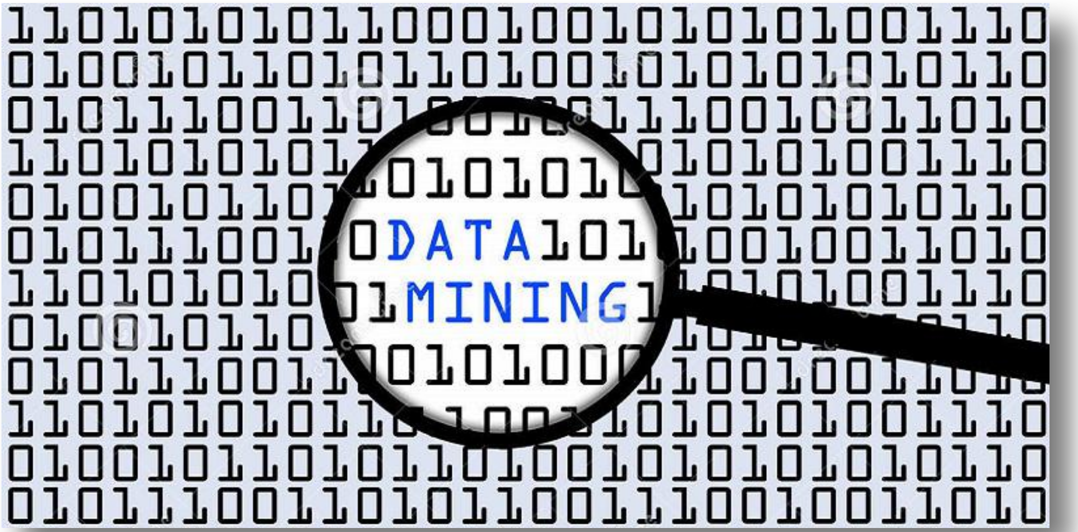
AGGLOMERATIVE CLUSTERING FOR AUDIO CLASSIFICATION USING LOW-LEVEL DESCRIPTORS

Frédéric Le Bel - http://repmus.ircam.fr/lebel - 2016.

ircam Centre Pompidou
UNIVERSITÉ PARIS 8
Fonds de recherche Société et culture Québec

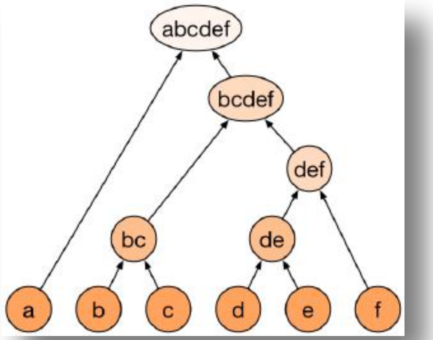
INTRODUCTION

- **Music information retrieval** (MIR)
  - Corpus-based concatenative synthesis [Schwarz, 2006]
  - Musical genre recognition [Peeters, 2007]
  - Computer-aided orchestration [Carpentier, 2008]
- Different framework
  - **Computer-aided composition**
    - Towards formalizing music
  - **Computational musicology**
    - Towards de-formalizing music
- Audio data mining
  - Analysis
  - Exploration
  - Understanding

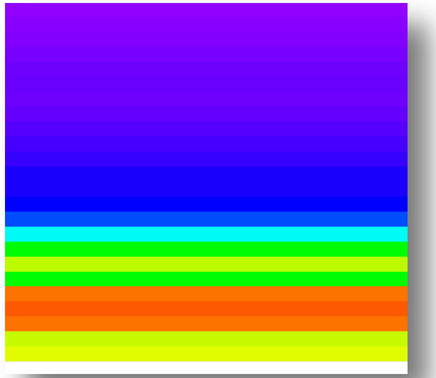


DEFINITIONS

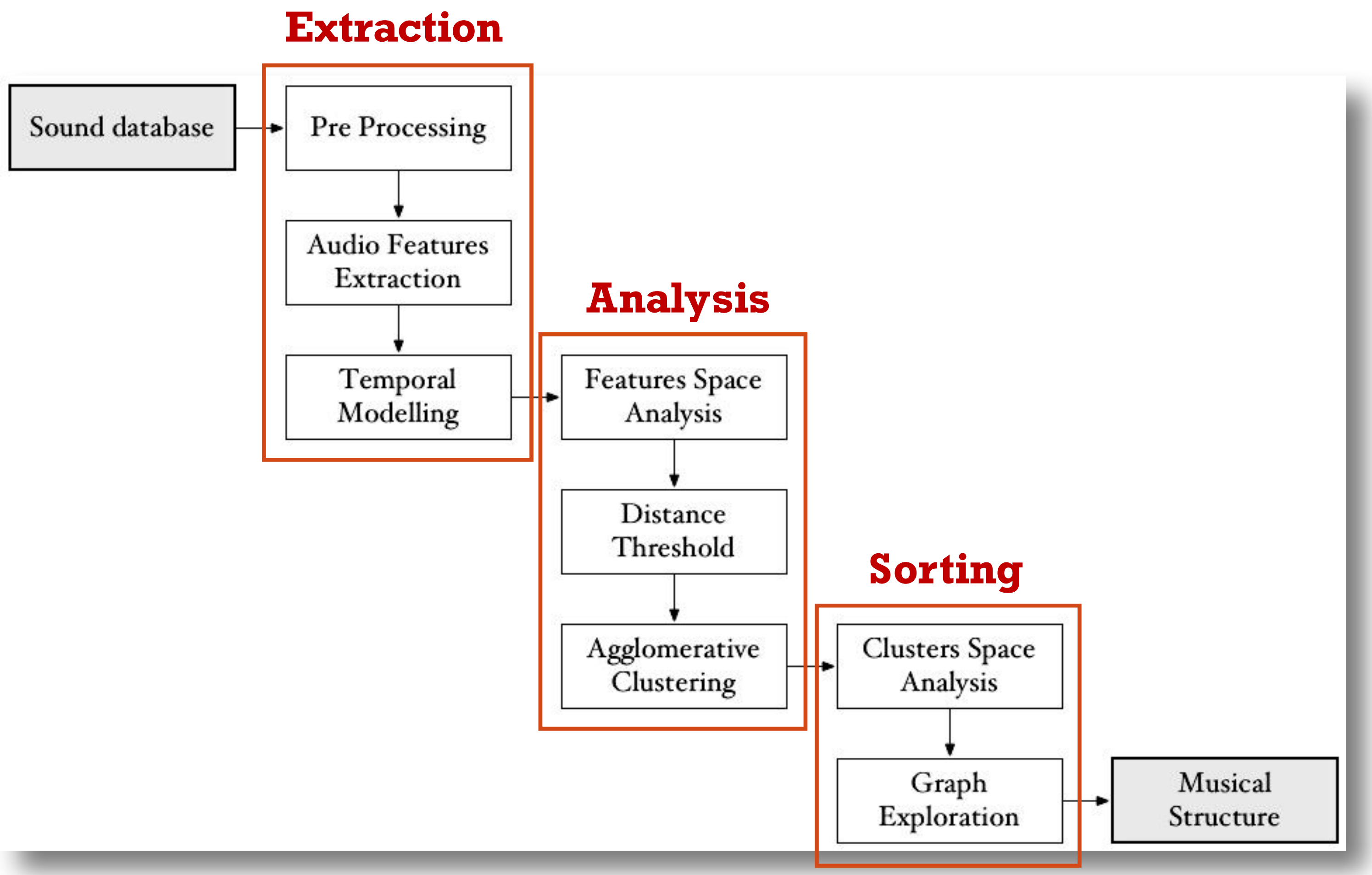
- **Agglomerative Clustering** [Defays, 1977]
  - Hierarchical cluster analysis (HCA)
  - **Unsupervised** learning method
  - Seeks to build a **hierarchy of clusters** (bottom-up)
  - Dendrogram (dendro = tree, gramma = drawing)

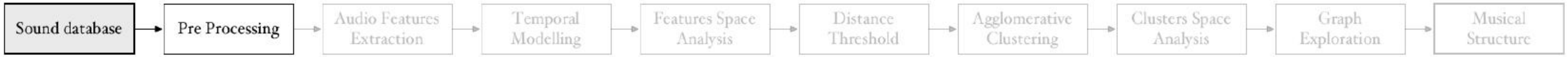


- **Low-level descriptors** and **Audio features** [Malt, 2012]
  - Mathematical operators
  - Transform a raw signal into a **smaller space of variables**
  - Specific audio features (loudness, sharpness, spectral variation, etc.)
  - Audio features = measurable **properties of sounds**
  - Information relevant for **pattern recognition**



STRUCTURAL OVERVIEW

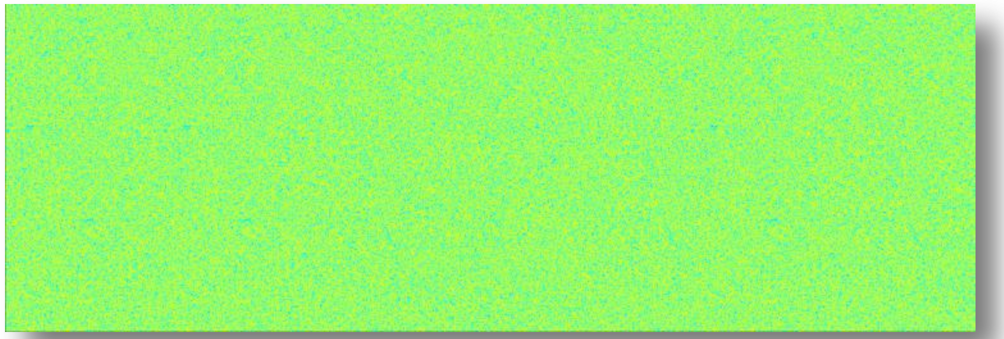




PRE-PROCESSING

Applying different kinds of filters, the idea is to emulate the human selective listening skill…

- Sampling rate: *signal smoothness*
- Number of channels: *auditory scene*
- De-noise, hum removal, […], spectral repair: ***patterns clarity***
- Segmentation, auto-trim, effective duration: ***patterns identification***
- Normalization: *data amplification*
- **Banal but crucial…**



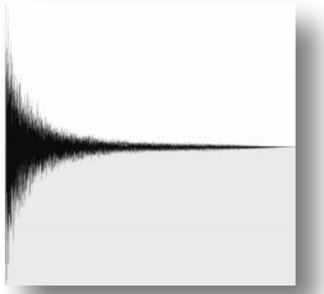
Raw signal (FFT)

Processed signal (FFT)

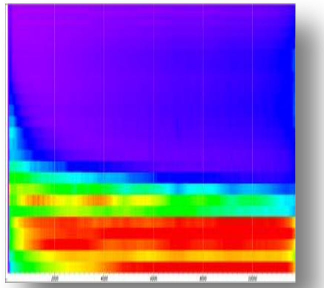
Sound database → Pre Processing → Audio Features Extraction → Temporal Modelling → Features Space Analysis → Distance Threshold → Agglomerative Clustering → Clusters Space Analysis → Graph Exploration → Musical Structure

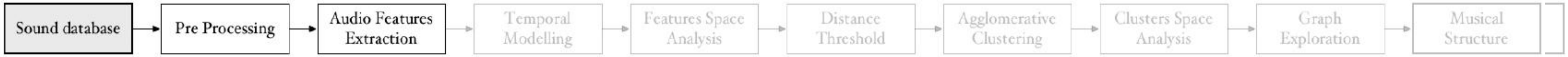
SIGNAL MODELLING

- **Physical model(s)** = raw signal
  - Energy envelope estimation (sampling)
  - Temporal segmentation (windowing)
  - Short-time Fourier Transform (STFT) [Cooley, 1965]
  - Harmonic sinusoid model approximation [Depalle, 1993]



- **Perceptual model(s)** = transformed signal
  - Mid-ear filtering (Fletcher-Munson curves) [Moore, 1997]
  - Mel scale conversion (critical bands filtering) [Rabiner, 1993]
  - Bark scale conversion (another type of critical bands filtering) [Zwicker, 1980]





LOW-LEVEL DESCRIPTORS

- **PHYSICAL MODEL based**
  - *Global* ***temporal*** *descriptors*
    - *Log attack time, temporal increase/decrease, amplitude modulation, MDF, EEV, EFD, TCN*
  - Instantaneous **temporal** descriptors
    - Energy envelope, auto-correlation, zero crossing rate

- **PERCEPTUAL MODEL based**
  - Instantaneous **energy** descriptors
    - Loudness, spread, relative specific loudness
  - Instantaneous **spectral** descriptors
    - MFCC, sharpness, spread, skewness, kurtosis, decrease, roll off, PVA, PSD, POE, PTR, SFM, SCM
  - Instantaneous **harmonic** descriptors
    - FQ0, inharmonicity, noisiness, Chroma, cs-analysis, partial tracking, masking effects

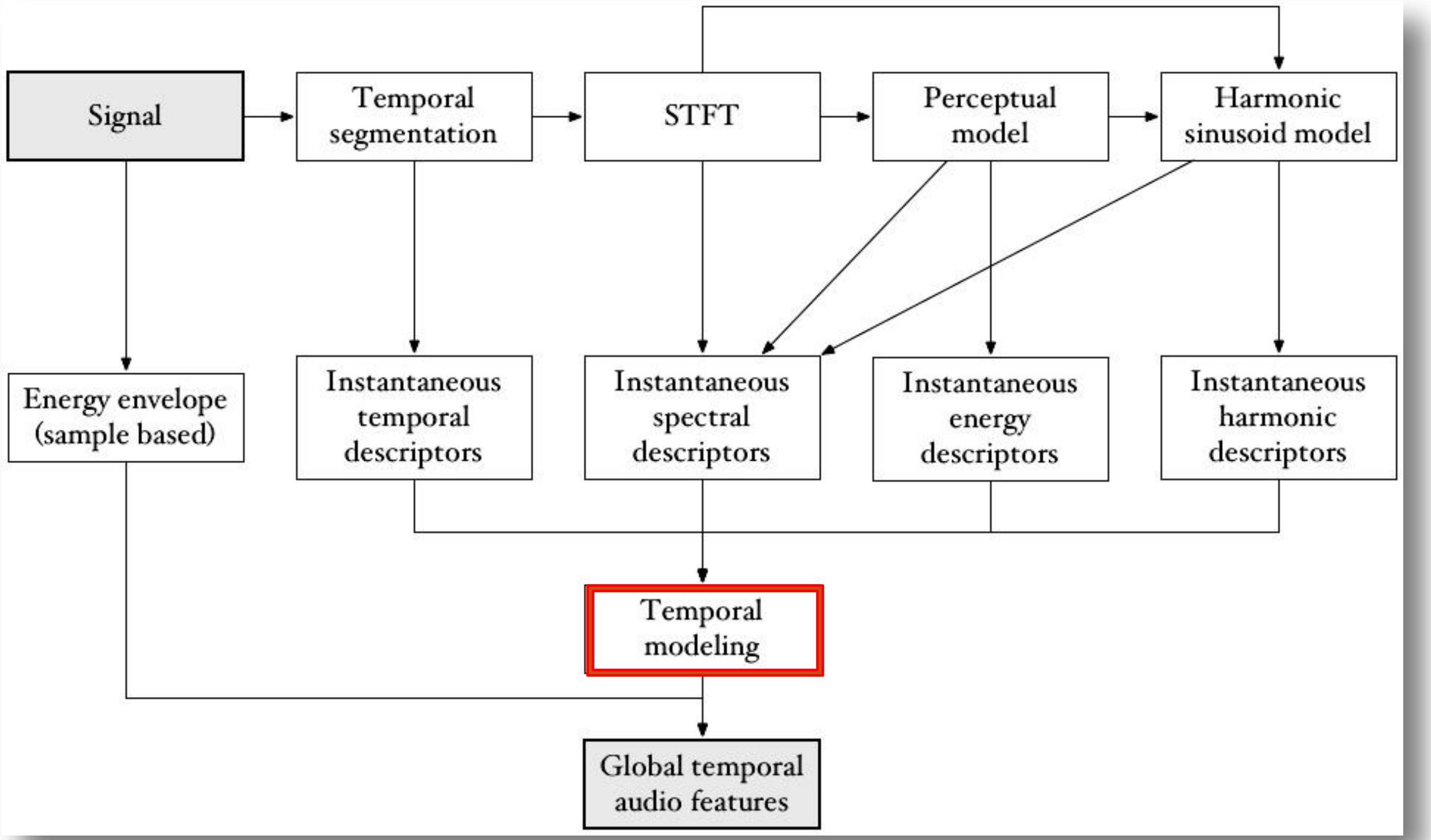
AUDIO FEATURES EXTRACTION

As one may listen to the same sound from different perspectives, segregating the different components, the idea is to project this ability into a computerized sound analysis…

Signal modelling →

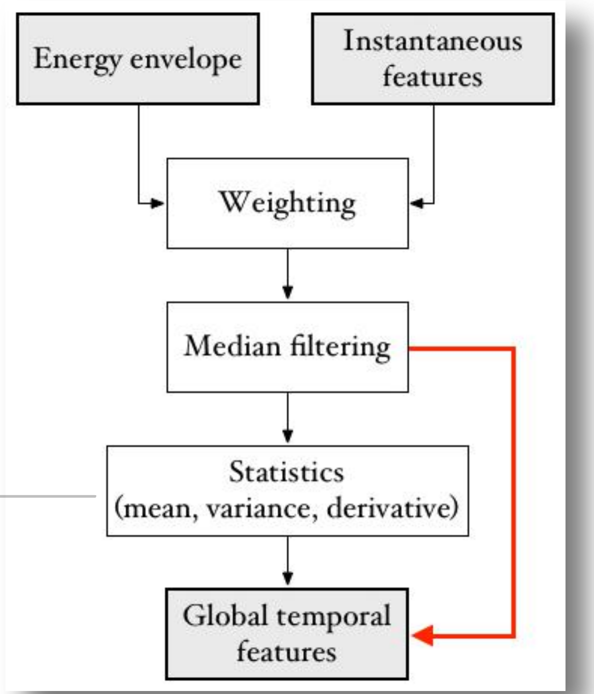
Low-level descriptors →

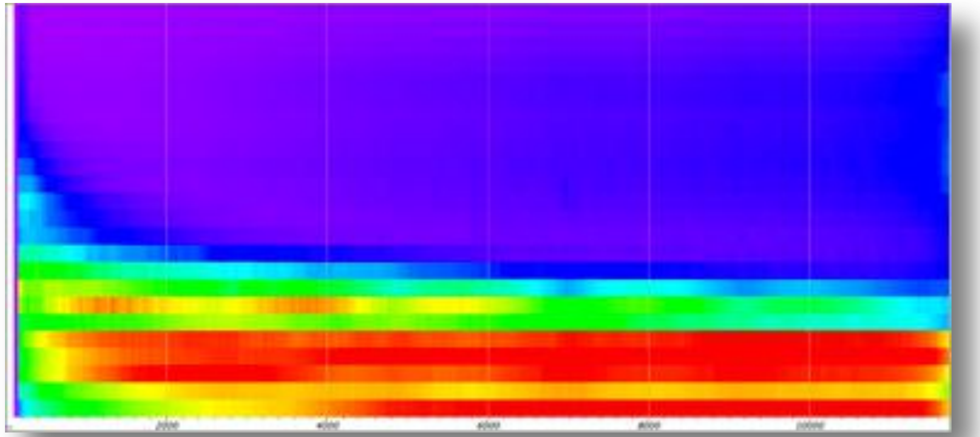
Audio features →



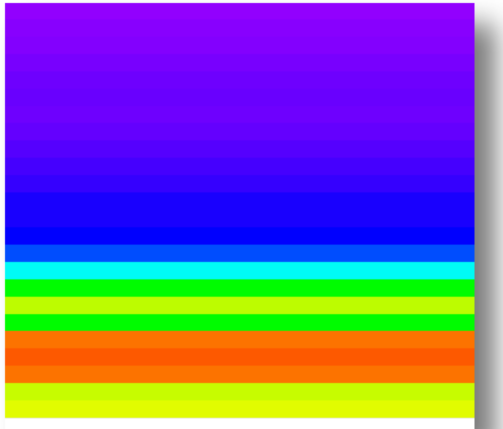
TEMPORAL MODELLING

- Global temporal model [Peeters, 2004]



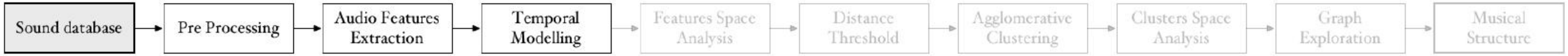


Instantaneous RSL



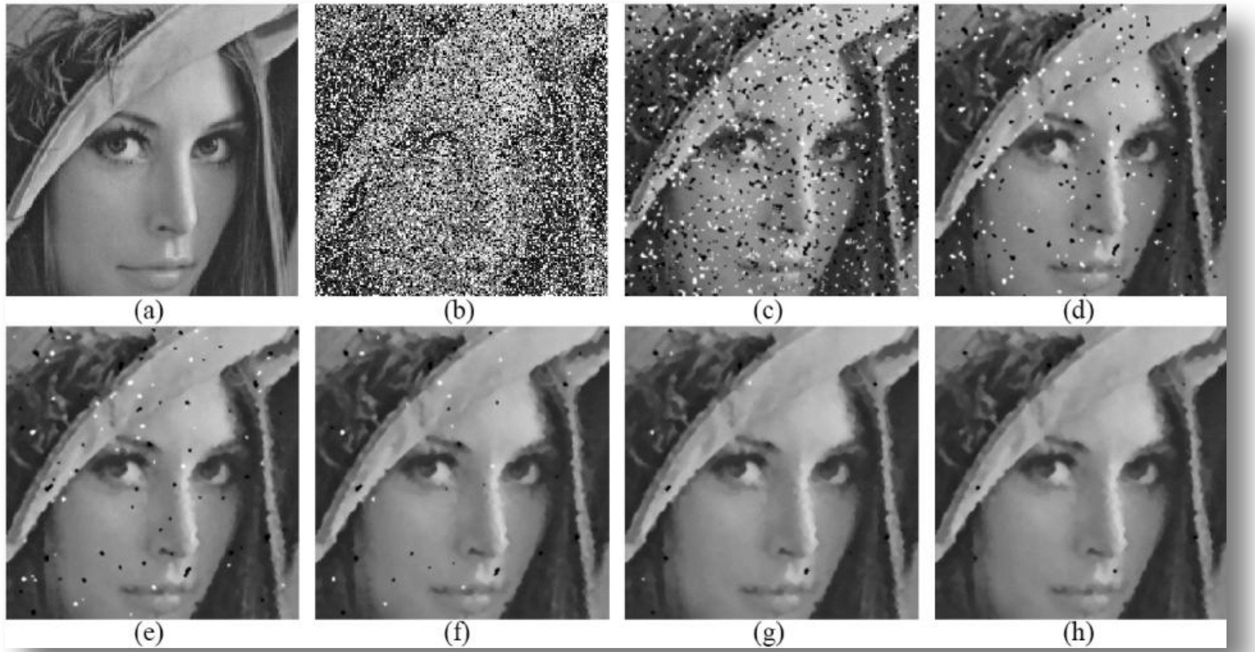
Global RSL (means/Bark bands)

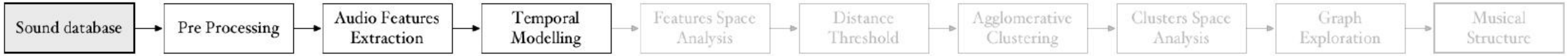
Potential loss of information…



MEDIAN FILTERING

- Nonlinear digital filtering technique [Huang, 1979]
- **Noise reduction on a signal**
- Sliding window
- Stream, $x = 5$
- Median

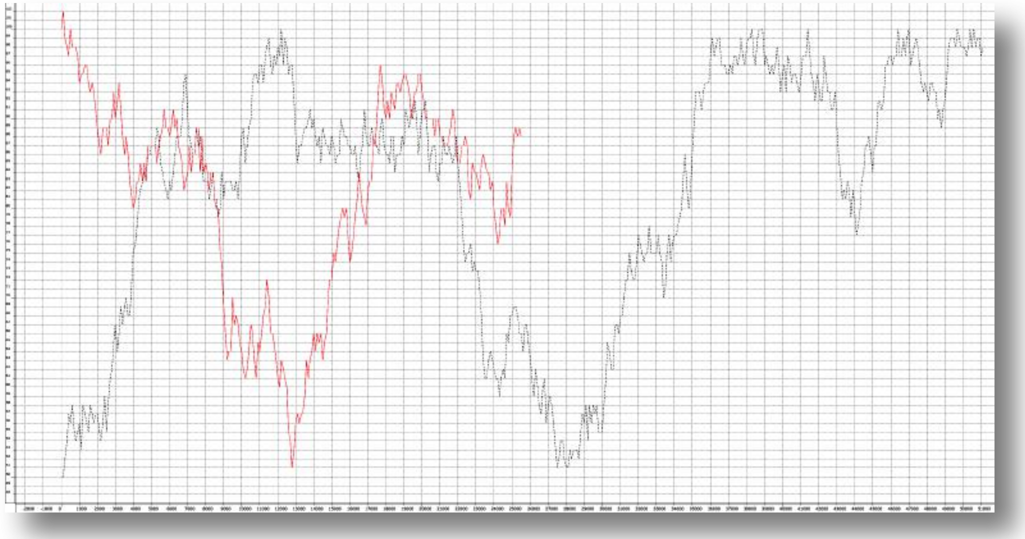




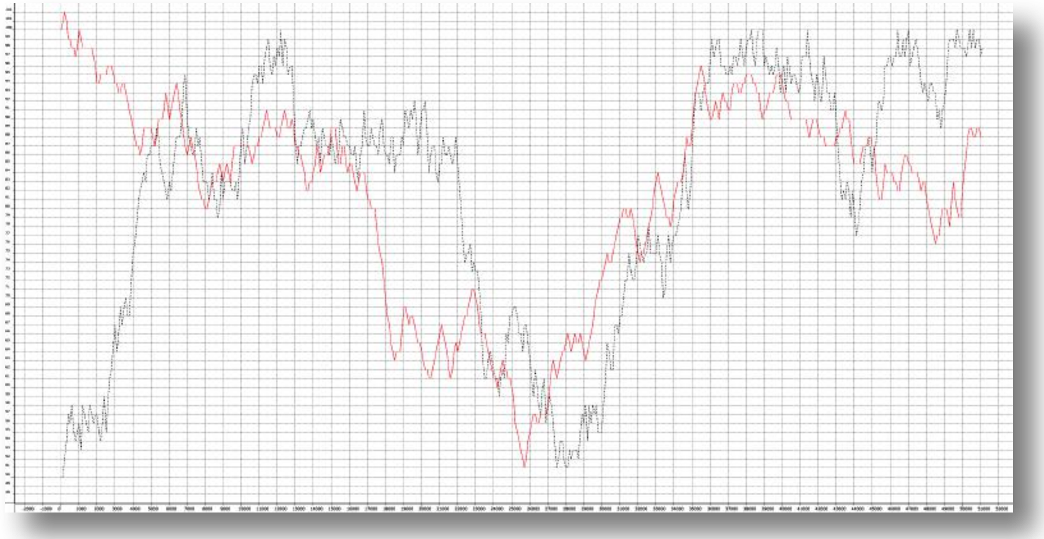
TEMPORAL ALIGNMENT

Here again, the idea is to mimic the selective listening skill, or the listening attitude in the time domain...

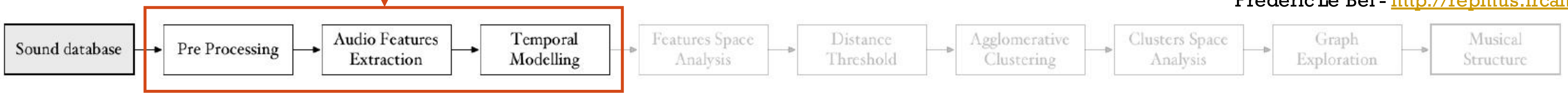
- Different assumptions [Lebel, 2016]: instantaneous OR **global** OR dynamic OR ??



Original lengths audio features



Resampled lengths audio features

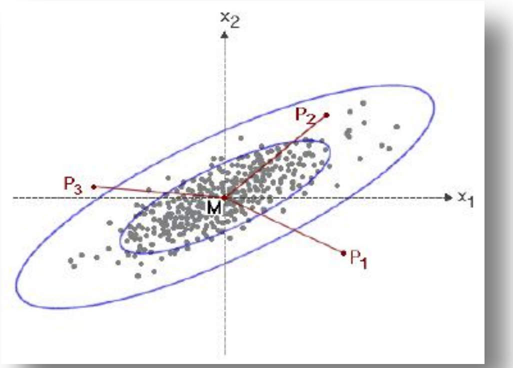


IN OTHER WORDS…

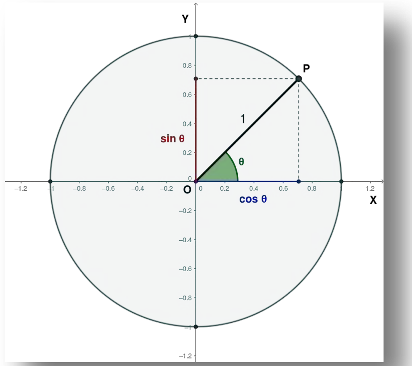
- This part of the framework focuses on
  - moulding the digital data through the perceptual data
  - in order to obtain clustering results that are consistent to the listeners…

LEVELS OF SIMILARITY

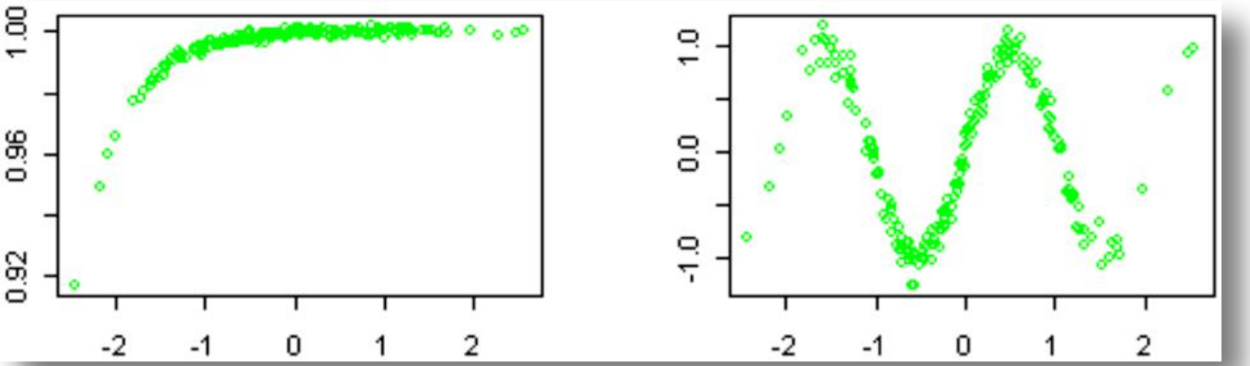
- **Distances** (magnitude)
  - Mahalanobis distance [0. 1.]
  - *Euclidean distance [0. +inf.]*
- **Similarity** (orientation)
  - |Cosine similarity| [0. 1.]
  - *Jaccard index [0. 1.]*
- **Correlation** (dependency)
  - |Spearman coefficient| [0. 1.]
  - *|Pearson coefficient| [0. 1.]*



Mahalanobis [Hazewinkel, 2002]



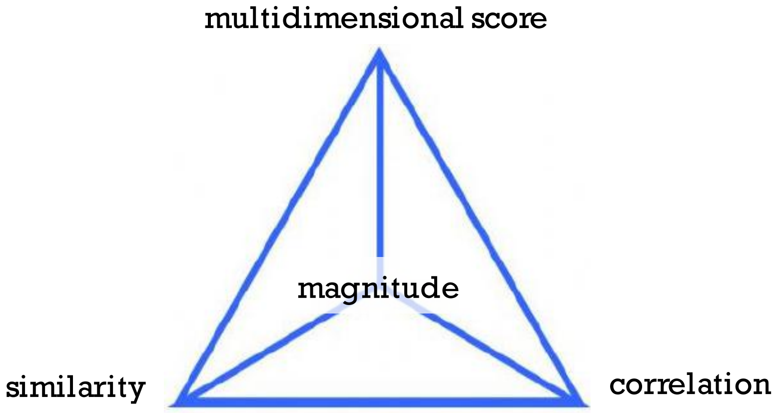
Cosine [Singhal, 2001]

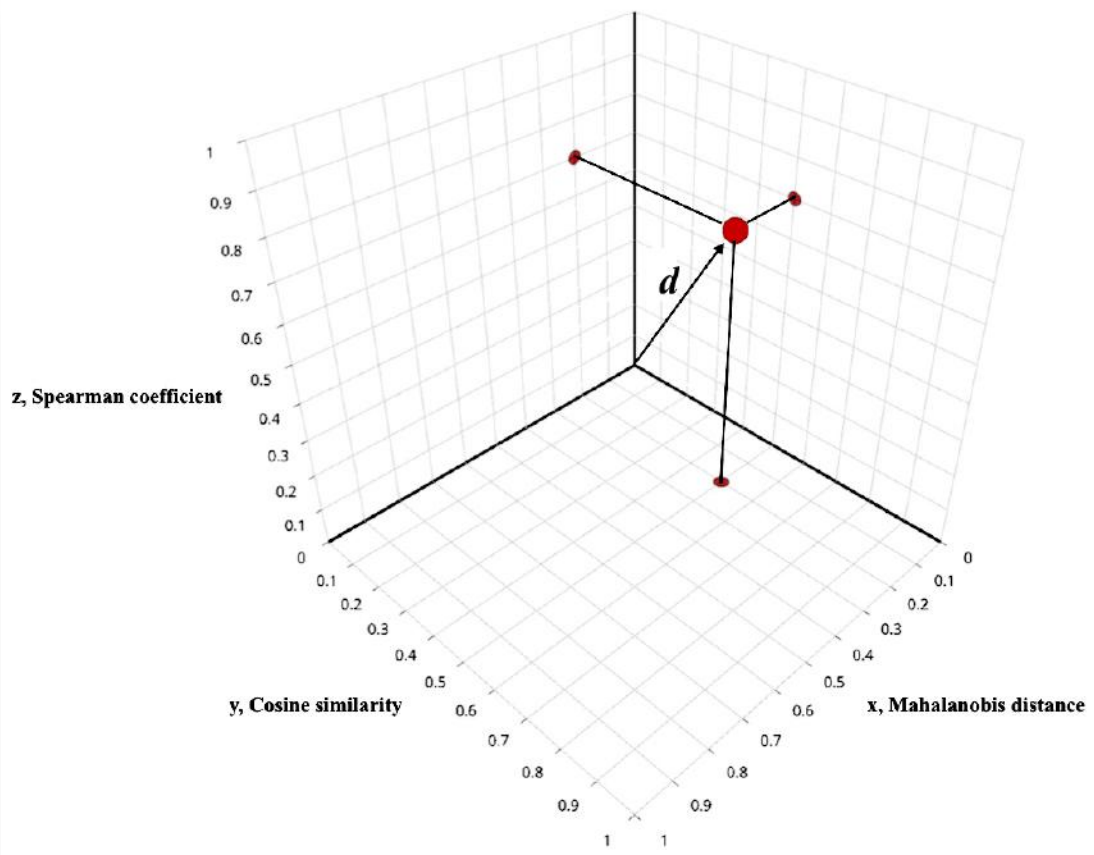


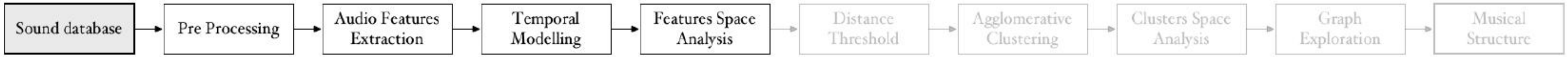
Spearman [Rakotomalala, 2015]

DISTANCE TRIANGULATION

- Single multidimensional score
- Including 3 **different perspectives**
- Higher level description
- Inverted and normalized values! [0. 1.]
- Shape of space/clusters [Lebel, 2016]

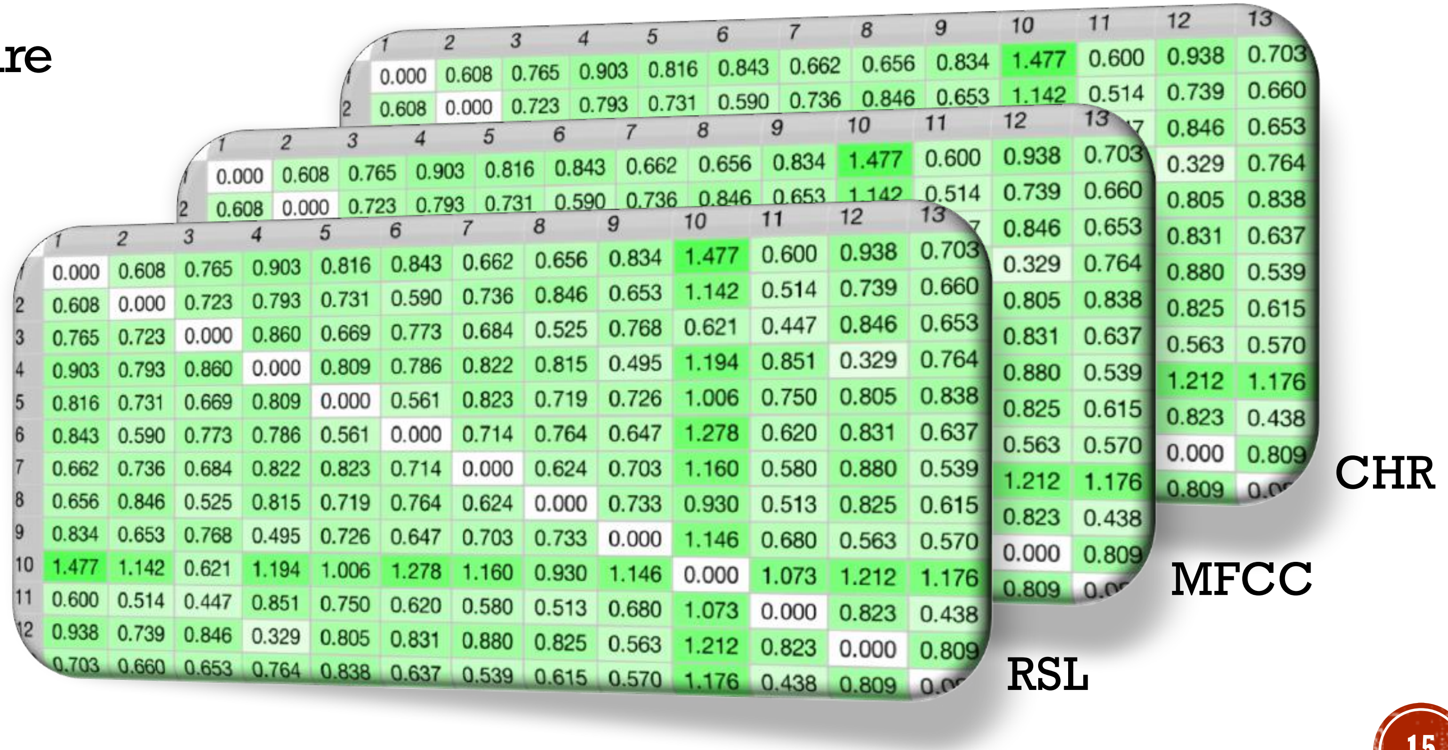


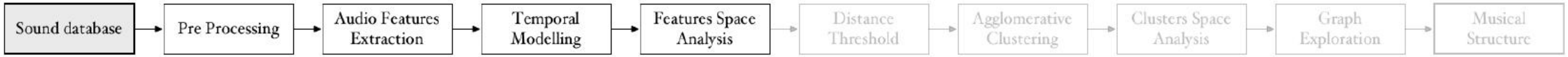




DISTANCE MATRICES

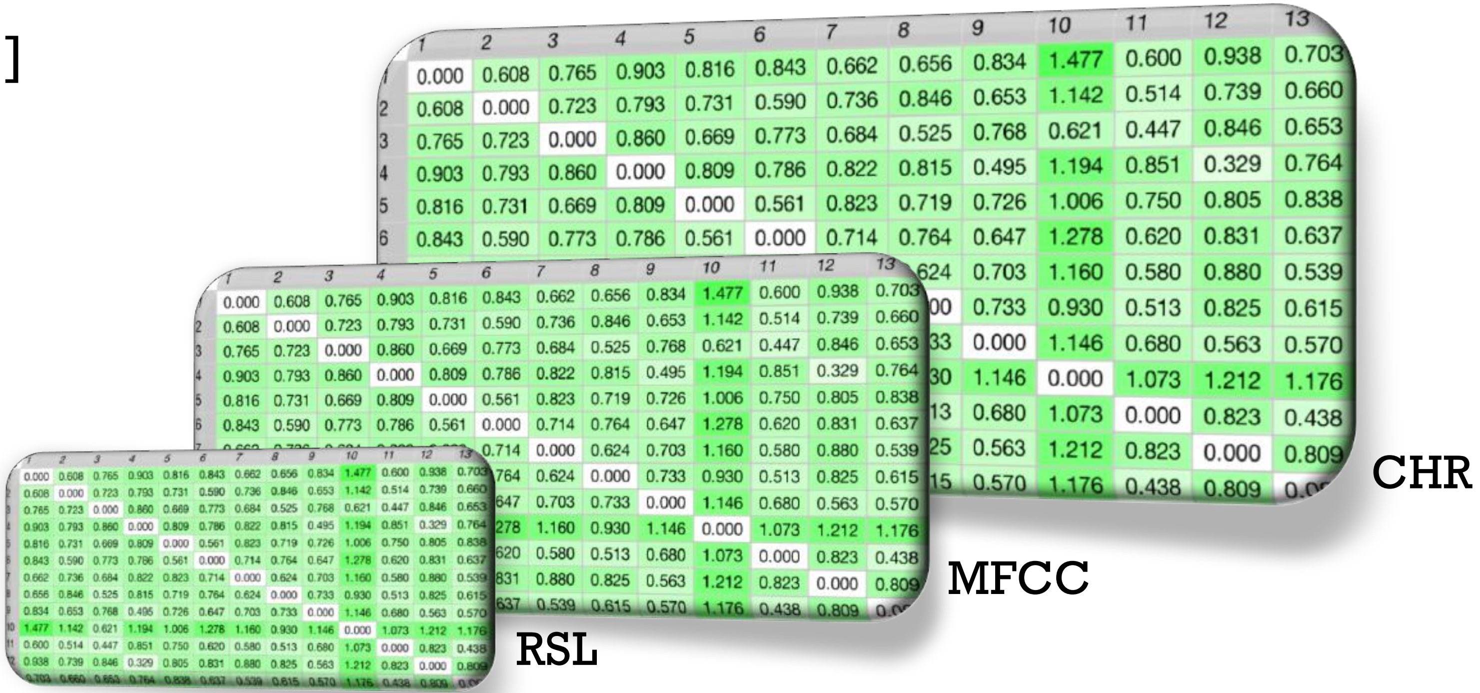
- 2D containers of distances taken pairwise [Gentle, 2007]
- **Quantifies all the connections**
- Each matrix = One audio feature
- *Thus, n* features = *n* matrices

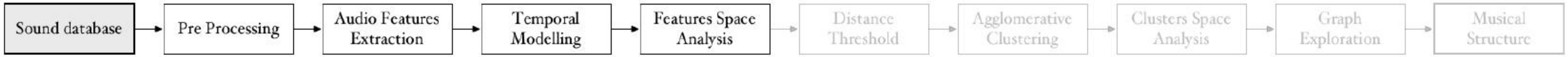




WEIGHTED MATRICES

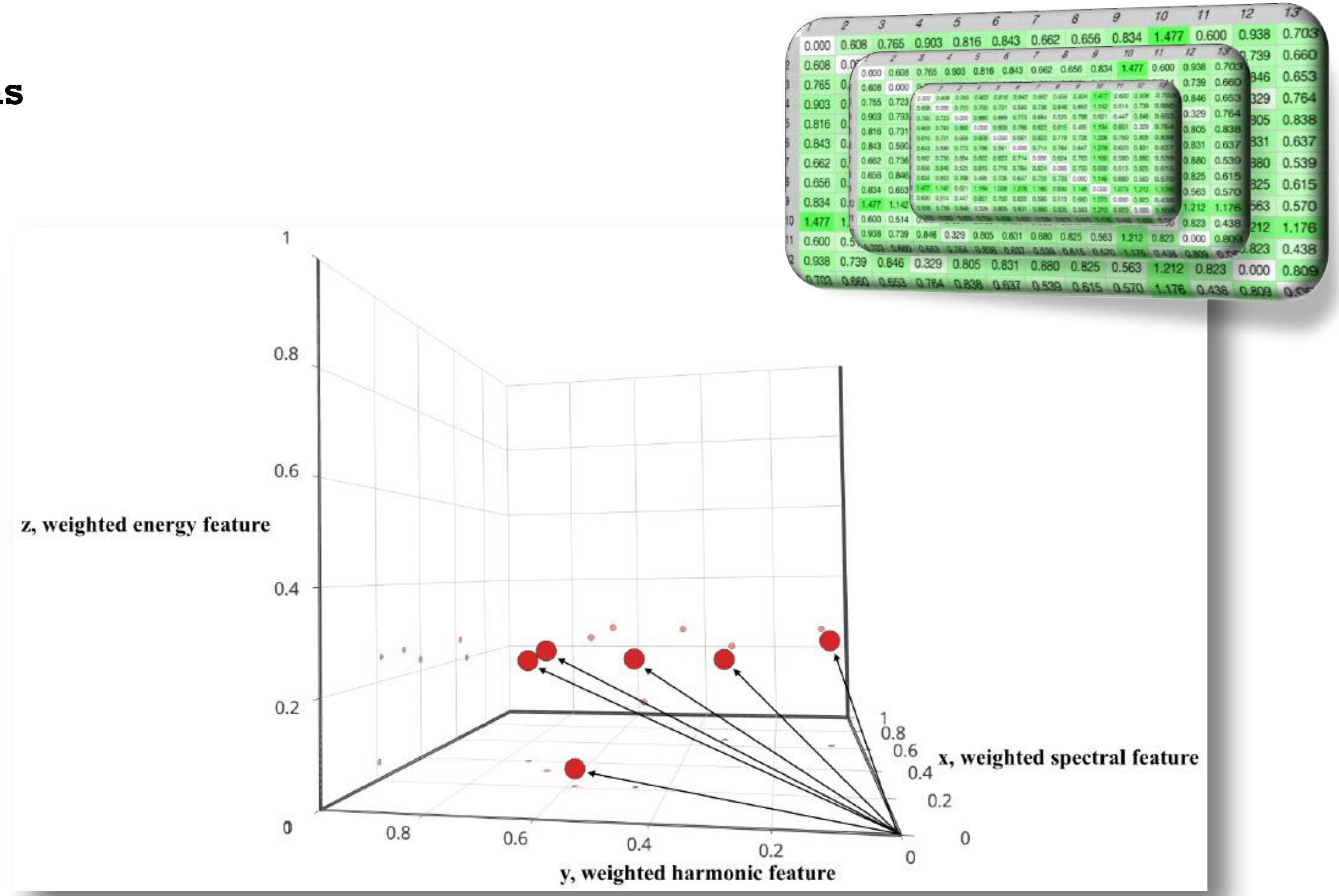
- Different features
- **Different levels of importance**
- Different weight factors [0. 1.]
- Simple multiplication: $x_i * w_i$
- User defined…

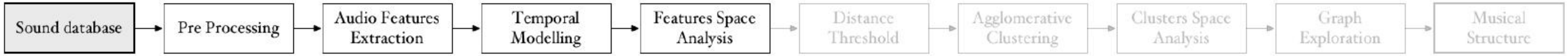




DIMENSIONALITY REDUCTION

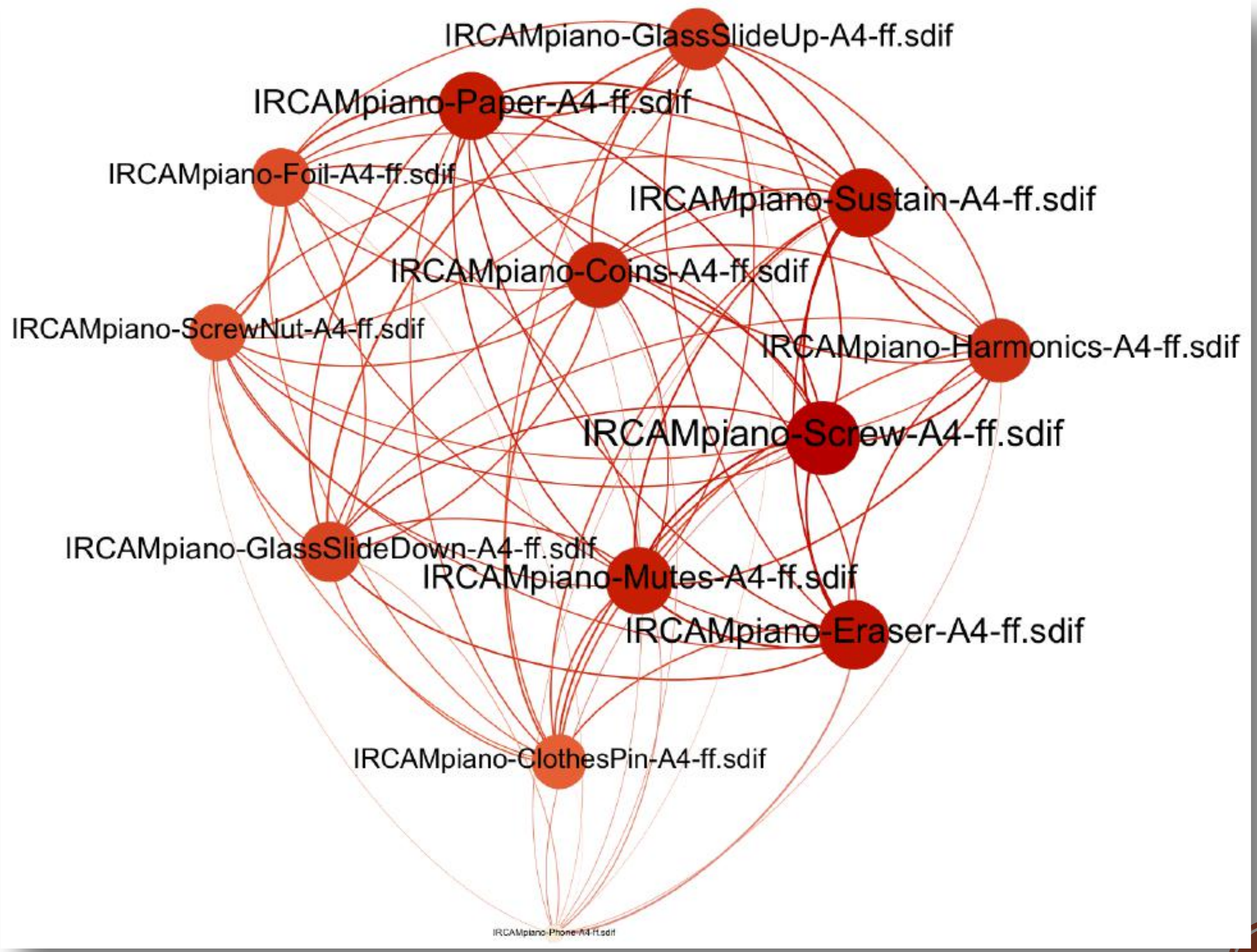
- ***n* features = *n* matrices = *n* dimensions**
- Need to simplify the data…
- …reduce the number of variables
- **P**rincipal **C**omponent **A**nalysis?
  - Features selection [Roweis, 2000]
  - Features extraction [Pudil, 1998]
- **No** because,
  - Low number of features… (< 100)
    - Curse of dimensionality [Bellman, 1957]
  - **Better interpretation of the output!**
- **Then** how,
  - Similar to distance triangulation…
  - Matrices triangulation!
- ***n* dimensions down to a single one**
  - Without any data transformation!



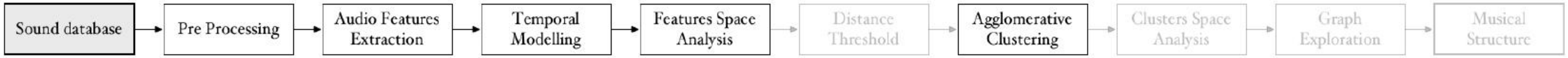


FEATURES SPACE NETWORK

- Multidimensional scaling [MDSJ, 2009]
  - 17D (audio features) → 2D (graph)
- Average weighted degrees
  - Big dark nodes = high degree
  - Small clear nodes = low degree
  - Centrality/Eccentricity
- **Just a representation…**
  - **Not a mathematical structure!**

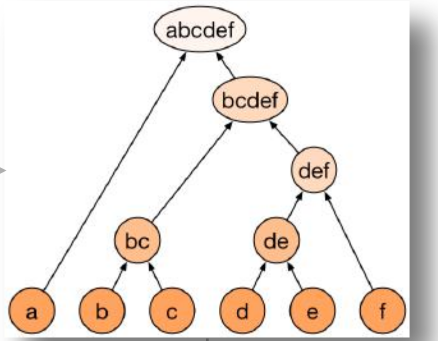


[Gephi, 2009]



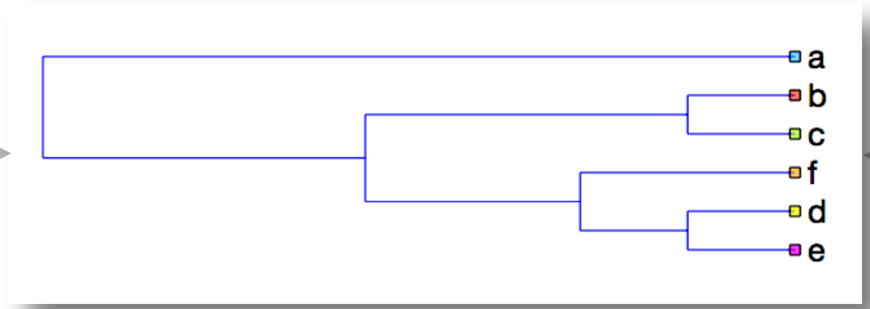
AGGLOMERATIVE CLUSTERING

- **H**ierarchical **C**luster **A**nalysis [Defays, 1977]
  - **Unsupervised learning** method (unlabelled data)
  - Seeks to build a hierarchy of clusters (bottom-up)
  - Based on **distances** + **linkage criterion** (*NN* algorithm)
  - Dendrogram (Greek: dendro = tree, gramma = drawing)
  - **Drawbacks**:
    1. consistency within and between clusters,
    2. tie-break problem… *(pairwise approach)*

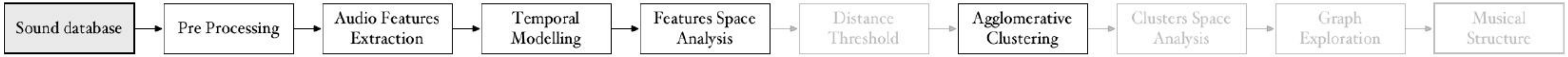


	a	b	c	d	e	f
a	0.000	0.700	0.600	0.500	0.400	0.300
b	0.700	0.000	0.100	0.200	0.300	0.400
c	0.600	0.100	0.000	0.100	0.200	0.300
d	0.500	0.200	0.100	0.000	0.100	0.200
e	0.400	0.300	0.200	0.100	0.000	0.100
f	0.300	0.400	0.300	0.200	0.100	0.000

Distance matrix

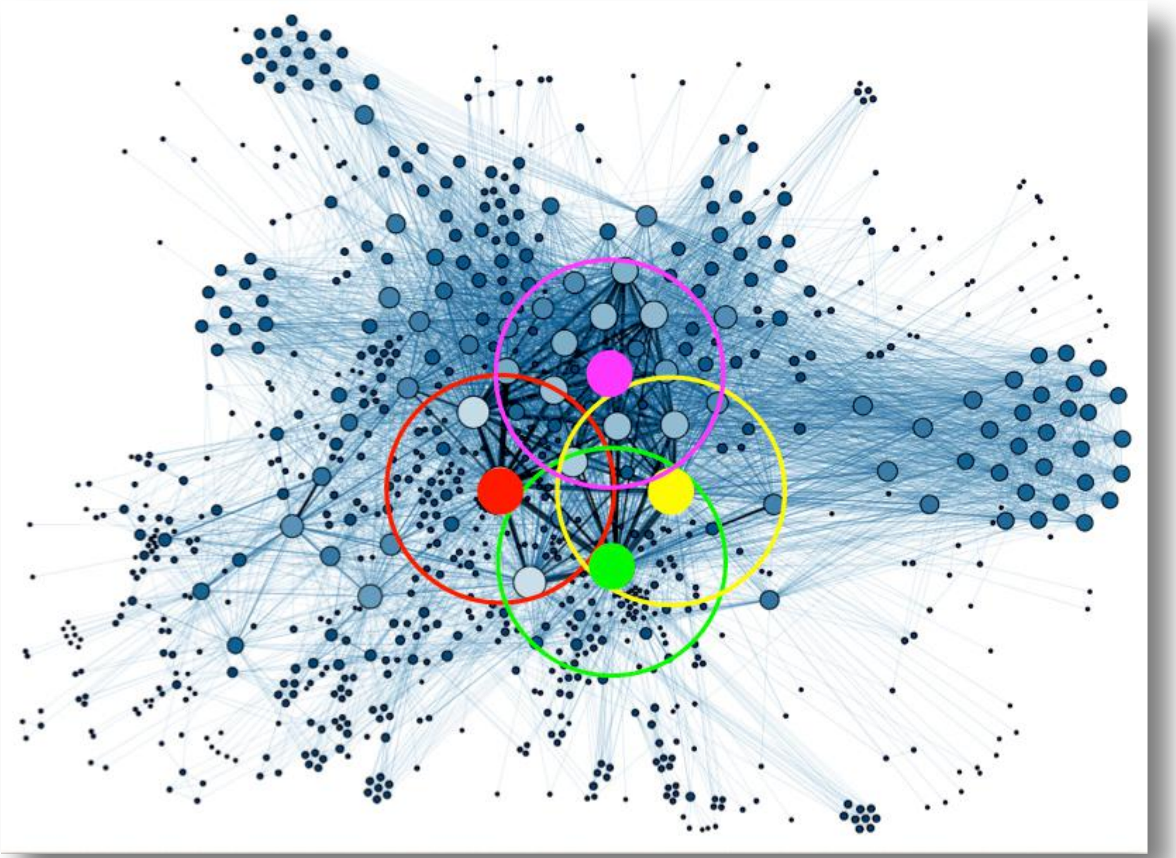


Dendrogram [Orange, 2013]



VARIATION ON HCA

- Instead of nearest neighbour…
- **Threshold constraint** (C_t)
  - Threshold = max distance (diametrical)
  - Kind of perceptual witness…
- If *d(x, y)* $\leq$ *C_t = true, then agglomerate,*
- else, *create new cluster.*
- Speed is traded for accuracy…
  - Global optimum!
  - Complexity = O(n!)…
- **Solves previous drawbacks**:
  1. consistency = isomorphic clusters,
  2. tie-break = overlapping clusters.

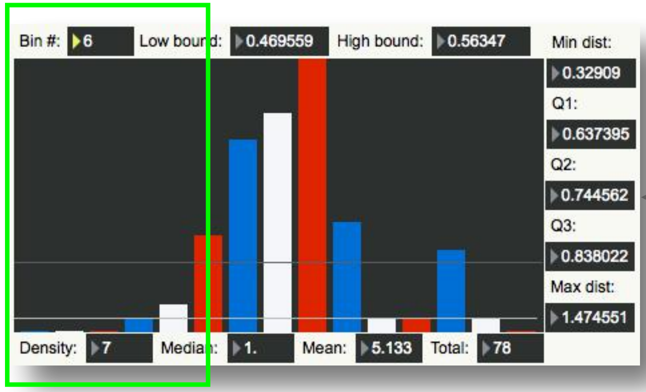


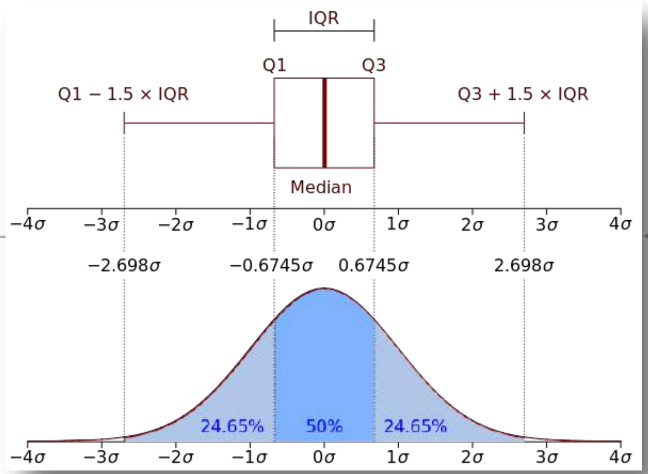
DISTANCE THRESHOLD

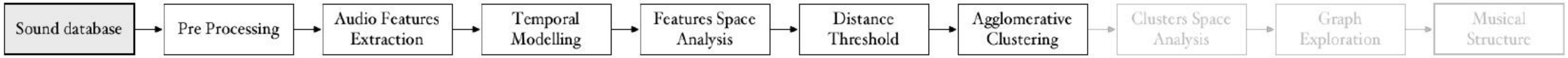
- **Descriptive statistics** [Mann, 2006]
  - Central tendency: median, mean and mode
  - Dispersion: extrema, standard deviation, kurtosis and skewness
  - Distribution: histogram and stem-and-leaf display
- **Histogram** (sparsity of *x*)
  - Freedman-Dicaonis rule [Freedman, 1981]
  - **Adaptive bin width** *(h)*
  - For number of bins *(k)*

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil$$

$$k = \left\lceil \frac{\max x - \min x}{2 \frac{\mathrm{IQR}(x)}{\sqrt[3]{n}}} \right\rceil$$

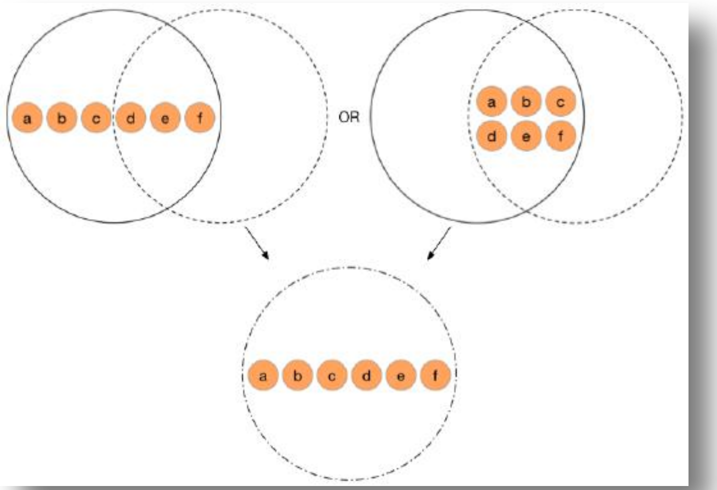




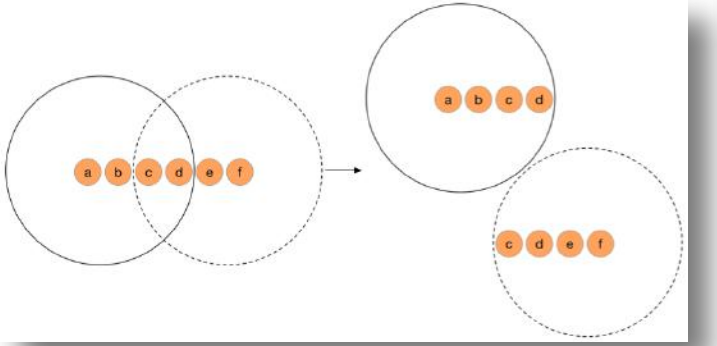


THE OUTCOME

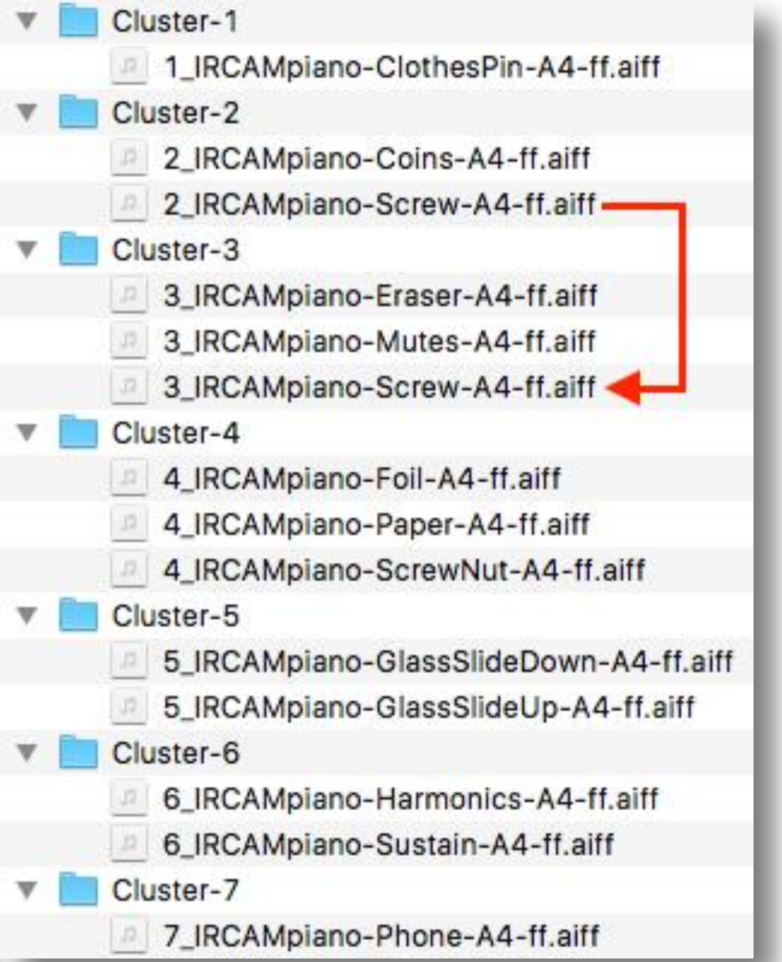
- **More accurate but more complex…**
  - Overlapping clusters = blured hierarchy
    - One component may belong to multiple clusters
    - More difficult to distinguish clusters
    - Fuzzy clustering…
- **Need to reduce the number of variables!**
  - Sub-clusters merging
  - Overlapping clusters split
- *Translated into musical terms [Huron, 2001]*
  - *Overlaps = voice-leading patterns*
  - *voice-leading patterns = timbral patterns*

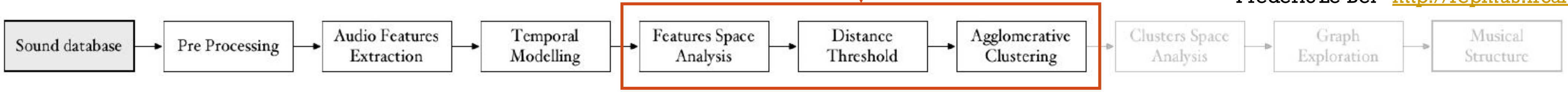


Sub-clusters merging



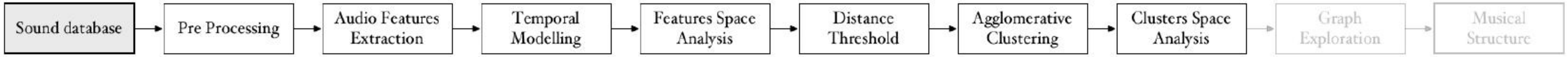
Overlapping clusters split





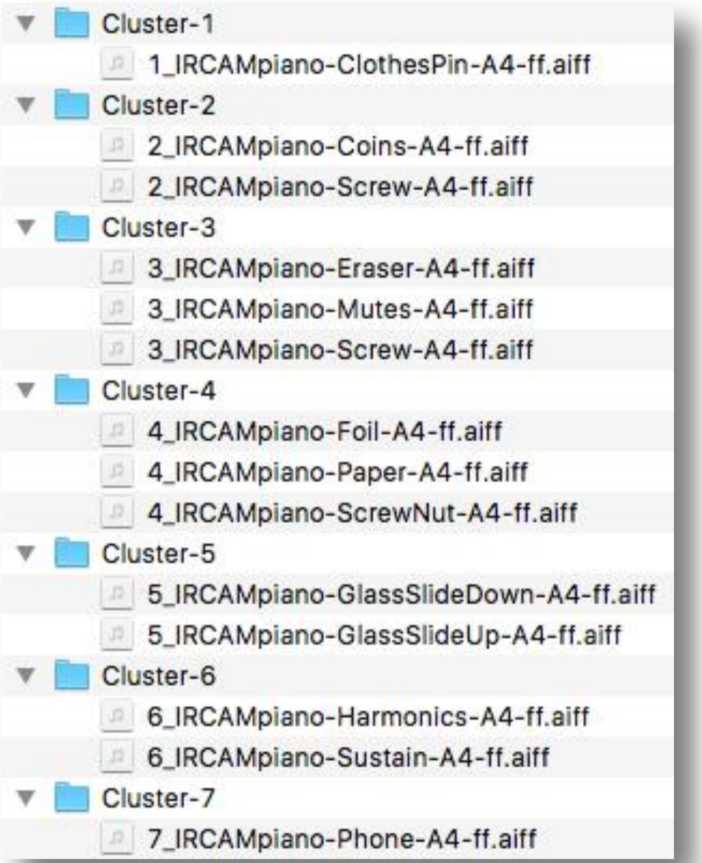
IN OTHER WORDS…

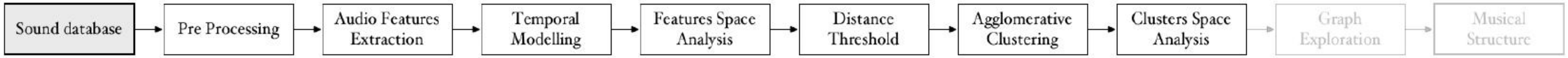
- This part of the framework strives to
  - computerize the way one would intuitively aggregate sounds by proximity, or similarity,
  - including the possibility of sounds to be part of multiple subpopulations at once.



CLUSTERS SPACE ANALYSIS

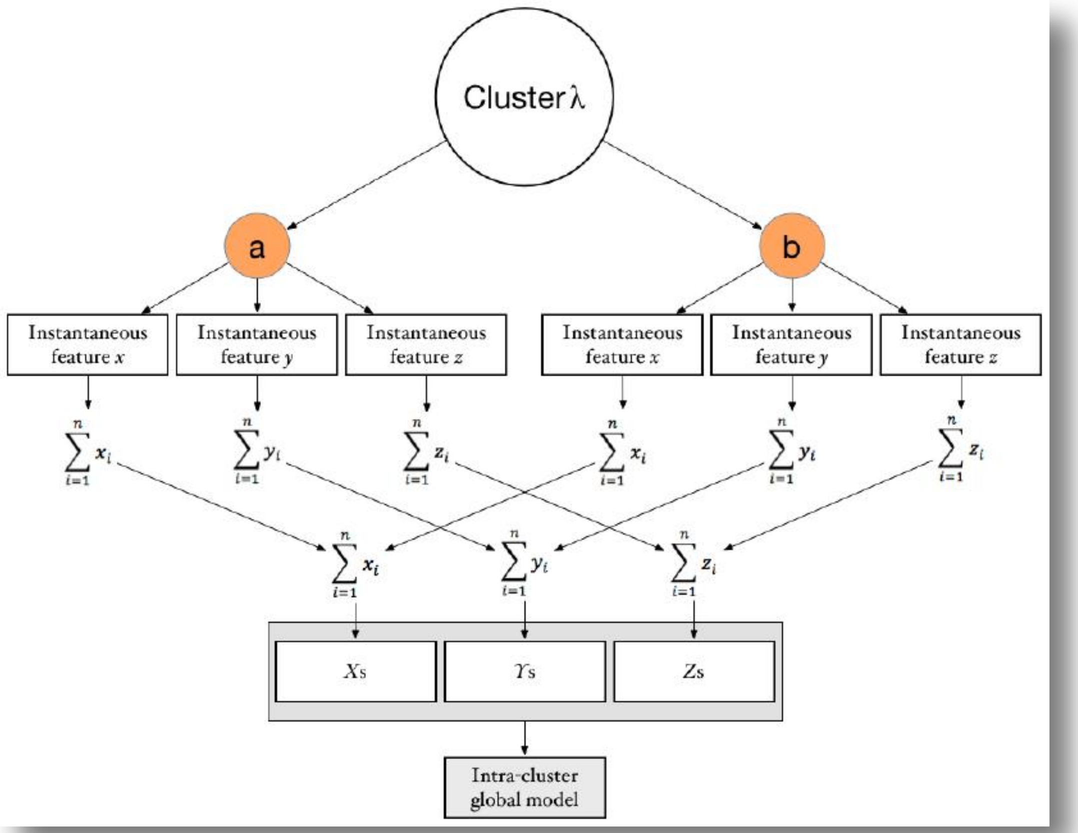
- To gain **more insight** on the resulting space
  - visualize the clustered space
    - looking at a down sampled dataset
    - using a larger bin width histogram
- Better understanding of the **core structure**
  1. Intra-cluster modelling
  2. Inter-cluster analysis



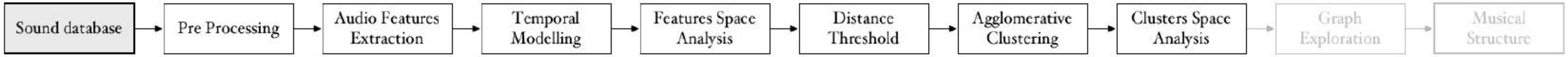


INTRA-CLUSTER ANALYSIS

- Similar to temporal modelling
  - **Generalizing audio features of each cluster**
    - Find the theoretical centre of mass (barycentre)
- Global model = multidimensional vector
  - **Each data point = sum of the sums of each audio features**
  - Summation and accumulation of audio features…
- In order to measure the distance between them…



Intra-cluster modelling



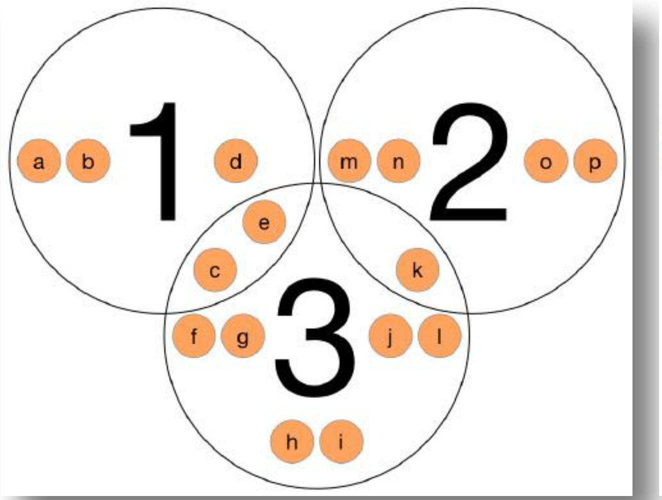
INTER-CLUSTER ANALYSIS

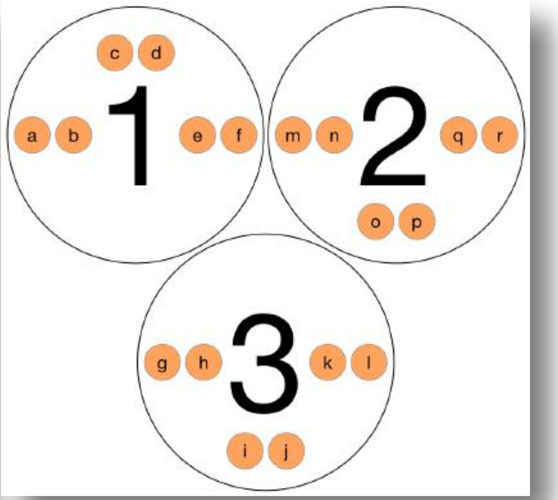
- Measuring the distance between each **clusters barycentre**
  - **Euclidean** distance [0. +inf.] (magnitude) * **Jaccard** index [0. 1.] (similarity)
    - single multidimensional score, including 2 different perspectives, higher level information…
- Gathered in a **distance matrix** to outline the resulting network…

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard index

[Jaccard, 1901]

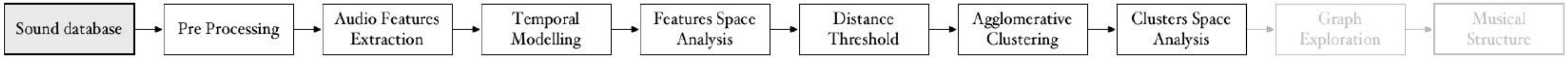




$$d(\vec{\lambda},\vec{\delta}) = \sqrt{\sum_{i=1}^{n}(\lambda_i - \delta_i)^2}$$

Euclidean distance

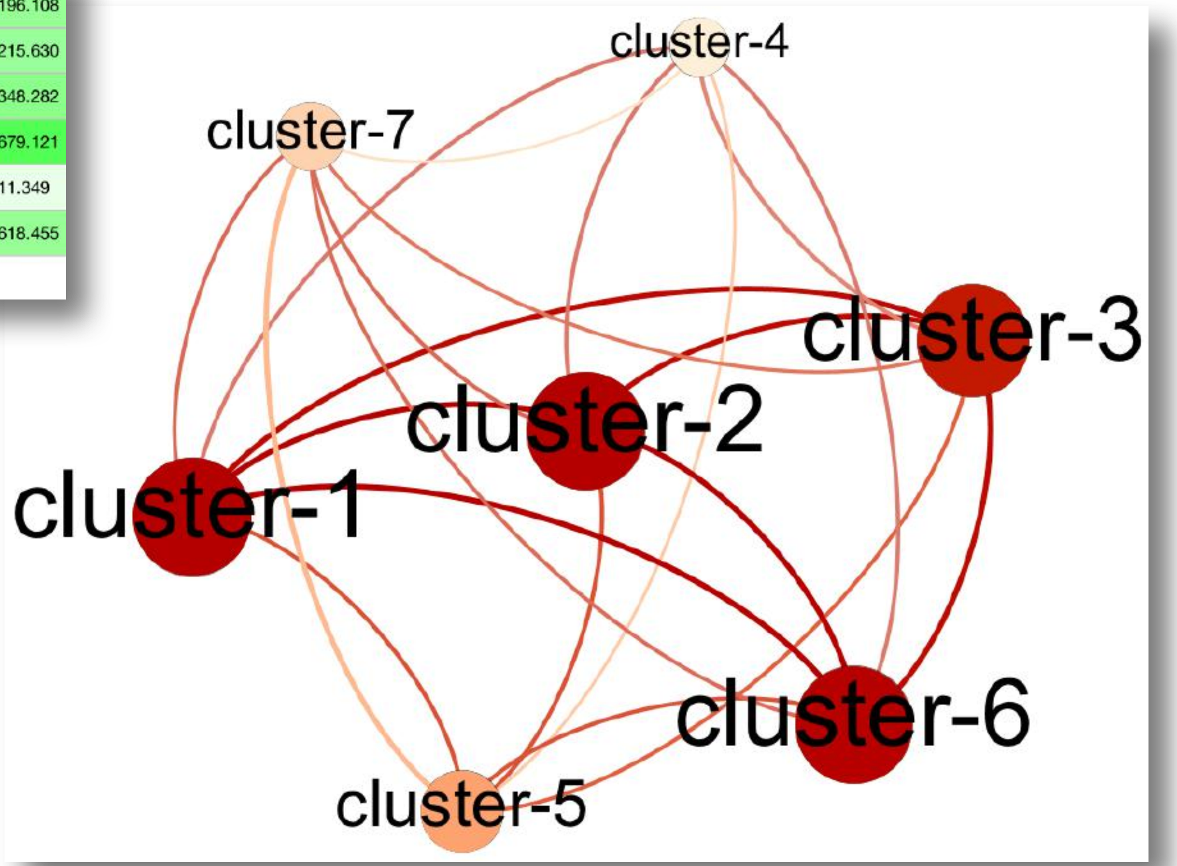
[Verley, 1997]



CLUSTERS SPACE NETWORK

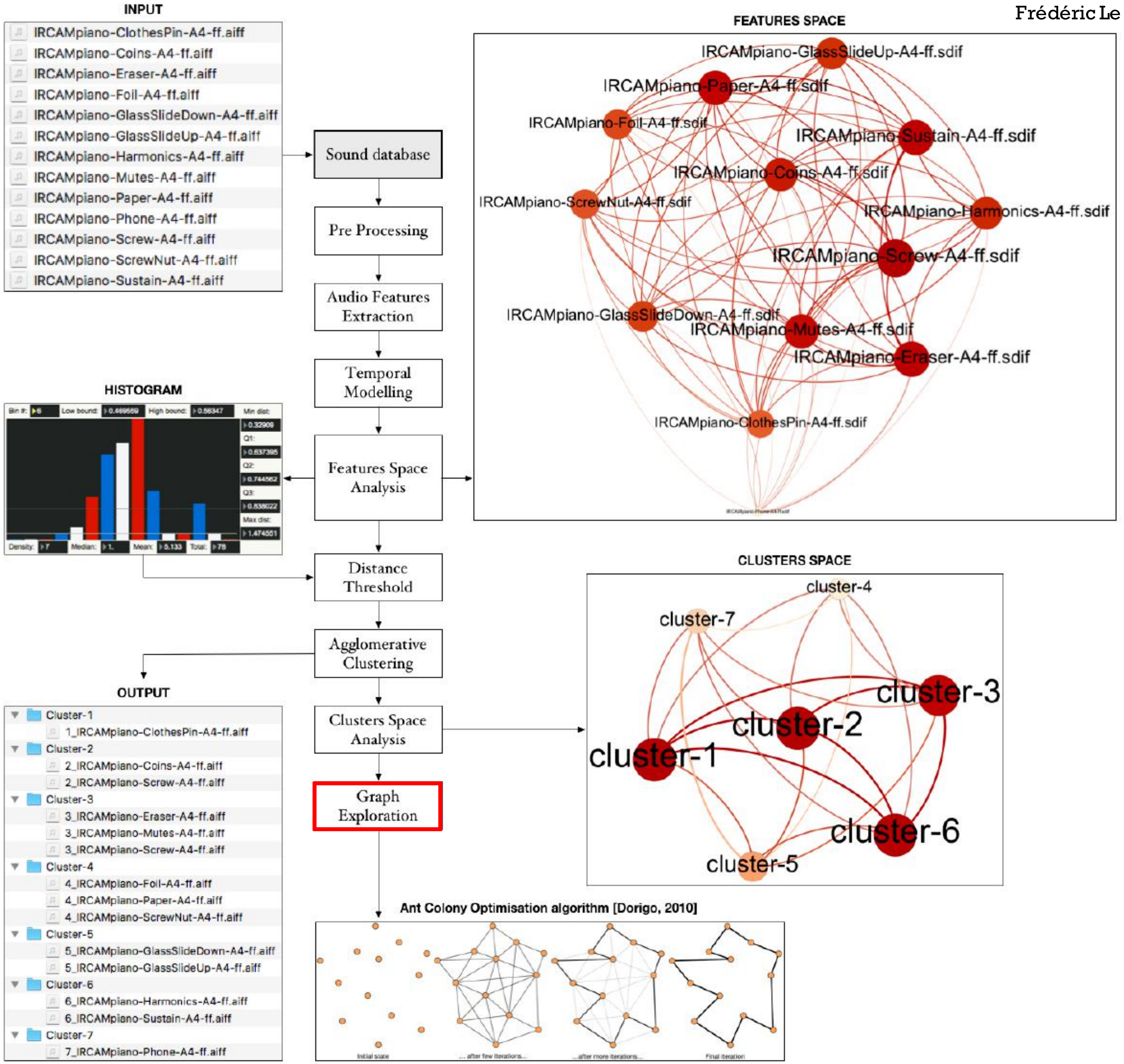
	"cluster-1"	"cluster-2"	"cluster-3"	"cluster-4"	"cluster-5"	"cluster-6"	"cluster-7"
"cluster-1"	0.000	24886673.802	38900706.651	297179876.332	286879316.637	10532180.324	339574196.108
"cluster-2"	24886673.802	0.000	56060666.279	282491866.688	274240613.812	35408076.384	330371215.630
"cluster-3"	38900706.651	56060666.279	0.000	285941699.637	325771919.827	35125811.841	378153348.282
"cluster-4"	297179876.332	282491866.688	285941699.637	0.000	476063442.497	304210404.150	546808679.121
"cluster-5"	286879316.637	274240613.812	325771919.827	476063442.497	0.000	292239273.422	72010411.349
"cluster-6"	10532180.324	35408076.384	35125811.841	304210404.150	292239273.422	0.000	343433618.455
"cluster-7"	339574196.108	330371215.630	378153348.282	546808679.121	72010411.349	343433618.455	0.000

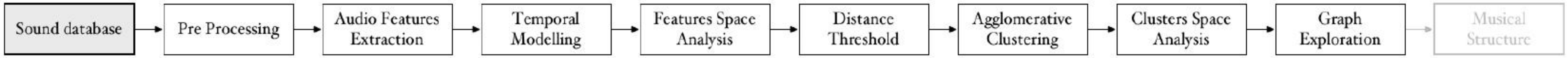
- Multidimensional scaling…
- Average weighted degrees
  - Big dark nodes = high degree
  - Small clear nodes = low degree
  - Centrality/Eccentricity
- **Just a representation…**
  - **Not a mathematical structure!**



Frédéric Le Bel - http://repmus.ircam.fr/lebel - 2016.

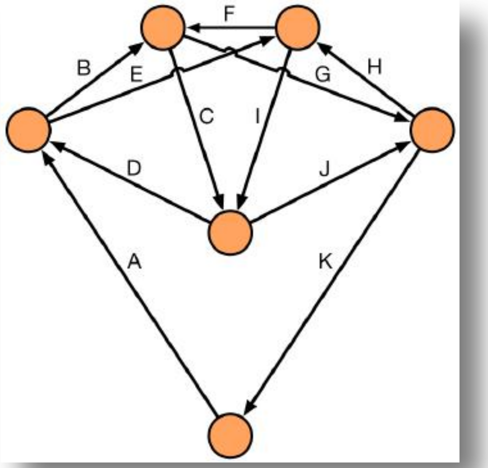
RECAP



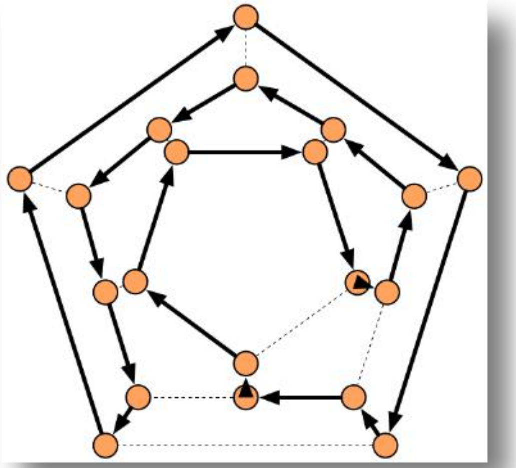


GRAPH THEORY (NOT EVEN AN OVERVIEW…)

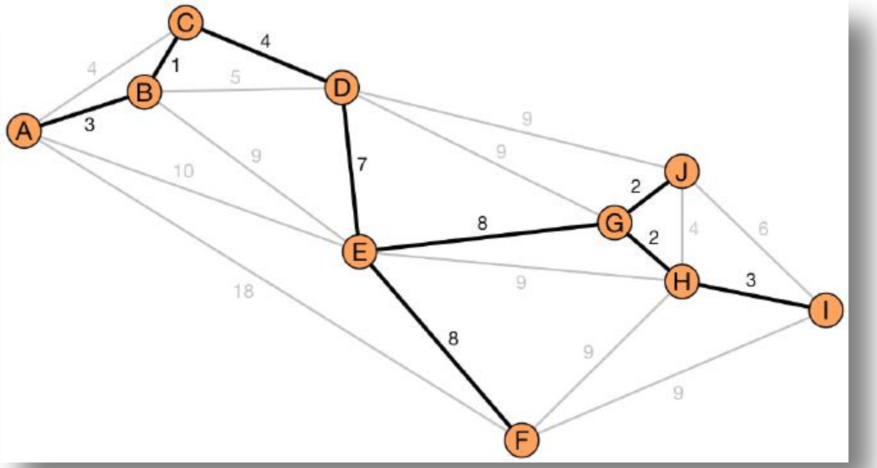
- Clearly suggests **ways of exploring the clusters space**!
- Eulerian cycles: visits each **edge** exactly once (start/end points = same)
- Hamiltonian cycles: visits each **node** exactly once (start/end points = same)
- Spanning trees: sets of **edges covering all nodes** (no path nor cycle)



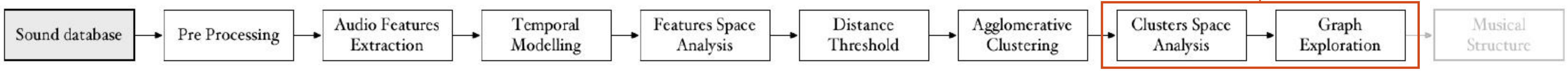
Eulerian cycle
[Euler, 1736]



Hamiltonian cycle
[Hamilton, 1857]

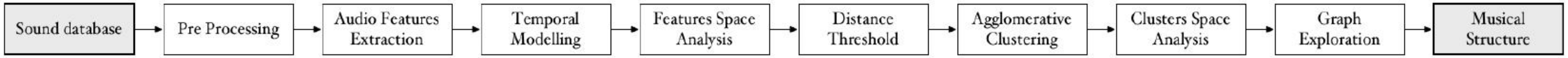


Minimum spanning tree
[Graham, 1985]



IN OTHER WORDS…

- This part of the framework aspires to
  - provide more insight on the resulting clusters space and to
  - find ways of navigating though it
  - towards formalizing music.

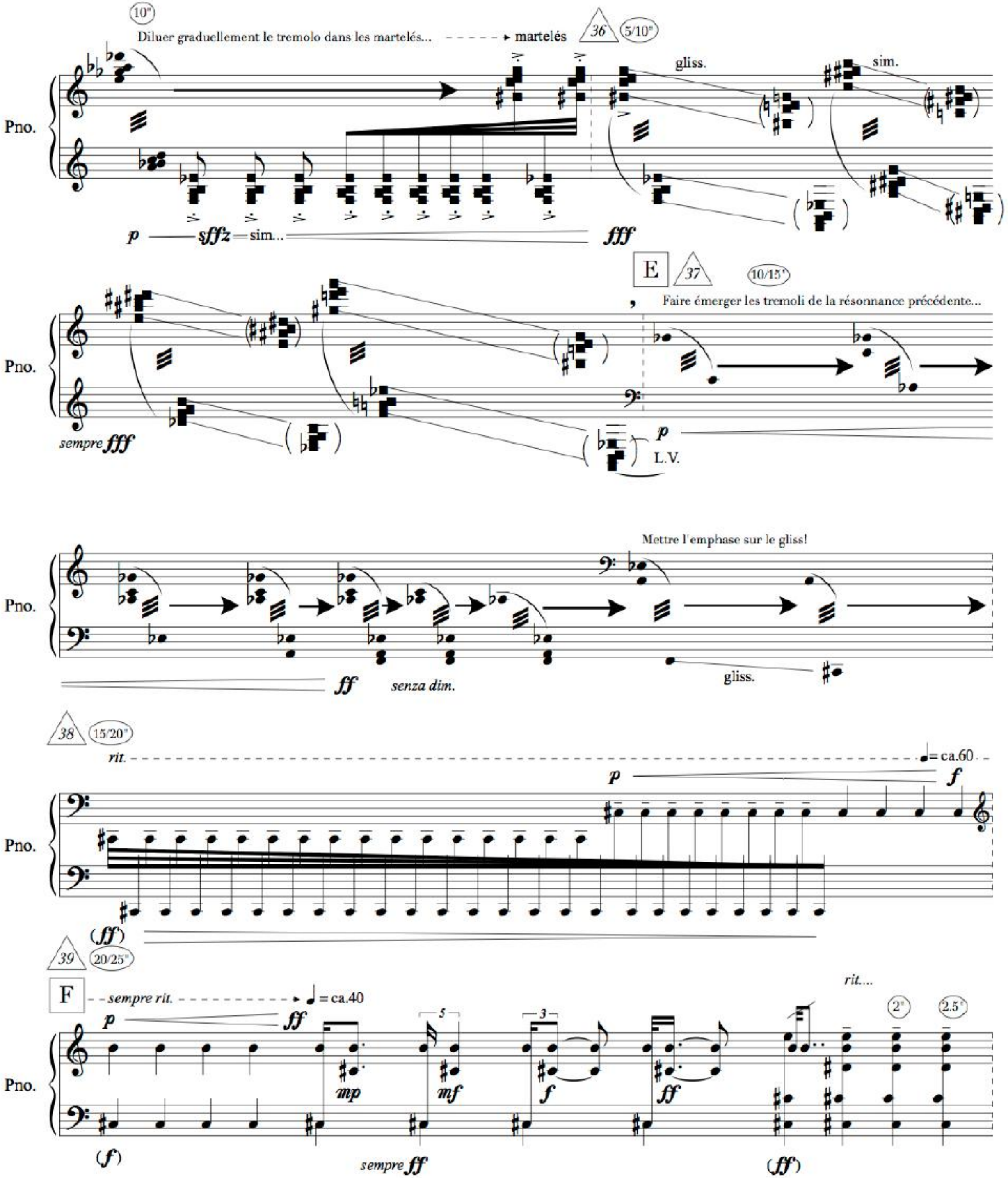


FROM SCIENCE TO MUSIC

- **Due to time constraint**…
  - could not make exhaustive use of the clustering process (nor the graph search algorithms)
- But many **concepts were translated**
  - to nourish musical ideas,
  - compositional techniques and
  - various scenarios of interaction.
- The idea was
  - not to mimic the algorithm itself
  - but to **expand the notion of distance/similarity** on multiple layers (locally and globally)
  - in a way to create different scenarios of interaction
- Each scenario
  - intuitively targets **specific audio features** (energy, pitch, spectrum, density, **time, behaviour, rhythm**, etc.)
  - and projects them into various **shifting processes**
  - in order to create some kind of a **smooth and perpetual drifting motion** through the music.
- That is to reflect the idea of
  - exploring the **shortest path(s)** within a graph.

EXEMPLIFICATION

3:56 – 5:02

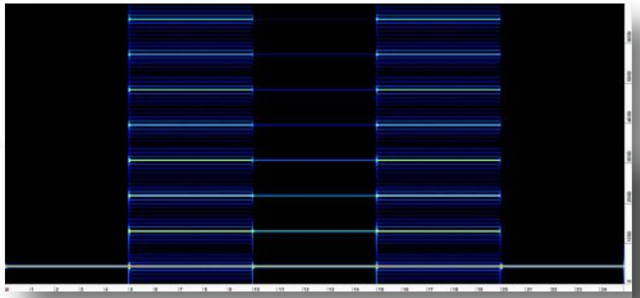


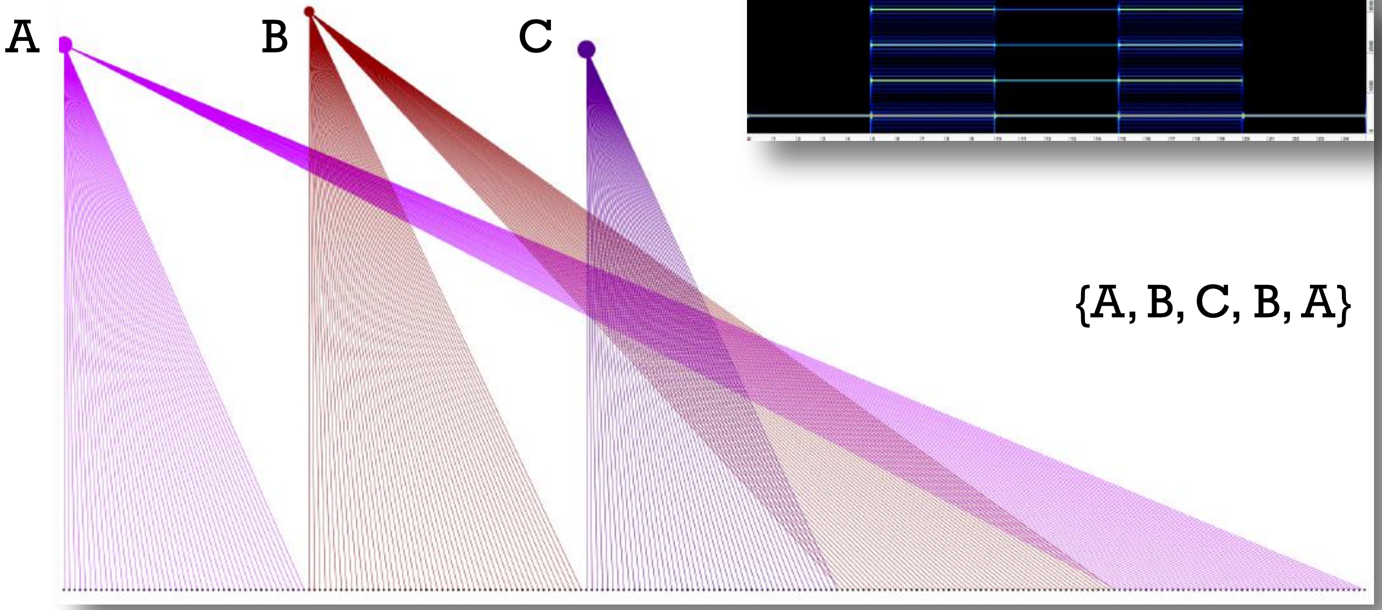
CONCLUSIONS…

- **Based on previous works** (MIR)
  - corpus-based concatenative synthesis [Schwarz, 2006]
  - musical genre recognition [Peeters, 2007]
  - computer-aided orchestration [Carpentier, 2008]
- **A different framework** for audio classification
  - with applications to **computer-aided composition** (graph theory)
  - And eventually for computational musicology…
- Contrary to its predecessors,
  - this framework is built **towards formalizing music**
  - rather than generating or labelling sound material
- In other words,
  - it is engineered to act on a **larger scale** than in the other cases
- Consequently,
  - it is designed to attempt rendering this level of **perspective through analysis and clustering**.

…PERSPECTIVES

- Developing **applications** for
  - **Computer-aided composition**
    - Graph search algorithms…
  - **Computational musicology**
    - Representations in the time domain…





Variation on k-medoids…

FOR MORE INFORMATION

- http://repmus.ircam.fr/lebel
  - Research report & audio examples
  - Score(s) & recording(s)

CLUSTERING EXAMPLES

PIANO	MULTIPHONICS	SOUND DESIGN	ABSTRACT
Cluster-x	Cluster-x	Cluster-x	Cluster-x
Cluster-y	Cluster-y	Cluster-y	Cluster-y
Cluster-z	Cluster-z	Cluster-z	Cluster-z