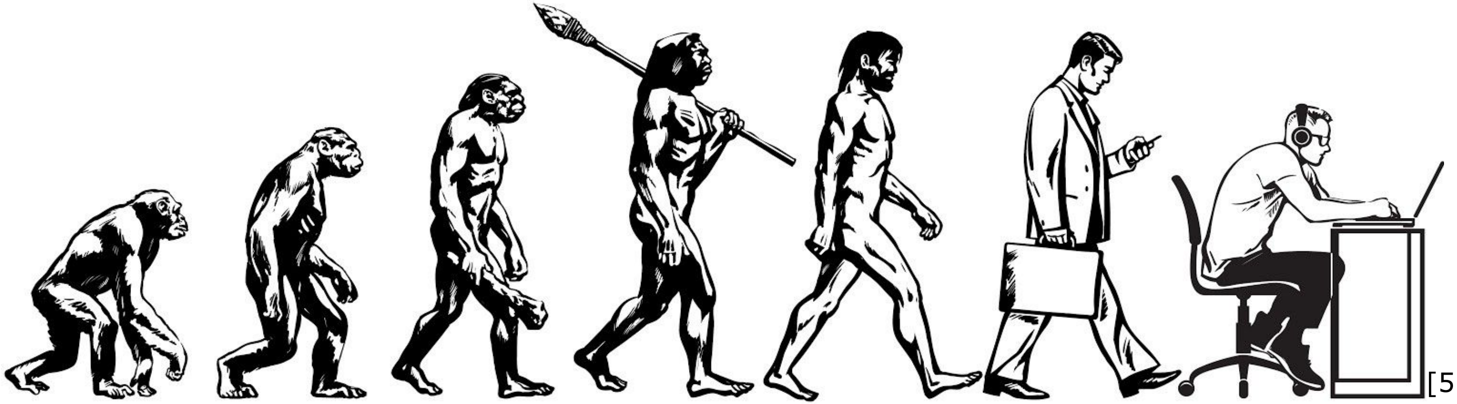

[5]

Discovering Process Models from unlabeled Event Logs

Pascal Schulze, Anjo Seidel

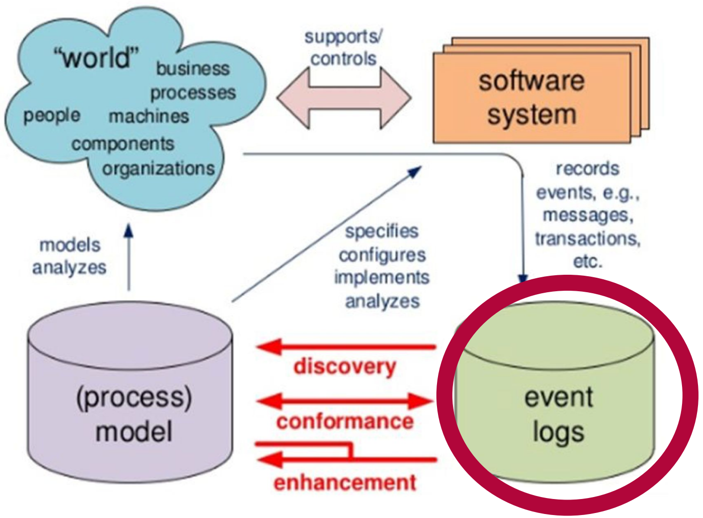
Data Extraction for Process Mining (ST-2020)

17.06.2020

The Problem

Unlabeled Event Log

- Events can be mapped to activities
- No knowledge of a data model
- Relation between events and cases is NOT known



How can we still do process mining with these logs?

The Approach

- If Case IDs are NOT given, try to **predict them**!
- Diogo R. Ferreira, Daniel Gillblad:
  "Discovering Process Models from Unlabelled Event Logs" [2]

Input: Unlabeled Event Log

A B D A E A C C D D A E F F F C D F . . .

Output: Estimated/Predicted Case ID

1 1 1 2 1 3 3 2 2 3 4 3 3 1 2 4 4 4 . .

The Approach - Estimating Case IDs

- Estimation based on a given Probability Matrix *M*
- Assign an event to the trace: - with the highest probability
  - which was last active

Log: A B D A E A C C D D A E F F F C D F ...

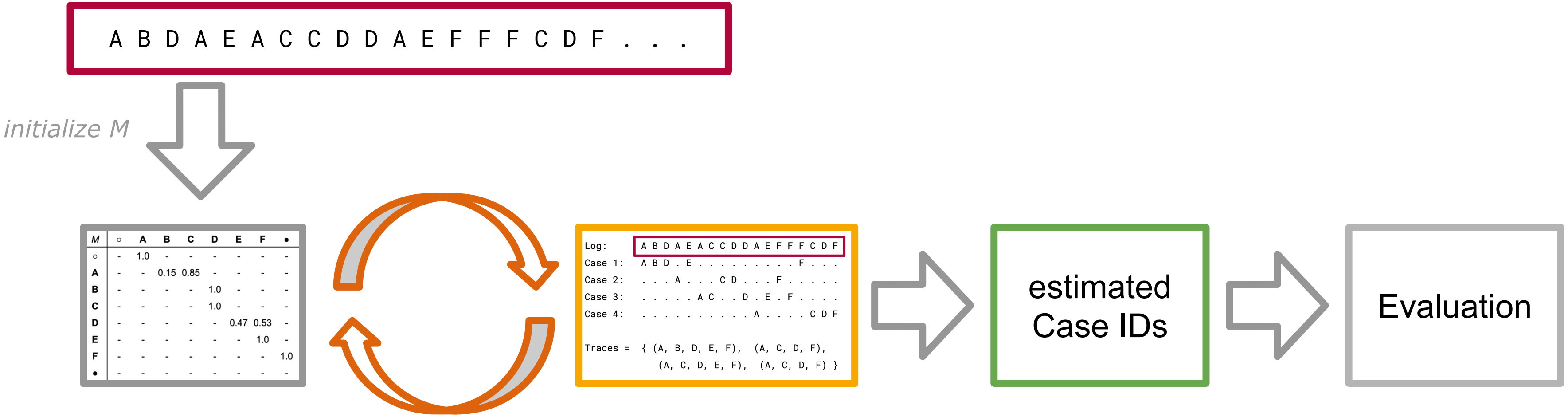
Case 1:

IDs:

M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.25	0.75	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.5	0.5	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

The Approach - Iterative Process

- M gets initialized: random *or*
  based on direct successorship (M+)
- Iteratively estimate Case IDs with M and estimate M with given Case IDs

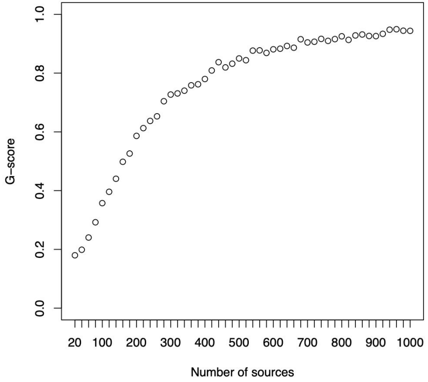


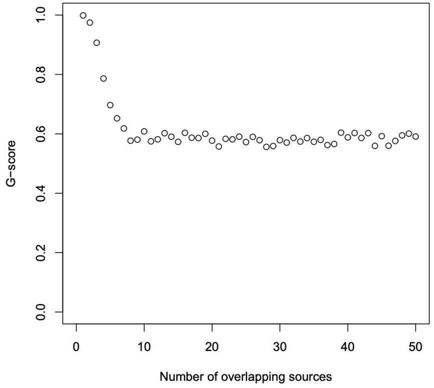
The Approach - Restrictions

- Metrics for Probability Matrix
  - M+ only with direct successorship

- Assumptions to the process
  - no loops
  - no parallelism
  - …

- Greedy Algorithm
  - “iterative Expectation–Maximization procedure”
  - always pick most likely candidates
  - leads to suboptimal solutions





[2]

Research Question

1. Can the results of this approach be improved by using a Genetic Programming Paradigm and other metrics?

2. Can assumptions for this approach be overcome with a Genetic Approach?

New Contribution

Steps

1. Genetic Extension to the iterative Algorithm [3]
   a. Implementation
   b. Evaluation and Comparison of both Approaches

2. Getting rid of assumptions: Testing on different input models
   a. Loops, parallel behaviour, only full recorded end-to-end instances

3. Extending Algorithm
   a. Dynamic model creation instead of greedy rule-based procedure
   b. Different estimation matrices for initialization
      i. Distance
      ii. Global strength of causality

Genetic Algorithm

Genetic Programing Paradigm

Searching for an optimal or at least suitable model among the space of all models, by evolving them, starting from a population of unfit (usually random) ones.

After every (training) epoch → Reproduction:

Selection

- select parents for next generation
- better performing individuals have a higher chance of getting selected

Crossover

- genetic operations to breed new individuals from selected parents

Mutation

- randomly change values

[RECALL] The Approach - Iterative Process [3]

A B D A E A C C D D A E F F F C D F . . .

initialize Matrix (M+)

M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

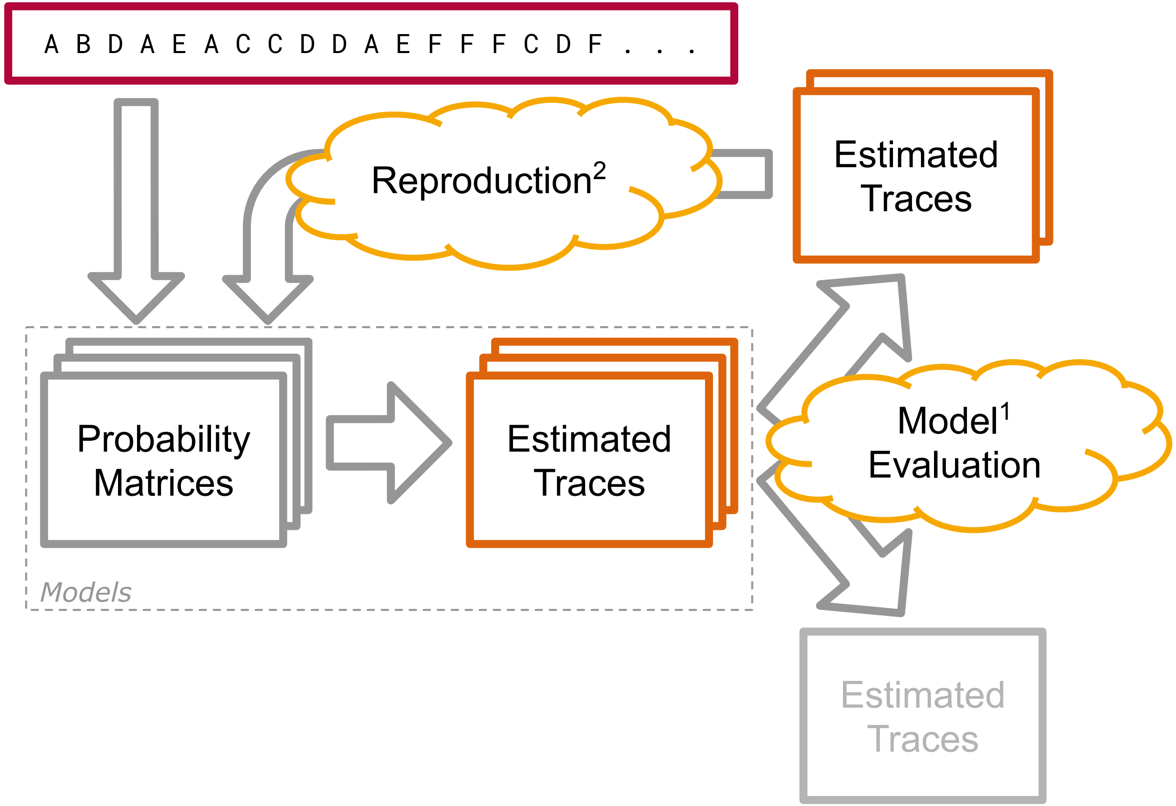
```
Log:     A B D A E A C C D D A E F F F C D F
Case 1:  A B D . E . . . . . . . . . F . . .
Case 2:  . . . A . . . C D . . . F . . . . .
Case 3:  . . . . . A C . . D . E . F . . . .
Case 4:  . . . . . . . . . . A . . . . C D F

Traces = { (A, B, D, E, F),  (A, C, D, F),
           (A, C, D, E, F),  (A, C, D, F) }
```

estimated Case IDs

Evaluation

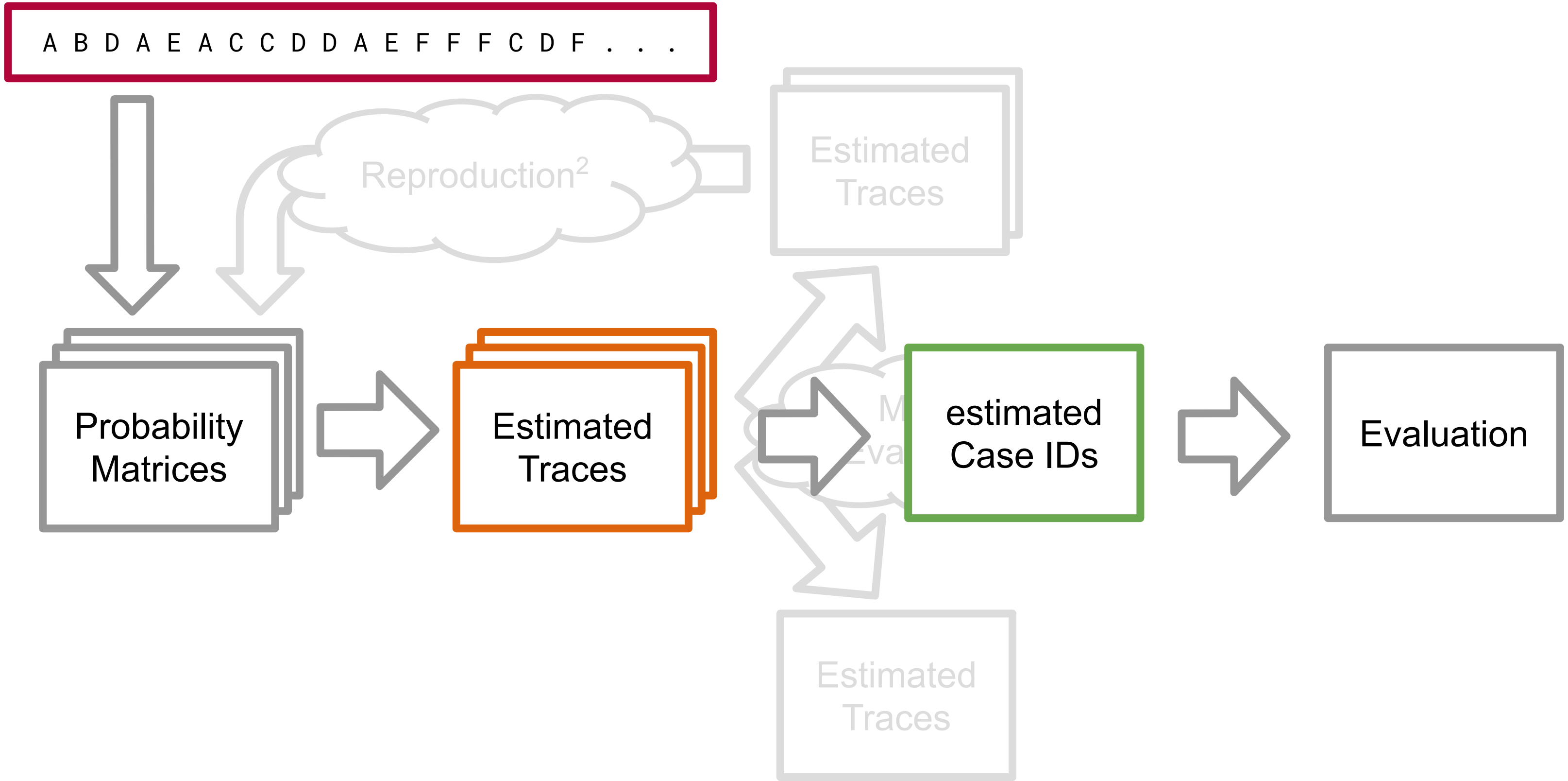
Genetic Programing Paradigm



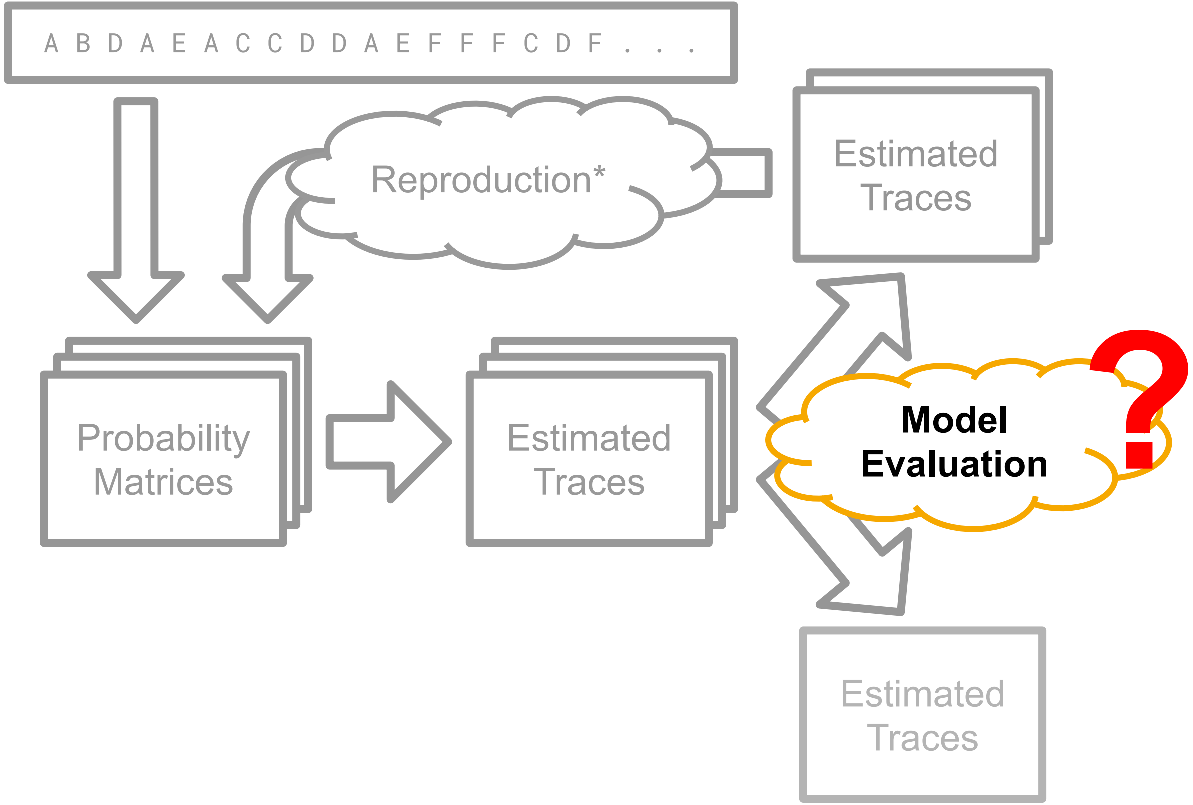
[1] The term 'Model' can be interpreted as a container for a Probability Matrix and the corresponding estimated traces.

[2] Breed new individuals through crossover and mutation operations from fittest individuals to replace the weakest ones.

Genetic Programing Paradigm



Genetic Programing Paradigm



Model Evaluation - Weight Functions

Weight Functions

- weight function $w \rightarrow w(M) \in [0, 1]$

Intuition

Compare Model to Real World Instances with Case IDs

- Not provided by unlabeled Event Logs

Problem

Evaluation of multiple model instances

- need for descriptive metrics
- need for computation of those metrics
- no Case IDs for comparison (no ground truth)

Research Question

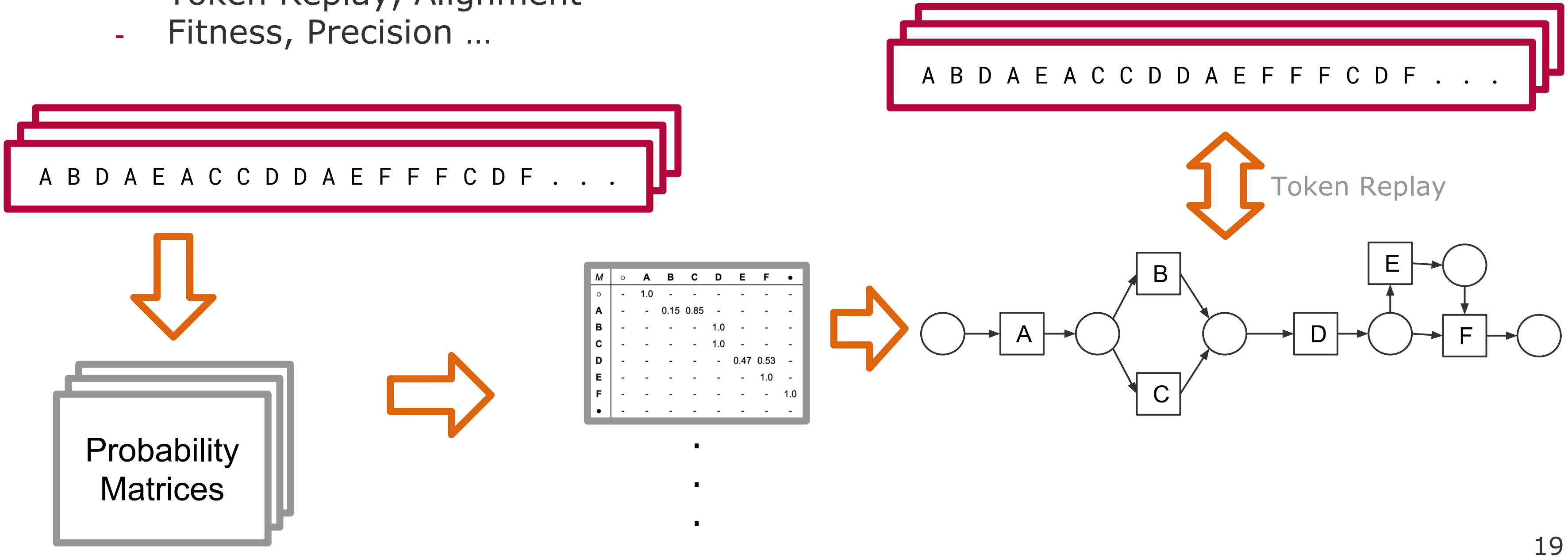
0. How can models be evaluated/ranked without further data (ground truth)?

1. Can the precision of this approach be improved by using a Genetic Programming Paradigm and other metrics?

2. Can assumptions for this approach be overcome with a Genetic Approach?

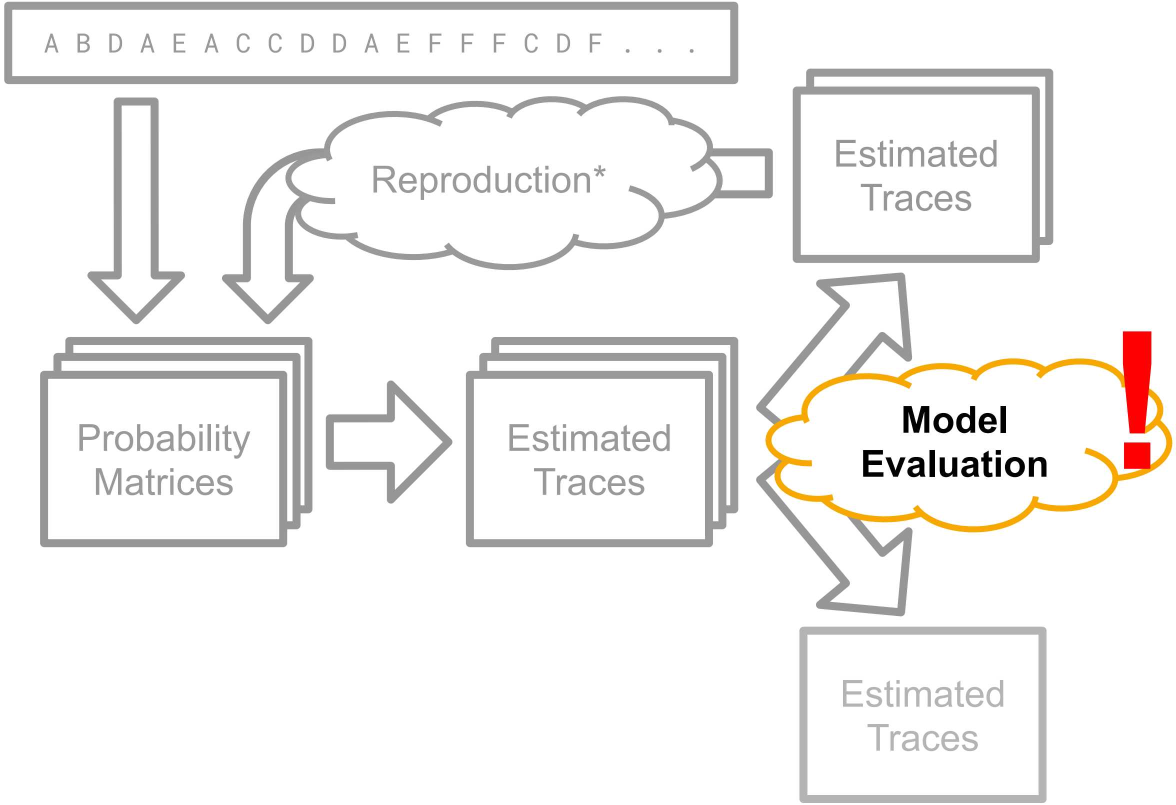
Weight Functions - Idea

- Take many unlabeled Event Logs as Input
- Initialize one Model each
- Compare one Matrix Instance to all other Input Logs
  - Token Replay, Alignment
  - Fitness, Precision …

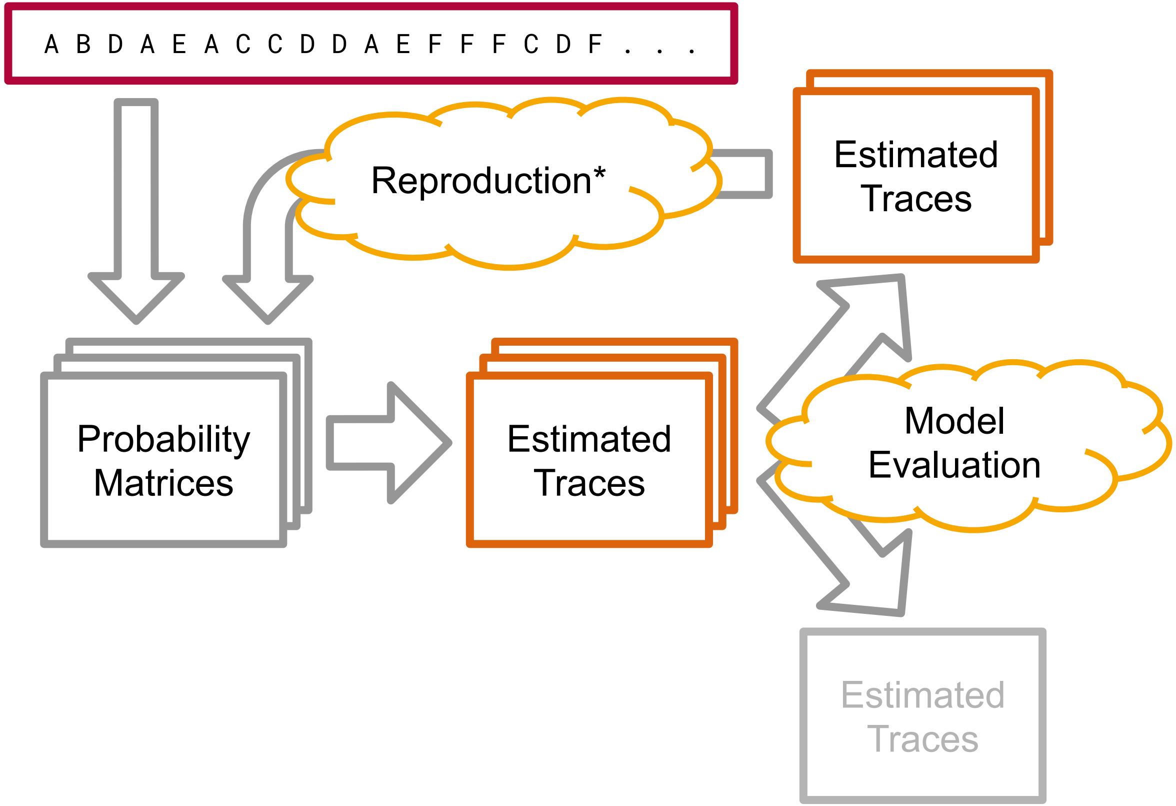


M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

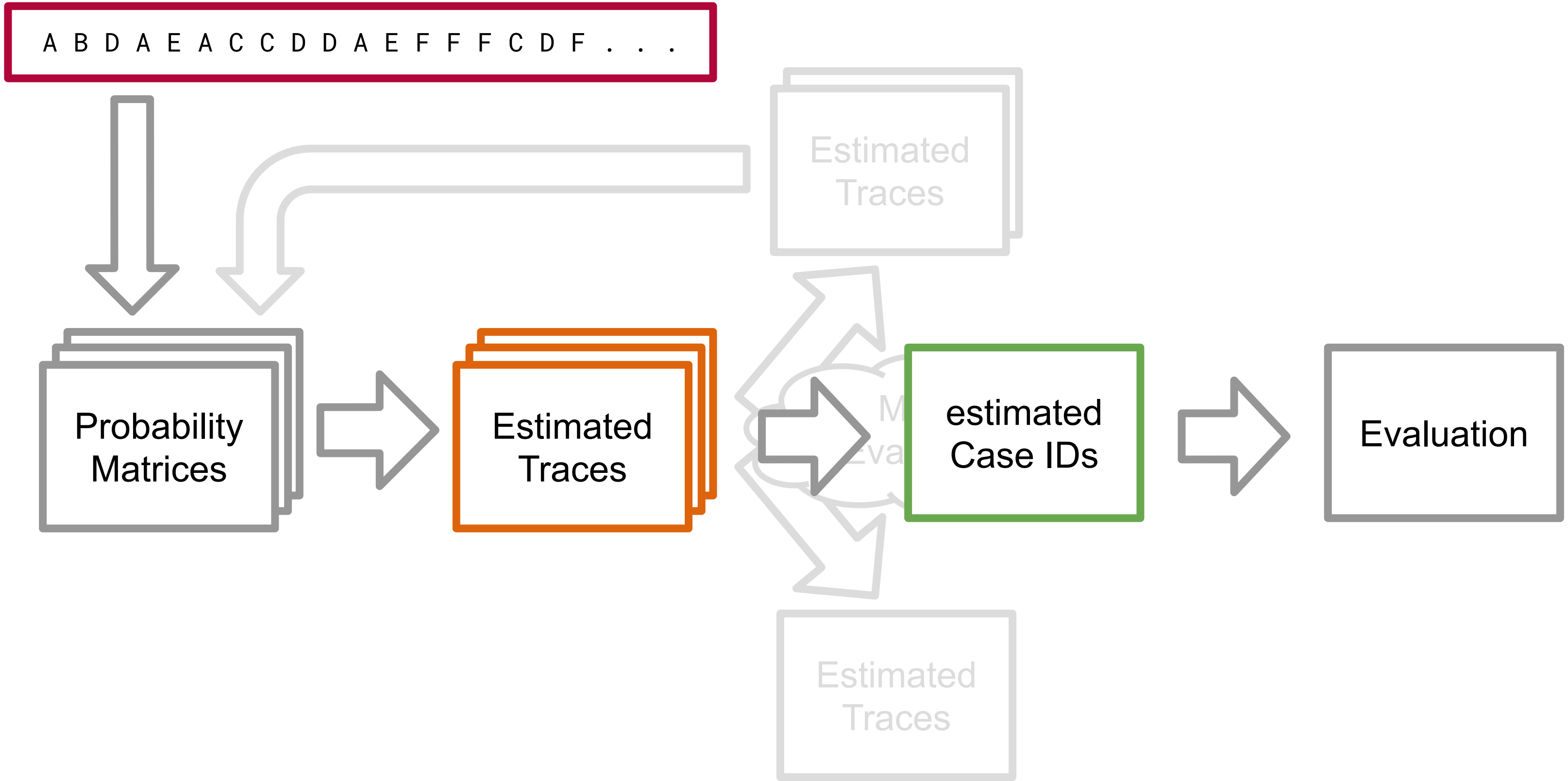
Genetic Programing Paradigm



Genetic Programing Paradigm



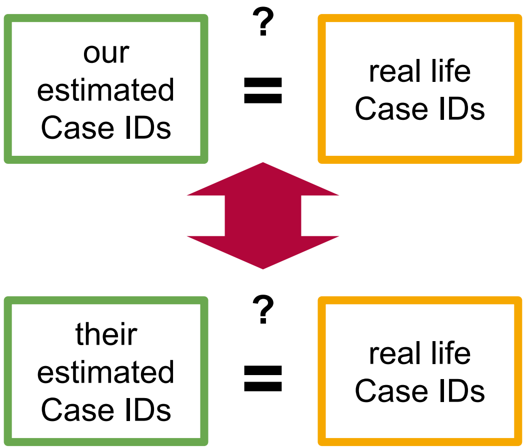
Genetic Programing Paradigm



Evaluation: Genetic Approach

Evaluation

- Comparison of present results and our own results
- Metrics:
  - G-score [2]
    - similarity of generated traces and real life traces
  - Fitness, Precision, Generalization, Simplicity [4]
  - Runtime & Space



Conclusion

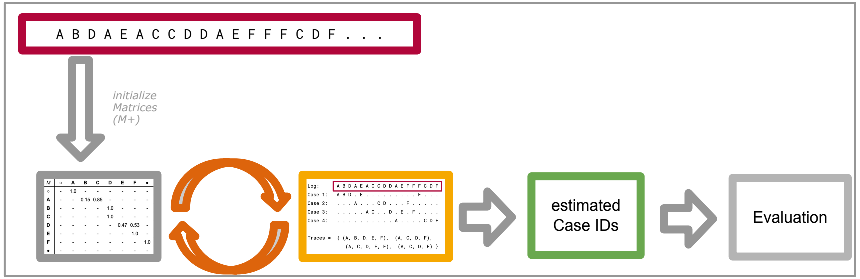
Next Steps

1. Genetic Extension to the iterative Algorithm [3]
   a. Define Weight Function
   b. Implementation
   c. Evaluation and Comparison of both Approaches

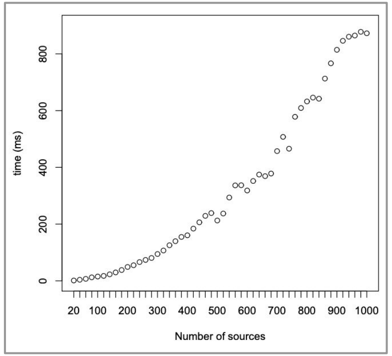
2. Getting rid of assumptions: Testing on different input models
   a. Loops, parallel behaviour, only full recorded end-to-end instances

3. Extending Algorithm
   a. Dynamic model creation instead of greedy rule-based procedure
   b. Different estimation matrices for initialization
      i. Distance
      ii. Global strength of causality

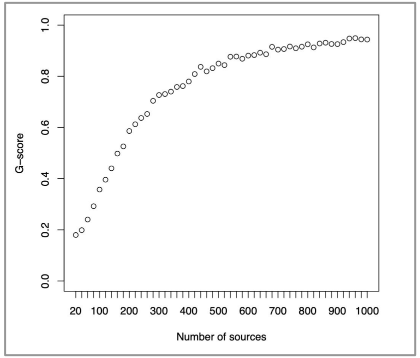
Summary



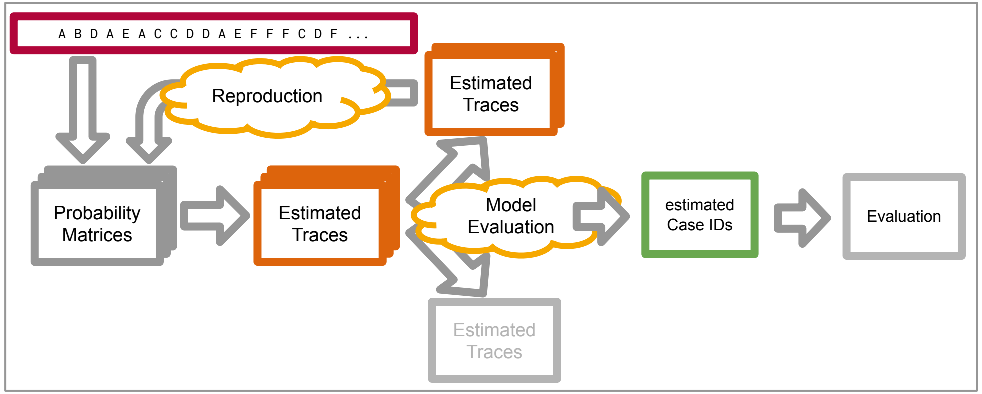
A B D A E A C C D D A E F F F C D F . . .
initialize Matrices (M+)
estimated Case IDs
Evaluation



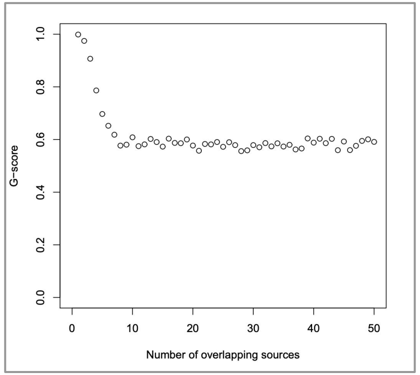
time (ms)
Number of sources



G−score
Number of sources

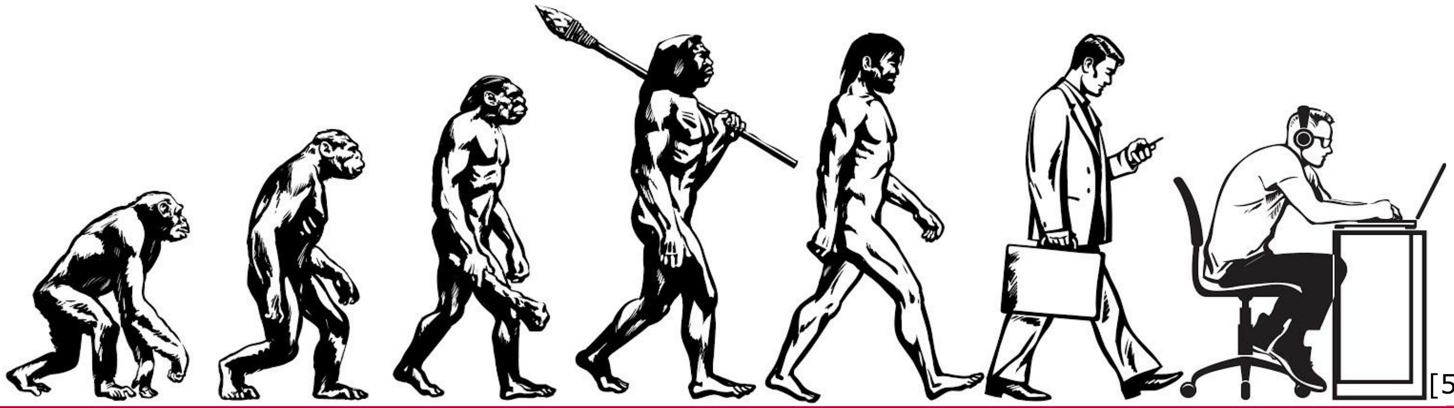


A B D A E A C C D D A E F F F C D F ...
Reproduction
Estimated Traces
Probability Matrices
Estimated Traces
Model Evaluation
estimated Case IDs
Evaluation
Estimated Traces



G−score
Number of overlapping sources

[2]


[5]

Discovering Process Models from unlabeled Event Logs

Pascal Schulze, Anjo Seidel
Data Extraction for Process Mining (ST-2020)
17.06.2020

Sources

[1] Diba, Kiarash & Batoulis, Kimon & Weidlich, Matthias & Weske, Mathias. (2019). Extraction, correlation, and abstraction of event data for process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 10. 10.1002/widm.1346.

[2] Ferreira D.R., Gillblad D. (2009) Discovering Process Models from Unlabelled Event Logs. In: Dayal U., Eder J., Koehler J., Reijers H.A. (eds) Business Process Management. BPM 2009. Lecture Notes in Computer Science, vol 5701. Springer, Berlin, Heidelberg

[3] Source code to accompany the paper “Discovering Process Models from Unlabelled Event Logs” [2] by Diogo R. Ferreira, Daniel Gillblad; Url: http://web.ist.utl.pt/diogo.ferreira/mimcode/

[4] Abbad Andaloussi A., Burattin A., Weber B. (2018) Toward an Automated Labeling of Event Log Attributes. In: Gulden J., Reinhartz-Berger I., Schmidt R., Guerreiro S., Guédria W., Bera P. (eds) Enterprise, Business-Process and Information Systems Modeling. BPMDS 2018, EMMSAD 2018. Lecture Notes in Business Information Processing, vol 318. Springer, Cham

[5] https://medium.com/ssense-tech/schema-evolution-in-data-lakes-f956c6f978d4

Appendix

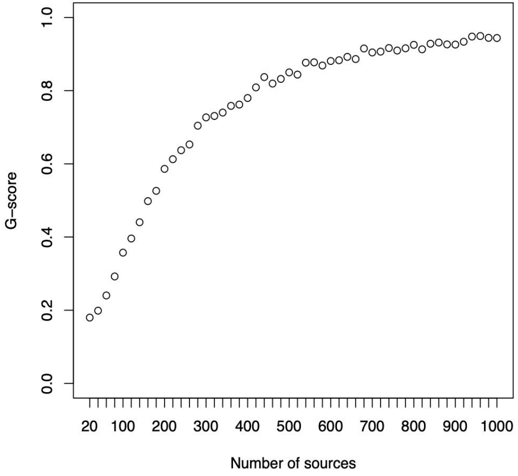
Real World Example

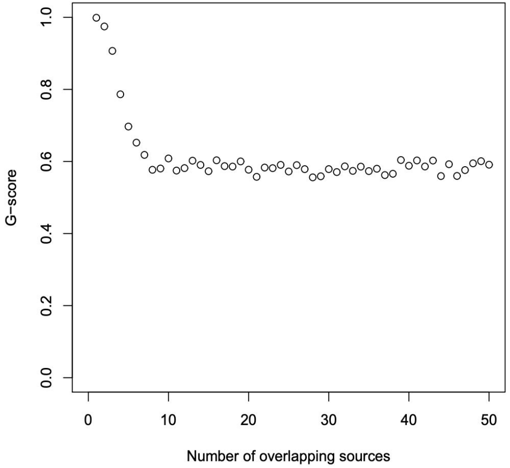
- Usage information needed improve (and automate) enterprise software
- Record user interaction in logs → User Behavior Mining
- “ERP Systems use the Business objects as the case identifier” [1]
- What if it is extremely difficult or even impossible to identify the Business objects?
  - companies often have old running on-premise systems (SAP Gui) → difficult
  - frontend gets rendered on server (SAP Screen Personas) → extremely difficult or impossible
- Updates or bigger Adjustments would be needed to get desired result

The Approach - G-score [2]

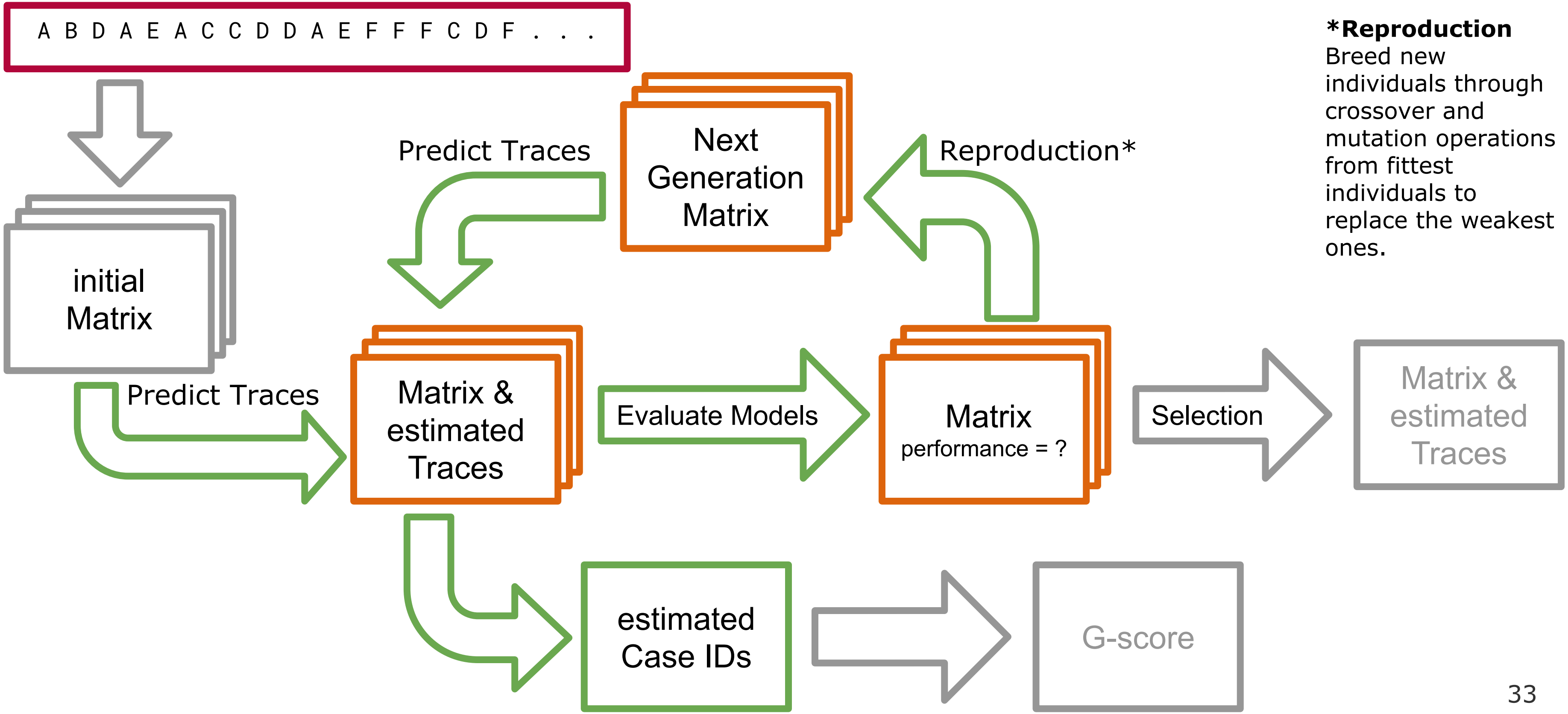
Scoring measure which evaluates the degree of similarity between a complete event log, where both x and s are known, and an incomplete event log x that has been labelled by the estimated source sequence s˜.

$$G^*\text{-score as } \sum_z \sqrt{p(z) \cdot q^*(z)}$$





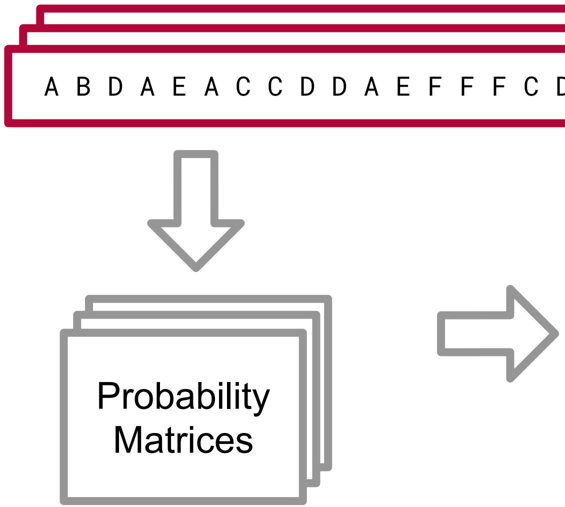
Genetic Programing Paradigm - Detailed



***Reproduction**
Breed new individuals through crossover and mutation operations from fittest individuals to replace the weakest ones.

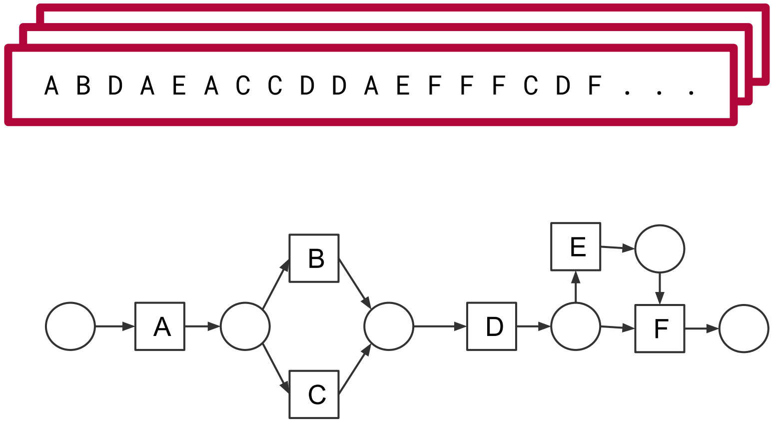
Weight Functions - Idea 1

- Take many unlabeled Event Logs as Input
- Initialize one Model each
- Compare one Matrix Instance to all other Input Logs
  - Token Replay, Alignment
  - Fitness, Precision …



M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

.
.
.



Weight Functions - Idea 2

- Compare one Matrix instance to all other estimated Traces
  - Token Replay, Alignment
  - Metrics: Fitness, Precision, Generalization, Simplicity
- Matrices resembling most consent get better weight

M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

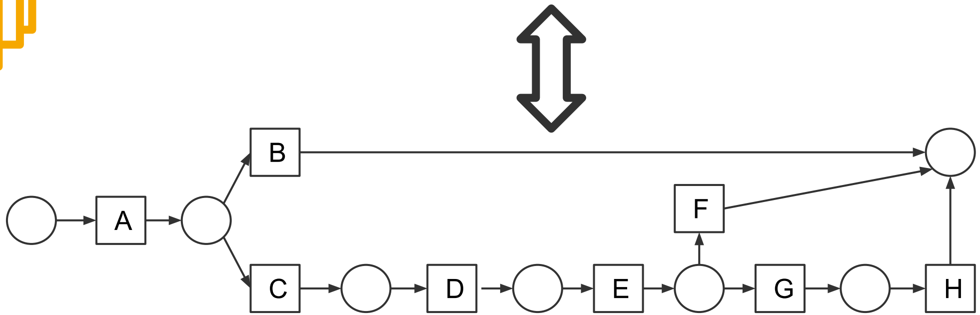
⋮

? =

```
Log:     A B D A E A C C D D A E F F F C D F
Case 1:  A B D . E . . . . . . . . F . . . .
Case 2:  . . . A . . . C D . . . F . . . . .
Case 3:  . . . . . A C . . D . E . F . . . .
Case 4:  . . . . . . . . . . A . . . . C D F

Traces = { (A, B, D, E, F),  (A, C, D, F),
           (A, C, D, E, F),  (A, C, D, F) }
```

Y = {(A,B),
(A,C,D,E,F),
(A,C,D,E,F,G,H)}



Weight Functions - Idea 3

- Take many unlabeled Event Logs as Input
- Initialize one model each
- for each model, compare pattern Y with patterns Y' of all other input strings
  - Idea: a good matrix produces similar patterns for event logs from the same process model