# Expediting the discovery process:

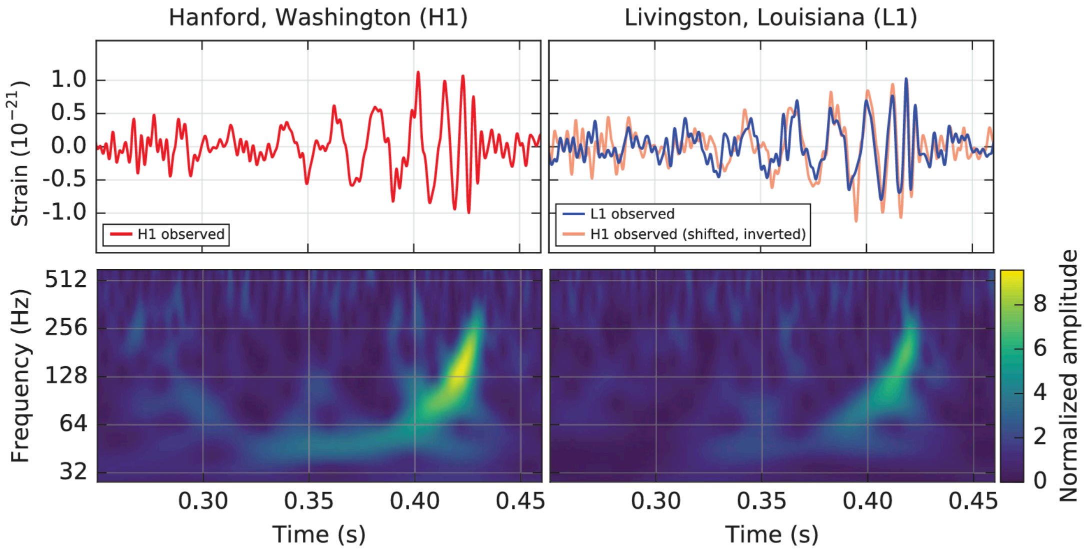
## Application of unsupervised machine learning algorithms to multi-wavelength astronomical datasets

Dalya Baron
Carnegie Observatories -> Stanford University
dalyabaron@gmail.com

MACHINE LEARNING
FOR ASTROPHYSICS
2ND EDITION CATANIA, 8-12 JULY, 2024

# New observations lead to new discoveries

## Gravitational waves & merging black holes



Abbott et al. (2016)

## Stellar streams around the MW

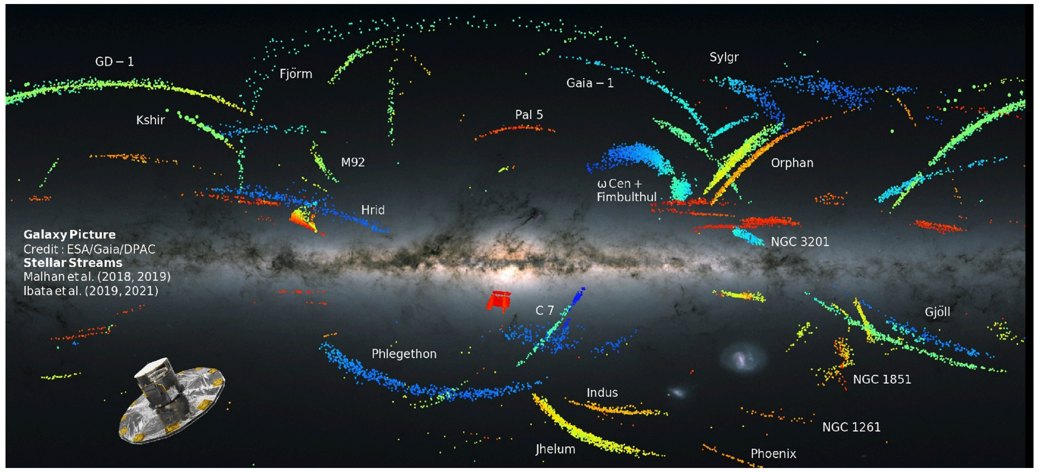


Image credit: K. Malhan

## Extrasolar planets

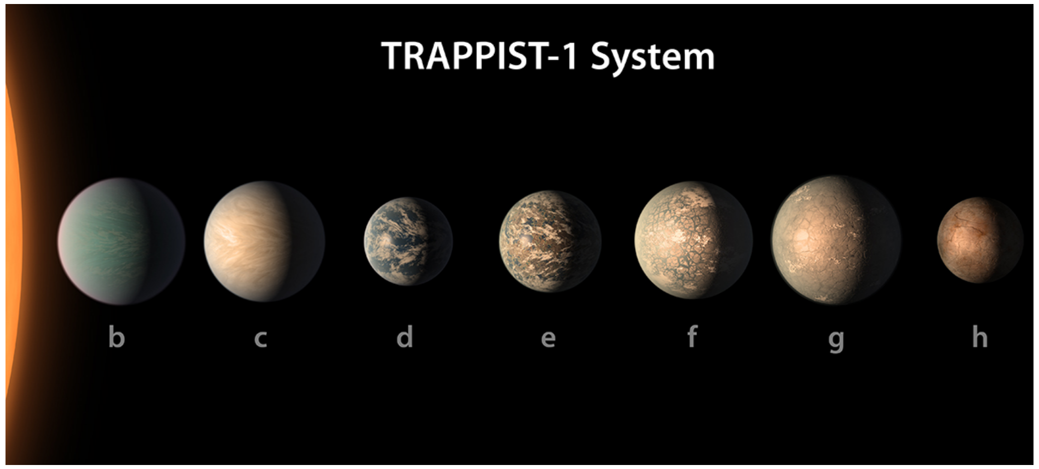
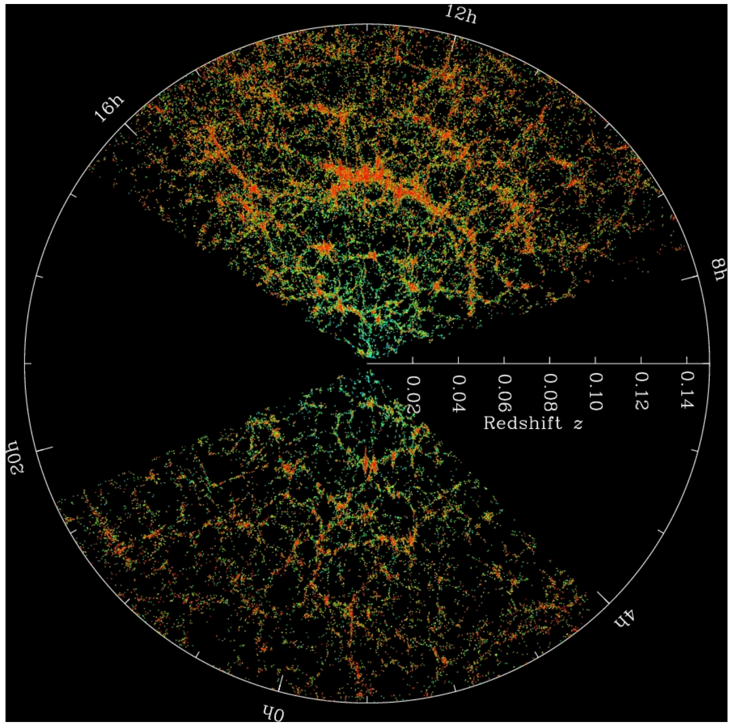


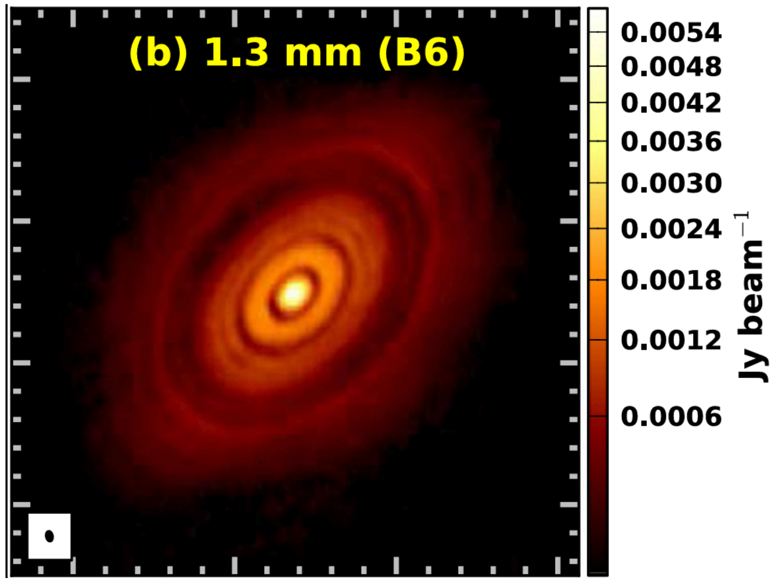
figure from Gillon et al. (2017)

## Large scale structure



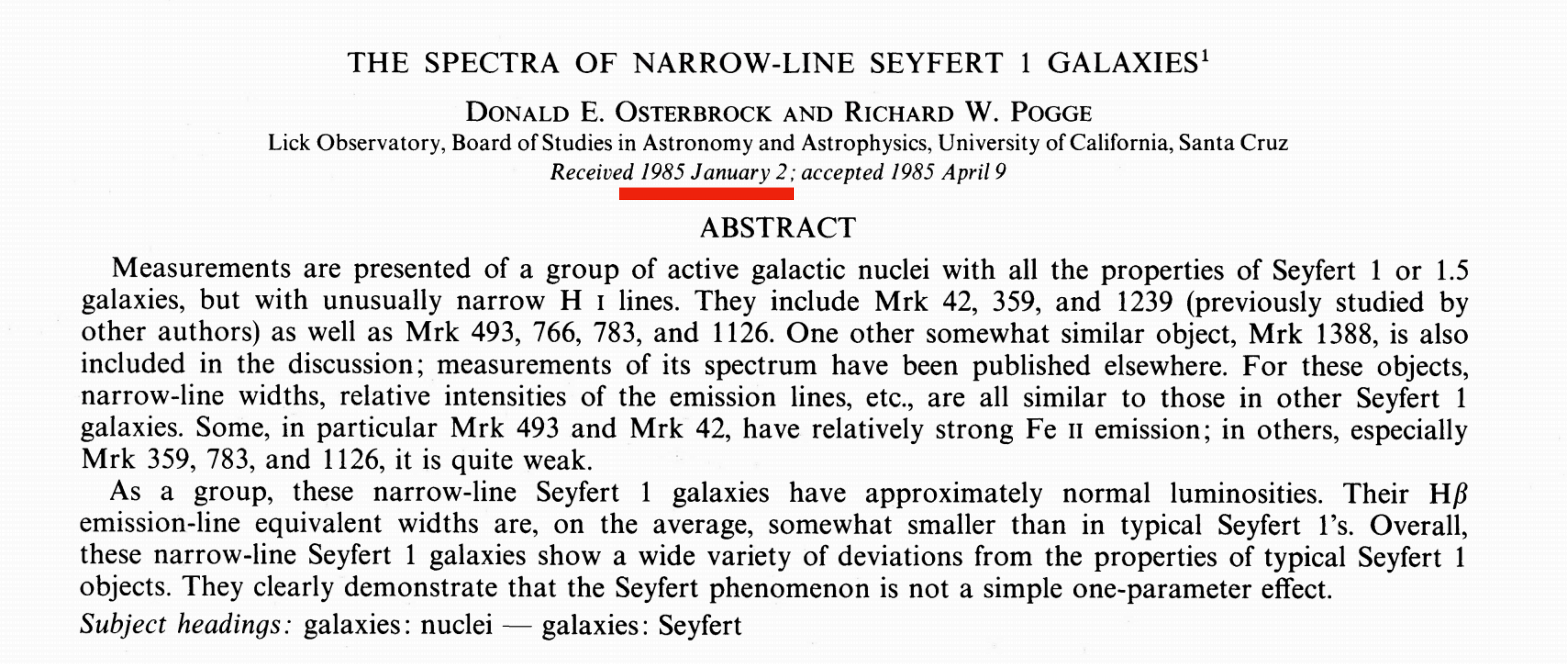
LCRS; 2dFGRS; SDSS

## Proto-planetary discs



Brogan et al. (2015)

# The astronomical data revolution



THE SPECTRA OF NARROW-LINE SEYFERT 1 GALAXIES[1]

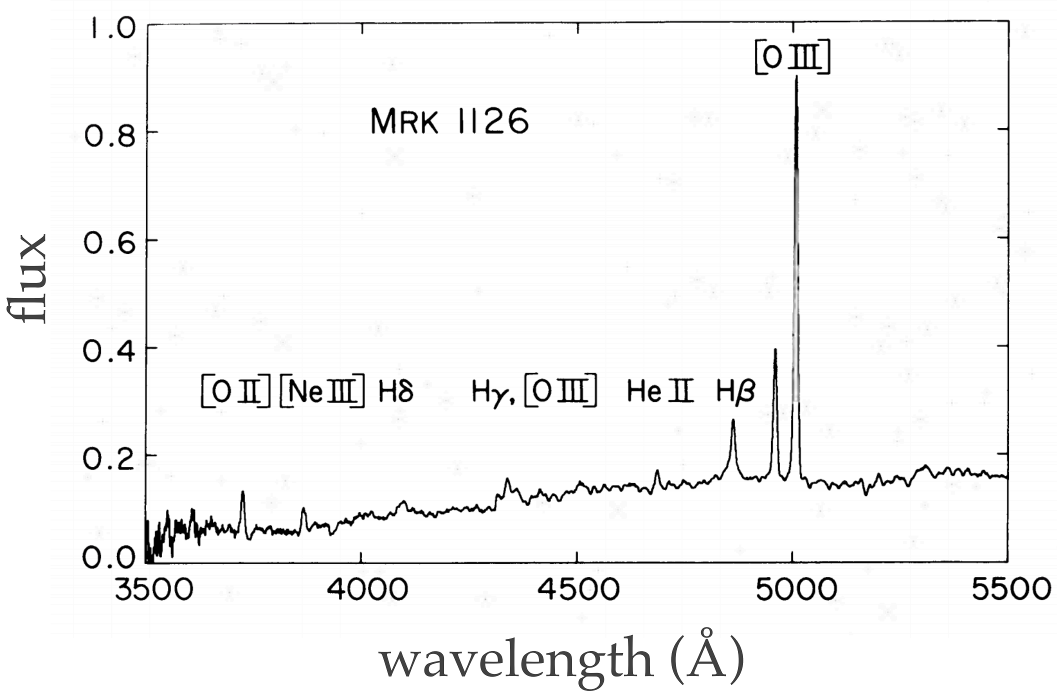
DONALD E. OSTERBROCK AND RICHARD W. POGGE
Lick Observatory, Board of Studies in Astronomy and Astrophysics, University of California, Santa Cruz
*Received 1985 January 2; accepted 1985 April 9*

ABSTRACT

Measurements are presented of a group of active galactic nuclei with all the properties of Seyfert 1 or 1.5 galaxies, but with unusually narrow H I lines. They include Mrk 42, 359, and 1239 (previously studied by other authors) as well as Mrk 493, 766, 783, and 1126. One other somewhat similar object, Mrk 1388, is also included in the discussion; measurements of its spectrum have been published elsewhere. For these objects, narrow-line widths, relative intensities of the emission lines, etc., are all similar to those in other Seyfert 1 galaxies. Some, in particular Mrk 493 and Mrk 42, have relatively strong Fe II emission; in others, especially Mrk 359, 783, and 1126, it is quite weak.

As a group, these narrow-line Seyfert 1 galaxies have approximately normal luminosities. Their H$\beta$ emission-line equivalent widths are, on the average, somewhat smaller than in typical Seyfert 1's. Overall, these narrow-line Seyfert 1 galaxies show a wide variety of deviations from the properties of typical Seyfert 1 objects. They clearly demonstrate that the Seyfert phenomenon is not a simple one-parameter effect.

*Subject headings:* galaxies: nuclei — galaxies: Seyfert



# The astronomical data revolution

(I) Survey mode: high-quality datasets made publicly-available.

Don't need a telescope to do science

# The astronomical data revolution

(I) Survey mode: high-quality datasets made publicly-available.

(II) Exponential increase in the number of observed sources.

- Past or current surveys: SDSS, Pan-STARRS, ZTF, DESI, Gaia.
- Near future: Rubin, Roman, Euclid, SDSS-V, SKA.

Enables a statistical view of the population. Allows to identify small groups & outliers.

# The astronomical data revolution

(I) Survey mode: high-quality datasets made publicly-available.

(II) Exponential increase in the number of observed sources.

- Past or current surveys: SDSS, Pan-STARRS, ZTF, DESI, Gaia.
- Near future: Rubin, Roman, Euclid, SDSS-V, SKA.

(III) Increased information content of a given observed source

- Higher spectral; angular; or temporal resolution.

Enables a more fine-grained view.
Leads to discoveries of phenomena
on smaller/shorter time scales.

# The astronomical data revolution

(I) Survey mode: high-quality datasets made publicly-available.

(II) Exponential increase in the number of observed sources.

- Past or current surveys: SDSS, Pan-STARRS, ZTF, DESI, Gaia.
- Near future: Rubin, Roman, Euclid, SDSS-V, SKA.

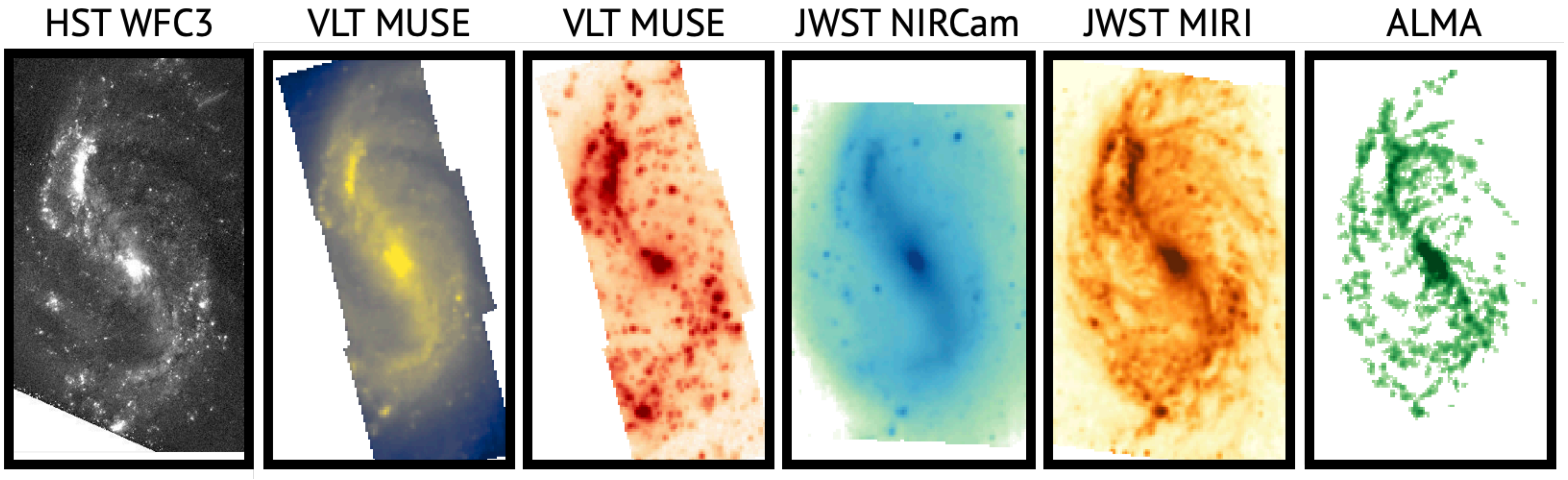
(III) Increased information content of a given observed source

- Higher spectral; angular; or temporal resolution.

(IV) Multi-wavelength coverage by multiple surveys covering radio to gamma-ray wavelengths.

**Dramatic increase of the discovery space**

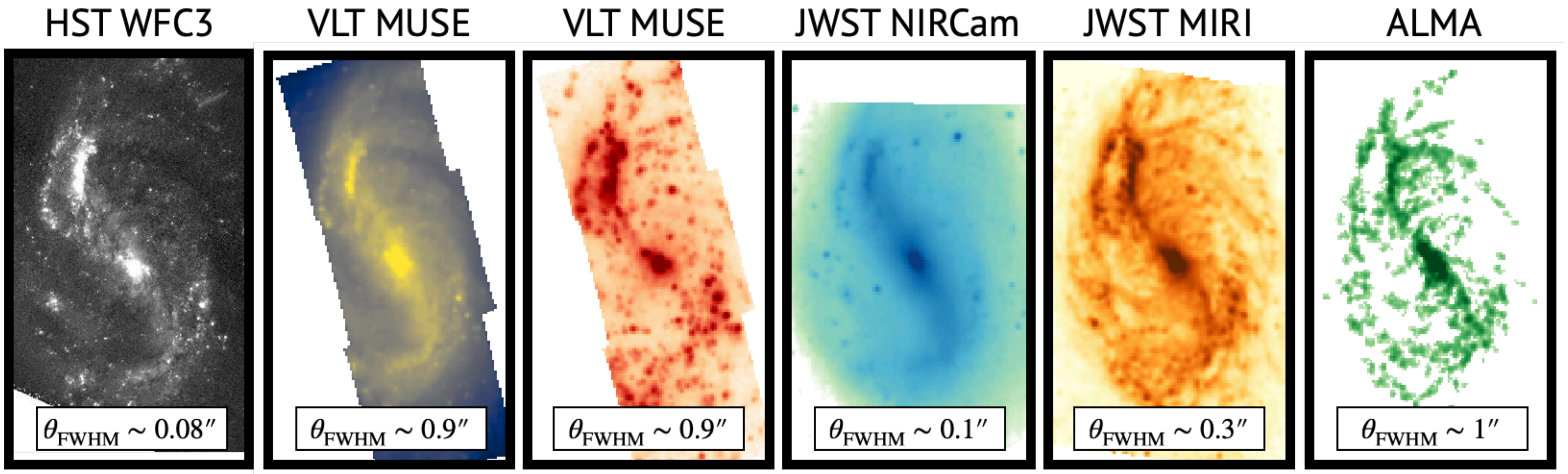
# Modern astronomical surveys produce complex and diverse datasets using observations across the electromagnetic spectrum



Observations of NGC 7496 by the **PHANGS** (The Physics at High Angular resolution in Nearby GalaxieS) survey.
The goal: probe the physics of gas, dust, and star formation, on scales of ~15-150 pc.

Figure from Baron et al. (2024)

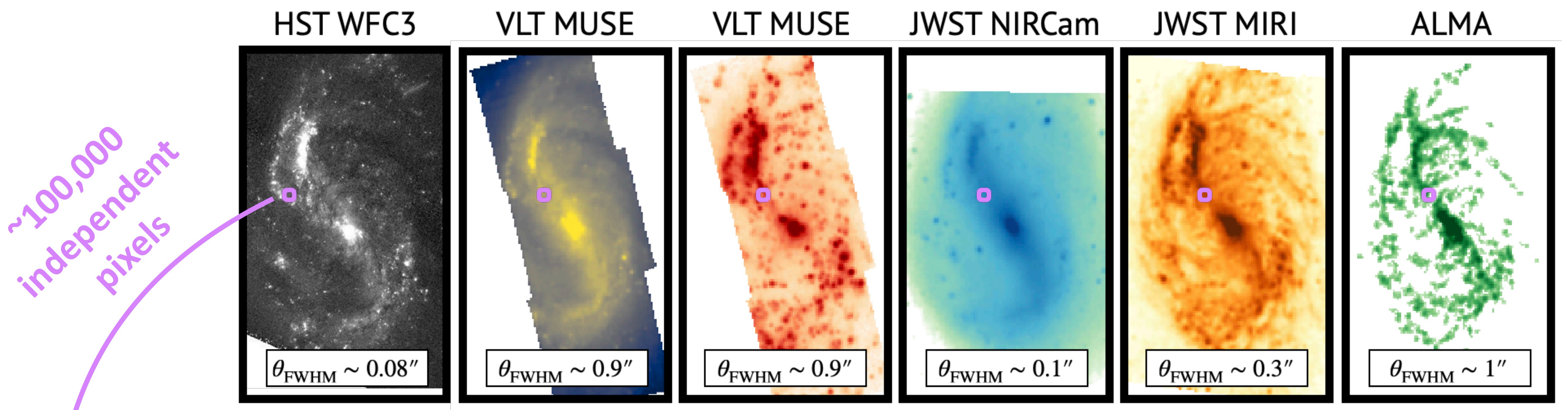
# Modern astronomical surveys produce complex and diverse datasets using observations across the electromagnetic spectrum


HST WFC3
VLT MUSE
VLT MUSE
JWST NIRCam
JWST MIRI
ALMA
$\theta_{\rm FWHM} \sim 0.08''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.1''$
$\theta_{\rm FWHM} \sim 0.3''$
$\theta_{\rm FWHM} \sim 1''$

Observations of NGC 7496 by the **PHANGS** (The Physics at High Angular resolution in Nearby GalaxieS) survey.
The goal: probe the physics of gas, dust, and star formation, on scales of ~15-150 pc.

Figure from Baron et al. (2024)

# Modern astronomical surveys produce complex and diverse datasets using observations across the electromagnetic spectrum


HST WFC3
VLT MUSE
VLT MUSE
JWST NIRCam
JWST MIRI
ALMA
~100,000 independent pixels
$\theta_{\mathrm{FWHM}} \sim 0.08''$
$\theta_{\mathrm{FWHM}} \sim 0.9''$
$\theta_{\mathrm{FWHM}} \sim 0.9''$
$\theta_{\mathrm{FWHM}} \sim 0.1''$
$\theta_{\mathrm{FWHM}} \sim 0.3''$
$\theta_{\mathrm{FWHM}} \sim 1''$

Observations of NGC 7496 by the **PHANGS** (The Physics at High Angular resolution in Nearby GalaxieS) survey.
The goal: probe the physics of gas, dust, and star formation, on scales of ~15-150 pc.

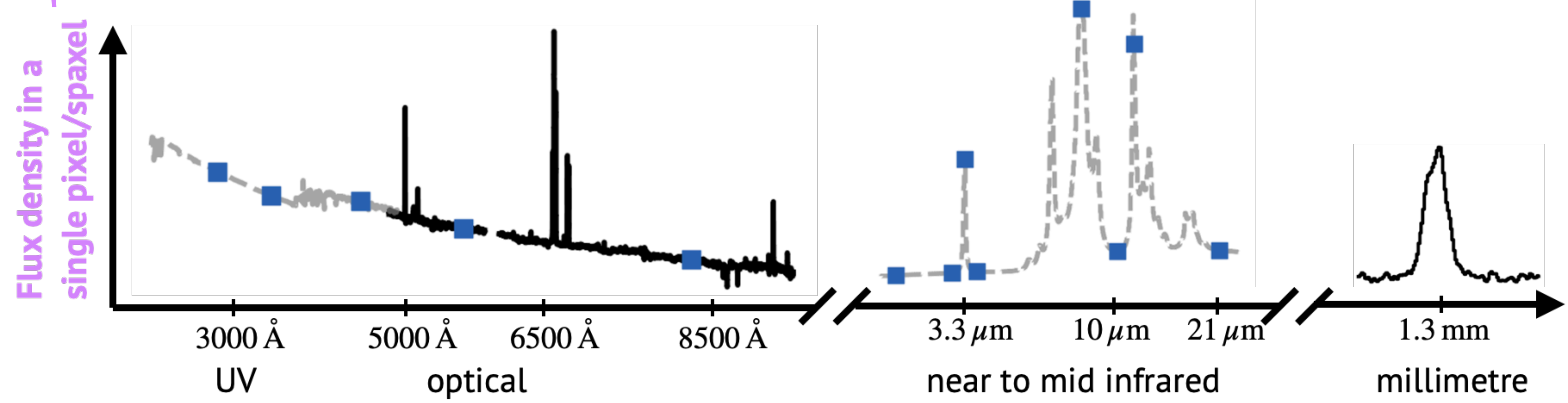


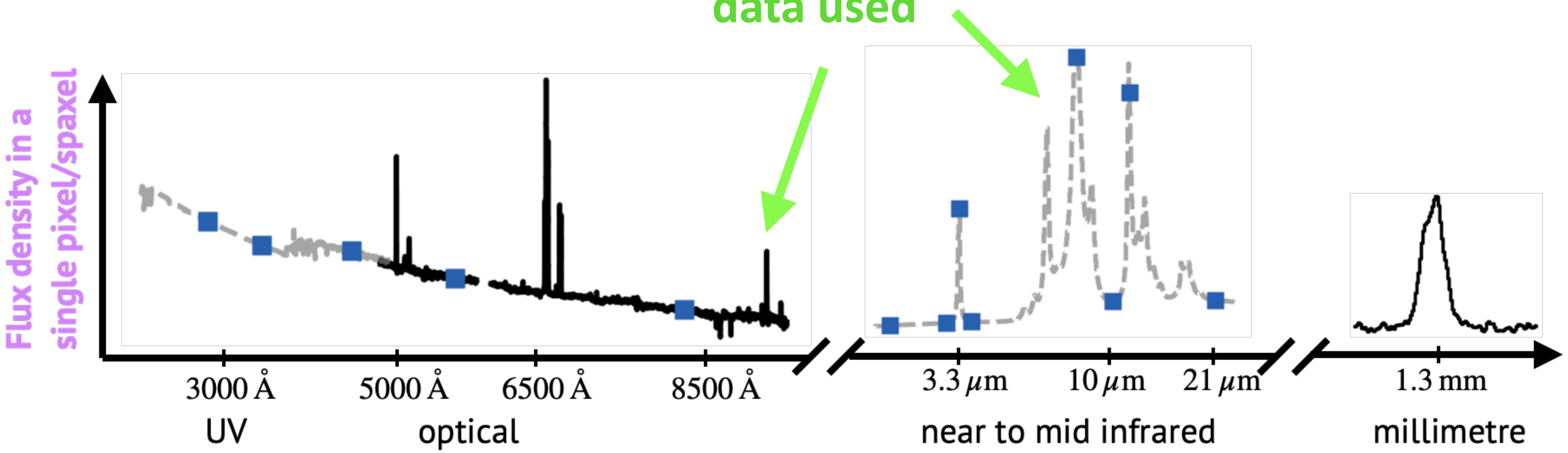
Figure from Baron et al. (2024)

PHANGS website (includes survey papers and public datasets)

# How to analyze such complex and high-dimensional datasets?

## Model-driven or theory-driven approach

1. Hypothesis (from theory or from a previous experiment).
2. Observations and analysis are conducted to test the hypothesis.
3. Analysis leads to new insight, often leading to new hypotheses.



data used

Inspired by Egorov et al. (2023)

# How to analyze such complex and high-dimensional datasets?

## Data-driven approach

1. Statistical tools are used to visualize and dissect the high-dimensional dataset.
2. Representation reveals trends; groups; or outliers, forming a new hypothesis.
3. Analysis of the data to test the hypothesis leads to new insight.

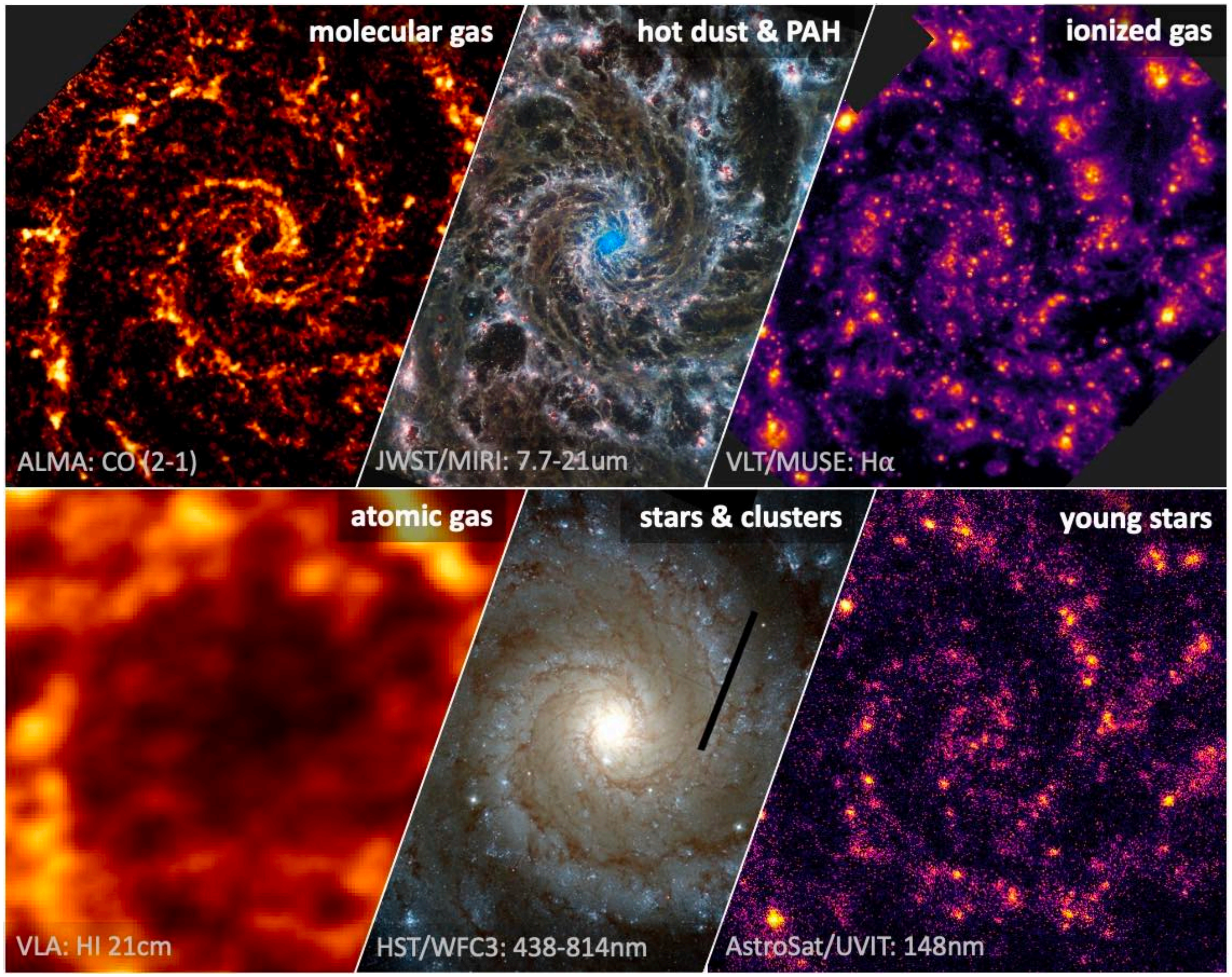
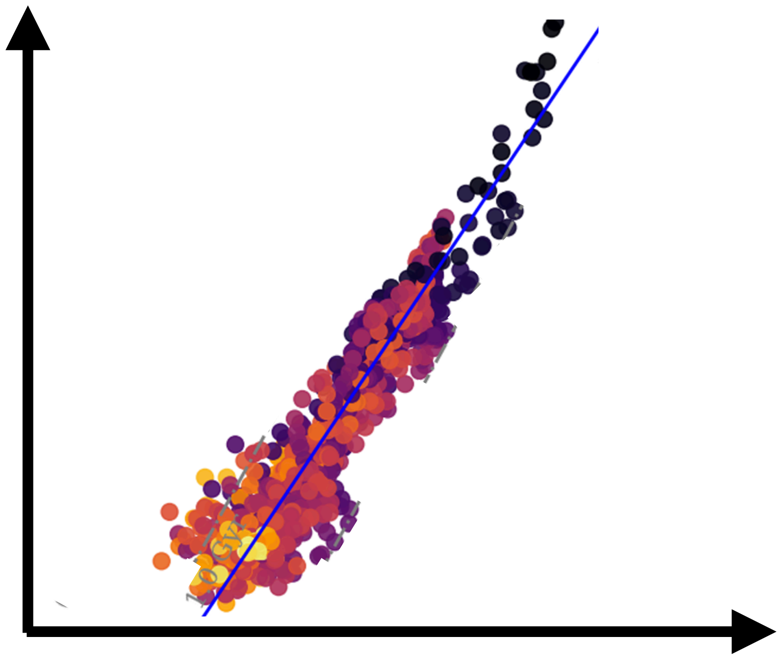


Image credit: Jiayi Sun



A space (or point of view) in which the data appear in the greatest simplicity (J W Gibbs 1881).

# How to analyze such complex and high-dimensional datasets?

## Data-driven approach

1. Statistical tools are used to visualize and dissect the high-dimensional dataset.
2. Representation reveals trends; groups; or outliers, forming a new hypothesis.
3. Analysis of the data to test the hypothesis leads to new insight.

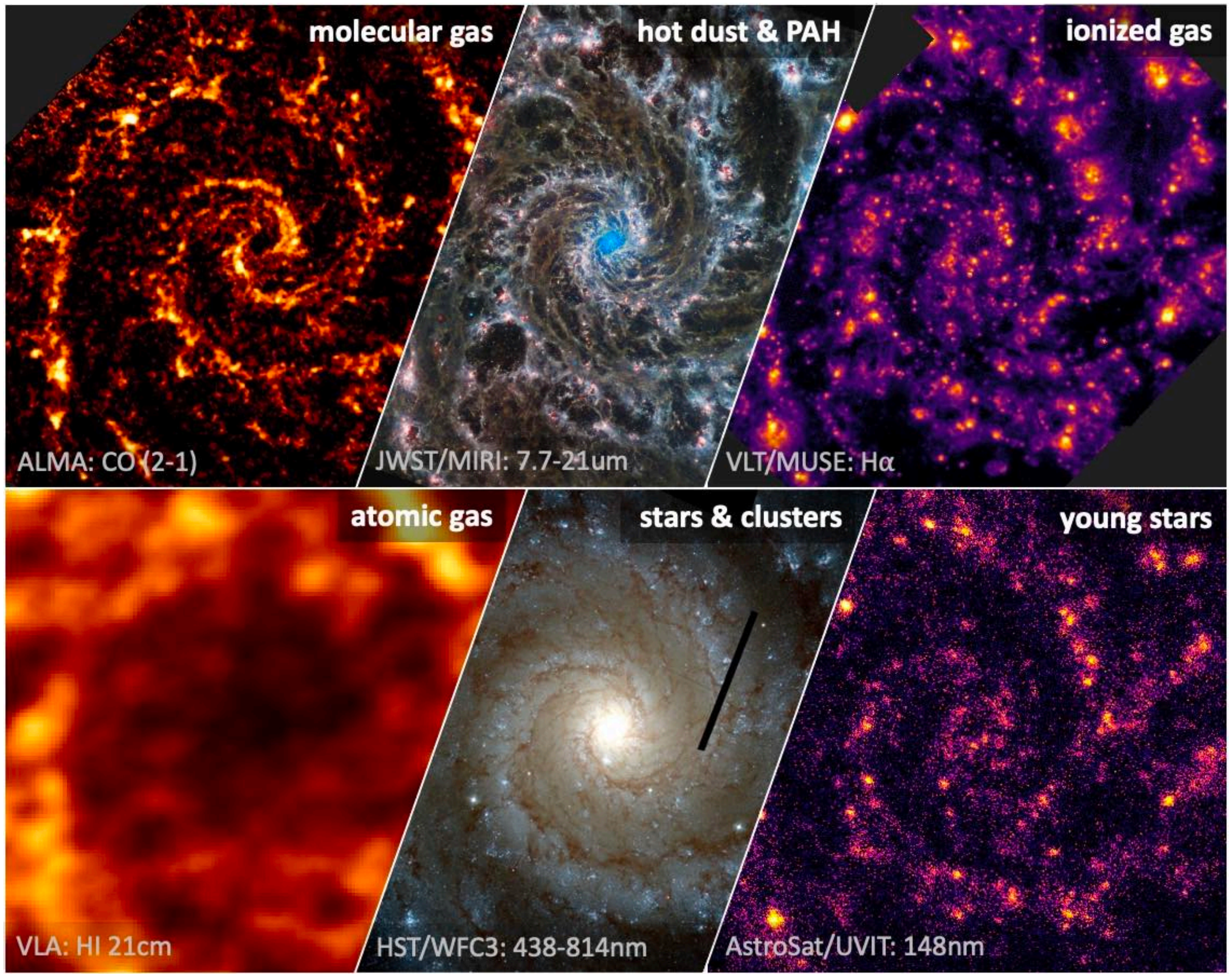
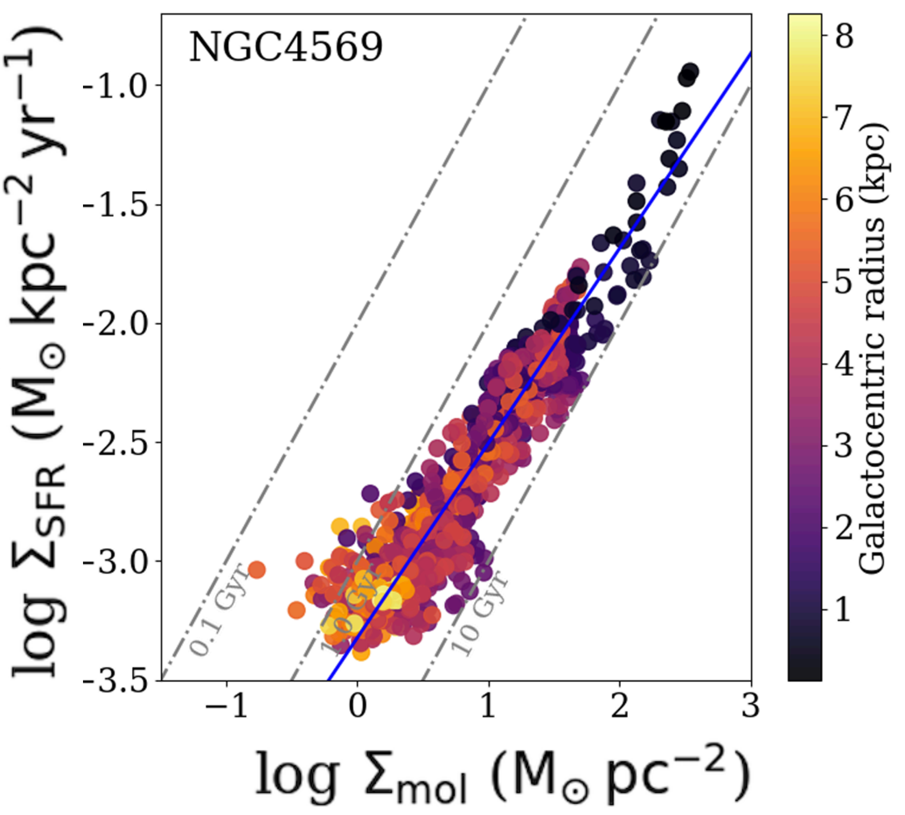


Image credit: Jiayi Sun



Kennicutt-Schmidt relation

(figure from Jiménez-Donaire et al. 2023)

# Simple manifolds in high-dimensional spaces lead to insight: the ultimate example

optical spectra of stars

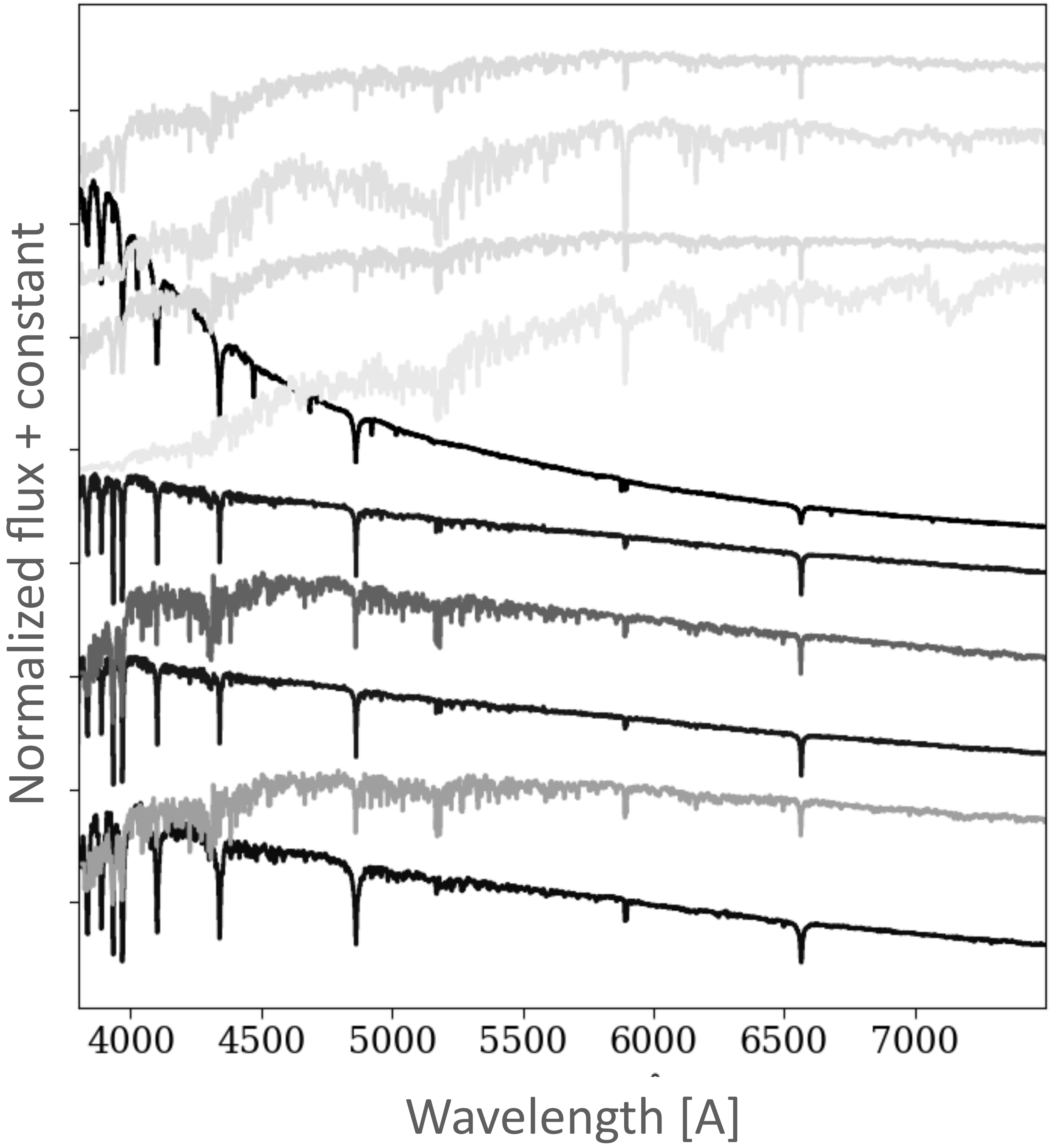
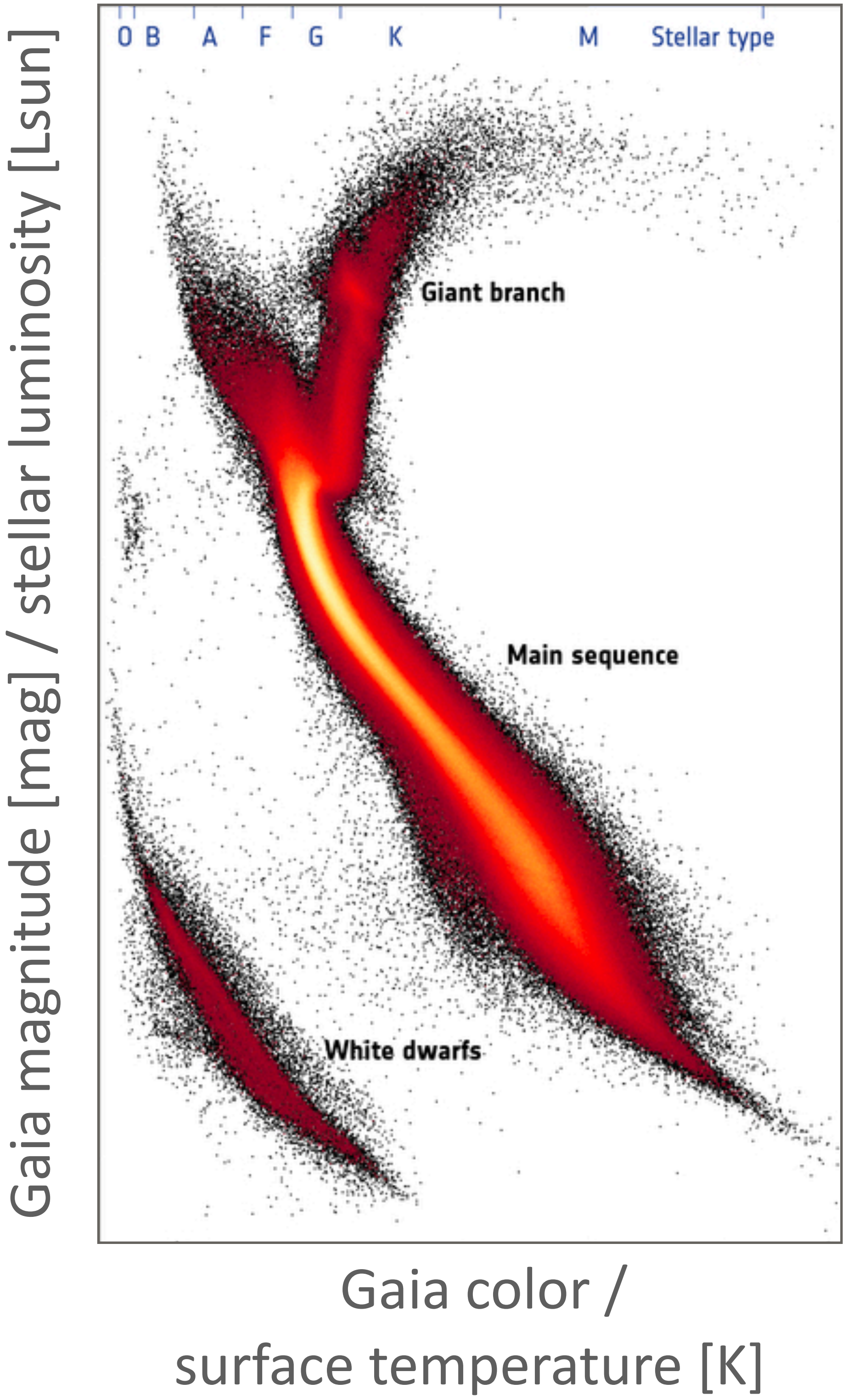


Image credit: Gaia collaboration (2018)



# Simple manifolds in high-dimensional spaces lead to insight: the ultimate example

optical spectra of stars

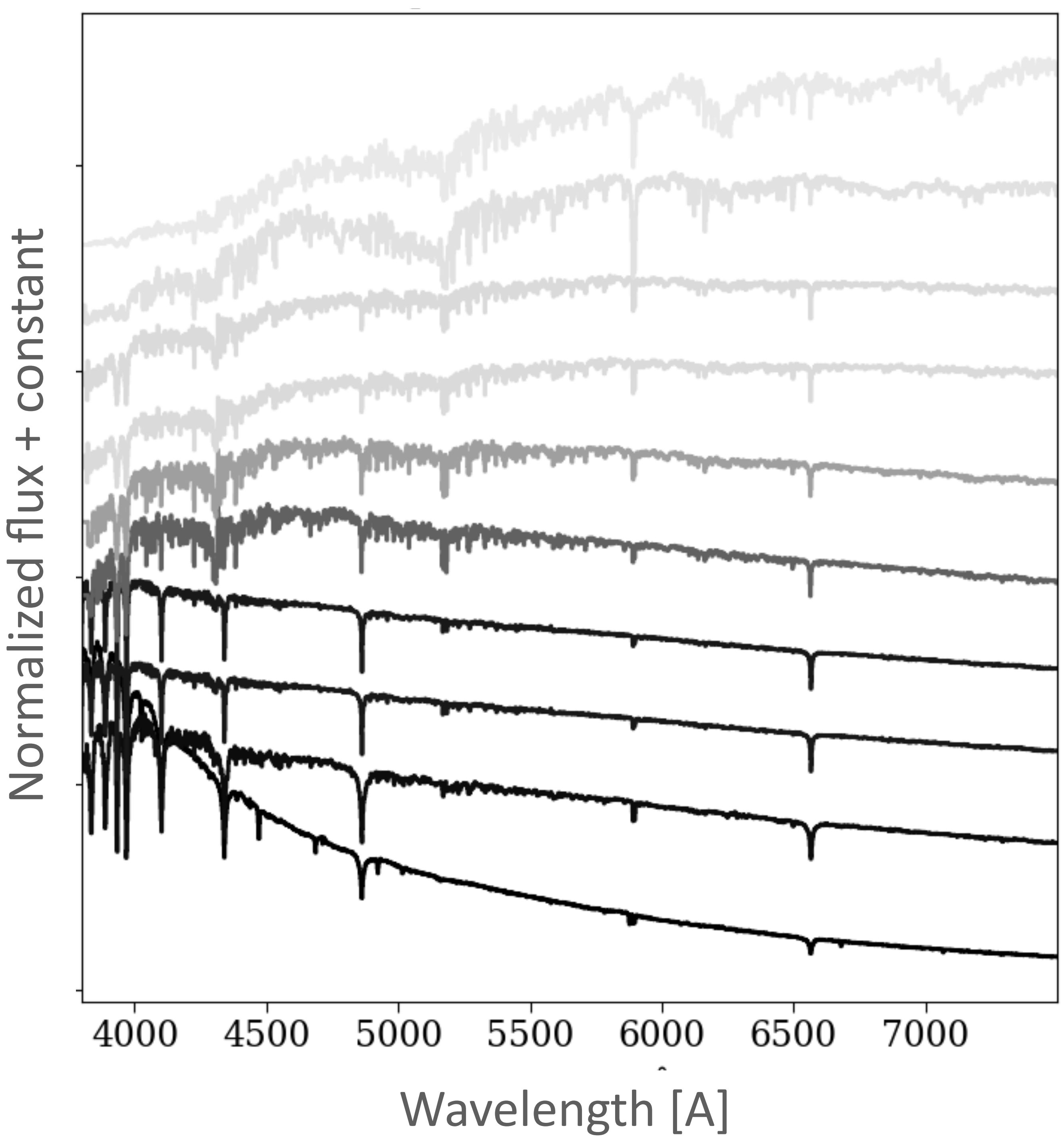
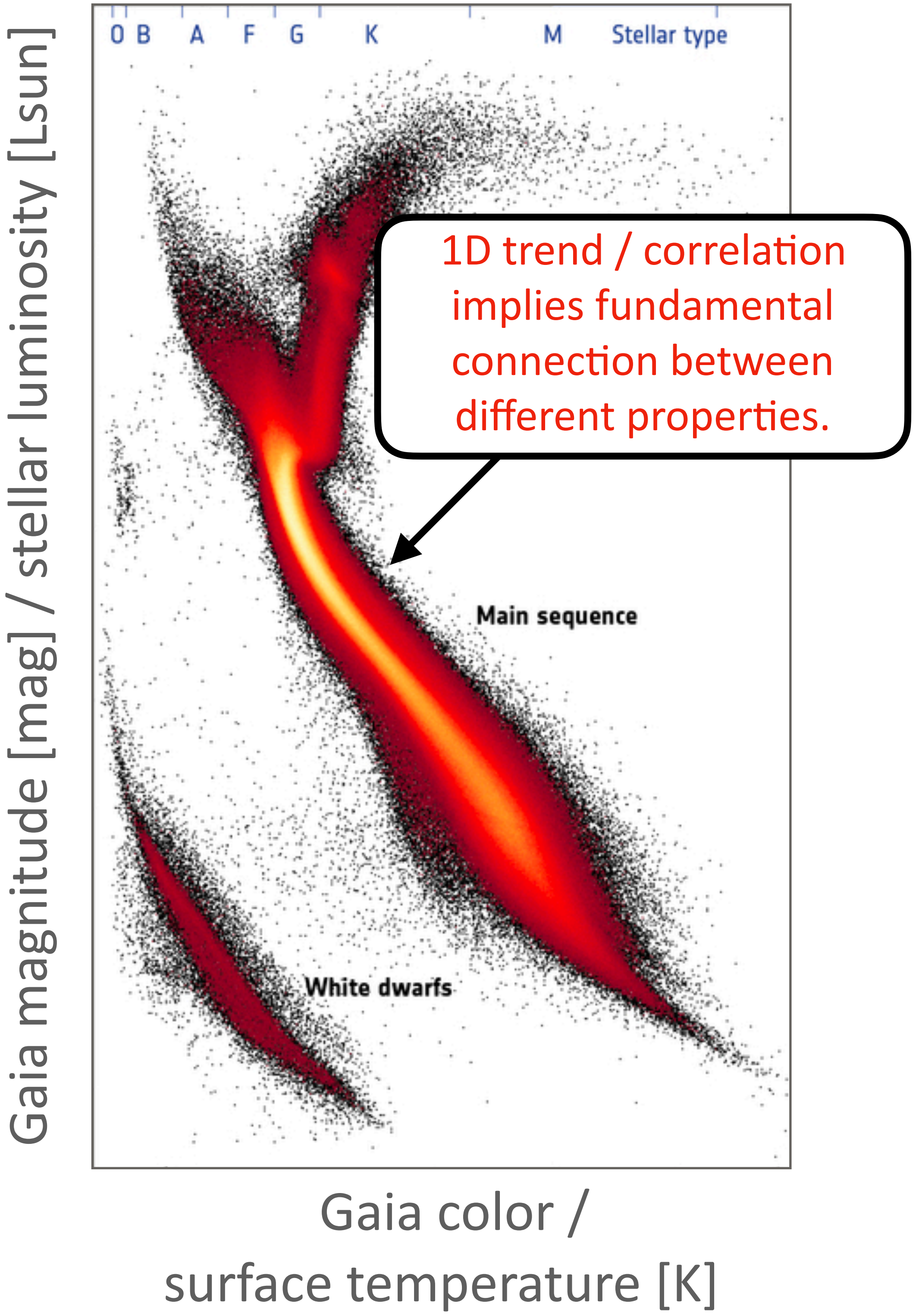


Image credit: Gaia collaboration (2018)



1D trend / correlation implies fundamental connection between different properties.

# Simple manifolds in high-dimensional spaces lead to insight: the ultimate example

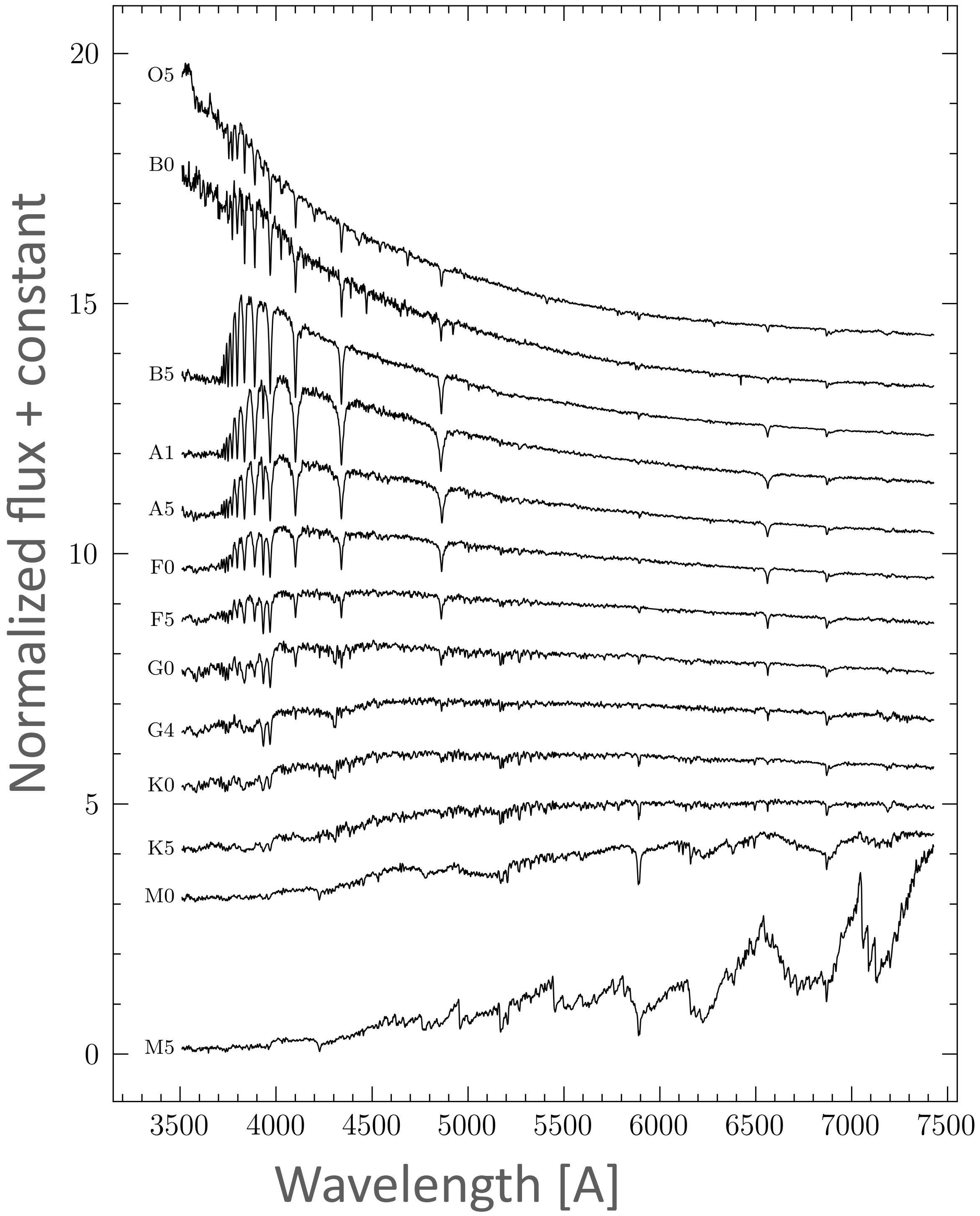
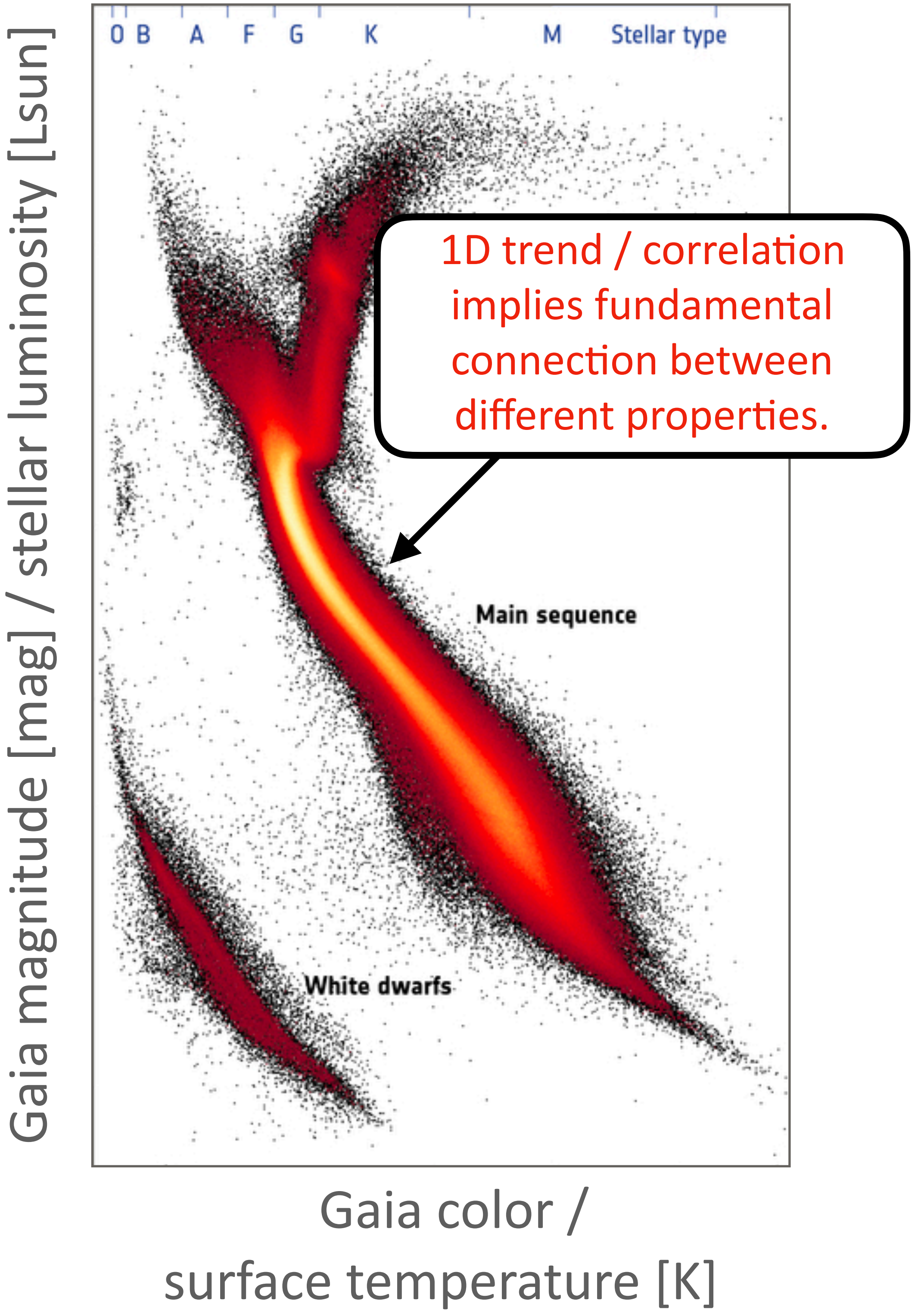


Image credit: Gaia collaboration (2018)



1D trend / correlation implies fundamental connection between different properties.

# Simple manifolds in high-dimensional spaces lead to insight: the ultimate example

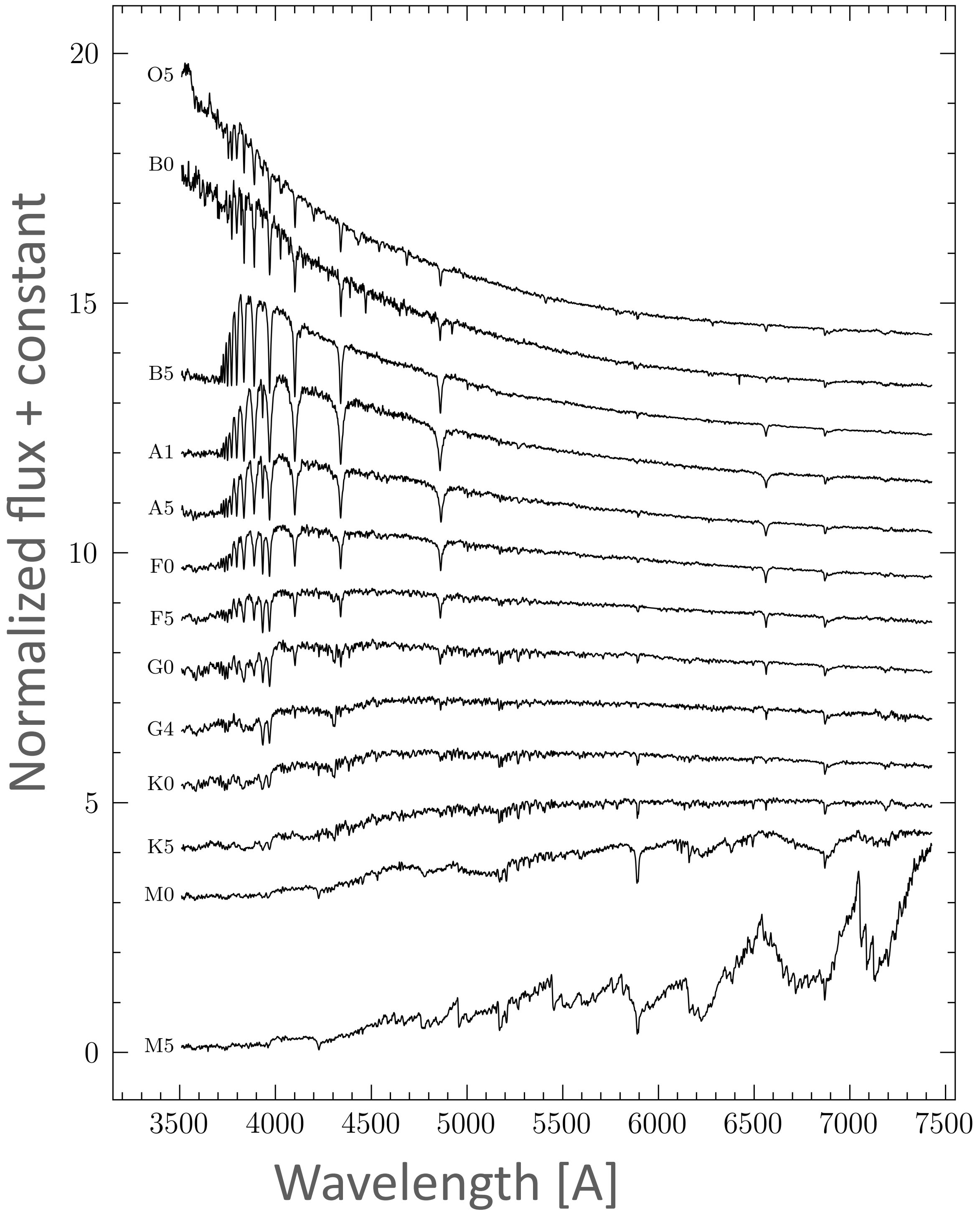
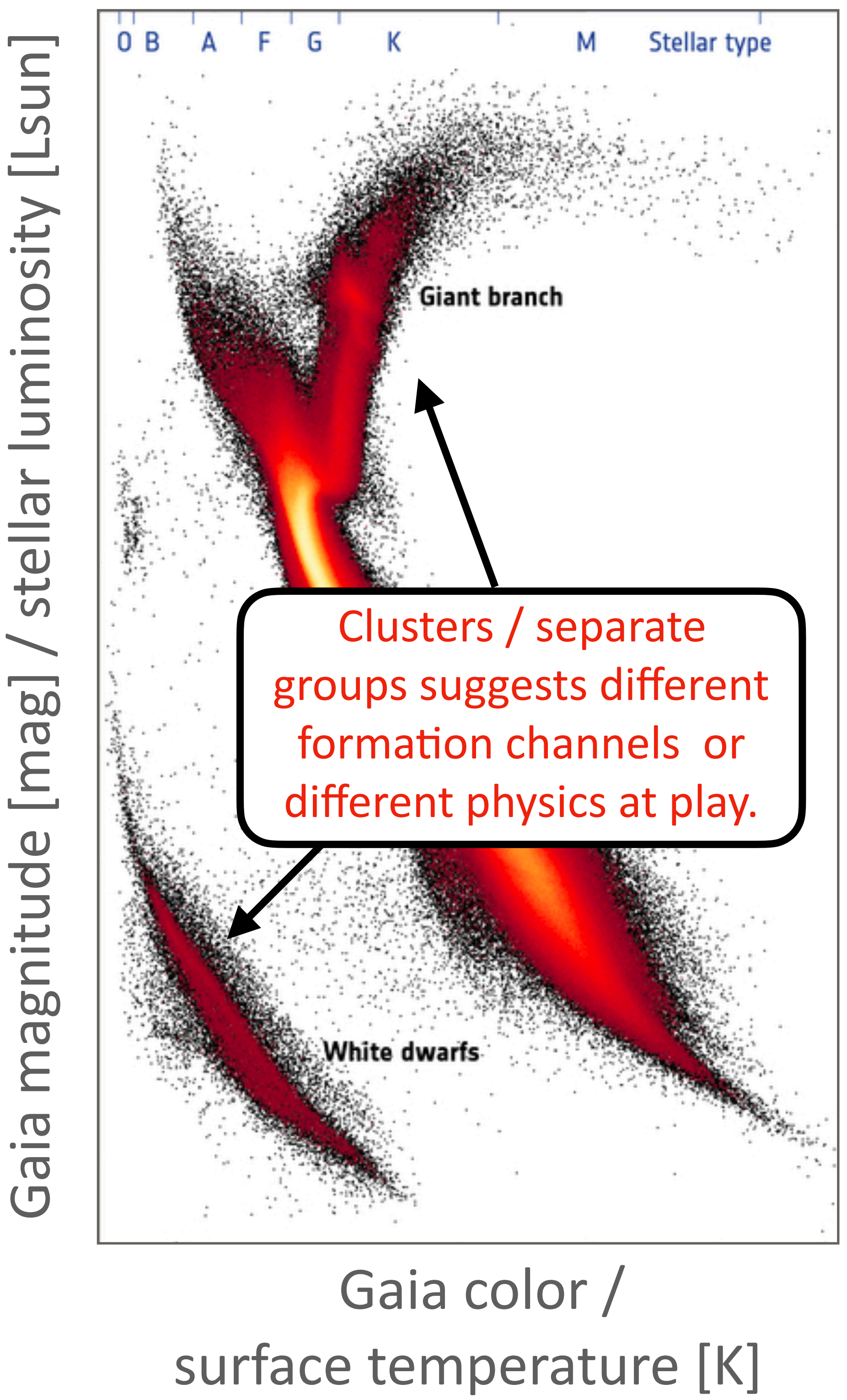


Image credit: Gaia collaboration (2018)



Clusters / separate groups suggests different formation channels or different physics at play.

# Simple manifolds in high-dimensional spaces lead to insight: the ultimate example

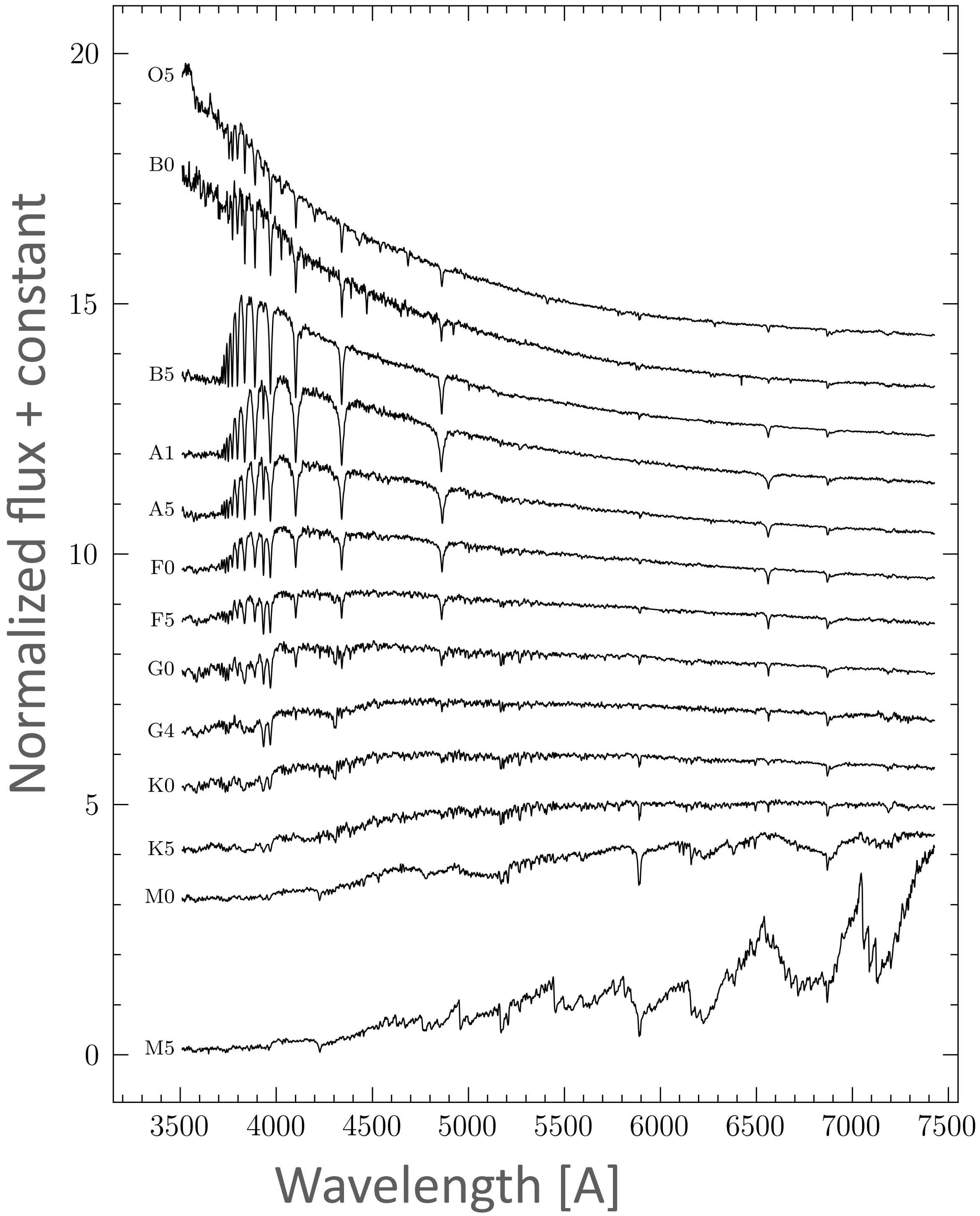
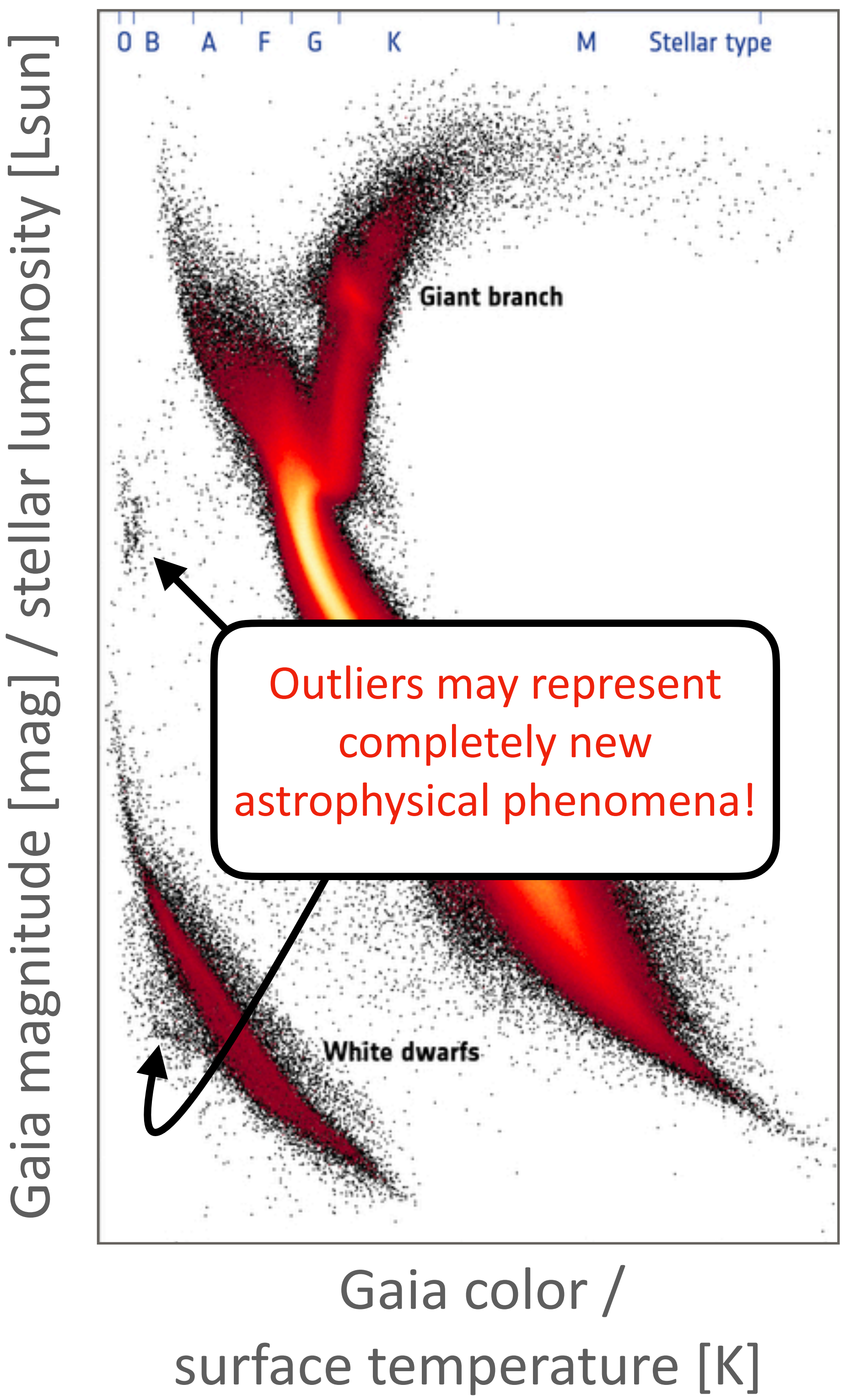


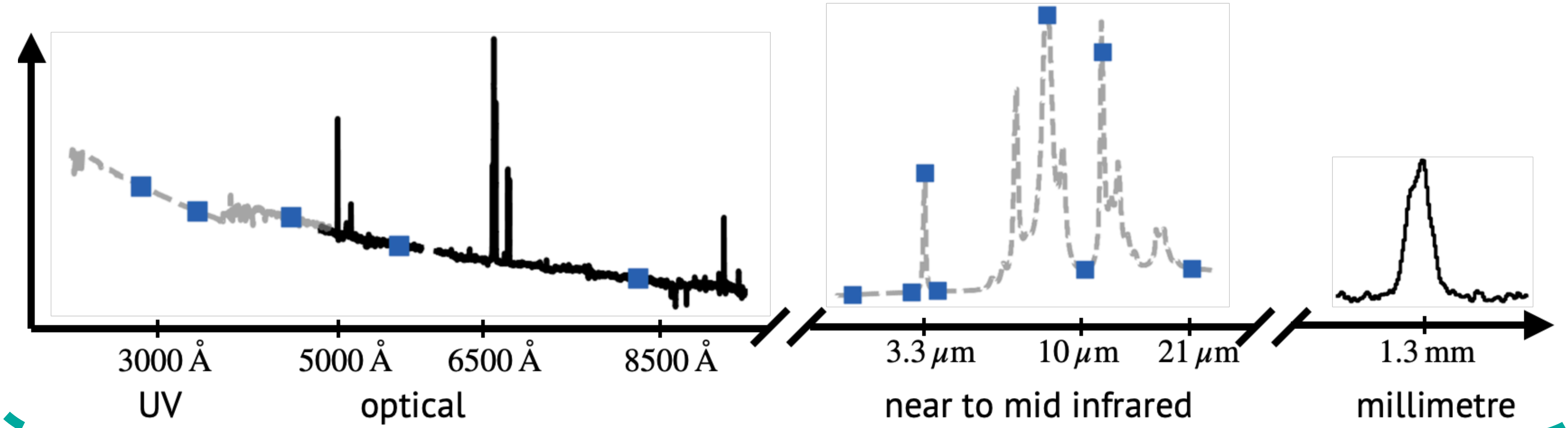
Image credit: Gaia collaboration (2018)



Outliers may represent completely new astrophysical phenomena!

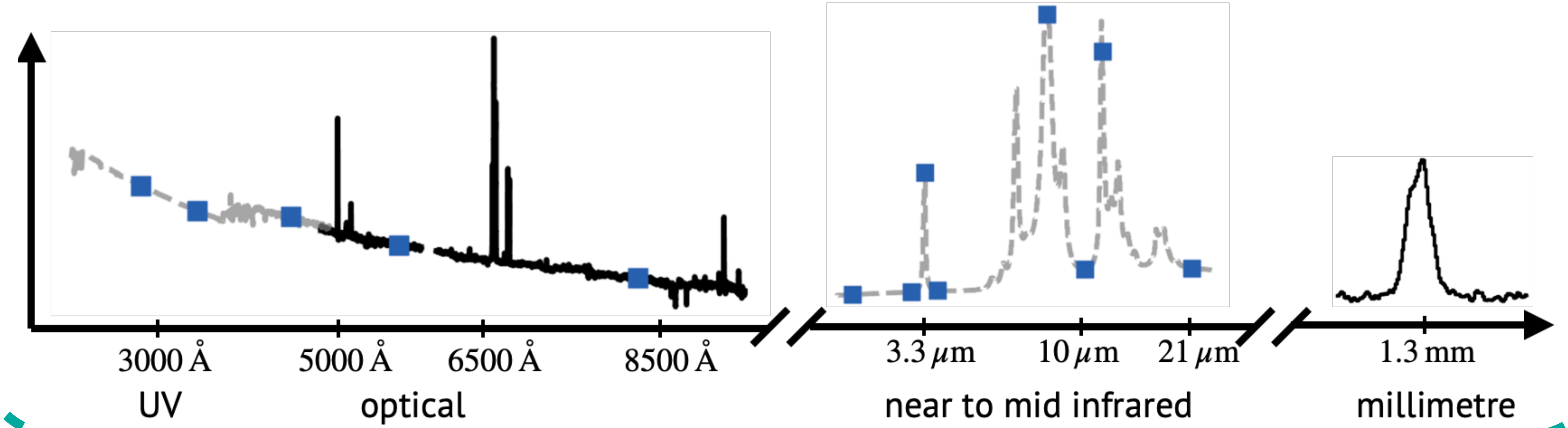
# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



# Using unsupervised machine learning tools to make discoveries

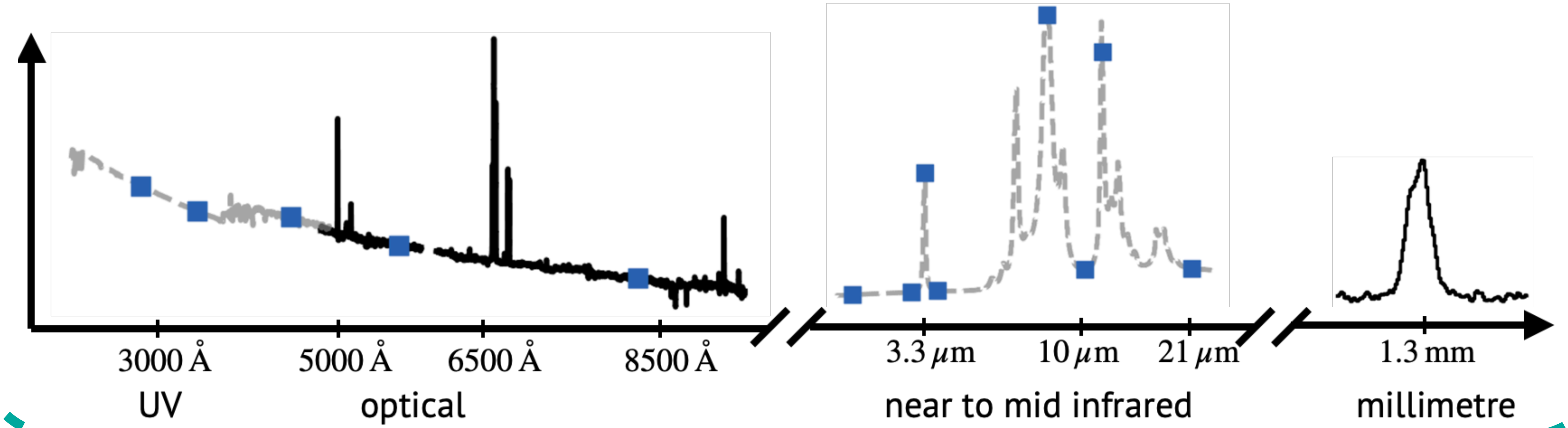
input data for example: ~100,000 objects with the following information:



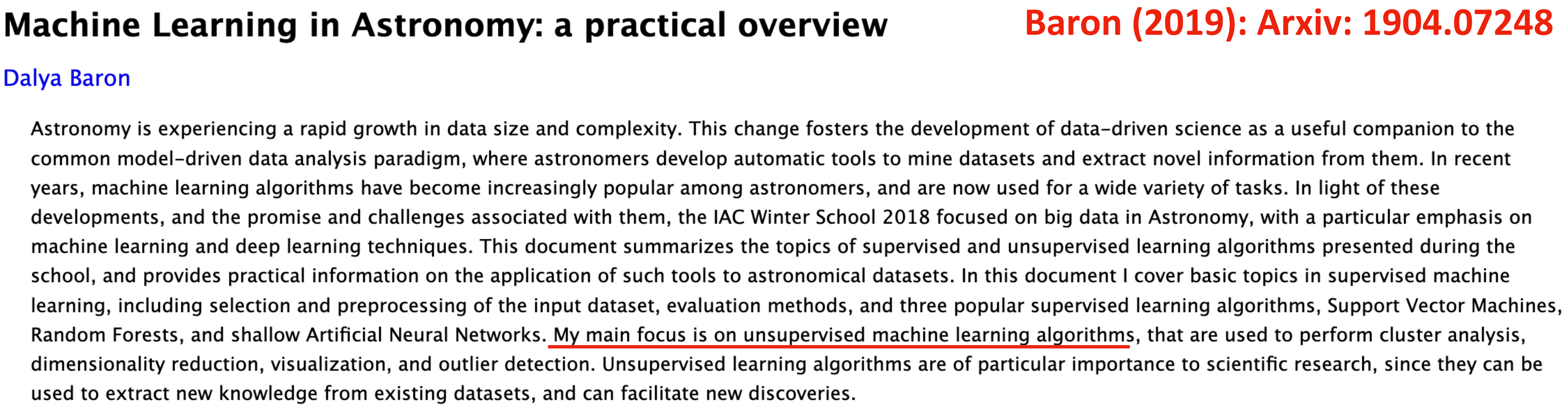
Bridging the Gap between Spectra and Light-Curves: Exploring Multimodal Learning for the Identification of Broad Absorption Line Quasars by N. Guerra-Varas (Tuesday)

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



Baron (2019): Arxiv: 1904.07248



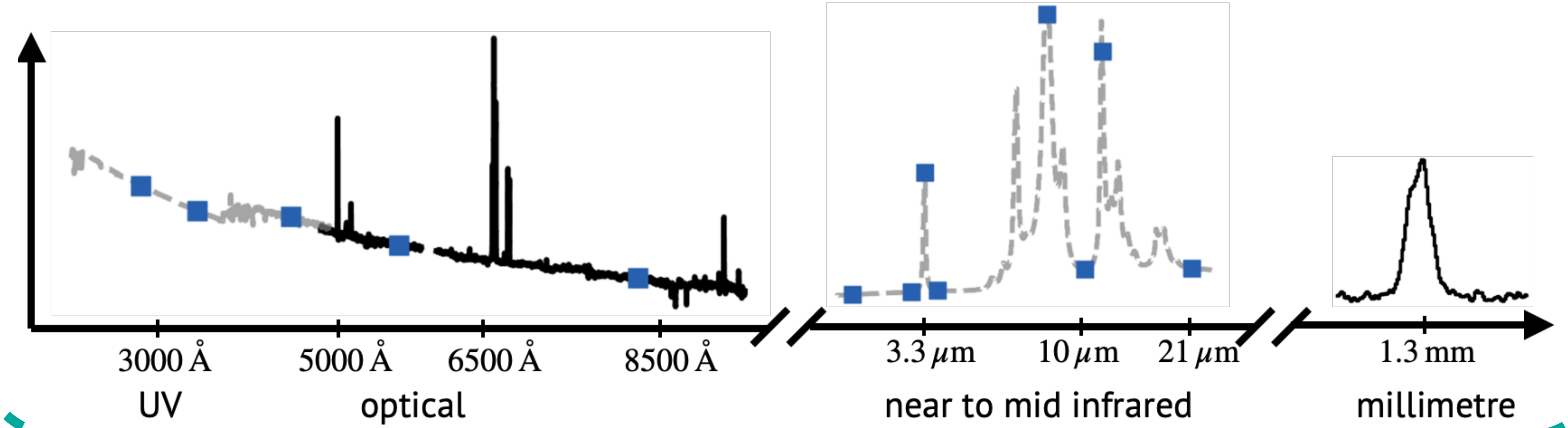
## Machine Learning in Astronomy: a practical overview

Dalya Baron

Astronomy is experiencing a rapid growth in data size and complexity. This change fosters the development of data-driven science as a useful companion to the common model-driven data analysis paradigm, where astronomers develop automatic tools to mine datasets and extract novel information from them. In recent years, machine learning algorithms have become increasingly popular among astronomers, and are now used for a wide variety of tasks. In light of these developments, and the promise and challenges associated with them, the IAC Winter School 2018 focused on big data in Astronomy, with a particular emphasis on machine learning and deep learning techniques. This document summarizes the topics of supervised and unsupervised learning algorithms presented during the school, and provides practical information on the application of such tools to astronomical datasets. In this document I cover basic topics in supervised machine learning, including selection and preprocessing of the input dataset, evaluation methods, and three popular supervised learning algorithms, Support Vector Machines, Random Forests, and shallow Artificial Neural Networks. My main focus is on unsupervised machine learning algorithms, that are used to perform cluster analysis, dimensionality reduction, visualization, and outlier detection. Unsupervised learning algorithms are of particular importance to scientific research, since they can be used to extract new knowledge from existing datasets, and can facilitate new discoveries.

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



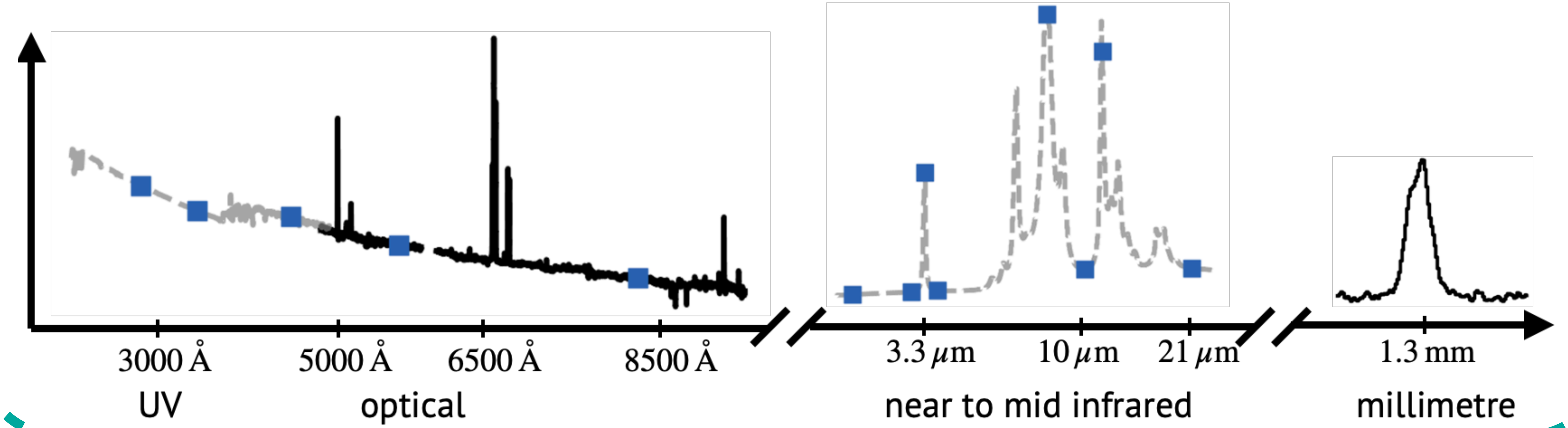
## Clustering algorithms

Objects in the sample are divided into a (typically small) number of groups.

Examples: KMeans; Hierarchical clustering; DBSCAN; Gaussian Mixture models; and using Self Organizing Maps (SOMs).

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



## Clustering algorithms

Objects in the sample are divided into a (typically small) number of groups.

Examples: KMeans; Hierarchical clustering; DBSCAN; Gaussian Mixture models; and using Self Organizing Maps (SOMs).

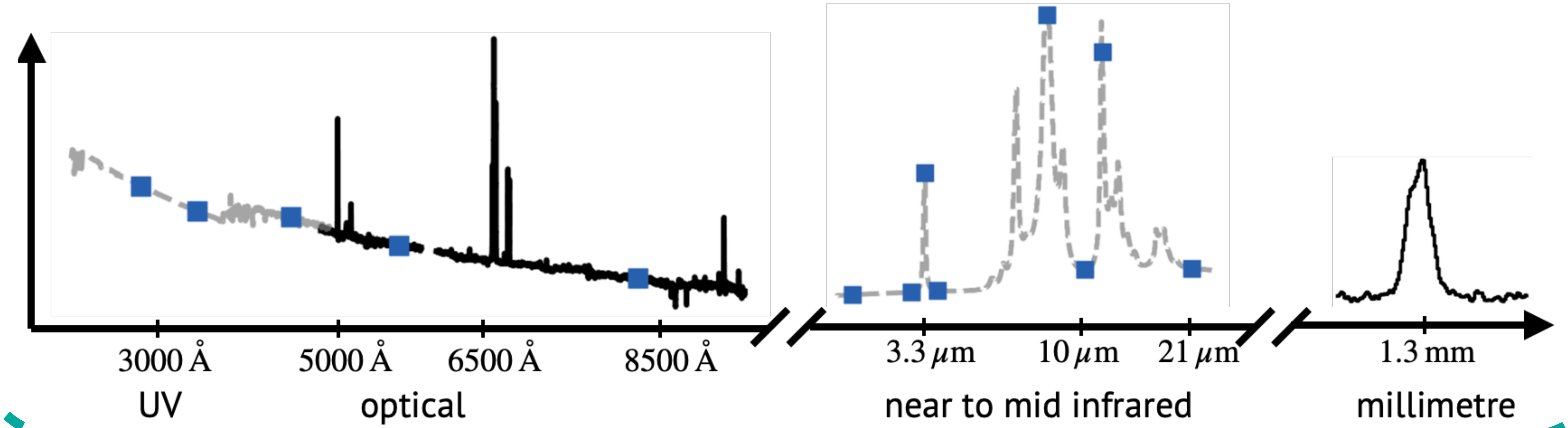
Galaxy Morphological Classification via Unsupervised Machine Learning in the Big Data Era led by JWST, EUCLID, LSST and SKA by I. Lazar (Tuesday).

Applying SOMs to identify and classify complex radio morphologies in next-generation radio surveys by A. Alam (today!)

Early-stopping SOMs as a tool to classify SSOs in space surveys by S. Sacquegna (Wednesday)

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



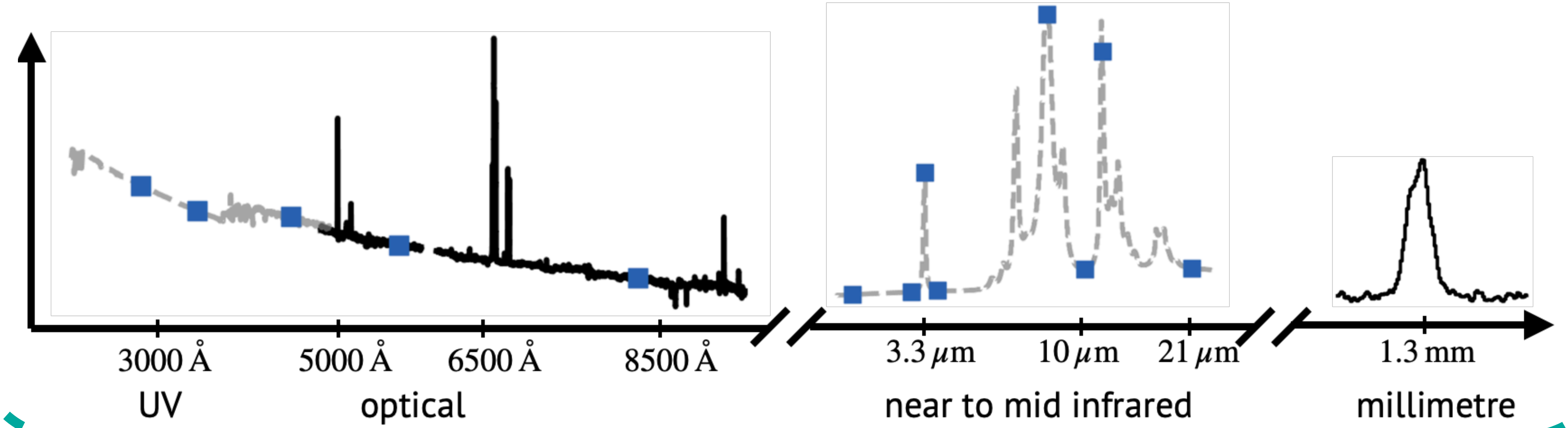
## Dimensionality Reduction algorithms

Objects are embedded into a lower-dimensional space (typically 2D or 3D).

Examples: Principal component analysis (PCA); Local linear embedding (LLE); tSNE; UMAP; SOMs; Autoencoders of various types.

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



## Dimensionality Reduction algorithms

Objects are embedded into a lower-dimensional space (typically 2D or 3D).

Examples: Principal component analysis (PCA); Local linear embedding (LLE); tSNE; UMAP; SOMs; Autoencoders of various types.

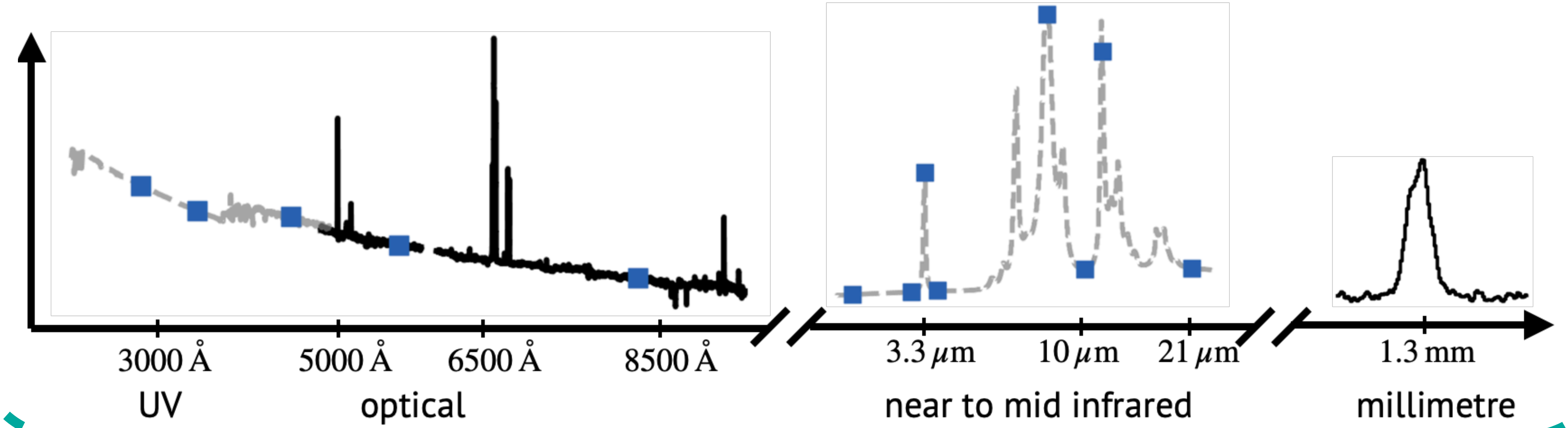
Multi-Task Neural Nets with Monte Carlo Dropout for Spectral Analysis of Galaxies by M. Ginolfi (Tuesday)

Exploring Multi-Band Imaging for Identifying z>6.5 Quasars: A Contrastive Learning Approach Using HSC Data by L. N. M. Ramirez (Tuesday)

Unsupervised Machine Learning Techniques for Young Stellar Object Light Curve Characterization by M. Madarász (Tuesday)

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



## Dimensionality Reduction algorithms

Objects are embedded into a lower-dimensional space (typically 2D or 3D).

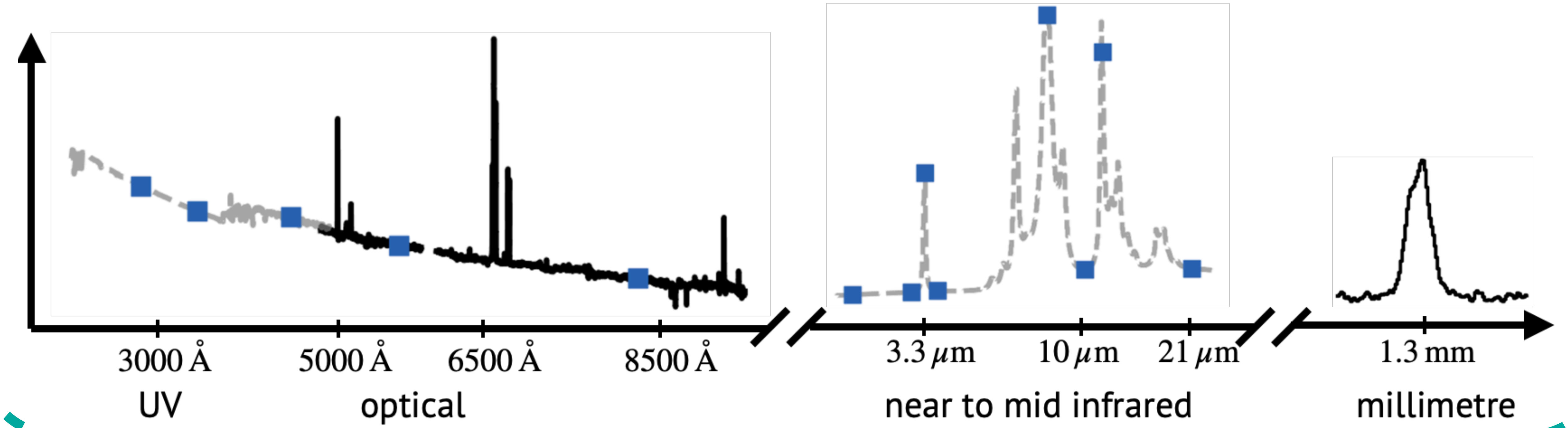
Examples: Principal component analysis (PCA); Local linear embedding (LLE); tSNE; UMAP; SOMs; Autoencoders of various types.

Unsupervised learning for GRBs classification by N. Cibrario (Wednesday)

Machine Learning Based Parametrization of Solar Active Regions Using Disentangled Variational Autoencoders by E. Dineva (Friday)

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



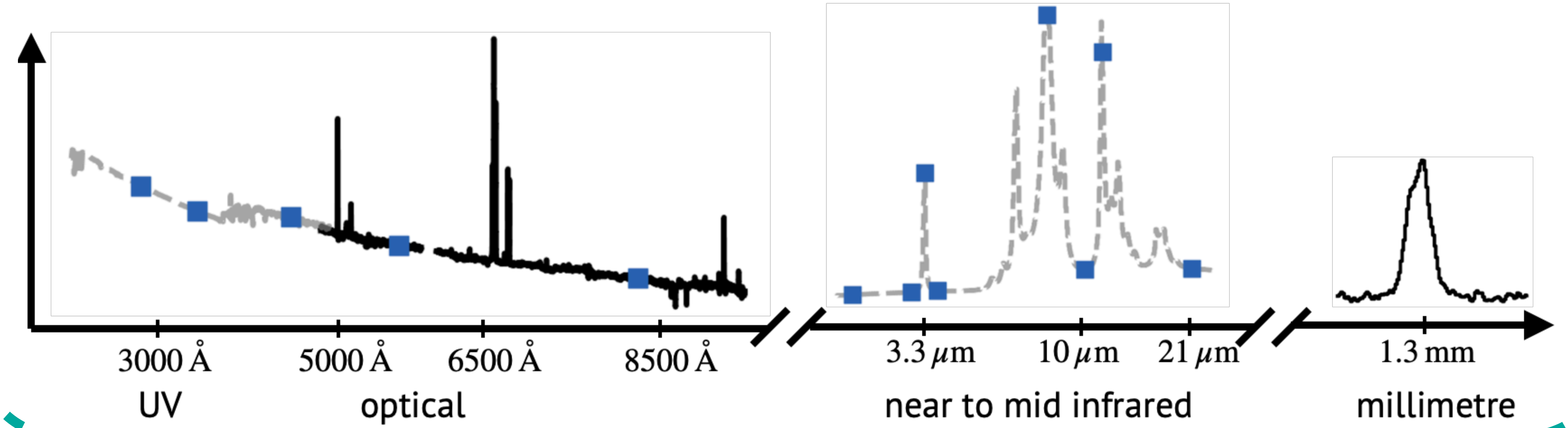
## Outlier Detection algorithms

A (very small) subset of the objects in the dataset is flagged as anomalies.

Examples: Isolation forests; Local outlier factor; using supervised learning algorithms; using dimensionality reduction algorithms.

# Using unsupervised machine learning tools to make discoveries

input data for example: ~100,000 objects with the following information:



## Outlier Detection algorithms

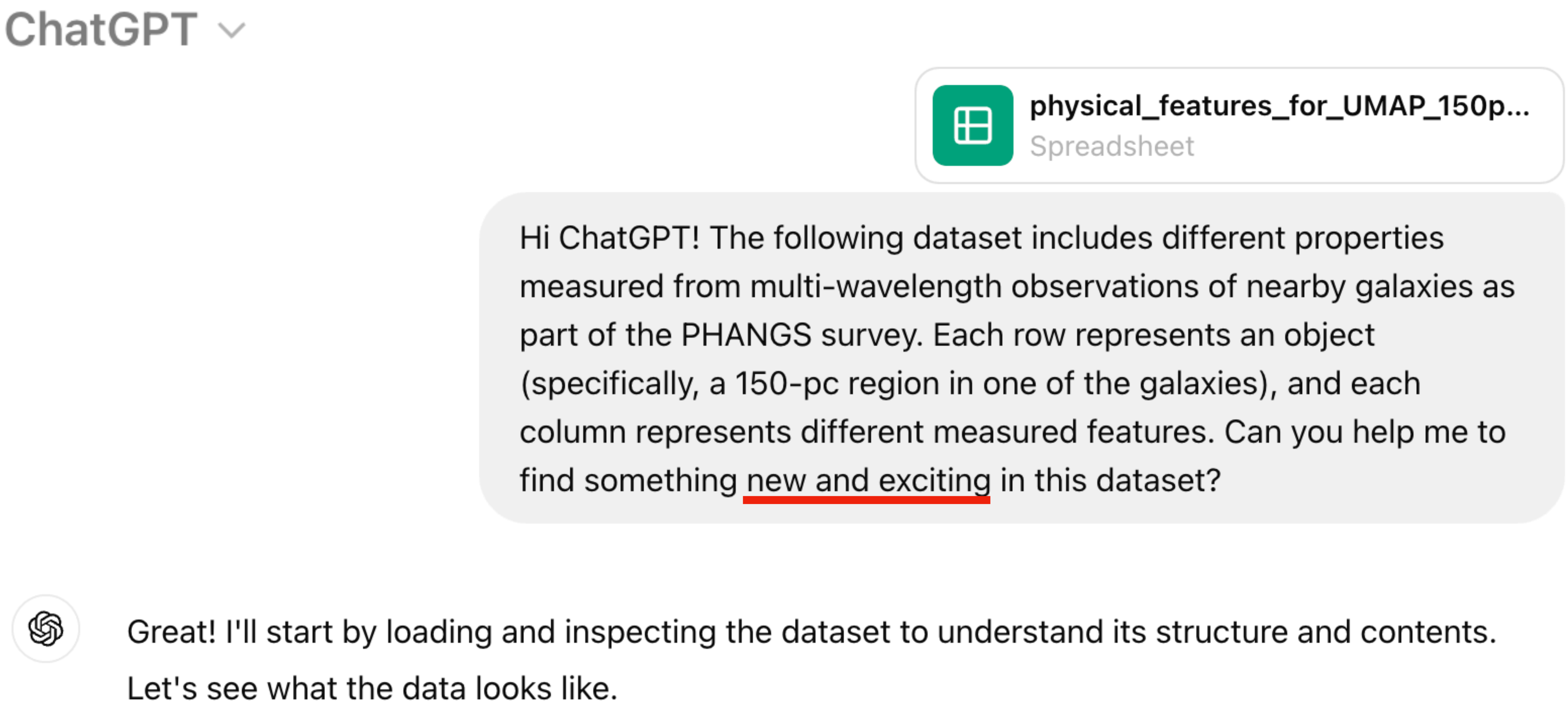
A (very small) subset of the objects in the dataset is flagged as anomalies.

Examples: Isolation forests; Local outlier factor; using supervised learning algorithms; using dimensionality reduction algorithms.

Signatures to help interpretability of anomalies in astronomical data by E. Gangler (Tuesday)

Unmasking the Hidden: Anomaly Detection in ASKAP's Monitoring by Z. Wang (Wednesday)

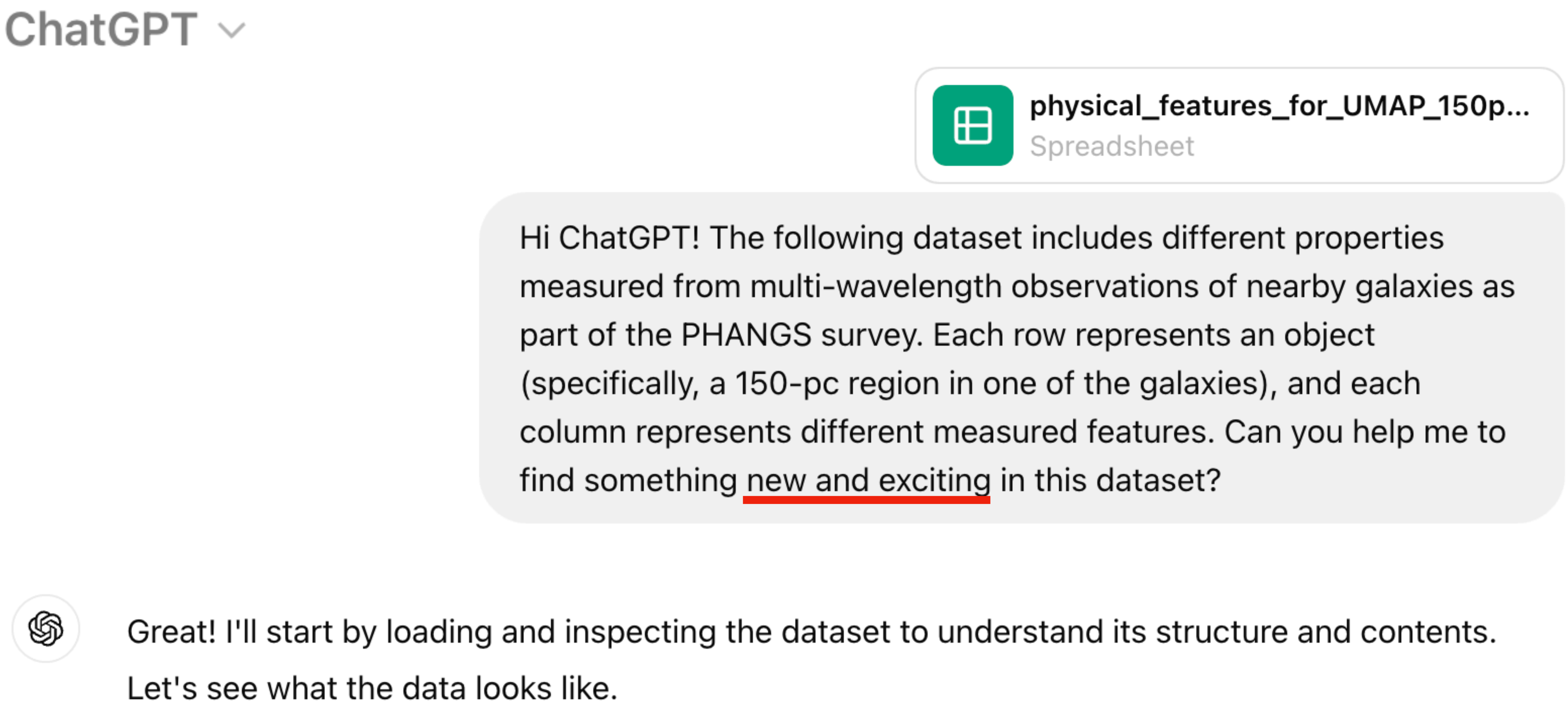
# Let’s discover!



Hi ChatGPT! The following dataset includes different properties measured from multi-wavelength observations of nearby galaxies as part of the PHANGS survey. Each row represents an object (specifically, a 150-pc region in one of the galaxies), and each column represents different measured features. Can you help me to find something new and exciting in this dataset?

Great! I'll start by loading and inspecting the dataset to understand its structure and contents. Let's see what the data looks like.

# Let’s discover!

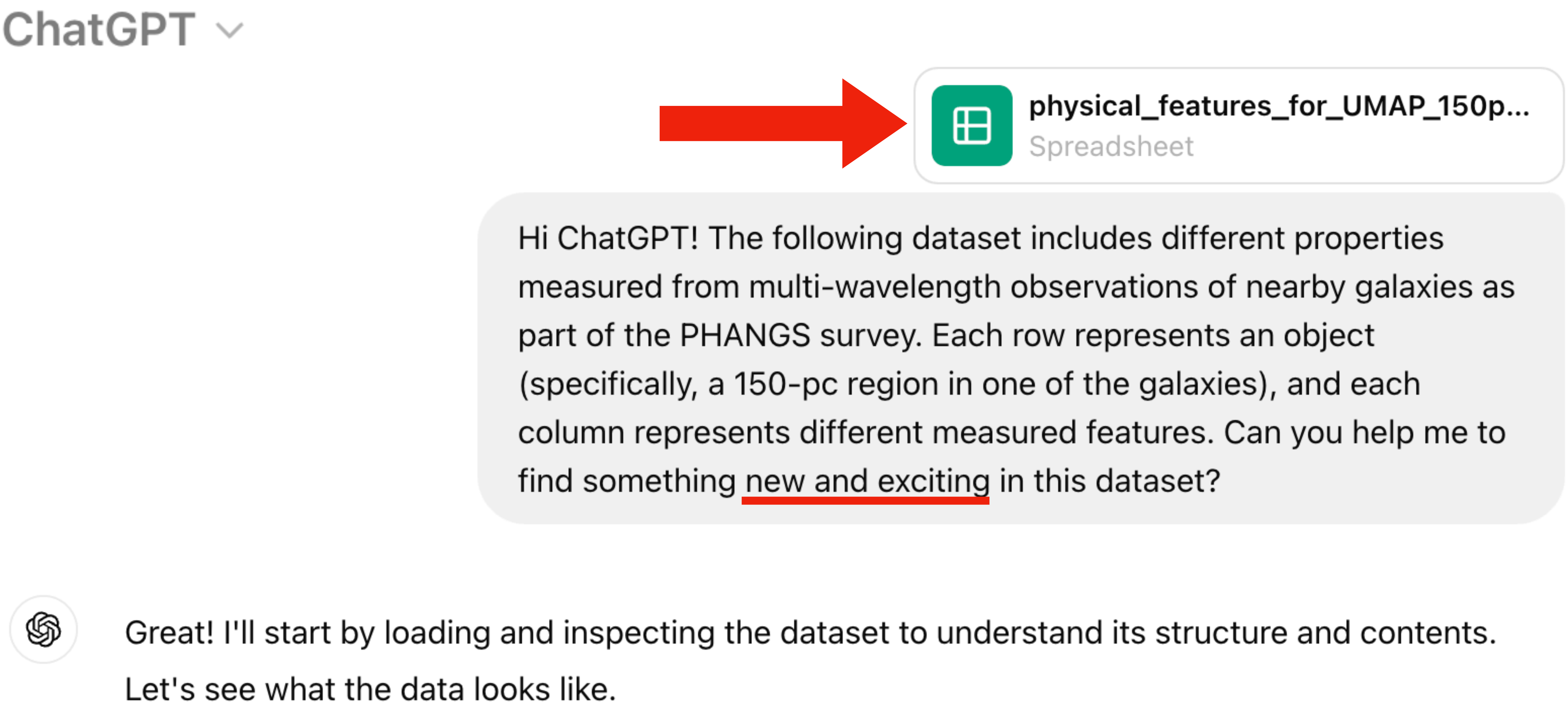


Hi ChatGPT! The following dataset includes different properties measured from multi-wavelength observations of nearby galaxies as part of the PHANGS survey. Each row represents an object (specifically, a 150-pc region in one of the galaxies), and each column represents different measured features. Can you help me to find something new and exciting in this dataset?

Great! I'll start by loading and inspecting the dataset to understand its structure and contents. Let's see what the data looks like.

The robot did really really well, much better than I have anticipated. It sent me into a small existential crisis… 🫠

# Let’s discover!



ChatGPT

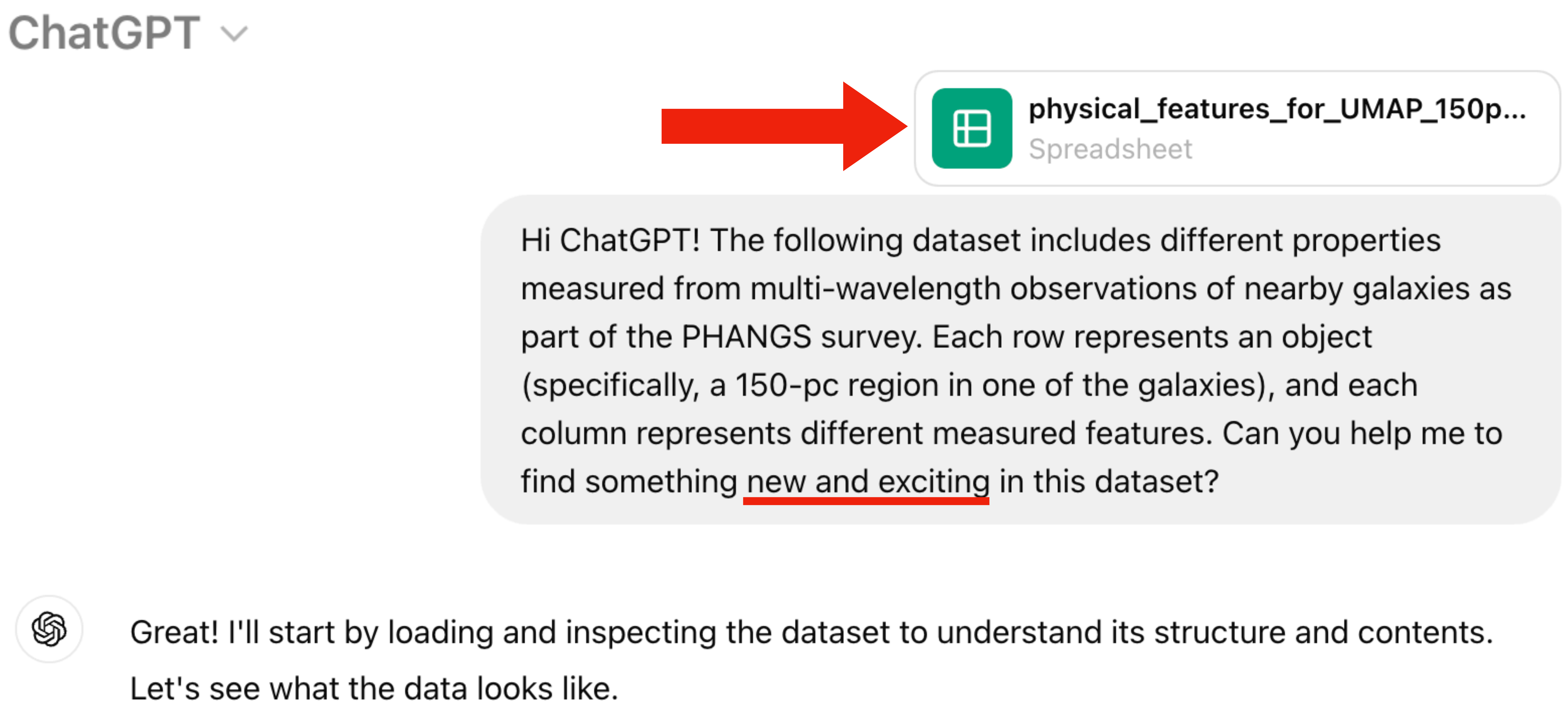
physical_features_for_UMAP_150p...
Spreadsheet

Hi ChatGPT! The following dataset includes different properties measured from multi-wavelength observations of nearby galaxies as part of the PHANGS survey. Each row represents an object (specifically, a 150-pc region in one of the galaxies), and each column represents different measured features. Can you help me to find something new and exciting in this dataset?

Great! I'll start by loading and inspecting the dataset to understand its structure and contents. Let's see what the data looks like.

The robot did really really well, much better than I have anticipated. It sent me into a small existential crisis... 🫠

# Let's discover!

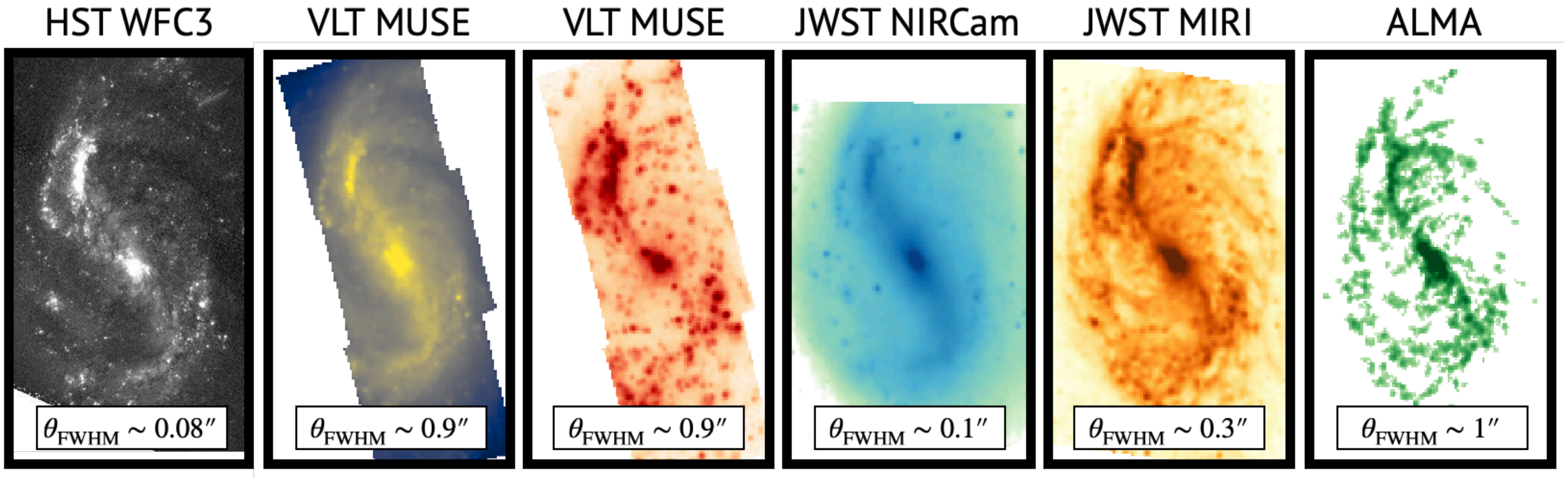


Hi ChatGPT! The following dataset includes different properties measured from multi-wavelength observations of nearby galaxies as part of the PHANGS survey. Each row represents an object (specifically, a 150-pc region in one of the galaxies), and each column represents different measured features. Can you help me to find something new and exciting in this dataset?

Great! I'll start by loading and inspecting the dataset to understand its structure and contents. Let's see what the data looks like.

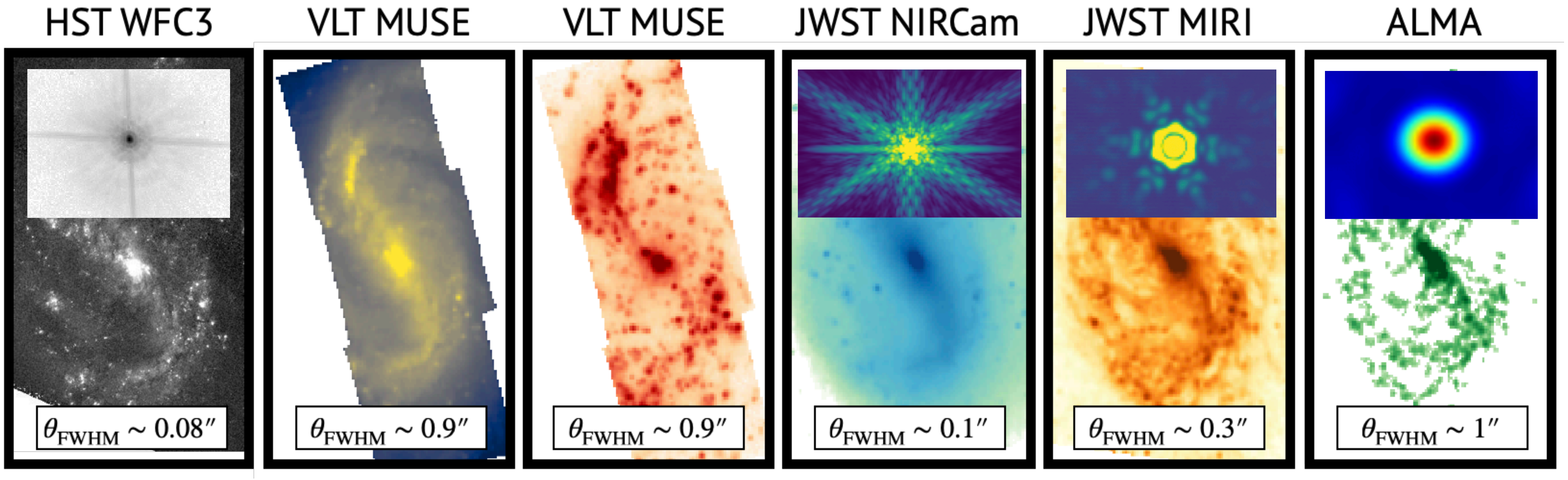
Expediting Astronomical Discovery with Large Language Models: Progress, Challenges, and Future Directions by Y. Sen-Ting (Thursday)

# Discovery with unsupervised learning: challenges/considerations


HST WFC3
VLT MUSE
VLT MUSE
JWST NIRCam
JWST MIRI
ALMA
$\theta_{\rm FWHM} \sim 0.08''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.1''$
$\theta_{\rm FWHM} \sim 0.3''$
$\theta_{\rm FWHM} \sim 1''$

**Data standardization**

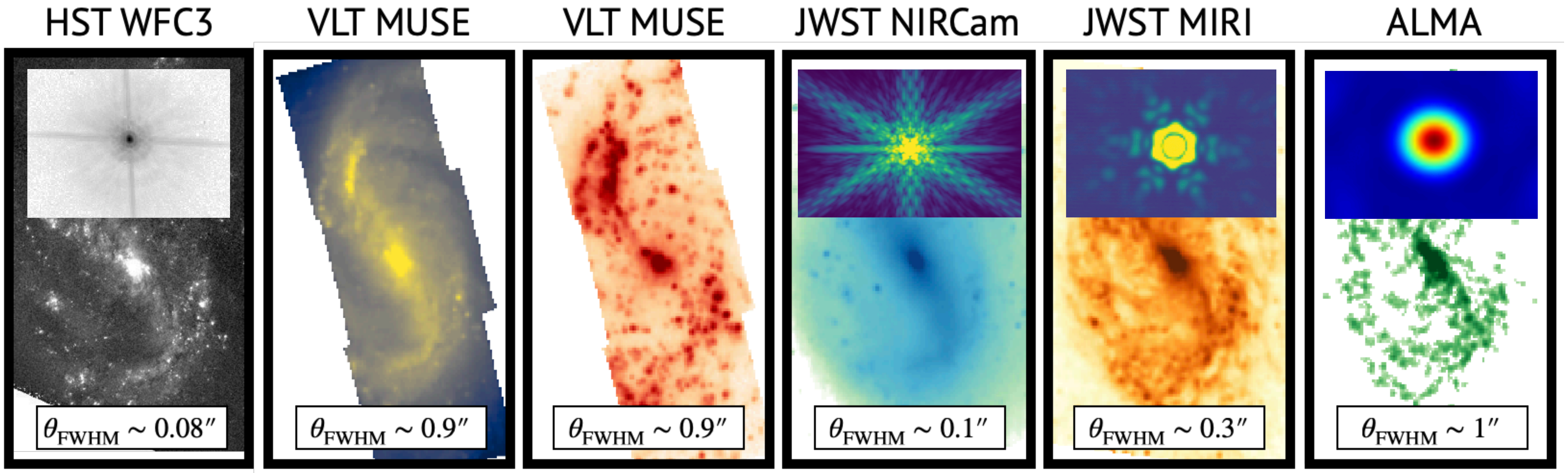
# Discovery with unsupervised learning: challenges/considerations


HST WFC3
VLT MUSE
VLT MUSE
JWST NIRCam
JWST MIRI
ALMA
$\theta_{\mathrm{FWHM}} \sim 0.08''$
$\theta_{\mathrm{FWHM}} \sim 0.9''$
$\theta_{\mathrm{FWHM}} \sim 0.9''$
$\theta_{\mathrm{FWHM}} \sim 0.1''$
$\theta_{\mathrm{FWHM}} \sim 0.3''$
$\theta_{\mathrm{FWHM}} \sim 1''$

## Data standardization

The most significant limitation in our ability to apply unsupervised learning algorithms to multi-wavelength datasets fused from different surveys. Different instruments have different resolutions, noise properties, FOVs, etc. We need to make sure that the data traces the same thing.

# Discovery with unsupervised learning: challenges/considerations

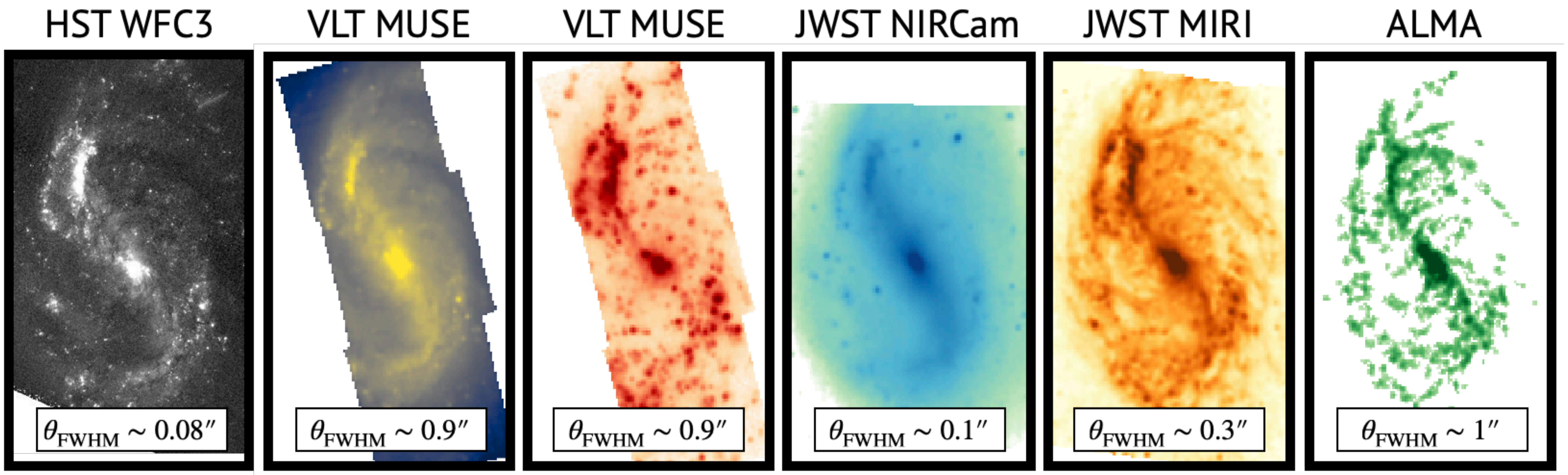

HST WFC3
VLT MUSE
VLT MUSE
JWST NIRCam
JWST MIRI
ALMA
$\theta_{\rm FWHM} \sim 0.08''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.1''$
$\theta_{\rm FWHM} \sim 0.3''$
$\theta_{\rm FWHM} \sim 1''$

## Data standardization

The most significant limitation in our ability to apply unsupervised learning algorithms to multi-wavelength datasets fused from different surveys. Different instruments have different resolutions, noise properties, FOVs, etc. We need to make sure that the data traces the same thing.

Surveys are doing a much better job in making sure that the data can be standardized wrt other surveys, either during the survey planning or post-survey. Examples: SDSS, PHANGS, the deep fields.

# Discovery with unsupervised learning: challenges/considerations


HST WFC3
VLT MUSE
VLT MUSE
JWST NIRCam
JWST MIRI
ALMA
$\theta_{\rm FWHM} \sim 0.08''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.9''$
$\theta_{\rm FWHM} \sim 0.1''$
$\theta_{\rm FWHM} \sim 0.3''$
$\theta_{\rm FWHM} \sim 1''$

## Data standardization

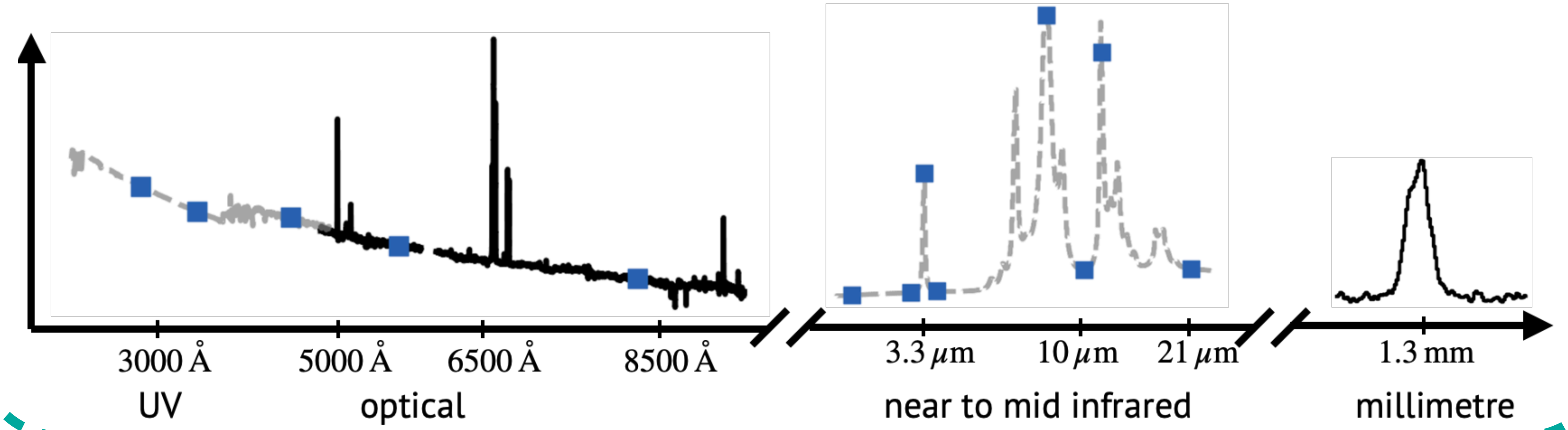
The most significant limitation in our ability to apply unsupervised learning algorithms to multi-wavelength datasets fused from different surveys. Different instruments have different resolutions, noise properties, FOVs, etc. We need to make sure that the data traces the same thing.

Surveys are doing a much better job in making sure that the data can be standardized wrt other surveys, either during the survey planning or post-survey. Examples: SDSS, PHANGS, the deep fields

Later today you will hear: “Machine learning for star parametrization” by A. Turchi, that will present “The Survey of Surveys”, homogenizing products from APOGEE, GALAH, Gaia-ESO, LAMOST, and RAVE, for radial velocity data.

# Discovery with unsupervised learning: challenges/considerations

input data for example: ~100,000 objects with the following information:



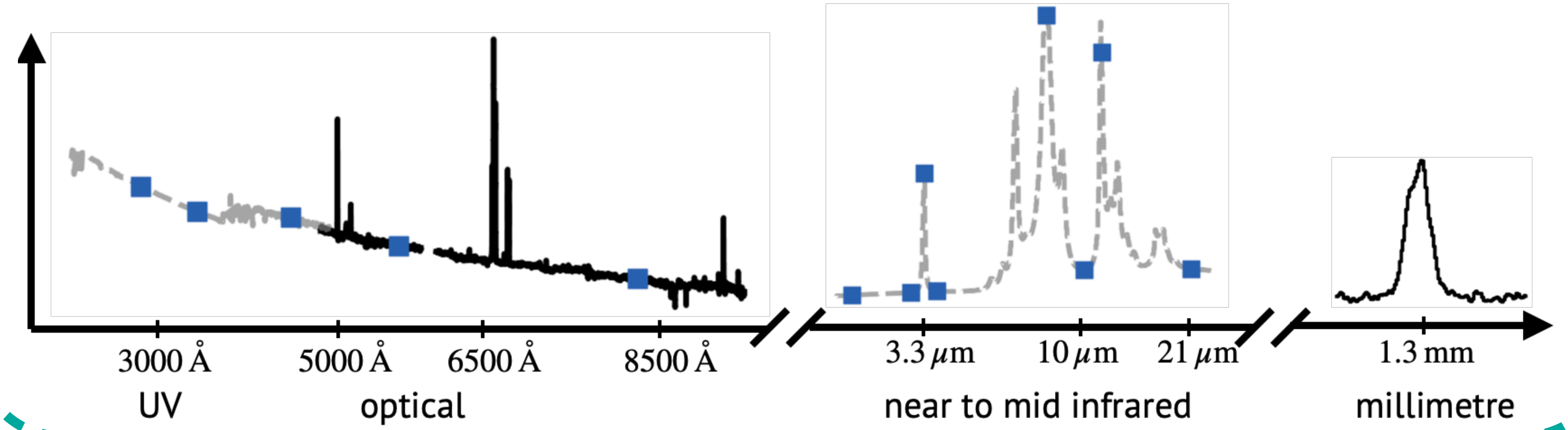
## What aspects of the data are more important?

Manual definition of features of interest by a domain expert

Apply deep networks to the raw data, letting the machine to define features

# Discovery with unsupervised learning: challenges/considerations

input data for example: ~100,000 objects with the following information:



## What aspects of the data are more important?

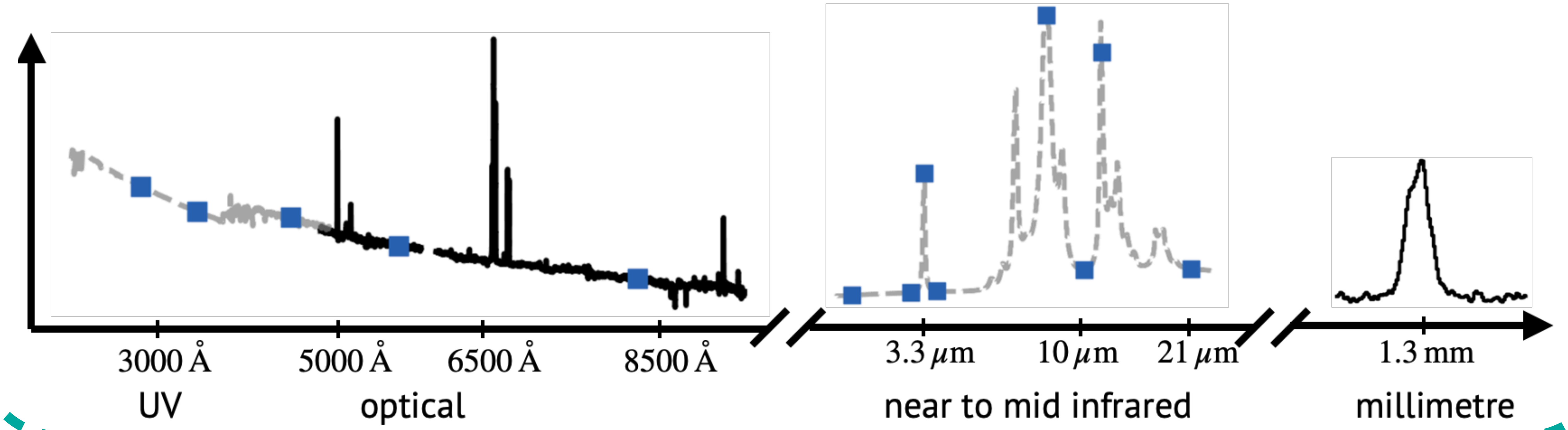
Manual definition of features of interest by a domain expert

**Advantage:** we define what features are of interest in the exploration process.

Apply deep networks to the raw data, letting the machine to define features

# Discovery with unsupervised learning: challenges/considerations

input data for example: ~100,000 objects with the following information:



## What aspects of the data are more important?

### Manual definition of features of interest by a domain expert

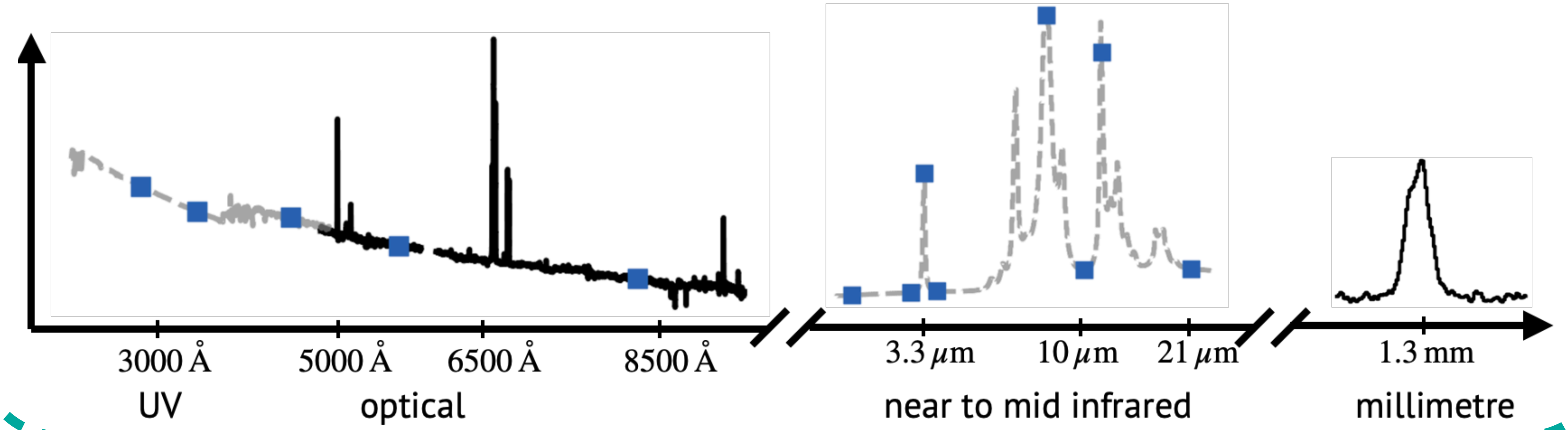
**Advantage:** we define what features are of interest in the exploration process.

**Disadvantage:** we define what features are of interest in the exploration process.

### Apply deep networks to the raw data, letting the machine to define features

# Discovery with unsupervised learning: challenges/considerations

input data for example: ~100,000 objects with the following information:



## What aspects of the data are more important?

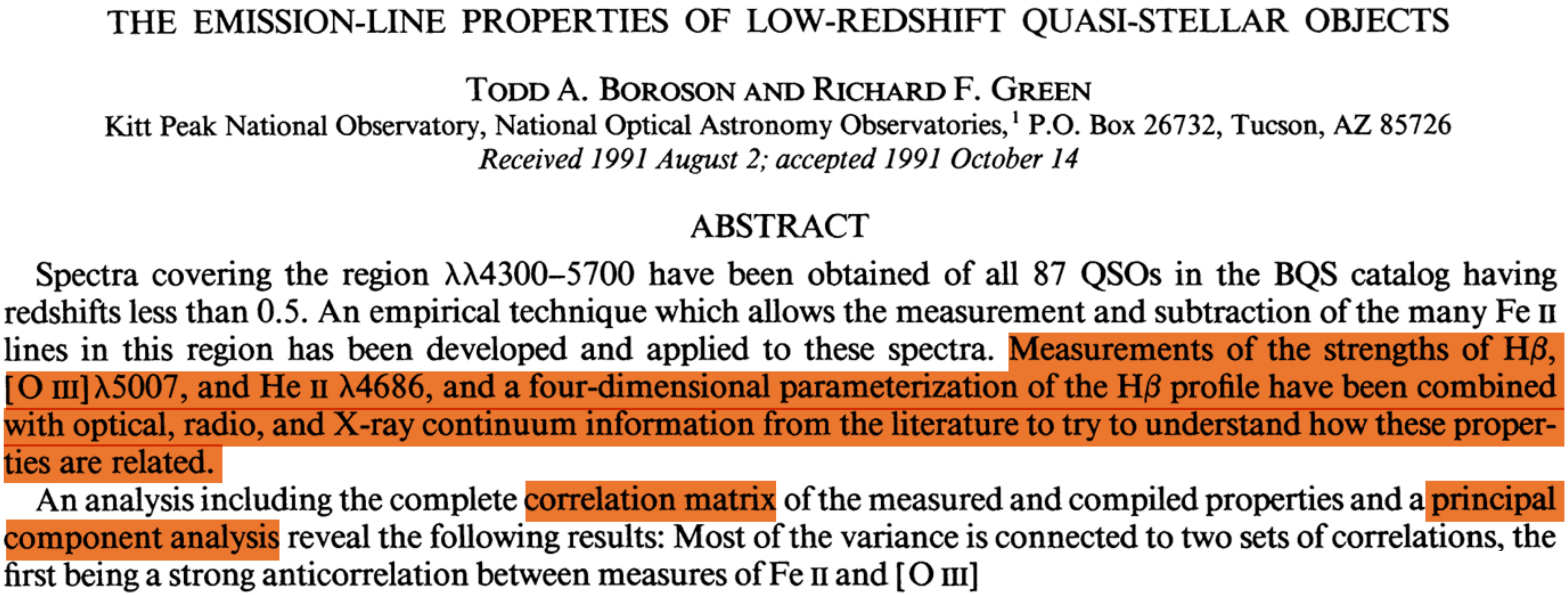
Manual definition of features of interest by a domain expert

**Advantage:** we define what features are of interest in the exploration process.

**Disadvantage:** we define what features are of interest in the exploration process.

**Important aspects to consider:** how to model missing values and upper limits?

Apply deep networks to the raw data, letting the machine to define features



# THE EMISSION-LINE PROPERTIES OF LOW-REDSHIFT QUASI-STELLAR OBJECTS

TODD A. BOROSON AND RICHARD F. GREEN
Kitt Peak National Observatory, National Optical Astronomy Observatories,[1] P.O. Box 26732, Tucson, AZ 85726

*Received 1991 August 2; accepted 1991 October 14*

## ABSTRACT

Spectra covering the region λλ4300–5700 have been obtained of all 87 QSOs in the BQS catalog having redshifts less than 0.5. An empirical technique which allows the measurement and subtraction of the many Fe II lines in this region has been developed and applied to these spectra. Measurements of the strengths of Hβ, [O III]λ5007, and He II λ4686, and a four-dimensional parameterization of the Hβ profile have been combined with optical, radio, and X-ray continuum information from the literature to try to understand how these properties are related.

An analysis including the complete correlation matrix of the measured and compiled properties and a principal component analysis reveal the following results: Most of the variance is connected to two sets of correlations, the first being a strong anticorrelation between measures of Fe II and [O III]

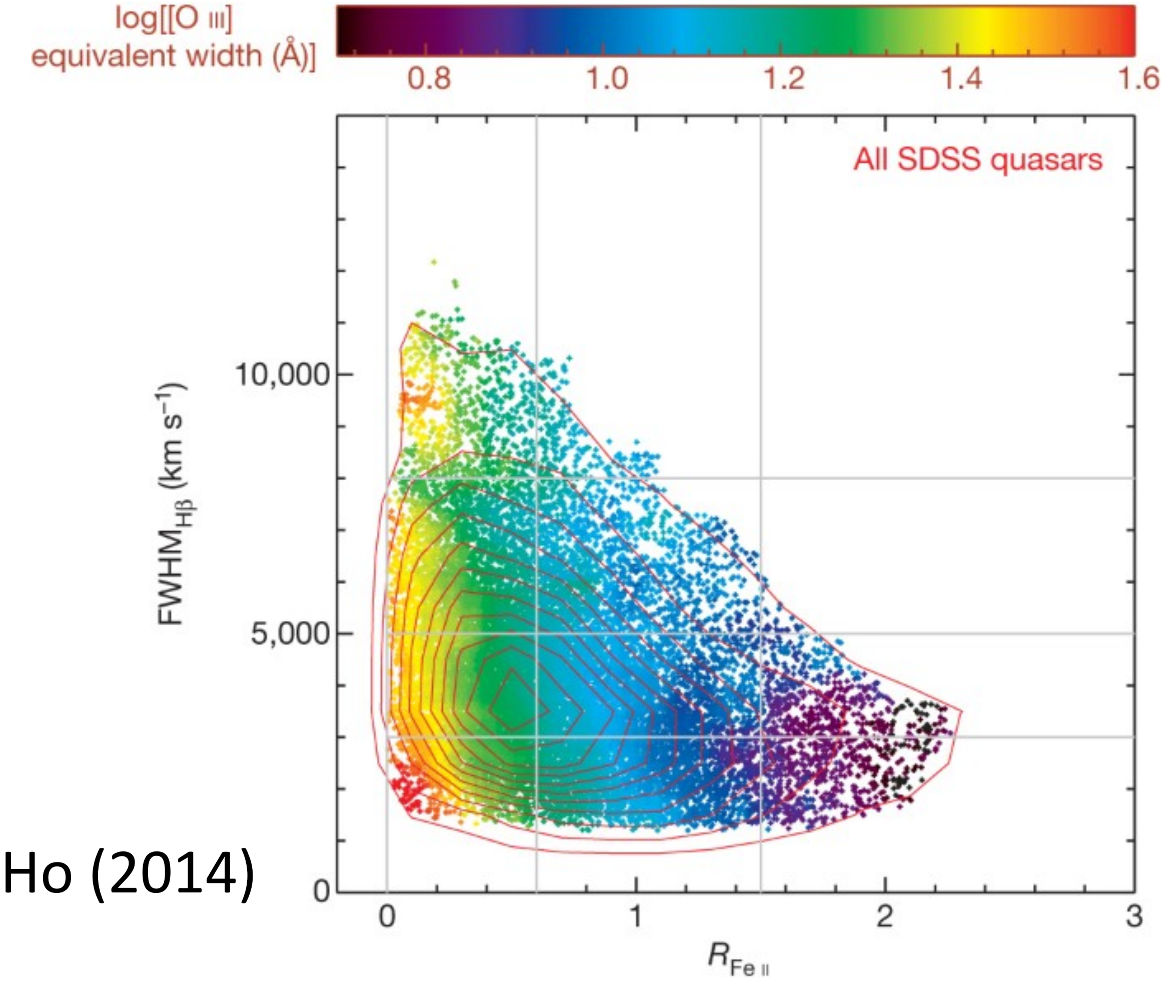
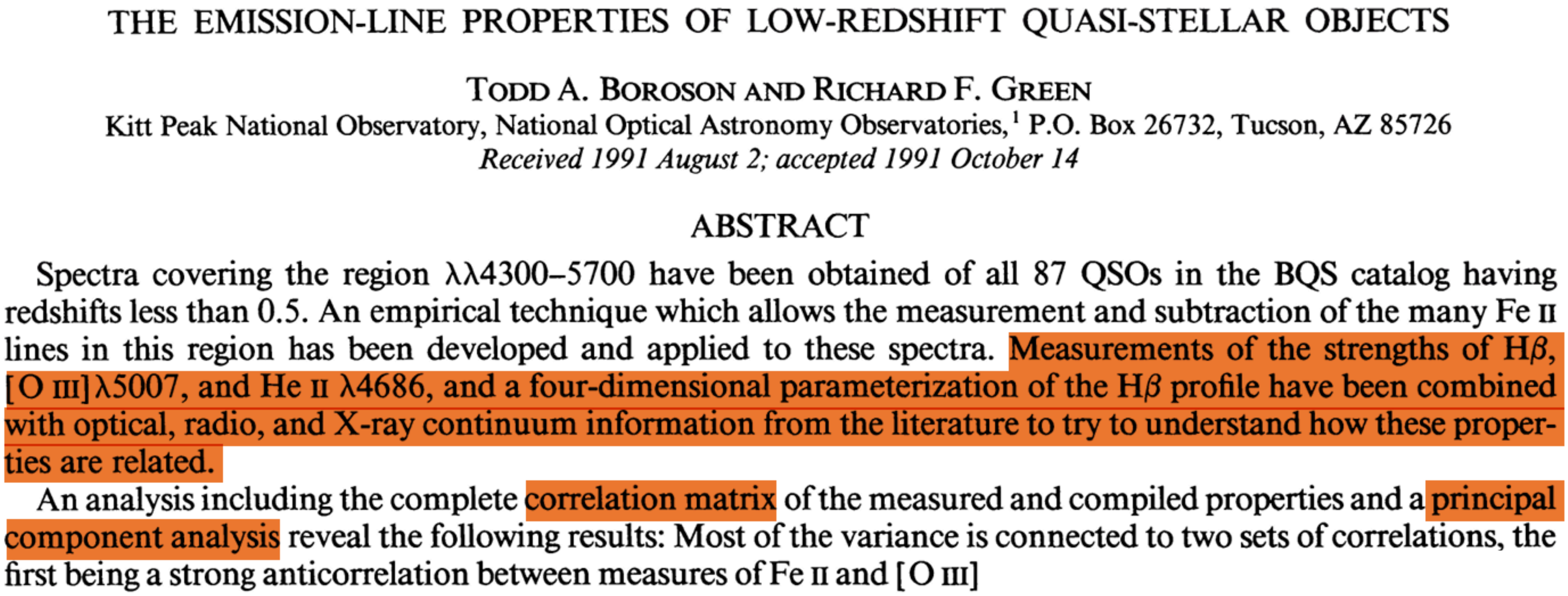


Figure from Shen & Ho (2014)



THE EMISSION-LINE PROPERTIES OF LOW-REDSHIFT QUASI-STELLAR OBJECTS

TODD A. BOROSON AND RICHARD F. GREEN
Kitt Peak National Observatory, National Optical Astronomy Observatories,[1] P.O. Box 26732, Tucson, AZ 85726

*Received 1991 August 2; accepted 1991 October 14*

ABSTRACT

Spectra covering the region λλ4300–5700 have been obtained of all 87 QSOs in the BQS catalog having redshifts less than 0.5. An empirical technique which allows the measurement and subtraction of the many Fe II lines in this region has been developed and applied to these spectra. Measurements of the strengths of H$\beta$, [O III] λ5007, and He II λ4686, and a four-dimensional parameterization of the H$\beta$ profile have been combined with optical, radio, and X-ray continuum information from the literature to try to understand how these properties are related.

An analysis including the complete correlation matrix of the measured and compiled properties and a principal component analysis reveal the following results: Most of the variance is connected to two sets of correlations, the first being a strong anticorrelation between measures of Fe II and [O III]

Predictive Modeling of Galaxy Spectra: A Comprehensive Framework Using Transformer Architecture, by G. Martínez Solaeche (Tuesday)

Predicting AGN Obscuration with Machine Learning by R. Silver (Wednesday)

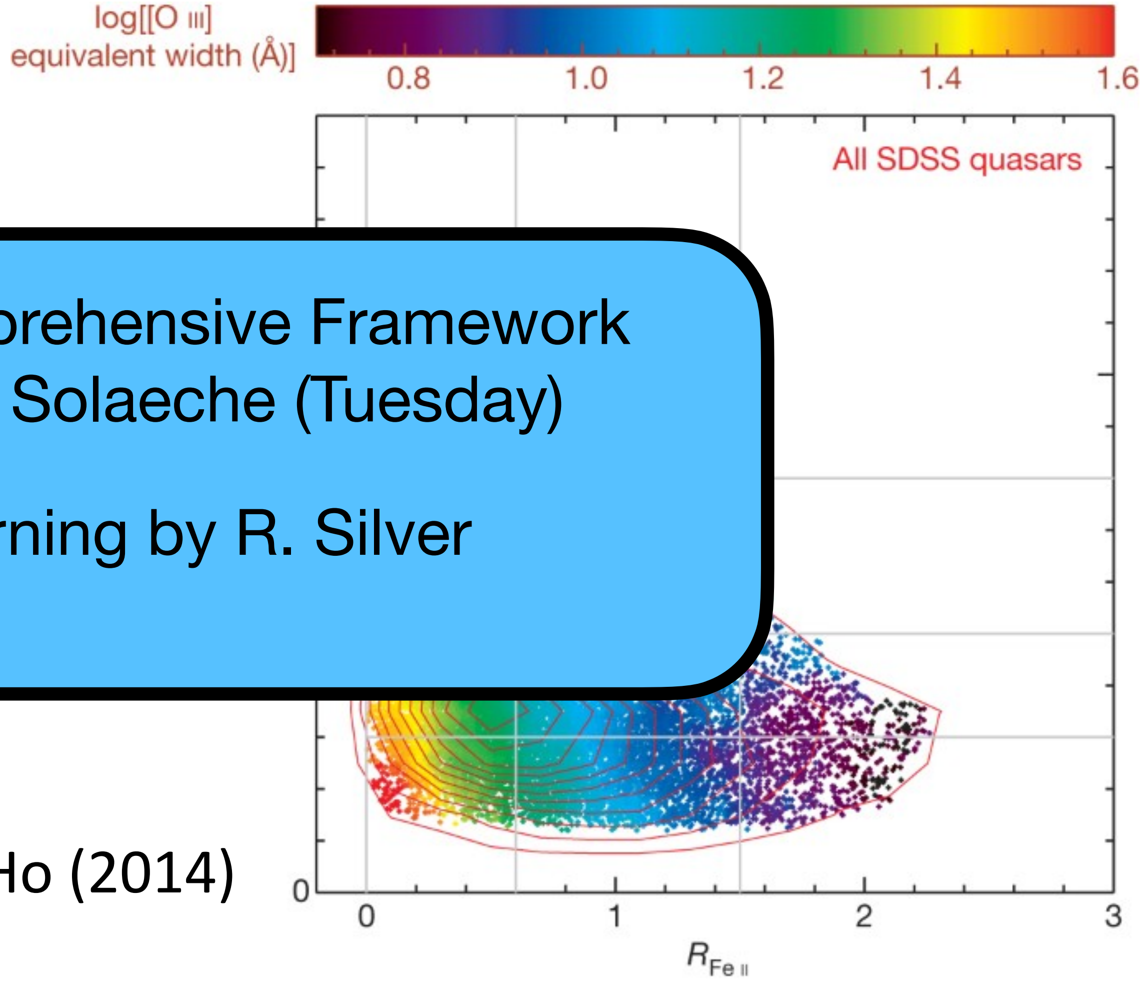
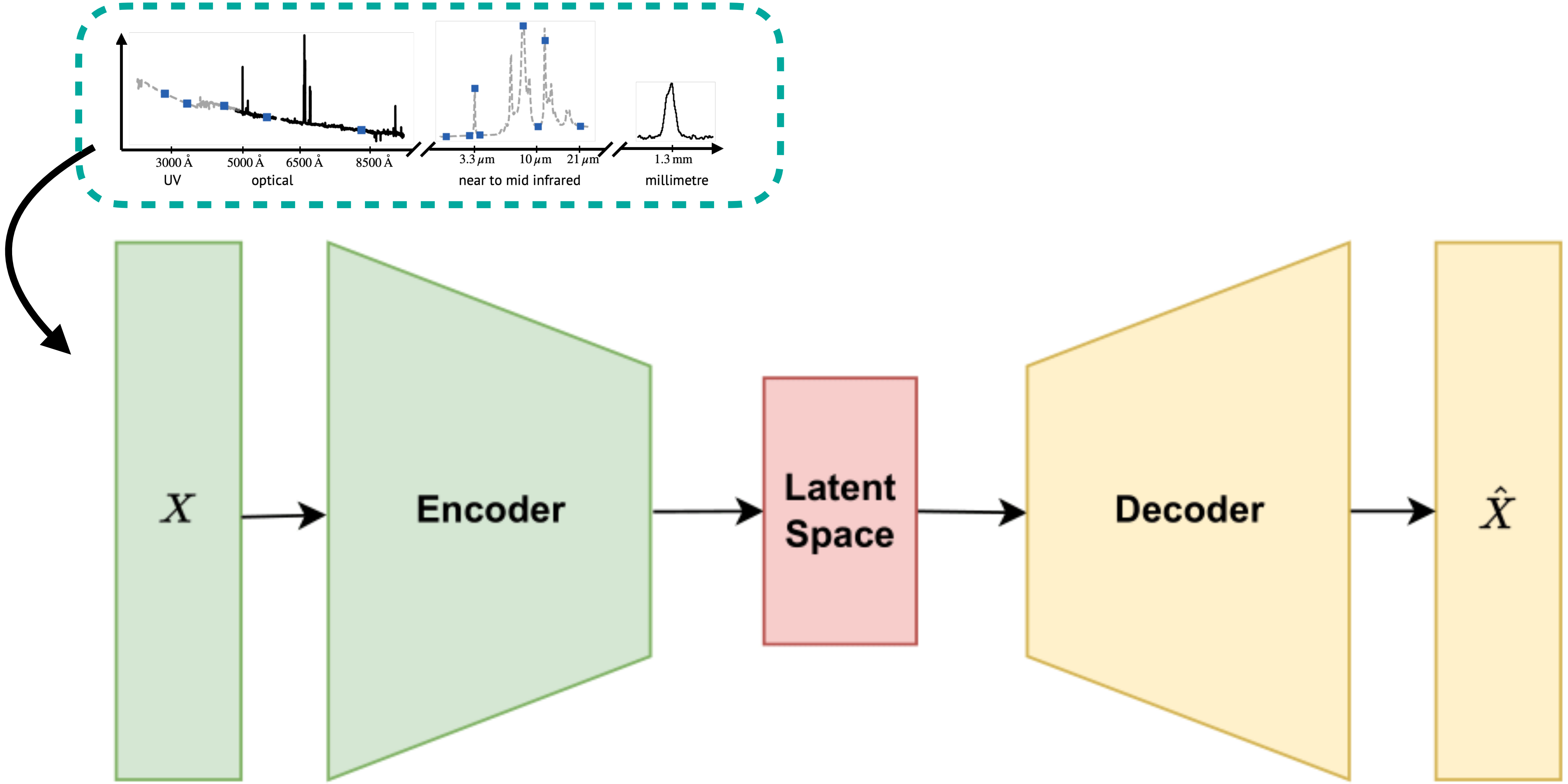


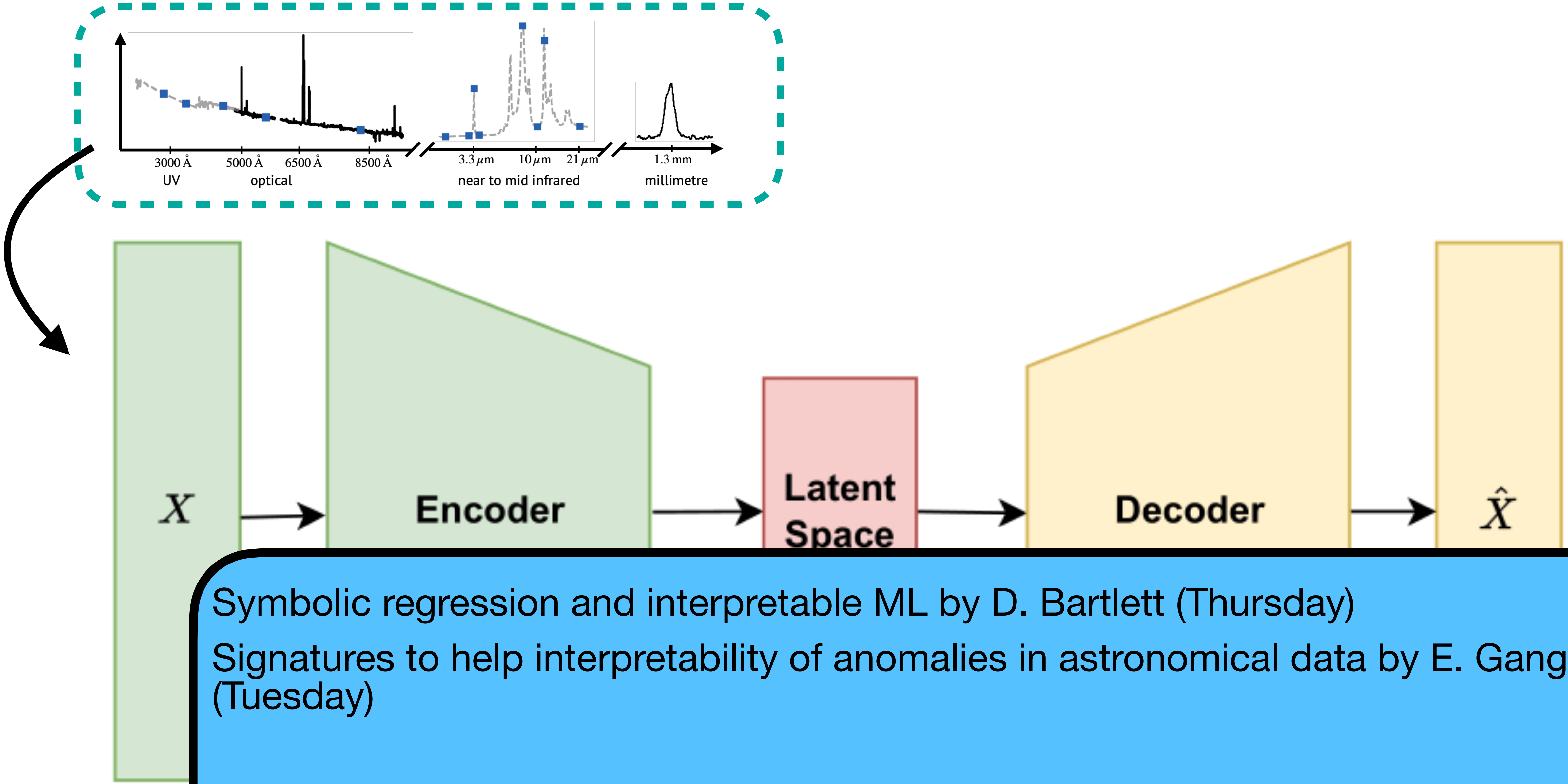
Figure from Shen & Ho (2014)

# Letting the machine derive the features: interpretability



The ability to interpret the resulting low-D embedding is of fundamental importance when using these tools to make scientific discoveries.

# Letting the machine derive the features: interpretability



Symbolic regression and interpretable ML by D. Bartlett (Thursday)

Signatures to help interpretability of anomalies in astronomical data by E. Gangler (Tuesday)

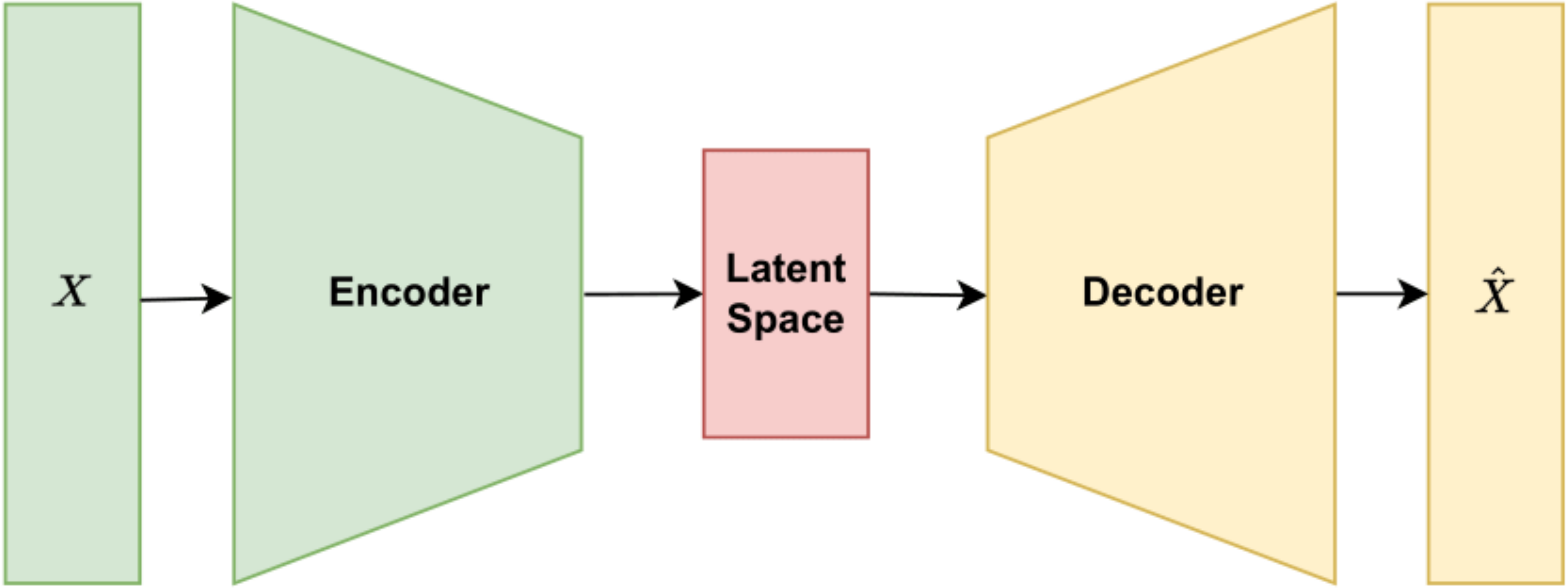
Multi-Task Neural Nets with Monte Carlo Dropout for Spectral Analysis of Galaxies by M. Ginolfi (Tuesday)

Inferring Galaxy Baryonic Properties from IllustrisTNG Dark Matter Merger Trees with Graph Neural Networks by N. Andreadis (Wednesday)

Interpreting the largest cosmological simulations using Representation Learning by S. Trujillo (Thursday)

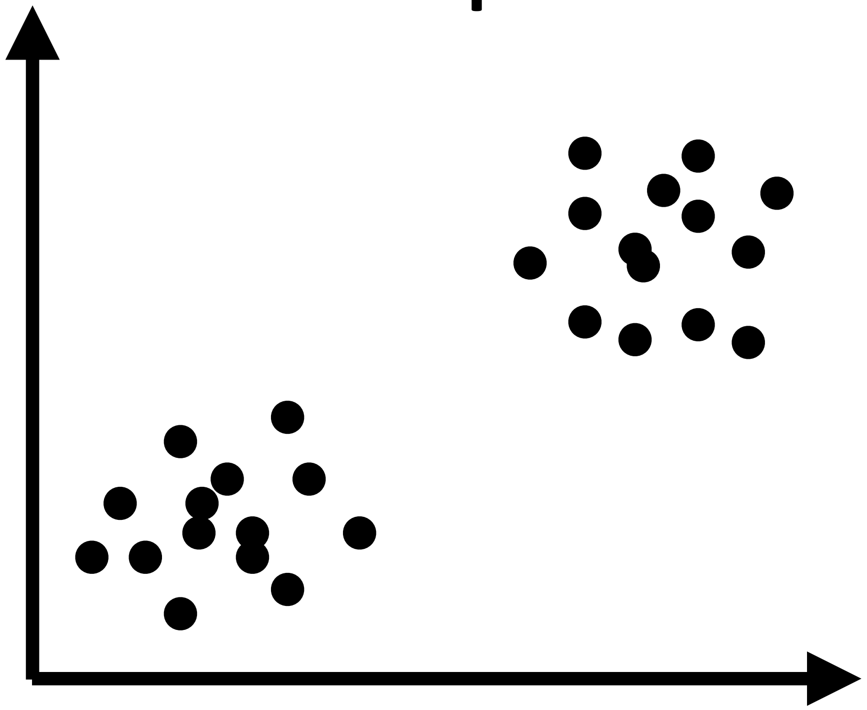
Machine Learning Based Parametrization of Solar Active Regions Using Disentangled Variational Autoencoders by E. Dineva (Friday)

# Letting the machine derive the features: interpretability

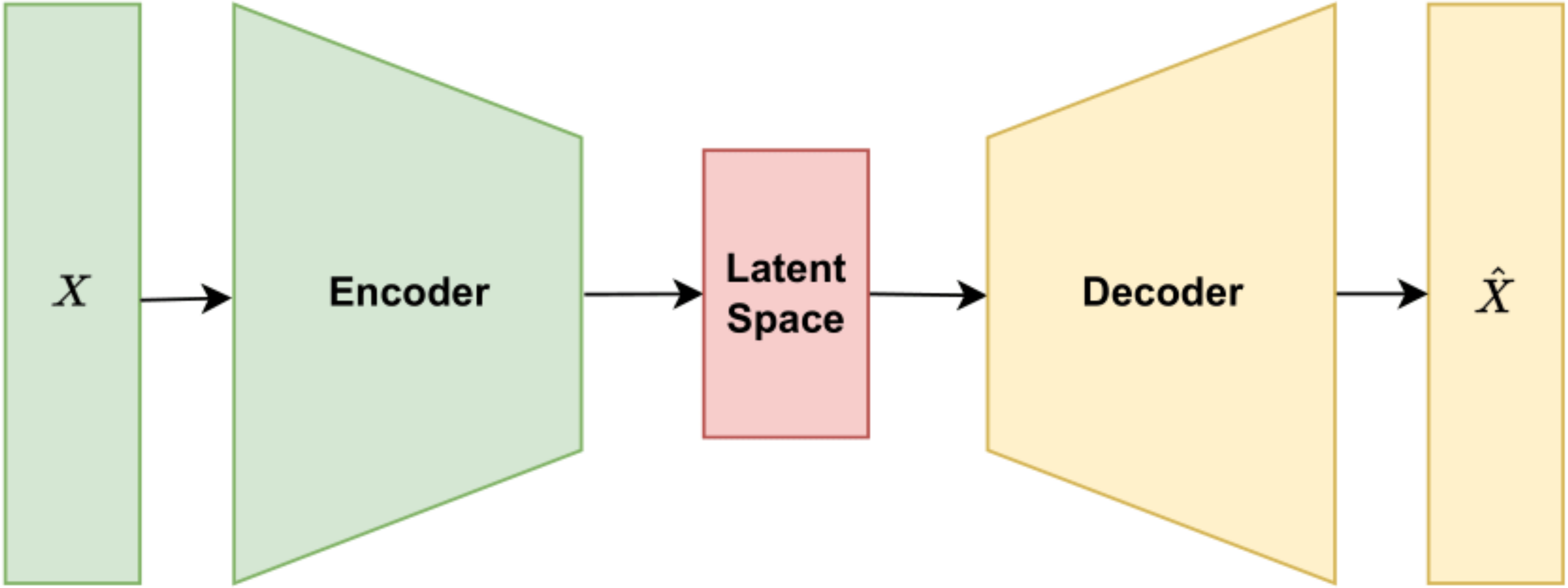


- It is not enough to identify which “features” the Encoder extracts.
- We need to understand the latent space and its properties:
  - Is the space Euclidean?
  - How to interpret short vs. long distances?
  - How to interpret different densities?
  - How to interpret straight vs. curved lines?

Latent space:



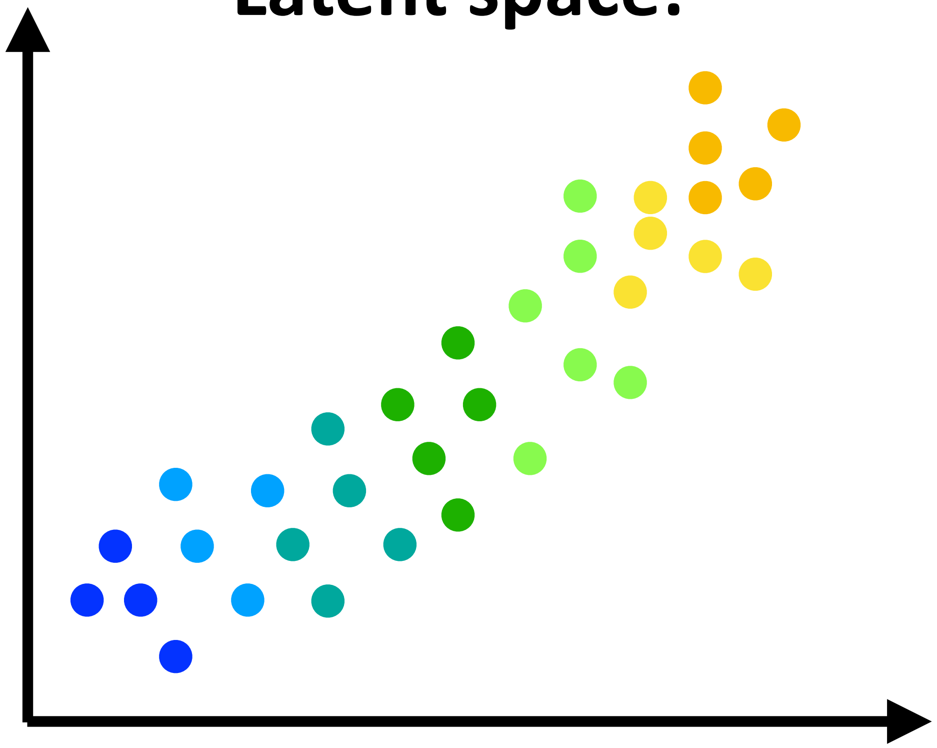
# Letting the machine derive the features: interpretability



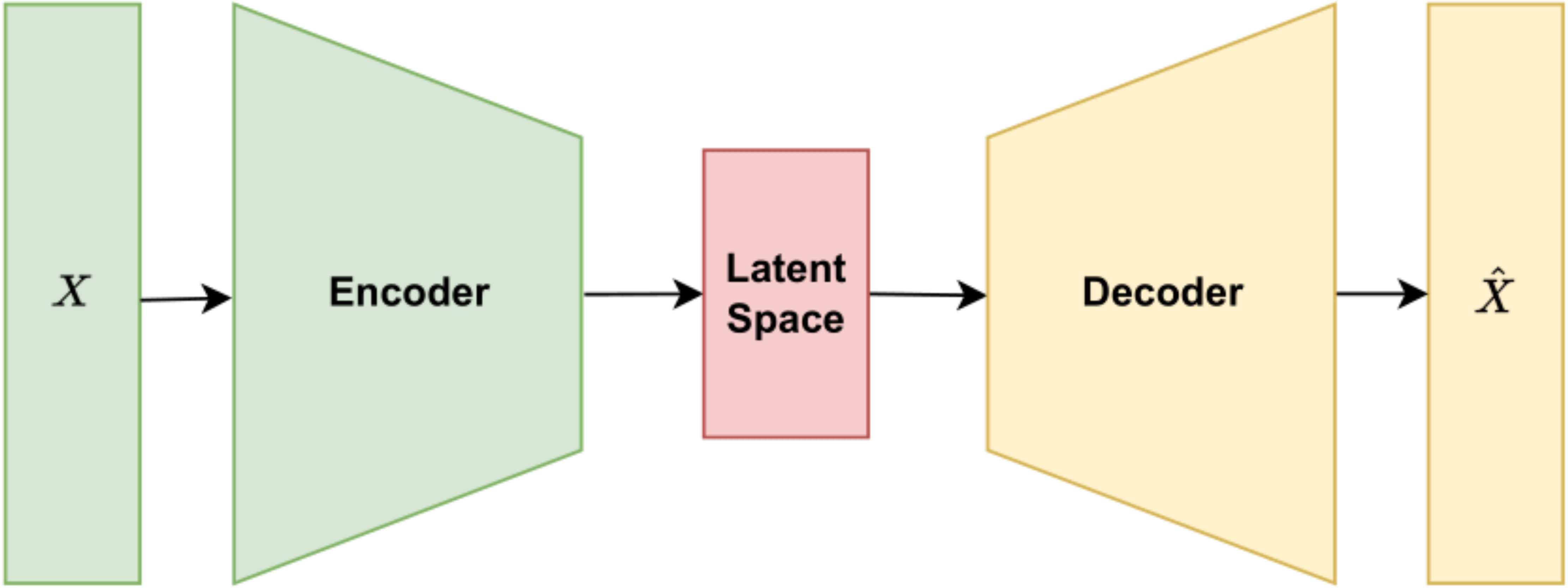
What drives the variation we see in the latent space?

- Using derived features to color-code points in the latent space: if our ability to interpret the results depends on a set of extracted features, then why not use derived features as the input in the first place?
- Plotting the “average” object in each bin in the latent space (like SOMs do) could be useful.

Latent space:



# Letting the machine derive the features: controlling the output



More often than not, the first representation will be dominated by:

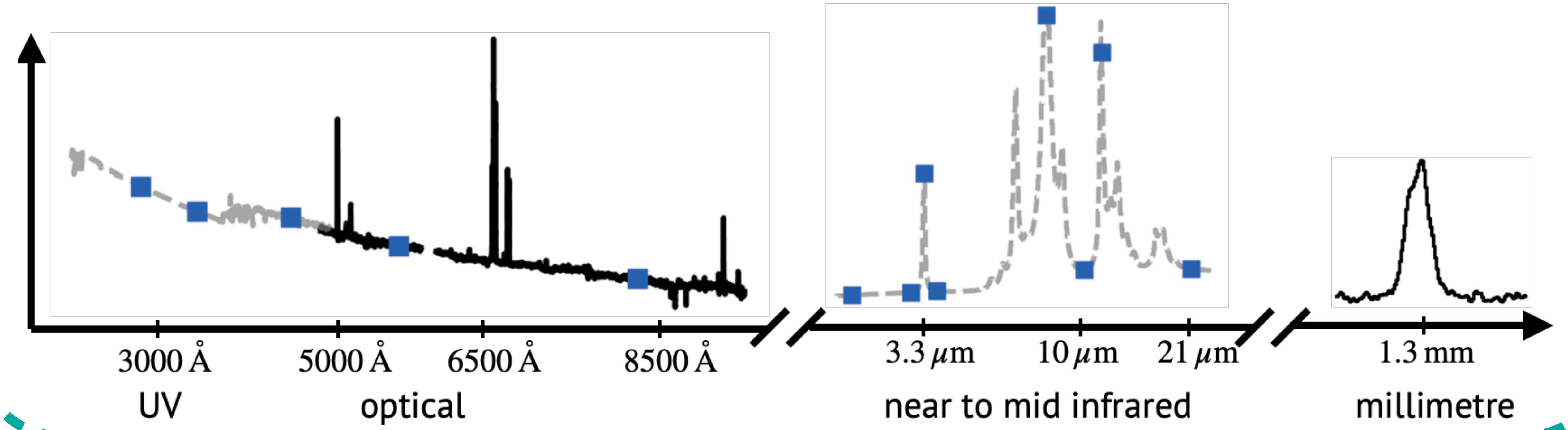
(i) Systematics; noisy objects; or objects that do not belong in the dataset.

(ii) Trends and/or group we already know: redshift; brightness/luminosity; the galaxy bimodality, SFR-stellar mass relation, SFR-molecular gas mass relation, etc.

We want to be able to go beyond that, either by having high-enough dimensions in the latent space, or by pre-processing the input data.

# Discovery with unsupervised learning: challenges/considerations

input data for example: ~100,000 objects with the following information:



## What aspects of the data are more important?

### Manual definition of features of interest by a domain expert

**Advantage:** we define what features are of interest in the exploration process.

**Disadvantage:** we define what features are of interest in the exploration process.

**Important aspects to consider:** how to model missing values and upper limits?

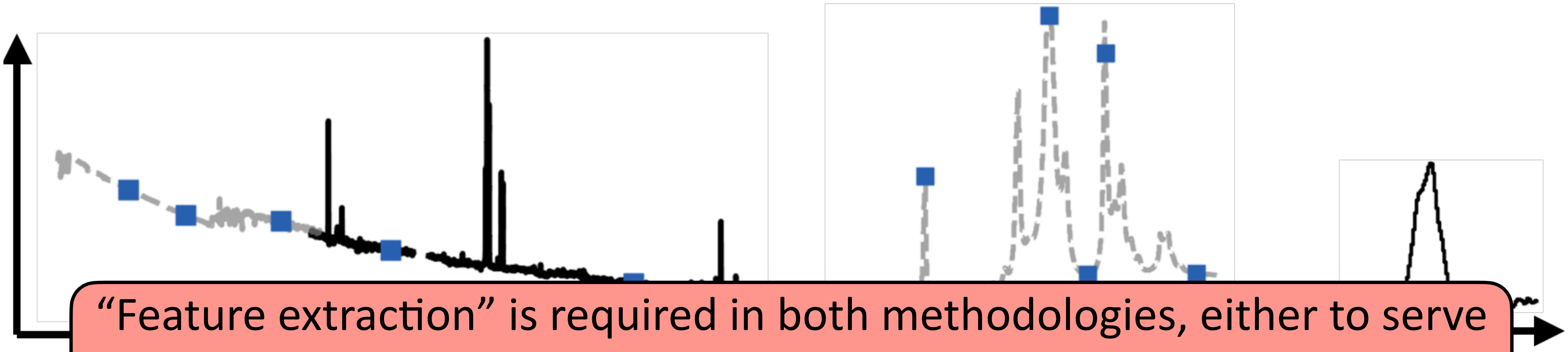
### Apply deep networks to the raw data, letting the machine to define features

**Interpretability:** we must be able to explain the latent space and the representation.

**Controllability:** we want to be able to control the output, and go beyond what we already know.

# Discovery with unsupervised learning: challenges/considerations

input data for example: ~100,000 objects with the following information:



“Feature extraction” is required in both methodologies, either to serve as the input data, or later, in the interpretation stage. In both cases, some back-and-forth is expected.

## What aspects of the data are more important?

### Manual definition of features of interest by a domain expert

**Advantage:** we define what features are of interest in the exploration process.

**Disadvantage:** we define what features are of interest in the exploration process.

**Important aspects to consider:** how to model missing values and upper limits?

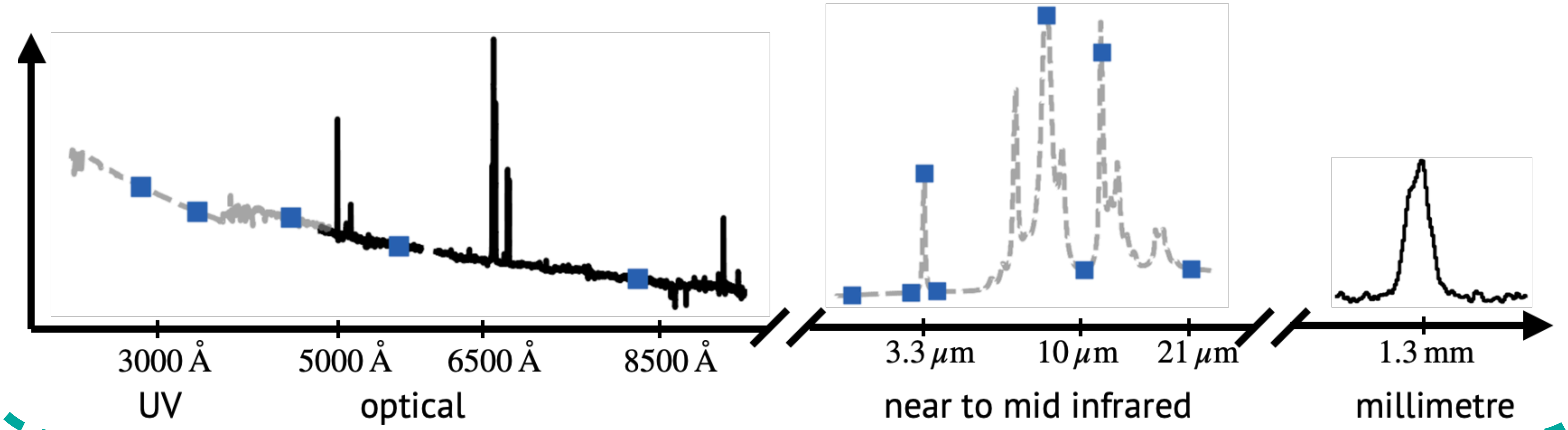
### Apply deep networks to the raw data, letting the machine to define features

**Interpretability:** we must be able to explain the latent space and the representation.

**Controllability:** we want to be able to control the output, and go beyond what we already know.

# Recent ML-assisted discovery in the PHANGS survey

input data for example: ~100,000 objects with the following information:

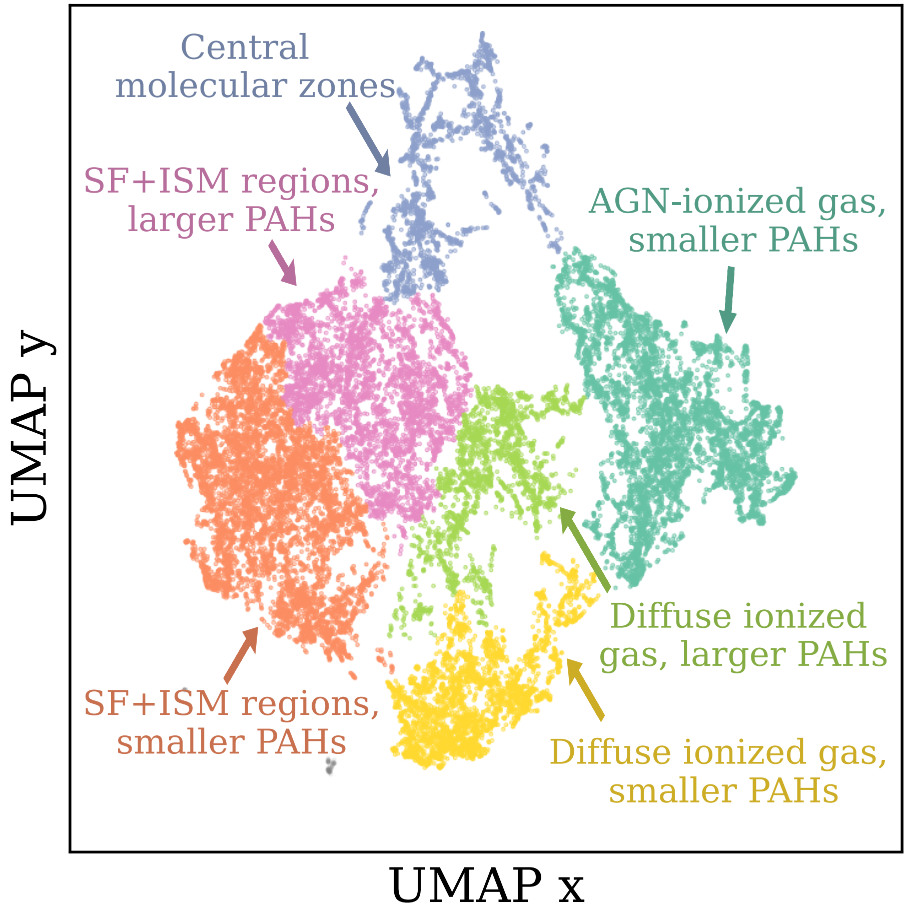


features of interest that trace multi-phase gas properties

$\mathrm{H}\alpha/\mathrm{CO}$
$\mathrm{H}\alpha/21\,\mu\mathrm{m}$
$\mathrm{CO}/21\,\mu\mathrm{m}$
$\mathrm{H}\alpha/\mathrm{PAH}\,7.7\,\mu\mathrm{m}$
$\mathrm{CO}/\mathrm{PAH}\,7.7\,\mu\mathrm{m}$

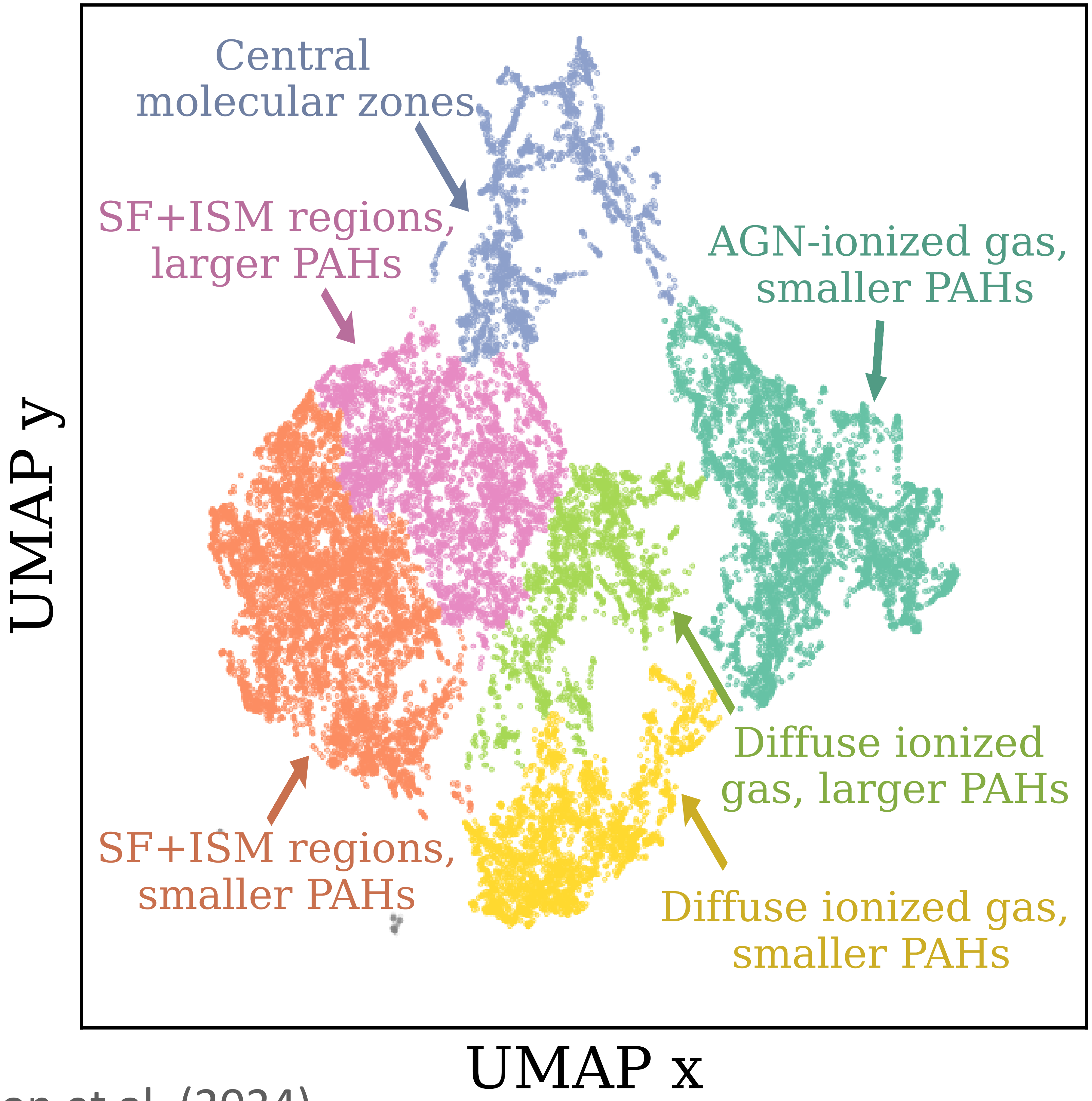
$[\mathrm{NII}]/\mathrm{H}\alpha$
$[\mathrm{SII}]/\mathrm{H}\alpha$
$[\mathrm{OI}]/\mathrm{H}\alpha$
$[\mathrm{OIII}]/\mathrm{H}\beta$

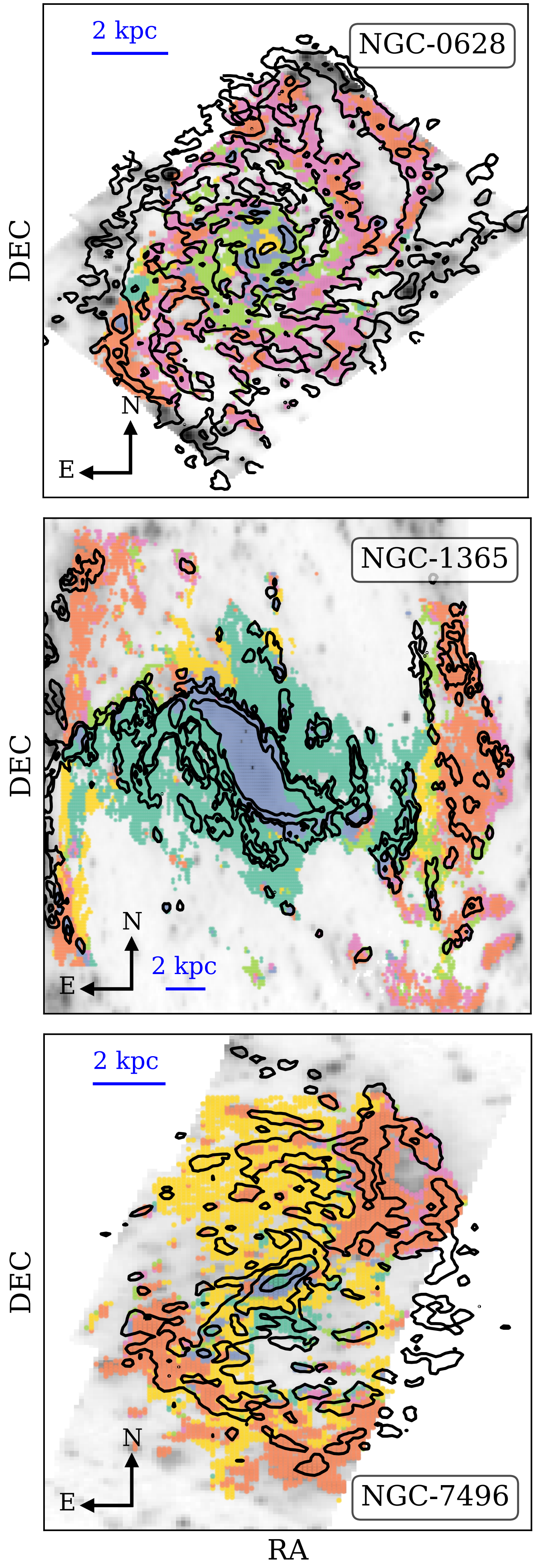
$\mathrm{PAH}\,11.3/7.7\,\mu\mathrm{m}$
$\mathrm{PAH}\,11.3/3.3\,\mu\mathrm{m}$



Baron et al. (2024)

# Six groups with distinct properties in their gas and dust properties:

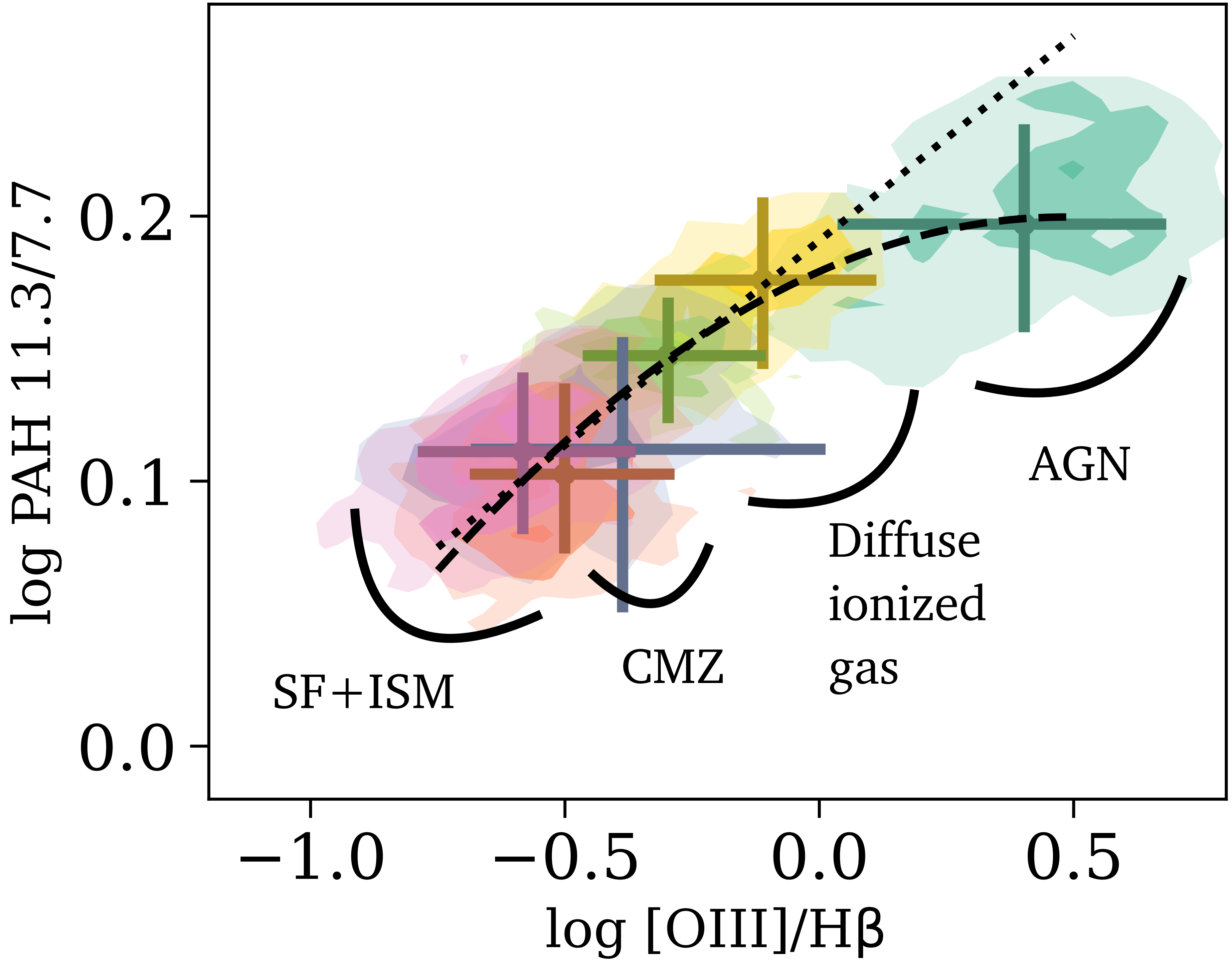




Baron et al. (2024)

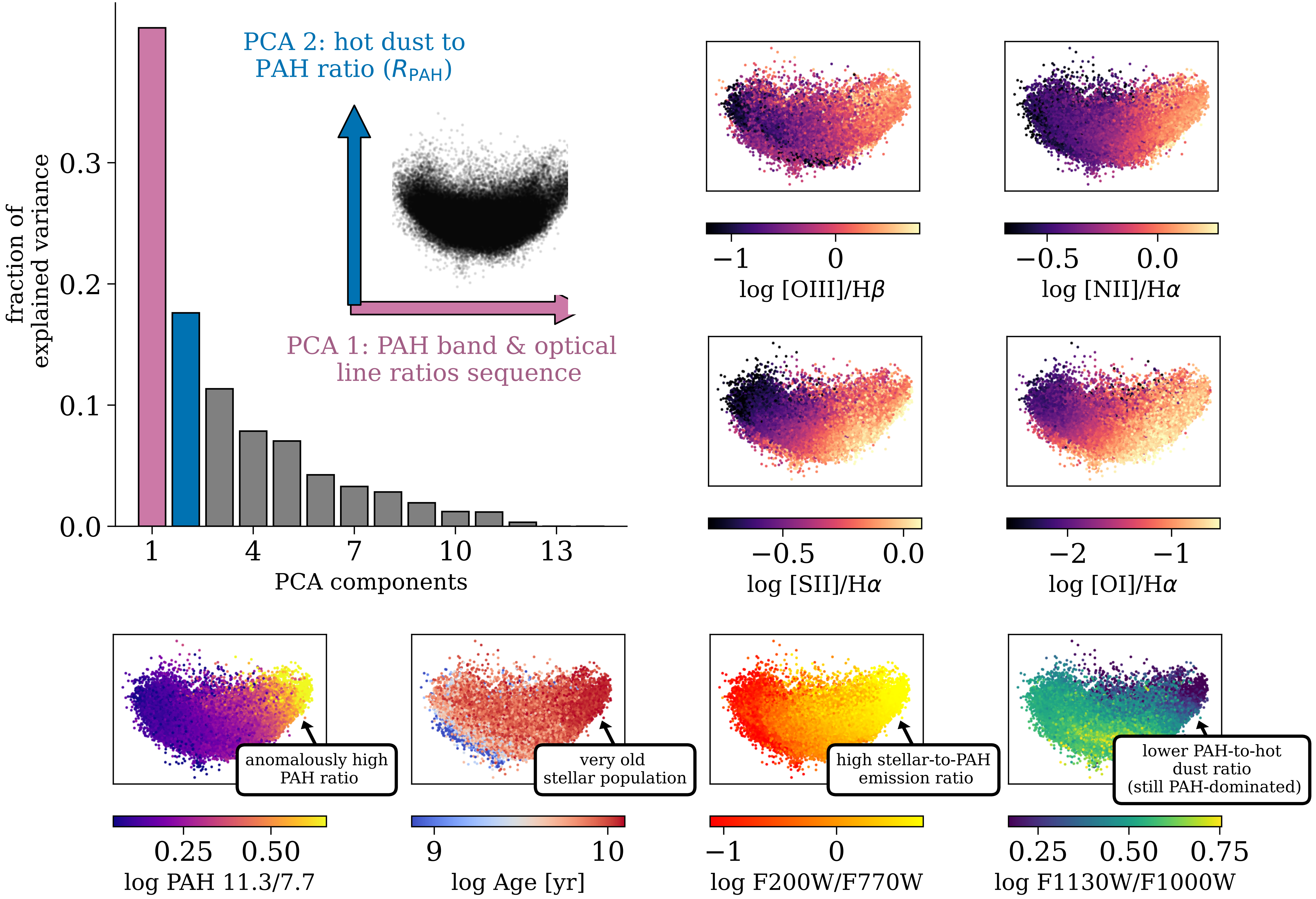
# A close connection between the heating of PAHs (residing in the neutral medium) and the ionization of the warm ionized gas

The correlations are essentially a sequence in gas & dust properties probed by the groups we found:

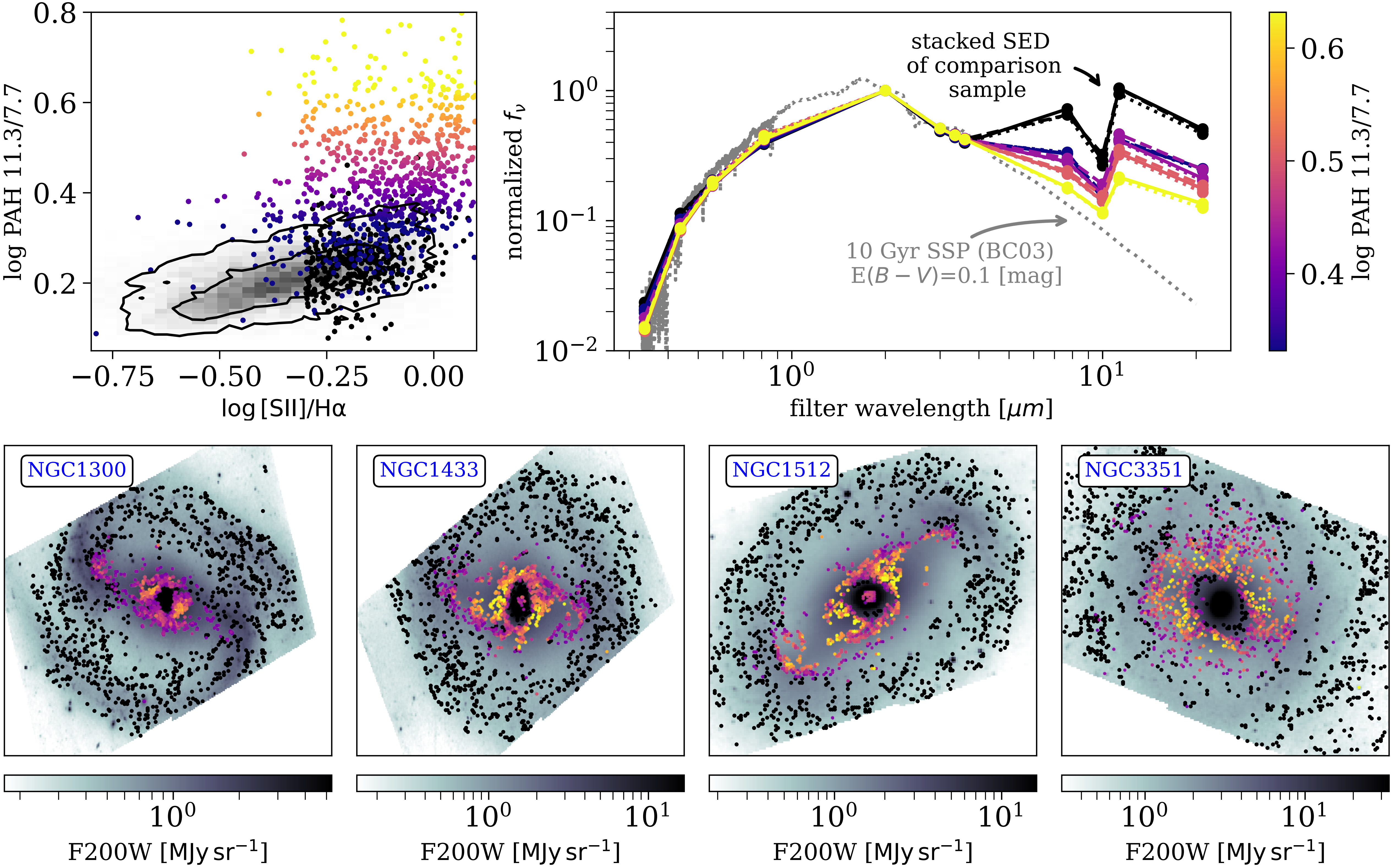


Baron et al. (2024)

# In a follow-up study of the full sample, PCA visualization reveals a group of anomalous pixels that do not follow the correlation

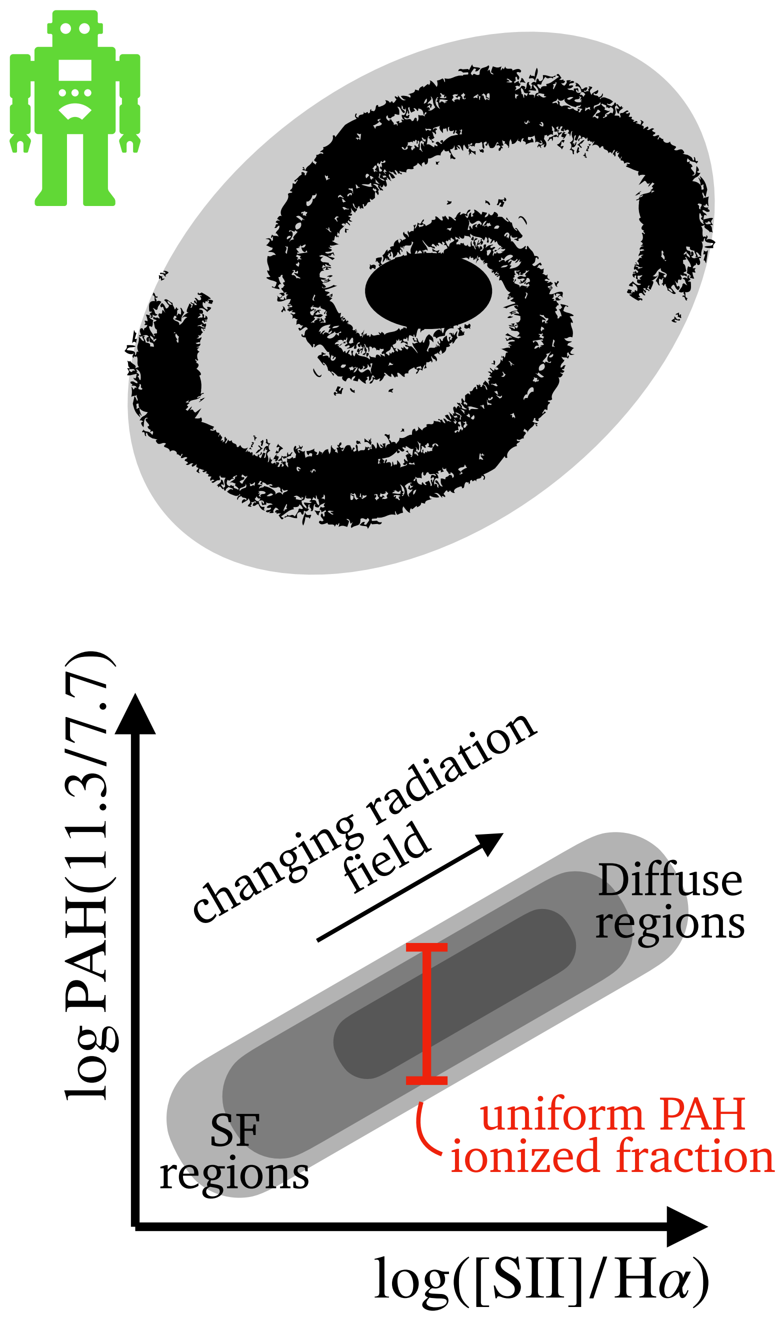


# In a follow-up study of the full sample, PCA visualization reveals a group of anomalous pixels that do not follow the correlation

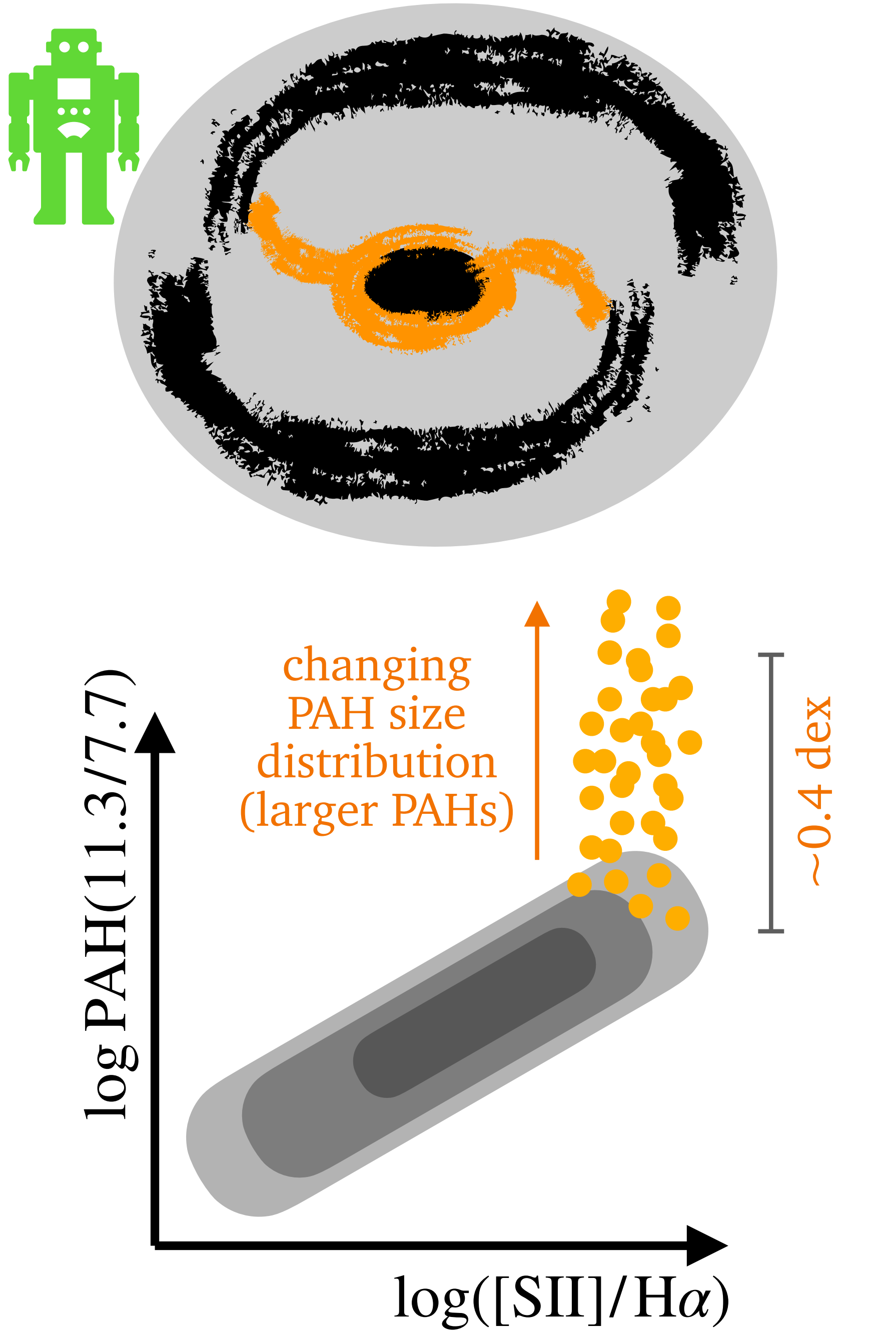


# Emerging picture of the connection between PAH heating and gas ionization

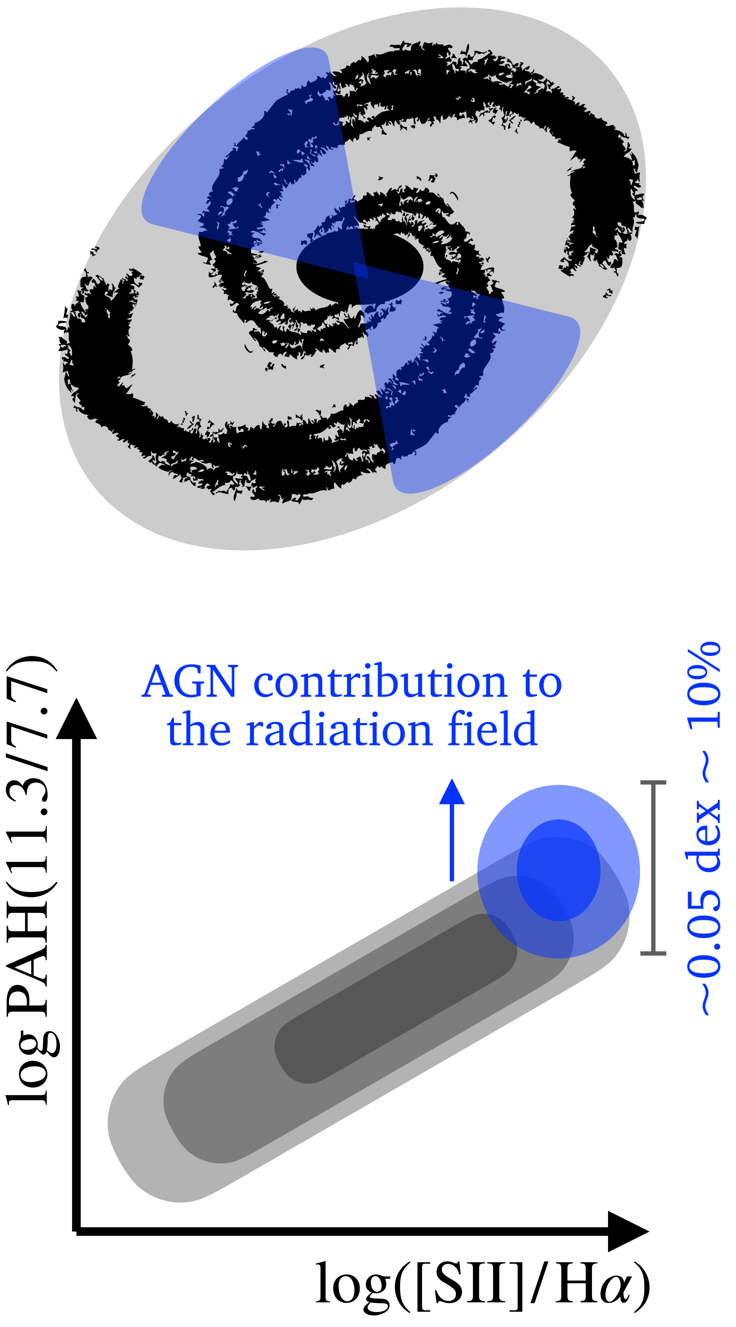
Typical Star-Forming Galaxies



Old and Bright Stars in the Bulge



Galaxies with AGN



Baron et al. (in prep.)

# Summary

In the Big Data era in astronomy, the abundance of multi-wavelength opportunities results in large and complex datasets. Machine Learning methods can be used to mine these information-rich datasets, facilitating new discoveries.

Physically-motivated features + simpler algorithms versus applying deep learning to the raw data: the two approaches have upsides and downsides, and I consider them complementary when exploring datasets for new trends.

Where can we find new multi-wavelength datasets to work with **NOW**?

- For stars: APOGEE-2 (near-infrared) & Gaia (astrometry) & TESS (RV) and more (e.g., Carrillo et al. 2020).
- For ISM: HI4PI (radio 21 cm) & LVM (optical IFU) & multi-wavelength photometry (here).
- Local galaxies: the PHANGS survey.
- JWST: NIRCam (near-infrared) and MIRI (mid-infrared) observations in ERS and Treasury programs (here and here). Usually with available multi-wavelength observations.

**Dalya Baron; dalyabaron@gmail.com**