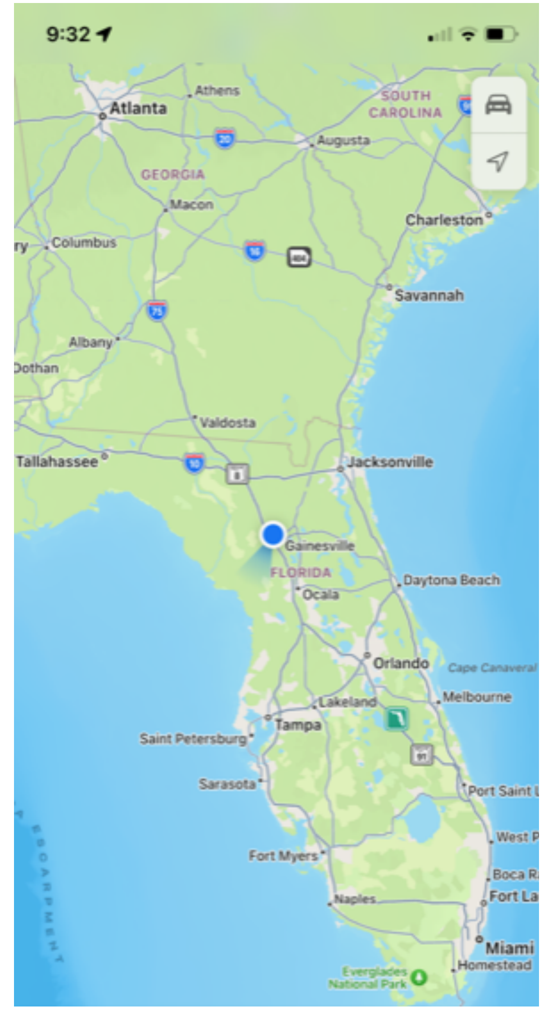
Machine Learning for Exoplanets (Unsupervised ML)

Katia Matcheva



University of Florida

ARES III School 09/11-16/2023

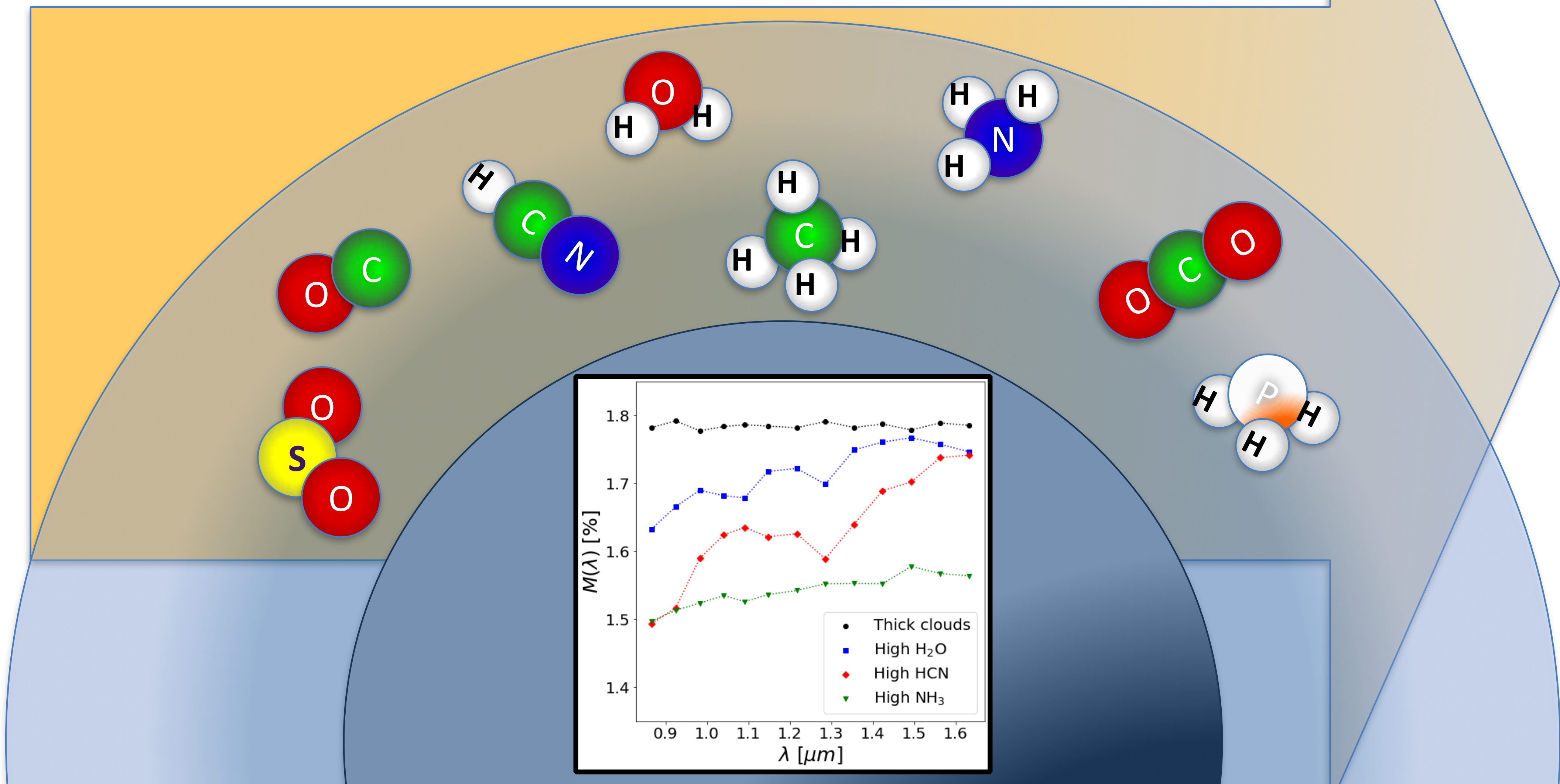
Outline

- The Role of ML
  - work with large volume of data
  - high dimensional data
  - speed up the simulation models
- ML Approaches
  - Unsupervised Learning
  - Supervised Learning
  - Ariel Data Challenge
- Searching for the unexpected
  - anomaly detection

SCAN ME

SCAN ME

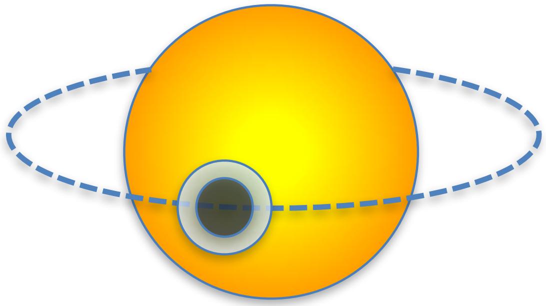
Transit Transmission Spectroscopy



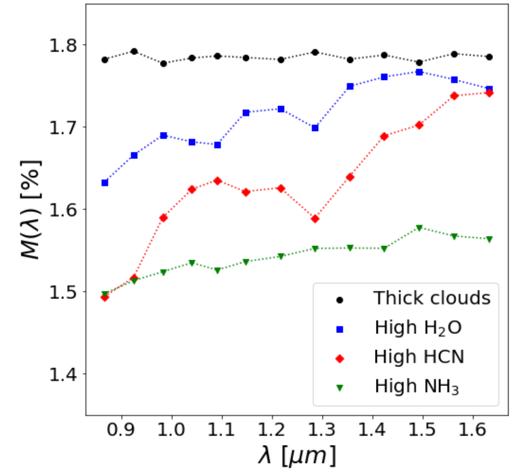
$$R_T(\lambda) = R_0 \left\{ 1 + \frac{H}{R_0} \left[\gamma_E + \ln \left(\frac{P_0 \kappa(\lambda)}{g} \sqrt{2\pi \frac{R_0}{H}} \right) \right] \right\}$$

Heng et al. 2017

Forward Radiative Transfer Models



Observation



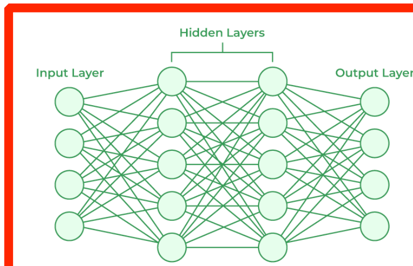
Inputs (features)

Planet	T (K)	R	R	X
1	1300	1.8	1.6	10
2	650	0.9	1.4	10
3	960	1.9	2.3	10
4	1150	2.0	1.5	10

Analytical

$$R_T(\lambda) = R_0\left\{1+\frac{H}{R_0}\left[\gamma_E+\ln\left(\frac{P_0\kappa(\lambda)}{g}\sqrt{2\pi\frac{R_0}{H}}\right)\right]\right\}.$$

RT model: TauREx

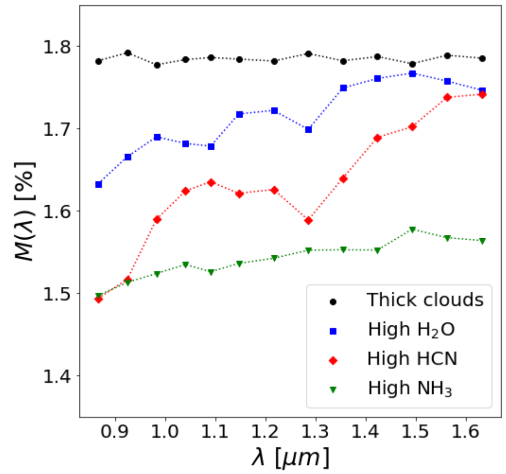


Outputs (targets)

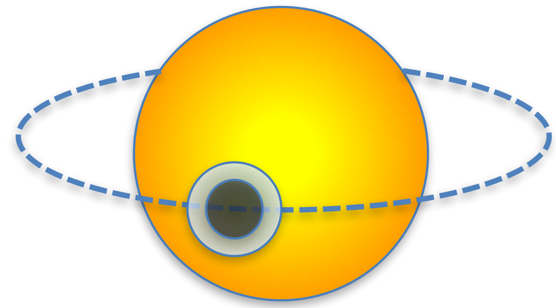
Planet	M	M	M	M
1	1.41	1.44	1.42	1.52
2	0.52	0.55	0.61	0.58
3	0.92	1.03	1.11	0.95
4	1.85	1.94	1.99	1.82

- Generate a training database of spectra **M** by scanning over the **input parameters** for the forward model.

Inverse Problem: parameter retrievals



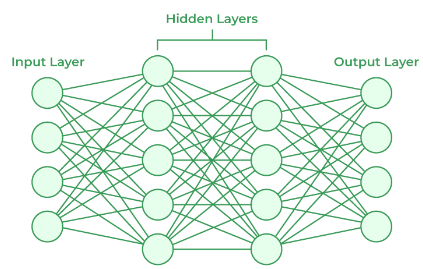
Retrieval Model



Inputs (features)

Planet	M	M	M	M
1	1.41	1.44	1.42	1.52
2	0.52	0.55	0.61	0.58
3	0.92	1.03	1.11	0.95
4	1.85	1.94	1.99	1.82

Retrieval: TauREx



Outputs (targets)

Planet	T (K)	R	R	X
1	1300	1.8	1.6	10
2	650	0.9	1.4	10
3	960	1.9	2.3	10
4	1150	2.0	1.5	10

- The Ariel Data challenge is **ML as a substitute for the Bayesian model**: Train on a database of solutions from TauREx with the goal of reproducing the TauREx predictions.

Meet the DATA!

- **HELA** database (Márquez-Neila P. et al., 2018, Nature,2, 719)

We use a public database[1] of 100,000 synthetic atmospheres created with an **analytical formula**:

- **Fixed parameters**: gravity, mean molecular mass, planetary radius, star radius, reference pressure (WASP-12b)
- **Scanned parameters**:
  - ✓ Temperature: 500 – 2900 K
  - ✓ H_2O volume mixing ratio: $10^{-13} – 1$
  - ✓ HCN volume mixing ratio: $10^{-13} – 1$
  - ✓ NH_3 volume mixing ratio: $10^{-13} – 1$
  - ✓ Cloud opacity: $10^{-13} – 10^{2}$
- Noise floor of 50 ppm on the transit depth (WFC3-like).
- **Spectral range**: 0.838-1.666 μm in 13 bins.

- **TRANSIT** database (with M. Himes, J. Harrington UCF)

We use full forward radiative transfer model (**TRANSIT**) with variable gravity , g, and self consistent mean molecular mass, μ.

- **Fixed parameters:** planetary radius, star radius, pressure grid of 100 layers
- **Scanned parameters**:
  - ✓ Temperature: 500 – 2900 K
  - ✓ H_2O volume mixing ratio: $10^{-13} – 10^{-2}$
  - ✓ HCN volume mixing ratio: $10^{-13} – 10^{-2}$
  - ✓ NH_3 volume mixing ratio: $10^{-13} – 10^{-2}$
  - ✓ Cloud opacity: $10^{-13} – 10^{2}$
  - ✓ Rayleigh Scattering and CIA
- No noise
- **Spectral range**: 0.838-1.666 μm in 13 bins

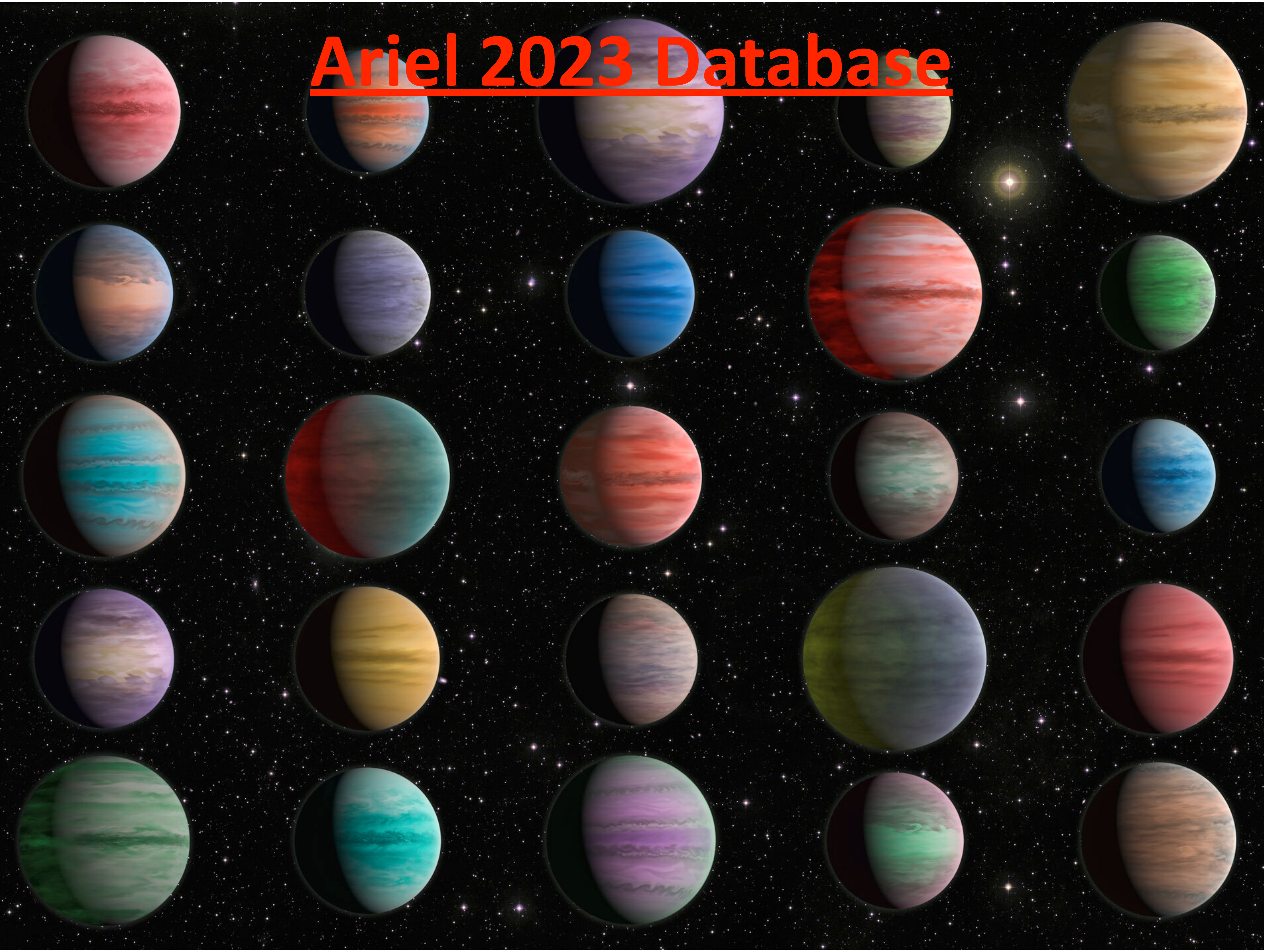
- **Ariel 2022** challenge database (Changeat and Yip, RASTI, 2023).

Ariel database (**TauRex**) with variable gravity , g

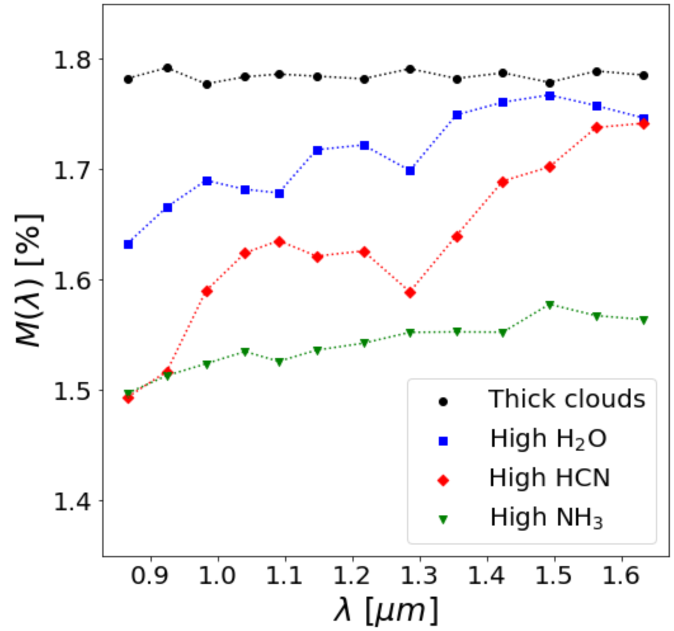
- **Fixed parameters:** pressure grid, mean molecular mass, μ
- **Varying parameters:** target planet/star: R_s, R_p, M_p, g, T_p.
- **Scanned parameters**:
  - ✓ H_2O volume mixing ratio: $10^{-9} – 10^{-3}$
  - ✓ CO_2 volume mixing ratio: $10^{-9} – 10^{-4}$
  - ✓ CH_4 volume mixing ratio: $10^{-9} – 10^{-3}$
  - ✓ CO volume mixing ratio: $10^{-6} – 10^{-3}$
  - ✓ NH_3 volume mixing ratio: $10^{-9} – 10^{-4}$
  - ✓ No clouds
  - ✓ Rayleigh Scattering and CIA
- Noise
- **Spectral range**: 0.5-7.5 μm in 52 bins

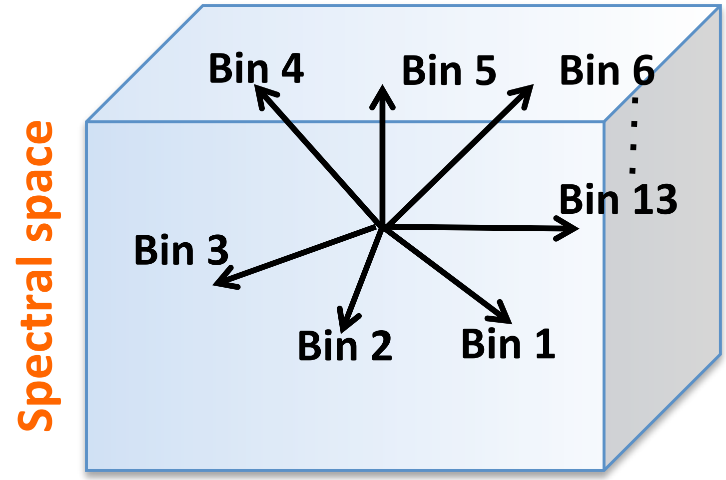
You can make your own database!

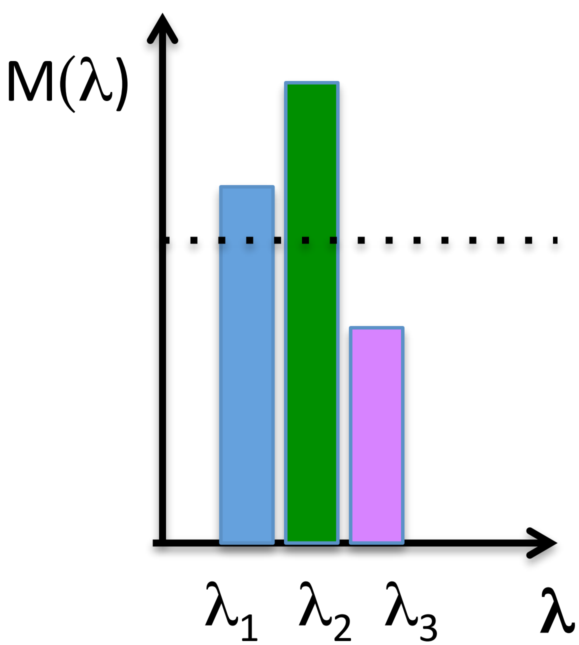
- spectral range, resolution, and noise.
- what are the fixed/varying parameters(T, R, g, m, clouds,…)?
- what are the ranges?
- what type of sampling?
- what optical processes are included?
- what are the physics approximations?


Ariel 2023 Database

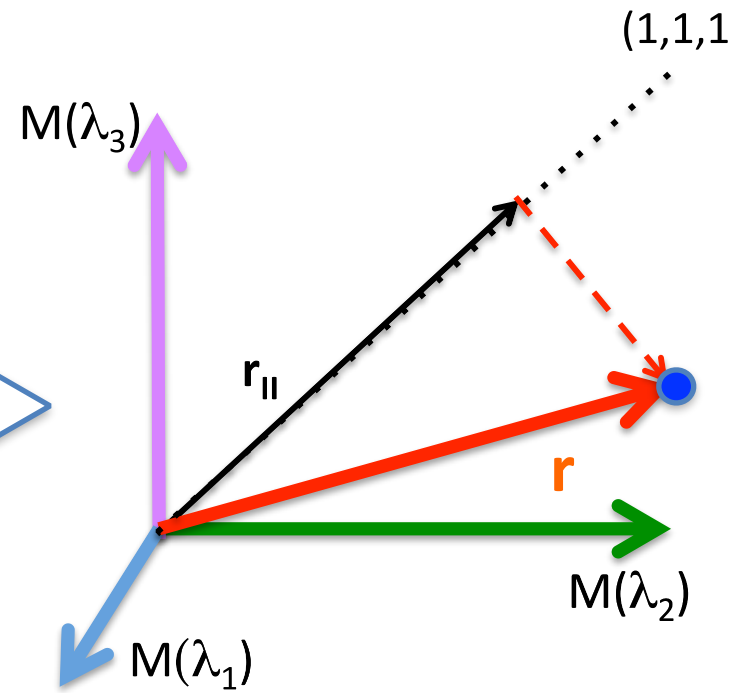
From 1D spectrum to N-dimensions







Each spectrum is a single point in N-dimensional spectral space.

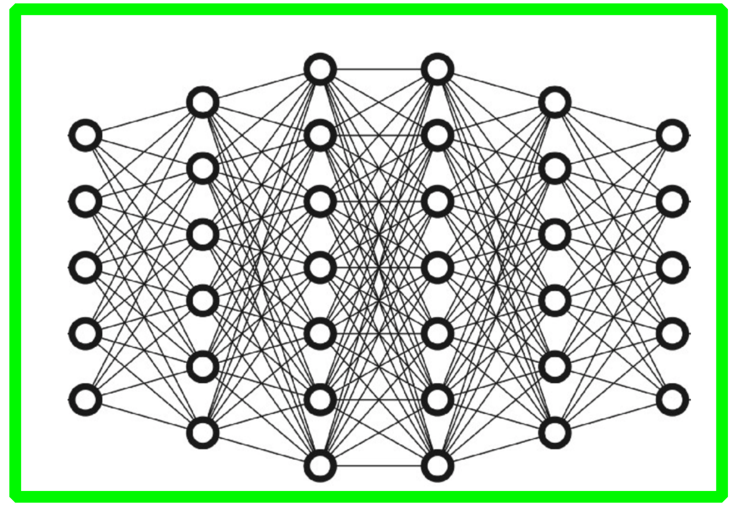


Supervised/Unsupervised Learning

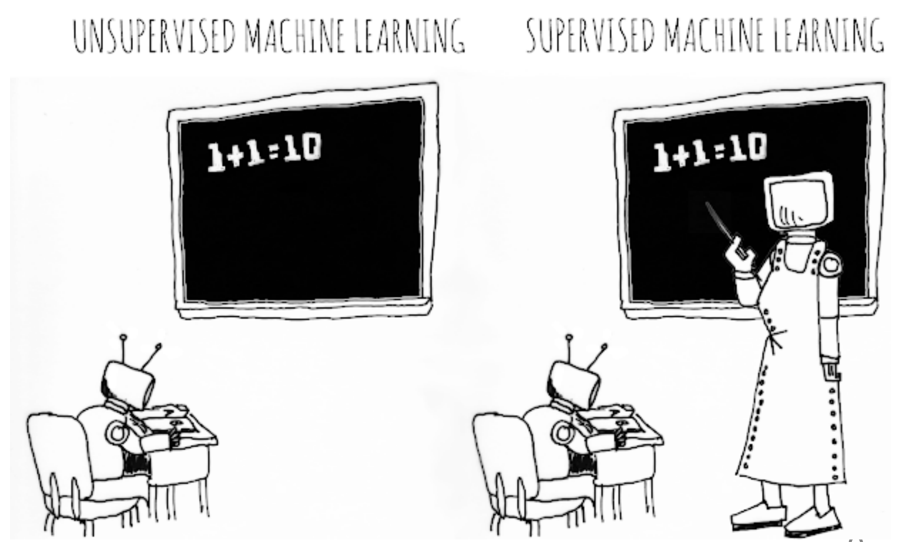
n **features** *k* **labels** **ML algorithm**

Planet *m* **samples**

$$\left\{\begin{array}{cccc|ccc} x_1^{(1)}, & x_1^{(2)}, & \ldots, & x_1^{(n)}; & y_1^{(1)}, & \ldots, & y_1^{(k)} \\ x_2^{(1)}, & x_2^{(2)}, & \ldots, & x_2^{(n)}; & y_2^{(1)}, & \ldots, & y_2^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_m^{(1)}, & x_m^{(2)}, & \ldots, & x_m^{(n)}; & y_m^{(1)}, & \ldots, & y_m^{(k)} \end{array}\right.$$



- **Supervised** Learning - using both features and labels
- **Unsupervised** Learning - using features only
- **Semi-supervised** Learning - Using some labels to label an unlabeled data set to increase the size of the available training dataset
- Reinforcement Learning
- …



Supervised Learning

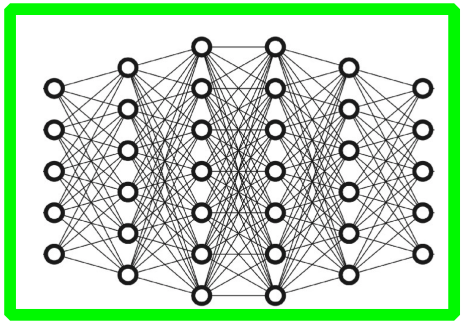
n **features**

k **labels**

Planet m **samples**

$$\begin{cases} x_1^{(1)}, x_1^{(2)}, \ldots, x_1^{(n)}; \\ x_2^{(1)}, x_2^{(2)}, \ldots, x_2^{(n)}; \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(n)}; \end{cases}$$

Build a model that maps {x} to {y}



$$\begin{matrix} y_1^{(1)}, \ldots, y_1^{(k)} \\ y_2^{(1)}, \ldots, y_2^{(k)} \\ \vdots \quad \ddots \quad \vdots \\ y_m^{(1)}, \ldots, y_m^{(k)} \end{matrix}$$

- Supervised ML uses both the **features** and the **labels** to train, validate, and test. typical Exoplanet tasks:
  - **Regression problems:**
    - given the planet/stellar parameters and composition - >predict the observed spectrum
    - given the observed spectrum -> predict the planet parameters and composition
  - **Categorization problem:**
    - given an observation (transit spectrum) what kind of planet that is (giant, terrestrial; cloudy or not; water rich or poor; …) as the training set is already split in **categories**.

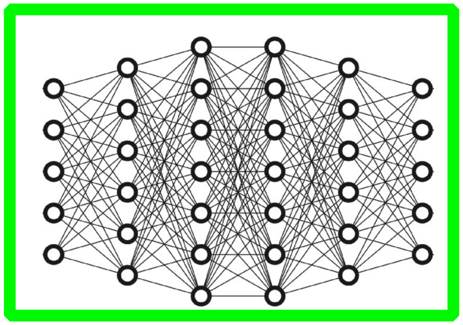
Unsupervised Learning

n **features**

Planet m **samples**

$$
\begin{cases}
x_1^{(1)}, x_1^{(2)}, \ldots, x_1^{(n)}; \\
x_2^{(1)}, x_2^{(2)}, \ldots, x_2^{(n)}; \\
\vdots \quad \vdots \quad \ddots \quad \vdots \\
x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(n)};
\end{cases}
$$

Build a model that maps {x} to {y}



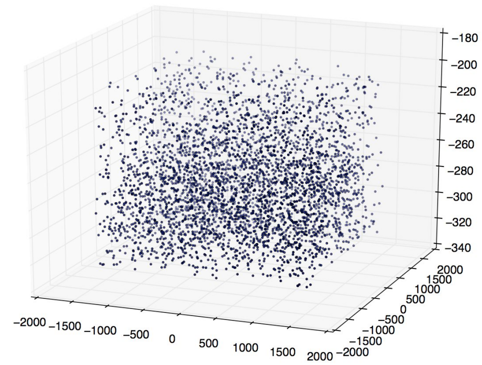
Questions???

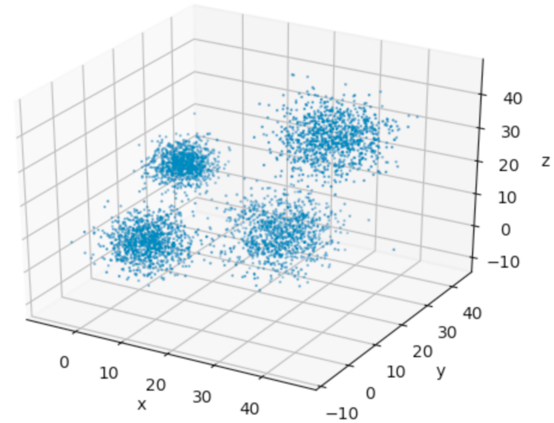
Why?
- You do not need labels.
- Getting labels is not easy.
- There is a lot of unlabeled data out there.
- First and last resort when you are clueless of what to do.

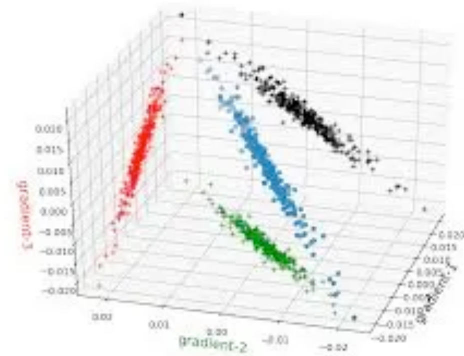
Common tasks:
- Clustering
- Outlier detection
- Dimensionality reduction
  - PCA
  - Manifold learning
  - Auto-encoders
- Feature engineering

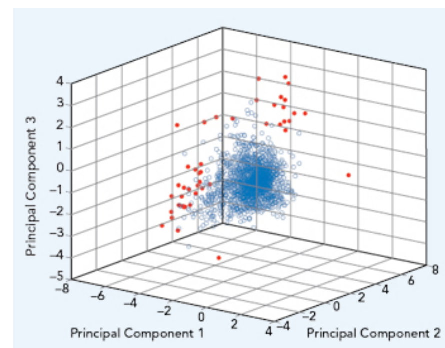
Visualization and Unsupervised ML

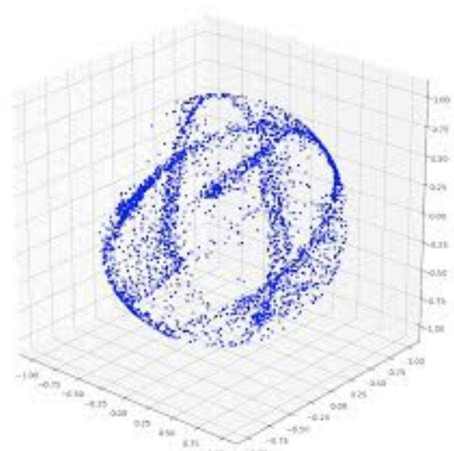
- **Unsupervised** ML uses only the **features** to answer **interesting questions**.
  - What is the **max/min** values of my data?
  - What is the variability (**standard deviation**)?
  - What is the **density** of the distribution?
  - What **type of distribution** it is?
  - Are there any **correlations**?
  - Are there any **clusters**?
  - Are there any **unusual data points**?
  - What is the true **dimensionality** of the dataset?
  - How many **free parameter** I need to describe the data?
  - Is there any **symmetry** in the data?

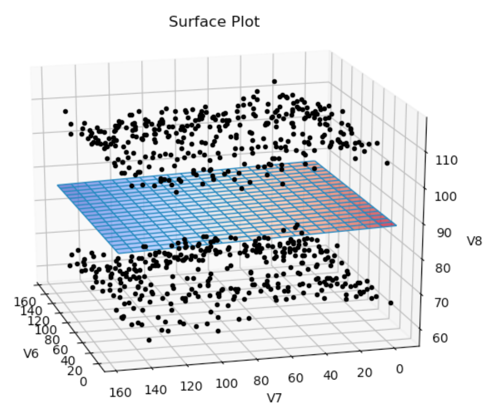








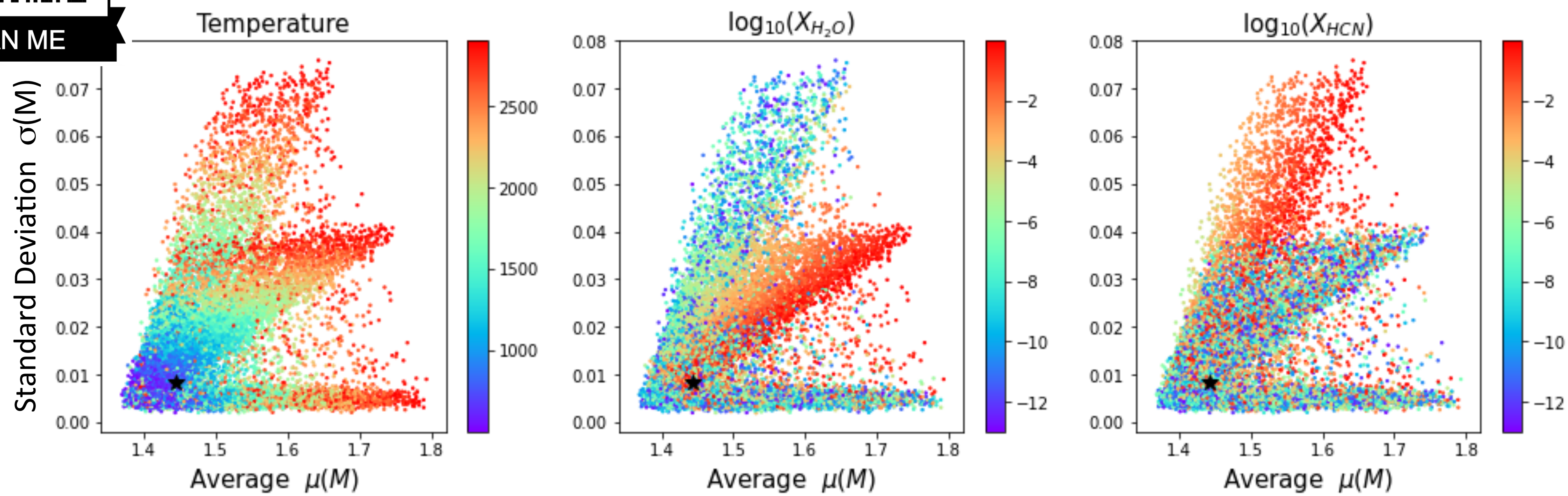




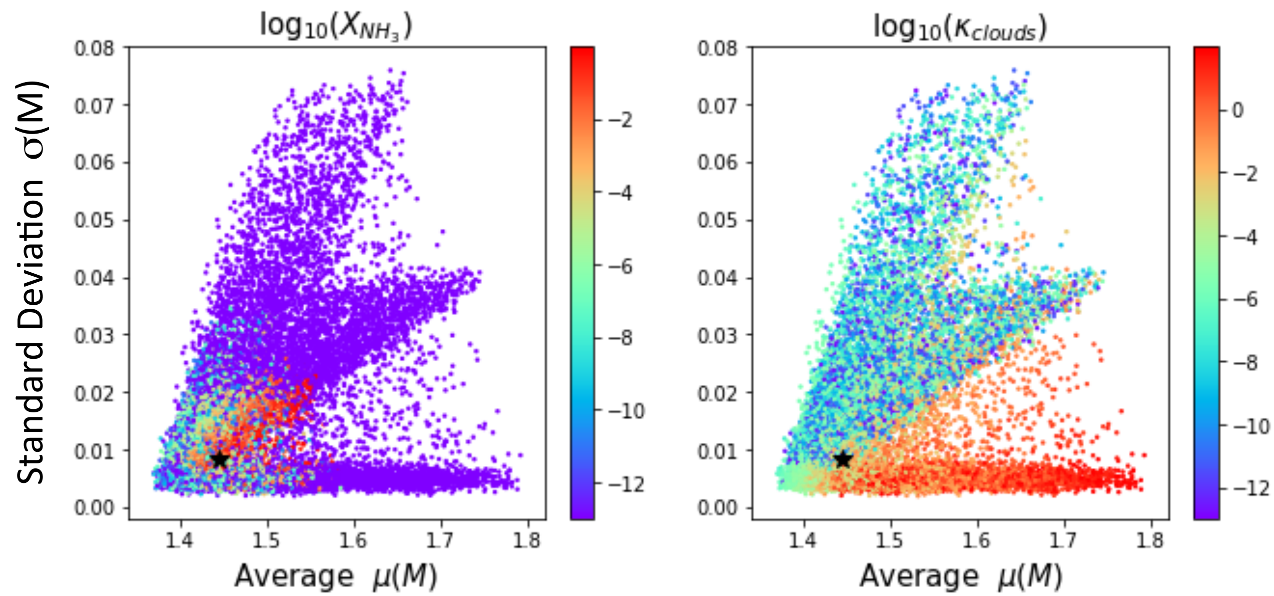
SCAN ME

Data Summary Statistics

● HELA database



2-D scatter plots of the **Standard Deviation** and the **Spectral Average** of the 13-dimentional spectral database. The points are color coded using the temperature, **T** or the composition, **X**. The black star * marks the position of **WASP-12b**

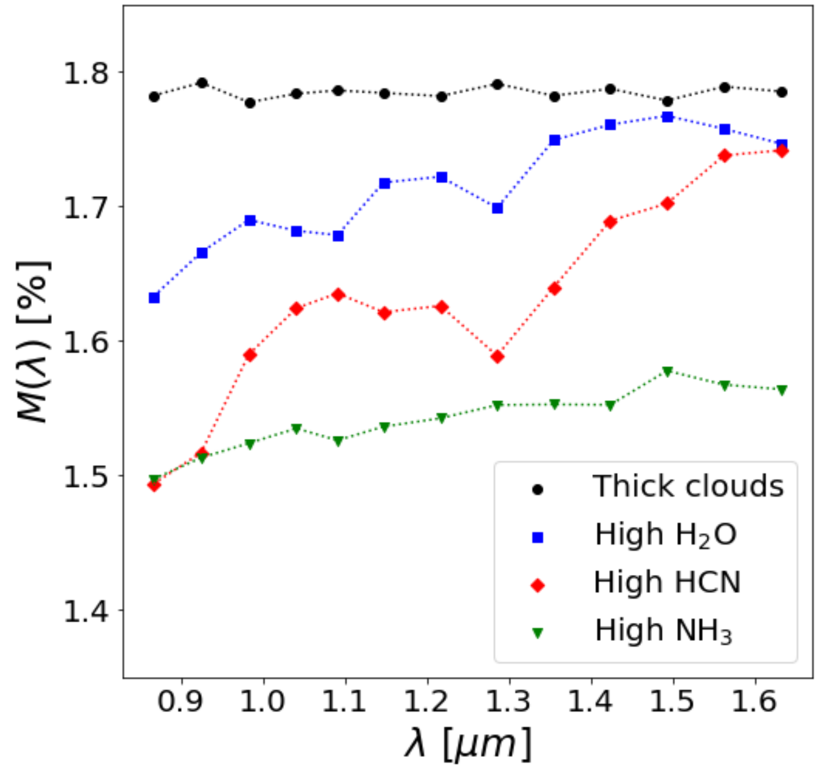


Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022

Information Content-Correlations

- **HELA** database

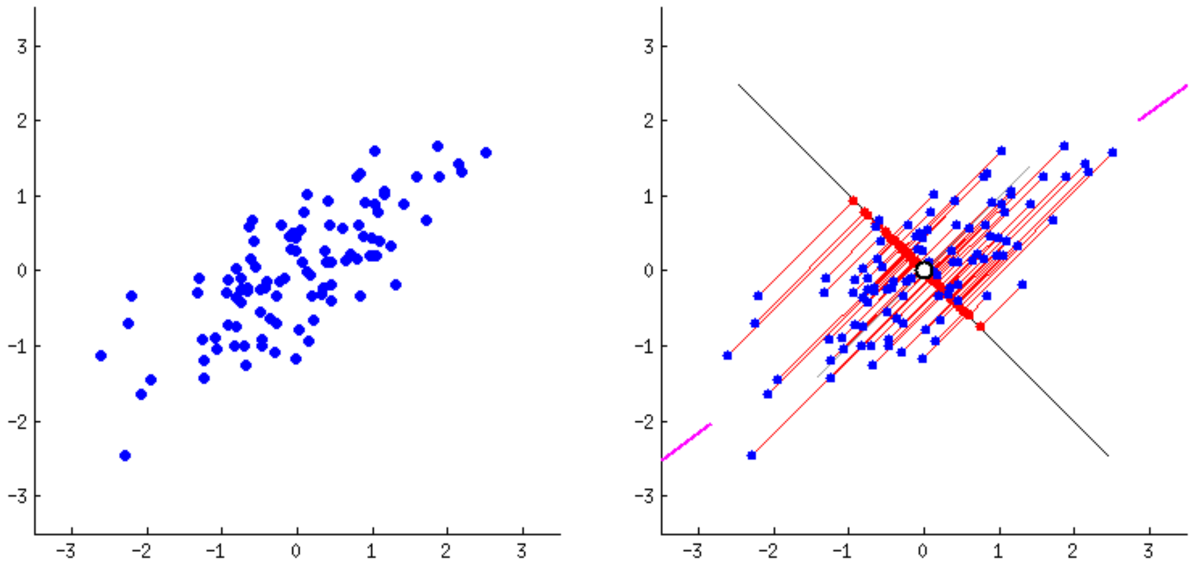
	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12	M13
M01	1	0.99	0.97	0.96	0.94	0.95	0.94	0.96	0.91	0.86	0.85	0.82	0.84
M02	0.99	1	0.98	0.96	0.95	0.96	0.96	0.97	0.94	0.89	0.87	0.84	0.85
M03	0.97	0.98	1	0.99	0.98	0.99	0.99	0.99	0.98	0.95	0.94	0.92	0.92
M04	0.96	0.96	0.99	1	0.99	0.99	0.99	0.99	0.97	0.95	0.95	0.94	0.95
M05	0.94	0.95	0.98	0.99	1	0.99	0.99	0.98	0.97	0.96	0.96	0.95	0.96
M06	0.95	0.96	0.99	0.99	0.99	1	0.99	0.99	0.99	0.97	0.96	0.94	0.95
M07	0.94	0.96	0.99	0.99	0.99	0.99	1	0.99	0.99	0.97	0.97	0.95	0.96
M08	0.96	0.97	0.99	0.99	0.98	0.99	0.99	1	0.98	0.95	0.95	0.93	0.93
M09	0.91	0.94	0.98	0.97	0.97	0.99	0.99	0.98	1	0.99	0.98	0.96	0.96
M10	0.86	0.89	0.95	0.95	0.96	0.97	0.97	0.95	0.99	1	0.99	0.98	0.98
M11	0.85	0.87	0.94	0.95	0.96	0.96	0.97	0.95	0.98	0.99	1	0.99	0.99
M12	0.82	0.84	0.92	0.94	0.95	0.94	0.95	0.93	0.96	0.98	0.99	1	1
M13	0.84	0.85	0.92	0.95	0.96	0.95	0.96	0.93	0.96	0.98	0.99	1	1



Correlation matrix of the 13 features (spectral bins) M_λ (λ=1…13). The matrix lists the Pearson correlation coefficient between any two features in the dataset. The information in the individual spectral bins is clearly **highly correlated**. This calls for **dimensionality reduction**.

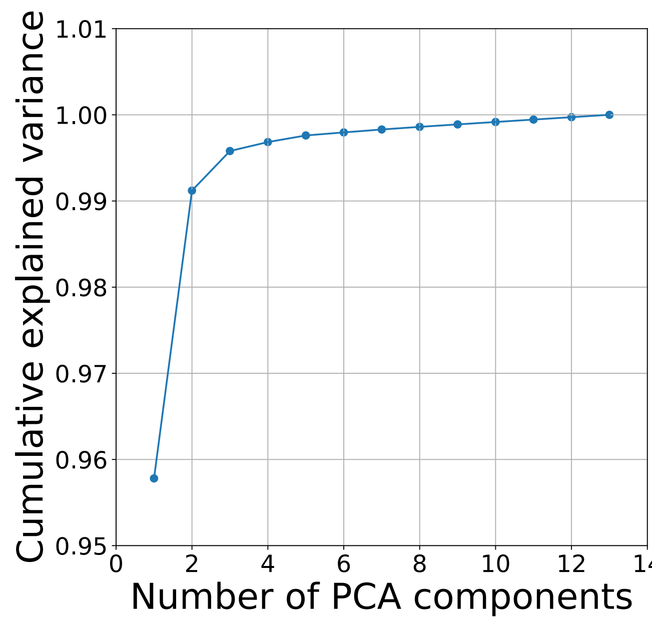
Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022

Principal Component Analysis



https://miro.medium.com/v2/resize:fit:1400/1*37a_i1t1tDxDYT3ZI6Yn8w.gif

● HELA database

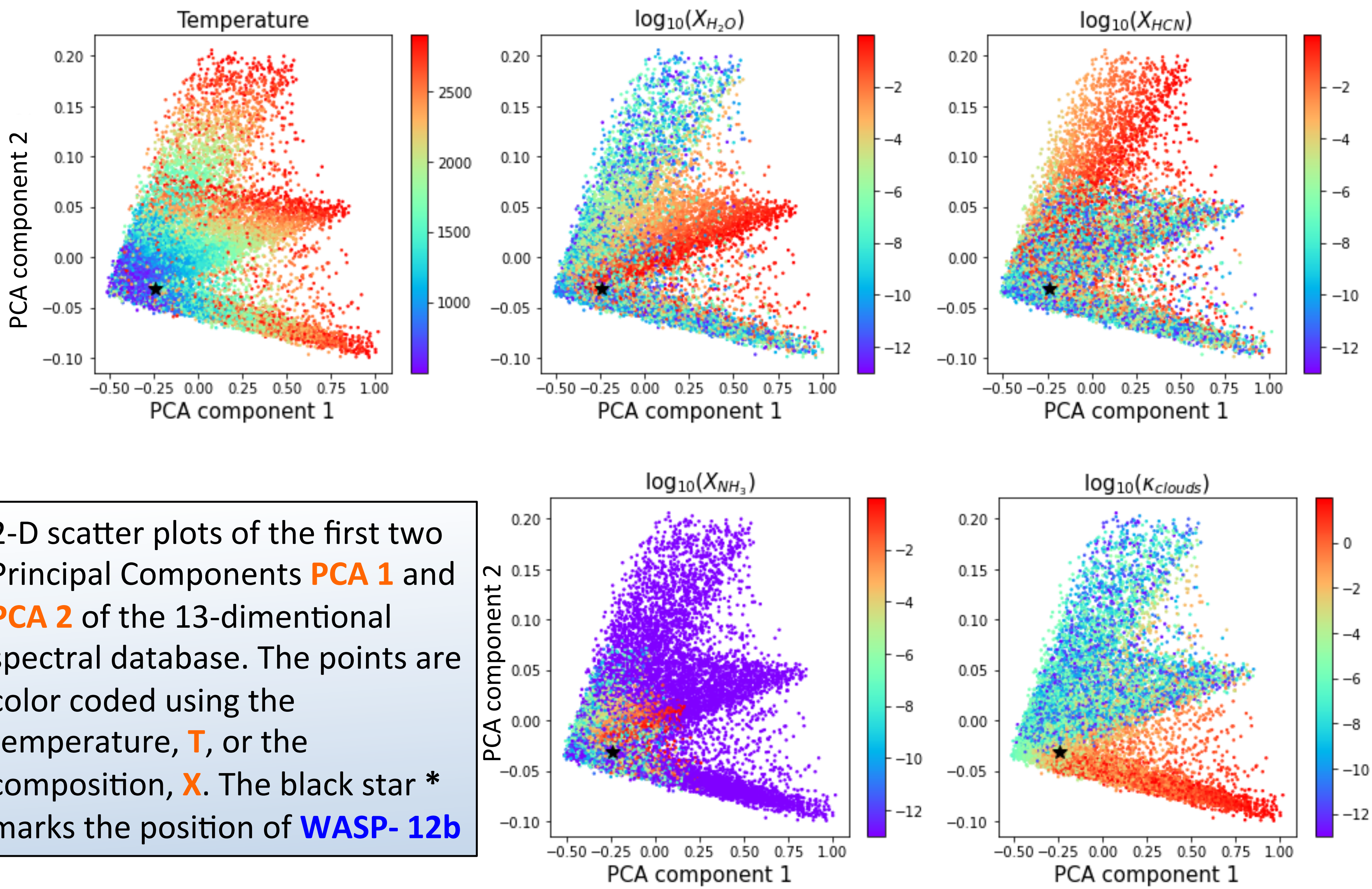


Principal Component Analysis: the plot on the right shows the cumulative explained variance ratio as a function of the number of included PCA components. The first three PCA components alone contain more than **99.5 %** of the variance in the data.

Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022

Principal Component Analysis

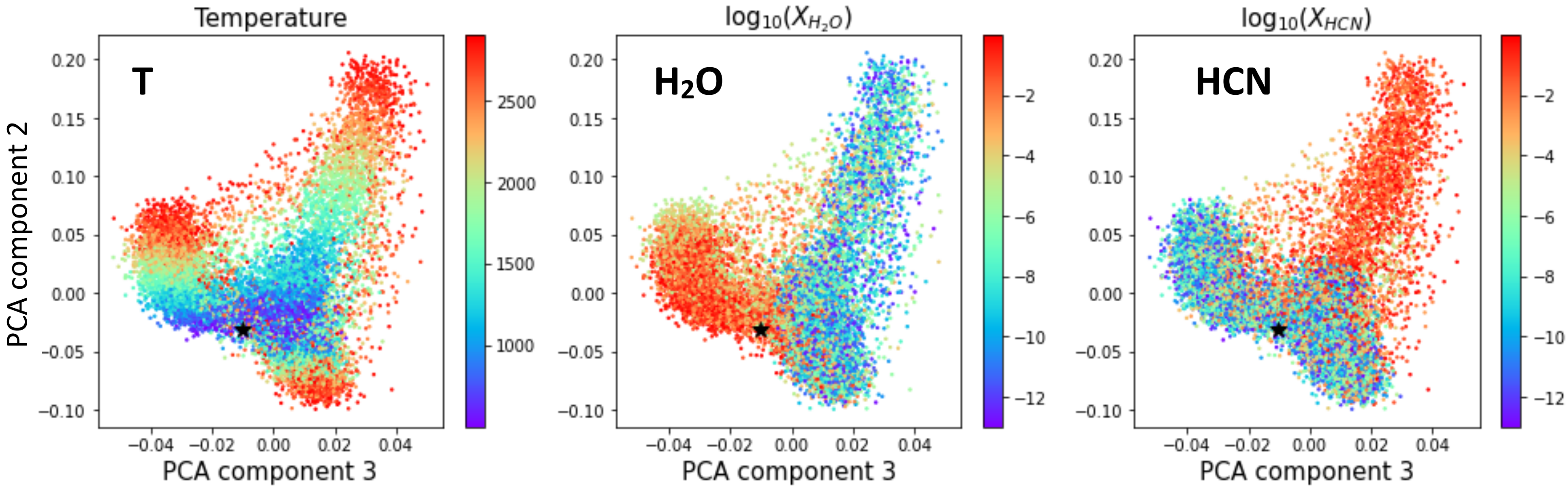
● **HELA** database

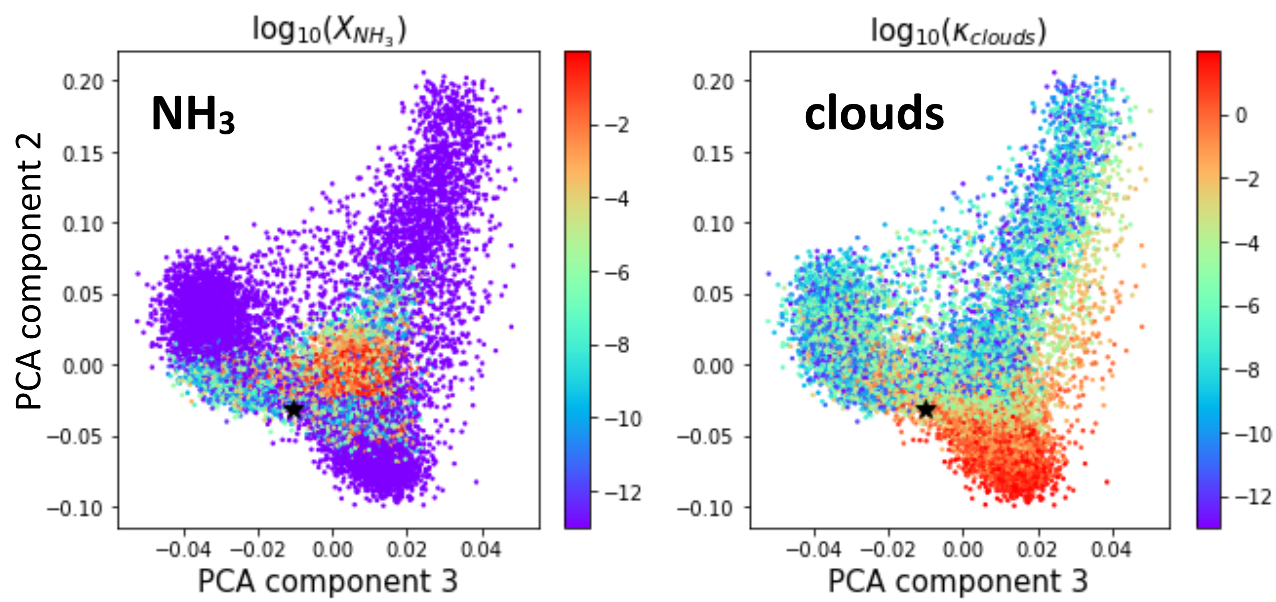


2-D scatter plots of the first two Principal Components **PCA 1** and **PCA 2** of the 13-dimentional spectral database. The points are color coded using the temperature, **T**, or the composition, **X**. The black star * marks the position of **WASP- 12b**

PCA 2D representation

- **HELA** database



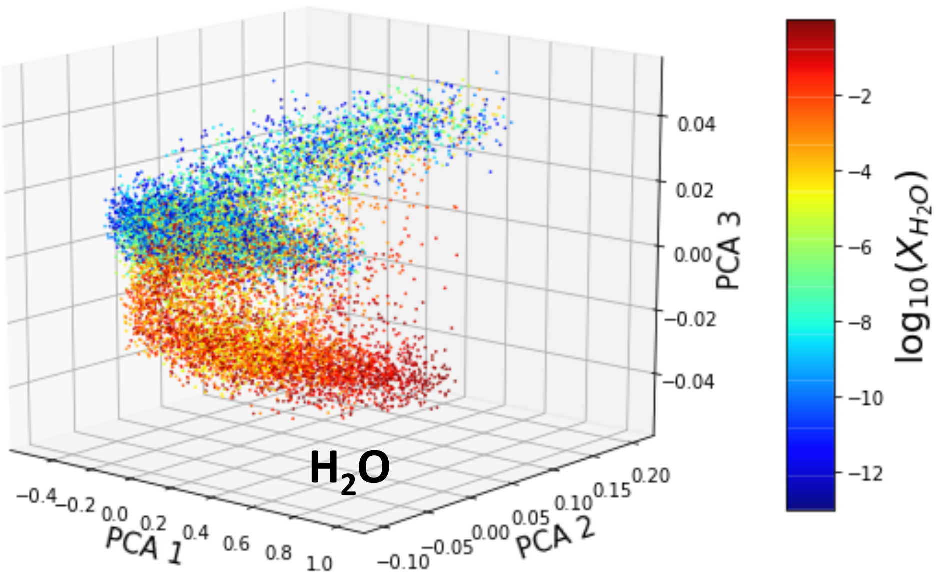


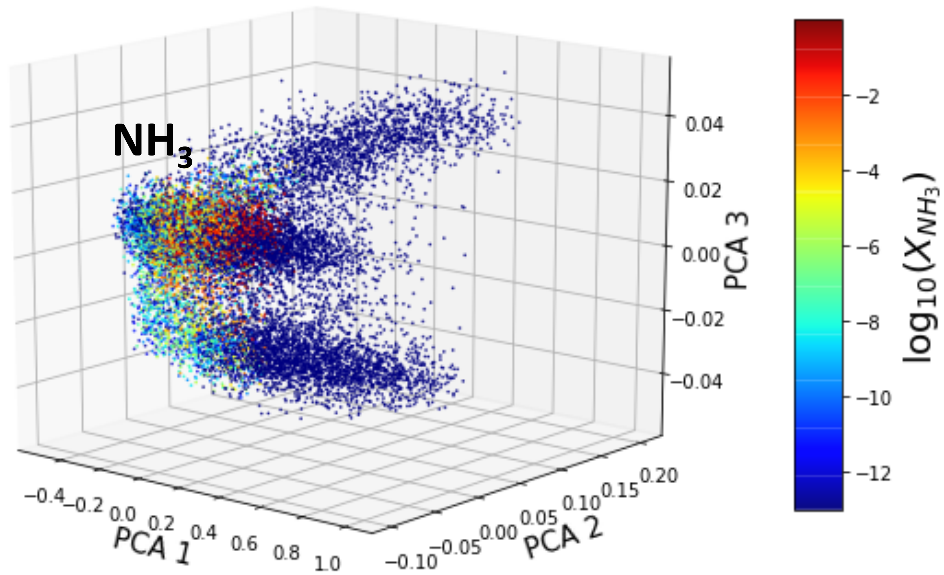
2-D scatter plots of the second and the third Principal Components, **PCA 2** and **PCA 3** of the 13-dimentional spectral database. The points are color coded using the temperature, **T**, or the composition, **X**. The black star * marks the position of **WASP- 12b**

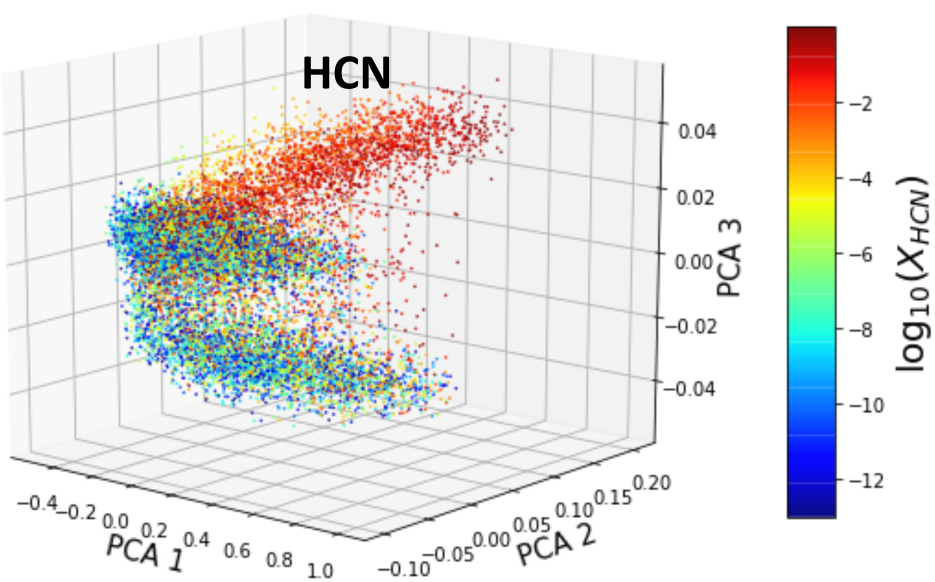
Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022

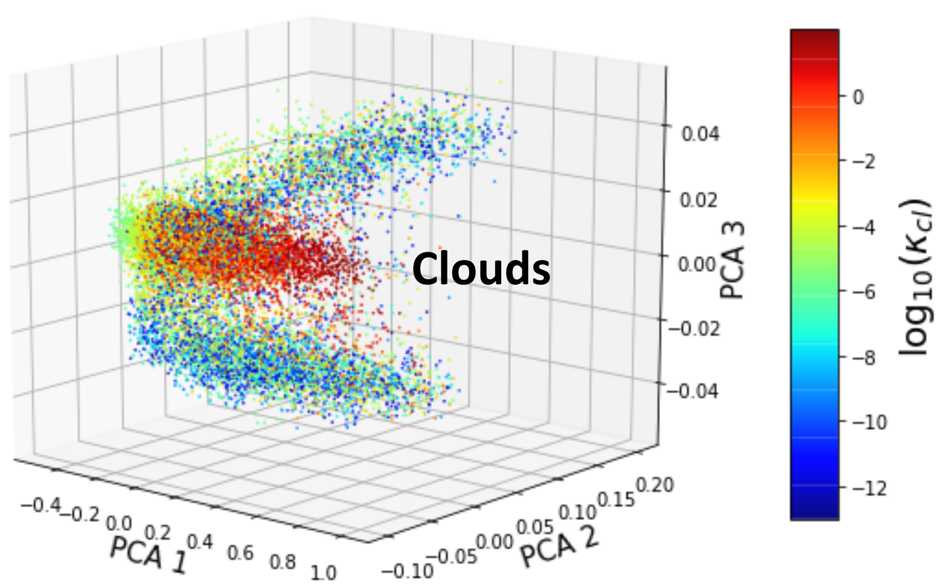
PCA 3D representation

Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022 ● **HELA** database



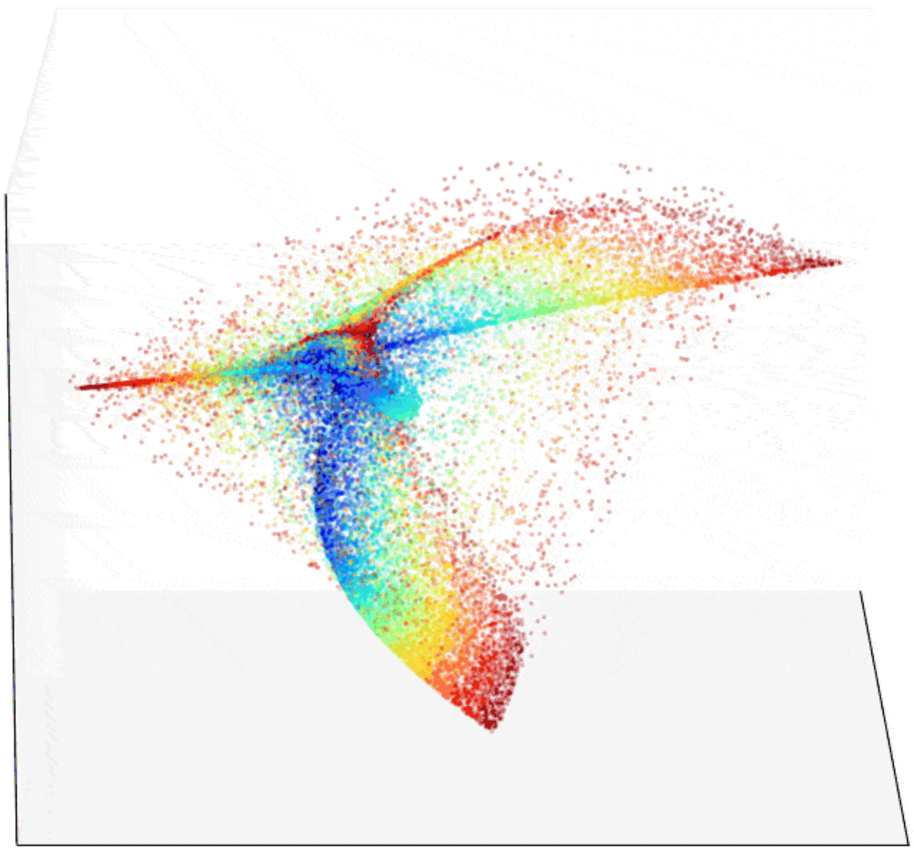






3-D representation of the first 3 Principal Components

- **TRANSIT** database (with M. Himes, J. Harrington, unpublished)



Color coding by temperature, T=500-2900K

Spectral classes of chemical regimes

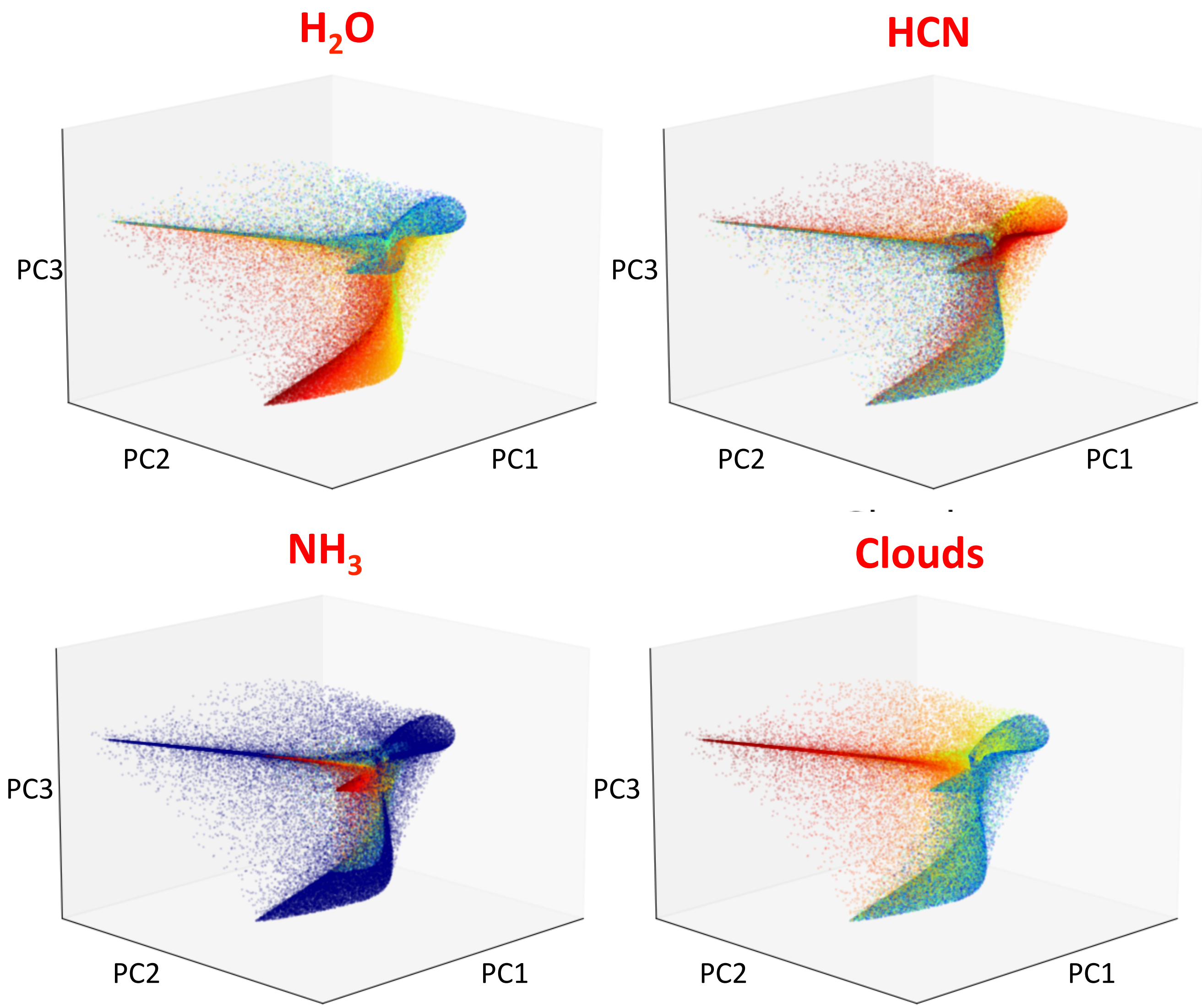
- ✓ H_2O branch
- ✓ NH_3 branch
- ✓ HCN branch
- ✓ Cloud branch

Distinct branch for each extinction

- ✓ Absorption due to distinct absorber
- ✓ Scattering
  - ✓ Grey clouds
  - ✓ Rayleigh scattering
- ✓ CIA (H_2-H_2, H_2-He)

PCA of Synthetic Spectra

- **TRANSIT** database (with M. Himes, J. Harrington, unpublished)

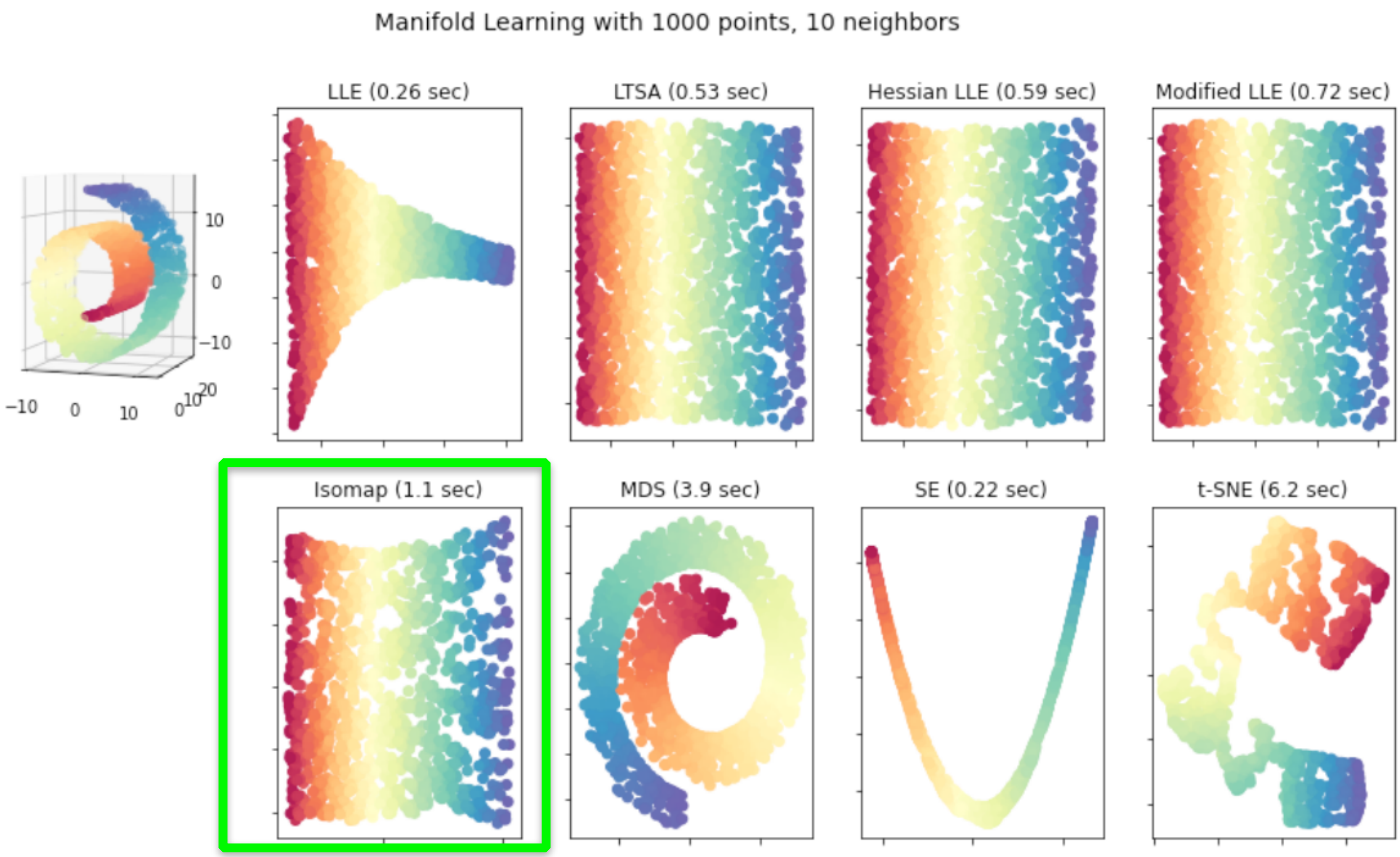


Synthetic Database

- ✓ 100,000 spectra
- ✓ for WASP-12b like planets
- ✓ full forward radiative transfer model (**TRANSIT**)
- ✓ *100 atmospheric pressure layers,
- ✓ variable gravity , g.
- ✓ self consistent mean molecular mass, μ.

Manifold Learning: Swiss Roll

Dimensionality reduction method using non-linear transformations.



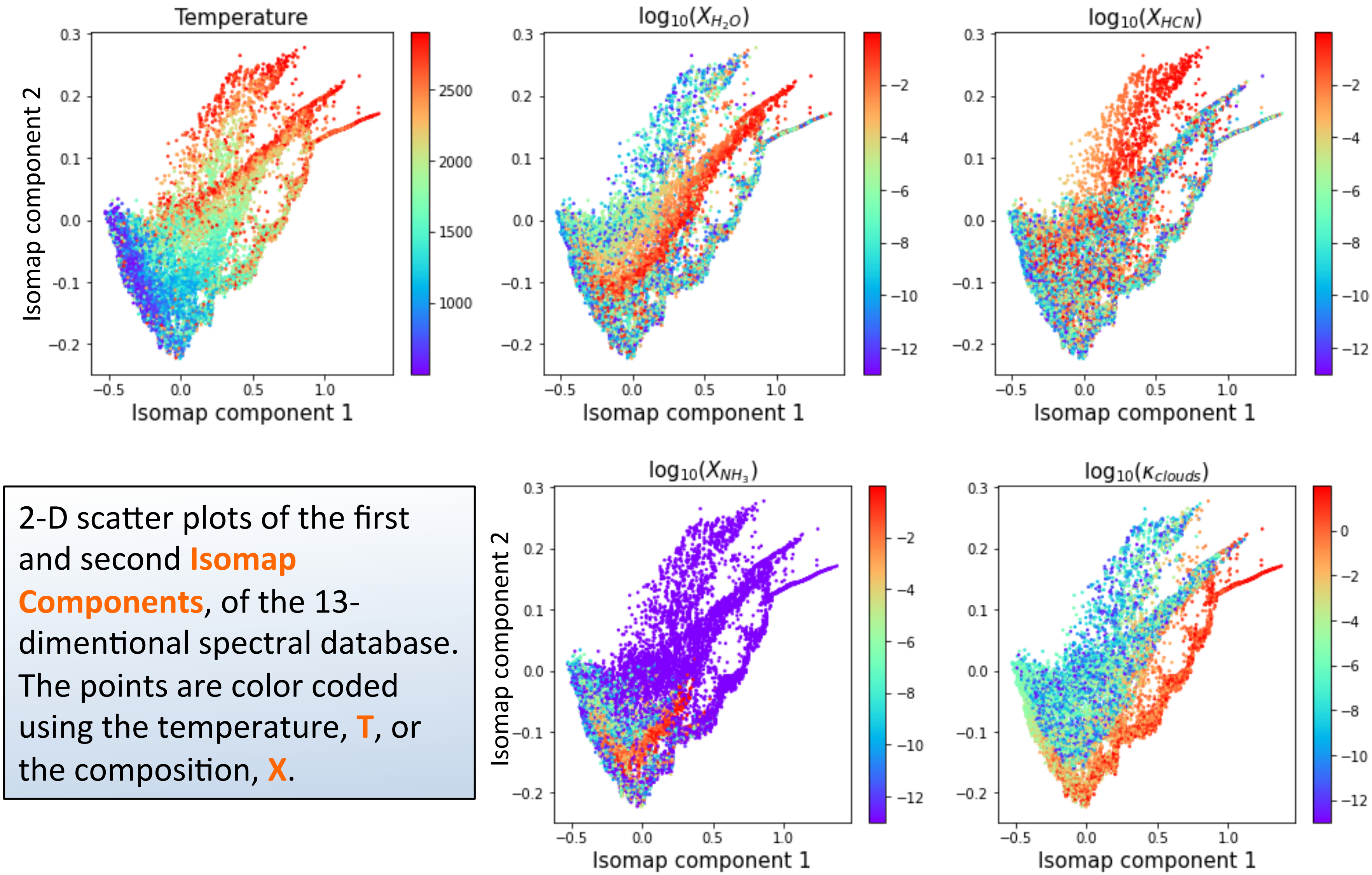
credit: https://parastoofalakaflaki.medium.com/manifold-learning-e8ca7b6df0f8

Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022

Non-linear dimensionality reduction

- **HELA** database

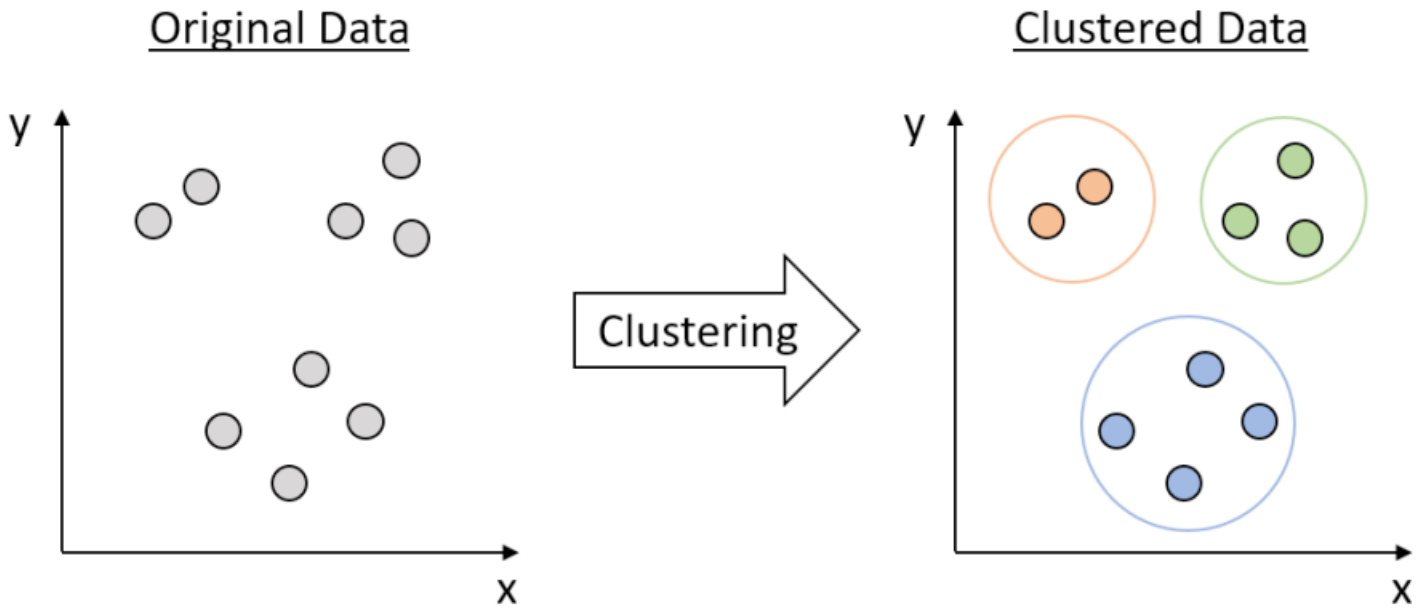
Isomap



2-D scatter plots of the first and second **Isomap Components**, of the 13-dimentional spectral database. The points are color coded using the temperature, **T**, or the composition, **X**.

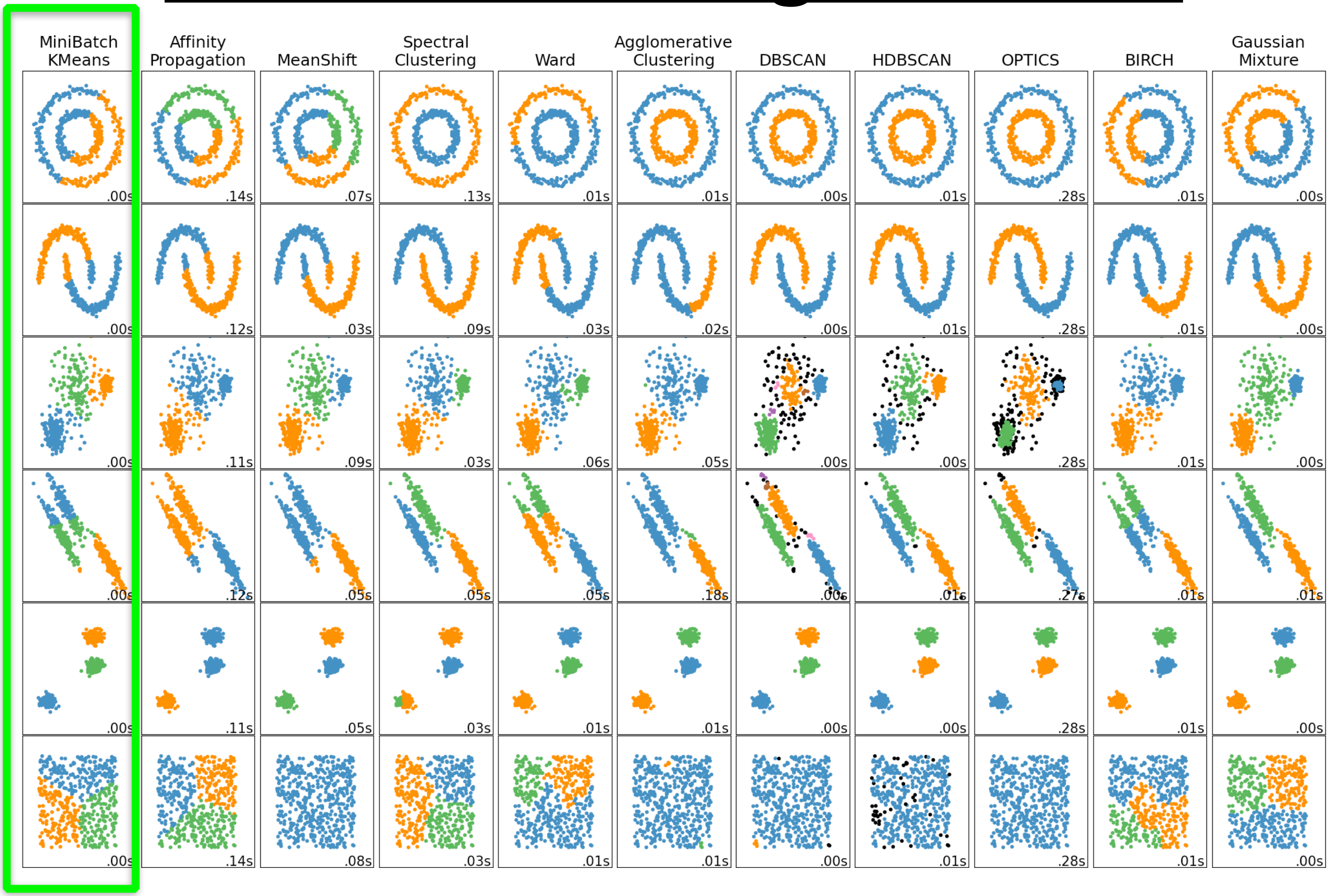
Clustering

- Unsupervised learning (**no labels**).
- Methods based on data **density estimation**.
- Large number of methods.
- Most methods require to **specify** the number of clusters.
- Significant number of "**hyper parameters**" that needs to be fine-tuned.



credit: https://waterprogramming.wordpress.com/2022/03/16/clustering-basics-and-a-demonstration-in-clustering-infrastructure-pathways/

Zoo of clustering methods



credit: https://scikit-learn.org/stable/modules/clustering.html

Clustering

Matchev, Matcheva, Roman, PSJ, v 3, id 205, 2022

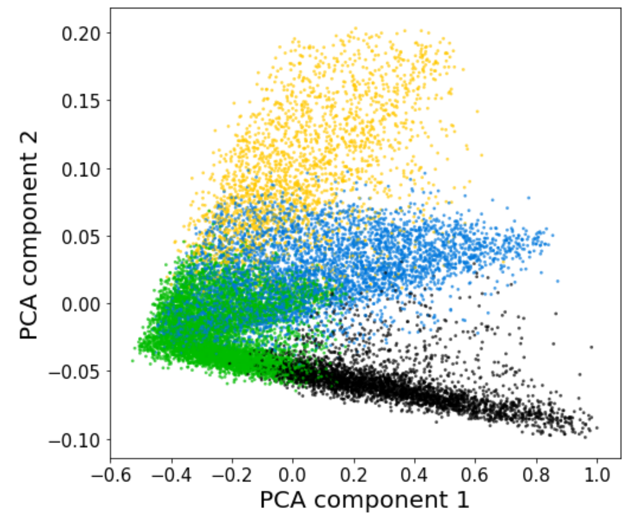
- **HELA** database

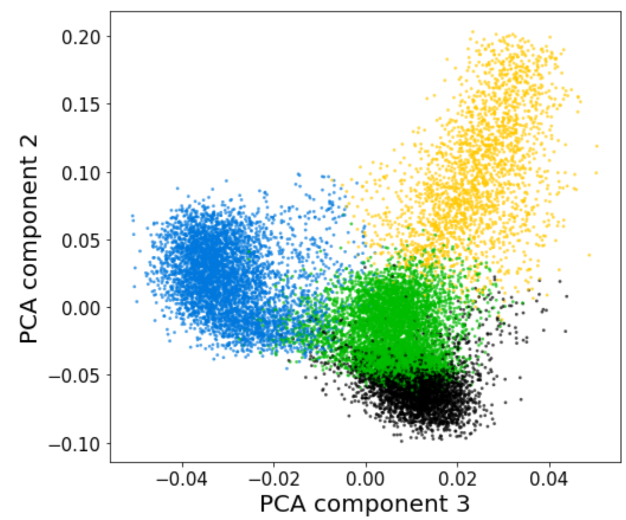
We use a public database[1] of 100,000 synthetic atmospheres:

- **Fixed parameters**: gravity, mean molecular mass, planetary radius, star radius, reference pressure (WASP-12b)
- **Scanned parameters**:
  - ✓ Temperature: 500 – 2900 K
  - ✓ H_2O volume mixing ratio: $10^{-13} – 1$
  - ✓ HCN volume mixing ratio: $10^{-13} – 1$
  - ✓ NH_3 volume mixing ratio: $10^{-13} – 1$
  - ✓ Cloud opacity: $10^{-13} – 10^2$
- Noise floor of 50 ppm on the transit depth (WFC3-like).
- **Spectral range**: 0.838-1.666 μm in 13 bins.

We perform several **unsupervised learning** tasks:

- Summary statistics
- Clustering (see figures on the right)
- Dimensionality reduction
- Manifold learning

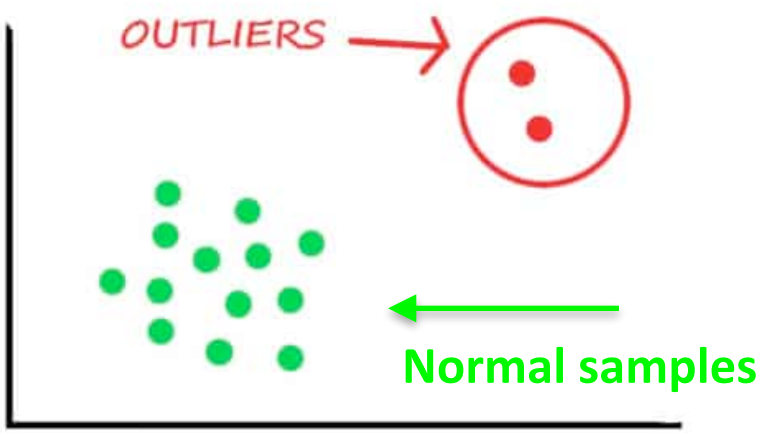


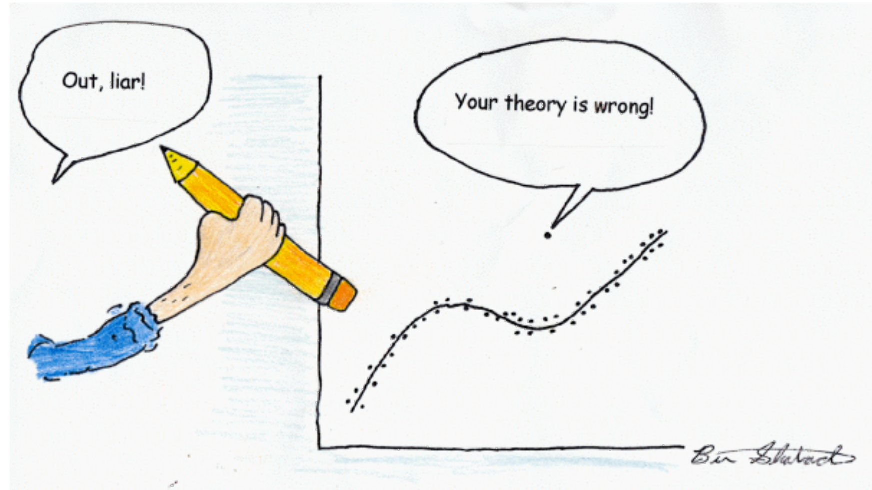


[1] Márquez-Neila P., Fisher C., Sznitman R., Heng K., 2018, Nature Astronomy,2, 719
https://github.com/exoclime/HELA

Results from **K-means** clustering of the synthetic atmospheres in the database

Anomaly Detection





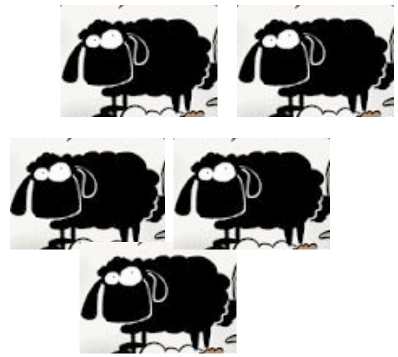
- **Basic question:** Does a given observation belong to the same distribution as the others (inlier) or is it different (outlier)?
- Some possible reasons for outliers:
  - Measurement or input error
  - Data corruption
  - True outlier observation (discovery!)
- **Anomaly detection** methods alert you to the presence of anomalous data, but do not tell you what to do with it - that is up to you.

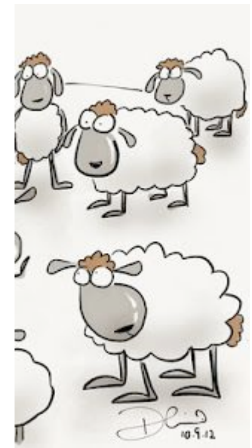
Outlier versus Novelty Detection

- **Outlier detection:** useful when we have an idea what anomalies might look like.

Training data

Testing data

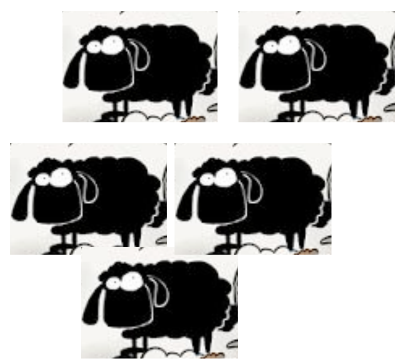


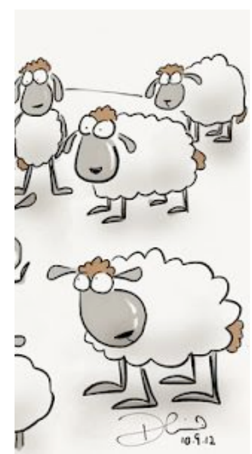


- **Novelty detection:** useful when we do not know what the potential anomalies look like.

Training data

Testing data





Novelty Detection using Ariel Data

Paper: motivation, approach, data preprocessing, method, results

(Forestano et al. 2023)

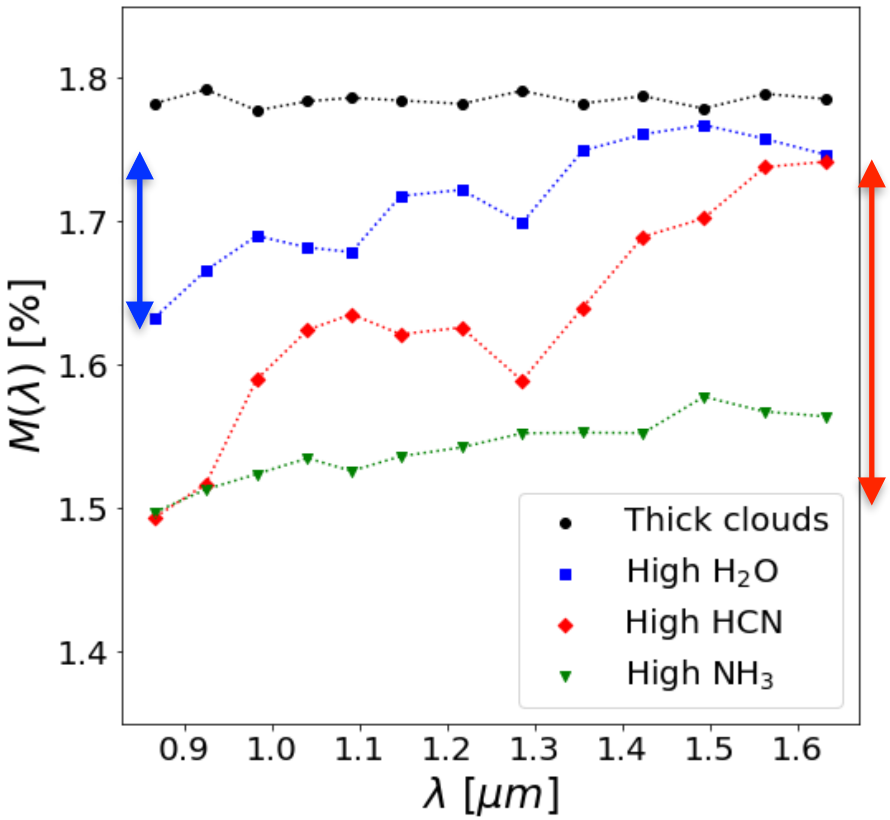
SCAN ME

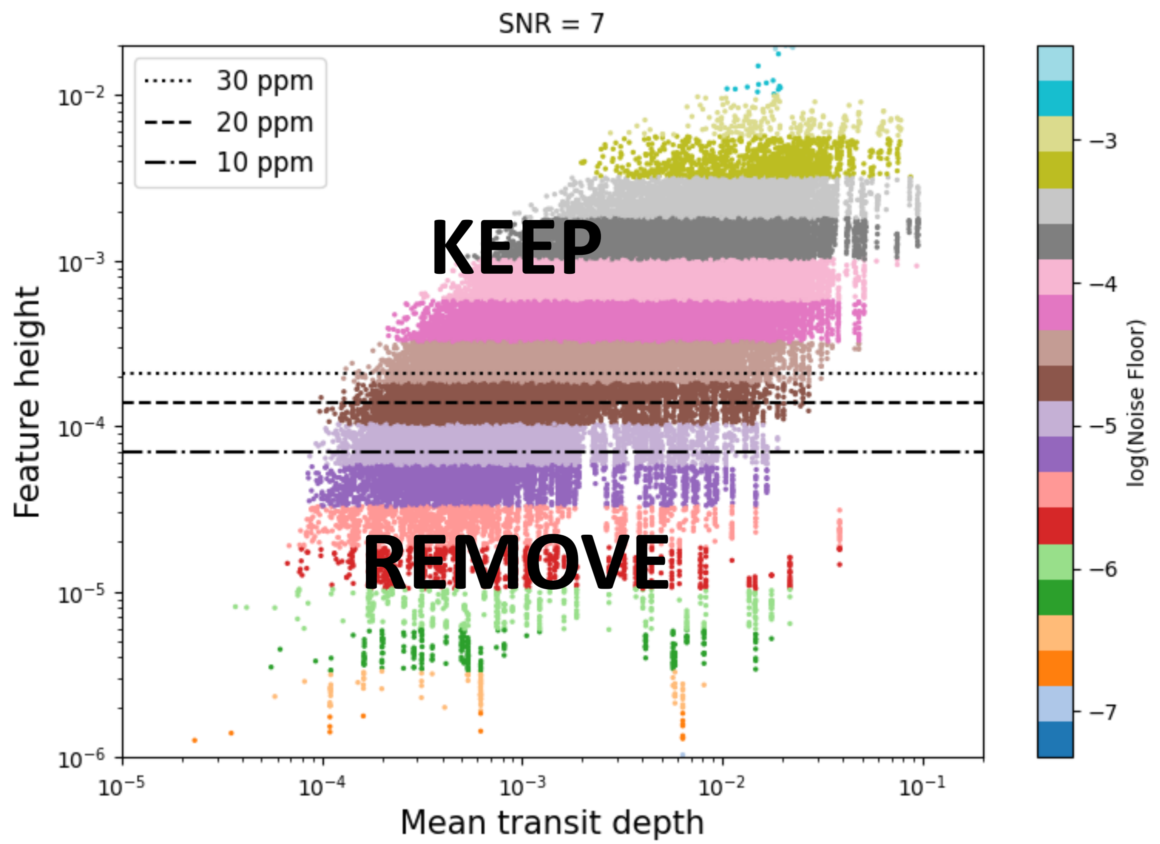
- The **Basic Questions**:
  - Can we identify planets with unusual or **unexpected chemical** composition?
  - Can we spot **alien life** as we do not know it?
  - Can we identify **new physics**?
  - Can we spot **glitches** with the instrument?
- Can we detect **anomalous** spectra?

- Starting from the generic data base let’s **reshape** it so that it fits our problem.

Spectra preselection

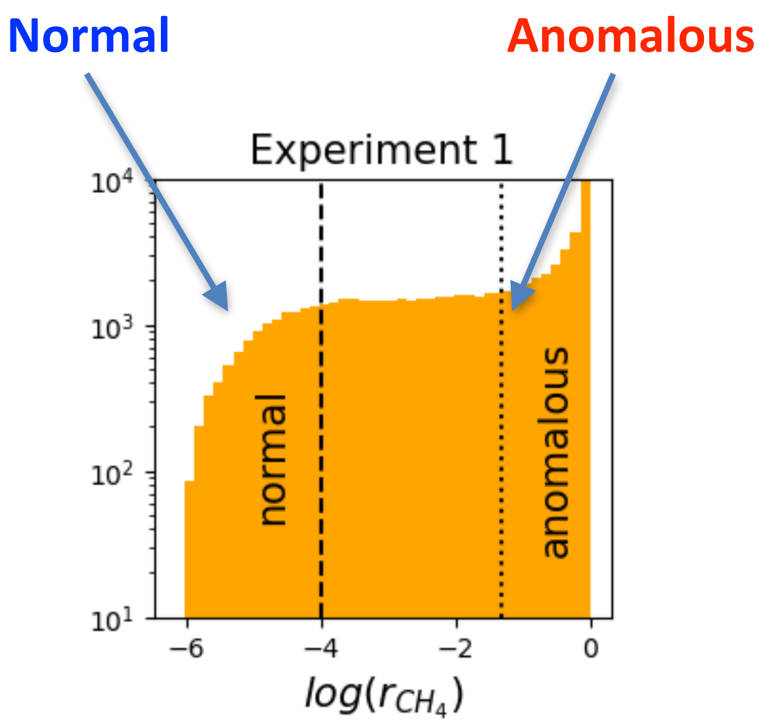
- A random scanning of the parameter space results in many **unobservable** transits or **featureless** spectra, which are **not interesting**.
- They can be removed by requiring a **minimum value for the feature height**
    - this cut is tied up to the **noise level**: large noise washes out small features
    - approximately 65,000 remaining spectra.



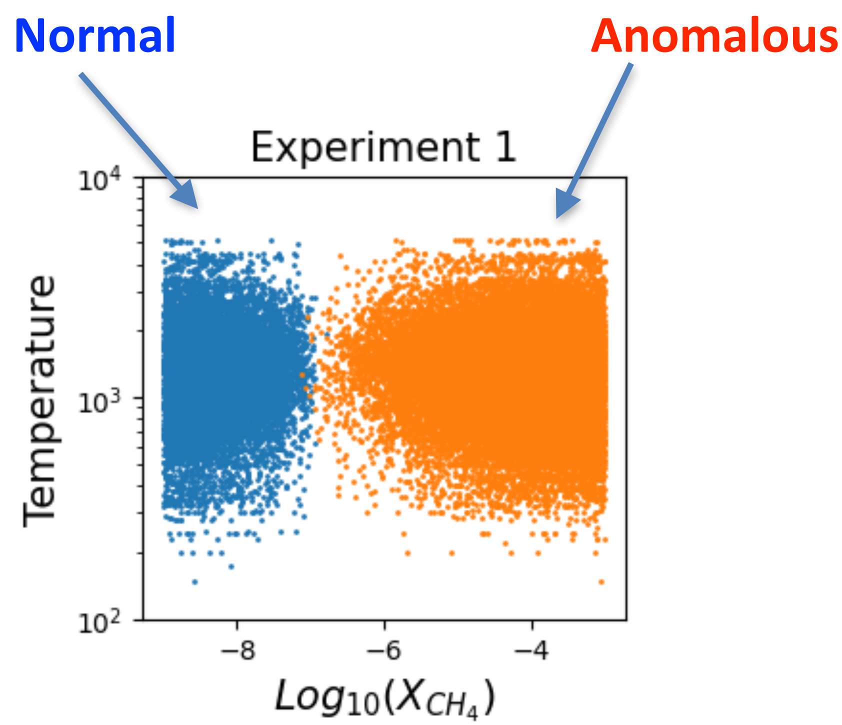


Defining "anomalous" atmospheres

- Since we do not know what types of surprises we can get, we want to **train** the model on **normal** samples only (white sheep) : **"novelty detection"**
- The **testing** is done on both **normal** and **anomalous** samples
- **Anomalous**: having an unexpected **mystery absorber**
    - Experiment 1: **CH_4**
- **Normal**: a mixture of the remaining four absorbers in the database, no mystery absorber
    - Experiment 1: **CO_2, H_2O, NH_3, CO**

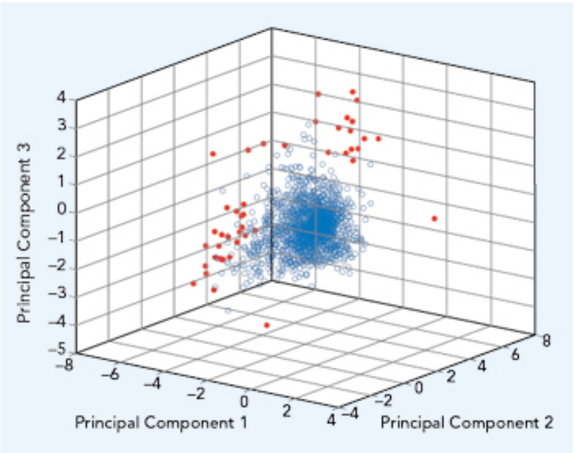


Log(Relative fraction of mystery absorber)

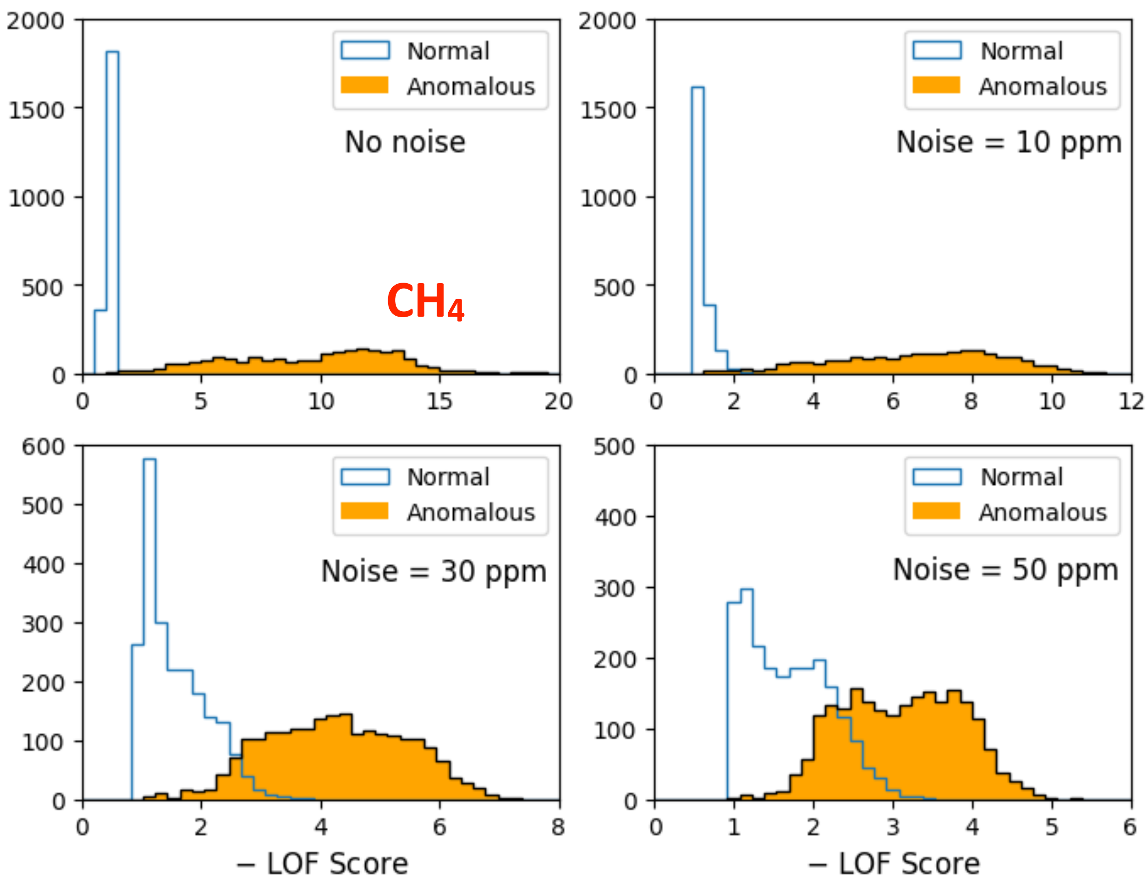


Log(Mixing ratio of mystery absorber)

Local Outlier Factor (LOF)



- Compares the **sample density** around a given point to the density around its neighbors
- Assigns an **LOF score**
  - small values (near zero) for **normal** samples
  - large values for **anomalous** samples
- The level of separation depends on the level of instrumental noise



ROC Curve

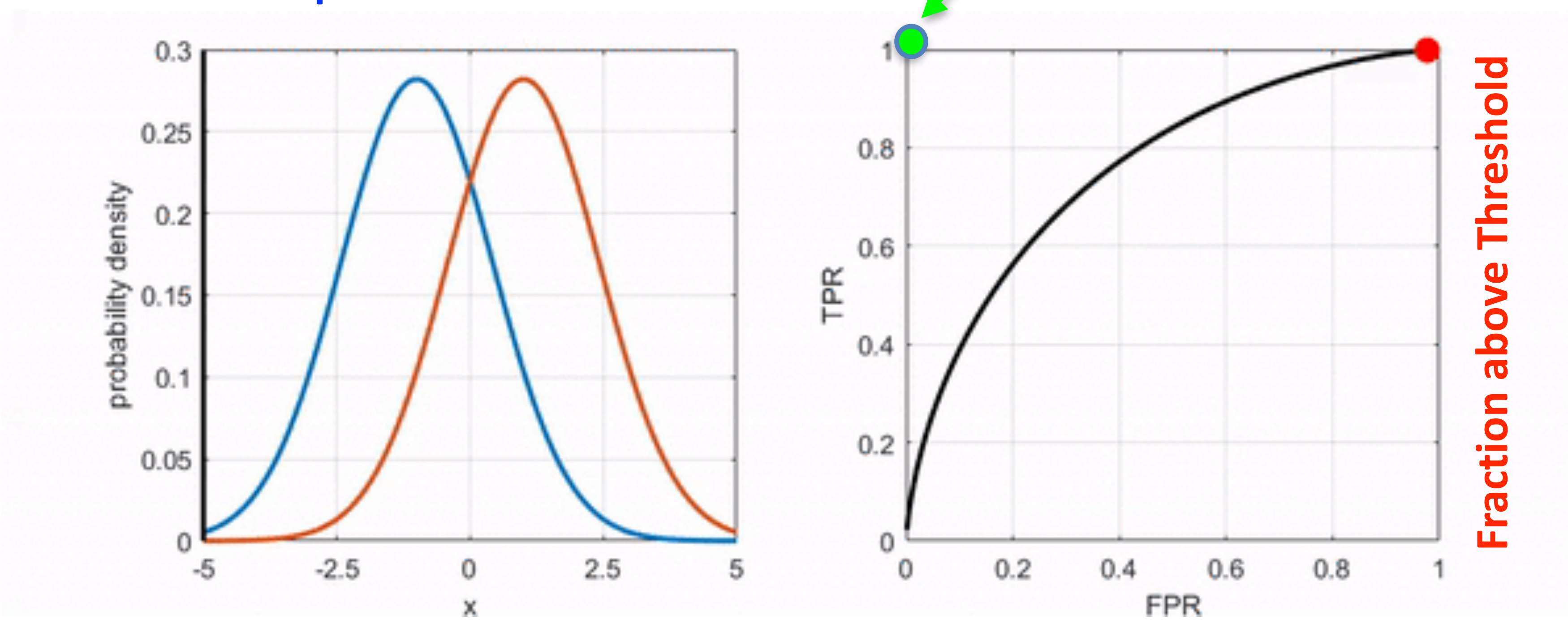
- A graph showing the **performance of a classifier** at all thresholds.
- Count the number of samples of each type to the right of the threshold

We want the ROC curve to be as close as possible to this point

Ideal classifier

Normal Samples

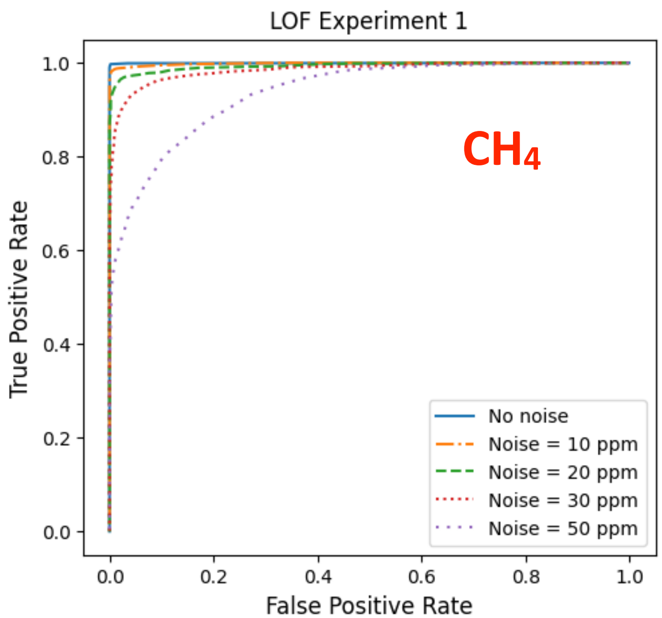
Anomalous Samples

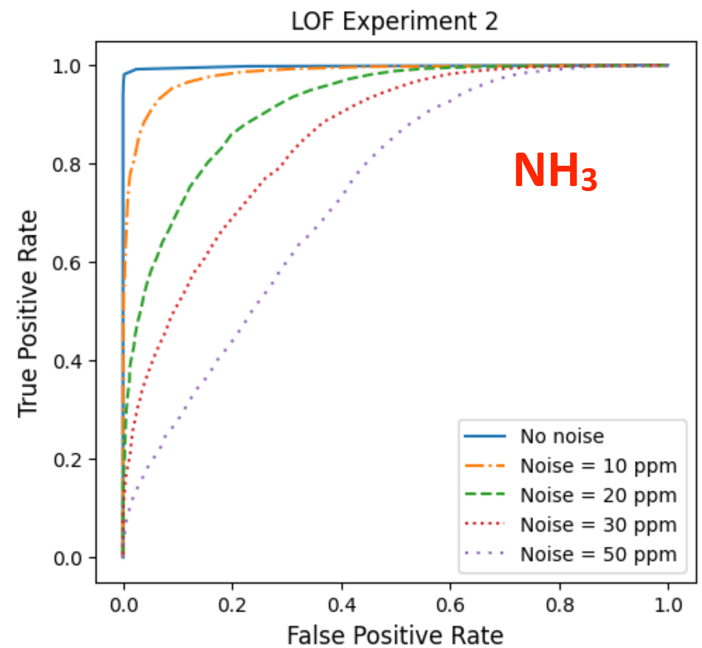


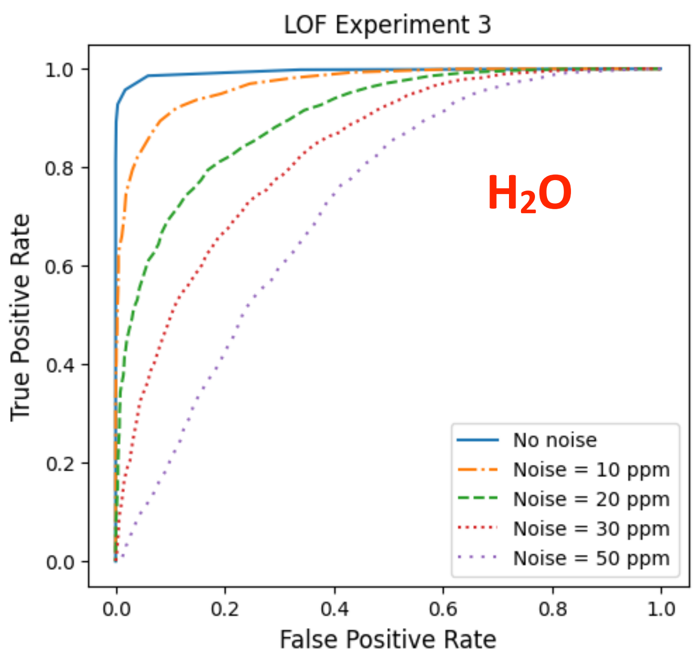
Fraction above Threshold

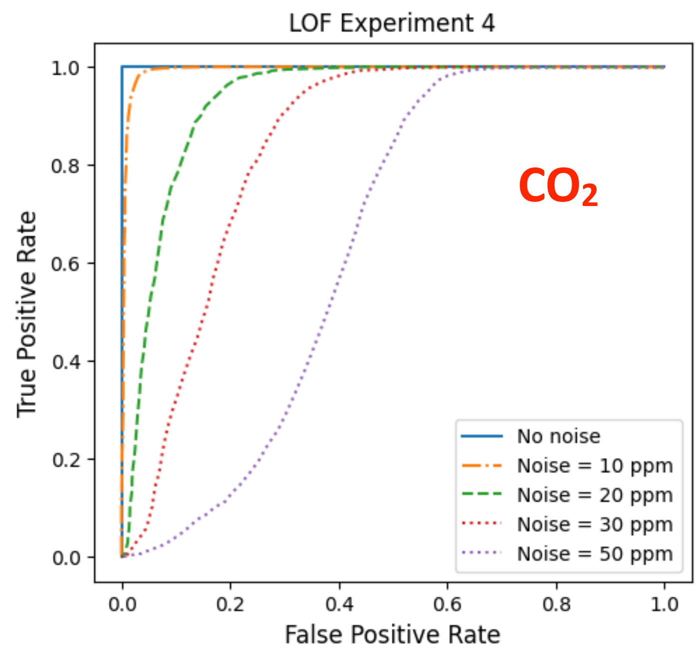
Fraction above Threshold

ROC curves for anomaly detection

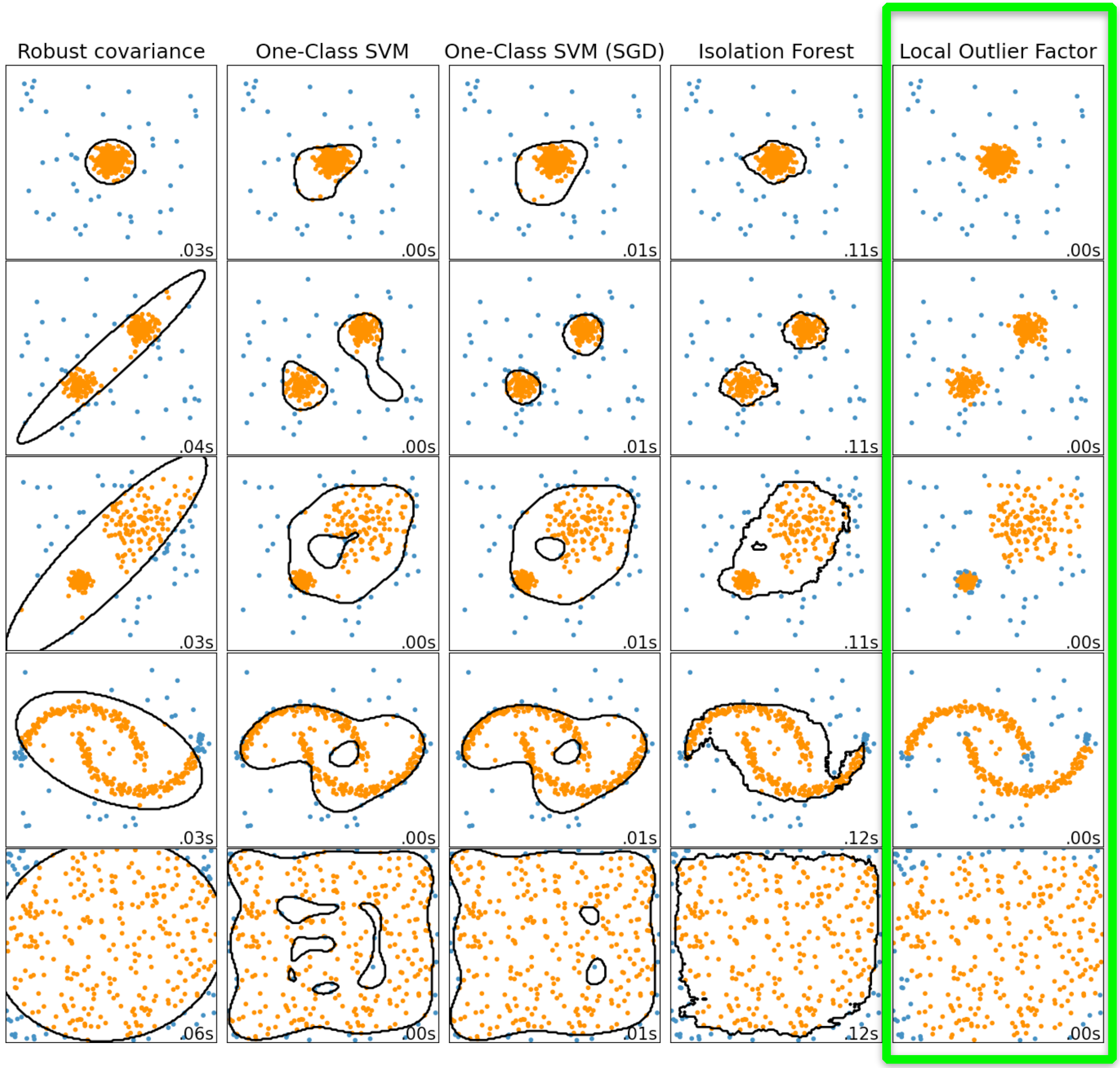








Anomaly Detection Methods



Looking for a needle in a haystack!

Searching for bio-signatures in spectroscopic data

- ✓ Know your haystack!
  Understand the data!

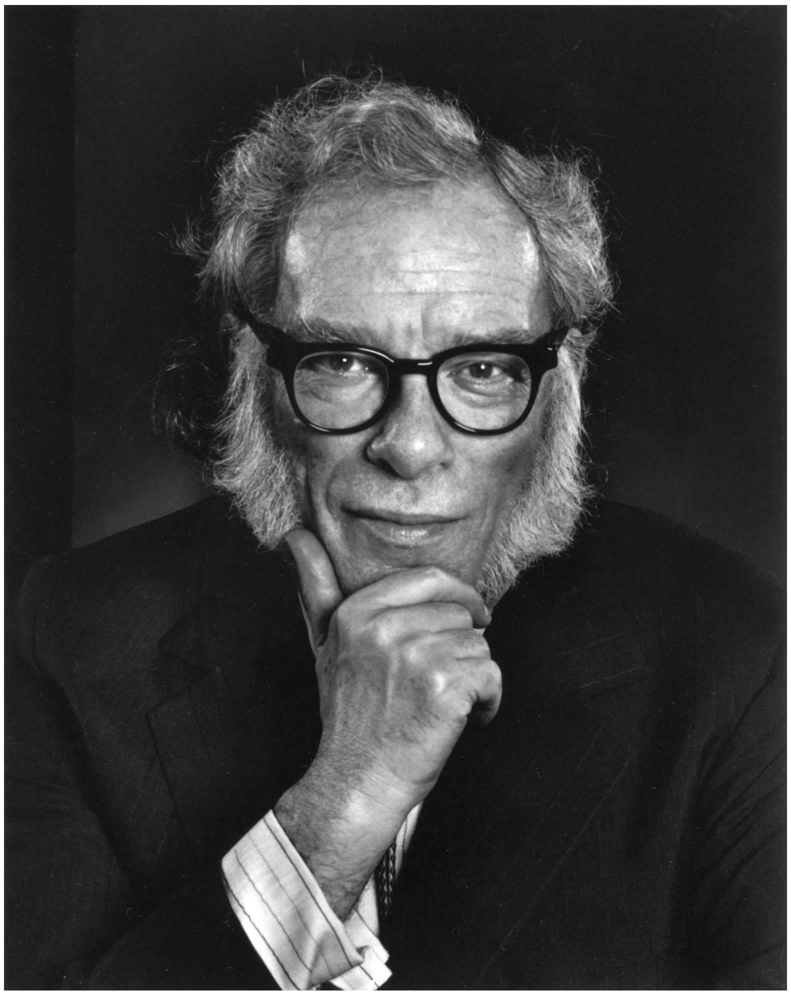
- ✓ Where to look?
  Dimensionality Reduction.

- ✓ What is the most contrasting property of the hay?
  Principal Component Analysis.

- ✓ Separate the stack in smaller distinct piles.
  Clustering and categorizing the data.

- ✓ Find the one that does not belong!
  Anomaly Detection

…The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but **"That's funny…"**



Isaac Asimov