

#### Automating Machine Learning Model Checking

Victoria Stodden Adhithya Bhaskar Department of Industrial and Systems Engineering

REAL@USC-META CENTER First Anniversary Conference University of Southern California Monday, October 31, 2022





- 1. ML Model Checking Value Proposition
- 2. Leveraging Steps in the Community
- 3. ReproScreener: A Tool for ML Model Checking (work in progress)



#### Model Checking Value Proposition: Potential Gains

Scalability to large-scale complex models

- e.g. integration of multiple large data sources, deployment on large scale / high throughput computing systems.
- Verification of model performance
  - Consistency of results / predictions across time, systems, data.

Transparency / interpretability

Efficiency in resource use / Discovery speedup

 Computing systems: compute time, appropriate benchmarking; Engineers: code re-use, reduction in effort duplication



#### Model Checking Value Proposition: Potential Drawbacks

Increased overhead

• Additional computational step(s) in model building, deployment

Culture change

• Increased emphasis on reproducibility, verification, correctness in ML models

Hewing to the wrong goals



## **Community Efforts**

#### ML model publication standards

- Gunderson (<u>AAAI 2018</u>)
- Pineau (JMLR 2020)

#### Formal Verification for ML models

- Abate (<u>MEMOCODE 2017</u>)
- Urban and Miné (arxiv 2021)

#### **Research Publication Standards**

- Willis and Stodden (HDSR 2020)
- ML Commons (<u>Github</u>)
- National Academies Reproducibility Report (<u>NASEM 2019</u>)



. . .

#### **ML Model Checking: A Novel Approach**

- Previous work in AI involves applying formal techniques using SMT (satisfiability modulo theory) solving, constraint solving, or abstract numerical interpretation.
- We exploit specialized features of ML pipelines and propose a *reproducibility* approach (<u>NASEM 2019</u>):
  - Exposure of methods
  - Well-defined guarantees in correctness of results



#### **ML Model Checking: ReproScreener**

- Automate ML model checking *at the point of publication*, to provide guarantees on correctness, scalability, and transparency.
- ReproScreener software tool verifies criteria and provides feedback.



### ReproScreener Development (work in progress)

- Create testbed of ML/AI publications (arχiv ML.stat and CS.GL)
- 2. Implement existing "Gunderson" ML model Criteria (Gunderson 2018)
- 3. Label testbed publications manually for Criteria
- 4. Run Reproscreener on testbed
- 5. Extend Criteria based on empirical findings



## **Criteria Adapted from "Gunderson" (2018)**

- Method Transparency:
  - Problem; Objective, Goal; Research Method; Pseudo Code.
- Data Transparency:
  - Training, Test, Validation Data; Results.
- We include:
- Code Transparency:
  - Github, Bitbucket, Gitlab



# Preliminary Results (Work in Progress)

problem: 0.94 objective : 0.8 research method : 0.88 research\_questions: 0.84 pseudocode: 0.5 training data: 0.36 validation data: 0.02 test data : 0.12 results: 0.0

**I ISC** 

hypothesis: 0.42 prediction: 0.68 method source code: 0.3 hardware specifications: 0.0 software\_dependencies : 0.0 experiment setup: 0.36 experiment source code: 0.0 affiliation: 0.3

#### **Extending "Gunderson" Criteria**

Implemented the Gunderson criteria on 8 publications and identified the following gaps:

- Measure only the availability of code and data.
- Do not account for any errors during the reproducibility process.
- Partial reproduction of results are not captured.



## **Extending "Gunderson" Criteria**

ReproScreener:

- Automatically checks specific guidances to improve correctness of ML models.
- Predicts (error bounds), captures and identifies differences in model output at scale (due to architecture, non-determinism, etc.)
- Enables comparison of model code through
  - Checking for modularity, file structure, dependencies.
  - Checking for steps/scripts to create figures & visualizations.
  - Tracking model benchmarks and provenance.
- Real world case studies to demonstrate ReproScreener's functionality

## **Open Source Development** (work in progress)

ReproScreener:

- Automatically checks specific guidances to improve correctness of ML models.
- Predicts (error bounds), captures and identifies differences in model output at scale (due to architecture, non-determinism, etc.)
- Enables comparison of model code through
  - Checking for modularity, file structure, dependencies.
  - Checking for steps/scripts to create figures & visualizations.
  - Tracking model benchmarks and provenance.
- Real world case studies to demonstrate ReproScreener's functionality



#### **Open Source Development** (work in progress)

hrome	File	Edit	View	History	Bookmarks	Profiles	Tab	Window	Н	lelp
0	Machin	e-Learni	ng-Pipel	ines/re ×	+					
C 🔒 github.com/Machine-Learning-Pipelines/repro-screener										
i≘ README.md										
	ReproScreener									
ReproScreener aims to address challenges in robustness, transparency and interpretability of ML models b automating verification of machine learning models at scale.										y and interpretability of ML models by
	Project structure									
	• c: • g	<ul> <li>case-studies contain the papers that ReproScreener is tested on</li> <li>guidance contain the set of metrics that ReproScreener will check for</li> </ul>								
Features										
	• Au • Pr • Er	utomat edict, o nable c o Che o Che	ically c capture ompari cks for cks for	heck spec and iden son of mo modularit steps/scr	tific guidanc tify difference del code the y, file struct tipts to creat	es to impr ces in mod rough ure, depen e figures 8	ove co el outp dencie k visua	rrectness out at scal es lizations	s of le (d	ML models due to architecture, non-determinism, etc.)
		hrome File Machine C a git image REAL Reprose automa Proje c c g Feat • Au • Pr • Er	hrome File Edit Machine-Learni C ■ github.cc ≡ README.m ReproScreen automating v Project s • case-st • guidance Features • Automat • Predict, d • Enable c • Che • Che	hrome       File       Edit       View         Image: Comparison of the system       Machine-Learning-Pipeling         Image: Comparison of the system       github.com/Mach         Image: Comparison of the system       README.md         Image: Comparison of the system       ReproScreener aims automating verification         Project struct       . case-studies         . case-studies       . guidance conto         Features       . Automatically cl         . Predict, capture       . Enable comparii         . Checks for       . Checks for	hrome       File       Edit       View       History	hrome       File       Edit       View       History       Bookmarks <ul> <li>Machine-Learning-Pipelines/re</li> <li>github.com/Machine-Learning-Pipelines</li> <li>Image: README.md</li> </ul> <ul> <li>README.md</li> </ul> ReproScreener aims to address challenge automating verification of machine learning                Project structure <ul> <li>case-studies</li> <li>contain the papers the guidance</li> <li>guidance</li> <li>contain the set of metrics</li> </ul> Features <ul> <li>Automatically check specific guidance</li> <li>Predict, capture and identify difference</li> <li>Checks for modularity, file struct</li> <li>Checks for steps/scripts to creat</li> </ul>	hrome       File       Edit       View       History       Bookmarks       Profiles <ul> <li>Machine-Learning-Pipelines/re</li> <li>github.com/Machine-Learning-Pipelines/repro-screetile</li> <li>README.md</li> </ul> <ul> <li>ReproScreener</li> <li>ReproScreener aims to address challenges in robust automating verification of machine learning models</li> </ul> Project structure <ul> <li>case-studies</li> <li>contain the papers that ReproS</li> <li>guidance</li> <li>contain the set of metrics that ReproS</li> <li>guidance</li> <li>Predict, capture and identify differences in mod</li> <li>Enable comparison of model code through             <ul> <li>Checks for steps/scripts to create figures 8</li> </ul></li></ul>	hrome       File       Edit       View       History       Bookmarks       Profiles       Tab <ul> <li>Machine-Learning-Pipelines/re</li> <li>github.com/Machine-Learning-Pipelines/repro-screener</li> <li>github.com/Machine-Learning-Pipelines/repro-screener</li> </ul> <ul> <li>github.com/Machine-Learning-Pipelines/repro-screener</li> <li>github.com/Machine-Learning-Pipelines/repro-screener</li> <li>README.md</li> </ul> ReproScreener aims to address challenges in robustness, t automating verification of machine learning models at scal            Project structure <ul> <li>case-studies</li> <li>contain the papers that ReproScreener</li> <li>guidance</li> <li>contain the set of metrics that ReproScreener</li> <li>guidance</li> <li>Predict, capture and identify differences in model outp</li> <li>Enable comparison of model code through                 <ul> <li>Checks for modularity, file structure, dependencie</li> <li>Checks for steps/scripts to create figures &amp; visua</li> </ul> </li> </ul>	hrome       File       Edit       View       History       Bookmarks       Profiles       Tab       Window         Image: Machine-Learning-Pipelines/repro-screener       Image: Machine-Learning-Pipelines/repro-screener       Image: Machine-Learning-Pipelines/repro-screener         Image: README.md       Image: Machine-Learning-Pipelines/repro-screener         Image: ReproScreener       Image: Machine-Learning-Pipelines/repro-screener         Image: ReproScreener       Image: Machine-Learning-Pipelines/repro-screener         Image: ReproScreener       ReproScreener         ReproScreener       ReproScreener         ReproScreener       Image: Machine-Learning models at scale.         Project structure       Image: Case-studies         Image: Contain the papers that ReproScreener is teste       Image: Gase-studies         Image: Case-studies       Image: Contain the papers that ReproScreener will class         Features       Image: Case-studies       Image: Case-studies         Image: Case-studies       Image: Case-studies       Image: Case-studies         Image: Case-studies	hrome       File       Edit       View       History       Bookmarks       Profiles       Tab       Window       H         Image: Comparison of the properties of

Track model benchmarks and provenance

**USC** 

#### **Thank you**



# This material is based upon work supported by the REAL@USC-META Center and National Science Foundation Grant No 2138776



