Activités de recherche et d'enseignement

Gautier Appert

mai, 2021

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).
   - Distorsion du type k-means généralisés
     $\mathbb{E}\left(\min\left\{\left\|X-\sum_{j\in A} c_j\right\|^2 : A\subset[\![1,k]\!], \bigsqcup_{j\in A}\mathrm{supp}(c_j)=[\![1,d]\!]\right\}\right)$.

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).
   - Distorsion du type k-means généralisés
     $\mathbb{E}\left(\min\left\{\left\|X - \sum_{j\in A} c_j\right\|^2 : A \subset [\![1,k]\!], \bigsqcup_{j\in A} \mathrm{supp}(c_j) = [\![1,d]\!]\right\}\right)$.
   - Bornes de généralisation.

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).
   - Distorsion du type k-means généralisés
     $\mathbb{E}\left(\min\left\{\left\|X - \sum_{j\in A} c_j\right\|^2 : A \subset [\![1,k]\!], \bigsqcup_{j\in A} \mathrm{supp}(c_j) = [\![1,d]\!]\right\}\right)$.
   - Bornes de généralisation.
   - Algorithme associé de compression avec perte.

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).
   - Distorsion du type k-means généralisés
     $\mathbb{E}\left(\min\left\{\left\|X-\sum_{j\in A} c_j\right\|^2 : A\subset[\![1,k]\!], \bigsqcup_{j\in A}\mathrm{supp}(c_j)=[\![1,d]\!]\right\}\right)$.
   - Bornes de généralisation.
   - Algorithme associé de compression avec perte.
   - Représentation des images par un ensemble aléatoire de fragments, package `R`/`C++`

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).
   - Distorsion du type k-means généralisés $\mathbb{E}\left(\min\left\{\left\|X-\sum_{j\in A} c_j\right\|^2 : A\subset[\![1,k]\!], \bigsqcup_{j\in A}\mathrm{supp}(c_j)=[\![1,d]\!]\right\}\right)$.
   - Bornes de généralisation.
   - Algorithme associé de compression avec perte.
   - Représentation des images par un ensemble aléatoire de fragments, package `R`/`C++`
3. Analyse syntaxique d'un ensemble aléatoire de labels.

Thèmes de recherche

Objectif : analyser la structure statistique de données de grande dimension.

Apprentissage non supervisé

1. Critère des k-means dans un Hilbert séparable et généralisations.
   - Bornes de généralisation non asymptotiques et inégalités PAC-Bayésiennes en dimension infinie.
   - Interprétations et extensions du critère. Relation avec l'estimation de la loi des données, quantification vectorielle de probabilités conditionnelles (clustering de bags of words), critère robuste
2. Etiquetage non supervisé des parties d'un signal $X \in \mathbb{R}^d$ (une image).
   - Distorsion du type k-means généralisés $\mathbb{E}\left(\min\left\{\left\|X-\sum_{j\in A} c_j\right\|^2 : A\subset[\![1,k]\!], \bigsqcup_{j\in A}\mathrm{supp}(c_j)=[\![1,d]\!]\right\}\right)$.
   - Bornes de généralisation.
   - Algorithme associé de compression avec perte.
   - Représentation des images par un ensemble aléatoire de fragments, package `R`/`C++`
3. Analyse syntaxique d'un ensemble aléatoire de labels.
   - Classification de labels et compression à base de grammaires.

Bornes de généralisations sur l'excès de risque des k-means

Minimiseur du risque empirique: $\widehat{C} \in \arg\min_{C \in H^k} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - C_j\|^2$ vérifie

Bornes de généralisations sur l'excès de risque des k-means

Minimiseur du risque empirique: $\widehat{C} \in \arg\min_{C \in H^k} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - C_j\|^2$ vérifie

$$\mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - \widehat{C}_j\|^2\right)\right]$$

$$\leq \inf_{C \in H^k} \mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\right) \underbrace{+16\, B^2 \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log(k)}{n}}}_{\text{Amélioration de [Biau et al 2008], [Fefferman et al, 2016] et [Klochkov et al 2020]}}.$$

Bornes non asymptotiques, indépendantes de la dimension

Bornes de généralisations sur l'excès de risque des k-means

Minimiseur du risque empirique: $\widehat{C} \in \arg\min_{C \in H^k} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - C_j\|^2$ vérifie

$$\mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - \widehat{C}_j\|^2\right)\right]$$

$$\leq \inf_{C \in H^k} \mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\right) \underbrace{+16\, B^2 \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log(k)}{n}}}.$$

Amélioration de [Biau et al 2008], [Fefferman et al, 2016] et [Klochkov et al 2020]

Bornes non asymptotiques, indépendantes de la dimension

Principes de la preuve:

Bornes de généralisations sur l'excès de risque des k-means

Minimiseur du risque empirique: $\widehat{C} \in \arg\min_{C \in H^k} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - C_j\|^2$ vérifie

$$\mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - \widehat{C}_j\|^2\right)\right]$$

$$\leq \inf_{C \in H^k} \mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\right) \underbrace{+16\, B^2 \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log(k)}{n}}}_{\text{Amélioration de [Biau et al 2008], [Fefferman et al, 2016] et [Klochkov et al 2020]}}.$$

Bornes non asymptotiques, indépendantes de la dimension

Principes de la preuve:

- Reparamétrisation linéaire de la fonction de perte dans un RKHS: $\|X - C_j\|^2 = \langle \theta_j(C_j), W(X) \rangle$

Bornes de généralisations sur l'excès de risque des k-means

Minimiseur du risque empirique: $\widehat{C} \in \arg\min_{C \in H^k} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - C_j\|^2$ vérifie

$$\mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - \widehat{C}_j\|^2\right)\right]$$

$$\leq \inf_{C \in H^k} \mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\right) \underbrace{+16\, B^2 \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log(k)}{n}}}_{\text{Amélioration de [Biau et al 2008], [Fefferman et al, 2016] et [Klochkov et al 2020]}}.$$

Bornes non asymptotiques, indépendantes de la dimension

Principes de la preuve:

- Reparamétrisation linéaire de la fonction de perte dans un RKHS: $\|X - C_j\|^2 = \langle \theta_j(C_j), W(X) \rangle$
- Inégalités PAC-Bayésiennes en dimension infinie

Bornes de généralisations sur l'excès de risque des k-means

Minimiseur du risque empirique: $\widehat{C} \in \arg\min_{C \in H^k} \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [\![1,k]\!]} \|X_i - C_j\|^2$ vérifie

$$\mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - \widehat{C}_j\|^2\right)\right]$$

$$\leq \inf_{C \in H^k} \mathbb{E}_X\left(\min_{j \in [\![1,k]\!]} \|X - C_j\|^2\right) \underbrace{+16\, B^2 \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log(k)}{n}}}_{\text{Amélioration de [Biau et al 2008], [Fefferman et al, 2016] et [Klochkov et al 2020]}}.$$

Bornes non asymptotiques, indépendantes de la dimension

Principes de la preuve:

- Reparamétrisation linéaire de la fonction de perte dans un RKHS: $\|X - C_j\|^2 = \langle \theta_j(C_j), W(X) \rangle$
- Inégalités PAC-Bayésiennes en dimension infinie
- Chaining PAC-Bayésien $\Longrightarrow$ séquence de perturbations gaussiennes $\rho_\theta \in \mathcal{M}_+^1(\mathbb{R}^{\mathbb{N}})$ indéxées par $\theta \in H$

Extension du critère quadratique des k-means

Extension du critère quadratique des k-means

- Information k-means $\Longrightarrow$ clustering d'un histogramme $p_X \sim \mathbb{P}_{p_X}$

$$\inf_{q_1,\dots,q_k \in \left(\mathbb{L}^1_{+,1}(\nu)\right)^k} \mathbb{E}\left(\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_X)\right),$$

où $\mathbb{L}^1_{+,1}(\nu) = \left\{q \in \mathbb{L}^1(\nu) : q \geq 0, \int q \,\mathrm{d}\nu = 1\right\}$

Extension du critère quadratique des k-means

- Information k-means $\Longrightarrow$ clustering d'un histogramme $p_X \sim \mathbb{P}_{p_X}$

$$\inf_{q_1,\dots,q_k \in \left(\mathbb{L}^1_{+,1}(\nu)\right)^k} \mathbb{E}\Big(\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_X)\Big),$$

où $\mathbb{L}^1_{+,1}(\nu) = \left\{ q \in \mathbb{L}^1(\nu) : q \geq 0, \int q \,\mathrm{d}\nu = 1 \right\}$

Algorithme de Lloyd:

Extension du critère quadratique des k-means

- Information k-means $\Longrightarrow$ clustering d'un histogramme $p_X \sim \mathbb{P}_{p_X}$

$$\inf_{q_1,\dots,q_k \in \left(\mathbb{L}^1_{+,1}(\nu)\right)^k} \mathbb{E}\Big(\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_X)\Big),$$

où $\mathbb{L}^1_{+,1}(\nu) = \Big\{q \in \mathbb{L}^1(\nu) : q \geq 0, \int q \,\mathrm{d}\nu = 1\Big\}$

Algorithme de Lloyd:

- Centroids optimaux

$$q_j^{\star,\ell} = Z_j^{-1} \exp\Big\{\mathbb{E}\big[\log(p_X) \,|\, \ell(X) = j\big]\Big\}, \qquad j \in [\![1,k]\!],$$

Extension du critère quadratique des k-means

- Information k-means $\Longrightarrow$ clustering d'un histogramme $p_X \sim \mathbb{P}_{p_X}$

$$\inf_{q_1,\dots,q_k \in \left(\mathbb{L}^1_{+,1}(\nu)\right)^k} \mathbb{E}\Big(\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_X)\Big),$$

où $\mathbb{L}^1_{+,1}(\nu) = \Big\{q \in \mathbb{L}^1(\nu) : q \geq 0, \int q \,\mathrm{d}\nu = 1\Big\}$

Algorithme de Lloyd:

- Centroids optimaux

$$q_j^{\star,\ell} = Z_j^{-1} \exp\Big\{\mathbb{E}\big[\log(p_X) \mid \ell(X) = j\big]\Big\}, \qquad j \in [\![1,k]\!],$$

- Classification optimale

$$\ell^\star_q(x) = \arg\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_x), \qquad x \in \mathcal{X}$$

Etude des k-means dans un Hilbert séparable

Extension du critère quadratique des k-means

- Information k-means $\Longrightarrow$ clustering d'un histogramme $p_X \sim \mathbb{P}_{p_X}$

$$\inf_{q_1,\dots,q_k \in \left(\mathbb{L}^1_{+,1}(\nu)\right)^k} \mathbb{E}\Big(\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_X)\Big),$$

où $\mathbb{L}^1_{+,1}(\nu) = \Big\{q \in \mathbb{L}^1(\nu) : q \geq 0, \int q \,\mathrm{d}\nu = 1\Big\}$

Algorithme de Lloyd:

- Centroids optimaux

$$q_j^{\star,\ell} = Z_j^{-1} \exp\Big\{\mathbb{E}\big[\log(p_X) \mid \ell(X) = j\big]\Big\}, \qquad j \in [\![1,k]\!],$$

- Classification optimale

$$\ell_q^\star(x) = \arg\min_{j\in[\![1,k]\!]} \mathcal{K}(q_j, p_x), \qquad x \in \mathcal{X}$$

Article

New bounds for k-means and Information k-means. Submitted. arXivpreprint, 2021.

Fragmentation

Fragmentation

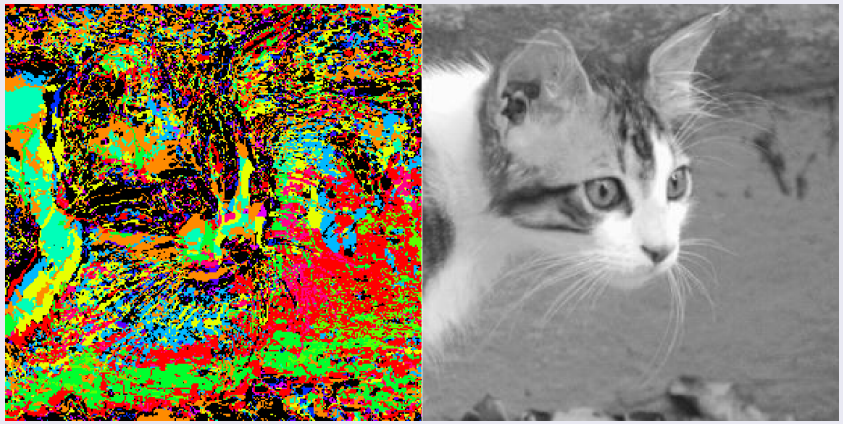
- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.

Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.

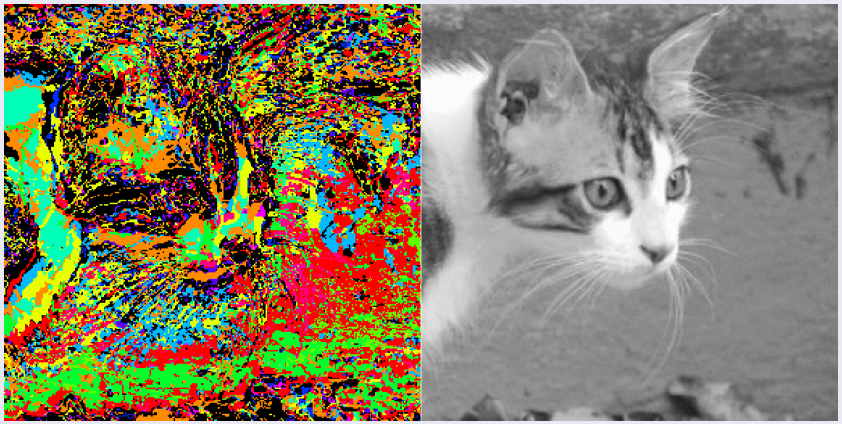
Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.



Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.



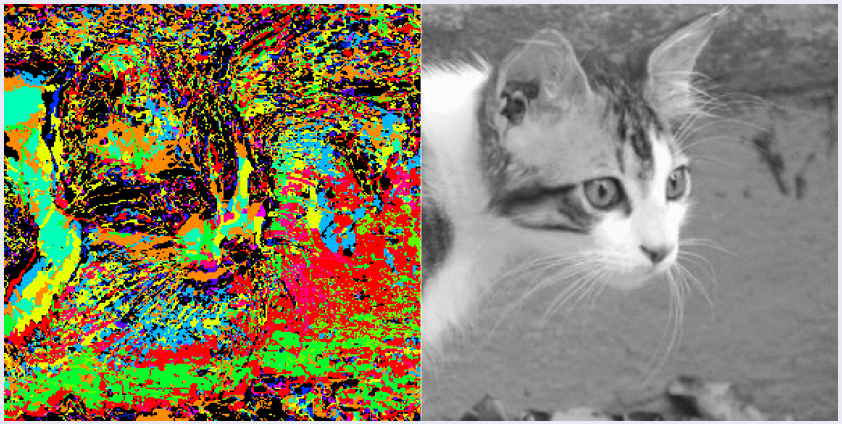
Distorsion:

$$\mathcal{R}(B, C) = \mathbb{E}\left(\min_{A \in \mathcal{T}_B} \left\| X - \sum_{j \in A} C_j \right\|^2 \right) \text{ avec}$$

$$\mathcal{T}_B = \left\{ A \subset [\![1, k]\!] : \bigsqcup_{j \in A} B_j = [\![1, d]\!] \right\}$$

Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.



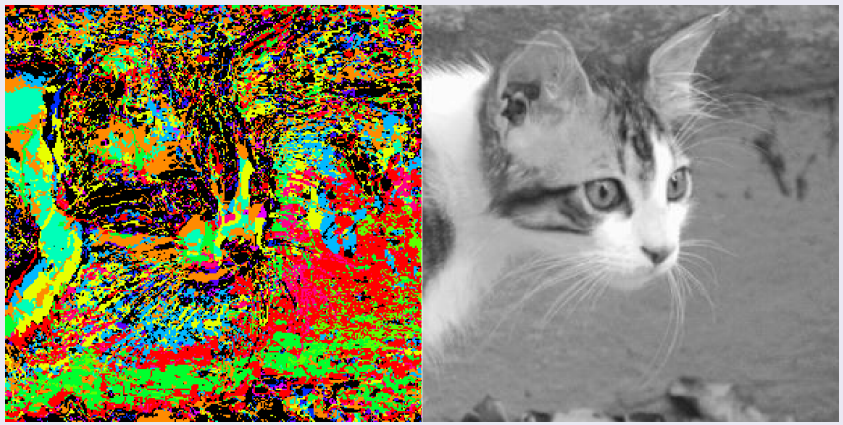
Distorsion:

$$\mathcal{R}(B, C) = \mathbb{E}\left(\min_{A \in \mathcal{T}_B} \left\| X - \sum_{j \in A} C_j \right\|^2 \right) \text{ avec}$$

$$\mathcal{T}_B = \left\{ A \subset [\![1, k]\!] : \bigsqcup_{j \in A} B_j = [\![1, d]\!] \right\}$$

Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.



Distorsion:

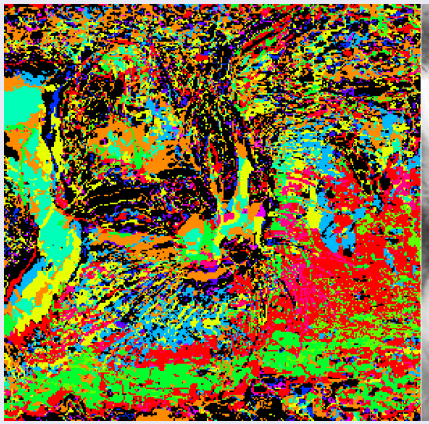
$$\mathcal{R}(B, C) = \mathbb{E}\left(\min_{A \in \mathcal{T}_B} \left\| X - \sum_{j \in A} C_j \right\|^2 \right) \text{ avec}$$

$$\mathcal{T}_B = \left\{ A \subset [\![1, k]\!] : \bigsqcup_{j \in A} B_j = [\![1, d]\!] \right\}$$

- Invention d'un nouvel algorithme de compression avec perte (du type Lempel Ziv):
  on minimise la surface totale de recouvrement $\sum_{j=1}^{k} \mathbb{P}_S(B_j)$ des fragments, à distorsion fixée $\mathcal{R}(B, C) \leq \alpha$.

Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.




Distorsion:

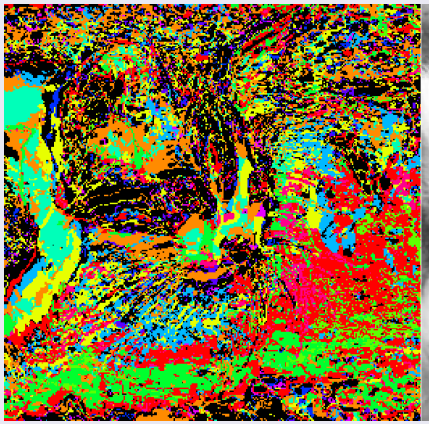
$$\mathcal{R}(B,C) = \mathbb{E}\Big(\min_{A\in\mathcal{T}_B} \big\|X - \sum_{j\in A} C_j\big\|^2\Big) \text{ avec}$$

$$\mathcal{T}_B = \Big\{A \subset [\![1,k]\!] : \bigsqcup_{j\in A} B_j = [\![1,d]\!]\Big\}$$

- Invention d'un nouvel algorithme de compression avec perte (du type Lempel Ziv):
  on minimise la surface totale de recouvrement $\sum_{j=1}^k \mathbb{P}_S(B_j)$ des fragments, à distorsion fixée $\mathcal{R}(B,C) \leq \alpha$.
  - Ne dépend pas de la géométrie du capteur.

Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.




Distorsion:

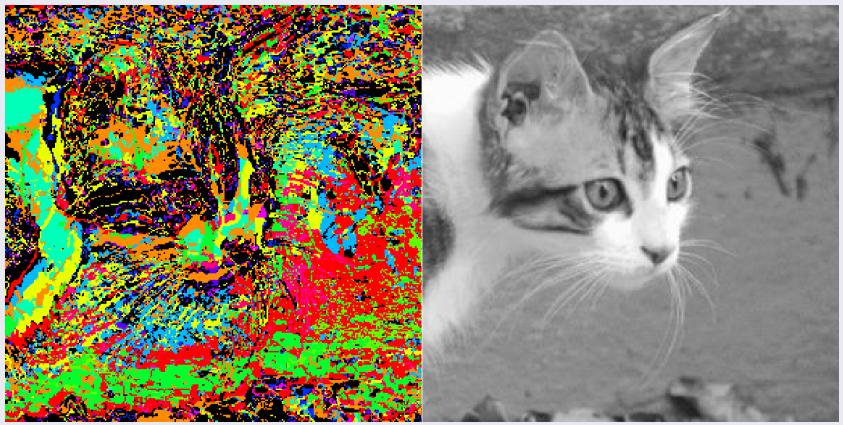
$$\mathcal{R}(B, C) = \mathbb{E}\left(\min_{A \in \mathcal{T}_B} \left\| X - \sum_{j \in A} C_j \right\|^2\right) \text{ avec}$$

$$\mathcal{T}_B = \left\{ A \subset [\![1, k]\!] : \bigsqcup_{j \in A} B_j = [\![1, d]\!] \right\}$$

- Invention d'un nouvel algorithme de compression avec perte (du type Lempel Ziv):
  on minimise la surface totale de recouvrement $\sum_{j=1}^{k} \mathbb{P}_S(B_j)$ des fragments, à distorsion fixée $\mathcal{R}(B, C) \leq \alpha$.
  - Ne dépend pas de la géométrie du capteur.
  - Forme aléatoire des fragments (arbitraire et dépendant de l'échantillon).

Fragmentation

- Allocation de plusieurs centroids (à supports disjoints) pour un même signal $X_i \implies$ **partitionnement/fragmentation du signal**.



Distorsion:

$$\mathcal{R}(B, C) = \mathbb{E}\left(\min_{A \in \mathcal{T}_B} \left\| X - \sum_{j \in A} C_j \right\|^2 \right) \text{ avec}$$

$$\mathcal{T}_B = \left\{ A \subset [\![1, k]\!] : \bigsqcup_{j \in A} B_j = [\![1, d]\!] \right\}$$

- Invention d'un nouvel algorithme de compression avec perte (du type Lempel Ziv):
  on minimise la surface totale de recouvrement $\sum_{j=1}^{k} \mathbb{P}_S(B_j)$ des fragments, à distorsion fixée $\mathcal{R}(B, C) \leq \alpha$.
  - Ne dépend pas de la géométrie du capteur.
  - Forme aléatoire des fragments (arbitraire et dépendant de l'échantillon).
  - Nombre de fragments paramétrable

Borne de généralisation sur la fragmentation

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B, C)_{j=1}^{k} : B_j \subset [\![1, d]\!], C_j \in [-a, a]^{B_j}, \sum_{j=1}^{k} \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B,C)_{j=1}^k : B_j \subset [\![1,d]\!], C_j \in [-a,a]^{B_j}, \sum_{j=1}^k \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

With probability at least $1-\delta$, for any $(B,C) \in \mathcal{M}(\mathcal{S})$,

$$\mathcal{R}(B,C) - \overline{\mathcal{R}}(B,C) \leq a^2 \mathcal{O}\left(\log\left(\frac{n}{\mathcal{S}K}\right) \sqrt{\frac{\mathcal{S}K^2 \log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \right),$$

avec $\overline{\mathcal{R}}(B,C)$ la distorsion empirique.

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B,C)_{j=1}^k : B_j \subset [\![1,d]\!], C_j \in [-a,a]^{B_j}, \sum_{j=1}^k \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

With probability at least $1-\delta$, for any $(B,C) \in \mathcal{M}(\mathcal{S})$,

$$\mathcal{R}(B,C) - \overline{\mathcal{R}}(B,C) \leq a^2 \mathcal{O}\left(\log\left(\frac{n}{\mathcal{S}K}\right) \sqrt{\frac{\mathcal{S}K^2 \log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \right),$$

avec $\overline{\mathcal{R}}(B,C)$ la distorsion empirique.

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B, C)_{j=1}^{k} : B_j \subset [\![1, d]\!], C_j \in [-a, a]^{B_j}, \sum_{j=1}^{k} \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

With probability at least $1 - \delta$, for any $(B, C) \in \mathcal{M}(\mathcal{S})$,

$$\mathcal{R}(B, C) - \overline{\mathcal{R}}(B, C) \leq a^2 \mathcal{O}\left(\log\left(\frac{n}{\mathcal{S}K} \right) \sqrt{\frac{\mathcal{S}K^2 \log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \right),$$

avec $\overline{\mathcal{R}}(B, C)$ la distorsion empirique.

- Bornes non asymptotiques, indépendantes de la dimension $\Longrightarrow$ application sur des images haute résolution, pas de réduction de dimension.

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B, C)_{j=1}^{k} : B_j \subset [\![1, d]\!], C_j \in [-a, a]^{B_j}, \sum_{j=1}^{k} \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

With probability at least $1 - \delta$, for any $(B, C) \in \mathcal{M}(\mathcal{S})$,

$$\mathcal{R}(B, C) - \overline{\mathcal{R}}(B, C) \leq a^2 \mathcal{O}\left(\log\left(\frac{n}{\mathcal{S}K} \right) \sqrt{\frac{\mathcal{S}K^2 \log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \right),$$

avec $\overline{\mathcal{R}}(B, C)$ la distorsion empirique.

- Bornes non asymptotiques, indépendantes de la dimension $\Longrightarrow$ application sur des images haute résolution, pas de réduction de dimension.
- Le rôle de $\mathcal{S}$ dans la borne offre une justification alternative à notre algorithme de compression avec perte.

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B, C)_{j=1}^{k} : B_j \subset [\![1, d]\!], C_j \in [-a, a]^{B_j}, \sum_{j=1}^{k} \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

With probability at least $1 - \delta$, for any $(B, C) \in \mathcal{M}(\mathcal{S})$,

$$\mathcal{R}(B, C) - \overline{\mathcal{R}}(B, C) \leq a^2 \mathcal{O}\left(\log\left(\frac{n}{\mathcal{S}K}\right) \sqrt{\frac{\mathcal{S}K^2 \log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \right),$$

avec $\overline{\mathcal{R}}(B, C)$ la distorsion empirique.

- Bornes non asymptotiques, indépendantes de la dimension $\Longrightarrow$ application sur des images haute résolution, pas de réduction de dimension.
- Le rôle de $\mathcal{S}$ dans la borne offre une justification alternative à notre algorithme de compression avec perte.
- Package R `"PatchProcess"` sur github: `https://github.com/GautierAppert/PatchProcess` , codé en `R` et `C++`

Borne de généralisation sur la fragmentation

$$\mathcal{M}(\mathcal{S}) = \left\{ (B, C)_{j=1}^{k} : B_j \subset [\![1, d]\!], C_j \in [-a, a]^{B_j}, \sum_{j=1}^{k} \mathbb{P}_S(B_j) \leq \mathcal{S}, |\mathcal{T}_{B,K}| \geq 2 \right\},$$

With probability at least $1 - \delta$, for any $(B, C) \in \mathcal{M}(\mathcal{S})$,

$$\mathcal{R}(B, C) - \overline{\mathcal{R}}(B, C) \leq a^2 \mathcal{O}\left(\log\left(\frac{n}{\mathcal{S}K}\right) \sqrt{\frac{\mathcal{S}K^2 \log(k/K)}{n}} + \sqrt{\frac{k^2 + \log(\delta^{-1})}{n}} \right),$$

avec $\overline{\mathcal{R}}(B, C)$ la distorsion empirique.

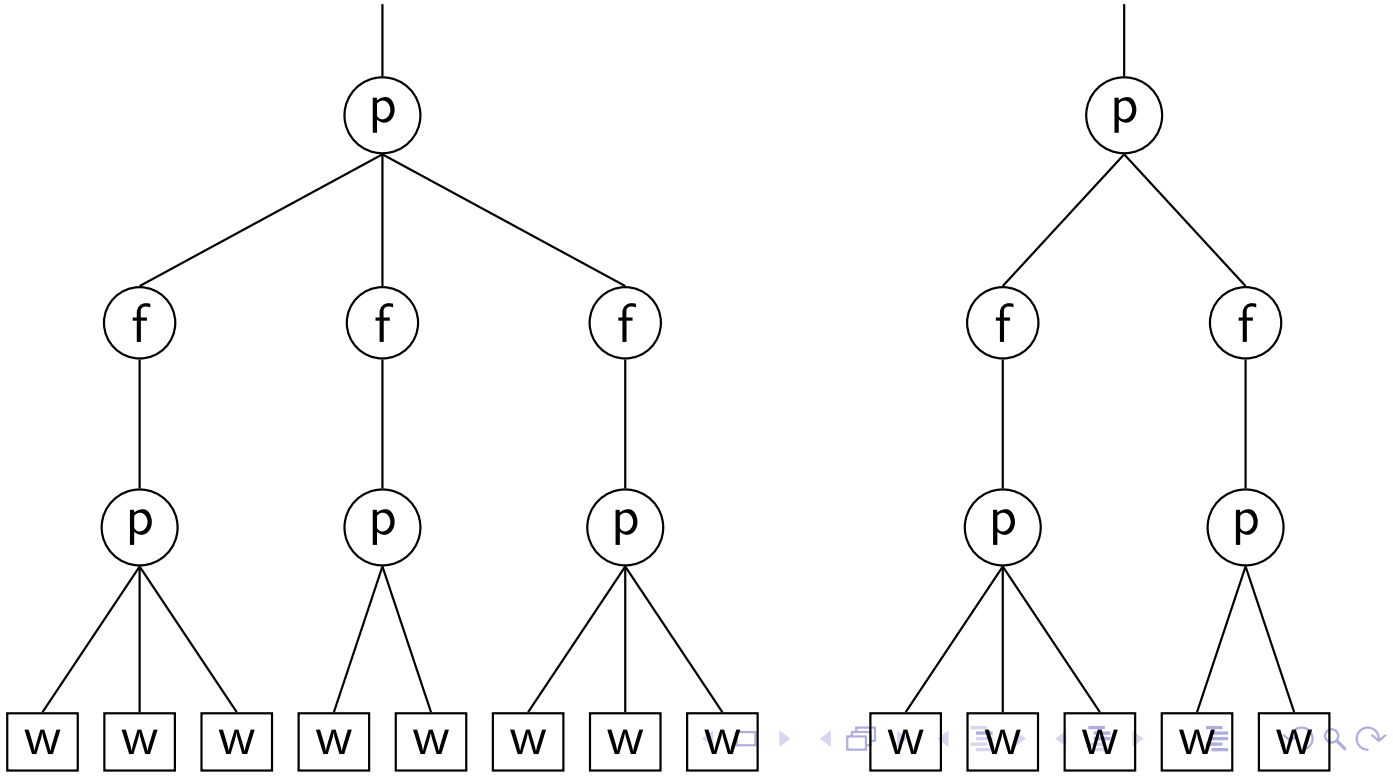
- Bornes non asymptotiques, indépendantes de la dimension $\Longrightarrow$ application sur des images haute résolution, pas de réduction de dimension.
- Le rôle de $\mathcal{S}$ dans la borne offre une justification alternative à notre algorithme de compression avec perte.
- Package R `"PatchProcess"` sur github: `https://github.com/GautierAppert/PatchProcess` , codé en `R` et `C++`

Article

From k-means to k-fragments : local vector quantization, 2021, en préparation.

Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

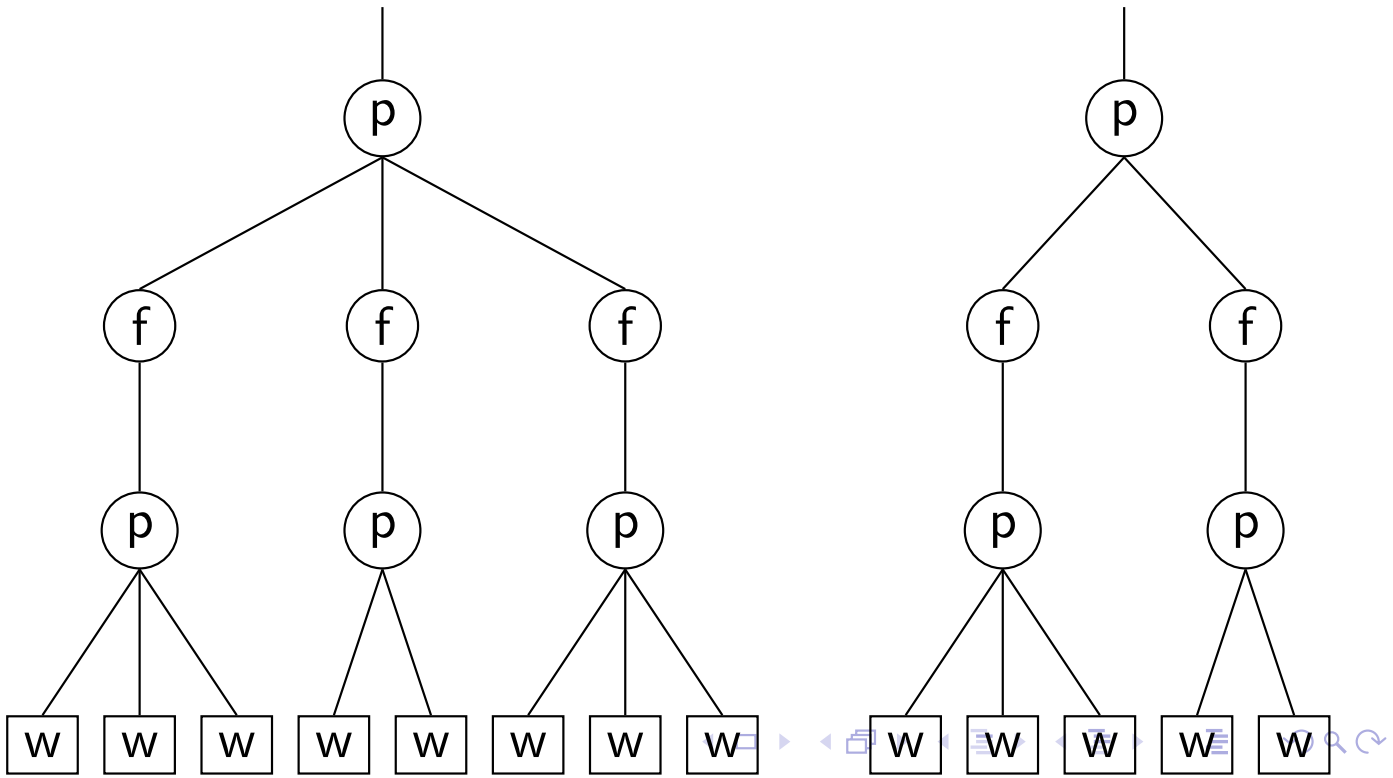


Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$



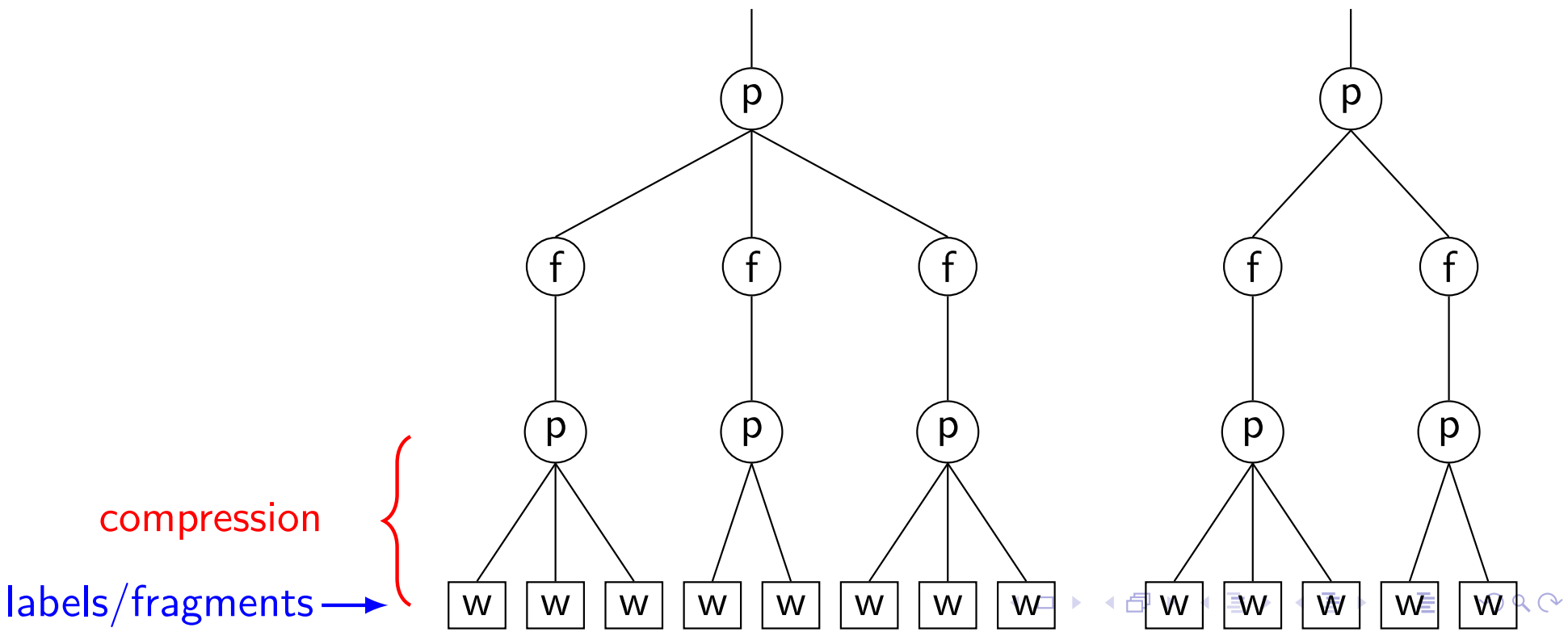
labels/fragments ⟶

Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$

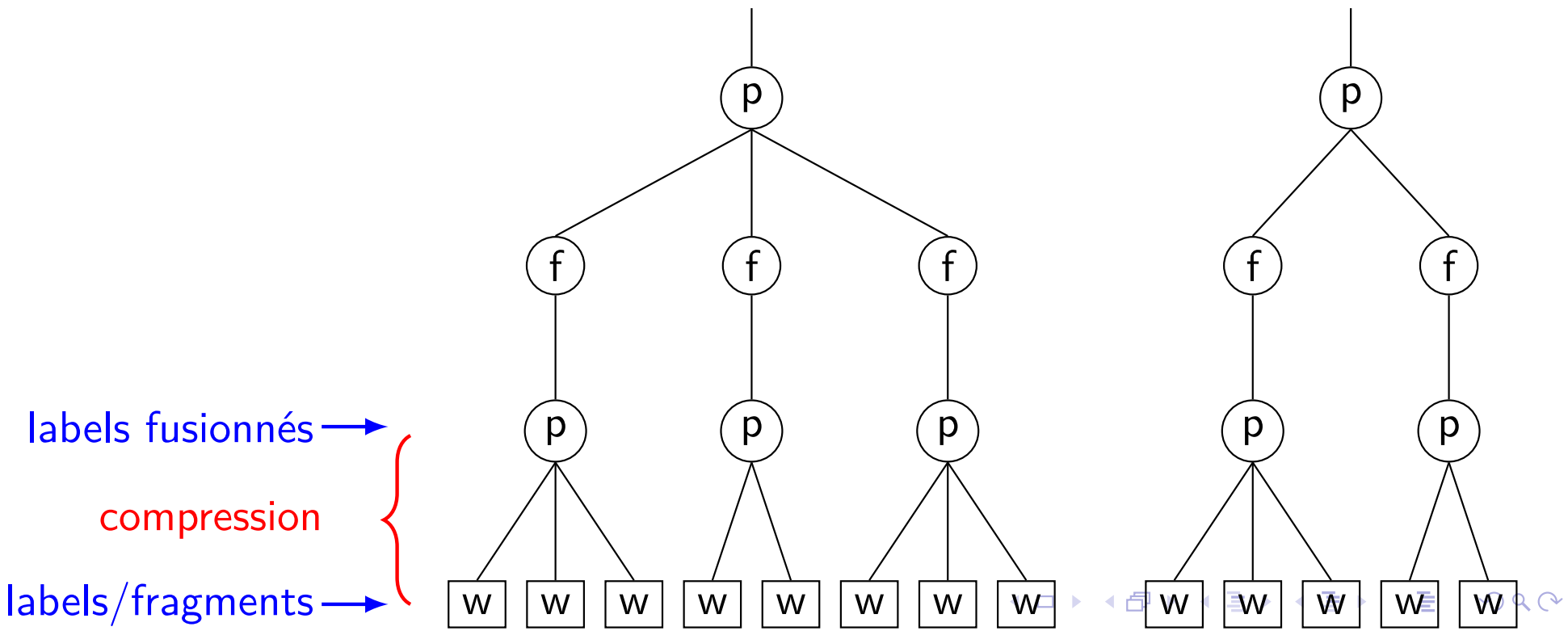


Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$

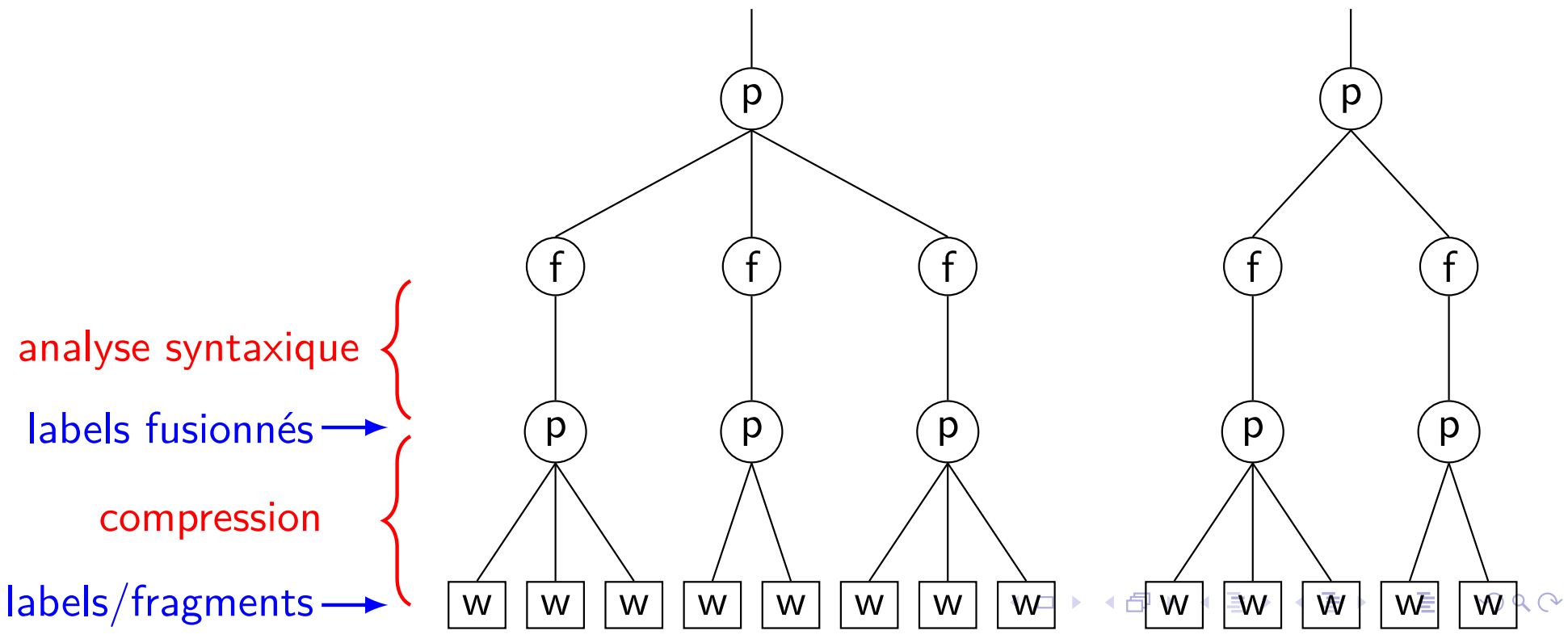


Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$

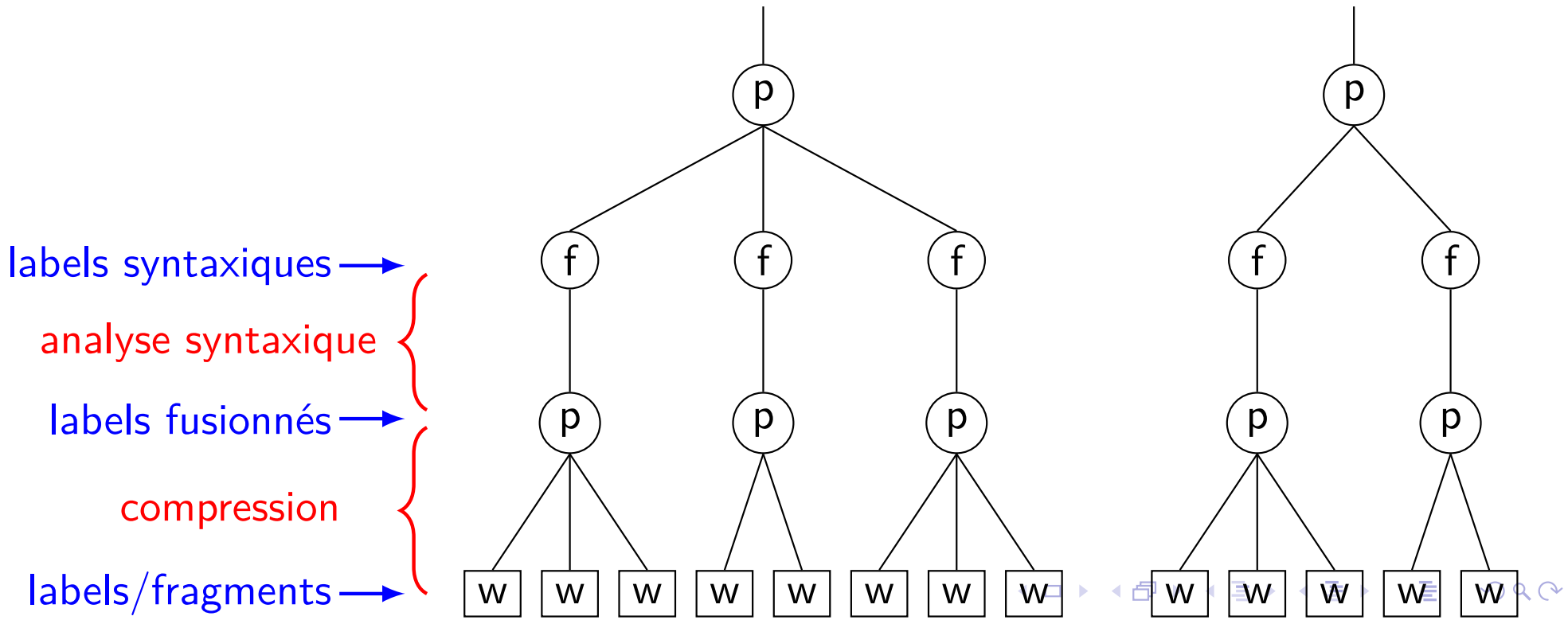


Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$

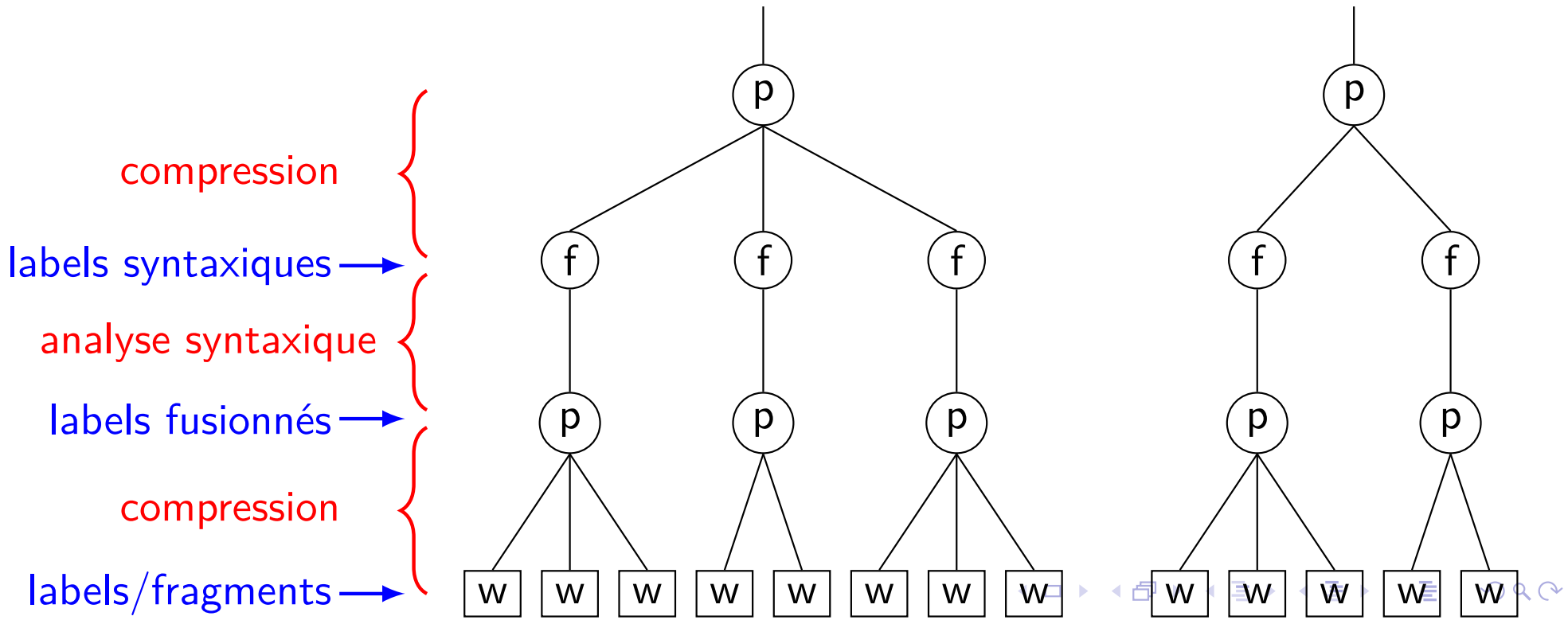


Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$

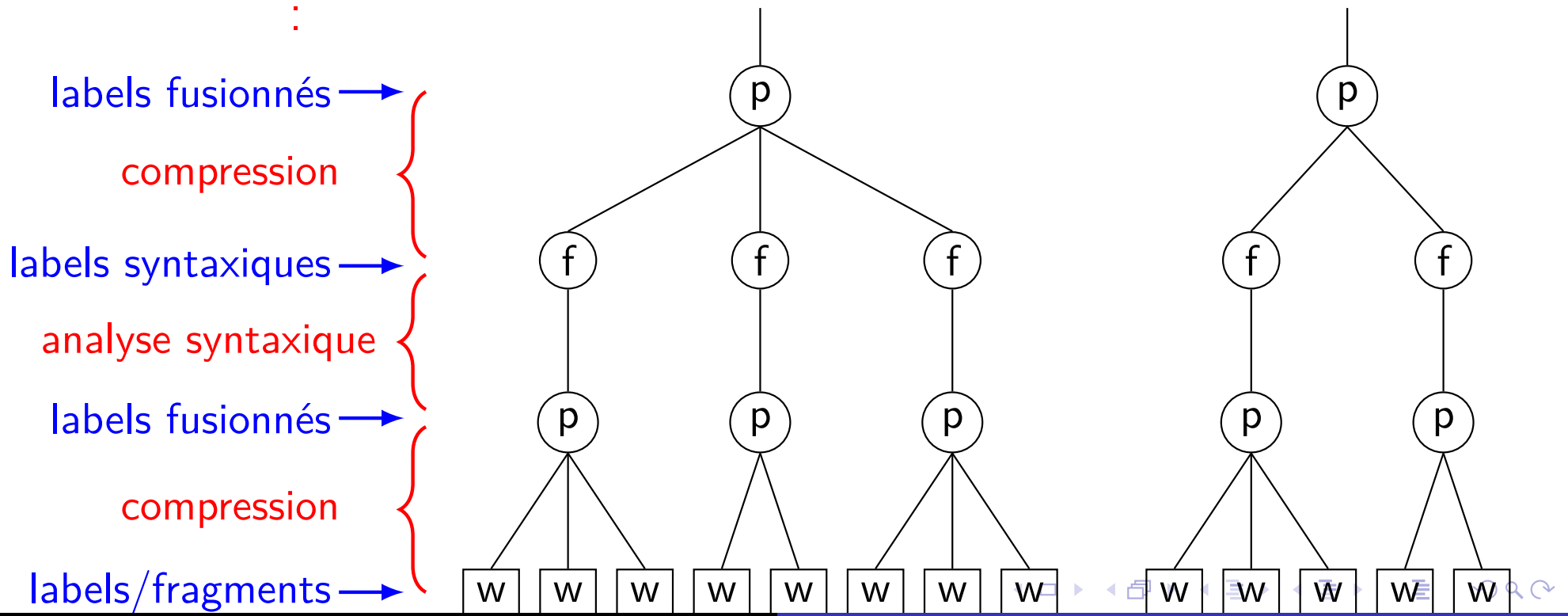


Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

Output de la fragmentation

Représentation:

$$w_{1,1} \dots w_{1,r_1} \wedge w_{2,1} \dots, w_{2,r_2} \wedge \dots \wedge w_{n,1} \dots w_{n,r_n} \wedge \quad \leftrightsquigarrow \quad \{\{w\}\wedge\}$$



Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

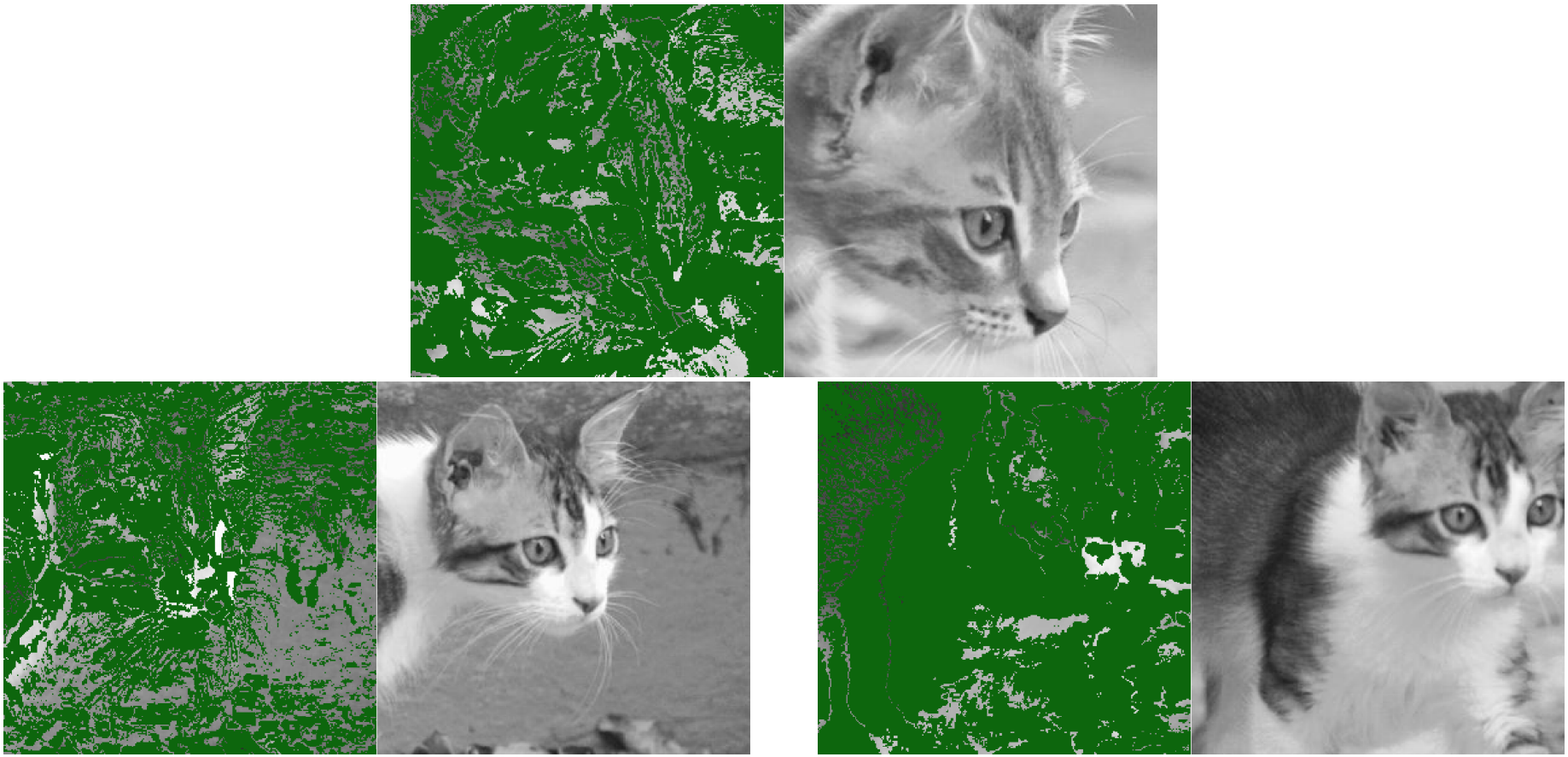


Figure: Label syntaxique permettant l'idenfication de translations.

Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

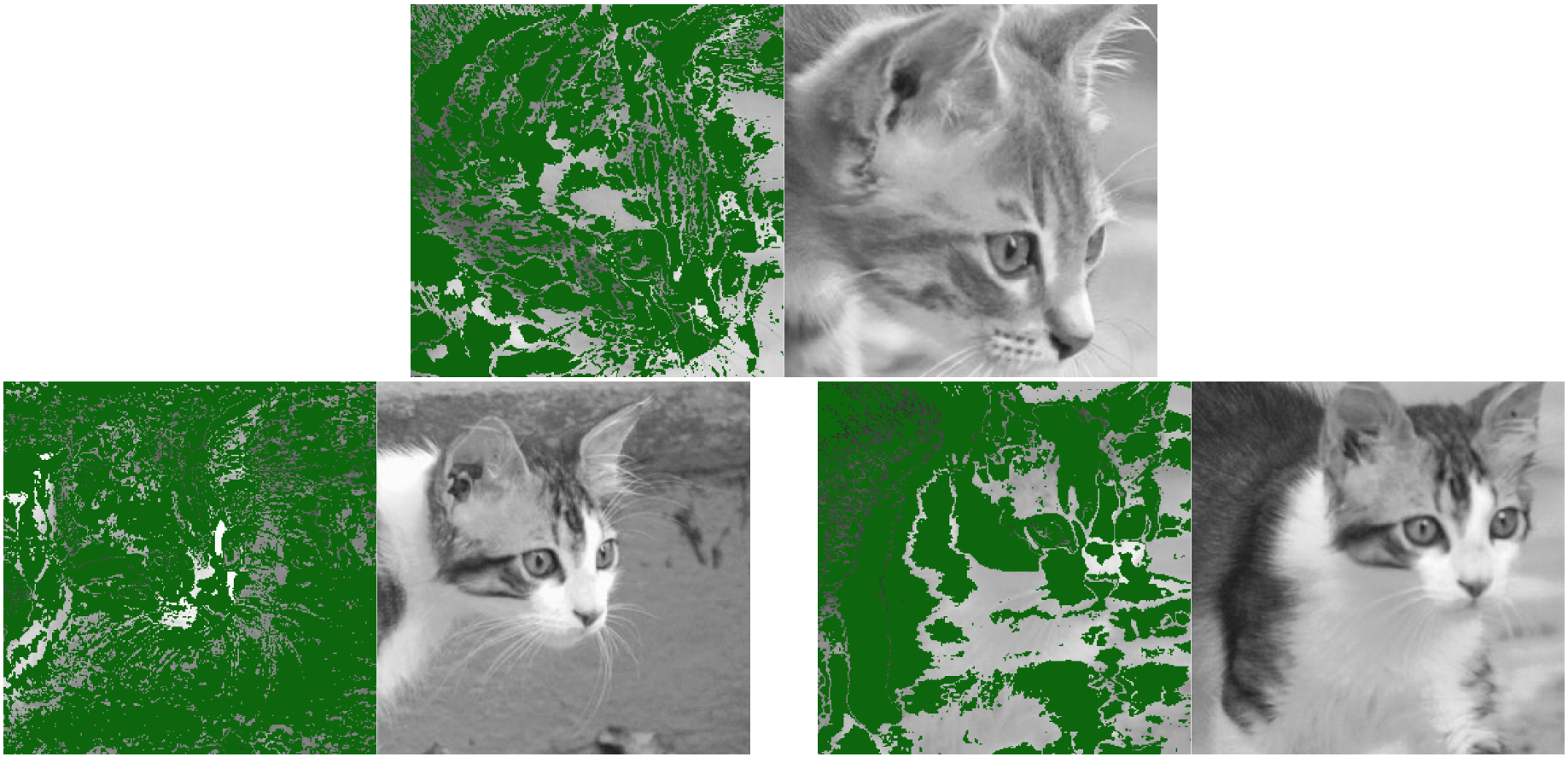


Figure: Label syntaxique permettant l'idenfication de translations.

Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

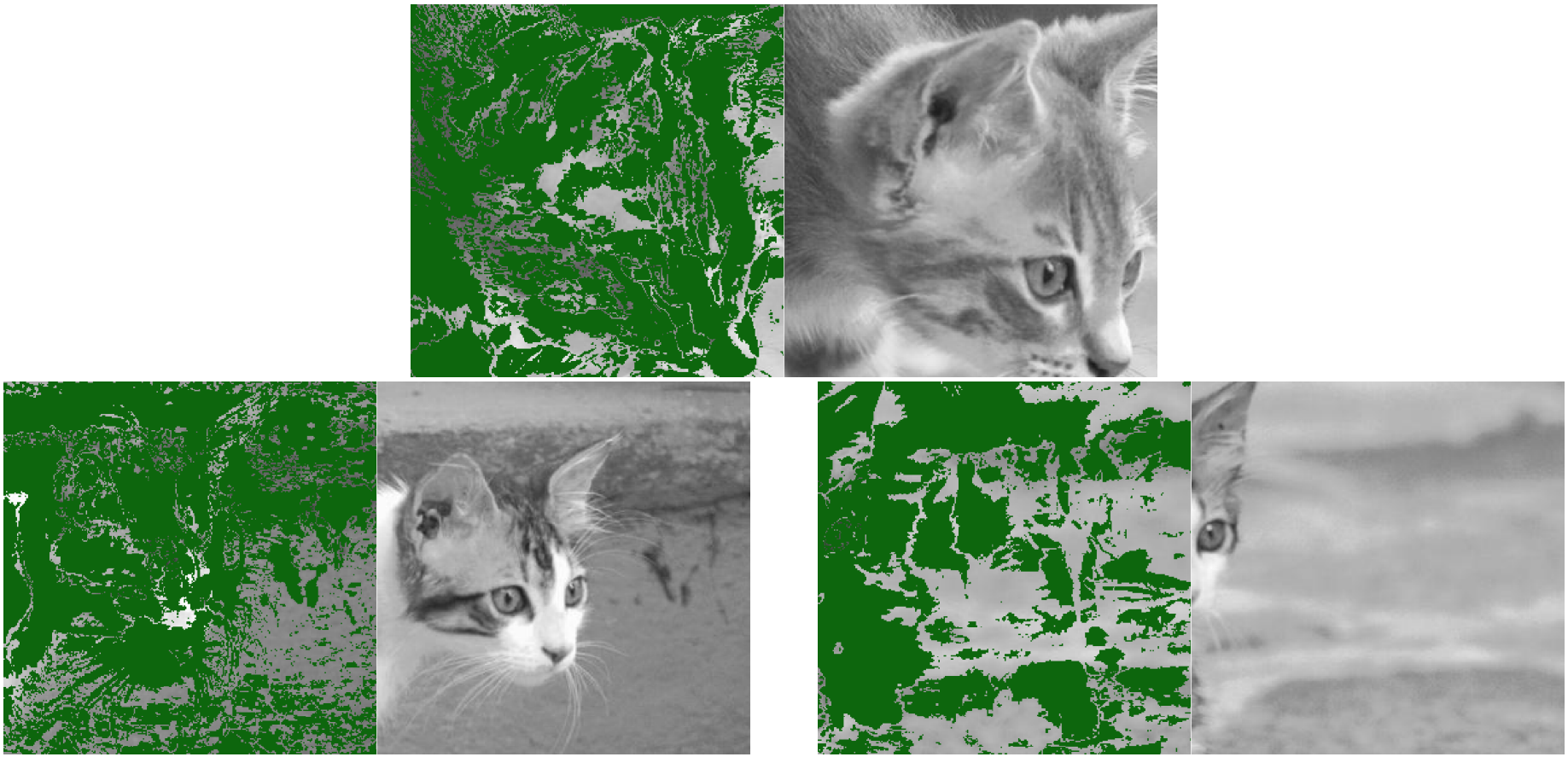


Figure: Label syntaxique permettant l'idenfication de translations.

Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

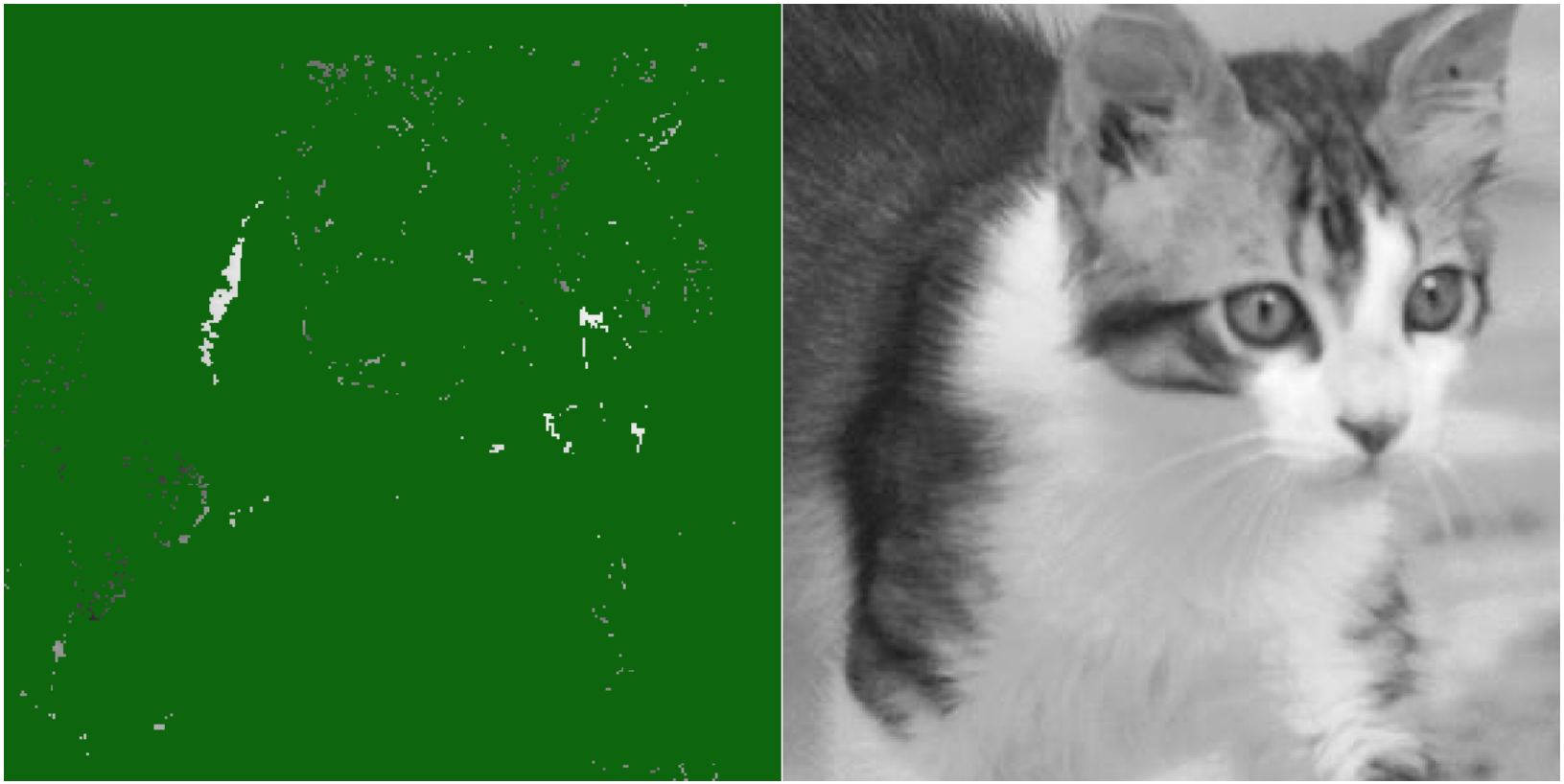


Figure: Identification d'une rotation à l'aide du label syntaxique.

Article

Syntax analysis for unsupervised signal classification, 2021, en préparation

Analyse syntaxique d'un signal composé d'ensembles aléatoires de labels, compression à base de grammaires

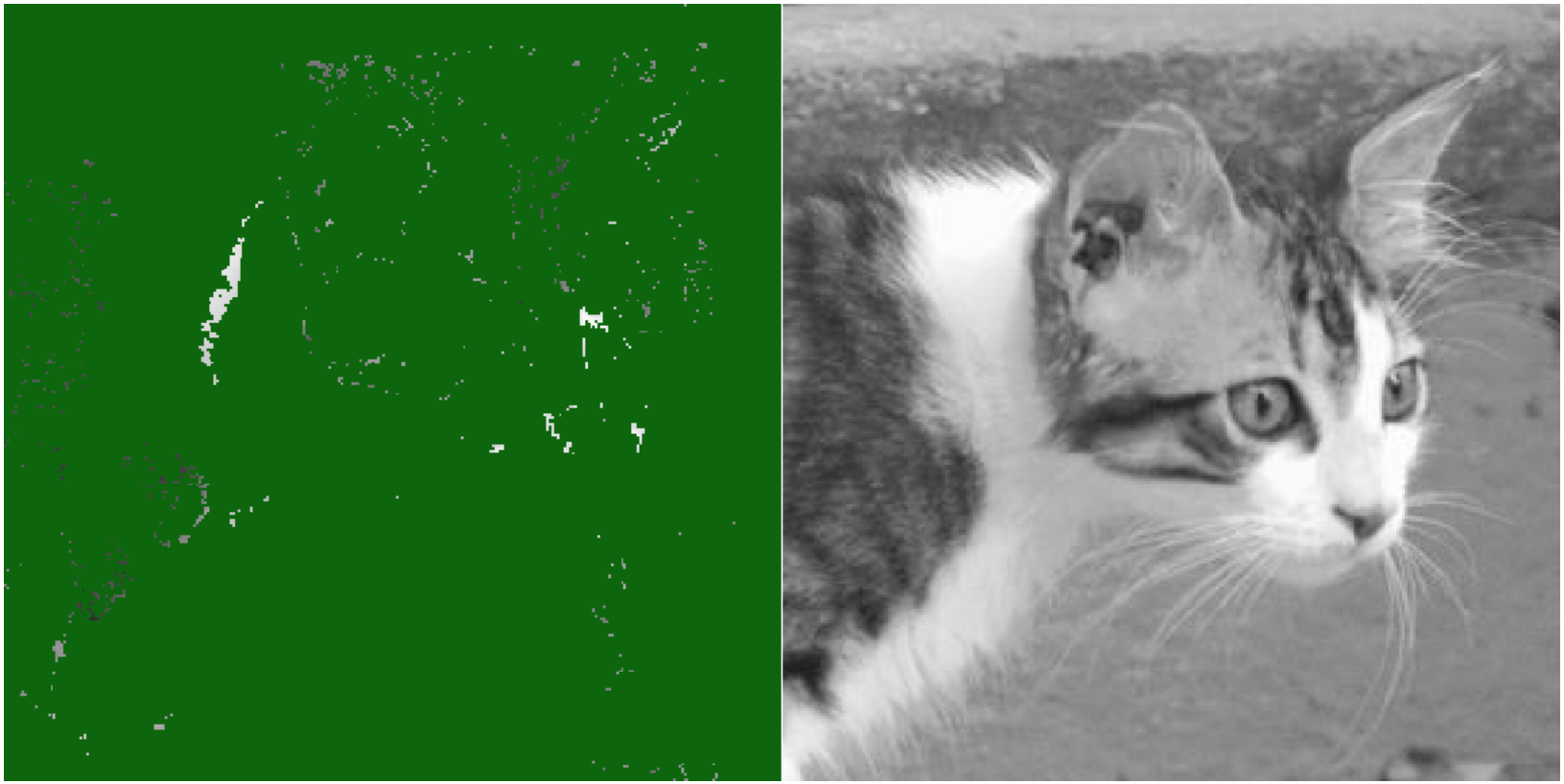


Figure: Identification d'une rotation à l'aide du label syntaxique.

Article

Syntax analysis for unsupervised signal classification, 2021, en préparation

Enseignements effectués

Thèmes enseignés

- Statistiques inférentielles (appli avec R) et intro au Machine learning (L2, L3, M1, 2ieme année école d'ingénieur).
- Probabilités et théorie de la mesure (L3, 1ère année école d'ingénieur).
- Analyse réelle (L1, L2).

Enseignements effectués

Thèmes enseignés

- Statistiques inférentielles (appli avec `R`) et intro au Machine learning (L2, L3, M1, 2ieme année école d'ingénieur).
- Probabilités et théorie de la mesure (L3, 1ère année école d'ingénieur).
- Analyse réelle (L1, L2).

Enseignements (travaux dirigés)

- **ENSAE ParisTech**:
  - Statistiques 1 et 2, 2ème année
  - Probabilités, 1ère année
  - Intro au Machine Learning, 2ème année
- **Université Paris Saclay**, Orsay:
  - Mesure, intégration et Probabilités, L3 maths (magistère)
  - Modélisation Statistique, 2ème année ENSTA ParisTech
  - Analyse réelle et Probabilités, L1 Bio
- **Université Paris 1 Panthéon-Sorbonne**
  - Statistiques, L2 MIASHS
  - Probabilités, L3 MIASHS
  - Méthodes Numériques, L2 MIASHS
  - Techniques de Calcul, L1 MIASHS