# Summary

## Hocking A. et al, 2017

- Unsupervised Machine Learning
- Use HST Frontier Fields, then applied to HST CANDELS
- Clearly separate early and late type galaxies
- Use Growing Neural Gas Algorithm (GNG), paper explains details
- Hierarchical Clustering for Clustering Analysis
- Connected Component Labelling, pretty much just labelling which involves looking at nearest neighbour/ next nearest neighbour
- The process is as follows
  - Use training data (which is unclassified) to train the algorithm
  - Check that the classified data actually makes sense
  - Using the gained classification weights, apply to general data
- This method is able to look identify peculiar objects using K-means instead of Hierachical clustering
- This method is compared to Galaxy Zoo (citizen science project)
- Conclusion that machines also classify galaxies like humans do
- Was performed on a desktop
- Can be made much more efficient, so is a very resource light way of classifying images
- Since the algorithm is very general, it can be used for other applications such as classifying spectral data
- **Looks like a promising method, but would be very difficult to implement from scratch**

## Aniyan A. K. & Thorat K. 2017

- Uses Supervised CNN
- Focuses on Radio Galaxies
- They had 3 categories FRI, FRII, Bent Tailed Radio Galaxies
- Sample Size ~200
- Accuracy of 79-95% depending on classification
- They didn't actually 'use' the classifier, as the data was already pre-classified
- All 3 categories had their own samples
- Their model had 5 HL and 3 pools
- They discuss preprocessing at length, discussing sigma clipping and rotating images
- Ultimately they really didn't have enough data to work with so they had to artificially inflate the number of images by rotating some and pretending it's new
- They used Caffe software
- The training was done on a machine with an Intel(R) Xeon(R) CPU, 260 GB memory and four TITAN-Black GPUs with 12GB RAM each.
- The conclusion is that this method is workable in principle and they have provided the source code for this https://github.com/ArunAniyan/RadioGalaxyClassification
- **Key Takeaways: they had insufficient data, they have a relatively robust discussion on data pre-processing, computational hurdles, as well as the mechanics of machine learning. Look at this for reference.**

## Kim E. J. & Brunner R. J. 2016

- Classifier based on traditional CNN
- Mostly a discussion based on a proof-of-concept
- Randomly selected sources that are either stars or galaxies (none which are neither)
- Data taken from CasJobs server
- Data is already labeled due to cross checking with other surveys and catalogues
- Data uses 5 bands u,g,r,i,z
- This paper explains in some detail how CNN's work
- They start with a 5x44x44 tensor
- Uses 8 ConHL, 3 pool, 2 Fully connected HL, ends with Softmax
- Training their network took 40 hours on a Tesla K40 GPU
- Code found https://github.com/EdwardJKim/dl4astro

- Since they had such a large NN, they were bound to overfit, they described methods in which overfitting can be avoided
- Rotation, reflection, translation(random crop), gaussian noise were used to artificially increase the size of the data set
- Dropout is used (however dropout parameter is not shown)
- Uses Bayesian combination to predict
- They discuss a lot about the detailed statistics, worth a detailed look when actually implementing
- CNN are especially good at image recognition
- Feature extraction is not required and hence feature engineering is not required to make the network work properly
- If this needs to be expanded to more training data, multi-GPU setup is required
- This method doesn't really work for spectroscopic data since it is too faint
- **Worth looking at if CNN is being used. However, this might be a dead end as further improvements are probably from hardware(?)**

# Kgalifa N. et al, 2017

- They used data from various sky surveys to get 13000 images for classification
- Their classifications were Elliptical, Spiral, Irregular
- They have 1 ConHL (96 filters), 27 fully connected HL, 1 pool, Ends with softmax
- ReLU activation
- Intel Xeon E5-2620 processor (2 GHz) and 96 GB Ram
- Trained over 1356 images, accuracy 97.272%
- **Probably not enough training data. Didn't really provide details on how exactly the images were processed and such. Not very useful.**

# Schutter A. & Shamir L. 2015

- Unsupervised Algorithm
- EFIGI Catalogue Data
- Data converted to TIFF images
- Wndchrm image recognition method (using 4027 numerical content descriptors)
- The morphological categories are vaguely consistent with the Hubble classifications
- Source code available in this paper
- **This paper seems to describe an inferior method compared to Hocking 2017, also it doesn't give detailed description into the mechanics of the algorithm**

# Lintott C. J. et al, 2008

- Explains how Galaxy Zoo was done
- Basically citizen science project, the precise mechanisms behind the classification not really important for our purposes
- **Most important thing is that these classifications are freely available and can be acquired online at any time**

# Banerji M. et al, 2010

- A pretty old paper on comparing machine learning with galaxy zoo
- They got 90% accuracy
- Doesn't clearly explain which type of NN they're using
- Using colours and profile-fitting parameters as inputs to the neural network, 87% of early type classifications, 86% of spiral classifications and 95% of point source/artifact classi- fications agree with those obtained by the human eye
- A combination of the profile fitting and adaptive weighted fitting parameters results in better than 90% agree- ment between classifications by humans and those by the neural network for all three morphological classes
- Their best sample reached 95% accuracy
- **Not very useful in terms of the technical aspects, but still gives a way of using Galaxy Zoo data to train the NN.**

# Huertas-Company M. et al, SSDS DR7

- Using SVM to separate galaxies into 4 categories (E, S0, Sab, Scd)
- SSDS DR7 Sample

- First separate E/S, then further split into 2 categories each
- From reading, this doesn't seem like simple SVM, but more like kernel SVM (non-linear)
- Better than galaxy zoo according to them since they split into 4 instead of 2
- http://gepicom04.obspm.fr/sdss_morphology/Morphology_2010.html classifier here
- 12% scatter, can be considered as accuracy
- 0.4% objects considered uncertain
- **A kernel SVM method, which is relatively older compared to CNN methods. However, this may present an alternative method of splitting 2 categories. However, I think this might have reached the maximum potential. Probably a dead end.**

# Huertas-Company M. et al, SVM on seeing images

- Using SVM to classify galaxies based on morphology
- Sample in Ks and i bands
- Source code here http://www.lesia.obspm.fr/~huertas/galsvm.html
- The code first separates the image into parameters, then from the 12-parameter space it classifies the data using SVM
- Accuracy higher than in SDSS sample (80%) and lower in WIRCam sample (60%)
- High redshift doesn't seem to be relevant.
- This classifies into 2: early and late types
- **An SVM method with defined parameters, the minutiae of this is is a good description of how to extract parameters from an n-dimentional space of images. However, this approach is considered outdated compared to CNN which doesn't need dimensional reduction**

# Conselice C. J., 2006

- 22121 galaxies in sample
- Uses PCA to identify the main features of data set, found that the main eigenvectors are: scale, star-forming regions, interactions/merger
- Sample nearby rather than distant sample
- They made a new method of classifying galaxies, based on the 3 eigenvectors
- They comment that the Hubble classification represent ideal galaxies, actual galaxies are much more messy
- **This gives a good idea how how to use PCA to reduce the dimensionality of the data. Depending on which method is used, this is a good source.**

# Calleja J. & Fuentes O., 2004

- Three stages: image analysis, data compression, machine learning
- Image analysis basically picks out galaxies and centres them to make them machine-learnable
- Data compression is a typical PCA method
- The machine learning is a comparison of 4 methods: Naiver Bayes Classifier, C4.5, Random Forest Predictor, Ensemble Method. Of these, they found RFP to be best
- Results for 3 class: 91.64%, 5 class: 54.72%, 7 class: 48,62%
- **Very early use of machine learning to produce classification, results not that much worse than modern methods**

# Abraham R. G. & Merrifield M. R., 2000

- The purpose was to make a morphology classification based on a reduced dimensional space: Hubble Space
- Hubble space is a 2D space of which, one dimension is early/type, one dimension is how barred it is
- Not really anything to do with machine learning, this represents a manual method of classification
- It uses various manual image classification algorithms to determine the Central concentration C and bar strength f-bar and plots all the galaxies in a plot
- The classification largely confirms the Hubble tuning fork, but that maybe due to the fundamental bias in choosing the two parameters
- **Not too useful, too old to represent any modern method of machine learning**