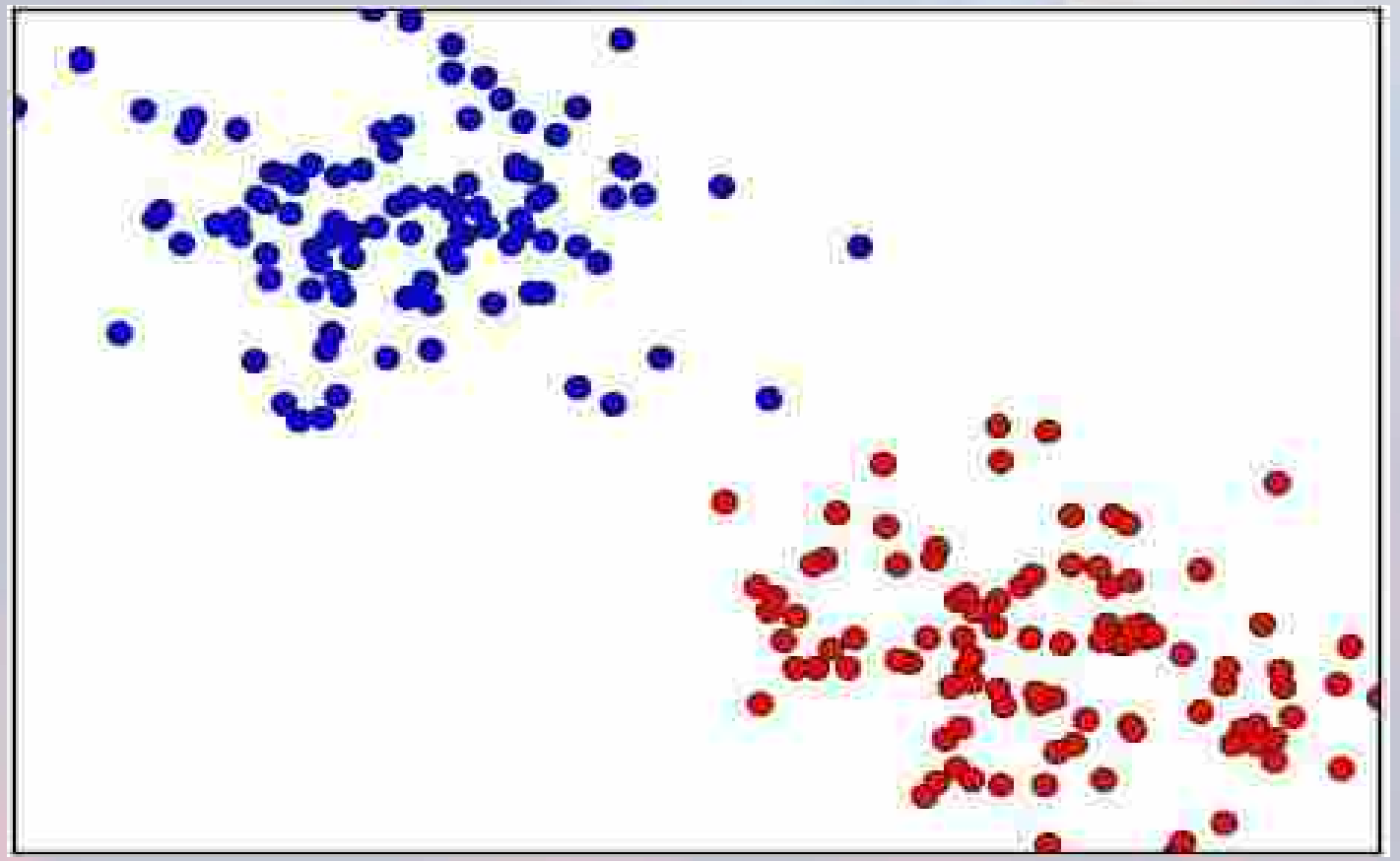
Pre-seminar

What is ML?

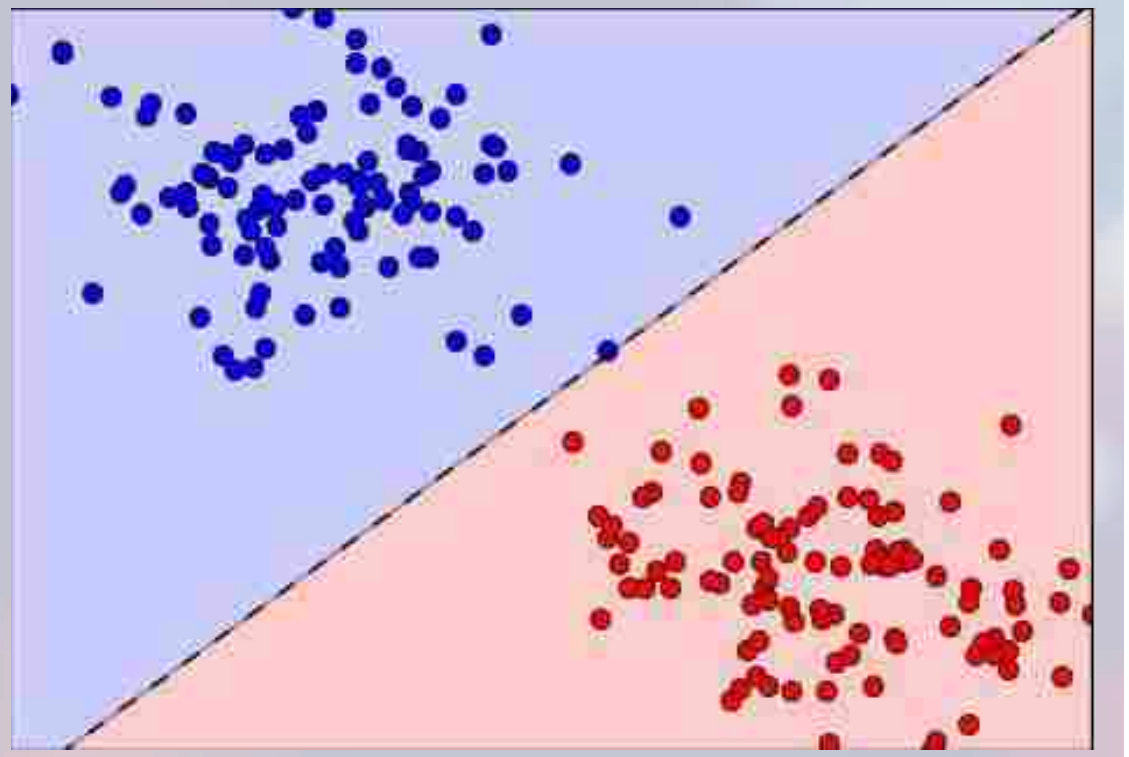
Geneva University , Jan, 2020

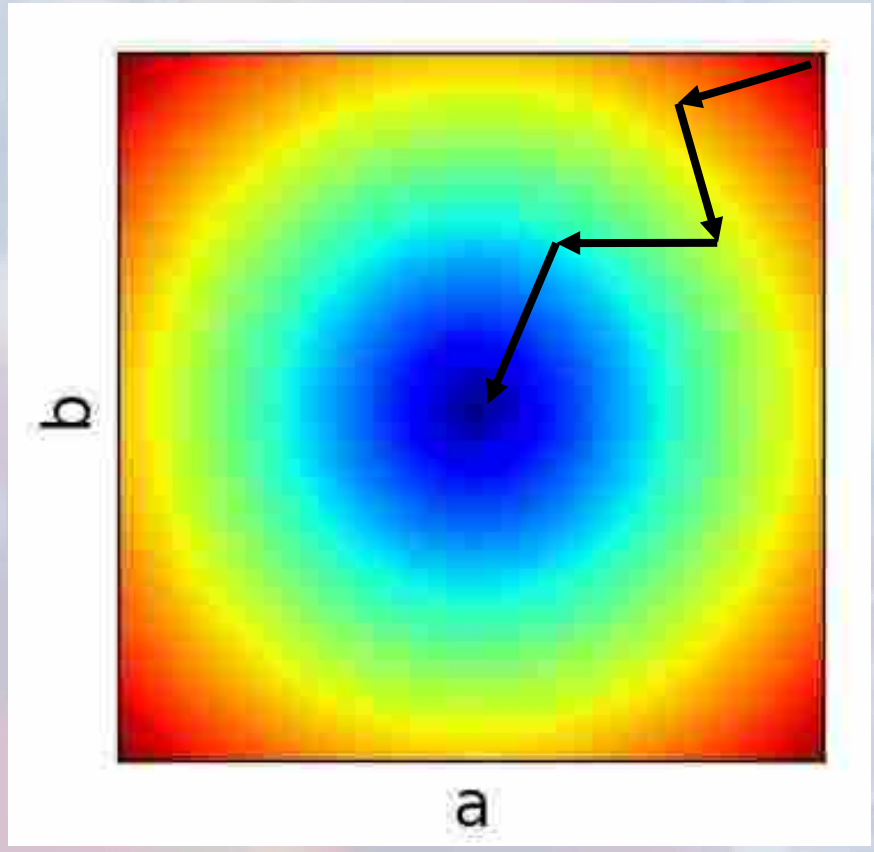
- What is Machine learning?
- What is Dimensionality reduction?
- What is Clustering Analysis?
- What am I doing with ML?

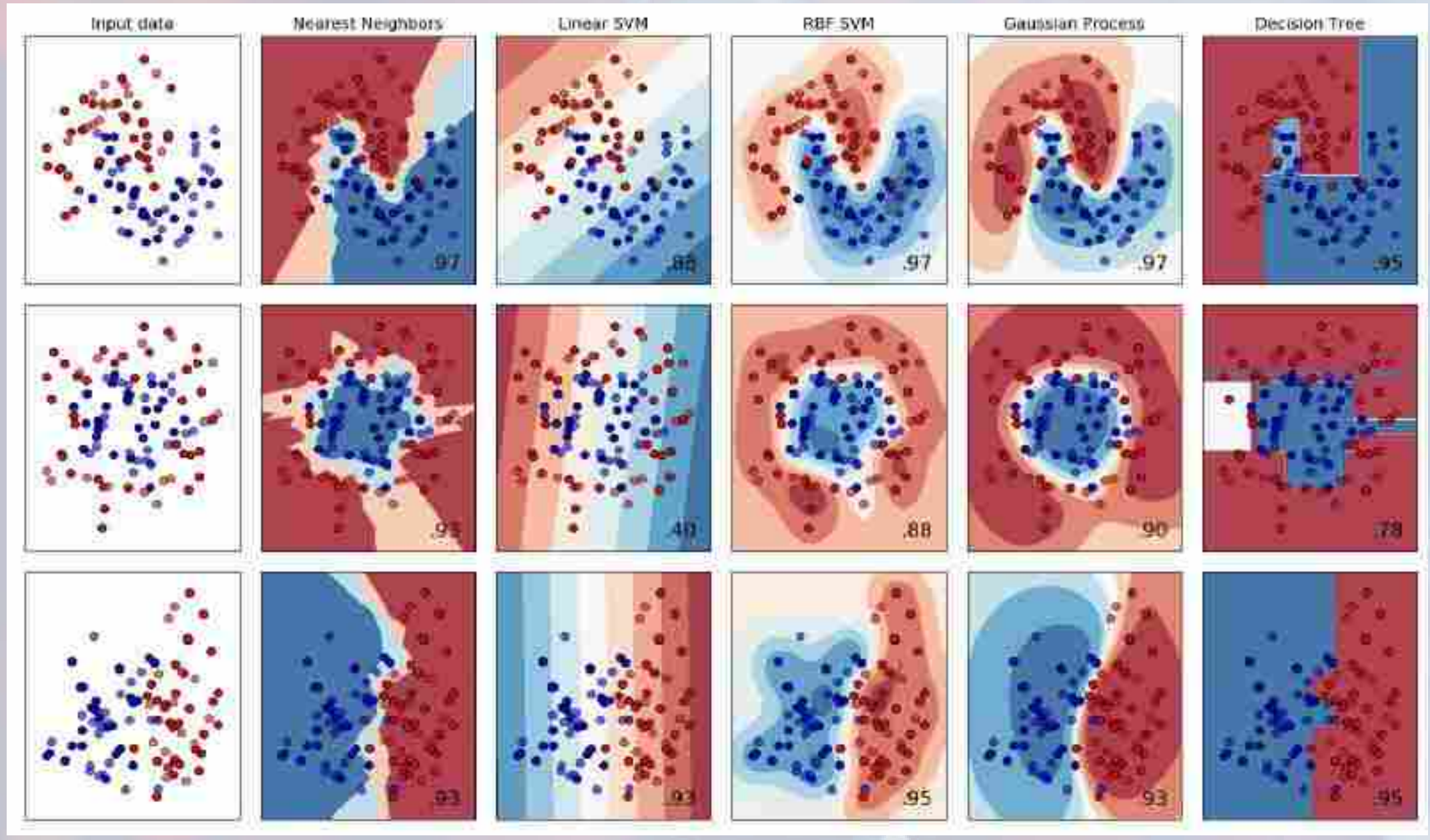
What is Machine learning?

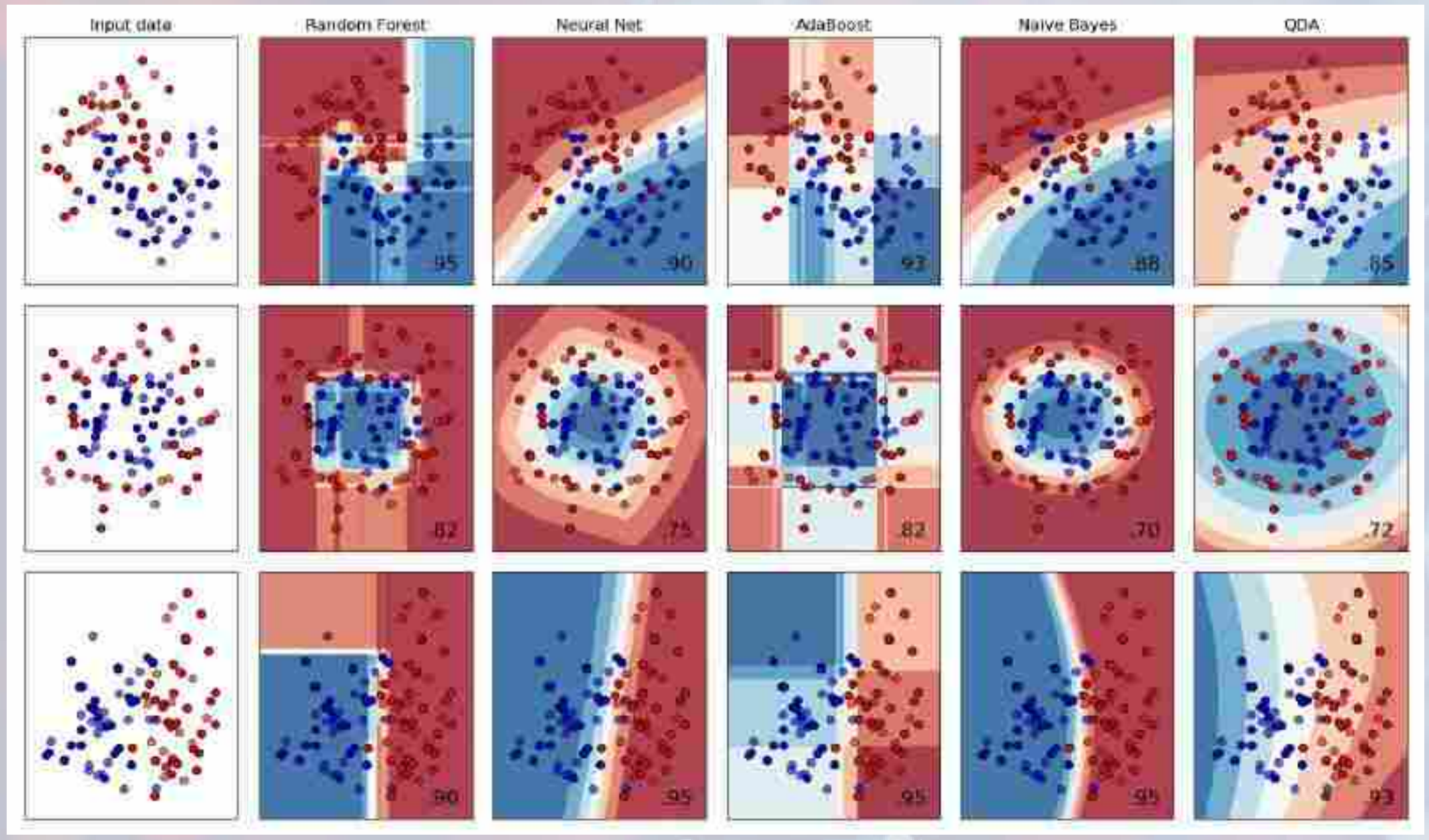


$$y = ax + b$$

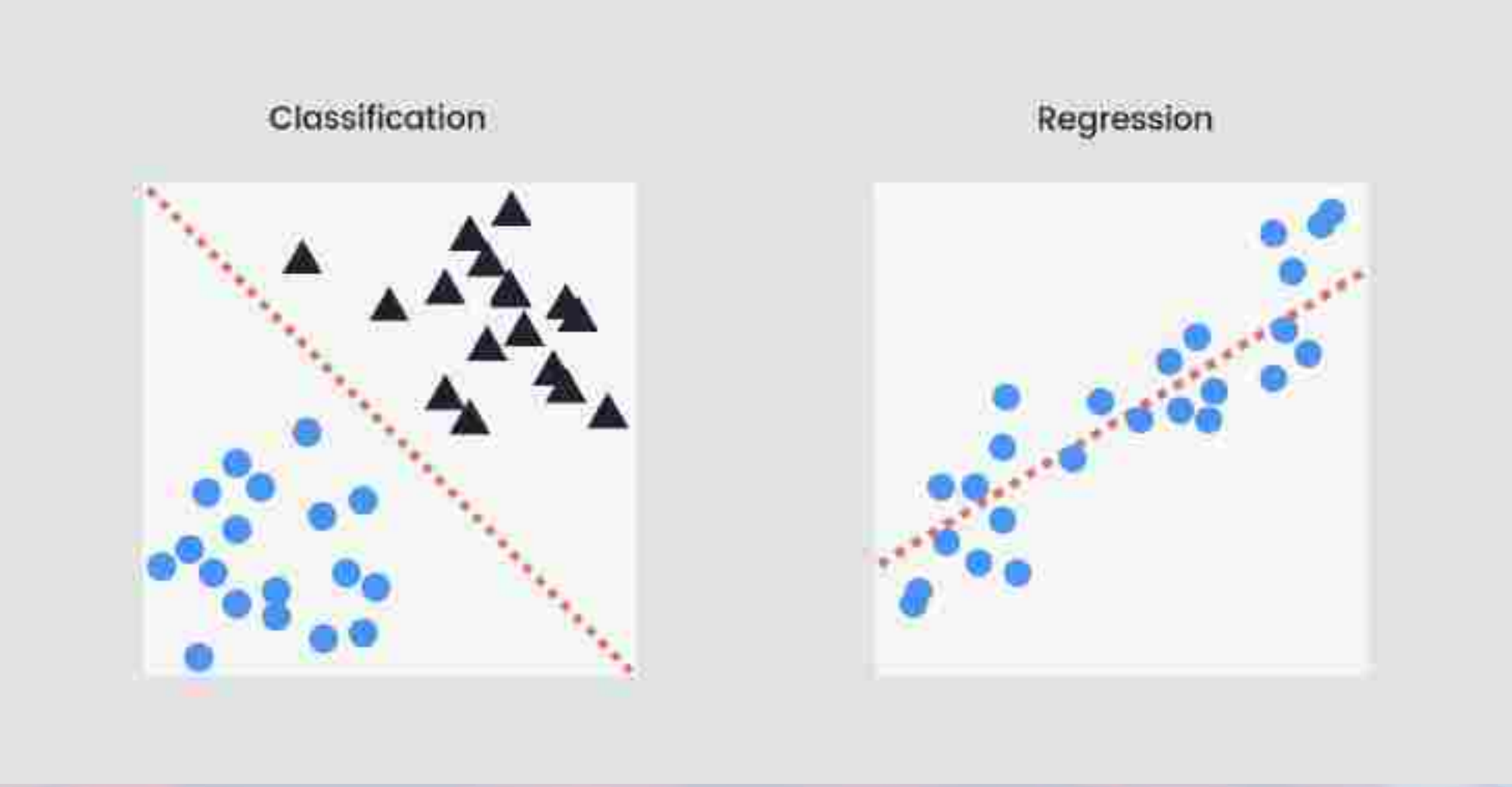




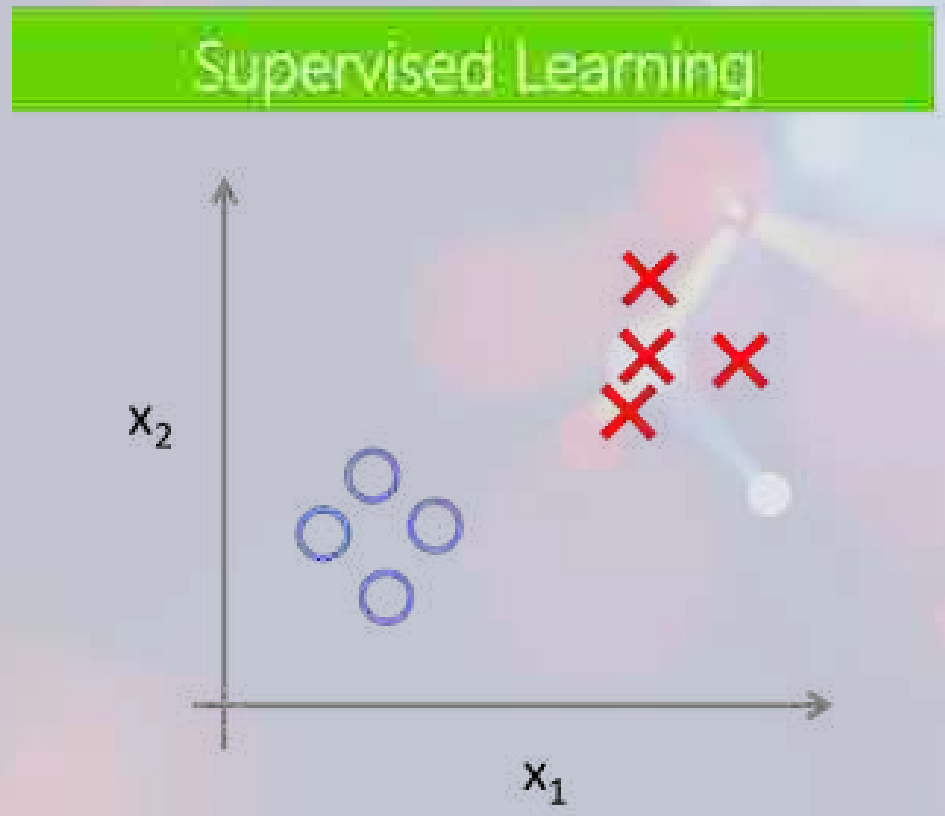

Input data
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
.97
.88
.97
.97
.95
.93
.40
.88
.90
.78
.93
.93
.95
.93
.95

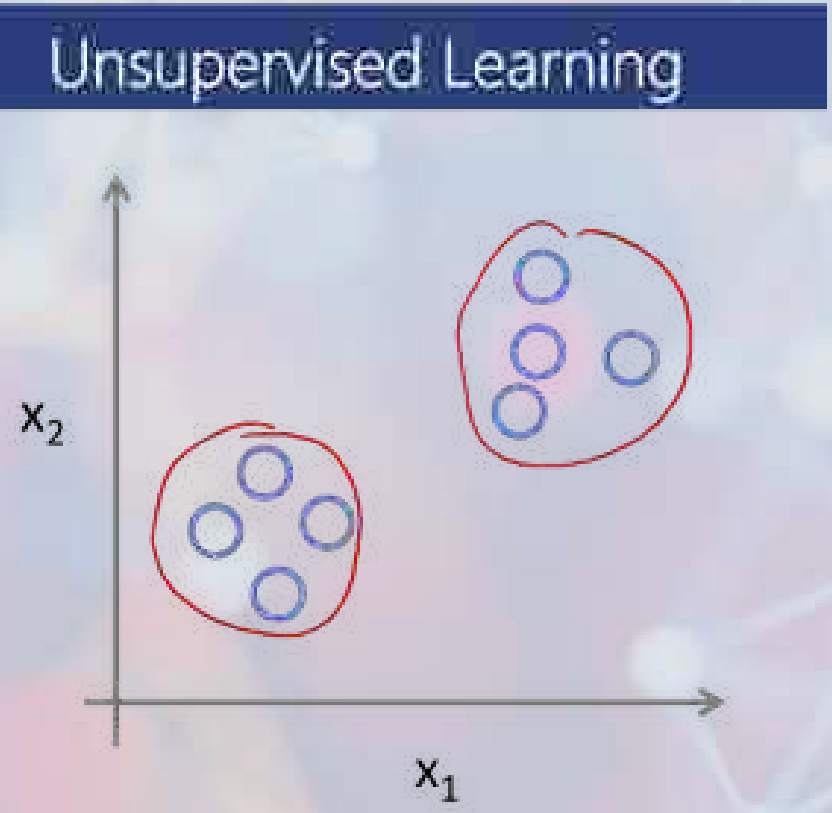

Input data
Random Forest
Neural Net
AdaBoost
Naive Bayes
QDA
.95
.90
.93
.88
.85
.82
.75
.82
.70
.72
.90
.95
.95
.95
.93

Classification vs. regression

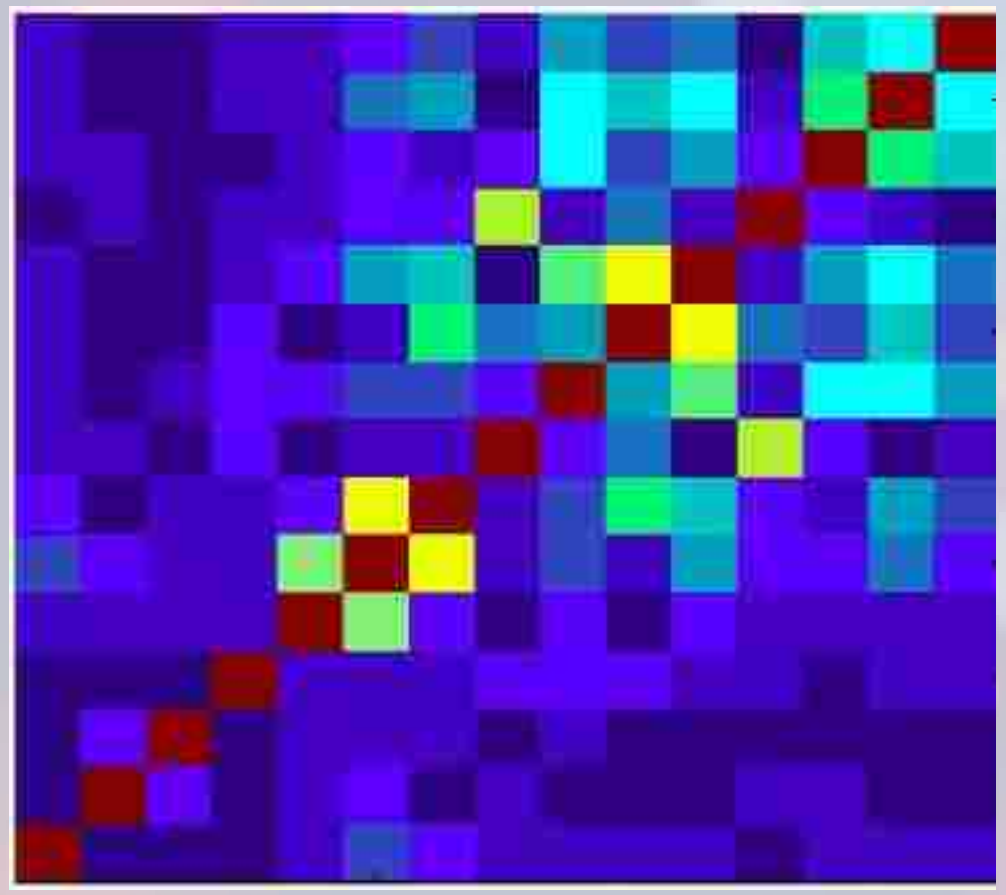


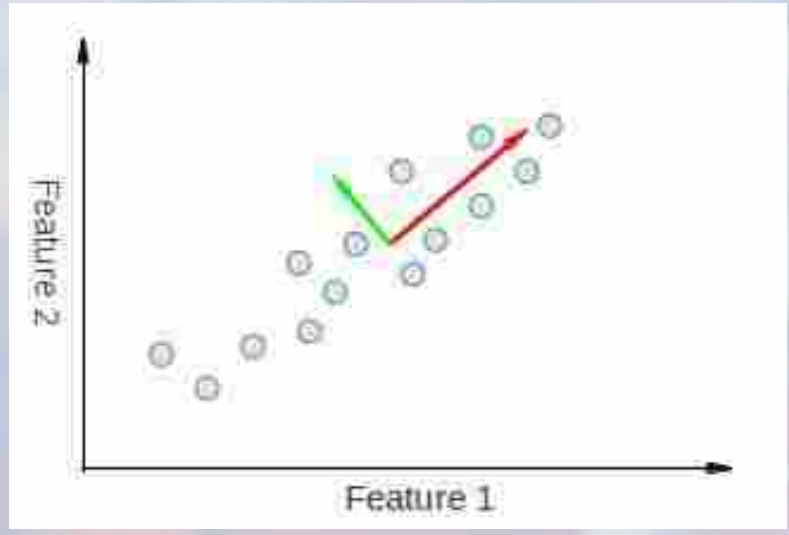
Supervised vs Unsupervised



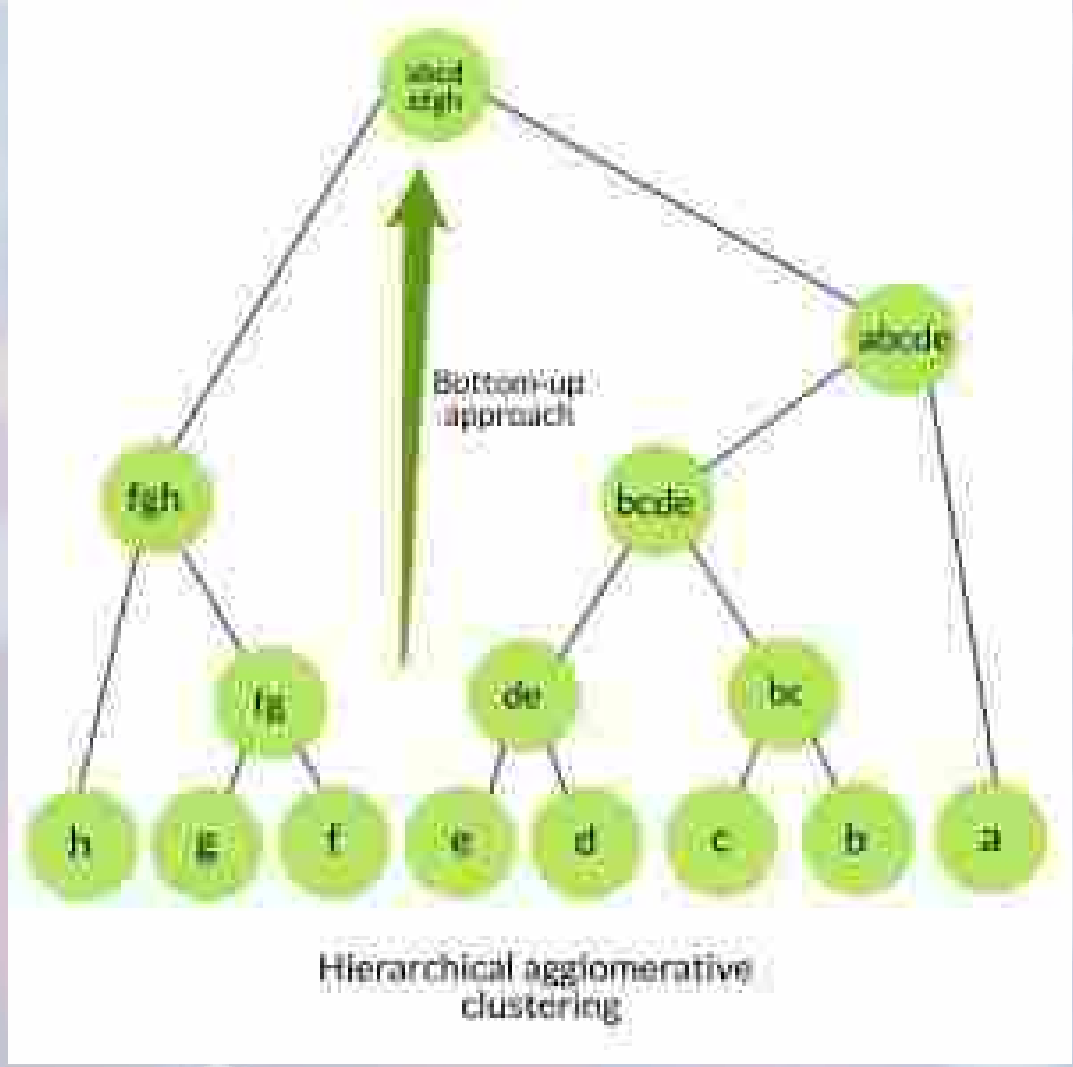


What is Dimensionality reduction?



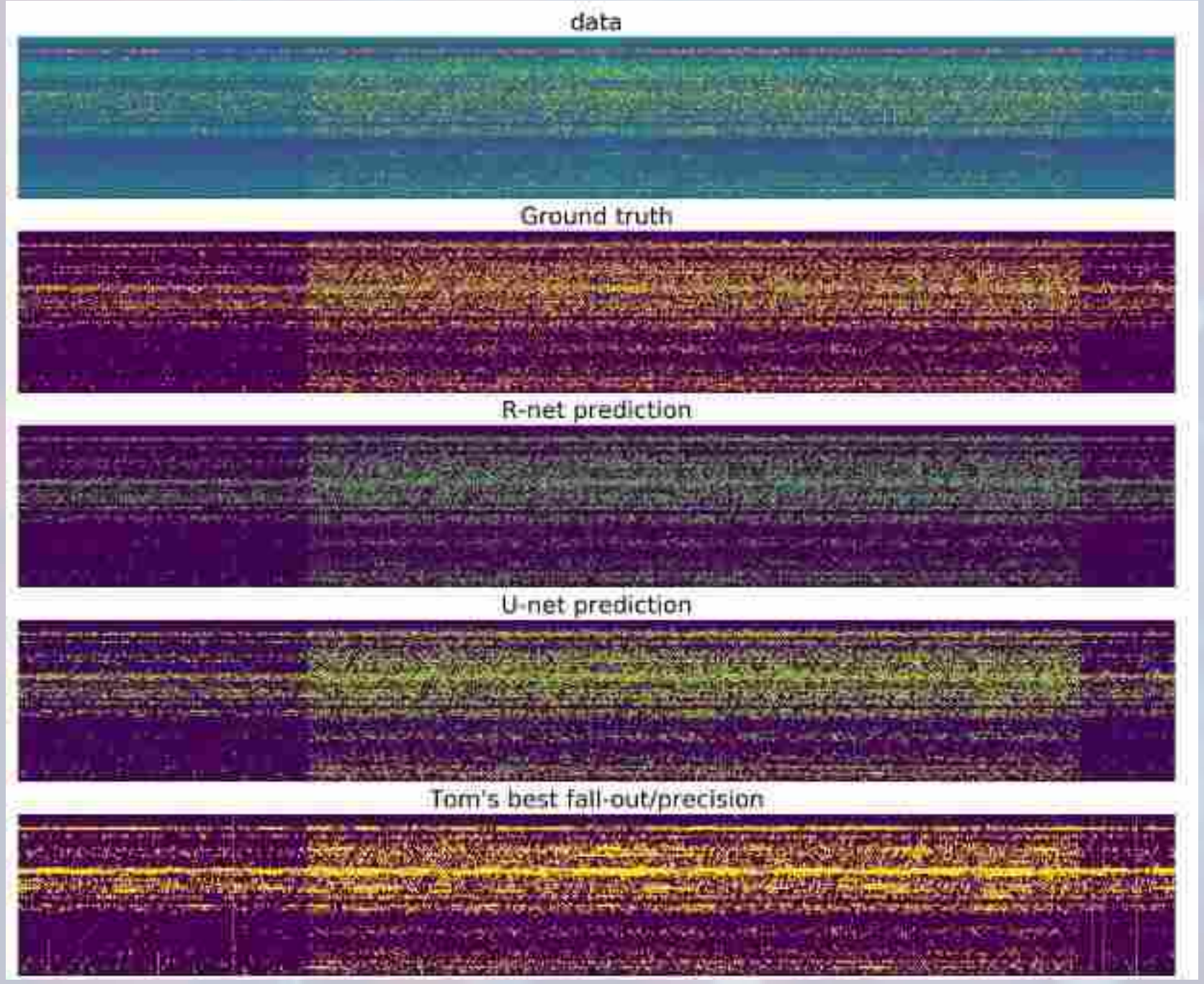

Feature 2
Feature 1

What is Clustering Analysis?

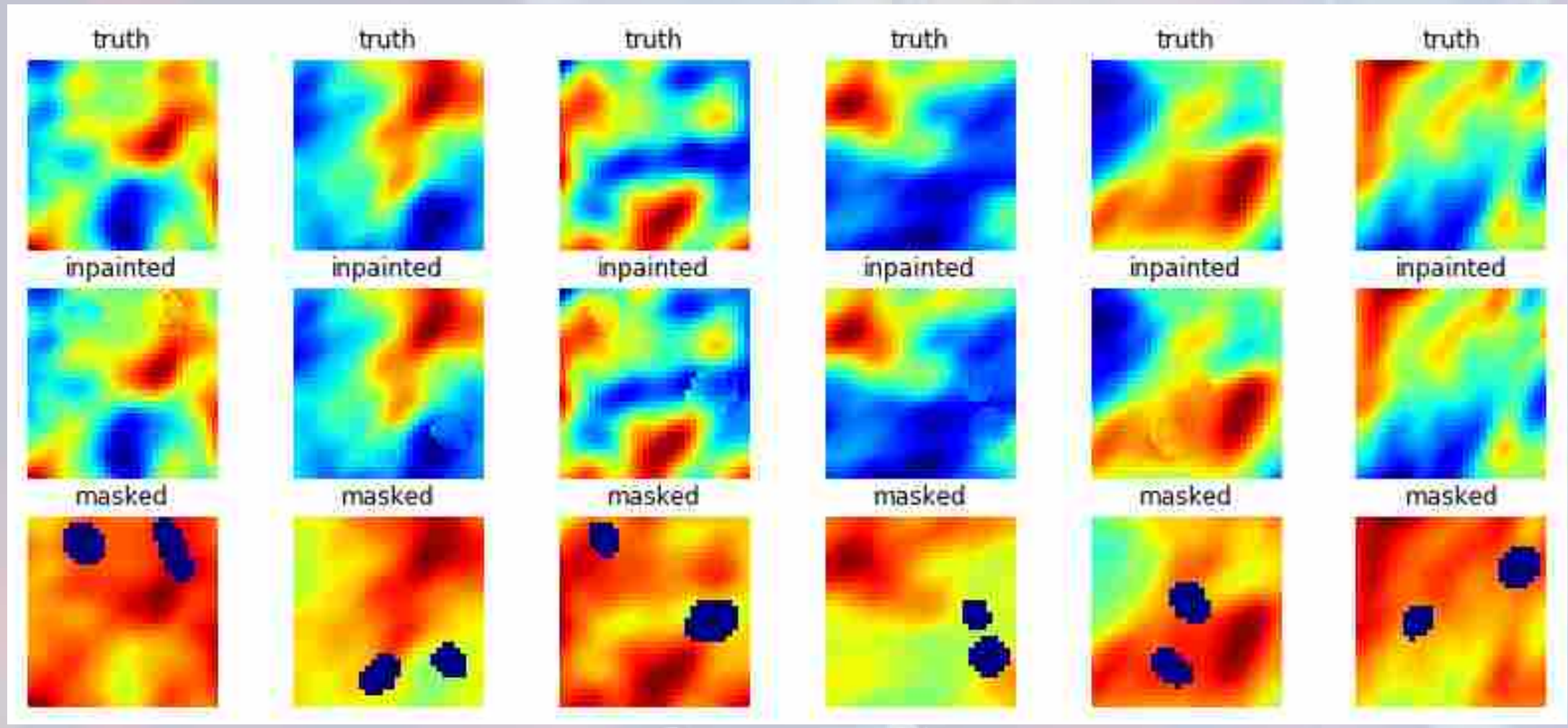


What am I doing with ML?

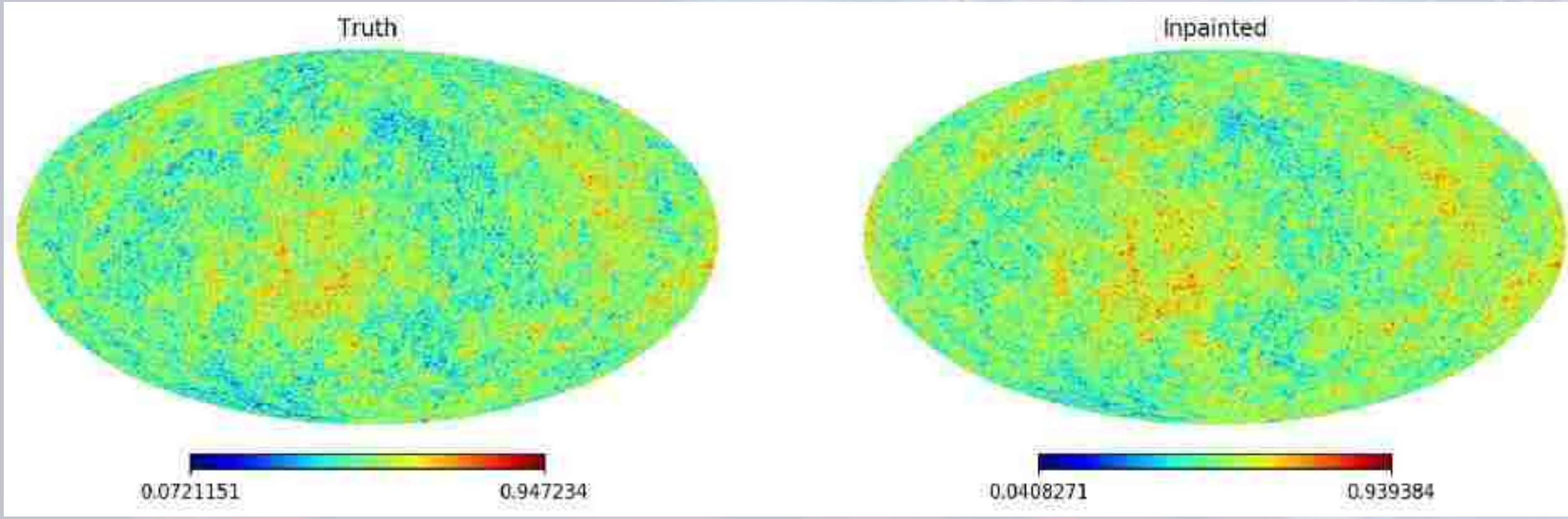
RFI mitigation



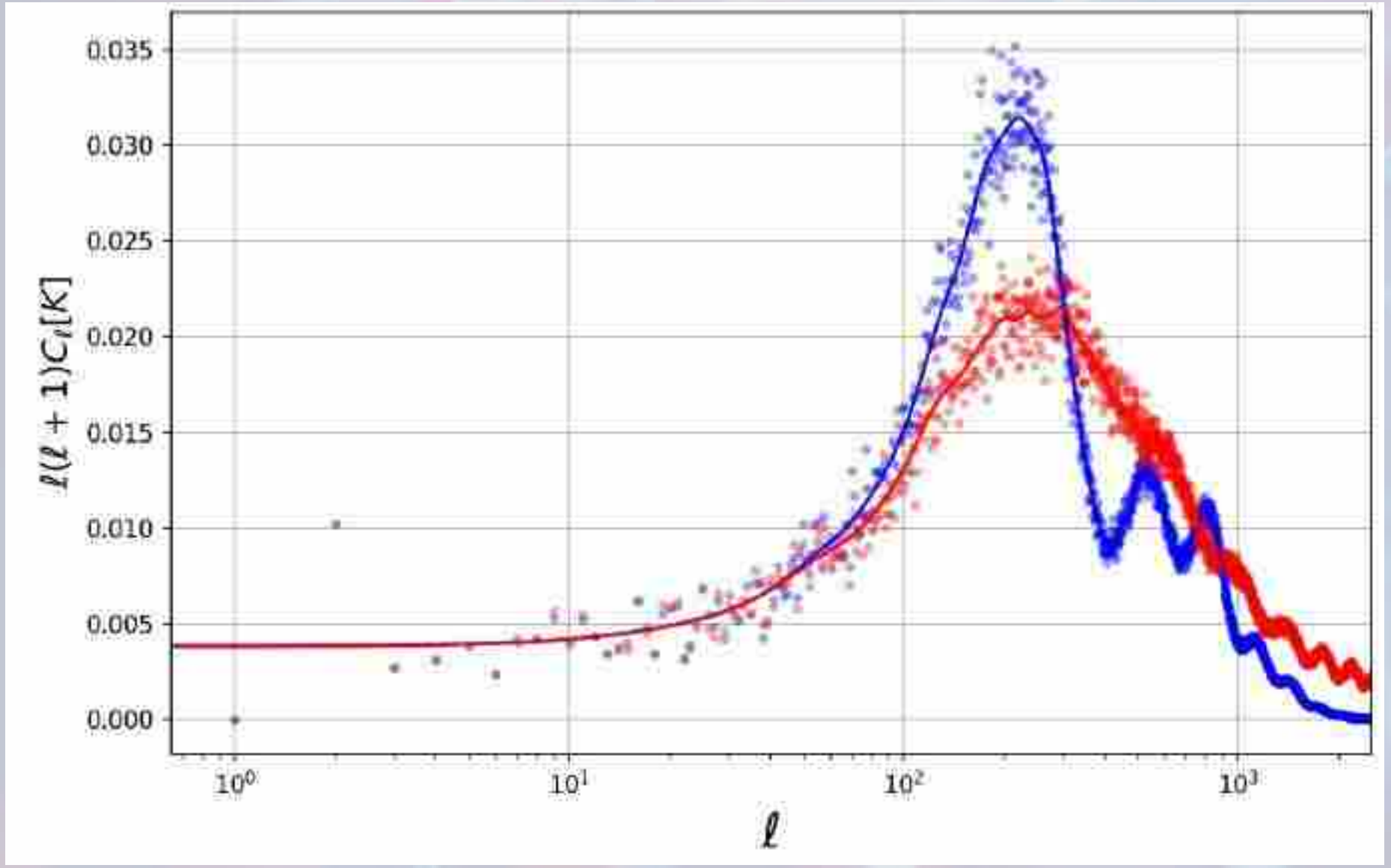
Inpainting CMB



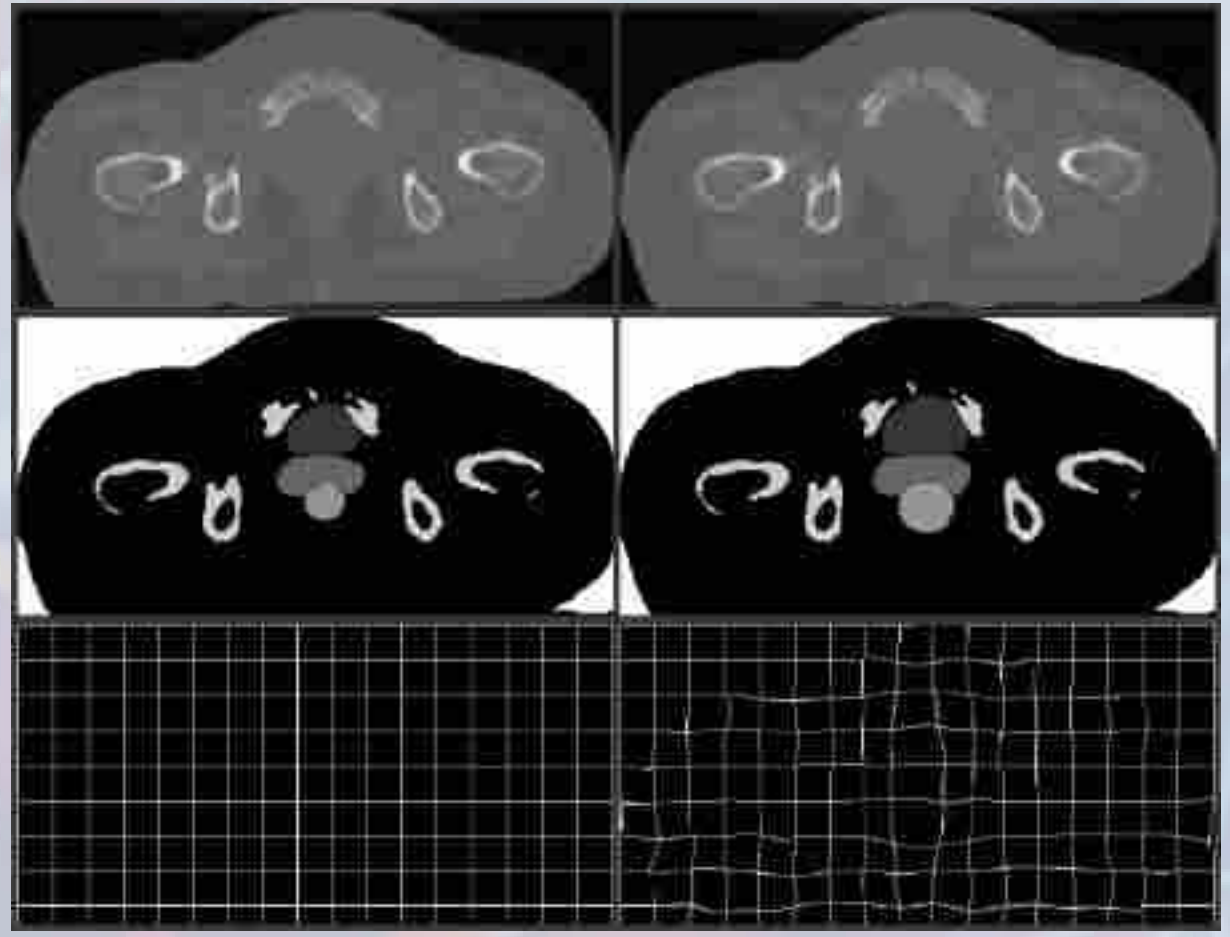
Inpainting CMB

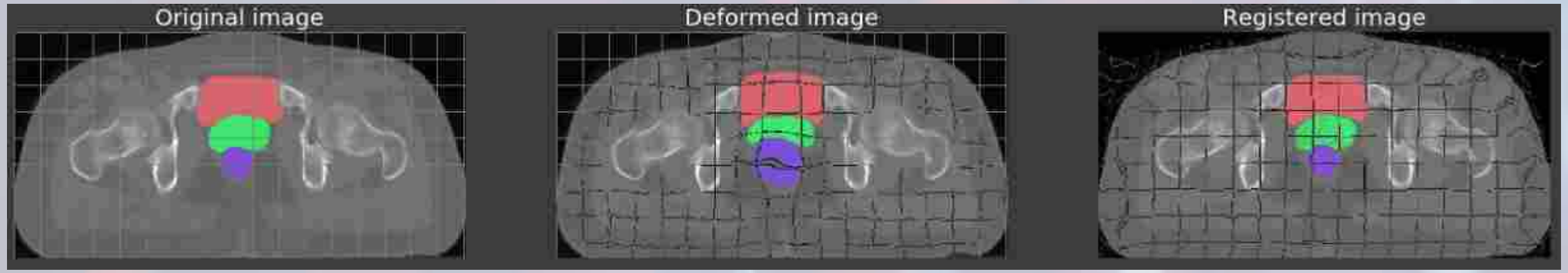


Inpainting CMB



Medical physics





Mining cosmic datasets

+some cool DS stuff

Alireza Vafaei Sadr

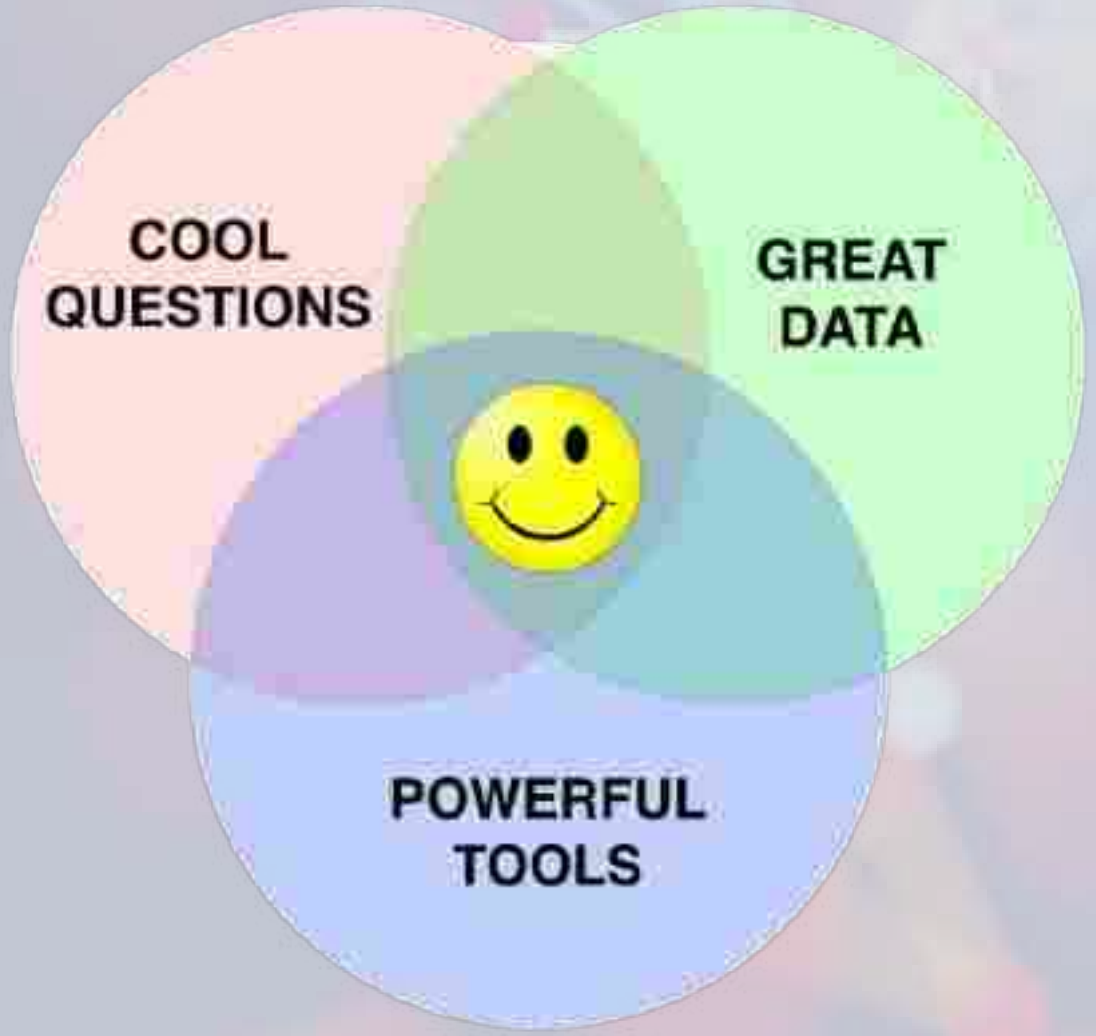
IPM, Tehran

Institute for Research in Fundamental Sciences

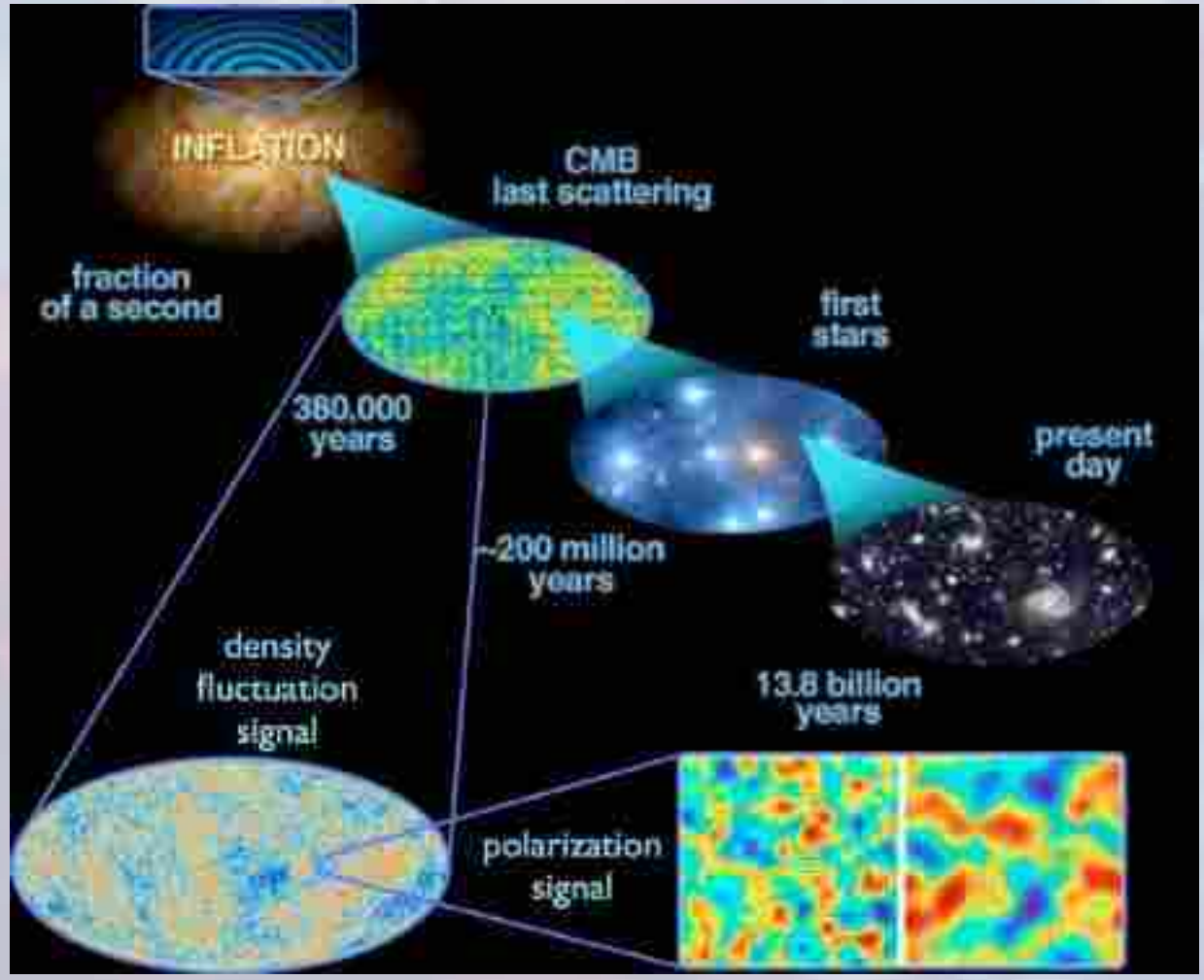
Geneva University , Jan, 2020

Outline

- ✓ Cosmology and BIG data
- ✓ A Quick Review of Applications
- ✓ Anomaly detection
- ✓ DRAMA
- ✓ Future directions

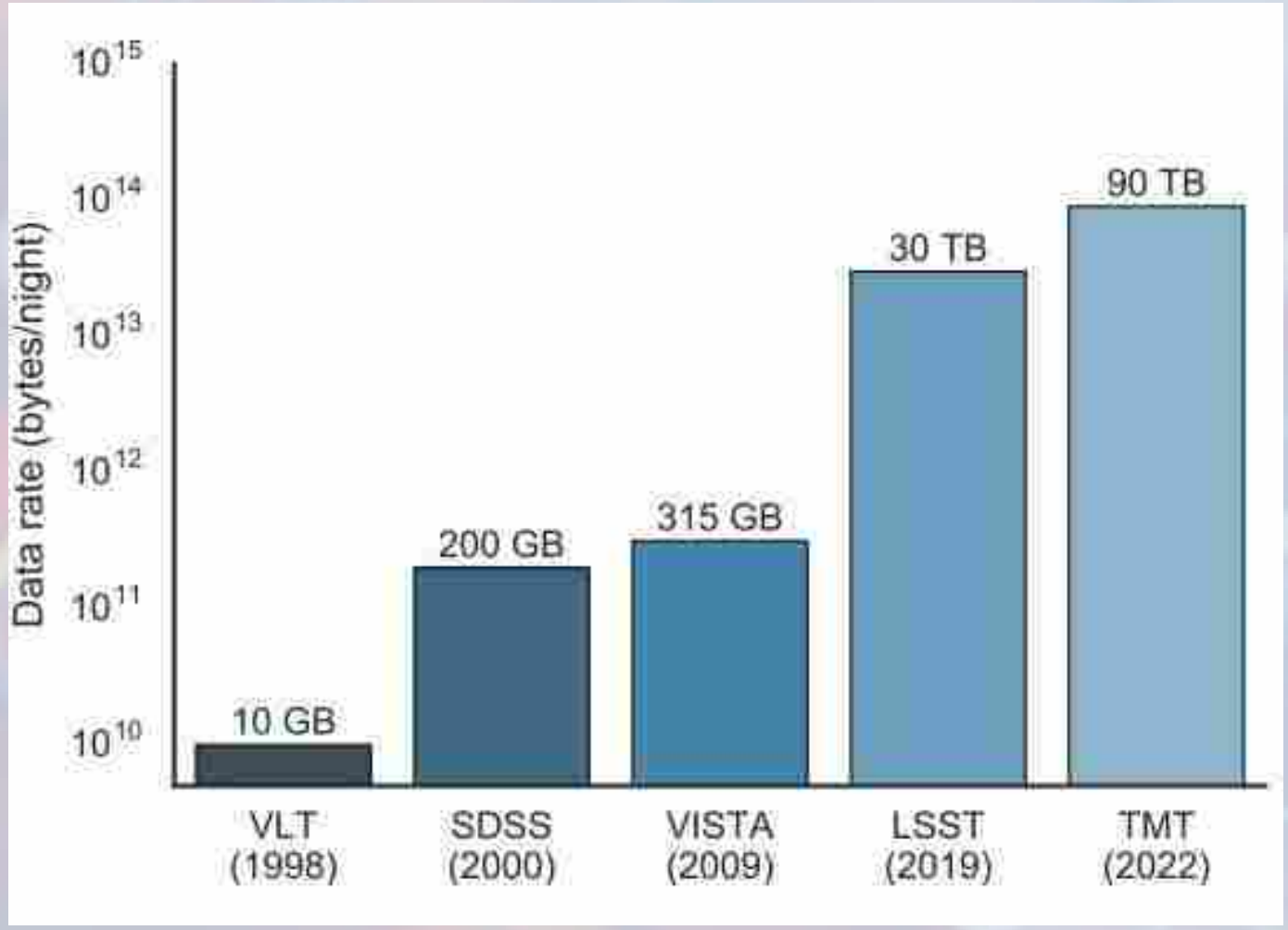

COOL
QUESTIONS
GREAT
DATA
POWERFUL
TOOLS

Cosmology/Astrophysics



E. Siegel, with images derived from ESA/Planck and the DoE/NASA/ NSF interagency task force on CMB research. From his book, Beyond The Galaxy.

Cosmology and Big data



Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy

Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel

LOFAR

ASKAP
Australian SKA Pathfinder

gaia

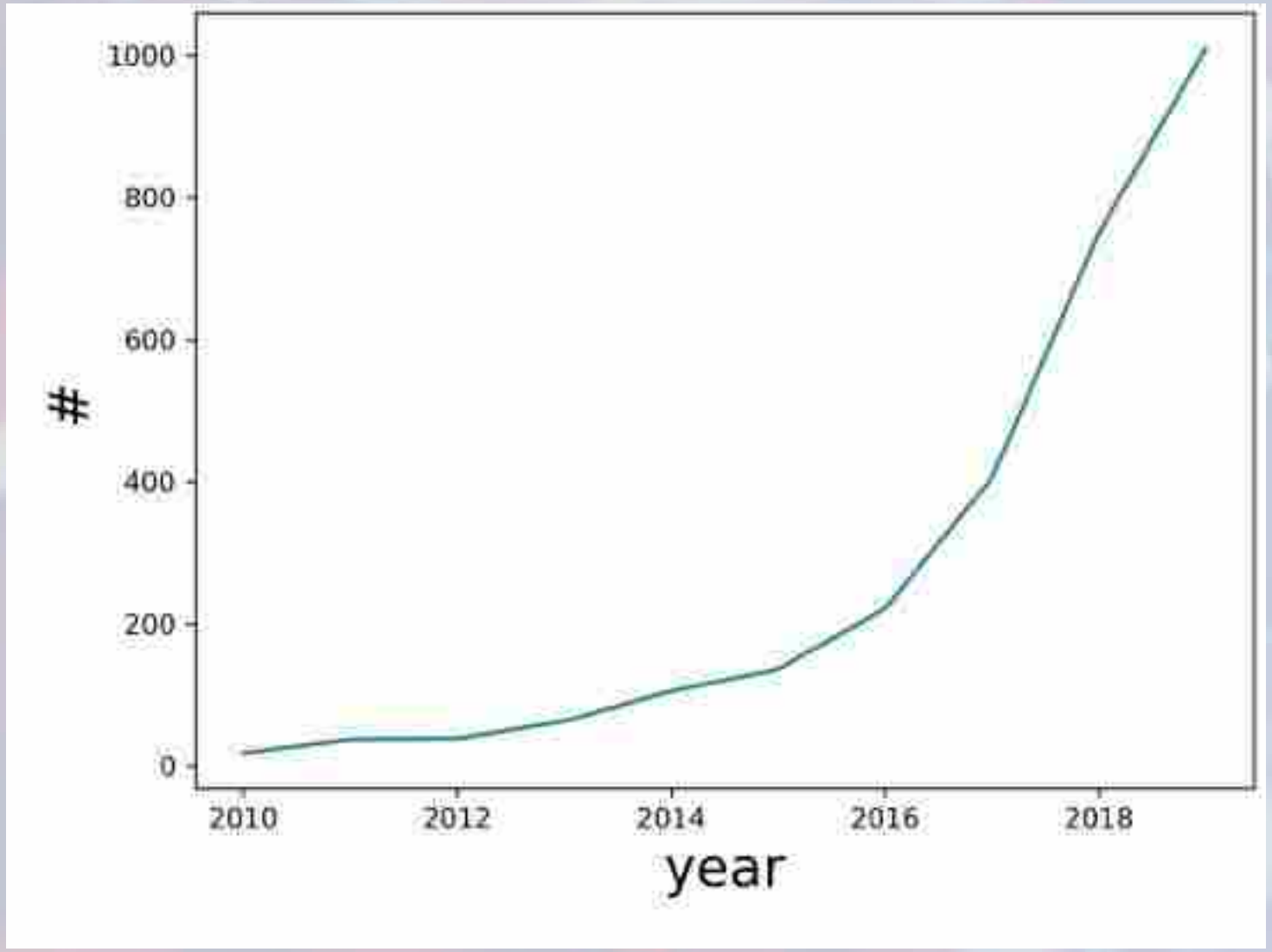
euclid

HIRAX

xmm-newton

LSST
Large Synoptic Survey Telescope

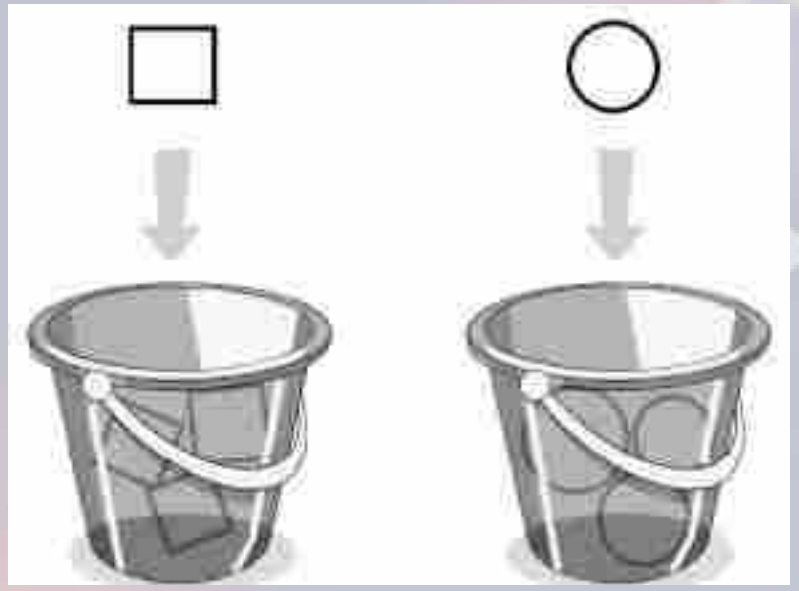
It is getting hotter!

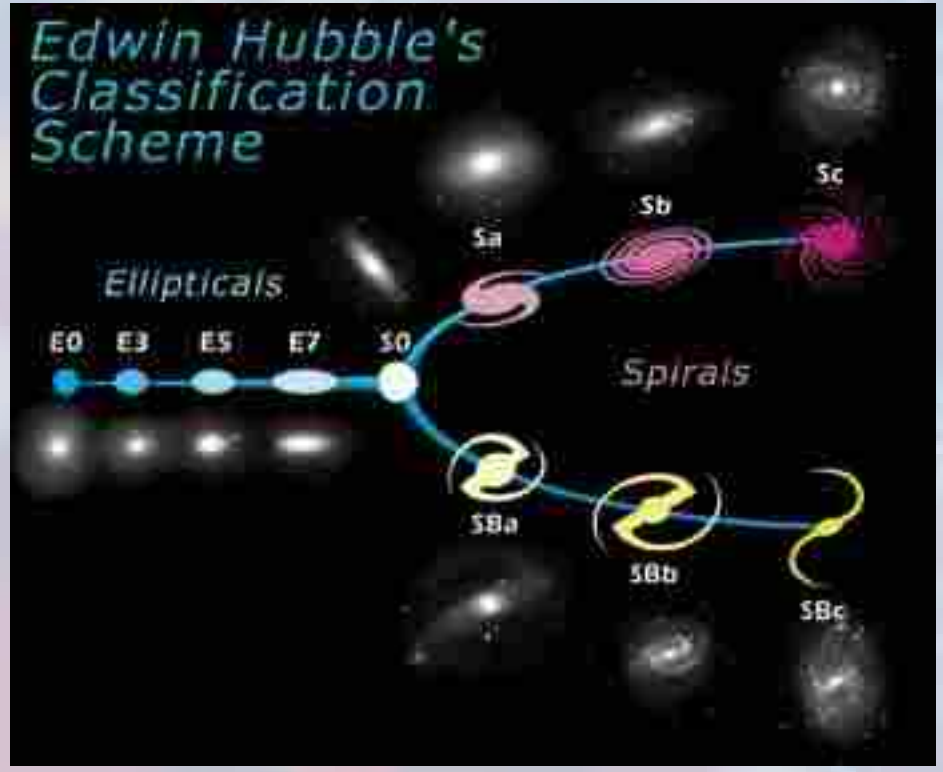


Number of physics submitted manuscripts that include “machine learning” in their abstracts.

A quick review on what people have done

Classification





http://cdn.spacetelescope.org/archives/images/screen/heic9902o.jpg

Galaxy zoo challenge

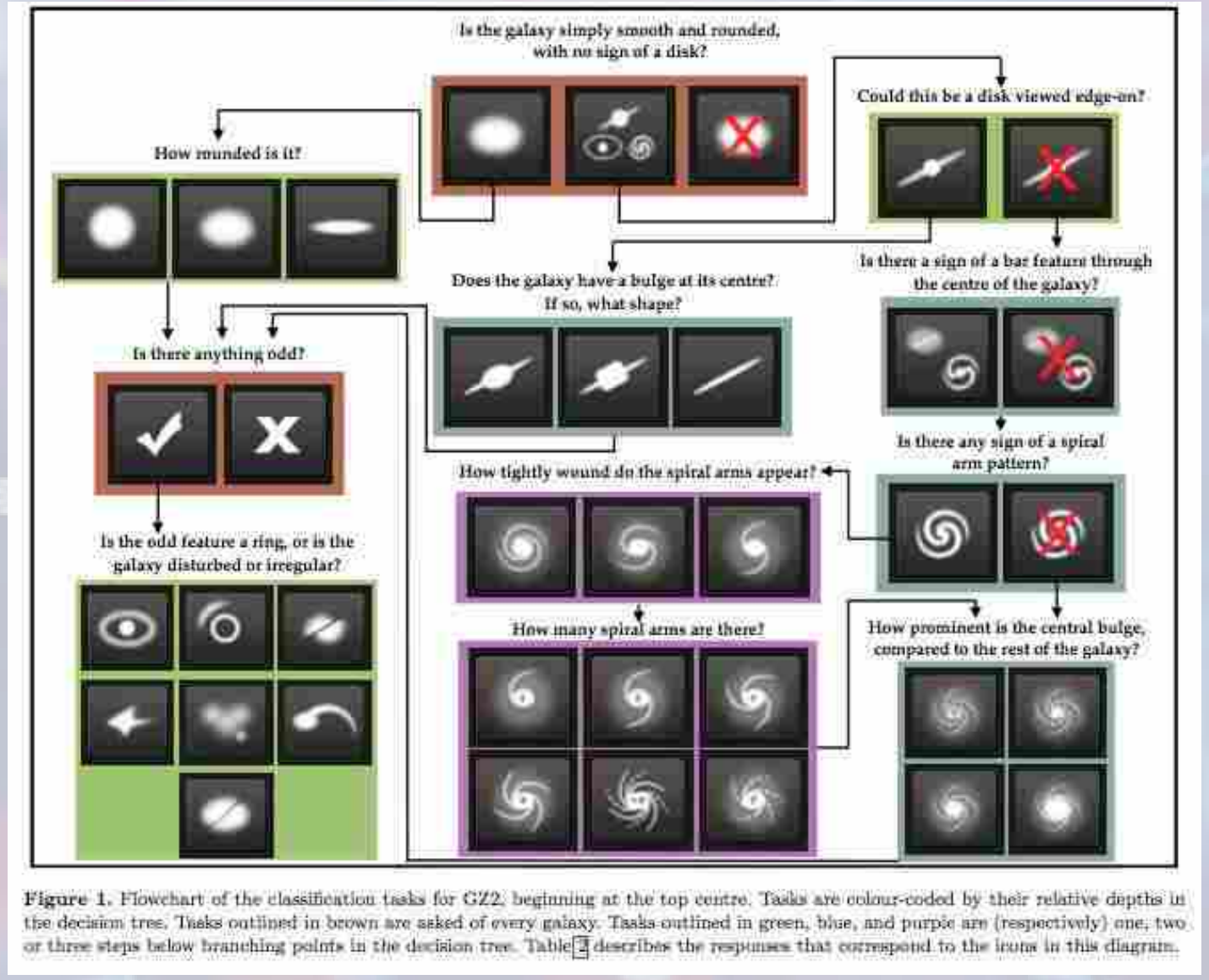
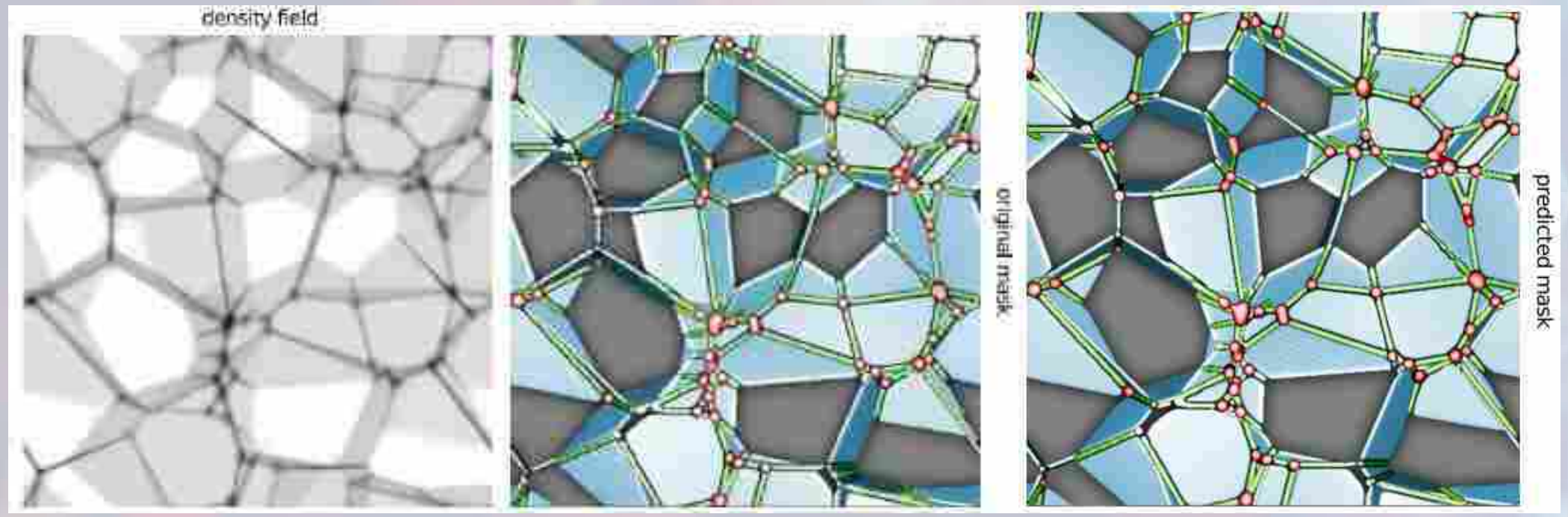


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 3 describes the responses that correspond to the icons in this diagram.

https://www.galaxyzoo.org/

Classifying the Large Scale Structure of the Universe with Deep Neural Networks

M.A. Aragon-Calvo[1] ⋆
[1] *Instituto de Astronomía, UNAM, Apdo. Postal 106, Ensenada 22800, B.C., México*



Detection

Galaxy detection and identification using deep learning and data augmentation.

Roberto E. González[a,b], Roberto P. Muñoz[a], Cristian A. Hernández[a]

Learn-As-You-Go Acceleration of Cosmological Parameter Estimates

Grigor Aslanyan, Richard Easther, and Layne C. Price

Deep Learning for Galaxy Mergers in the Galaxy Main Sequence

William J. Pearson[1,2], Lingyu Wang[1,2], James Trayford[3], Carlo E. Petrillo[1], and Floris F.S. van der Tak[1,2]

Deep Learning for Real-time Gravitational Wave Detection and Parameter Estimation with LIGO Data

Daniel George E. A. Huerta

Identifying Galaxy Mergers in Observations and Simulations with Deep Learning

W. J. Pearson[1,2], L. Wang[1,2], J.W. Trayford[3], C.E. Petrillo[1], F.F.S. van der Tak[1,2]

Estimating Photometric Redshifts for X-ray sources in the X-ATLAS field, using machine-learning techniques

G. Mountrichas[1], A. Corral[2,1], V. A. Masoura[3,1], I. Georgantopoulos[1], A. Ruiz[1], A. Georgakakis[1], F. J. Carrera[2], S. Fotopoulou[4]

Deterministic and Bayesian Neural Networks for Low-latency Gravitational Wave Parameter Estimation of Binary Black Hole Mergers

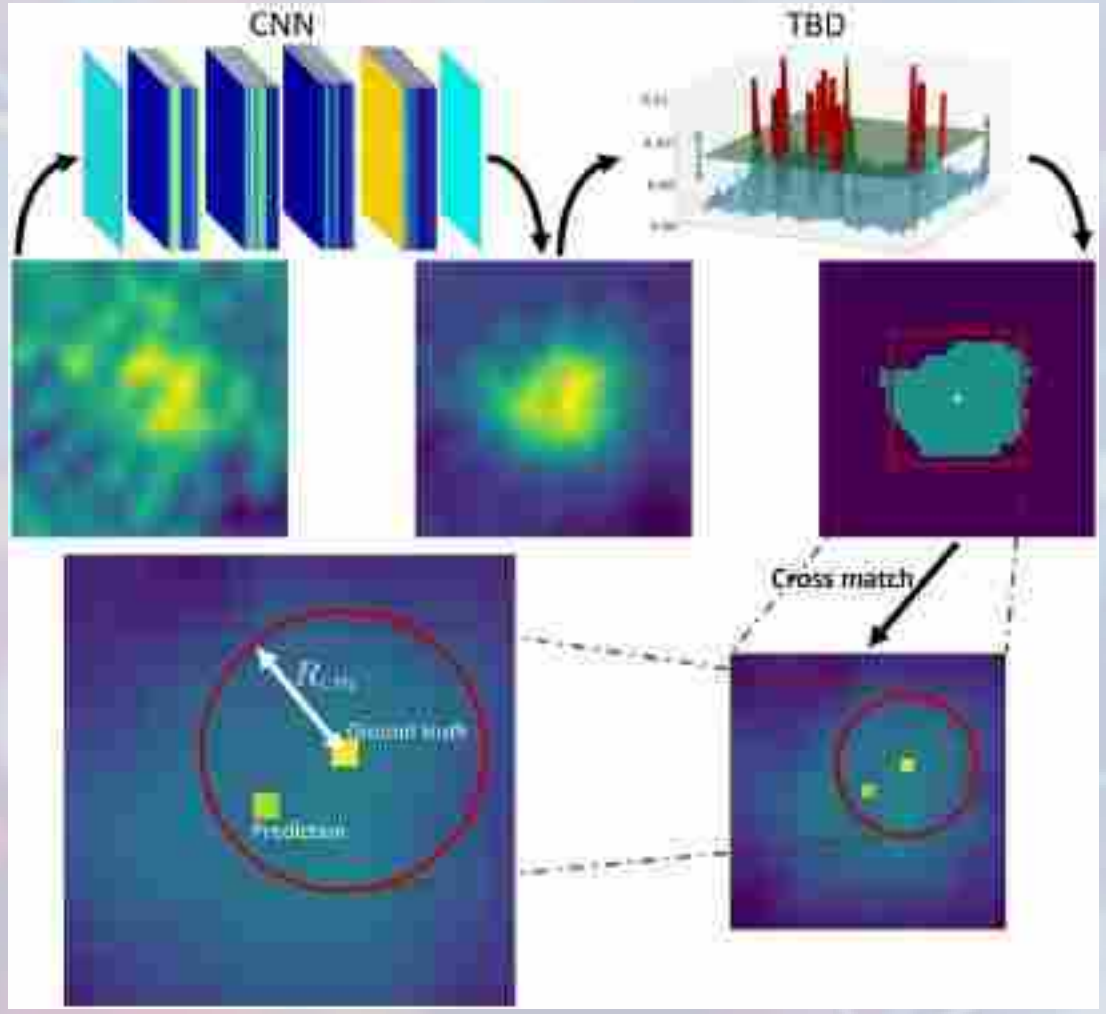
Hongyu Shen,[1,2] E. A. Huerta,[1,2] Zhizhen Zhao,[1,2,3] Eliu Jennings,[2] and Himanshu Sharma[2]

Cosmic String Detection with Tree-Based Machine Learning

A. Vafaei Sadr[1,2,3], M. Farhang[4], S. M. S. Movahed[1,4,*], B. Bassett[2,5,6,7], M. Kunz[2]

Morpheus: A Deep Learning Framework For Pixel-Level Analysis of Astronomical Image Data

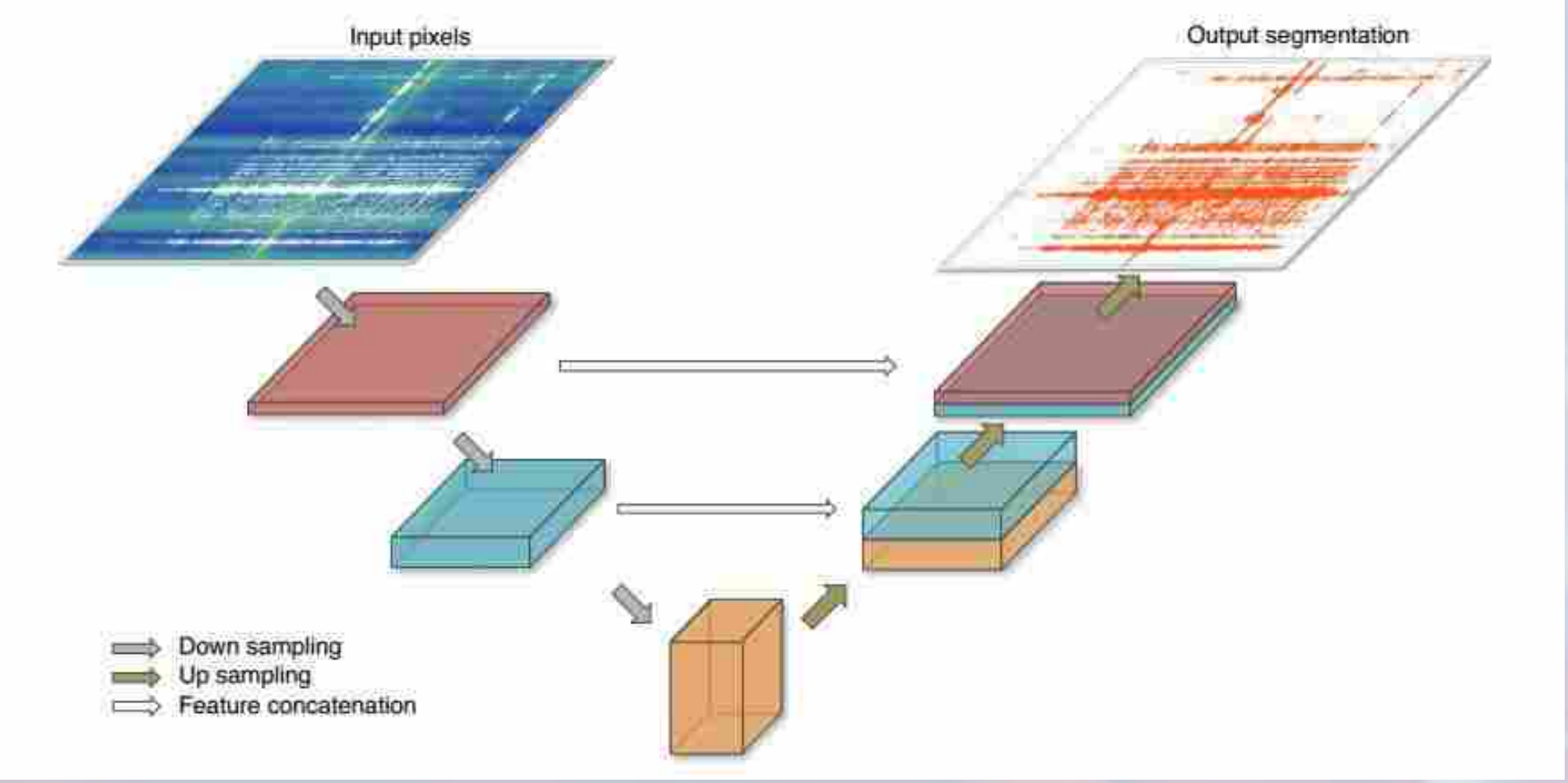
RYAN HAUSEN[1] AND BRANT E. ROBERTSON[2,1]



DeepSource: Point Source Detection using Deep Learning

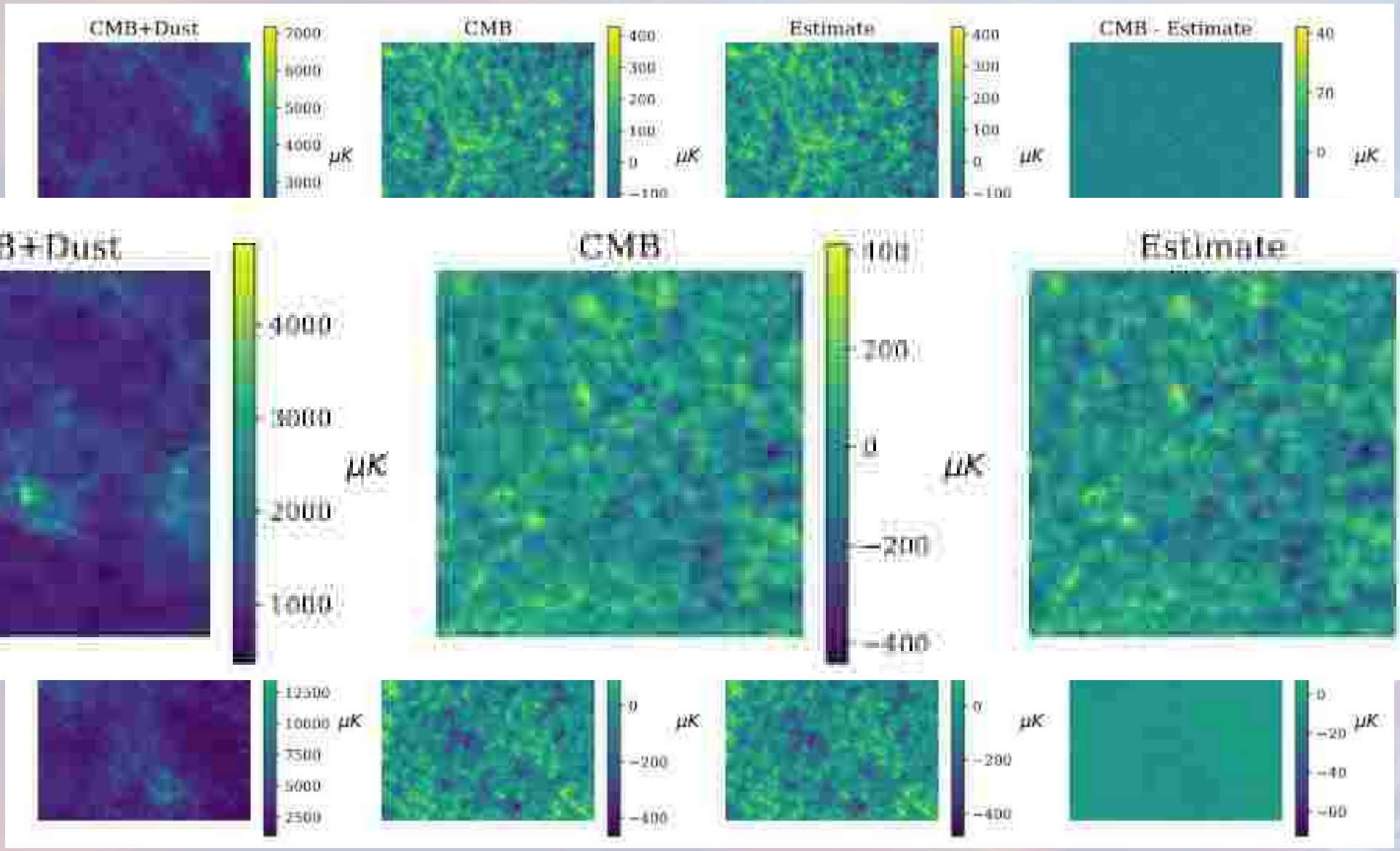
A. Vafaei Sadr,[1,2,3*] Etienne E. Vos,[2,4,5†] Bruce A. Bassett,[2,3,5,6†] Zafiirah Hosenie,[2,5,7] N. Oozeer,[2,3] and Michelle Lochner[2,3]

Data cleansing



Radio frequency interference mitigation using deep convolutional neural networks

Joël Akeret[a,*], Chihway Chang[a], Aurelien Lucchi[b], Alexandre Refregier[a]



Cleaning our own Dust: Simulating and Separating Galactic Dust Foregrounds with Neural Networks

K. Aylor,[1] M. Haq,[2] L. Knox,[1] Y. Hezaveh,[3,4] and L. Perreault-Levasseur[3,5,4]

Gravitational Wave Denoising of Binary Black Hole Mergers with Deep Learning

Wei Wei[a,b], E. A. Huerta[a,c]

Denoising Weak Lensing Mass Maps with Deep Learning

Masato Shirasaki,[1] Naoki Yoshida,[2,3,4] and Shiro Ikeda[5,6]

DENOISING GRAVITATIONAL WAVES WITH ENHANCED DEEP RECURRENT DENOISING AUTO-ENCODERS

Hongyu Shen[1] *Daniel George*[2] *Eliu. A. Huerta*[2,3] *Zhizhen Zhao*[1,3]

Separating the EoR signal with a convolutional denoising autoencoder: a deep-learning-based method

Weitian Li,[1★] Haiguang Xu,[1,2★] Zhixian Ma,[3] Ruimin Zhu,[4] Dan Hu,[1] Zhenghao Zhu,[1] Junhua Gu,[5] Chenxi Shan,[1] Jie Zhu[3] and Xiang-Ping Wu[5]

Solar image denoising with convolutional neural networks

C. J. Díaz Baso[1], J. de la Cruz Rodríguez[1], and S. Danilovic[1]

Simulation

CAN: Creative Adversarial Networks
Generating "Art" by Learning About Styles and
Deviating from Style Norms*

Ahmed Elgammal[1] Bingchen Liu[1] Mohamed Elhoseiny[2] Marian Mazzone[3]

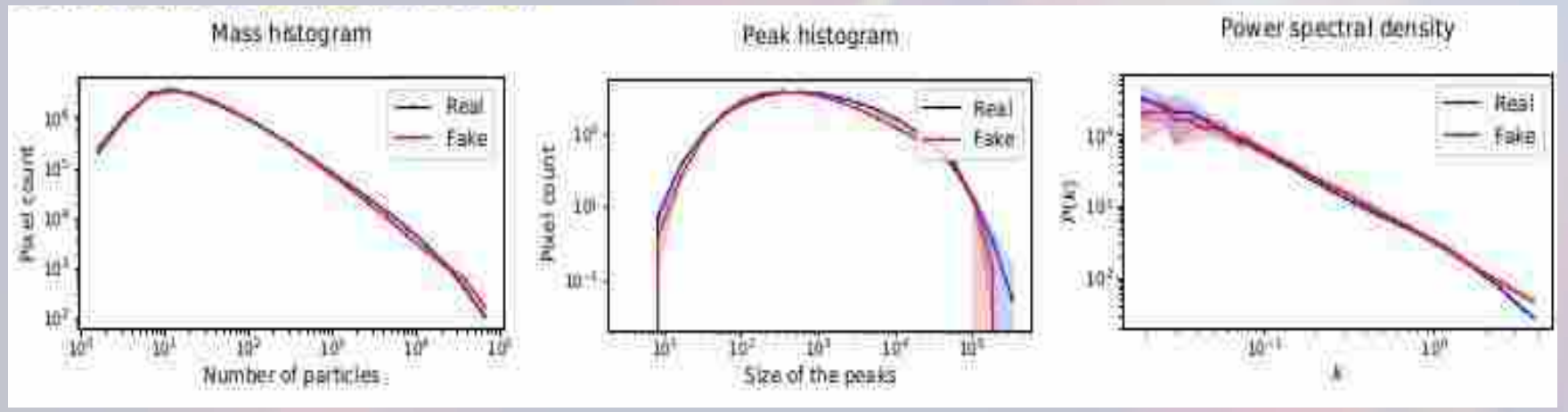
PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION

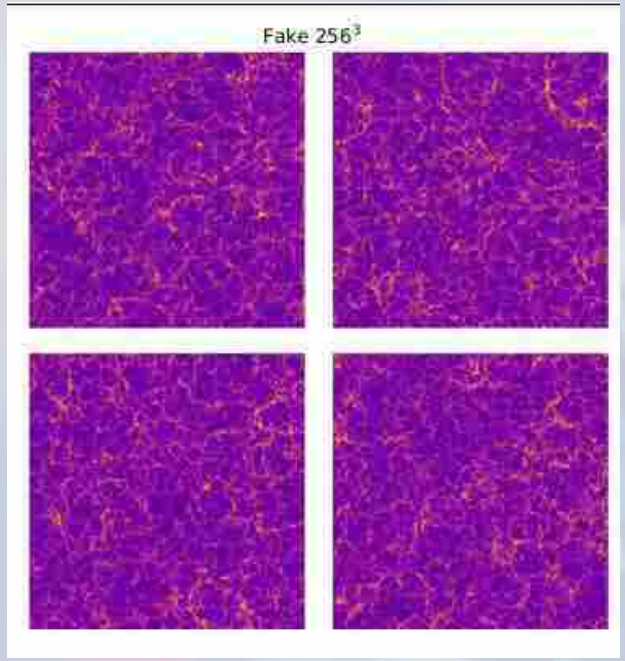
Tero Karras NVIDIA
Timo Aila NVIDIA
Samuli Laine NVIDIA
Jaakko Lehtinen NVIDIA and Aalto University

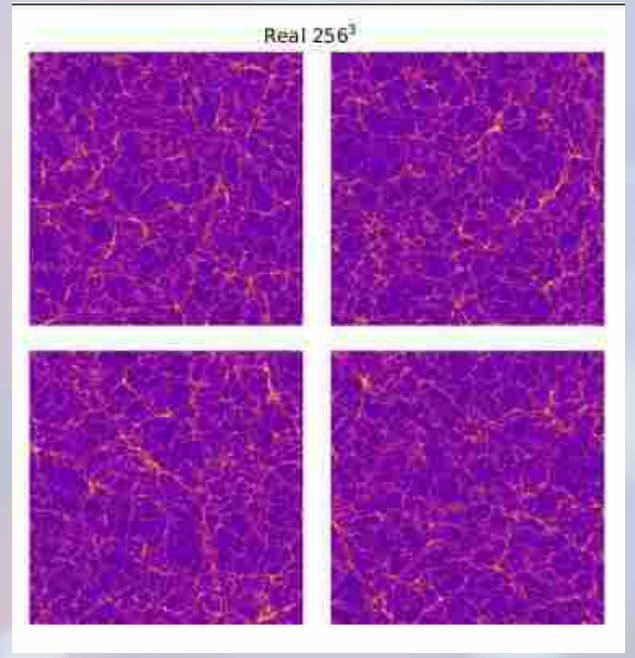
https://towardsdatascience.com/do-gans-really-model-the-true-data-distribution-or-are-they-just-cleverly-fooling-us-d08df69f25eb

Cosmological N-body simulations: a challenge for scalable generative models

Nathanaël Perraudin[1*], Ankit Srivastava[1], Aurelien Lucchi[2], Tomasz Kacprzak[3], Thomas Hofmann[2] and Alexandre Réfrégier[3]







From Dark Matter to Galaxies with Convolutional Neural Networks

Jacky H. T. Yip[1]
Xinyue Zhang[2], Yanfang Wang[2], Wei Zhang[2], Yueqiu Sun[2]
Gabriella Contardo, Francisco Villaescusa-Navarro, Siyu He, Shy Genel, Shirley Ho

Deep learning dark matter map reconstructions from DES SV weak lensing data

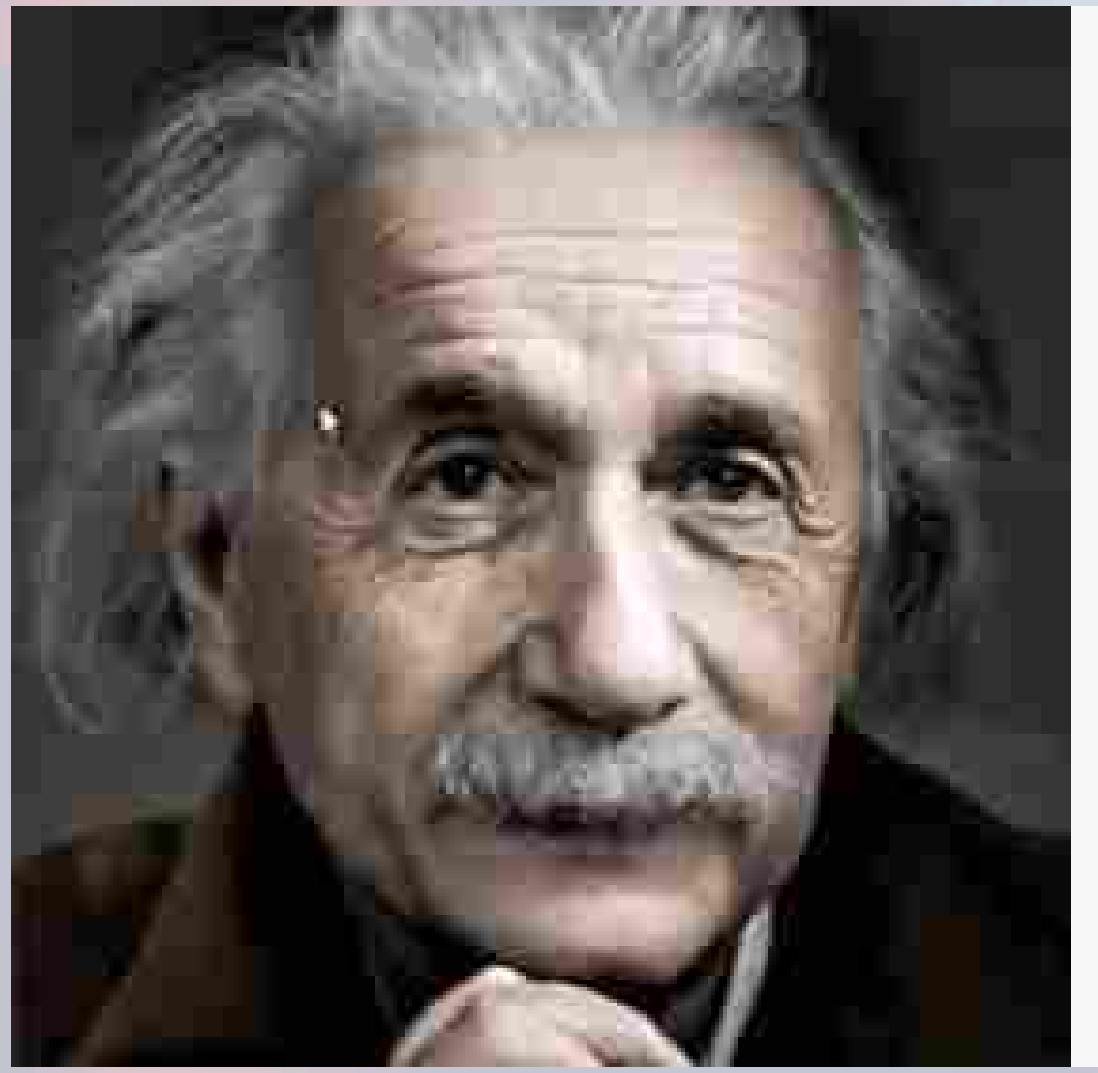
Niall Jeffrey,[1]* François Lanusse[2], Ofer Lahav[1], Jean-Luc Starck[3]

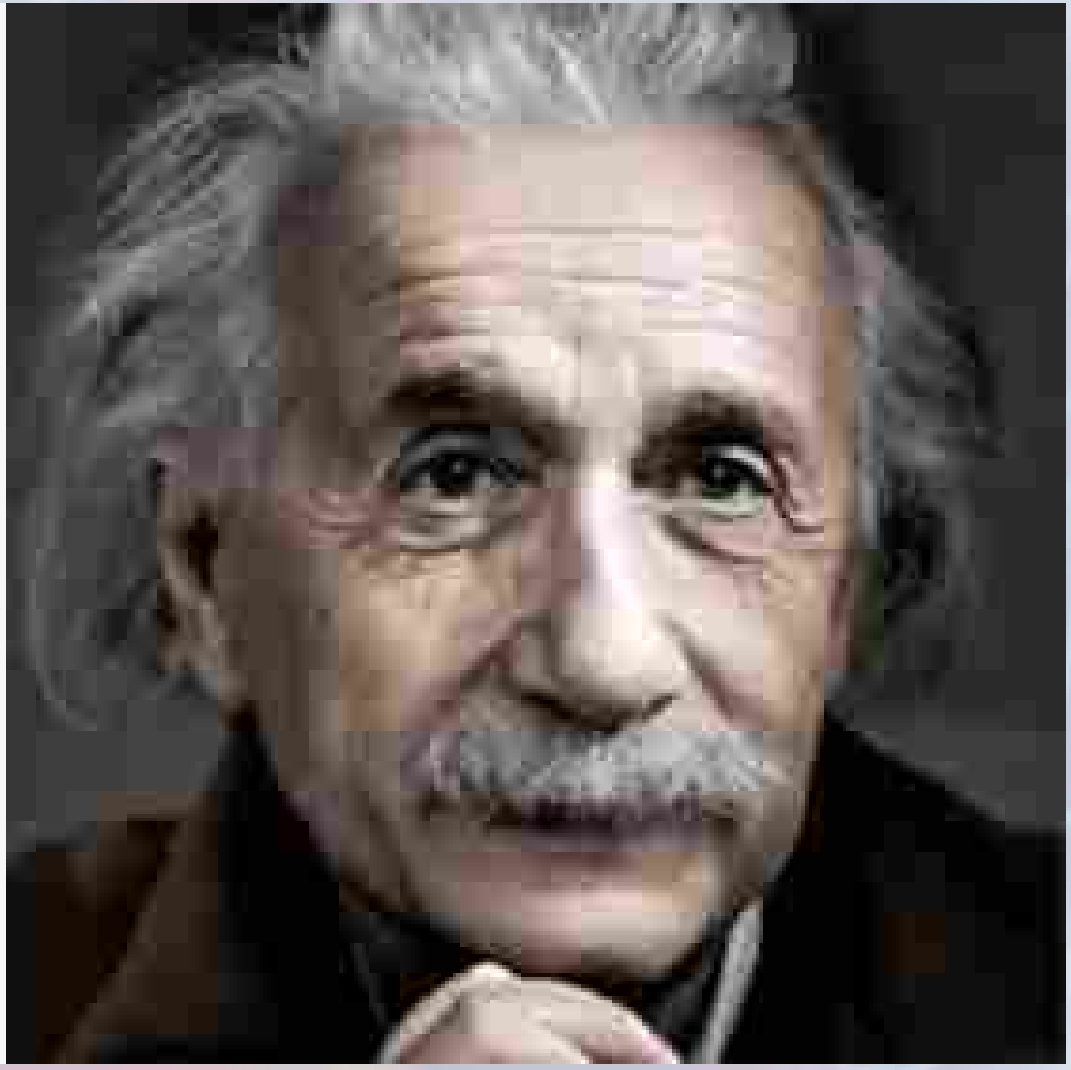
Learning to Predict the Cosmological Structure Formation

Siyu He, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos

CMB-GAN: Fast Simulations of Cosmic Microwave background Anisotropy maps using Deep Learning

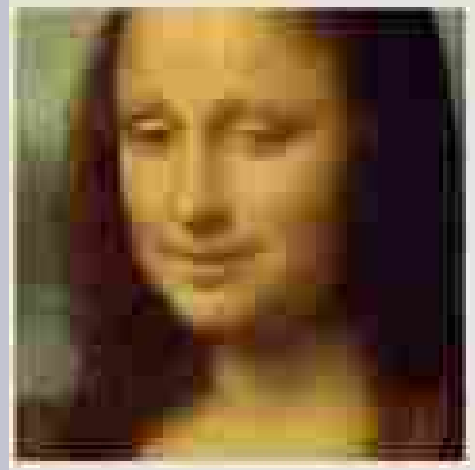
Amit Mishra[a], Pranath Reddy[a], Rahul Nigam[a]





SC-FEGAN: Face Editing Generative Adversarial Network with User's Sketch and Color

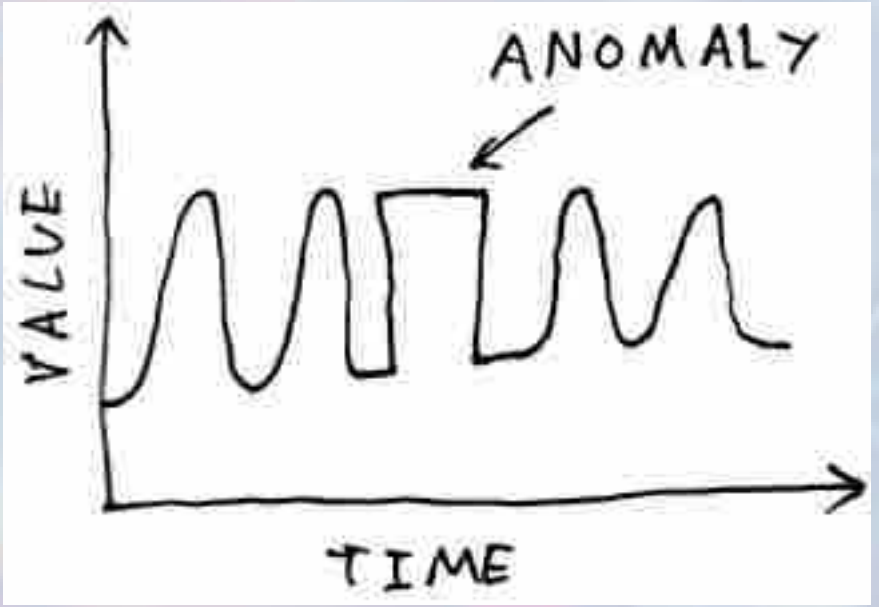
Youngjoo Jo Jongyoul Park

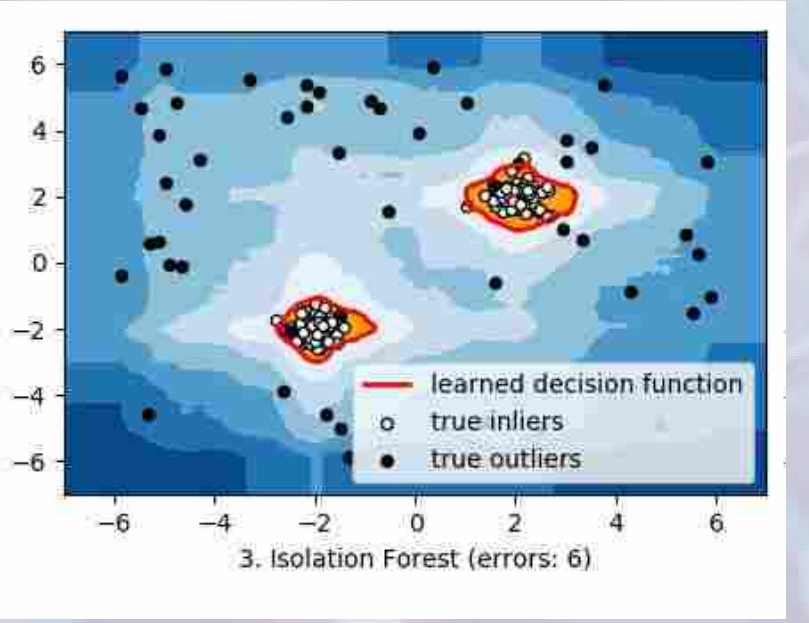



https://arxiv.org/pdf/1905.08233v1.pdf

Anomaly detection

The light from quasar pairs reach Earth, although some were absorbed by the gas in the cosmic web, Springel et al. (2005) (cosmic web) / J. Neidel, MPIA


ANOMALY
VALUE
TIME


6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
learned decision function
true inliers
true outliers
3. Isolation Forest (errors: 6)

FIND THE ODD ONE IN ~~60~~ 10 SECONDS

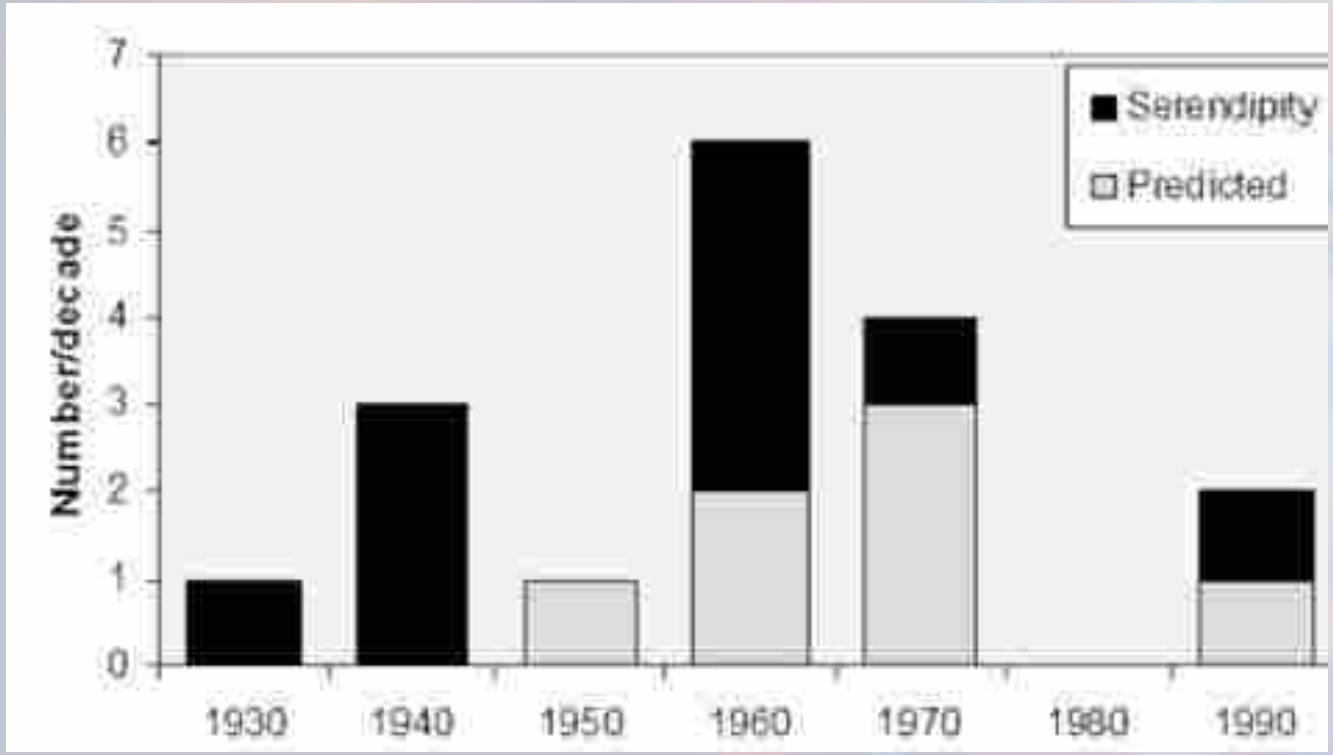


Figure 1 A plot of recent major astronomical discoveries, taken from (Ekers 2009), of which seven were "known-unknowns" (i.e. discoveries made by testing a prediction) and ten were "unknown-unknowns" (ie. a serendipitous result found by chance while performing an experiment with different goals). The data in this plot are taken from Wilkinson et al. (2004).

Norris, R. P. (2017). Discovering the unexpected in astronomical survey data. *Publications of the Astronomical Society of Australia*, *34*.

Table 1 Major discoveries made by the Hubble Space Telescope (*HST*). Of the *HST*'s "top ten" discoveries (as ranked by National Geographic magazine), only one was a key project used in the *HST* funding proposal (Lallo 2012). A further four projects were planned in advance by individual scientists but not listed as key projects in the *HST* proposal. Half the "top ten" *HST* discoveries were unplanned, including two of the three most cited discoveries, and including the only *HST* discovery (Dark Energy) to win a Nobel prize. This Table was previously published by Norris et al. (2015).

Project	Key Project?	Planned?	Nat Geo top ten?	Highly cited?	Nobel Prize?
Use cepheids to improve value of H_0	✓	✓	✓	✓	
UV spectroscopy of ig medium	✓	✓			
Medium-deep survey	✓	✓			
Image quasar host galaxies		✓	✓		
Measure SMBH masses		✓	✓		
Exoplanet atmospheres		✓	✓		
Planetary Nebulae		✓	✓		
Discover Dark Energy			✓	✓	✓
Comet Shoemaker-Levy			✓		
Deep fields (HDF, HDFS, GOODS, FF, etc)			✓	✓	
Proplyds in Orion			✓		
GRB Hosts			✓		

Neural network-based anomaly detection for high-resolution X-ray spectroscopy

Y. Ichinohe,[1,2*] and S. Yamada,[3]

Search for unusual objects in the WISE Survey

Aleksandra Solarz[1], Maciej Bilicki[2,1,3] and Agnieszka Pollo[1,4]

SELF-SUPERVISED ANOMALY DETECTION FOR NARROWBAND SETI

Yuhan Gerry Zhang[1], Kl Hyun Won[2], Seung Woo Son[3], Andrew Siemion[1,4,5], Steve Croft[1]

Anomaly detection for machine learning redshifts applied to SDSS galaxies

Ben Hoyle[1,2], Markus Michael Rau[1,4], Kerstin Paech[1,2], Christopher Bonnett[3] Stella Seitz[1,4], Jochen Weller[1,2,4]

Active Anomaly Detection for time-domain discoveries

E. E. O. Ishida[1*], M. V. Kornilov[2,3†], K. L. Malanchev[2,3], M. V. Pruzhinskaya[2], A. A. Volnova[4], V. S. Korolev[5,6], F. Mondon[1], S. Sreejith[1], A. Malancheva[7] and S. Das[8]

Anomaly Detection in the Open Supernova Catalog

M. V. Pruzhinskaya,[1*] K. L. Malanchev,[1,2†] M. V. Kornilov,[1,2] E. E. O. Ishida,[3] F. Mondon,[3] A. A. Volnova[4] and V. S. Korolev[5,6]

Current projects in SKA (MeerKAT) DS team:

- Source detection
  (Mightee, SKAch-I)
- Anomaly detection
  WTF, PLAsTiCC
- RFI simulation/mitigation
- Extended source simulation
- Observation strategy

A Flexible Framework for Anomaly Detection via Dimensionality Reduction

A. Vafaei Sadr, B. Bassett, M. Kunz

ISCMI-2019

No Free Lunch Theorems

- Any two optimization algorithms are equivalent when their performance is averaged across all possible problems
- No anomaly detection algorithm works for all anomalies
- No anomaly algorithm is “best” **on average.**
- Different algorithms work for different anomalies
- **So lets consider families of anomaly algorithms**

Dimensionality

Reduction

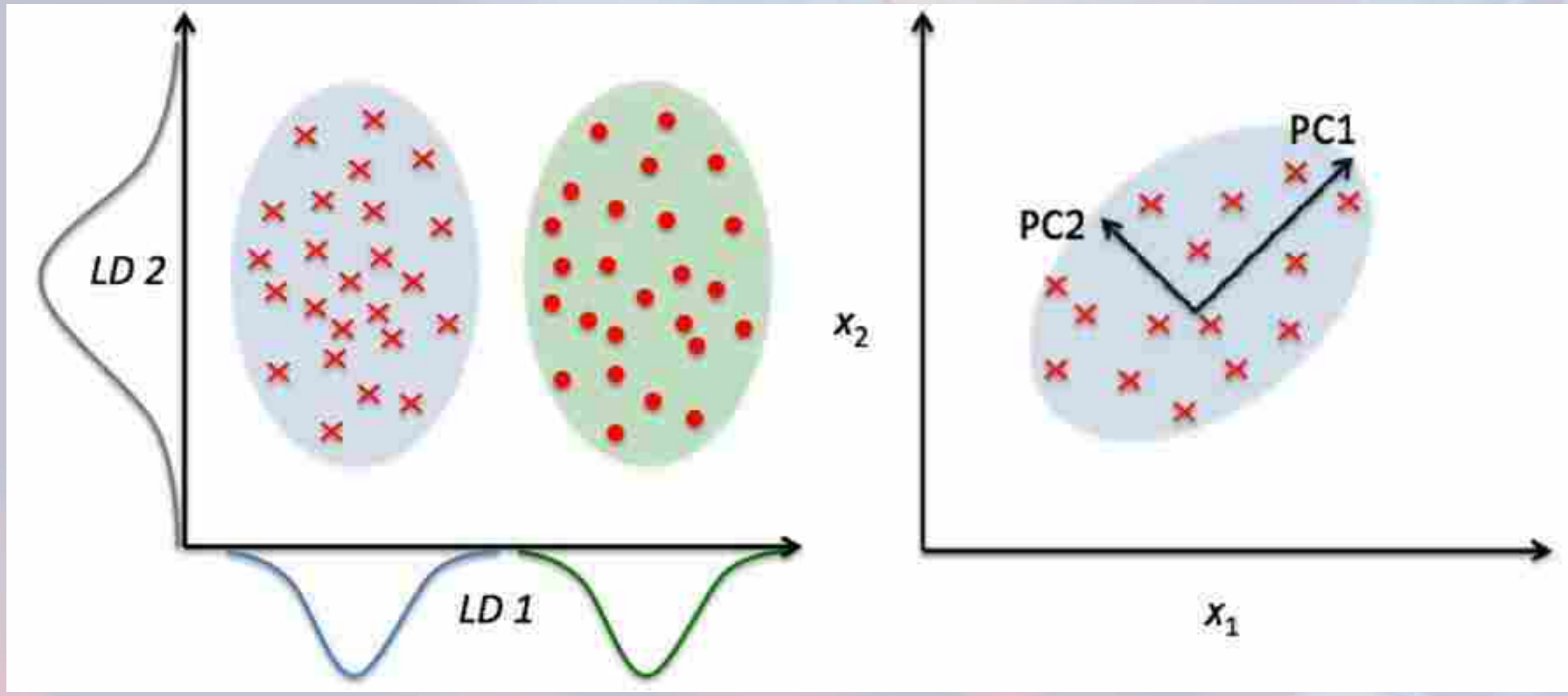
Anomaly

Meta

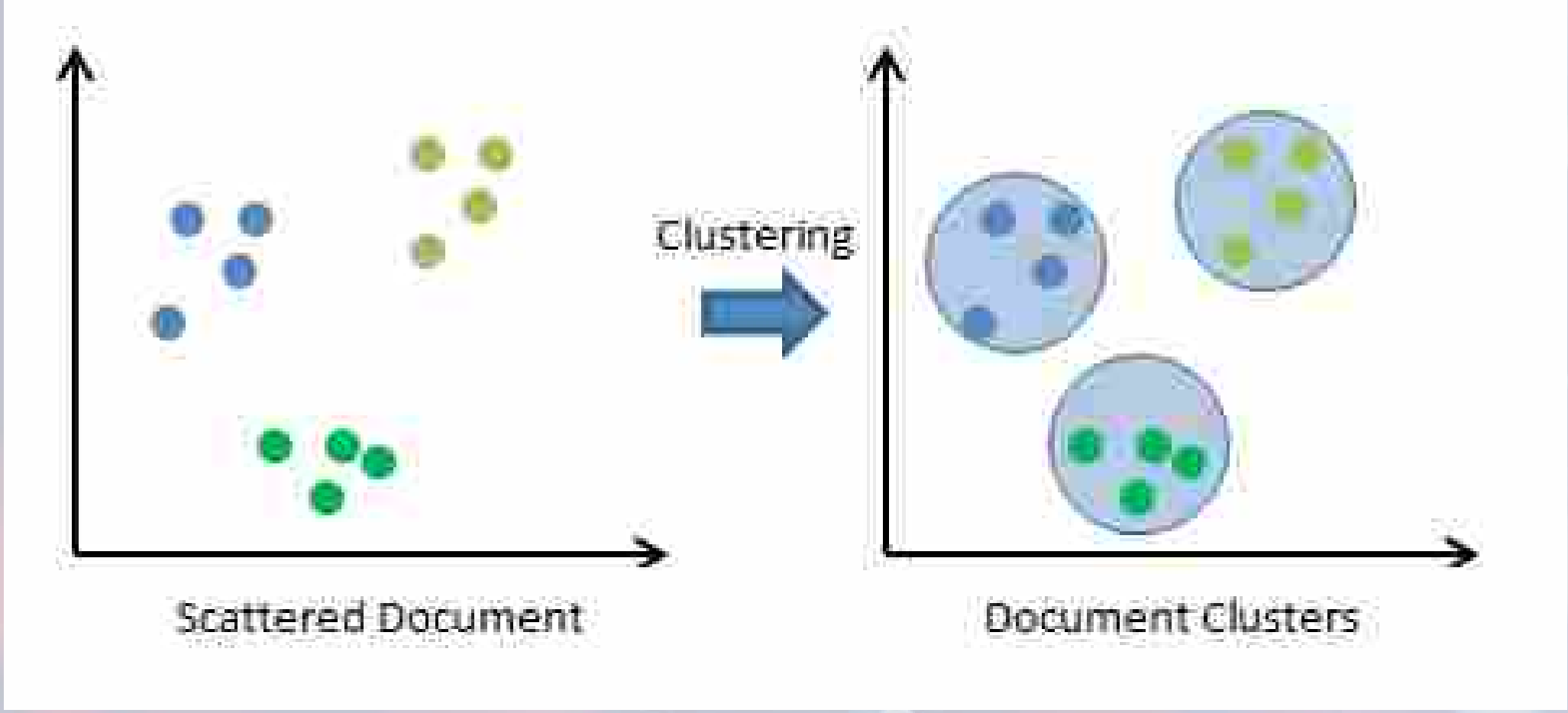
Algorithm

https://github.com/vafaei-ar/drama

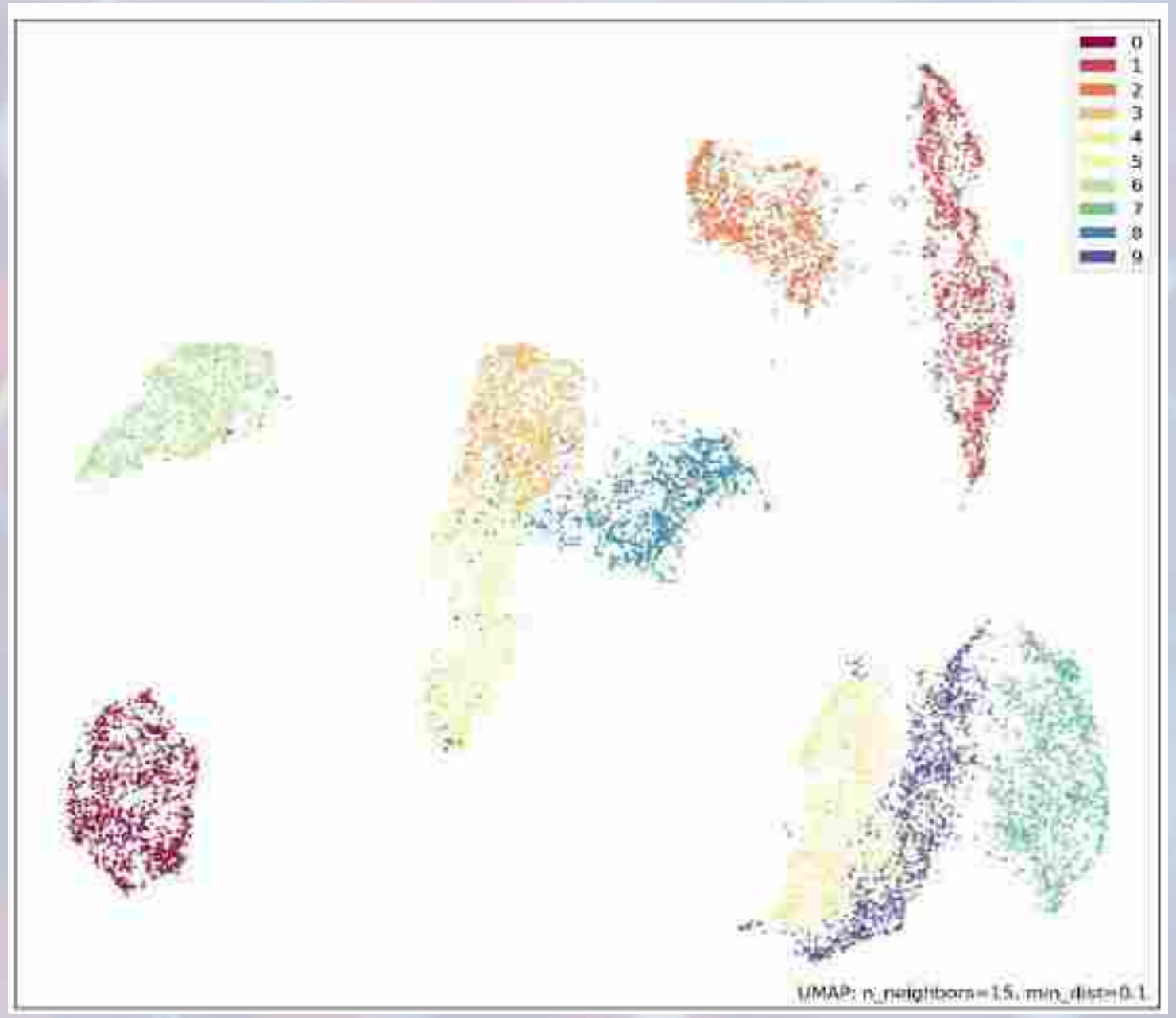
Dimensionality reduction



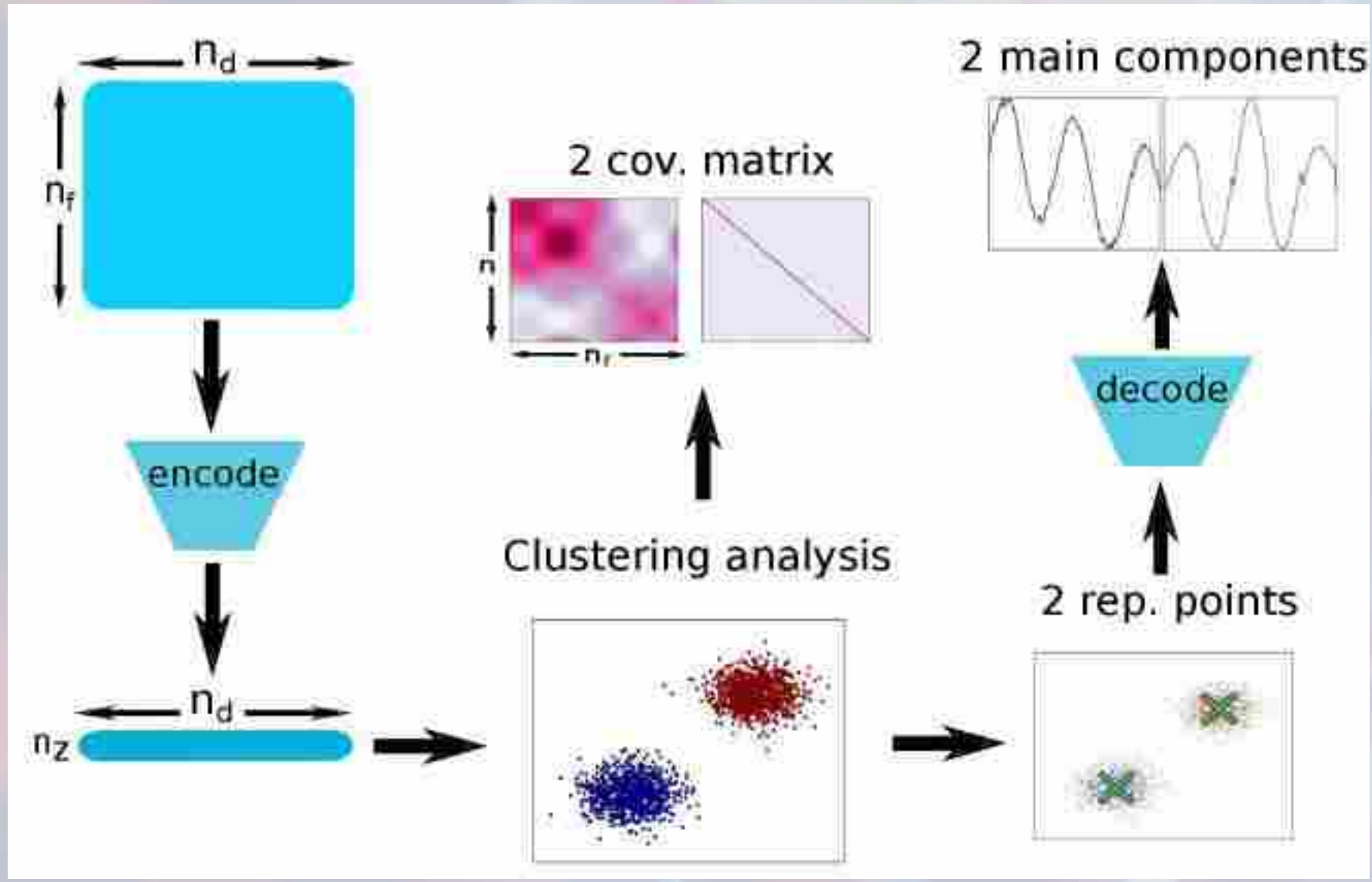
Clustering

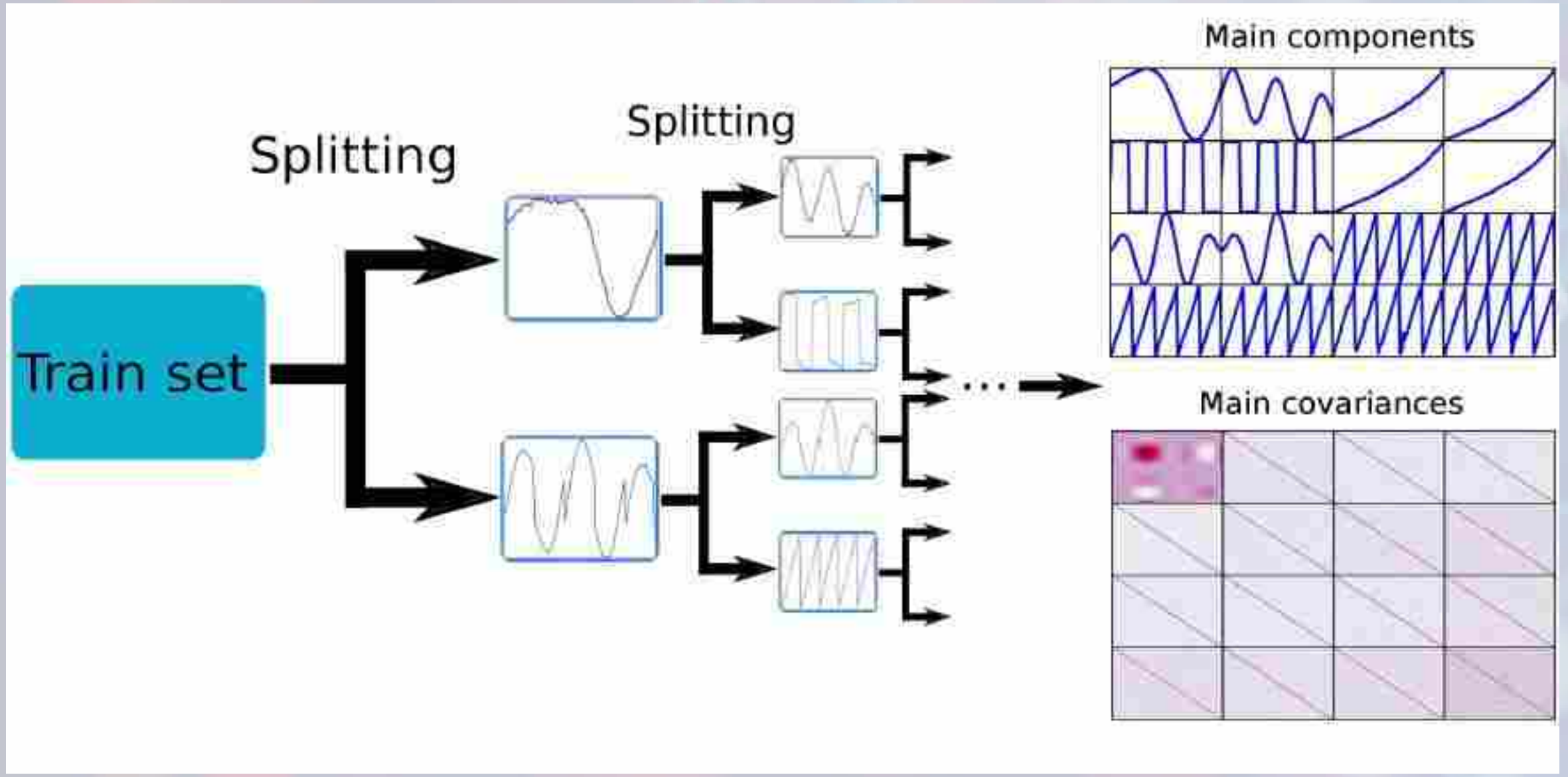


As an example: MNIST



DRAMA (Dimensionality Reduction Anomaly Meta-Algorithm):




Train set
Splitting
Splitting
Main components
Main covariances

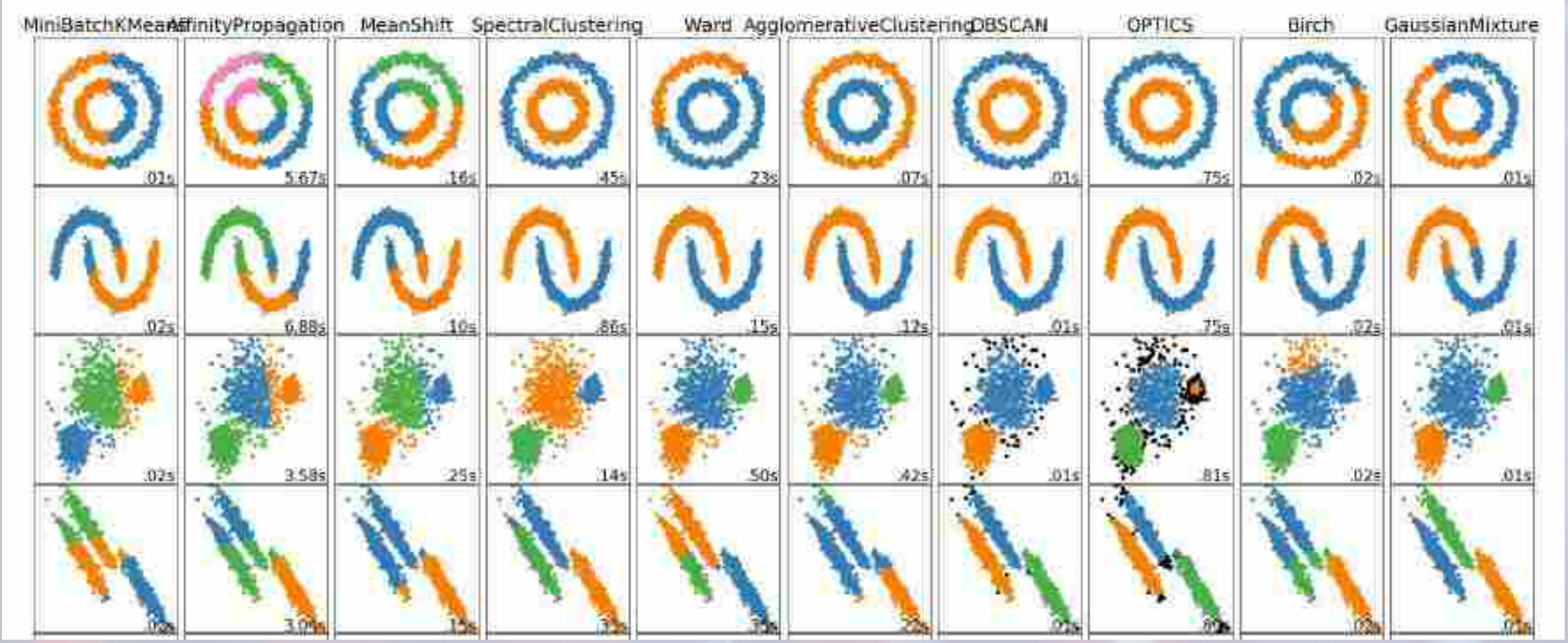
Dimensionality Reduction Technique

- Autoencoder
- Variational autoencoder
- principal component analysis
- independent component analysis
- non negative matrix factorization

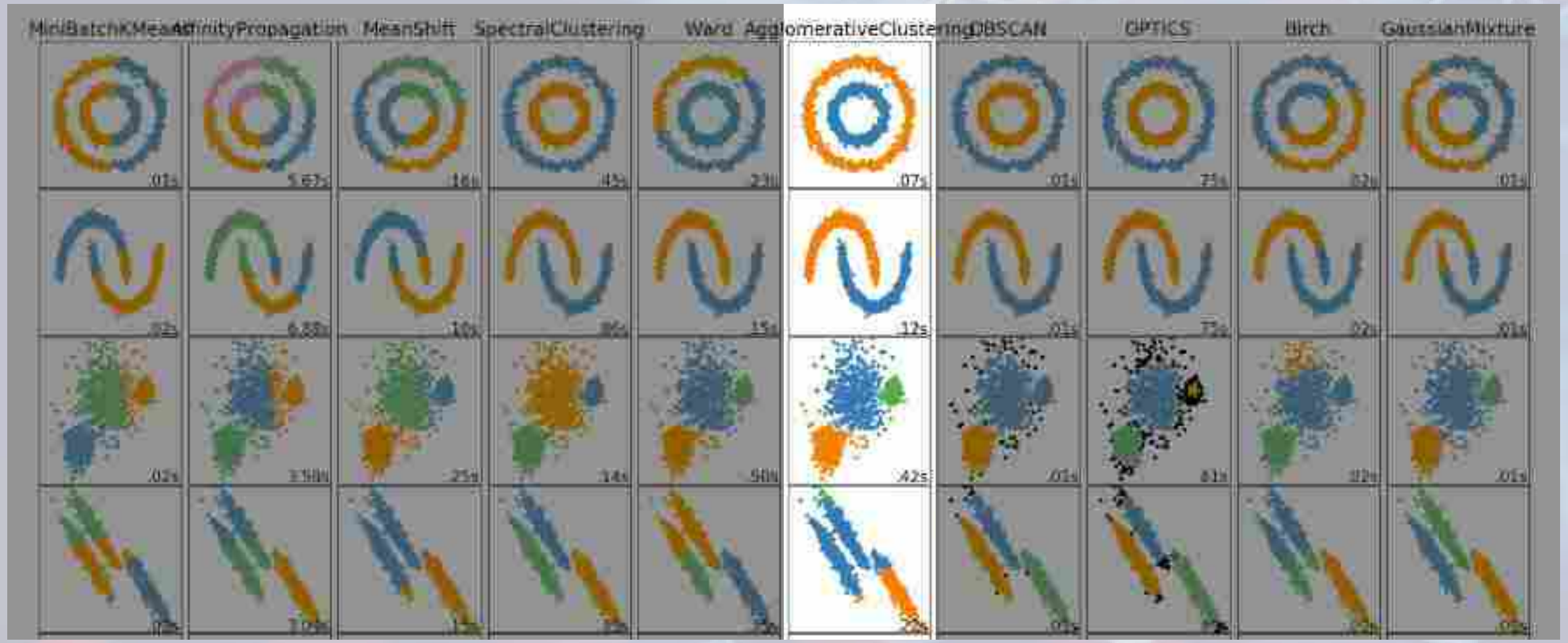
Also newly added:

- Convolutional (V)AE (1D)
- Convolutional (V)AE (2D)
- Convolutional UMAP

Clustering


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture

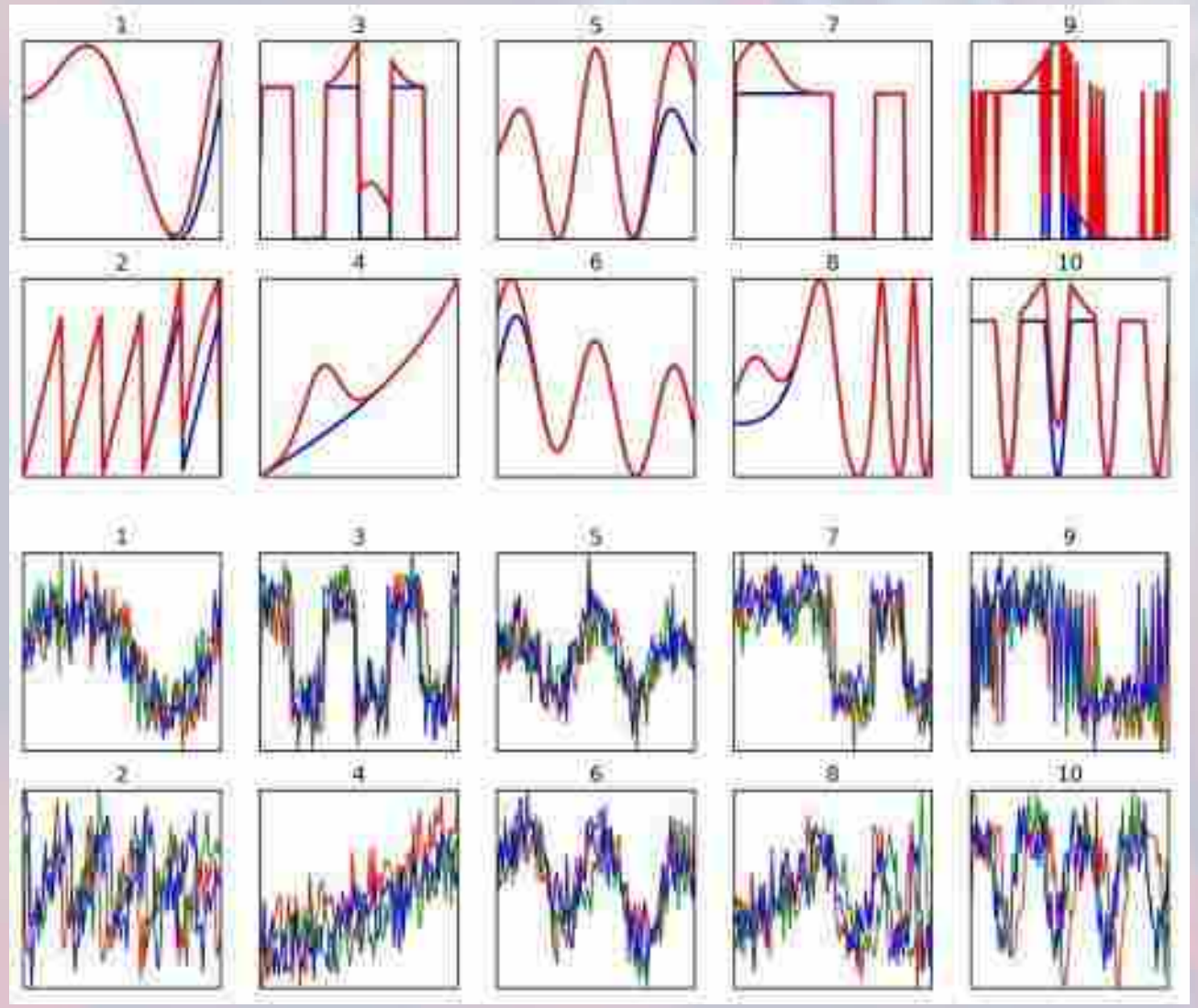
Clustering


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture

Distance metrics

Metric	definition
L1	$\sum_i \lvert u_i - v_i \rvert$
L2	$\lvert\lvert u - v \rvert\rvert_2$
L4	$\lvert\lvert u - v \rvert\rvert_4$
wL2	$\lvert\lvert \frac{u-v}{\sigma} \rvert\rvert_2$
wL4	$\lvert\lvert \frac{u-v}{\sigma} \rvert\rvert_4$
Bray-Curtis	$\sum \lvert u_i - v_i \rvert / \sum \lvert u_i + v_i \rvert$
Chebyshev	$\max_i \lvert u_i - v_i \rvert$
Canberra	$\sum_i \frac{\lvert u_i - v_i \rvert}{\lvert u_i \rvert + \lvert v_i \rvert}$
correlation	$1 - \frac{(u-\bar{u}) \cdot (v-\bar{v})}{\lvert\lvert (u-\bar{u}) \rvert\rvert_2 \lvert\lvert (v-\bar{v}) \rvert\rvert_2}$
Mahalanobis	$\sqrt{(u-v)C^{-1}(u-v)^T}$

Comparison with LOF and i-forest:



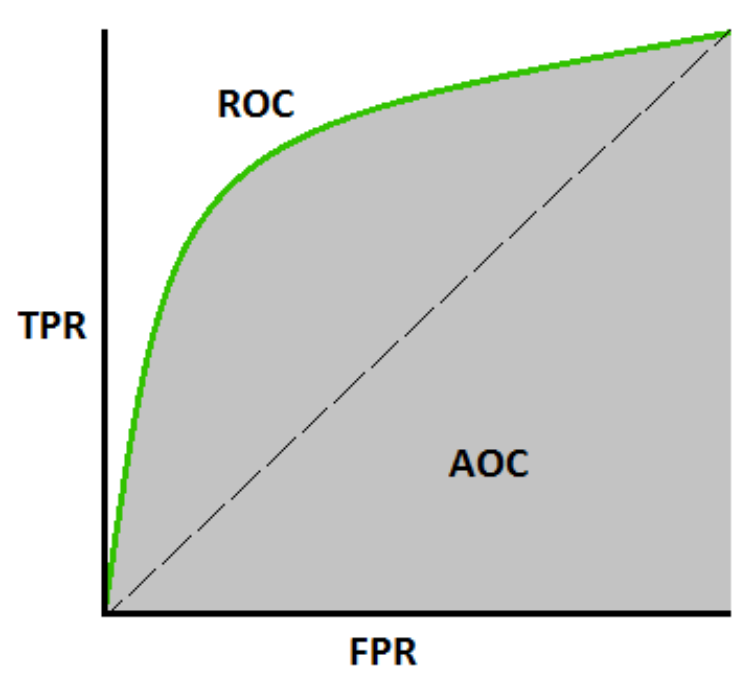
Simulated

Table 2: Real data summary.

Data set	# points	# dim.	# outliers
lympho	148	18	6
breastw	683	9	239
wine	129	13	10
vertebral	240	6	30
glass	214	9	9
pima	768	8	268
letter	1600	32	100
thyroid	3772	6	93
ionosphere	351	33	126
cardio	1831	21	176
wbc	378	30	21
annthyroid	7200	6	534
arrhythmia	452	274	66
vowels	1456	12	50
satellite	6435	36	2036
satimage-2	5803	36	71
optdigits	5216	64	150
mammography	11183	6	260
shuttle	49097	9	3511
mnist	7603	100	700
pendigits	6870	16	156
musk	3062	166	97
smtp	95156	3	30
http	567498	3	2211
cover	286048	10	2747
speech	3686	400	61

Real

Benchmark metrics:



$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$RWS = \frac{1}{N(N+1)} \sum_{i=1}^{N} w_i I_i$$

Table 6: Real dataset scores (all ×100). You can see ensemble method results for each DRTs as first five columns (best is yellow). The next four columns are the recommended DRT, splitting level and metrics. LOF results (best for number of neighbors equals 5,10 and 35) and isolated forest, respectively (best is blue). Errors are 1 standard deviation and are only shown if larger than unity.

		AUC(%)			MCC(%)			RWS(%)		
Dataset	DRT-metric	DRAMA	LOF	i-Forest	DRAMA	LOF	i-Forest	DRAMA	LOF	i-Forest
arrhythmia	NMF-L2	83	80	80	41	40 ± 3	40 ± 3	37	42 ± 3	42 ± 3
shuttle	NMF-L2	99	95 ± 13	100	92	87 ± 27	96 ± 1	99	85 ± 25	93 ± 2
smtp	NMF-L2	86	91	91	47	0	0	25	0	0
lympho	NMF-wL2	99	100	100	83	86 ± 10	88 ± 8	81	79 ± 14	80 ± 13
thyroid	NMF-wL2	96	97 ± 2	98	43	52 ± 11	55 ± 5	42	49 ± 10	52 ± 4
annthyroid	NMF-wL2	69	81 ± 3	82 ± 2	18	26 ± 2	26 ± 1	18	24 ± 1	24 ± 1
musk	NMF-wL2	100	95 ± 14	100	100	88 ± 25	97 ± 3	100	85 ± 27	94 ± 5
letter	NMF-Mah.	88	65 ± 9	62 ± 1	12	7 ± 14	3 ± 2	33	12 ± 8	10 ± 3
vowels	NMF-Mah.	95	77 ± 7	75 ± 3	36	19 ± 7	17 ± 6	33	18 ± 7	15 ± 5
wine	ICA-L2	100	82 ± 6	79 ± 2	89	17 ± 25	8 ± 7	82	28 ± 22	21 ± 13
cardio	ICA-L2	93	90 ± 10	93 ± 1	50	44 ± 11	48 ± 3	42	43 ± 10	46 ± 4
optdigits	ICA-corr.	81	71 ± 5	71 ± 5	0	0 ± 2	0 ± 2	4	4 ± 3	3 ± 3
mnist	ICA-corr.	93	79 ± 3	80 ± 2	52	22 ± 5	23 ± 5	48	26 ± 3	27 ± 3
glass	ICA-canb.	93	70 ± 3	69 ± 1	30	8 ± 3	7	31	10 ± 4	9 ± 4
http	AE-L2	100	94 ± 19	100	89	87 ± 29	97 ± 2	98	88 ± 30	98
wbc	AE-L2	94 ± 0	94	94	64 ± 9	51 ± 3	51 ± 2	53 ± 10	41 ± 5	41 ± 5
mammog.	AE-L2	88 ± 2	85 ± 4	86	30 ± 7	20 ± 3	20 ± 3	26 ± 8	15 ± 3	15 ± 2
breastw	AE-L2	99	94 ± 16	99	93	78 ± 32	89	92	84 ± 16	89
pima	AE-wL2	75 ± 1	67 ± 3	68	35 ± 6	25 ± 4	26 ± 2	53 ± 4	49 ± 2	50 ± 2
vertebral	VAE-cheb.	79 ± 5	37 ± 6	35 ± 1	20 ± 4	-10 ± 2	-11 ± 2	46 ± 6	5 ± 3	4 ± 3
ionosphere	VAE-Mah.	93 ± 0	86 ± 2	85	76 ± 2	50 ± 8	47 ± 2	70 ± 3	53 ± 8	50 ± 2
pendigits	none-L4	99	91 ± 12	95	46	31 ± 9	34 ± 3	53	28 ± 8	31 ± 3
satimage-2	none-L4	100	96 ± 11	99	91	80 ± 23	87 ± 2	85	70 ± 22	77 ± 4
cover	none-wL4	96	86 ± 10	89 ± 3	6	8 ± 3	9 ± 2	9	7 ± 2	8 ± 2
satellite	none-canb.	77	69 ± 4	70 ± 1	38	33 ± 8	36 ± 2	39	39 ± 2	40 ± 1
speech	none-canb.	50	49 ± 5	48 ± 1	2	1 ± 3	0 ± 1	0	2 ± 2	1 ± 1

AE, VAE, PCA, ICA, NMF, ...?

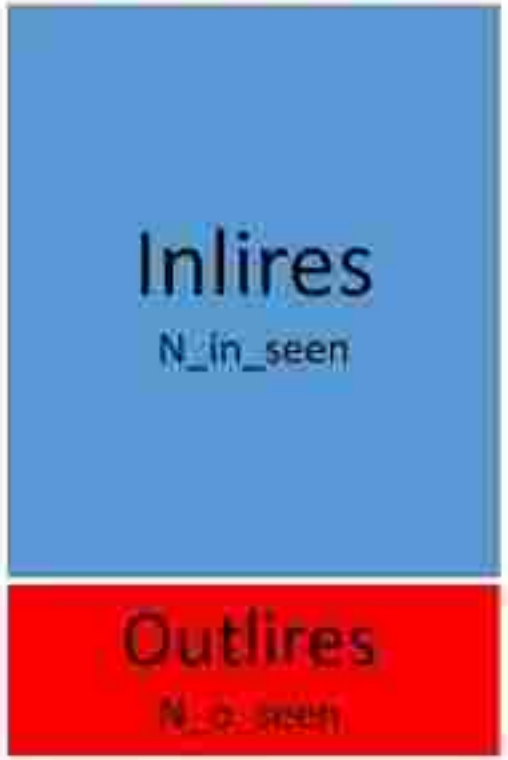
L1, L2, L4, Chebyshev, Canbera, ...?

Semi-supervised or active learning

DRAMA vs. LOF vs. iforest

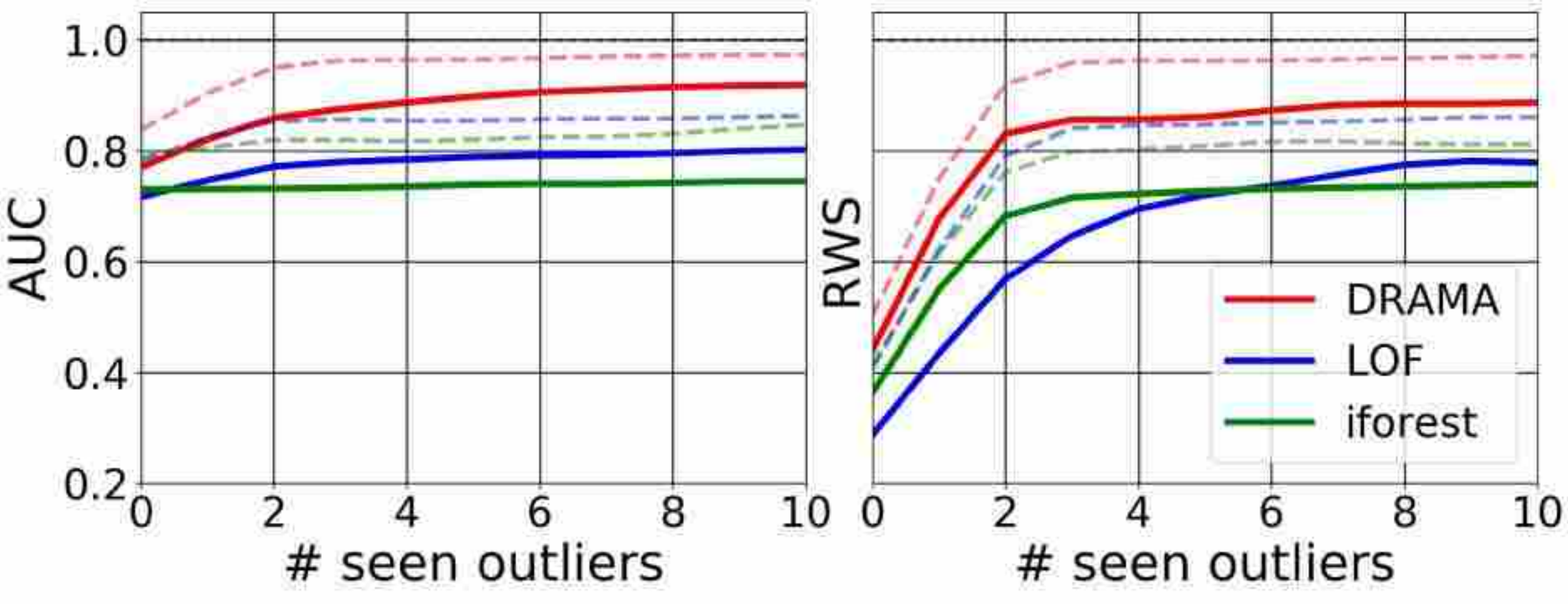
Semi-supervised approach:

Train:
picking the best DRT+metric



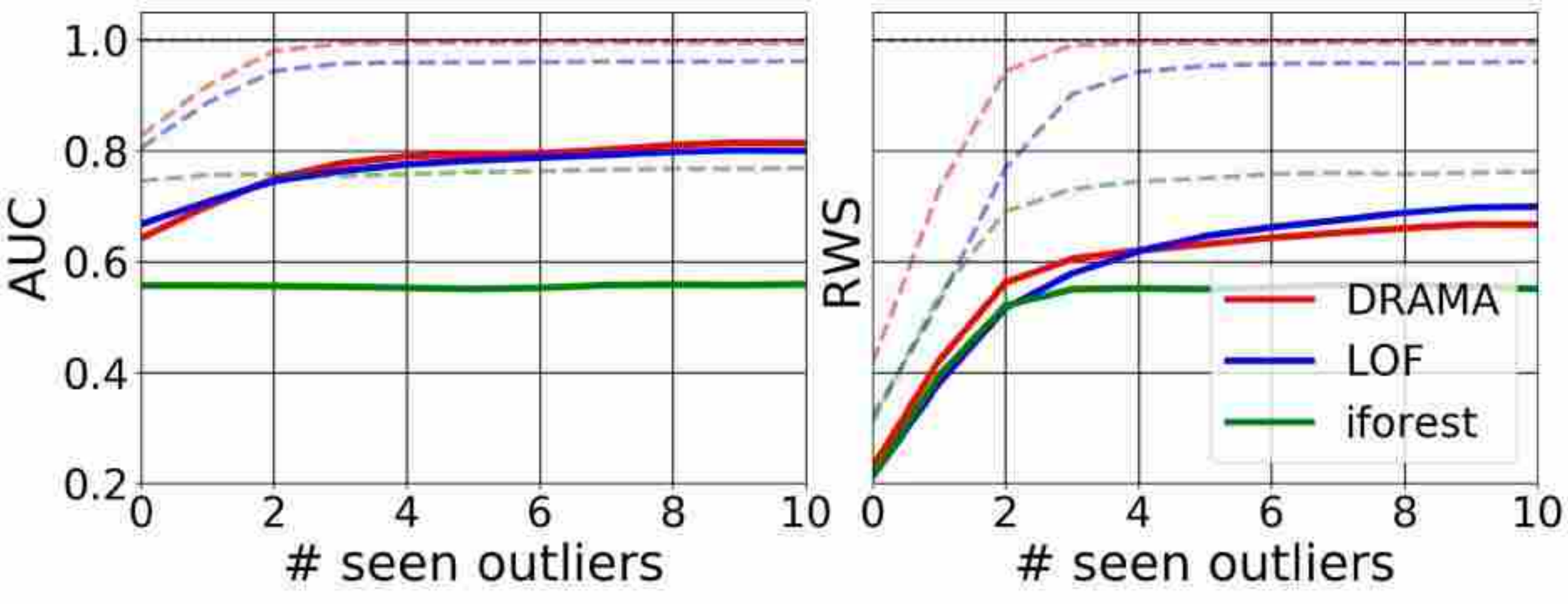
Prediction

Local anomaly in low dimension



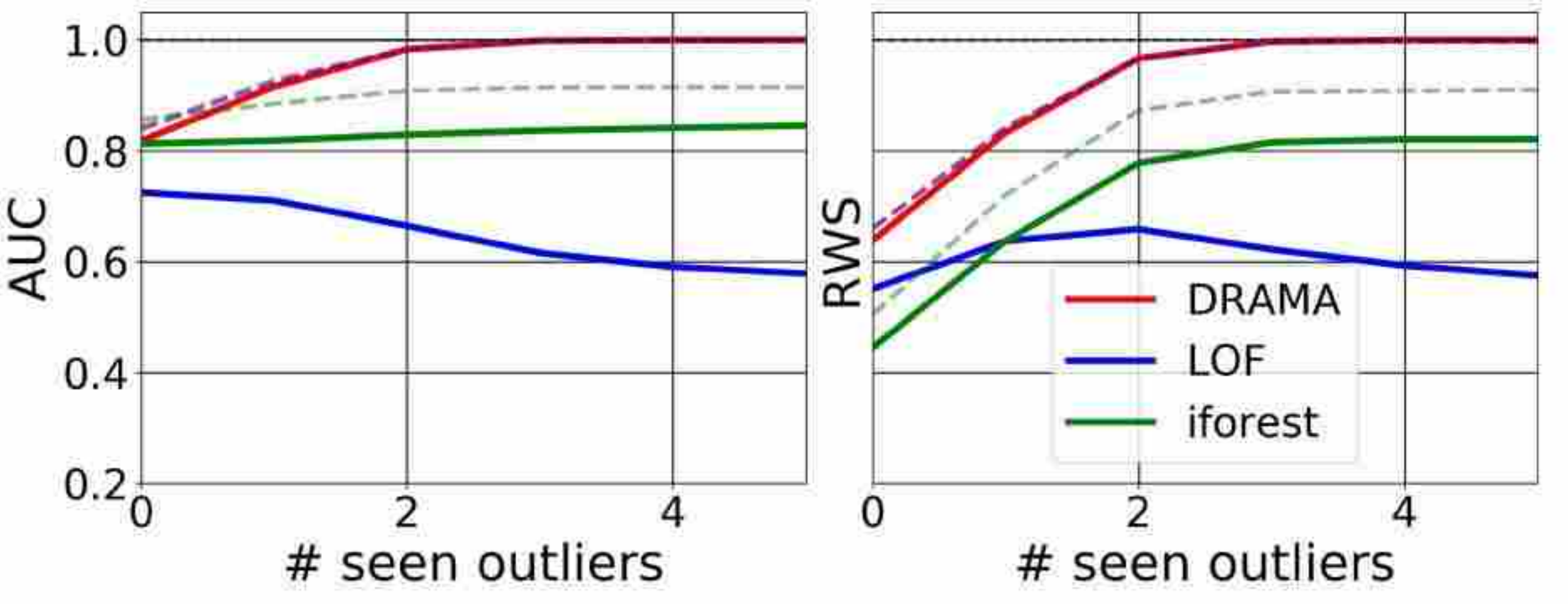
n_f=100

New class anomaly in low dimension



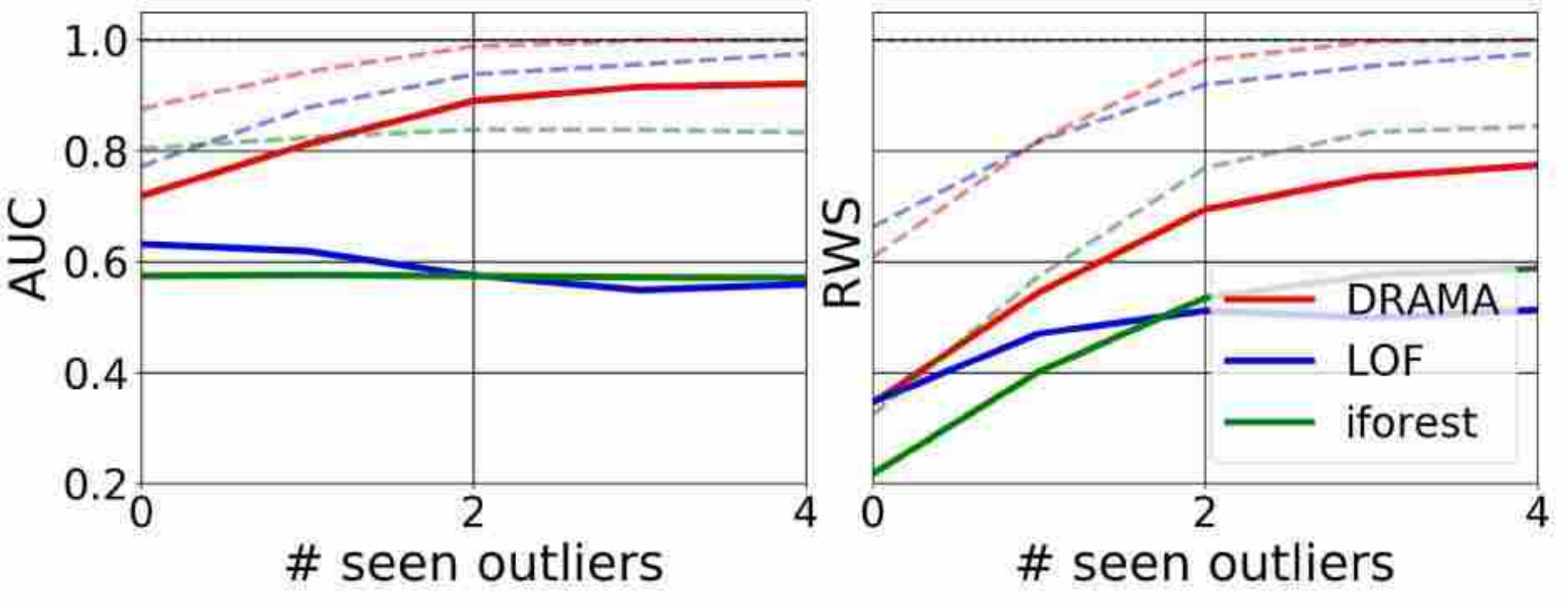
n_f=100

Local anomaly in high dimension



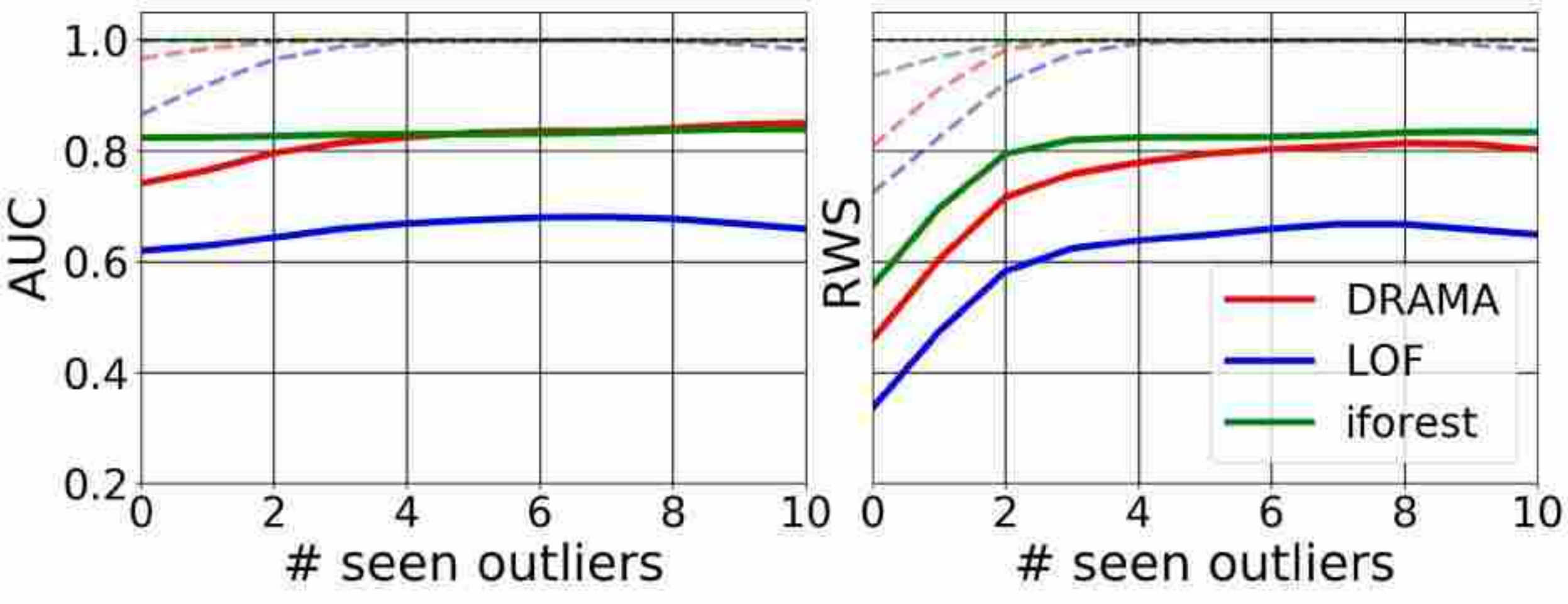
n_f=3000

New class anomaly in high dimension



n_f=3000

Averaged on real data

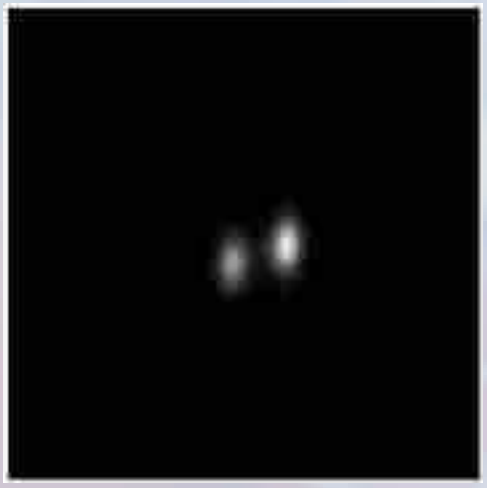


Widefield ouTlier Finder (WTF)

Includes 337 boring objects and 17 interesting

AUC: 87
MCC: 26
RWS: 31

boring



interesting

Future directions

- 2D convolutions
- PLAsTiCC (Kaggle competitions)
- Deep learning based clustering
- Graphical user interface
- Active learning mode
- Detect and classify
- …


Thanks for your attention.
You can find DRAMA here
https://github.com/vafaei-ar/drama