Big Data and Deep Learnings in Astronomy

Sungryong Hong

2/20/2019

The 8th SSG Workshop

Outlines

- Big Data in Astronomy
  - Apache Spark
  - Examples and My Work
- Deep Learnings in Astronomy
  - Keras/Tensorflow
  - Examples and My Work

Why Big Data in Astronomy ?

Literally, astronomers need to handle "Astronomical" scales of data.

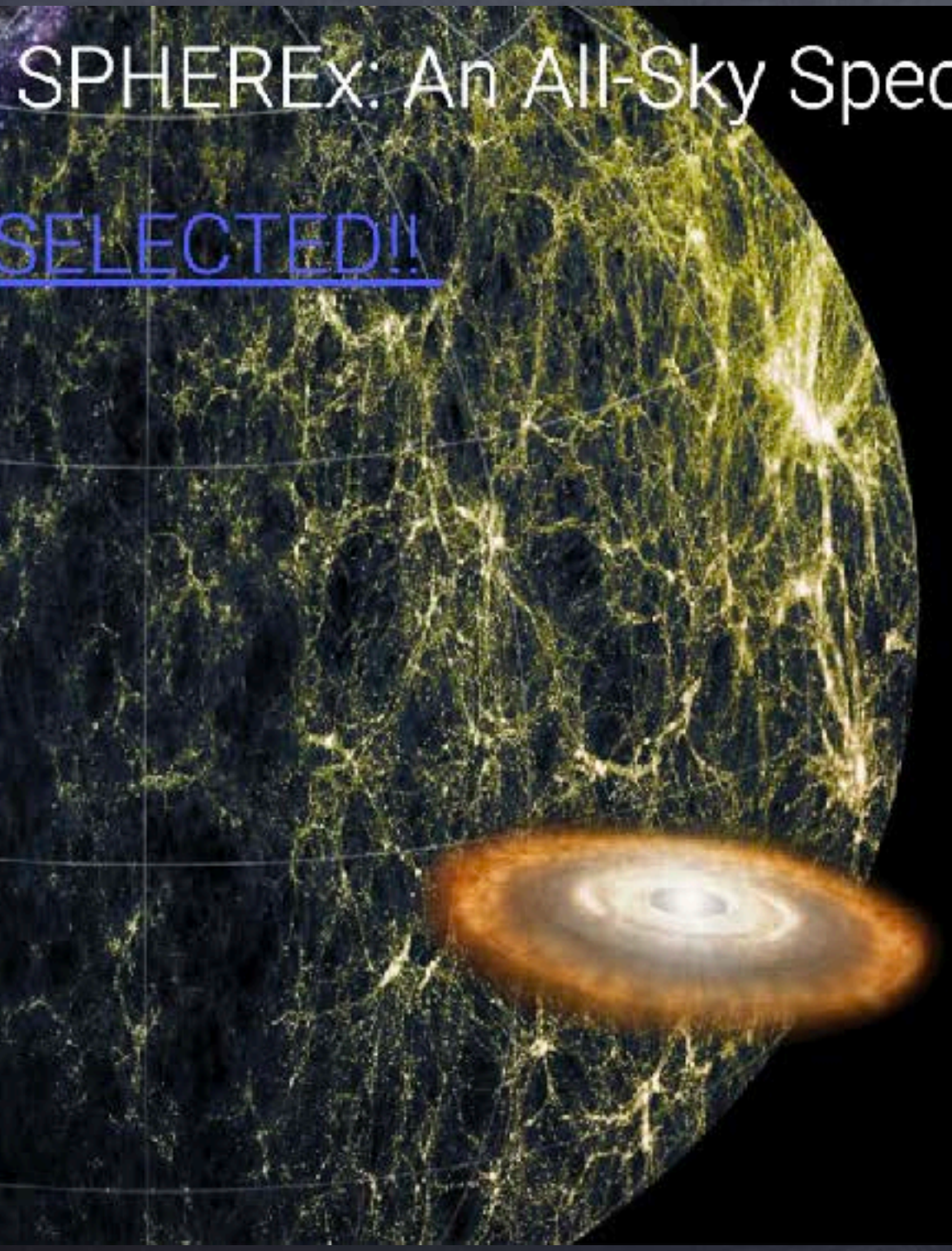
SDSS, DESI, ...

Gaia DR2

Below an overview of the planned Gaia Data Release 2 in numbers:

	# sources in Gaia DR2	# sources in Gaia DR1
Total number of sources	**1,692,919,135**	**1,142,679,769**
Number of 5-parameter sources	1,331,909,727	2,057,050
Number of 2-parameter sources	361,009,408	1,140,622,719
Sources with mean G magnitude	1,692,919,135	1,142,679,769
Sources with mean G_{BP}-band photometry	1,381,964,755	-
Sources with mean G_{RP}-band photometry	1,383,551,713	-
Sources with radial velocities	7,224,631	-
Variable sources	550,737	3,194
Known asteroids with epoch data	14,099	-
Gaia-CRF sources	556,869	2,191
Effective temperatures (T_{eff})	161,497,595	-
Extinction (A_G) and reddening ($E(G_{BP}-G_{RP})$)	87,733,672	-
Sources with radius and luminosity	76,956,778	-

SPHEREx: An All-Sky Spectral Survey



SELECTED!!

Designed to Explore

The Origin of the Universe

The Origin and History of Galaxies

The Origin of Water in Planetary Systems

The First All-Sky Spectral Survey

A Rich Legacy Archive
for the Astronomy Community
with 100's of Millions
of Stars and Galaxies

Low-Risk Implementation

No Moving Parts

Single Observing Mode

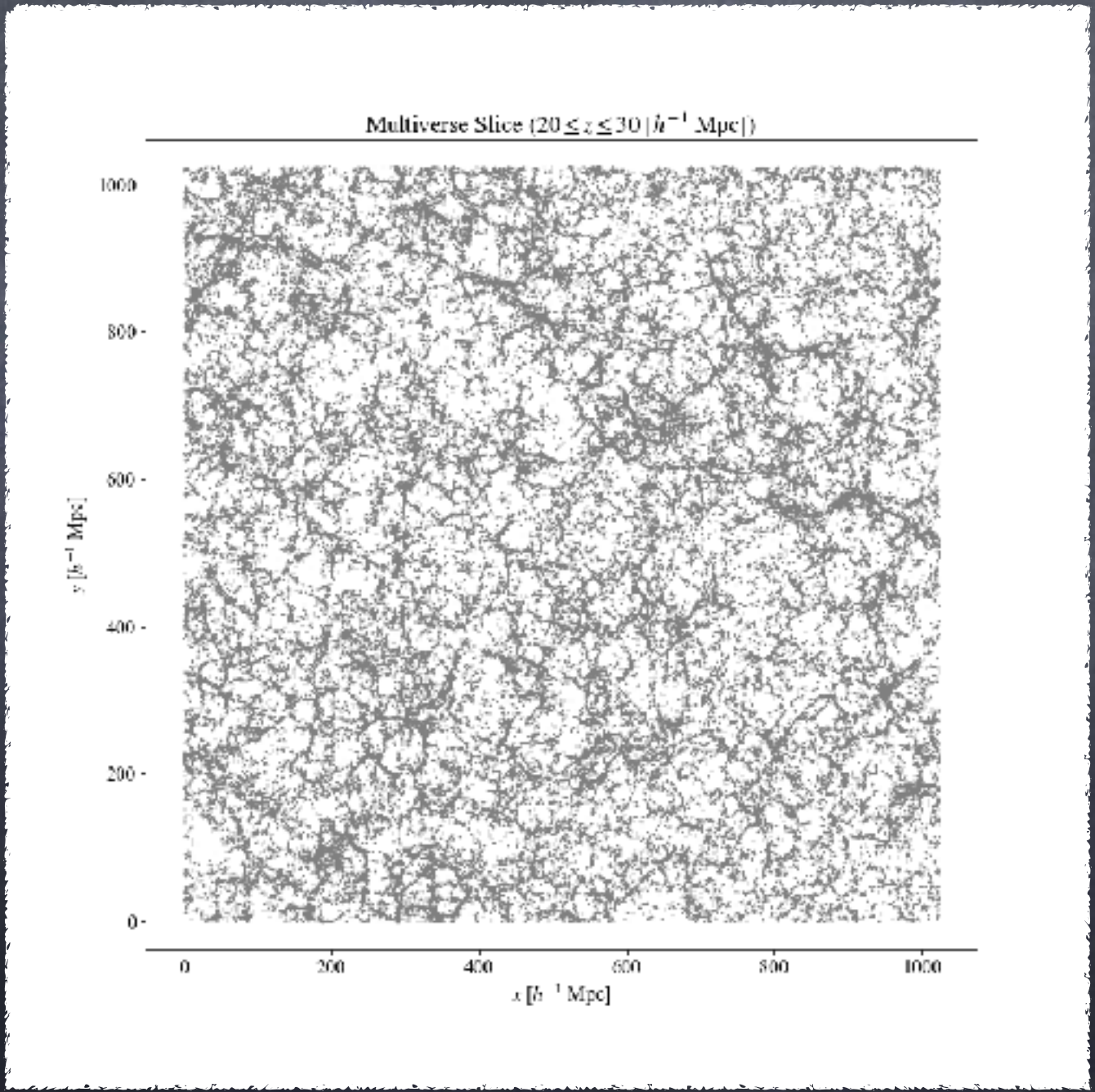
Large Technical & Scientific Margins

Follows successful CIT/JPL mgt. model of NuSTAR

JPL R·I·T ipac

CENTER FOR ASTROPHYSICS

My Work: Multiverse Simulations



My Work: Multiverse Simulations

Table 2. Sample Selections

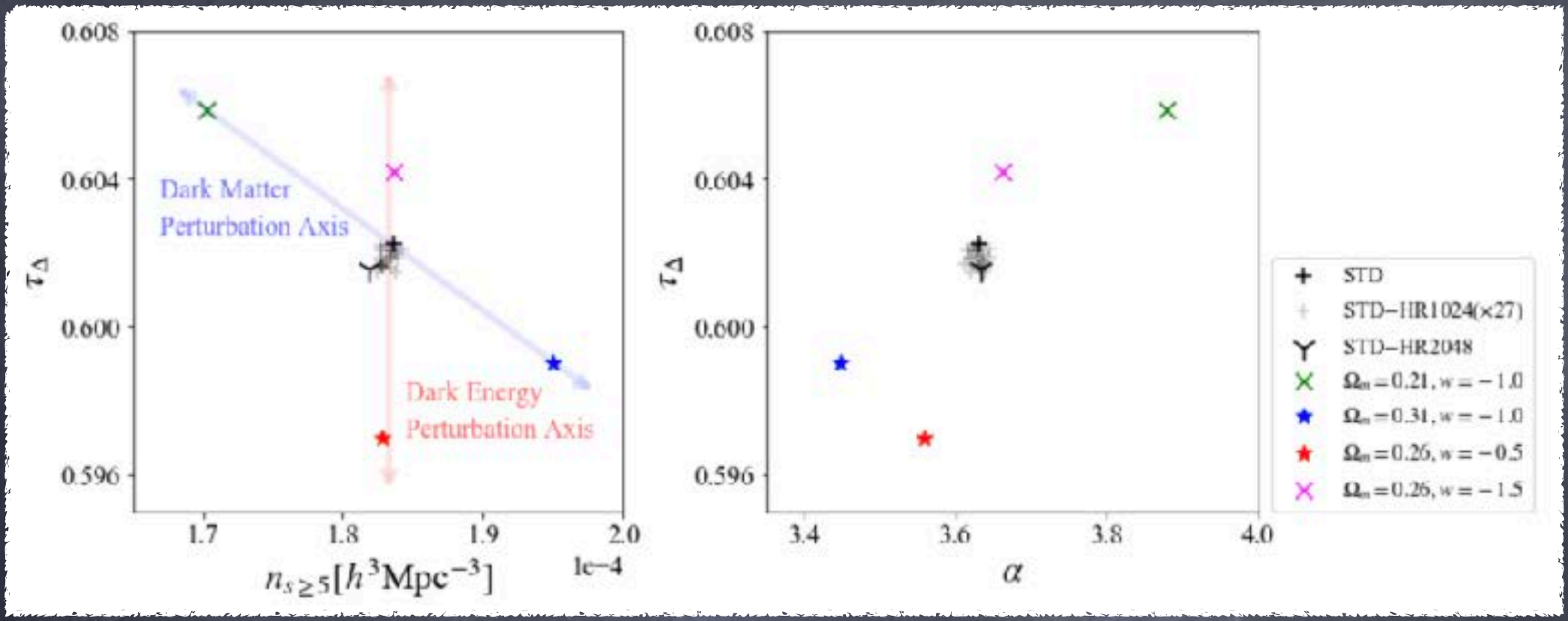
Multiverses		Equal Mass Cut Sample		Equal Abundance Sample[a]	
Name	Cosmological Parameters	N_h	$M_{cut}(M_\odot)$	N_h	$M_{min}(M_\odot)$
STD	$\Omega_m = 0.26, w = -1.0$	7,086,717	5.00×10^{11}	7,086,717	5.05×10^{11}
DE1	$\Omega_m = 0.26, w = -0.5$	7,806,135	5.00×10^{11}	7,086,717	5.59×10^{11}
DE2	$\Omega_m = 0.26, w = -1.5$	6,886,870	5.00×10^{11}	7,086,717	4.87×10^{11}
DM1	$\Omega_m = 0.31, w = -1.0$	8,595,923	5.00×10^{11}	7,086,717	6.24×10^{11}
DM2	$\Omega_m = 0.21, w = -1.0$	5,579,491	5.00×10^{11}	7,086,717	3.86×10^{11}
STD-HR	Horizon Run[†]	206,140,716	5.00×10^{11}	206,140,716	5.05×10^{11}

Table 2. Sample Selections

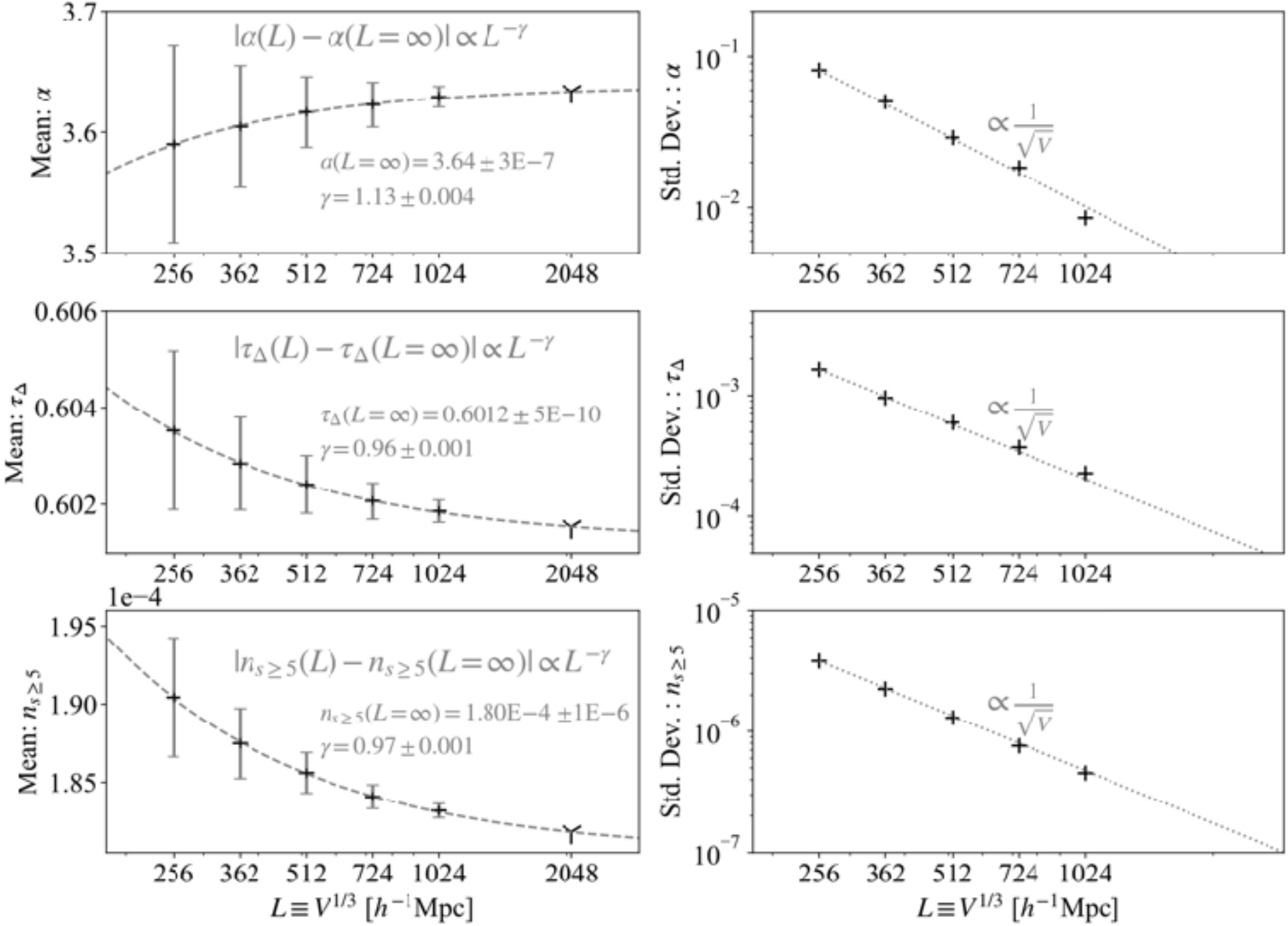
Multiverses		Equal Mass Cut Sample		Equal Abundance Sample[a]	
Name	Cosmological Parameters	N_h	$M_{cut}(M_\odot)$	N_h	$M_{min}(M_\odot)$
STD	$\Omega_m = 0.26, w = -1.0$	7,086,717	5.00×10^{11}	7,086,717	5.05×10^{11}
DE1	$\Omega_m = 0.26, w = -0.5$	7,806,135	5.00×10^{11}	7,086,717	5.59×10^{11}
DE2	$\Omega_m = 0.26, w = -1.5$	6,886,870	5.00×10^{11}	7,086,717	4.87×10^{11}
DM1	$\Omega_m = 0.31, w = -1.0$	8,595,923	5.00×10^{11}	7,086,717	6.24×10^{11}
DM2	$\Omega_m = 0.21, w = -1.0$	5,579,491	5.00×10^{11}	7,086,717	3.86×10^{11}
STD-HR	Horizon Run[†]	206,140,716	5.00×10^{11}	206,140,716	5.05×10^{11}

Table 1. Hardware Configurations for the Spark Clusters[†]

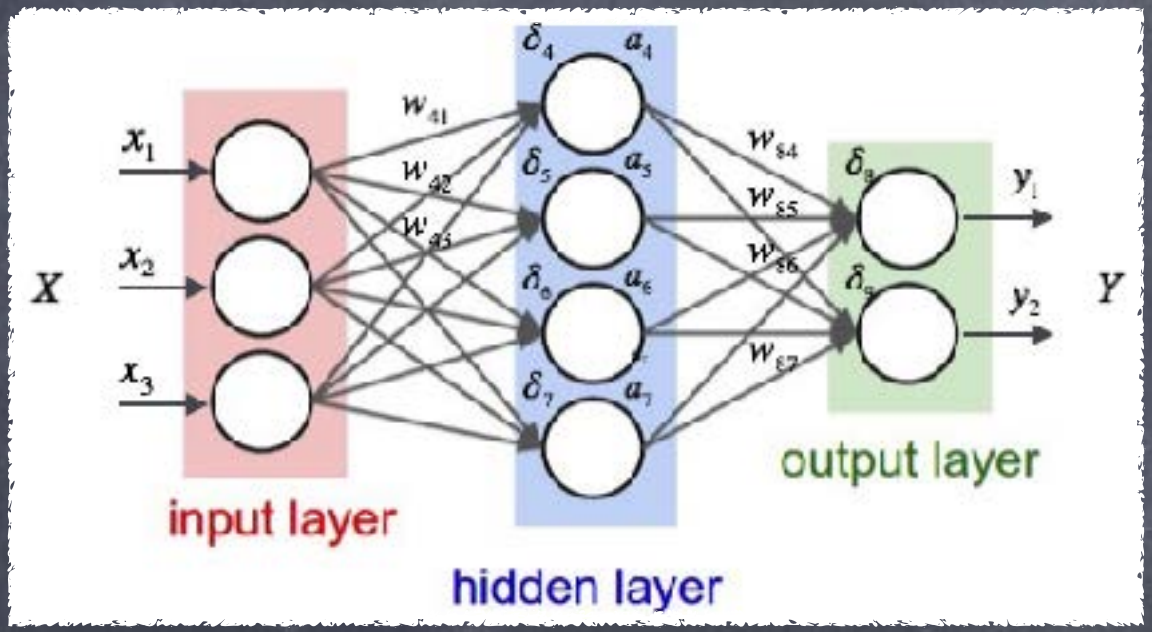
	Driver Node		Worker Node		
Cluster Name	vCPUs[†]	Memory	vCPUs[†]	Memory	nWorkers[†]
KIAS Standalone[a]	4	32GB	16	52GB	3
Google Cloud Dataproc[b]	16	104GB	32	208GB	5



STD-HR2048: 57 millions halos with 206 millions connections
I paid $30 for this single point.


$|\alpha(L) - \alpha(L=\infty)| \propto L^{-\gamma}$
$\alpha(L=\infty) = 3.64 \pm 3\text{E}{-7}$
$\gamma = 1.13 \pm 0.004$
Mean: α
Std. Dev.: α
$\propto \frac{1}{\sqrt{V}}$
$|\tau_\Delta(L) - \tau_\Delta(L=\infty)| \propto L^{-\gamma}$
$\tau_\Delta(L=\infty) = 0.6012 \pm 5\text{E}{-10}$
$\gamma = 0.96 \pm 0.001$
Mean: τ_Δ
Std. Dev.: τ_Δ
1e−4
$|n_{s\geq 5}(L) - n_{s\geq 5}(L=\infty)| \propto L^{-\gamma}$
$n_{s\geq 5}(L=\infty) = 1.80\text{E}{-4} \pm 1\text{E}{-6}$
$\gamma = 0.97 \pm 0.001$
Mean: $n_{s\geq 5}$
Std. Dev.: $n_{s\geq 5}$
256 362 512 724 1024 2048
$L \equiv V^{1/3}$ [h^{-1}Mpc]

Deep Learnings in Astronomy



Multilayer Perceptrons (MLP)

Example (1) : BPT Classifications

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **478,** 3177–3188 (2018)
Advance Access publication 2018 May 18

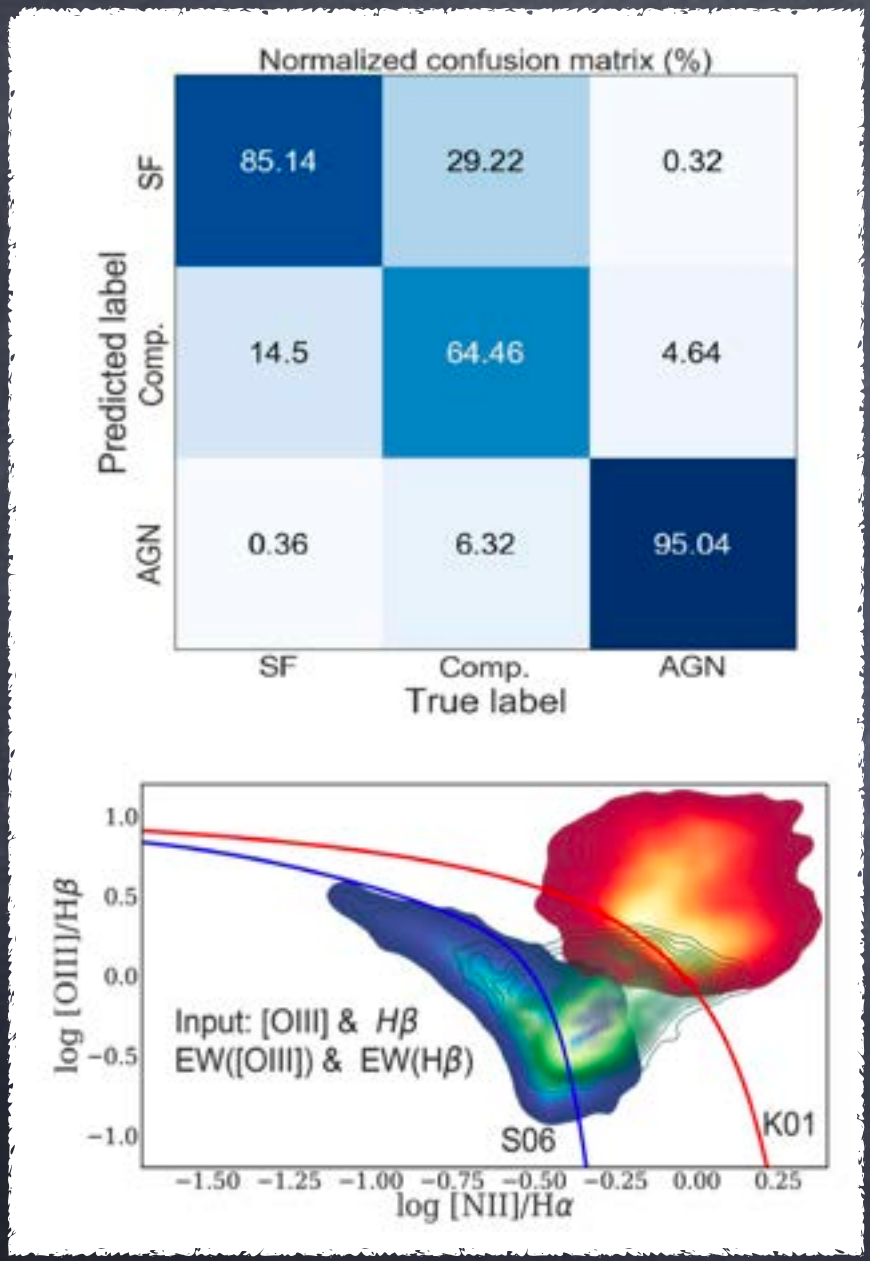
doi:10.1093/mnras/sty1331

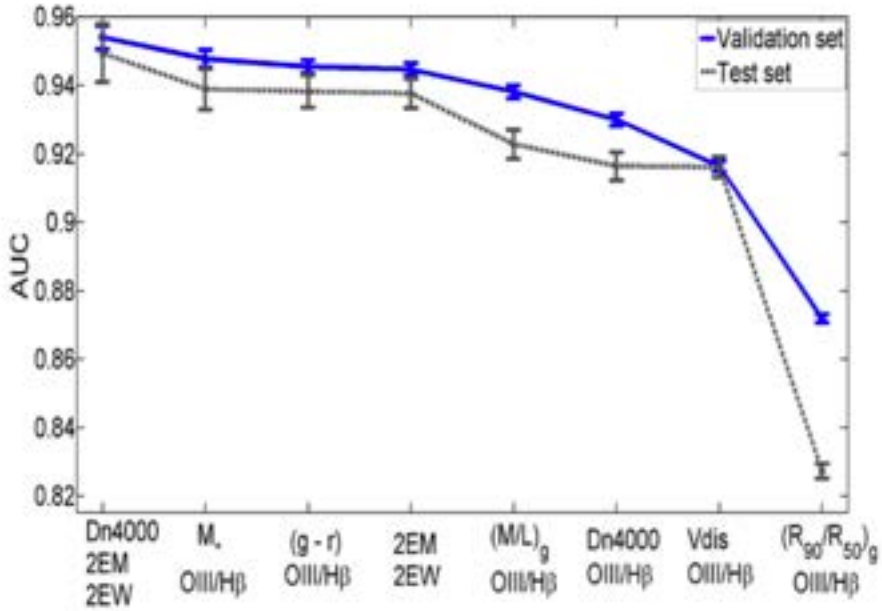
The discrimination between star-forming and AGN galaxies in the absence of H α and [NII]: a machine -learning approach

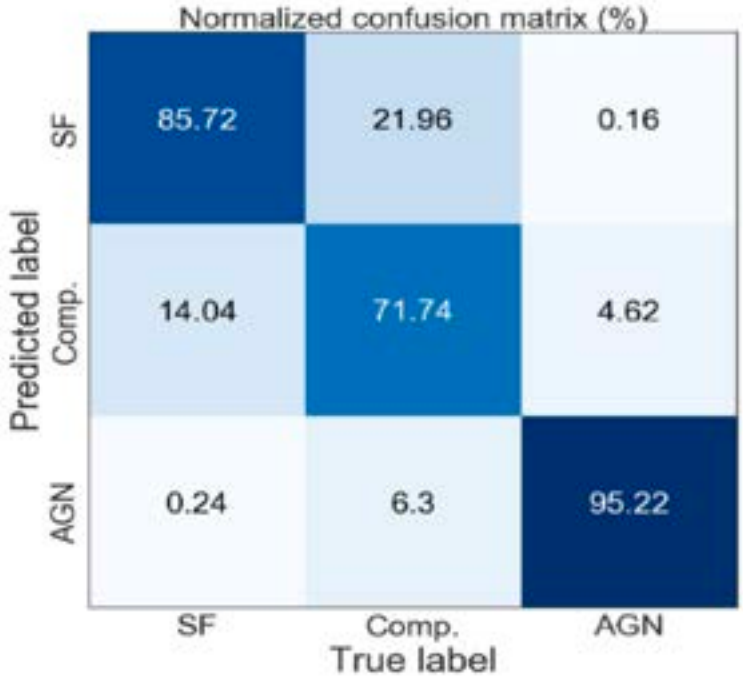
H. Teimoorinia[1]★ and J. Keown[2]

[1]*NRC Herzberg Astronomy and Astrophysics, 5071 West Saanich Road, Victoria, BC, V9E 2E7, Canada*
[2]*Department of Physics and Astronomy, University of Victoria, Victoria, BC, V8P 5C2, Canada*

Using Artificial Neural Network (ANN)
to Mimic BPT classifications
without H α and [N II] emissions


Normalized confusion matrix (%)
Predicted label
SF
Comp.
AGN
85.14
29.22
0.32
14.5
64.46
4.64
0.36
6.32
95.04
SF
Comp.
AGN
True label
log [OIII]/Hβ
log [NII]/Hα
Input: [OIII] & Hβ
EW([OIII]) & EW(Hβ)
S06
K01
1.0
0.5
0.0
−0.5
−1.0
−1.50 −1.25 −1.00 −0.75 −0.50 −0.25 0.00 0.25


AUC
Validation set
Test set
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
Dn4000
2EM
2EW
M*
OIII/Hβ
(g - r)
OIII/Hβ
2EM
2EW
(M/L)g
OIII/Hβ
Dn4000
OIII/Hβ
Vdis
OIII/Hβ
(R90/R50)g
OIII/Hβ


Normalized confusion matrix (%)
Predicted label
SF
Comp.
AGN
85.72
21.96
0.16
14.04
71.74
4.62
0.24
6.3
95.22
SF
Comp.
AGN
True label

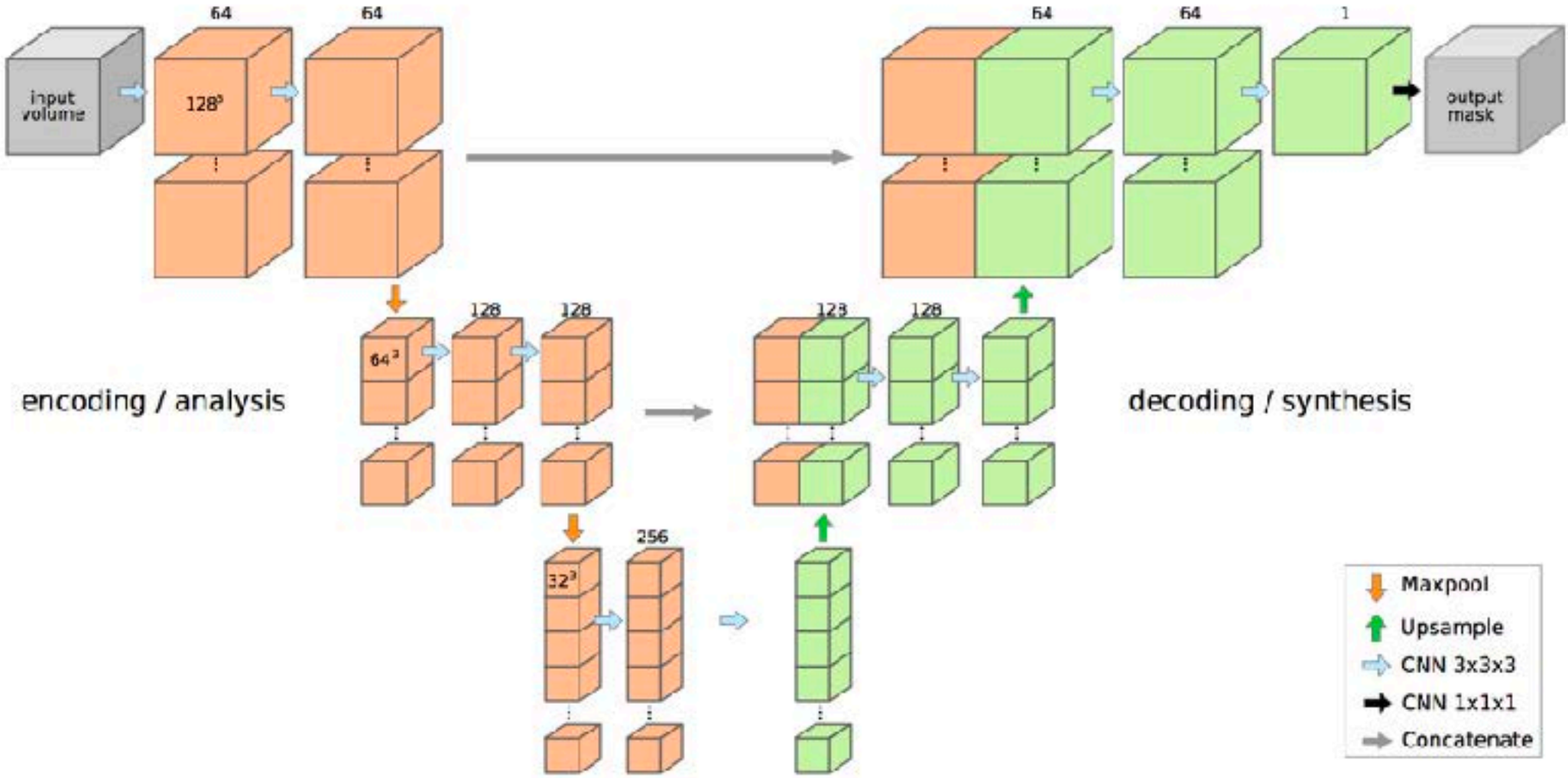
Example (2) : Cosmic Patterns

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

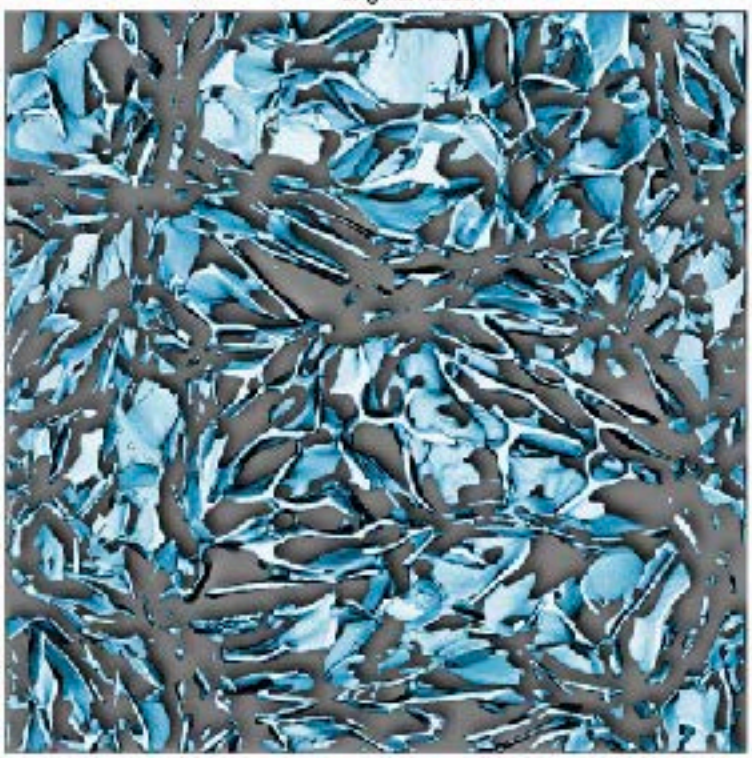
M.A. Aragon-Calvo[1] *
[1] *Instituto de Astronomía, UNAM, Apdo. Postal 106, Ensenada 22800, B.C., México*

Using Convolutional Neural Network (CNN)
to Mimic the Pattern Finder,
called Multi-scale Morphology Finder (MMF).

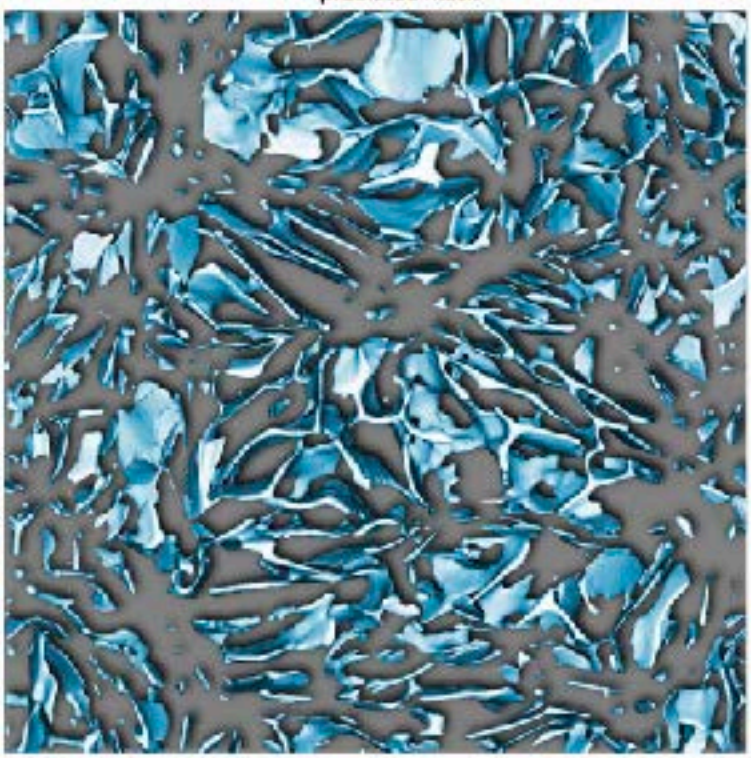
Example (2) : Cosmic Patterns



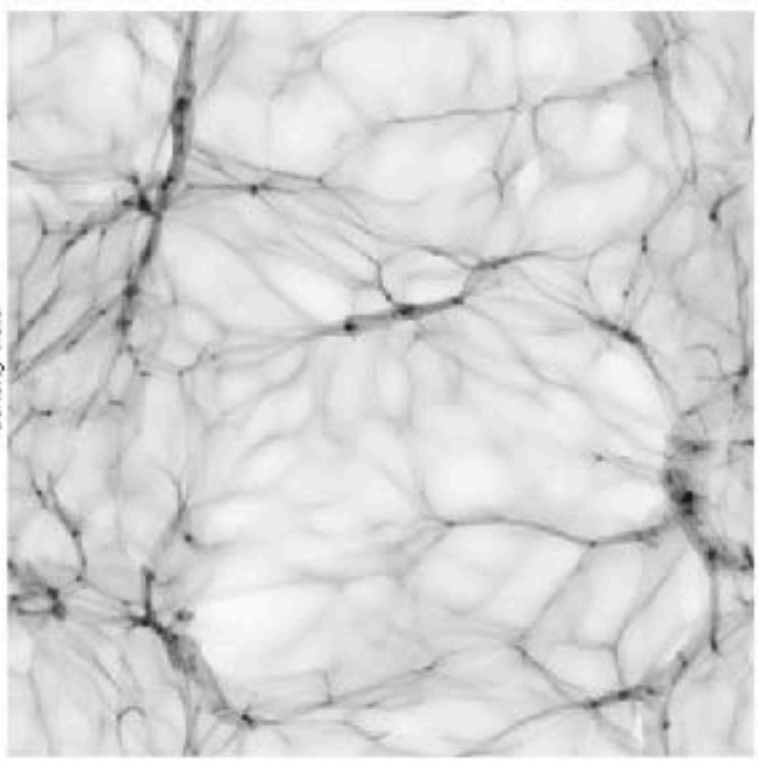
original mask



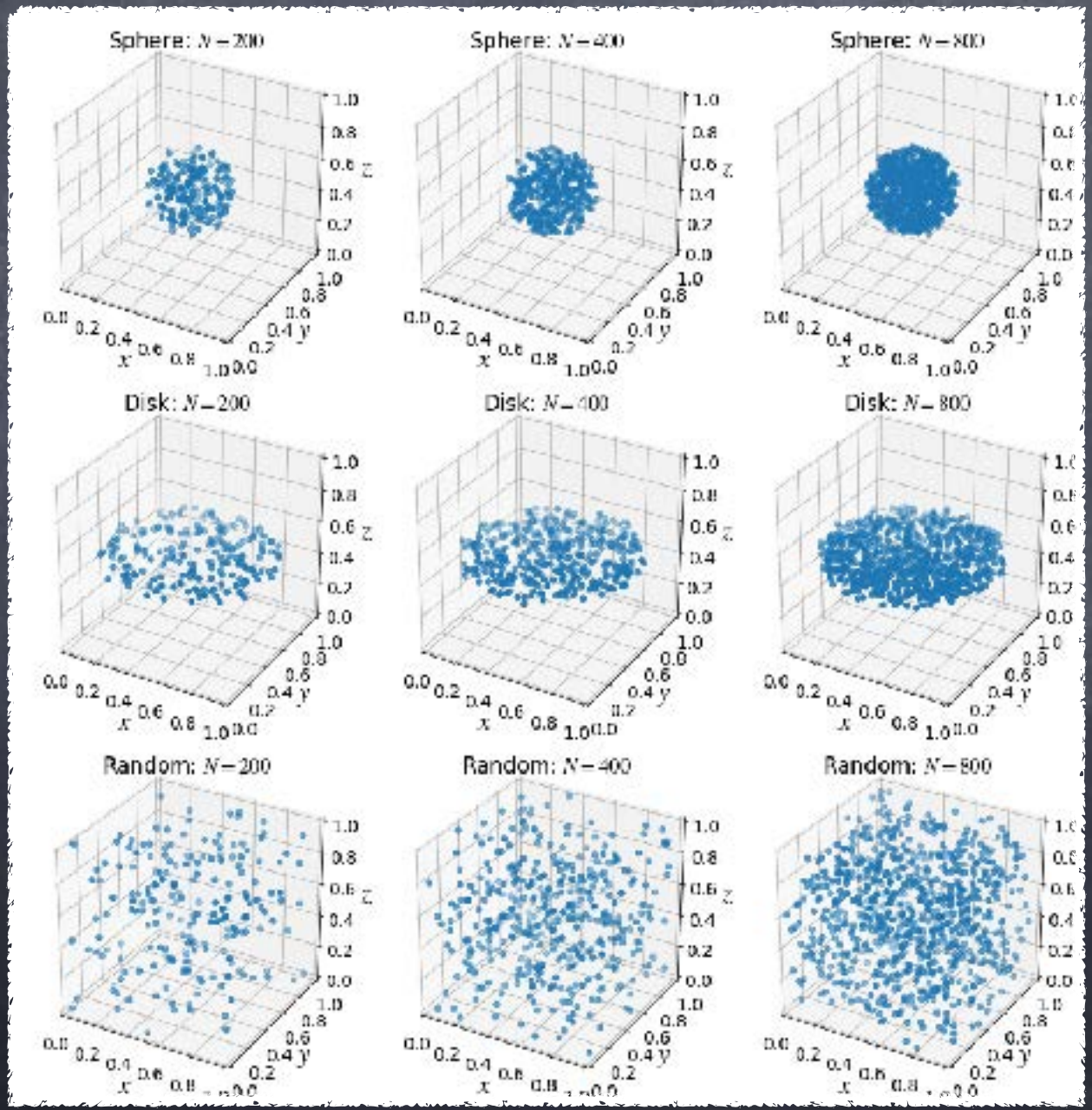
predicted mask



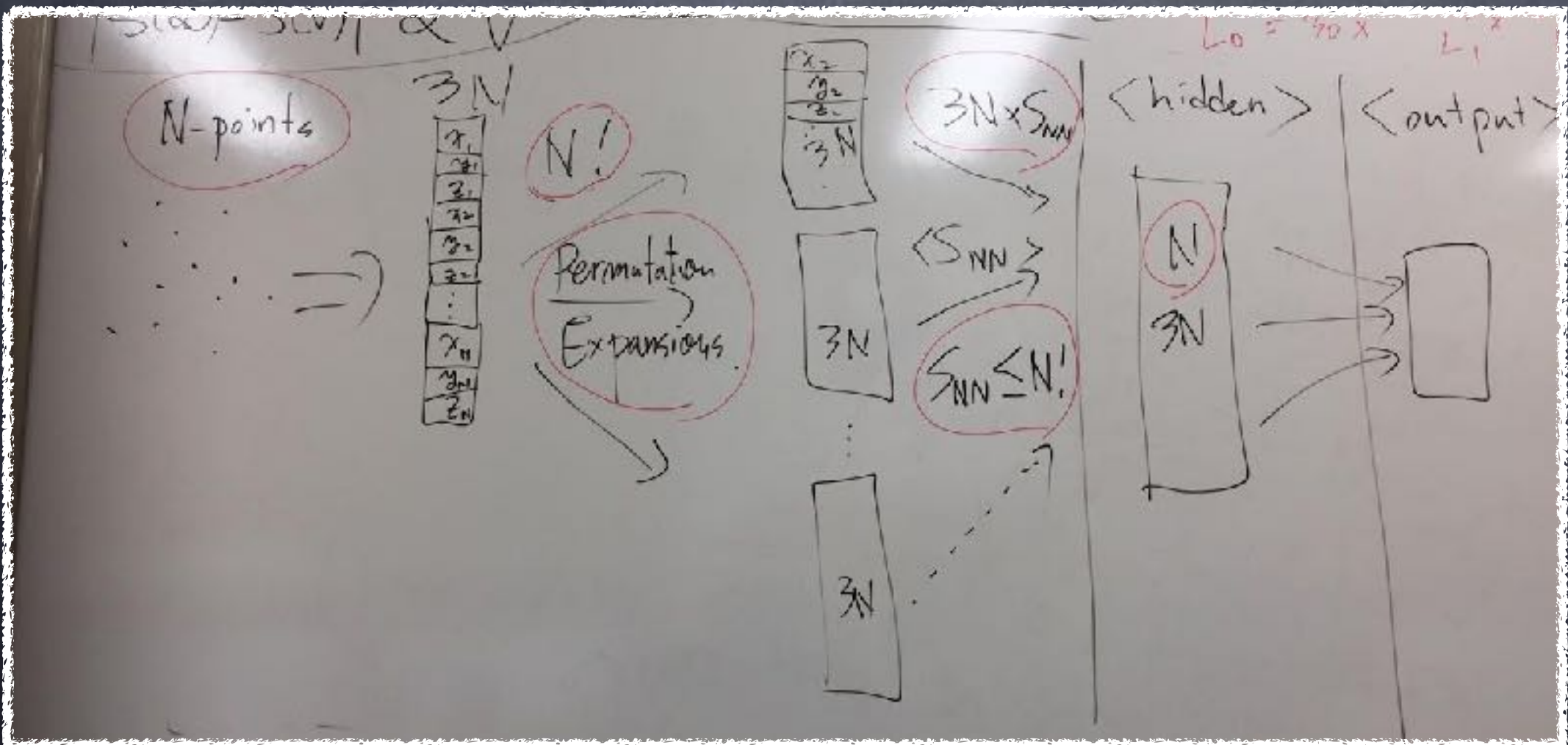
density field

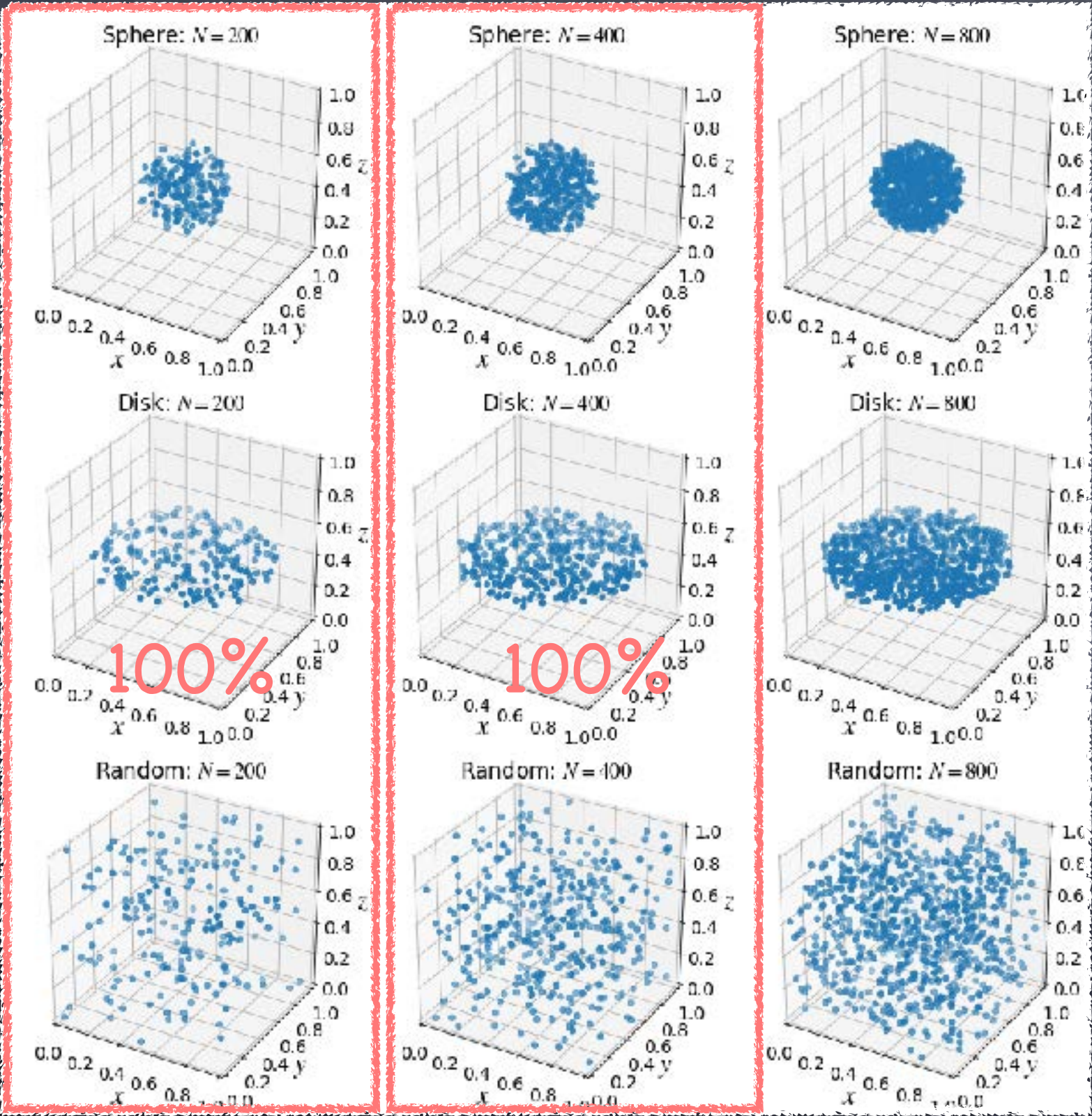


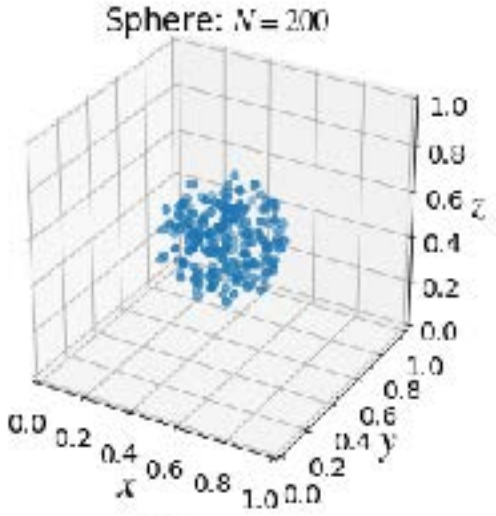
My Work : Point Patterns

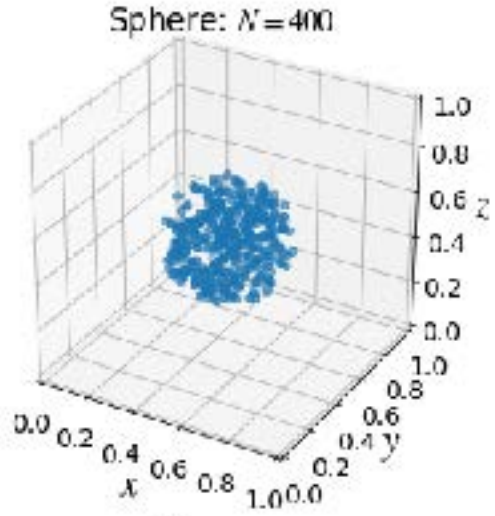


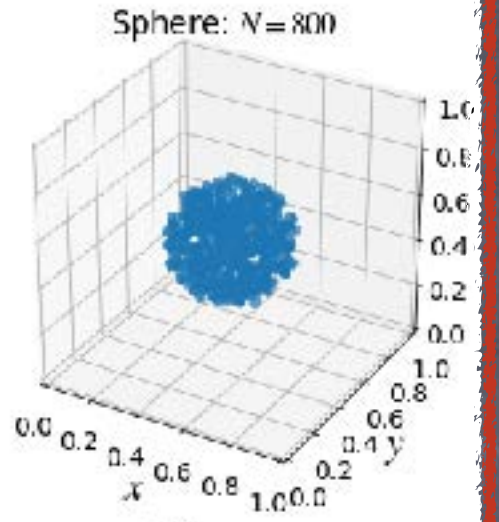
My Work : Point Patterns



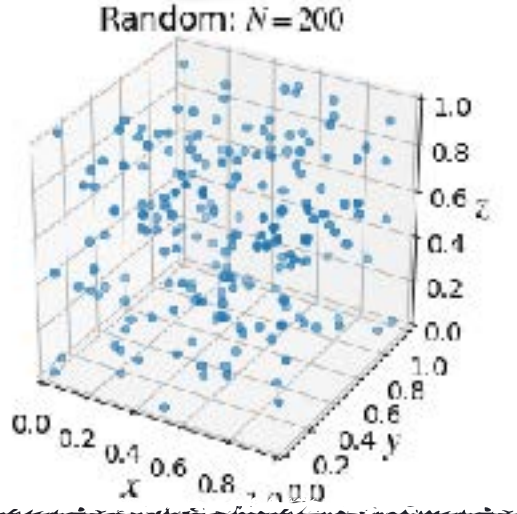

Sphere: $N = 200$
Sphere: $N = 400$
Sphere: $N = 800$
Disk: $N = 200$
Disk: $N = 400$
Disk: $N = 800$
100%
100%
Random: $N = 200$
Random: $N = 400$
Random: $N = 800$
x
y
z

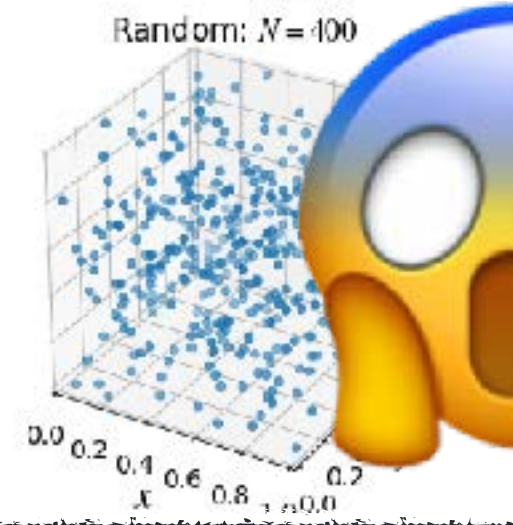


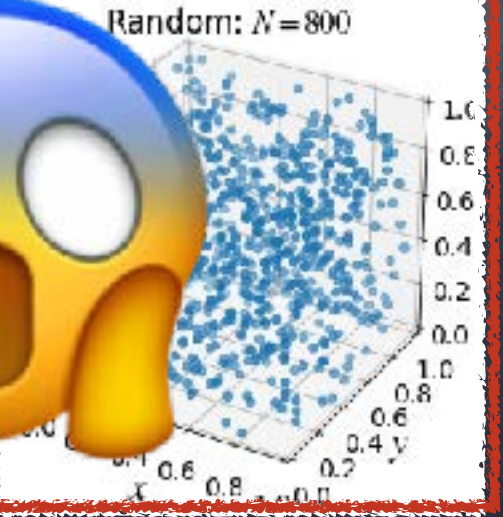




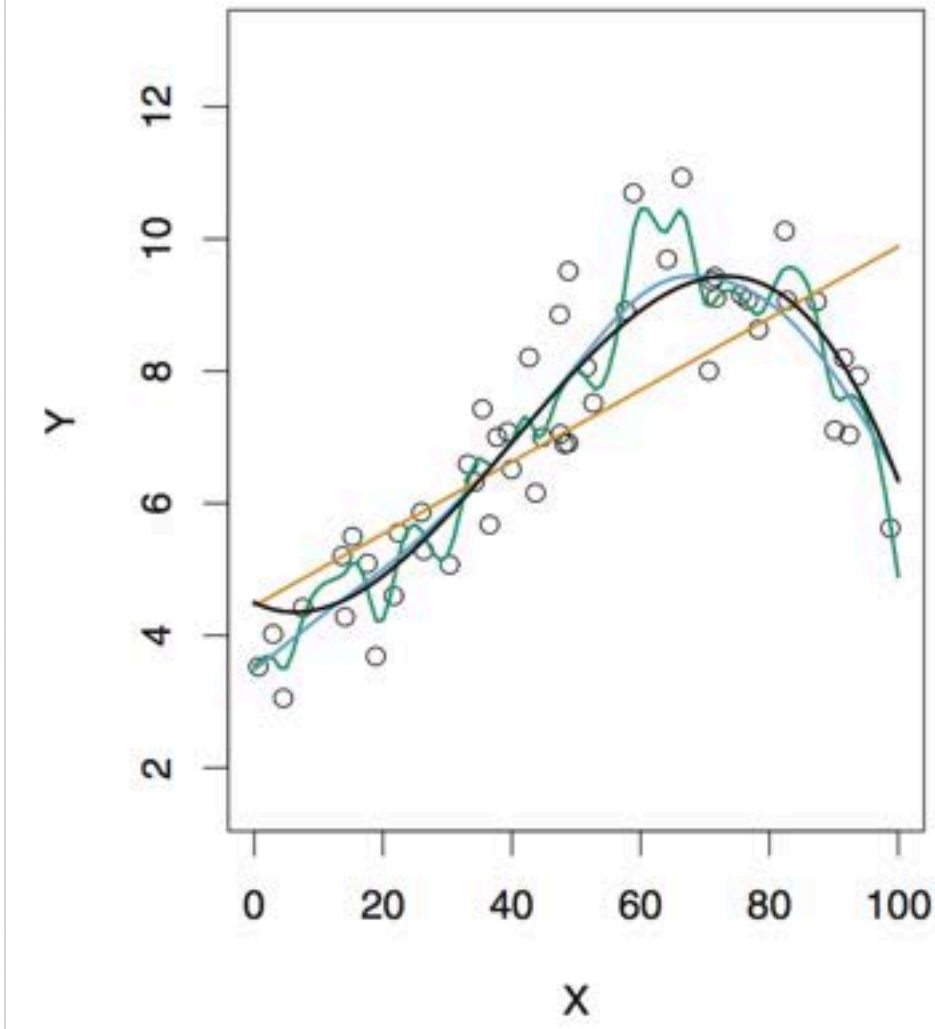

```
Epoch 17/20
 - 7s - loss: 5.3727 - acc: 0.6667 - val_loss: 5.3727 - val_acc: 0.6667
Epoch 18/20
 - 7s - loss: 5.3727 - acc: 0.6667 - val_loss: 5.3727 - val_acc: 0.6667
Epoch 19/20
 - 7s - loss: 5.3727 - acc: 0.6667 - val_loss: 5.3727 - val_acc: 0.6667
Epoch 20/20
 - 7s - loss: 5.3727 - acc: 0.6667 - val_loss: 5.3727 - val_acc: 0.6667
Baseline Error: 33.33%
```

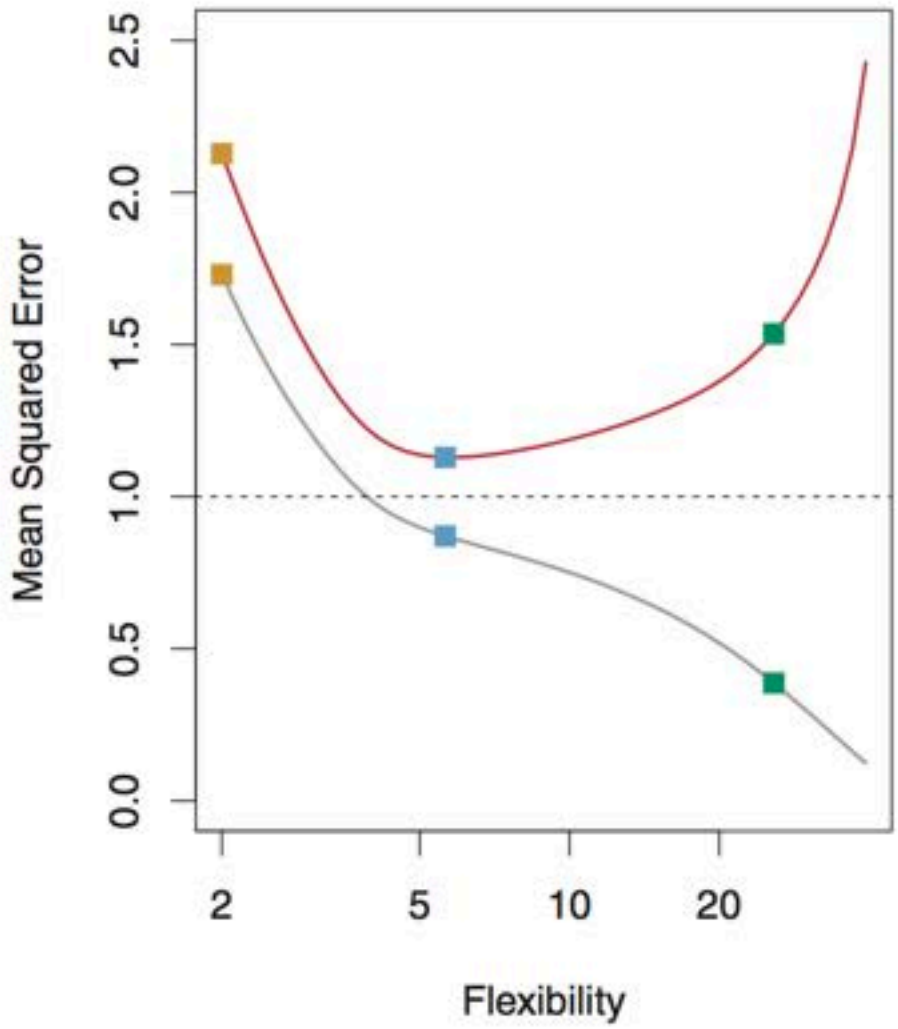


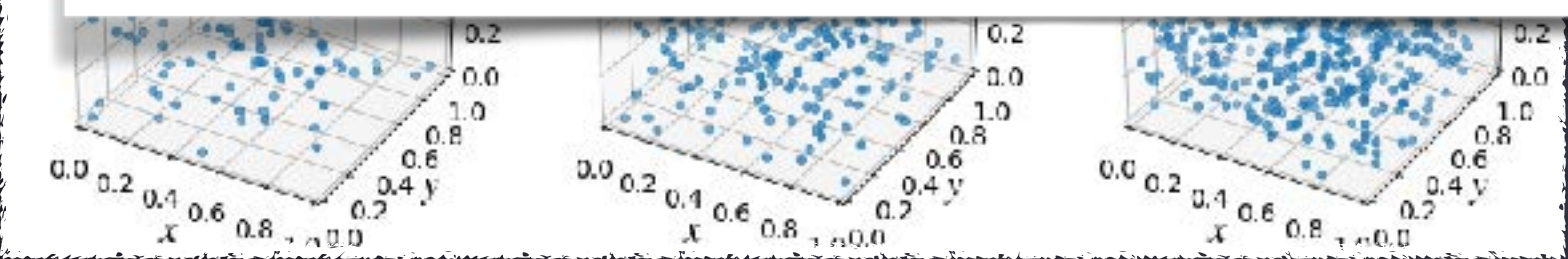


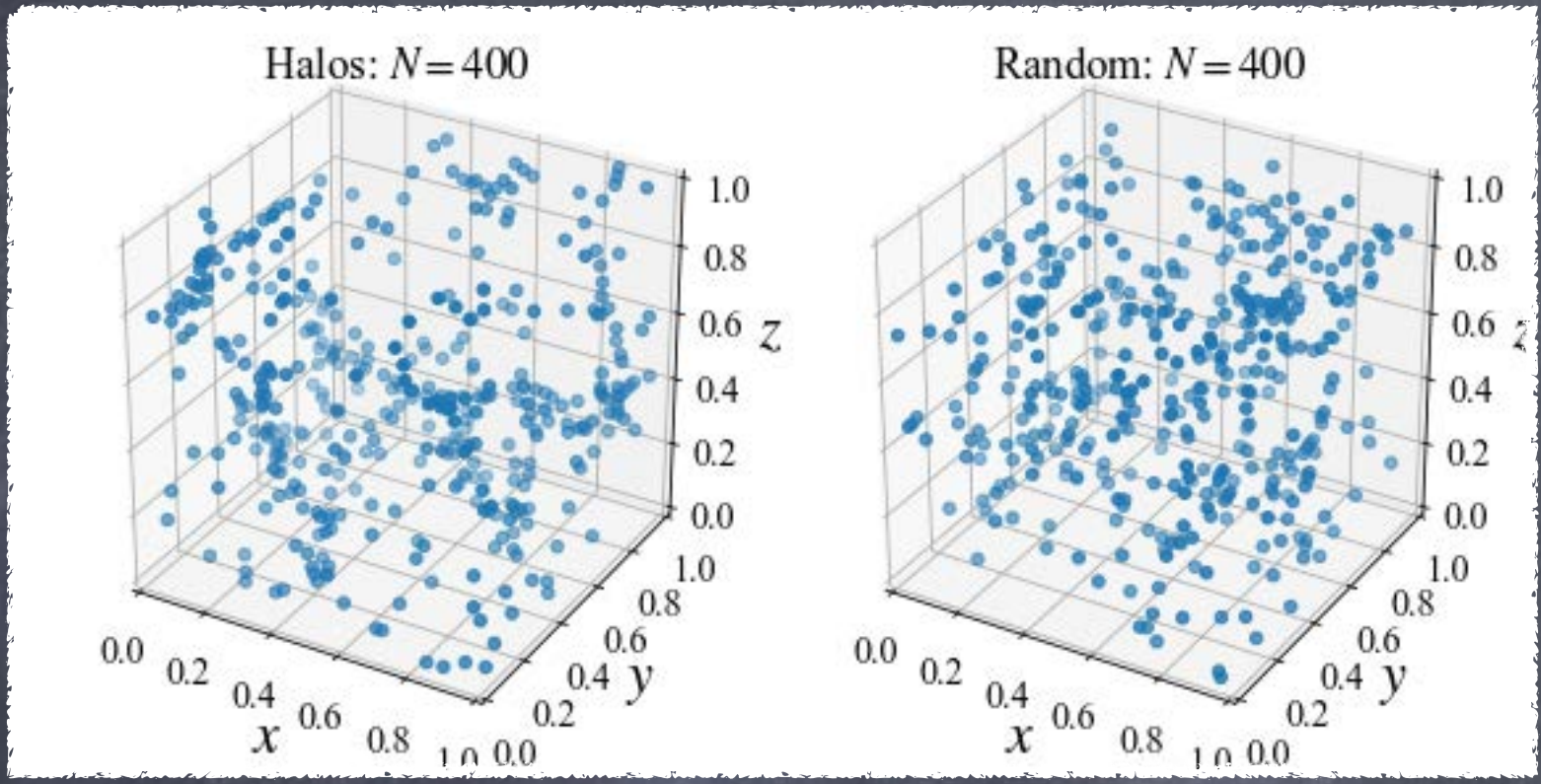



Sphere: N = 200
Sphere: N = 400
Sphere: N = 800

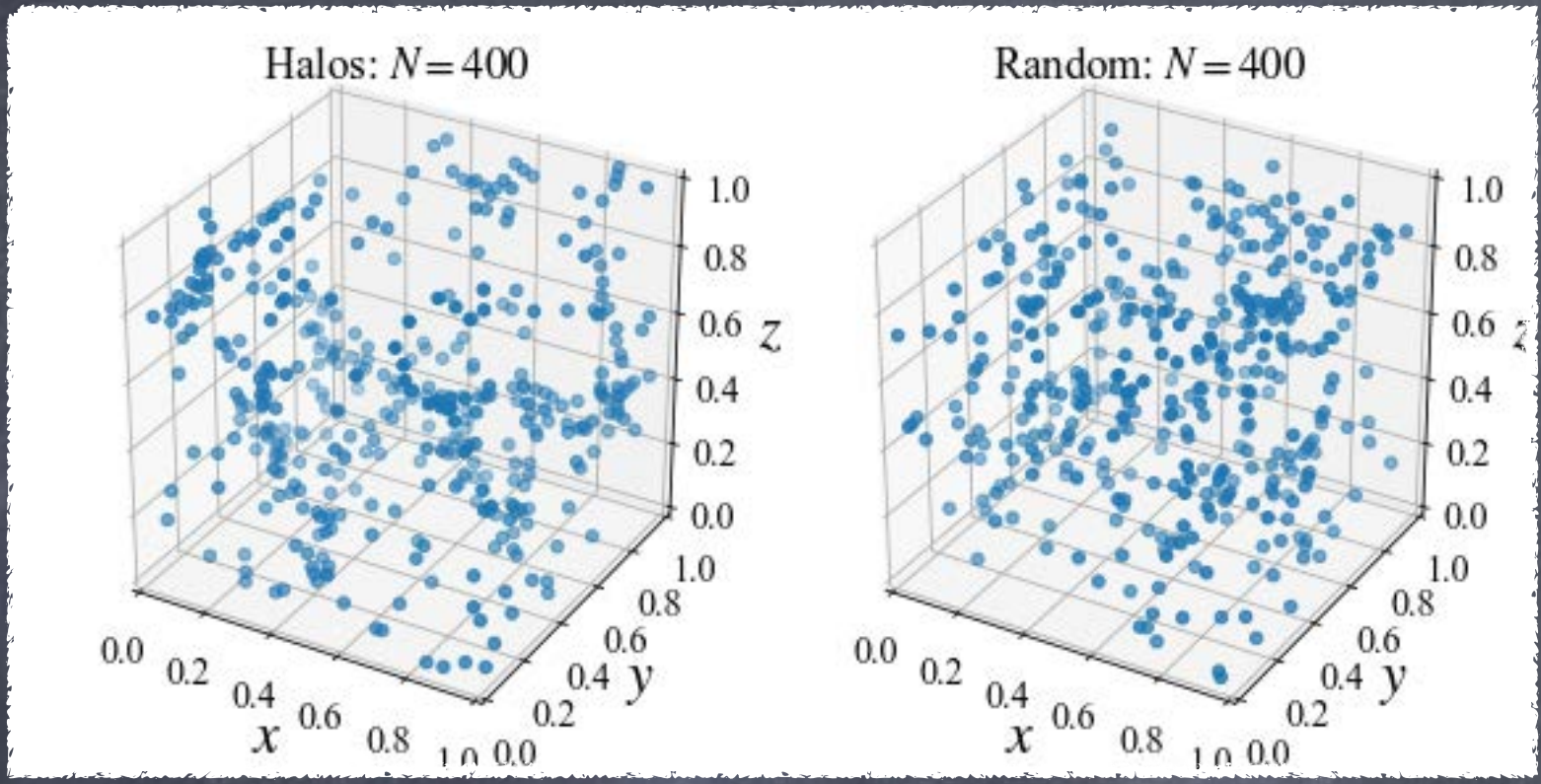

Y
X
0
20
40
60
80
100
2
4
6
8
10
12


Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20

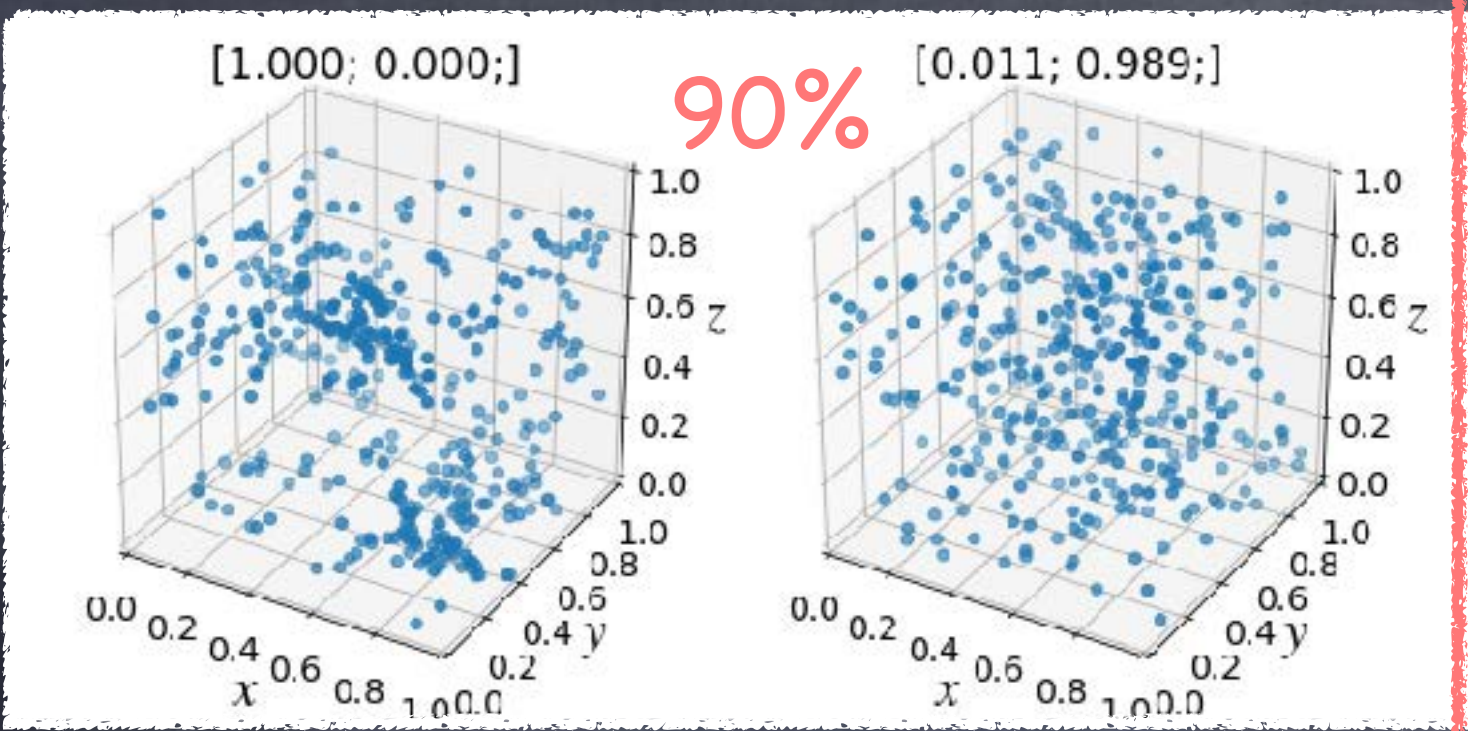

Sphere: N = 200
Sphere: N = 400
Sphere: N = 800

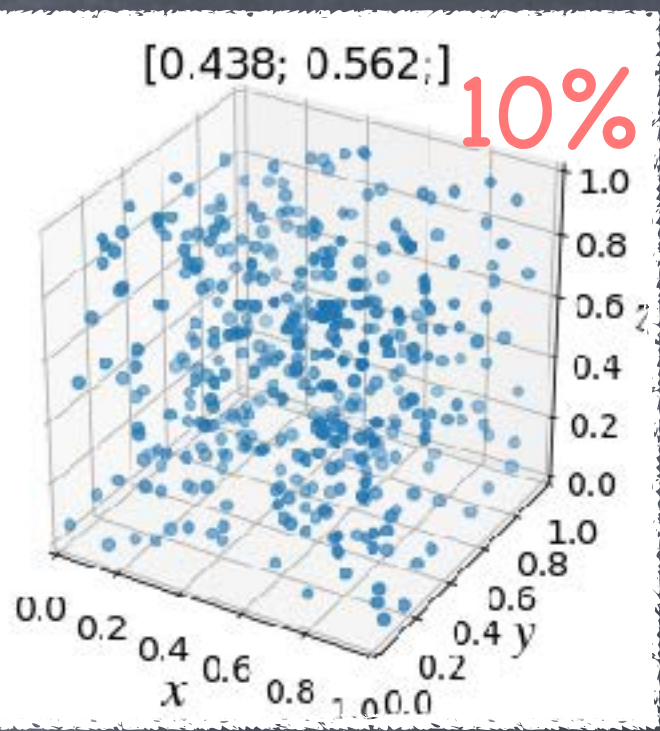


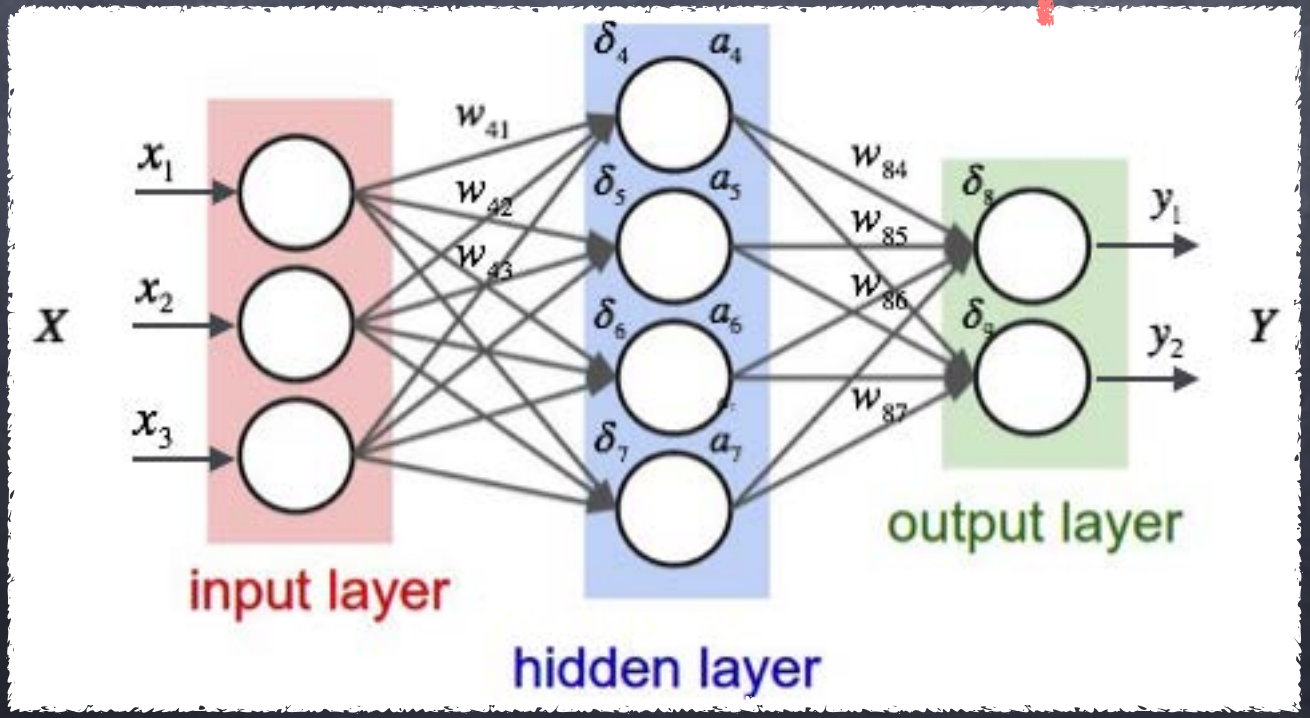

```
Epoch 47/50
 - 1s - loss: 0.0385 - acc: 0.9973 - val_loss: 0.3019 - val_acc: 0.8891
Epoch 48/50
 - 1s - loss: 0.0403 - acc: 0.9962 - val_loss: 0.3056 - val_acc: 0.8838
Epoch 49/50
 - 1s - loss: 0.0332 - acc: 0.9980 - val_loss: 0.3048 - val_acc: 0.8893
Epoch 50/50
 - 1s - loss: 0.0296 - acc: 0.9993 - val_loss: 0.3057 - val_acc: 0.8877
Baseline Error: 11.23%
CPU times: user 4min 20s, sys: 18.3 s, total: 4min 39s
Wall time: 51 s
```



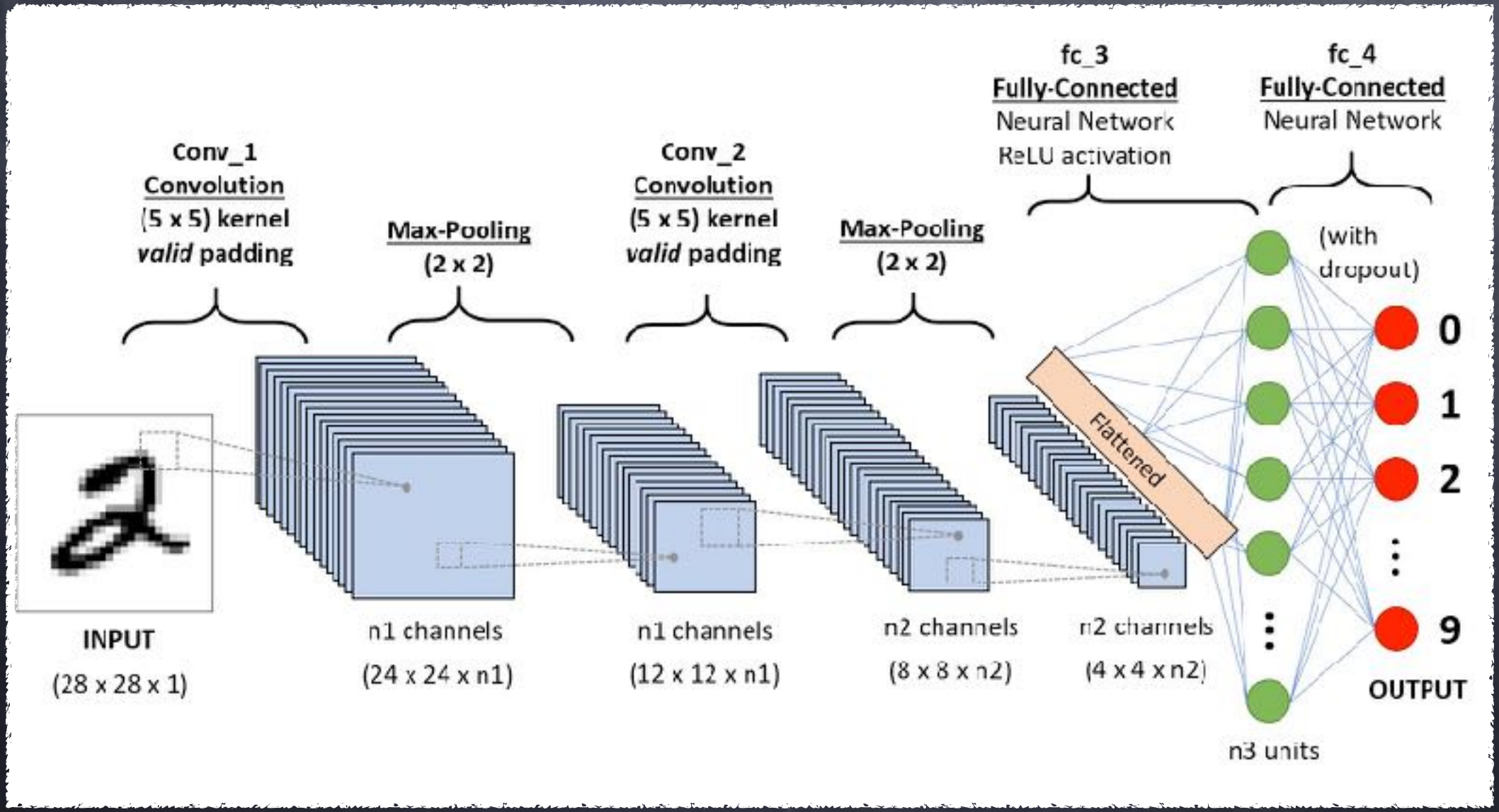

```
Epoch 47/50
 - 0s - loss: 0.0330 - acc: 0.9970 - val_loss: 0.2899 - val_acc: 0.8903
Epoch 48/50
 - 0s - loss: 0.0212 - acc: 1.0000 - val_loss: 0.2917 - val_acc: 0.8911
Epoch 49/50
 - 0s - loss: 0.0192 - acc: 1.0000 - val_loss: 0.3073 - val_acc: 0.8924
Epoch 50/50
 - 0s - loss: 0.0169 - acc: 1.0000 - val_loss: 0.2986 - val_acc: 0.8937
Baseline Error: 10.63%
CPU times: user 38.1 s, sys: 3.01 s, total: 41.1 s
Wall time: 24.9 s
```


[1.000; 0.000;]
90%
[0.011; 0.989;]
x
y
z

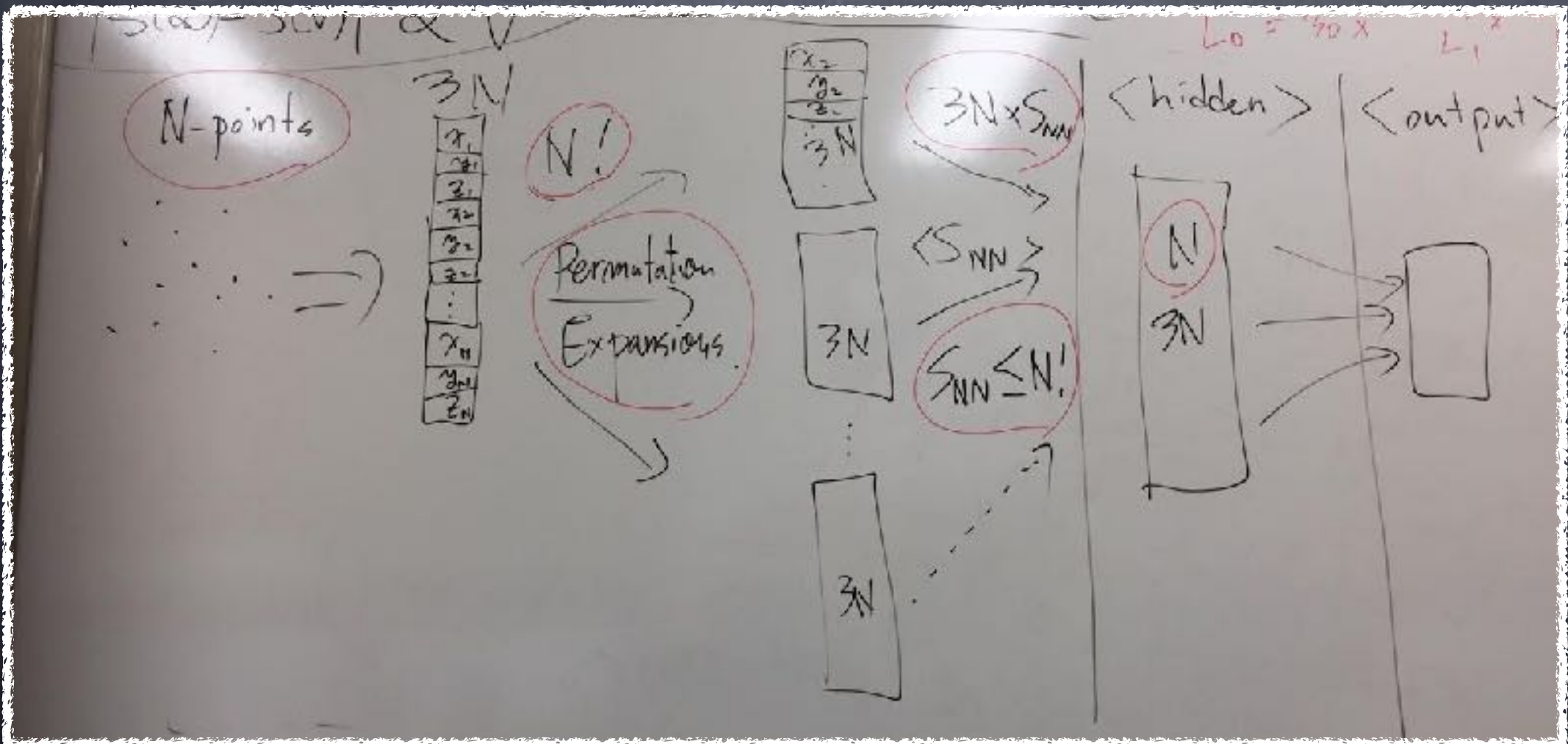

[0.438; 0.562;]
10%
x
y
z


X
Y
input layer
hidden layer
output layer

How to break the 90% barrier?



How to break the 90% barrier?




Point Pattern Classifier ...
Any Practical Application?
"Drone Shows"
generated by GANs ?
Thank you !