Berkeley
UNIVERSITY OF CALIFORNIA

CNMAT

CARMINE-EMANUELE CELLA

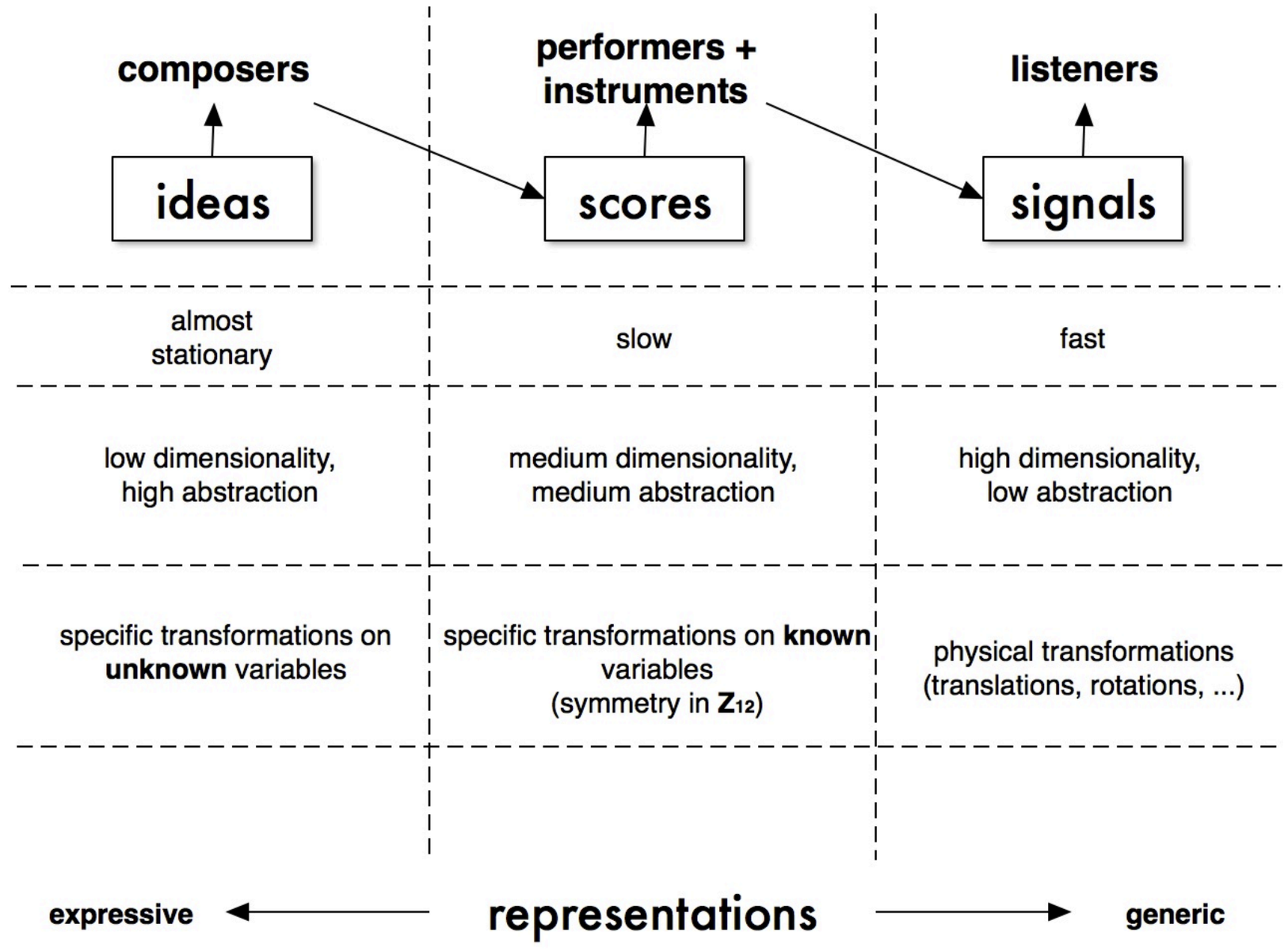
CONVOLUTIONAL REPRESENTATIONS AND SOUND-TYPES

MUSIC 159

THE OPEN PROBLEM (OBSESSION)

Which connections can we make between the symbolic space and the signal space?

THE OPEN PROBLEM



BETWEEN SYMBOLS AND SIGNALS: A FIRST APPROACH

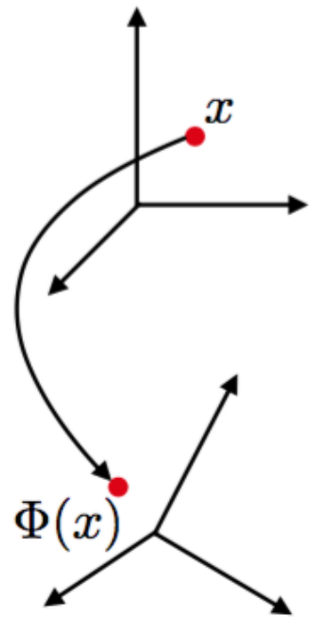
Symbolic space

(scores)

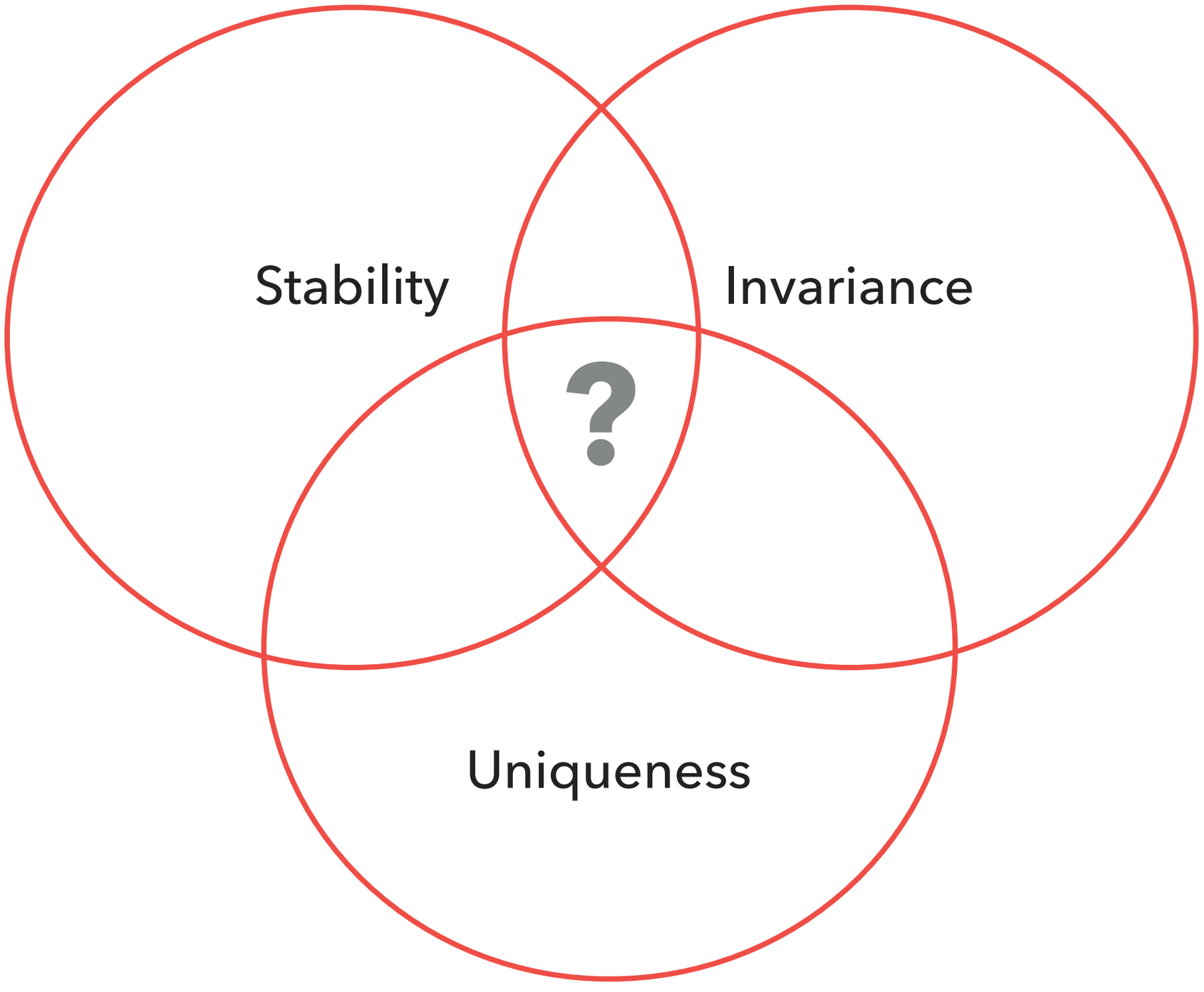
Signal space

(timbre)

PROJECTIVE REPRESENTATIONS



GOOD PROPERTIES FOR A REPRESENTATION



GOOD PROPERTIES FOR A REPRESENTATION

Let $\Phi(x)$ be a representation of x; it can have the following properties:

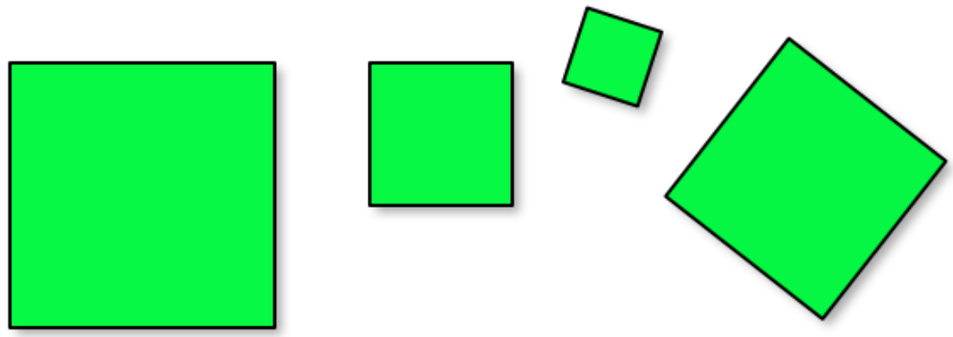
- **uniqueness**: $\Phi(x) \neq \Phi(y) \iff x \neq y$;
- **stability**: $\|\Phi(x) - \Phi(y)\|_2 \leq C\|x - y\|_2$;
- **invariance** (to group of transformations G): $\forall g \in G, \Phi(g.x) = \Phi(x)$;
- **reconstruction**: $y = \Phi(x) \iff \widetilde{x} = \widetilde{\Phi}^{-1}(y)$

Don't bother (too much) with equations...

TRANSFORMATIONS/DEFORMATIONS

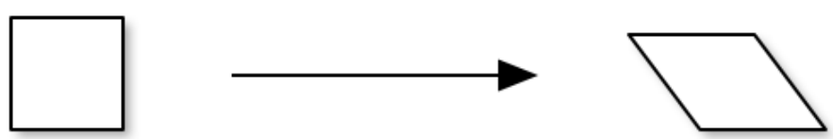
A good representation should be able to handle fundamental transformations.

translations, scaling, rotations

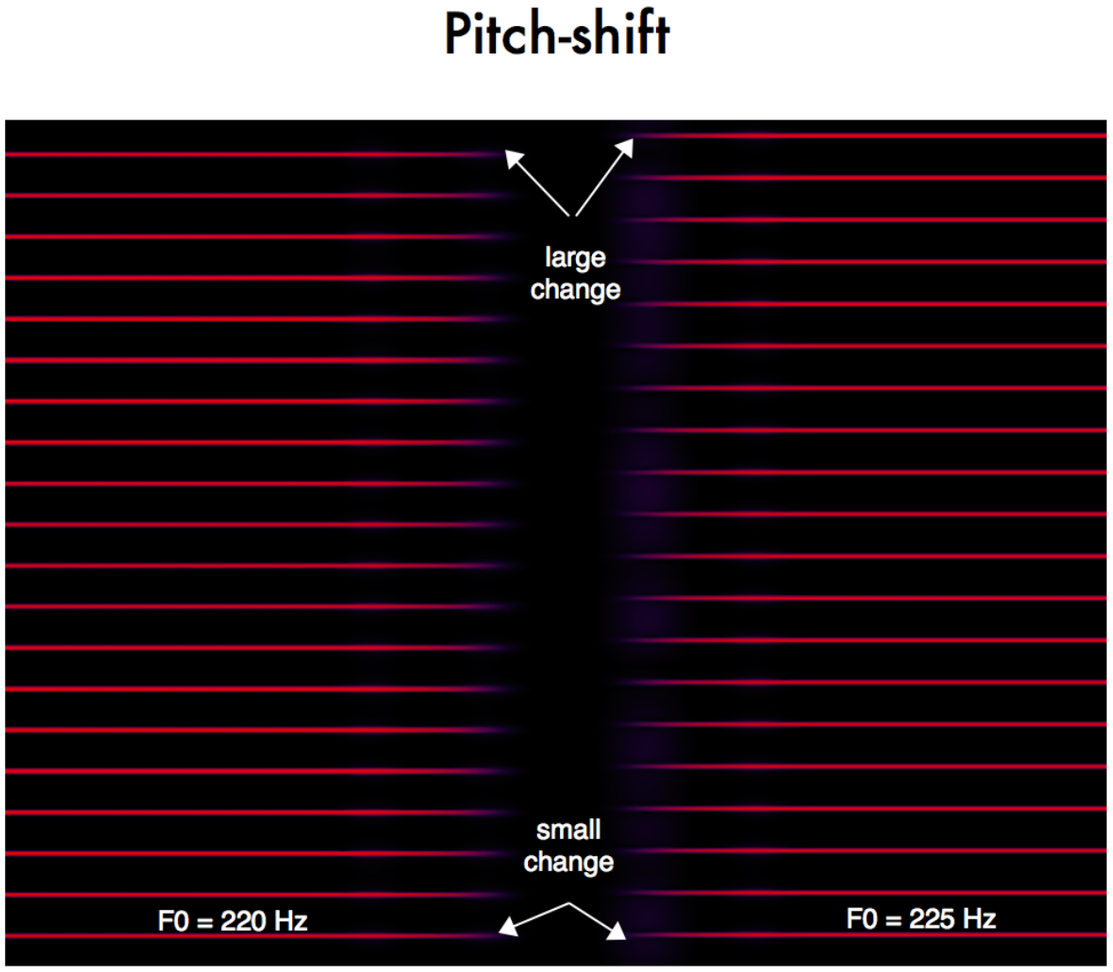


diffeomorphisms (small deformations)

$$x_\tau(u) = x(u - \tau(u)), \tau \in C^\infty$$



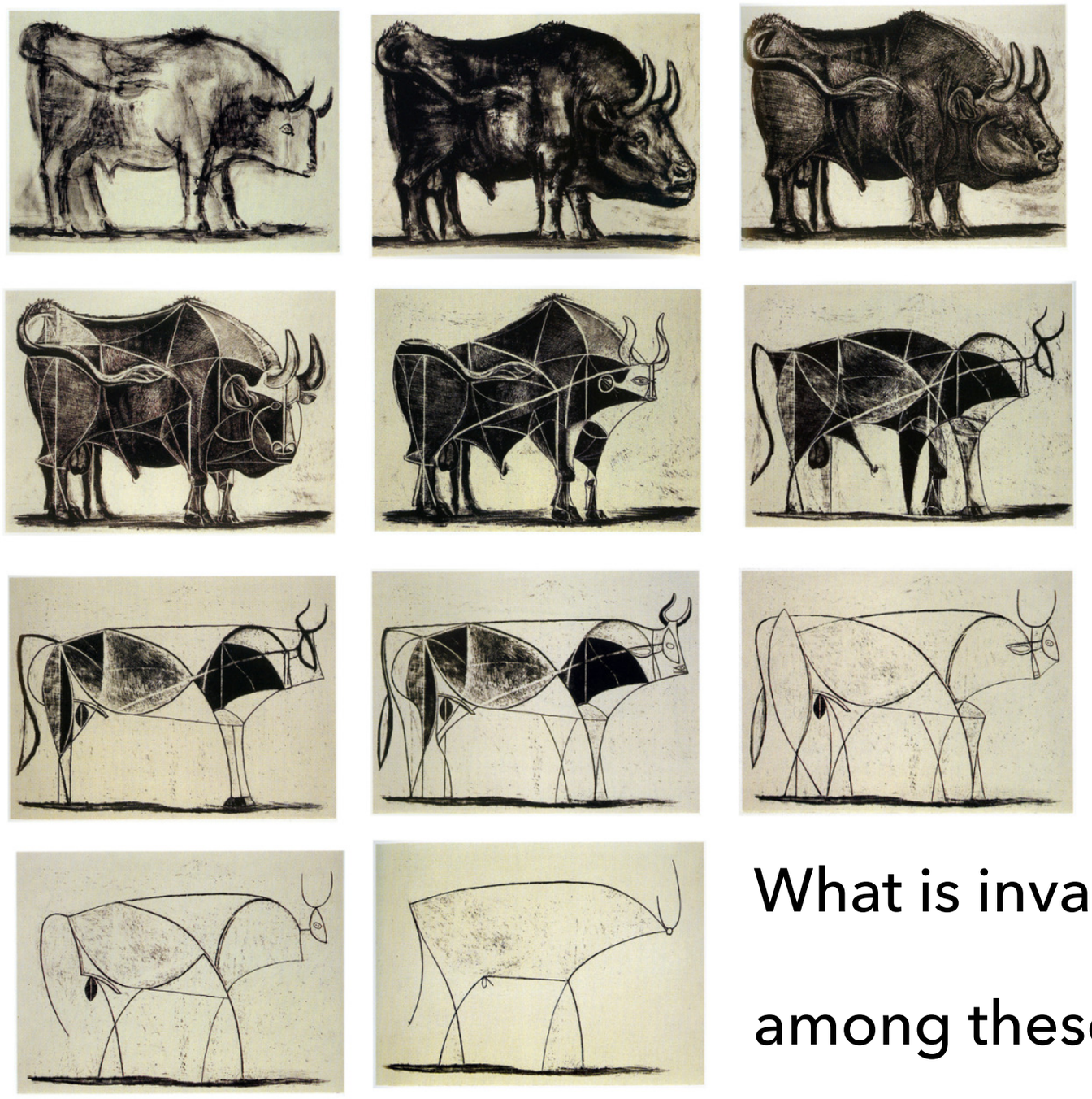
A SIMPLE DIFFEOMORPHISM FOR SOUND



- The modulus of DFT in not stable to pitch-shift;
- constant-Q transforms (for example wavelets) add stability by averaging, therefore loosing uniqueness.

HUMANS ARE GOOD AT HANDLING DEFORMATIONS…

. . .AND PICASSO IS VERY GOOD



What is invariant

among these bulls?

A MORE COMPLEX DIFFEOMORPHISM

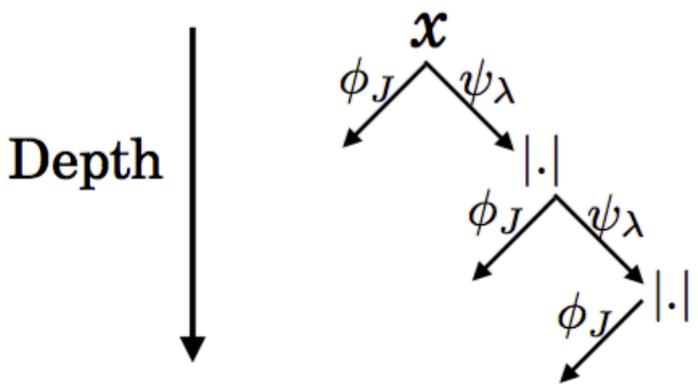
P. S. Johnson, 500 years of female portraits in Western Art

SCATTERING TRANSFORM

Definition

- The *scattering transform* (by S. Mallat) at scale J is a convolutional representation made of a cascade of complex wavelet transforms followed by a modulus non-linearity and averaged by a low-pass filter:

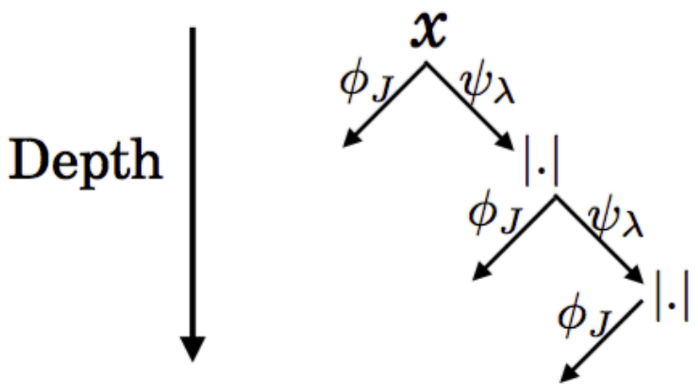
$$S_J x = \{x \star \phi_J, \quad \text{with } \lambda_i = \{j_i, \theta_i\}, j_i \leq J$$
$$|x \star \psi_{\lambda_1}| \star \phi_J,$$
$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J\}$$



Definition

- The *scattering transform* (by S. Mallat) at scale J is a convolutional representation made of a cascade of complex wavelet transforms followed by a modulus non-linearity and averaged by a low-pass filter:

$$S_J x = \{x \star \phi_J, \quad \text{with } \lambda_i = \{j_i, \theta_i\}, j_i \leq J$$
$$|x \star \psi_{\lambda_1}| \star \phi_J,$$
$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J\}$$

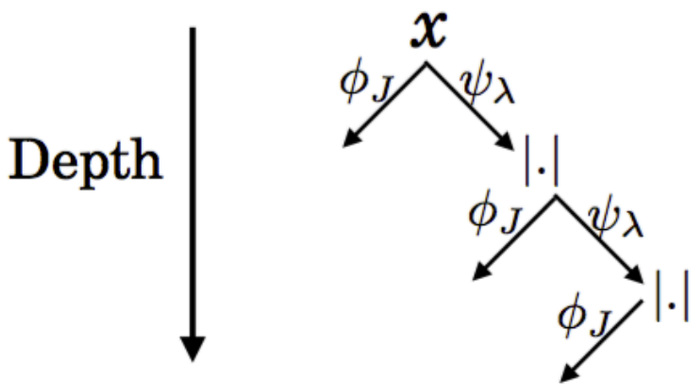


- it has multi-layer architecture, where higher layers recover the information lost in previous ones;

Definition

- The *scattering transform* (by S. Mallat) at scale J is a convolutional representation made of a cascade of complex wavelet transforms followed by a modulus non-linearity and averaged by a low-pass filter:

$$S_J x = \{x \star \phi_J, \quad \text{with } \lambda_i = \{j_i, \theta_i\}, j_i \leq J$$
$$|x \star \psi_{\lambda_1}| \star \phi_J,$$
$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J\}$$



- it has multi-layer architecture, where higher layers recover the information lost in previous ones;
- it is invariant up to J, stable to small diffeomorphisms and unique.

The scattering tree

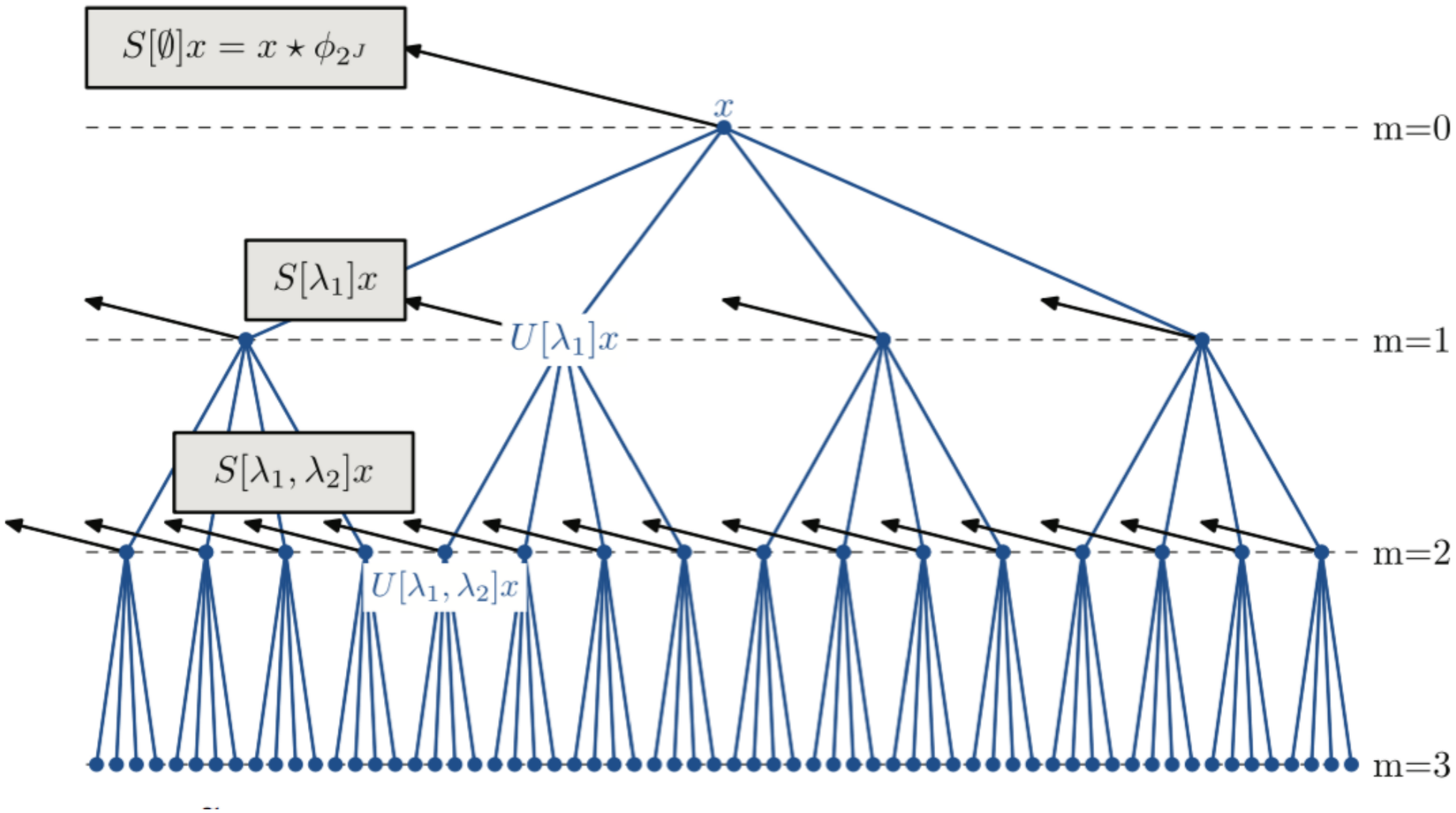


Image by J. Bruna and S. Mallat

Scattering on images

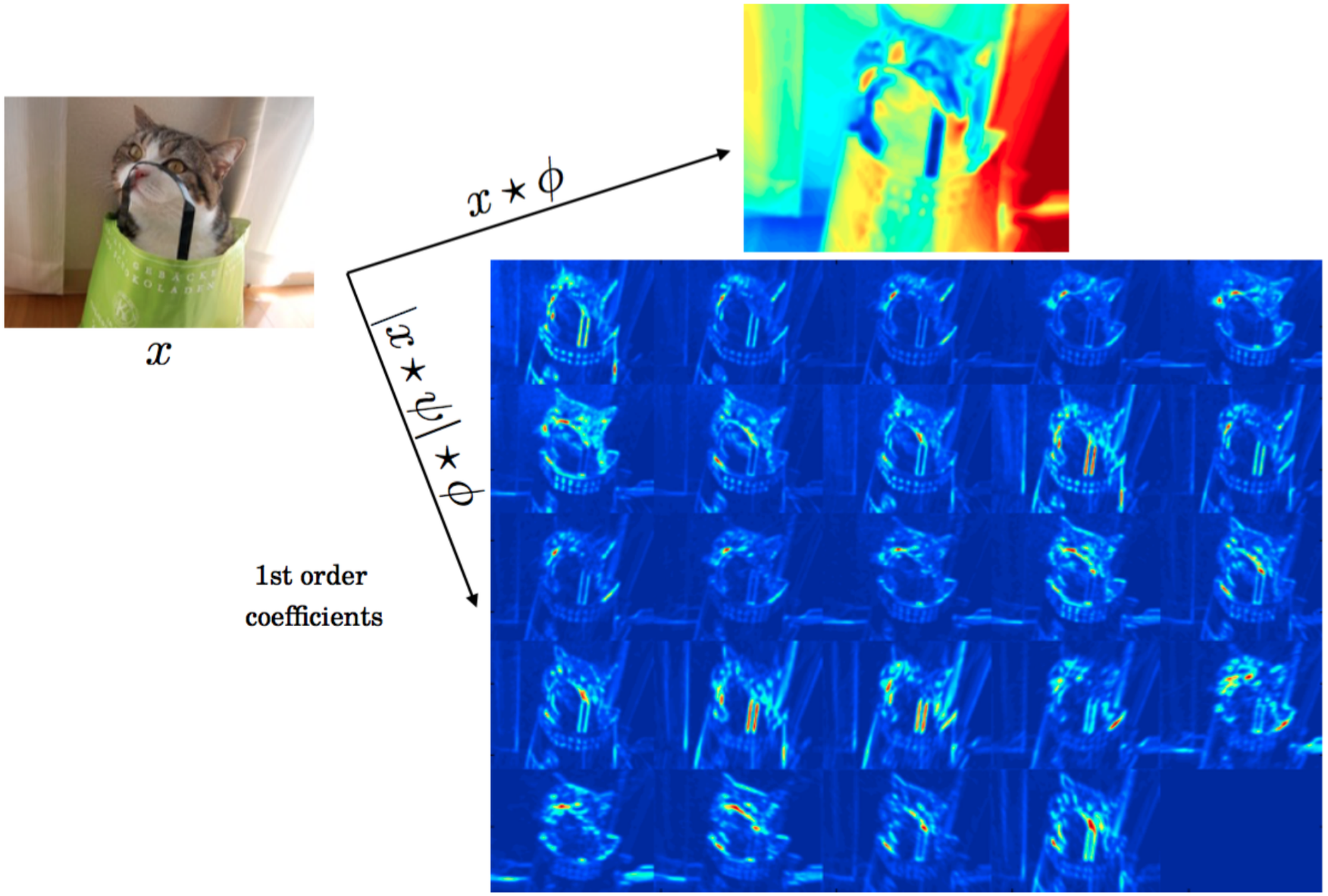


Image by E. Oyallon

Scattering on audio

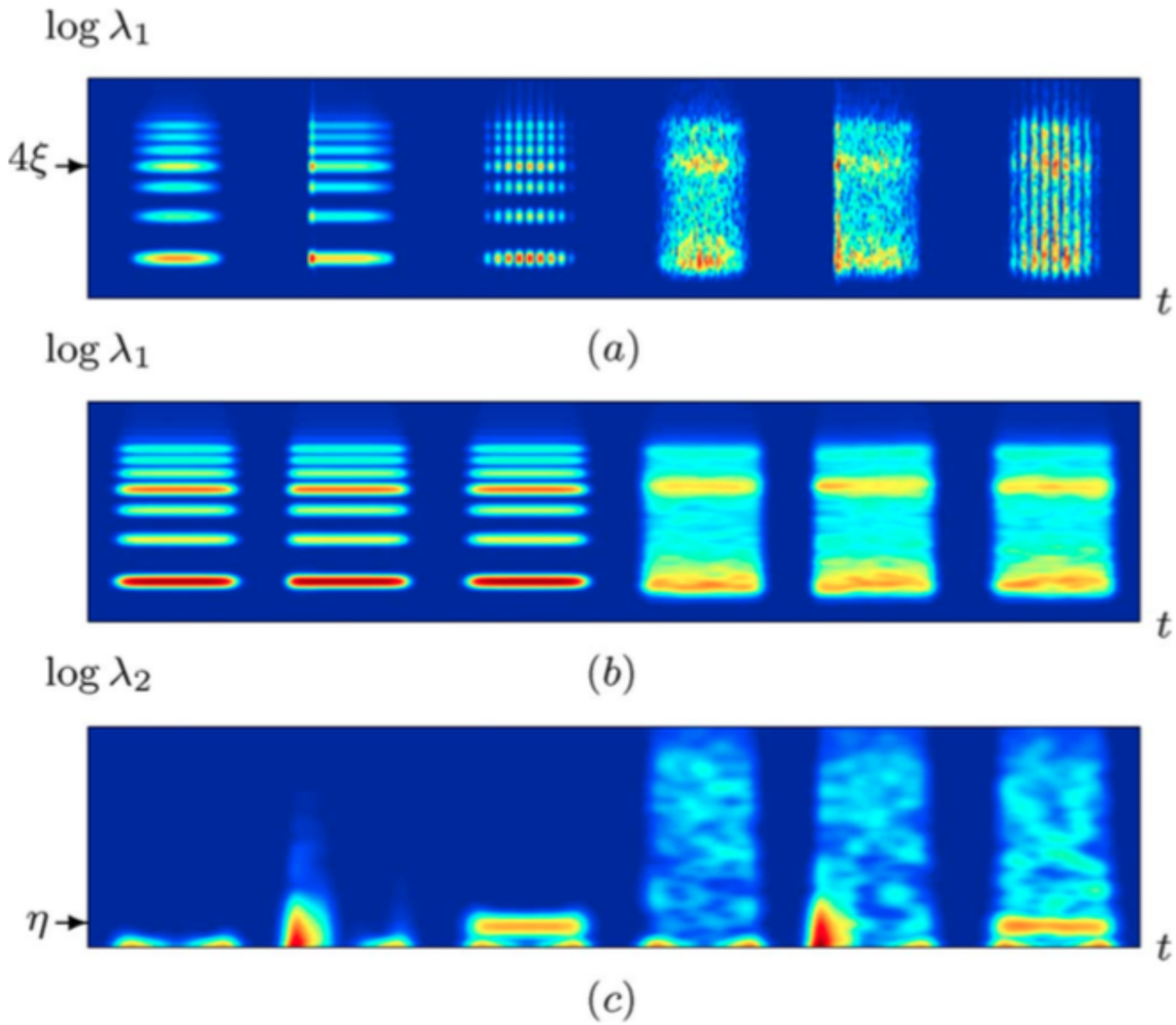
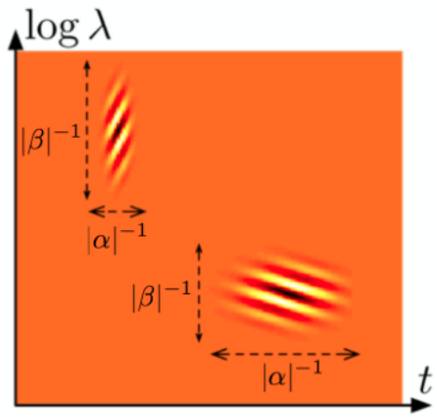


Image by J. Andén and S. Mallat

Multi-variable scattering

Capture complex structures in signals looking at time-frequency coherence:

- **joint-scattering**: $||x \overset{t}{\star} \psi_\lambda| \overset{t}{\star} \psi_\alpha \overset{log\lambda}{\star} \psi_\beta|$

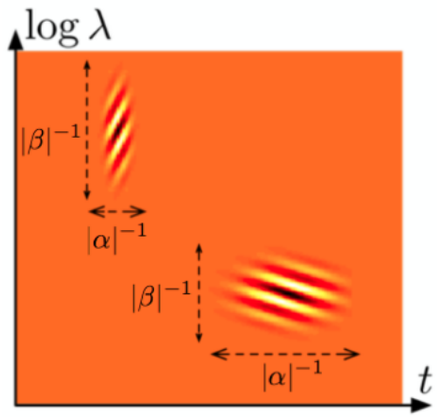


Images by V. Lostanlen

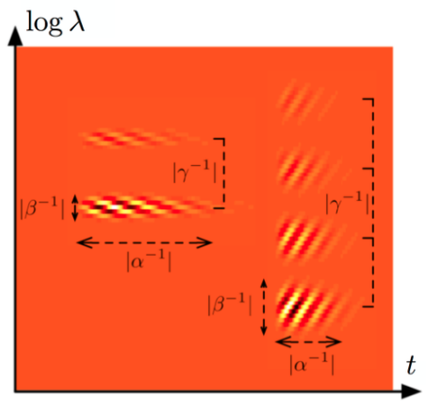
Multi-variable scattering

Capture complex structures in signals looking at time-frequency coherence:

- **joint-scattering**: $||x \overset{t}{\star} \psi_\lambda| \overset{t}{\star} \psi_\alpha \overset{log\lambda}{\star} \psi_\beta|$



- **spiral-scattering**: $||x \overset{t}{\star} \psi_\lambda| \overset{t}{\star} \psi_\alpha \overset{log\lambda}{\star} \psi_\beta \overset{oct}{\star} \psi_\gamma|$



Images by V. Lostanlen

Inverse scattering (1)

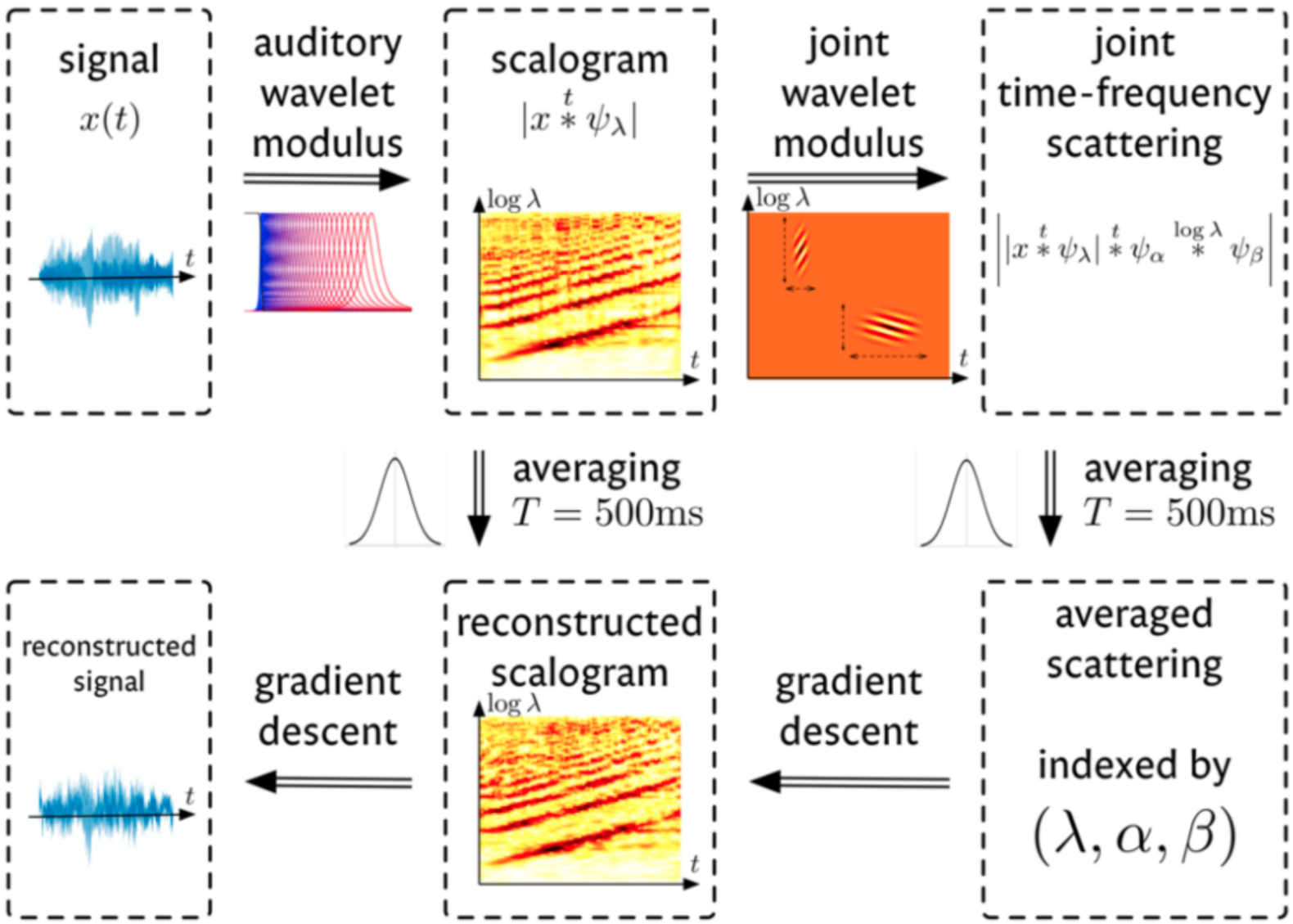
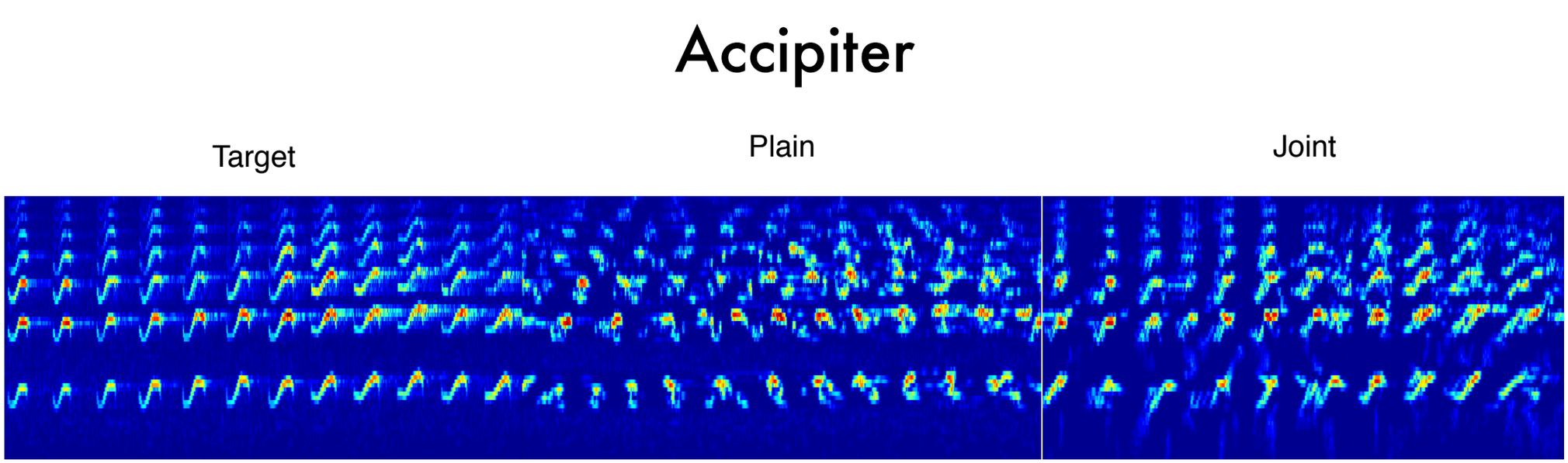


Image by V. Lostanlen

Inver scattering (2)

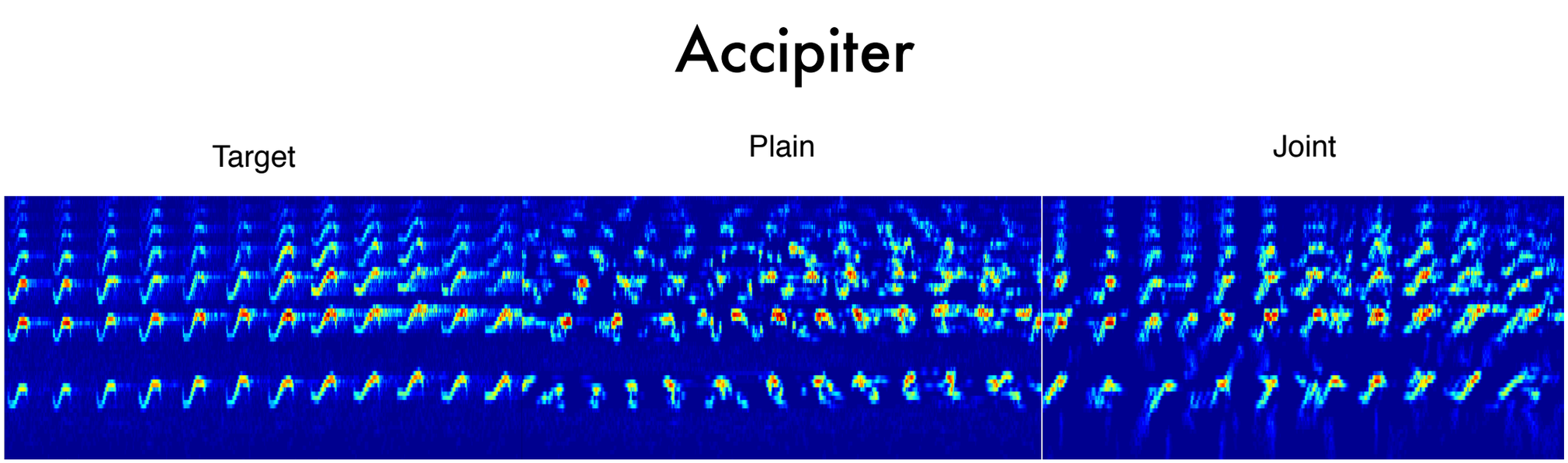


Examples

- Accipiter (original file).

Reconstructions by V. Lostanlen

Inver scattering (2)

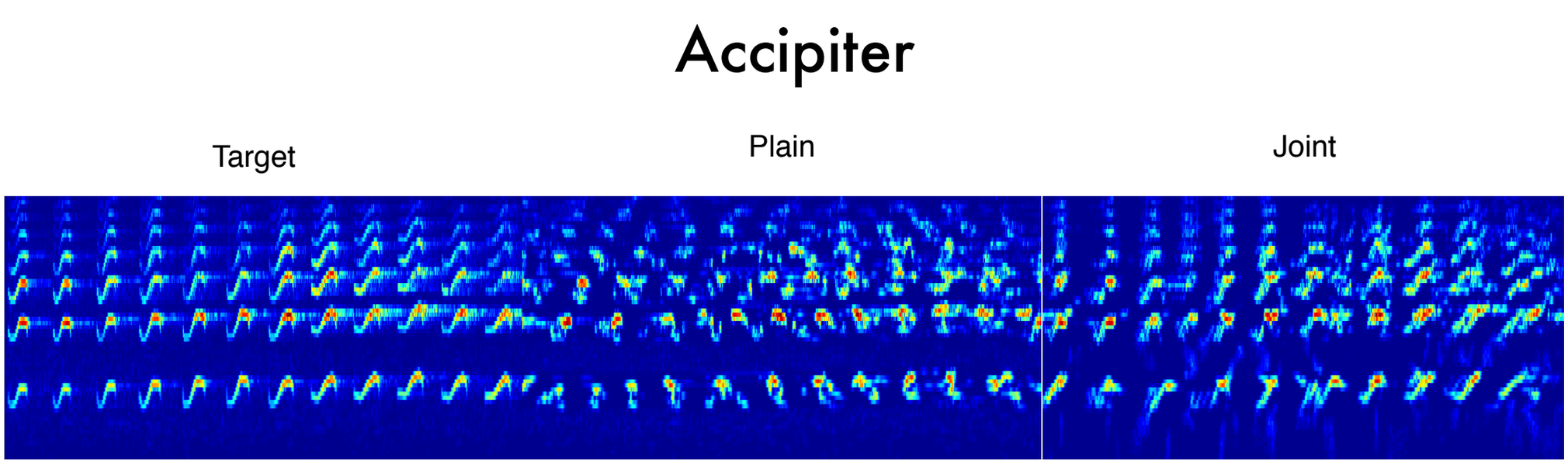


Examples

- Accipiter (original file).
- Second order plain scattering reconstruction.

Reconstructions by V. Lostanlen

Inver scattering (2)



Examples

- Accipiter (original file).
- Second order plain scattering reconstruction.
- Joint-scattering reconstruction.

Reconstructions by V. Lostanlen

Complexity of ideas

- Discover and represent meaningful concepts is extremely complex.

Complexity of ideas

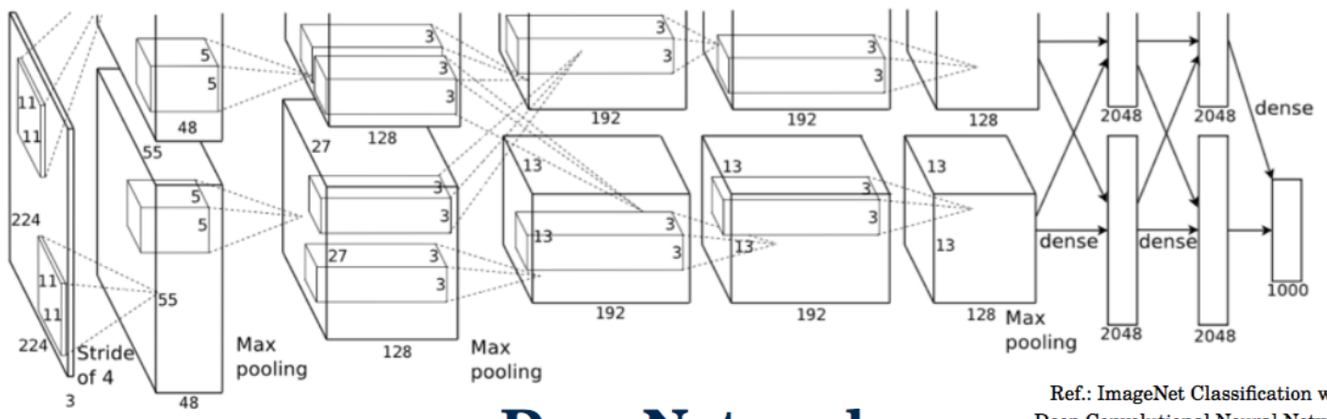
- Discover and represent meaningful concepts is extremely complex.
- Only geometric transformations are mathematically handled by convolutional representations.

Complexity of ideas

- Discover and represent meaningful concepts is extremely complex.
- Only geometric transformations are mathematically handled by convolutional representations.
- Higher-level transformations are either to be learned by specific algorithms or imposed beforehand by domain experts.

Complexity of ideas

- Discover and represent meaningful concepts is extremely complex.
- Only geometric transformations are mathematically handled by convolutional representations.
- Higher-level transformations are either to be learned by specific algorithms or imposed beforehand by domain experts.
- Supervised learning by deep networks seems to be able to spot out such high-level concepts, but there is no mathematical definition for the moment.



DeepNetwork

Ref.: ImageNet Classification with Deep Convolutional Neural Networks. A Krizhevsky et al.

SOUND-TYPES

AT A GLANCE

- The *theory of sound-types* is a multi-layer framework for sound representation with multiple abstraction levels (such as CNN)

- Sounds are described by **equivalence classes** and **probabilities:**

- This is realised by the **sound-types transform** (STT) and by joint probabilistic models and machine learning

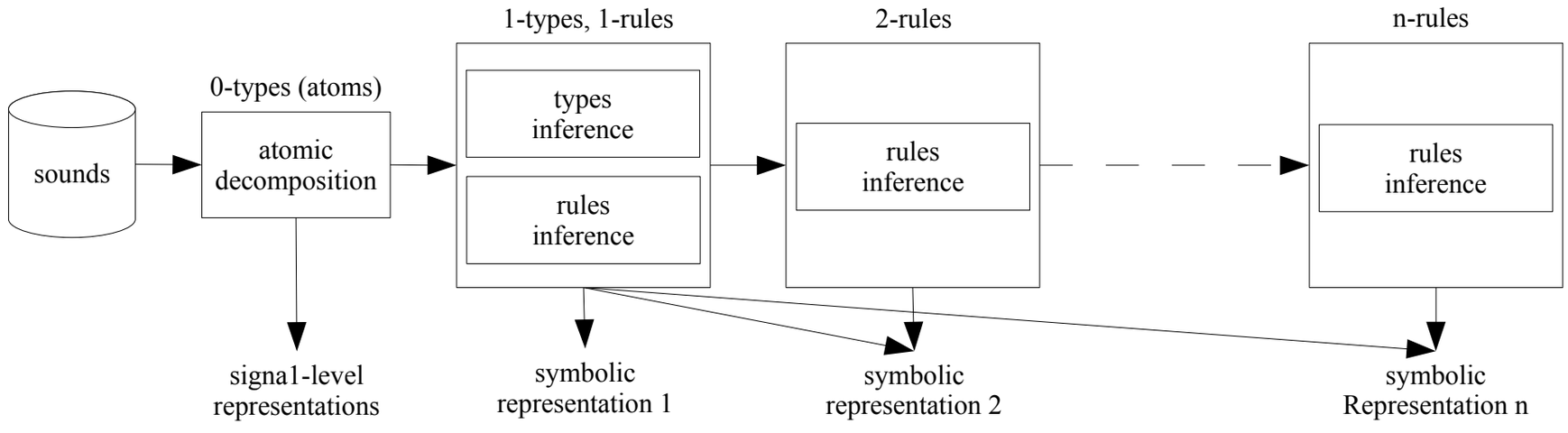
(spoiler alert: next slides will be less fun!!)

CREATION OF SOUND-TYPES 1/2

- **atomic decomposition**: subdivide a sound into small grains of approximately 40 ms called *atoms* or *0-types* overlapping in time and frequency;
- **1-types inference**: compute a set of low-level descriptors on the atoms, project the descriptors in a multi-dimensional space and compute the *clusters*; each cluster will represent a *1-type*;
- **1-rules inference**: estimate a Markov model to describe the sequences of types present in the analysed sound (*1-rules)*;
- **1-level representation**: represent the sound in a symbolic language using the discovered 1-types and 1-rules;

CREATION OF SOUND-TYPES 2/2

- **n-rules inference**: estimate a Markov model of order n to describe the sequences of 1-types;
- **n-level representation**: represent the sound in a symbolic language using the discovered 1-types and n-rules;
- **repeat n-level rules and n-level representations**: until the desired level number has been reached.



SOUND-TYPES TRANSFORM 1/3

- Given a signal x of length N-samples and a window $\vec{h}^{n}$ of length n-samples, it is possible to define an **atom** as a windowed chunk of the signal of length n-samples:

Cut a signal into pieces...

$$\vec{a}^{n} = \vec{h}^{n} \cdot \vec{x}^{n} .$$

- A **sound-cluster** as a set of atoms that *lie* in a defined area of a feature space (ie. that share a *similar* set of features):

...then group pieces that are *similar*

$$\vec{c}_r^{k_r} = \{\vec{a}_{r,1}^{n}, \ldots, \vec{a}_{r,k_r}^{n}\}.$$

The content of $\vec{c}_r^{k_r}$ is given by a statistical analysis applied on the feature space.

SOUND-TYPES TRANSFORM 2/3

- A sound-cluster has an associate **sound-type** $\vec[n]{\tau}_r$, defined as the weighted sum of all the atoms in the sound-cluster where the weights $\vec[k_r]{\omega}_r$ are the distances of each atom to the center of the cluster:

$$\vec[n]{\tau}_r = \sum_{j=1}^{k_r} \vec[n]{a}_{r,j} \cdot \omega_{r,j}$$

Create an intermediate representation of each *type* of sound discovered

with $\omega_{r,j} \in \vec[k_r]{\omega}_r$.

- The set of sound-types in the signal $\vec[N]{x}$ is called **dictionary**:

$$\mathcal{D}_{\vec[N]{x}} = \{\vec[n]{\tau}_1, \ldots, \vec[n]{\tau}_r\}.$$

SOUND-TYPES TRANSFORM 3/3

- Finally, it is possible to define the *sound-types transform* as a function of time and frequency obtained by multiplying the sound-types in a given dictionary with complex sinusoids:

$$\vec{\Phi}^{N}_{\vec{k}_n} = \sum_{i=0}^{N/t} \vec{\tau}^{n}_{r,p} \cdot e^{-j \cdot \frac{2 \cdot \pi}{n} \cdot \vec{k}^{n}}$$

Create the final projective representation at the appropriate abstraction level

where $\vec{k}^{n} = \{ f_1, \ldots, f_n \}$ is a vector of frequencies.

PROPERTIES OF SOUND-TYPES

- The extreme case for $|\mathcal{M}| = N/t$ is interesting: for that abstraction level, each sound-cluster is a singleton made of a single atom and consequently each sound-type reduces to that single atom scaled in amplitude:

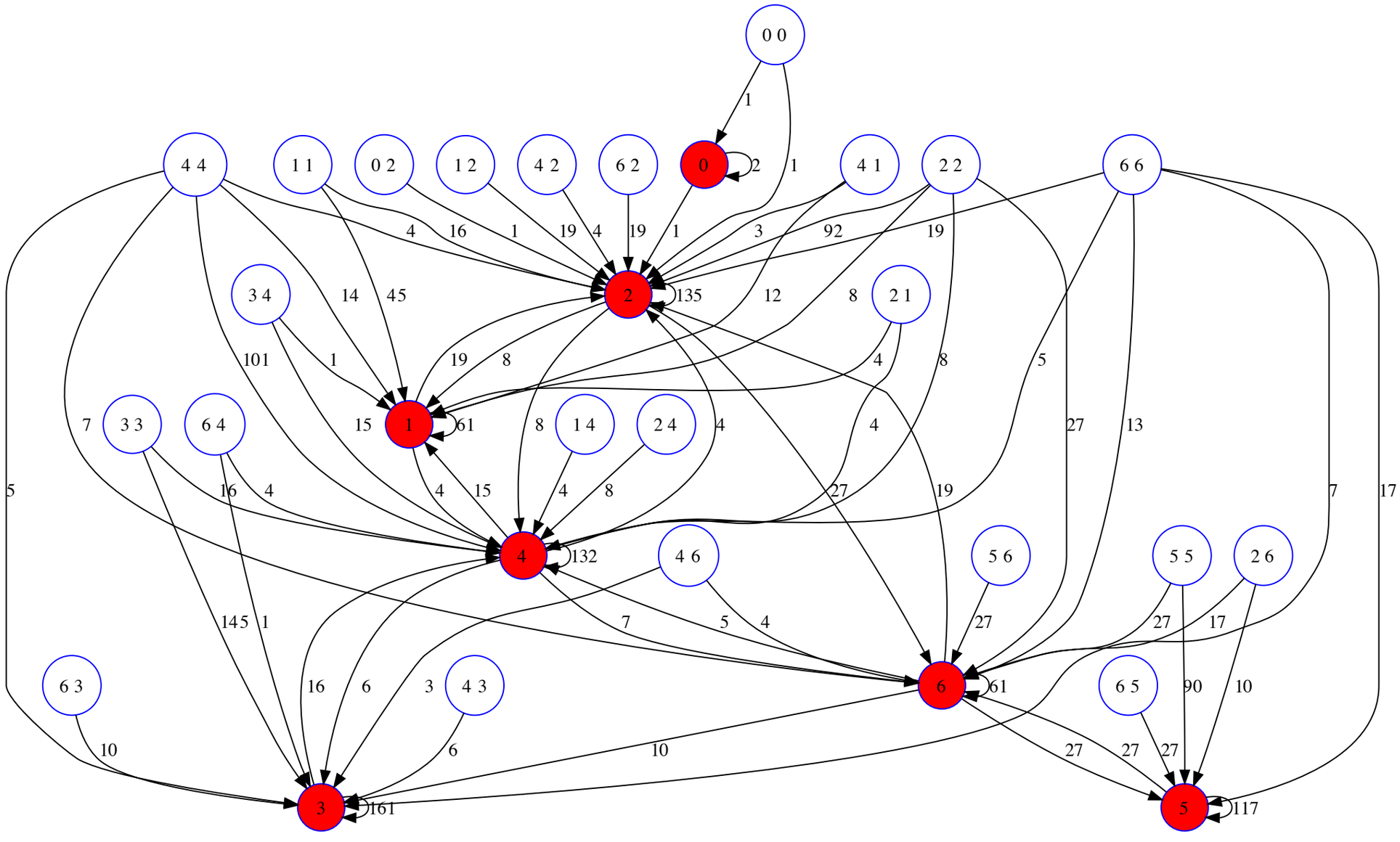
$$|\mathcal{M}| = N/t \implies \vec{c}_r = \{\vec{a}_1\} \implies \vec{\tau}_r = \vec{a}_r \cdot \omega_{r,1}.$$

- This leads to the important consequence that STT is a **generalization** of STFT:

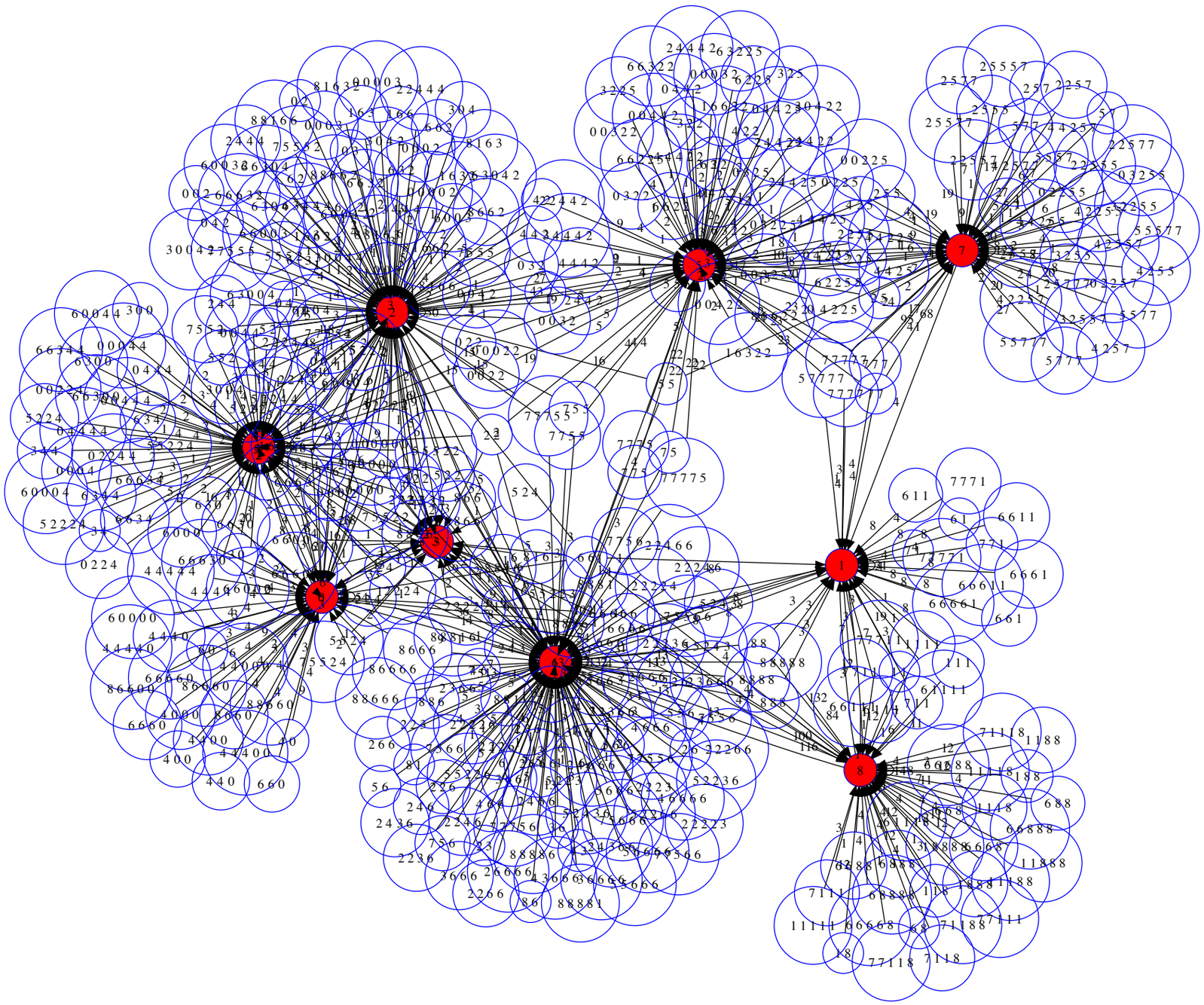
$$\overset{n}{\vec{\tau}}_r = \overset{n}{\vec{a}}_r = \overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}} \implies \sum_{i=0}^{N/t} \overset{n}{\vec{\tau}}_{r,p} \cdot e^{-j \cdot \frac{2 \cdot \pi}{n} \cdot \overset{n}{\vec{k}}} = \sum_{i=0}^{N/t} \overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}}_{i \cdot t} \cdot e^{-j \cdot \frac{2 \cdot \pi}{n} \cdot \overset{n}{\vec{k}}}$$

with p defined as above.

SOUND-TYPES AT ABSTRACTION LEVEL 2



SOUND-TYPES AT ABSTRACTION LEVEL 5



EXAMPLES

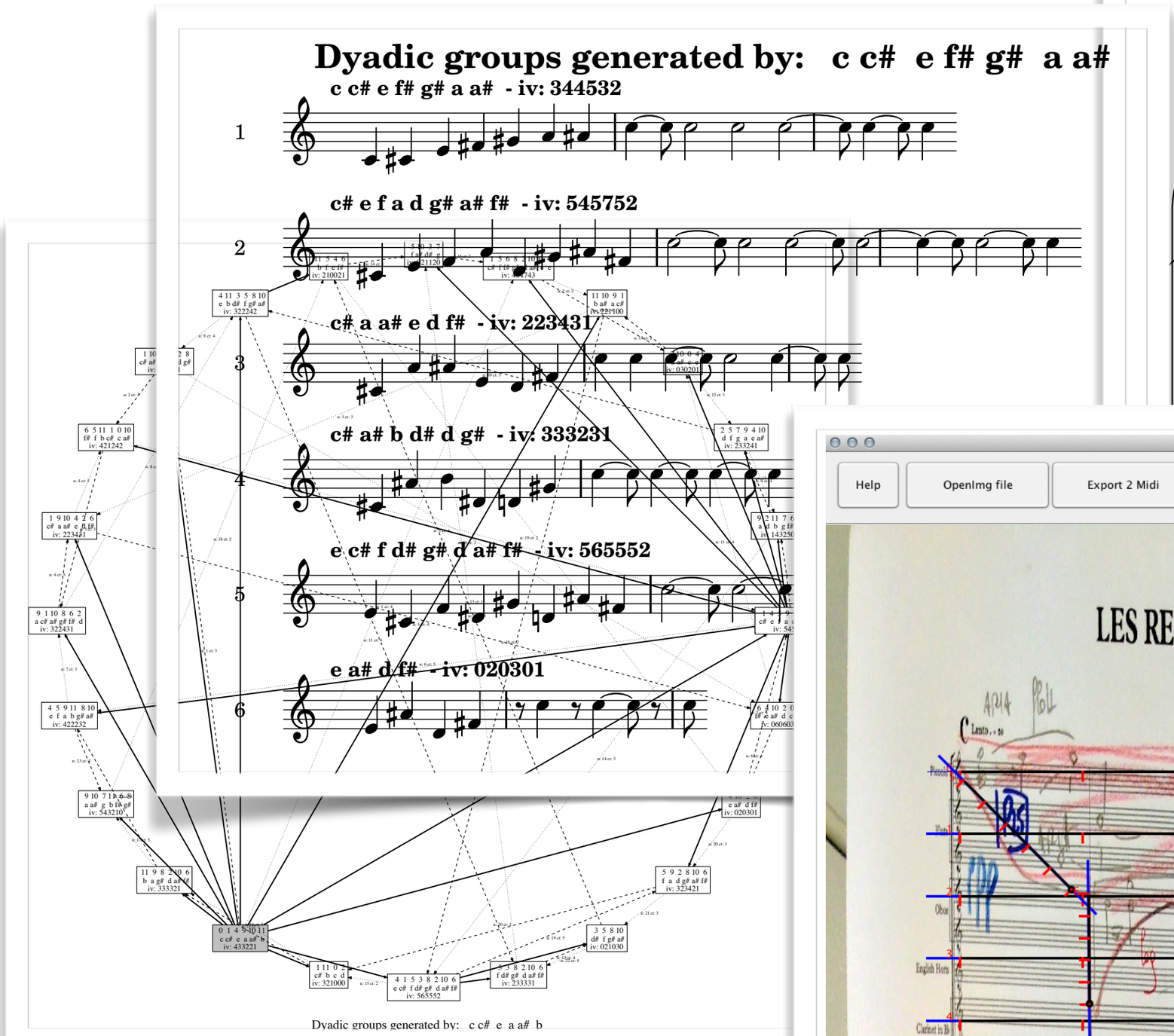
Probabilistic generation / sound-types matching

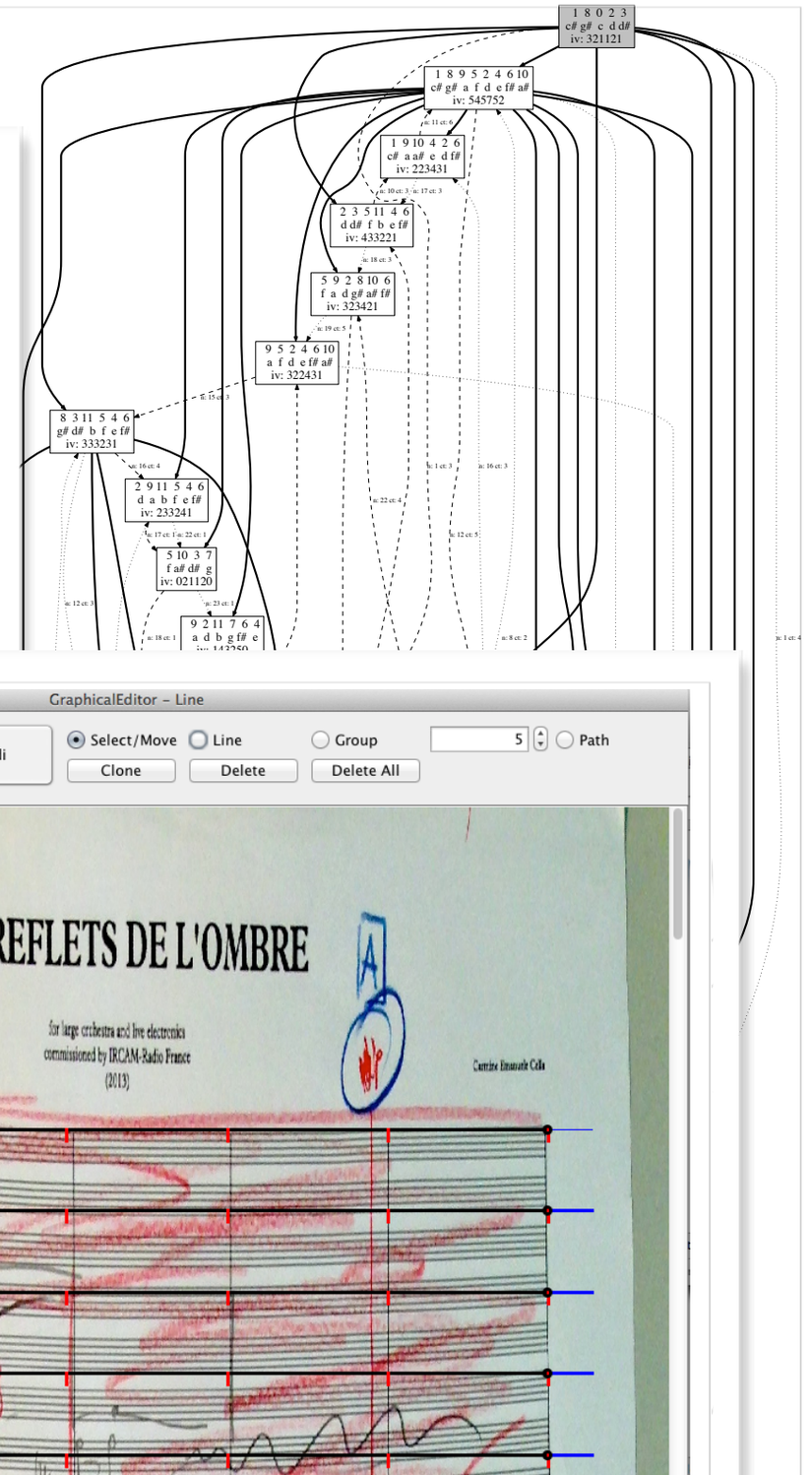
- Bass sample and its probabilistic generation.
- A probabilistic generation of an orchestral sample.
- Hybridization between instruments and voices (Gervasoni).
- Hybridization between Cage and Beach boys.
- Hybridization between Cage and Lachenmann's piece.

REFLETS DE L'OMBRE (2013)

- **Instrumentation**: large orchestra and live electronics
- **Commission**: IRCAM - Radio France
- **First performance**: 6 june 2013, Salle Pleyel - Paris, OPRF, Jukka-Pekka Saraste
- **Duration**: 18 minutes
- **Production**: jan-june 2012 musical research, oct 2012-june 2013 studio work

EXCERPT


Dyadic groups generated by: c c# e f# g# a a#
c c# e f# g# a a# - iv: 344532
c# e f a d g# a# f# - iv: 545752
c# a a# e d f# - iv: 223431
c# a# b d# d g# - iv: 333231
e c# f d# g# d a# f# - iv: 565552
e a# d f# - iv: 020301
Dyadic groups generated by: c c# e a a# b



GraphicalEditor – Line
Help
OpenImg file
Export 2 Midi
Select/Move
Line
Group
Path
Clone
Delete
Delete All
LES REFLETS DE L'OMBRE
for large orchestra and live electronics
commissioned by IRCAM-Radio France
(2013)
Carrière Emanuele Cella
Flutes
Oboe
English Horn
Clarinet in Bb

THANK YOU!

Suggested exercise: try to use the code on github to make some cool sounds!