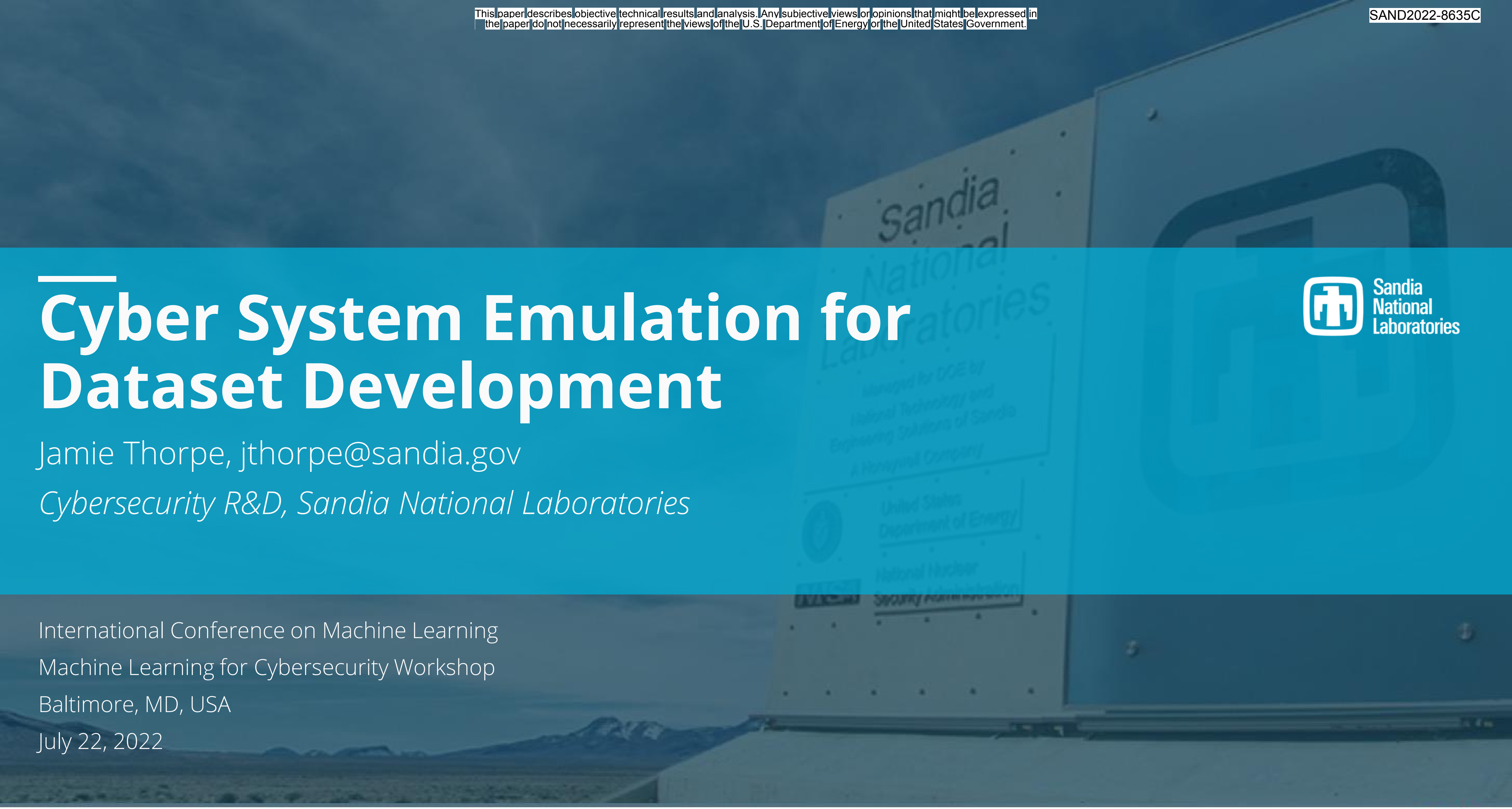
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-8635C



Cyber System Emulation for Dataset Development

Jamie Thorpe, jthorpe@sandia.gov

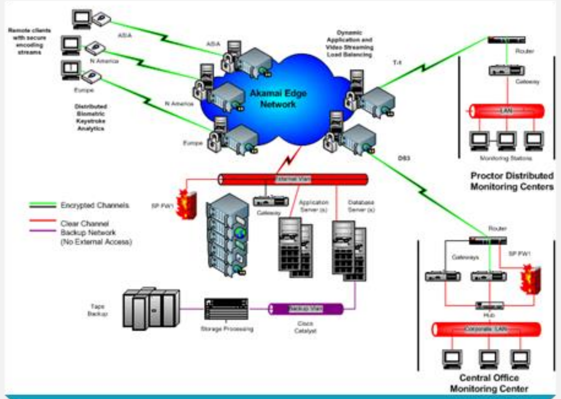
Cybersecurity R&D, Sandia National Laboratories

International Conference on Machine Learning
Machine Learning for Cybersecurity Workshop
Baltimore, MD, USA
July 22, 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

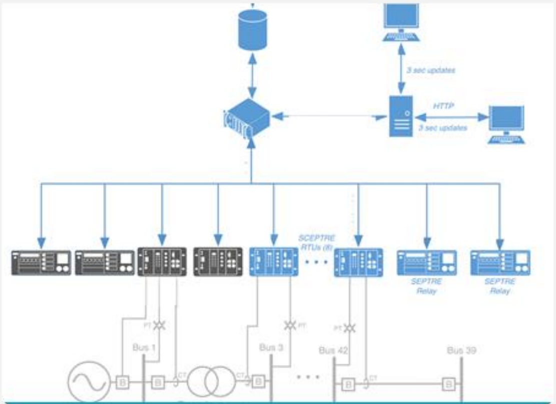
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

What is "Emulation"?



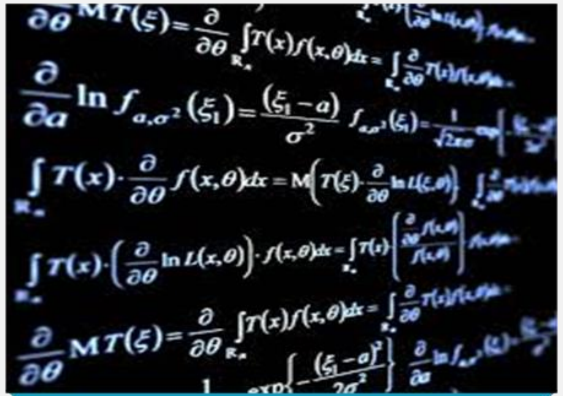
ACTUAL SYSTEM

REAL HARDWARE
REAL SOFTWARE



VIRTUALIZED TESTBED

ABSTRACT HARDWARE
REAL SOFTWARE



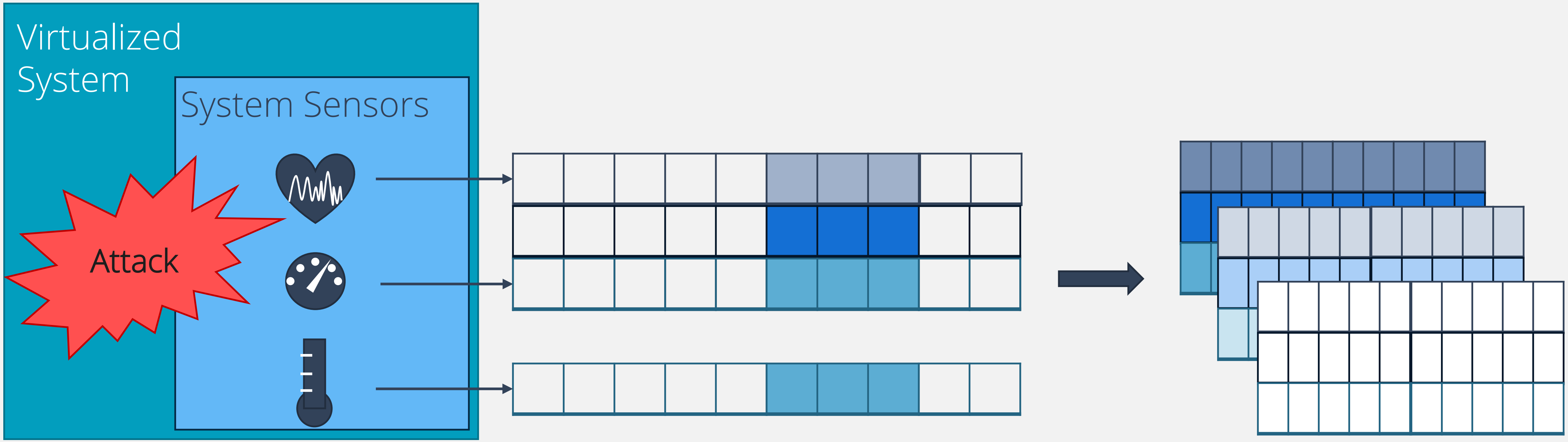
SIMULATION

ABSTRACT HARDWARE
ABSTRACT SOFTWARE

"BAD DAY" BRAINSTORMING

SUBJECT MATTER
EXPERT-DRIVEN

Emulation and Data Collection



Emulation for Dataset Development

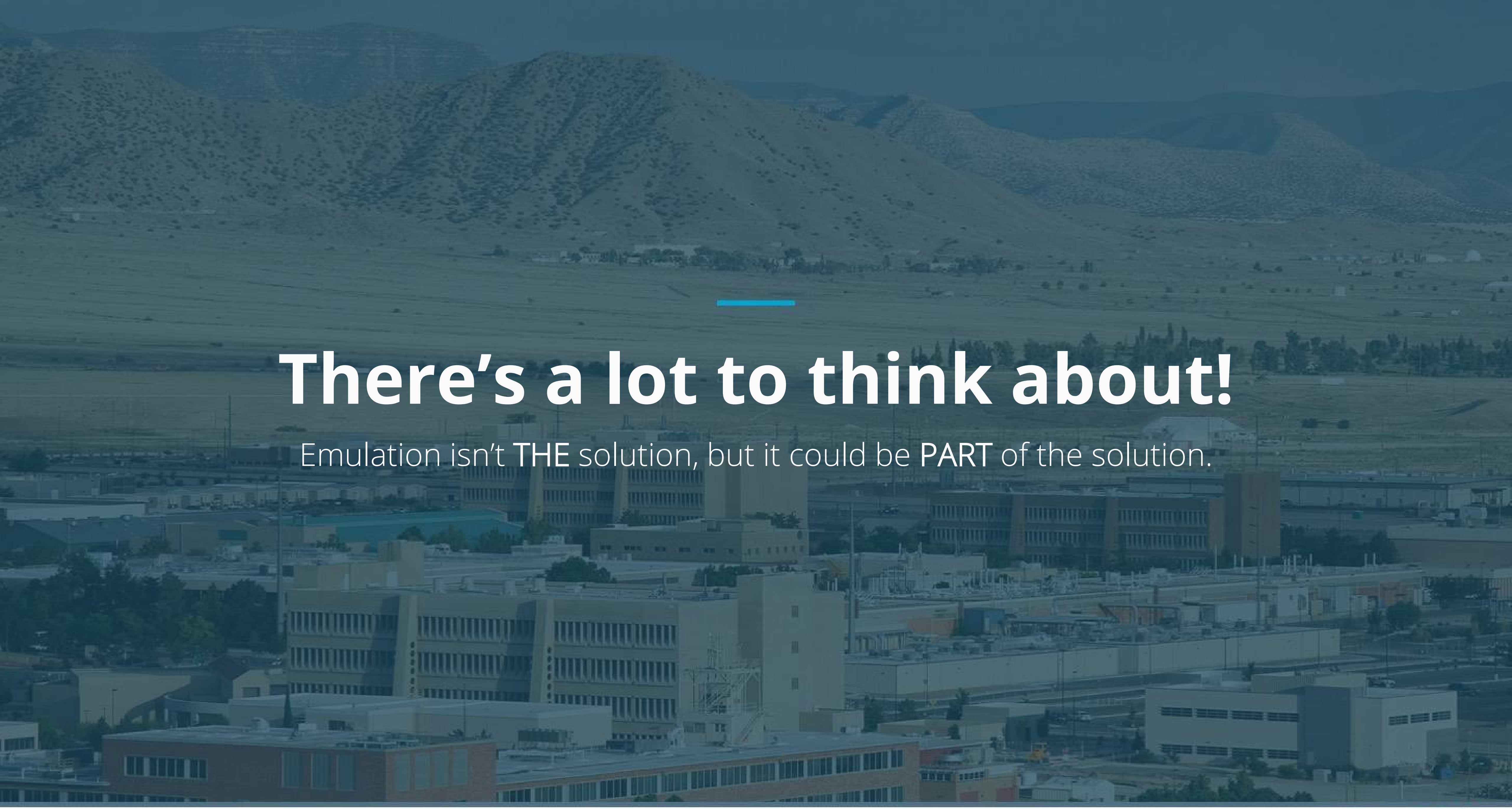
What are some of the big challenges for system emulation?

- Emulation model creation and validation
- Emulation verification
- Data collection
  - What data can be collected?
  - How much data do we collect?
  - How do we know that the data collection process doesn't disrupt the emulation?

Emulation for Dataset Development

Why isn't emulation the "simple" solution to developing datasets?

- Real-time runs
- Can Generative ML be applied here? If so, we need to consider…
    - How to apply common generative algorithms to multivariate timeseries
        - Different structure
        - Different datatypes
        - Different data relationships
    - How many samples do we need from the emulation in order to accurately represent each class?



There’s a lot to think about!

Emulation isn’t THE solution, but it could be PART of the solution.