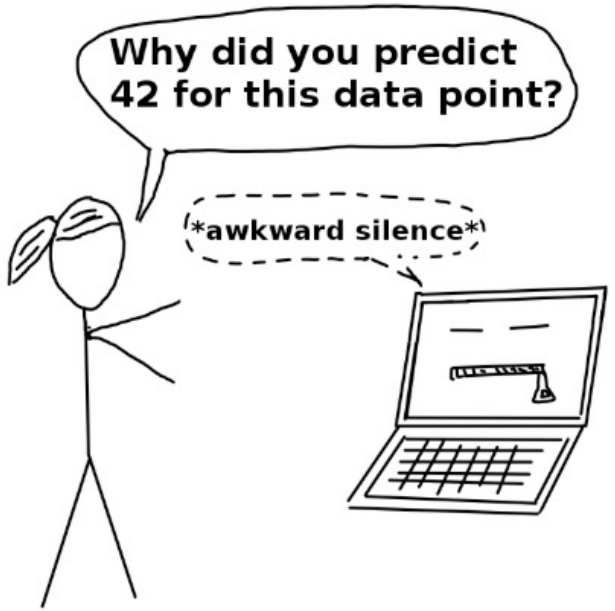
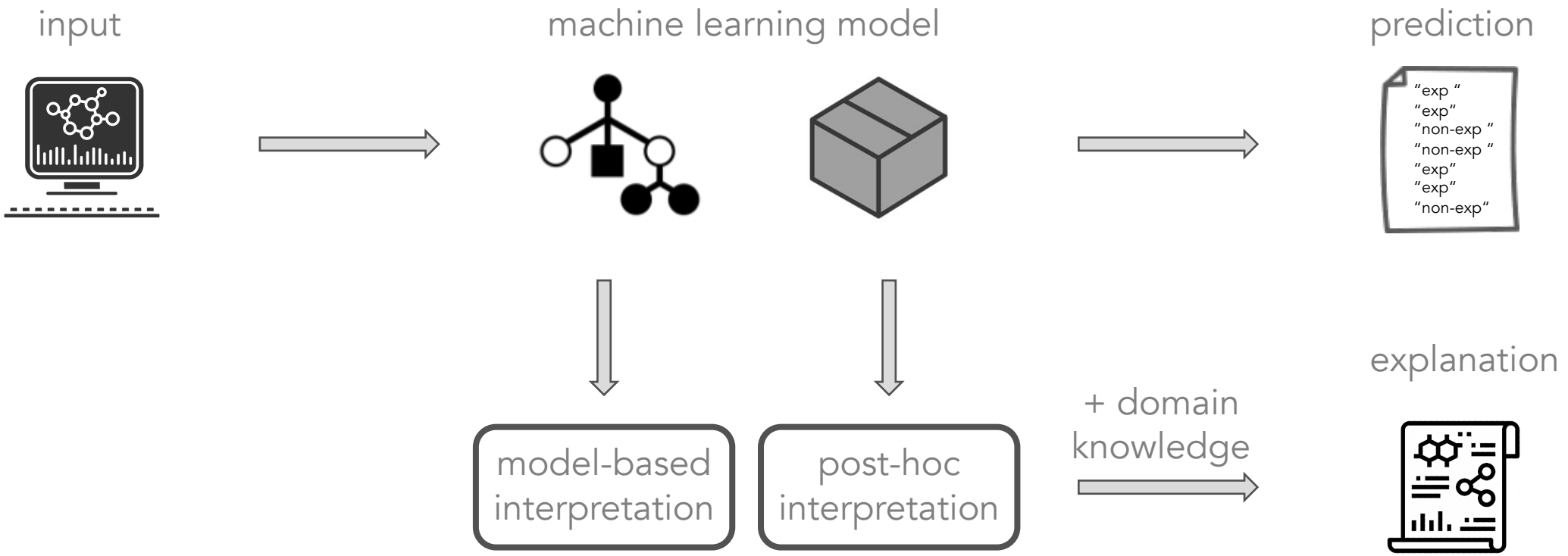
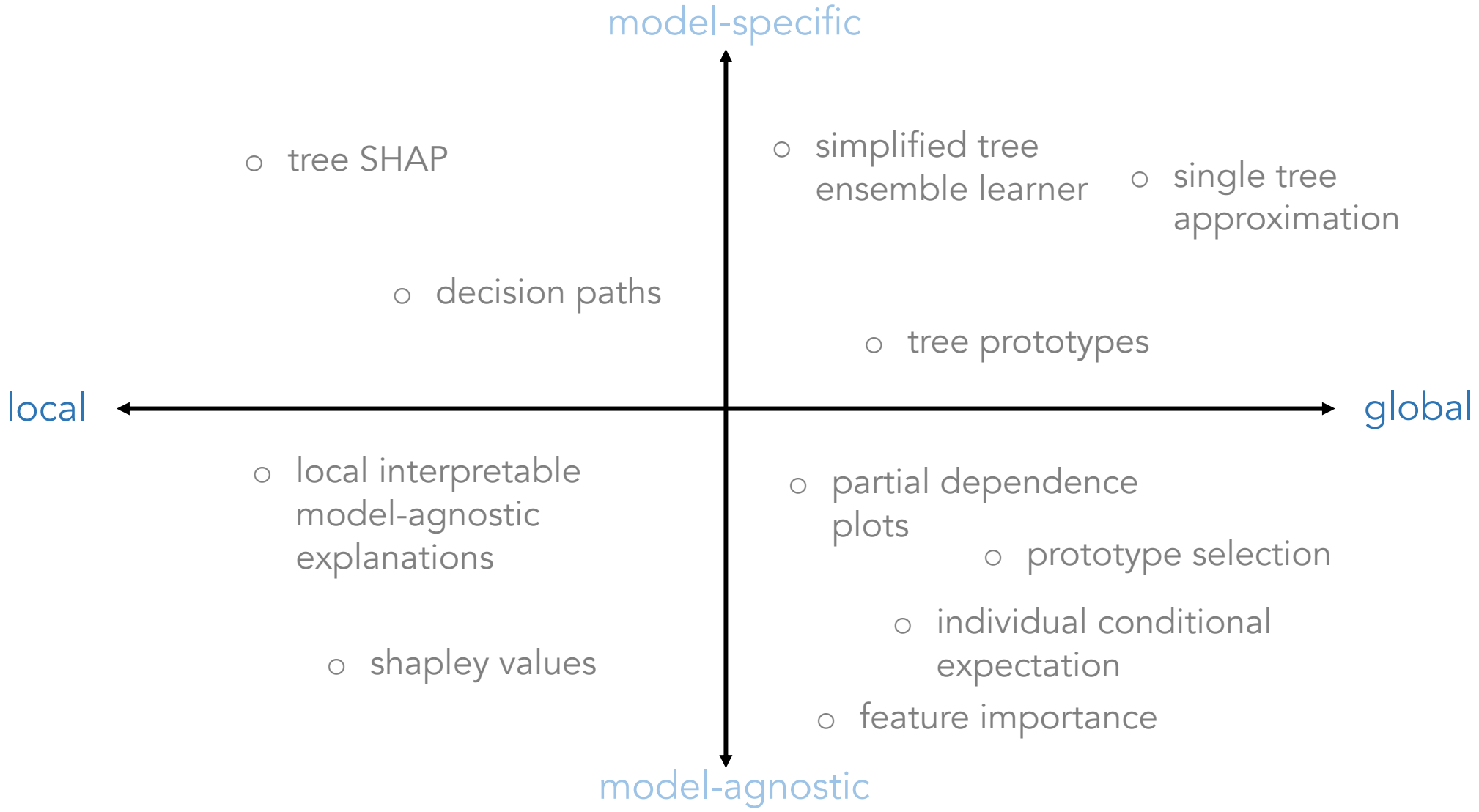
XAI for Random Forest



How is interpretability defined?

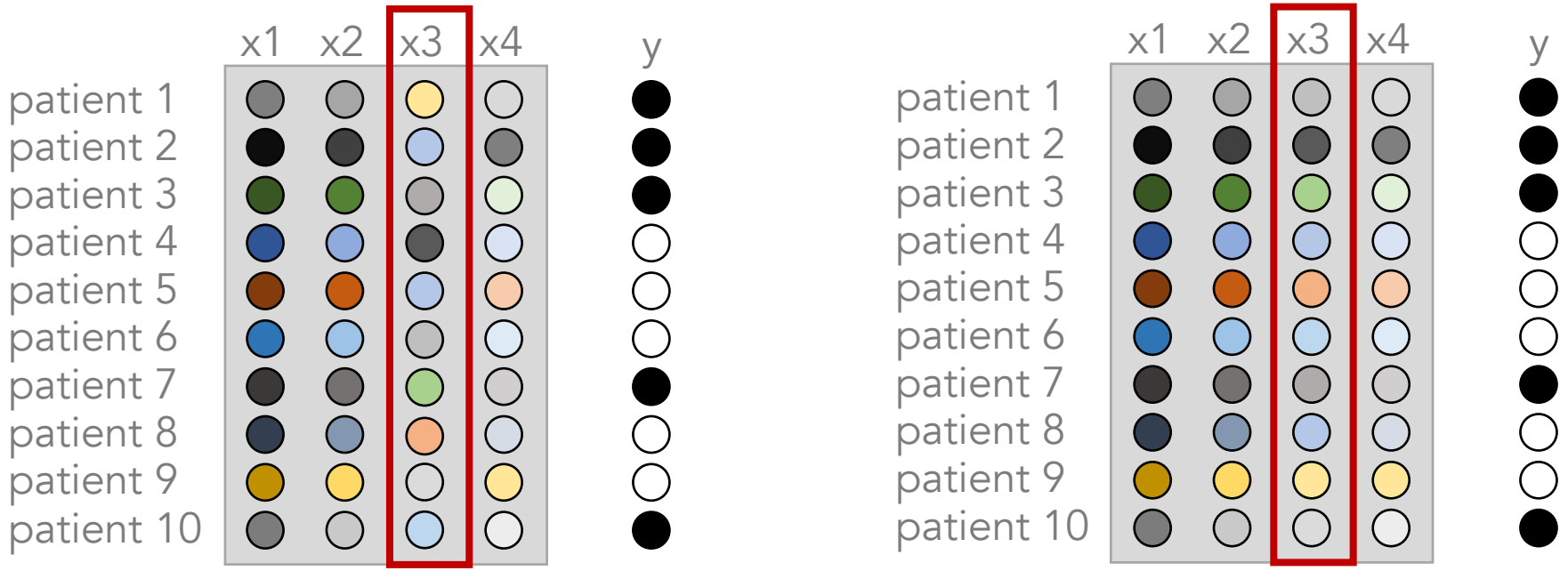


Post-hoc interpretation methods



Feature importance – a global model-agnostic method

idea: measure feature importance through feature permutation (Breiman, 2001)

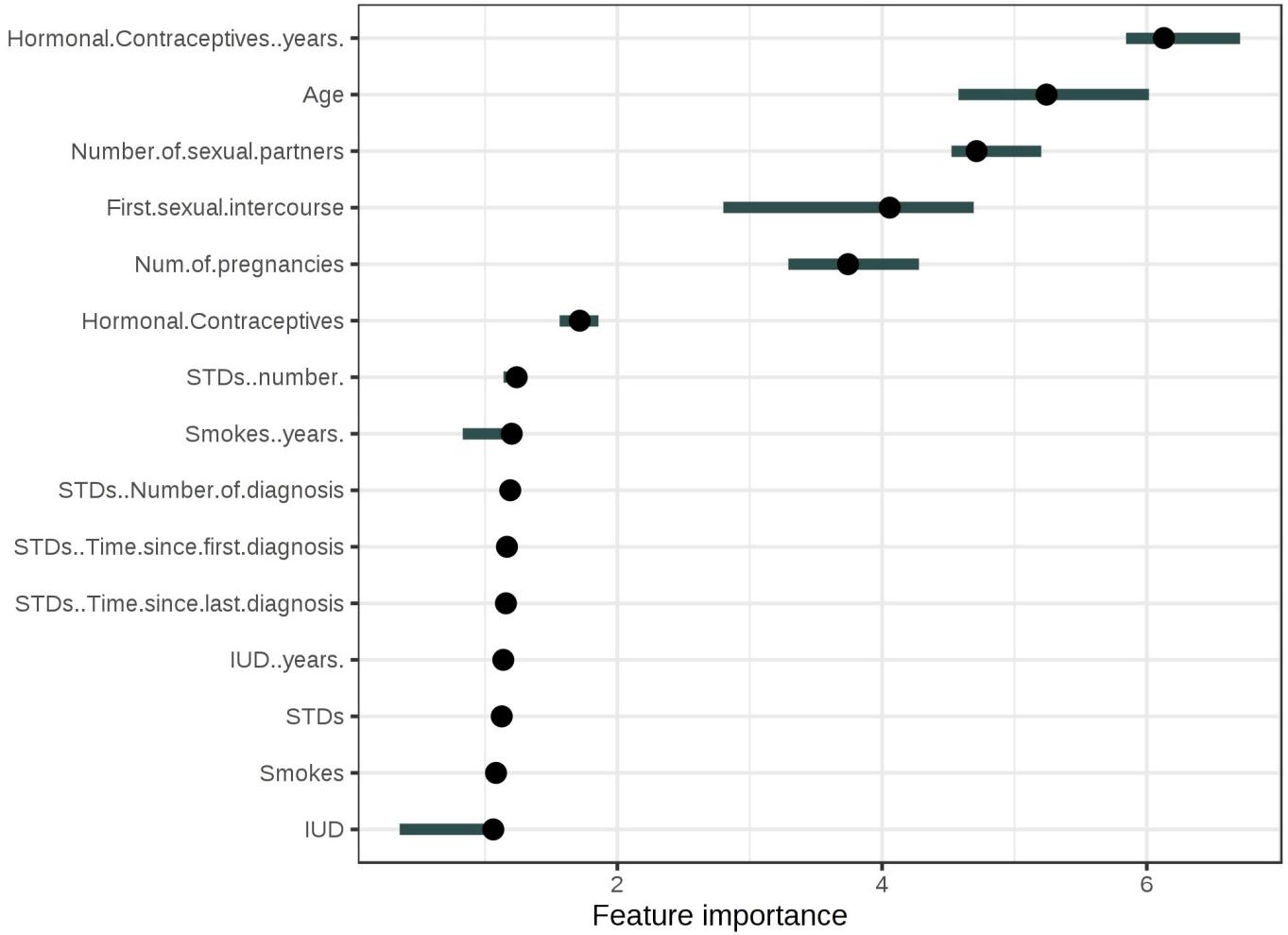


$$FI_{x3} \qquad e_{perm} = L(y, f(X_{perm})) \qquad - \qquad e_{orig} = L(y, f(X))$$

Feature importance – a global model-agnostic method

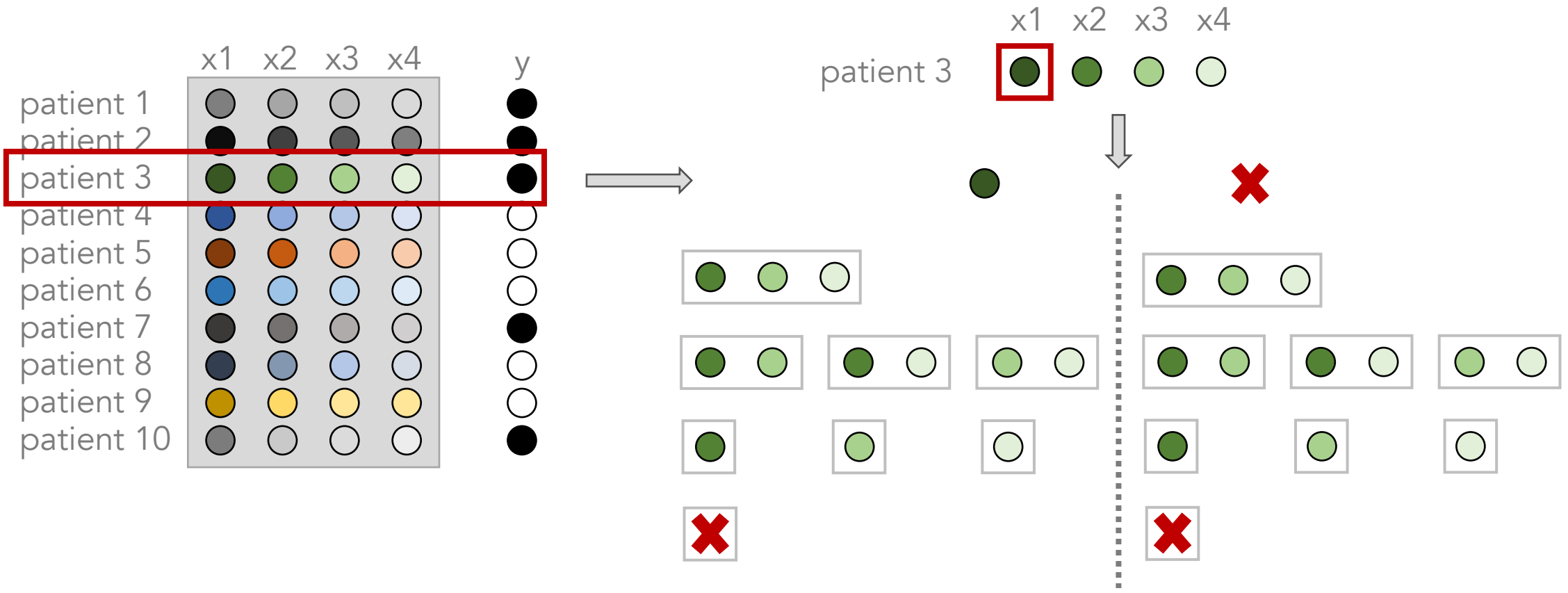
idea: measure feature importance through feature permutation (Breiman, 2001)

example for cervical cancer classification (Molnar, 2019):



Shapley values – a local model-agnostic method

idea: compute feature contributions for single predictions (Shapley, 1953)

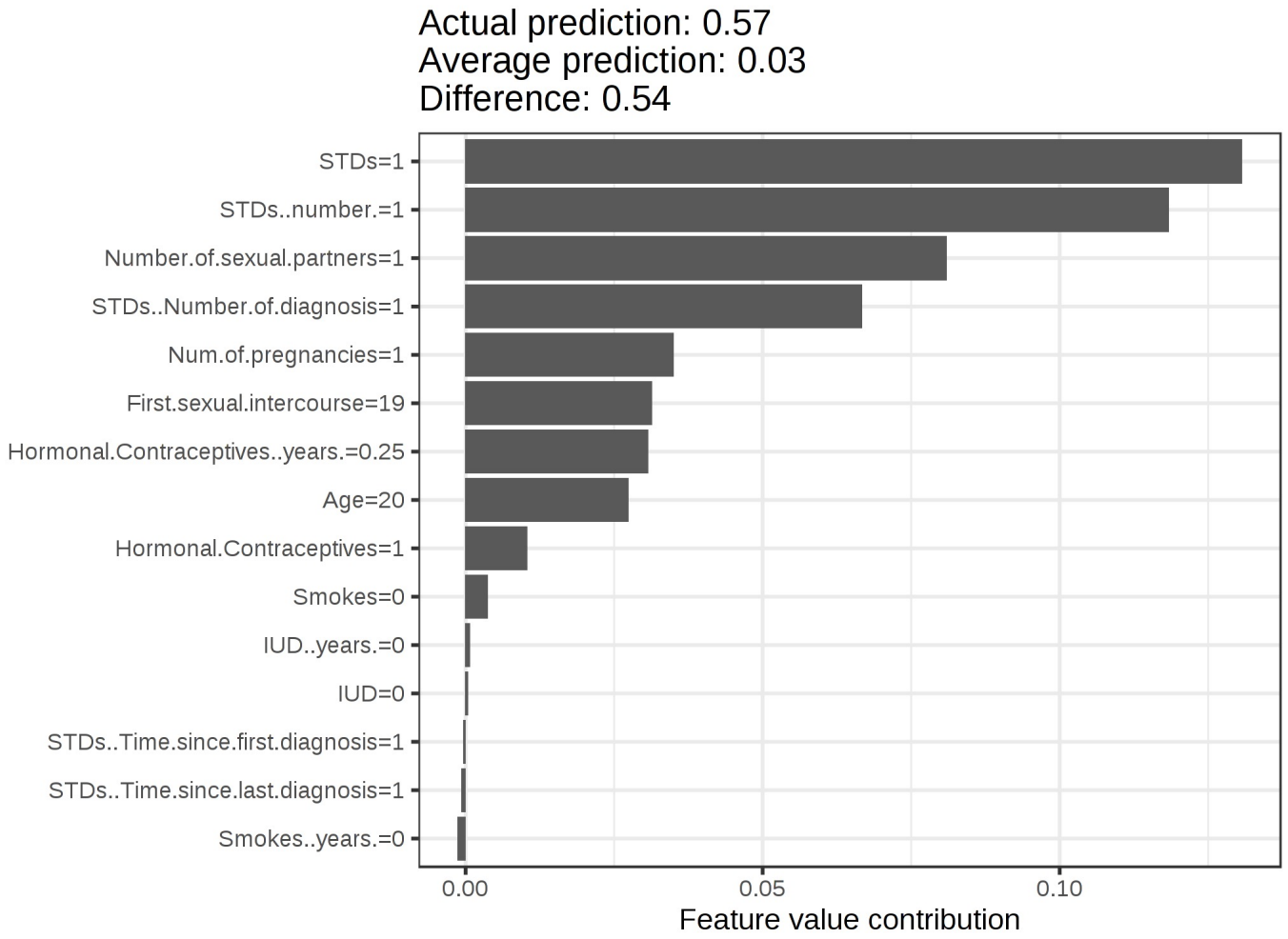


marginal contribution: $\phi_s(x_1) = f(s_{+x_1}) - f(s_{-x_1})$

Shapley values – a local model-agnostic method

idea: compute feature contributions for single predictions (Shapley, 1953)

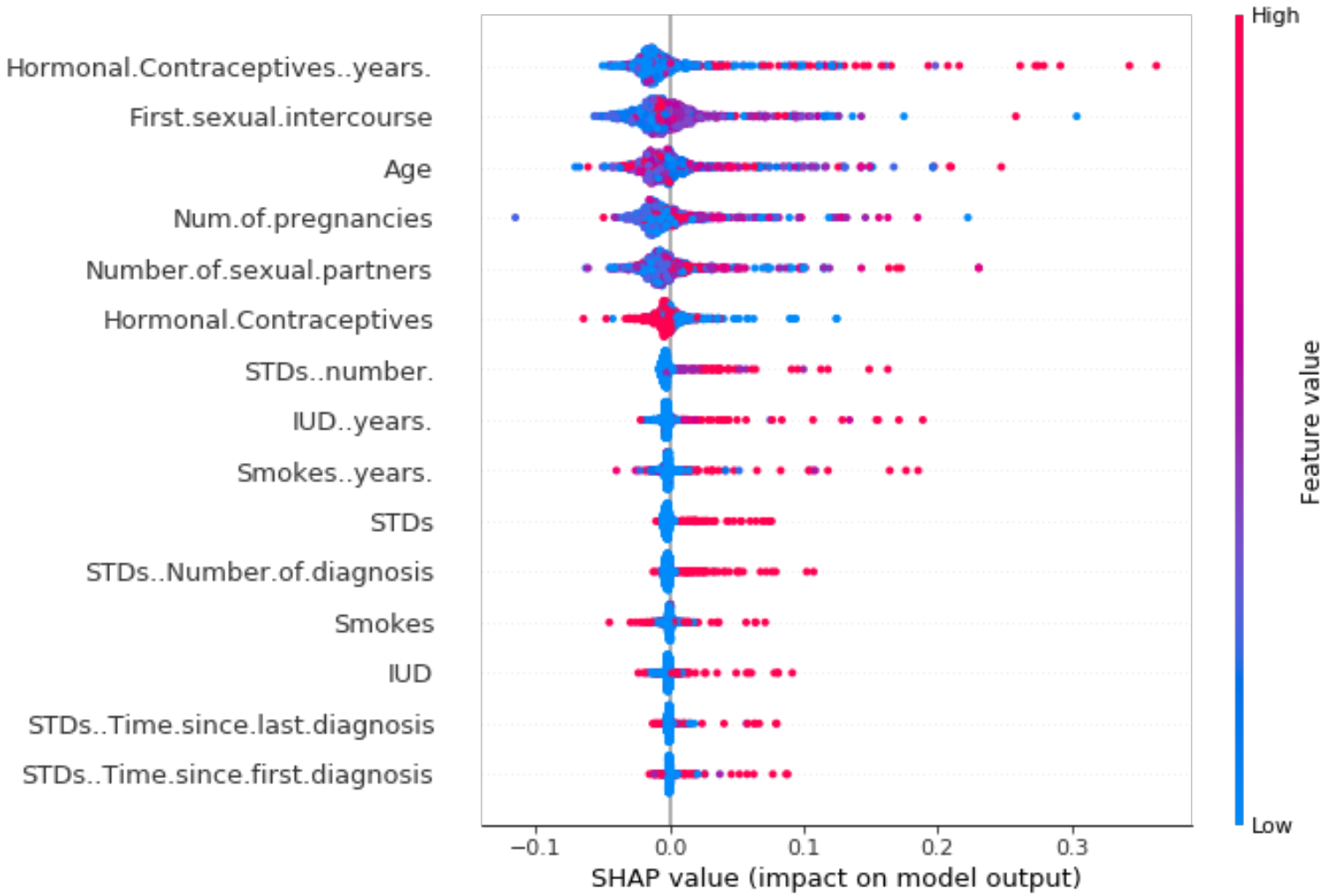
example for cervical cancer classification (Molnar, 2019):



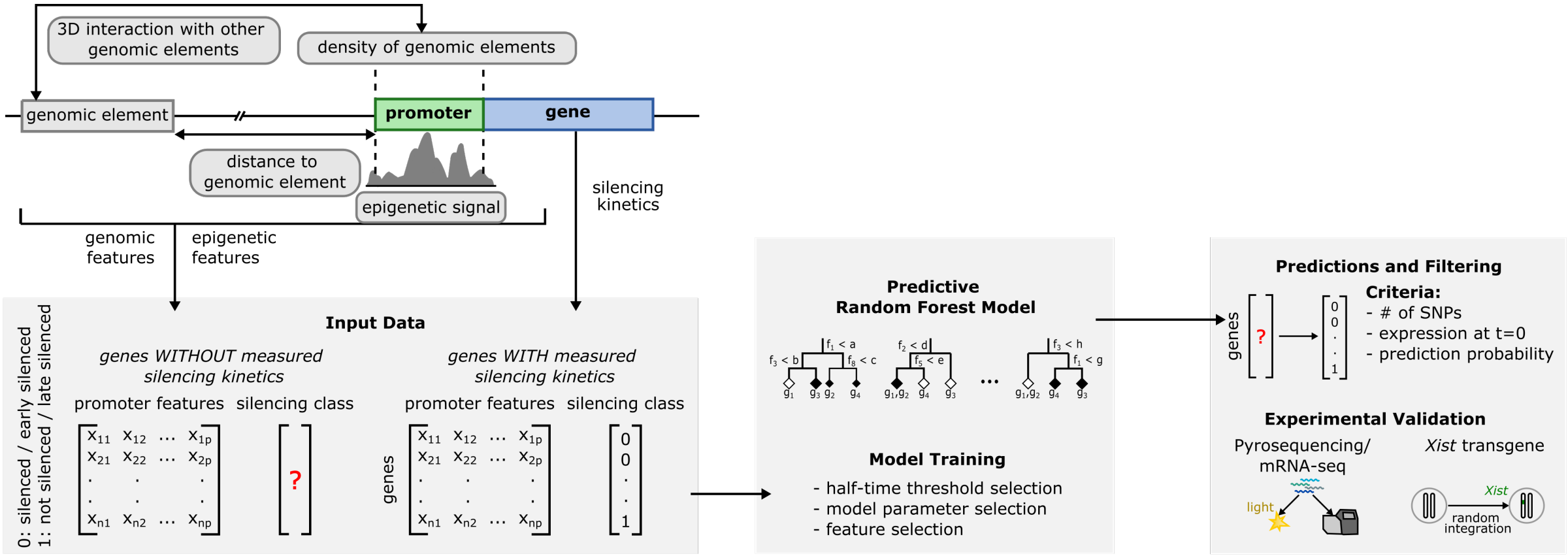
SHapley Additive exPlanations (SHAP)

idea: alternative estimation method for shapley values to reduce computing time (Lundberg et al., 2017)

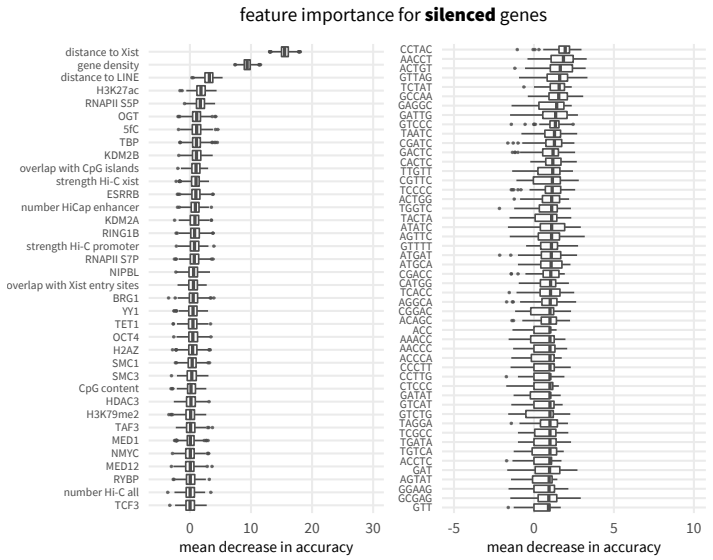
example for cervical cancer classification (Molnar, 2019):



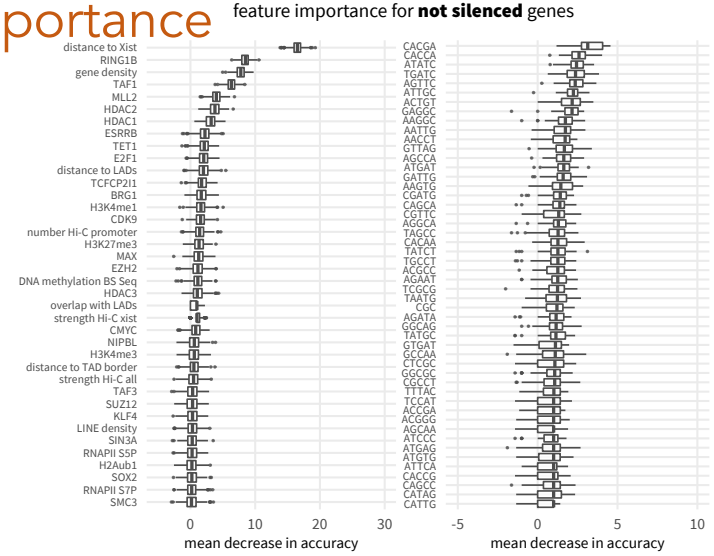
Modelling approach

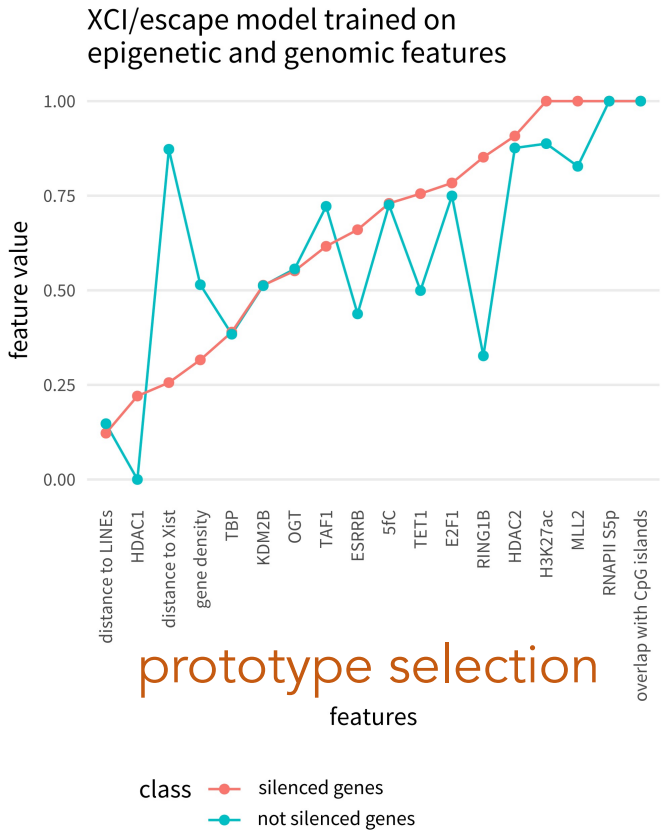


Model Interpretation



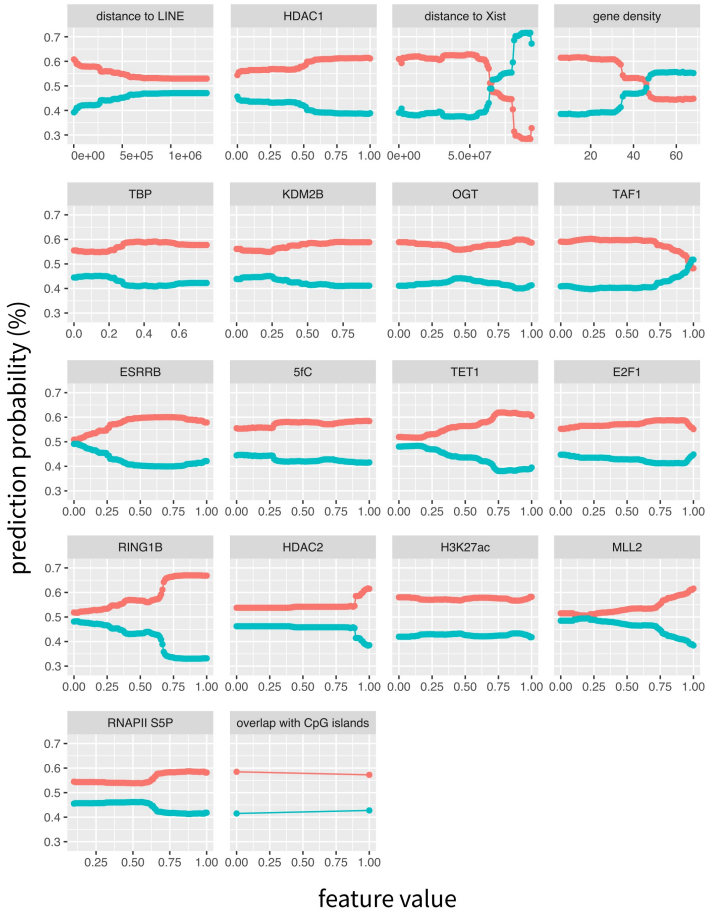
feature importance



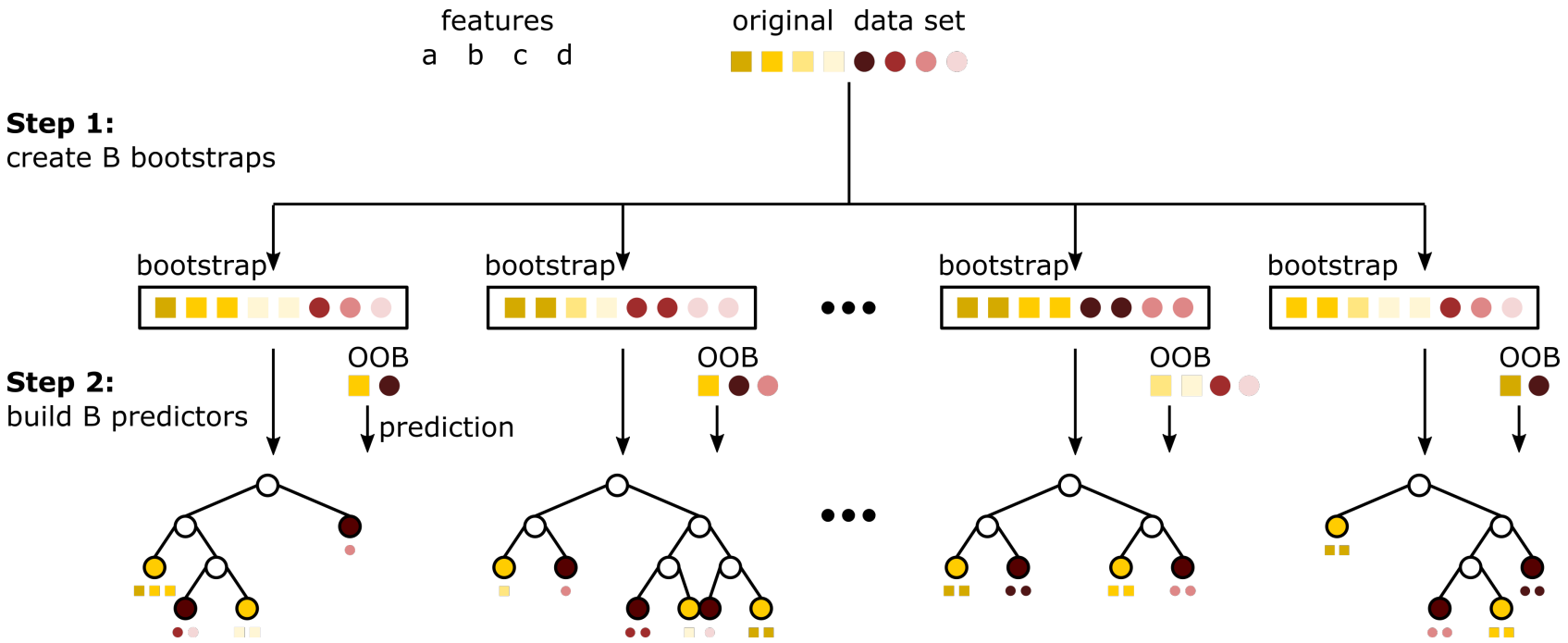


prototype selection

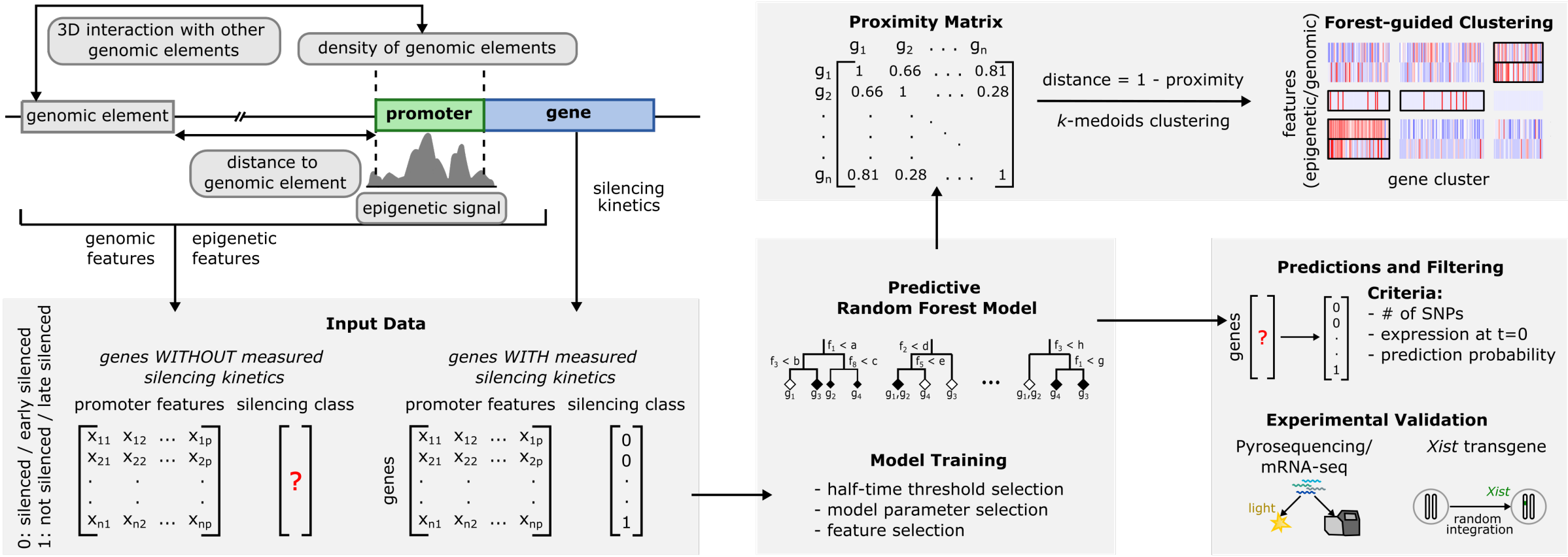
partial dependence



Random Forest Models



Modelling approach



Forest-guided clustering of XCI/escape model