CS 4072 – Topics in CS
Process Mining

Lecture # 12

April 04, 2022

Spring 2022

FAST – NUCES, CFD Campus

Dr. Rabia Maqsood

rabia.maqsood@nu.edu.pk

Today's Topics

- Quality issues related to Event logs
  - Noise and Incompleteness
- Quality issues related to Event logs
  - Fitness, Simplicity, Precision and Generalization

Quality issues related to Event logs

- To discover a suitable process model, it is assumed that the event log contains a *representative sample of behavior*.

- Two related phenomena:
  - **Noise:** the event log contains rare and infrequent behavior not representative for the typical behavior of the process.

  - **Incompleteness:** the event log contains too few events to be able to discover some of the underlying control-flow structures.

Noise

- In model discovery, “noise” to refer to rare and infrequent behavior (or “outliers”) rather than errors related to event logging.

- An intuitive idea is to filter out noise before model discovery.
  - Algorithms that support this: heuristic mining, genetic mining and fuzzy mining.

Noise

- Starting point for the α-algorithm is the $>_L$ relation.
- Recall that a $>_L$ b if and only if there is a trace in L in which a is directly followed by b.
- To quantify noise, we can define two measures:
  - Support
    - Support of a $>_L$ b based on number of times the pattern ⟨...,a,b,...⟩ appears in the log, e.g., the fraction of cases in which the pattern occurs.
  - Confidence
    - Confidence of a $>_L$ b can be defined by comparing the number of times the pattern ⟨...,a,b,...⟩ appears in the log divided by the frequency of a and b.

Incompleteness

- Like in any data mining or machine learning context one cannot assume to have seen all possibilities in the “training material” (i.e., the event log at hand).
- Process models typically allow for an exponential or even infinite number of different traces (in case of loops).
- Some traces may have a much lower probability than others.
- Therefore, it is unrealistic to assume that every possible trace is present in the event log.

Noise refers to the problem of having “too much data” (describing rare behavior), completeness refers to the problem of having “too little data”.

Incompleteness

- The α-algorithm assumes a relatively weak notion of completeness to avoid this problem.

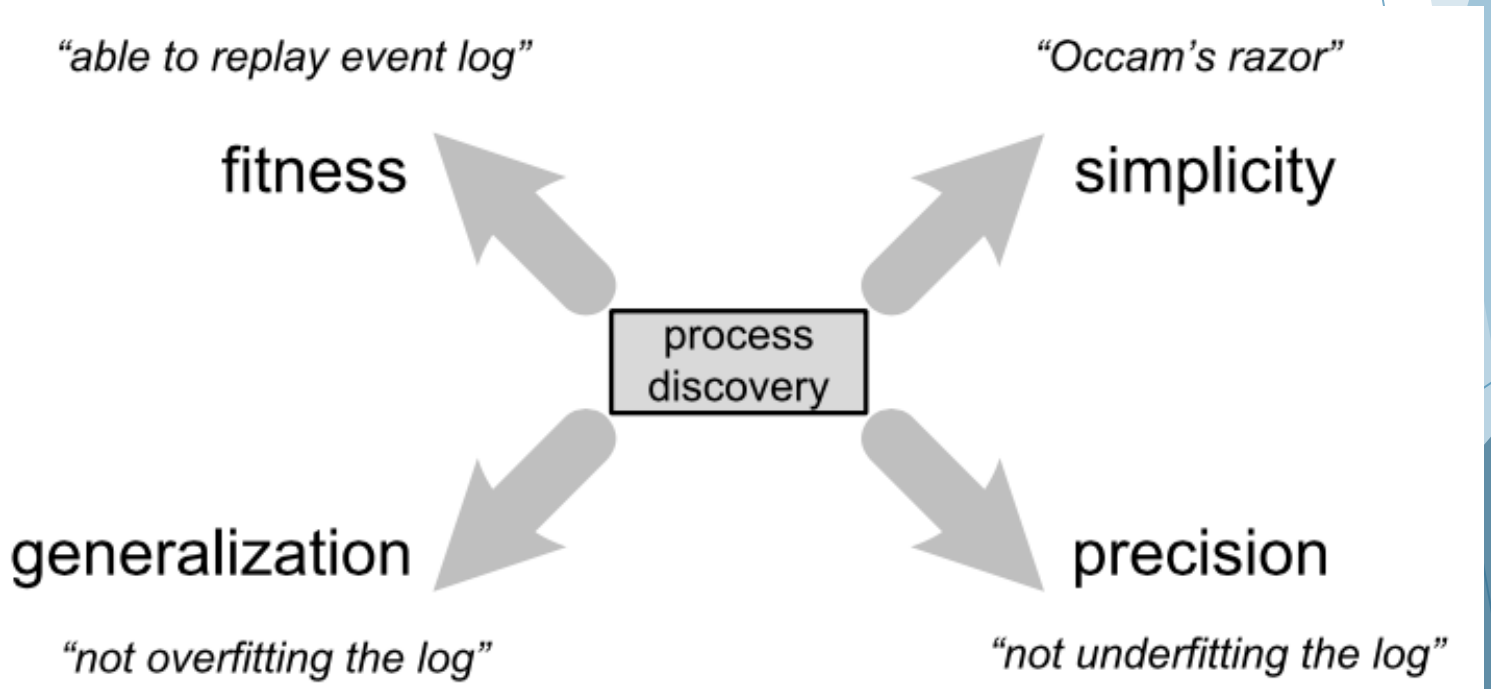
- The α-algorithm uses a local completeness notion based on $>_L$, i.e., if there are two activities a and b, and a can be directly followed by b, then this should be observed at least once in the log.

Incompleteness

- Consider a process consisting of 10 activities that can be executed in parallel and a corresponding log that contains information about 10,000 cases.
- The total number of possible interleaving in the model with 10 concurrent activities is 10! = 3,628,800.
- It is impossible that each interleaving is present in the log as there are fewer cases (10,000) than potential traces (3,628,800).
- Even if there are 3,628,800 cases in the log, it is extremely unlikely that all possible variations are present.
- Therefore, weaker completeness notions are needed.
  - Local completeness can reduce the required number of observations dramatically.
  - For example, for the α-algorithm only $10 \times (10 - 1) = 90$ rather than 3,628,800 different observations are needed to construct the model.

Quality issues related to Process Models

- Four main quality dimensions are defined to refer to quality of the discovered model:
  - *fitness*,
  - *simplicity*,
  - *precision*, and
  - *generalization*



Quality issues related to Process Models

- Fitness
  - Fitness can be defined at case or event level.
- Simplicity
  - the complexity of the model could be defined by the number of nodes and arcs in the underlying graph.
  - Also more sophisticated metrics can be used, e.g., metrics that take the “structuredness” or “entropy” of the model into account.

Quality issues related to Process Models

- Fitness and simplicity alone are not adequate.

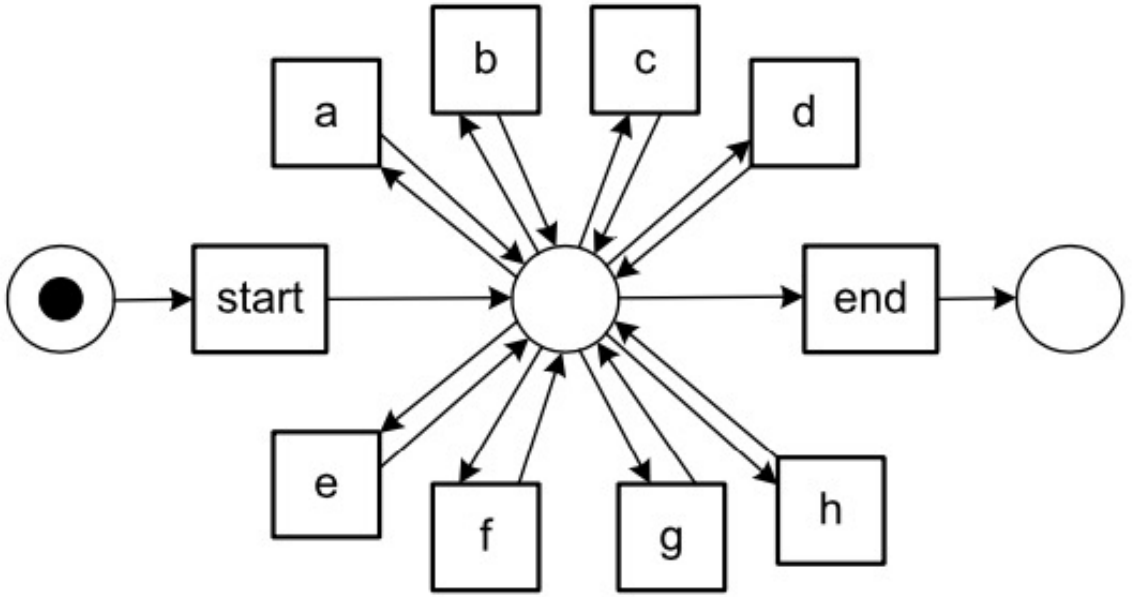


Fig. 6.23 The so-called "flower Petri net" allowing for any log containing activities {a,b,...,h}

Quality issues related to Process Models

- Precision
  - A model is *precise* if it does not allow for "too much" behavior (e.g., flower model).
  - A model that is not precise is "underfitting" (i.e., the model allows for behaviors very different from what was seen in the log).
- Generalize
  - A model should *generalize* and not restrict behavior to the examples seen in the log (e.g., enumerating model).
  - A model that does not generalize is "overfitting".

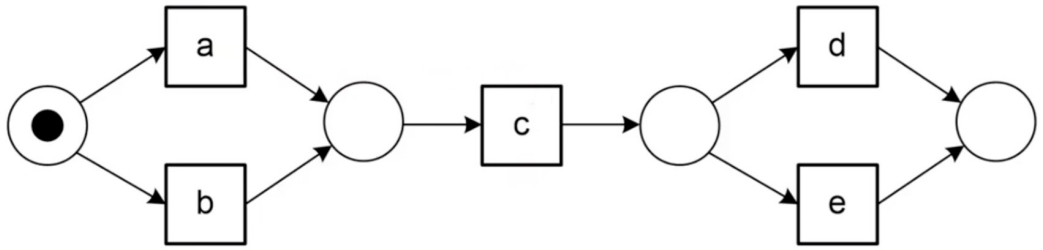
Process mining algorithms need to strike a balance between "overfitting" and "underfitting".

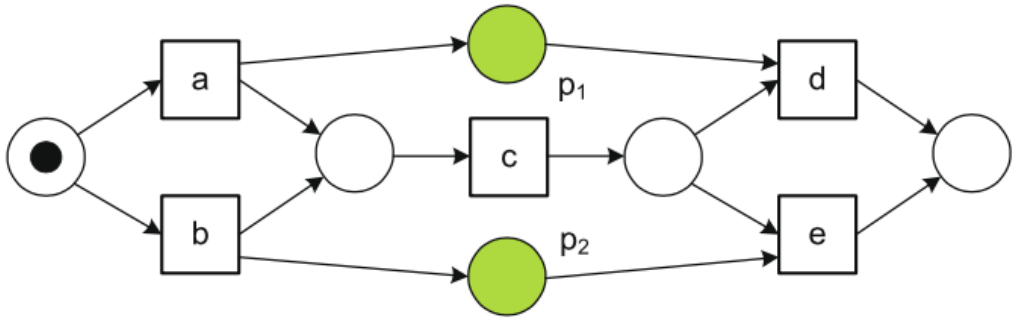
Quality issues related to Process Models

- Sometimes it is difficult to balance between being too general and too specific.

$L_4 = [\langle a, c, d\rangle^{45}, \langle b, c, d\rangle^{42}, \langle a, c, e\rangle^{38}, \langle b, c, e\rangle^{22}]$

$L_9 = [\langle a,c,d\rangle^{45}, \langle b,c,e\rangle^{42}]$

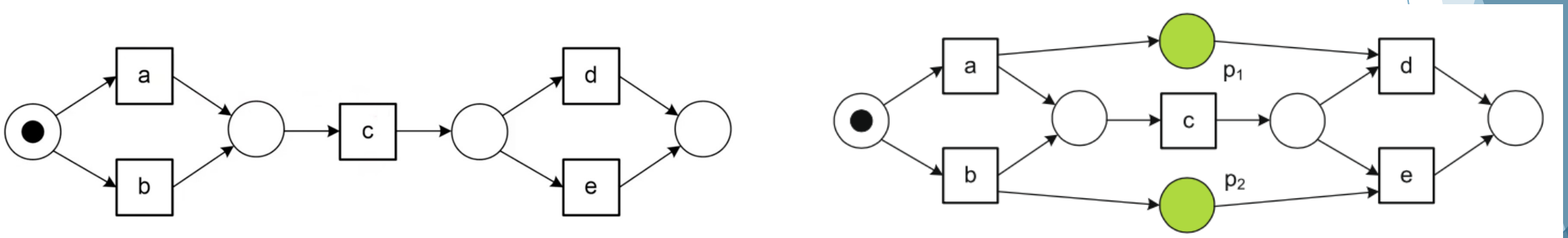




Both WF-nets can produce the log L_9 but L_4 can only be produced by the left WF-net. WF-net on the right is a better choice for L_9.

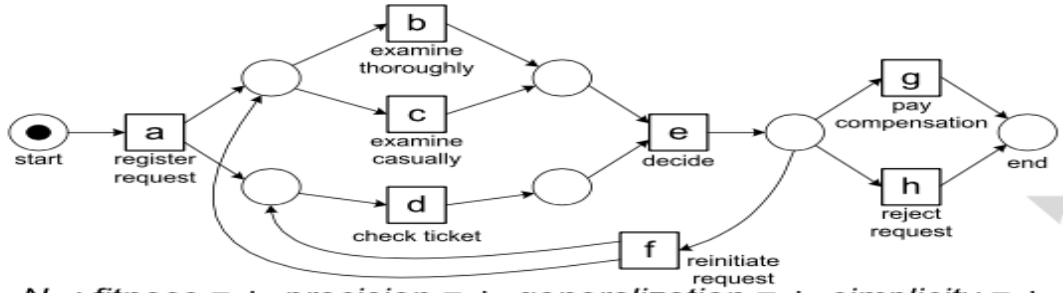
Quality issues related to Process Models

$L_{12} = [\langle a, c, d\rangle^{99}, \langle b, c, d\rangle^{1}, \langle a, c, e\rangle^{2}, \langle b, c, e\rangle^{98}]$

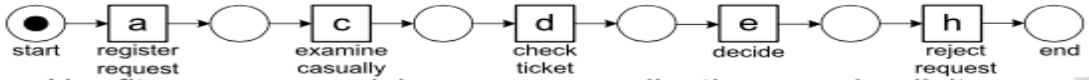


Argument 1: N_4 (on the left) is a better model for L_{12} as all traces can be reproduced
Argument 2: 197 out of 200 traces can be explained by the more precise model N9.

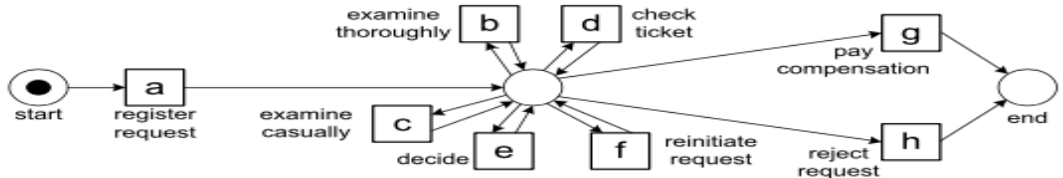
There is a delicate balance between “overfitting” and “underfitting”.
Hence, it is difficult, if not impossible, to select “the best” model.



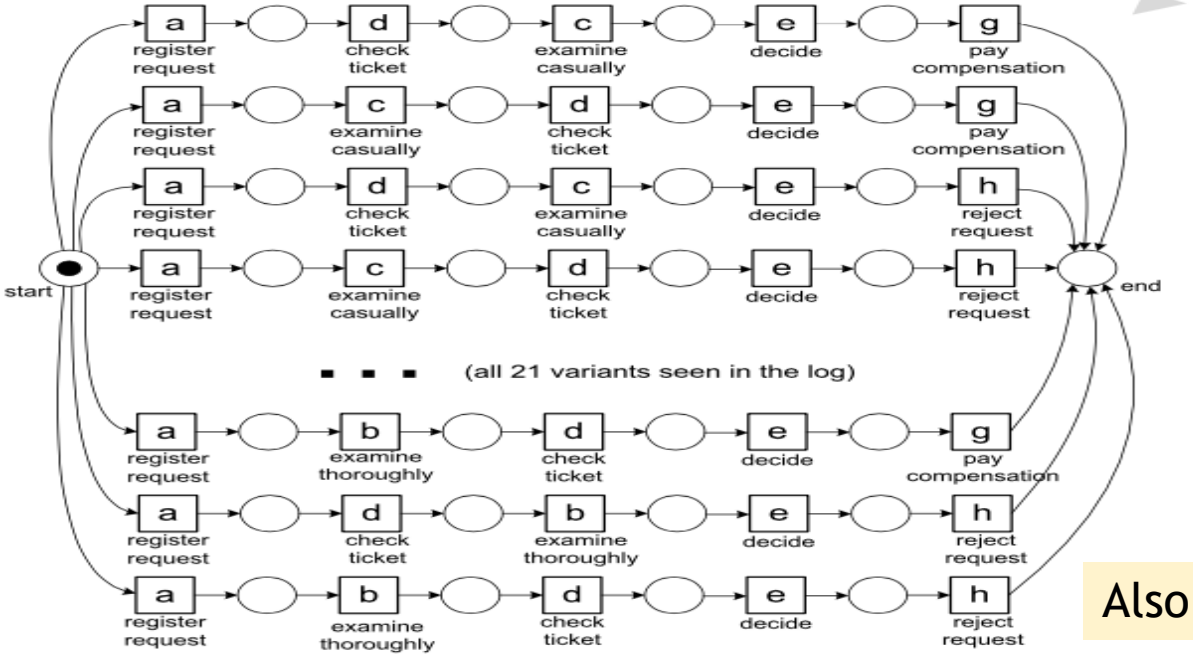
N_1 : *fitness* = +, *precision* = +, *generalization* = +, *simplicity* = +



N_2 : *fitness* = -, *precision* = +, *generalization* = -, *simplicity* = +



N_3 : *fitness* = +, *precision* = -, *generalization* = +, *simplicity* = +

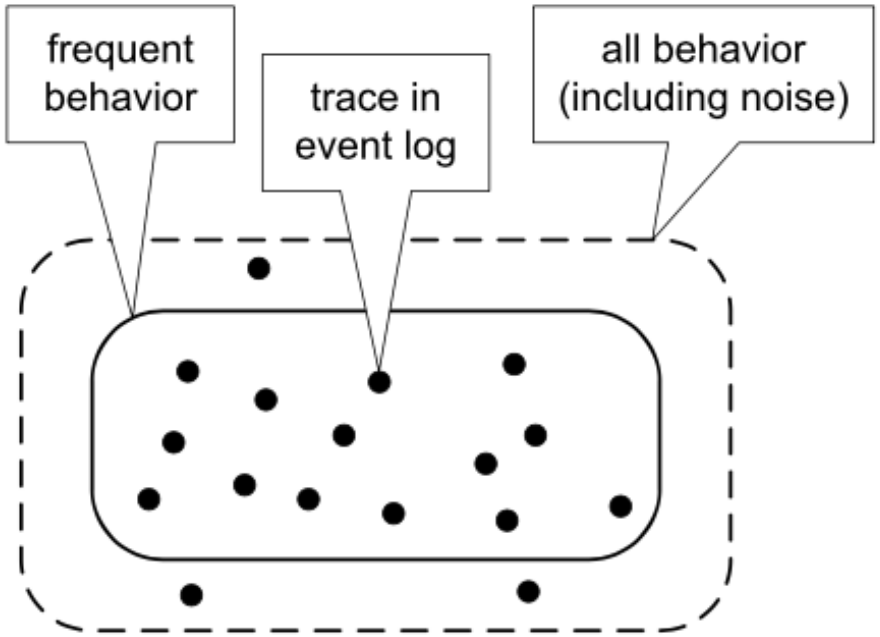


N_4 : *fitness* = +, *precision* = +, *generalization* = -, *simplicity* = -

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

Also known as Enumeration model

Challenges for Process Discovery Algorithms



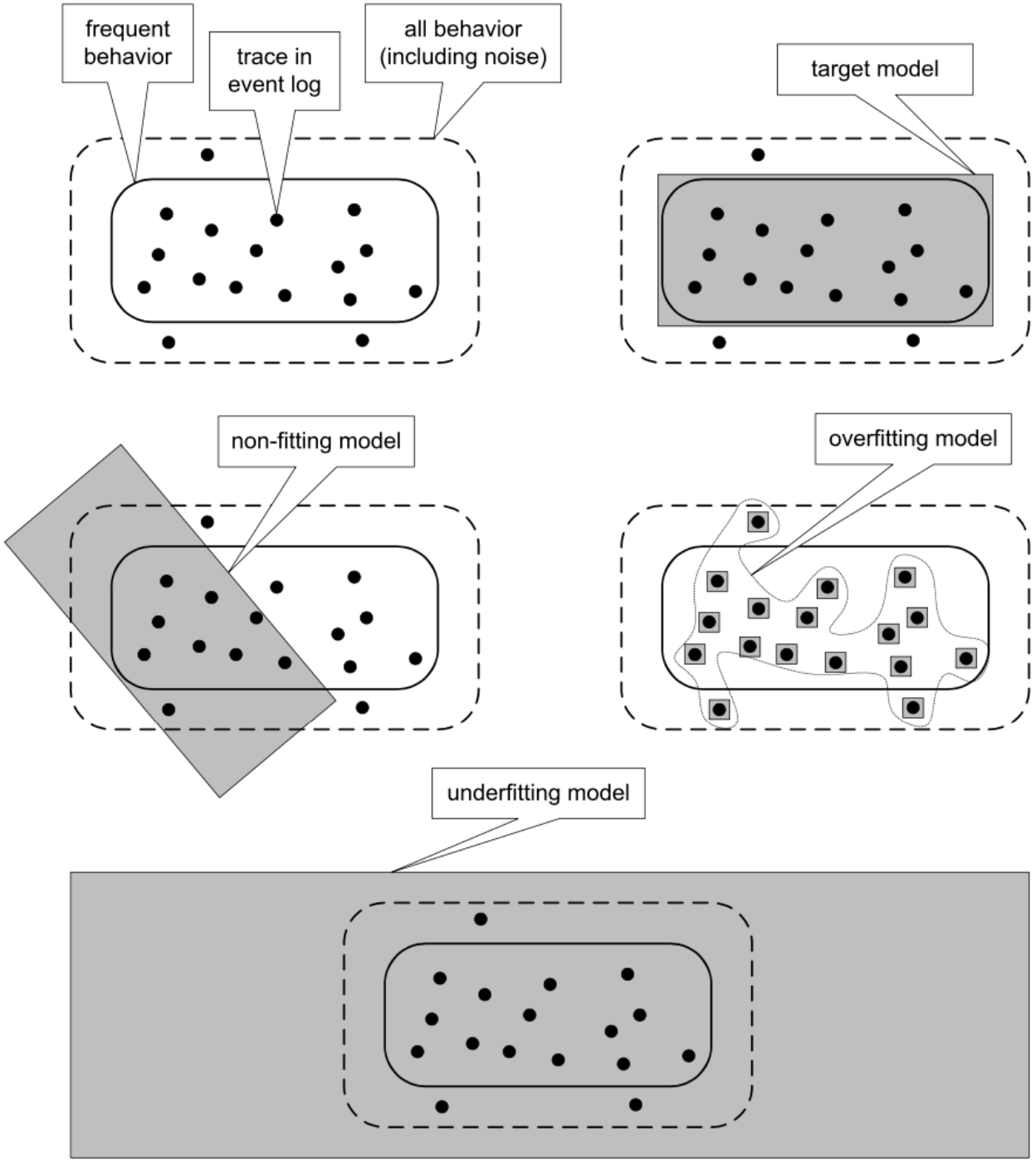


Fig. 7.1 Overview of the challenges that process discovery techniques need to address

Challenges for Process Discovery Algorithms

- **Fitness** can be defined as a measure between 0 (very poor) to 1 (perfect).
- A notion called “structural appropriateness” considers the **simplicity** dimension; the model is analyzed to see whether it is “minimal in structure”.
- Another notion called “behavioral appropriateness” analyzes the balance between **overfitting** and **underfitting**.

Reading Material

- Chapter 6 & 7: Aalst