Expectation-Maximization Techniques for Process Mining

Diogo R. Ferreira

http://web.tagus.ist.utl.pt/~diogo.ferreira/

LOIS Workshop: Process Mining meets Data Mining

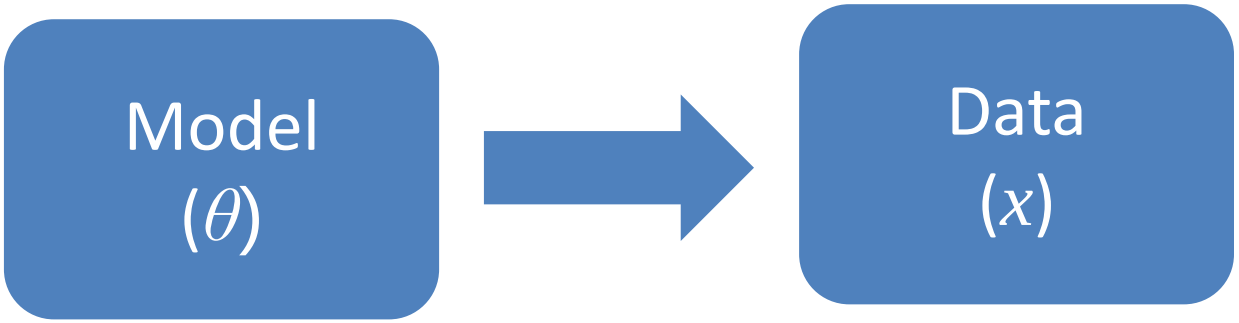
October 28, 2009 | Technische Universiteit Eindhoven (TU/e)

Agenda

- Overview of EM
- Applications to process mining
  - sequence clustering
  - mining without case ids

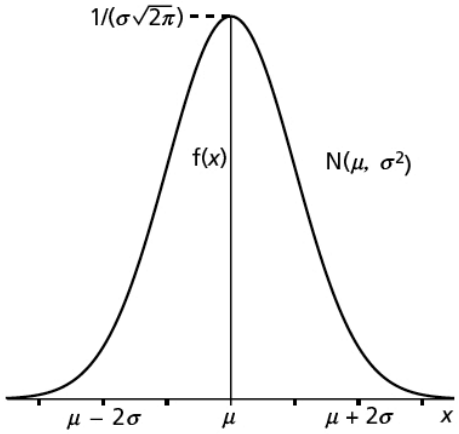
Expectation-Maximization

- What is EM used for?

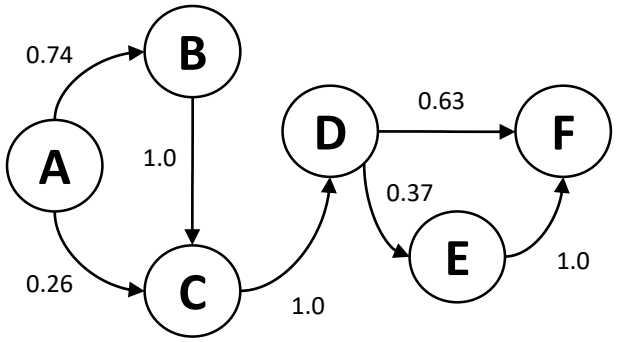


$$\theta \Longleftarrow x$$

Expectation-Maximization



$$\begin{cases} x_1 = 2.5 \\ x_2 = 2.45 \\ x_3 = 2.53 \\ x_4 = 1.99 \\ x_5 = 2.62 \\ \dots \end{cases}$$



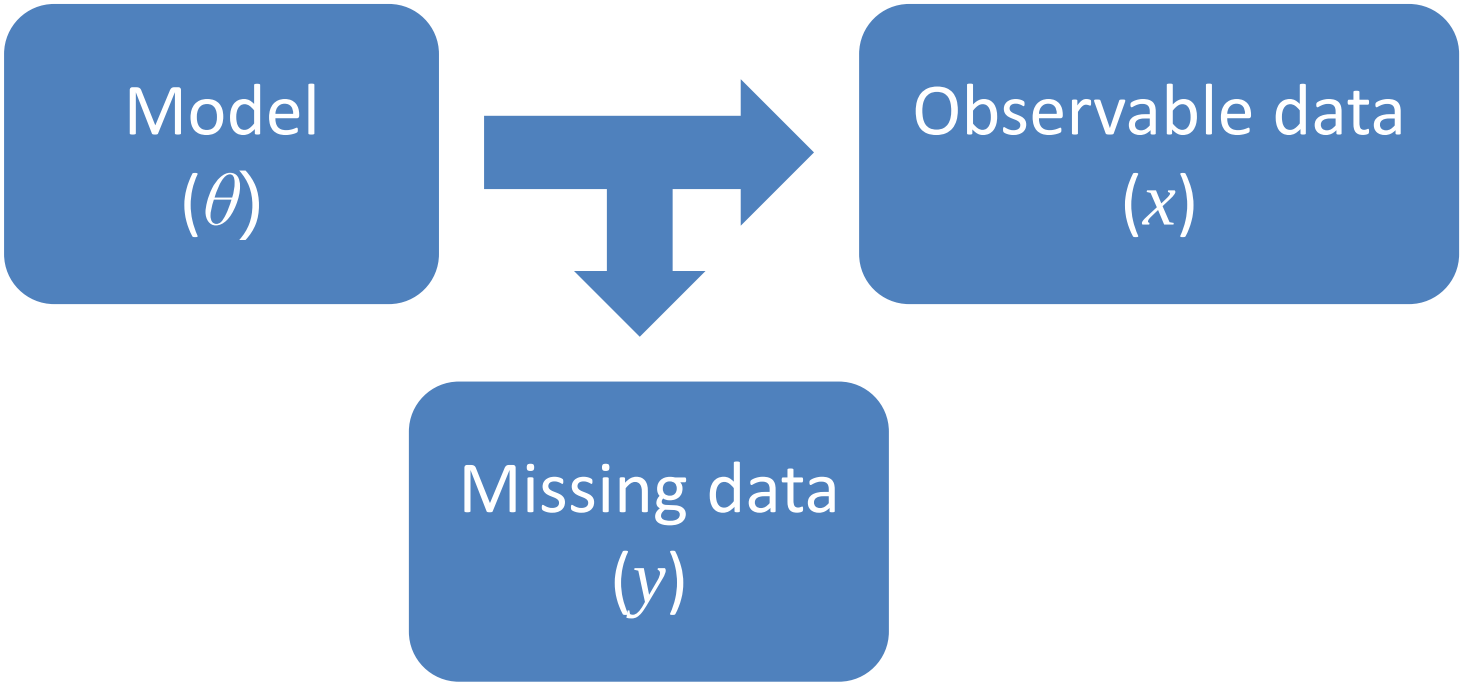
$$\begin{cases} x_1 = ABCDEF \\ x_2 = ACDE \\ x_3 = BCDF \\ x_4 = CDF \\ x_5 = DE \\ \dots \end{cases}$$

Expectation-Maximization

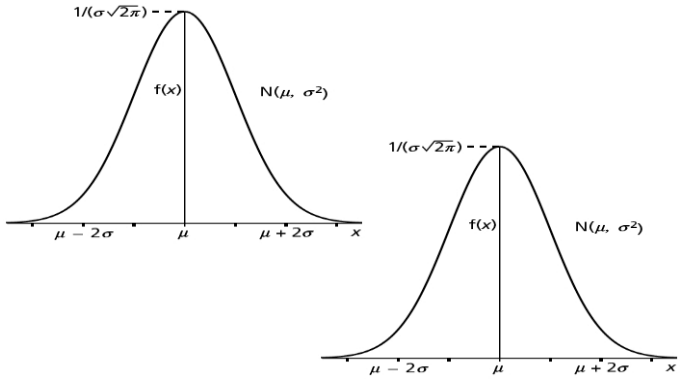
- Probability $P(x|\theta)$
- Likelihood $L(\theta) = \log P(x|\theta)$
- Goal find θ that maximizes $L(\theta)$

Expectation-Maximization

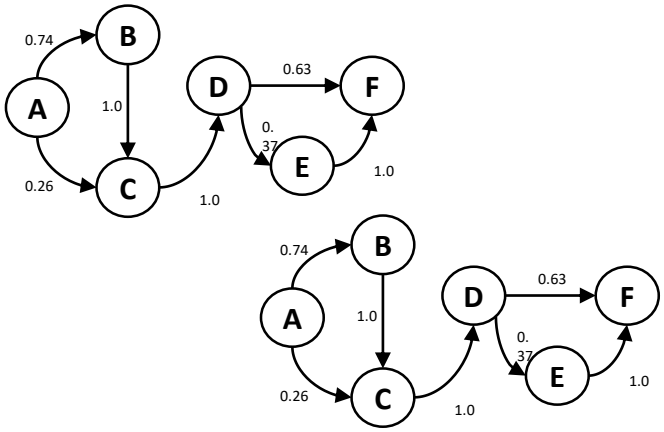
- The missing data y



Expectation-Maximization



$$\begin{cases} x_1 = 2.5 \\ x_2 = 2.45 \\ x_3 = 2.53 \\ x_4 = 1.99 \\ x_5 = 2.62 \\ \dots \end{cases}$$



$$\begin{cases} x_1 = ABCDEF \\ x_2 = ACDE \\ x_3 = BCDF \\ x_4 = CDF \\ x_5 = DE \\ \dots \end{cases}$$

Expectation-Maximization

- Problem would be easy with complete data

$$x, y \rightarrow \theta \qquad P(x, y \mid \theta) \qquad L_c(\theta) = \log P(x, y \mid \theta)$$

- Incomplete-data likelihood

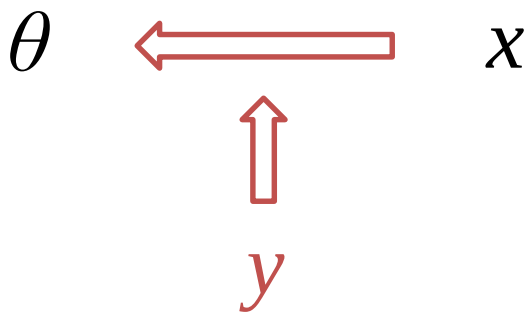
$$L(\theta) = \log P(x \mid \theta)$$

- Same goal: find θ that maximizes $L(\theta)$

Expectation-Maximization

- Problem

How to get from x to θ
when y is necessary
but y is unknown?



Expectation-Maximization

- Solution by EM
  - initial estimate θ'

$$\left\{ \begin{array}{l} x, \theta' \rightarrow y' \\ x, y' \rightarrow \theta'' \end{array} \right.$$

e-step

m-step

  - initial estimate y'

$$\left\{ \begin{array}{l} x, y' \rightarrow \theta' \\ x, \theta' \rightarrow y'' \end{array} \right.$$

m-step

e-step

Expectation-Maximization

- The math
  - with an initial estimate θ' the ***expectation*** is defined as

$$Q(\theta \mid \theta') = E_{y|x;\theta'}[L_c(\theta)]$$
$$= E_{y|x;\theta'}[\log P(x, y \mid \theta)]$$
$$= \sum_y P(y \mid x; \theta') \cdot \log P(x, y \mid \theta)$$

  - ***maximization*** means finding θ'' such as: $\frac{\partial Q}{\partial \theta} = 0$

Expectation-Maximization

- But there is a problem…

$$Q(\theta \mid \theta') = \sum_{y} P(y \mid x; \theta') \cdot \log P(x, y \mid \theta)$$

This is a sum over all possible values of y !!...

Expectation-Maximization

- so we approximate and consider only

$$\hat{y} = \arg\max_{y} \{P(y \mid x; \theta')\}$$

- giving

$$Q(\theta \mid \theta') = \sum_{y} P(y \mid x; \theta') \cdot \log P(x, y \mid \theta)$$

$$\cong P(\hat{y} \mid x; \theta') \cdot \log P(x, \hat{y} \mid \theta)$$

Expectation-Maximization

- and now we have

$$\hat{y} = \arg\max_{y} \{P(y \mid x; \theta')\}$$

E-step

$x, \theta' \rightarrow y'$

$$\theta'' = \arg\max_{\theta} \{Q(\theta \mid \theta')\}$$

$$= \arg\max_{\theta} \{P(\hat{y} \mid x; \theta') \cdot \log P(x, \hat{y} \mid \theta)\}$$

$$= \arg\max_{\theta} \{\log P(x, \hat{y} \mid \theta)\}$$

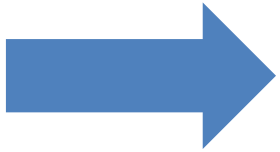
M-step

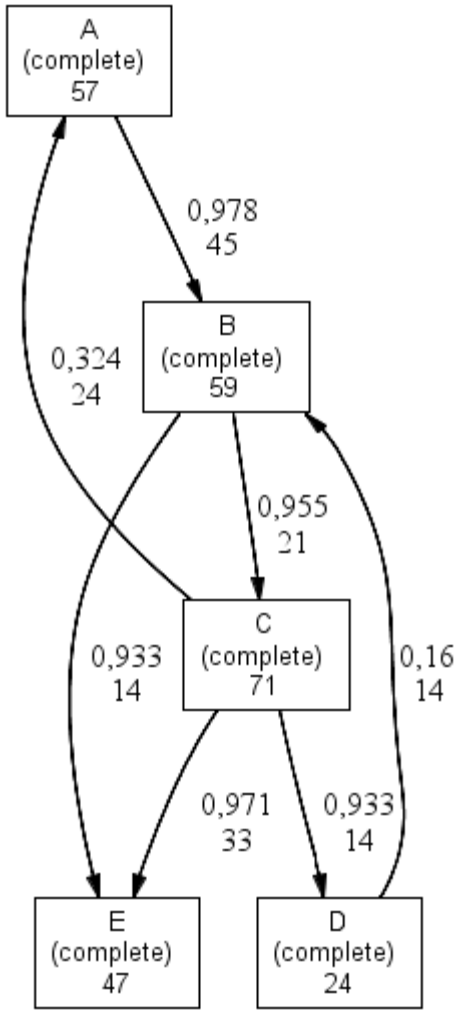
$x, y' \rightarrow \theta''$

Application: Sequence Clustering

Process Mining

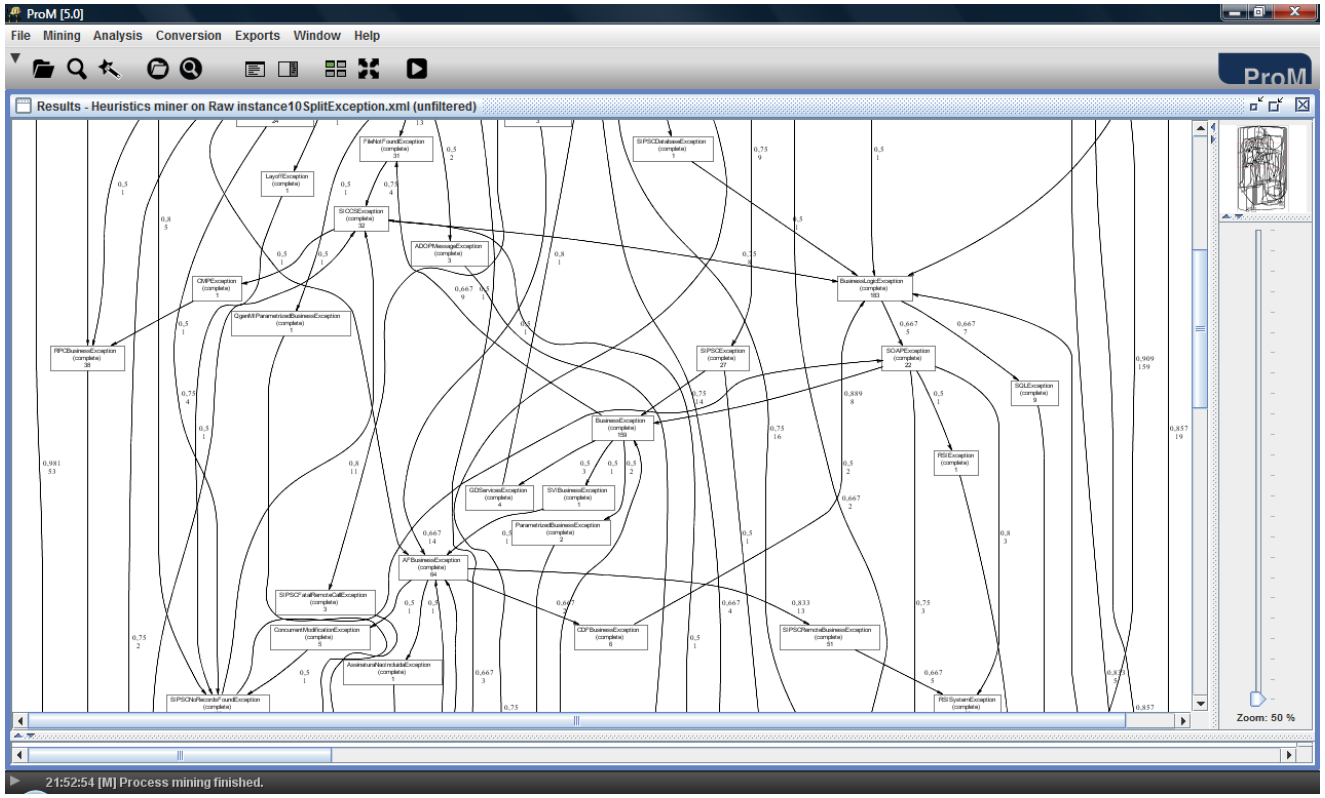
sequence	*no. cases*
ABCE	21
ACE	12
CABD	10
CAB	14
CBDE	14



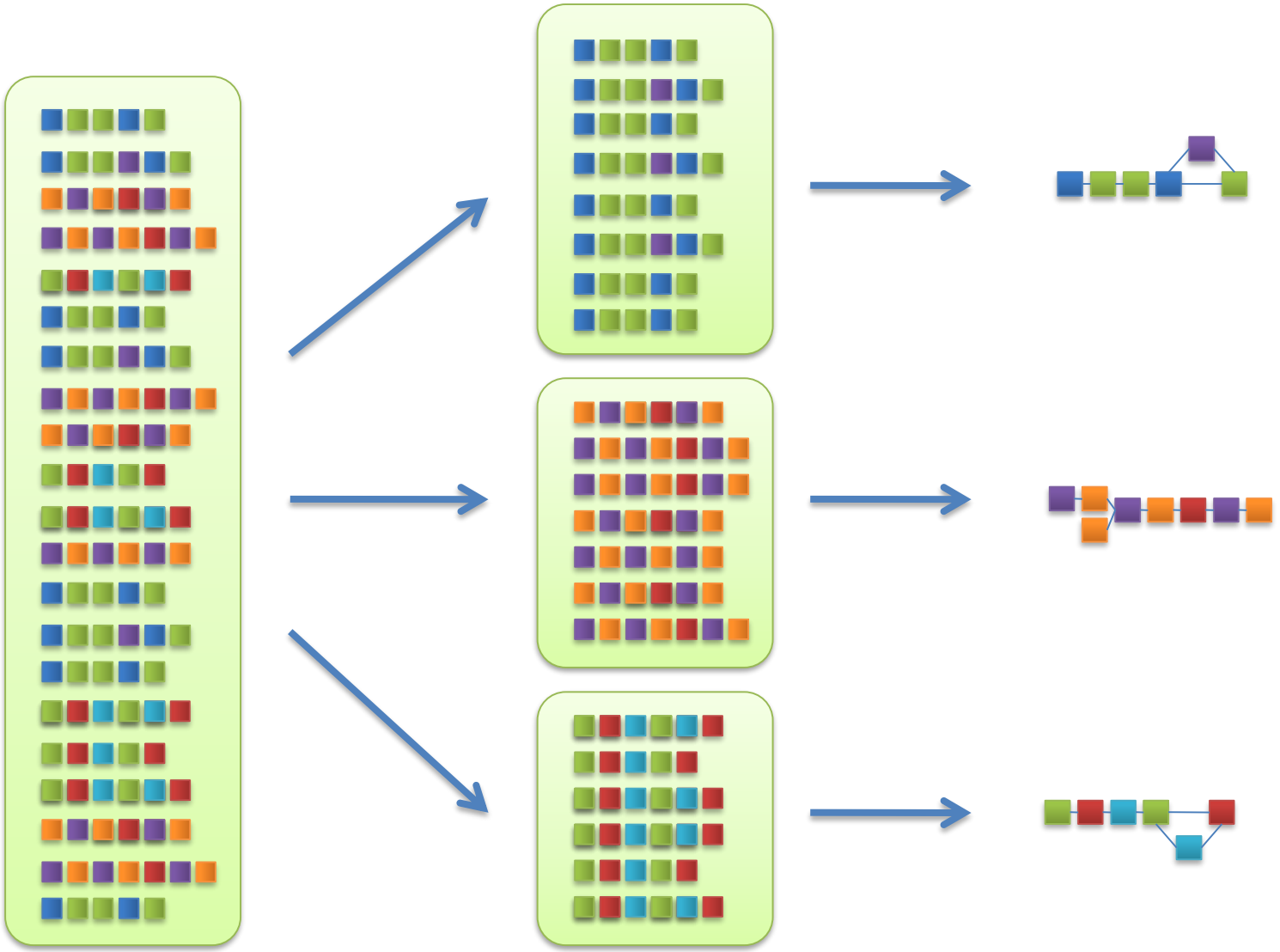


Process Mining

- Spaghetti models



Clustering

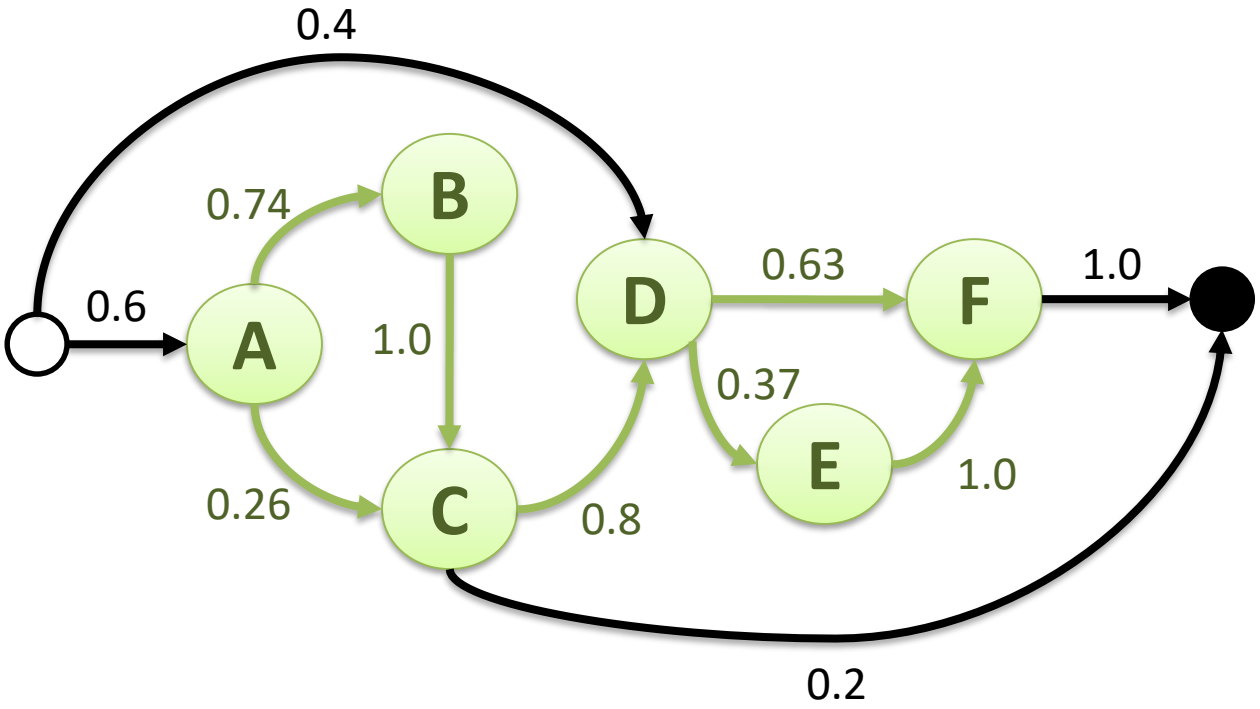


Clustering

- DWS mining plug-in [Medeiros et al., 2007]
  - heuristics miner + k-means clustering
- Trace clustering [Song et al., 2008]
  - feature-based clustering (k-means, SOM, etc.)
  - string-based edit distance [Bose/Aalst, 2009]
- Sequence clustering [Ferreira et al., 2007]
  - clustering based on first-order Markov chains

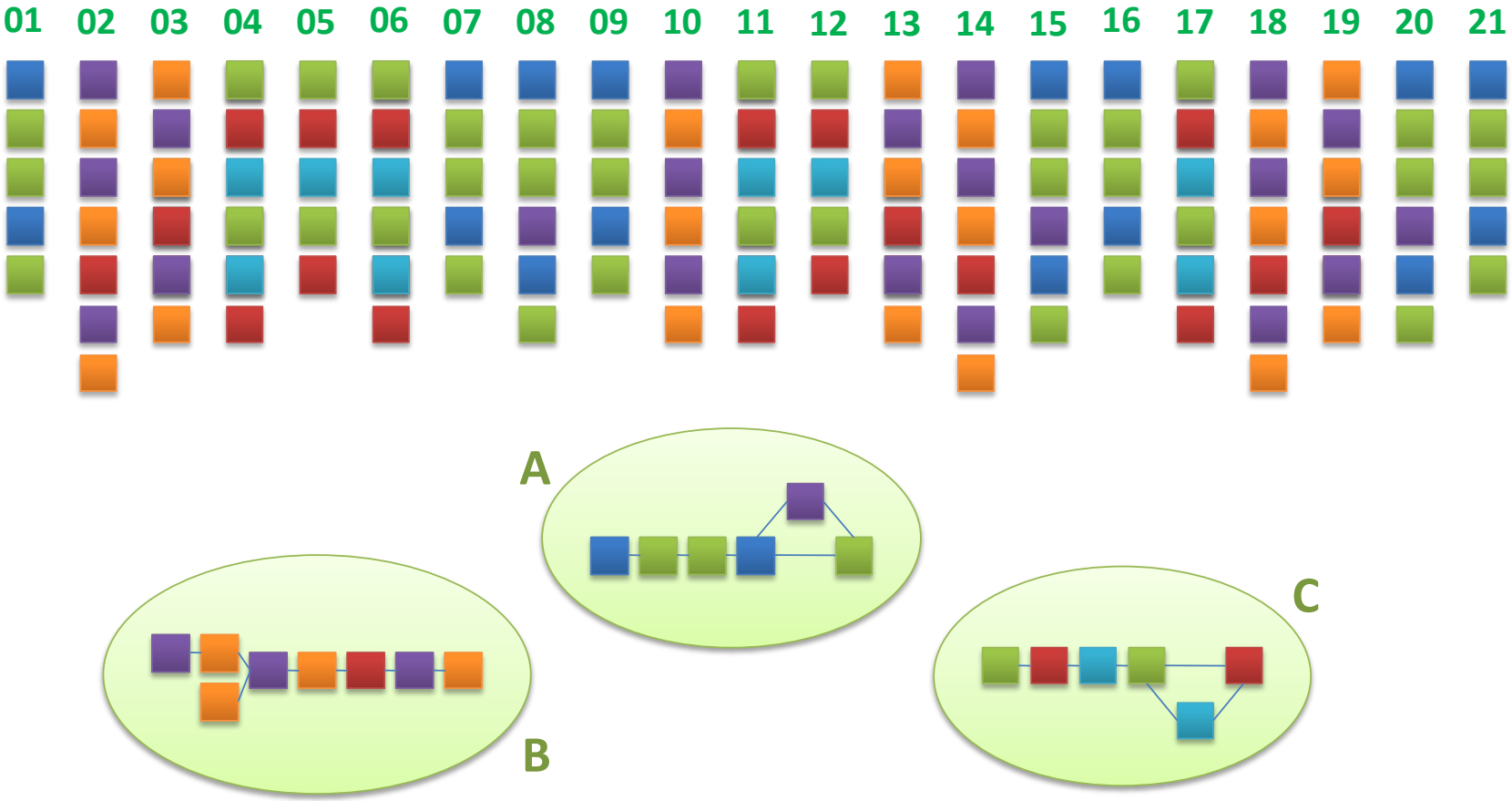
Sequence Clustering

- Cluster model



- AC●
- ABC●
- ABC●
- ACDF●
- ACDEF●
- ABCDF●
- ABCDEF●
- DF●
- DEF●

Sequence Clustering



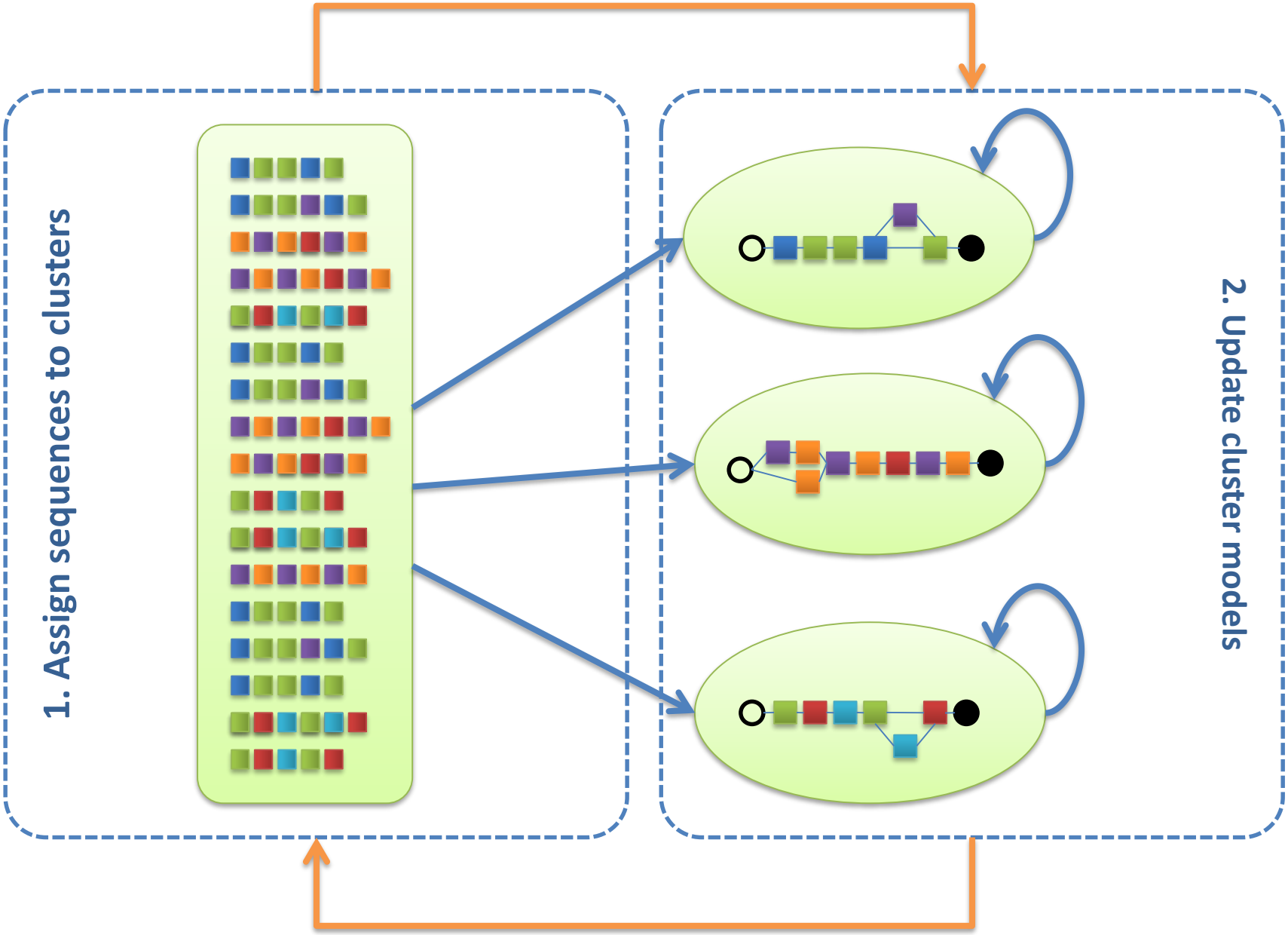
Sequence Clustering

- Missing data y

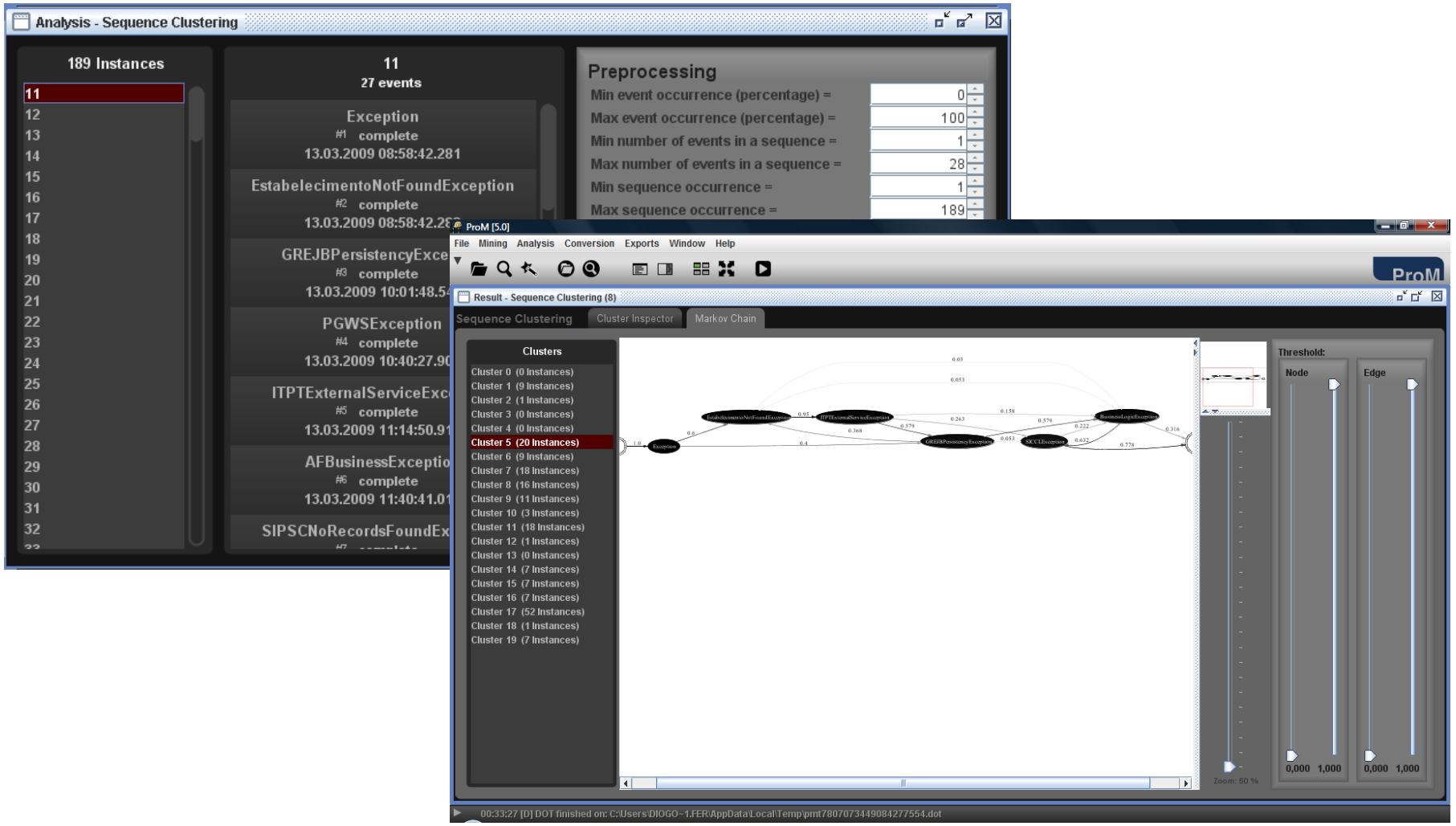
sequence	*no. cases*	*cluster*
ABCE	21	?
ACE	12	?
CABD	10	?
CAB	14	?
CBDE	14	?

$$\begin{cases} x, \theta' \rightarrow y' \\ x, y' \rightarrow \theta'' \end{cases}$$

$$\begin{cases} x, y' \rightarrow \theta' \\ x, \theta' \rightarrow y'' \end{cases}$$


1. Assign sequences to clusters
2. Update cluster models

Sequence Clustering



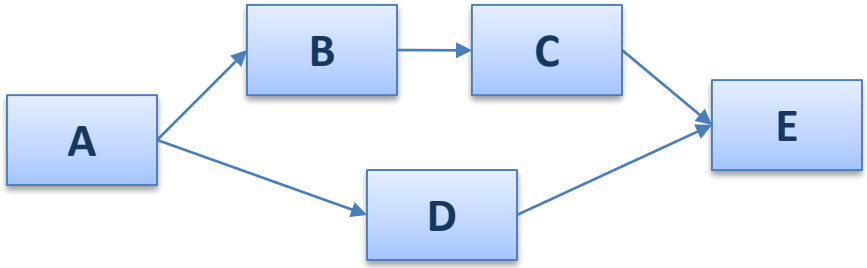
Application: Mining without case ids

Process Mining

event log

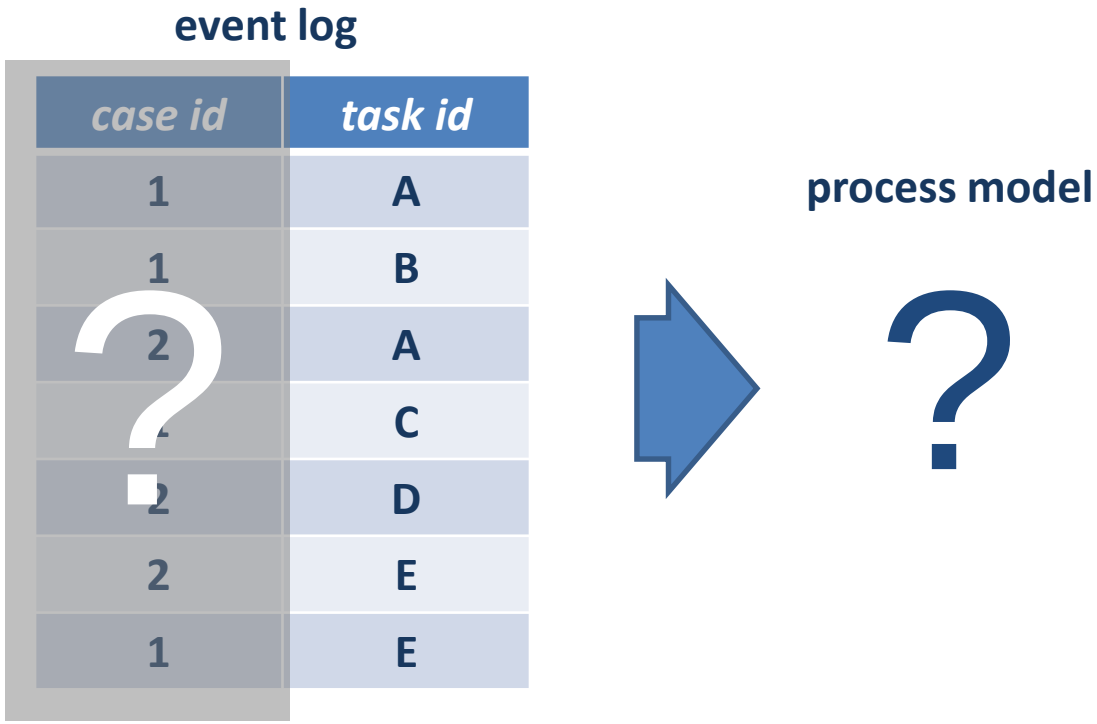
case id	*task id*
1	A
1	B
2	A
1	C
2	D
2	E
1	E
…	…

process model

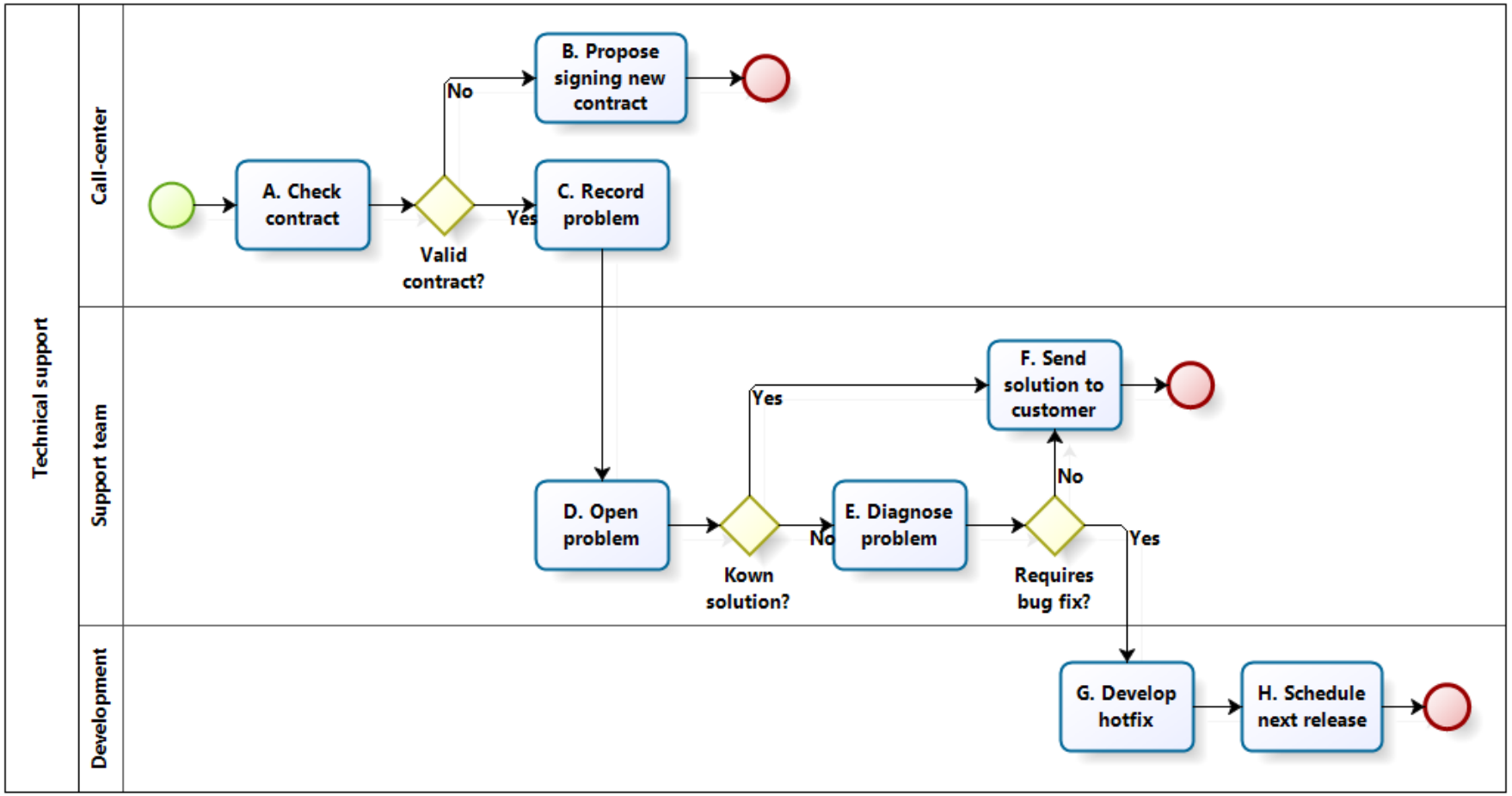


Process Mining

- Missing data y

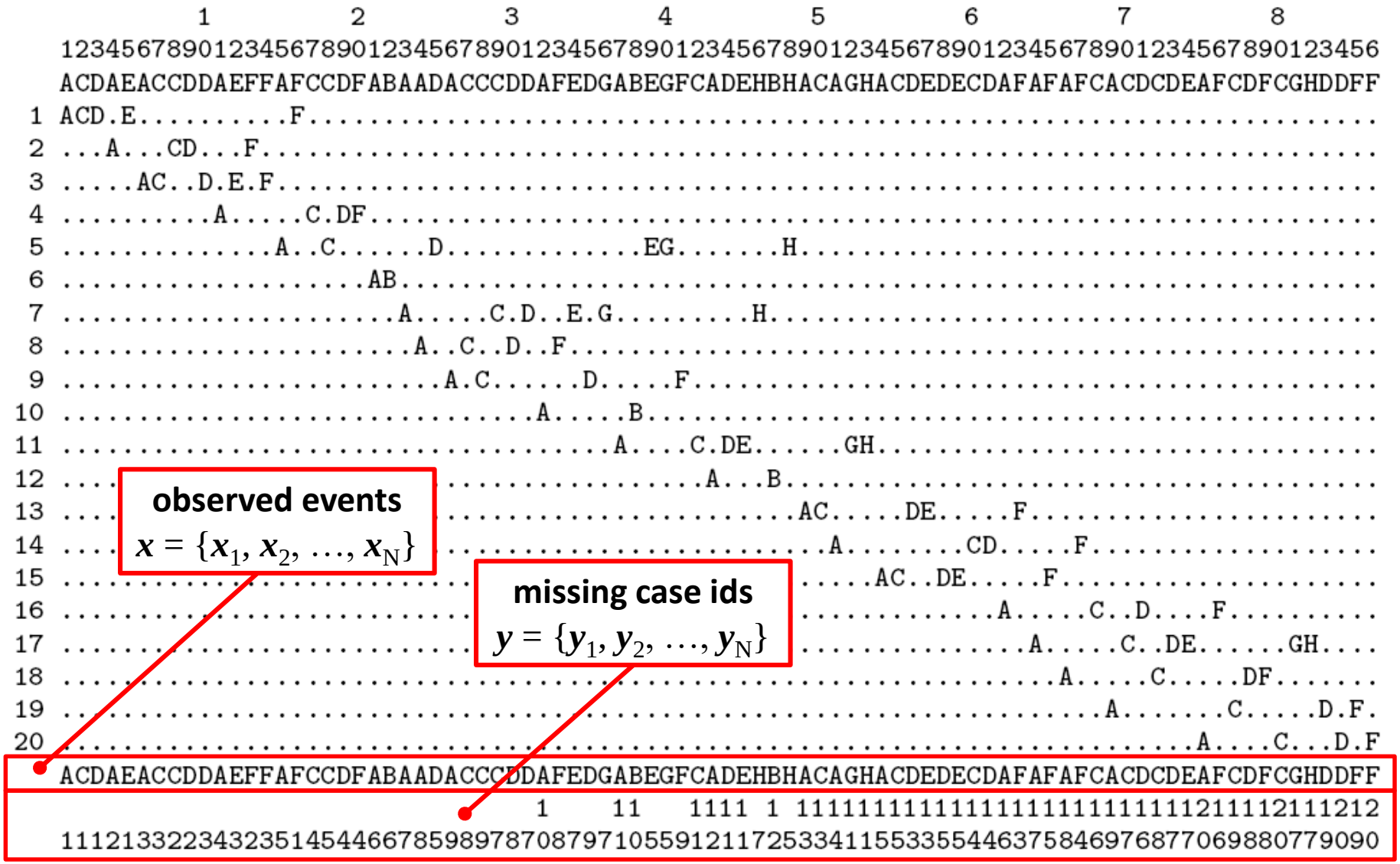


Example



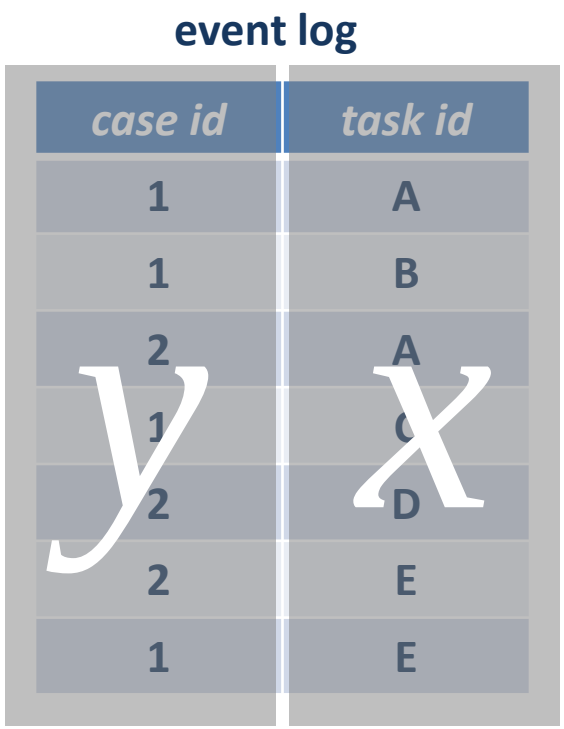
AB ACDF ACDEF ACDEGH

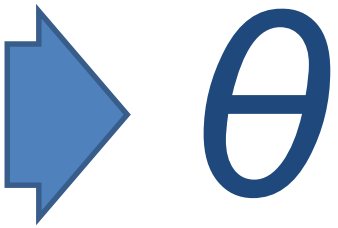
Example


observed events
$x = \{x_1, x_2, \ldots, x_N\}$
missing case ids
$y = \{y_1, y_2, \ldots, y_N\}$

Missing case ids

- Solution by EM





process model

θ

$$\left\{ \begin{array}{l} x, \theta' \rightarrow y' \\ x, y' \rightarrow \theta'' \end{array} \right.$$

$$\left\{ \begin{array}{l} x, y' \rightarrow \theta' \\ x, \theta' \rightarrow y'' \end{array} \right.$$

Approach

- First problem $x, y \rightarrow \theta$ **M**-step
- Second problem $x, \theta \rightarrow y$ **E**-step

First problem: $x, y \rightarrow \theta$

- Same as process mining with case id

x ACDAEACCDDAEFFAFCCDFABAADACCCDDAFEDGABEGFCADEHBHACAGHACDEDECDAFAFAFCACDCDEAFCDFCGHDDFF

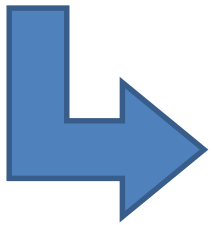
y 1112133223432351454466785989787087971055912117253341155335544637584697687706988077909 0

1 11 1111 1 1111111111111111111111111111211112111212

y	z
1, 3, 13, 15	○ACDEF●
2, 4, 8, 9, 14, 16, 18, 19, 20	○ACDF●
5, 7, 11, 17	○ACDEGH●
6, 10, 12	○AB●

First problem: $x, y \rightarrow \theta$

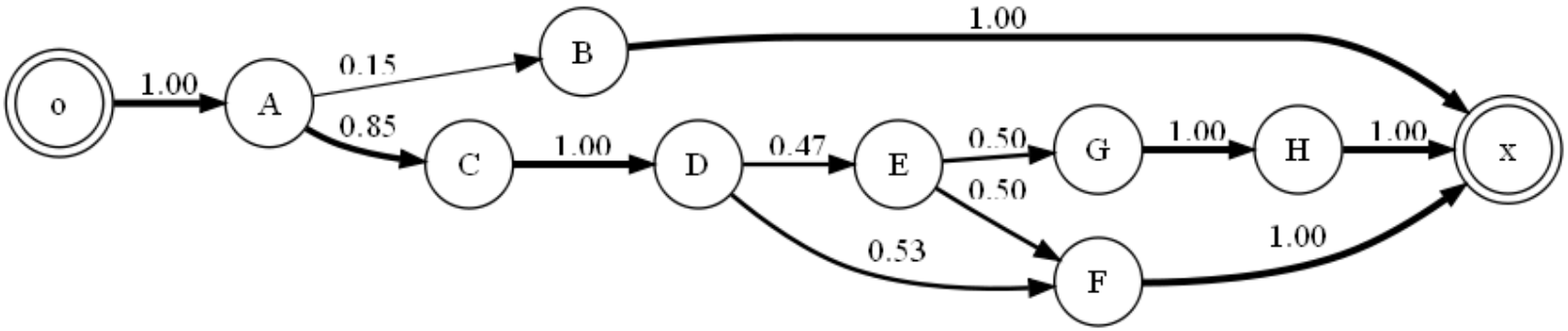
y	z
1, 3, 13, 15	○ACDEF●
2, 4, 8, 9, 14, 16, 18, 19, 20	○ACDF●
5, 7, 11, 17	○ACDEGH●
6, 10, 12	○AB●



θ	○	A	B	C	D	E	F	G	H	●
○	-	1.0	-	-	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-	-	-
B	-	-	-	-	-	-	-	-	-	1.0
C	-	-	-	-	1.0	-	-	-	-	-
D	-	-	-	-	-	0.47	0.53	-	-	-
E	-	-	-	-	-	-	0.5	0.5	-	-
F	-	-	-	-	-	-	-	-	-	1.0
G	-	-	-	-	-	-	-	-	1.0	-
H	-	-	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-	-	-

First problem: $x, y \rightarrow \theta$

θ	○	A	B	C	D	E	F	G	H	●
○	-	1.0	-	-	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-	-	-
B	-	-	-	-	-	-	-	-	-	1.0
C	-	-	-	-	1.0	-	-	-	-	-
D	-	-	-	-	-	0.47	0.53	-	-	-
E	-	-	-	-	-	-	0.5	0.5	-	-
F	-	-	-	-	-	-	-	-	-	1.0
G	-	-	-	-	-	-	-	-	1.0	-
H	-	-	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-	-	-



Approach

- First problem $x, y \rightarrow \theta$

- Second problem $x, \theta \rightarrow y$

Second problem: $x, \theta \rightarrow y$

$\boldsymbol{x}$ = A C D A E A C C D D A E F F A F C C D F …

$\boldsymbol{y}$ = 1 1 1 2 1 3 3 2 2 3 4 3 2 …

θ	○	A	B	C	D	E	F	G	H	●
○	-	1.0	-	-	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-	-	-
B	-	-	-	-	-	-	-	-	-	1.0
C	-	-	-	-	1.0	-	-	-	-	-
D	-	-	-	-	-	0.47	0.53	-	-	-
E	-	-	-	-	-	-	0.5	0.5	-	-
F	-	-	-	-	-	-	-	-	-	1.0
G	-	-	-	-	-	-	-	-	1.0	-
H	-	-	-	-	-	-	-	-	-	1.0
●	-	-	-	-	-	-	-	-	-	-

$$\hat{y} = \arg\max_{y} \{P(y \mid x; \theta')\}$$

$$x, \theta \rightarrow y$$

- Greedy algorithm

θ	○	A	B	C	D	E	F	G	H	●
○	-	1.0	-	-	-	-	-	-	-	-
A	-	-	0.15	0.85	-	-	-	-	-	-
B	-	-	-	-	-	-	-	-	-	1.0
C	-	-	-	-	1.0	-	-	-	-	-
D	-	-	-	-	-	0.47	0.53	-	-	-
E	-	-	-	-	-	-	0.5	0.5	-	-
F	-	-	-	-	-	-	-	-	-	1.0
G	-	-	-	-	-	-	-	-	1.0	-
H	-	-	-	-	-	-	-	-	-	1.0
	-	-	-	-	-	-	-	-	-	-

x	active	y	1	2	3	…
A	{ }	1	○A	-	-	-
C	{ 1}	1	○AC	-	-	-
A	{ 1}	2	○AC	○A	-	-
D	{ 1, 2}	1	○ACD	○A	-	-
B	{ 1, 2}	2	○ACD	○AB●	-	-
E	{ 1}	1	○ACDE	○AB●	-	-
A	{ 1}	3	○ACDE	○AB●	○A	-
F	{ 1,3}	1	○ACDEF●	○AB●	○A	-
…	…	…	…	…	…	…

Approach

- First problem $x, y \rightarrow \theta$ 
- Second problem $x, \theta \rightarrow y$ 
- how to get the initial estimate θ' or y' ? 

Initialization

- initialize $\theta = \theta'$ directly from x

```
ACAABAACCACDDDFCEFDFEFDAFCDEGHAAAAABBCACCDDFDACAEGDCFDEEAGCGHHCDEADCACDEFDFEHAGAFACACHDCDECFADAFDCAFEGCHDAECF
AADDCBCFDAFEDAEABCFDACADCECFDFDEFEFGHABACDEAGAACBACDHECADCFADCDAFACDFCDFFADEEBFGAAAACHCCCDDDFFFDEFACADECDEGHA
AACDEFGCHACADDCEADFECACFADECGDEGHFHDACCDFAFDCDFABAAACAFCCDADEECGCDDEEHFDFAGAFCADCDACCDHEDAFAEBFACFAEDCDAECBAF
GECDHDFFAEFACDABCDAECAADFCDAECFEFFDFAABACDCBACDEEFFDEGACACHAAADDBFAAECCFCCDADDDEFACFGHCEDFAEGADHFBCDEACDAAEFA
AFCDACCCCDFEFADBAACDEDFAEBGCACDHEADFDEFFCFAACDEBDAFCDACDGHFEGHACAABADABCAEFBDEACGCAADCDBEAGHCDAHFECDDFEAACFCD
DEEFEGHABGAAACAHDCDBBACEAFCDAEFCDFDEEGHABGHAAACACACDDCECDDDEEEFFFGABHFAACCDDAFAECCFDEFDAAABECFBDACFDEACDEFACF
DFAACACAACCCDEDDDFEEABADGFCACDFFHDAAAECCFCDFACDEAFDFFDFCAADCAECAFDCADDEEBCFAADCFDCEFEGHAACACFDDDAFECEGGDFAHCH
AACDEGHACCCADFDACFDCDDDAEFAFECFACACCGHDDFDEFDFAAACCFAACDDEDECFGAHDBCDEGAHCDEEFFFAACDACFABBDAFAACDBBF
```

θ' =

	○	A	B	C	D	E	F	G	H	●
○	-	-	-	-	-	-	-	-	-	-
A	-	0.23	**0.07**	**0.32**	0.13	0.10	0.10	0.03	0.03	-
B	-	0.30	0.11	0.22	0.11	0.05	0.11	0.08	0.03	-
C	-	0.20	0.02	0.13	**0.36**	0.05	0.17	0.03	0.04	-
D	-	0.20	0.04	0.13	0.16	**0.24**	**0.19**	0.01	0.03	-
E	-	0.14	0.05	0.22	0.07	0.12	**0.24**	**0.14**	0.02	-
F	-	0.30	0.02	0.13	0.24	0.13	0.10	0.06	0.02	-
G	-	0.27	-	0.16	0.08	0.03	0.03	0.03	**0.41**	-
H	-	0.35	-	0.16	0.22	0.08	0.16	-	0.03	-
●	-	-	-	-	-	-	-	-	-	-

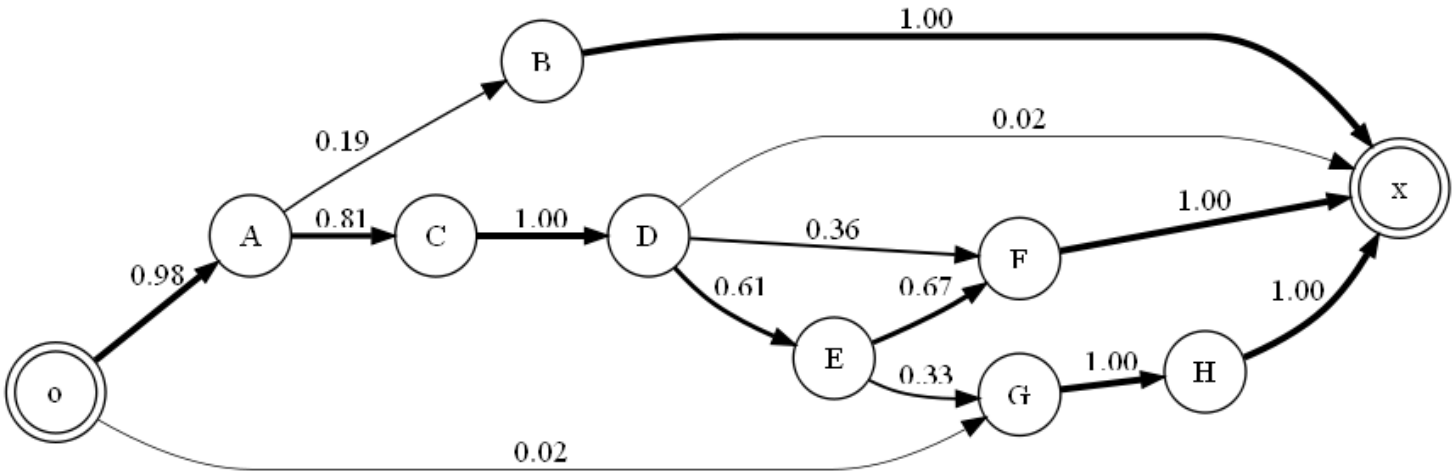
Mining without case ids

- EM approach:
  1. initialize θ
  2. repeat
  3. from $\boldsymbol{x}$ and θ estimate $\boldsymbol{y}$
  4. from $\boldsymbol{x}$ and $\boldsymbol{y}$ estimate θ
  5. until θ converges

Example

θ =

	○	A	B	C	D	E	F	G	H	●
○	-	0.98	-	-	-	-	-	0.02	-	-
A	-	-	0.19	0.81	-	-	-	-	-	-
B	-	-	-	-	-	-	-	-	-	1.00
C	-	-	-	-	1.00	-	-	-	-	-
D	-	-	-	-	-	0.61	0.36	-	-	0.02
E	-	-	-	-	-	-	0.67	0.33	-	-
F	-	-	-	-	-	-	-	-	-	1.00
G	-	-	-	-	-	-	-	-	1.00	-
H	-	-	-	-	-	-	-	-	-	1.00
●	-	-	-	-	-	-	-	-	-	-



Example

- comparing the results

true model

z	$p(z)$
ACDEF	30.0 %
ACDF	30.0 %
AB	20.0 %
ACDEGH	20.0 %

estimated model

z	$q(z)$
ACDEF	32.8 %
ACDF	28.9 %
AB	18.1 %
ACDEGH	16.2 %
ACD	2.0 %
GH	2.0 %

Evaluation the results

- define the G-score as

$$G(p \parallel q) \triangleq \sum_{z \in \mathbb{Z}} \sqrt{p(z) \cdot q(z)}$$

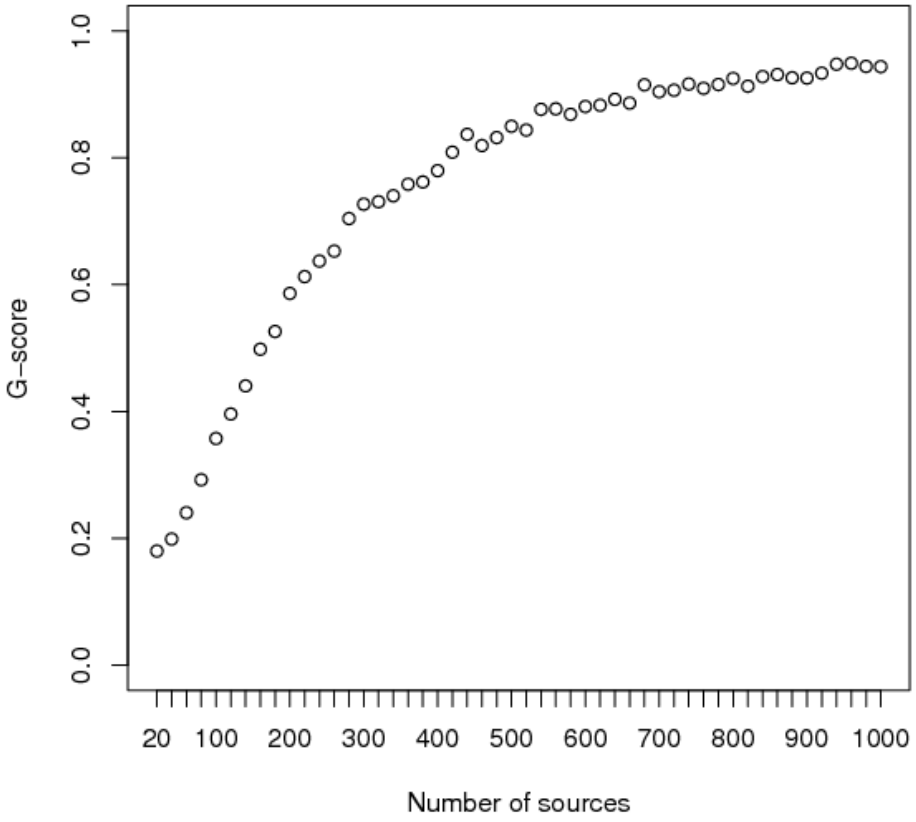
z	$p(z)$
ACDEF	30.0 %
ACDF	30.0 %
AB	20.0 %
ACDEGH	20.0 %

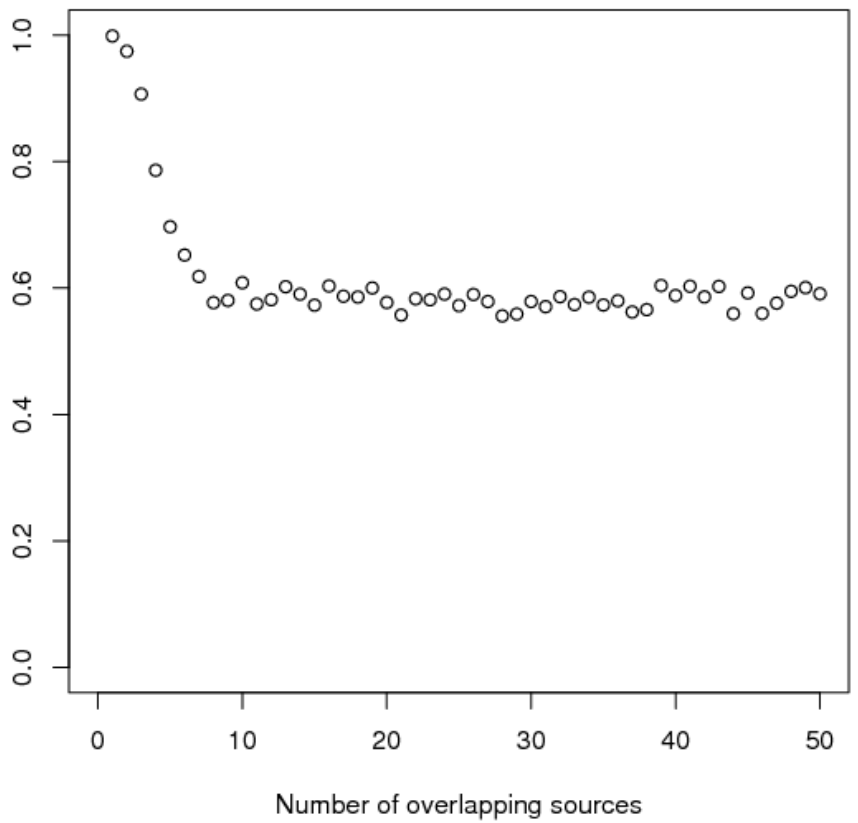
z	$q(z)$
ACDEF	32.8 %
ACDF	28.9 %
AB	18.1 %
ACDEGH	16.2 %
ACD	2.0 %
GH	2.0 %

97.8 %

Evaluating the results

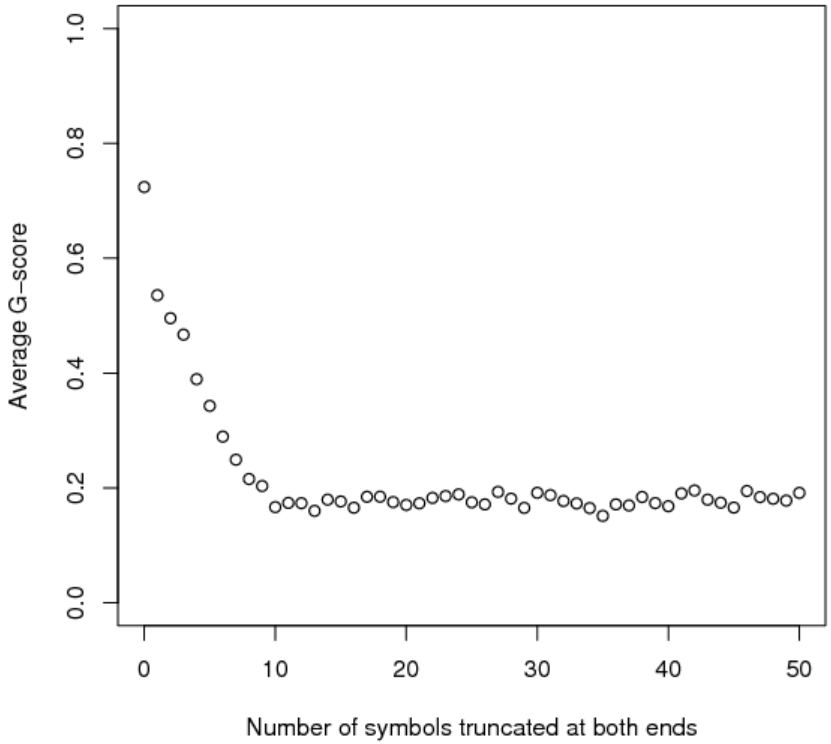
- varying number of instances and overlap





Evaluating the results

- truncation of symbols in x



Results on other test sets

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

(using G*, an extension to the G-score)

Duplicate tasks

- in the estimated model θ, no instance is allowed to produce the same event twice
  - reduces the search space
  - easier for EM to converge

Pattern	$p(\boldsymbol{z})$	No. symbol sequences	Average G-score	Best G-score	Best $q(\boldsymbol{z})$
Duplicate tasks	BDE : 24 / 61 $\simeq$ 0.393 AABHF : 7 / 61 $\simeq$ 0.115 CHF : 15 / 61 $\simeq$ 0.246 ADBE : 6 / 61 $\simeq$ 0.098 ACBGDFAA : 1 / 61 $\simeq$ 0.016 ABEDA : 8 / 61 $\simeq$ 0.131	1000	0.196	0.591	BDE : 0.381 A : 0.355 CHF : 0.169 BHF : 0.056 BD : 0.010 B : 0.009 F : 0.009 G : 0.009 DE : 0.002 BE : 0.001

(test set from Rozinat/Aalst)

Conclusion

Conclusion

- EM is a powerful framework to deal with missing data
  - many problems can be cast into such form
  - other authors have been using EM techniques as well without explicitly referring it (e.g. Buffett & Geng, BPI 2008)

More info…

- G. M. Veiga, D. R. Ferreira, “Understanding Spaghetti Models with Sequence Clustering for ProM”, BPI 2009 Workshop, Ulm, Germany
- D. R. Ferreira, D. Gillblad, “Discovering Process Models from Unlabelled Event Logs”, BPM 2009, Ulm, Germany