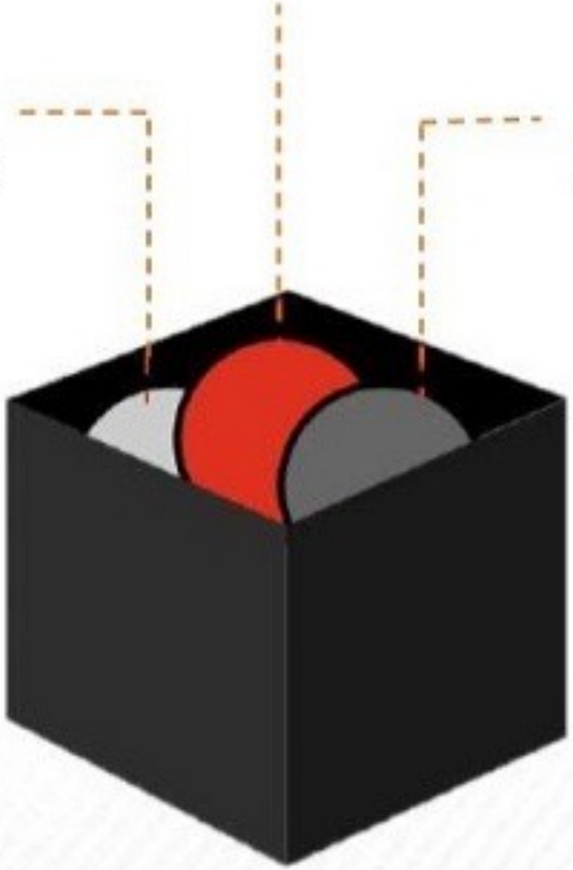
Universitat de les Illes Balears

Machine Learning

Explainable AI

Shapley values



Explainability

Can I explain the decision?

Robustness

Is it secure and compliant?

Bias

Have I been unfair?

Black-box Decisioning Models

(Statistical models, ML Models, Rules)

If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

- **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against underrepresented groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.

- **Privacy**: Ensuring that sensitive information in the data is protected.

- **Reliability or Robustness**: Ensuring that small changes in the input do not lead to large changes in the prediction.

- **Causality**: Check that only causal relationships are picked up.

- **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

Interpretable models

- The easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models:
- Linear regression
- logistic regression
- decision tree

- However, these models are either high bias - low variance (prone to underfitting) or low bias – high variance (prone to overfitting)
- What about model agnostic interpretability methods ?

Model Agnostic Interpretability methods

- Typically, not just one, but many types of machine learning models are evaluated to solve a task. When model agnostic interpretability methods are used, machine learning developers are free to use any machine learning model they like.
- Anything that builds on an interpretation of a machine learning model, such as a graphic or user interface, also becomes independent of the underlying machine learning model.

- **Model flexibility:** The interpretation method can work with any machine learning model, such as random forests and deep neural networks.
- **Explanation flexibility:** You are not limited to a certain form of explanation. In some cases it might be useful to have a linear formula, in other cases a graphic with feature importances.
- **Representation flexibility:** The explanation system should be able to use a different feature representation as the model being explained. For a text classifier that uses abstract word embedding vectors, it might be preferable to use the presence of individual words for the explanation.

Model interpretability types

- Global methods:
    - Describe the average behaviour of a machine learning model
    - The modeler wants to understand the general mechanisms in the data or debug a model

- Local methods:
    - Local interpretation methods explain individual predictions.
    - The modeler wants to understand why a prediction has a particular value given an input X

Shapley values and SHAP

Shapley values

- Is a local attribution method that fairly assigns the prediction to individual features.
- attribution methods: the prediction of a single instance is described as the **sum** of feature effects

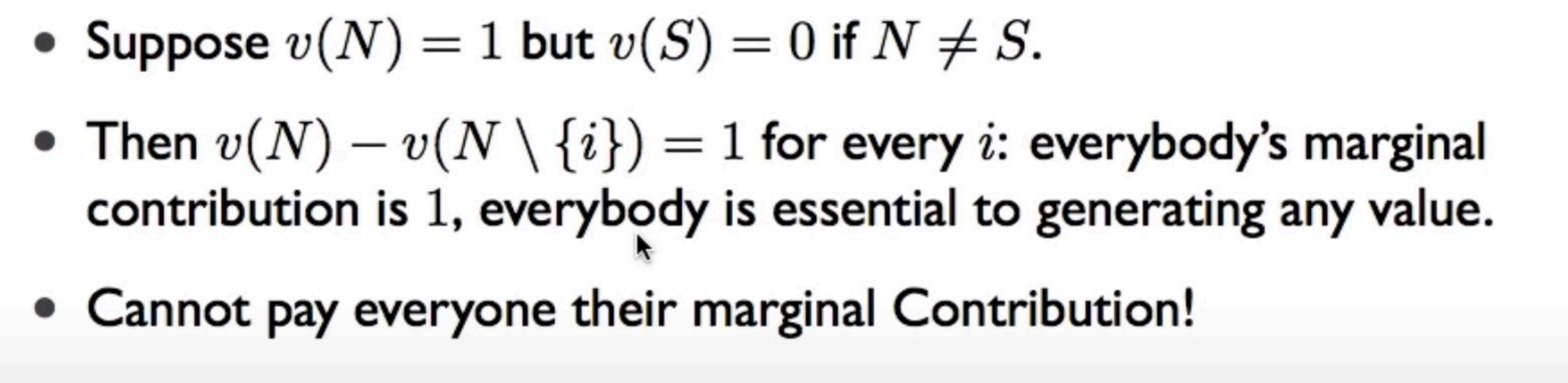
SHAP

- Another computation method for Shapley values
- Proposes global interpretation methods based on combinations of Shapley values across the data.

Shapley values first originated in the field of game theory in 1953 with the purpose of answering…

A group of differently skilled participants are all cooperating with each other for a collective reward.

How should the **reward** be **fairly divided** amongst the group?



- Suppose $v(N) = 1$ but $v(S) = 0$ if $N \neq S$.
- Then $v(N) - v(N \setminus \{i\}) = 1$ for every i: everybody's marginal contribution is 1, everybody is essential to generating any value.
- Cannot pay everyone their marginal Contribution!

Weighting the contributions should satisfy…

- **Symmetry**: This can be thought of as giving equal treatment to interchangeable agents
- **Dummy Player :** if a player does not contribute to the output it should receive nothing
- **Additivity**: If a game can be decomposed in two parts (v = v1 + v2), we should be able to decompose the retribution as well

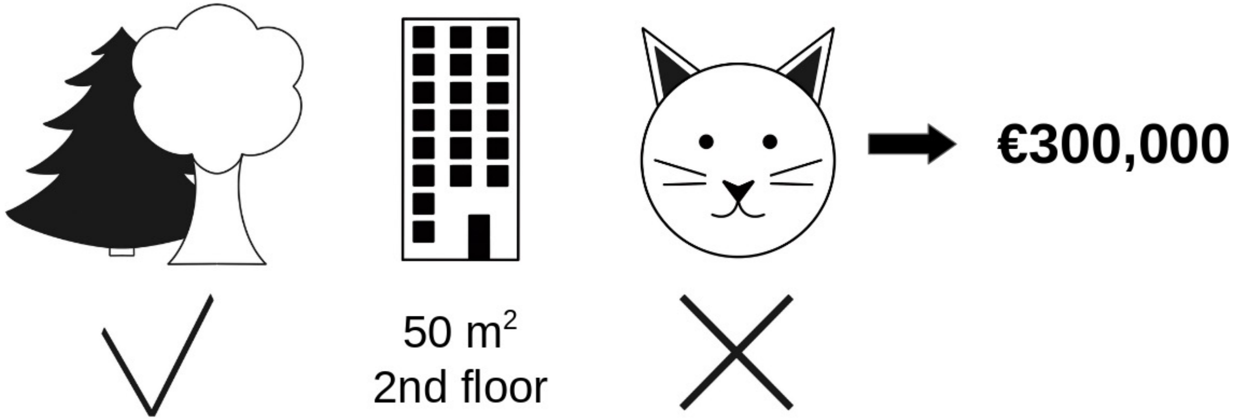
Marginal contribution of player i

$$\phi_i(N,v) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)! \Big[v(S \cup \{i\}) - v(S) \Big].$$

for each player i.

Weighted by all the possible “societies” we can build before adding player i

https://www.cse.iitk.ac.in/users/swaprava/courses/cs698a/scribe/lec08.pdf



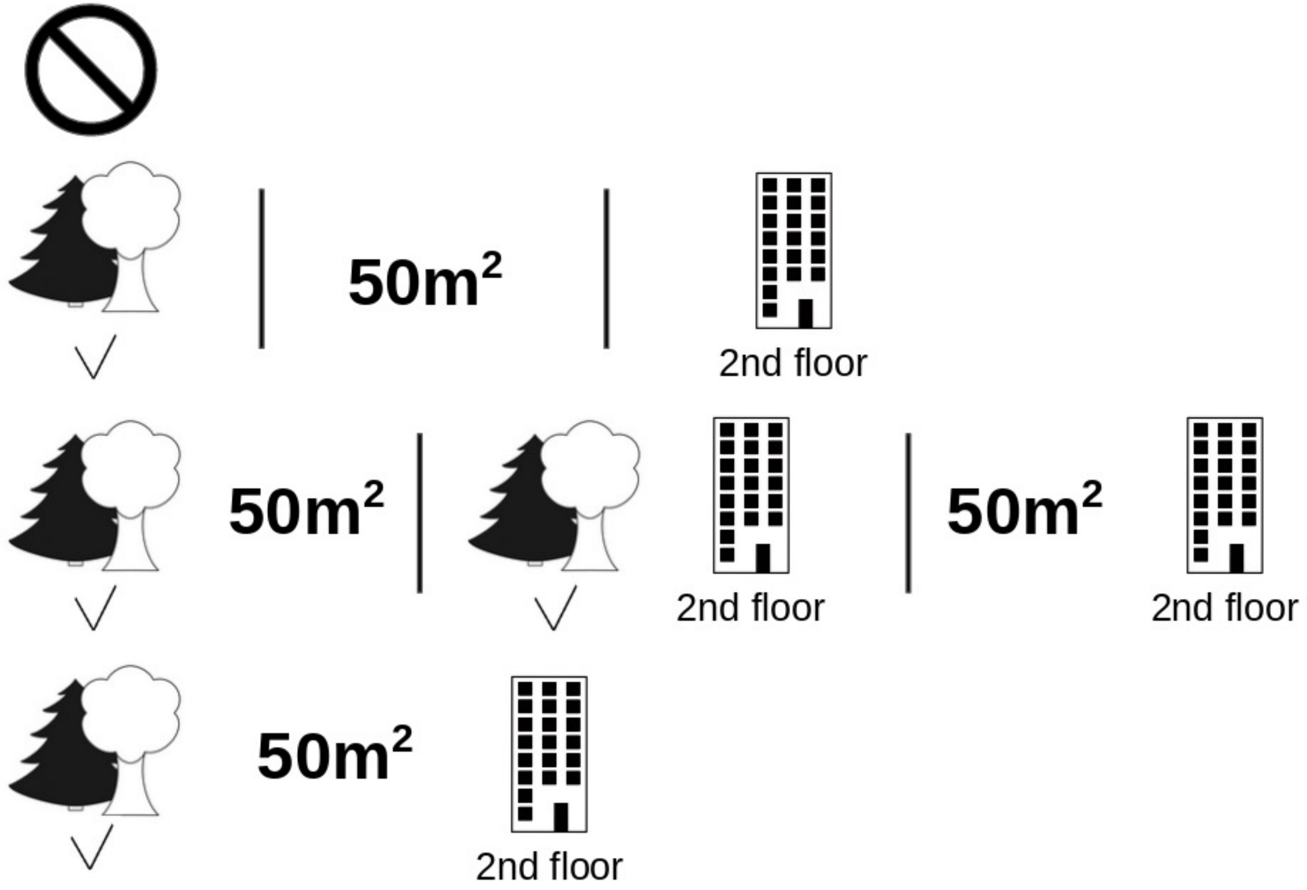
What is the connection to machine learning predictions and interpretability

The “game” is the prediction task for a single instance of the dataset.

The “gain” is the actual prediction for this instance minus the average prediction for all instances.

The “players” are the feature values of the instance that collaborate to receive the gain (= predict a certain value).

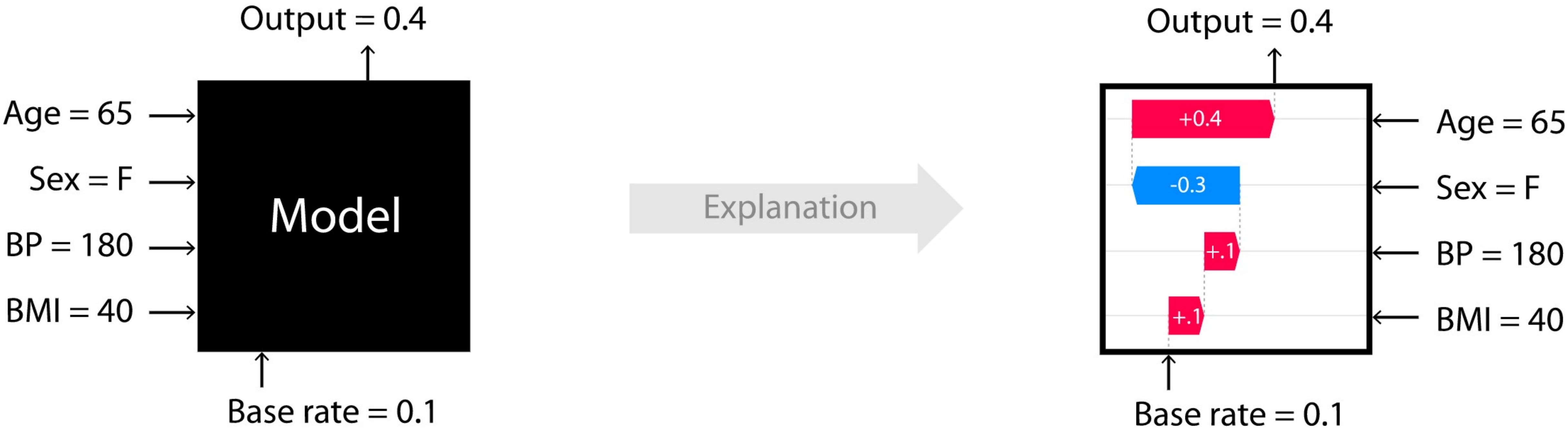
All possible coalitions (sets) of feature values have to be evaluated with and without the j-th feature to calculate the exact Shapley value



For a set of n players participating in a game you will have 2^n subsets that you will need to analyse in order to compute the Shapley values…

SHAP framework and its main strength is that it enables more computationally efficient calculation of Shapley values when applying them to machine learning

SHAP



Output = 0.4

Age = 65

Sex = F

BP = 180

BMI = 40

Model

Base rate = 0.1

Explanation

Output = 0.4

+0.4

-0.3

+.1

+.1

Age = 65

Sex = F

BP = 180

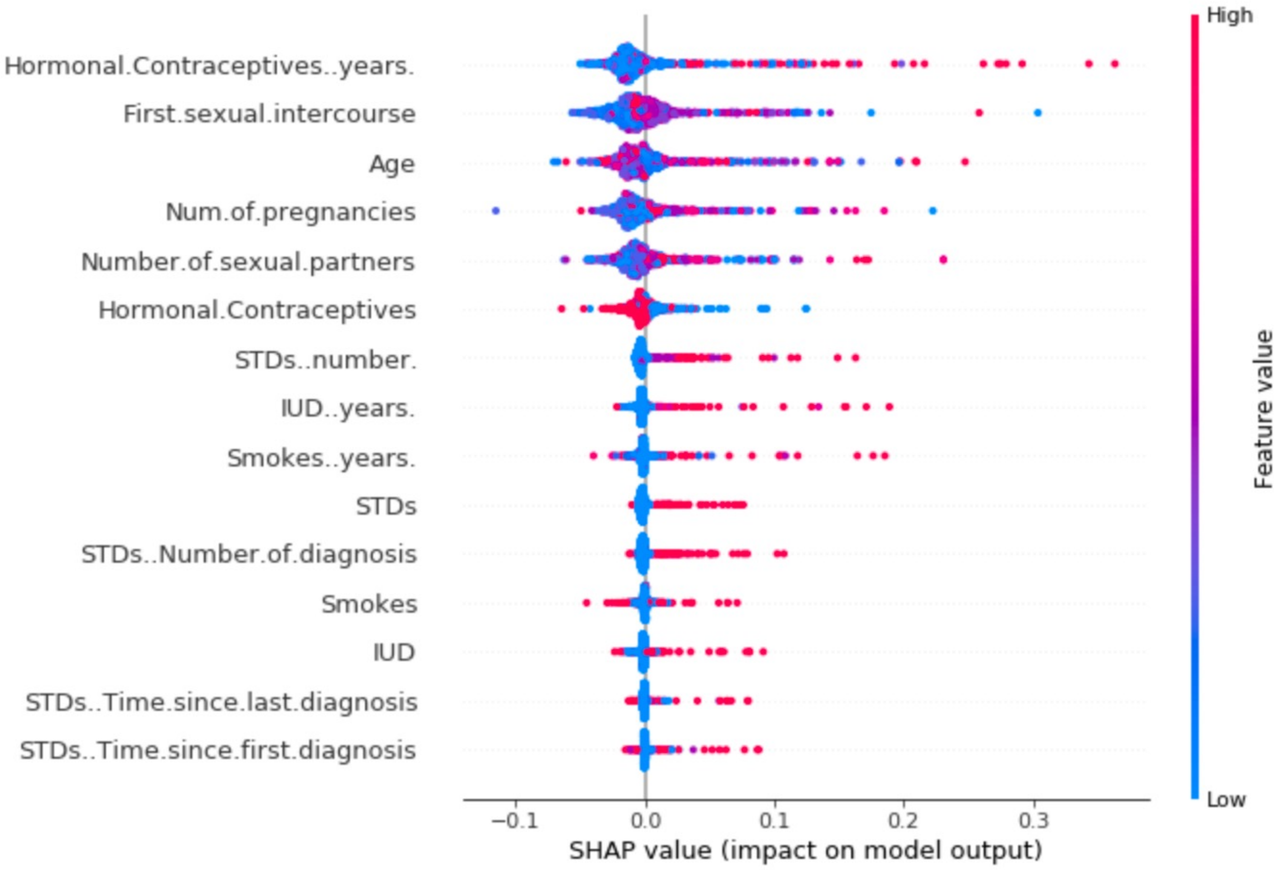
BMI = 40

Base rate = 0.1

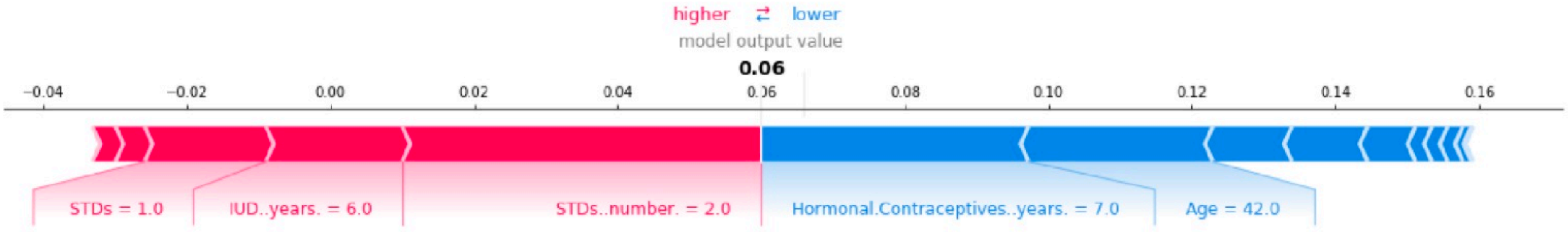
SHAP – summary plot

Each point on the summary plot is a Shapley value for a feature and an instance.

- y-axis: determined by the feature
- x-axis: the Shapley value.
- colour: represents the value of the feature from low to high



SHAP – force plot



* Low predicted risk of breast cancer (0.06)

* Risk increasing effects such as STDs (red) are compensated by decreasing effects such as age (blue)