Trespassing Random Forests

with a pointed stick for self defence

Kainat Khowaja

Wolfgang Karl Härdle

IRTG 1792 High Dimensional
Non-Stationary Time Series
Humboldt-Universität zu Berlin
IRTG1792.HU-Berlin.de

The fable of bundle of sticks — reversed

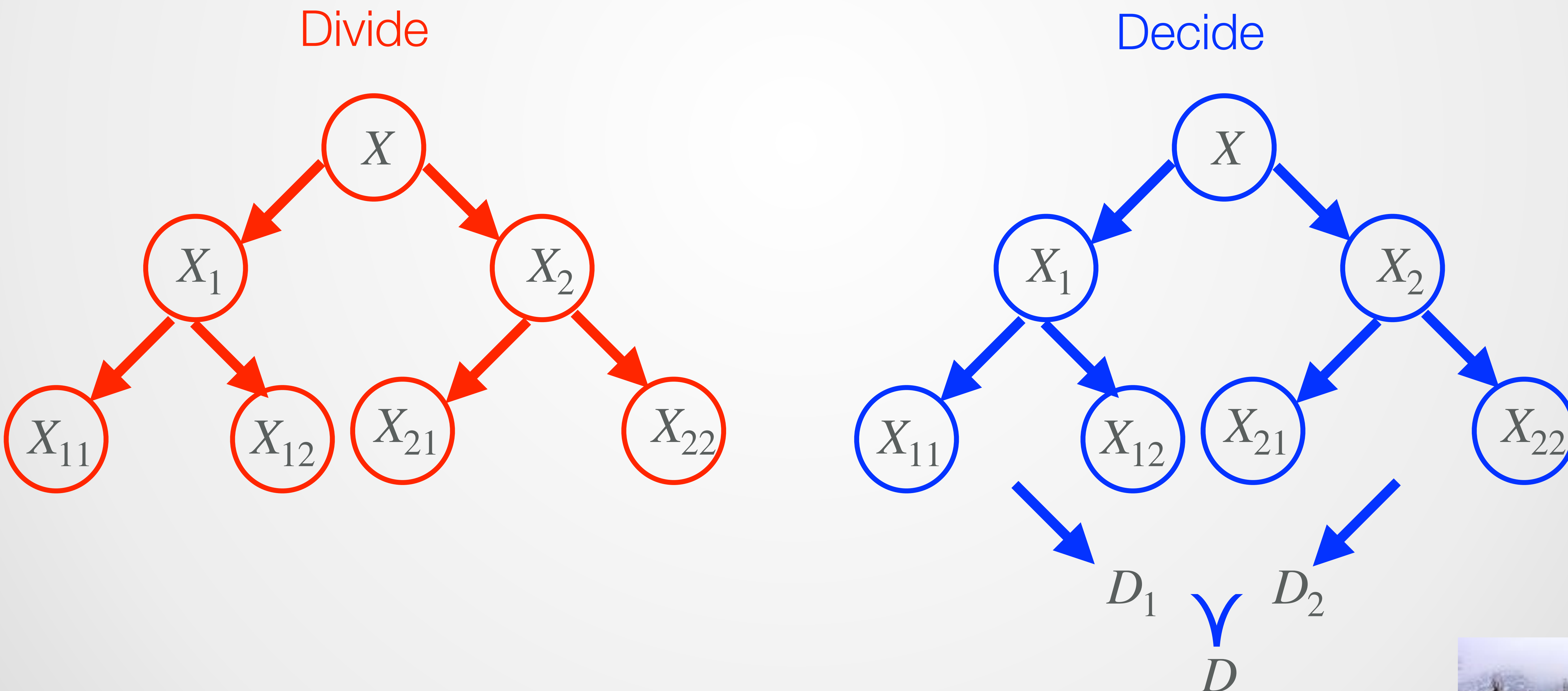
A 17th century illustration of the fable by Jacob Gole from Pieter de la Court's Sinryke Fabulen

A bunch of sticks is difficult to break

So break one stick at a time

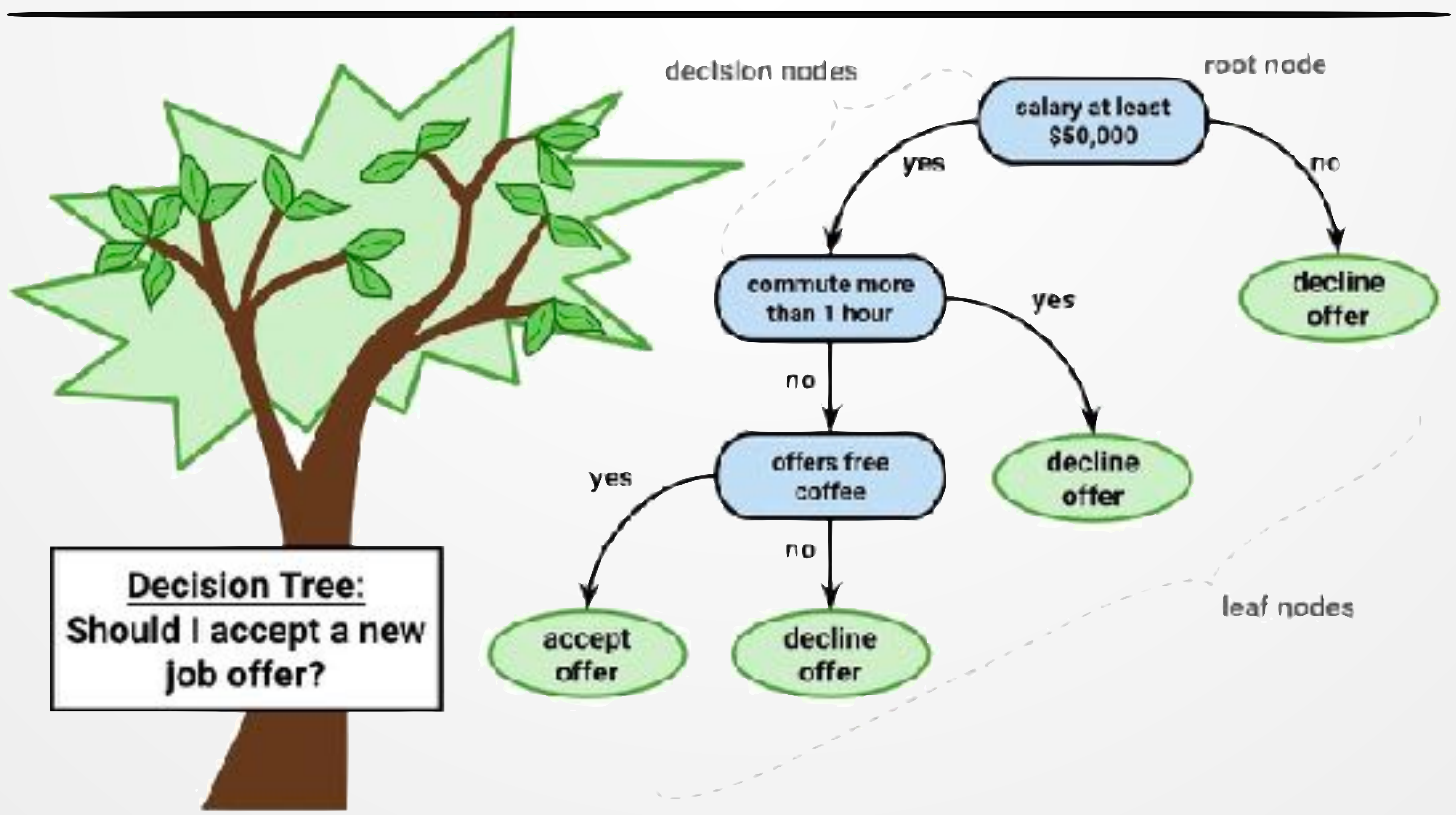
Divide et decide!

- We need to make sure they do not all just learn the same
- $X \succ$ Data, $D \succ$ Decision

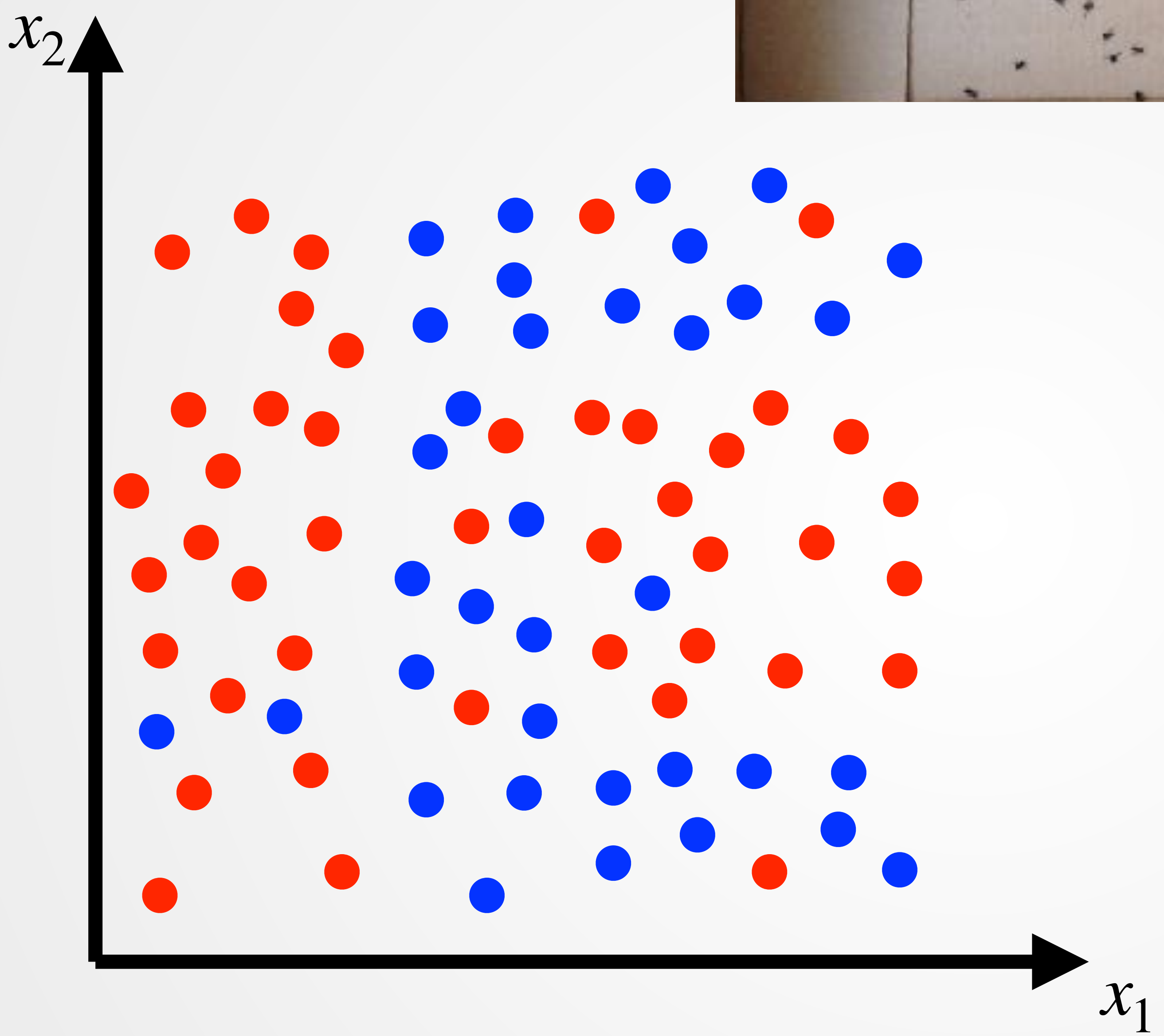


Random Forests (RF)

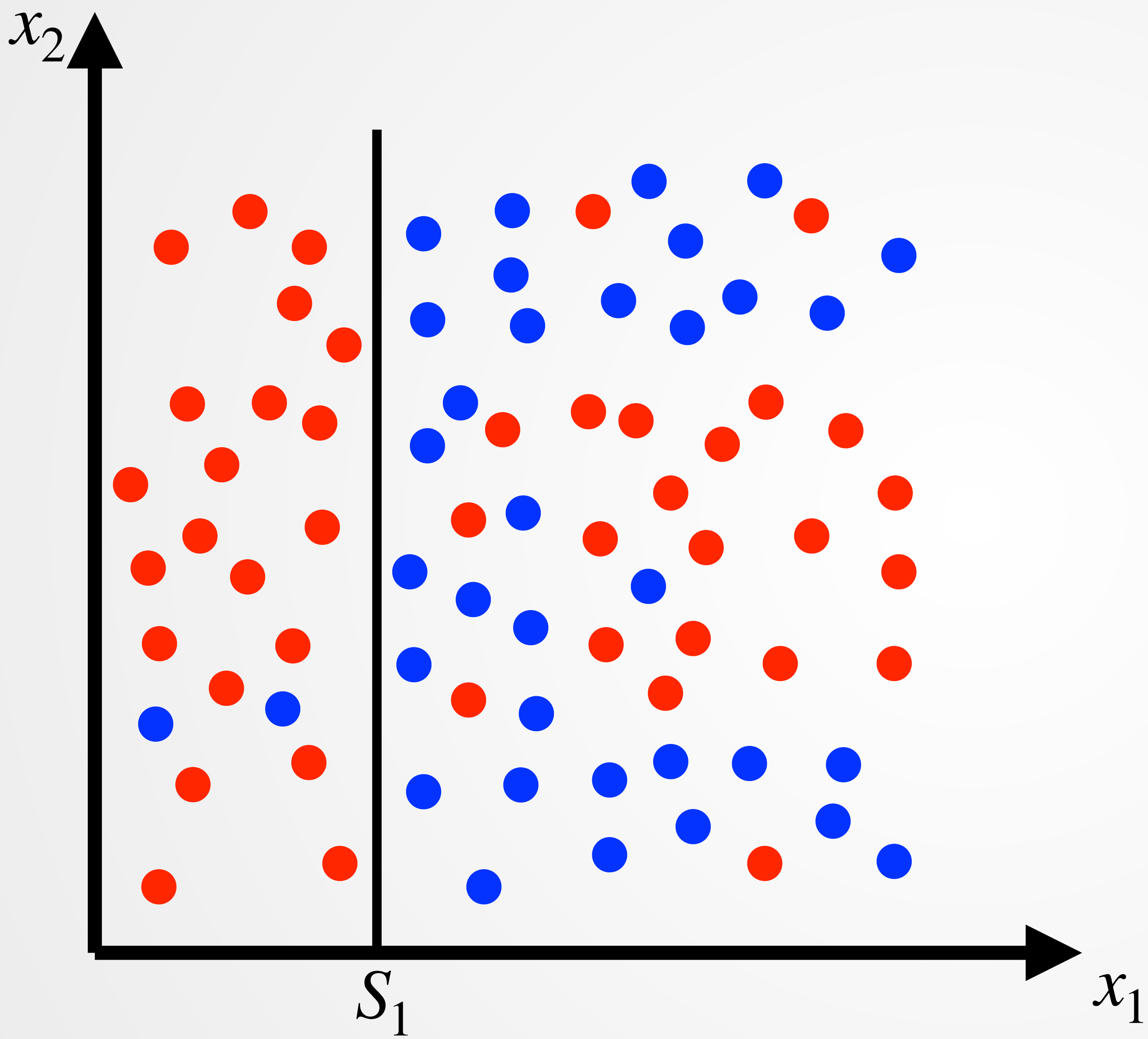
- Leo Breiman 2001
- Supervised learning for classification and regression
- Divide and decide (conquer)
- Ensemble Method which grows trees as base learners
- Combines randomised **decision trees, aggregates** the prediction

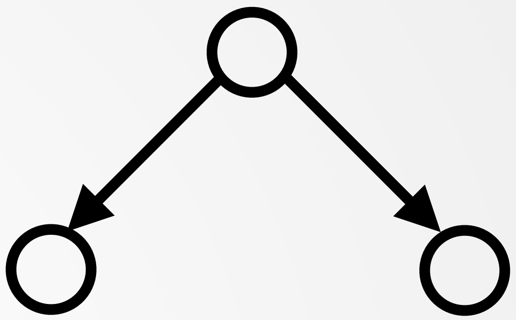


Decision Trees

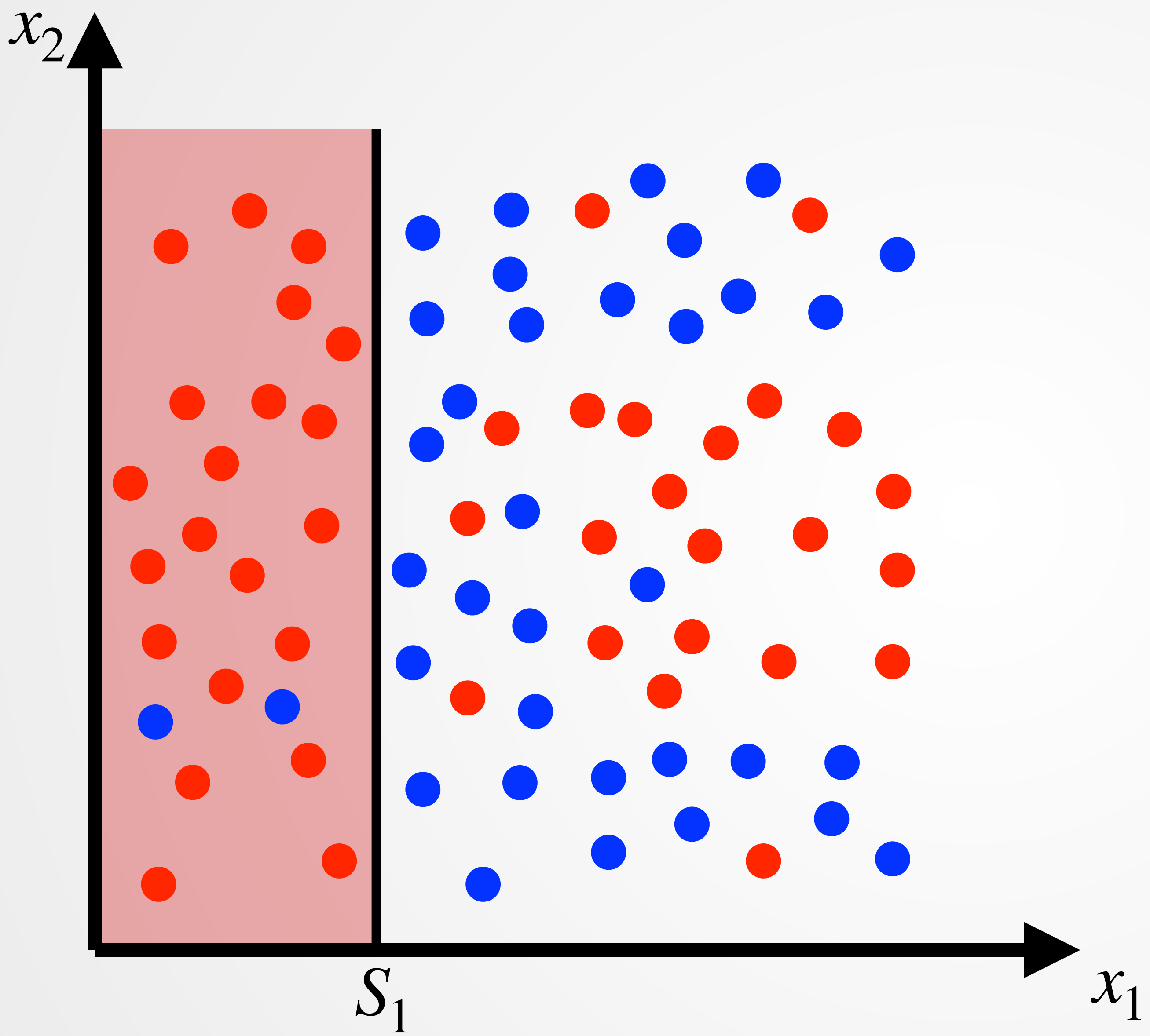


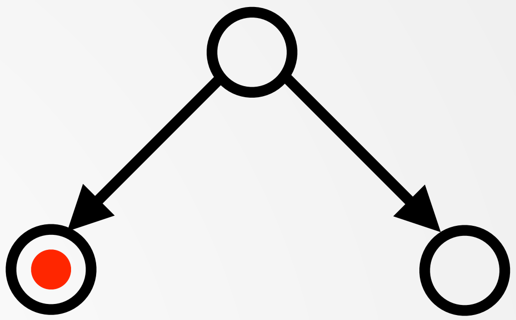
Decision Trees



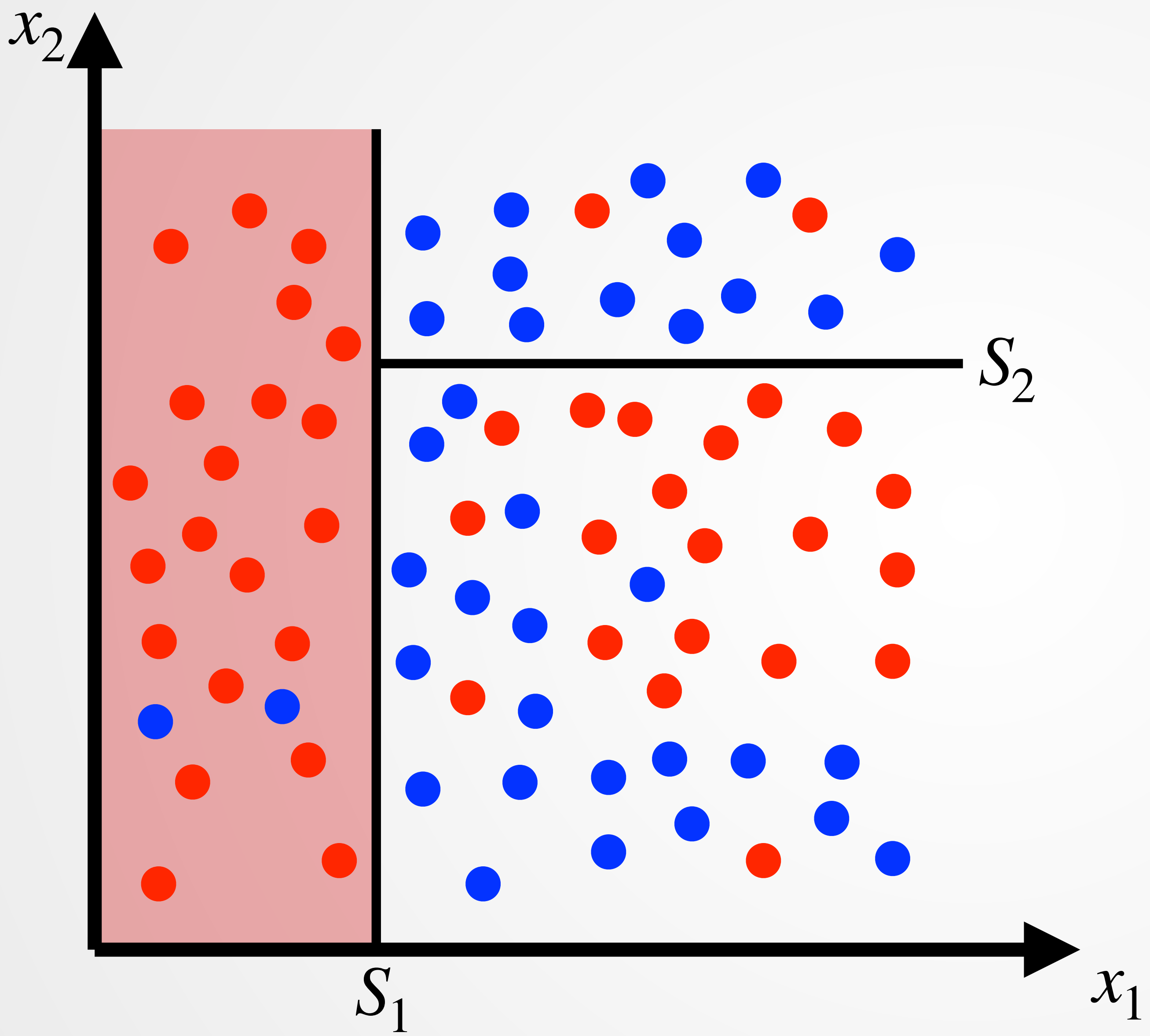


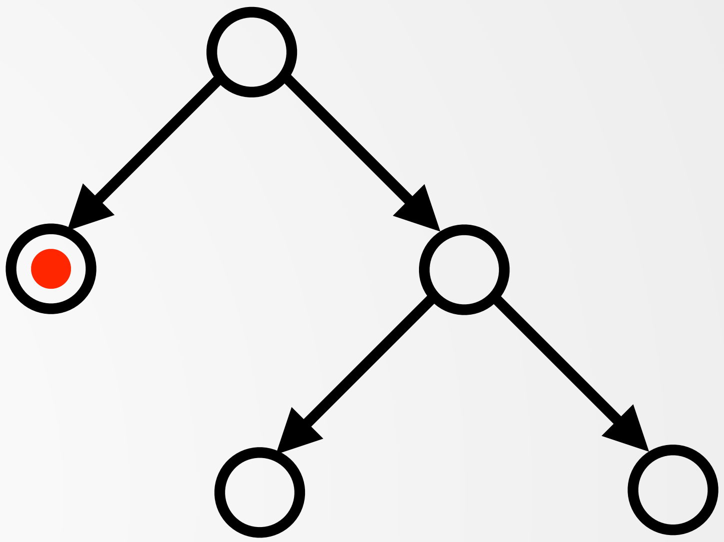
Decision Trees



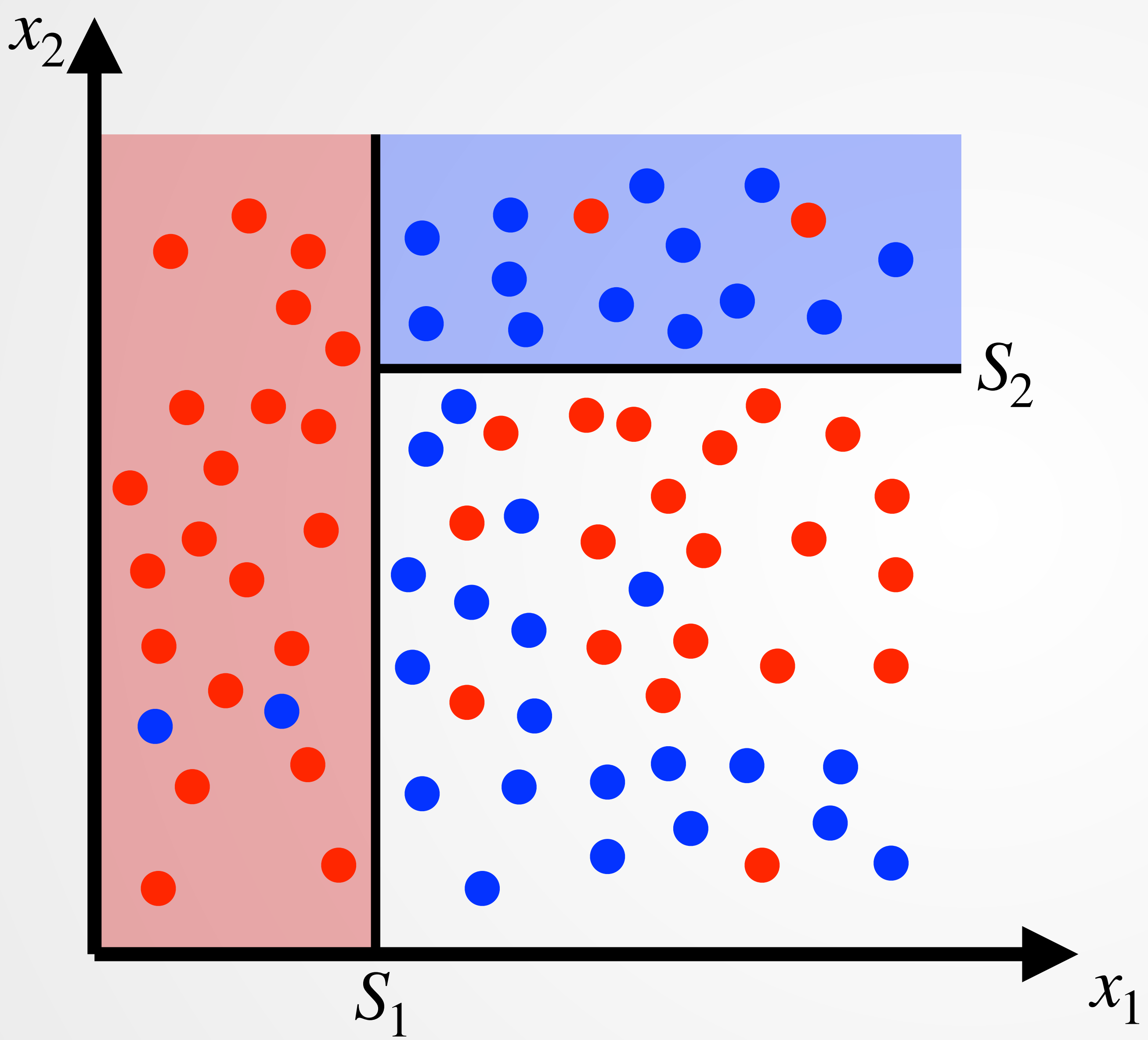


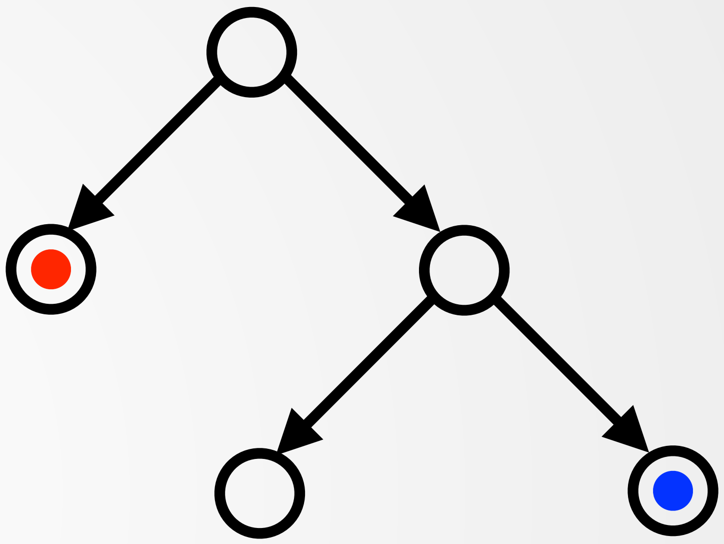
Decision Trees



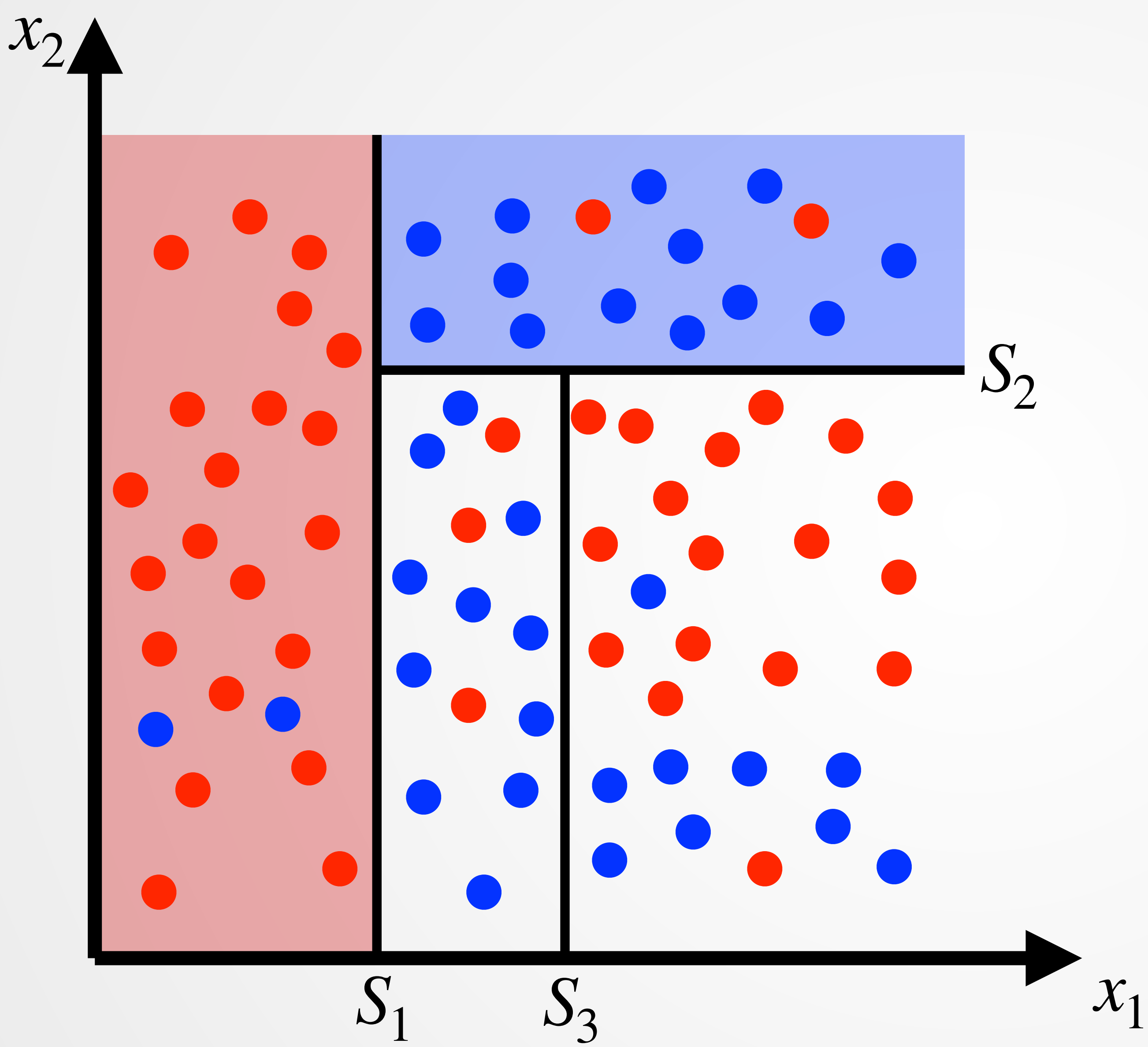


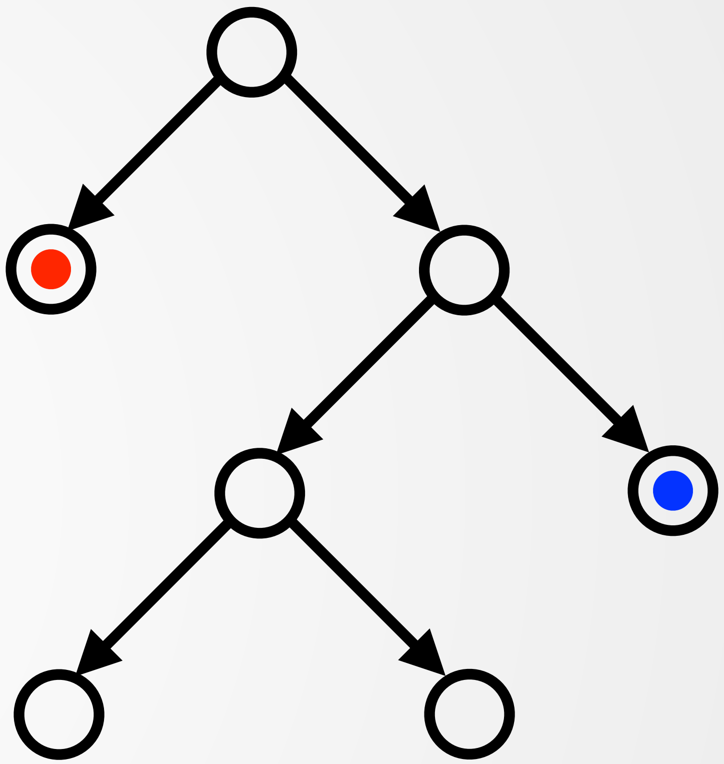
Decision Trees



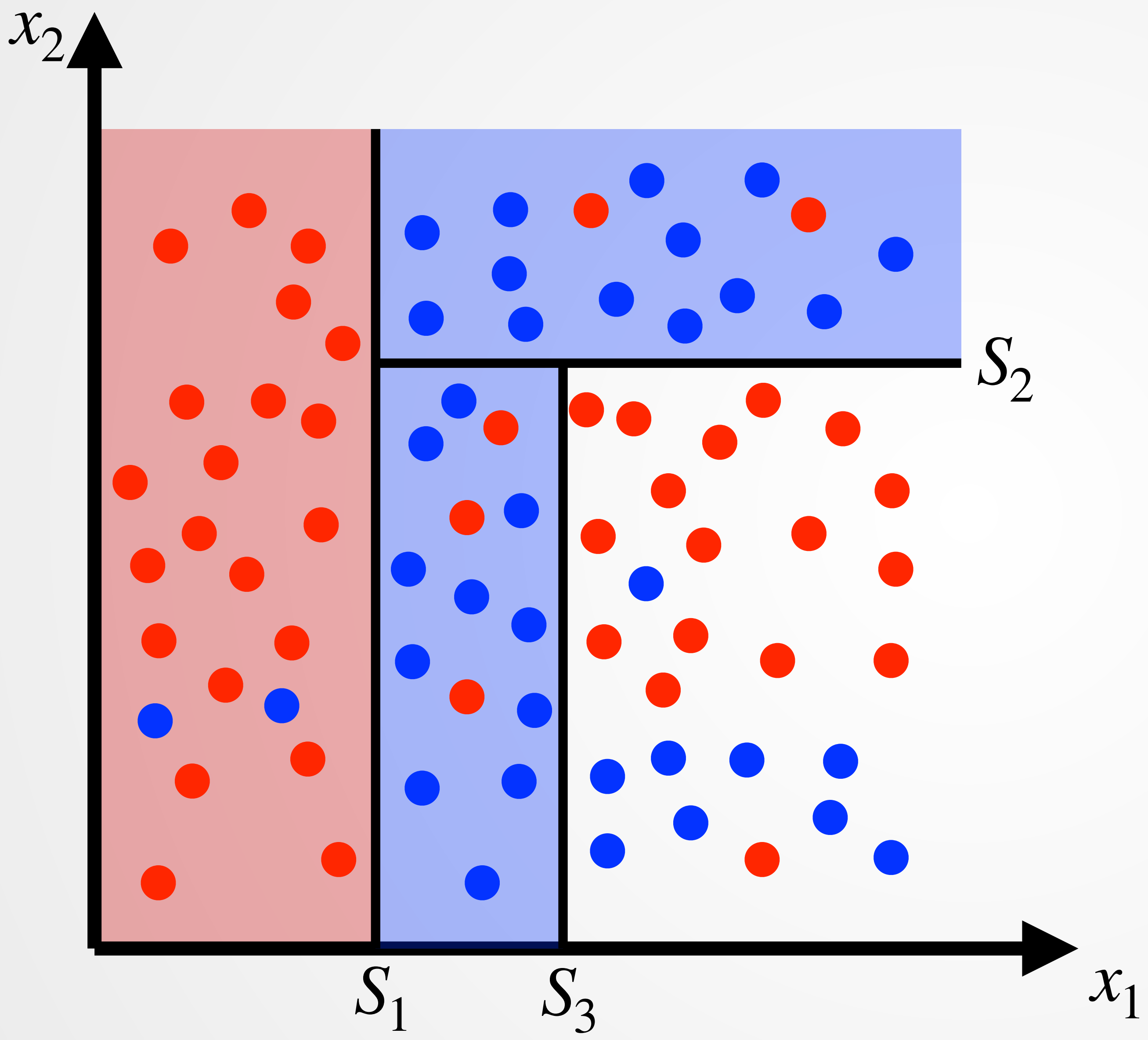


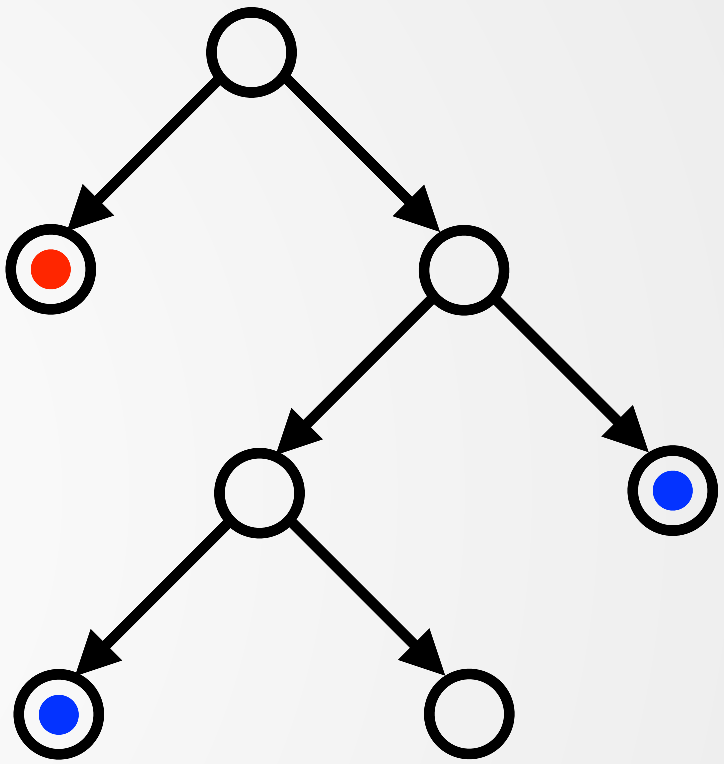
Decision Trees



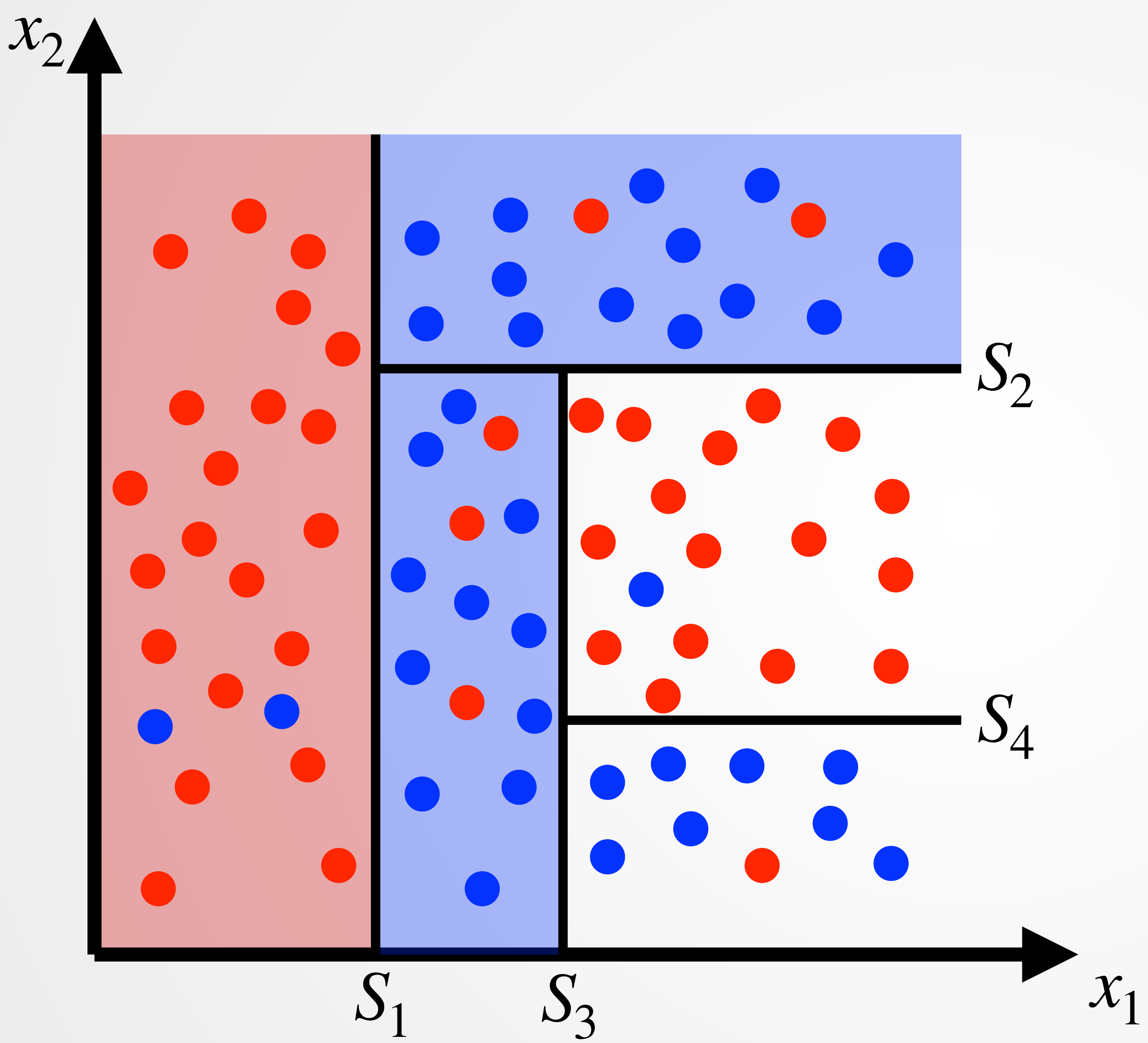


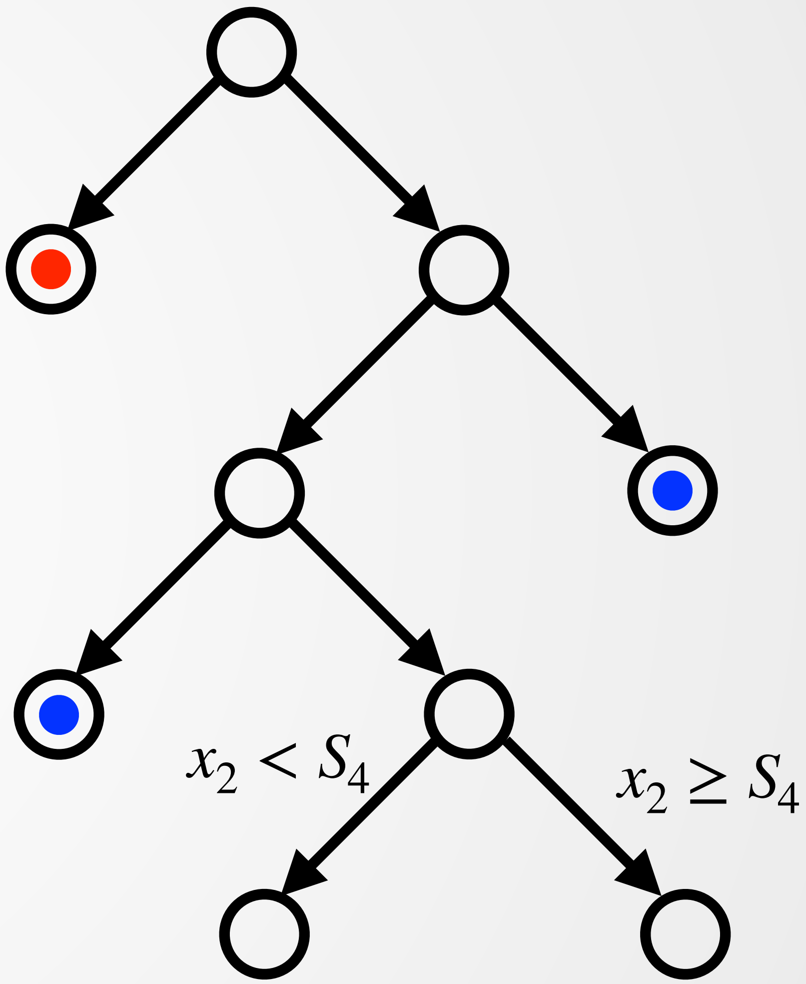
Decision Trees



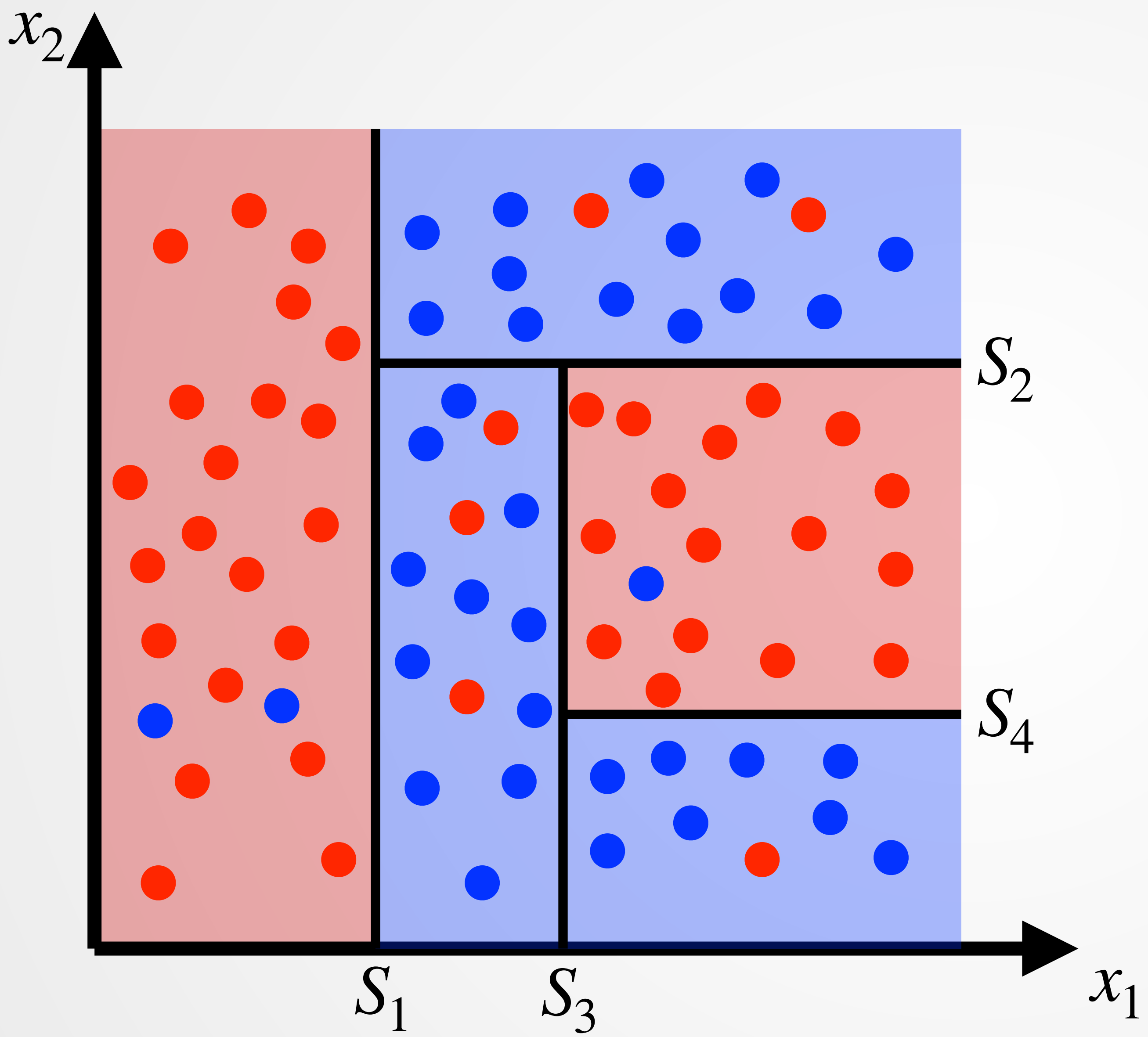


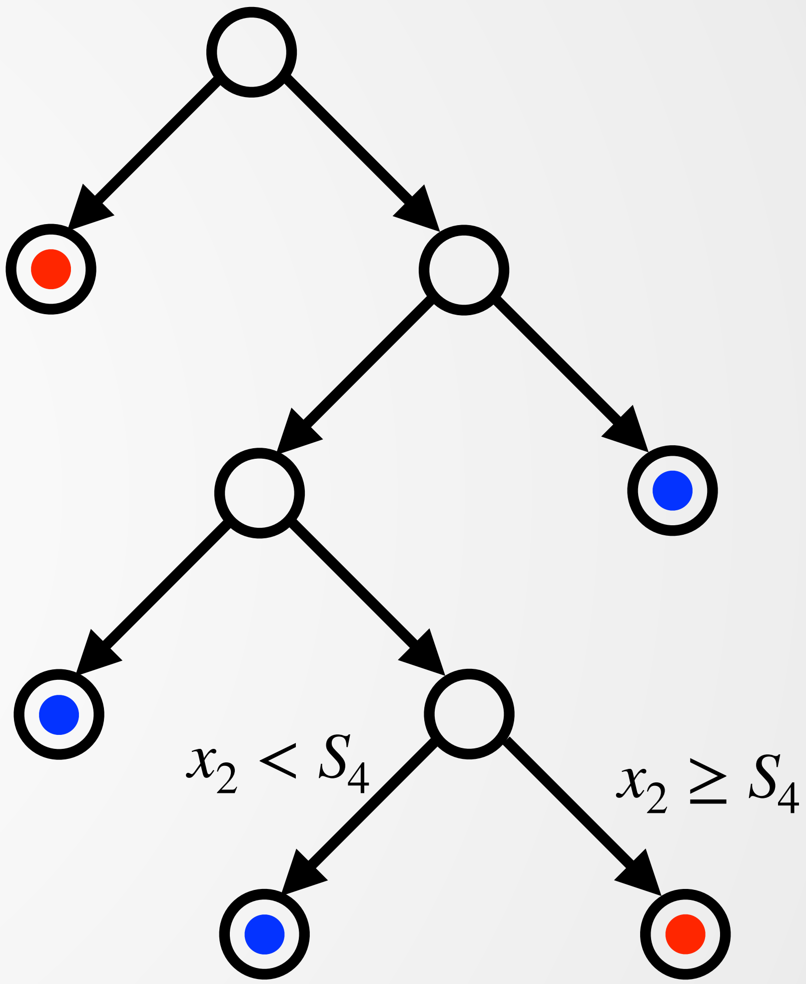
Decision Trees





Decision Trees





Decision Trees

- Let $n(m) = \#\{x_i \in R^{(m)} \,|\, i = 1,\ldots,n\}$ be the # of obs in region $R^{(m)}$
- Defining the classification accuracy of node m classifying class k:

$$\hat{p}_k^{(m)} = \frac{1}{n(m)} \sum_{(x_i,y_i)\in R^{(m)}} \mathbf{I}(y_i = k),$$

- Possible choices for the impurity measure:
  - Misclassification error:

$n(m_L)$ # of pts on the left side

$$Q^{(m)}(T) = \frac{n(m_L)}{n(m)}\left(1 - \hat{p}_k^{(m_L)}\right) + \frac{n(m_R)}{n(m)}\left(1 - \hat{p}_k^{(m_R)}\right)$$

  - Gini index:

$$Q^{(m)}(T) = \frac{n(m_L)}{n(m)} 2\hat{p}_k^{(m_L)}\left(1 - \hat{p}_k^{(m_L)}\right) + \frac{n(m_R)}{n(m)} 2\hat{p}_k^{(m_R)}\left(1 - \hat{p}_k^{(m_R)}\right)$$

  - Cross-entropy:

$$Q^{(m)}(T) = -\sum_{m_i\in\{m_L,m_R\}} \frac{n(m_i)}{n(m)}\left\{\hat{p}_k^{(m_i)} \log \hat{p}_k^{(m_i)} + \left(1 - \hat{p}_k^{(m_i)}\right) \log\left(1 - \hat{p}_k^{(m_i)}\right)\right\}$$

    - Generally, no difference between Gini impurity and entropy wrt performance, see Raileanu and Stoffel

CART

- Classification And Regression Trees
- Decision Tree algorithms that are used for classification
- Regression trees for predictive modelling (very old pb)
- Choosing cuts perpendicular to the axes by optimising criteria
- Split criterion: Gini impurity (for classification) and prediction squared error (for regression)

Engel (1857)
Regressogram: step functions as approximations

Chakrabarty et al (2009)
Engel's Law reconsidered

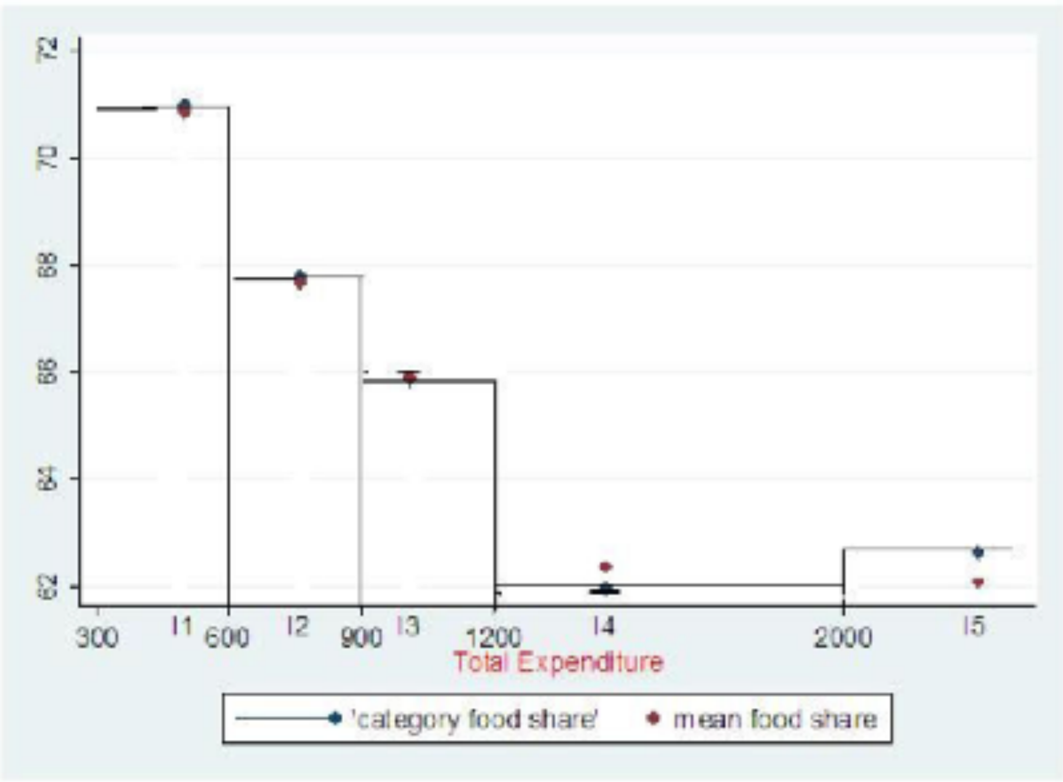
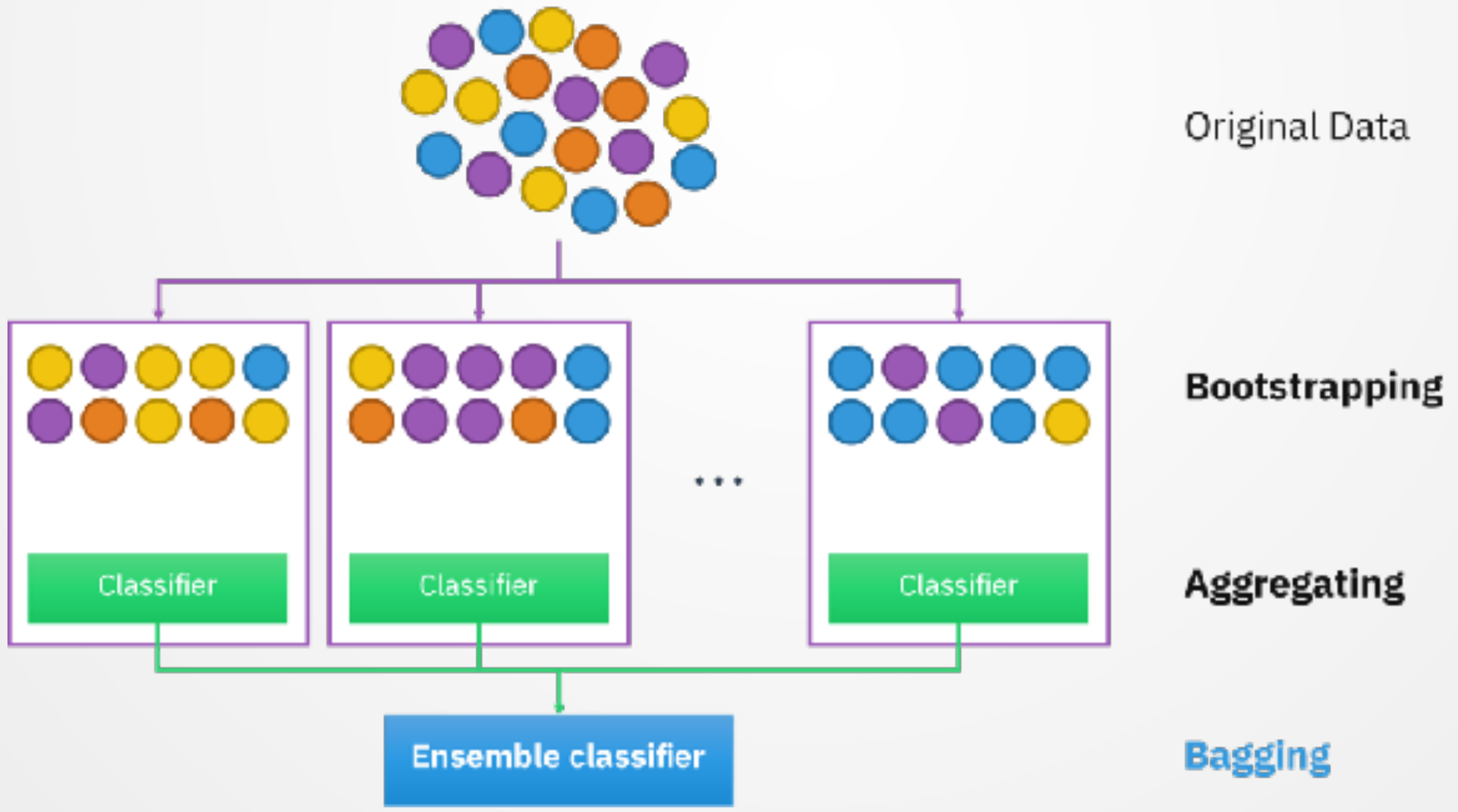


Figure 3.1d: Comparison of Regressogram and Engel's smoother

".. je aermer eine Familie ist, einen desto groesseren Antheil von der Gesamtausgabe muss zur Beschaffung der Nahrung aufgewendet werden ..."

Bagging

- Bootstrap aggregating
- Generate bootstrap samples from original dataset $S_1, S_2, \ldots, S_n$
- Construct predictor for each sample $P_1, P_2, \ldots, P_n$
- Decide by averaging $\dfrac{P_1 + P_2 + \ldots + P_n}{n}$



Algorithm

Algorithm 1: Breiman's random forest predicted value at $\mathbf{x}$.

Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1, \dots, n\}$, `mtry` $\in \{1, \dots, p\}$, `nodesize` $\in \{1, \dots, a_n\}$, and $\mathbf{x} \in \mathcal{X}$.

Output: Prediction of the random forest at $\mathbf{x}$.

```
for j = 1, ..., M do
    Select a_n points, with (or without) replacement, uniformly in D_n. In the
    following steps, only these a_n observations are used.
    Set P = (X) the list containing the cell associated with the root of the
    tree.
    Set P_final = ∅ an empty list.
    while P ≠ ∅ do
        Let A be the first element of P.
        if A contains less than nodesize points or if all X_i ∈ A are equal
        then
            Remove the cell A from the list P.
            P_final ← Concatenate(P_final, A).
        else
            Select uniformly, without replacement, a subset M_try ⊂ {1, ..., p}
            of cardinality mtry.
            Select the best split in A by optimizing the CART-split criterion
            along the coordinates in M_try (see text for details).
            Cut the cell A according to the best split. Call A_L and A_R the
            two resulting cells.
            Remove the cell A from the list P.
            P ← Concatenate(P, A_L, A_R).
        end
    end
    Compute the predicted value m_n(x; Θ_j, D_n) at x equal to the average of
    the Y_i falling in the cell of x in partition P_final.
end
Compute the random forest estimate m_{M,n}(x; Θ_1, ..., Θ_M, D_n) at the query
point x according to (1).
```

Bootstrapping (1st randomization)

(2nd randomization)

CART

Aggregating

$$m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n). \quad (1)$$

Advantages

Howard (Kaggle) and Bowles (Biomatica) *’ensemble of decision trees (random forests) have been the most successful general-purpose algorithm in modern times ’*

- Performs well when # variables exceeds # of observations
- Very few parameters to tune
- Can be applied to large scale problems/ high dim feature spaces
- Easily adaptable to ad-hoc learning tasks and return measures
- High accuracy
- Easily parallelizable

Outline

1. Introduction ✔
2. Babylon
3. Trespassing Random Forests
4. Pointed Sticks for self defence
5. What to do next?

12. Philosophy. Breiman passionately believed that statistics should be motivated by problems in data analysis. Comments such as

> If statistics is an applied field and not a minor branch of mathematics, then more than 99% of the published papers are useless exercises. [Breiman (1995b)]

Theory speaks and Practice follows ?

"Despite their widespread use, a gap remains between the theoretical understanding of random forests and their practical performance. This algorithm, which relies on complex data-dependent mechanisms, is difficult to analyze and its basic mathematical properties are still not well understood.
As observed by Denil et al. (2014), this state of affairs has led to polarization between theoretical and empirical contributions to the literature. Empirically focused papers describe elaborate extensions to the basic random forest framework but come with no clear guarantees. In contrast, most theoretical papers focus on simplifications or stylized versions of the standard algorithm, where the mathematical analysis is more tractable."

Biau G. and Scornet E. (2016)

Random Forests Lingo

- CART
- Bagging
- Neural Random Forests
- Purely RF/ central RF
- Median RF
- Quantile RF
- Generalized RF
- Dynamic RF
- Local linear forests

https://www.historyextra.com/period/ancient-history/babylon-babylonia-tower-babel-hanging-gardens-hammurabi/

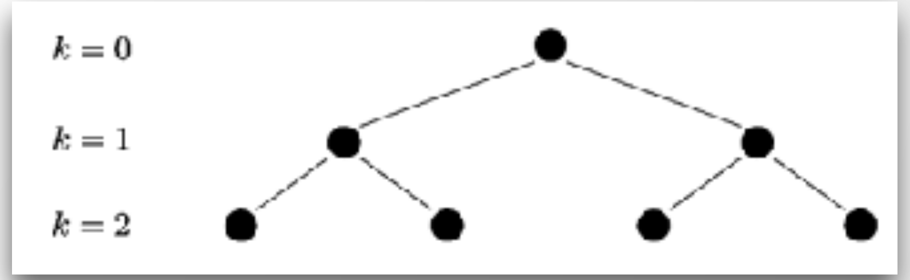
Purely RF (Breiman 2001)

- A family of simplified models
- Basic framework for accessing theoretical properties

Model (RF)

- The root $\mathscr{X}$
- At each leaf
  - choose *mtry* variables uniformly
  - Find the best split using CART, data dependent

Model (PRF)



- The root $\mathscr{X} = [0,1]^d$
- Select smoothness parameter k
- Repeat $k \in \mathbb{N}$ times (k controls the size of terminal node)
  - Randomly choose a node, to be split, uniformly among all terminal nodes. Randomly choose split variable
  - Randomly choose split point - data independent

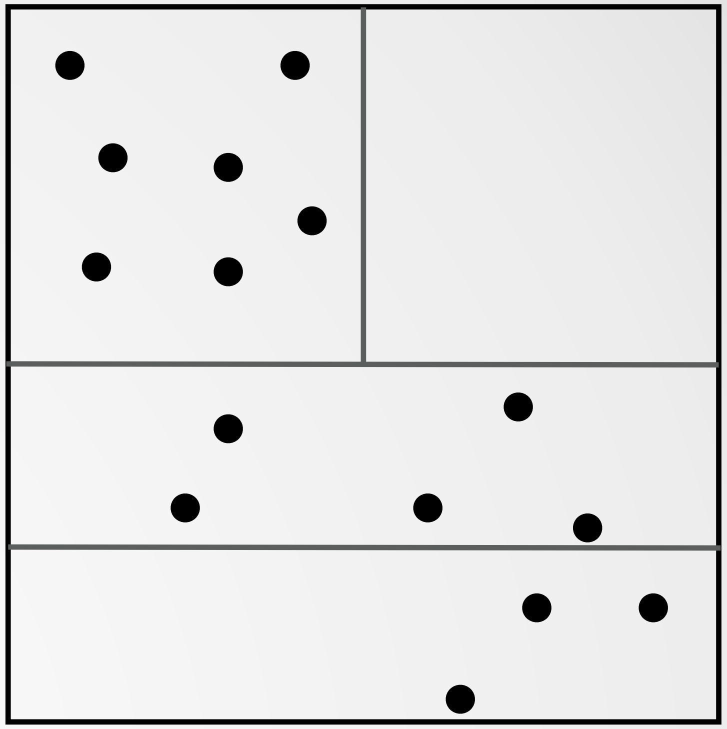
Purely uniform RF (Genuer 2012)

- An alternative to PRF
- For $d = 1$

Model (PURF)

- The root $\mathcal{X} = [0,1]$
- Select smoothness parameter k controlling the size of terminal node
- Repeat $k \in \mathbb{N}$ times (for tree with level k)
  - Randomly choose a node, to be split, uniformly among all terminal nodes
  - Randomly choose split variable
  - Randomly choose split point - data independent
- Consistent under Lipschitz assumptions (Genuer 2012)

Centered forests (Breiman 2004)



- Example of PRF
- Independent of whole data

Model (Centered forests)

- The root $\mathscr{X} = [0,1]^d$
- No resampling step
- Select smoothness param k and repeat k times
  - Randomly choose a node, to be split, uniformly among all terminal nodes
  - Randomly choose split variable
  - Split in the centre
- Each tree ends up with 2^k leaves
- Consistent as $k \to \infty$ and $\frac{n}{2^k} \to \infty$ (Scornet 2015)

Median RF (Devroye et al. 1996)

- Good trade-off between CRF and Breiman's RF
- Independent of response variable

Model (MRF)

- The root $\mathscr{X} = [0,1]^d$
- No resampling step
- Repeat until there is only one observation in each cell
  - Randomly split a node, uniformly among all terminal nodes
  - Randomly choose split variable
  - Split in the empirical median of data in the cell
- In general not consistent (Györfi et al. 2002)

observations without replacement among the original sample

- If $a_n \to \infty$ and $\dfrac{a_n}{n} \to 0$, then median RF are consistent even though individual trees are not (Scornet 2016)

Orthogonal decision trees (Kargupta et al. 2006)

- Way to construct redundancy free decision trees
- Trees are functionally orthogonal to each other and correspond to PC of underlying functional space

Model (ODT)

- The root $\mathscr{X}$
- Construct Fourier spectrum of the tree (algebraic representation of the trees)
- Perform Eigenanalysis and PCA
- Convert PCs to trees in original space
- Apply RF algorithm on these trees

Quantile regression forests (Meinshausen 2006)

- Estimates conditional quantiles instead of conditional mean
- Computes the whole conditional distribution of response var

Model (QRF)

- The root $\mathscr{X}$
- Select k, the tree level
- For each leaf of each tree
  - Note all observations (not just their average)
  - Split using CART
- Consistent for $\mathscr{X} = [0,1]^d$ with additional assumptions such as Lipschitz continuity of conditional cdf

Generalized RF (Athey et al. 2018)

- Estimates params that are identified via local moments condition
- Develops robust regression procedures via Huberization

Model (GRF)

- The root $\mathscr{X}$
- choose k and resampling rate
- Repeat k times
  - Labeling step $\succ$ calculate pseudo outcomes, define the forest-based adaptive neighborhood for each datapoint
  - Regression step $\succ$ Split using CART
- Consistent and asymptotically normal

Dynamic RF (Bernard et al. 2012)

- Unlike original RF where trees are uncorrelated, here trees are grown by taking into account the sub-forest already built
- Guides the tree induction so that each tree compliments the existing trees as much as possible
- Only reliable trees are allowed to grow in the forest
- Inspired by boosting (manipulates the importance through assigning weights)

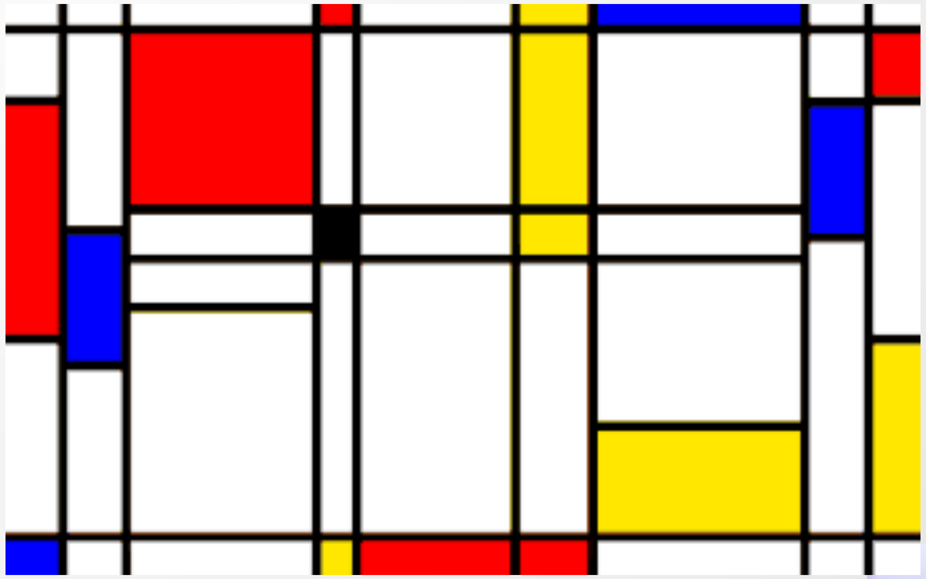
Model (DRF)

- The root $\mathscr{X}$
- choose k and resampling rate. Assign same weight to all training instances
- Repeat k times
  - Same as RF
  - Update the weights on class counts acc. to importance

Online RF (Saffari et al. 2009, Denil et al. 2013 ..)

- Do not require accessibility to entire training set at once
- Data incorporated in the model with time
- Trees are dropped from the forest based on performance and replaced by new ungrown trees
- Approximately: sample independent partitions $\Lambda_\lambda^{(1)}, \ldots, \Lambda_\lambda^{(M)} \sim MP(\lambda, [0,1]^d)$, fit them and average their partitions, where MP is Mondrian Process (Roy and Teh, 2008)
- Example: Mondrian forests, Information forests
- Proven to be consistent
- Choice of complexity param λ ?

Dirichlet in BBI



Consistency of PRF (base forests)

- Consider an estimate of the form

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i$$

Theorem (Stone, 1977). Weights W_{ni} non negative and sum to one. Then the estimate $m_n(x)$ is consistent $m(x) = \mathrm{E}[Y \mid X = x]$ iff

1. There is a constant C such that, for every measurable function $g : [0,1]^d \to \mathbb{R}$ with $\mathrm{E}|g(X)| < \infty$,
$$\mathrm{E}\sum_{i=1}^{n} [W_{ni}(X)\,|\,g(X_i)] \leq C\,\mathrm{E}|g(X)|, \text{ for all } n \geq 1$$
2. For all $a > 0$. $\sum_{i=1}^{n} W_{ni}(X)\mathbf{I}\{\|X_i - X\| > a\} \to 0$, in probability
3. $\max_{1 \leq i \leq n} W_{ni}(X) \to 0$, in probability

Stone conditions for RF

Choosing number of trees in a Mondrian forest

- Denote $m_{\lambda,M,n}$ the (randomized) Mondrian forest estimator with M trees and parameter λ Let

$$(*) \qquad \mathrm{Var}(Y|X) \leq \sigma^2 < \infty \text{ a.s.}$$

Theorem (Mourtada, Gaiffas). Assume (*) and that the regression function *m* is *l*-Lipschitz. Then:

$$\mathcal{R}(m_{\lambda,M,n}) \leq \frac{4dl^2}{\lambda^2} + \frac{(1+\lambda)^2}{n}(2\sigma^2 + 9\|m\|_\infty^2)$$

In particular, $\lambda = \lambda_n \approx n^{1/(d+2)}$ gives

$$\mathcal{R}(m_{\lambda,M,n}) = \mathcal{O}(n^{-2/(d+2)})$$

which is „Chuck's speed" for Lipschitz (*p*=1) „balls" in *d* dimensions

Chuck speed limiz

- True for every $M \geq 1$. But in practice more trees perform better, why? How to choose M?

Neural Random Forests

- Perceptron to the rescue
- this CART tree is actually a 2 layer NN !

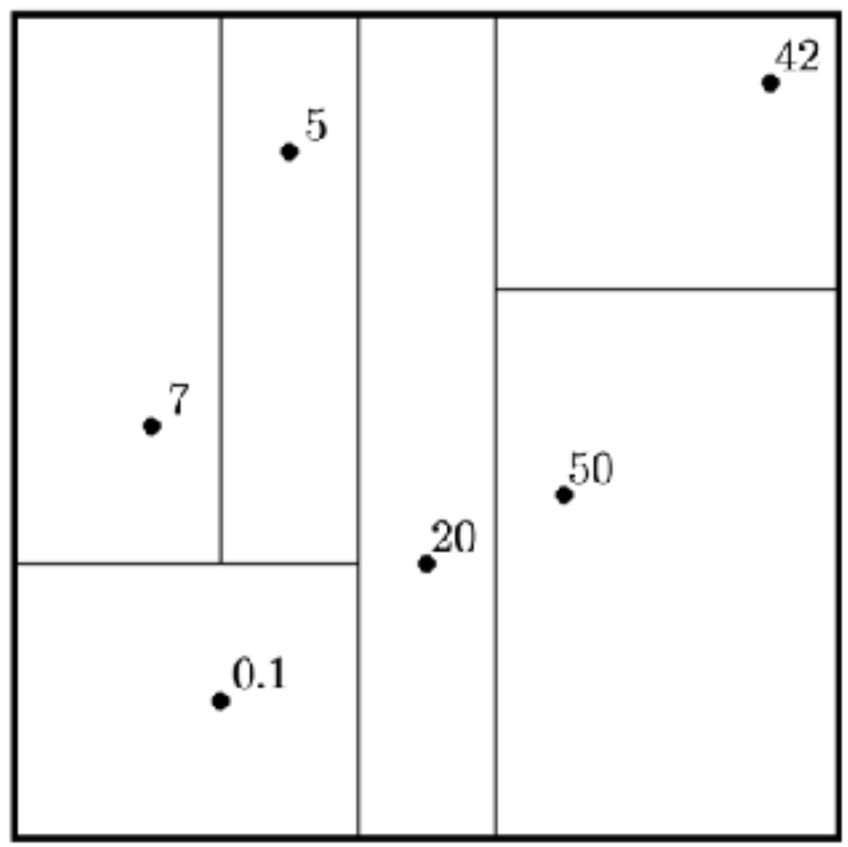
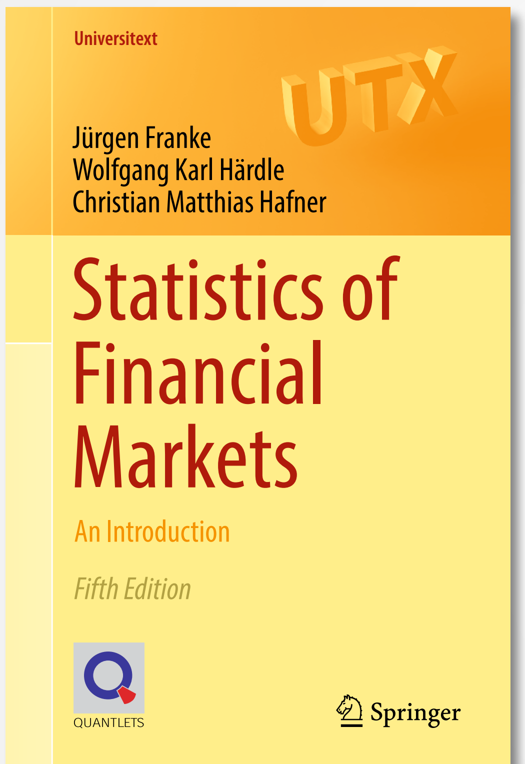


Figure 1: Tree partitioning in dimension $d = 2$, with $n = 6$ data points.

Neural Random Forests

- Perceptron, SFM Book !



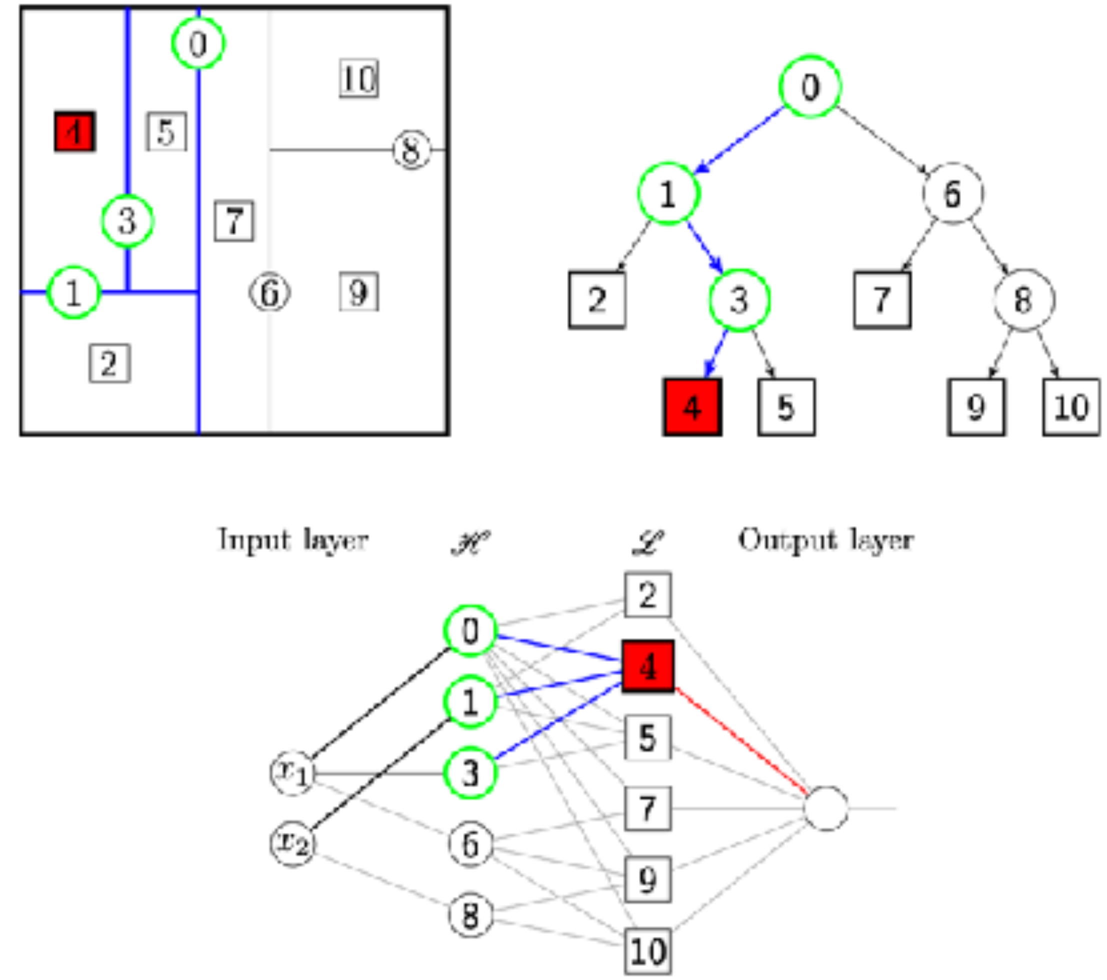


Figure 2: An example of regression tree (top) and the corresponding neural network (down).

First hidden layer. The first hidden layer of neurons corresponds to $K-1$ perceptrons (one for each inner tree node), whose activation is defined as

$$\tau(h_k(\mathbf{x})) = \tau(x^{(j_k)} - \alpha_{j_k}),$$

where $\tau(u) = 2\mathbb{1}_{u\geq 0} - 1$ is a threshold activation function. The weight vector is merely a single one-hot vector for feature j_k, and $-\alpha_{j_k}$ is the bias value. So, for each split in the tree, there is a neuron in layer 1 whose activity encodes the relative position of an input $\mathbf{x}$ with respect to the concerned split. In total, the first layer outputs the ± 1-vector $(\tau(h_1(\mathbf{x})), \ldots, \tau(h_{K-1}(\mathbf{x})))$, which describes all decisions of the inner tree nodes (including nodes off the tree

Neural Random Forests

- Independent Training
- Each tree calculated independently
- Resulting regression fct

$$|r_{M,n}(X) - m(X)|^2 \to 0?$$

To allow for training based on gradient backpropagation, the activation functions must be differentiable. A natural idea is to replace the original relay-type activation function $\tau(u) = 2\mathbb{1}_{u \geq 0} - 1$ with a smooth approximation of it; for this the hyperbolic tangent activation function

$$\sigma(u) := \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} = \frac{e^{2u} - 1}{e^{2u} + 1},$$

which has a range from -1 to 1 is chosen. More precisely, we use $\sigma_1(u) = \sigma(\gamma_1 u)$ at every neuron of the first hidden layer and $\sigma_2(u) = \sigma(\gamma_2 u)$ at every neuron of the second one. Here, γ_1 and γ_2 are positive hyperparameters that determine the contrast of the hyperbolic tangent activation: the larger γ_1 and γ_2, the sharper the transition from -1 to 1. Of course, as γ_1 and γ_2 approach infinity, the continuous functions σ_1 and σ_2 converge to the

hyper params

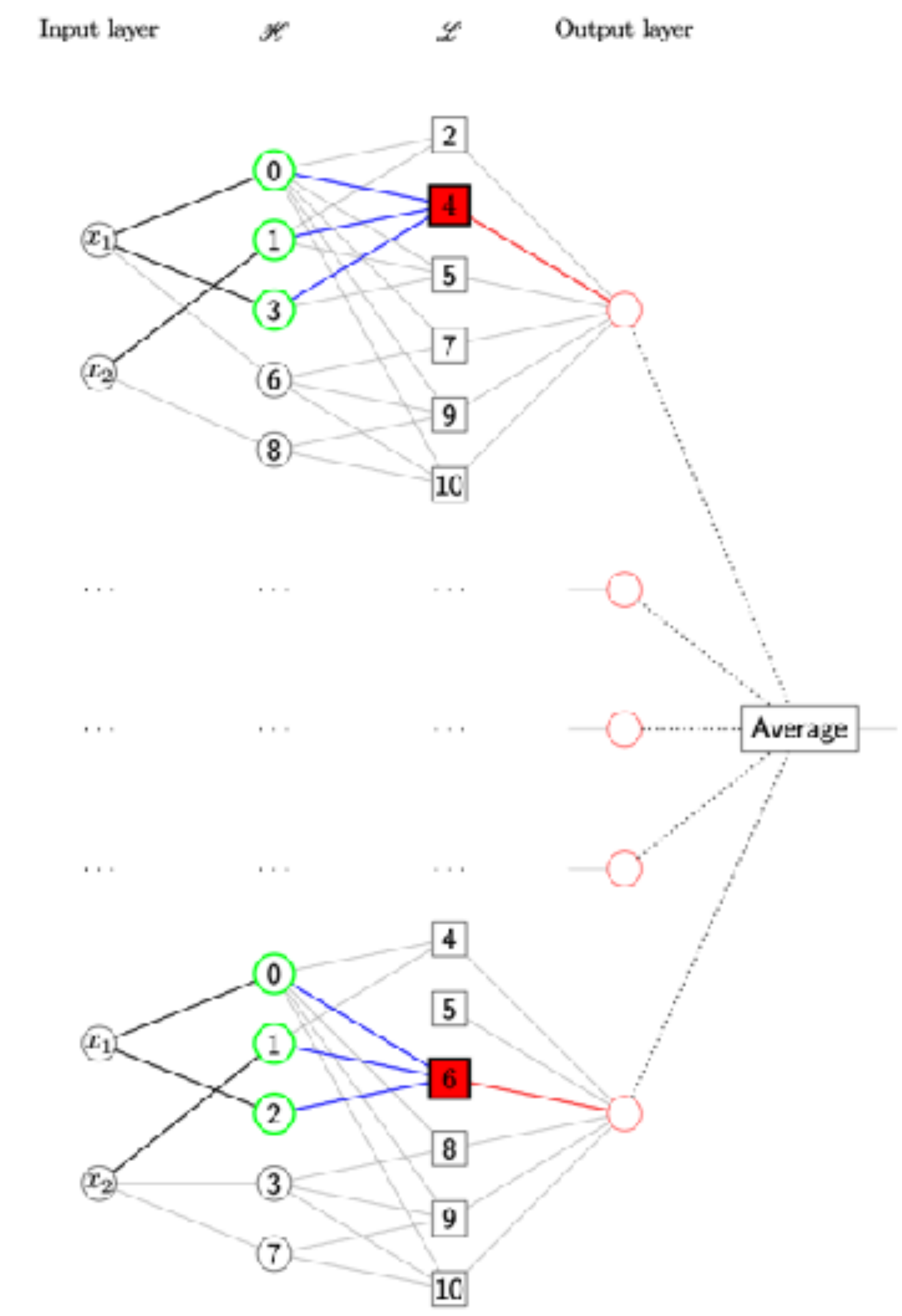


Figure 3: Method 1: Independent training.

Neural Random Forests

- Perceptron to the rescue ?

$$|s_{M,n}(X) - m(X)|^2 \to 0?$$

For a hyperrectangle $A = [a_1, b_1] \times \cdots \times [a_d, b_d] \subseteq [0,1]^d$, we let $A^{\setminus j} = \prod_{i \neq j}[a_i, b_i]$ and $\mathrm{d}\mathbf{x}^{\setminus j} = \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1}\mathrm{d}x_{j+1} \ldots \mathrm{d}x_d$. Assume we are given a measurable function $f : [0,1]^d \to \mathbb{R}$ together with $A = [a_1, b_1] \times \cdots \times [a_d, b_d] \subseteq [0,1]^d$, and consider the following two statements:

(*i*) For any $j \in \{1, \ldots, d\}$, the function

$$x_j \mapsto \int_{A^{\setminus j}} f(\mathbf{x})\mathrm{d}\mathbf{x}^{\setminus j}$$

is constant on $[a_j, b_j]$;

$\vee$

(*ii*) The function f is constant on A.

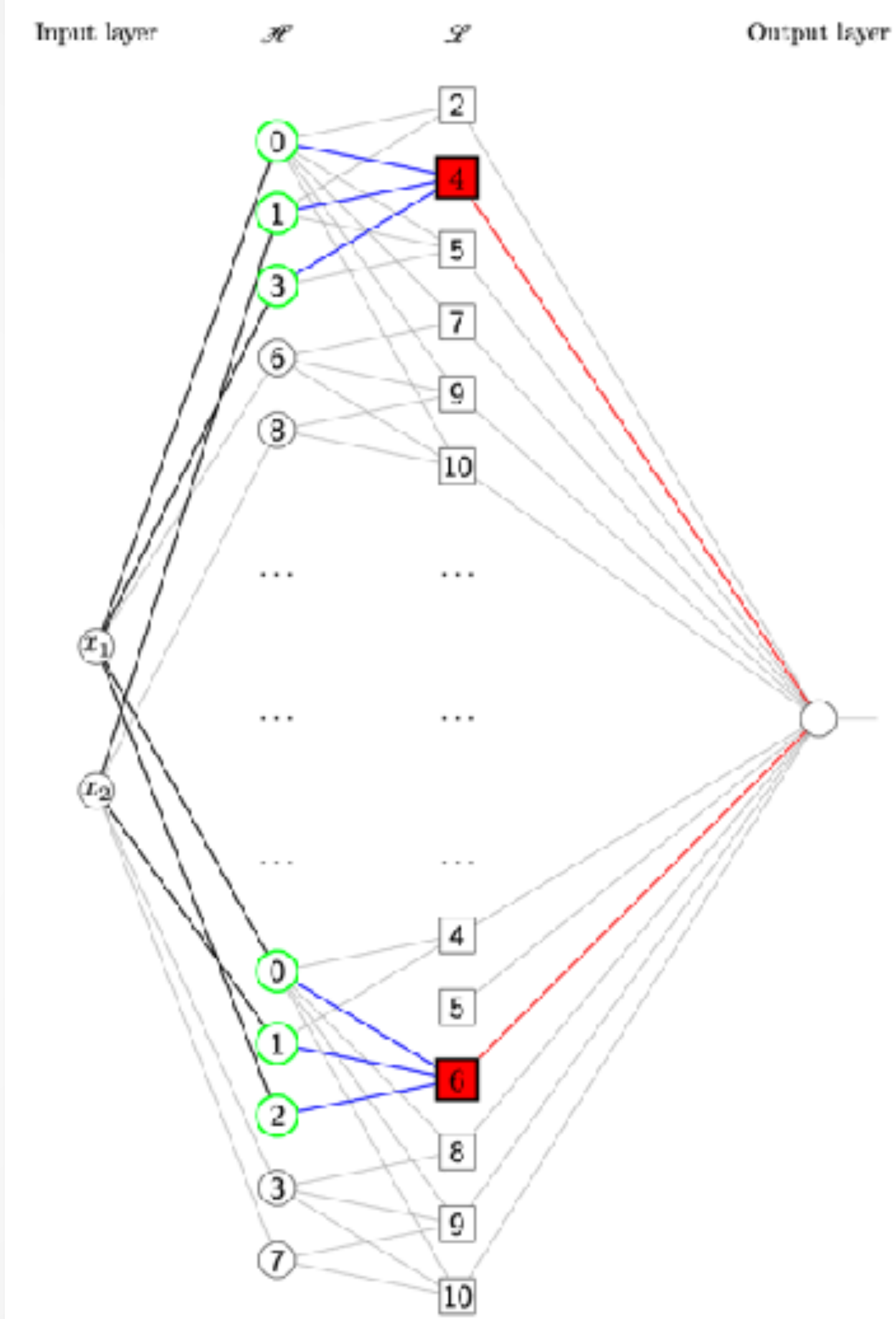


Figure 4: Method 2: Joint training.

Neural Random Forests

- Consistency

Theorem (Consistency of $r_{M,n}$ and $s_{M,n}$). Assume that X is uniformly distributed in $[0,1]^d$, $\|Y\|_\infty < \infty$, and $r \in \mathscr{F}$. Assume, in addition, that $K_n, \gamma_1, \gamma_2 \to \infty$ such that, as n tends to infinity,

$$\frac{K_n^6 \log(\gamma_2 K_n^5)}{n} \to \infty, \quad K_n^2 e - 2\gamma_2 \to 0, \quad \text{and} \quad \frac{K_n^4 \gamma_2^2 \log(\gamma_1)}{\gamma_1} \to 0$$

Then, as $n \to \infty$,

$$\mathrm{E}\,|\, r_{M,n}(X) - r(X)\,|^2 \to 0 \quad \text{and} \quad \mathrm{E}\,|\, s_{M,n}(X) - s(X)\,|^2 \to 0$$

Mathematical framework for NRF

- Additive models (AM) satisfy the condition (*)

$$f(x) = \sum_{j=1}^{d} f_j(x^{(j)})$$

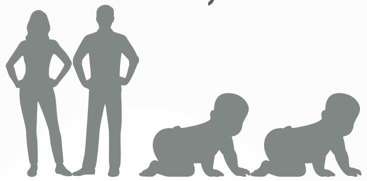
- Additive models have been extensively studied eg:
  - Härdle WK, Hall P (1993) study the backfitting algorithm for AM along with its convergence properties and consistency of its estimators
  - Härdle WK, Tsybakov AB (1995) consider additive nonparametric regression on principal components
  - Fan J, Härdle WK, Mammen E (1998) estimate the low dim components in AM
  - Härdle WK et al.(2001) developed structural tests for AM
  - Yang L, Sperlich S, Härdle WK (2003) developed tests for generalised AM
  - Härdle WK et al. (2004) provided bootstrap inference in semiparam. gen. AM
  - Liu R, Yang L, Härdle WK (2013) provide efficient estimation of gen. AM

Implied functionals

- Model free causal inference with binary treatment effects
- Generalized RF (GRF) by Athey et al. (2016) tackle the problem via generalised method of moments (GMM), e.g. for
  - Quantile regression
  - Treatment effect estimation
  - Instrumental variables
- GRF can estimate functions with different loss functions

Application

- Treatment effect analysis of number of children on labor force participation of mothers in the US in Athey et al. (2019)

- Data:
  - Subset of 1980 US census data, including only married mothers with ≥ 2 children  ....
  - Target variable: Did the mother work in the year before the census? 
- Analysis of labor-force participation of mothers with ≥ 3 children
  - Treatment effect: Does mother have ≥ 3
  - Instrumental variable: Do first children have different gender?
- Covariates:
  - Age of mother at birth of first child
  - Age of mother at census
  - Years of education of mother
  - Race of mother
  - Income of father

IV Regression

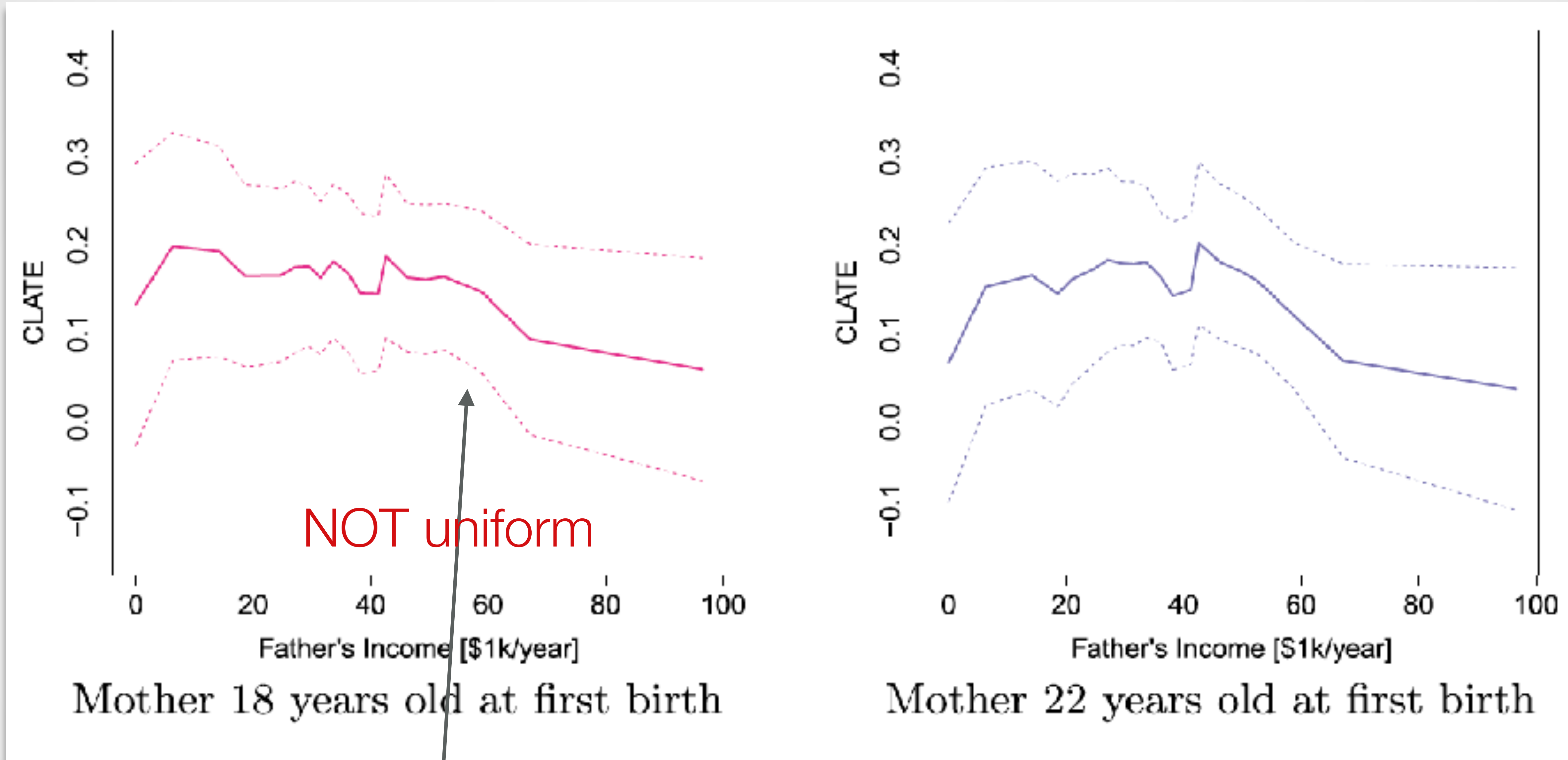


Fig.: GRF estimates with pointwise 95% confidence intervals for causal effect of having a their child on probability that mother works for pay. CLATE ↑ ≻ Probability mother working ↓
Source: Fig. 3 in Athey et al 2019

design of the following 4 slides by Marius Sterling

Effective weights

$n = 500, 1000, 2000$

$X_i = -1 + 2i/n,\ i = 1,\ldots,n$

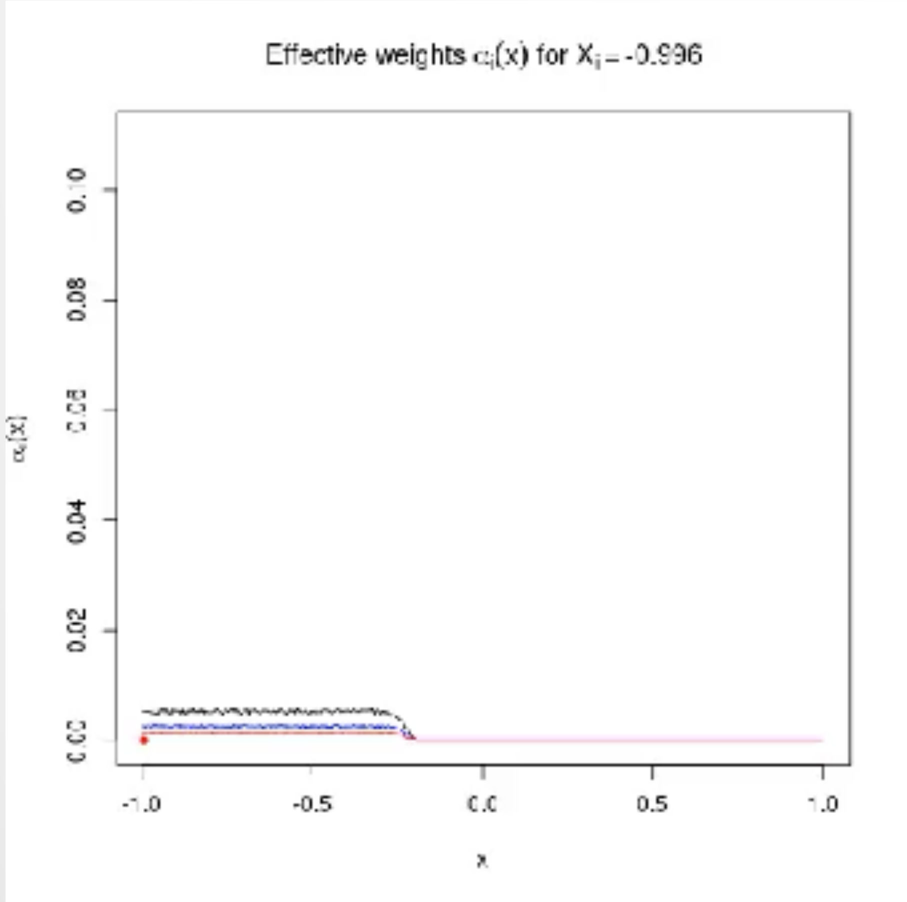
$\theta(x_1, x_2) = \max(0, 1 - |x_1|/\eta),\ \eta = 0.2$

$Y_i = \theta(X_i) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

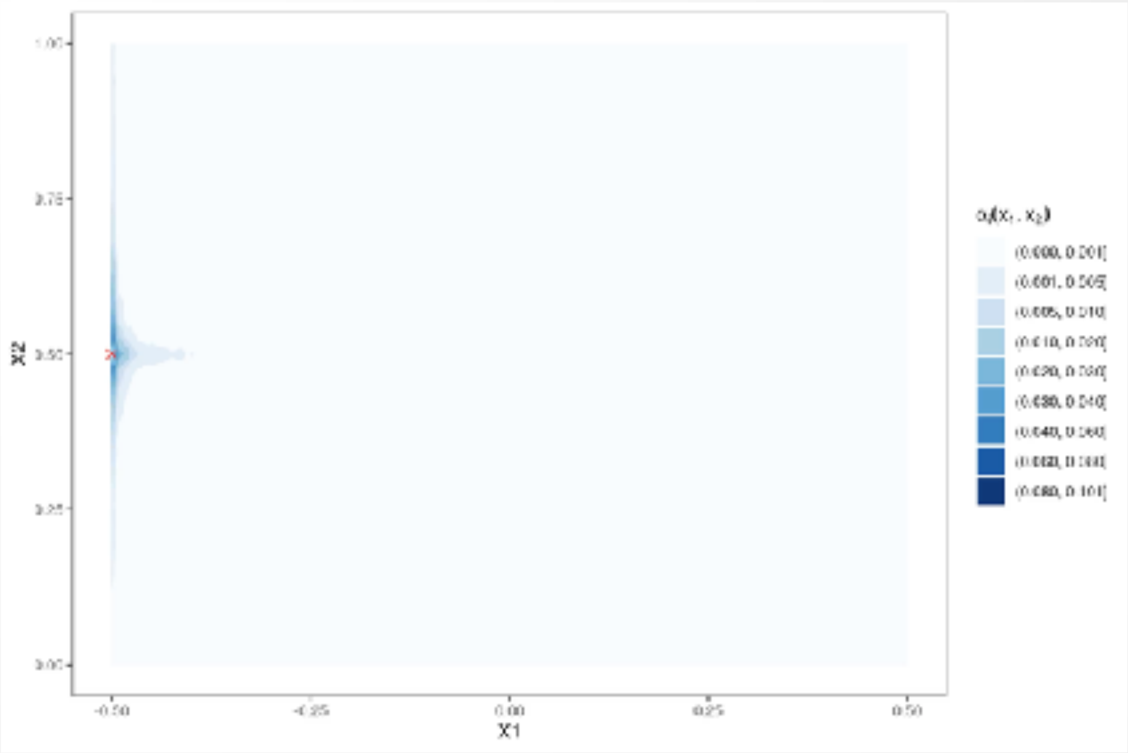
$$\tilde{\theta}(x) = \theta(x) + \sum_{i=1}^{n} \alpha_i(x)\tilde{\varepsilon}_i(x)$$

$\alpha_i(x)$ based on RF trained on $(X_i, Y_i)_i$

$\sigma_\varepsilon = 0$



eff. weights at $x_i = (0, 0.5)$ $\sigma_\varepsilon = 0.1$



GRF_effective_weights2D

RFs are locally adaptive Smoothers

- **(H1)** Fix $x \in [0,1]^d$, and assume that $\mathscr{D}_n = (X, Y)$, $Y \geq 0$ a.s.. and $N_n\left(x, \Theta_j\right) = \sum_{i=1}^{n} \mathbf{I}_{\mathbf{X}_i \in A_n\left(x, \Theta_j\right)}$ and $A_n\left(x, \Theta_j\right)$ is the cell containing x
- Then, one of the following two conditions holds:
  - **(H1.1)**There exist sequences $\left(a_n\right), \left(b_n\right)$ such that, a.s.

$$a_n \leq N_n(x, \Theta) \leq b_n \quad \text{and} \quad a_n \leq \frac{1}{M} \sum^{M} N_n\left(x, \Theta_m\right) \leq b_n$$

  - **(H1.2)** There exist sequences $\left(\varepsilon_n\right), \left(a_n\right), \left(b_n\right)$ such that, a.s.

1. $\mathrm{E}_\Theta \left[N_n(x, \Theta)\right] \geq 1$
2. $\mathbb{P}\left[a_n \leq N_n(x, \Theta) \leq b_n \mid \mathscr{D}_n\right] \geq 1 - \varepsilon_n/2$
3. $\mathbb{P}\left[a_n \leq \mathrm{E}_\Theta \left[N_n(x, \Theta)\right] \leq b_n \mid \mathscr{D}_n\right] \geq 1 - \varepsilon_n/2$

Kernel based on random forests (KeRF) (Scornet 2015)

- RF estimates

$$m_{M,n}\left(x,\Theta_1,\ldots,\Theta_M\right) = \frac{1}{M}\sum_{j=1}^{M}\left(\sum_{i=1}^{n}\frac{Y_i \mathbf{I}_{\mathbf{X}_i\in A_n\left(x,\Theta_j\right)}}{N_n\left(x,\Theta_j\right)}\right)$$

- KeRF estimates

$$\tilde{m}_{M,n}\left(x,\Theta_1,\ldots,\Theta_M\right) = \frac{1}{\sum_{j=1}^{M} N_n\left(x,\Theta_j\right)}\sum_{j=1}^{M}\sum_{i=1}^{n} Y_i \mathbf{I}_{\mathbf{X}_i\in A_n\left(x,\Theta_j\right)}$$

Proposition: Assume that (H1.1) is satisfied. Thus, almost surely,

$$\left|m_{M,n}(x) - \tilde{m}_{M,n}(x)\right| \le \frac{b_n - a_n}{a_n}\tilde{m}_{M,n}(x)$$

- Hence RFs are kernel estimates, if # obs in each cell is controlled
- (H1.1) holds true for some type of random forests

„Chuck's speed limiz"

- RFs are local kernel estimators
- Convergence rates calculated in min max framework
- All smoothers follow the eternal Charles Stone rule

A very small „ball" yields increased precision but dimension hits you exponentially hard.

Chuck's speed limiz: „ball" = functional class (p), dimension (d)

$$\min\max MISE(m_n) = \mathcal{O}(n^{-2p/(2p+d)})$$

- The Olymp is right: RFs cannot escape the eternal rules!
- Non-eternal optimists: smaller balls (like AMs) do the job!

▶ Return to (Mourtada, Gaiffas)

In the words of the founder

'But the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem.'

'Take the output of random forests not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem.'

Pointed Stick

- Monty Python Flying Circus

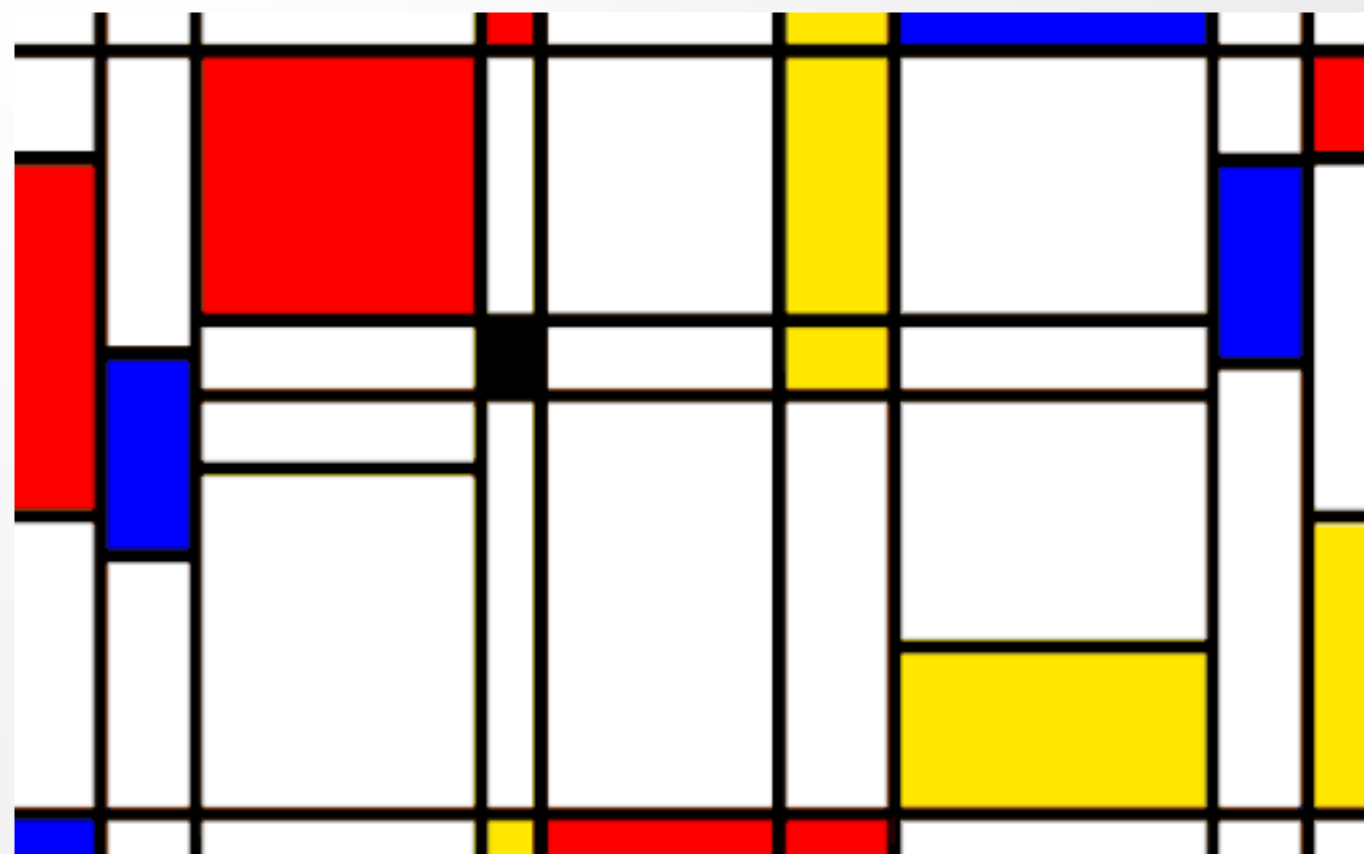
Trespassing Random Forests

with a pointed stick for self defence

Kainat Khowaja

Wolfgang Karl Härdle

IRTG 1792 High Dimensional Non-Stationary Time Series
Humboldt-Universität zu Berlin
IRTG1792.HU-Berlin.de



Important Literature

Biau G, Scornet E (2016)
A random forest guided tour
TEST, Vol. 25, 197-227, DOI 10.1007/s11749-016-0481-7

Athey S, Tibshirani J, Wager S (2019)
Generalized Random Forests
Annals of Statistics, Vol. 47(2), 1148-1178 , DOI: 10.1214/18-AOS1709

Mozharovskyi P (2017)
Classification tree, bagging, and random forest
https://perso.telecom-paristech.fr/mozharovskyi/resources/Lecture_CART_RF_handout.pdf

Härdle WK, Huet S, Mammen E, Sperlich S. (2004) *Bootstrap Inference in Semiparametric Generalized Additive Models,*
Econometric Theory, 20(2), 265-300.

Biau G, Scornet E, Welbl J (2018)
Neural Random Forests
Sankhya A 81, 347–386 DOI 10.1007/s13171-018-0133-y

Engel E (1857). *Die Productions- und Consumptionsverhältnisse des Königreichs Sachsen*. Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren. 8–9: 28–29. "

Chakrabarty M, Hildenbrand W (2009) *Engel's Law Reconsidered*
Discussion Paper 22/2009 , Bonn Graduate School

Härdle WK, Hall P (1993) *On the Backfitting Algorithm for additive regression models.*
Statistica Neerlandica, 47, 43-57.

Härdle WK, Tsybakov AB (1995)
Additive Nonparametric Regression on Principal Components.
Journal of Nonparametric Statistics, 5, 157-184.

Fan J, Härdle WK, Mammen E (1998) *Direct Estimation of Low Dimensional Components in Additive Models.* Annals of Statistics, 26, 943-971

Härdle WK, Sperlich S, Spokoiny V (2001) *Structural Tests in Additive Regression.* J. Amer. Stat. Assoc., 96, 1333-1347.

Yang L, Sperlich S, Härdle WK (2003) *Derivative Estimation and Testing in Generalized Additive Models.* J Statistical Planning and Inference, 115, 521-542.

Liu R, Yang L, Härdle WK (2013) *Oracally Efficient Two-Step Estimation of Generalized Additive Model.* Journal of the American Statistical Association, Vol 108, Issue 502, 619-631.

Stone Theorem for single trees

Condition 1: Set $W_{ni}(x) = \dfrac{\mathbf{I}\{X_i \in A_n(x,\theta)\}}{N_n(x,\theta)}$ in tree estimate $m_n(x)$

Condition 2: Note that for all $a > 0$

$$\mathrm{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(X)\mathbf{I}\{\,\|X_i - X\|_\infty > a\}\right] = \mathrm{E}\left[\sum_{i=1}^{n} \frac{\mathbf{I}\{X_i \in A_n(x,\theta)\}}{N_n(x,\theta)}\mathbf{I}\{\,\|X_i - X\|_\infty > a\}\right]$$

$$= \mathrm{E}\left[\sum_{i=1}^{n} \frac{\mathbf{I}\{X_i \in A_n(x,\theta)\}}{N_n(x,\theta)}\mathbf{I}\{\,\|X_i - X\|_\infty > a\} \times \mathbf{I}_{diam}\{A_n(X,\theta) \geq a/2\right]$$

Because $\mathbf{I}\{\,\|X_i - X\|_\infty > a\} \times \mathbf{I}_{diam}\{A_n(X,\theta) < a/2 = 0$. Thus,

$$\mathrm{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(X)\mathbf{I}\{\,\|X_i - X\|_\infty > a\}\right] \leq \mathrm{E}\left[\mathbf{I}_{diam}\{A_n(X,\theta) \geq a/2 \times \sum_{i=1}^{n} \mathbf{I}\{X_i \in A_n(x,\theta)\}\mathbf{I}\{\,\|X_i - X\|_\infty > a\}\right]$$

$$\leq \mathbb{P}\left[\mathrm{diam}\{A_n(X,\theta) \geq a/2\right] \to 0 \quad \text{as } n \to \infty \text{ (per assumption)}$$

Condition 3: The tree partition has 2^k cells, denoted by $A_1, \ldots, A_{2^k}$. For $1 \le i \le 2^k$, let N_i be the number of points among $X, X_1, \ldots, X_n$ falling into A_i. Finally, set $S = \{X, X_1, \ldots, X_n\}$. Since these points are independent and identically distributed, fixing the set S (but not the order of the points) and Θ, the probability that X falls in the ith cell is $\dfrac{N_i}{n+1}$ Thus, for every fixed $t > 0$,

$$\mathbb{P}\left[N_n(X,\Theta) < t\right] = \mathbb{E}\left[\mathbb{P}\left[N_n(X,\Theta) < t \,|\, S, \Theta\right]\right] = \mathbb{E}\left[\sum_{i:N<t+1} \frac{N_i}{n+1}\right] \le \frac{2^k}{n+1} t$$

Thus, by assumption, $N_n(X,\Theta) \to \infty$ as $n \to \infty$

Note:

$$\mathbb{E}\left[\max_{1 \le i \le n} W_{ni}^{\infty}(X)\right] \le \mathbb{E}\left[\max_{1 \le i \le n} \frac{\mathbf{I}\{X_i \in A_n(x,\theta)\}}{N_n(x,\theta)}\right] \le \mathbb{E}\left[\frac{\mathbf{I}\{X_i \in A_n(x,\theta)\}}{N_n(x,\theta)}\right] \to 0 \quad \text{as } n \to \infty$$

Since $N_n(X,\Theta) \to \infty$ in probability, as $n \to \infty$

Imp: Forest consistency results from the consistency of each tree.

▶ Return to Stone theorem