CLOUD-BASED CROWD MONITORING

MAARTEN VAN STEEN

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Valeriu-Daniel Stanciu

Ciprian Dobre

Andreas Peter

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

WIFI-BASED PEDESTRIAN MONITORING IN A NUTSHELL

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

MOBILE SENDER

MESSAGES

FIXED RECEIVER

38:f9:d3:51:0e:de
(source address)

b0:be:76:e3:17:2b
(destination address)

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

MOBILE SENDER

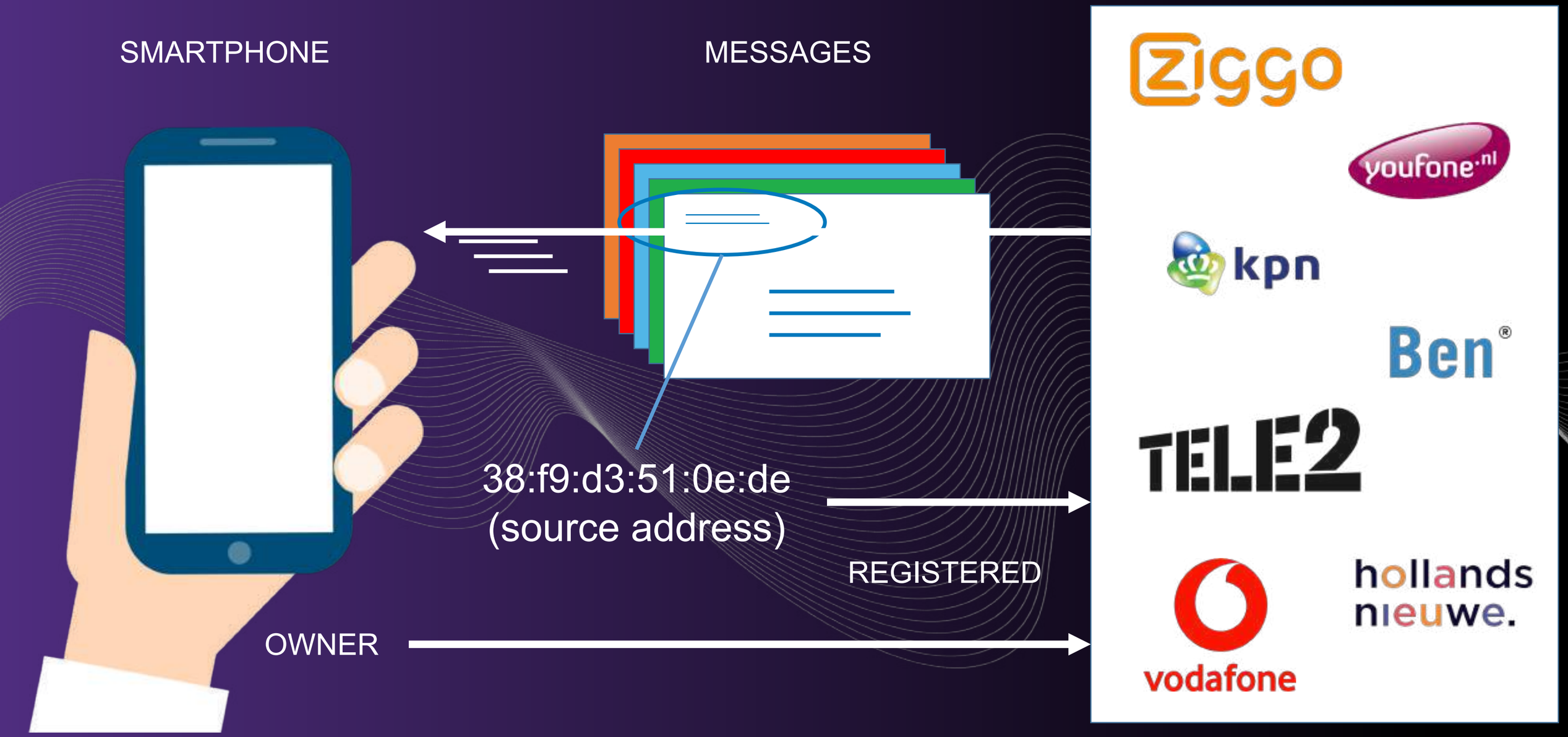
MESSAGES

FIXED RECEIVER

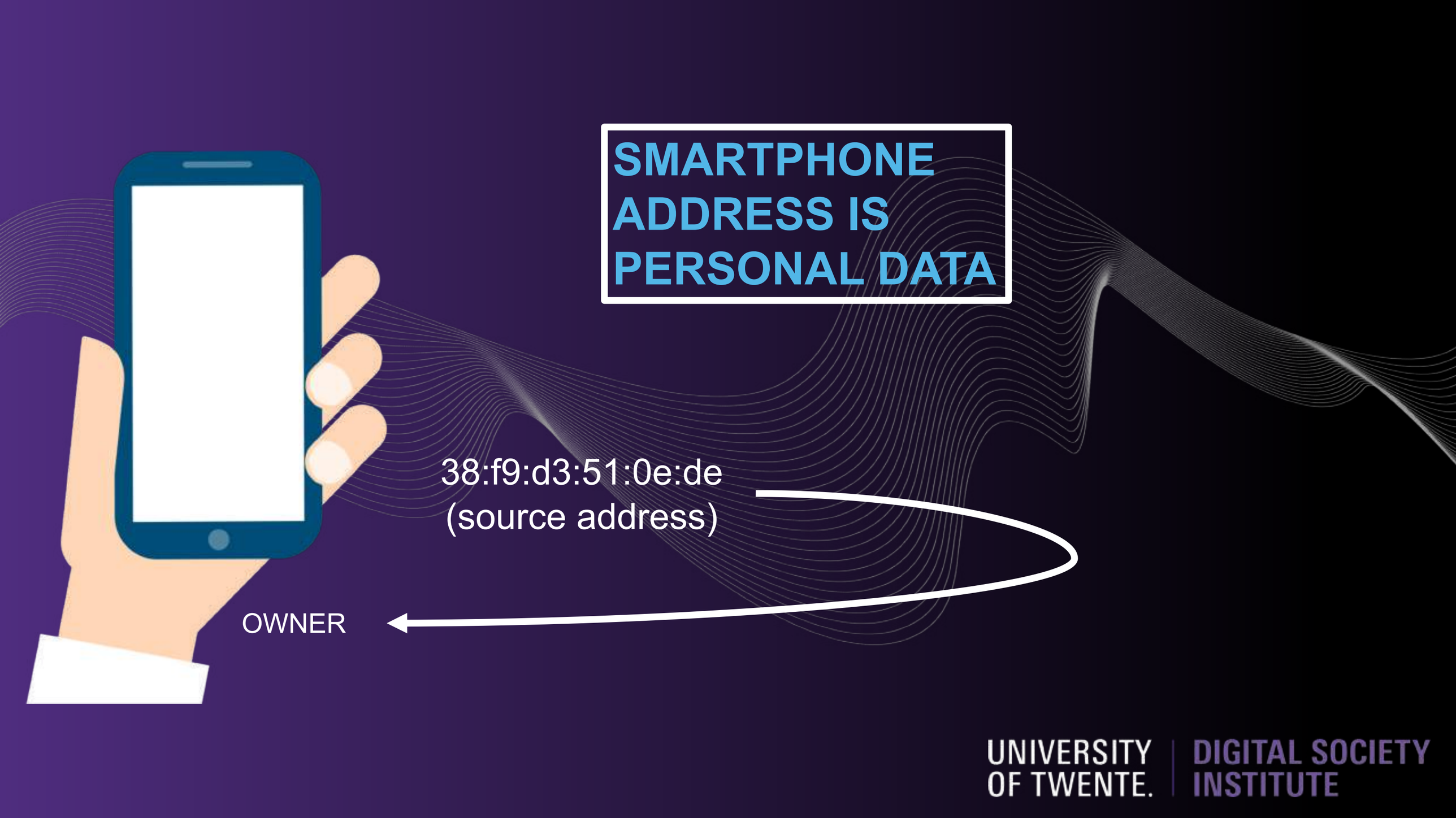
38:f9:d3:51:0e:de
(source address)

AT RECEIVER (WITH KNOWN LOCATION):
- RECORD SOURCE ADDRESS
- RECORD TIME
AND YOU KNOW WHERE AND WHEN A PHONE WAS DETECTED

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE



UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE



UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

WORKAROUND?

COMPUTATIONALLY EASY

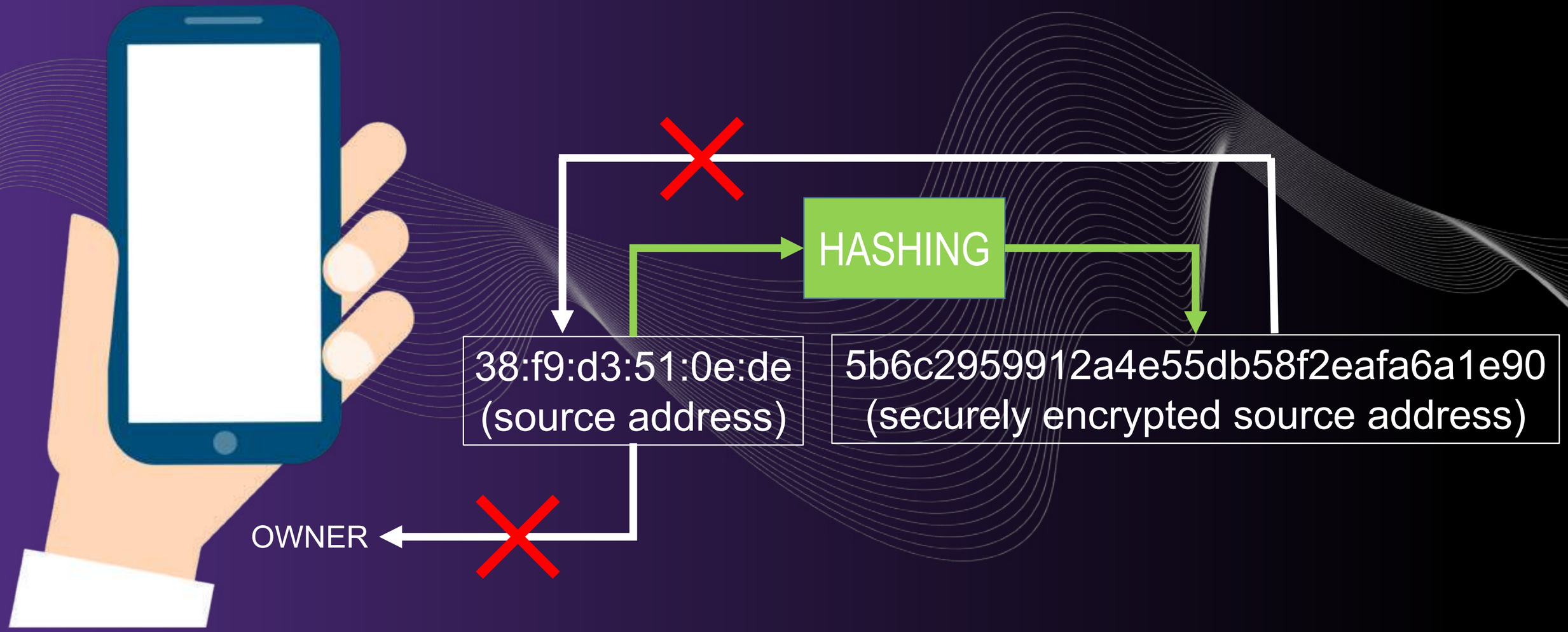
INPUT → HASHING → HASH

COMPUTATIONALLY HARD
(BRUTE FORCE NEEDED)

38:f9:d3:51:0e:de
(source address)

OWNER

WORKAROUND?


HASHING
38:f9:d3:51:0e:de
(source address)
5b6c2959912a4e55db58f2eafa6a1e90
(securely encrypted source address)
OWNER

WORKAROUND?

GDPR: AN INDIVIDUAL MAY NOT BE IDENTIFIABLE FROM A DATA SET

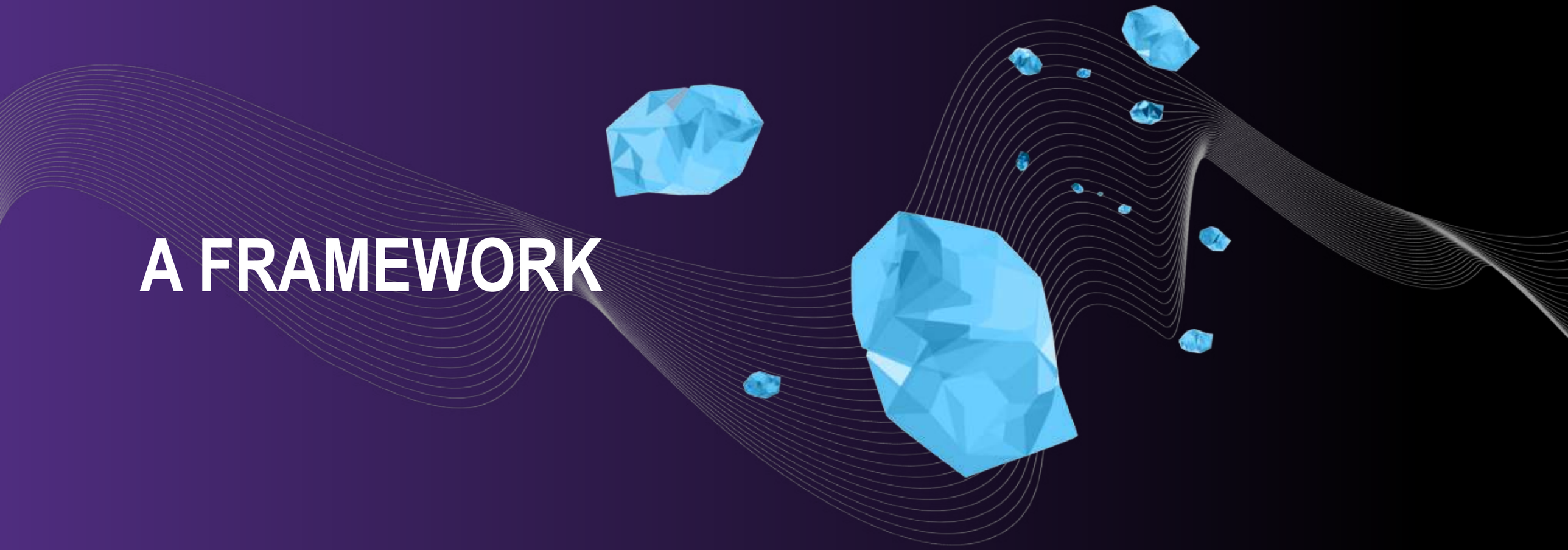
HASHING

38:f9:d3:51:0e:de
(source address)

5b6c2959912a4e55db58f2eafa6a1e90
(securely encrypted source address)

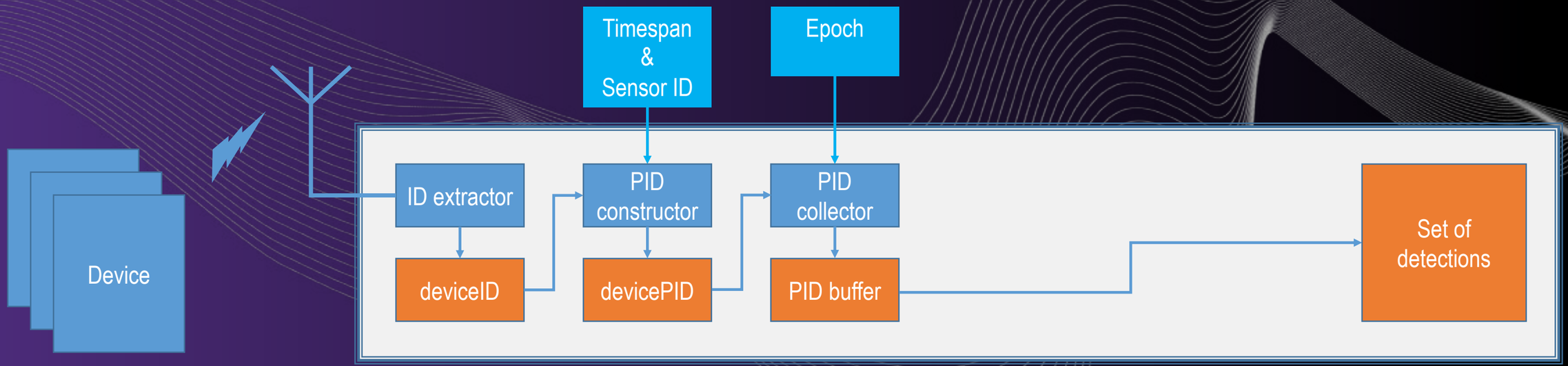
OWNER

A FRAMEWORK

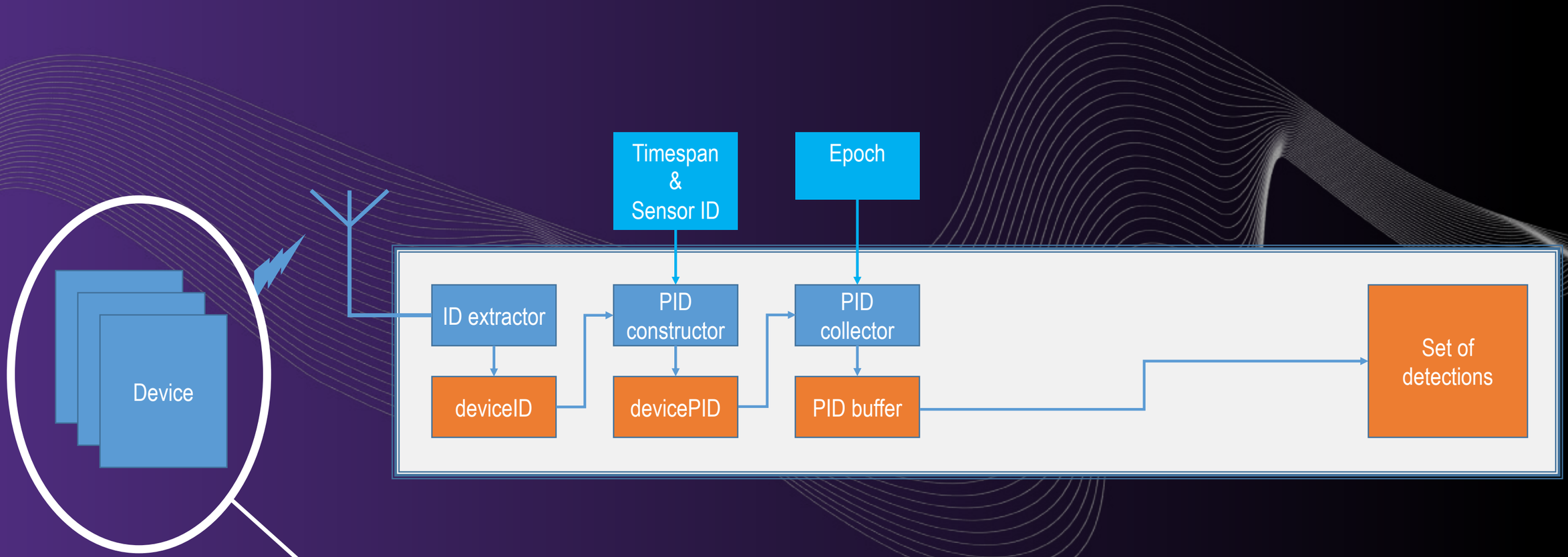


UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

TAKING A STEP BACK: A FRAMEWORK



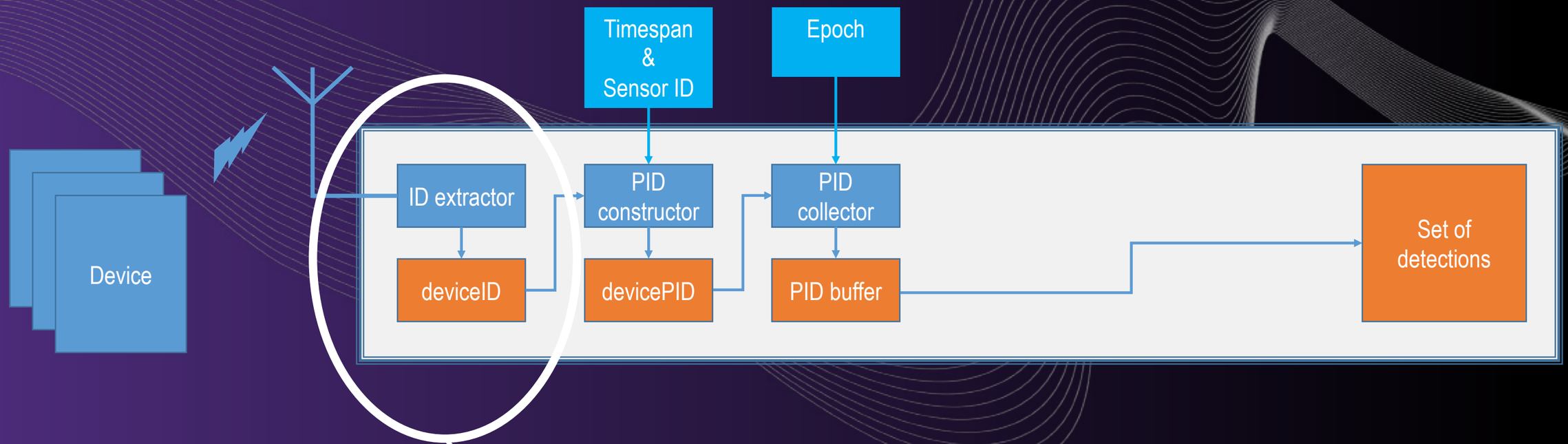
TAKING A STEP BACK: A FRAMEWORK



TYPICALLY WIFI-ENABLED SMARTPHONES AND OTHER CARRY-ON DEVICES

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

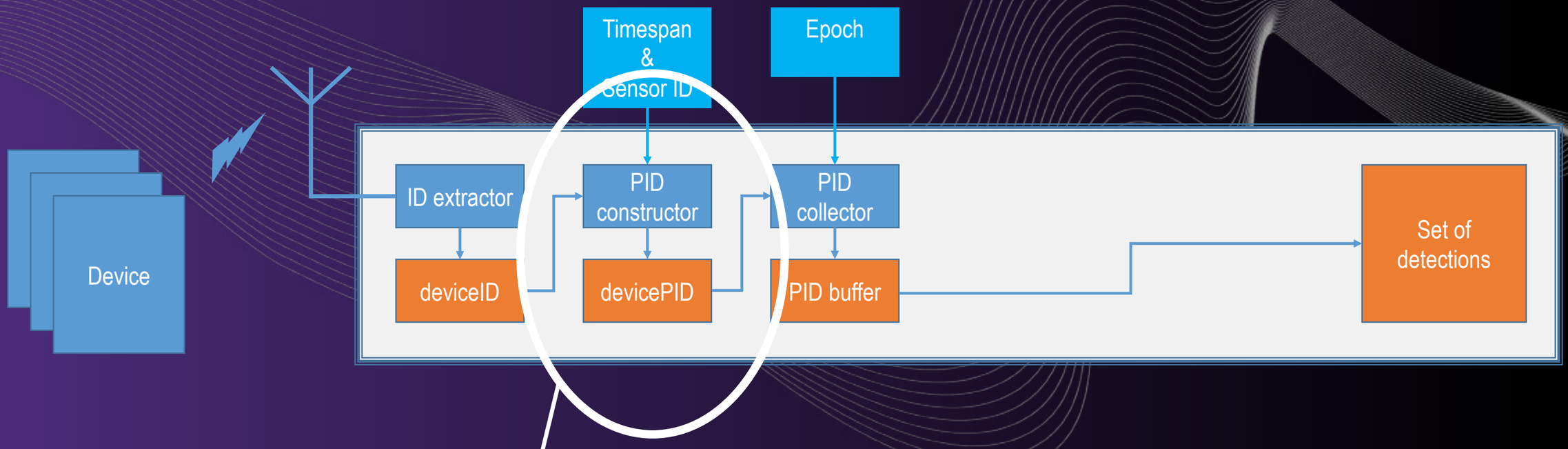
TAKING A STEP BACK: A FRAMEWORK



A WIFI INTERFACE THAT RECEIVES INCOMING FRAMES
AND EXTRACTS A DEVICE IDENTIFIER FROM THOSE FRAMES
(USUALLY A MAC ADDRESS)

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

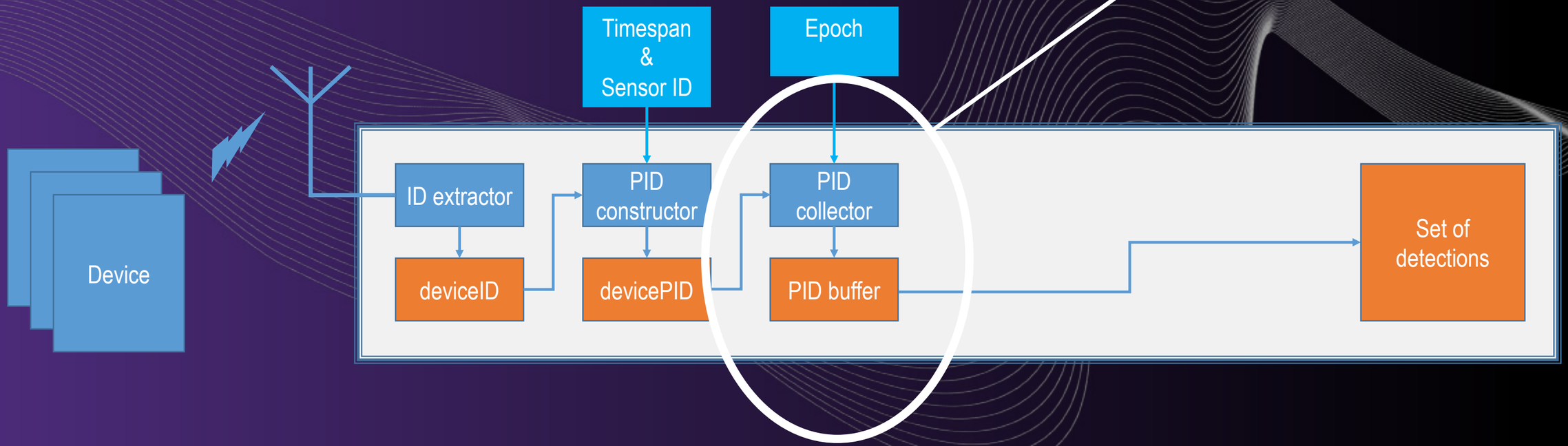
TAKING A STEP BACK: A FRAMEWORK



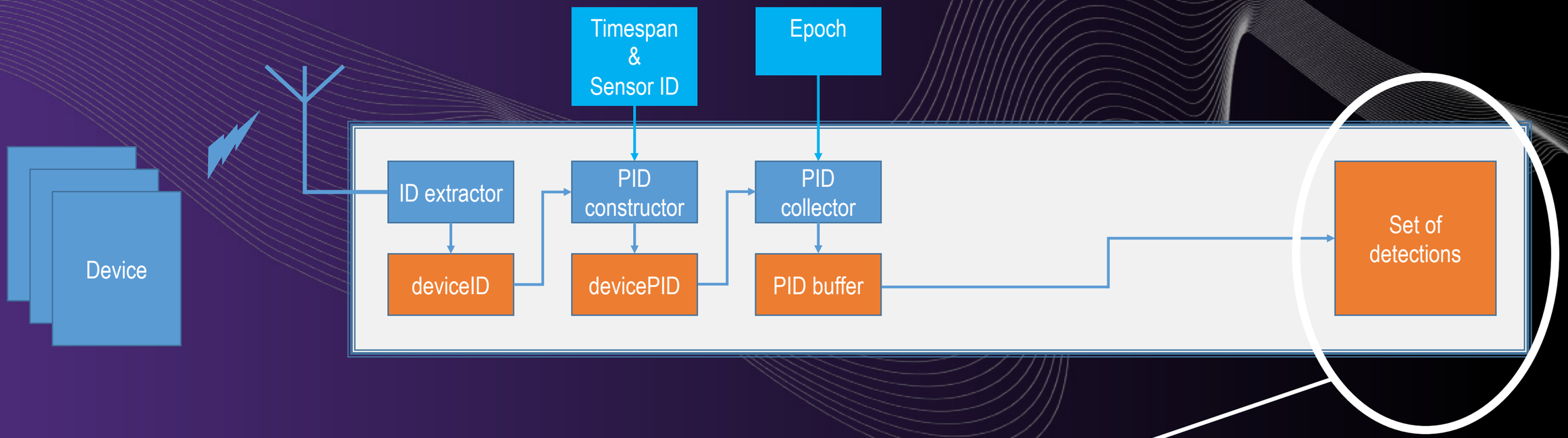
TYPICALLY AN ALGORITHM THAT TRANSFORMS A DEVICE ID TO AN IRREVERSIBLE PSEUDONYM (E.G. THROUGH HASHING)

TAKING A STEP BACK: A FRAMEWORK

MULTIPLE PSEUDONYMS ARE COLLECTED DURING A RELATIVELY SMALL EPOCH, REMOVING DUPLICATES (THE SYSTEM ACCUMULATES INCOMING PSEUDONYMS)

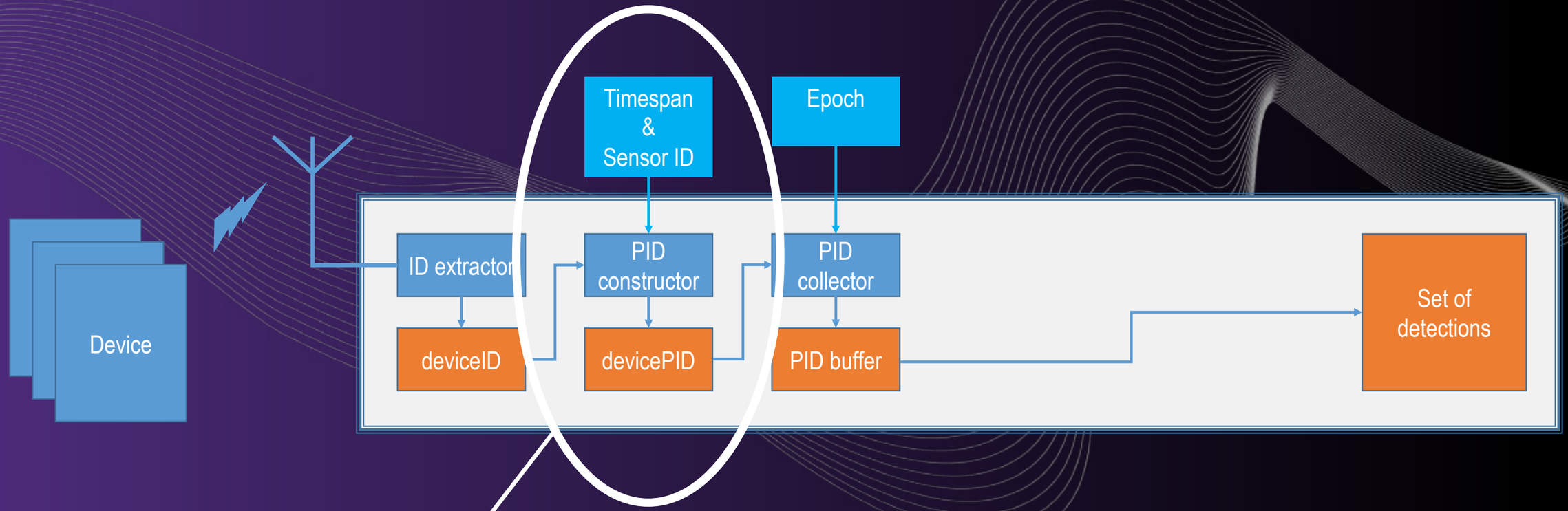


TAKING A STEP BACK: A FRAMEWORK



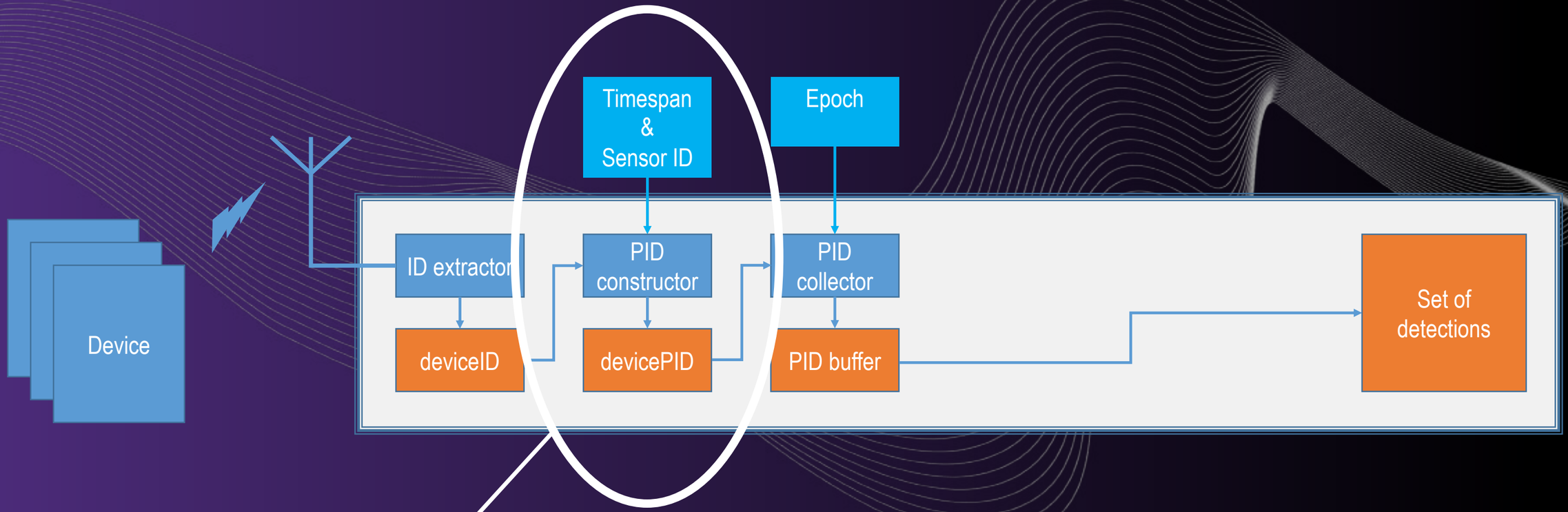
DETECTIONS OF THE FORM ⟨SENSOR, EPOCH, {PIDS}⟩

TAKING A STEP BACK: A FRAMEWORK



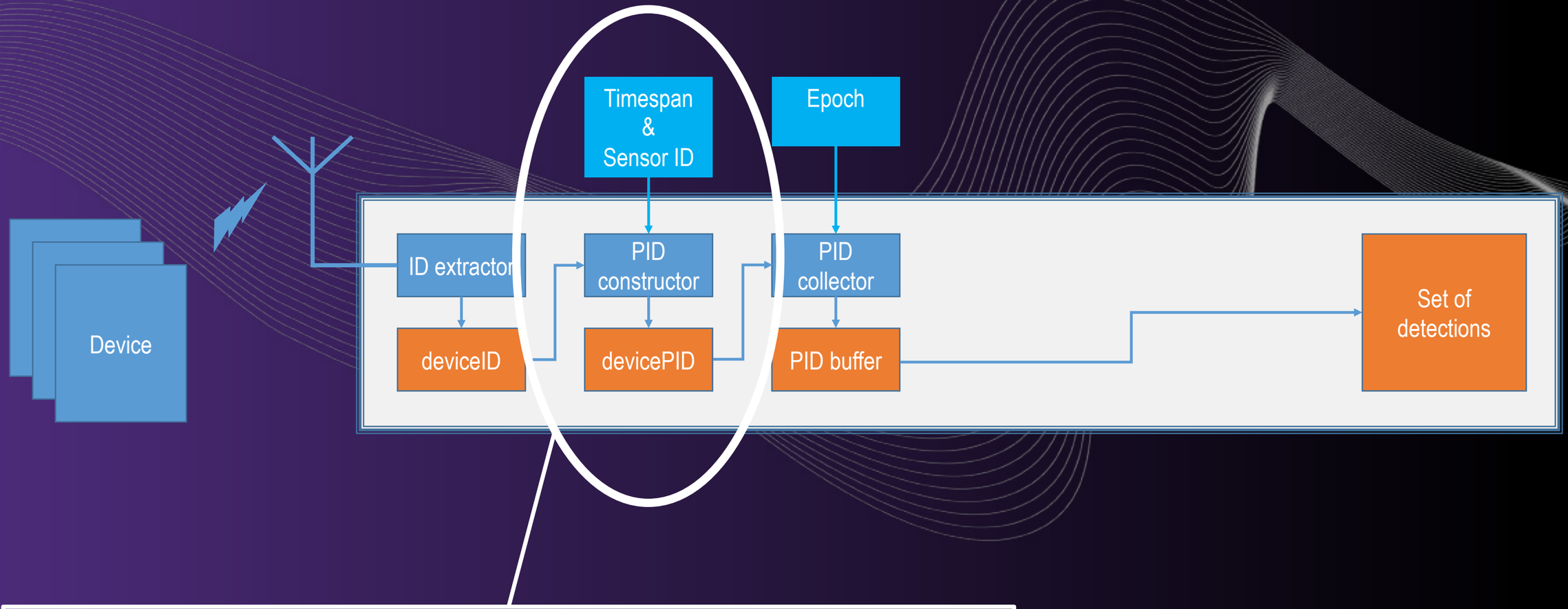
A PID MAY BE DEPENDENT ON TIME AND SENSOR

TAKING A STEP BACK: A FRAMEWORK



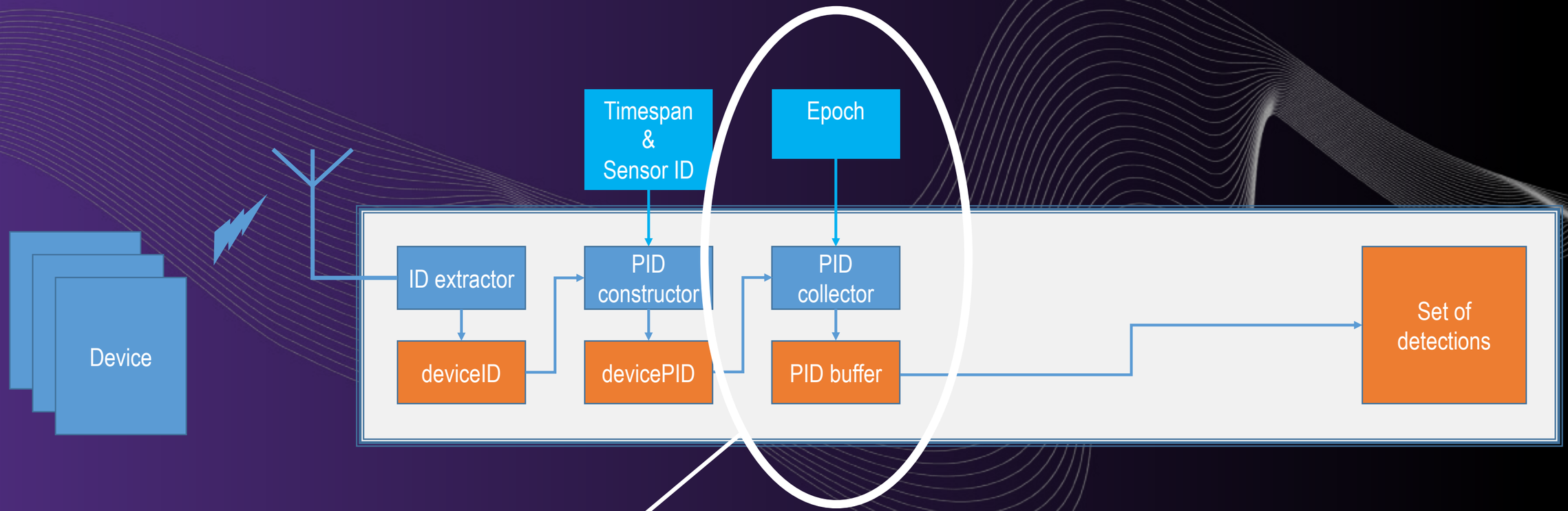
PER-SENSOR PID: IMPOSSIBLE TO TRACK MOVEMENTS BETWEEN DIFFERENT SENSORS

TAKING A STEP BACK: A FRAMEWORK



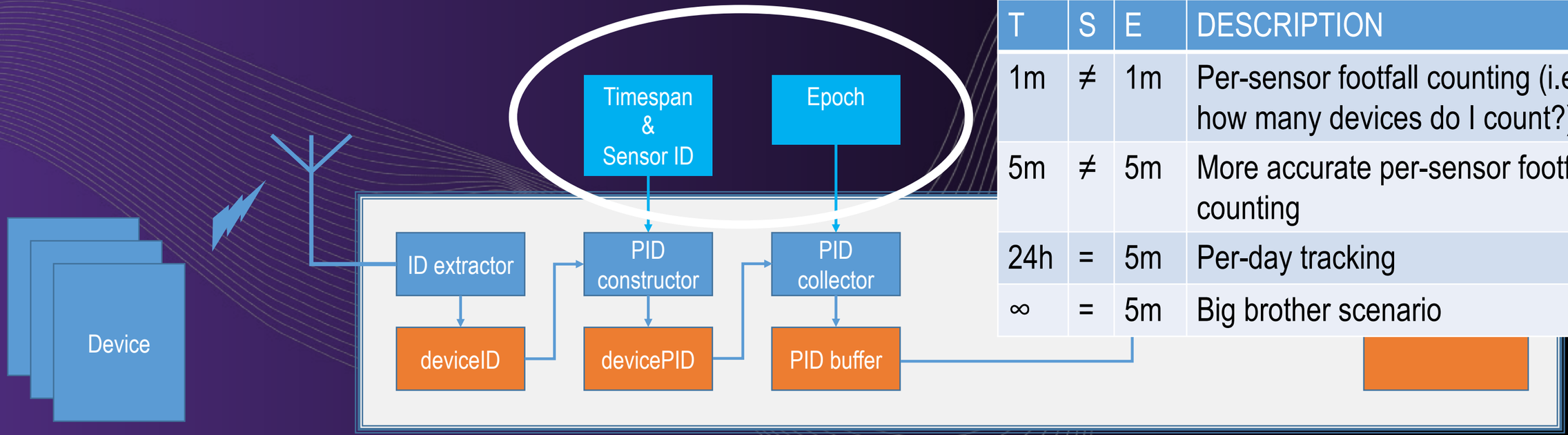
TIME-DEPENDENT PID: IMPOSSIBLE TO TRACK RECURRENT BEHAVIORS THAT SPAN OVER A SPECIFIC TIME (E.G. A DAY)

TAKING A STEP BACK: A FRAMEWORK

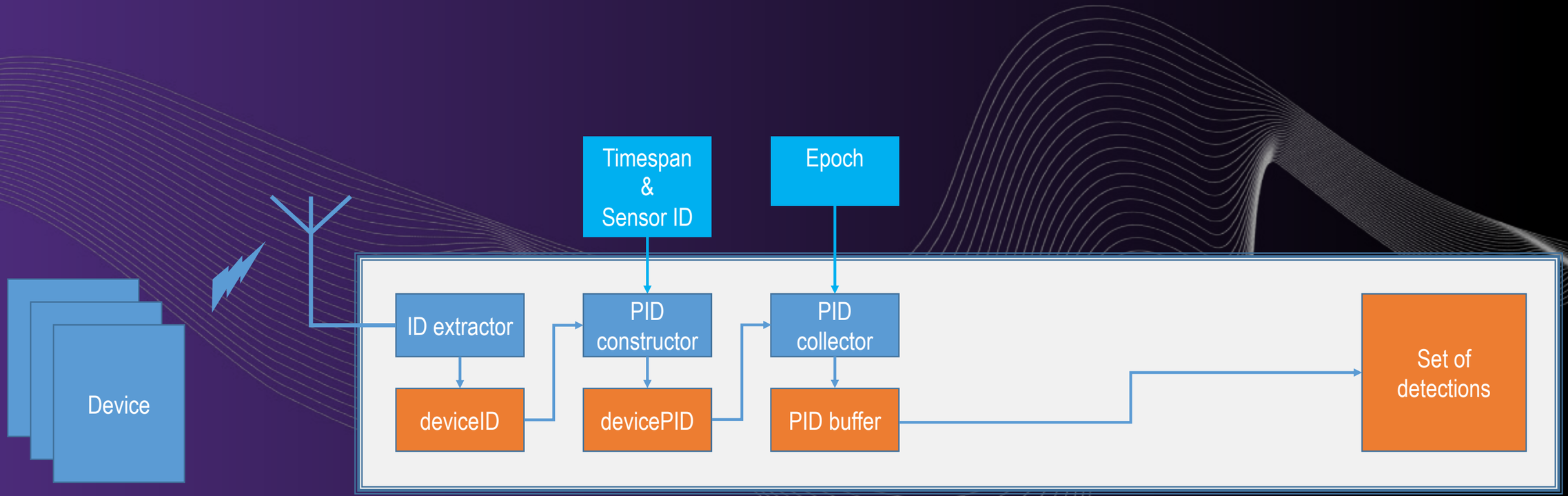


EPOCH DETERMINES THE ACCUMULATION PERIOD. VARIES BETWEEN A FEW MINUTES AND MAXIMUM CHOSEN TIMESPAN

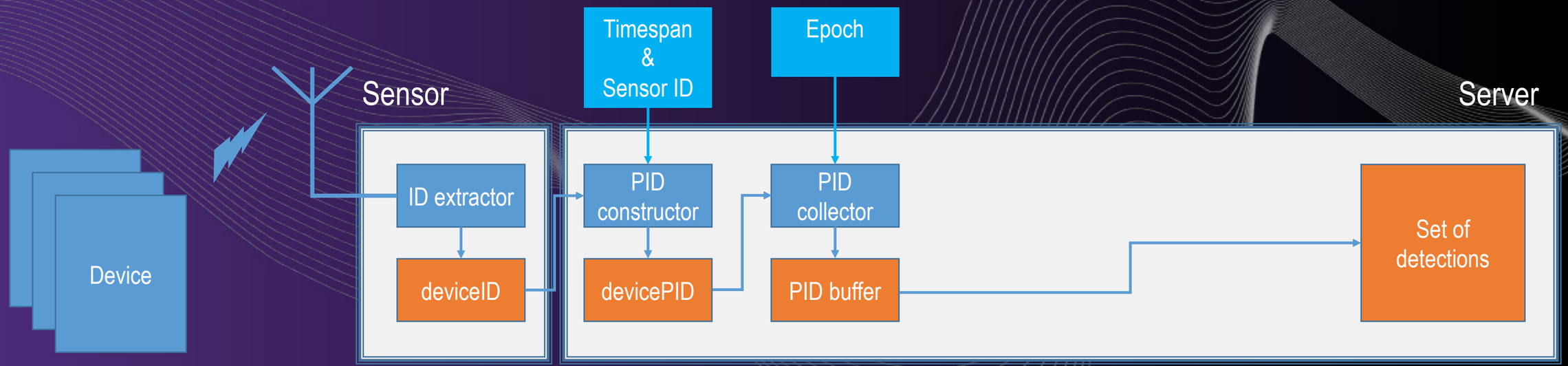
TAKING A STEP BACK: A FRAMEWORK

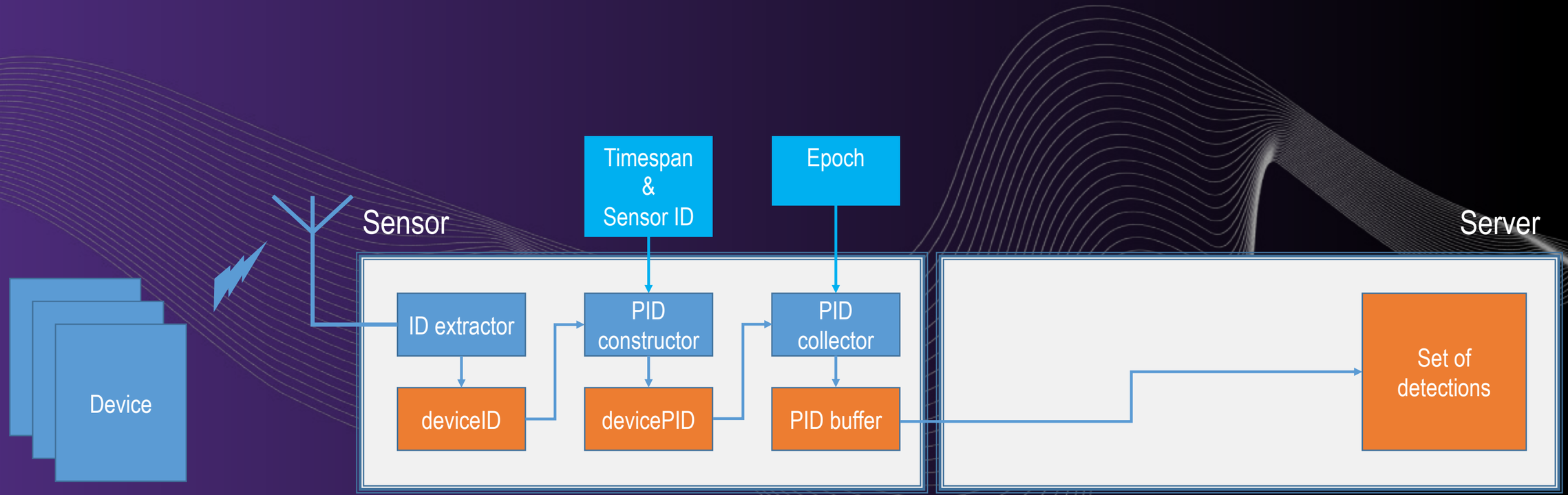


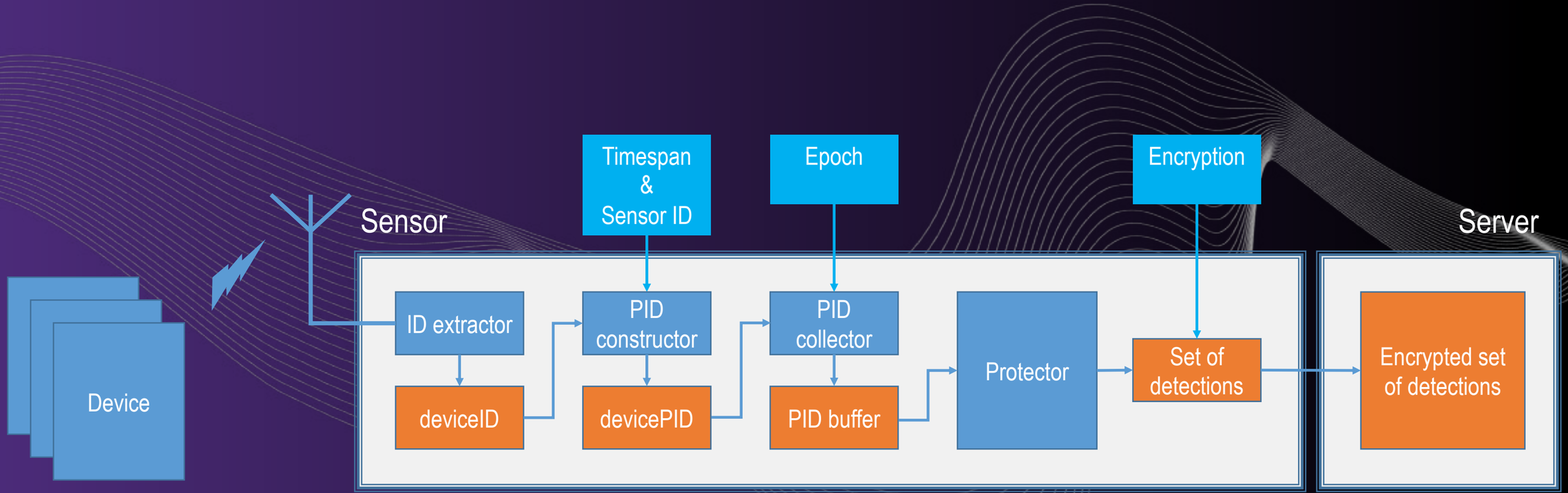
T	S	E	DESCRIPTION
1m	≠	1m	Per-sensor footfall counting (i.e., how many devices do I count?)
5m	≠	5m	More accurate per-sensor footfall counting
24h	=	5m	Per-day tracking
∞	=	5m	Big brother scenario

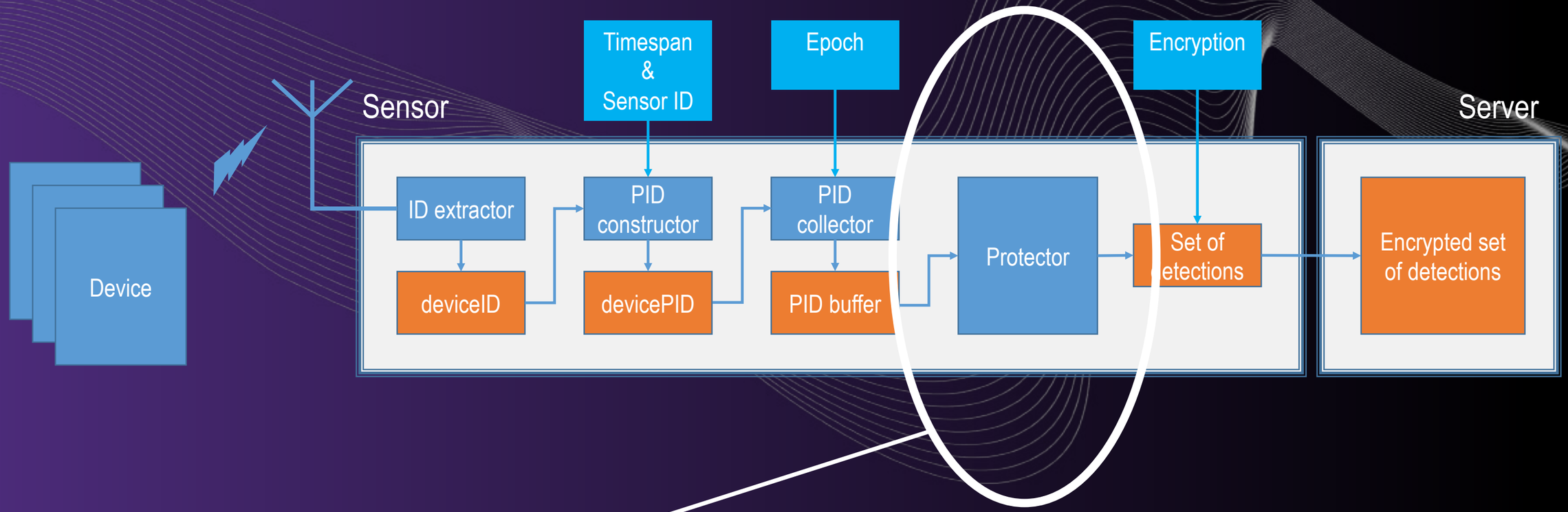


UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE


Timespan
&
Sensor ID
Epoch
Sensor
Server
Device
ID extractor
deviceID
PID
constructor
devicePID
PID
collector
PID buffer
Set of
detections

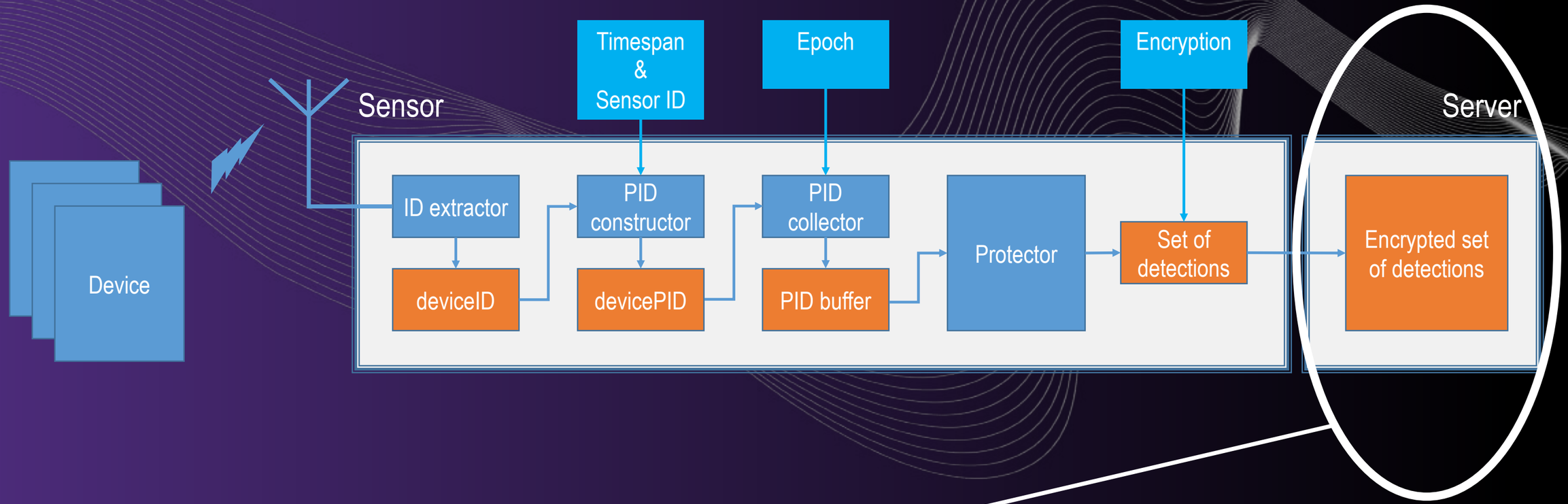

Timespan
&
Sensor ID
Epoch
Sensor
Server
Device
ID extractor
PID
constructor
PID
collector
deviceID
devicePID
PID buffer
Set of
detections


Timespan
&
Sensor ID
Epoch
Encryption
Sensor
Server
Device
ID extractor
deviceID
PID
constructor
devicePID
PID
collector
PID buffer
Protector
Set of
detections
Encrypted set
of detections



A DATA PROTECTOR, ENSURING THAT DATA ASSOCIATED WITH AN INDIVIDUAL CANNOT BE TRACED BACK TO THAT INDIVIDUAL

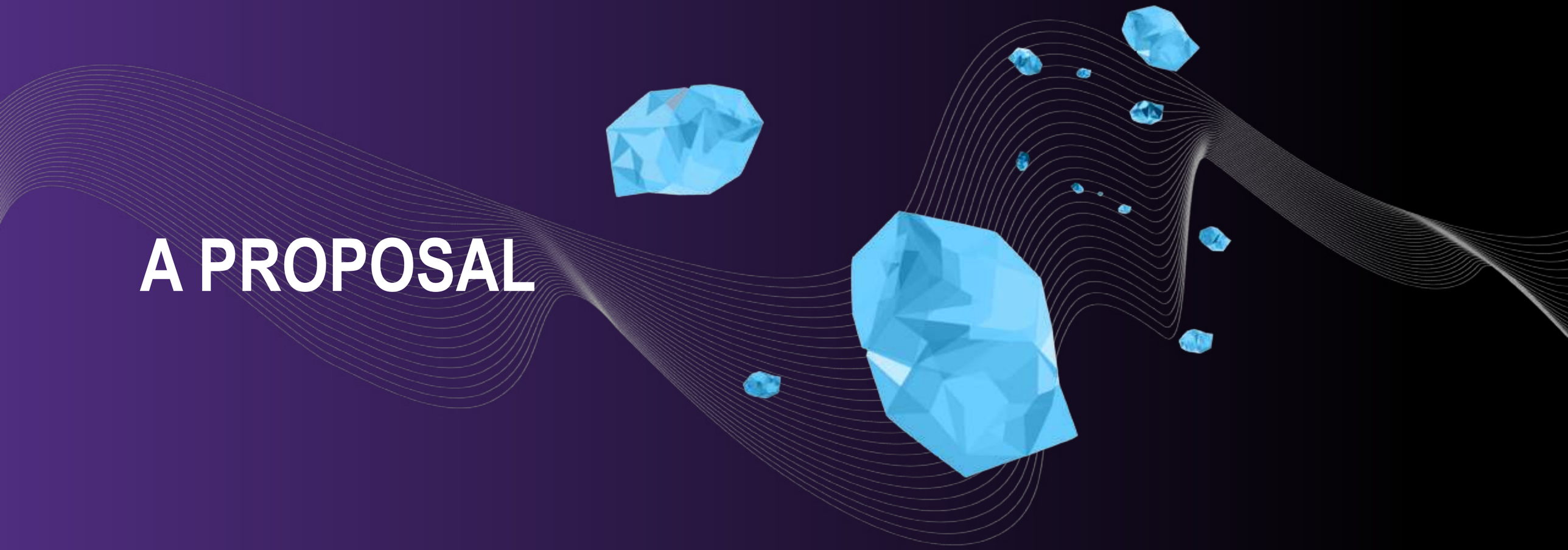
UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE



AN OBLIVIOUS SERVER: IF ATTACKED, DATA ALWAYS REMAINS PROTECTED

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

A PROPOSAL



UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

OBSERVATION

- The GDPR is (going to be) extended with rules that allow for **statistical counting**
- Collect signals, process them for counting purposes, and dismiss the data once the results have been established

PRINCIPLES:

- Data minimization
- Minimal trusted computing base

CONSEQUENCES (STRICT APPROACH):

- Data is collected only when it is known what to count
- Measured data are discarded asap
- Minimal sharing of data between sensors
- Server is minimized, if needed at all

BASIC QUERIES:

- How many devices detected by sensor S during epoch E?
- How many devices detected by sensor S1 during epoch E1 are detected by sensor S2 during epoch E2?

ASSUMPTION:

- Ranges of different sensors do not overlap

COMPOSITE QUERIES:

- How many devices detected by sensor S during timespan T?

- How many devices moved from sensor S1 to sensor S2 during timespan T?

IMPORTANT OBSERVATION:

- Many composite queries can be answered by taking the intersection of sets of detections

OBSERVATION:

- To count the number of devices all detected by several sensors, it suffices to compute only the size of intersections of sets of detected devices.
- No need to know the detected devices

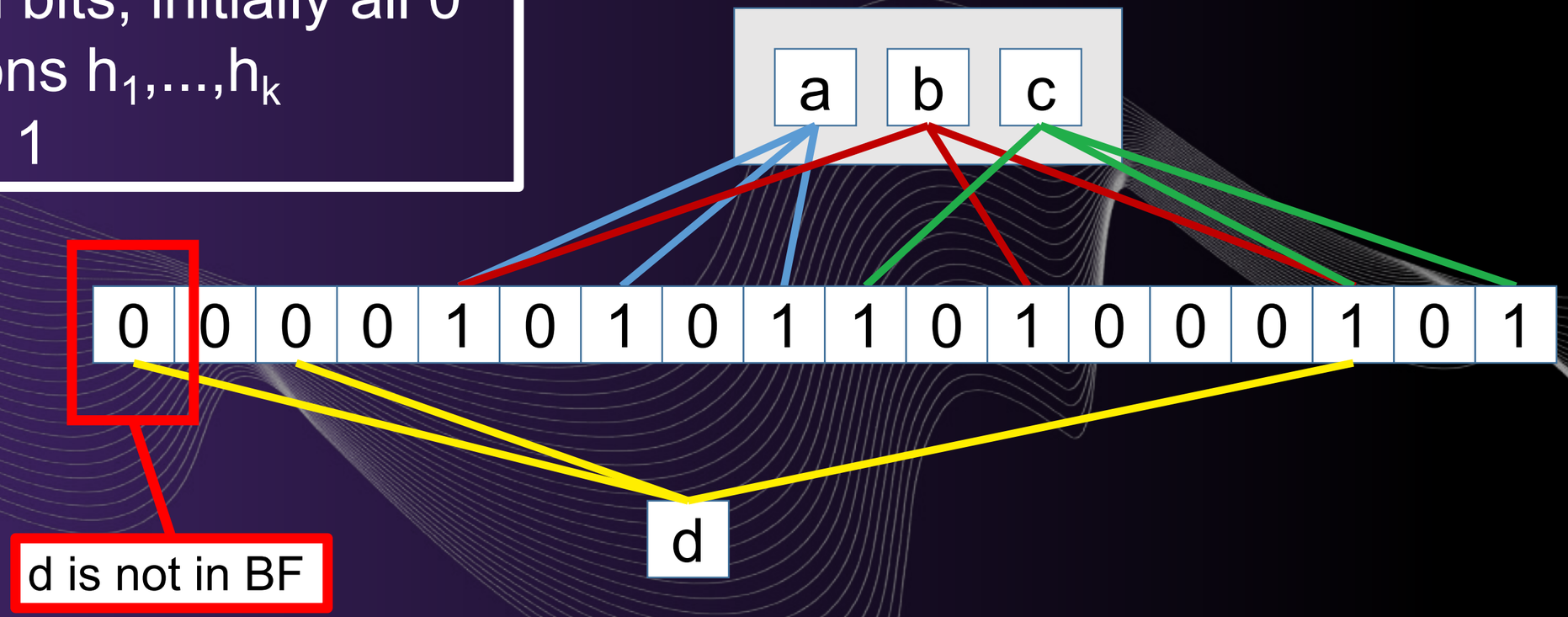
IMPORTANT OBSERVATION:

- Bloom filters are ideal for this purpose

BLOOM FILTERS

- Array BF of m bits, initially all 0
- k hash functions $h_1,...,h_k$
- $BF[\, h_i(a) \,] \leftarrow 1$

m = 18 bits
k = 3 hash functions
BF contains 3 elements



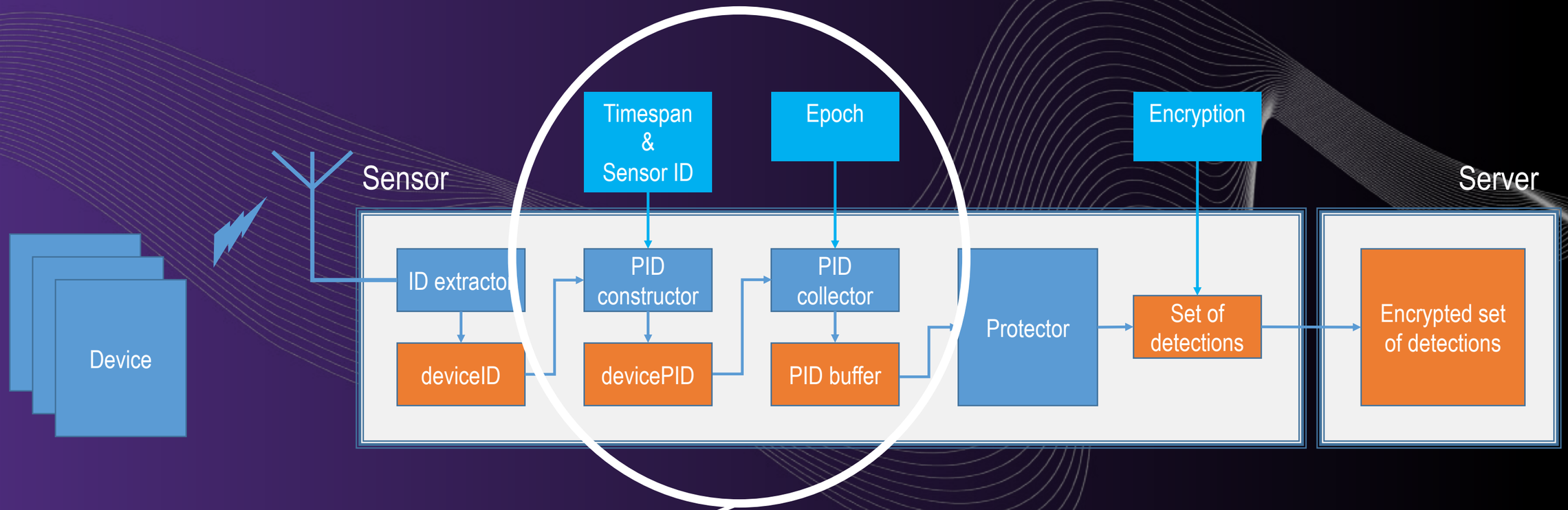
To know the elements in a BF, requires exhaustive membership testing

BLOOM FILTERS: COMPUTE INTERSECTION

BLOOM FILTERS: ESTIMATE SIZE

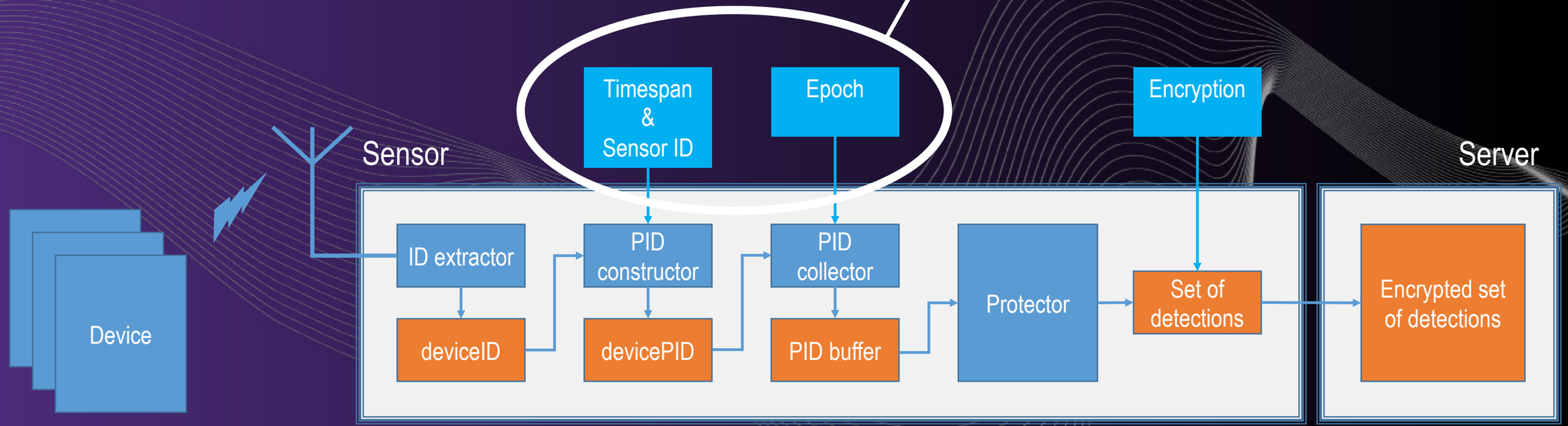
$$n^* = -\frac{m}{k} \ln\left[1 - \frac{X}{m}\right]$$

3 (3)	0	0	0	0	1	0	1	0	1	1	0	1	0	0	0	1	0	1
5 (5)	0	1	0	0	1	1	0	1	0	1	0	1	0	0	1	1	1	1
2 (2)	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	1

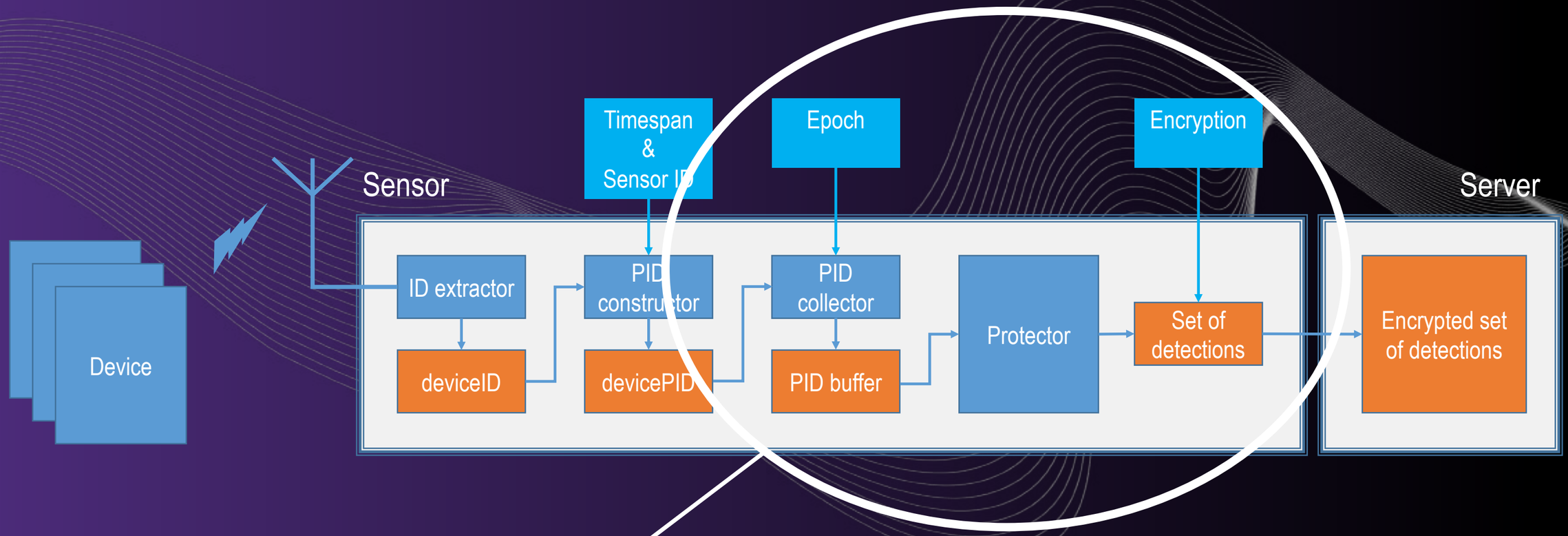


EACH SENSOR SIMPLY COLLECTS MAC ADDRESSES, HASHES THEM AND ADDS THESE TO A LOCAL BUFFER.

ALL SENSORS USE THE SAME (LARGE) TIMESPAN AND THE SAME HASH FUNCTION, AND SAME EPOCH LENGTH



UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE



COLLECTED PSEUDONYMS ARE ADDED TO A LOCAL BLOOM FILTER, OF WHICH EACH ENTRY IS ENCRYPTED WITH A GIVEN PUBLIC KEY

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

HOMOMORPHIC ENCRYPTION:

- We homomorphically encrypt each entry of a Bloom filter
- Bitwise multiplication is unaffected:
  [0] * [0] = [0]
  [0] * [1] = [0]
  [1] * [0] = [0]
  [1] * [1] = [1]

- [x] = encrypted entry

- Encrypted Bloom filters are stored at a server

ABOUT ENCRYPTION KEYS:

- We assume there is an external consumer interested in the statistical counting of pedestrians
- The consumer provides a public key, and keeps the associated private key to itself
- The server is assumed to
  - compute intersections (on encrypted data)
  - shuffle the entries of an intersection before handing it to the consumer
- The consumer knows m, k, and can compute X (through decryption) and can thus estimate the size of the intersection

ABOUT THE SERVER:

- Sees only encrypted Bloom filters, which it cannot decrypt
- Is required to compute intersections and shuffle entries
- Is assumed not to collude with a consumer

IF NECESSARY:

- Let the server only store encrypted Bloom filters
- Let sensors compute intersections (and store at the server)
- Let a specific sensor shuffle before handing over to consumer

OBSERVATIONS

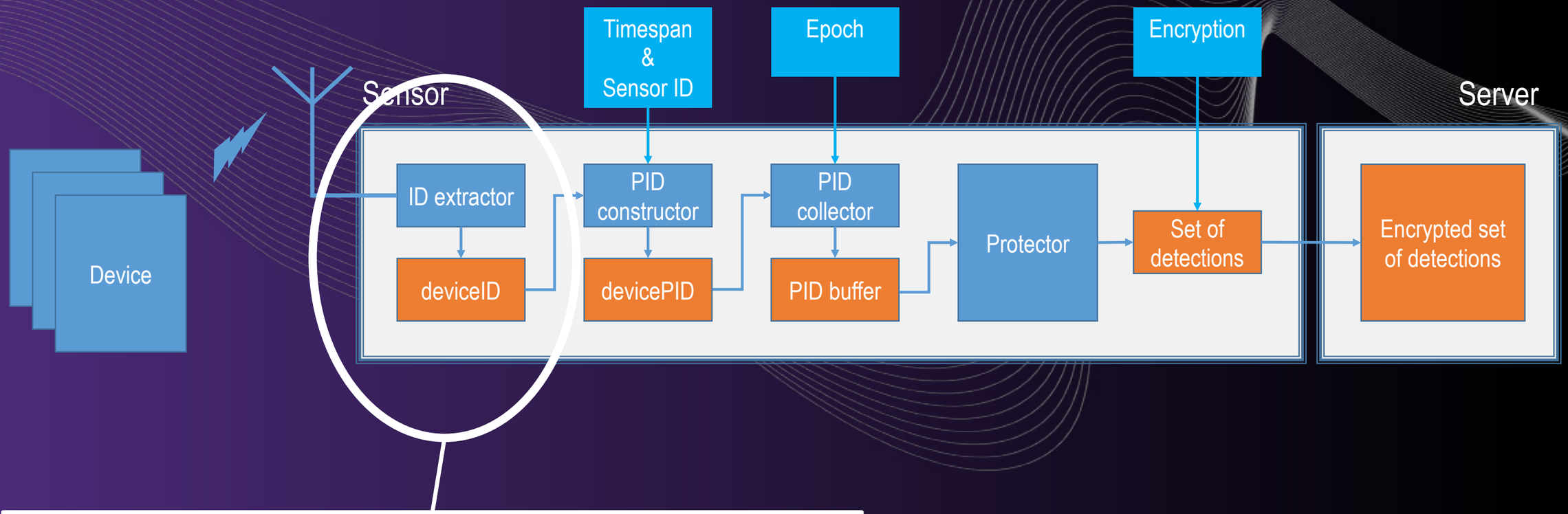
UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

OBSERVATION:

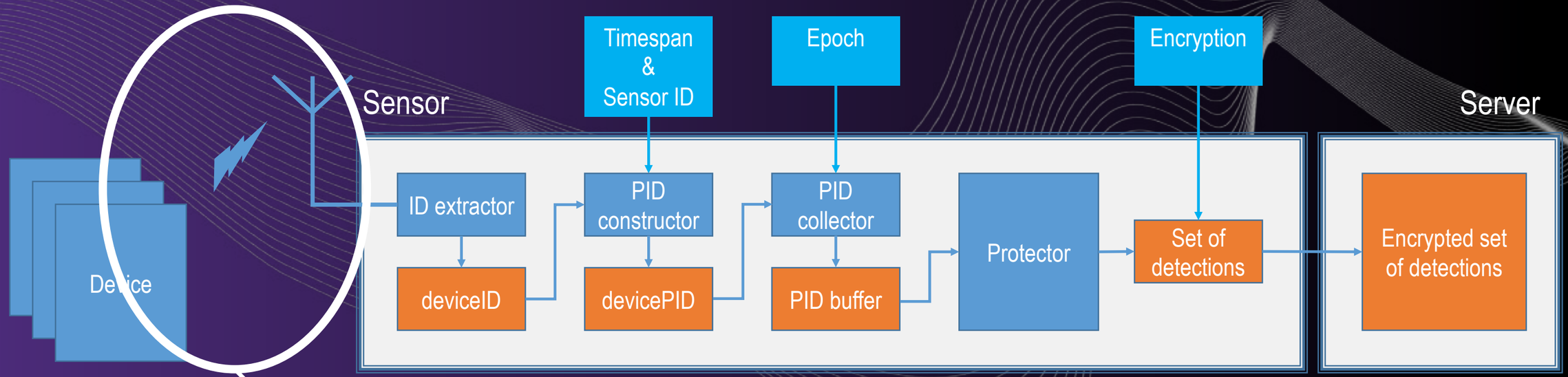
- Sensor nodes need to be trusted, the server only when it comes to shuffling, unless sensor does shuffling.
- The only information that is revealed are statistical counts.
- Complexity is dictated by composite queries. Simple queries (with only single epochs) are computationally easy.
- Theoretical accuracy is dictated by probabilistic properties of Bloom filters
- Practical accuracy by ability to sample wireless network packets: devices are known to behave very differently.
- We count devices, which is not the same as people: correction will always be needed.

OBSERVATION:

- Sensor nodes need to be trusted, the server only when it comes to shuffling, unless sensor does shuffling.
- The only information that is revealed are statistical counts.
- Complexity is dictated by composite queries. Simple queries (with only single epochs) are computationally easy.
- Theoretical accuracy is dictated by probabilistic properties of Bloom filters
- **Practical accuracy by ability to sample wireless network packets: devices are known to behave very differently.**
- We count devices, which is not the same as people: correction will always be needed.



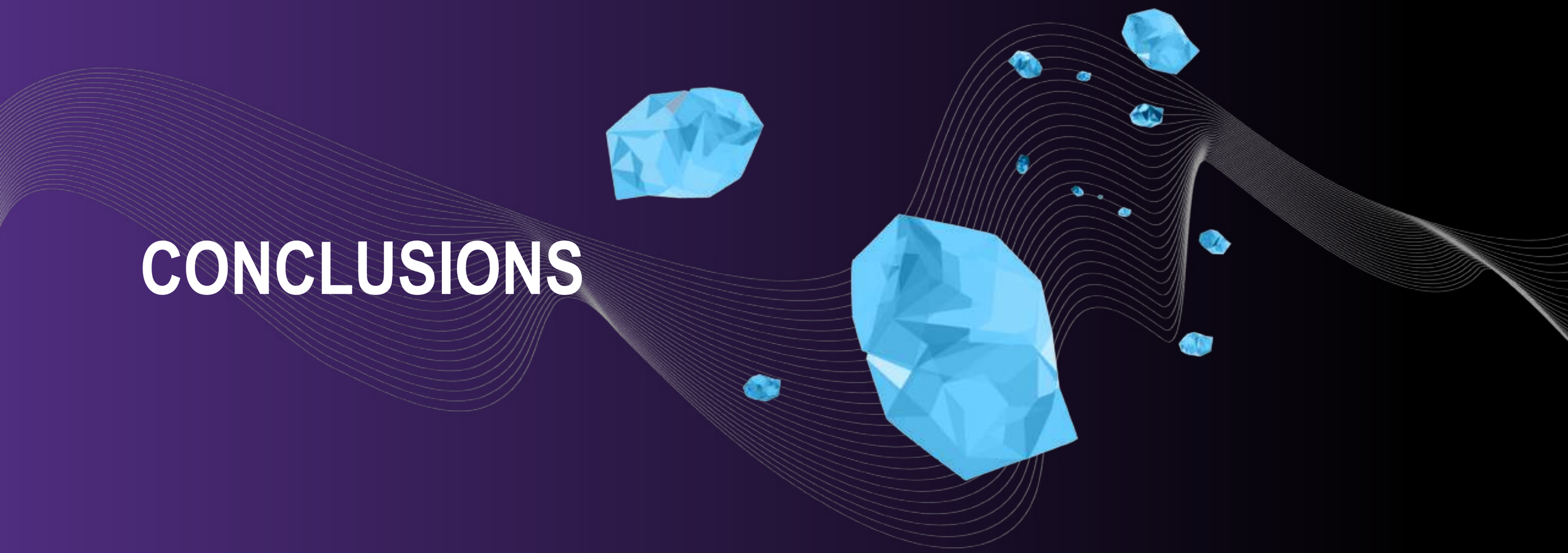
RANDOMIZATION IS GETTING IN OUR WAY:
WE NEED TO DEVELOP ALTERNATIVE TECHNIQUES



WIRELESS COMMUNICATION IS SUBJECT TO MANY DISTURBANCES AND INTERFERENCES

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

CONCLUSIONS



UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

- Purposefully design systems for data minimization:
  - Minimize Trusted Computing Base
  - Minimize needed functionality of the cloud

- Privacy because of the edge?
  - A solution that guarantees privacy only because of the edge should be distrusted: privacy is location-independent
  - Data protection is not the same as privacy protection

Thank you!