MoDaL

Multi-scale Data Links

Aurélien Cornet, Christian Barillot, Olivier Dameron, Alban Gaignard, Camille Maumet, Richard Redon, Anne Siegel

Contexte et enjeux

Constat

La recherche en biologie est confrontée de manière croissante au besoin de devoir lier des données issues de différentes sources :

- Des technologies OMICS
- De l'imagerie
- Du phénotypage

➢ De nombreux projets nécessitent une coopération multi-disciplinaire et ce besoin se fait ressentir dans tous les domaines

Problématique

- Il est aujourd'hui difficile d'exploiter de façon croisée ces données hétérogènes multi-échelles

  - Objets d'étude distincts
  - Échelles différentes
  - Représentation et analyse des données différentes

➢ Cette problématique se retrouvant dans tous les domaines, l'enjeu serait de mutualiser les efforts

Des enjeux pour le domaine végétal

Le projet EUCLEG vise à réduire la dépendance à l'import de protéines de l'Europe et la Chine

- Diversité des sujets d'études: 5 espèces fourragères et légumineuses différentes étudiées

➢ Sélection de caractères morphologiques avantageux

- Besoin d'associer les variants génétiques aux caractères morphologiques

➢ Combinaison de génomique et d'imagerie

- Challenges :

➢ Faire le lien entre les ontologies des différentes espèces

➢ Intégrer les différents types de données : séquençage et imagerie

➢ Collaboration multi-sites entre plusieurs pays

Des enjeux pour le domaine marin

Le projet OCEANOMICS a pour objectif de mieux comprendre la nature et la fonction du plancton océanique

- Données de prélèvements sur 150 sites à travers le monde, suite du projet TARA OCEAN
  - ➢ Combinaison de séquençage génomique et d'imagerie à haut débit

- Comparaison avec les connaissances actuelles sur la plancton
  - ➢ Besoin d'aligner les formats de données et les ontologies

- Challenges :
  - ➢ Stocker et structurer les données pour assurer leur bonne utilisation
  - ➢ Assurer que les ontologies utilisées soient compatibles avec celles déjà existantes
  - ➢ Partager les données entre plusieurs sites

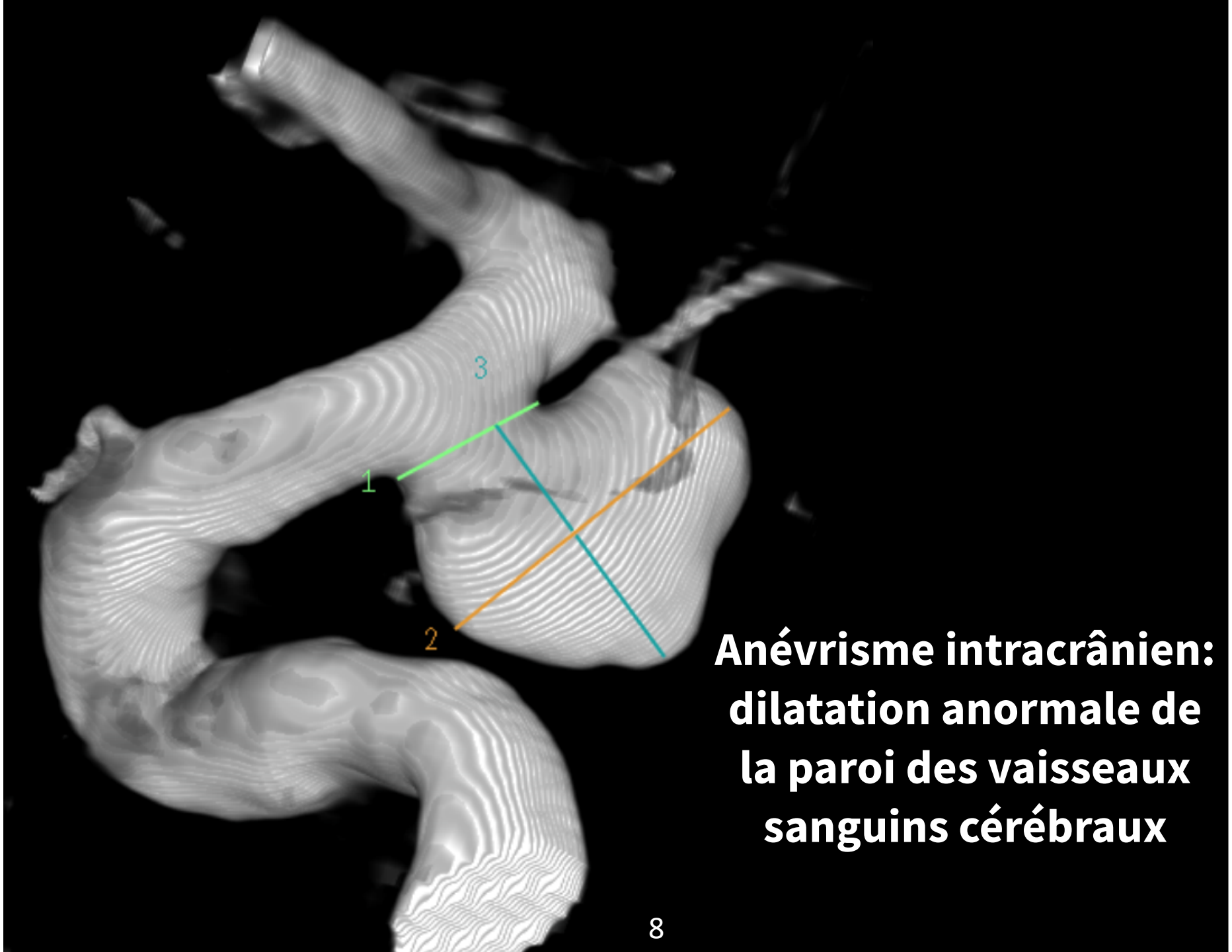
Des enjeux pour le domaine médical

Le but du projet INEX-MED est de développer une infrastructure pour l'exploitation de données hétérogènes dans le cadre des anévrismes intracrâniens

- Différentes échelles: de l'organe à la cellule

- Croisement de données génomiques, cliniques et d'imagerie

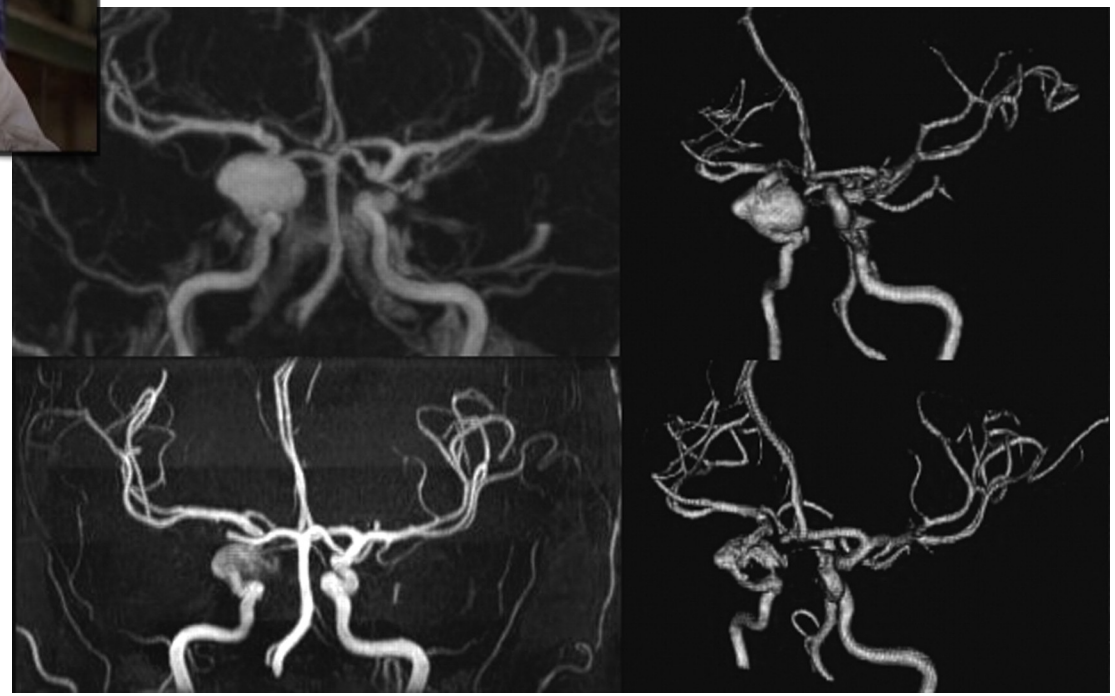
Rentrons un peu plus dans les détails



Anévrisme intracrânien: dilatation anormale de la paroi des vaisseaux sanguins cérébraux

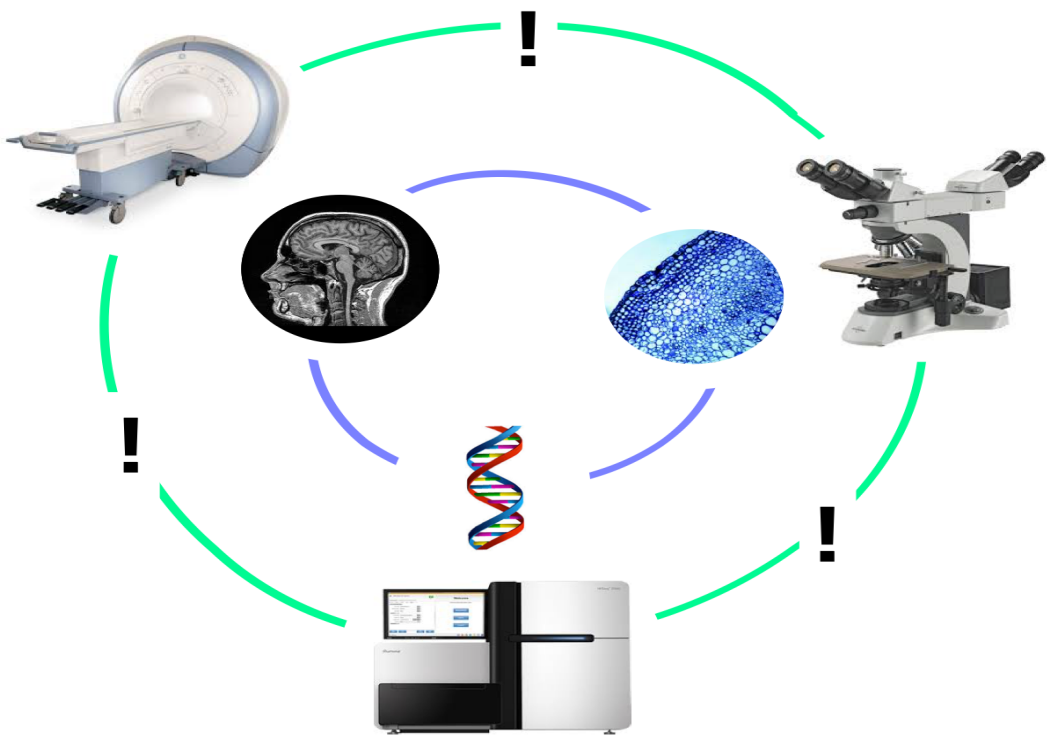
```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 39458_2001 39458_2003 39458_2005 39458_2007 39458_2009 39458_2012 39458_3003
39458_39461 39458_39462 39458_39465 39458_39466 39458_CD12099 39458_CD12100 39458_CD12252 39458_CD12350 39458_CD12380 39458_CD12406 39458_CD12466
39458_CD12480 39458_CD12501 39458_CD12586 39458_CD12587 39458_CD12588 39458_CD12704 39458_CD13016 39458_CD13018 39458_CD13178 39458_CD13363
1 752566 rs3094315 A G . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/1 0/0 0/1 0/1 0/1 0/0 0/1 0/1 0/1 0/1 0/1 0/1 0/0 0/0
0/1 0/1 1/1
1 754105 rs12184325 C . . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0
0/0 0/0 0/0
1 754182 rs3131969 G A . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/1 0/0 0/1 0/1 0/1 0/0 0/1 0/1 0/1 0/1 0/1 0/1 0/0 0/0
0/1 0/1 1/1
1 758626 rs3131954 T . . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0
0/0 0/0 0/0
1 768448 rs12562034 G A . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/0 0/0 0/0 0/1 0/0 0/1 0/1 0/0 0/0 0/0 0/1 0/0 0/0 0/0
0/0 0/0 0/0
1 792534 rs2518996 G . . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0 0/0
0/0 0/0 0/0
1 798801 rs12132517 G A . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/1 0/0 0/0 0/1 0/0 1/1 0/1 0/0 0/0 0/0 0/1 0/1 0/0 0/0
0/0 0/0 0/0
1 798959 rs11240777 G A . . PR GT ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. ./. 0/1 0/0 0/0 0/1 0/1 1/1 0/1 0/1 0/1 0/1 0/1 0/1 0/0 0/0
0/0 0/1 0/0
```

ente atteir	circonstances d	consommation c	quantite alcool p	statut phenotypi	nb d anevrismes	date du 1er diag	AIC 1 Grand axe	AIC 1 Localisatic	AIC 2 Grand axe	AIC 2 L			
	Fortuite	Sevrage > 3ans	96.6	certain	1.0	1/22/16	7.0	5					
	Dépistage familial	Tabagisme actif	277.2	certain	1.0	3/3/16	4.0	18					
	Fortuite	Sevrage < 3 ans	109.2	certain	3.0	12/29/15	6.0	4	14.0	4			
	Rupture AIC	Jamais fumé	0.0	certain	2.0	4/3/09	7.0	M.C.A.Gche	5.0	M.C.A.Dte			
	Rupture AIC	Sevrage > 3ans		certain	4.0	1/1/97		18		17			
	Fortuite	Jamais fumé	25.2	certain	1.0	1/22/16	5.0	21					
	Fortuite	Sevrage < 3 ans	50.4	certain	3.0	9/22/15	5.0	18	6.9	118			
	Rupture AIC	Jamais fumé		certain	1.0	5/1/05	6.0	118					
	Rupture AIC	Jamais fumé		certain	1.0	10/16/10	9.0	107					
	Fortuite	Tabagisme actif		certain	1.0		7.0	105					
	Rupture AIC	Sevrage > 3ans	58.8	certain	1.0	1/9/10	2.0	10					
	Rupture AIC	Jamais fumé		certain	1.0		4.0	10					
	Fortuite	Jamais fumé	46.2	certain	1.0		9.0	10					
	Fortuite	Sevrage > 3ans		certain	2.0		9.0	18	6.0	10			
	Fortuite	Jamais fumé		certain	1.0	8/10/16	26.0	104					
	Dépistage familial	Jamais fumé		certain	1.0	1/1/98	6.0	118					
	Fortuite	Tabagisme actif		certain	1.0		6.0	20					
	Rupture AIC	Tabagisme actif	25.2	certain	4.0	3/3/16	7.0	17	4.0	104	18.0	105.0	3.0
	Rupture AIC	Tabagisme actif	336.0	certain	1.0		5.0	132					
	Rupture AIC	Jamais fumé		certain	3.0	1/1/10	3.0	10	4.0	10	3.0	10.0	
	Rupture AIC	Tabagisme actif		certain	2.0	5/23/05	12.0	10	3.0	106			
	Dépistage familial	Sevrage < 3 ans	378.0	certain	1.0	1/1/16	2.0	18					
	Rupture AIC	Jamais fumé		certain	1.0	2/5/12	25.0	6					
	Rupture AIC	Jamais fumé		certain	1.0	1/1/03	6.0	6					
	Rupture AIC	Tabagisme actif	84.0	certain	3.0	10/1/16	4.0	10	2.0	118	2.0	105.0	
	Rupture AIC	Jamais fumé		certain	2.0	3/15/16	5.0	118	1.5	18			
	Dépistage familial	Sevrage > 3ans		certain	1.0		2.0	18					
	Rupture AIC	Tabagisme actif	0.0	certain	3.0	6/14/16	2.0	18	5.0	21	15.0	121.0	
		Sevrage > 3ans	12.6	certain	3.0	11/1/14	8.0	10	2.0	18	3.0	118.0	
	Rupture AIC	Sevrage > 3ans	25.2	certain	1.0	8/21/16	2.0	10					
	Fortuite	Jamais fumé		certain	1.0	12/8/16							
	Rupture AIC	Sevrage < 3 ans	470.4	certain	2.0	3/24/15	2.5	10	3.0	12			
	Fortuite	Tabagisme actif	12.6	certain	1.0	1/1/99	8.0	10					
	Fortuite	Tabagisme actif	0.0	certain	2.0	7/16/15	11.0	118	5.5	18			
	Fortuite	Jamais fumé		certain	1.0	12/18/14	7.0	10					
	Rupture AIC	Jamais fumé	0.0	certain	4.0			18	4.0	106	4.0	118.0	3.0
		Jamais fumé	0.0	certain	1.0								
		Jamais fumé	0.0	certain	2.0		3.5	12	2.5	18			
	Fortuite	Jamais fumé		certain	1.0	7/23/12	5.0	132					
	Rupture AIC	Jamais fumé	0.0	certain	2.0	4/24/15	5.5	105	13.0	5			
	Fortuite	Tabagisme actif	42.0	certain	1.0	1/1/15	7.0	105					
	Dépistage familial	Sevrage < 3 ans	21.0	certain	1.0	6/20/16	9.0	109					
	Fortuite	Tabagisme actif	11.76	certain	2.0	6/9/16	2.8	5	3.0	105			
	Rupture AIC	Sevrage < 3 ans	35.28	certain	1.0	9/4/16	5.9	111					
	Fortuite	Sevrage > 3ans	0.0	certain	1.0	4/1/16	8.5	121					
	Rupture AIC	Jamais fumé	11.76	certain	1.0	2/19/14	6.5	10					
	Rupture AIC	Sevrage > 3ans	35.28	certain	1.0	9/5/15	6.7	10					
	Fortuite	Jamais fumé	11.76	certain	1.0	2/22/16	6.0	111					
	Dépistage familial	Tabagisme actif	11.76	certain	1.0	4/1/16	9.0	107					
	Fortuite	Sevrage > 3ans	203.28	certain	2.0	1/1/14	4.4	17	5.2	18			



Anévrismes intracrâniens : challenge

Objectif : mieux **comprendre** les mécanismes et causes des anévrismes pour **prévenir** les risques de rupture



Challenges

- Interopérabilité des outils
- Diversité des données
- Différentes ontologies
- Collaboration multi-site, décentralisation

Un contexte local unique

- Le Grand-Ouest est le seul site français qui héberge des nœuds nationaux en imagerie (in-vivo, in-vitro) et génomique

France Life Imaging

Neurinfo

Imagerie *in vivo*

France Bio Imaging

MicroPicell

Imagerie *in vitro*

France Bioinformatique

GenoBiRD

GenOuest

Génomique

MoDal « Multi-Scale Data Links »

Objectifs

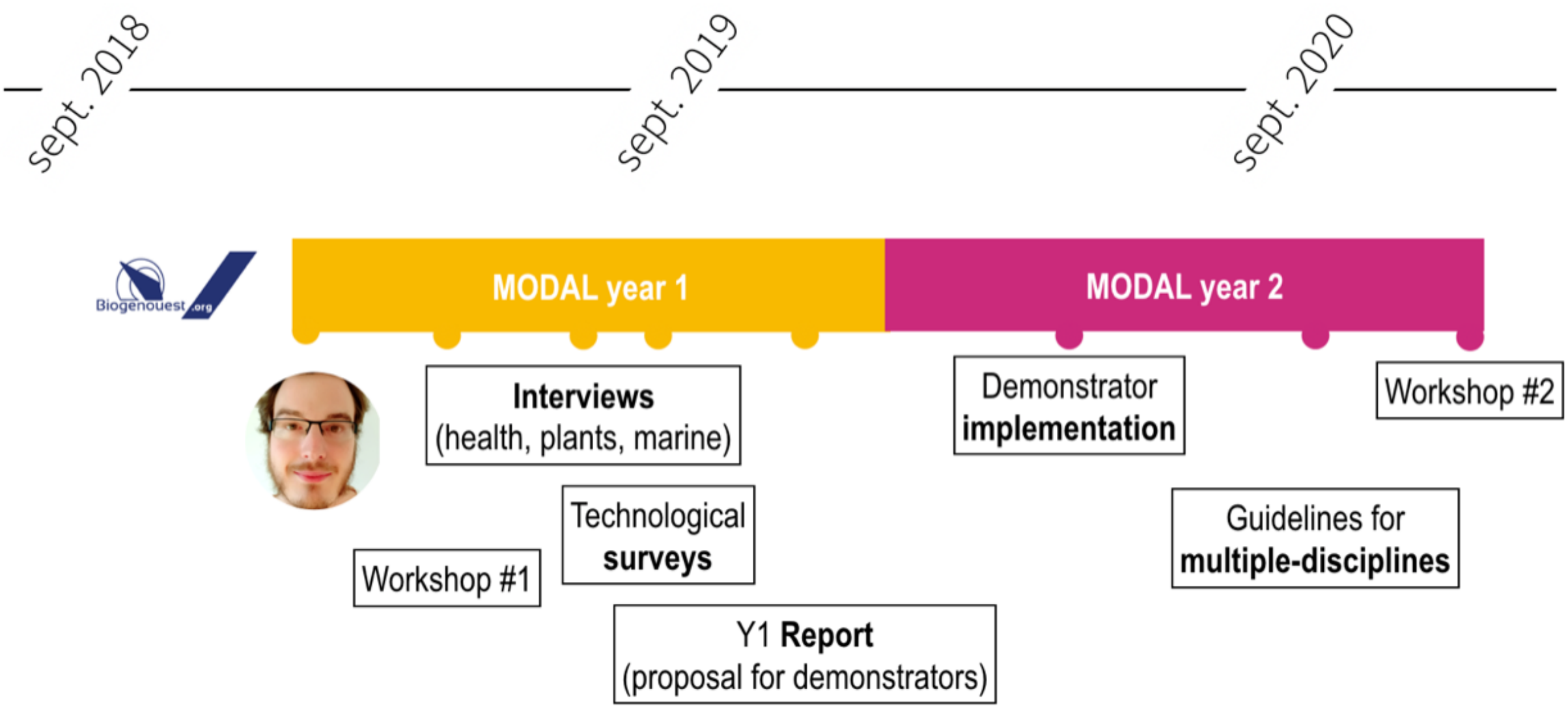
Décloisonner les acteurs et les ressources à l'echelle du grand ouest pour envisager **l'exploitation algorithmique** (sémantique et statistique) de **données hétérogènes multi-échelles**

Résultats

1: Etat des lieux des acteurs et infrastructures du Grand-Ouest

2: Synthèse des problématiques et verrous communs

MoDaL : 1ère année


sept. 2018
sept. 2019
sept. 2020
Biogenouest.org
MODAL year 1
MODAL year 2
Interviews
(health, plants, marine)
Demonstrator
implementation
Workshop #2
Technological
surveys
Guidelines for
multiple-disciplines
Workshop #1
Y1 Report
(proposal for demonstrators)

Workshop MoDaL 11/07/19 à l'INRIA Rennes

- 26 participant-e-s venant de Rennes, Nantes, Brest, Nice et Paris

- 7 présentations et des discussions libres autour de MoDaL

- Large diversité de domaines de compétence, d'échelles d'étude et de modèles

- Rapport détaillé de la journée :

Quelques sujets abordés

- Institut du cerveau et de la moelle épinière, Paris

➢ Mise en place d'un « Datalake », une base de données pour stocker et analyser les données issues de l'imagerie, des différentes technologies OMICS, humaines et animales, et ce pour les données brutes d'acquisitions ou les données traitées.

- Laboratoire I3S, Nice

➢ Utilisation du machine learning pour récupérer des mots clés dans les données de patients afin de les croiser avec des connaissances issues de la littérature pour améliorer les diagnostics pour la médecine de précision

- IFREMER, Brest

➢Analyse croisée sur les espèces marines à l'aide d'imageries satellite, de données de prélèvement et de recensement, et de données de séquençage génomique

Des défis spécifiques

- Taille importante des données, limite de stockage
- Limitation par la puissance de calcul
- Confidentialité des données humaines
- ...

Beaucoup des verrous qui ressortent sont spécifiques d'un ou deux domaines

Un verrou commun possible ?

Les données et outils ne sont pas assez réutilisées en dehors de leur but initial

- Faciliter la collaboration inter et intra-instituts

➢ Augmenter l'interopérabilité des outils et des ontologies

- Favoriser des outils communautaires dédiés et interopérables plutôt qu'un unique outil généraliste « usine à gaz » mais peu précis

- Utiliser des ontologies standardisées ou des versions « maisons » spécifiques mais connectées

➢ Éviter les erreurs d'interprétations en utilisant les mêmes termes

Extension du spectre des domaines

- Bioinformatique
- Neuroimagerie
- Clinique
- Végétal
- Marin

Interview de 39 acteurs sur les problématiques rencontrées

Questions et sujets de discussion
Pouvez-vous vous présenter et me dire quels sont vos sujets de recherches actuels ?
Quels sont vos domaines de compétences ?
Que fait votre équipe/laboratoire en général ?
Pouvez-vous me parler de projets passés ou en cours auquels vous avez pris part ?
Quels types de données vous ou votre équipe gérez-vous ?
Avez-vous à croiser des données de différents types au cours de vos recherches ?
Si oui, rencontrez-vous des difficultés ?
Comment sont stockées ces données ?
Pouvez-vous me parler d'outils que vous utilisez ?
Discussion libre
Intérêt des données de nos jours
Pourquoi croiser les données ?
Les verrous actuels du domaine d'étude (de l'interviewé) ou de la biologie en général

Résumé des acteurs et synthèse

- Echanges avec l'IRISA, le LTSI et l'IGDR à Rennes
- L'institut du Thorax et au LS2N à Nantes
- L'IRHS d'Angers
- La station Biologique de Roscoff

Synthèse des interviews pour recouper les challenges communs

- Les nouveaux éléments se recoupent avec la piste du Workshop
- Favoriser la réutilisation des ressources en dehors du projet associé

Synthèse générale MoDaL

Ontologies	Données	Outils
Les ontologies sont conçues dans le cadre du projet mais n'ont pas de connecteurs	Les structures des bases de données et des métadonnées utilisées ne sont pas assez explicites	Les outils d'analyse utilisent des formats bien spécifiques et répondent à des problèmes précis du sujet d'étude en cours
Difficulté de croiser les données annotées avec celles d'autres projets	Une fois le projet terminé, il est difficile de remobiliser les données sans les informations pour les décrypter	Manque de flexibilité, difficile à utiliser sur d'autres cas

La plupart des ressources déployées le sont dans le cadre d'un projet spécifique, et leurs utilisations ne dépassent pas l'horizon de ce projet

Suite de MoDaL

Démonstrateur technologique

Sur la base de cet état des lieux, le prochain objectif de MoDaL est le développement d'un démonstrateur technologique adressant les verrous identifiés

Les différentes pistes :

- Favoriser la remobilisation des données et des outils
  - Guide d'utilisation sur les formats de fichiers et les ontologies
  - Exemple d'intégration mettant à profit des données d'anciens projets dans la santé, le végétal et le marin

Feuille de route

- Terminer et publier le document de synthèse de l’état des lieux

- Développer le démonstrateur technologique

- Continuer à solliciter la communauté scientifique pour favoriser la coopération autour du projet MoDaL

➢ Organiser le second Workshop

Questions ?

Biogenouest

cnrs

Inserm

Inria

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES