Manage and Mine Geoscience Data for *Your* CEDAR Science Breakthroughs

Tomoko Matsuo
Ann and H.J. Smead Department of Aerospace Engineering Sciences
University of Colorado at Boulder


ITALY

Machine Learning

In machine learning, **statistical learning** techniques are used to automatically identify patterns in data. Most statistical learning problems fall into one of two categories:

- Supervised Learning

  Learning process is guided by a set of labeled samples $\{x_i, y_i\}$ *(training data), where* x_i *is the predictor measurement and* y_i *is an associated response measurement.*

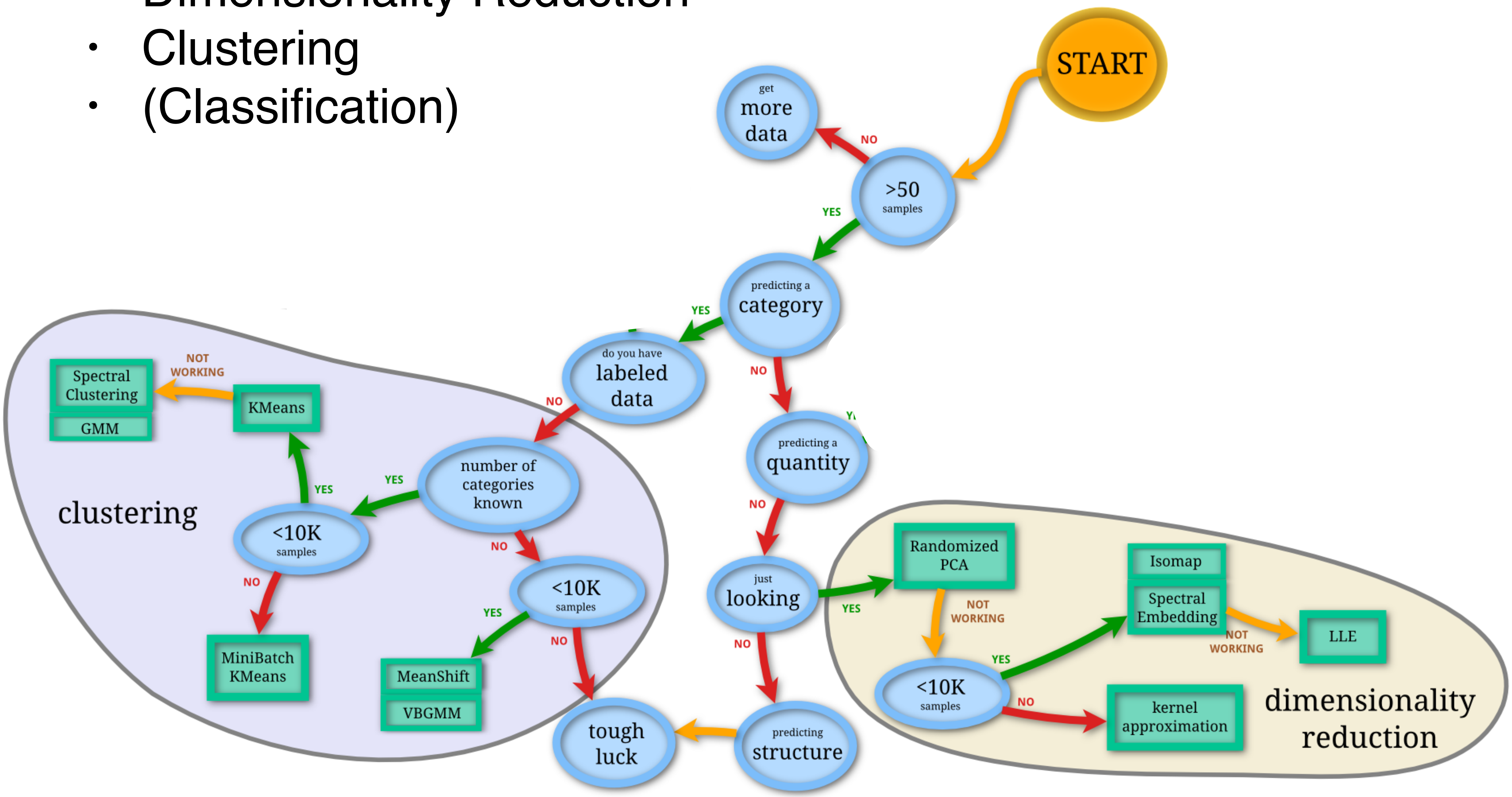
- Unsupervised Learning

  No training data is used to supervise learning process.
  Only $\{x_i\}$ *is known.*

Big Picture: Learning Techniques

Unsupervised Methods

- Dimensionality Reduction
- Clustering
- (Classification)

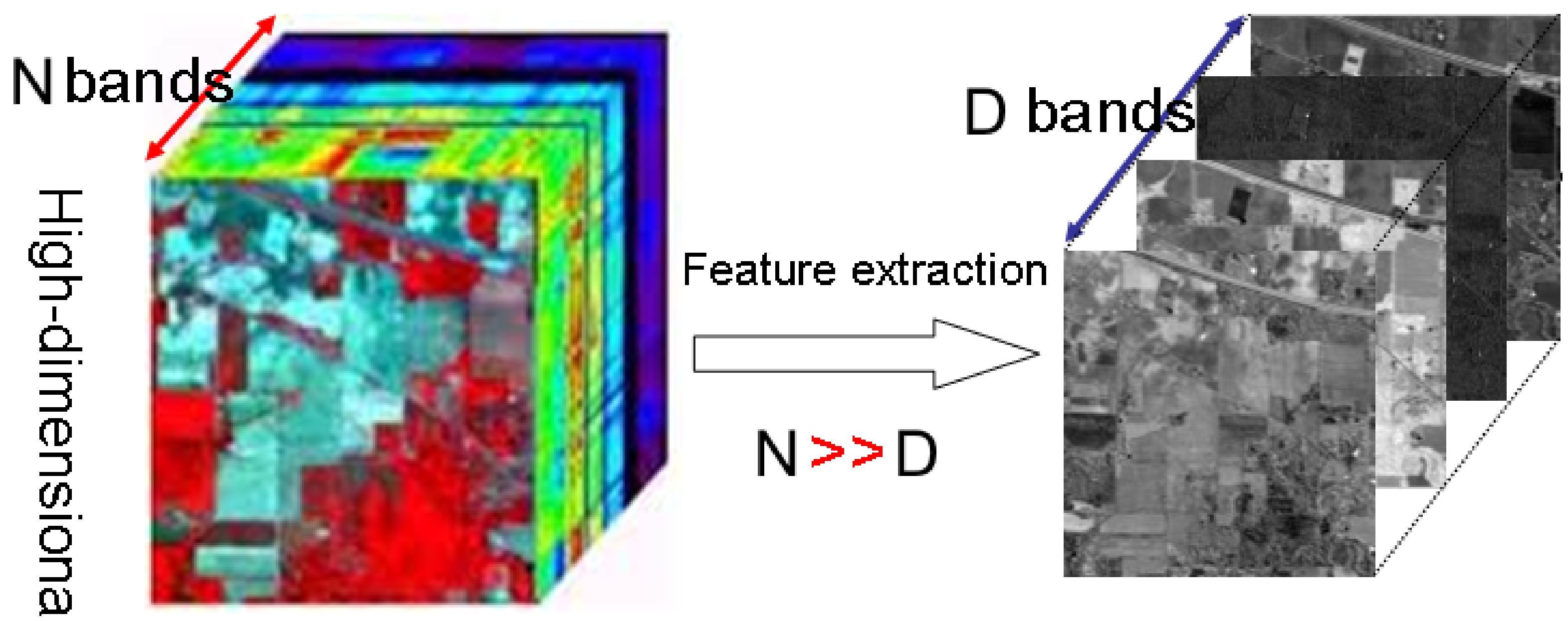


http://scikit-learn.org/stable/tutorial/machine_learning_map/

Data Dimensionality Reduction

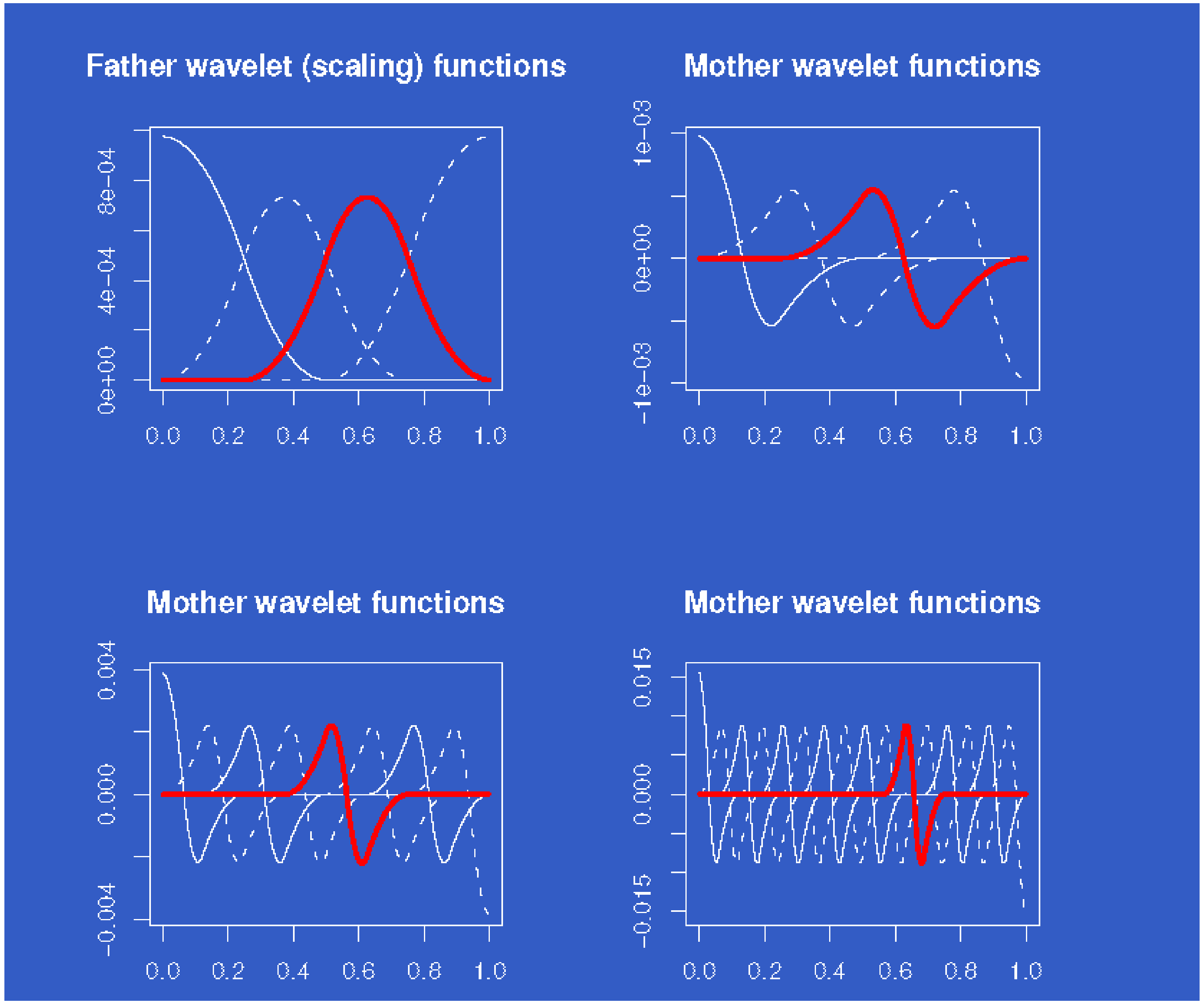
Examples of Linear Projection Methods

- Wavelet Compression
- Principal Component Analysis (PCA)



Wavelet Compression

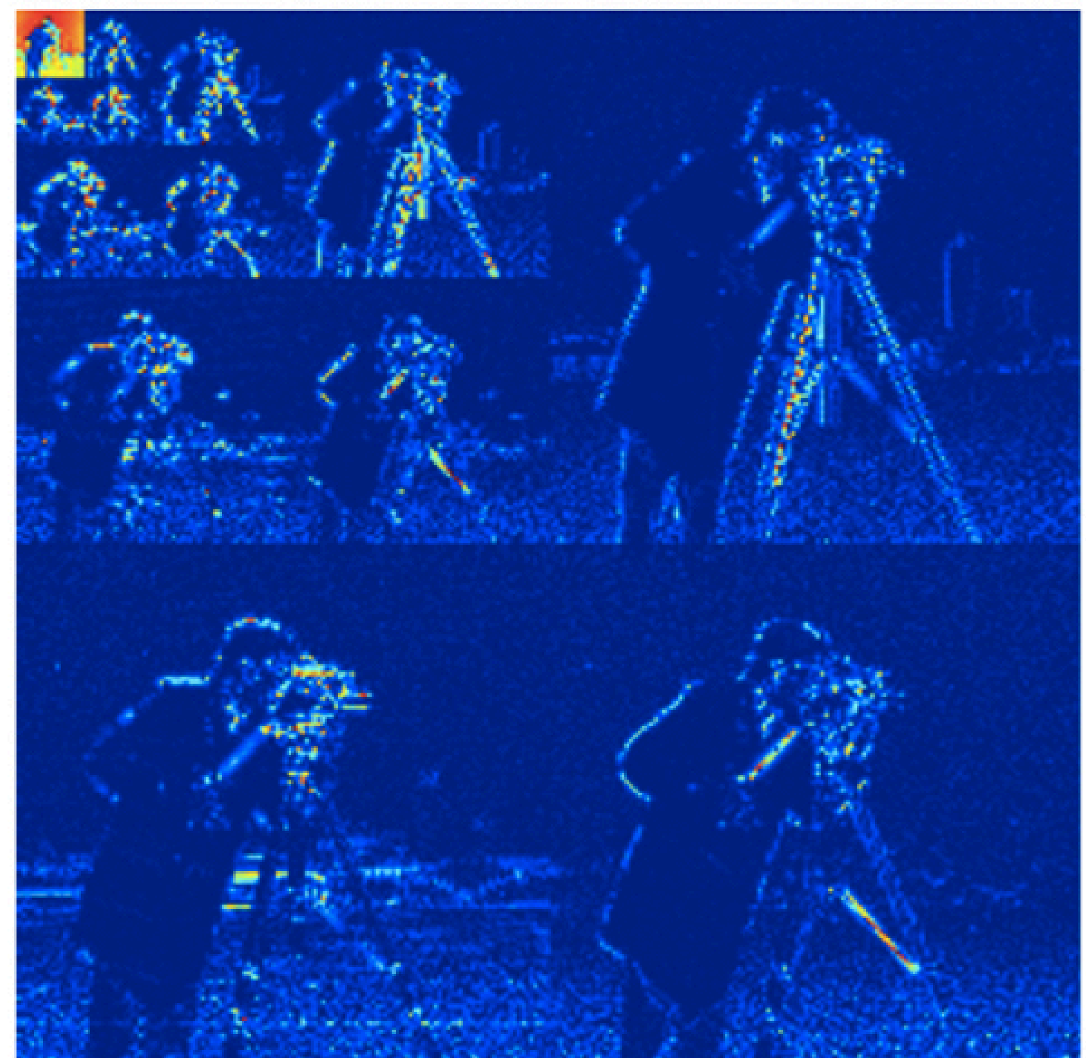
An orthonormal basis vector $\mathbf{\Psi} \in \mathbb{R}^{N \times N}$ where $\mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{I}$



Wavelet Compression

- Suppose $\mathbf{x} \in \mathbb{R}^N$ can be expanded to $\mathbf{\Psi}$ as
$$\mathbf{x} = \mathbf{\Psi}\mathbf{c}$$
where $\mathbf{c}$ contains coefficients.

256×256 image

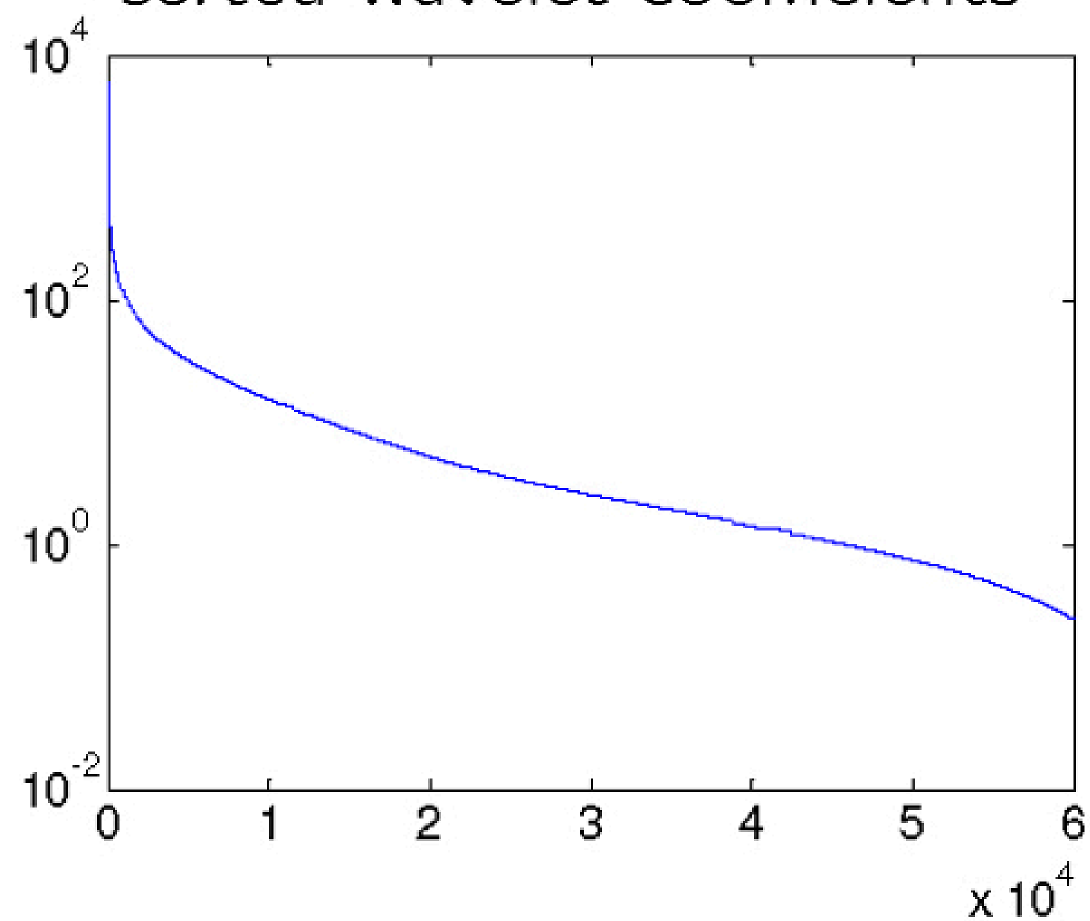


Wavelet Compression

- Compressible signals are well approximated by D-sparse representations, meaning that only D of $\{c_i\}_{i=1}^N$ are nonzero.

256×256 image

sorted wavelet coefficients



Wavelet Compression

original

D-term approximation

$D = 0.1\ N$

Representation of a Stochastic Process Using $\boldsymbol{\Psi}$

- Suppose that $\mathbf{x} = \{x_n\}_{n=1}^{N}$ is now a centered Gaussian stochastic process.

Representation of a Stochastic Process Using $\boldsymbol{\Psi}$

- Suppose that $\mathbf{x} = \{x_n\}_{n=1}^{N}$ is now a centered Gaussian stochastic process.
- Karhunen-Loéve transform

$$x_n = \sum_{i=1}^{\infty} c_i \Psi_i(n)$$

where coefficients c_i are independent Gaussian random variables.

Representation of a Stochastic Process Using $\boldsymbol{\Psi}$

- In the discrete case, KarhunenLoéve transform can be approximated as

$$\mathbf{x} \approx \boldsymbol{\Psi}\mathbf{c}$$

Representation of a Stochastic Process Using $\boldsymbol{\Psi}$

- In the discrete case, KarhunenLoéve transform can be approximated as

$$\mathbf{x} \approx \boldsymbol{\Psi}\mathbf{c}$$

- Covariance of $\mathbf{x}$ is then

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] \approx \boldsymbol{\Psi}\mathbb{E}[\mathbf{c}\mathbf{c}^T]\boldsymbol{\Psi}^T$$

Representation of a Stochastic Process Using $\boldsymbol{\Psi}$

- In the discrete case, KarhunenLoéve transform can be approximated as

$$\mathbf{x} \approx \boldsymbol{\Psi}\mathbf{c}$$

- Covariance of $\mathbf{x}$ is then

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] \approx \boldsymbol{\Psi}\mathbb{E}[\mathbf{c}\mathbf{c}^T]\boldsymbol{\Psi}^T$$

- The columns of $\boldsymbol{\Psi}$ are *principal components* if $\mathbb{E}[\mathbf{c}\mathbf{c}^T]$ is diagonal (close to diagonal) so that $\mathbf{c}$ is uncorrelated.

Representation of a Stochastic Process Using $\boldsymbol{\Psi}$

- In the discrete case, KarhunenLoéve transform can be approximated as

$$\mathbf{x} \approx \boldsymbol{\Psi}\mathbf{c}$$

- Covariance of $\mathbf{x}$ is then

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] \approx \boldsymbol{\Psi}\mathbb{E}[\mathbf{c}\mathbf{c}^T]\boldsymbol{\Psi}^T$$

- The columns of $\boldsymbol{\Psi}$ are *principal components* if $\mathbb{E}[\mathbf{c}\mathbf{c}^T]$ is diagonal (close to diagonal) so that $\mathbf{c}$ is uncorrelated.

- Usually *principal components* can be estimated from factorization of a sample covariance, e.g., by eigenvalue decomposition,

$$\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$$

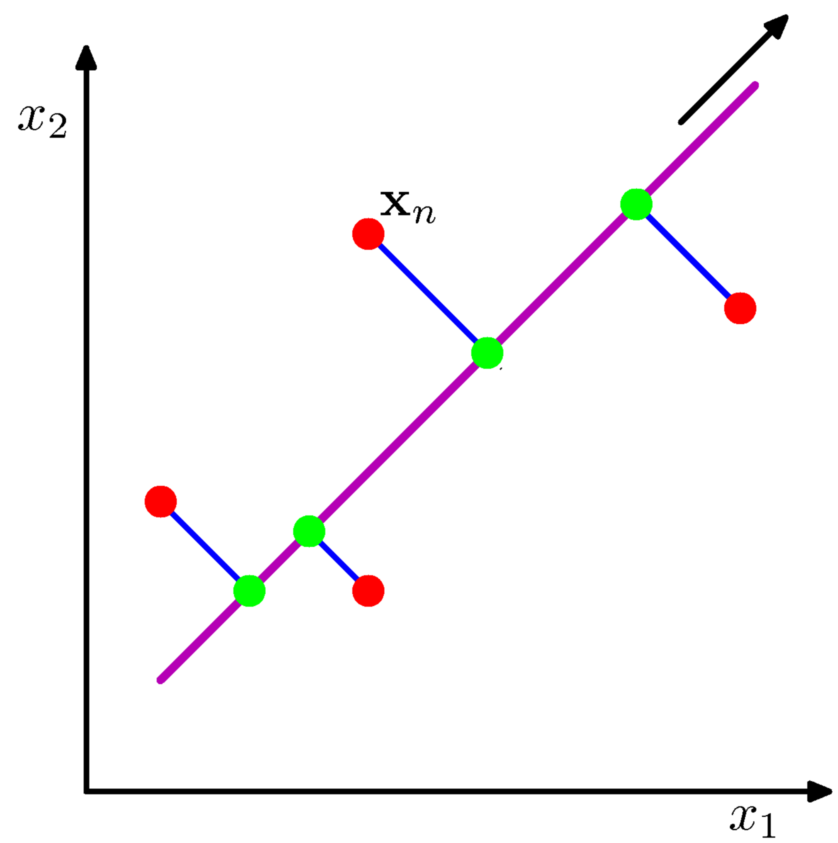
where $\boldsymbol{\Sigma}$ is symmetric, $\mathbf{V}$ is orthogonal, and $\boldsymbol{\Lambda}$ is diagonal.

Principal Component Analysis

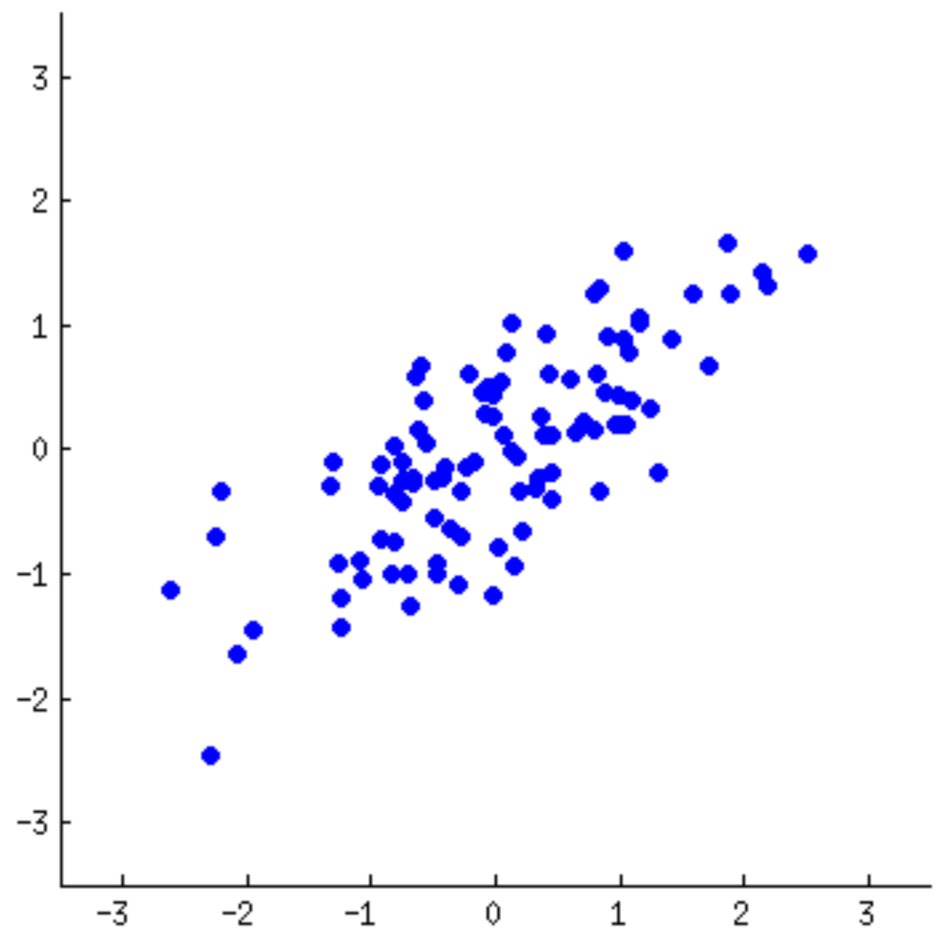
- *Principal components* are essentially eigenvectors.

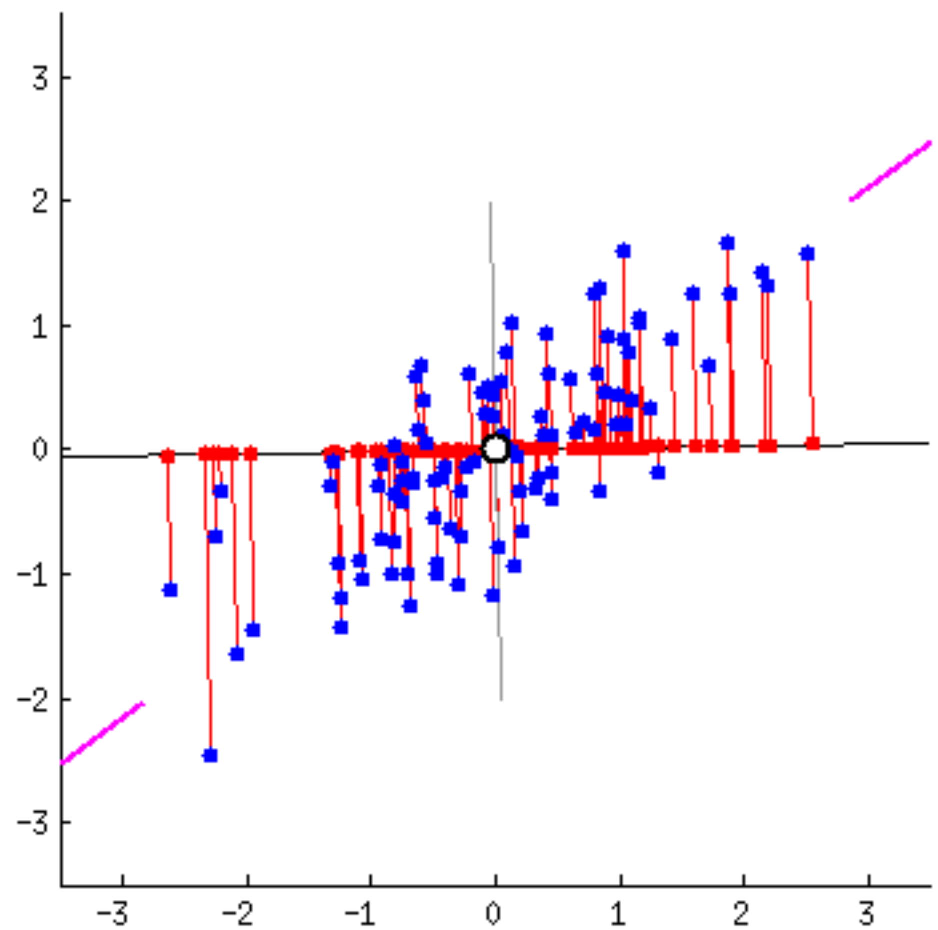
Principal Component Analysis

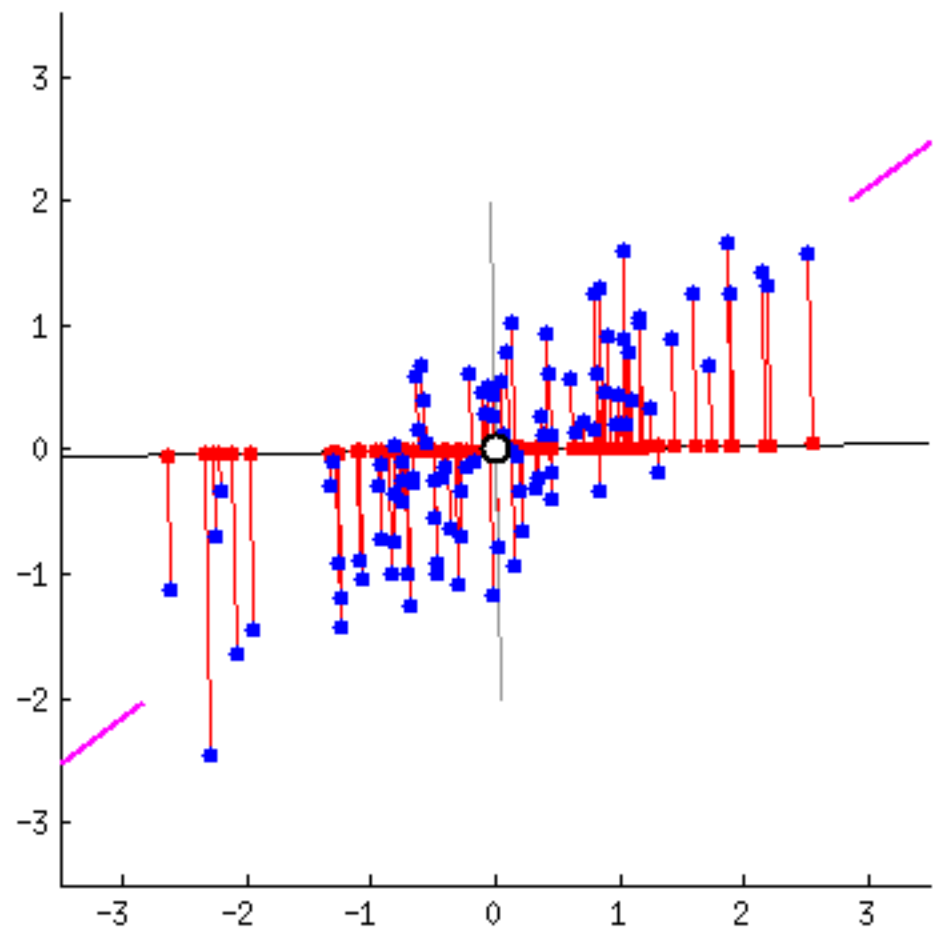
- *Principal components* are essentially eigenvectors.
- The Principal Component Analysis (PCA) referers to orthogonal projection of the data onto a lower-dimensional space spanned by these eigenvectors that maximizes the variance of projected data.



Bishop, 2006


3
2
1
0
-1
-2
-3
-3 -2 -1 0 1 2 3


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3



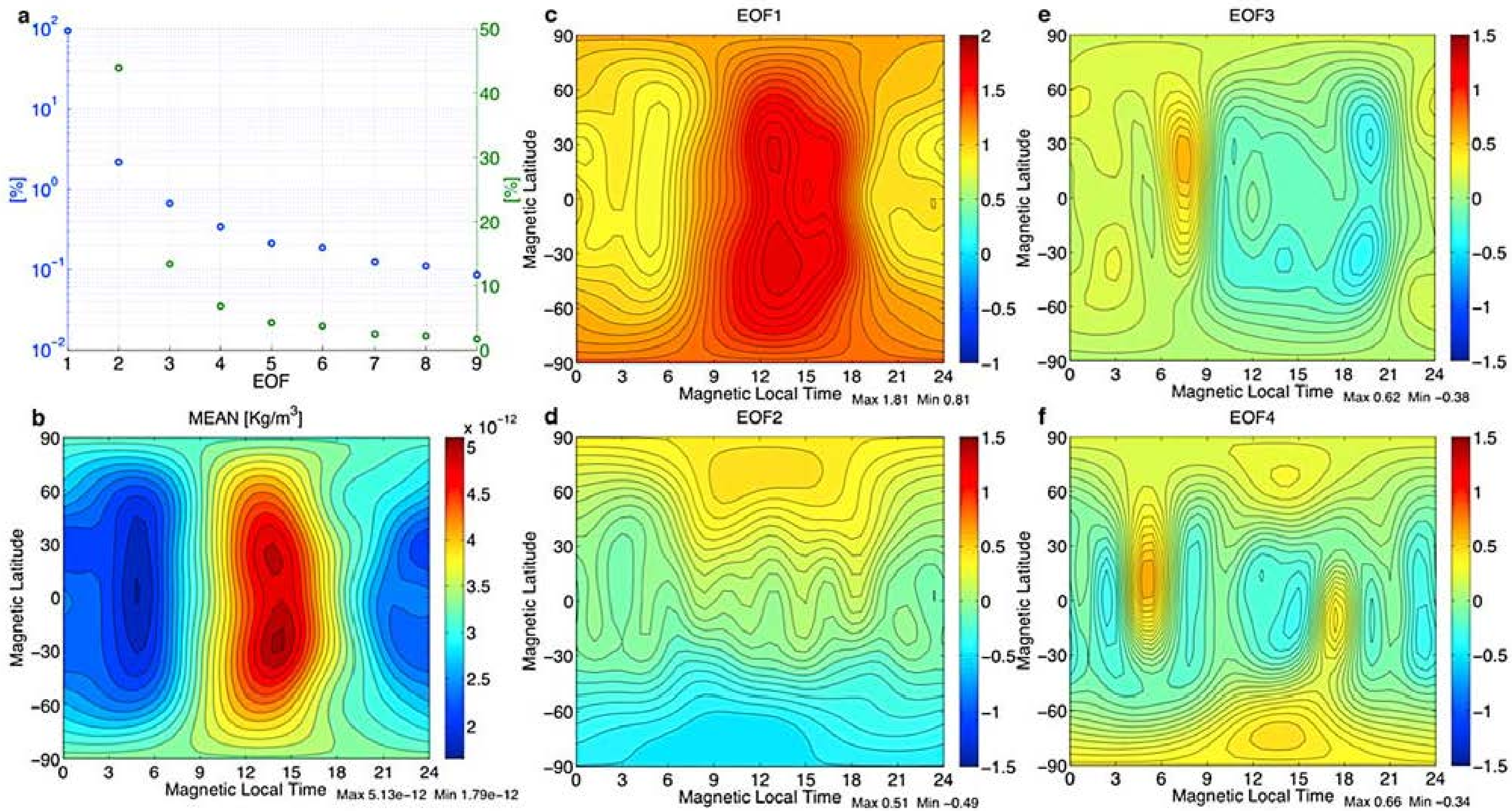
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

PCA Example - Thermospheric Mass Density

PCA of 9-year CHAMP mass density data

- Let's suppose that the data can be decomposed with respect to principal components $\mathbf{\Psi}$ as:

$$x(s,t) \approx c_1(t)\Psi_1(s) + c_2(t)\Psi_2(s) + c_3(t)\Psi_3(s) + \cdots$$



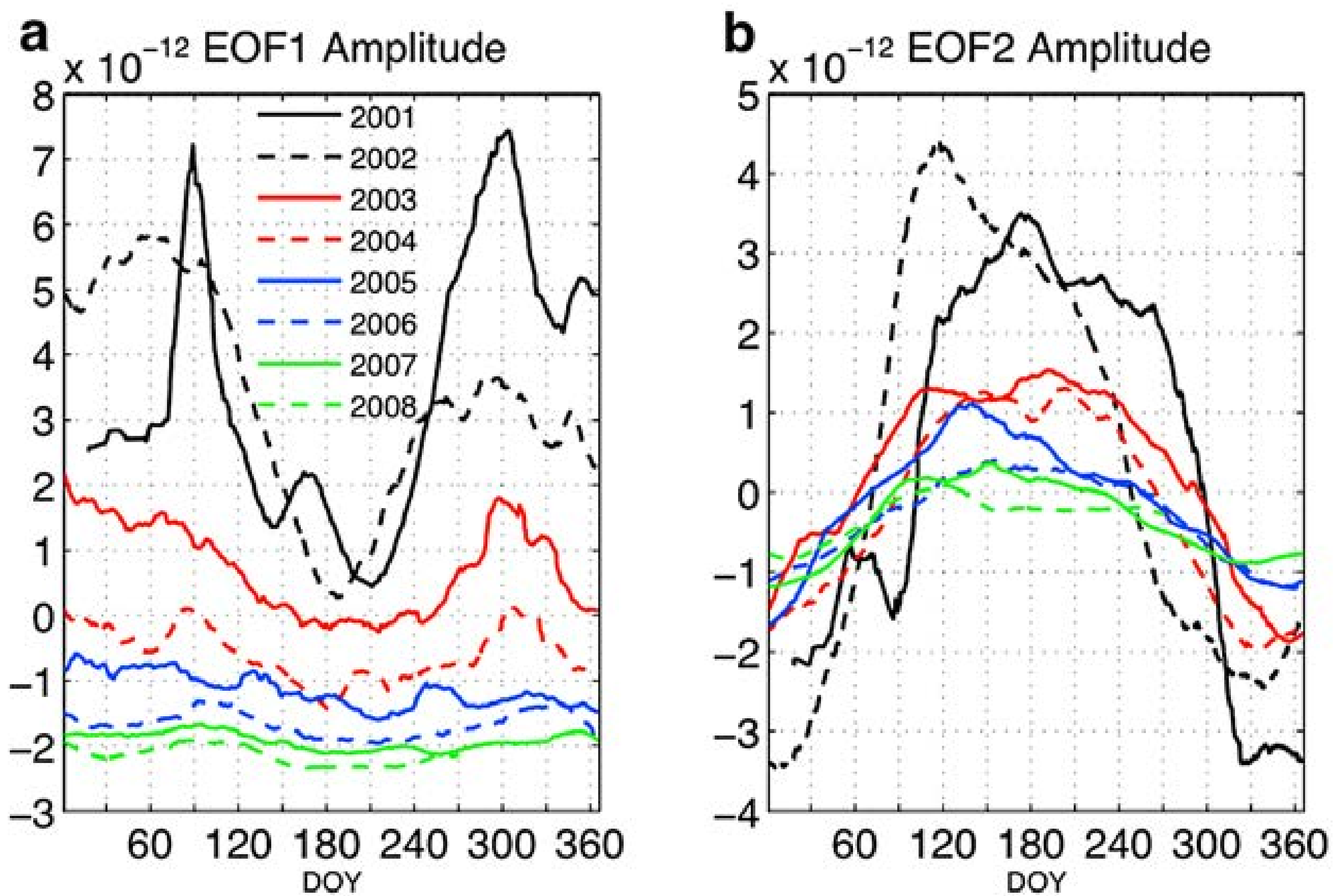
Matsuo and Forbes, 2008

PCA Example - Thermospheric Mass Density

PCA of 9-year CHAMP mass density data

- Let's suppose that the data can be decomposed with respect to principal components $\mathbf{\Psi}$ as:

$$x(s,t) \approx c_1(t)\Psi_1(s) + c_2(t)\Psi_2(s) + c_3(t)\Psi_3(s) + \cdots$$



Matsuo and Forbes, 2008

Some Remarks on PCA

- The PCA helps to extract a relevant representation of the data in a low-dimensional space and select a subset of relevant features. So, it is closely related to *dimensionality reduction*.

Some Remarks on PCA

- The PCA helps to extract a relevant representation of the data in a low-dimensional space and select a subset of relevant features. So, it is closely related to *dimensionality reduction*.
- The problem to figure out how many components contain physically relevant information is a open question.

Some Remarks on PCA

- The PCA helps to extract a relevant representation of the data in a low-dimensional space and select a subset of relevant features. So, it is closely related to *dimensionality reduction*.
- The problem to figure out how many components contain physically relevant information is a open question.
- The PCA is a linear dimensional reduction method. PCA can be extended to nonlinear by using kernel methods.

Clustering

Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set, such that those within each cluster are more closely related to one another than objects assigned to different clusters.

- Image example (2 segments or clusters)

Original Image

Segmented Image

Clustering Algorithms

- The objective is to identify data structure such as natural groups or clusters by measuring similarities between different data, i.e., find a mapping operator $\mathcal{F}$

$$\mathcal{F} : \mathbf{x} \in \mathbb{R}^N \mapsto \mathbf{C} \in \mathbb{N}$$

where classes $\mathbf{C} = \{\mathcal{C}_1 = 1, \mathcal{C}_2 = 2, \cdots\}$.

Clustering Algorithms

- The objective is to identify data structure such as natural groups or clusters by measuring similarities between different data, i.e., find a mapping operator $\mathcal{F}$

$$\mathcal{F} : \mathbf{x} \in \mathbb{R}^N \mapsto \mathbf{C} \in \mathbb{N}$$

  where classes $\mathbf{C} = \{\mathcal{C}_1 = 1, \mathcal{C}_2 = 2, \cdots\}$.

- Clustering techniques include K-means, hierarchical, Gaussian mixture models, hidden Markov models.

Clustering Algorithms

- The objective is to identify data structure such as natural groups or clusters by measuring similarities between different data, i.e., find a mapping operator $\mathcal{F}$

  $$\mathcal{F} : \mathbf{x} \in \mathbb{R}^N \mapsto \mathbf{C} \in \mathbb{N}$$

  where classes $\mathbf{C} = \{\mathcal{C}_1 = 1, \mathcal{C}_2 = 2, \cdots\}$.

- Clustering techniques include K-means, hierarchical, Gaussian mixture models, hidden Markov models.

- Clustering algorithms are used for *unsupervised classification*.

K-means clustering

- The objective is to assign each observation, uniquely labeled by an integer $n \in \{1, \cdots, N\}$, to one and one only cluster $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$.
  - The total number of clusters is fixed $(K < N)$.
  - The number of data points in the k-th cluster is N_k

K-means Clustering

- The most common choice for $W(\mathcal{C}_k)$ involves squared Euclidean distance, and for D-dimensional space

$$d(\mathbf{x}_n, \mathbf{x}_{n'}) = \|\mathbf{x}_n - \mathbf{x}_{n'}\|^2 = \sum_{j=1}^{D} (x_{nj} - x_{n'j})^2$$

K-means Clustering

- The most common choice for $W(\mathcal{C}_k)$ involves squared Euclidean distance, and for D-dimensional space

$$d(\mathbf{x}_n, \mathbf{x}_{n'}) = \|\mathbf{x}_n - \mathbf{x}_{n'}\|^2 = \sum_{j=1}^{D} (x_{nj} - x_{n'j})^2$$

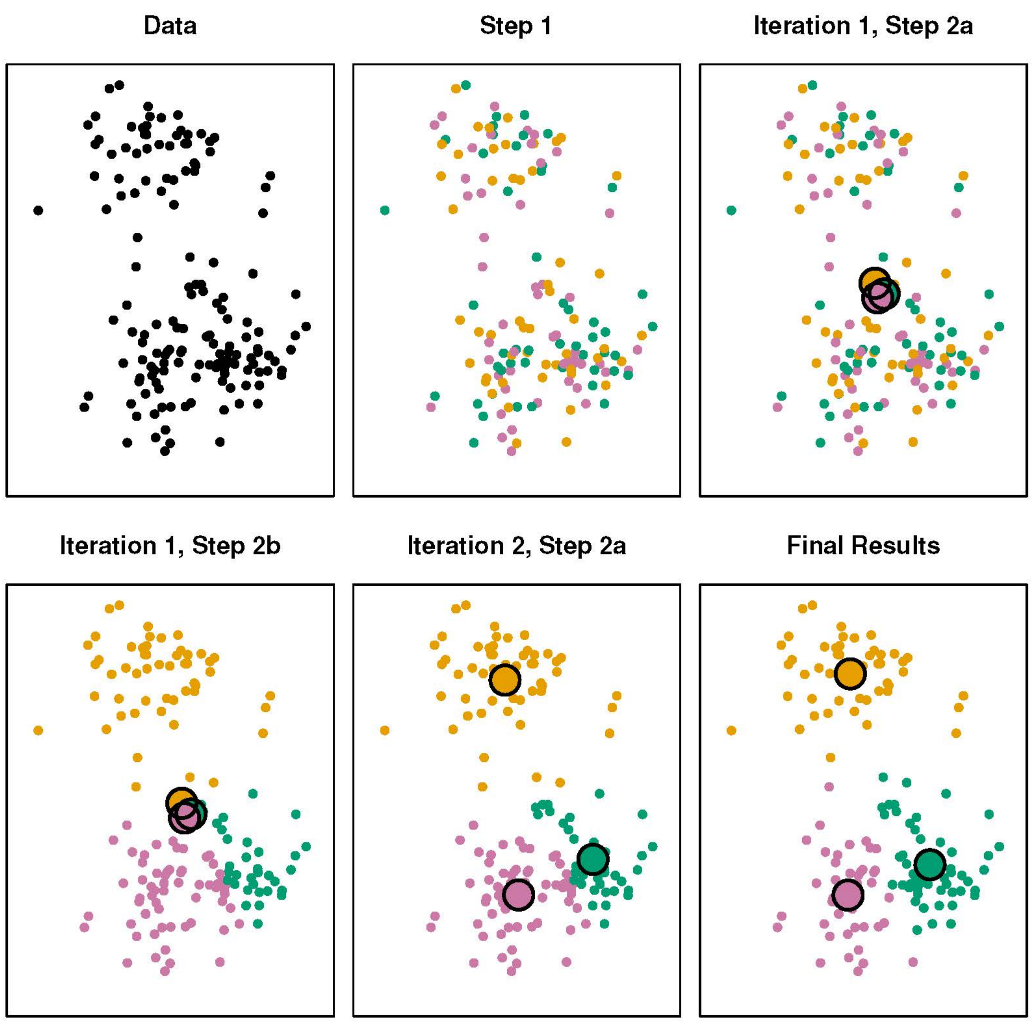
- The within-cluster variation $W(\mathcal{C}_k)$ is

$$\begin{aligned} W(\mathcal{C}_k) &= \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \sum_{n' \in \mathcal{C}_k} d(x_n, x_{n'}) \\ &= 2 \sum_{n \in \mathcal{C}_k} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \end{aligned}$$

where $\overline{x_{kj}} = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} x_{nj}$ and $\boldsymbol{\mu}_k = \{\overline{x_{k1}}, \cdots, \overline{x_{kj}}\}$ (kth cluster centroid)

K-means Clustering Algorithm Steps

1. Randomly assign a number, from 1 to K, to each data.
2. Iterate until the cluster assignments stop changing:
   - For each of the K clusters, compute the cluster centroid $\boldsymbol{\mu}_k$
   - Assign each data to the cluster whose centroid is closest.



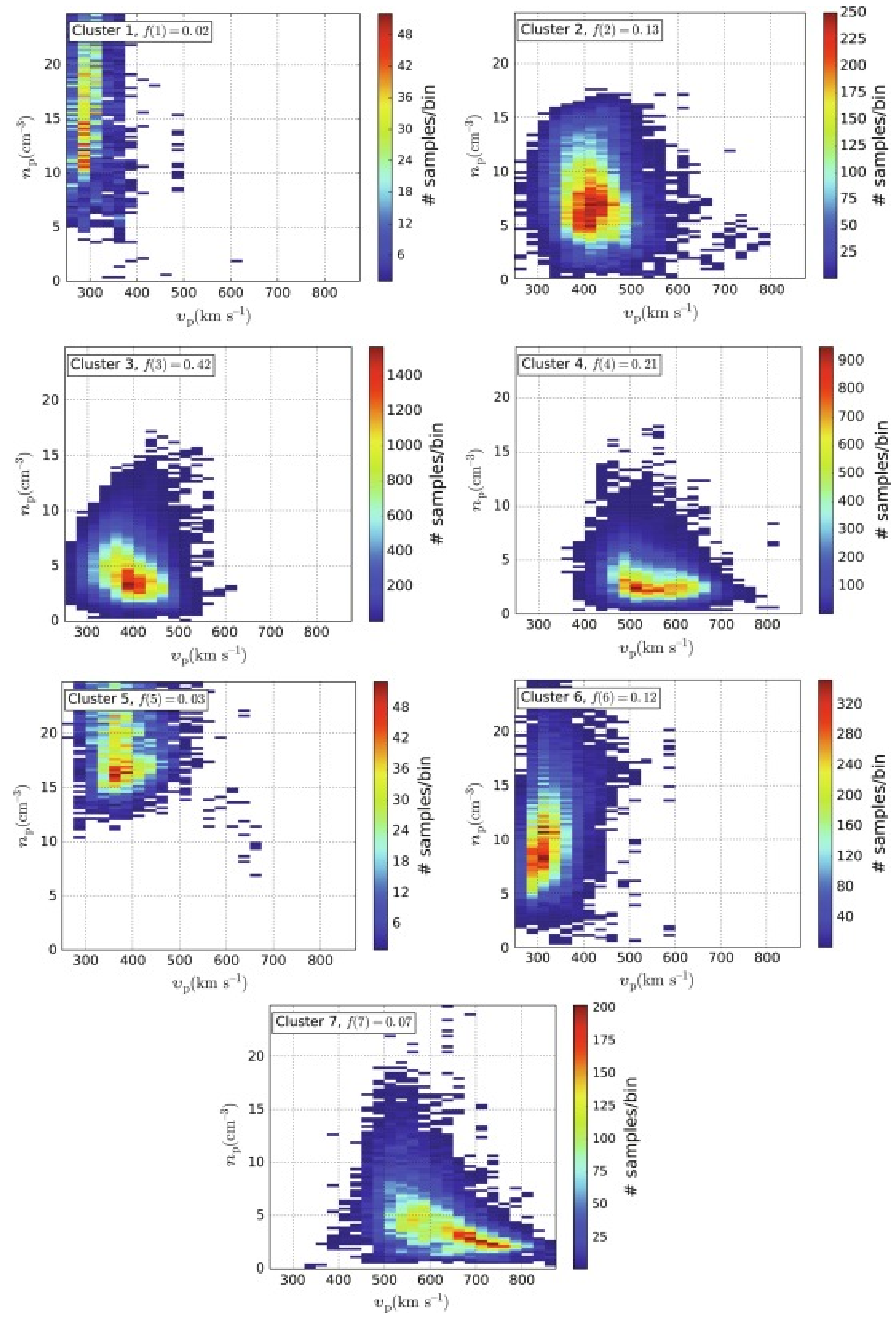
James, 2013

K-means Clustering Example

K = 19 clustering of RGB image of aurora (D = 3)

K-means Clustering Example - Solar Wind

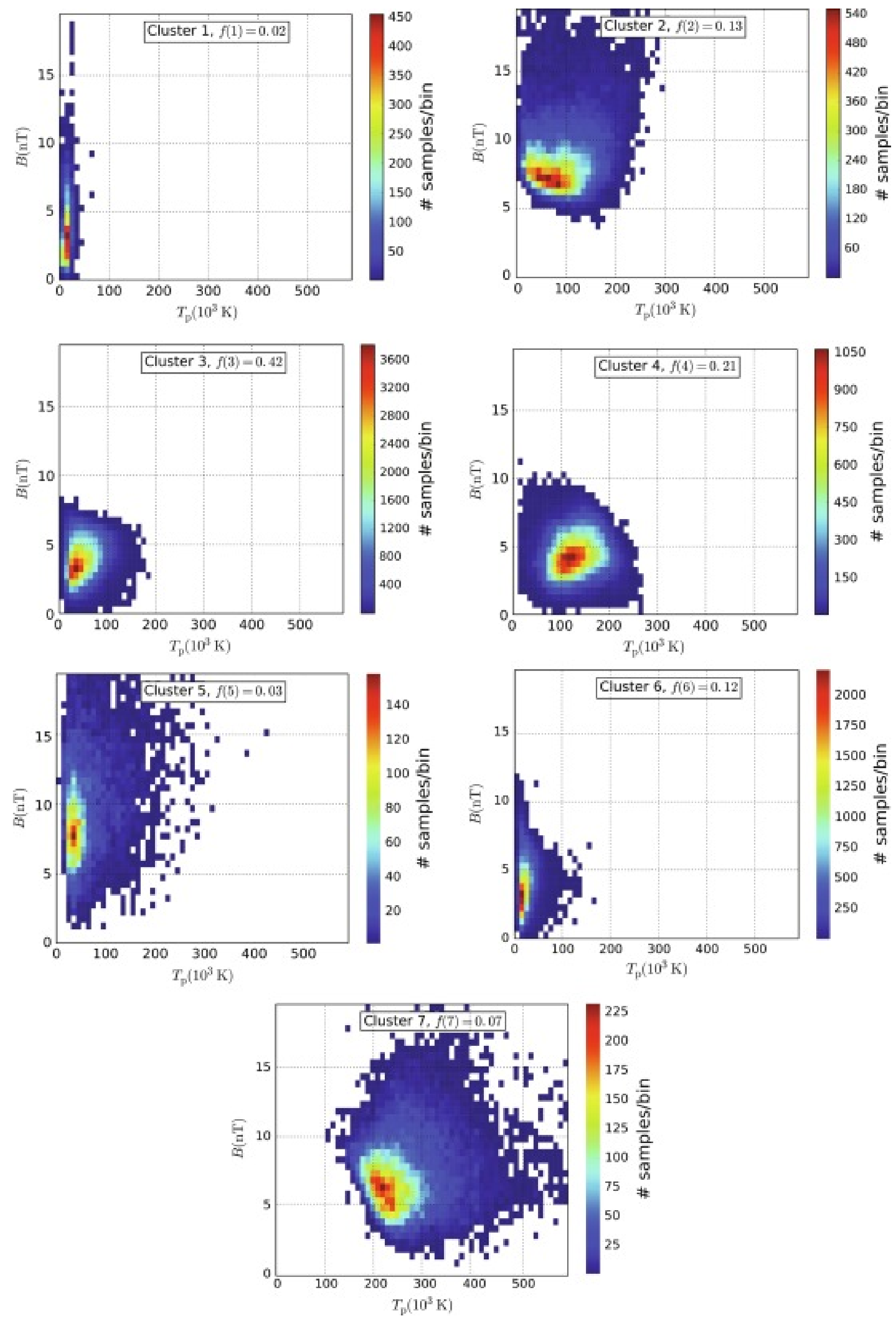
K = 7 clustering of 10-year solar wind data (D = 7)



Heidrich-Meisner and Wimmer-Schweingruber, 2018

K-means Clustering Example - Solar Wind

K = 7 clustering of 10-year solar wind data (D = 7)



Heidrich-Meisner and Wimmer-Schweingruber, 2018

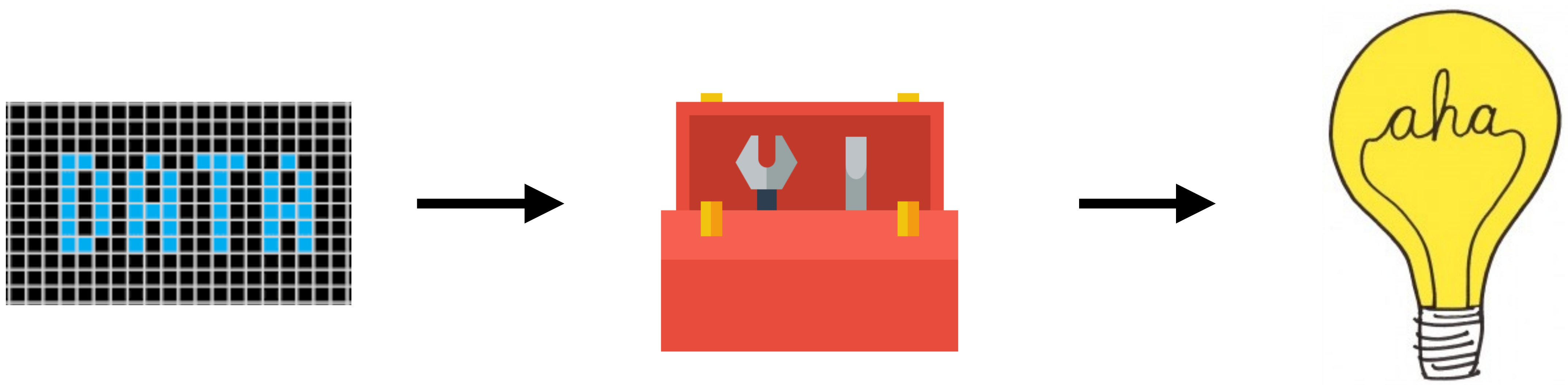
Some Remarks on K-means Clustering

- The within-cluster variation decreases with each iteration of the algorithm.
- The K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial (random) cluster assignment.

Some Remarks on K-means Clustering

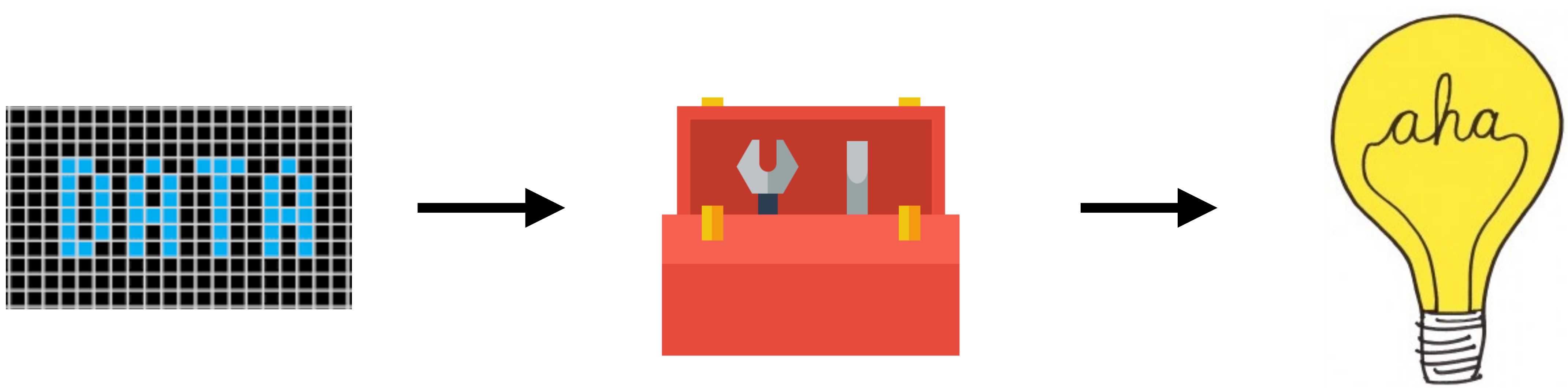
- The within-cluster variation decreases with each iteration of the algorithm.
- The K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial (random) cluster assignment.
- The problem of selecting K is far from simple.

What do These Tools do for us?



What do These Tools do for us?

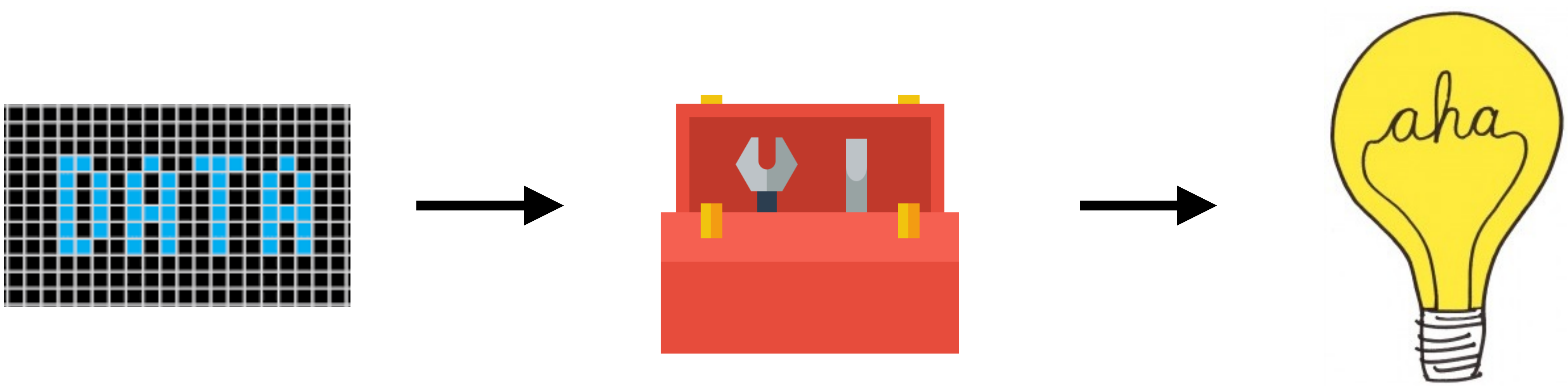
Statistical Learning Techniques help *you*



What do These Tools do for us?

Statistical Learning Techniques help *you*

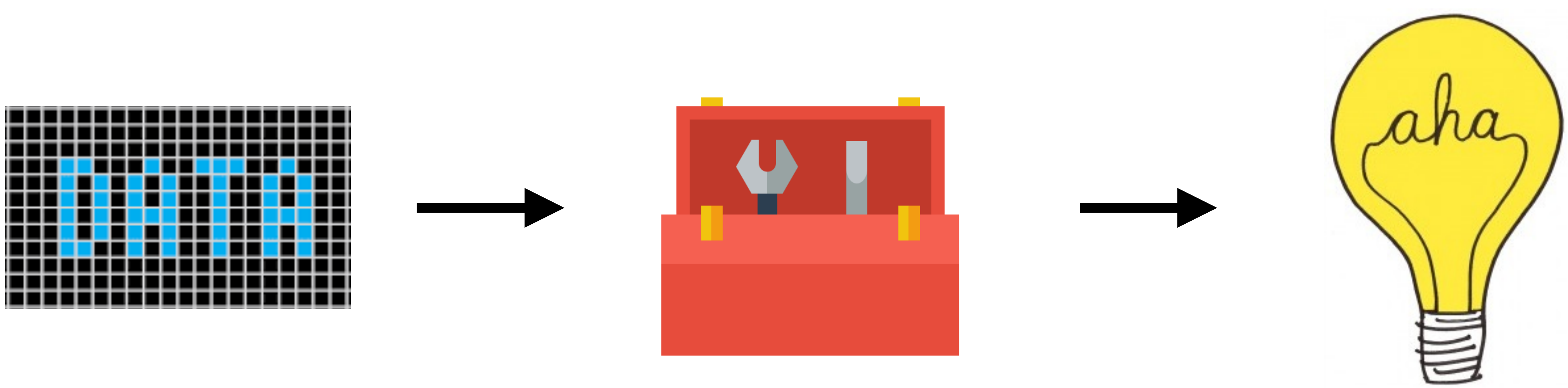
- Manage large volumes of data by dimensionality reduction



What do These Tools do for us?

Statistical Learning Techniques help *you*

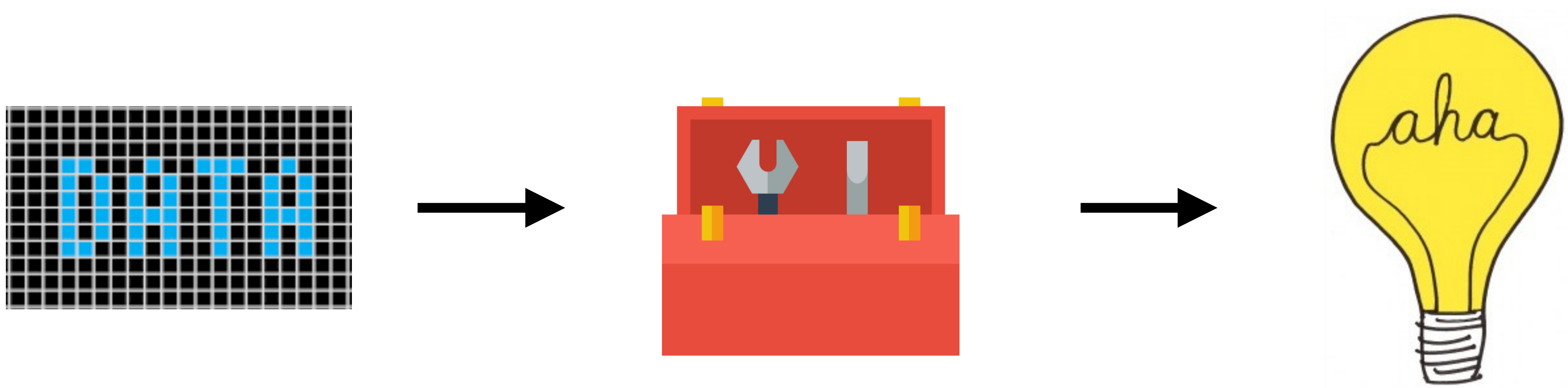
- Manage large volumes of data by dimensionality reduction
- Extract characteristic patterns and trends from data



What do These Tools do for us?

Statistical Learning Techniques help *you*

- Manage large volumes of data by dimensionality reduction
- Extract characteristic patterns and trends from data
- Identify the way data are naturally grouped together



What do These Tools do for us?

Statistical Learning Techniques help *you*

- Manage large volumes of data by dimensionality reduction
- Extract characteristic patterns and trends from data
- Identify the way data are naturally grouped together
- Gain scientific insights