SAND2020-7861C

Parsimonious Rational Belief Foundations for Trusted AI

Jed A. Duersch and Thomas A. Catanach

Sandia Machine Learning and Deep Learning Conference, 2020

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.

Trusted Artificial Intelligence

What makes a prediction trustworthy?

1. Scientific Method & Machine Learning
2. Naturalist Epistemology & Rational Belief
3. Bayesian Inference, Prior Belief, & Machine Learning
4. A Universal Formulation of Complexity
5. Numerical Experiments & Results

Scientific Method & Machine Learning

Standard Machine Learning (ML) training practices are analogous to the scientific method.

Scientific Method:

1. Gather evidence.
2. Formulate a hypothesis.
3. Test predictions with experiments.
4. Accept or reject the model.

Machine Learning:

1. Curate training data.
2. Adjust model parameters to predict training labels.
3. Monitor predictions using validation dataset.
4. Select model producing best validation predictions.

Unfortunately, standard approaches do not use data efficiently.
Wasteful methods require a lot of evidence to obtain reliable results.

Cross-Validation and Limited Data

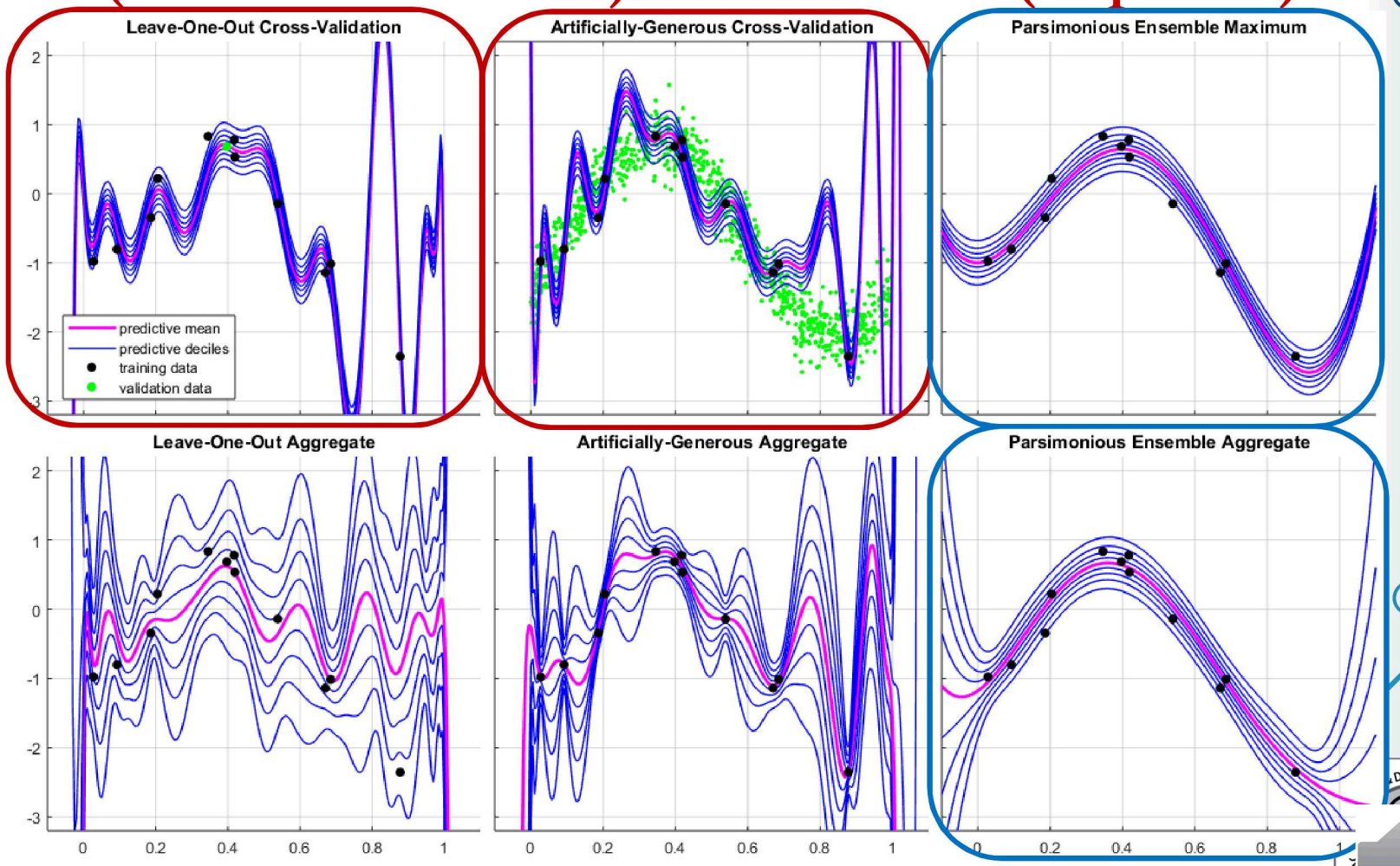
Underdetermined polynomial regression demonstrates the problem with having **many parameters (21 basis functions) and few data (12 points)**.

Not only is it **difficult to identify** the optimal model,

training fails to propose credible models.

By **controlling model complexity** from first principles,

we obtain **rigorously justified uncertainty** in predictions.



Rational Belief & Uncertainty Quantification

The subtext of **uncertainty quantification for machine learning** is the desire for **a clearer understanding of what may be true**.

The **scientific method** is based upon the **naturalist view of epistemology**.

Validity derives from consistency.

1. Rational beliefs must **avoid internal contradictions**.
2. Rational beliefs must **account for all past observations**.
3. As additional data become available, credible past beliefs yield **predictions matching additional evidence**.
4. As additional data become available, **rational beliefs evolve**.

The first point allows us to place the remaining points within a mathematically rigorous extended logic, Bayesian inference.

Belief as an Extended Logic

Building on the work of Keynes (1929) and Jeffreys (1939), Cox (1946) uses **binary logic** to derive the laws of probability as **an extended logic** representing degrees of truth.

Logic:

$$\sim\sim a = a, \quad (1)$$

$$a \cdot b = b \cdot a, \quad (2) \qquad a \vee b = b \vee a, \quad (2')$$

$$a \cdot a = a, \quad (3) \qquad a \vee a = a, \quad (3')$$

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c = a\ b \cdot c, \quad (4)$$

$$a \vee (b \vee c) = (a \vee b) \vee c = a \vee b \vee c, \quad (4')$$

$$\sim(a \cdot b) = \sim a \vee \sim b, \quad (5)$$

$$\sim(a \vee b) = \sim a \cdot \sim b, \quad (5')$$

$$a \cdot (a \vee b) = a, \quad (6) \qquad a \vee (a \cdot b) = a. \quad (6')$$

Extended logic:

- Probability is nonnegative.
- Impossibility has probability zero.
- Certainty has maximum probability, normalized to one.
- **Bayes' theorem conditions belief on evidence.**

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)}$$

Bayesian Inference

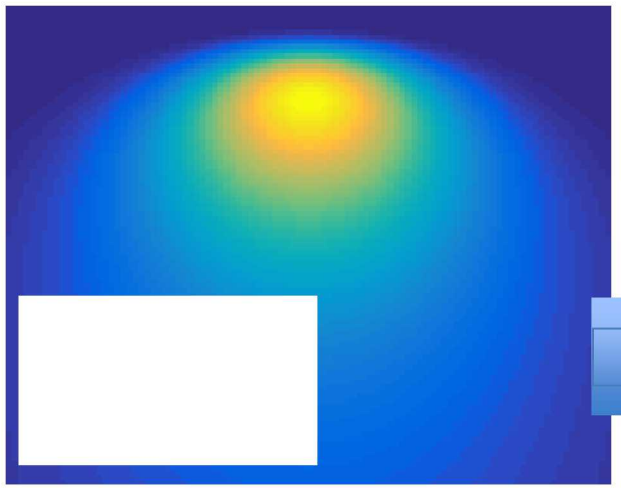
features ↘ ↙ labels

Empirical Data: $\mathcal{D} = \{(x_i, y_i) \mid i \in [n]\}$

parameters ↙

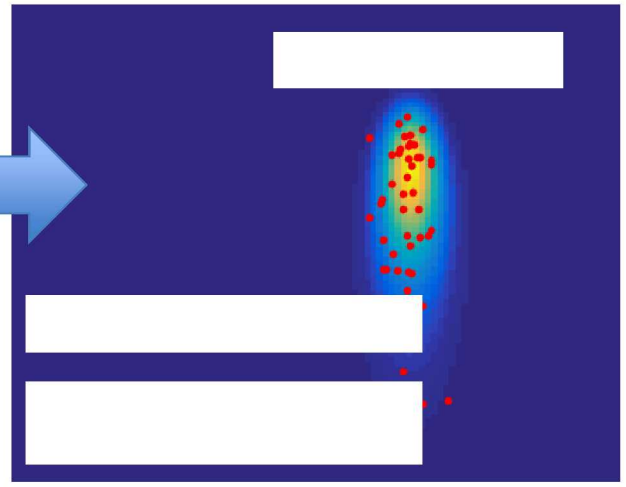
Likelihood: $p(\mathcal{D} \mid \theta)$, or more precisely $p(y \mid x, \theta)$

Evidence: $p(\mathcal{D} \mid \mathcal{M}) = \int d\theta \, p(\mathcal{D} \mid \theta) p(\theta \mid \mathcal{M})$



Bayes' Theorem

$$p(\theta \mid \mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D} \mid \theta) p(\theta \mid \mathcal{M})}{p(\mathcal{D} \mid \mathcal{M})}$$

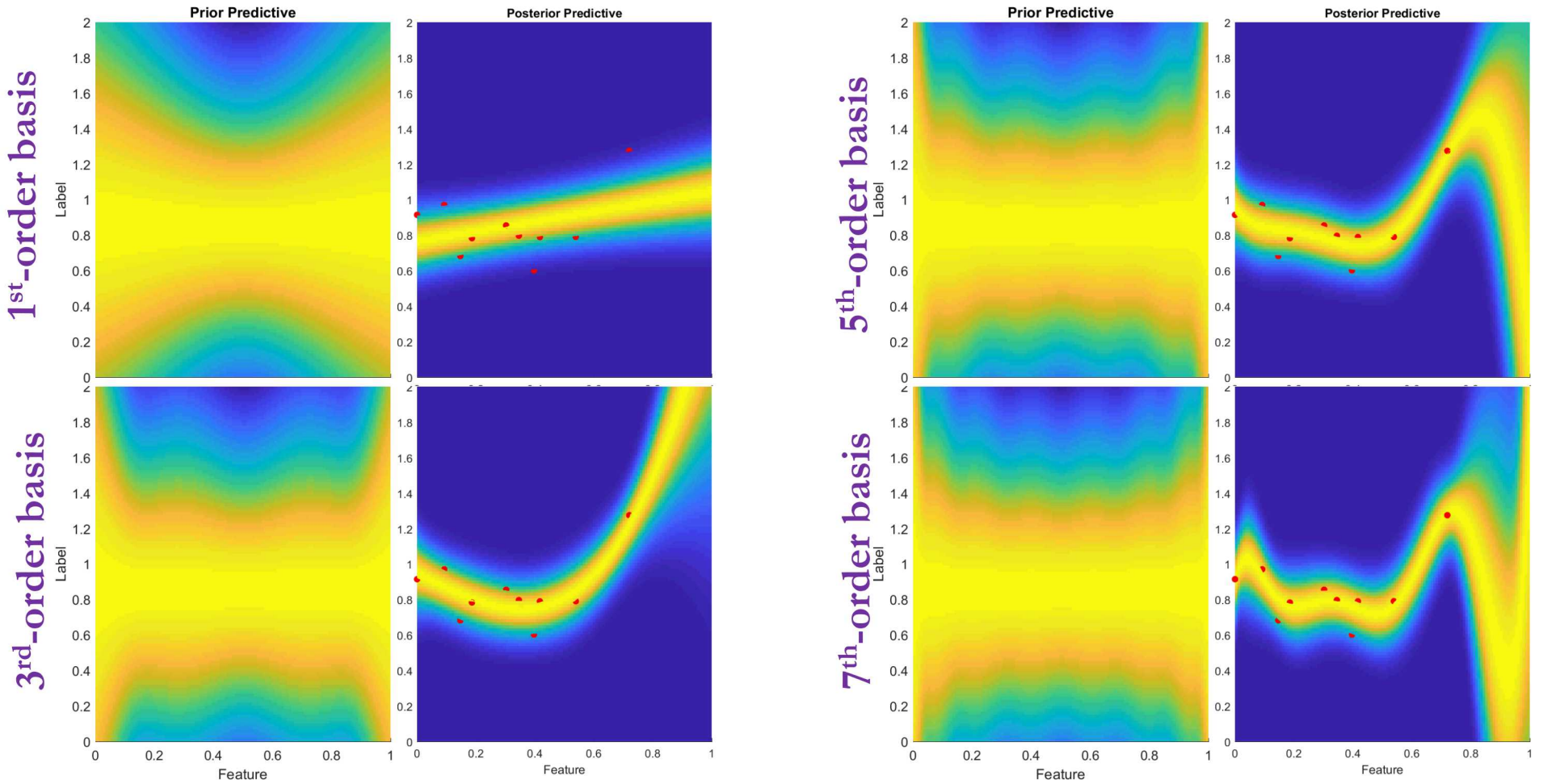


Rational Predictions: $p(\hat{y} \mid \hat{x}, \mathcal{D}, \mathcal{M}) = \int d\theta \, p(\hat{y} \mid \hat{x}, \theta) p(\theta \mid \mathcal{D}, \mathcal{M})$

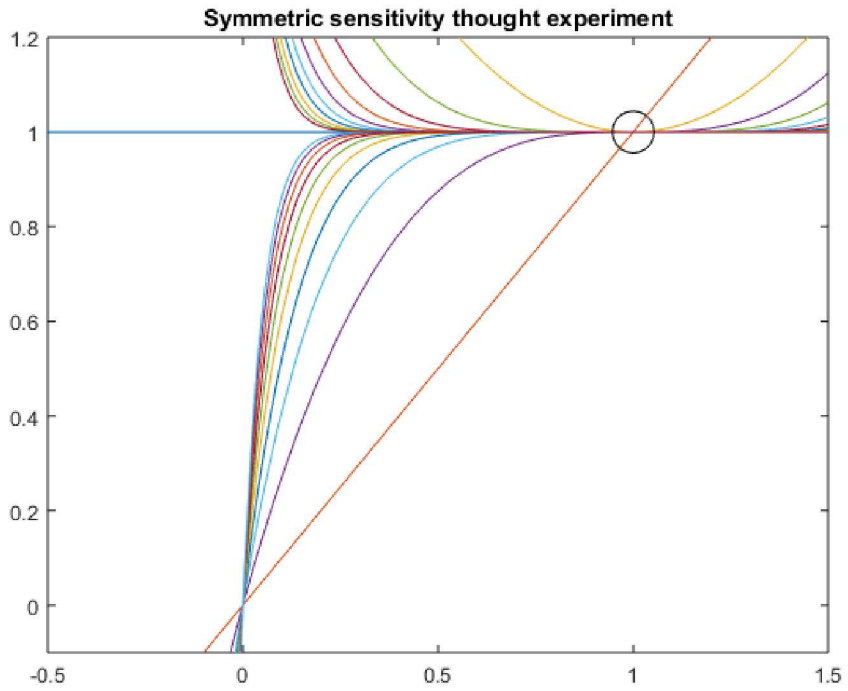
potential labels ↗ ↖ new features

The Problem of Prior Belief

Inference requires prior belief. When data are limited, predictions are highly sensitive to prior belief (Owhadi, 2015). **Maximum entropy priors** demonstrate this:



Scope of the Machine Learning Problem



- **Any convex combination** of basis functions **explains this point.**
- This is **every machine learning problem**.
- With n observations, d differentiable parameters, we have at least $d - n$ degenerate parameter dimensions; **prior belief totally determines posterior belief on this manifold.**

The **model universe** is the set of **all coherent predictive models**

$$\Theta = \{\theta \,|\, p(y \mid x, \theta) \text{ predicts } y \text{ from } x\}\,.$$

In this perspective, **prior belief subsumes computational architecture and regularization.**

Information as Change in Belief

Postulates of **information as a rational measure of change in belief:**

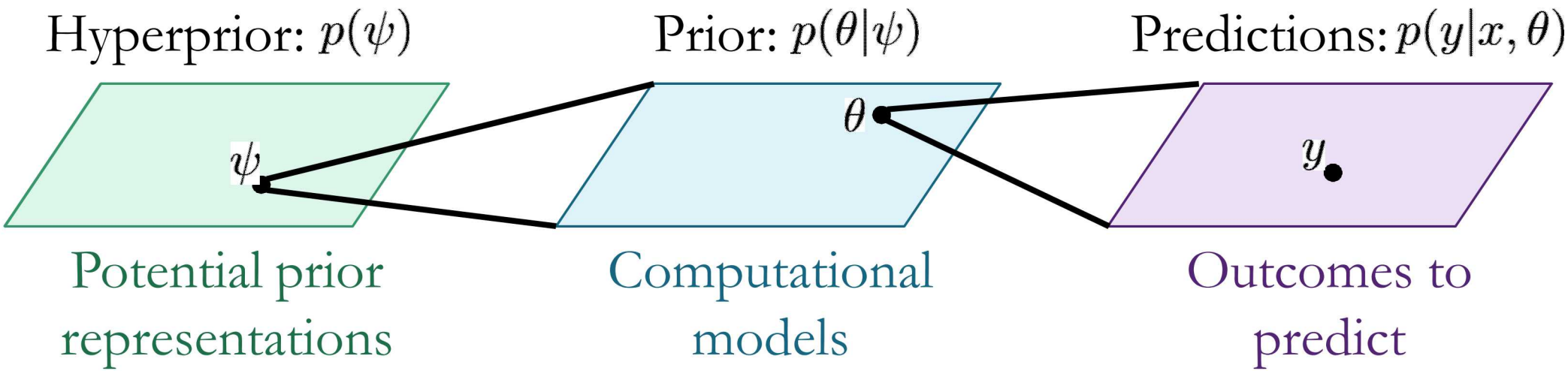
1. Information is a **reasonable expectation** over **rational belief** or a **hypothetical choice** measuring a change in belief.
2. Information is **additive** over **independent processes**.
3. When we have **no change** in belief, we have **zero information**.
4. The **information gained** from any hypothetical belief **to rational belief** is **nonnegative**.

Theorem 1. Information satisfying these postulates is computed as

$$\mathbb{I}_{r(z)}[\, q_1(z) \,\|\, q_0(z) \,] = \alpha \int dz\, r(z) \log\left(\frac{q_1(z)}{q_0(z)}\right).$$

Duersch, J.A.; Catanach, T.A., Generalizing Information to the Evolution of Rational Belief. Entropy Journal, 2020.

Theoretical Framework to Control Complexity



- **A prior representation $\boldsymbol{\psi}$ fits in your computer**; it is a discrete random variable and corresponds to some sequence of symbols.
- **Prior complexity is the amount of information generated** when a specific representation is realized.

$$\chi(\check{\psi}) = \mathbb{I}_{r(\psi|\check{\psi})}\left[r(\psi|\check{\psi})\|p(\psi)\right] = -\log\left(p(\check{\psi})\right)$$

Program Length & Algorithmic Probability

- Solomonoff (1960) used program length to derive algorithmic probability:

  As a program, $\psi(\cdot)$ has length $\ell(\psi)$ given by the number of bits used to encode it with some interpreter.

  $$\text{Given} \quad \Psi(x,y) = \{\psi_i | \psi_i(x) = y, i = 1,2,\ldots\} \quad \text{then} \quad p(\psi_i) \propto 2^{-\ell(\psi_i)}.$$

- Kolmogorov (1968) used program length to define information-theoretic complexity of a mapping:

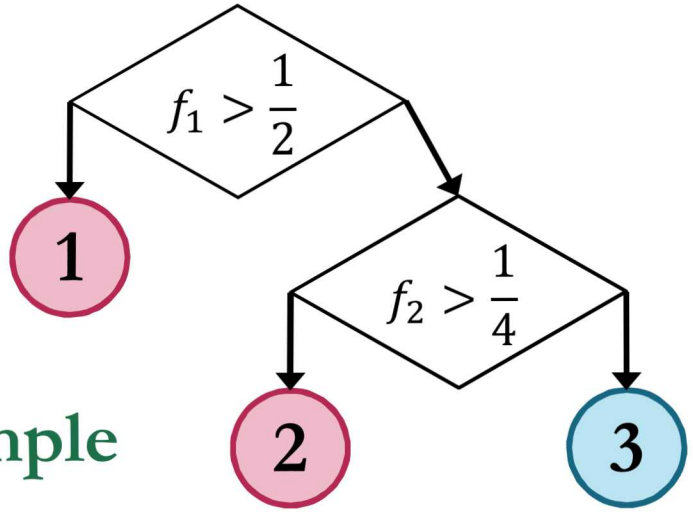
  $$K(x,y) = \min_{\psi \in \Psi(x,y)} \ell(\psi).$$

For our purposes, **a sequence of symbols ψ specifies prior belief**, including both a manifold of potential models and the parameter distribution over that manifold.

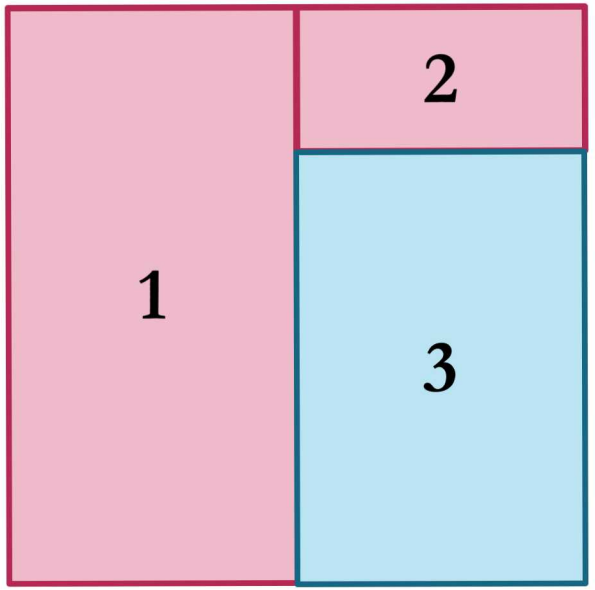
A **universal hyperprior over arbitrary architectures** easily follows by regarding **prior complexity as program length**.

$$\chi(\psi) \equiv \ell(\psi).$$

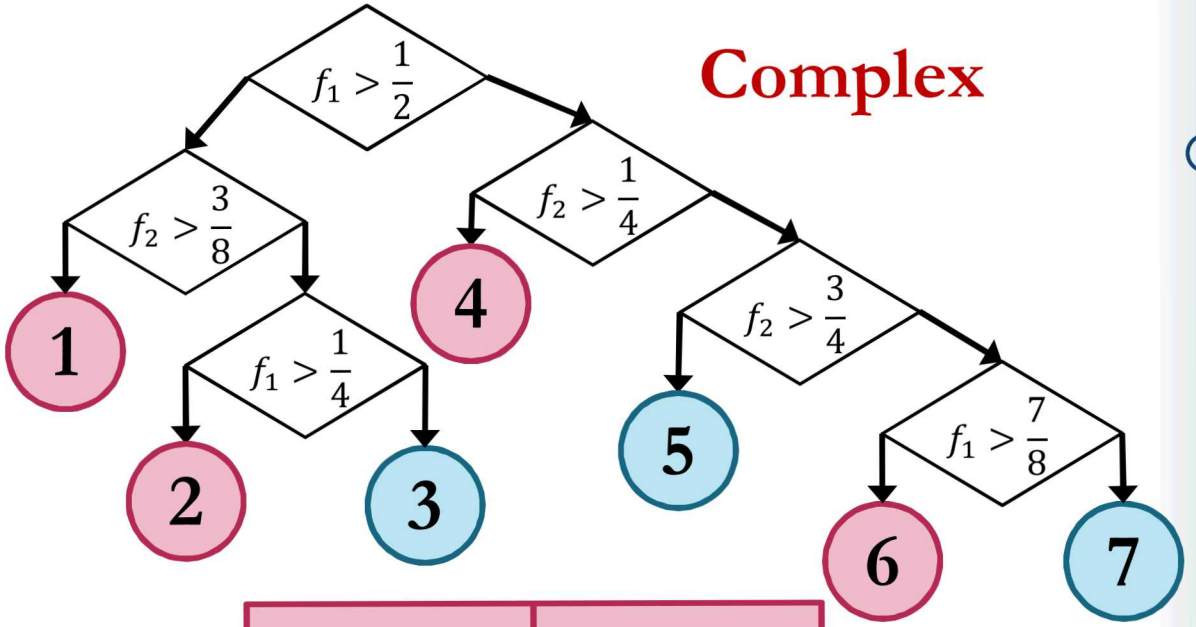
Complexity in Decision Trees

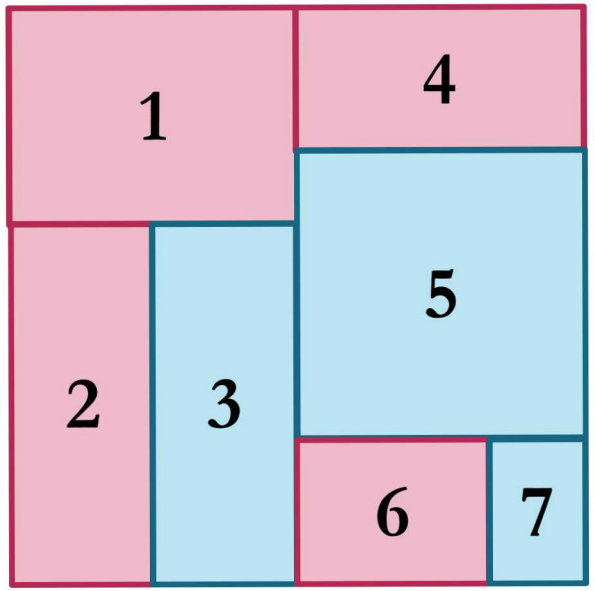


Simple

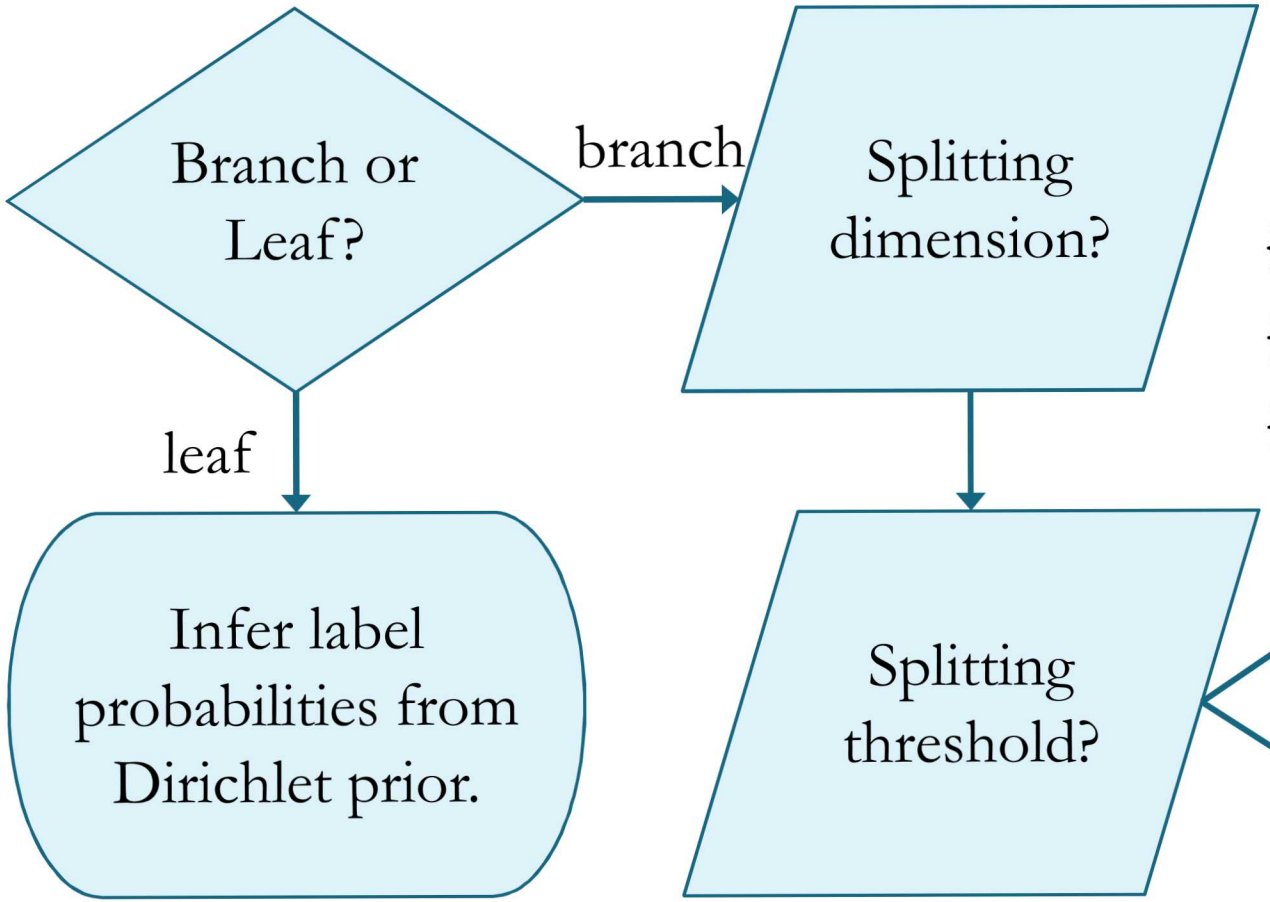


Complex





Example Encoding, Decision Tree Node



Example code: (branch, 2, 1 0 1)

{branch, leaf} $\log_2(2) = 1$ bit

{1, 2, 3} $\log_2(3) = 1.58$ bits

{½ : 0, ¼ : 1 0 0, ¾ : 1 0 1, … }

1 bit 3 bits 3 bits

Information Minimization Objective

How much does our belief change in $\boldsymbol{y}, \boldsymbol{\psi}$, and $\boldsymbol{\theta}$ when we observe data $\check{\boldsymbol{y}}$, select a prior representation $\check{\boldsymbol{\psi}}$, and infer model belief?

$$\check{\psi}^* = \arg\min_{\check{\psi}} \mathbb{I}_{r(y|\check{y})r(\psi|\check{\psi})p(\theta|\check{y},\check{\psi})} \left[r(y|\check{y})r(\psi|\check{\psi})p(\theta|\check{y},\check{\psi}) \,\|\, p(\psi,\theta,y) \right]$$

$$\check{\psi}^* = \arg\max_{\check{\psi}} \omega(\check{\psi})$$

$$\omega(\check{\psi}) = \mathbb{E}_{p(\theta|\check{y},\check{\psi})} \mathbb{I}_{r(y|\check{y})} \left[p(y|\theta) \,\|\, q_0(y) \right]$$

Expected info gained about data.

$$- \mathbb{I}_{p(\theta|\check{y},\check{\psi})} \left[p(\theta|\check{y},\check{\psi}) \,\|\, p(\theta|\check{\psi}) \right]$$

Model info due to inference.

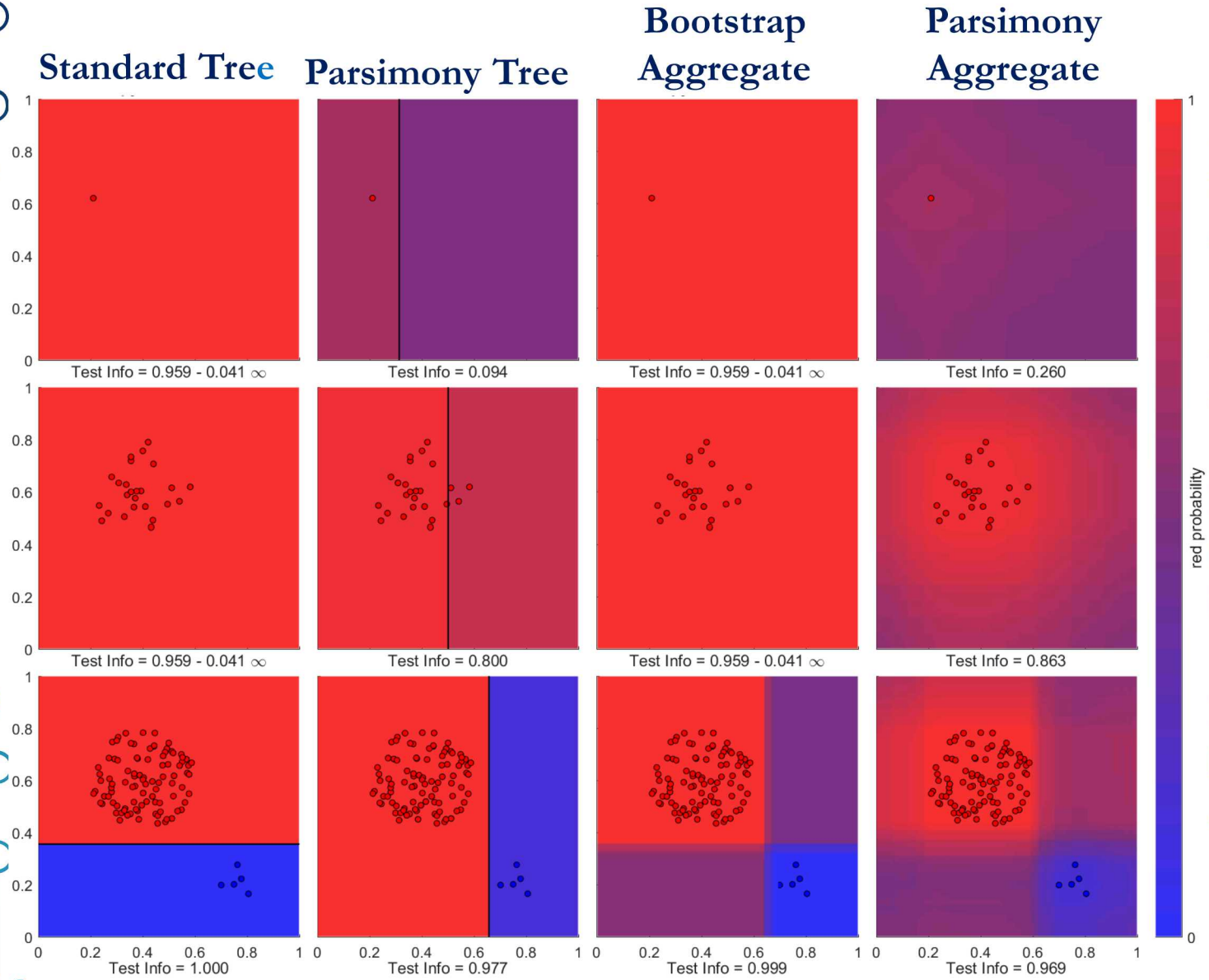
$$- \mathbb{I}_{r(\psi|\check{\psi})} \left[r(\psi|\check{\psi}) \,\|\, p(\psi) \right] .$$

Representation info due to selection.

$$= \log \left(p(\check{\psi}|\check{y}) \right) + \mathbb{I}_{r(y|\check{y})} \left[p(y) \| q_0(y) \right] .$$

Representation posterior.

Extreme Data Limitations & Uncertainty



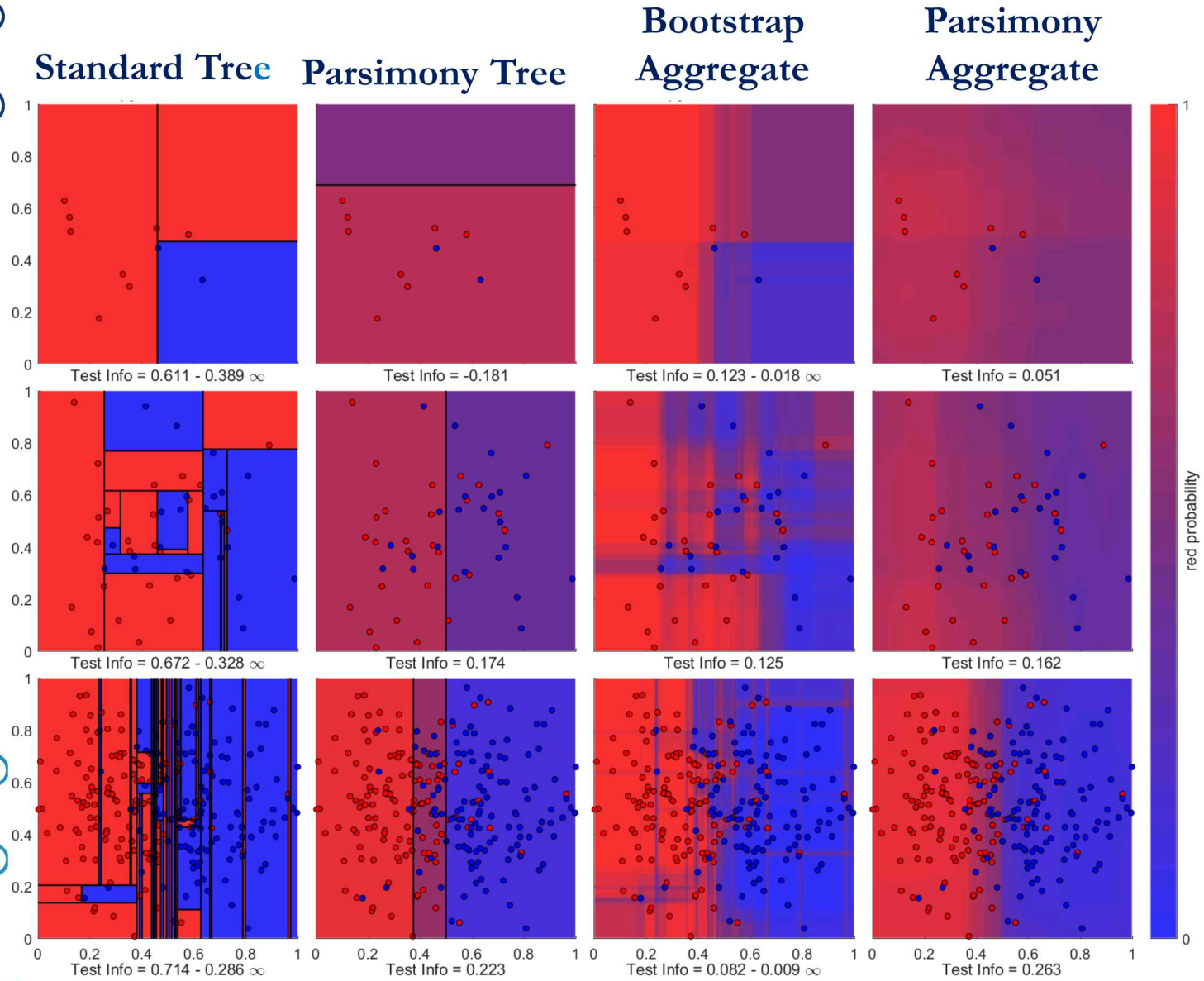
All three experiments are generated from the same, **heavily skewed**, distribution.

Row 1: Learning from a **single point** results in extreme uncertainty.

Row 2: A **group with the same label** contributes confidence. Uncertainty increases as we move away from data.

Row 3: Learning from **heavily skewed labels** is possible. Our model **avoids confidence in regions without data.**

Overlapping Labels and Artifacts



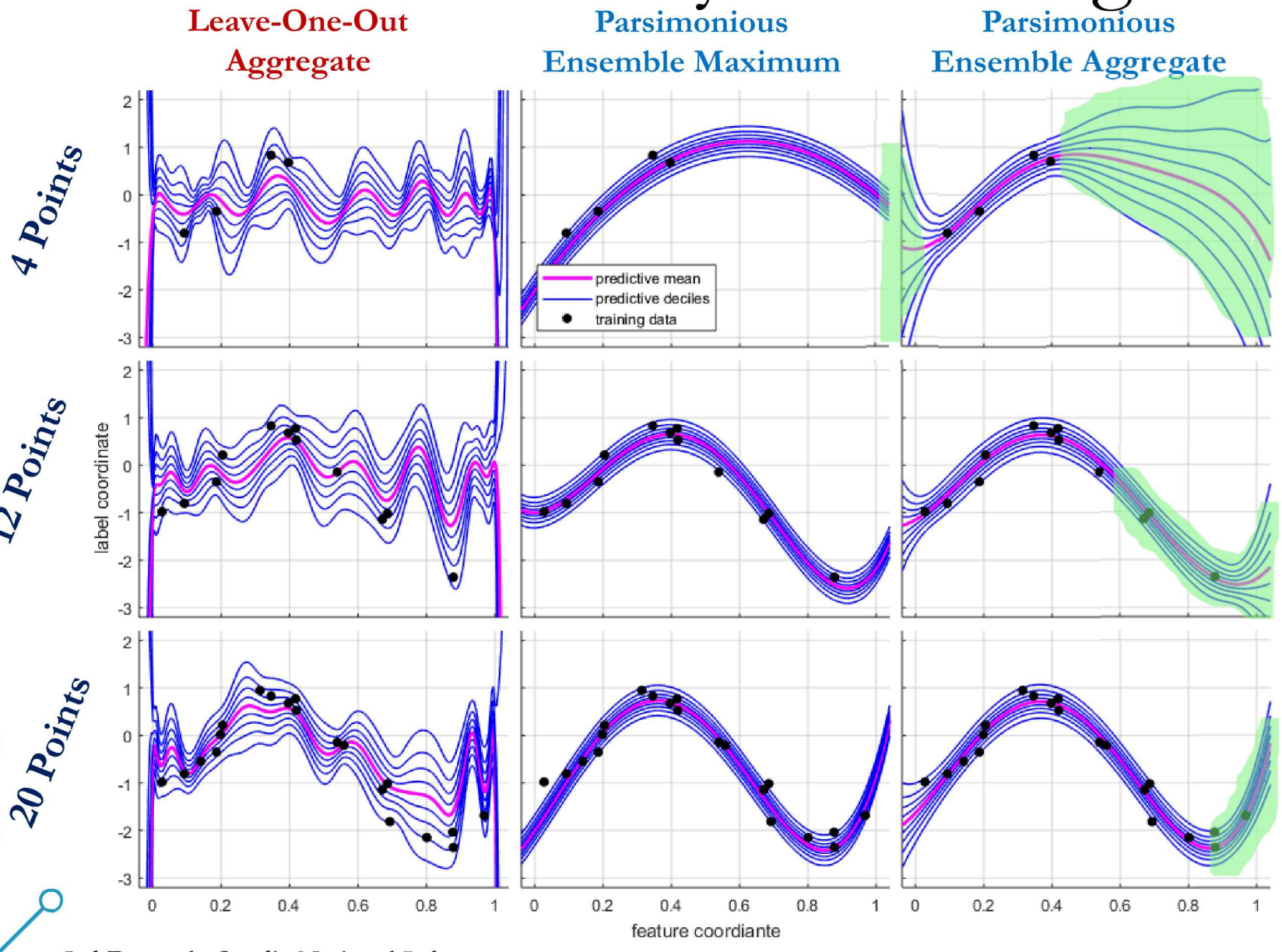
This distribution **generates both labels in the middle region.**

Row 1: The parsimonious aggregate **avoids confidence with few data.**

Row 2: Bootstrap aggregation may produce **artifacts that artificially hew to the data.**

Row 3: Parsimony trees **resist increases in complexity** as the dataset grows.

Parsimonious Polynomial Regression



Accounting for many explanations gives us **natural extrapolation uncertainty.**

Additional data provide specificity.

Complexity justifiably increases with the size of the dataset.

Summary

- We are developing machine learning inference methods that are **rooted in the principles of science.**
- **Controlling complexity** with information is **more reliable than cross-validation**.
- Our next challenge is to **optimize efficiently in high parameter dimensions.**
- We believe this theoretical framework will support **1. well-founded learning from limited data, 2. firm differential privacy guarantees, 3. robust anomaly detection.**

Thank you!