laboratoire eau environnement systemes urbains

WaterOmics

Analyse non-ciblée : Traitement et stockage de données

Café science 27/11/2020

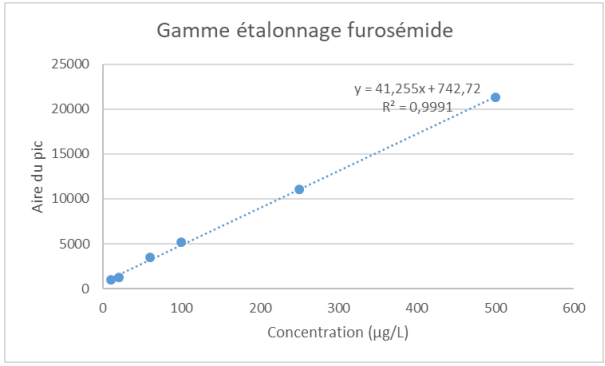
Nina Huynh & Julien Le Roux

L'analyse non-ciblée

Analyse ciblée

On sait ce qu'on cherche et comment le trouver

QUANTIFICATION



Analyse en suspect

On sait ce qu'on cherche

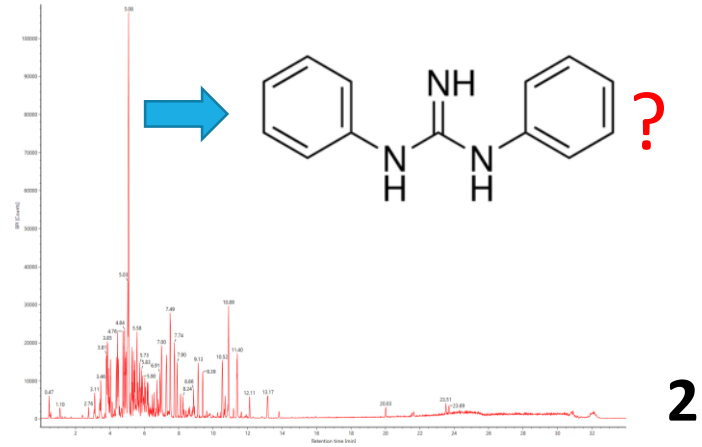
DÉTECTION (absence / présence)

Molécule	Sc1	Sc1d	Sc2	Sc2d	V01	V01(	V02	V02(
4-Aminoantipyrine	N	N	N	N	O	O	O	N
Acetaminophen	O	O	O	O	O	O	O	O
Aciclovir	O	O	O	O	O	O	O	N
Aténolol	N	N	N	N	N	N	N	O
Atorvastatin	O	O	O	O	N	N	N	N
Bezafibrate	N	O	N	N	O	N	O	N
Canrenone	N	O	N	O	O	O	O	O
Clopidogrel	N	N	N	N	O	O	O	N
Codeine	O	O	O	O	O	N	O	O
Dicyclanil	O	O	O	N	O	O	O	O
Doxepin	N	N	N	N	O	O	O	O
Enrofloxacine	O	O	O	N	N	N	N	O

Analyse non ciblée

On ne cherche rien, on essaie d'attribuer une identité à ce qu'on voit

IDENTIFICATION



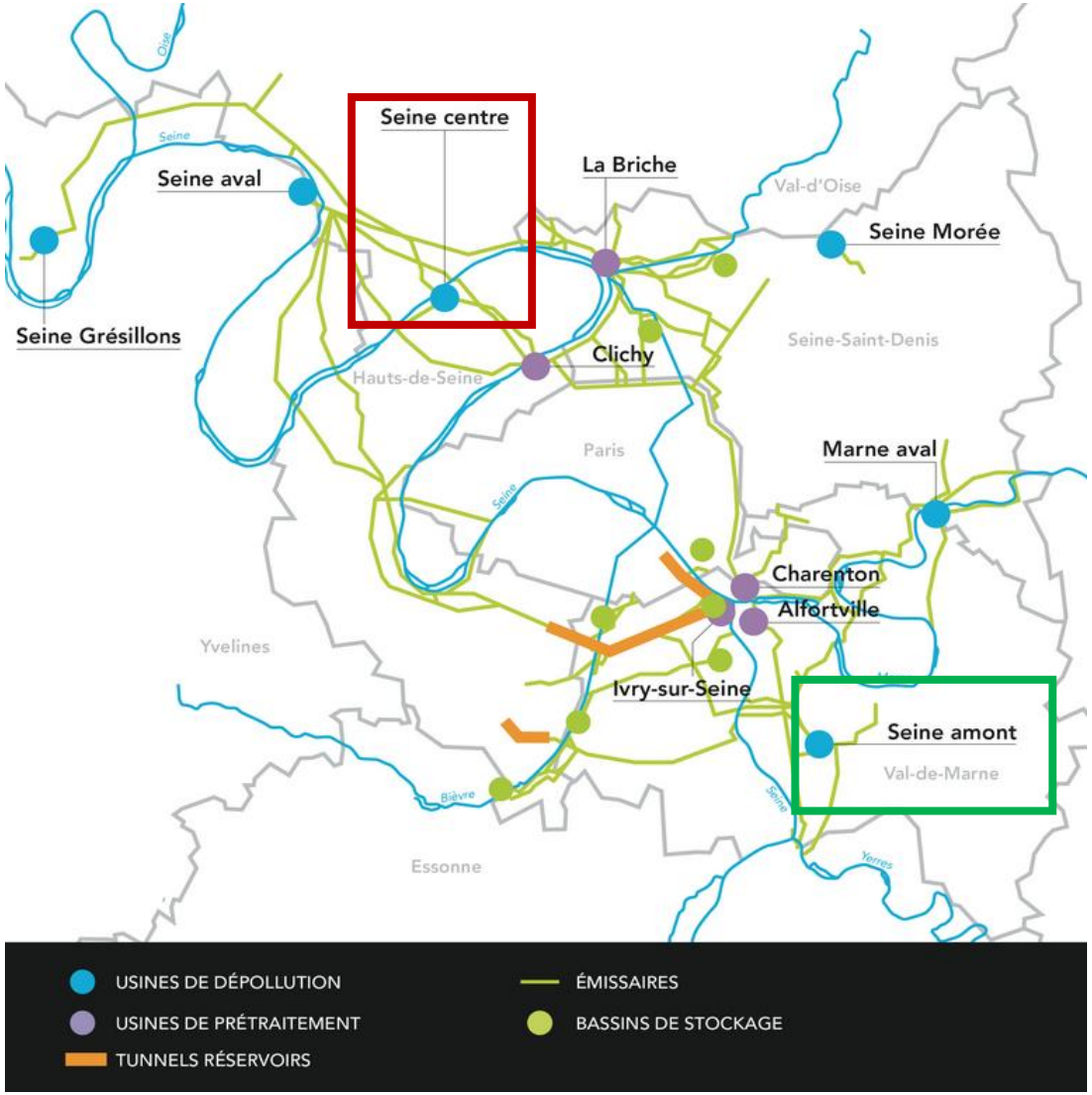
Observatoire des micropolluants organiques dans les eaux résiduaires urbaines

OPUR

SIAAP
Service public de l'assainissement francilien

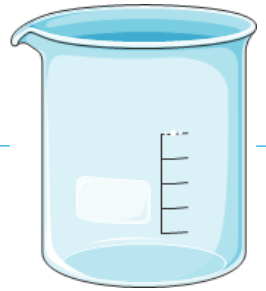
Objectifs :

- Suivre l'**évolution** des eaux résiduaires urbaines sur le **long terme**
- Etudier la différence spatiale de ces eaux
- Acquérir une **base de donnée** HRMS pour d'éventuelles **études rétrospectives**



http://www.siaap.fr/fr/equipements/le-reseau/les-emissaires/

Analyses réalisées



SIAAP
Service public de l'assainissement francilien

Analyse des paramètres globaux :

- pH
- Conductivité
- Turbidité
- MES
- COD
- DBO
- DCO
- Ammonium
- Nitrites
- Nitrates
- Azote Kjedahl
- Phosphore
- Orthophosphate

leesu
laboratoire eau environnement systemes urbains

Ecotoxicologie

Fluo 3D

Micropolluants organiques

Ciblé
Biocides

Non-ciblé

Suspect
Biocides
Pharmaceutiques

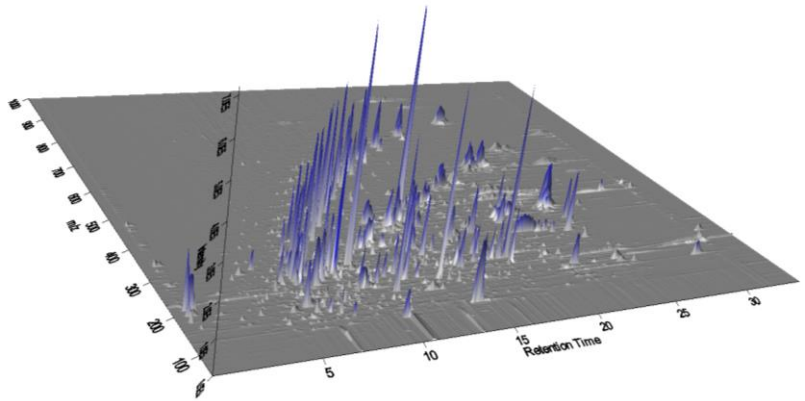
Analyse par spectrométrie de masse haute résolution (HRMS)

Présentation des données HRMS

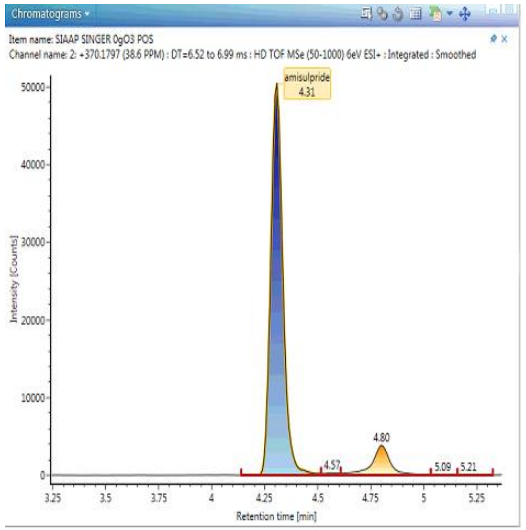
1 échantillon

~ 120 échantillons attendus sur cette 1ère année d'observatoire

6 injections (Triplicata dans 2 modes d'ionisation)

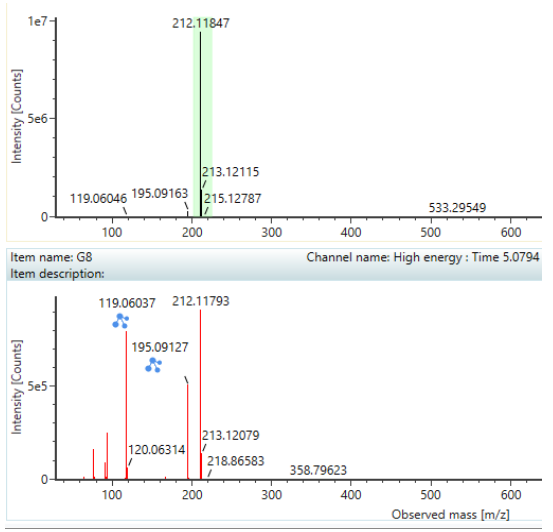


~ 20 000 pics par chromatogramme

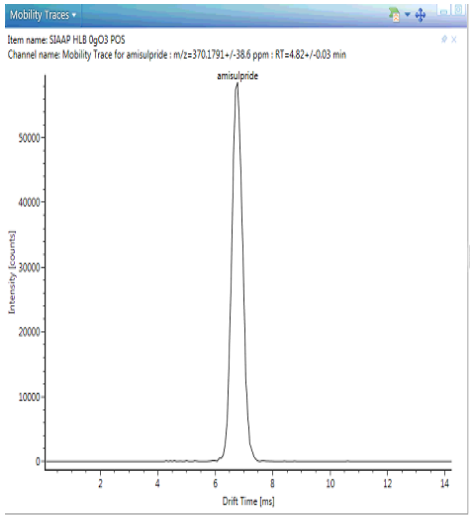


Temps de rétention

Séparation physico-chimique



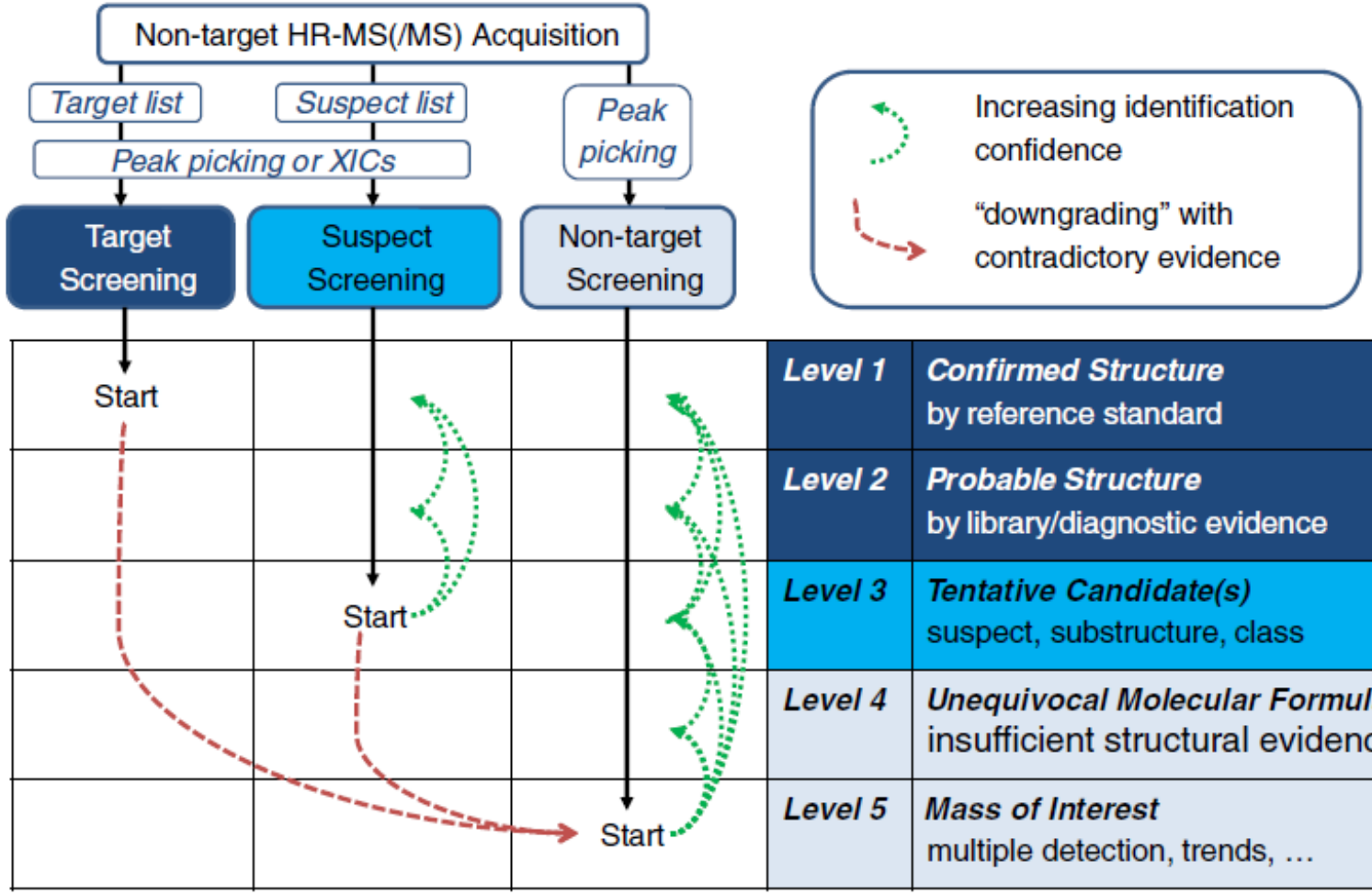
Spectres de masse



Temps de drift

Séparation par taille et forme

Identification de molécules inconnues



Achat d'un **standard analytique** pour confirmation

Comparaison avec des **bases de données**

Déduction de **structure** grâce aux **fragments**

Calcul de **formule brute** grâce au **massif isotopique**

Schymanski et al, Environmental Science and Technology, 2014

Format et taille des données

Méthode 1

Méthode 2

UNIFI SCIENTIFIC INFORMATION SYSTEM

Acquisition des données

proteowizard

Conversion des données en format universel (*.mzXML)

2 à 6 Go par injection

1 à 2 Go par injection

UNIFI SCIENTIFIC INFORMATION SYSTEM

Prétraitement des données sous format fournisseur (*.uep)

Prétraitement des données sous format *.rds

~ 600 Ko par injection

Export de données en *.csv

~ 500 Ko par injection

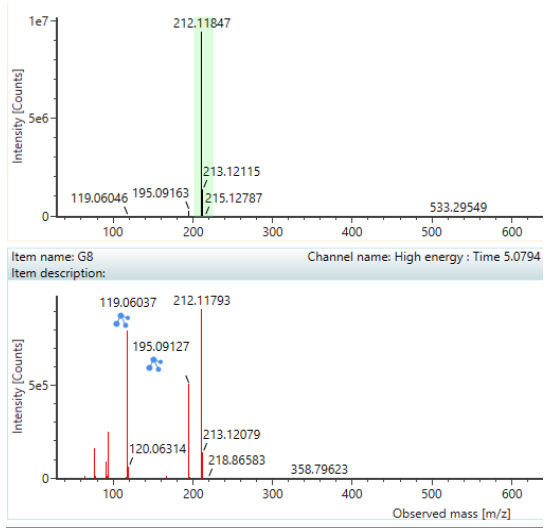
RStudio®

Traitement des données

Résultats exportables en *.csv ou *.png

Quelle différence entre les méthodes?

Méthode 1 Pas de drift time!



Méthode 2 Pas de données spectrales!

PrimaryId	m/z	RetentionTime	DriftTime	Comment	020317 Compans brut_replicate_1	020317 Compans brut_replicate_2	020317 Compans brut_replicate_3	020317 Compans Noue_replicate_1	0203 Noue
212.11777_5.020_4.52	212.1178	5.020	4.52	NA	3245065.2	3112848.8	3088330.2	4837.55	11547
389.23306_8.412_6.09	389.2331	8.412	6.09	NA	1525821.9	1444938.0	1488661.7	1155.88	1210.
447.29234_9.753_6.65	447.2923	9.753	6.65	NA	1420615.2	1327129.0	1262714.4	512.71	607.8
317.18756_8.452_5.46	317.1876	8.452	5.46	NA	1408639.8	1453012.8	1346995.9	107.75	329.2
277.19022_6.901_5.42	277.1902	6.901	5.42	NA	1383569.6	1392927.0	1339906.1	17009.77	23157
331.19673_7.700_5.64	331.1967	7.700	5.64	NA	1380783.6	1334140.2	1232307.6	907.86	1305.
273.16556_7.471_5.00	273.1656	7.471	5.00	NA	1128124.1	1098739.8	1105216.2	484.85	1798.
319.16759_5.538_5.17	319.1676	5.538	5.17	NA	1036348.4	993732.7	992760.3	265.15	316.1
287.14586_7.661_5.09	287.1459	7.661	5.09	NA	958771.2	954936.3	930564.4	33025.18	40706
481.24925_6.103_6.66	481.2493	6.103	6.66	NA	871858.3	854690.7	872194.9	4104.96	5449.
275.14599_4.311_4.69	275.1460	4.311	4.69	NA	851128.5	840000.0	849433.5	469.63	1547.
361.18180_8.073_5.78	361.1818	8.073	5.78	NA	744457.3	755816.0	728493.8	0.00	0.00

Quelle molécule se cache derrière chaque ligne?

Possibilité d'**analyse rétrospective** pour chercher de **« nouveaux » polluants** dans de **« vieux » échantillons**!

Exemple de fichiers

*.csv (52.1 Mo)

132 499 colonnes

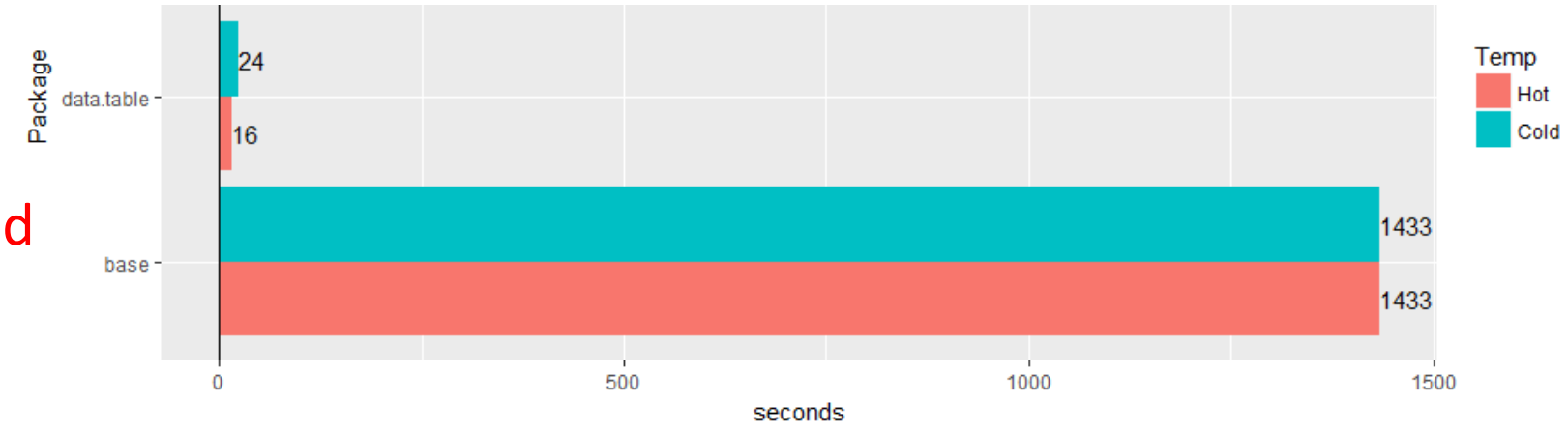
73 lignes

	PrimaryId	Item description	Gender	Group ID	201.88722_0.039_4.36	246.86258_0.044_4.26	217.85990_0.046_4.14
1	m/z	NA	NA		201.8872	246.8626	217.8599
2	RetentionTime	NA	NA		0.0390	0.0440	0.0460
3	DriftTime	NA	NA		4.3600	4.2600	4.1400
4	Comment	NA	NA		NA	NA	NA
5	Blanc EUP 2 POS_replicate_1	NA	NA	blanc	539.3400	0.0000	0.0000
6	SIAAP 0gO3 POS_replicate_1	NA	NA	brut	0.0000	0.0000	0.0000
7	SIAAP 0.2gO3 POS_replicate_1	NA	NA	brut	0.0000	164.8100	0.0000

read.csv : ~ 880 sec = 15 min

data.table::fread : ~2.1 s

hospital data: 30 Millions rows × 125 columns



Data.table → fread

Exemple de fichier

*.csv (9 Mo)

59 589 colonnes

29 lignes

	PrimaryId	Gender	Group ID	Sample type	113.96347_0.026_2.90	141.95804_0.040_3.25	261.07795_0.041_4.44	87.92408_0.043_3
1	m/z	NA			113.9635	141.958	261.0779	87.92408
2	RetentionTime	NA			0.0260	0.040	0.0410	0.04300
3	DriftTime	NA			2.9000	3.250	4.4400	3.00000
4	Comment	NA			NA	NA	NA	NA
5	Blanc_replicate_1	NA	Blanc	Unknown	0.0000	0.000	0.0000	0.00000
6	Blanc extraction-1_replicate_1	NA	BlkExtr	Unknown	0.0000	0.000	0.0000	0.00000
7	Blanc extraction-2_replicate_1	NA	BlkExtr	Unknown	0.0000	0.000	0.0000	0.00000

fread ~ 2.9 sec

*.rds (= objets R : 235 Ko)

+ 3 autres RDS
listes de data.table associées
(total de 12 Mo)

Liste de data.table

readRDS ~ 6.5 ms

- 19 * 1 910 (aires des marqueurs dans les échantillons)
- 19 * 4 (métadonnées des échantillons)
- 1 910 * 2 (propriétés des marqueurs)
- Liste de 19 data.table : nb de marker dans l'échantillon * 10
- 19 * 1 910 (présence/absence)

Stockage et accessibilité des données

Analyse non-ciblée et suspect => Possibilité d'**analyse rétrospective**

Stockage des données brutes (*.uep ou *.mzXML) nécessaire

Actuellement…

Export des données en *.uep

Ordinateur de pilotage

Uniquement pour acquisition

Ordinateur de retraitement

Actuellement utilisé pour le stockage

~ 3,2 To de données *.uep

(en moins de 2 ans)

Base OSU

(en cours d'installation)

Ré-import des données en *.uep

Export/Import très long !!

Problèmes à résoudre…

- Optimiser les étapes de traitement
- Trouver un moyen de stocker des données sur le long terme (observatoire de micropolluants prévu sur 10 ans)
- Trouver une façon d'accéder rapidement aux données pour les retravailler

Merci pour votre attention!