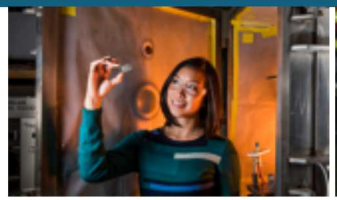
SAND2019-5290C

Sandia National Laboratories

A Causal Perspective on Data Integration



PRESENTED BY

Benjamin B. Schroeder and Lauren Hund

ASME V&V Symposium
May 15-18, 2019
Las Vegas, NV

Built on LDRD Laboratory Directed Research and Development

SAND2019-???? C

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.

Funding acknowledgement

This work was supported by a Sandia National Laboratories Laboratory Directed Research and Development (LDRD) grant. 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.

Outline

- **Define data integration**
- **Define structural causal modeling (SCM)**
- **Introduce SCM for data integration**
  - Integration steps
- **Pedagogical problem**
  - Walk through SCM data integration steps
  - When not to integrate
  - How to deal with integration complications
- **Conclusions**

What is Data Integration

Data integration is the use of multiple data sources to inform inferences (also known as data fusion).

Drivers for data integration include:

- Increased reliance on computational simulation based analyses;
- Lack of data for exact conditions/configuration/system of interest;
- Programmatic limitations for additional data.

Current challenge is establishing a process to ensure that the data is correctly integrated. If done incorrectly biases can be erroneously incorporated into models, leading to incorrect predictions.

What is Causal Modeling

Causal modeling refers to modeling how a system output changes due to controlled changes in system inputs.

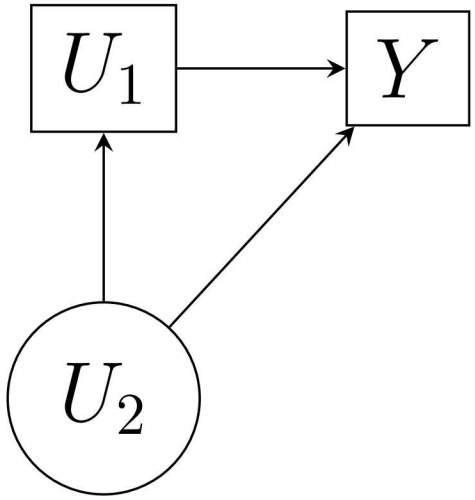
- This type of thinking is already common in engineering modeling.

Basic causal inference process relies on **counterfactuals.**

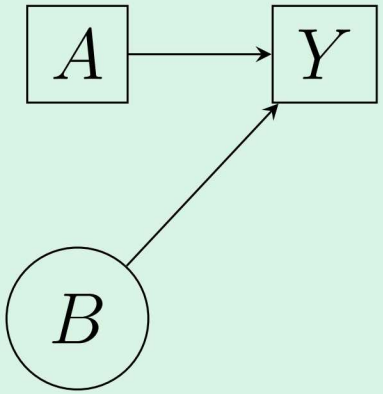
- Inference of events given hypothetical interventions based on knowledge of the underlying causal structure and information gained from other conditions.

Structural causal modeling can use causal diagrams to mathematically specify causal relationships between model inputs (U) and outputs (Y).

- SCM comprised of {F, U, Y}.
- True model not known, but hypothesized model structure is utilized based on observations and subject matter expertise.



How SCM Differs from a Standard Probabilistic Approach



The structural causal modeling approach identifies the **data generation mechanism** in terms of a structural model.

Assuming a phenomena can be described by the following function.

$$Y = \alpha_0 + \alpha_1 A + \alpha_2 B + \epsilon$$

$$\epsilon = \mathcal{N}(0, \sigma)$$

$$P(\alpha_0, \alpha_1, \alpha_2, \sigma | Y_{\text{obs}})$$

Is desired prediction estimable based on the available data? Probabilistic update of the underlying functional relationships.

Ultimately, the model with updated parameters is used to infer for conditions not observed (**counterfactual**).

$$P(Y | do(A = 5))$$

notation for fixing value in test

SCM Emphasis

Standard Probabilistic Approach

Steps to Causal Modeling for Data Integration

1. **Define a causal query**
   - What is the quantity of interest (QoI) or prediction scenario?
2. **Identify the data generation mechanism**
   - What is the underlying structural causal model?
   - What was observed?
   - What was not observed?
3. **Determine if causal query is identifiable from data**
   - Will the unobserved variables cause bias in the estimate of underlying causal relationships?
4. **Data integration**
   - Inform underlying functional relationships based on available data.
5. **Estimate causal query**
   - Make inferences.

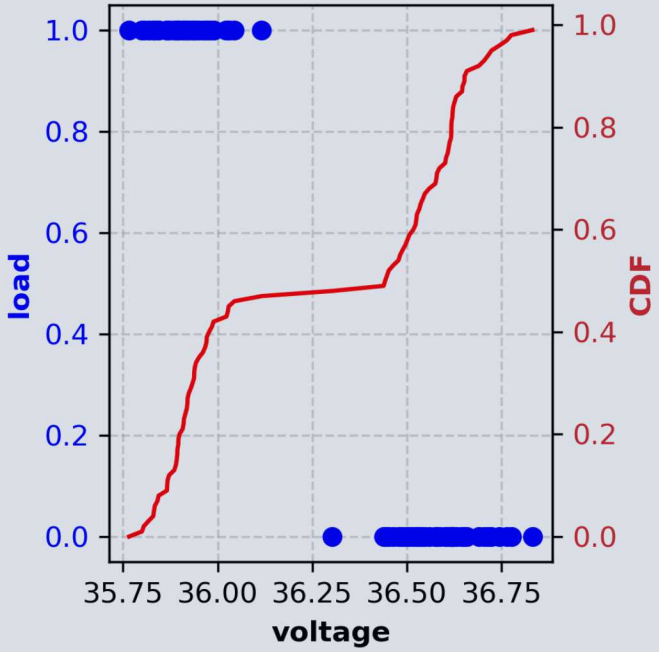
Example Description

Interested in **baseline voltage performance** of an **aging population** of thermal batteries for range of electrical current **loading profiles.**

- Load profiles (load) represented as intensities between 0 and 1.

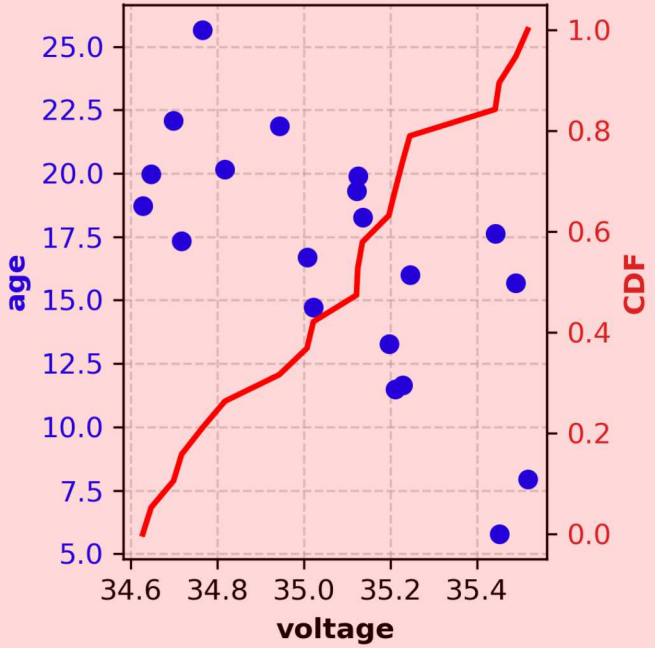
Two datasets are available to infer battery performance.

Production data



- Samples taken during production for two extremes of possible load profiles ***(n=200 tests)***.

Surveillance data



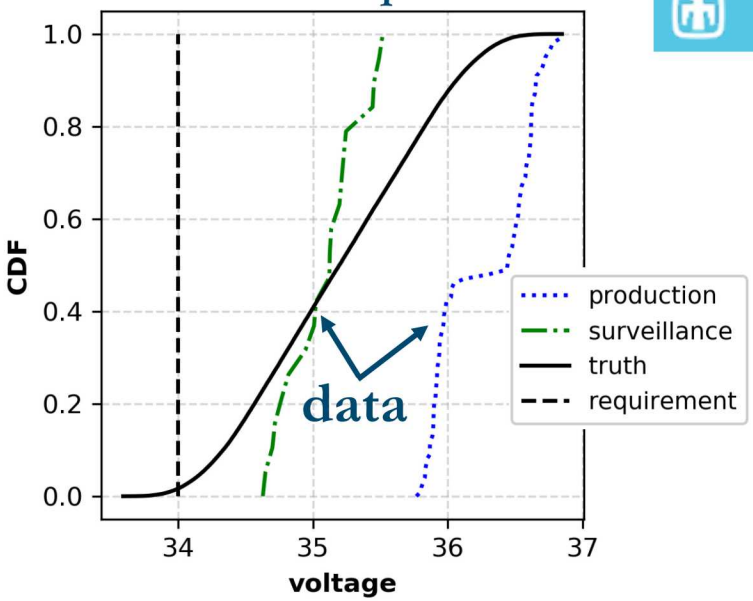
- Samples take during surveillance of the population, but loading profiles not reported ***(n=20 tests)***.

Ex: Define Causal Query (Step 1)

What is the baseline voltage for the population over its **lifetime** and **anticipated load** profiles?

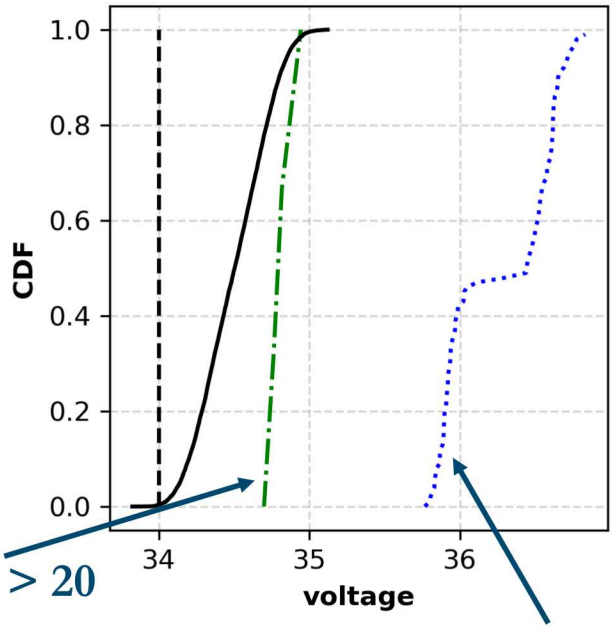
$$P(Y|do(A = a, L = l))$$

Available data most relevant to queries



What is the baseline voltage for the population **25 years** after production?

$$P(Y|do(A = 25))$$



surveillance data age > 20

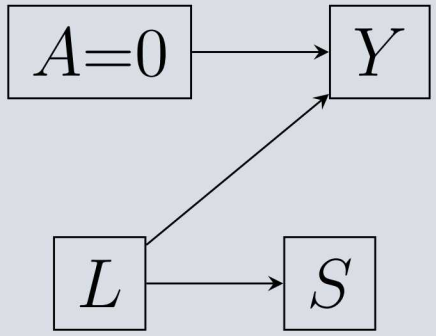
production data contains no aging information

Ex: Identify Data Generation Mechanism (Step 2)

Production data

Loading profiles applied were biased, but observed.

Age was fixed.

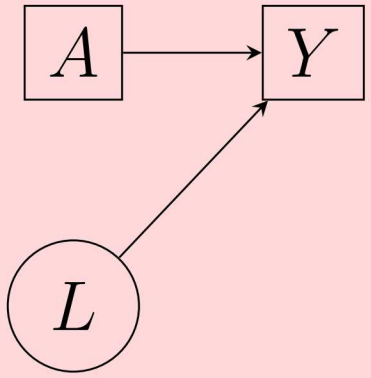


Surveillance data

Age was observed.

Loading profiles were not recorded, but assumed to not be biasedly sampled.

No confounding between age and load.



Ex: Determine if Causal Query is Estimable from Data

(Step 3)

Functional assumptions

- Linear relationship between E[Y] and A & L
- Normally distributed variability in Y

$$Y|do(A,L) = f(A,L)$$
$$= \alpha_0 + \alpha_1 A + \alpha_2 L + \epsilon$$
$$\epsilon = \mathcal{N}(0,\sigma)$$

Production data has sample bias in load, but can estimate a linear relationship between Y and L. No aging estimation.

- Estimate of $\alpha_0, \alpha_2, \sigma$

Surveillance data allows estimation of aging effect.

- Estimate of α_1
- Leverages estimates of $\alpha_0, \alpha_2, \sigma$ from production data

Integration of data should allow estimation of underling model given assumptions.

Ex: Data Integration (Step 4)

Probabilistically solved with **Pystan** (Stan library for Python).

Estimating $\alpha_0, \alpha_1, \alpha_2, \sigma$

- Linking observed data with counterfactual model
- Uninformed priors

$$P(Y|do(A=a, L=l))$$

$$Y_i^p \sim \mathcal{N}(\alpha_0 + \alpha_2 l_i, \sigma)$$

$$Y_i^s \sim \mathcal{N}(\alpha_0 + \alpha_1 a_i + \alpha_2 L, \sigma)$$

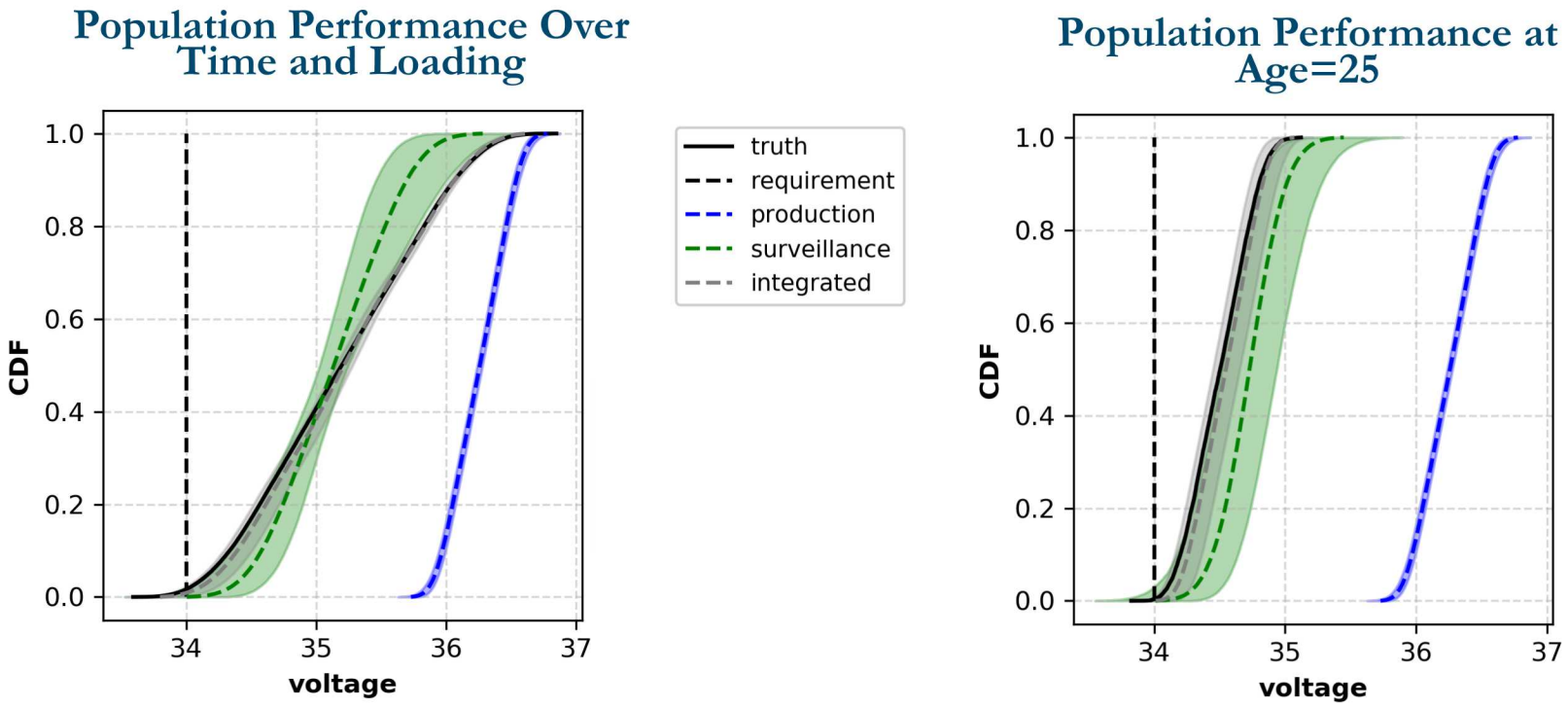
$$P(Y|do(A=25)) = \int P(Y|do(A=a, L=l)P(L)\mathrm{d}L$$

Ex: Estimate Causal Query (Step 5)

Comparing inference based on training counterfactual model to separate datasets or integrated dataset.

More accurate prediction with less uncertainty when using integrated inference.

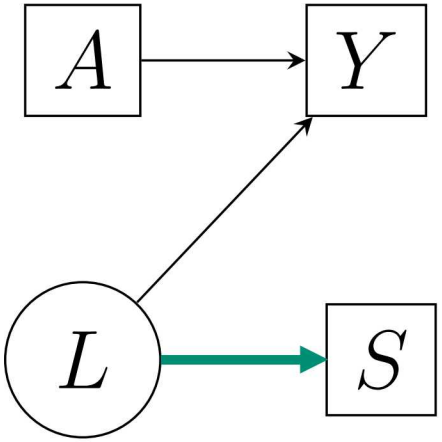
- Sample size reduces uncertainty in integrated versus surveillance counterfactual inferences.



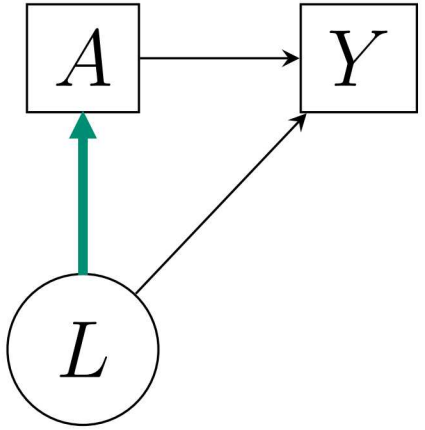
Shaded region shows 95% prediction bounds

Pedagogical Problem: When Data Can't be Integrated

Examples of alterations to **surveillance data** that would bias inferences in previous data integration.



Unobserved **sample bias** on loading.
Any estimate of A effects will be biased.



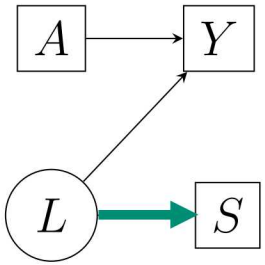
Confounding bias of unobserved L on A.
Any estimate of A effects will be biased.

Pedagogical Problem: How to Deal with Complications

Potential presence of additional biases/complications in data may required inclusion of auxiliary information to make unbiased inferences.

Ex. Suspected Sample Bias

- Potential sample bias in surveillance data due to tester error.



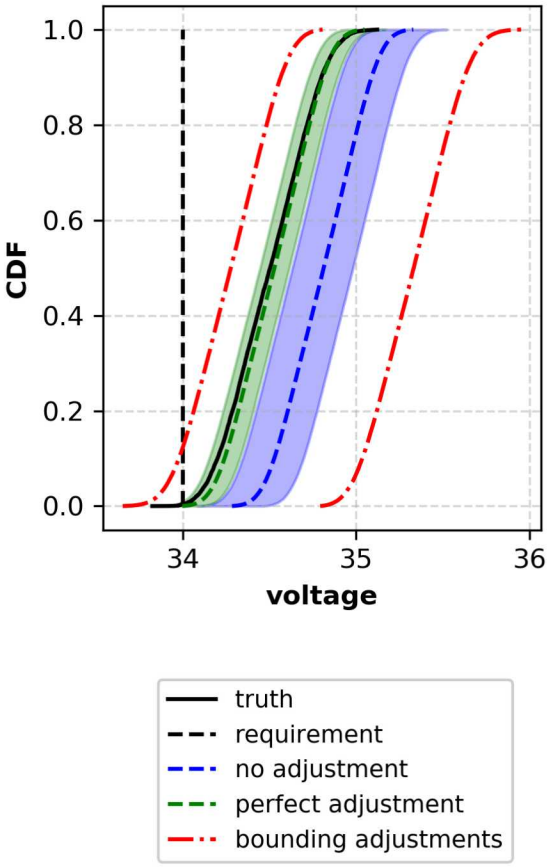
- Tester weighted applied loading profiles.

$$P(L) = \mathcal{N}(0, 0.25) \quad T\,[0, 1]$$

- Study the sensitivity to loading profiles to assess potential impact on inferences.

$$L = [0, 1]$$

Population Performance at Age=25



Conclusions

Structural Causal Modeling forces us to consider the data generation mechanism behind the data.

Reduces likelihood of bias infiltrating inferences.

We have proposed steps to applying structural causal modeling to data integration:

1. **Define a causal query,**
2. **Identify the data generation mechanism,**
3. **Determine if causal query is identifiable from data,**
4. **Data integration, and**
5. **Estimate causal query.**

Data integration can lead to better inferences, but sensitivity to assumptions should be considered.

Thank You for Your Attention

Presenter's Contact Information:

Benjamin Schroeder

Sandia National Laboratories

V&V, UQ, and Credibility Processes Department

bbschro@sandia.gov