On the Gold Standard

Uncertainty in Computer Science

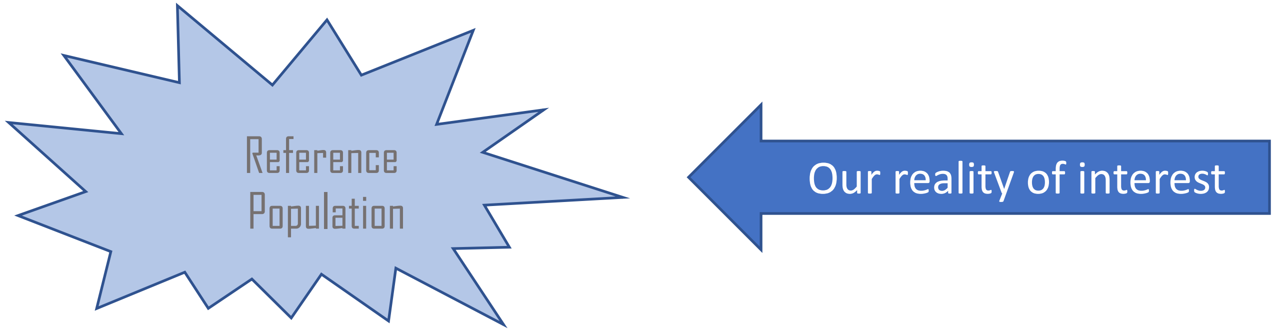
PhD course, 2021/22

Problem: how much *golden* is the gold standard?

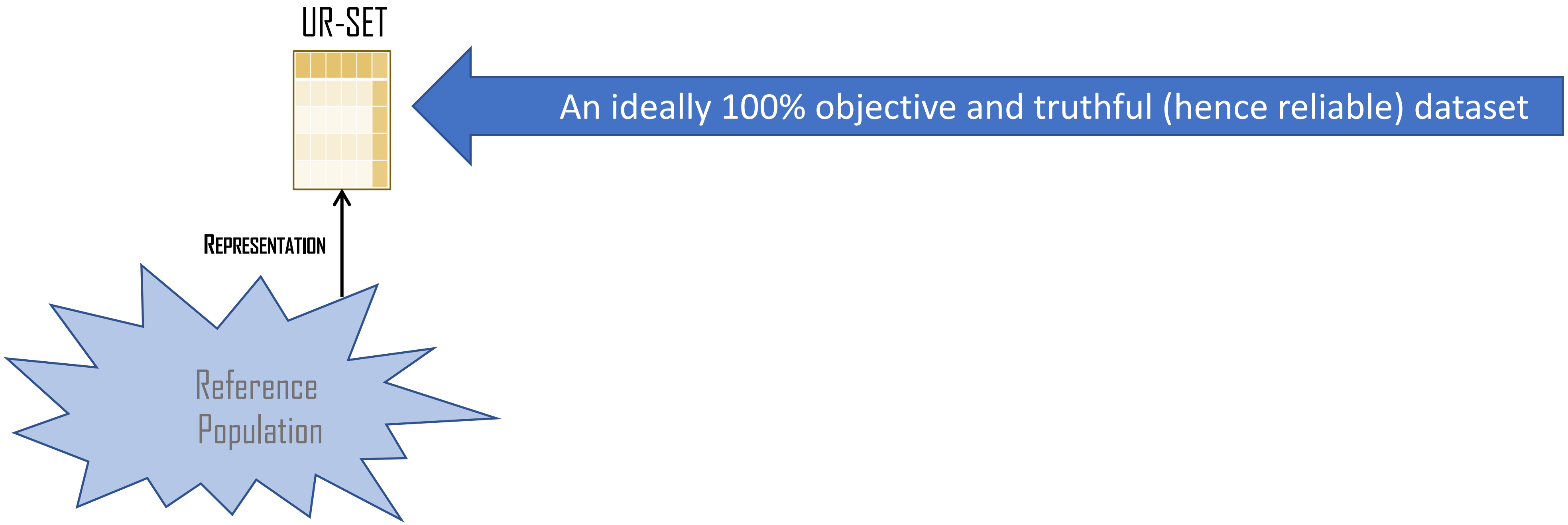
Being able to address questions like these:

1) How much reliable is the ground truth?
2) How much representative is the training wrt reference population?
3) How many annotators do we need?
4) …

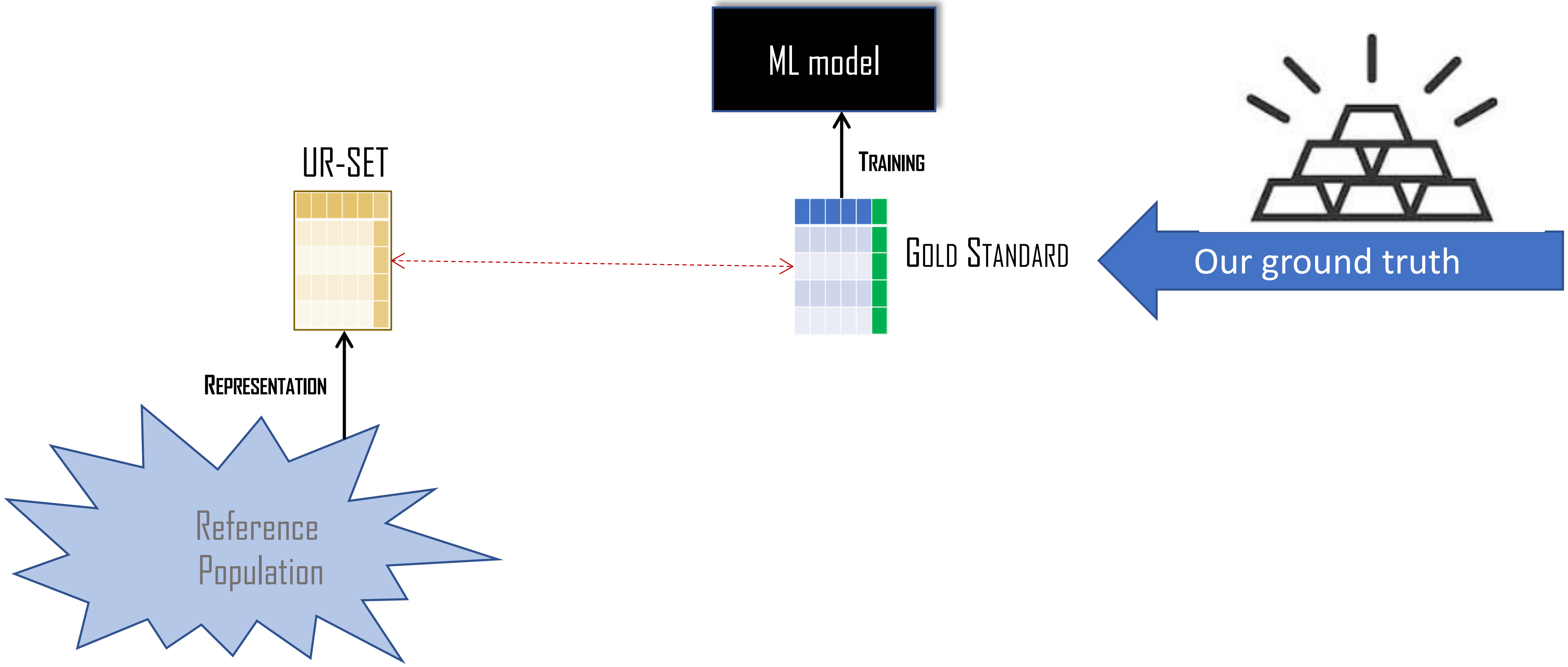
A provocation: how much objective is your dataset?



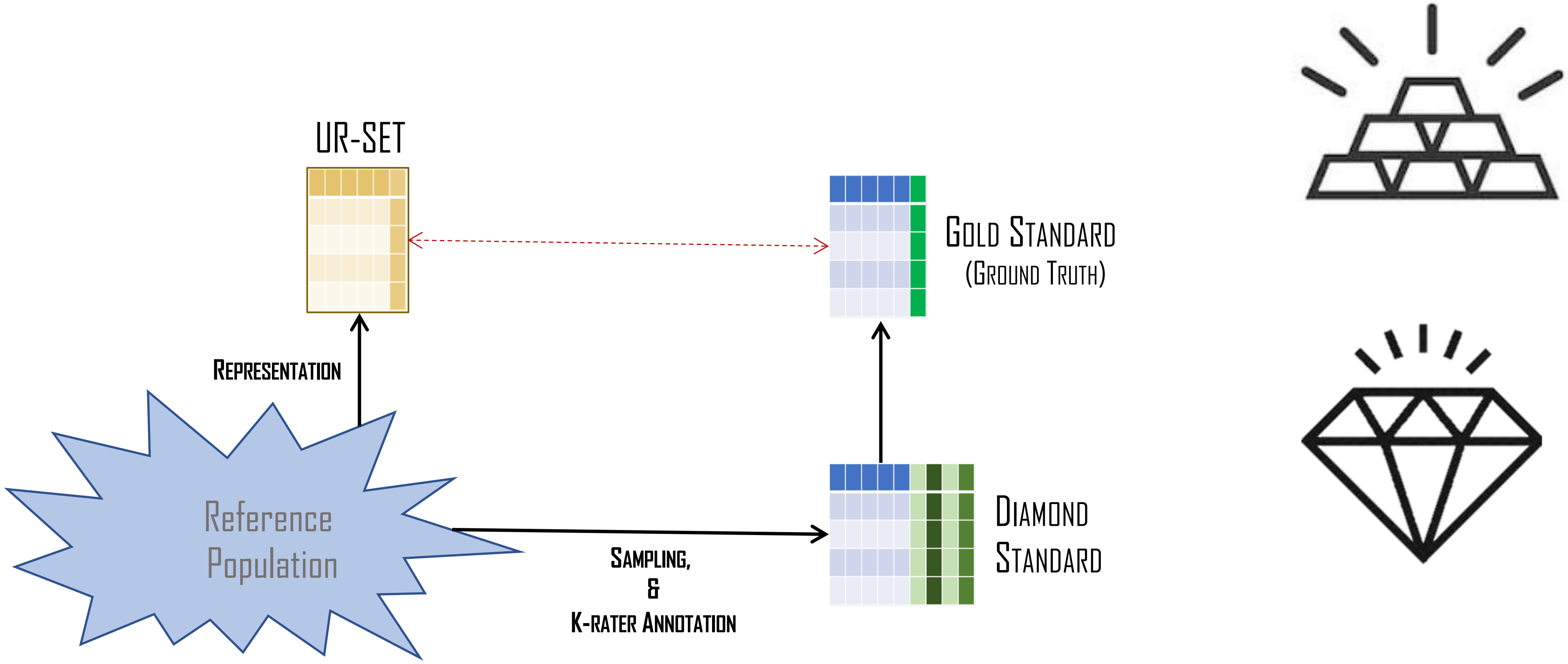
OUR FRAMEWORK



OUR FRAMEWORK

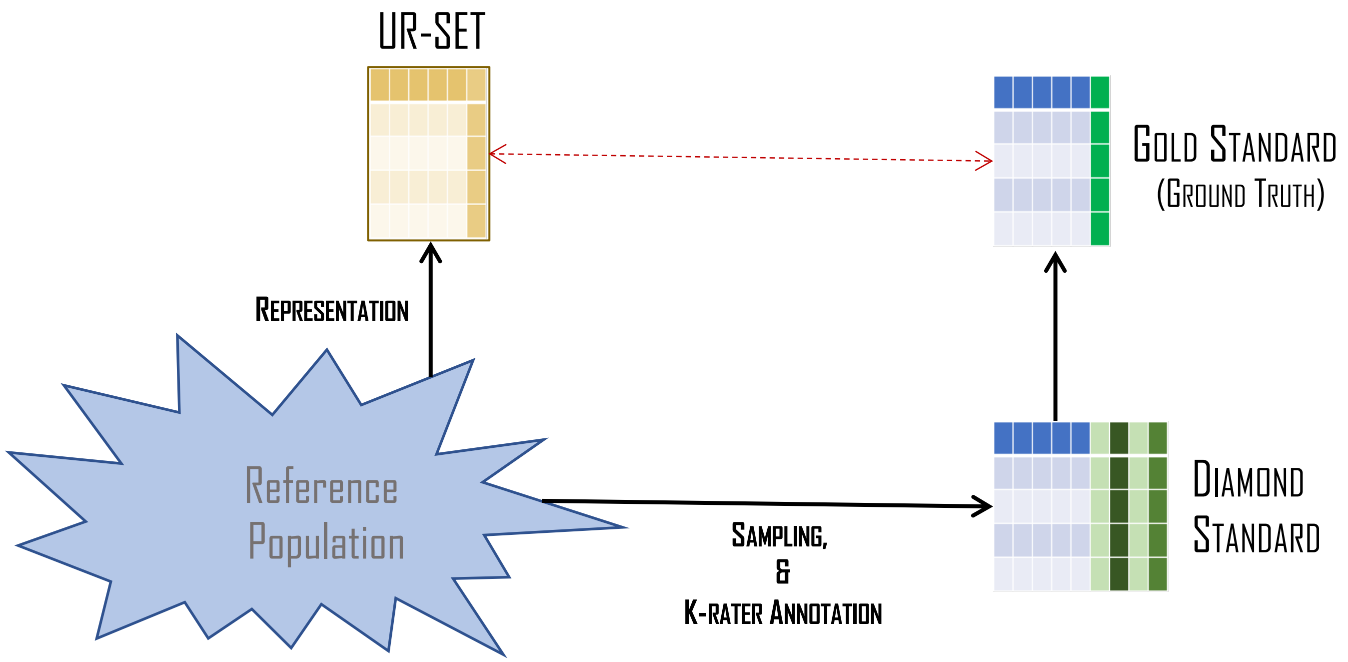


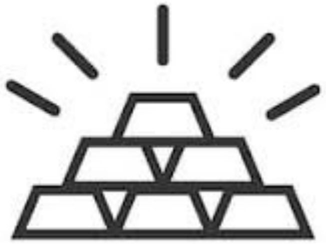
Our framework



Our framework

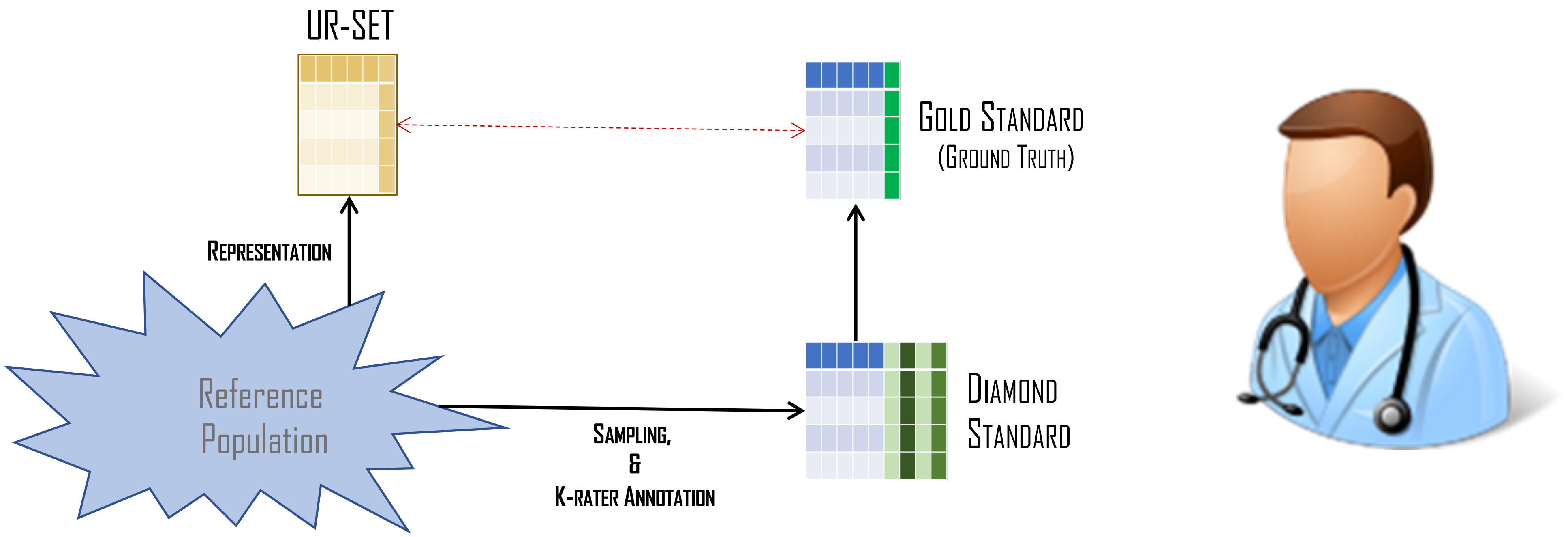
How to transform Diamond (multi-rater labels) into Gold (reliable target labels)?



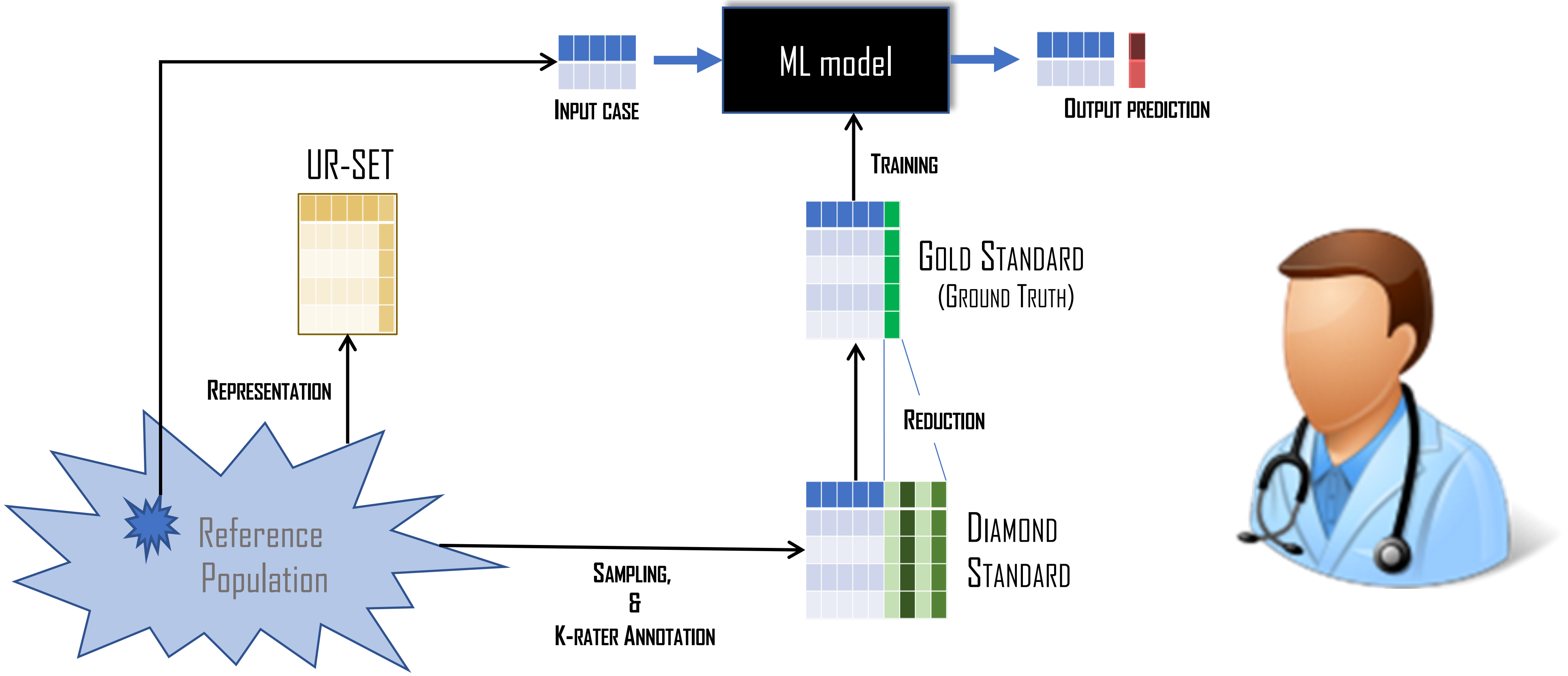


Our framework

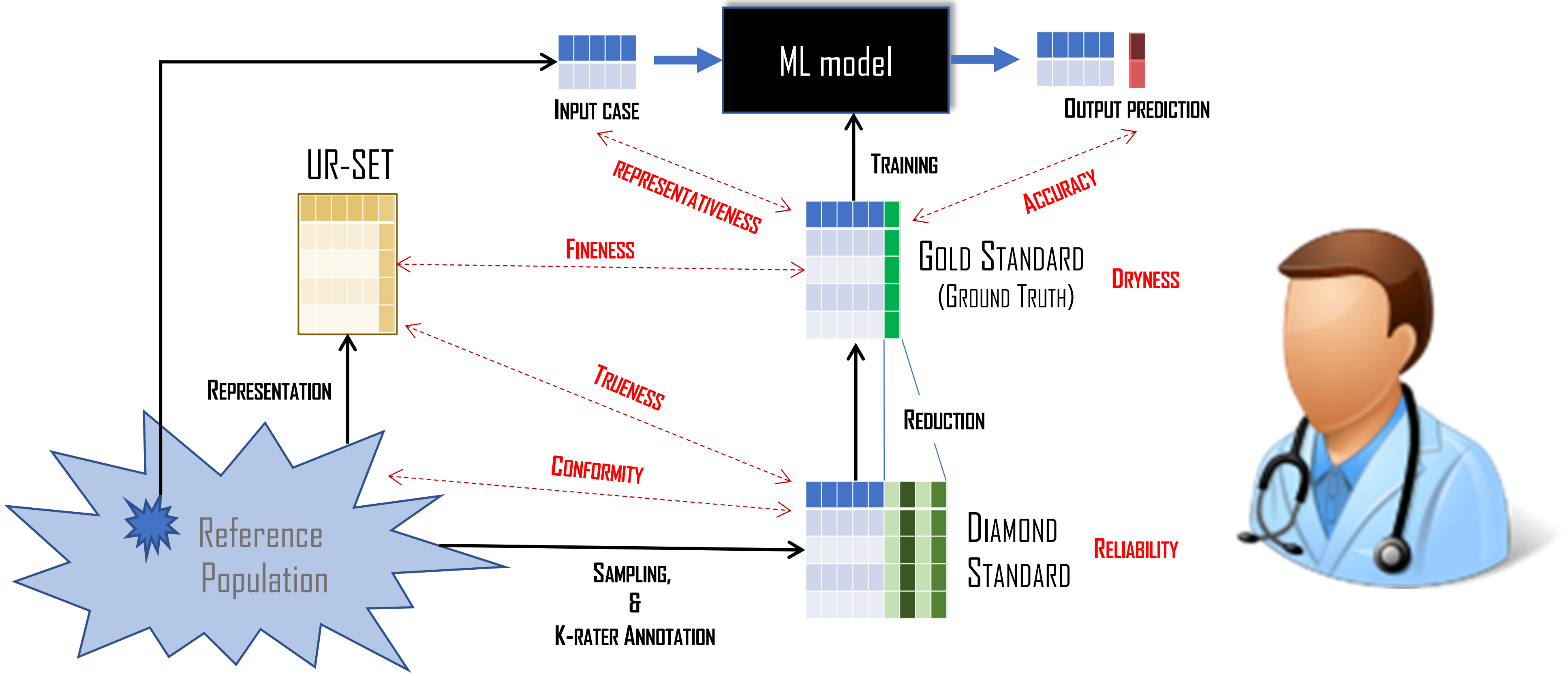
A framework (incl. quality dimensions, assessment methods & improvement methods) to have decision makers become more aware of how much objective/reliable **input** data is, and help them put **output** in context (i.e., interpret it).



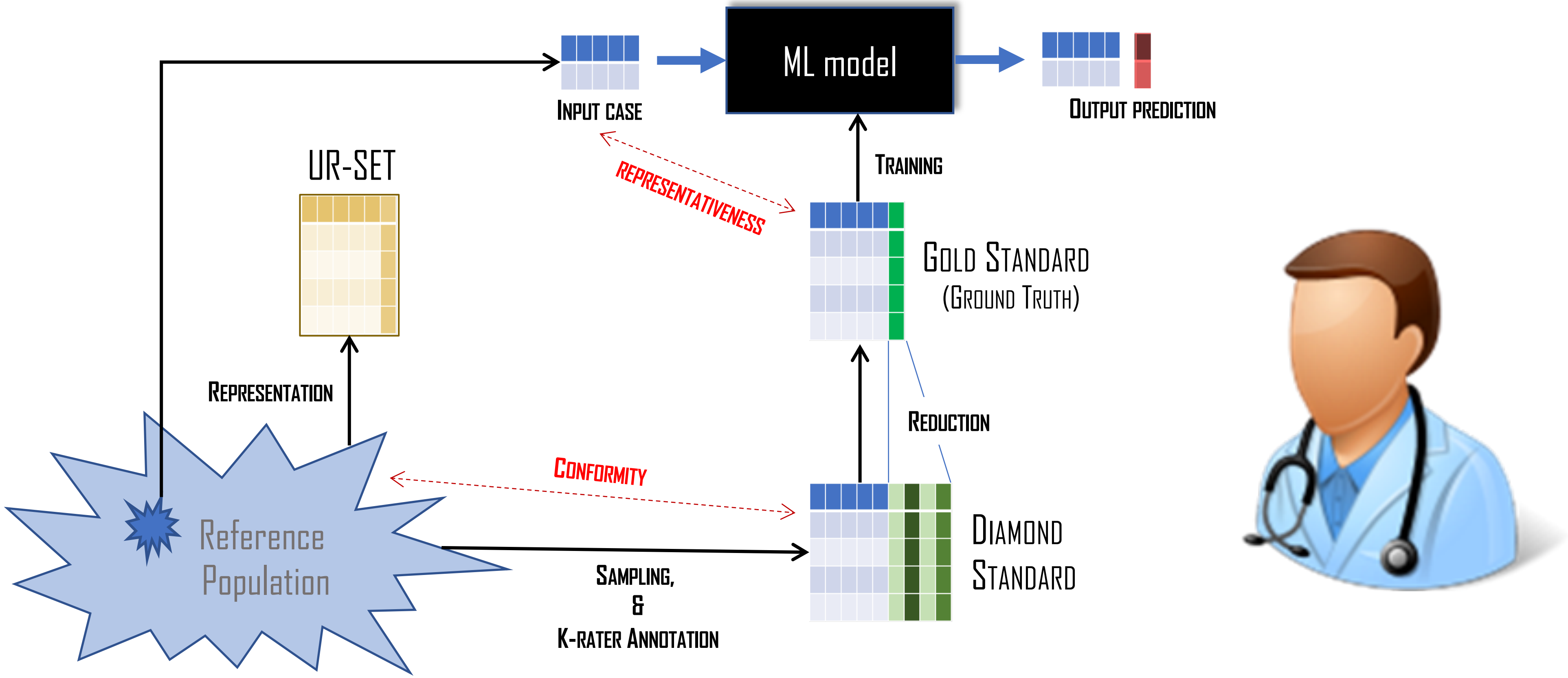
Our framework



Our framework



Our framework

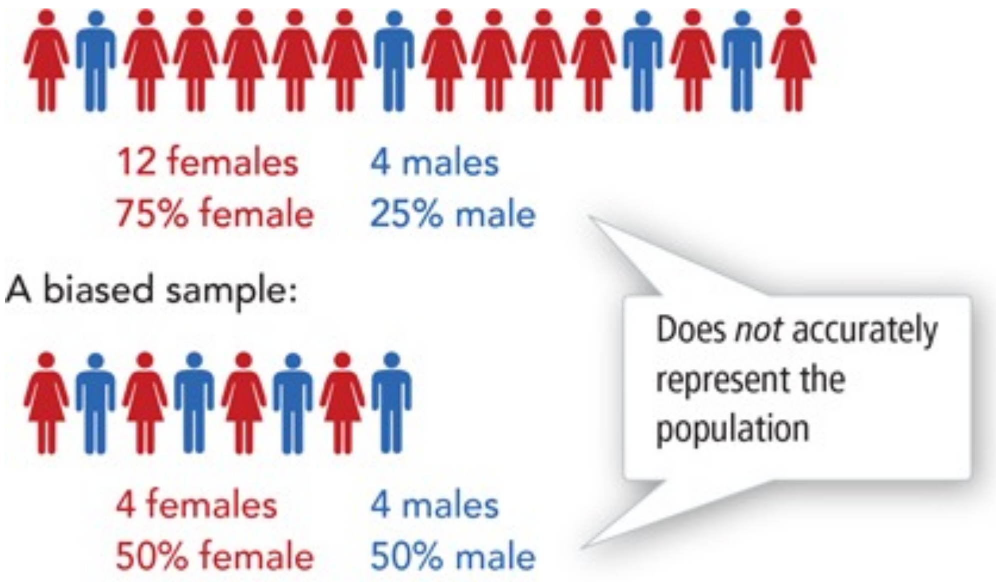


Representative for a group/individual

CONFORMITY

The sample (Diamond Standard) *conforms to* the real population?

- Can be assessed if we have *metadata* about the real population distributions of features (e.g. census data)
- Most simple approach uses standard *goodness-of-fit* tests (e.g. Kolmogorov-Smirnov) w.r.t. the univariate or multivariate distributions.



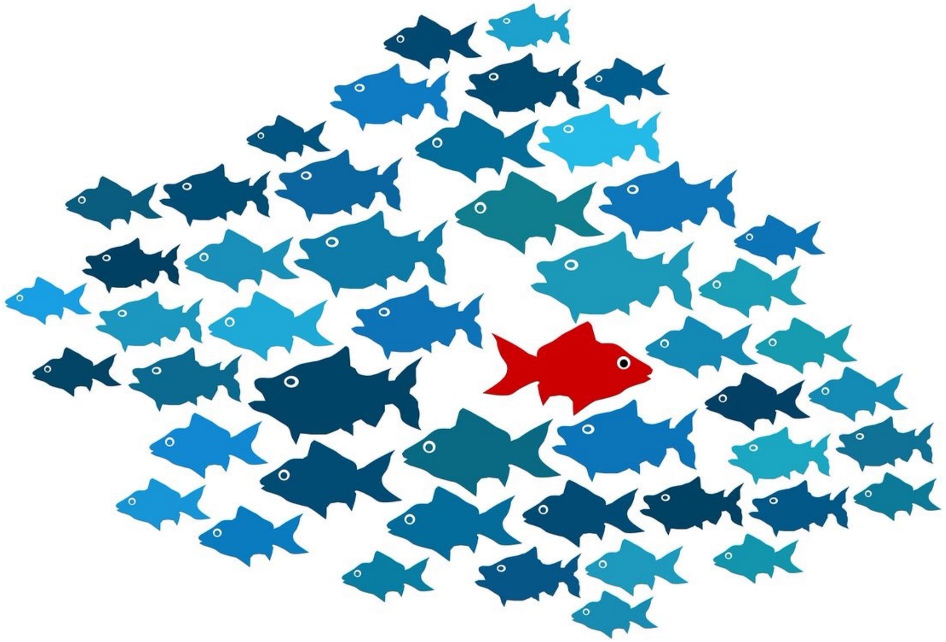
DIMENSIONS

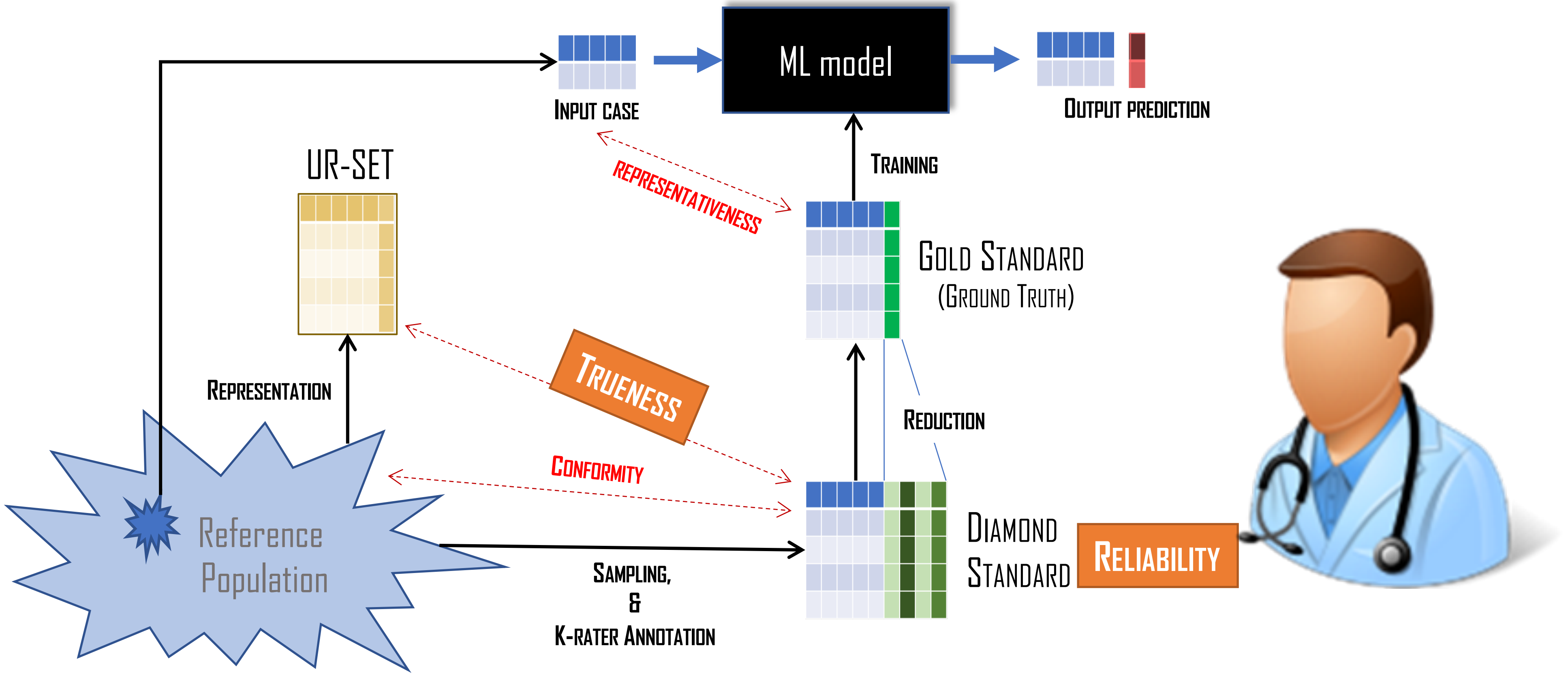
REPRESENTATIVENESS

Is the gold standard representative of a new instance x?

- Naive approach: compare the new instance with the centroid of the training set
  [does not take into account the whole distribution]

- More robust techniques inspired by outlier-detection algorithms
  [probability of obtaining form the Gold Standard a point similar to x]





ML model
Input case
Output prediction
Training
Representativeness
UR-SET
Gold Standard
(Ground Truth)
Trueness
Representation
Reduction
Conformity
Reference
Population
Diamond
Standard
Reliability
Sampling,
&
K-rater Annotation

FRAMEWORK

RELIABILITY

How much the raters offer a *unitary view*? How much do they agree?

Despite being an important dimension to understand how much can we trust our data, it is not widely reported, even in popular ML studies!

RELIABLE NOT VALID	NOT RELIABLE VALID	NOT RELIABLE NOT VALID	RELIABLE VALID

- ***The naive measure (proportion of matched pairs) is problematic: no chance effects!***
- ***The most well-used alternative Fleiss' Kappa is considered by experts as similarly affected by methodological issues (arbitrary threshold, poor chance model, …)***

Cohen's Kappa	*Degree of Agreement*
< 0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

Source: Landis & Koch, 1977.

DIMENSIONS

Krippendorff's Alpha

- robust and more realistic modeling of chance effect
- suitability also for non-nominal data (ordinal, numeric, ...) and missing data
- robust acceptability criteria
- ... also widely implemented software-wise

$$\alpha_{\text{metric}} = 1 - \frac{D_o}{D_e}$$

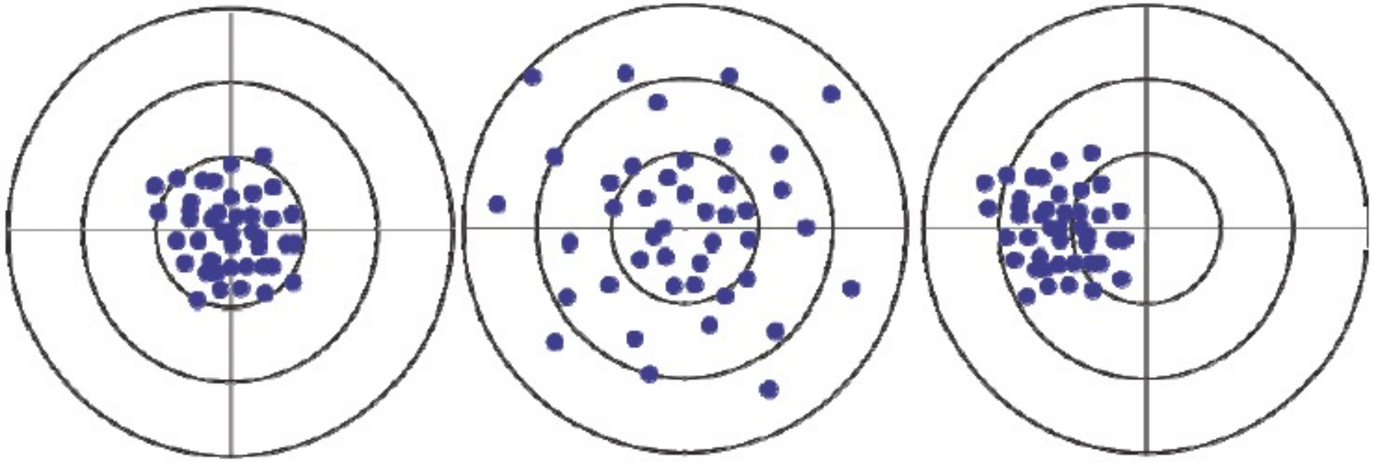
$$= 1 - \frac{\sum_{u=1}^{u=r} \frac{m_u}{n} \sum_{i=1}^{i=m} \sum_{j=1}^{j=m} \frac{{}_{\text{metric}}\delta^2_{c_{iu}k_{ju}}}{m_u(m_u - 1)}}{\sum_{c=1}^{c=n} \sum_{k=1}^{k=n} \frac{{}_{\text{metric}}\delta^2_{ck}}{n(n-1)}}.$$

DIMENSIONS

TRUENESS

Probability that an instance's multi-rater label is the true/correct one?

Similarly to reliability, should be maximized when all raters agree: together with reliability can be considered as a proxy for objectivity of the dataset.



High trueness
High precision

High trueness
Low precision

Low trueness
High precision

ACCEPTABLE TRUENESS

- Assumption: raters err independently
- The *most probably correct* label is the majority one
- its observed proportion is an estimate of the real success rate
- *acceptable trueness* if inf(trueness(o(x))>k

$$trueness'_c(\mathbf{o}(x)) = p \pm 1.96\sqrt{\frac{p(1-p)}{m}}$$

DIMENSIONS

DISAGREEMENT TRUENESS

Information-theoretic definition
Od number of disagreement
Md maximum number of possible disagreement
ε **smoothing factor (if Od=0, then trueness is not 1)**

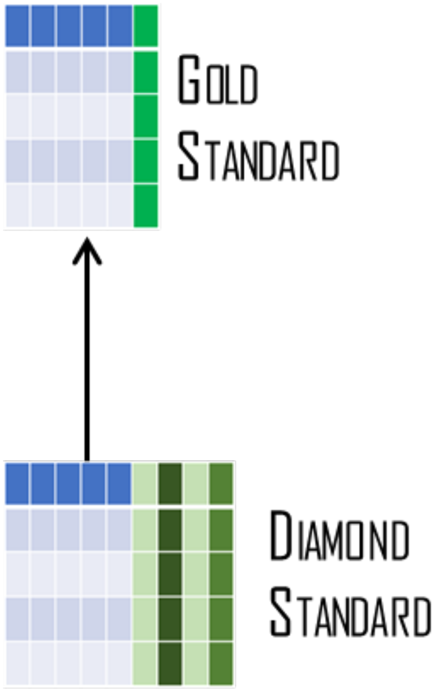
$$trueness''_c(\mathbf{o}(x)) = 1 - \frac{O_d + \epsilon}{M_d + \epsilon}$$

DIMENSIONS

DRYNESS

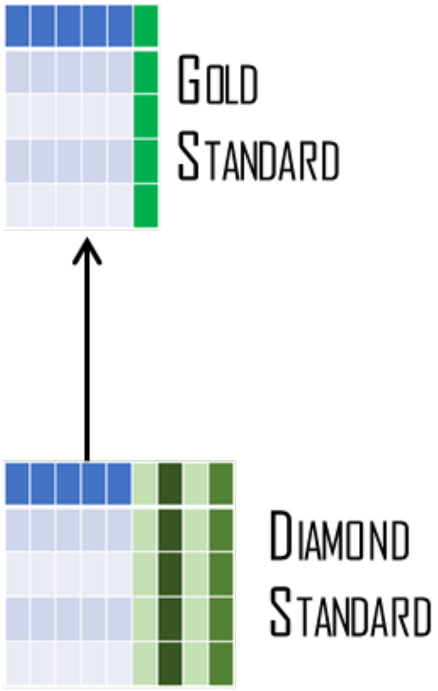
Going from the Diamond Standard to the Gold Standard involves an information aggregation (*reduction*) that leads to information loss

Standard approach: take majority label...
Is this warranted when the margin is small?



DRYNESS

On high-uncertainty instances we could employ more sophisticated *reduction rules*, inspired by the ensemble learning and uncertainty representation (*fuzzy sets, three-way decision, probability theory*) literatures.

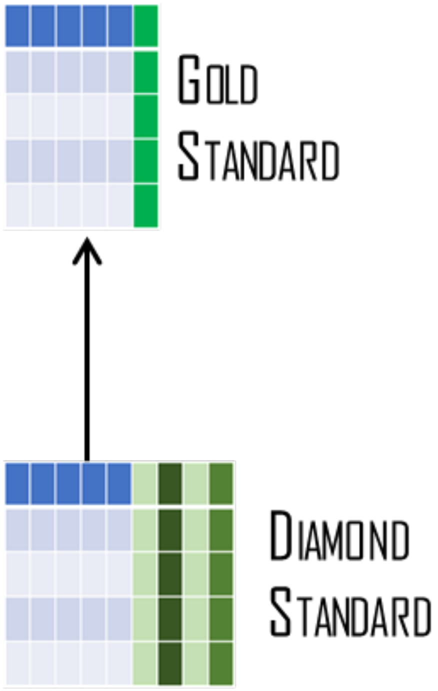


DIMENSIONS

DRYNESS

It measures the amount of information loss when applying a specific reduction.

The idea is that reductions with lower dryness (hence preserving more information) could be useful in situations where simply applying the majority rule would be *too risky* (small margin).



DIMENSIONS

Probabilistic reduction: maps each possible label to its relative frequency. Models degree of belief in the alternatives

$$freq(\mathbf{o}(x)) = \langle \frac{m_1}{m}, ..., \frac{m_{|Y|}}{m} \rangle$$

Fuzzy reduction: normalize the frequency of the alternatives by the maximum one *m**.
Gives a preference/plausibility ordering between the alternatives

$$fuzzy(\mathbf{o}(x)) = \langle \frac{m_1}{m^*}, ..., \frac{m_{|Y|}}{m^*} \rangle$$

Three-way reduction: set of labels that cannot be excluded under a decision-theoretic analysis.
Simply tells which labels are not totally implausible giving no quantitative information.

$$tw_d(\mathbf{o}(x), \epsilon, \alpha) = \begin{cases} \{\sigma_1, ..., \sigma_j\} & \alpha \cdot \sum_{i=1}^{j} \sigma_i + \epsilon \cdot \sum_{i=j+1}^{k} \sigma_i < \epsilon * (1 - \sigma_1) \\ \sigma_1 & \text{the inequality has no solution} \end{cases}$$

$$D(S) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0/1 decision, three instances, five raters

$$D(S) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0/1 decision, three instances, five raters

$$maj[D(S)] = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

Majority reduction

$$D(S) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0/1 decision, three instances, five raters

$$maj[D(S)] = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

Majority reduction

$$prob[D(S)] = \begin{bmatrix} (0 : 3/5, 1 : 2/5) \\ (0 : 1/5, 1 : 4/5) \\ (0 : 4/5, 1 : 1/5) \end{bmatrix}$$

Probabilistic reduction

$$D(S) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0/1 decision, three instances, five raters

$$maj[D(S)] = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

Majority reduction

$$prob[D(S)] = \begin{bmatrix} (0:3/5, 1:2/5) \\ (0:1/5, 1:4/5) \\ (0:4/5, 1:1/5) \end{bmatrix}$$

Probabilistic reduction

$$fuzzy[D(S)] = \begin{bmatrix} (0:1, 1:2/3) \\ (0:1/4, 1:1) \\ (0:1, 1:1/4) \end{bmatrix}$$

Fuzzy reduction

$$D(S) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0/1 decision, three instances, five raters

$$maj[D(S)] = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

Majority reduction

$$prob[D(S)] = \begin{bmatrix} (0 : 3/5, 1 : 2/5) \\ (0 : 1/5, 1 : 4/5) \\ (0 : 4/5, 1 : 1/5) \end{bmatrix}$$

Probabilistic reduction

$$fuzzy[D(S)] = \begin{bmatrix} (0 : 1, 1 : 2/3) \\ (0 : 1/4, 1 : 1) \\ (0 : 1, 1 : 1/4) \end{bmatrix}$$

Fuzzy reduction

$$tw[D(S)] = \begin{bmatrix} \{0, 1\} & 1 & 0 \end{bmatrix}$$

Three-way reduction

Probabilistic reduction: maps each possible label to its relative frequency. Models degree of be
alternatives

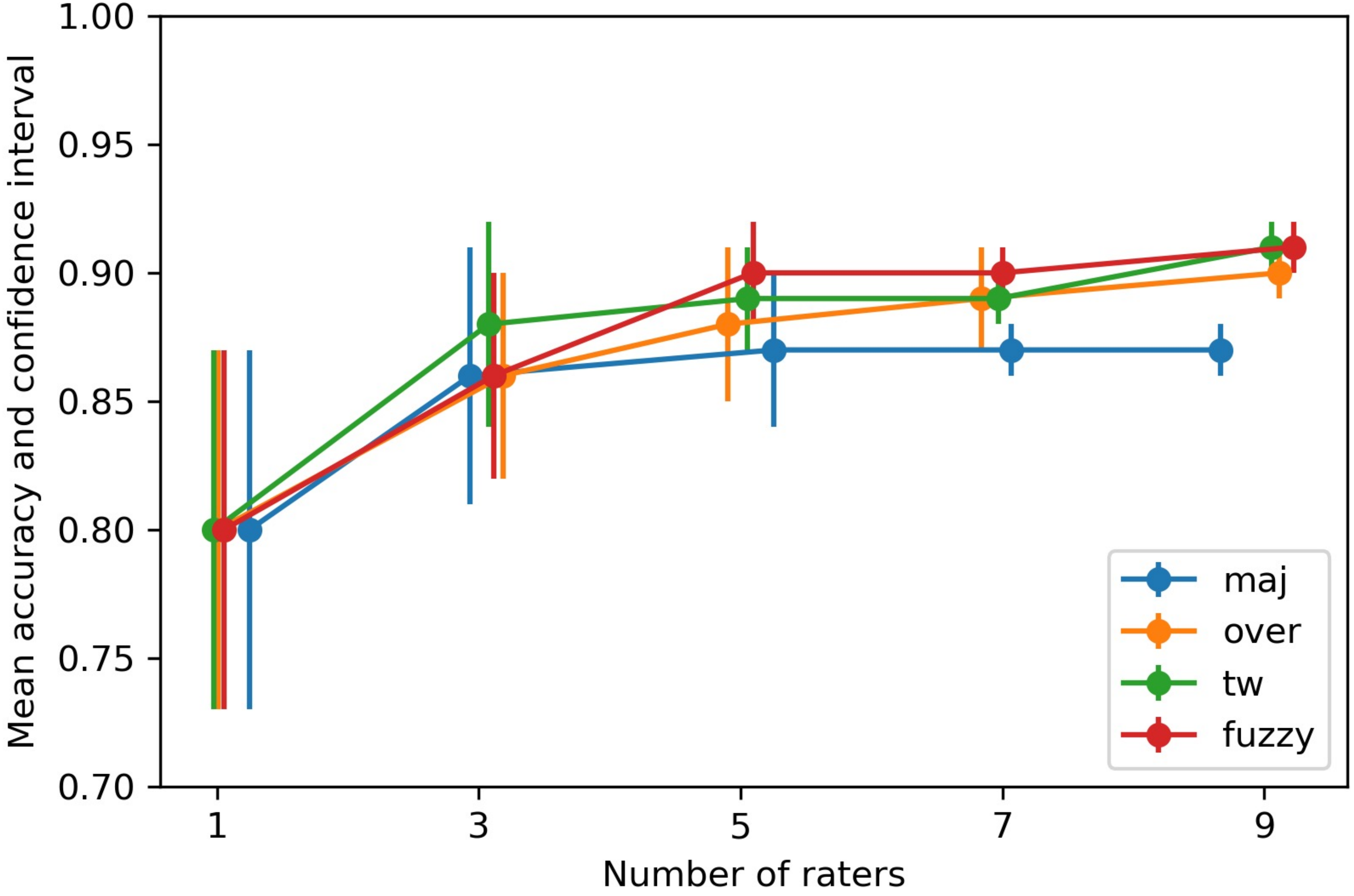
Fuzzy reduction: no
frequency of the alt
maximum one. Give
preference/plausibi
between the alterna

Three-way reductio
cannot be excluded
theoretic analysis. Simply tells which labels are not totally implausible giving no quantitative information.

Notice that each reduction corresponds to different ML settings requiring different classes of models and strategies:

1. Supervised learning (majority reduction)
2. Superset learning (Three-way reduction)
3. Learning on Fuzzy Data

From our experiments we observed that on high uncertainty/low reliability cases the three-way and fuzzy reductions result in better performances than standard majority



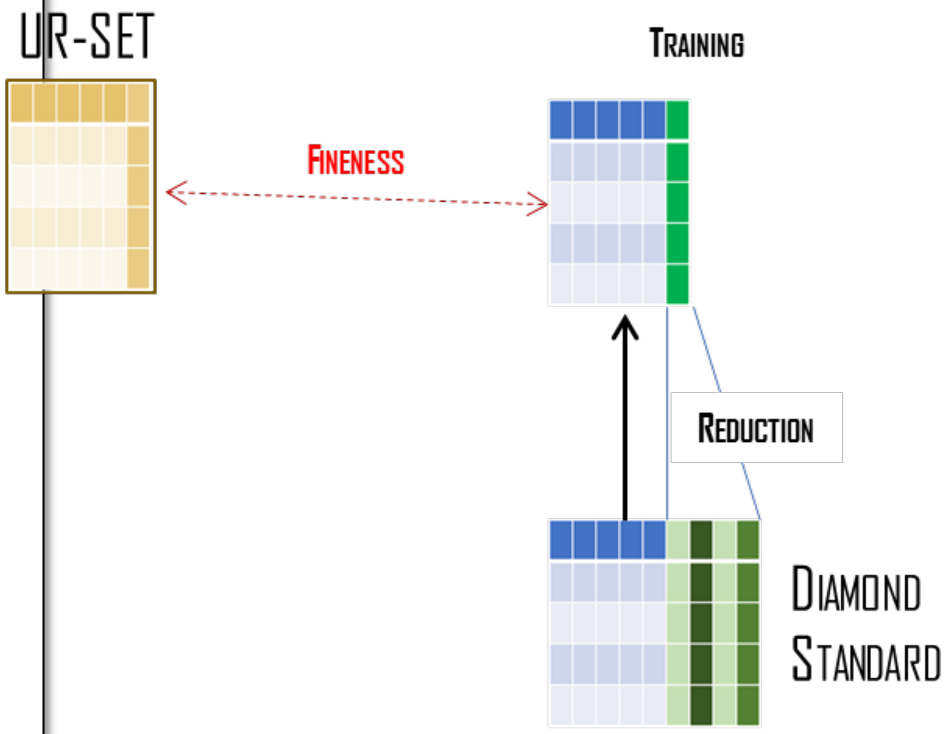
Mean accuracy and confidence interval
1.00
0.95
0.90
0.85
0.80
0.75
0.70
1
3
5
7
9
Number of raters
maj
over
tw
fuzzy

FINENESS

what is the probability that the gold-standard labels are equal to the correct (and unknown) labels in the UR-SET?

Via *Computational Learning Theory* (PAC Learning and VC dimension) this quality dimension is strictly related to *performance bounds* for the predictive model

- How many samples to get a fixed error?
- How many raters to obtain a fixed fineness?



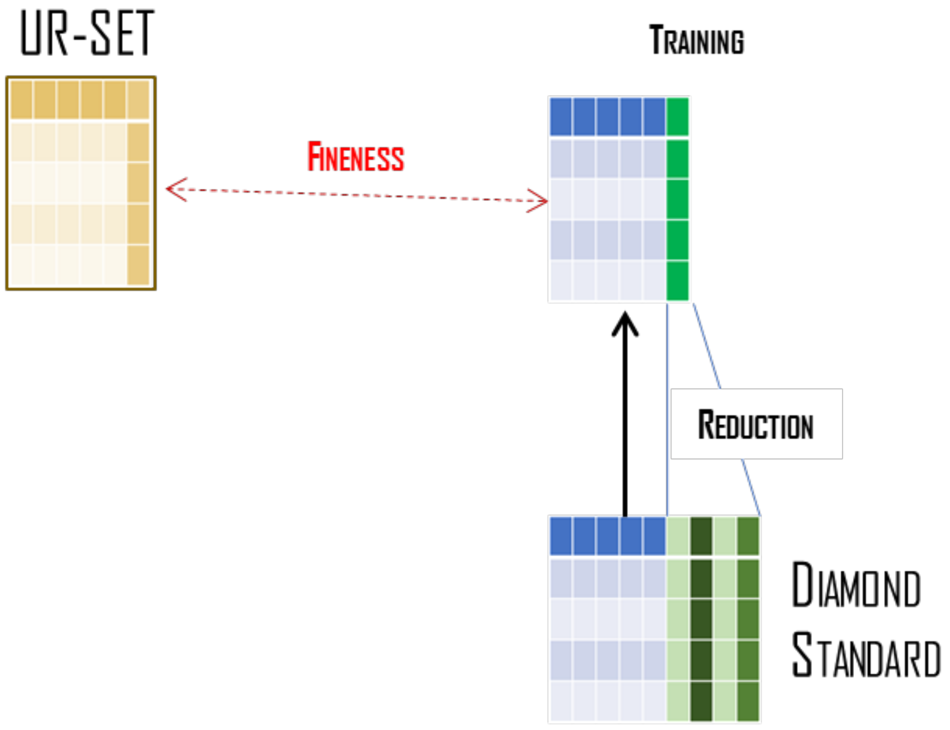
DIMENSIONS

Notably, we can bound the *number of raters needed to achieve a desired level of fineness*

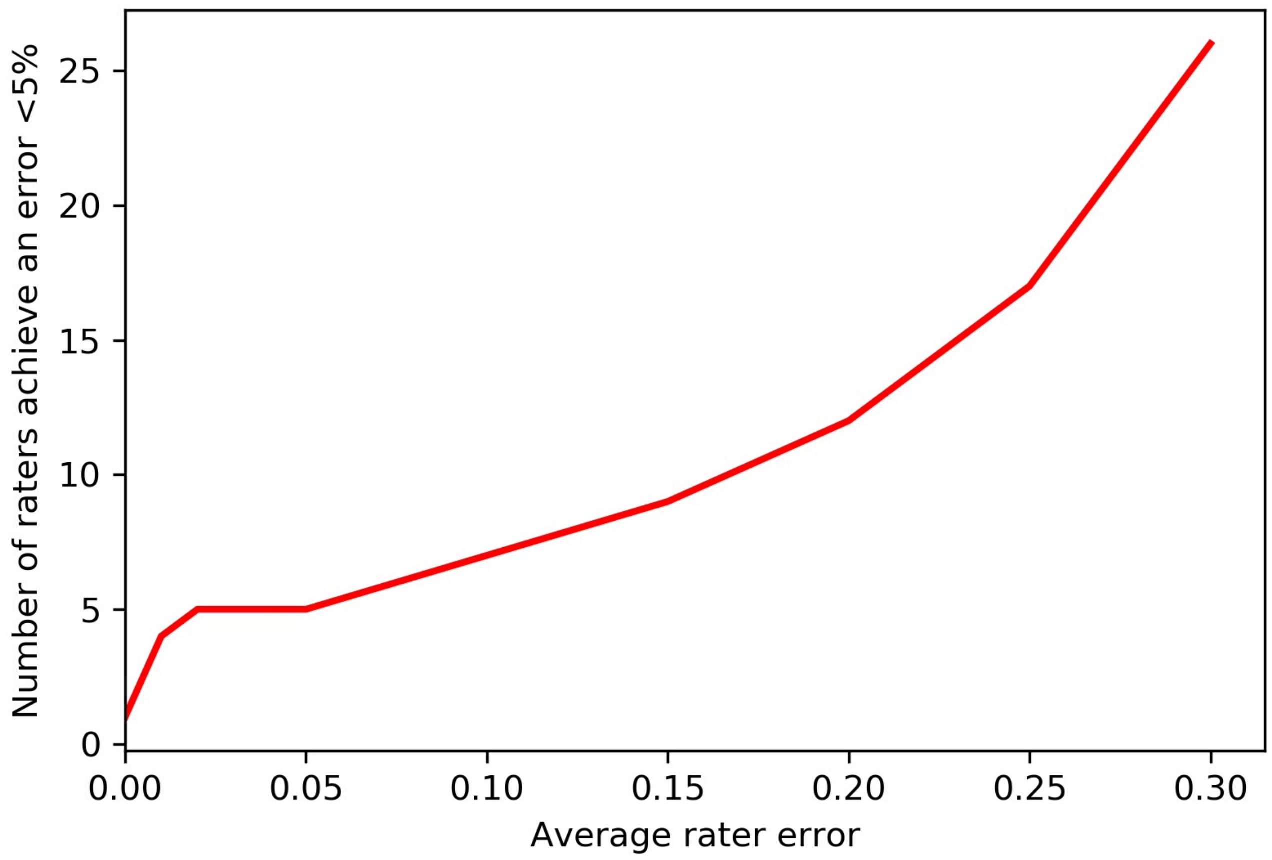
$$\mathcal{O}\left(\frac{log\frac{|D|}{\delta}}{(1-2\eta_O)^2}\right)$$

Also, we can obtain the *sample complexity* (number instances required to correctly learn the target concept with high probability and low approximation error)

$$\mathcal{O}\left(\frac{d\cdot log\frac{1}{\delta}}{\epsilon(1-2e^{-\frac{m+1}{2}log\frac{m+1}{2\mu}})^2}\right)$$



DIMENSIONS



Bound of the number of raters needed to obtain a labeling error $\delta \leq 0.05$ at a fixed average rater error rate on a dataset of size $|S| = 771$

Summarizing

- The number and expertise of the raters have a critical influence on accuracy and generalization capacity of the trained models
- New reduction methods can achieve higher accuracy and higher robustness when the accuracy of the raters decreases

To conclude

It would be good to have more transparency in the AI/ML community and availability to share the original multi-rater datasets (i.e.. Diamond Standards) along with Gold Standards and reduction techniques adopted,

To conclude

It would be good to have more transparency in the AI/ML community and availability to share the original multi-rater datasets (i.e.. Diamond Standards) along with Gold Standards and reduction techniques adopted,

or at least publish some quality measures re the dimensions mentioned above,

To conclude

It would be good to have more transparency in the AI/ML community and availability to share the original multi-rater datasets (i.e.. Diamond Standards) along with Gold Standards and reduction techniques adopted,

or at least publish some quality measures re the dimensions mentioned above,

or at the very least reliability measures like kappa or alpha.