Un système agnostique de détection et de diagnostic d'intrusion radio pour l'IoT

Pierre-François Gimenez, Jonathan Roux, Éric Alata, Guillaume Auriol, Mohamed Kaâniche, Vincent Nicomette

Nouvelles Avancées en Sécurité des Systèmes d'Information
22 janvier 2020

L'IoT explose

L'IoT se développe...

- L'IoT (Internet of Things) se développe rapidement
- Tout devient connecté (« smart »): les télévisions, les périphériques, les ampoules, les fourchettes...
- Environnements connectés: ville, bâtiment, usine, domicile...

... et les menaces aussi

- Réduire une fonctionnalité. Exemple: couper le refroidissement d'un frigo
- Manipuler la fonctionnalité. Exemple: ouvrir des volets de l'extérieur
- Étendre la fonctionnalité. Exemple: fuiter de données avec une ampoule
- Utiliser l'objet pour lui-même. Exemple: créer un botnet (Mirai)

Spécificités

Comment protéger l'IoT ? Ce n'est pas facile:

Spécificités de l'IoT vis-à-vis de la sécurité

Hétérogénéité protocoles (propriétaires ou non) et architectures matérielles en constante évolution

Mobilité réseaux dynamiques (ex: smartphones, montres connectées...)

Nombre ces objets se multiplient, ce qui augmente les interactions

Vulnérabilité manque d'expertise et de sensibilisation

Approches existantes

Pare-feu Filtre les communications

VPN Isole des réseaux

Système de détection d'intrusion réseau (IDS) Surveille les communications réseaux

→ Solutions partielles à cause de l'hétérogénéité, inadaptées aux réseaux décentralisés

→ IDS: diagnostic spécialisé pour certains types d'attaque ou certaines technologies

Sommaire

Radio IDS

Notre approche: un IDS...

- basé sur les communications radio (couche physique) → fonctionne avec les protocoles propriétaires (**agnostique**)
- qui surveille plusieurs larges bandes de fréquences ($\sim$ 100MHz) → indépendant des protocoles (y compris futurs)
- qui modélise le comportement normal et n'utilise pas de signatures → indépendant des attaques
- qui aide au diagnostic → pour faciliter le traitement de l'anomalie

Remarques

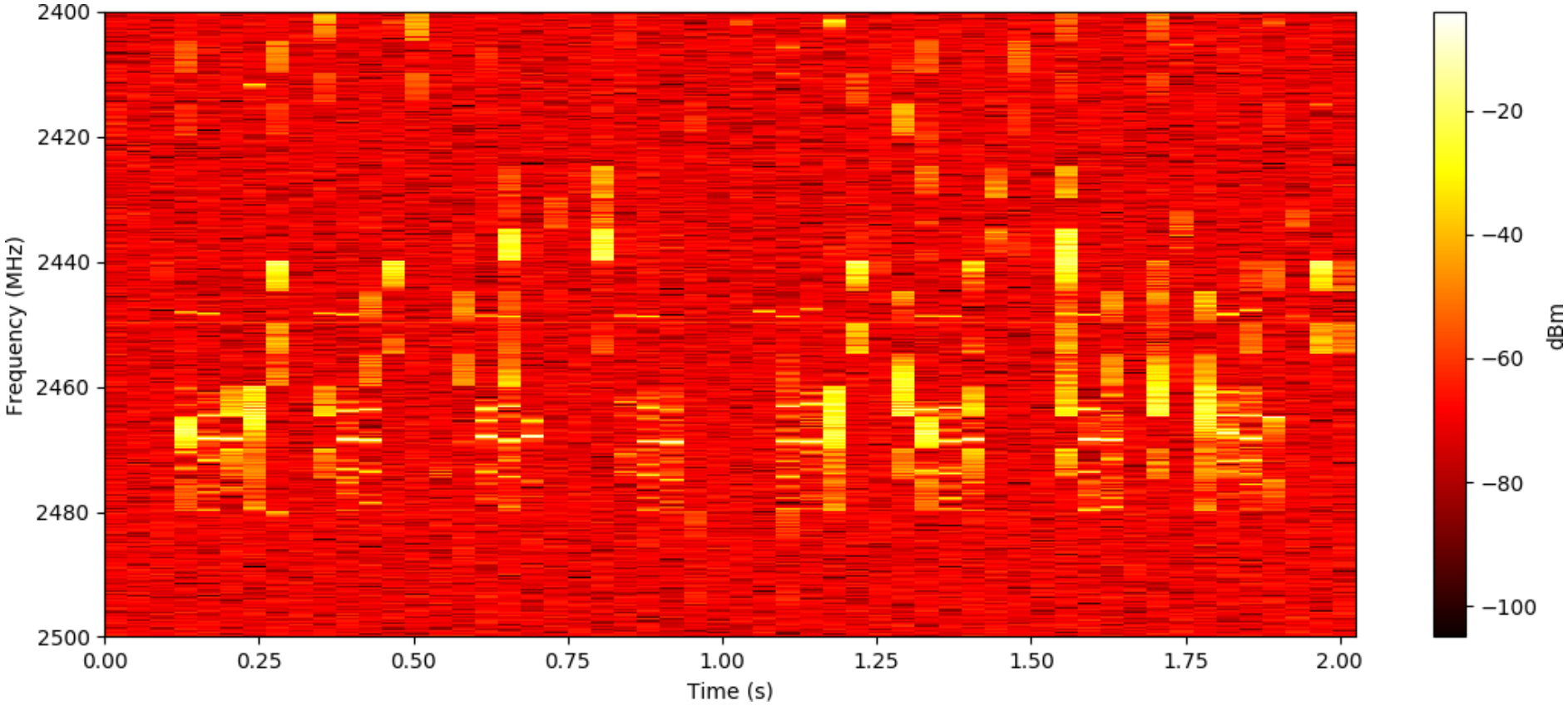
- On ne couvre pas les communications filaires (déjà surveillées et moins utilisées dans l'IoT)
- On suppose qu'une attaque est perceptible dès la couche physique

IDS basé sur SDR

- Besoin: mesurer les puissances radio sur un large spectre
- Outil: SDR (software defined radio) qui peut balayer de larges bandes de fréquence (jusqu'à 6 GHz)
- On obtient la puissance en chaque fréquence en fonction du temps (par FFT): un **spectrogramme** (ou un waterfall)
- Nombre de mesures à la seconde fixé: compromis entre résolution temporelle et fréquentielle

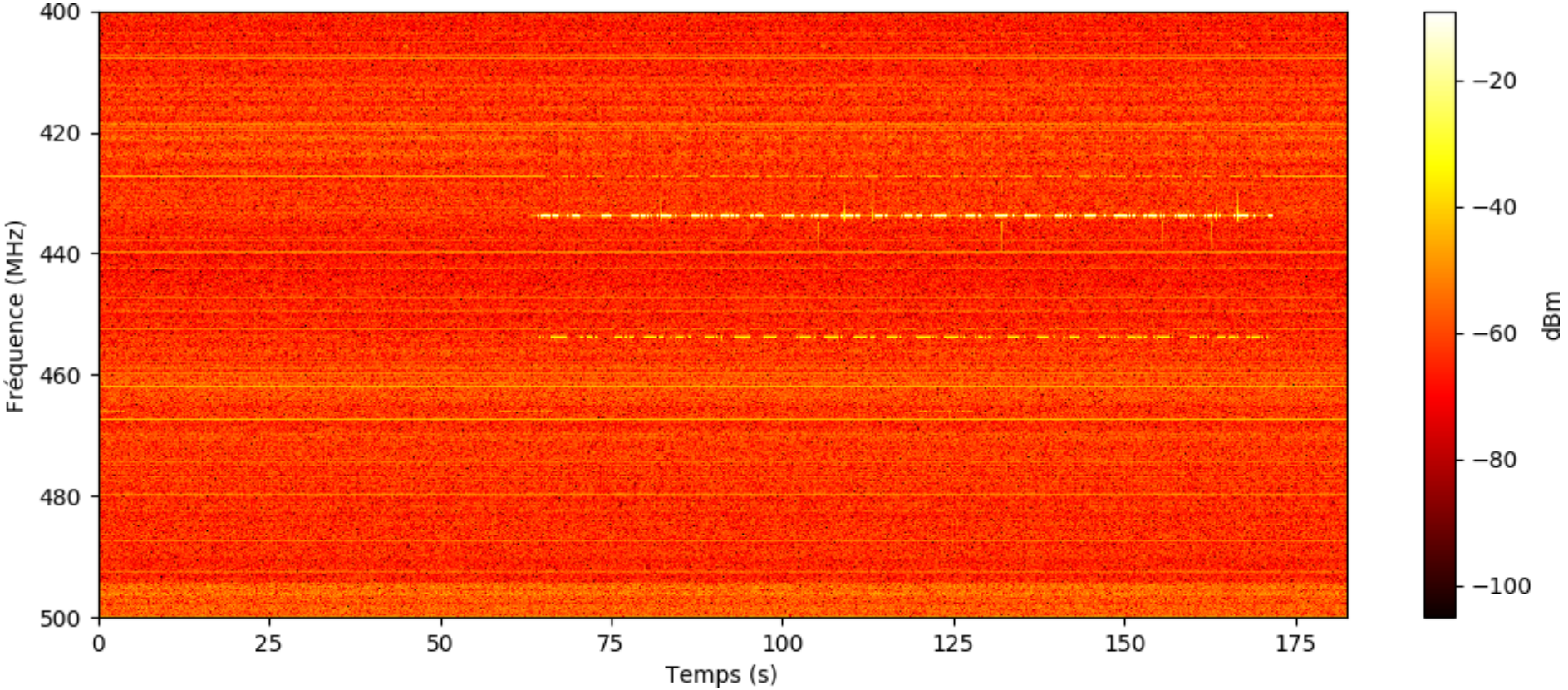
Modèle utilisé : HackRF One (280 euros)

Exemple de spectrogramme



Perte d'information → impossible de récupérer les communications haut niveau à partir d'un spectrogramme (démodulation impossible)

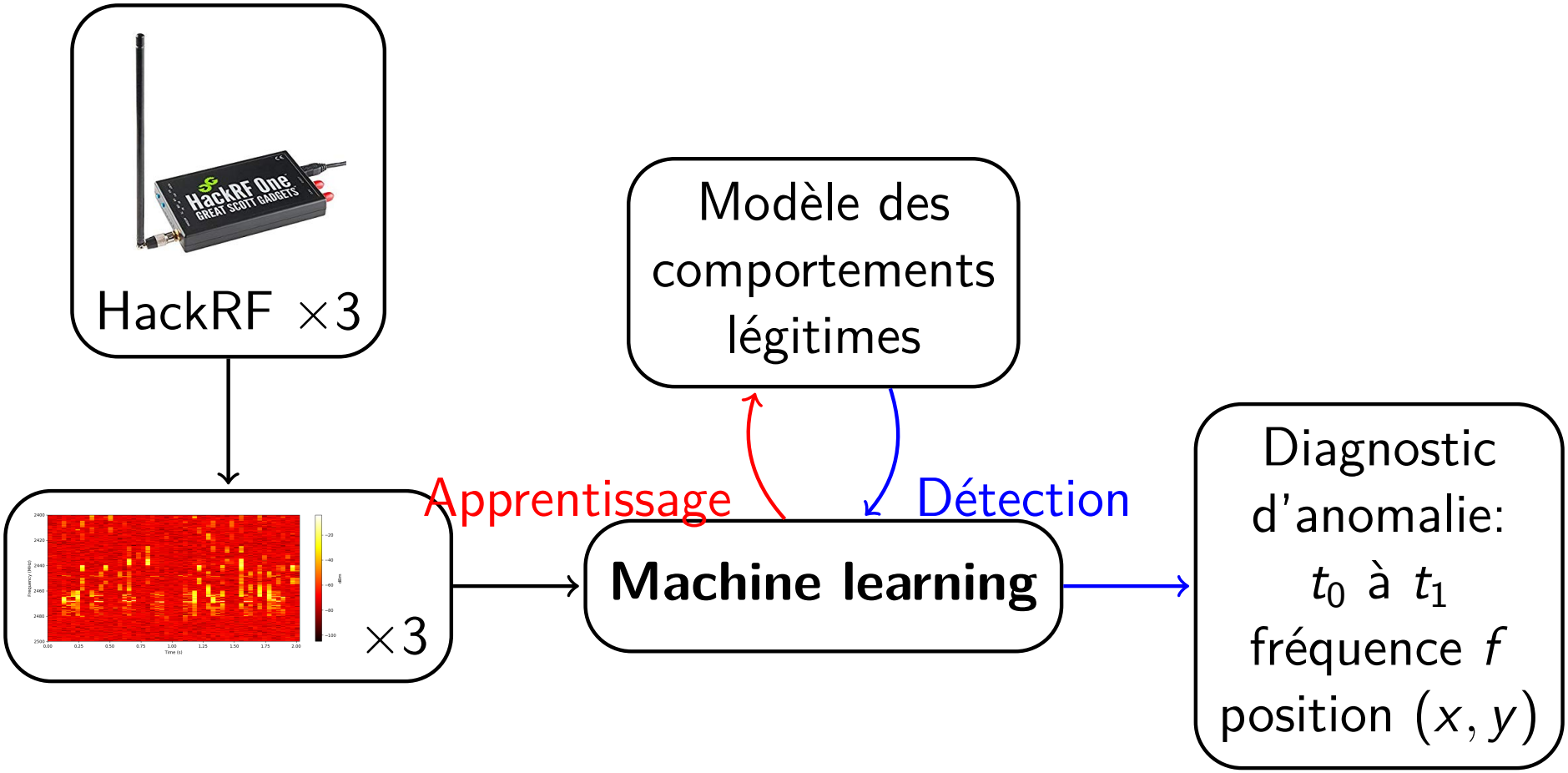
Exemple d'anomalie



Une attaque par séquence de Bruijn → elle est visible

Schéma général de l'approche

- Phase d'apprentissage du comportement normal
- Phase de détection d'anomalie



Hypothèses

Pour simplifier le problème, on établit certaines hypothèses.

Hypothèses sur les anomalies

- Les anomalies peuvent être détectées dès la couche physique
- Si on surveille plusieurs larges bandes de fréquence, on suppose qu'une anomalie peut être détectée en traitant chaque large bande indépendamment
- Si on a plusieurs sondes, on suppose qu'une anomalie peut être détectée en traitant chaque sonde indépendamment
- Une anomalie peut être détectée sur une courte durée: on ne traite pas de cas complexe comme « l'utilisateur allume toujours la lumière avant de faire un café »

Par contre, on s'impose de ne pas perdre d'information en « résumant » le spectrogramme

Quel modèle utiliser ?

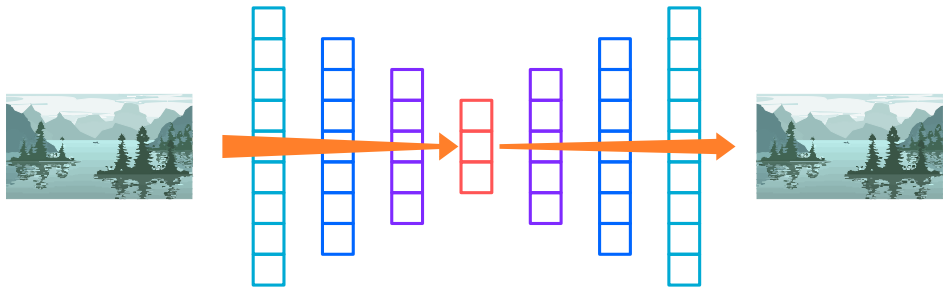
Les contraintes du problème

- Pas d'expertise sur la forme des attaques dans le spectre radio
- Les spectrogrammes sont de grandes tailles ($\sim 10^4$ points)
- Plusieurs Go de données d'apprentissage
- L'inférence doit être faisable en temps-réel
- Le modèle doit permettre de localiser l'anomalie

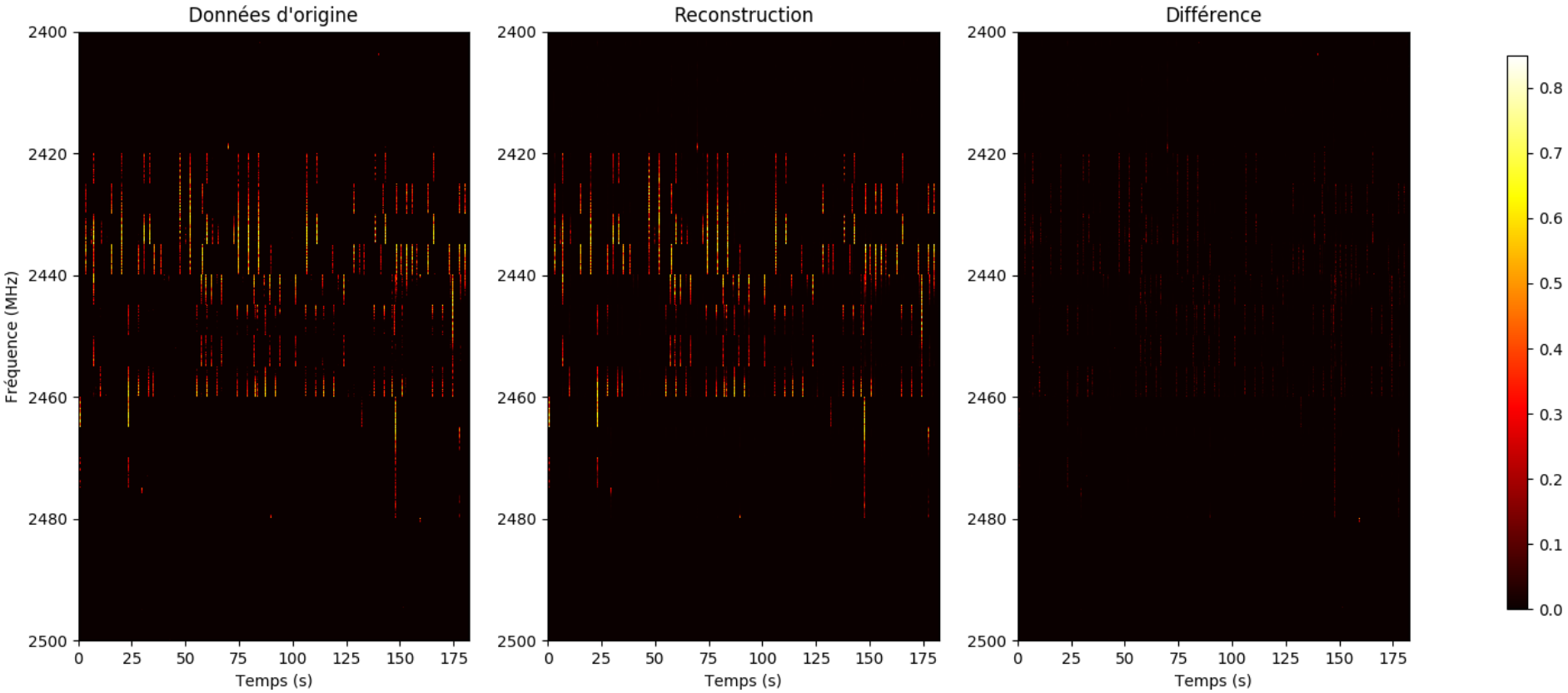
$\rightarrow$ Les réseaux de neurones sont adaptés à ce problème, plus précisément les auto-encodeurs

Auto-encodeur

- Un réseau de neurones qui cherche à reconstruire ses entrées
- Le réseau a un goulot d'étranglement: il doit compresser l'entrée intelligemment
- Cette compression est adaptée aux **exemples d'apprentissage légitimes**
- Une fois appris: s'il reconstruit bien l'entrée, c'est qu'elle est légitime. Sinon, c'est que c'est une anomalie.
- On a une **erreur de reconstruction**: on fixe un seuil qui délimite les entrées légitimes des anomalies

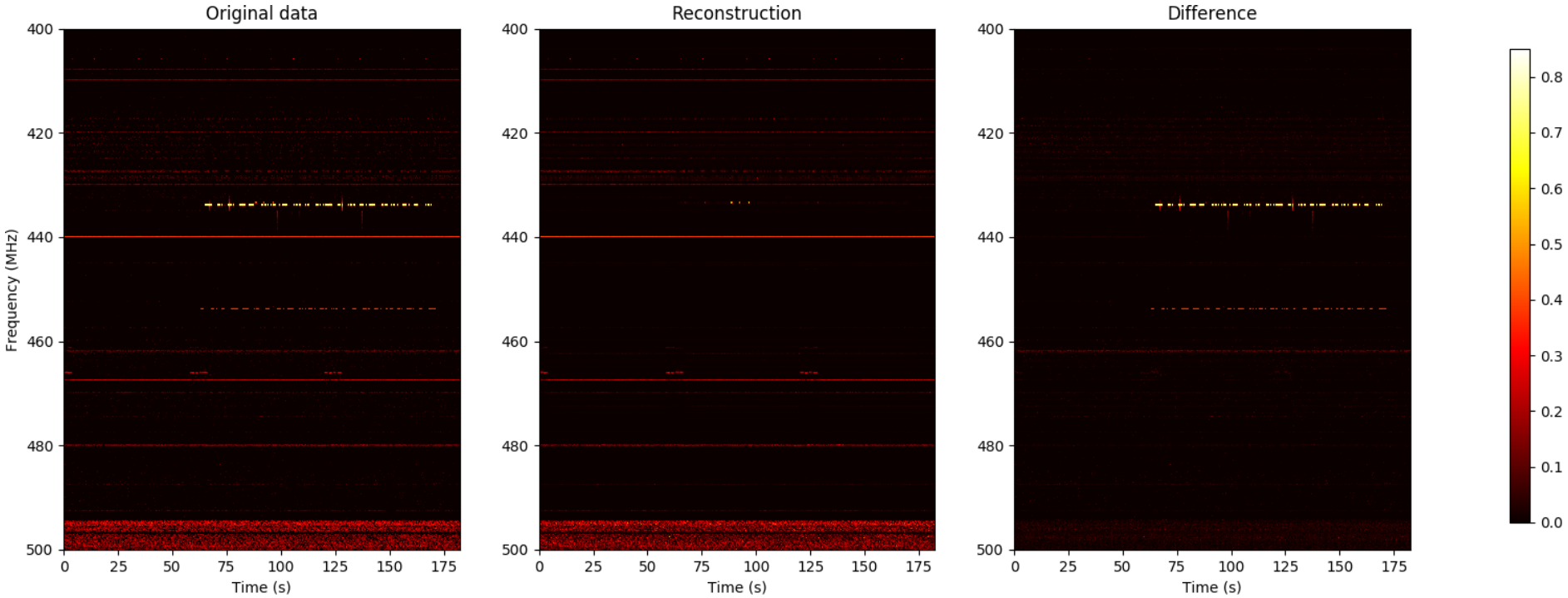


Exemple de reconstruction sans anomalie



Faible erreur de reconstruction (image de droite)

Exemple de reconstruction avec anomalie



Grande erreur de reconstruction (image de droite)

Paramètres de la SDR pour l'auto-encodeur

SDR paramétrable

- Compromis entre résolution temporelle et fréquentielle
- Choix: résolution fréquentielle de 0.1MHz et résolution temporelle de 37.5ms
- On balaye 300 MHz, ce qui donne 3000 mesures par balayage

Longueur temporelle à choisir

- Nombre de paramètres de l'auto-encodeur $\sim$ quadratique en la taille de l'entrée
- Beaucoup de paramètres $\rightarrow$ besoin de plus de données, de puissance de calcul et de temps d'apprentissage (et peut mener à du surapprentissage)
- Compromis entre taille des données et prise en compte du comportement temporel
- Choix: 600ms de mémoire (16 balayages)

Diagnostic

Motivation

- Large surface surveillée : il ne suffit pas de dire qu'on a vu, il faut dire où
- En milieu professionnel, on peut donner des informations techniques à l'utilisateur

→ Traitable par un SOC (security operation center)

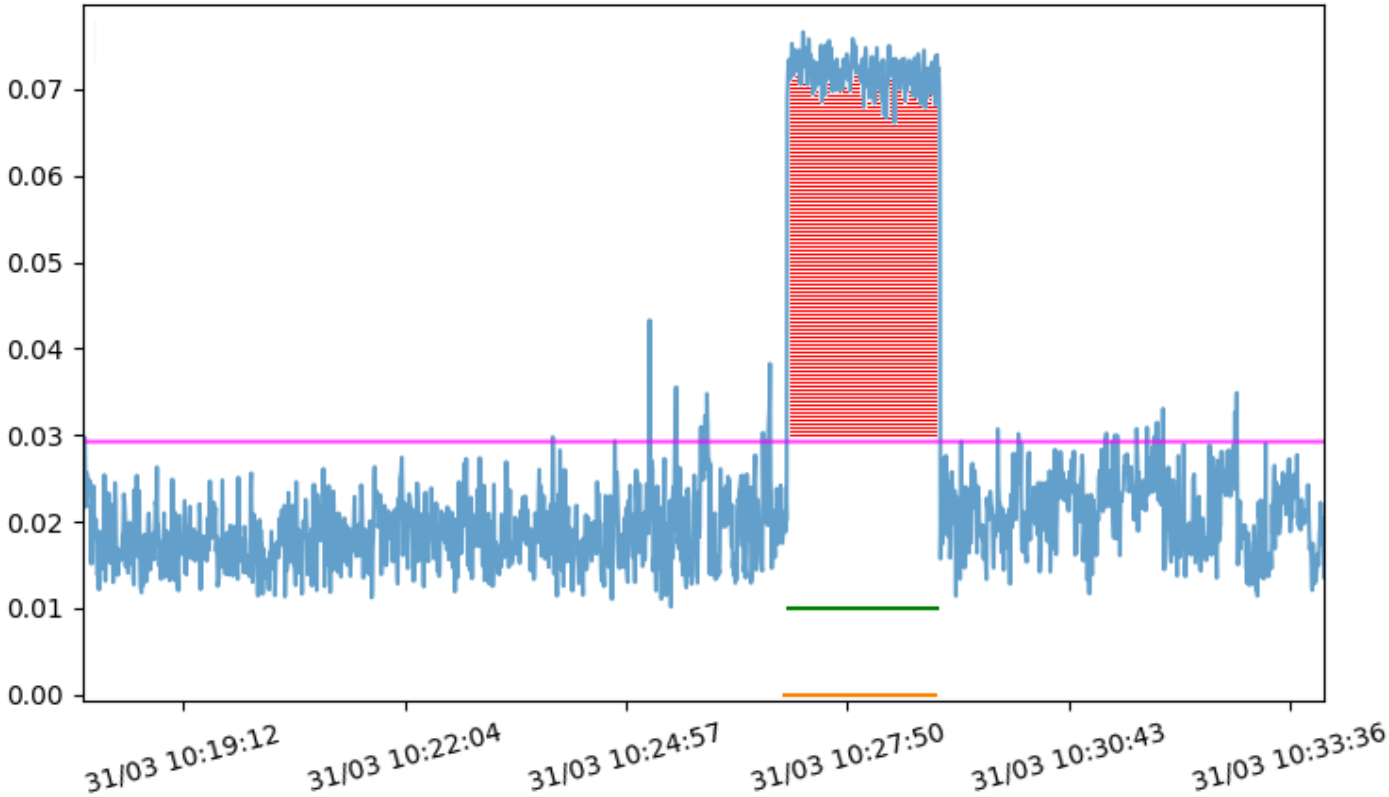
Objectif

On travaille toujours sur la couche physique ! (pas de démodulation possible)

- Les dates de début et de fin de l'anomalie
- La fréquence de l'anomalie
- La position géographique de l'anomalie (approximative)

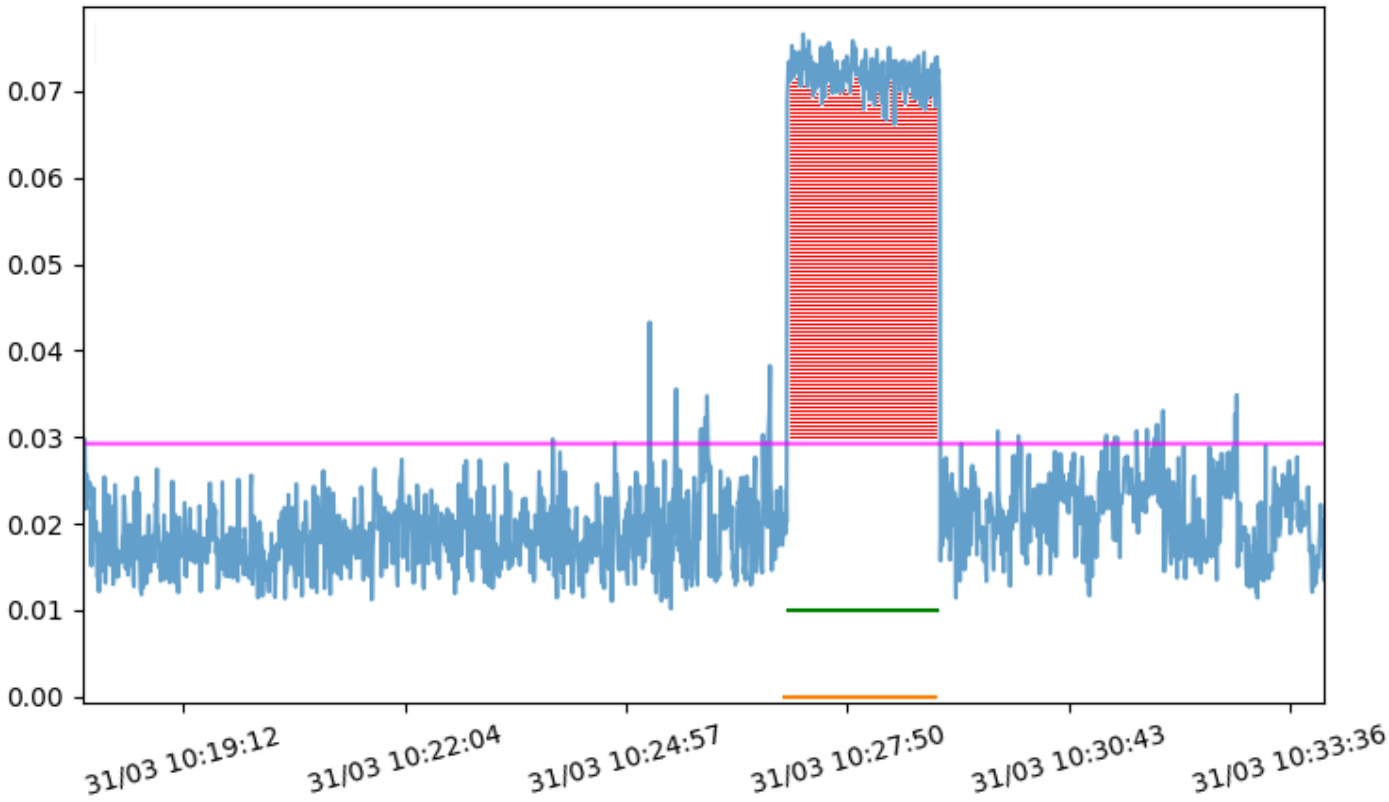
Détection

Erreur de reconstruction en fonction du temps



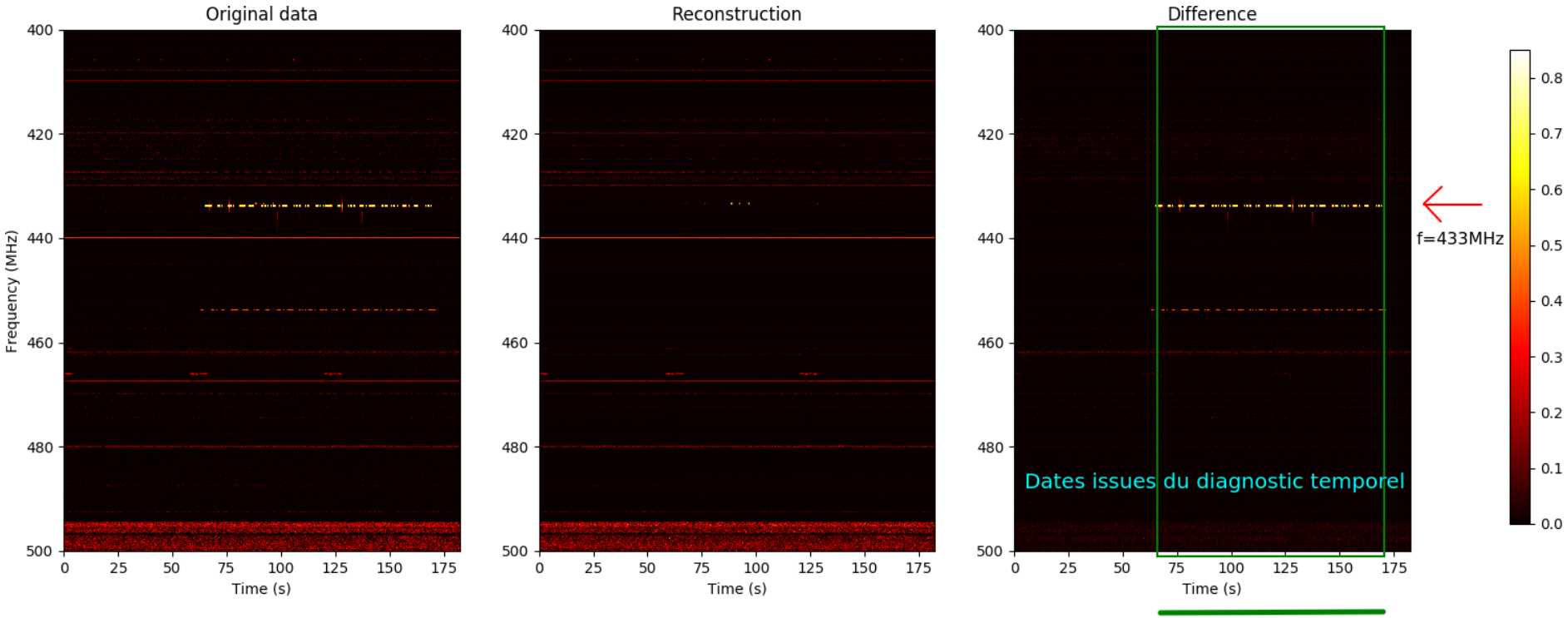
Pour éviter les faux positifs: on regarde l'erreur cumulative (en rouge)
Les seuils sont appris

Diagnostic temporel



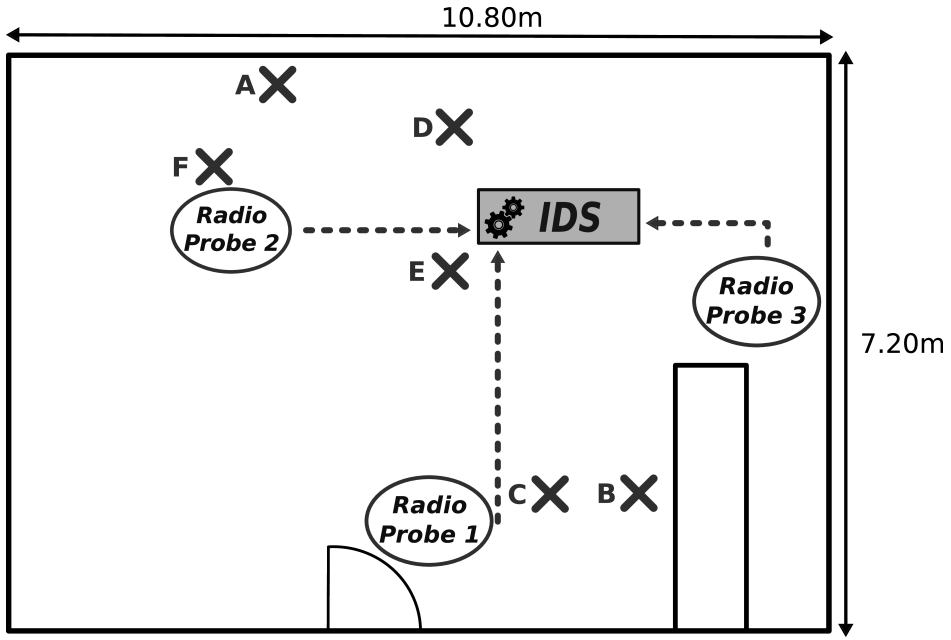
Intervalle de temps déduit de la surface rouge.
Ici: notre détection et l'attaque

Diagnostic fréquentiel



Diagnostic fréquentiel: la fréquence avec l'erreur maximale
Repose sur le diagnostic temporel!

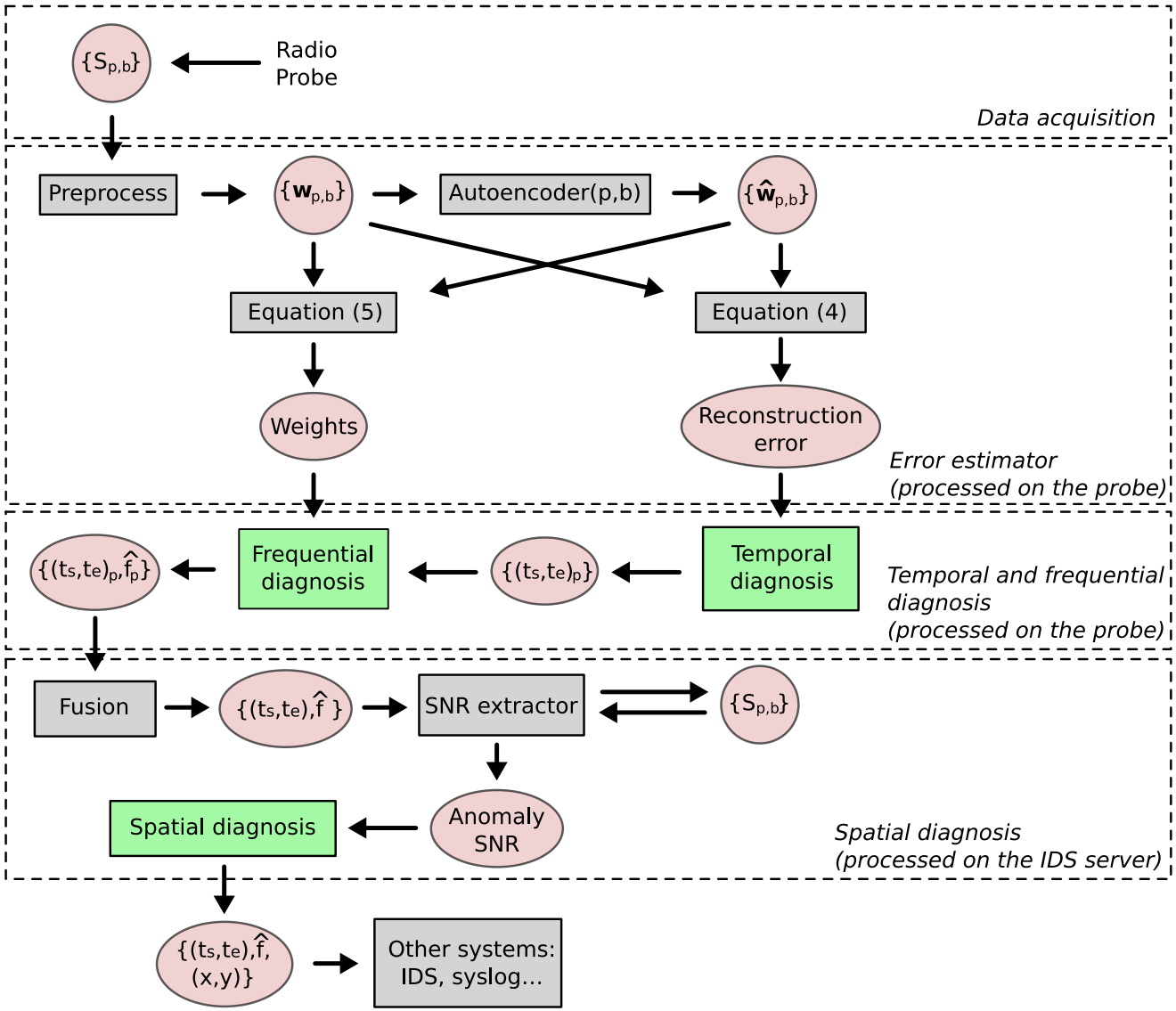
Diagnostic spatial



Principe

- Dates et fréquence de l'anomalie connues : on peut estimer la puissance du signal
- Estimation de la position grâce à un barycentre des puissances de l'anomalie perçues par chaque sonde

Architecture de l'IDS



Salle d'expérimentation

Objets connectés

- Ampoule, sonnette, clavier, téléphones, ...
- Environnement pas complètement contrôlé : c'est une salle de pause avec des visiteurs et des stagiaires

Matériel IDS

- Bandes surveillées : 400-500MHz, 800-900MHz, 2400-2500MHz (bandes libres)
- 3 HackRF avec des antennes différentes (2 Go mesurés /jour /sonde)
- 9 jours d'apprentissage, 11 jours de tests
- Apprentissage et inférence faits avec un serveur de calcul (hors ligne)

Attaques lancées avec Mirage

Mirage : framework modulaire dédié à l'analyse des réseaux sans-fils
https://redmine.laas.fr/projects/mirage

ID	Name	Technology	Type	Freq./Band	#Inj.
			400–500 MHz		
1	scan433-17	433 MHz	Scan 433 17 dBm	433 MHz	256
2	DoS433-27	433 MHz	DoS 27 dBm	433 MHz	50
3	DoS433-40	433 MHz	DoS 40 dBm	433 MHz	13
4	TV-spoofing	DVB-T	TV spoofing	485–499 MHz	12
5	bruijn	433 MHz	De Bruijn injection	433 MHz	13
			800–900 MHz		
9	scan868	868 MHz	Scan 868 20 dBm	868 MHz	279
10	DoS868	868 MHz	DoS 35 dBm	868 MHz	80
			2.4–2.5 GHz		
13	blescan	BLE	BLE Scan	2.4–2.5 GHz	177
14	zigbeescan	Zigbee	Zigbee Scan	2.4–2.5 GHz	25
15	deauth	WiFi	Deauthentication	2.451–2.473 GHz	33
16	rogueAP	WiFi	RogueAP	2.461–2.483 GHz	83
17	esbinject	ESB	Injection ESB	2.4–2.5 GHz	66
18	injectzigbee	Zigbee	Injection Zigbee	2.48 GHz	33

Métriques de performances

Précision

Proportion du temps d'alarme qui correspond à de vraies anomalies
$\rightarrow$ bonne précision = peu de fausses alarmes

Rappel

Proportions du temps d'anomalies correctement détecté
$\rightarrow$ bon rappel = peu d'anomalies ratées

Accuracy

Proportion des prédictions (alarmes ou non) qui sont correctes
$\rightarrow$ bonne accuracy = peu d'erreurs

Résultat du diagnostic temporel (rappel)

Name	Probe 1	Probe 2	Probe 3
400–500 MHz			
scan433-17	51.52%	96.63%	0%
DoS433-27	99.19%	99.21%	4.00%
DoS433-40	99.19%	99.87%	95.77%
TV-spoofing	99.20%	99.85%	99.00%
bruijn	77.88%	96.60%	84.48%
probfail-1	99.83%	99.88%	9.59%
anomaly462	3.51%	100%	5.36%
anomaly467	70.75%	29.70%	6.18%
800–900 MHz			
scan868	64.97%	74.13%	64.56%
DoS868	82.09%	93.33%	63.95%
harmo433	40.02%	40.73%	3.35%
probfail-2	0%	0%	3.32%
2.4–2.5 GHz			
ID 13–18	$\leq 1\%$	$\leq 1\%$	$\leq 1\%$

Résultat du diagnostic temporel (précision et accuracy)

Précision:

Bande	Sonde 1	Sonde 2	Sonde 3
400–500 MHz	99.13%	79.79%	87.54%
800–900 MHz	97.76%	96.38%	97.40%
2.4–2.5 GHz	15.90%	8.03%	5.43%

Accuracy:

Bande	Sonde 1	Sonde 2	Sonde 3
400–500 MHz	81.94%	93.93%	78.32%
800–900 MHz	96.99%	97.44%	96.61%
2.4–2.5 GHz	93.57%	93.54%	93.60%

Résultat du diagnostic fréquentiel: erreur médiane

Name	Frequency	Probe 1	Probe 2	Probe 3
400–500 MHz				
scan433-17	433 MHz	0.1 MHz	0.1 MHz	×
DoS433-27	433 MHz	0.1 MHz	0.1 MHz	63.4 MHz
DoS433-40	433 MHz	0.1 MHz	0 MHz	0 MHz
TV-spoofing	485–499 MHz	0 MHz	0 MHz	0 MHz
bruijn	433.8 MHz	0 MHz	0 MHz	0 MHz
anomaly462	462 MHz	×	0.1 MHz	×
anomaly467	467 MHz	0 MHz	0 MHz	×
800–900 MHz				
scan868	868 MHz	0.1 MHz	0.1 MHz	0.1 MHz
DoS868	868 MHz	0.1 MHz	0.1 MHz	0.2 MHz
harmo433	867.7 MHz	0 MHz	0 MHz	0 MHz

Résultat du diagnostic spatial : distance médiane

Name	Pos. A	Pos. B	Pos. C	Pos. D	Pos. E	Pos. F	Moy.
scan433-17	-	-	-	-	-	3.63 m	3.63 m
DoS433-27	-	-	-	-	-	3.96 m	3.96 m
DoS433-40	-	0.82 m	0.51 m	-	1.24 m	-	0.86 m
TV-spoofing	-	0.45 m	0.35 m	-	1.92 m	-	0.91 m
bruijn	-	0.95 m	0.32 m	-	2.12 m	-	1.13 m
scan868	2.45 m	-	-	-	-	2.91 m	2.68 m
DoS868	3.16 m	2.47 m	2.67 m	3.13 m	1.15 m	2.19 m	2.46 m
harmo433	2.13 m	-	-	2.40 m	-	-	2.27 m
Moyenne	2.68 m	1.17 m	0.91 m	2.77 m	1.61 m	3.17 m	**1.95 m**

Problème: antenne différente par sonde et pas omnidirectionnelle

Conclusion

Performances

- Détection efficace sur les protocoles peu utilisés, inexistante sur 2.4–2.5 GHz (mais plein d'outils de détection pour le 2.4–2.4 GHz)
- Diagnostic temporel et fréquentiel précis, diagnostic spatial peu satisfaisant

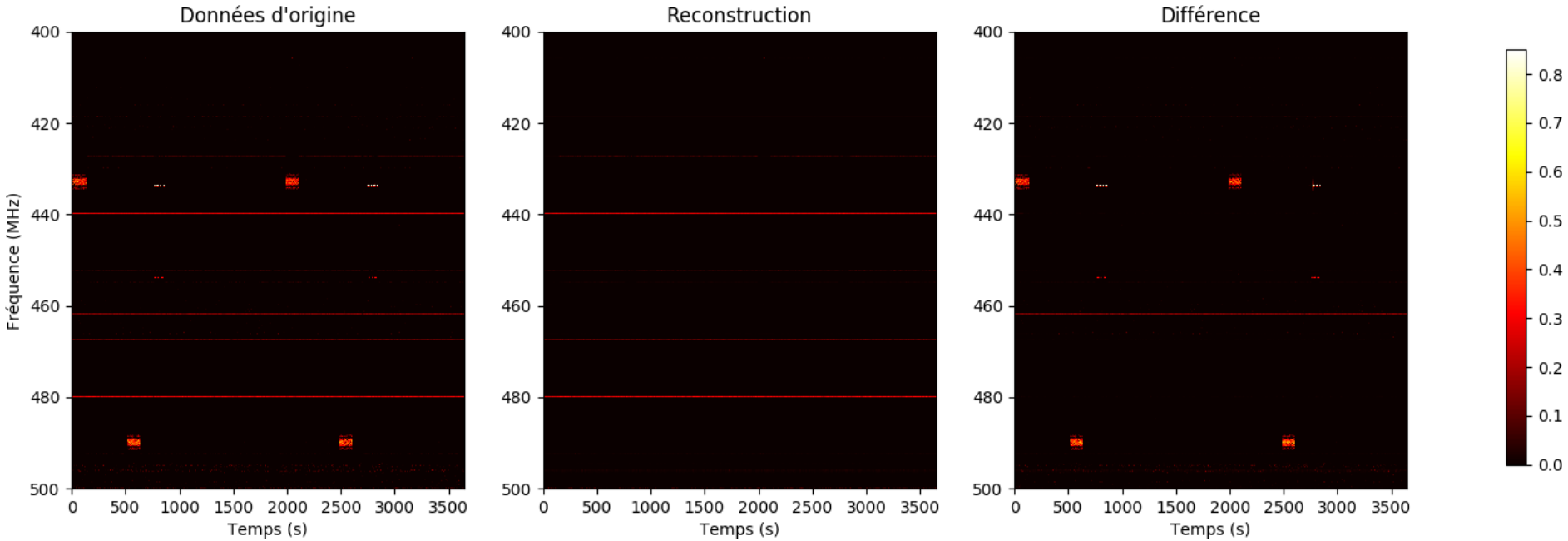
Expérience en ligne

- Sonde pas chère: Raspberry Pi 4 + HackRF
- 40% CPU utilisé
- 2Go RAM utilisé
- Consommation électrique: 4.5W (moins qu'une ampoule LED)

Et quelques surprises...

Anomalie 462 MHz

- Signal puissant pendant 33h sur la fréquence 462 MHz
- Signal retracé jusqu'à l'écran d'un doctorant avec un HackRF → signal non-malveillant
- Montre la pertinence du diagnostic



Résumé et conclusion

Approche

IDS radio qui surveille de larges bandes de fréquences sans hypothèses sur les attaques et qui aide au diagnostic grâce au machine learning

Performances

- Matériel pas cher (environ 350 euros)
- Traite efficacement des bandes peu surveillées: complémentaire aux IDS existants

Perspectives

- Améliorer la détection 2.4–2.5 GHz
- Détecter une fuite de données radio
- Détecter des anomalies longues