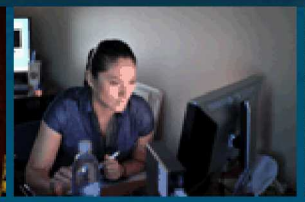
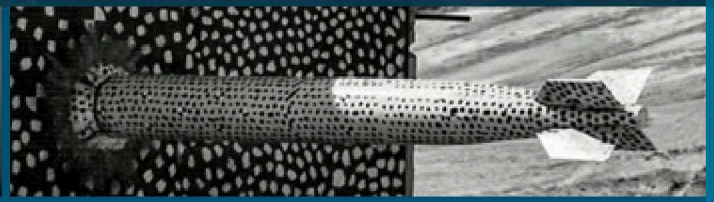
SAND2020-9499PE

Assessing Global Sensitivity Analysis for Credibility in Machine Learning Explainability





SAMSI Presentation
Numerical Analysis for Data Sciences Program
Global Sensitivity Analysis Working Group
2020-09-17

PRESENTED BY

Erin Acquesta and Mike Smith

Rich Field, Trevor Maxfield, Ahmad Rushdi,
and many others…

The Need for Credible Explainability

ML is being used in an increasingly number of high-consequence applications.

ML explainability has emerged as field that seeks to build trust.

The Need for Credible Explainability

ML is being used in an increasingly number of high-consequence applications.

ML explainability has emerged as field that seeks to build trust.

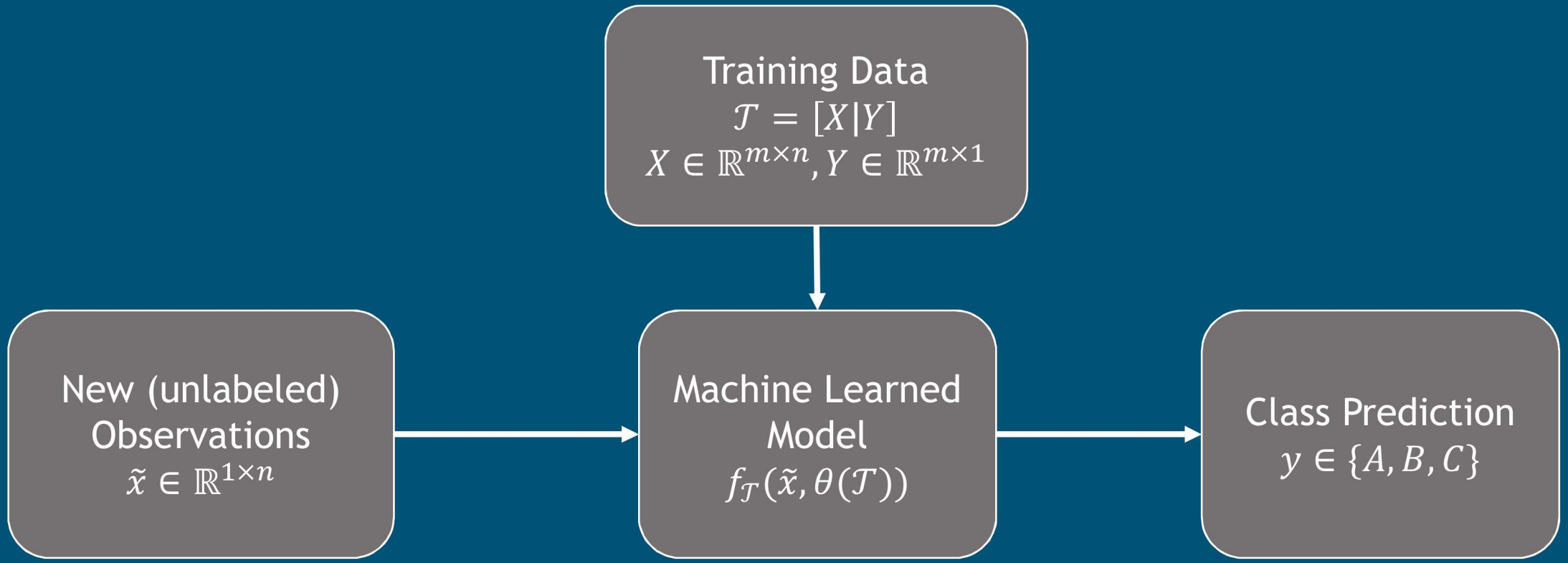
Can we trust the explanation?

- Computational shortcuts
- Assumes some understanding of machine learning
- Lack verifiable foundations

Current ML Explainability Methods

- Sensitivity Analysis Guided Explainability
- Correlated Feature
- Training Data Statistics Preserving Sampling
- Gaps and Limitations

Machine Learning (Supervised) Classification Model Diagram



ML Explainability

Attempt to describe the decision process that a machine learned model uses to make a prediction

Interpretable models

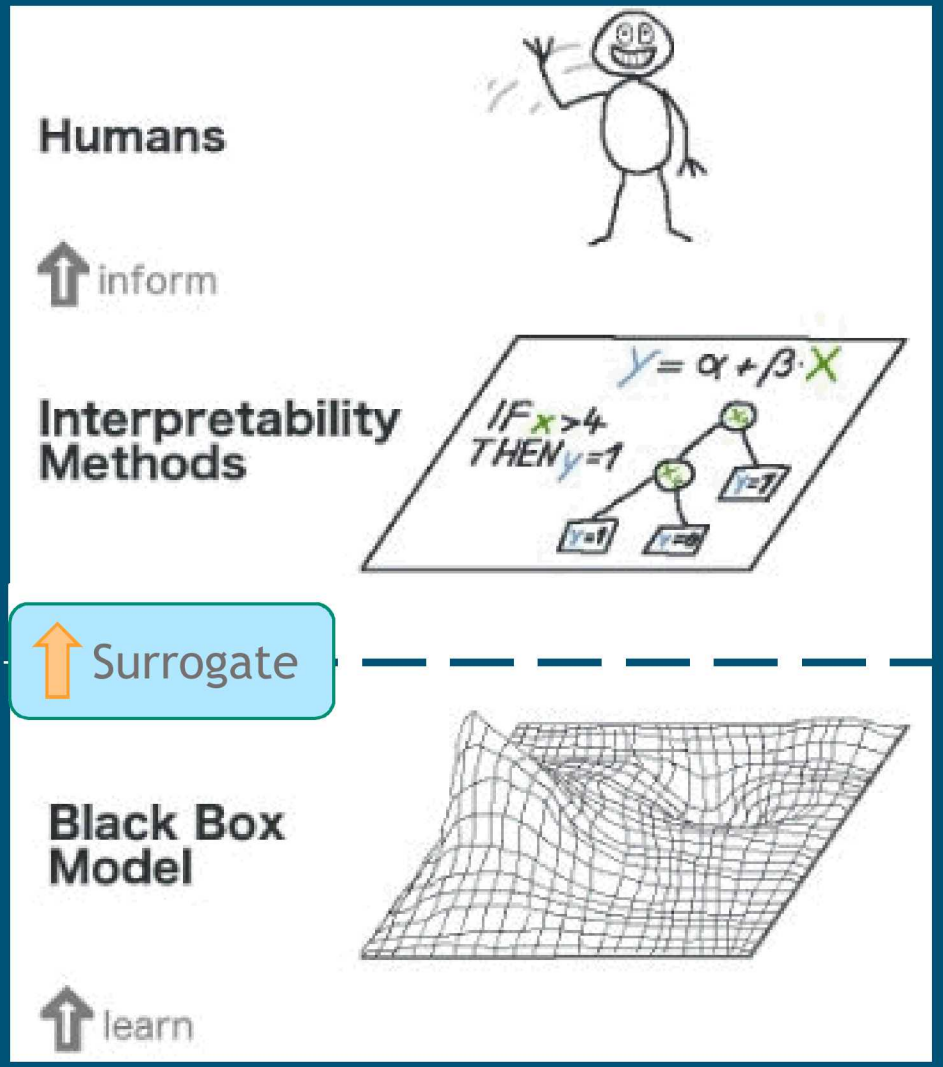
- Can inspect the model
- Models that are relatively easy to interpret (linear regression models, shallow decision trees)

White-box/Integrated

- Can inspect the model, but the model is sufficiently complex
- Gini-importance for decision trees
- Gradient-based methods for deep learning models (Saliency maps)

Black-box/Post-hoc

- Do not inspect the model, instead evaluates feature importance
- Create a surrogate model that is interpretable
- Often perturb the data and observe how the output changes



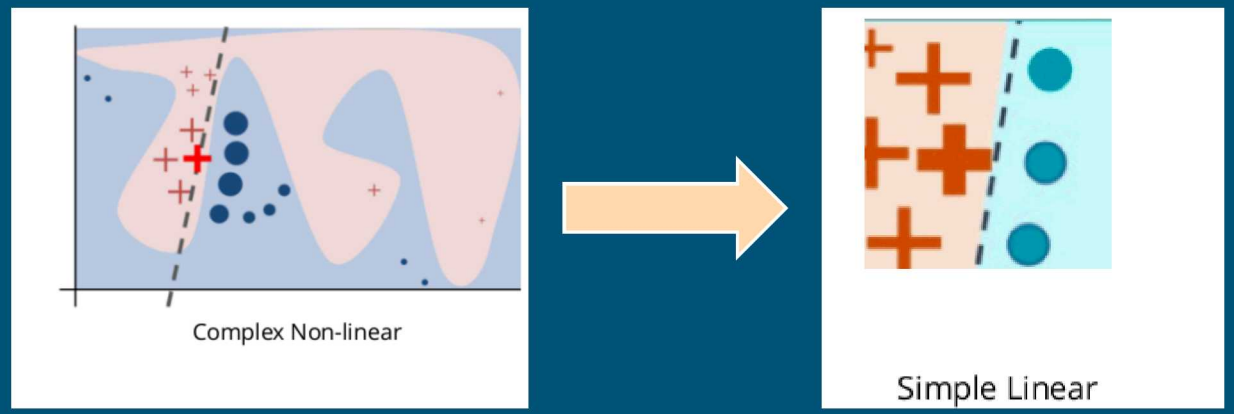
LIME: Local Interpretable Model-agnostic Explanation [1]

Perturbation Based Method

For $\tilde{x} \in \mathbb{R}^{1\times n}$ (a new realization)

1. Sample $\tilde{m}$ observation in a neighborhood of $\tilde{x}$, assuming that each feature of $\tilde{x}$ is independent and normally distributed. Resulting in: $\{\tilde{z}^i\}$ for $i = 1 \dots \tilde{m}$
2. For each of the samples, get the original ML model prediction $f_{\mathcal{T}}(\tilde{z}^i) = \tilde{y}^i$
3. With the derived sample dataset $\tilde{\mathcal{T}} = \{\tilde{z}^i | \tilde{y}^i\}$, learn a linear regression model,

$$g_{\tilde{\mathcal{T}}}\left(\tilde{x}, \theta(\tilde{\mathcal{T}})\right) = \sum_{j\in\{1,\dots,\tilde{m}\}} \alpha_j \tilde{x}_j$$



LIME feature importance index =

$$\phi_j = \alpha_j * \tilde{x}_j$$

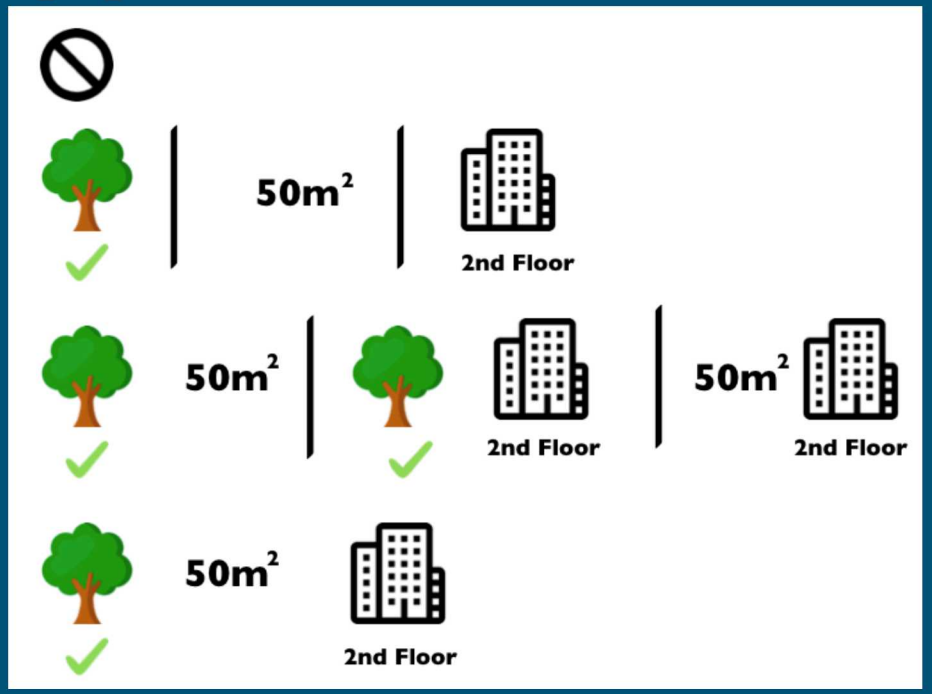
SHAP: SHapley Additive exPlanations [2]

Based on Cooperative Game Theory

N : the set of features
$j \in N$
$S \subseteq N$

$v(S)$: the total value of the S features

$(v(S \cup \{j\}) - v(S))$ for $S \subseteq N \backslash \{j\}$:
marginal contribution of feature j to set S



SHAP feature importance index:

$$\phi_j = \sum_{S \subseteq N \backslash \{j\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{j\}) - v(S))$$

One Computational Approximation:

- To approximate $(v(S \cup \{j\}) - v(S))$ without retraining the model, “noise” is used as a surrogate for feature removal.
- “noise” is replicated as independent bootstrap samples from the training data

Summary of Black-Box Explanation Methods

Dependencies and Assumptions

- Dependent on a process for sampling the data
- Require distance on output—how much it changes
- Assumes independence and linearity

Observed Deficiencies [3]

- Descriptive Accuracy: Match when features are removed
- Instability: Produces different explanations on the same input
- Completeness: Generate explanations for all possible input vectors
- Efficiency: Can be slow to calculate especially as the dimensionality increases

No agreed upon definition of explainability or what constitutes an explanation

Can We Trust the ML Explanation??

Goal of explainability: provide credibility evidence by describing the decision process that a model considers for a prediction.

BUT….

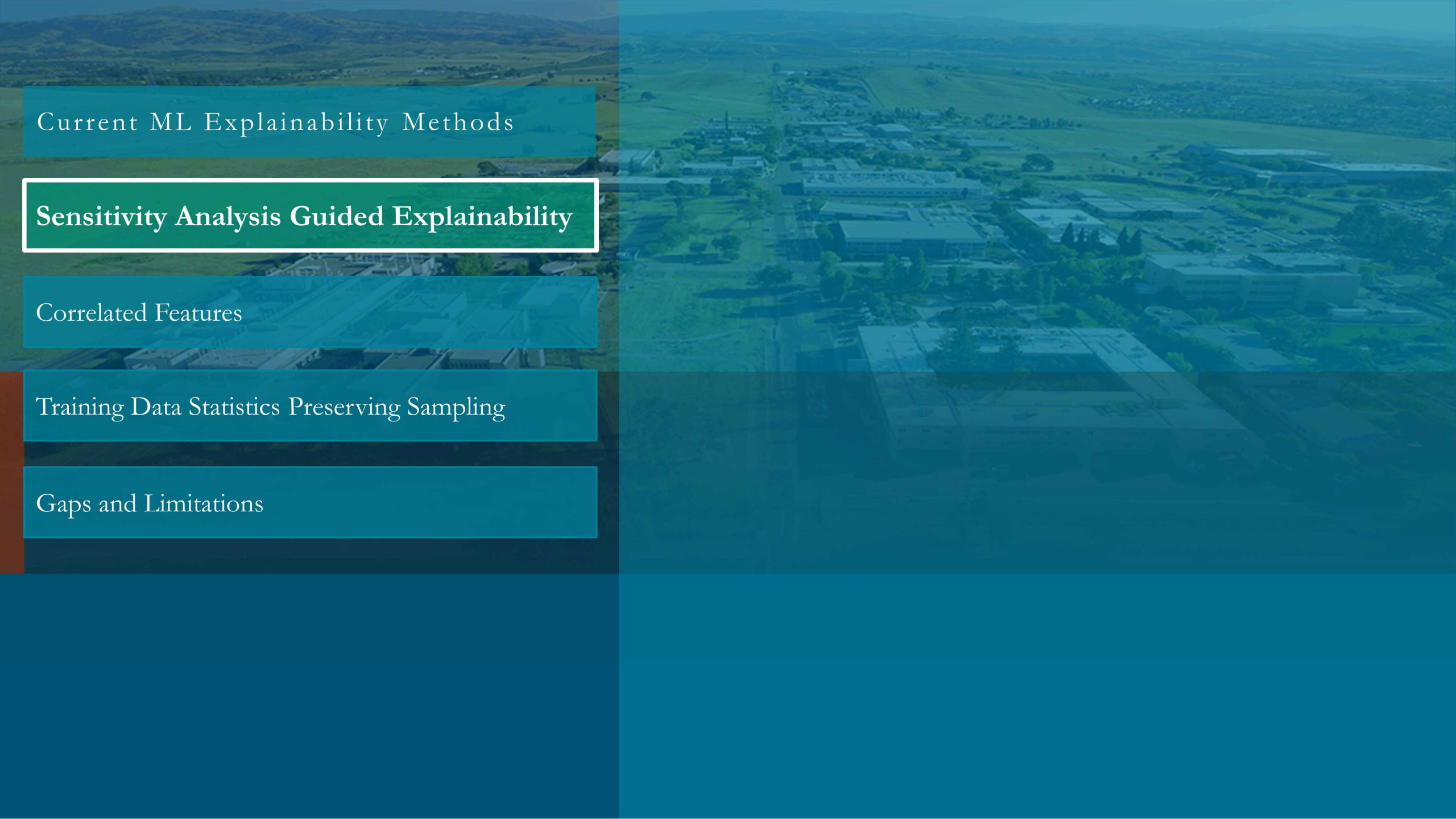
LIME

- Introduces a surrogate model without assessing the fidelity of that surrogate in relation to the black-box model.
- Implements perturbation of the input uncertainty assuming independence and normally distributed random variates…which is uncharacteristic of the true statistics from the training dataset.

SHAP

- Has the most theoretical guarantees available to us today, but computational heuristics for the approximations break those theoretical guarantees.
- Bootstrapping from the training dataset to replicate ”noise”…but is ”noise” an appropriate replicate for removing a feature??

Can we use Sensitivity Analysis Guided Explainability to mitigate these challenges??



Current ML Explainability Methods

Sensitivity Analysis Guided Explainability

Correlated Features

Training Data Statistics Preserving Sampling

Gaps and Limitations

Global Sensitivity Analysis: Notational Example

The apportionment for the contributions of input uncertainties on output uncertainty. [4]

Modeling Flow Diagram

Nominal Parameter Values $\theta = \{\alpha, \beta, \gamma\}$ → Model of the Real World $f(x, \theta)$ → Deterministic Outcome $y \in \mathbb{R}^v$

Example: Infectious Disease Modeling

α = seroprevalence
β = infection rate
γ = [infectious duration]$^{-1}$

→

$$x = \begin{bmatrix} S \\ I \\ R \end{bmatrix} \quad \frac{dx}{dt} = f(x, \theta), \quad x_0 = g(\theta)$$

→

$$y = F(x)$$

Uncertainty Quantification of Model Forecast for the Infection Rate Curve

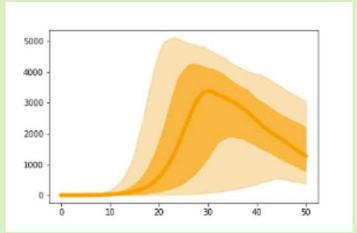
→

$$\begin{bmatrix} [\alpha_1, \beta_1, \gamma_1] \\ [\alpha_2, \beta_2, \gamma_2] \\ \cdots \\ [\alpha_s, \beta_s, \gamma_s] \end{bmatrix}$$

→ $f(x, \theta)$ →

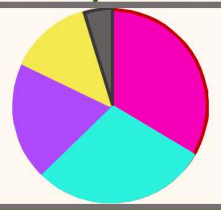
$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_s \end{bmatrix}$$

→



Sensitivity of the Infection Rate Curve with respect to the Parameters of the Model

Total Variance of y

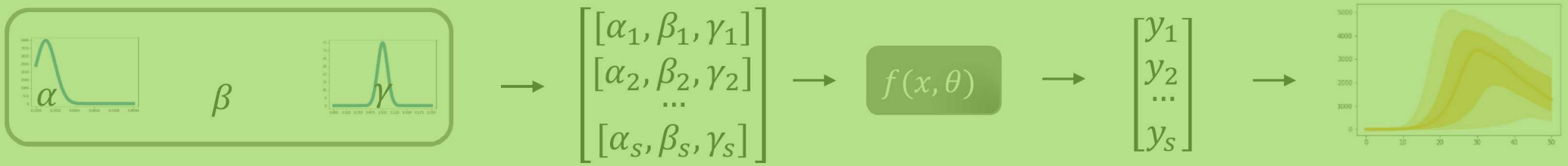


$S_{\beta\gamma}$
S_{β}
$S_{\beta\alpha}$
S_{α}

Sum of ALL other combined contributions

Experimental Design for Sensitivity Analysis

Experimental Design is a scientific approach for identifying the inputs to a *process* that are most influential to the outcome of that process; following particular design decisions.



Inputs:
Uncertainty in Parameters

Process:
Mathematical Model

Outcome:
Uncertainty in Model Output

Design Decision I

Sampling

Sampling sufficient discrete realizations that preserve the statistics and introduces only marginal standard error.

Design Decision II

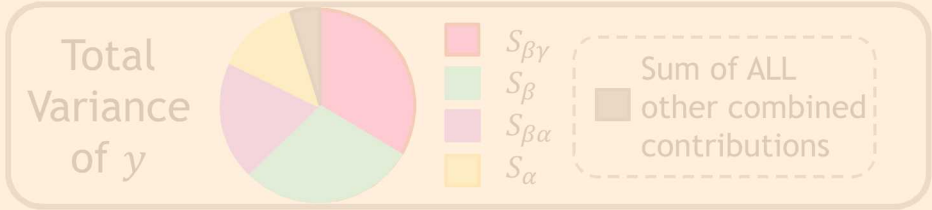
Controlled/Uncontrolled Random Behavior

Controlled: Sources of Variance
Uncontrolled: Random behaviors inherent to the model

Design Decision III

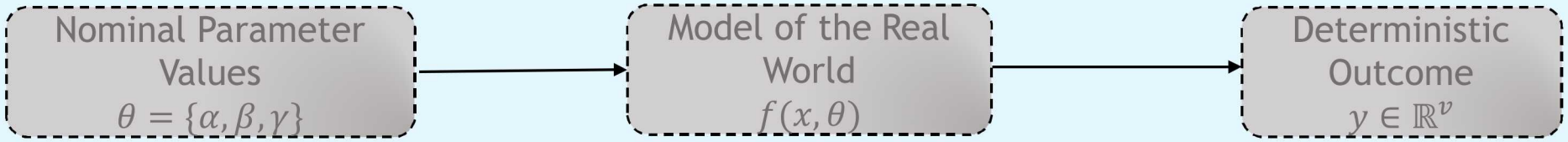
Quantity of Interest (QoI)

For the intended use case, what output from the model maps to quantitative metric for that intended purpose.

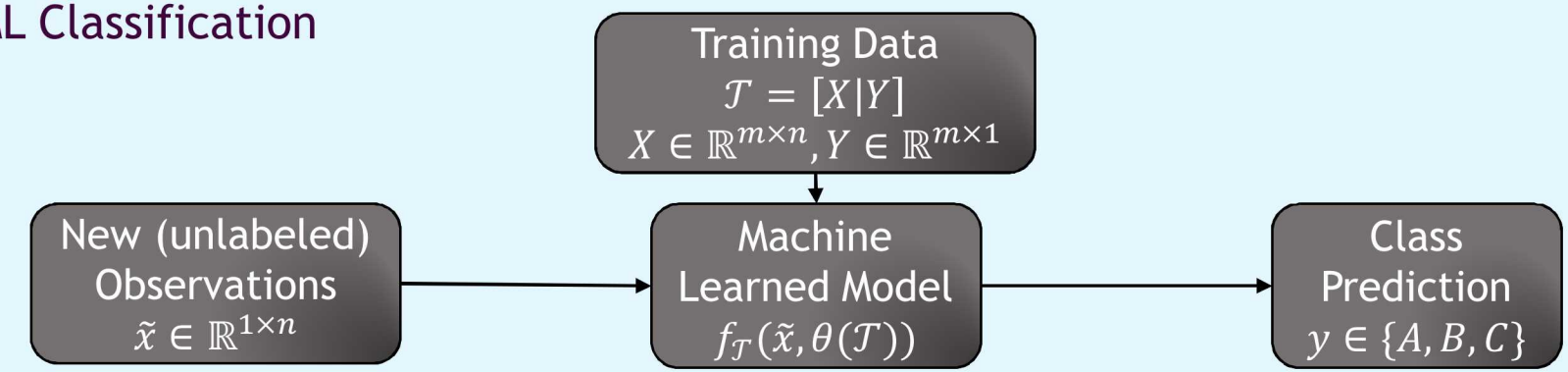


Experimental Design for ML Explainability

Original Modeling Flow Diagram



Translation to ML Classification



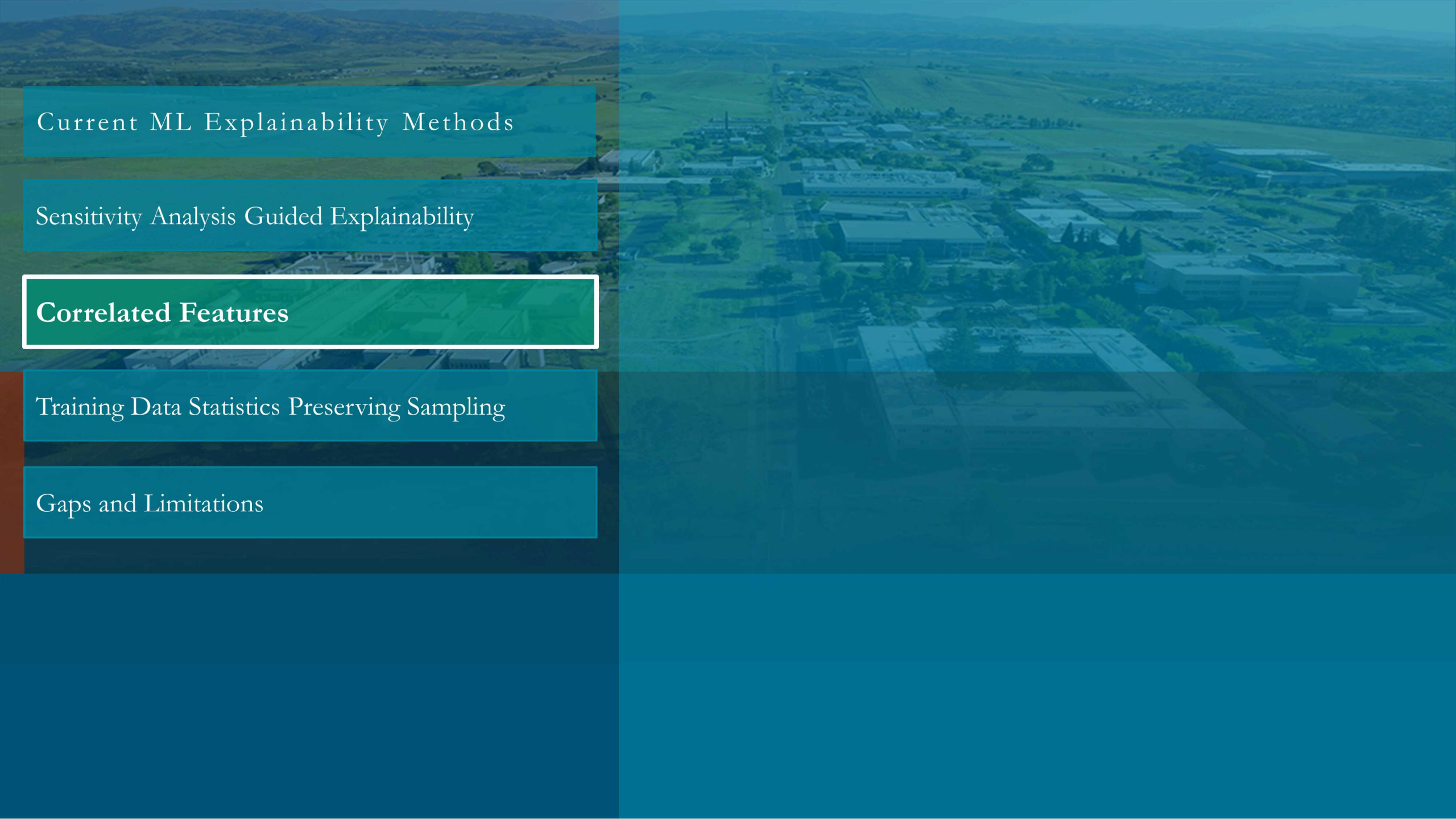
Inputs:
Uncertainty in Features

Process:
Machine Learned Model

Outcome:
Uncertain Model Predictions

Sampling	**Controlled/Uncontrolled Random Behavior**	**Quantity of Interest (QoI)**
Preserving the statistical properties of the training data: non-Gaussian, discrete, correlated, and sparse	Running sufficient replicates for the random behavior of stochastic machine learned models.	What is the appropriate QoI for which a sensitivity analysis will provide insight for ML explainability?

Methods to apportion the influence of sources of input uncertainty across output uncertainty, accounting for higher-order interactions in a model and input correlations.



Current ML Explainability Methods

Sensitivity Analysis Guided Explainability

Correlated Features

Training Data Statistics Preserving Sampling

Gaps and Limitations

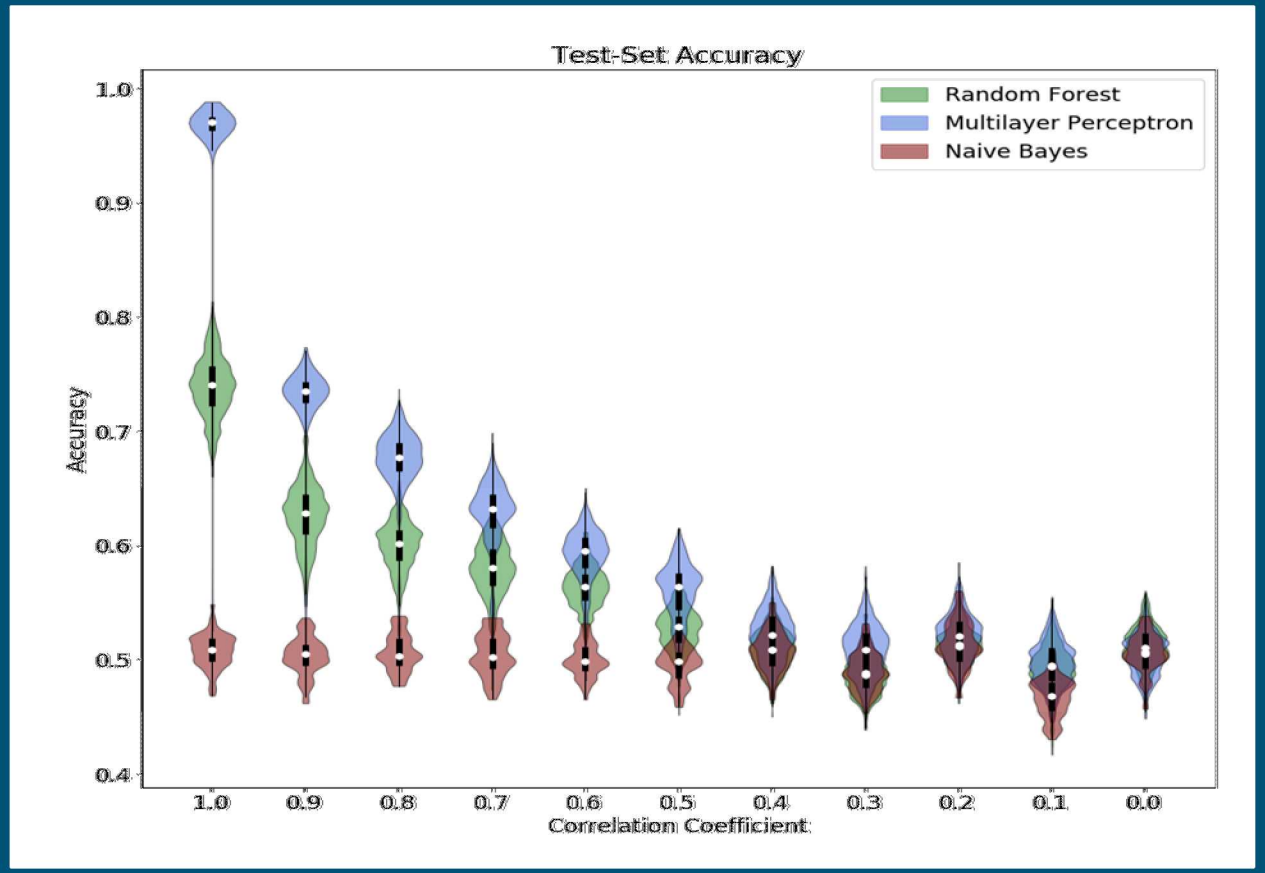
Correlated Features Provide Discriminative Power

Design of Experiments was used to the debunking the myth that correlated variables only provide redundant information.

- Used synthetic data to control the amount of correlation as the distinguishing characteristic between classes
- Naïve Bayes is the baseline for linear relationships

Most explainability methods assume independence

- Incongruent explanations for the learned model
- LIME uses a linear model
- SHAP makes independence and linear assumptions
- Tested with quadradic regression



Current ML Explainability Methods

Sensitivity Analysis Guided Explainability

Correlated Features

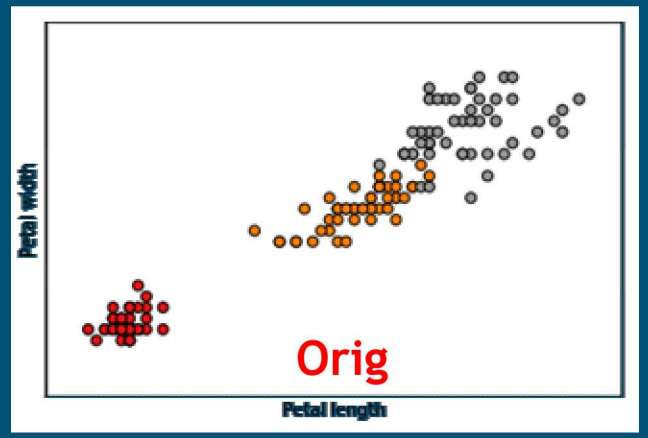
Training Data Statistics Preserving Sampling

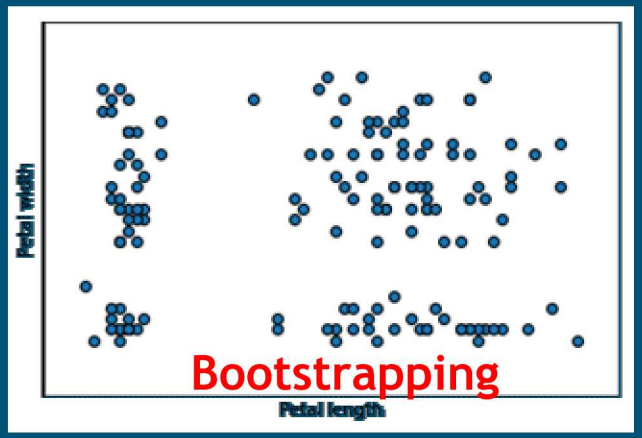
Gaps and Limitations

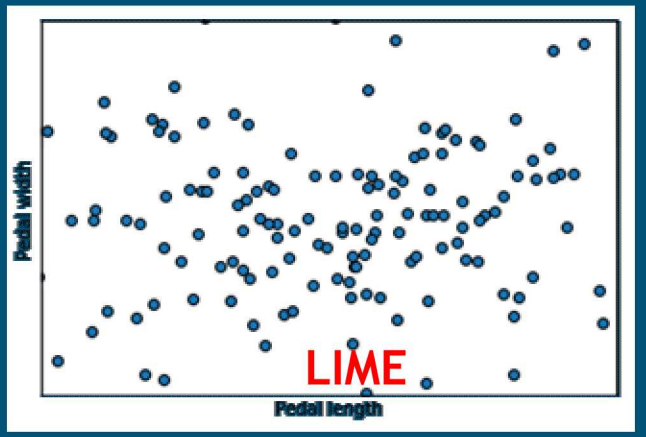
Training Data Statistics Preserving Sampling

Current sampling approaches can create unrealistic data points (Do not preserve correlations or distributions)

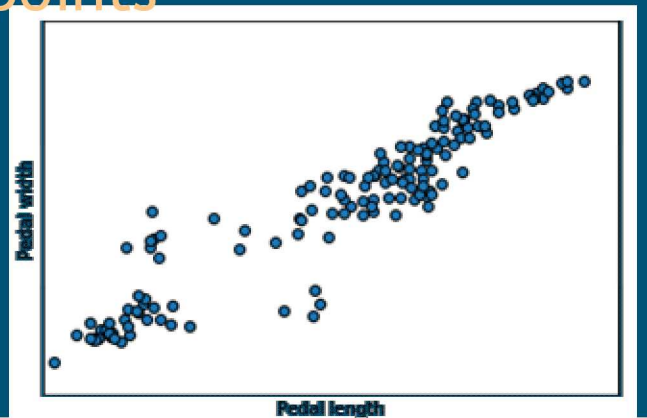
Features 3 and 4 from the iris data set (correlated features)

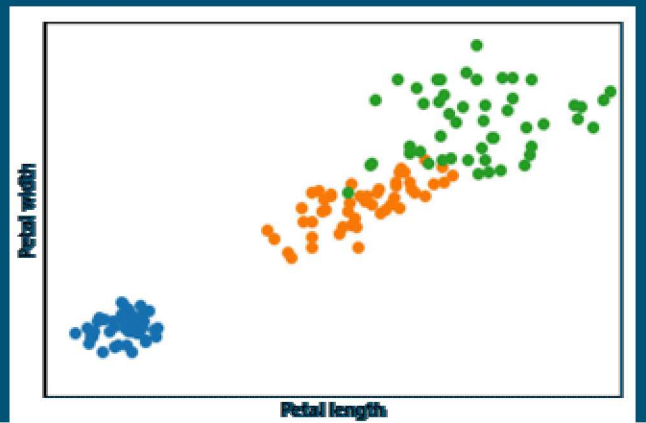


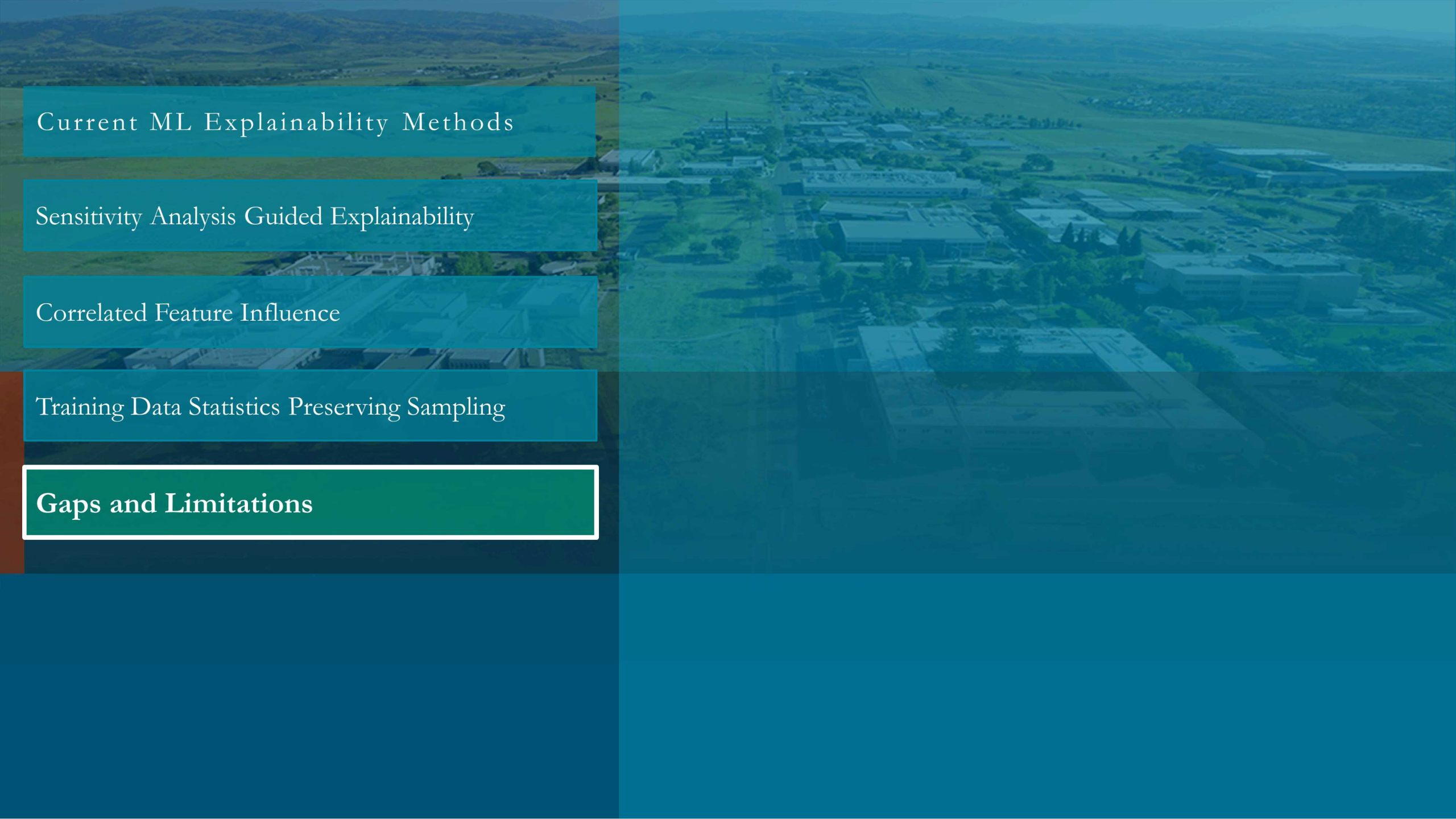




Developed Sampling methods that preserve correlations and generate realistic data points







Current ML Explainability Methods

Sensitivity Analysis Guided Explainability

Correlated Feature Influence

Training Data Statistics Preserving Sampling

Gaps and Limitations

Blockers to using GSA on ML Models for Explainability

- Dependence in our feature space, which defines our sources of uncertainty
- Appropriate Quantity of Interest that will map to an explanation.
- Expedient Computational Methods for our correlation and distribution preserving sampling technique.

Questions to the Working Group:

- Are there scalable methods for approximating the Shapley indices?
- How are dependent sources of uncertainty currently handled by the research community?

Continued Research Directions:

- Extensions of Sobol indices for dependent features. [5,6]
- Verification of computational estimators for SHAP feature importance indexes.

References

1. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

2. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774).

3. Warnecke, A., Arp, D., Wressnegger, C. and Rieck, K., 2019. Evaluating explanation methods for deep learning in security. *arXiv preprint arXiv:1906.02108*.

4. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S., 2008. *Global sensitivity analysis: the primer*. John Wiley & Sons.

5. Hart, J. and Gremaud, P.A., 2018. An approximation theoretic perspective of Sobol'indices with dependent variables. *International Journal for Uncertainty Quantification*, *8*(6).

6. Chastaing, G., Gamboa, F. and Prieur, C., 2012. Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, *6*, pp.2420-2448.