Who's responsible? Jointly quantifying the contribution of data and algorithm
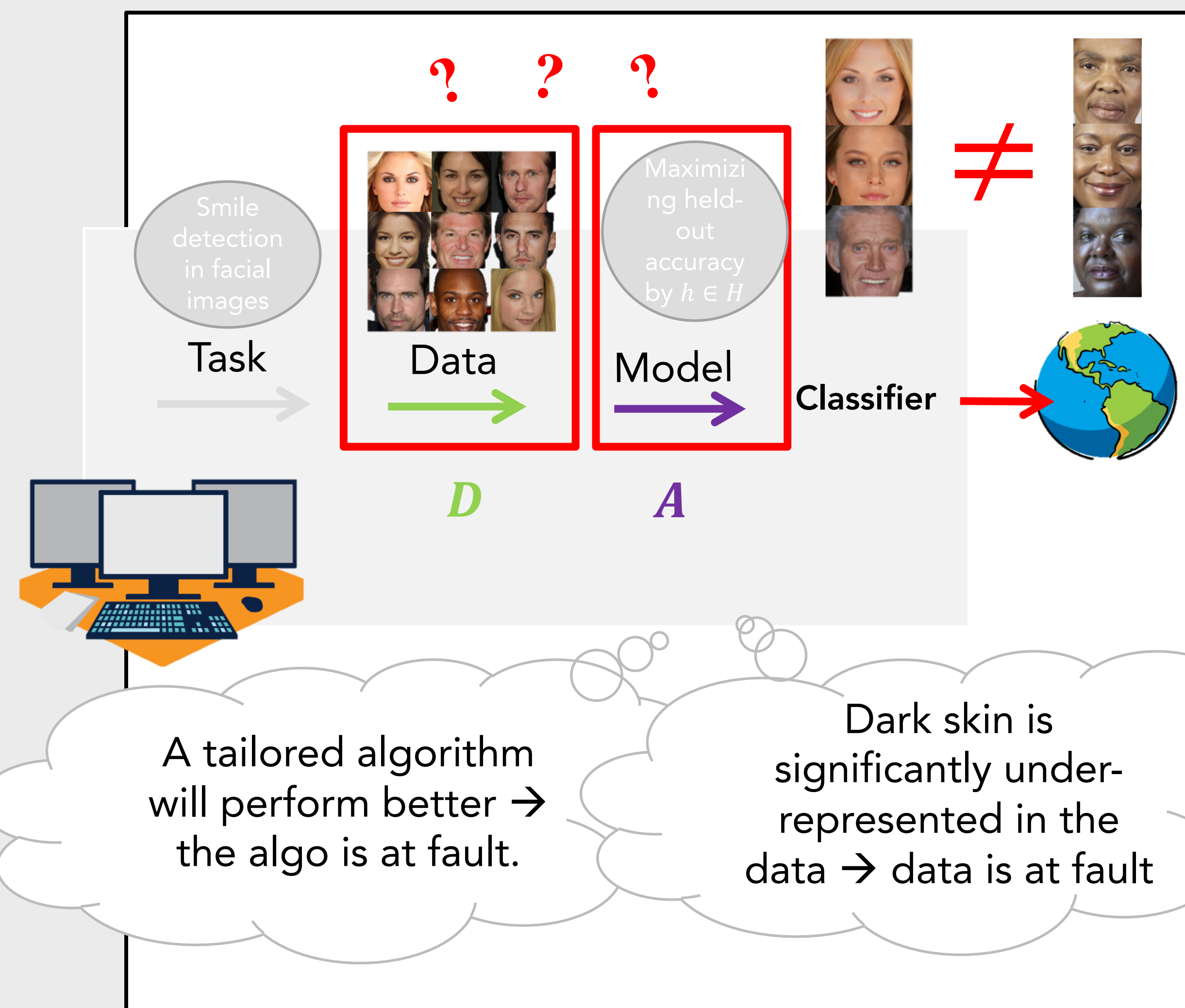
Gal Yona, Amirata Ghorbani & James Zou
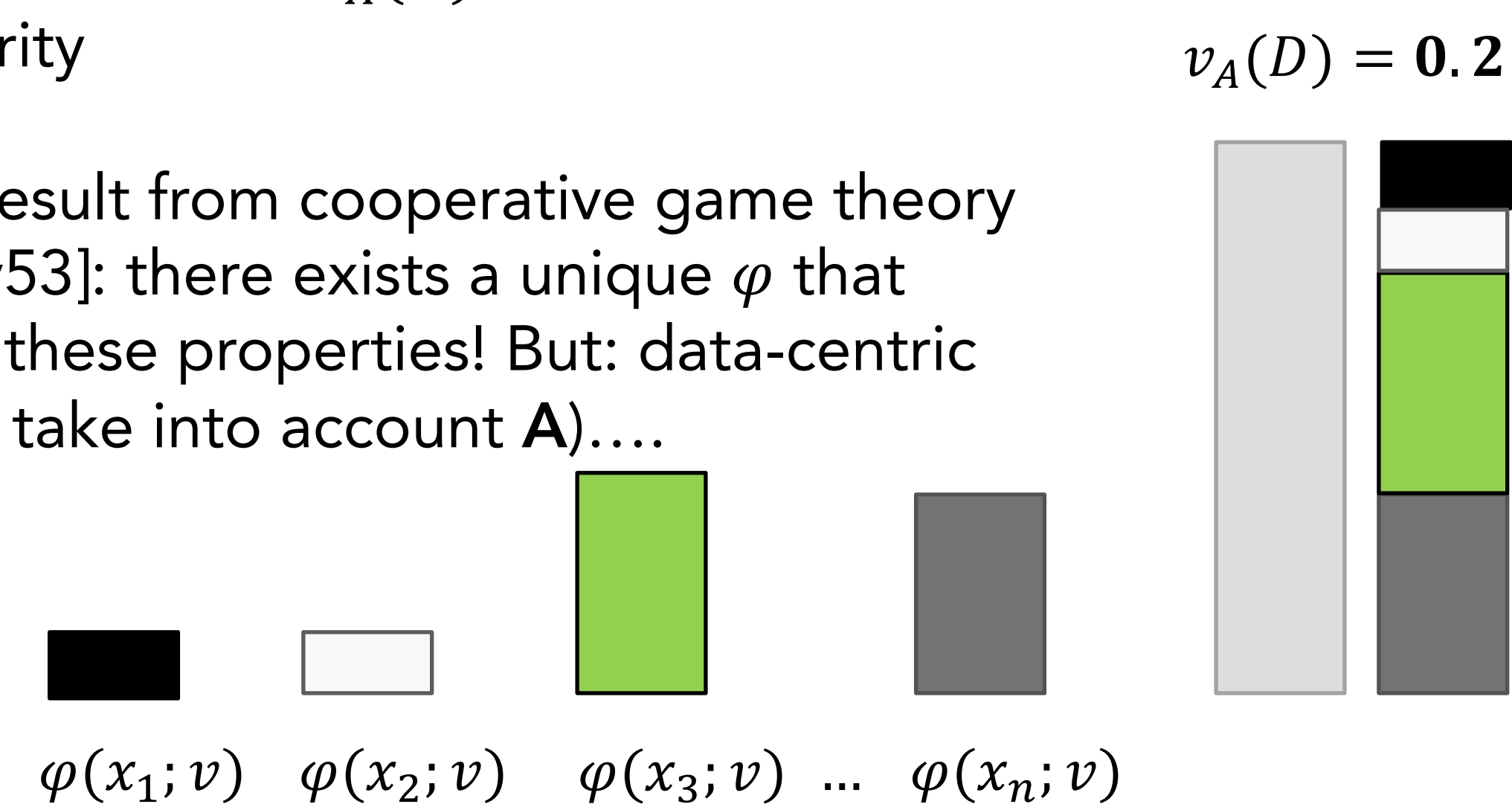
Task → Data D → Model A → Classifier → 🌍

$\neq$

A tailored algorithm will perform better → the algo is at fault.

Dark skin is significantly under-represented in the data → data is at fault

Q: recognizing the interaction between existing biases in data and different (potentially subtle) modeling choices, can we disentangle their effect on the overall performance?

Joint data-algorithm valuation problem

A: A reduction to the data valuation problem (recently studied e.g. in [GZ19], [ADS19], [JDW+19])
Specifically the approach in [GZ19]

acc = 0.95   acc = 0.75   = 0.2

Dataset $D = \{x_1, \ldots, x_n\}$

Maximizing held-out accuracy by $h \in H$

Algorithm **A**   Evaluation criteria v

How to "distribute" $v_A(D)$ among the individual data points $x_1, \ldots, x_n$?

Data Shapley [GZ19]

Specifies *four natural conditions* for an equitable data valuation $\varphi: \{x_1, \ldots, x_n\} \to \mathbb{R}^n$:
1. Null player receives *zero value*
2. Symmetric players receives *equal value*
3. Sum of values is $v_A(D)$
4. Linearity

Classic result from cooperative game theory [Shapley53]: there exists a unique φ that satisfies these properties! But: data-centric (doesn't take into account **A**)....

$v_A(D) = \mathbf{0.2}$

$\varphi(x_1; v)$ $\varphi(x_2; v)$ $\varphi(x_3; v)$ ... $\varphi(x_n; v)$

This work: Extended Shapley [YGZ19]

Fix a benchmark algorithm B and add the algorithm **A** as an "additional" $n + 1$ player

Specify *five natural conditions* for an equitable data-algorithm valuation $\varphi: \{A, x_1, \ldots, x_n\} \to \mathbb{R}^{n+1}$:
1. Null datum receives *zero value*; if A is identical to B, algo receives *zero value*
2. Symmetric players receives *equal value*
3. Sum of values is $v_A(D)$
4. Linearity

Existence of a unique solution extends! These are the Extended Shapley values.

$v_A(D) = \mathbf{0.2}$

$\varphi(A)$ $\varphi(x_1)$ $\varphi(x_2)$ $\varphi(x_3)$ ... $\varphi(x_n)$
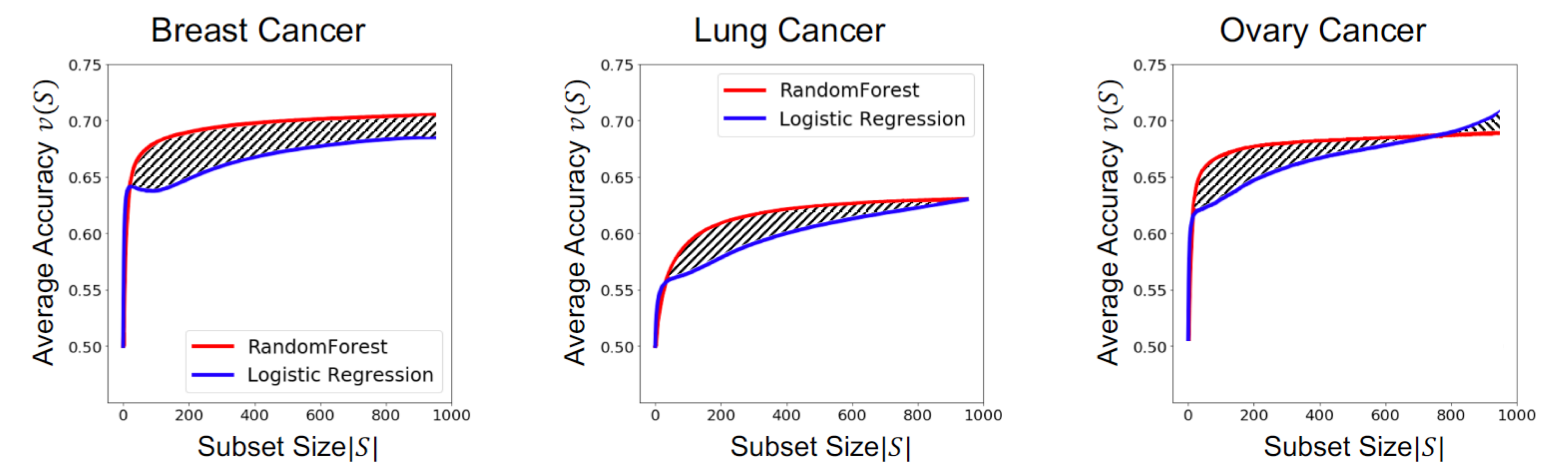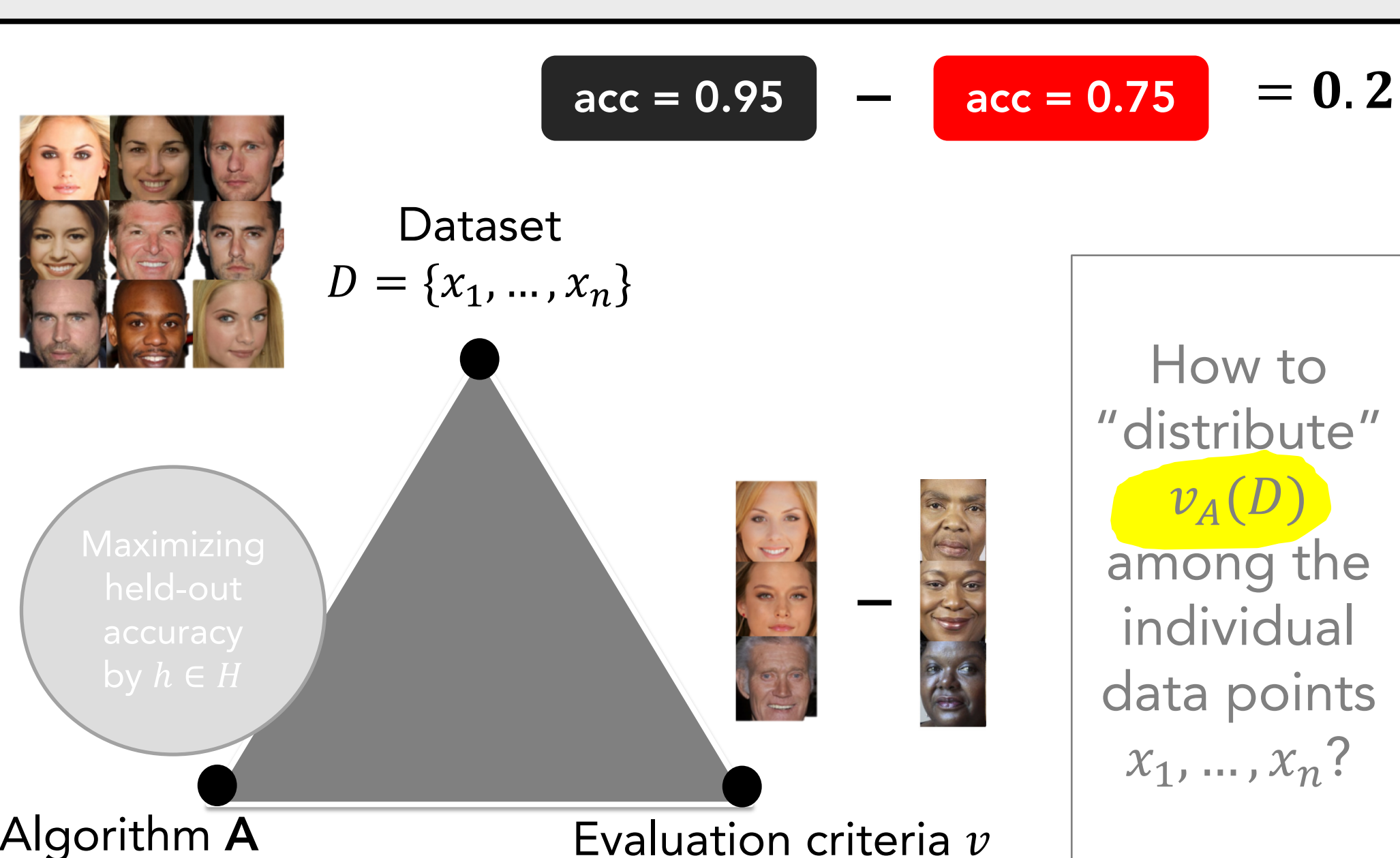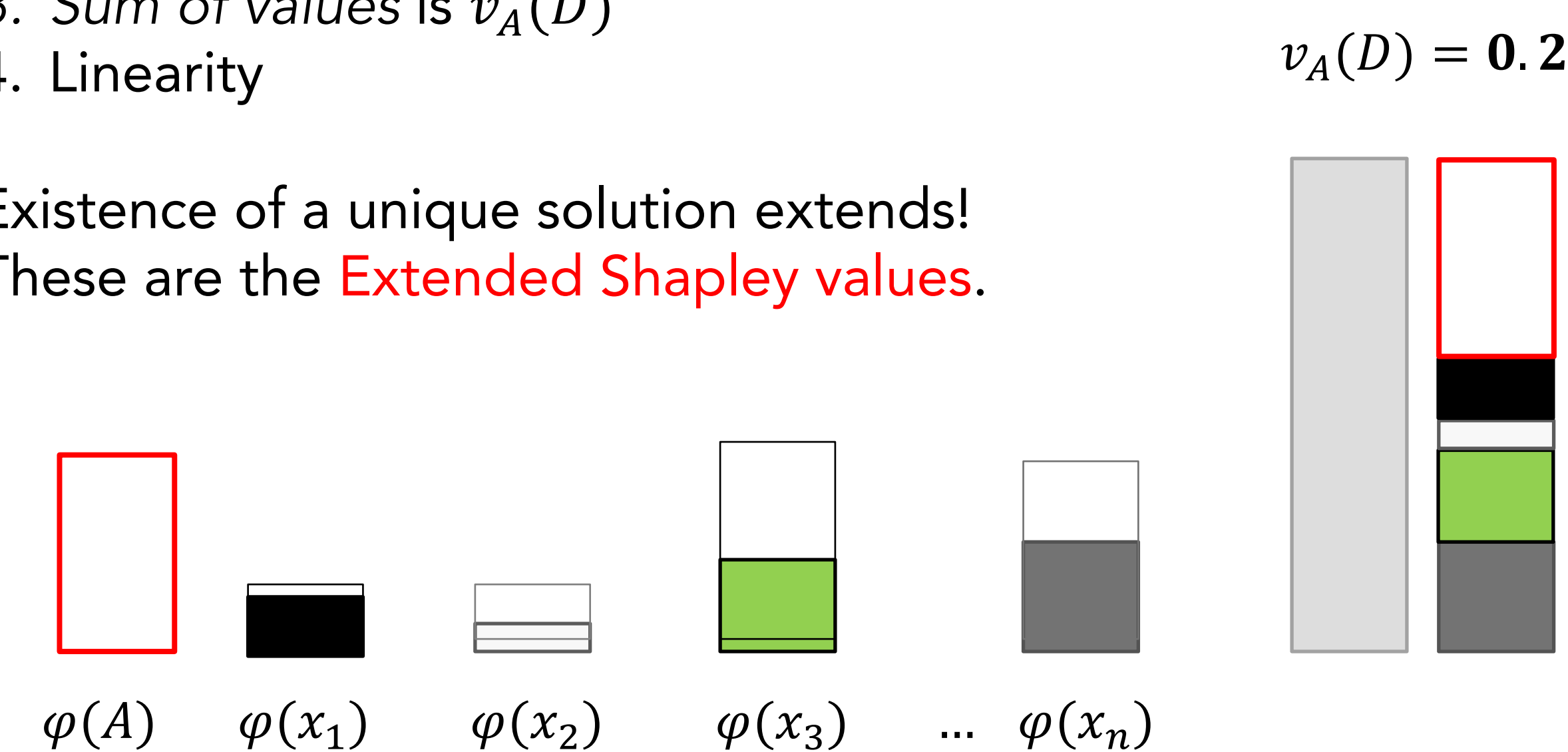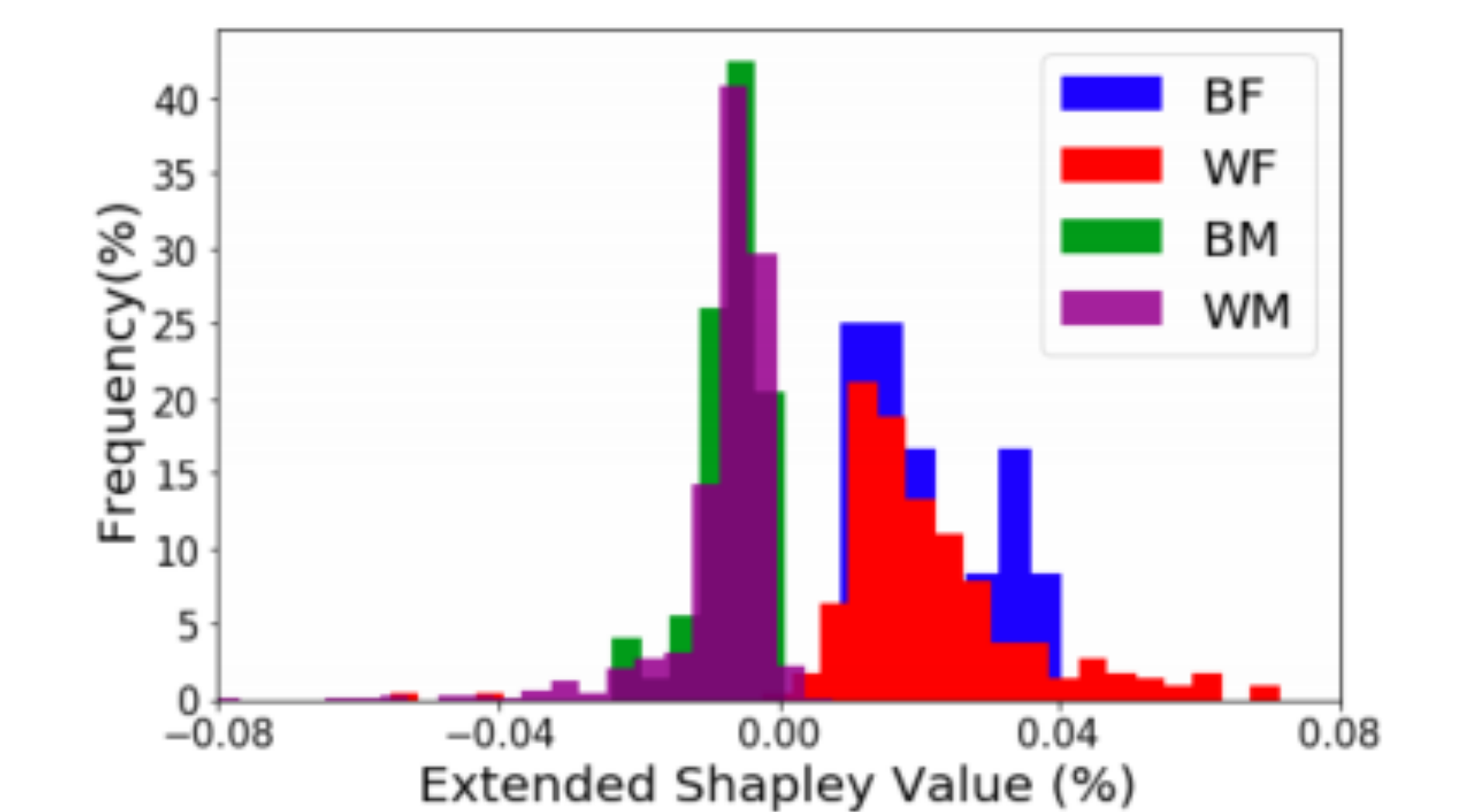
A practical framework!

#1: Closed-form expressions for the value of data and algorithm. E.g. value of algo is $E_S[v_A(S) - v_B(S)]$

Intuition:
If **A** is consistently doing worse off than baseline **B** (irrespective of the specific training set they are trained on), then **A** "deserves" a large portion of the blame.
If very specific combinations of datapoints contribute to A's failure, then these carry more of the "blame".

#2: Can re-use heuristics for efficient computation from [GZ19].

#1: Understanding Extended Shapley as a new performance metric

Breast Cancer   Lung Cancer   Ovary Cancer

Average Accuracy $v(S)$ — Subset Size $|S|$
RandomForest / Logistic Regression

#2: Allocating responsibility for unfairness

- Training data: 1000 images from LFW+A dataset (imbalanced: 21% female, 5% black)

- Performance measure: maximal accuracy gap among groups {WM, WF, BM, BF} on the balanced PPB dataset

- Algorithm **A**: Logistic regression applied to 128-dimensional feature vectors obtained by passing the images through a ResNet-V1 pre-trained on CelebA.

- Benchmark **B**: a constant classifier (perfectly fair: $v_B(D) = 0$)

$v_A(D) = 22.9$ (WM-BF) $\varphi_A = 22.1$

Frequency(%) — Extended Shapley Value (%)
BF / WF / BM / WM

Arxiv: https://arxiv.org/abs/1910.04214.   **Contact:** gal.yona@gmail.com