Attribute Interactions

in Machine Learning

Aleks Jakulin

Faculty of Computer and Information Science

University of Ljubljana

A Classification Problem

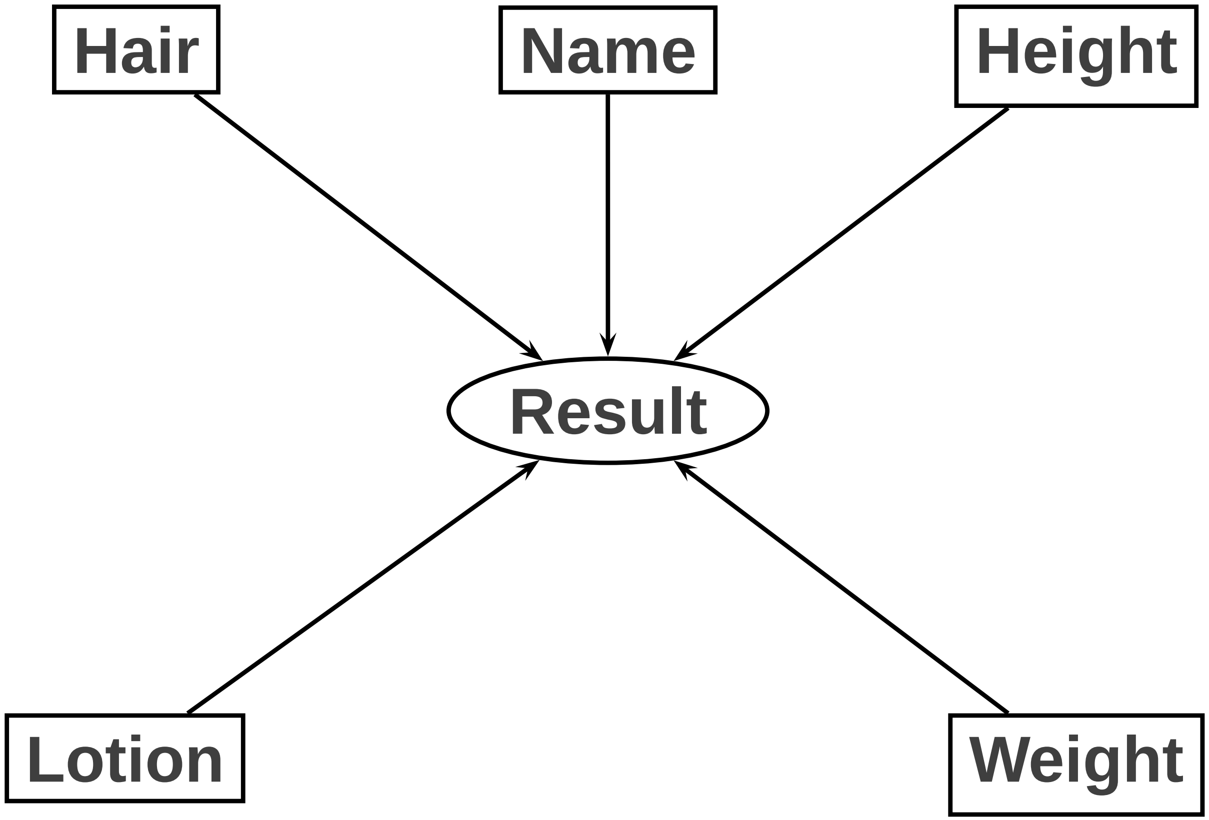
	ATTRIBUTES				LABEL
Name	**Hair**	**Height**	**Weight**	**Lotion**	**Result**
Sarah	blonde	average	light	no	*sunburned*
Dana	blonde	tall	average	yes	*tanned*
Alex	brown	short	average	yes	*tanned*
Annie	blonde	short	average	no	*sunburned*
Emily	red	average	heavy	no	*sunburned*
Pete	brown	tall	heavy	no	*tanned*
John	brown	average	heavy	no	*tanned*
Katie	blonde	short	light	yes	*tanned*

TASK: PREDICT AN INSTANCE'S CLASS GIVEN THE ATTRIBUTE VALUES.

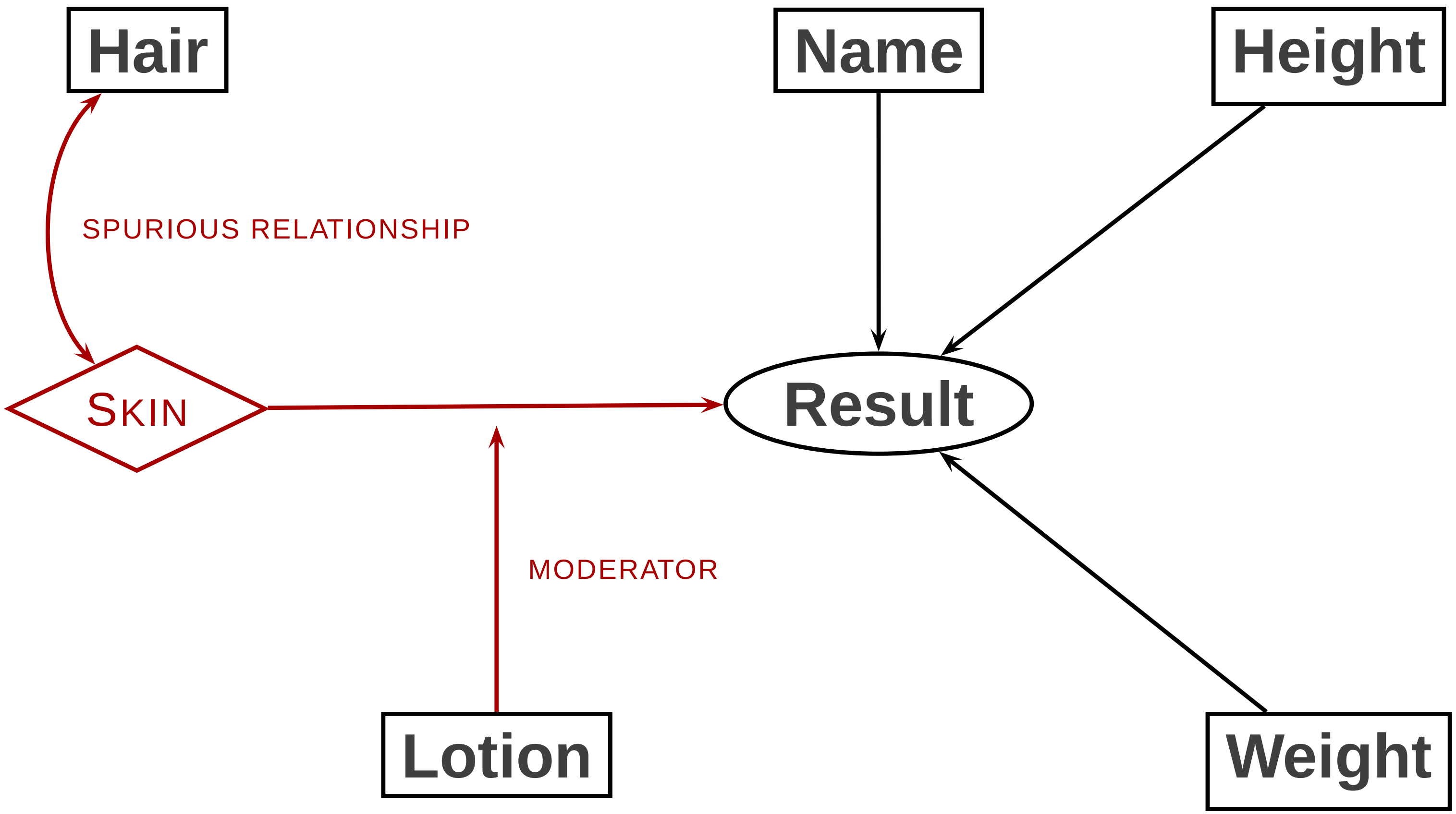
Interactions

- "We cannot conquer a group of interacting attributes by dividing them."
- Most machine learning algorithms assume either
  - that all attributes are independent (naïve Bayes, logistic regression, linear SVM, perceptron),
  - or that all attributes are dependent (classification trees, constructive induction, rules, kernel methods, instance-based methods).
- However, voting ensembles, where a number of classifiers trained on subsets of attributes or instances vote to predict the label (attribute decomposition, random forests, decision graphs, subspace methods), yield good results. Why?

Voting

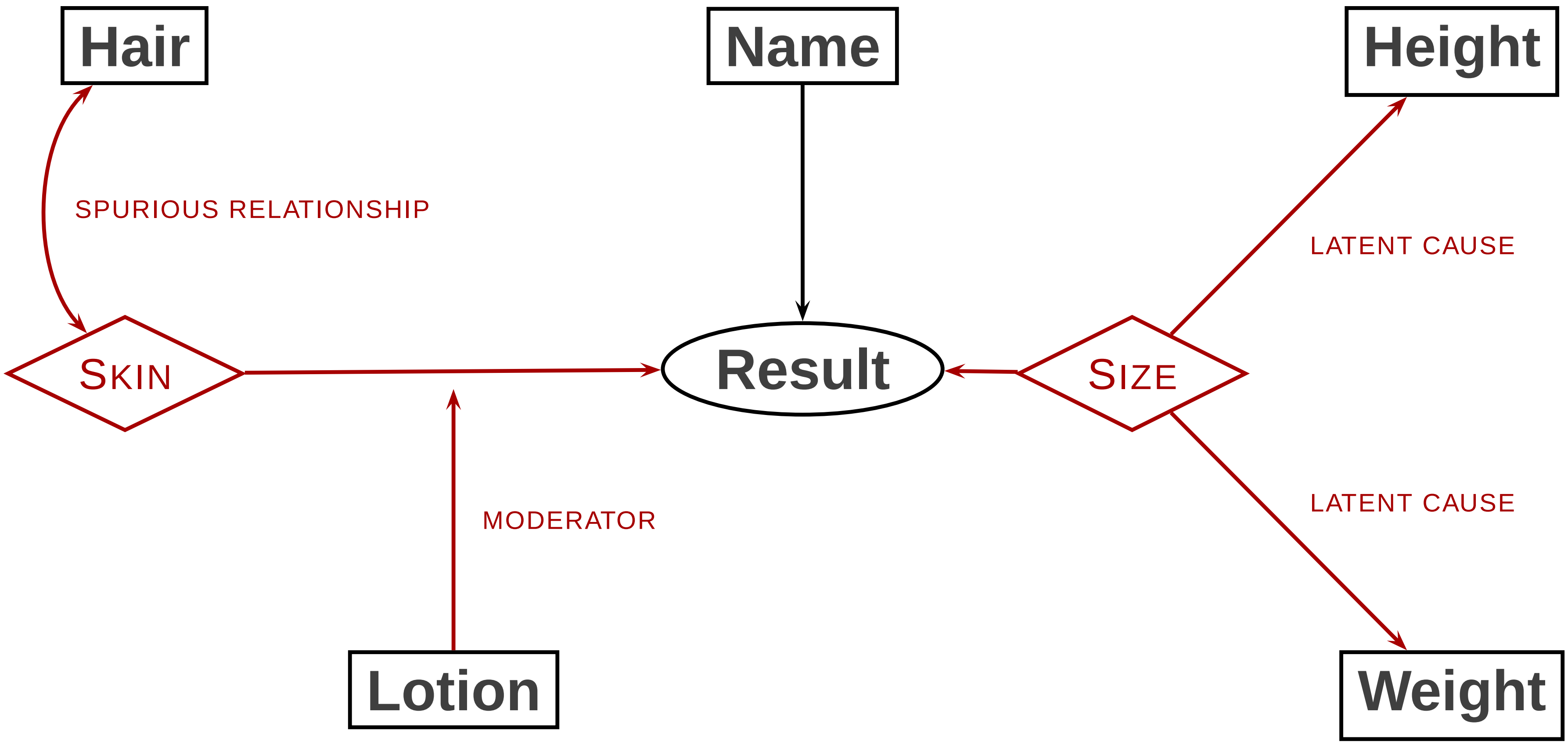


Voting



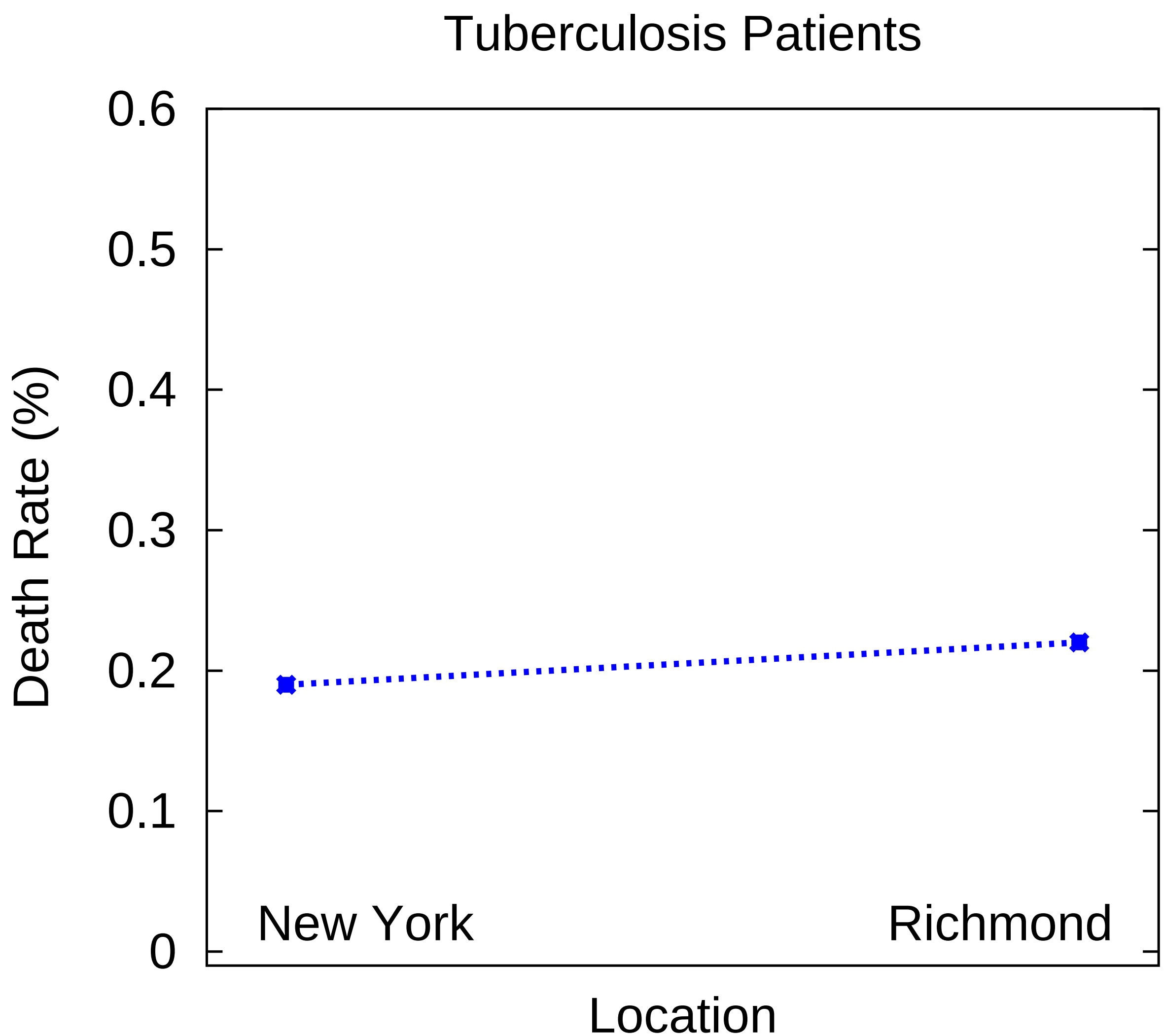
WE DECLARE THIS TO BE: A TRUE INTERACTION

Voting

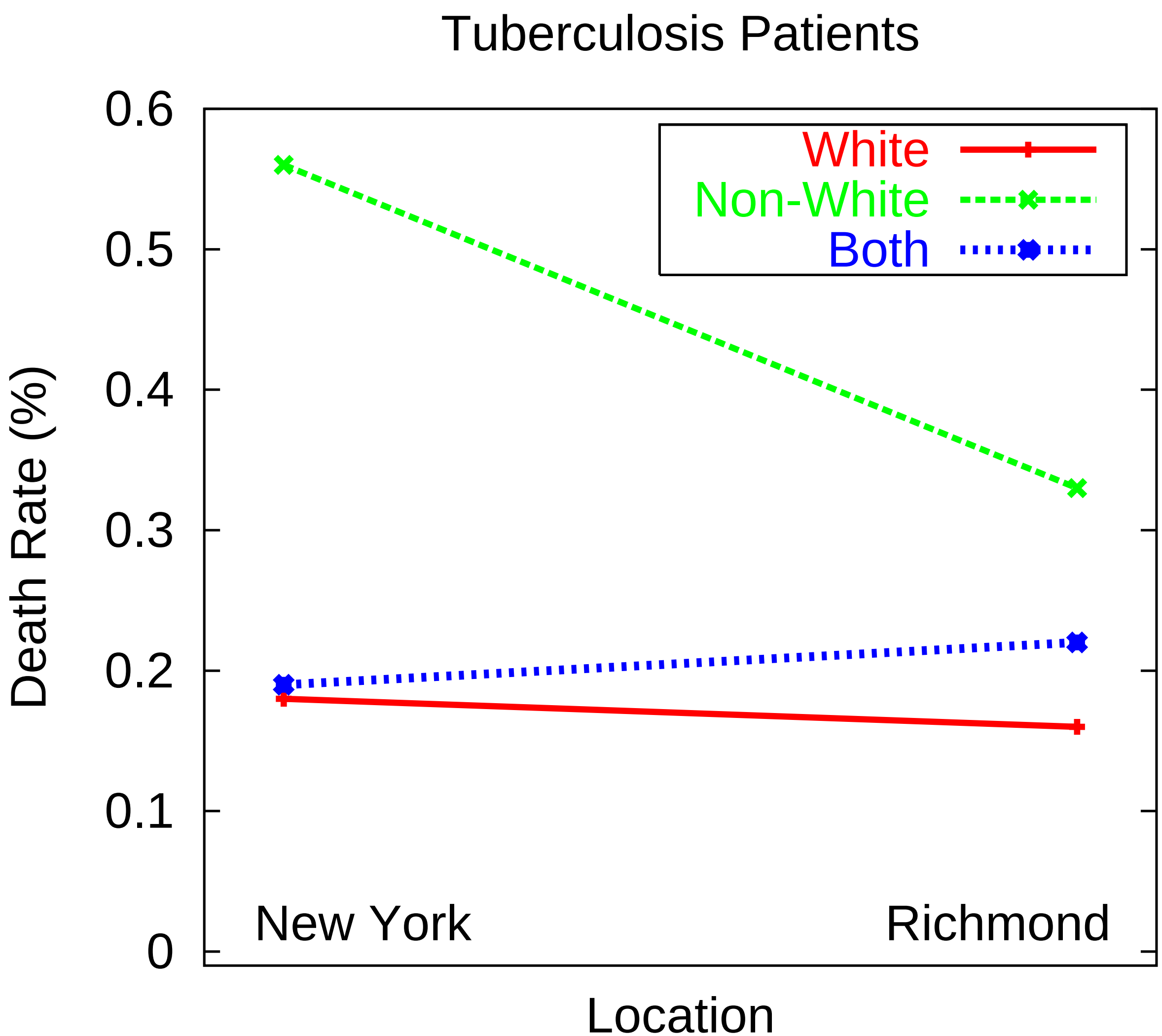


WE DECLARE THIS TO BE: | A TRUE INTERACTION | A FALSE INTERACTION

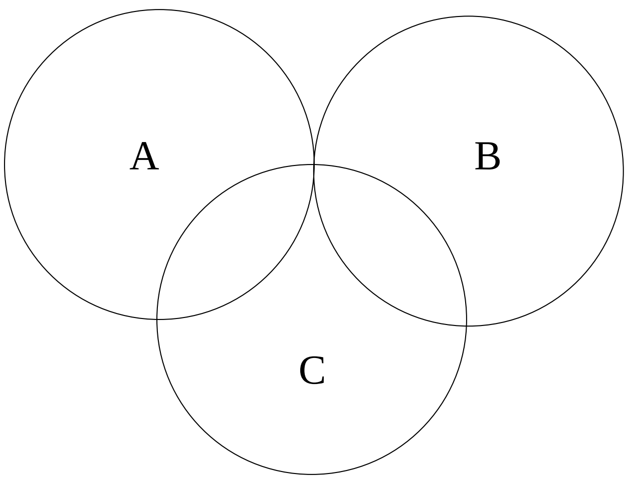
Simpson's Paradox



Simpson’s Paradox

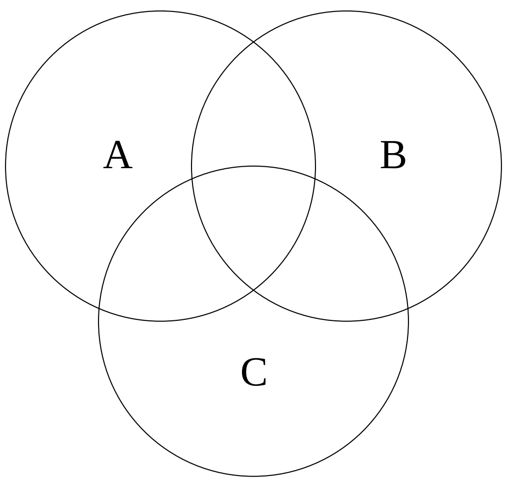


Information Gain



- An attribute is an information source. We want to estimate the amount of information shared between two sources.
- The amount learned about a label C from an attribute A is quantified by *information gain*: $\mathrm{Gain}_C(A) := H(A) + H(C) - H(AC)$.
- Interpretation: our ignorance about an unknown C reduces by $\mathrm{Gain}_C(A)$ given the knowledge of A.
- Sufficient, if all attributes are conditionally independent with respect to the label, when there are only 2-way interactions.

Interaction Gain

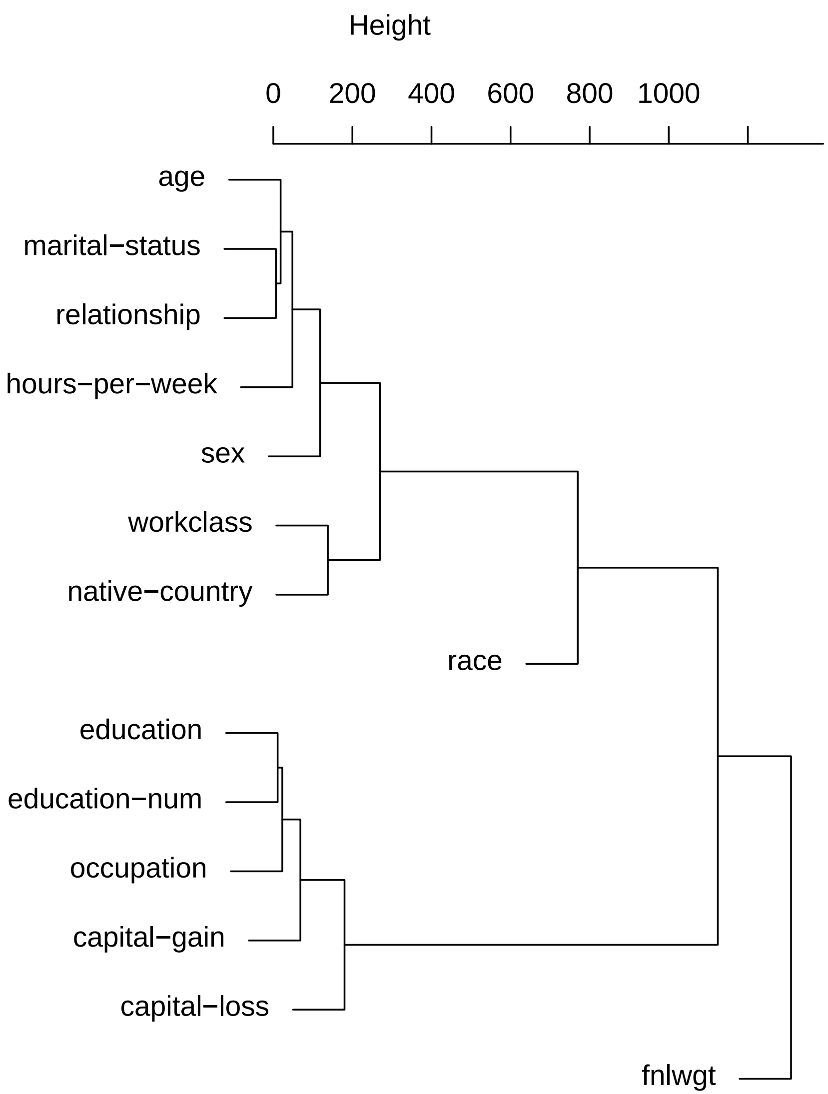


- How to estimate the amount of information shared among three attributes?
- Generalization of information gain for 3-way interactions is *interaction gain*:

$$\begin{aligned} IG_3(ABC) &:= H(AB) + H(AC) + H(BC) - H(A) \\ &\qquad -H(B) - H(C) - H(ABC) \\ &= \mathrm{Gain}_C(AB) - \mathrm{Gain}_C(A) - \mathrm{Gain}_C(B). \end{aligned}$$

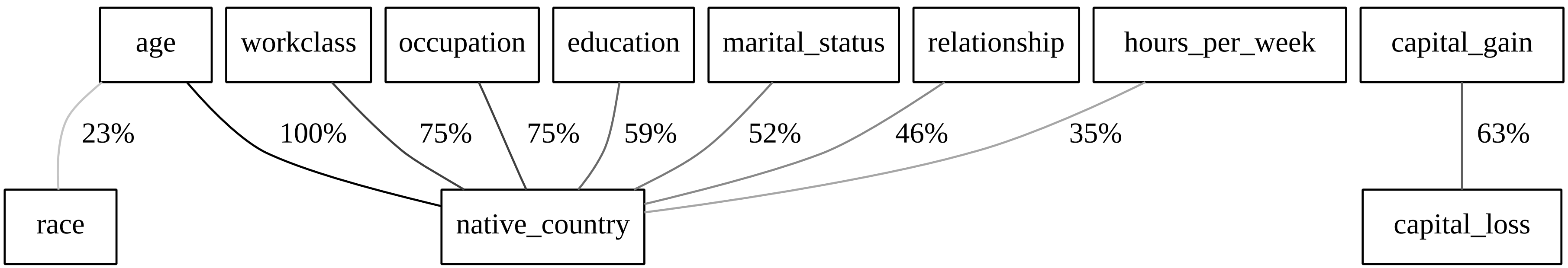
  - If IG negative: a false interaction.
  - If IG positive: a true interaction.
  - If IG zero: no 3-way interaction.

False Interaction Analysis



- The Census/Adult domain from UCI, 2-classes of individuals: rich, poor.
- Similarity between two attributes is proportional to negated 3-interaction gain between them and the label.
- Only false interactions were included into consideration.
- Agglomerative clustering was used to create the interaction dendrogram.

True Interaction Analysis



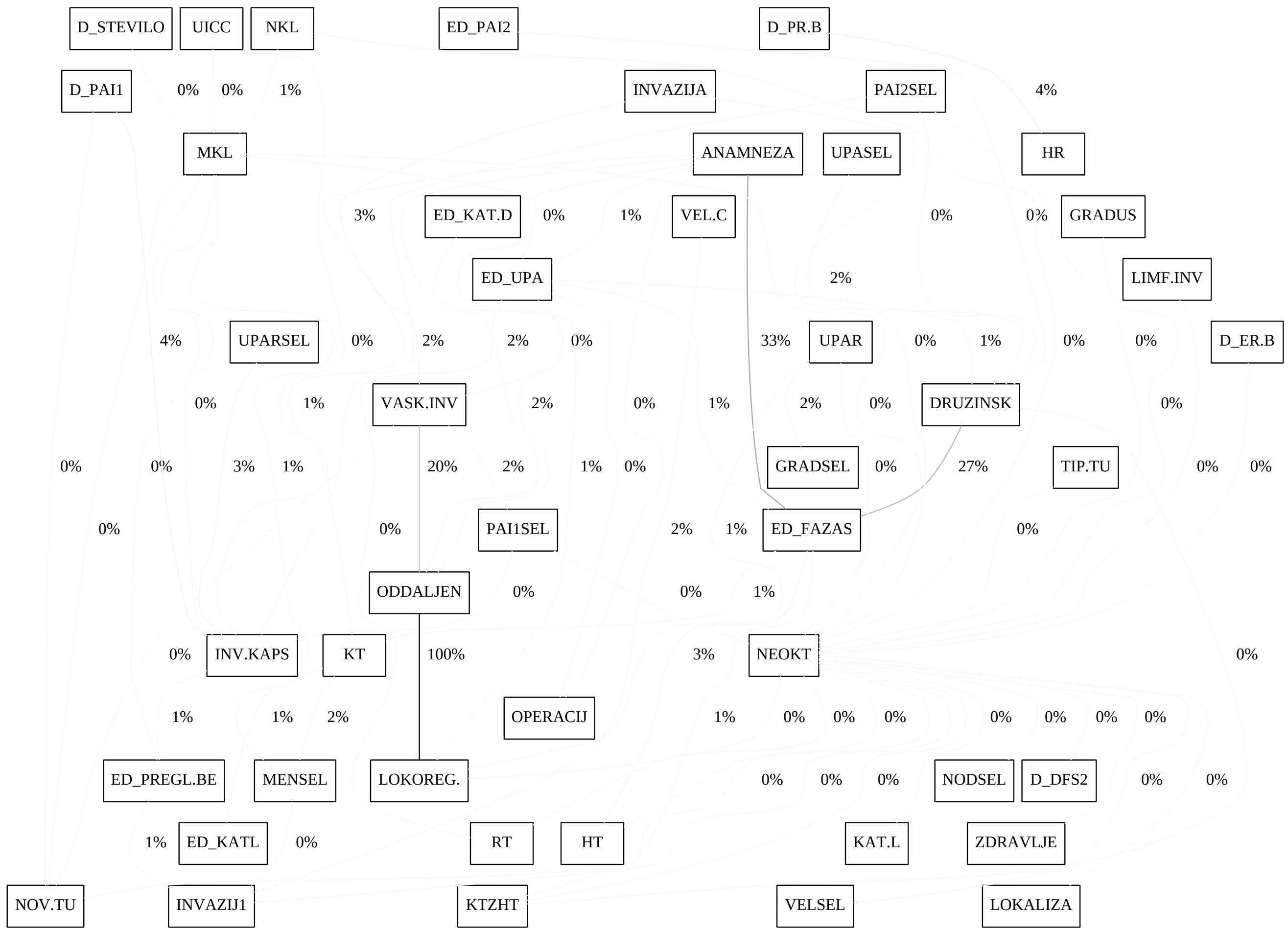
- A percentage on an interaction graph edge indicates the *strength* of a true interaction.
- Native country appears to be an important moderator, moderating a large number of 2-way interactions.
- True interactions are rarely transitive relations.
- True interactions are a forest of trees, not a single tree.

Interaction Significance (1)

When is an interaction significant?

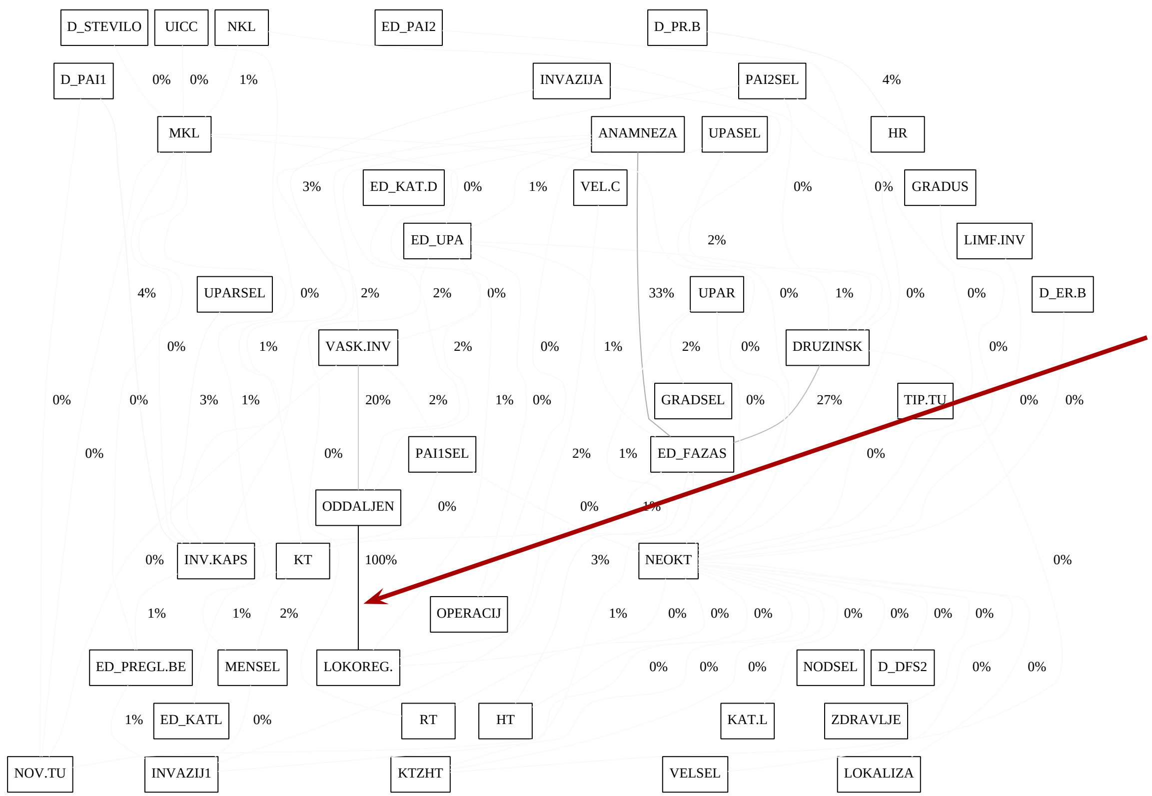
- Special statistics for conditional dependence and independence tests, e.g., Cochran-Mantel-Haenszel.
- Evaluate classifier performance on unseen data by comparing:
  - A classifier assuming independence between two attributes (voting).
  - A classifier exploiting dependence between two attributes via interaction resolution (segmentation).

Interaction Significance (2)



- There are generally few significant interactions.

Interaction Significance (2)




```
ODDALJEN >  0:  y
ODDALJEN <= 0:
:...LOKOREG. <= 0:  n
    LOKOREG. >  0:  y
```

A perfect classification tree for the 'breast' domain, induced by C4.5.

- But they matter: non-myopic feature selection, non-myopic split selection, non-myopic discretization, rules, trees, constructive induction.

Classification Performance

'adult'	Base	False	True
NBC	0.416	0.352	0.392
LR	1.562	0.418	1.564
SVM	—	—	—

'breast'	Base	False	True
NBC	0.262	0.187	0.171
LR	0.016	0.016	0.016
SVM	0.032	0.032	0.016

- A wrapper algorithm detects true or false interactions with interaction gain and uses minimal-error attribute reduction to resolve them. No feature selection and no parameter tuning was used.
- It improves results with logistic regression, SVM, and the naïve Bayesian classifier.
- There must be enough data!

Applications

- **Prediction:**
  - Resolving significant interactions helps improve classification performance.
  - Interactions limit or prevent myopia in discretization and feature selection.
  - Interactions justify constructive induction.
- **Analysis:**
  Interactions are interesting, especially if unexpected: interactions between treatments, symptoms, etc.

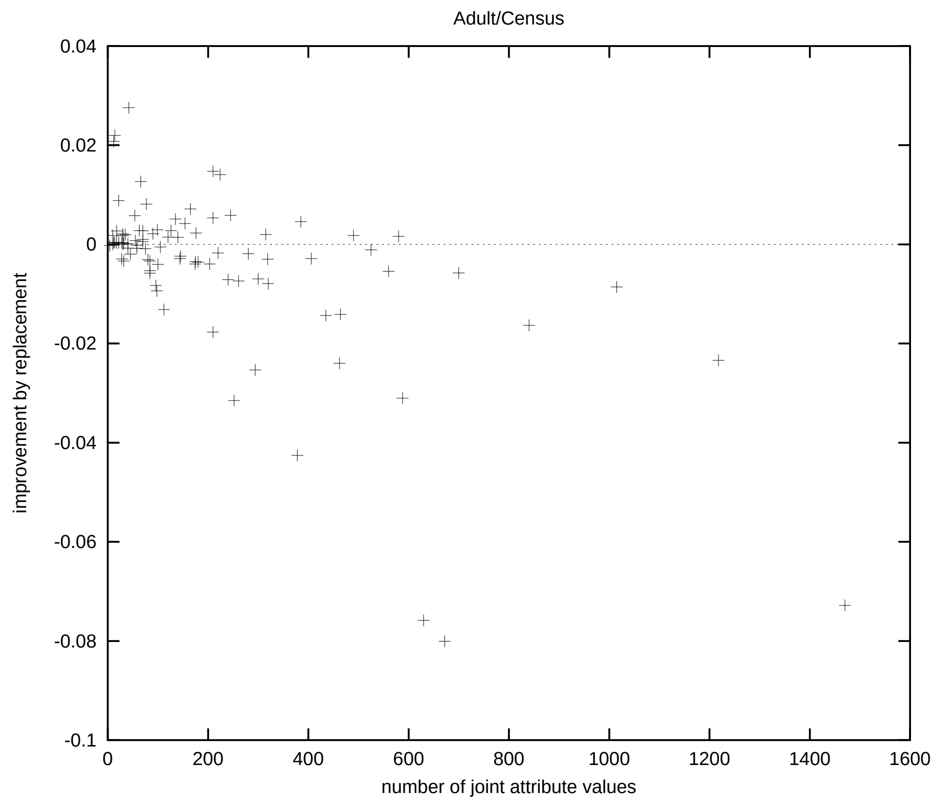
Summary of Contributions

- Two kinds of interactions: true and false interactions.
- Interaction gain is an interaction probe, able to detect and classify 3-way interactions.
- The pragmatic interaction significance test, based on comparison of classification performance on unseen data.
- True and false interaction analysis methodology, with interaction graphs and interaction dendrograms.
- Improving classification performance with interaction resolution.

Further Work

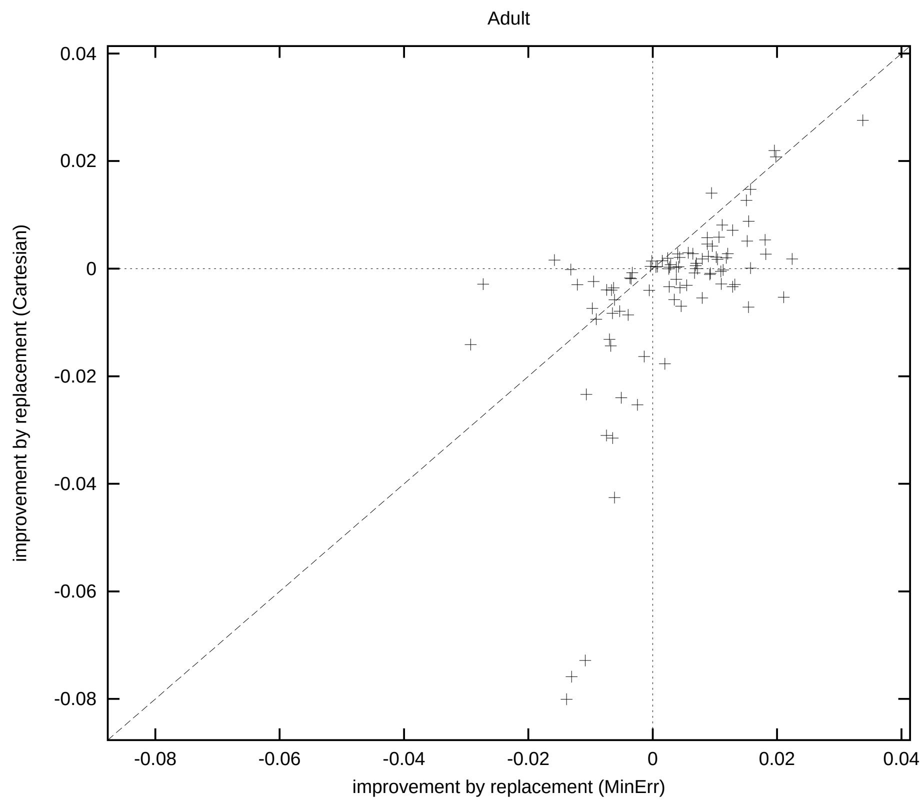
- A full-fledged tool for interaction analysis.
- Support for numerical and ordered attributes.
- Generalization to k-way interactions.
- Improved methods of resolution, especially of false interactions.
- Exploration of implications of interactions to discretization, split selection, etc.
- Applications.

Cardinality of Attributes



The greater the number of values in the constituent attributes, the lower the chances of the interaction between them to be significant.

Attribute Reduction



Minimal-error attribute reduction often yields better results than using the non-reduced Cartesian product of attributes.