Applications of Deep Learning in Astronomy and Biology

Ajit Kembhavi
IUCAA

Ajit Kembhavi

Ninan Sajeeth Philip

Sheelu Abraham

Kaushal Sharma

Kaustubh Vaghmare

Aniruddha Kembhavi

Ashish Mahabal

Arun Aniyan

Rohan Pattnaik

Janesh, Radha, Anshul, Blesson, Kriti

How Does One Use Large Data Sets?

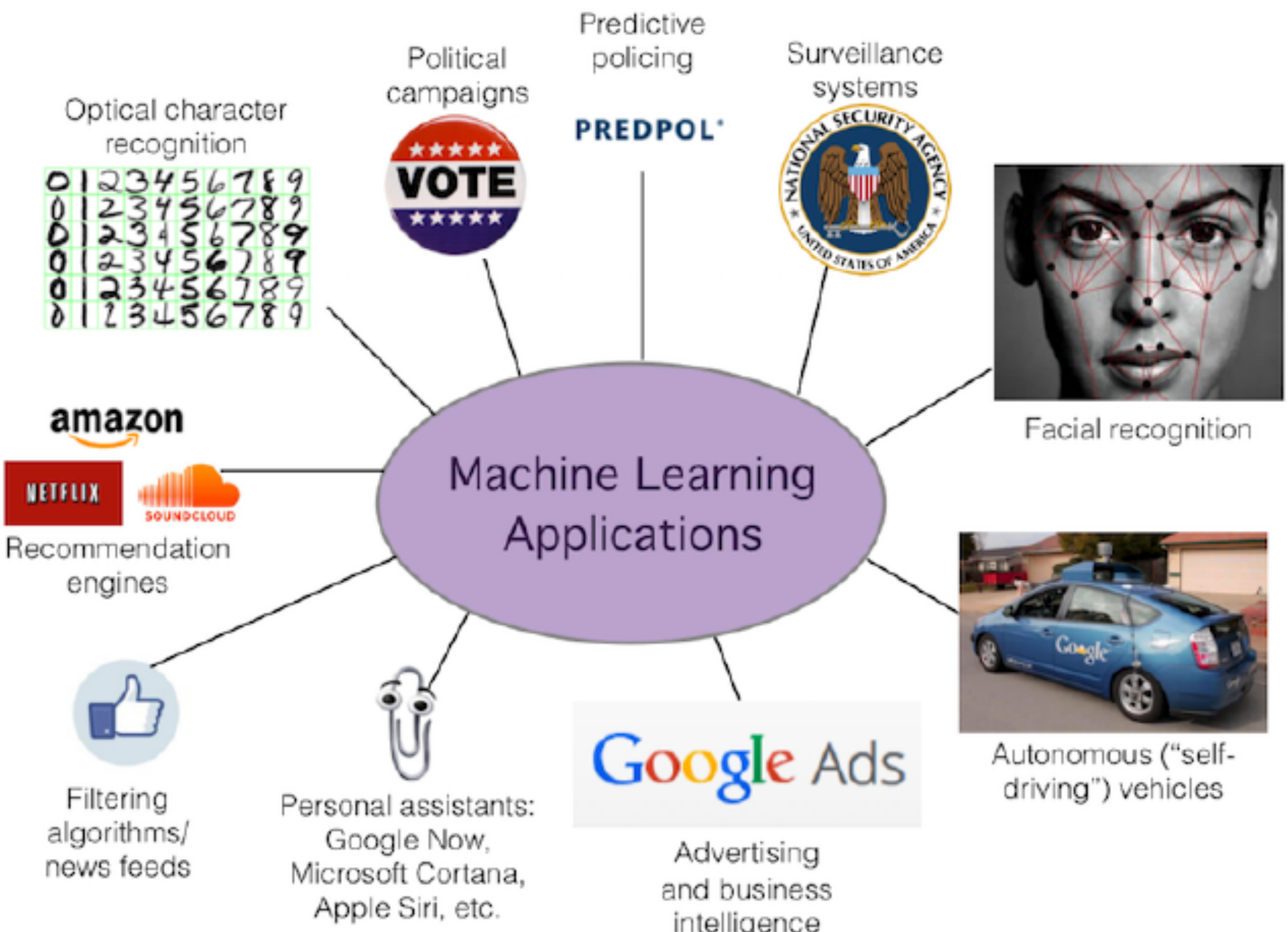
Find patterns, correlations, classes, outliers and meaning in the data

Data Analytics:

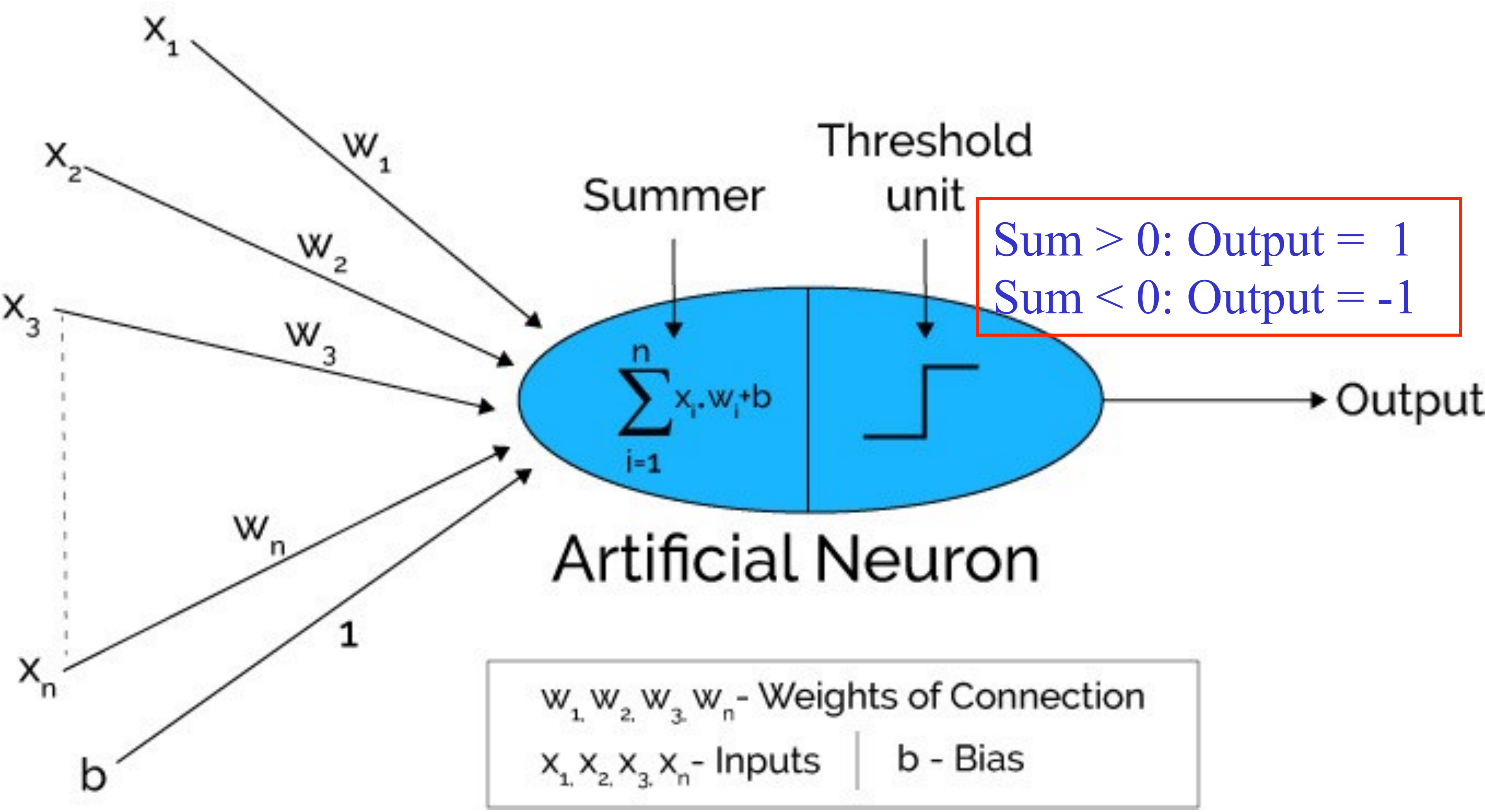
Domain Knowledge
Mathematics
Statistics

Visualisation
Data Mining

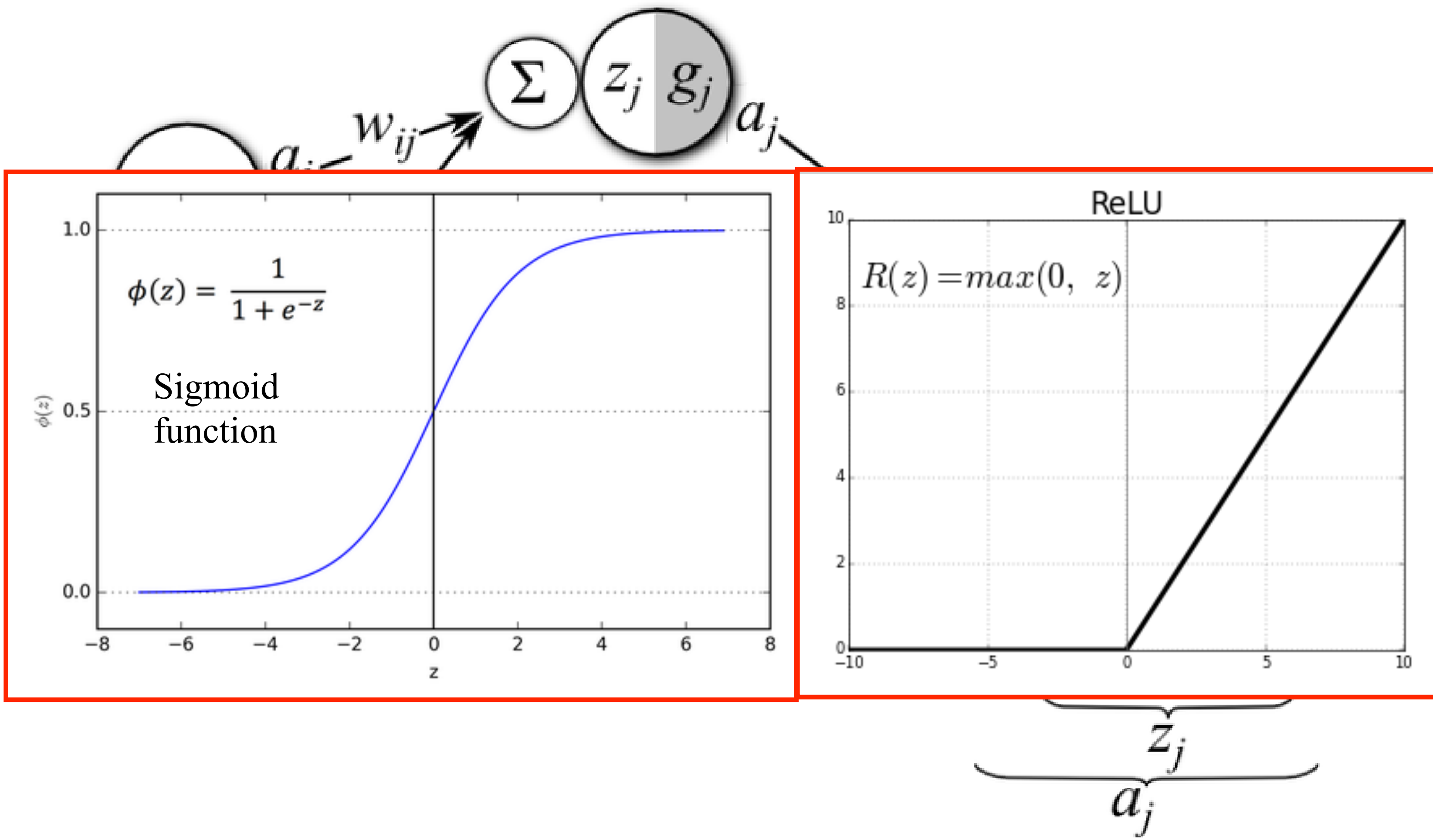
Machine Learning
Deep learning


Machine Learning Applications
Optical character recognition
Political campaigns
VOTE
Predictive policing
PREDPOL
Surveillance systems
NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA
Facial recognition
amazon
NETFLIX
SOUNDCLOUD
Recommendation engines
Autonomous ("self-driving") vehicles
Filtering algorithms/ news feeds
Personal assistants: Google Now, Microsoft Cortana, Apple Siri, etc.
Google Ads
Advertising and business intelligence

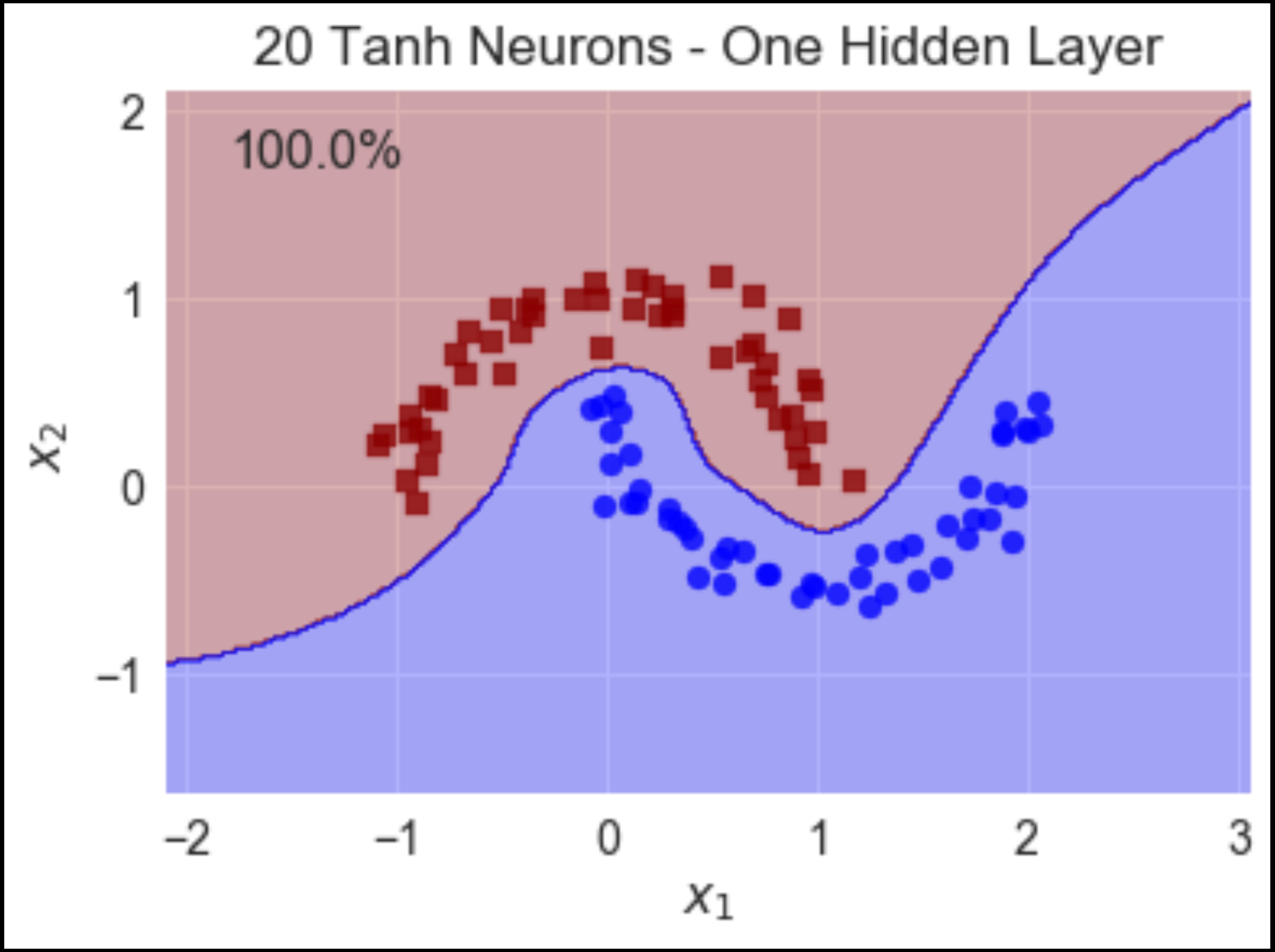
Machine Learning: Perceptrons



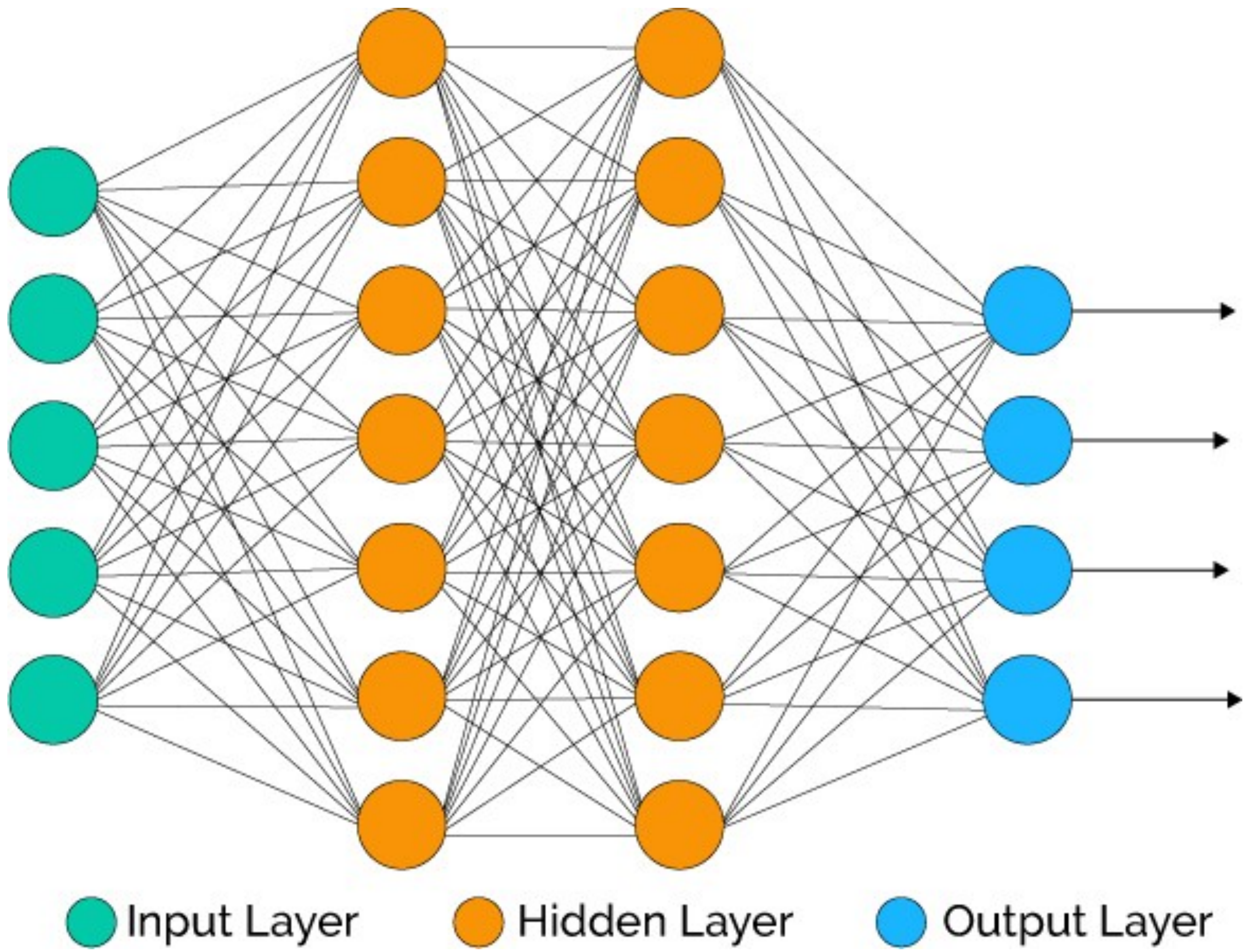
Sum > 0: Output = 1
Sum < 0: Output = -1



Σ
z_j g_j
w_{ij}
a_j
$\phi(z) = \frac{1}{1+e^{-z}}$
Sigmoid function
$\phi(z)$
z
ReLU
$R(z) = max(0, z)$
z_j
a_j


20 Tanh Neurons - One Hidden Layer
100.0%
2
1
0
−1
−2
−1
0
1
2
3
x_1
x_2

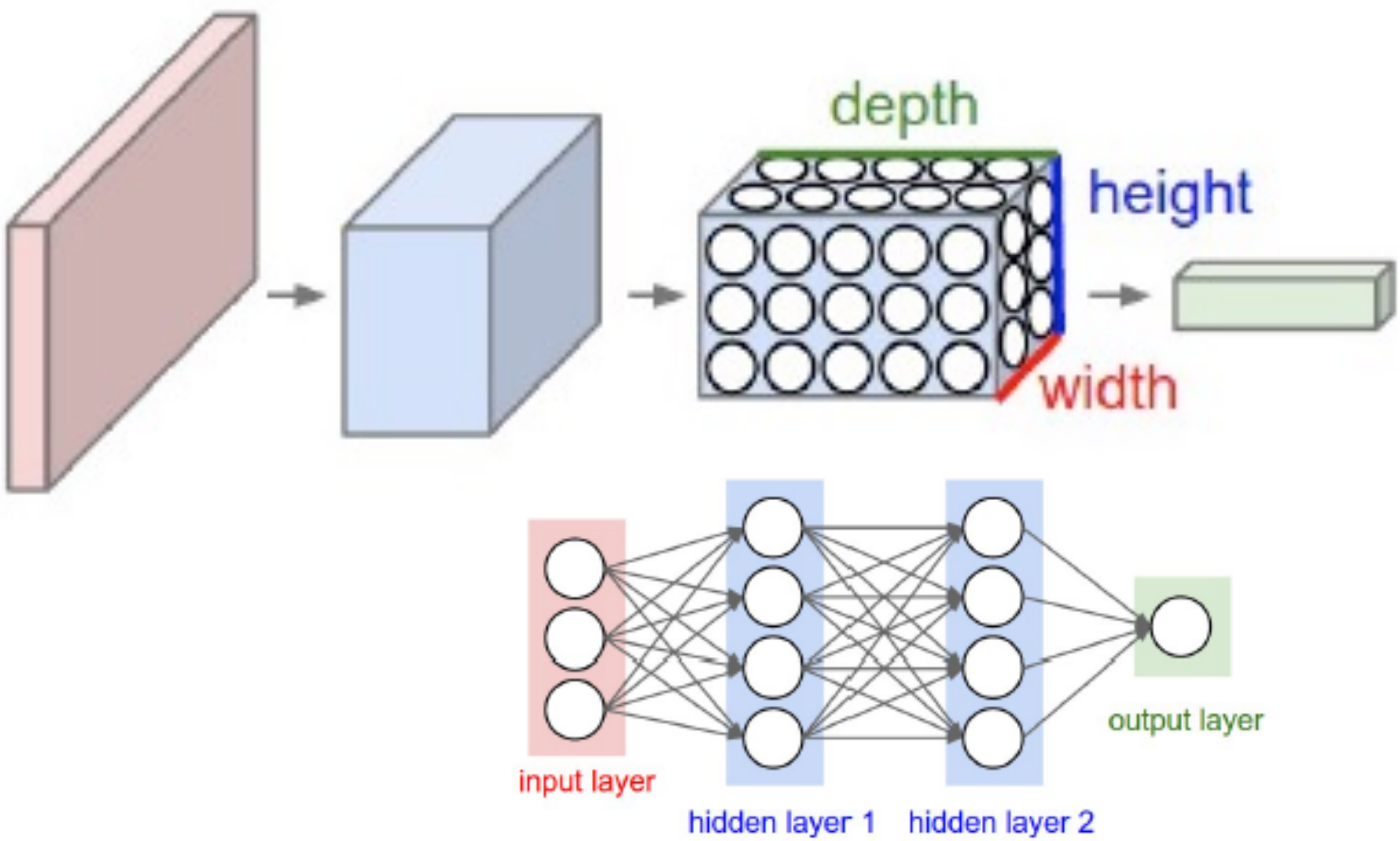
Multilayer Network



Deep Learning

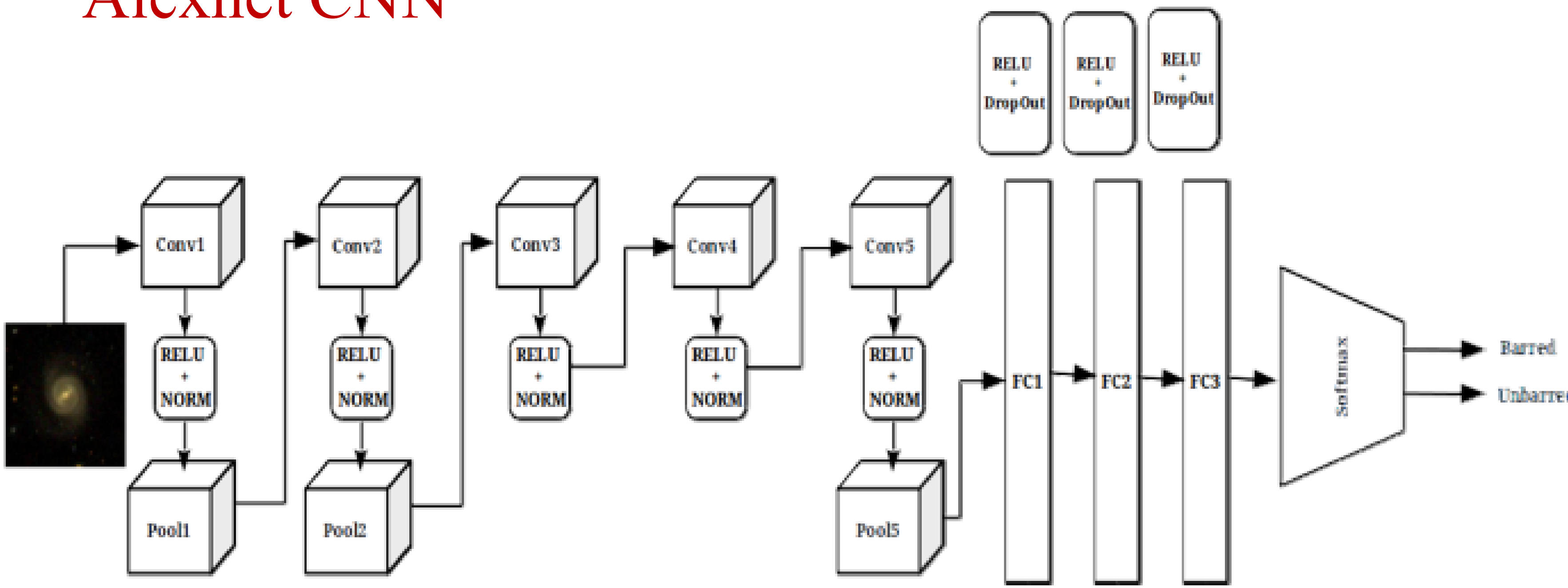
- Artificial neural networks work on features extracted from the data, for example images.
- Deep learning networks work directly on the data, extracting useful features from the data and downsizing it.
- Deep learning networks can therefore address very complex data which would be intractable for the conventional networks.

Convolutional Neural Network



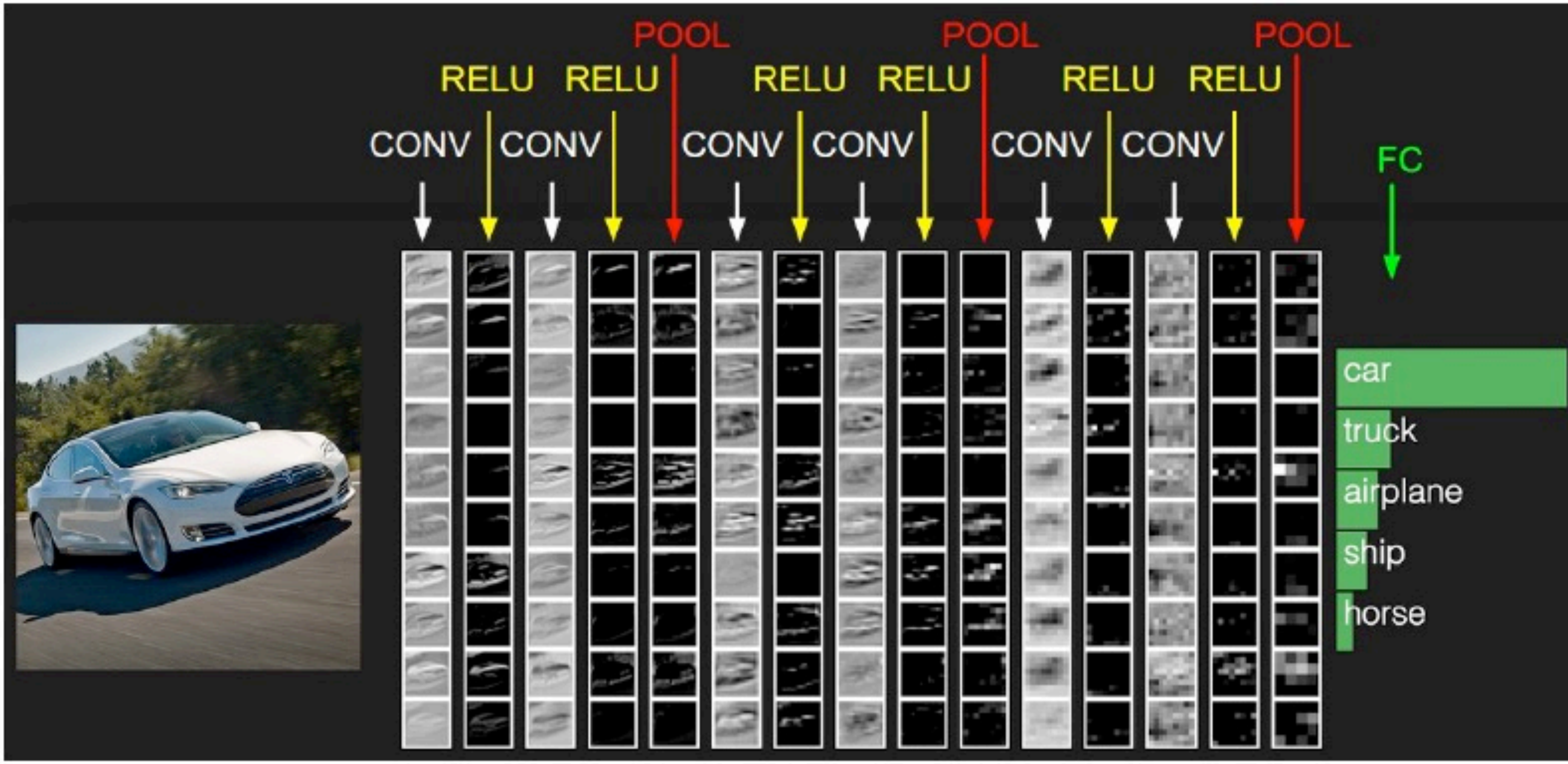
Network Architecture

Alexnet CNN



12 layers with
5 convolutional
layers

Sheelu Abraham+ 2017


POOL
POOL
POOL
RELU
RELU
RELU
RELU
RELU
RELU
CONV
CONV
CONV
CONV
CONV
CONV
FC
car
truck
airplane
ship
horse

Bar Detection in Galaxies

M 51

Barred Galaxies

NGC 1300, HST

Bars are important dynamical features in galaxies. They break axial symmetry and lead to flow of stars and gas towards the centre, leading to build up of the bulge. How frequent are bars and how are influenced by galaxy type and environment?

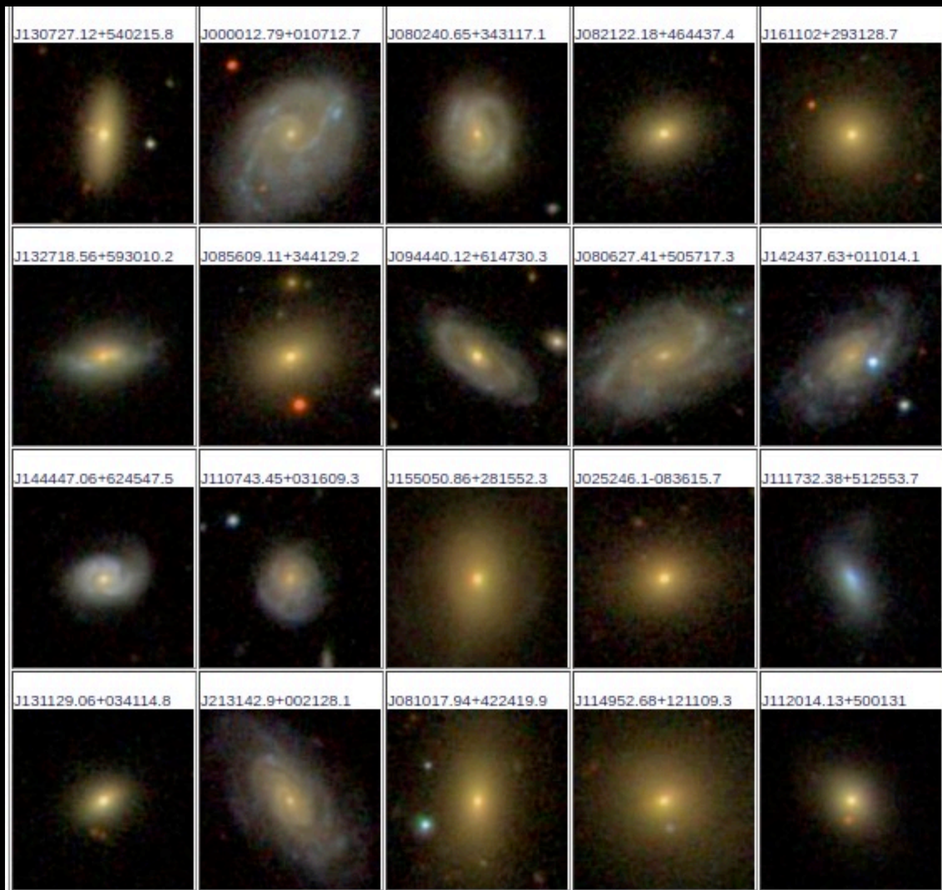
Discover Barred Galaxies Using CNN

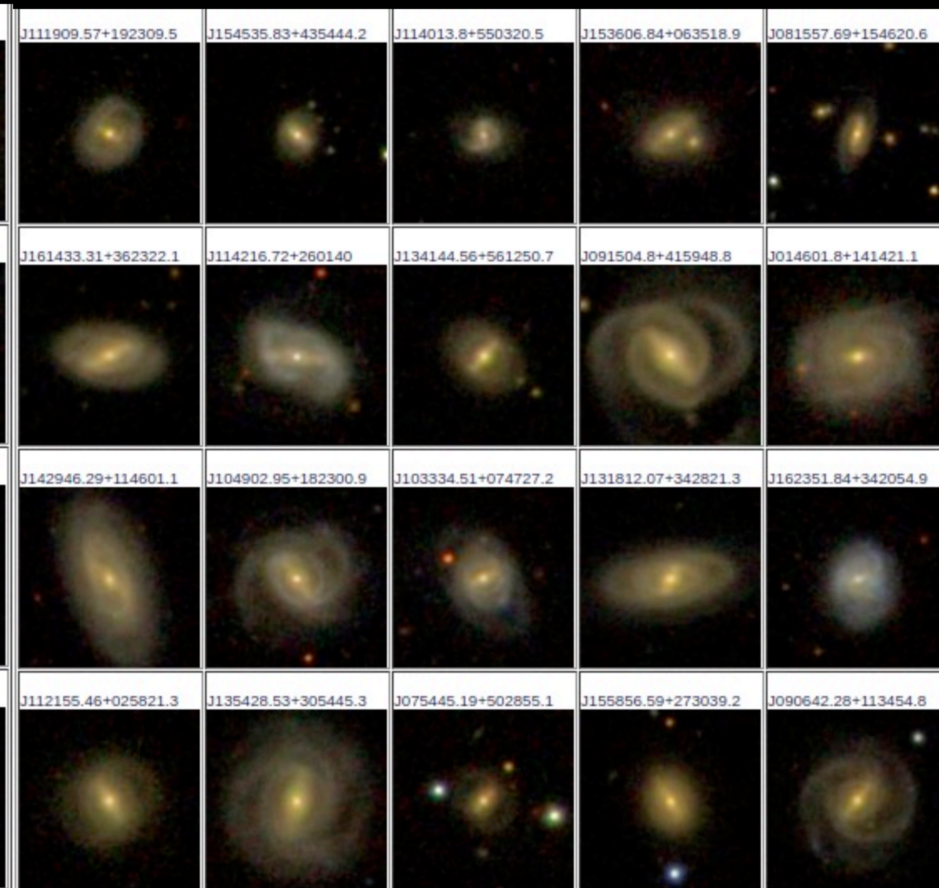
- Bars in galaxies are discovered through visual inspection or detailed quantitative study of galaxy morphology.
- Process is time consuming, and would be impossible to apply to millions of galaxies in large surveys.
- Use Deep Learning with a large training sample of known barred and unbarred galaxies.

Galaxy Sample Selection

- A sample of galaxies is first selected from the Sloan Digital Sky Survey DR13
- Selected galaxies have r magnitude in the range $14 < r < 17.4$, redshift $z < 0.2$ and half light radius between 5 and 30 arcsec.
- gri colour composite images are used.
- Galaxies are cross matched catalogues of galaxies which have barred and unbarred galaxies (Nair & Abraham 2010, Galaxy Zoo DR2 Willett+ 2013).

Images scaled to de Vaucouleurs radius



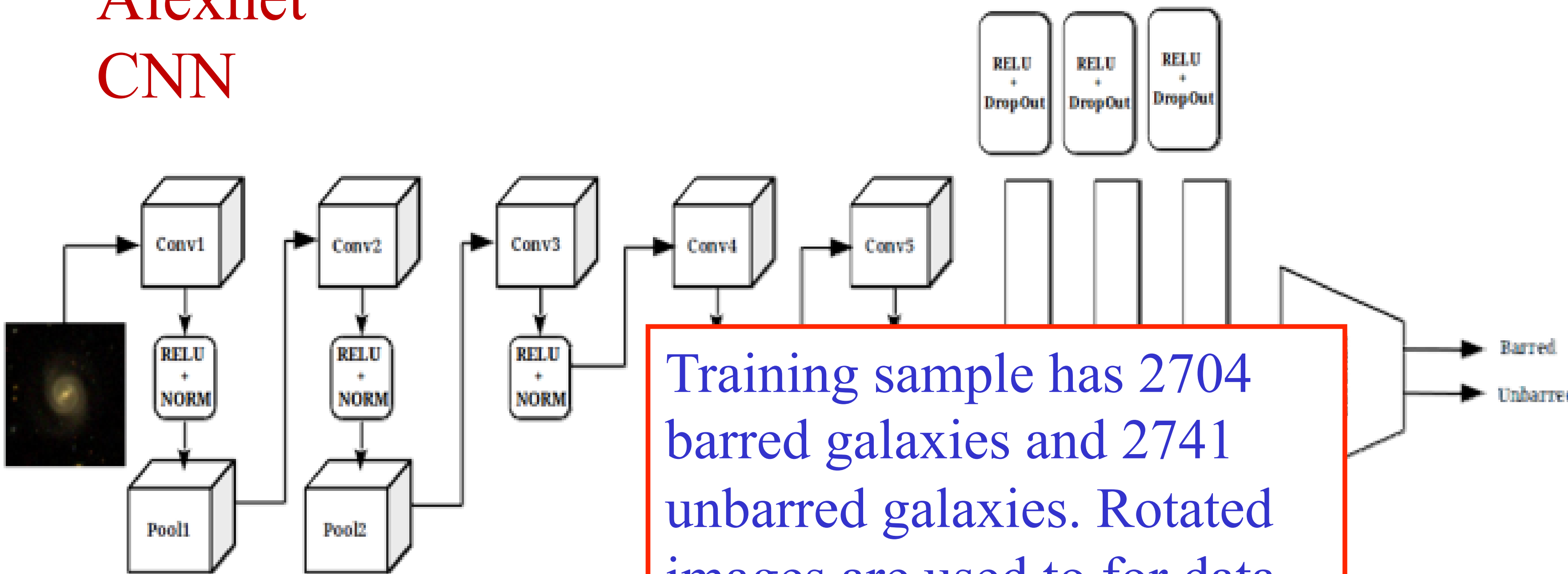


Unbarred Galaxies

Barred Galaxies

Network Architecture

Alexnet
CNN

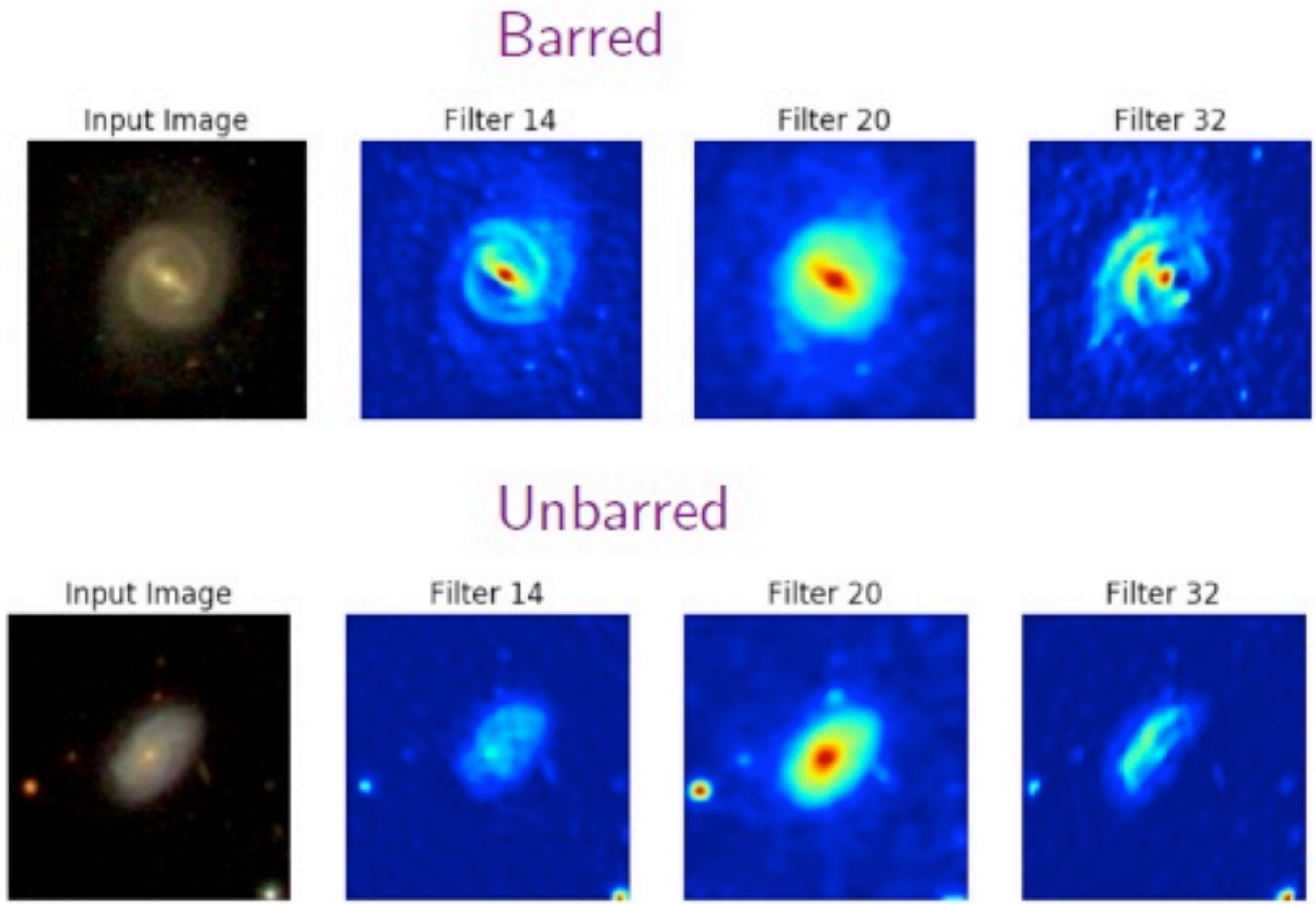


Training sample has 2704 barred galaxies and 2741 unbarred galaxies. Rotated images are used to for data augmentation.

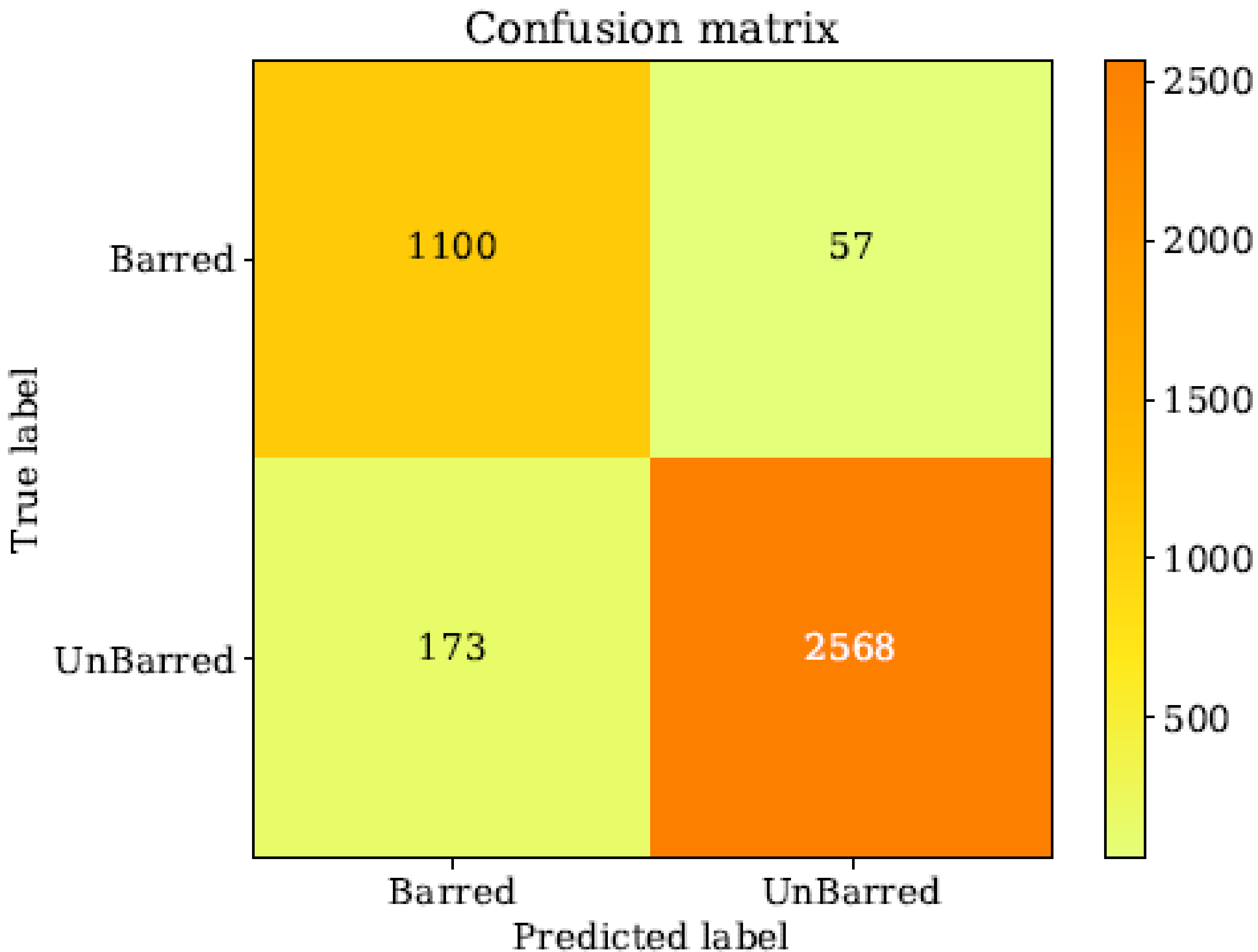
Twelve layers with five convolutional layers

Sheelu Abraham+ 2017

Visualisation of Layers



How Good is the Network?



Sheelu Abraham+ 2017

	Precision %	Recall %	Number in Sample
Barred	86.41	95.07	1157
Unbarred	97.83	93.69	2741
Average	94.1	94.1	3898

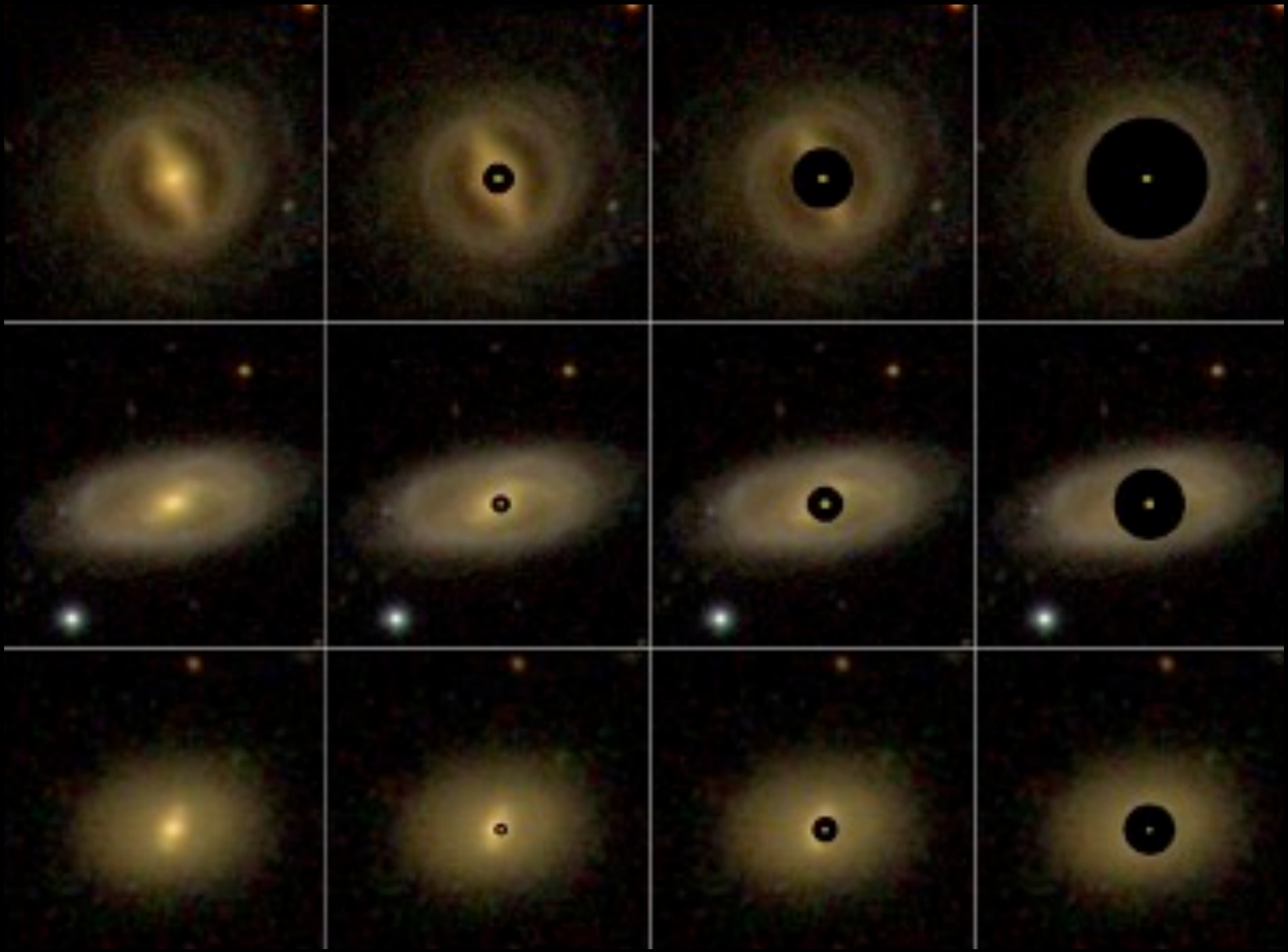
$$\text{Precision} = \frac{\text{Correctly classified as barred}}{\text{Total classified as barred}}$$

$$\text{Recall} = \frac{\text{Correctly classified as barred}}{\text{Actual number of barred}}$$

Observationally unbarred,
classified as barred

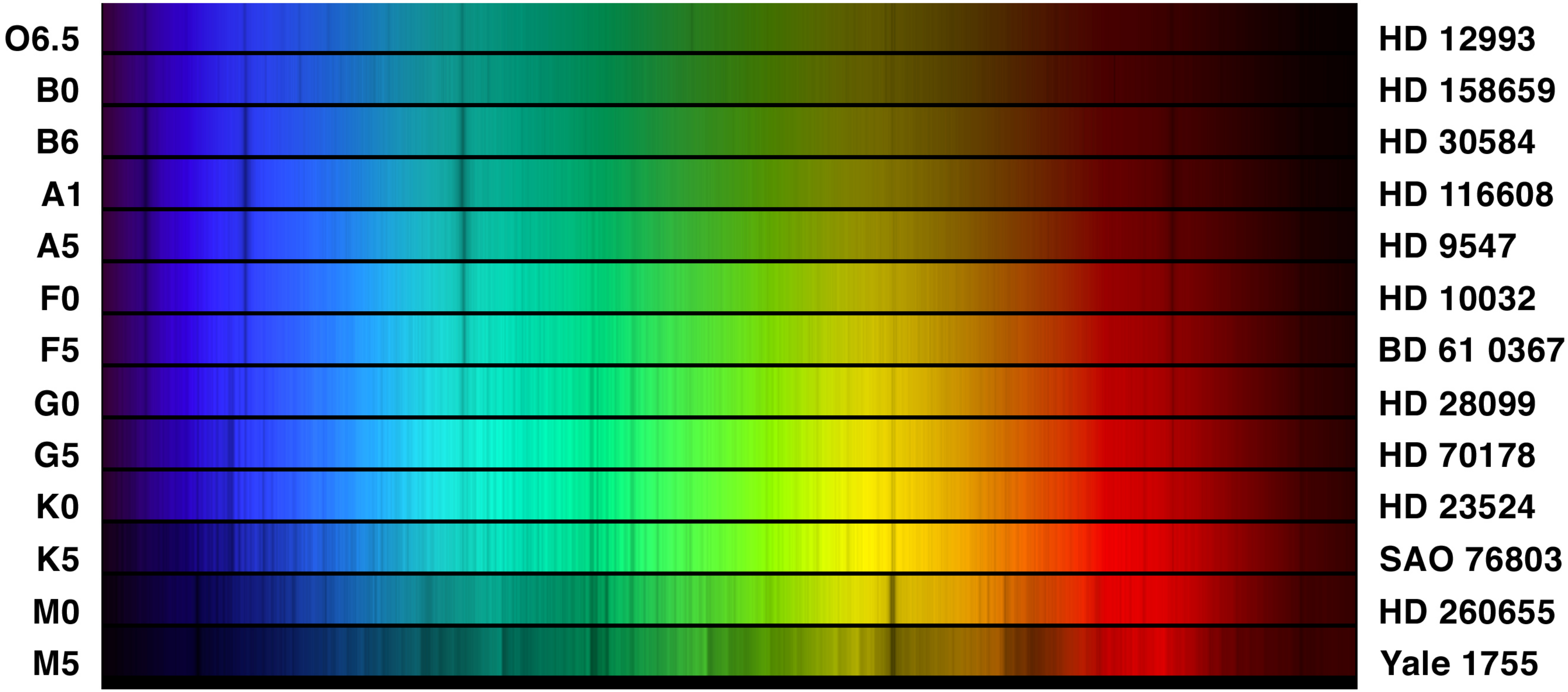
Observationally barred,
classified as unbarred

Occlusion Test Covering Barred Region



Spectral Classification of Stars

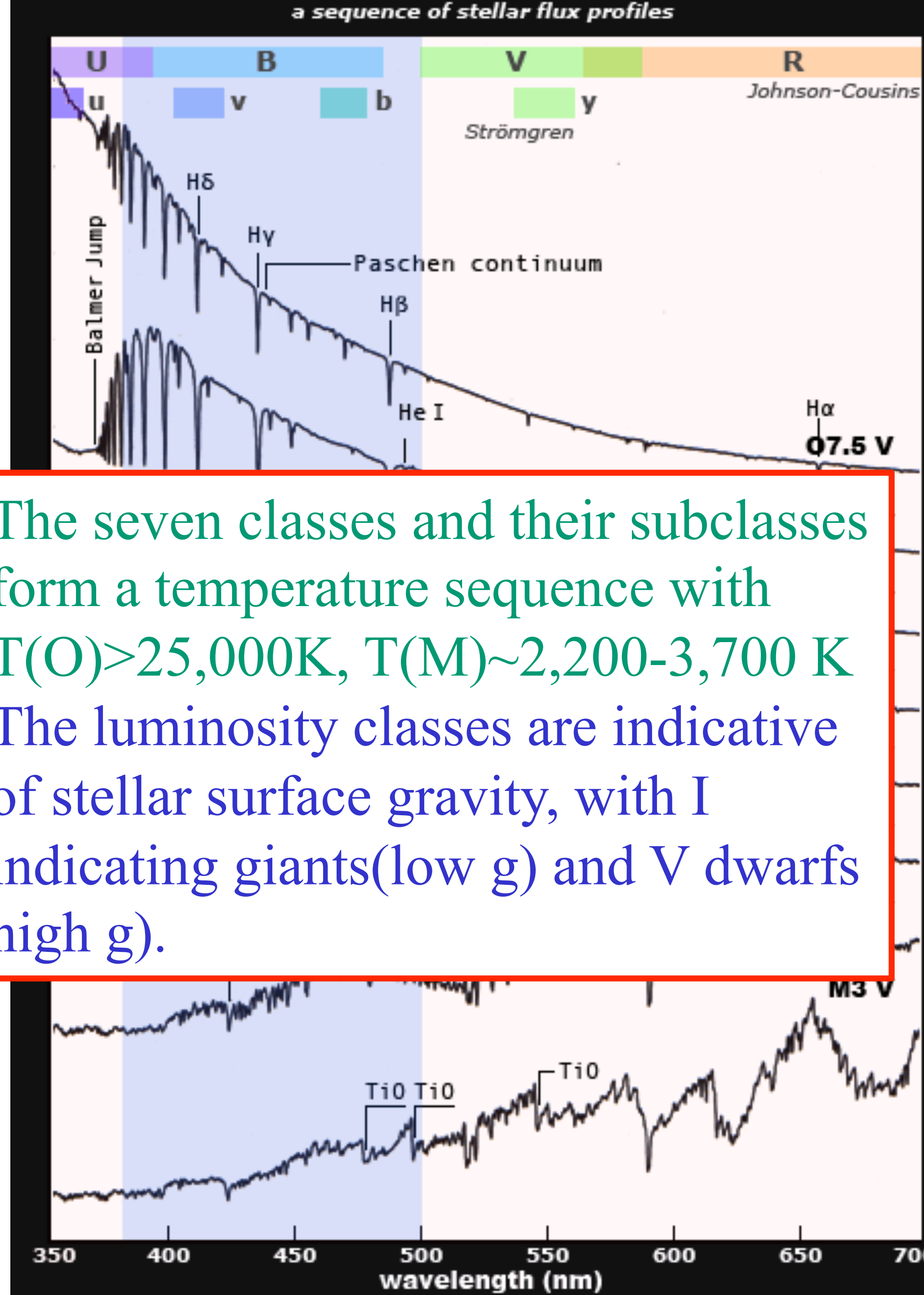
Stellar Spectra



Stellar Spectral Classes

Harvard Classification System:

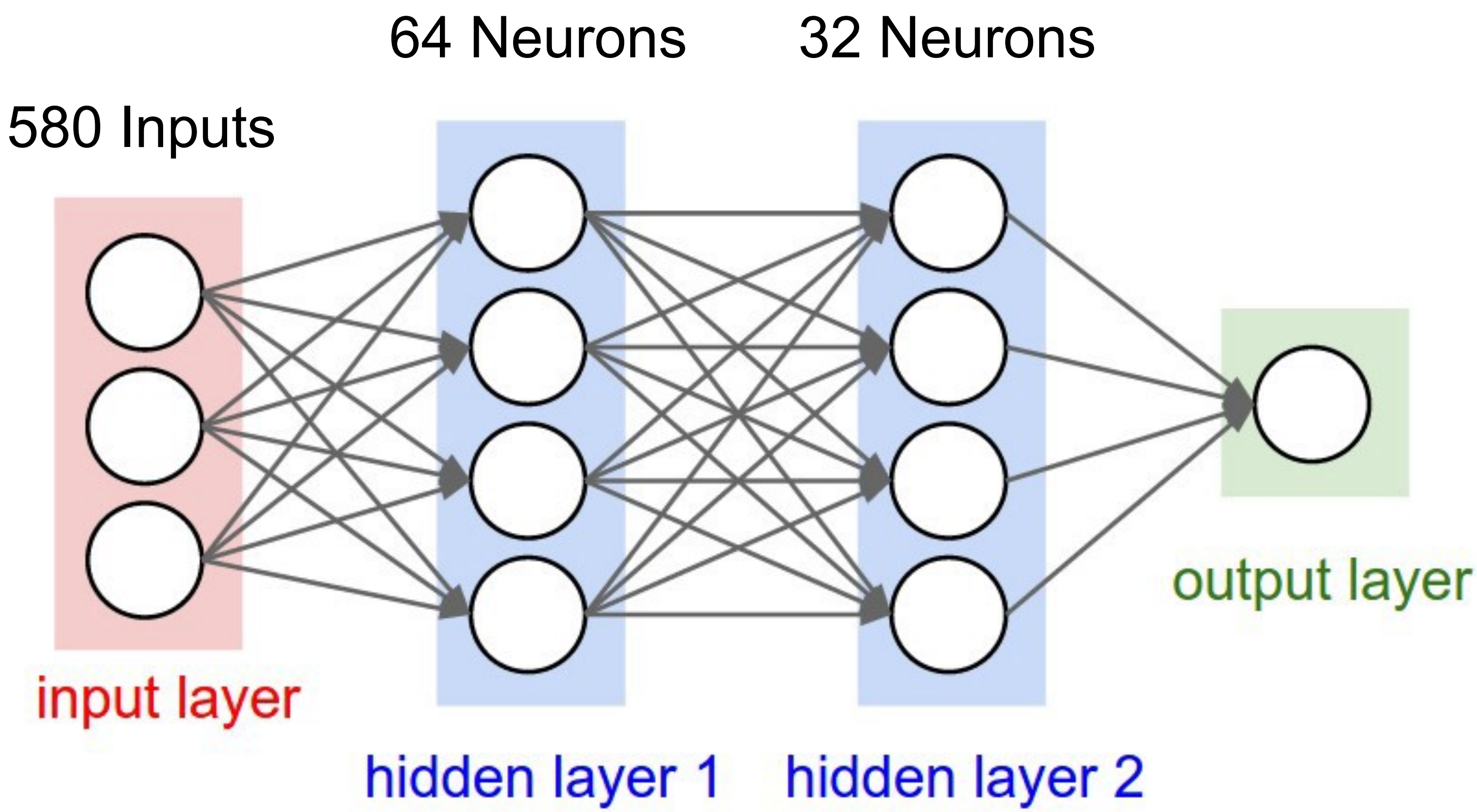
- Seven main classes O, B, A, F, G, K, M
- Ten subclasses in each class
- Five luminosity classes I-V



The seven classes and their subclasses form a temperature sequence with T(O)>25,000K, T(M)~2,200-3,700 K The luminosity classes are indicative of stellar surface gravity, with I indicating giants(low g) and V dwarfs high g).

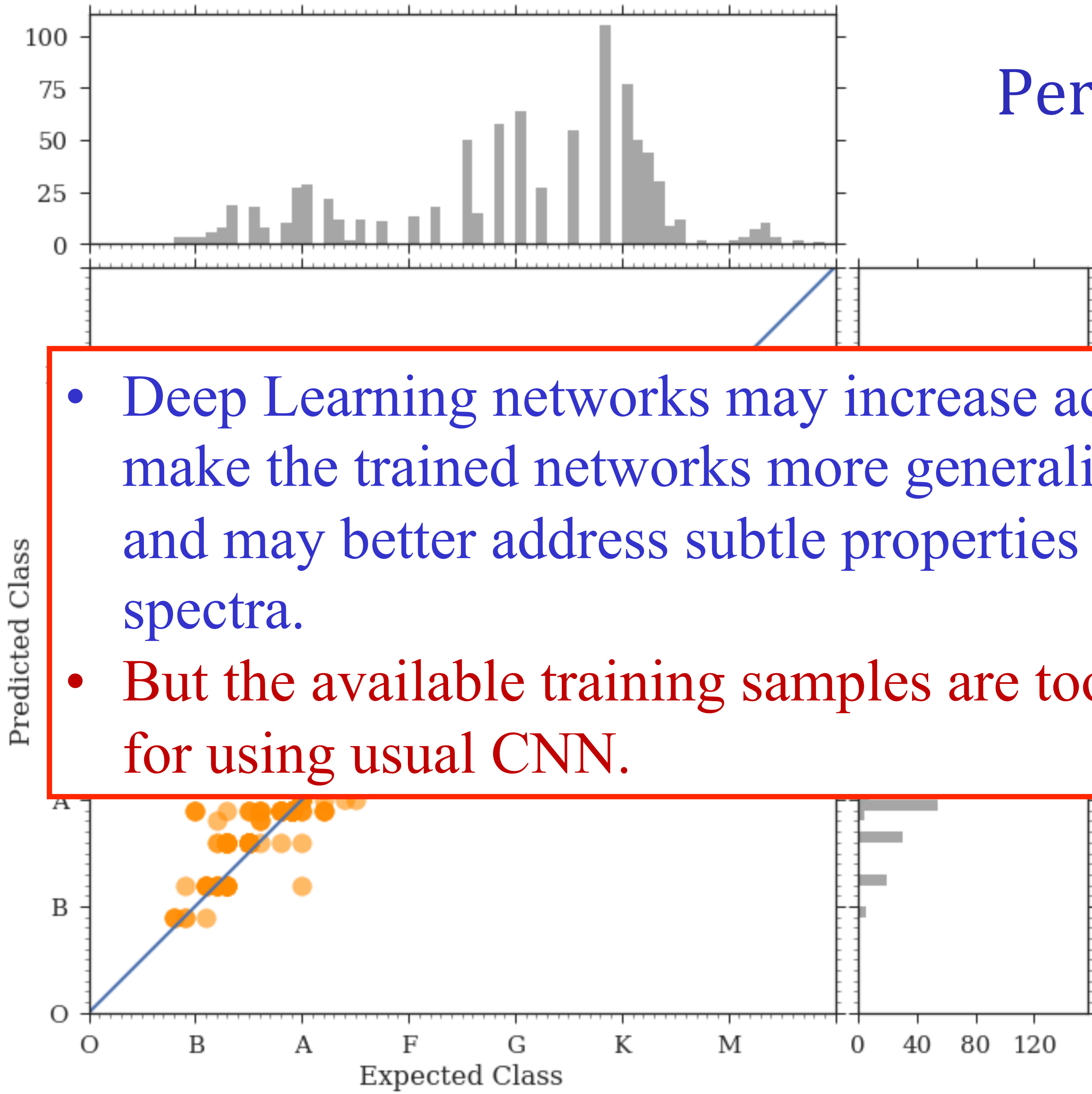
ANN

The classification problem is converted to a regression problem using
spectral code = 1000*A1 + 100*A2 +2*A3 + 1.5



Keras ANN

Performance



- Deep Learning networks may increase accuracy, make the trained networks more generalizable, and may better address subtle properties of the spectra.
- But the available training samples are too small for using usual CNN.

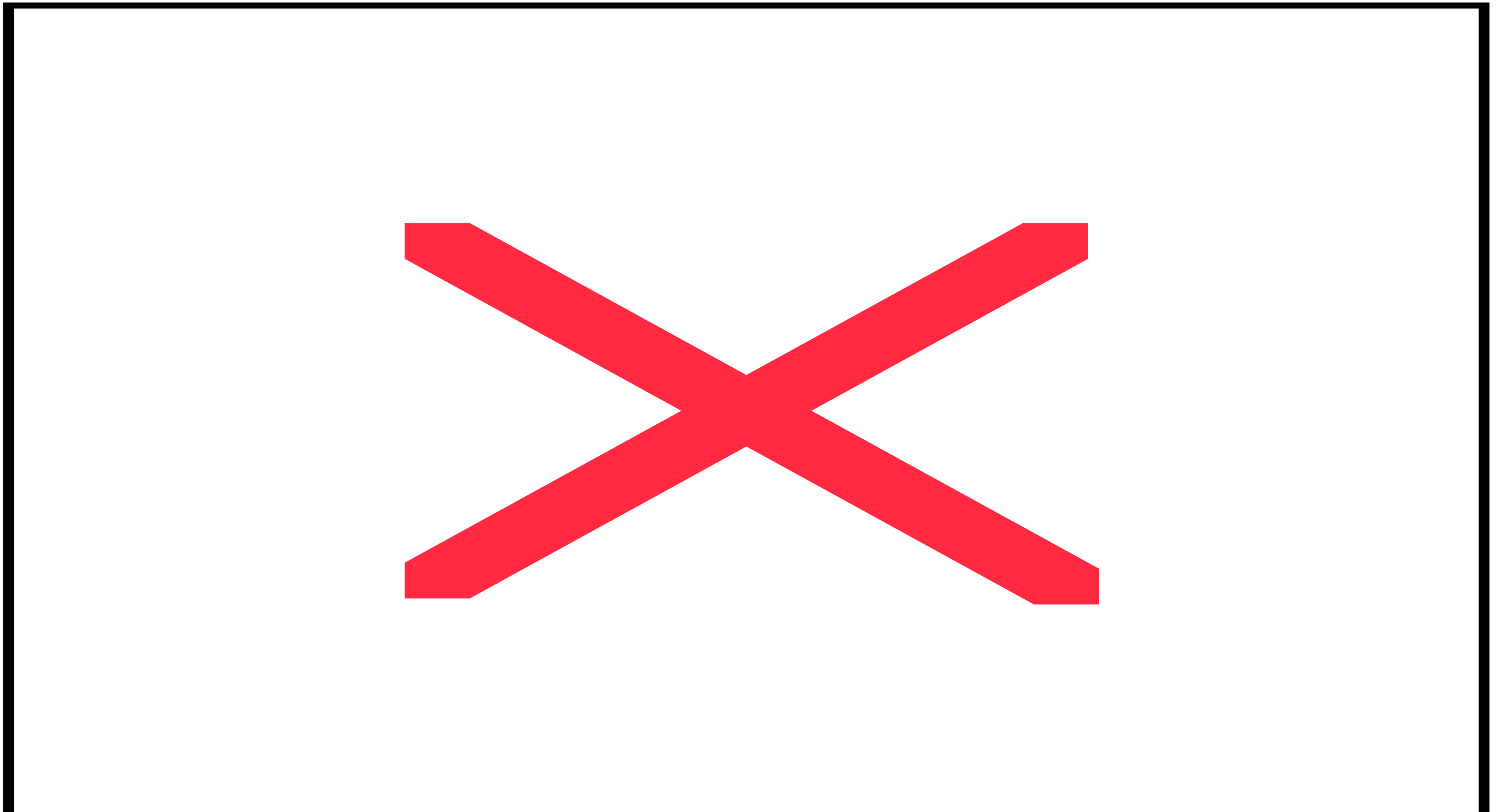
Autoencoders have many applications including

- Dimensionality Reduction
- Denoising
- Data Compression
- Outlier Detection

- networks where the output is the same as the input.
- The encoder compresses the input into a lower-dimensional code. Then the decoder reconstructs the input using only the code.
- An Autoencoder is an *unsupervised* learning technique as it does not need labels to train on. But they can be considered to be *self-supervised* as they generate their own labels from the training data.
- A loss function compares the output with the target.

Autoencoder as Stellar Classifier

- First train an autoencoder with ~60,000 stellar spectra from SDSS.
- Remove the decoding layer, append a fully connected ANN classifier to the trained encoding layer.
- Train this model with labelled spectral data from training set.
- With this supervised training, the encoding layers are fine tuned and the weights are readjusted to classify stellar spectra.

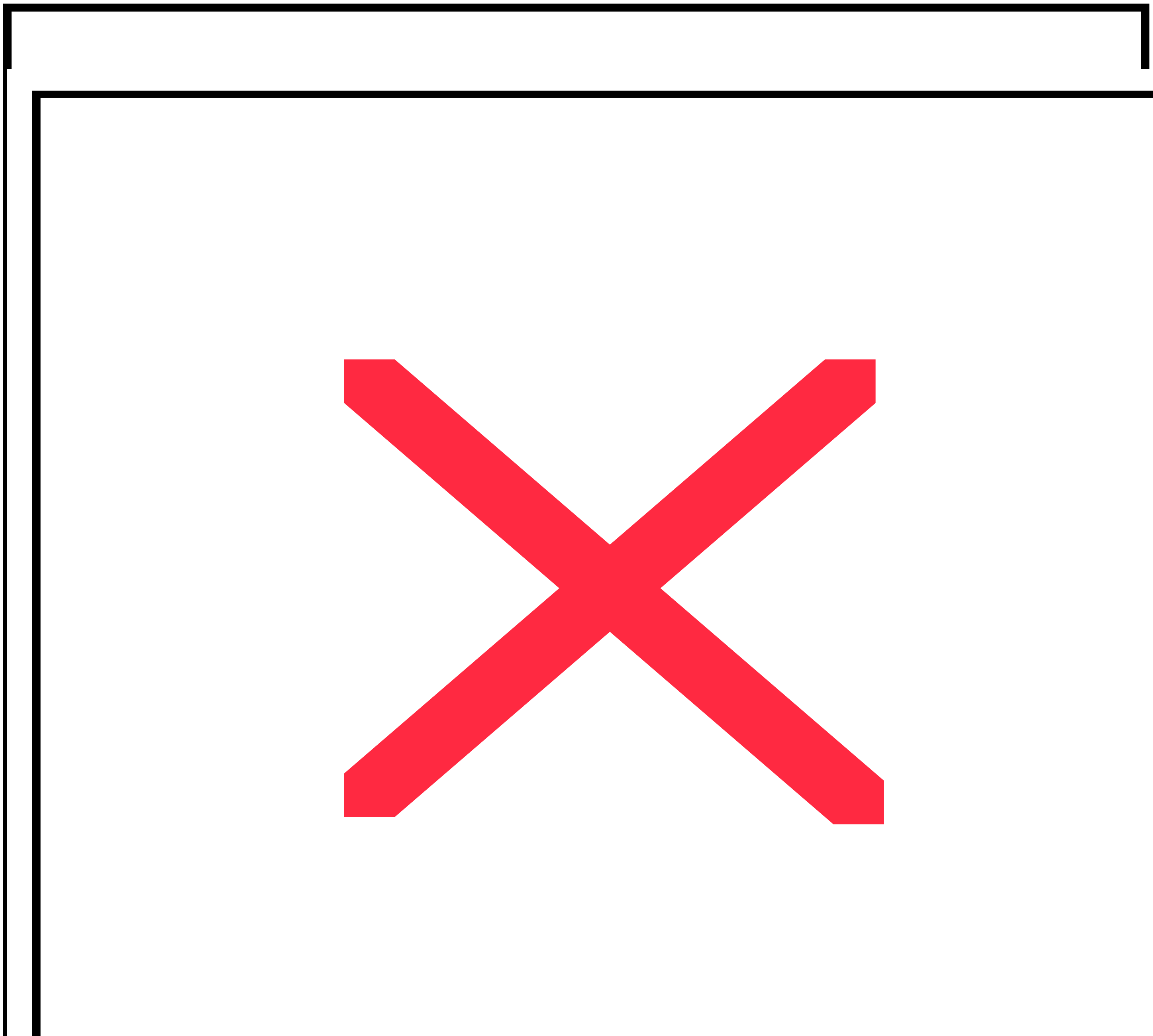


Training and Test Samples

Database	No. of stars/ Selected sample	λ Coverage (Å)	FWHM Resolution (Å) ($R = \lambda/\Delta\lambda$)	Reference
JHC Atlas ●	161/158	3510 - 7427	4.50 ($R \sim 1200$)	Jacoby et al. (1984)
ELODIE.3.1 ●	1959/1248	3900 - 6800	0.57 ($R \sim 10000$)	Prugniel et al. (2007)
Indo - US ●	1273/850	3460 - 9464	1.00 ($R \sim 5000$)	Valdes et al. (2004)
MILES ●	985/453	3536 - 7410	2.56 ($R \sim 2000$)	Sánchez-Blázquez et al. (2006)
Kesseli Templates	324/319	3650 - 10200	2.5 ($R \sim 2000$)	Kesseli et al. (2017)
Kesseli Original Sample	5630/4888	3650 - 10200	2.5 ($R \sim 2000$)	Kesseli et al. (2017)

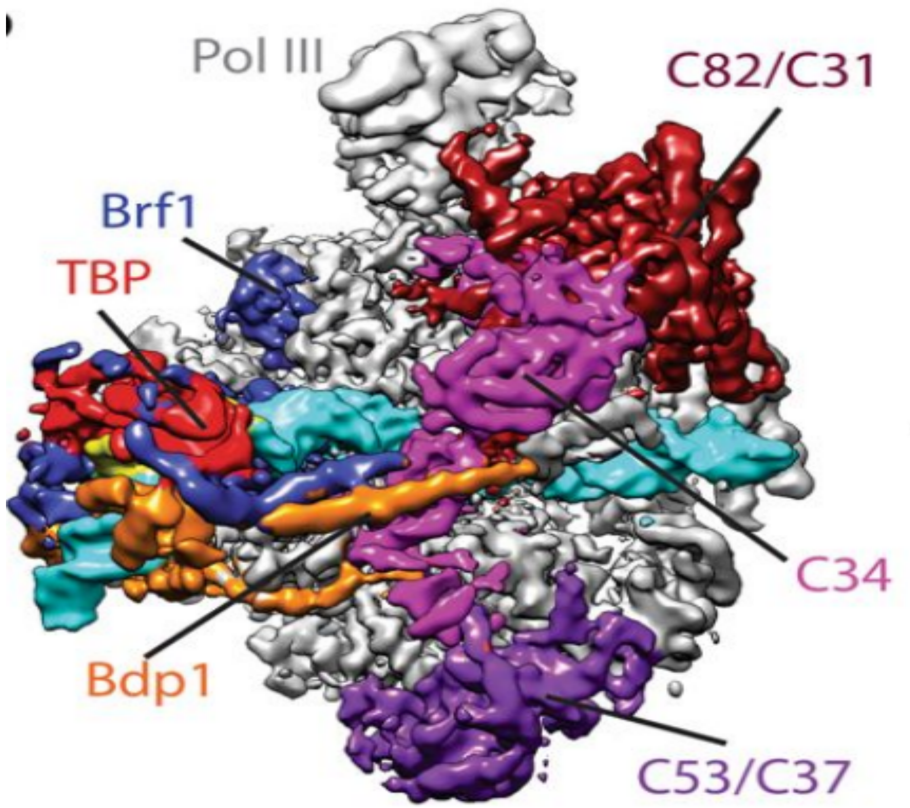
	Feature Matrix Size	Label matrix Size	Test Matrix (CFLIB) Size
● Train Set A	1859×580	1859×1	850×580
Train Set B	4886×1900	4886×1	850×1900

- Test



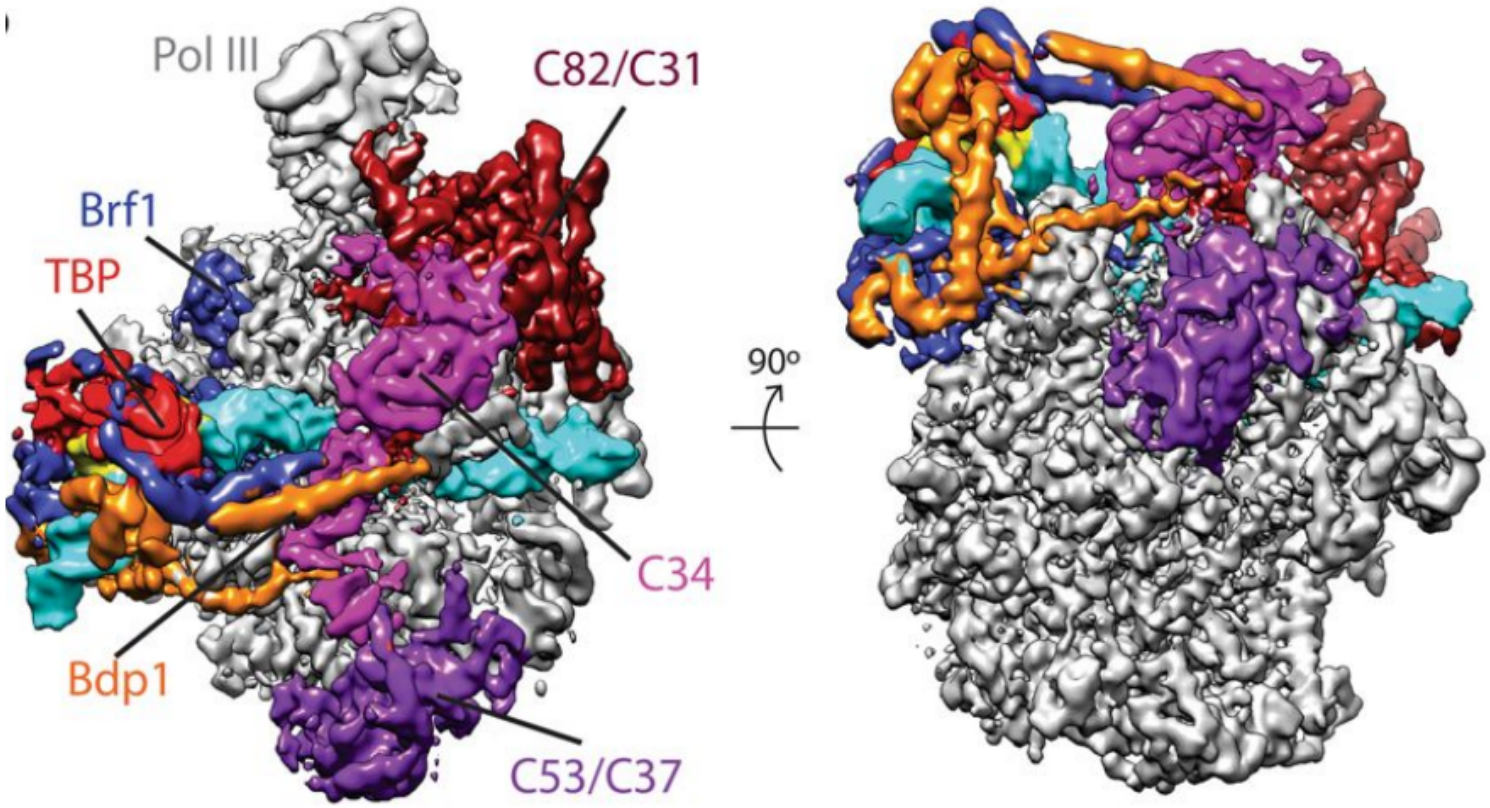
Protein identification in 2D images

RNA Polymerase III



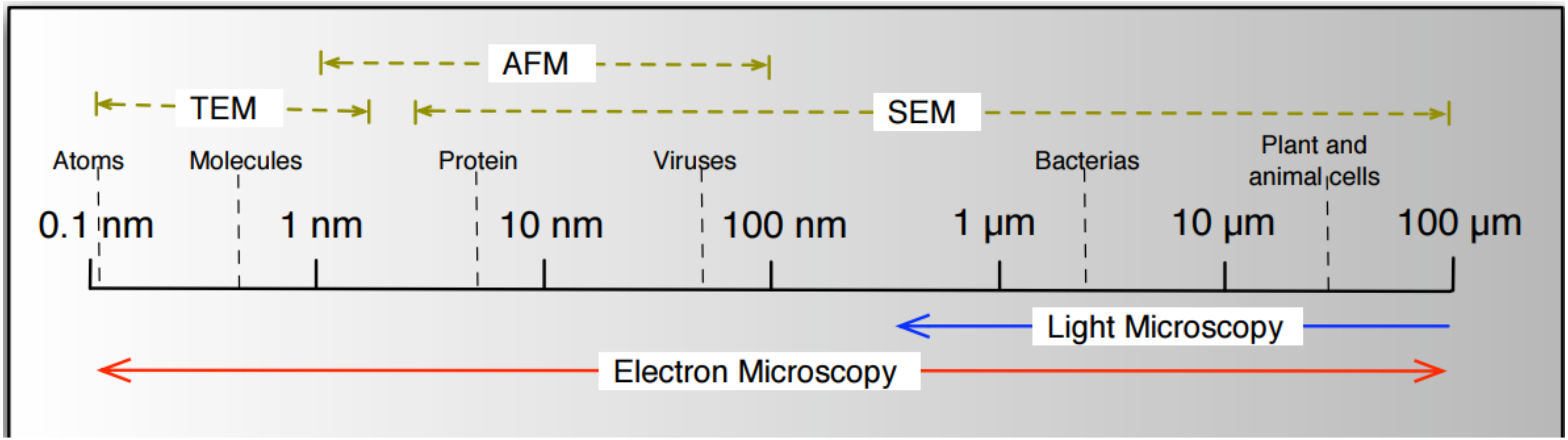
The challenge is to construct the 3D structure from the 2D projections in the SEM image.

Depending on the orientation, the 3D structure may appear differently in the Scanning Electron Microscope



Sajeeth Philip+ 2019

Electron Microscopy

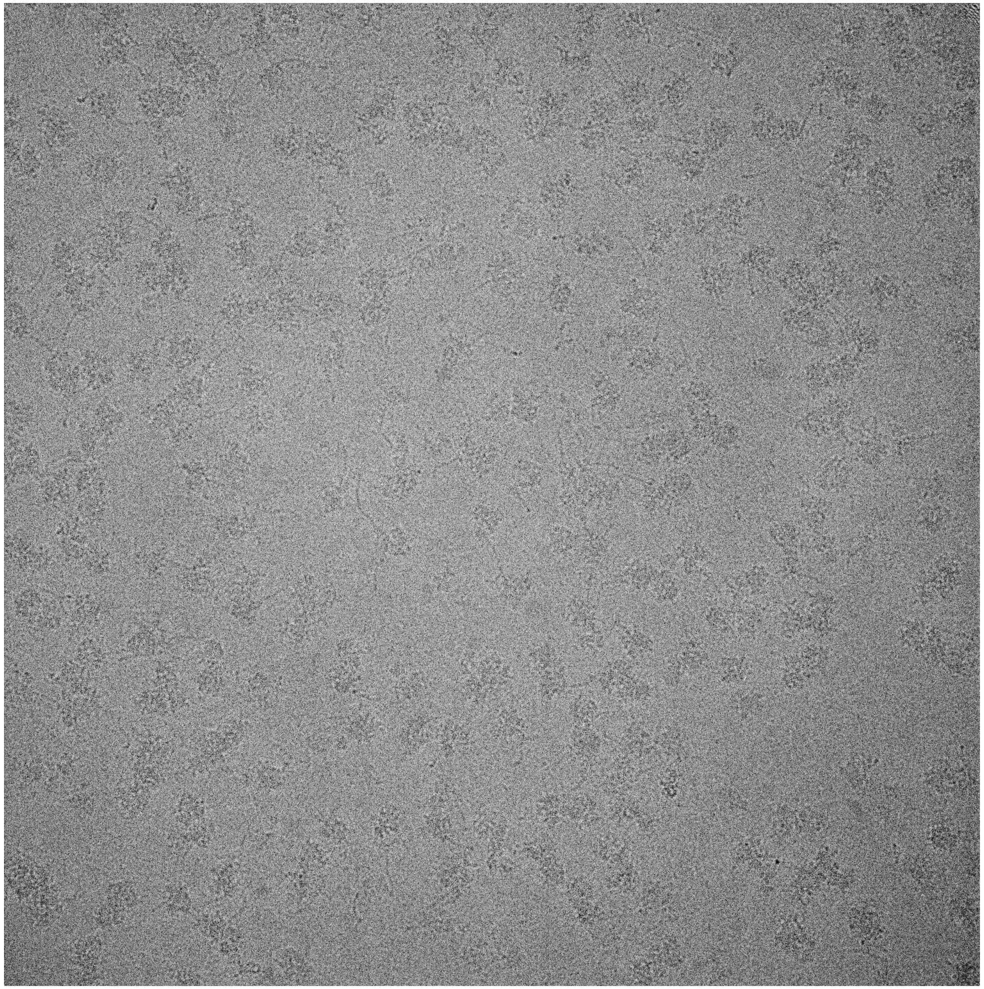


Low SNR

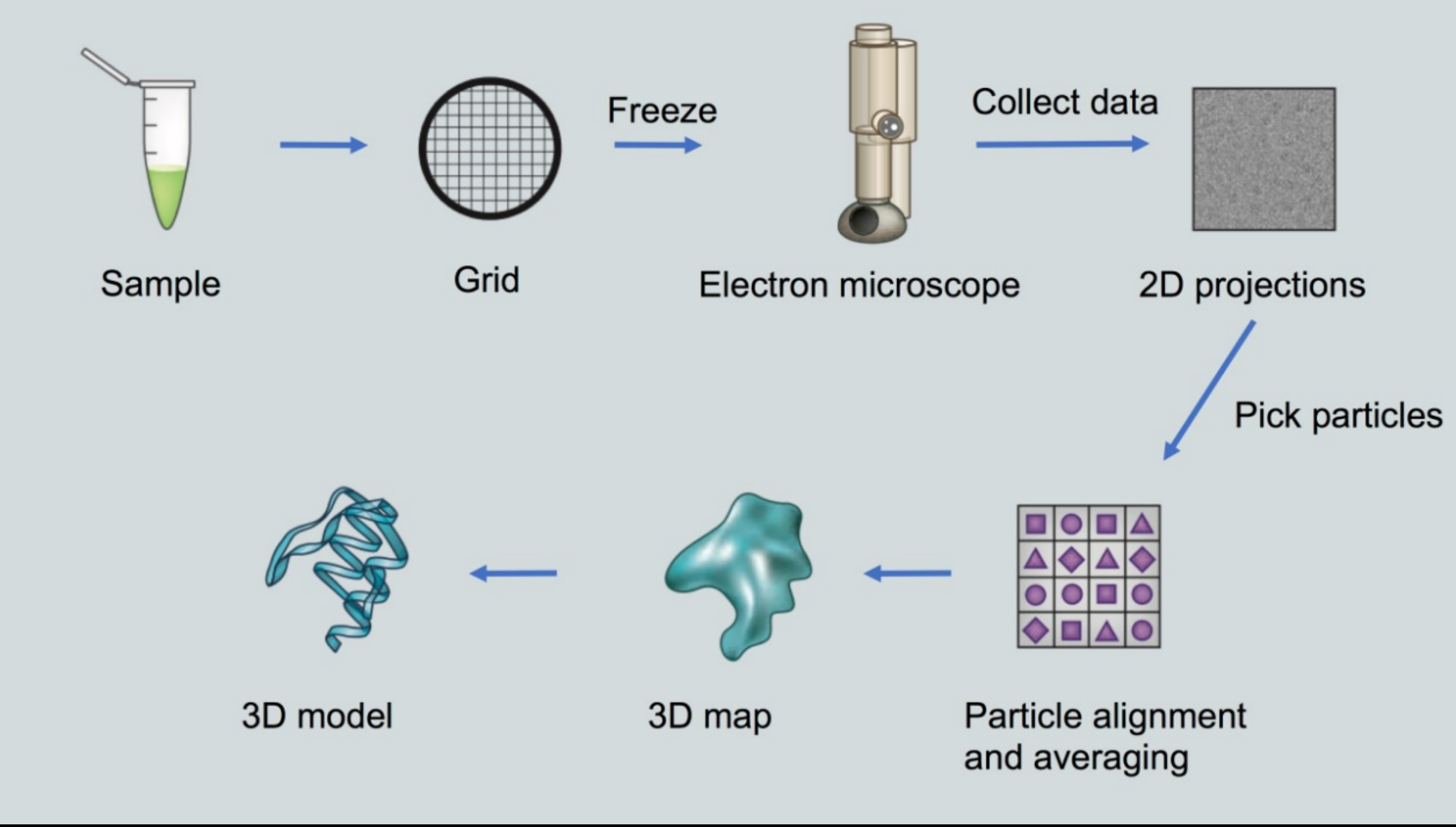
Even when generated using the most sophisticated devices such as Scanning Electron Microscope (SEM) or Transmission Electron Microscopes (TEM), nanoscale protein images are extremely noisy making it one of the hardest challenges for Computer Vision Algorithms.

The challenge in determining the structure is mostly in identifying the particles and its different orientations in the SEM so that they may be integrated to build the 3D structure using epipolar geometric constraints.

The procedure is called Particle Picking

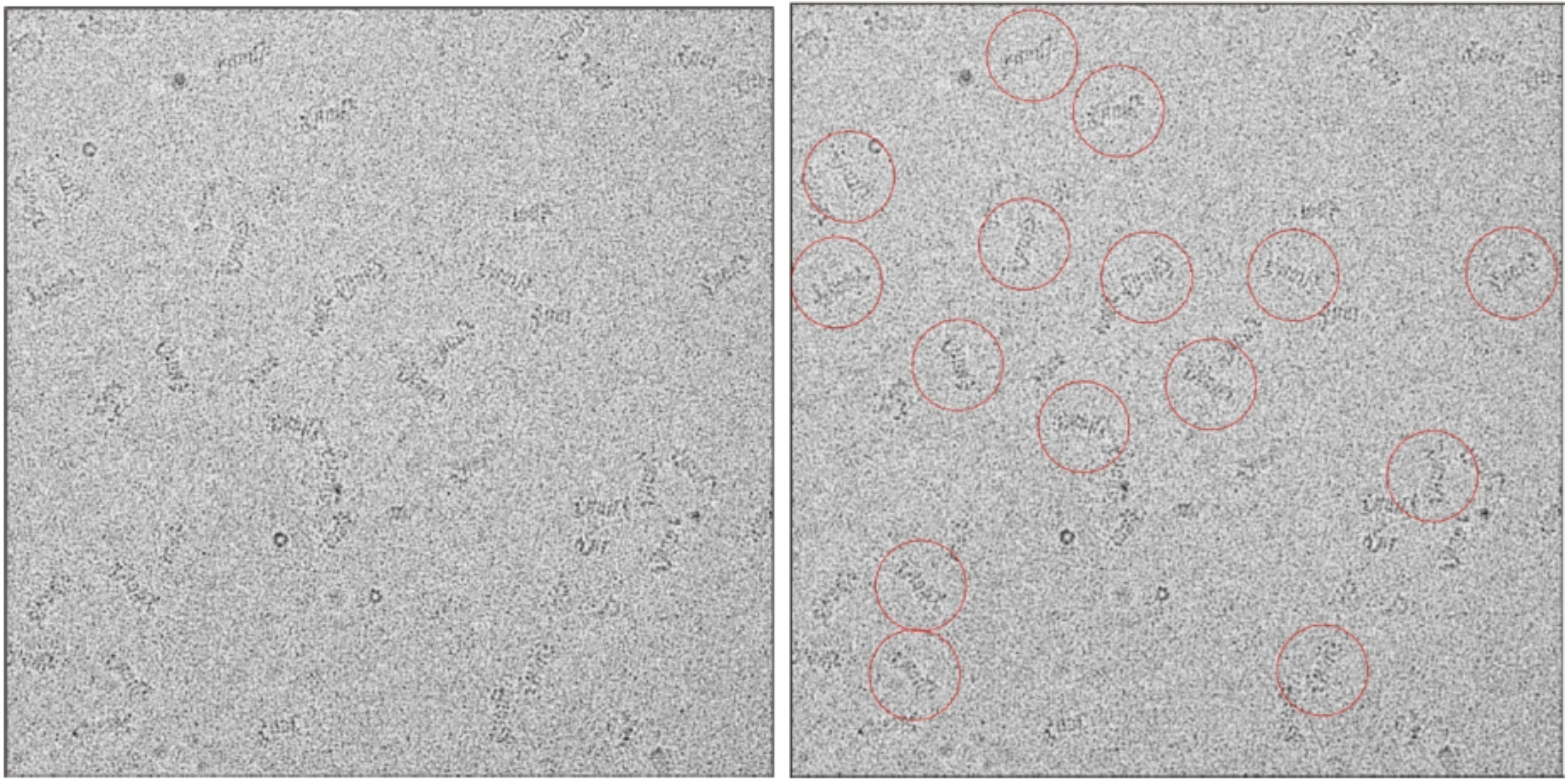


Steps in Protein Identification



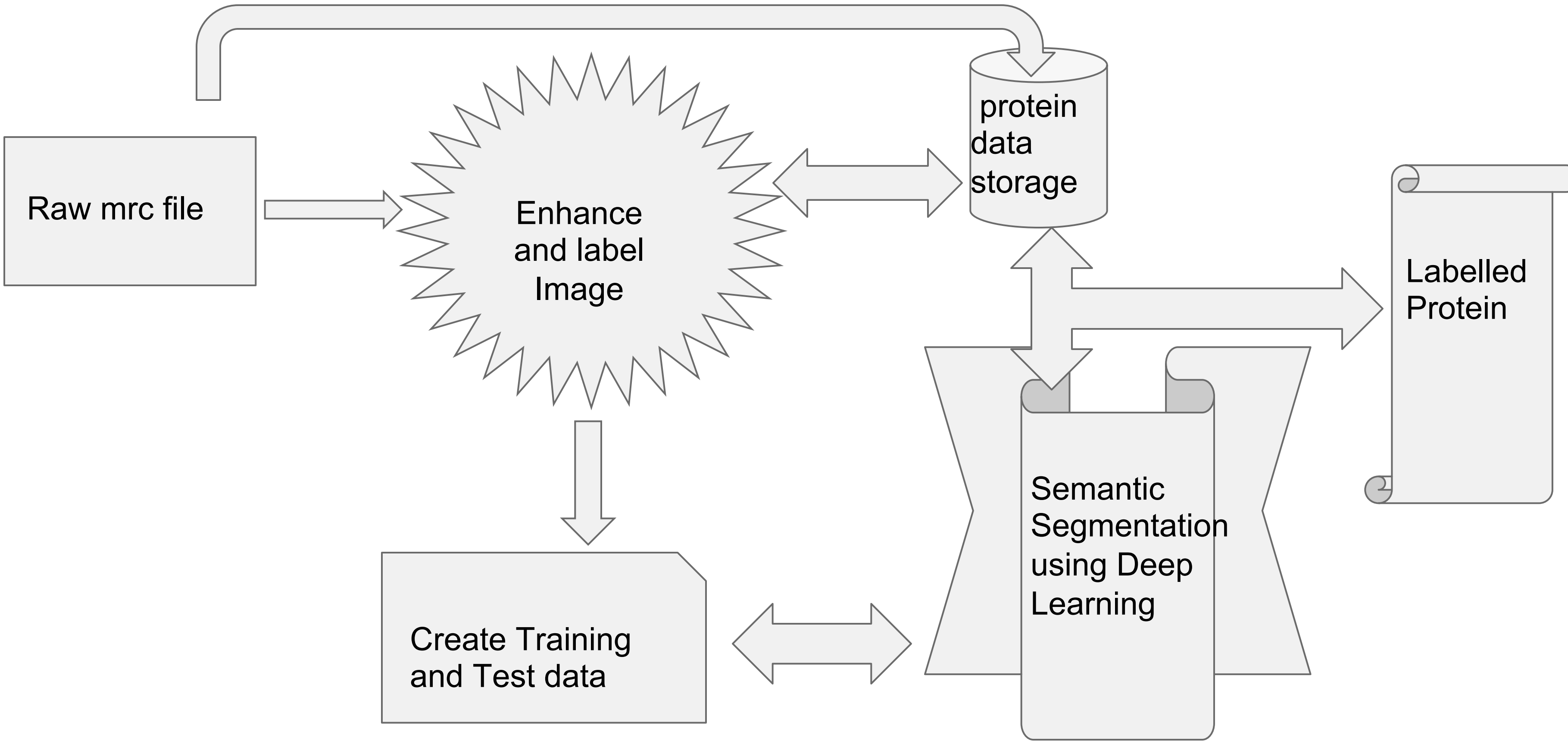
Sajeeth Philip+ 2019

Because of low SNR, Particle picking is often done manually

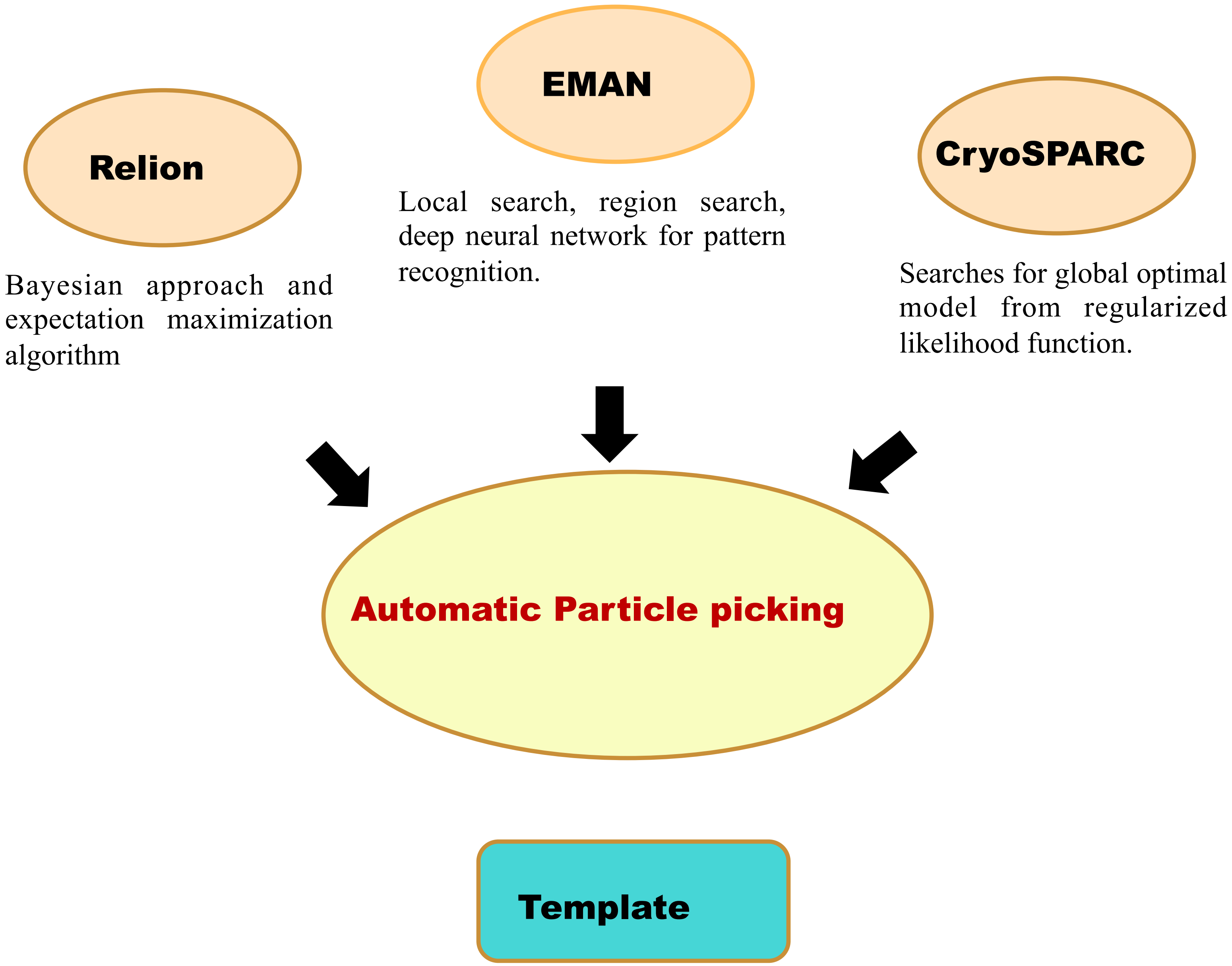


Sajeeth Philip+ 2019

Using Deep Learning for Particle Picking

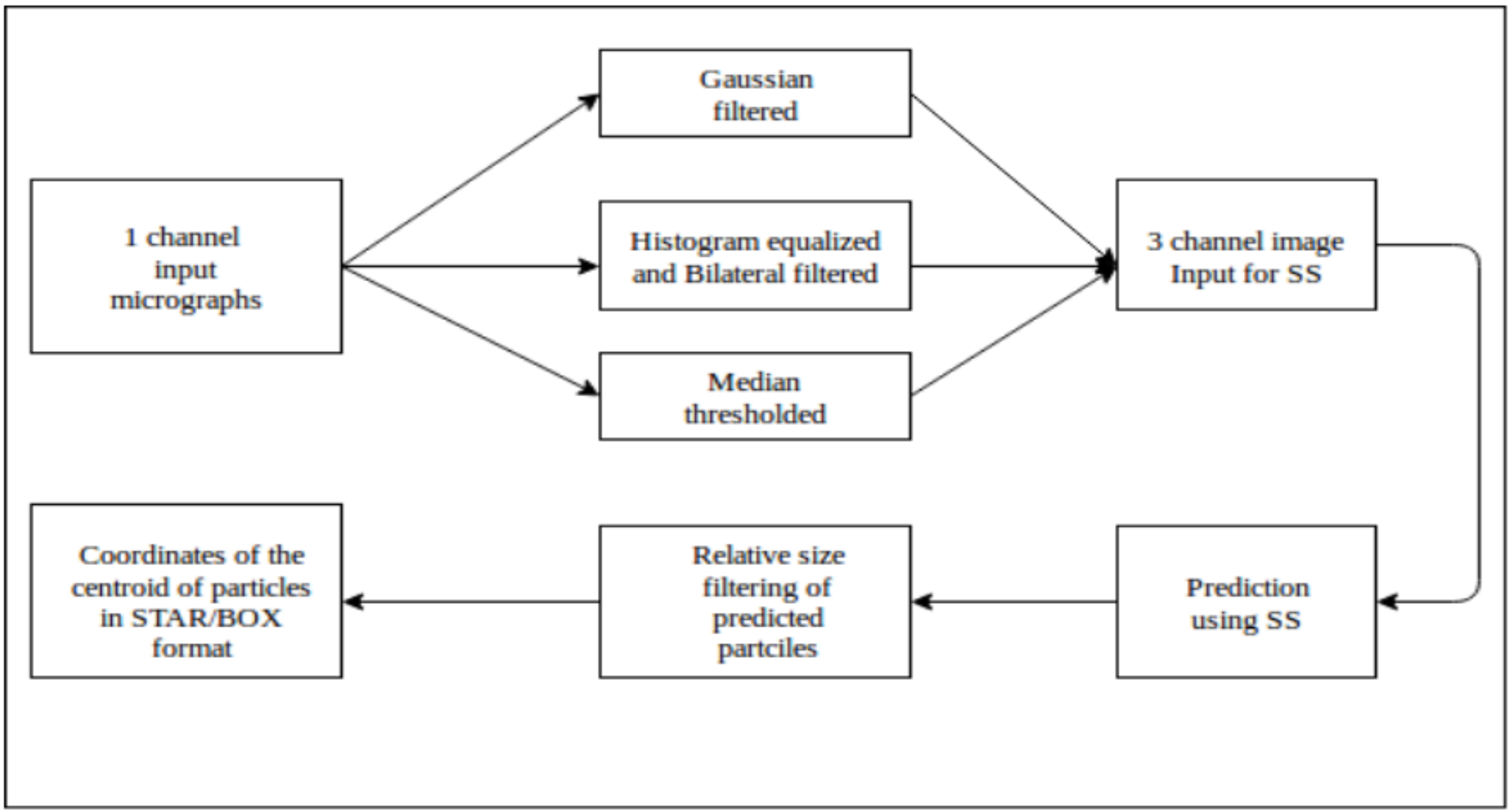


Sajeeth Philip+ 2019


EMAN
Relion
CryoSPARC
Local search, region search, deep neural network for pattern recognition.
Bayesian approach and expectation maximization algorithm
Searches for global optimal model from regularized likelihood function.
Automatic Particle picking
Template

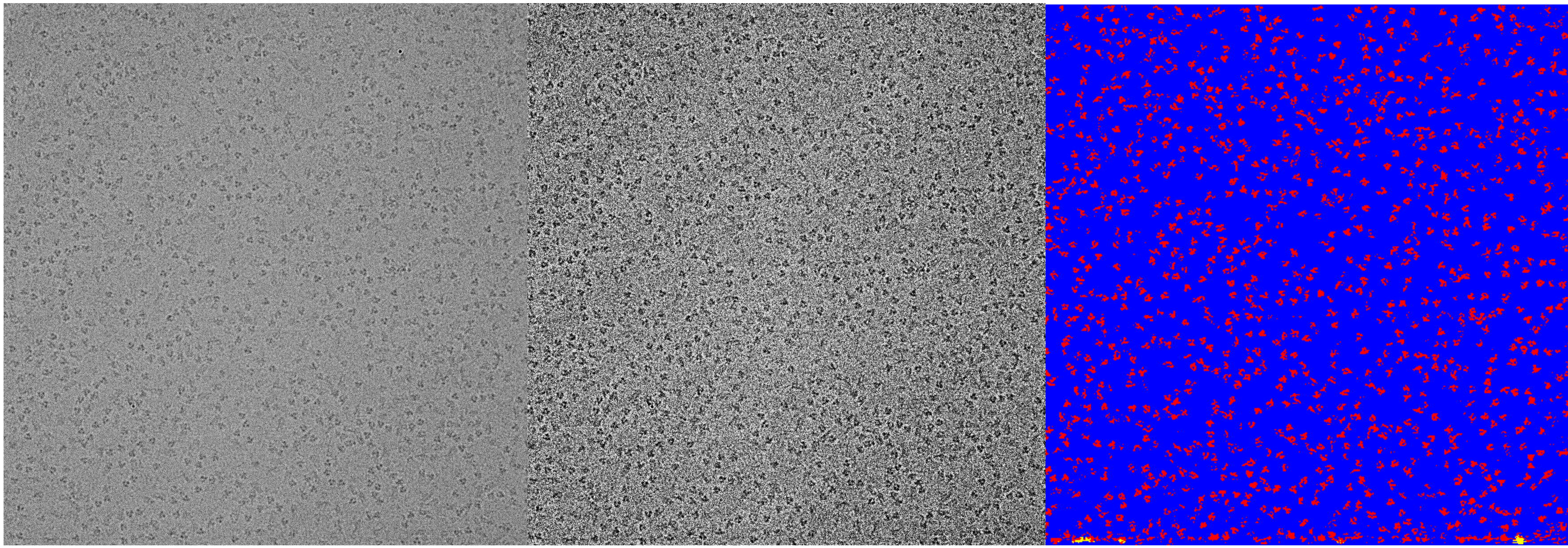
Semantic Segmentation

In house developed tool for Automated particle picking



Schematic representation of particle picking pipeline by Semantic Segmentation

Result



Original

Contrast Enhanced

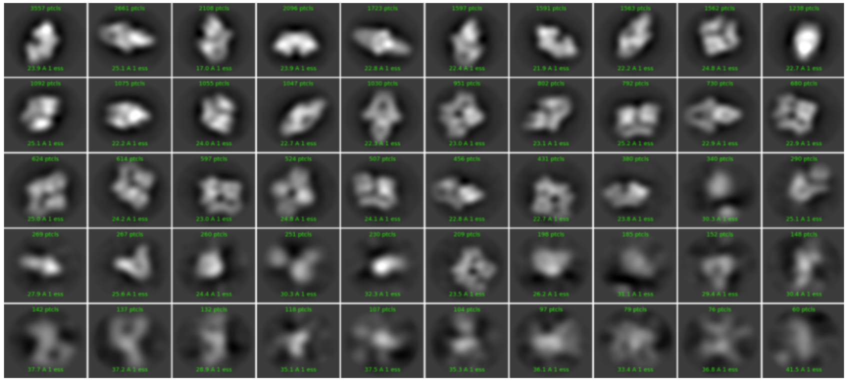
Automated Particle picking

Contrast Enhancement and Particle picking are automated. Given the raw microgram, the deep learning tool will label the particles and provide the mask

Sajeeth Philip+ 2019

Particles picked for Beta galactosidase (EMPIAR-10017) by different particle picking tools

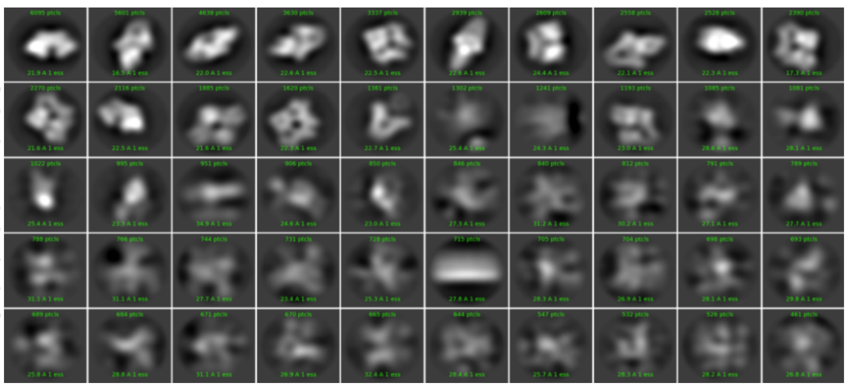
SS



3557	2661	2108	2096	1723	1597	1591	1563	1562	1238
1092	1075	1055	1047	1030	951	802	792	730	680
624	614	597	524	507	456	431	380	340	290
269	267	260	251	230	209	198	185	152	148
142	137	132	118	107	104	97	79	76	60

Total no of particles : 36934
False picked: 2404
Accuracy: 93.5%

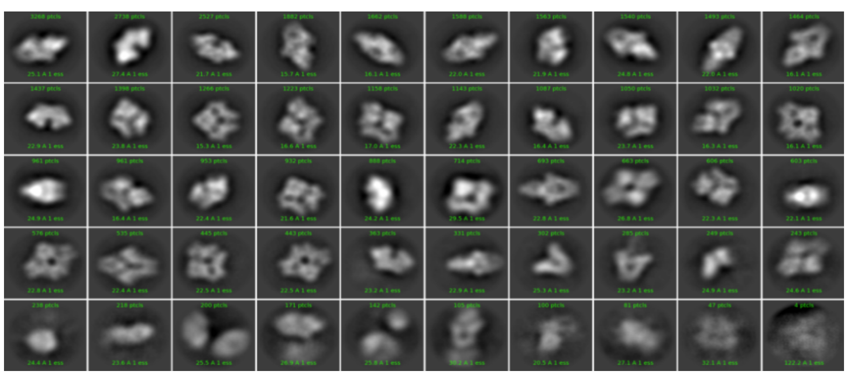
Gautomatch



6095	5601	4638	3630	3337	2939	2609	2558	2528	2390
2270	2116	1885	1620	1381	1302	1241	1193	1085	1081
1022	995	951	906	850	846	840	812	791	789
788	766	744	731	728	715	705	704	698	693
689	684	671	670	665	644	547	532	526	461

Total no of particles: 73662
False picked : 18541
Accuracy: 74.8%

crYOLO

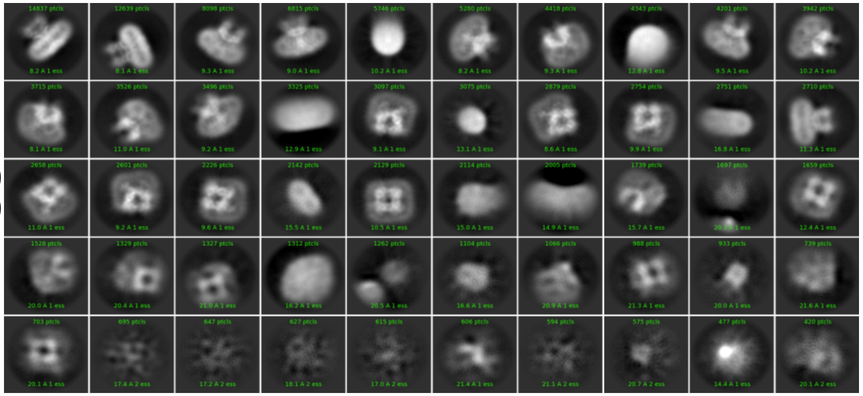


3268	2738	2527	1882	1662	1588	1563	1540	1493	1464
1437	1398	1266	1223	1158	1143	1087	1050	1032	1020
961	961	953	932	888	714	693	663	606	603
576	535	445	443	363	331	302	285	249	243
238	218	200	171	142	105	100	81	47	4

Total no of particles: 44591
False picked: 1787
Accuracy: 96%

Particles picked for (HCN1 EMPIAR-10081) by different particle picking tools

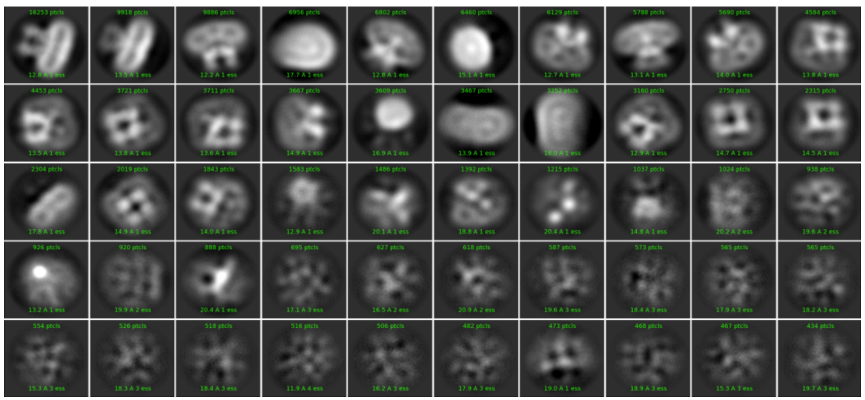
SS



14837	12639	8098	6815	5746	5280	4418	4343	4201	3942
3715	3526	3496	3325	3097	3075	2879	2754	2751	2710
2658	2601	2226	2142	2129	2114	2005	1739	1697	1659
1528	1329	1327	1312	1262	1104	1066	988	933	739
703	695	647	627	615	606	594	575	477	420

Total no of particles: 140164
False picked 33598
Accuracy: 77%

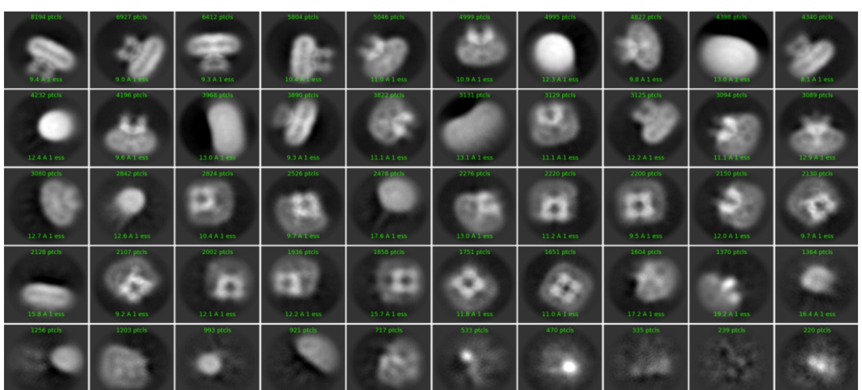
Gautomatch



16253	9918	9886	6956	6802	6460	6129	5788	5690	4584
4453	3721	3711	3667	3609	3467	3252	3160	2750	2315
2304	2019	1843	1583	1486	1392	1215	1037	1024	938
926	920	888	695	627	618	587	573	565	565
554	526	518	516	506	482	473	468	467	434

Total no of particles picked : 139320
False picked: 35652
Accuracy 75%

crYOLO



8194	6927	6412	5804	5046	4999	4995	4827	4398	4340
4232	4196	3968	3890	3822	3131	3129	3125	3094	3089
3080	2842	2824	2526	2478	2276	2220	2200	2150	2130
2128	2107	2002	1936	1858	1751	1651	1604	1370	1364
1256	1203	993	921	717	533	470	335	239	220

Total no of particles: 141002
False picked: 32375
Accuracy: 77%

Thank you!