Secrypt 2024

# Explainability and privacy-preserving data-driven models

Vicenç Torra

July, 2024

Dept. CS, Umeå University, Sweden

# Motivation

## Motivation[1]

- Is explainability still possible for privacy-preserving models?

[1]Talk based on (1) A. Bozorgpanah, V. Torra, L Aliahmadipour, Privacy and Explainability: The Effects of Data Protection on Shapley Values, Tech. 2022; (2) V. Torra, unpublished results

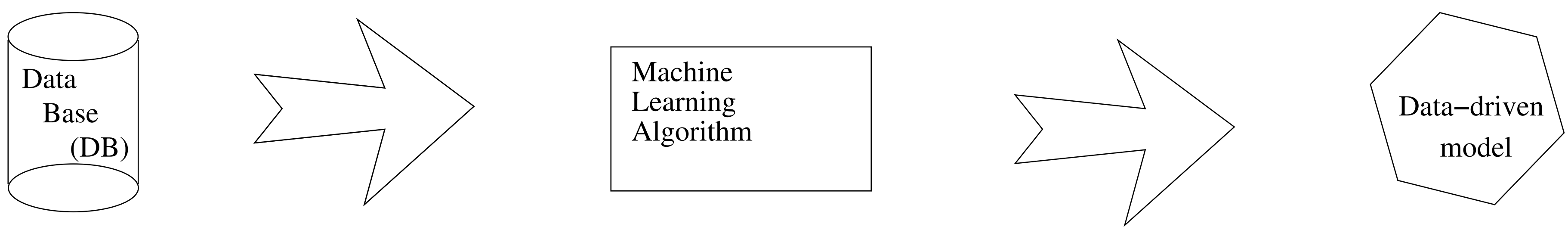
# Outline

1. Introduction
   - A context: Data-driven ML
   - Privacy for machine learning and statistics
   - Privacy models and masking methods
2. Explainability
   - AI and Explainability
   - Shapley values
3. Experiments and analysis
   - Methodology
   - Analysis
4. Future directions

# Introduction

# A context: Data-driven machine learning/statistical models

- From huge databases, build the "decision maker"
  - Use (logistic) regression, deep learning, neural networks, . . .

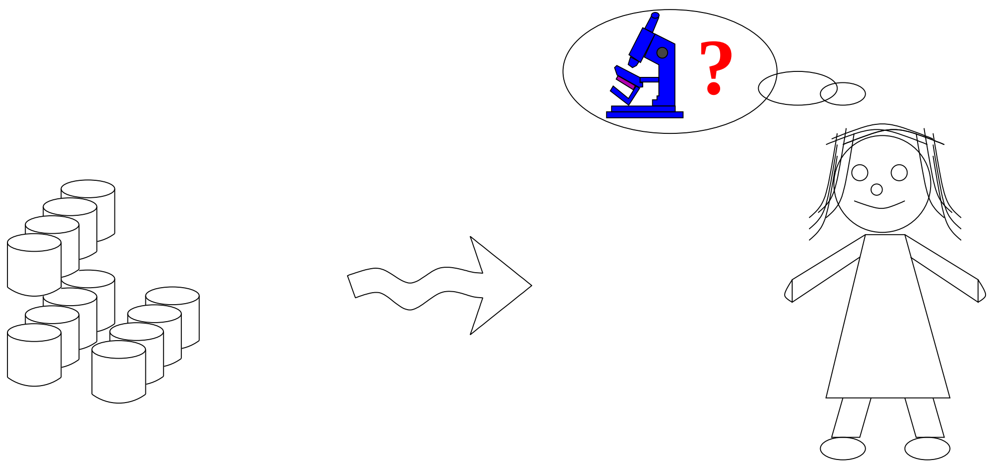



- Example: build a predictor from hospital historical data about length-of-stay at admission

# Privacy for machine learning and statistics

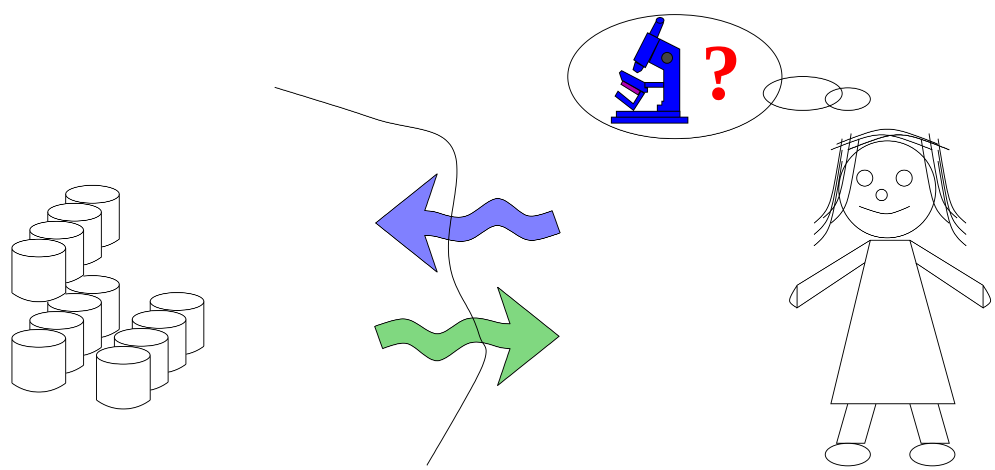
# Data is sensitive

- Who/how is going to create this model (this "decision maker")?
- Case #1. Sharing (part of the data)



# Data is sensitive

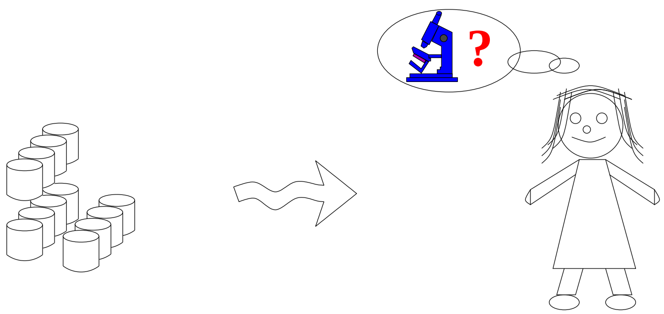
- Who/how is going to create this model (this "decision maker")?
- Case #2. Not sharing data, only querying data



# Data is sensitive

- Case #1. Sharing (part of the data)

- Naive anonymization does not work[2]. Few attributes cause disclosure.



- Predict length-of-stay, database with only *(year-birth, town, illness/ICD-9 codes)*
  1967, Umeå, circulatory system
  1957, Umeå, digestive system
  1964, Umeå, congenital anomalies
  1997, Umeå, injury and poisoning
  1986, Täfteå, injury and poisoning
  ...

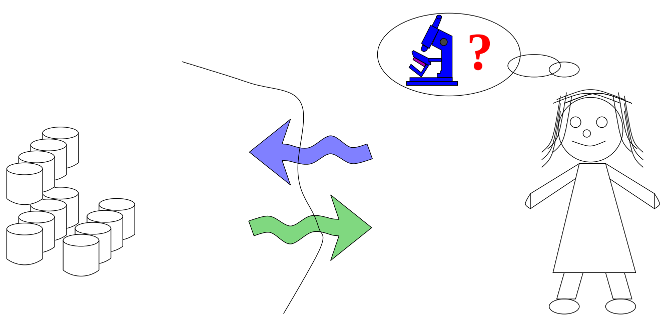
  However:
  1984, Holmöns distrikt, xxx

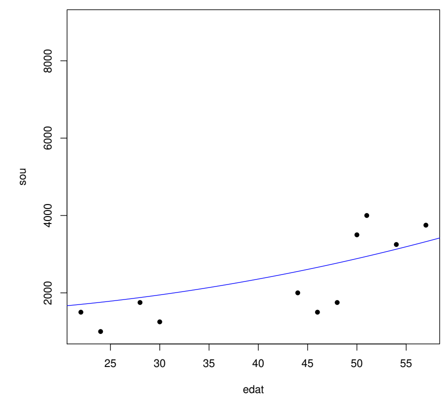
[2] Folkmängd: 62 (https://sv.wikipedia.org/wiki/Holm%C3%B6ns_distrikt)

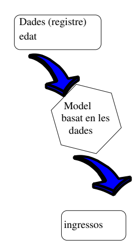
# Model is sensitive

- Case #2. Not sharing data, only querying data, sharing the model

- Models may reveal sensitive information
  - Income prediction vs. age for a town



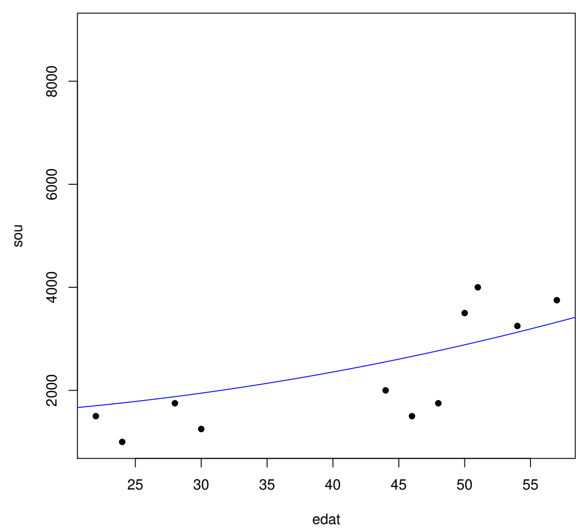


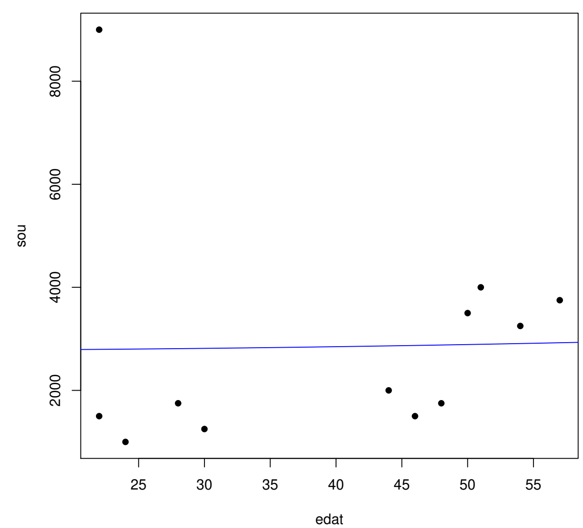




$$\text{income} = 1418.63 + 0.5864 * \text{age}^2$$

# Model is sensitive

- Case #2. Not sharing data, only querying data, sharing the model
- Models may reveal sensitive information
  Did they use my data (without permission)??
  - Membership inference attacks:
    We add Dona Obdúlia (who is very very rich and young)



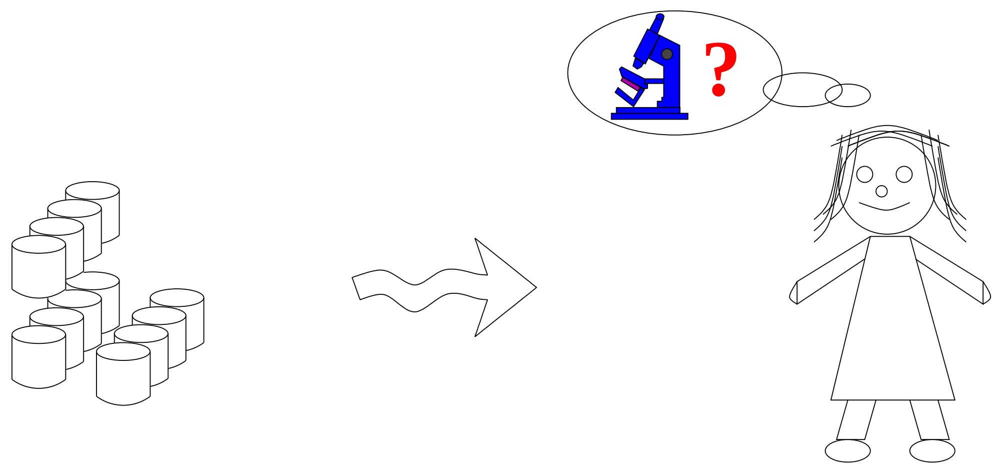




income $= 1418.63 + 0.5864$ * age$^2$ vs. income $= 2774 + 0.04639$ * age$^2$

# So, then, how?
# Privacy models and privacy solutions

# Data is sensitive: How to make ML possible?

- Who/how is going to create this model (this "decision maker")?
- Case #1. Sharing (part of the data)

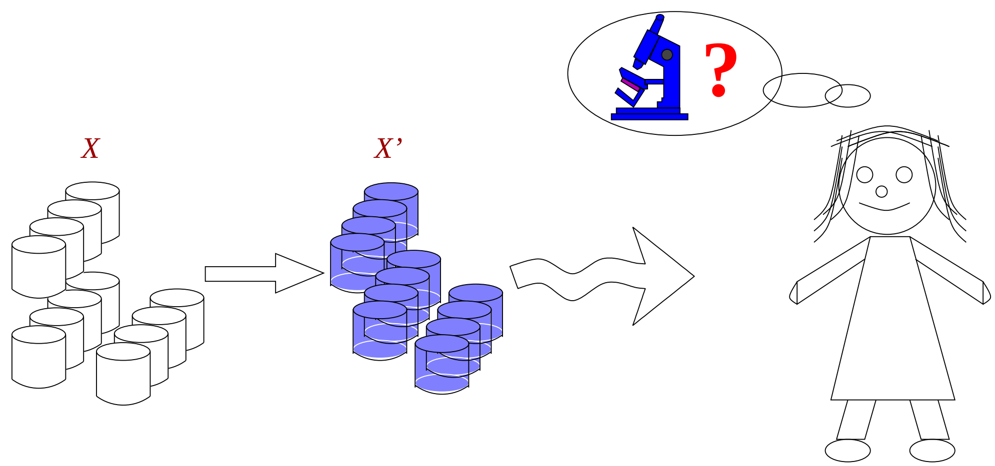



- Why data sharing?
  - Data scientists, statisticians, and ML researchers want the data.
  - Explore the data, apply several algorithms, test different parameters.
  - Other approaches (DP) properly applied degrade utility too much.

# Data is sensitive: How to make ML possible?

- Case #1. Sharing (part of the data)

- How ML is possible:
  - **Privacy models**. Computational definitions of privacy. E.g.,
    - $k$-Anonymity (Samarati, 2001)
    - reidentification privacy (Dalenius, 1986)
  - **Data protection mechanisms: masking methods**.
    to provide files with privacy guarantees
  - Remark. If $DB'$ is safe, any $f(DB')$ is safe.



# Data is sensitive: How to make ML possible?

- Case #1. Sharing (part of the data)

- Masking methods.
  - Methods $\rho$ to construct $DB'$ from $DB$.
  - Some examples (used in our experiments):
    - ▷ Microaggregation
    - ▷ Noise addition
    - ▷ Lossy compression and other transform-based methods

# Masking methods: Microaggregation

- Microaggregation (provides $k$-anonymity):
  - Group (cluster) a few (at least $k$) people with similar characteristics,
  - provide **safe** summaries of these people.
- Implementations
  - Different clustering / different summaries lead to different results
  - Examples: MDAV, Mondrian, others

# Masking methods: Noise addition

- Noise addition (to avoid re-identification, LDP):
  - replace $x$ by $x + r$
    with $r$ following an *appropriate distribution*
- Examples:
  - $\epsilon$ according to Normal distribution,
    zero mean, standard deviation as $\sqrt{(\text{variance} \cdot k)}$
  - $\epsilon$ according to Laplacian distribution (provides some LDP),
    zero mean, standard deviation as $\sqrt{(\text{variance} \cdot k)}$

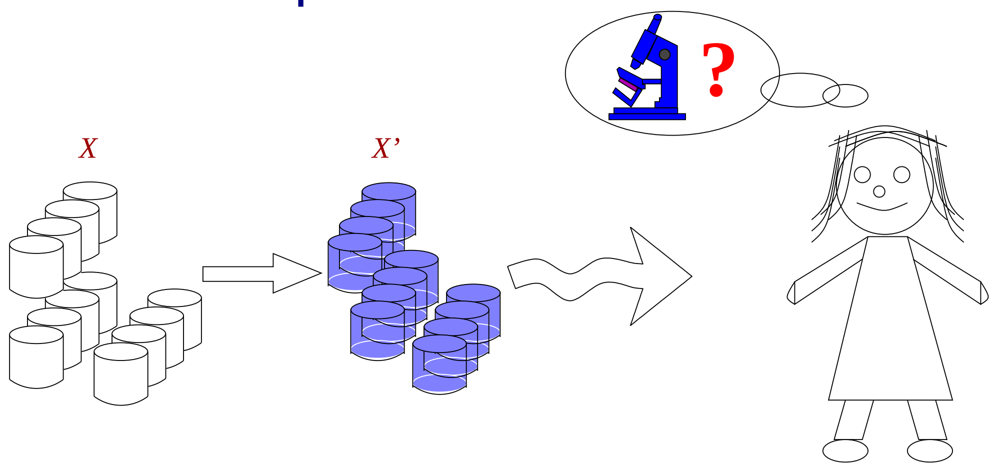
# Masking methods: Lossy compression / transform-based protection

- Compression (to avoid re-identification):
  - Apply a transformation
  - Select $main$ components
  - Undo the transformation

- Examples
  - SVD. Singular value decomposition. Select $k$ components.
  - PCA. Principal components. Select $k$ principal components.
  - NMF. Non-negative matrix factorization. Select $k$ components.

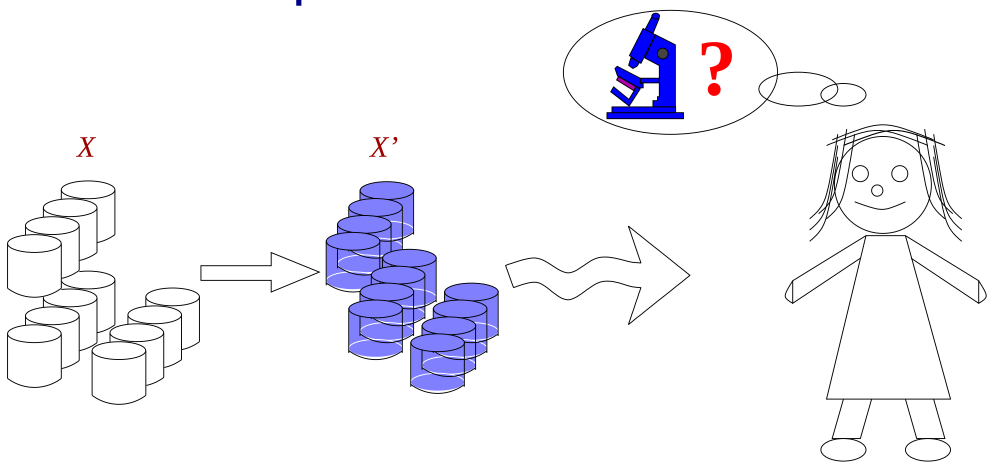
# Masking methods

- Masking methods cause a distortion to the data
  - Distortion depends on the parameter selected



# Masking methods

- Masking methods cause a distortion to the data
  - Distortion depends on the parameter selected




  - Quite a few studies on the effects of distortion on information loss some show that a small distortion may have no effect on IL
  - Quite a few studies on the effects of distortion on some disclosure risk measures (attribute disclosure, identity disclosure)

# Explainability

# Is explainability still possible for privacy-preserving models?

# Explainability?

# AI and explainability

- European regulation (GDPR) not only supports data protection and privacy, but also requirements on how decision making affecting people should be done.
  - Automated decisions should be explainable
- So, models need to be accurate, unbiased, etc. but also explainable

# AI and explainability

- Interpretable vs. explainable
  - Interpretable model: it is about the model itself. E.g., can we understand the model by inspection (e.g. decision trees)?
  - Explainable model: it is about the outputs.
- So, explainability, is for all models, including black-box models.

# AI and explainability

- Explainability
  - Model specific vs model agnostic
    - ▷ Model specific. When the explanation is based on the model itself
    - ▷ Model agnostic. The method is applicable to any kind of model.

# AI and explainability

- Explainability
  - Model specific vs model agnostic
    - ▷ Model specific. When the explanation is based on the model itself
    - ▷ Model agnostic. The method is applicable to any kind of model.
  - Global vs local methods
    - ▷ Global: Average behavior of the model. General mechanism behind
    - ▷ Local: Model's individual prediction

# AI and explainability

- Local model-agnostic methods. Examples.
  - Individual Conditional Expectation (ICE), Local Interpretable Model-agnostic Explanations (LIME), counterfactual explanation, Scoped Rules (Anchors), Shapley values (e.g., SHapley Additive exPlanations: SHAP).

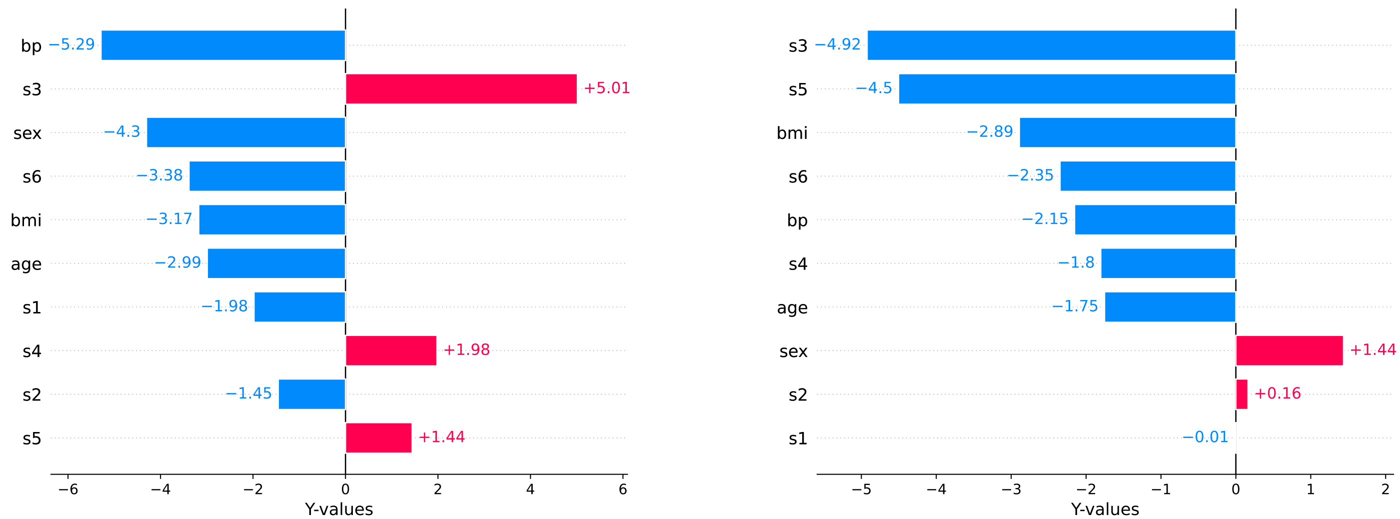
# AI and explainability

- Back to our questions
  - Is explainability still possible for privacy-preserving models?
- Why this question?
  - These methods for explainability are based on the data-driven model
  - If the data is perturbed, is the explanation still valid?

# Explainability: Shapley values

# Explainability: Shapley values and XAI

- Local model-agnostic methods: using Shapley values
  - a data-driven model $M$, applied to an example $u$
  - Why do we get $M(u)$?
  - Which variables contribute to $M(u)$? How much they contribute?




  - Figures. Shapley values for a record of the Diabetes data set (records 1 and 4 in the test set) computed from a model = SVM/SVR.

# Explainability: Shapley values

- Shapley values. An index from game theory.
  - We have a set function (a game) which provides values for coalitions
    A simple case, is a coalition a winning coalition?
  - Let $X$ be a set,
    parties in coalitions
    in our context the set of all variables

# Explainability: Shapley values

- Shapley values. An index from game theory.
  - We have a set function (a game) which provides values for coalitions
    A simple case, is a coalition a winning coalition?
  - Let $X$ be a set,
    parties in coalitions
    in our context the set of all variables
  - $\mu(S)$ for $S \subset X$ is the contribution of $S$.
    is $S$ a winning coalition, $\mu(S) = 1$; otherwise $\mu(S) = 0$
    considering only the variables of $S$, not the others

# Explainability: Shapley values

- Shapley values. An index from game theory.
  - From $\mu$ compute Shapley values $\phi$ for each $x$.
    $\phi_x$ represents the power/relevance of $x \in X$.
    For $X = \{x_1, \ldots, x_n\}$ we have values $\phi_{x_1}(\mu), \ldots, \phi_{x_n}(\mu)$.
  - These values are computed as

$$\phi_{x_i}(\mu) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}(\mu(S \cup \{i\}) - \mu(S)).$$

  - $\phi_{x_i}$ is the average contribution of $i$ when incorporated to a set

# Explainability: Shapley values

- Shapley values. An index from game theory.
  - From $\mu$ compute Shapley values $\phi$ for each $x$.
    $\phi_x$ represents the power/relevance of $x \in X$.
    For $X = \{x_1, \ldots, x_n\}$ we have values $\phi_{x_1}(\mu), \ldots, \phi_{x_n}(\mu)$.
  - These values are computed as

$$\phi_{x_i}(\mu) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!}(\mu(S \cup \{i\}) - \mu(S)).$$

  - $\phi_{x_i}$ is the average contribution of $i$ when incorporated to a set
    Example. If $x_i$ is a required party in any winning coalition, $\phi_{x_i}(\mu) = 1$.

# Explainability: Shapley values

- Shapley values. An index from game theory.
  - From $\mu$ compute Shapley values $\phi$ for each $x$.
    $\phi_x$ represents the power/relevance of $x \in X$.
    For $X = \{x_1, \ldots, x_n\}$ we have values $\phi_{x_1}(\mu), \ldots, \phi_{x_n}(\mu)$.
  - These values are computed as

$$\phi_{x_i}(\mu) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (\mu(S \cup \{i\}) - \mu(S)).$$

  - $\phi_{x_i}$ is the average contribution of $i$ when incorporated to a set
    Example. If $x_i$ is a required party in any winning coalition, $\phi_{x_i}(\mu) = 1$.
  - The Shapley value is the only power index that satisfies the dummy player condition, additivity, anonymity, and efficiency conditions.

# Explainability: Shapley values

- Shapley values. A power index from game theory.
  - Distributing $\mu(X)$ in a fair manner between the elements in $X$.
  - Efficiency condition. $\sum_{x \in X} \phi_x = \mu(X)$

# Explainability: Shapley values

- Shapley values. A power index from game theory.
  - Distributing $\mu(X)$ in a fair manner between the elements in $X$.
  - Efficiency condition. $\sum_{x \in X} \phi_x = \mu(X)$

- $\mu$ can be non-linear, and include interactions between the variables. So, $\phi$ is linear and removes interactions.

# Explainability: Shapley in XAI

- Shapley values in explainability, main idea
  - $\mu(S)$: Difference with mean output when only variables $S$ are known

Example. Extreme case, nothing is known $\mu(\emptyset) = 0$

# Explainability: Shapley in XAI

- Shapley values in explainability, and partially undefined inputs
  - Recall. Particular input/instance $u$ and model $M$

# Explainability: Shapley in XAI

- Shapley values in explainability, and partially undefined inputs
  - Recall. Particular input/instance $u$ and model $M$
  - Instance $u$ with partial information for only variables $S \subset X$:
    $$u^S$$
    where
    $$u_i^S = u_i \text{ if } x_i \in S$$
    and then, in principle, $u_i^S = \perp$ (undefined) if $x_i \notin S$.
    Example. Something like $u^S = (u_1, u_2, \perp, \perp, u_5, \perp, u_7, u_8)$.

# Explainability: Shapley in XAI

- Shapley values in explainability, and partially undefined inputs
  - Recall. Particular input/instance $u$ and model $M$
  - Instance $u$ with partial information for only variables $S \subset X$:
    $$u^S$$
    where
    $$u_i^S = u_i \text{ if } x_i \in S$$
    and then, in principle, $u_i^S = \perp$ (undefined) if $x_i \notin S$.
    Example. Something like $u^S = (u_1, u_2, \perp, \perp, u_5, \perp, u_7, u_8)$.
  - Most models are numerical and numbers are expected in inputs, then, the mean input is often used for $u_i^S$ when $x_i \notin S$. Mathematically, using $\bar{X}_i$ to represent the mean of variable $x_i$ it is common to use
    $$u_i^S = \bar{X}_i \text{ if } x_i \notin S.$$

# Explainability: Shapley in XAI

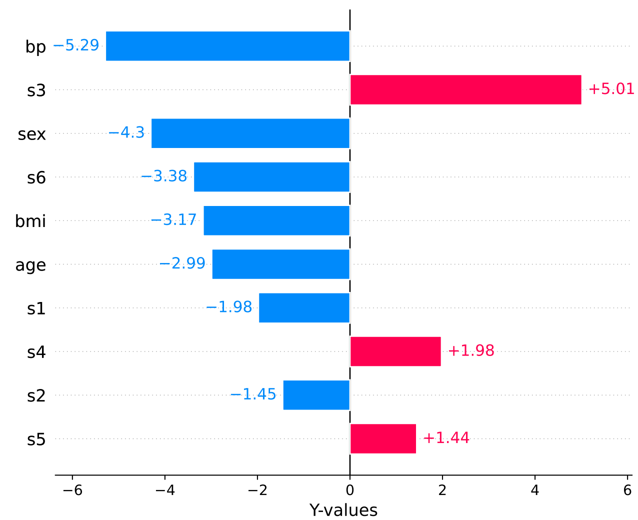
- Shapley values in explainability, and definition of $\mu$
  - Now, the game $\mu$ from $M(u)$ is defined as

    $$\mu(S) = M(u^S) - M(u^{\emptyset}) \tag{1}$$

    for all $S \subseteq X$.
    $\mu(\{x_1, x_2, x_5, x_7, x_8\}) = M(u_1, u_2, \perp, \perp, u_5, \perp, u_7, u_8) - M(\perp, \ldots, \perp)$
  - Note: As we substract $\mu(u^{\emptyset})$ we have $\mu(\emptyset) = 0$. We could just use $\mu(S) = M(u^S)$, then all Shapley values are shifted by a constant.

# Explainability: Shapley in XAI

- Shapley values in explainability, from $\mu$ to $\phi$
  - Given $M(u)$, we compute $\phi_x$ rellevance and importance for $x \in X$
  - So, in this example,




  - ▷ If age = -2.99 means that (in average) adding the variable age to any set of variables decreases the output for this instance in 2.99.
  - ▷ $M(u) = \mu(X) + M(u^{\emptyset}) = \sum_x \phi_x(\mu) + M(u^{\emptyset})$

# Experiments and analysis

# Explainability

- Back to our questions
  - Is explainability still possible for privacy-preserving models?
- Evaluation
  - How data protection affect Shapley values?

# Methodology

# Explainability

- How data protection affect Shapley values?
- Comparison of Shapley values
  - Local vs global: One or a set of Shapley values
    - ▷ Individual comparison of Shapley values
    - ▷ Comparison of mean Shapley values for a set of instances $u$ (test set, global importance)
  - Shapley values or ranks of variables
    - ▷ Compare numerical values (Shapley values themselves)
    - ▷ Compare ranking of variables: Spearman's rank correlation
- So, 4 comparisons

# Explainability

- Local vs global: One or a set of Shapley values (test set $X^{te}$): Case: rank correlation $CORR$.
  - Individual comparison of Shapley values (and their mean)

$$\frac{\sum_{x \in X^{te}} Corr(\phi_{ML_0}(x), \phi_{ML_{\rho_p}}(x))}{|X^{te}|},$$

  - Comparison of mean Shapley values for a set of instances $u$

$$Corr(\bar{\phi}_{ML_0,X^{te}}, \bar{\phi}_{ML_{\rho_p},X^{te}}).$$

where

▷ $\bar{\phi}_{ML_0,X^{te}} = \frac{\sum_{x \in X^{te}} \phi_{ML_0}(x)}{|X^{te}|}$

▷ $\bar{\phi}_{ML_{\rho_p},X^{te}} = \frac{\sum_{x \in \rho_p(X)^{te}} \phi_{ML_{\rho_p}}(x)}{|X^{te}|}$

# Methodology

Masking methods $\rho$ with parameters $p_\rho$.

- Split the data set $X$ in training $X^{tr}$ and testing $X^{te}$
- $ML_o := A(X^{tr})$, the ML model built from original data
- For each $x \in X^{te}$, define game $\mu_{ML_o,x}$.
  Compute Shapley values $\phi_{ML_o}(x)$.

  Compute the mean Shapley value of $X^{te}$: $\bar{\phi}_{ML_o,X^{te}}$.

# Methodology

Masking methods $\rho$ with parameters $p_\rho$.

- Split the data set $X$ in training $X^{tr}$ and testing $X^{te}$
- $ML_o := A(X^{tr})$, the ML model built from original data
- For each $x \in X^{te}$, define game $\mu_{ML_o,x}$.
    Compute Shapley values $\phi_{ML_o}(x)$.

  Compute the mean Shapley value of $X^{te}$: $\bar{\phi}_{ML_o,X^{te}}$.
- $X_{\rho_p} := \rho_p(X^{tr})$ (protected versions using $\rho$ and $p_\rho$)
- $ML_{\rho_p} := A(X_{\rho_p})$, the ML model built from $X_{\rho_p}$
- For each $x \in X^{te}$, define game $\mu_{ML_{\rho_p},x}$
    Compute Shapley values $\phi_{ML_{\rho_p}}(x)$

  Compute the mean Shapley value of $X^{te}$: $\bar{\phi}_{ML_{\rho_p},X^{te}}$.

# Methodology

Masking methods $\rho$ with parameters $p_\rho$.

- Split the data set $X$ in training $X^{tr}$ and testing $X^{te}$
- $ML_o := A(X^{tr})$, the ML model built from original data
- For each $x \in X^{te}$, define game $\mu_{ML_o,x}$.
  Compute Shapley values $\phi_{ML_o}(x)$.
  Compute the mean Shapley value of $X^{te}$: $\bar{\phi}_{ML_o,X^{te}}$.
- $X_{\rho_p} := \rho_p(X^{tr})$ (protected versions using $\rho$ and $p_\rho$)
- $ML_{\rho_p} := A(X_{\rho_p})$, the ML model built from $X_{\rho_p}$
- For each $x \in X^{te}$, define game $\mu_{ML_{\rho_p},x}$
  Compute Shapley values $\phi_{ML_{\rho_p}}(x)$
  Compute the mean Shapley value of $X^{te}$: $\bar{\phi}_{ML_{\rho_p},X^{te}}$.
- Compare the Shapley values (four comparisons)

# Methodology

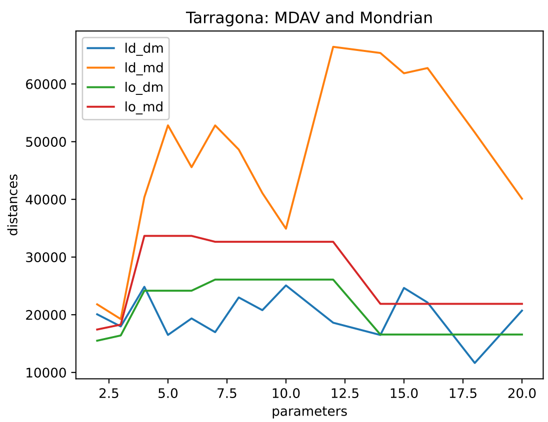
## Methodology

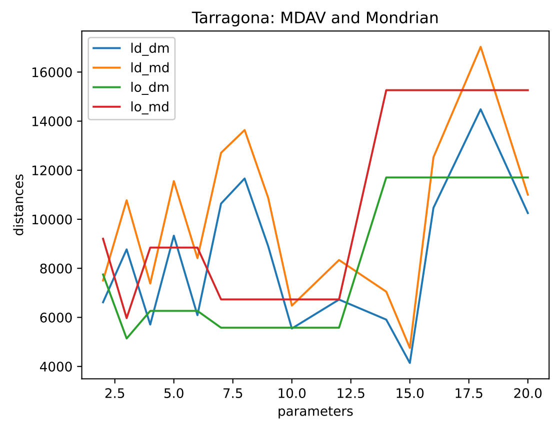
- Data sets: Tarragona (834x12+1), Diabetes (442x10+1), Iris (150x4+1), Cervical cancer (858x35+1), Breast cancer (116x9+1)
- ML algorithms (python `sklearn`):
  - `linear_model.LinearRegression` (linear regression),
  - `sklearn.linear_model.SGDRegressor` (linear model implemented with stochastic gradient descent),
  - `sklearn.kernel_ridge.KernelRidge` (linear least squares with l2-norm regularization, with the kernel trick),
  - `sklearn.svm.SVR` (Epsilon-Support Vector Regression).
- Masking methods
  - Microaggregation (MDAV, Mondrian)
  - Noise addition (Gaussian, Laplacian)
  - Lossy compression/transform-based methods (SVD, PCA, NMF)
- Explainability (own implementation + SHAP for num. vars $> 10$)

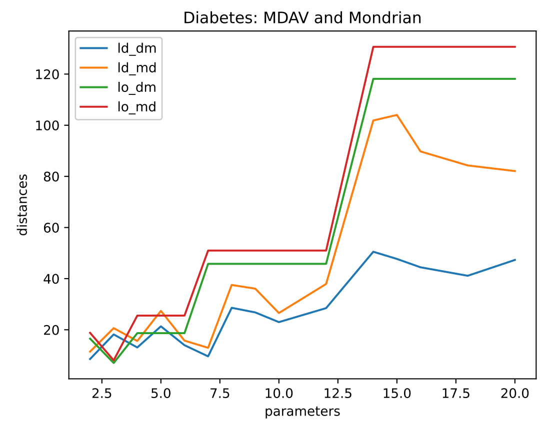
# Analysis

# Analysis (I)

- Distances can be very large, comparisons cumbersome.
  - The game, defined for ML is unbounded (arbitrarily large)
  - $Small$ changes on the model affect the game.



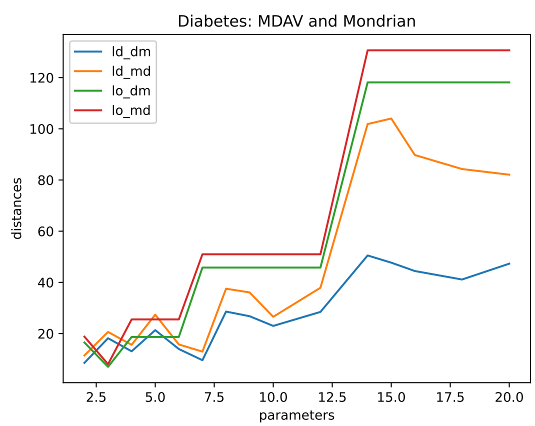



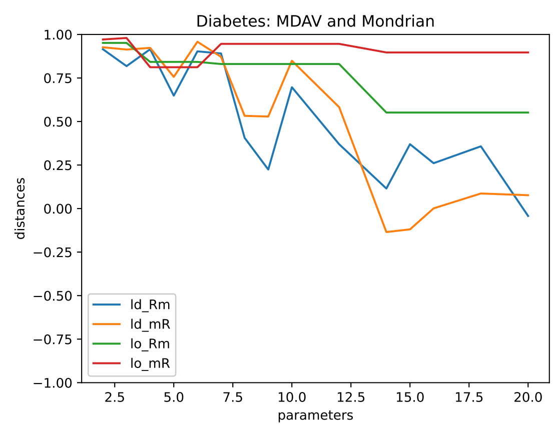




_dm: mean Shapley values, _md: mean distance of Shapley values. d: mdav, o:mondrian. Linear regression. DB: 11 and 12 inputs (left, middle)

# Analysis (II)

- In contrast, rank correlation is always in [-1,1].
  - Larger distances do not mean larger rank correlation. Large distances between Shapley values do not imply changes in values order.
  - Mondrian give larger distances than MDAV, but MDAV shows a worse performance as Mondrian has a rank correlation near to 1 for larger parameters.

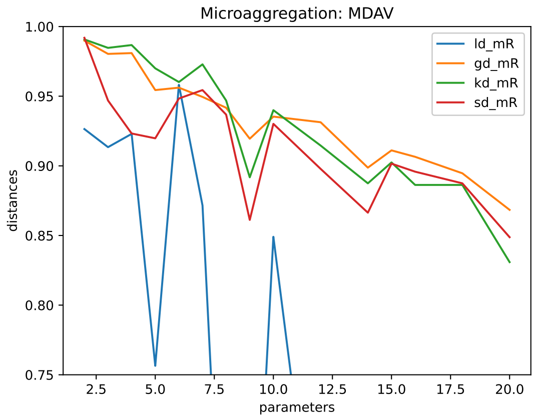


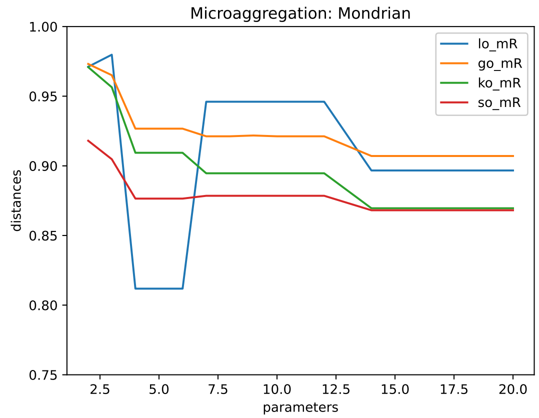




_dm: mean Shapley values, _md: mean distance of Shapley values, _Rm: Rank correlation of mean Shapley values, _mR: mean correlation of Shapley values. d: mdav, o:mondrian. Linear regression.

# Analysis (III)

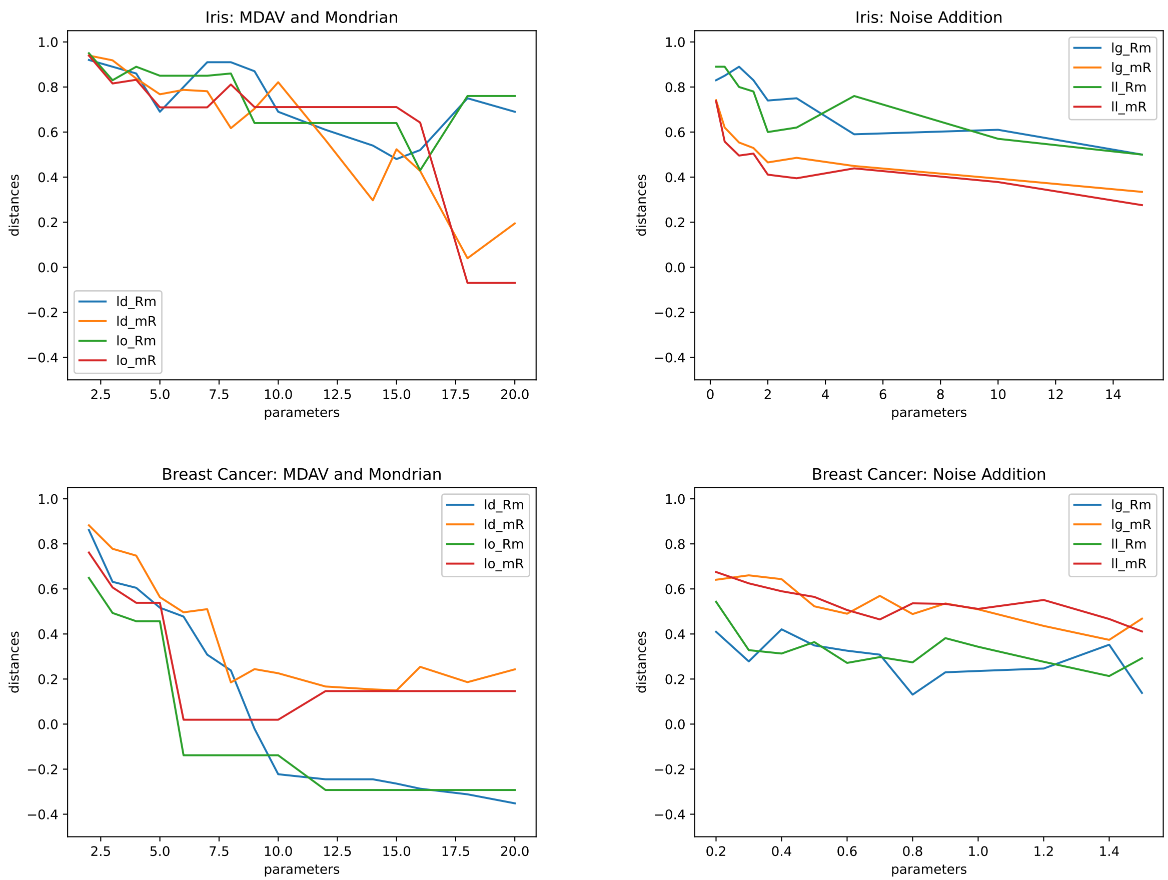
- For rank correlation, similar tendency results independent of ML. Mean rank correlation for MDAV and Mondrian



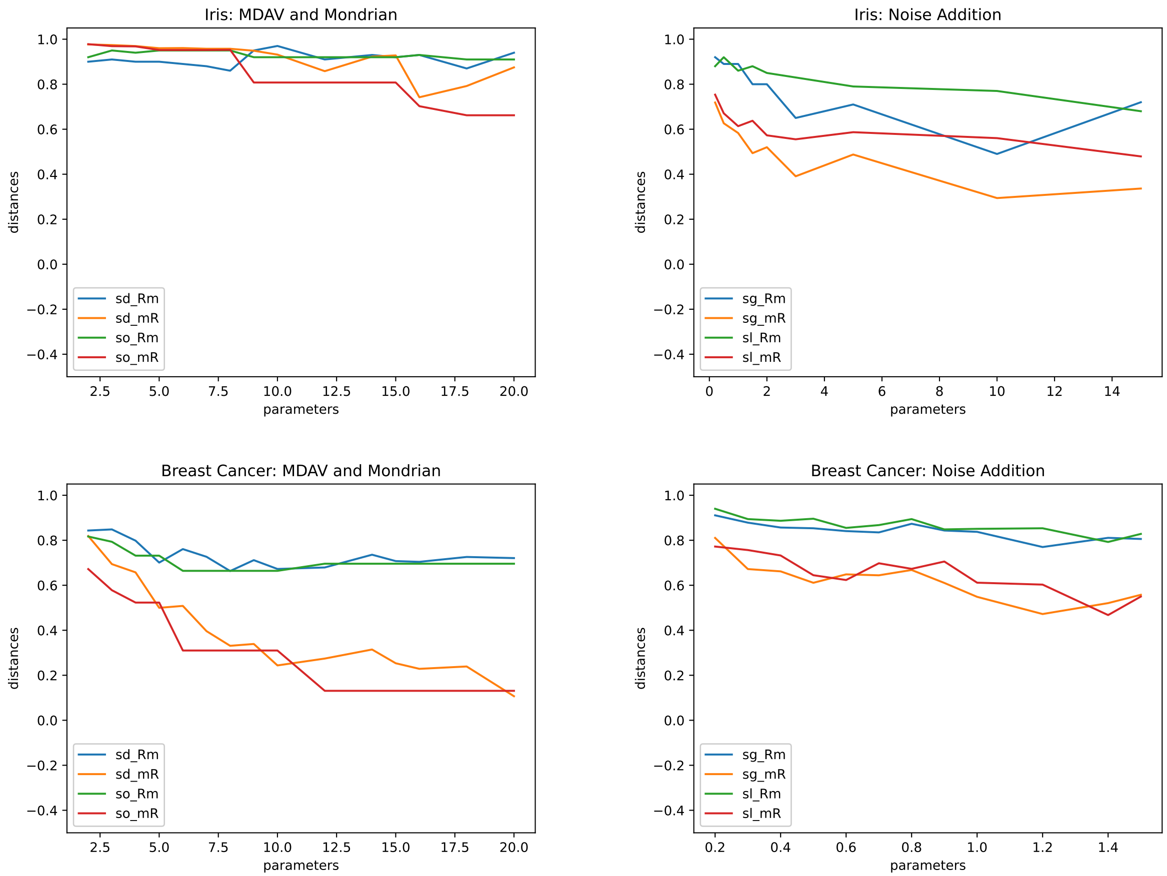
# Analysis (IVa)

- Seems, microaggregation leads to better results than noise addition.
  - Linear regression. $d$: MDAV, $o$: Mondrian, $g$: Gaussian, $l$: Laplace



# Analysis (IVb)

- SVM-regression. $d$: MDAV, $o$: Mondrian, $g$: Gaussian, $l$: Laplace



# Analysis (IVa)

- Seems, microaggregation leads to better results than noise addition.
  - This is also supported by privacy protection level.
  - For $k = 1.5$, from $\epsilon$-LDP-perspective we have $\epsilon$ values of
    - Breast cancer: $\epsilon = (4.56, 1.48, 10.38, 2.90, 1.48, 6.26, 2.68, 3.71, 121.77)$
    - Iris: $\epsilon = (4.94, 3.02, 8.10, 3.29)$
  - For $k = 15$
    - Iris: $\epsilon = (1.56, 0.95, 2.56, 1.04)$

# Analysis (V)

- Summary.
  - Protection does not prevent explainability (Shapley values). Not incompatible
  - Results based on rank correlation have a sounder behavior change more smoothly w.r.t. protection, similar behavior for diff. ML
  - Among the four machine learning models, the linear model is the one that has the worst performance with respect to the Shapley value.
  - Microaggregation ($k$-anonymity) seems better

# Future Directions

# Work in progress

- Research directions related to Shapley values
  - Games are set functions, and information on the model is rich e.g. interactions
  - Shapley values are just summaries
  - We need to further exploit the game

# Work in progress

- Exploiting the game[3]
  - Interactions. E.g., $I(age, sex)$? (interaction index)
  - Other indices. E.g., $\Upsilon$-values[4]
  - Not all coalitions are possible. E.g., either we know both variables $x_1$ and $x_2$, or we know none.
  - The game itself. $\mu(S) = M(u^S) - M(u^{\emptyset})$

[3] V. Torra, Games, fuzzy measures, indices, and explainable ML: exploiting the game, INFUS 2024

[4] V. Torra (2024) $\Upsilon$-values: power indices Ã**?** la orness for non-additive measures, IEEETFS.

# Thank you