Big Data Course

Chap 2-
Dimension Reduction

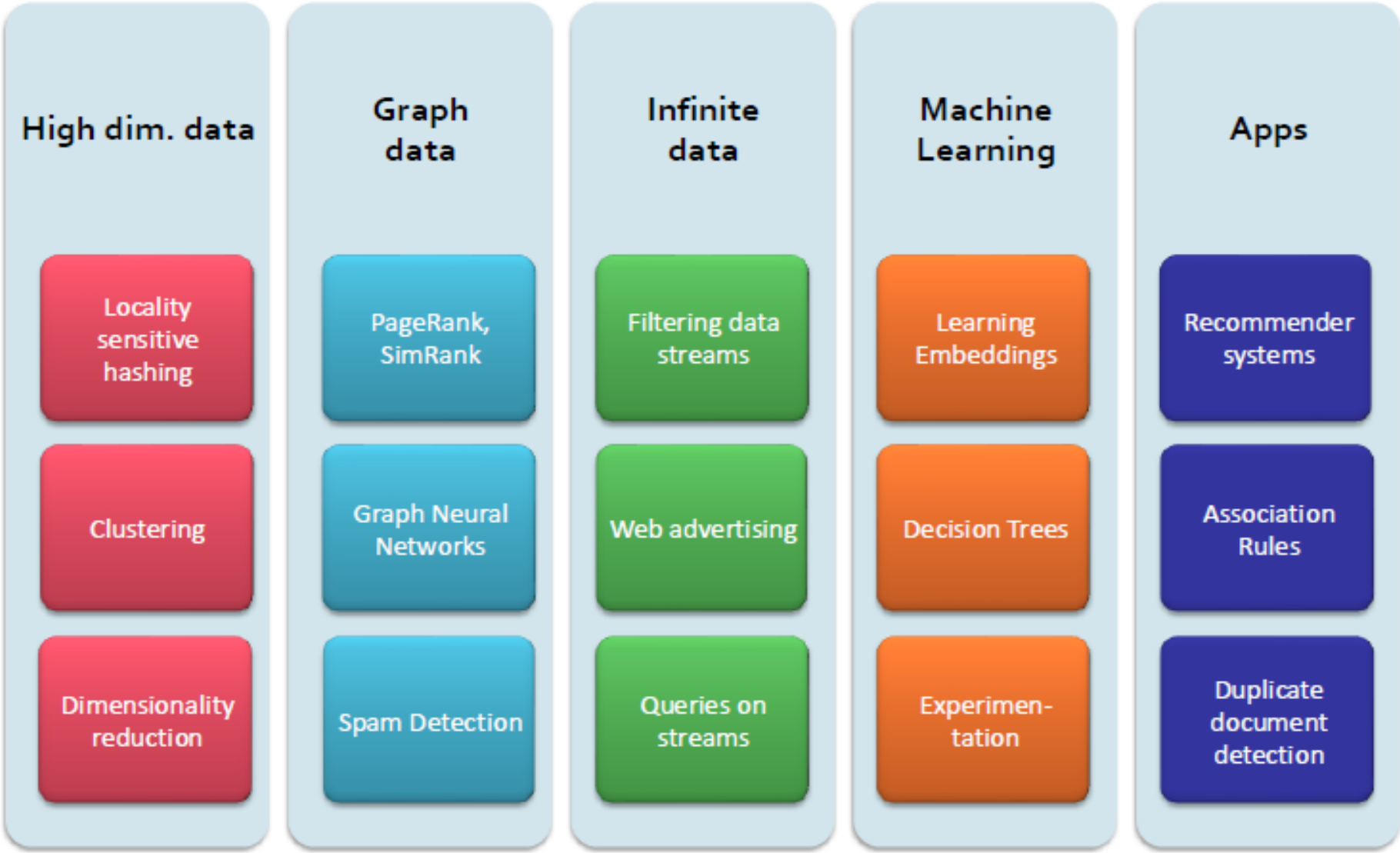
Electrical Engineering department of

Amirkabir University of technology

Dr. Mohammadreza Pourfard

August 2025

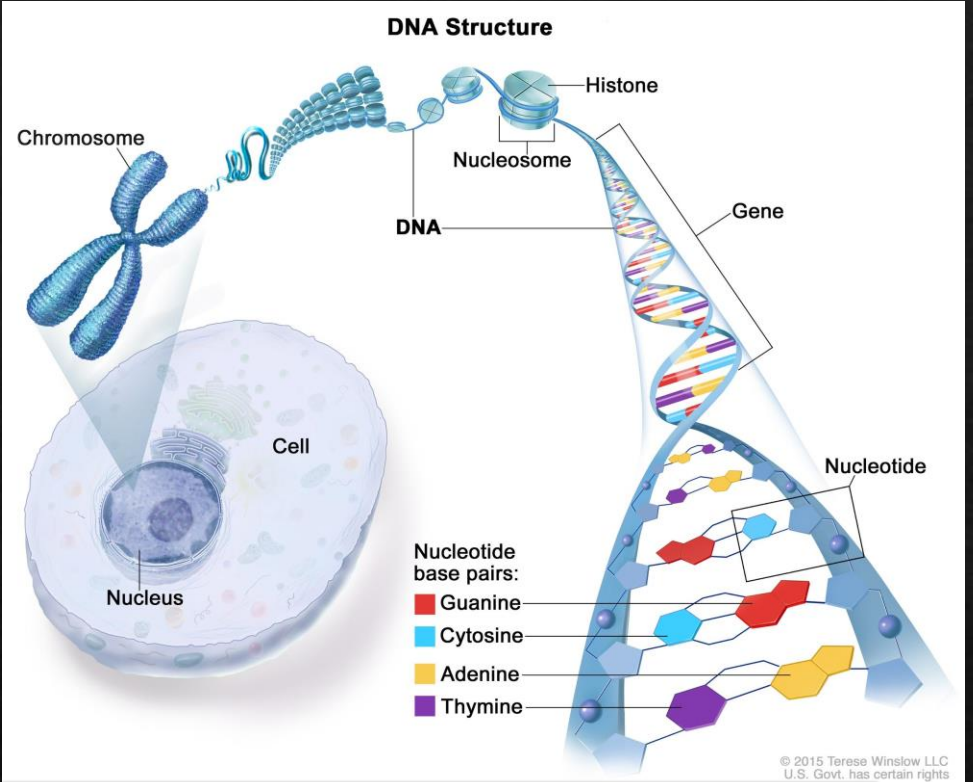
How the Class Fits Together

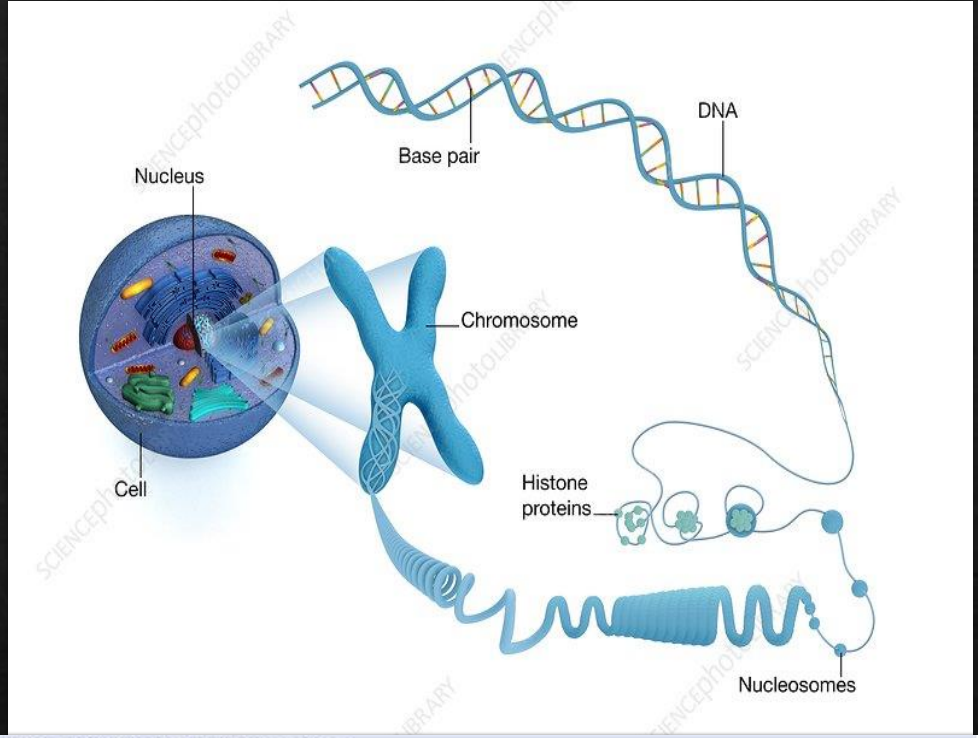


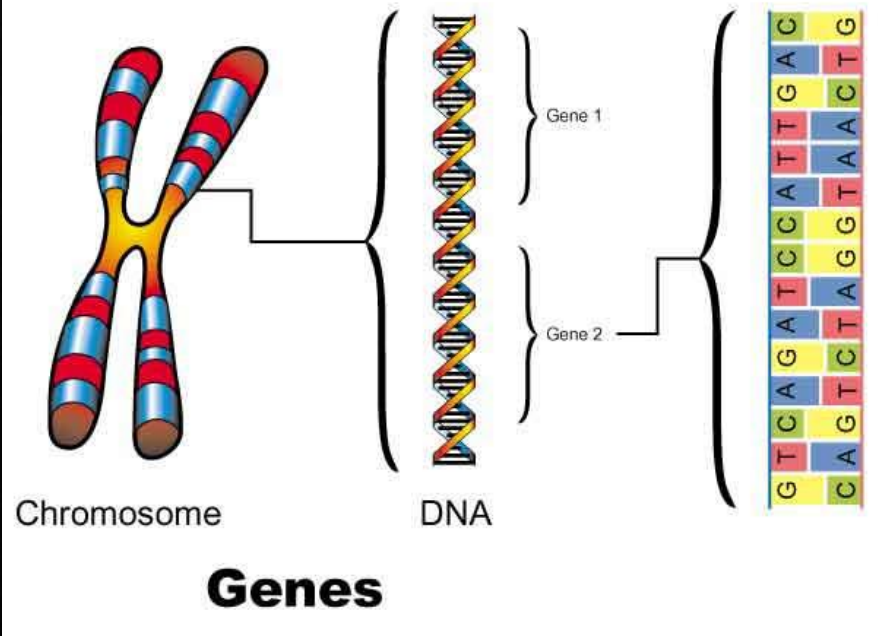
حجم بالای داده‌ها در حوزه پزشکی

و لزوم کاهش بعد

DNA


DNA Structure
Chromosome
Histone
Nucleosome
DNA
Gene
Cell
Nucleotide
Nucleotide base pairs:
Guanine
Cytosine
Adenine
Thymine
Nucleus
© 2015 Terese Winslow LLC
U.S. Govt. has certain rights


DNA
Base pair
Nucleus
Chromosome
Cell
Histone proteins
Nucleosomes


Gene 1
Gene 2
Chromosome
DNA
Genes

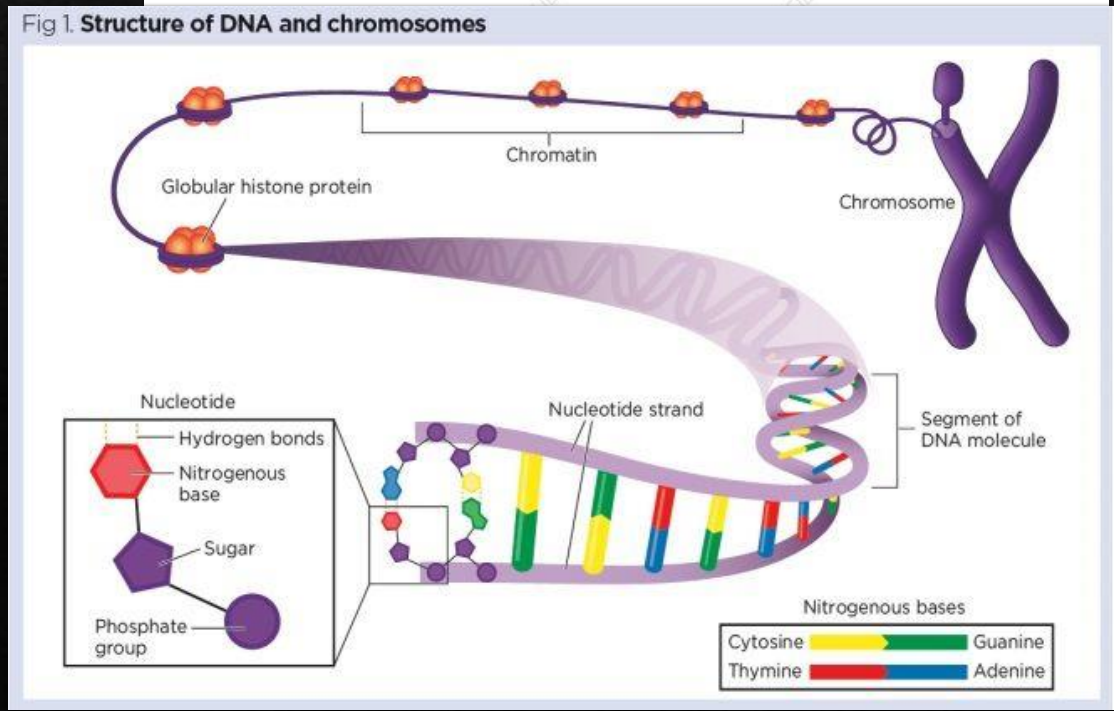
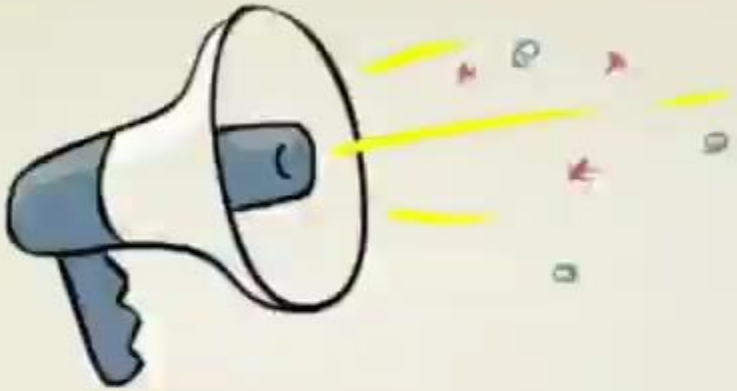
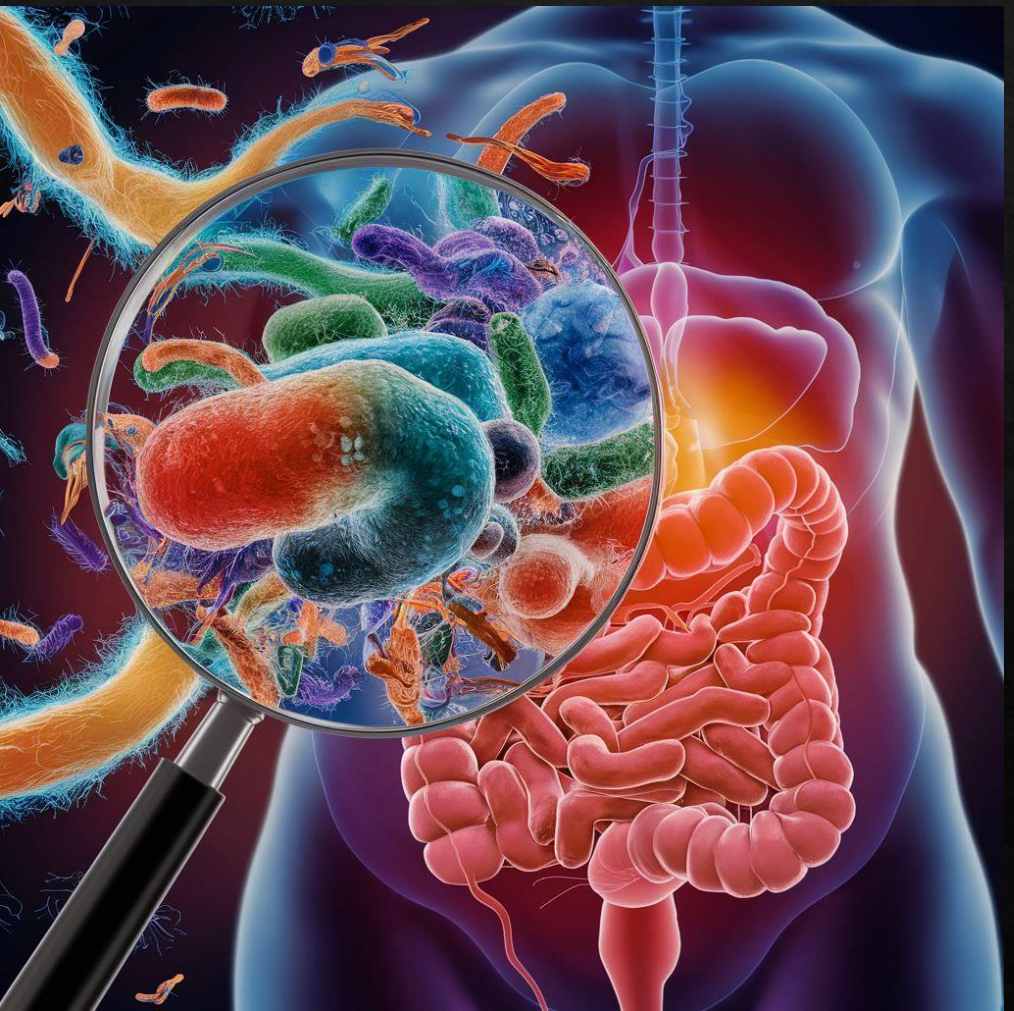


Fig 1. Structure of DNA and chromosomes
Chromatin
Globular histone protein
Chromosome
Nucleotide
Hydrogen bonds
Nitrogenous base
Sugar
Phosphate group
Nucleotide strand
Segment of DNA molecule
Nitrogenous bases
Cytosine
Guanine
Thymine
Adenine

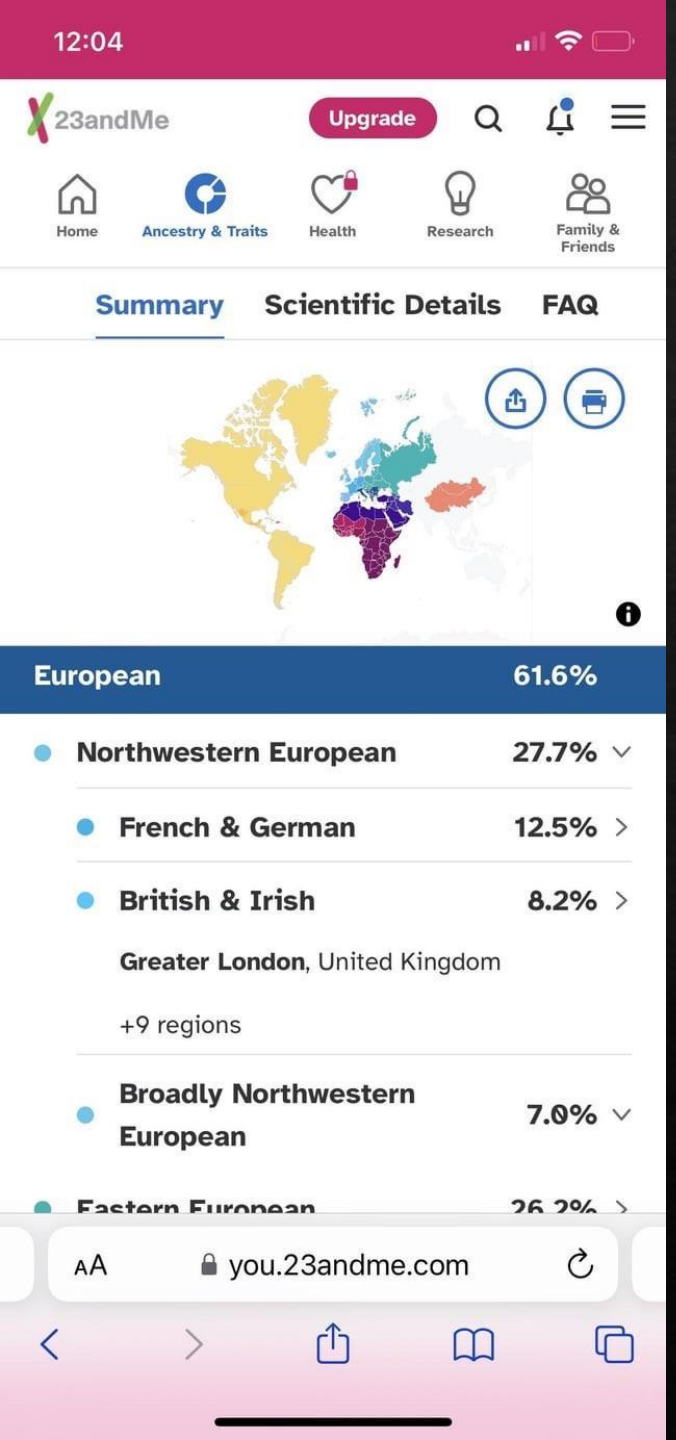


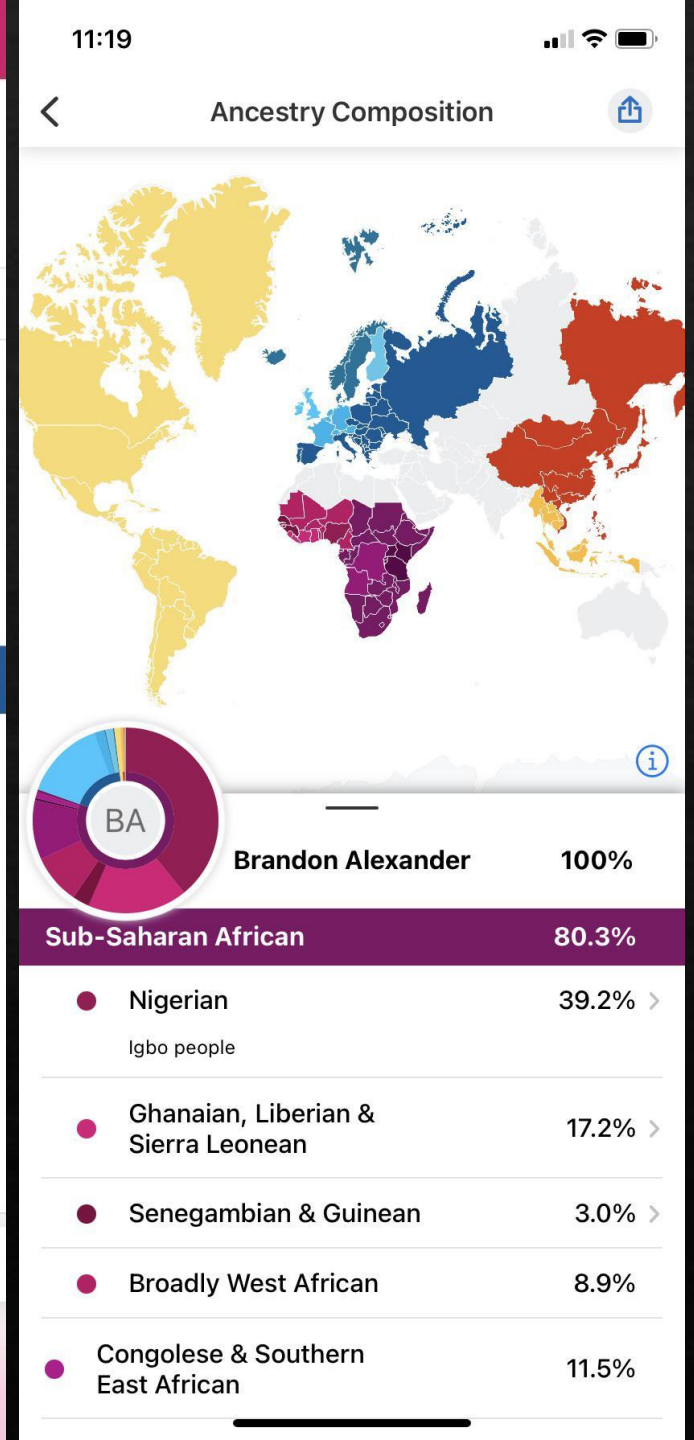
Stated Clearly

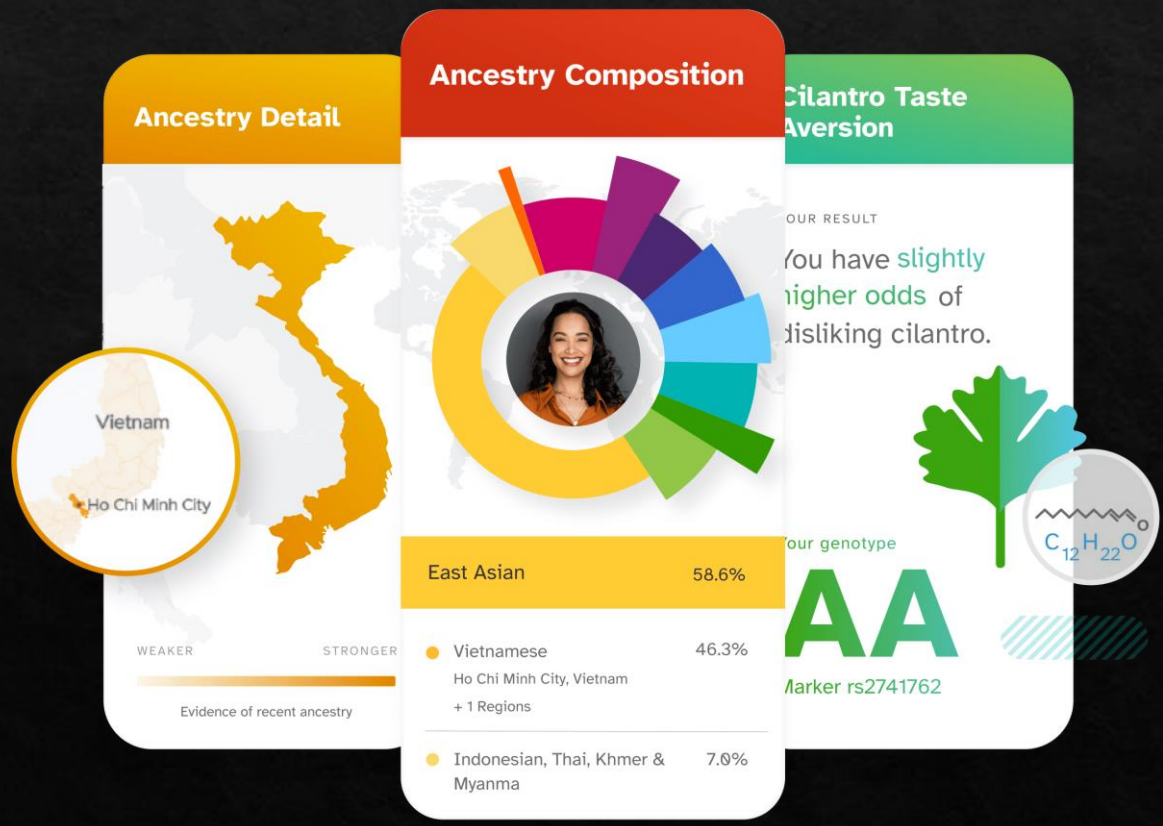
حوزه Gut Microbiota
بر اساس داده های ژنتیک



تحلیل جمعیت اجدادی با داده ژنتیک

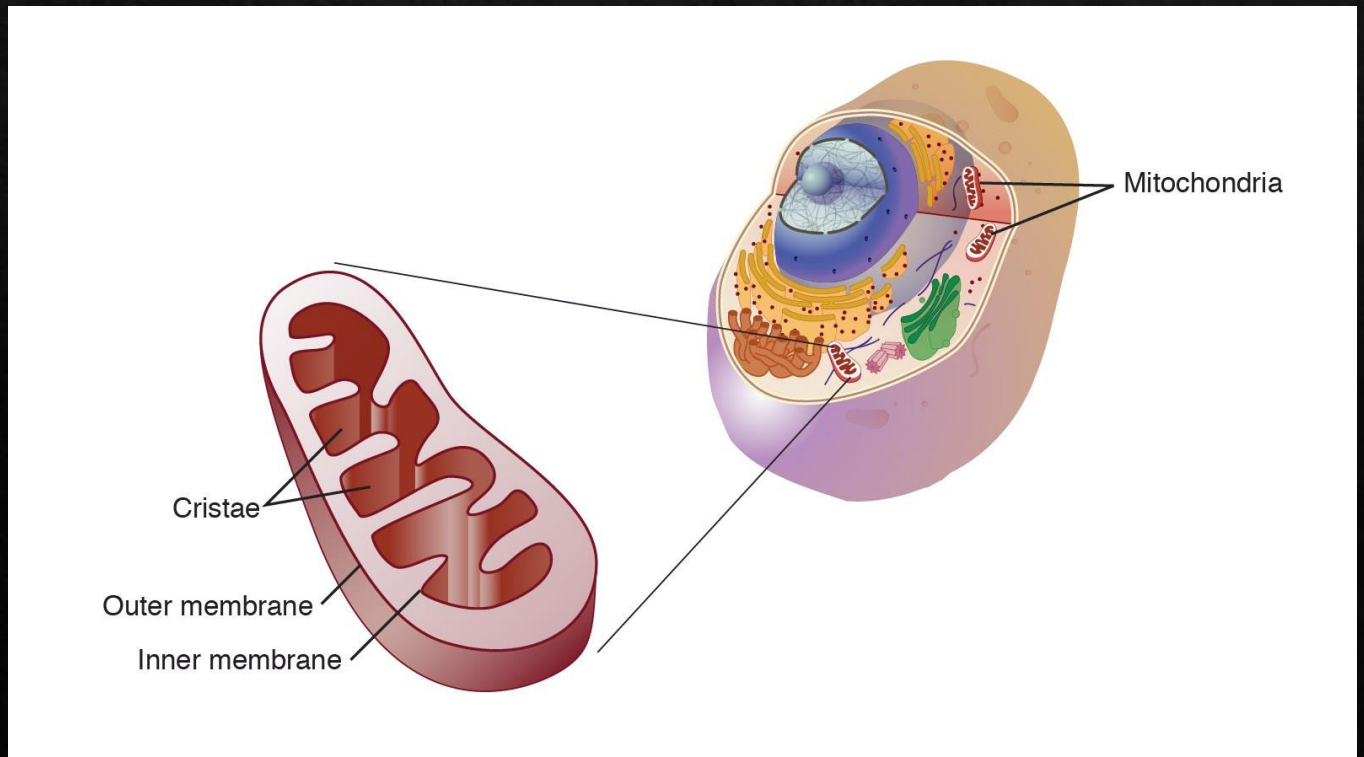


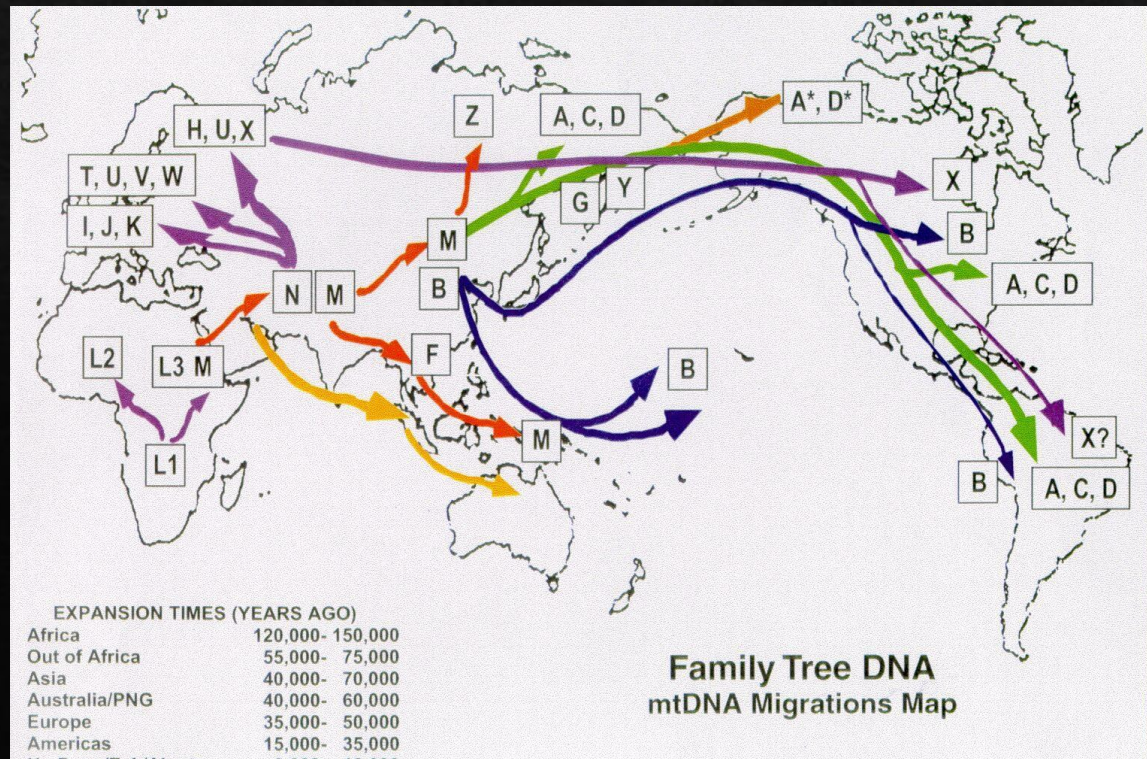


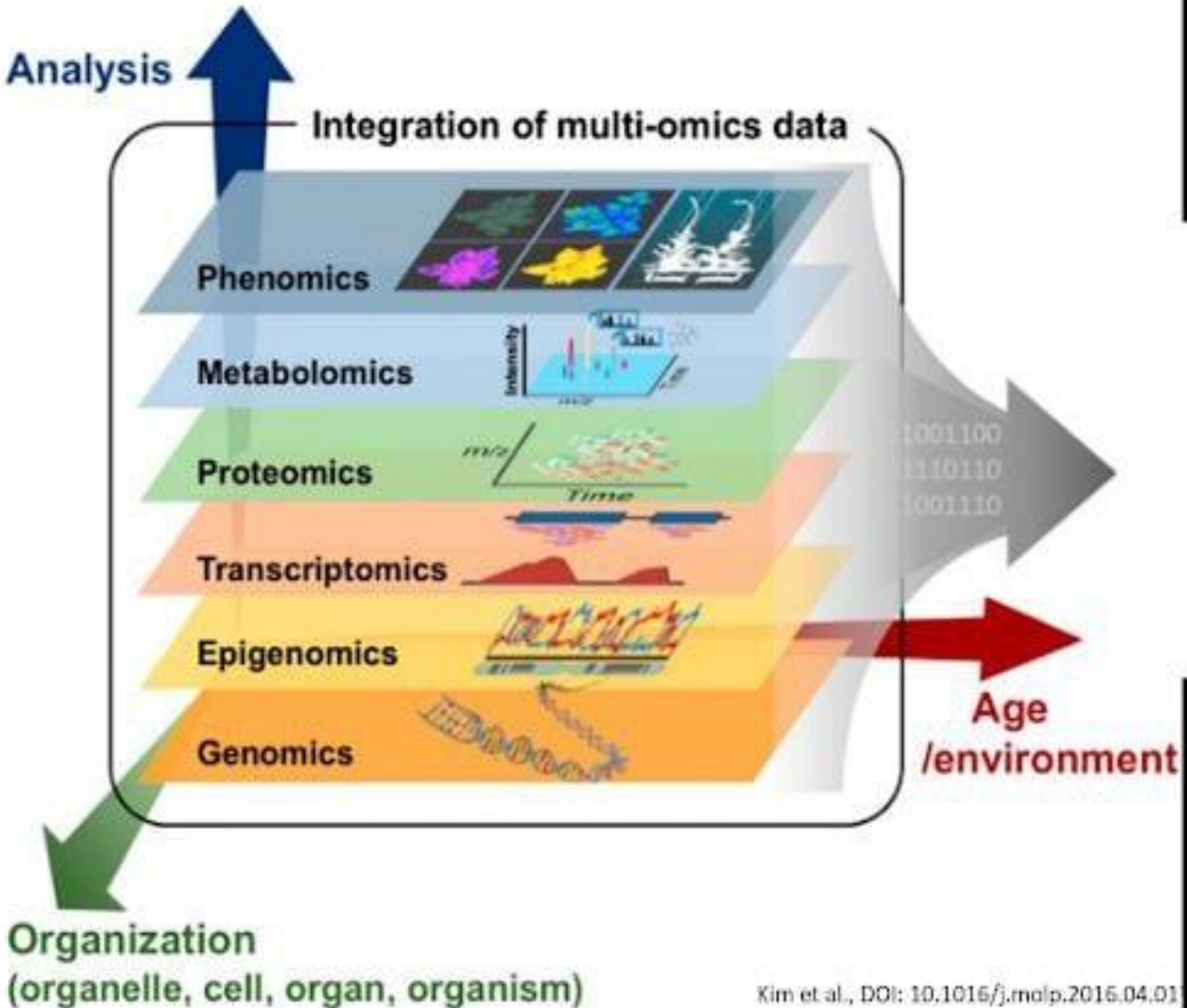


تشخیص مسیر مهاجرت مادری

- تعیین مسیر مهاجرت اجداد به کمک داده ژنتیک میتوکندری





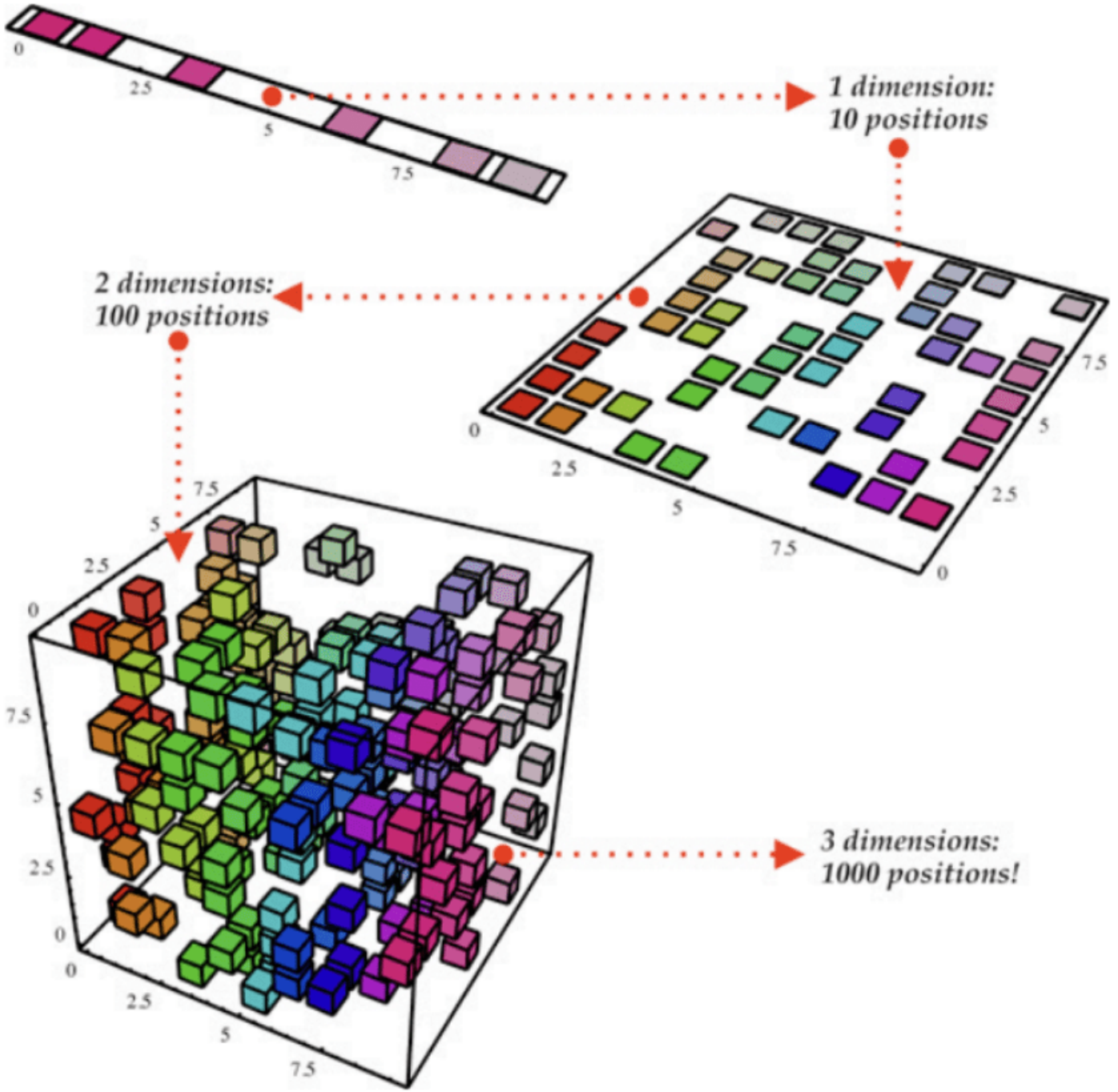


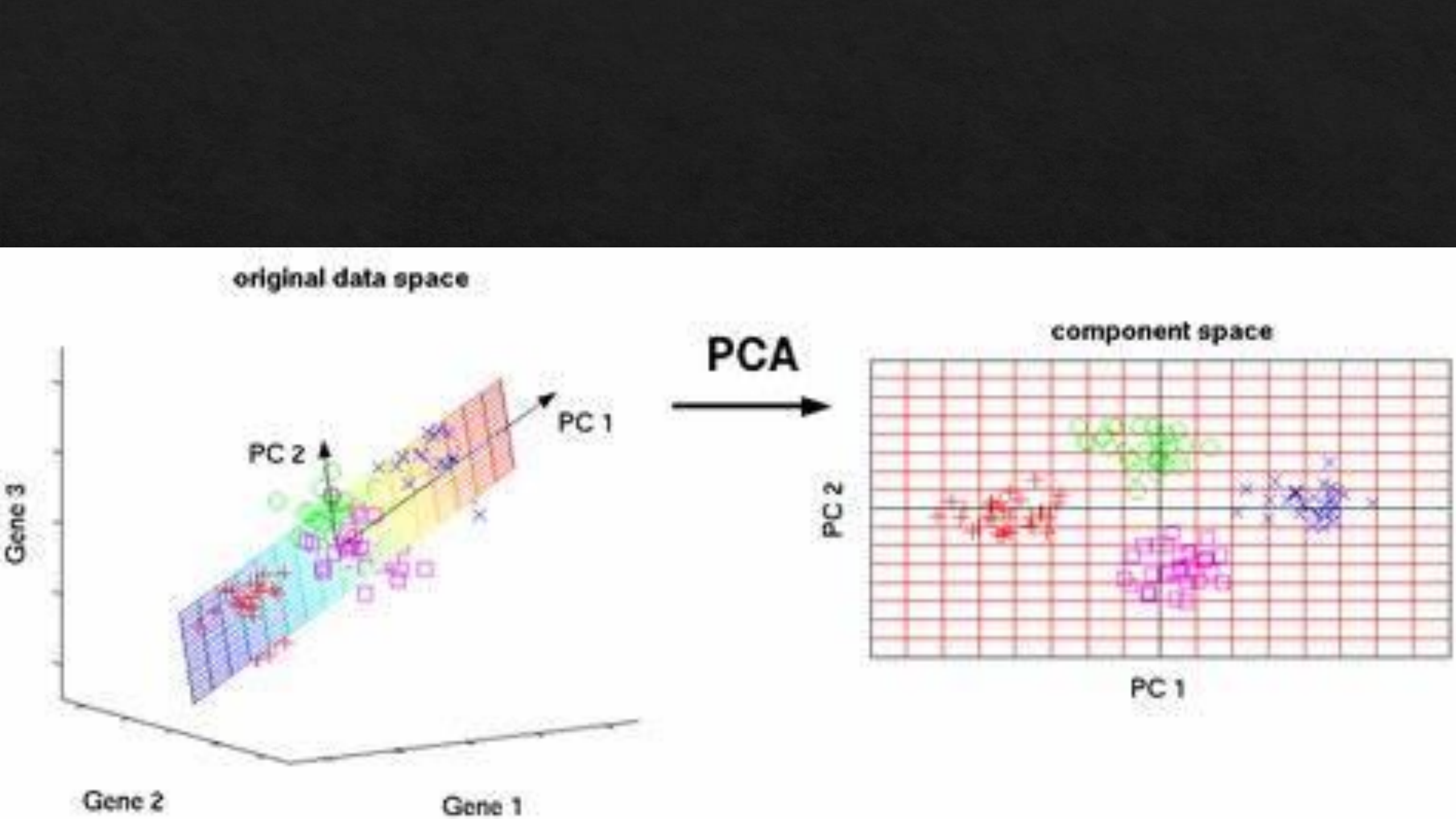
سطوح تشخیص بیماری (شروع از پیچیده ترین سطح)

Kim et al., DOI: 10.1016/j.molp.2016.04.01

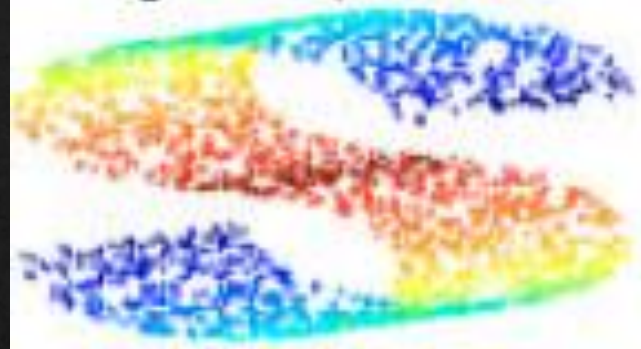
ضرورت کاہش بعد

Dimensionreduction

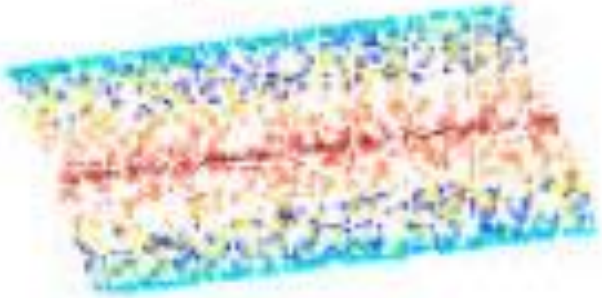

1 dimension:
10 positions
2 dimensions:
100 positions
3 dimensions:
1000 positions!


original data space
PC 1
PC 2
Gene 3
Gene 2
Gene 1
PCA
component space
PC 2
PC 1

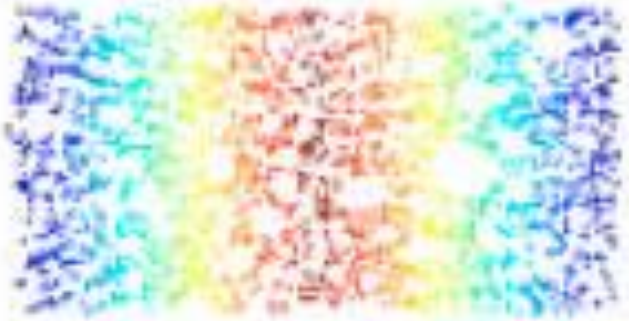
Original data, N = 2000



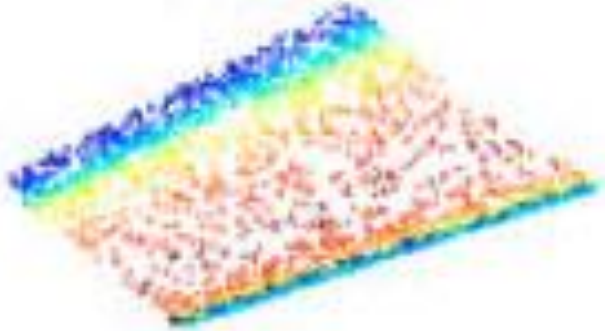
MDS, corr = 0.7911



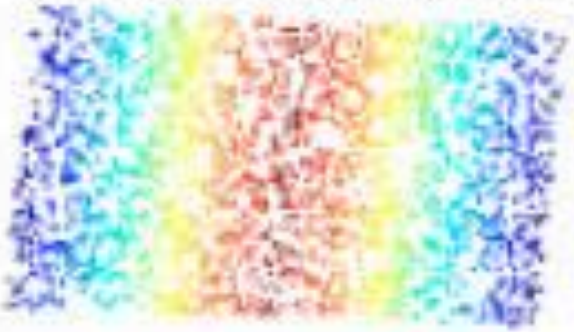
Isomap, k = 12, corr = 0.9995



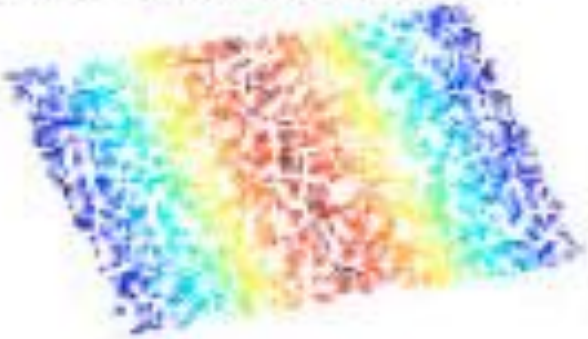
LLE, k = 12, corr = 0.8261



HLLE, k = 12, corr 0.9069

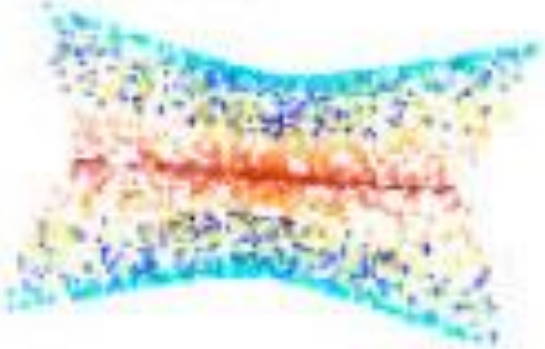


LTSA, k = 12, corr = 0.9068



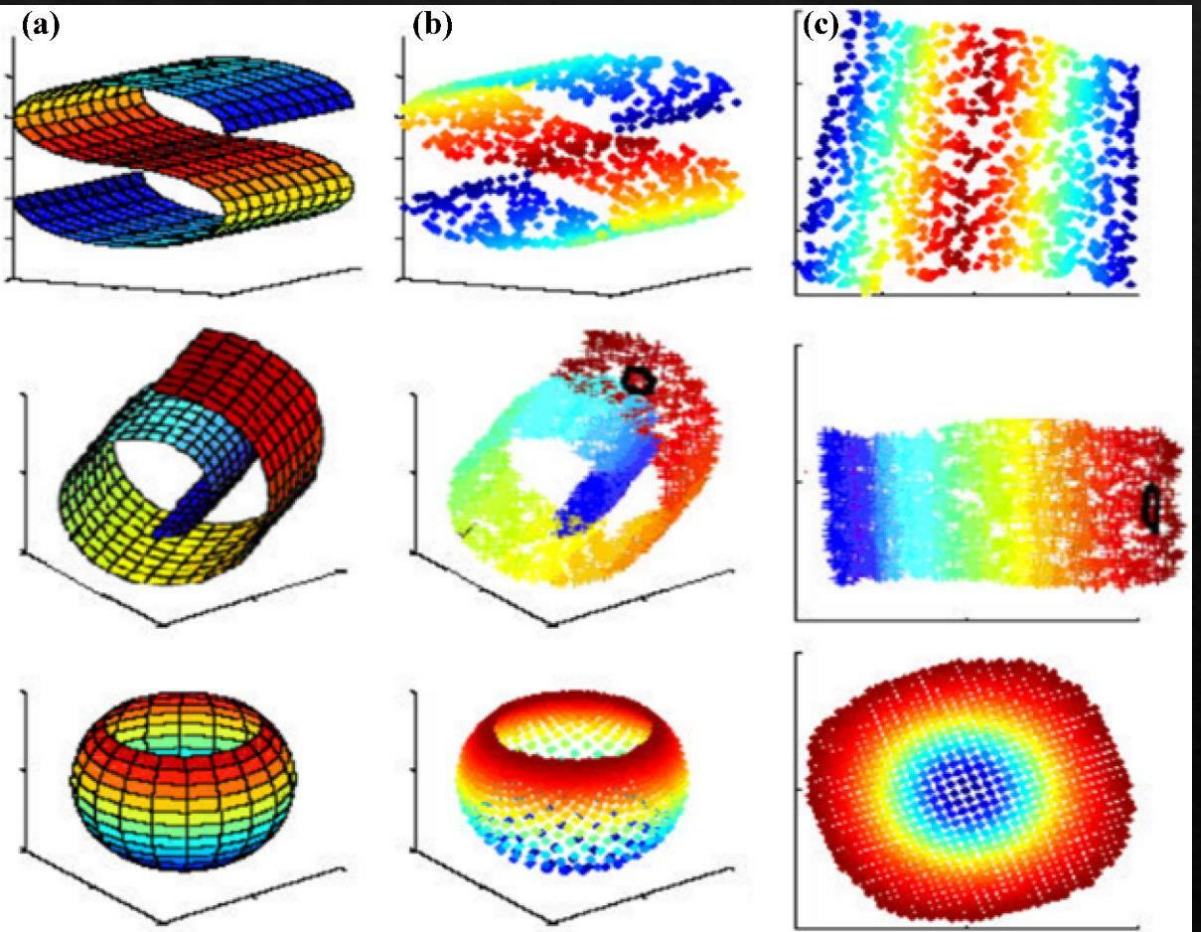
Kernal PCA, poly, corr = 0.4487

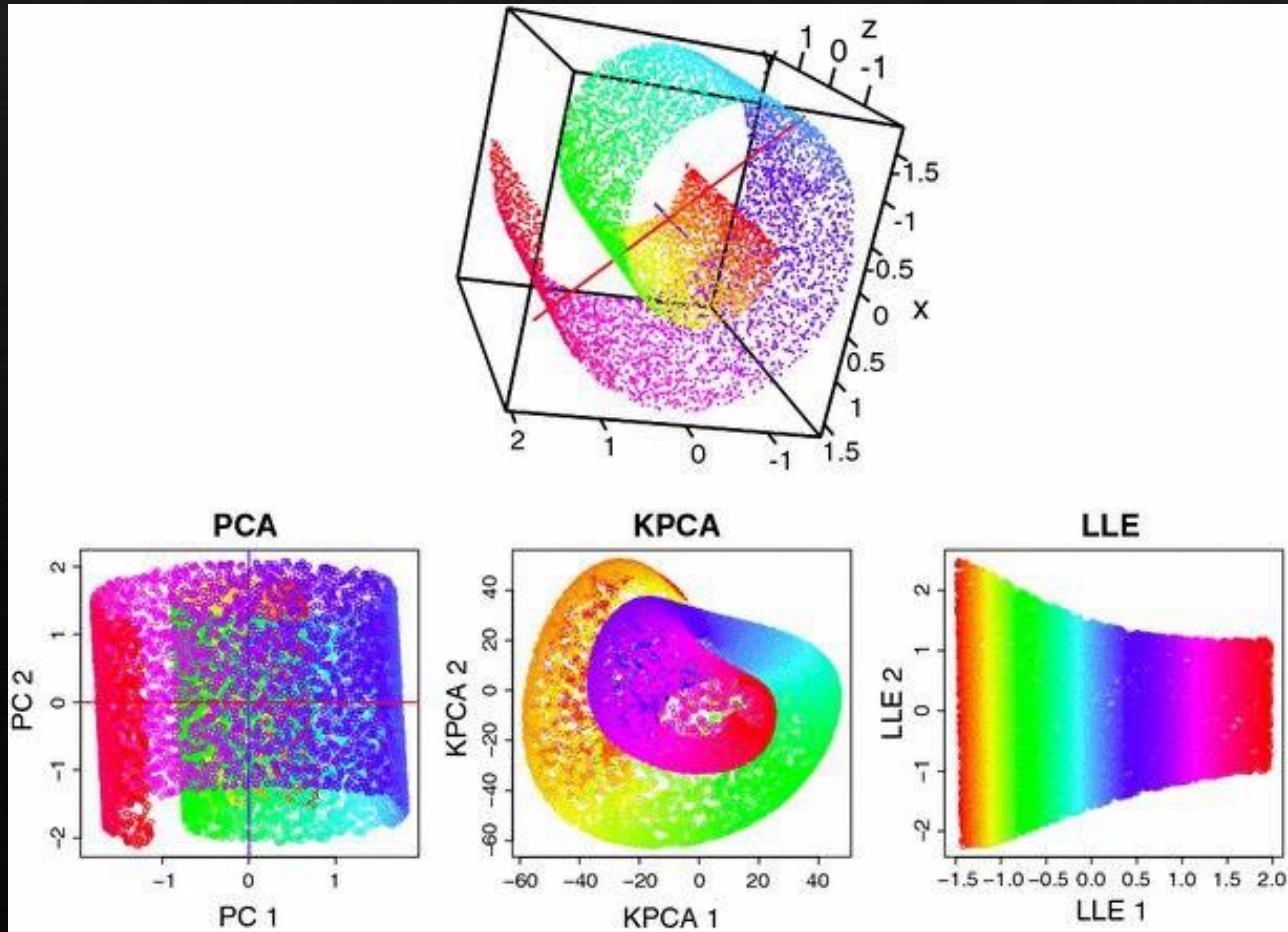
Diffusion maps, corr = 0.6995

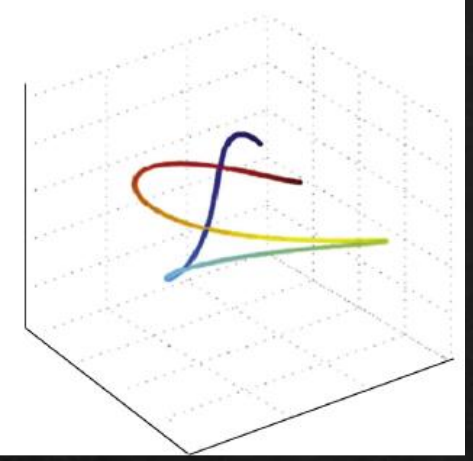


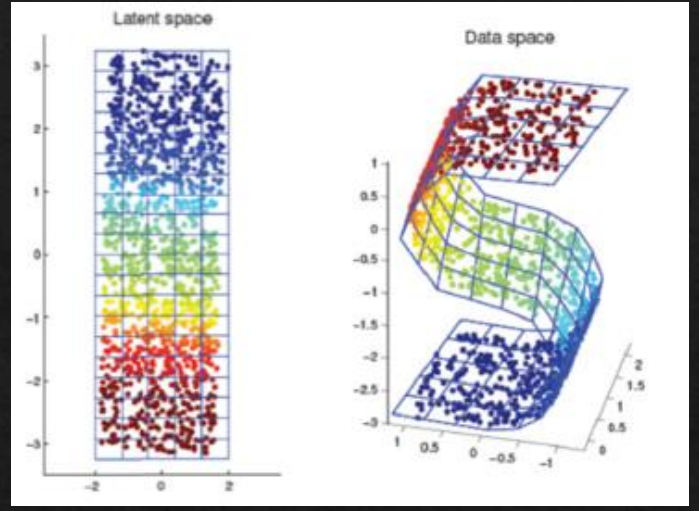
Auto encoder RBM, corr = 0.6441

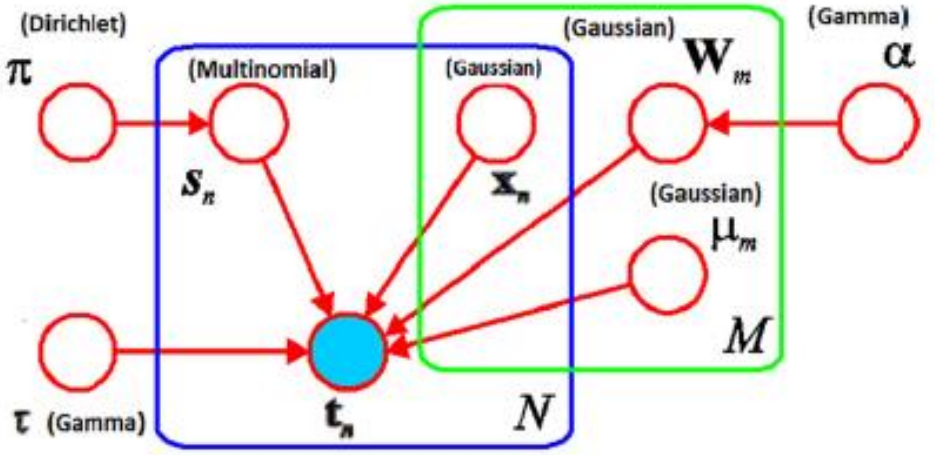
Advanced Dimension reduction techniques to Intrinsic dimension



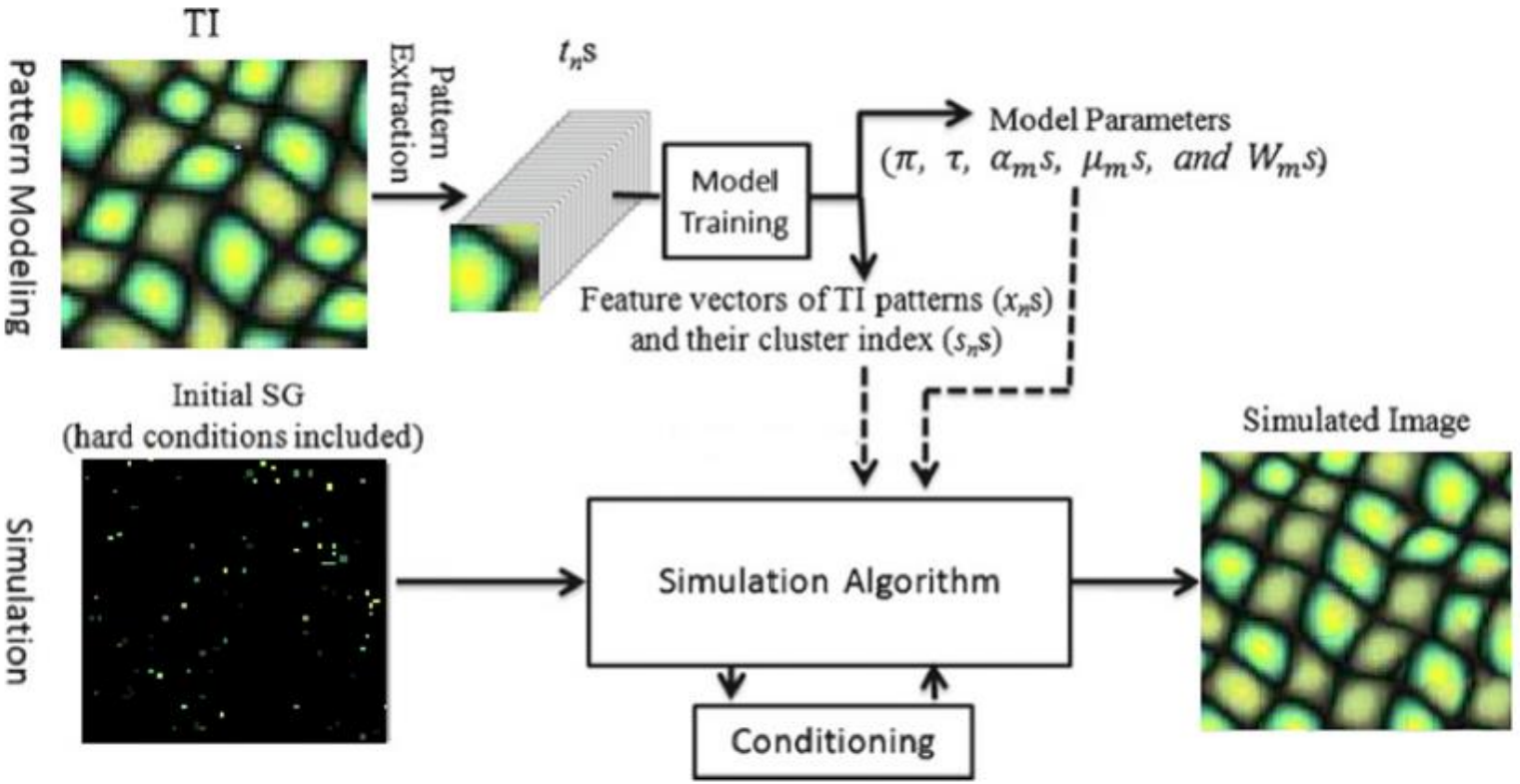


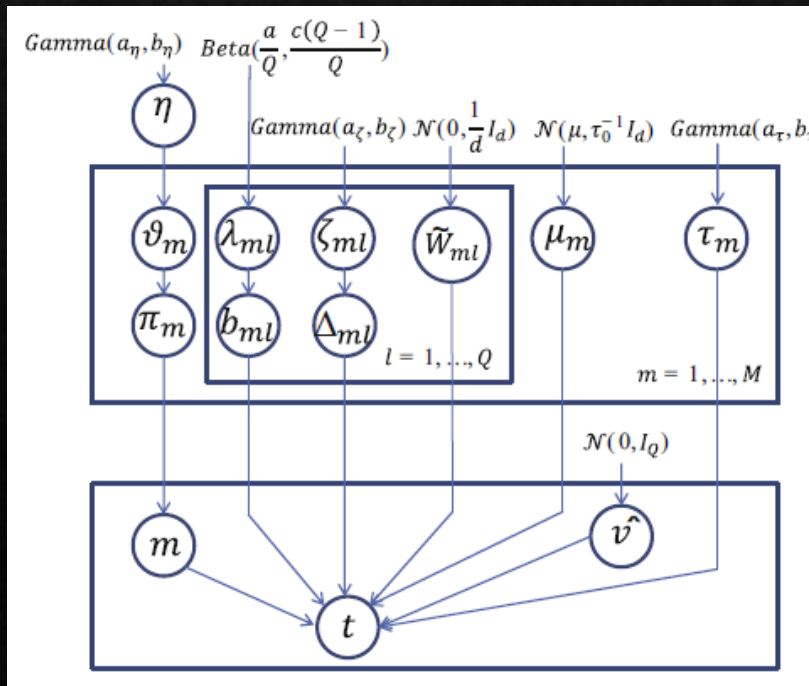


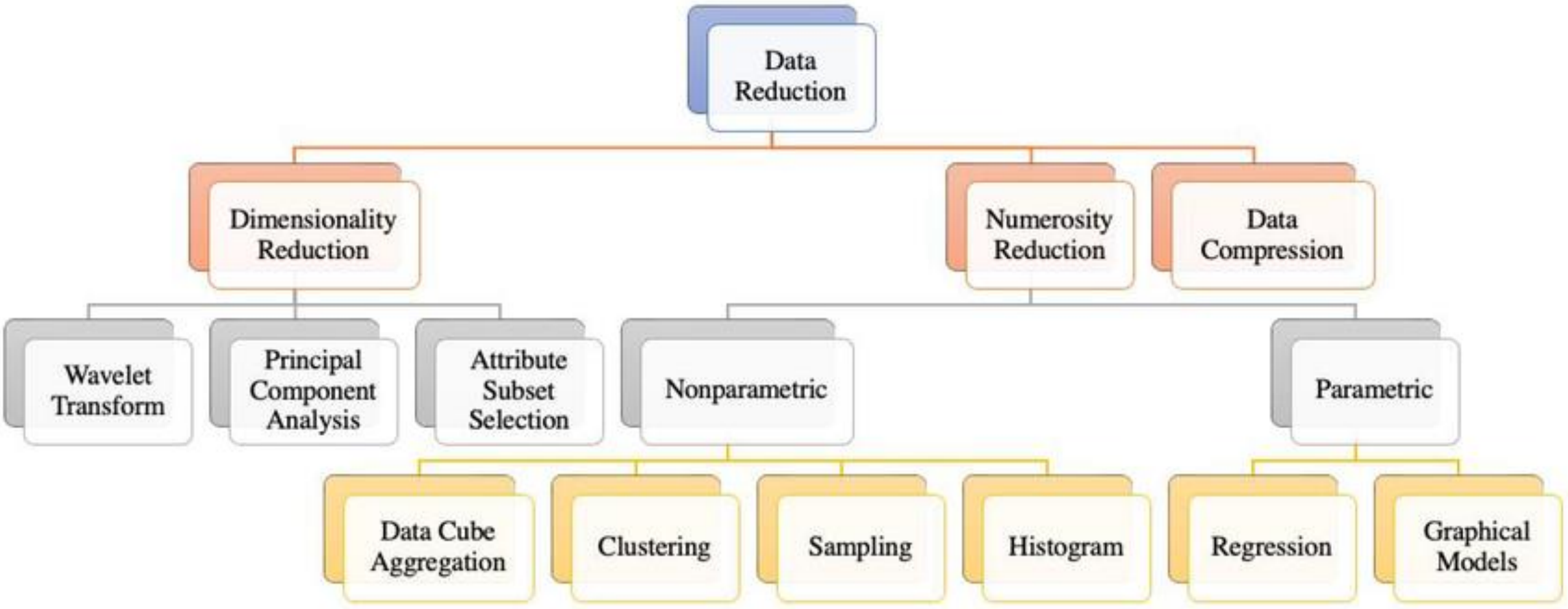


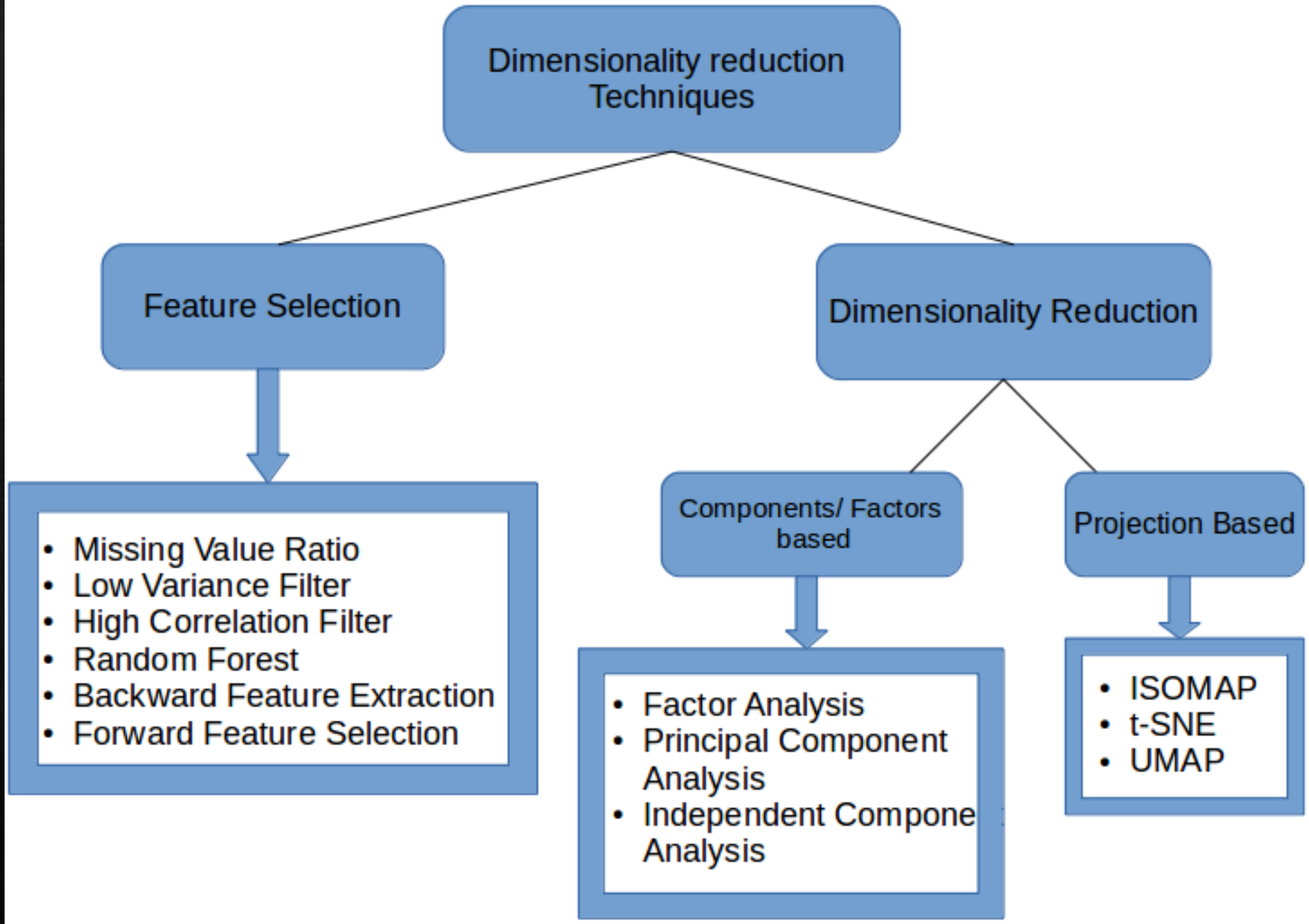


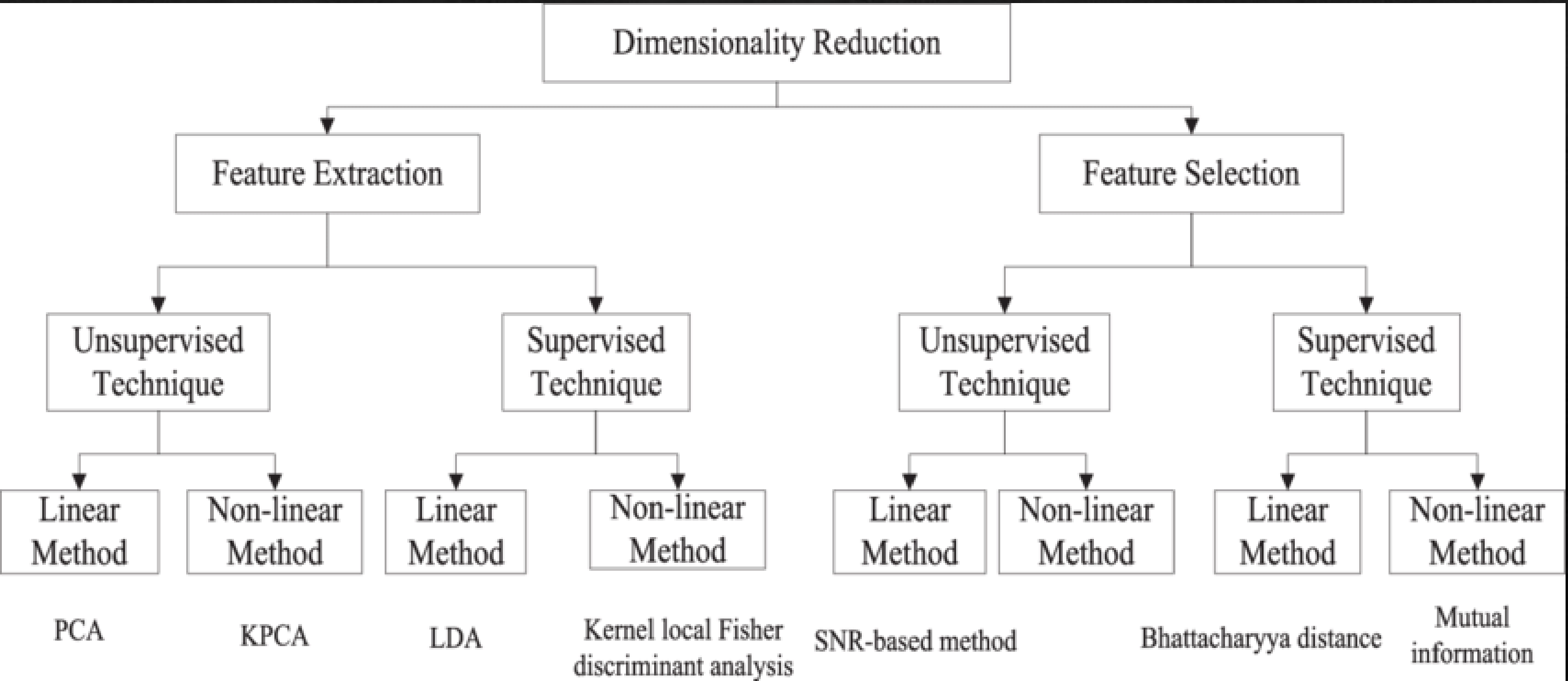
Bayesian MPPCA (





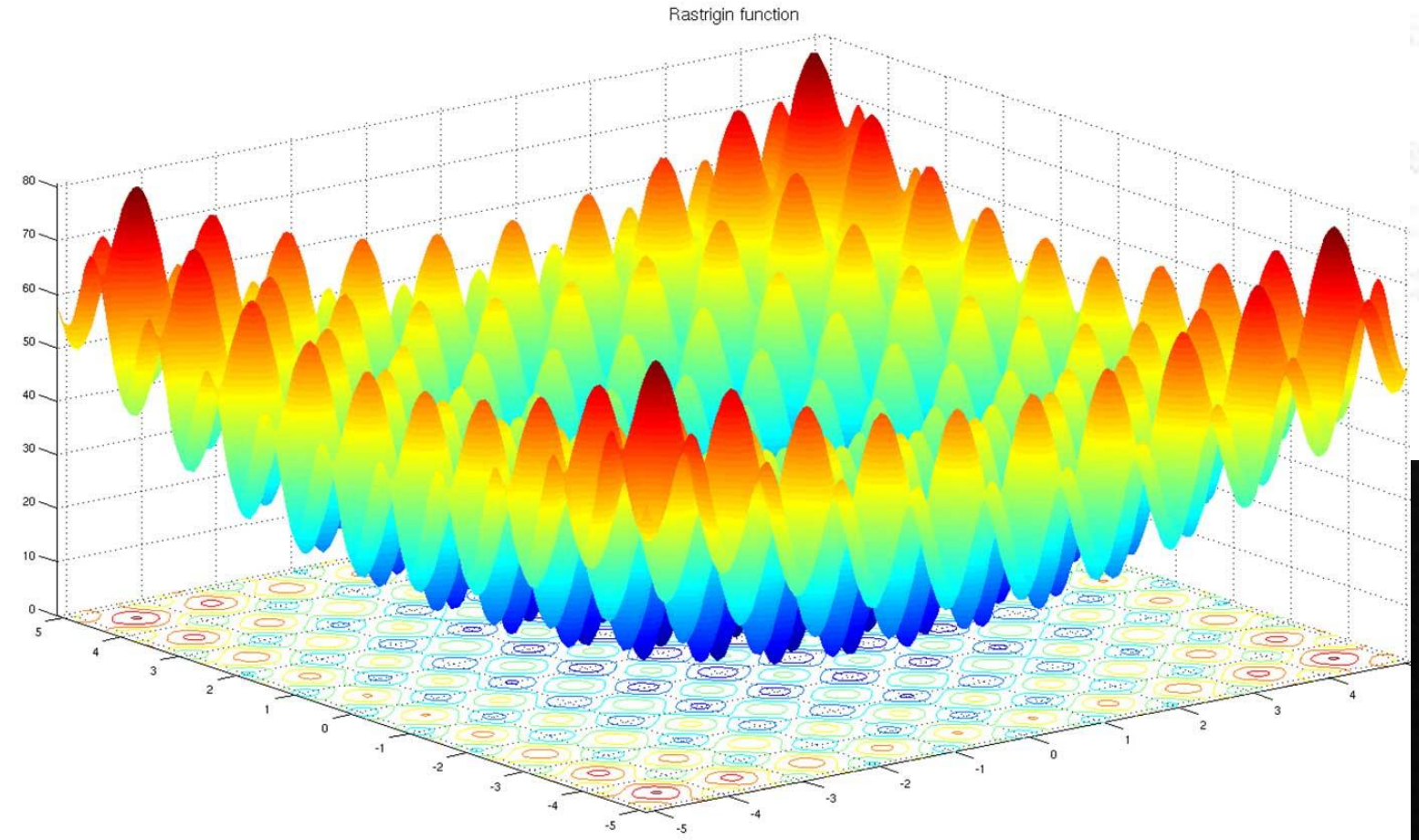

Data Reduction
Dimensionality Reduction
Numerosity Reduction
Data Compression
Wavelet Transform
Principal Component Analysis
Attribute Subset Selection
Nonparametric
Parametric
Data Cube Aggregation
Clustering
Sampling
Histogram
Regression
Graphical Models

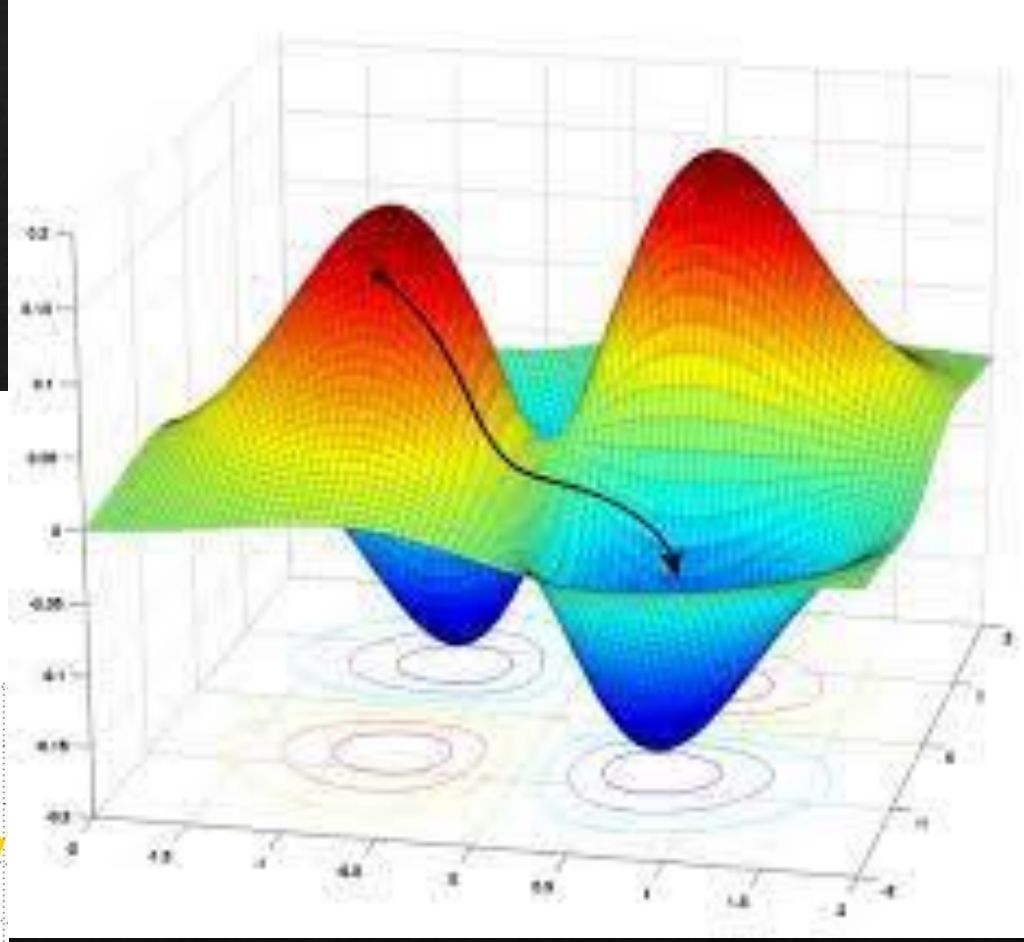

Dimensionality reduction Techniques
Feature Selection
Dimensionality Reduction
Missing Value Ratio
Low Variance Filter
High Correlation Filter
Random Forest
Backward Feature Extraction
Forward Feature Selection
Components/ Factors based
Projection Based
Factor Analysis
Principal Component Analysis
Independent Compone Analysis
ISOMAP
t-SNE
UMAP


Dimensionality Reduction
Feature Extraction
Feature Selection
Unsupervised Technique
Supervised Technique
Unsupervised Technique
Supervised Technique
Linear Method
Non-linear Method
Linear Method
Non-linear Method
Linear Method
Non-linear Method
Linear Method
Non-linear Method
PCA
KPCA
LDA
Kernel local Fisher discriminant analysis
SNR-based method
Bhattacharyya distance
Mutual information

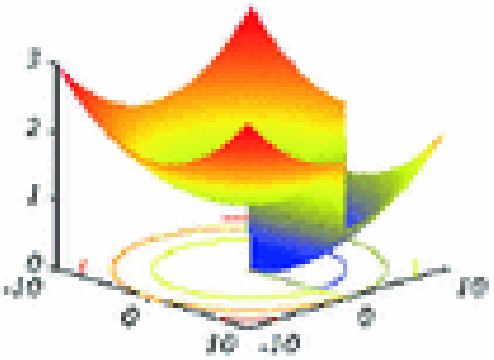
جایگاه روشهای فرا ابتکاری در Optimization

Optimization

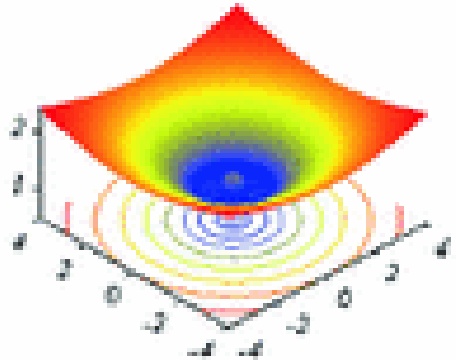




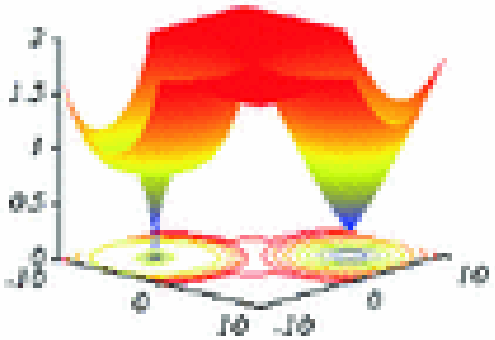
TP1



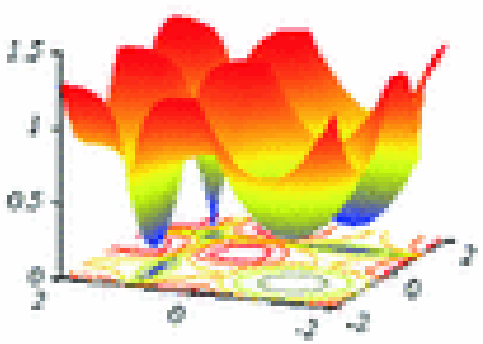
TP2



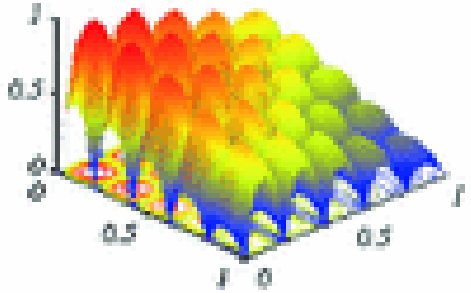
TP3



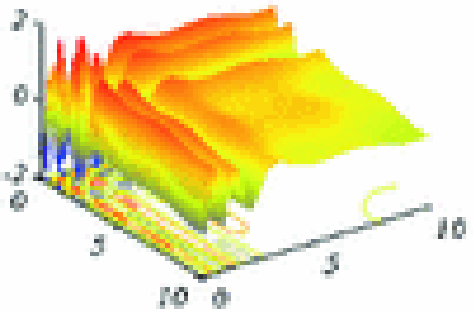
TP4



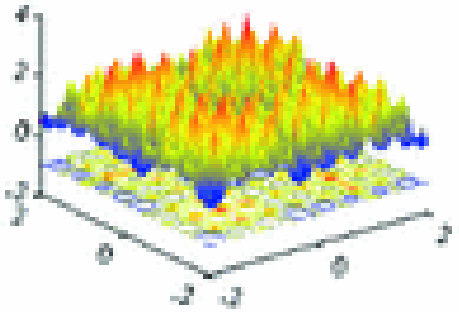
TP5



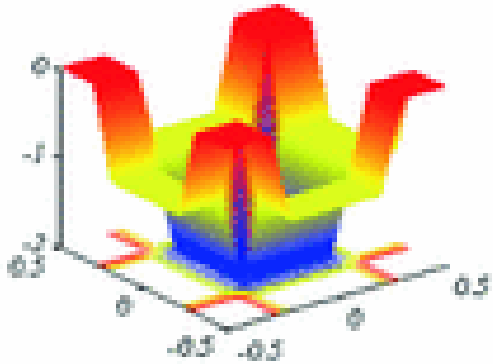
TP6



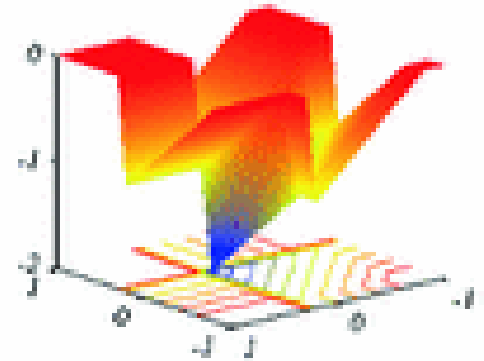
TP7



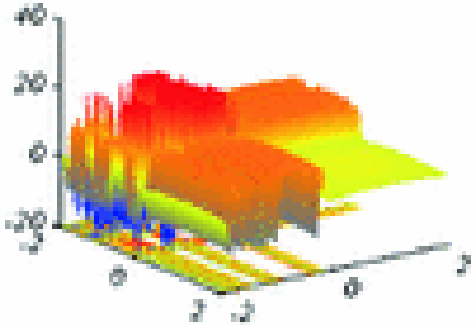
TP8



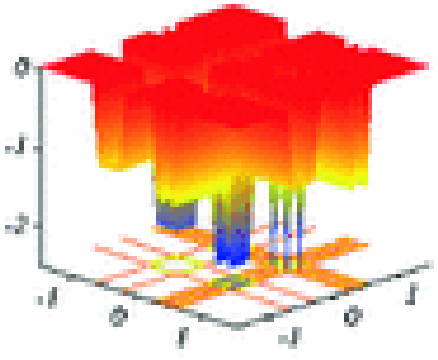
TP9



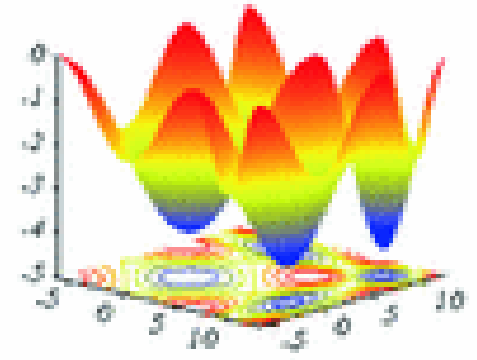
TP10

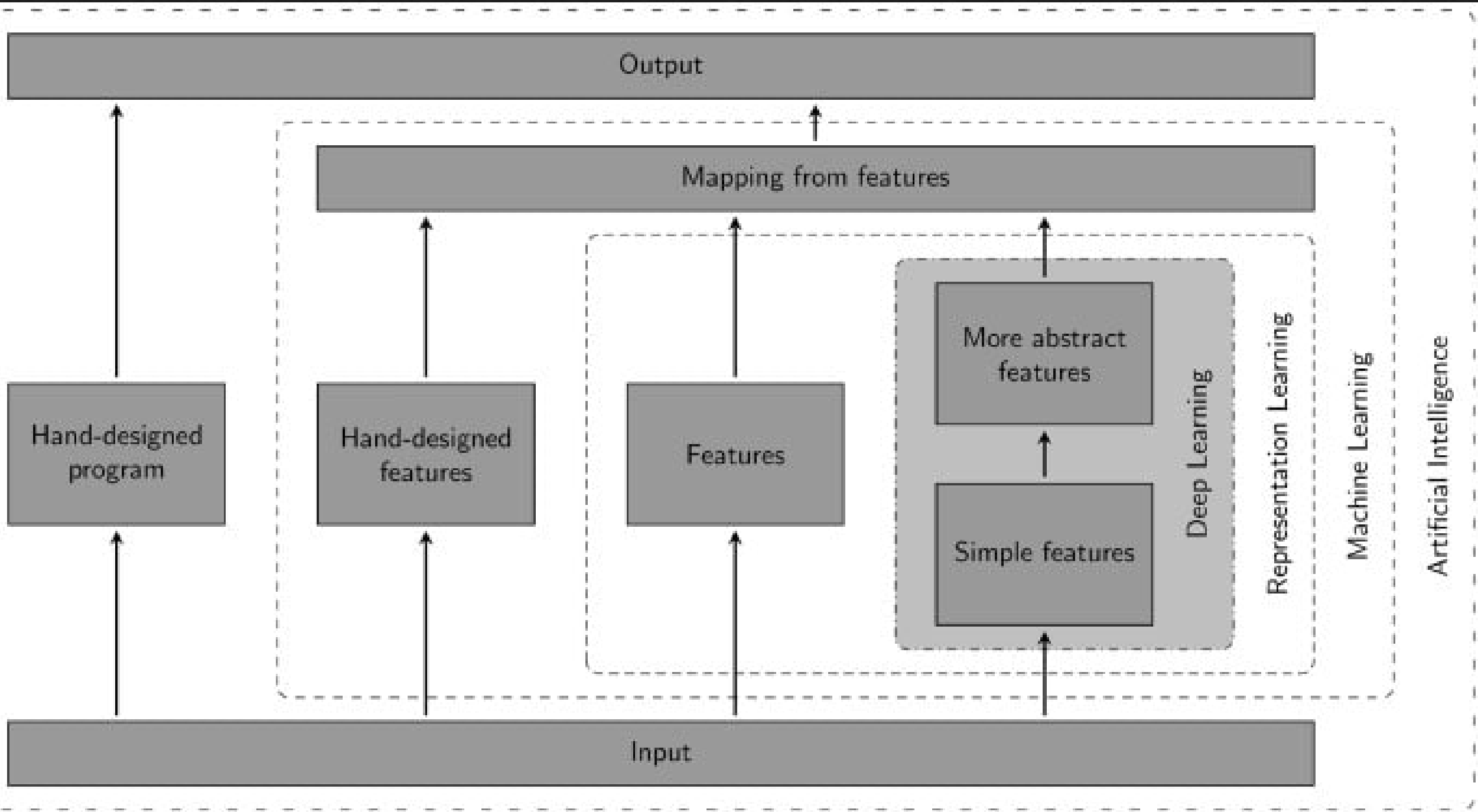


TP11



TP12

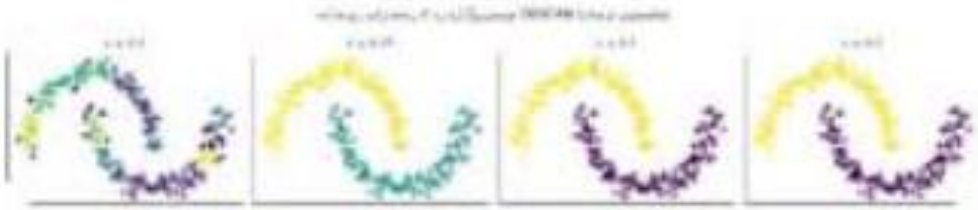



Output
Mapping from features
Hand-designed program
Hand-designed features
Features
More abstract features
Simple features
Deep Learning
Representation Learning
Machine Learning
Artificial Intelligence
Input

Curse of Dimensionality

🔸 تأثیر در الگوریتم‌های مختلف

نوع الگوریتم	اثر نفرین بُعدی
K-Means, KNN	فاصله‌ها معنی‌دار نیستند
Decision Trees	شاخه‌های بیش‌ازحد زیاد ← Overfitting
DBSCAN	چگالی واقعی پیدا نمی‌شود
SVM	نیاز به کرنل مناسب یا کاهش بُعد
Neural Networks	نیاز به داده‌ی زیاد و زمان آموزش طولانی



در این تصویر می‌بینی که مقدار **ε (epsilon)** تأثیر مستقیمی بر نحوه‌ی تشکیل خوشه‌ها در DBSCAN دارد:

- **ε = 0.1:** خوشه‌ها شکسته و خیلی کوچک است ← نقاط زیادی به عنوان نویز (رنگ خاکستری) در نظر گرفته می‌شوند.
- **ε = 0.15:** شروع به اتصال نقاط هم‌جوار می‌کند، اما هنوز تعدادی نویز وجود دارد.
- **ε = 0.2:** بهترین حالت؛ دو خوشه‌ی نعل‌اسبی درست شناسایی شده‌اند.
- **ε = 0.3:** بیش از حد بزرگ ← دو خوشه به هم چسبیده‌اند و به صورت یک خوشه‌ی بزرگ دیده می‌شوند.

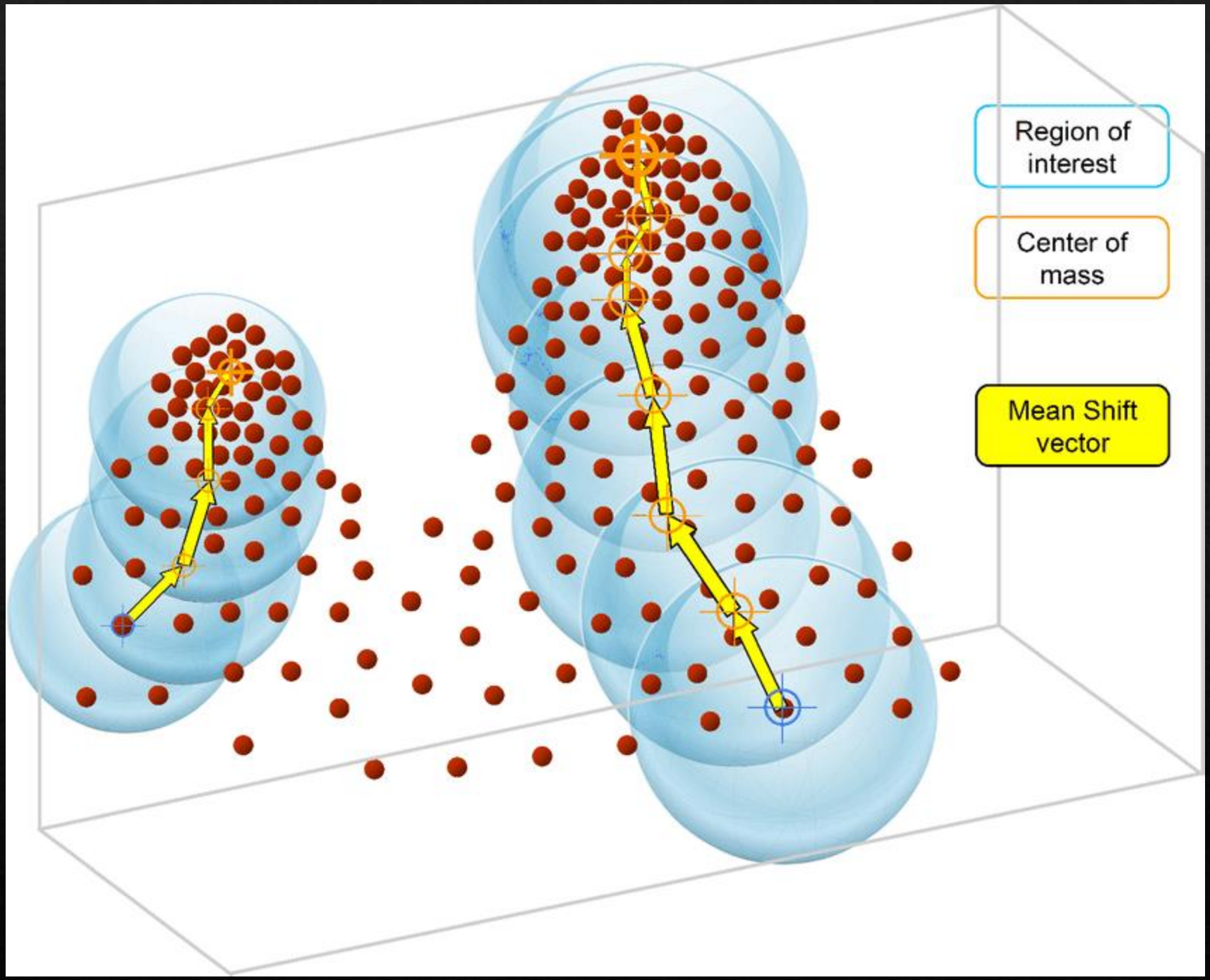
📊 تفاوت CLARANS با CLARA و K-Medoids

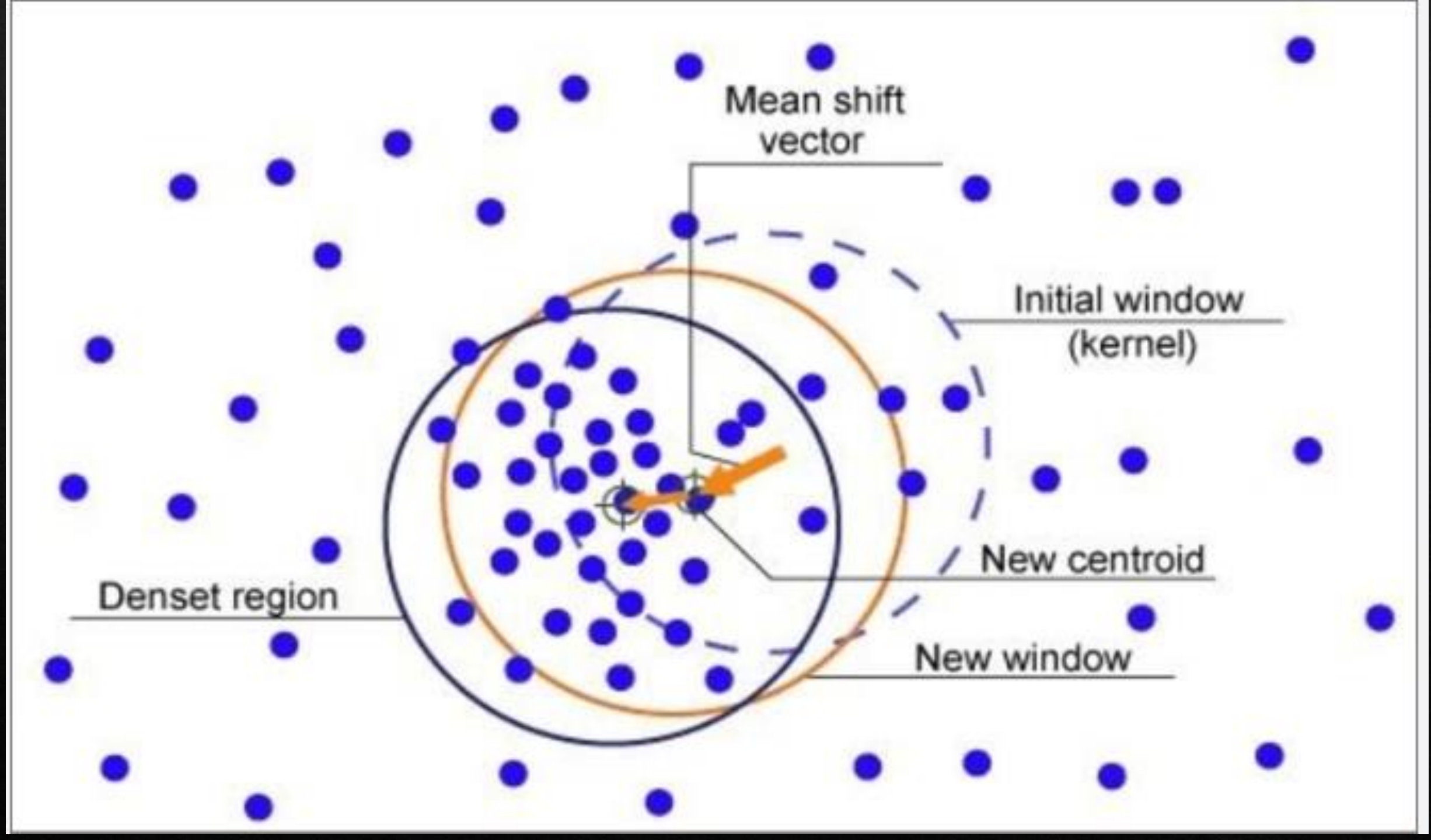
ویژگی	K-Medoids	CLARA
داده‌ی ورودی	کل داده‌ها	نمونه‌گیری از داده‌ها
سرعت	کند	سریع‌تر
دقت	بالا	بستگی به نمونه دارد
روش جستجو	کامل (exhaustive)	روی نمونه‌ها

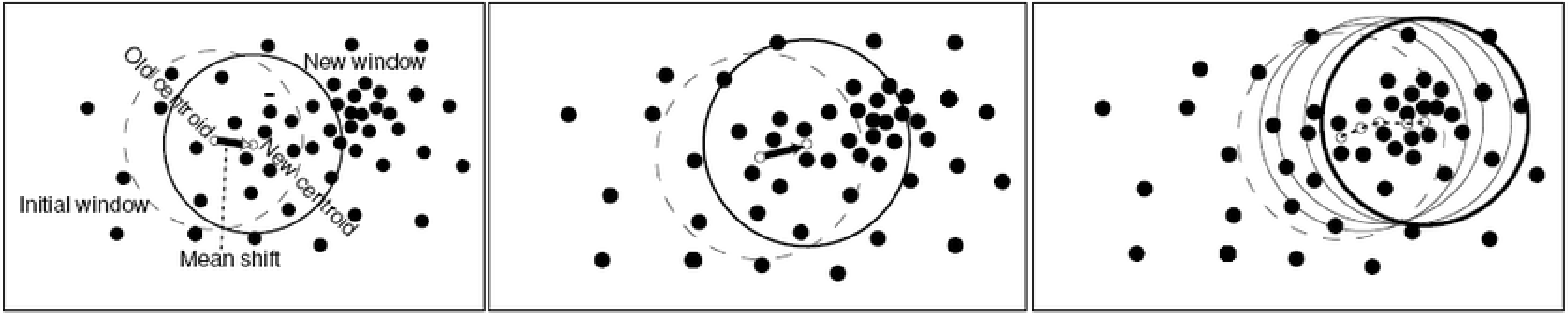
🔸 خلاصه مقایسه

نوع الگوریتم	نیاز به تعیین k	برای شکل‌های غیرکروی
K-Means	✅ بله	❌ خیر
Hierarchical	❌ اختیاری	✅ گاهی
DBSCAN	❌ خیر	✅ بله
GMM	✅ بله	✅ بله
Spectral	✅ بله	✅ بله

Clustering methods
Mean shift


Region of interest
Center of mass
Mean Shift vector

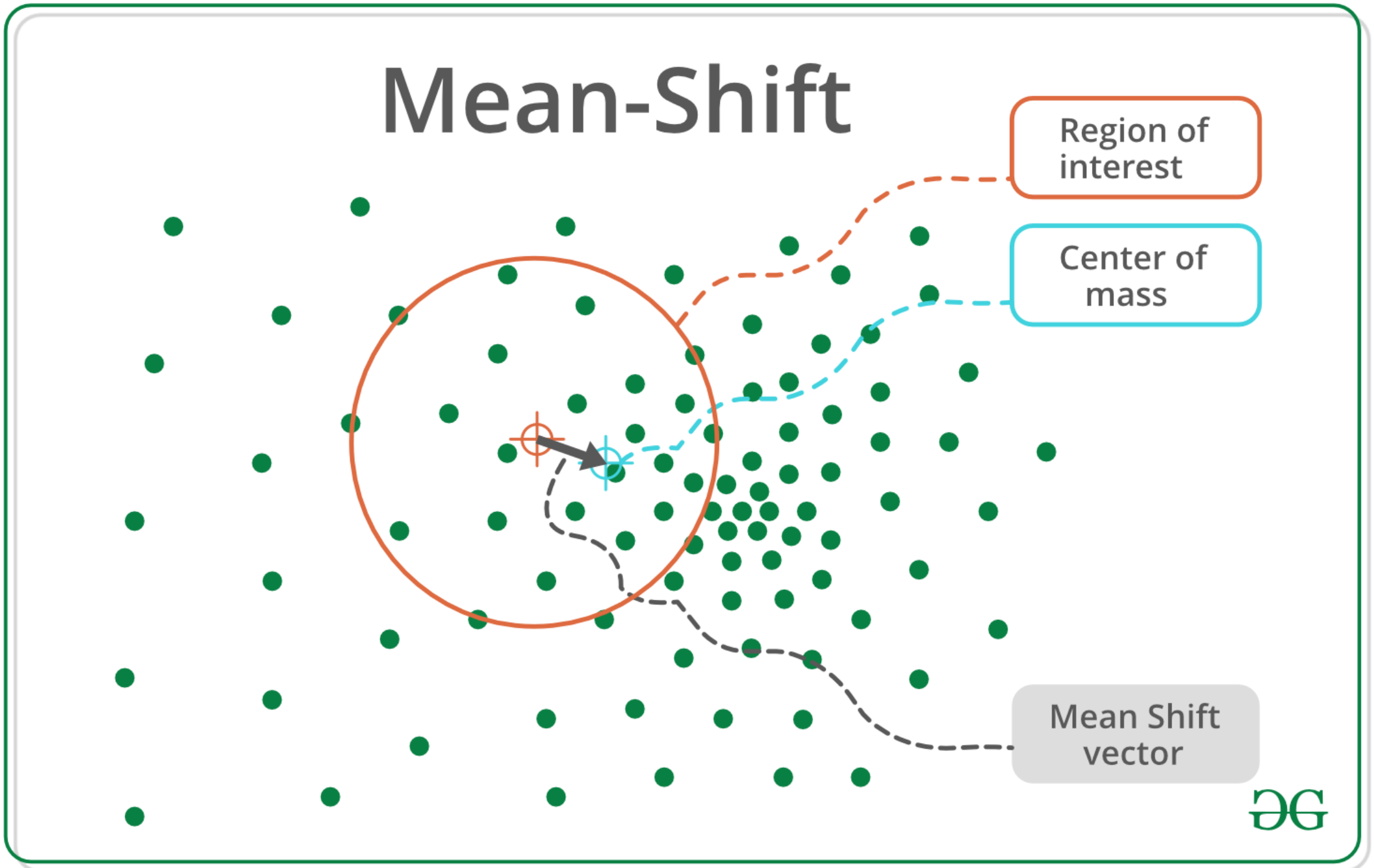

Mean shift
vector
Initial window
(kernel)
New centroid
Denset region
New window

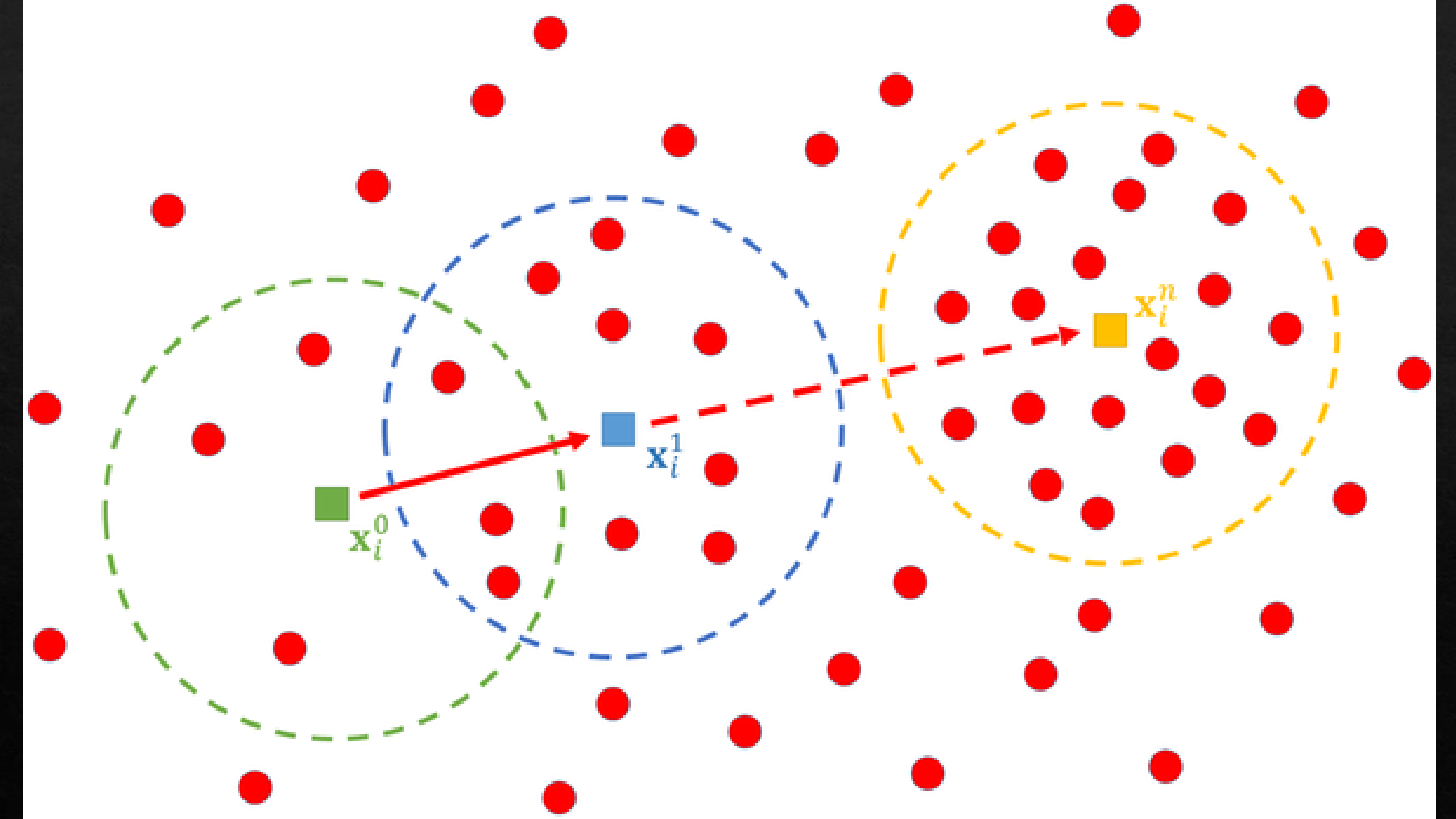


1st Iteration

2nd Iteration

Nth iteration


Mean-Shift
Region of interest
Center of mass
Mean Shift vector


$\mathbf{x}_i^0$
$\mathbf{x}_i^1$
$\mathbf{x}_i^n$

A. Gaussian mean-shift (MS) algorithm

for $n \in \{1, \ldots, N\}$
  $\mathbf{x} \leftarrow \mathbf{x}_n$
  repeat
    $\forall n\colon\ p(n|\mathbf{x}) \leftarrow \frac{\exp\left(-\frac{1}{2}\|(\mathbf{x}-\mathbf{x}_n)/\sigma\|^2\right)}{\sum_{n'=1}^{N}\exp\left(-\frac{1}{2}\|(\mathbf{x}-\mathbf{x}_{n'})/\sigma\|^2\right)}$
    $\mathbf{x} \leftarrow \sum_{n=1}^{N} p(n|\mathbf{x})\mathbf{x}_n$
  until stop
  $\mathbf{z}_n \leftarrow \mathbf{x}$
end
connected-components$(\{\mathbf{z}_n\}_{n=1}^{N},\epsilon)$

B. Gaussian blurring mean-shift (BMS) algorithm

repeat
  for $m \in \{1, \ldots, N\}$
    $\forall n\colon\ p(n|\mathbf{x}) \leftarrow \frac{\exp\left(-\frac{1}{2}\|(\mathbf{x}_m-\mathbf{x}_n)/\sigma\|^2\right)}{\sum_{n'=1}^{N}\exp\left(-\frac{1}{2}\|(\mathbf{x}_m-\mathbf{x}_{n'})/\sigma\|^2\right)}$
    $\mathbf{y}_m \leftarrow \sum_{n=1}^{N} p(n|\mathbf{x}_m)\mathbf{x}_n$
  end
  $\forall m\colon\ \mathbf{x}_m \leftarrow \mathbf{y}_m$
until stop
connected-components$(\{\mathbf{x}_n\}_{n=1}^{N},\epsilon)$

C. Gaussian MS algorithm in matrix form

$\mathbf{Z} = \mathbf{X}$
repeat
  $\mathbf{W} = \left(\exp\left(-\frac{1}{2}\left\|(\mathbf{z}_m - \mathbf{x}_n)/\sigma\right\|^2\right)\right)_{nm}$
  $\mathbf{D} = \operatorname{diag}\left(\sum_{n=1}^{N} w_{nm}\right)$
  $\mathbf{Q} = \mathbf{W}\mathbf{D}^{-1}$
  $\mathbf{Z} = \mathbf{X}\mathbf{Q}$
until stop
connected-components$(\{\mathbf{z}_n\}_{n=1}^{N},\epsilon)$

D. Gaussian BMS algorithm in matrix form

repeat
  $\mathbf{W} = \left(\exp\left(-\frac{1}{2}\left\|(\mathbf{x}_m - \mathbf{x}_n)/\sigma\right\|^2\right)\right)_{nm}$
  $\mathbf{D} = \operatorname{diag}\left(\sum_{n=1}^{N} w_{nm}\right)$
  $\mathbf{P} = \mathbf{W}\mathbf{D}^{-1}$
  $\mathbf{X} = \mathbf{X}\mathbf{P}$
until stop
connected-components$(\{\mathbf{x}_n\}_{n=1}^{N},\epsilon)$

Algorithm 1 Structure of the Mean Shift algorithm.

Data: $\boldsymbol{X} = \{\boldsymbol{x}_i\}$ - input image, $\boldsymbol{Y} = \{\boldsymbol{y}_i\}$ - output image, $\boldsymbol{x}_i^{(t)} = \left(\boldsymbol{\xi}_i^{(t)}, \boldsymbol{\eta}_i^{(t)}\right)$ - pixel at position $\boldsymbol{\xi}_i^{(t)}$ with color vector $\boldsymbol{\eta}_i^{(t)}$,
$\mathscr{B}_i^{(t)}$ - processing block centered at $\boldsymbol{\xi}_i^{(t)}$, $\varepsilon = 10^{-3}$ - stopping criterion

for each image pixel $\boldsymbol{x}_i \in \boldsymbol{X}$ **do**
  $t = 1, \quad \boldsymbol{x}_i^{(1)} = \boldsymbol{x}_i$ (initialization)
  do
    $t{+}{+}$
    calculate $\boldsymbol{x}_i^{(t)} = \left(\boldsymbol{\xi}_i^{(t)}, \boldsymbol{\eta}_i^{(t)}\right)$ according to (13)
    move center of $\mathscr{B}_i^{(t+1)}$ to $\left\lfloor \boldsymbol{\xi}_i^{(t)} \right\rceil$ (rounding to the nearest integer)
  while $\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\| \geq \varepsilon$
  $\boldsymbol{y}_i = \left(\boldsymbol{\xi}_i^{(1)}, \boldsymbol{\eta}_i^{(t)}\right)$
end for

Mean-shift offset

Spatial kernel

Color weight

Local pixel offset

$$\Delta x = \frac{\sum_i K(x_i - x)\, w(x_i)\, (x_i - x)}{\sum_i K(x_i - x)\, w(x_i)}$$

Normalization factor

The Mean-Shift

$$\nabla P(x) = \frac{1}{N} 2c \sum_n g_n \left(\frac{\sum_n x_n g_n}{\sum_n g_n} - x \right)$$

Mean

Shift