



**Métaprogramme DIGIT-BIO**

*Biologie numérique pour explorer et prédire le vivant*

# ***L'IA en Sciences du Vivant***

Séance introductive

## **14h - Présentation du cycle d'animations**

*Julien Chiquet, UMR MIA Paris Saclay, INRAE*

*Marie-Laure Martin, Institut of Plants Sciences Paris-Saclay & UMR MIA Paris Saclay, INRAE*

*Christèle Robert-Granié, UMR GenPhySE, INRAE*

## **14h30 - Introduction à l'IA et au Machine learning**

*Liva Railavola, Head of AI Research at Criteo AI Lab*

 **Livestorm**

Chat

**Questions**





# Une avancée remarquable grâce à l'IA

## La prédiction de la structure des protéines (depuis 2018)

AlphaFold a des résultats presque aussi bon que ceux qu'il est possible d'atteindre avec des observations expérimentales

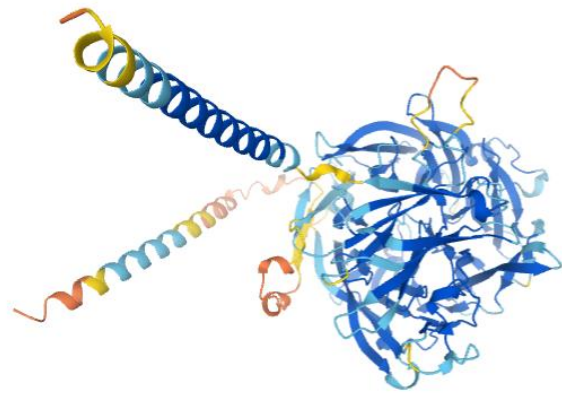
### 3D viewer

Model Confidence:

-  Very high (pLDDT > 90)
-  Confident (90 > pLDDT > 70)
-  Low (70 > pLDDT > 50)
-  Very low (pLDDT < 50)

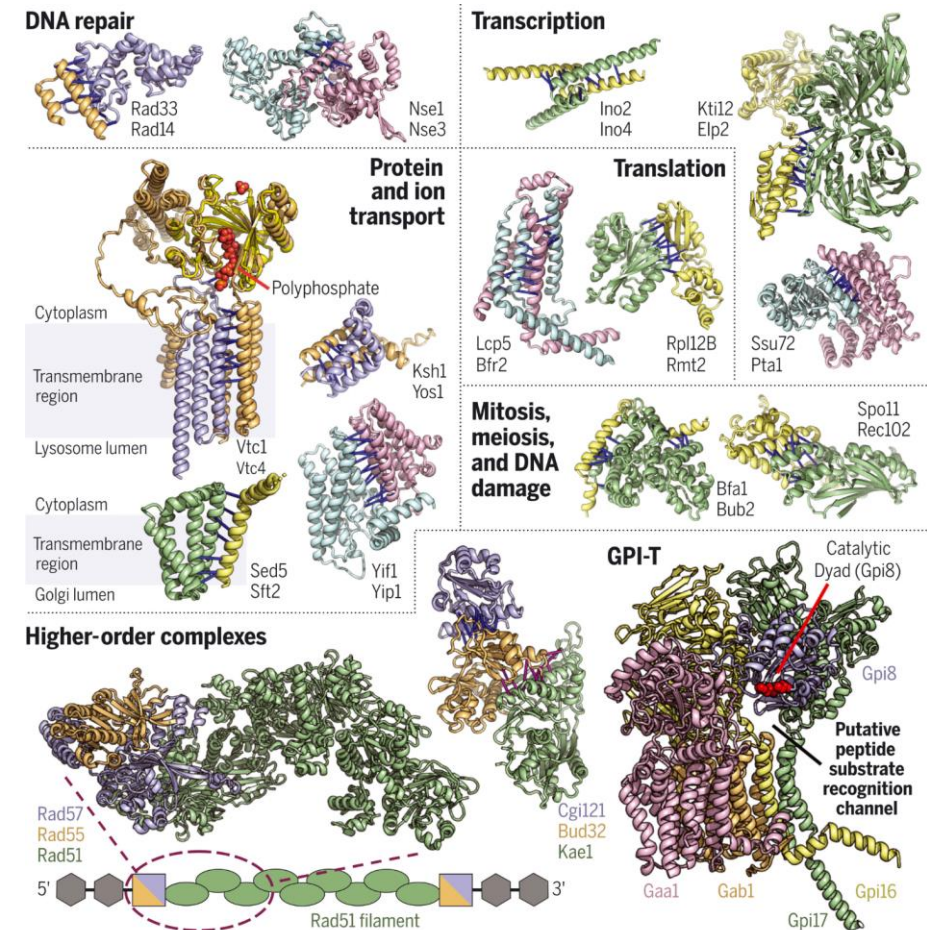
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

## T-cell immunomodulatory protein homolog



## La prédiction d'interaction protéine-protéine chez la levure

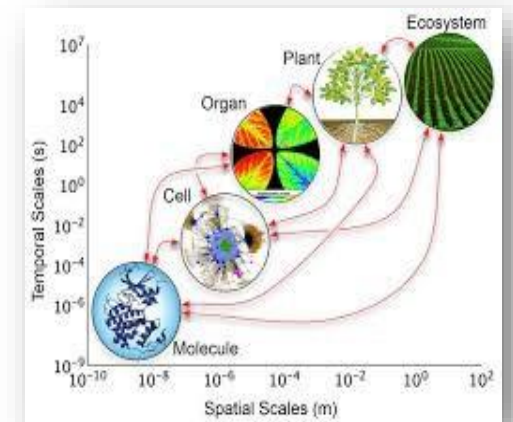
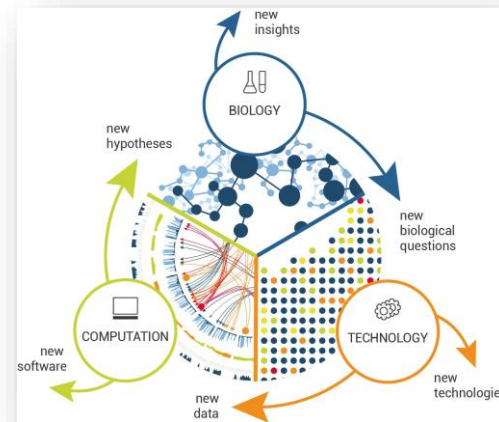
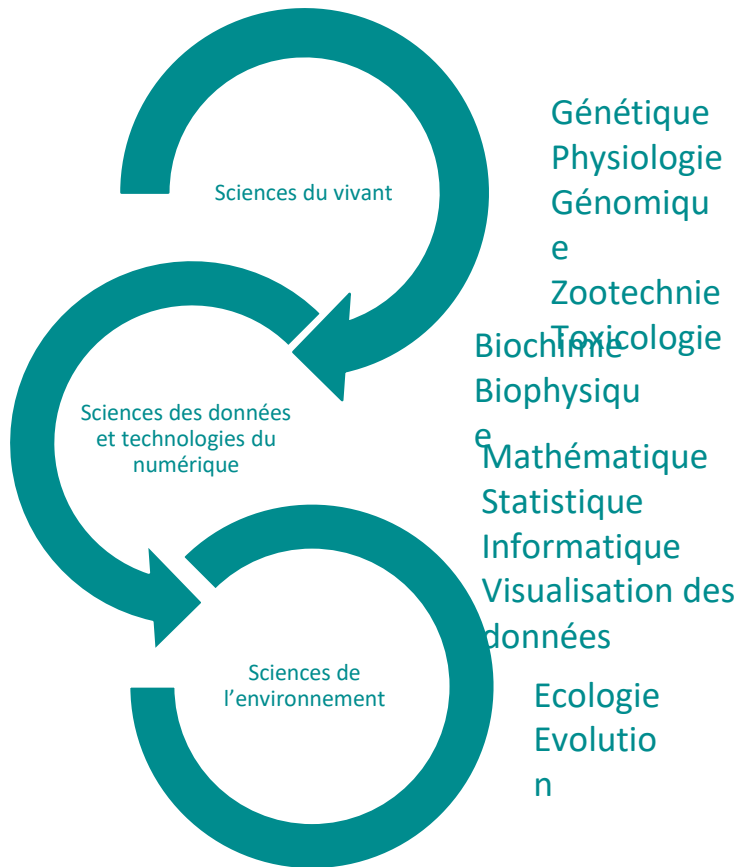
Humphreys et al (2021)



<https://alphafold.ebi.ac.uk/>

# DIGIT-BIO

## Biologie numérique pour explorer et prédire le vivant



- Décloisonner les communautés : règnes/objets, échelles, disciplines
- Soutenir les recherches sur des objets d'étude plus complexes, en mobilisant les approches interdisciplinaires et dans un cadre *open/fair* science
- Accompagner la « mathématisation » de la biologie en combinant défis en biologie et défis en sciences des données et du numérique

# La cellule d'animation IA de DIGIT-BIO



**Christèle Robert-Granié**  
DR1 INRAE, GA  
UMR GenPhySE



**Julien Chiquet**  
DR2 INRAE, MathNum  
UMR MIA Paris-Saclay



**Marie-Laure Martin**  
DR2 INRAE, BAP  
Institut of Plants Sciences Paris-Saclay  
UMR MIA Paris-Saclay

## Nos objectifs

- Mettre en place un vocabulaire partagé sur les méthodes de l'IA
- Sensibiliser aux questions liées à l'IA propres aux sciences du vivant
- Identifier des questions biologiques pour lesquelles des développements en IA méritent d'être poursuivis

# Une animation en trois temps pour construire une communauté

## **Séquence 1: concepts en IA**

Mettre en place un vocabulaire partagé sur les méthodes de l'IA  
Sensibiliser aux questions liées à l'IA propres aux sciences du vivant

## **Séquence 2: tour d'horizon et illustrations (2ème semestre 2022)**

Présenter des projets en science du vivant et utilisant l'IA

## **Résidentiel : pour aller plus loin ensemble (2023)**

Identifier des questions biologiques pour lesquelles des développements en IA méritent d'être poursuivis

# Exemple en apprentissage supervisé (prédiction/régression)

## Projet digit-bio GenIALearn

### Consortium:

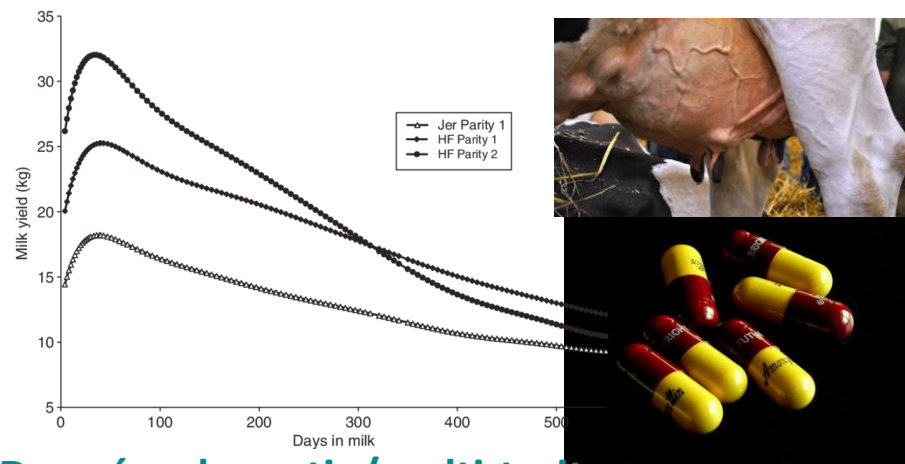
- Généticiens GA – GABI (équipe BIGE et G2B)
- Statisticiens/informaticiens (MathNum - MIA Paris, IBISC - Univ. d'Évry)

**Objectif:** Évaluer les performances des méthodes de Machine Learning (méthodes d'ensemble et apprentissage profond) pour la prédiction conjointe de multiples caractères, par intégration de données massives de génotypage.

**Jeu de données:** 100,000 Bovins x 50,000 SNP x 30 caractères

### Compétiteurs:

- État de l'art: BLUP (modèle mixte); méthodes bayésiennes
- Méthodes ensemblistes: random forest, gradient boosting
- Apprentissage profond : réseaux de neurones



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
0.12	0.08	0.05	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
0.15	0.10	0.07	0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	
0.18	0.12	0.09	0.06	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	
0.21	0.14	0.11	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	
0.24	0.16	0.13	0.10	0.08	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	

100,000 lignes,  
30 colonnes,  
corrélation,  
normalisation

**Données de sortie/multi-traits**  
traitement, production, bien être



**Apprentissage supervisé**

→ construction d'un sous espace de représentation des entrées calibrés pour optimiser la prédiction des sorties

- gourmand en données
- gourmand en temps de calcul

**Méthodes ensemblistes**

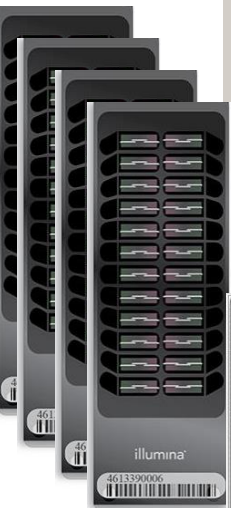
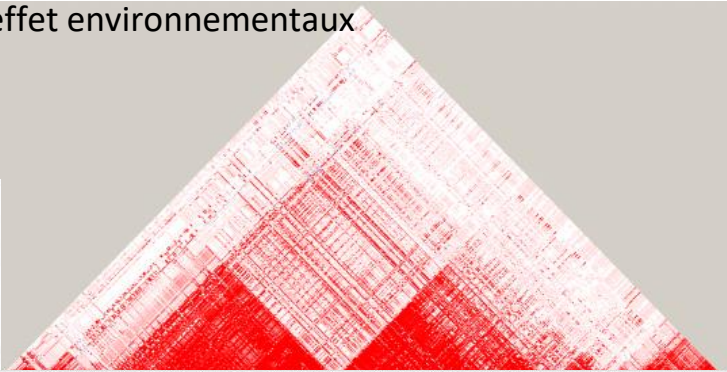
- génération de combinaisons d'effet non linéaires
- sélection de variable
- rééchantillonnage

**Apprentissage profond**

- choix d'une architecture
- espace de représentation non-linéaire

**Données d'entrée/prédicteurs**

Génotypage, effet environnementaux



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
0.12	0.08	0.05	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
0.15	0.10	0.07	0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
0.18	0.12	0.09	0.06	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
0.21	0.14	0.11	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
0.24	0.16	0.13	0.10	0.08	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

100,000 lignes, 50,000 colonnes, structure cachée

# Exemple en apprentissage supervisé (prédiction/classification)

## Projet digit-bio BovMovie2Pred

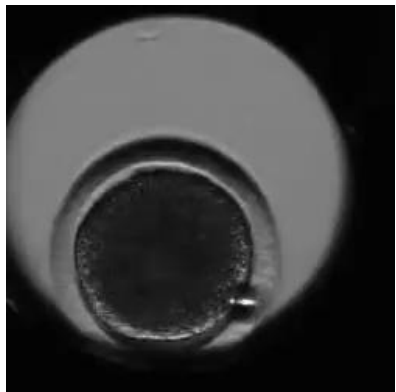
### Consortium:

- Biologiste reproduction animale (PHASE – UMR BREED)
- Statisticiens/informaticiens (MathNum - MIA Paris-Saclay, MaIAGE, INRIA - SERPICO)

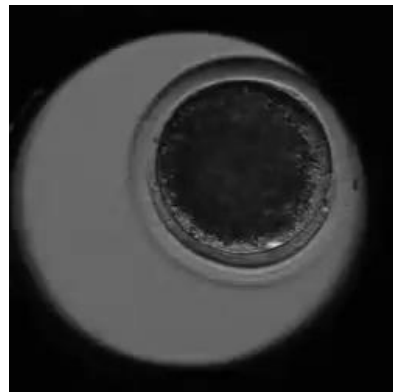
**Contexte:** Les performances actuelles de la FIV et du transfert embryonnaire chez les bovins avoisinent 30 %. La sélection des embryons est basée sur une classification à J7 après fécondation.

**Objectif:** aide à la sélection des embryons par classification précoce de vidéos sans annotations expertes

**Jeu de données:** 300 vidéos d'embryogenèse labélisée en 8 groupes



+... +



Apprentissage (profond)  
(représentation + prédiction)

data challenge

Classifieur à J+1 ... J+7

- évaluation ensemble test



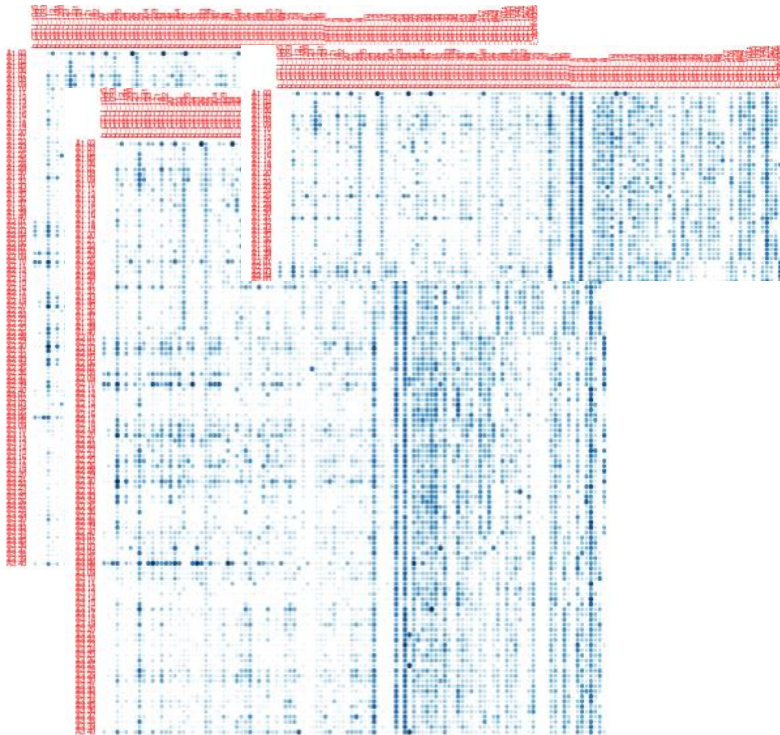
# Exemple en apprentissage non supervisé

## Clustering, réduction de dimension, visualisation

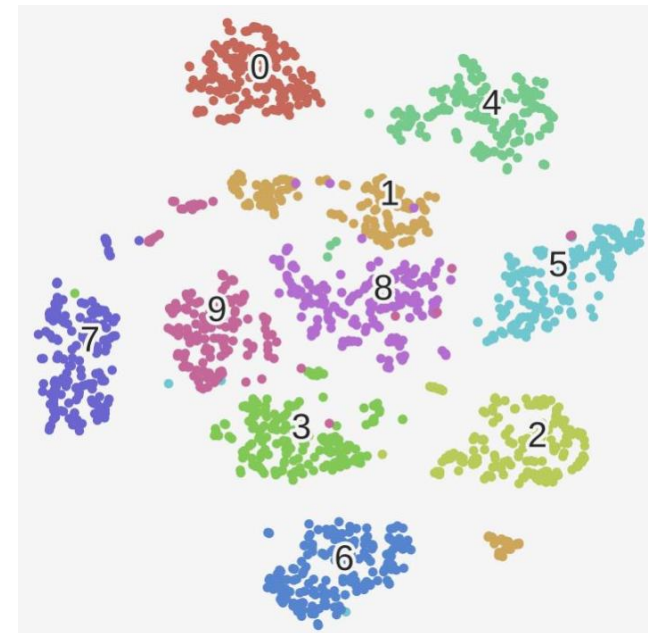
**Contexte:** méta-analyse de communautés bactériennes de sols observés par séquençage

**Objectif:** clustering; visualisation; étude de la spécificité des sols; groupe/pattern de communautés bactériennes partagées dans différents environnements

**Jeu de données:** 8 études de sols (5,000 échantillons, 1,800 OTU communes)



Apprentissage profond  
(représentation)  
e.g. auto-encoder



+ modèle génératif (nouveau "sol" plausible)

# Exemple en Natural Language Processing (NLP) et web sémantique



**CovidOnTheWeb**

<https://github.com/Wimmics/CovidOnTheWeb>

**SPARQL** <https://covidontheweb.inria.fr/sparql>

<https://covidontheweb.inria.fr/ict/>

<https://doi.org/10.5281/zenodo.3833753>

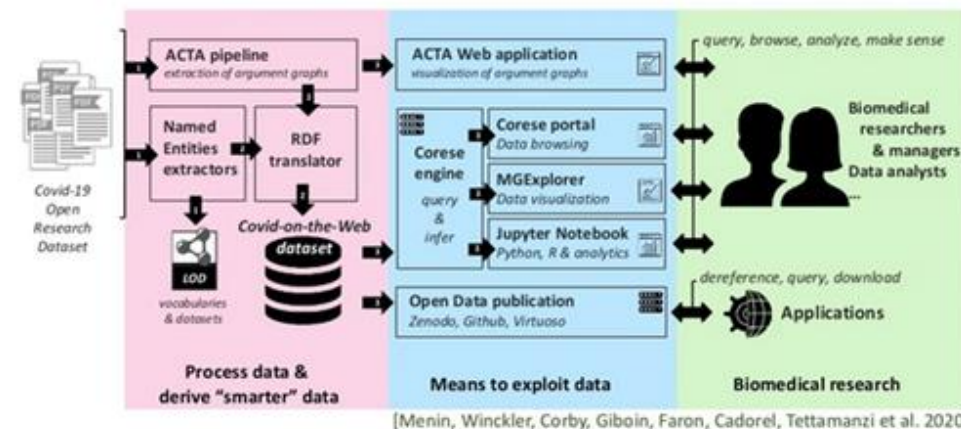
Fabien Gandon, Franck Michel, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Catherine Faron, Raphaël Gazzotti, Alain Giboin, Santiago Marro, Tobias Mayer, Aline Menin, Mathieu Simon, Serena Villata, and Marco Winckler

*inria* UNIVERSITÉ CÔTE D'AZUR *oip* *ias*

F A I R LOD W3C

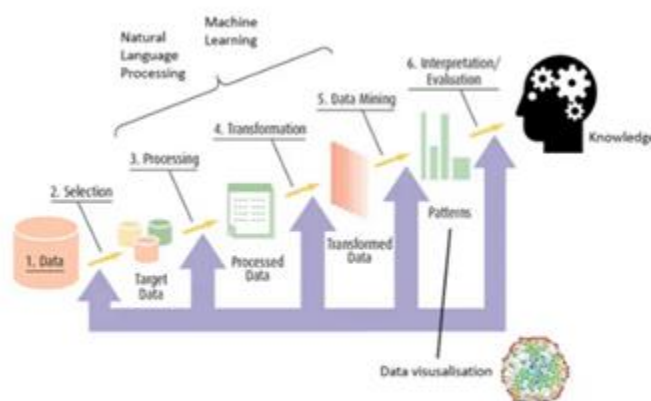



## COVID ON THE WEB [ISWC 2020, IC 2021]



## Domaines d'applications

- Biologie
- Génomique
- Santé
- Bibliométrie : extraire la connaissance
- Analyse de sons, langage
- ....



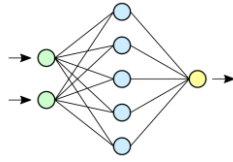
- Phase de prétraitement** : du texte aux données
- Phase de standardisation** : interopérabilité des différents corpus de données
- Phase d'apprentissage** : des données au modèle, donner du sens aux données, visualisation
- Phase de validation**

# Séquence 1: concepts en IA

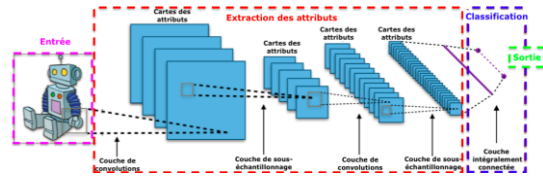
Date		Lieu	Orateur	Titre
31/01/22	14:00 – 17:00	distanciel	<b>Séance 0</b>	Séance introductive
	14:00 – 14:30		<b>Ouverture</b>	Organisateurs
	14:30 – 17:00		<b>Cours</b>	Liva Ravailova
				Introduction à l'IA et au machine learning
07/04/22	14:00 – 17:00	hybride	<b>Séance 1</b>	Machine learning pour la classification supervisée
	14:00 – 15:30		<b>Cours</b>	Blaise Hanczar
				De la régression logistique aux réseaux de neurones
	16:00 – 17:00		<b>Étude de cas</b>	Fadwa Fatmaoui , Emmanuel Moebel
				Le deep learning et la cryo-tomographie électronique permettent l'identification et la localisation de nucléosomes in situ
23/05/22	14:00 – 17:00	distanciel	<b>Séance 2</b>	Identification de structure par classification non-supervisée
	14:00 – 15:30		<b>Cours</b>	Cathy Maugis
				Du modèle de mélanges aux réseaux de neurones
	16:00 – 17:00		<b>Étude de cas</b>	Sophie Donnet, François Massol
				Apprentissage non supervisé de structures de réseaux écologiques
juillet		distanciel	<b>Séance 3</b>	Identification de structure par réduction de dimension
			<b>Cours</b>	Stéphanie Allasonnière
				De l'ACP au VAE
			<b>Étude de cas</b>	Olivier Gandrillon, Franck Picard
				TBA

En construction pour septembre/octobre : une séance sur le web sémantique et Natural Language Processing (NLP)

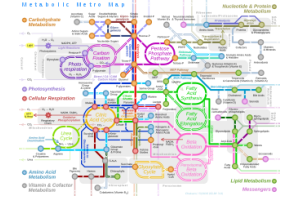
# L'IA nécessite une forte interdisciplinarité



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



IA pour les sciences du vivant



Ne pensez que les autres savent mieux que vous  
Posez des questions  
Faites attention au vocabulaire que vous employez  
Essayez de changer de point de vue  
Apprenons ensemble !

## Séance introductive

# 14h30 - Introduction à l'IA et au Machine learning


*Liva Railavola, Head of AI Research at Criteo AI Lab*



Liva Ralaivola

Head of AI Research at Criteo AI Lab

[Criteo AI Lab](#)

 **Livestorm**  
Chat  
**Questions**