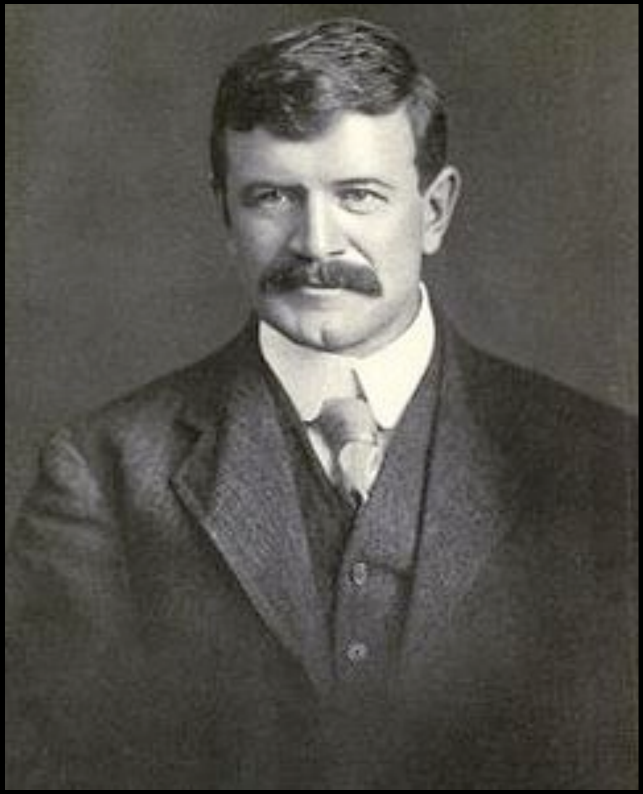
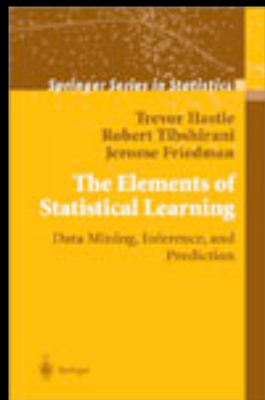
Lecture Two: Working with high dimensional data

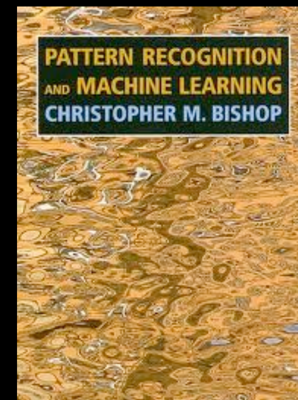


"In ancient times they had no statistics so they had to fall back on lies." **Stephen Leacock**

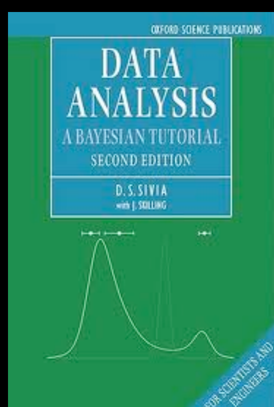
Recommended books



“The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Hastie et al



“Pattern Recognition and Machine Learning”, Bishop



“Data Analysis: A Bayesian Tutorial”, Sivia

Python based machine learning tool kit.

What is the science we want to do?

- **Finding the unusual**
  - Nova, supernova, GRBs
  - Source characterization
  - Instantaneous discovery
- **Finding moving sources**
  - Asteroids and comets
  - Proper motions of stars
- **Mapping the Milky Way**
  - Tidal streams
  - Galactic structure
- **Dark energy and dark matter**
  - Gravitational lensing
  - Slight distortion in shape
  - Trace the nature of dark energy

What are the operations we want to do?

- Finding the unusual
  - Anomaly detection
  - Dimensionality reduction
  - Cross-matching data
- Finding moving sources
  - Tracking algorithms
  - Kalman filters
- Mapping the Milky Way
  - Density estimation
  - Clustering (n-tuples)
- Dark energy and dark matter
  - Computer vision
  - Weak Classifiers
  - High-D Model fitting

Science is driven by precision we need to tackle issues of complexity:

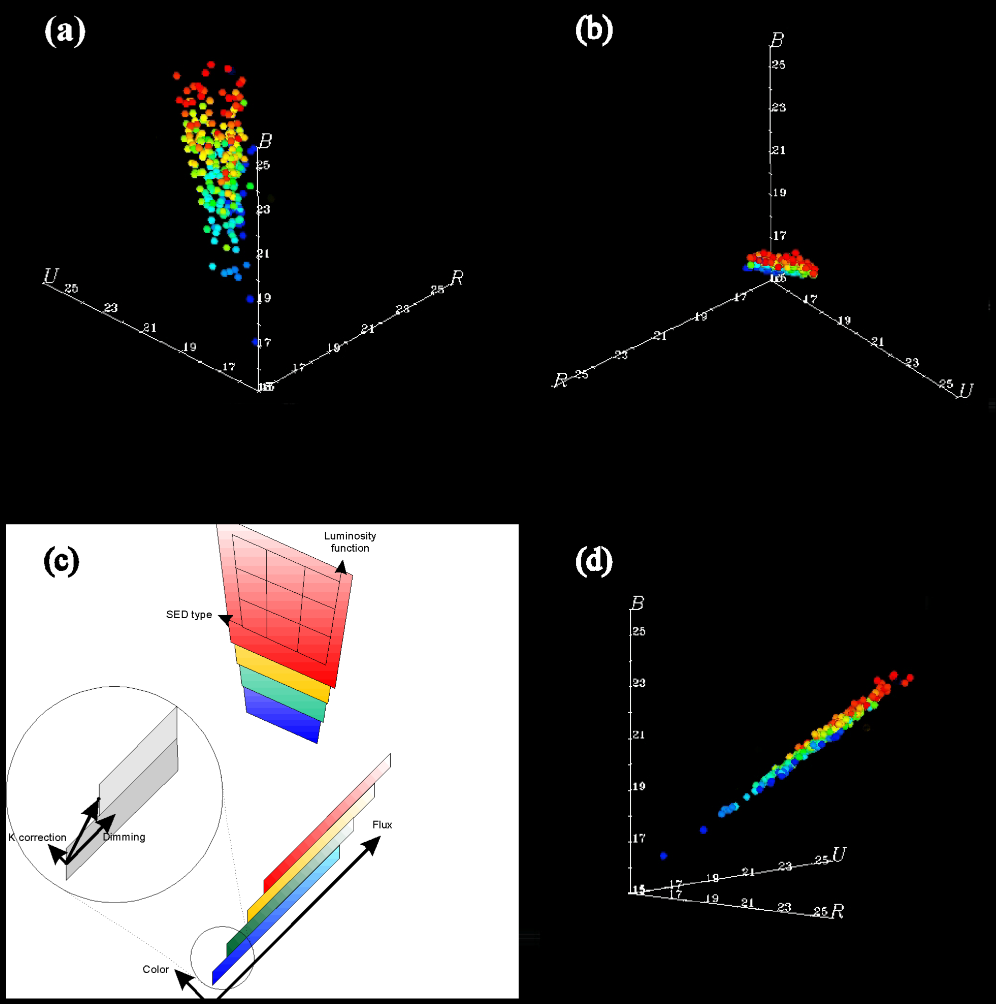
1. **Complex models of the universe**
   What is the density distribution and how does it evolve
   What processes describe star formation and evolution

2. **Complex data streams**
   Observations provide a noisy representation of the sky

3. **Complex scaling of the science**
   Scaling science to the petabyte era
   Learning how to do science without needing a CS major

There are no black boxes

How complex is our view of the universe?



We can measure many attributes about sources we detect…

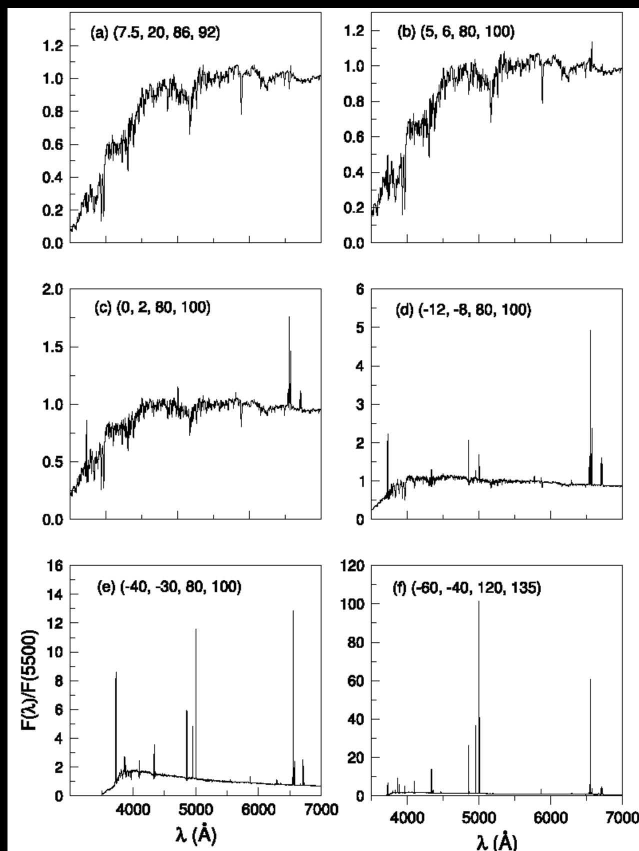
… which ones are important and why (what is the dimensionality of the data and the physics)

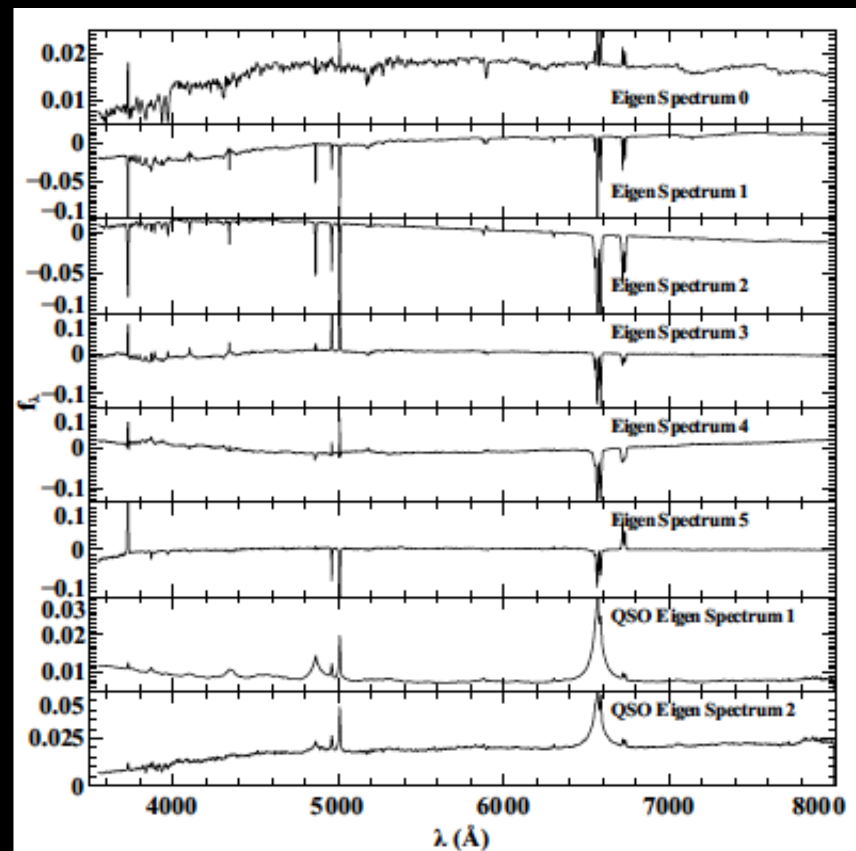
Connolly et al 1995

Low dimensionality remains even with more complex data

Old

Young



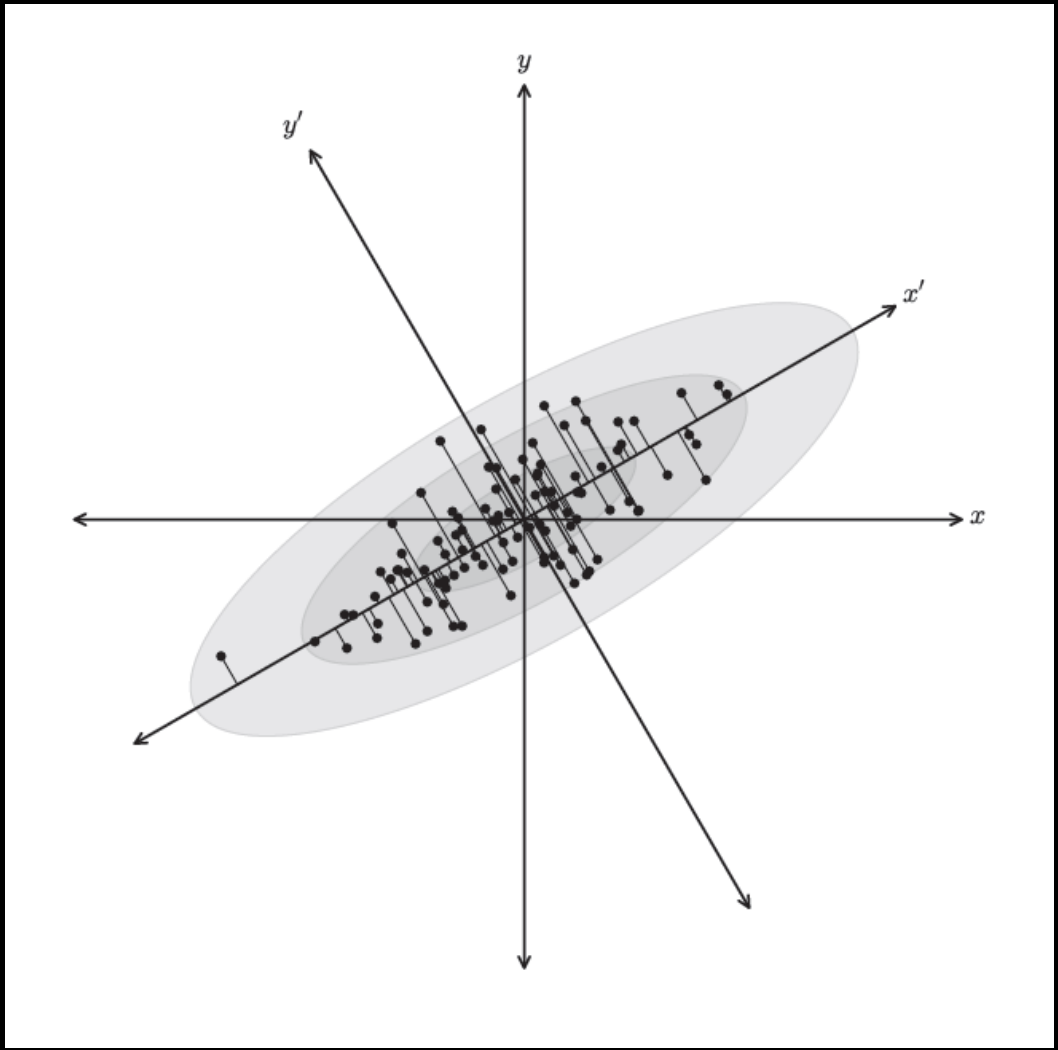


4000-dimensional (λ's)

$$f(\lambda) = \sum_{i<N} a_i e_i(\lambda)$$

10 components Ξ >99% of variance

Principal Components



PCA in a Nutshell

- We can define a covariance matrix for the data (centered)

$$C_X = \frac{1}{N-1} X^T X$$

- We want a new set of axes where the covariance matrix is diagonal

$$C_Y = \frac{1}{N-1} Y^T Y,$$

- What is the appropriate transform?

$$C_X = R^T C_Y R$$

Simply the definition of an eigensystem

PCA in a Nutshell

- Singular Valued Decomposition decomposes a matrix as

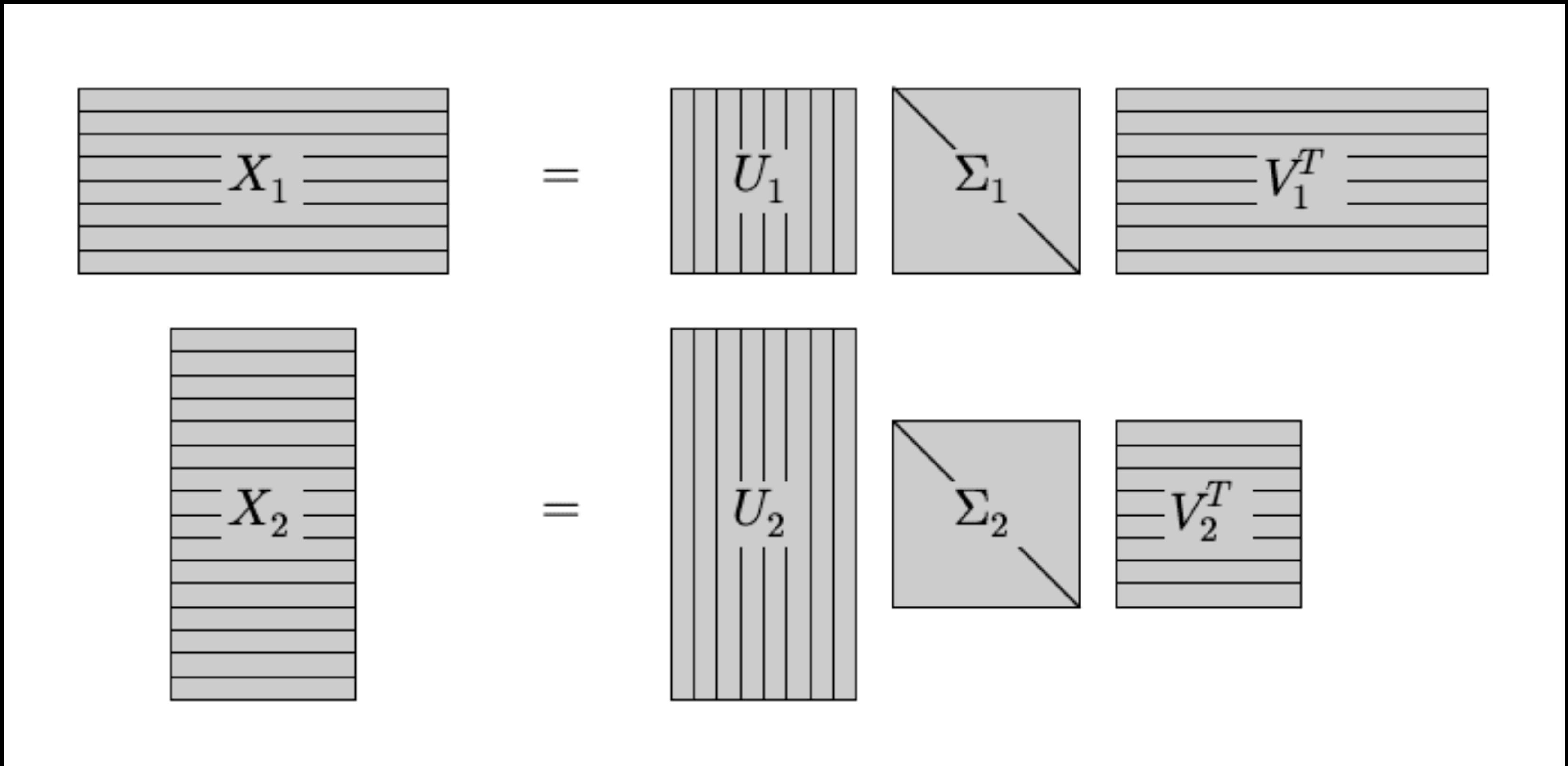
$$X = U\Sigma V^T$$

- Decomposing the correlation matrix

$$\begin{aligned} C_X &= \frac{1}{N-1} X^T X \\ &= V\Sigma U^T U \Sigma V^T \\ &= V\Sigma^2 V^T, \end{aligned}$$

- We see that V=R and so SVD results in the eigenvectors of the system

Quick note on speed



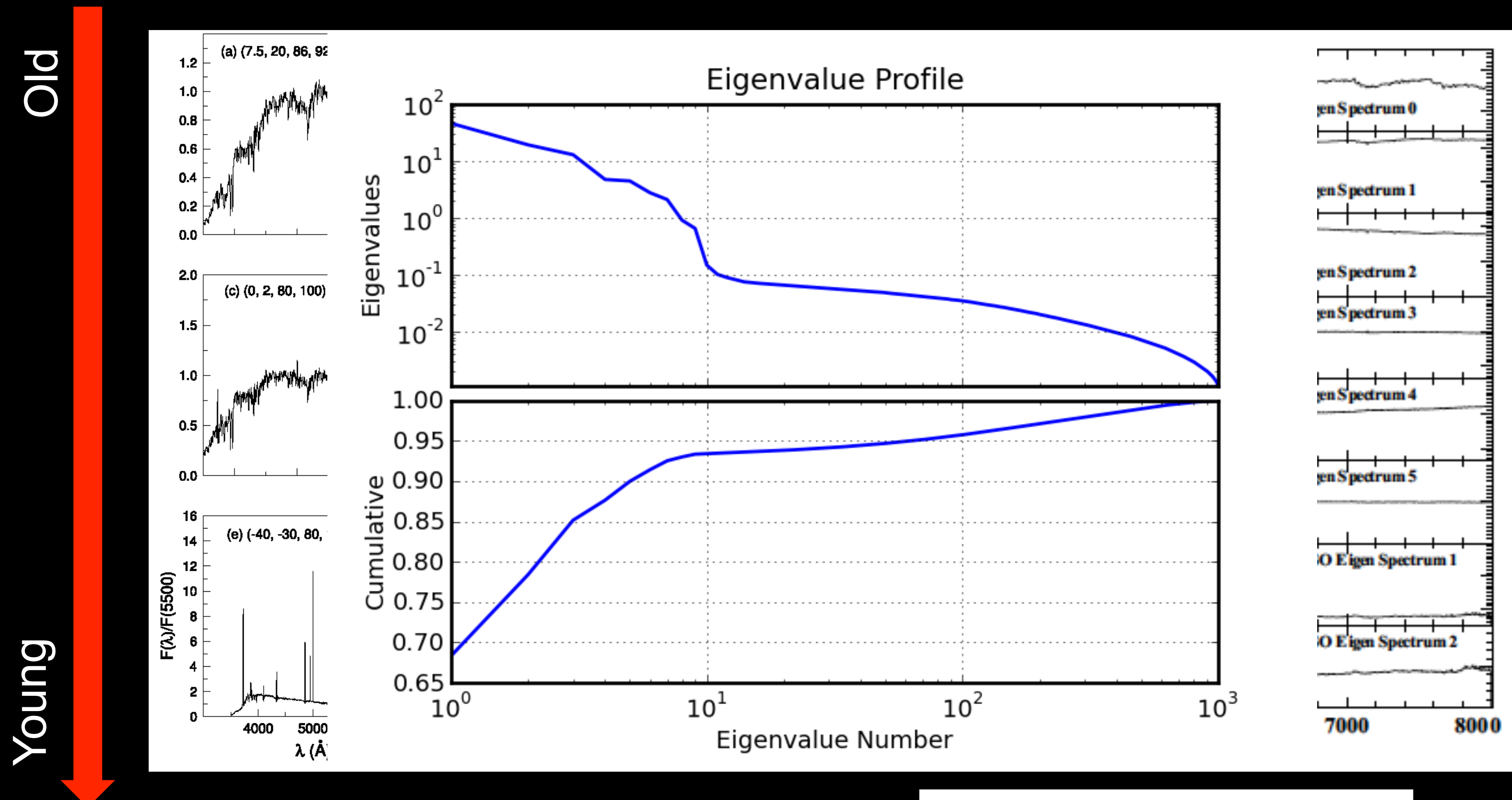
$X^T X = V\Sigma^2 V^T$ Is equivalent to $X X^T = U\Sigma^2 U^T$

Use the covariance or correlation matrix depending on the rank of the system

PCA with Python

```
from sklearn.decomposition import RandomizedPCA

n_components = 5
# Compute PCA components

spec_mean = spectra.mean(0)

# use randomized PCA for speed
pca = RandomizedPCA(n_components - 1)
pca.fit(spectra)
pca_comp = np.vstack([spec_mean,
                      pca.components_])
```

Low dimensionality remains even with more complex data

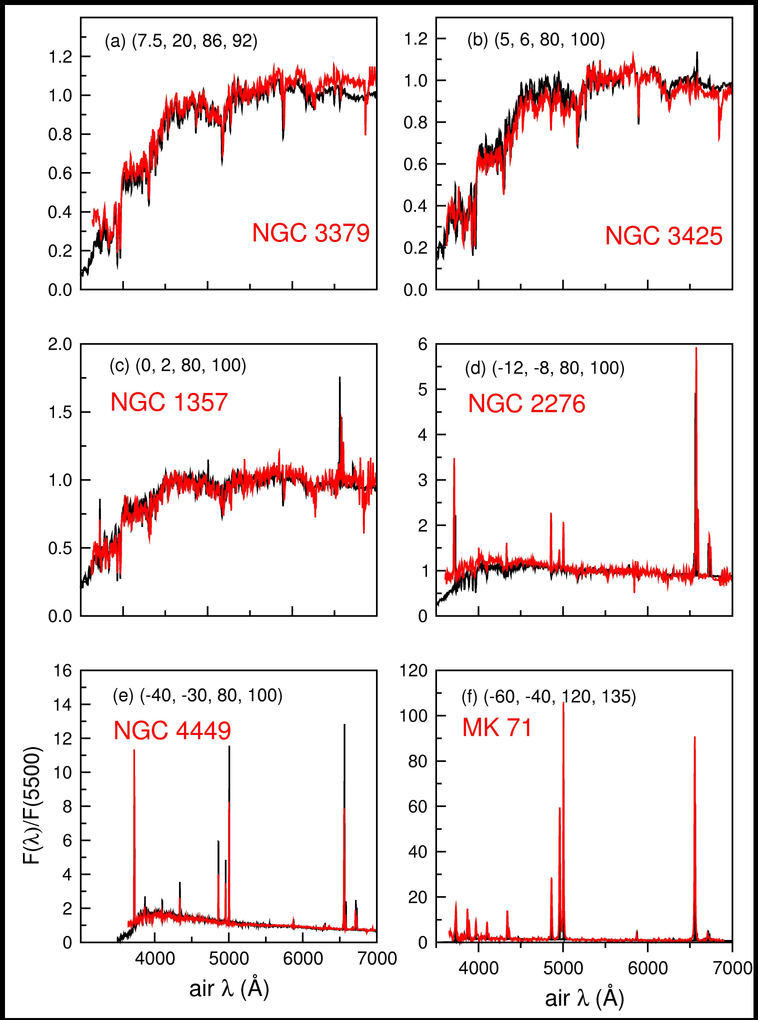


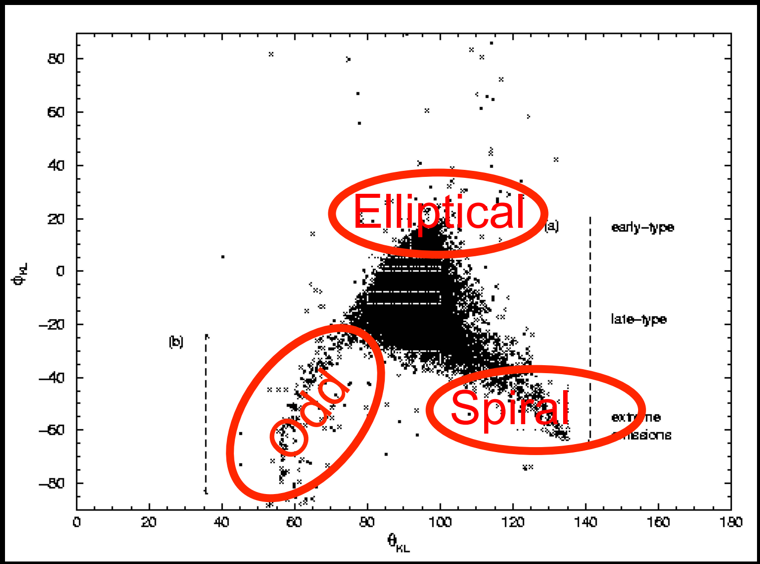
4000-dimensional (λ's)

$$f(\lambda) = \sum_{i<N} a_i e_i(\lambda)$$

10 components Ξ >99% of variance

Dimensionality relates to physics





400-fold compression
Signal-to-noise weighted
Accounts for gaps and noise
Compression contains physics
Not good at non-linear features

Independent Component Analysis

The cocktail party problem

$$\begin{aligned}
x_1(\lambda) &= a_{11}s_1(\lambda) + a_{12}s_2(\lambda) + a_{13}s_3(\lambda) + ... \\
x_2(\lambda) &= a_{21}s_1(\lambda) + a_{22}s_2(\lambda) + a_{23}s_3(\lambda) + ... \\
x_3(\lambda) &= a_{31}s_1(\lambda) + a_{32}s_2(\lambda) + a_{33}s_3(\lambda) + ...
\end{aligned}$$

We want to extract the independent components (to find the mixing matrix W)

$$S(\lambda) = WX(\lambda)$$

Statistical independence

$$f(x^p, y^q) = f(x^p)f(y^q)$$

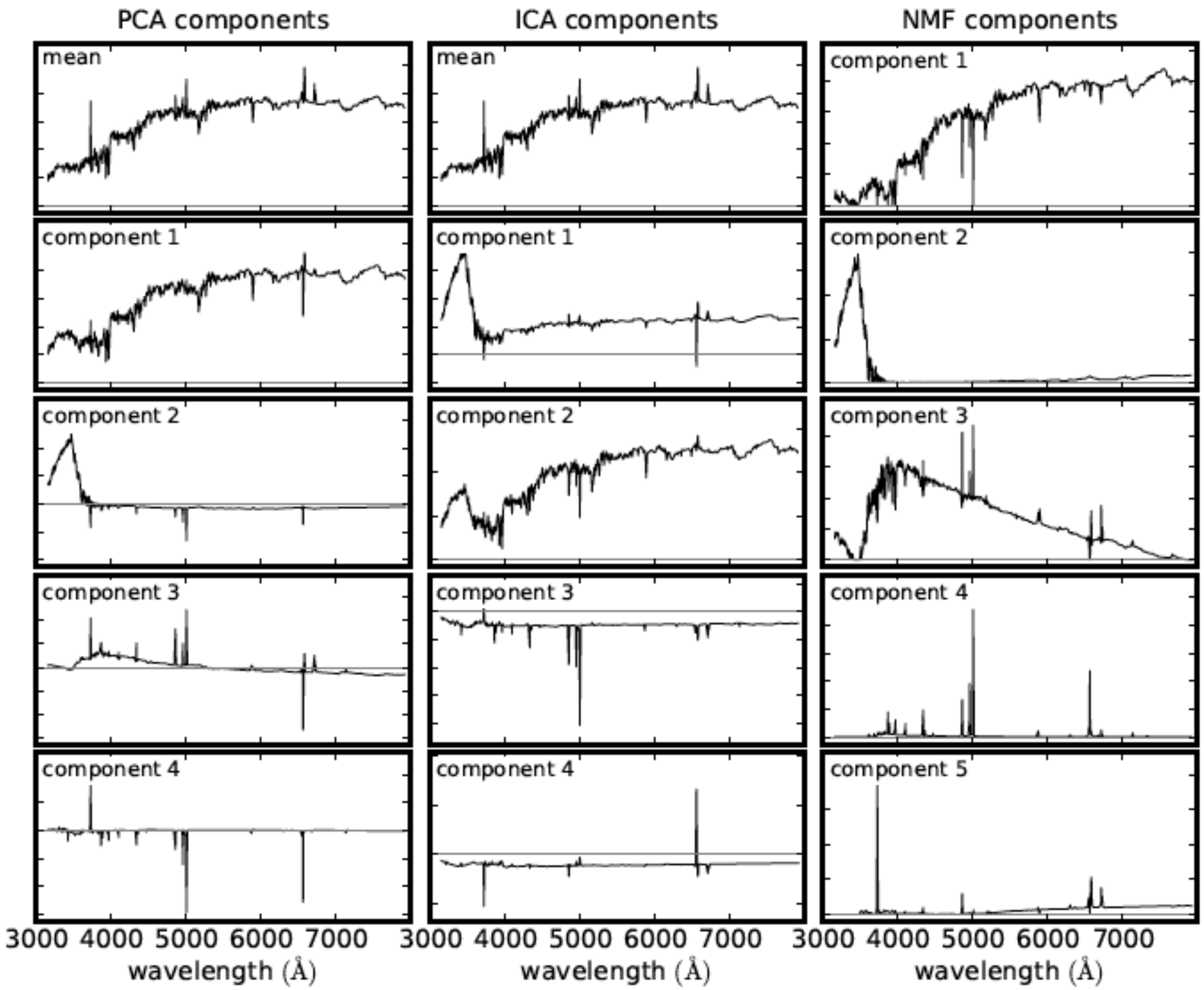
For PCA p=q=1

Search for non-Gaussian signal with the rationale being that the sum of two independent random variables will be more Gaussian that either individual component.

Non-Gaussianity defined by Kurtosis and negentropy,

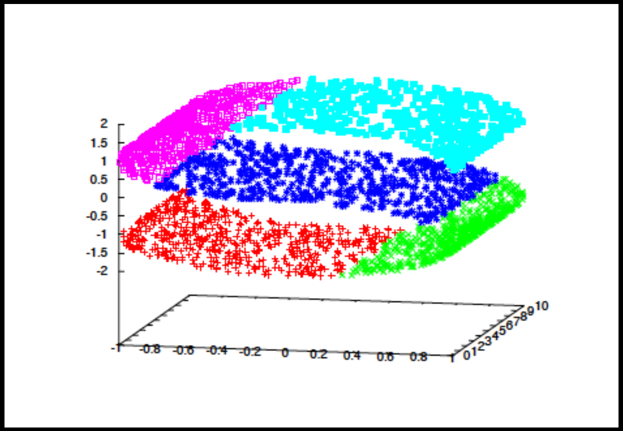
ICA in Python

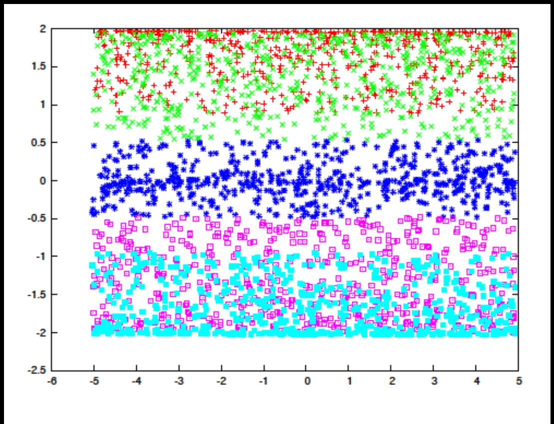
```
from sklearn.decomposition import FastICA

n_components = 5

# ICA treats sequential observations as related.
# Because of this, we need to fit with the transpose of the spectra
ica = FastICA(n_components - 1)
ica.fit(spectra.T)
ica_comp = np.vstack([spec_mean, ica.transform(spectra.T).T])
```



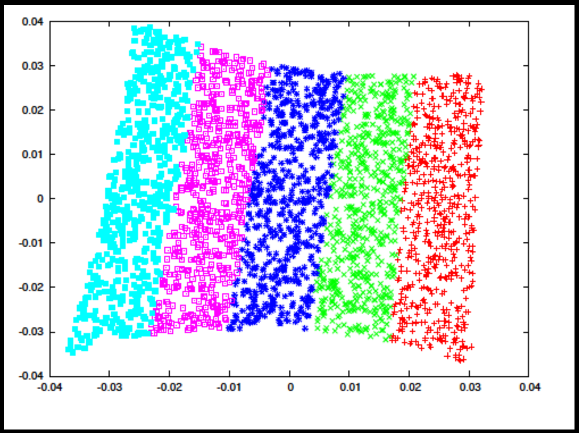
PCA components
ICA components
NMF components
mean
component 1
component 2
component 3
component 4
component 5
3000 4000 5000 6000 7000
wavelength (Å)

Responding to non-linear processes





PCA



LLE

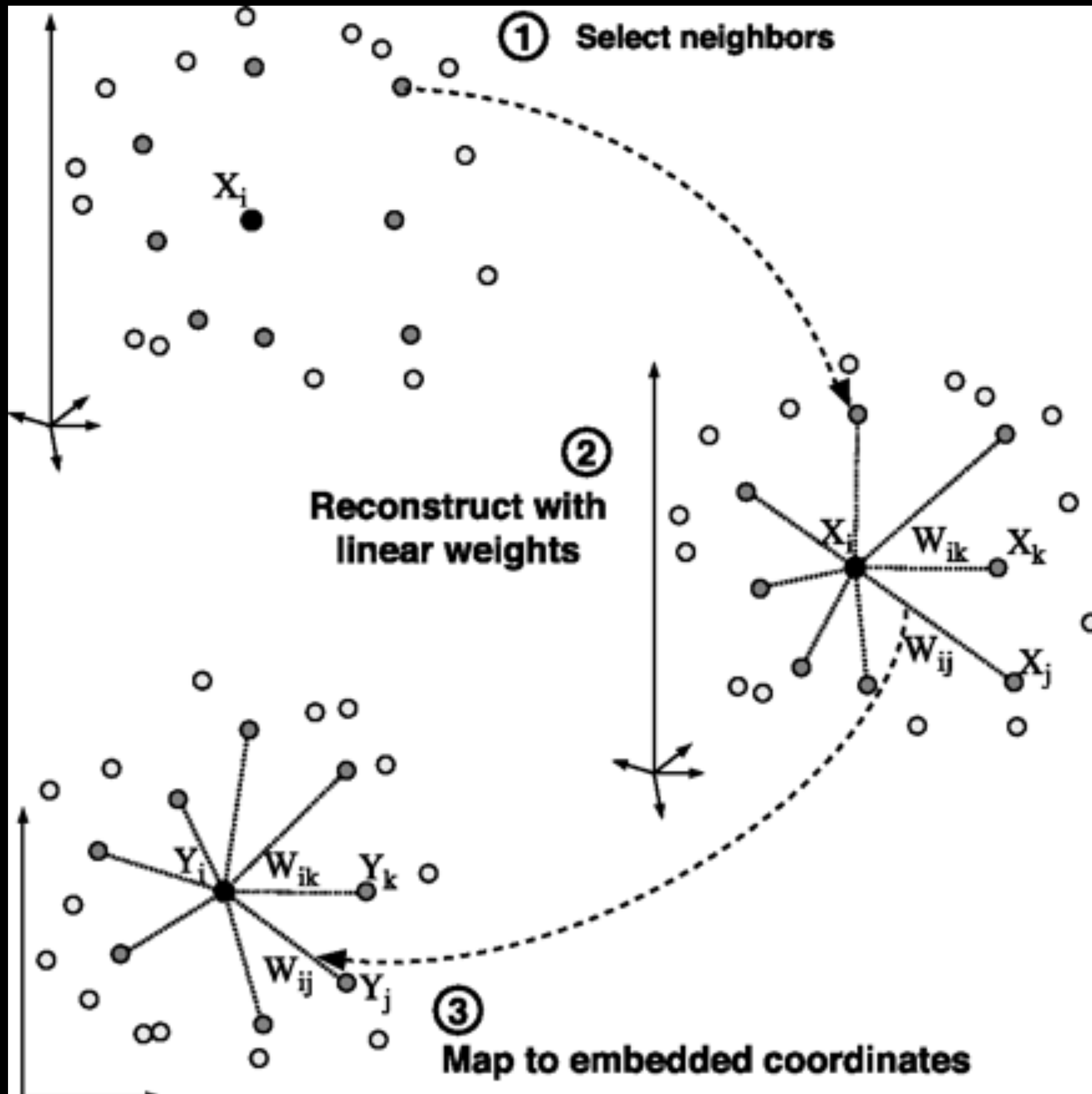
Local Linear Embedding (Roweis and Saul, 2000)

$$\mathcal{E}_1^{(i)}(\mathbf{w}^{(i)}) = \left| \mathbf{x_i} - \sum_{j=1}^{K} w_j^{(i)} \mathbf{x}_{n_j^{(i)}} \right|^2$$

$$\mathcal{E}_2(\mathbf{Y}) = \sum_{i=1}^{N} \left| \mathbf{y_i} - \sum_{j=1}^{K} w_j^{(i)} \mathbf{y}_{n_j^{(i)}} \right|$$

Preserves local structure

Slow and not always robust to outliers

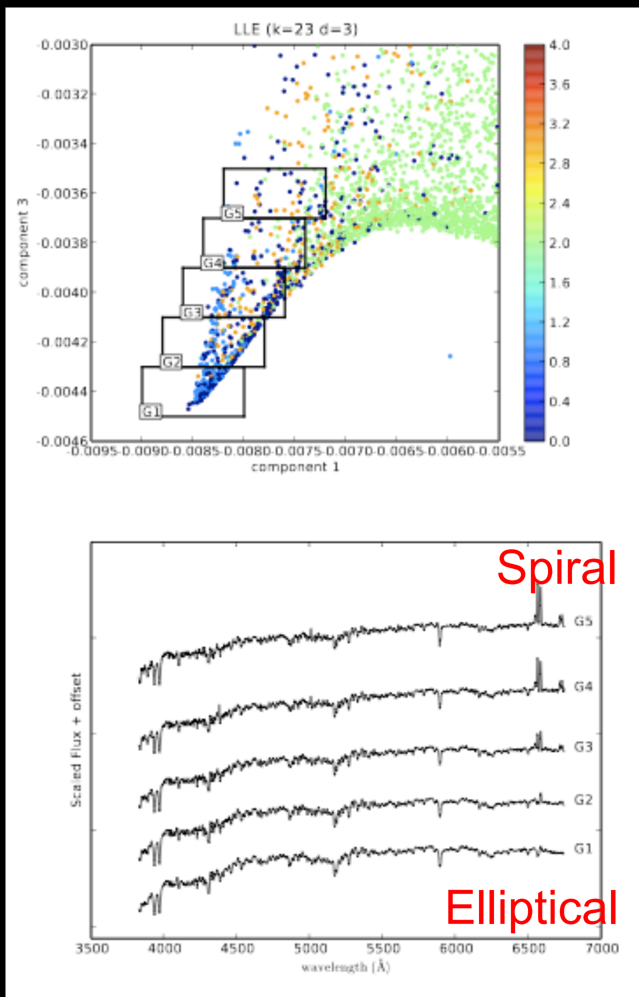


1 Select neighbors
X_i
2 Reconstruct with linear weights
X_i
W_{ik}
X_k
W_{ij}
X_j
Y_i
W_{ik}
Y_k
W_{ij}
Y_j
3 Map to embedded coordinates

LLE with Python

```
from sklearn import manifold, neighbors
n_neighbors = 10
out_dim = 3

LLE = manifold.LocallyLinearEmbedding(n_neighbors, out_dim,
                                      method='modified',
                                      eigen_solver='dense’)
Y_LLE = LLE.fit_transform(spec_train)

flag = flag_outliers(Y_LLE, nsig=0.25)
coeffs = Y_LLE[~flag]
```

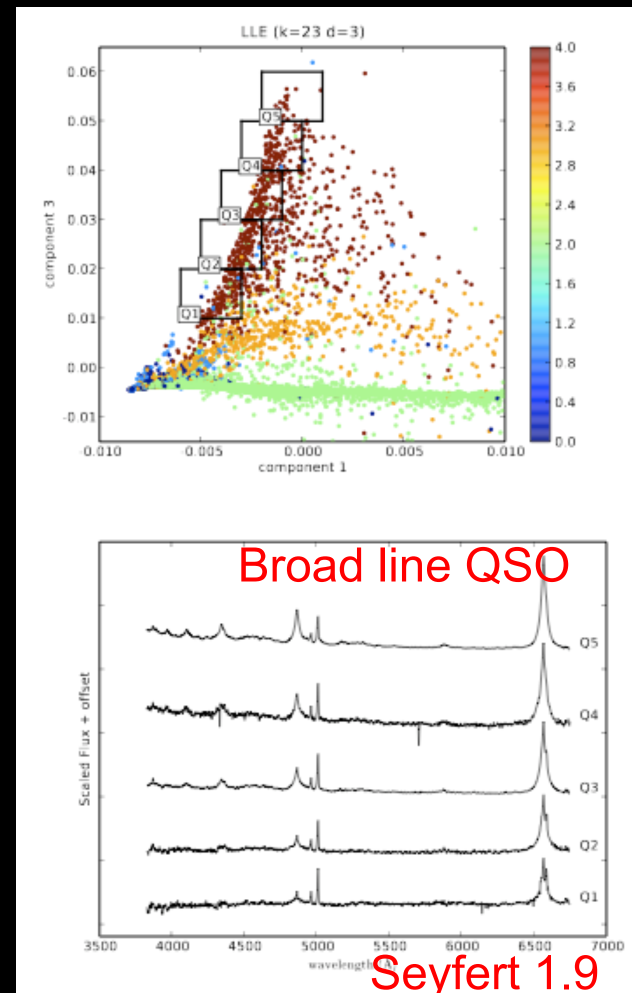
A compact representation accounting for broad lines



No preprocessing

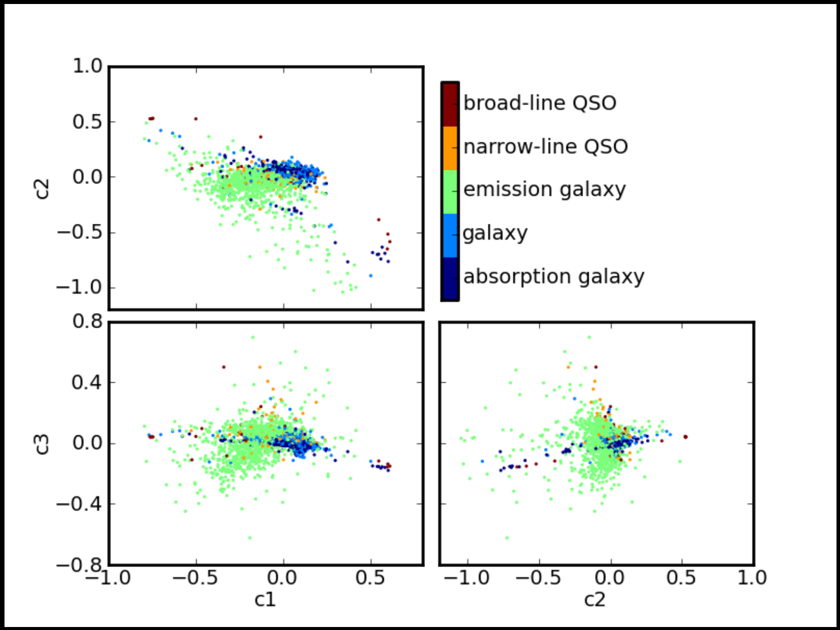
Continuous Classification

Maps to a physical space

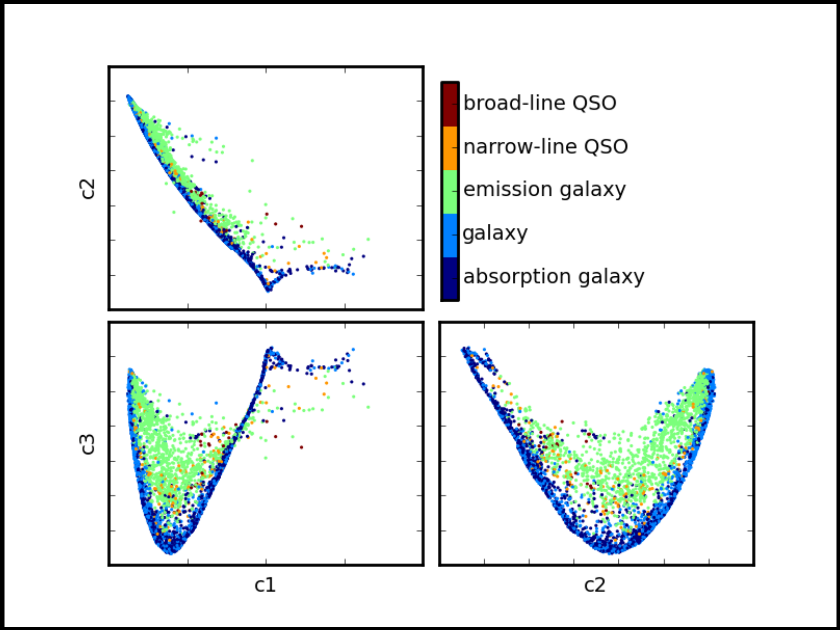


VanderPlas and Connolly 2009

PCA vs LLE

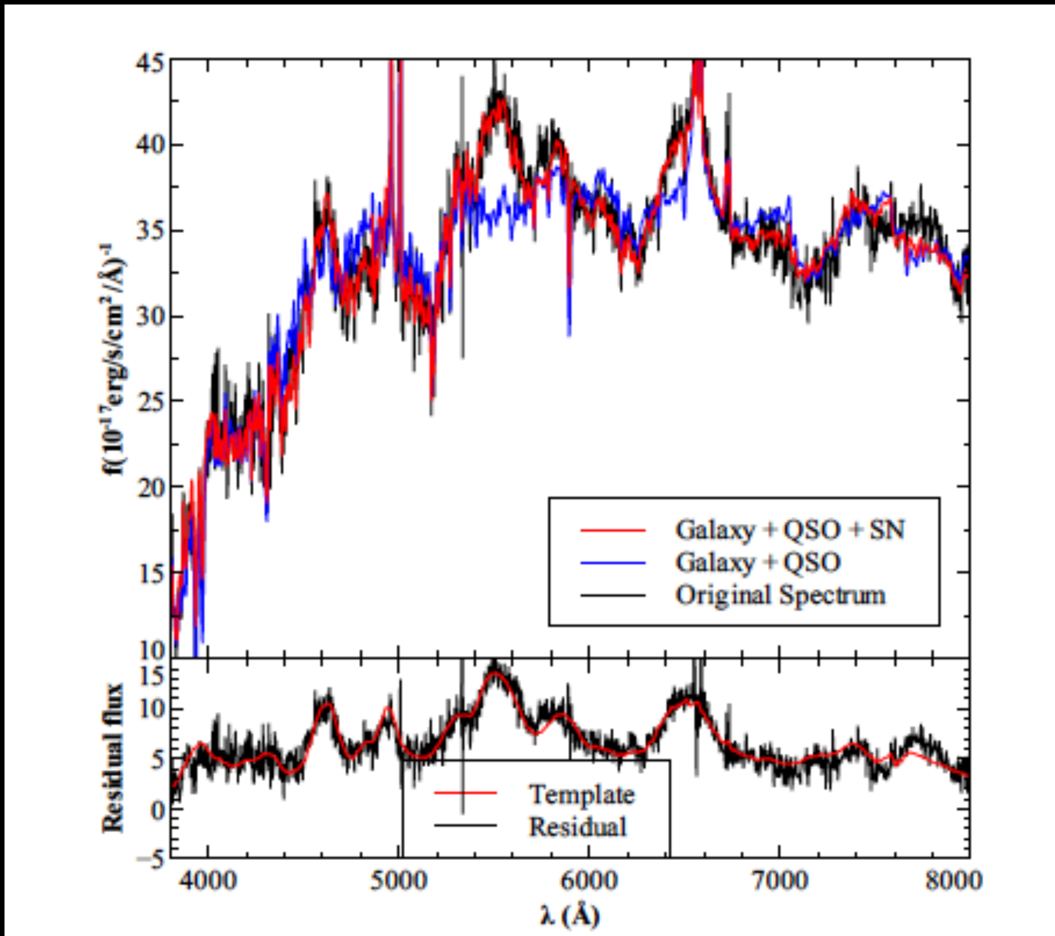


PCA



LLE

Using structure to detect outliers



Type Ia supernovae
0.01% contamination
to SDSS spectra

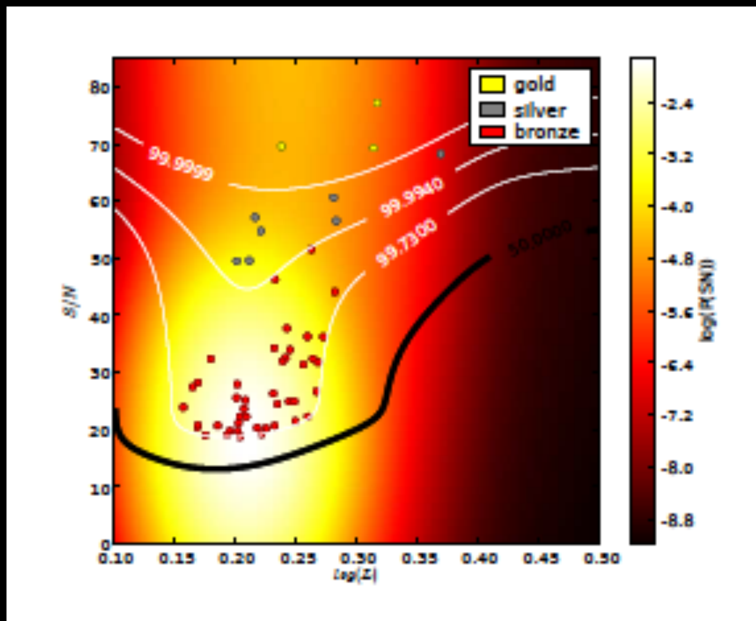
Type Ia supernovae
Visible for long
(-15 to 40 days)

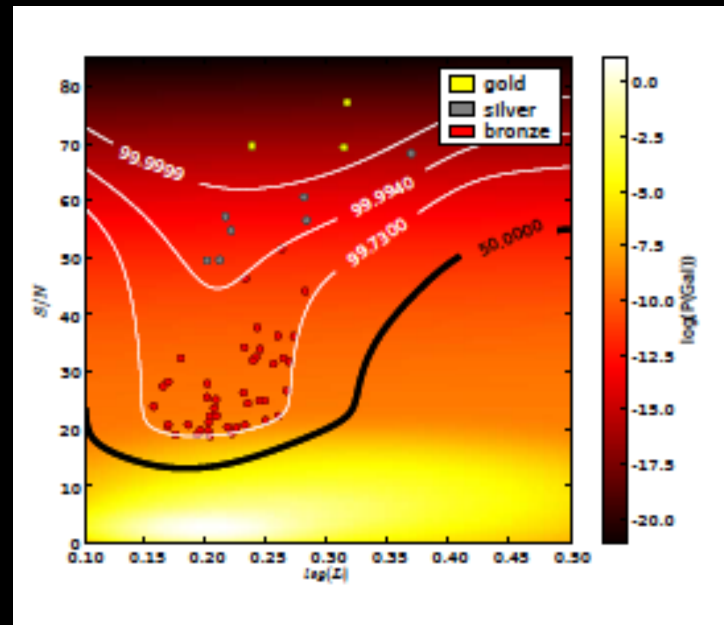
Well defined spectral
signatures

Magwick et al 2003

$$SN(\lambda) = f(\lambda) - \sum_{i<N} a_i e_{g_i}(\lambda) - \sum_{i<N} q_i e_{q_i}(\lambda)$$

Bayesian Classification of outliers

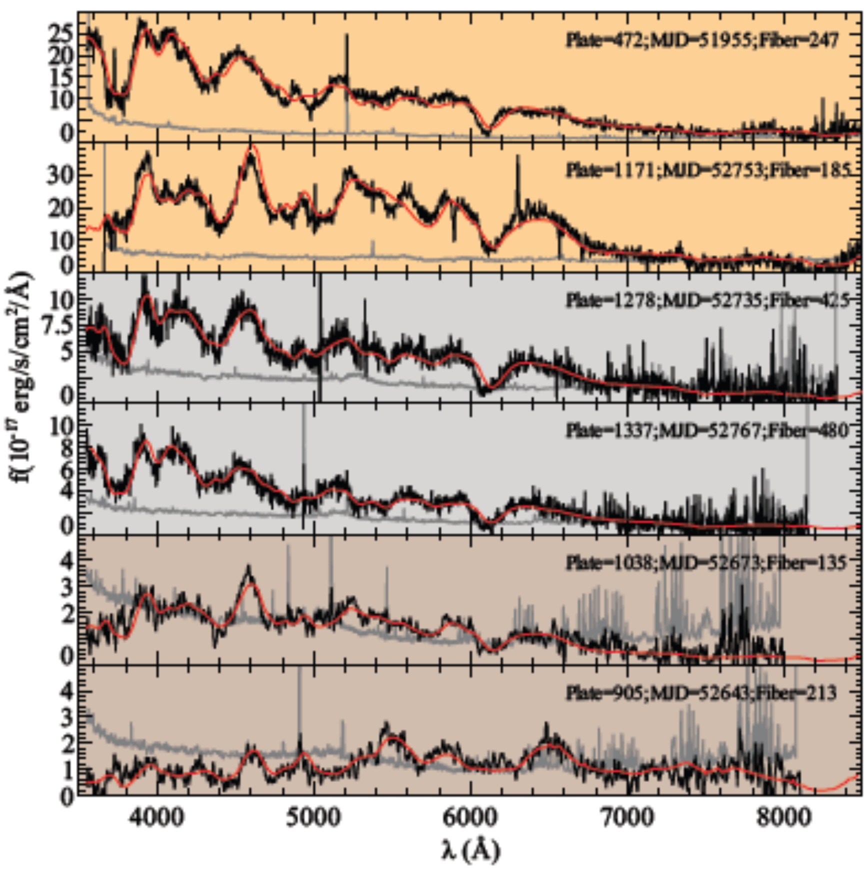


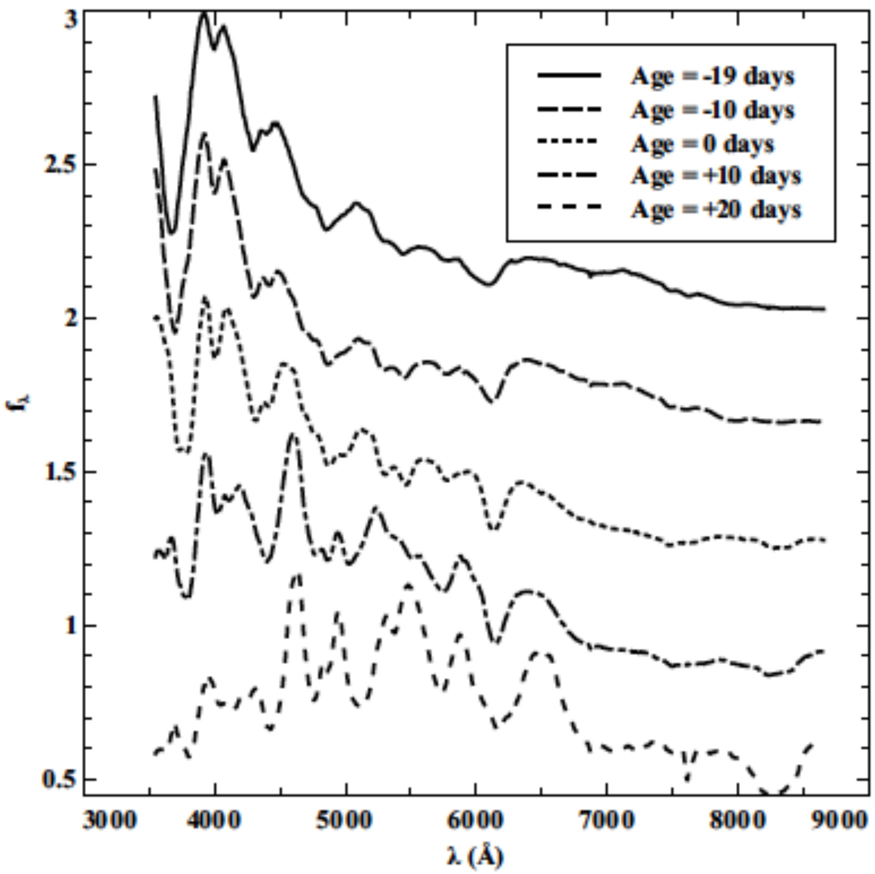


$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$

Density estimation using a mixture of Gaussians gives P(x|C): likelihood vs signal-to-noise of anomaly

Probabilistic identification with no visual inspection



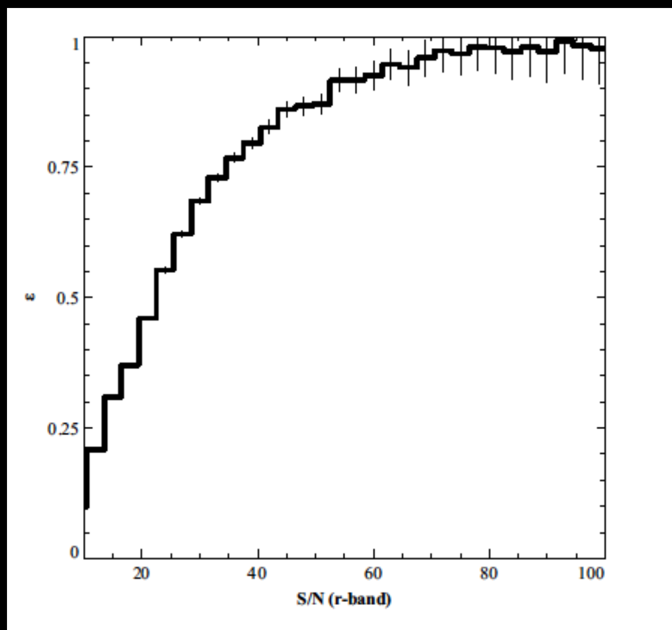


Krughoff et al 2011

Nugent et al 1994

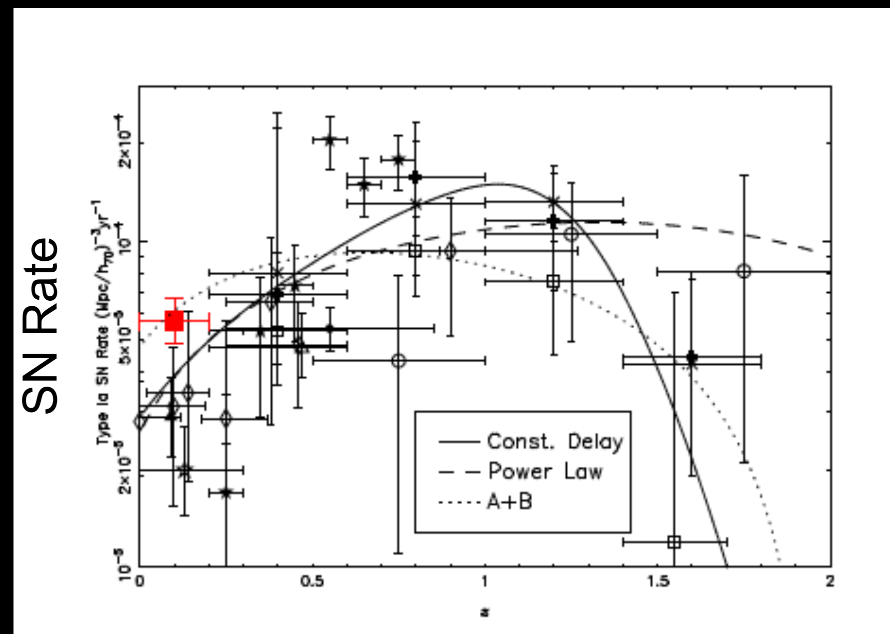
A serendipitous way to measure supernovae rates

Efficiency



S/N galaxy

SN Rate

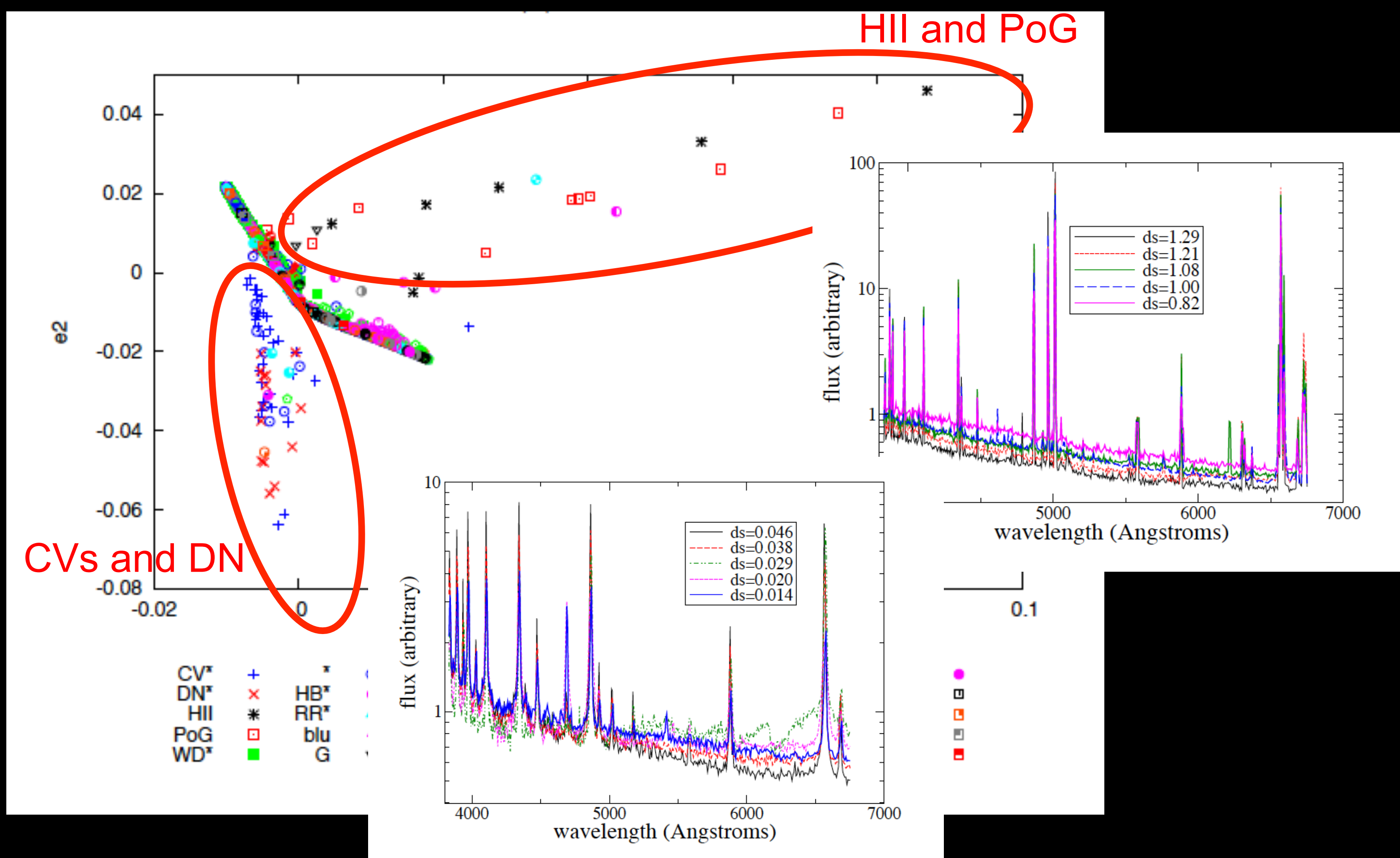


Redshift

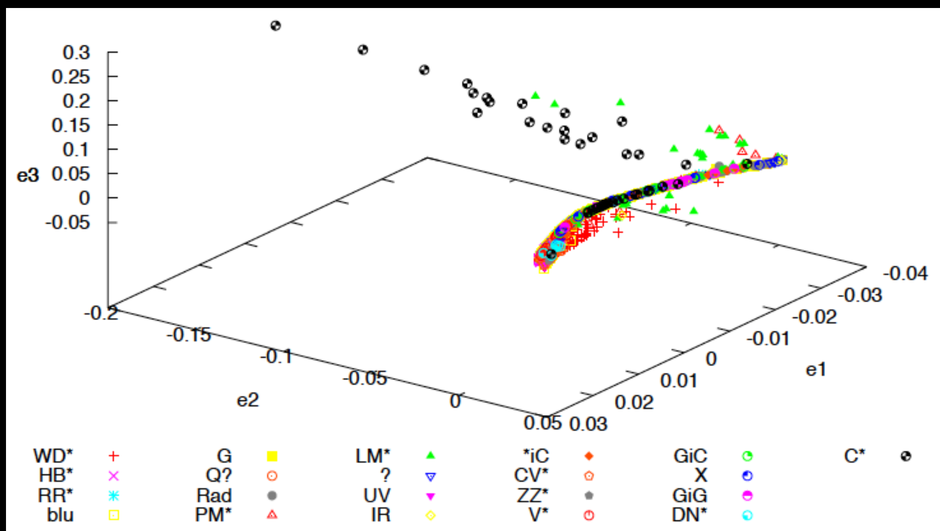
350K SDSS spectra, 52 SN Ia, $z \sim 0.1011$

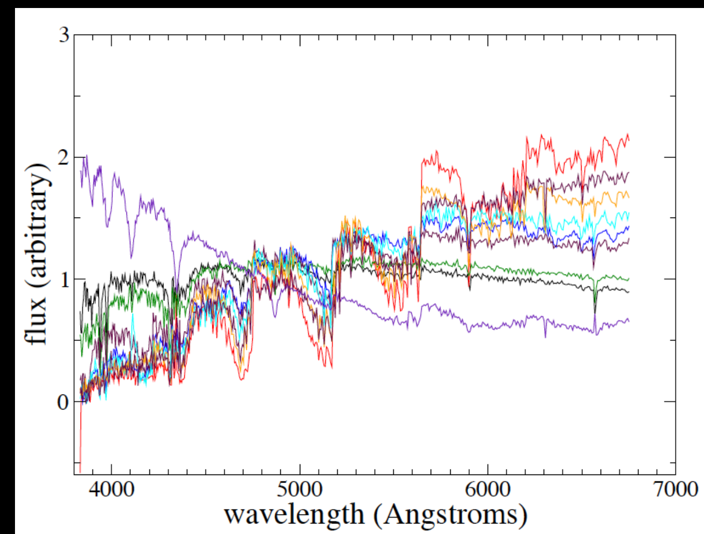
0.470 ± 0.08 Snu (1 SNu = 10^{10} $L_\odot$ per century)

How to find anomalies when we don't have a model for them



Anomaly discovery from a progressive refinement of the subspace





Outliers impact the local subspace determination (dependent on number on nearest neighbors). Progressive pruning identifies new components (e.g. Carbon stars).

Need to decouple anomalies from overall subspace

Quantifying the outliers and subspaces

Decompose into principal subspace and noise subspace (SVD)

$$x_i = \sum_{j=1}^{k} u_j s_j v_{ij} + \sum_{j=k+1}^{d} u_j s_j v_{ij}$$

Accumulate the errors given a truncation (or over all truncations)

$$\varepsilon_{ad}(x_i) = \sum_{1}^{d} \sum_{j=k+1}^{d} \frac{s_j^2 v_{ij}^2}{s_j^2 / n}$$

Extend to non negative matrix factorization (a more physical basis)

$$U,V = \arg\min_{U,V} \| X - U^T V \|^2, U \ge 0, V \ge 0$$

Robust low rank detectors

Decompose into Gaussian noise and outliers

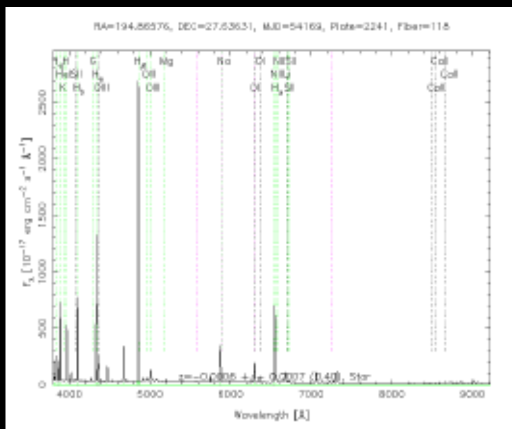
$$X = U^T V + E + O$$

Mixed matrix factorization (iteratively decompose matrix then solve for outliers). Using the L_1 norm as the error measure

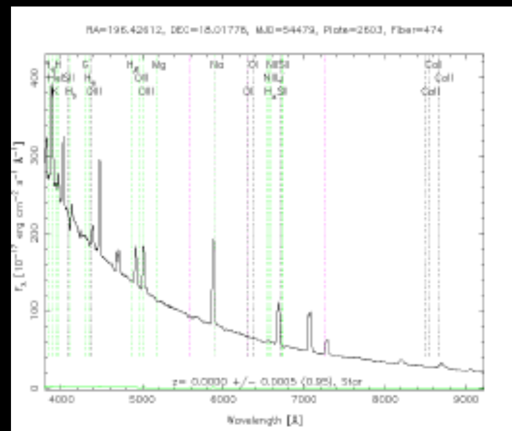
$$\min_{U,V,O} \frac{1}{2} \| X - U^T V - O \|^2 + \lambda \| O \|_r$$

How to choose λ is an open question (set to produce % of outliers)

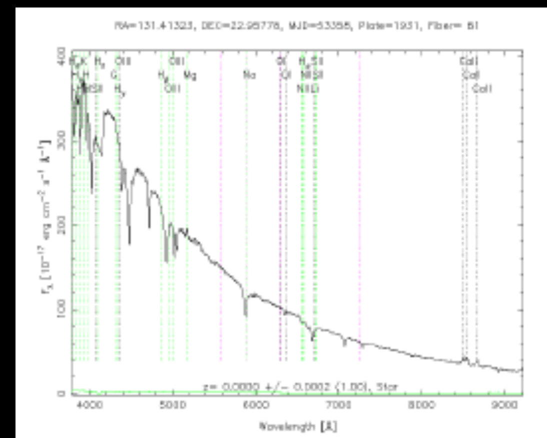
Anomalies within the SDSS spectral data



PN G049.3+88.1
Ranked first
Expect 1-3 PNE
Found 2



CV-AM
2 orbiting WDs
Ranked top 10



WD with debris disk
Ranked top 30
Only 3 known in SDSS

Xiong et al 2011

Expert user tagging (http://autonlab.org/sdss)

Xiong et al 2011