Abstract

Type 2 Active Galactic Nuclei (AGNs) are a key component for understanding black hole growth because dust obscuration can hide signatures that are easier to detect in unobscured AGNs, biasing many large-sample evolution studies. This project develops an observation-driven method to track how Type 2 AGN populations change over cosmic time using only directly measured quantities. We construct a clean sample of local Type 2 AGNs ($z < 0.2$) using optical spectroscopy from the Sloan Digital Sky Survey (SDSS DR16), cross-matched with mid-infrared photometry from the Wide-field Infrared Survey Explorer (WISE), and supplemented with public catalog information. From these data we compute emission-line ratios, [O III] luminosity, and WISE color indices, then apply unsupervised machine learning—Principal Component Analysis (PCA) and UMAP for dimensionality reduction, followed by K-Means clustering—to identify distinct groupings ("stages") in observable parameter space. We compare cluster abundances and cluster-averaged properties across redshift bins and validate interpretations with BPT diagnostic diagrams to separate AGN activity from star-formation contamination. Our approach reveals multiple Type 2 AGN clusters with different power and dust/torus signatures and shows measurable shifts in cluster composition with redshift, indicating evolutionary trends even within the nearby universe. Scientifically, this provides astronomers a utilizable framework for discovering population evolution in survey-scale datasets and prioritizing targets for follow-up spectroscopy. More broadly, it demonstrates how machine learning can convert large catalogs into interpretable insight, supporting the public understanding of black hole growth and cosmic history.

Rationale *

Active galactic nuclei (AGNs) are powered by accretion onto supermassive black holes and can outshine their host galaxies. This eventually affects gas, star formation, and the long-term evolution of galaxies that researchers look at. In the standard "unified" picture, observed AGN types are fundamentally similar, but there are still many questions about how their observable properties change over cosmic time.

Large sky surveys like the Sloan Digital Sky Survey (SSS DR16) and the Wide-field Infrared Survey Explorer (WISE) give us huge samples of galaxies and AGNs with optical spectra and mid-infrared photometry. These surveys make it possible for us to study thousands of Type 2 AGNs with their characteristics too.

In the realm of science, this project is important because it uses only directly observed quantities like line fluxes, colors, and redshift as well as unsupervised machine learning to search for possible "stages" in Type 2 AGN populations across lookback time. Instead of attempting complex radiation-transfer modeling, we're going to treat redshift as a time axis and delve deeper into how clusters in the observable parameter space shift as the universe evolves. This will provide insight into how AGN power varies over billions of years and how it may connect to galaxy evolution.

Societally, this project is useful to demonstrate how modern astronomy intertwines with data science. We'll be using real survey data, coding, and machine learning on a large dataset, directly showing how these tools are transferable to other big-data problems in science. The interactive visualization that we plan to build can also

help make abstract concepts like cosmic time and black hole growth more accessible to students and the public.

Research Question/Hypothesis(es)/Engineering Goal(s) *

Research Question:

How do the observable properties of local Type 2 AGNs (like luminosity, emission-line ratios, and WISE mid-infrared colors) cluster in parameter space, and how do these clusters change as a function of redshift for z < 0.2?

Hypotheses:

Type 2 AGNs occupy distinct properties that correspond to different levels of AGN power and dust/torus activity. The relative abundance and average properties of these clusters change with redshift (which reflects evolution in AGN activity over cosmic time). Clusters identified by unsupervised machine learning will roughly align with known diagnostic boundaries in standard emission-line diagrams and also could reveal additional trends.

Engineering Goals:

Build a reproducible Python pipeline that can select a clean sample of Type 2 AGN data from astronomical catalogs, cross-match these objects with WISE All-Sky data, extract key features, and apply dimensionality reduction. We also want to develop an interactive notebook that lets users explore AGN clusters by color-coding points by physical properties.

Expected Outcomes:

Identification of one or more clusters of Type 2 AGNs in observable parameter space. UMAP/PCA plots and BPT diagrams annotated with cluster identities. Quantitative descriptions of how cluster averages change ind different redshift bins.

Materials *

Hardware: personal laptop (16 GB RAM, 50 GB free disk space)

Software: Operating system, Python (3.10+) with libraries (numpy, pandas, matplotlib, astropy, scikit-learn, umap-learn, astroquery, jupyter, streamlit), Git, and GitHub.

Astronomical Data Sources: SDSS DR16/MPA-JHU catalogs, WISE All-Sky catalogs, Veron-Cetty and Veron AGN catalog, and 4XMM-DR11 X-ray catalog.

Procedures *

We will first use the SDSS DR16 and the Veron-Cetty catalog to select Type 2 AGNs with reliable emission-line measurements and z < 0.2. After that, we'll export a combined table containing redshift, key line fluxes, and basic photometry.

Then, we're going to cross-match the AGN sample with the WISE All-Sky/AllWISE catalog using coordinates to obtain W1, W2, and W3 magnitudes.

For our third step, we're going to apply simple quality cuts and compute emission-line ratios, luminosity, WISE colors, and lookback time from redshifts using standard cosmology tools.

Then, we'll prepare for machine learning by assembling a feature matrix (line ratios, luminosities, colors, and an optional X-ray). With that, we'll standardize all features in Python using scikit-learn.

After that, we'll use PCA and/or UMAP to create 2D embeddings of the AGN sample and then apply K-Means to group AGNS into clusters in feature space. We'll store a cluster label for each object.

For the sixth step, we'll divide the sample into several redshift/lookback-time bins and, for each bin, calculate basic statistics of each cluster's properties.

Finally, we'll create PCA/UMAP scatter plots, BPT diagrams, and redshift-binned trends in Jupyter Notebooks. We'll also package the full workflow into a documented Python pipeline.

Risk Assessment *

There are no physical, chemical, biological, or electrical hazards because all the work is done on a computer using public astronomical data. There are minimal data-security risks (using public survey data, no personal or human subject data).

Some safety precautions that we could take are: taking regular breaks from the screen to reduce eye strain, maintaining good posture, and using version control so there is no data loss.

No Institutional Review Board (IRB) or animal care approvals are required because there are no human or vertebrate animal subjects.

Data Analysis *

The first step in our data analysis was to perform data wrangling and to remove the unnecessary data like the Right Ascension (RA), Declination (Dec), LII (Galactic Longitude), and the BII (Galactic Latitude), and the NaN values with the averages to ensure that the data is streamlined for the machine learning model to be applied.

The second step was then to compute basic statistics (mean, median, standard deviation) for each of the major features ([O III] luminosity, line ratios, WISE colors).

The third step in the process was to create a Principal Component Analysis (PCA), which is a machine learning technique that transforms a dataset's features into a new set of uncorrelated variables called principal components.

In the fourth step, we had to use a K-Means clustering algorithm to create a prediction algorithm to predict the evolution of characteristics based on the physical properties provided in the dataset. This is set up by the redshift bins, which are vital in the creation of the cluster and groups for the K-means clustering machine learning algorithm.

In the fifth step, we plotted the cluster of the type 2 AGNs and the physical properties like the [O III] luminosity, line ratios, and WISE colors, which show the specific. trends. Finally, we created BPT diagrams to distinguish between the start formation and AGN activity to ensure that there is high accuracy and there isn't any star formation data affecting our results in AGN evolution.

Bibliography *

Asmus, D., Gandhi, P., Hönig, S. F., Smette, A., & Tristram, K. R. W. (2020). Local AGN Survey (LASr) I: Galaxy sample, infrared colour, and volume statistics. Monthly Notices of the Royal Astronomical Society, 494(2), 1784–1802. https://doi.org/10.1093/mnras/staa760

Astropy Project. (n.d.). Astroquery documentation. https://astroquery.readthedocs.io/en/latest/

Ballantyne, D. R. (2006). Does the AGN unified model evolve with redshift? Using the X-ray background to predict the fraction of obscured AGN at all redshifts. The Astrophysical Journal, 653, 1070–1081. https://doi.org/10.1086/508600

Hasinger, G. (2008). Absorption properties and evolution of active galactic nuclei (AGN). arXiv:0808.0260. [*https://arxiv.org/abs/0808.0260*](https://arxiv.org/abs/0808.0260)

McInnes, L., Healy, J., & Melville, J. (n.d.). UMAP: Uniform Manifold Approximation and Projection documentation. [*https://umap-learn.readthedocs.io/en/latest/*](https://umap-learn.readthedocs.io/en/latest/)

NASA HEASARC. (n.d.). Véron-Cetty and Véron active galactic nuclei catalog. [*https://heasarc.gsfc.nasa.gov/w3browse/all/veroncat.html*](https://heasarc.gsfc.nasa.gov/w3browse/all/veroncat.html)

NASA/IPAC Infrared Science Archive. (n.d.). WISE All-Sky data release documentation. [*https://irsa.ipac.caltech.edu/data/WISE/docs/release/All-Sky/*](https://irsa.ipac.caltech.edu/data/WISE/docs/release/All-Sky/)

Sloan Digital Sky Survey. (n.d.). DR16 value-added catalogs. [*https://www.sdss4.org/dr16/data_access/value-added-catalogs/*](https://www.sdss4.org/dr16/data_access/value-added-catalogs/)

Past Abstract:

The exploration of Active Galactic Nuclei (AGNs), the highly energetic centers of galaxies hosting supermassive black holes, has been a cornerstone of modern astrophysics. Milestones such as the discovery of quasars in the 1960s and the identification of the first AGN in a Seyfert galaxy have established AGNs as key tools for studying galaxy evolution and black hole growth. Despite decades of research, the mechanisms linking AGN properties to cosmic distances remain untouched. This project investigates the evolution of Type 2 AGNs across the redshift spectrum, where redshift serves as a critical measure of distance and cosmic history. We utilized data from NASA's HEASARC, specifically the XMM-CTY2AGN database, comprising multi-wavelength spectroscopic observations of 255 Type 2 AGNs. Employing advanced machine learning models like Gaussian processes and x-y kernel methods, we analyzed non-linear relationships between redshift and AGN properties. Our findings reveal that Type 2 AGNs at higher redshifts exhibit significantly greater luminosities and elevated U-V color indices. These results suggest an evolutionary trend in AGN characteristics, shedding light on their role in shaping the early universe and advancing our understanding of supermassive black hole activity.