Sequence Synopsis: Optimize Visual Summary of Temporal Event Data

IEEE VAST 2017 (TVCG)

Yuanzhe Chen
Hong Kong University of Science and Technology

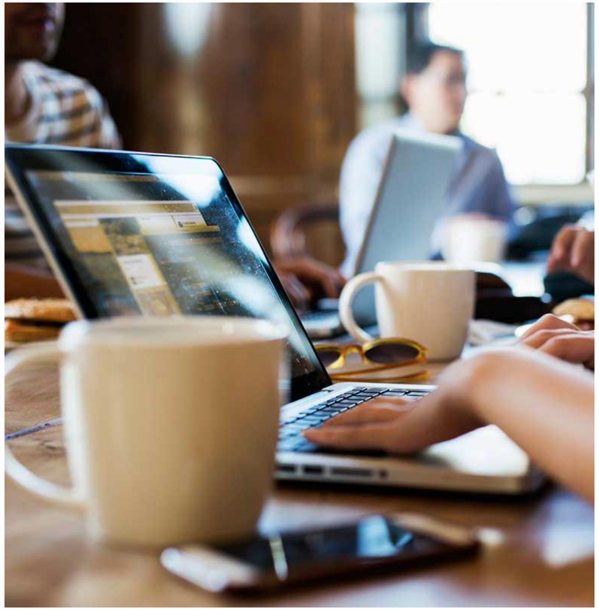
Panpan Xu, Liu Ren
Bosch Research North America
Palo Alto, CA

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

BOSCH

MOTIVATION

Event Sequences
Use Case: Human Activities Analysis

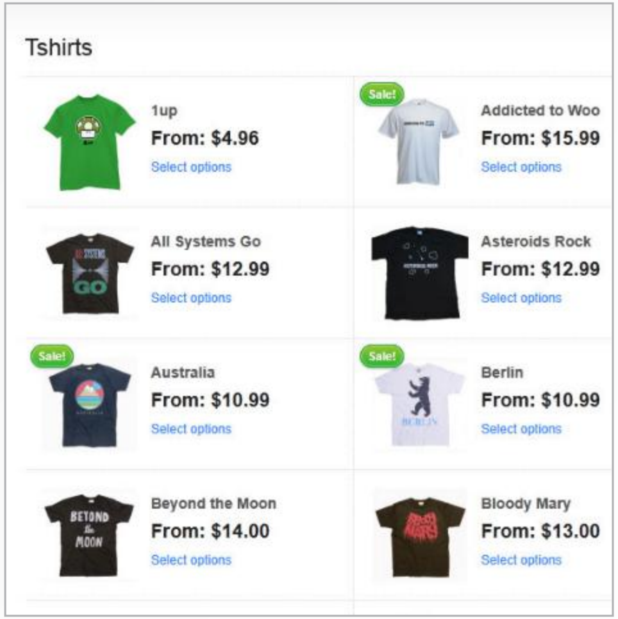


wake up → breakfast → start work

Research and Technology Center North America | CR/RTC-HMI1 | 8/25/2017

Event Sequences

Use Case: Website Click Streams Analysis



log in → browse products → checkout

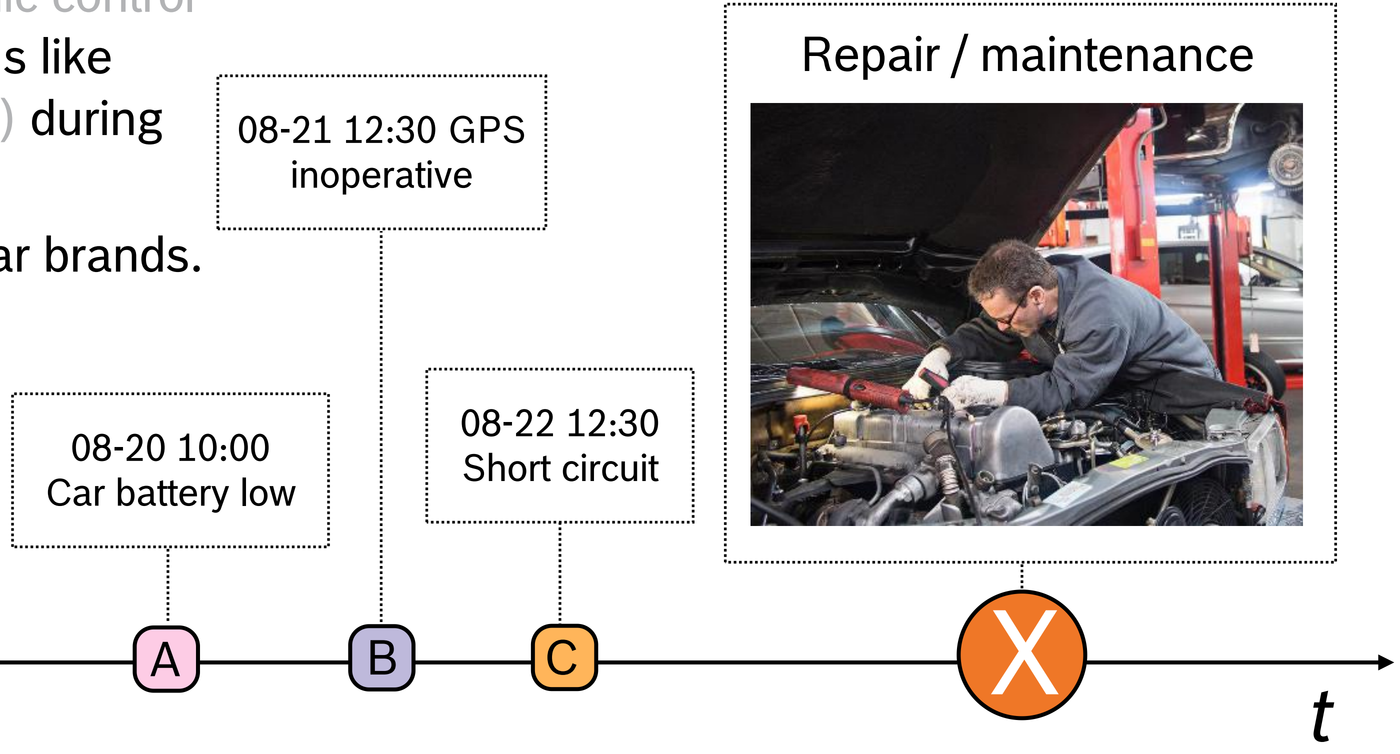
Understand customer behavior

Adjust UI design & improve customer experience

BOSCH

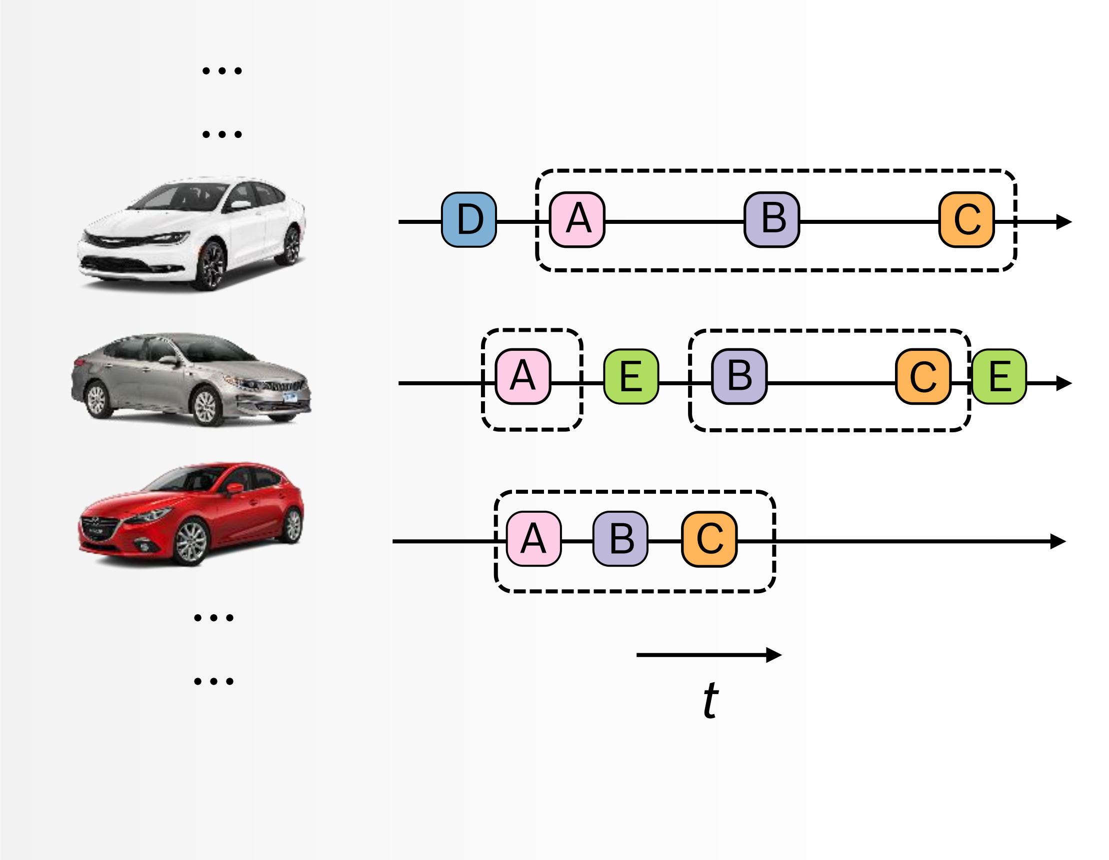
Event Sequences
Use Case: Car Faults Analysis

- Car modules like ECUs (electronic control units) / sensors emits fault signals like DTCs (diagnostics trouble codes) during operation.
- Fault data is archived for most car brands.



BOSCH

Event Sequences
Use Case: Car Faults Analysis

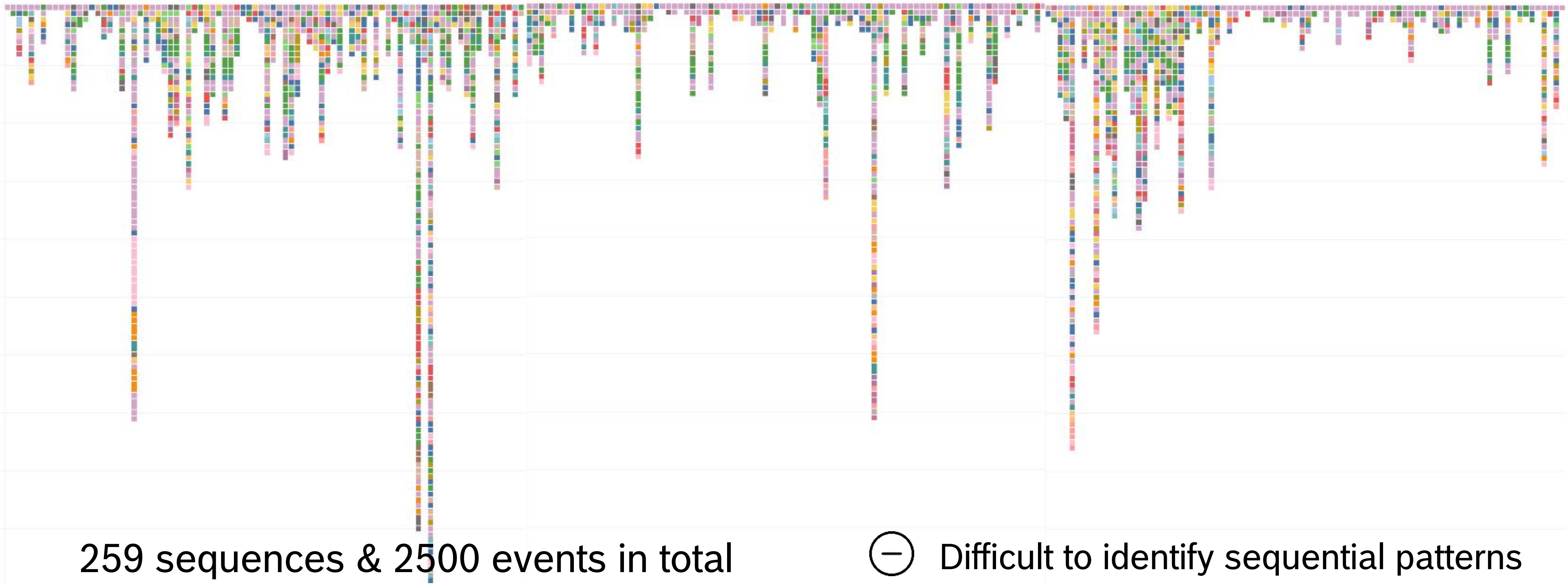


- What are the **typical development paths of faults**? (Identify sequential patterns)
- Do cars matched to the same pattern come from the same country? (correlation analysis)

Insights support predictive diagnostics (i.e. identify faults likely to happen in the future).

Better driving experience & warranty cost saving.

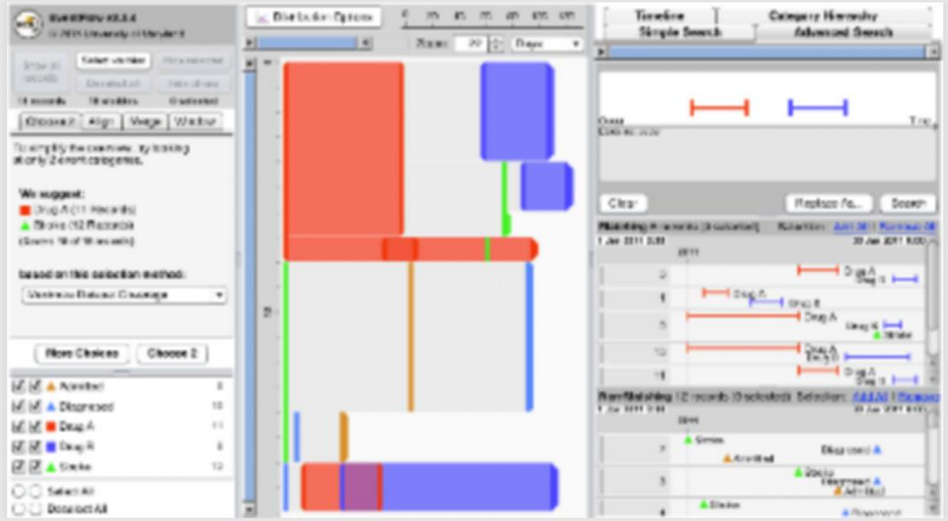
Visualize Event Sequences
Plotting Raw Data



259 sequences & 2500 events in total

⊖ Difficult to identify sequential patterns

Visualizing Event Sequences
Aggregation and Interaction



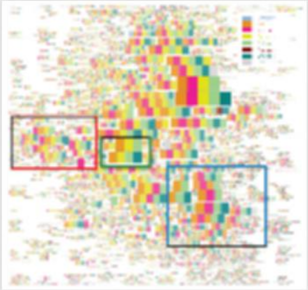
EventFlow
Monroe et. al. 2013

Outflow
Wongsuphasawat and Gotz, 2015

⊕ Provide succinct overview of sequences

⊖ Not robust to noisy data

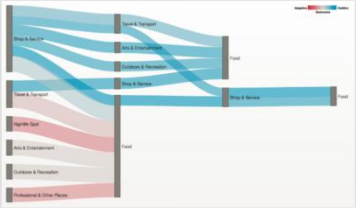
Visualizing Event Sequences
Visual Summary through Sequential Pattern Mining / Clustering



Unsupervised clickstream clustering, *Wang et. al. 2016*

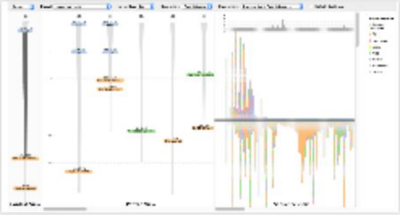
Visual cluster exploration, *Wei et. al. 2012*

- Sequence Clustering
  - ⊕ Robust to noisy data
  - ⊖ Interpretation of clusters: How to characterize each sequence cluster



Frequence, *Perer and Wang, 2014*

Peekquence, Kwon *et. al. 2016*



Patterns&Sequences,

- Sequential Pattern Mining
  - ⊕ Interpretable algorithmic parameters and results
  - ⊖ Large number of patterns: Need to be pruned based on heuristics

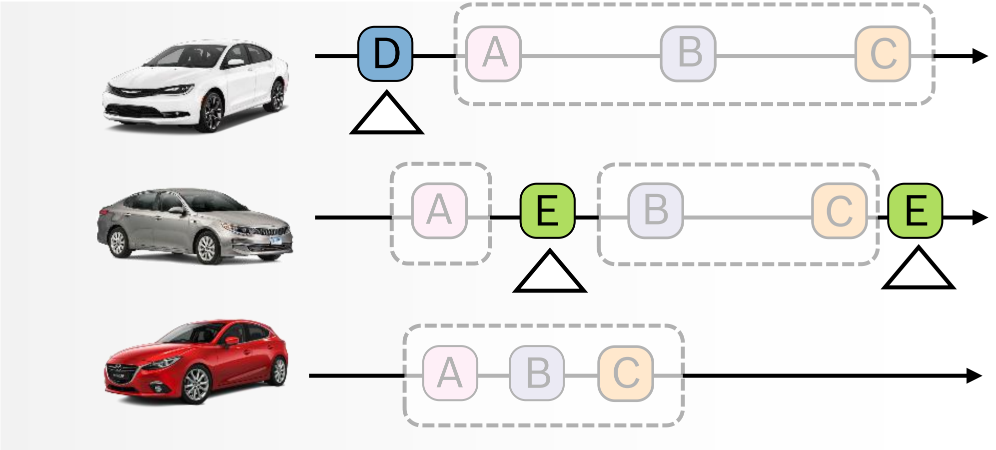
ents

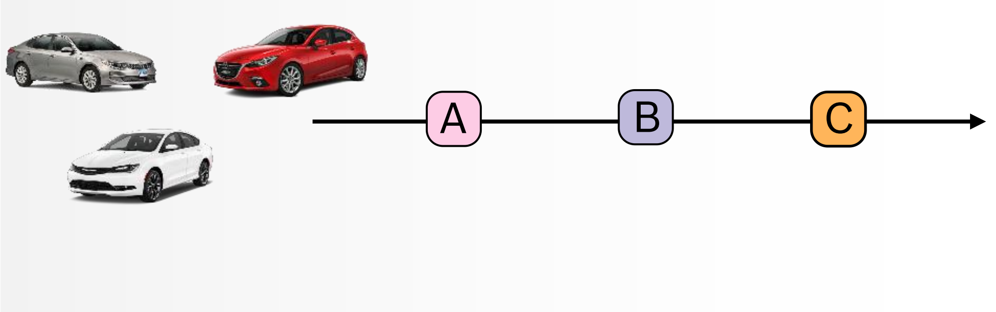
We need to have an **interpretable**, **noise tolerant**, **principled** approach for event sequence summarization.

OUR APPROACH

Our Approach – Sequence Synopsis
Overview

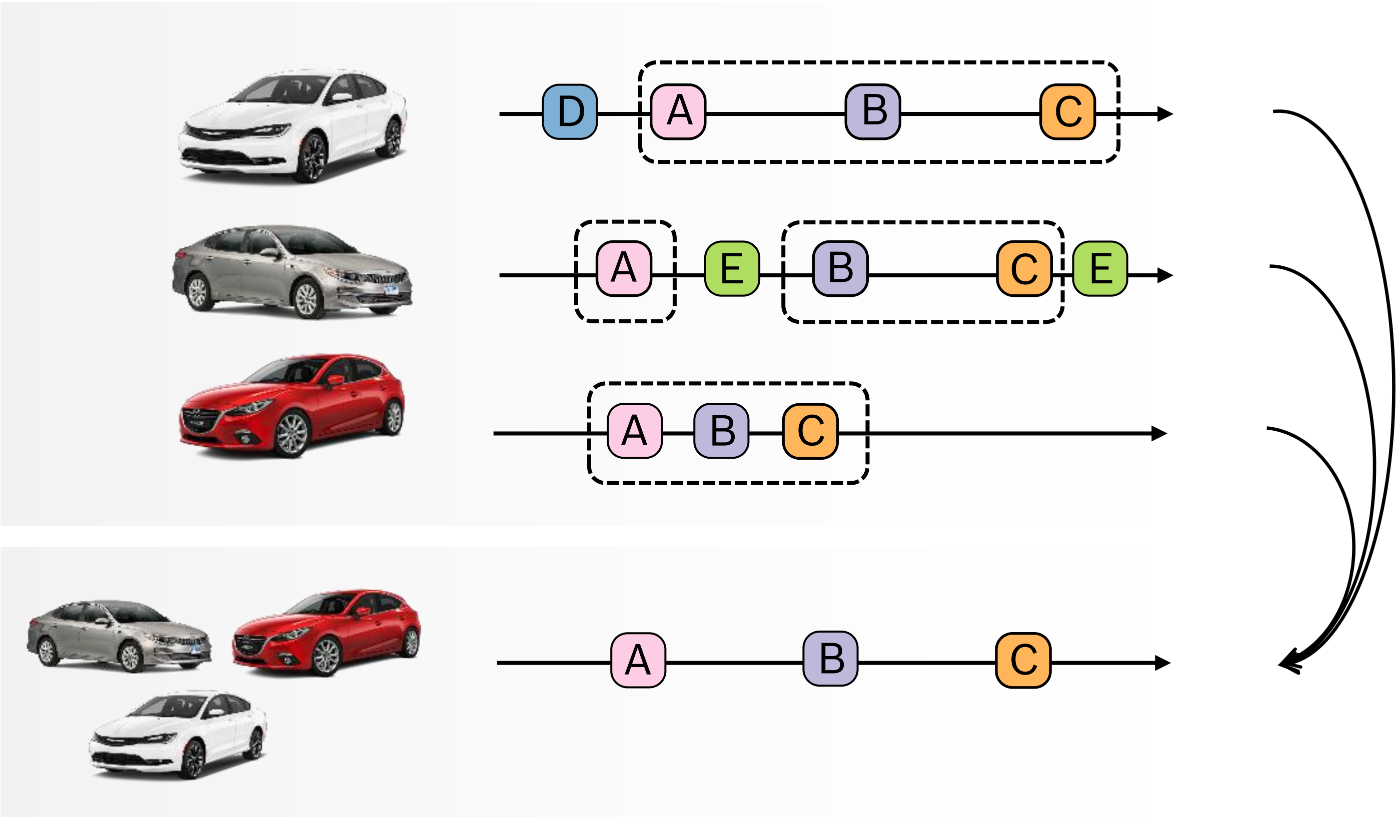
- Two-part representation of event sequences as lossless compression of the data
- Optimal pattern set selection for visual summary based on the Minimum Description Length (MDL) principle
  - Optimization algorithm
  - Speedup with locality sensitive hashing





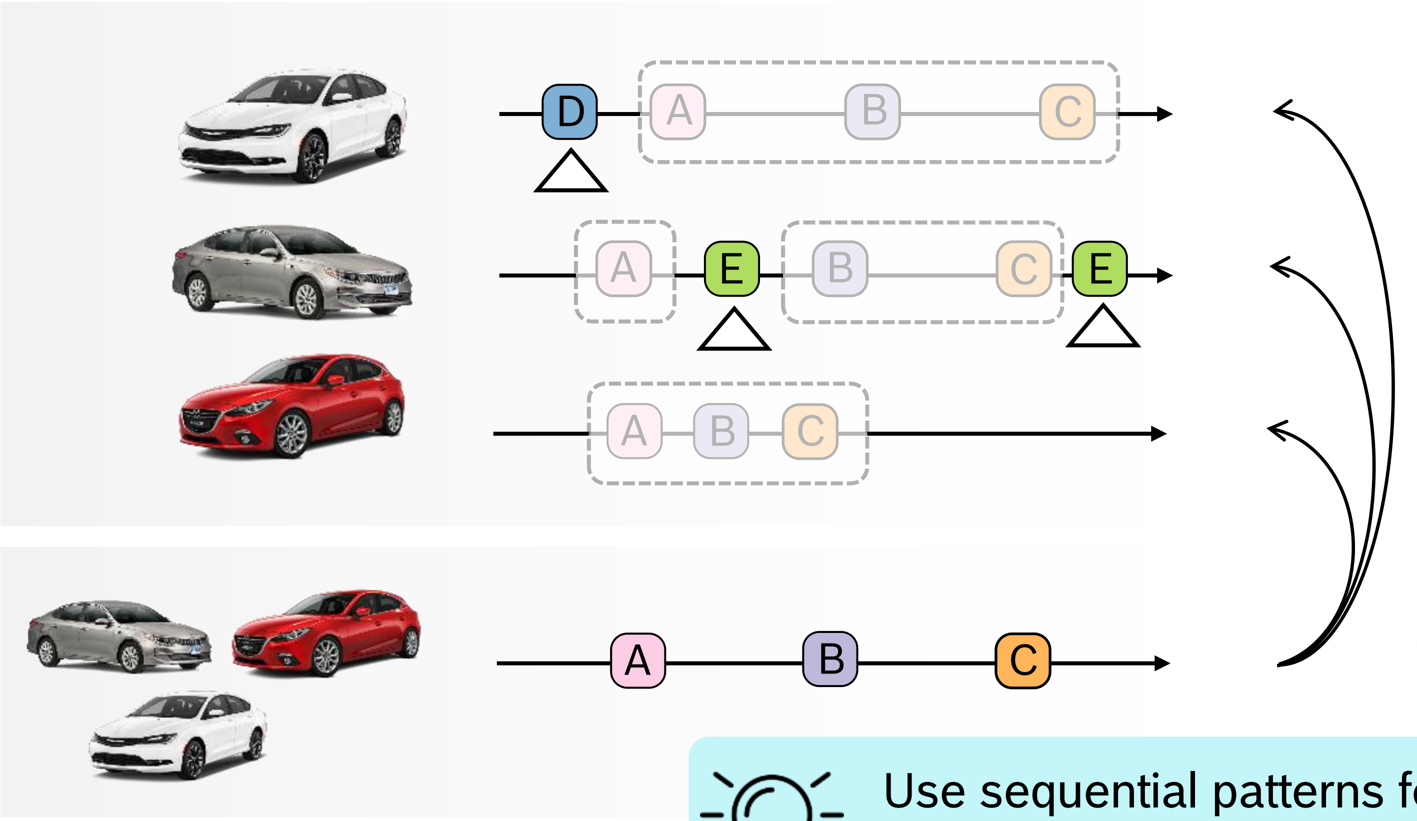
Our Approach – Sequence Synopsis

Two-Part Representation of Event Sequences



*Representative **pattern** summarizes multiple sequences.*

Our Approach – Sequence Synopsis
Two-Part Representation of Event Sequences



Corrections *- event* ***insertions*** *(edits) recover the original sequences from the pattern.*

Representative ***pattern*** *summarizes multiple sequences.*

Use sequential patterns for **visual summary**.

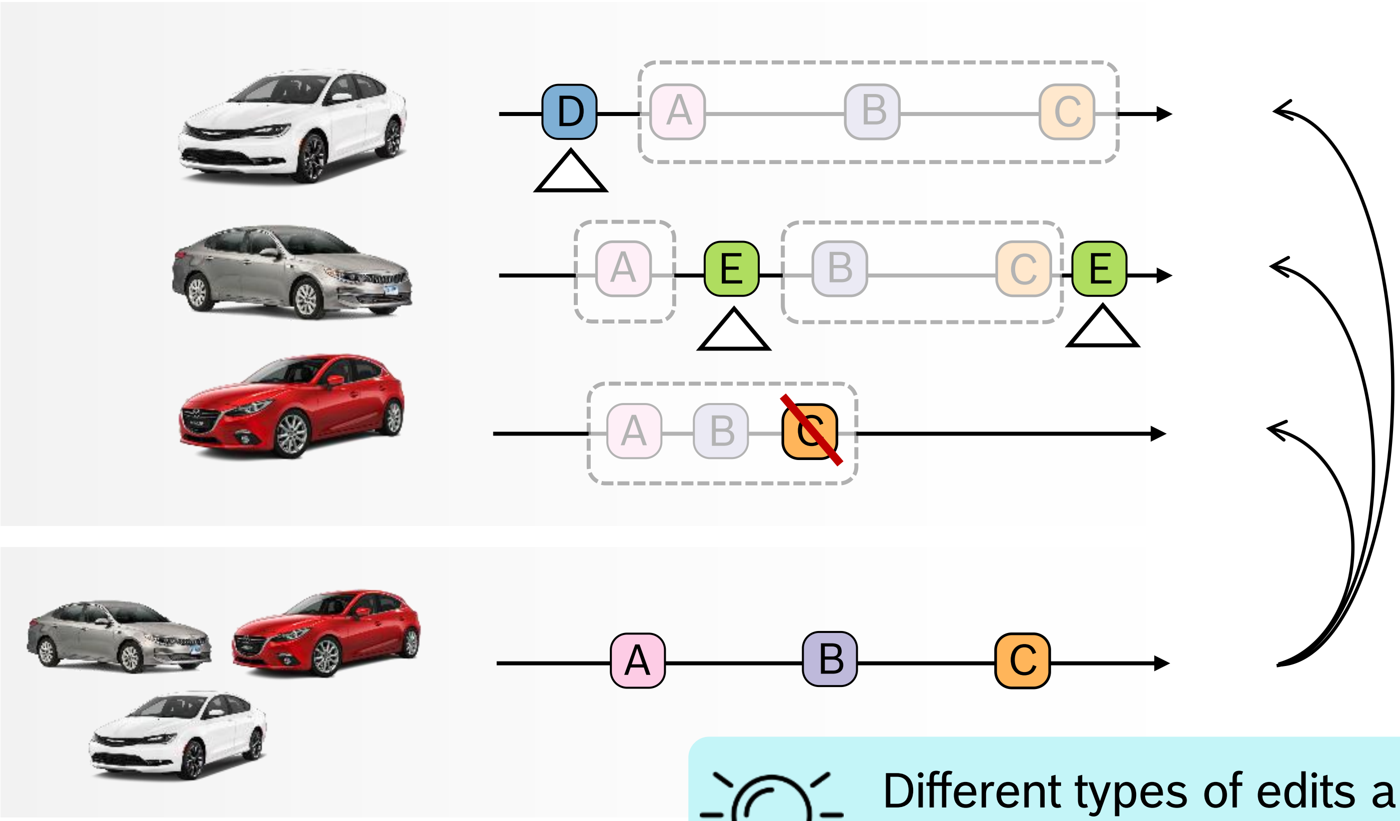
Model **information loss** with the required edits (corrections).

 Research and Technology Center North America | CR/RTC1.4-NA | 8/25/2017

BOSCH

Our Approach – Sequence Synopsis

Two-Part Representation of Event Sequences



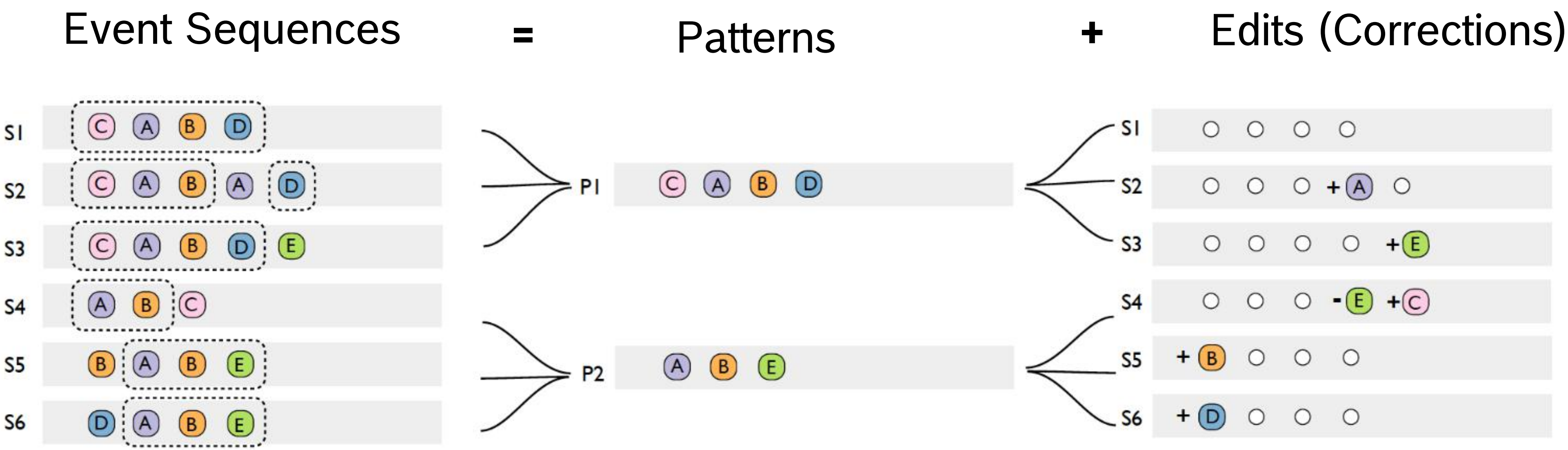
*Event **deletion** is another possible type of edit.*

*Representative **pattern** summarizes multiple sequences.*

Different types of edits allow different variations from the pattern. Enable **noise tolerant & robust** pattern matching.

BOSCH

Our Approach – Sequence Synopsis
Two-Part Representation of Event Sequences



What can be considered as a good set of patterns to summarize a collection of event sequences?

BOSCH

Our Approach – Sequence Synopsis
The Minimum Description Length (MDL) Principle

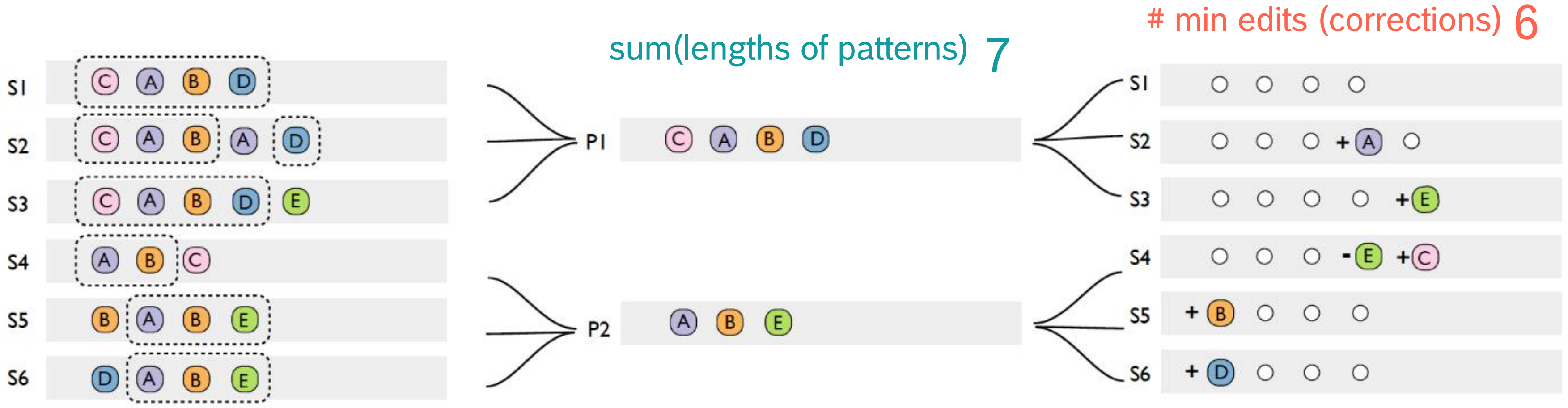
- The best model (or hypothesis) of a data set should minimize its **total description length**:

$$L = \underbrace{L(M)}_{\text{Model description length}} + \underbrace{L(D|M)}_{\text{Data description length with the help of the model}}$$

- Widely used **information-theoretic** criteria for model selection
- Introduced by Jorma Rissanen in 1978
- Formalizes "**Occam's Razor**"

Our Approach – Sequence Synopsis
Description Length of Event Sequences

$$\underbrace{L = L(M)}_{} + \underbrace{L(D|M)}_{} \longrightarrow L(\mathcal{P}, f) = \underbrace{\Sigma_{P\in\mathcal{P}}\, len(P)}_{} + \alpha\, \underbrace{\Sigma_{S\in\mathcal{S}} \|edits(S, f(S))\|}_{} + \lambda\|\mathcal{P}\|$$



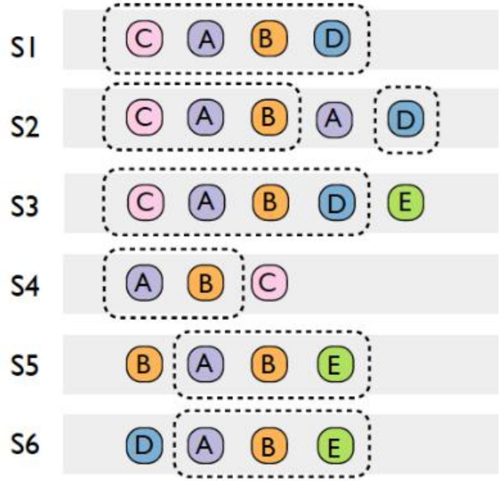
Trade-off between **reducing visual complexity** & **minimizing information loss.**

Our Approach – Sequence Synopsis
Optimize Description Length for the Best Set of Patterns

- Basic Idea: **iteratively find & merge** two groups of sequences with maximum description length reduction
- How to calculate description length reduction?
  - Find **representative sequence** for the merged group
  - Calculate the **minimum number of edits** (insertion, deletion, swapping event positions) needed to transform the representative sequence to the individual sequence in the merged group
    - Assuming insertion & deletion are allowed. Longest common subsequence (LCS) algorithm can be applied to calculate min #edits
  - **Sum up** the description length

Our Approach – Sequence Synopsis
Optimize Description Length for Best Set of Patterns

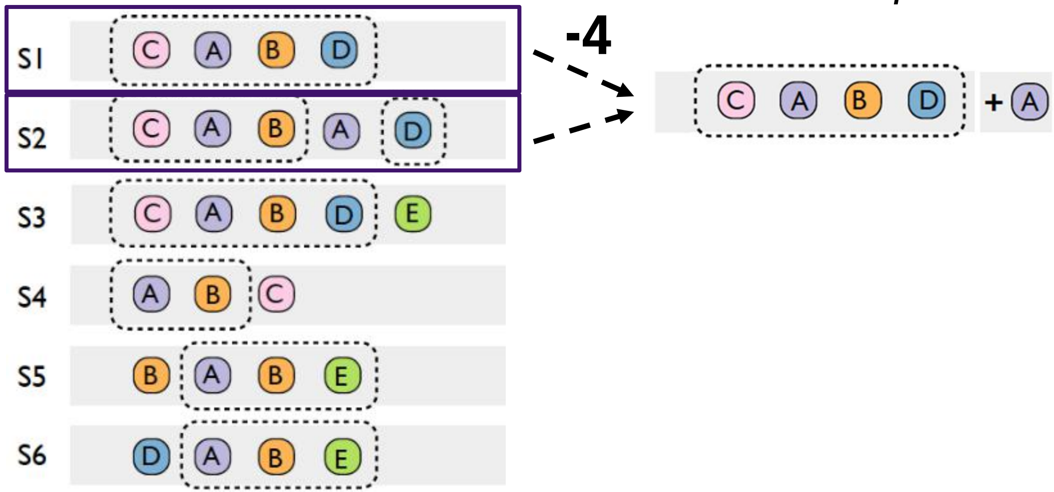
- Basic Idea: **iteratively find & merge** two groups of sequences with maximum description length reduction



Our Approach – Sequence Synopsis
Optimize Description Length for Best Set of Patterns

- Basic Idea: **iteratively find & merge** two groups of sequences with maximum description length reduction

Calculate description length reduction

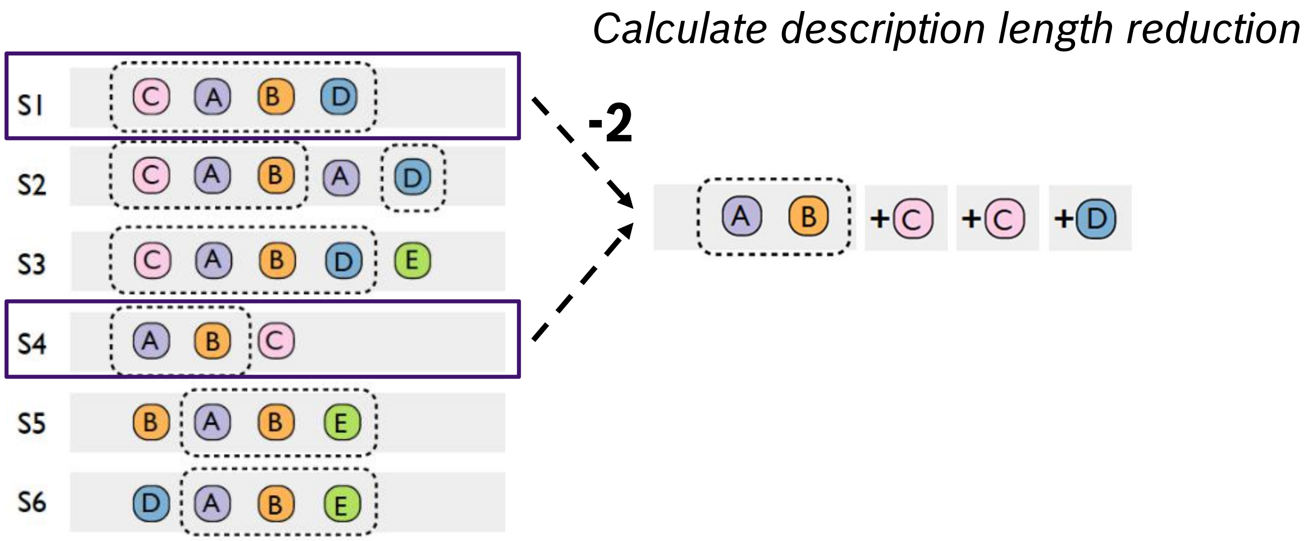


Try to merge each pair of sequences/patterns

BOSCH

Our Approach – Sequence Synopsis
Optimize Description Length for Best Set of Patterns

- Basic Idea: **iteratively find & merge** two groups of sequences with maximum description length reduction



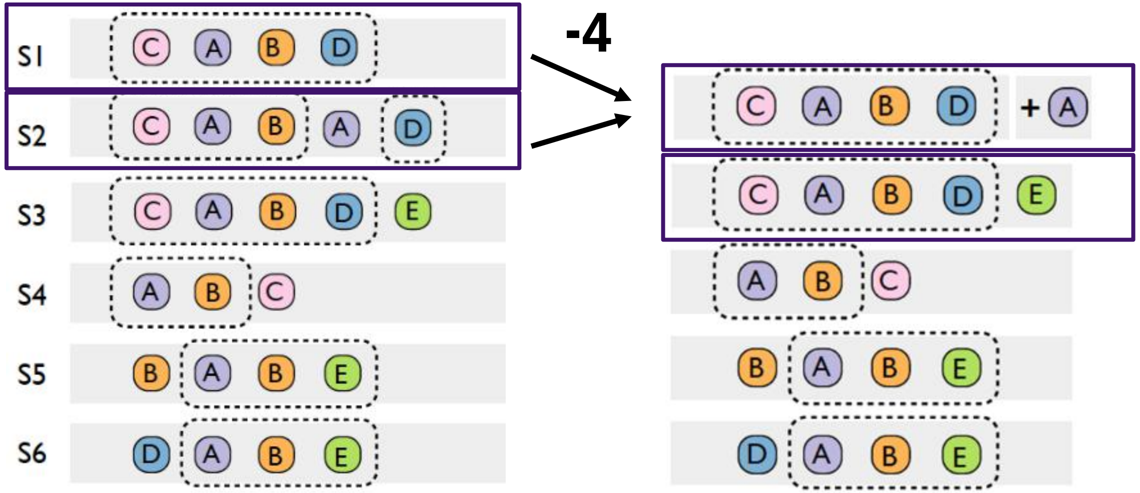
Try to merge each pair of sequences/patterns

 Research and Technology Center North America | CR/RTC1.4-NA | 8/25/2017

BOSCH

Our Approach – Sequence Synopsis
Optimize Description Length for Best Set of Patterns

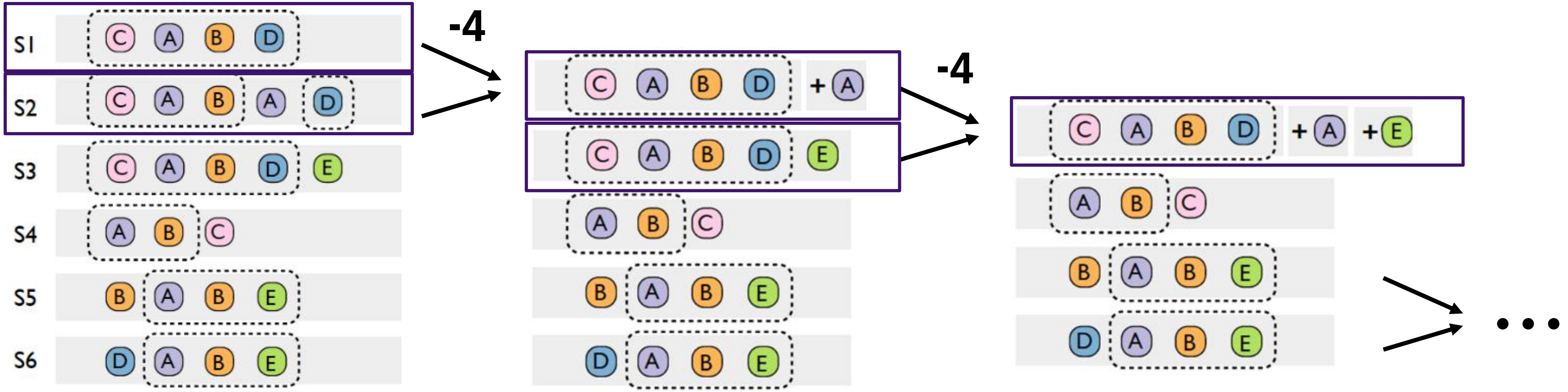
- Basic Idea: **iteratively find & merge** two groups of sequences with maximum description length reduction



Merge the pair with maximum description length reduction

Our Approach – Sequence Synopsis
Optimize Description Length for Best Set of Patterns

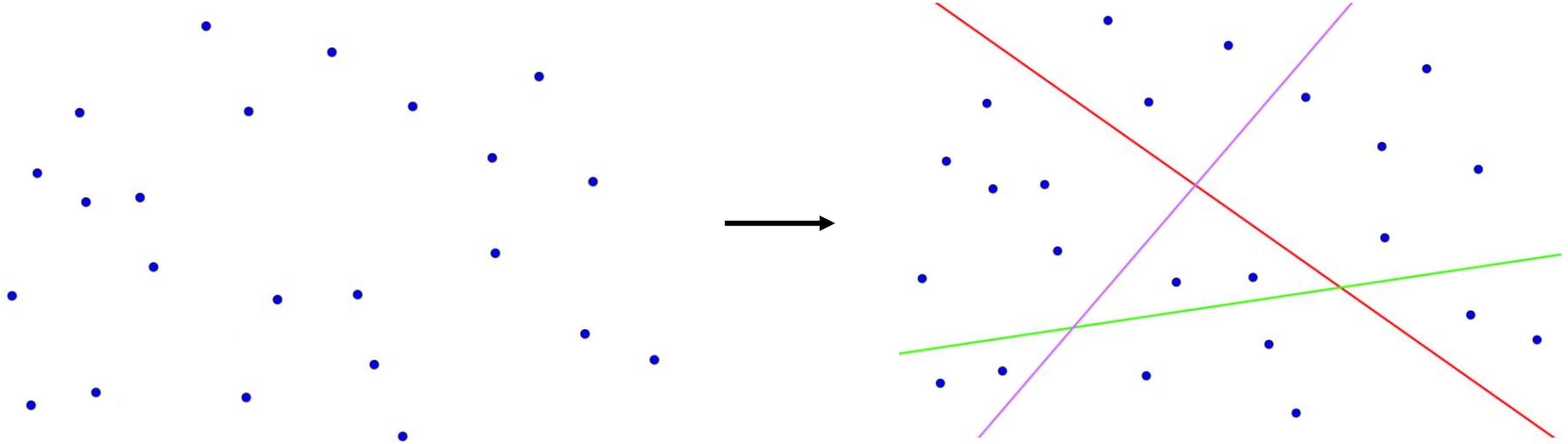
- Basic Idea: **iteratively find & merge** two groups of sequences with maximum description length reduction



Need to perform pairwise comparison at each iteration

Our Approach – Sequence Synopsis
Algorithm Speedup through Locality Sensitive Hashing (LSH)

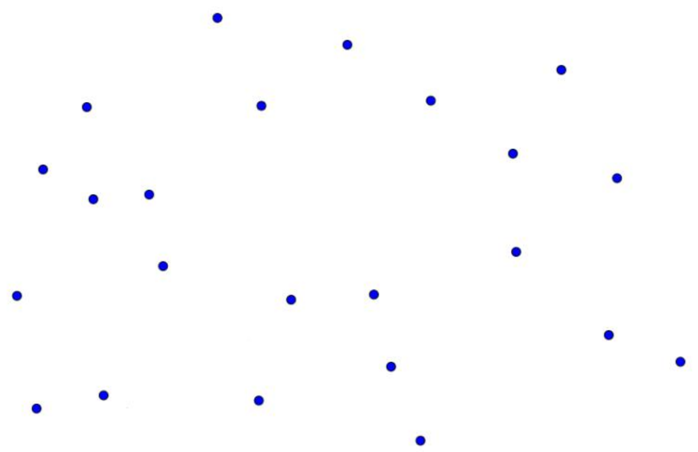
- **Bottleneck of the approach:** find best pair of event sequence groups to merge
- **Locality sensitive hashing**: algorithm for fast approximate neighbor search



BOSCH

Our Approach – Sequence Synopsis
Algorithm Speedup through Locality Sensitive Hashing (LSH)

- **Bottleneck of the approach:** find best pair of event sequence groups to merge
- **Locality sensitive hashing:** algorithm for fast approximate neighbor search

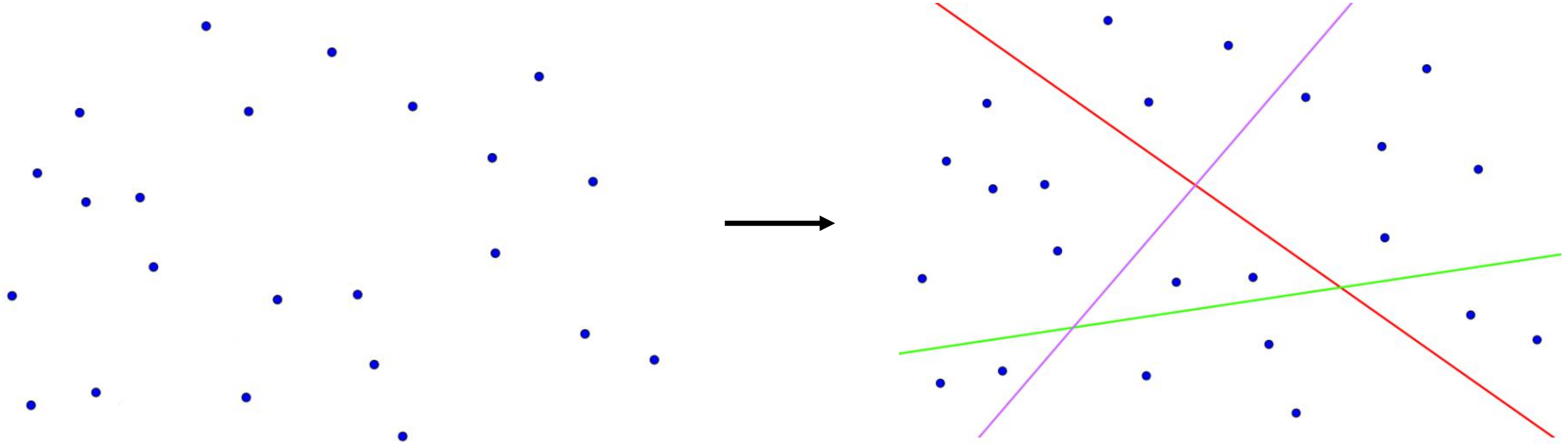


Simplified similarity measure with set relation

Research and Technology Center North America | CR/RTC1.4-NA | 8/25/2017

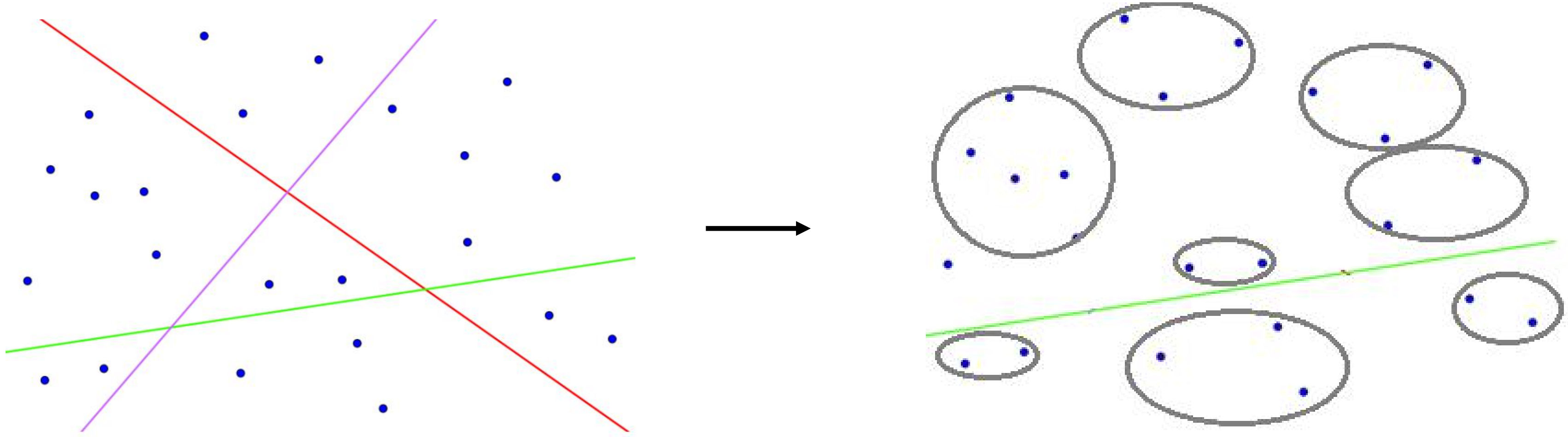
Our Approach – Sequence Synopsis
Algorithm Speedup through Locality Sensitive Hashing (LSH)

- **Bottleneck of the approach:** find best pair of event sequence groups to merge
- **Locality sensitive hashing**: algorithm for fast approximate neighbor search



Our Approach – Sequence Synopsis
Algorithm Speedup through Locality Sensitive Hashing (LSH)

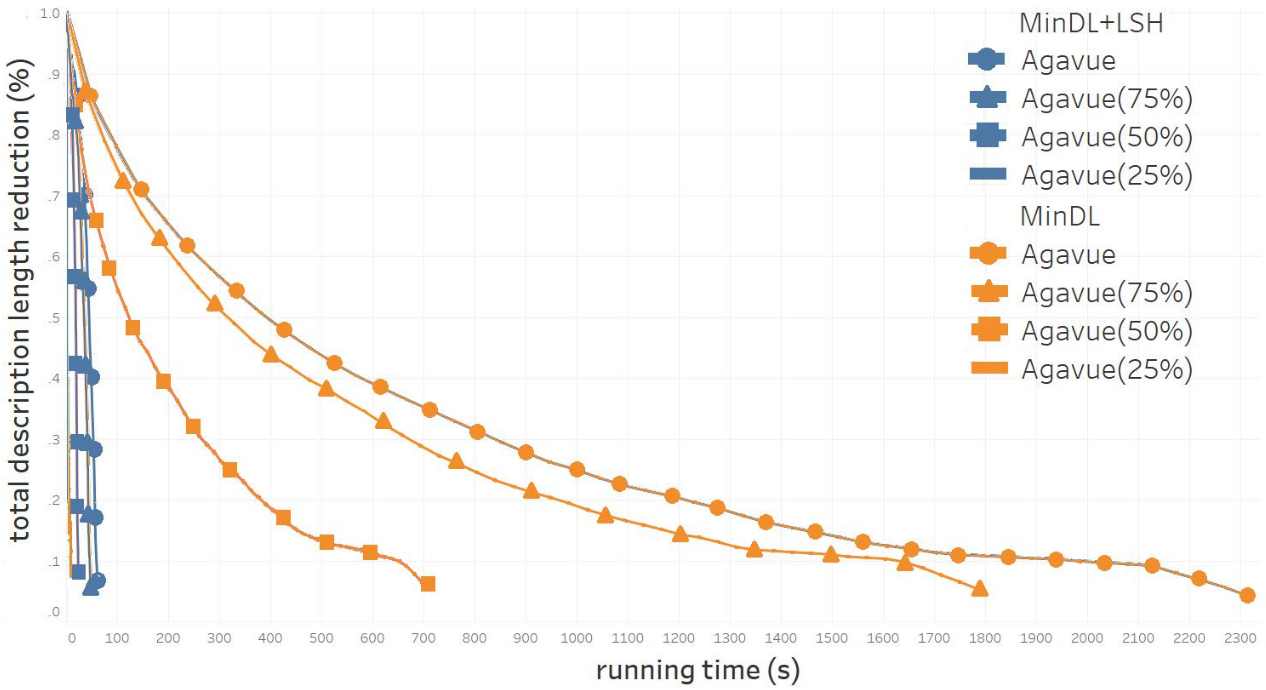
- **Bottleneck of the approach:** find best pair of event sequence groups to merge
- **Locality sensitive hashing:** algorithm for fast approximate neighbor search



20x ~ 50x speed gain

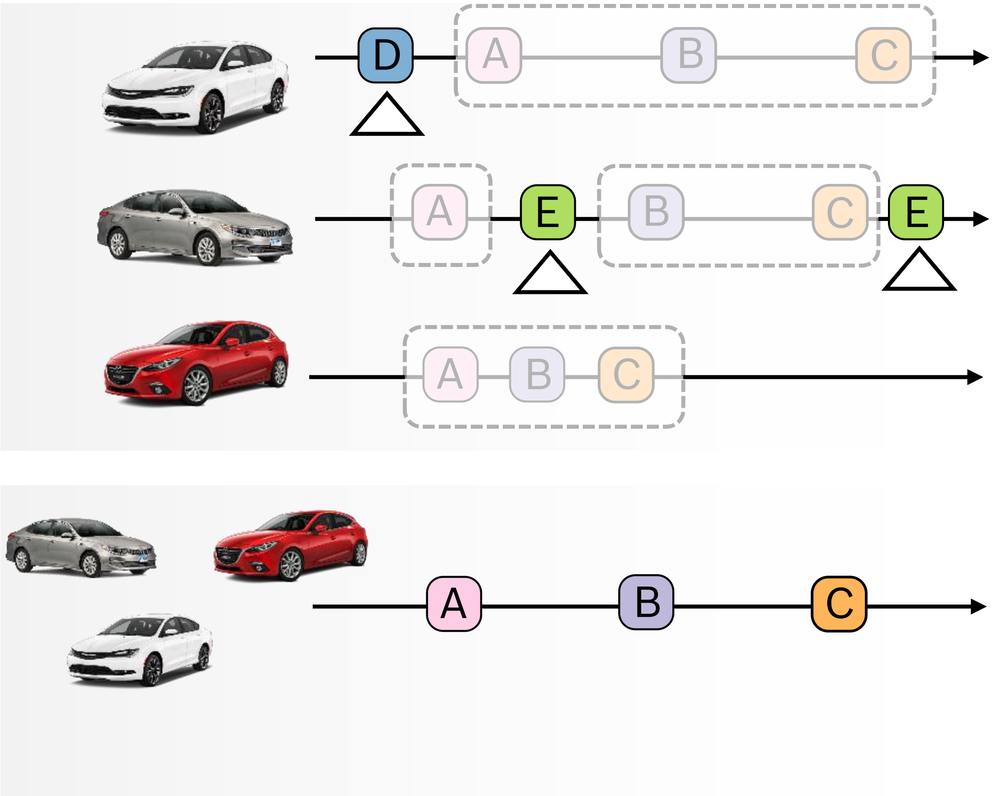
Our Approach – Sequence Synopsis
Algorithm Speedup through Locality Sensitive Hashing (LSH)

- **Bottleneck of the approach:** find best pair of event sequence groups to merge
- **Locality sensitive hashing**: algorithm for fast approximate neighbor search



Our Approach – Sequence Synopsis
Advantages

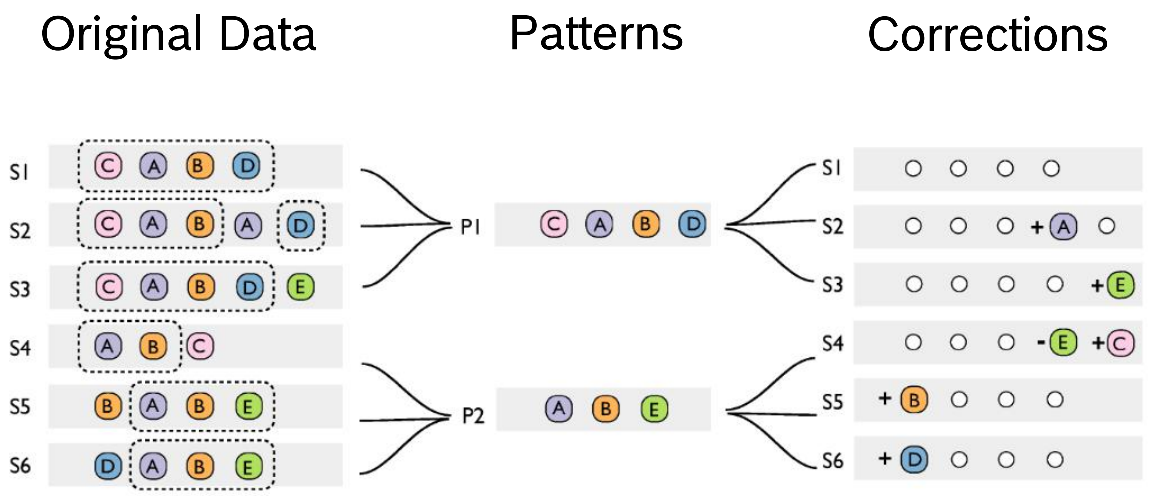
- **Simultaneous** event sequence clustering and pattern extraction
- **Soft constraints** on pattern matching, therefore robust to noisy data
- **Generalizability**: possibility to include different sequence editing operations (e.g. event insertion, deletion, swapping positions)



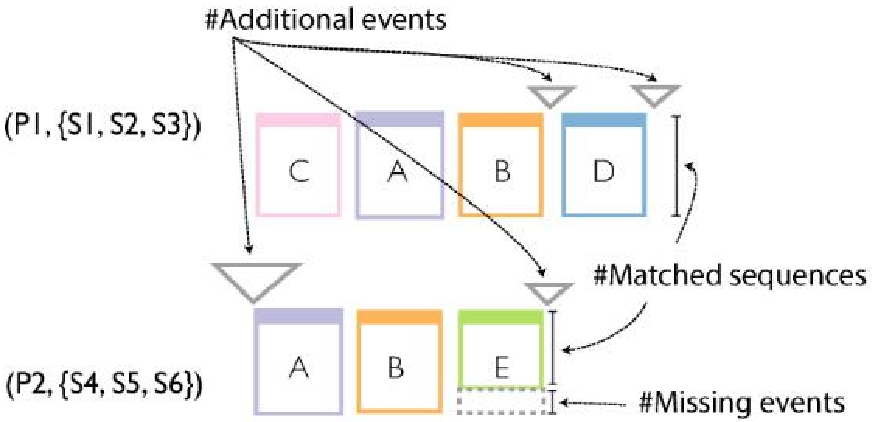
BOSCH

SYSTEM

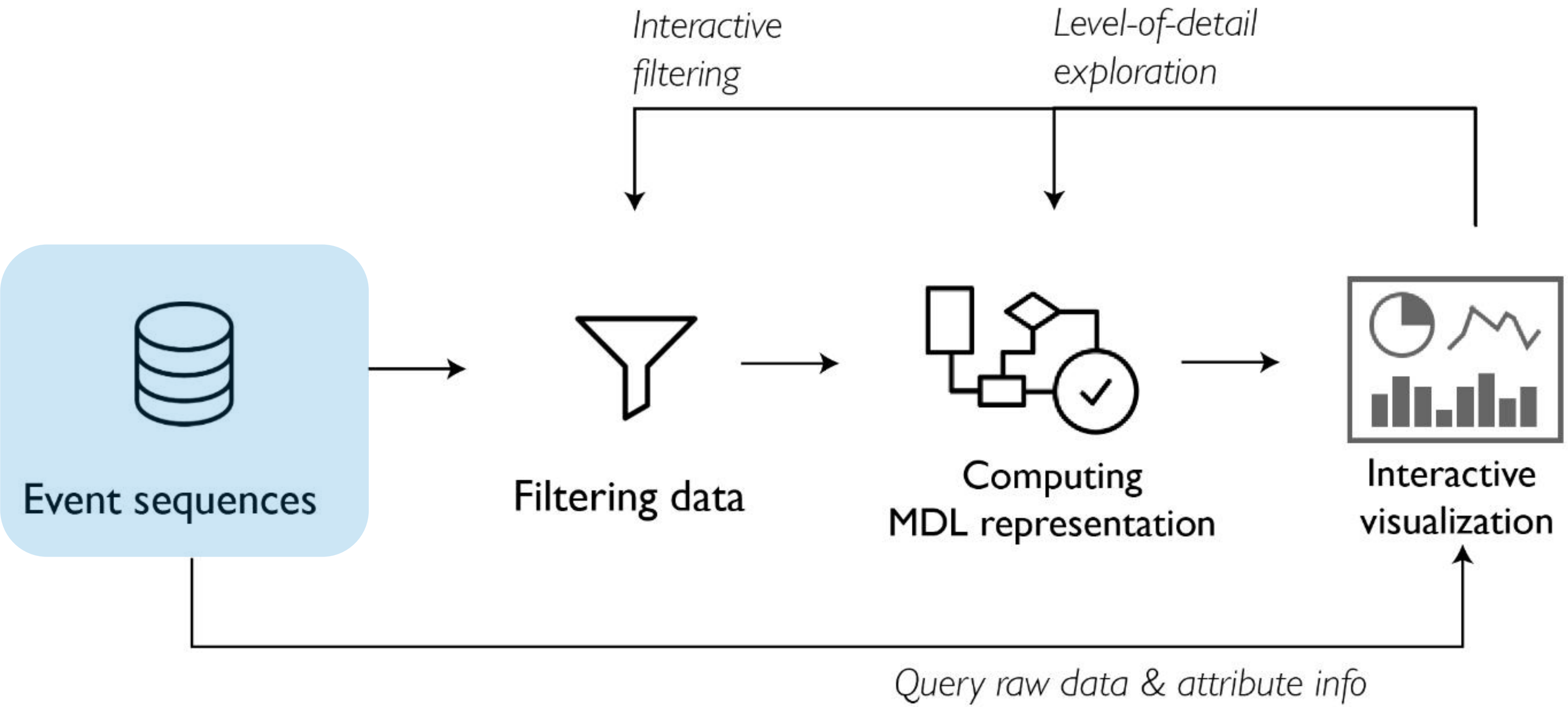
System
Visual Design



Visual Design

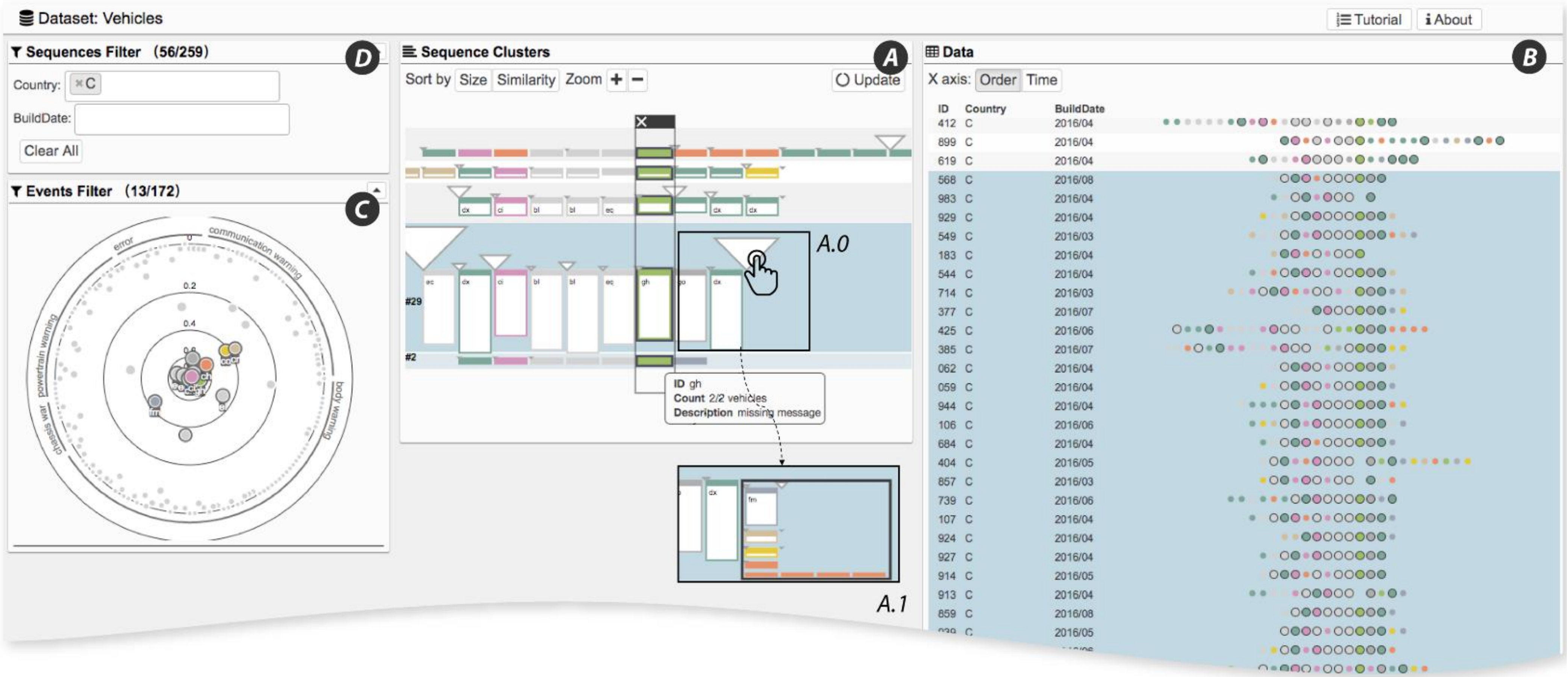


System
Architecture



System

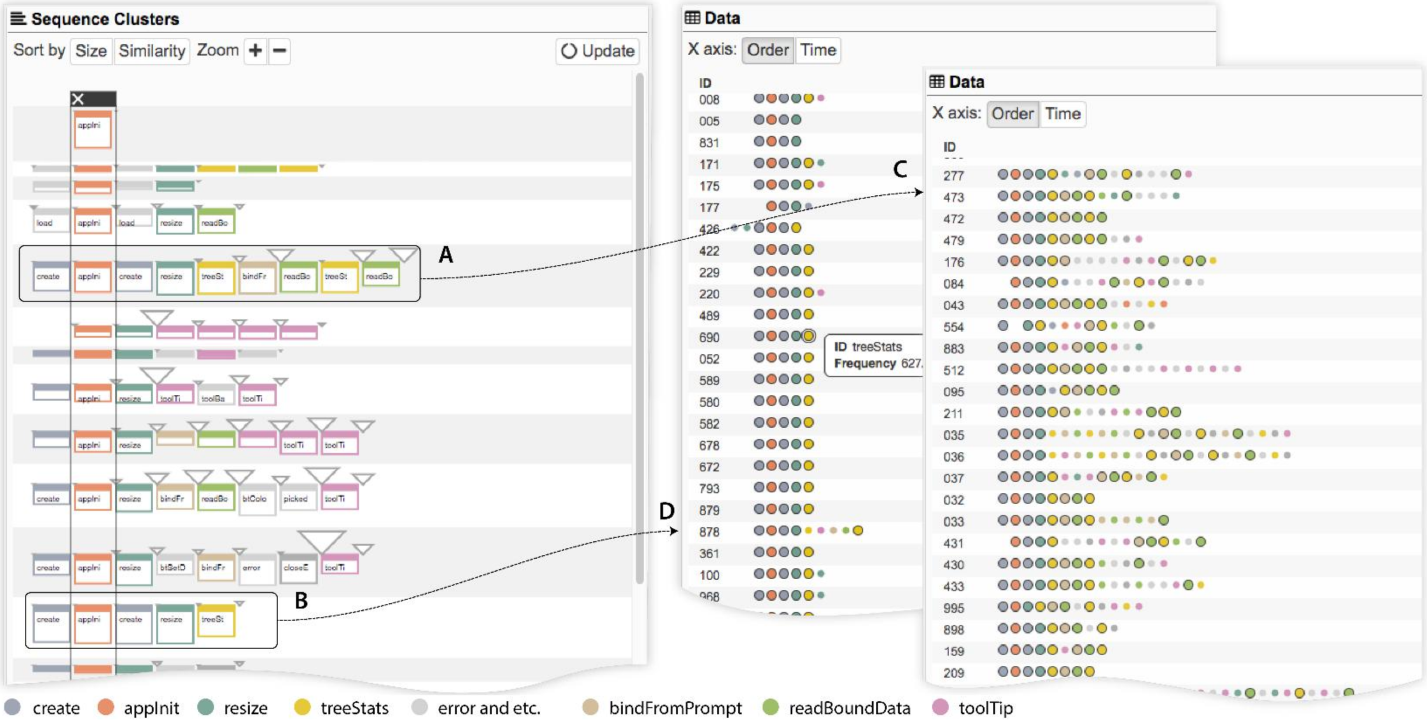
Supportive Views, UI, Case Study – Vehicle Fault Analysis



BOSCH

System
Case Study – Application Log Analysis

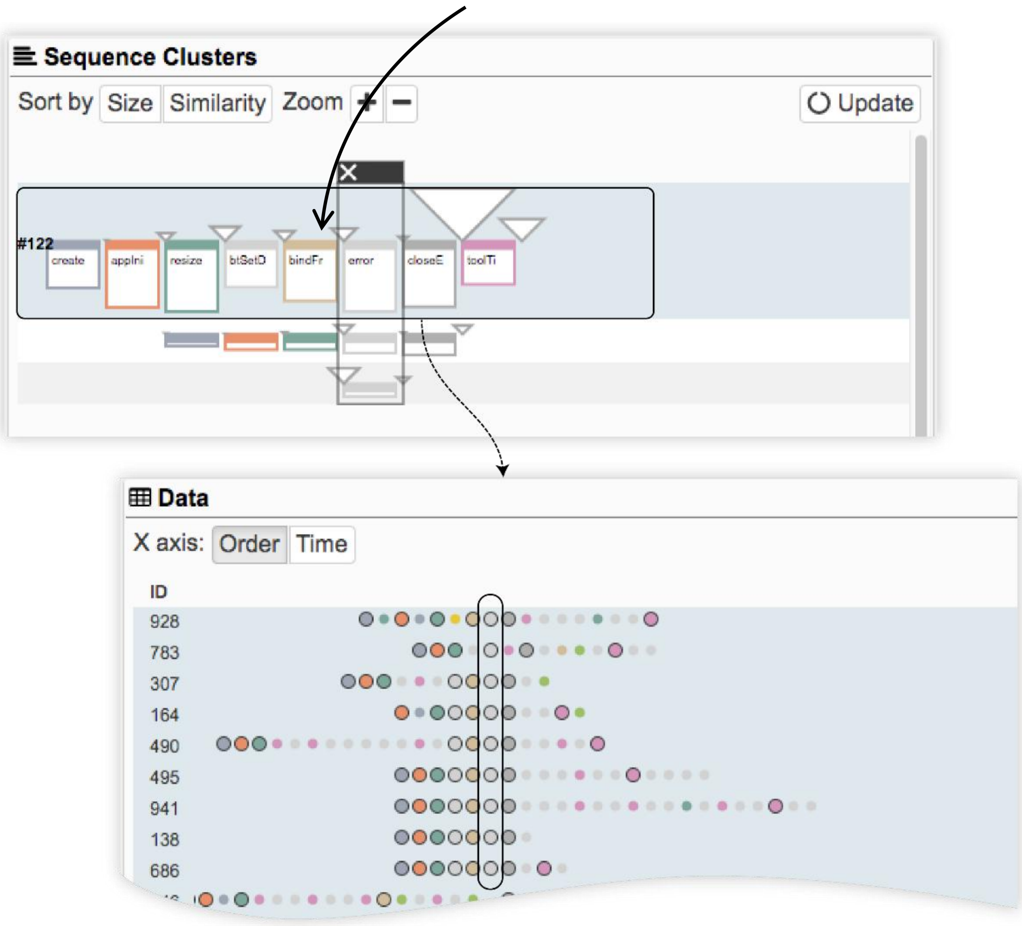
- D. Fisher. Agavue event data sample
- ~2000 user sessions
- Interaction log of using a data visualization application



System
Case Study – Application Log Analysis

- D. Fisher. Agavue event data sample
- ~2000 user sessions
- Interaction log of using a data visualization application

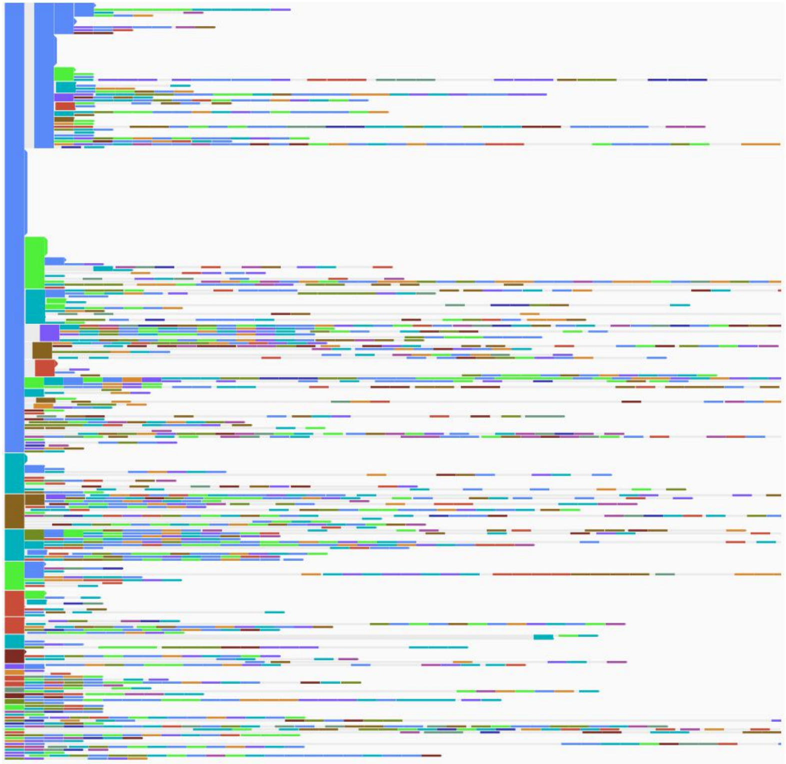
Binding data



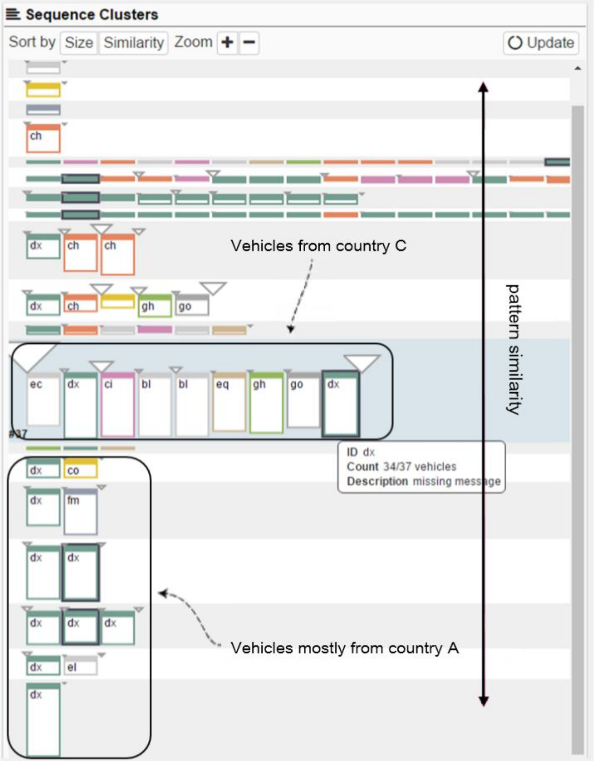
EVALUATION & SUMMARY

Evaluation & Summary
Comparative Experiment

- Vehicle Fault Sequence
- 259 cars & 2500 events



EventFlow
Monroe et. al. 2013

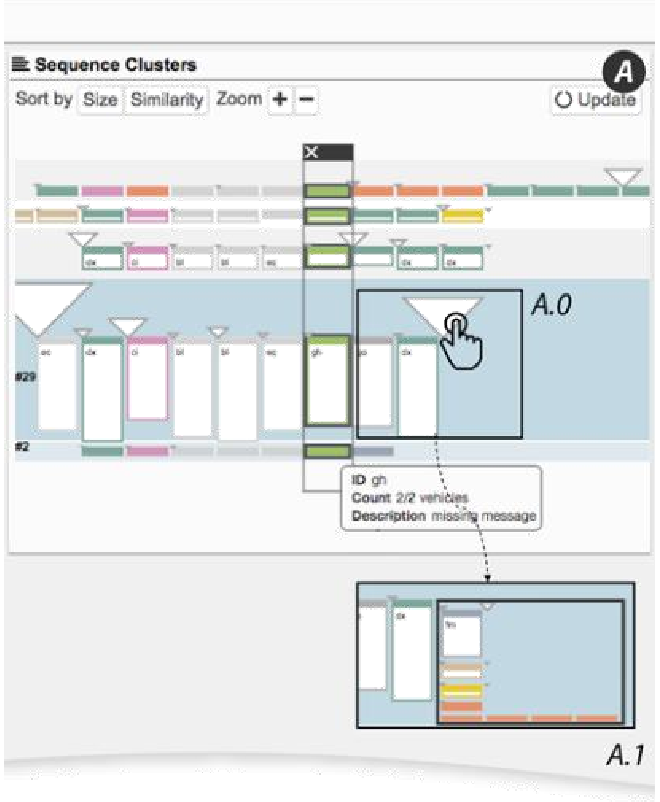


Our method

BOSCH

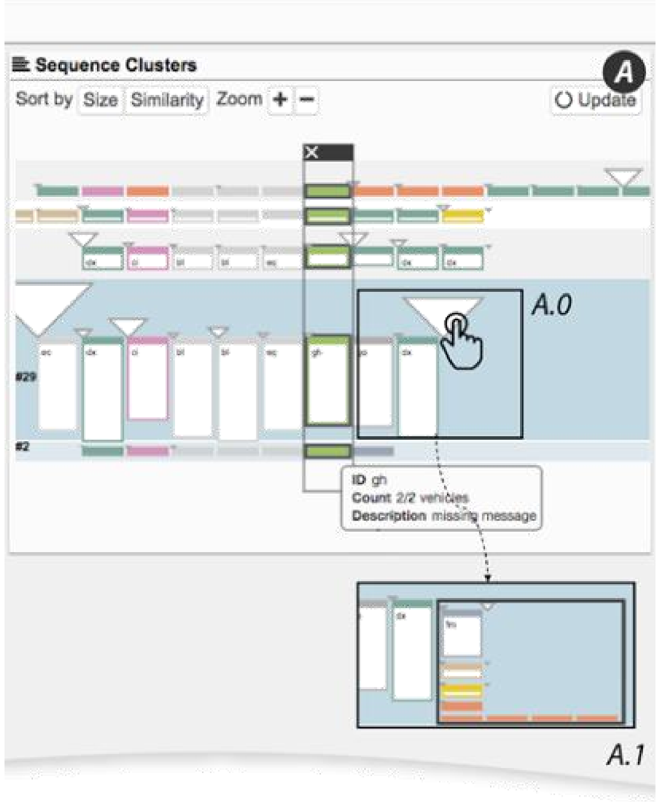
Evaluation & Summary
Contributions

- A new application domain of event sequence visualization
- A generic **two-part representation** of event sequences that:
  - **Quantifies visual complexity & information loss** in visual summaries
  - Combined with the **MDL principle**, defines an optimal set of patterns for summary
- An efficient algorithm to optimize visual summary using LSH
- A visual analytics system that supports interactive analysis of **real-world** event sequences from **different application domains**



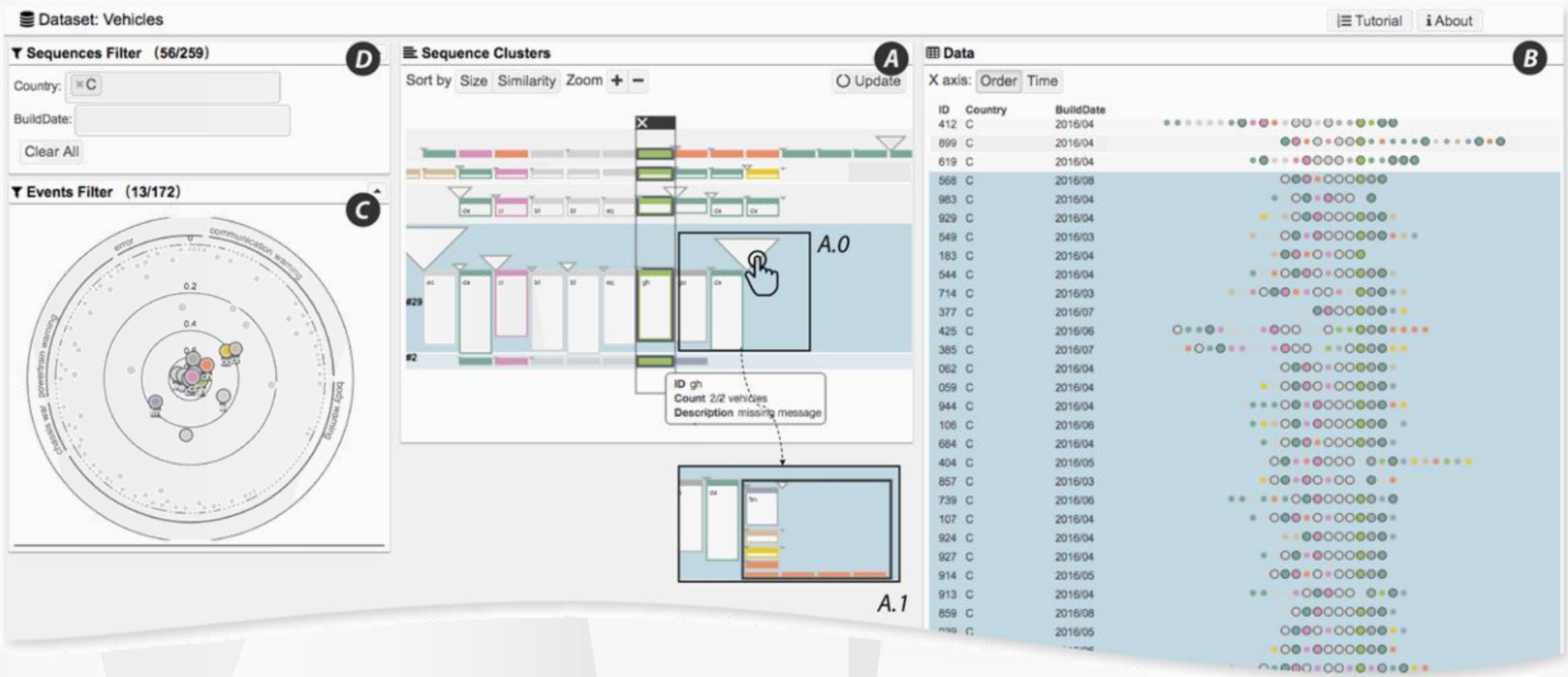
Evaluation & Summary
Future Work

- Revise model representation to discover multiple patterns in a single sequence
- Towards **quantifiable visual designs** by applying the MDL principle to different types of data: graph/networks, time series …



BOSCH

THANK YOU!
Q&A



Dataset: Vehicles
Tutorial
About
Sequences Filter (56/259)
D
Country:
C
BuildDate:
Clear All
Events Filter (13/172)
C
Sequence Clusters
A
Sort by Size Similarity Zoom
Update
A.0
A.1
ID gh
Count 2/2 vehicles
Description missing message
Data
B
X axis: Order Time
ID Country BuildDate