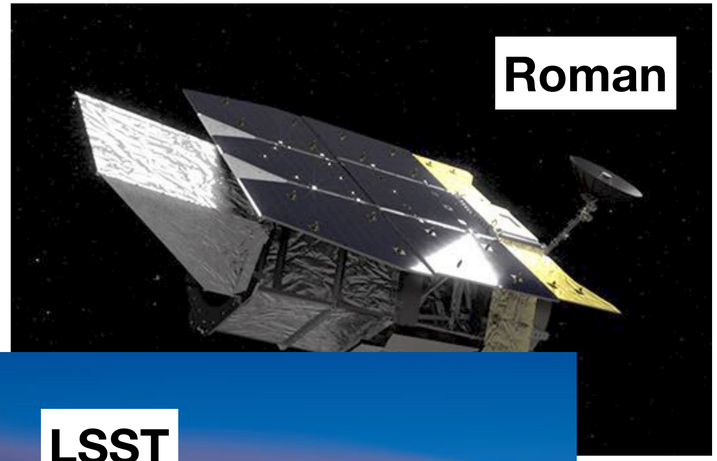


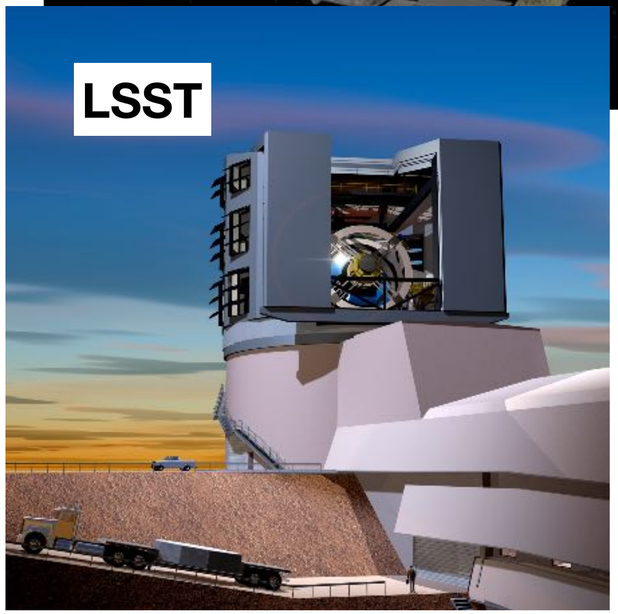
Reionization (Moriwaki+19)

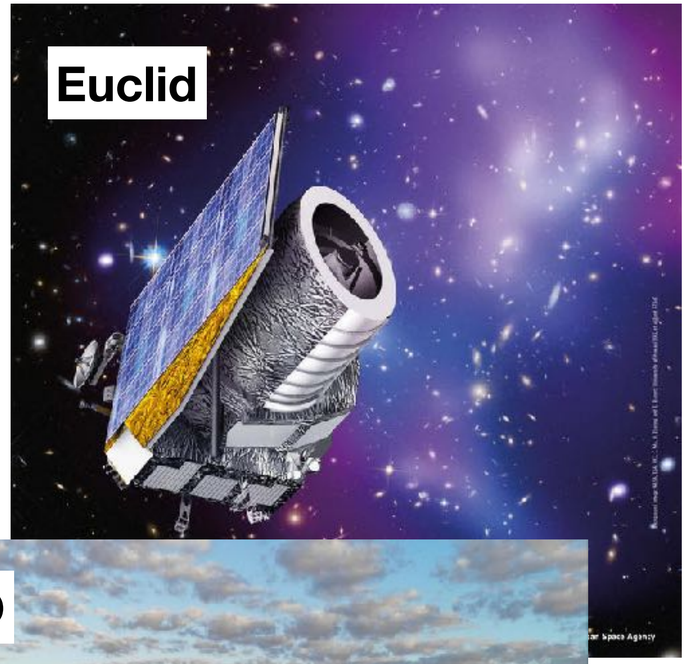
現在・今後の銀河サーベイ











数億〜数十億個もの銀河が今後見つかる → 機械学習を用いた自動処理がより重要に

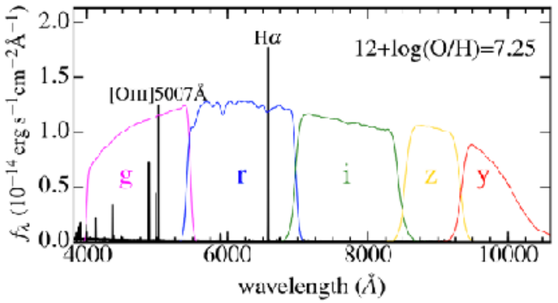
機械学習の様々な応用例

Images (Dieleman+15)

↓

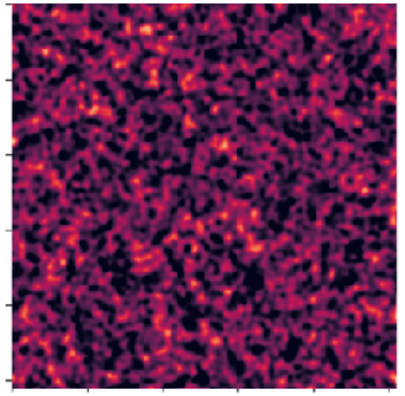
渦巻き？楕円？
合体？
(分類 / Classification)

Photometry (Kojima+20)



↓

パラメータ推定
赤方偏移、金属量、etc.
(回帰 / regression)



Weak lensing map (Shirasaki+19)

↓

宇宙論パラメータ
(回帰 / regression)

ノイズ除去
(生成 / Generation)

Random noise

↓

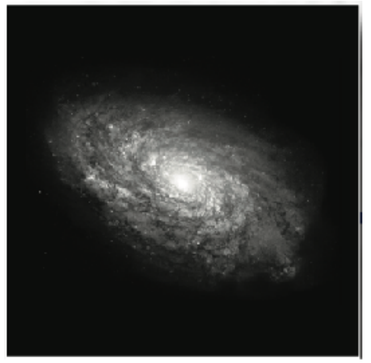


Image of galaxy (Ravanbakhsh+16)

(生成 / Generation)

今日話す内容

- イントロダクション ― 機械学習とは
- 観測データへの適用例
  1. 分類（Classification）― 銀河の形態分類
  2. 回帰（Regression）― 銀河の測光赤方偏移推定
  3. 発見（Discovery）― 強重力レンズ天体、超低金属量銀河候補
  4. 生成（Generation）― 模擬観測データの生成、ノイズ除去
- シミュレーションデータへの応用例

今日話す内容のまとめ（最後にまた繰り返します）

- 機械学習とは、データから自動的に最適解を見つけること
- 特に深層学習では大量のパラメータを最適化することで複雑な課題もこなすことができる

- 機械学習を使うメリット：
  - 高速に大量のデータを処理できる
  - 解析的に処理するのが難しいタスクをこなせる

  → 今後の銀河サーベイプロジェクトにおいて機械学習が重要となる

- 特に CNN を用いた画像解析での成功例が多い（e.g., 形態分類、photoz、重力レンズ）
- 目的によって評価指標が異なる（e.g., レアな天体の発見は recall が大きければOK）
- 新しいモデルも次々と用いられている（e.g., GAN）
- 学習データをどのように増やすかが大きな課題（e.g., photoz、CAMELS project）

イントロダクション

ー 機械学習とは ー

例：回帰問題

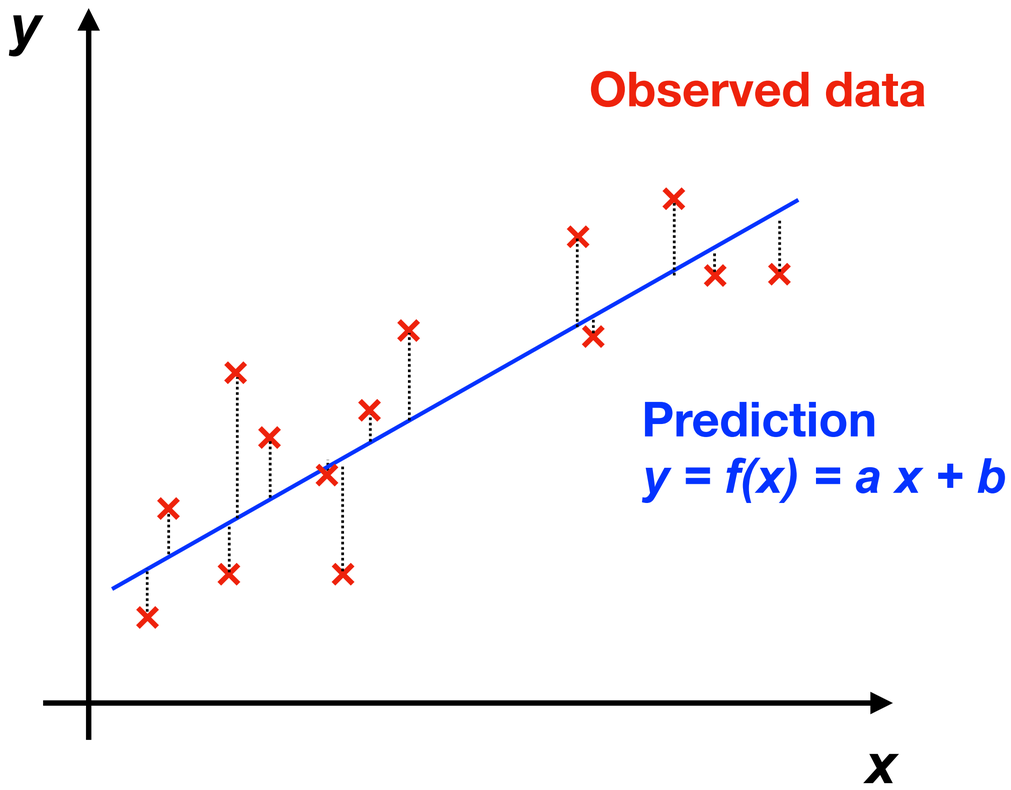
例：観測量 x = color から物理パラメータ y = redshift を推定したい

→ Regression task（回帰問題）

Steps:

1. モデルを構築（例：y = ax + b）

2. モデルのパラメータ（例：a, b）を動かして観測データ（x, y）との差が最も小さくなるパラメータを探す（最適化/Optimization）



最適化（Optimization）

例）線形モデルの場合：

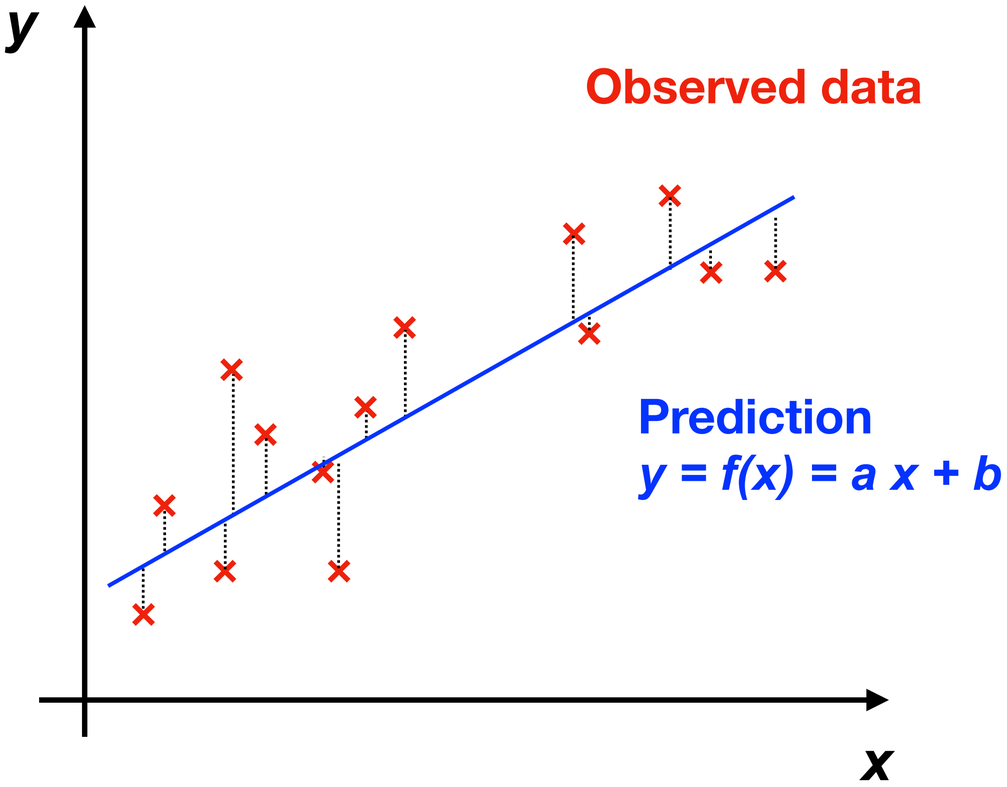
(a,b) の関数（残差） $L(a,b) = \sum_i (ax_i + b - y_i)^2$ が極小になるところを探せば良いので

$$\frac{\partial L}{\partial a} = \sum_i 2(ax_i + b - y_i)x_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_i 2(ax_i + b - y_i) = 0$$

これらを (a, b) について解けば良い。

$$a = \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} \qquad b = \frac{1}{N}\left(\sum_i y_i - a\sum_i x_i\right)$$



ここまでは、機械学習に限らず、「経験則に基づいた」パラメータ推定に共通の話。

より複雑なモデルを作るには？

入力
例：測光データ

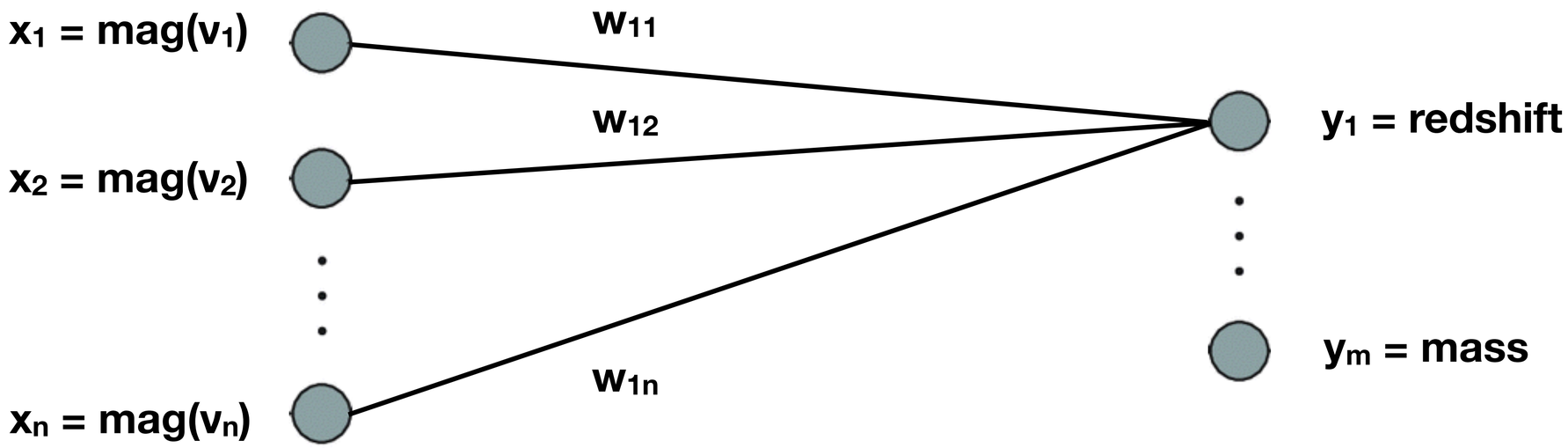
出力
例：銀河のパラメータ

より複雑なモデルを作るには？

入力
例：測光データ

重み

出力
例：銀河のパラメータ



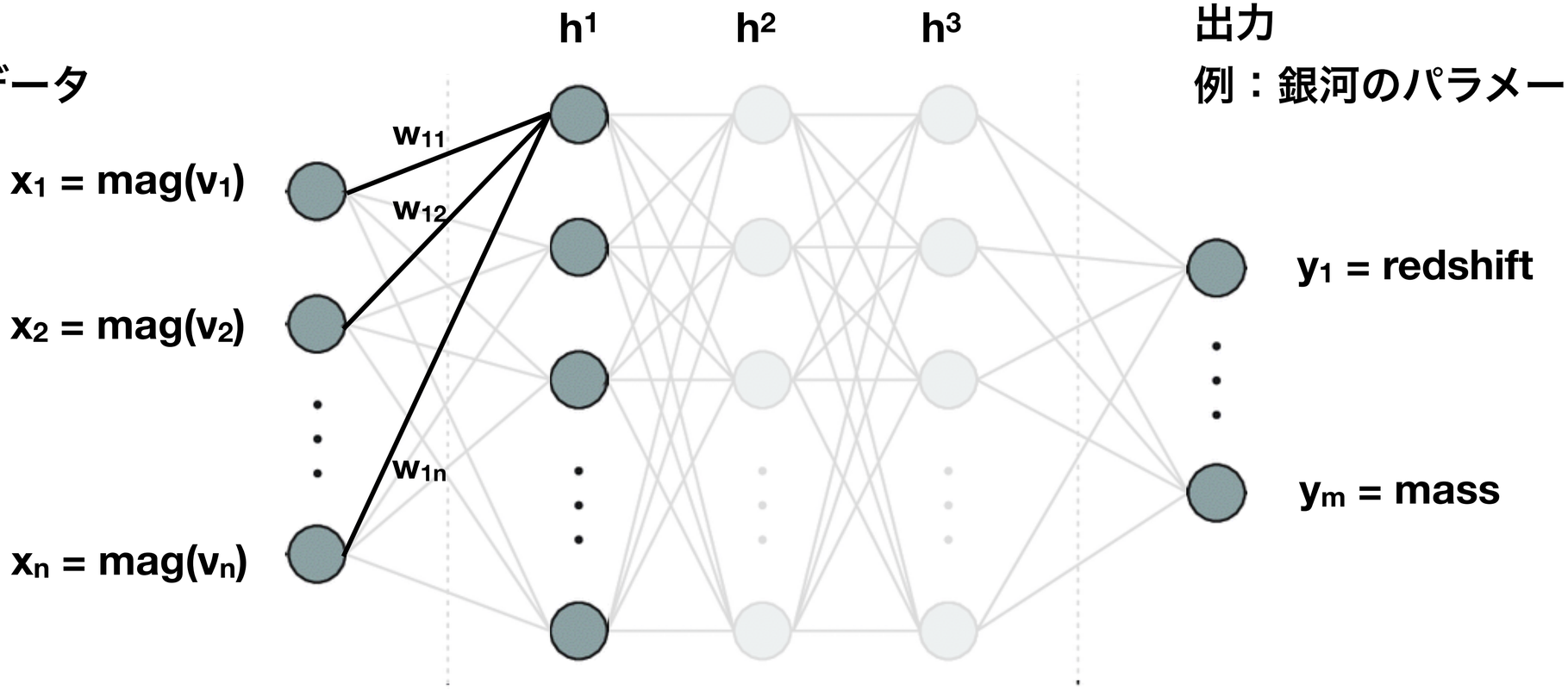
最も単純なモデル：線形モデル

$$y_i = w_{i1}x_1 + w_{i2}x_2 + \ldots + w_{in}x_n + b_i$$

より複雑なモデルを作るには？

入力

例：測光データ



出力

例：銀河のパラメータ

$$h_i^1 = a(w_{i1}^0 x_1 + \ldots + w_{in}^0 x_n + b_i^0)$$

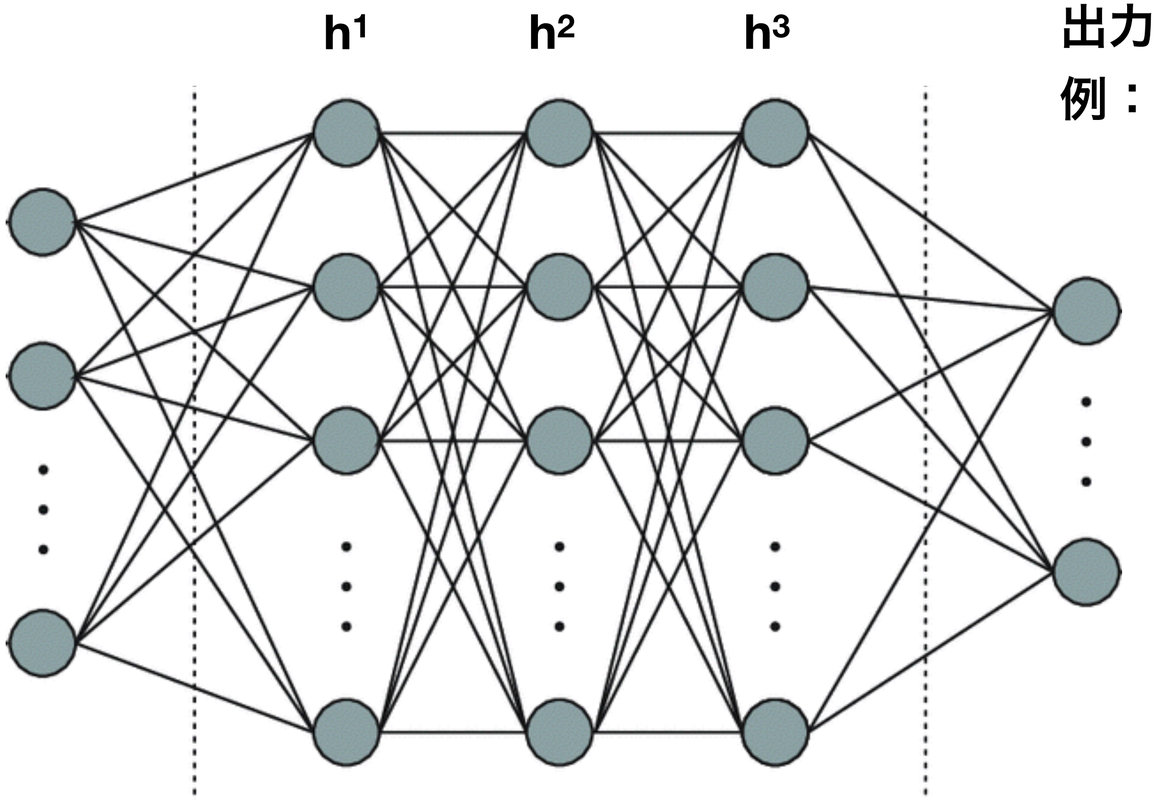
より複雑なモデルを作るには？

入力
例：測光データ

$x_1 = \mathrm{mag}(v_1)$

$x_2 = \mathrm{mag}(v_2)$

$x_n = \mathrm{mag}(v_n)$



出力
例：銀河のパラメータ

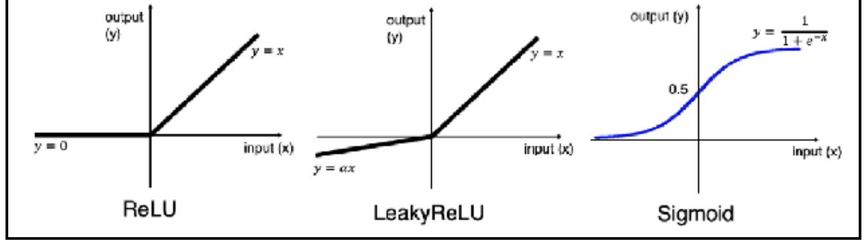
$y_1 = \mathrm{redshift}$

$y_m = \mathrm{mass}$

線型変換+非線形変換を繰り返すことで複雑なモデルを構築
→ （深層）ニューラルネットワーク

$$h_i^1 = a(w_{i1}^0 x_1 + \ldots + w_{in}^0 x_n + b_i^0)$$
$$h_i^2 = a(w_{i1}^1 h_1^1 + \ldots + w_{in}^1 h_n^1 + b_i^1)$$
$$h_i^3 = a(w_{i1}^2 h_1^2 + \ldots + w_{in}^2 h_n^2 + b_i^2)$$
$$y_i = a(w_{i1}^3 h_1^3 + \ldots + w_{in}^3 h_n^3 + b_i^3)$$

activation function（非線形変換）



最適化（Optimization）

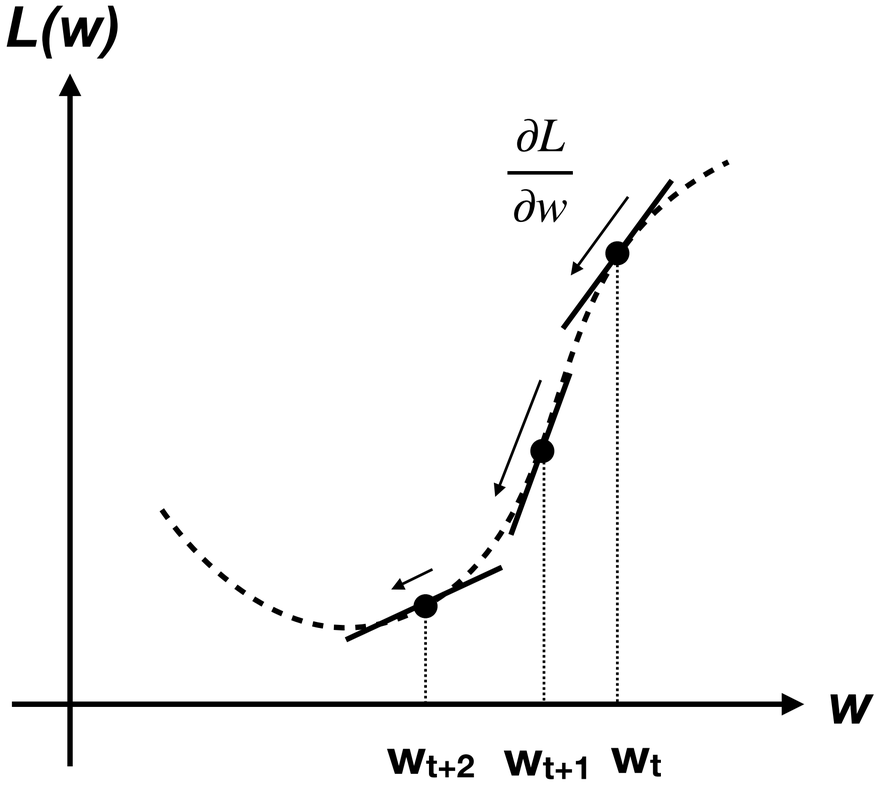
- モデル　$y = f^w(x)$　（w: モデルパラメータ）
- 残差　$L(w) = \sum_i (y_i - f^w(x_i))$

関数（モデル）が複雑な時は数値的に最適化する

例：導関数（勾配）を使った最適化方法：
とある点 w_t 付近で微小変動させた時にどちらの方向
動けば残差が小さくなるかを計算する

坂を「くだる」方向にパラメータを少しずつ動かしていけば良い

$$w_{t+1} = w_t - \alpha \frac{\partial L}{\partial w}$$

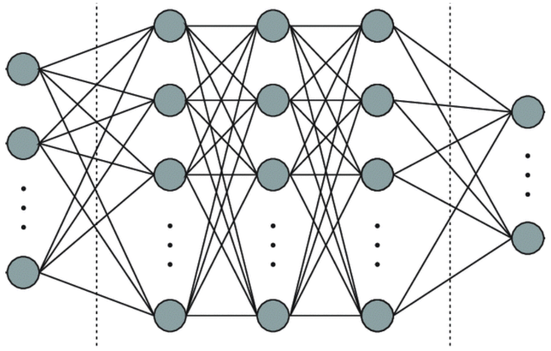


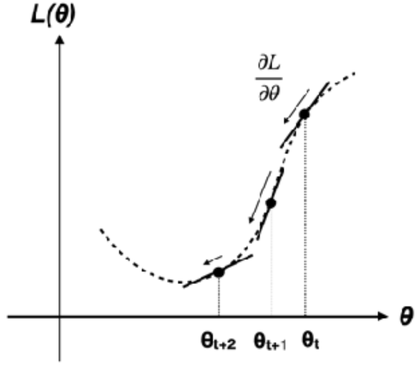
機械学習とは

機械学習とは「学習データをもとに**自動的に**モデルを構築するプロセス」

特に近年注目されている深層学習では**大量のモデルパラメータ**を自動的に最適化する

→ これによって複雑な課題もこなすことができる





さまざまな機械学習モデル

- **Principal component analysis (PCA; ~1980s-)**
- **Artificial neural network (ANN; ~1990s-)**
- **Decision tree (DT; ~1990s-) — cf. Random Forest (RF)**
- **Support vector machine (SVM; ~2000s-)**

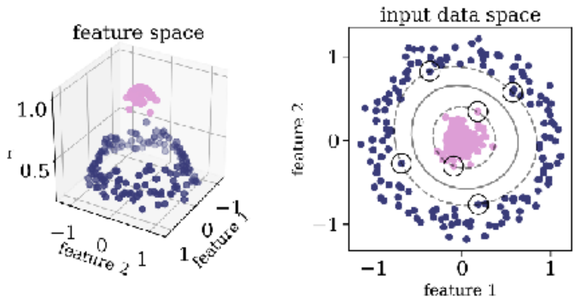
天文学では比較的早期から機械学習の手法が取り入れられてきた

（Baron 2019 for review）

さらに近年ではより高コストなモデルも

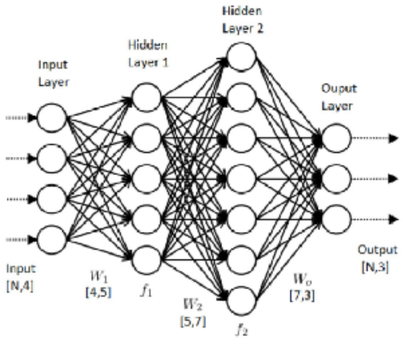
- **Convolutional neural network (CNN)**
- **Recurrent neural network (RNN)**
- **Graph neural network (GNN)**
- **Transformer**

Support vector machine

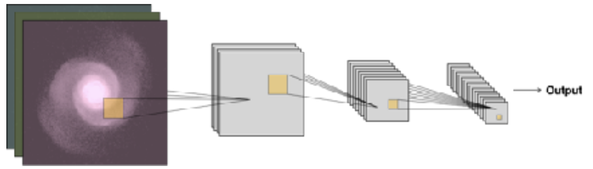


Decision tree

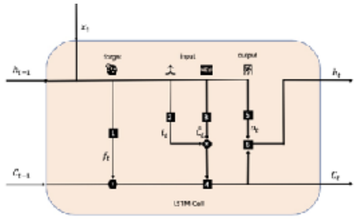
Artificial neural network



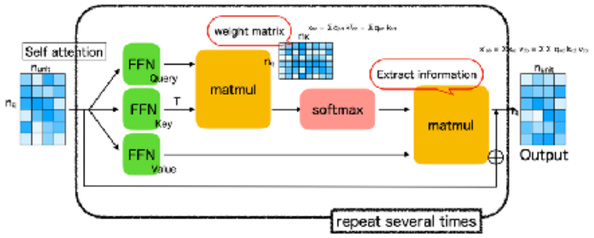
CNN



RNN



Transformer

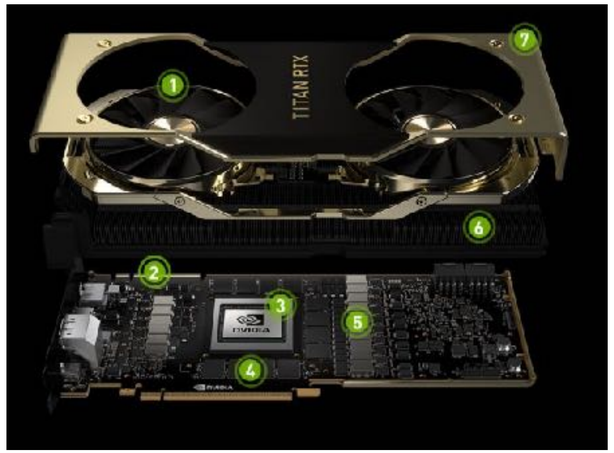


Figures from Baron (2019)

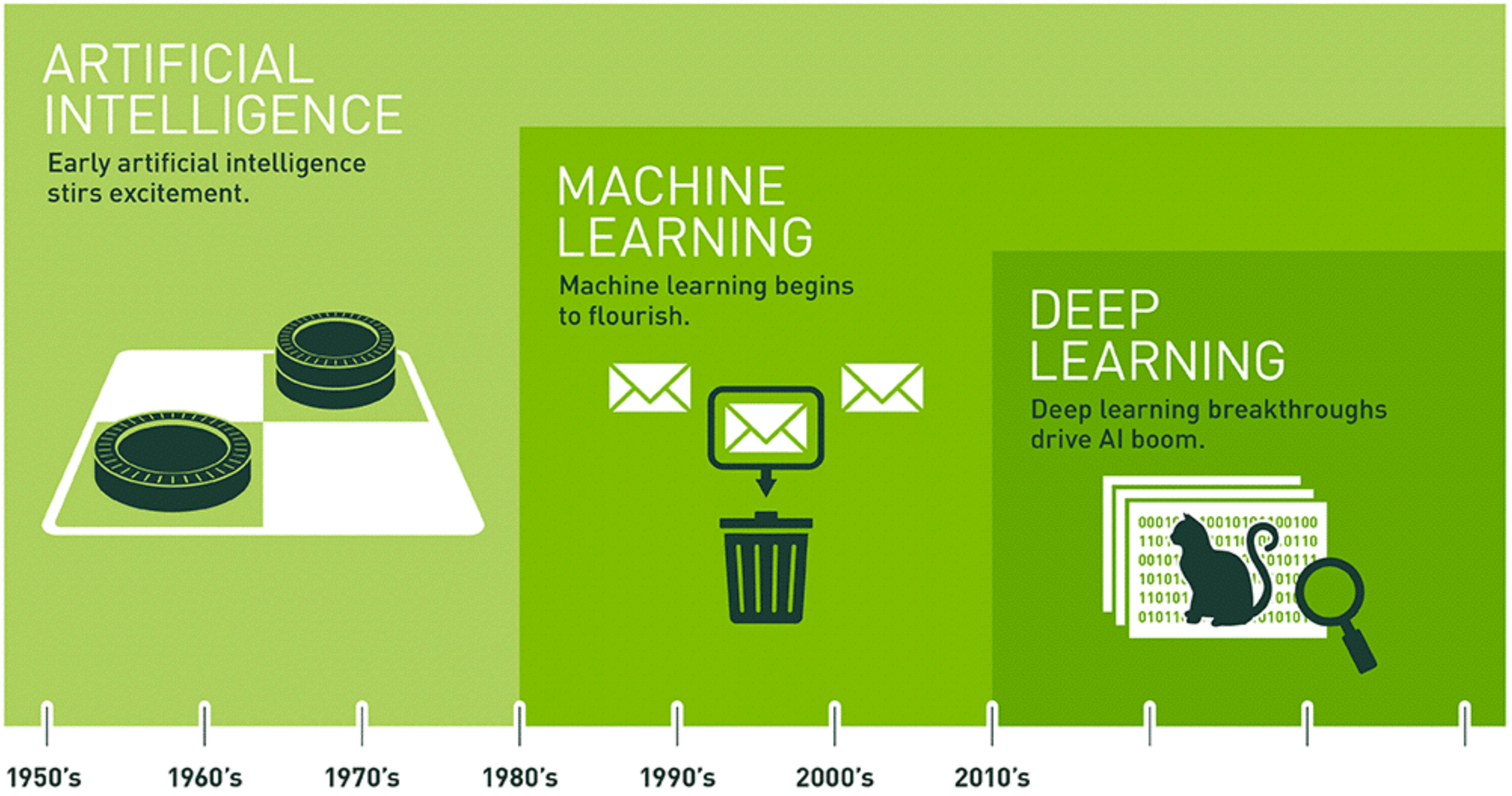
さまざまな機械学習モデル

GPU の登場後、より複雑なモデルの学習（深層学習）が可能に

→ ここ数年のAIブームの要因の一つ



GPU (NVIDIA TITAN RTX)

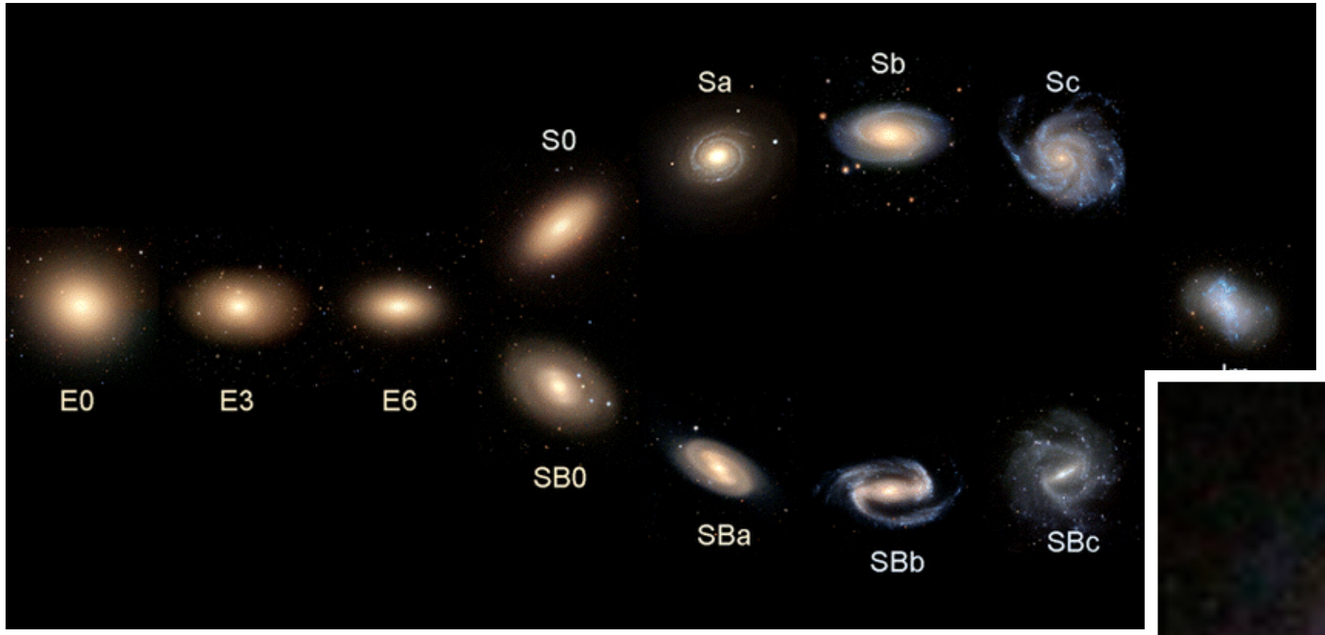


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Image from [NVIDIA “What’s the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?”]

観測データへの適用例
1. 分類（Classification）

銀河の形態分類



渦巻き？
楕円？
合体？

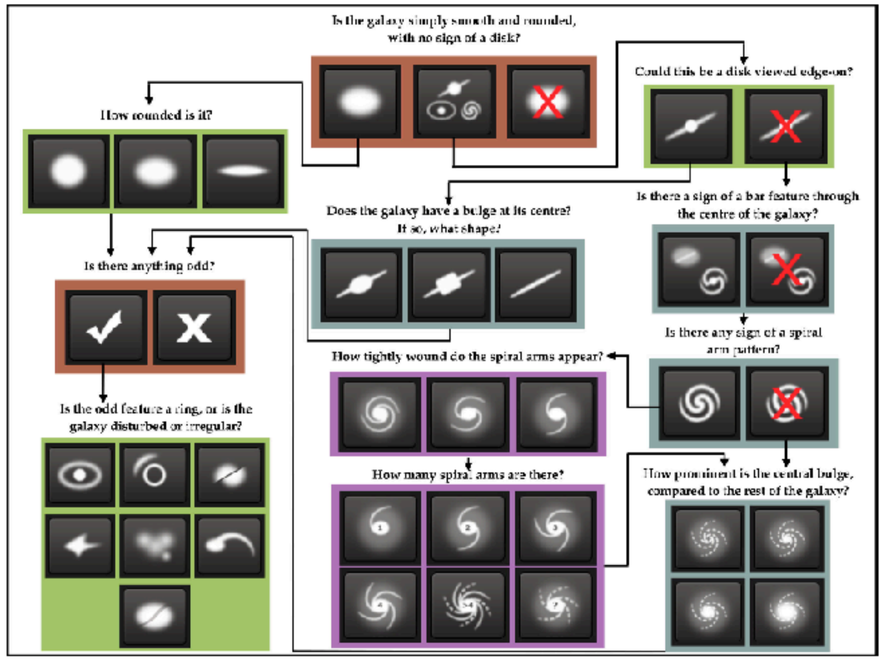
銀河の形態は銀河形成と密接に関連している

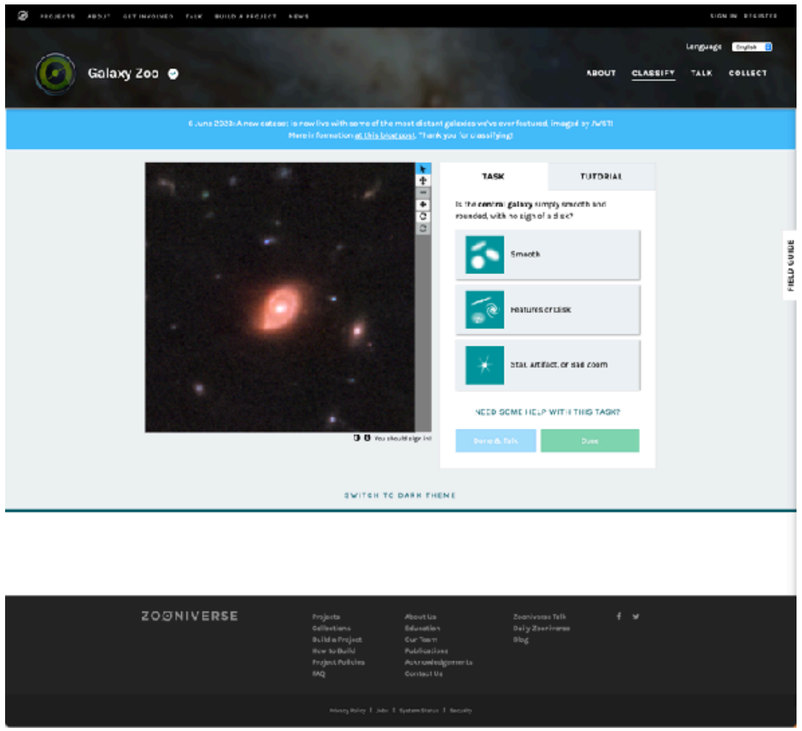
将来観測で得られる膨大な画像データを分類することで統計的な議論が可能となる

銀河の形態分類

これまでは人の目（含 Citizen science）や non-parametric method で主に分類されてきたが、データが多くなるとこうした手法を用いるのはもはや不可能

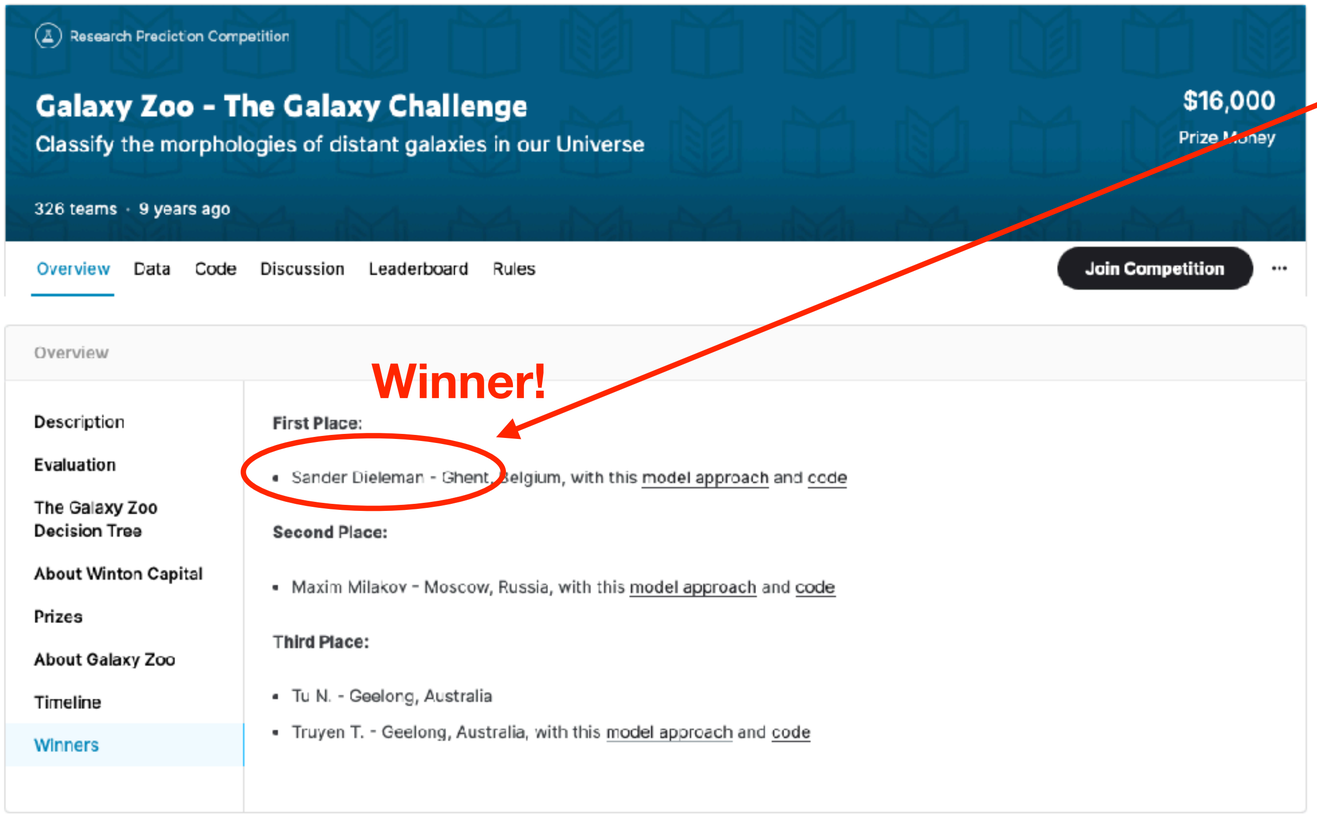
Galaxy Zoo (Willett+13): The images are classified by many (amateur) volunteers





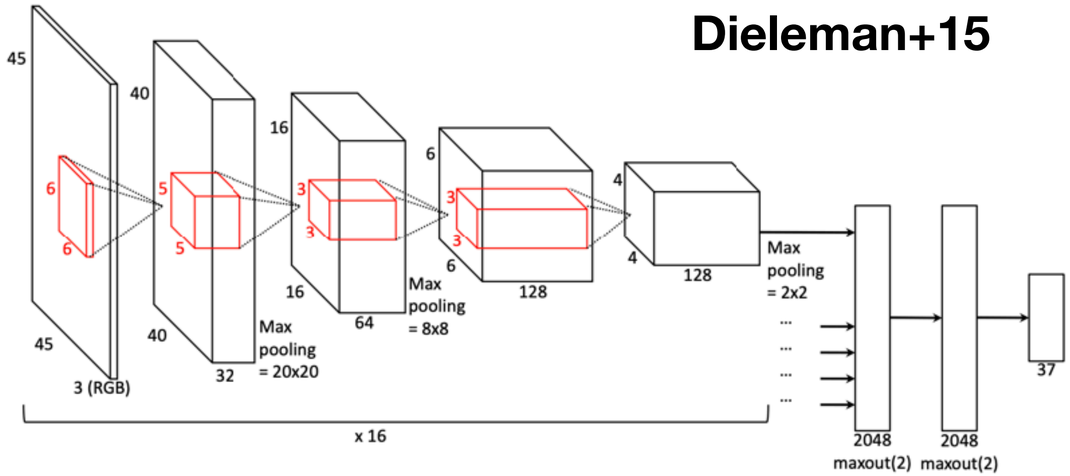
銀河の形態分類

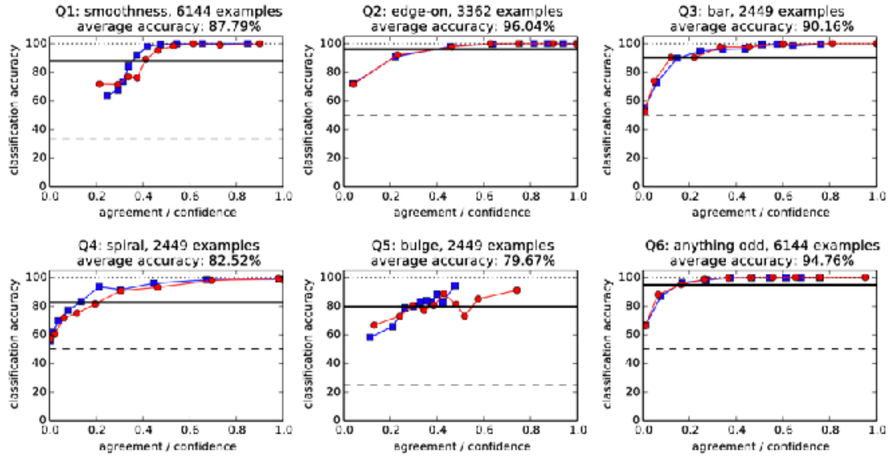
形態分類では早くから機械学習の手法が取り入れられてきた。
さらに、近年では畳み込みを用いた深層学習により高い精度での分類が可能となってきている



Morphological classification challenge on the Kaggle platform

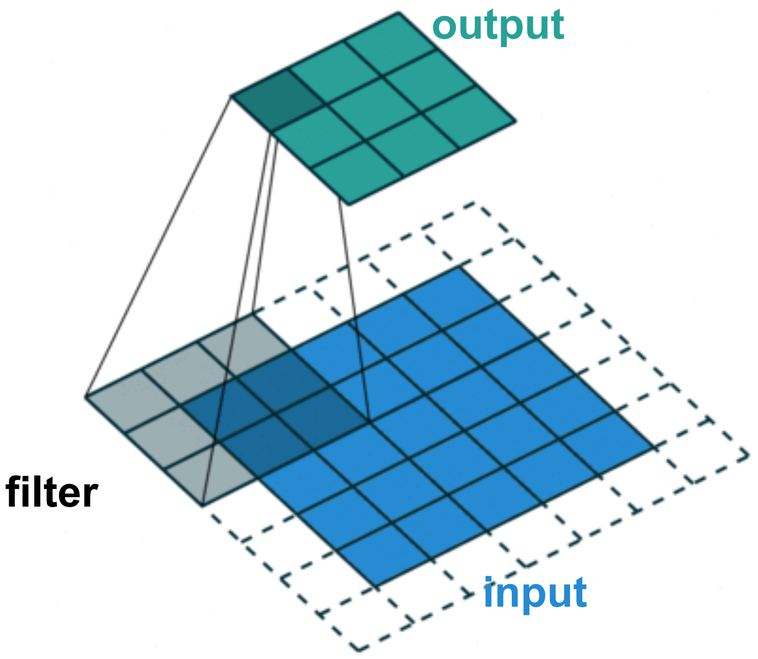
Dieleman+15

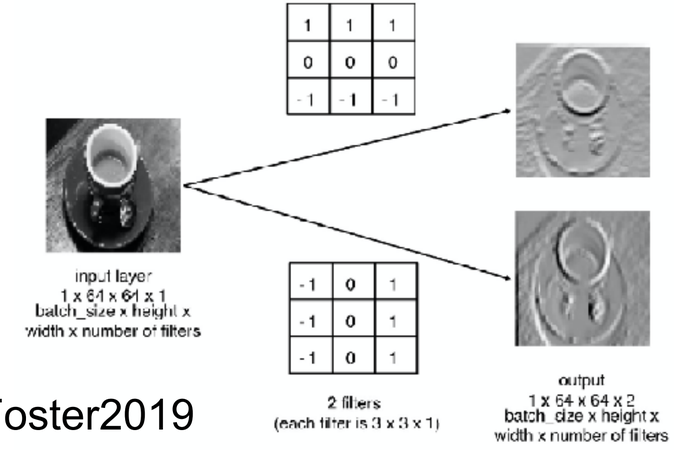




The first usage of CNN in astronomy

畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）





Foster2019

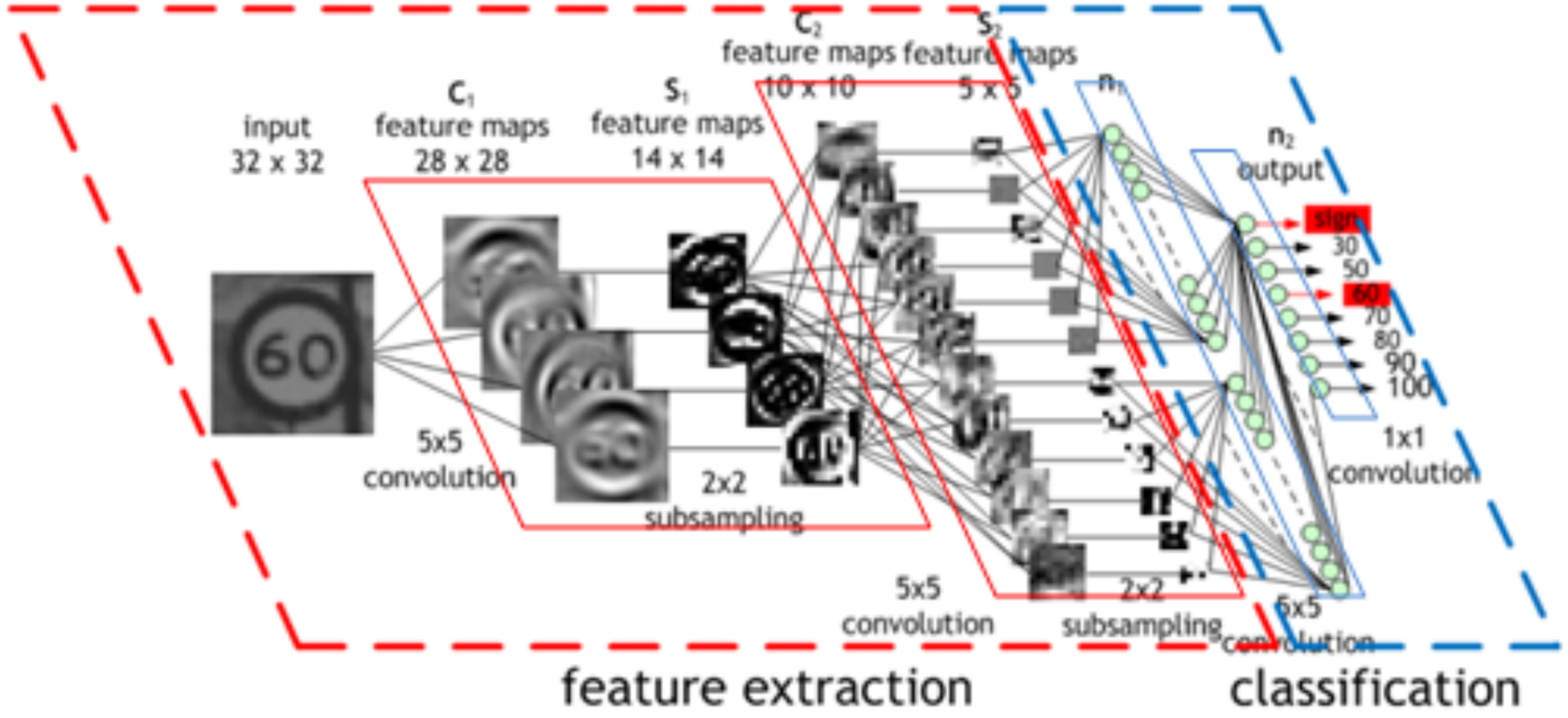
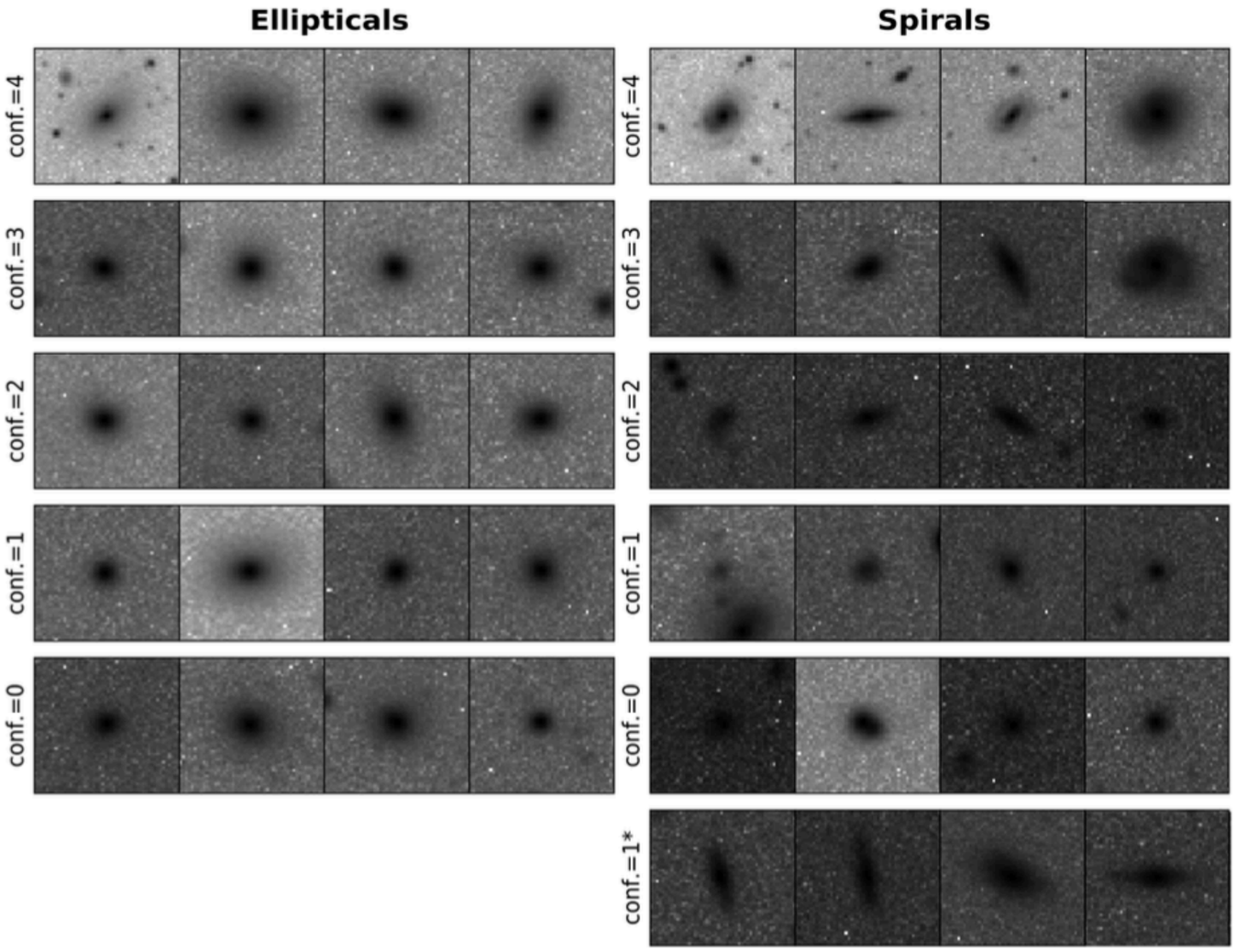


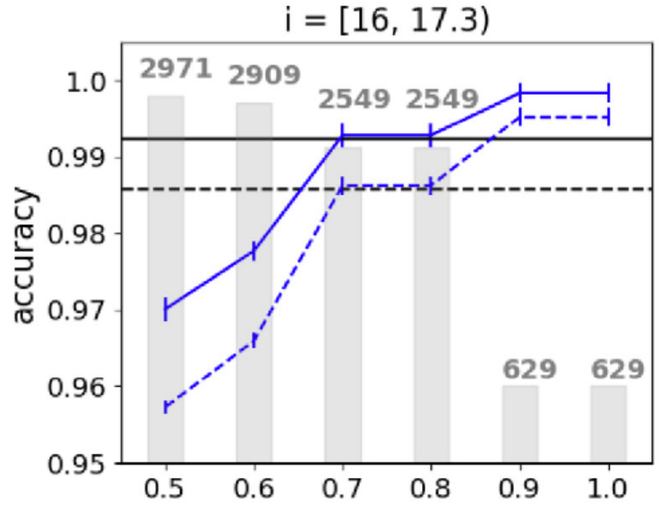
Image from https://developer.nvidia.com/discover/convolutional-neural-network

銀河の形態分類



Cheng+2021

- Galaxy Zoo の2862個の銀河を学習データとして用い、 DES で得られた二千万個もの銀河の形態カタログを作成
- 99% 以上の精度での分類に成功



これを仮に従来の方法で行った場合、少なくとも100年はかかる見積もり

観測データへの適用例
2. 回帰（Regression）

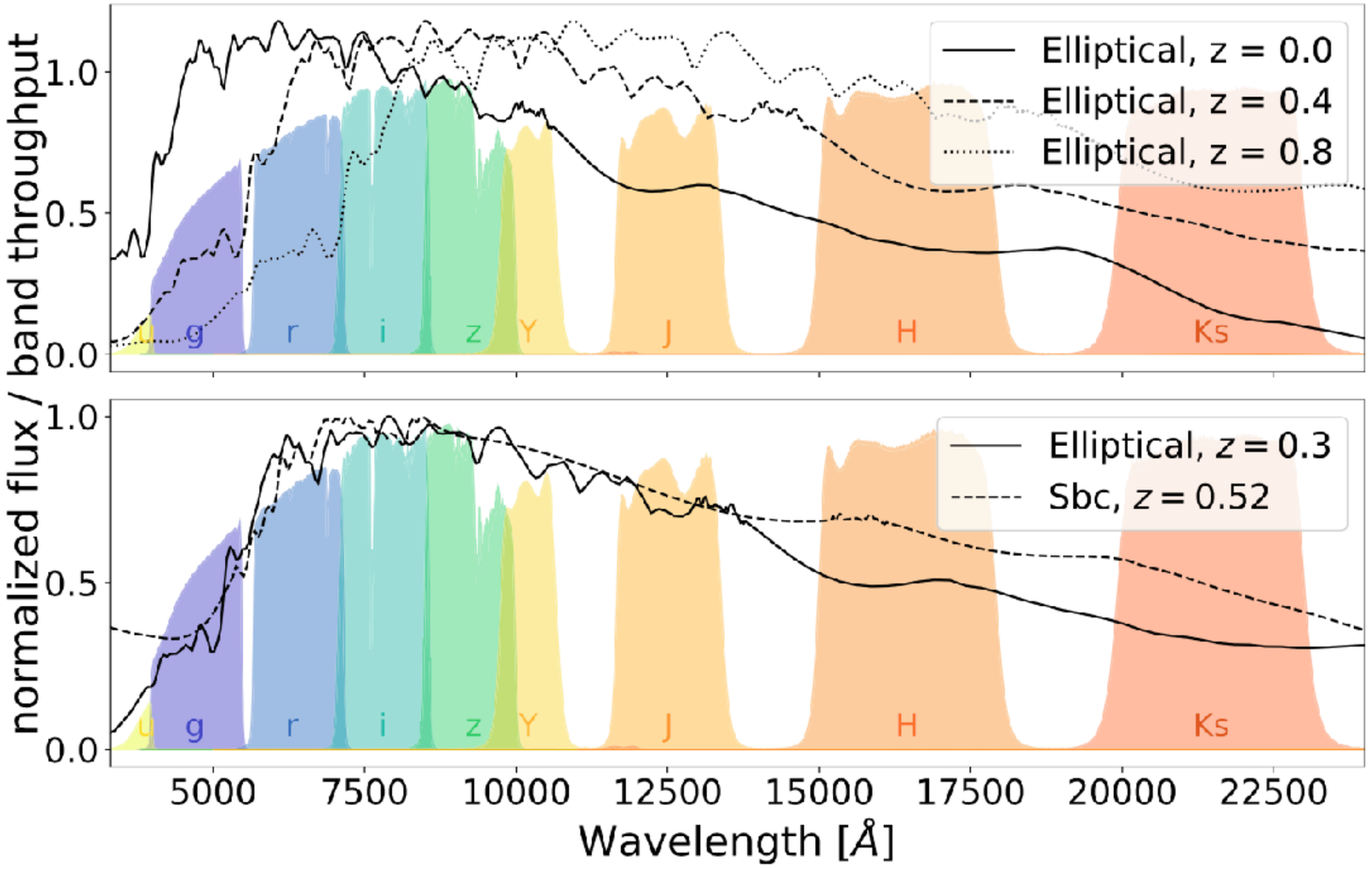
銀河の photometric redshift 推定

分光観測は非常にコストが高いため、測光観測データのみから赤方偏移を推定したい

（Cf. Photo-z 推定手法に関する最近のレビュー：Brescia et al. 2021, Newman & Gruen 2022）

従来の手法： SED テンプレートを用いたフィッティング（e.g., BPZ; Benítez 2000）

テンプレートは Stellar Synthesis Model か観測データを利用



Buchs+ (DES Collaboration) 2019

銀河の photometric redshift 推定

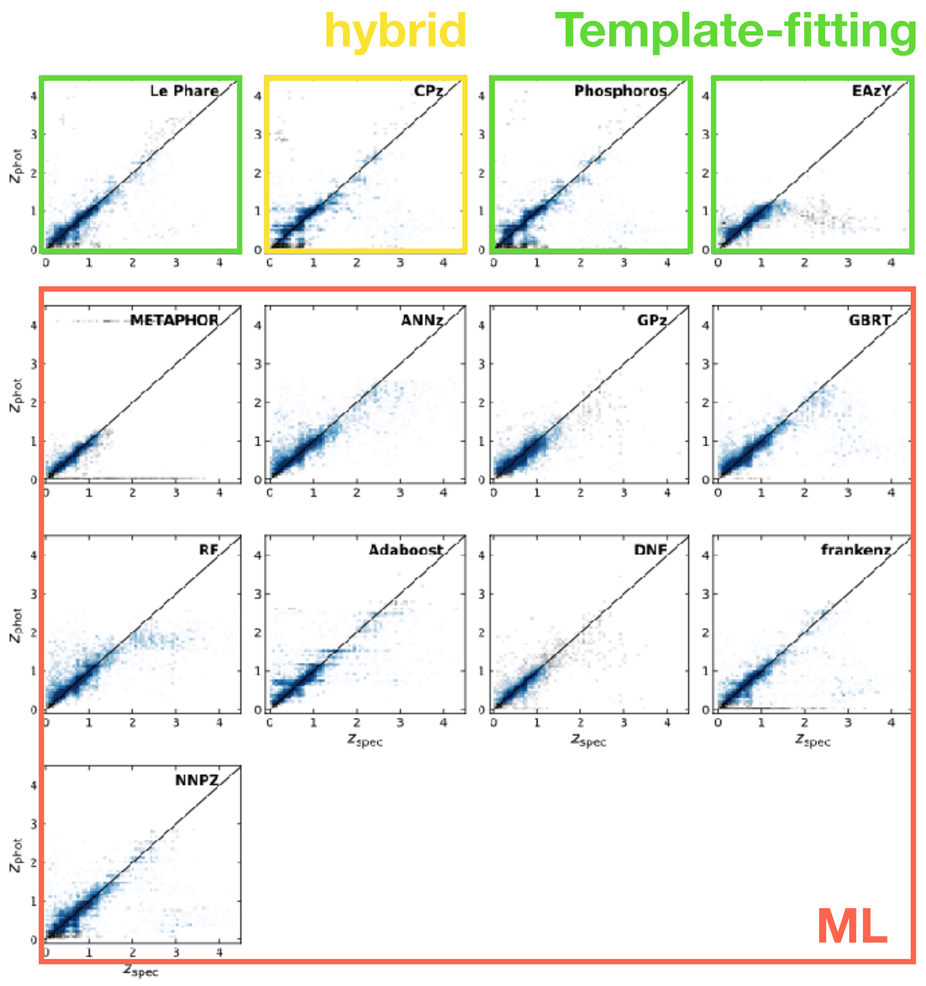
比較的早期から機械学習の手法も取り入れられてきた
（e.g., ANNz; Collister & Lahav 2004）。
分光観測で赤方偏移がわかっている銀河を学習に用いる

今のところ従来の手法と同程度の精度（Cavuoti+2012,
Henghes+2021, Tanaka+2018 (HSC), Schmidt+2020 (LSST)
Desprez+2020 (Euclid)）

それぞれの手法に長所・短所がある

Pros: 観測データを学習に用いることでより現実的なスペクトル・ノイズモデルを構築できる。テンプレートフィッティングほどは prior に依存しない

Cons: 学習データとターゲットデータの分布が異なる場合、バイアスが生じうる。学習データに内在するエラーを組み込みにくい



Desprez+2020

銀河の photo-z 推定

従来の手法と機械学習を組み合わせて、すでにさまざまなサーベイの photo-z カタログが作られている

- ~40 million KiDS galaxies (deJong+17, Bilicki+18)
- ~100 million HSC galaxies (Tanaka+18, Schuldt+21)
- 1 billion DESI galaxies (Duncan+22)
- 3 billion Pan-STARRS1 galaxies (Beck+21)

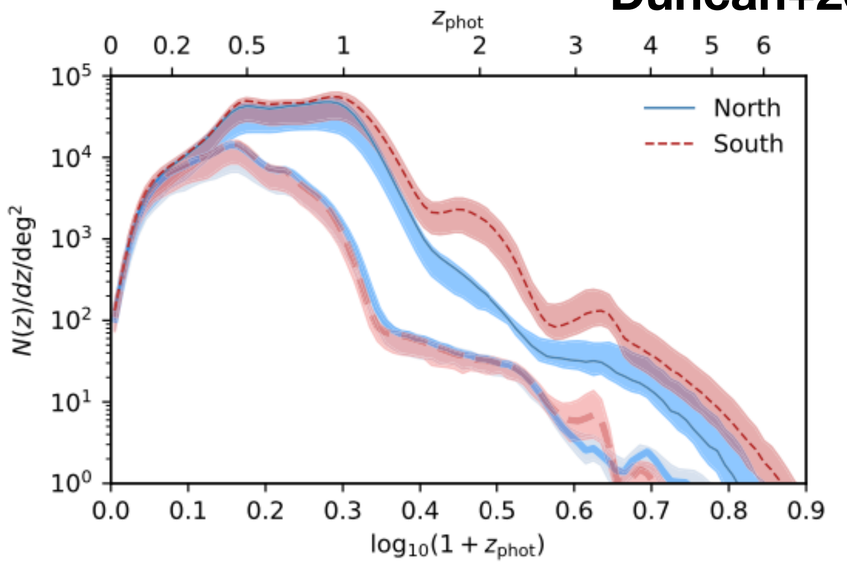
さらに、将来のサーベイプロジェクトに向けた準備もなされている

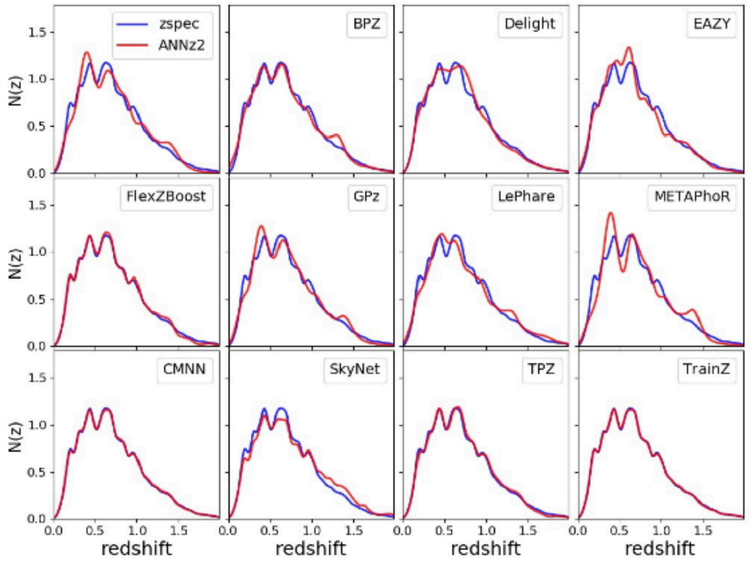
- LSST (Schmidt+20)
- Euclid (Desprez+20)

必要な精度：

- $\sigma_z < 0.02\,(1+z)$
- 3σ outlier rate < 10 %
- 多くの場合、統計的な再現性（赤方偏移PDFの精度）が重要となる

Duncan+2022



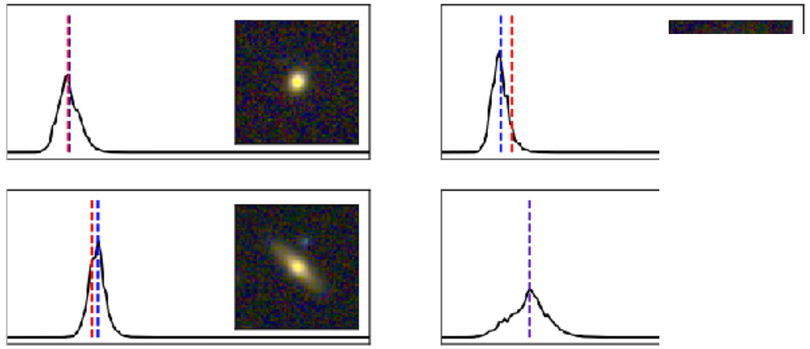


Schmidt+2020

銀河の photo-z 推定：今後の課題

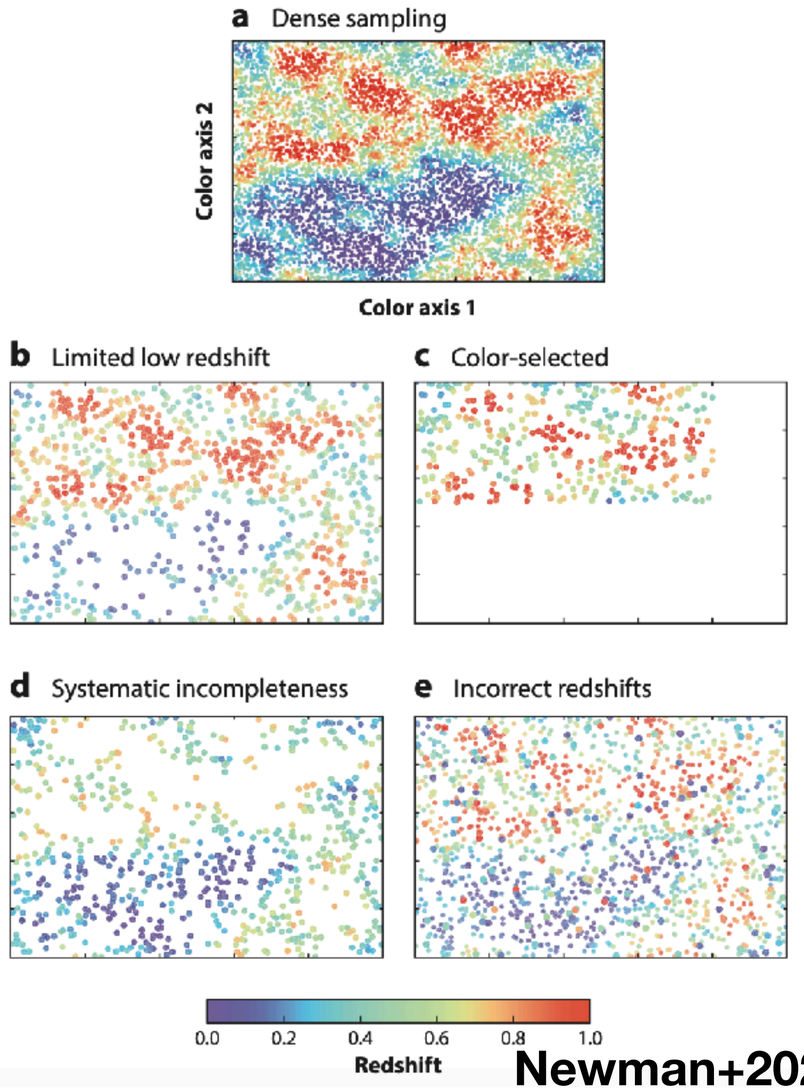
今後の課題

- より精度を上げるには？
  → 画像も機械に学ばせる（表面輝度、
  サイズ、形態、傾き、etc.）



Pasquet+2019

- 学習データの不足
  → 分光サンプルを増やす必要がある。
  どの手法も、典型的には数万個程度のデータが必要。
  （ただし画像を用いる場合はこれの10倍以上）
  学習データの中のバイアスを減らすように増やす必要がある。
  複数の離れた領域で分光観測をする（Newman+15）などして、
  field variance の影響を減らすことも必要

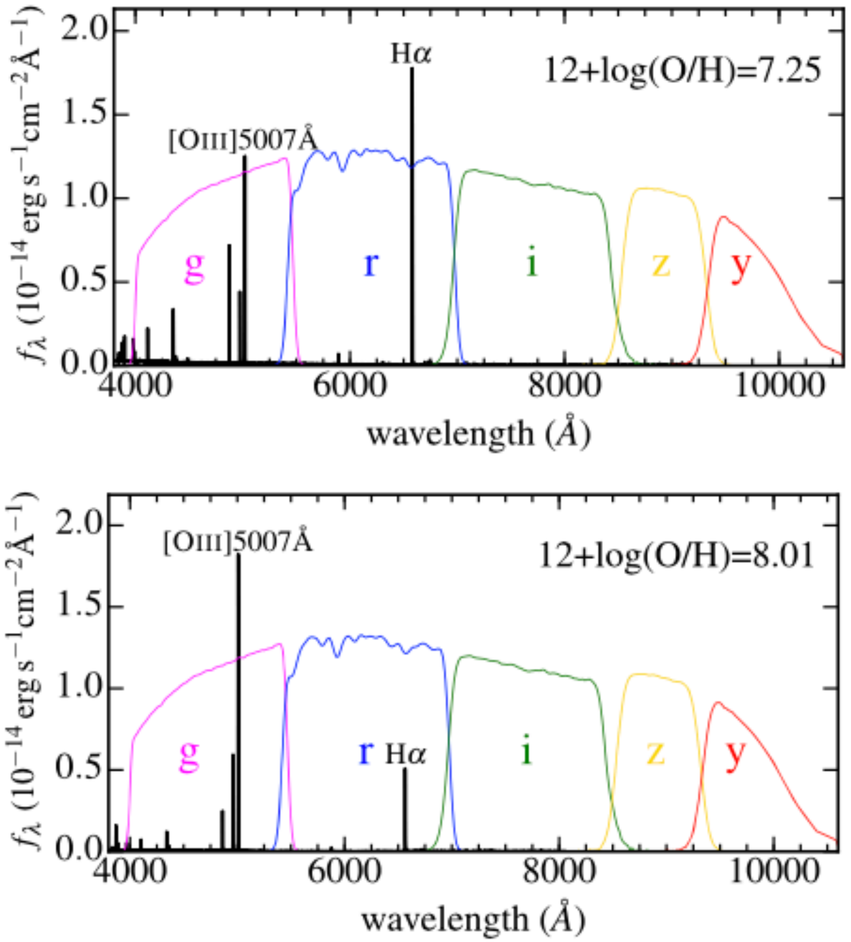


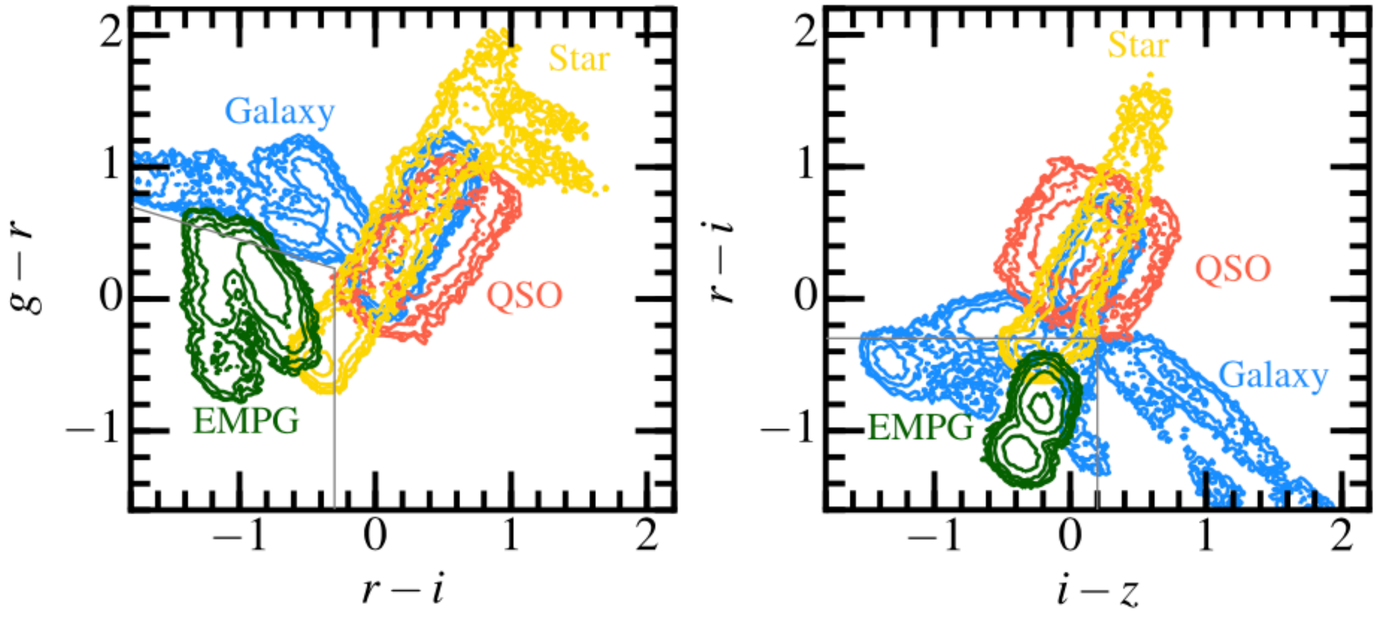
Newman+2022

観測データへの適用例
3. 発見（Discovery）

超低金属量銀河の発見

測光観測カタログから超低金属量銀河（Extremely Metal-Poor Galaxy; EMPG）を見つけたい

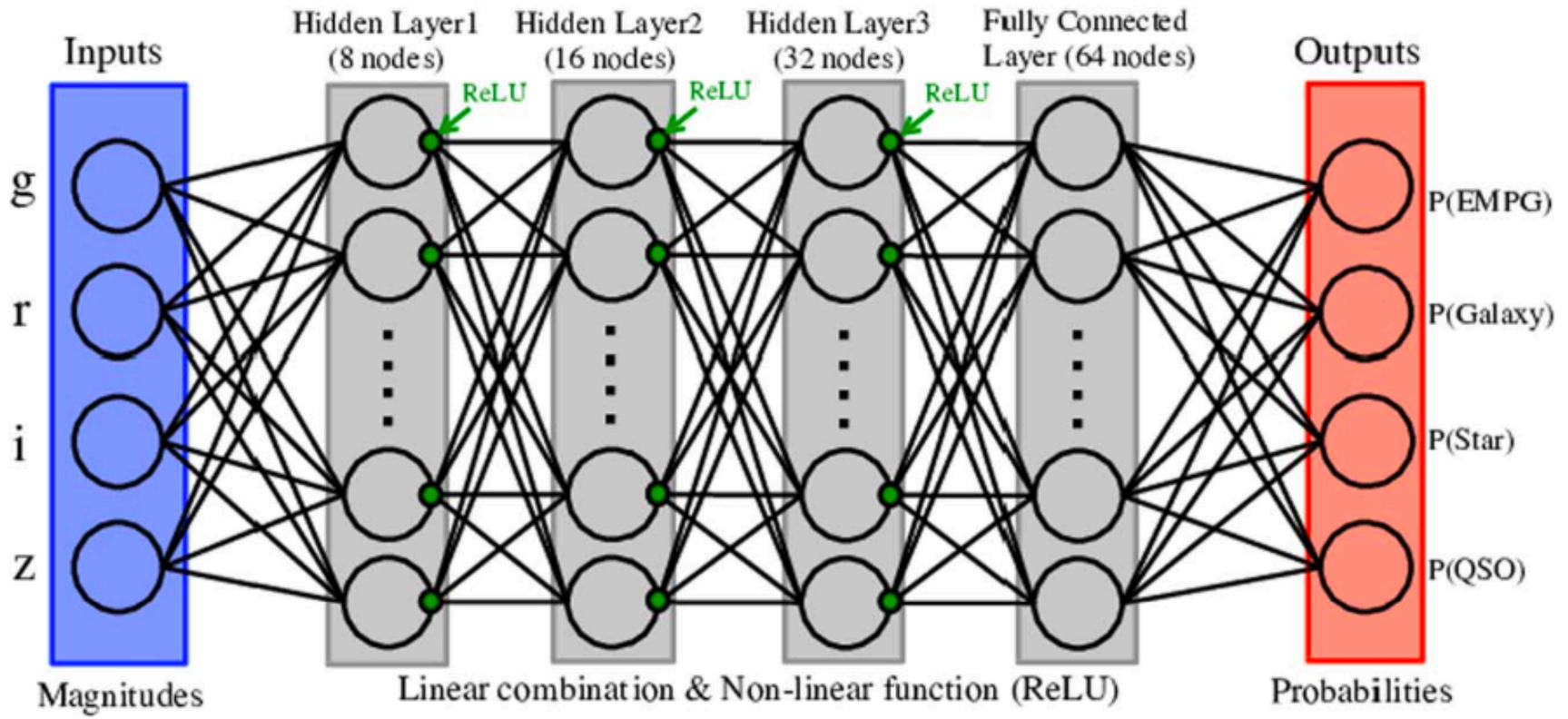




Kojima+2020

超低金属量銀河の発見

測光観測カタログから超低金属量銀河（Extremely Metal-Poor Galaxy; EMPG）を見つけたい



Kojima+2020

超低金属量銀河の発見

分類の評価尺度

本当はEMPGなのに
検出できなかった

Actual class \ Predicted class	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

間違ってEMPGと判定してしまった

accuracy (精度・正解率)

$$\frac{TP + TN}{TP + FP + TN + FN}$$

全体的なモデルの
「良さ」

precision / purity (適合度)

$$\frac{TP}{TP + FP}$$

間違った検出を
少なくしたい時

recall / sensitivity / completeness (再現度/完全性)

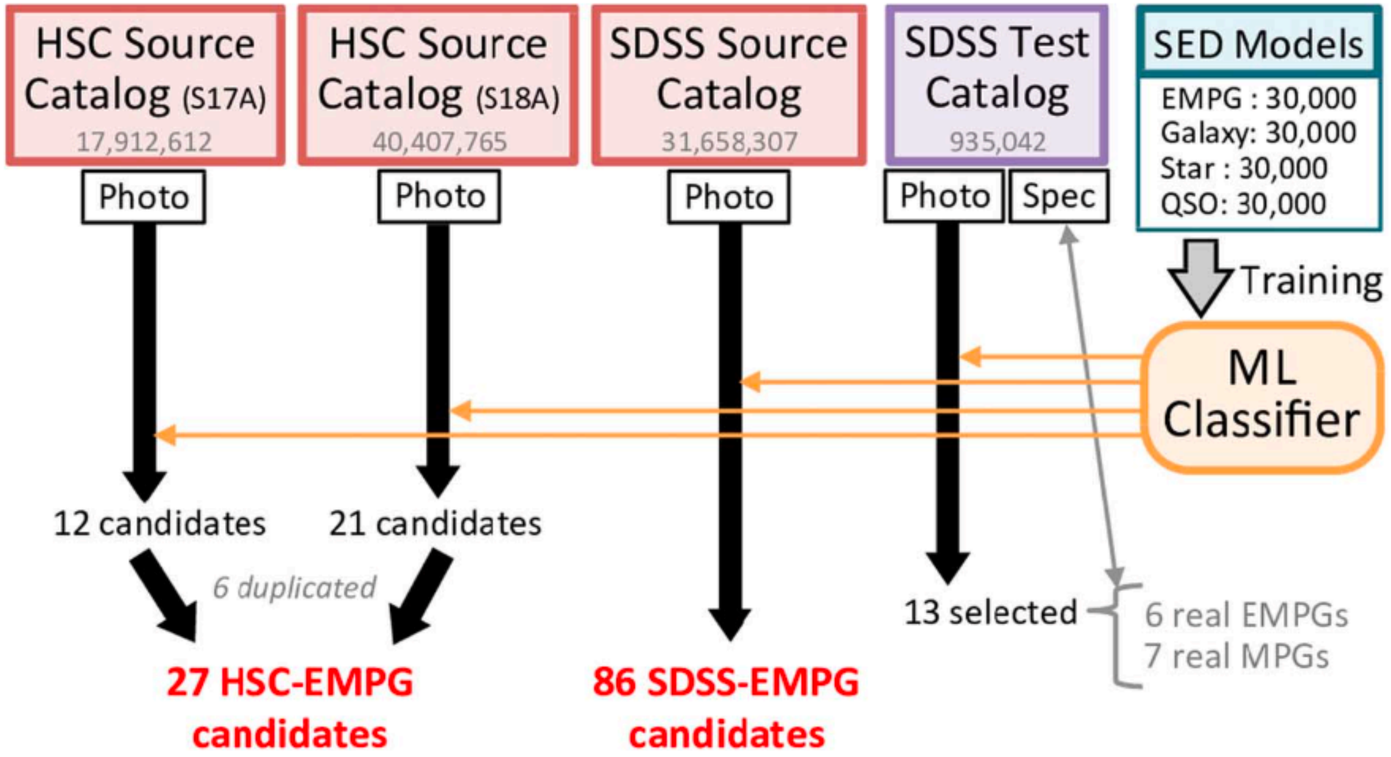
$$\frac{TP}{TP + FN}$$

取りこぼしを
減らしたい時

超低金属量銀河の発見

そもそも超レアな天体を探している

→ ある程度間違ってもいいのでできるだけ取りこぼさないようにしたい



SDSS でのテスト：

935,042 sources in total
only 7 among them are EMPG

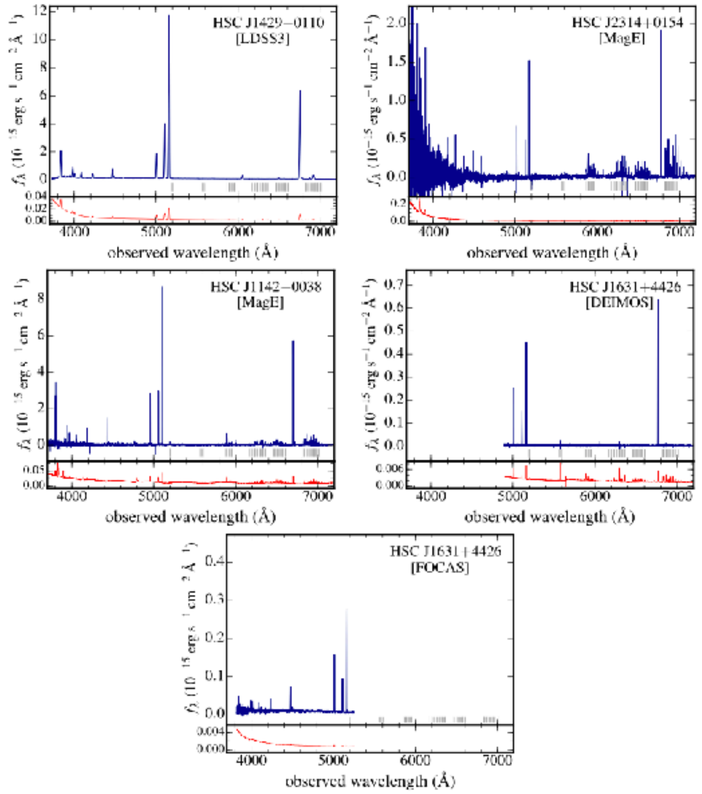
13 EMPG candidates are ditected
6 among them are indeed EMPG

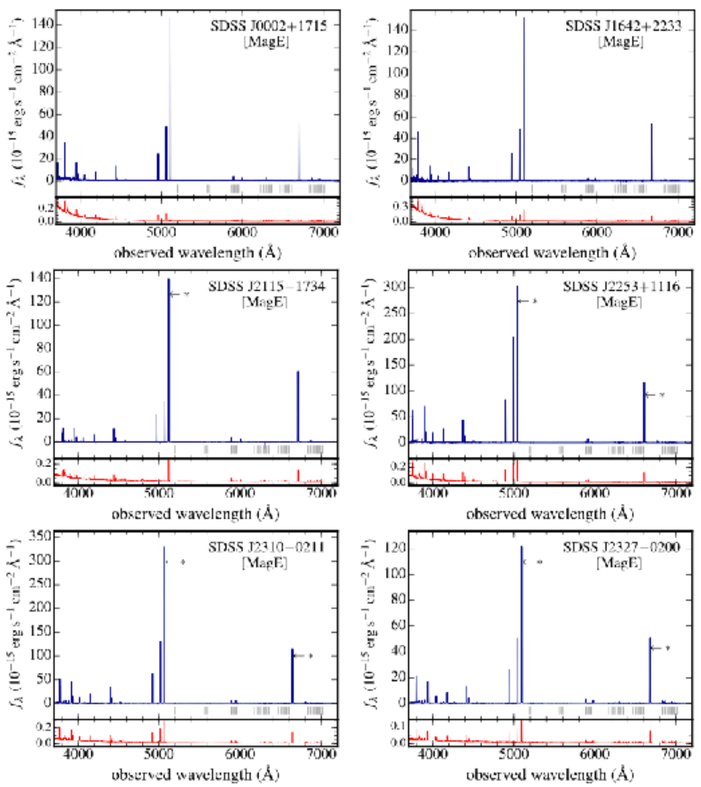
→ The purity is 54%, but completeness is as high as 86%

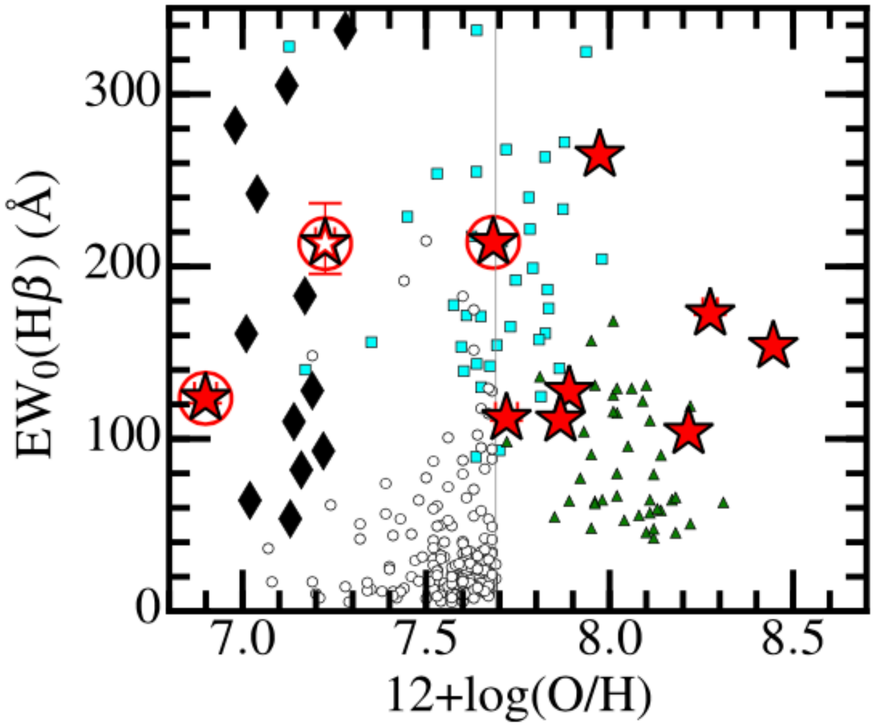
Kojima+2020

超低金属量銀河の発見

候補天体を絞ることができたので、追観測としてより高コストな分光観測もできる



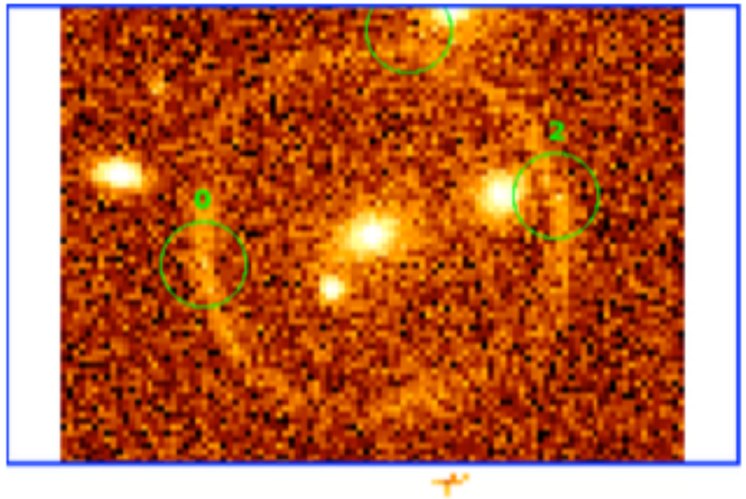




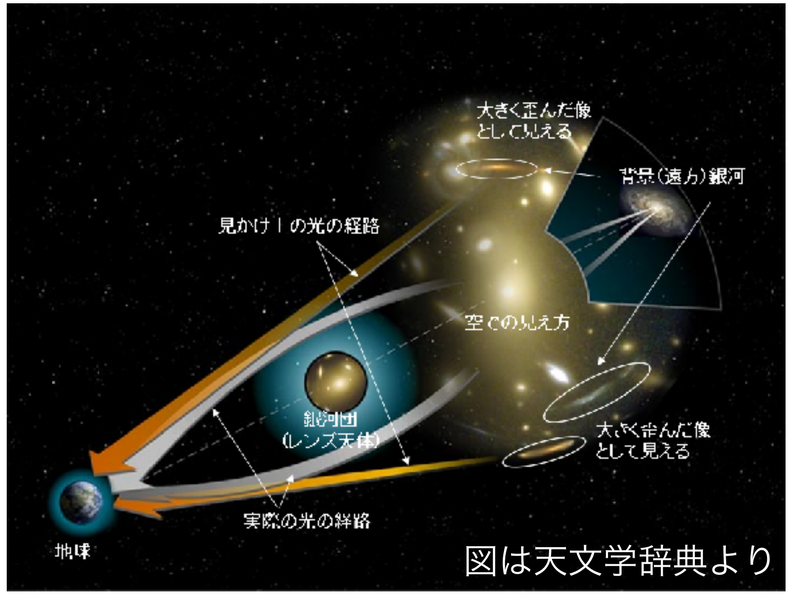
Kojima+2020

10個中3個が実際に EMPG であることを確認できた

強重力レンズ検出



→ Is there a gravitational lens?



図は天文学辞典より

これまで見つかった強重力レンズ天体は < 1000 個程度。

SKA, LSST, Euclid などでは数十〜百億天体のうち数十万個程度検出可能なレンズがあると予測される

これまでの最も主力な方法は目視による確認

SPACE WARPS project (Marshall+16, More+16, Geach+15)

最近では自動アルゴリズムや機械学習を用いた方法も使われ始めている（e.g., Paraficz+16）

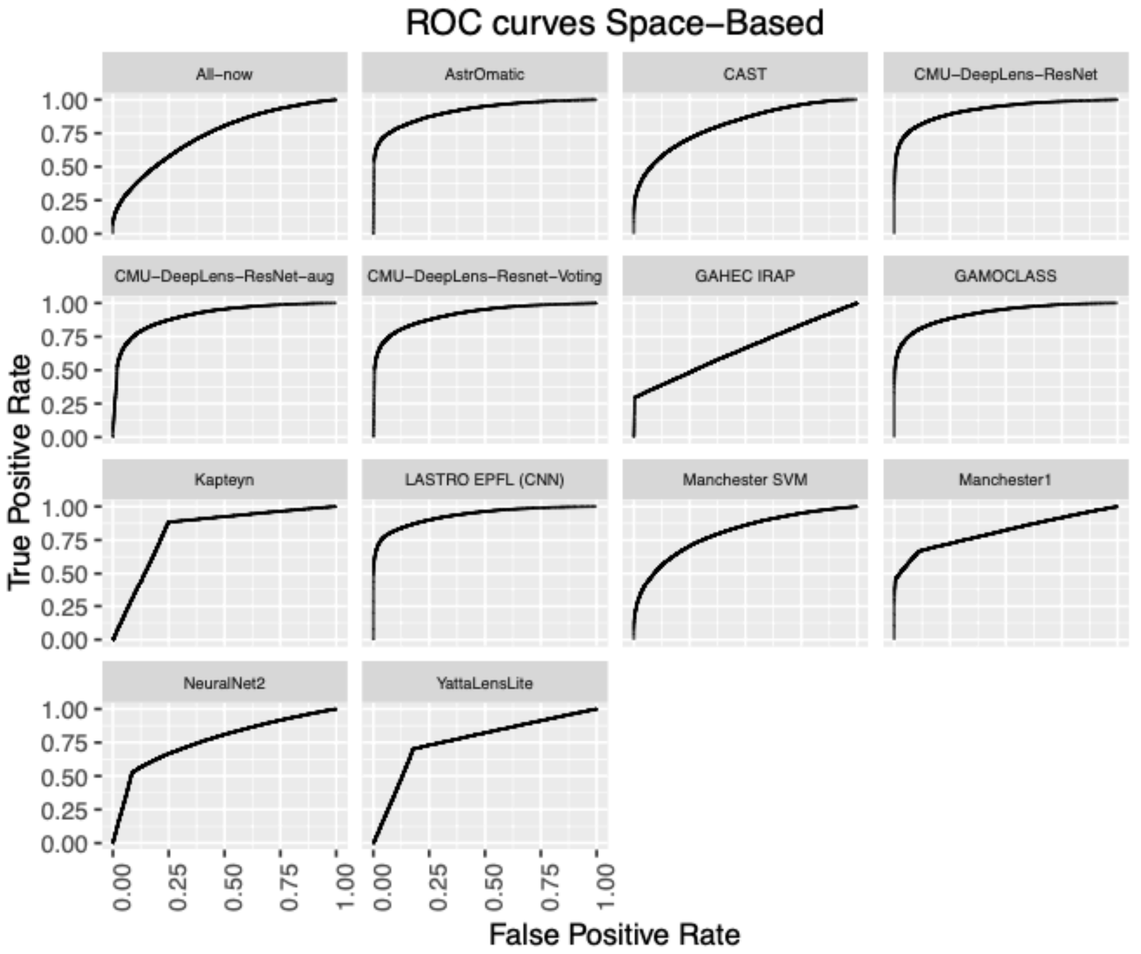
強重力レンズ検出

“The strong gravitational lens finding challenge” (Metcalf+19)

Area under ROC これが大きいほど良い

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

Receiver operation characteristics (ROC)



観測データへの適用例
4. 生成（Generation）

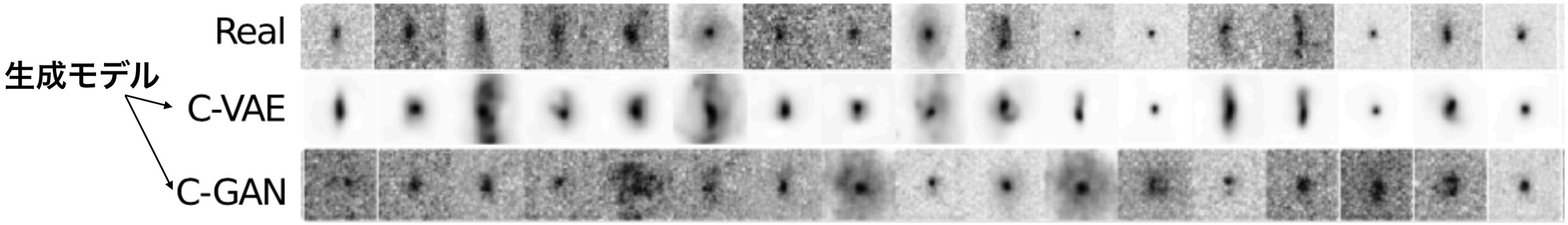
模擬観測データの大量生成

観測データの解析には、模擬観測データが大量に必要な場合がある
例）

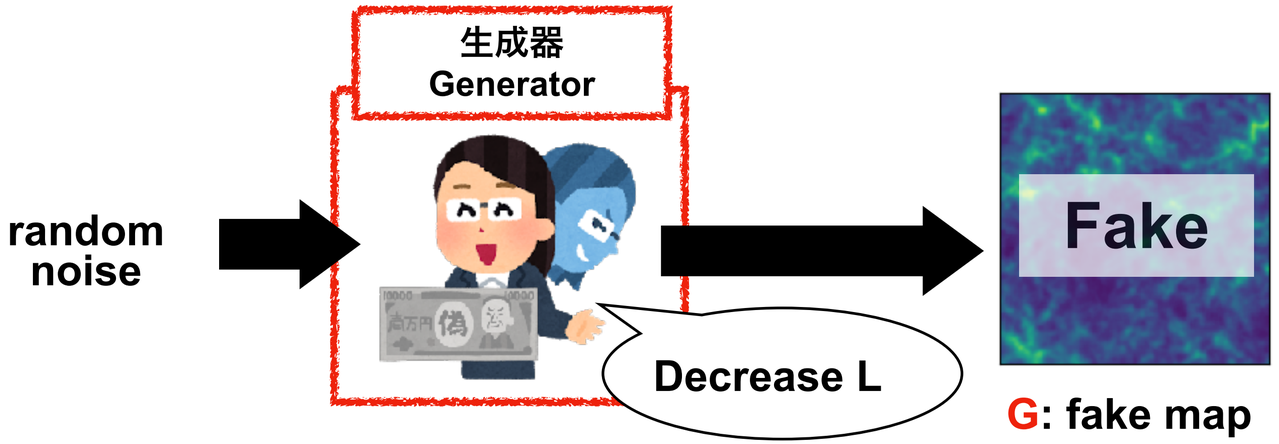
- パラメータ推定する際の共分散（Covariance）・系統誤差の見積もり
- 機械学習に用いる学習データ

機械学習は、模擬観測データを大量に生成することにも使える

重力レンズ解析に必要となる模擬銀河画像の生成（Ravanbakhsh+16）

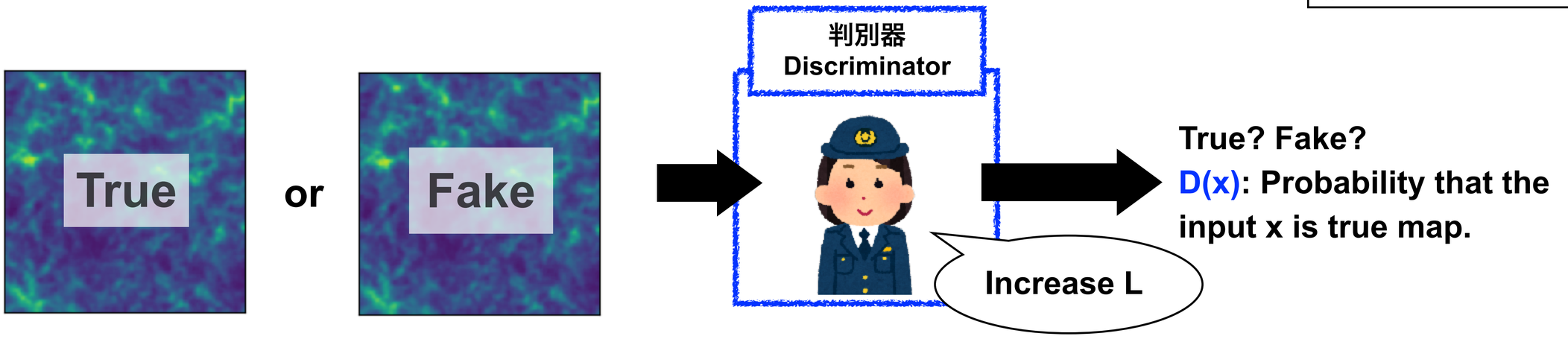


敵対的生成ネットワーク（Generative Adversarial Network; GAN）



G: fake map

二つのネットワークが敵対しながら（adversarially）学習していく



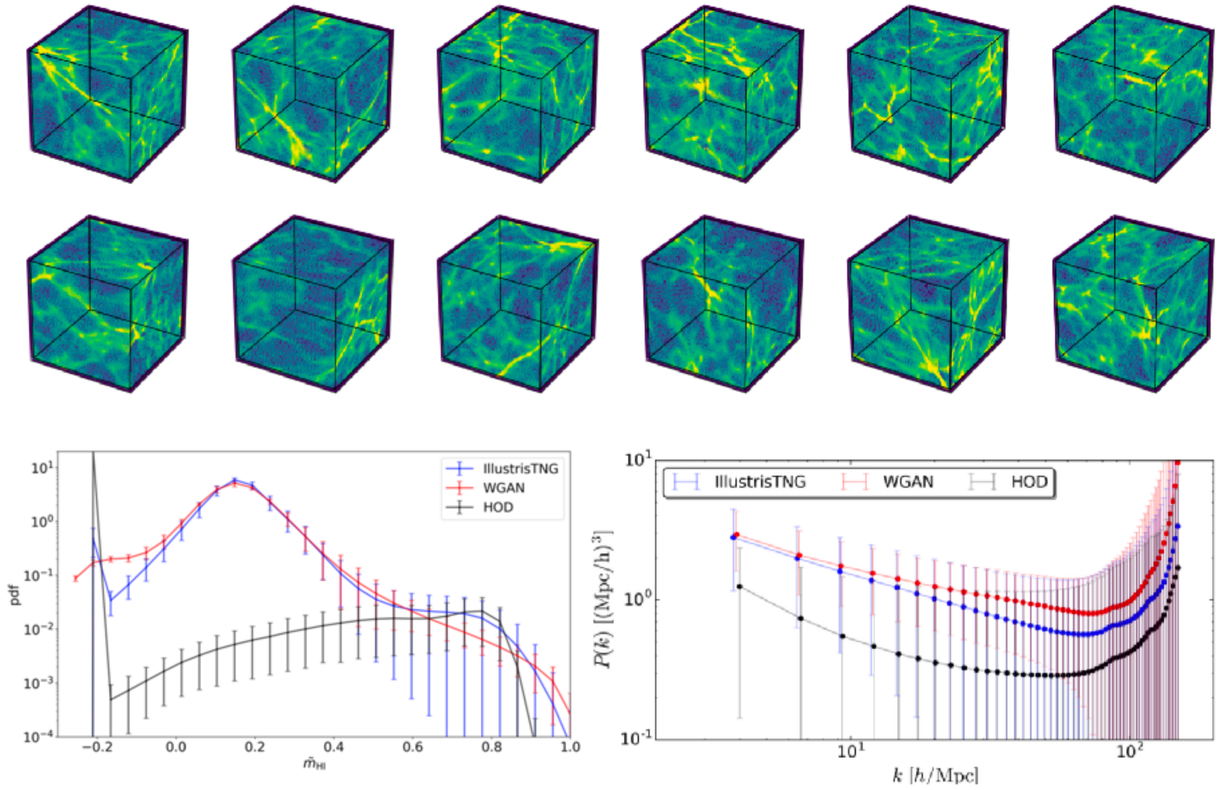
True? Fake?

D(x): Probability that the input x is true map.

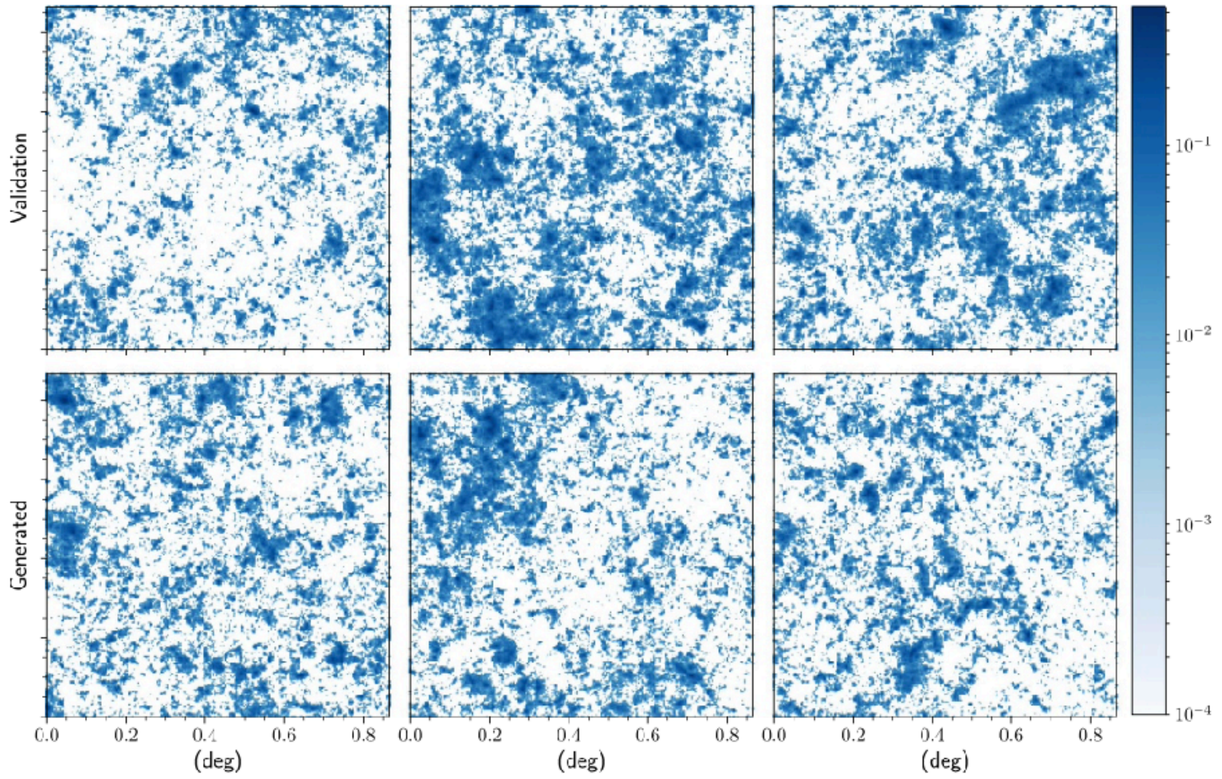
Loss function: $L[G, D] = \log D(X_{\text{true}}) + \log[1 - D(G)]$

模擬観測データの大量生成

大規模構造マップや、三次元データの生成も可能
パワースペクトルやPDFなどの統計量も再現



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



Mock weak lensing convergence map (CosmoGAN; Mustafa+19)

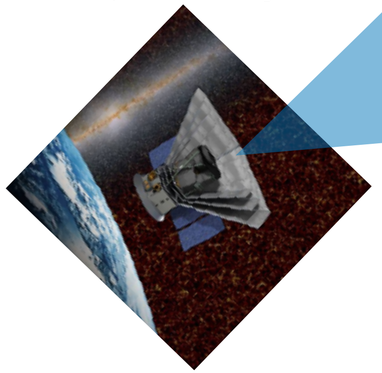
大規模構造マップからのノイズ除去

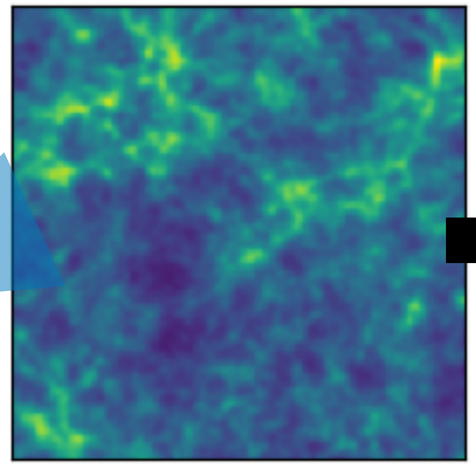
observed data
at λ = 1.5μm

Hα intensity
at z = 1.3

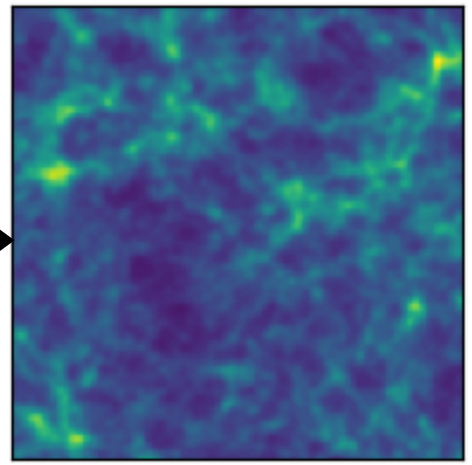
[OIII] intensity
at z = 2.0

SPHEREx
(NASA; 2024-)

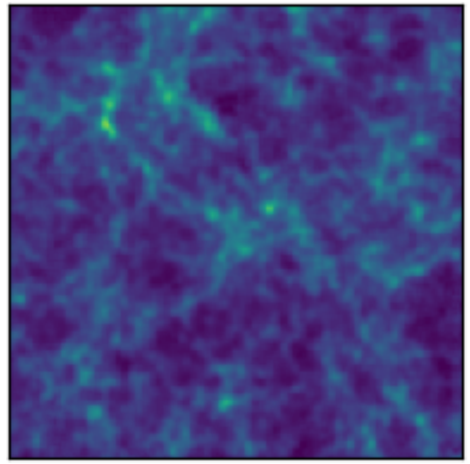




CNN



+



Traditional solutions:
- Cross-correlation
- Anisotropy in power spectrum

シミュレーションデータを用いて学習

- **Line confusion problem**

Multiple emission lines contaminate each other when they satisfy

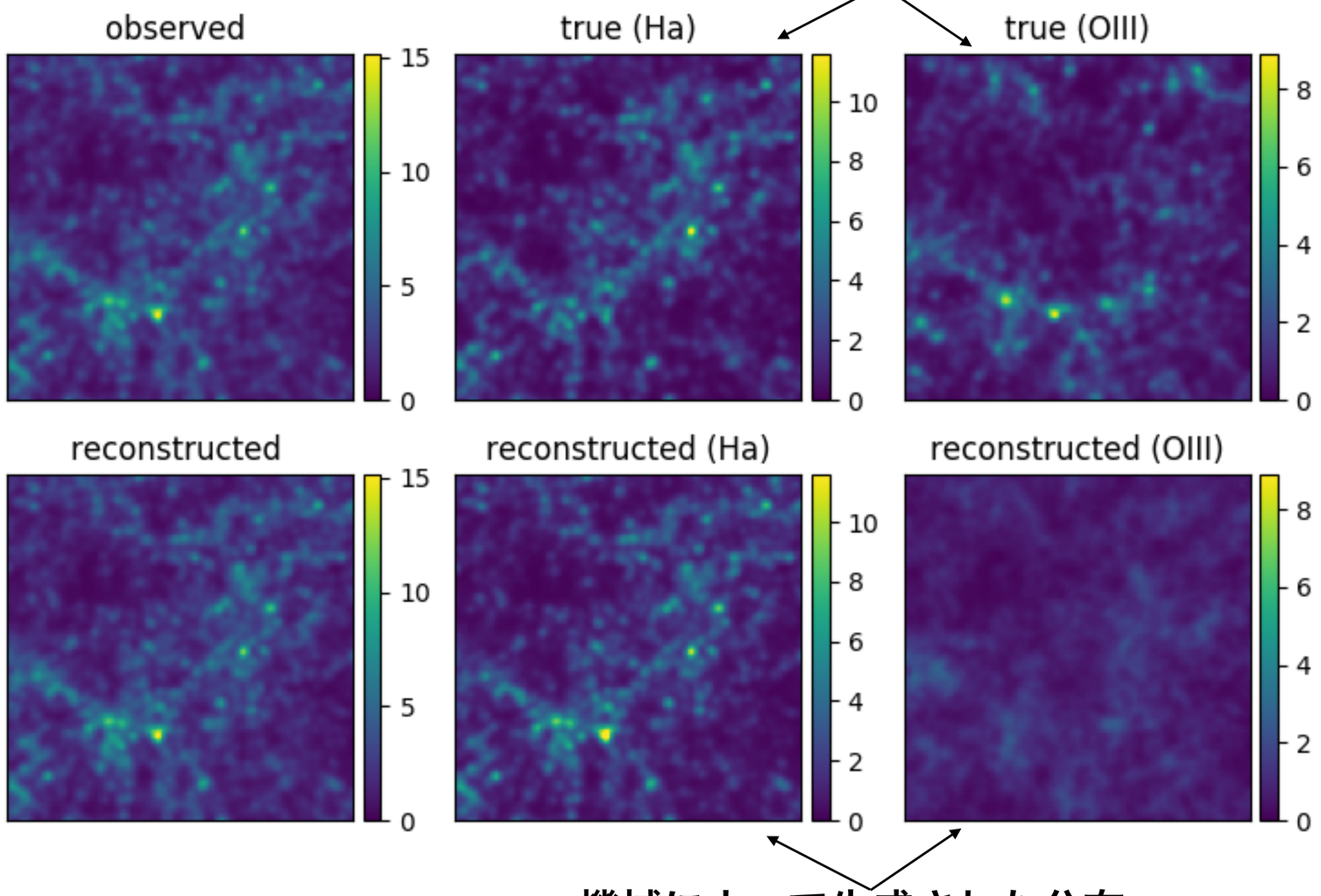
$\lambda_{obs} = \lambda_1 (1 + z_1) = \lambda_2 (1 + z_2)$

KM+2020

大規模構造マップからのノイズ除去

観測データ

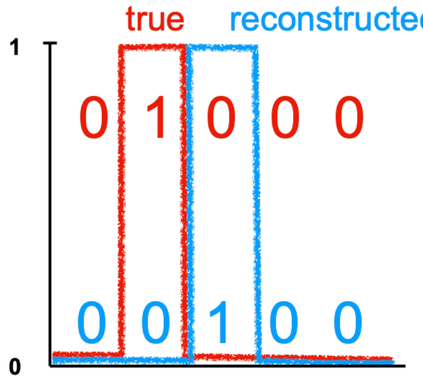
真の分布



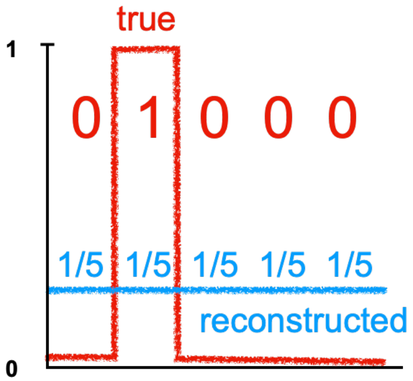
機械によって生成された分布

単純なノルムを損失関数に用いるとうまく分離ができない。

$$L_{L1} = \lambda \langle |X_{\mathrm{true}} - G(X_{\mathrm{obs}})| \rangle$$



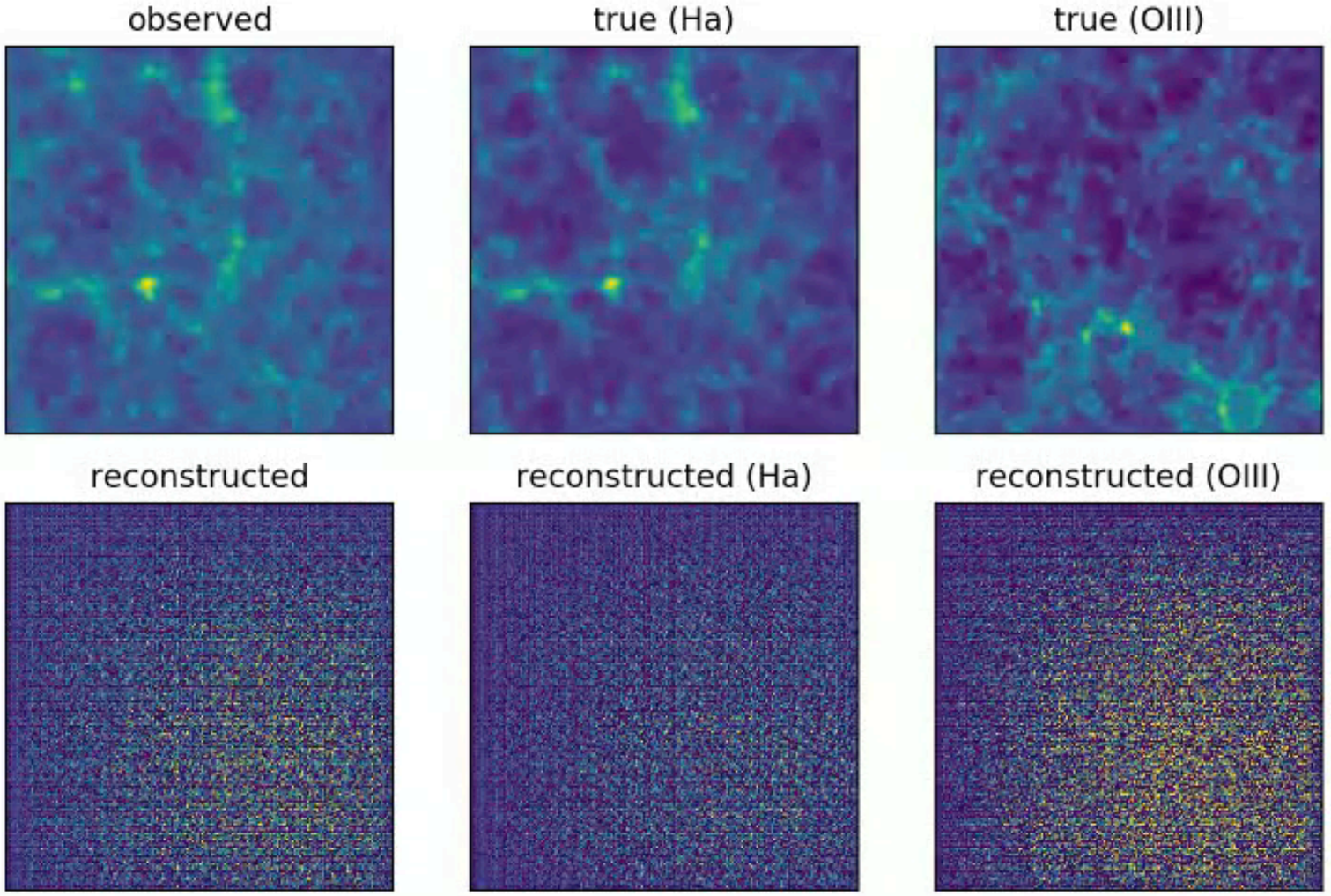
$L_{L1} = 3\times0 + 2\times1 = 2$



$L_{L1} = 4\times1/5 + 4/5 = 8/5$

大規模構造マップからのノイズ除去

GAN による学習

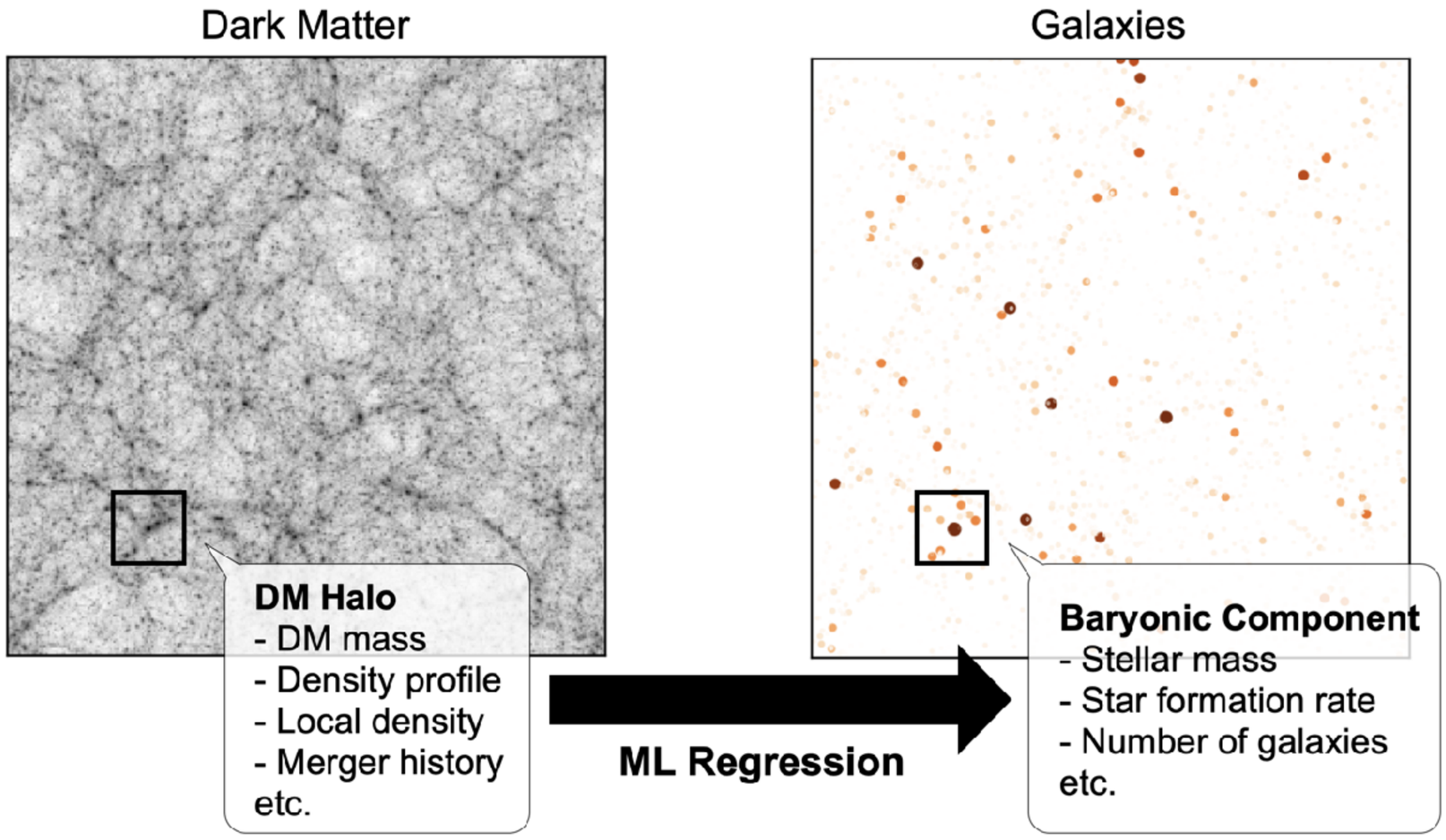


個々の赤方偏移の大規模構造を抽出することができた

シミュレーションデータへの適用例

銀河とハローの関係

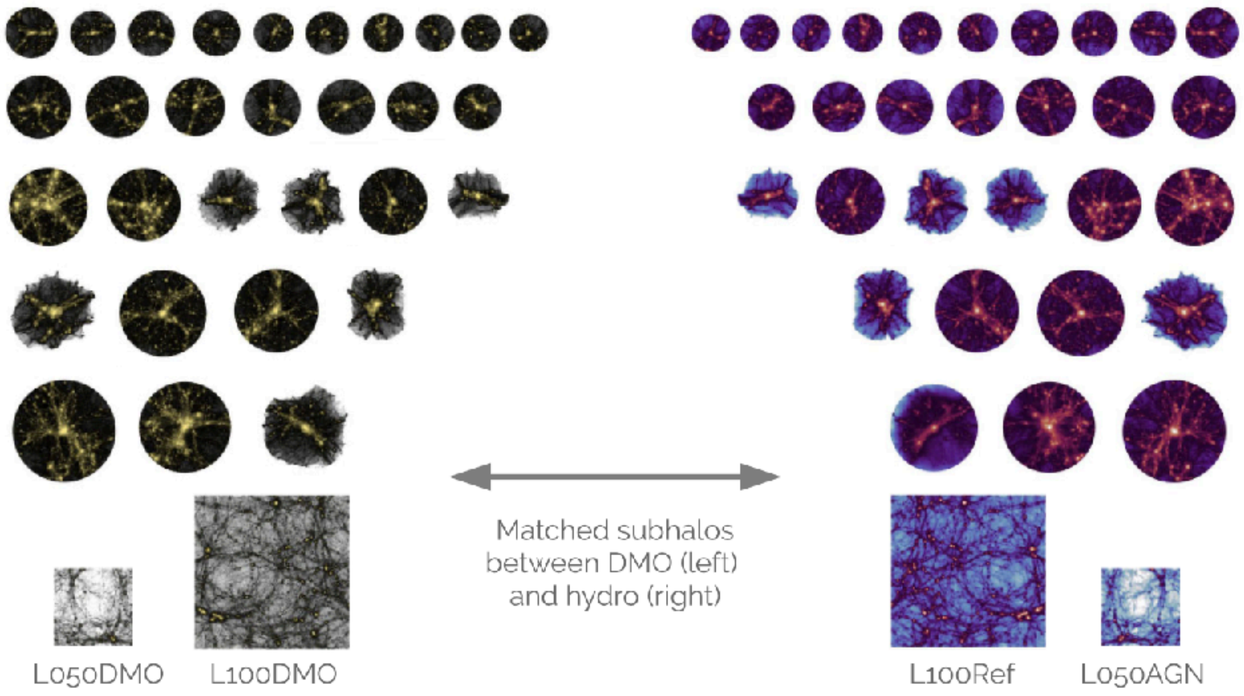
ダークマターのみのシミュレーションは低コストだけど銀河に関する情報を得られない
一方、宇宙論的流体シミュレーションは高コストで大量に・大領域でやるのは難しい



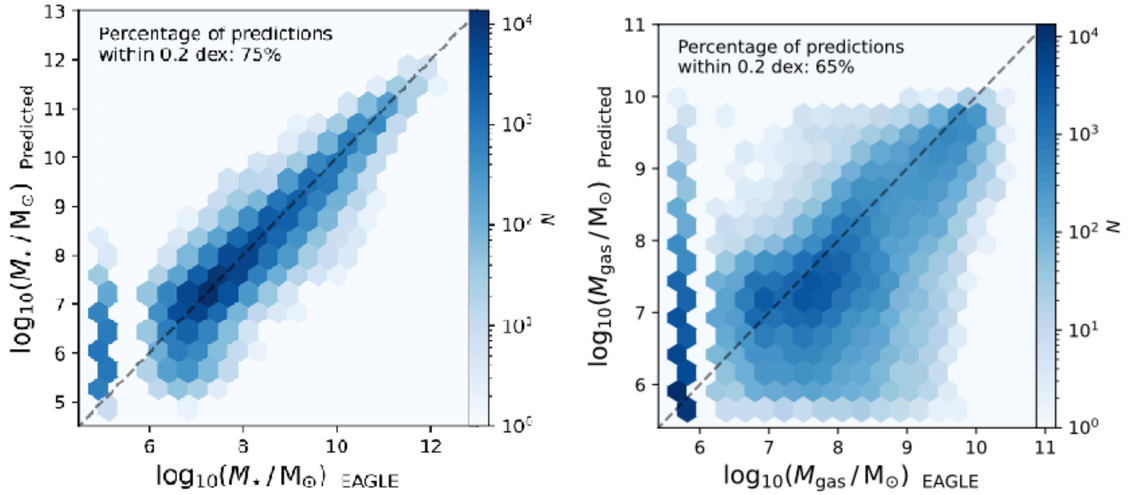
Moriwaki+23

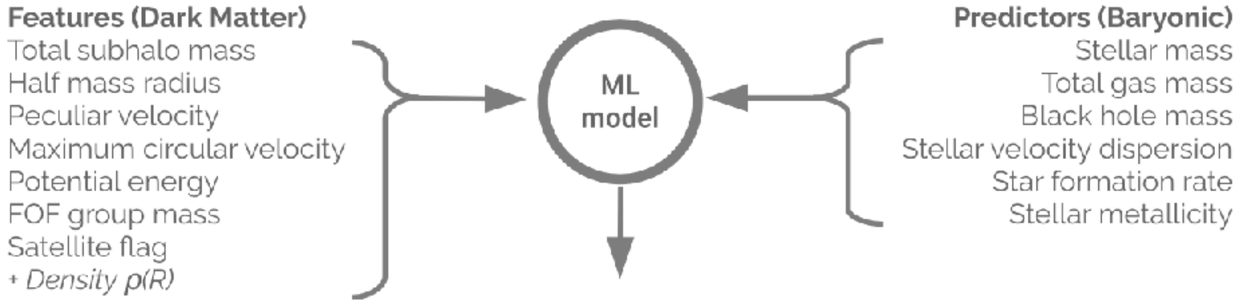
銀河とハローの関係

Lovell+22



宙論的流体シミュレーションを用いてダークハローと銀河との関係を学習させる





様々なシミュレーションが用いられている

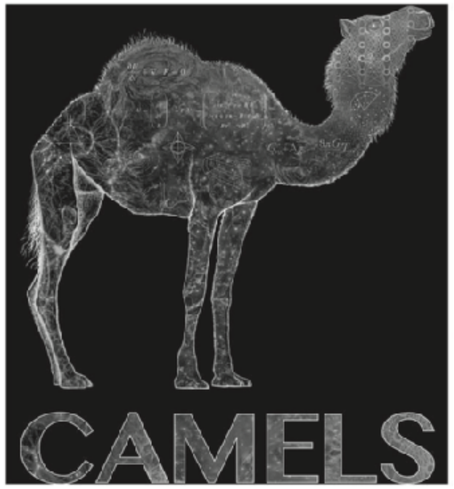
- **Illustris (Kamdar+16)**
- **MUFASA (Agarwal+18)**
- **IllustrisTNG (Jo & Kim 19)**
- **EAGLE (Lovell+22)**

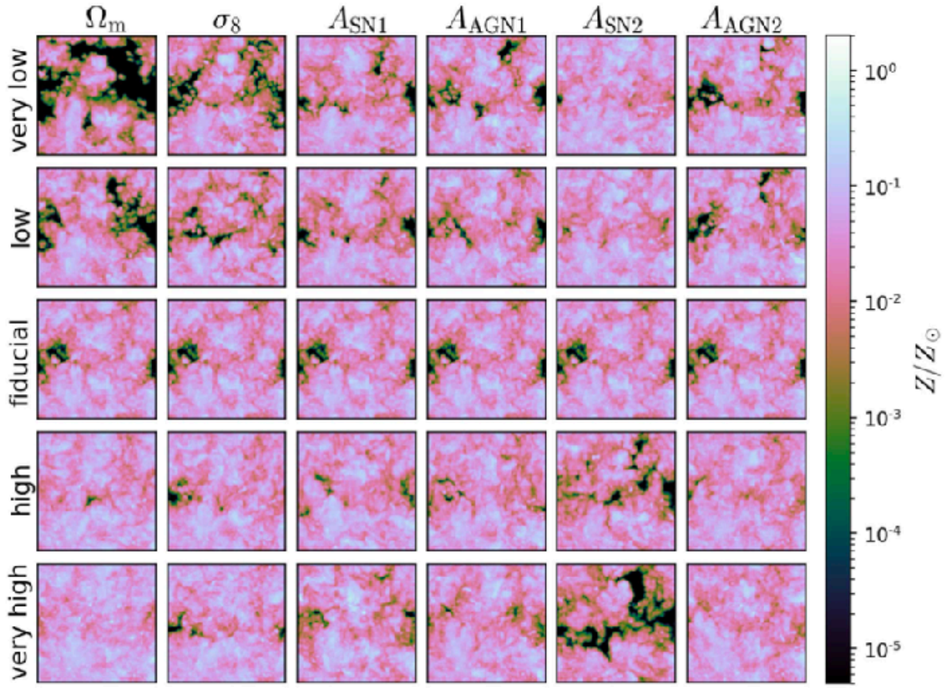
宇宙論的流体シミュレーションデータ

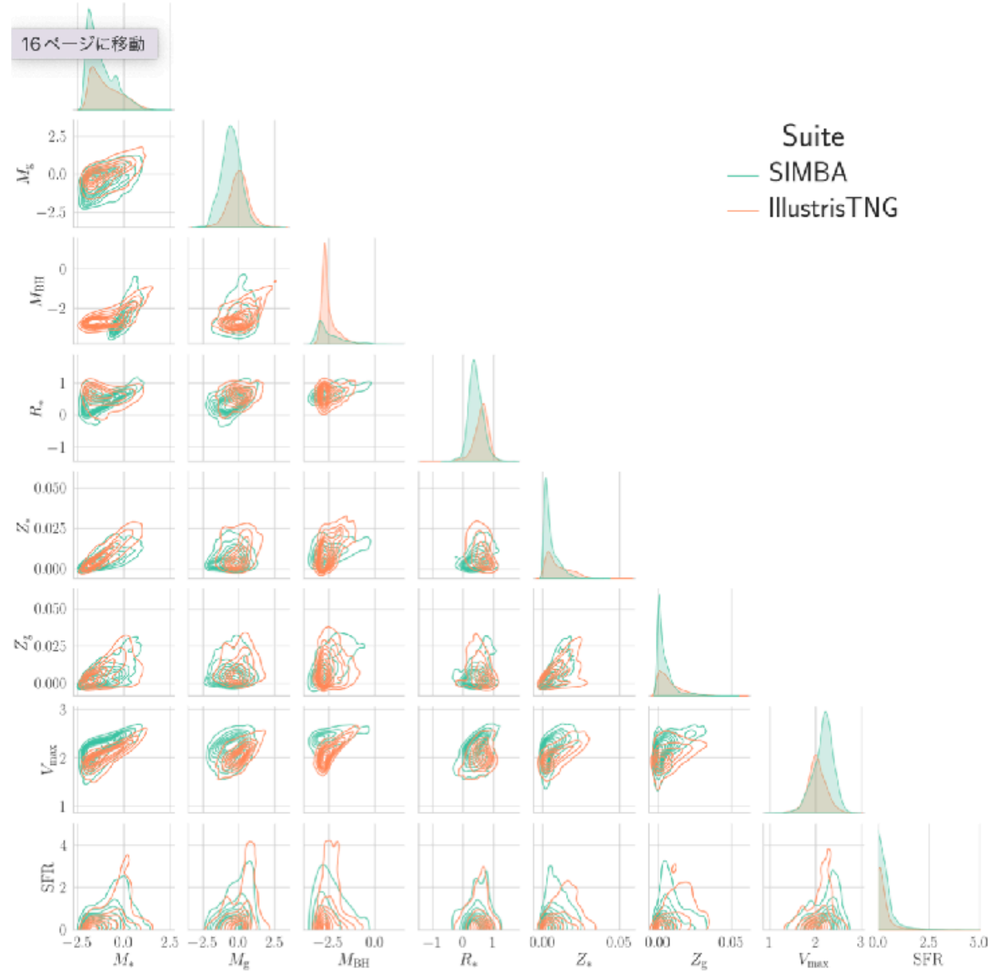
Detailed cosmological hydrodynamics simulation is costly

- **Lack of training data**
- **Lack of diversity in training data**

What shall we do?





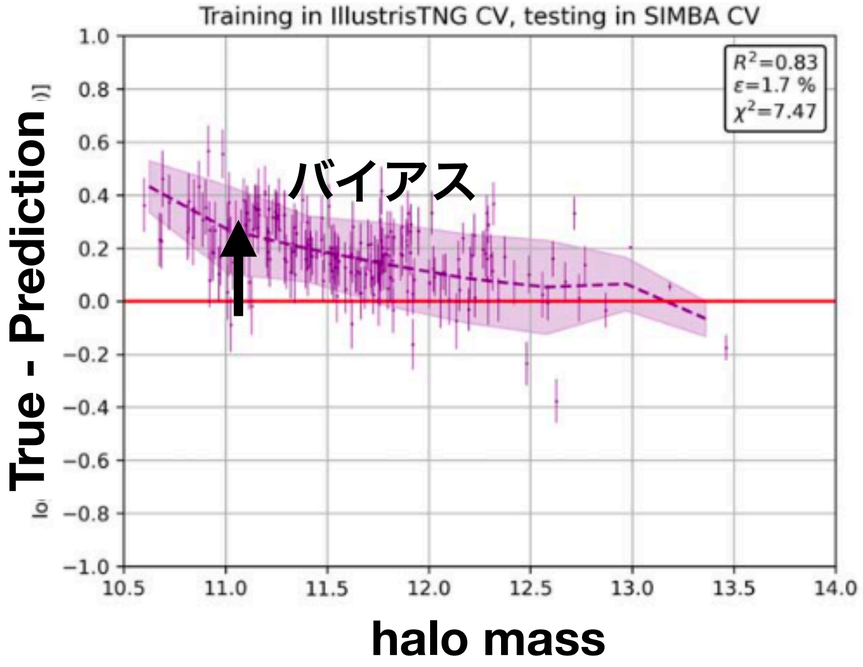


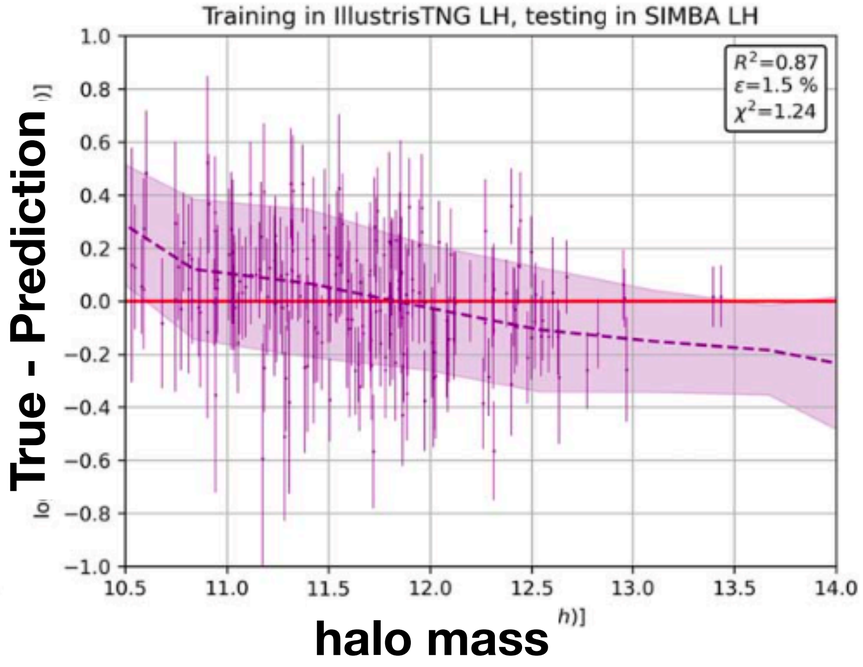
CAMELS project (Villaescusa-Navarro+2021)

宇宙論的流体シミュレーションデータ

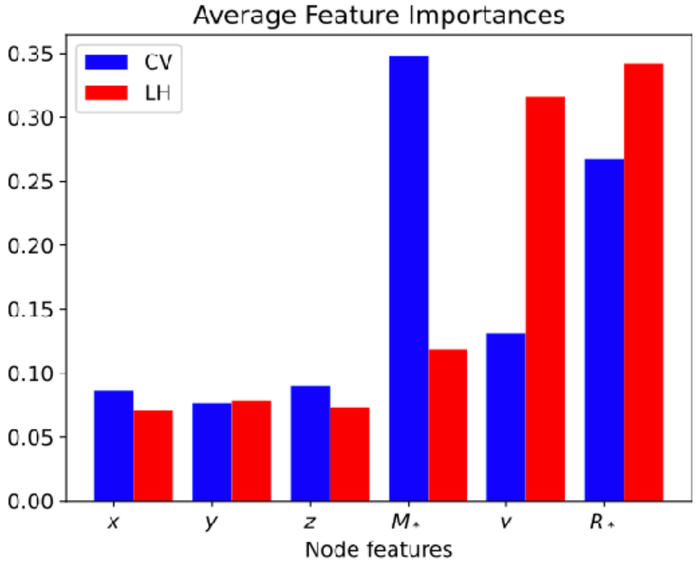
特定の astrophysical/cosmological パラメータを用いたシミュレーションで学習した場合

さまざまなパラメータを用いたシミュレーションで学習した場合





Feature importance
（どの物理量が重要か）



学習データに多様性を持たせることでバイアスのない推論ができる

まとめ

- 機械学習とは、データから自動的に最適解を見つけること
- 特に深層学習では大量のパラメータを最適化することで複雑な課題もこなすことができる

- 機械学習を使うメリット：
  - 高速に大量のデータを処理できる
  - 解析的に処理するのが難しいタスクをこなせる

  → 今後の銀河サーベイプロジェクトにおいて機械学習が重要となる

- 特に CNN を用いた画像解析での成功例が多い（e.g., 形態分類、photoz、重力レンズ）
- 目的によって評価指標が異なる（e.g., レアな天体の発見は recall が大きければOK）
- 新しいモデルも次々と用いられている（e.g., GAN）
- 学習データをどのように増やすかが大きな課題（e.g., photoz、CAMELS project）
- そのほかモデル出力の不定性など、解決すべき課題は色々とある

機械学習の応用における課題とその解決策

Challenge 1 Small (and biased) 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)
Solution 1.D Active Learning and similar	Walmsley et al. (2020)
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); Bowles et al. (2021); Bhambra et al. (2022)
Solution 3.B Symbolic regression	Cranmer et al. (2020)
Solution 3.C Physics informed	Scaife & Porter (2021); Villar et al. (2021a); Charnock et al. (2019)
Challenge 4 Domain shift	
Solution 4.A Transfer Learning	Tuccillo et al. (2018); Domínguez Sánchez et al. (2019); Ghosh et al. (2020)
Solution 4.A Domain Adaptation	Ćiprijanović et al. (2021b)
Challenge 5 Benchmarking	
Solution 5.A Standardized datasets	PLAsTiCC, SKA data challenge, Galaxy Zoo

Table 2 Major challenges that deep learning works applied to astronomy might suffer and that will need to be addressed in the coming years. We also provide elements of solutions already being explored along with the corresponding references.

学習データが大量に必要

出力の不定性

「ブラックボックス」性

学習データと実際のデータが異なる

手法間の比較

Huertas-Company & Lanusse (2022)

参考文献

- **Ball & Brunner (2010) “Data Mining and Machine Learning in Astronomy”, International Journal of Modern Physics D, 19, 1049**
  早期の機械学習の天文研究への応用例
- **Baron (2019) “MACHINE LEARNING IN ASTRONOMY: A PRACTICAL OVERVIEW”, arXiv:1904.07248**
  機械学習における一般的な手法論や SVM, RF, ANN の具体的な構造
- **Ntampaka et al. (2019) “The Role of Machine Learning in the Next Decade of Cosmology”, BAAS, 51, 14 (Astro2020 Science White Paper)**
  今後の大規模観測に機械学習がどう貢献できるか
- **Fluke & Jacobs (2020) “Surveying the reach and maturity of machine learning and artificial intelligence in astronomy”, WIREs Data Mining and Knowledge Discovery, 10, e1349**
  天文学における機械学習活用についてデータの型や手法・モデルごとに網羅的にまとめられている
- **Huertas-Company & Lanusse (2022) “The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys”**
  銀河サーベイにおける深層学習のインパクト
- **Moriwaki et al. (2023) “Machine Learning for Observational Cosmology”, Rep. Prog. Phys. 86 076901**
  宇宙論（含銀河サーベイ）における機械学習の応用例が網羅的にまとめられている
- **https://github.com/georgestein/ml-in-cosmology**
  宇宙論分野での機械学習の大量の応用例を arXiv から網羅的にリスト化してくれている

機械学習をこれから取り入れてみたいという方に

- 技術的なことは参考書なども活用しつつ、各フレームワークのチュートリアルをやってみるのが手早い。非深層学習であれば Scikit-learn 、深層学習に関してはPyTorch が使いやすいように感じる。他にもTensorFlow, Keras, JAX など。そのほか基本的なところや最新の情報を勉強するには YouTube チャンネルなども便利（e.g., @stanfordonline, @AndrejKarpathy, @Alcia_Solid, etc.）
- 各プラットフォーム上でシェアされているコードを一旦そのまま使ってみるのも有用。From scratch で書くのは勉強にはなるが無駄も多い。特に新しいモデルは次々と出てくるので追いつくのはなかなか大変。データ科学の分野の人と協力することも大事
- 天文・天体物理への理解を深めた上で以下の点などを検討する
  - 従来の手法の課題は何か？（**データ量が問題？複雑さが問題？**）
  - 最終目的は何か？（**統計的な議論？特定の天体の発見？**）
  - **そもそも原理的に可能なタスクか？学習データは適切か？データにバイアスはないか？etc.**
  - 物理的な情報を活用してモデルを改良できないか？
- 特に深層学習はしばしばハイパーパラメータのチューニングが大変。大まかな傾向はあるけど、タスク・データによってうまくいくものはさまざま。データの事前処理やモデル自体を変えることが必要な場合も。従来の手法の方が良い場合もあるので、必ずしも機械学習にこだわりすぎず柔軟に考えられるとベストかなと個人的には思う