PÔLE

MACHINE LEARNING / DEEP LEARNING

du CeSAM (LAM)

PRÉSENTATION
&
EXEMPLES DE
DEUX PROBLÉMATIQUES

Morgan Gray, Simon Conseil, Jean-Charles Lambert,
Jean-Charles Meunier, Christian Surace, Didier Vibert

Missions du pôle

Personnels du pôle

Infrastructure de calcul intensif (HPC)

École thématique CNRS Astro-Informatique 2021

Anatomie d'un réseau de neurones

2 exemples de problématiques

- Régression :
estimation de paramètres des PSFs en optique adaptative
pour l' E-ELT (**APPLY**) *ou* pour un nano-satellite (**AZIMOV**)

- Classification :
indice de fiabilité des estimations de redshifts spectroscopiques
du pipeline 1D du satellite **EUCLID**
ou classification Etoiles/Galaxies/Quasar à partir des spectres **EUCLID**

MISSIONS du PÔLE ML/DL

Pôle créé en novembre 2020

- **Apporter une expertise**

  sur les méthodes de Machine Learning (ML) et de Deep Learning (DL)
  pour les ***problématiques astrophysiques et instrumentales***,
  au sein des ***différents projets et équipes*** du LAM :
  - développement de codes, suivis & conseils,
  - ressources personnels (2 ETP en permanents & 2 futurs CDD),
  - ressources informatiques (18 GPUs sur le cluster du LAM & aide à l'utilisation des GPUs du Mésocentre Aix-Marseille Université)

- **Faciliter les interactions**

  entre les équipes au sein même du LAM,
  établir des liens avec les équipes d'autres laboratoires (équipe QARMA du LIS) & des entreprises (start-up SpaceAble...) travaillant sur ces méthodes de ML/DL

- **Animer des rencontres, transmettre des annonces de formation**

  sur des thématiques relatives au ML/DL

- **Organisation d'une école thématique CNRS AstroInformatique**

  2 éditions 2018 & 2021 (prochaine en 2023)

  Page du pôle ML/DL : https://projets.lam.fr/projects/mldl/wiki

PERSONNELS du PÔLE ML/DL

Morgan Gray (*100 %*) IR, responsable pôle ML / DL

Développement de codes numériques pour les projets APPLY (GRD) ; EUCLID (GECO)
Conseil/support aux projets BigSF (GECO) ; AZIMOV (GRD)

Simon Conseil (50 %) IR (*arrivé le 1er fev. 2022*)

Développement de codes numériques pour le projet EUCLID (GECO)

Jean-Charles Lambert (*5 %*) IR, responsable pôle Infrastructure

Développement & Maintenance des GPUs du cluster de calcul du LAM

Jean-Charles Meunier (25 %) IR, technologies du BigData

Pré traitement & visualisation de données multisources pour le projet BigSF (A. Zavagno)

Christian Surace (*5%*) IR, responsable du CeSAM

Organisation de l'école thématique CNRS Astro-Informatique (2018 & 2021)

Didier Vibert (5 *%*) IR, responsable pôle TARDIS

Expertise scientifique pour EUCLID (GECO)

François-Xavier Dupé (50 %) MdC, détachement (LIS-QARMA)

Développement de codes numériques pour le projet BigSF (GECO)

Charleston Chauvet (50 *%*) CDD-IE 2 ans (*début le 1er juin 2022*)

Collaboration GRD / start-up SpaceAble (plan de relance France 2020-2022)

Raissa Camelo (50 *%*) CDD-IE 2 ans (*début le 1er sept. 2022*)

Collaboration GRD / start-up SpaceAble (plan de relance France 2020-2022)

INFRASTRUCTURE DE CALCUL INTENSIF

Matériel

Système linux Almalinux 8
Gestionnaire de jobs **SLURM**
Total de **42** noeuds de calcul (**1120** cpus)
→ calcul **CPU** : **2.5 M** heures (année 2021)

6 noeuds dédiés au **Machine Learning / Deep Learning** (**18** cartes GPU)
→ calcul **GPU** : **75 K** heures (année 2021)

Partitions de calcul slurm : **batch** : jobs séquentiels + parallèles partagés / OpenMP
mpi : jobs parallèles distribués / MPI
mem : jobs interactifs + jobs mémoire
gpu : jobs ML/DL & cuda

Stockage Systeme TrueNAS
480 TB d'espace partagés NFS

Réseau d'interconnexion **1** Gigabit/s ethernet entre les nœuds
100 GB/s réseau faible latence infiniband
pour le calcul parallèle MPI

INFRASTRUCTURE DE CALCUL INTENSIF

Logiciel

Compilateurs GNU et Intel

Librairies scientifiques

Environnement conda

Python

Jupyter notebook

Travaux python interactifs sur le cluster

Matlab compiler runtime

Execution de jobs MATLAB sur le cluster

IDL

Environment pour le Machine Learning /Deep Learning

CUDA 11.5

Tensorflow

Pytorch

Page du pôle HPC : https://projets.lam.fr/projects/cluster-de-calcul-du-lam/wiki#Presentation

ÉCOLE THÉMATIQUE CNRS ASTROINFORMATIQUE 2021

Objectifs

- **Former des membres** de la communauté INSU Astrophysique aux méthodes d'analyse des grandes masses de données avec les outils les plus récents
- **Faire intervenir** des formateurs reconnus pour leurs expertises en ML/DL
- **Définir** des projets et **établir** des liens plus forts entre les communautés informatique et astrophysique afin d'assurer une meilleure cohésion interdisciplinaire
- **Réaliser** un projet concret au travers d'un **Hackathon**

Instituts scientifiques : **INSU / IN2P3 / INS2I** *Nb de participants* : 35 (dont 23 CNRS)

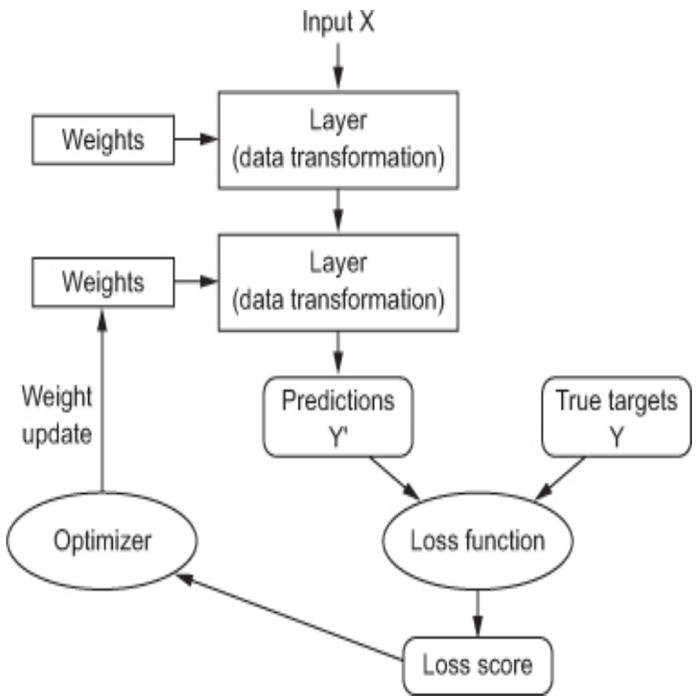
Déroulement

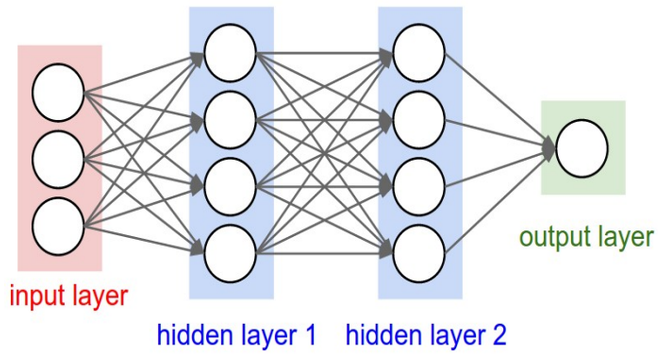
- **Semaine 1** : Alternance de **cours**, **TP** & **discussions informelles**
  Classification Supervisée / Non supervisée Pré traitement des données
  Réseaux de neurones (bases, réseaux convolutifs, récurrents, probabilistes...)
- **Semaine 2** : - Hackathon sur 2 sujets différents (petits groupes et encadrés par des experts)
  - Recherche & expérimentation, utilisation du super calculateur du CNRS Jean Zay
  - Séances d'approfondissement de certaines notion en DL, publication en préparation

Bilan très positif : introduction graduelle des notions, interactions & collaborations efficaces
Prochaine session : juin 2023

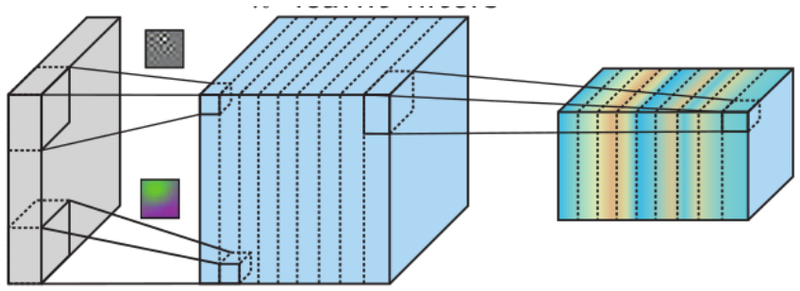
Page de l'École Thématique : https://astroinfo2021.sciencesconf.org/program

ANATOMIE D' UN RÉSEAU DE NEURONES





Couches Denses



Couche de Convolution

Couche (dense / de convolution) Tenseur → Tenseur

neurones, poids, # filtres, kernels, strides, padding,
fonctions d'activation, initialiseurs…

Fonction de Perte avec Labels "Prédits" & Labels "Vrais"

calcul d'une Distance / Mesure de perte

Optimizer Algorithme de rétroprogation du gradient des erreurs

permet de *déterminer* les valeurs des paramètres du réseau
minimisant la Fonction de Perte

Métrique évaluation de l' évolution des performances

Modèle architecture du réseau de neurones
& choix des différents paramétrages

ENTRAÎNER le modèle d'un réseau de neurones

Répéter un "nombre suffisant'' de fois (# d'époques)
la boucle d'apprentissage,
- *en propageant* des batchs (petites parties)
d'un ensemble d'apprentissage (Training Set)
- *en minimisant* la fonction perte,

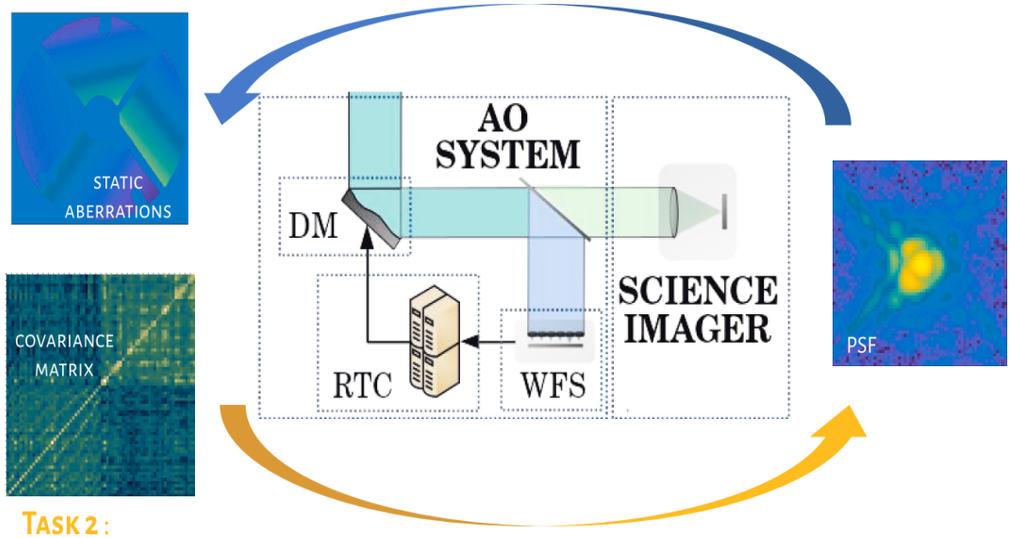
⇒ **détermination les valeurs des paramètres du réseau**

APPLY : PROBLÉMATIQUE DE RÉGRESSION

But

- *Caractériser conjointement* la **turbulence atmosphérique** et les **aberrations optiques du télescope** *à partir de* la **PSF (Point Spread Function)** résultante,

- *Réduire la dimensionnalité* de la description des phénomènes physiques en utilisant un **modèle avec seulement 15 paramètres.**



TASK 2 :

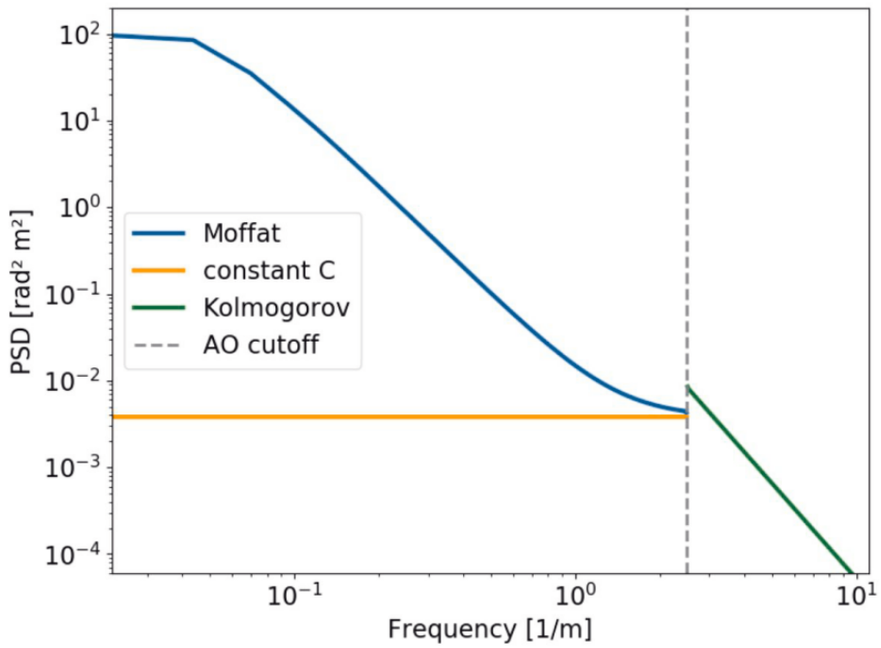
Model of the wavefront spatial PSD

$W(k<=k_{AO}) = M(k, A, \alpha, \rho, \beta) + C$

$W(k>k_{AO}) = 0.023 r_0^{-5/3} k^{-11/3}$

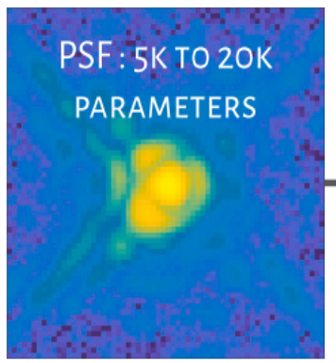
$PSF_{atm} = F^{-1}[\exp(F[W])]$

Atmospheric component (maoppy)



Méthode

- Apprentissage d'un **réseau de neurones CNN** sur des simulations de PSFs correspondant à 2 instruments : NIRC2 / Keck (10m) ; IRDIS / VLT (8m)
- Comparer les performances avec les méthodes d'estimations classiques de psf-fitting



PSFAO21 model : **6 to 15 parameters**

$PSF(\mu_{stat}, r_0, C, A, \alpha, \rho, \beta) = PSF_{stat}(\mu_{stat}) * PSF_{atm}(r_0, C, A, \alpha, \rho, \beta)$

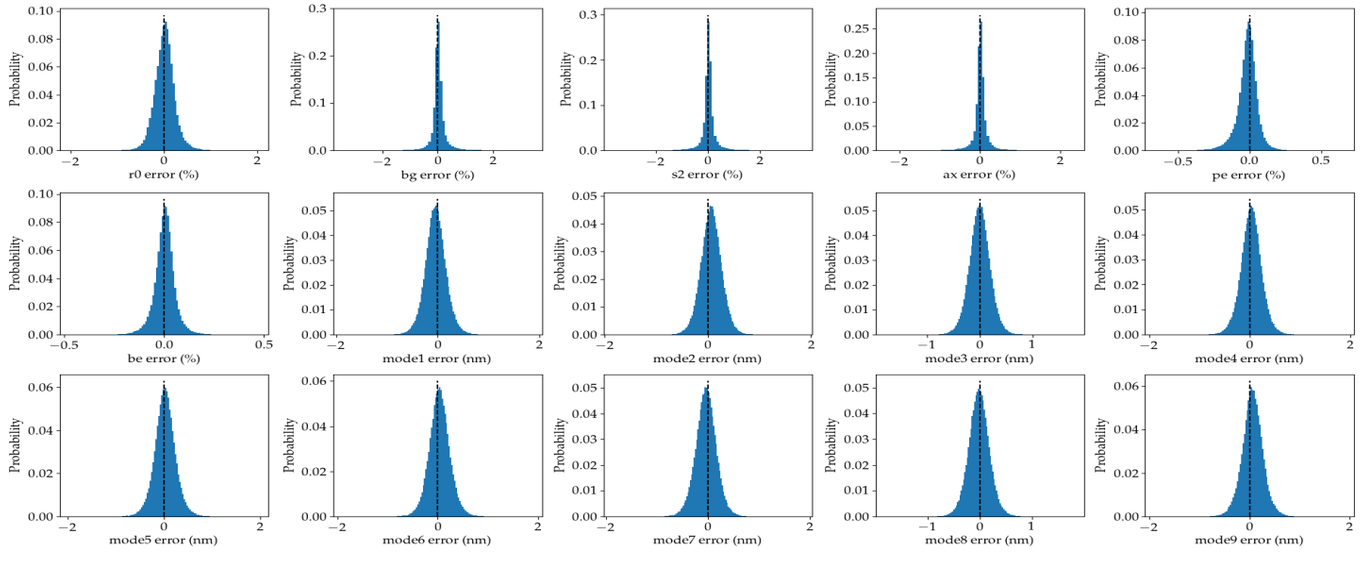
APPLY : PROBLÉMATIQUE DE RÉGRESSION

Résultats - PSF *non bruitées* :

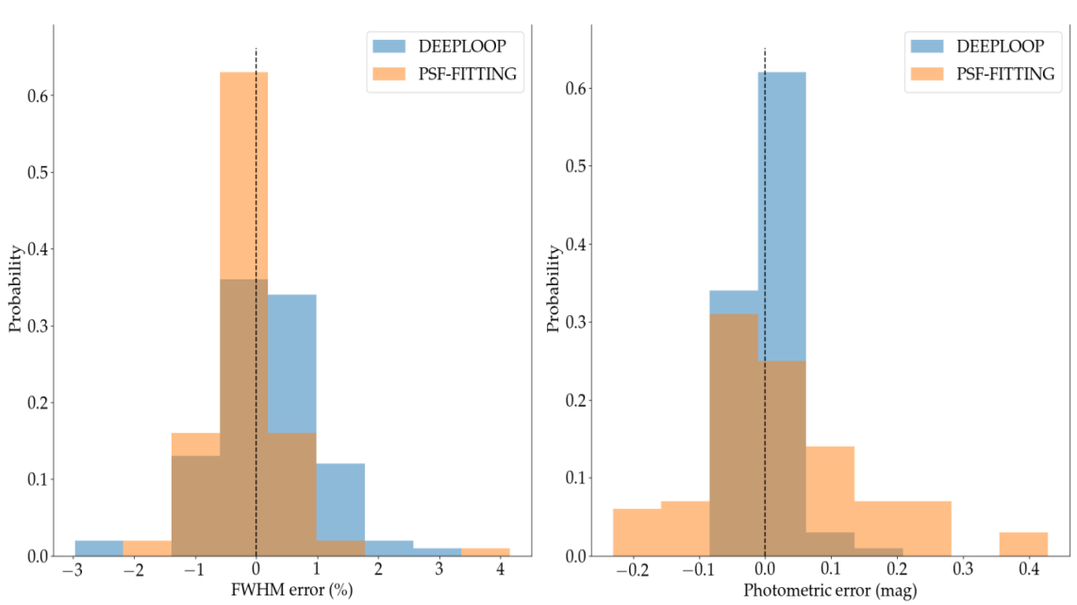
Résidu de turbulence **au % près**,

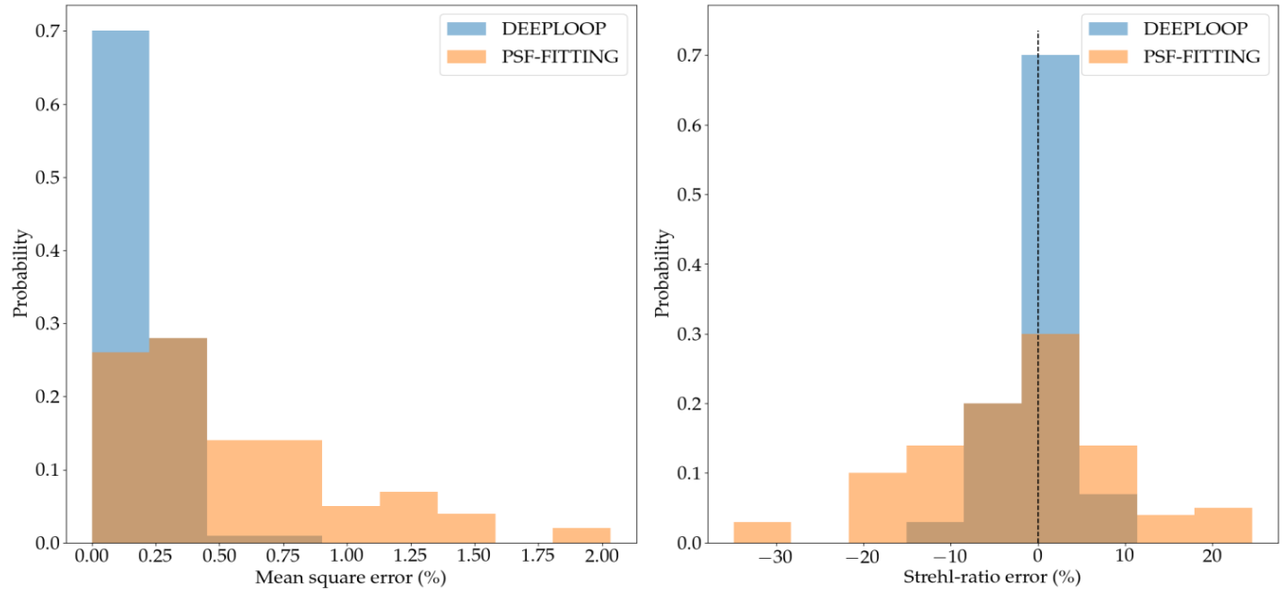
Alignement des segments **au nm près**

Performances du réseau de neurones **supérieures** aux méthodes classiques d'ajustement de modèles (pb de couplage entre les paramètres atmosphériques et ceux du télescope)



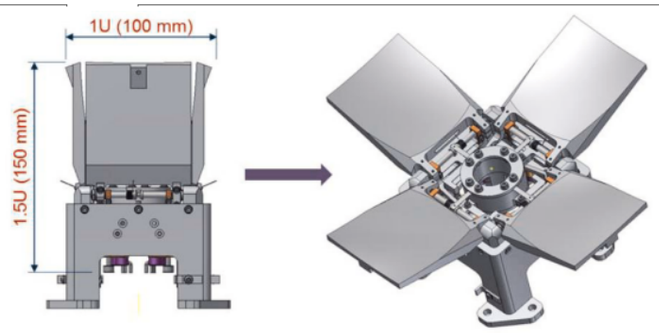
- PSF *bruitées* : **Amélioration d'un facteur 2** par rapport aux méthodes d'ajustement de modèles





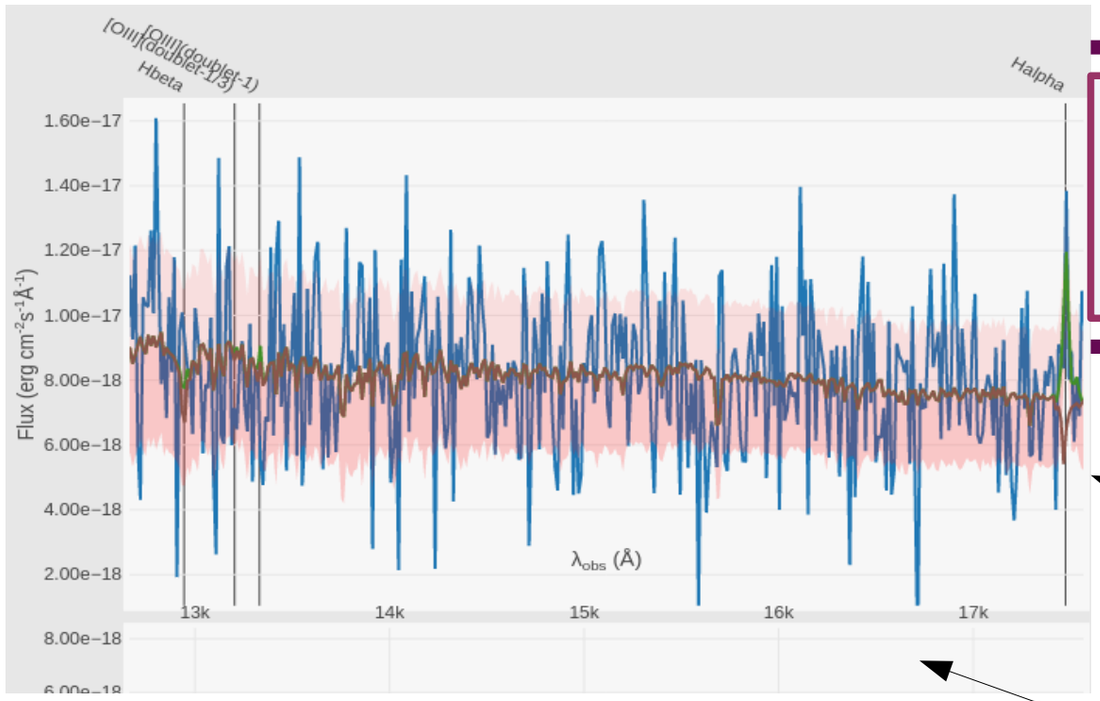
Autre Exemple Projet **AZIMOV**

Estimation de la PSF (12 modes de Zernike) pour le cophasage des 4 pétales du miroir d'un nanosatellite d'observation de la terre

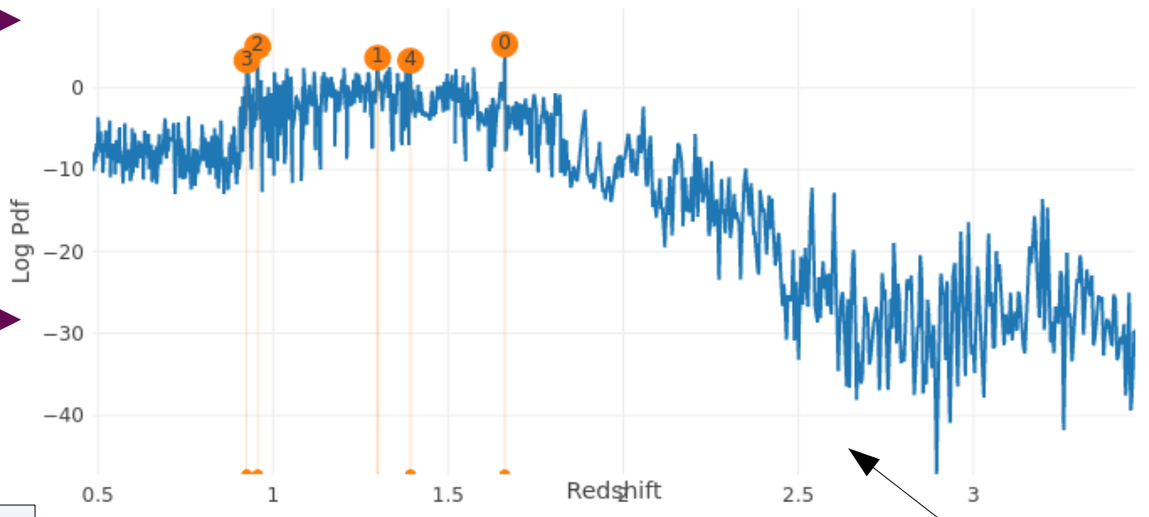


EUCLID : PROBLÉMATIQUE DE CLASSIFICATION

But - *Classer* **suivant un indice de fiabilité** les estimations de redshifts spectroscopiques *à partir* **des LogPDF** fournies par le pipeline 1D Amazed du satellite EUCLID (étude de l'énergie noire),
- **Automatiser la classification** pour un **traitement sur toute la ''chaine de calcul''** d'1 spectre/minute



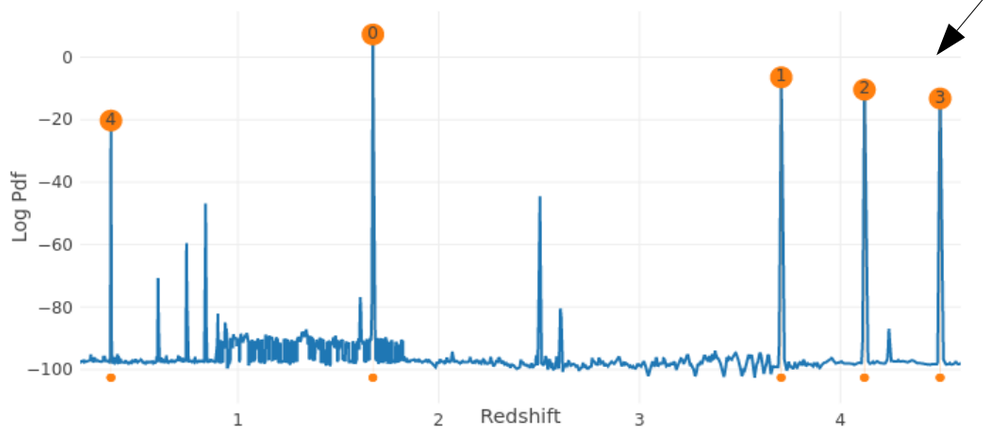
Pipeline Amazed

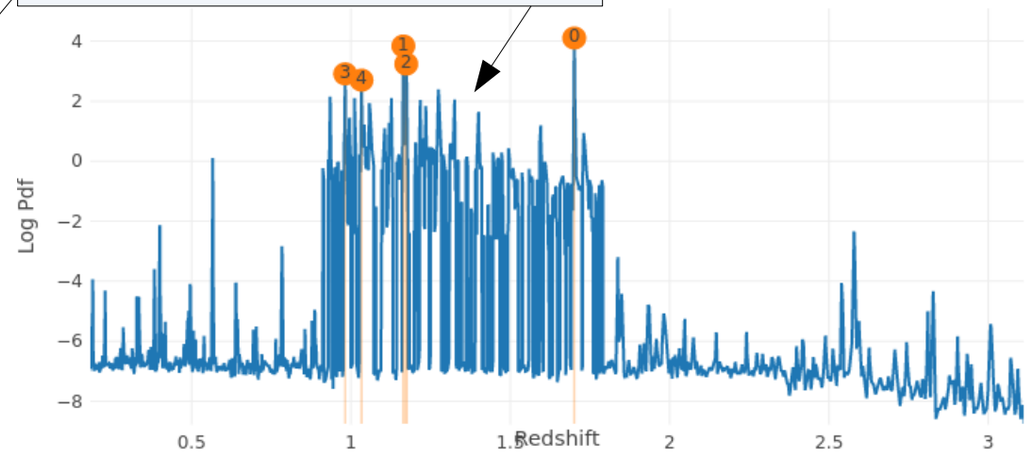


Spectre 1D fourni par EUCLID

LogPDF

Autres ex. de LogPDF





- *Proposer différents types de classifieurs* suivant **différents critères de classification.**

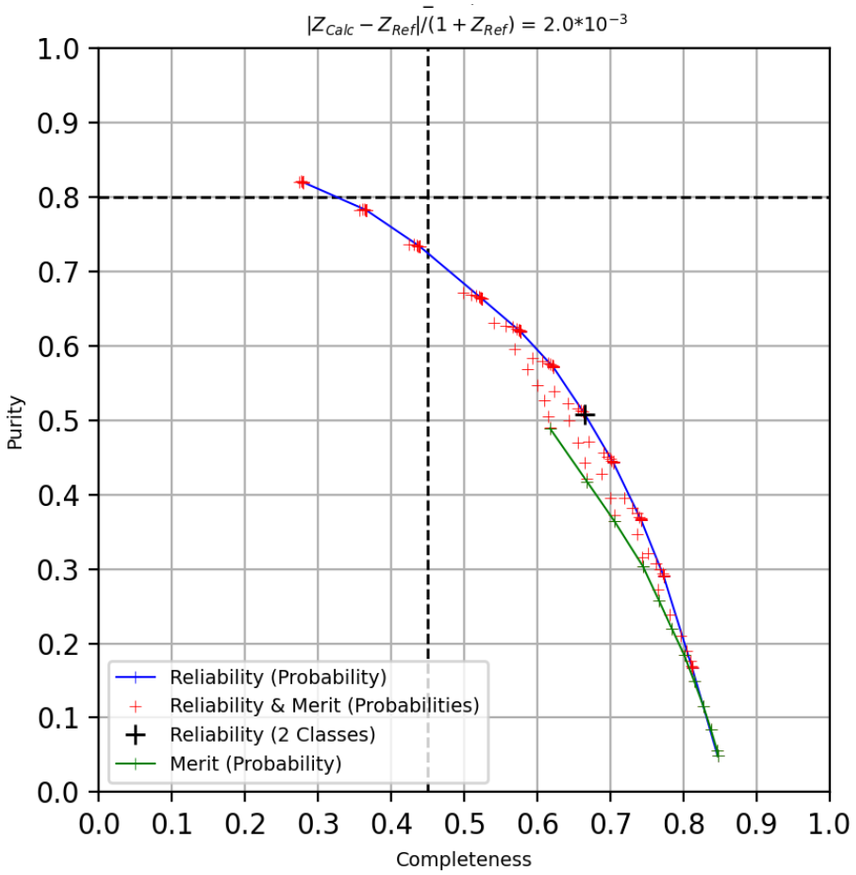
https://euclid-france.fr/contexte-et-objectifs/

EUCLID : PROBLÉMATIQUE DE CLASSIFICATION

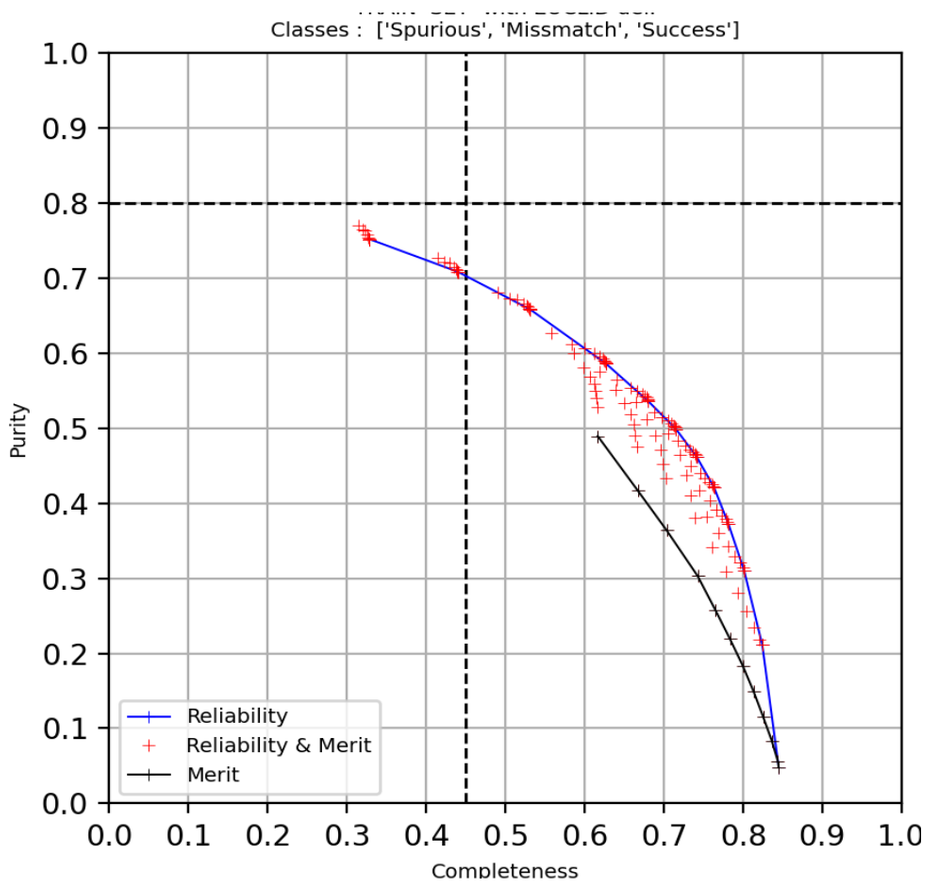
Méthode - Apprentissage d'un **réseau de neurones CNN-1D** sur des simulations et données réelles
- Comparaison les performances avec la ‘’redshift probability’’ fournie par le pipeline Amazed

Résultats - Utilisation des **courbes de Purity (Precision) / Completeness (Recall)** définies avec la classe "success"
(plusieurs définitions possibles : classiques, adaptées aux spécificités EUCLID)
- l'objectif est de déterminer des seuils pour atteindre :
Purity > 0.8 & Completeness > 0.45

Critère avec **2** classes



Critère avec **3** classes



Autre Exemple Projet **EUCLID** : classification Etoiles / Quasars / Galaxies à partir des spectres
Travail de Simon Conseil (démarré récemment, phase exploratoire)

CeSAM

Pôle Machine Learning / Deep Learning

(responsable : morgan.gray@lam.fr)

https://projets.lam.fr/projects/mldl/wiki

Pôle Infrastructure HPC

(responsable : jean-charles.lambert@lam.fr)

https://projets.lam.fr/projects/cluster-de-calcul-du-lam/wiki#Presentation