Sandia National Laboratories

Exceptional service in the national interest

FORESTR

Searching for patterns in random forests

Katherine Goode

Joint work with J. Derek Tucker

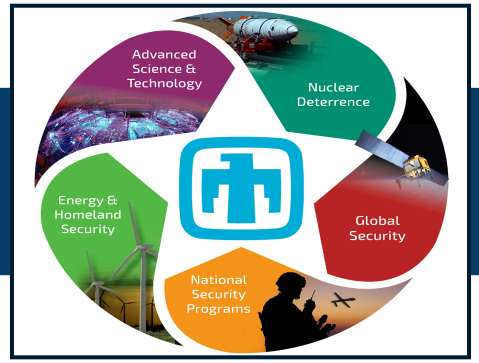
March 12, 2025

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

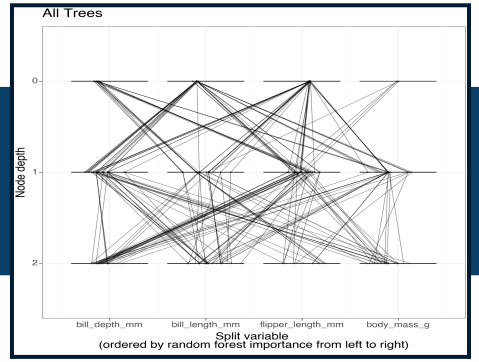
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525

SAND2025-02862PE

OVERVIEW



Background: Sandia & Machine Learning



FORESTR: Topology Patterns in Random Forests

BACKGROUND

SANDIA & MACHINE LEARNING

SANDIA IS A FEDERALLY FUNDED RESEARCH AND DEVELOPMENT CENTER MANAGED AND OPERATED BY

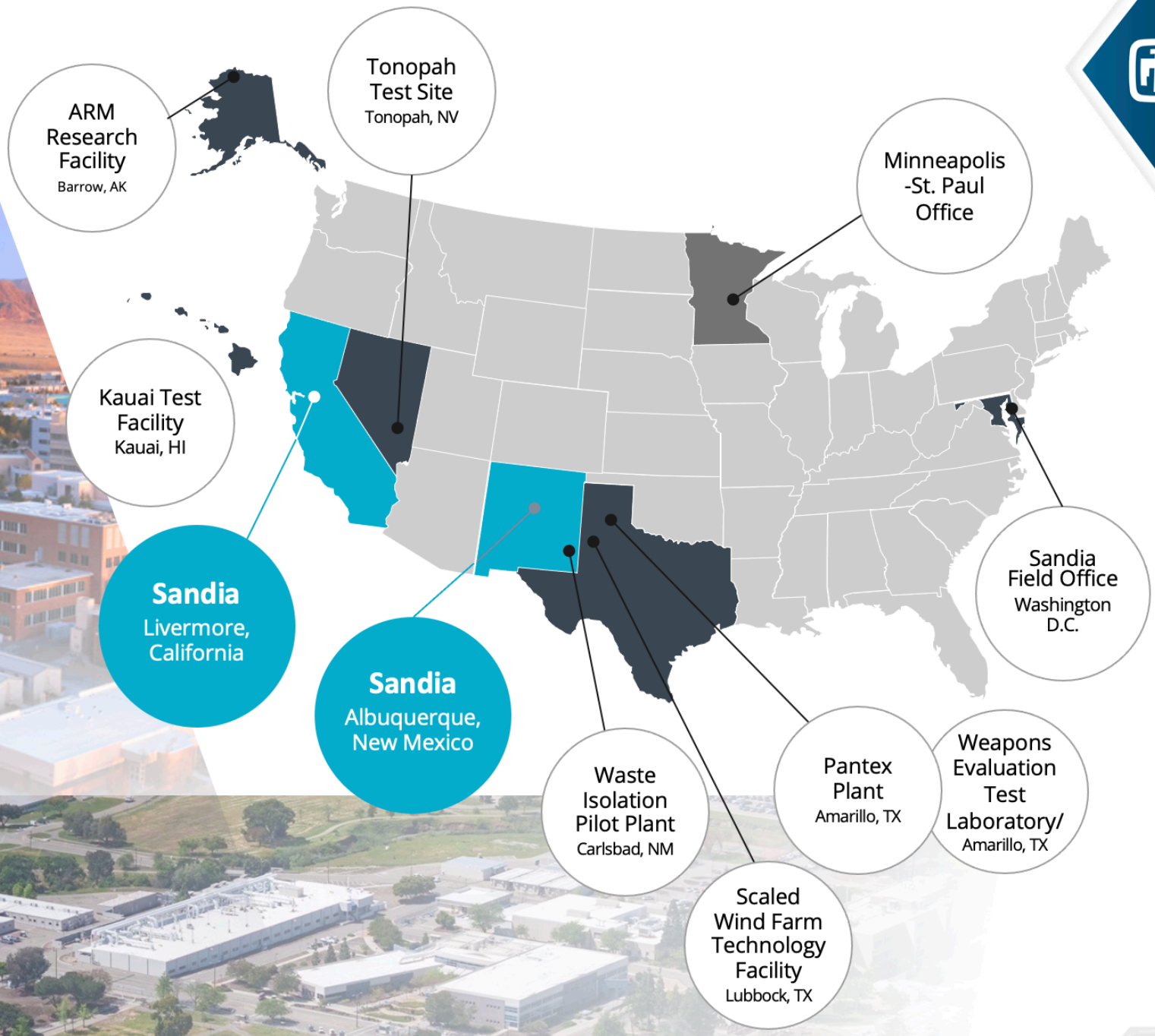
National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc.

Government-owned, contractor-operated

FFRDCs are long-term strategic partners to the federal government, operating in the public interest with objectivity and independence and maintaining core competencies in missions of national significance

WE HAVE FACILITIES ACROSS THE NATION

Main Sites
Albuquerque, New Mexico
Livermore, California



NATIONAL SECURITY IS OUR MISSION

Sandia delivers essential science and technology
to address the nation's most challenging security issues

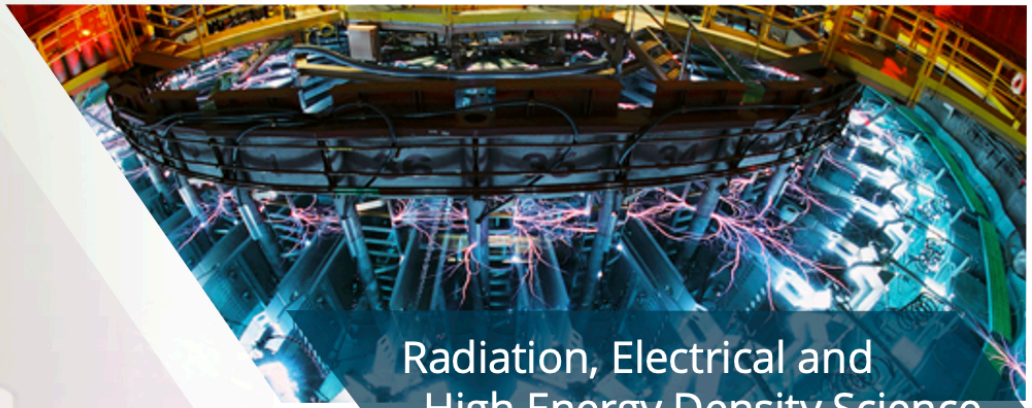
PURPOSE	We render exceptional service in the national interest
VISION	We make Sandia a leader in keeping the world safe and secure
MISSION	We use innovative science and engineering to anticipate and solve the most challenging national security problems
OBJECTIVE	In 10 years, we will have unleashed high-velocity engineering to counter global threats

SANDIA HAS FIVE MAJOR PROGRAM PORTFOLIOS

ADVANCED SCIENCE & TECHNOLOGY

Research Foundations play an integral role in mission delivery

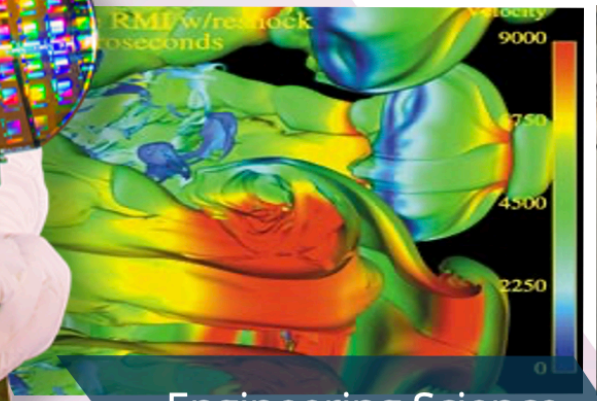
Nanodevices & Microsystems



Radiation, Electrical and High Energy Density Science

Materials Science

Computing Information Science



Engineering Science

Earth Science

Bioscience

MACHINE LEARNING AT SANDIA: A BALANCING ACT

MACHINE LEARNING AT SANDIA: A BALANCING ACT

The National Nuclear Security Administration (NNSA) Labs emphasize **trusted artificial intelligence (AI)** as a necessity to meet national security mission delivery.

MACHINE LEARNING AT SANDIA: A BALANCING ACT

The National Nuclear Security Administration (NNSA) Labs emphasize **trusted artificial intelligence (AI)** as a necessity to meet national security mission delivery.

- Recognize the value AI could provide to achieve mission goals
- Evaluating the credibility of current techniques poses challenges that may hinder widespread acceptance and use
- Sandia's mission needs set us apart from industry and academia due to reasons such as high-consequence applications

Sandia's five major mission portfolios

MACHINE LEARNING AT SANDIA: A BALANCING ACT

The National Nuclear Security Administration (NNSA) Labs emphasize **trusted artificial intelligence (AI)** as a necessity to meet national security mission delivery.

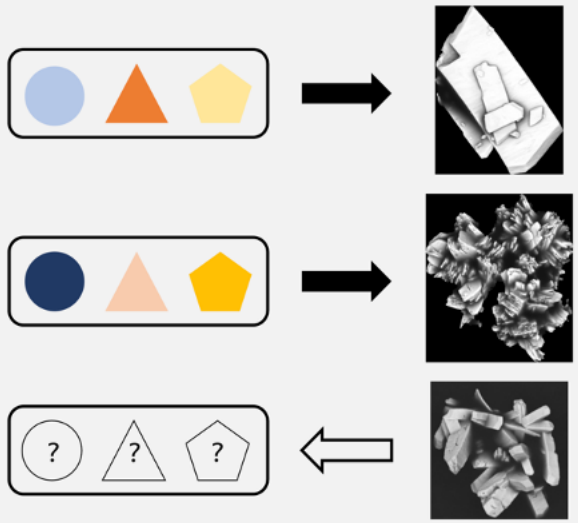
- Recognize the value AI could provide to achieve mission goals
- Evaluating the credibility of current techniques poses challenges that may hinder widespread acceptance and use
- Sandia's mission needs set us apart from industry and academia due to reasons such as high-consequence applications

Sandia's five major mission portfolios

The NNSA Labs must strike a balance between leveraging the advantages of ML while ensuring its responsible use for national security purposes.

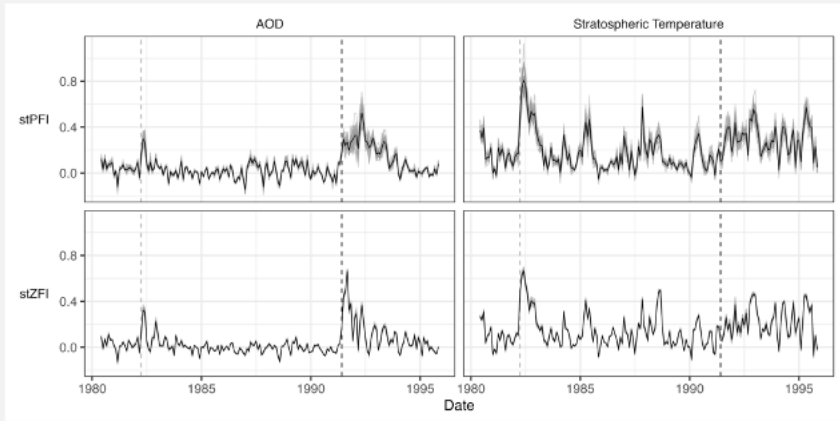
MACHINE LEARNING AT SANDIA: EXAMPLE PROJECTS

Predicting Processing Conditions of Nuclear Particulates

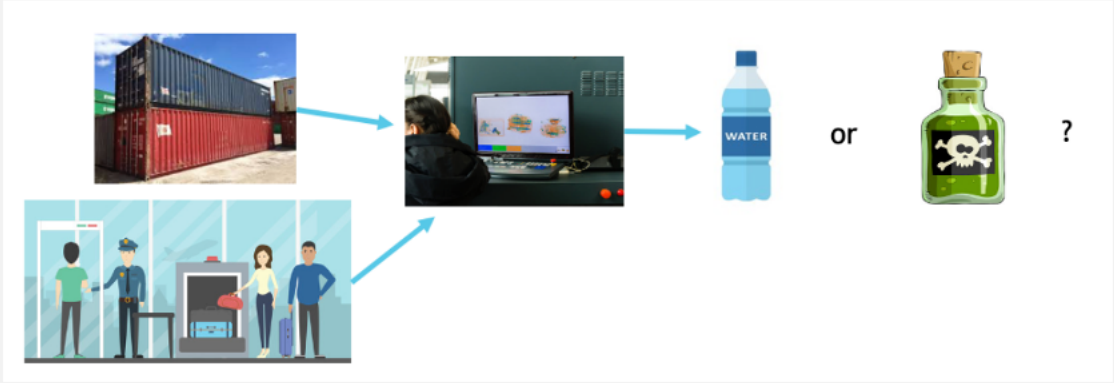


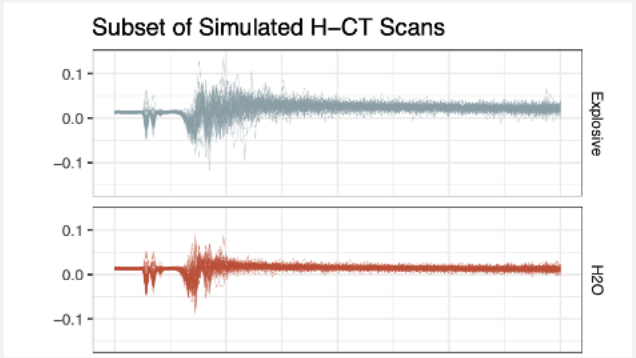
Characterizing Variable Pathways Associated with a Volcanic Eruption

Image source: https://pubs.usgs.gov/fs/1997/fs113-97/



Predicting Explosive Materials Using H-CT Scans





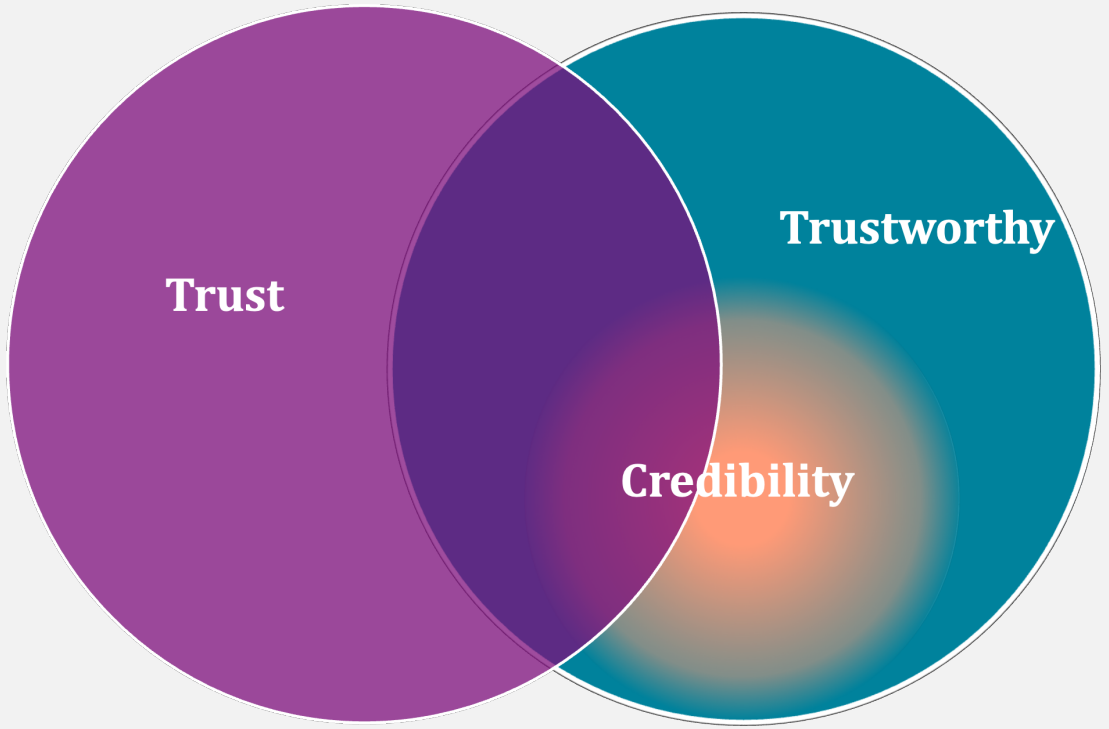
MACHINE LEARNING AT SANDIA: CONSIDERATIONS FOR TRUSTED AI

MACHINE LEARNING AT SANDIA: CONSIDERATIONS FOR TRUSTED AI

Trust defines the state of the decision maker.

Trustworthy defines the state of the model.

Credibility defines the technical basis of the model.



An example of a distinction in trust AI terminology at Sandia.

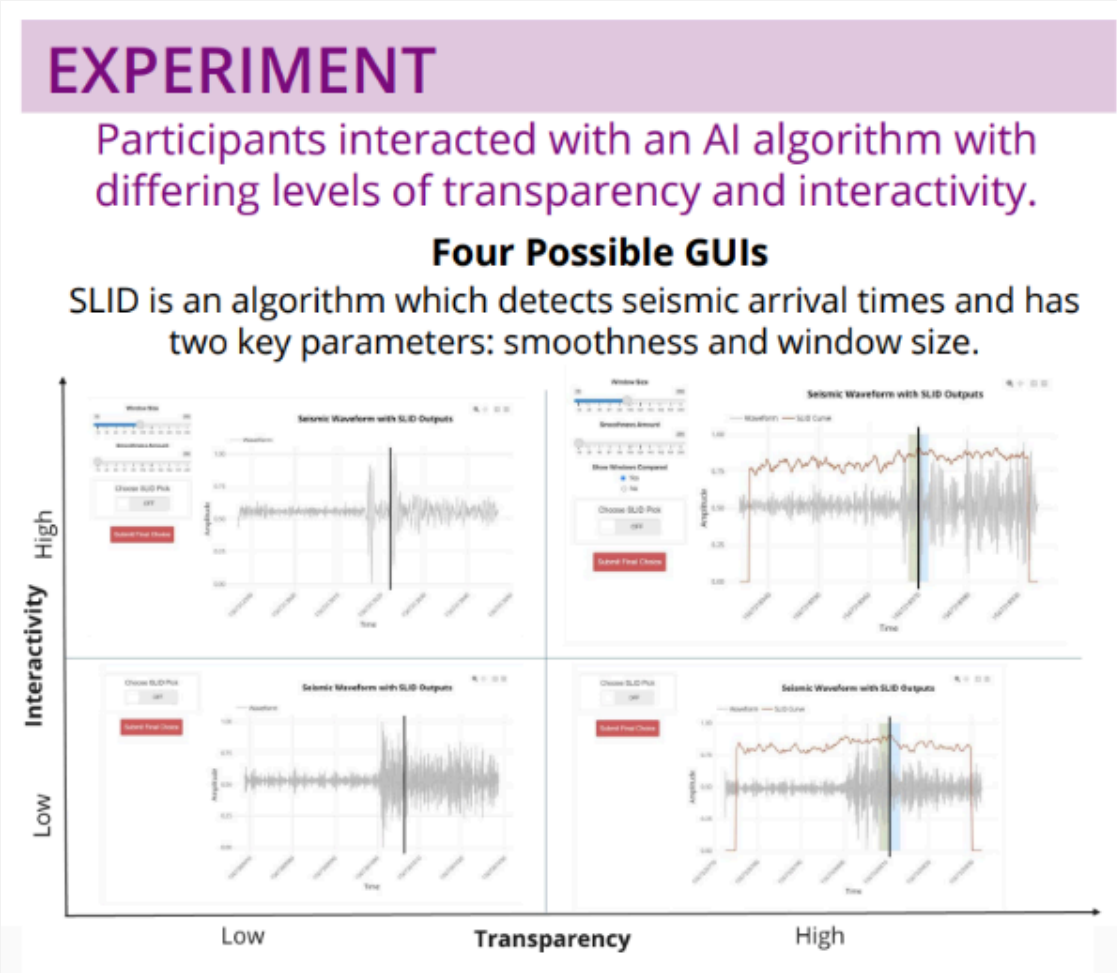
MACHINE LEARNING AT SANDIA: CONSIDERATIONS FOR TRUSTED AI

Trust defines the state of the decision maker.

- *Human studies investigating how transparency and interactivity affect users' trust and performance with an algorithm*

Trustworthy defines the state of the model.

Credibility defines the technical basis of the model.



MACHINE LEARNING AT SANDIA: CONSIDERATIONS FOR TRUSTED AI

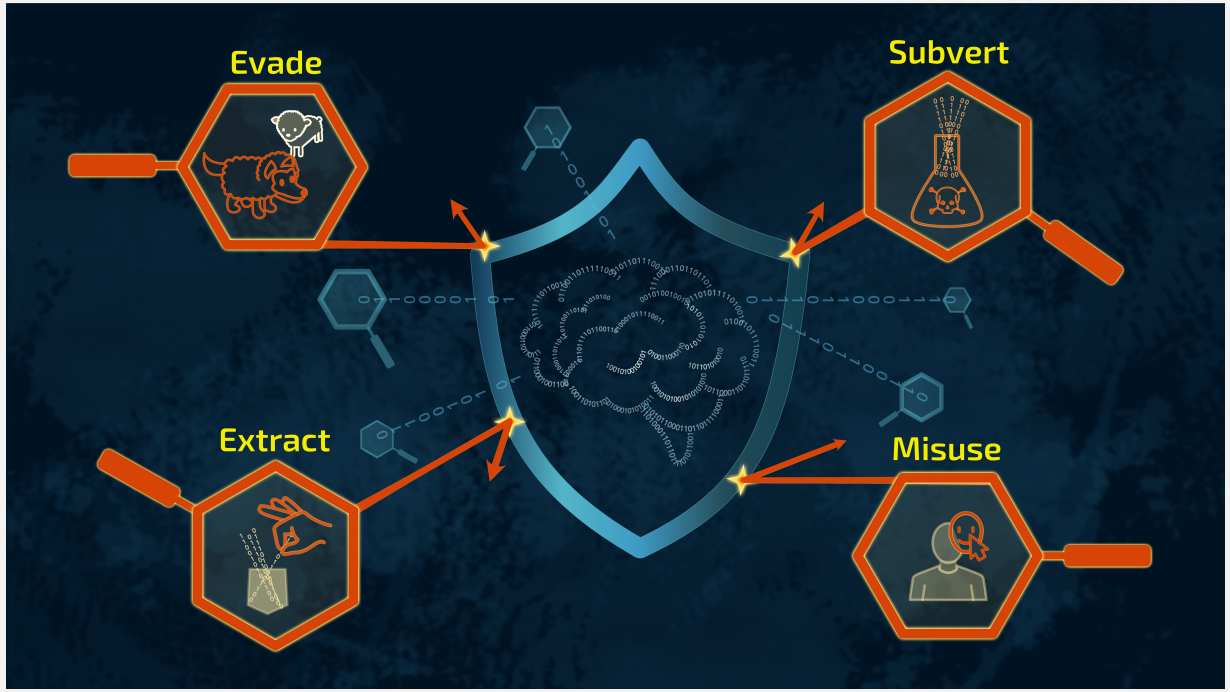
Trust defines the state of the decision maker.

- *Human studies investigating how transparency and interactivity affect users' trust and performance with an algorithm*

Trustworthy defines the state of the model.

- *Projects considering counter-adversarial ML*

Credibility defines the technical basis of the model.



MACHINE LEARNING AT SANDIA: CONSIDERATIONS FOR TRUSTED AI

Trust defines the state of the decision maker.

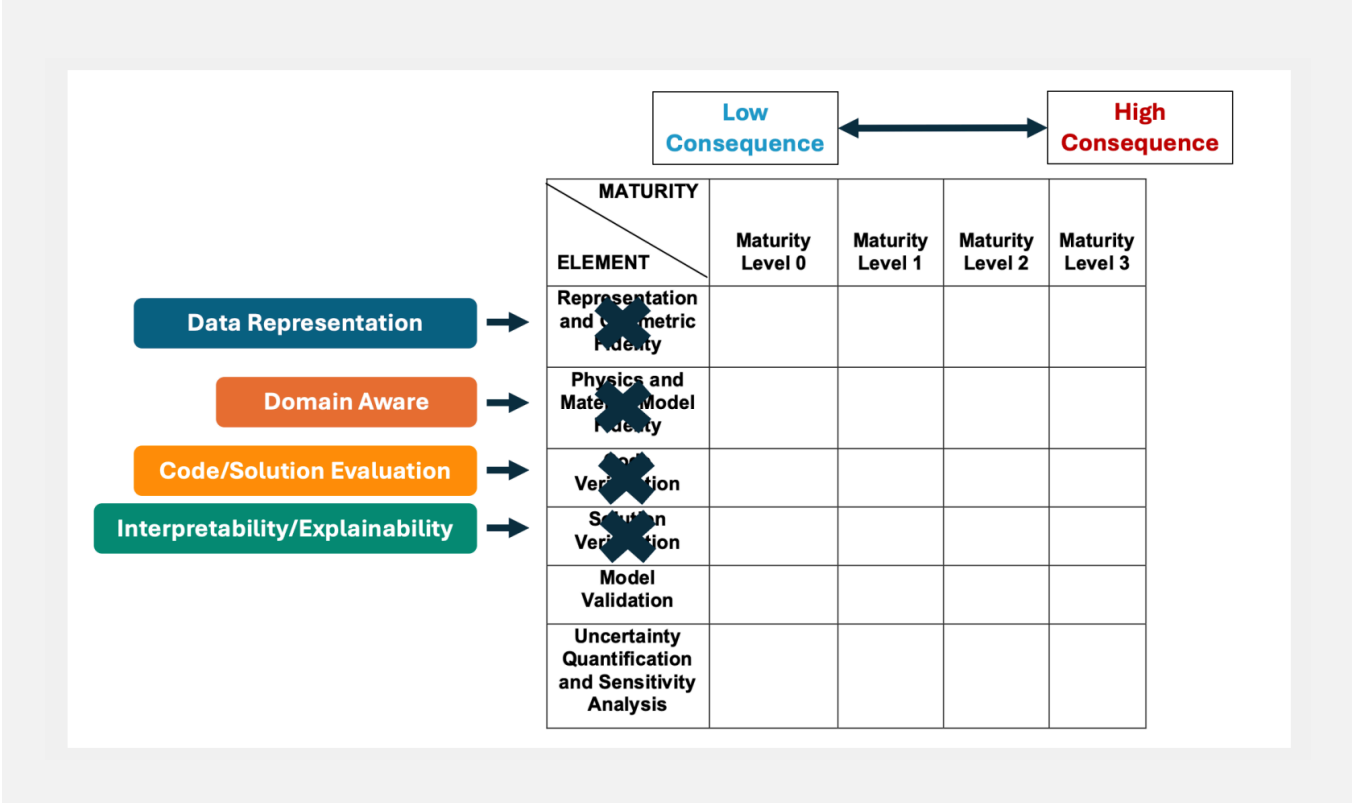
- *Human studies investigating how transparency and interactivity affect users' trust and performance with an algorithm*

Trustworthy defines the state of the model.

- *Projects considering counter-adversarial ML*

Credibility defines the technical basis of the model.

- *Investigations into uncertainty quantification methods for ML and technical evaluations of explainability methods*



MACHINE LEARNING AT SANDIA: ELEMENTS OF TRANSPARENCY

One area emphasized at Sandia is AI model transparency, which includes model interpretability and explainability

MACHINE LEARNING AT SANDIA: ELEMENTS OF TRANSPARENCY

One area emphasized at Sandia is AI model transparency, which includes model interpretability and explainability

Interpretability

Ability to directly use model to understand how algorithm makes decisions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Using interpretable model or adjusting black box models to contain interpretable parameters

MACHINE LEARNING AT SANDIA: ELEMENTS OF TRANSPARENCY

One area emphasized at Sandia is AI model transparency, which includes model interpretability and explainability

Interpretability

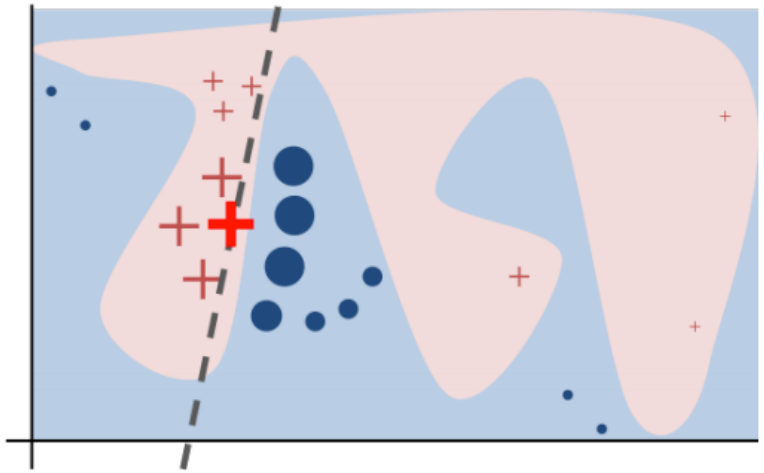
Ability to directly use model to understand how algorithm makes decisions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Using interpretable model or adjusting black box models to contain interpretable parameters

Explainability

Ability to indirectly use model to understand how algorithm makes decisions



Often post-hoc techniques

Figure from LIME paper (Ribeiro 2016)

FORESTR

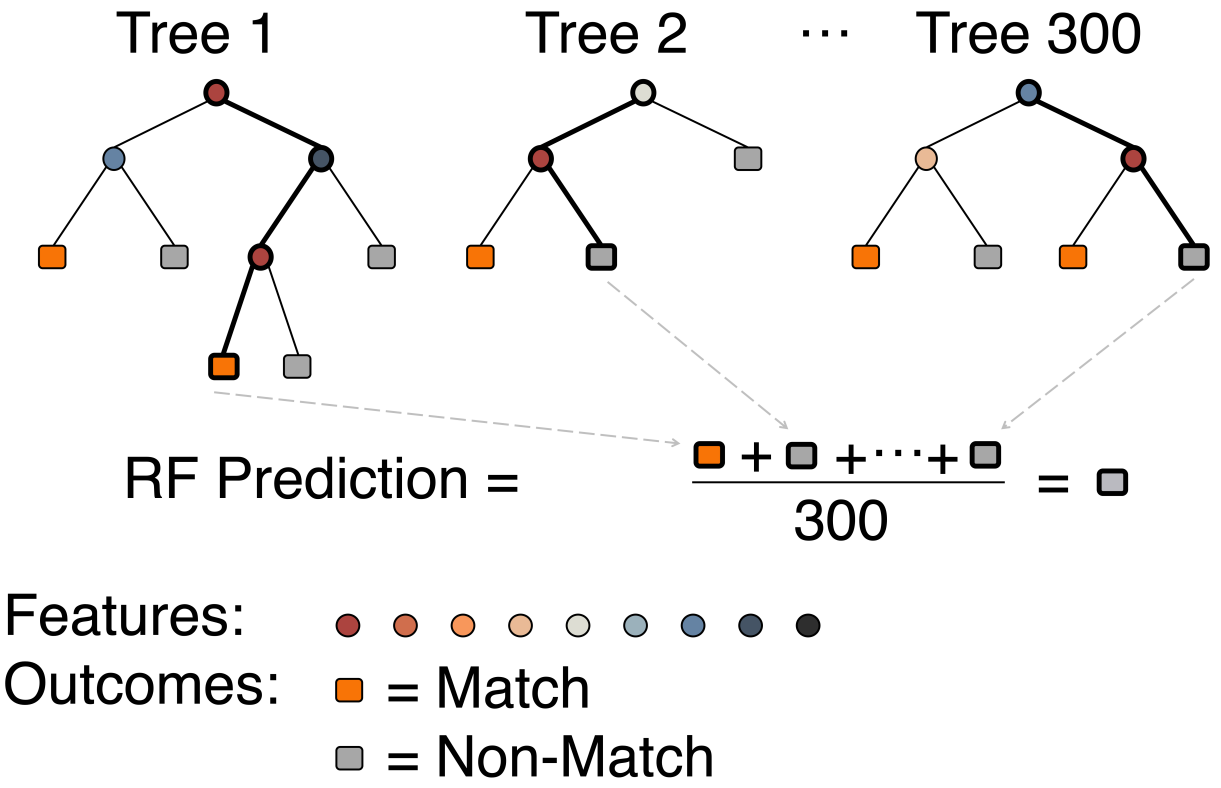
TOPOLOGY PATTERNS IN RANDOM FORESTS

MOTIVATION

MOTIVATION: RANDOM FORESTS

Random forests have become a common model used/considered in applications across Sandia

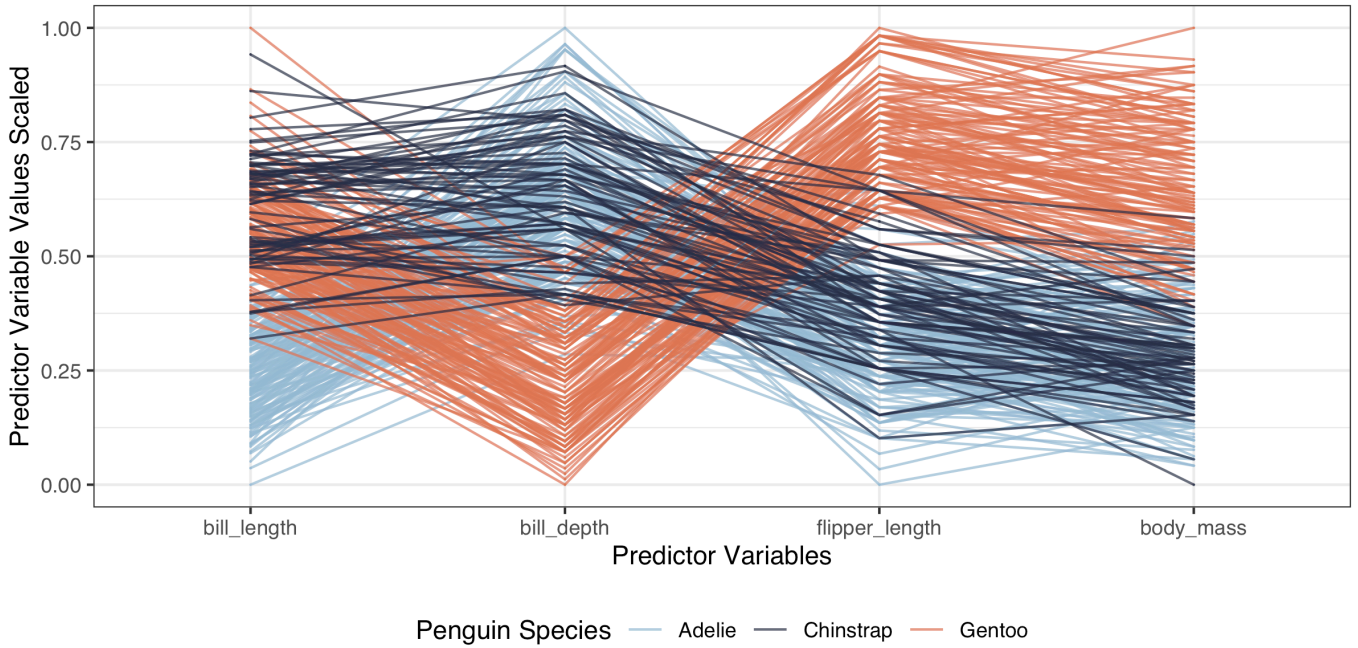
- Ensemble of decision/regression trees
- Introduces randomness in two places:
  - Each tree trained on a bootstrap sample of training data
  - Each split considers a random subset of features
- Advantages over a single tree:
  - Helps prevent overfitting
  - More robust to small variations in data

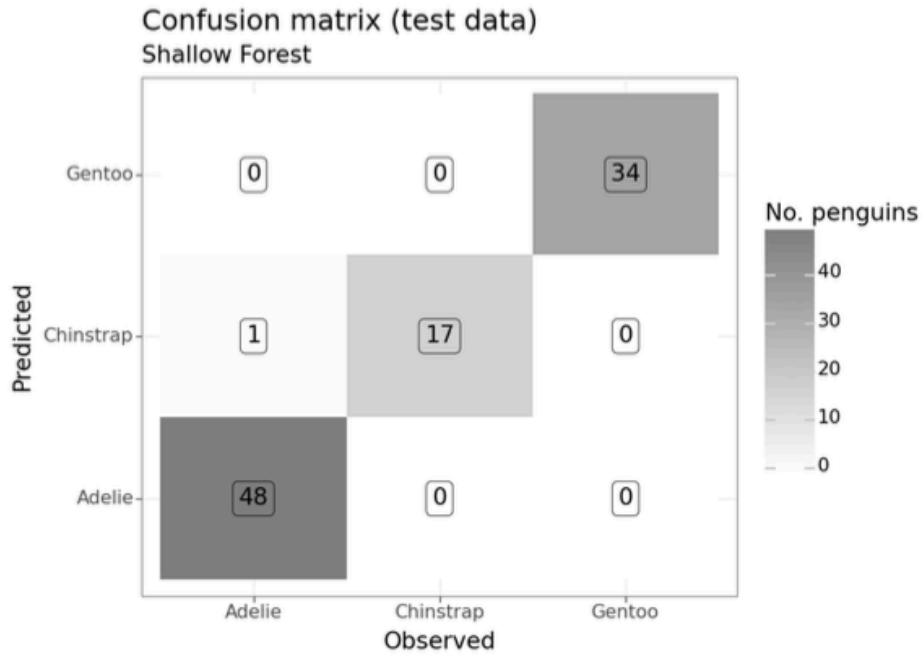


MOTIVATION: EXAMPLE DATA

Palmer Penguins

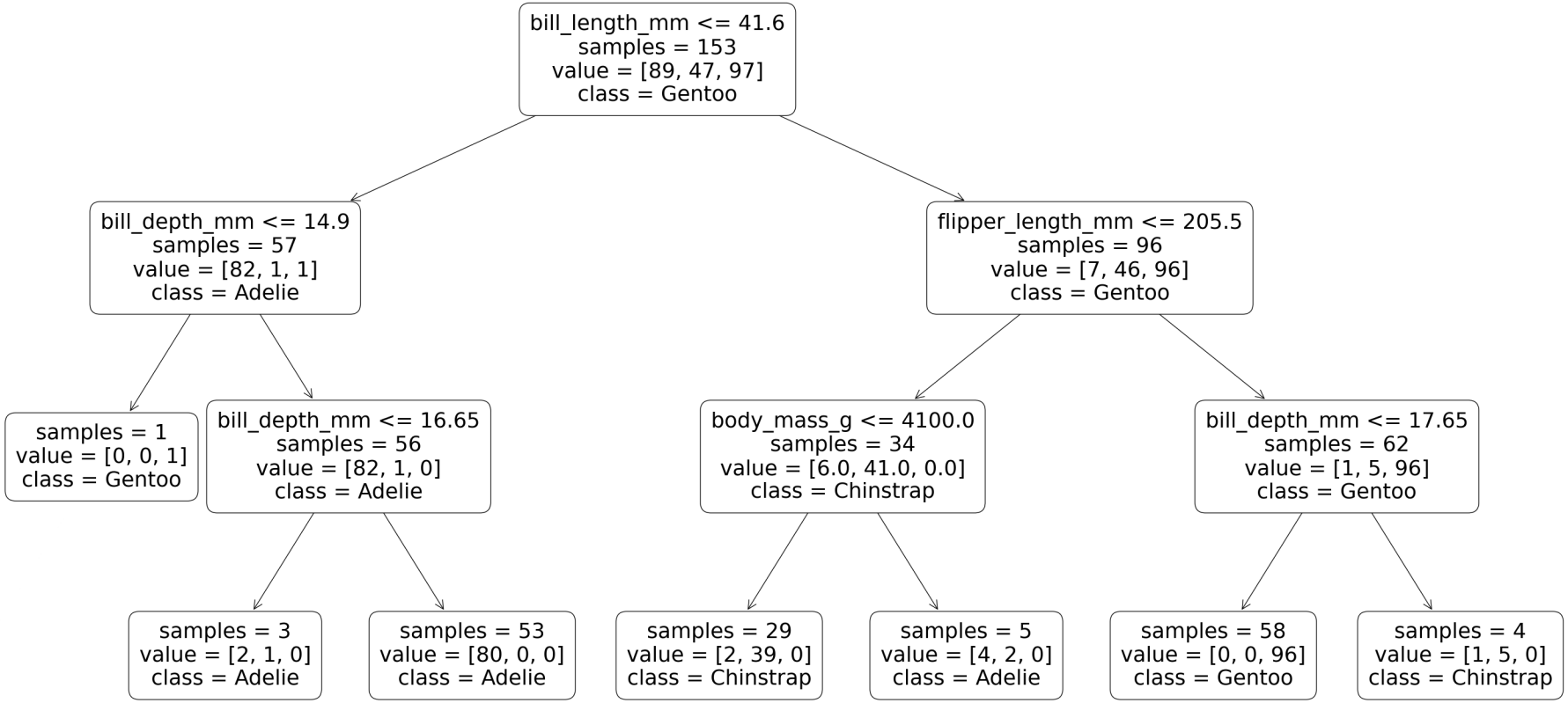
- **Data**: 342 penguins from Palmer Archipelago in Antarctica
- **Goal**: Predict species (Adelie, chinstrap, or gentoo)
  - Given four body measurements (bill length, bill depth, flipper length, body mass)
- **Model**: Random forests with 50 trees (max depth of 4 + full depth)





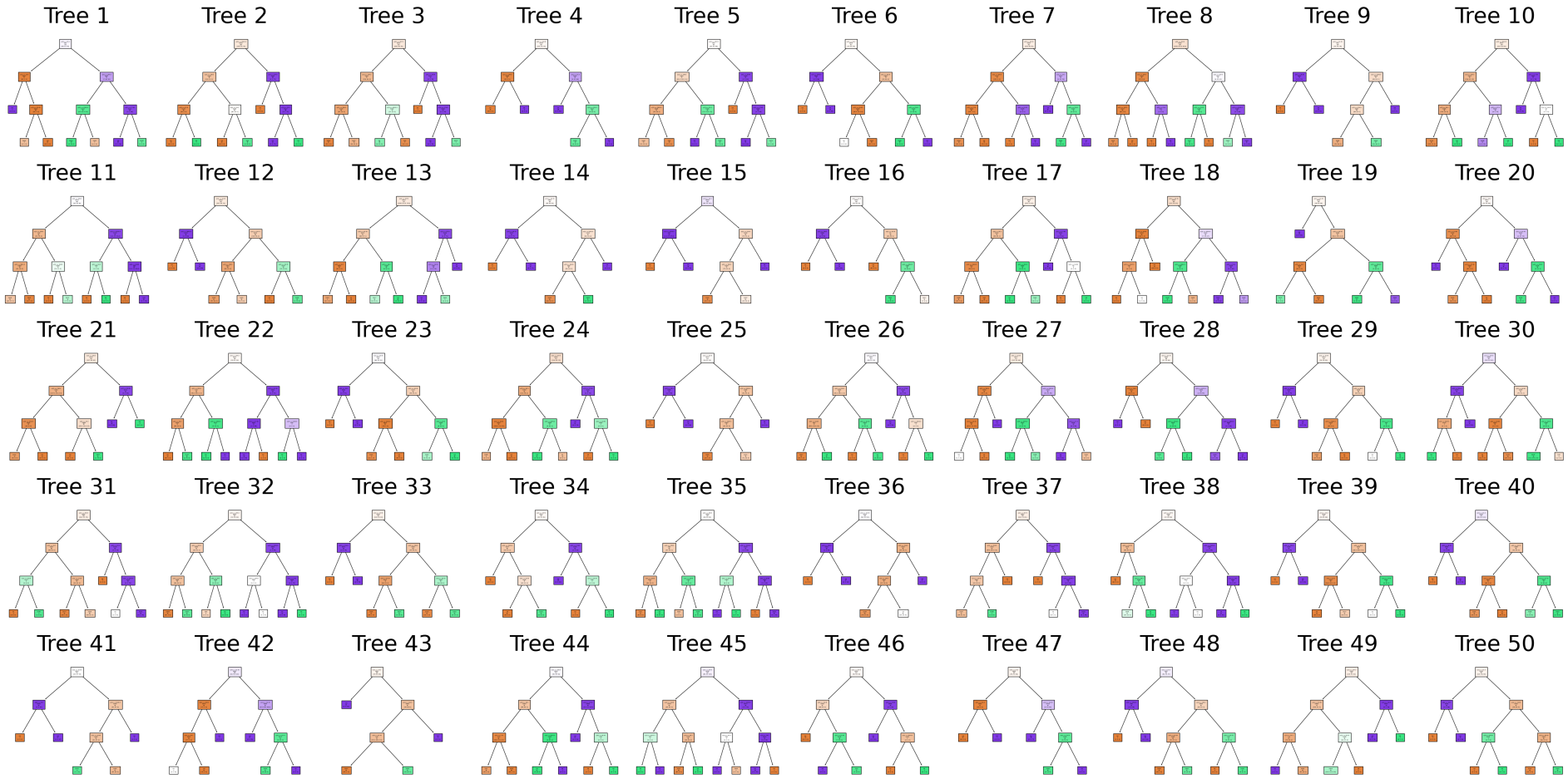
MOTIVATION: INTERPRETING TREE MODELS

- **Single tree** Typically considered interpretable



MOTIVATION: INTERPRETING TREE MODELS

- **Single tree** Typically considered interpretable
- **Ensemble of trees** Naturally becomes more difficult to interpret due to cognitive load



PREVIOUS WORK

PREVIOUS WORK: EXPLAINING RANDOM FORESTS

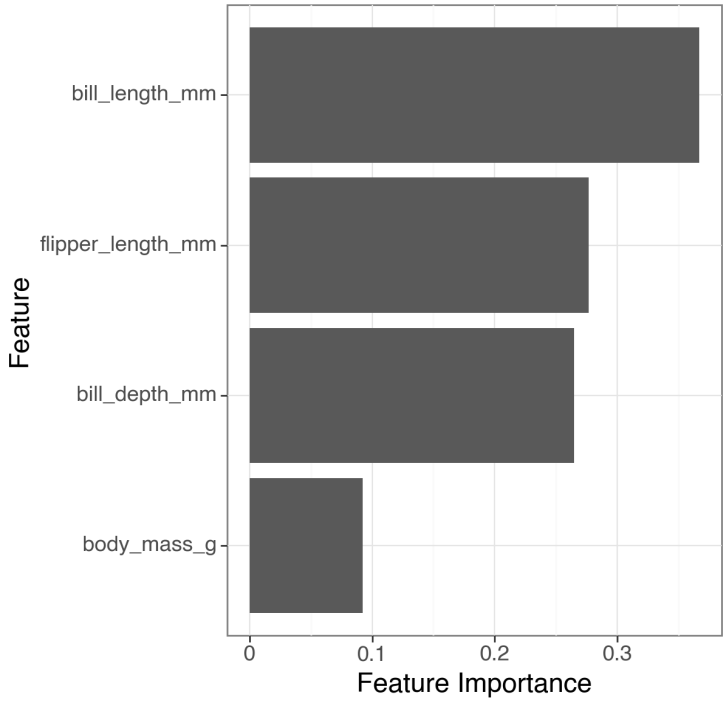
Various approaches have been developed for gaining insight into how random forests use data for predictions

PREVIOUS WORK: EXPLAINING RANDOM FORESTS

Various approaches have been developed for gaining insight into how random forests use data for predictions

Permutation Feature Importance (Breiman 2001)

Permute an input variable and quantify change in model performance



Terms:

- Model: f
- Data: $\mathbf{X}$ with n obs and p variables
- Variables: $X_1, X_2, \ldots, X_p$ columns of $\mathbf{X}$
- Metric: m computed with $\mathbf{X}$ and f (s.t. larger indicates better performance)

Procedure:

For variable $j \in \{1, \ldots, p\}$ and repetition $k \in \{1, \ldots, K\}$:

1. Create $\tilde{X}_{j,k}$ by randomly permuting X_j
2. Create $\tilde{\mathbf{X}}_{j,k}$ by replacing X_j with $\tilde{X}_{j,k}$ in $\mathbf{X}$
3. Compute $m_{j,k}$ with $\tilde{\mathbf{X}}_{j,k}$ and f

PFI for j: Average change in model performance when j is randomly permuted

$$\mathscr{I}_j = m - \frac{1}{K}\sum_{k=1}^{K} m_{j,k} = \frac{1}{K}\sum_{k=1}^{K}\left(m - m_{j,k}\right) = \frac{1}{K}\sum_{k=1}^{K}\mathscr{I}_{j,k}$$

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

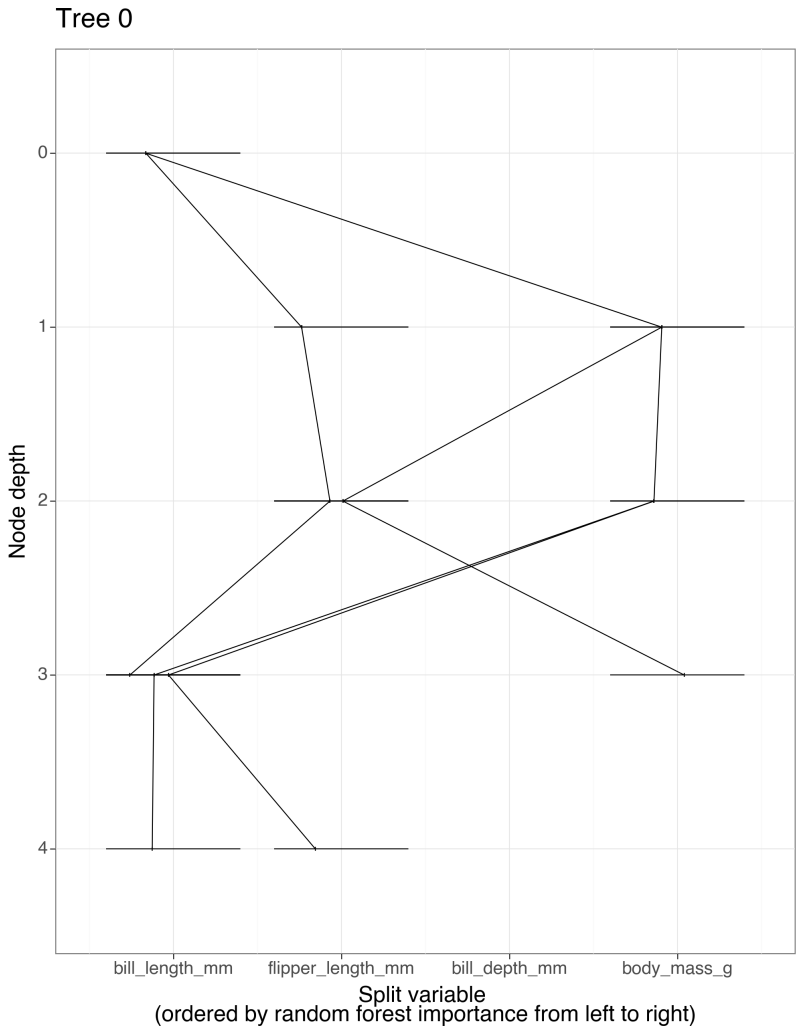
PREVIOUS WORK: EXPLAINING RANDOM FORESTS

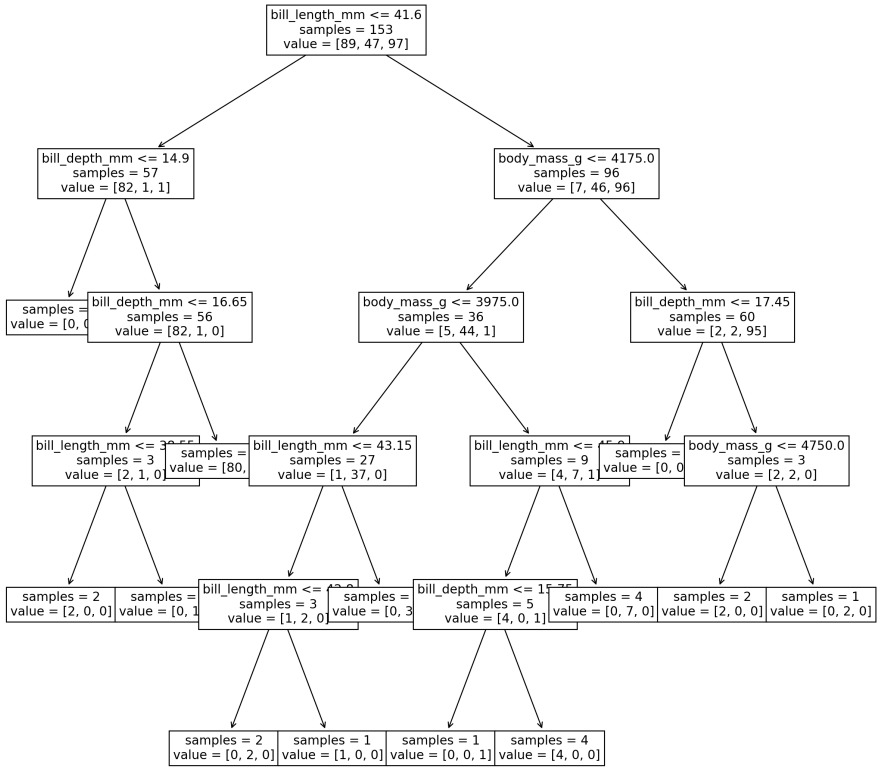
Various approaches have been developed for gaining insight into how random forests use data for predictions

Trace Plots (Urbanek 2010)

Designed to compare

1. variables used for splitting,
2. location of split points, and
3. hierarchical structure





Urbanek, S. (2008). Visualizing Trees and Forests. In: Handbook of Data Visualization. Springer Handbooks Comp.Statistics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_11

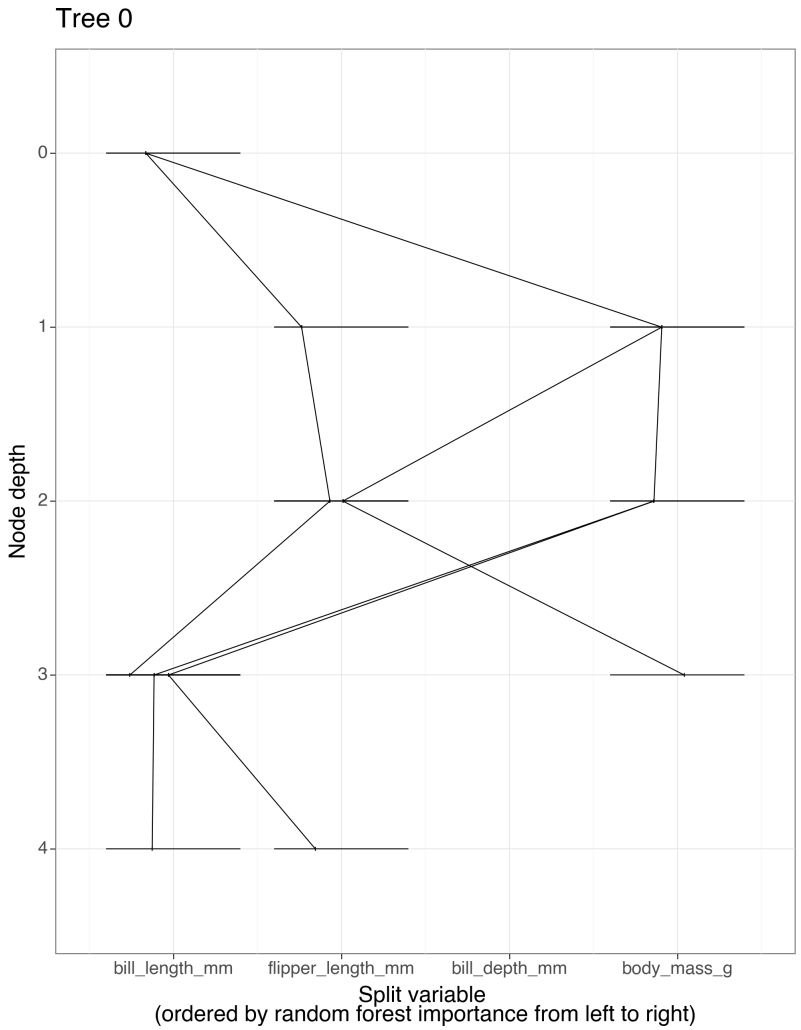
PREVIOUS WORK: EXPLAINING RANDOM FORESTS

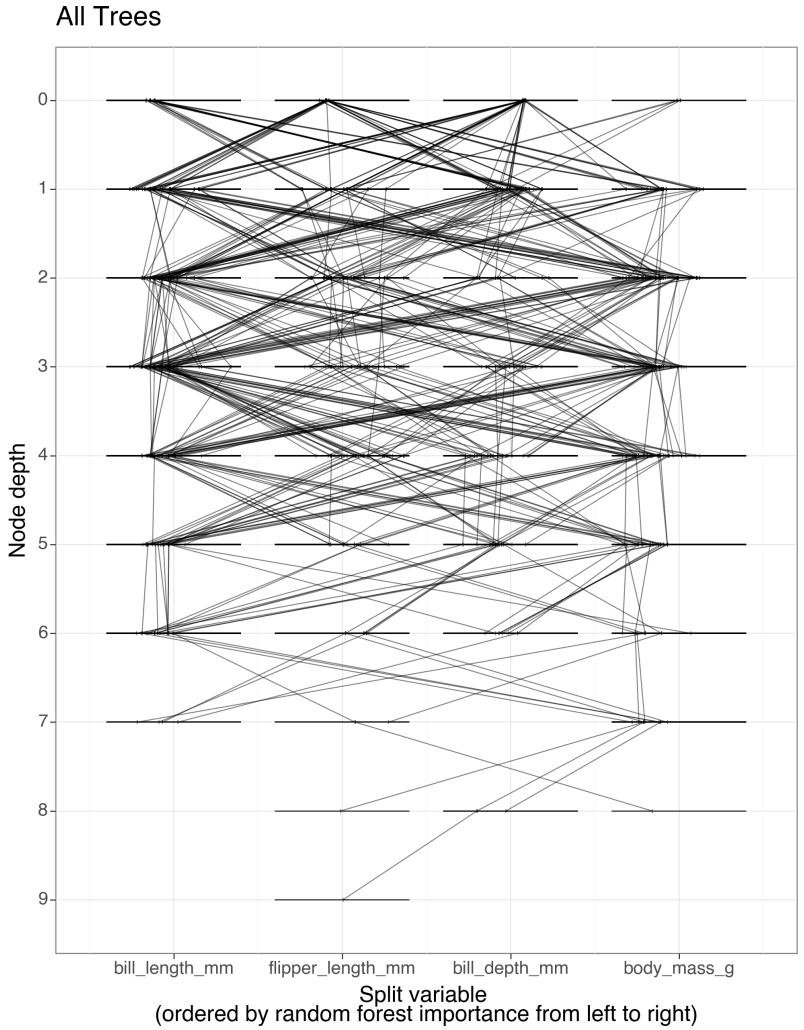
Various approaches have been developed for gaining insight into how random forests use data for predictions

Trace Plots (Urbanek 2010)

Designed to compare

1. variables used for splitting,
2. location of split points, and
3. hierarchical structure





Urbanek, S. (2008). Visualizing Trees and Forests. In: Handbook of Data Visualization. Springer Handbooks Comp.Statistics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_11

PREVIOUS WORK: EXPLAINING RANDOM FORESTS

Various approaches have been developed for gaining insight into how random forests use data for predictions

Representative tree

- Identify a tree that is representative of the forest
- One approach: Find tree that has smallest average distance to all other trees

Clusters of trees

- Compute distances between trees
- Identify clusters via MDS, K-means, etc.

Banerjee, M., Y. Ding, and A. Noone (2012). "Identifying representative trees from ensembles". In: *Statistics in Medicine* 31.15, pp. 1601-1616. ISSN: 1097-0258. 10.1002/sim.4492.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Making sense of a forest of trees". In: *Proceedings of the 30th Symposium on the Interface*. , pp. 84-92. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2598.

Shannon, W.D. and Banks, D. (1999), Combining classification trees using MLE. Statist. Med., 18: 727-740. https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6<727::AID-SIM61>3.0.CO;2-2

Sies, A. and I. V. Mechelen (2020). "C443: a Methodology to See a Forest for the Trees". In: *Journal of Classification* 37.3, pp. 730-753. ISSN: 0176-4268. https://link.springer.com/article/10.1007/s00357-019-09350-4.

Weinberg, A. I. and M. Last (2019). "Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification". In: *Journal of Big Data* 6.1, p. 23. 10.1186/s40537-019-0186-3.

PREVIOUS WORK: REPRESENTATIVE AND CLUSTERING OF TREES

PREVIOUS WORK: REPRESENTATIVE AND CLUSTERING OF TREES

Identifying a representative tree or clusters of trees both require a distance metric

Banerjee, M., Y. Ding, and A. Noone (2012). "Identifying representative trees from ensembles". In: *Statistics in Medicine* 31.15, pp. 1601-1616. ISSN: 1097-0258. 10.1002/sim.4492.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Making sense of a forest of trees". In: *Proceedings of the 30th Symposium on the Interface.* , pp. 84-92. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2598.

PREVIOUS WORK: REPRESENTATIVE AND CLUSTERING OF TREES

Identifying a representative tree or clusters of trees both require a distance metric

Two approaches:

Comparing Predictions

Comparing Topology

Banerjee, M., Y. Ding, and A. Noone (2012). "Identifying representative trees from ensembles". In: *Statistics in Medicine* 31.15, pp. 1601-1616. ISSN: 1097-0258. 10.1002/sim.4492.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Making sense of a forest of trees". In: *Proceedings of the 30th Symposium on the Interface.* , pp. 84-92. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2598.

PREVIOUS WORK: REPRESENTATIVE AND CLUSTERING OF TREES

Identifying a representative tree or clusters of trees both require a distance metric

Two approaches:

Comparing Predictions

Fit metric Compares predictions from two trees (Chipman, George, and McCulloch (1998)

$$d_{FM}\left(T_1, T_2\right) = \frac{1}{n}\sum_{i=1}^{n} m\left(\hat{y}_{i1}, \hat{y}_{i2}\right)$$

where T_t is tree t with $t \in \{1, 2\}$, y_i is response, $\hat{y}_{it}$ is fitted value, and m is a metric such as

- Regression metric: $m\left(\hat{y}_{i1}, \hat{y}_{i2}\right) = \left(\hat{y}_{i1} - \hat{y}_{i2}\right)^2$
- Classification metric: $m\left(\hat{y}_{i1}, \hat{y}_{i2}\right) = \begin{cases} 1 & \text{if } \hat{y}_{i1} \neq \hat{y}_{i2} \\ 0 & \text{o.w.} \end{cases}$

Comparing Topology

Banerjee, M., Y. Ding, and A. Noone (2012). "Identifying representative trees from ensembles". In: *Statistics in Medicine* 31.15, pp. 1601-1616. ISSN: 1097-0258. 10.1002/sim.4492.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Making sense of a forest of trees". In: *Proceedings of the 30th Symposium on the Interface.* , pp. 84-92. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2598.

PREVIOUS WORK: REPRESENTATIVE AND CLUSTERING OF TREES

Identifying a representative tree or clusters of trees both require a distance metric

Two approaches:

Comparing Predictions

Fit metric Compares predictions from two trees (Chipman, George, and McCulloch (1998)

$$d_{FM}\left(T_1, T_2\right) = \frac{1}{n}\sum_{i=1}^{n} m\left(\hat{y}_{i1}, \hat{y}_{i2}\right)$$

Comparing Topology

Covariate metric Compares split variables from two trees (Banerjee, Ding, and Noone (2012))

$$d_{CM}(T_1, T_2) = \frac{\text{Number of covariate mismatches for } T_1 \text{ and } T_2}{k}.$$

where T_t is tree t with $t \in \{1, 2\}$, y_i is response, $\hat{y}_{it}$ is fitted value, and m is a metric such as

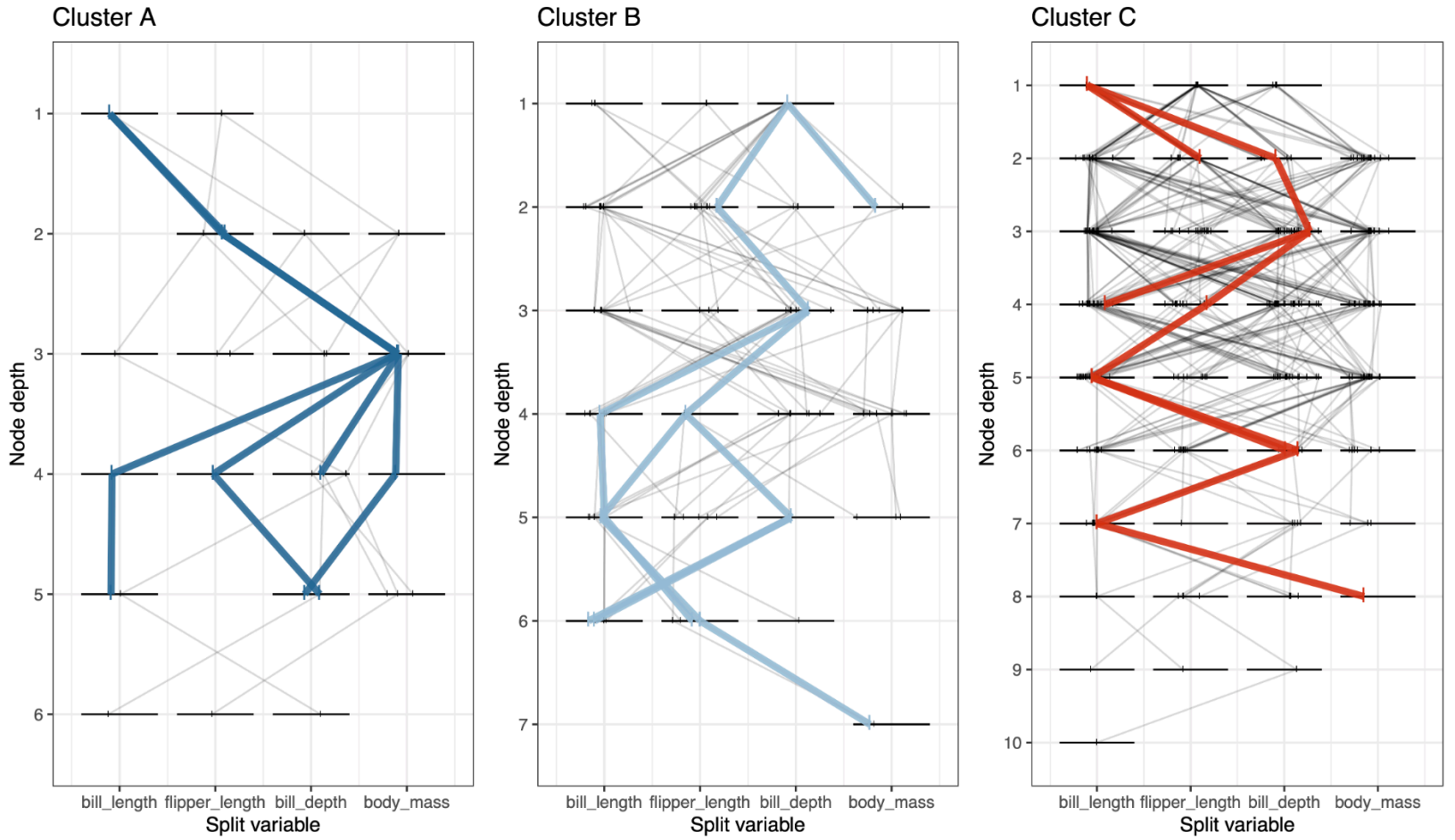
- Regression metric: $m\left(\hat{y}_{i1}, \hat{y}_{i2}\right) = \left(\hat{y}_{i1} - \hat{y}_{i2}\right)^2$
- Classification metric: $m\left(\hat{y}_{i1}, \hat{y}_{i2}\right) = \begin{cases} 1 & \text{if } \hat{y}_{i1} \neq \hat{y}_{i2} \\ 0 & \text{o.w.} \end{cases}$

Banerjee, M., Y. Ding, and A. Noone (2012). "Identifying representative trees from ensembles". In: *Statistics in Medicine* 31.15, pp. 1601-1616. ISSN: 1097-0258. 10.1002/sim.4492.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Making sense of a forest of trees". In: *Proceedings of the 30th Symposium on the Interface.* , pp. 84-92. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.2598.

PREVIOUS WORK: REPRESENTATIVE AND CLUSTERING OF TREES

Example of random forest tree clusters and corresponding central trees



Clusters identified using multidimensional scaling with fit metric and representative trees determined by tree with smallest average fit metric distance to all other trees in cluster.

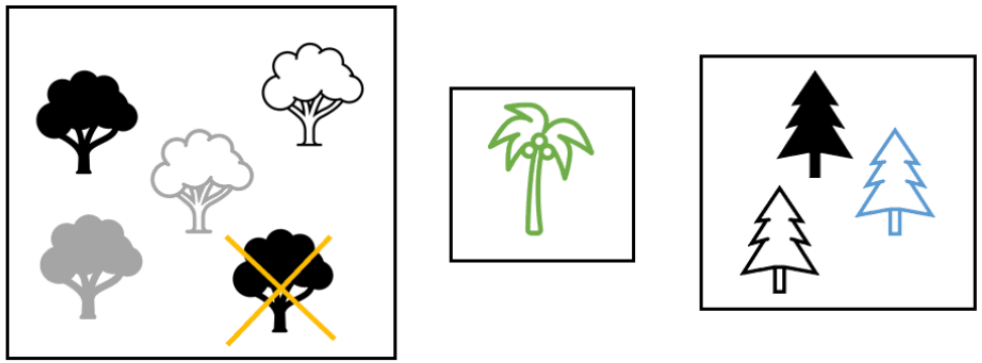
FORESTR

FORESTR: PROJECT GOALS

FORESTR: Finding, **O**rganizing, **R**epresenting, **E**xplaining, **S**ummarizing, and **T**hinning **R**andom forests

Overview

- Use graph topology distance metric to identify tree topology patters patterns in forests
- Would like a distance metric for comparing tree topologies that is a *proper mathematical distance* to allow for the *computation of summary statistics*
- Future work could consider using metric to create ensemble with a reduced number of trees



FORESTR: TREE DISTANCE METRIC

Guo, X., Srivastava, A. & Sarkar, S. A Quotient Space Formulation for Generative Statistical Analysis of Graphical Data. *J Math Imaging Vis* 63, 735–752 (2021). https://doi.org/10.1007/s10851-021-01027-1
Bal, Aditi Basu, et al. "Statistical Shape Analysis of Shape Graphs with Applications to Retinal Blood-Vessel Networks." *arXiv preprint arXiv:2211.15514* (2022).

FORESTR: TREE DISTANCE METRIC

Adapt graph metric from Guo (2021) to trees, which has nice properties:

Proper mathematical distance, invariant to transformations, registration between graphs, and computes topological evolutions between graphs

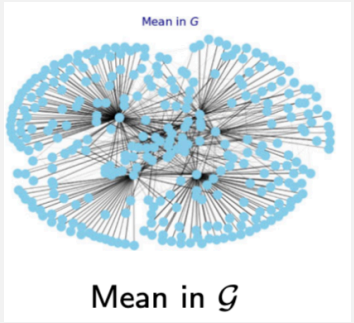
A Quotient Space Formulation for Generative Statistical Analysis of Graphical Data

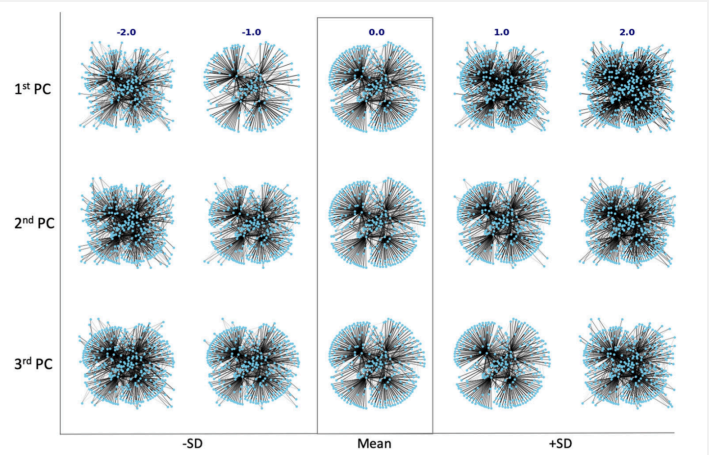
Xiaoyang Guo[1] · Anuj Srivastava[1] · Sudeep Sarkar[2]

Received: 19 May 2020 / Accepted: 9 March 2021 / Published online: 31 March 2021

Definition 1 (*Graph Metric*) Define a metric on the graph space $\mathcal{G}$ according to:

$$d_g([A_1], [A_2]) = \min_{P \in \mathcal{P}} d_a(A_1, PA_2P^T) = \min_{P \in \mathcal{P}} d_a(A_2, PA_1P^T) \quad (4)$$



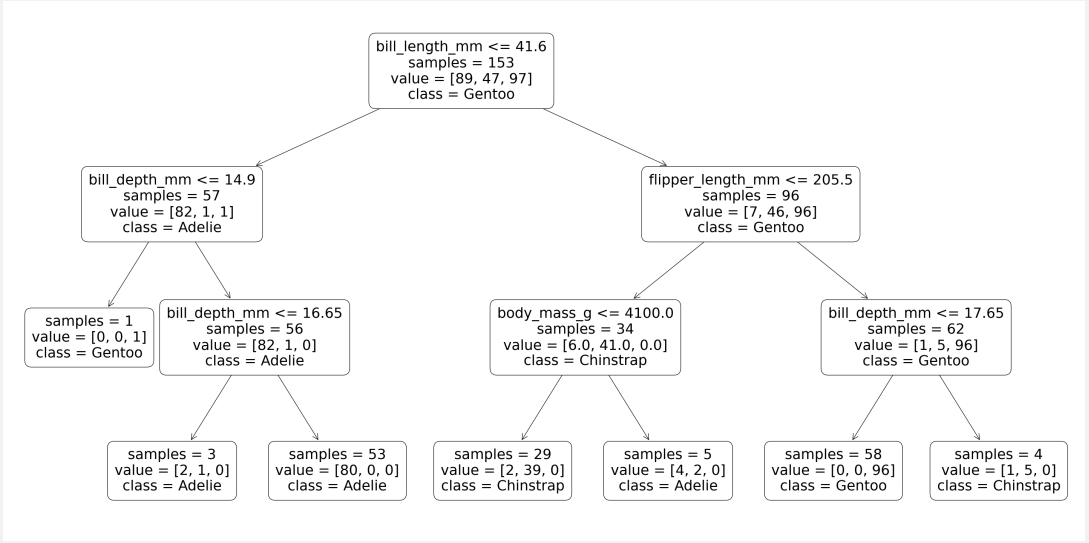


Guo, X., Srivastava, A. & Sarkar, S. A Quotient Space Formulation for Generative Statistical Analysis of Graphical Data. *J Math Imaging Vis* 63, 735–752 (2021). https://doi.org/10.1007/s10851-021-01027-1
Bal, Aditi Basu, et al. "Statistical Shape Analysis of Shape Graphs with Applications to Retinal Blood-Vessel Networks." *arXiv preprint arXiv:2211.15514* (2022).

FORESTR: TREE DISTANCE METRIC

For an ensemble of trees $t = 1, \ldots, T$, let

- G_t = weighted graph associated with tree t
- (V_t, e_t) = nodes and edges associated with G_t

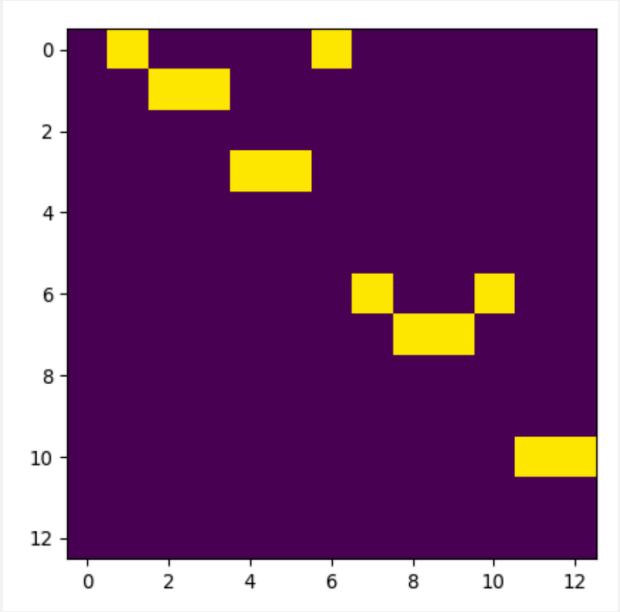


A tree in penguin random forest

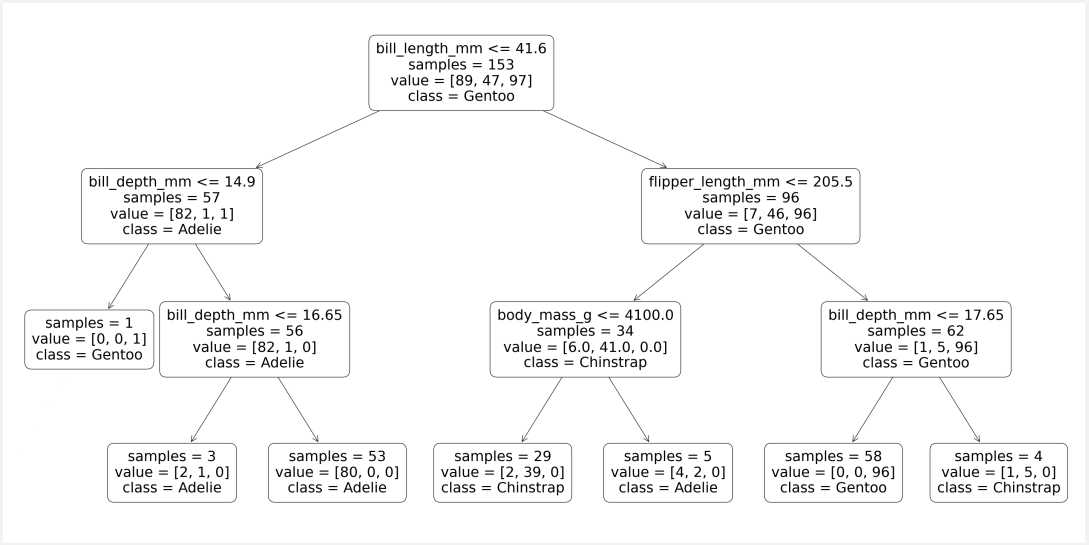
FORESTR: TREE DISTANCE METRIC

For an ensemble of trees $t = 1, \ldots, T$, let

- G_t = weighted graph associated with tree t
- (V_t, e_t) = nodes and edges associated with G_t
- A_t = adjacency matrix associated with tree t



Corresponding adjacency matrix

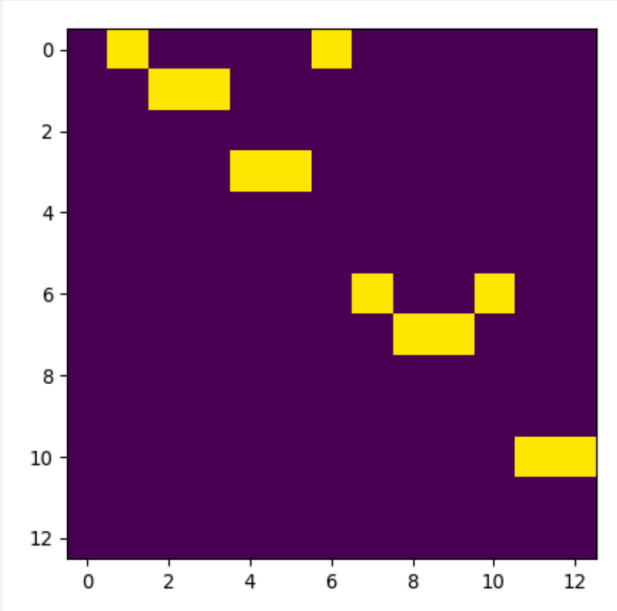


A tree in penguin random forest

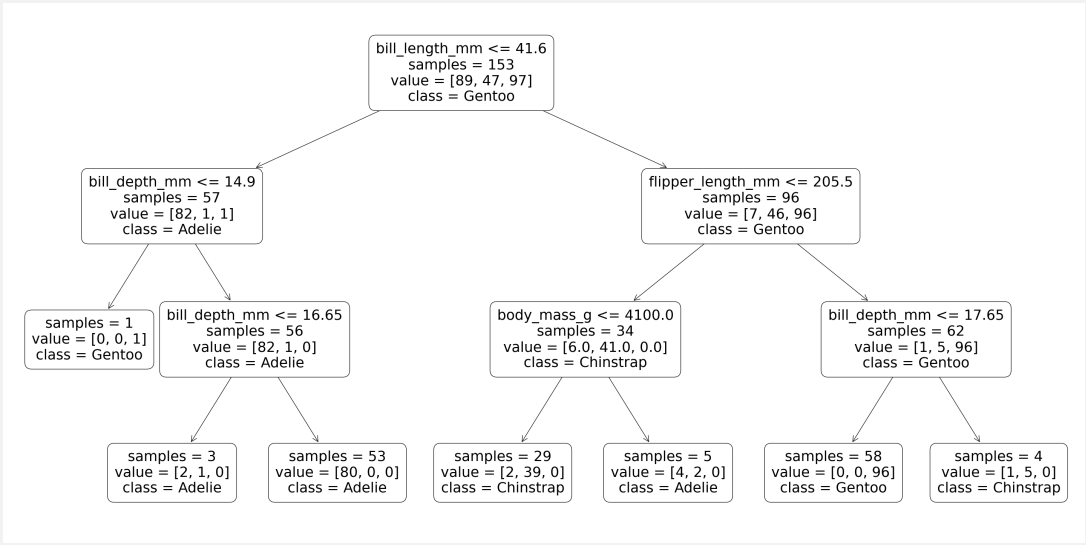
FORESTR: TREE DISTANCE METRIC

For an ensemble of trees $t = 1, \ldots, T$, let

- G_t = weighted graph associated with tree t
- (V_t, e_t) = nodes and edges associated with G_t
- A_t = adjacency matrix associated with tree t



Corresponding adjacency matrix



A tree in penguin random forest

GOAL For all pairs of trees t and u, want to compute distance between trees:

$$d_g([A_t], [A_u])$$

FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

$$d_g([A_t],[A_u]) = \min_{P \in \mathcal{P}} \|PA_tP^T - A_u\|^2 + \lambda \mathrm{Tr}(PD_{t,u})$$

FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

$$d_g([A_t],[A_u]) = \min_{P \in \mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \text{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$

where the optimization is implemented using Umeyama algorithm or fast approximate quadratic programming and

- P = permutation matrix
- $\mathcal{P}$ = set of all permutation matrices of dimension $n \times n$
- λ = tuning parameter specifying weight to place on attributes
- $D_{t,u}$ = distance between node attributes of graphs t and u

FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

$$d_g([A_t],[A_u]) = \min_{P \in \mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \text{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$

where the optimization is implemented using Umeyama algorithm or fast approximate quadratic programming and

- P = permutation matrix
- $\mathcal{P}$ = set of all permutation matrices of dimension $n \times n$
- λ = tuning parameter specifying weight to place on attributes
- $D_{t,u}$ = distance between node attributes of graphs t and u

For trees G_t and G_u,

$$D_{t,u} = [d_{ij} = d(\alpha_t(v_i^t), \alpha_u(v_i^u))] \in \mathbb{R}^{n \times n},$$

where $\alpha_t(v_i^t)$ represents a node attribute and d is some distance metric.

FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

$$d_g([A_t], [A_u]) = \min_{P \in \mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \text{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$

where the optimization is implemented using Umeyama algorithm or fast approximate quadratic programming and

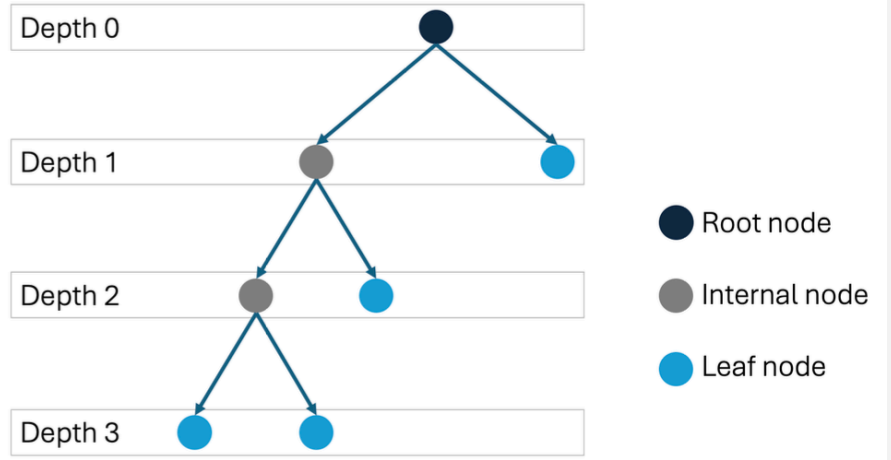
- P = permutation matrix
- $\mathcal{P}$ = set of all permutation matrices of dimension $n \times n$
- λ = tuning parameter specifying weight to place on attributes
- $D_{t,u}$ = distance between node attributes of graphs t and u

For trees G_t and G_u,

$$D_{t,u} = [d_{ij} = d(\alpha_t(v_i^t), \alpha_u(v_i^u))] \in \mathbb{R}^{n \times n},$$

where $\alpha_t(v_i^t)$ represents a node attribute and d is some distance metric.

To account for the hierarchical nature of trees, we consider node depth as an attribute.



FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

$$d_g([A_t],[A_u]) = \min_{P\in\mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \mathrm{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$

where the optimization is implemented using Umeyama algorithm or fast approximate quadratic programming and

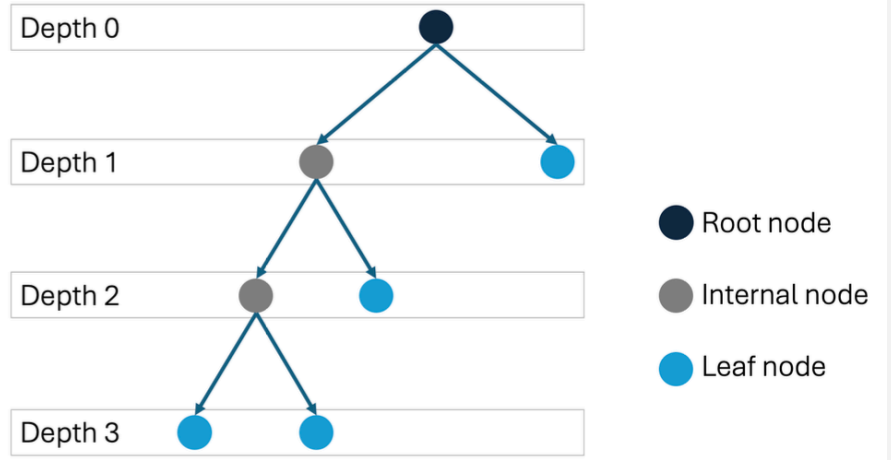
- P = permutation matrix
- $\mathcal{P}$ = set of all permutation matrices of dimension $n \times n$
- λ = tuning parameter specifying weight to place on attributes
- $D_{t,u}$ = distance between node attributes of graphs t and u

For trees G_t and G_u,

$$D_{t,u} = [d_{ij} = (\alpha_t(v_i^t) - \alpha_u(v_i^u))^2] \in \mathbb{R}^{n\times n},$$

where $\alpha_t(v_i^t)$ represents the node depth associated with node i in tree t.

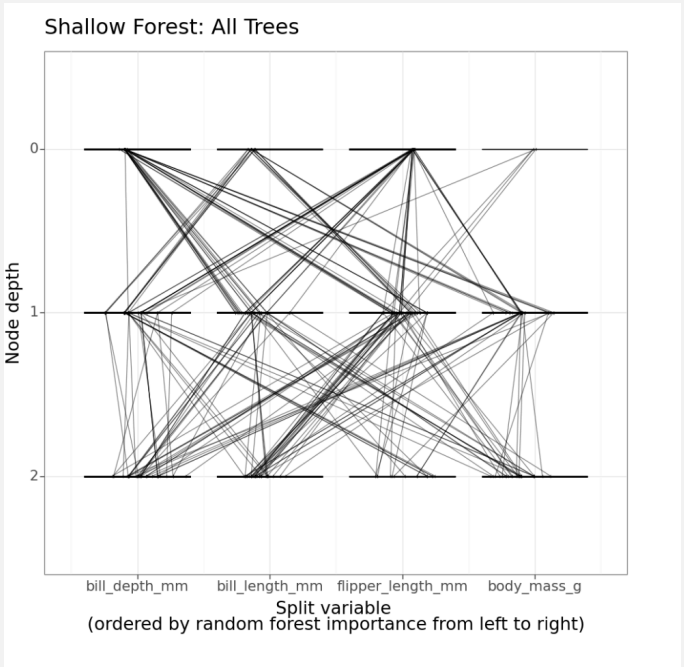
To account for the hierarchical nature of trees, we consider node depth as an attribute.

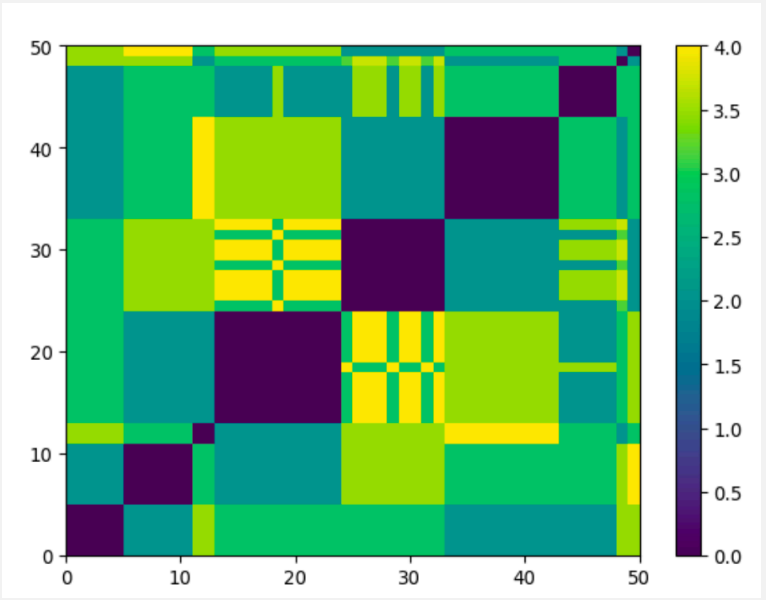


FORESTR: TREE DISTANCE METRIC

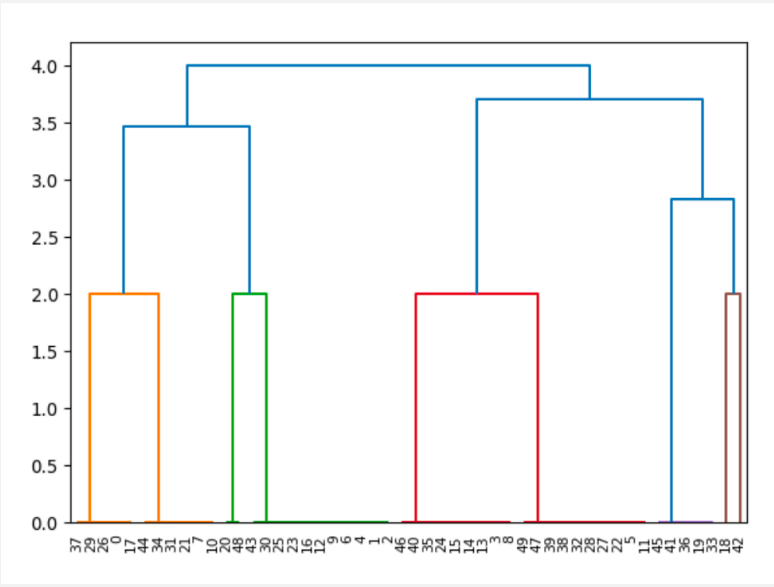
Compute the distance between trees G_t and G_u as

$$d_g([A_t],[A_u]) = \min_{P\in\mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \mathrm{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$





All pairwise distances

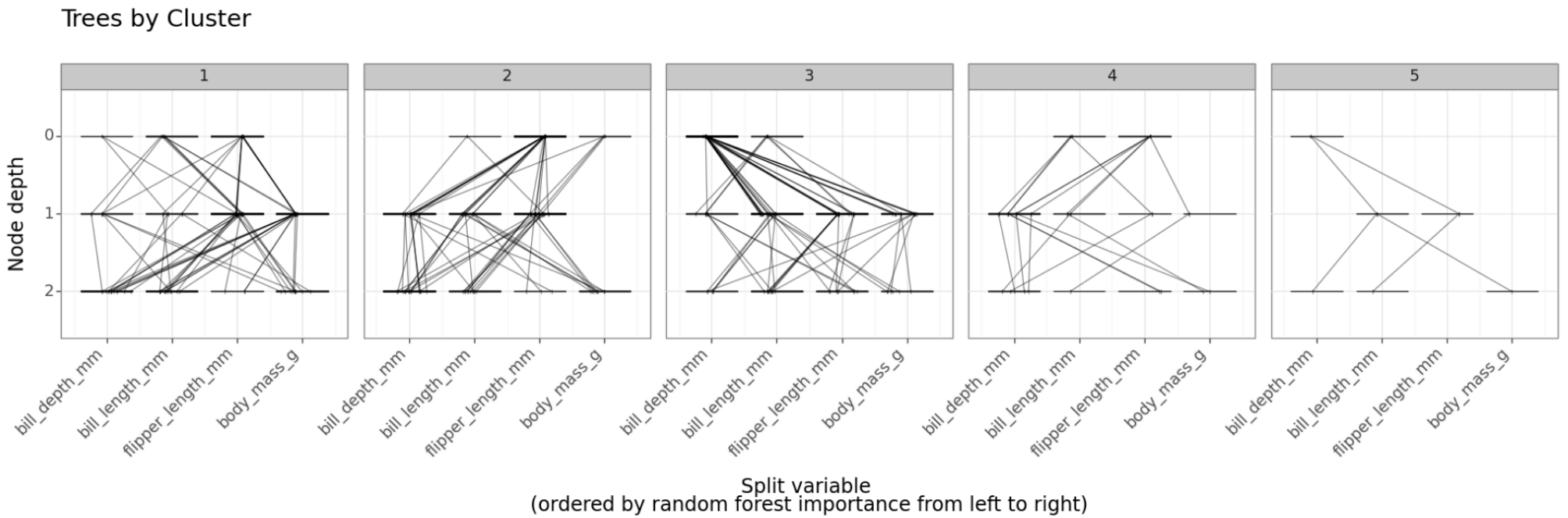


Dendrogram of complete linkage clusters based on pairwise distances

FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

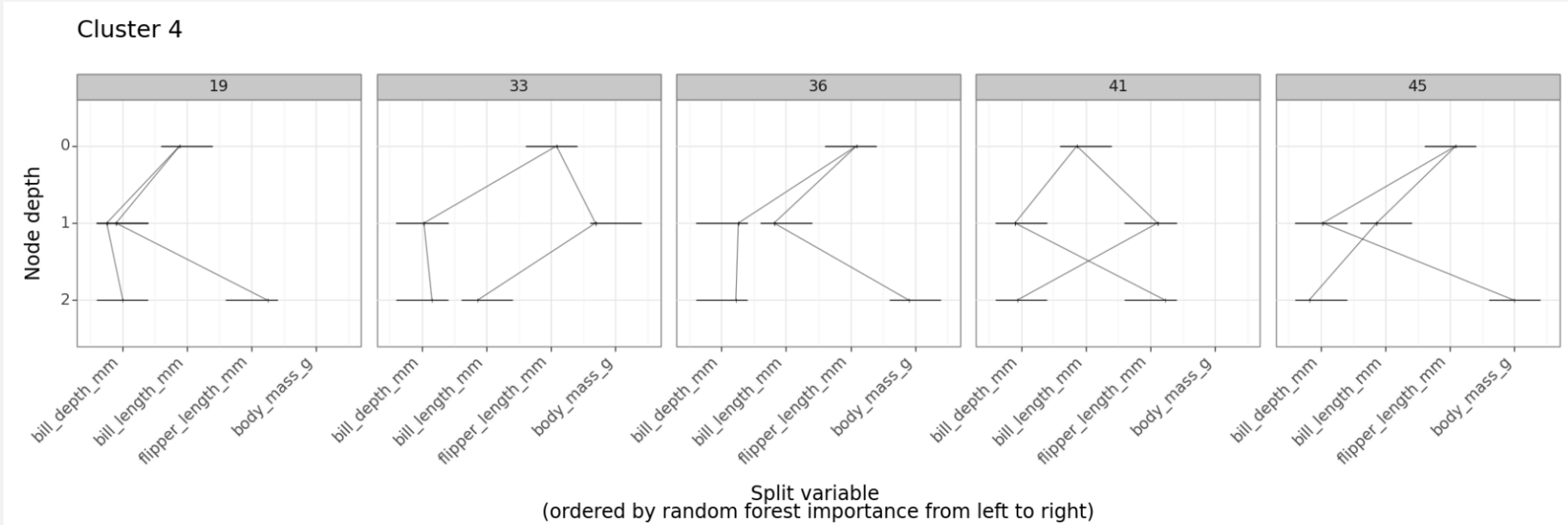
$$d_g([A_t],[A_u]) = \min_{P\in\mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \mathrm{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$



FORESTR: TREE DISTANCE METRIC

Compute the distance between trees G_t and G_u as

$$d_g([A_t],[A_u]) = \min_{P\in\mathcal{P}} \underbrace{\|PA_tP^T - A_u\|^2}_{\text{Accounts for edges}} + \underbrace{\lambda \mathrm{Tr}(PD_{t,u})}_{\text{Accounts for nodes}}$$



FORESTR: CENTRAL TREE

FORESTR: CENTRAL TREE

Given a set of T graphs, denoted as $G_i \in \mathcal{G}, t = 1, \ldots, m$, and their respective adjacency matrices, $A_t \in \mathbb{R}^{n \times n}$, the mean graph is obtained by minimizing the sum of squared distances:

$$[A_\mu] = \underset{A \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} \sum_{t=1}^{T} d_g([A], [A_i])^2$$

The mean graph is guaranteed to be unique, and the optimization is performed using a greedy optimization algorithm.

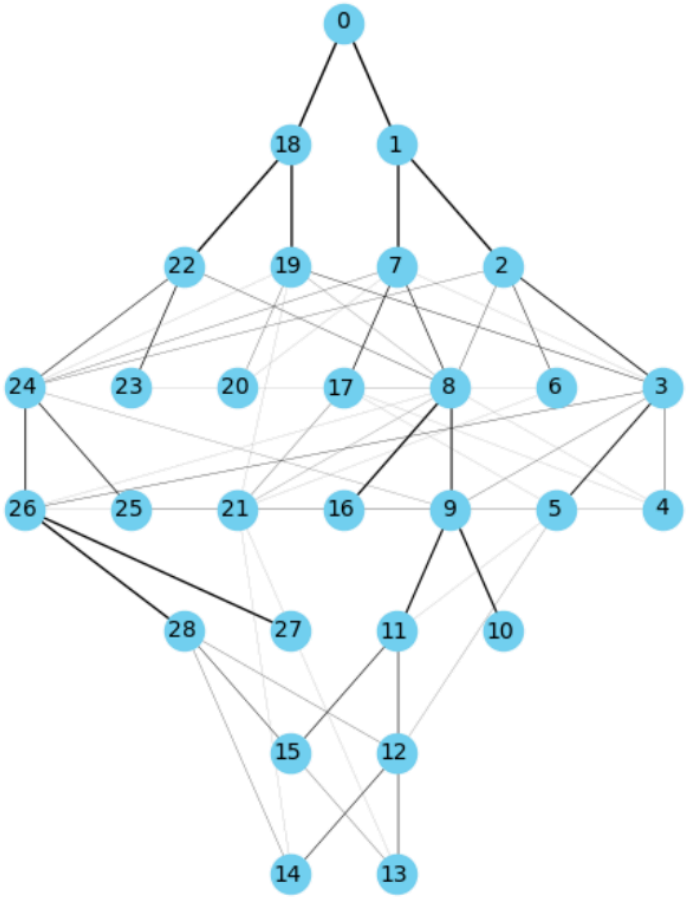
FORESTR: CENTRAL TREE

Given a set of T graphs, denoted as $G_i \in \mathcal{G}, t = 1, \ldots, m$, and their respective adjacency matrices, $A_t \in \mathbb{R}^{n \times n}$, the mean graph is obtained by minimizing the sum of squared distances:

$$[A_\mu] = \underset{A \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} \sum_{t=1}^{T} d_g([A], [A_i])^2$$

The mean graph is guaranteed to be unique, and the optimization is performed using a greedy optimization algorithm.

This optimization results in violations of the properties of trees:



FORESTR: CENTRAL TREE

Given a set of T graphs, denoted as $G_i \in \mathcal{G}, t = 1, \ldots, m$, and their respective adjacency matrices, $A_t \in \mathbb{R}^{n \times n}$, the mean graph is obtained by minimizing the sum of squared distances:

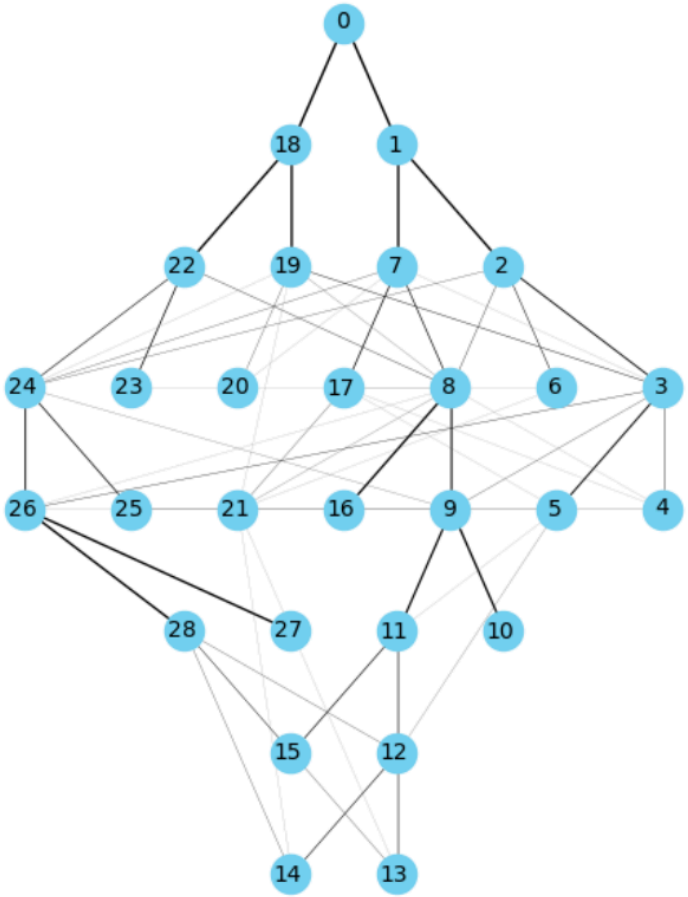
$$[A_\mu] = \underset{A \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} \sum_{t=1}^{T} d_g([A], [A_i])^2$$

The mean graph is guaranteed to be unique, and the optimization is performed using a greedy optimization algorithm.

This optimization results in violations of the properties of trees:

Currently, we compute a central tree as the three with the smallest distance to the mean graph. That is, let A_μ represent the mean graph, the central tree of a set of trees is computed as

$$A_{central} = \underset{A \in \{A_1, \ldots, A_T\}}{\arg\min} \; d_g(A, A_\mu).$$



FORESTR: CENTRAL TREE

Given a set of T graphs, denoted as $G_i \in \mathcal{G}, t = 1, \ldots, m$, and their respective adjacency matrices, $A_t \in \mathbb{R}^{n \times n}$, the mean graph is obtained by minimizing the sum of squared distances:

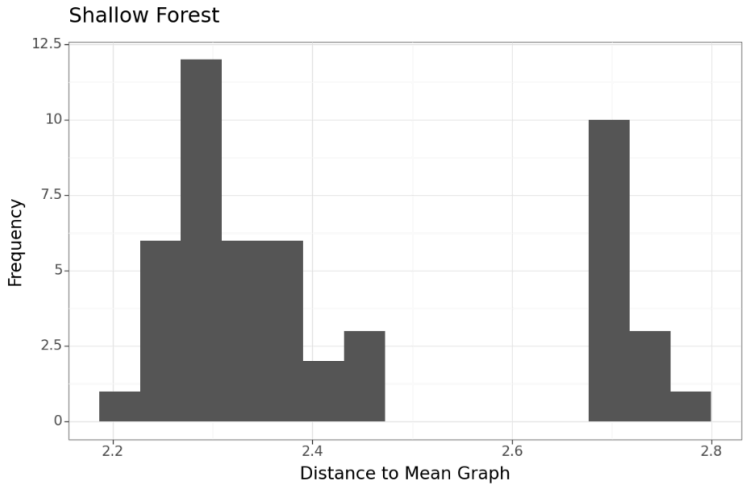
$$[A_\mu] = \underset{A \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} \sum_{t=1}^{T} d_g([A], [A_i])^2$$

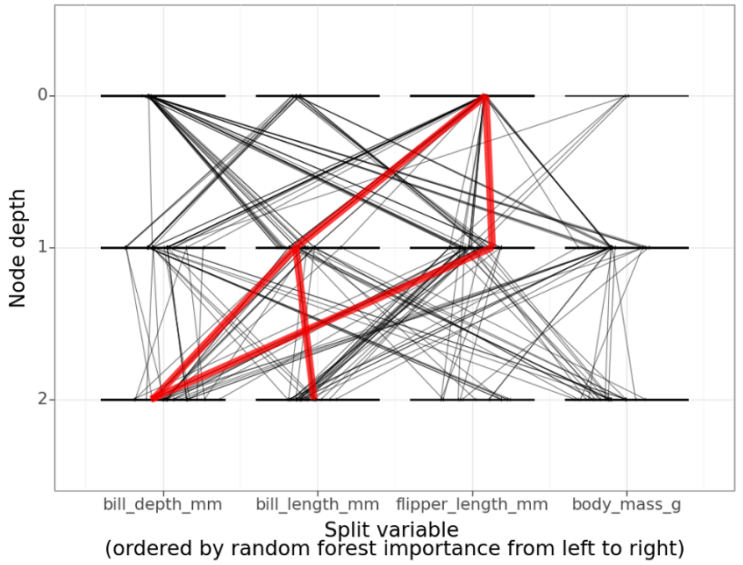
The mean graph is guaranteed to be unique, and the optimization is performed using a greedy optimization algorithm.

This optimization results in violations of the properties of trees:

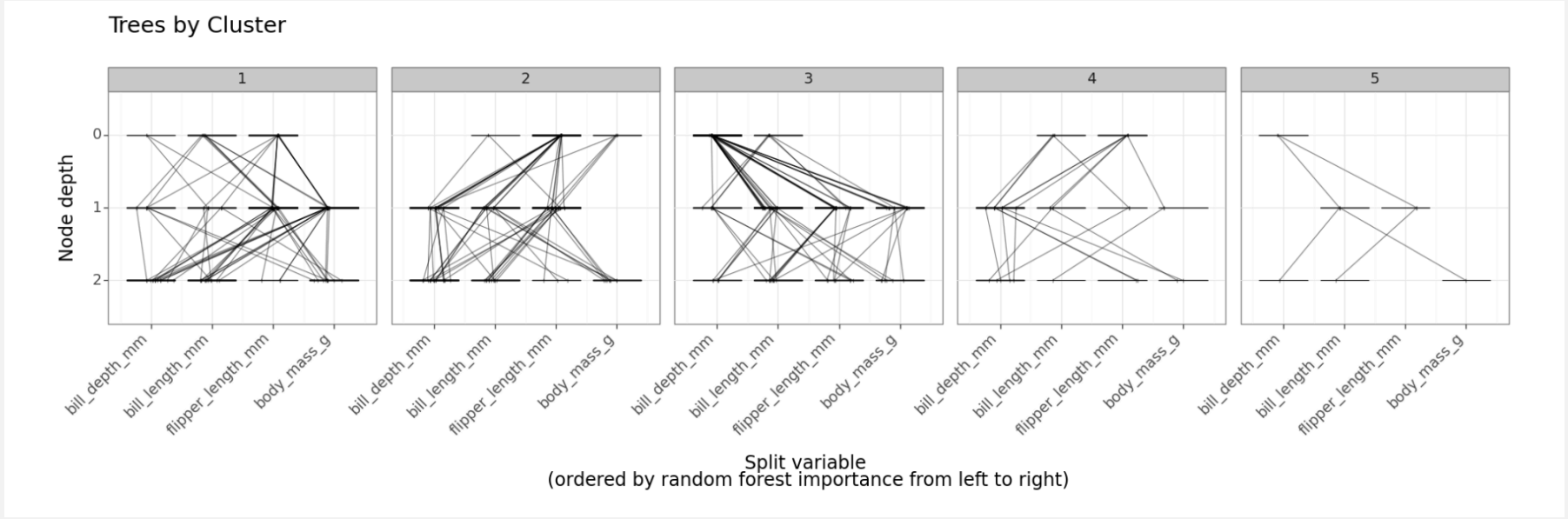
Currently, we compute a central tree as the three with the smallest distance to the mean graph. That is, let A_μ represent the mean graph, the central tree of a set of trees is computed as

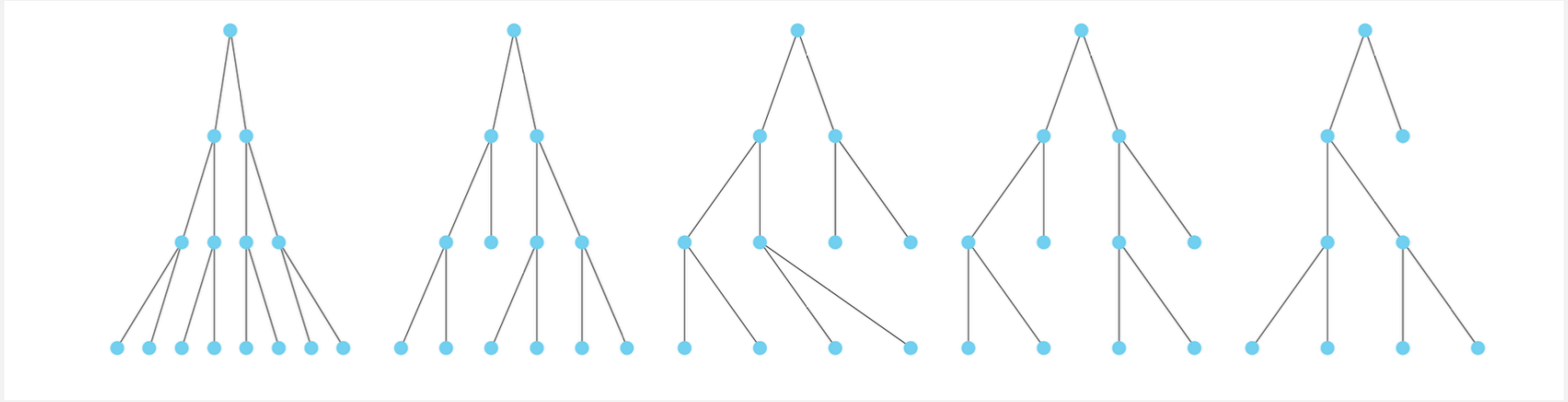
$$A_{central} = \underset{A \in \{A_1, \ldots, A_T\}}{\arg\min} \; d_g(A, A_\mu).$$





FORESTR: CENTRAL TREE WITHIN CLUSTERS





Central trees from each cluster

FORESTR: PRODUCT INSPECTION APPLICATION

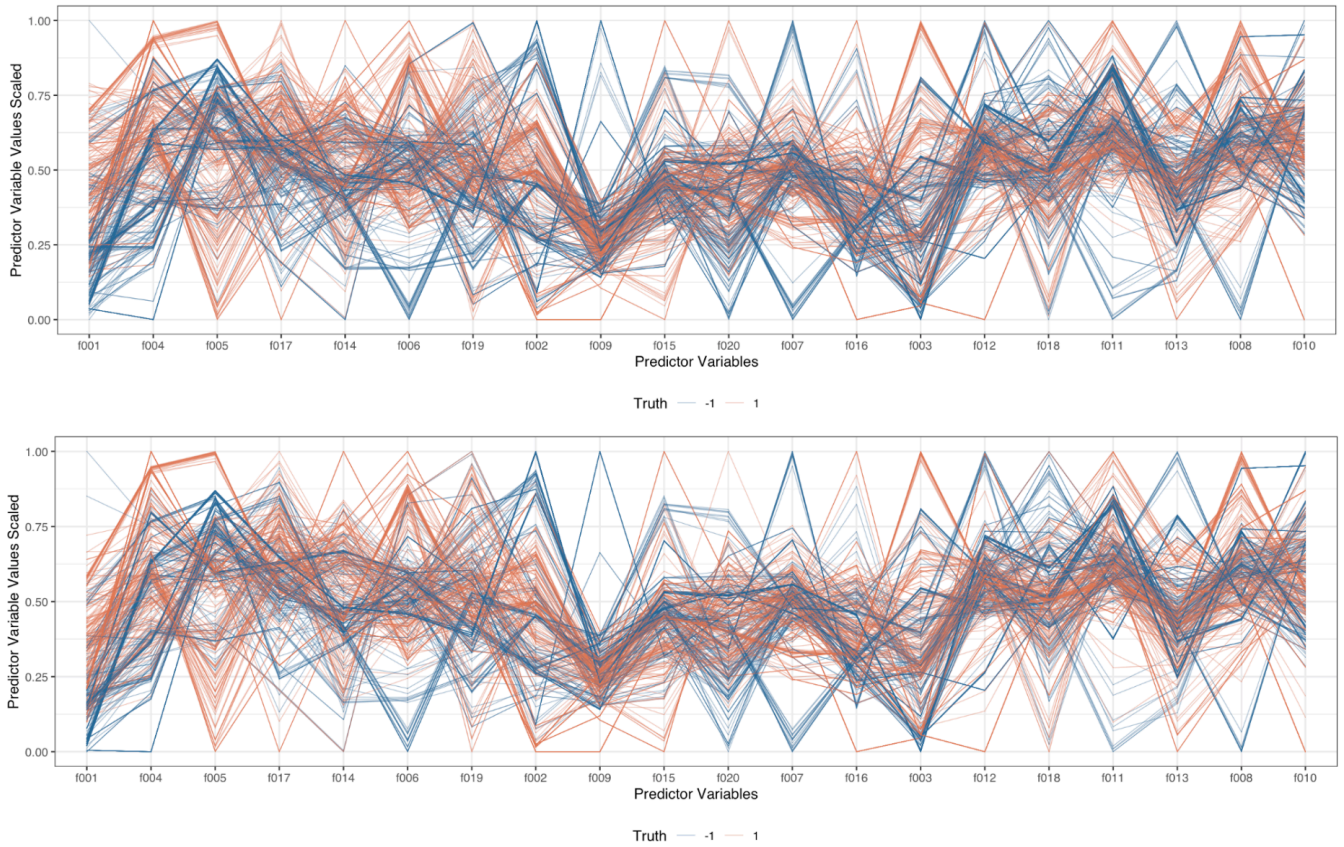
FORESTR: PRODUCT INSPECTION APPLICATION

s500 data

- Information from a product inspection application (Kegelmeyer 2015)
- 20 numeric features associated with 1000 products identified as good or bad
- Separated into training and testing sets with 500 observations in each set

Predict good/bad

- 25 trees
- Maximum depth 5
- Fit using scikit-learn and all other default parameters



Parallel coordinate plots of s500 training (top) and testing (bottom) data. Lines are colored by the inspection labels, and the features are ordered by Gini importance.

P. Kegelmeyer, T. M. Shead, J. Crussell, K. Rodhouse, D. Robinson, C. Johnson, D. Zage, W. Davis, J. Wendt, T. Cayton J. Doak, R. Colbaugh, K. Glass, B. Jones, and J. Shelburg. Counter adversarial data analytics Technical report, Sandia National Labs (SAND2015-3711), 2015.

FORESTR: PRODUCT INSPECTION APPLICATION

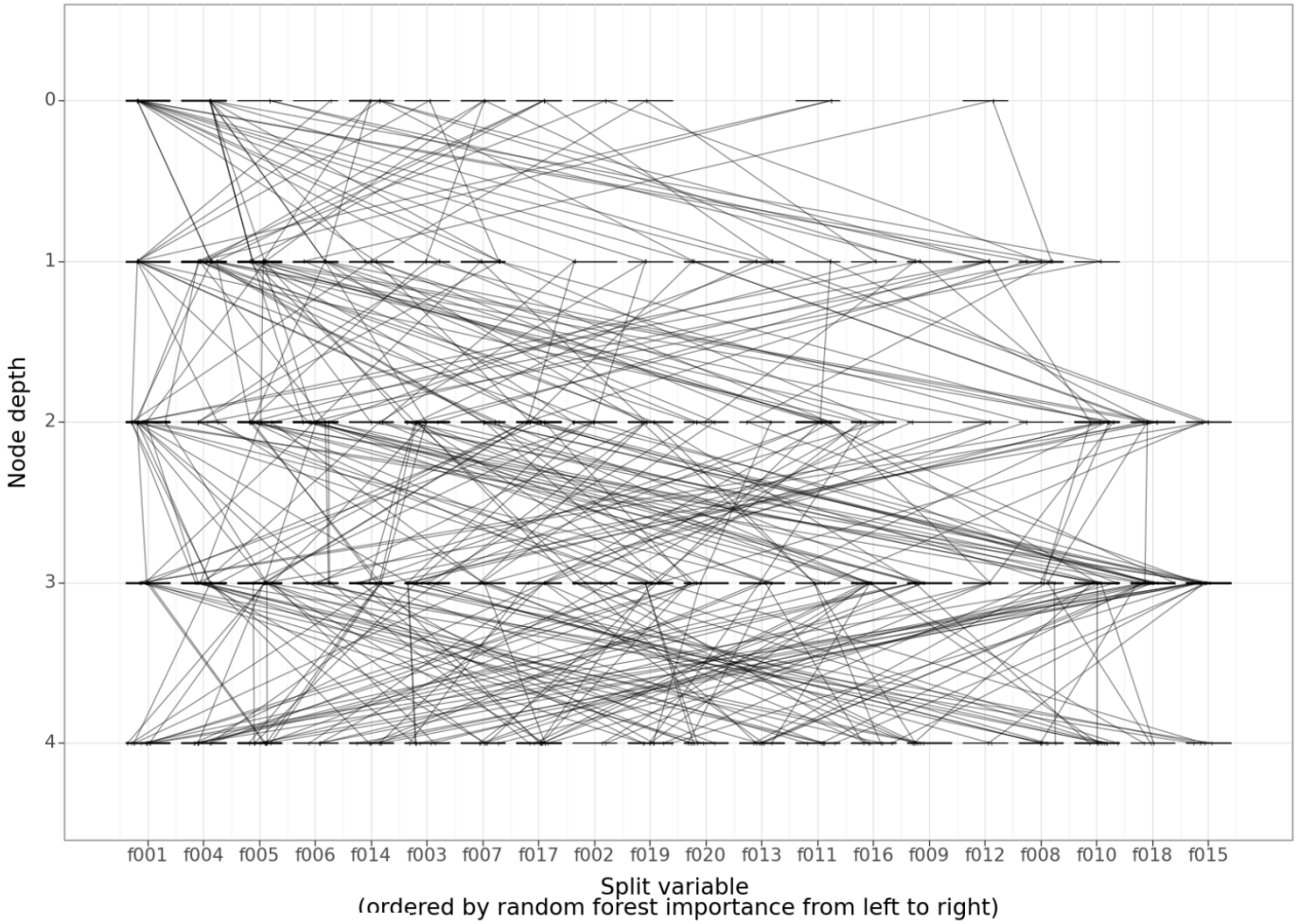
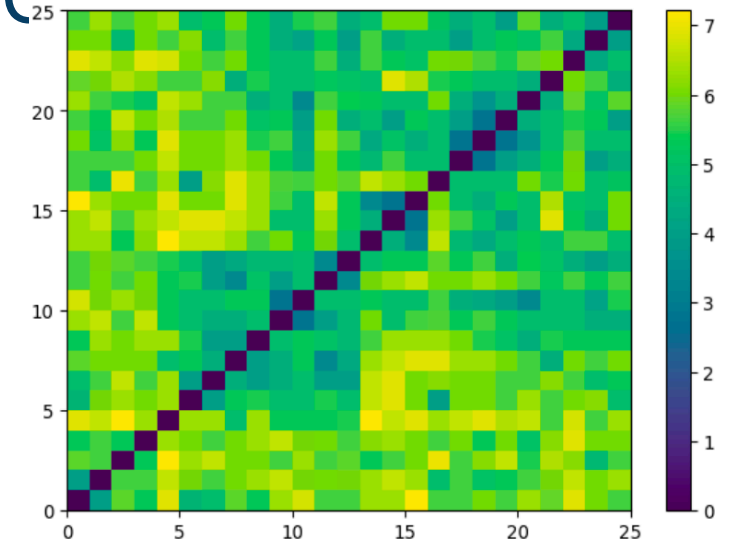
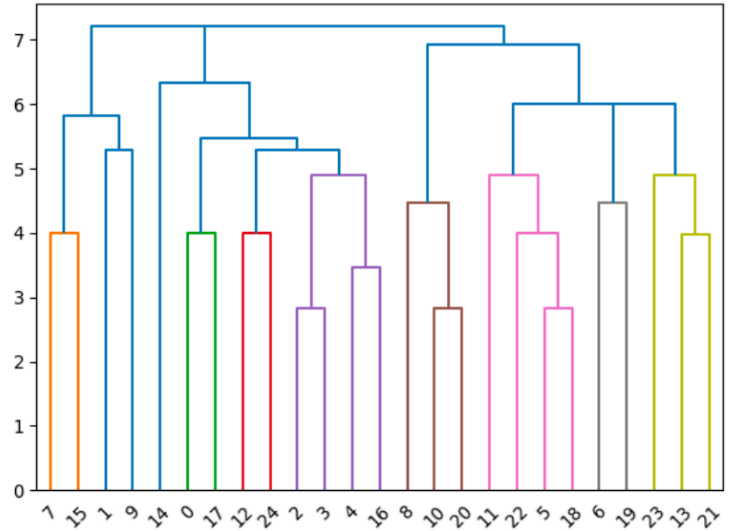


Figure 4-27.: Trace plot of the trees in the s500 random forest.



All pairwise distances

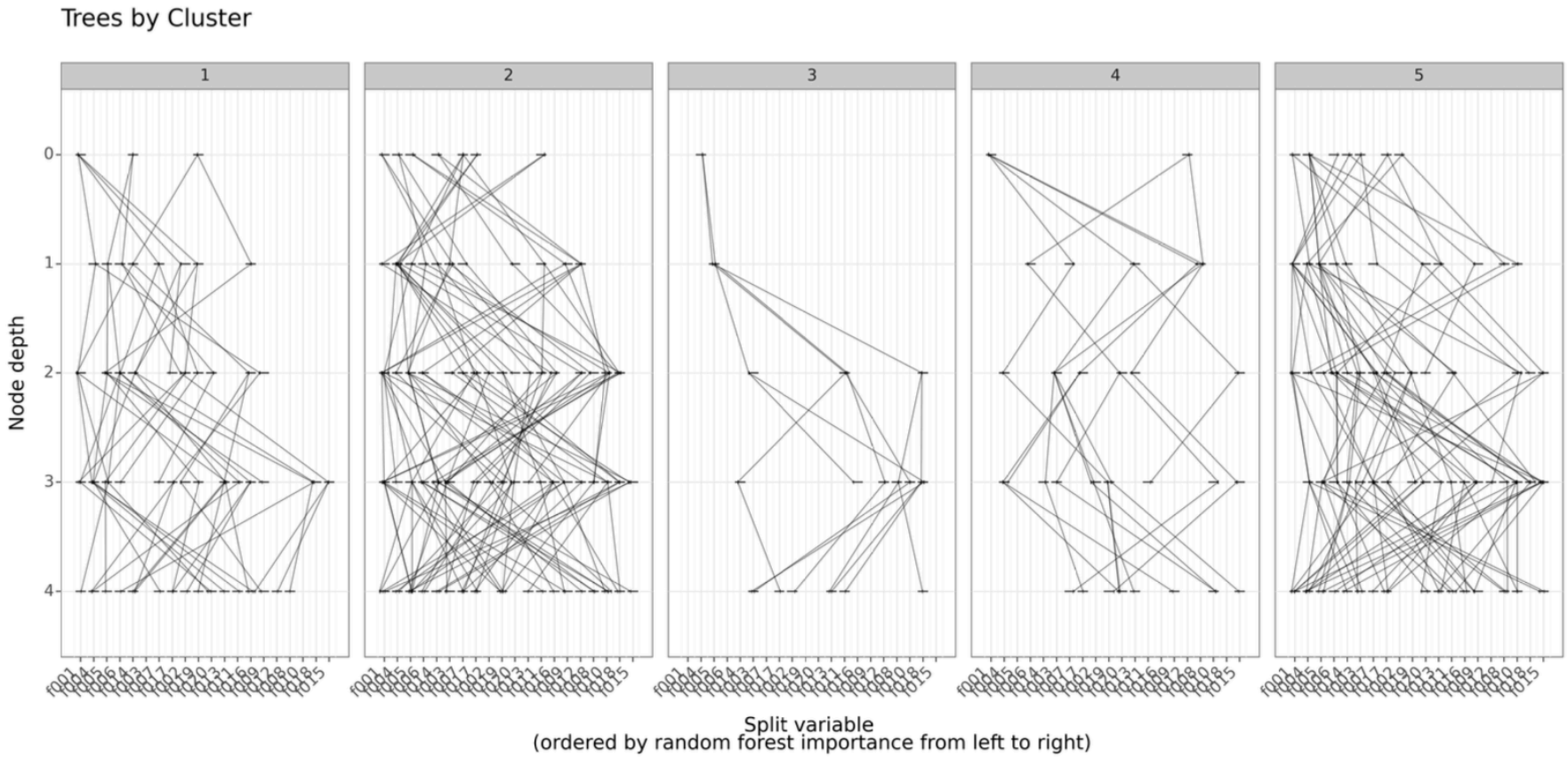


Dendrogram of complete linkage clusters based on pairwise distances

FORESTR: PRODUCT INSPECTION APPLICATION

The number of trees in clusters 3 and 4 is much less than the other clusters (e.g., only 1 tree in cluster 3).

- Could these be outlier trees?
- How would removing them from the model affect model performance?

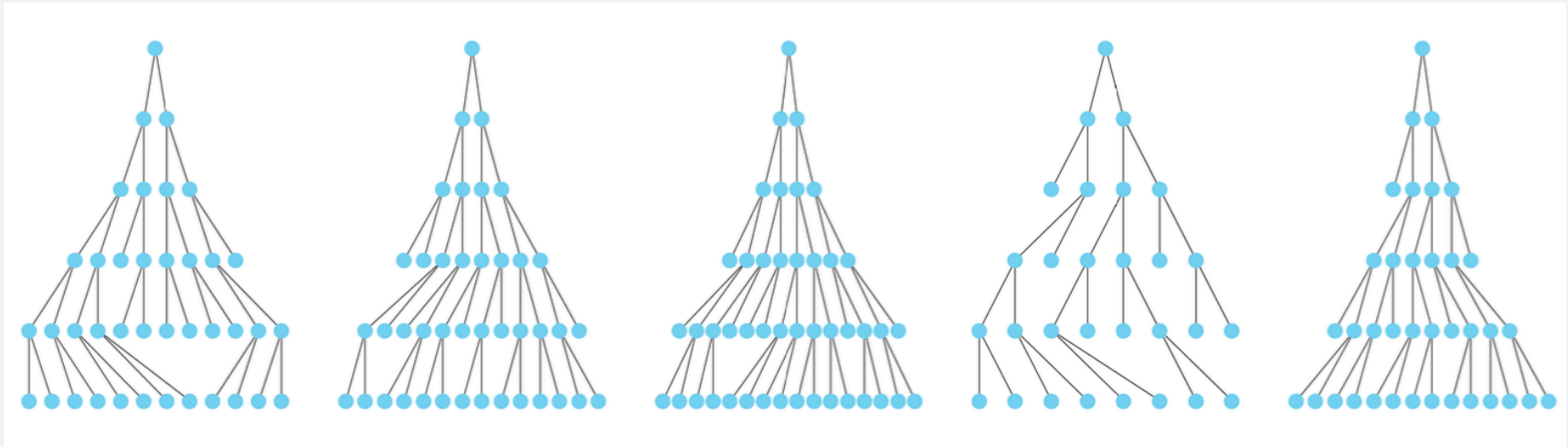


Trace plots of the tree clusters from the s500 random forest.

FORESTR: PRODUCT INSPECTION APPLICATION

The central trees help identify differences between the clusters:

- the depths at which leaf nodes begin to occur and
- the number of leaf nodes at a depth.

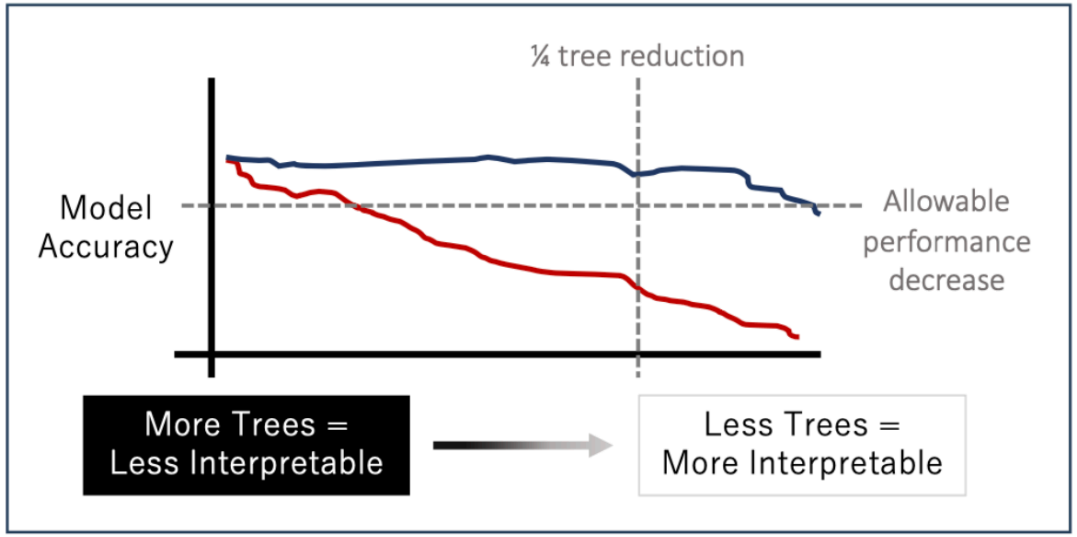


Central trees from each cluster

FORESTR: FUTURE RESEARCH DIRECTIONS

FORESTR: FUTURE RESEARCH DIRECTIONS

- Account for **additional node attributes**
  - (e.g., split feature)
- Consider approaches to identify **repeating sub-graphs**
  - (i.e., split patterns that are common among the forest)
- Use distances to **reduce ensemble size** and consider affect on predictive performance
  - Build up a model using a central tree or central trees from identified clusters
  - Trim down a model based on their distance to central tree or cluster central trees



THANK YOU

KJGOODE@SANDIA.GOV
GOODEKAT.GITHUB.IO

Technical Report Goode, Katherine Jean and Tucker, James Derek. "FORESTR: Finding, Organizing, Representing, Explaining, Summarizing, and Thinning Random forests." Sep. 2024. https://doi.org/10.2172/2472741.