Corentin Cadiou – CR Section 55

Laboratoire d’accueil : Institut d'Astrophysique de Paris

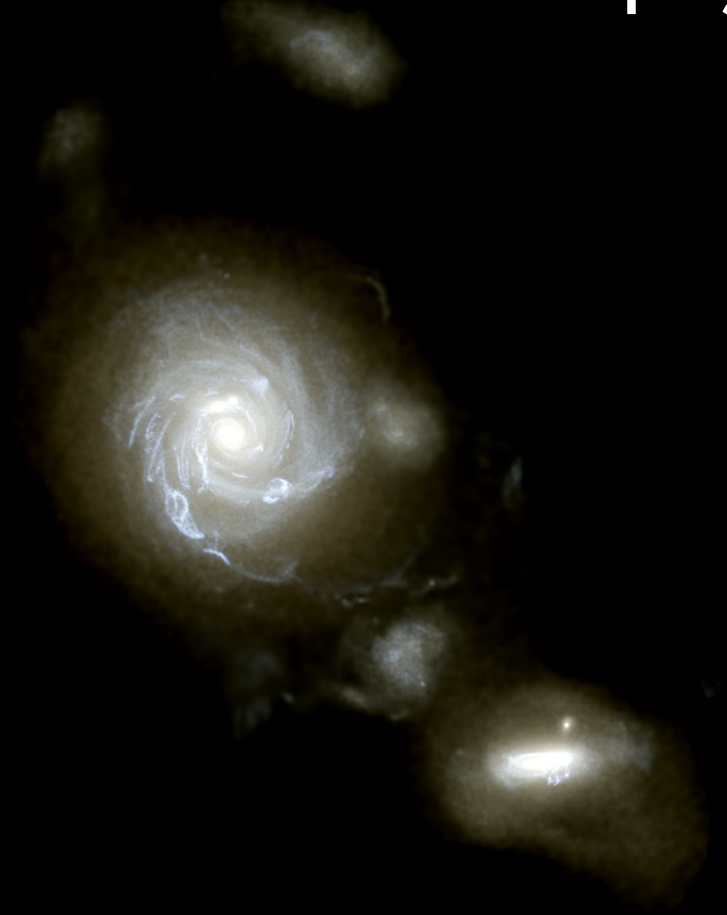


Image simulée © Cadiou

Corentin Cadiou – CR Section 55

Laboratoire d'accueil : Institut d'Astrophysique de Paris

— Compétences

Simulation astrophysique & cosmologie :

Calcul haute performance

Analyse grand volume de données

Visualisation : ex. rendu volumétrique, images simulées

— Financements & awards

100Mh de temps de calcul

(investigateur principal : 12Mh)

Développeur : codes simulations + post-traitement

Reviewer pour la conférence Scipy (21, 22)

NumFOCUS Contributor Award 2018

Membre du *Sustainability Committee* de l'EAS

Élève normalien
Majeure : Physique
Mineure : Informatique

2016

Thèse IAP
Bourse ILP

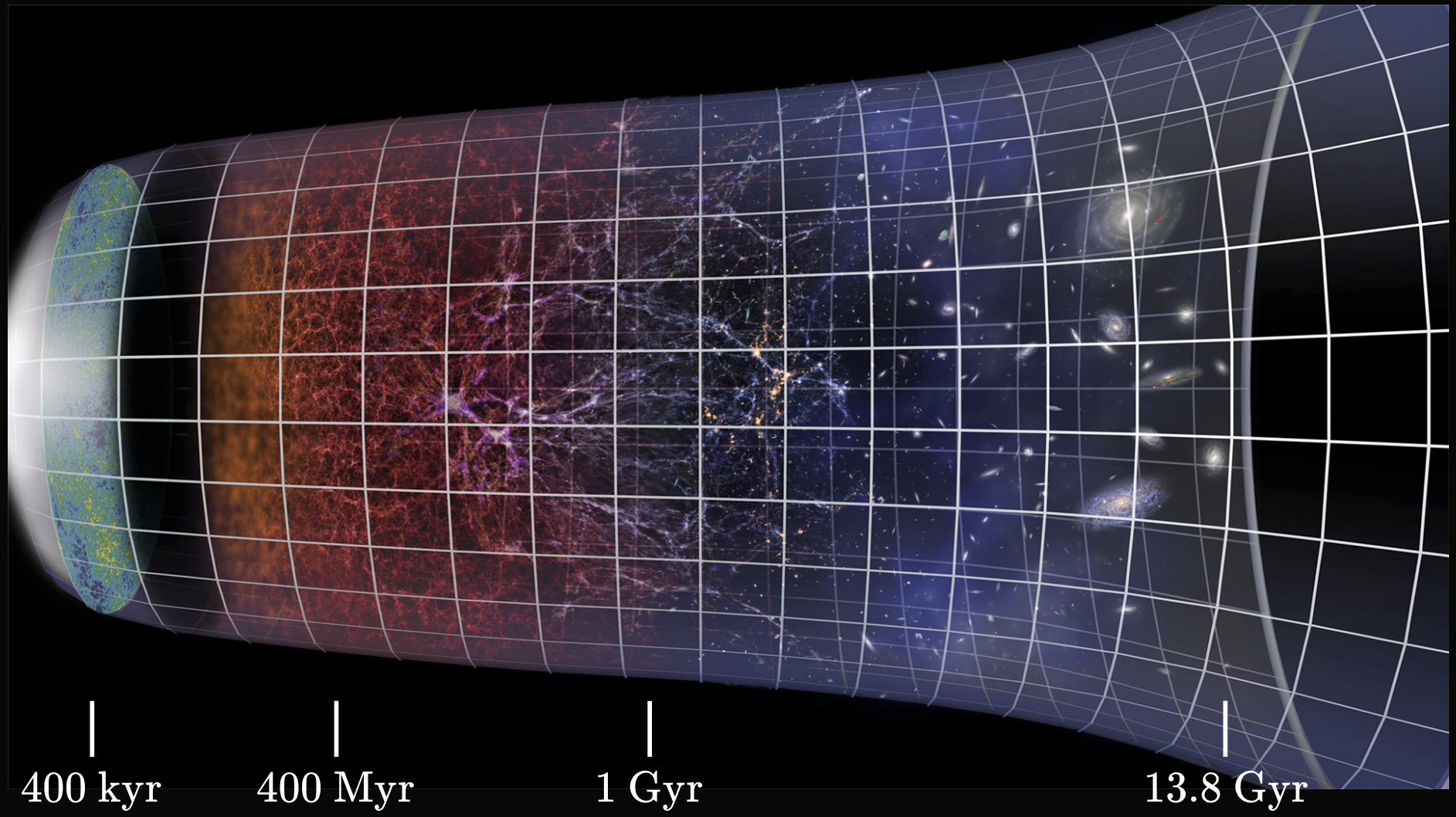
2019

Postdoc UCL
Angleterre

2022

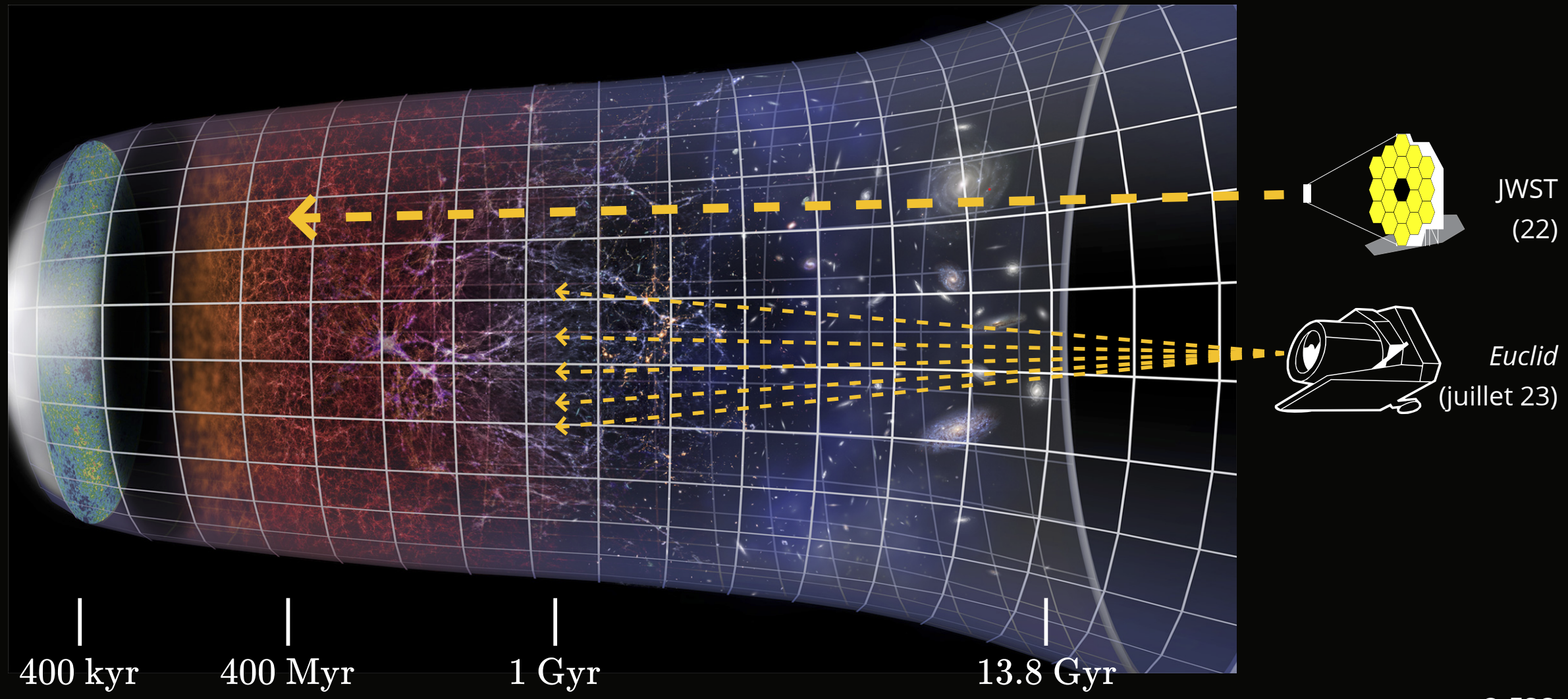
Postdoc Lund
Suède

Comprendre l'histoire des galaxies



© ESO

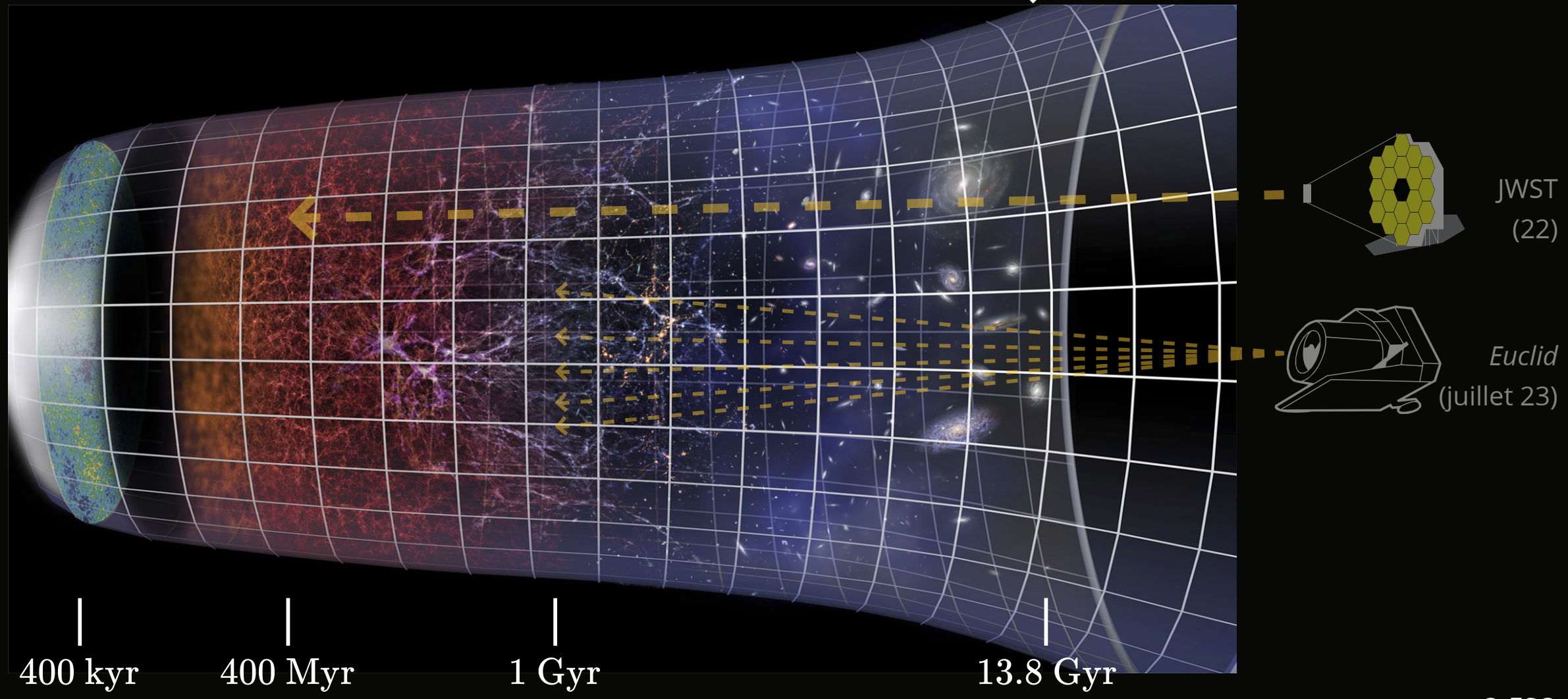
Comprendre l'histoire des galaxies



© ESO

Comprendre l'histoire des galaxies

Mes simulations (cosmologiques)

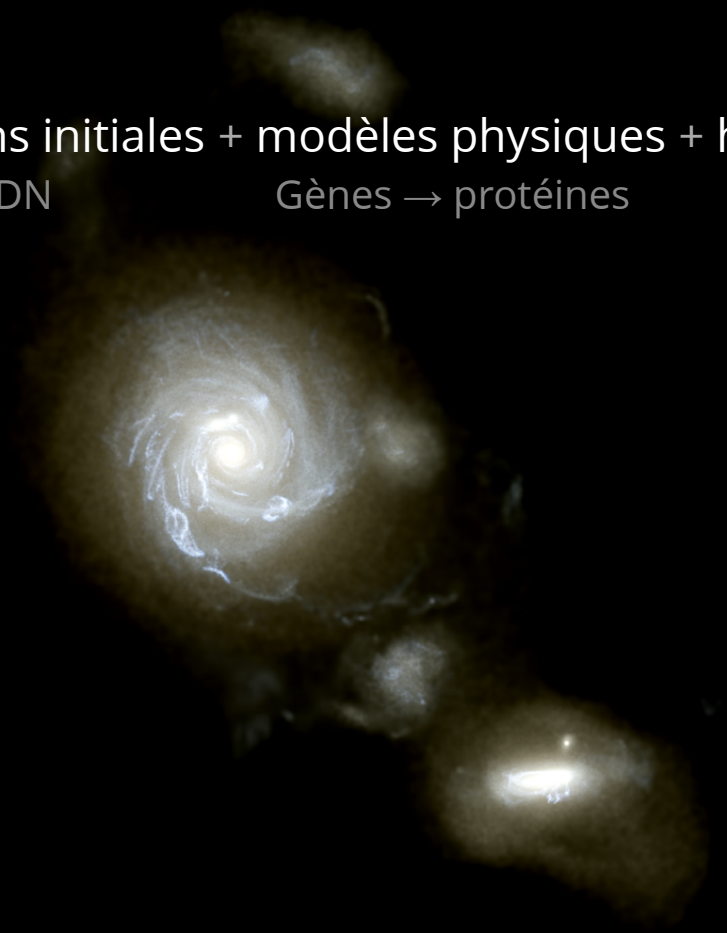


© ESO

Origine des propriétés des galaxies

Propriétés des galaxies = conditions initiales + modèles physiques + histoire de formation

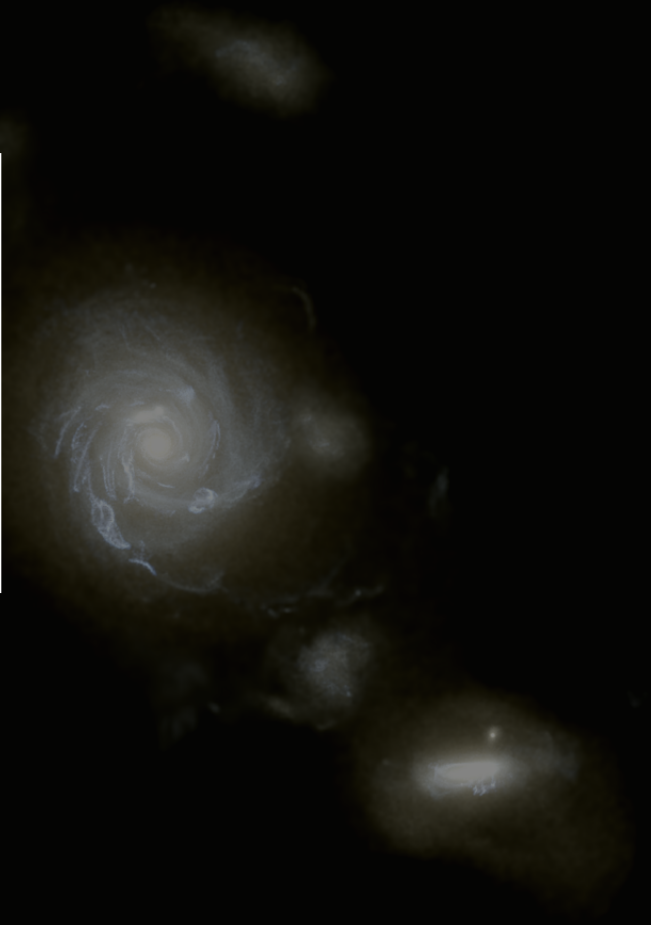
ADN Gènes → protéines Epigénétique



Propriétés des galaxies = conditions initiales + modèles physiques + histoire de formation

ADN | Gènes → protéines | Epigénétique

Simulations cosmologiques standard



Très grand échantillon de galaxies

- conditions initiales **aléatoires**
- histoire de formation **aléatoire**

Focus sur **statistique**

Propriétés des galaxies = conditions initiales + modèles physiques + histoire de formation

ADN — Gènes → protéines — Epigénétique

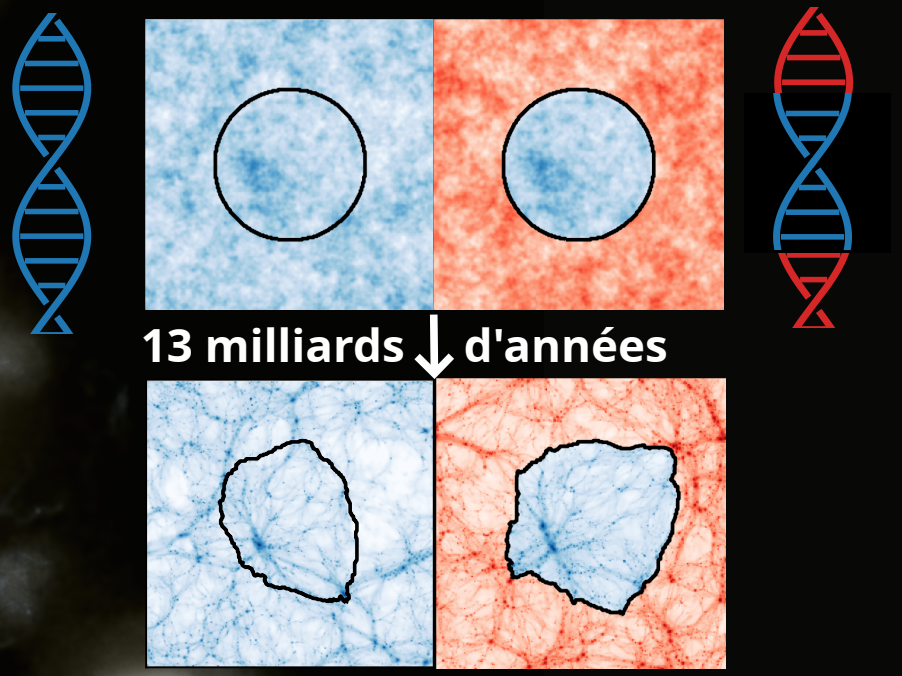
Simulations cosmologiques standard

Très grand échantillon de galaxies

- conditions initiales **aléatoires**
- histoire de formation **aléatoire**

Focus sur **statistique**

Mon approche : modifications génétiques



Petit échantillon de galaxies sous contrôle

Focus sur l'**interprétabilité**

Cadiou+21a,b, 22

Propriétés des galaxies = conditions initiales + modèles physiques + histoire de formation

ADN — Gènes → protéines — Epigénétique

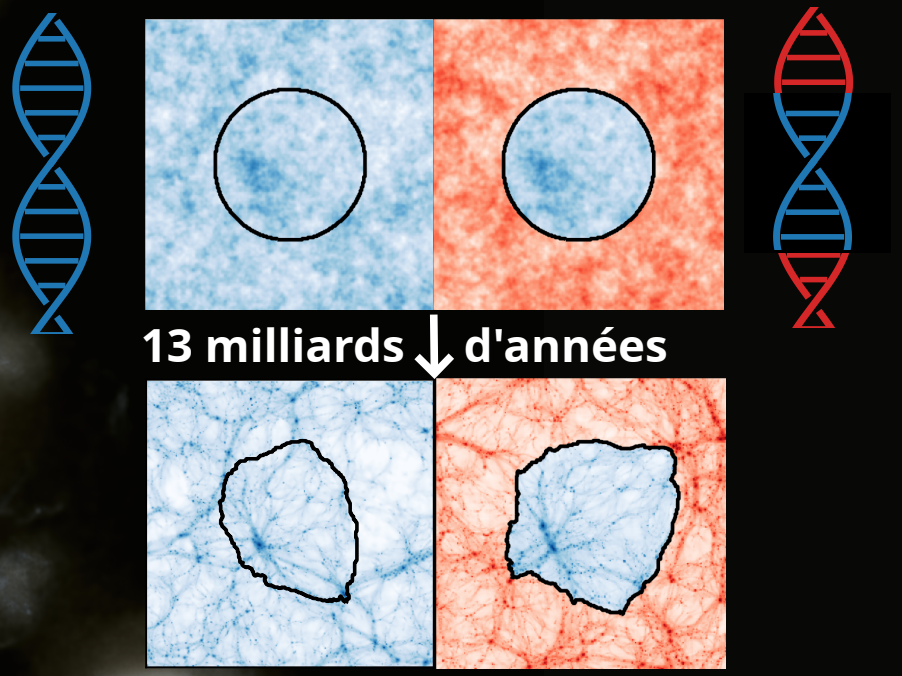
Simulations cosmologiques standard

Très grand échantillon de galaxies

- conditions initiales **aléatoires**
- histoire de formation **aléatoire**

Focus sur **statistique**

Mon approche : modifications génétiques



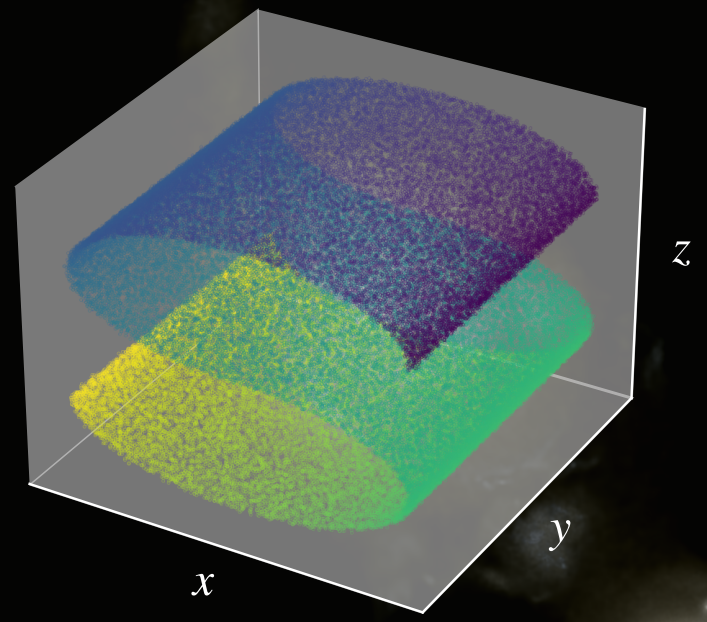
Petit échantillon de galaxies sous contrôle

Focus sur l'**interprétabilité**

Cadiou+21a,b, 22

⇒ Certaines propriétés des galaxies ne sont **pas** stochastiques (moment angulaire, concentration)

Mesure de la dimensionnalité des données

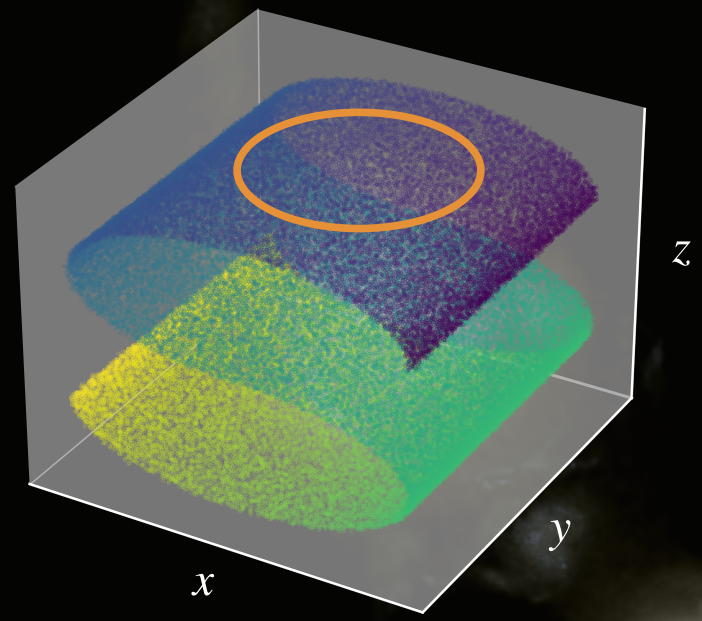


Dimensions intrinsèques : 2

Plongées dans 3 dimensions

Cadiou+in prep

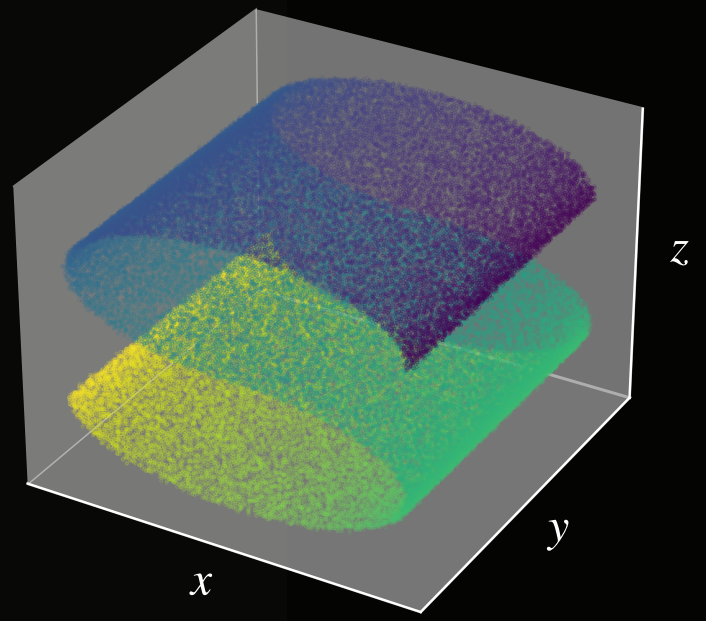
Mesure de la dimensionnalité des données



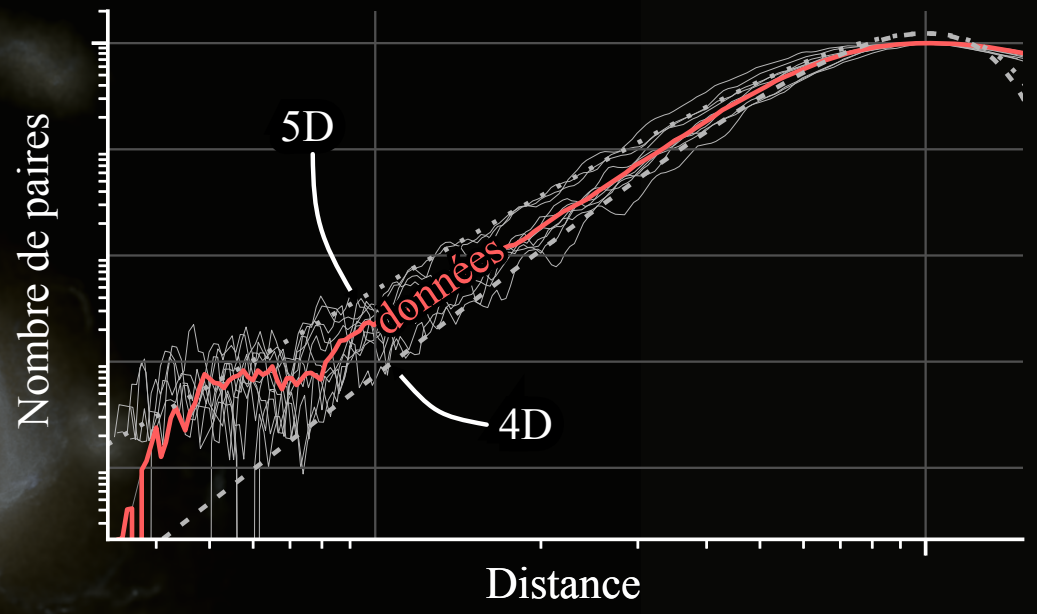
Dimensions intrinsèques : 2
Plongées dans 3 dimensions

Cadiou+in prep

Dimensionnalité des données



Dimensions intrinsèques : 2
Plongées dans 3 dimensions

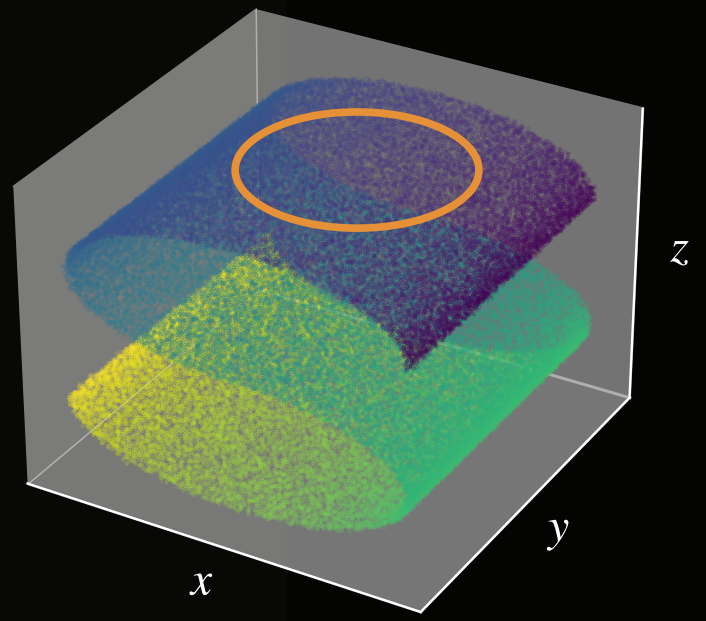


Dimensions intrinsèques : 4~5
Plongées dans 14 dimensions

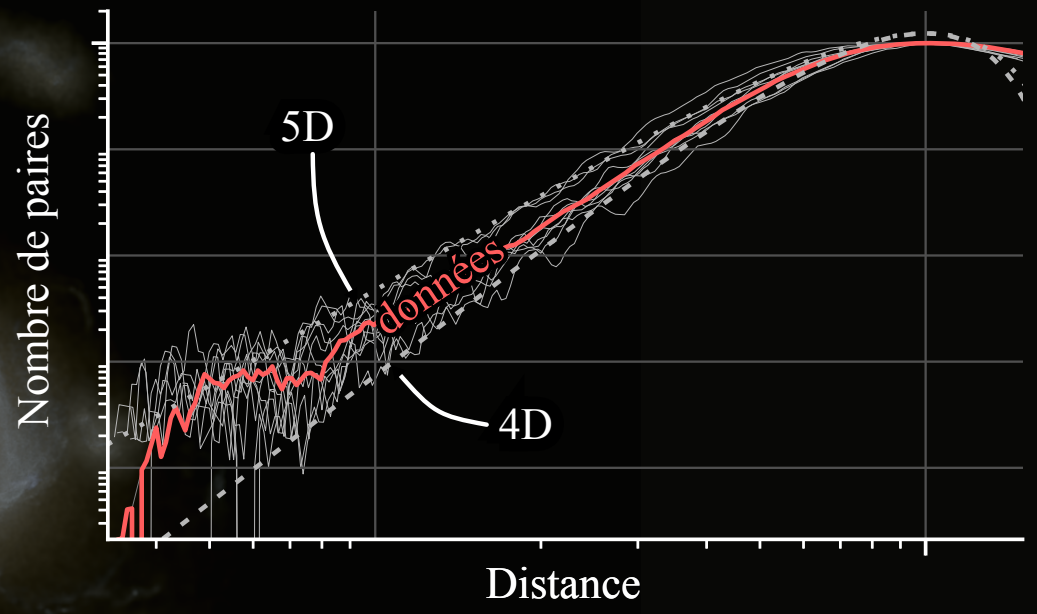
⇒ Les observations des galaxies ont une **faible dimension intrinsèque**
(données photométriques 14 bandes larges dans le catalogue COSMOS)

Cadiou+in prep

Dimensionnalité des données



Dimensions intrinsèques : 2
Plongées dans 3 dimensions



Dimensions intrinsèques : 4~5
Plongées dans 14 dimensions

⇒ Les observations des galaxies ont une **faible dimension intrinsèque**
(données photométriques 14 bandes larges dans le catalogue COSMOS)

Cadiou+in prep

Analyse en série temporelle

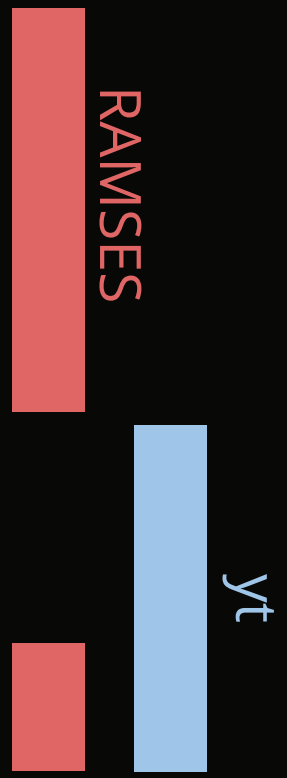
Combinaison entre particules Lagrangiennes et grille Eulérienne

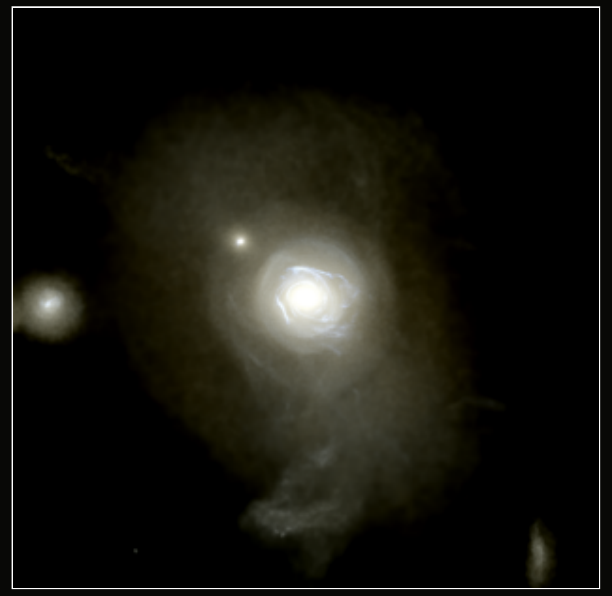
2018
familles de particules

2019
particles traceuses

2020
support yt (NumFOCUS award)

Aujourd'hui
3.5 articles / an utilisent RAMSES + yt





Cadiou+18; Kocjan, Cadiou+in prep. (étudiante)

⇒ Code réutilisable autres domaines aux volume finis : météorologie (CFradial), nucléaire (Pyne)

Analyse en série temporelle

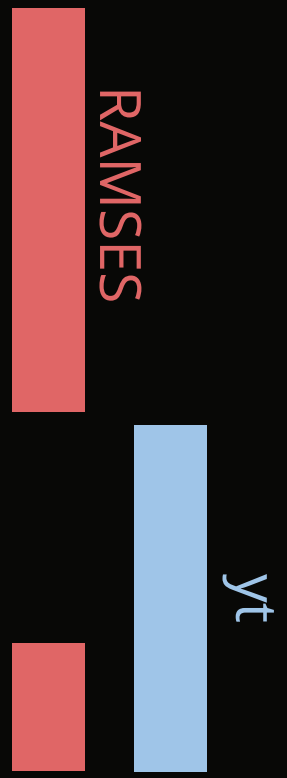
Combinaison entre particules Lagrangiennes et grille Eulérienne

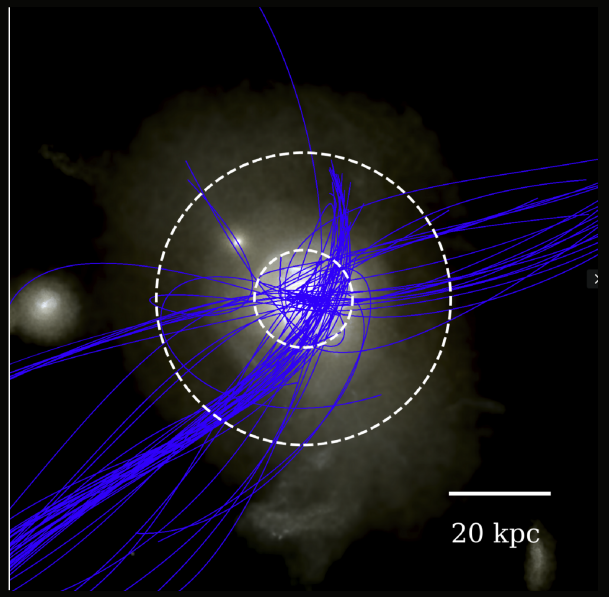
2018
familles de particules

2019
particles traceuses

2020
support yt (NumFOCUS award)

Aujourd'hui
3.5 articles / an utilisent RAMSES + yt





Cadiou+18; Kocjan, Cadiou+in prep. (étudiante)

⇒ Code réutilisable autres domaines aux volume finis :
météorologie (CFradial), nucléaire (Pyne)

Analyse en série temporelle

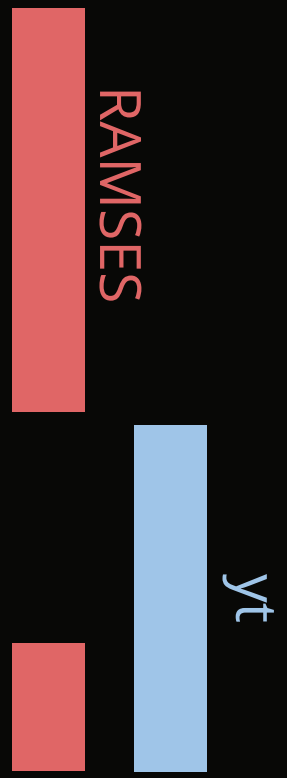
Combinaison entre particules Lagrangiennes et grille Eulérienne

2018
familles de particules

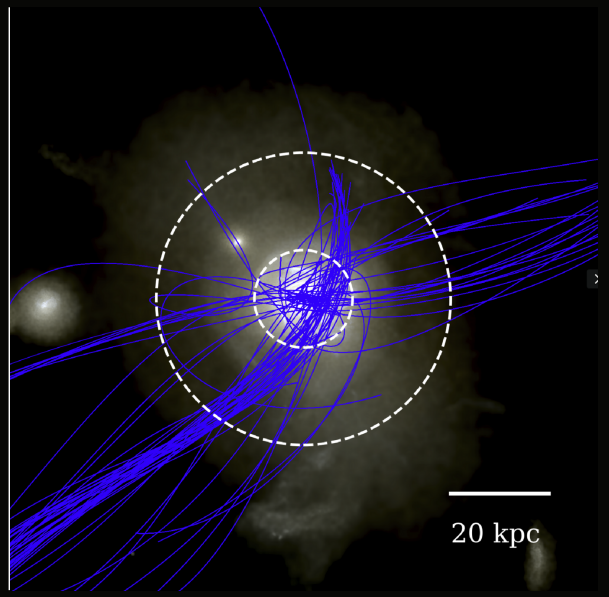
2019
particles traceuses

2020
support yt (NumFOCUS award)

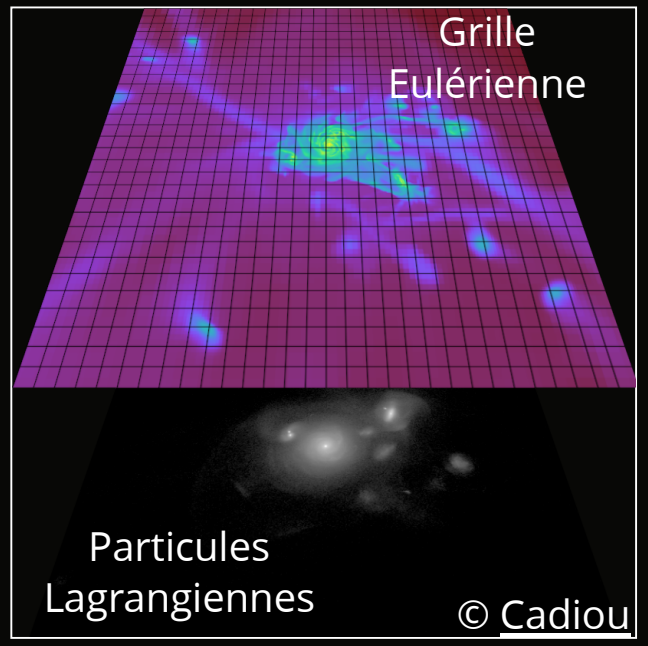
Aujourd'hui
3.5 articles / an utilisent RAMSES + yt



⇒ Code réutilisable autres domaines aux volume finis : météorologie (CFradial), nucléaire (Pyne)



Cadiou+18; Kocjan, Cadiou+in prep. (étudiante)



Analyse en série temporelle

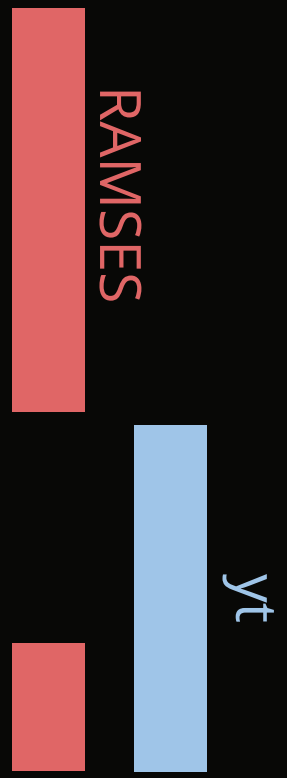
Combinaison entre particules Lagrangiennes et grille Eulérienne

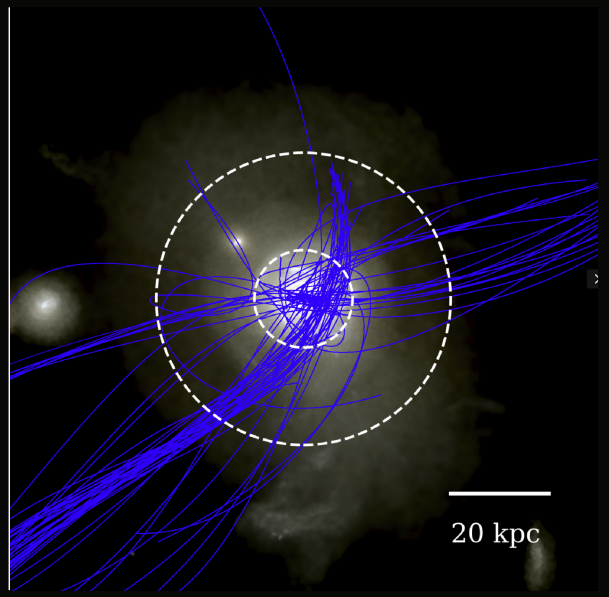
2018
familles de particules

2019
particles traceuses

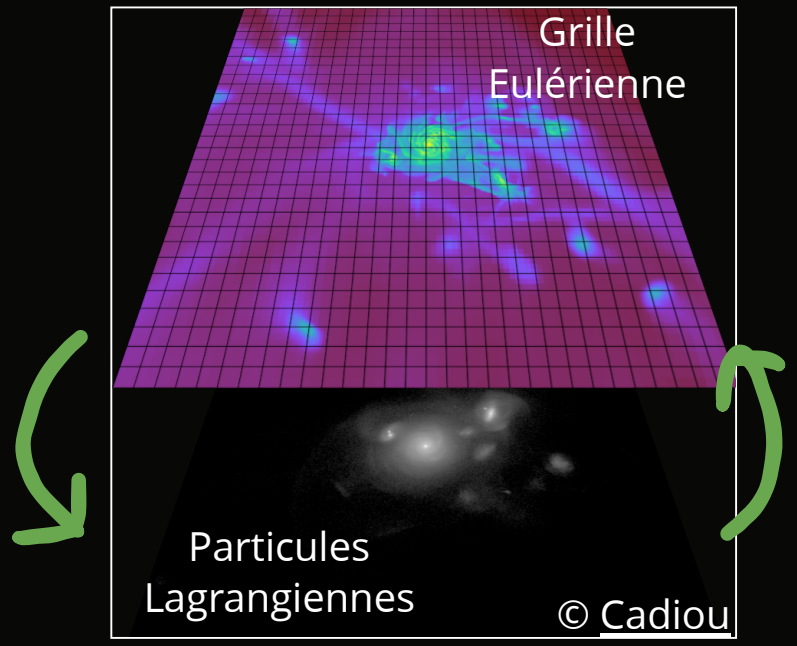
2020
support yt (NumFOCUS award)

Aujourd'hui
3.5 articles / an utilisent RAMSES + yt



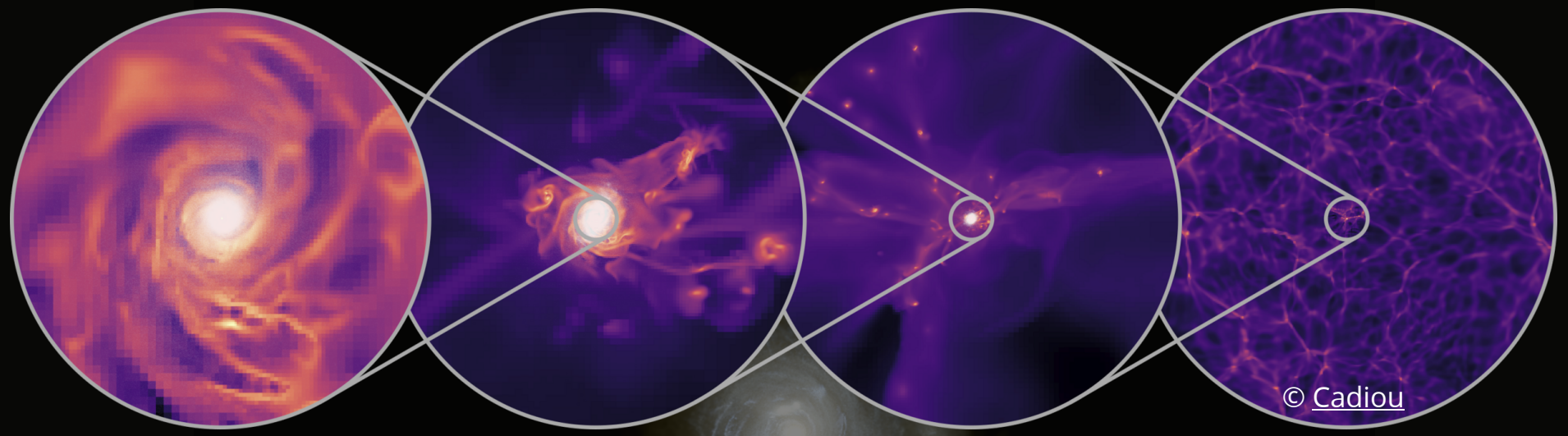


Cadiou+18; Kocjan, Cadiou+in prep. (étudiante)



⇒ Code réutilisable autres domaines aux volume finis : météorologie (CFradial), nucléaire (Pyne)

Passage à l'exascale : + résolution, + volume



Émergence
< 100 ly

+6 ordres de grandeurs

Cosmologie
> 100 Mly

Grand volume, faible résolution

Mes simulations :
petits volumes
forte résolution

Euclid flagship

Modifications génétiques (**Cadiou**+22 ; 12Mh calcul)

Vintergatan-GM (Rey, ..., **Cadiou** 23 ; 50 Mh calcul)

PRISM-XL (Katz, Rey et **Cadiou** 23 ; 50Mh calcul)

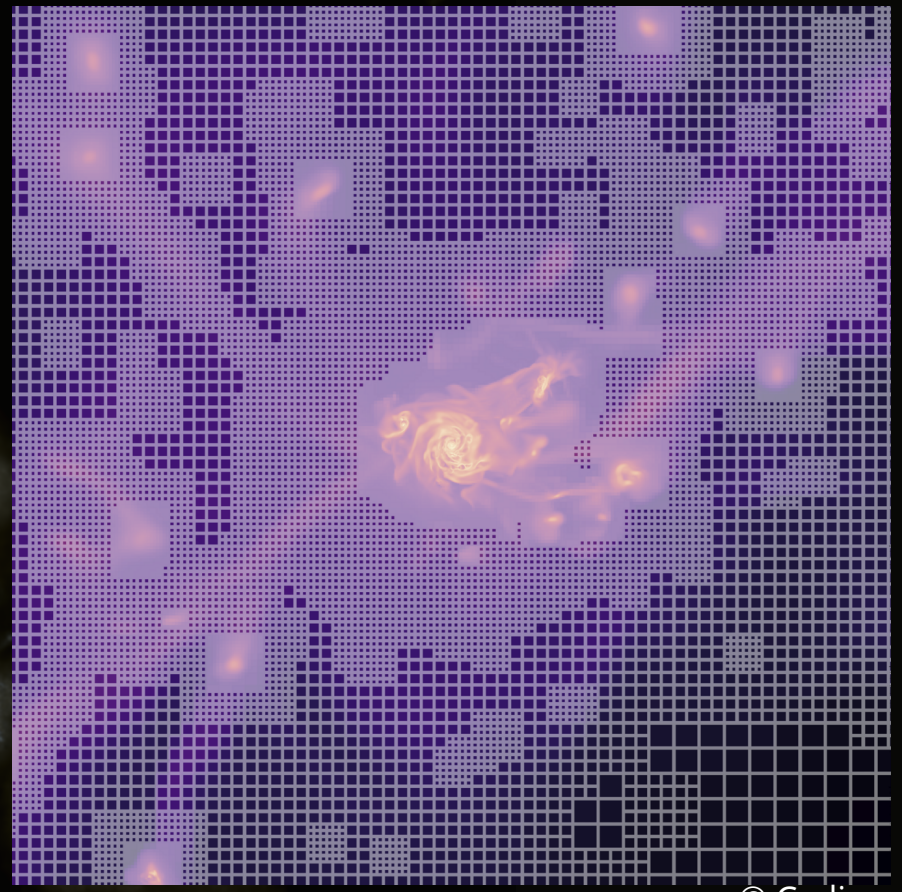
Dyablo : un code adapté à l'exascale

Horizon : premières simulations 1 à 2 ans

⇒ Intégration dès **aujourd'hui** dans développement du code **Dyablo**
Court terme : traduire modèle physiques RAMSES → Dyablo

Difficultés :

- architectures hétérogènes (x86/GPU/ARM*) ;
- répartition de la charge ;
- communications ;
- utilisation mémoire ;
- stockage et analyse des données.



© Cadiou

Géometrie complexe : difficile de répartir le calcul & limiter les communications

Dyablo : développé par collaboration Ginea, soutenue par l'Action Numérique Spécifique

Convergences possibles avec projets Wholesun, Idefix (physique des étoiles)
Intégration dans le cadre d'EuroHPC

*Architectures RISC développées dans le cadre d'EuroHPC

Compression des données pour le calcul sur GPU

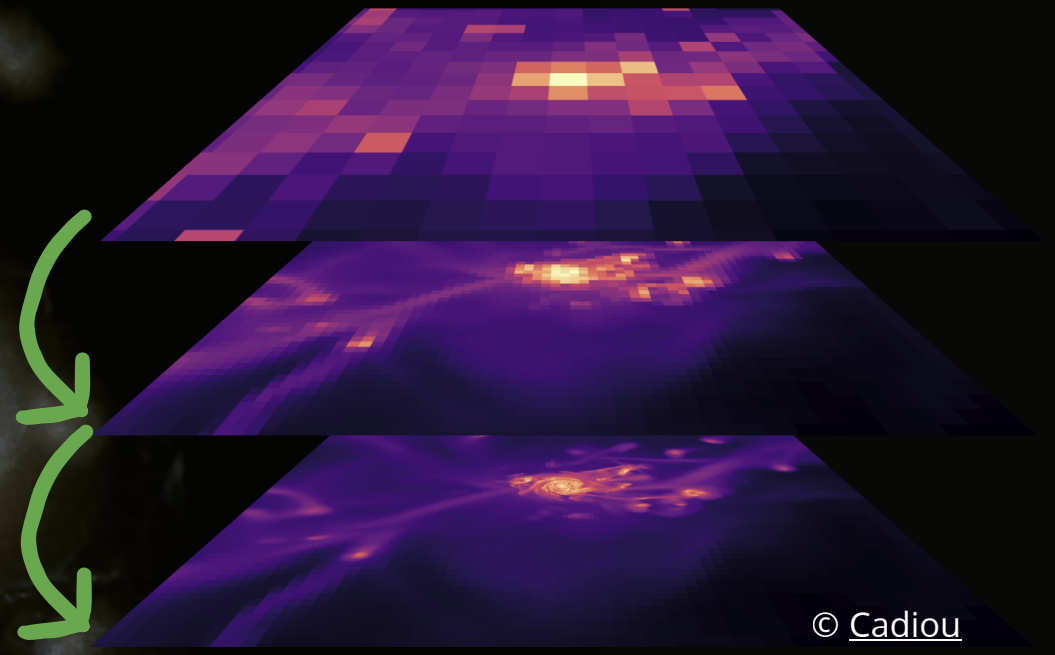
Horizon : 5 ans

Comment augmenter le nombre d'éléments simulés avec moins de mémoire ?

Difficultés :

- architectures hétérogènes (x86/GPU/ARM*) ;
- répartition de la charge ;
- communications ;
- **utilisation mémoire ;**
- stockage et analyse des données.

→ Compression des données en mémoire



© Cadiou

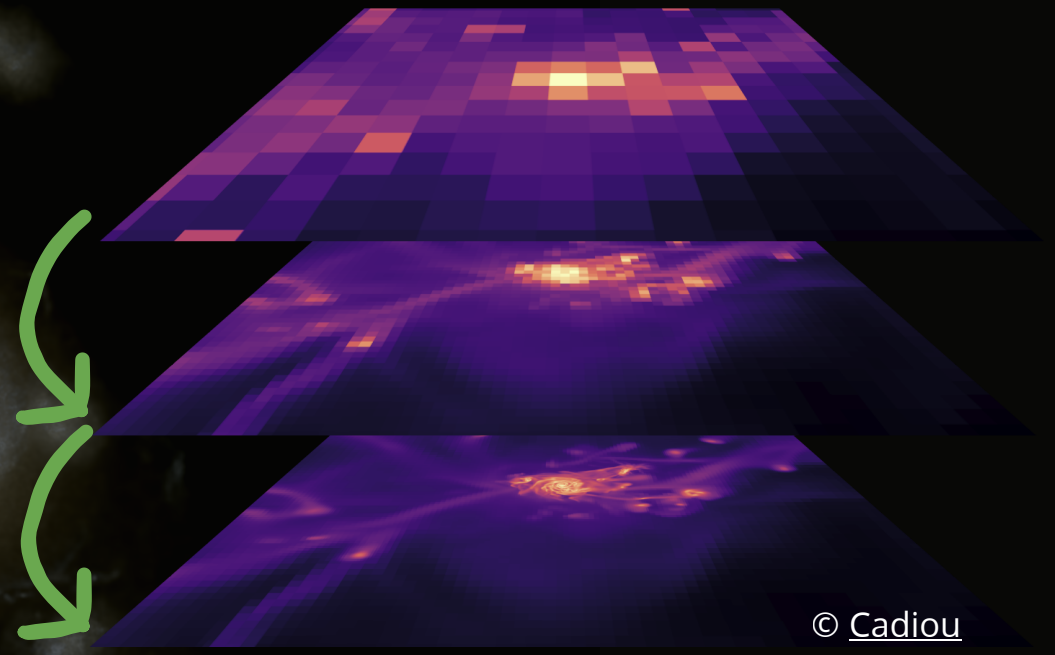
Compression des données pour le calcul sur GPU

Horizon : 5 ans

Comment augmenter le nombre d'éléments simulés avec moins de mémoire ?

Difficultés :

- architectures hétérogènes (x86/GPU/ARM*) ;
- répartition de la charge ;
- communications ;
- **utilisation mémoire ;**
- stockage et analyse des données.



© Cadiou

→ Compression des données en mémoire

1245			
124*5*	**124***1*	**124***6*	**124***9*

Données brutes : 160 bits

Économie de 50% de mémoire ? →

1245			
+0	-4	+1	+4

Compression “parent-enfant” : 45 bits

Nécessité de post-traiter & indexer les données de l'exascale

Horizon : 5-10 ans

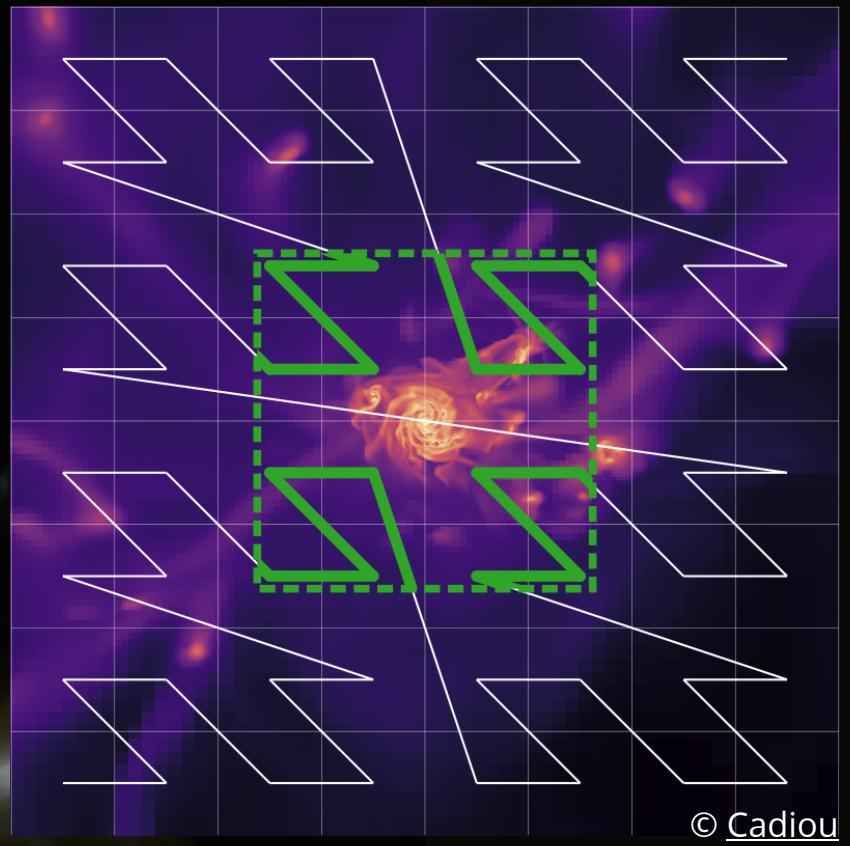
Comment réduire le volume des données et y accéder efficacement ?

Difficultés :

- architectures hétérogènes (x86/GPU/ARM*) ;
- répartition de la charge ;
- communications ;
- utilisation mémoire ;
- **stockage et analyse des données.**

Directions à explorer :

1. analyse à la volée ;
2. indexation des données.



Indexage par une courbe “Z” (de Morton) : accès rapide aux données

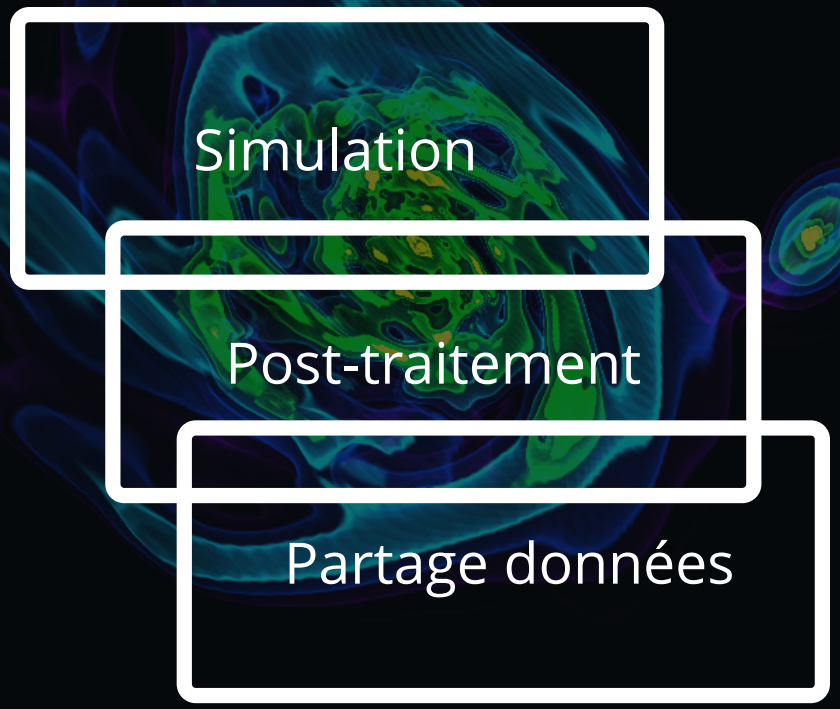
Applications possibles aux codes à Raffinement de Maillage Adaptatif (AMR) : Mécanique des Fluides Numériques (CFD), combustion, climatologie, ...

Dissémination des simulations

Horizon : 10 ans

Comment rendre accessible des centaines de TiB de donnée ?

Nécessité de séparation des tâches (*separation of concerns*)



Rendu volumétrique avec yt, PR#2610 © Cadiou

Dissémination des simulations

Horizon : 10 ans

Comment rendre accessible des centaines de TiB de donnée ?

Simulation

↓

Post-traitement

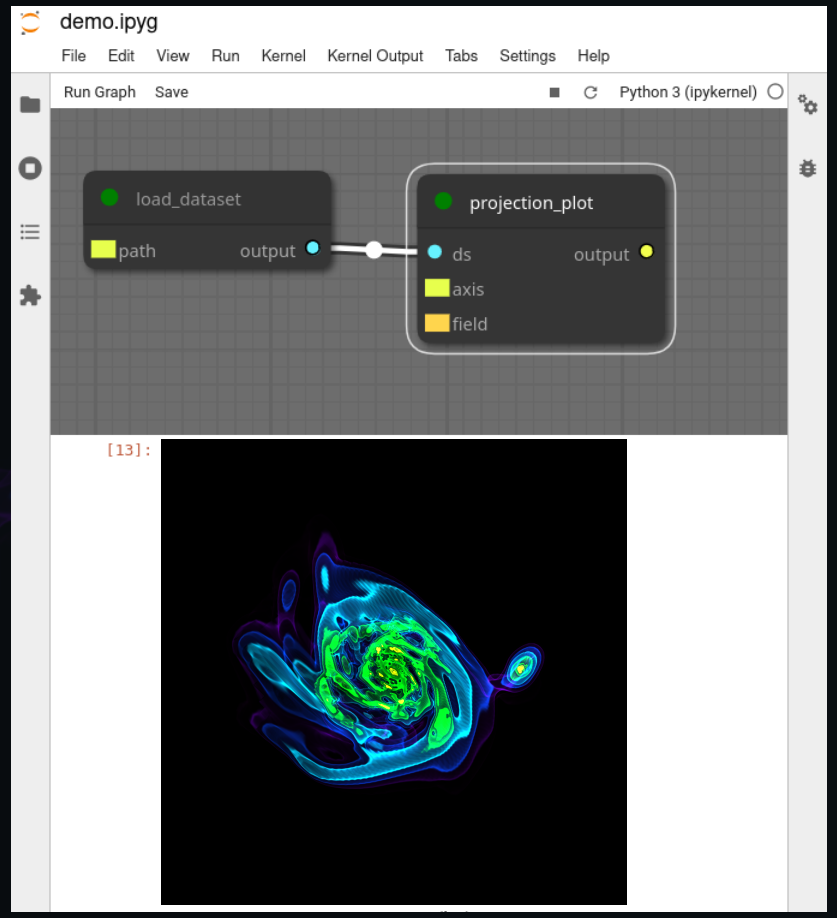
↓

Partage données

Normalisation des données
Librairies génériques (yt)

Rendre outils modulaires
Nécessité de représentation intermédiaires

Partage *via* base de donnée
(tangos; Pontzen [...] Cadiou+22)



ipyspaghetti © Cadiou
Grammaire de l'analyse : Walkow, Turk, Cadiou
Présentés à la conférence Scipy 2021

Principe FAIR : ***F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets*
Collaborations avec plateformes Galaxy (génomique, ...), Galactica (CEA), CDS Strasbourg

Conclusion

Quelle est l'origine des propriétés des galaxies ?

- Nouveau regard sur les simulations
  (modifications génétiques, analyse en série temporelle) ;
- Approche transdisciplinaire
  (analyse de dimensionnalité) ;
- Franchissement de l'exascale
  (challenges associés, traitement et accès aux données).

Science des données à différentes échelles :

- (astro)physique
  (RAMSES, Dyablo, yt) ;
- *data science*
  (reviewer conférence Scipy, librairies open-source) ;
- réflexion globale
  (grammaire de l'analyse, changement climatique).

Laboratoire d'accueil : IAP
Science ouverte : github.com/cphyc
Médiation : Astronomy on Tap London, Pint of Science

100Mh de temps de calcul (investigateur principal : 12Mh)
Développeur : RAMSES, Dyablo, yt
Reviewer pour la conférence Scipy (2021, 2022)
NumFOCUS New Contributor Award (2018)
Membre du *Sustainability Committee* de l'EAS (2023-)