COIN'S

Dracula

Dimensionality Reduction And Clustering
for Unsupervised Learning in Astronomy

DRACULA

Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy

arXiv:1512.06810v1

Sasdelli et al., MNRAS,2016, 461, Issue 2, p.2044-2059

[ascl:1512.009]

Emille E. O. Ishida

Laboratoire de Physique Corpusculaire - Universite Clermont-Auvergne
Clermont Ferrand, France

MAESTRO, Marseilles – June/2017

Type Ia Sne can be used as standard candles

Image credit: NASA/JPL-Caltech

Type Ia Sne can be used as standard candles

The expansion of the Universe
is accelerating!

Image credit: NASA/JPL-Caltech

Type Ia Sne can be used as standard candles

The expansion of the Universe
is accelerating!

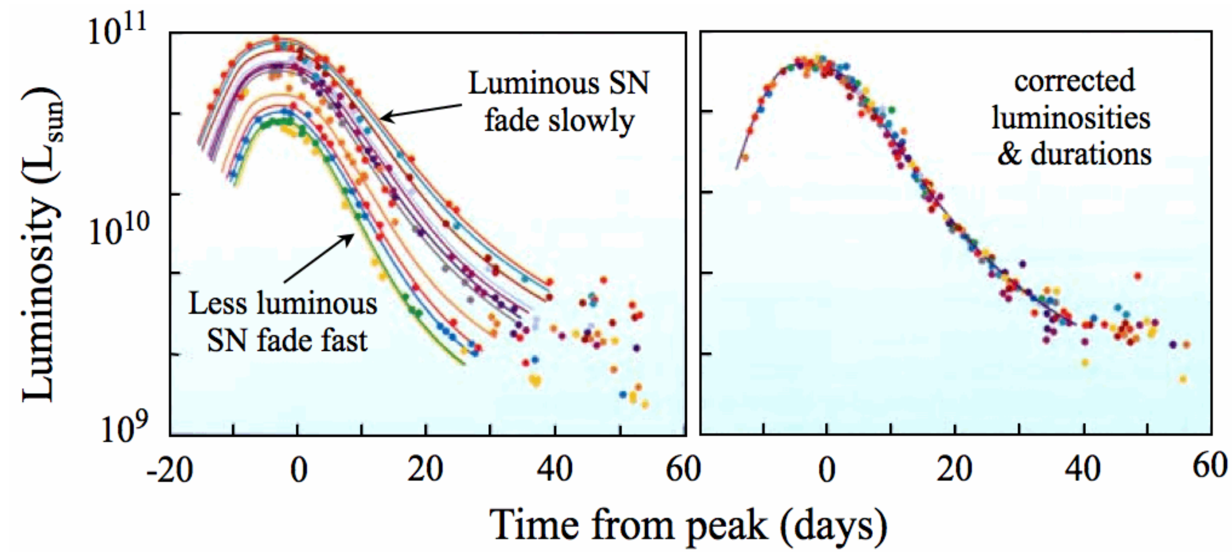
2011

Photo: Lawrence Berkeley National Lab	Photo: Belinda Pratten, Australian National University	Photo: Scanpix/AFP
Saul Perlmutter	Brian P. Schmidt	Adam G. Riess

The Nobel Prize in Physics 2011 was awarded *"for the discovery of the accelerating expansion of the Universe through observations of distant supernovae"* with one half to Saul Perlmutter and the other half jointly to Brian P. Schmidt and Adam G. Riess.

Image credit: NASA/JPL-Caltech

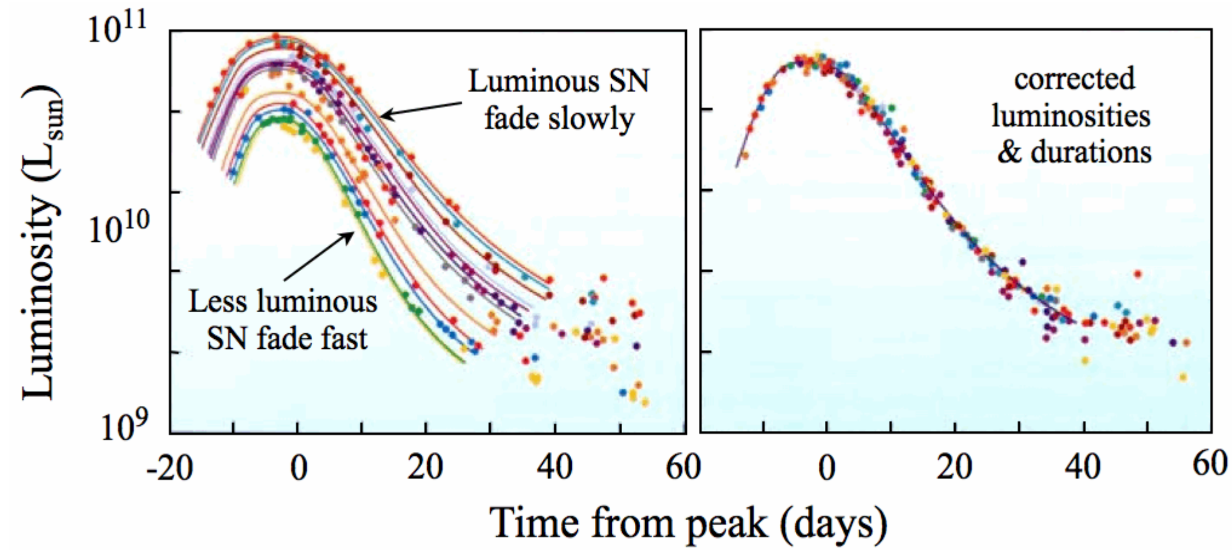
Type Ia Sne are standard*izable* candles



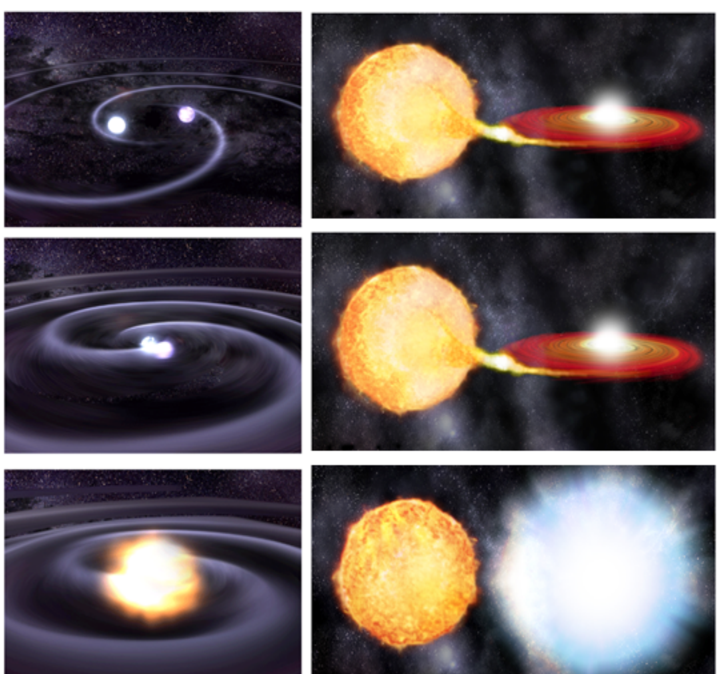
http://community.dur.ac.uk/john.lucey/bridge/SN2015M_bridge.html

Type Ia Sne are standard*izable* candles

Can we use machine learning to distinguish multiple sub-classes of SN Ia ?



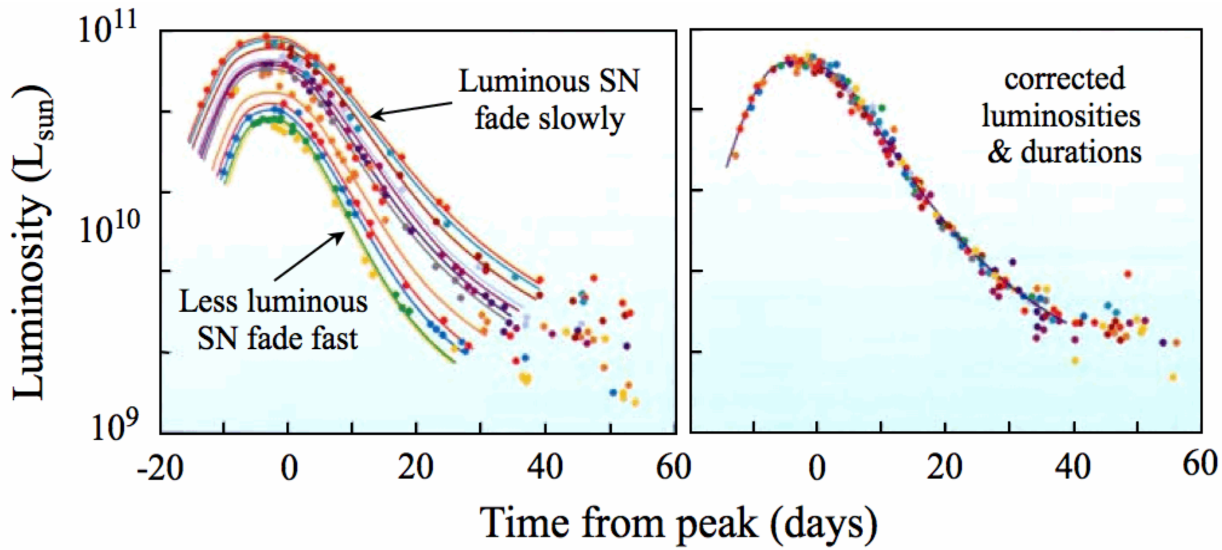
http://community.dur.ac.uk/john.lucey/bridge/SN2015M_bridge.html



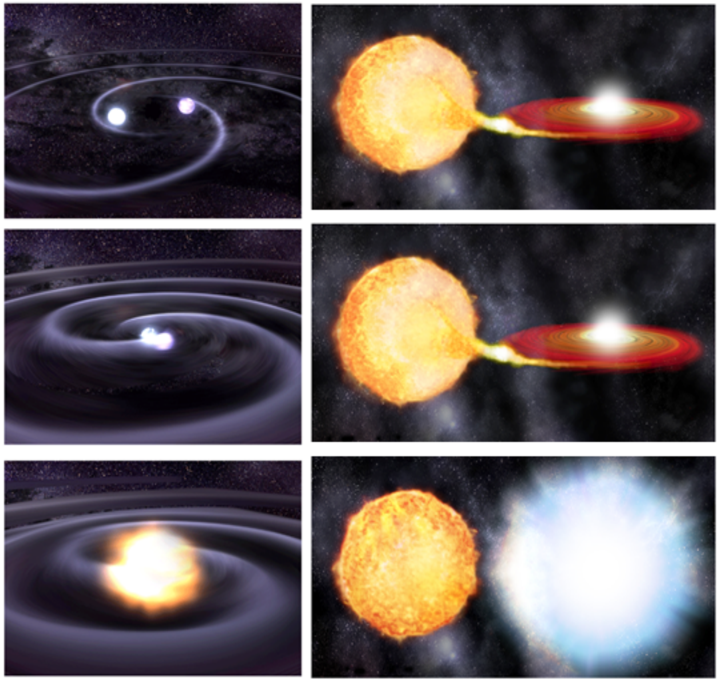
Images taken from Wikipedia Commons and Discover Magazine

Type Ia Sne are standard*izable* candles

Can we use machine learning to distinguish multiple sub-classes of SN Ia ?



http://community.dur.ac.uk/john.lucey/bridge/SN2015M_bridge.html



If so, are the resulting classes physically meaningful?

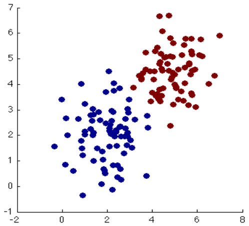
Images taken from Wikipedia Commons and Discover Magazine

IDEALLY

Build data matrix

IDEALLY

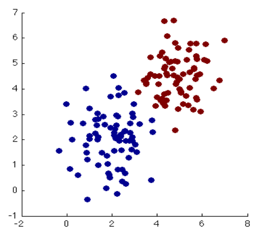
Build data matrix



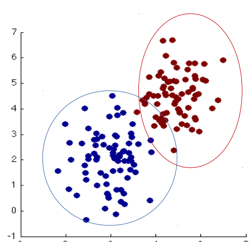
Dimensionality Reduction

IDEALLY

Build data matrix



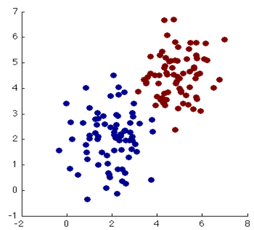
Dimensionality Reduction



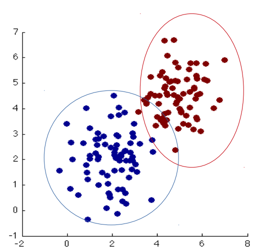
Unsupervised clustering

IDEALLY

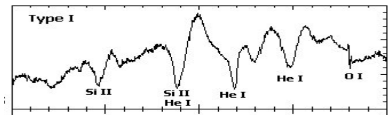
Build data matrix

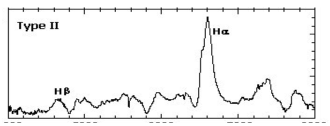


Dimensionality Reduction



Unsupervised clustering

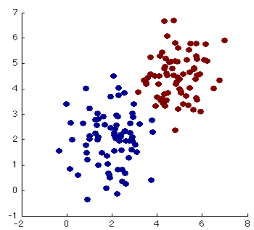




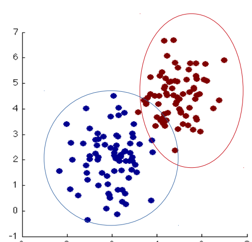
subclasses

IDEALLY

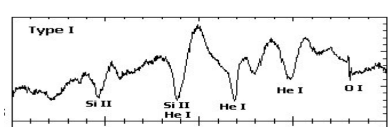
Build data matrix

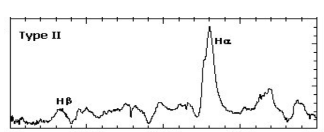


Dimensionality Reduction



Unsupervised clustering





subclasses

-REALISTICALLY-

Build data matrix ← Problem 1:
Not enough data for training

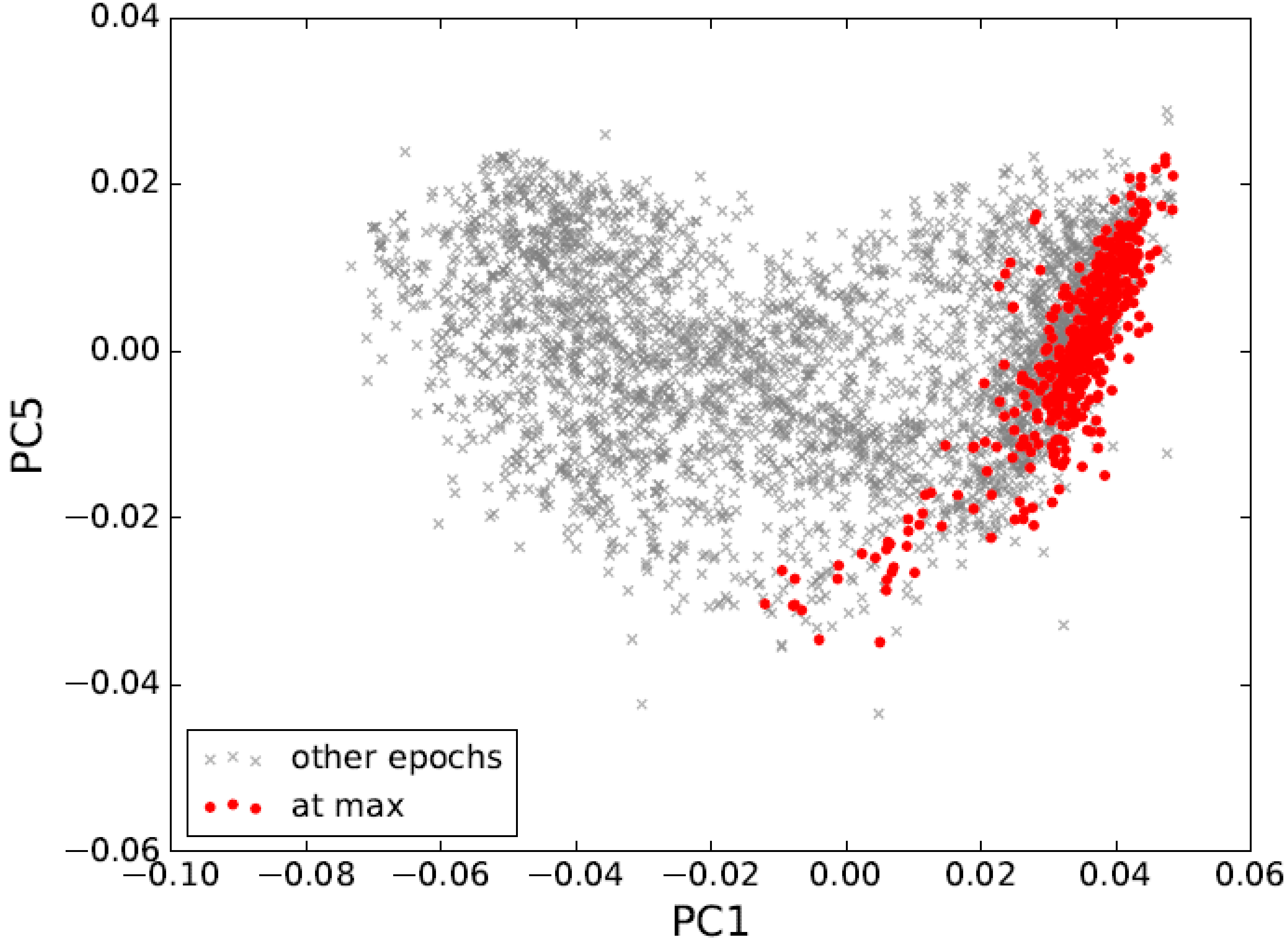
Dimensionality Reduction

Unsupervised clustering

subclasses

PCA parameter space built with all available spectra

even those without an epoch determination
3677 spectra, of which 486 at max



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

-REALISTICALLY-

Build data matrix ◄—— Problem 1:
Not enough data for training
→ Transfer Learning

↓

Dimensionality Reduction

↓

Unsupervised clustering

↓

subclasses

-REALISTICALLY-

Build data matrix ← Problem 1:
Not enough data for training
→ Transfer Learning

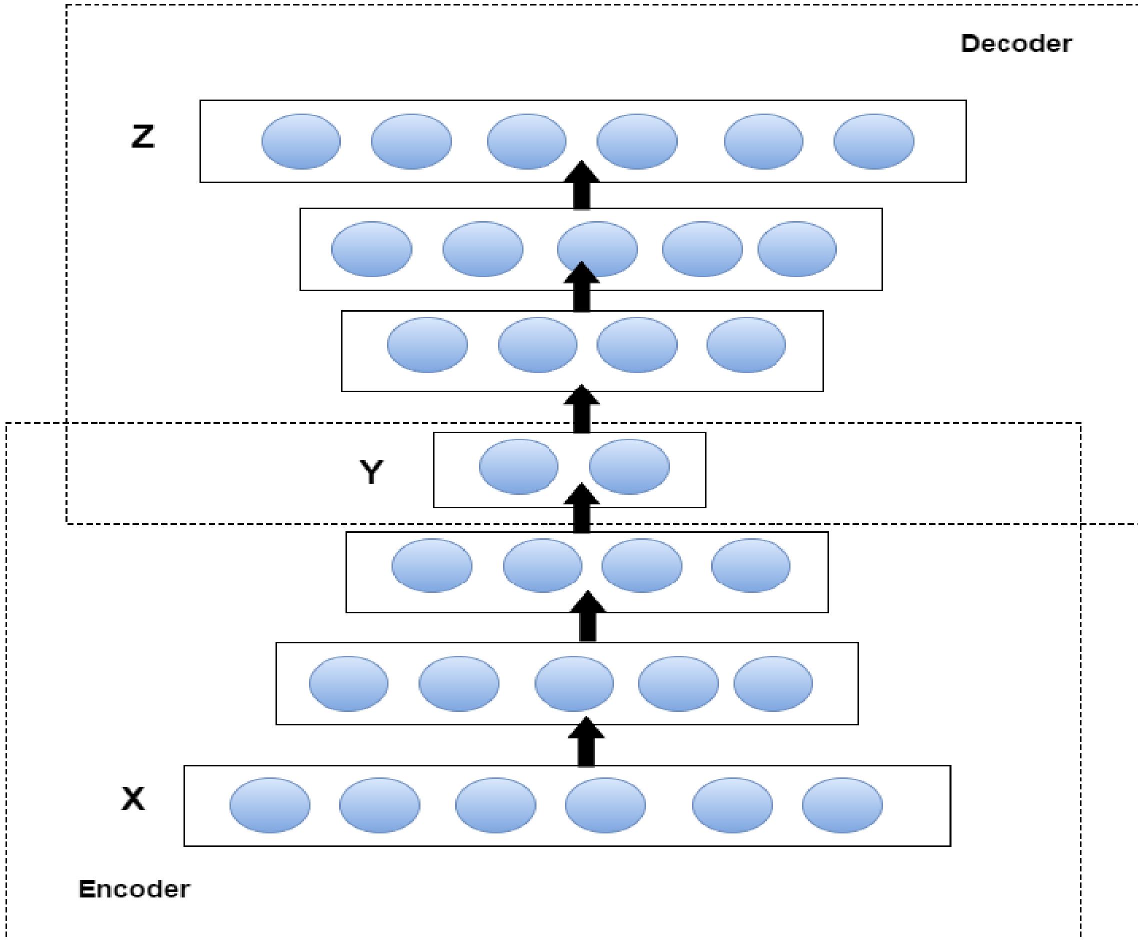
Dimensionality Reduction ← Problem 2:
Highly non-linear problem

Unsupervised clustering

subclasses

Deep Learning for Dimensionality Reduction

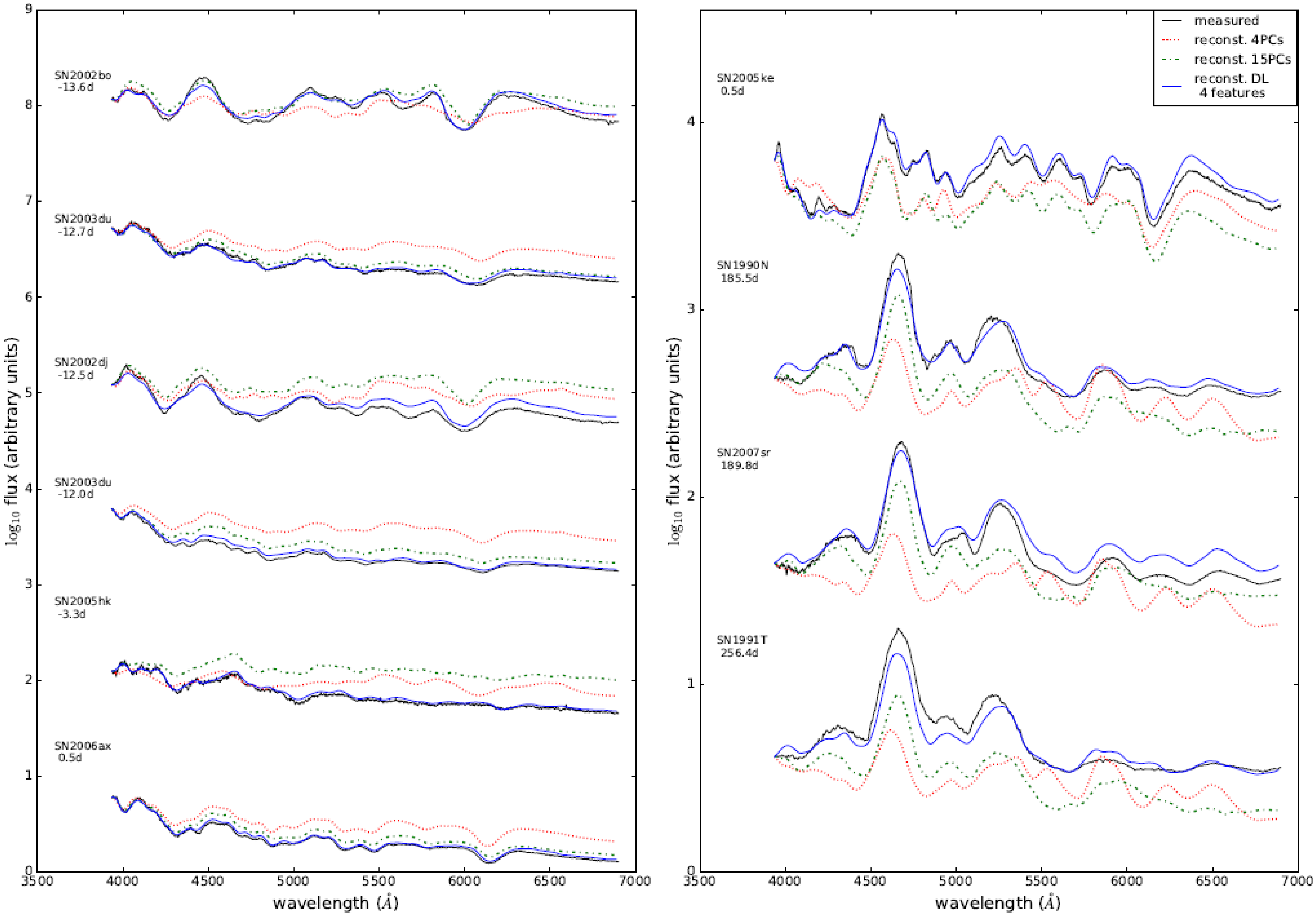
Layers= (120,100,90,50,30,20,4,20,30,50,90,100,120)



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

Deep Learning x PCA reconstruction power

DL results are good even at late epochs



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

-REALISTICALLY-

Build data matrix ← Problem 1: Not enough data for training

↓

Dimensionality Reduction ← Problem 2: Highly non-linear problem

→ Deep Learning

300 → 4

↓

Unsupervised clustering

↓

subclasses

Build data matrix

Problem 1:
Not enough data for training
→ Transfer Learning

Dimensionality Reduction

Problem 2:
Highly non-linear problem
→ Deep Learning

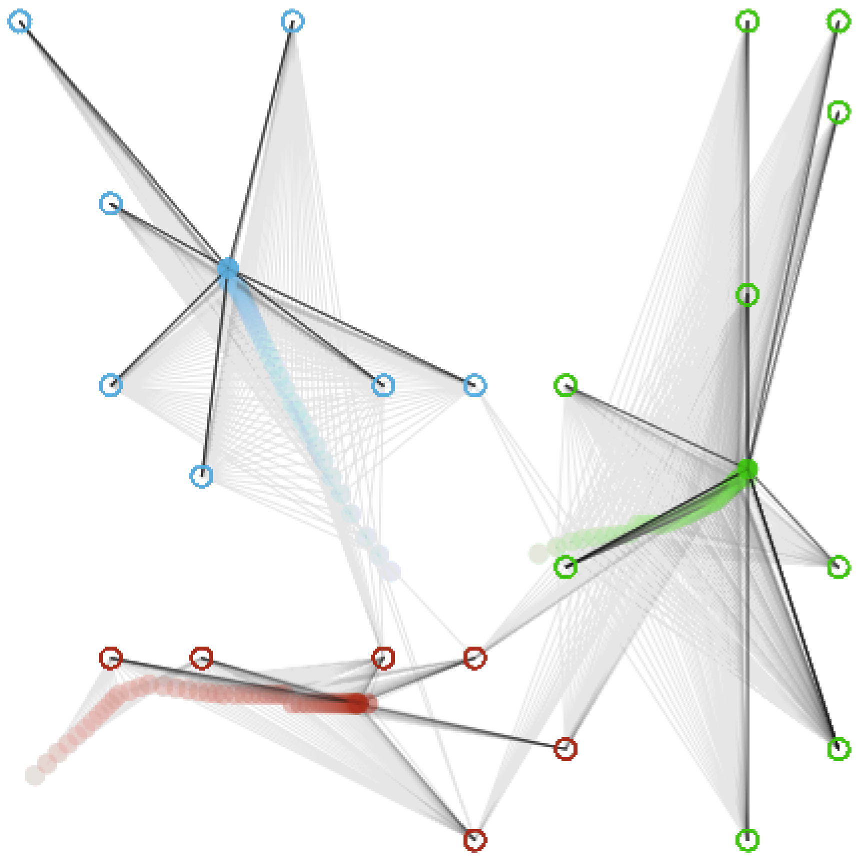
Unsupervised clustering

Problem 3:
We search for a data-driven approach

subclasses

K-means clustering

Number of clusters as input



https://miguelmota.com/blog/k-means-clustering-in-javascript/

REALISTICALLY

Build data matrix

Problem 1:
Not enough data for training
→ Transfer Learning

Dimensionality Reduction

Problem 2:
Highly non-linear problem
→ Deep Learning

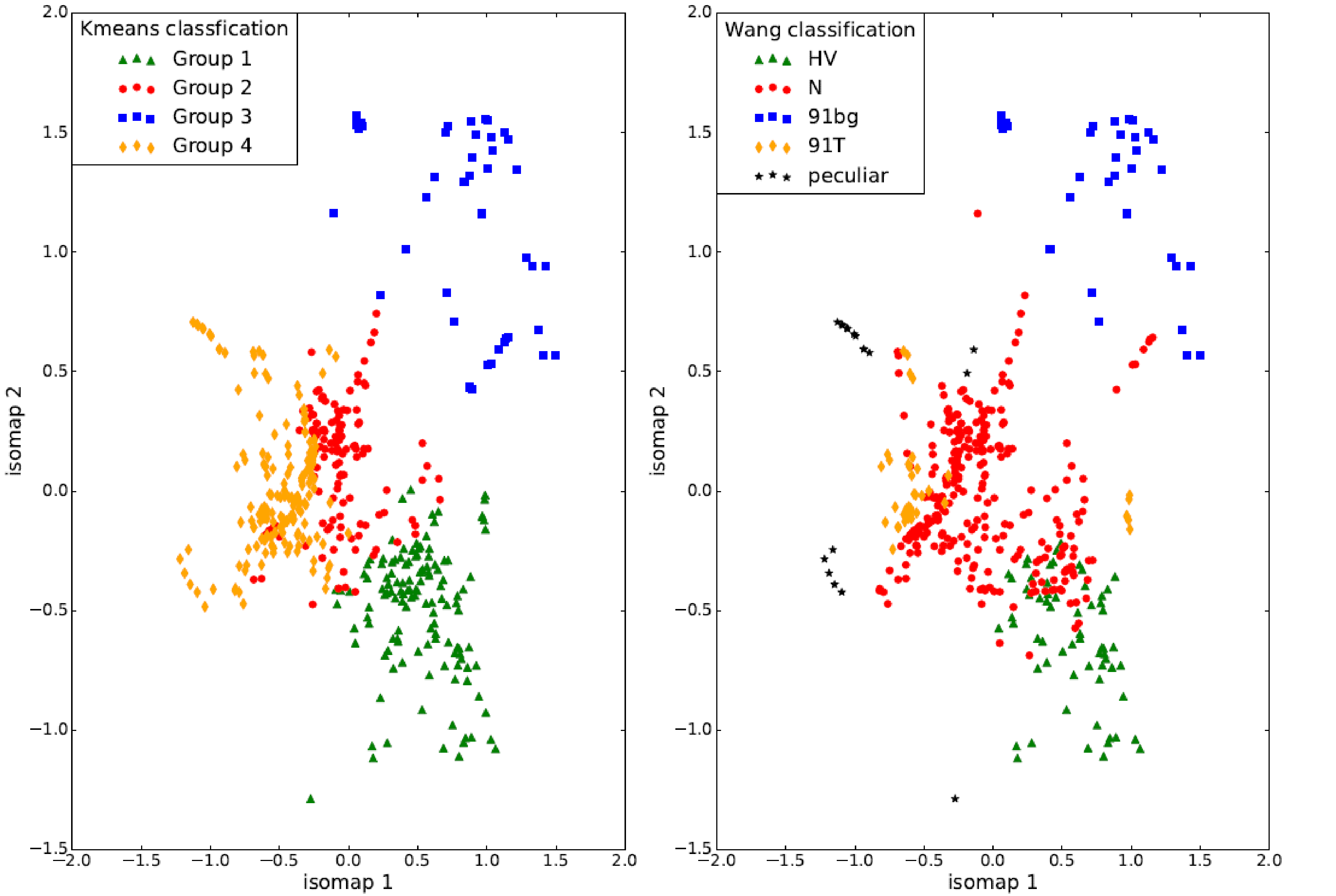
Unsupervised clustering

Problem 3:
We search for a data-driven approach

subclasses

2D visualization of 4D Deep Learning parameter space

Results from K-means



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

REALISTICALLY

Build data matrix

↓

Dimensionality Reduction

↓

Unsupervised clustering

↓

subclasses

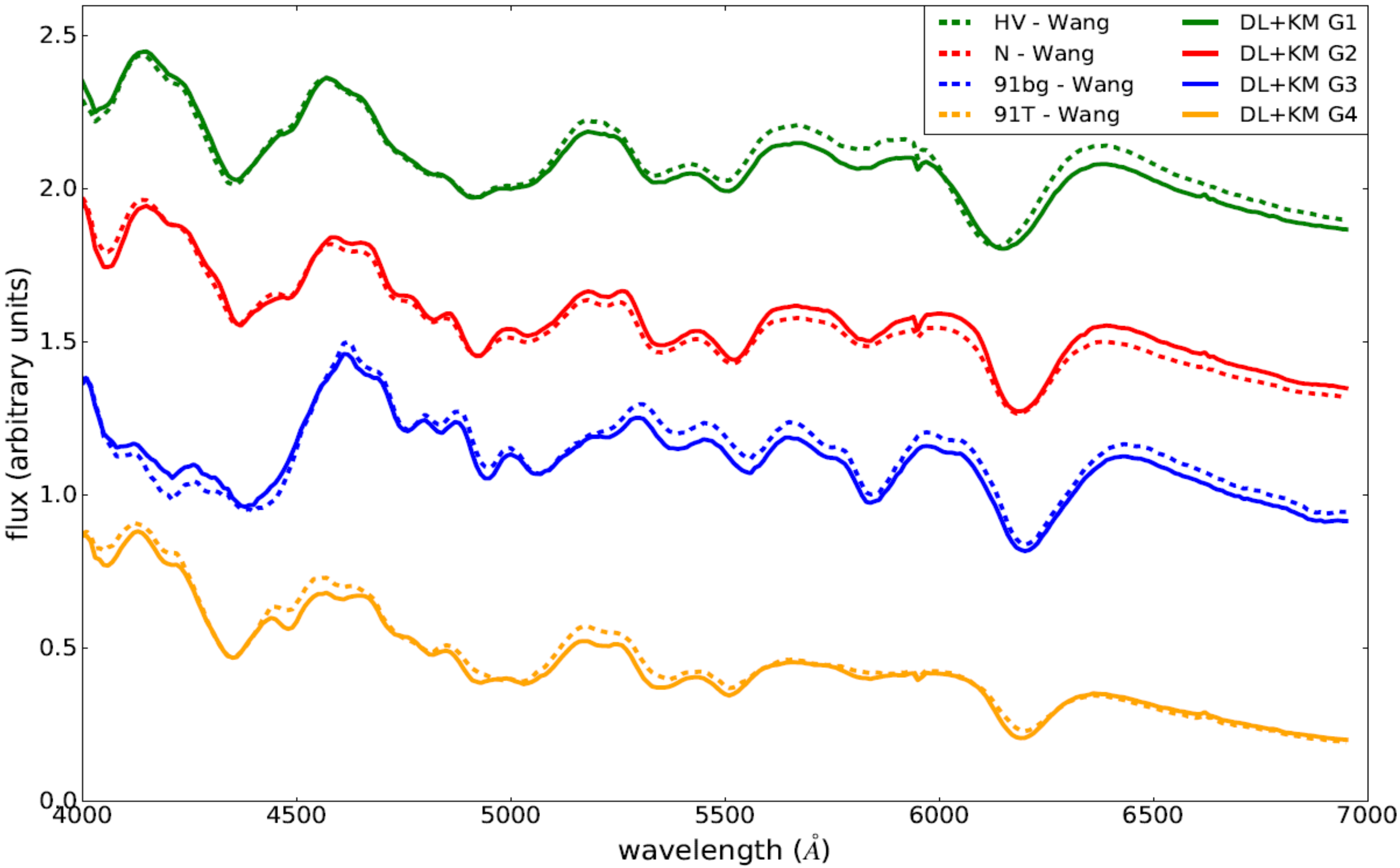
Problem 1:
Not enough data for training
→ Transfer Learning

Problem 2:
Highly non-linear problem
→ Deep Learning

Problem 3:
We search for a data-driven approach
→ K-means

Mean spectra by ML x human

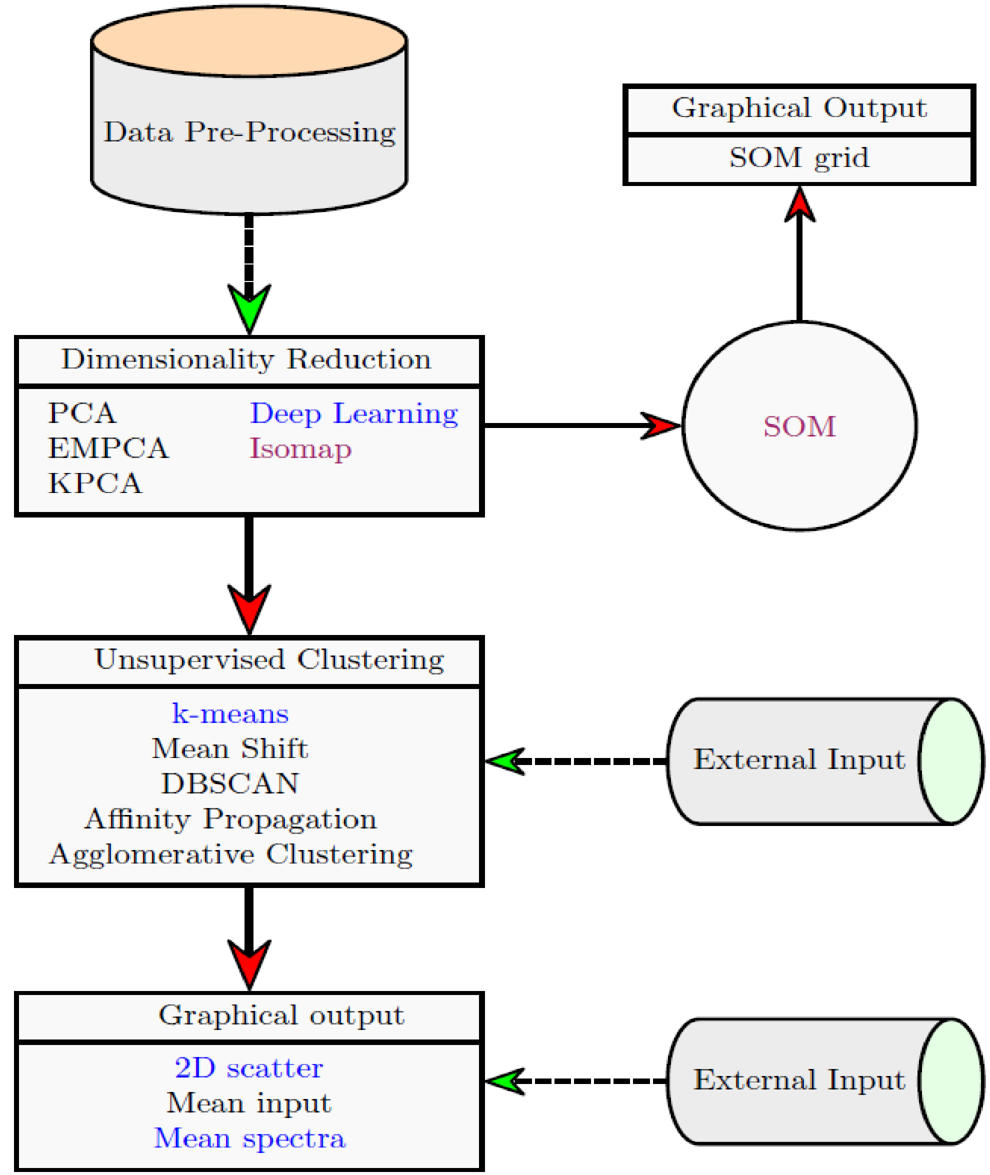
K-means with 4 groups x Wang



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

https://github.com/COINtoolbox/DRACULA

COIN'S Dracula Workflow



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

COIN'S Dracula TEAM

Michele Sasdelli
Project leader
Liverpool John Moores University, UK
Postdoc - on site
Now in Cortexica

Emille Ishida
U. Clermont-Auvergne, France
Postdoc - on site

Ricardo Vilalta
U. Houston, USA
Senior - remotely

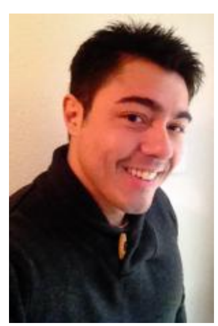
Michel Aguena
Head of software development
U. of Sao Paulo, Brazil
PhD student - remotely

Vinicius Busti
U. of Sao Paulo, Brazil
Postdoc - remotely

Hugo Camacho
U. of Sao Paulo, Brazil
PhD student - remotely

Arlindo Trindade
U. of Porto, Portugal
PhD student – on site
Now in Rolls-Royce

Fabian Gieseke
Radboud University Nijmegen, Netherlands
Postdoc – remotely
Now in Denmark



Rafael de Souza
ELTE, Hungary
Postdoc – on site

Yabebal Fantaye
U. Rome Tor Vergata, Italy
Postdoc – on site
Now in South Africa

Paollo Mazzali
Liverpool John Moores University, UK
Senior - remotely

What now?

Is there a future in this kind unsupervised learning?

What now?

Is there a future in this kind unsupervised learning?

Maybe - if the data is abundance and of good quality

Human classification is not based on current measurement alone

What now?

Is there a future in this kind unsupervised learning?

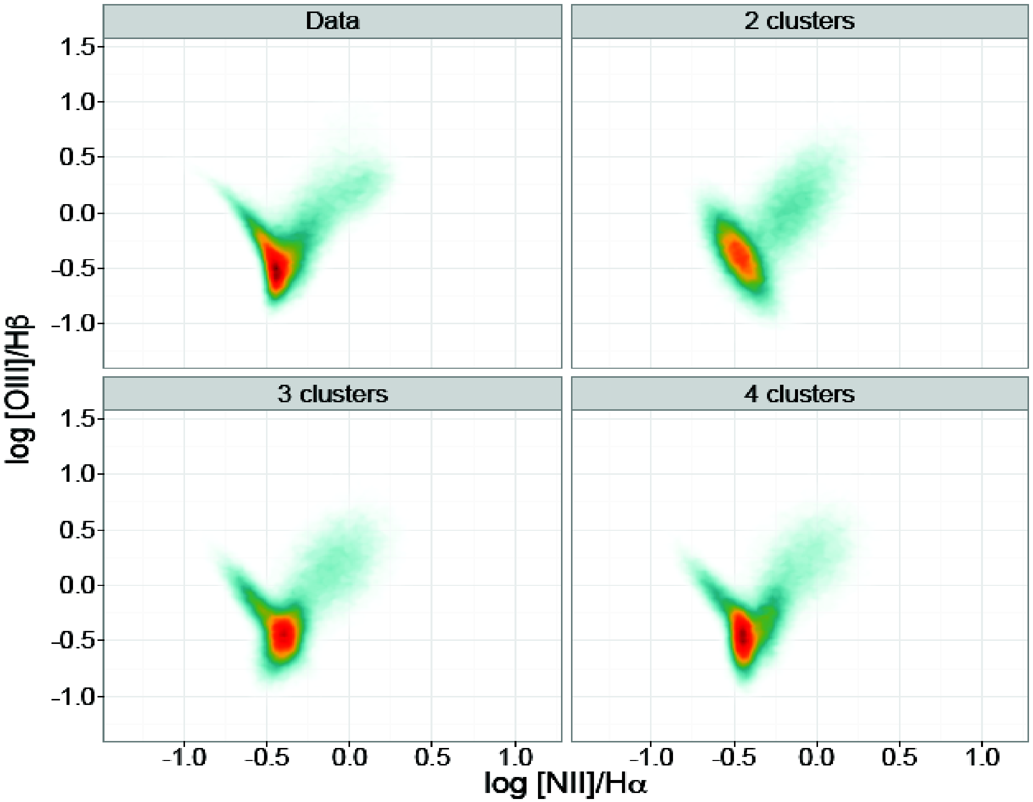
Maybe - if the data is abundance and of good quality

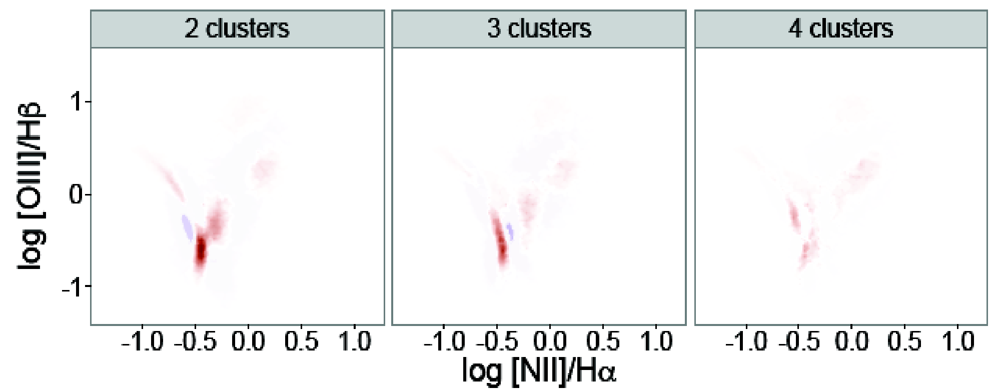
Human classification is not based on current measurement alone

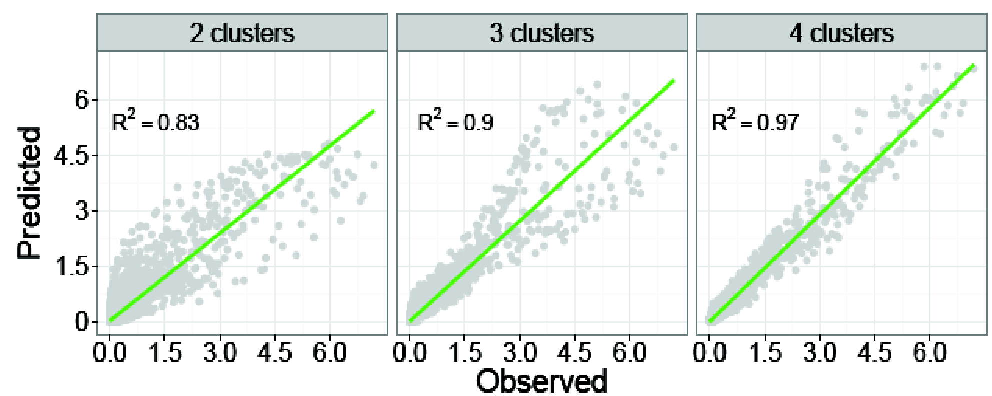
Number of clusters - statistics might help

External clustering validation

CRP #4: External Cluster Validation







de Souza et al, 2017 - A probabilistic approach for emission-line galaxy classification – arXiv:astro-ph/1703.07607

What now?

Is there a future in this kind unsupervised learning?

Maybe - if the data is abundance and of good quality

Human classification is not based on current measurement alone

Number of clusters - statistics might help

External clustering validation

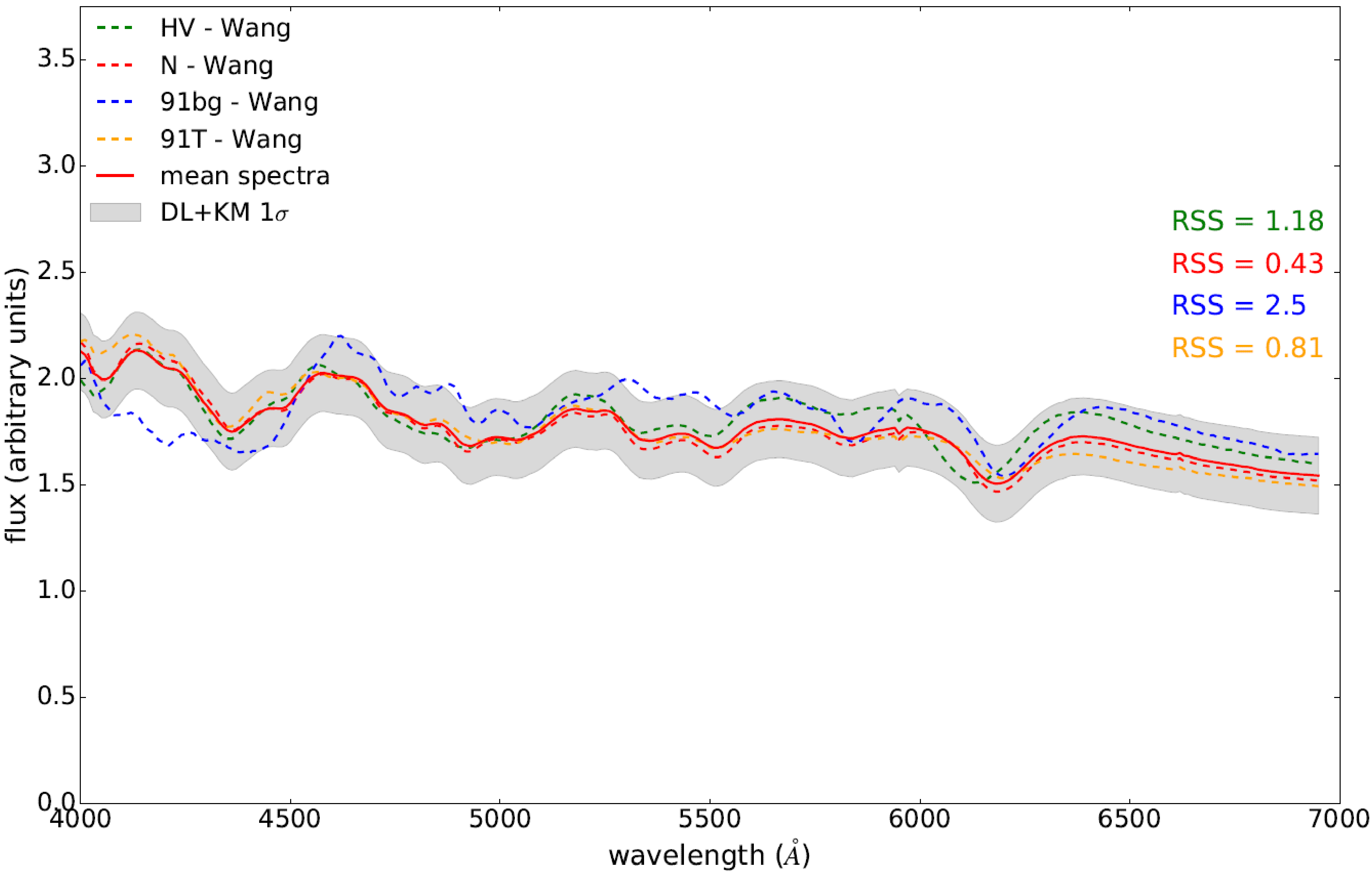
Can data guide the theoretical modeling?

To be continued...


Thank You!

Mean spectra by ML x human

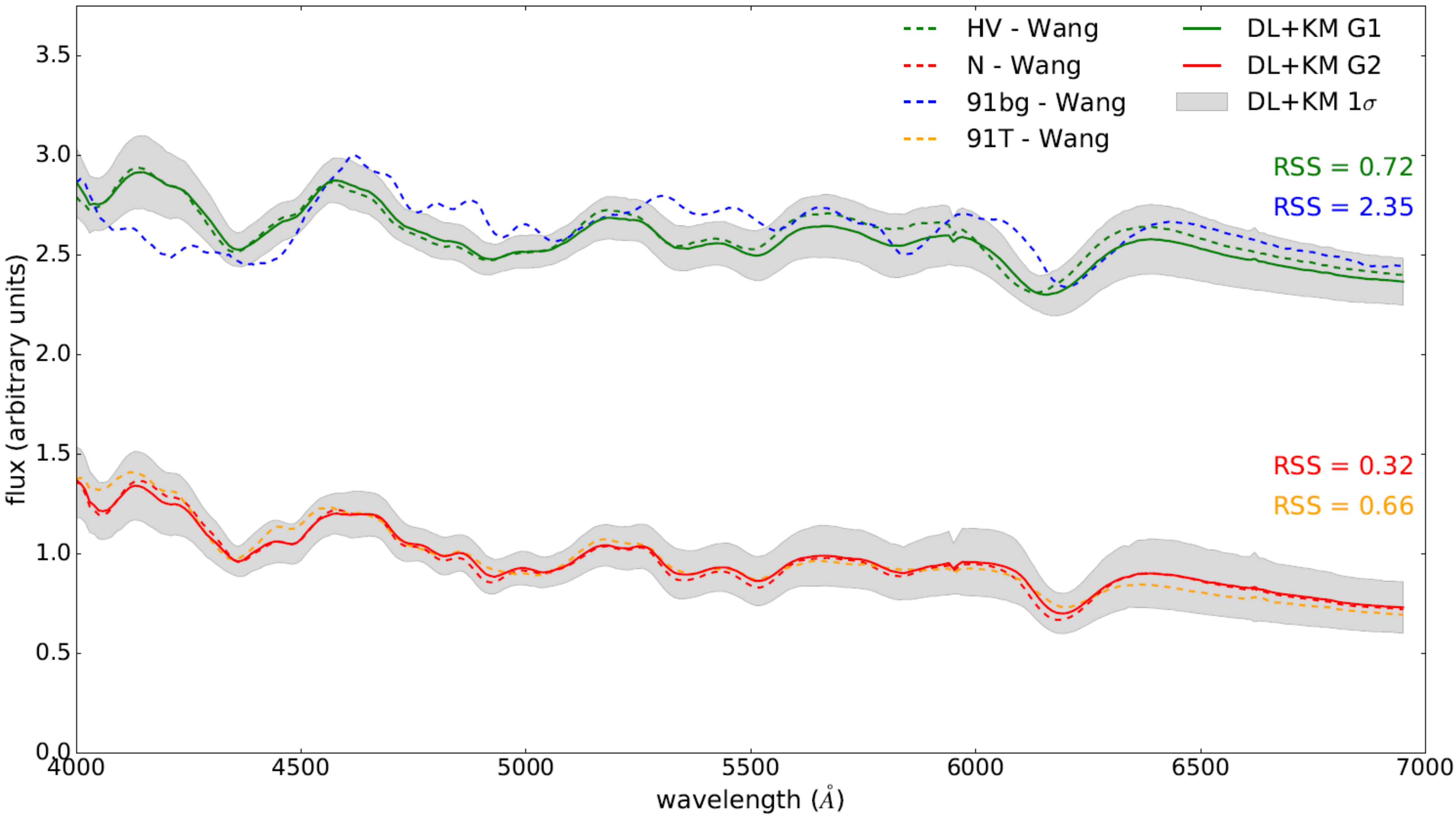
K-means with 1 group



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

Mean spectra by ML x human

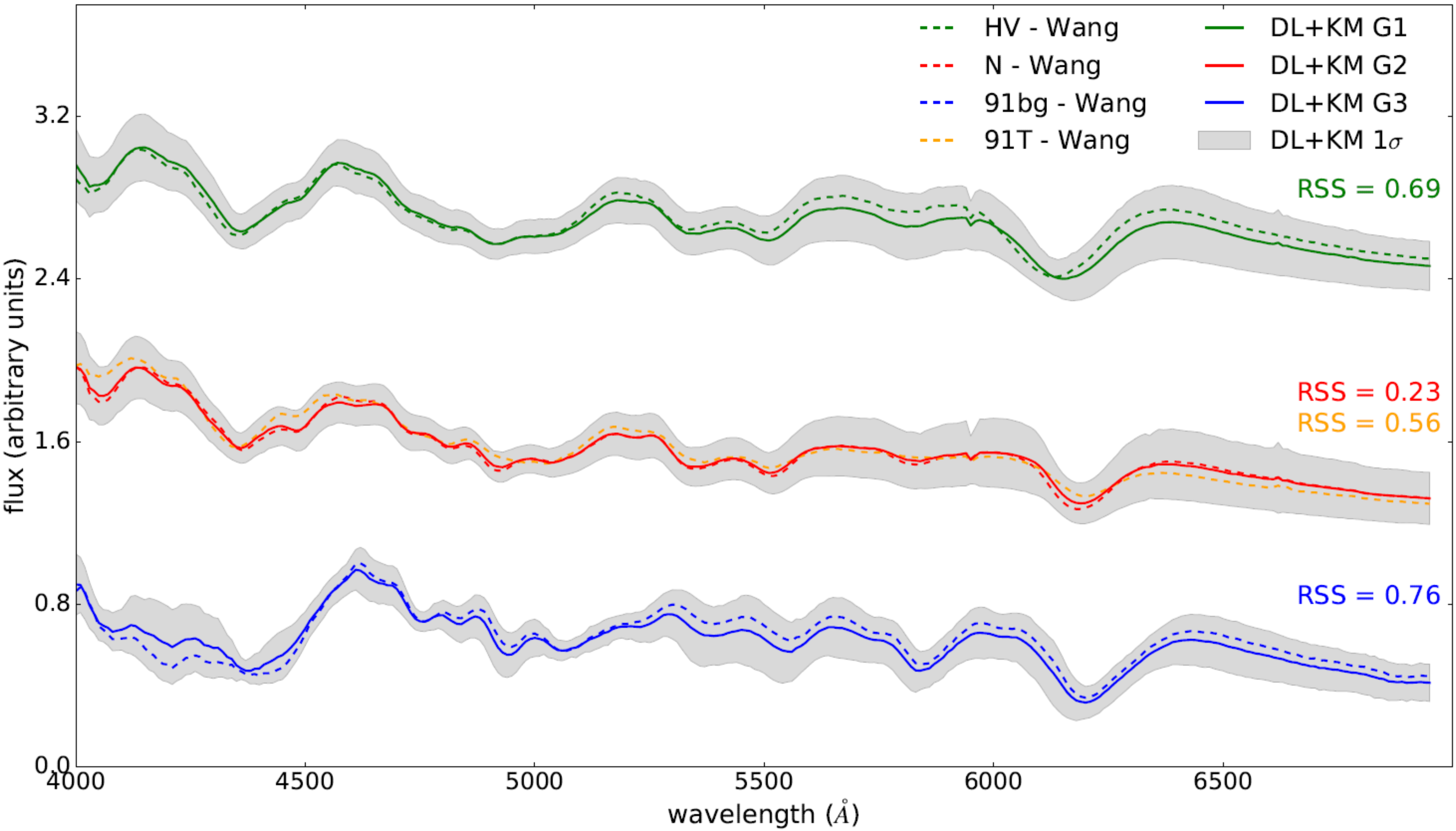
K-means with 2 groups



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

Mean spectra by ML x human

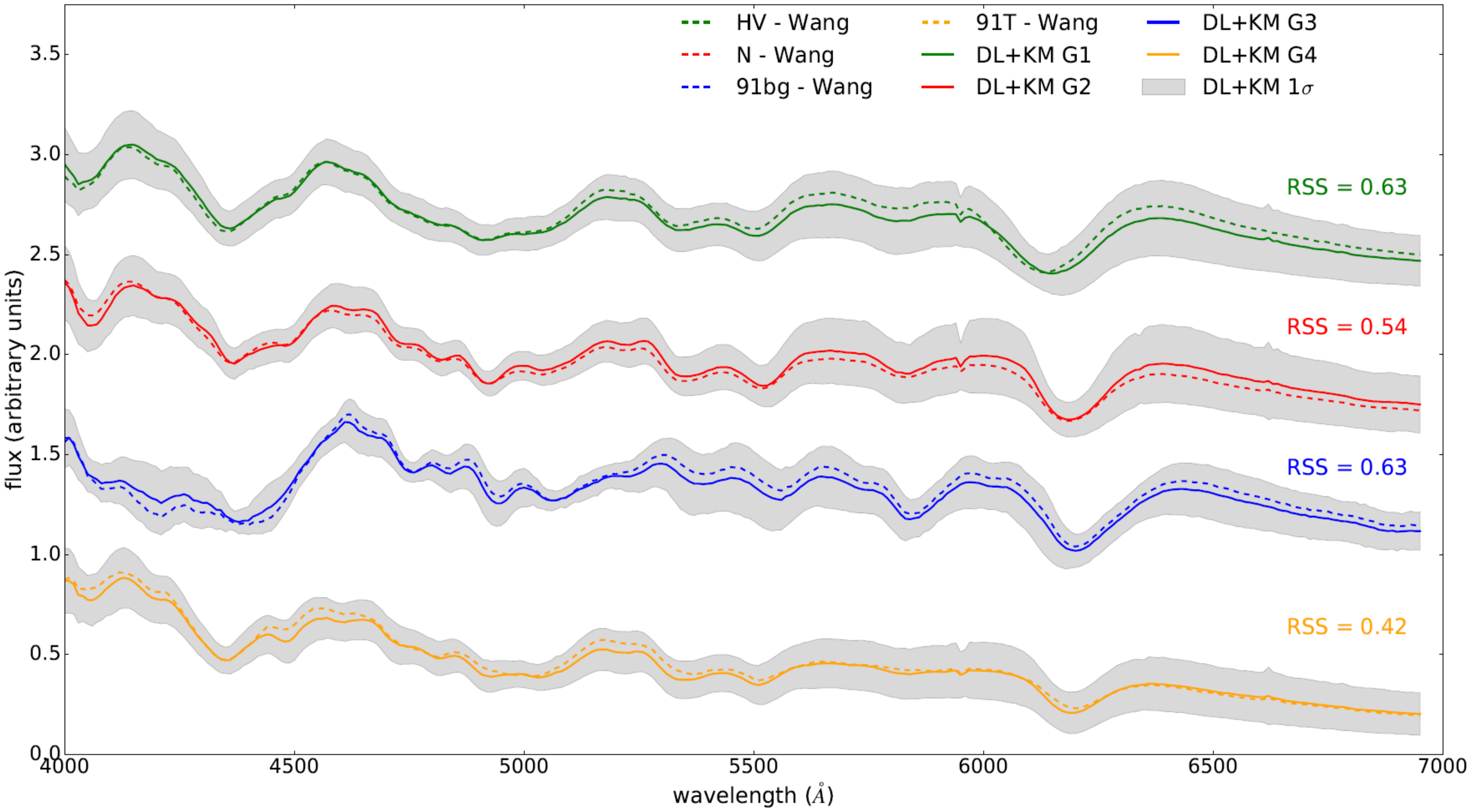
K-means with 3 groups



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

Mean spectra by ML x human

K-means with 4 groups



Sasdelli *et al.*, 2016, MNRAS, *461, Issue 2, p.2044-2059*

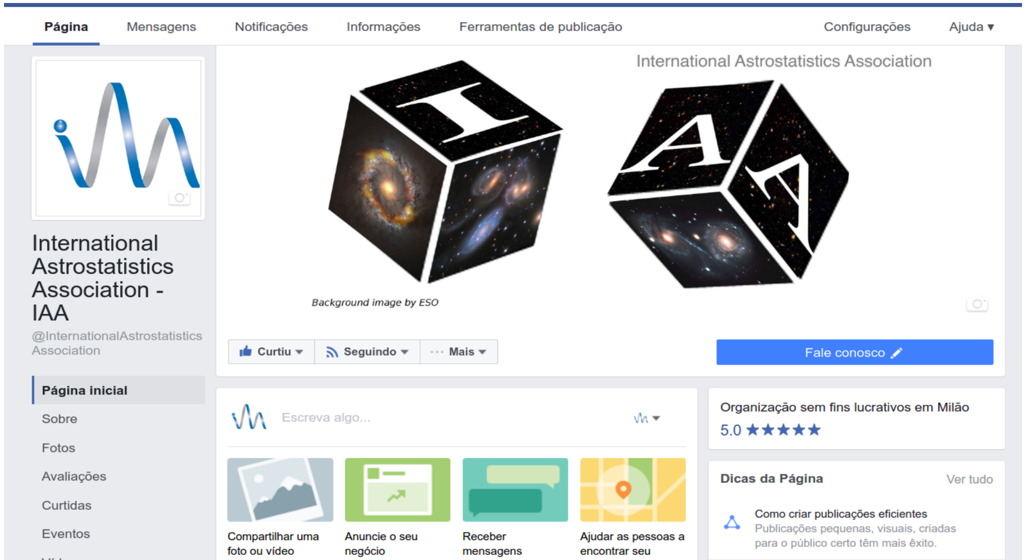
In CRP #3

Budapest, 2016

COIN Residence Program #4

Clermont Ferrand, 20-27 August 2017

IAA facebook page



COIN on twitter

Registrations open in July

Deep Learning + Self-Organized Maps

Investigating the parameter space

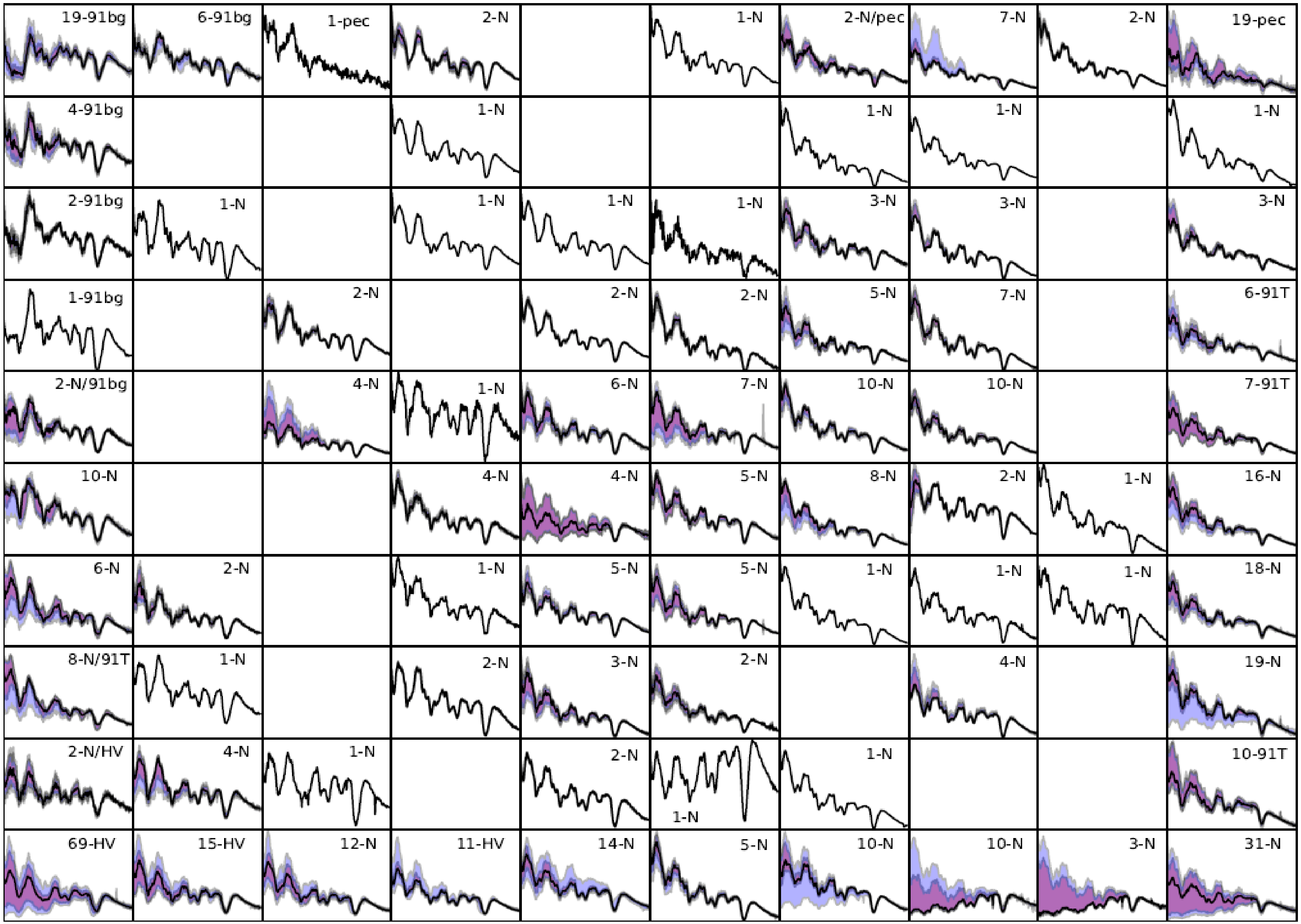


Figure 7. A self organized map of SN Ia spectra at maximum, constructed from the Deep Learning 4-dimensional feature space. Black lines correspond to the mean spectra of each cell, purple and blue bands correspond 1σ and 2σ respectively. Also shown are the number of spectra allocated in each individual cell and the subtype of the majority of SNe populating each cell according to the classification proposed by Wang et al. (2009b). In case there are exactly the same number of objects of different subtypes both labels are shown.

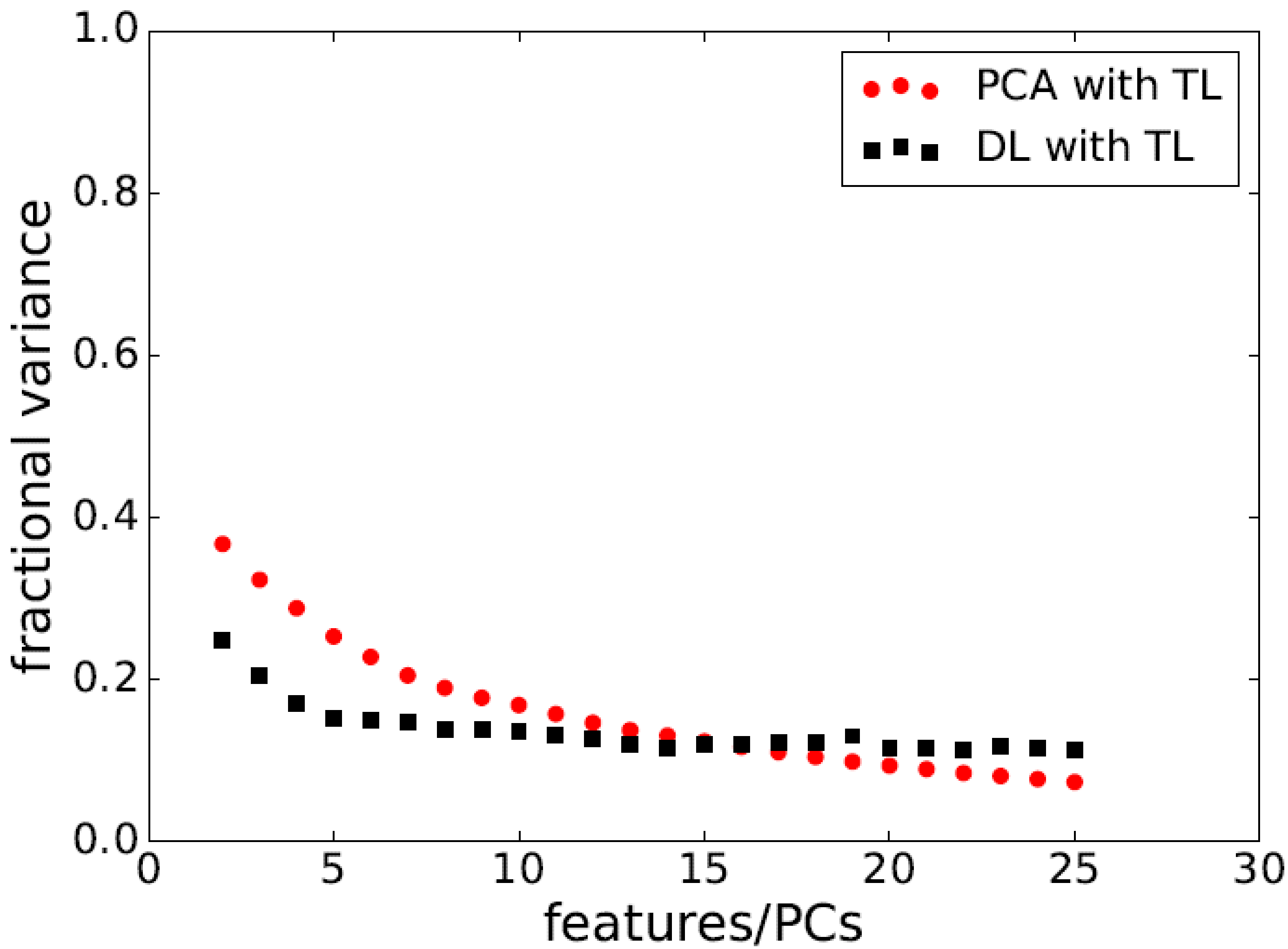
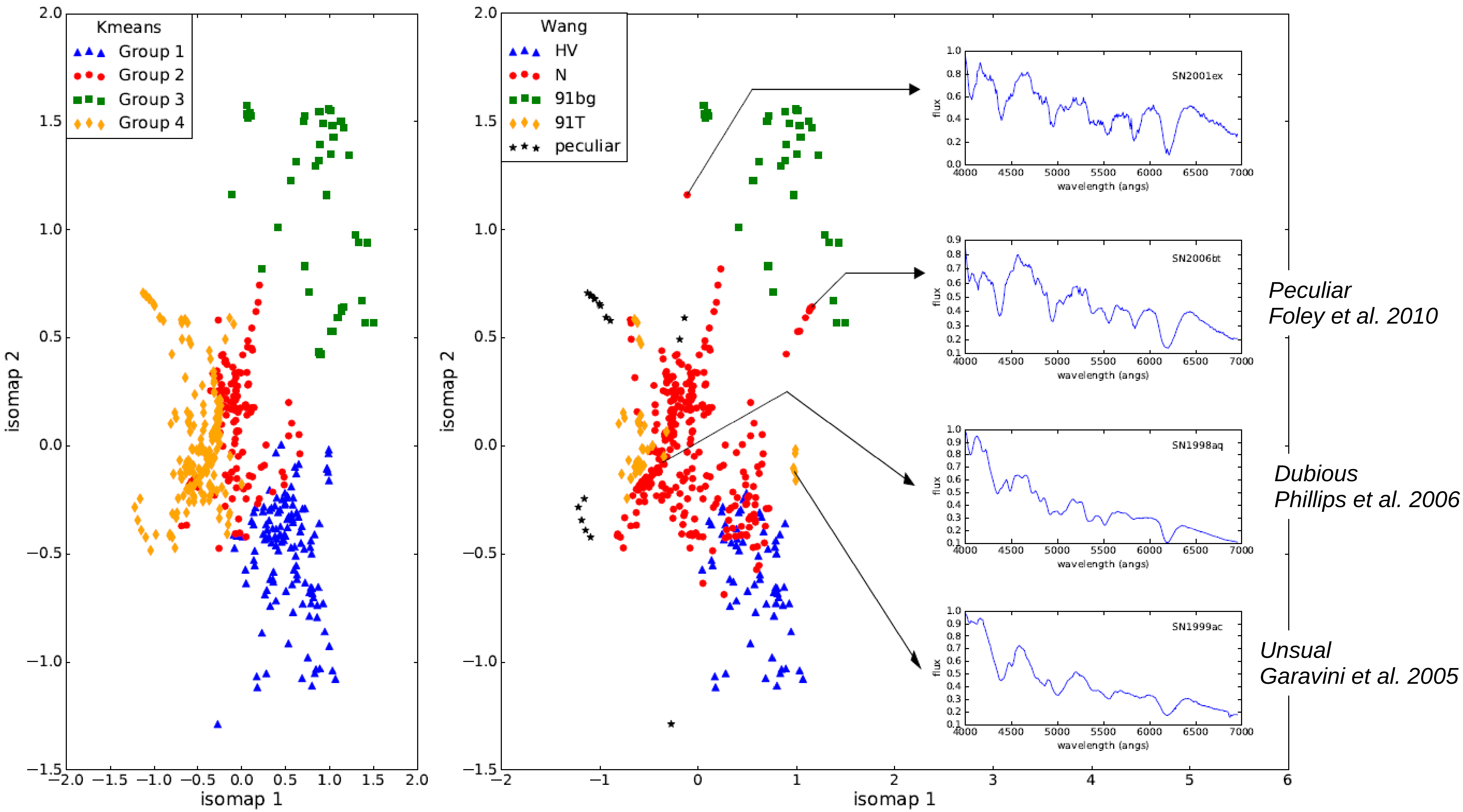


Figure 6. Comparison between Deep Learning and PCA in their capacity to reconstruct the original spectra and robustness to overfitting. Horizontal axis stands for the number of PCs/features and vertical axis shows the deviation between real data and reconstruction.

2D visualization of 4D Deep Learning parameter space

Results from K-means



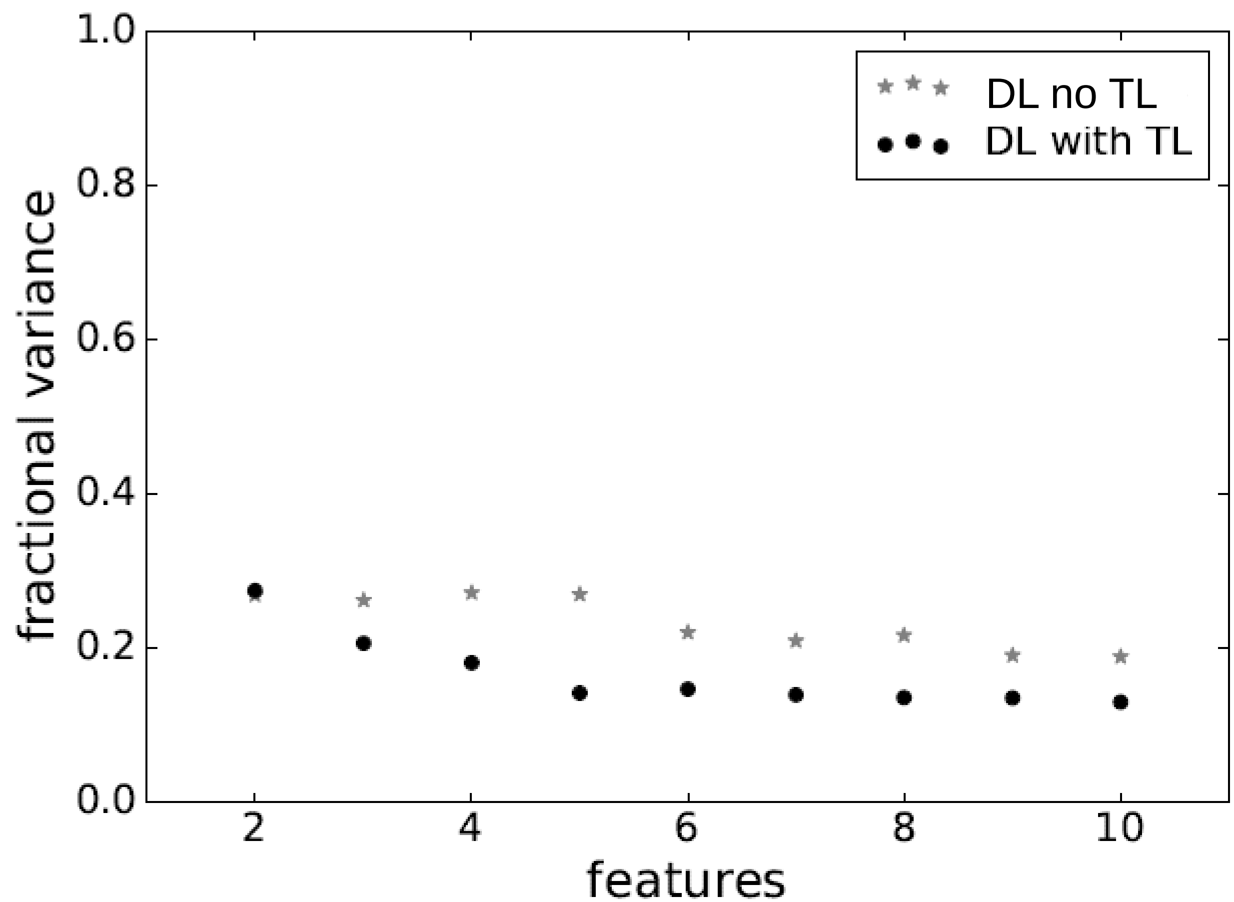


Figure 4. Variance between the deep learning reconstructions and observed SN Ia spectra at B max. Gray crosses correspond to a data configuration using only SNe Ia at maximum (no transfer learning) and black circles denote results from an initial data matrix containing spectra from all epochs (with transfer learning). The horizontal axis stands for the number of features; the vertical axis shows the deviation between real data and reconstruction.