How Much Event Data Is Enough?

A Statistical Framework for Process Model Discovery

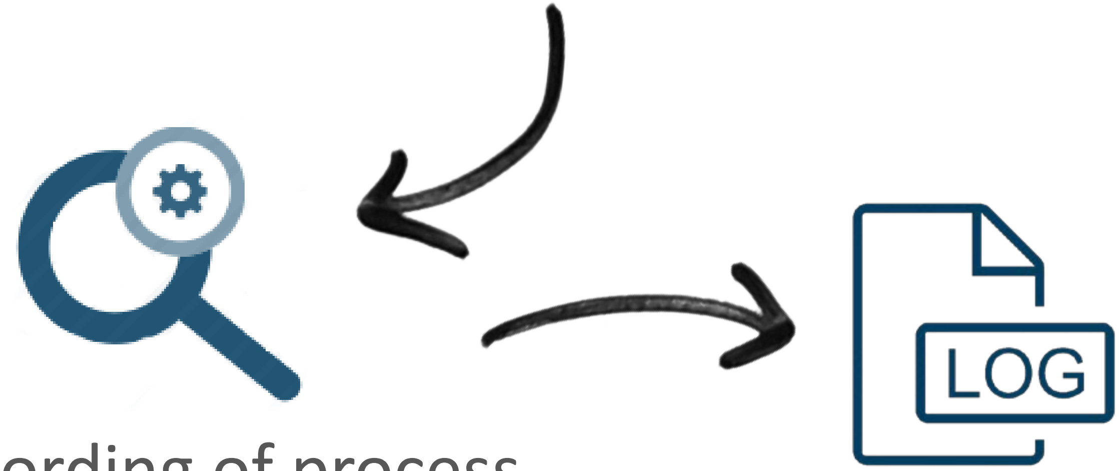
Martin Bauer, Arik Senderovich, Avigdor Gal,
Lars Grunske, Matthias Weidlich

HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

Technion
Israel Institute of
Technology

Processes and Events

Automated control of business processes

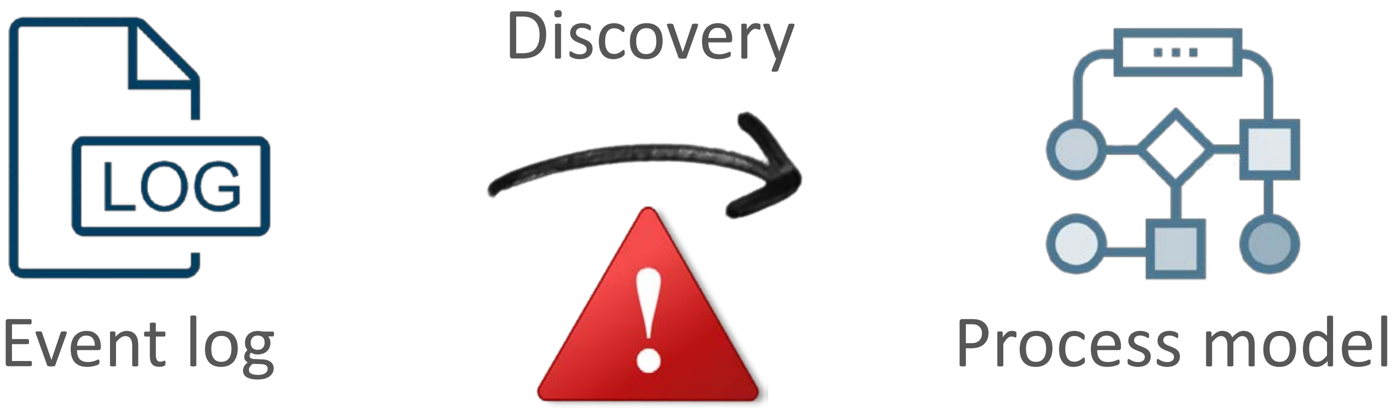


Recording of process execution information

Event logs:

- Timestamps
- Case IDs
- Activity IDs
- …

The Question of Process Discovery



Event log

Discovery

Process model

Efficiency of process discovery becomes increasingly important

- Pervasiveness of data sensing/logging
  => Large-scale event logs
- Tuning a large range of parameters of discovery algorithms
  => Repeated, exploratory analysis
- Software-as-a-Service solutions for process discovery
  => Online handling of event logs

A View on Related Work

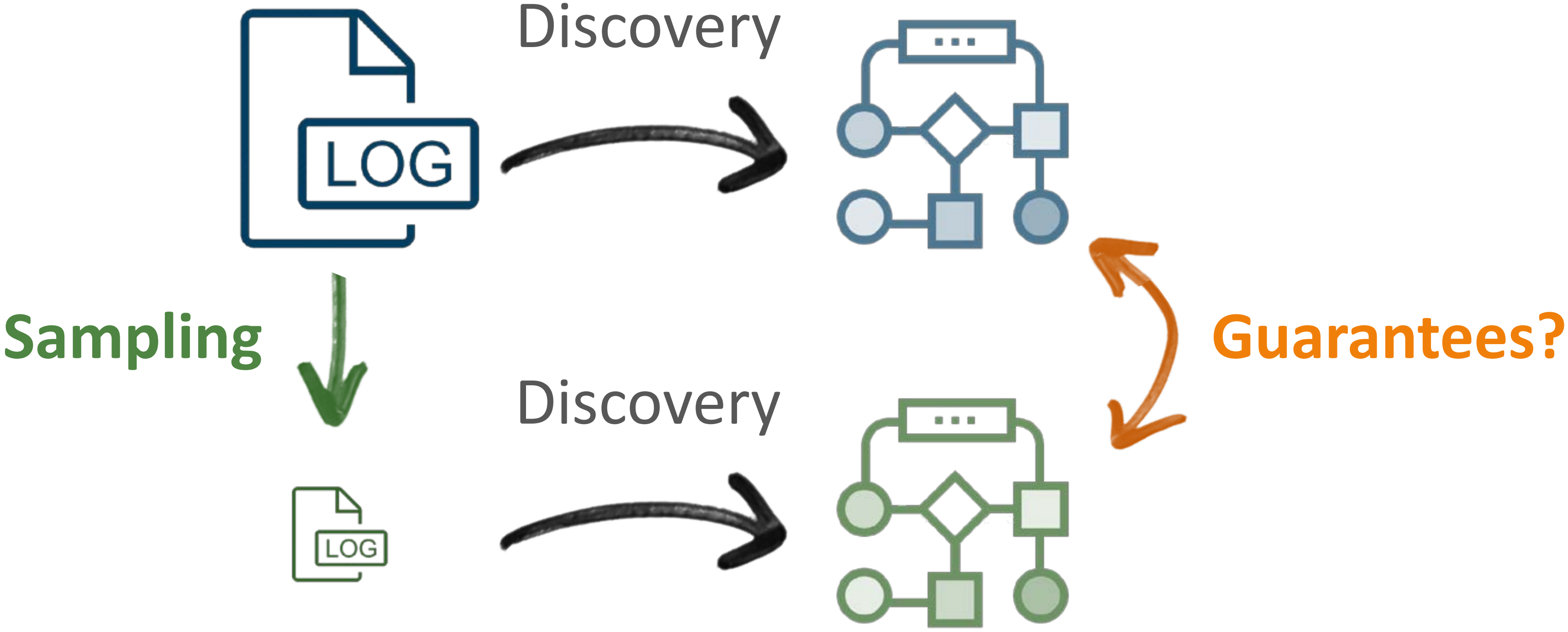
Plethora of discovery algorithms [Augusto et al. 2017]

Striving for scalability

- By divide-and-conquer: Decompose the discovery problem [van der Aalst & Verbeek 2015]
- By parallelization and distribution [Wang et al. 2015, Evermann 2016]

Recently: Idea of sampling event data [Busany & Maoz 2016]

Daring the Gap



Discovery

Sampling

Guarantees?

Discovery

How to determine how much of an event log to use to discover a process model?

Agenda

Background and Related Work on Process Discovery

A Statistical Framework for Process Discovery

- Log sampling
- Framework definition

Instantiating the Framework

- For control-flow discovery
- For performance discovery

Experimental Results

Log Sampling

TODO: Discovery sufficiency etc

Minimum Sample Size

TODO: how to determine N
Include example

Framework

TODO: main steps of the algorithm

Agenda

Background and Related Work on Process Discovery

A Statistical Framework for Process Discovery

- Log sampling
- Framework definition

Instantiating the Framework

- For control-flow discovery
- For performance discovery

Experimental Results

Control-Flow Perspective

A notion of “new control-flow information”

- New activity
- New directly-follows dependency
- New initial or final activity

TODO: EXAMPLE

What about frequencies?

- Determine on sample (no guarantee on δ-similarity)
- Changes in relative frequencies are “new information”

Performance Perspective

Focus on cycle time of a process, a fine-grained numerical value

A notion of “new cycle-time information”

- Cycle time is more than ε-different
- Measuring granularity:
  - Per complete process
  - Per individual activities

TODO: EXAMPLE

Agenda

Background and Related Work on Process Discovery

A Statistical Framework for Process Discovery

- Log sampling
- Framework definition

Instantiating the Framework

- For control-flow discovery
- For performance discovery

Experimental Results

Setup

Datasets

- BPI Challenge 2012
  - Loan/overdraft applications
  - >262k events of >13k traces
- BPI Challenge 2014
  - Incident management at Rabobank Group ICT,
  - >343k events of >46 traces

Discovery algorithm

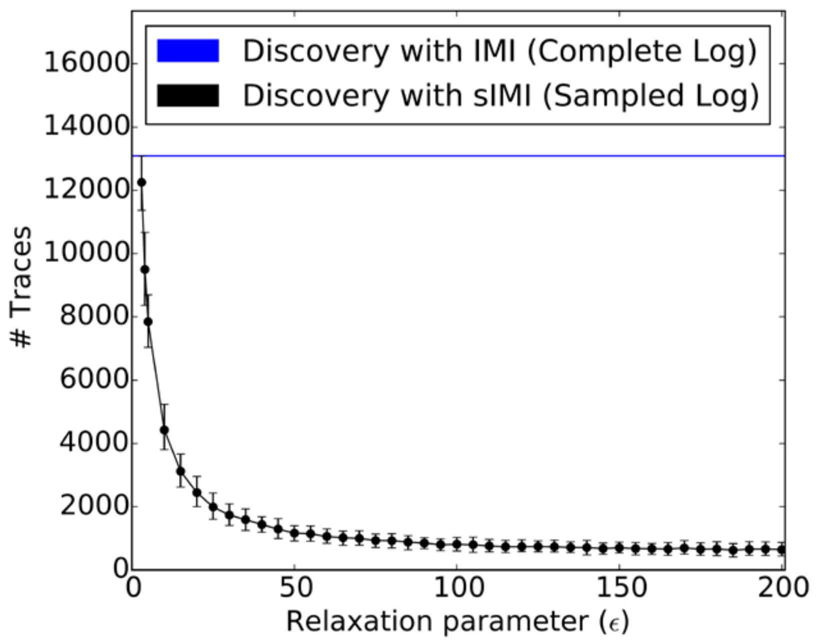
- Inductive Miner Infrequent [Leemans et al. 2013]
- Noise threshold set to 20%

Approach implemented as a ProM plugin (@Github)

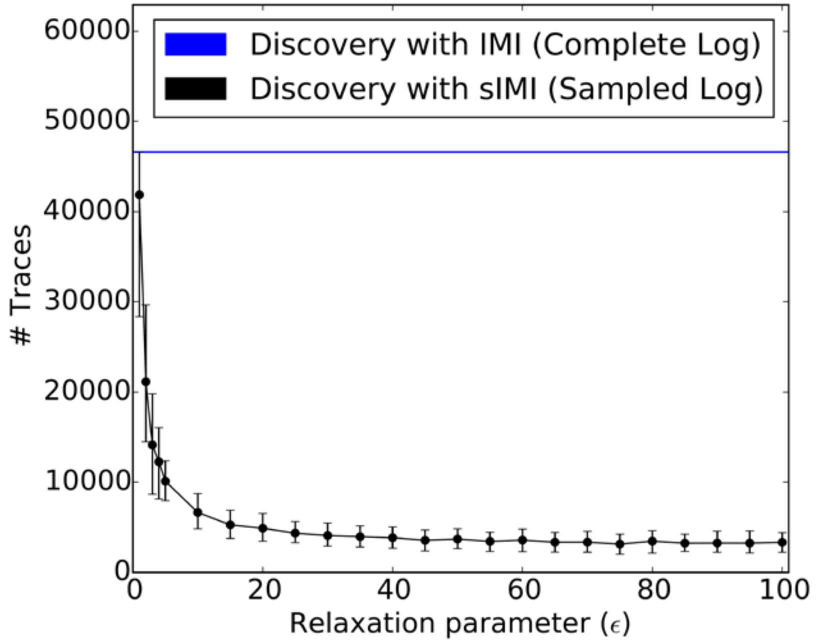
Measures

- Pre-processing effectiveness: #traces sampled
- Actual efficiency: runtime, memory footprint
- Discovery effectiveness: fitness, approximated cycle time

Pre-Processing Effectiveness



BPI-2012

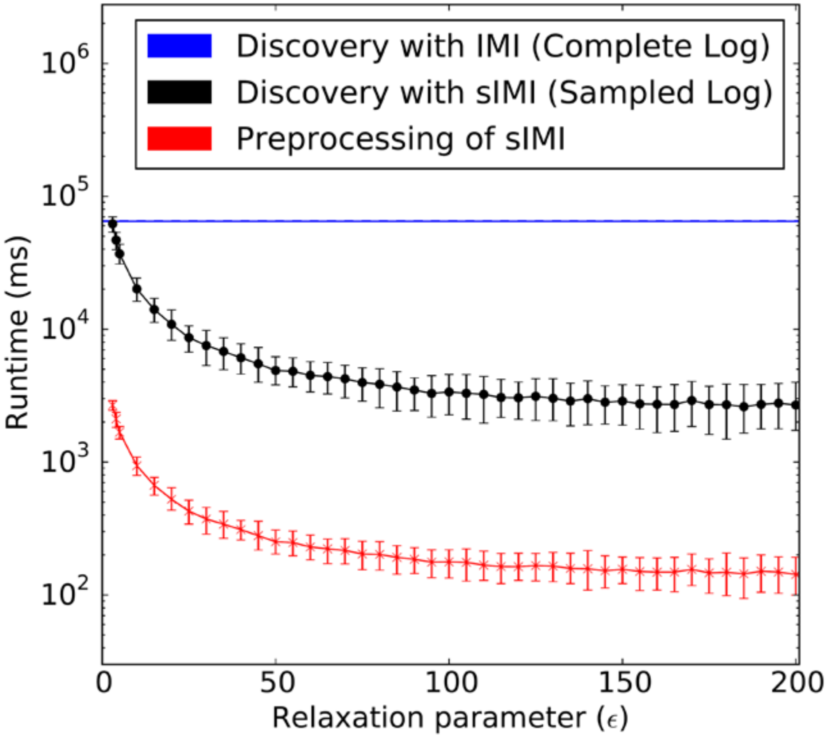


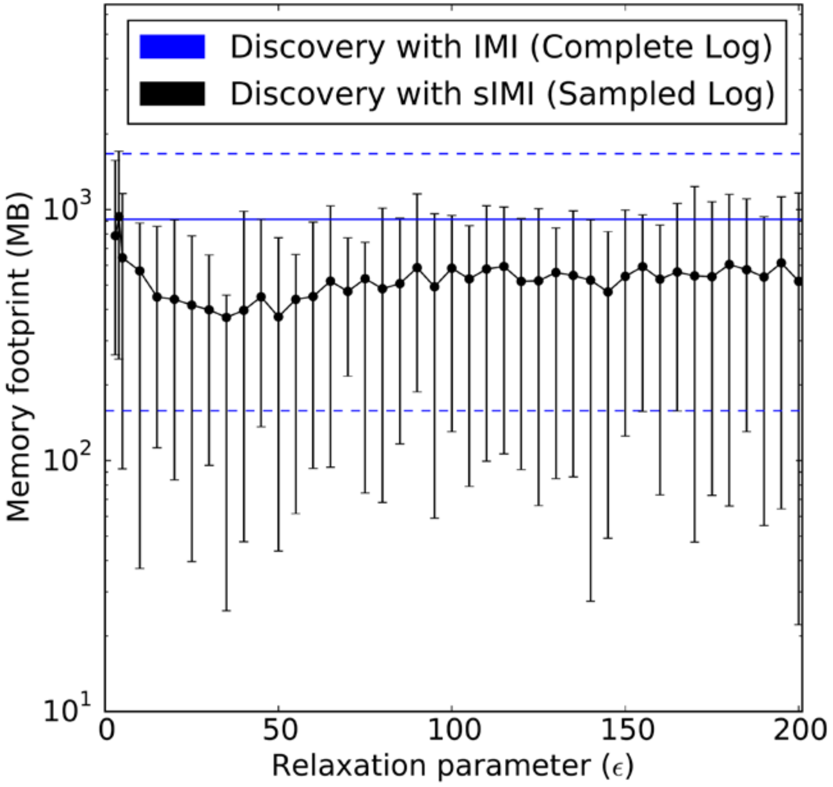
BPI-2014

Drastic reduction of number of traces considered for discovery

Trend is consistent for different datasets

Runtime and Memory Footprint

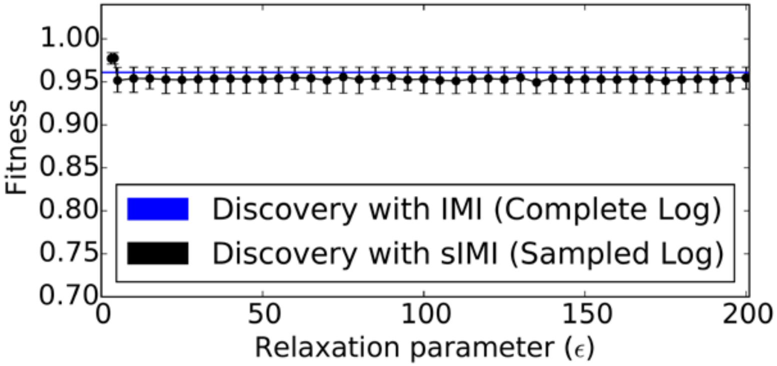


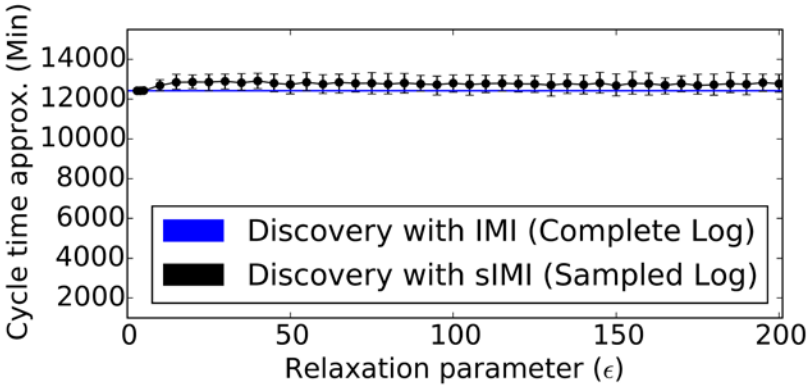


Pre-processing is efficient

Significant reduction of overall resource utilisation

Discovery Effectiveness





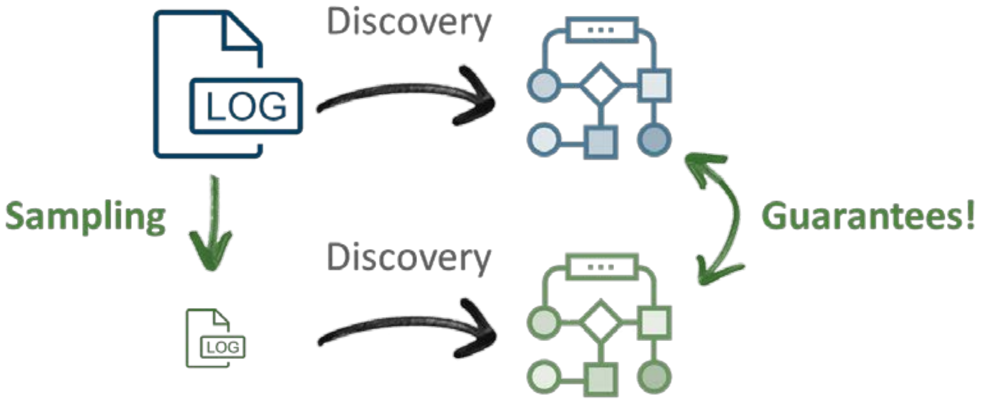
Negligible degradation of discovery quality

- For control-flow fitness
- For the cycle time approximation

Conclusions

Framework for statistical process discovery

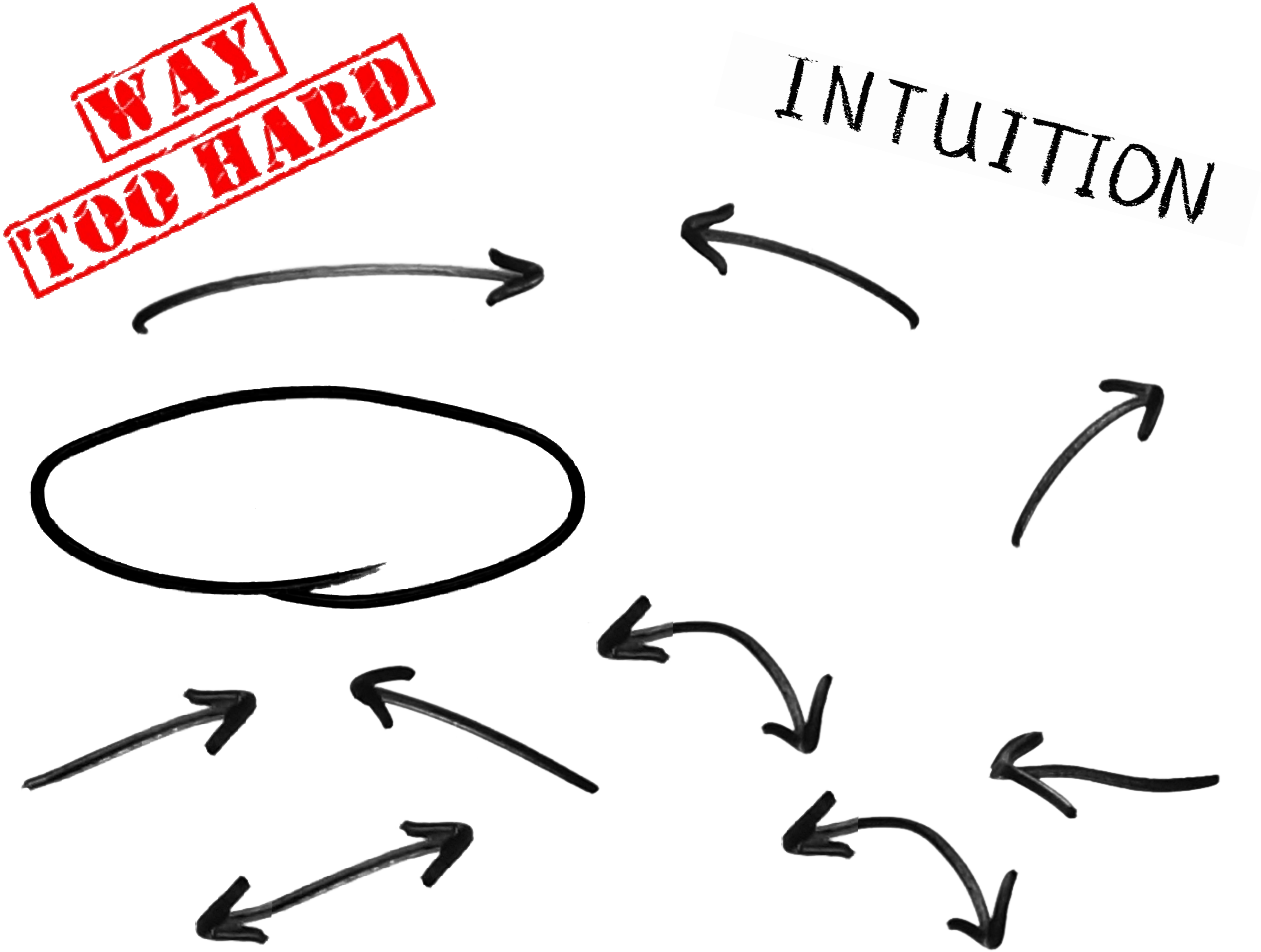
- Sample an event log
- Guarantees on the introduced error



Instantiation for control-flow and performance aspects

Next: Additional model perspectives

Thank you!



WAY TOO HARD

INTUITION