

Data Processing

Técnicas de Perceção de Redes

**Mestrado Integrado em
Engenharia de Computadores e Telemática
DETI-UA**

Qualitative Data

- Most monitored data is qualitative.
 - An event (with description) at a specific time (with a time-stamp).
 - ➔ 00:01:23.4566 – IP Packet [from A to B with 64 bytes]
 - ➔ 21:04:23.4566 – Error [id 404]
 - ➔ ...
- Must be converted to quantitative data.
- Some is pre-processed and it is already presented as quantitative.
 - Packets sent: 5467.
 - Bytes seen in the last 10 minutes: 18471947.
 - May require some additional processing.
 - ➔ Packets sent at 1s: 300pkts, Packets sent at 2s: 350pkts → Packets sent between 1s-2s: $350 - 300 = 50$ pkts.

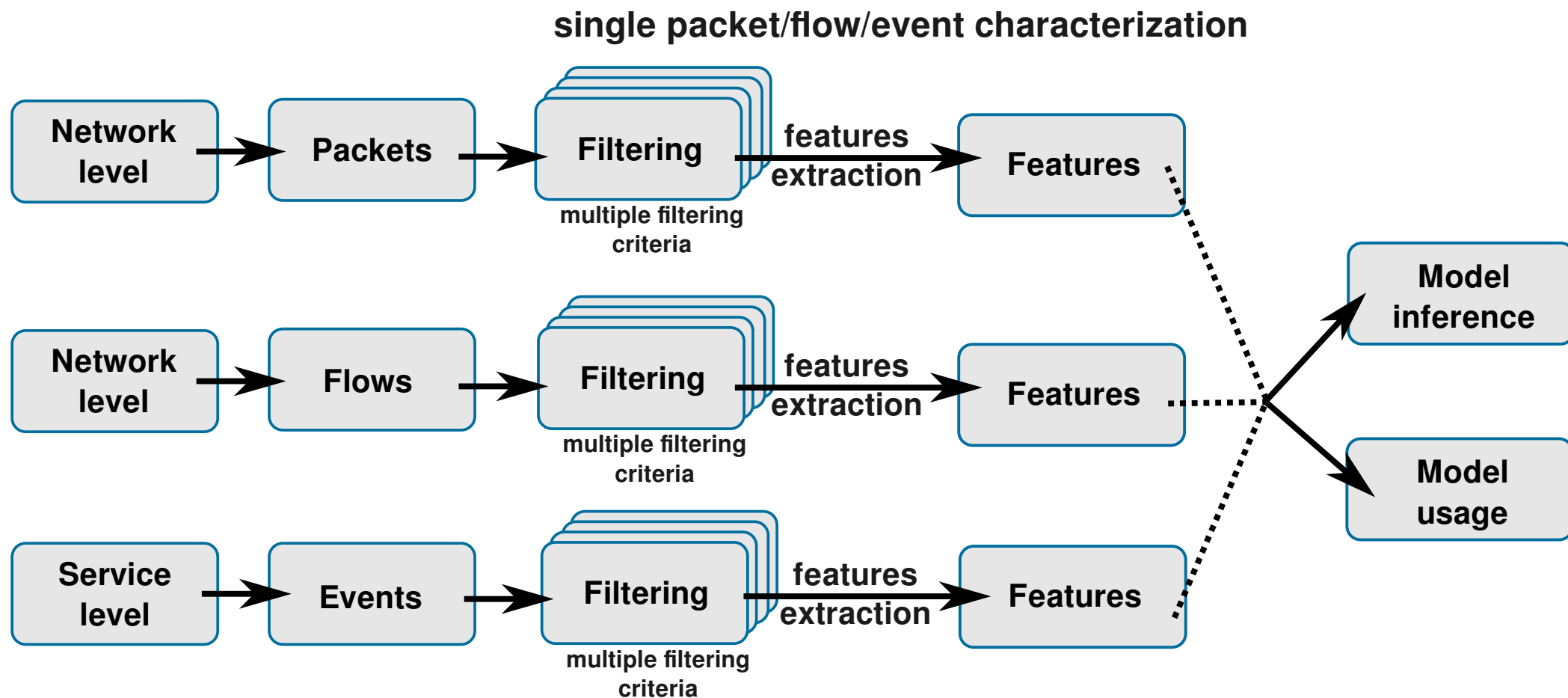


Modeling Single Actions vs. Behaviors

- Models require numerical/quantitative data.
 - ◆ Inputs for any descriptive model are usually called features.
 - ◆ Features should describe the modeling object.
 - ◆ Excluding NLM models.
- Models may describe single events:
 - ◆ One packet sent/received, one conversation (flow), one service request...
 - ◆ For network awareness the information of a single event is not enough to describe and detect advanced anomalies. Specially stealth anomalies.
 - Allow the detection of port scans, unusual service requests, etc...
- Any complex interaction generates multiple events.
- Behavior models describe the characteristics of a set of complex interactions over time.
 - ◆ Requires the aggregation of single events data over time.
 - ◆ Models are constructed based on historic data.
 - Using a set of multiple observations (time windows).
 - ◆ This models should allow to test new data and make decisions periodically.



Single Event Model Features

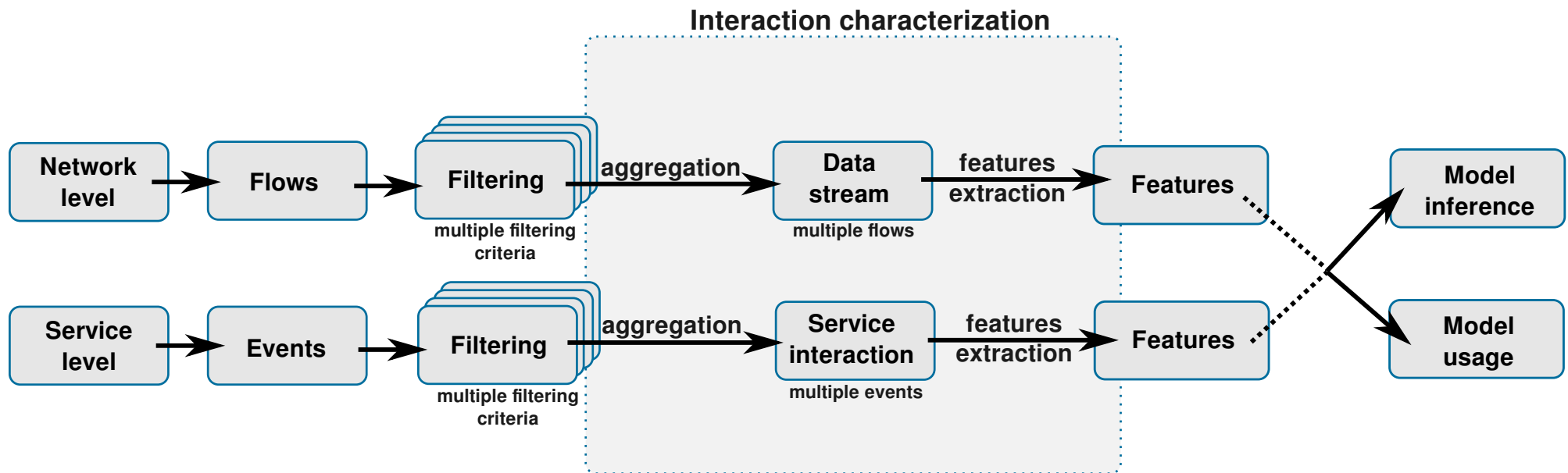


Data Filtering and Aggregation (1)

- Filtering is the method by the raw data is selected according to multiple criteria.
 - Some data may be discarded.
 - Data may be divided to created differentiated and relevant features.
 - ➔ E.g., download/upload, to TCP port 443, 80 or 22, with destination/source in different IP networks, etc...
- Aggregation is the first step to create a behavior model.
 - Any network entity (human or non-human) interaction results in multiple observable events.
 - ➔ Multiple packets, multiple flows, multiple service requests, etc...
 - A set of aggregated network flows is usually called datastream.
 - ➔ A flow is described by a 5-tuple (transport protocol and source IP addresses and ports) .
 - ➔ A datastream is any aggregation of flow using a tuple smaller than 5.
 - Ex1 - all traffic from a specific terminal (1-tuple: source address);
 - Ex2 - all traffic from a specific terminal to port TCP 443 (3-tuple: TCP transport, source address, destination port);
 - Ex2 - all traffic from a specific terminal using port TCP 443 to a specific service (4-tuple: TCP transport, source address, destination port, destination server addresses);
 - A set of service events is called a service interaction.

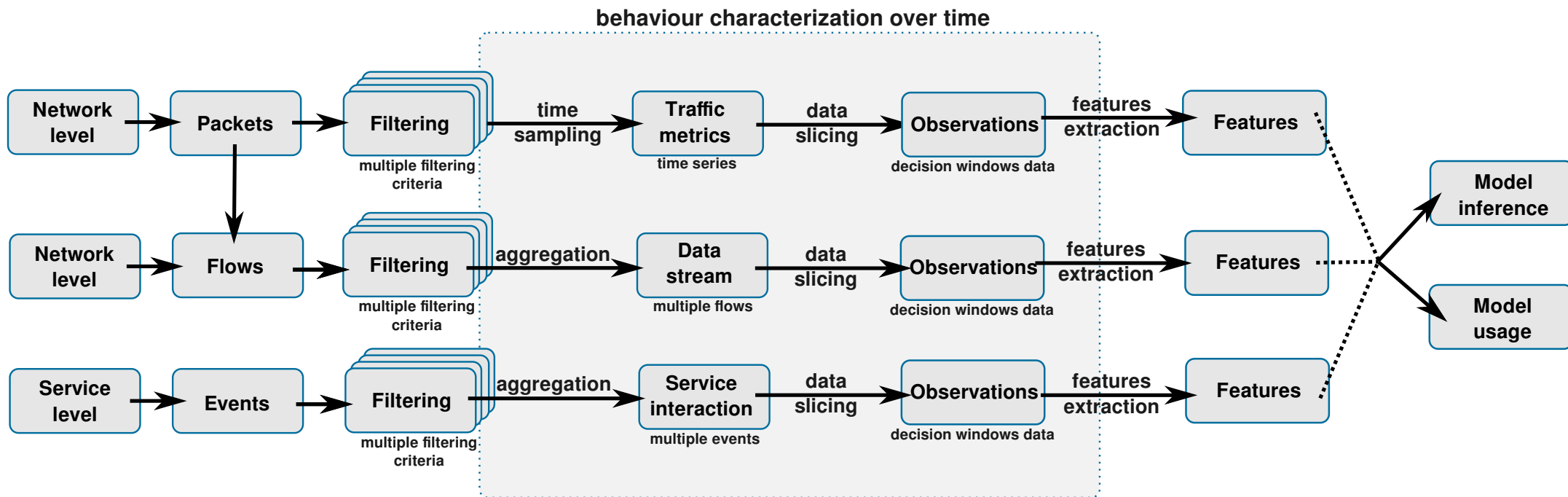


Data Filtering and Aggregation (2)



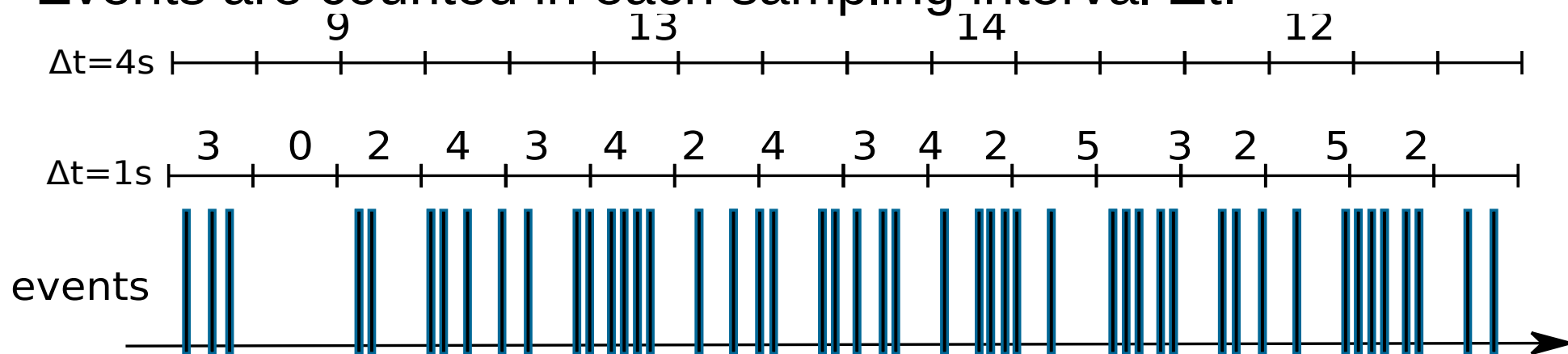
Behavior Models

- Any behavior should be **described over time**.
 - Data must be sliced over time, to create sub-sets of data that will allow to describe complex interactions within a time window, over time.
 - Behavior may change from observation window to observation window.
 - Model should be created from a set of historic observation windows.
 - Model is applied to any new data (new observation window data).
- Raw packet data is qualitative and must be sampled (a simplified aggregation process).



Data Sampling (1)

- Sampling transforms Qualitative into Quantitative data.
- Events must be defined, identified and grouped:
 - All packets from IP 10.0.0.1,
 - All 400 errors accessing site X, etc...
- Sampling/Counting Interval
 - Time window in each the number of a specific event is counted, associated with a time index, and stored.
 - Minimum timescale.
- Events are counted in each sampling interval Δt .



Data Sampling (2)

- Results in discrete time sequences for event:

- For $\Delta t=1$: $X_k=\{3,0,2,4,3,4,2,4,3,4,2,5,3,2,5,2\}$

- $X_0=3, X_1=0, \dots, X_{12}=2$

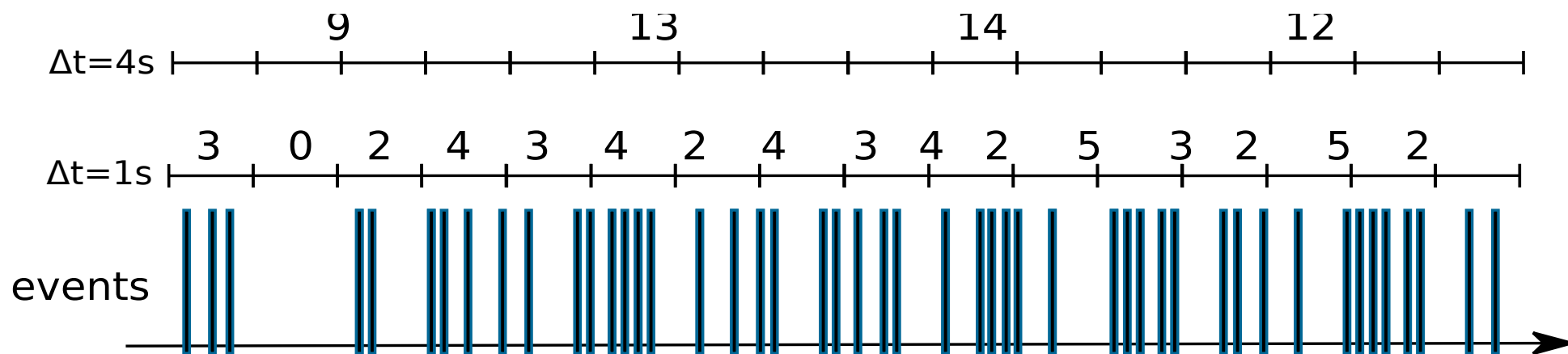
- For $\Delta t=4$: $Y_k=\{9,13,14,12\}$

- Time sequences may be multi-dimensional:

- Time sequences of n-tuples.

- e.g., Number of packets, upload e download.

- $Z_k=\{(3,9),(0,45),\dots(67,90)\}$



Time Windows and Entity Behavior Profile

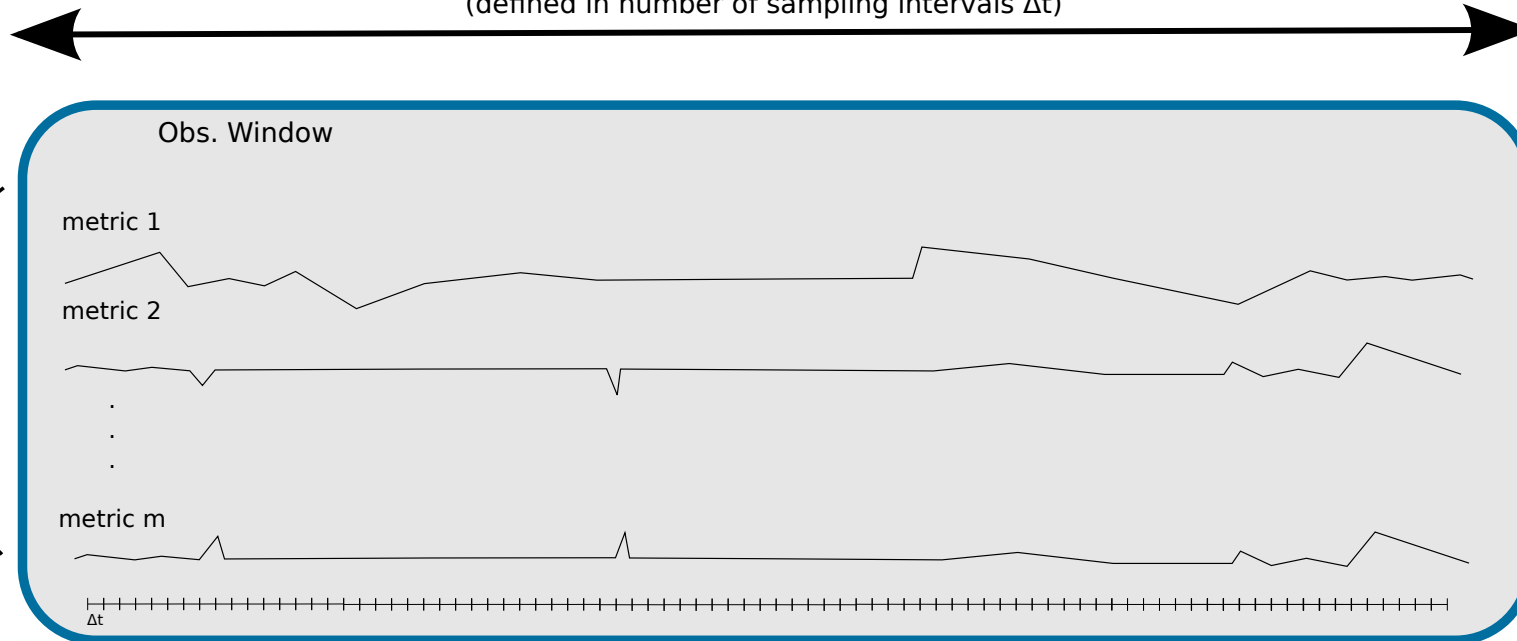
- Sampling/Counting Window.
 - Provides time series of multiple metrics.
 - e.g., number of packets received by a terminal each second.
- Observation Window.
 - Features/Characteristics extraction Window.
 - Uses multiple Sampling/Counting Windows,
 - Statistics of respective time series.
 - Provides a n-tuple characterizing an entity behavior at a specif time.
 - e.g., 2-tuple with mean and variance of the number of packets received by a terminal in 30 seconds (30 counting 1s windows).
- Entity Behavior Profile
 - Pattern from multiple Observation Windows.
 - Provides a model to classify entities and detect anomalies.
 - May include time dynamics over time.



Observation Window (1)

- An observation is constructed based on multiple sampling/counting metrics.
- Sampling/counting metrics should quantify activity events:
 - ◆ Start/End of activity.
 - Traffic Flows, Calls, Service usage, etc...
 - ◆ Amount of activity.
 - Traffic per sampling interval, activity duration, actions per sampling interval, etc...
 - ◆ Activity targets
 - IP addresses contacted, UCP/TCP ports used, services user IDs, points of access, etc...

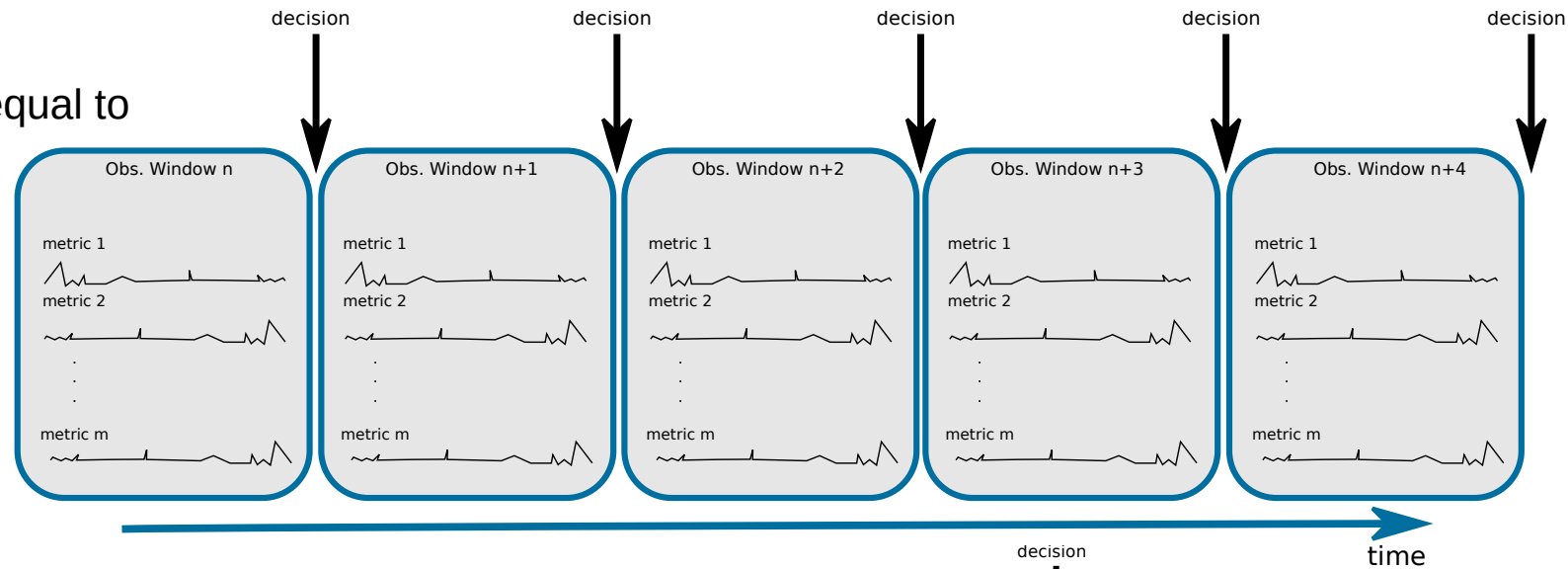
Fixed Length
(defined in number of sampling intervals Δt)



Observation Window (2)

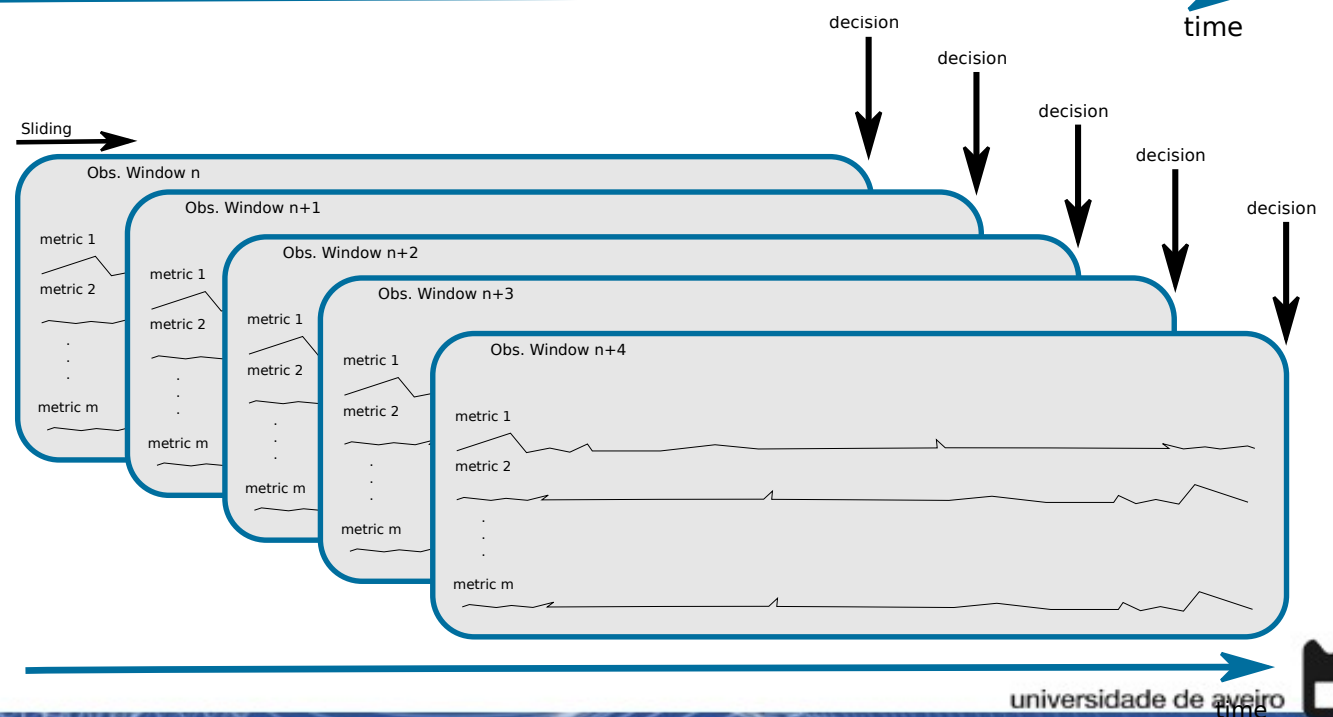
- Sequential

- Decision interval is equal to window size.



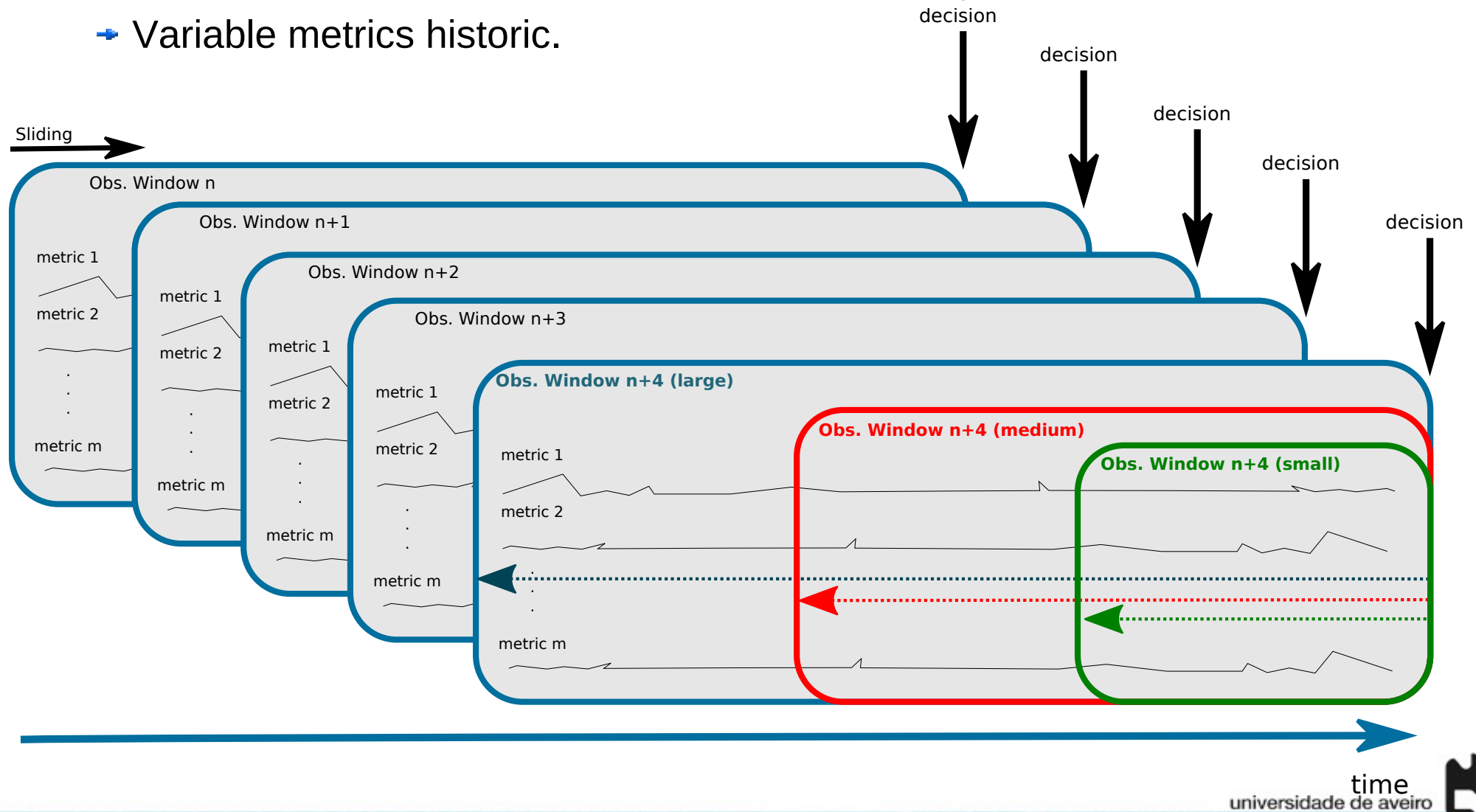
- Sliding

- Allows for longer periods of observation, while maintaining a short period of decision.



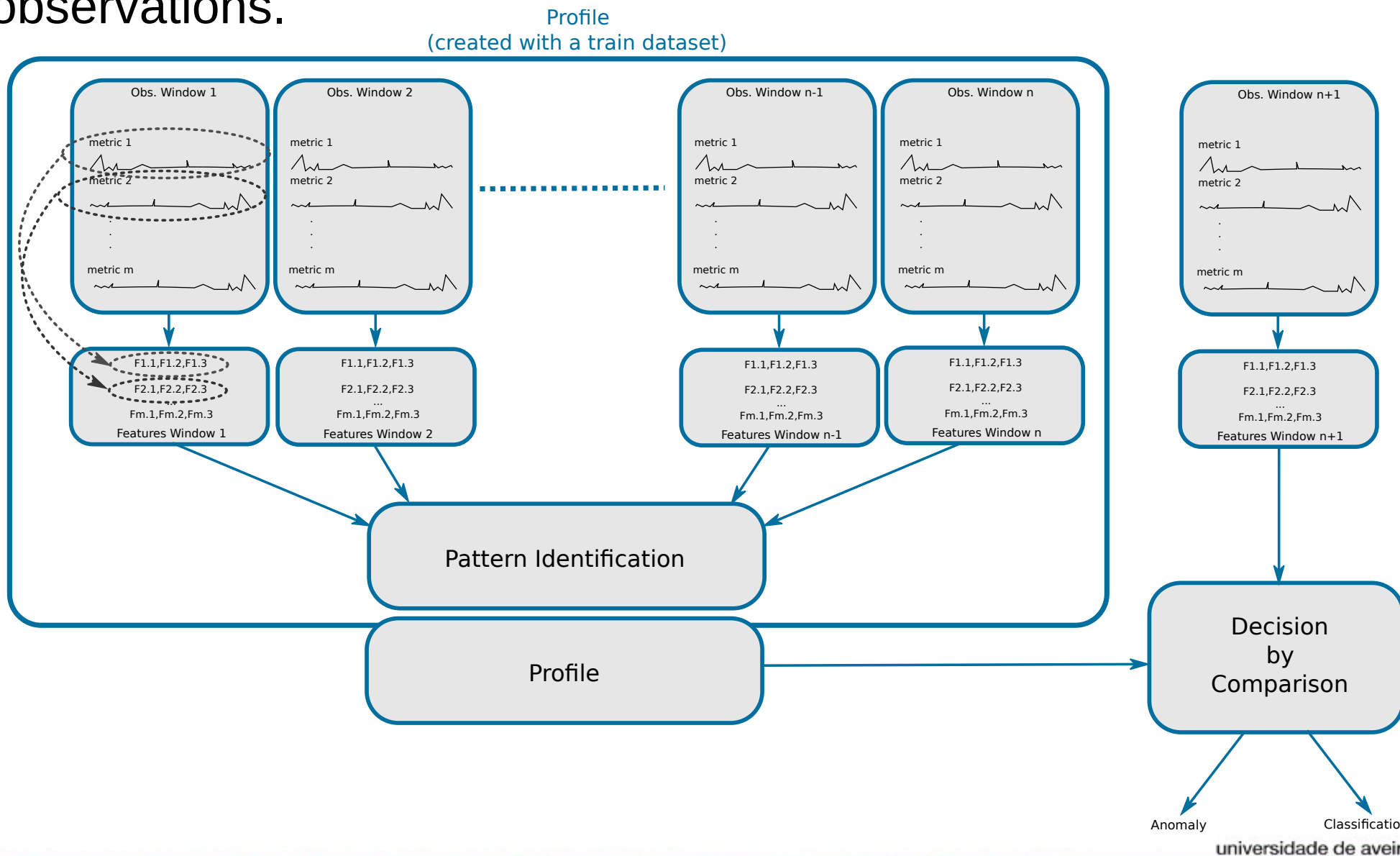
Multiple Observation Windows

- At each decision time point.
 - Construct observation windows with different lengths.
 - Variable metrics historic.



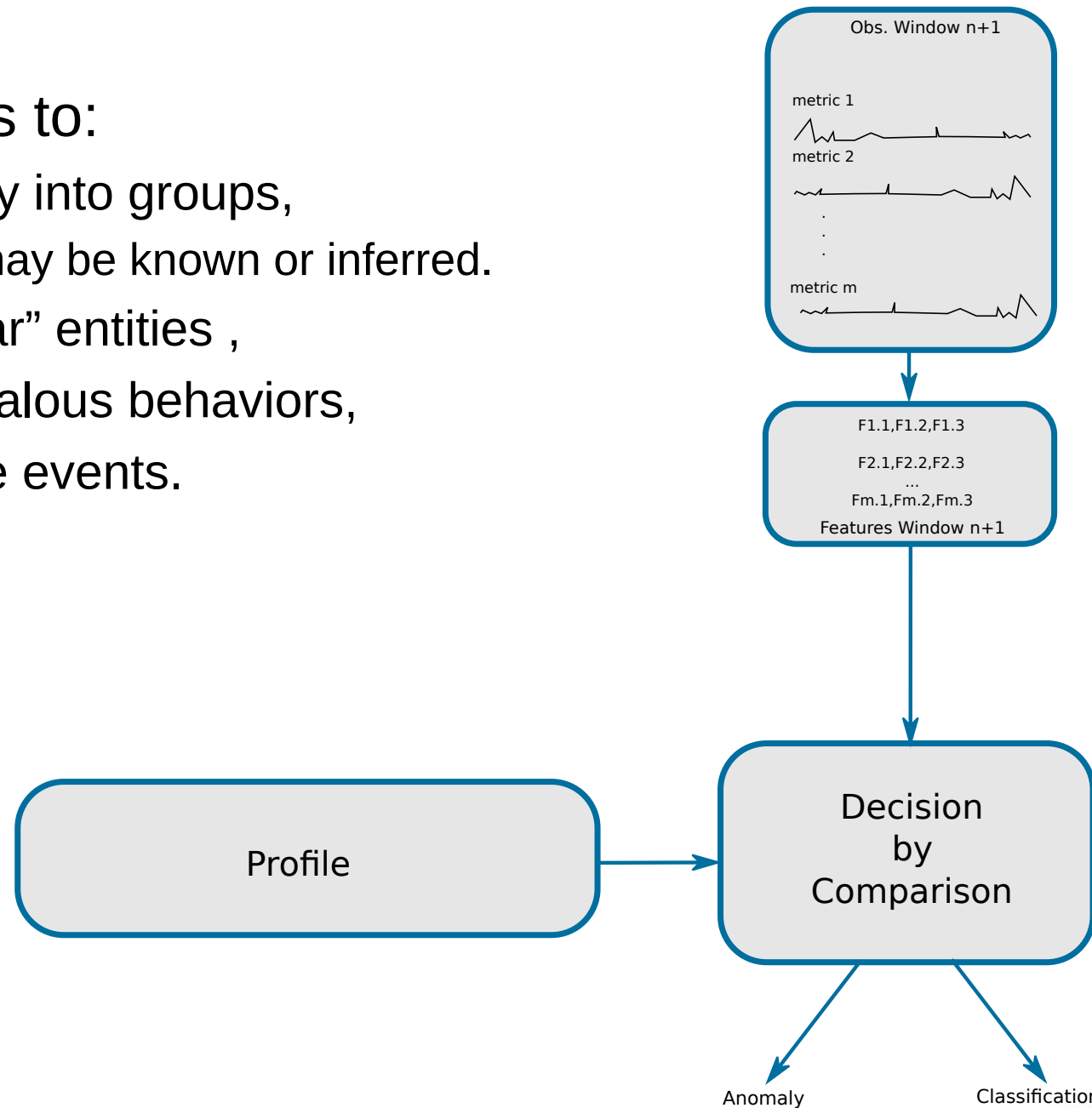
Entity Profiling

- Characterization of the observation windows after multiple observations.



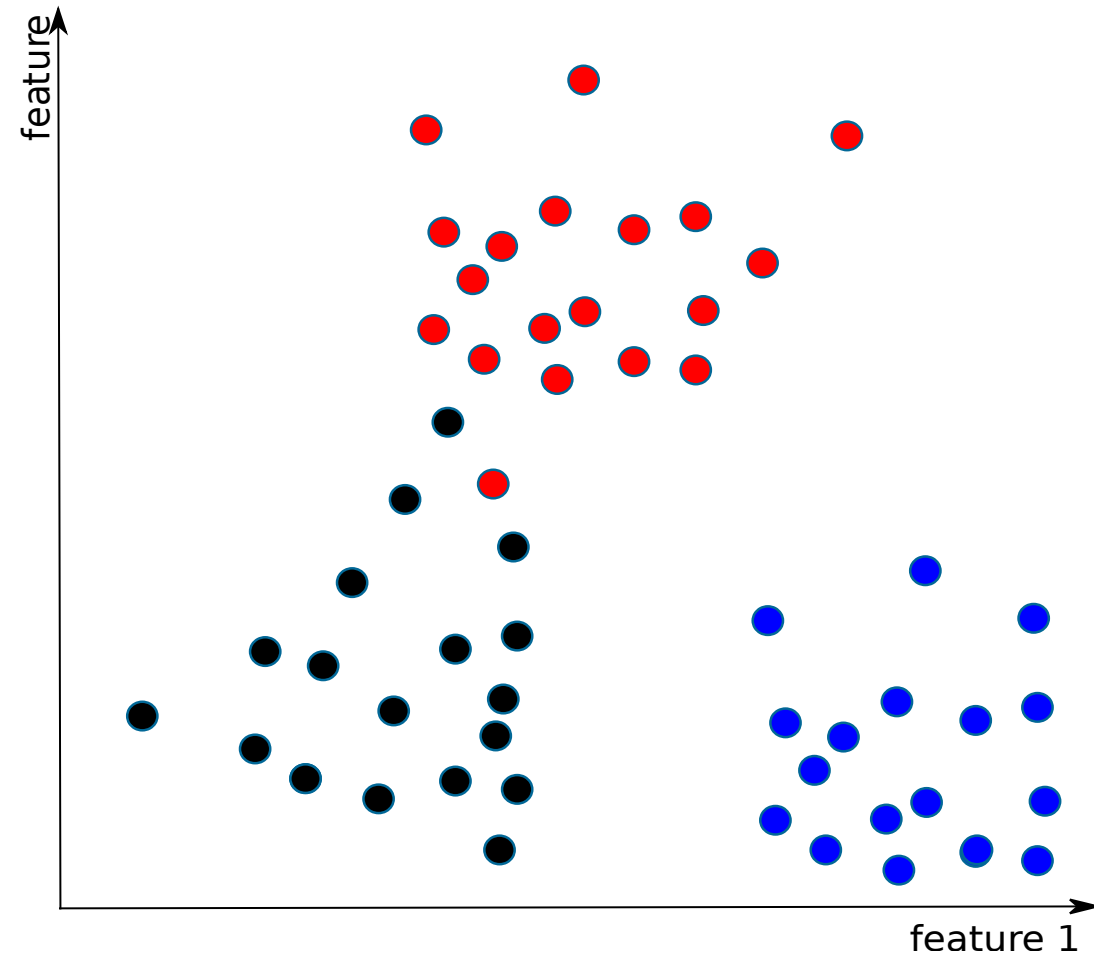
Profile Comparison

- A profile allows to:
 - Classify entity into groups,
 - Groups may be known or inferred.
 - Group “similar” entities ,
 - Detect anomalous behaviors,
 - Predict future events.



N-Dimensional Features Space

- A features' n-tuple defines a point in a N-Dimensional space that describes an entity behavior at a specific time.
- Allows to detect and define repetitive events and evolution over time.
- Allows to classify and discriminate behaviors.
- Allows to detect anomalies.



Data Formats

- The ideal data format is a n-tuple per time interval.
 - n metrics measured over time (n per observation).
$$(x_1, x_2, x_3, x_4, \dots, x_n)_k$$
 - Bi-dimensional data structure (time x metrics).
 - Optimal storing digital format:
 - ➔ Binary storage (array/matrix).
 - ➔ Sparse matrices could be advantageous.
 - ➔ Usage of fixed formats with integer indexes.
 - Avoid complex data structures with complex indexing of data, e.g.: python dictionaries.
 - ➔ Text formats are acceptable only in test scenarios.
 - ➔ Non-relational databases could also be an option.



Observation Features

- Time-independent descriptive statistics.
 - ♦ Mean, variance, standard deviation, quantiles, etc...
- Time-dependent descriptive statistics.
 - ♦ Time-relations between metrics over time
 - E.g., mean/std of length of silences [number of sampling slots with metric equal to zero], mean/std of length of activity [number of sampling slots with metric greater than zero], etc...
 - ♦ (Pseudo-)Periodicity components.
 - Time dependent.
 - Time multi-fractality (repetition of “similar events” in multiple time-scale).
 - Auto-correlation, FFT, CWT, DWT, and other spectral/frequency analysis.
- (Parameters of) Probabilistic functions/models.
 - ♦ Base function/model may be time independent or time dependent.



Descriptive Statistics (1)

- For a (equally) sampled-continuous time process:

$$X = \{x'_t = x_k, T_0 + k\Delta t \leq t < T_0 + (k+1)\Delta t, k = 1, 2, \dots, N\}$$

- **Mean:** $\mu = \frac{1}{N} \sum_{i=1}^N x_i$

- **Median:** $m_d = F^{-1}(0.5)$

- **Variance:** $Var(X) = \sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$

- **Standard deviation:** $SQRT(Var(X))$

- **Quantiles/Percentiles**

$$Y = \{y_j\}_{1 \leq j \leq N} = \text{sorted}(\{x_k\}_{1 \leq k \leq N})$$

- ♦ 64th percentile (64%)=0.64 quantile

- ♦ Quartiles: 25%, 50%, and 75%

$$\pi_p = \min(y_{j \geq pN})$$



Descriptive Statistics (3)

- Covariance

- Metric that quantifies how much two random variables have simultaneous variations:

$$\text{Cov}_{X,Y} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)$$

- Correlation coefficient

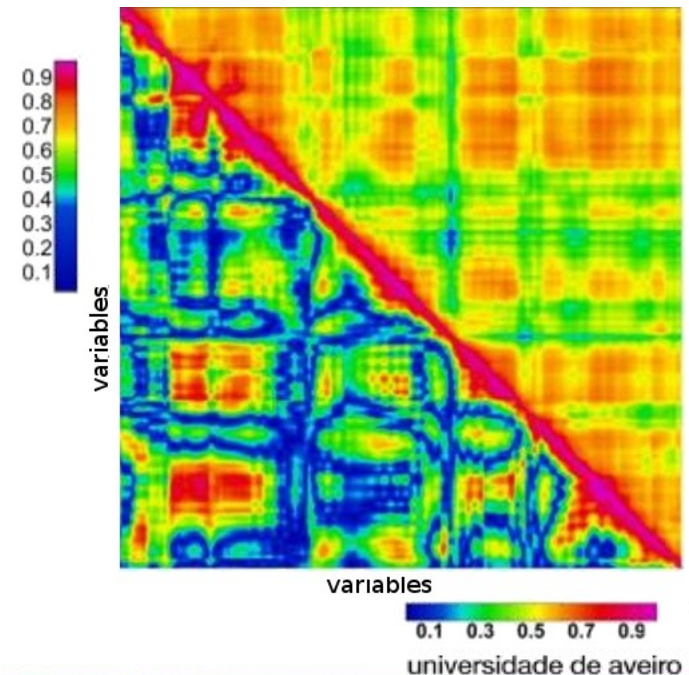
- Normalized covariance, varies between -1 and 1:

$$\rho_{X,Y} = \frac{\text{Cov}_{X,Y}}{\sigma_X \sigma_Y} \quad \sigma_X = \sqrt{\text{Var}(X)}$$

- Correlation matrix

- Defined by a (MxM) matrix, to quantify the correlation between M variables X_i :

$$C = \{c_{i,j}\}, i, j = 1, \dots, M$$
$$c_{i,j} = \rho_{X_i, X_j}$$



Periodicity Analysis (1)

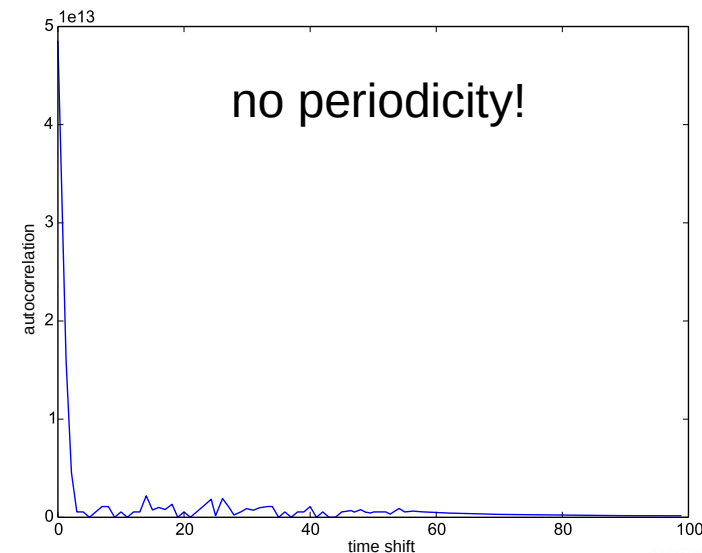
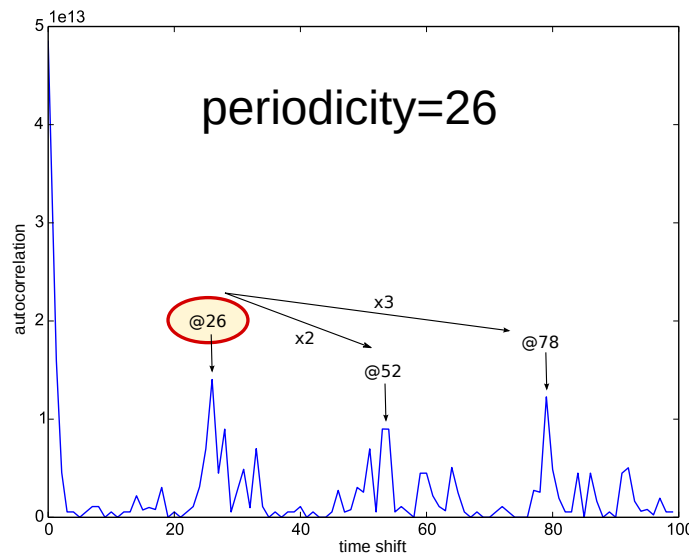
Autocorrelation

- Autocorrelation

- Correlation between the process and a shifted version (in time, by k samples) of the same process:

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \mu_X)(x_{i+k} - \mu_X)}{\sum_{i=1}^N (x_i - \mu_X)^2}$$

- Autocorrelation local maximums (peaks), reveal periodicity.
 - Differences between positions (k) of local maximums give periodicity.



Periodicity Analysis (2)

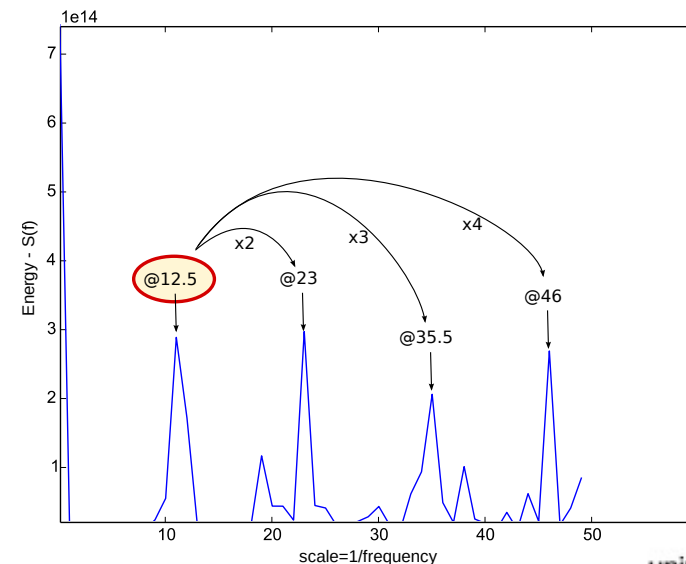
Periodograms

- Periodogram

- Frequency analysis → Spectral density estimation: Energy per frequency.
- Given by the modulus squared of the discrete Fourier transform.
 - For a signal x_i sampled every Δt :

$$S(f) = \frac{\Delta t}{N} \left| \sum_{n=1}^N x_n e^{-j2\pi n f} \right|^2, \quad -\frac{1}{2\Delta t} < t \leq \frac{1}{2\Delta t}$$

- The inverse of the frequencies with higher energy give the different periods (of periodicity).



Periodicity Analysis (3)

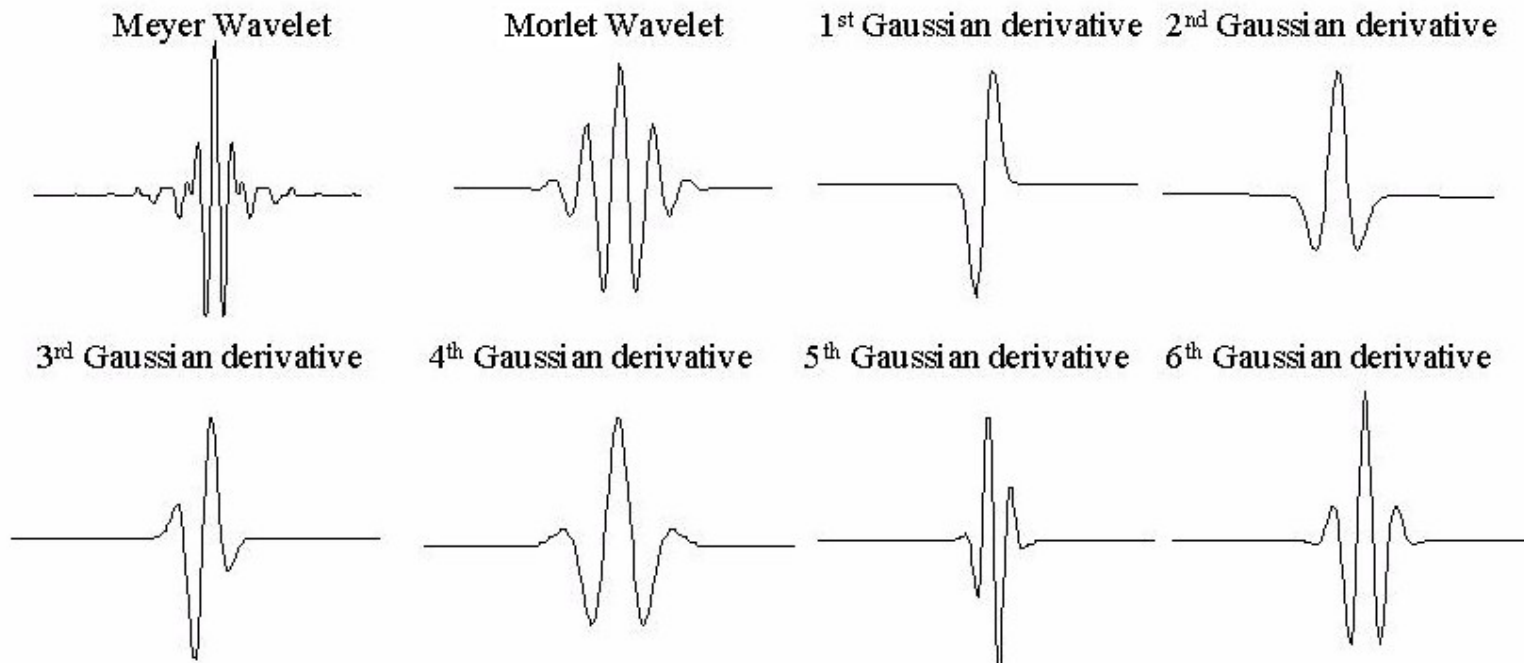
Scalograms

- Scalogram

- Joint Frequency/Time analysis → Wavelet Analysis
 - Energy per frequency/time.

$$\Psi_x^\psi(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} x(t) \psi^*\left(\frac{t - \tau}{s}\right) dt$$

Wavelet
functions
 $\psi^*(t)$



Periodicity Analysis (4)

Scalograms

- Given by the normalized modulus squared of the Wavelet transform.

$$\hat{E}_x(\tau, s) = \frac{|\Psi_x^\psi(\tau, s)|^2}{\sum_{\tau' \in \mathbf{T}} \sum_{s' \in \mathbf{S}} |\Psi_x^\psi(\tau', s')|^2}$$

→ Averaged over time.

$$\bar{e}_x(s) = \frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} \hat{E}_x(\tau, s), \forall s \in \mathbf{S}$$

