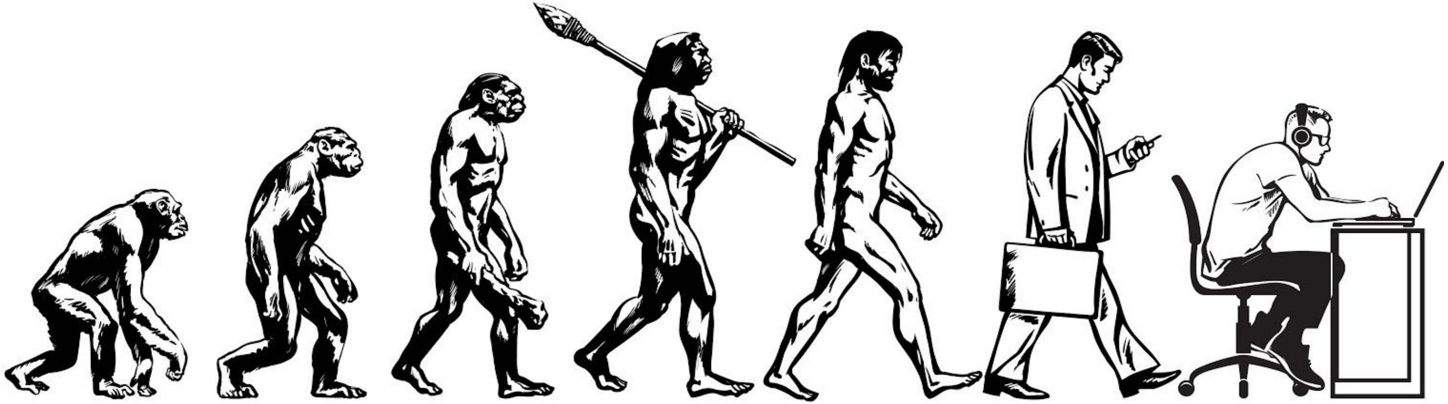


Genetic Correlation Discovery for unlabeled Event Logs

Pascal Schulze, Anjo Seidel

Data Extraction for Process Mining (ST-2020)

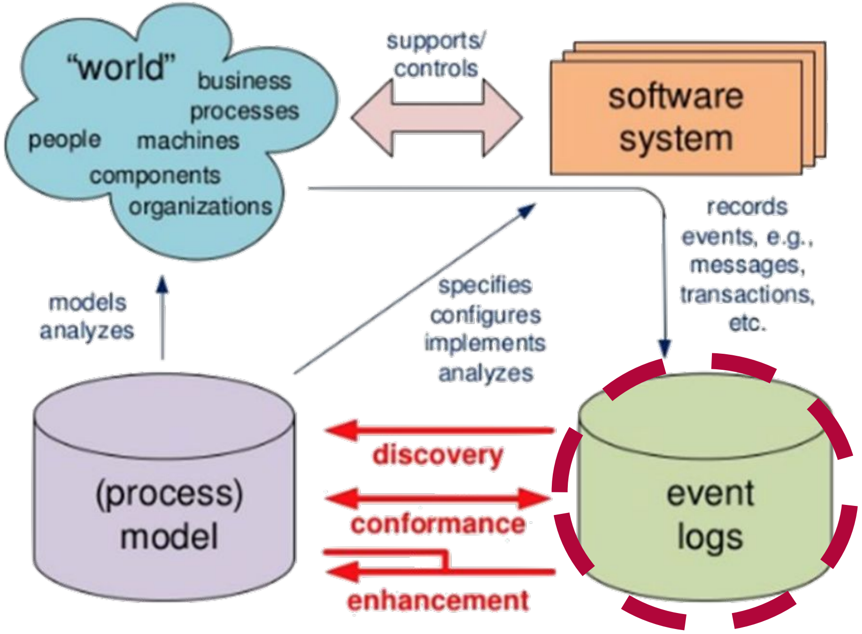
Supervisor: Simon Remy

13.08.2020

(Quick) Recap

Context

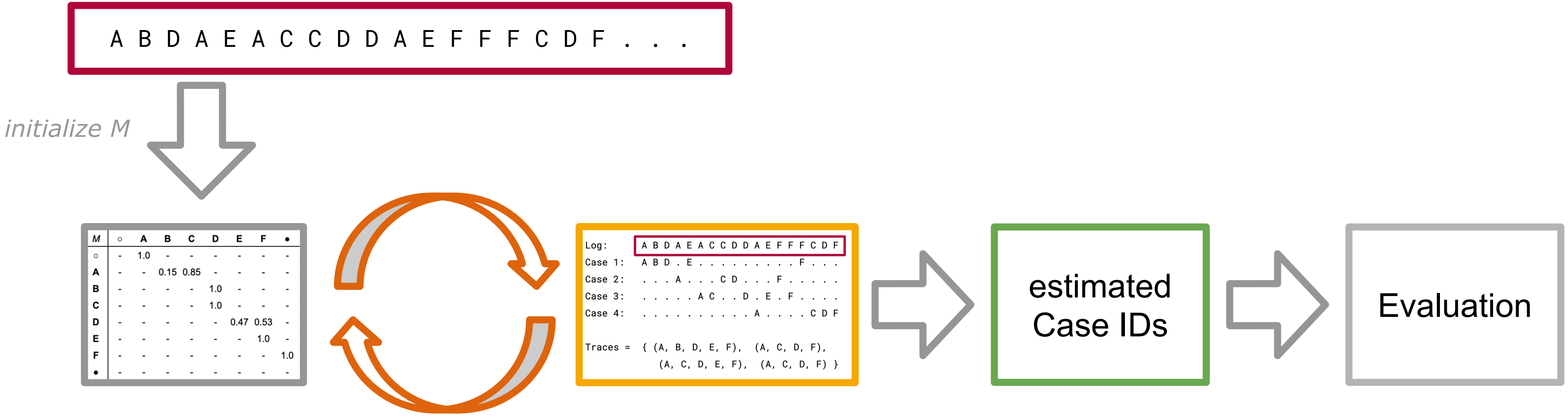
- Need for event logs
- Event logs are generated from data within software systems
- Events are extracted, correlated and abstracted
- Normal procedure: find correlating events by
  - existing IDs
  - attributes
  - domain knowledge
- **How to find correlation without that information?**



Present Approach [2]

- M gets initialized based on direct successorship (M+)

- Iteratively estimate case IDs with M and estimate M with given case IDs



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

initialize 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
●	-	-	-	-	-	-	-	-

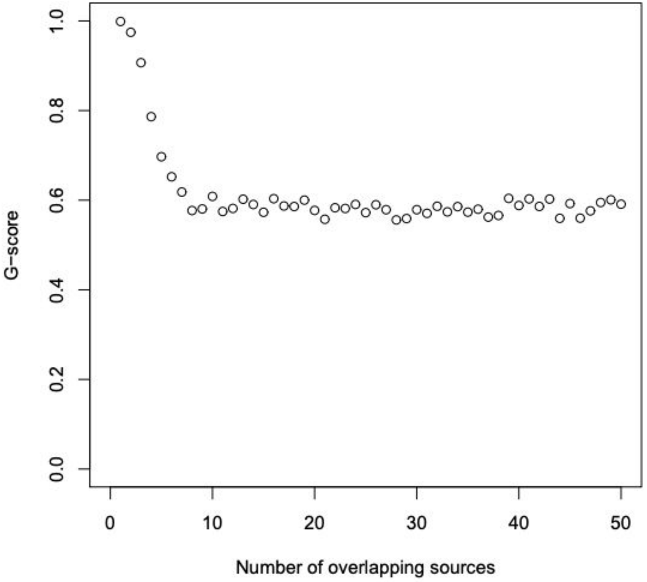
```
Log:     A B D A E A C C D D A E F F F C D F
Case 1:  A B D . E . . . . . . . . . F . . .
Case 2:  . . . A . . . C D . . . F . . . . .
Case 3:  . . . . . A C . . D . E . F . . . .
Case 4:  . . . . . . . . . . A . . . . C D F


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

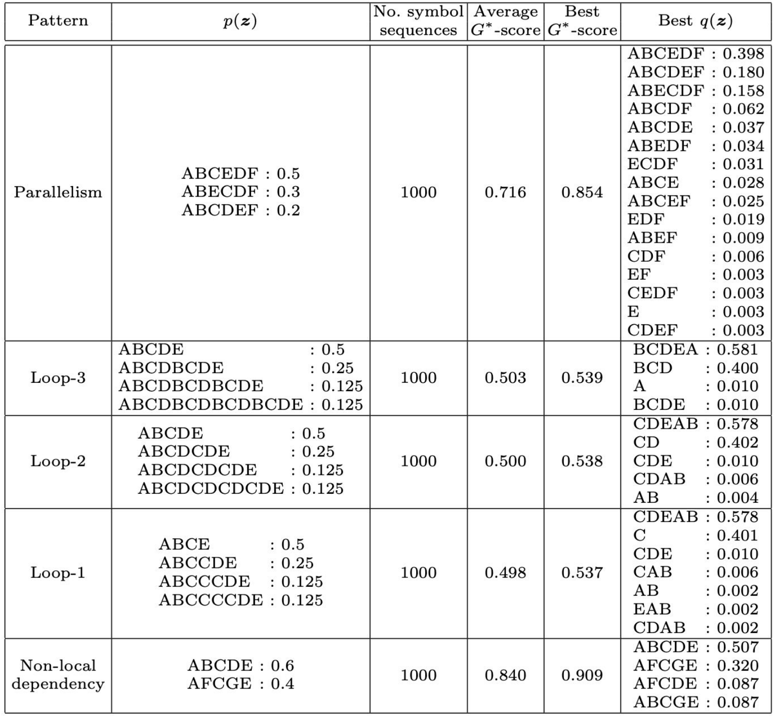
estimated Case IDs

Evaluation

Motivation

- Improve the outcome accuracy (G/G*-Score) with a genetic extension
- Greedy-Algorithm with strong assumptions



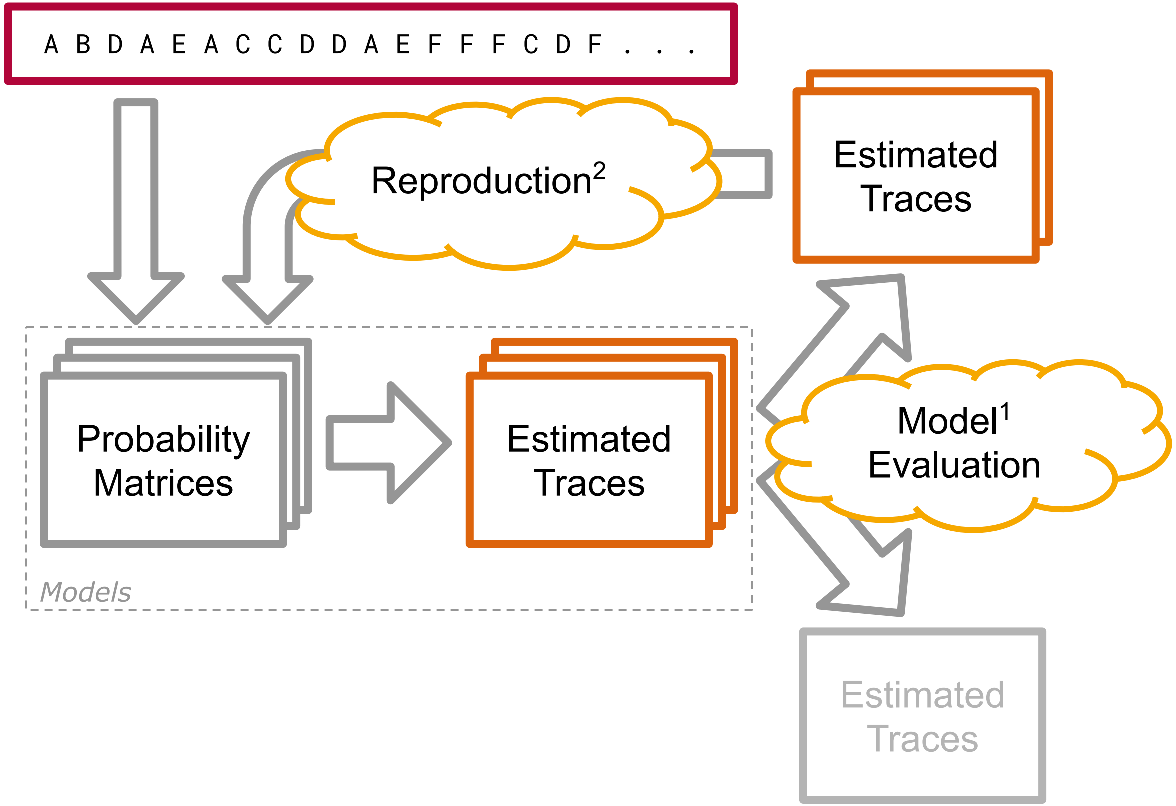


Pattern	$p(\boldsymbol{z})$	No. symbol sequences	Average G^*-score	Best G^*-score	Best $q(\boldsymbol{z})$
Parallelism	ABCEDF : 0.5 ABECDF : 0.3 ABCDEF : 0.2	1000	0.716	0.854	ABCEDF : 0.398 ABCDEF : 0.180 ABECDF : 0.158 ABCDF : 0.062 ABCDE : 0.037 ABEDF : 0.034 ECDF : 0.031 ABCE : 0.028 ABCEF : 0.025 EDF : 0.019 ABEF : 0.009 CDF : 0.006 EF : 0.003 CEDF : 0.003 E : 0.003 CDEF : 0.003
Loop-3	ABCDE : 0.5 ABCDBCDE : 0.25 ABCDBCDBCDE : 0.125 ABCDBCDBCDBCDE : 0.125	1000	0.503	0.539	BCDEA : 0.581 BCD : 0.400 A : 0.010 BCDE : 0.010
Loop-2	ABCDE : 0.5 ABCDCDE : 0.25 ABCDCDCDE : 0.125 ABCDCDCDCDE : 0.125	1000	0.500	0.538	CDEAB : 0.578 CD : 0.402 CDE : 0.010 CDAB : 0.006 AB : 0.004
Loop-1	ABCE : 0.5 ABCCDE : 0.25 ABCCCDE : 0.125 ABCCCCDE : 0.125	1000	0.498	0.537	CDEAB : 0.578 C : 0.401 CDE : 0.010 CAB : 0.006 AB : 0.002 EAB : 0.002 CDAB : 0.002
Non-local dependency	ABCDE : 0.6 AFCGE : 0.4	1000	0.840	0.909	ABCDE : 0.507 AFCGE : 0.320 AFCDE : 0.087 ABCGE : 0.087

[2]

Loops and parallelism are decreasing the accuracy significantly.

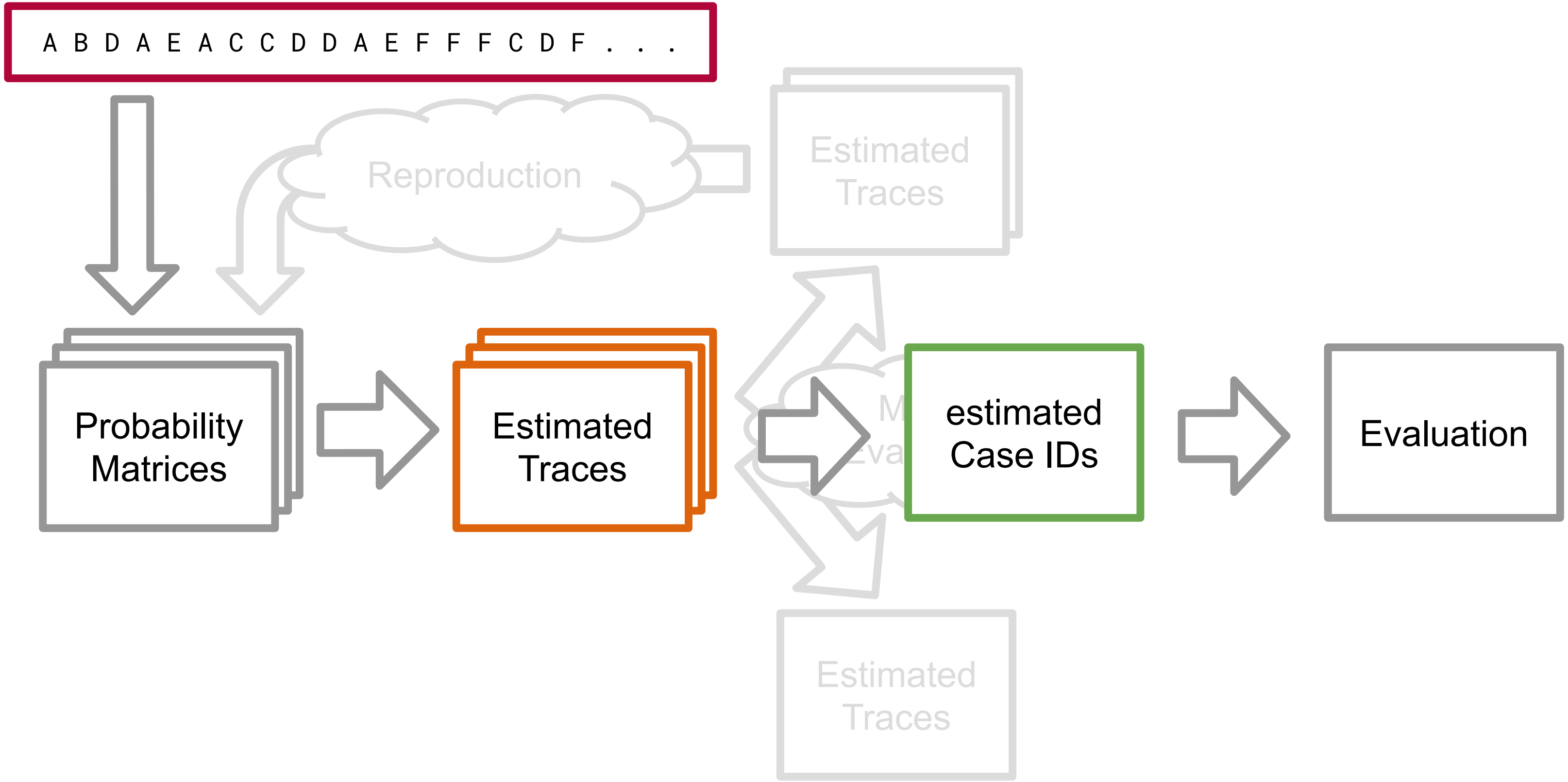
Genetic Extension



[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. Also includes random mutations.

Genetic Extension



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

M

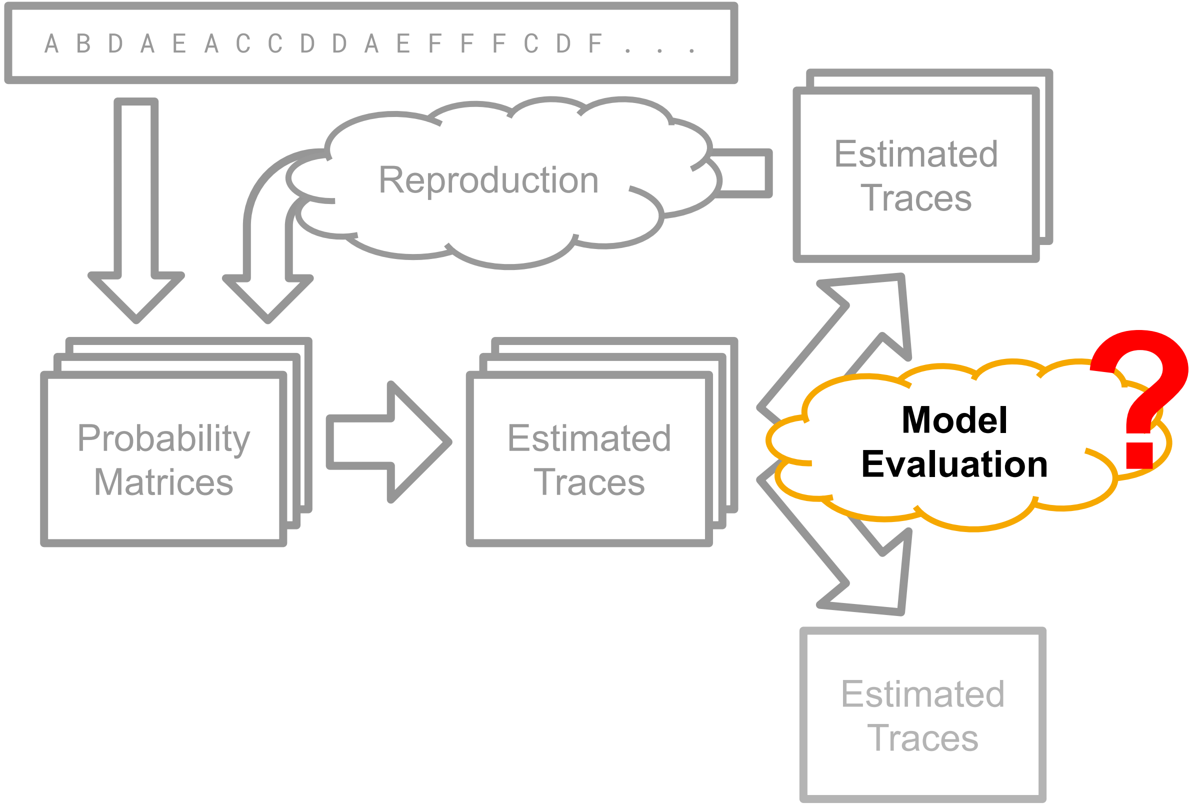
Eva

estimated Case IDs

Evaluation

Estimated Traces

Genetic Extension



Research Questions

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?

Fitness (Approximation) Functions

Fitness Functions

$m \in$ Models M

fitness function f: m → [0, 1]

Intuition

- Compare Model to Real World Instances with Case IDs
  - Not provided by unlabeled Event Logs

Problem

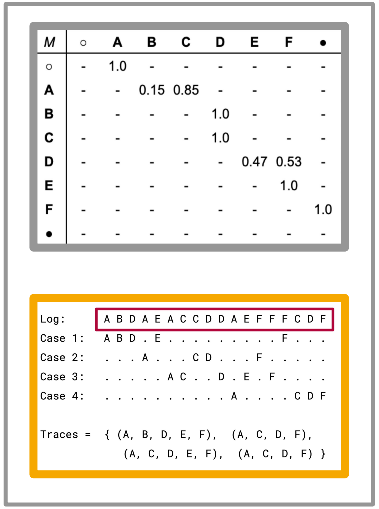
- Evaluation of multiple model instances
- No Case IDs for comparison (no ground truth)

Fitness Approximation Function

Ideas general

- compare one model instance to all other model instances
- All models depict the same event log/ business process
- Find the best consensus of all models

Model m:



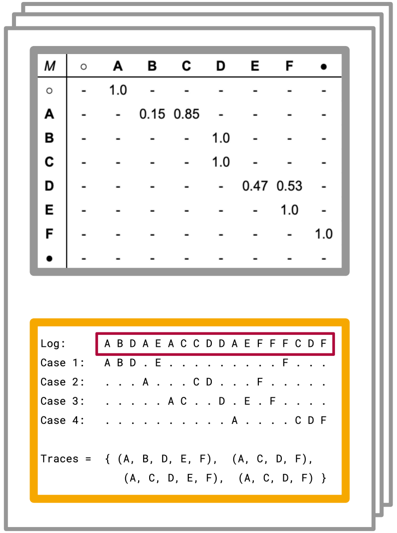
M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

```
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) }
```

?

=

All models M:



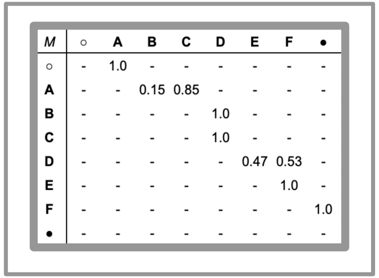
M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

```
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) }
```

Ideas general

- compare one model instance to all other model instances
- All models depict the same event log/ business process
- Find the best consensus of all models

Model m:

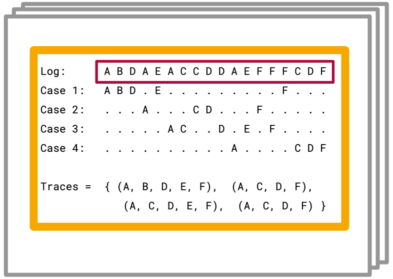


M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

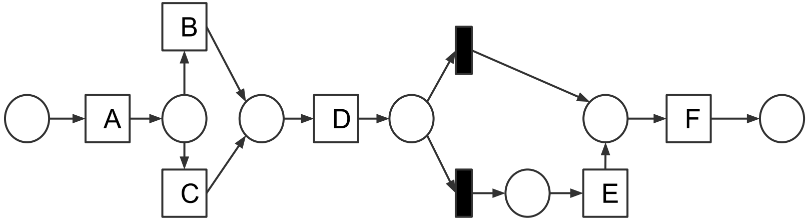
?

=

All models M:

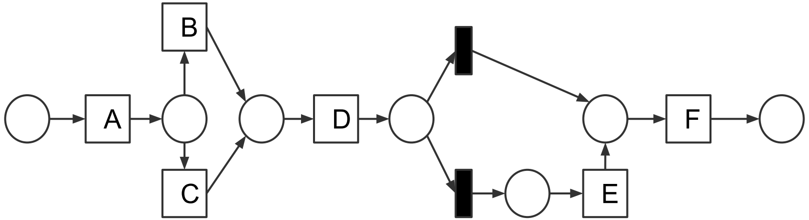


```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .
Case 4:   . . . . . . . . . . A . . . . C D F


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)}

Idea 1 - Alignment



A	B	⊥	D	E	F
A	B	C	⊥	E	F

A B C E F

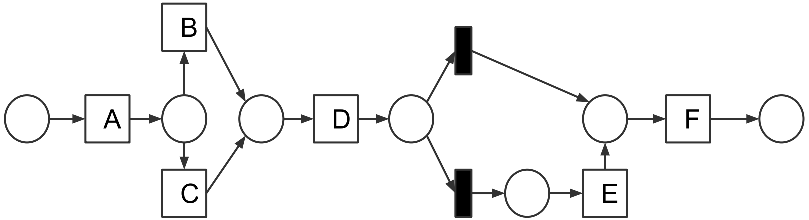
?

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

Challenges

- Multiple Trace Alignment is NP-complete
- Not tested/implement by us

Idea 2 - Token Replay



A B D E F

c	6
p	6
m	1
r	1

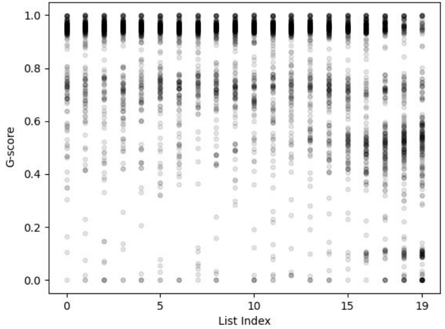
$$f = \frac{1}{2}(1 - \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i c_i}) + \frac{1}{2}(1 - \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i p_i})$$

Challenges

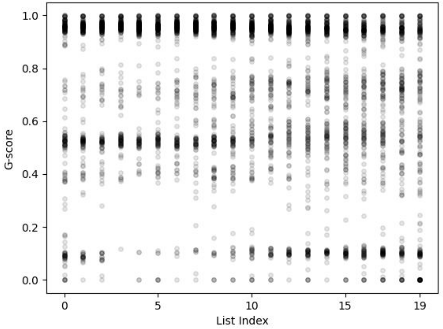
- many parallel traces
- probabilities/ weighted edges

Idea 2 - Token Replay

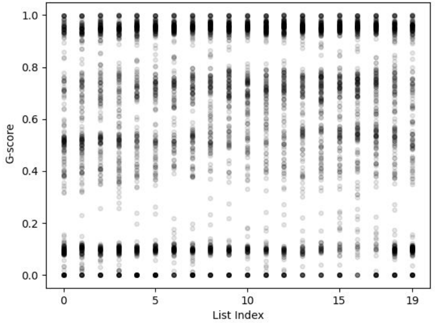
Token Replay on Symbol Sequences



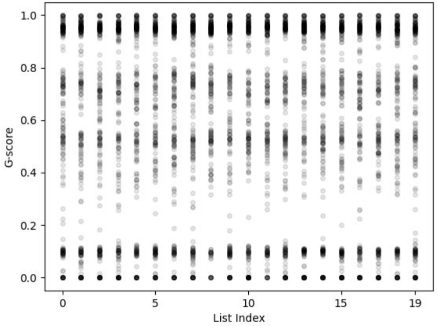
Weighted-Linked Token Replay on Symbol Sequences



Token Replay on Models



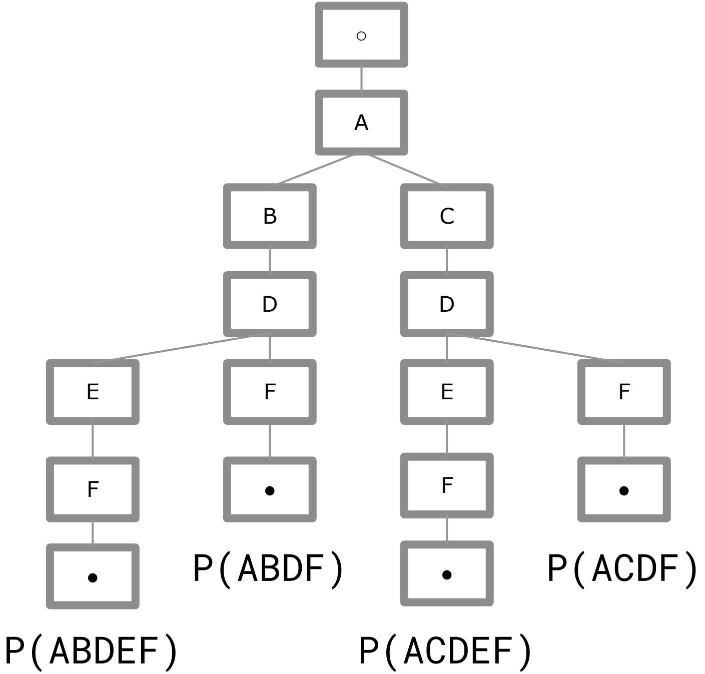
Random



Idea 3 - Summed Probabilities

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
●	-	-	-	-	-	-	-	-

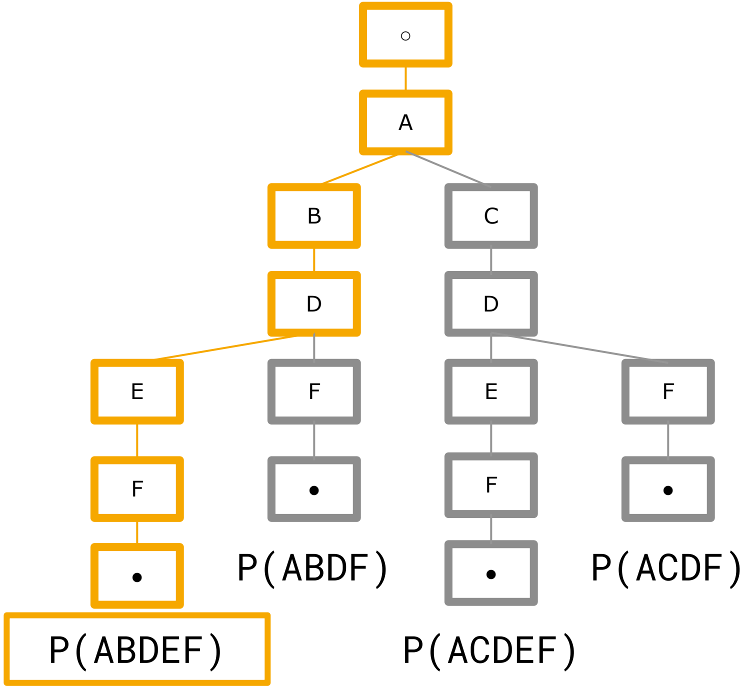
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .
Case 4:   . . . . . . . . . . A . . . . C D F

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



Idea 3 - Summed Probabilities

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
●	-	-	-	-	-	-	-	-

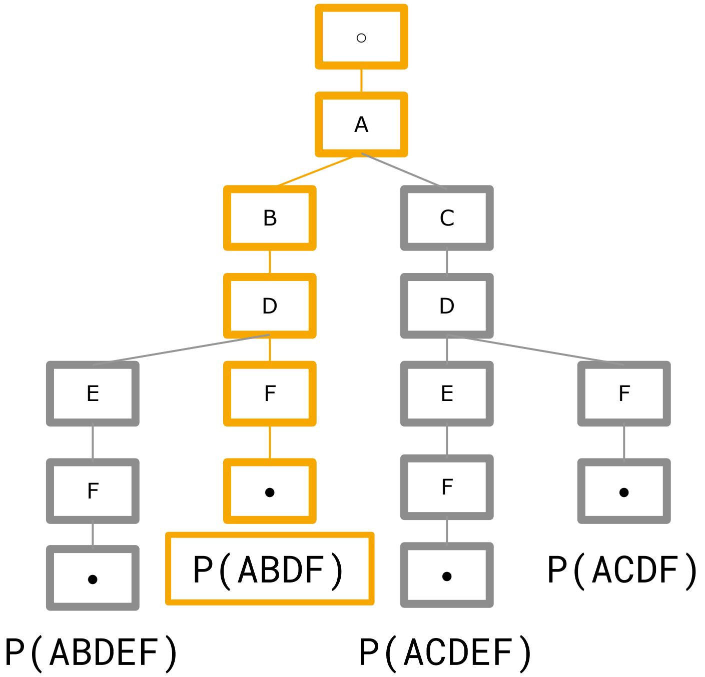
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .
Case 4:   . . . . . . . . . . A . . . . C D F


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



Idea 3 - Summed Probabilities

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
●	-	-	-	-	-	-	-	-

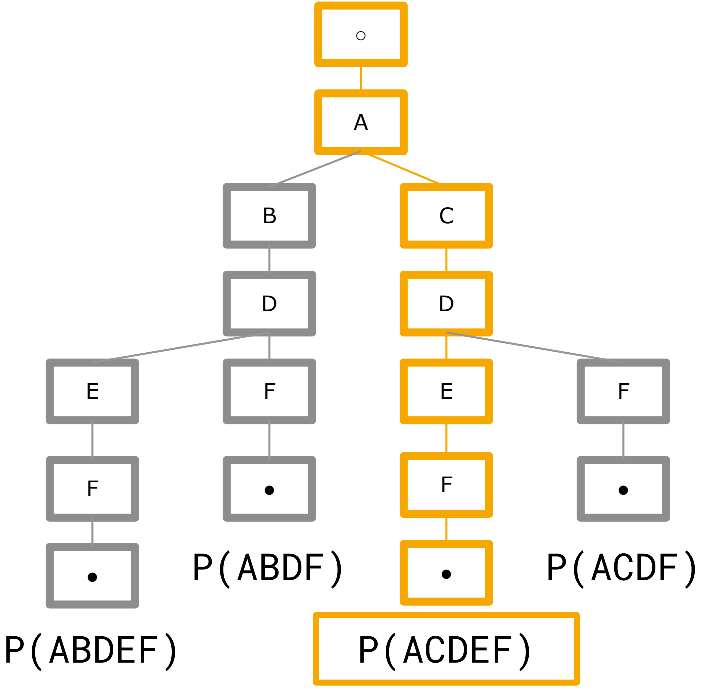
```
Log:     A B D A E A C C D D A E F F F C D F
Case 1:  A B D . E . . . . . . . . . F . . .
Case 2:  . . . A . . . C D . . . F . . . . .
Case 3:  . . . . . A C . . D . E . F . . . .
Case 4:  . . . . . . . . . . A . . . . C D F

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



Idea 3 - Summed Probabilities

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
●	-	-	-	-	-	-	-	-

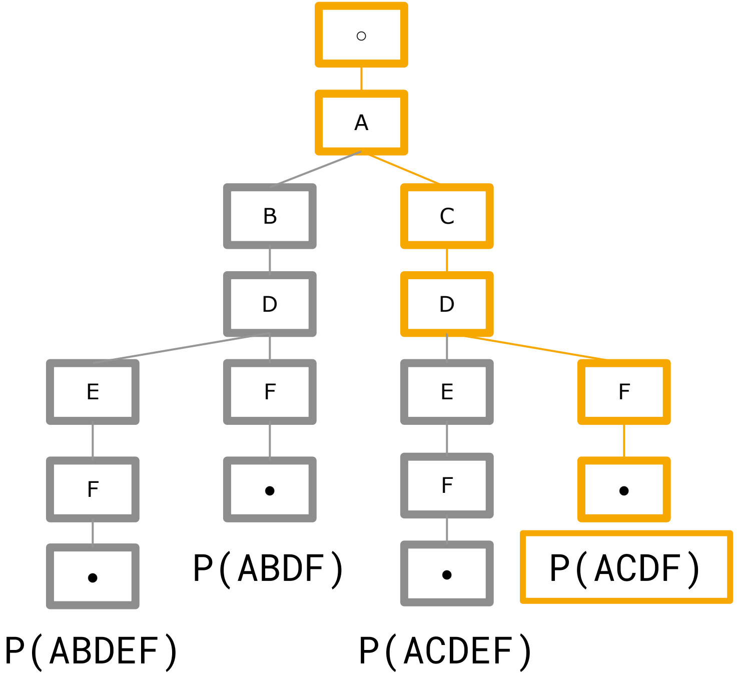
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .
Case 4:   . . . . . . . . . . A . . . . C D F


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



Idea 3 - Summed Probabilities

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
●	-	-	-	-	-	-	-	-

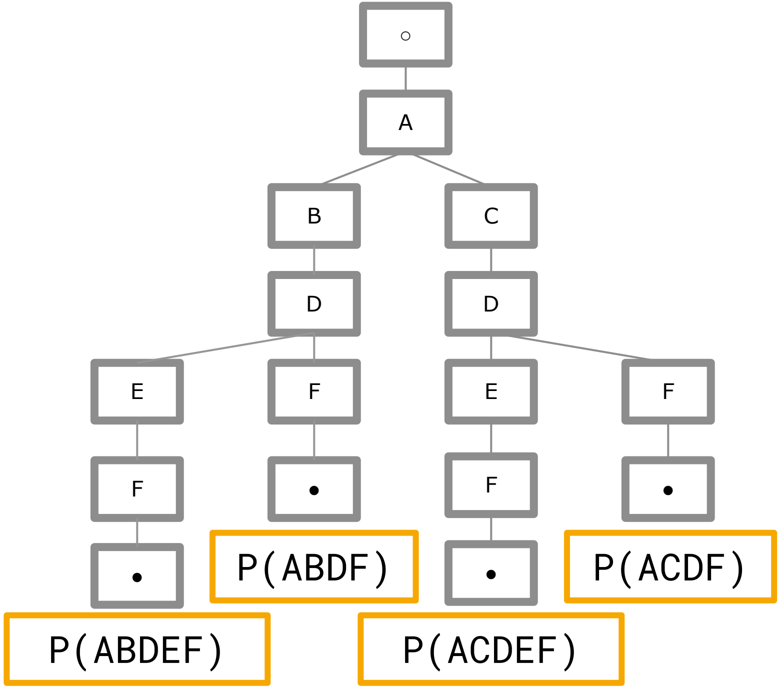
```
Log:     A B D A E A C C D D A E F F F C D F
Case 1:  A B D . E . . . . . . . . . F . . .
Case 2:  . . . A . . . C D . . . F . . . . .
Case 3:  . . . . . A C . . D . E . F . . . .
Case 4:  . . . . . . . . . . A . . . . C D F

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



Idea 3 - Summed Probabilities

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) }
```



P(ABDF)

P(ACDF)

P(ABDEF)

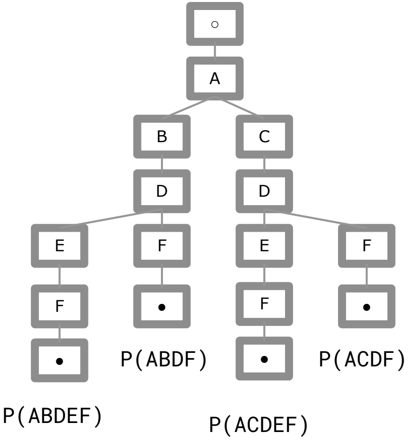
P(ACDEF)

$$\sum_{y \in Y} P(y) = 1$$

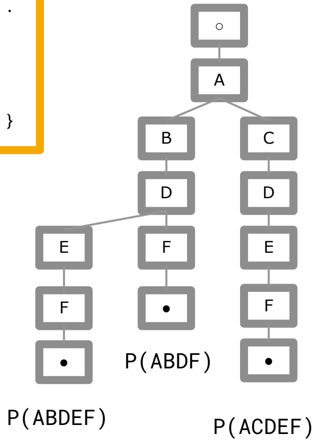
Idea 3 - Summed Probabilities

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
●	-	-	-	-	-	-	-	-

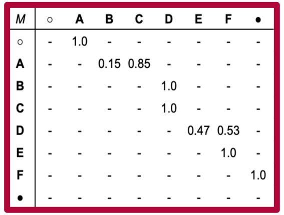
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .
Case 4:   . . . . . . . . . . A . . . . C D F

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



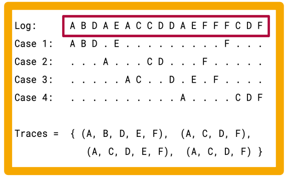
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .


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



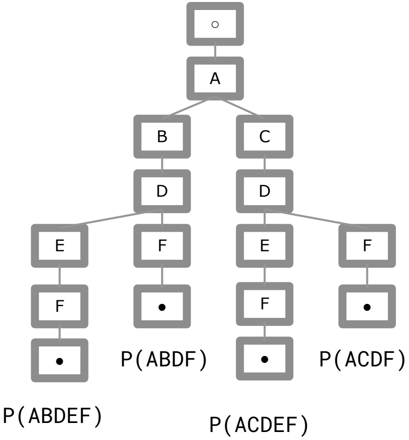
Idea 3 - Summed Probabilities

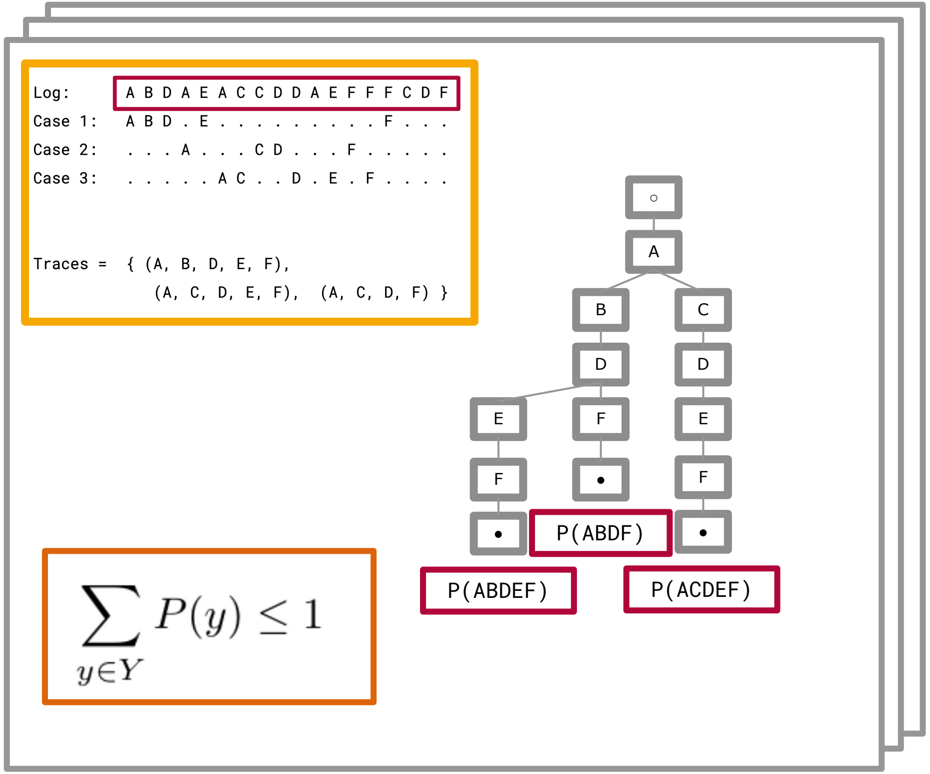


M	○	**A**	**B**	**C**	**D**	**E**	**F**	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-



```
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) }
```





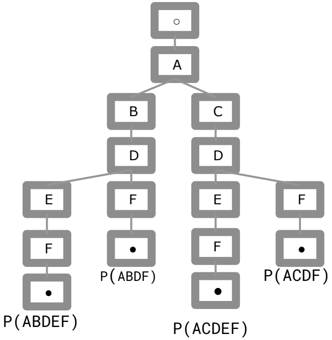
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .


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

$$\sum_{y \in Y} P(y) \leq 1$$

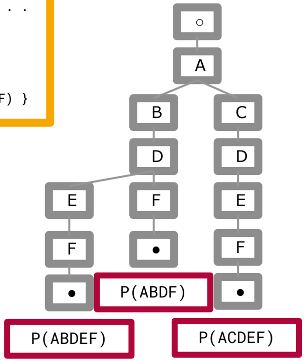
Idea 3 - Summed Probabilities

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) }
```



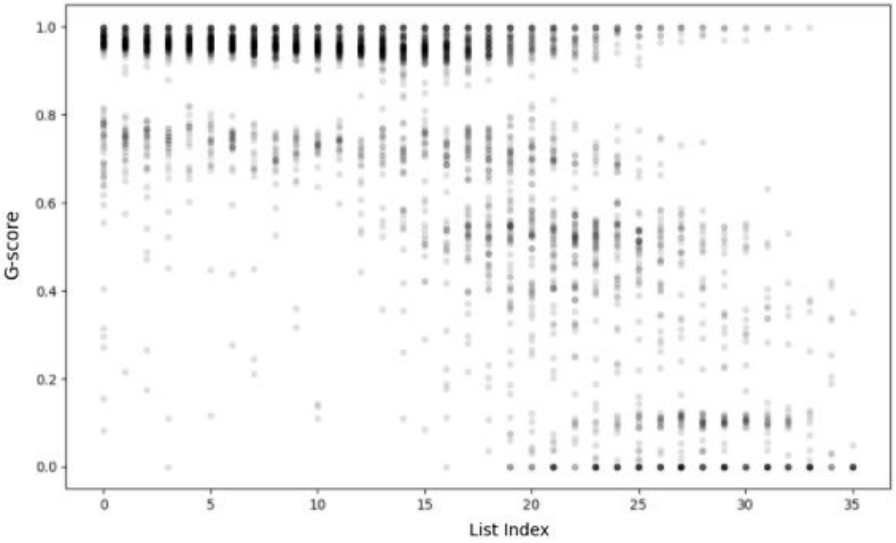
```
Log:      A B D A E A C C D D A E F F F C D F
Case 1:   A B D . E . . . . . . . . . F . . .
Case 2:   . . . A . . . C D . . . F . . . . .
Case 3:   . . . . . A C . . D . E . F . . . .


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



$$\sum_{y\in Y} P(y) \leq 1$$

$$f(m,M) = \frac{1}{|M|} \cdot \sum_{m_i\in M} \Big(\sum_{y\in Y_{m_i}} P_{m_i}(y) \Big)$$

Idea 3 - Summed Probabilities



$$f(m, M) = \frac{1}{|M|} \cdot \sum_{m_i \in M} \left(\sum_{y \in Y_{m_i}} P_{m_i}(y) \right)$$

Final Idea - G-Score

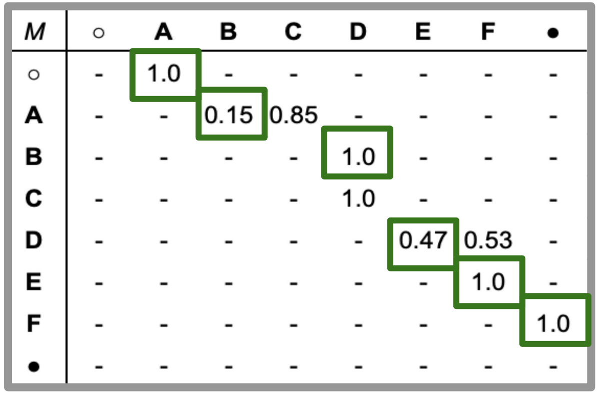
- Idea 3 is basically the G-Score!

- G-Score:

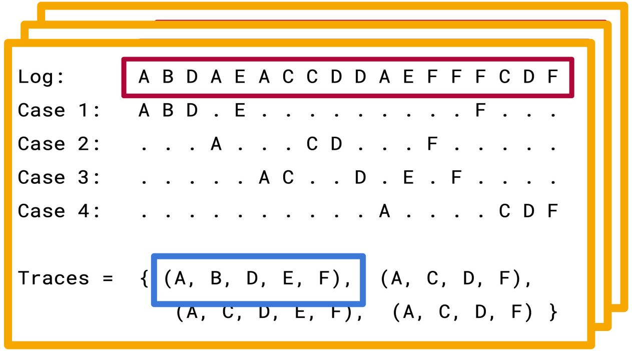
$$G(p\,||\,q) \mathrel{\widehat{=}} \sum_{z\in Z} \sqrt{\boxed{p(z)}\cdot\boxed{q(z)}}$$

- Fitness Approximation Function:

$$f(\boxed{m},\boxed{M}) = \frac{1}{\boxed{|M|}} \cdot \sum_{m_i\in M} G(\boxed{m},\boxed{m_i})$$



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) }
```

Final Idea - G-Score

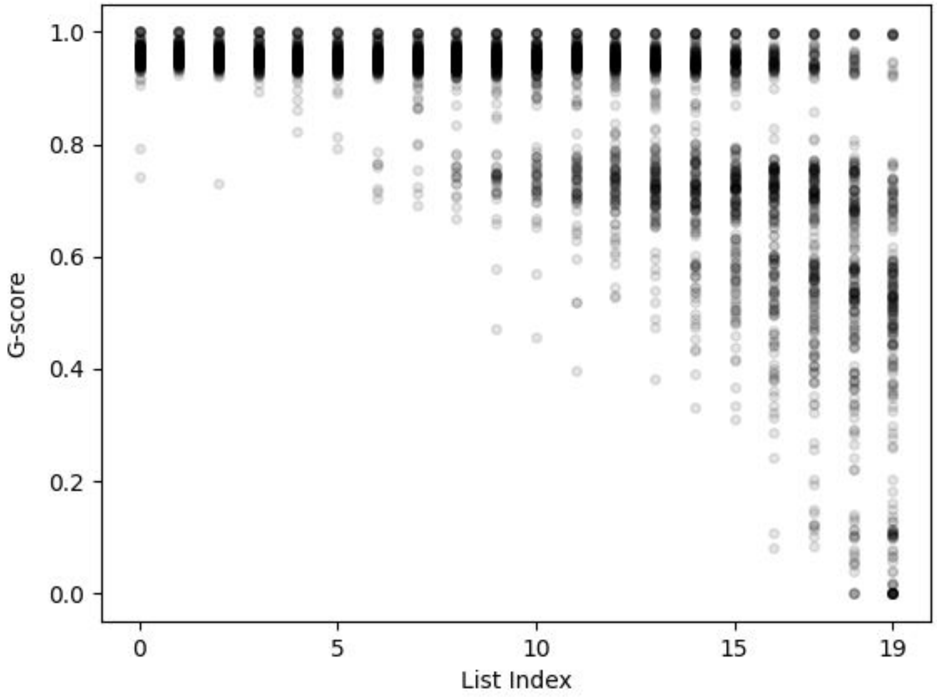
- Idea 3 is basically the G-Score!

- G-Score:

$$G(p\,||\,q) \widehat{=} \sum_{z \in Z} \sqrt{p(z) \cdot q(z)}$$

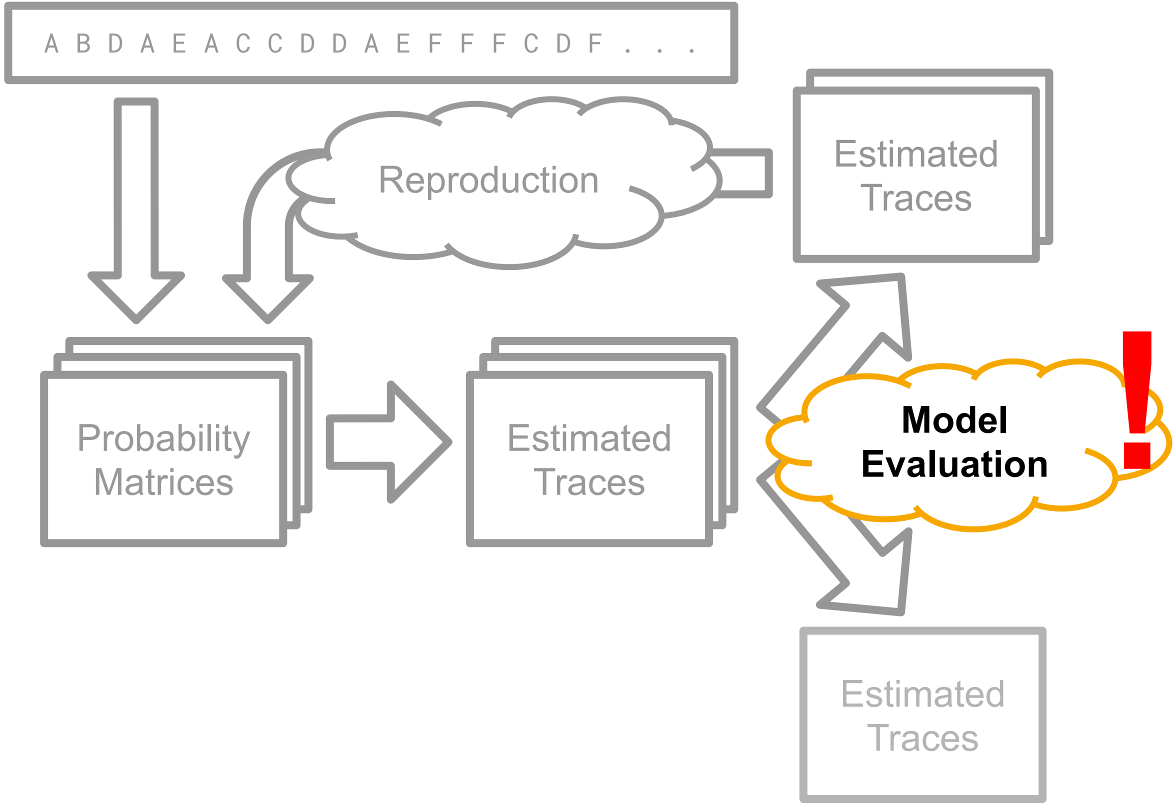
- Fitness Approximation Function:

$$f(m, M) = \frac{1}{|M|} \cdot \sum_{m_i \in M} G(m, m_i)$$

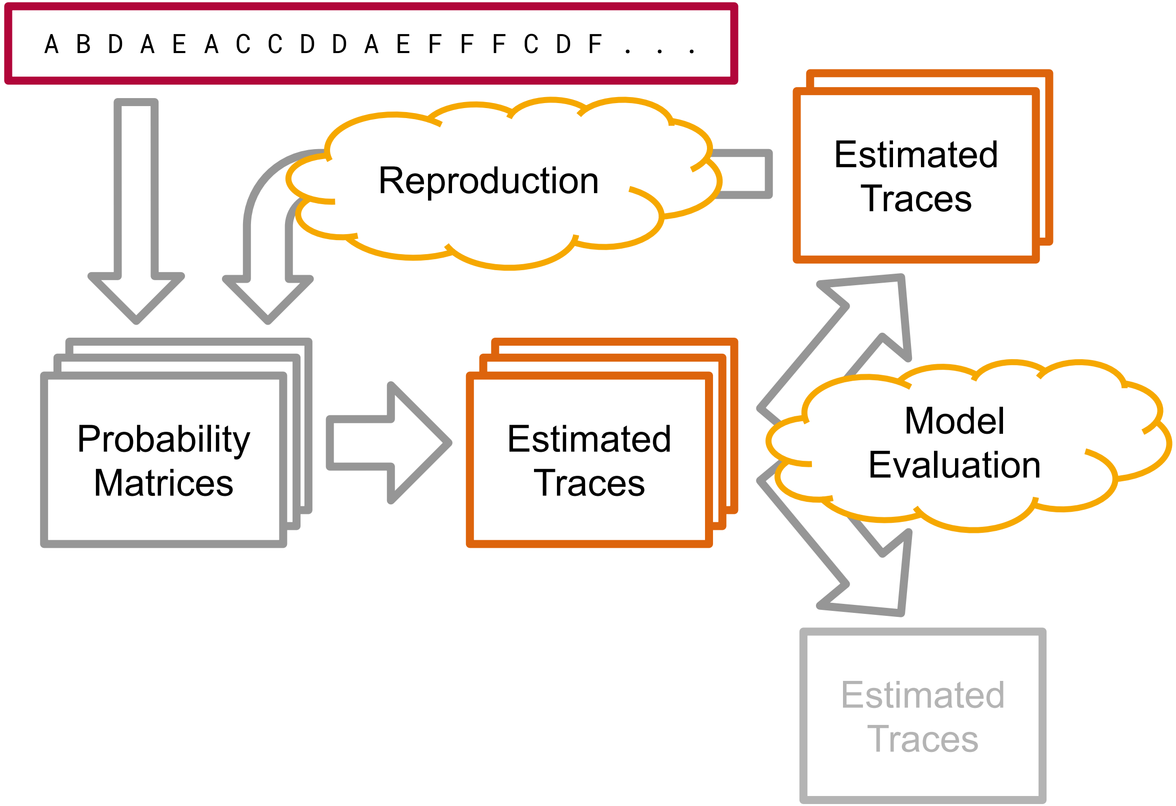


The diagram displays 100 sorted model-lists with 20 models each. For each index per individual list the relating model is then evaluated to get its real g-score.

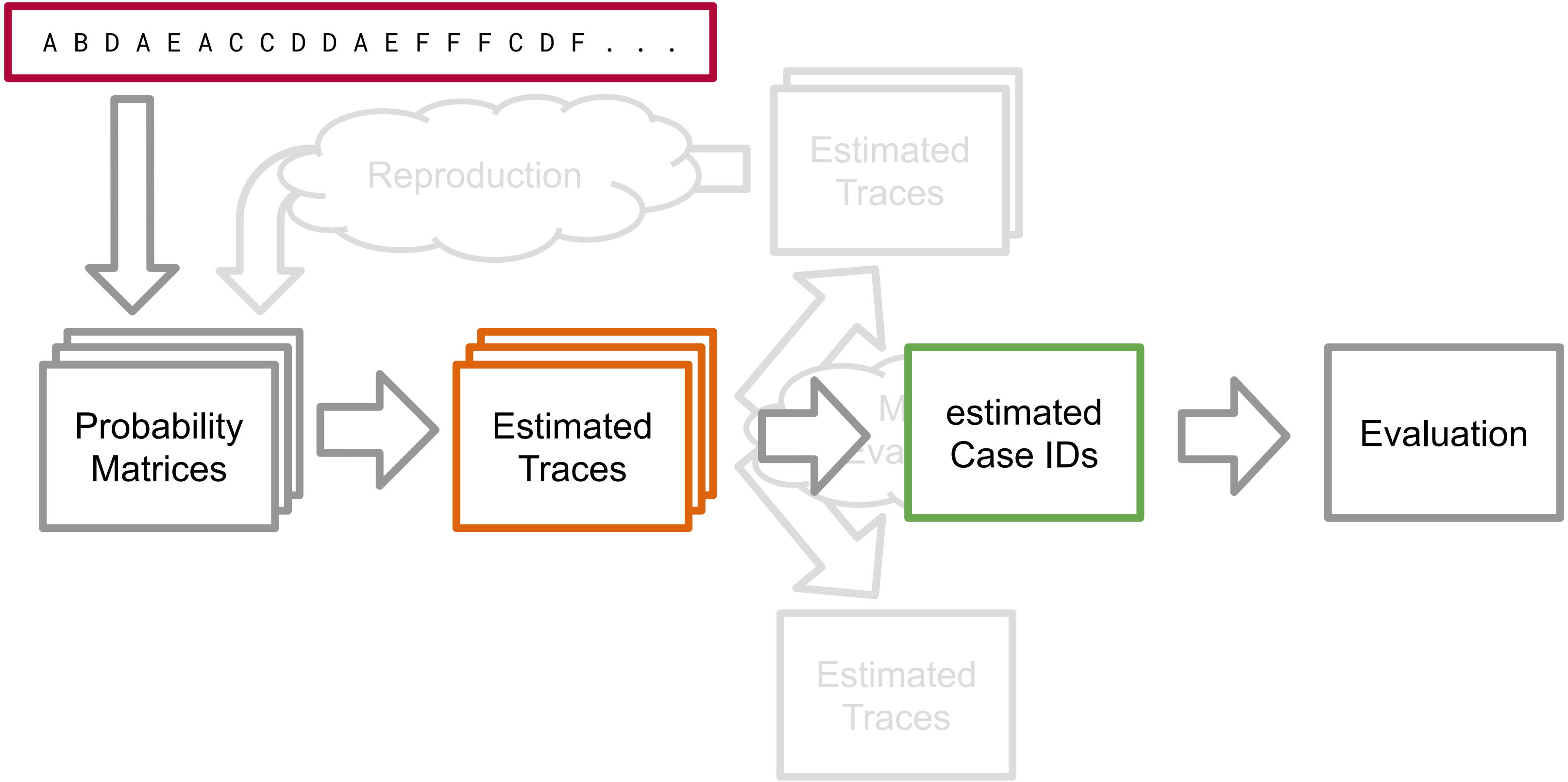
Genetic Programing Paradigm



Genetic Programing Paradigm



Genetic Programing Paradigm



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

M

Eva

estimated Case IDs

Evaluation

Estimated Traces

Implementation

Implementation Disclaimer

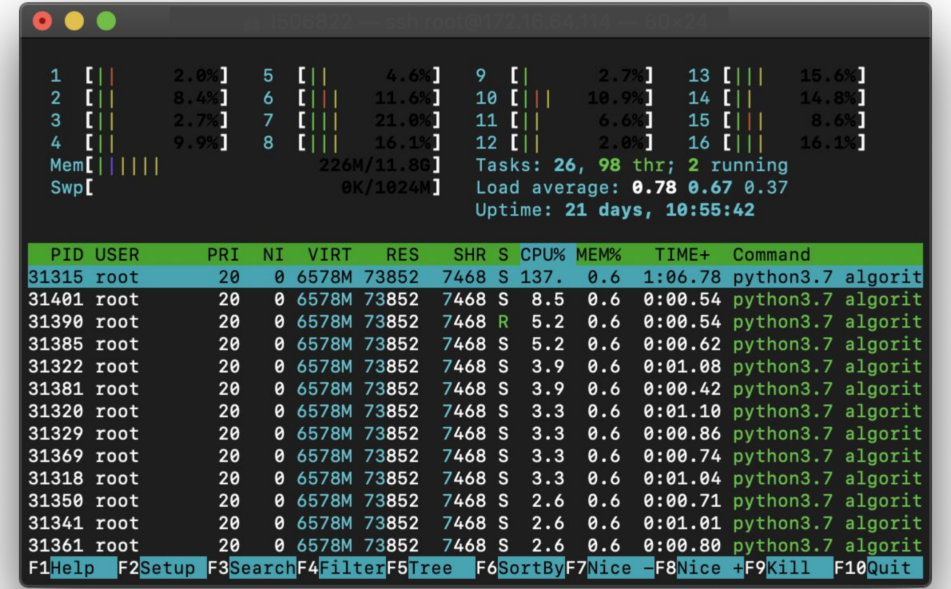
- project core: base algorithm from “Process Discovery of unlabeled Event Logs” [3]
  - minimal edited to make it usable
  - no changes in functionality and logic

- genetic programming extension[1] on the base algorithm [2]

- multiple short and long version experiments were run on the BPT-Chair Server

[1] https://github.com/pscls/genetic-process-discovery

Execution on Server



i506822 — ssh root@172.16.64.114 — 80×24

```
  1  [||       2.0%]   5  [||       4.6%]   9  [|        2.7%]   13 [|||     15.6%]
  2  [||       8.4%]   6  [|||     11.6%]   10 [|||     10.9%]   14 [||      14.8%]
  3  [||       2.7%]   7  [|||     21.0%]   11 [||       6.6%]   15 [|||      8.6%]
  4  [||       9.9%]   8  [|||     16.1%]   12 [||       2.0%]   16 [|||     16.1%]
  Mem[||||||                   226M/11.8G]   Tasks: 26, 98 thr; 2 running
  Swp[                          0K/1024M]   Load average: 0.78 0.67 0.37
                                             Uptime: 21 days, 10:55:42

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
31315 root       20   0 6578M 73852  7468 S 137.  0.6  1:06.78 python3.7 algorit
31401 root       20   0 6578M 73852  7468 S  8.5  0.6  0:00.54 python3.7 algorit
31390 root       20   0 6578M 73852  7468 R  5.2  0.6  0:00.54 python3.7 algorit
31385 root       20   0 6578M 73852  7468 S  5.2  0.6  0:00.62 python3.7 algorit
31322 root       20   0 6578M 73852  7468 S  3.9  0.6  0:01.08 python3.7 algorit
31381 root       20   0 6578M 73852  7468 S  3.9  0.6  0:00.42 python3.7 algorit
31320 root       20   0 6578M 73852  7468 S  3.3  0.6  0:01.10 python3.7 algorit
31329 root       20   0 6578M 73852  7468 S  3.3  0.6  0:00.86 python3.7 algorit
31369 root       20   0 6578M 73852  7468 S  3.3  0.6  0:00.74 python3.7 algorit
31318 root       20   0 6578M 73852  7468 S  3.3  0.6  0:01.04 python3.7 algorit
31350 root       20   0 6578M 73852  7468 S  2.6  0.6  0:00.71 python3.7 algorit
31341 root       20   0 6578M 73852  7468 S  2.6  0.6  0:01.01 python3.7 algorit
31361 root       20   0 6578M 73852  7468 S  2.6  0.6  0:00.80 python3.7 algorit
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

Base Paper Experiment

~ 1h

Paper Equivalent Genetic* Experiment (multithreading)

> 36h

* 10 models and 10 epochs per symbol sequence run

Evaluation

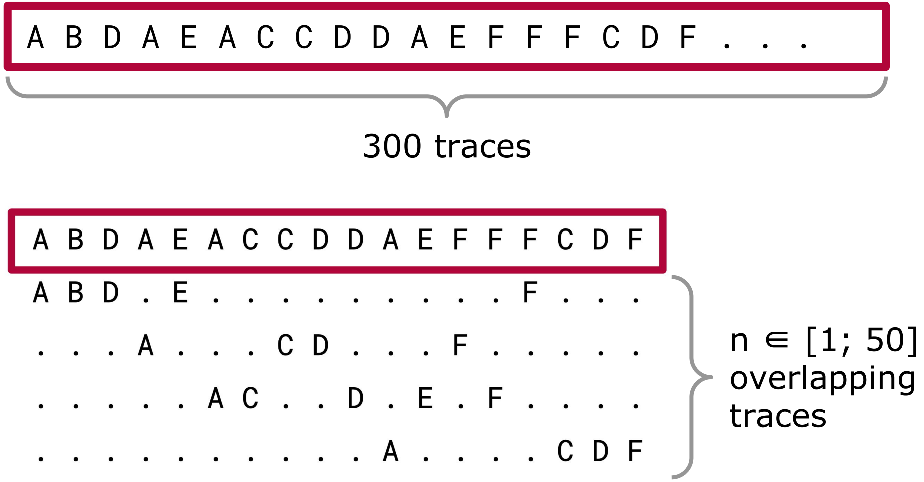
Experiments

Experiment Input

- number of symbol sequences[1]
- number of concurrent overlapping traces[1]
- different process models[1]
- genetic parameters

Output

- trace predictions
- g-score
- g*-score[2]
- performance



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

300 traces

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

n ∈ [1; 50] overlapping traces

$$G(p \,||\, q) \widehat{=} \sum_{z \in Z} \sqrt{p(z) \cdot q(z)}$$

[2]

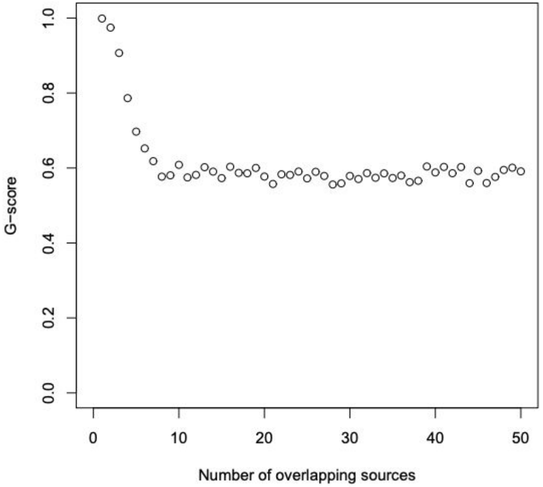
[1] used only for generation or evaluation steps
[2] a more tolerant version of the g-score

Base Algorithm Disclaimer

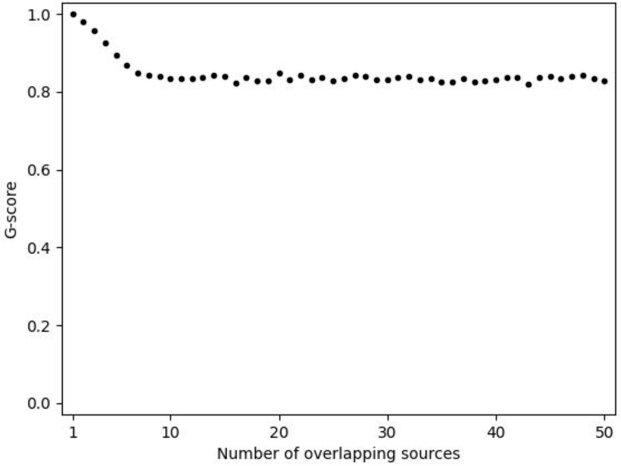
- our vs. paper results [2] differ by about 20% using the same code [3]

- discrepancy was confirmed by the author Ferreira D.R. [2]

Paper Results



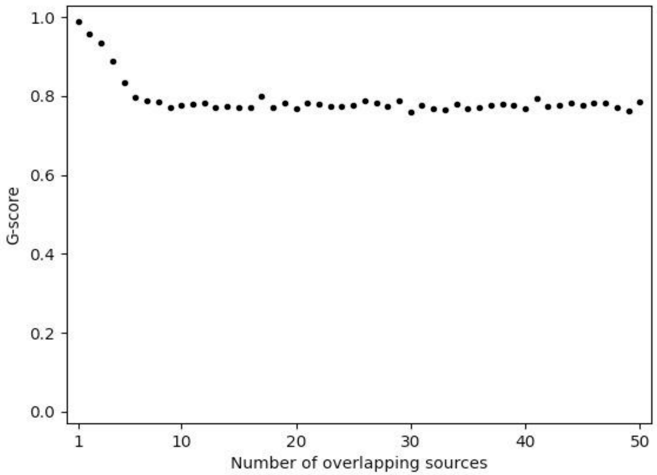
Our Results



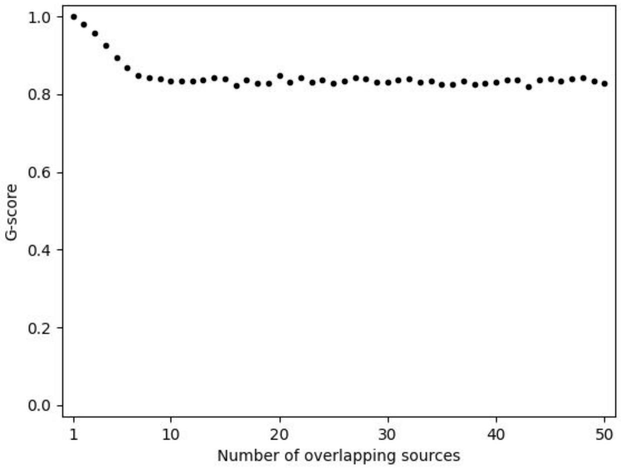
Genetic Algorithm

- ran a dozen of experiments to find best hyperparameter setup (reproduction, mutation, initialization, #models, #epochs) → 4 to 24h each
- still slightly inferior

Genetic Results



Base Algorithm



Explanations

Maybe the base algorithm is as good as it gets

or

Genetic Approach has to be improved

more models?

more epochs?

better initialization of matrices?

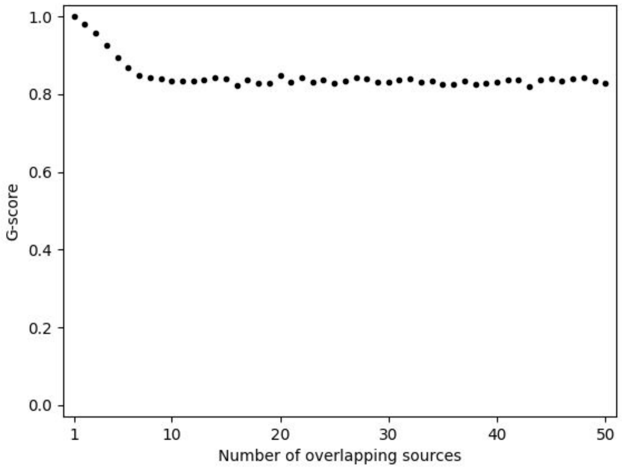
BUT

immensely time-consuming

Experiments would need weeks instead of days

out of scope at this point (future work)

Base Algorithm

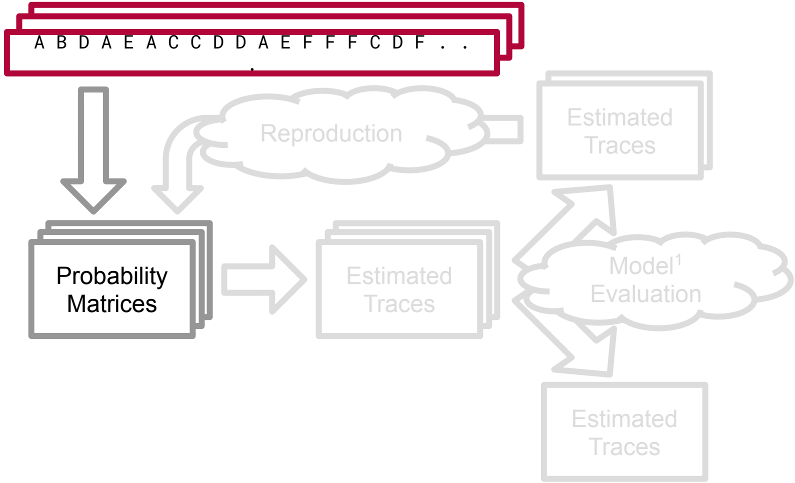


Improvements

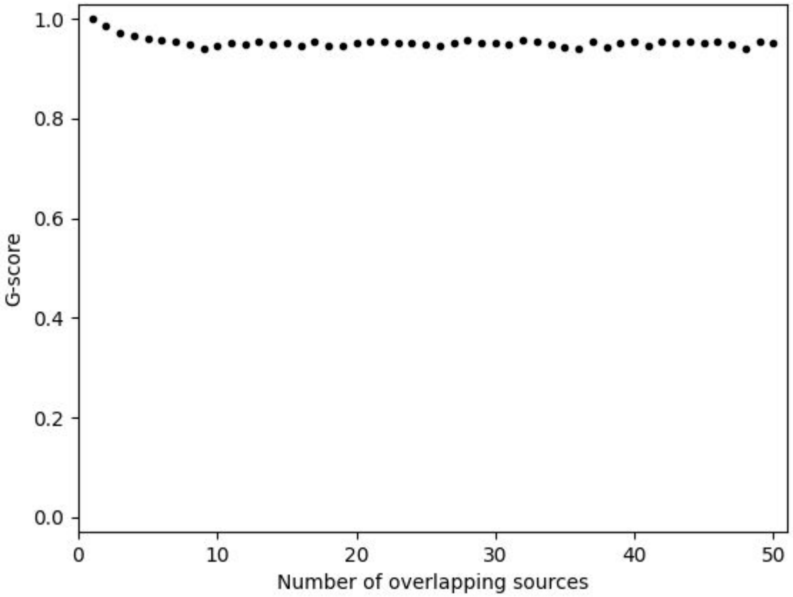
New assumption

- Input of many unlabeled event logs

- Initialization models on different event logs



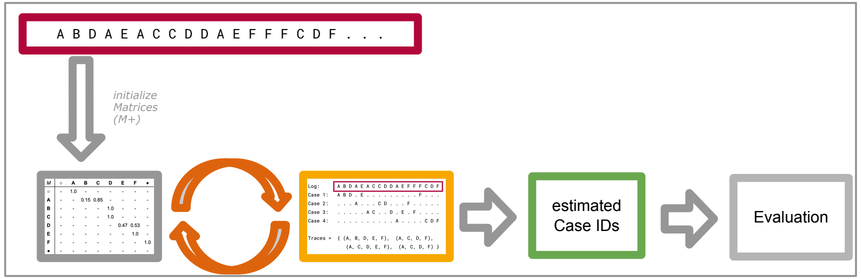
Genetic Algorithm on multiple symbol sequences



Summary

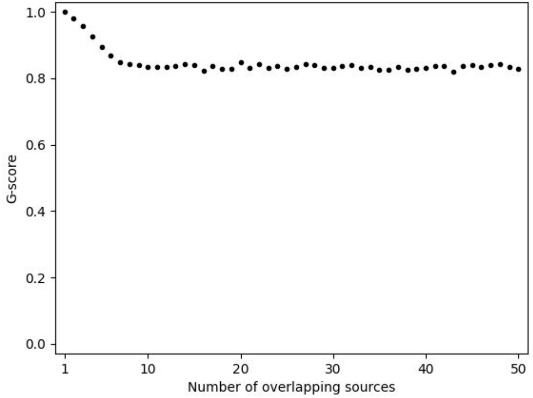
- Conclusion and Findings -

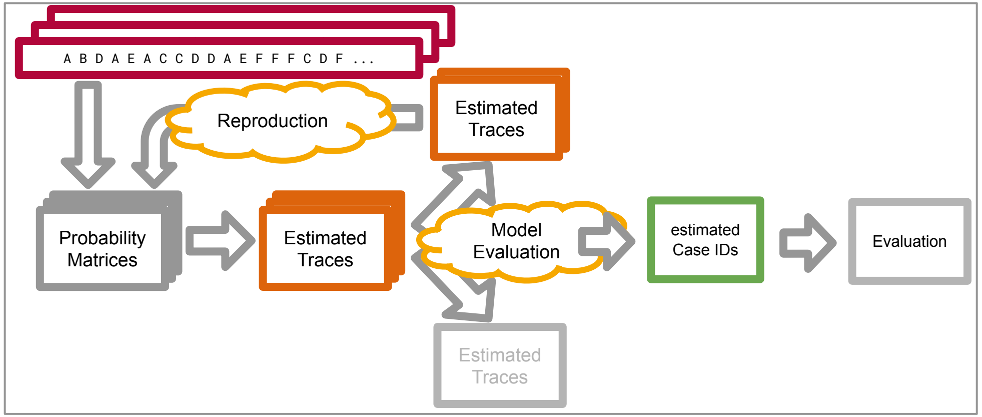
Solving the Correlation Problem

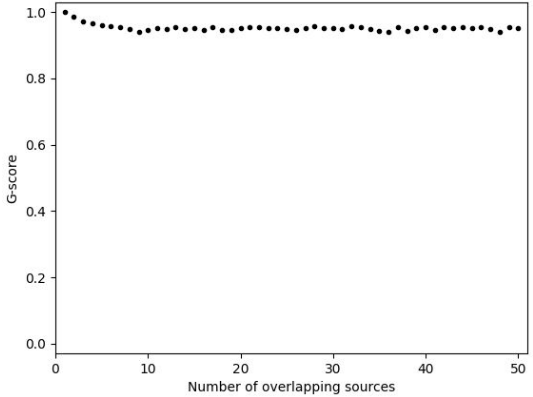


M	○	A	B	C	D	E	F	●
○	-	1.0	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-
B	-	-	-	-	1.0	-	-	-
C	-	-	-	-	1.0	-	-	-
D	-	-	-	-	-	0.47	0.53	-
E	-	-	-	-	-	-	1.0	-
F	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-

```
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) }
```







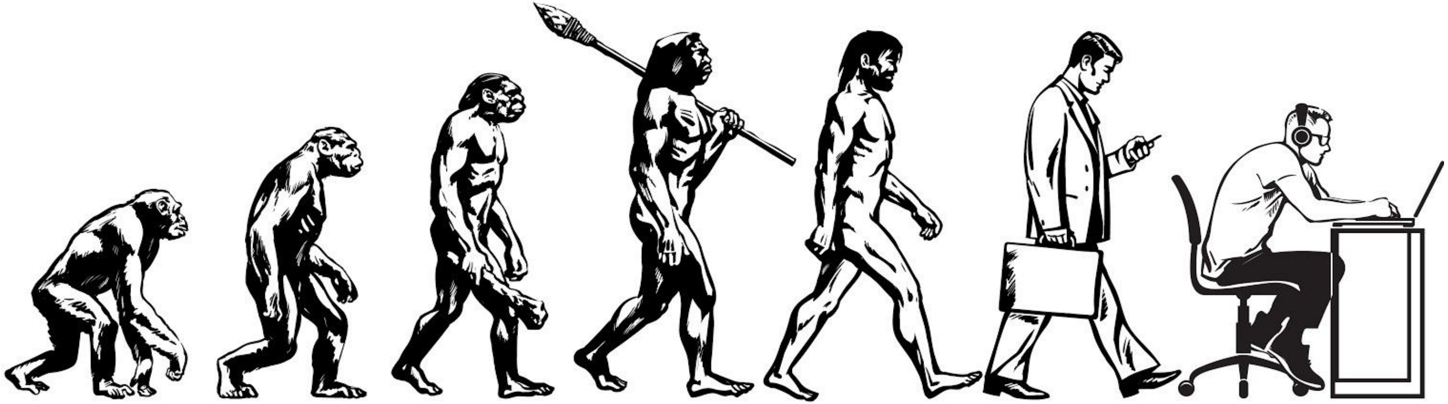
Findings

1. Fitness approximation function for genetic algorithm:
   a. Find best consensus of all models
   b. Metric: g-score
   → Answer to RQ0

2. Accuracy of base algorithm [2] is about 20% better than described

3. Genetic extension is currently slightly inferior
   a. hyperparameter configuration
   b. random initializations?
   c. Problem: runtime
   → RQ1 not confirmed, future work

4. Genetic algorithm can be improved and extended
   → RQ2 future work



Genetic Correlation Discovery for unlabeled Event Logs

Pascal Schulze, Anjo Seidel
Data Extraction for Process Mining (ST-2020)
Supervisor: Simon Remy
13.08.2020

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

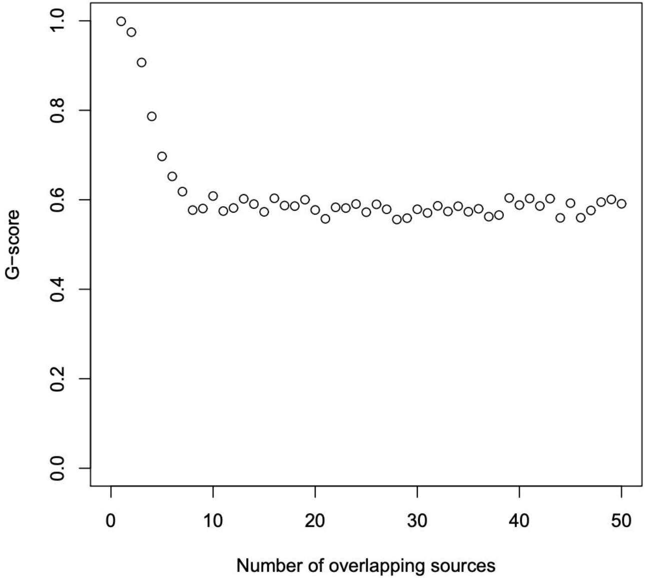
[5] https://medium.com/ssense-tech/schema-evolution-in-data-lakes-f956c6f978d4

Apendix

The Approach - Restrictions

- Assumptions to the process
  - no loops
  - no parallelism
  - start and end event is fixed

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



[2]

Motivation

ABDAEACCDDAEFFFCDF...

Case ID	Event	Attributes
?	A	…
?	B	…
?	D	…
?	A	…
?	E	…
?	A	…
?	C	…
?	C	…
?	D	…
?	D	…
?	A	…

Motivation

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

```
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) }
```

Case ID	Event	Attributes
1	A	…
1	B	…
1	D	…
2	A	…
1	E	…
3	A	…
3	C	…
2	C	…
2	D	…
3	D	…
4	A	…

Other interesting Findings

- Some bad input sequences lower the G-Score
- Some bad input sequences can not be estimated correctly
    - if the unlabelled event log does not represent the process correctly → longer input sequences

Metrics

- Input:
  - length of symbol sequence
  - number of overlapping traces
  - genetic parameters

- Output:
  - G-Score
  - G*-Score
  - Performance

Experiments

- Input
  - 1000 symbol sequences each contains 300 sources
  - [1; 50] overlapping traces
  - → 50,000 individual runs
    - Base: 50,000 Models
    - Genetic: 500,000 Models over 10 epochs (+ reproduction and fitness operations)
  - Different Models
    - without loops
    - with loops

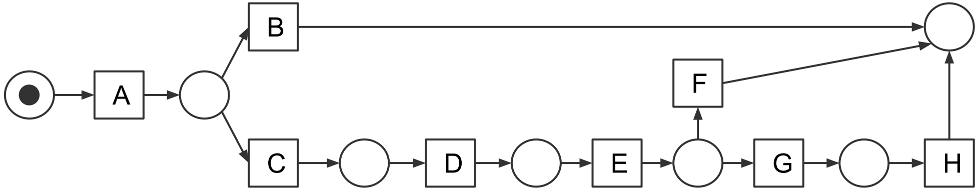
Research Questions

0. **How can models be evaluated/ranked without further data (ground truth)?**
   - with a fitness function comparing one instance with all other instances based on the G-Score

1. **Can the precision of this approach be improved by using a Genetic Programming Paradigm and other metrics?**
   - The present approach is better than initially assumed
   - A genetic approach is assumably not better in an acceptable running time

2. **Can assumptions for this approach be overcome with a Genetic Approach?**
   - Future Work (Paper)

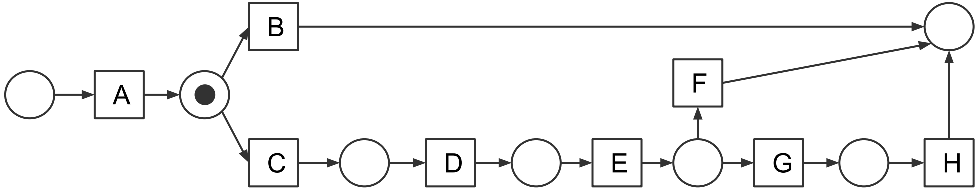
Idea 2 - Token Replay



A B D E F

c	0
p	1
m	0
r	0

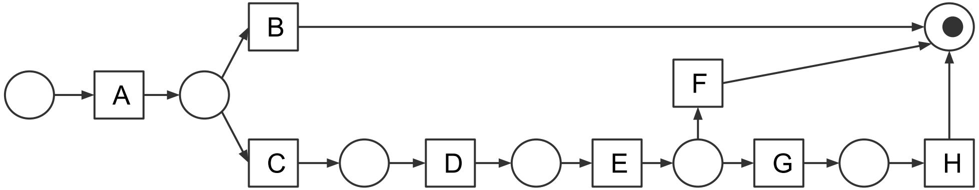
Idea 2 - Token Replay



A B D E F

c	1
p	2
m	0
r	0

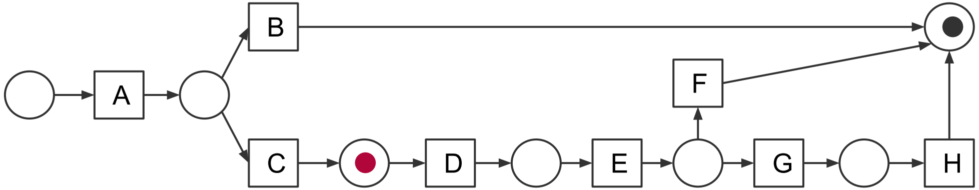
Idea 2 - Token Replay



A B D E F

c	2
p	3
m	0
r	0

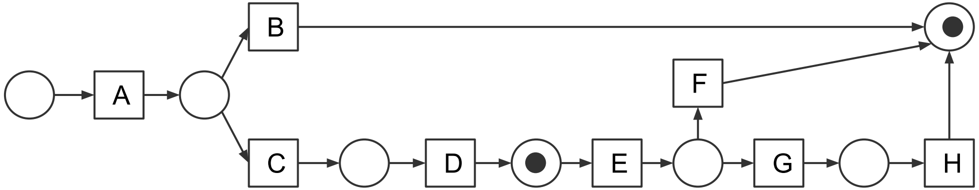
Idea 2 - Token Replay



A B D E F

c	2
p	3
m	1
r	0

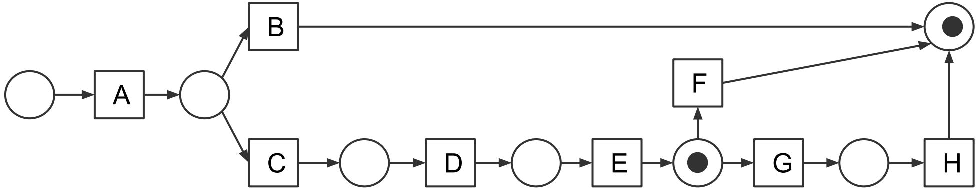
Idea 2 - Token Replay



A B D E F

c	3
p	4
m	1
r	0

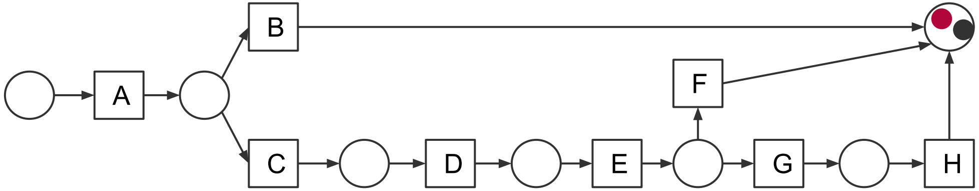
Idea 2 - Token Replay



c	4
p	5
m	1
r	0

A B D E F

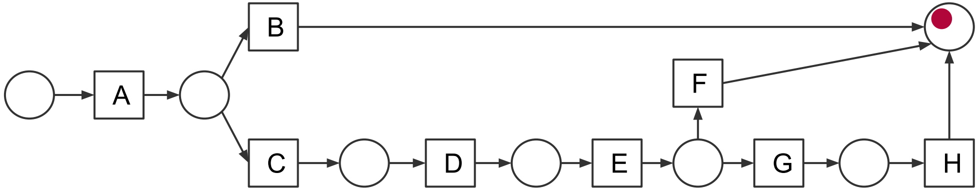
Idea 2 - Token Replay



A B D E F

c	5
p	6
m	1
r	0

Idea 2 - Token Replay



c	6
p	6
m	1
r	1

A B D E F

$$f = \frac{1}{2}(1 - \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i c_i}) + \frac{1}{2}(1 - \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i p_i})$$

- We have many traces
- We have probabilities/ weighted edges