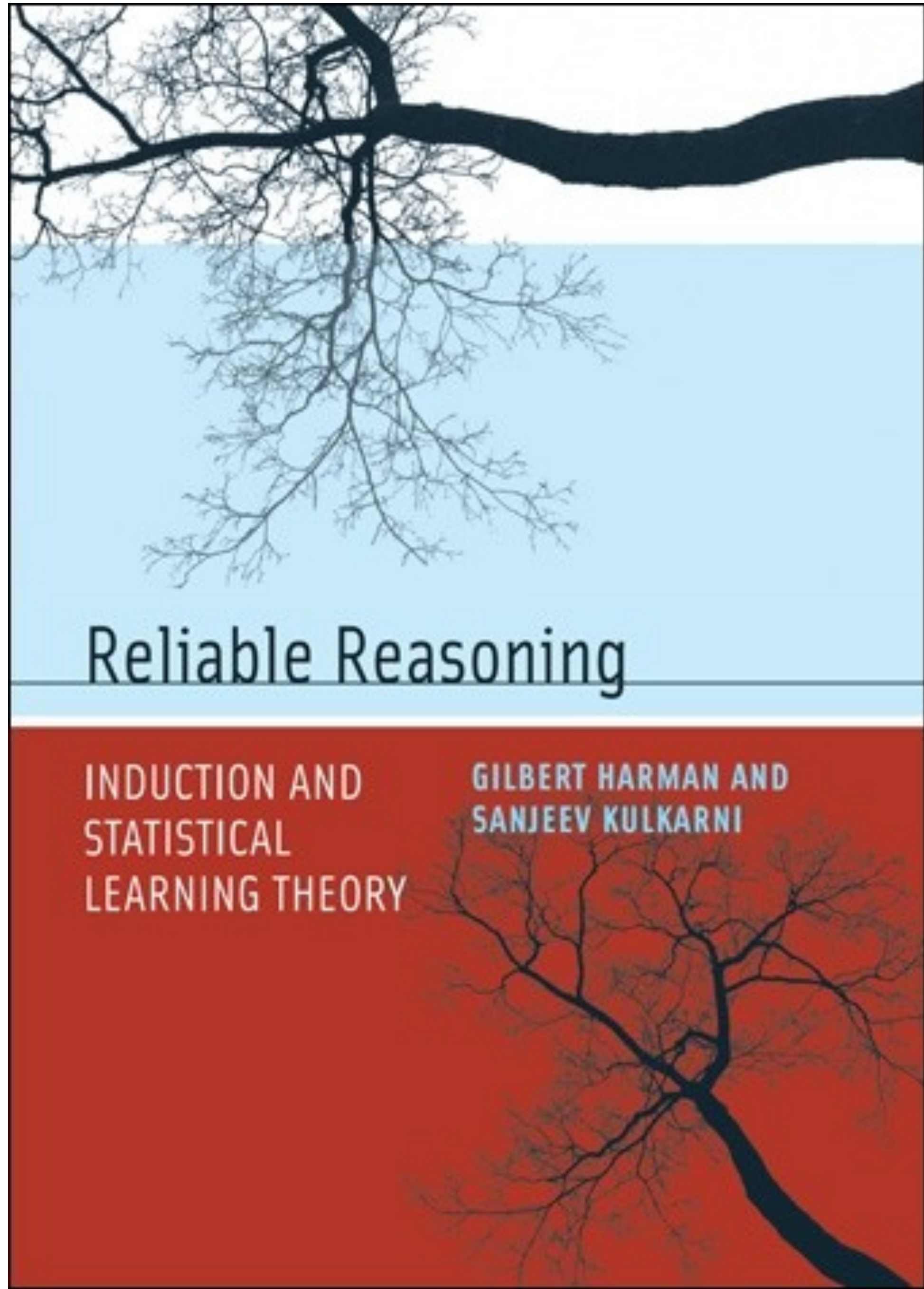


Reliable Reasoning and the Interpretability of ML Models

Pierre-Edouard Portier, INSA Lyon, Monday, February, 2021



Reliable Reasoning

INDUCTION AND
STATISTICAL
LEARNING THEORY

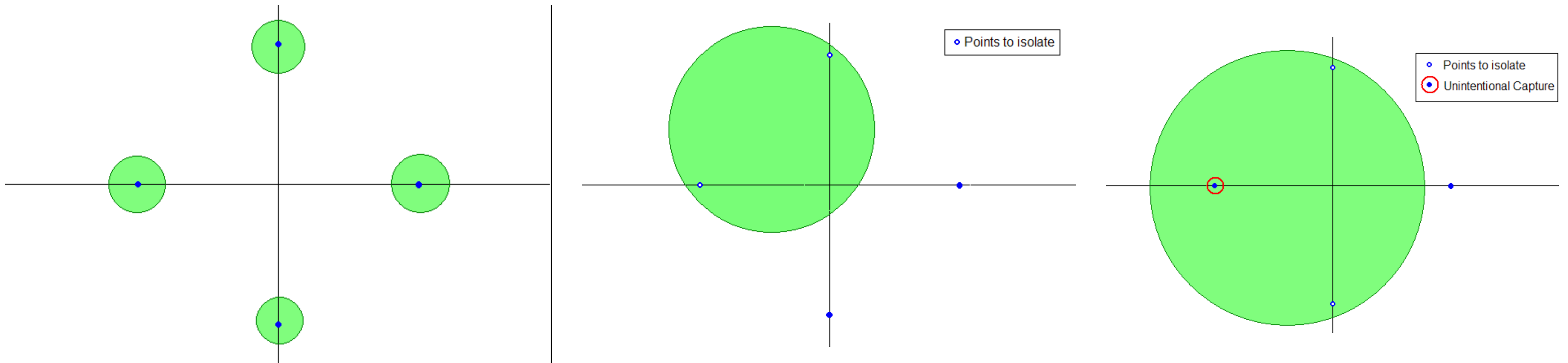
GILBERT HARMAN AND
SANJEEV KULKARNI

Deductive Logic and Induction

- Deductive rule: If “all F are G” then “the next F will be G”
 - Reliable: It never leads from true premises to a false conclusion.
- Inductive rule: If “many F have been found to be G” and “until now, no F have been found not to be G”, then “the next F will be G”.
- Inductive reasoning is a process of change in view.
- When can an inductive rule be considered reliable?

Vapnik and Chervonenkis (VC) Dimension

- The VC dimension of a set of rules C is the maximum number N of data points that can be arranged so that, for every one of the 2^N ways of assigning values to each of those points, there is a rule in C that is in accord with that assignment.



https://en.wikipedia.org/wiki/Shattered_set

Vapnik and Chervonenkis (VC) Dimension

- The VC dimension of a set of rules C is the maximum number N of data points that can be arranged so that, for every one of the 2^N ways of assigning values to each of those points, there is a rule in C that is in accord with that assignment.
- Relationship with Karl Popper's methodology: evidence cannot establish a scientific hypothesis, it can only falsify it.
- For finite VC dimension V , there is a function $m(V, \epsilon, \delta)$ that indicates the maximum amount of data needed to ensure that the probability is less than δ that enumerative induction (or, empirical risk minimization) will endorse a hypothesis with an expected error rate that exceeds the minimum expected error rate for rules in C by more than ϵ .

Limitation of Models with Finite VC Dimension

- Linear models in a D -dimensional feature space have VC dimension $D+1$.
- Even with enough data, the best linear rule can have a high expected error.
- We can use a richer class of rules.
- With a finite VC dimension, no guarantee for the expected error of the best rule in C to be close to the expected error of the overall best rule (i.e., the Bayes rule).

EPISTEMOLOGY
and the
PSYCHOLOGY
of HUMAN JUDGMENT



MICHAEL A BISHOP
J. D. TROUT

“Golden Rule of Predictive Modeling”

- Based on the same evidence, the predictions of (very simple) statistical models are typically more reliable than the predictions of human experts.
- This is true for proper linear models with the weights learned to best fit data.
- Also true for improper linear models based on bootstrapping (i.e., proper linear models of an expert’s judgments)...!
- Also true for random linear models (where the variables are defined to be positively correlated with the target)...?!
- Possible explanations: Flat maximum principle (Einhorn and Hogarth, 1975) and the nature of human decisions.

The “Broken Leg” Problem

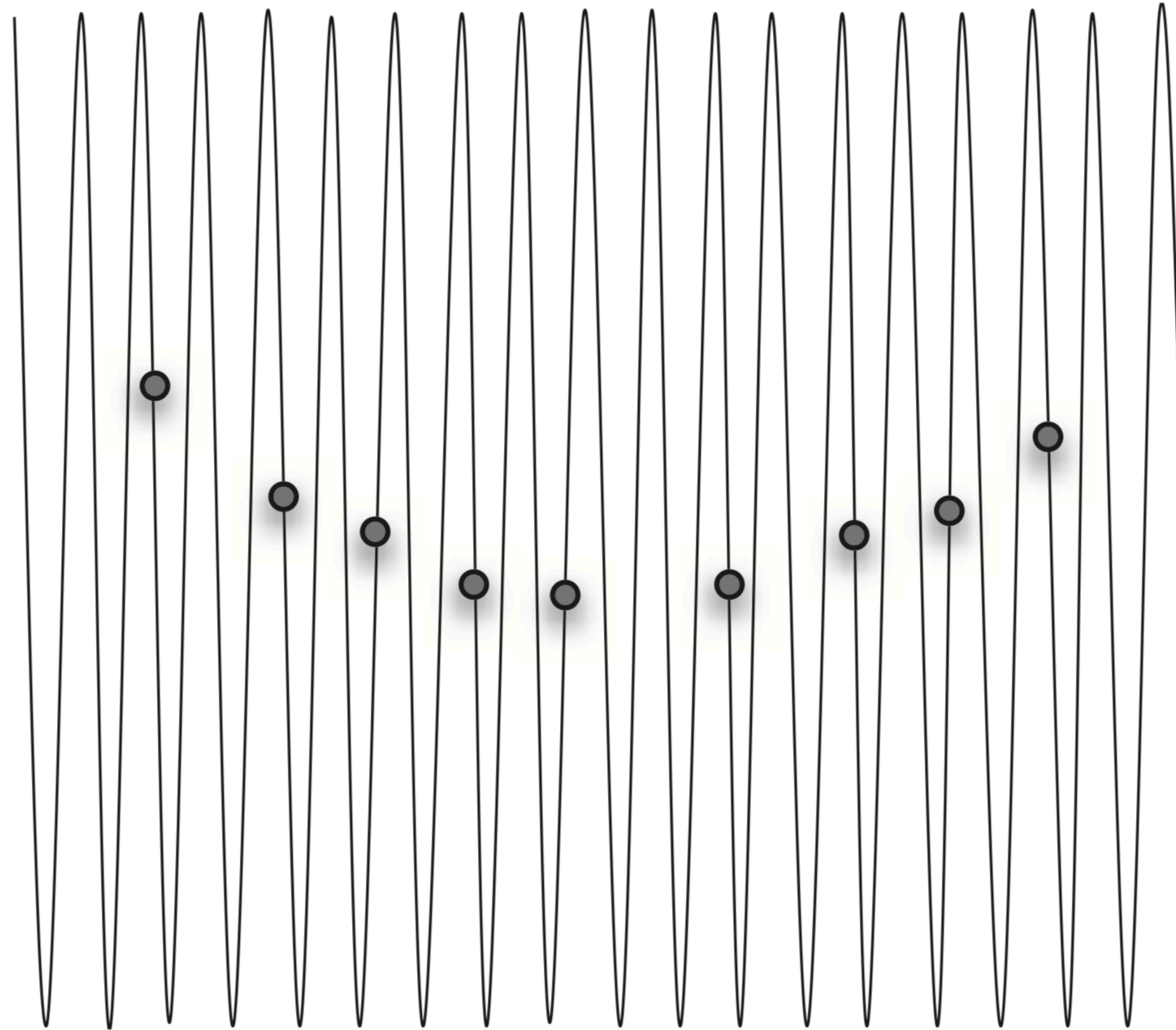
- A formula successful at predicting an individual’s weekly attendance to a movie should be discarded upon discovering that the subject just had a broken leg.
- However, most often we observe a failure of human defection strategy.
- Only for statistical models for which we have a theoretical explanation of their success, a human expert can apply her additional theoretical knowledge, and defecting from the model can be successful.
- In that case, the decision is often based on additional cues currently unknown to the statistical model.

Universal Consistency

- A method is universally consistent when, for any background distribution, as more data are obtained, the expected error of the learned model approaches the expected error of the best rule.
- When the VC dimension is infinite, no method can guarantee a rate of convergence.
- For example, with n the amount of data, a \sqrt{n} -nearest-neighbors model is universally consistent.

Inductive Bias and Structural Risk Minimization

- A kind of universally consistent method with a trade-off between empirical adequacy to available data and another factor, the “simplicity”.
- For a class of rules $C = C_1 \cup C_2 \cup \dots \cup C_n \cup \dots$ where $C_1 \subset C_2 \subset \dots \subset C_n \subset \dots$, SRM is to minimize a function of both the empirical error of the rule on the data and the VC dimension of the smallest class containing the rule.
- Cross-validation is another way to mitigate overfitting.
- “Simplicity” is hard to define. For example, measuring it by the number of parameters (as Popper did) is not satisfying. Take for example, the class of sine curves $y = a \sin(bx)$. Is it simple because it has two parameters?



“Reliable Reasoning”, Figure 3.3 p. 72

Universal Approximation For Neural Network

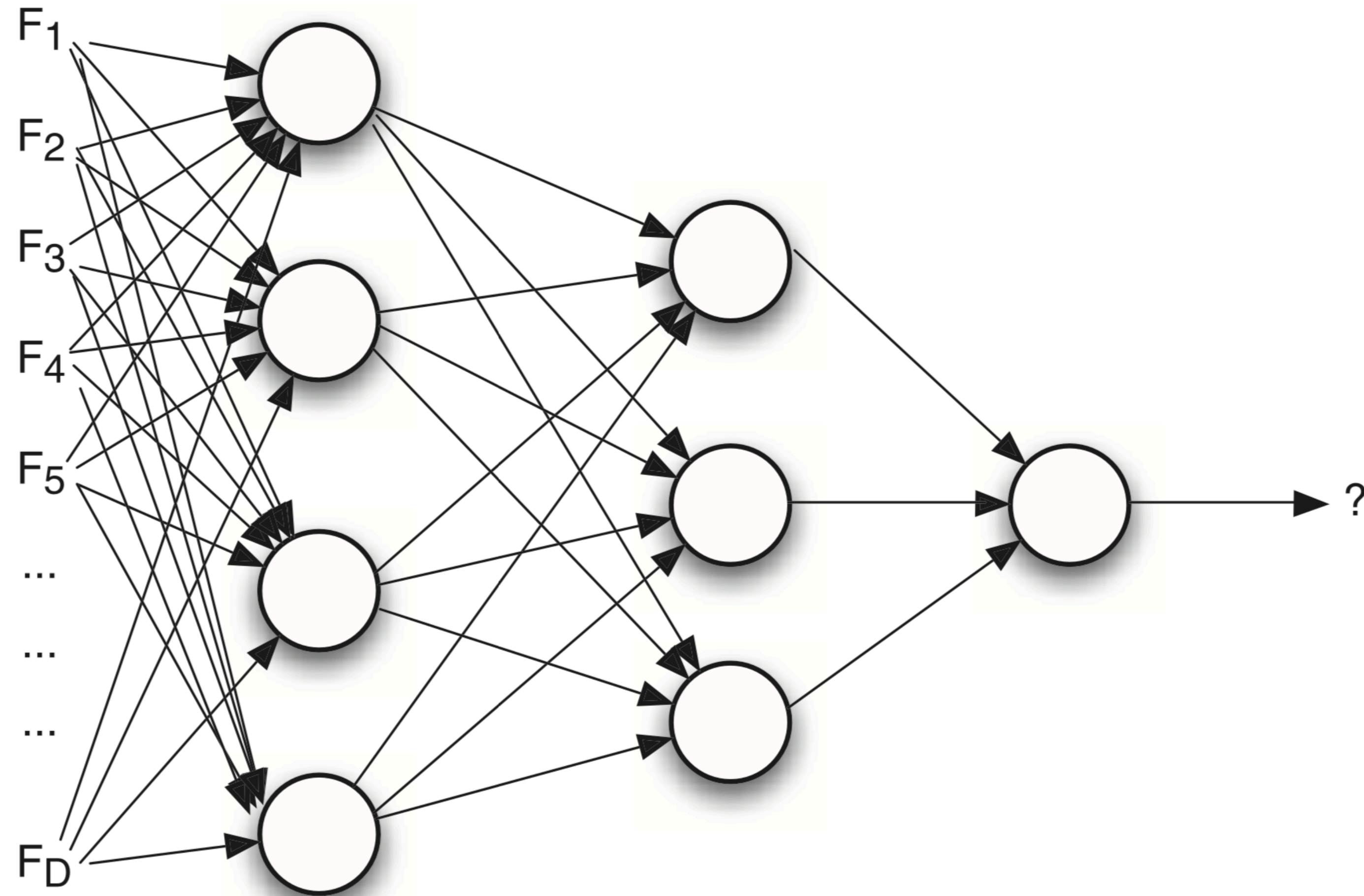


Figure 4.3

A feed-forward network.

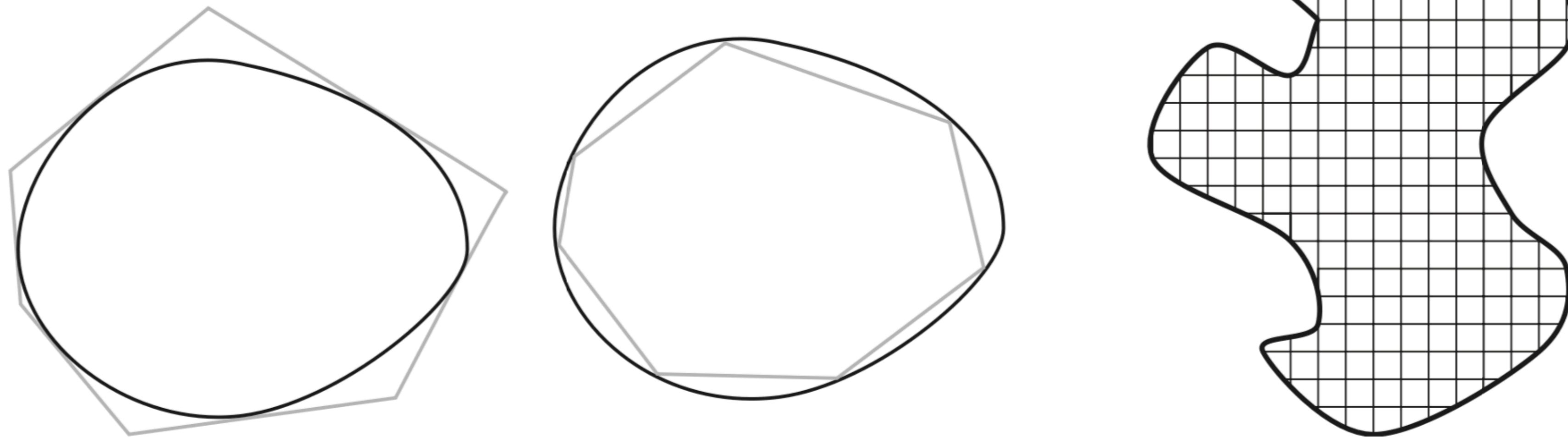


Figure 4.4

Approximating a convex hypervolume.

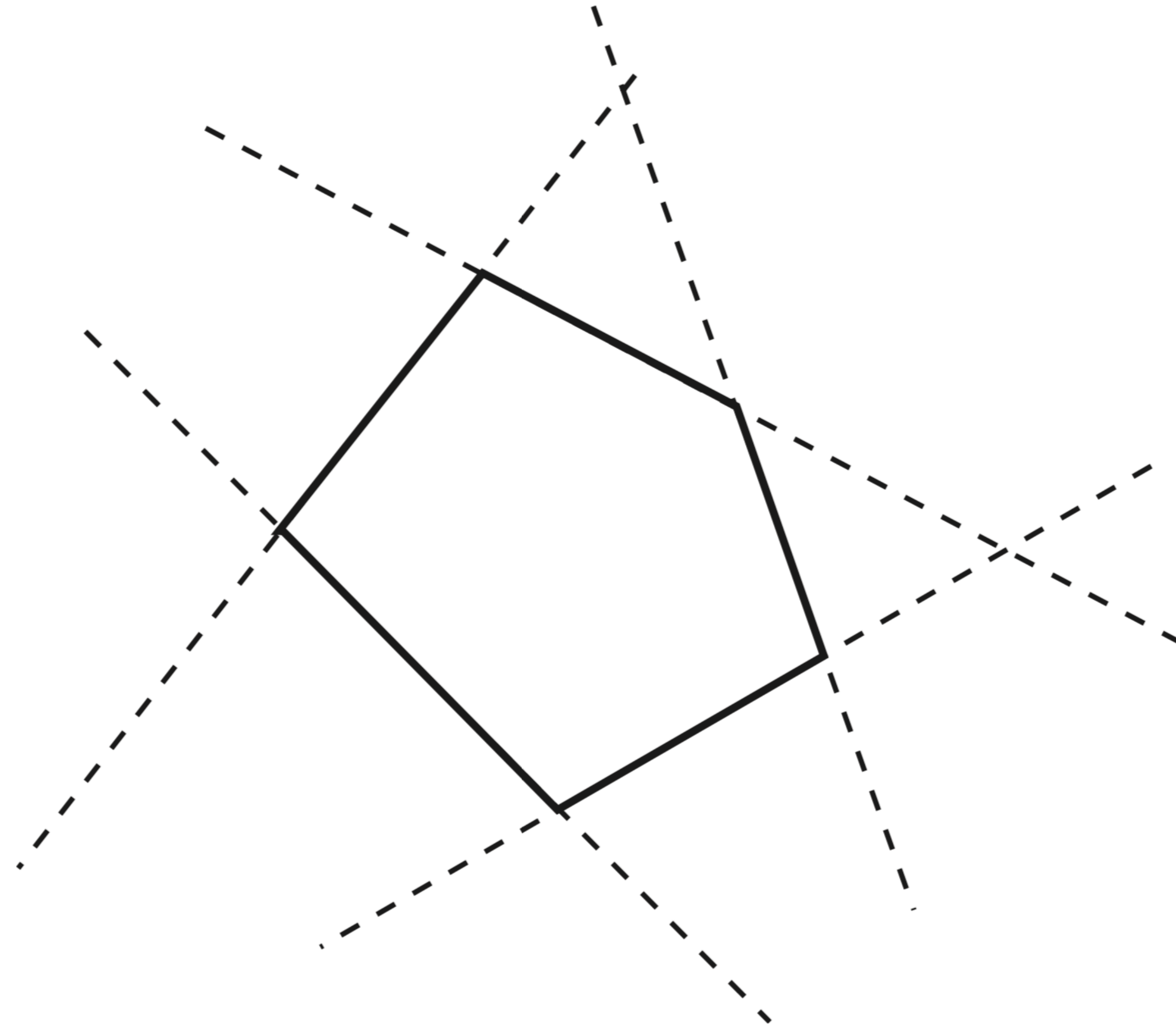


Figure 4.5
Intersecting half-spaces.

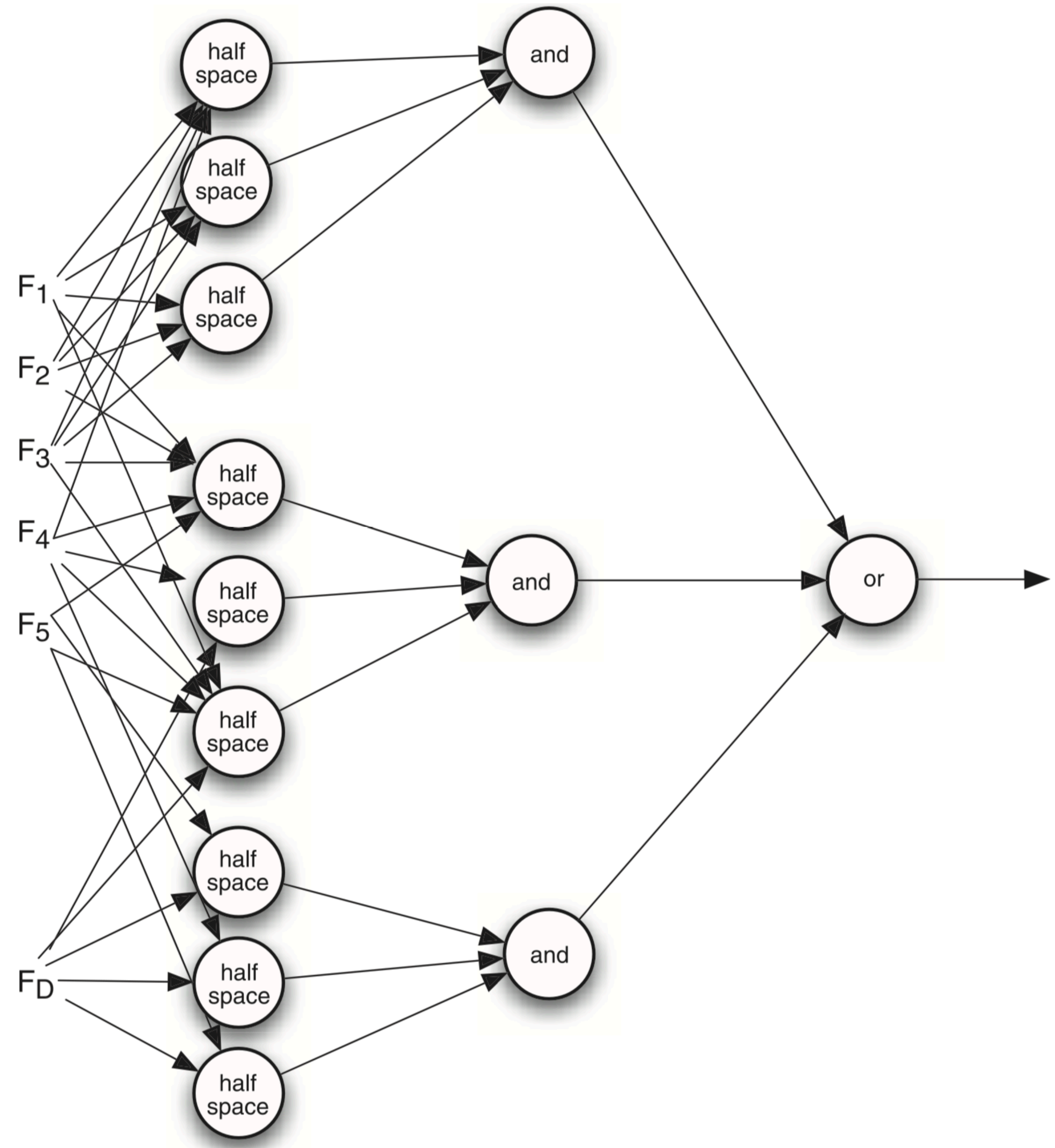


Figure 4.6
Taking a union of intersections.

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin 

Nature Machine Intelligence **1**, 206–215(2019) | [Cite this article](#)

Interpretability, a domain-specific notion

- Constraints in model form: monotonicity, giving preference to variables identified by domain experts...
- Sparsity: in particular for structured data, it allows for understanding how a handful of variables interact jointly.
- Incorporating application-specific constraints can lead to computationally hard problems.
- Making interpretable models requires specific skills for the data scientists.

Black-box models

- An metamodel “explanation” is an understanding of how the model works and not necessarily an explanation of how the world works.
- A metamodel can show trends in how predictions are related to the features.
 - It can be inaccurate in parts of the feature space.
 - It can reproduce accurately the predictions of the original model but using completely different features.
- It can be difficult to fine-tune the importance given to contextual information.
- The sensibility to “noise” of black box models can be difficult to assess.
- There can be an incentive to monetize a black box model.
- Counterfactual explanations may not be sufficient.

Trade-off between accuracy and interpretability

A myth?

- For problems with structured data and meaningful features, there is often no difference in performance between more complex classifiers and simpler ones, after preprocessing.
- The standard process for knowledge discovery (KDD, CRISP-DM...) is more essential than the differences between algorithms.
- Uninterpretable methods can provide baseline levels of performance.

Algorithmic challenges in interpretable ML

- Falling Rule Lists, with Bayesian learning.
- Optimized risk scores, with Integer Linear Program (ILP) solvers.
- Generalised Additive models (GAM).
- Symbolic regression, with genetic programming, simulated annealing or even gradient descent given a recent parametrization.
- Interpretable deep-learning for case-based reasoning with prototypes.
- Etc.

[Submitted on 3 Dec 2020]

Interpretability and Explainability: A Machine Learning Zoo Mini-tour

Ričards Marcinkevičs, Julia E. Vogt







In this review, we examine the problem of designing interpretable and explainable machine learning models. Interpretability and explainability lie at the core of many machine learning and statistical applications in medicine, economics, law, and natural sciences. Although interpretability and explainability have escaped a clear universal definition, many techniques motivated by these properties have been developed over the recent 30 years with the focus currently shifting towards deep learning methods. In this review, we emphasise the divide between interpretability and explainability and illustrate these two different research directions with concrete examples of the state-of-the-art. The review is intended for a general machine learning audience with interest in exploring the problems of interpretation and explanation beyond logistic regression or random forest variable importance. This work is not an exhaustive literature survey, but rather a primer focusing selectively on certain lines of research which the authors found interesting or informative.

Subjects: **Machine Learning (cs.LG)**

Cite as: [arXiv:2012.01805](https://arxiv.org/abs/2012.01805) [cs.LG]

Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning



Authors:  [Harmanpreet Kaur](#),  [Harsha Nori](#),  [Samuel Jenkins](#),  [Rich Caruana](#),  [Hanna Wallach](#),
 [Jennifer Wortman Vaughan](#) [Authors Info & Affiliations](#)

Publication: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems • April 2020

• Pages 1–14 • <https://doi.org/10.1145/3313831.3376219>

ML as an engineering discipline

- Data scientists cannot debug their models if they do not understand their behavior.
- Interactive ML (iML), a new challenge in Human Computer Interaction (HCI).
- How to evaluate interpretable or explainable models? Very few user studies.

Common issues faced by Data Scientists

As captured by conducting pilot interviews

- Missing values
- Changes in data over time (e.g., new categories for an existing feature)
- Duplicate data
- Redundant features
- Ad-hoc categorization of continuous features.
- Debugging difficulties. Identifying potential model improvements based on a small number of data points is difficult.

Contextual inquiry

- 11 participants
- “Adult” dataset, 1994 US census data, a data point is a person, the features are her age, education, marital status, native country, occupation, etc. The label is binary: income >\$50k.
- 2 interpretability tools: GAM or the SHAP Python package.
- Each participant used only one tool selected at random.
- Dataset modified to evaluate how the common issues faced by data scientists are managed in practice.
- Conclusion: useful (e.g., for identifying missing values) but over-trust and misuse. No deep understanding of the visualizations. A bias toward model deployment.

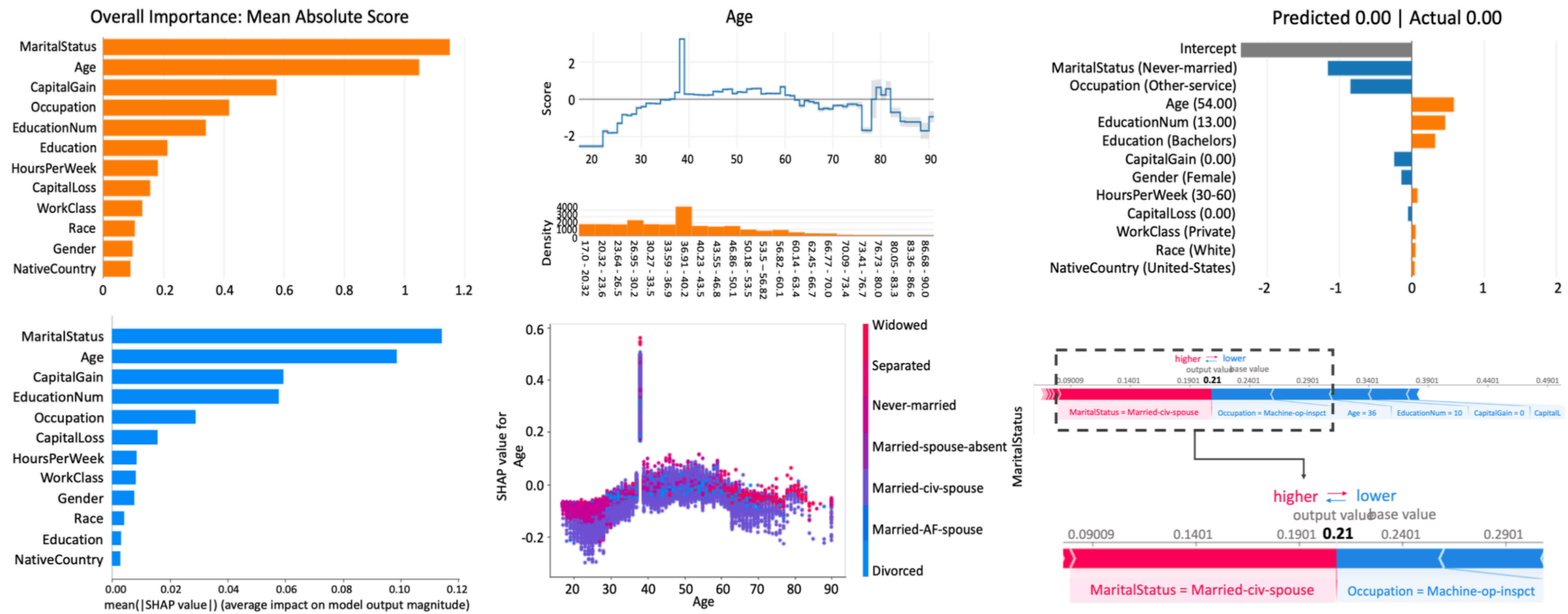


Figure 1: Visualizations output by the InterpretML implementation of GAMs (top) and the SHAP Python package (bottom). Left column: global explanations. Middle column: component (GAMs) or dependence plot (SHAP). Right column: local explanations.

Thank you!