# New things in wavelet analysis and clustering

# What are continuous wavelets?

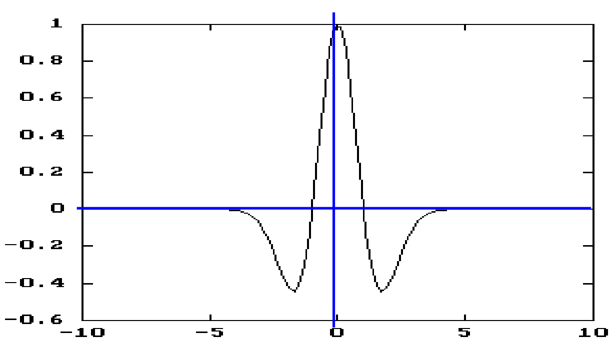
**In contrast to the most known mean of signal analysis as Fourier transform, one-dimensional wavelet transform (WT) of the signal f(x) has 2D form**

$$W_{\psi}(a,b)f = \frac{1}{\sqrt{C_{\Psi}}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{|a|}}\Psi\left(\frac{b-x}{a}\right)f(x)dx,$$

**where the function $\Psi$ is the wavelet, *b* is a displacement (time shift), and *a* is a scale (or frequency). Condition $C_{\psi} < \infty$ guarantees the existence of $\Psi$ and the wavelet inverse transform. Due to the freedom in $\Psi$ choice, many different wavelets were invented.**

The family of **continuous wavelets** with vanishing momenta is presented here by **Gaussian wavelets**, which are generated by **derivatives of Gaussian function**

$$g_n(x) = (-1)^{n+1}\frac{d^n}{dx^n}e^{-x^2/2},$$



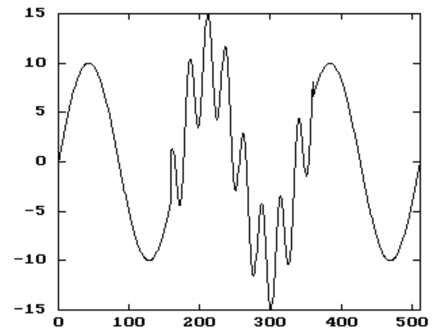

Most known wavelet **G2** is named **"the Mexican hat"**

$$g_2(x) = (1-x^2)e^{-\frac{x^2}{2}}$$

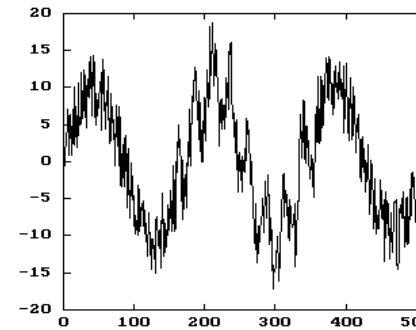
**The biparametric nature of wavelets renders it possible to analyze simultaneously both time and frequency characteristics of signals. So wavelet analysis is used as a mean for smoothing signals, filtering them from noise and, in particular, looking for some tiny artefacts of signals hidden in a heavy background.**

# Wavelets can be applied for extracting very special features of mixed and contaminated signal

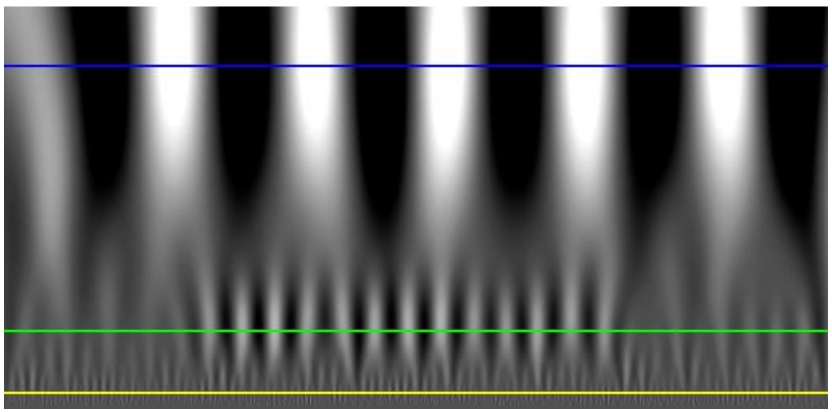
An example of the signal with a localized high frequency part and considerable contamination




Source sample

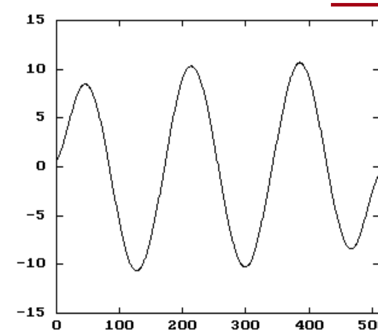



Noise added

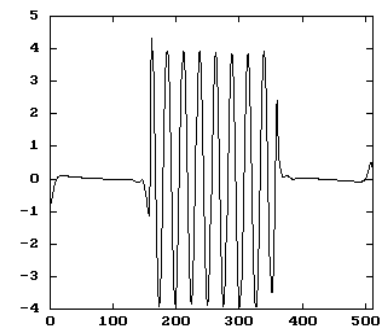


**$G_2$** wavelet spectrum of this signal

then wavelet **filtering** is applied




Low frequency



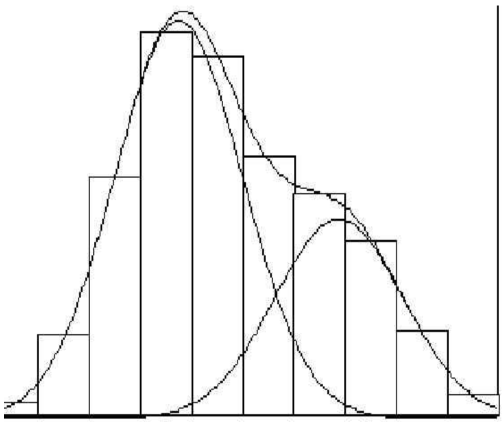

High frequency

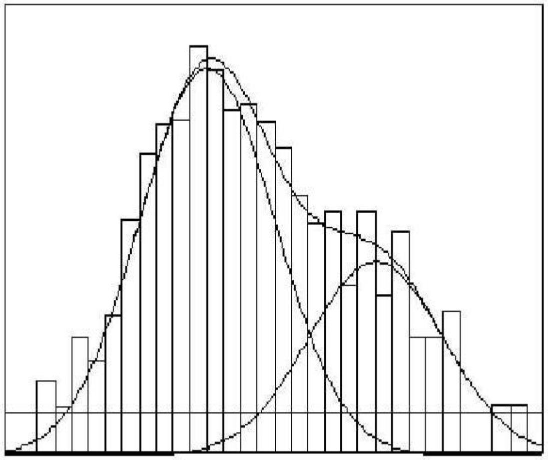
**Filtering** works in the wavelet domain by thresholding of scales, to be eliminated or extracted, and then by making the inverse transform

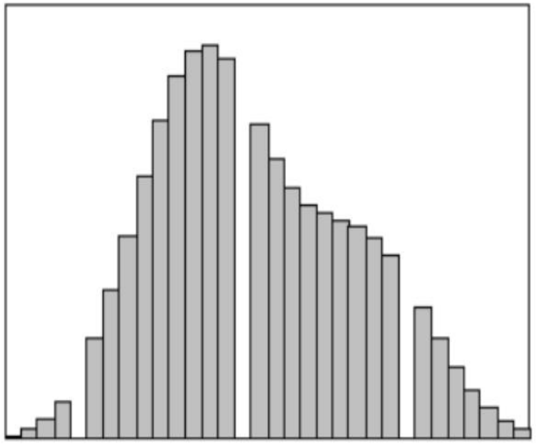
**Filtering results. Noise is removed and high frequency part perfectly localized.**

**NOTE: that is impossible by Fourier transform**

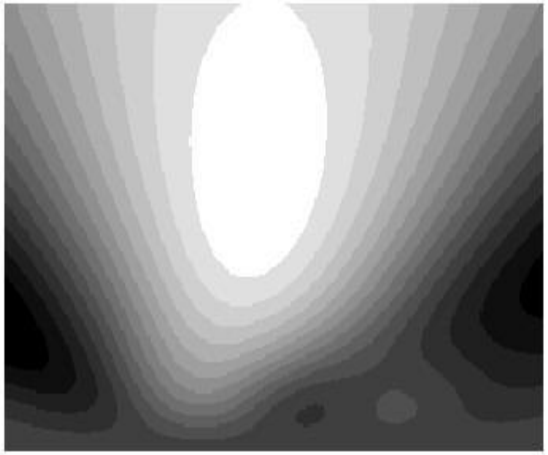
# Нечувствительность вейвлет-спектров к дисторсиям сигналов

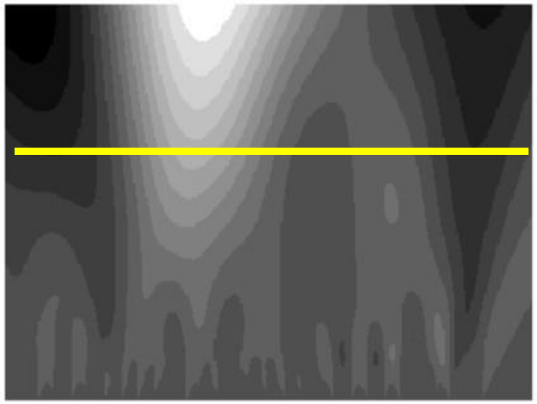


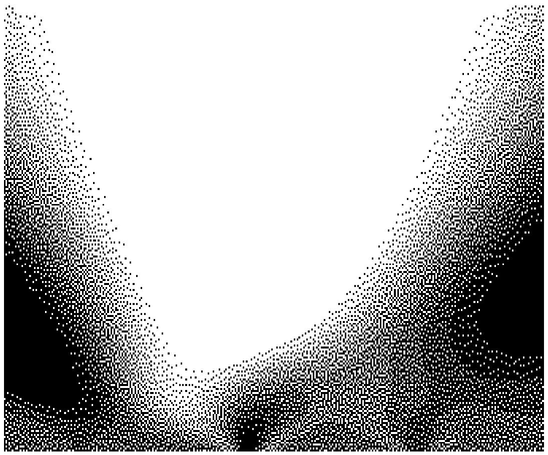




Различие в грануляции и зашумленности                    пропуски данных







Вверху – сигналы, внизу - соответствующие им вейвлет-спектры

# Continuous wavelets: pro and contra

**PRO**: - **Using wavelets we overcome background estimation**

- **Wavelets are resistant to noise (robust)**

**CONTRA**: - **redundancy** → slow speed of calculations

- **nonorthogonality** (**signal distotres after inverse transform!**)

Besides, **real signals** to be analysed by computer are **discrete,**

So **orthogonal discrete wavelets** should be preferable.

**However there are some special feature of continuos wavelets which allows us to avoid inverse transfom, but make our analysis directly in the wavelet domain**

# Back to continuous wavelets

## Peak parameter estimating by gaussian wavelets

When a signal is bell-shaped one, it can be approximated by a gaussian

$$g(x; A, x_0) = A \exp\left(-\frac{(x - x_0)^2}{2\sigma^2}\right).$$

Then it can be derived analytically that its wavelet transformation looks as the corresponding wavelet. For instance, for $G_2(x)$

$$= (1 - x^2)e^{-\frac{x^2}{2}}$$

one has

$$W_{G2}(a,b)g = \frac{Aa^{\frac{5}{2}}\sigma}{(a^2+\sigma^2)^{\frac{3}{2}}} G_2\left(\frac{b - x_0}{\sqrt{(a^2+b^2)}}\right)$$

Considering $W_{G2}$ as a function of the dilation $b$ we obtain its maximum

$$\max_b W_{G2}(a,b) = \frac{Aa^{\frac{5}{2}}\sigma}{(a^2+\sigma^2)^{\frac{3}{2}}}$$

and then solving

the equation $\frac{\partial \max_b(a)}{\partial a} = 0$ we obtain $a_{\max} = \sqrt{5}\sigma$

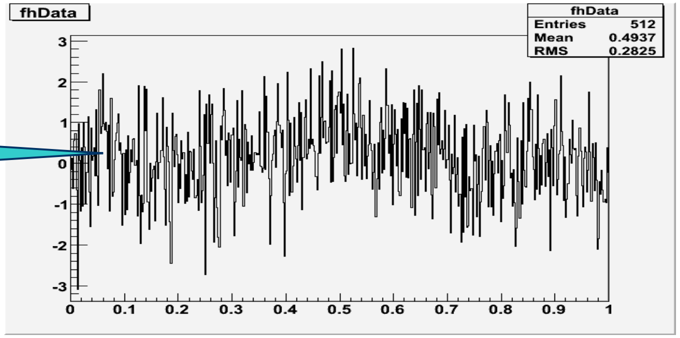
Thus, we can work directly in the wavelet domain instead of time/space domain and use this analytical formula for $W_{G2}(a,b;x_0,\sigma)g$ surface in order to fit it to the surface, obtained for a real invariant mass spectrum.

The most remarkable point is: since the fitting parameters $x_0$ and $\sigma$, can be estimated directly in the $G_2$ domain, we do not need the inverse transform!

# Estimating peak parameters in $G_2$ wavelet domain

***How it works?***

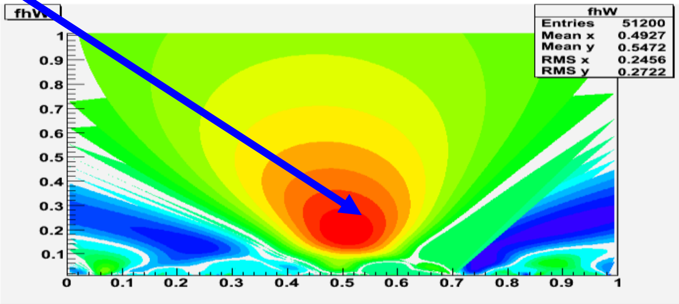
*Let us have a noisy **invariant mass spectrum***




*peak has bell-shape form*

1. *transform it by $G_2$ into wavelet domain*
2. *2. look for the wavelet* **surface maximum**

$b_{max}$, $a_{max}$. 3. *From the formula for $W_{G2}(a,b;x_0,\sigma)g$ one can derive analytical expressions for its maximum $x_0$ and* $a_{max} = \sqrt{5}\sigma$ *which should correspond to the found $b_{max}$, $a_{max}$. Thus we can use coordinates of the maximum as estimations of wanted peak parameters* $\hat{x}_0, \hat{a}$




*4. From them we can obtain halfwidth*

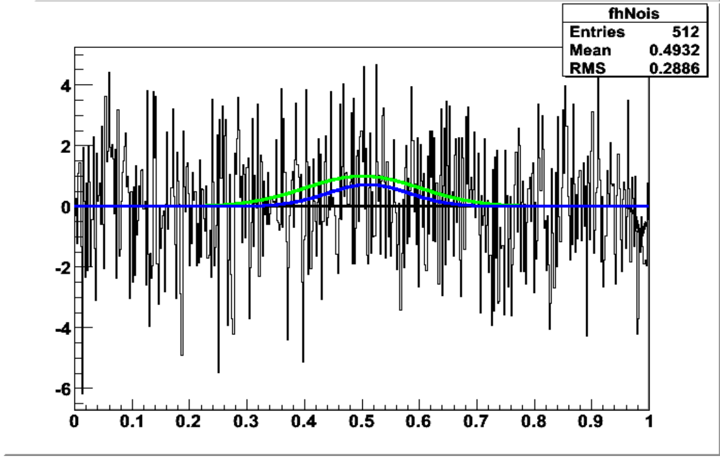
$$\hat{\sigma} = \frac{\hat{a}}{\sqrt{5}}$$

*amplitude*

$$\hat{A} = \frac{\max W}{\hat{a}^{\frac{5}{2}}\hat{\sigma}}(\hat{a}^2 + \hat{\sigma}^2)^{\frac{3}{2}}$$

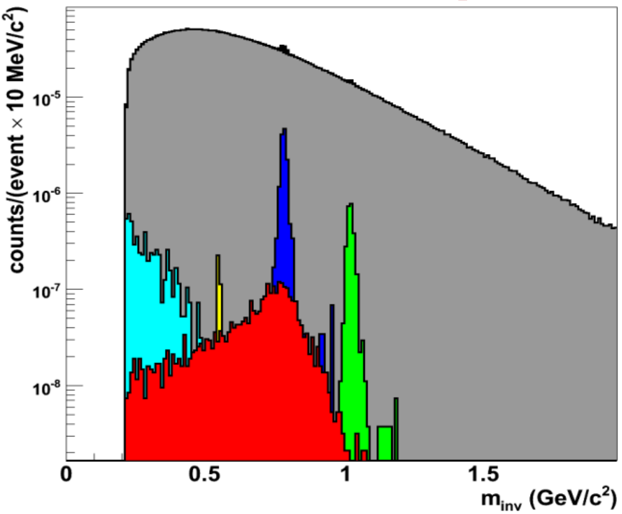
*and even the integral*

$$I = A\sigma\sqrt{2\pi}$$

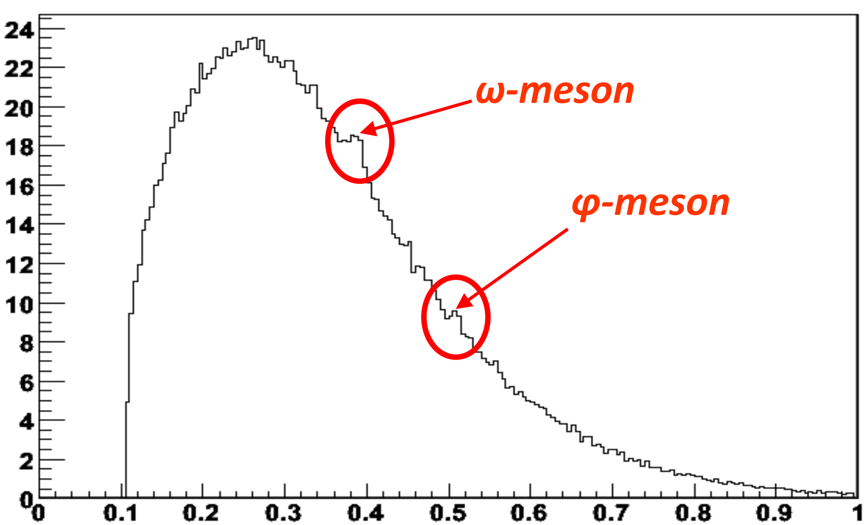


# Application results to CBM spectra

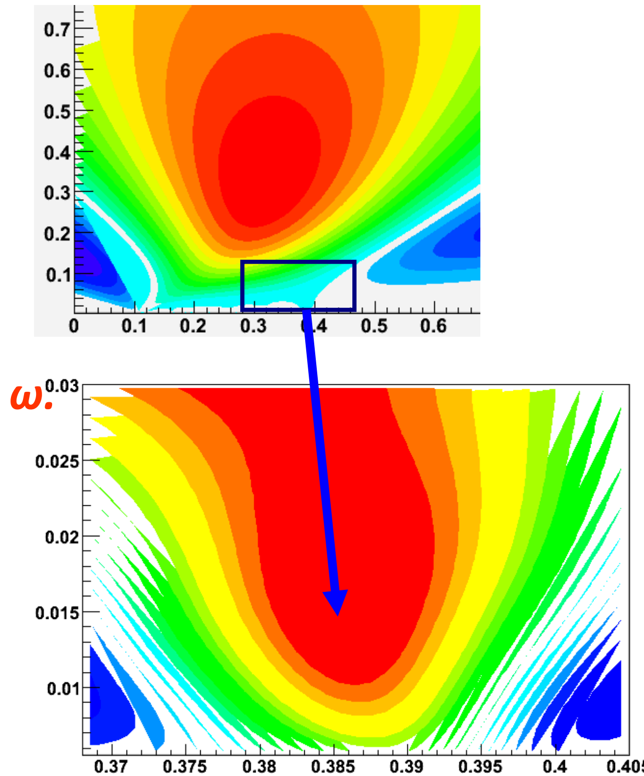
## *Low-mass dileptons (muon channel)*




*Thanks to Anna Kiseleva*




- *ω– wavelet spectrum*

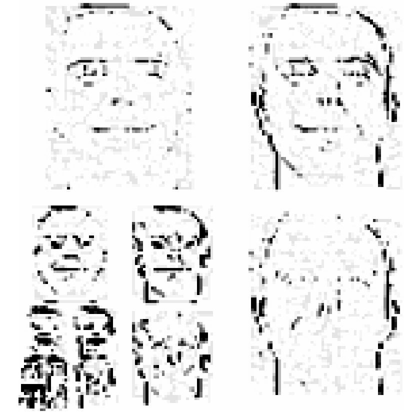


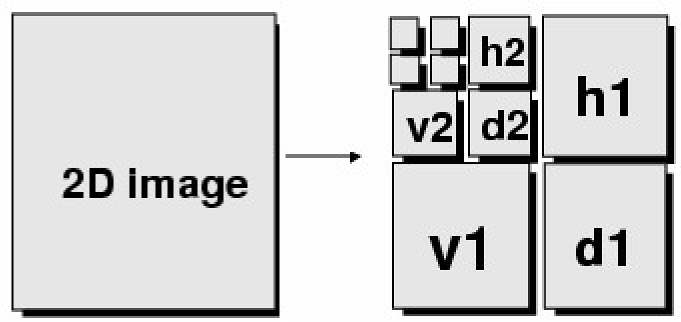

***ω.** Gauss fit*
*<u>of reco signal</u>*
***M=0.7785***
***σ =0.0125***
***A=1.8166***
***$I_g$=0.0569***
***ω.** <u>Wavelets</u>*
***M=0.7700***
***σ =0.0143***
***A=1.8430***
***$I_w$=0.0598***

***Even φ- and ψ′ mesons have been visible in the wavelet space, so we could extract their parameters.***

# Wavelet preprocessing for 2D images

A fast algorithm was developed for 2D-wavelet.
Applying Daubechies wavelets to the image on
the left we obtain the following wavelet expansion






Summarizing three 2D-wavelet components
– vertical, horizontal and diagonal
we obtain the wavelet transform independent
on the image variability of lightening, background
and size.

Lower row shows results of applying  2-d order
2D-wavelets to face images of the upper row

# Image compression

**Ingrid Dobeshi picture restored after wavelet compression up to 3% of original**

Fingerprint compression renders it possible to store in DB 2% of originals only

Oriinal

restored after 26:1 compression

# Application to the hadronic jets reconstruction

## Description of the algorithm

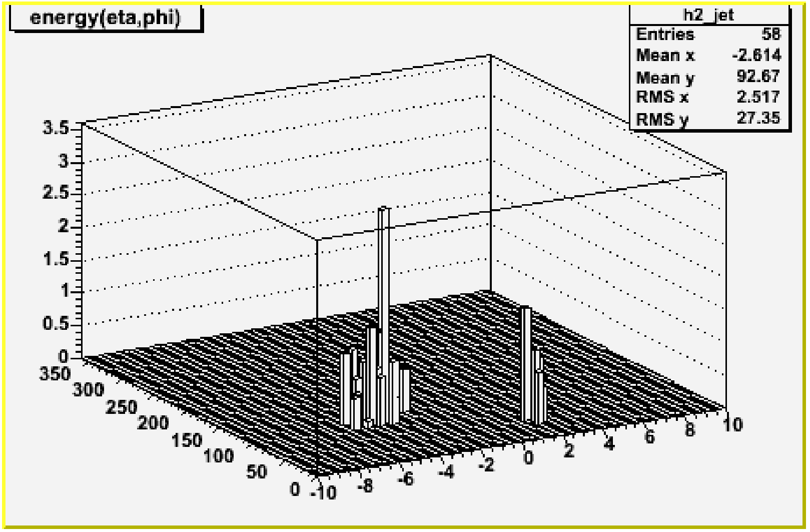
- decompose event into a set of wavelet layers (2-dimensional DWT)
- calculate for each layer RMS.
- apply “hard” rule with threshold value equal $\boldsymbol{\lambda}$***RMS**$_{\mathbf{layer}}$ for each layer of decomposition individually**,** where $\boldsymbol{\lambda}$ is a global control parameter for all layers;
- make the inverse transformation (IDWT);
- accept all residuary peaks as possible jet directions

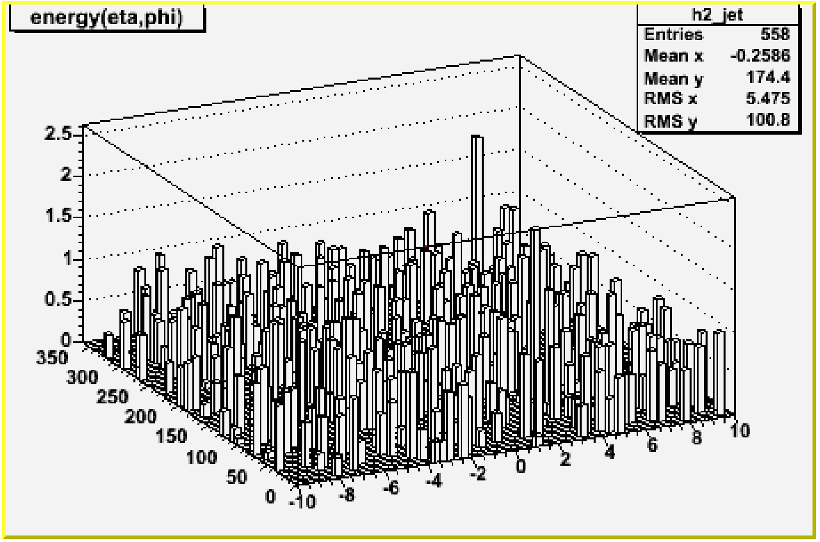
## Wavelet basis

For de-noising orthogonal wavelets are used (“coiflets”).

- Works better compare to another ones.
- The most symmetric from the orthogonal ones.
- We use coiflet with minimal filter length and one vanishing moment.

# Reconstruction of two jets with  different width



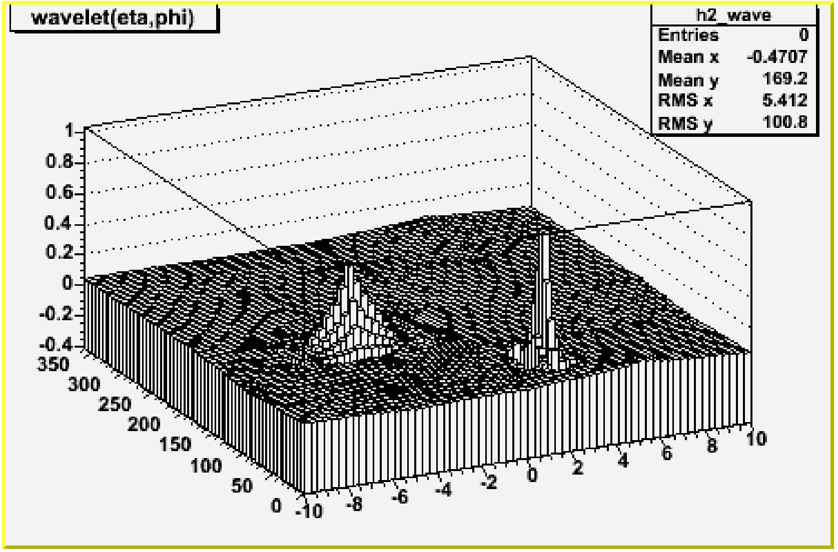




An example with two simulated jets with different width (cone size)

: Two simulated jets before adding background;

: Uniform noise added.

: Two peaks with different width after wavelet filtering.



# clustering

# How big data could be clustered

In many fields of today's science – biology, physics, geology, etc researchers deal with so-called **big data** when the amount of input data is especially large causing such difficulties as:

- the number of measurements to be processed is extremely large – $\mathbf{10^6}$ and more;
- the feature space has many dimensions;
- no preliminary information about the number and locations of the sought-for regions.

Disadvantages of *k*-means clustering in this case

  – fixed number of clusters in the feature space
  – changing number of clusters results in completely different clustering - no sign of succession

On the other hand, there are algorithms that have no disadvantages like these, although they have a much higher complexity and, therefore, unsuitable for processing large amounts of input data.

# New strategy of clustering – two steps

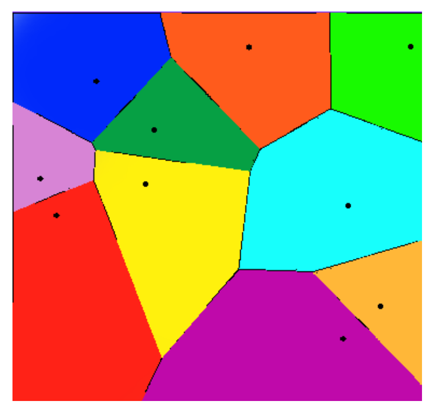
**In the first step** the data undergoes **intermediate clustering** producing clusters which number is much smaller than the number of original objects.

For clustering on the first step we choose **Voronoi partition.** It divides the vector space in sets of points so that for each subset $S_j$ of the partition one can choose such reference vector $C_j$ that all objects $x \in S_i$ of the subset are nearer to it than to any other reference vector $C_j$ *(i≠j).*

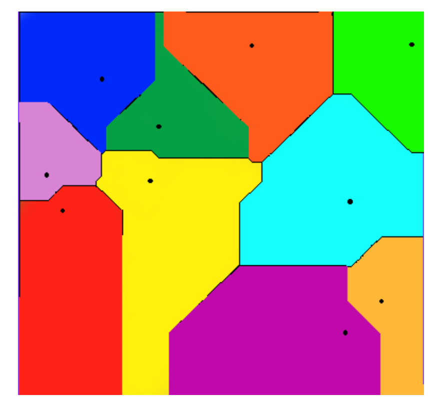
One should keep in mind that that the Voronoi cells depend significantly on the metric used.

**One example**

Estimation of the number of customers of a given shop by the nearest distance considerations. When customers go to the shop on foot by shortest way, Euclidean distance is used, but if they go by a vehicle and the traffic paths are parallel, then a more realistic distance function will be the Manhattan distance



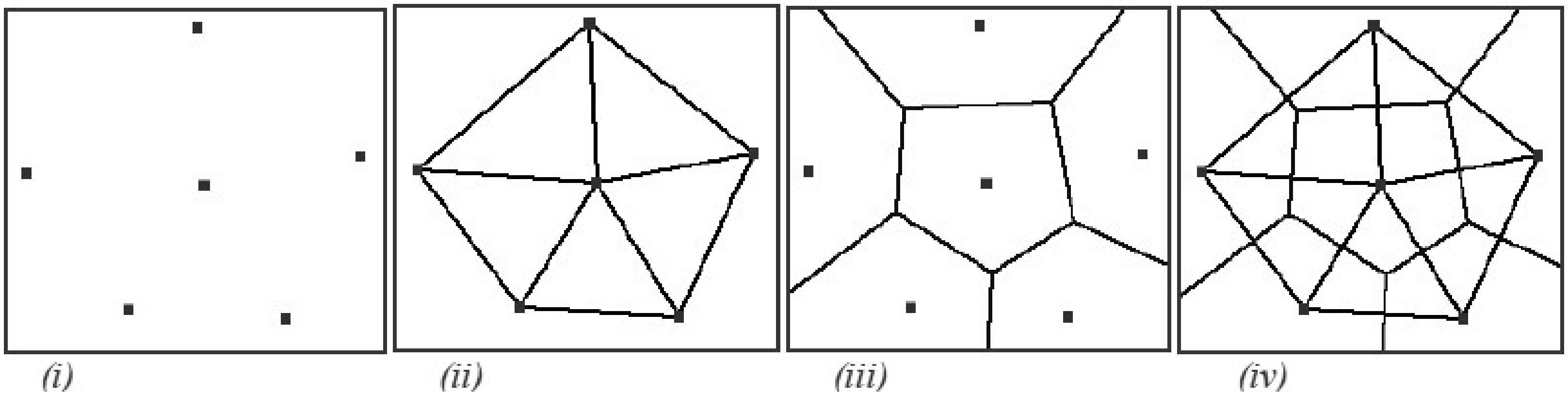
10 shops in a flat city and their Voronoi cells (Euclidean distance).



The same 10 shops, now under Manhattan distance.

# Delaunay triangulation and Voronoi diagram correspondence

The Delaunay triangulation corresponds to the Voronoi diagram in a one-to-one manner: the triangulation links the reference vectors whose Voronoi regions have common boundaries



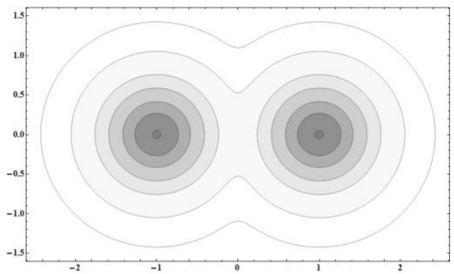

Formation of a Voronoi diagram on a plane: (i) nods on the plane, (ii) Delaunay triangulation, (iii) Voronoi diagram, (iv) superposition of the Delaunay triangulation and the resulting Voronoi diagram.

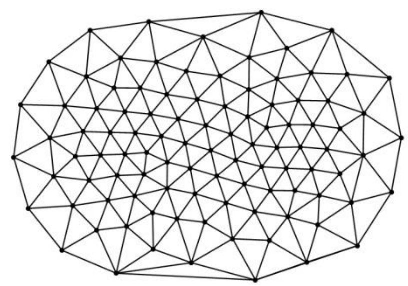
# How it works by Growing Neural Gas

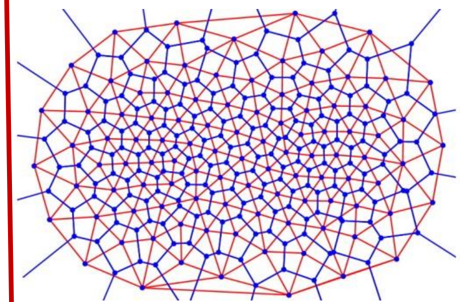
**S.V. Mitsyn, G.A. Ososkov, The Growing Neural Gas and Clustering of Large Amounts of Data, Optical Memory and Neural Networks (Information Optics), 2011, Vol. 20, No. 4, pp. 260–270.**

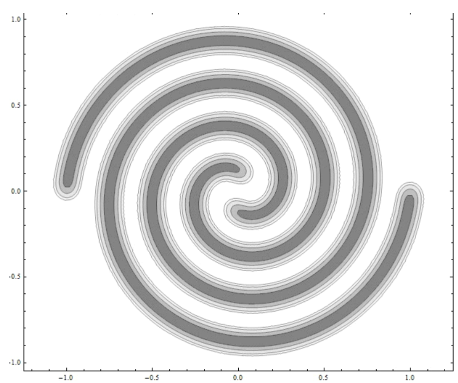
Two examples of objects to be partitioned into Voronoi mosaic

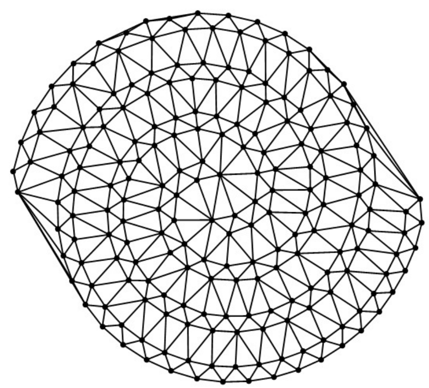
| source data | Delaney triangulation | Voronoi mosaic |
|---|---|---|

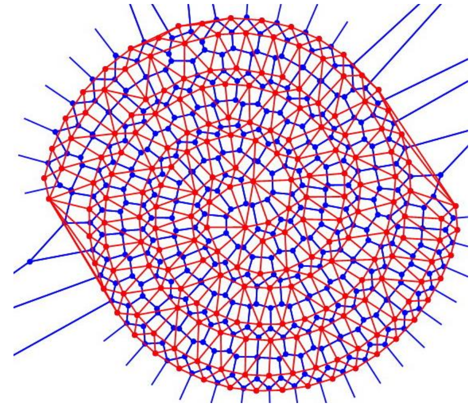






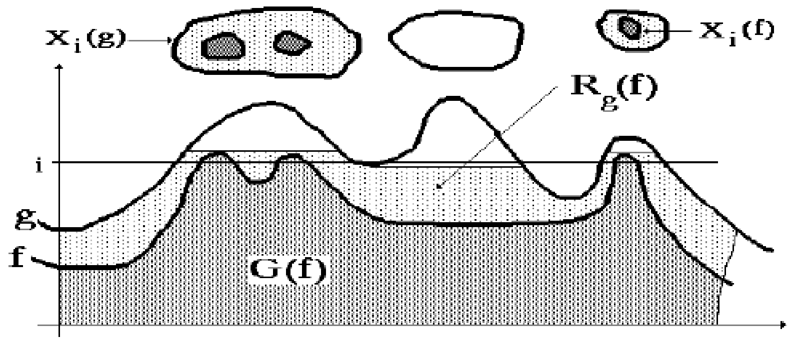



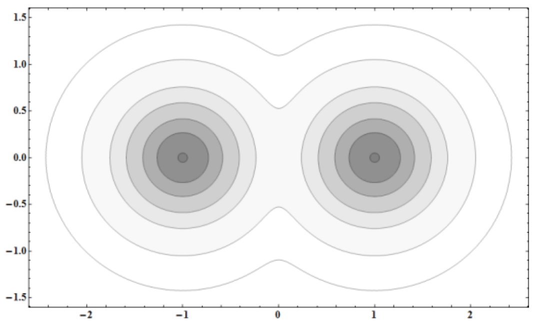


# The second step of clustering

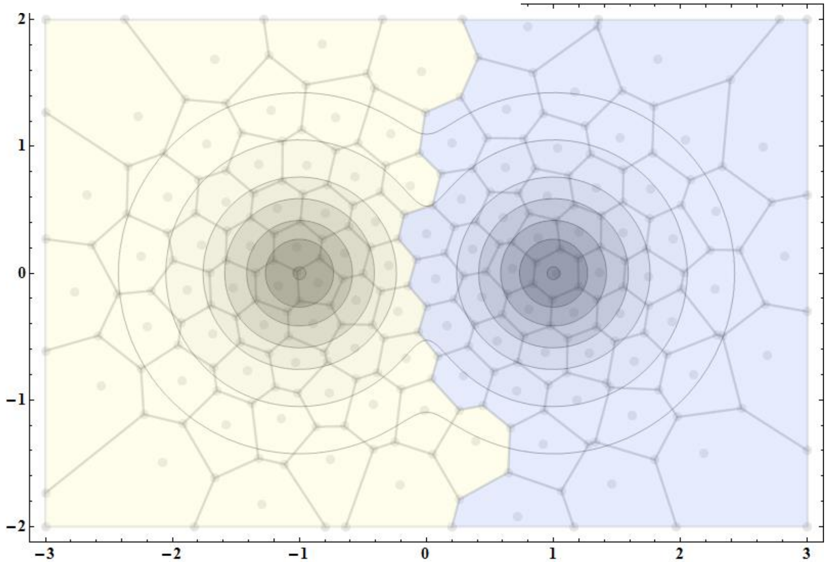
**Final clustering by watershed**

**watershed as geodesic reconstruction**



Initial distribution



Result of watershed clustering

**Thanks to Serge Mitsyn**