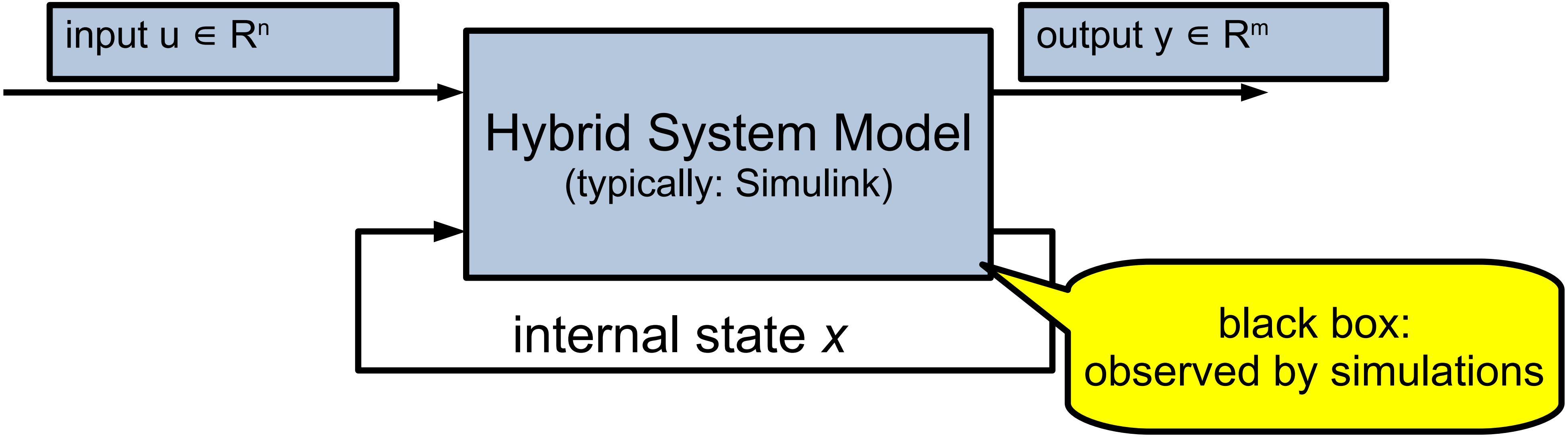
ARCH Competition 2019
Falsification Category

participants | organization | benchmarks | outcome

Paolo Arcaini, Alexandre Donze, Gidon Ernst,
Georgios Fainekos, Logan Mathesen,Giulia Pedrielli,
Shakiba Yaghoubi, Yoriyuki Yamagata, Zhenya Zhang

ASU, USA | Decyphir, France | LMU, Germany | AIST&NII, Japan
contact: gidon.ernst@sosy.ifi.lmu.de

Falsification



Goal:

- find an input u
- such that the output y
- violates a given specification in temporal logic (STL/MTL)

Participating Tools

Breach	(Alexandre Donze)
S-TaLiRo	(Shakiba Yaghoubi, Logan Mathesen, Georgios Fainekos)
falsify	(Yoriyuki Yamagata, Shuang Liu)
FalStar	(Gidon Ernst, Zhenya Zhang, Paolo Arcaini)

Organization

- 2017: 1 tool, 1 benchmark
- 2018: 2 tools, same 1 benchmark
- **2019**: 4 tools, 6 models, 24 requirements
  - two sets of results
    - arbitrary inputs → can achieve best results
    - fixed constrained inputs → better for direct comparison
  - Goal: validate all results (not really achieved)

Benchmarks

- Source
  - standard from the literature (e.g. automatic transmission)
  - new ones provided by participants
- Important
  - test cases: how to initialize and run the models
  - precise (informal) input and requirement specifications

Evaluation

- Setup
  - max number of simulations per trial: 300
  - stochastic algorithms, hence multiple trials: 50

    → running all benchmarks takes several days
- Metrics:
  - falsification rate
  - average/median required simulation (over successful trials)

Highlights

- Breach/FalStar: good success with extreme values and random sampling → benchmarks too easy
- S-TaLiRo: only tool to falsify steam condenser benchmark (by combination of techniques)
- falsify: many counterexamples from a **single simulation** **(**online, grey box: learns system dynamics from trace prefix)

→ Different approaches have different strengths

Conclusion & Outlook

- need harder benchmarks
- need a maintained benchmark repository
  (talk to Gidon if interested)
- need a standardized result format for validation
  (fairly straight forward, but ran out of time)
- next steps: in-depth analysis of results (also no time)

→ **hard but rewarding work for all participants**

→ **made lots of progress this year :-)**