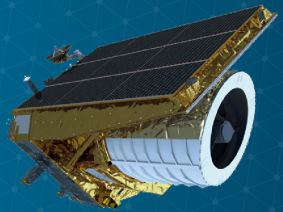
Du machine learning pour calculer la fonction de transfert du télescope spatial Euclid



Jérôme Odier

Journées Informatiques, Jeudi 26 Septembre 2024

LPSC Grenoble

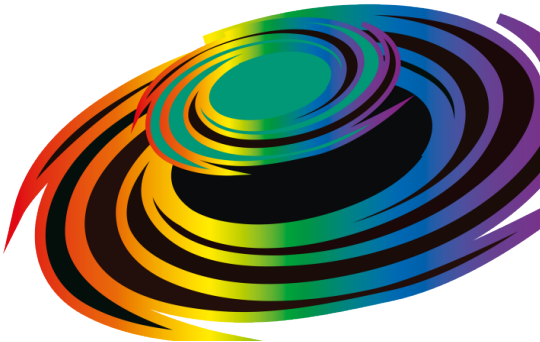
Pour l'équipe VMPZ-ID : Gaël Alguero and Juan F. Macías-Pérez

Euclid

Problématique

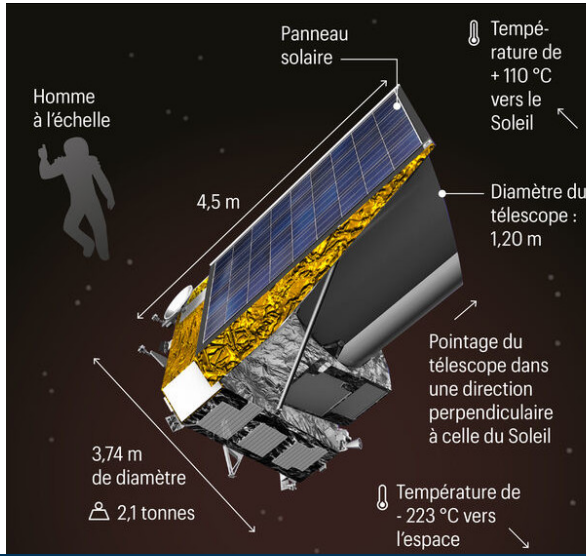
Les cartes auto-organisatrices

Une bibliothèque open-source



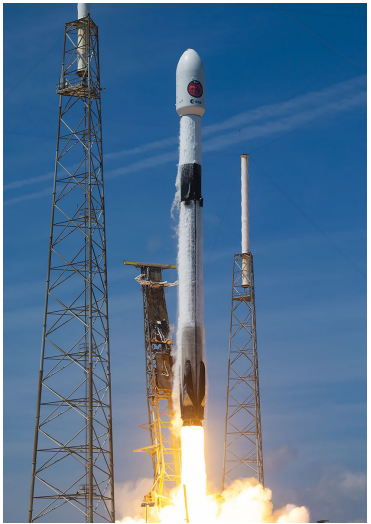
Le télescope Euclid

- Objectif de la mission :
  Effectuer une carte 3D (longitude, latitude et resdhift) des galaxies la plus précise au monde.
- Améliorer la compréhension des problématiques liées à la matière noire et à l'énergie noire.
- Contraindre les paramètres libres de la cosmologie.
- Observation de milliards de galaxies sur 30% du ciel ($\sim$15 000 deg^2).

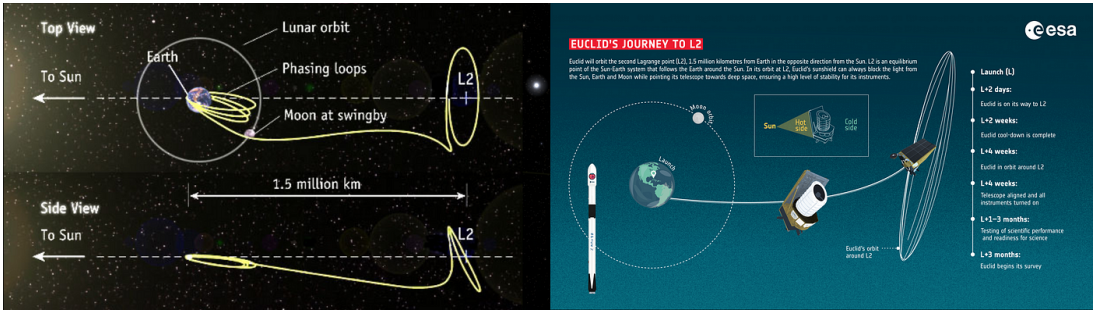


Le télescope Euclid

Lancement d'Euclid le 1er juillet 2023 par une Falcon 9 à Cap Cavaveral.



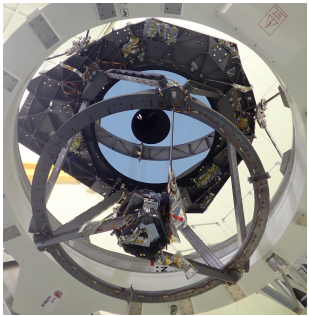
Le télescope Euclid



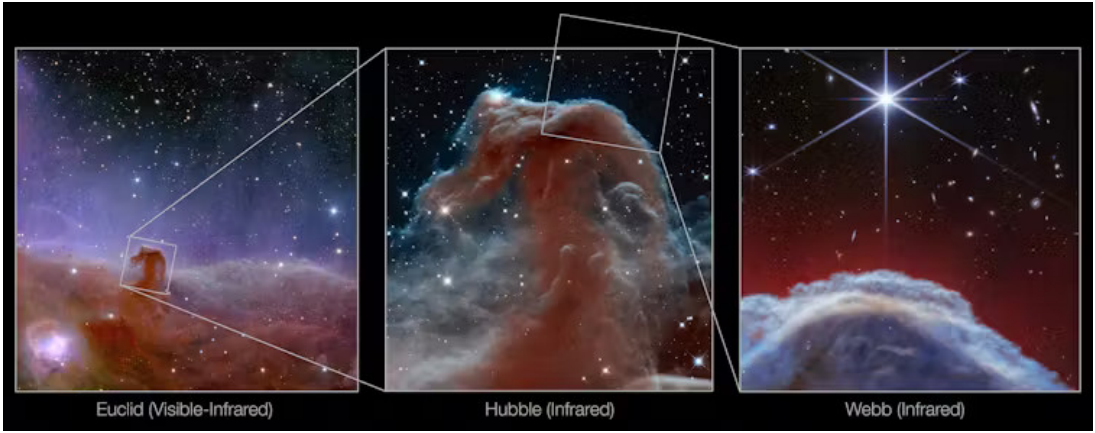
Euclid est placé au point de Lagrange L2 (comme le JWST).

Le télescope Euclid

- Diamète de la pupille: 1.2m (Hubble: 2.4m, JWST: 6.5m).
- Focale: 24.5m, instrument extrêmement lumineux.
- Champ 0.47°: Euclid peut voir d'un seul coup une région de taille similaire à la pleine lune !
- Imageur visible (VIS) et spectro-imageur proche infrarouge (NISP) en 19200x19200 pixels.

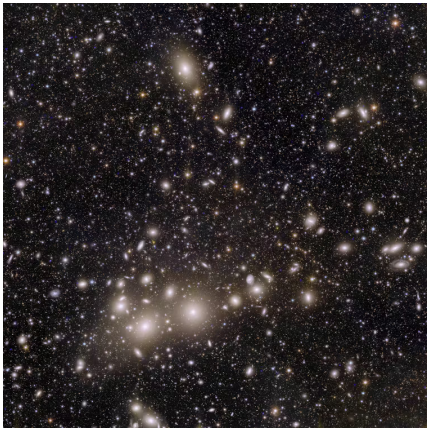


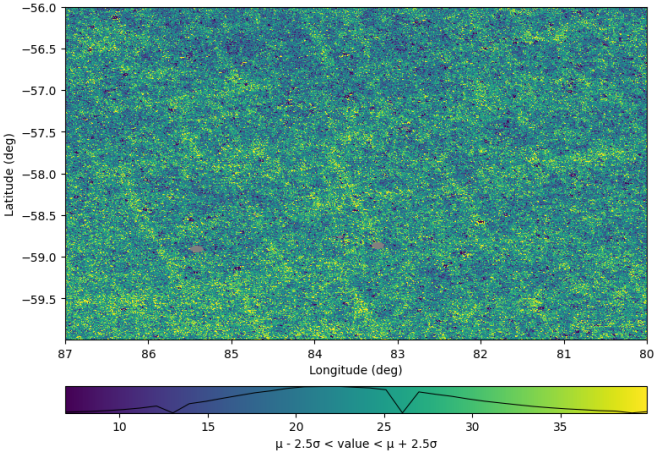
Le télescope Euclid



Comparaison des champs de : Euclid, Hubble et JWST.

Une myriade de galaxies



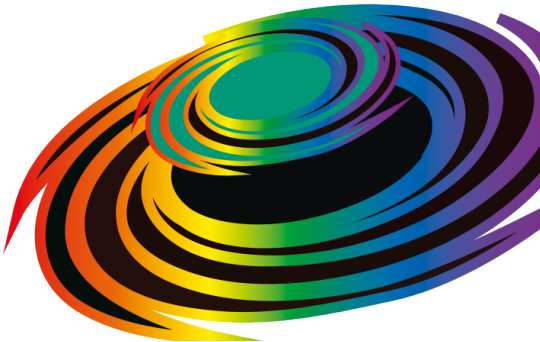


Euclid

Problématique

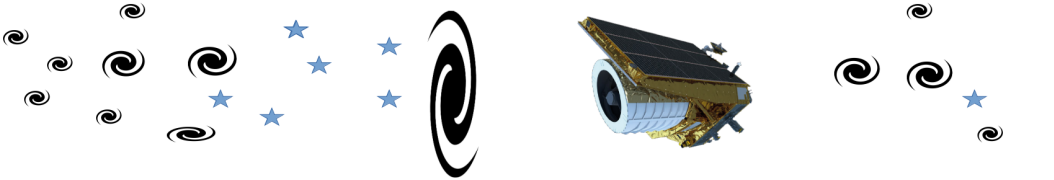
Les cartes auto-organisatrices

Une bibliothèque open-source



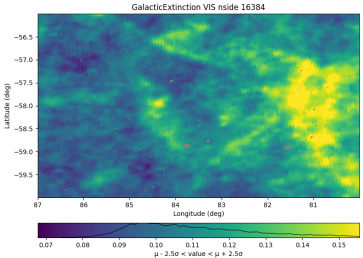
Propriétés affectant la densité de galaxies

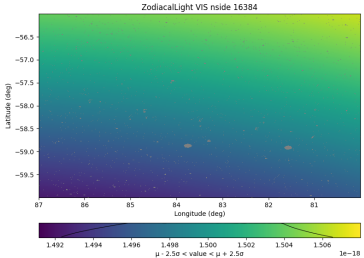
- Ensemble de facteurs (propriétés du ciel et performances instrumentales) impactant la détection de galaxies ⇒ erreurs systématiques.

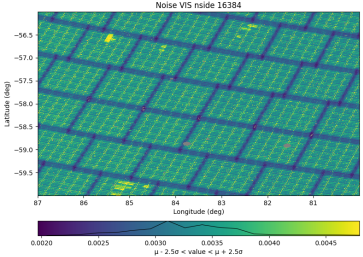


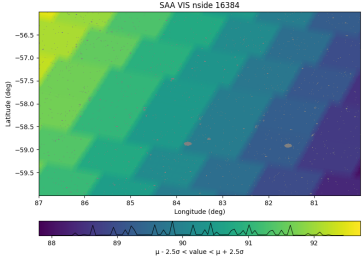
⇒ Calculs de physique fortement impactés par les systématiques.

Exemples de systématiques ($\sim$20 au total)

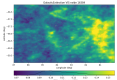
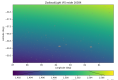
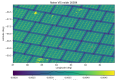
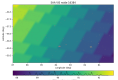








Fonction de transfert d'Euclid

$$\xi = H\left(\text{  }, \text{  }, \text{  }, \text{  }, \cdots \right)$$

Le constraste ξ représente l'excès ou le déficit en galaxies observées, par rapport à la densité moyenne dans l'univers, en fonction des systématiques.

Plusieurs solutions:

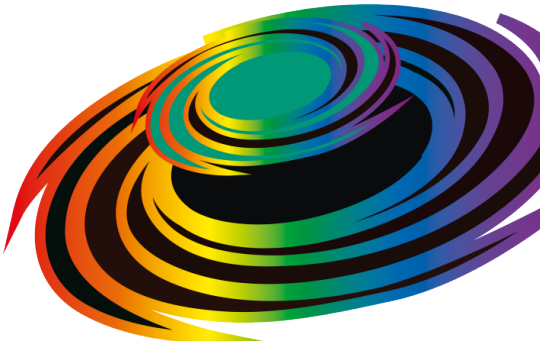
- Des régressions linéaires itératives.
- Des régressions linéaires avec pénalités L1 et L2.
- **Une méthode non linéaire basée sur des cartes auto-organisatrices (SOM).**

Euclid

Problématique

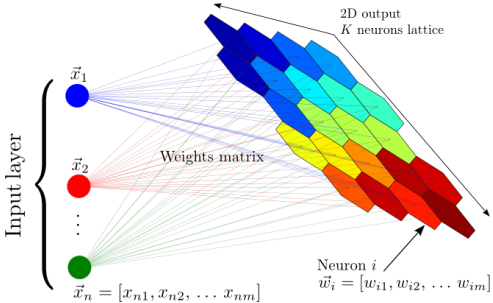
Les cartes auto-organisatrices

Une bibliothèque open-source



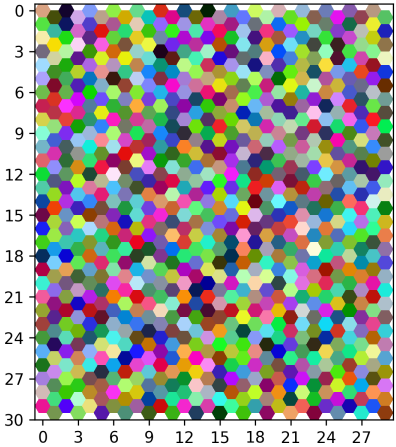
Un exemple avec des couleurs

	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	...
R	0	255	255	0	0	255	0	255	128	0	128	0	255	255	255	...
G	0	255	0	255	0	255	255	0	128	128	0	0	69	165	215	...
B	0	255	0	0	255	0	255	255	0	0	128	128	0	0	0	...

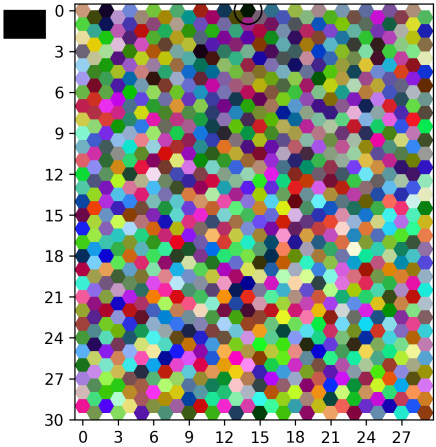


- Non supervisé.
- Une seule couche de neurones.
- Principe : encoder les informations d'entrée (systématiques) dans un espace latent de petite taille en limitant la perte d'information.

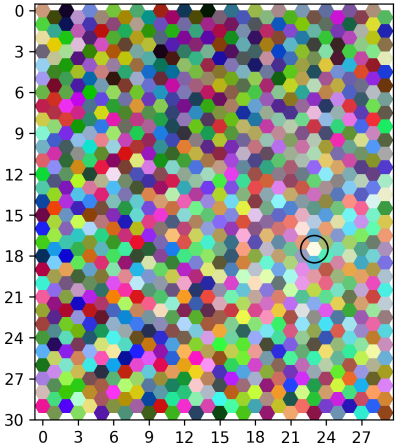
Principe du SOM



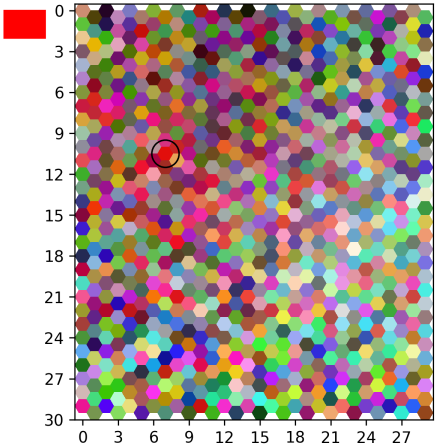
Principe du SOM



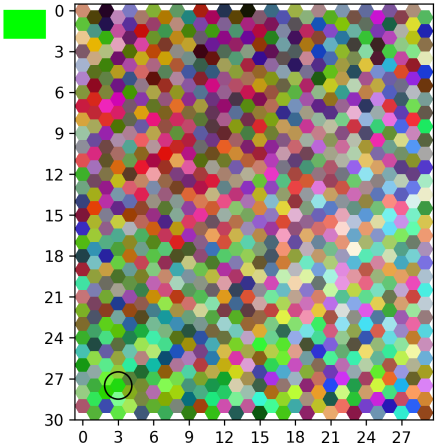
Principe du SOM



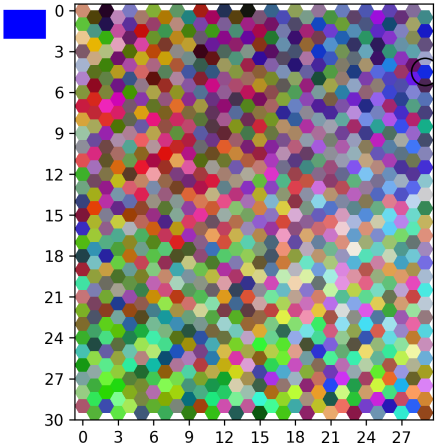
Principe du SOM



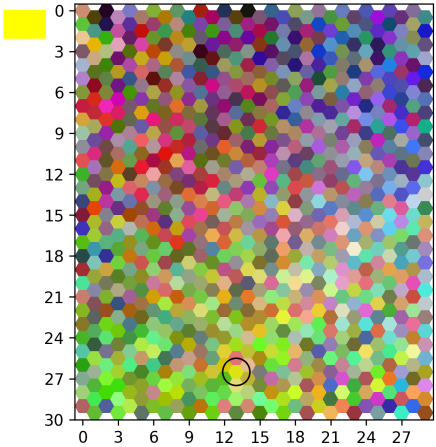
Principe du SOM



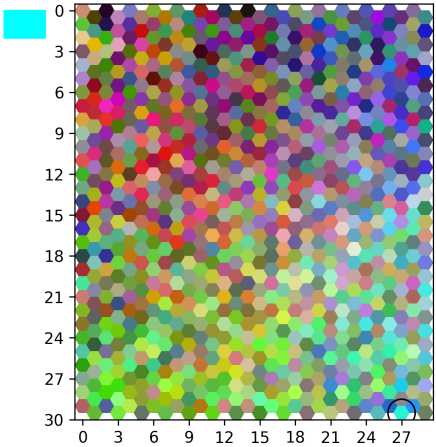
Principe du SOM



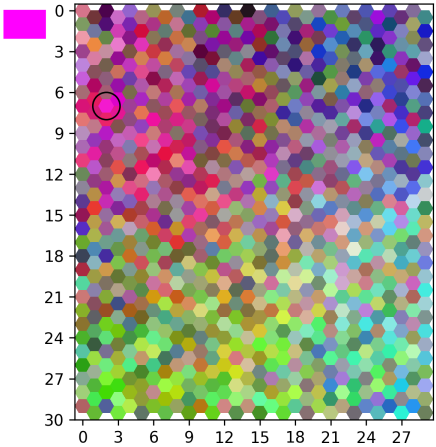
Principe du SOM



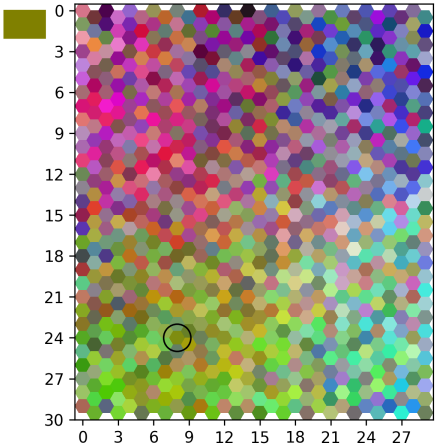
Principe du SOM



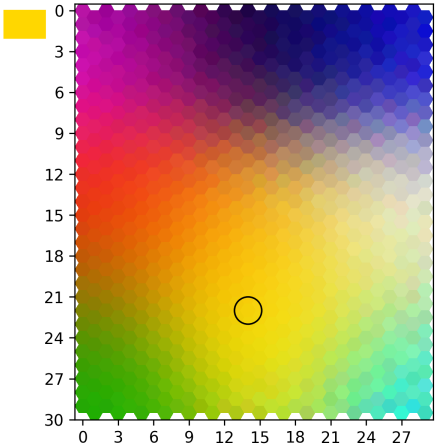
Principe du SOM



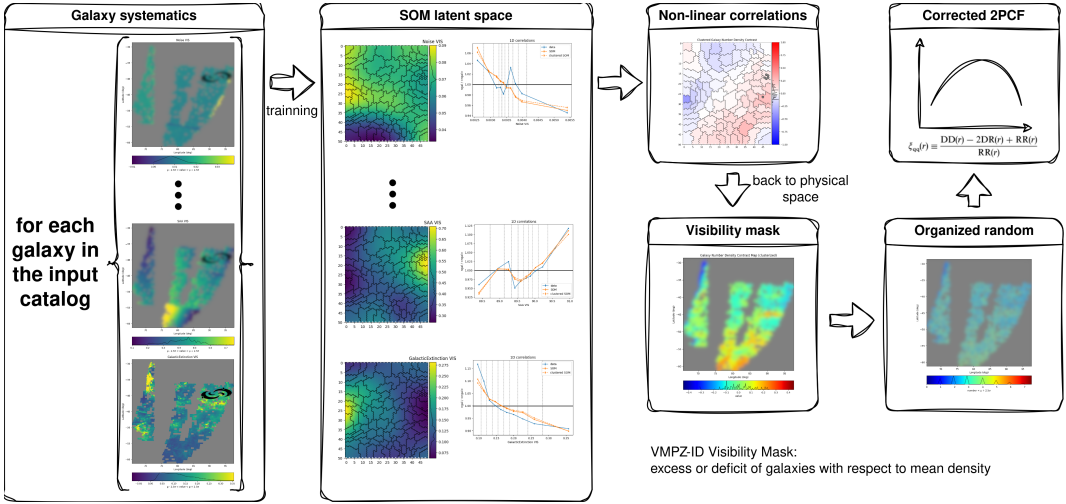
Principe du SOM



Principe du SOM



Le modèle ou la grande moulinette du SOM



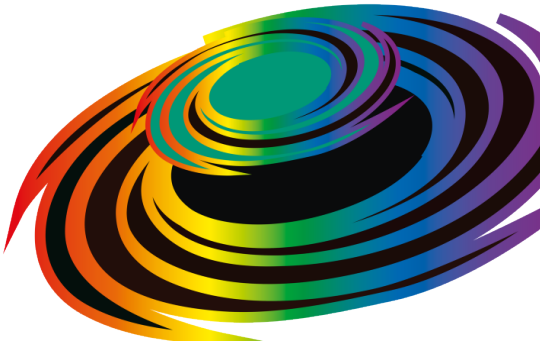
VMPZ-ID Visibility Mask:
excess or deficit of galaxies with respect to mean density

Euclid

Problématique

Les cartes auto-organisatrices

Une bibliothèque open-source



L'outil et les contraintes techniques

- Développement d'un outil générique pour décontaminer la détection de galaxies des effets systématiques, et pour simuler des catalogues.
  https://gitlab.in2p3.fr/lpsc-euclid/decontamination
- Contraintes:
  - plusieurs méthodes de décontamination,
  - python3 et peu de dépendances,
  - doit fonctionner sur CPU et GPU,
  - doit fonctionner à mémoire constante,
  - doit pouvoir être entrainé sur des milliards de galaxies,
  - doit avoir >70% de code coverage ($\rightarrow$ imposé par le CNES),
  - doit être validé par SonarCube (0 bug et vulnérabilité $\rightarrow$ imposé par le CNES),
  - documentation auto-générée.

Solution techniques

- Utilisation du compilateur *Just-In-Time* Numba avec une surcouche maison pour écrire un unique code CPU et GPU/Cuda (voir slide suivant).
- Utilisation systématique de générateurs pour l'accès aux données et développement d'algorithmes tous parallélisables.
- Tests unitaires et code coverage avec *pytest* via GitLab CI au CC-IN2P3 et déploiement sur PyPi.
- Code auto-documenté, documentation générée via GitLab CI au CC-IN2P3 et déploiement sur GitLab Pages.
- Utilisation du SonarCube (développement) du CC-IN2P3 et de la collaboration Euclid (intégration).

Surcouche numba

- Un décorateurs maison (pour les kernels et les fonctions device).
- Balises CPU/GPU pour simplifier la maintenabilité et éviter la redondance.
- Le trainning sur GPU (NVIDIA T2) ∼100x plus rapide qu'en single thread sur CPU.

```
################################################################
# !--BEGIN-CPU--

sigma = sigma0 * asymptotic_decay_cpu(cur_epoch, n_epochs)

for i in nb.prange(vectors.shape[0]):

    _train_step2_xpu(
        numerator,
        denominator,
        weights,
        topography,
        vectors[i],
        sigma,
        mn
    )

# !--END-CPU--
################################################################
# !--BEGIN-GPU--

i = jit.grid(1)

if i < vectors.shape[0]:
```

Conclusion : un outil générique visant à être open source

- Développement (depuis l'été 2023) d'un outil générique pour décontamination de la détection des galaxies des effets systématiques et pour simuler des catalogues.
- Nous avons pu faire une première estimation de la fonction de transfert d'Euclid (donnée non publique).
  - Besoin des 15 000 deg2 observés pour un résultat final dans 6 ans.
- Nous avons essayé de respecter les recommandation F.A.I.R.
- Quid de la licence ?

Thanks for your attention!

Deux implémentations du SOM

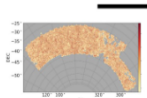
On-line SOM

- Vecteurs d'entrée présentés séquentiellement
- Espace latent du SOM mis à jour à chaque itération
- Non parallélisable
- $\vec{w}_k(t+1) = \vec{w}_k(t) + \alpha(e) h_{ck}(e) [\vec{x}(t) - \vec{w}_k(t)]$

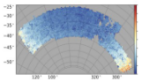
Batch SOM

- Vecteurs d'entrée présentés simultanément
- Espace latent du SOM une fois par époque
- Autement parallélisable
- Idéal pour une grosse quantité de données
- $\vec{w}_k(t+n) = \frac{\sum_{t'=t}^{t+n} h_{ck}(t') \vec{x}(t')}{\sum_{t'=t}^{t+n} h_{ck}(t')}$

Décontamination itérative



Galaxy density

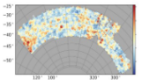


Template 1

+

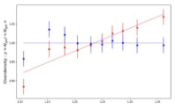
⋮

+

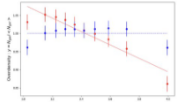


Template k

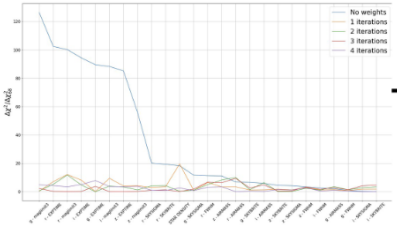
1D correlation figures



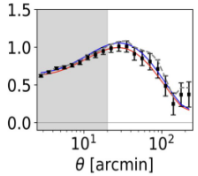
+ ... +



ISD principle



Apply weights



Decontaminated sample