

# Applications of machine learning and neural networks to data processing

Carlos José Díaz Baso

Rosseland Centre for Solar Physics, Institute of Theoretical Astrophysics, University of Oslo, N-0315 Oslo, Norway

[carlos.diaz@astro.uio.no](mailto:carlos.diaz@astro.uio.no)

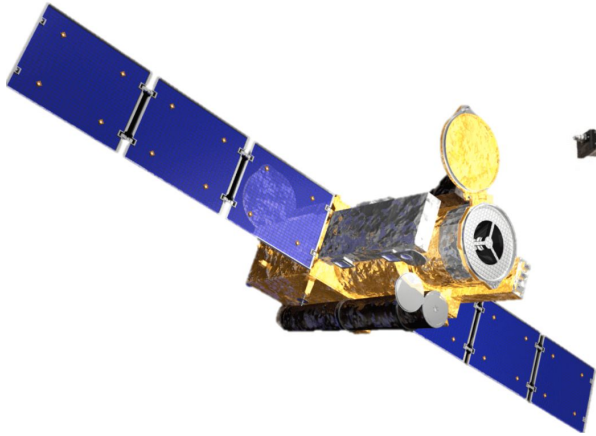


Rosseland  
Centre  
for Solar  
Physics

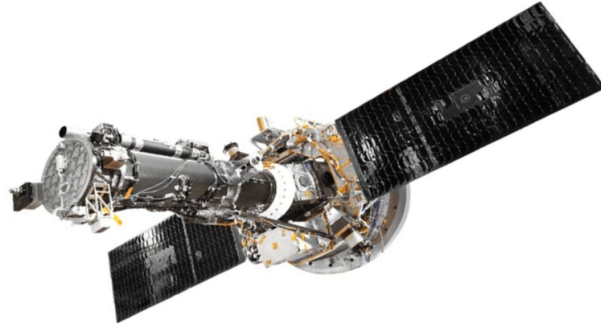
# Why do we need a review?

(Pietarila et al. 2007; Viticchié & Sánchez Almeida 2011; Panos et al. 2018; Sainz Dalda et al. 2019; Bose et al. 2019; Rouppe van der Voort et al. 2021; Robustini et al. 2019; Joshi & Rouppe van der Voort 2020b; Kuckein et al. 2020; Bose et al. 2021a,b; Barczynski et al. 2021; Nóbrega-Siverio et al. 2021; Kleint & Panos 2022; Joshi et al. 2022 Asensio Ramos et al. 2007; Asensio Ramos & López Ariste 2010 Quintero Noda et al. 2015, 2016; Felipe et al 2016, Griñón-Marín 2021; Rees et al. 2000; López Ariste & Casini 2002; Skumanich & López Ariste 2002; Casini et al. 2005; Casini et al. 2009, 2013; Ruiz Cobo & Asensio Ramos (2013); Socas-Navarro et al. (2001); Yuan et al. 2010; Nishizuka et al. 2017; Florios et al. 2018; Bobra et al. (2015, 2016); Carroll et al. 2001, 2008; Socas-Navarro, H. 2003, 2005; Sainz Dalda et al. 2019; Milić et al. 2020; Gafeira et al. 2021; Centeno et al. 2022); Kianfar S. et al. (2019); Morosin R. et al. 2022; Skumanich & López Ariste 2002; Vishal Upendran 2020; Armstrong et al. 2021; Wang et al 2021; Deng et al 2021; Armstrong & Fletcher (2019); Ahmadzadeh et al 2019; Zhu et al. 2019; Diercke et al 2022; Felipe et al. 2019; Broock et al 2021; Pietarila et al. 2007; Viticchié & Sánchez Almeida 2011; Panos et al. 2018 2020; Sainz Dalda et al. 2019; Bose et al. 2019; Rouppe van der Voort et al. 2021; Robustini et al. 2019; Joshi & Rouppe van der Voort 2020b; Kuckein et al. 2020; Bose et al. 2021a,b; Barczynski et al. 2021; Nóbrega-Siverio et al. 2021; Kleint & Panos 2022; Joshi et al. 2022 Asensio Ramos et al. 2007; Asensio Ramos & López Ariste 2010 Quintero Noda et al. 2015, 2016; Felipe et al 2016, Griñón-Marín 2021; Rees et al. 2000; López Ariste & Casini 2002; Skumanich & López Ariste 2002; Casini et al. 2005; Casini et al. 2009, 2013; Ruiz Cobo & Asensio Ramos (2013); Socas-Navarro et al. (2001); Yuan et al. 2010; Nishizuka et al. 2017; Florios et al. 2018; Bobra et al. (2015, 2016); Carroll et al. 2001, 2008; Socas-Navarro, H. 2003, 2005; Szenicer et al 2019; Lim et al 2021; Salvatelli et al 2022; [Yes, your paper should be also around here](#); Milić et al. 2020; Gafeira et al. 2021; Centeno et al. 2022); Kianfar S. et al. (2019); Morosin R. et al. (2022); Skumanich & López Ariste 2002; Asensio Ramos 2018, 2021; Armstrong et al. 2021; Wang et al 2021; Deng et al 2021; Armstrong & Fletcher (2019); Ahmadzadeh et al 2019; Zhu et al. 2019; Diercke et al 2022; Felipe et al. 2019; Broock et al 2021; Pietarila et al. 2007; Viticchié & Sánchez Almeida 2011; Panos et al. 2018; Sainz Dalda et al. 2019; Bose et al. 2019; Rouppe van der Voort et al. 2021; Robustini et al. 2019; Joshi & Rouppe van der Voort 2020b; Kuckein et al. 2020; Bose et al. 2021a,b; Barczynski et al. 2021; Nóbrega-Siverio et al. 2021; Kleint & Panos 2022; Joshi et al. 2022 Asensio Ramos et al. 2007; Thoen Faber 2022; Quintero Noda et al. 2015, 2016; Felipe et al 2016, Griñón-Marín 2021; Rees et al. 2000; López Ariste & Casini 2002; Skumanich & López Ariste 2002; Casini et al. 2005; Casini et al. 2009, 2013; Ruiz Cobo & Asensio Ramos 2013; Socas-Navarro et al. 2001; Yuan et al. 2010; Nishizuka et al. 2017; Florios et al. 2018; Bobra et al. (2015, 2016); Carroll et al. 2001, 2008; Socas-Navarro, H. 2003, 2005; Sainz Dalda et al. 2019; Milić et al. 2020; Gafeira et al. 2021; Centeno et al. 2022); Kianfar S. et al. (2019); Morosin R. et al. (2022); Skumanich & López Ariste 2002; Vishal Upendran 2020; Armstrong et al. 2021; Wang et al 2021; Deng et al 2021; Armstrong & Fletcher (2019); Ahmadzadeh et al 2019; Zhu et al. 2019; Diercke et al 2022; Felipe et al. 2019; Broock et al 2021; Pietarila et al. 2007; Viticchié & Sánchez Almeida 2011; Panos et al. 2018; Sainz Dalda et al. 2019; Bose et al. 2019; Rouppe van der Voort et al. 2021; Robustini et al. 2019; Joshi & Rouppe van der Voort 2020b; Kuckein et al. 2020; Bose et al. 2021a,b; Barczynski et al. 2021; Nóbrega-Siverio et al. 2021; Kleint & Panos 2022; Joshi et al. 2022 Asensio Ramos et al. 2007; Asensio Ramos & López Ariste 2010; Quintero Noda et al. 2015, 2016; Felipe et al 2016, Griñón-Marín 2021; Rees et al. 2000; López Ariste & Casini 2002; Skumanich & López Ariste 2002; Thoen Faber 2022; Casini et al. 2005; Casini et al. 2009, 2013; Ruiz Cobo & Asensio Ramos (2013); Socas-Navarro et al. (2001); Yuan et al. 2010; Nishizuka et al. 2017; Florios et al. 2018; Bobra et al. (2015, 2016); Carroll et al. 2001, 2008; Socas-Navarro, H. 2003, 2005 2009; Sainz Dalda et al. 2019; Milić et al. 2020; Gafeira et al. 2021; Centeno et al. 2022); Kianfar S. et al. (2019); Morosin R. et al. (2022); Skumanich & López Ariste 2002; Asensio Ramos 2018,2021; Armstrong et al. 2021; Wang et al 2021; Deng et al 2021; Armstrong & Fletcher 2019, 2021; Ahmadzadeh et al 2019; Zhu et al. 2019; Diercke et al 2022; Szenicer et al 2019; Lim et al 2021; Salvatelli et al 2022)

# ¿Big Data?



Hinode/SOT (2006-now) ~ 35 TB\*  
\*Level 1 (FG) + 1&2 (SP)



IRIS (2013-now) ~ 61 TB\*  
\*Level 2

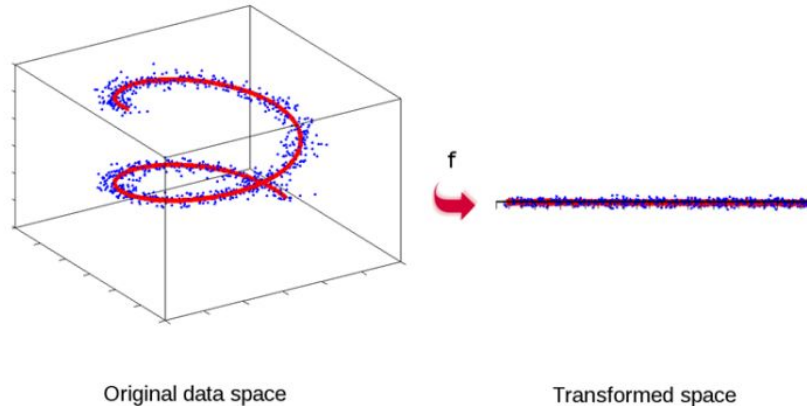


DKIST, EST, SST ~ TB/h/instr

# Exploration and dimensionality reduction

Typical questions of someone that recently got a big dataset:

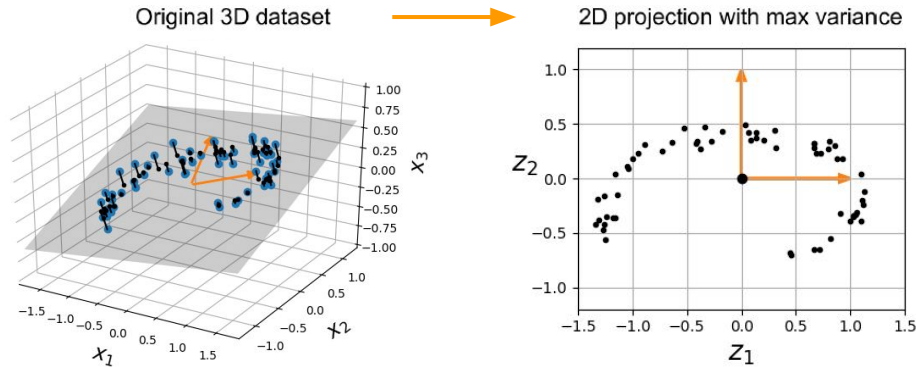
- How diverse vs big is my dataset?
- How many of these features are actually really **informative** or contain just **redundant** information?



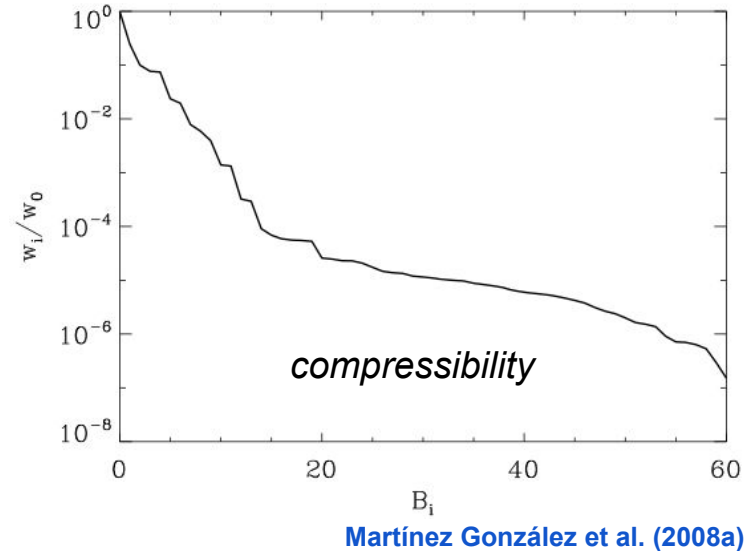
# Dimensionality reduction: Principal Component Analysis

What is the basic idea behind PCA?

- 1.- Find the “principal components” basis
- 2.- Project the data into a truncated version of it.



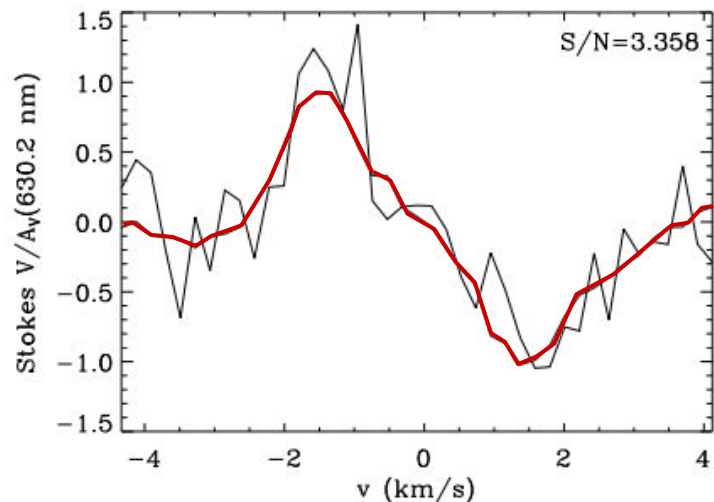
Why is it useful in solar observations?



(e.g. Asensio Ramos et al. 2007; Asensio Ramos & López Ariste 2010)

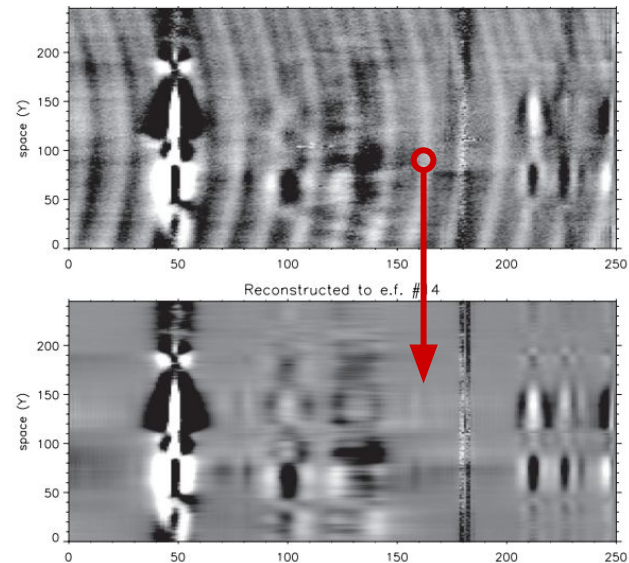
# Dimensionality reduction: PCA applications

## Denoising



Martínez González et al. (2008a,b)

## Removal of fringes

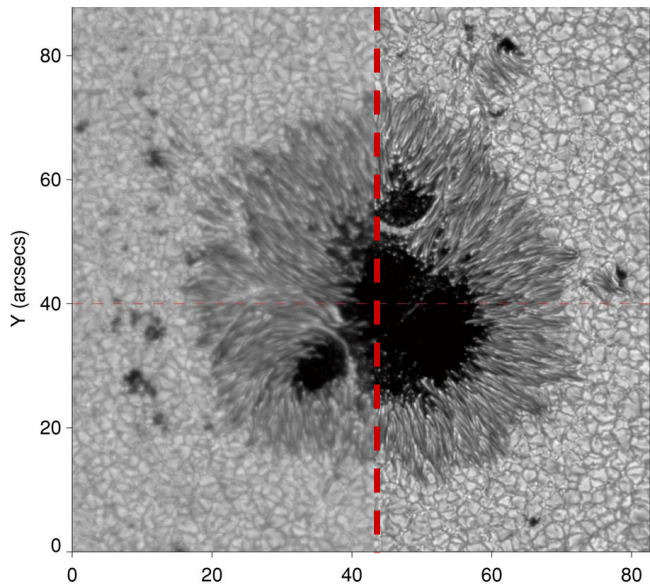


Casini et al. (2012, 2021)

(e.g. Asensio Ramos et al. 2007; Asensio Ramos & López Ariste 2010; Paletou 2012; Pastor Yabar et al. 2018; Trelles Arjona et al. 2021)

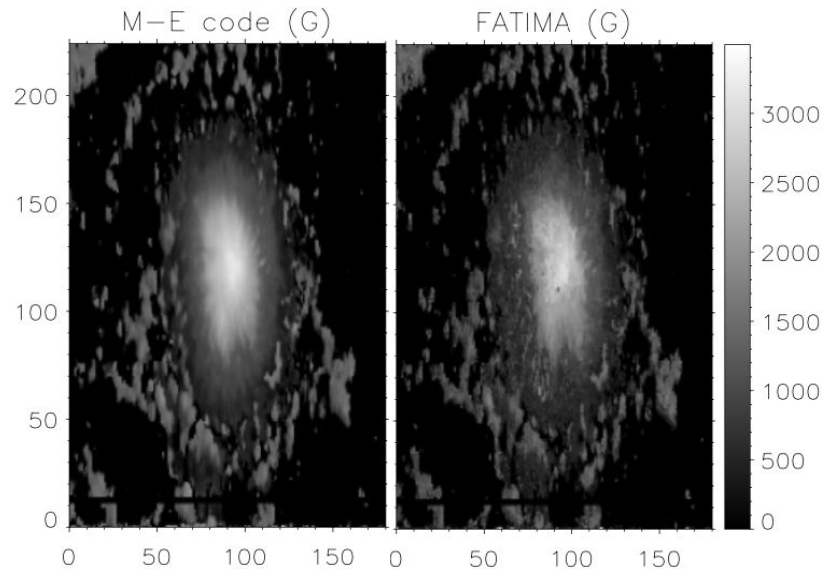
# Dimensionality reduction: PCA applications

## PCA deconvolution



Ruiz Cobo & Asensio Ramos (2013)

## PCA inversion



Socas-Navarro et al. (2001)

(e.g. Quintero Noda et al. 2015, 2016; Felipe et al. 2016, Griñón-Marín 2021)

(e.g. Rees et al. 2000; López Ariste & Casini 2002; Skumanich & López Ariste 2002; Casini et al. 2005; Casini et al. 2009, 2013; Sainz Dalda et al. 2019)

# Finding patterns

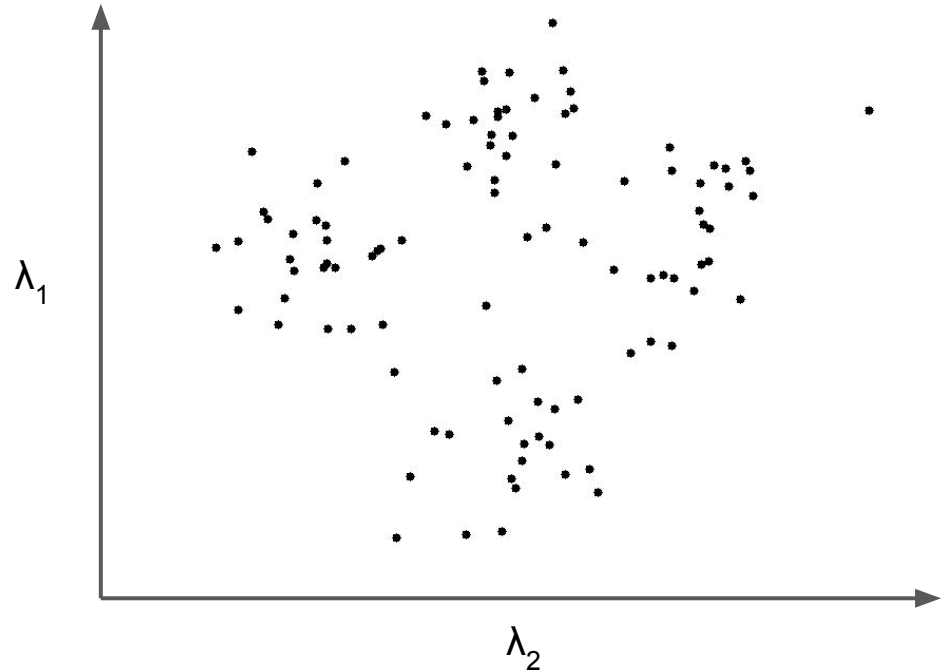
Once we have processed our dataset ...

- Can I find a way to distinguish “groups” of similar properties?
- The conclusions will be the same for all the examples within the group!
  - Where are all these “groups” located in the solar surface?



# Clustering: K-means algorithm

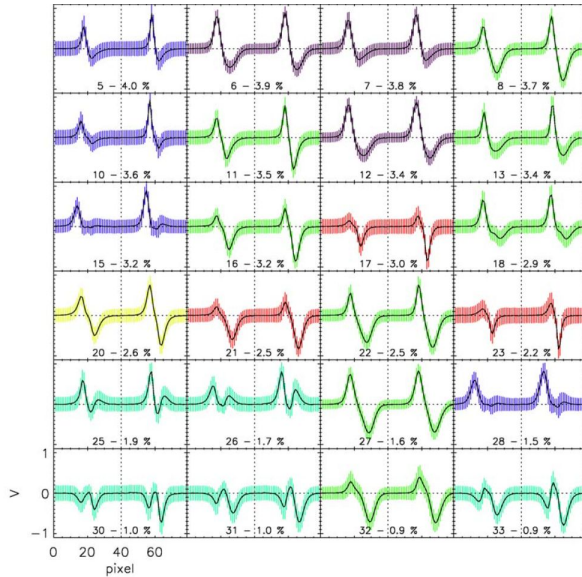
- 1.- Define the K clusters and draw the **centroids**
- 2.- Assign each point to the closest centroid (Euclidean **distance**)
- 3.- The centroids are updated as the average of their cluster



(e.g., Pietarila et al. 2007; Viticchié & Sánchez Almeida 2011; Panos et al. 2018; Sainz Dalda et al. 2019; Bose et al. 2019; Rouppe van der Voort et al. 2021; Robustini et al. 2019; Joshi & Rouppe van der Voort 2020b; Kuckein et al. 2020; Bose et al. 2021a,b; Barczynski et al. 2021; Nóbrega-Siverio et al. 2021; Kleint & Panos 2022; Joshi et al. 2022; Thoen Faber 2022).

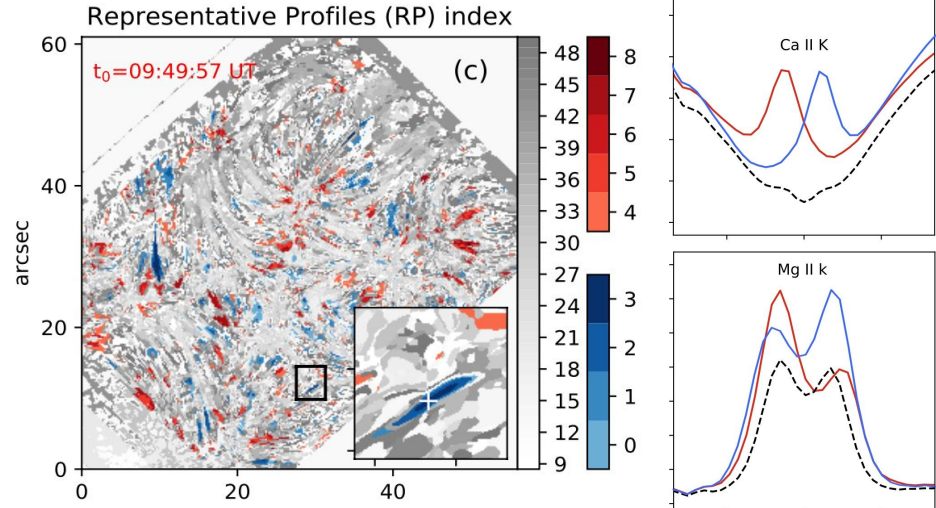
# K-means applications

## QS magnetic field



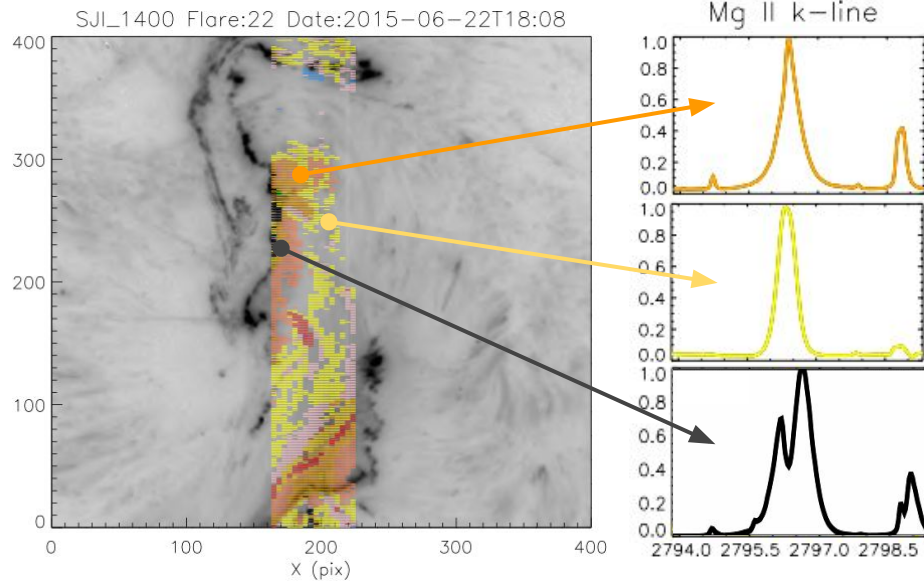
Viticchié & Sánchez Almeida (2011)

## Type-II Spicules

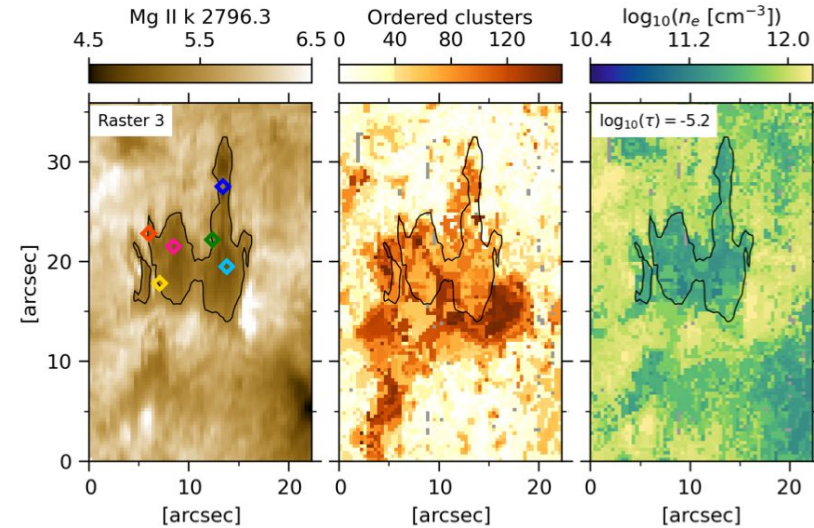


Bose et al. (2019, 2021a,b)

# K-means applications



Brandon Panos et al. (2018)



Nóbrega-Siverio et al. (2021)

(e.g. Magnus Woods et al. 2021)

# PCA and k-means are not the only ones

## Dimensionality reduction

- Feature selection
- Principal Component Analysis (PCA)
- Non-negative matrix factorization (NMF)
- Linear discriminant analysis (LDA)
- Generalized discriminant analysis (GDA)
- Missing Values Ratio
- Low Variance Filter
- High Correlation Filter
- Backward Feature Elimination
- Forward Feature Construction
- t-SNE (T-distributed stochastic neighbour embedding)

## Clustering techniques

- Affinity Propagation
- Agglomerative Hierarchical Clustering
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Gaussian Mixture Models (GMM)
- K-Means
- Mean Shift Clustering
- Mini-Batch K-Means
- OPTICS
- Spectral Clustering

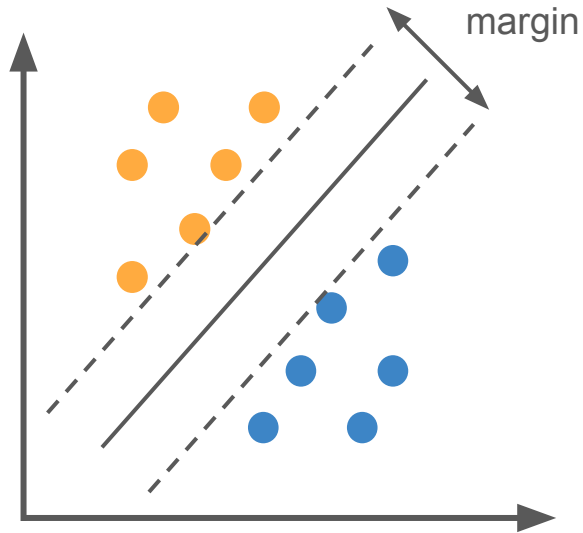
# Classification and prediction

Once you know the interesting part in our dataset ...

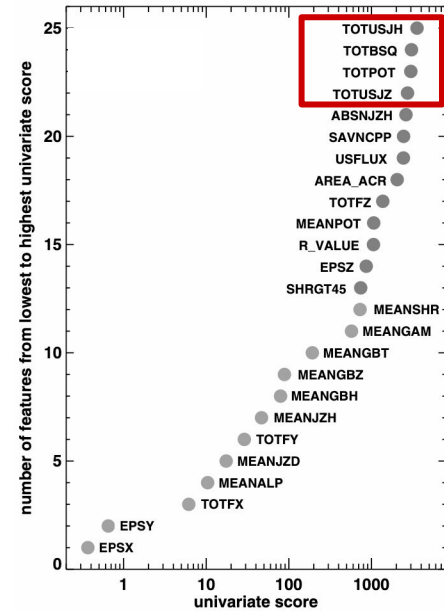
- Can we find a recipe that links our **data** with some **properties**?
  - they can come from different sources
  - just very computationally expensive
  - very difficult to manually find a way
- Can it be general enough to be applied to future data?
- Even more important, can we learn something from it?

# Classification and prediction

## Support Vector Machines



## Flare prediction

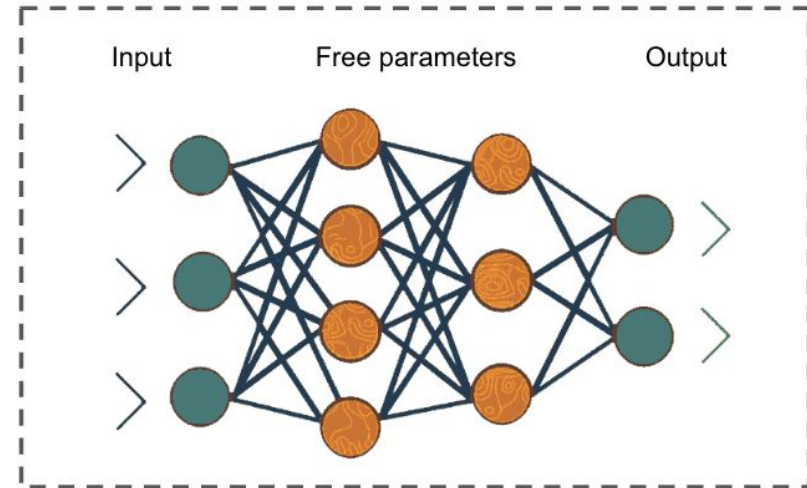
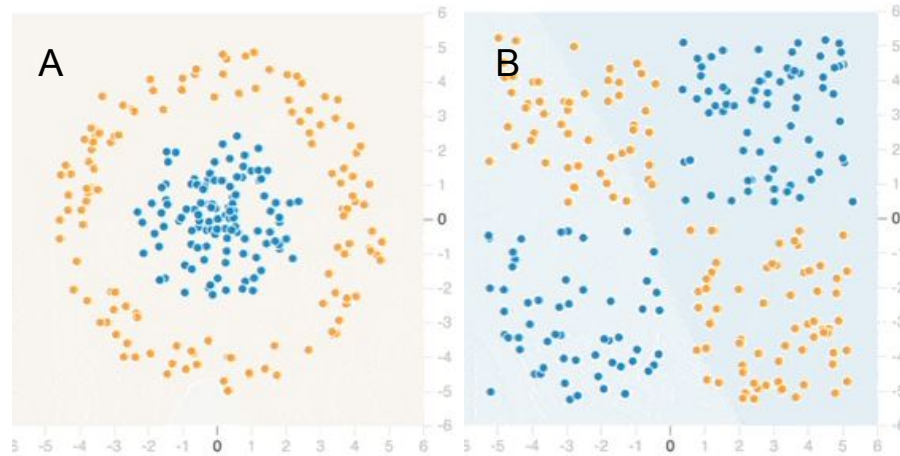


Bobra et al. (2015, 2016)

(e.g. Yuan et al. 2010; Nishizuka et al. 2017; Florios et al. 2018)

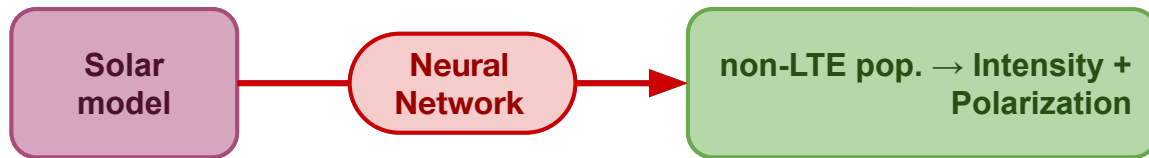
# Nonlinear modeling → Neural networks

What happens if this relation that we try to model is very non-linear?

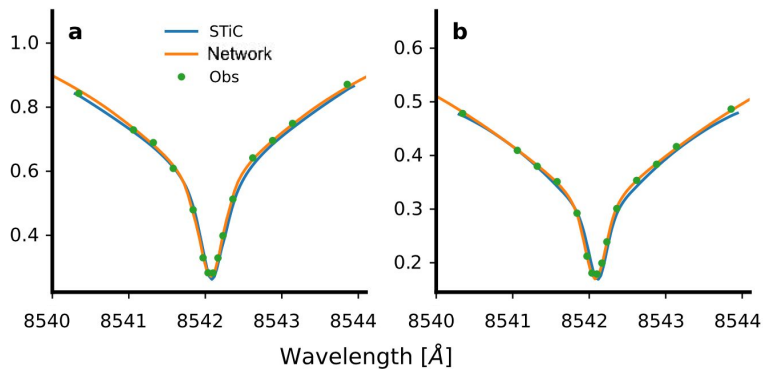


$$o_i = f(\sum_j x_j \cdot w_j + b_j)$$

# NLTE Radiative transfer calculations

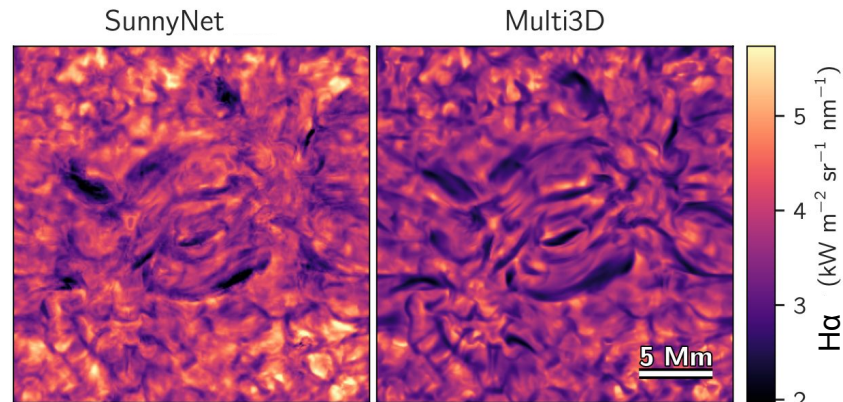


1D / departure coefficients



Vicente Arévalo et al. (2021)

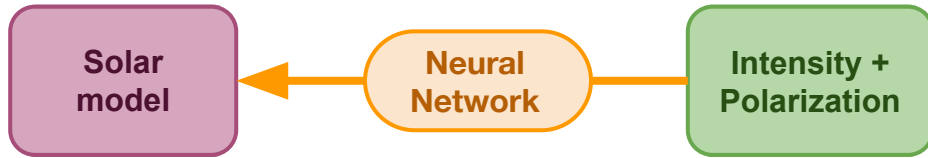
3D / LTE → non-LTE populations



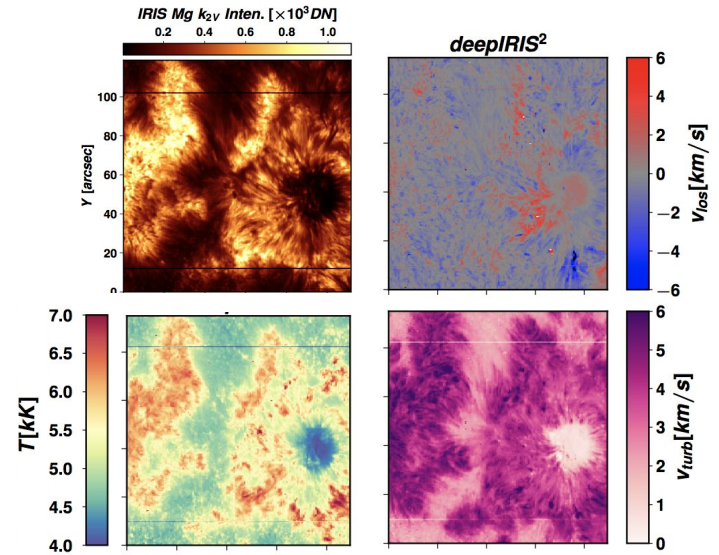
Chappell & Pereira (2021)



# Spectropolarimetric inversions



Synthesis + Inversions  $\sim 10^3 - 10^6$  faster

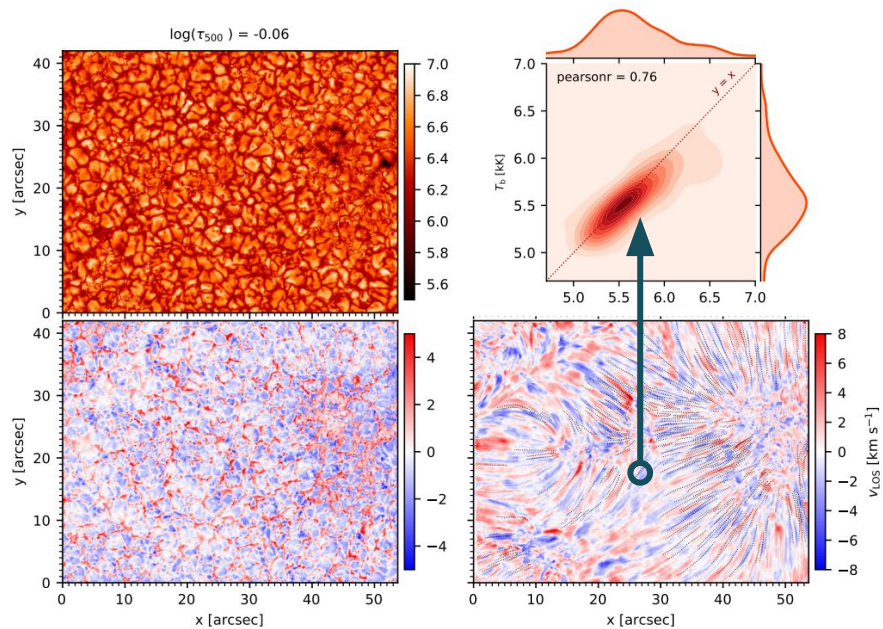


Sainz Dalda et al. (2019)

(e.g. Carroll et al. 2001, 2008; Socas-Navarro, H. 2003, 2005; Sainz Dalda et al. 2019; Milić et al. 2020; Gafeira et al. 2021; Centeno et al. 2022)

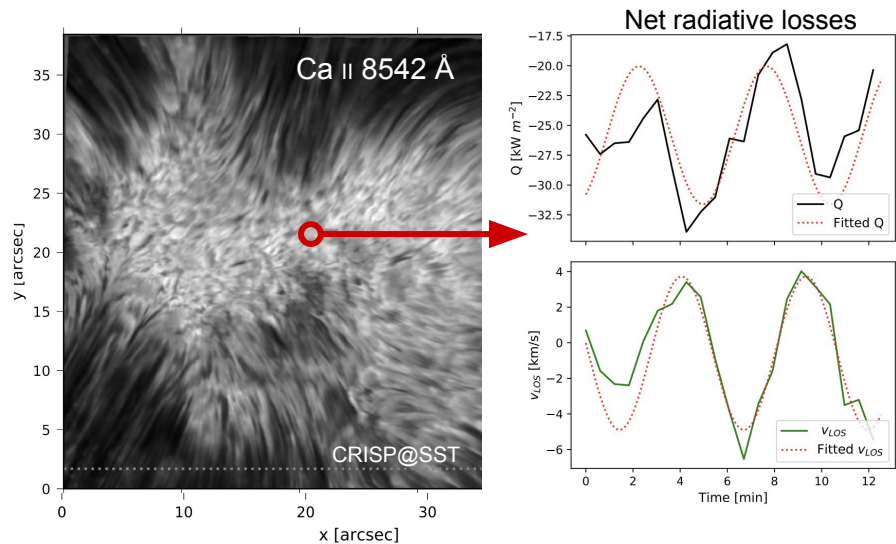
# Accelerating the inference ...

... in large FOVs



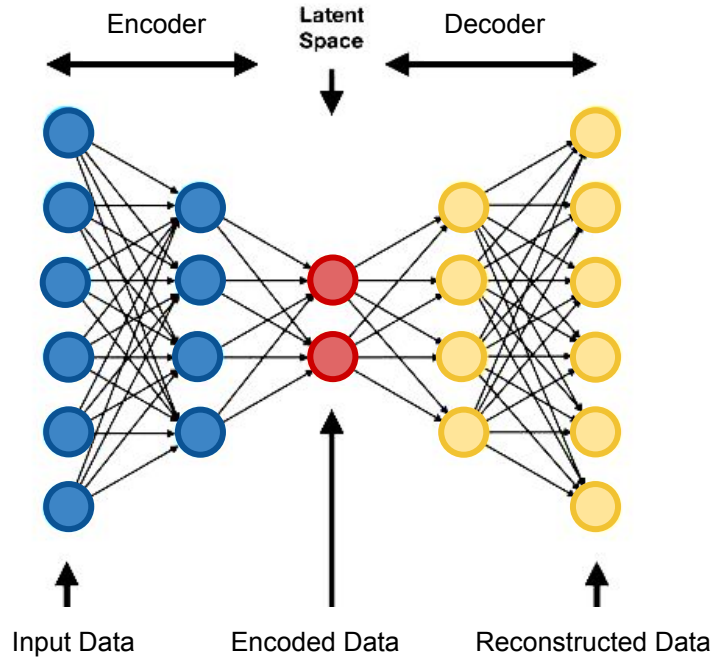
Kianfar S. et al. (2019)

... in long time-series

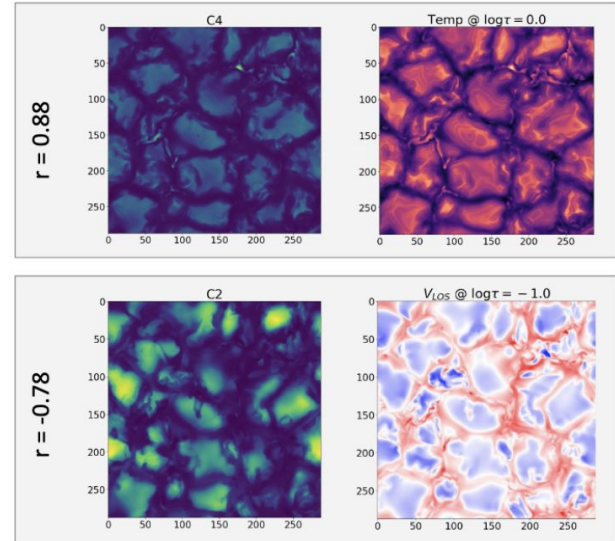


Morosin R. et al. (2022)

# Autoencoders (the non-linear PCA)



## Sparse Representation



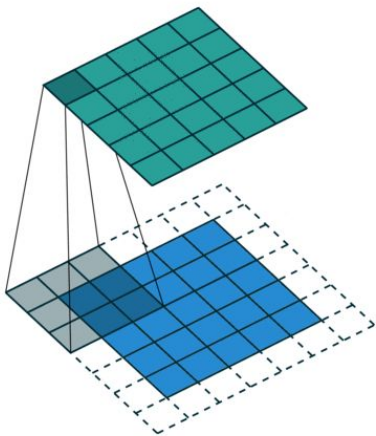
Flint S. & Milić I. (2021)

(e.g. Skumanich & López Ariste 2002; Sadykov et al. 2021; Sergey Ivanov et al. 2021)

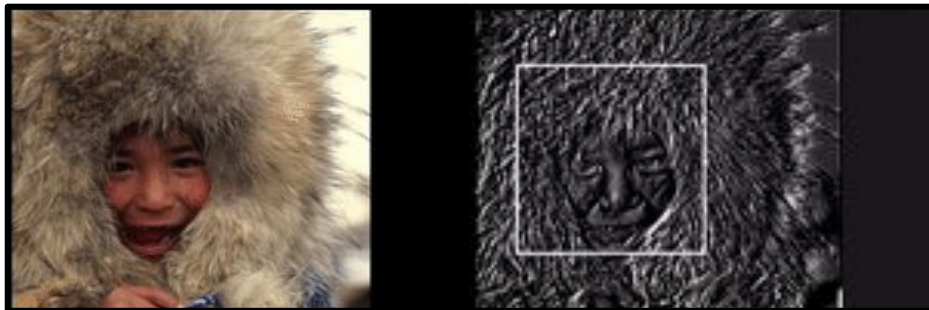
If I want to analyze a megapixel image ( $10^6$ ), do I need a neural network with  $O(>10^6)$  learnable parameters?

## Convolutional Neural Networks

[github.com/vdumoulin/](https://github.com/vdumoulin/)



Input data



Output data

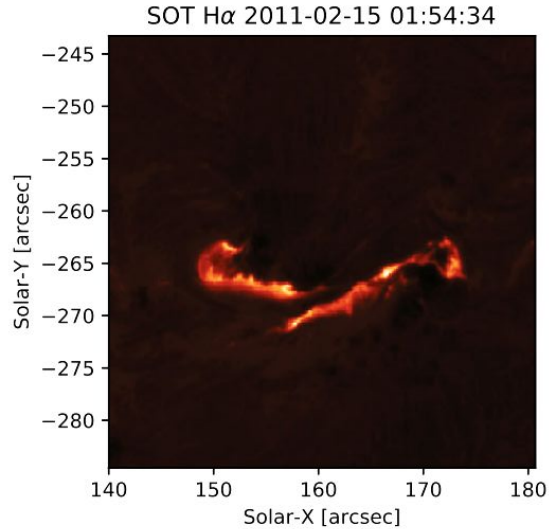
[visual.cs.ucl.ac.uk/pubs/harmonicNets](http://visual.cs.ucl.ac.uk/pubs/harmonicNets)

Translational Equivariance

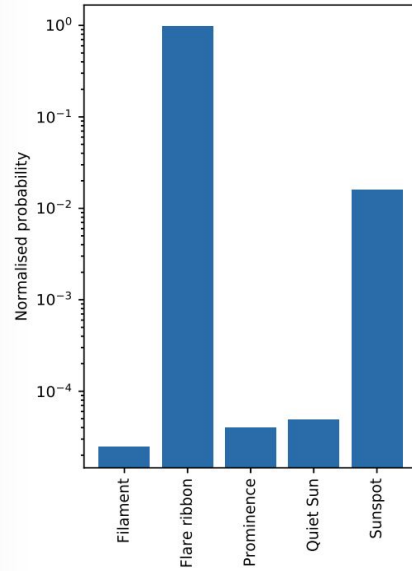
$$f(g(x))=g(f(x))$$

# The image as a “whole”

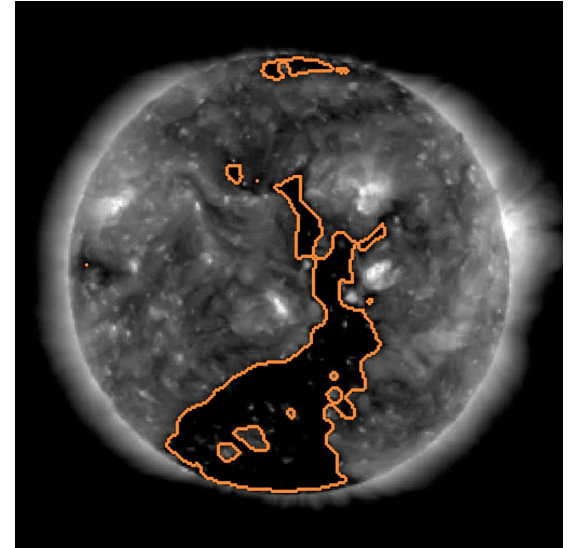
## Automatic catalogues



Armstrong & Fletcher (2019)



## Automatic segmentation

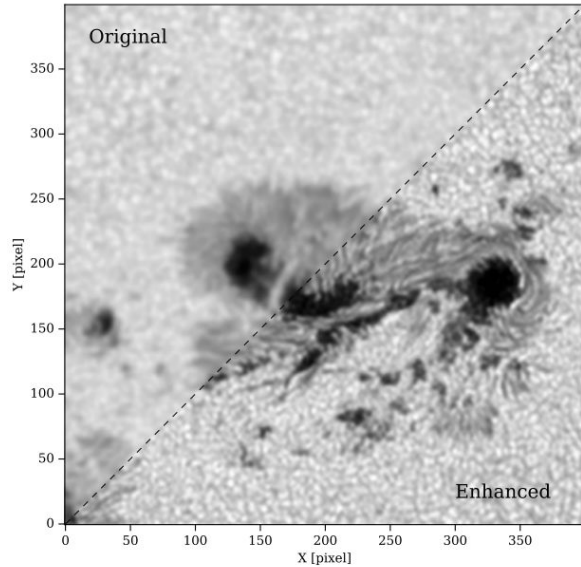


Illarionov & Tlatov (2018)

(e.g. Ahmadzadeh et al 2019; Zhu et al. 2019; Diercke et al 2022)

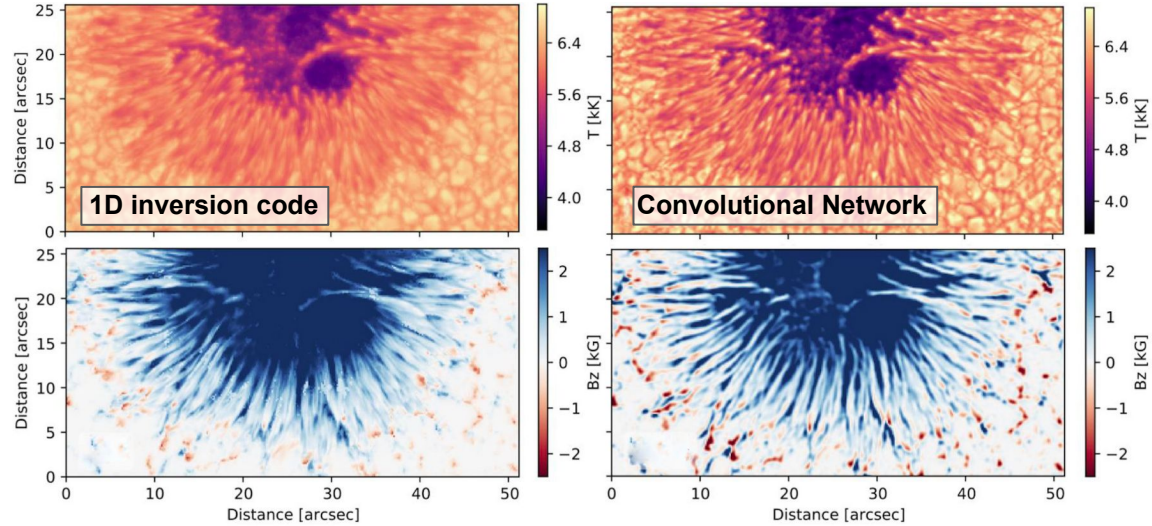
# Using the information from nearby pixels

## Image deconvolution



Díaz Baso et al. (2018)

## Hinode PSF-compensated Stokes inversions

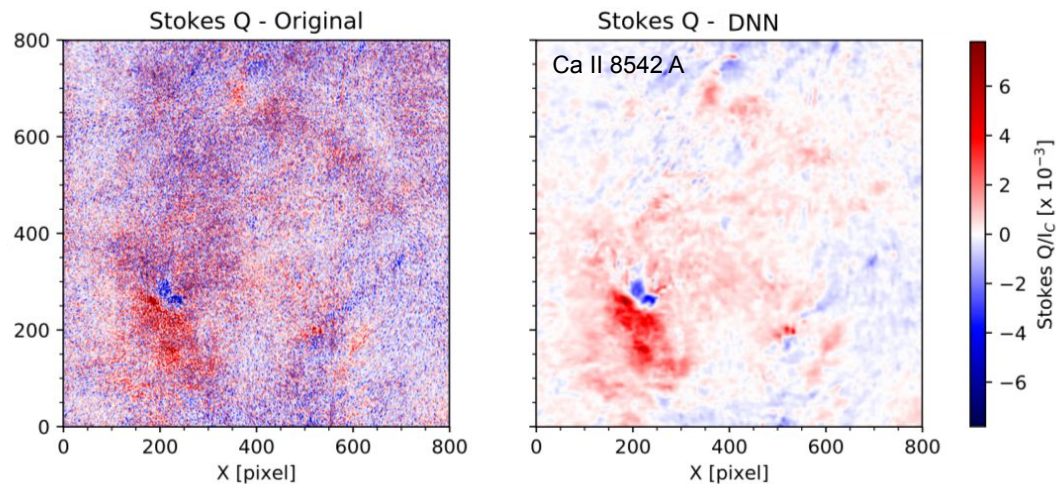


Asensio Ramos & Díaz Baso (2019)

(e.g. Asensio Ramos et al. 2018, 2021; Armstrong et al. 2021; Wang et al 2021; Deng et al 2021)

# CNNs applications

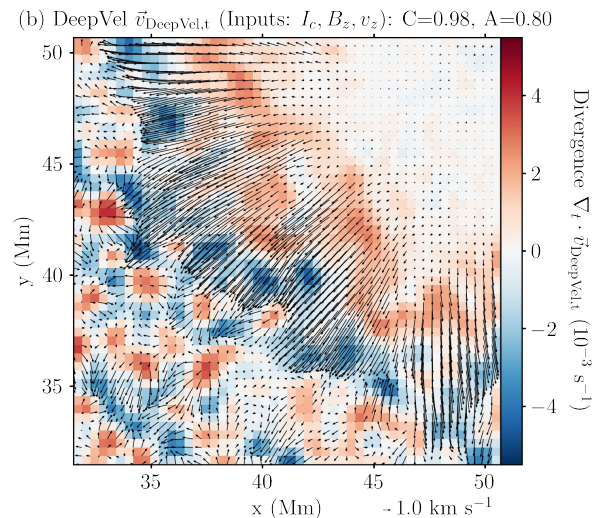
## Solar image denoising



Díaz Baso et. al (2019)

(e.g. Eunsu Park et al. 2020)

## Horizontal velocity fields

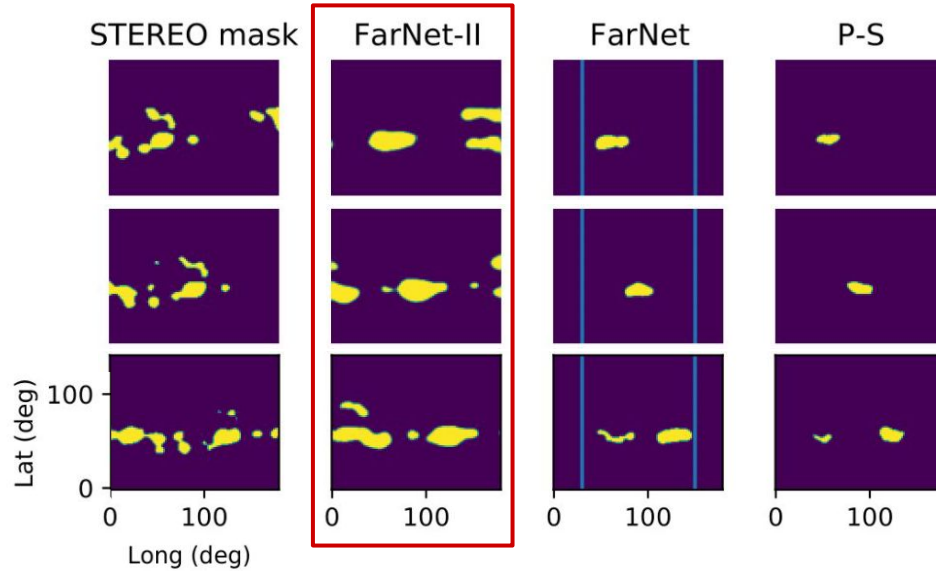


Benoit Tremblay et al. (2020, 2021)

(e.g. Asensio Ramos et al. 2017; Ishikawa et al., 2022)

# Enhancing CNNs with temporal information

## Far-side activity detection



[Broock et al. \(2022\)](#)

(e.g. Felipe et al. 2019; Broock et al 2021; Zeyu Sun et al. 2022)



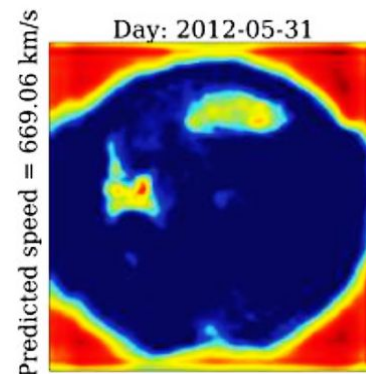
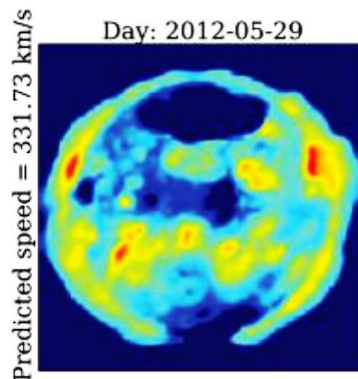
# Challenges and future directions

- Why is this AR classified as a flare-producer?

## Interpretability

To name a few methods:

- Learned Features
- Pixel Attribution (Saliency Maps)
- Testing Concepts
- Adversarial Examples
- Influential Instances
- Symbolic regression



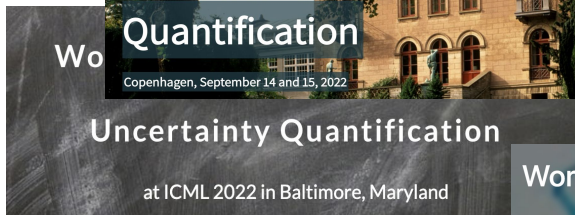
Vishal Upendran et al. (2020)

(e.g. Kangwoo Yi et al. 2021)

# Does your method know when it doesn't know?

## Uncertainty quantification

38th Conference on Uncertainty in Artificial Intelligence  
Eindhoven, Netherlands  
August 1-5, 2022



Symposium on  
Advances in Approximate Bayesian Inference



Bayesian Deep Learning



Workshop on Uncertainty Estimation in Neural Networks

September 14, 2021, Virtual

ICDLBM 2022: 16. International Conference on Deep Learning and Bayesian Machine Learning  
October 13-14, 2022 in London, United Kingdom

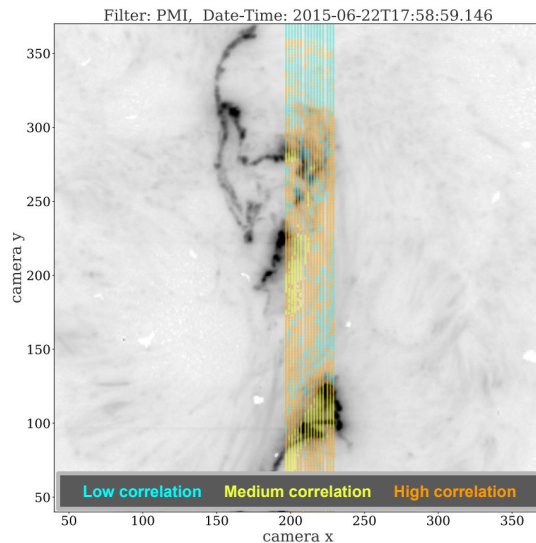
Artificial Intelligence and Statistics  
2022



21 Workshop on Information-Theoretic Methods for Rigorous, Interpretable, and Reliable Machine Learning

# A probabilistic perspective: conditional information

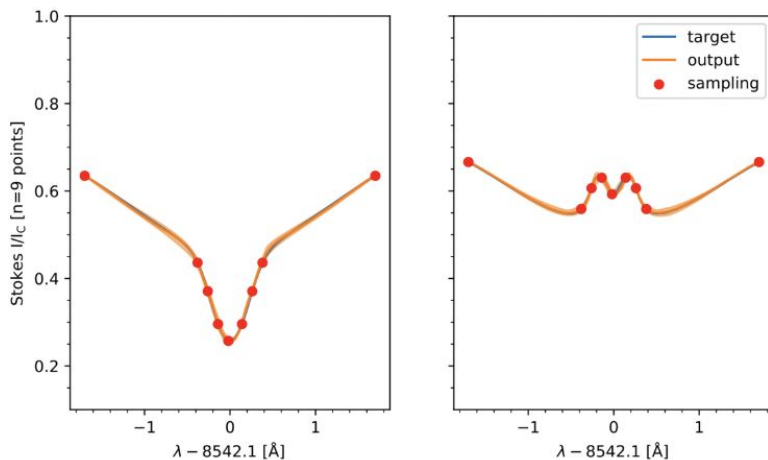
## Mutual information



Panos et al. (2021a,b)

(e.g. Snelling et al. 2020)

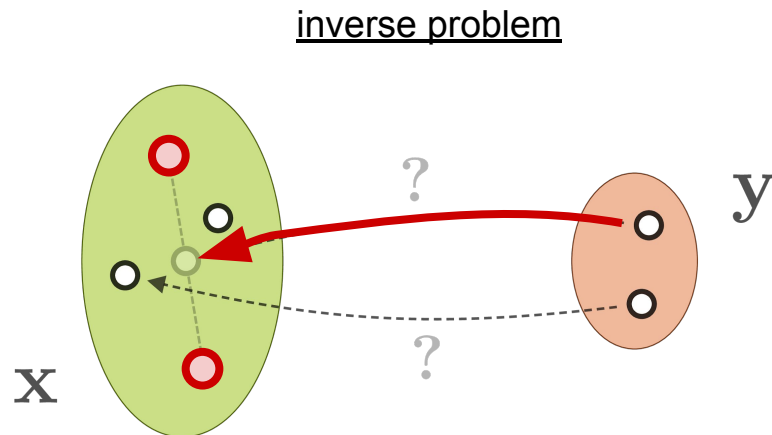
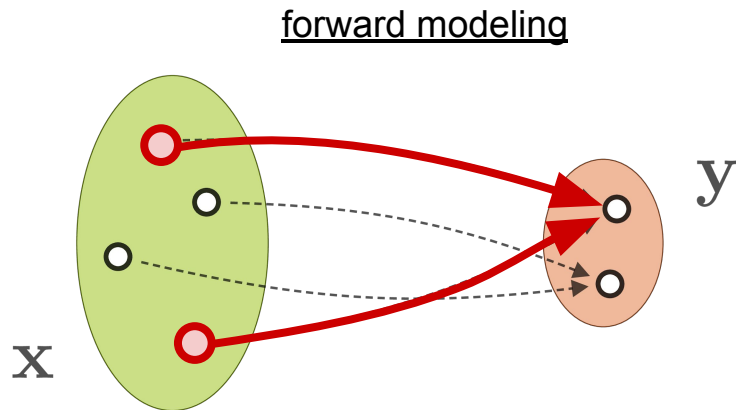
## Designing observing sampling



Díaz Baso et al. (in prep)

(e.g. Szenicer et al. 2019; Lim et al. 2021; Salvatelli et al. 2022)

# A probabilistic perspective: inverse problems

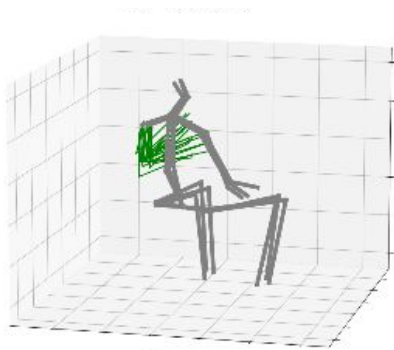


# Inherent in many problems

## Pose estimation



Input image



3D Pose

Wehrbein et al. (2021)

## Super-Resolution



LR Input

Input image



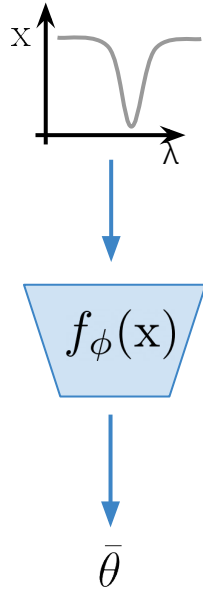
Output: **Distribution**

Lugmayr A. et al. (2020)

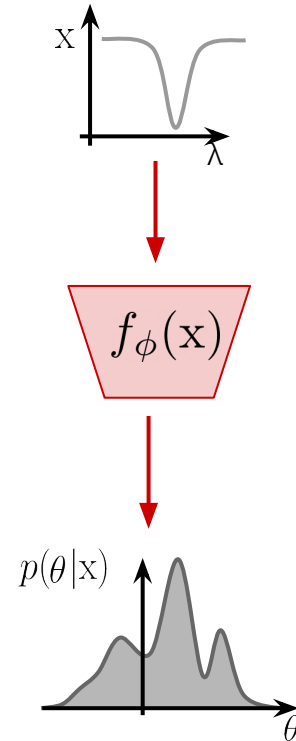
# Normalizing flows

Rezende & Mohamed (2015), Dinh et al. (2016)

NNs

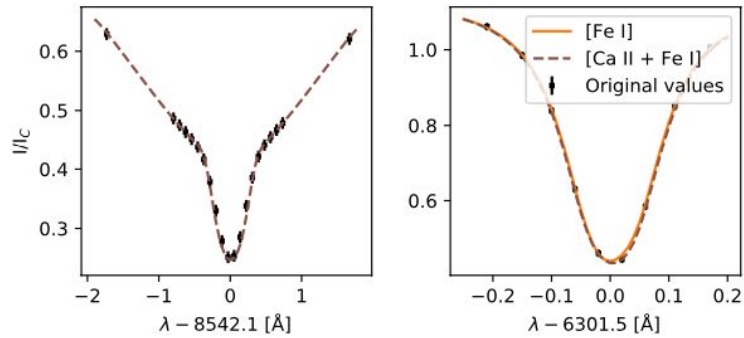


NFlows



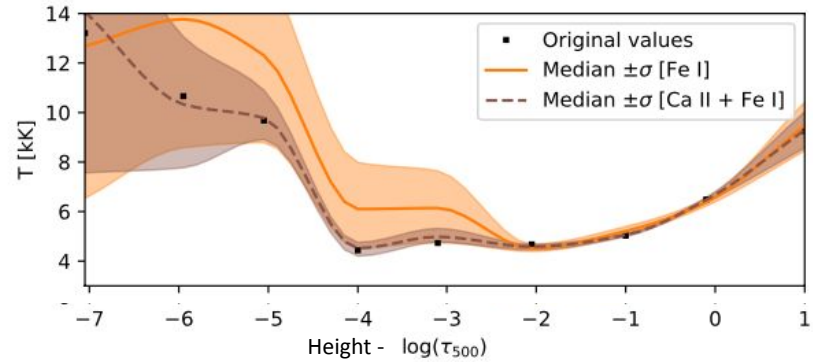
# Normalizing flows

Rezende & Mohamed (2015), Dinh et al. (2016)



Díaz Baso et al. (2021)

## N-LTE inversion



→ Only using the Fe I 6301 line

→ Also using the Ca II 8542 profile

(e.g. Osborne et al. 2019, Asensio Ramos et al. 2021)

# Normalizing flows → Diffusion models

Sohl-Dickstein et al. (2015), Yang & Ermon (2019), Ho et al. (2020)



Irish Terrier riding a horse in Patagonia and playing the harmonica



Cat with a yellow hat going down the stairs under water



Panda mad scientist mixing sparking chemicals, artstation



Grizzly bear taking a selfie on the Golden Gate bridge on a windy day

**Ramesh et al. (2022)**

→ valuable effort in complex inverse problems.



# Summary and conclusions

- Machine learning can help in many different ways: **explore patterns, image reconstruction, compression, denoising, parameter inference, classification, and tracking**, etc. Inference is extremely fast.

# Summary and conclusions

- Machine learning can help in many different ways: **explore patterns, image reconstruction, compression, denoising, parameter inference, classification, and tracking**, etc. Inference is extremely fast.
- The question you want to address is as important as the method. Depending on the goal, no need to always "reinvent the wheel" (**literature vs own design**).

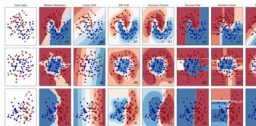
(e.g. **scikit-learn** - python)

## Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, and more...

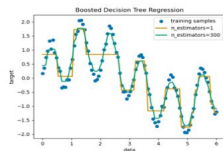


## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, nearest neighbors, random forest, and more...

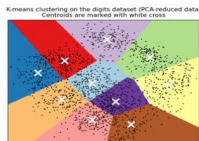


## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, and more...

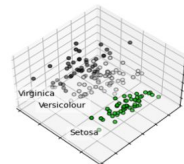


## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization, and more...

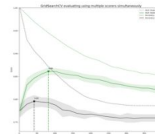


## Model selection

Comparing, validating and choosing parameters and models.

**Applications:** Improved accuracy via parameter tuning

**Algorithms:** grid search, cross validation, metrics, and more...

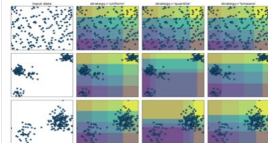


## Preprocessing

Feature extraction and normalization.

**Applications:** Transforming input data such as text for use with machine learning algorithms.

**Algorithms:** preprocessing, feature extraction, and more...



# Summary and conclusions

- Machine learning can help in many different ways: **explore patterns, image reconstruction, compression, denoising, parameter inference, classification, and tracking**, etc. Inference is extremely fast.
- The question you want to address is as important as the method. Depending on the goal, no need to always "reinvent the wheel" (**literature vs own design**).
- There are still many ways to improve them: incorporating **physical constraints** (e.g. symmetries, conservation laws), making them **interpretable**, quantifying **uncertainty**, enabling **multimodal** solutions, etc.

*What a time to be alive!*