

HUMAN CENTERED MACHINE LEARNING

Lecture 4: Introduction to Explainable Machine Learning & Interpretable Models

Lecturer: Heysem Kaya

Reminder



Project groups are not randomly assigned. You form your own!



Deadline for project proposals is Friday May 28, 17:00!



See the course syllabus for details!



Bridge

- Former Lectures
 - Introduction to HCML: Bias / Fairness in Society and Data
 - Measuring and Mitigating Bias
- Today
 - Introduction to Explainable Machine Learning
 - Properties of (good) explanations
 - Intrinsically Interpretable Models

Outline



What is Explainable AI/ML?



Motivation, Taxonomy and Properties



Interpretable Models

Linear Models

Decision Trees

RuleFit



Examples

What is Explainable AI/ML



- No consensus on a universal definition: definitions are domain-specific
- **Interpretability:**
 - *ability to explain or to present in understandable terms to a human* [D]
 - the degree to which a human can understand the cause of a decision
- **Explanation:** Answer to a *WHY* question
 - An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way.
- Interpretability and explainability: sometimes used interchangeably
- Molnar [M]: **model interpretability (global)** vs **explanation** of an individual **prediction (local)**
- Ribeiro [R]: **explainable** models are **interpretable** if they use a small set of features
 - «an **explanation** is a local linear approximation of the model's behaviour.»
- Prediction model vs explanation model

[D] Doshi-Velez and Kim, Towards A Rigorous Science of Interpretable Machine Learning, ArXiv 2017

[M] Molnar, Interpretable Machine Learning, <https://christophm.github.io/interpretable-ml-book/>

[R] Ribeiro et al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier, KDD 2016

Motivation: Why do we need XAI?

Scientific Understanding

- In search of causal factors / effects

Bias / fairness issues

- Does my model discriminate?

Model debugging and auditing

- Why did my model make this mistake?

Human-AI cooperation / acceptance

- How can I understand / interfere with the model?

Regulatory compliance

- Does my model satisfy legal requirements? E.g. GDPR*

High-risk applications & regulated industries

- Healthcare, finance / banking, insurance

* <https://www.privacy-regulation.eu/en/22.htm>

Class discussion

- When do you think we need interpretability?

Application-wise? (must have – nice to have)

- Affect recognition in video games / intelligent tutoring systems
- Bank loan decision
- Bail / parole decisions
- Critical healthcare predictions (e.g. cancer, major depression)
- Film / music recommendation
- Job interview recommendation / job offer
- Personality impression prediction for job interview recommendation
- Tax exemption

Taxonomy of XAI Methods

Intrinsic or post-hoc

- Intrinsic: model is interpretable due to simplicity
 - shallow **decision trees** or sparse **linear models**
- Post-hoc: another method/model is required to explain the black-box model

Model Agnostic or Model Specific

- Agnostic: explains any model no matter the type
- Specific: accesses and uses the model internals

Local (instance) or Global (model)

- Or in between: a group of instances

Result of the interpretation method

- on the next slide

Taxonomy: Result of the interpretation method



Feature summary statistic

$E[\text{feature importance}]$,
feature interaction strengths



Feature summary
visualization

Partial dependence plot,
feature importance plot



Model internals

Linear model weights,
decision tree structure,
filters



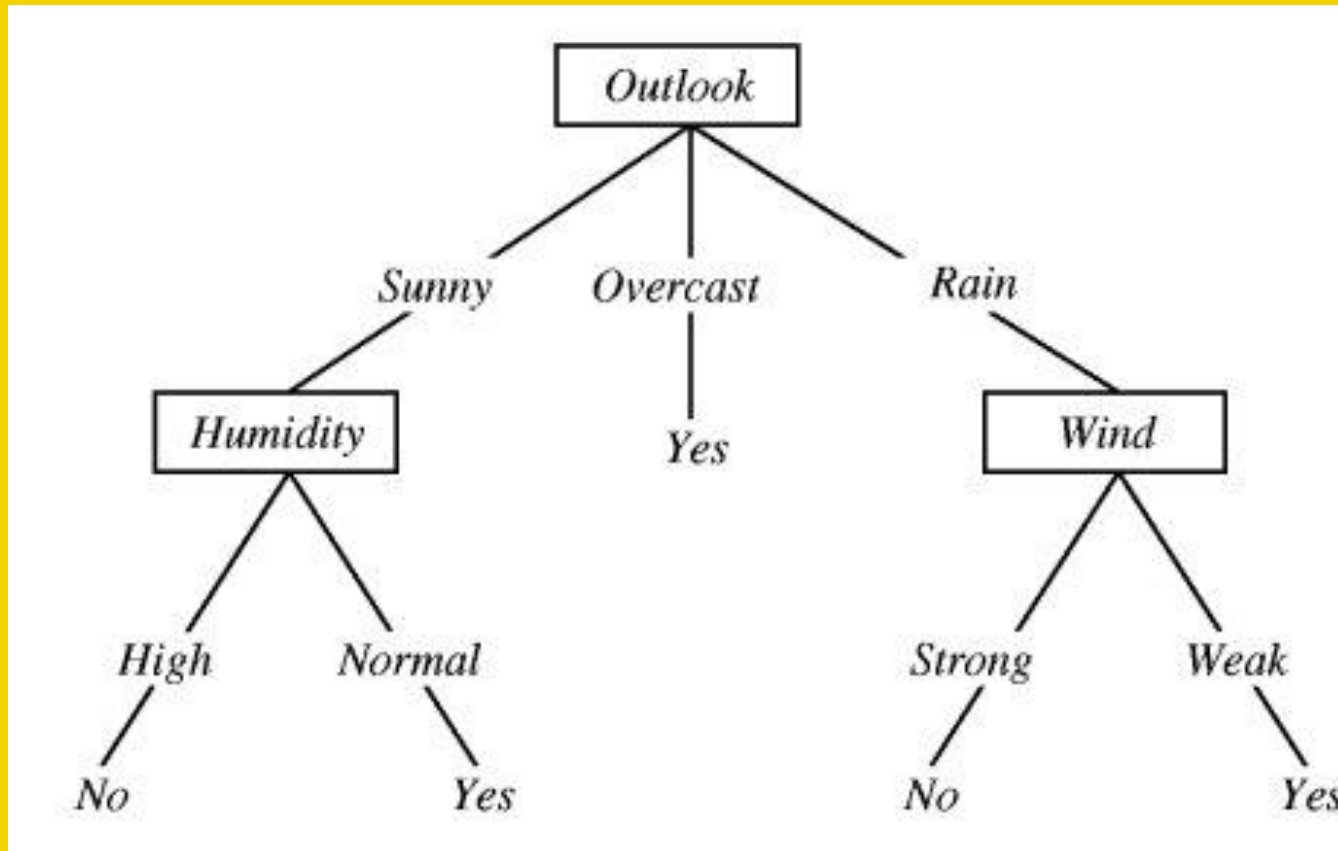
Data points

Exemplars in counterfactual
explanations



Global or local surrogates via intrinsically
interpretable models

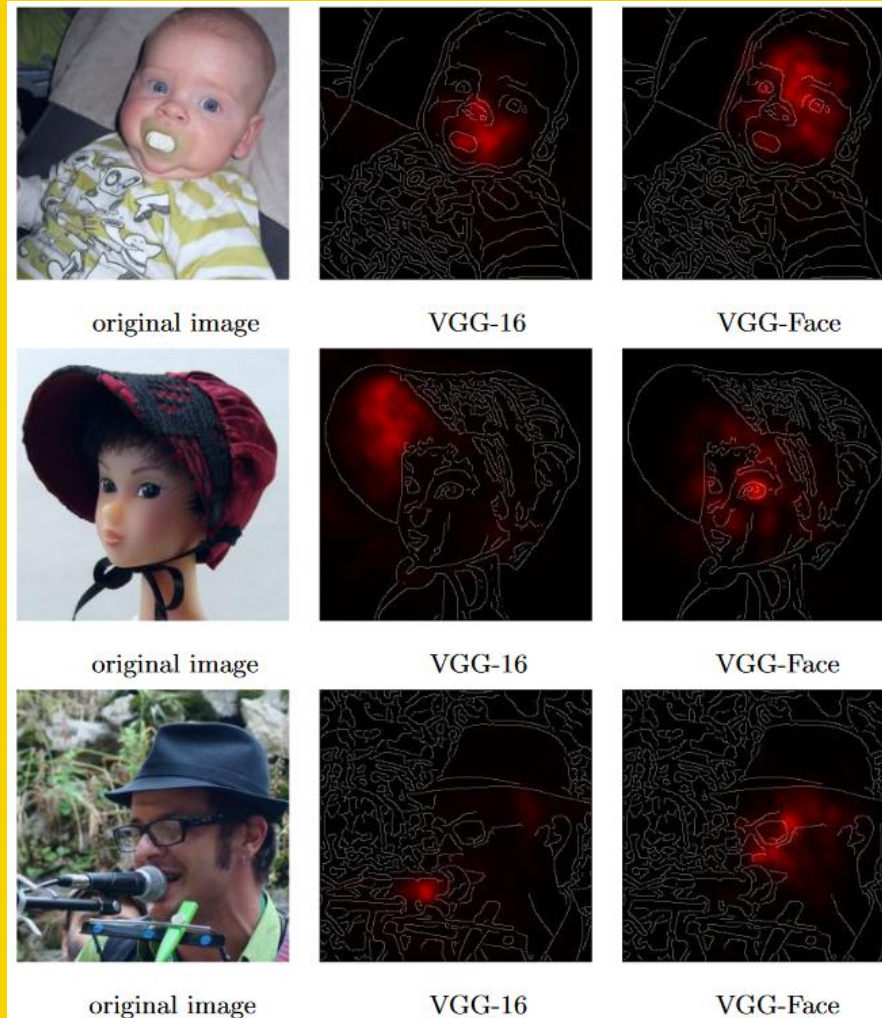
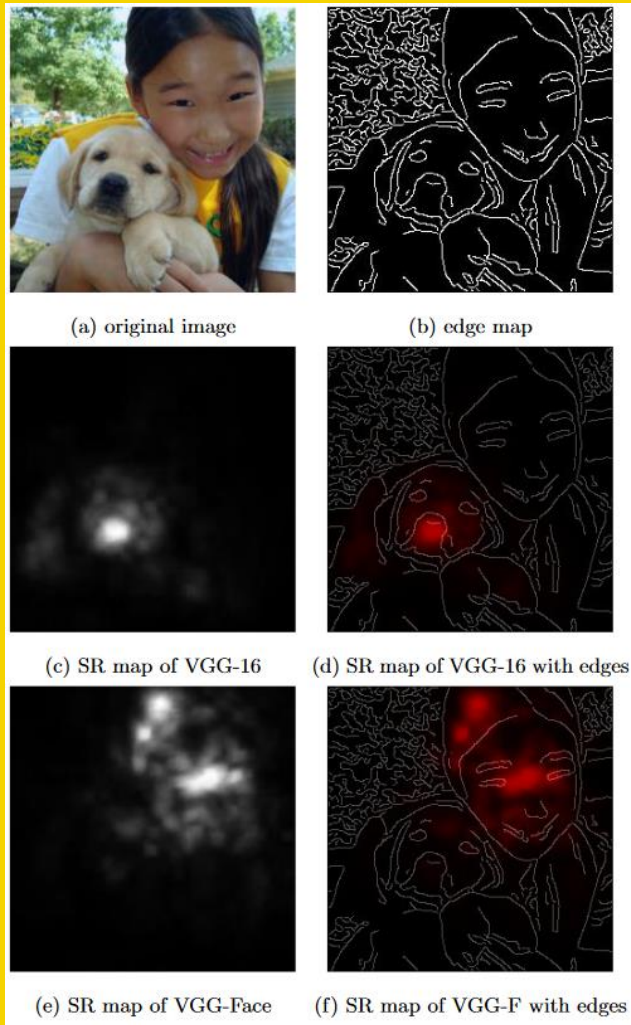
Ex. 1: Play Tennis Decision Tree



- Intrinsic or post-hoc
- Model-specific or -agnostic
- Global or Local
- Method Result

- Intrinsic
- Model specific
- Global & Local
- Model internals

Ex. 2: CNN Decision Areas^{*,#}



- Intrinsic or post-hoc
- Model-specific or -agnostic
- Global or Local
- Method Result

- Post-hoc
- Model specific
- Local
- Model internals

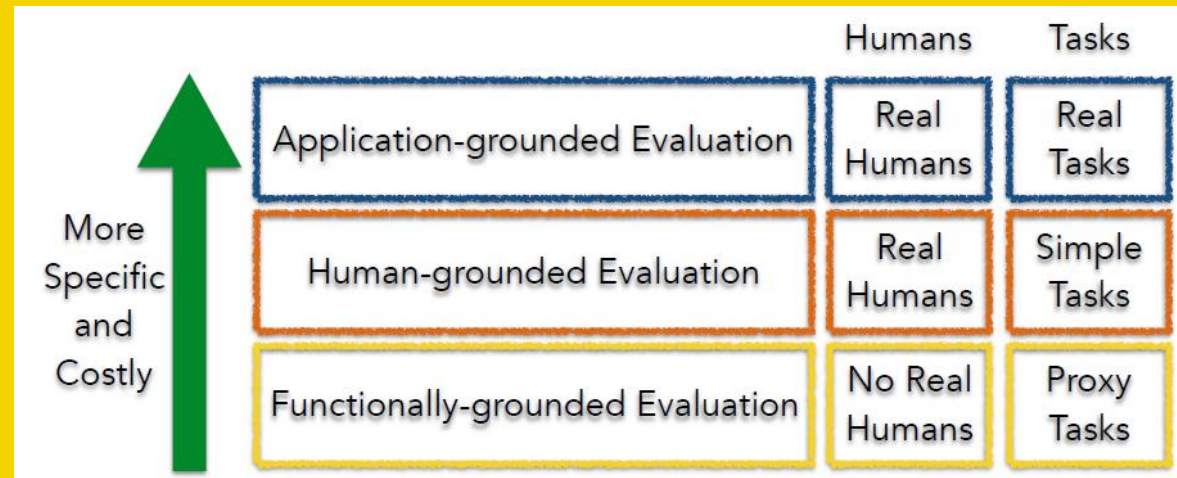


Scope of Interpretability

- Algorithmic Transparency
 - How does the algorithm generate the model?
- Global, Holistic Model Interpretability
 - How does the trained model make predictions?
 - Can we comprehend the entire model at once?
- Global Model Interpretability on a Modular Level
 - How do parts of the model affect predictions?
- Local Interpretability for a Single Prediction
 - Why did the model make a certain prediction for an instance?
- Local Interpretability for a Group of Predictions
 - Why did the model make specific predictions for a group of instances?
 - may be used for analyzing group-wise bias



Evaluation of interpretability



Taxonomy of evaluation approaches for interpretability. Adapted from [D]

- Application-level evaluation (real task)
 - Deploy the interpretation method on the application
 - Let the experts experiment and provide feedback
- Human-level evaluation (simple task)
 - During development, by lay people (AMT)
- Function-level evaluation (proxy task)
 - Does not use humans directly
 - Uses measures from a previous human evaluation
- All of above can be used for evaluating model interpretability as well as individual explanations



Properties of Explanation Methods

- Expressive Power
 - the "language" or structure of the explanations
 - E.g. IF-THEN rules, tree itself, natural language etc.
- Translucency
 - describes how much the explanation method relies on looking into the machine learning model
- Portability
 - describes the range of machine learning models with which the explanation method can be used
- Algorithmic Complexity
 - computational complexity of the explanation method



Properties of Individual Explanations

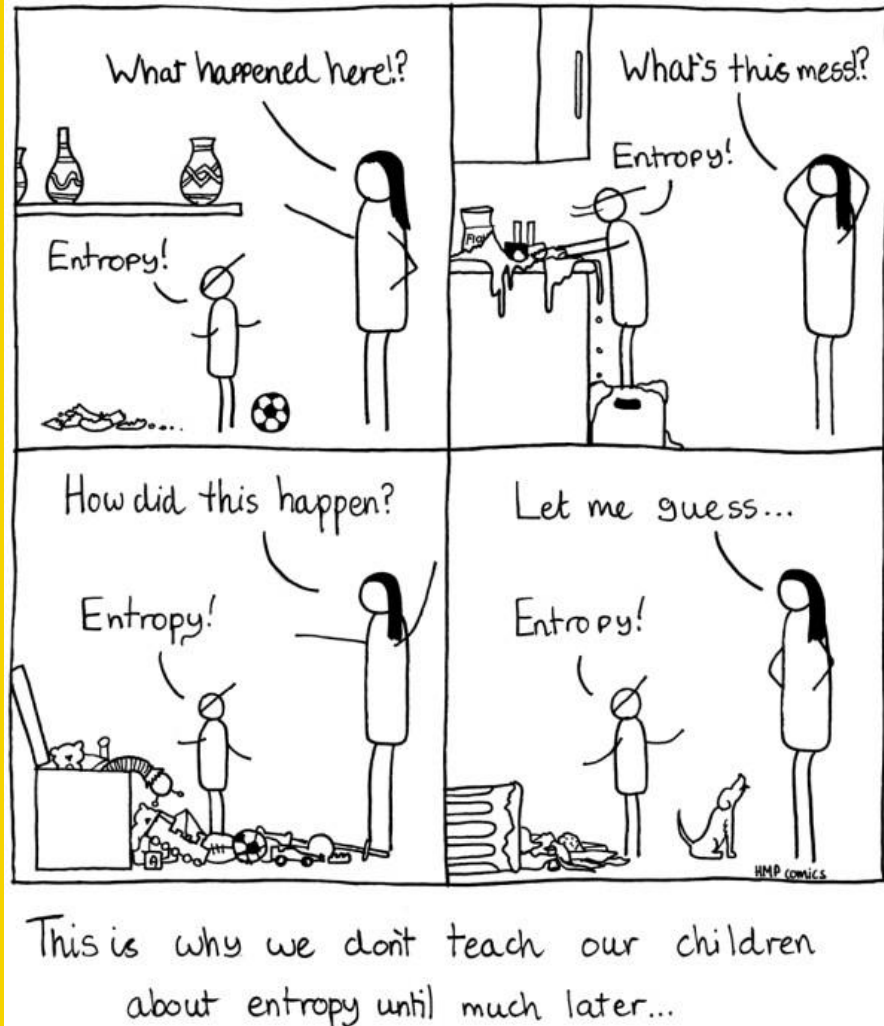
- **Accuracy**
 - How well does an explanation predict unseen data?
- **Fidelity**
 - How well does the explanation approximate the prediction of the black box model?
- **Certainty / confidence**
 - Does the explanation reflect the certainty of the machine learning model?
- **Comprehensibility / Plausibility**
 - How well do humans understand the explanations?
 - How convincing (trust building) are they?
 - Difficult to define and measure, but extremely important to get right. Any ideas for measuring?


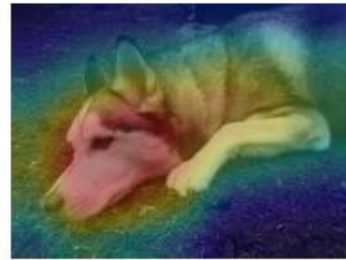



How is explainability usually measured?

- **Fidelity**
 - Should be measured objectively
 - Not all explanations can be checked for fidelity
- **Plausibility**
 - Requires a user study
- **Simulatability**
 - Measures the degree that a human can calculate / predict the model's outcome, given the explanation

What Is a Good Explanation?



	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

The same 'explanation' for different problems is not always good!
Figure from [R, C]

[R] Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 2019

[C] Checkermallow. Canis lupus winstonii (Siberian Husky); [Public domain image link](#)

What Is a Good Explanation?

- **Contrastive**
 - requires a point of reference for comparison
- **Selective**
 - precise, a small set of most important factors
 - *humans can handle at most 7 ± 2 cognitive entities at once* [R, M]
- **Social**
 - considers the social context (environment / audience)
- **Truthful (scientifically sound)**
 - Good explanations prove to be true in reality
- **General and probable**
 - A cause **that** can explain many events is very general and could be considered a good explanation.
- **Consistent with prior beliefs of the explainee**

Example Explanation: Real Estate Appraisal

Objectkenmerken:		Verschil	Toelichting taxateur waarbij de geconstateerde verschillen met het getaxeerde object worden weergegeven
Woningtype:	benedenwoning	vergelijkbaar	
Bouwjaar:	2008	-5	
Gebruiksoppervlakte wonen (m2):	83	0,0%	
Bruto inhoud (m3):	314	9,8%	
Perceeloppervlakte (m2):	---	---	
Energie label:	A (definitief)	vergelijkbaar	
Bij-, op- of aanbouwen:	- Berging / schuur (vrijst.)	vergelijkbaar	
Ligging:		vergelijkbaar	in woonwijk
Onderhoudssituatie:		vergelijkbaar	goed, netjes onderhouden
Mate van luxe en doelmatigheid:		beter	referentie heeft een hoogwaardiger afwerkingsniveau
Kwaliteit en conditie:		vergelijkbaar	goed, netjes afgewerkt en goed onderhouden

- Example comparison from a real estate appraisal (Dutch: taxatie) report.
- Comparison is made with a recently sold real estate in the neighborhood having similar conditions.
- Notice the contrastive and partly selective nature of the justification.

Why should we hire you?

- Consider the properties of good explanations and give a brief answer

Interpretability vs Explainability



- Individual terms may not have no precise definition, but
- **Interpretable models**
 - are transparent and simple enough to understand
 - stand for their own explanation
 - thus, their explanations reflect perfect fidelity*
- **Black-box models**
 - require post-hoc explanations (*as an excuse to their opacity [R1]*)
 - *cannot have perfect fidelity with respect to the original model [R2]*
 - *their explanations often do not make sense, or do not provide enough detail to understand what the black box is doing [R2]*
 - are often not compatible with situations where information outside the database needs to be combined with a risk assessment [R2]

[R1] Rudin et al., Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, ArXiv 2021

[R2] Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 2019

Interpretability as a Design Choice

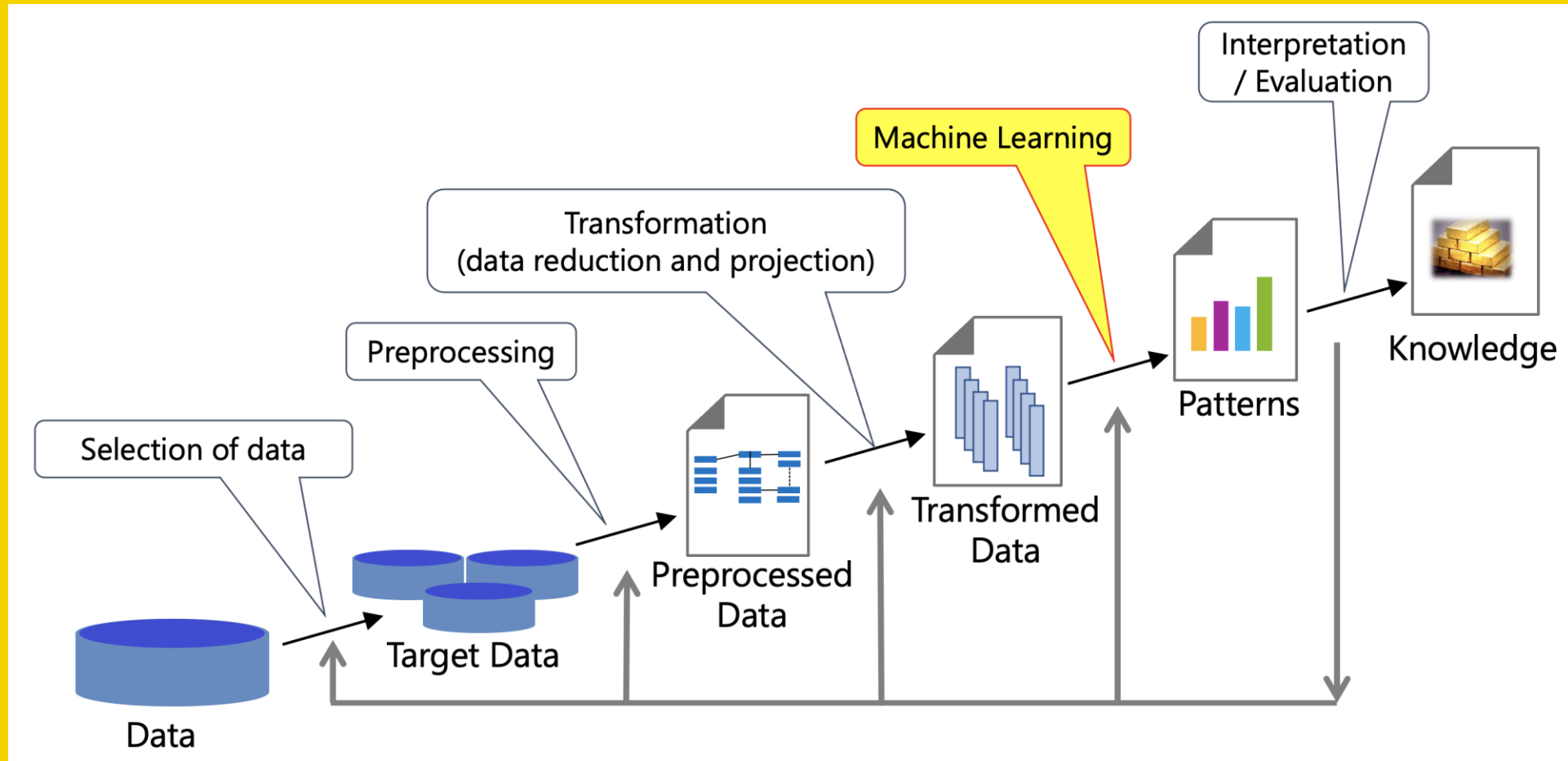


Figure: Knowledge Discovery Process (figure from [R], adapted from [F])

[F] Fayyad et al., From data mining to knowledge discovery in databases. AI Magazine, 1996

[R] Rudin et al., Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, ArXiv 2021

Rule of Thumb: When to use what?

Models	Data Type
Decision Trees / Decision Lists / Decision Sets	somewhat clean tabular data with interactions, including multiclass problems. Particularly useful for categorical data with complex interactions (i.e., > quadratic).
Scoring Systems	somewhat clean tabular data, typically used in medicine and criminal justice. The models are small enough that they can be memorized by humans.
Generalized Additive Models (GAMs)	continuous data with at most quadratic interactions, useful for raw medical records.
Case-based reasoning	any data type (different methods exist for different data types), including multiclass problems.
Disentangled Neural Networks	data with raw inputs (computer vision, time series, textual data), suitable for multiclass problems.

Table: Rule of thumb for the types of data that naturally apply to various supervised learning algorithms. Adapted from [R]

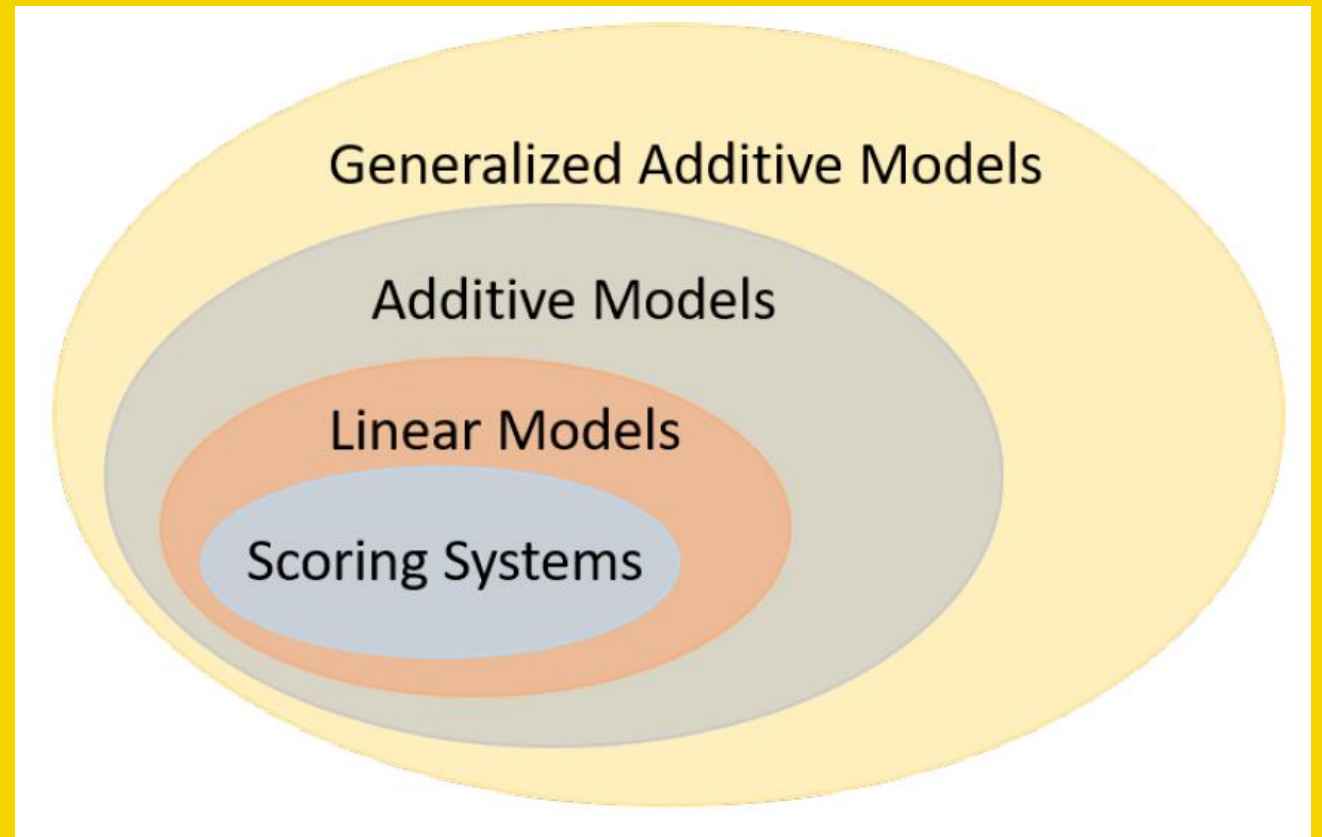


Interpretable Models

- The set is growing with new methods
- We will focus on two families
 - Linear models
 - **Linear regression**, perceptron, SVM with a linear kernel, logistic regression, **GAMs** etc.
 - **Decision trees** and decision rules
- And a combination of the two, namely
 - **RuleFit**



Linear Models

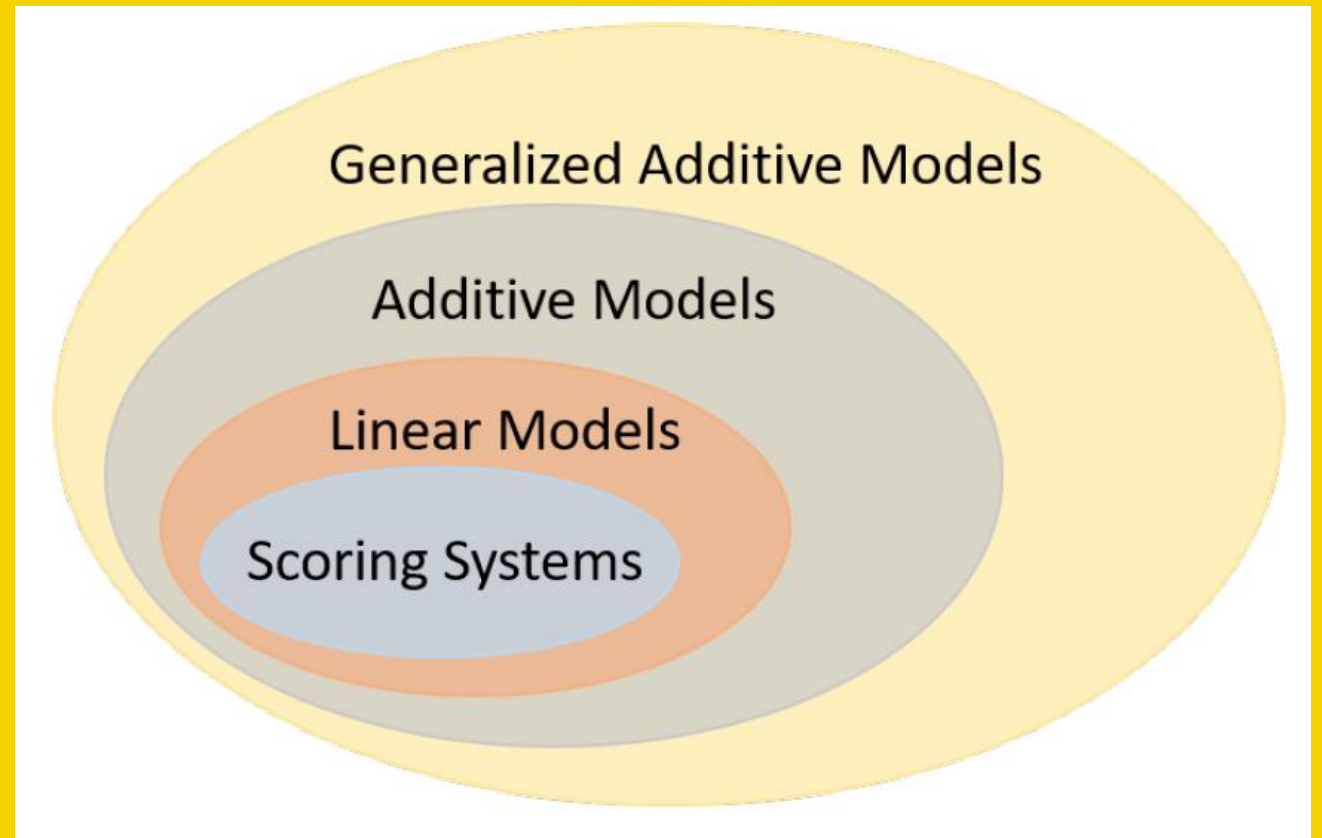


Hierarchy of linear, generalized linear and generalized additive models. From [R]

- **Linear models:** $g(E[y^t]) = g(x^t) = w_0 + w_1 x_1^t + w_2 x_2^t + \dots + w_d x_d^t$
where x_i^t is i^{th} feature of instance t
- **Scoring systems:** A model whose linear weights are constrained to be integers (e.g. credit risk scoring)



Generalized Additive Models



Hierarchy of linear, generalized linear and generalized additive models. From [R]

- **Linear models:** $g(E[y^t]) = g(x^t) = w_0 + w_1 x_1^t + w_2 x_2^t + \dots + w_d x_d^t$
- **GAM:** $h(E[y^t]) = h(x^t) = w_0 + f_1(x_1^t) + f_2(x_2^t) + \dots + f_d(x_d^t)$

where $h(\cdot)$ is a link function and the f_i 's are univariate component functions that are possibly nonlinear [R]

Multivariate Regression*

- Multivariate linear model $r^t = g(x^t | w_0, w_1, \dots, w_d) + \varepsilon$

$$g(x^t) = w_0 + w_1 x_1^t + w_2 x_2^t + \dots + w_d x_d^t$$

- Multivariate polynomial model:

Define new higher-order variables

$$z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 x_2$$

and use the linear model in this new z space

The transformations of the latter form: **Generalized Linear Models**



Assumptions

- Linearity
 - $f(x+y)=f(x)+f(y)$, $f(cx)=cf(x)$
- Normality of the target variable
- Homoscedasticity: constant variance
- Independent instance distribution
- Absence of multicollinearity
 - No pairs of strongly correlated features
 - E.g. house size and number of rooms

Interpretation of Linear Models

- Modular view: we assume all remaining feature values are fixed
- Numerical feature weight
 - Increasing the numerical feature by one unit changes the estimated outcome by its weight.
- Binary feature weight
 - The contribution of the feature when it is set to 1.
- Categorical feature with L categories
 - Carry out one-hot-encoding into L binary features
 - Eg. 3 levels: $1 \rightarrow [1\ 0\ 0]$, $3 \rightarrow [0\ 0\ 1]$
 - Remove a (default) category (redundance)
 - Interpret the remaining $L-1$ features as binary
- Intercept: Prediction value when all features are 0.

Feature Importance

- Feature importances are calculated over t-statistics: $f_i = w_i / SE(w_i)$, where $SE()$ is standard error of estimation.
- The features are ranked w.r.t. absolute value of the t-statistics

Example: Combined Cycle Power Plant Data

Task: Predicting the Gas Power Plant's Net Energy Yield using Ambient Variables

Attributes	w	SE	t-Statistic
Intercept (bias var)	519,42	0,834	622,830
Ambient Temperature	-2,07	0,030	-68,108
Exhaust Vacuum	-0,21	0,016	-13,631
Relative Humidity	-0,18	0,009	-20,158

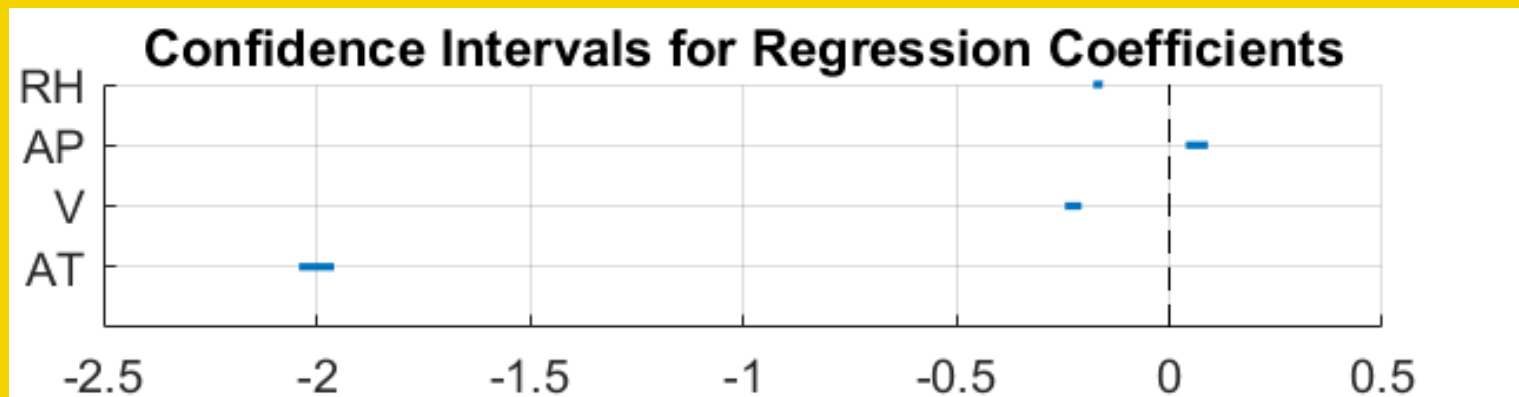
Confidence Interval Calculation

- Confidence Interval of estimated weights is a function of the estimated weights, SE and the z-scores: $CI(w_i)_{\alpha} = [w_i - z_{\alpha/2} \times SE(w_i) , w_i + z_{\alpha/2} \times SE(w_i)]$

Confidence Level	α (level of significance)	$z_{\alpha/2}$
99%	0.01	2.575
95%	0.05	1.960
90%	0.10	1.645

Therefore, for the 95% confidence case, we have:

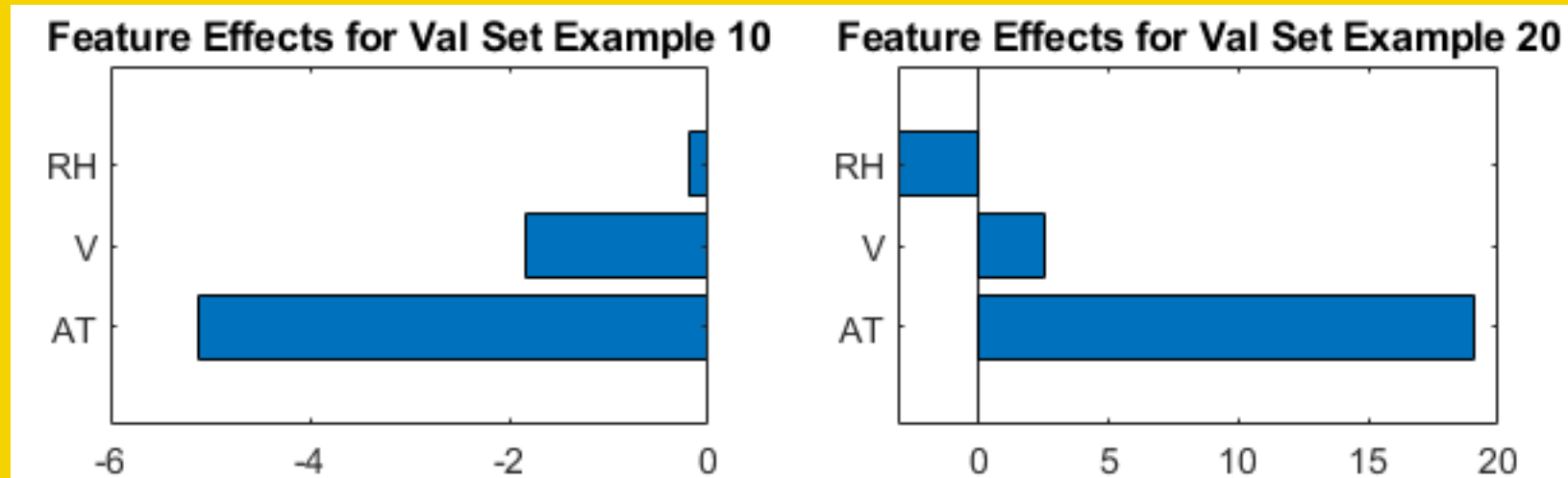
$$CI(w_i)_{0.05} = [w_i - 1.96 \times SE(w_i) , w_i + 1.96 \times SE(w_i)]$$



Combined Cycle Power Plant Data
Linear Model Confidence Intervals
with $\alpha = 0.05$ (i.e. 95% confidence)

Explaining a Prediction: Feature Effects

- Feature Effects are multiplication of the estimated weight and the (mean normalized) feature value: $FE_i = w_i x_i$
- NB: In case the features are very close to mean values, mean normalized vector is close to zero
 - this makes the intercept term interpretation meaningful



Generalized Additive Models (GAMs)

- GAM: $h(E[y^t]) = h(x^t) = w_0 + f_1(x_1^t) + f_2(x_2^t) + \dots + f_d(x_d^t)$
- How to model nonlinear component functions $f_j(x_j)$?

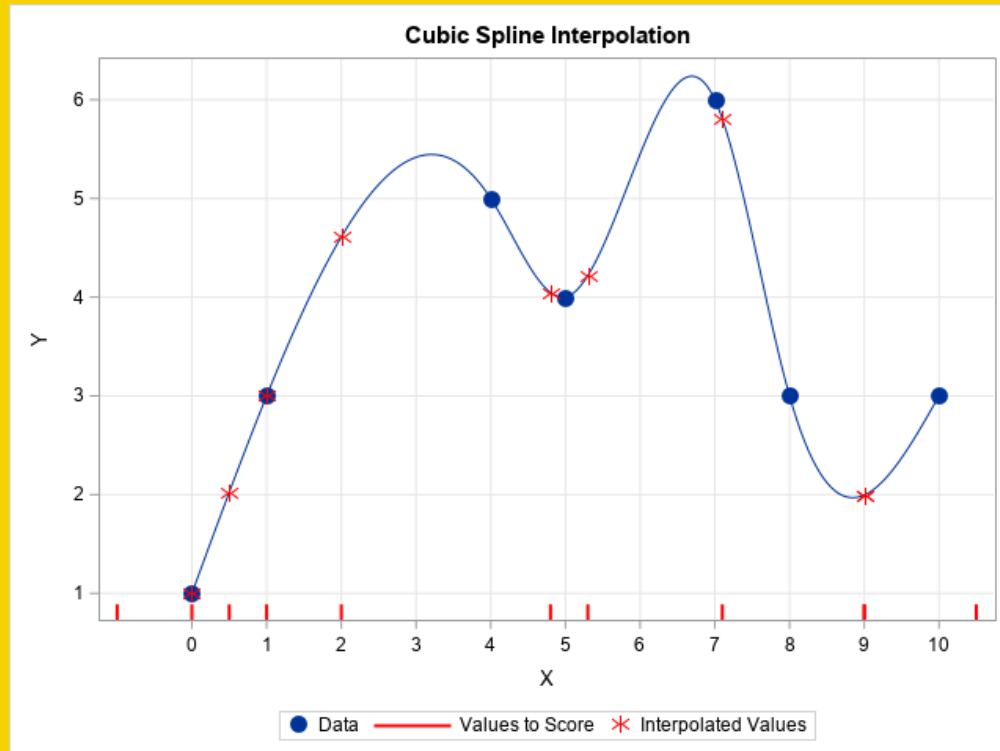


Figure from [SAS]

[SAS] <https://blogs.sas.com/content/iml/2020/05/11/cubic-interpolation-sas.html>

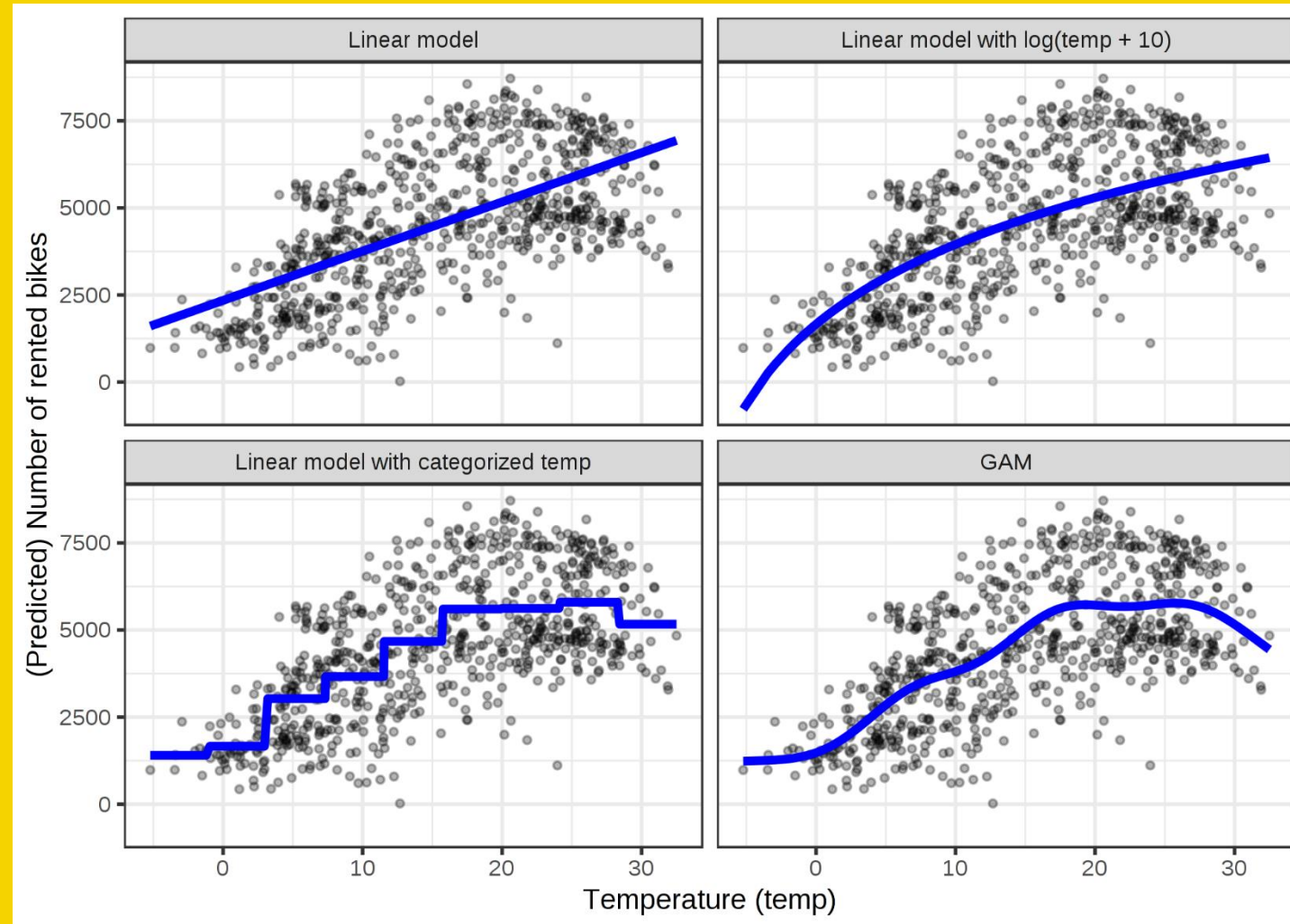
Continuous variables

- Splines
 - Smooth, continuous nonlinear modeling
 - May not model surges/jumps well
- Weighted sum of indicator functions:
$$f_j(x_j) = \sum_{thresholds\ j'} c_{j,j'} \mathbf{1}[x_j > \theta_{j'}]$$
 - can be trained using decision stumps
 - If weights are integer and sparse?

Generalized Additive Models (GAMs)

- How to model nonlinear component functions $f_j(x_j)$?
 - How to model discrete variables?
 - How to model feature interactions $f_{jk}(x_j, x_k)$?
- How to learn the GAM as a whole?
 - We iteratively add new components to minimize the residual error
 - w_0 : can be set to $E[y]$ (estimated from the training set)
 - Step K: calculate the residual: $r = y - (w_0 + \sum_{j=1}^{K-1} f_j(x_j))$
 - Fit a model to the residual (a DT with limited depth or splines)
- Very popular: explainable and accurate implementations exist
 - See e.g. EBM in interpretml [1]

Comparative GAM Example



Predicting the number of rented bicycles using only temperature

- Top right: GLM
 - Bottom left: GAM
- GAM is powerful and eases feature-wise interpretation
 - GAM can easily benefit from Partial Dependence Plots (we will cover in L5)



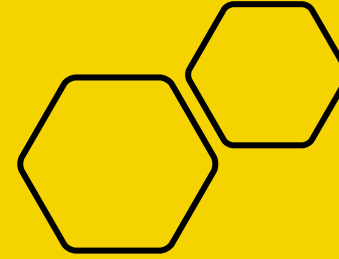
Interpretable Models

- The set is growing with new methods
- We will focus on two families
 - Linear models
 - Linear regression, perceptron, SVM with a linear kernel, logistic regression, GAMs etc.
 - **Decision trees** and decision rules
- And a combination of the two, namely
 - RuleFit

Decision Tree Interpretation

- Model Interpretation
 - Visual: Visualization of the tree itself
 - Textual: The set of rules (conjunction of the nodes from root to each leaf)
- Individual Explanation
 - Output the rule that applies (as IF-THEN or natural language)
- Feature Importances
 - Sum the *split merit* (e.g. *information gain*) for each feature
 - Normalize the overall sum to 1 or 100
- Pruning may be needed to improve generalization and intelligibility

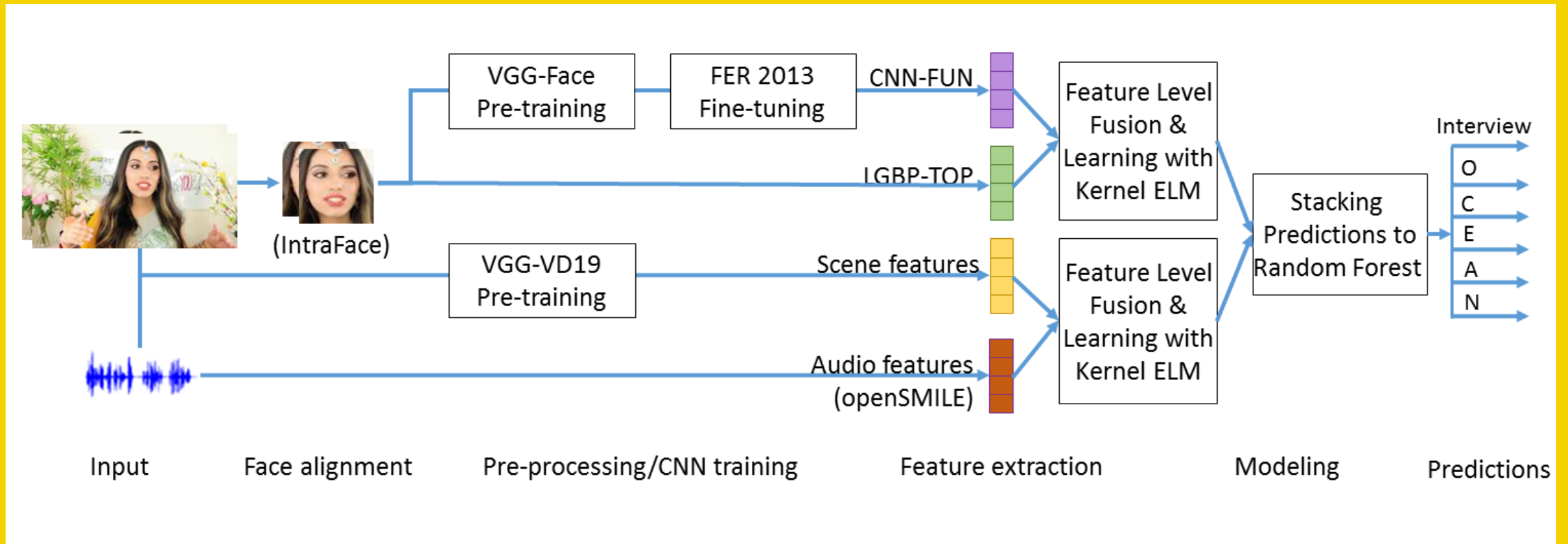
Explainability Example with Decision Trees



Explainable Job Interview
Recommendations

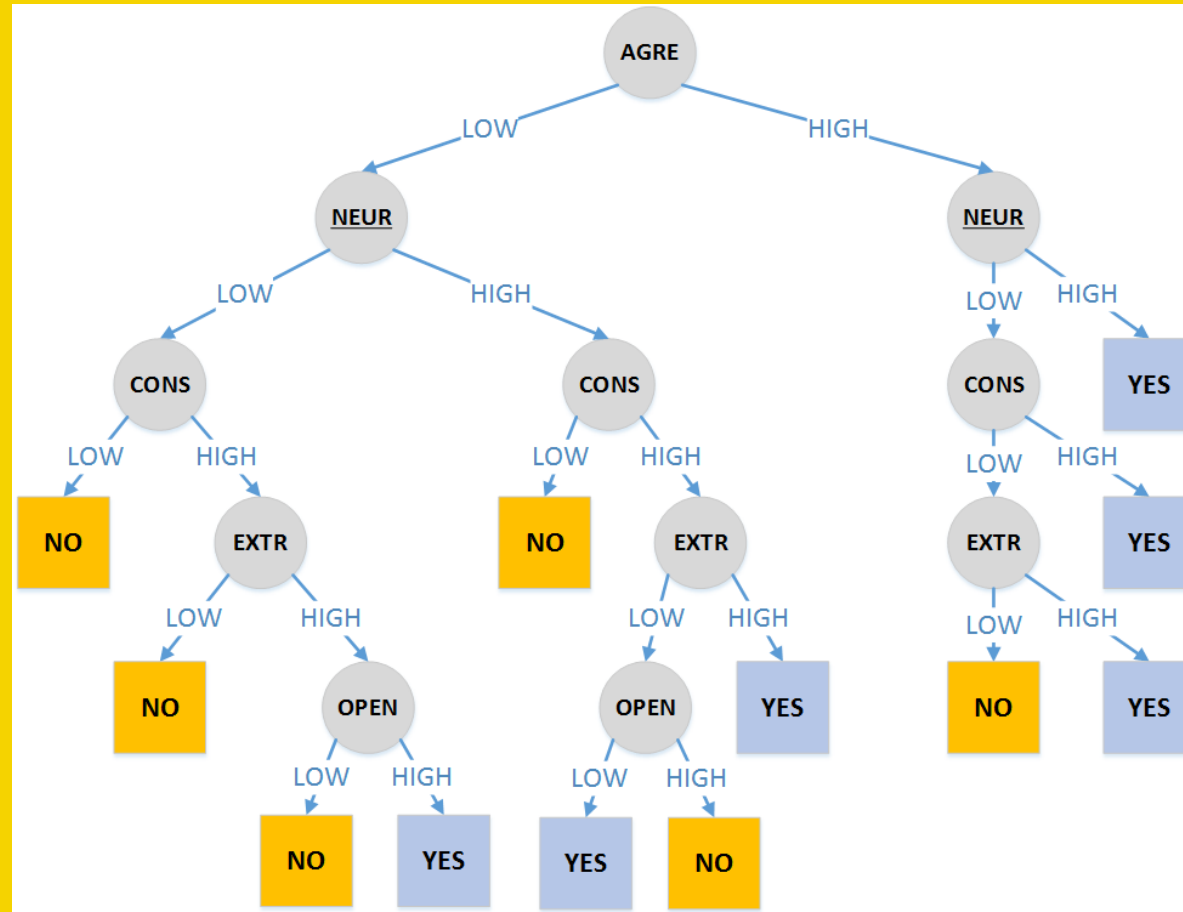
Predicting the Personality Impressions

- The winner system of CVPRW'17 ChaLearn Explainable Job Interview Recommendation Competition*



*Kaya et al., Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. *CVPRW 2017*.

Interview Invitation Explanation Model



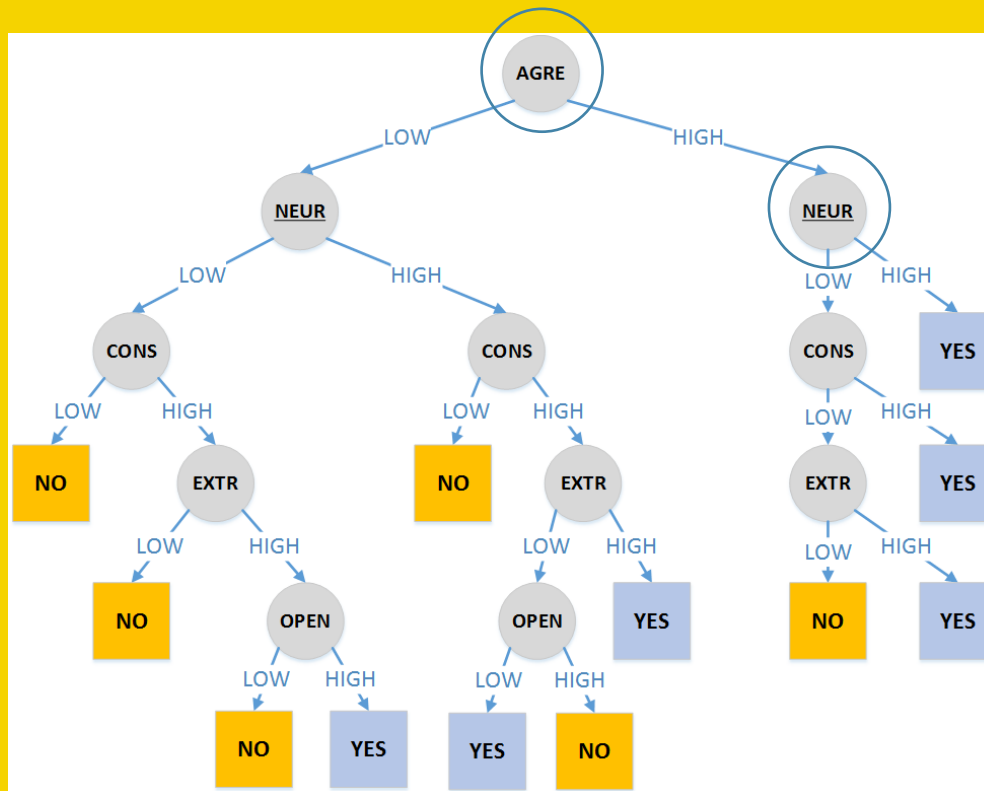
*Kaya et al., Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. *CVPRW* 2017.

Explanations

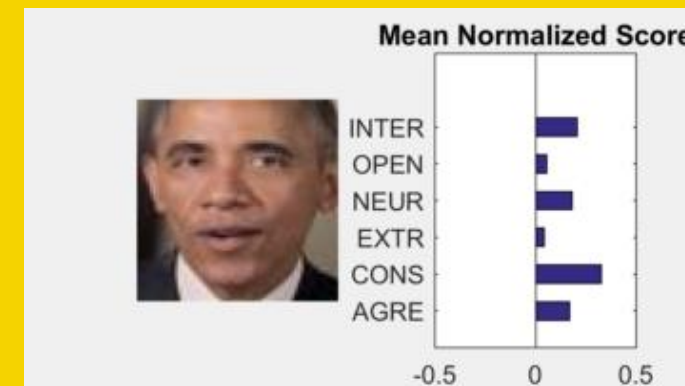
- If invite decision is 'YES'
 - 'This [gentleman/lady] is invited due to [his/her] high apparent *{list of high scores on the trace}*' [optional depending on path: ', although low *{list of low scores on the trace}* is observed.']
- If invite decision is 'NO'
 - This [gentleman/lady] is not invited due to [his/her] low apparent *{list of low scores on the trace}*' [optional depending on path: ', although high *{list of high scores on the trace}* is observed.']
- If the direct and indirect predictions get in conflict
 - The directly predicted interview score and the classification based on traits are not consistent, the [gentleman/lady] may be re-evaluated. Following explanation is based on predicted traits.
- We also check which modality is dominant
 - If the face system has the same sign the the final results it is visual
 - Else (speech has higher effect than scene) it is audio system

Explaining Decision Trees

- model explanation \rightarrow the illustration of the tree
- instance explanation \rightarrow conjunction of the nodes from root to leaf



Visual Explanation



Automatic Verbal Explanation

This gentleman is invited for an interview due to his high apparent agreeableness and non-neuroticism impression.

Bias Problem

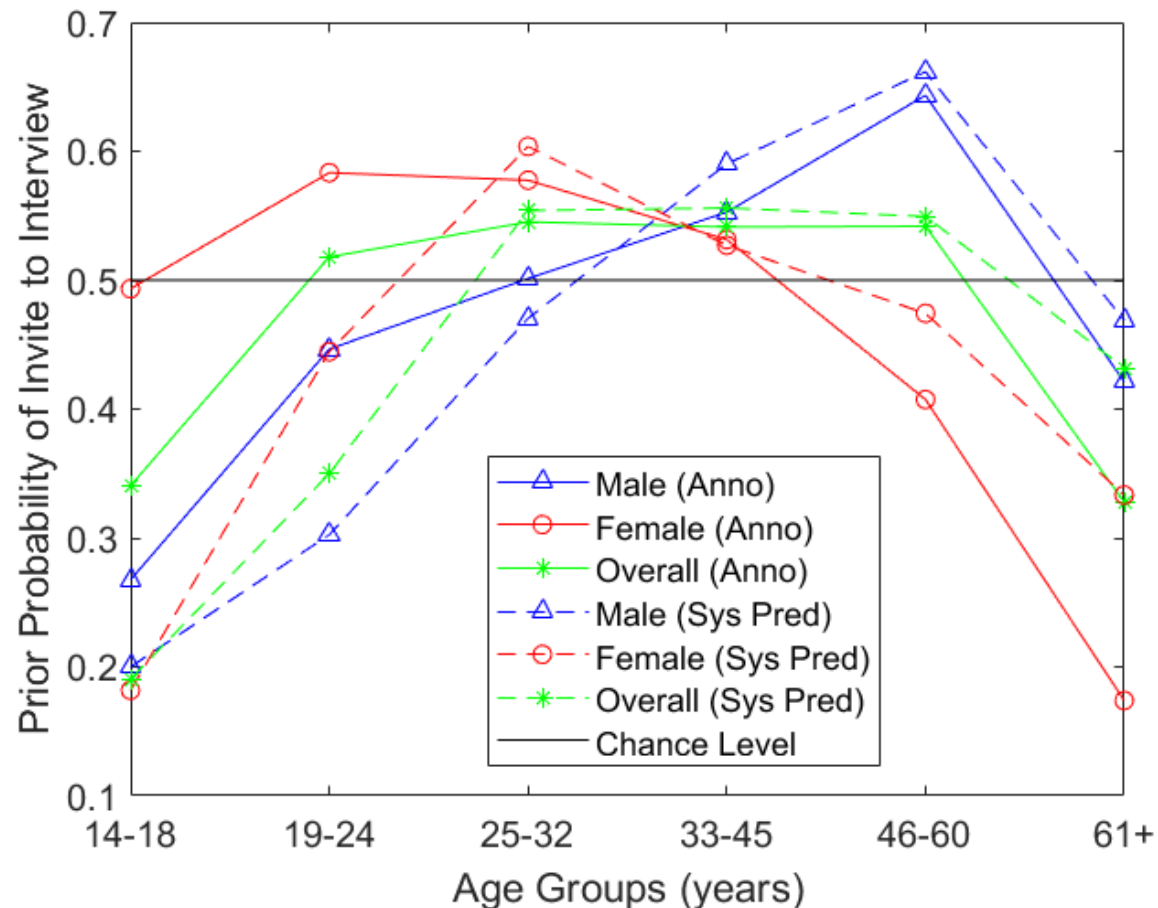
Pearson Correlations Among Traits and Personality Impressions

Correlation	Gender	Ethnicity		
Dimension	Female	Asian	Caucasian	Afro-American
<i>Agreeableness</i>	-0.023	-0.002	0.061**	-0.068**
<i>Conscientiousness</i>	0.081**	0.018	0.056**	-0.074**
<i>Extroversion</i>	0.207**	0.039*	0.039*	-0.068**
<i>Neuroticism</i>	0.054*	-0.002	0.047*	-0.053**
<i>Openness</i>	0.169**	0.010	0.083**	-0.100**
Interview	0.069**	0.015	0.052*	-0.068**

Prior Probabilities of «Invite to Interview»

	Male	Female	Asian	Caucasian	Afro-American
mean scores	0.539	0.589	0.515	0.507	0.475
p(invite trait)	0.495	0.560	0.562	0.539	0.444

Accurate, explainable but clearly biased!



- Notice the age-gender bias patterns carried on!
- Both human annotations and the model favors
 - Younger female candidates and
 - Older male candidates
- In working age range ([18, 60]), increasing age disfavors females while favoring males

Figure from Escalante et al. [E]

Recent papers also analyzed the bias in this system [Y, K]

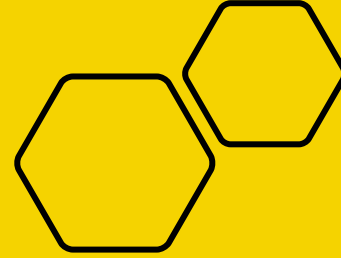
[E] Escalante, Kaya, Salah et al., Modeling, recognizing, and explaining apparent personality from videos, *Transactions on Affective Computing*, 2020

[Y] Yan et al., Mitigating Biases in Multimodal Personality Assessment, *ICMI 2020*

[K] Köchling et al., Highly Accurate, But Still Discriminatory, *Business & Information Systems Engineering*, 2021

RuleFit

Combining Strengths of
Trees and Linear Models



Useful Remarks

	Modeling Linearity	Modeling Feature Interactions	Multiclass classification	Interpretable Modeling of Regression
Linear Models	Powerful	Poor	Poor	Powerful
Decision Trees	Poor	Powerful	Powerful	Poor

Can we combine the strengths of linear models and decision trees / rules?



RuleFit [F]

- The main idea is to learn **binary rules** from data and use them as features together with the original features in a **sparse linear model**
- Rules are generated using ensemble of decision trees (e. g. Boosting)
- Rules mentioned here are the conjunctions from root to any node:
 - $r_m(x) = \prod_{j \in T_m} I(x_j \in s_{jm})$
- T_m : the set of features used in r_m
- s_{jm} : a subset or range of feature j

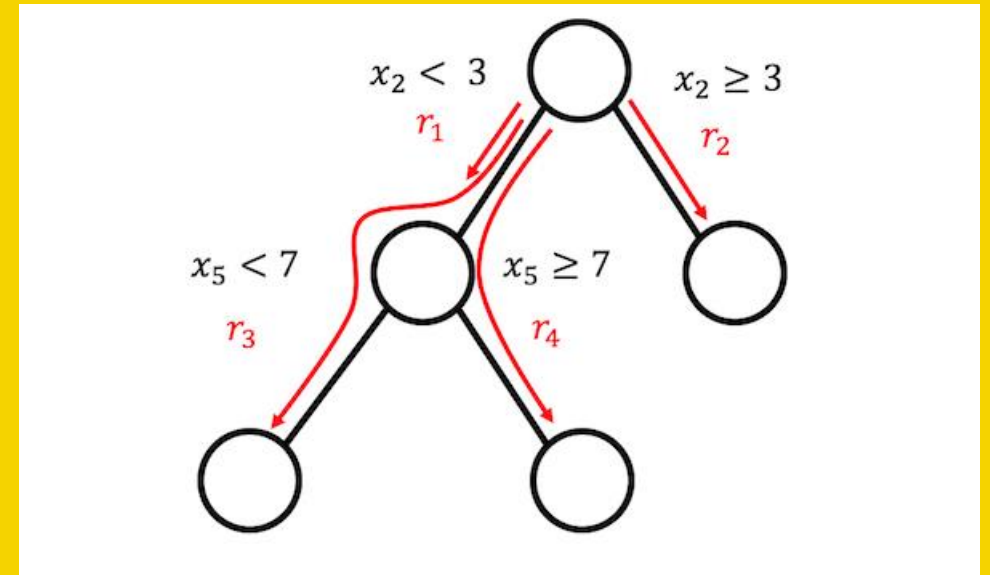


Figure: Illustration of rules in RuleFit. From [M]

[F] Friedman and Popescu, Predictive learning via rule ensembles. Annals of Applied Statistics, 2008.

[M] Molnar, Interpretml book, <https://christophm.github.io/interpretable-ml-book/rulefit.html>

RuleFit - Example

Table: An illustration of learned rules using 'Bike Rental Dataset'. (from [M])

Description	Weight	Importance
days_since_2011 > 111 & weathersit in ("GOOD", "MISTY")	793	303
37.25 <= hum <= 90	-20	272
temp > 13 & days_since_2011 > 554	676	239
4 <= windspeed <= 24	-41	202
days_since_2011 > 428 & temp > 5	366	179

RuleFit – Preprocessing for the Linear Model

- Rules $r_k(x)$ are standardized to ensure equal apriori effect,
 - $r_k(x) = \frac{r_k(x)}{t_k}$, where $t_k = \sqrt{s_k(1 - s_k)}$ and $s_k = \frac{1}{n} \sum_{i=1}^n r_k(x^{(i)})$
 - s_k is support of r_k and t_k is its scale (standard deviation)
- Linear terms x_j are preprocessed with the following steps:
 - 1- Sanitization (Winsorization): $l_j^*(x_j) = \min(\delta_j^-, \max(\delta_j^+, x_j))$
where δ_j^+ and δ_j^- are δ and $1 - \delta$ quantiles of feature x_j , respectively
 - 2- Standardization:
$$l_j(x_j) = \frac{l_j^*(x_j)}{\text{std}(l_j^*(x_j))}, \text{ if rules are standardized}$$
$$l_j(x_j) = 0.4 \frac{l_j^*(x_j)}{\text{std}(l_j^*(x_j))}, \text{ if rules are not standardized (0.4 is average } t_k)$$

RuleFit – Learning the Sparse Linear Model

- The rules and original features are used to learn a sparse linear model

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{k=1}^K \hat{\alpha}_k r_k(x) + \sum_{j=1}^p \hat{\beta}_j l_j(x_j)$$

using Least Absolute Shrinkage Operator (LASSO):

$$(\{\hat{\alpha}\}_1^K, \{\hat{\beta}\}_0^p) = \underset{\{\hat{\alpha}\}_1^K, \{\hat{\beta}\}_0^p}{\operatorname{argmin}} \sum_{i=1}^n L(y^{(i)}, f(x^{(i)})) + \lambda \cdot \left(\sum_{k=1}^K |\alpha_k| + \sum_{j=1}^p |b_j| \right)$$

where λ is the regularization parameter.

RuleFit Interpretation: Feature Importance

- We can have two options to interpret the weights $\hat{\alpha}_k$ and $\hat{\beta}_j$
 - Mixed: interpret weights from rule and linear term together
 - Back to original: breakdown the rules and obtain original feature importances
- In RuleFit feature importance is a function of feature weight and scale
 - Linear term importance: $I_j = |\hat{\beta}_j| * std(l_j(x_j))$
 - Rule importance: $I_k = |\hat{\alpha}_k| * \sqrt{s_k(1 - s_k)}$
- For total feature importance, we first calculate individual effects
 - $J_j(x^{(i)}) = I_j(x^{(i)}) + \sum_{x_j^{(i)} \in r_k} \frac{I_k}{m_k}$, where m_k is the # of features used in r_k
 - Then the total (original feature) importance is $J_j(X) = \sum_{i=1}^n J_j(x^{(i)})$

Outstanding Questions

- So how do we explain individual prediction in RuleFit?
- How can we alternatively calculate the feature importances?
 - Consider feature importance calculations we use in Linear models and DTs.

Literature for today

- C. Molnar's online book, [Chapter 2](#) and [Chapter 4](#)
- Doshi-Velez and Kim, Towards A Rigorous Science of Interpretable Machine Learning, ArXiv 2017
- Rudin et al., Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, ArXiv 2021
 - (particularly up to and including Section 3)
- If you are interested you are encouraged to read the papers referenced under respective slides

Literature for the next lecture

- Lecture 5: Model Agnostic Interpretability Methods
- Required Reading
 - C. Molnar's book, [Chapter 5](#)
 - [LIME] "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al. KDD 2016 (<https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>)
- Suggested Reading
 - [SHAP] A Unified Approach to Interpreting Model Predictions, Lundberg and Lee, NeurIPS 2017 (<http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>)

Quiz for today



A quiz for this lecture is available on course Blackboard page.



This is not graded, but meant to provide formative evaluation.



The deadline for submitting the quiz is Friday May 21, 19:00!



We will be discussing the answers next Tuesday.