LEESURIALES

Variabilité spatio-temporelle des sources de contaminations microbiologiques des eaux de surface lors des évènements pluvieux

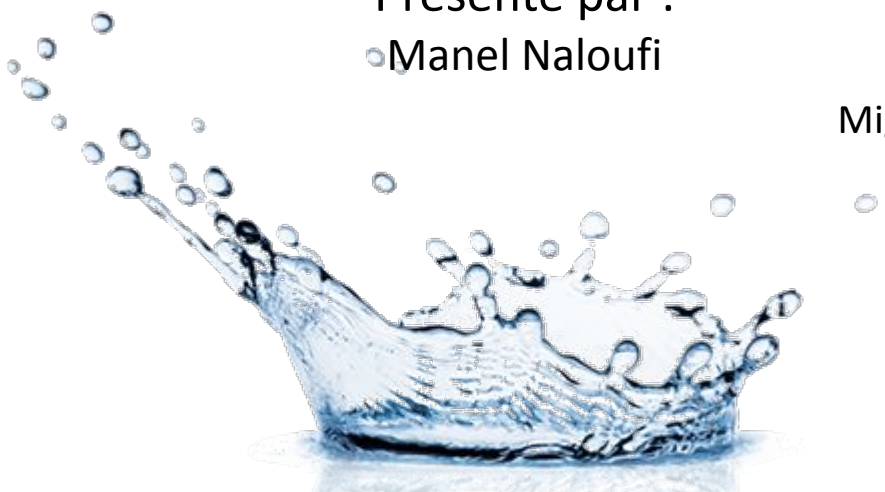
Présenté par :
Manel Naloufi

Encadrants :

Françoise LUCAS (Leesu)

Miguel GILLON-RITZ et Marion DELARBRE (STEA)

Thiago ABREU (LiSSi)



2022

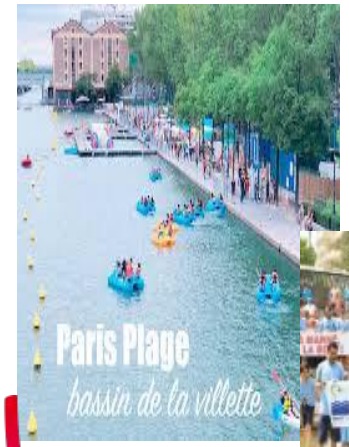
Contexte

Changement climatique

- Baignades sauvages
- Nouveaux sites de baignade (Bassin de la Villette)

⇩

Demande politique & sociétale forte



En Europe.
USA.
Australie…

PARIS 2024

Baignades en rivières urbaines

Risques sanitaire?

Pathogènes émergents

Sources multiples

Introduction

Réglementation 2006/7/CE

Ouvrir une baignade

Suivi de la qualité

Bactéries indicatrices fécales (BIF)

(E. coli et entérocoques intestinaux (EI))

- Collecte d'eau
- Analyse en laboratoire

→ Qualité suffisante ↓ Valeur Seuil

→ Suivi microbiologique

Introduction

Réglementation 2006/7/CE

Ouvrir une baignade

Suivi de la qualité

Bactéries indicatrices fécales (BIF)

(E. coli et entérocoques intestinaux (EI))

Collecte d'eau

Analyse en laboratoire

→ Qualité suffisante → Valeur Seuil

→ Suivi microbiologique

Gérer une baignade

Gestion quotidienne

Alerte et prédiction

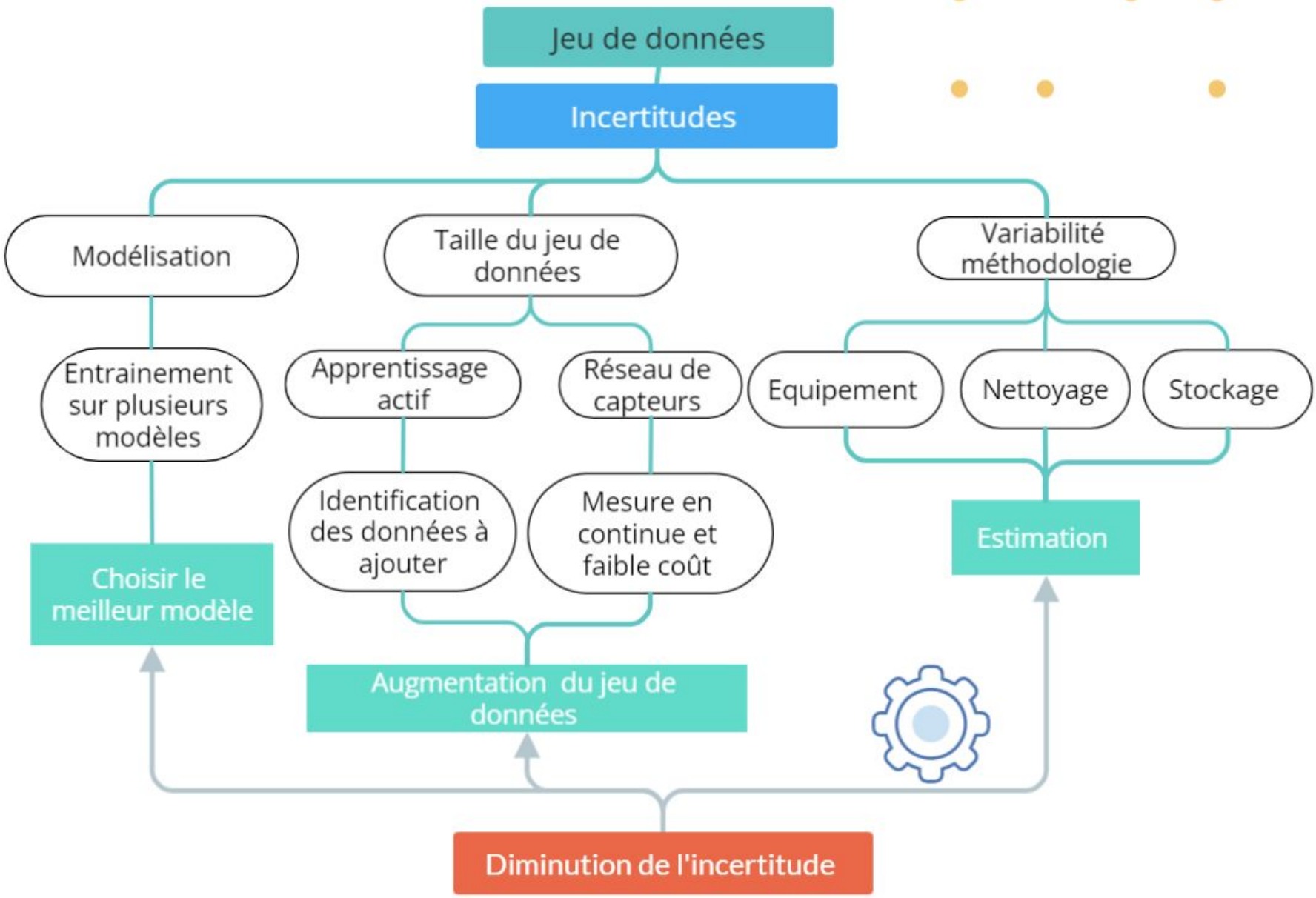
↓

Modélisation

(OMS. 2018)

→ Suffisamment de données

Optimisation de la collecte de données pour la modélisation



Optimisation de la collecte de données pour la modélisation

Base de données de prédicteurs

Jeu de données

Incertitudes

Modélisation

Taille du jeu de données

Variabilité méthodologie

Entrainement sur plusieurs modèles

Apprentissage actif

Réseau de capteurs

Equipement

Nettoyage

Stockage

Identification des données à ajouter

Mesure en continue et faible coût

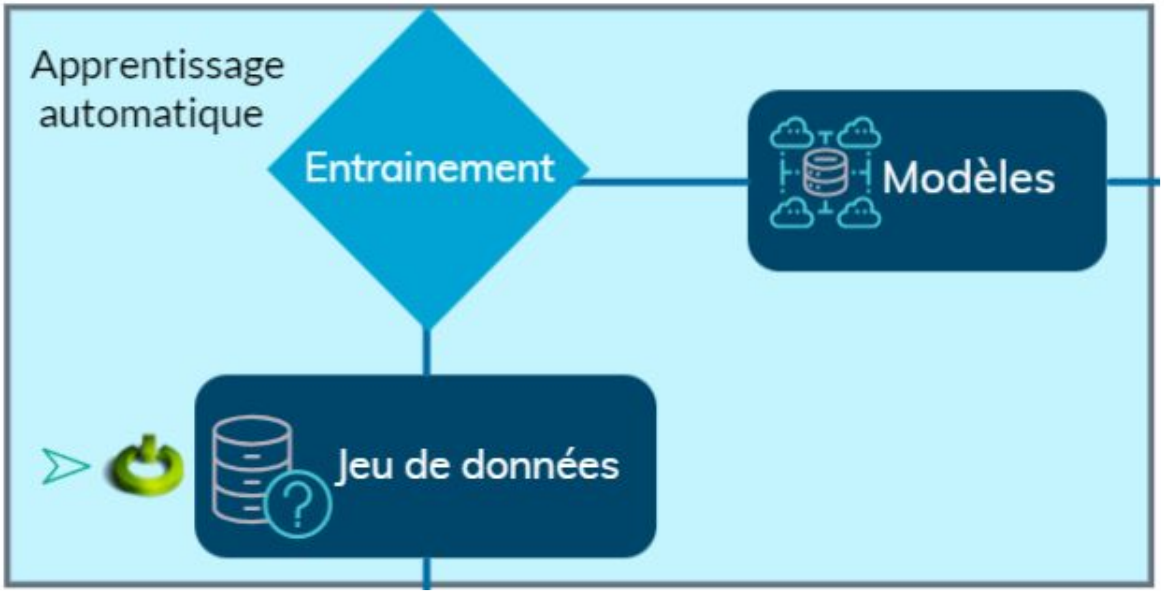
Choisir le meilleur modèle

Estimation

Augmentation du jeu de données

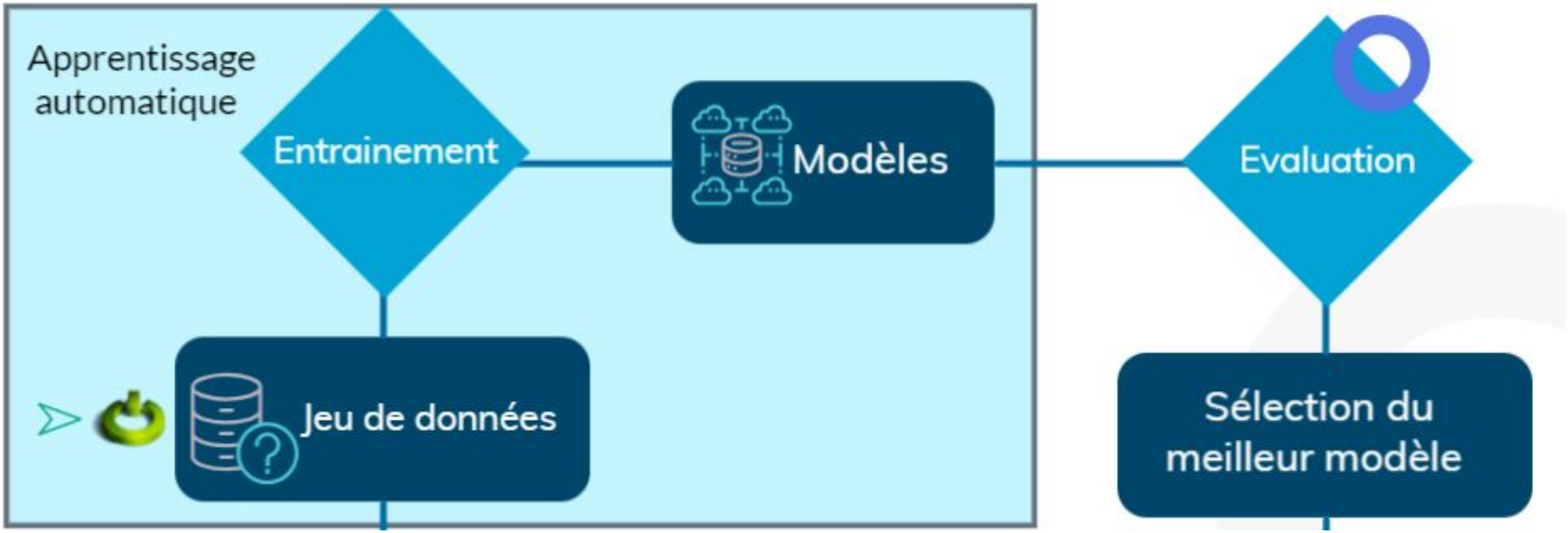
Diminution de l'incertitude

Optimisation des bases de données



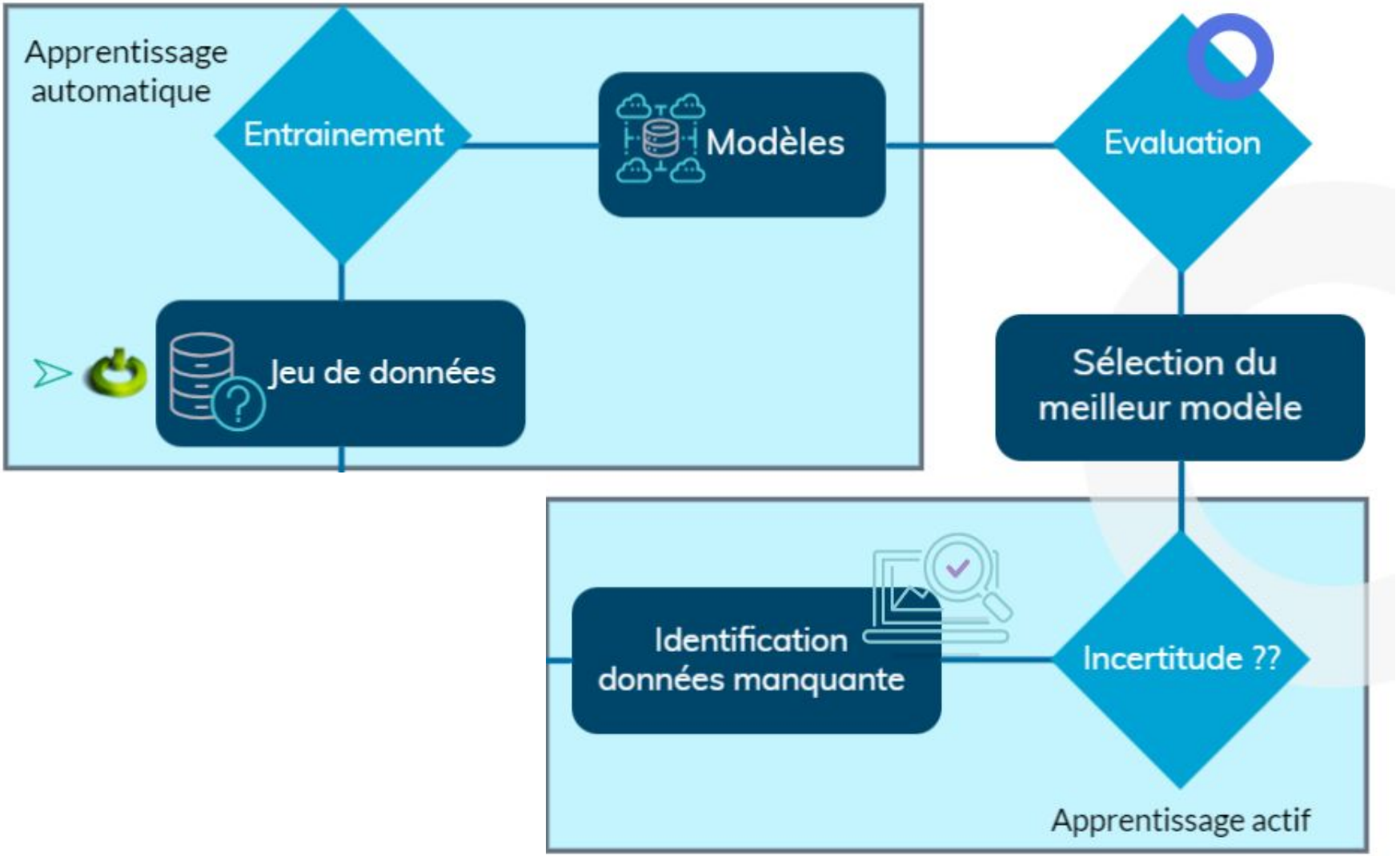
(Naloufi et al., 2021)

Optimisation des bases de données



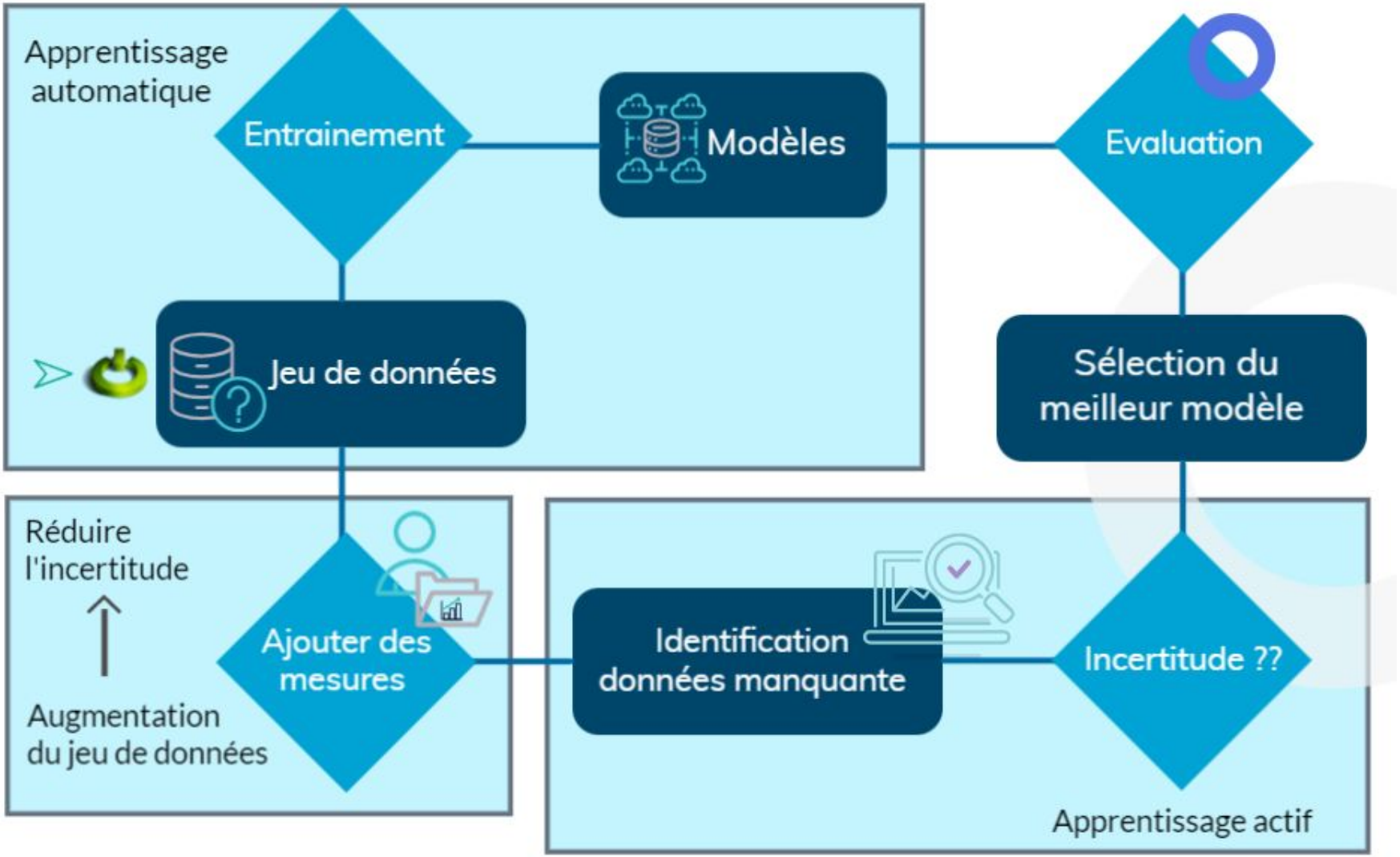
(Naloufi et al., 2021)

Optimisation des bases de données



(Naloufi et al., 2021)

Optimisation des bases de données



(Naloufi et al., 2021)

Optimisation des bases de données

Bases de données

- **Marne**
- **Seine**

Paramètres proxy

ID stations
Température
Conductivité
Turbidité
Matières en suspension (MES)
NH_4^+
Azote total (NTK)
Nombre de jours secs
Pluviométrie du jour
Pluviométrie de la veille
Débit (à Gournay/Marne ou à Austerlitz)

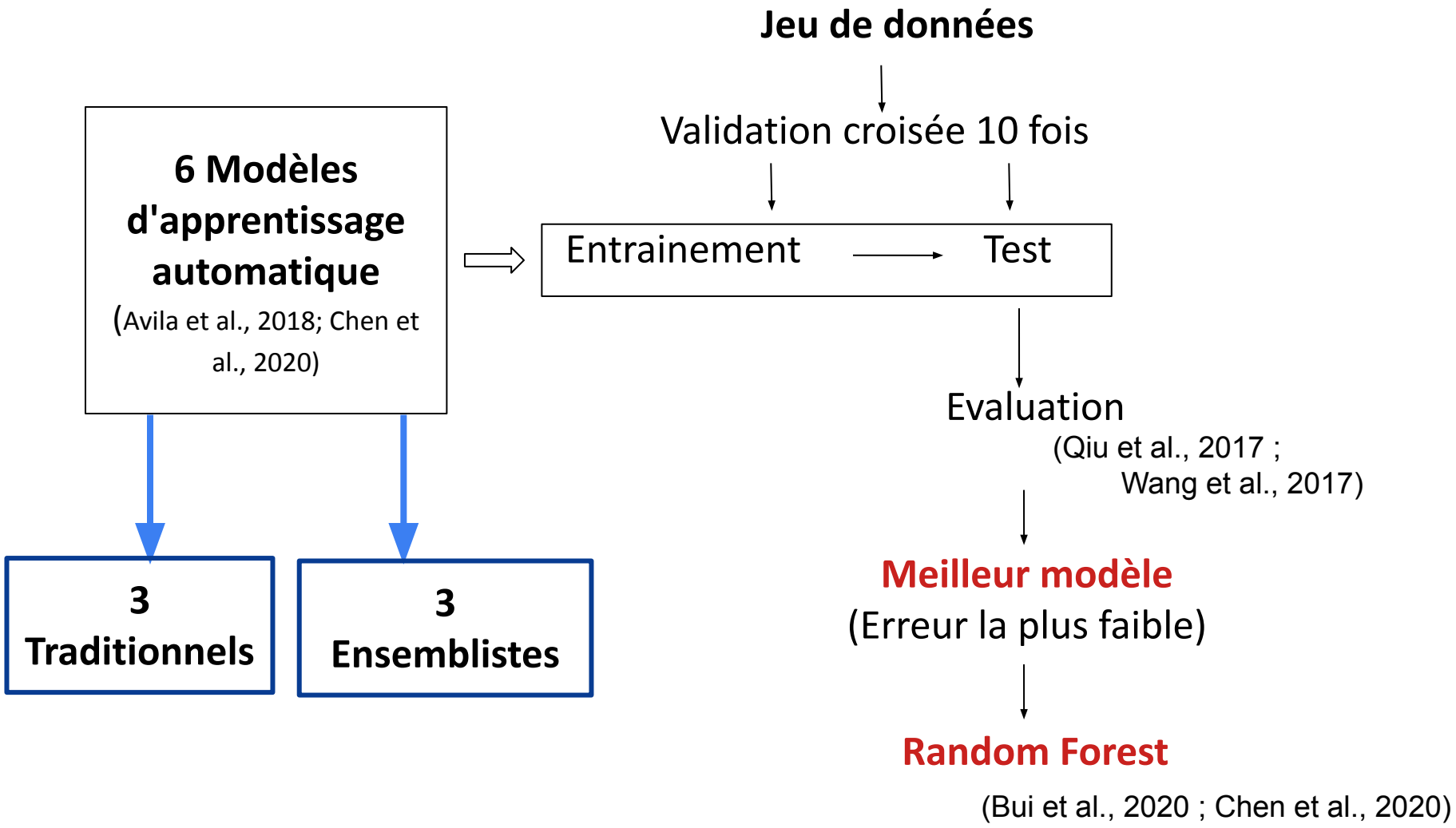
Concentration en *E. coli*

CAPGEO

PARIS

Marne Vive
Syndicat mixte

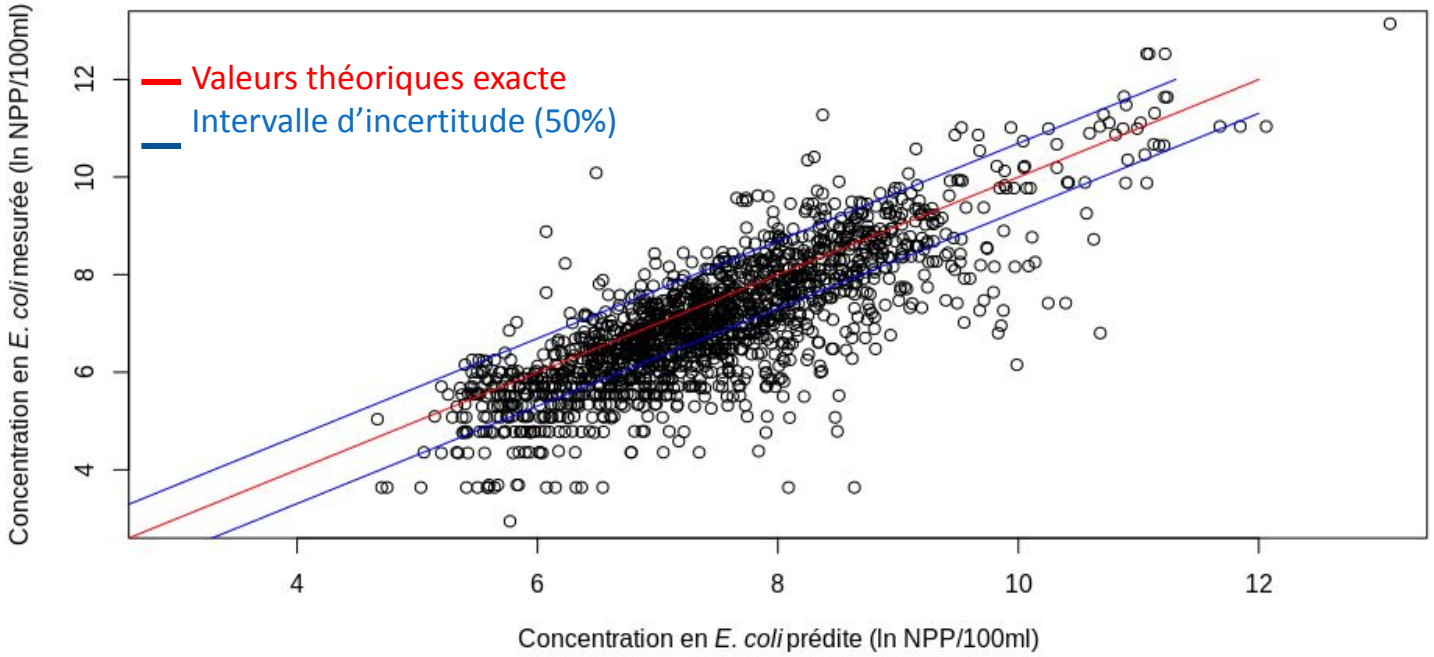
Optimisation des bases de données



Optimisation des bases de données

Analyse de la prédiction d'*E.coli* :

Ex. Marne



- Incertitude dans la prédiction (MAPE* ≥ 50%) :
  Prédiction inexacte : **53±4%** (Marne) et **64±3%** (Seine)

*Pourcentage d'erreur absolu moyen par observation (Yan et al., 2020)

Optimisation des bases de données

Hypothèse :

- L'apprentissage actif permet une amélioration des modèles de prédiction ?

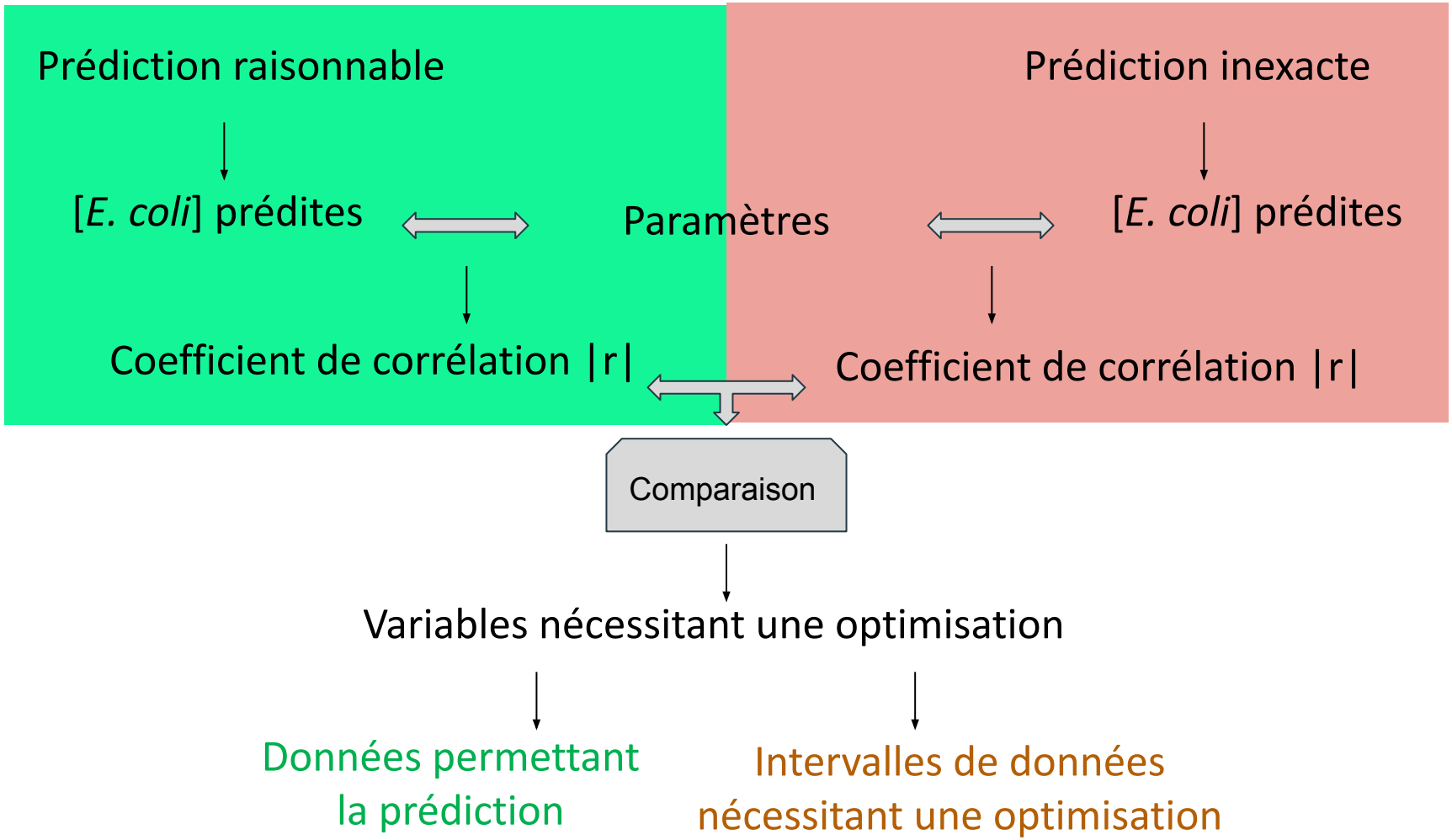
- **C'est quoi l'apprentissage actif ?**

❖ Sélectionner les données de manière à apprendre une bonne hypothèse avec moins d'entraînement (Qian et al., 2020).

➢ Stratégie : échantillonnage d'incertitude (Bouneffouf, 2016)

Optimisation des bases de données

Stratégie : Identification des paramètres à optimiser :



Optimisation des bases de données

(Naloufi et al., 2021)

Ex. Marne

Paramètres à optimiser

- Température
- Conductivité
- Pluviométrie de la veille
- Débit

Température (°C)

(r=-0.17)

Valeurs permettant une prédiction raisonnable

Valeurs entraînant une prédiction inexacte ou valeurs manquantes

Optimisation des bases de données

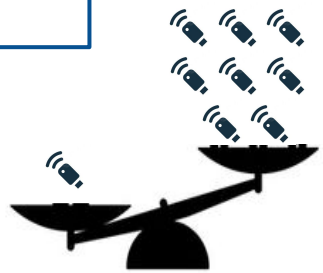
Forbath

Hypothèse :

- Le réseau de capteurs à faible coût permet de combler les manques des modèles ?

Marge d'erreur légèrement supérieure au équipements de haute précision

Réseau dense, en moyenne, est capable de fournir suffisamment d'informations pour les modèles (Wang et al., 2019).



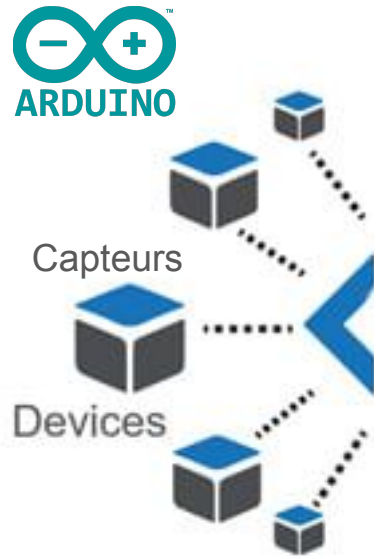
Couverture spatiale

❖ Surveillance de l'eau

➢ Acquisition de données en continue (Jiang et al., 2020)

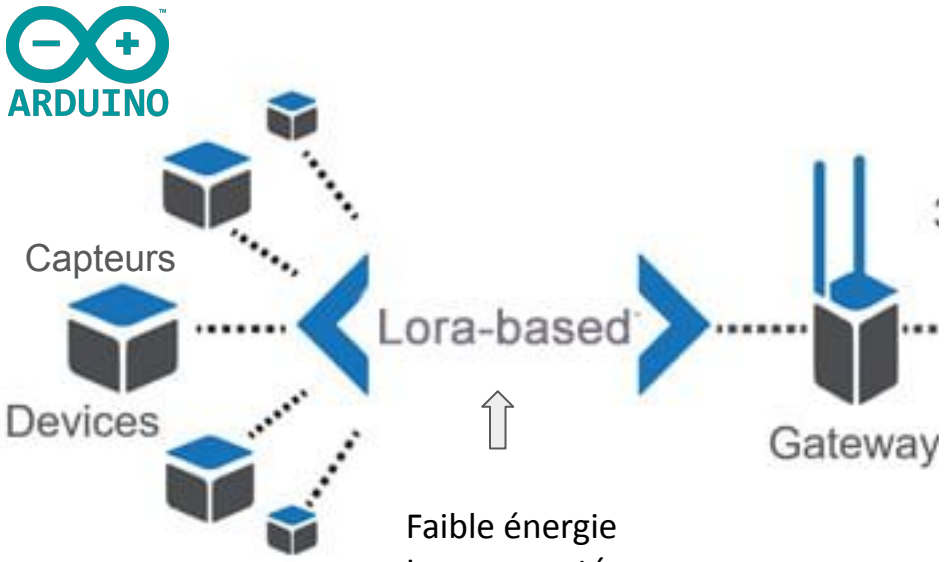
Optimisation des bases de données

Architecture du réseau



Optimisation des bases de données

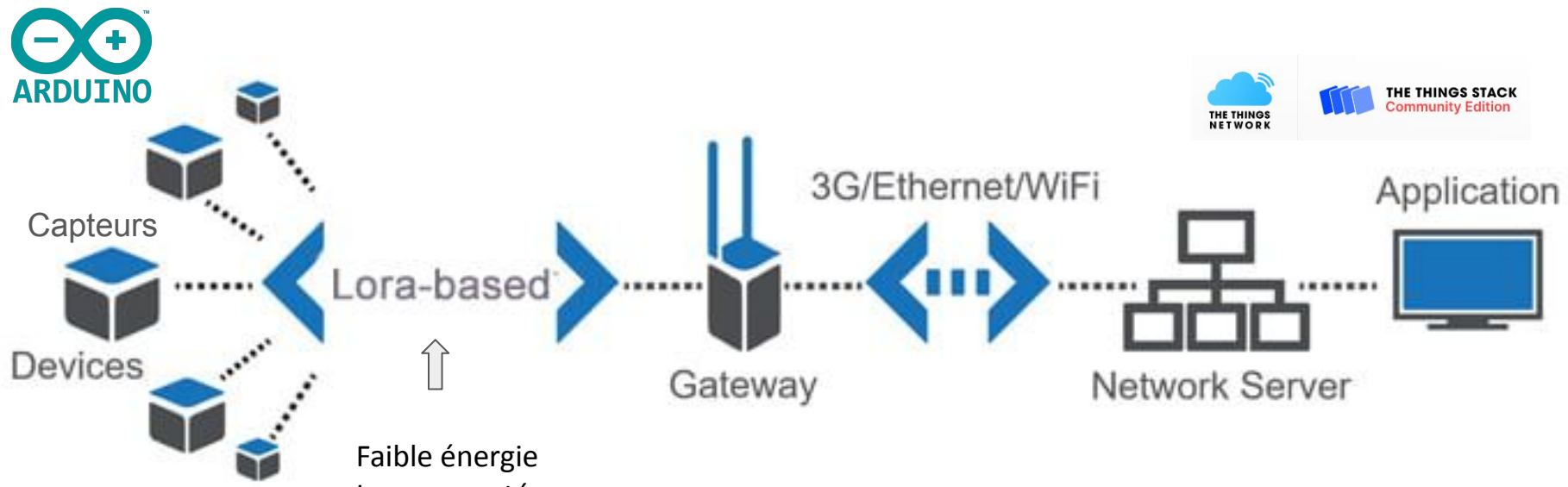
Architecture du réseau



Faible énergie
Longue portée
(Jiang et al.. 2020)

Optimisation des bases de données

Architecture du réseau



Faible énergie
Longue portée
(Jiang et al.. 2020)

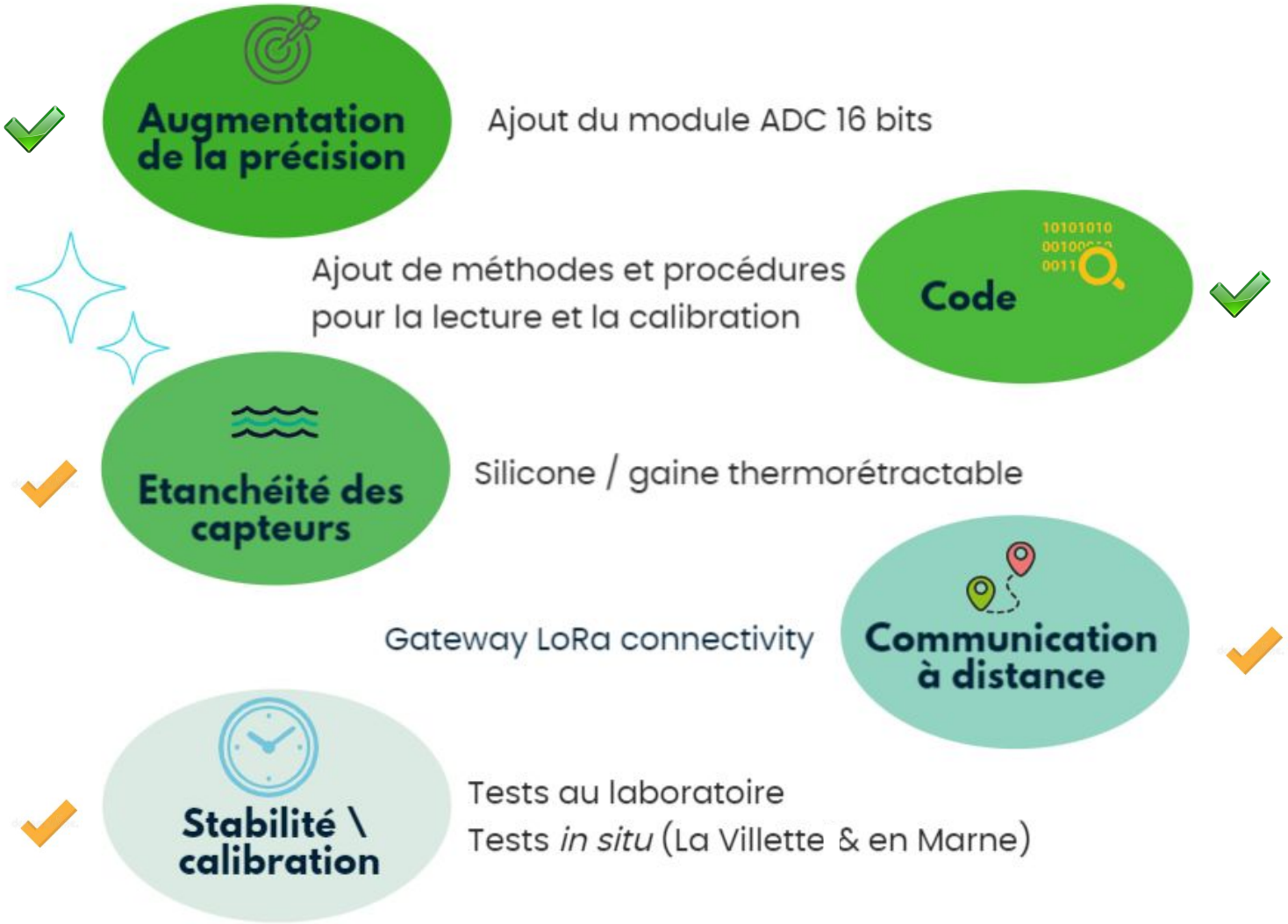
Optimisation des bases de données

Fait

En cours

Pas encore

Solutions mises en place



Optimisation de la collecte de données pour la modélisation

Jeu de données

Incertitudes

Bactéries

Modélisation

Taille du jeu de données

Variabilité méthodologie

Entrainement sur plusieurs modèles

Apprentissage actif

Réseau de capteurs

Equipement

Nettoyage

Stockage

Identification des données à ajouter

Mesure en continue et faible coût

Estimation

Choisir le meilleur modèle

Augmentation du jeu de données

Diminution de l'incertitude

Variabilité liée à la méthodologie

Bactéries indicatrices fécales (BIF)

Incertitudes de mesure

Variabilité méthodologie

- Temps d'incubation → Dépend de la concentration du site
- Equipements → Similaire
- Nettoyage → Rinçage
- Stockage → Jusqu'à 6 heures

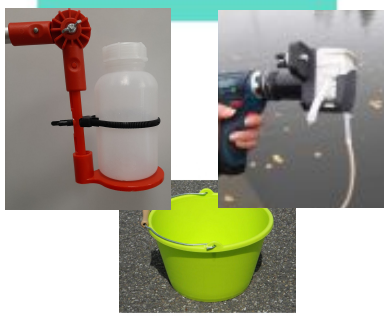
Variabilité liée à la méthodologie

Bactéries indicatrices fécales (BIF)

Incertitudes de mesure

Variabilité méthodologie

- Temps d'incubation → Dépend de la concentration du site
- Equipements → Similaire
- Nettoyage → Rinçage
- Stockage → Jusqu'à 6 heures



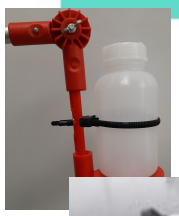
Variabilité liée à la méthodologie

Bactéries indicatrices fécales (BIF)

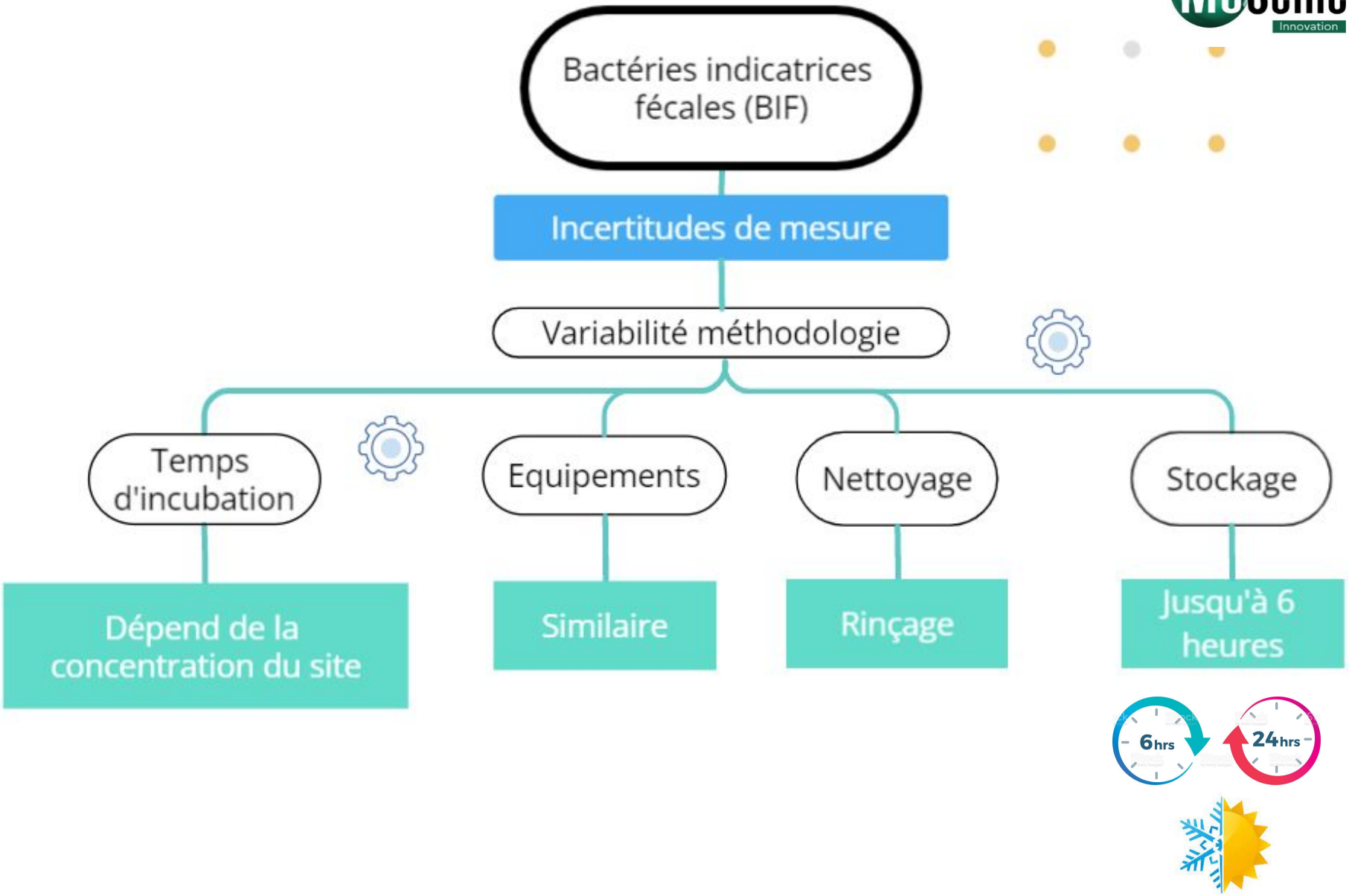
Incertitudes de mesure

Variabilité méthodologie

- Temps d'incubation → Dépend de la concentration du site
- Equipements → Similaire
- Nettoyage → Rinçage
- Stockage → Jusqu'à 6 heures



Variabilité liée à la méthodologie



Conclusion et perspectives

Conclusion et perspectives

- Mise en place et installation du réseau de capteurs

- Variabilité lié à la méthodologie
  - Préleveur automatique 
  - Groupe de travail

- Prélèvements et analyses de l'eau de surface et des rejets

UPEC
Connaissance - Action
UNIVERSITÉ PARIS-EST CRÉTEIL VAL DE MARNE

PARIS EST SUP ÉCOLE DOCTORALE
Sciences, Ingénierie et Environnement

PARIS

École des Ponts ParisTech

leesu
laboratoire eau environnement systemes urbains

LiSSiX
Laboratoire Images, Signaux et Systèmes Intelligents

Merci de votre attention

PiREN Seine

OPUR

Marne Vive
Syndicat mixte

eau de Paris
service public de l'eau

MeSeine
Innovation

eurostars™

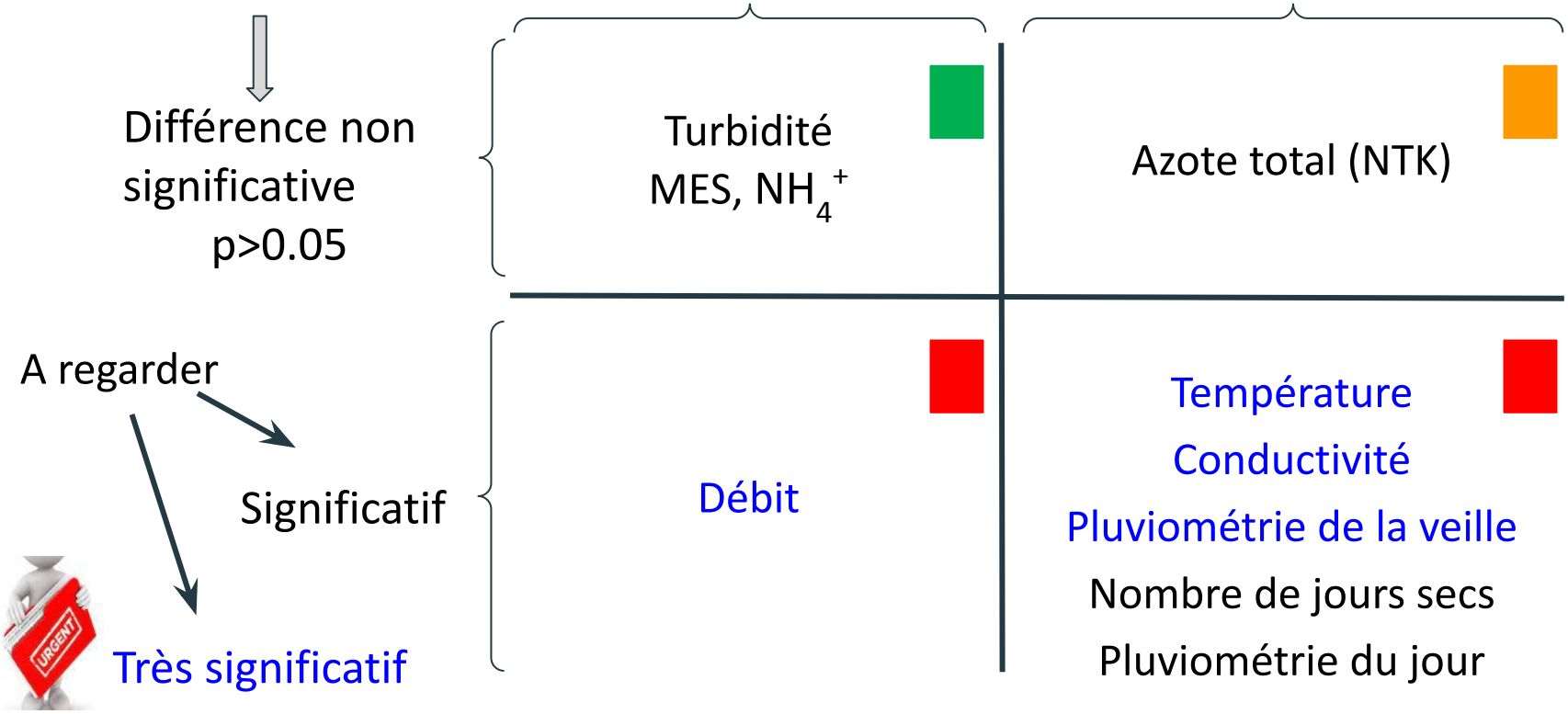
Naloufi, M.; Lucas, F.S.; Souihi, S.; Servais, P.; Janne, A.; Wanderley Matos De Abreu, T. Evaluating the Performance of Machine Learning Approaches to Predict the Microbial Quality of Surface Waters and to Optimize the Sampling Effort. *Water* **2021**, *13*, 2457. https://doi.org/10.3390/w13182457

Optimisation des bases de données

Corrélation avec la concentrations en *E. coli*

Comparaison des corrélations entre les prédictions raisonnables et inexactes

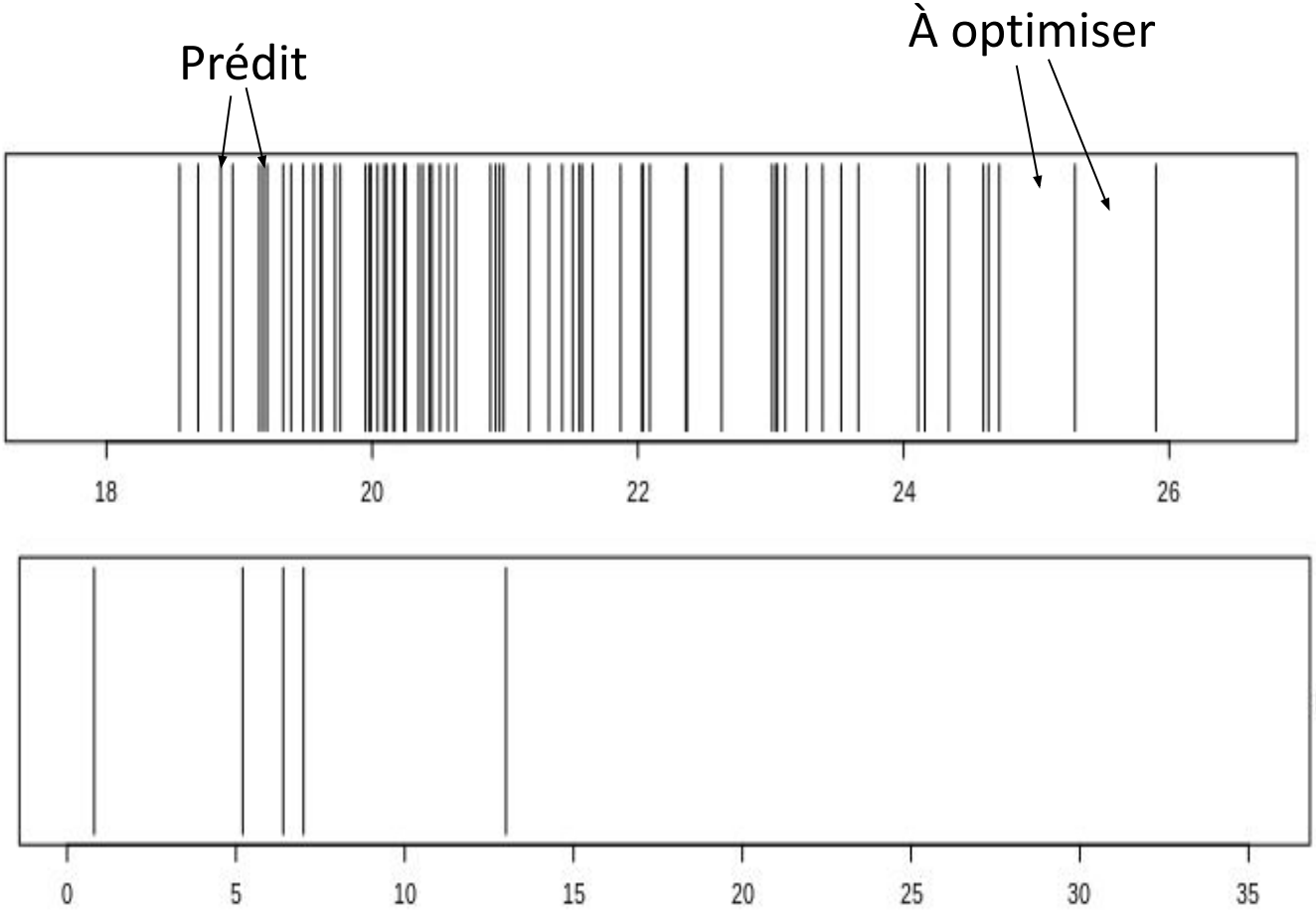
Ex. Marne



Optimisation des bases de données

Paramètres clés pour l'optimisation de la collecte

Ex. Marne



Température (°C)

(r=-0.17)

Pluviométrie de la veille (mm)

(r=0.17)

Valeurs permettant une prédiction raisonnable

Valeurs entraînant une prédiction inexacte ou valeurs manquantes

Variabilité liée à la méthodologie

Variabilité liée à la méthodologie

Impact du temps d'incubation sur la lecture

- Incubation pendant 36 à 72 heures (Norme ISO 9308-3 et ISO 7899-1)
- 333 mesures (La Villette, St Thibault des Vignes et Créteil)

Lecture

après 24h

après 48h

après 72h

Analyse statistique
(Wilcoxon et test t apparié)

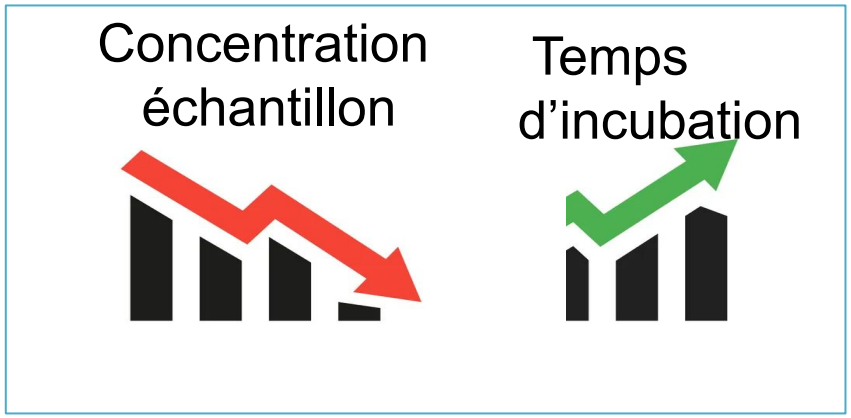
Concentration échantillon

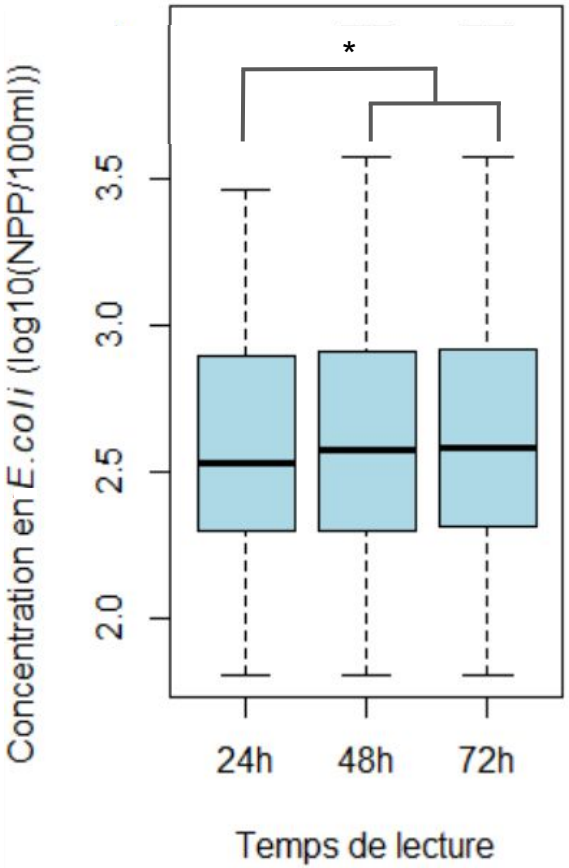
Temps d'incubation

Variabilité liée à la méthodologie

Impact du temps d'incubation sur la lecture

Analyse statistique
(Wilcoxon et test t apparié)





N= 144
*: p<0.05

Echantillons :
[*E. coli*] = 701 ± 881 NPP/100ml

Variabilité liée à la méthodologie

Lac de Créteil

Comparaison des équipements

Equipement

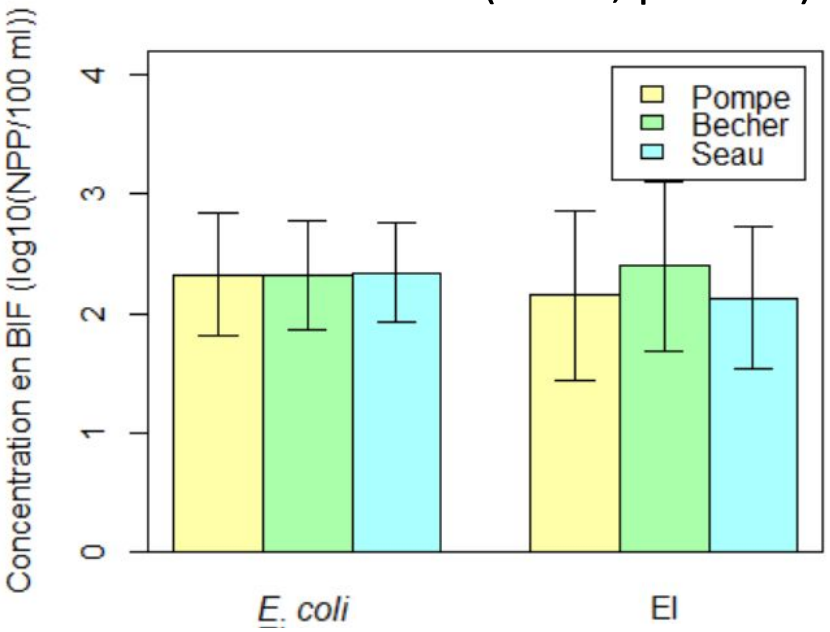
Becher et perche

Pompe

Seau

Analyse statistique :

- *E. coli* et EI :
  Test de Wilcoxon apparié
  (N=30, p> 0.05)



⇒ Les 3 systèmes sont similaires

Variabilité liée à la méthodologie

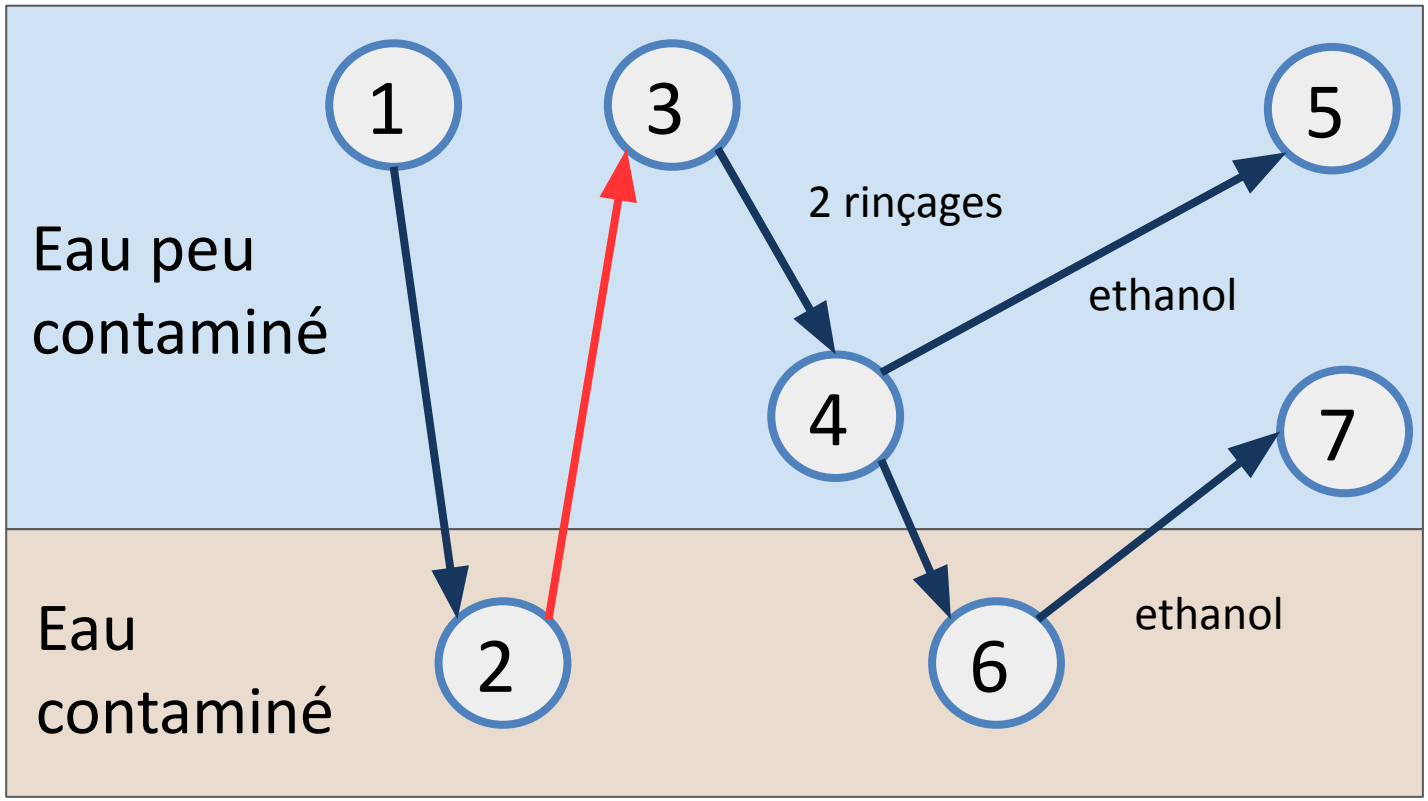
Nettoyage : Rincer, stériliser ou non ?

Lac de Créteil

Nettoyage

Becher

Tuyau



1 Référence

Variabilité liée à la méthodologie

Nettoyage : Rincer, stériliser ou non ?

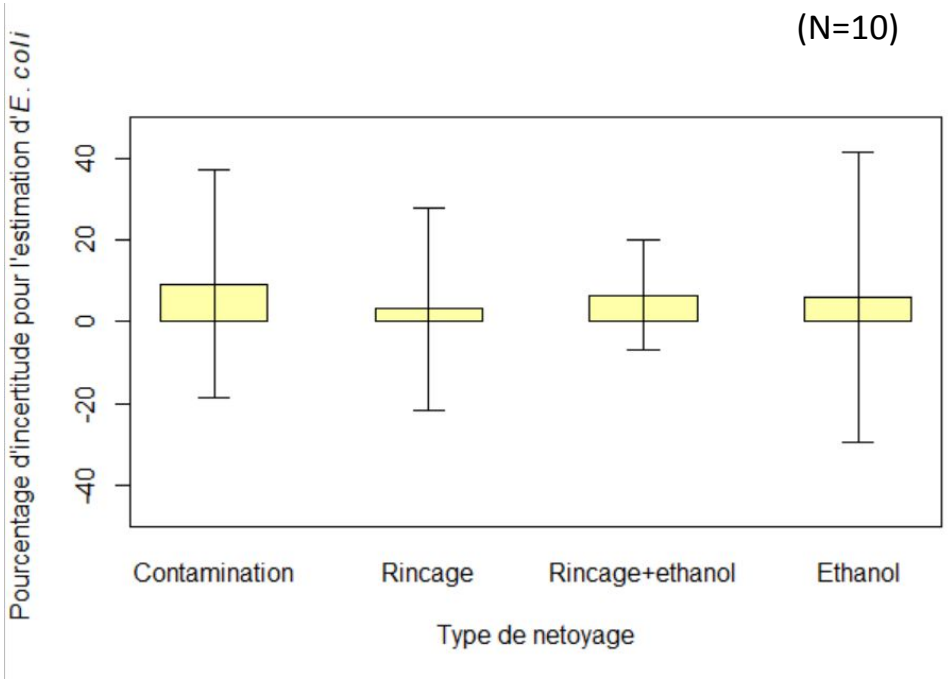
Prélèvement ponctuelle :

Estimation de l'incertitude :
(erreur relative d'échantillonnage)

$$unc = \frac{X2-X1}{X1} \quad Eq(1)$$

X1 : Valeur de référence (échantillon 1)

(Esbensen et Wagner, 2014;
Harmel et al., 2016)

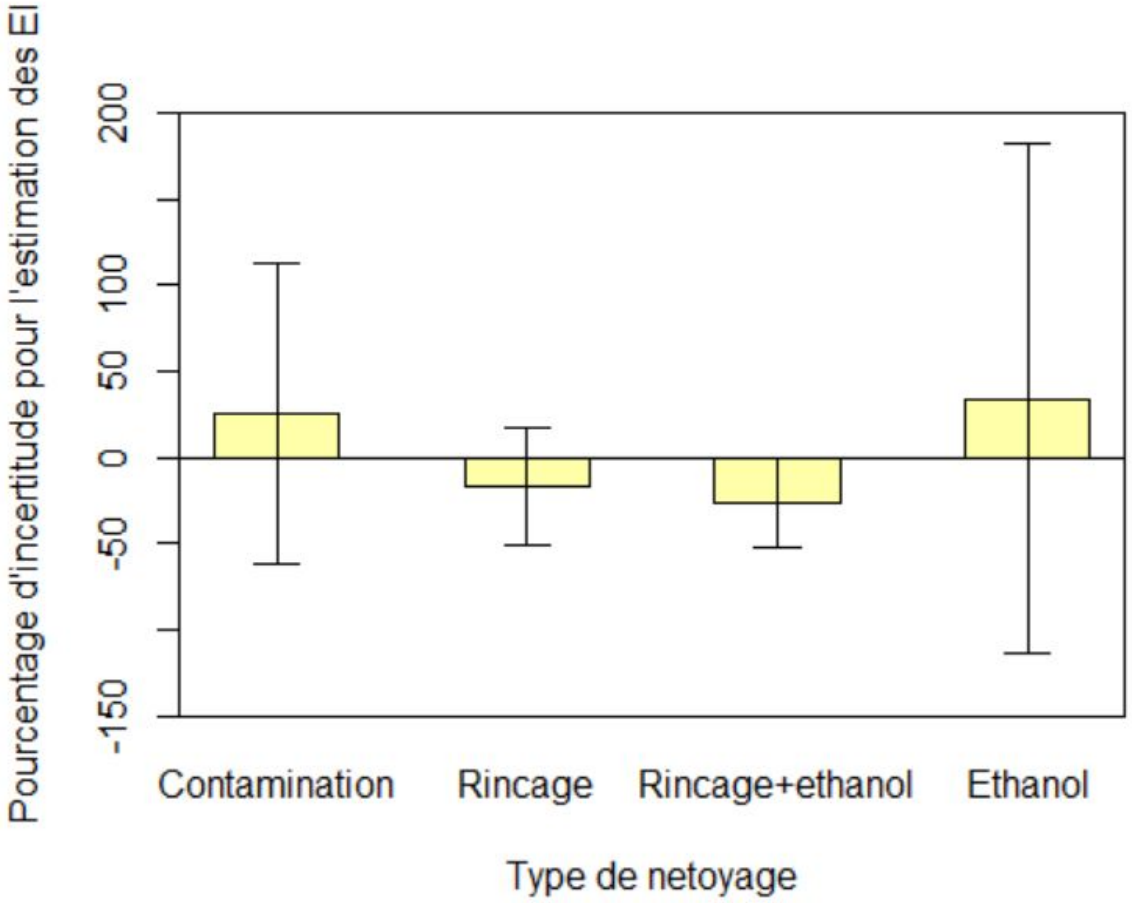


Variabilité liée à la méthodologie

Nettoyage : Rincer, stériliser ou non ?

Prélèvement ponctuelle :

(N=10)



Variabilité liée à la méthodologie

Transport et stockage : Température et temps ?

➢ À T°C$_{amb}$: augmentation à partir de 4h
(McCarthy el al., 2008)

➢ A 24h : diminution de la concentration
(Harmel et al., 2016 ; McCarthy el al., 2008)

Effet Transport, Stockage ?

Température
(5°C, ambiante)

Froid durant le transport?

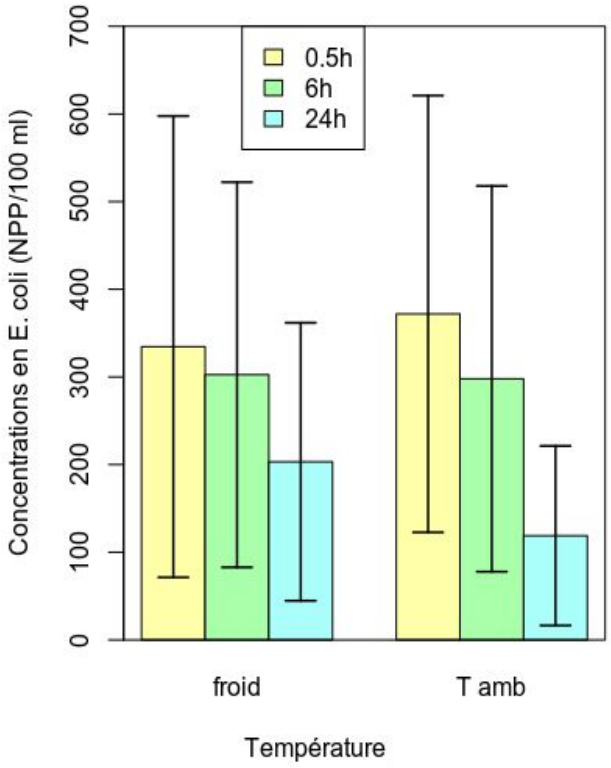
Temps de stockage?

Prélèvement 1: OUI
glaciaire
➡ ***Stockage au frigo*** ➡ 0.5h, 6h, 24h

Prélèvement 2: NON
T°C ambiante
0.5h, 6h ➡ ***Stockage au frigo*** ➡ 24h

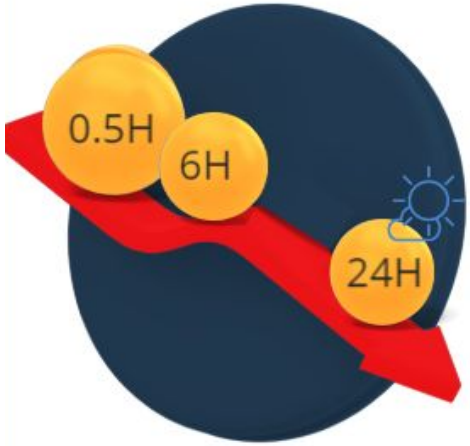
Variabilité liée à la méthodologie

Transport et stockage : Température et temps ?



Analyse statistique :

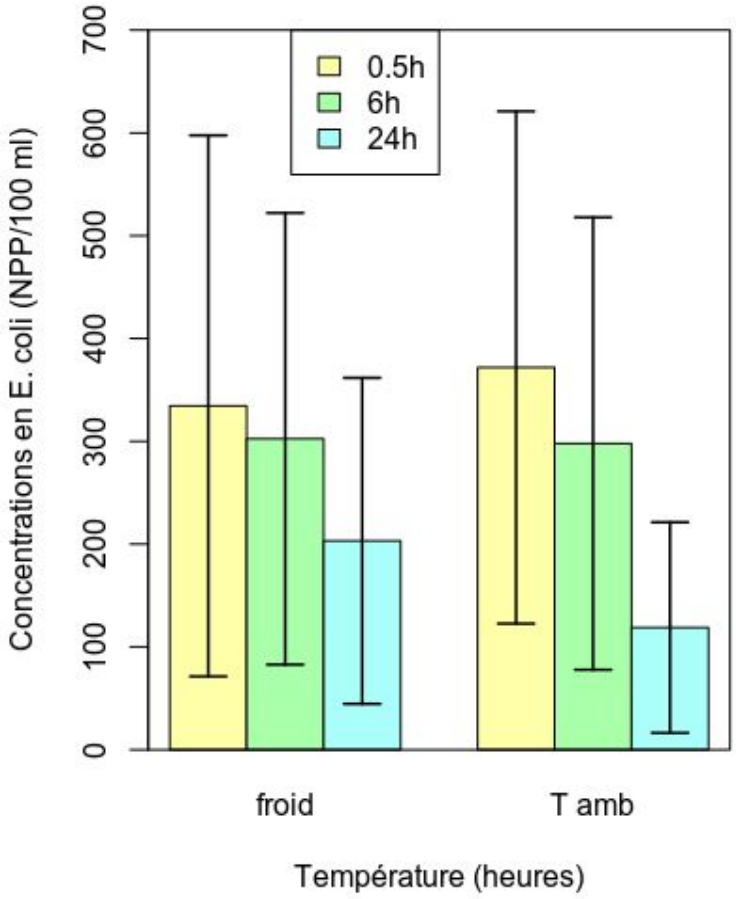
Test t et Wilcoxon apparié
(correction Bonferroni)
(N=12, p> 0.05)

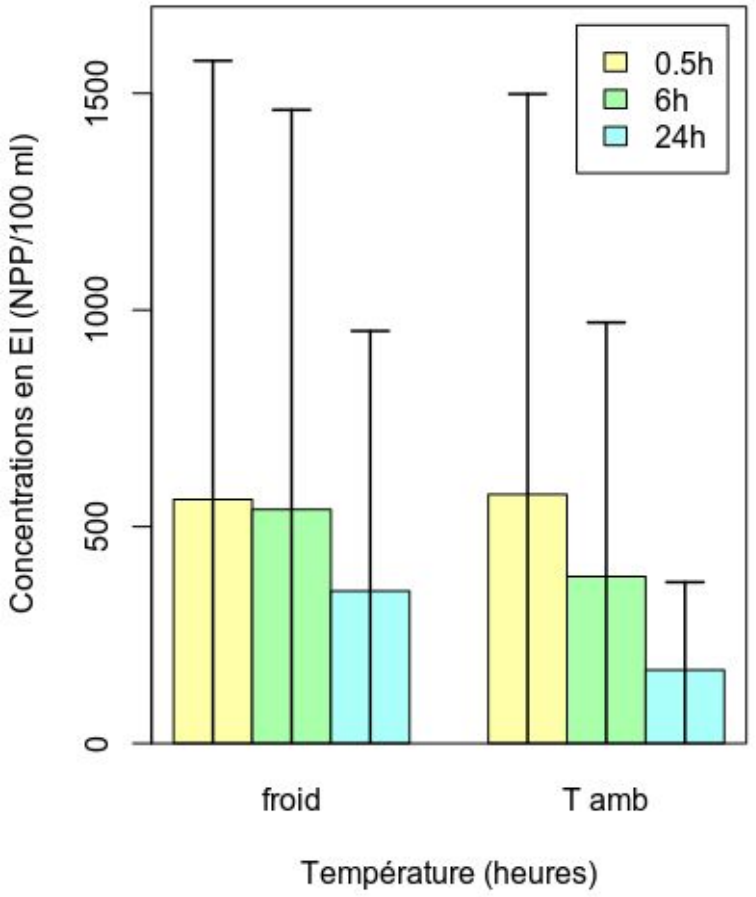


⇒ Faible décroissance jusqu'à 6h et au delà plus forte (augmente avec la température)

Variabilité liée à la méthodologie

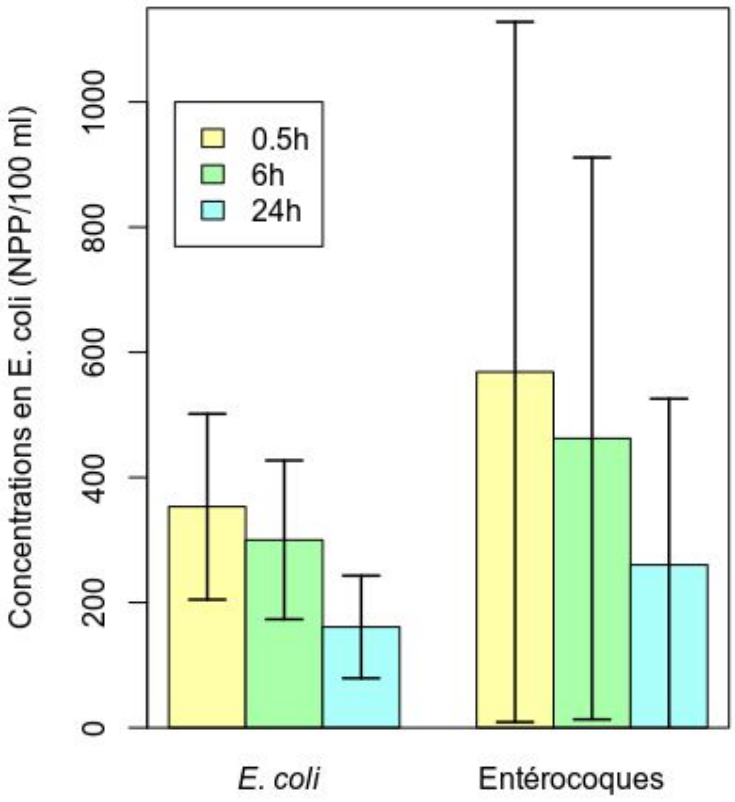
Transport et stockage : Température et temps ?





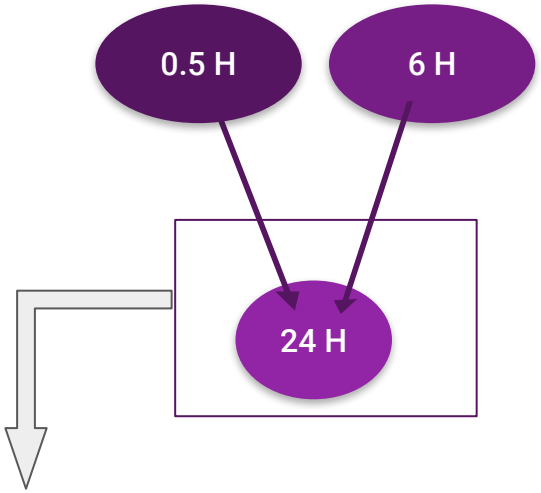
Variabilité liée à la méthodologie

Transport et stockage : Température et temps ?



Analyse statistique :

- *E. coli* et EI :
  Test t et Wilcoxon apparié
  (correction Bonferroni)
  (N=36 ; p=0.007; p=0.006)



Décroissance significative
(N=24; p<0.05)

Variabilité liée à la méthodologie

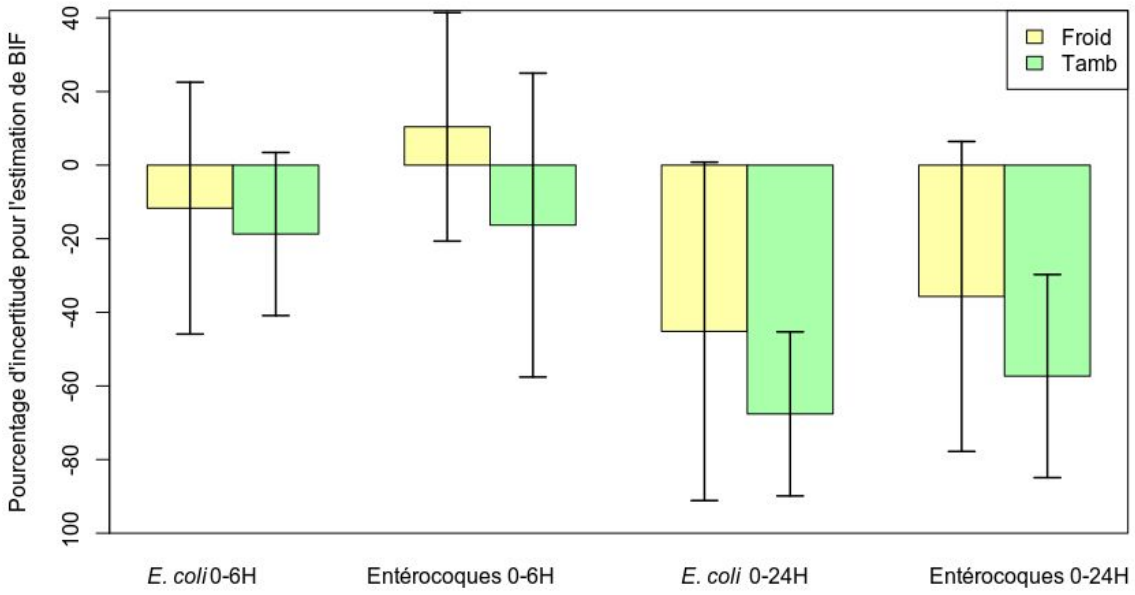
Transport et stockage : Température et temps ?

Estimation de l'incertitude :
(erreur relative d'échantillonnage)

$$unc = \frac{X2-X1}{X1}$$

X1 : Valeur de référence (échantillon 1)

(Esbensen et Wagner, 2014; Harmel et al., 2016)



⇒ Conservation possible jusqu'à 6h et au delà plus forte décroissance (qui augmente avec la température)

Réseau de capteurs

Kit KnowFlow (coût)

Boite étanche : 10,13€

Batterie externe (20000mAh) : 33,68 €

Conductivité

192.47€

Oxygène dissous

147.85€

pH

Température

8.18€

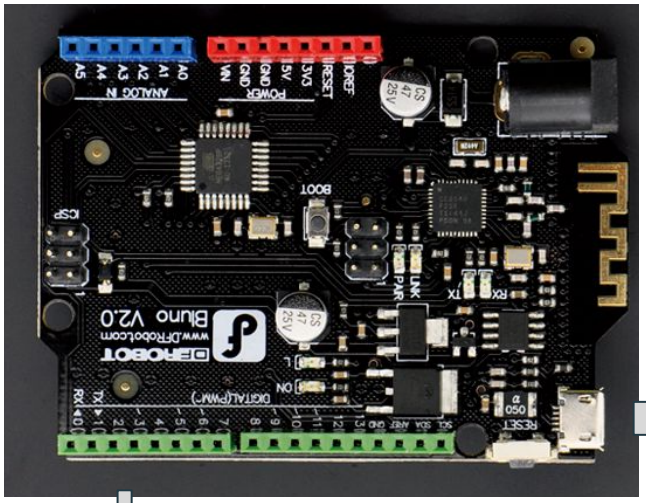
Turbidité

8.66€

Surveillance de l'eau (Arduino)

=> kit : 400.97€

Structure des capteurs



Etalonnage des capteurs

	Température (°C)	pH	Conductivité (µS\cm)	Turbidité (NTU)	Oxygène dissous	
					(µg\L)	(%)
N	362	30	44	77	20	
Etalons	5 à 30	4 à 10	220 à 1418	0 à 800	0 à 100%	
linéarité	**R^2>0.99**					
Répétabilité (écart-type)	0.004	0.02	21.14	3.66	142.46	1.74
Précision constructeur	±0.5	±0.1	±1000	±3.6	±400	±4

✔ Etalonnage

⇒ Meilleur précision : comparaison aux précisions du constructeur

Stabilité temporelle des capteurs

Lac de Créteil
Marne

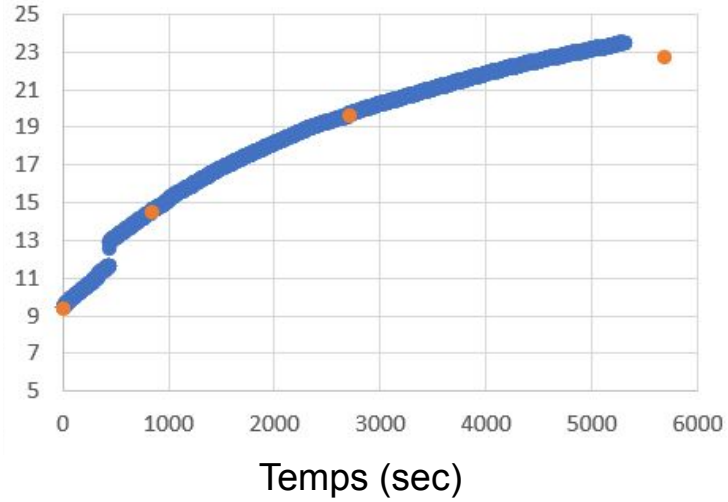
→ Mesure toutes les 2 secondes pendant 3h à 6h

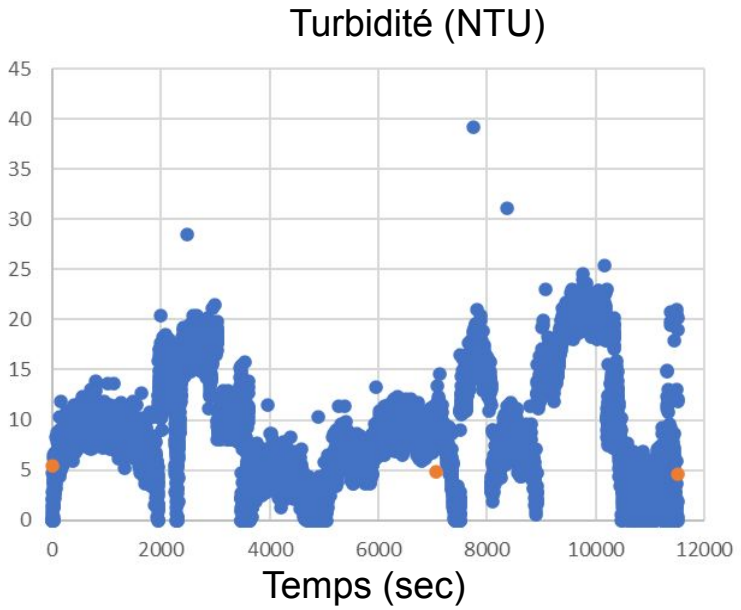
	Température (°C)	**pH**	**Conductivité (µS\cm)**	**Turbidité (NTU)**	**Oxygène dissous**	
					(µg\L)	(%)
Ecart-type moyen	1.82	0.095	43.90	**4.68**	266.5	2.33
Précision constructeur	±0.5	±0.1	±1000	±3.6	±400	±4

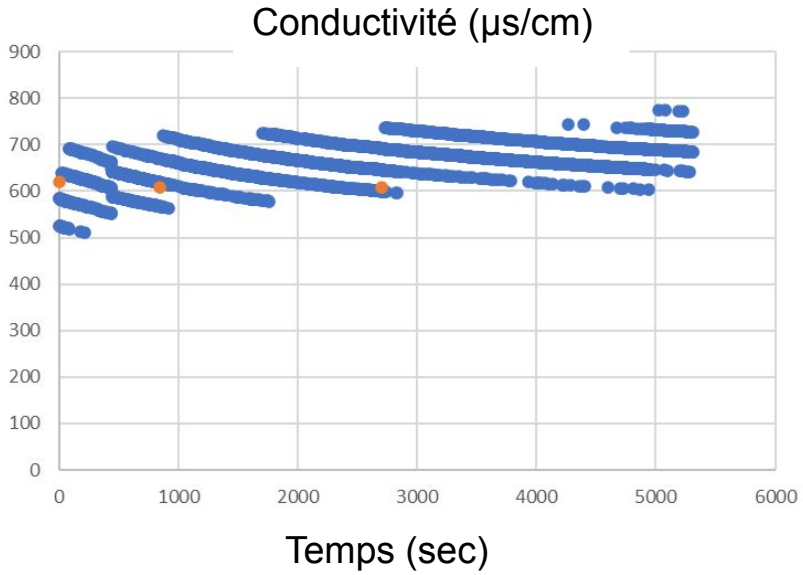
⇒ Stabilité temporelle relativement satisfaisante

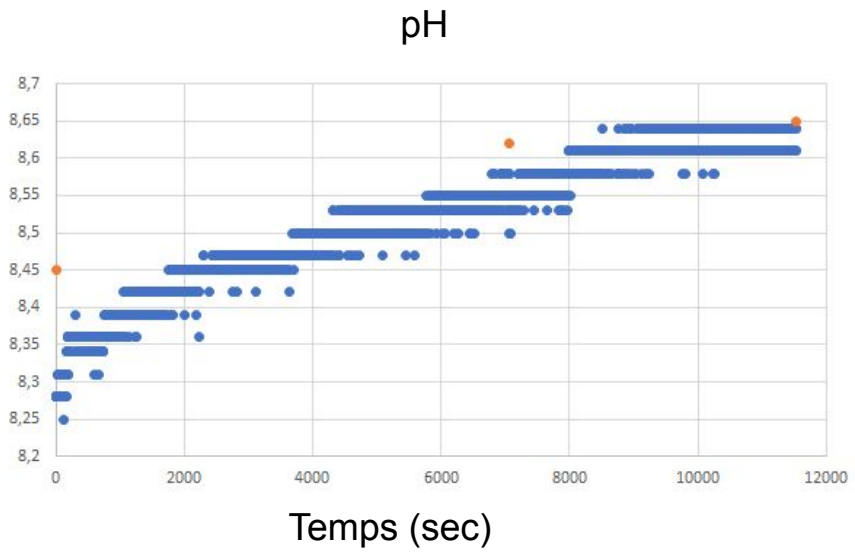
Stabilité temporelle des capteurs

Sonde capteur
Sondes (Eutech) et turbidimètre (HACH)









ML

1- Optimisation de la collecte de donnée

Les modèles d'apprentissage automatique :

Traditionnel		Ensembliste	
KNN	K-nearest neighbors	Bagging	Bootstrap aggregating
SVM	Support vector machines	RF	Random forest
DT	Decision tree	AdaBoost	Adaptive boosting

(Hastie, 2009)

1- Optimisation de la collecte de donnée

Evaluation des modèles :

- Erreur quadratique moyenne (RMSE) : $$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(yi - y'i)^2} \quad (1)$$

  (Qiu et al., 2017)

- Erreur absolue moyenne (MAE) : $$MAE = \frac{1}{N}\sum_{i=1}^{N}|yi - y'i| \quad (2)$$

  (Bui et al., 2020)

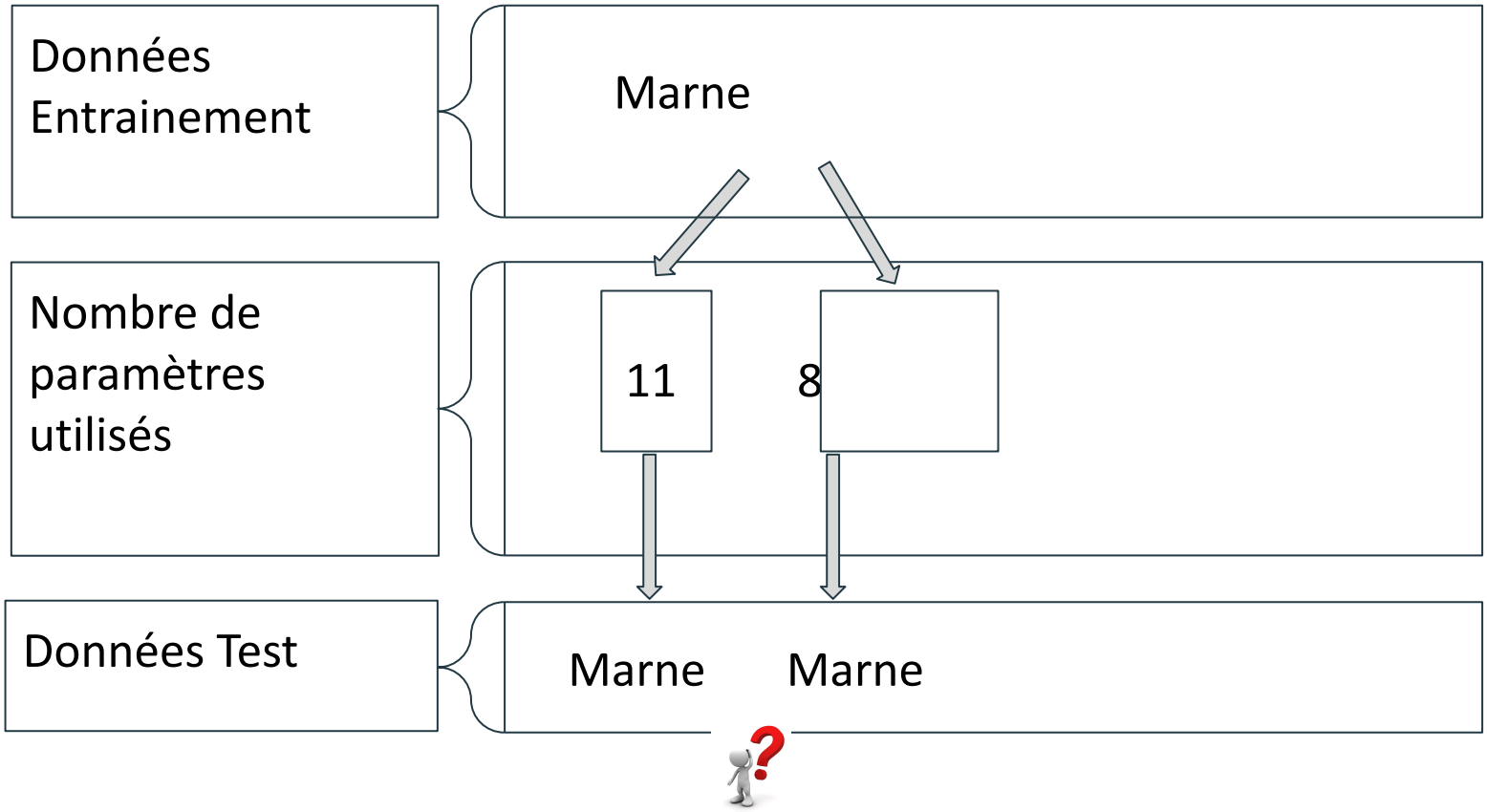
- Rapport performance/déviation (RPD) : $$RPD = \frac{SD}{RMSE} \quad (3)$$

yi : Mesurée
y'i : prédite

- **RPD < 1,4 ⇒ Modèle non fiable**
- **1,4 < RPD < 2 ⇒ Modèle modérément précis**
- **RPD > 2 ⇒ Haut niveau de capacité prédictive**

1- Optimisation de la collecte de donnée

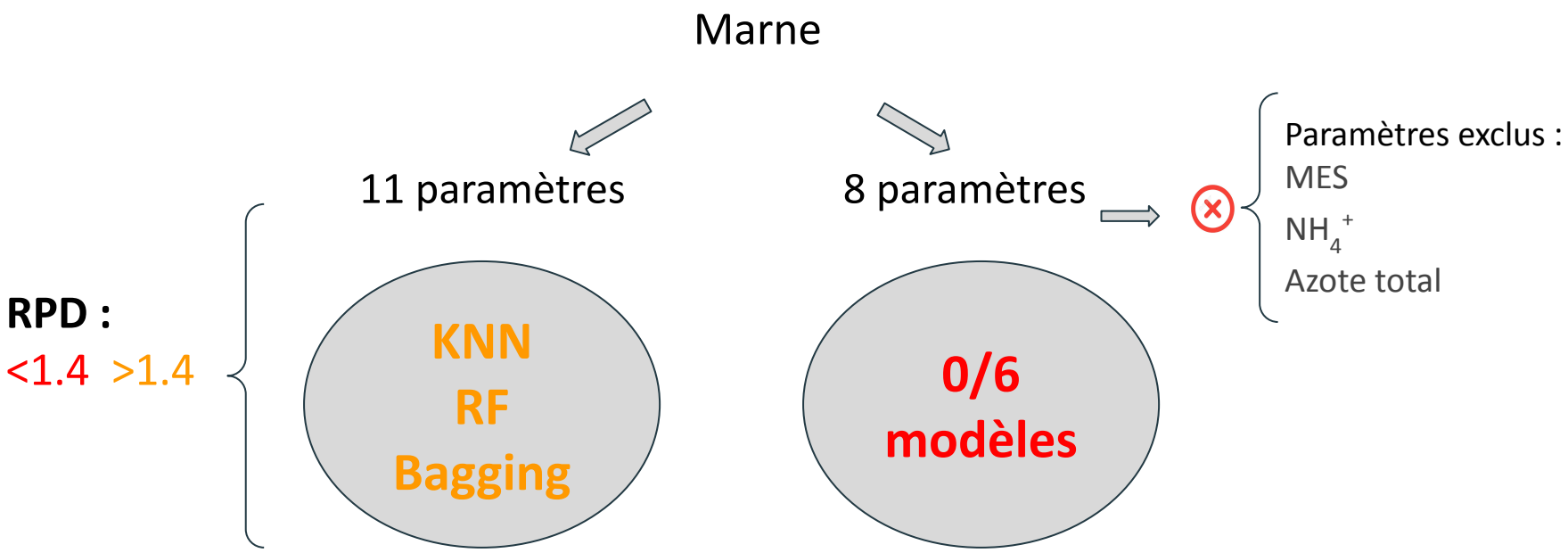
Stratégie pour l'entraînement et le test :



Quel est l'effet du nombre de paramètres ?

1- Optimisation de la collecte de donnée

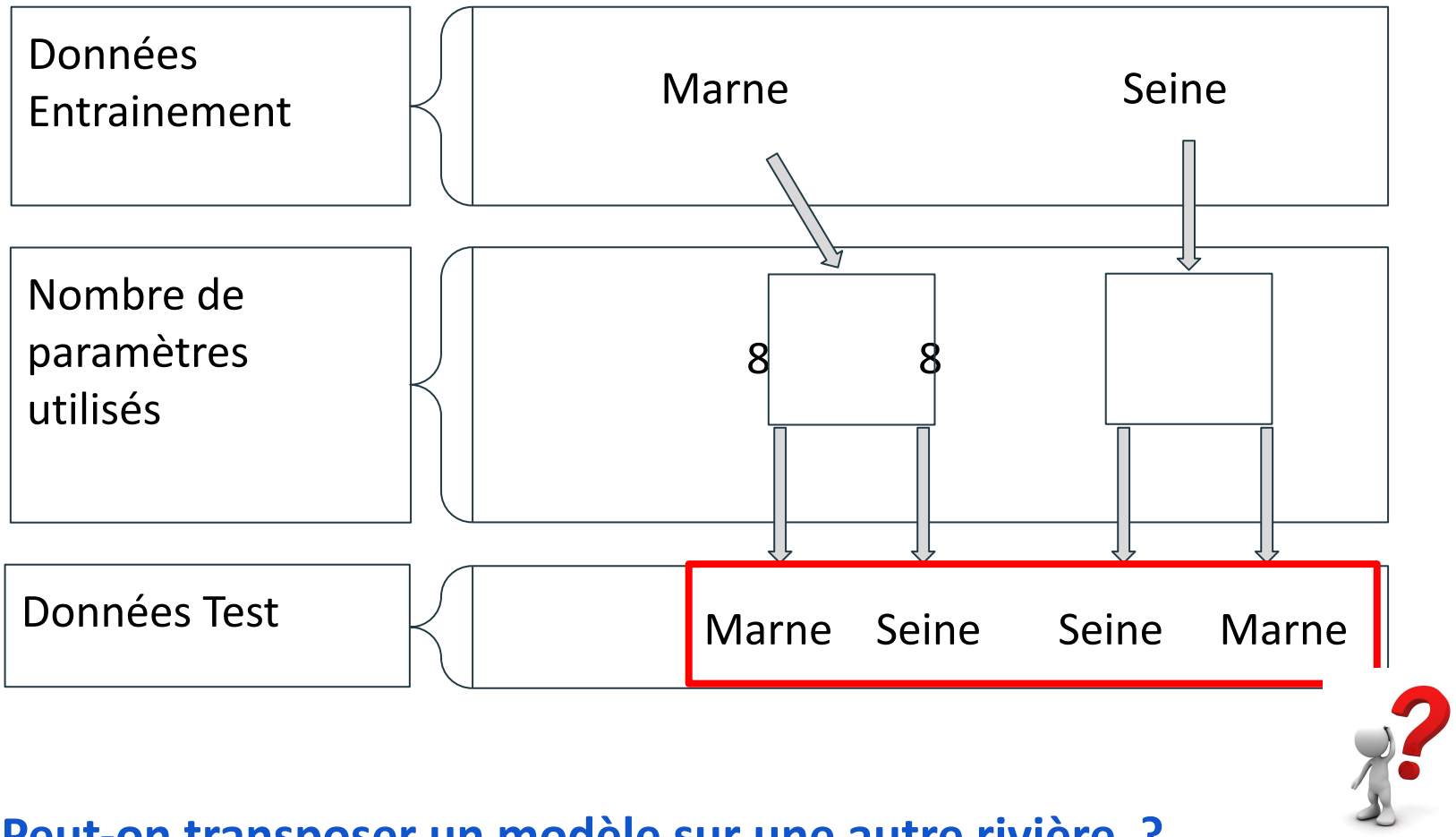
Quel est l'effet du nombre de paramètres ?



- Enlever trois paramètres lors de l'entraînement
  ⇒ Diminution de la performance des modèles (Chen et al., 2020)
  ⇒ Paramètres pertinents?

1- Optimisation de la collecte de donnée

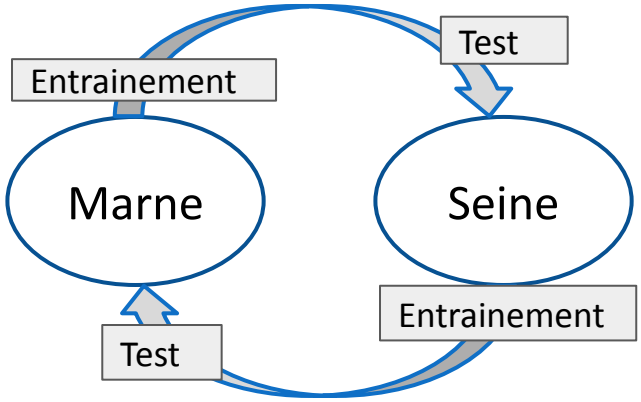
Stratégie pour l'entraînement et le test :



Peut-on transposer un modèle sur une autre rivière ?

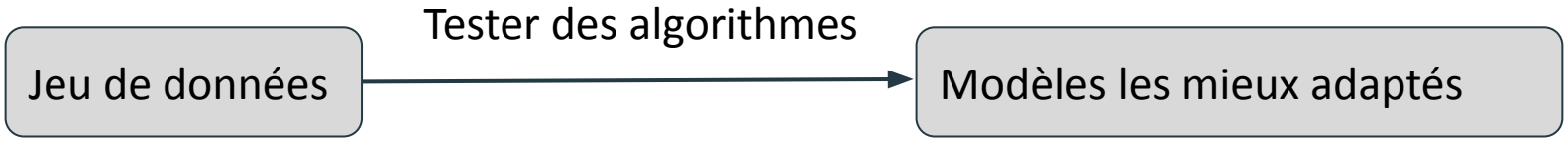
1- Optimisation de la collecte de donnée

Peut-on transposer un modèle sur une autre rivière ?



⇒ **RPD** : Aucun modèle fiable

- Performance : entraînement avec les données de la même rivière
  - Performance diffère d'un site à l'autre (Mälzer et al., 2016)



1- Optimisation de la collecte de donnée

Méthodologie : Évaluer la prédiction

Meilleur modèle → MAPE → Pourcentage d'erreur absolu moyen par observation :

(Yan et al., 2020)

$$MAPE = \frac{|yi - y'i|}{yi} * 100 \ (4)$$

yi : Mesurée
y'i : prédite

Qualité d'ajustement :

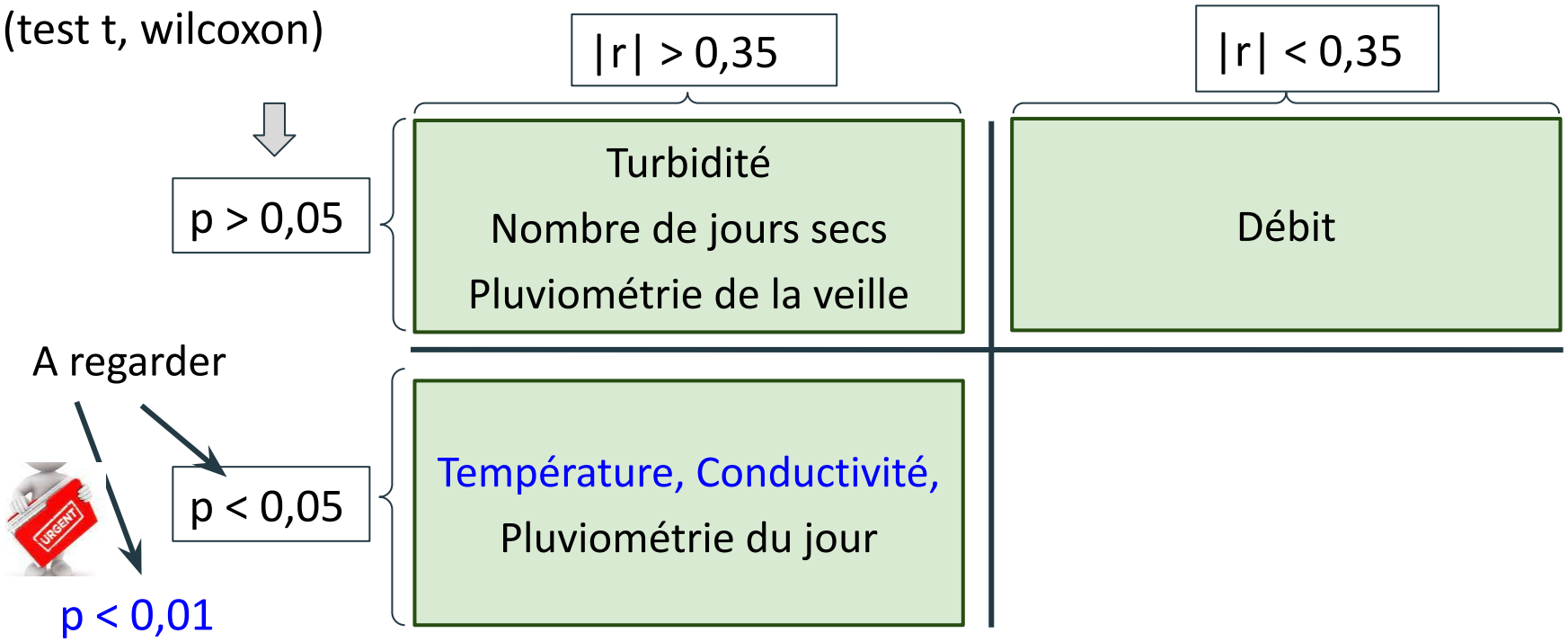
- MAPE < 50% ⇒ "raisonnable"
- MAPE > 50% ⇒ "inexact"

(Lu et Ma, 2020)

Corrélation avec la concentrations en E. coli

Seine

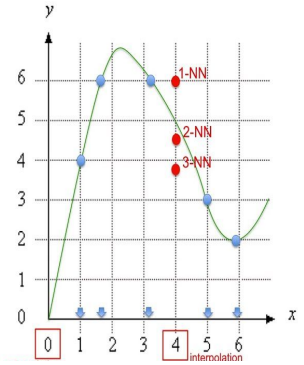
Entre raisonnable et inexacte



Modéles

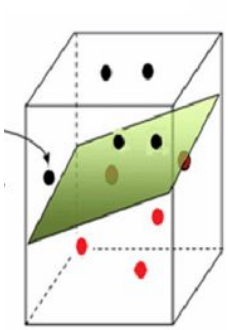
Les modèles d'apprentissage automatique (Hastie, 2009)

KNN



Trouver les k observations les plus proches dans l'espace d'entrée

SVM

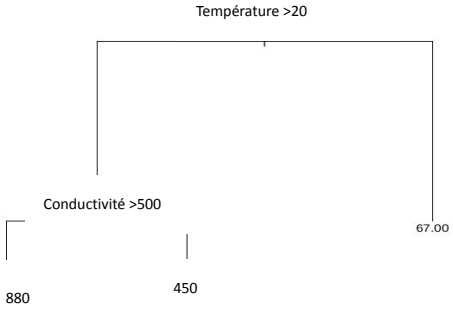


L'espace des variables

(González et al., 2018)

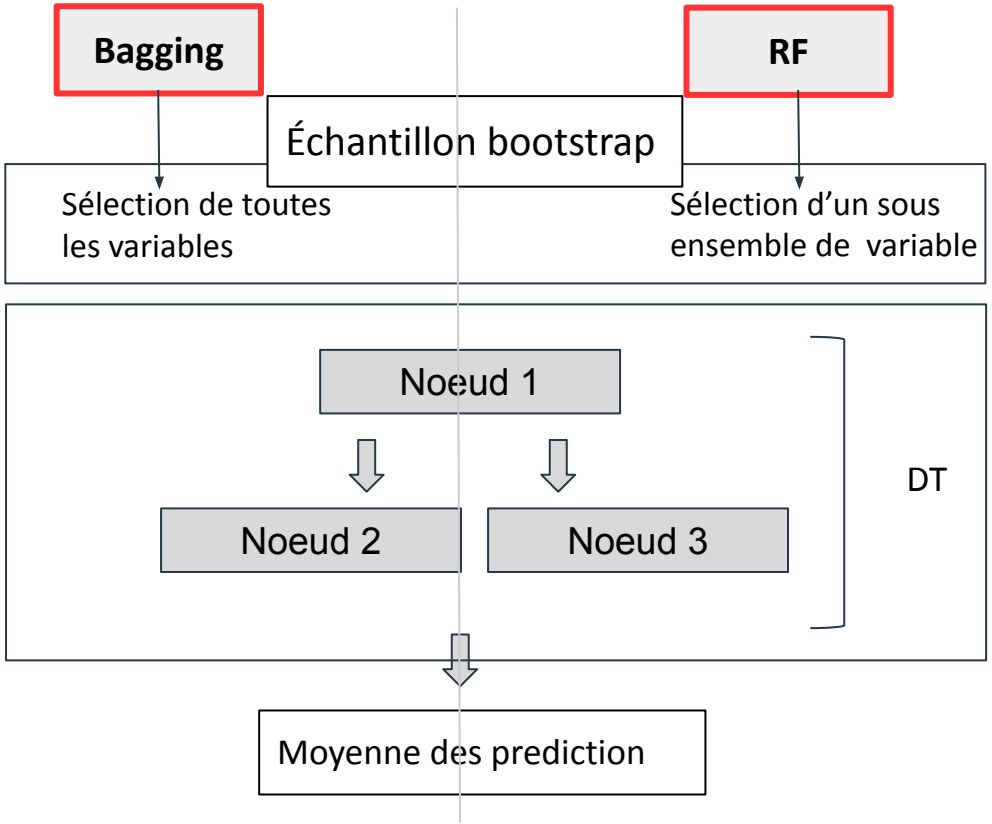
Représentation dans un espace multidimensionnelle

DT



Découper l'espace des variables partitionnement récursif

Les modèles d'apprentissage automatique



AdaBoost

DT

⇩

Ajuste le poids des variables (Sélection itérative)

⇩

Moyenne des poids

Entraînement sur la Marne

Model	KNN	RF	DT	SVM	AdaBoost	Bagging
RMSE	0.41±0.28	0.37±0.20	0.54±0.29	0.53±0.48	0.53±0.28	0.38±0.19
MAE	0.09±0.03	0.09±0.02	0.14±0.05	0.13±0.05	0.10±0.03	0.14±0.06
RDP	1.60±0.49	1.91±1.65	1.12±0.36	1.32±0.22	1.28±0.62	1.77± 1.62

- KNN, Bagging et RF :
  - Erreur RMSE et MAE faible : Pouvoir de prédiction le plus élevé
  - RPD > 1.4 : Modèles modérément précis

⇒ Le modèle RF présente le pouvoir de prédiction le plus élevé.

Entraînement sur la Seine

Model	KNN	RF	DT	SVM	AdaBoost	Bagging
RMSE	0.69±0.09	0.67±0.09	0.77±0.12	0.75±0.12	0.73±0.10	0.68±0.09
MAE	0.34±0.03	0.33±0.02	0.40±0.03	0.32±0.04	0.35±0.02	0.30±0.03
RDP	1.43±0.21	1.47±0.23	1.30±0.13	1.33±0.11	1.37±0.25	1.44± 0.18

⇒ Le modèle RF présente le pouvoir de prédiction le plus élevé.

- Performance : entraînement avec les données de la Seine > la Marne

En Marne

Corrélation avec la concentrations en *E. coli*

(test t)

Paramètres	Prédictions raisonnables	Prédictions inexactes	p-valeur
Température	-0.17 ± 0.05	-0.28 ± 0.07	0.001
Conductivité	-0.05 ± 0.11	-0.18 ± 0.09	0.009
Turbidité	0.42 ± 0.07	0.39 ± 0.08	0.43
MES	0.43 ± 0.09	0.40 ± 0.04	0.42
NH_4^+	0.54 ± 0.06	0.48 ± 0.07	0.05
NTK	-0.03 ± 0.08	0.001 ± 0.06	0.26
Nombre de jours secs	-0.10 ± 0.09	-0.01 ± 0.09	0.02
Pluviométrie du jour	0.09 ± 0.10	-0.02 ± 0.11	0.02
Pluviométrie de la veille	0.17 ± 0.08	0.03 ± 0.10	0.002
Débit à Gournay/Marne	0.54 ± 0.09	0.39 ± 0.09	0.001

En Seine

(test t, Wilcoxon)

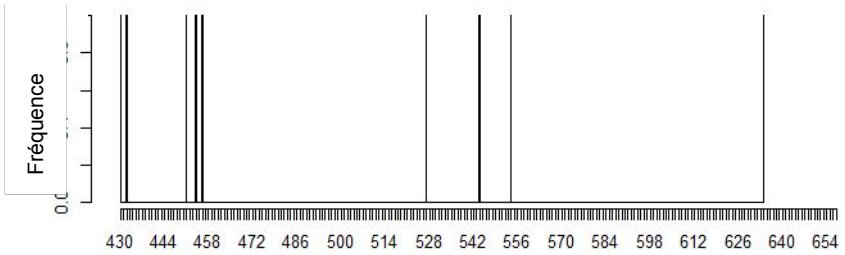
Paramètres	Prédictions raisonnables	Prédictions inexactes	p-valeur
Température	-0.51 ± 0.06	-0.37 ± 0.06	0.002
Conductivité	-0.45 ± 0.10	-0.26 ± 0.08	0.002
Turbidité	0.39 ± 0.15	0.28 ± 0.09	0.173
Nombre de jours secs	-0.45 ± 0.16	-0.32 ± 0.09	0.139
Pluviométrie du jour	0.44 ± 0.15	0.27 ± 0.09	0.01
Pluviométrie de la veille	0.56 ± 0.08	0.45 ± 0.08	0.10
Débit à Austerlitz	0.26 ± 0.10	0.33 ± 0.08	0.18

En Marne

A

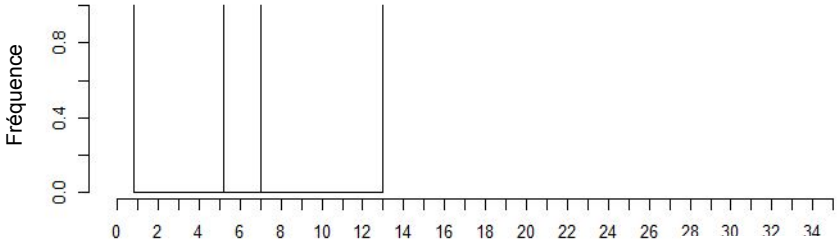
Température (°C (r=-0.17)

B



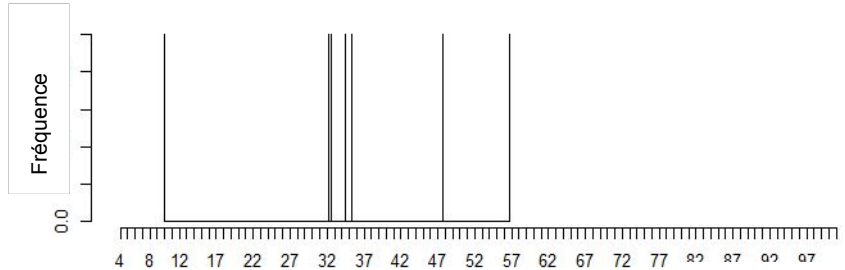
Conductivité (µS/cm) (r=-0.05)

C



Pluviométrie de la veille (mm (r=0.17)

D



Débit à Gournay/Marne (m³/s) (r=0.54)