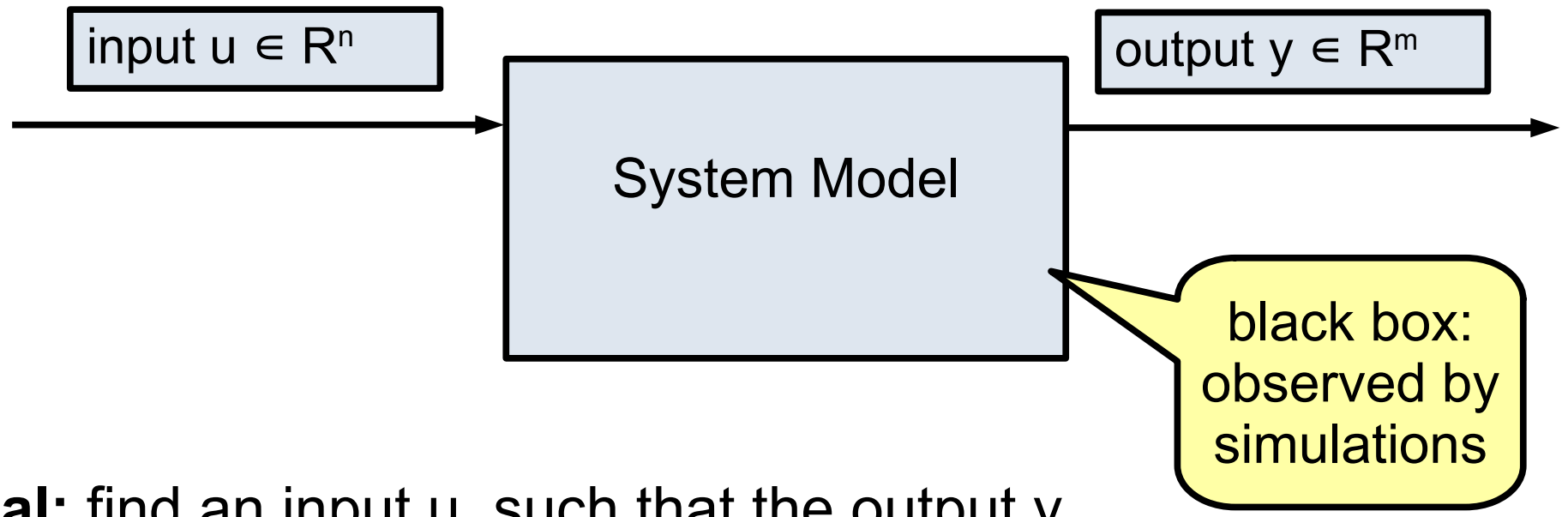
ARCH-COMP 2021: Falsification

Gidon Ernst <gidon.ernst@lmu.de>, Paolo Arcaini, Ismail Bennani, Aniruddh Chandratre, Alexandre Donze,Georgios Fainekos,Goran Frehse, Khouloud Gaaloul, Jun Inoue, Tanmay Khandait, Logan Mathesen, Claudio Menghi, Giulia Pedrielli, Marc Pouzet, Masaki Waga, Shakiba Yaghoubi, Yoriyuki Yamagata, and Zhenya Zhang

AIST/NII/Kyushu/Kyoto, Japan | ASU, USA | ENS/ENSTA/Decyphir Fance | LMU, Germany | University of Luxembourg

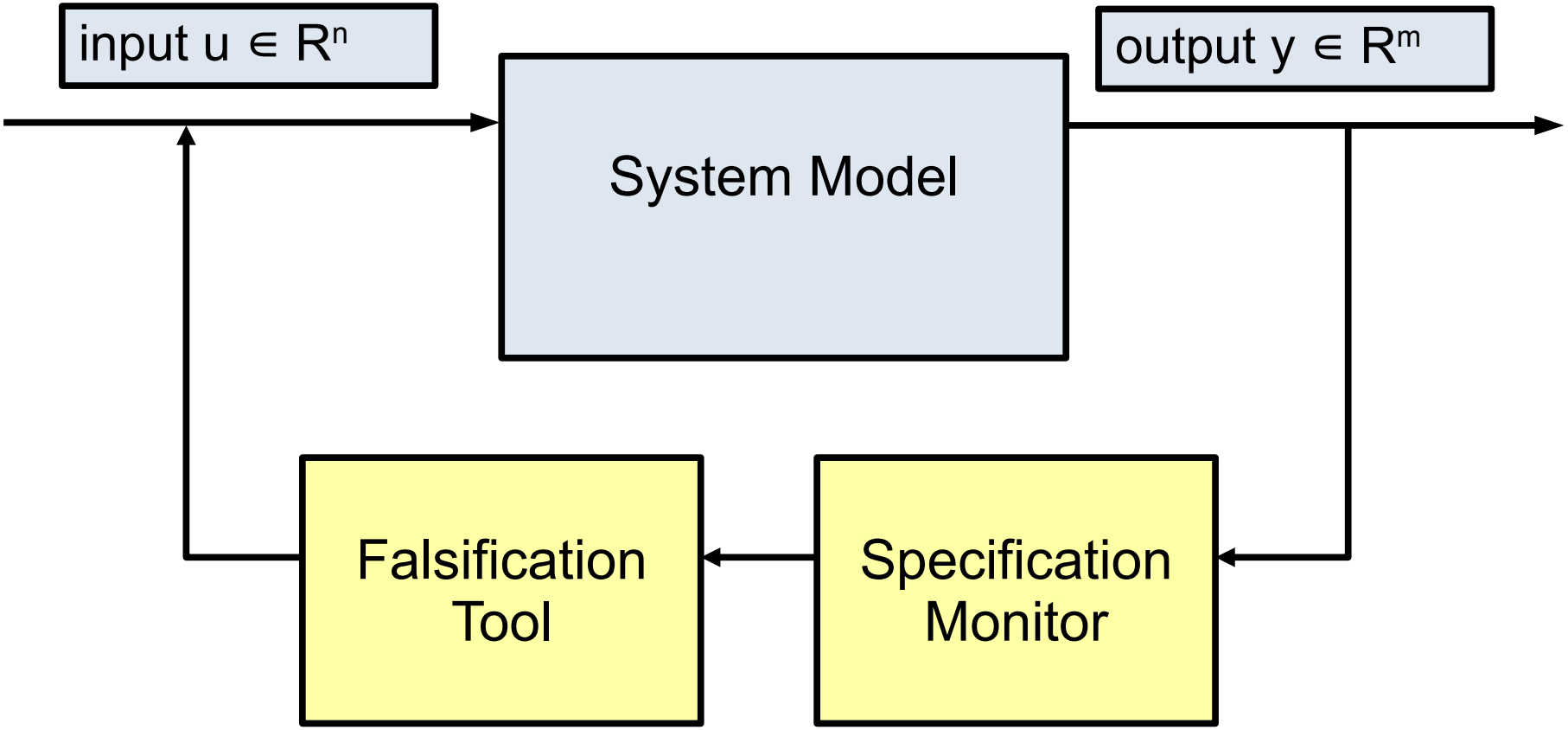
Falsification



Goal: find an input u, such that the output y violates a given specification in temporal logic

how to crash a car into a wall

Falsification



Δ 2021

1) two new tools, some new people (welcome!)
2) new conjunctive requirements (thanks, Logan!)
3) effort towards validating results (thanks everyone!)

Participants

• FalCAuN	(M. Waga)	discrete abstraction + model cecking
• ForSee	(Z. Zhang, P. Arcaini, I. Hasuo)	MCTS, robustness, optimization
• ARIsTEO	(K. Gaaloul, C. Menghi)	surrogate model, refinement
• Breach	(A. Donze)	robustness + stoch. optimization
• falsify	(J. Inoue, Y. Yamagata)	reinforcement learning
• FalStar	(G. Ernst, S. Sedwards)	Las-Vegas tree search
• S-TaLiRo	(S. Yaghoubi, L. Mathesen, A. Chandratre, T. Khandait, G. Pedrielli, G. Fainekos)	robustness + stoch. optimization
• zlscheck	(I. Bennani, M. Pouzet, G. Frehse)	whitebox gradient descent

Requirements

- Example (automatic transmission):
  - within 20s the speed remains below 120
  - formally: $\Box_{[0,20]}\ (v < 120)$
- Example (magnetic levitation controller):
  - the position stabilizes in a certain way
  - $\Diamond_{[0.0,\ 1.0]}\ (Pos > 3.2)\ \wedge$

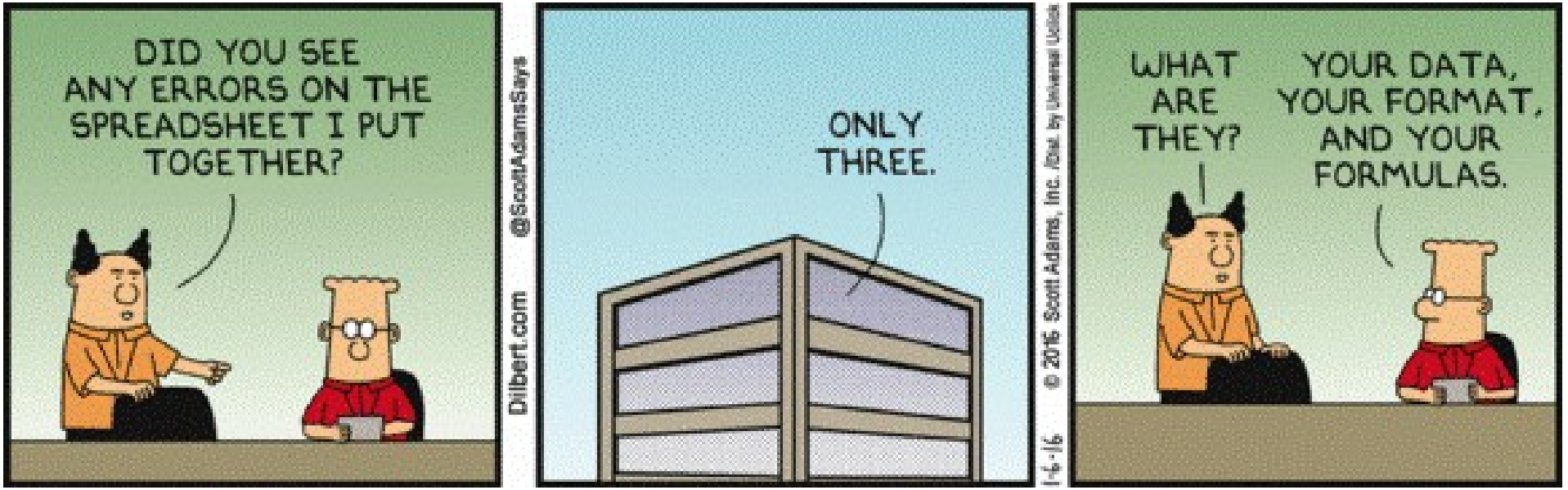
    $\Diamond_{[1.0,\ 1.5]}\ (\Box_{[0,\ 0.5]}\ 1.75 < Pos < 2.25)\ \wedge$

    $\Box_{[2.0,\ 3.0]}\ (1.825 < Pos < 2.175)$

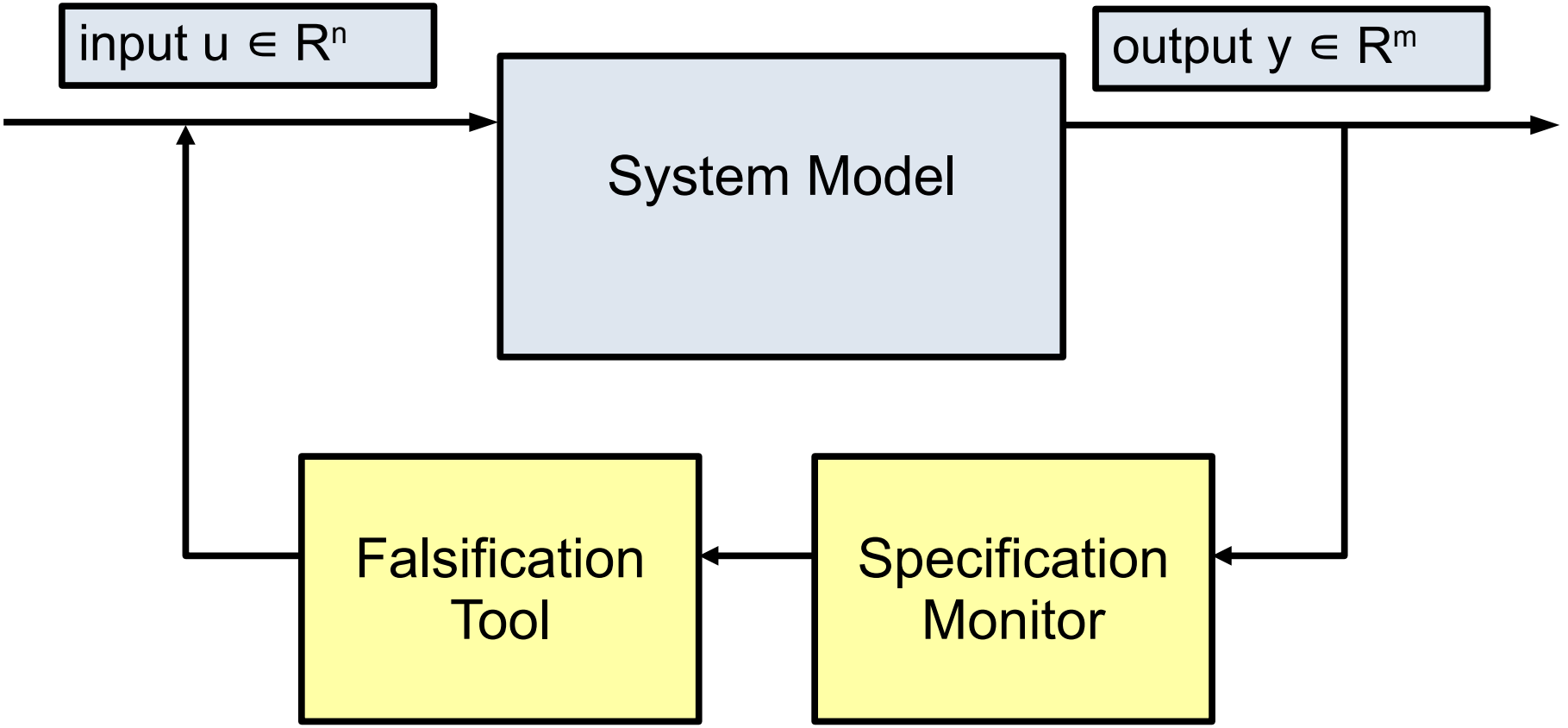
check multiple requirements at once

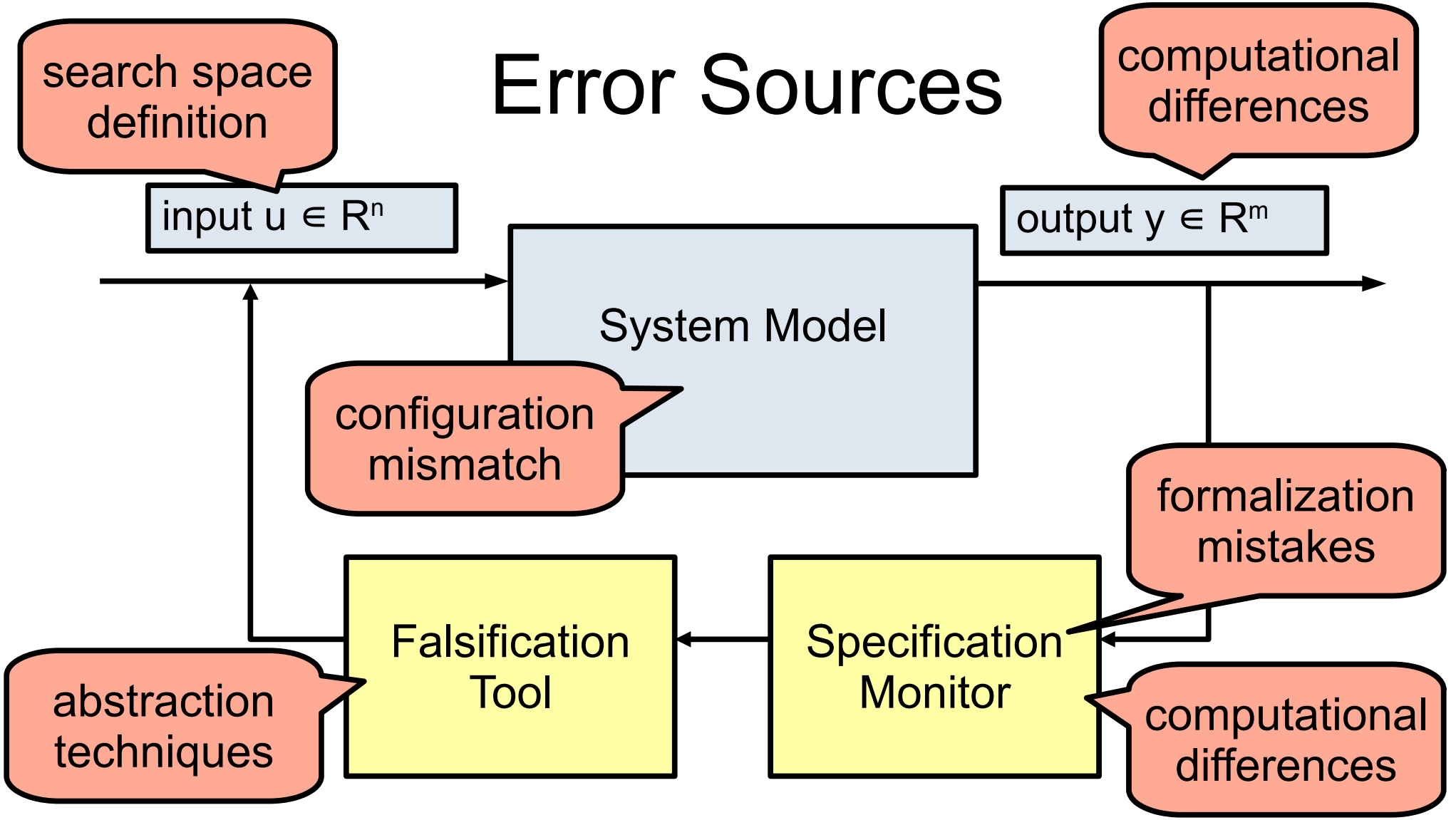
but possibly conflicting search guidance

Validation


DID YOU SEE ANY ERRORS ON THE SPREADSHEET I PUT TOGETHER?
Dilbert.com @ScottAdamsSays
ONLY THREE.
1-6-16 © 2016 Scott Adams, Inc. /Dist. by Universal Uclick
WHAT ARE THEY?
YOUR DATA, YOUR FORMAT, AND YOUR FORMULAS.

Error Sources




Error Sources
search space definition
computational differences
input u ∈ Rⁿ
output y ∈ Rᵐ
System Model
configuration mismatch
formalization mistakes
Falsification Tool
Specification Monitor
abstraction techniques
computational differences

2021/FALS/Validation.md

formalization

search space

reported verdict

simulation & computation

```
property: AT1
formula: □_[0, 20] (speed < 120)
input is within bounds
using stop time as provided: 20.0
falsified is correct
expected robustness -0.014
computed robustness -0.013
robustness error 0.001
[...]
```

Validity via Robustness Score

by how much did we miss the wall?

reported (by participants)	computed (by validator)	confirmed falsification?
r < 0	r < 0	
r < 0	r < threshold	
r >= 0	_	

Erwin Wurm: Truck, 2015

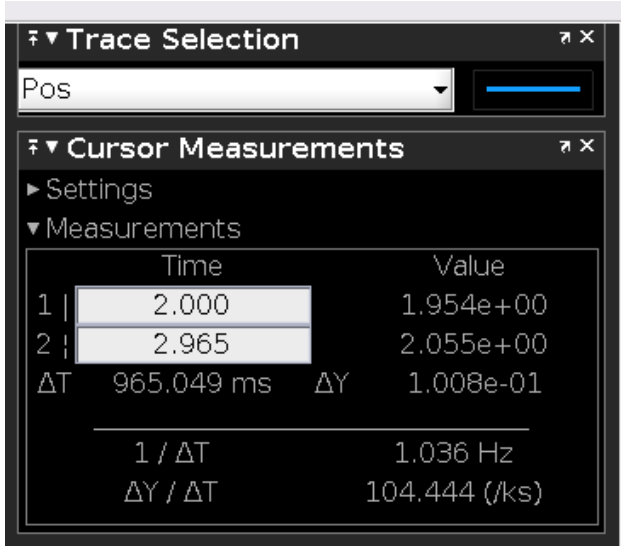
Findings

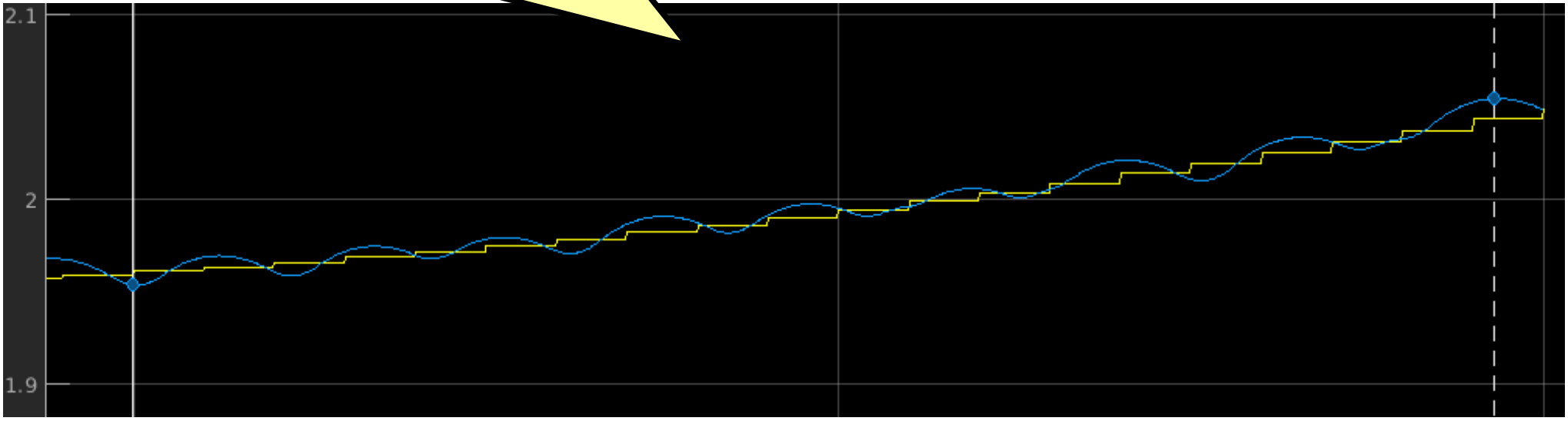
- Discrepancies discovered everywhere
  - documentation somewhat lacking/imprecise/inconsistent
  - some mistakes: formalization, input ranges
  - signal sampling and reporting (granularity, interpolation)

- Summary
  - majority of new results confirmed | some old results wrong
  - obvious & large discrepancies now fixed
  - small discrepancies: computational mismatch or actual errors?

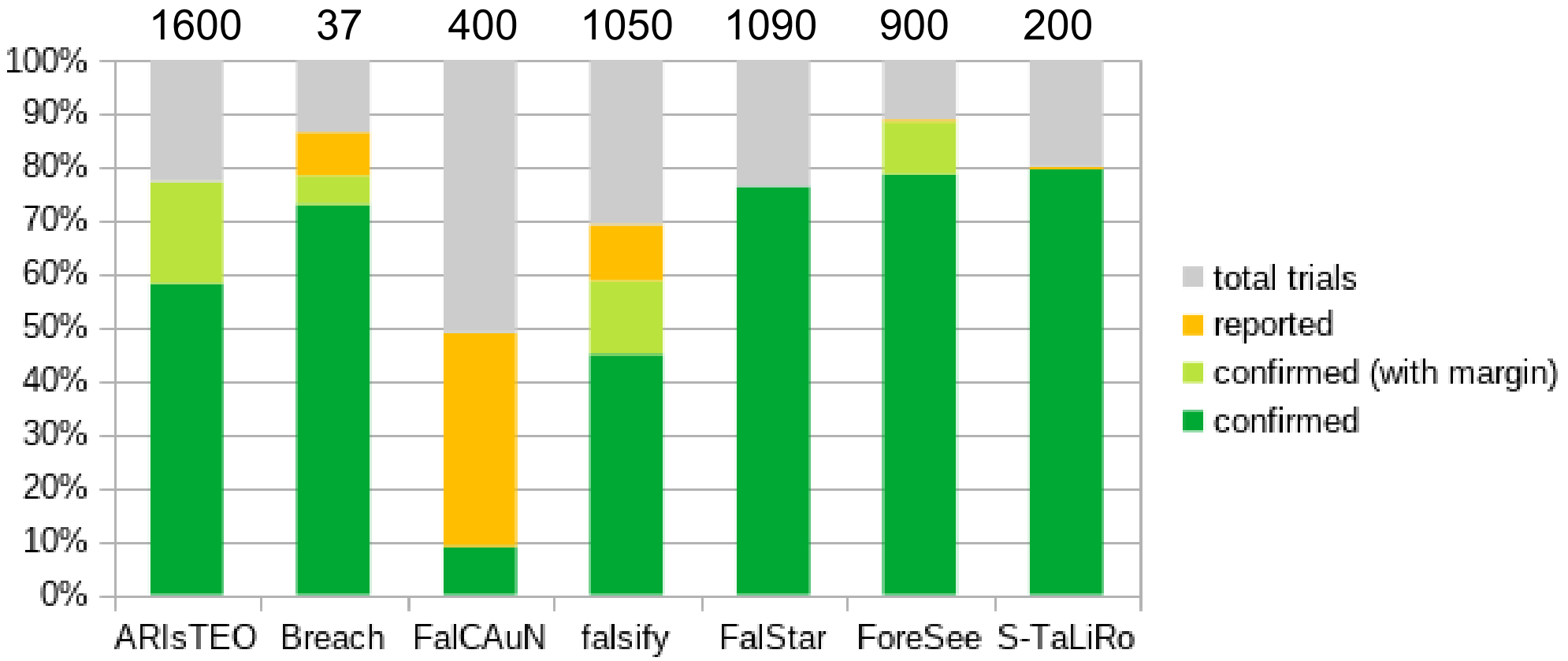
Troubleshooting

check manually:
does this signal falsify?

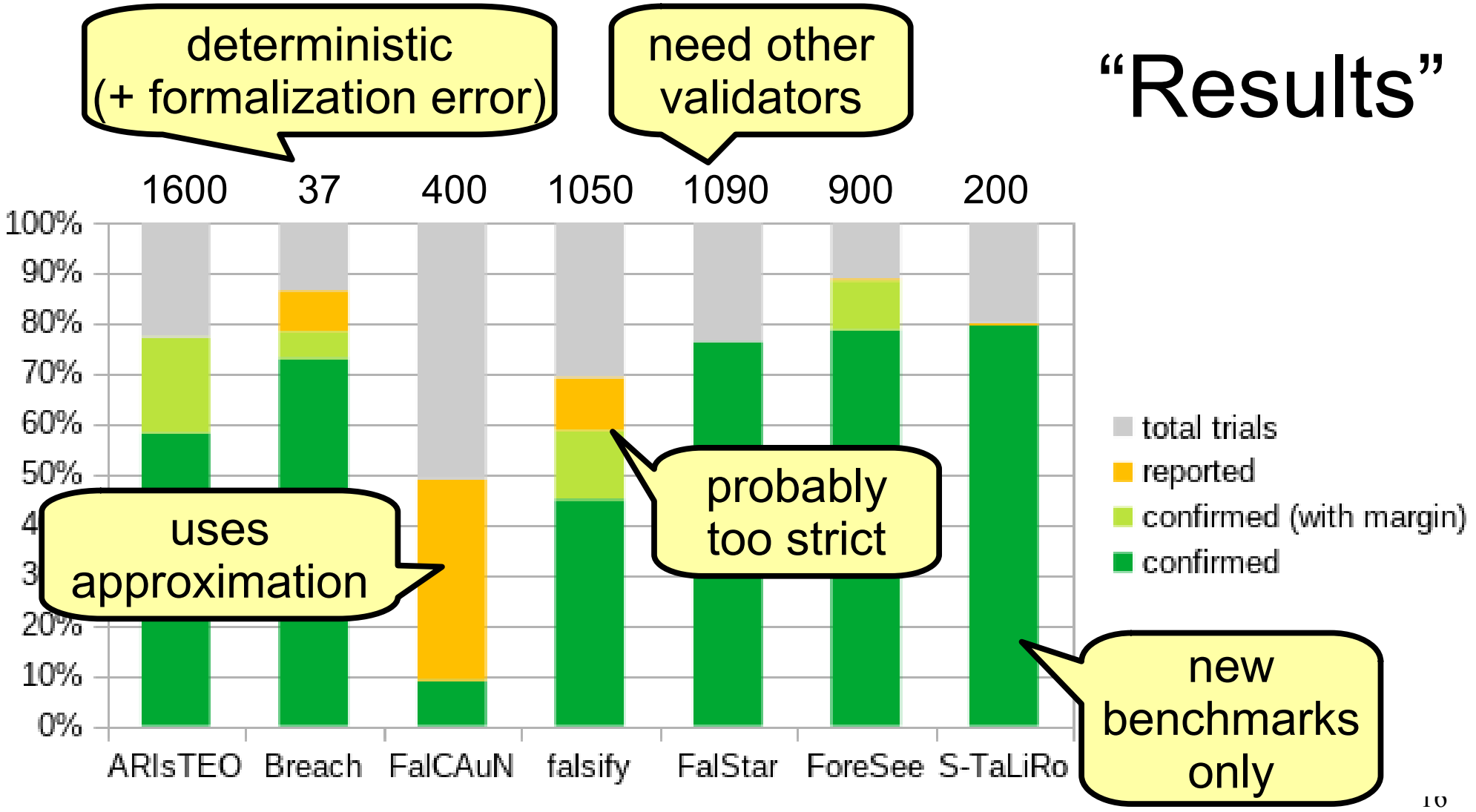




“Results”


1600
37
400
1050
1090
900
200
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
ARIsTEO
Breach
FalCAuN
falsify
FalStar
ForeSee
S-TaLiRo
total trials
reported
confirmed (with margin)
confirmed

“Results”


deterministic
(+ formalization error)
need other
validators
1600
37
400
1050
1090
900
200
100%
90%
80%
70%
60%
50%
20%
10%
0%
uses
approximation
probably
too strict
new
benchmarks
only
total trials
reported
confirmed (with margin)
confirmed
ARIsTEO
Breach
FalCAuN
falsify
FalStar
ForeSee
S-TaLiRo

Conclusion & Outlook

- collection of results still ongoing

- both needed
  - Validation: reported results are plausible
  - Repeatability: performance is plausible

- next year
  - start much earlier, converge with reporting format
  - repeatability (on CodeOcean)
  - non-Matlab models (Zélus, Python), harder benchmarks