University of Ljubljana, Faculty of Computer and Information Science

Inference and explanation of prediction models

Prof Dr Marko Robnik-Šikonja

Intelligent Systems, Edition 2025

Overview of topics

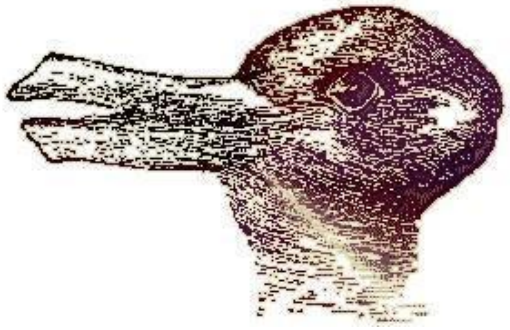
- Visualization and knowledge discovery.
- General methodology for explaining predictive models.
- Model level and instance level explanations, methods EXPLAIN and IME.

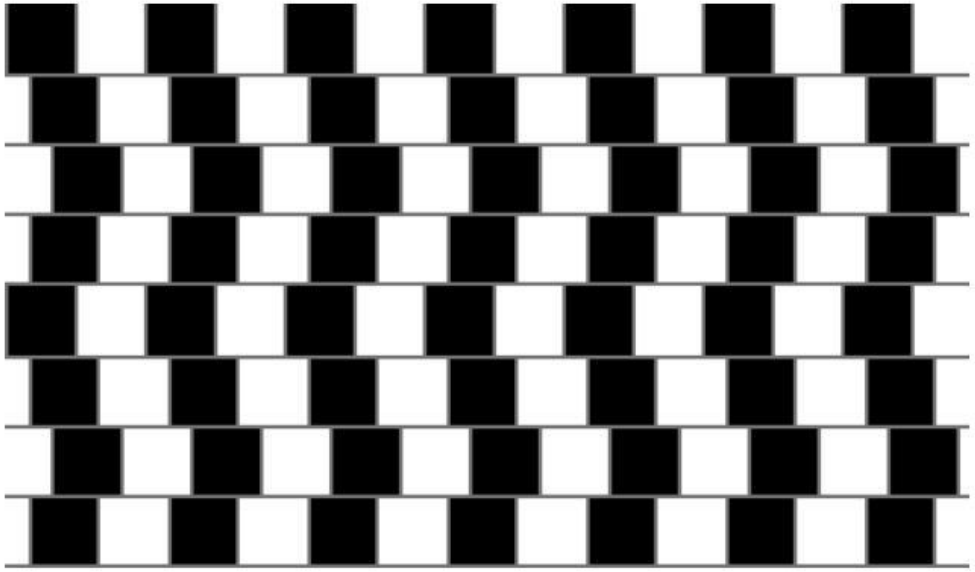
Visualization

- 1st rule of data mining: know your data.
- Therefore: visualizations, getting background data.
- Visualize: distributions of individual variables, their relations, etc.
- For high dimensional data sets one can use scaling, e.g. UMAP or t-SNE
- Clustering is useful in supervised tasks to get insight into the relation between predicted values Y and basic groups in the data. If unrelated, feature set might need amendments.

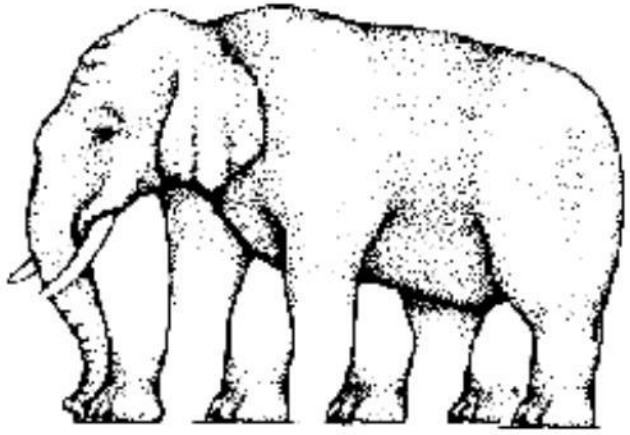
Visualizations

- Human visual perception has certain limitations:
  - we see what we want to see
  - we see what we see often
  - it is more difficult to notice unexpected patterns
- practice in detection of unknown
- use visualizations which expose “the unknown”





Are the horizontal lines parallel or do they slope?



How many legs does this elephant have?

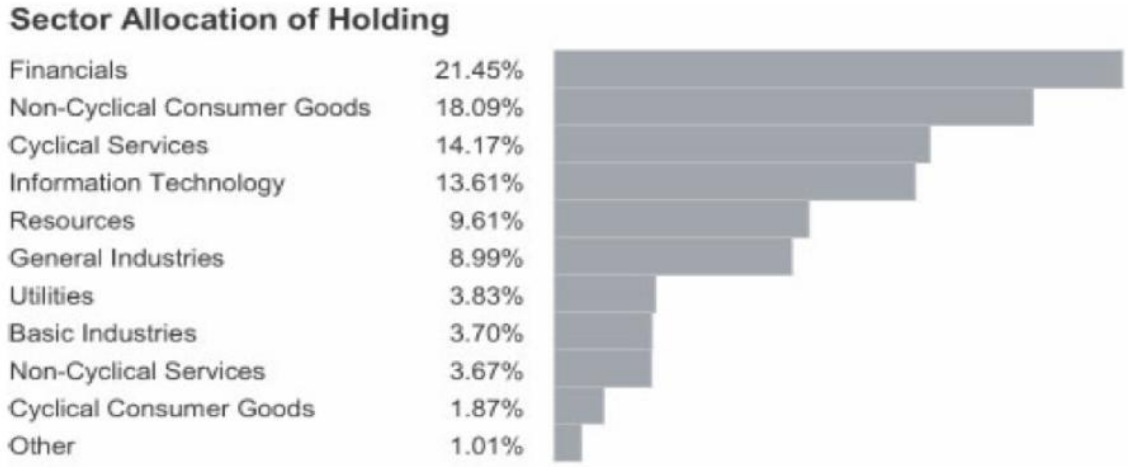
Human pattern recognition

- We see inexistent patterns because we WANT to see them (we feel lost without them).

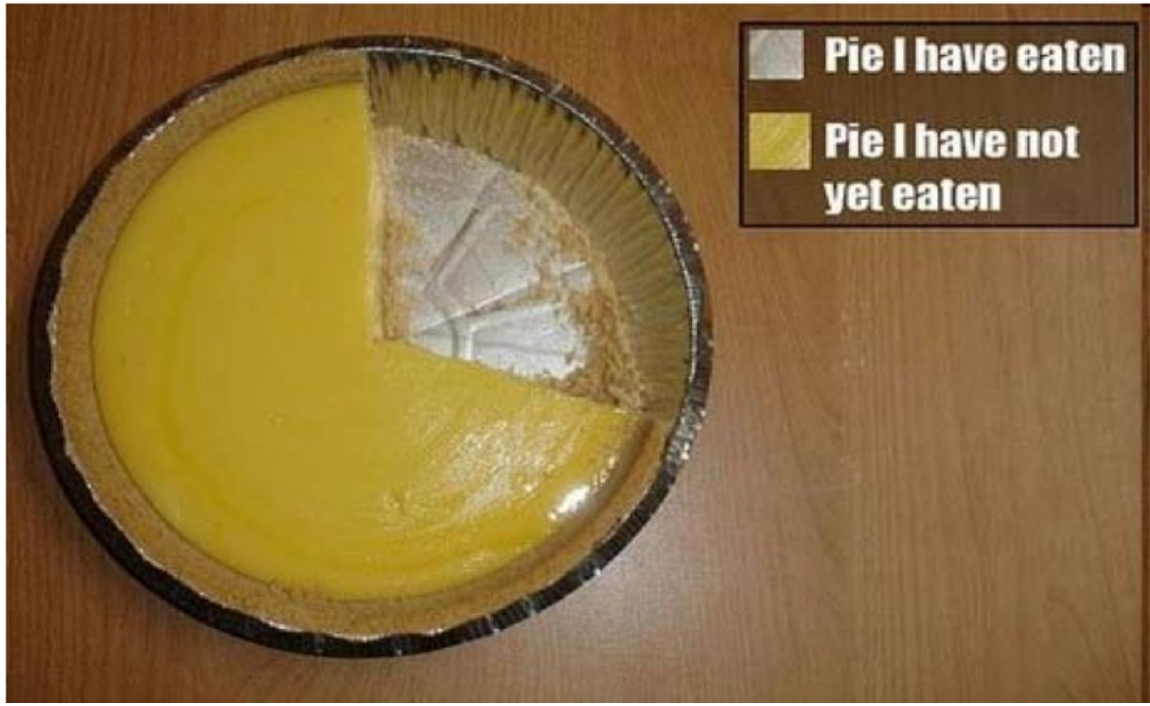
"The researchers found that when people were primed to feel out of control, they were more likely to see patterns where none exist." (**See a Pattern on Wall Street?,** *John Tierney*, *Science*)

Facts about simple visualizations

- Pie charts are a bad choice: hard to read, similar colors, slope, legend is too far away
- Bar chart is much better

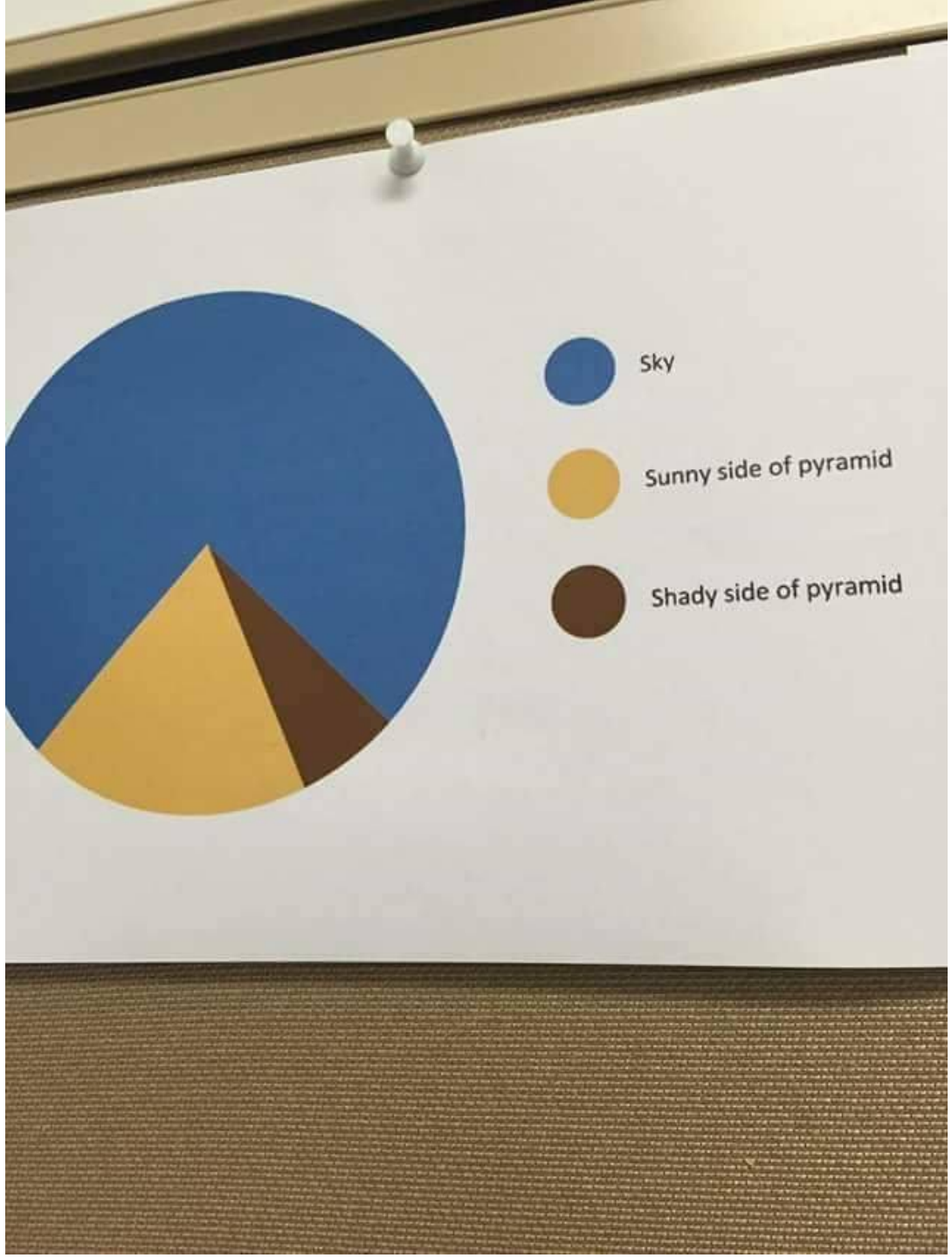


The best pie chart



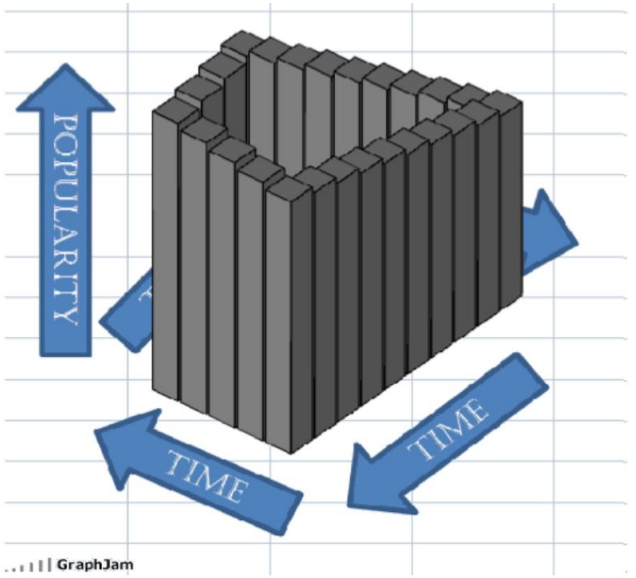
Pie charts jokes

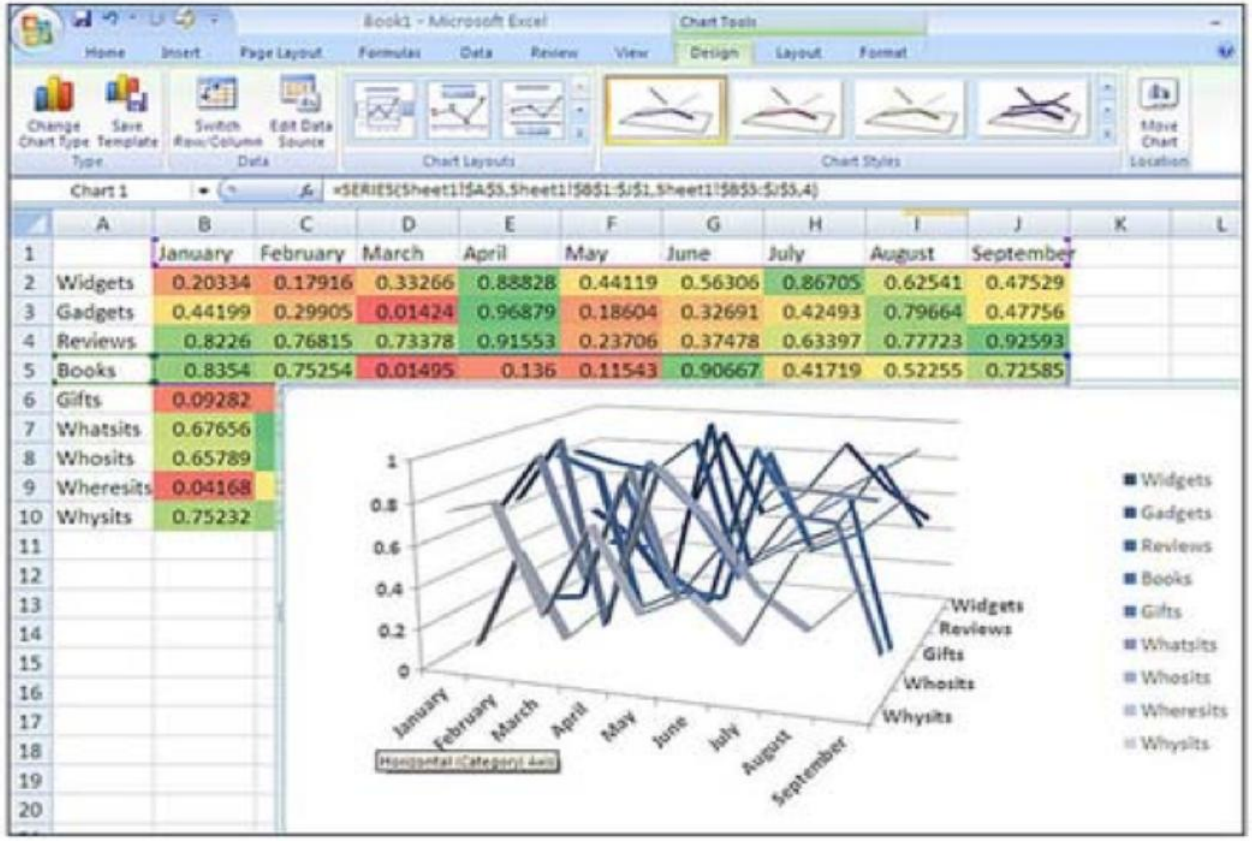
- notoriously bad

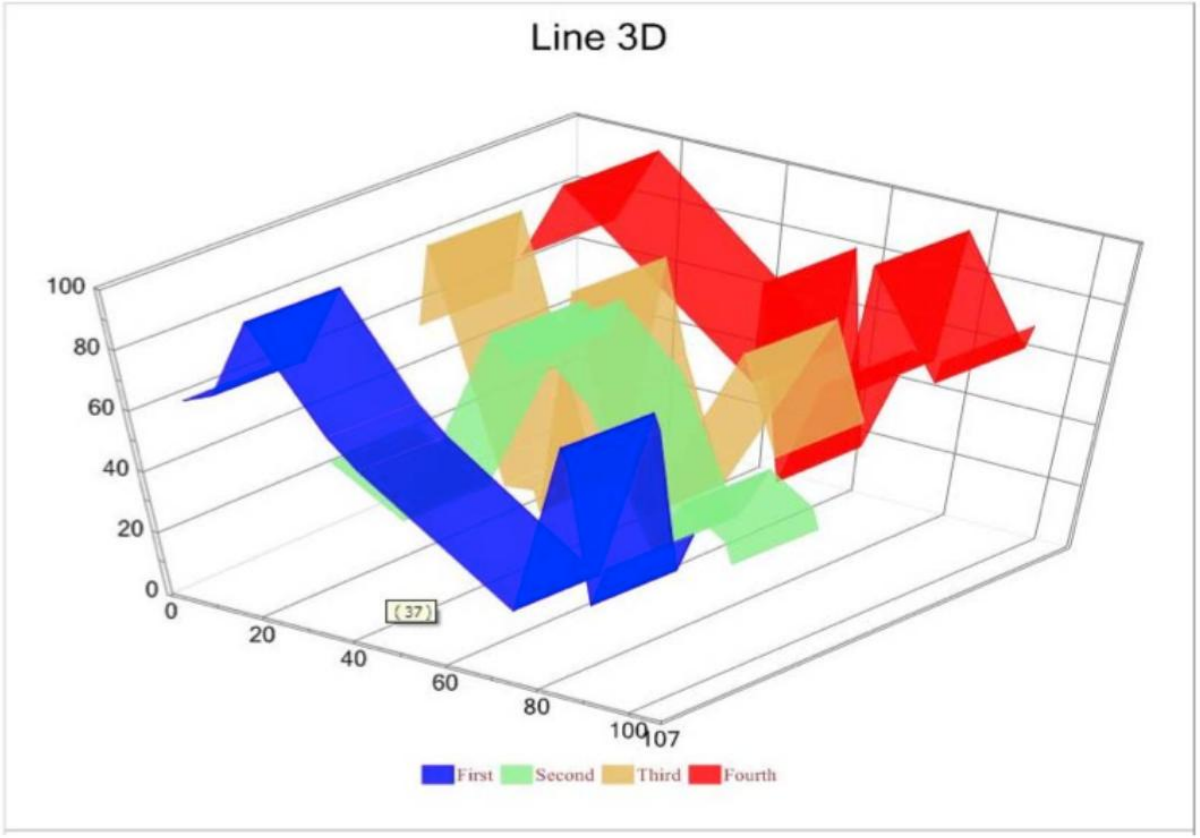


Facts about simple visualizations

- Bar charts, box plots can be OK
- 3D graphs are almost never OK for 2D info: spider plot, bowl of noodles
- Take care to be clear and do not manipulate
- A more detailed examples and recommendations https://github.com/cxli233/FriendsDontLetFriends

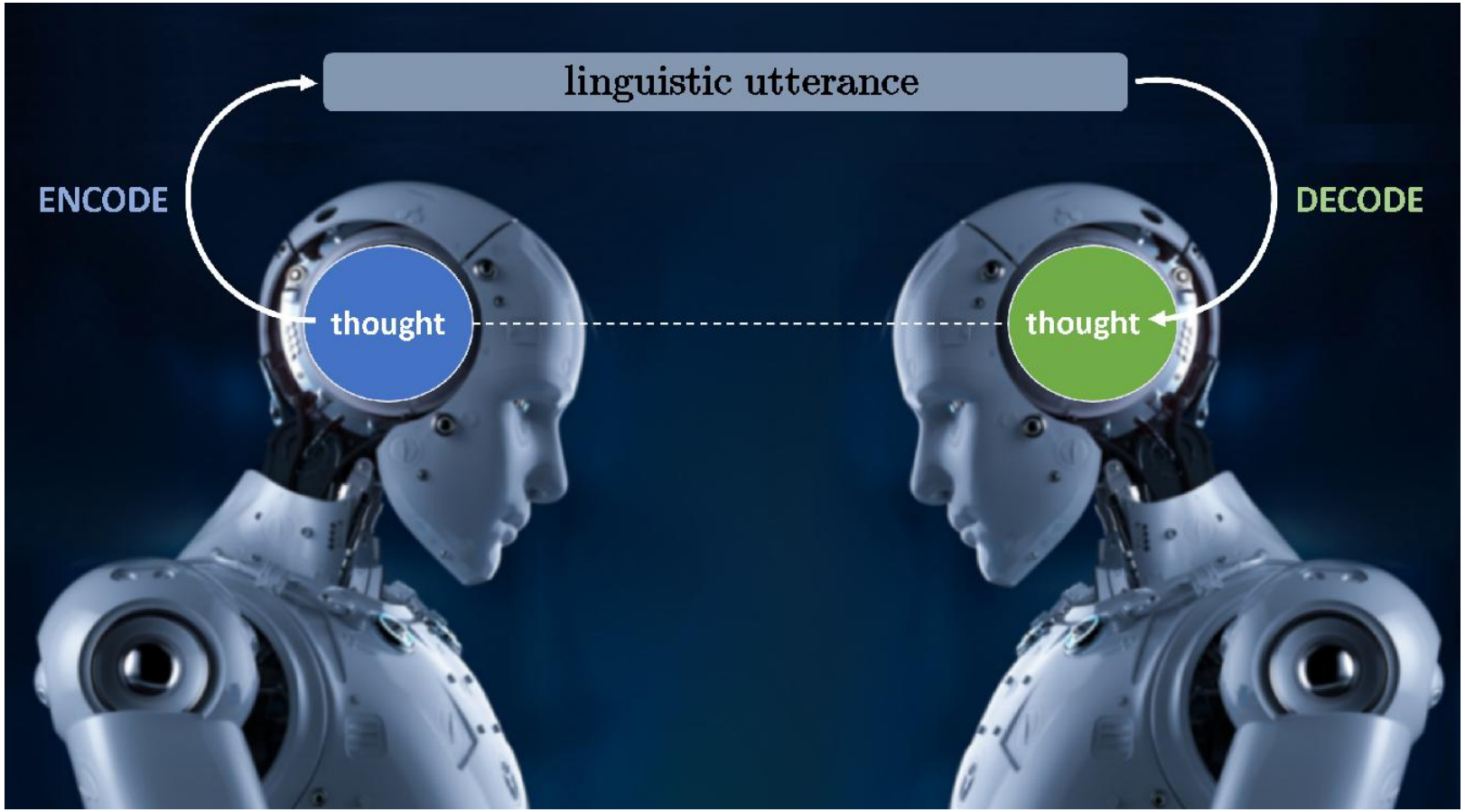






Understanding

Walid Saba, "Machine Learning Won't Solve Natural Language Understanding", The Gradient, 2021.

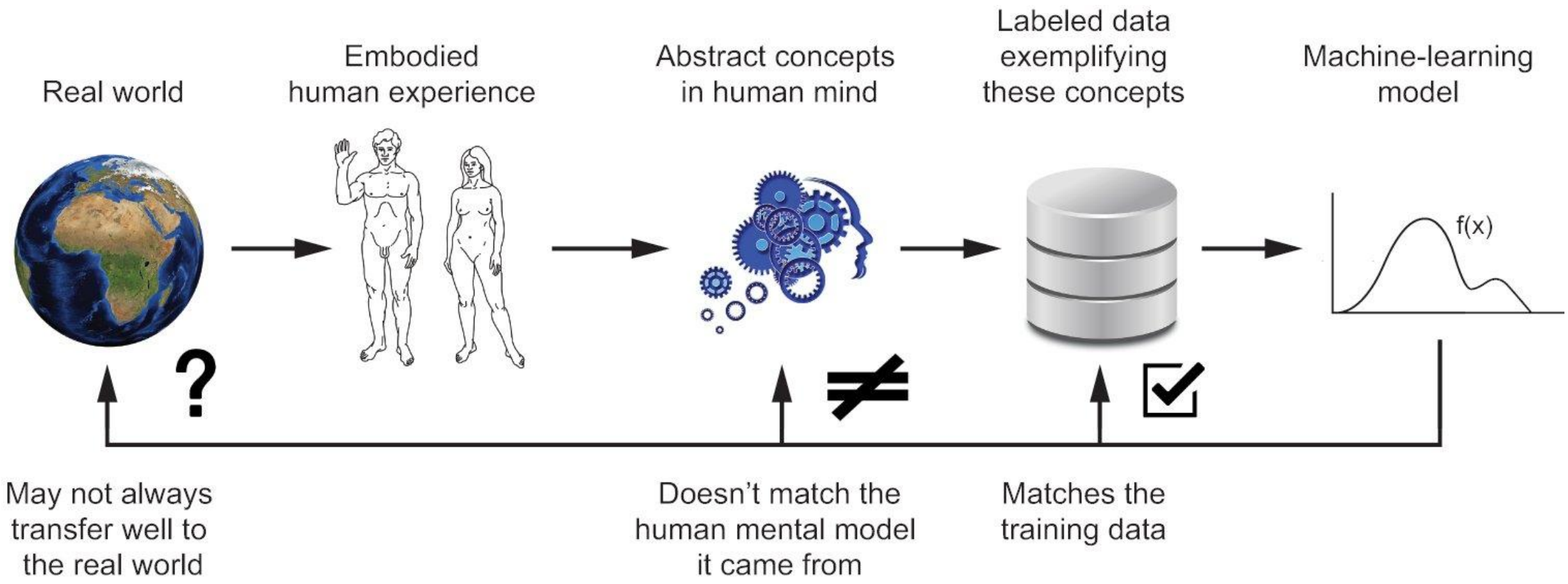


Xanadu, who is a living young human adult, and who was in graduate school, quit graduate school to join a software company that had a need for a new employee.

≈

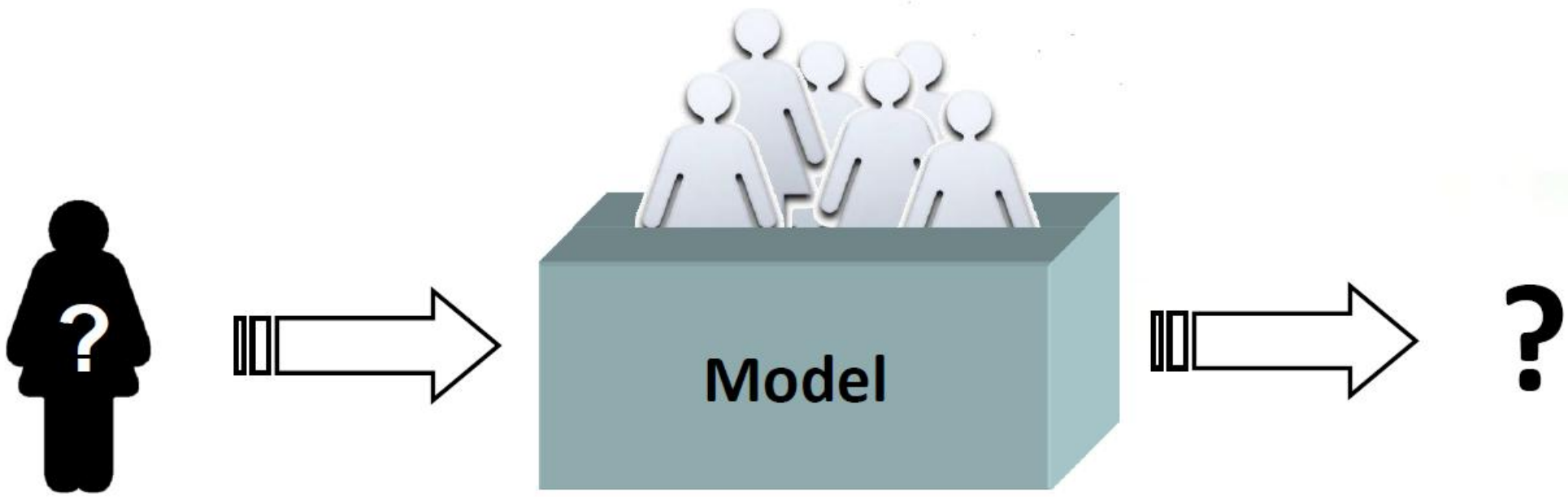
Xanadu quit graduate school to join a software company.

Understanding ML models is difficult



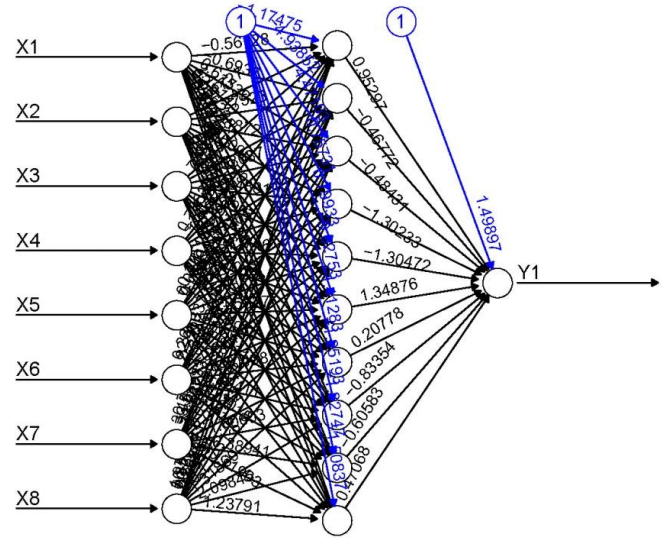
Predictive modeling scenario

We want to learn from past examples,with known outcomes.



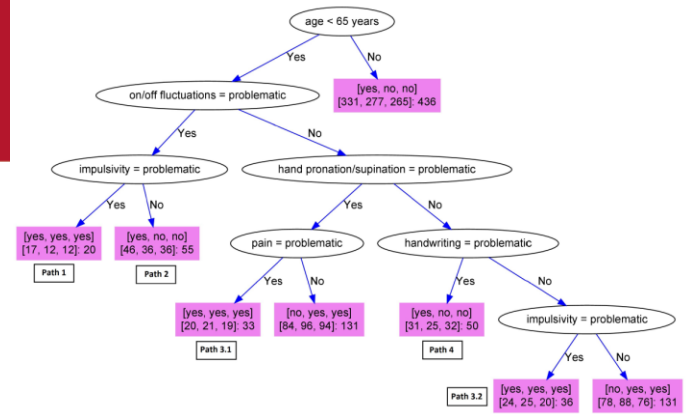
To predict the outcome for a new patient.

Explanation of predictions



- a number of successful prediction algorithms exist (SVM, boosting, random forests, neural networks), but to a user they are [redacted]
- many fields where users are very much concerned with the transparency of the models: medicine, law, consultancy, public services, etc.
- Some explanation methods are applicable to arbitrary predictors

Model comprehensibility



- decision support: model comprehensibility is important to gain users‘ trust
- knowledge acquisition
- some models are inherently interpretable and comprehensible
- decision and regression trees, classification and regression rules, linear and logistic regression $1/(1+\exp(-(b_0+b_1x_1+...b_px_p))$
- really?

Domain level explanation

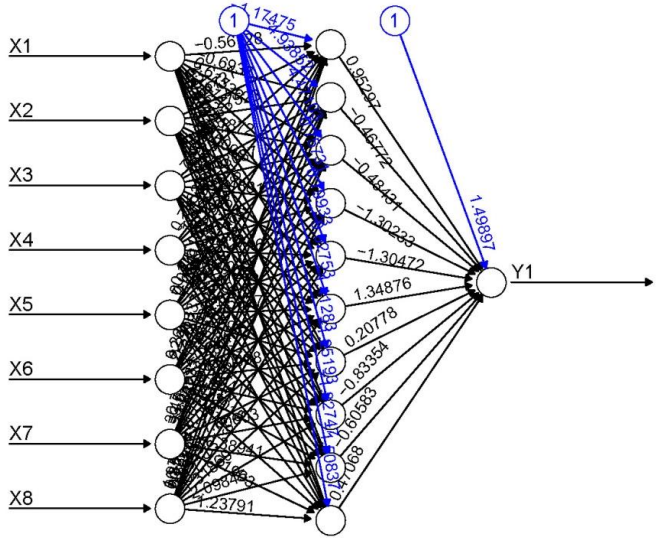
- trying to explain the "true causes and effects"
  - physical processes
  - stock exchange events

- usually unreachable except for artificial problems with known relations and generator function
- some aspects are covered with attribute evaluation, detection of redundancies, ...
- targeted indirectly through the models

Model-based explanations

All models are wrong,
but some are useful.

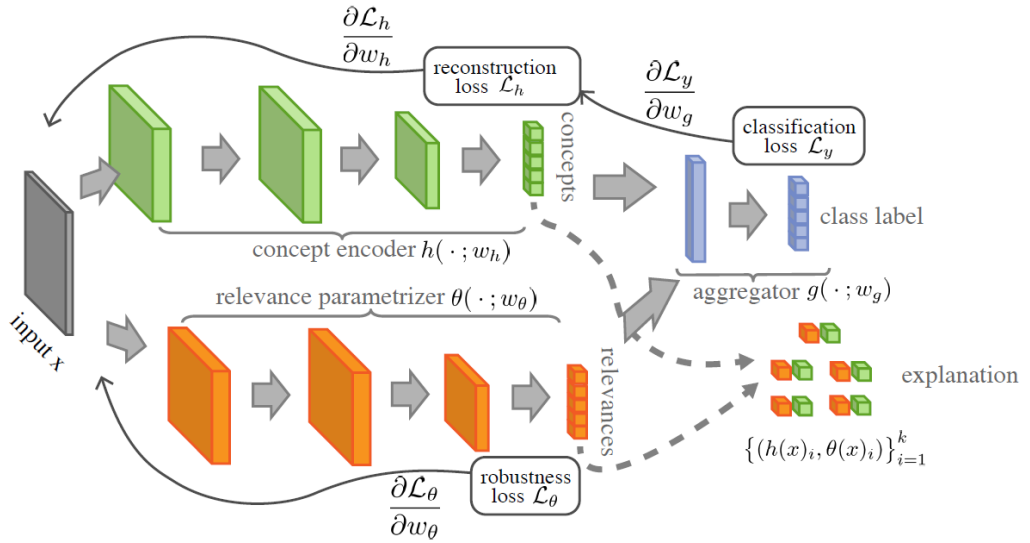
George Box, British statistician (1919 – 2013)



- make transparent the prediction process of a particular model
- the correctness of the explanation is independent of the correctness of the prediction but
- better models (with higher prediction accuracy) enable in principle better explanation at the domain level
- explanation methods are interested only in the explanation at the model level and leave to the developer of the model the responsibility for its prediction accuracy

Two flavours of explanation techniques

- model specific
  - especially used for deep neural networks

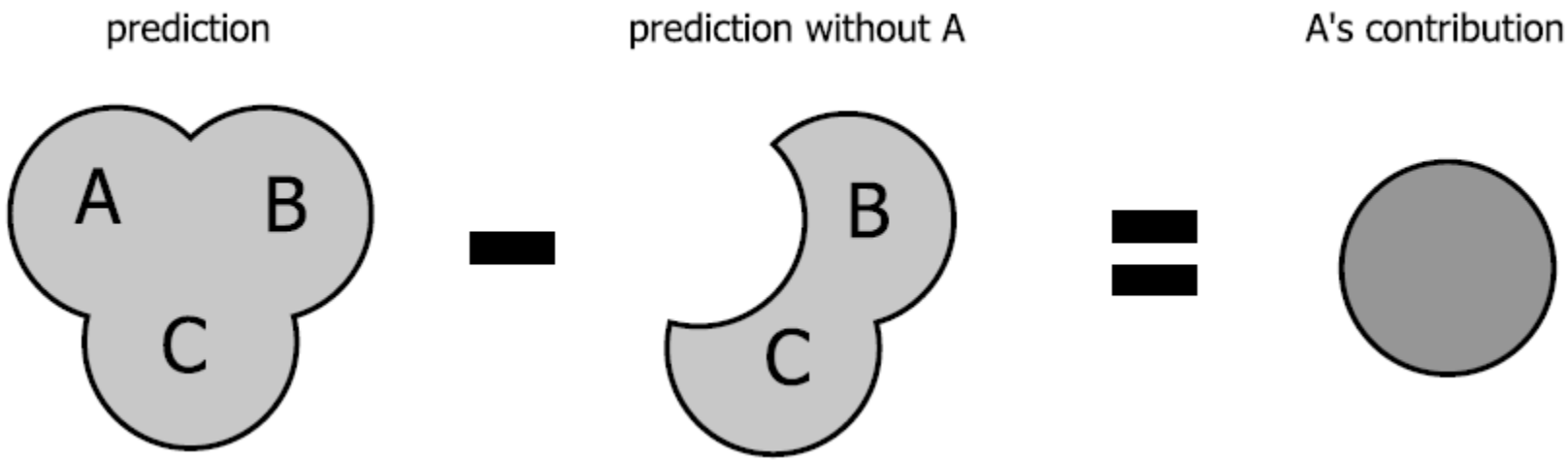


Melis, D.A. and Jaakkola, T., 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems* (pp. 7786-7795).

- model agnostic
  - can be used for any predictor,
  - based on perturbation of the inputs

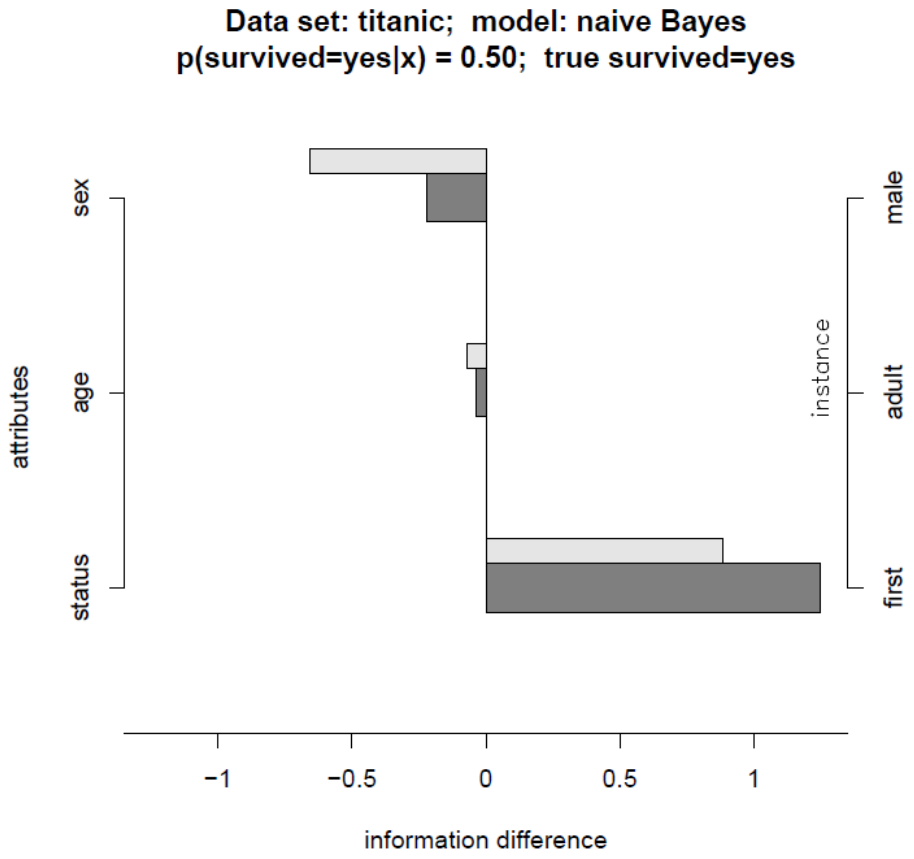
Idea of perturbtion-based explanations

- importance of a feature or a group of features in a specific model can be estimated by simulating lack of knowledge about the values of the feature(s)



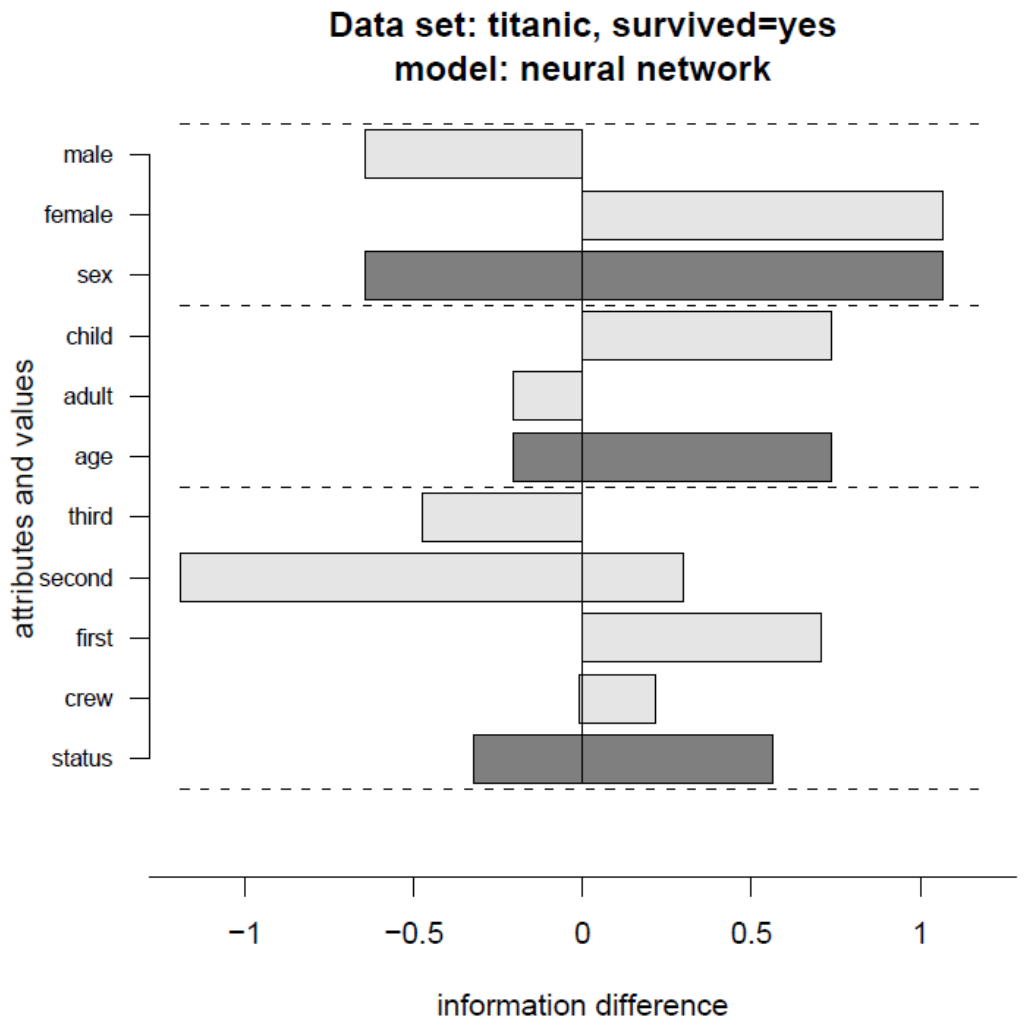
Instance-level explanation

- explain predictions for each instance separately
  - this is what practitioners applying models are interested in
  - presentation format: impact of each feature on the prediction value
- model-based



Model-level explanation

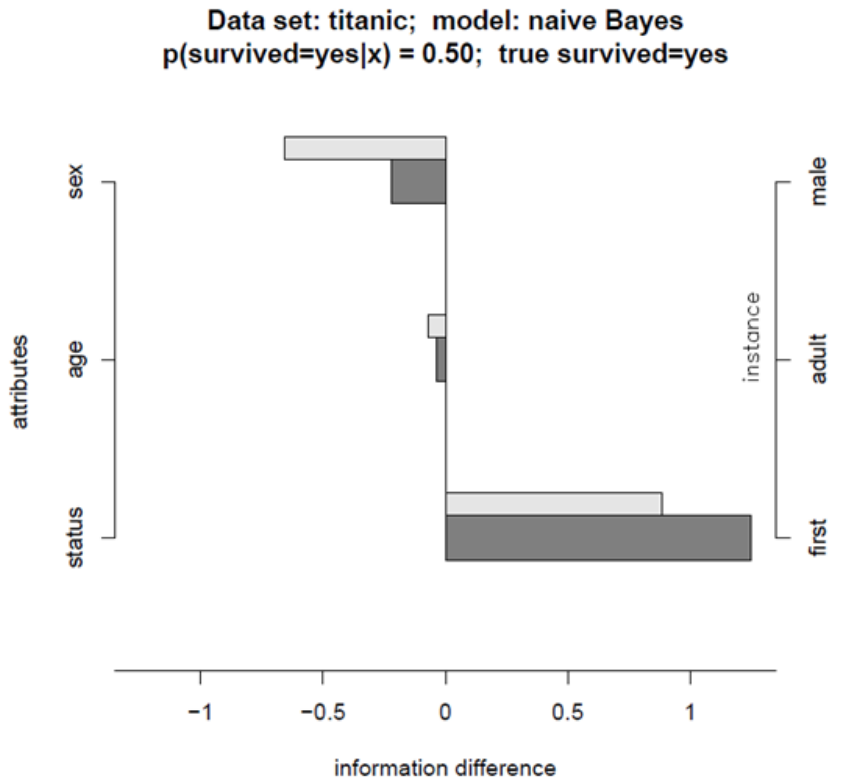
- the overall picture of a problem the model conveys
  - this is what knowledge extractors are interested in
  - presentation format: overall importance of each feature,
    but also rules, trees
- model-based



The method EXPLAIN

- "hide" one attribute at a time
- estimate contribution of attribute from

$$p(y_k|x) - p_{S\setminus\{i\}}(y_k|x)$$



Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering,*, *20*(5), 589-600.

Explaining EXPLAIN

- assume an instance (**x**, y); components of **x** are values of attributes A_i
- for a new instance **x**, we want to know what role each attribute's value play in the prediction model f, i.e. to what extend it contributed to the classification f(**x**)
- for that purpose
  - we compute f(**x** \ A_i), the model's prediction for **x** without the knowledge of the event $A_i = a_k$ (marginal prediction)
  - we comparing f(**x**) and f(**x** \ A_i) to assess importance of $A_i = a_k$
  - the larger the the difference the more important the role of $A_i=a_k$ in the model
- f(**x**) and f(**x** \ A_i) are source of explanations

Evaluation of prediction differences

- how to evaluate $f(\mathbf{x}) - f(\mathbf{x} \setminus A_i)$
- in classification, we take $f(\mathbf{x})$ in the form of probability

1. **difference of probabilities**

   $probDiff_i(y|\mathbf{x}) = p(y|\mathbf{x}) - p(y|\mathbf{x}\setminus A_i)$

2. **information gain** (Shannon, 1948)

   $infGain_i(y|\mathbf{x}) = \log_2 p(y|\mathbf{x}) - \log_2 p(y|\mathbf{x}\setminus A_i)$

3. **weight of evidence** also log odds ratio (Good, 1950)

   $odds(z) = p(z) / (1 - p(z))$

   $WE_i(y|\mathbf{x}) = \log_2 odds(y|\mathbf{x}) - \log_2 odds(y|\mathbf{x}\setminus A_i)$

Implementation

- $p(y|\mathbf{x})$: classify $\mathbf{x}$ with the model
- $p(y|\mathbf{x}\setminus A_i)$ – simmulate lack of knowledge of A_i in the model
  - replace with special NA value: good for some, mostly bad, left to the mercy of model's internal mechanism
  - average prediction across perturbations of A_i

    $p(y|\mathbf{x}\setminus A_i) = \sum_a p(A_i=a_s)\, p(y|\mathbf{x} \leftarrow A_i = a_s)$

    - use discretization for numeric attributes
    - use Laplace correction for probability estimation
  - we could build a separate model for each $p(y|\mathbf{x}\setminus A_i)$

Weaknes of EXPLAIN

- "hide" one attribute at a time
- estimate contribution of attribute from

$$p(y_k|x) - p_{S\setminus\{i\}}(y_k|x)$$

- weakness: if there are redundant ways to express concept, credit is not assigned
- example:

  $C = A_1 \vee A_2A_3$

  explanation for instance ($A_1=A_2=A_3=1$)

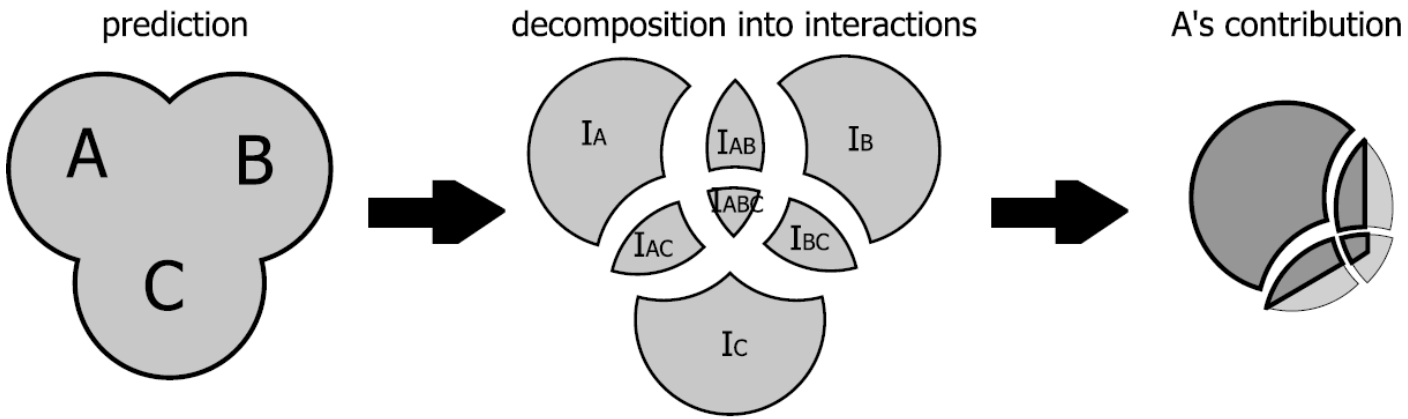
University of Ljubljana
Faculty of Computer and Information Science

The method IME

- (Interactions-based Method for Explanation)
- “hide” any subset of attributes at a time (2^a subsets!)
- the source of explanations is the difference in prediction using a subset of features Q and an empty set of features {}

$$\Delta_Q = h(x_Q) - h(x_{\{\}})$$

- the feature gets some credit for standalone contributions and for contributions in interactions



IME: sum over all subsets

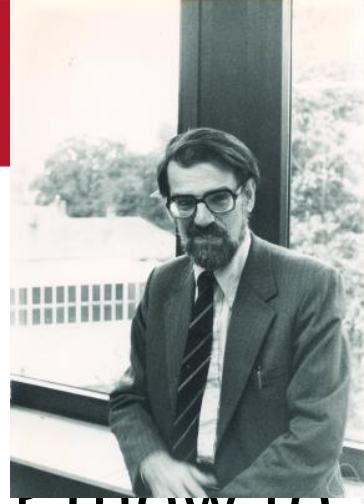
- the contributions are

$$\pi_i = \sum_{Q \subseteq \{1,2,\ldots,a\} - \{i\}} \frac{1}{a \binom{a-1}{a-|Q|-1}} (\Delta_{Q \cup \{i\}} - \Delta_Q)$$

Štrumbelj, E., Kononenko, I. & Robnik-Šikonja, M., Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, Oct. 2009, 68(10):886-904

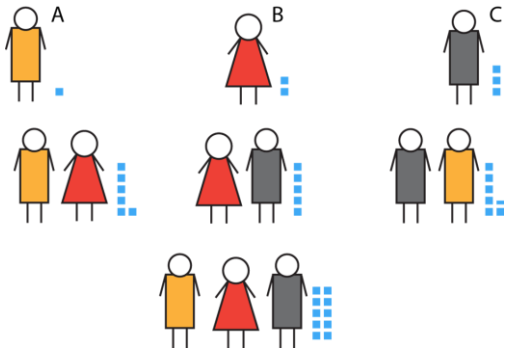
University of Ljubljana
Faculty of Computer and Information Science

Game theory analogy



- coalitional game of *a* players (attributes)
- players form coalitions (i.e. interactions)
- how to distribute the payout to the members of a coalition? (how to assign the credit for prediction)
- The Shapley value is the unique payoff vector that is
  - efficient (exactly splits payoff value),
  - symmetric (equal payments to equivalent players)
  - additive (overall credit is a sum of participating in coalitions), and
  - assigns zero payoffs to dummy players (no contribution to any coalition).

Shapley value



$$Sh_i(v) = \sum_{S \subseteq N \setminus \{i\}, s=|S|} \frac{(n-s-1)!s!}{n!}(v(S \cup \{i\}) - v(S)), \quad i = 1, \ldots, n.$$

$$\pi_i = \sum_{Q \subseteq \{1,2,\ldots,a\} - \{i\}} \frac{1}{a\binom{a-1}{a-|Q|-1}}(\Delta_{Q \cup \{i\}} - \Delta_Q)$$

- Shapley value can be efficiently approximated

University of Ljubljana
Faculty of Computer and Information Science

Solution for IME: sampling

- Shapley value can expressed in an alternative formulation
- $\pi(a)$is the set of all ordered permutations of *a*
- Pre^i(O) is the set of players which are predecessors of player *i* in the order O ∈ $\pi(a)$

$$\varphi_i(k,x) = \frac{1}{a!} \sum_{\mathcal{O}\in\pi(a)} \left(\Delta(Pre^i(\mathcal{O}) \cup \{i\})(k,x) - \Delta(Pre^i(\mathcal{O}))(k,x)\right) =$$

$$= \frac{1}{a!} \sum_{\mathcal{O}\in\pi(a)} \left(p_{Pre^i(\mathcal{O})\cup\{i\}}(y_k|x) - p_{Pre^i(\mathcal{O})}(y_k|x)\right),$$

- smart sampling over subsets of attributes
- computationally feasible approach

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11, 1-18

IME algorithm

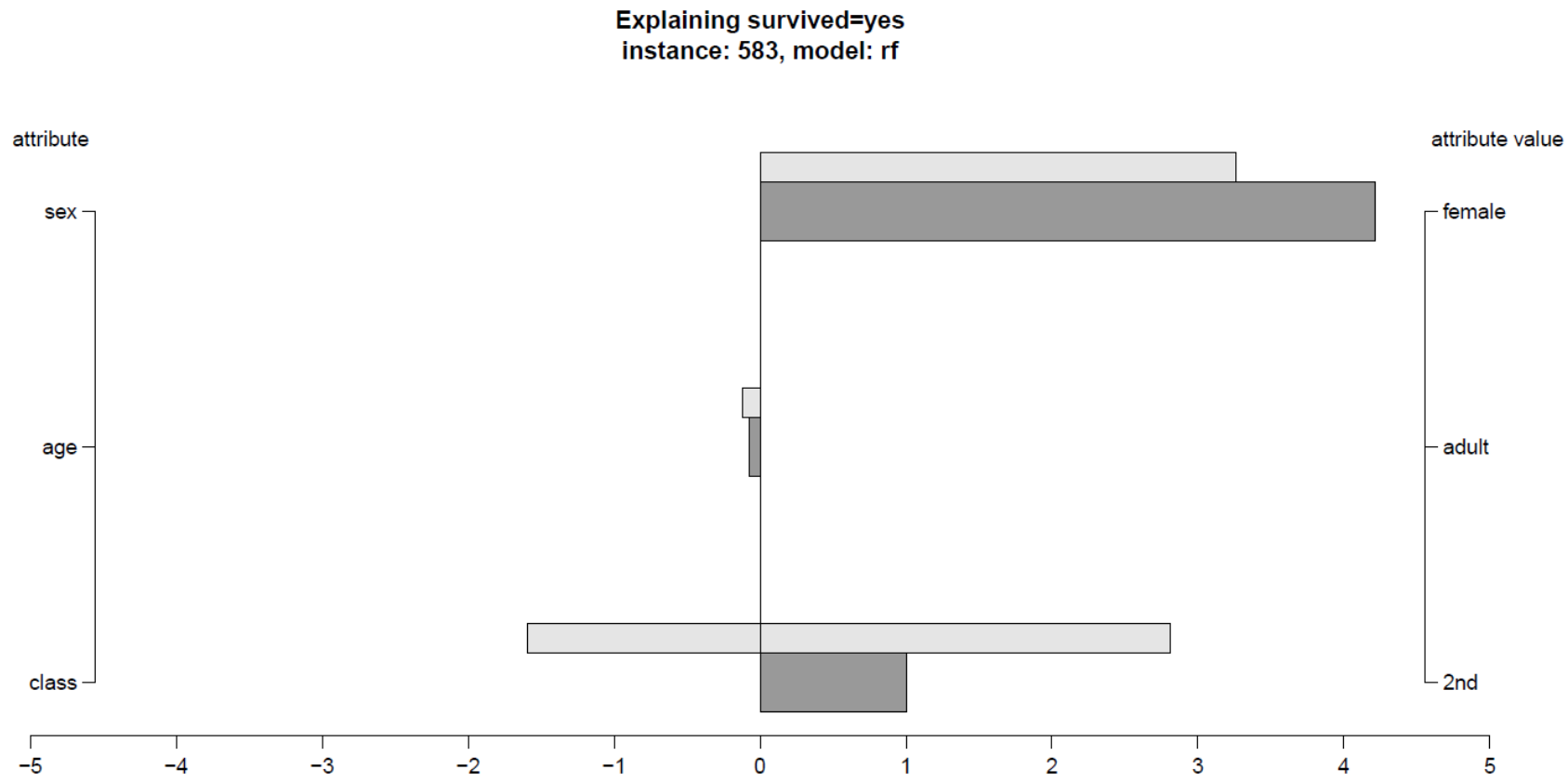
Algorithm 1 Approximating the contribution of the i-th feature's value, φ_i, for instance $x \in \mathcal{A}$.

determine m, the desired number of samples
$\varphi_i \leftarrow 0$
for $j = 1$ to m **do**
  choose a random permutation of features $O \in \pi(N)$
  choose a random instance $y \in \mathcal{A}$
  $v_1 \leftarrow f(\tau(x, y, Pre^i(O) \cup \{i\}))$
  $v_2 \leftarrow f(\tau(x, y, Pre^i(O)))$
  $\varphi_i \leftarrow \varphi_i + (v_1 - v_2)$
end for
$\varphi_i \leftarrow \frac{\varphi_i}{m}$

- by measuring the variance of contributions, we can determine the necessary number of samples for each attribute

Visualization of explanations

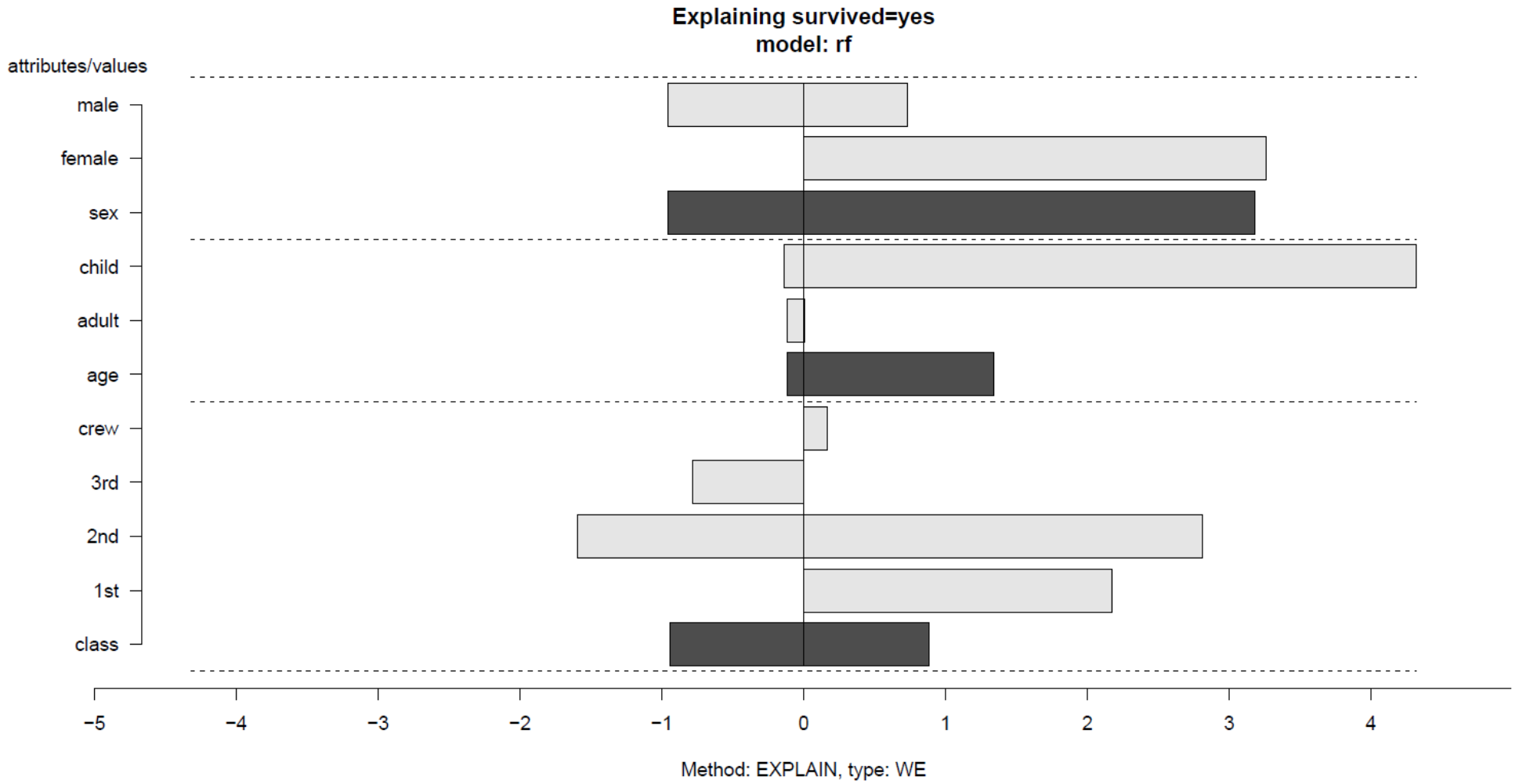
- instance-level explanation on Titanic data set



Robnik-Šikonja, M. (2015), ExplainPrediction: Explanation of Predictions for Classification and Regression.
R package version 1.3.0. http://cran.r-project.org/package=ExplainPrediction

Visualization of explanations

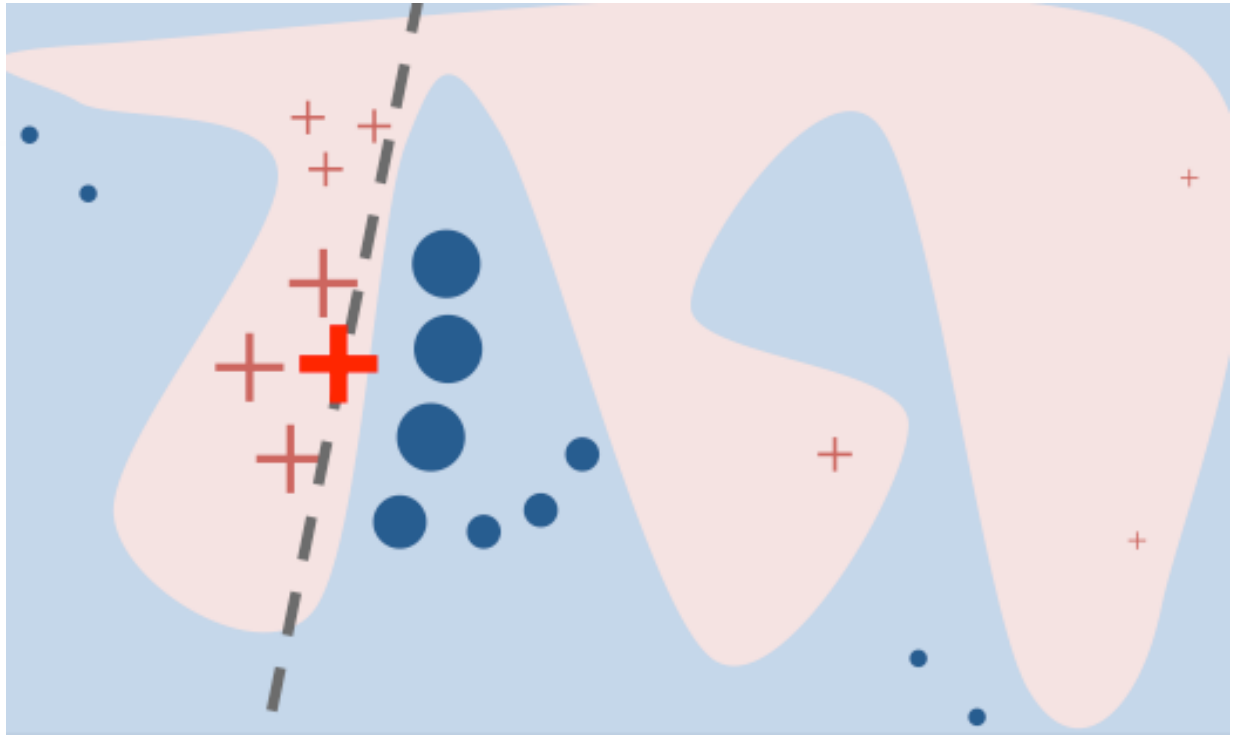
- model-level explanation on Titanic data set



University of Ljubljana
Faculty of Computer and Information Science

LIME explanation method

- Local Interpretable Model-agnostic Explanations)
- perturbations in the locality of an explained instance

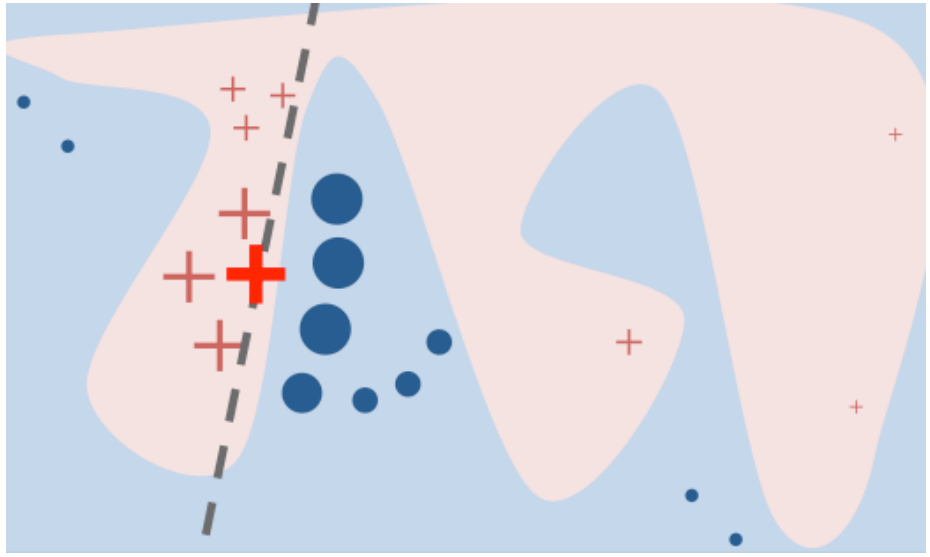


Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144.

LIME explanation method

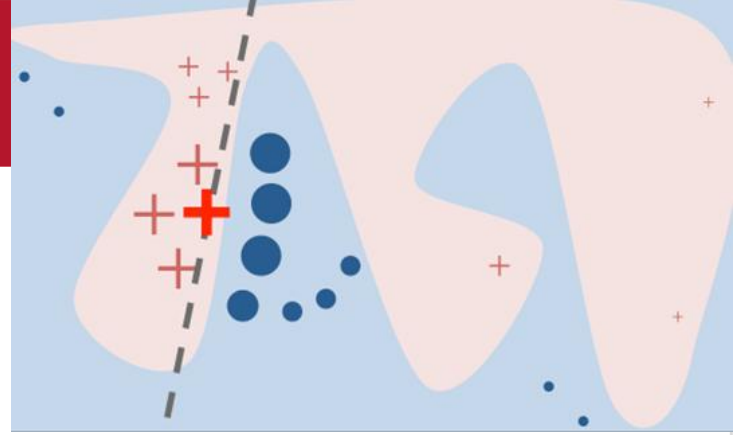
- optimize a trade-off between local fidelity of explanation and its interpretability

$$e(x) = \arg\min_{g \in G} L(f, g, \pi) + \Omega(g)$$



- L is a local fidelity function, f is a model to be explained, g is an interpretable local model g (i.e. linear model), $\pi(x, z)$ is proximity measure between the explained instance *x* and perturbed points *z* in its neighborhood, Ω is a model complexity measure

LIME details



- LIME samples around the explanation instance *x* to draw samples *z* weighted by the distance $\pi(\mathrm{x}, \mathrm{z})$
- samples *z* are used to training an interpretable model *g* (linear model)
- the squared loss measures local infidelity
- number of non-zero weights is complexity
- samples are weighted according to the Gaussian distribution of the distance between *x* and *z*

LIME strengths and weaknesses

- faster than IME
- works for many features, including text and images

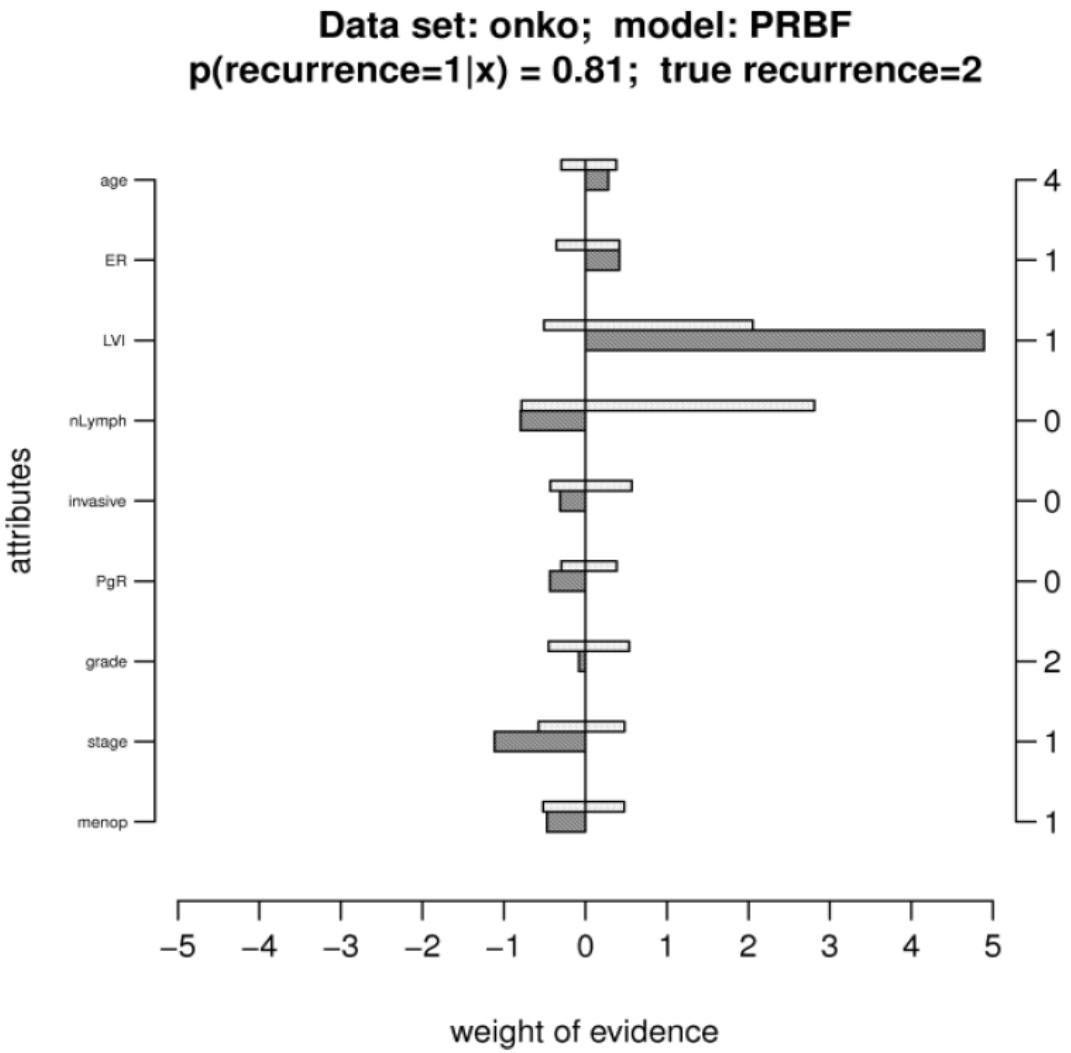
- no guarantees that the explanations are faithful and stable
- neighborhood based: a curse of dimensionality
- may not detect interactions due to (too) simple interpretable local model (linear model)

SHAP

- SHapley Additive exPlanation
- unification of several explanation methods, including IME and LIME
- KernelSHAP: based on Shapley values which are estimated using a LIME style linear regression
- faster then IME but
- still uses linear model with all its strengths and weaknesses

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Use case: breast cancer recurrence



Cancer recurrence within 10 years

menop binary feature indicating menopausal status

stage tumor stage 1: less than 20mm, 2: between 20mm and 50mm, 3: over 50mm

grade tumor grade 1: good, 2: medium, 3: poor, 4: not applicable, 9: not determined

histType histological type of the tumor 1: ductal, 2: lobular, 3: other

PgR level of progesterone receptors in tumor (in fmol per mg of protein) 0: less than 10, 1: more than 10, 9: unknown

invasive invasiveness of the tumor 0: no, 1: invades the skin, 2: the mamilla, 3: skin and mamilla, 4: wall or muscle

nLymph number of involved lymph nodes 0: 0, 1: between 1 and 3, 2: between 4 and 9, 3: 10 or more

famHist medical history 0: no cancer, 1: 1st generation breast, ovarian or prostate cancer 2: 2nd generation breast, ovarian or prostate cancer, 3: unknown gynecological cancer 4: colon or pancreas cancer, 5: other or unknown cancers, 9: not determined

LVI binary feature indicating lymphatic or vascular invasion

ER level of estrogen receptors in tumor (in fmol per mg of protein) 1: less than 5, 2: 5 to 10, 3: 10 to 30, 4: more than 30, 9: not determined

maxNode diameter of the largest removed lymph node 1: less than 15mm, 2: between 15 and 20mm, 3: more than 20mm

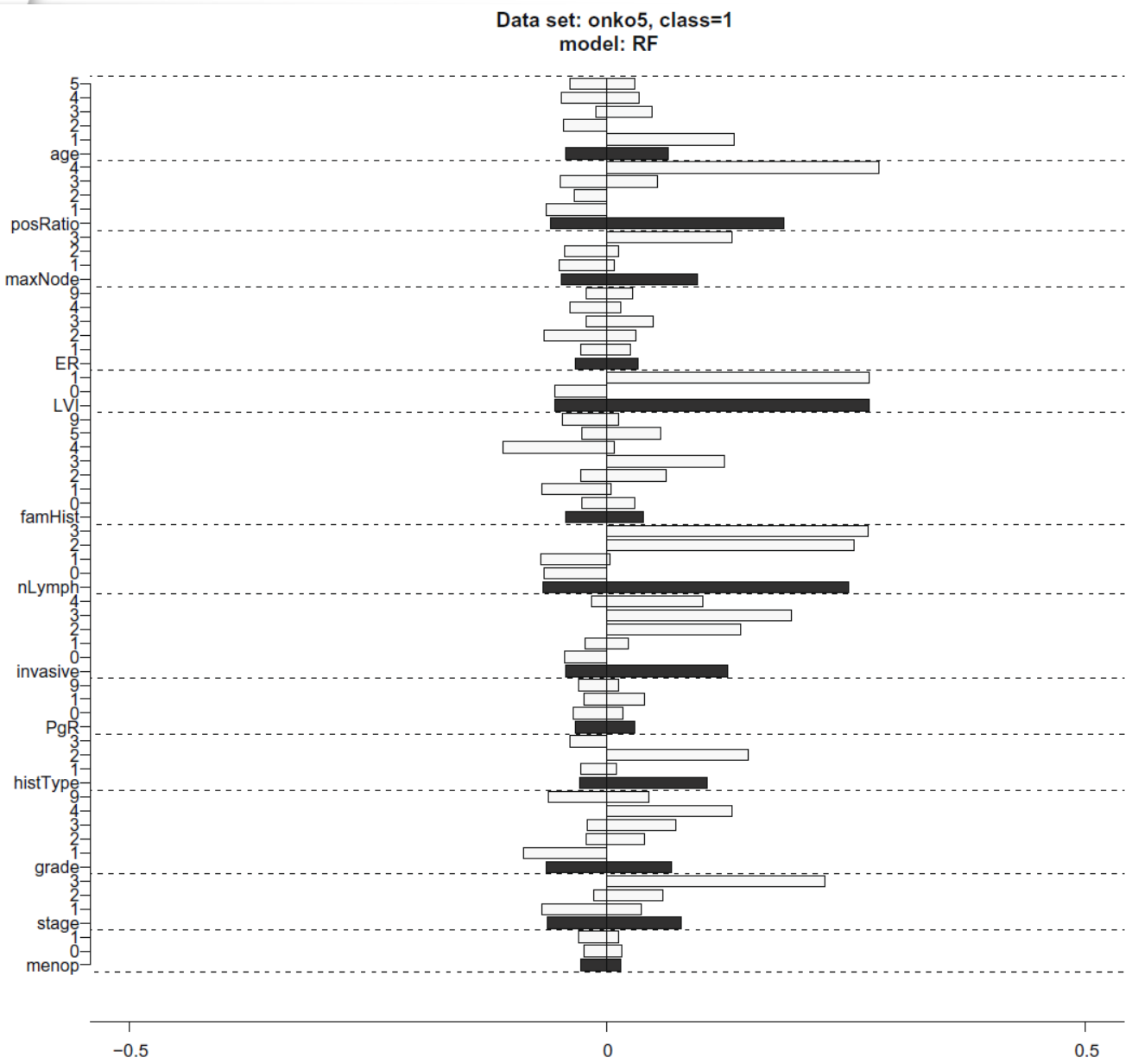
posRatio ratio between involved and total lymph nodes removed 1: 0, 2: less that 10%, 3: between 10% and 30%, 4: over 30%

age patient age group 1: under 40, 2: 40-50, 3: 50-60, 4: 60-70, 5: over 70 years

Robnik-Šikonja, M., Kononenko, I., & Štrumbelj, E. (2012). Quality of classification explanations with PRBF. *Neurocomputing*, *96*, 37-46.

University of Ljubljana
Faculty of Computer and Information Science

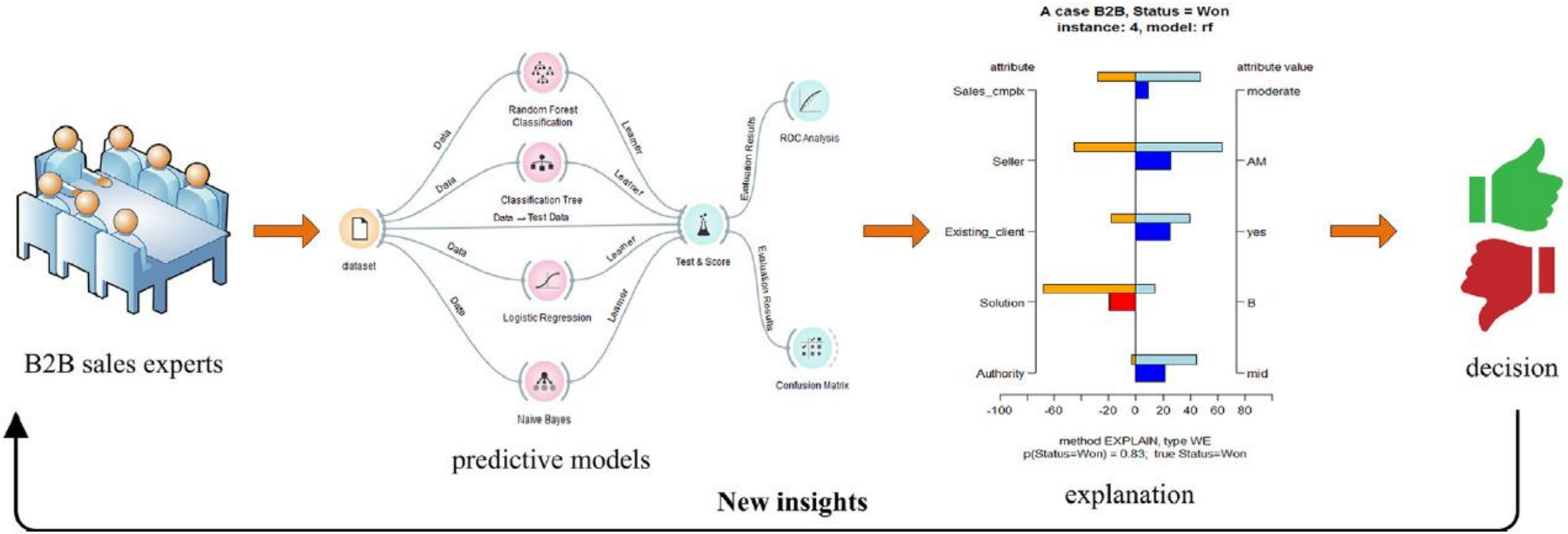
Use case: breast cancer recurrence



University of Ljubljana
Faculty of Computer and Information Science

Use case: B2B sales forecasting

- Goals: improve understanding of factors influencing the outcome and improve the sales performance



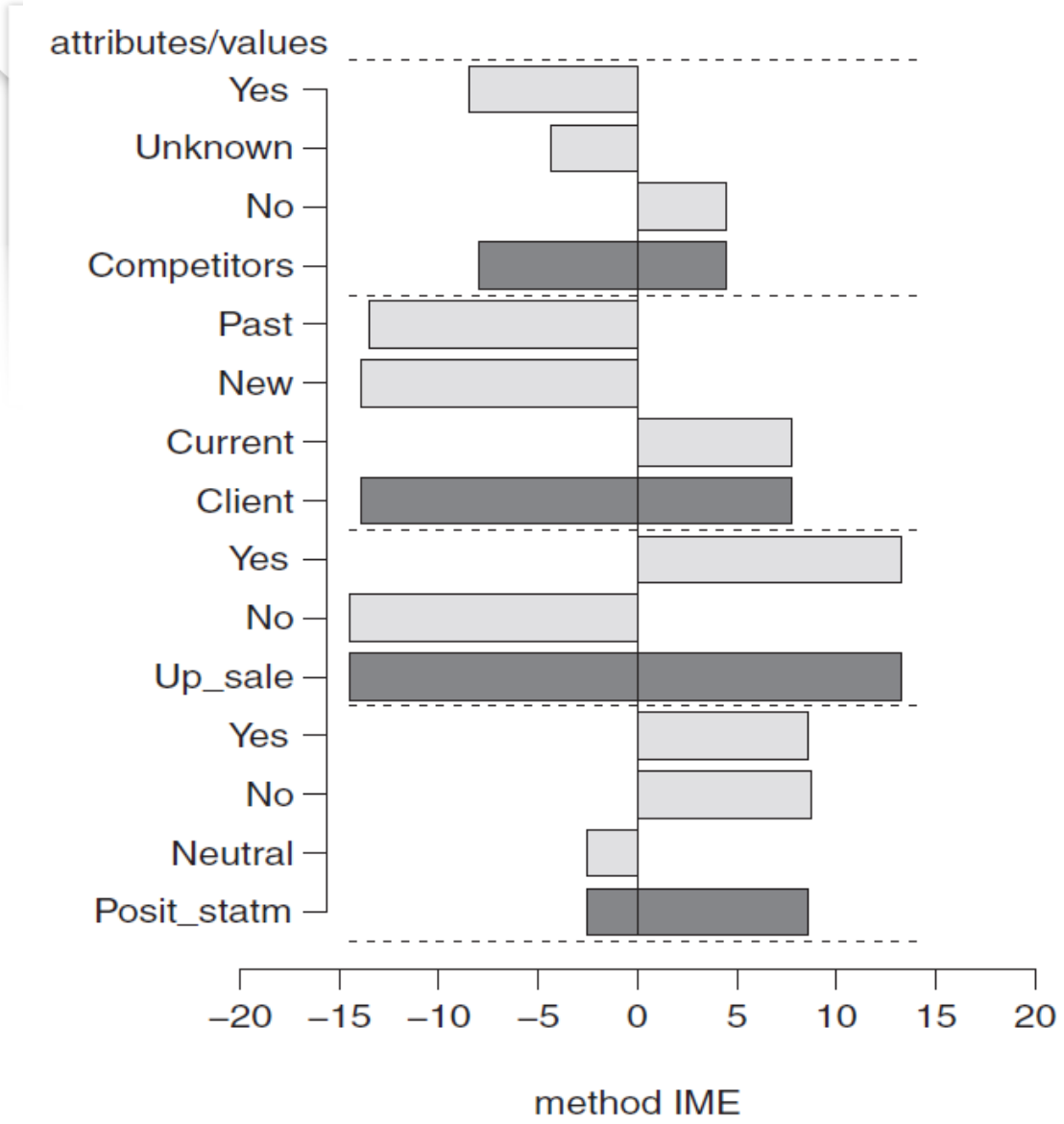
Bohanec, M., Borštnar, M. K., & Robnik-Sikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, *71*, 416-428.

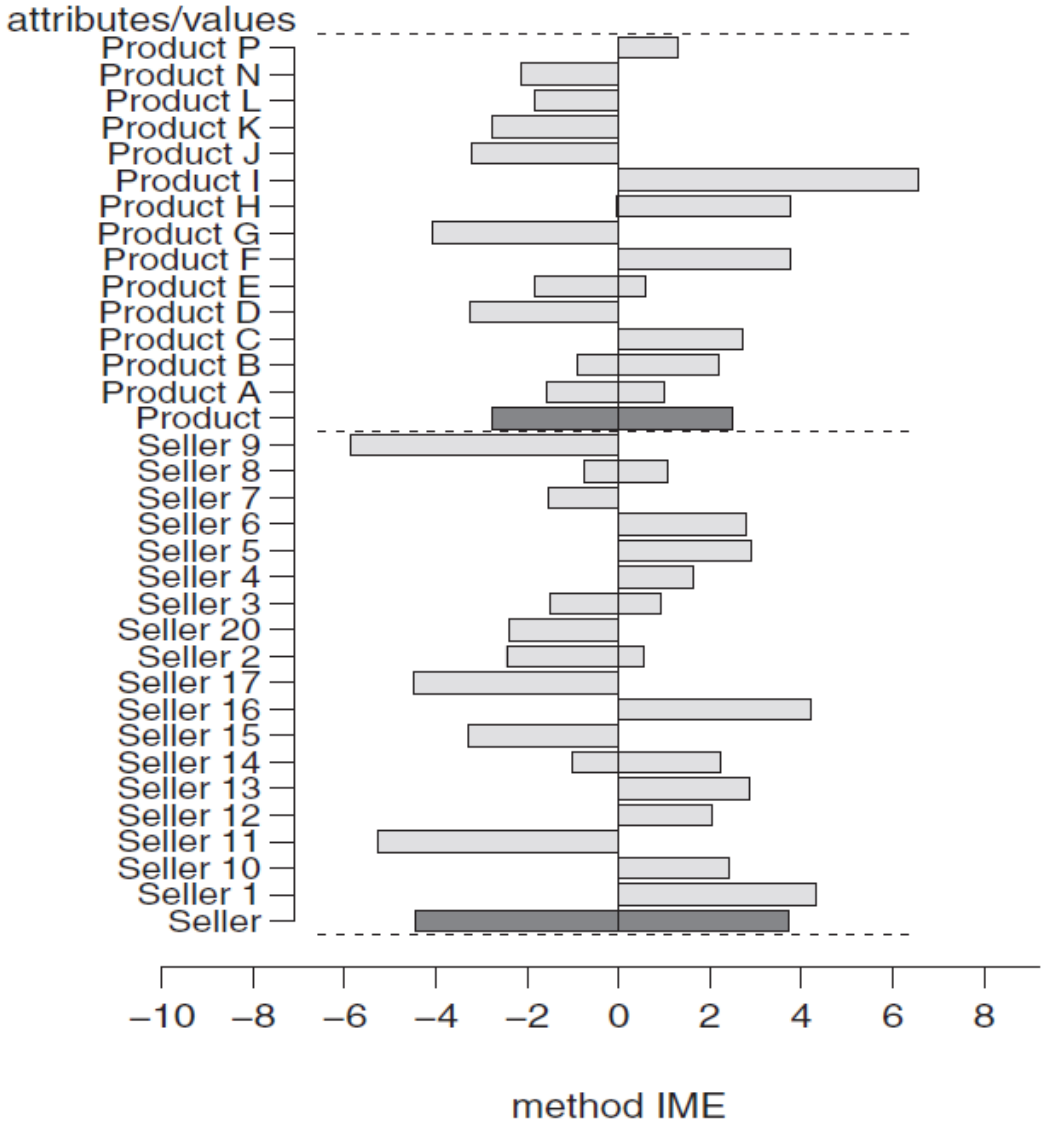
B2B sales attributes

Attribute	Description	Values
Authority	Authority level at a client side	Low, mid, high
Product	Offered product	e.g. A, B, C, etc.
Seller	Seller's name	Seller's name
Competitors	Do we have competitors?	No, yes, unknown
Company size	Size of a company	Big, mid, small
Purchasing department	Is the purchasing department involved?	No, yes, unknown
Partnership	Selling in partnership?	No, yes
Budget allocated	Did the client reserve the budget?	No, yes, unknown
Formal tender	Is a tendering procedure required?	No, yes
RFI	Did we get request for information?	No, yes
RFP	Did we get request for proposal?	No, yes
Growth	Growth of a client?	Growth, stable, etc.
Positive statements	Positive attitude expressed?	No, yes, neutral
Source	Source of the opportunity	e.g. referral, web, etc.
Client	Type of a client	New, current, past
Cross sale	A different product to existing client?	No, yes
Scope clarity	Implementation scope defined?	Clear, few questions, etc.
Strategic deal	Does this deal have a strategic value?	Very important, etc.
Up sale	Increasing sales of existing products?	No, yes
Deal type	Type of a sale	Consulting, project, etc.
Needs defined	Is client clear in expressing the needs?	Info gathering, etc.
Attention to client	Attention to a client	First deal, normal, etc.
Status	An outcome of sales opportunity	Lost, won

University of Ljubljana
Faculty of Computer and Information Science

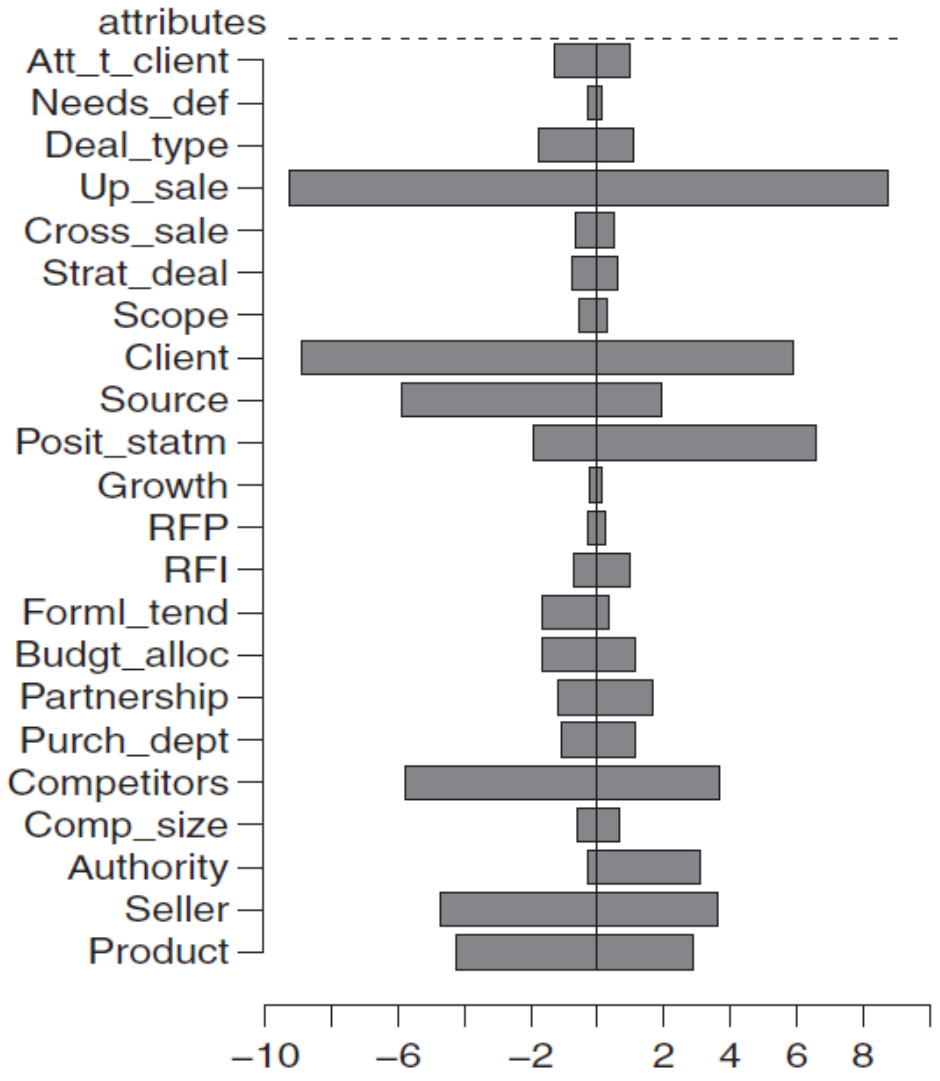
B2B sales: drill in



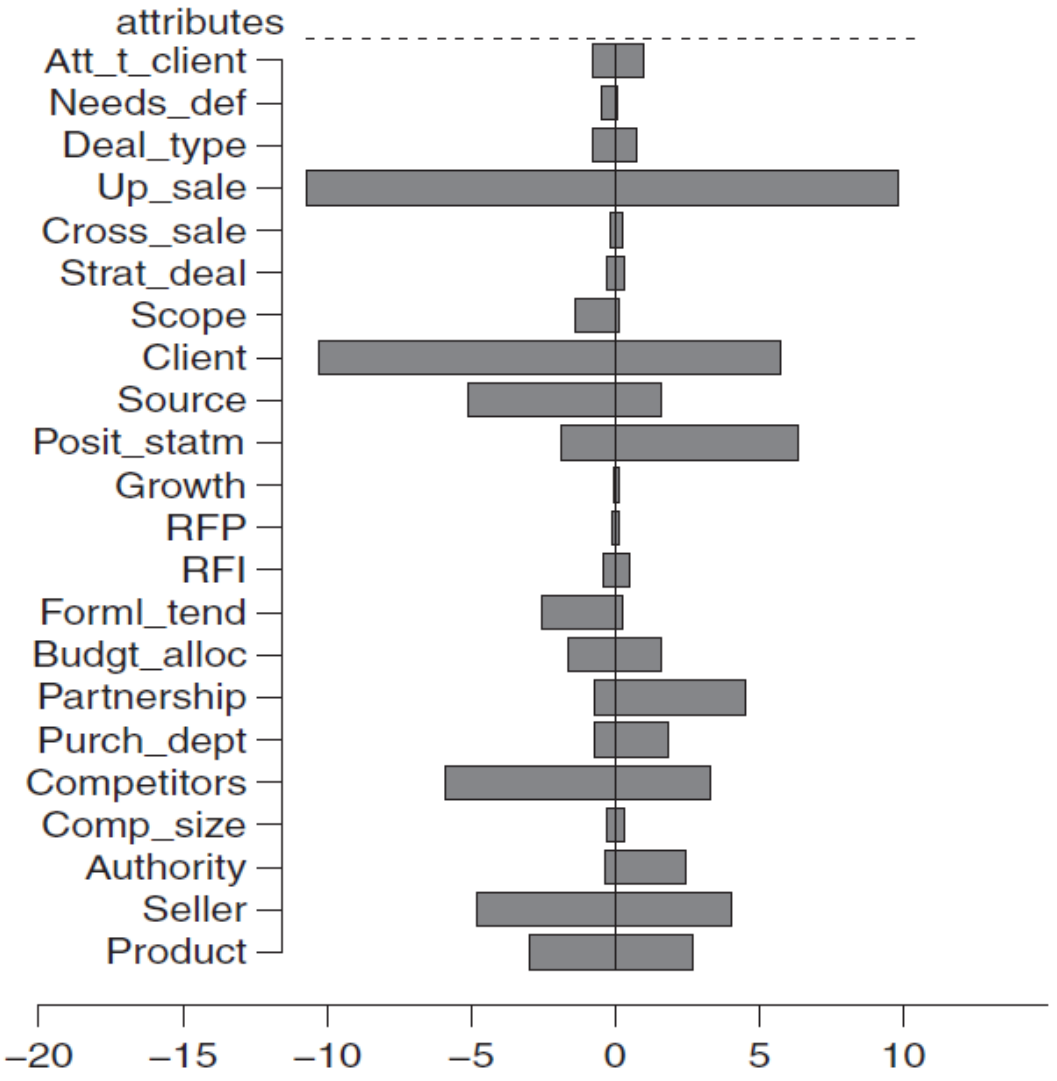


University of Ljubljana
Faculty of Computer and Information Science

B2B sales: EXPLAIN and IME



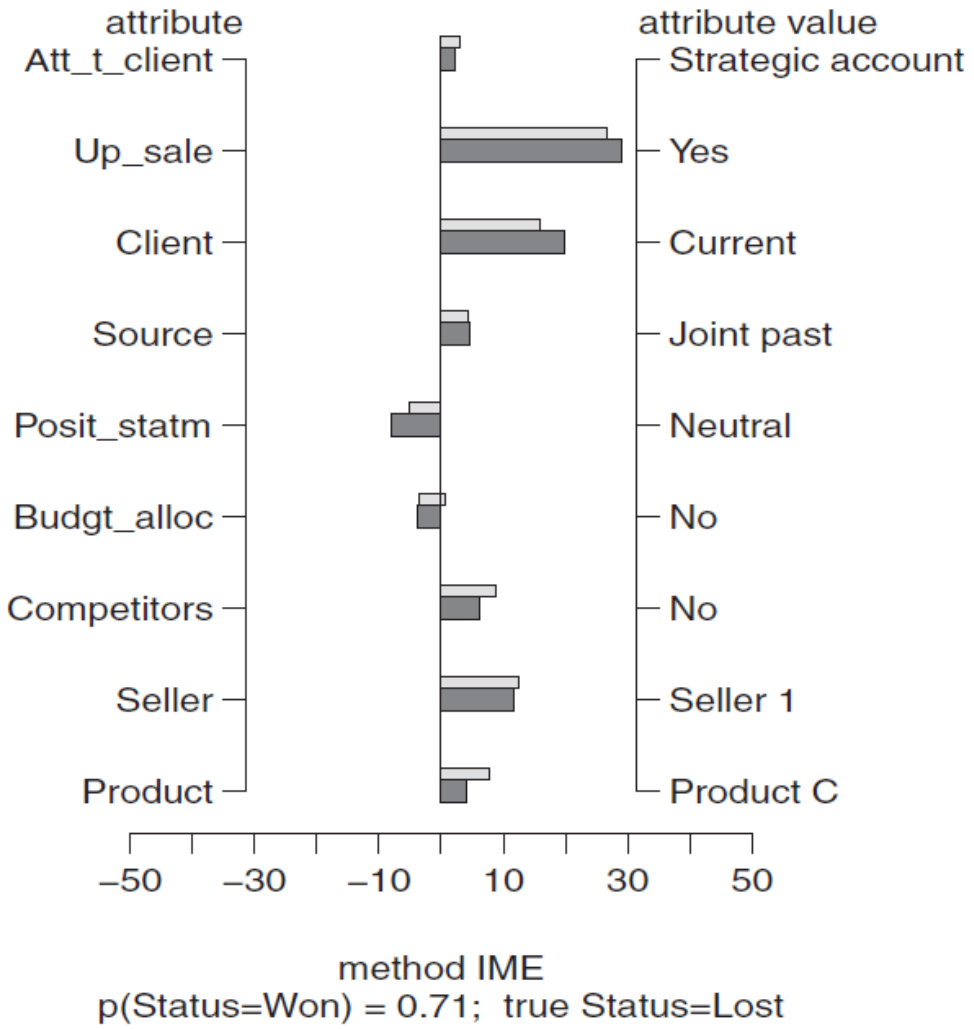
method EXPLAIN, type WE

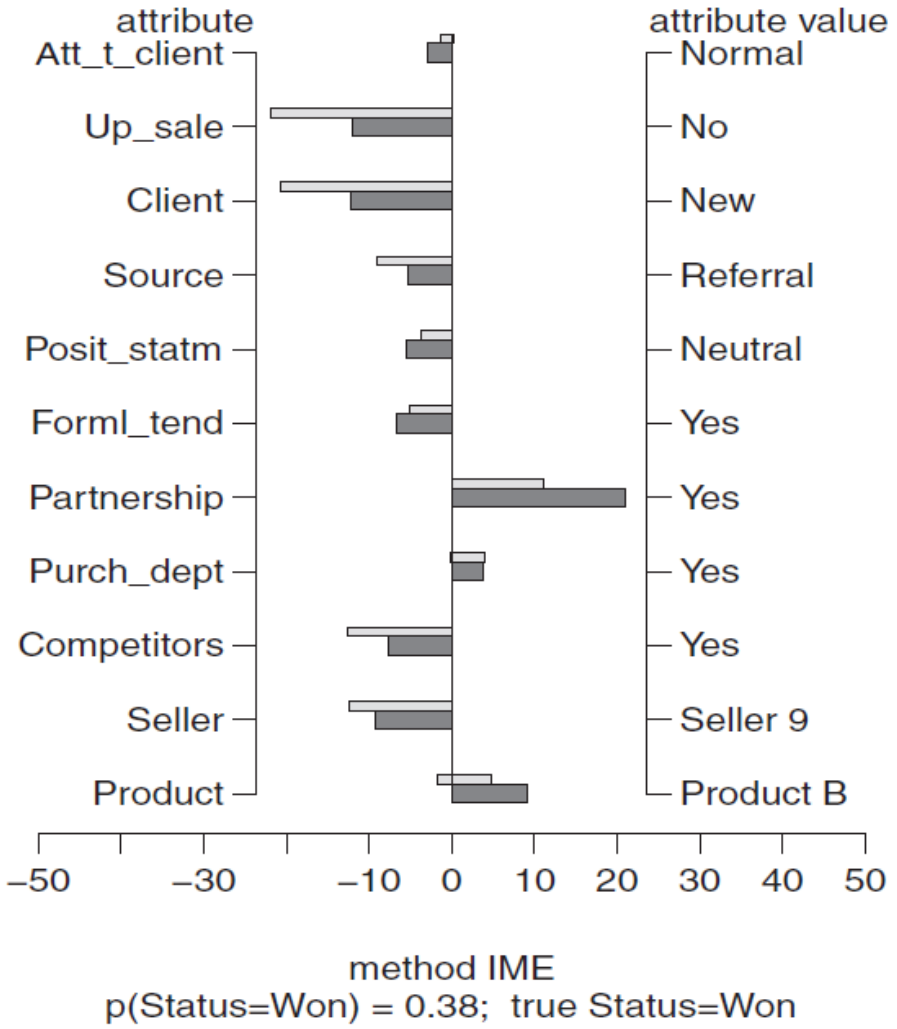


method IME

University of Ljubljana
Faculty of Computer and Information Science

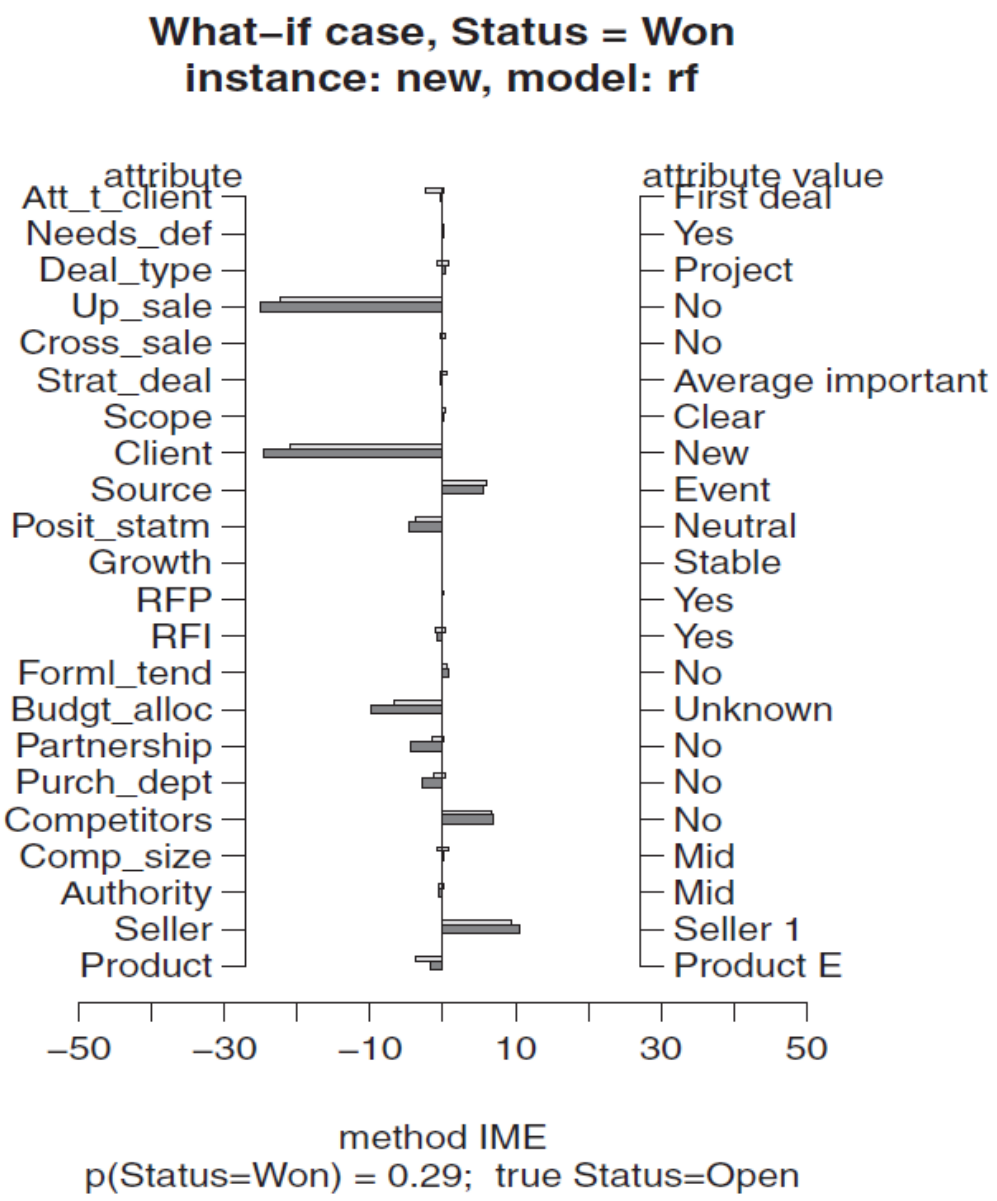
B2B sales: learning from errors





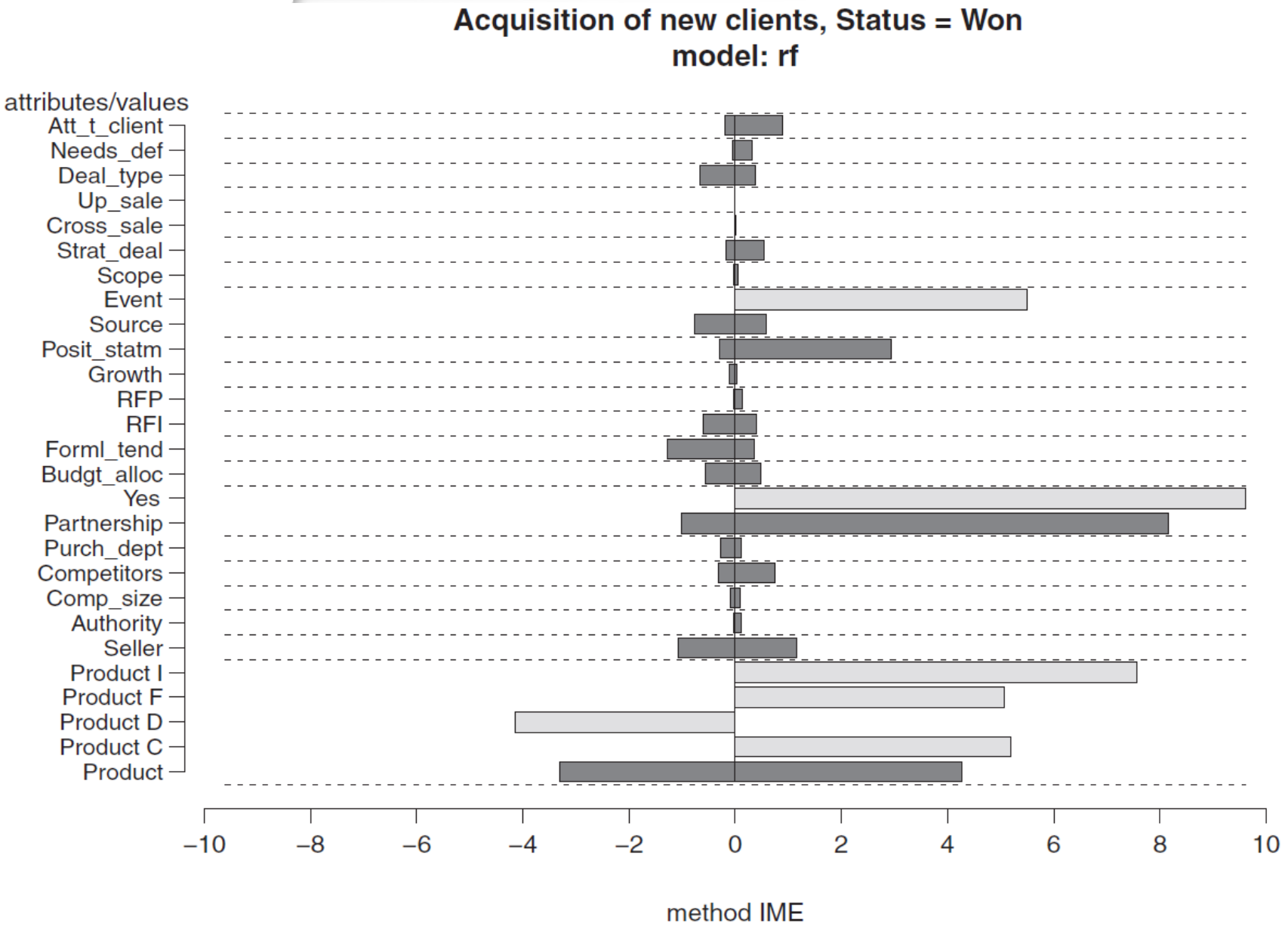
University of Ljubljana
Faculty of Computer and Information Science

B2B: what if

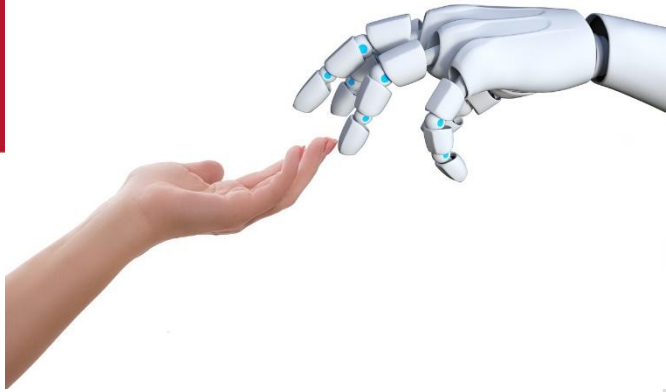


University of Ljubljana
Faculty of Computer and Information Science

B2B: change of distribution



University of Ljubljana
Faculty of Computer and Information Science



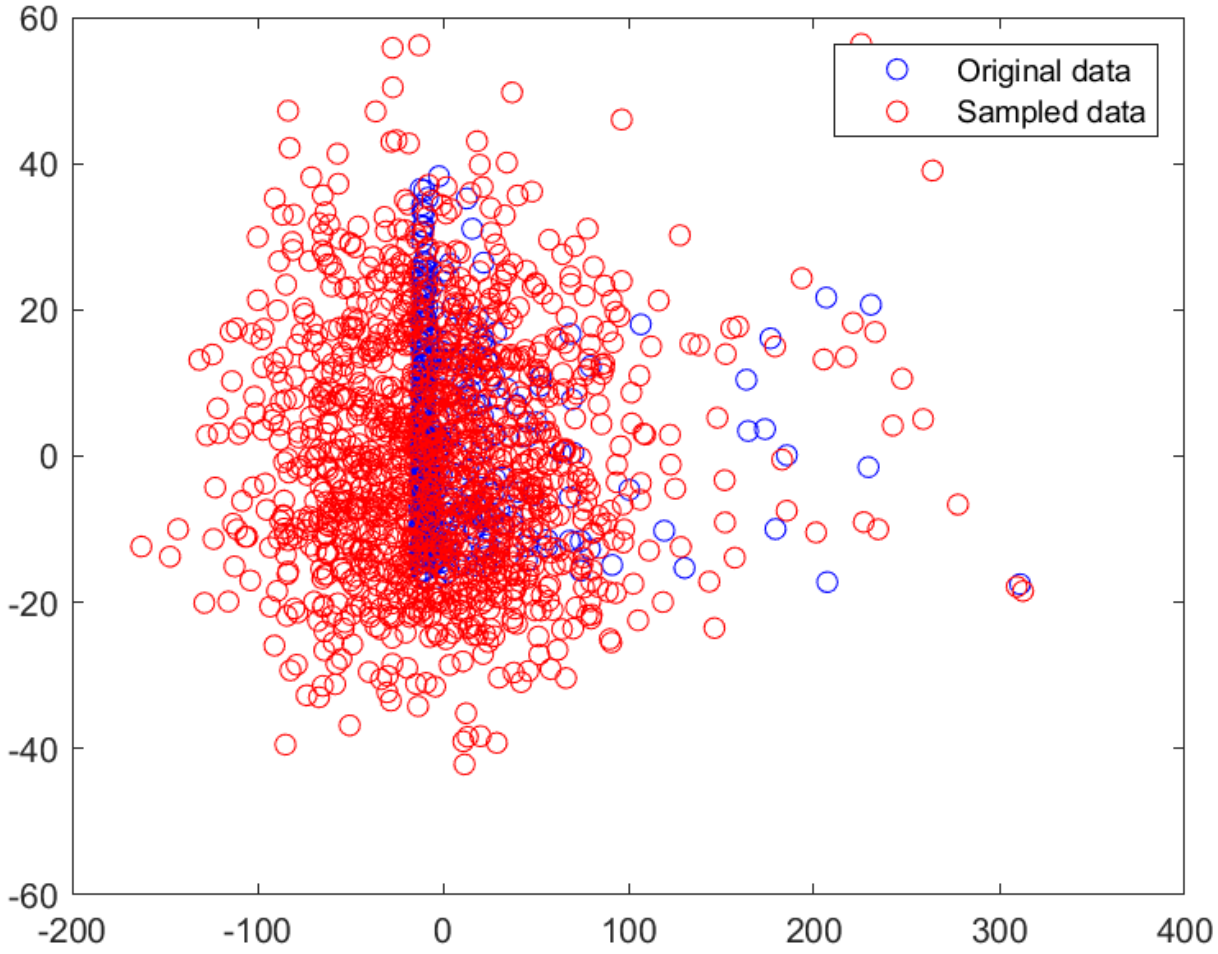
Lessons learned in B2B

- an effort needed to overcome the users' resistance
- human-in-the-loop is necessary to train, discuss, clean data, introduce explanations
- with an increased use, users gain trust in the methodology
- human mental models tend to be biased
- joint interactive approach beats both humans and ML models

University of Ljubljana
Faculty of Computer and Information Science

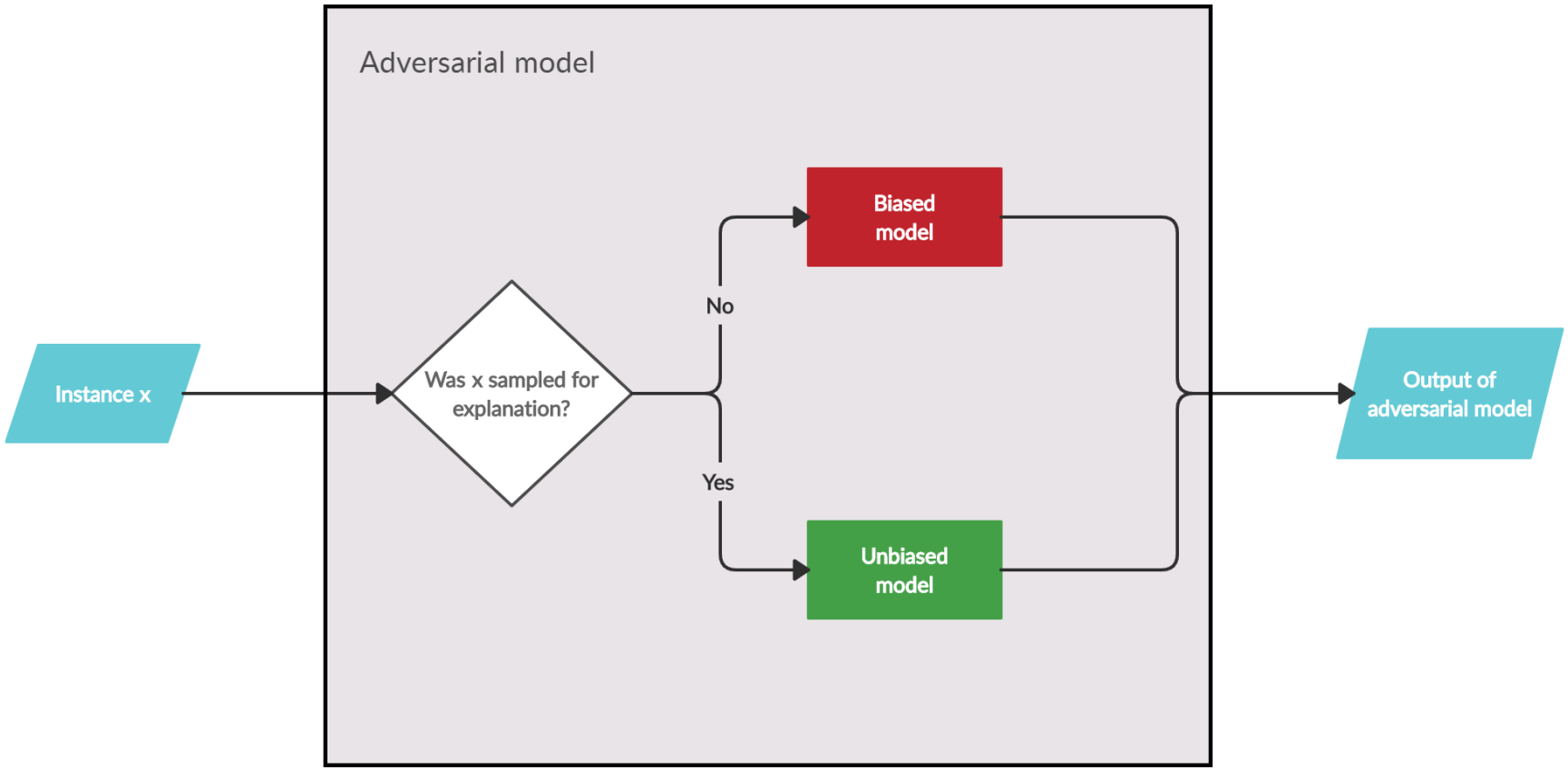
Attacks on explanations

- Poor sampling in explanation approaches makes them vulnerable
- Example: PCA based visualization of a part of the COMPAS dataset; the red dots were generated by LIME



University of Ljubljana
Faculty of Computer and Information Science

Dieselgate attacks on explanations

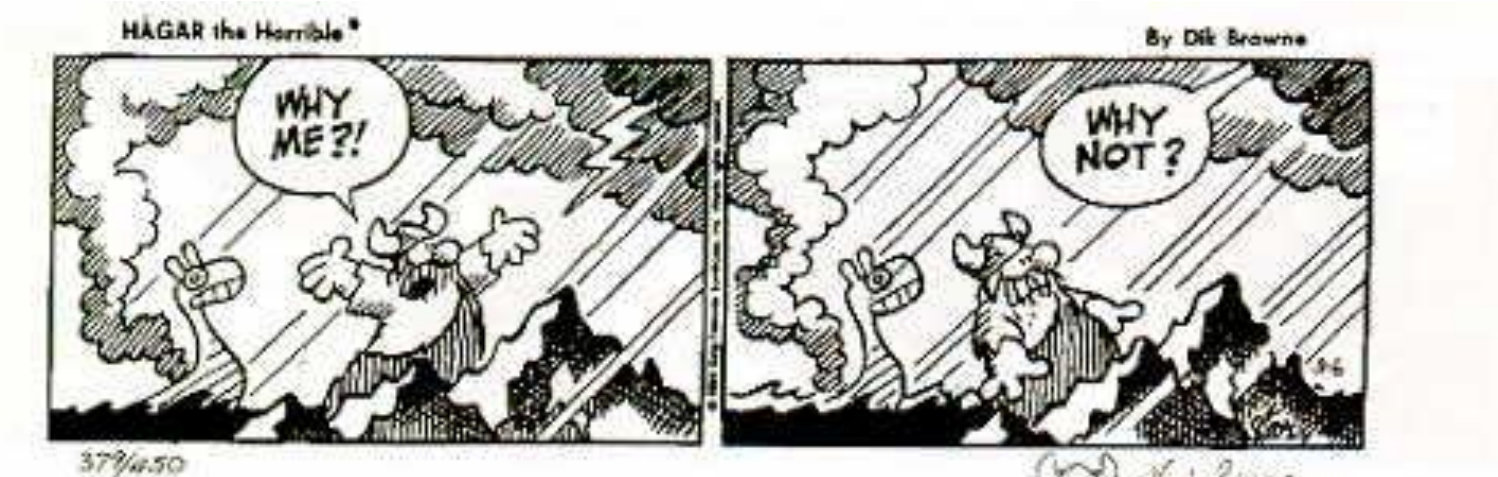
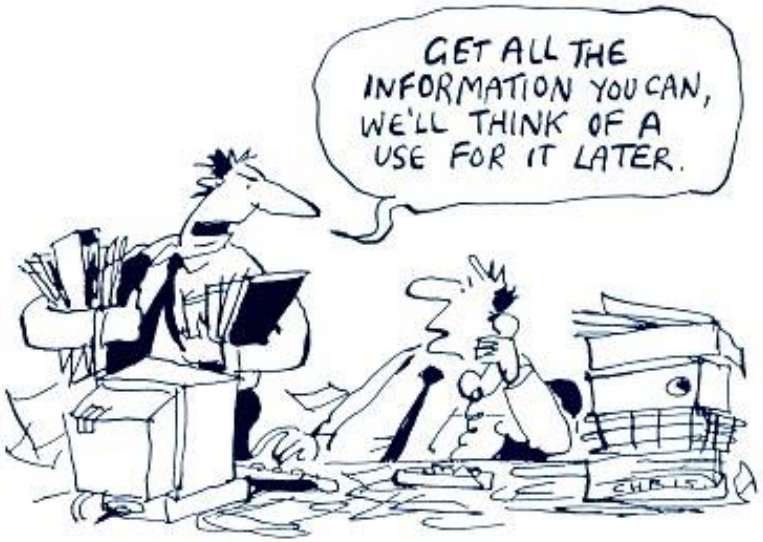


- Defence: better sampling

Domen Vreš and Marko Robnik-Šikonja, 2021. Better sampling in explanation methods can prevent dieselgate-like deception. *arXiv preprint arXiv:2101.11702*.

Opportunities

- better and more focused sampling
- better local explanation models
- interactions: detect and describe
- sequences: the order of attributes is important!
- images: decison areas, super-pixels
- better visualizations: human cognitive limitations
- explanations is also domain specific, we need explanation datasets

Summary of explanation approaches

- many successful mechanistic explanation approaches, mostly for tabular classification problems
- LLMs are trained to explain their behavior for particular important problems
- lots of opportunities for improvements
- even human explanations are not necessarily comprehensible
- humans often explain by providing background or additional knowledge
- legal and practical need for explanations of ML models