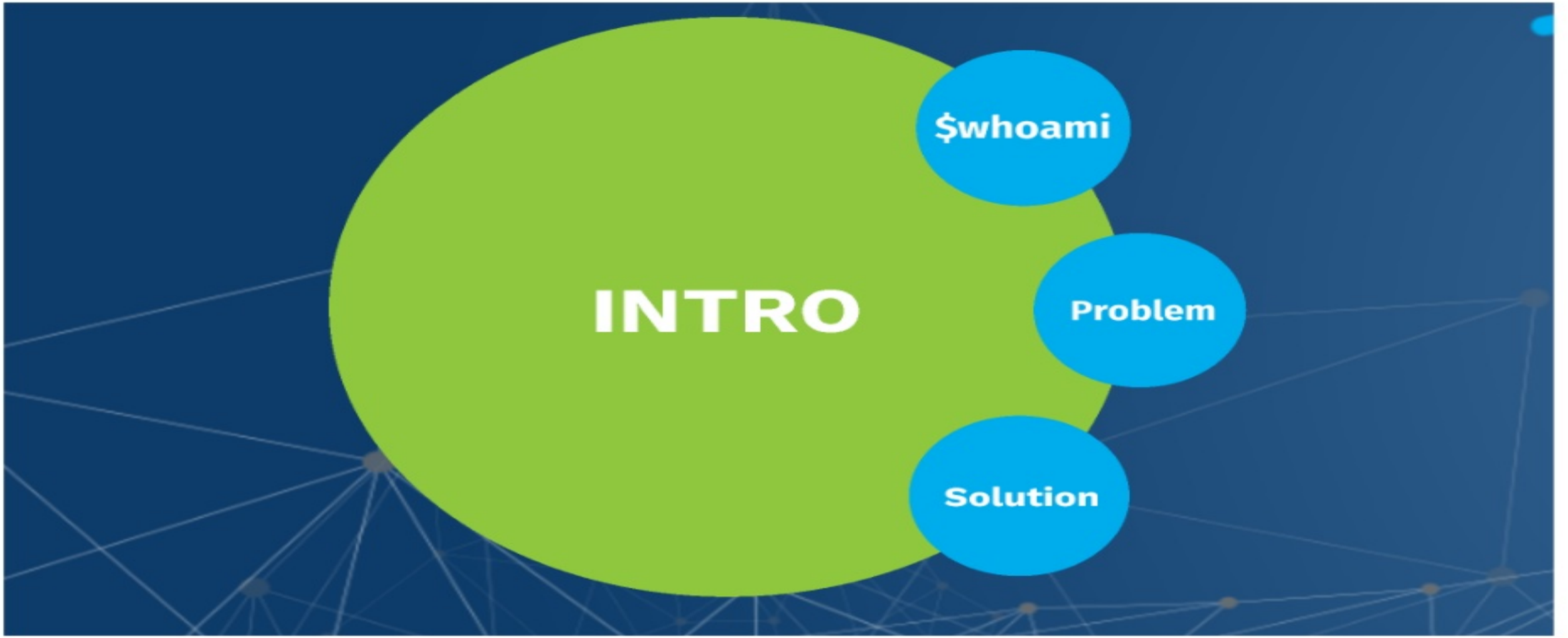

STREAMING
DIGITAL
FINGERPRINTS
INTRO
GOALS
CASE
STUDY
THANKS
Sebastian Czarnota
Digital Fingerprints


INTRO
$whoami
Problem
Solution

$whoami

Sebastian Czarnota

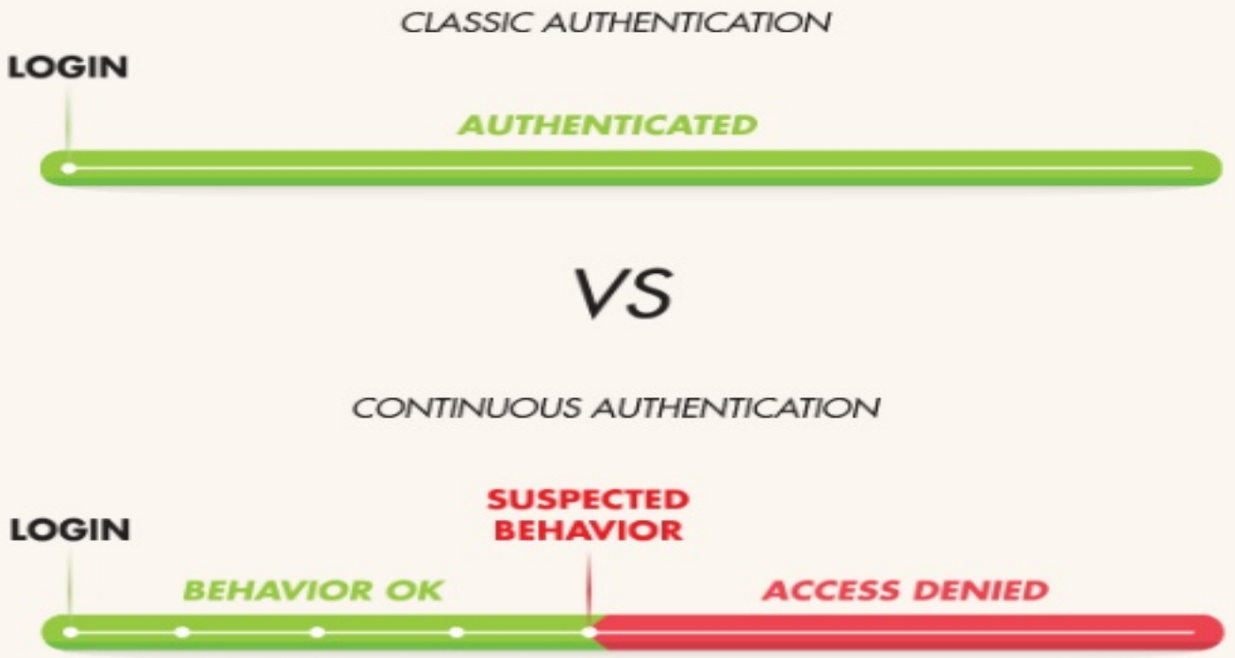
Principal Streaming Architect

sebastian@fingerprints.digital

https://fingerprints.digital/

Problem

Solution


CLASSIC AUTHENTICATION
LOGIN
AUTHENTICATED
VS
CONTINUOUS AUTHENTICATION
LOGIN
SUSPECTED BEHAVIOR
BEHAVIOR OK
ACCESS DENIED


STREAMING
DIGITAL
FINGERPRINTS
INTRO
GOALS
CASE
STUDY
THANKS
Sebastian Czarnota
Digital Fingerprints


GOALS
User Anonymity
User Experience
Simple Integration
Fast reaction time

User Anonymity as First Citizen Feature

User Anonymity as First Citizen Feature

- We **care** for users

User Anonymity as First Citizen Feature

- We **care** for users
- GDPR **compliant**

User Anonymity as First Citizen Feature

- We **care** for users
- GDPR **compliant**
- Personal information **free**

User Anonymity as First Citizen Feature

- We **care** for users
- GDPR **compliant**
- Personal information **free**
- **Limited** consent requirements

We still remember "Don't be evil"...

Seamless User experience

Seamless User experience

- No visible changes in **UI**

Seamless User experience

- No visible changes in **UI**
- **Low** memory footprint

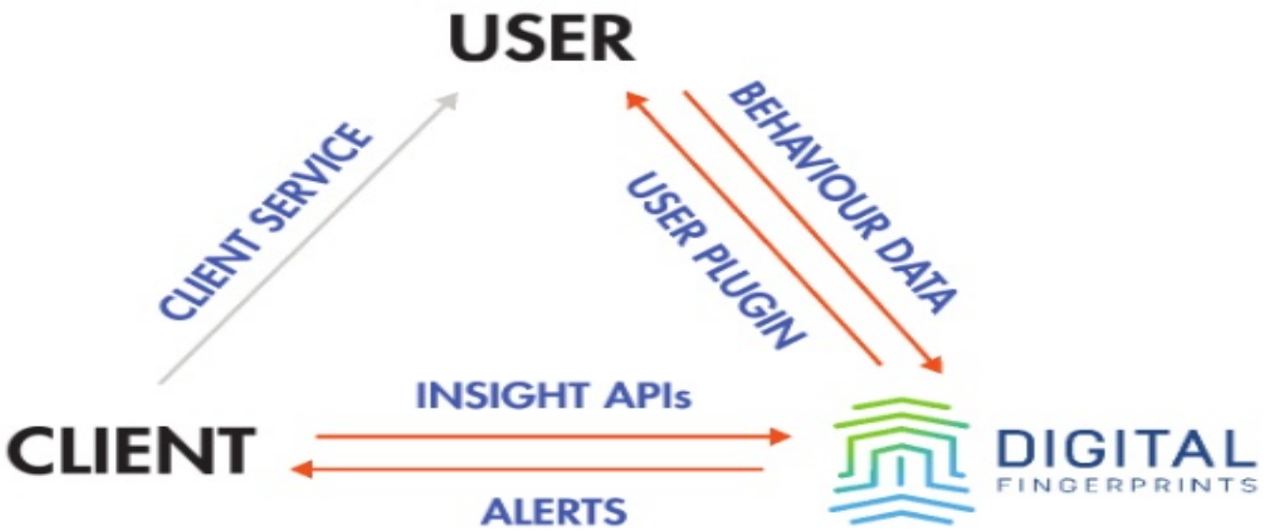
Seamless User experience

- No visible changes in **UI**
- **Low** memory footprint
- **Optimized** network traffic

Seamless User experience

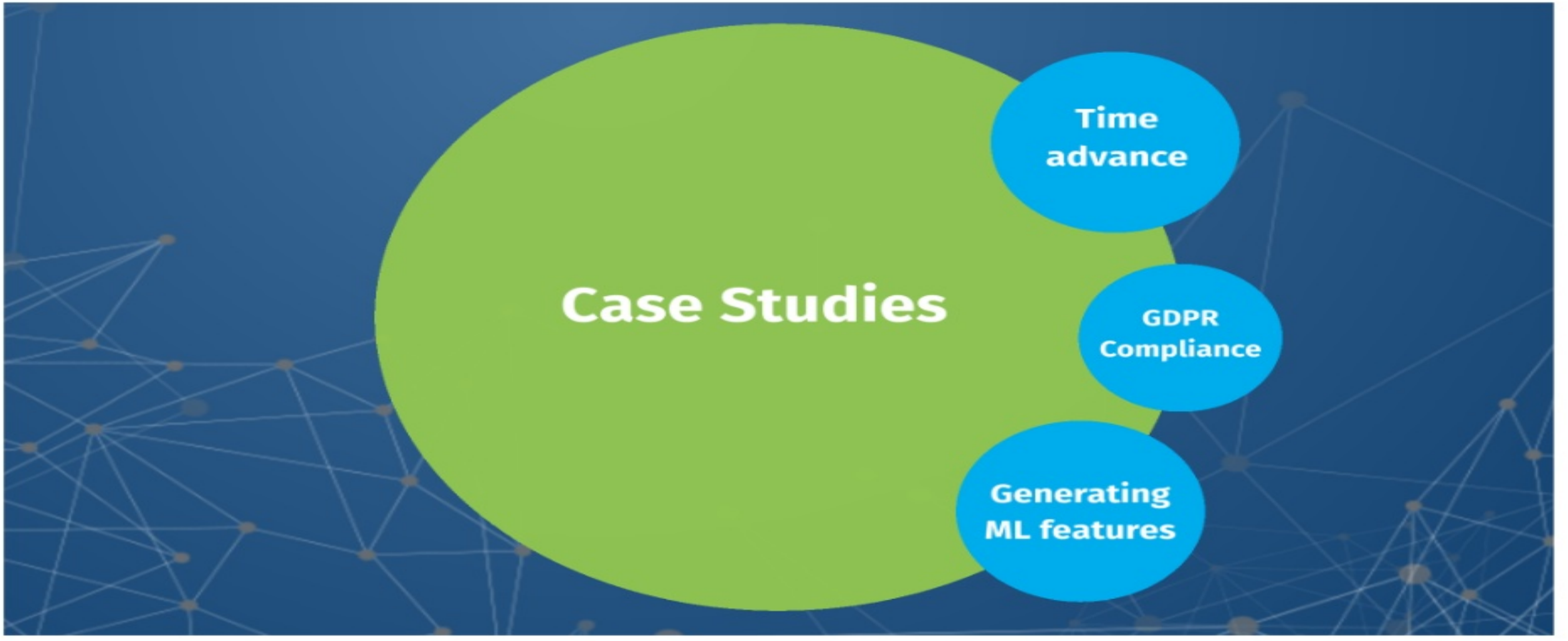
- No visible changes in **UI**
- **Low** memory footprint
- **Optimized** network traffic
- Handle **wide** range of systems

Simple integration



Provide fast reaction time


STREAMING
DIGITAL
FINGERPRINTS
INTRO
GOALS
CASE
STUDY
THANKS
Sebastian Czarnota
Digital Fingerprints


Case Studies
Time advance
GDPR Compliance
Generating ML features

How to advance time?

- Background
- Processing time
- Event time
- Ingestion time
- Our approach

Key requirements

Key requirements

- **Low** latency -> **Fast** reaction time

Key requirements

- **Low** latency -> **Fast** reaction time
- **Scalability** to met growing load

Key requirements

- **Low** latency -> **Fast** reaction time
- **Scalability** to met growing load
- Reprocessing and **time-travel**

Reprocessing

Reprocessing

- Process **lots** of stored data **fast**

Reprocessing

- Process **lots** of stored data **fast**

Reprocessing

- Process **lots** of stored data **fast**

- For **experimenting** and machine learning model **improvement**

Message lifetime

Processing Time Characteristic

Time advances according to system's time

Processing Time Characteristic

Time advances according to system's time

+ Simplicity - wall clock for processing

Processing Time Characteristic

Time advances according to system's time

+ Simplicity - wall clock for processing
+ Low latency footprint

Processing Time Characteristic

Time advances according to system's time

+ Simplicity - wall clock for processing
+ Low latency footprint

- Non-deterministic results!!!

Processing Time Characteristic

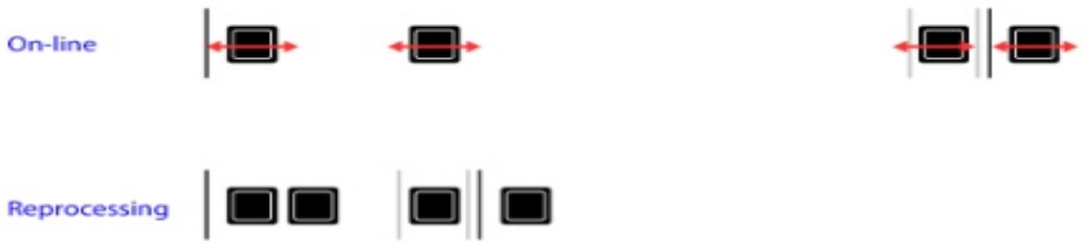
Time advances according to system's time

+ Simplicity - wall clock for processing
+ Low latency footprint

- Non-deterministic results!!!
- Hard to test processing time timers

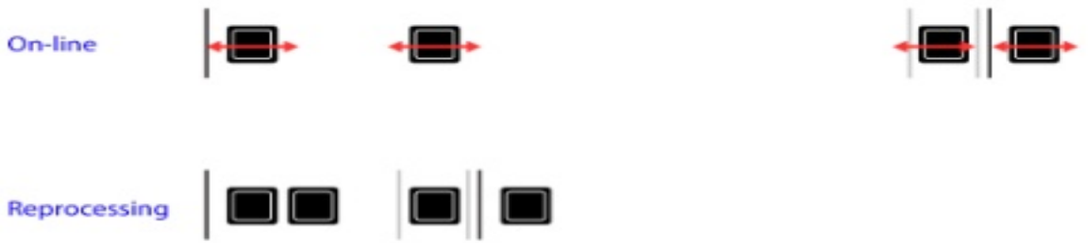
Event Time Characteristic

Time advances according to external timestamps - event time



Event Time Characteristic

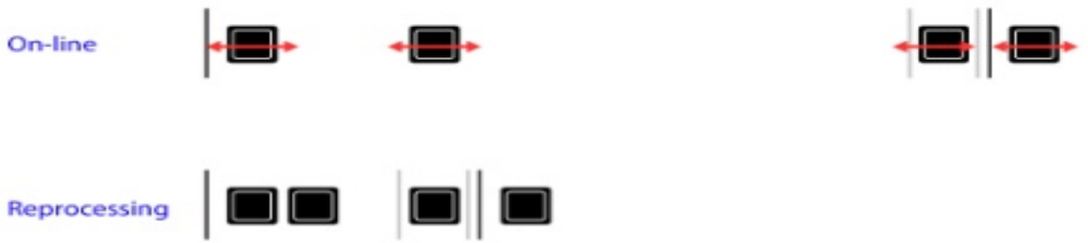
Time advances according to external timestamps - event time



+ Deterministic results

Event Time Characteristic

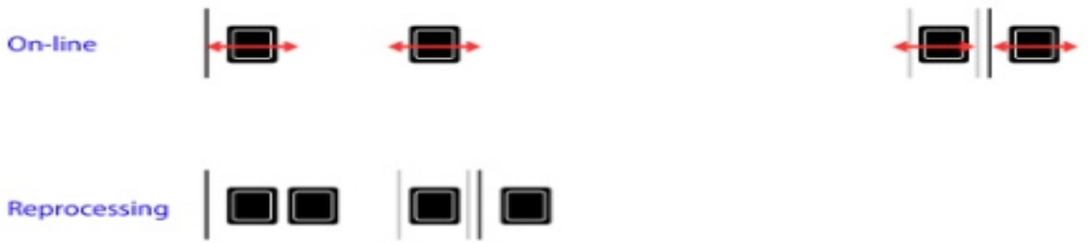
Time advances according to external timestamps - event time



+ Deterministic results
+ Easy to test

Event Time Characteristic

Time advances according to external timestamps - event time

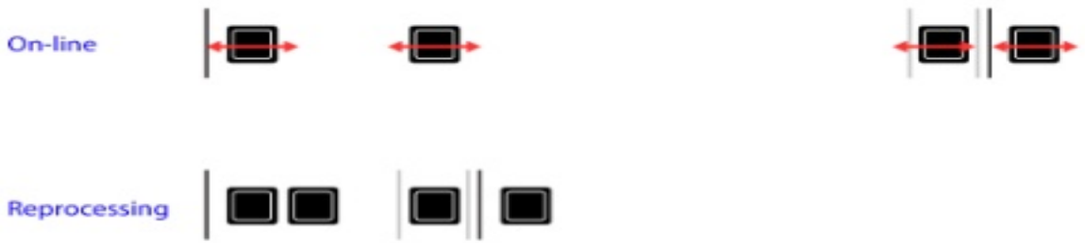


+ Deterministic results
+ Easy to test

- Every user may have different time...

Event Time Characteristic

Time advances according to external timestamps - event time

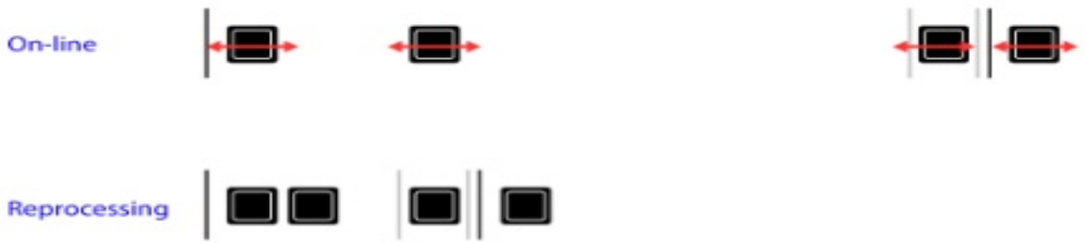


+ Deterministic results
+ Easy to test

- Every user may have different time...
  which requires delaying time advance by lag...

Event Time Characteristic

Time advances according to external timestamps - event time

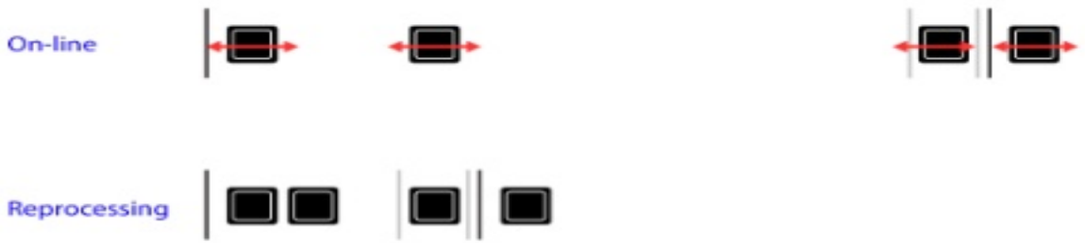


+ Deterministic results
+ Easy to test

- Every user may have different time...
  which requires delaying time advance by lag...
  thus increases latency and reaction time

Event Time Characteristic

Time advances according to external timestamps - event time



+ Deterministic results
+ Easy to test

- Every user may have different time...
  which requires delaying time advance by lag...
  thus increases latency and reaction time
- Risks due to timestamp tampering

Ingestion Time Characteristic

Time advances according to message's ingestion time

Ingestion Time Characteristic

Time advances according to message's ingestion time

+ We control source of time

Ingestion Time Characteristic

Time advances according to message's ingestion time

+ We control source of time

- Non-deterministic reprocessing

Ingestion Time Characteristic

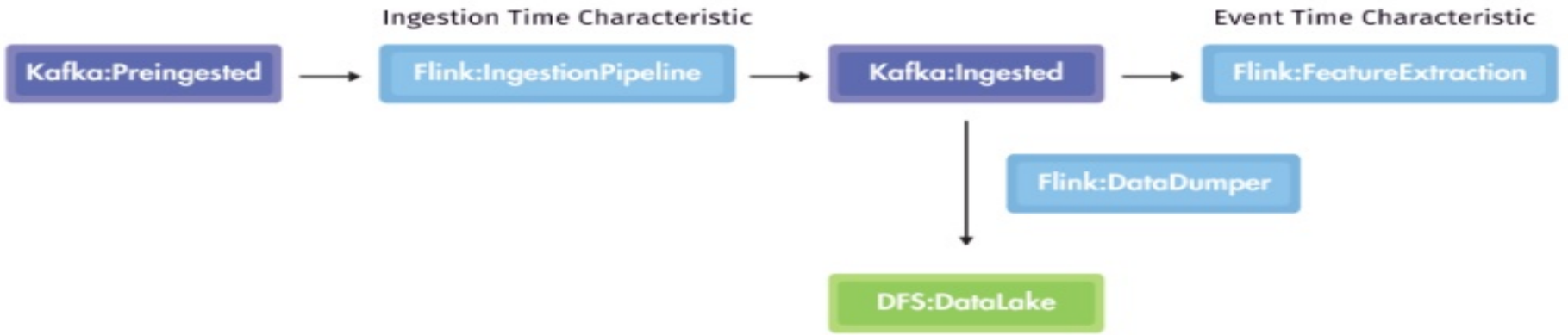
Time advances according to message's ingestion time

+ We control source of time

- Non-deterministic reprocessing
- Whole pipeline tests are hard

Our approach to time advancing

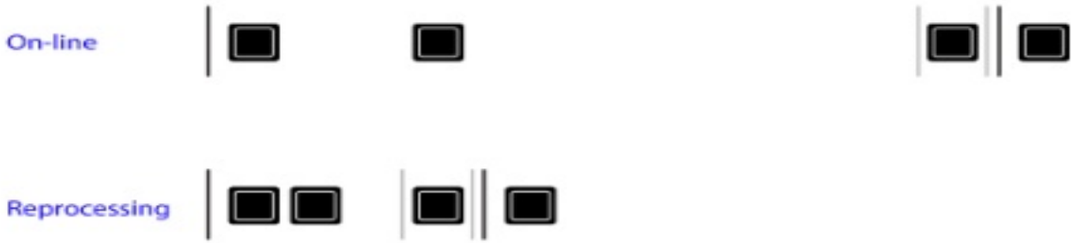
Time advances according to message's arrival time on boundary system



- We **combined** two time characteristics to achieve desired result

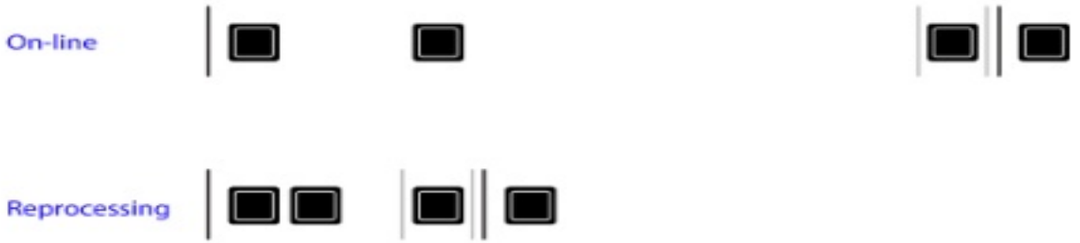
Our approach to time advancing

Time advances according to message's arrival time on boundary system



Our approach to time advancing

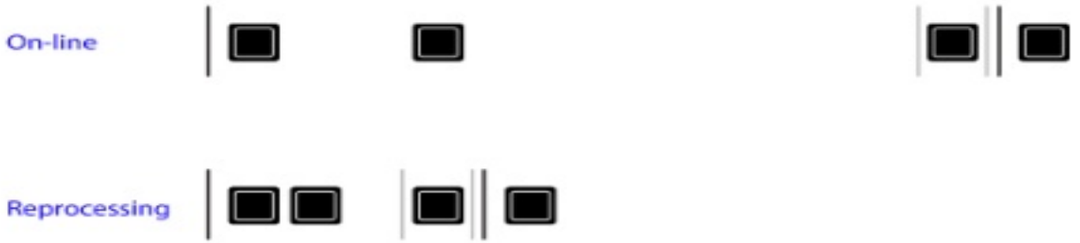
Time advances according to message's arrival time on boundary system



+ We control source of time

Our approach to time advancing

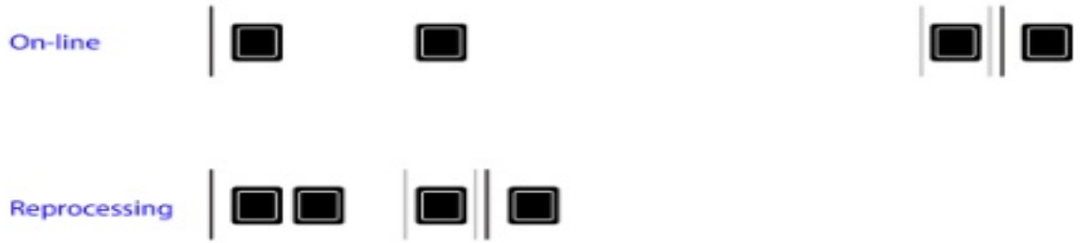
Time advances according to message's arrival time on boundary system



+ We control source of time
+ Stable reprocessing

Our approach to time advancing

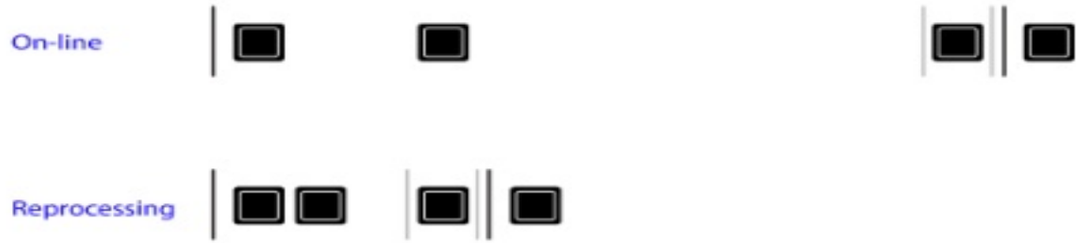
Time advances according to message's arrival time on boundary system



+ We control source of time
+ Stable reprocessing
+ Archive data with fixed timestamps

Our approach to time advancing

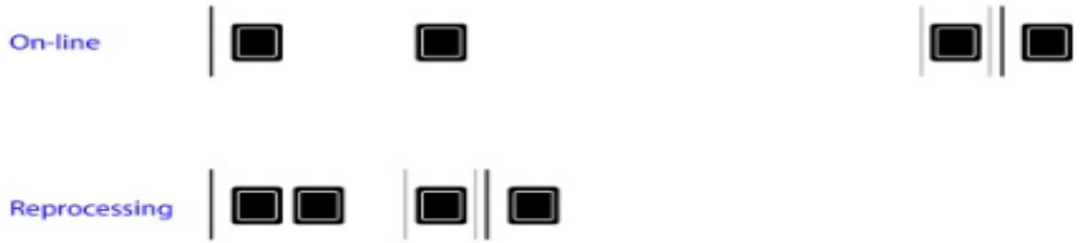
Time advances according to message's arrival time on boundary system



+ We control source of time
+ Stable reprocessing
+ Archive data with fixed timestamps
+ Easy track reaction time in on-line

Our approach to time advancing

Time advances according to message's arrival time on boundary system



+ We control source of time
+ Stable reprocessing
+ Archive data with fixed timestamps
+ Easy track reaction time in on-line

– Reprocessing reaction time incorrect

Advancing Time Summary

- Always keep in mind your **requirements**
- **Know** drawbacks of each time characteristic
- You can **combine** time characteristics to fit your needs
- You can (and **should**!) unittest Flink applications

- **Watch out** for checkpointing interval!


Technical aspects of GDPR compliance
EU GDPR COMPLIANT
Requirements
Anonymization
Flink state weeding
Data lake
Machine Learning

GDPR Requirements

GDPR imposes requirements to:
- **delete** every bit of user's data when asked
- **move** user's data to another service provider
- **hold** raw data for extended time
- have **consent** for behaviour data processing

Anonymization

The less personal data you process the less consent you need...

NO! We are not gathering user's logins or passwords

Flink state weeding

Flink state weeding

- Remember **clearing** operator state!

Flink state weeding

- Remember **clearing** operator state!
- Use **timers** and inactivity markers

Flink state weeding

- Remember **clearing** operator state!
- Use **timers** and inactivity markers
- We can't wait for **Time-to-live** option on state variables

Flink state weeding

- Remember **clearing** operator state!
- Use **timers** and inactivity markers
- We can't wait for **Time-to-live** option on state variables
- It would be nice to have timer **unregistering**

GDPR and data lake

GDPR and data lake

- We **gather** behavioral data for:
  - ML methods research
  - Model training
  - Legal purposes - automatic decision making

GDPR and data lake

- We **gather** behavioral data for:
  - ML methods research
  - Model training
  - Legal purposes - automatic decision making
- **Index** data by user_id

GDPR and data lake

- We **gather** behavioral data for:
  - ML methods research
  - Model training
  - Legal purposes - automatic decision making
- **Index** data by user_id
- **Remove** it when asked

GDPR and data lake

- We **gather** behavioral data for:
  - ML methods research
  - Model training
  - Legal purposes - automatic decision making
- **Index** data by user_id
- **Remove** it when asked
- You need **indexable** data lake

GDPR and Machine Learning models

GDPR and Machine Learning models

- We are building Machine Learning models **per user**

GDPR and Machine Learning models

- We are building Machine Learning models **per user**
- We had to **index** by user_id besides GDPR

GDPR and Machine Learning models

- We are building Machine Learning models **per user**
- We had to **index** by user_id besides GDPR
- Even after removing user's data there are **echos**:
  - In metrics and summaries
  - **Residual** impact on other user's models

GDPR and Machine Learning models

- We are building Machine Learning models **per user**
- We had to **index** by user_id besides GDPR
- Even after removing user's data there are **echos**:
  - In metrics and summaries
  - **Residual** impact on other user's models
- Residual impact **cannot** help find user

GDPR and Machine Learning models

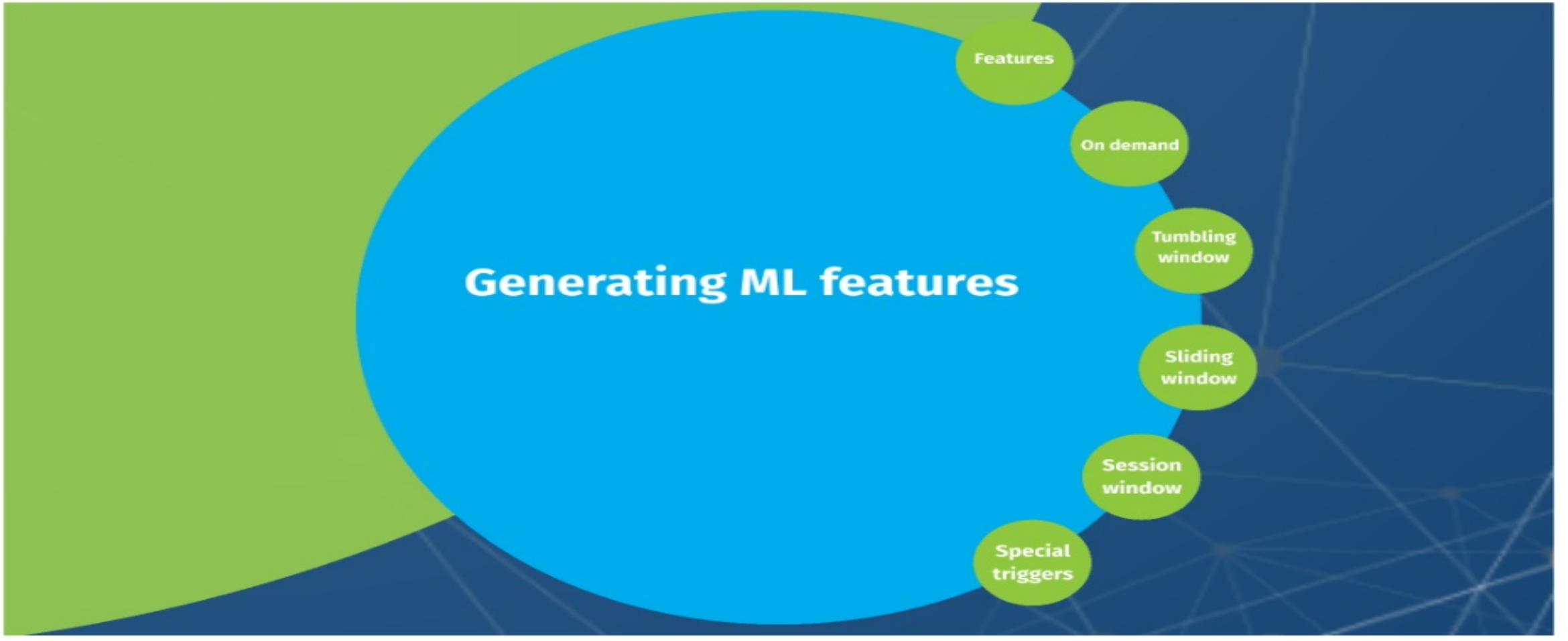
- We are building Machine Learning models **per user**
- We had to **index** by user_id besides GDPR
- Even after removing user's data there are **echos**:
  - In metrics and summaries
  - **Residual** impact on other user's models
- Residual impact **cannot** help find user
- Residual impact **fades** with time

GDPR and Machine Learning models

- We are building Machine Learning models **per user**
- We had to **index** by user_id besides GDPR
- Even after removing user's data there are **echos**:
  - In metrics and summaries
  - **Residual** impact on other user's models
- Residual impact **cannot** help find user
- Residual impact **fades** with time
- Residual impact are **NOT** against GDPR

Summary

- Index your data by user ID
- Be able to move and remove data by user Id


Generating ML features
Features
On demand
Tumbling window
Sliding window
Session window
Special triggers

Generating Machine Learning Features

- Low latency
- Deterministic response time
- Catch behaviour dynamics

Generating features on demand

Generating features on demand

Generating features on demand

+ Cheap computationally

Generating features on demand

+ Cheap computationally

- High latency

Generating features on demand

+ Cheap computationally

- High latency
- Non deterministic latency

Generating features on demand

+ Cheap computationally

- High latency
- Non deterministic latency
- Cannot do pushing mode

Tumbling window

Tumbling window

+ Quite light computationally

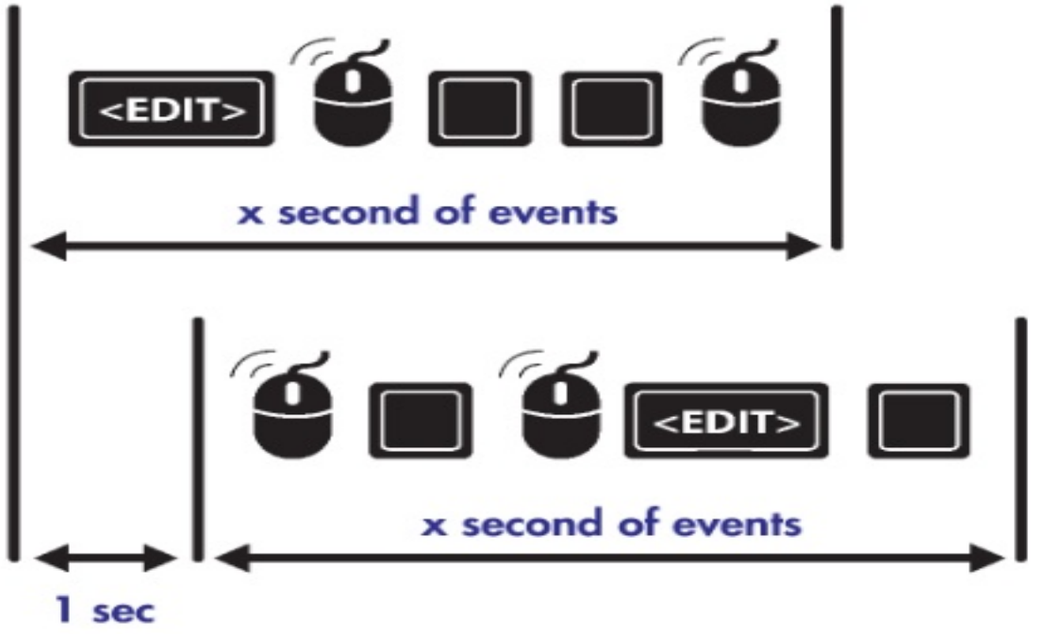
Tumbling window

+ Quite light computationally
- Variable events in feature - poor ML models

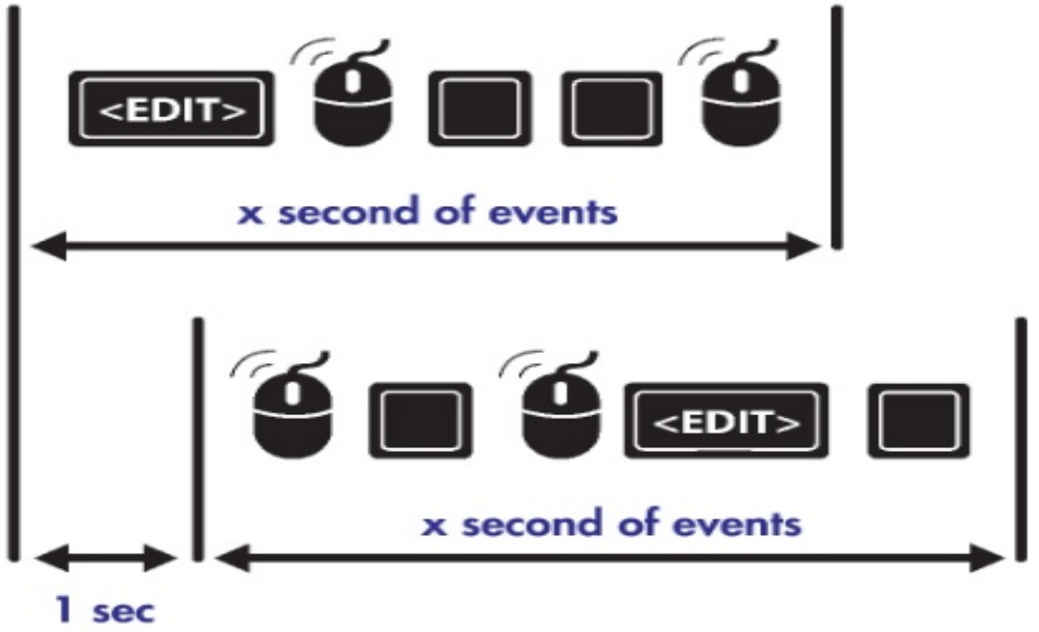
Tumbling window

+ Quite light computationally
- Variable events in feature - poor ML models
- Very few features

Sliding window

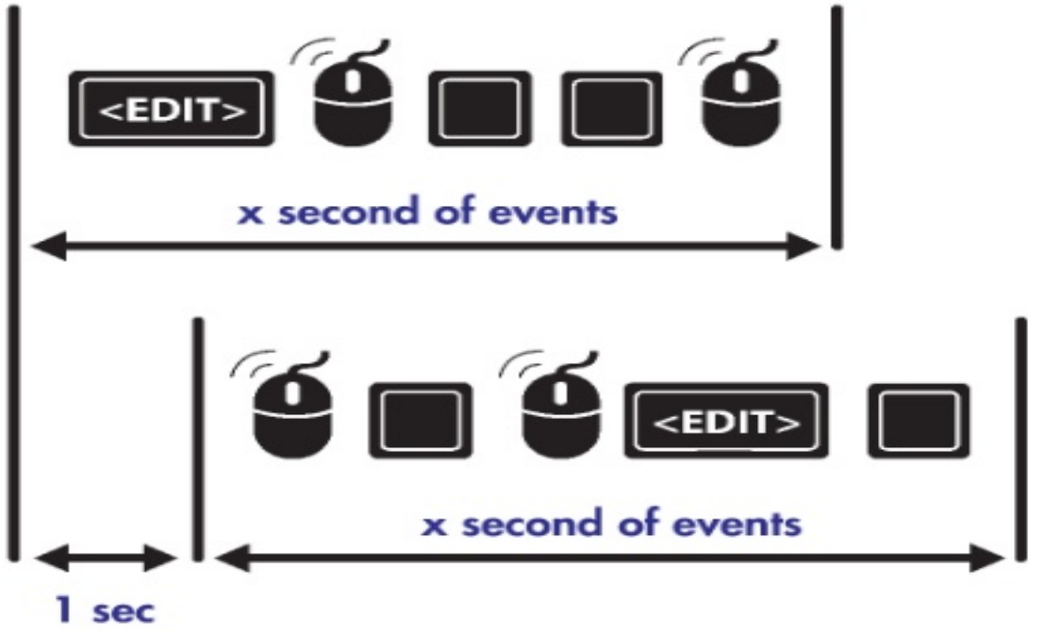


Sliding window



+ More features

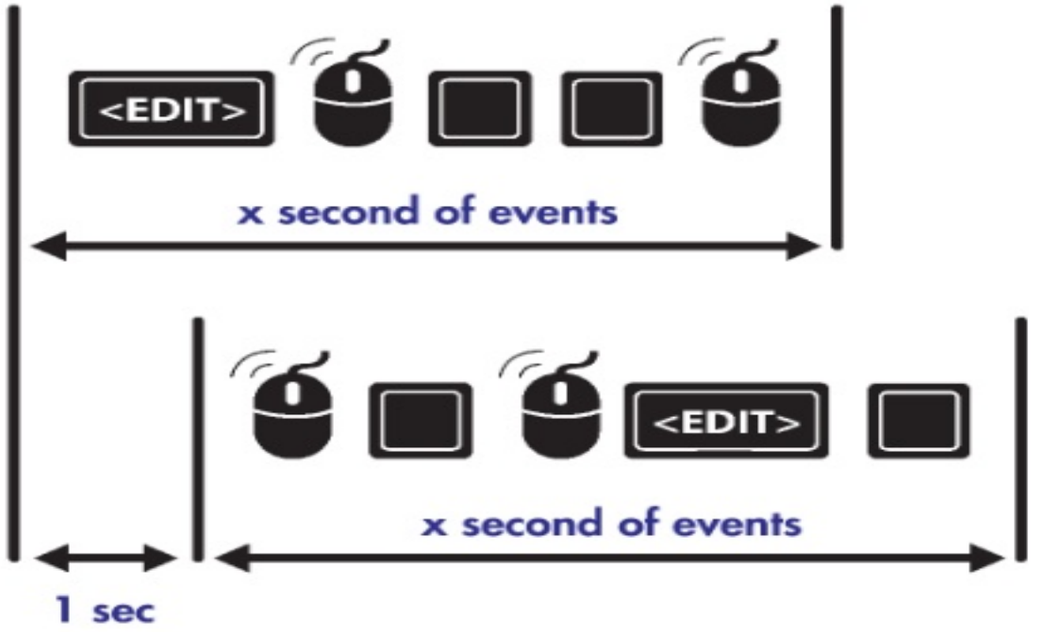
Sliding window



+ More features

- Computationally more intensive

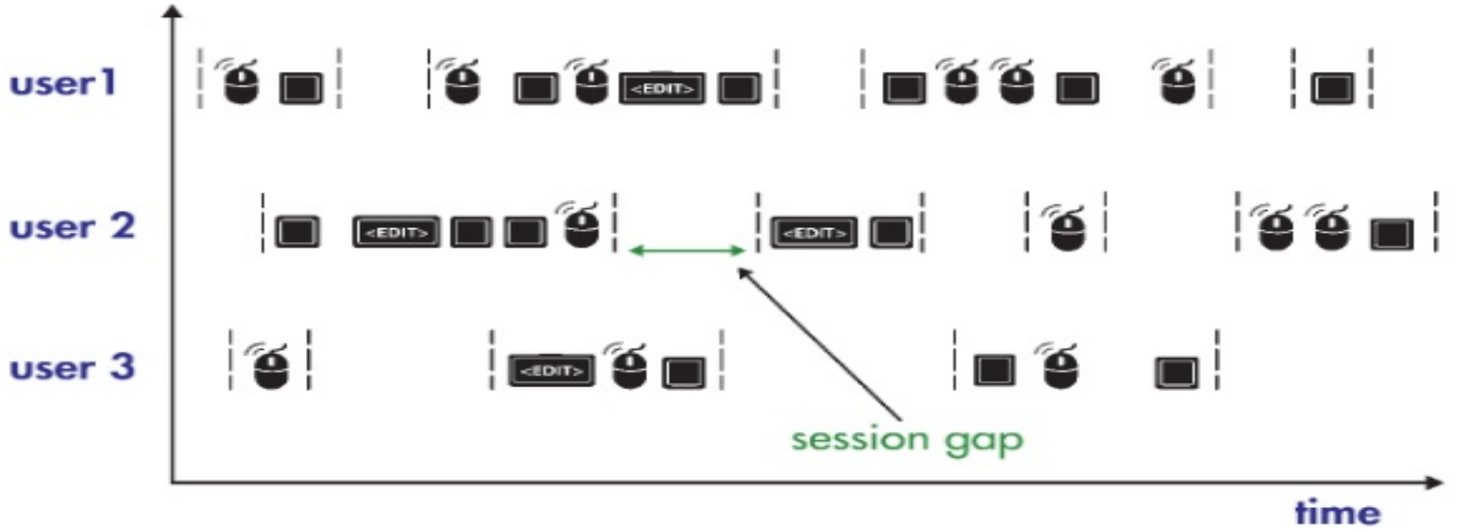
Sliding window



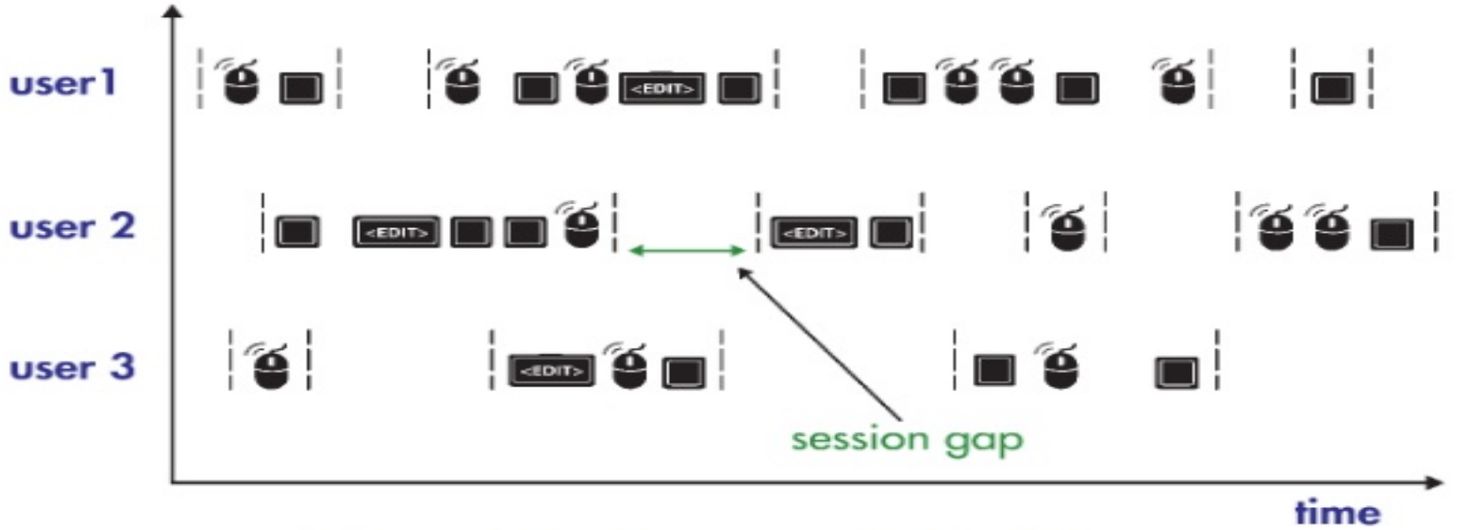
+ More features

- Computationally more intensive
- As poor results as in tumbling window

Session window

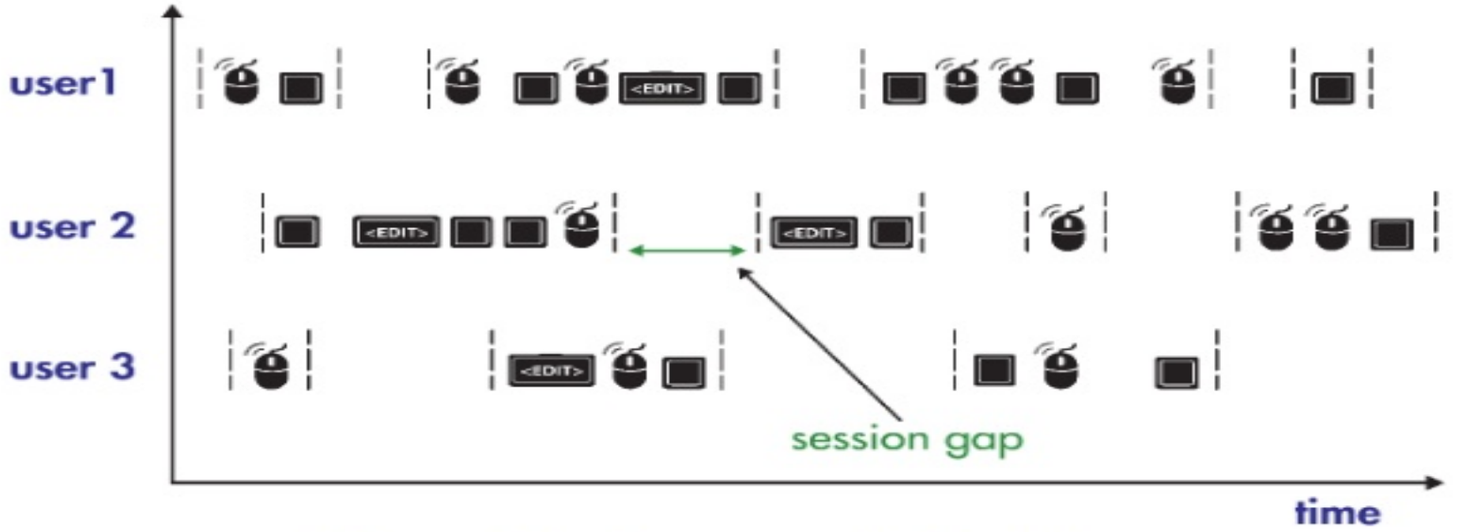


Session window



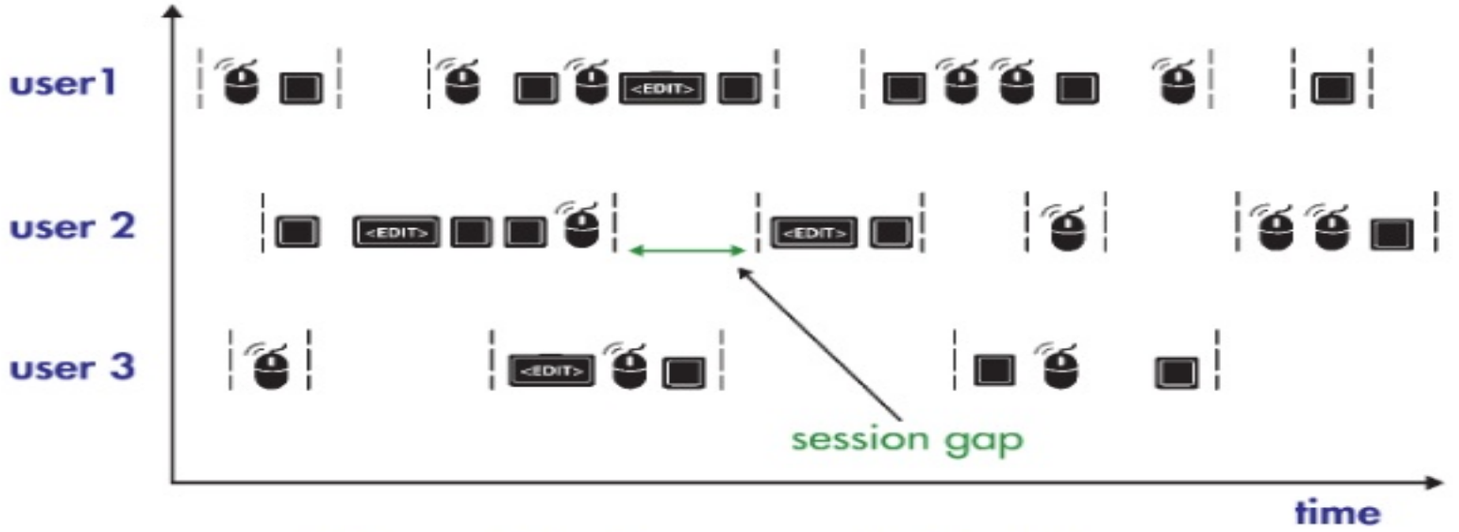
+ Not so intensive computationally

Session window



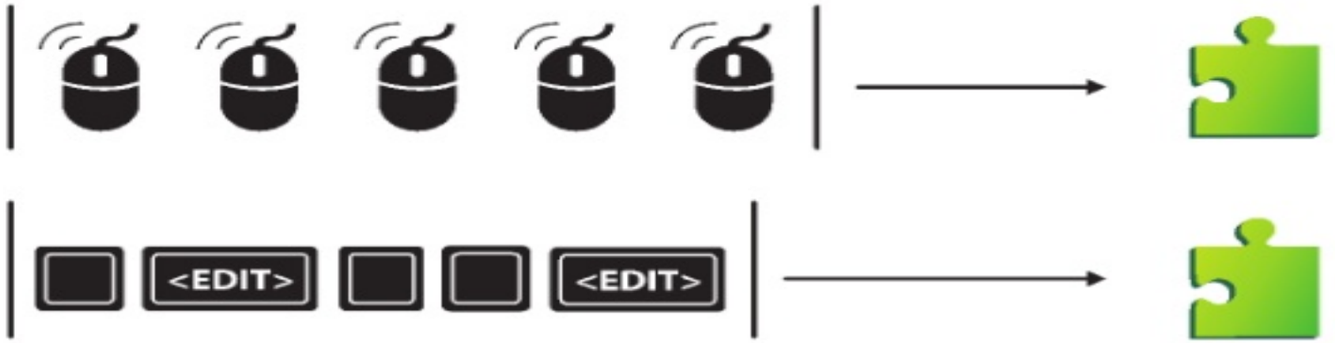
+ Not so intensive computationally
- Extends latency by session gap (sic!)

Session window

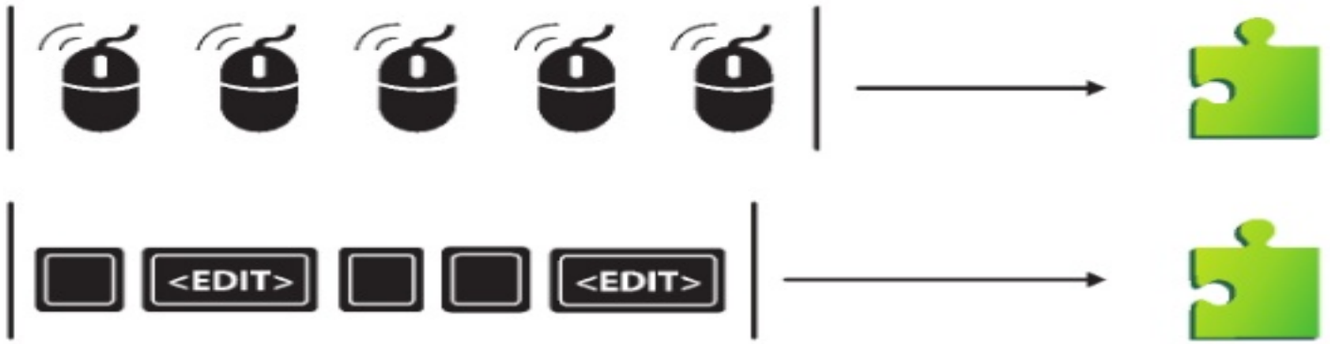


+ Not so intensive computationally
- Extends latency by session gap (sic!)
- Again variable number of events

Special triggers

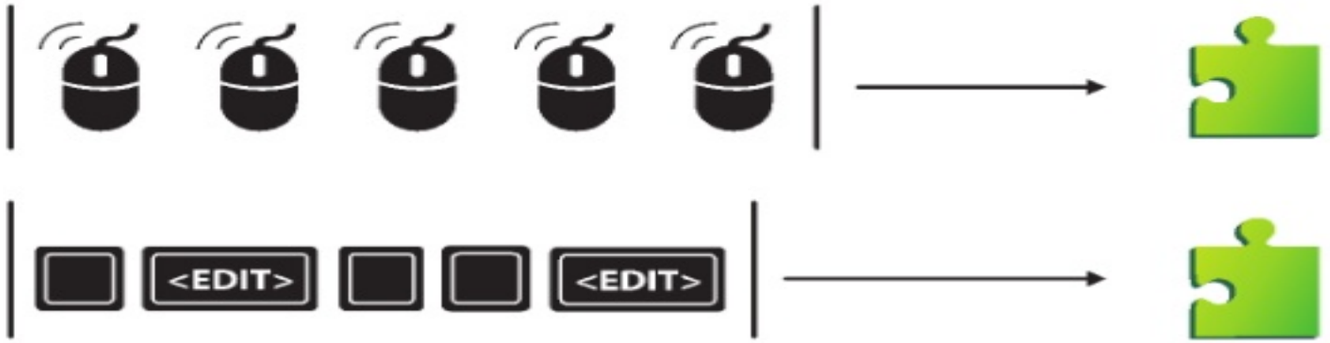


Special triggers



- Computationally intensive

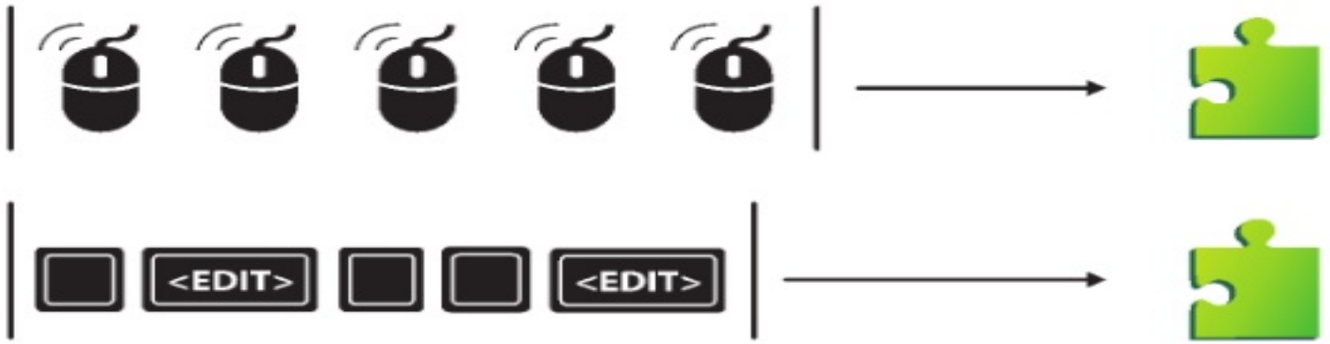
Special triggers



- Computationally intensive

+ Constant number of events

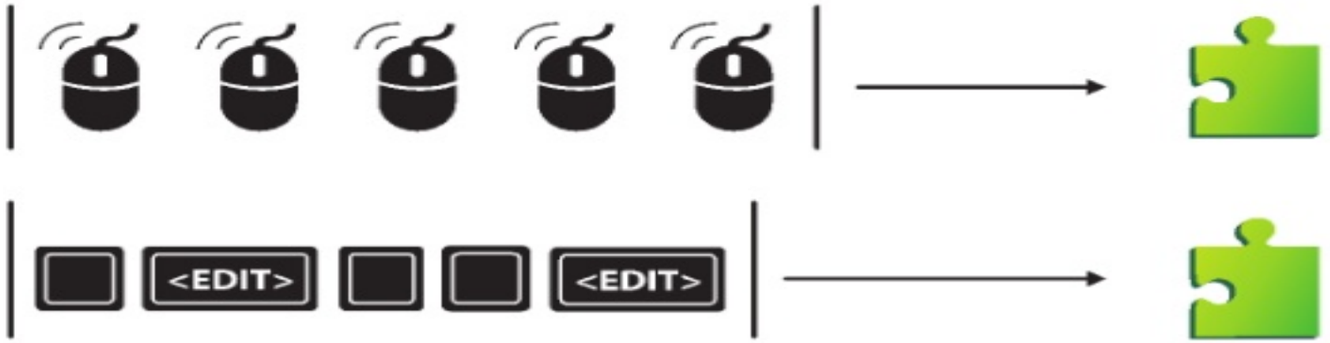
Special triggers



- Computationally intensive

+ Constant number of events
+ No latency slow downs

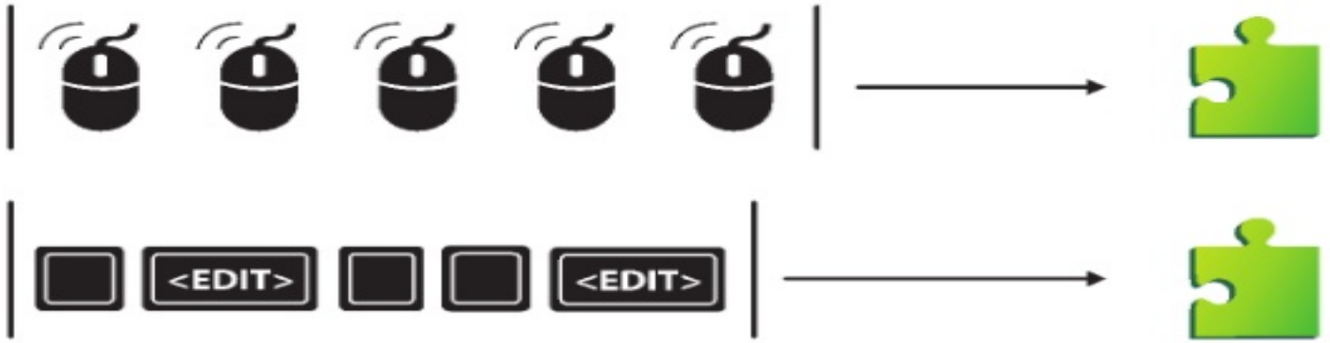
Special triggers



- Computationally intensive

+ Constant number of events
+ No latency slow downs
+ Lots of features

Special triggers



- Computationally intensive

+ Constant number of events
+ No latency slow downs
+ Lots of features
+ Good ML results

Generating ML Features Summary

- Event grouping doesn't end at standard Flink windows
- Test everything - results more then often are surprising


STREAMING
DIGITAL
FINGERPRINTS
INTRO
GOALS
CASE
STUDY
THANKS
Sebastian Czarnota
Digital Fingerprints

Thanks and wishes

- Deployment much improved!
- Waiting for state Time-To-Live
- Nice to have: efficient unregistering of many timers
- Need to have: testing utilities for processing time applications

- Conquer the world


STREAMING
DIGITAL
FINGERPRINTS
INTRO
GOALS
CASE
STUDY
THANKS
Sebastian Czarnota
Digital Fingerprints