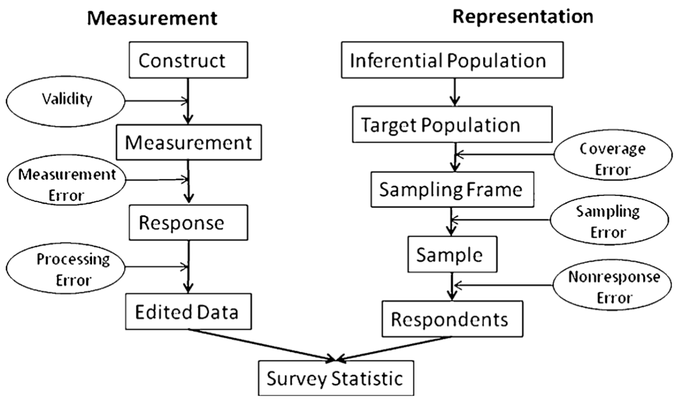
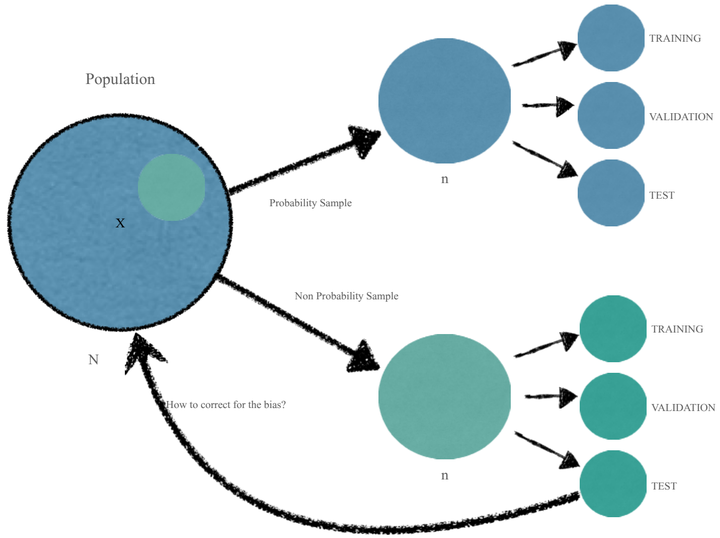
Quality of Training Data

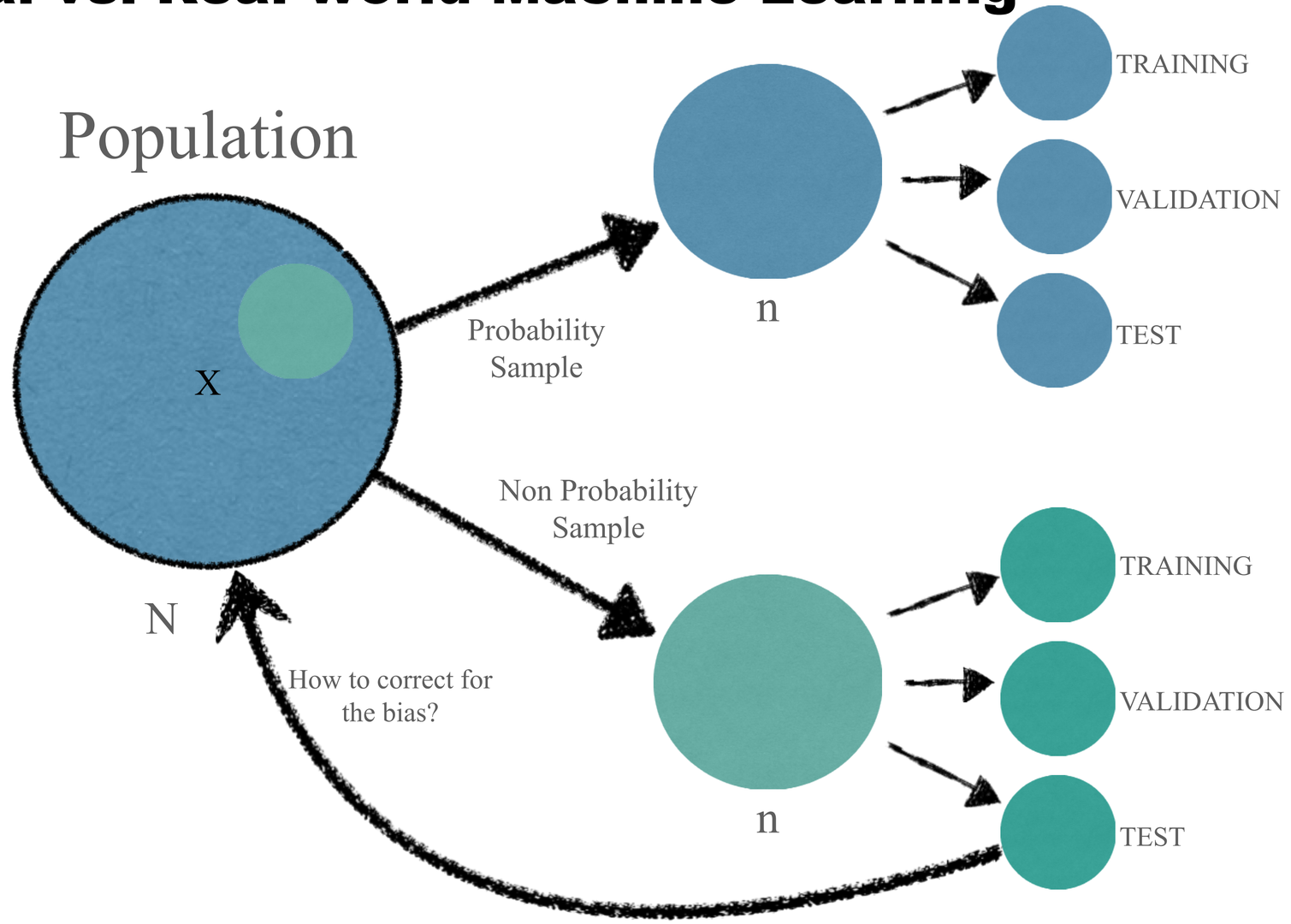




Marco Puts

ML Quality of training data: Representativity

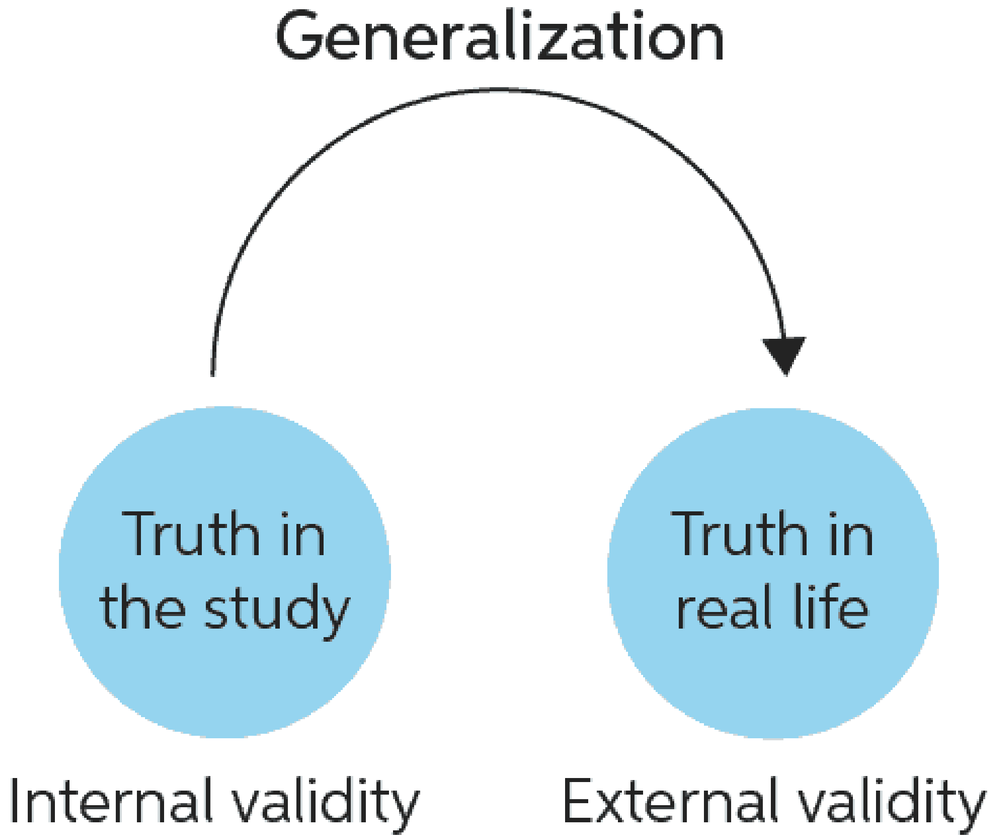
Ideal vs. Real world Machine Learning



- Finite Populations
- Sampling Error
- Bias Estimations

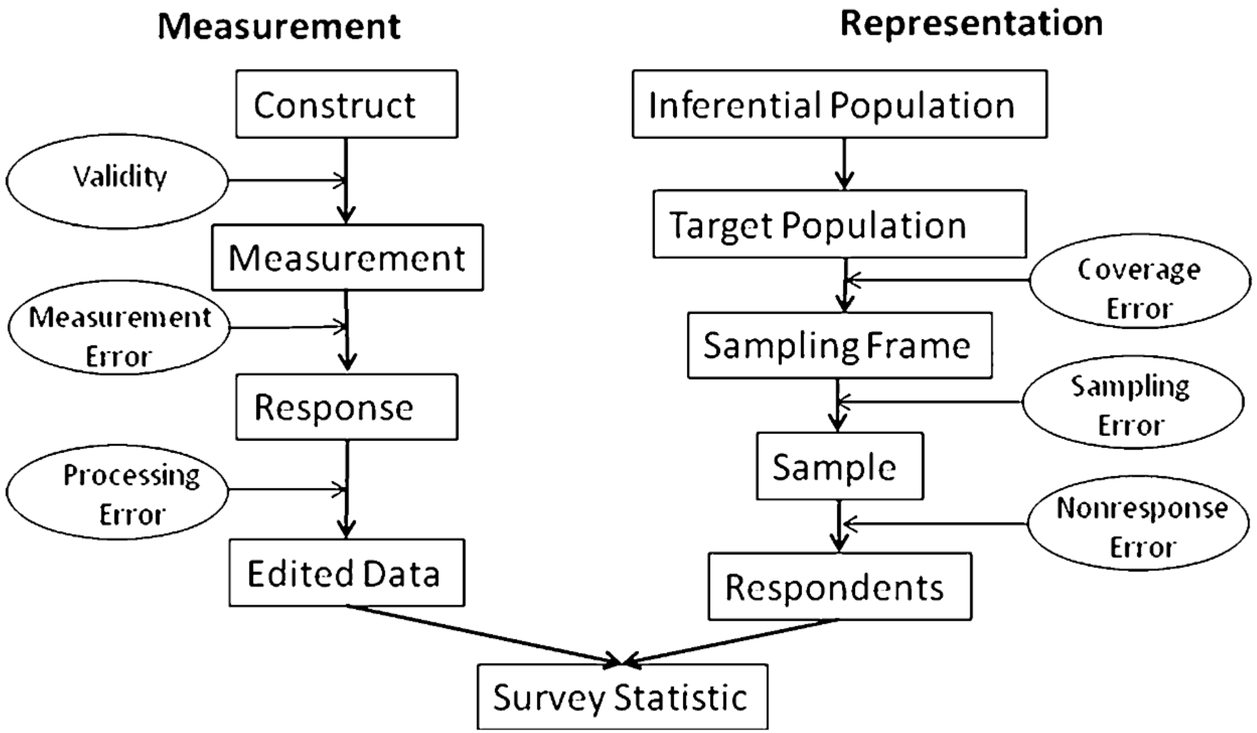
- Confusion Matrix
- Performance Indicators
- K-fold Cross validation
- …

Internal vs external validity



Total Survey Error model

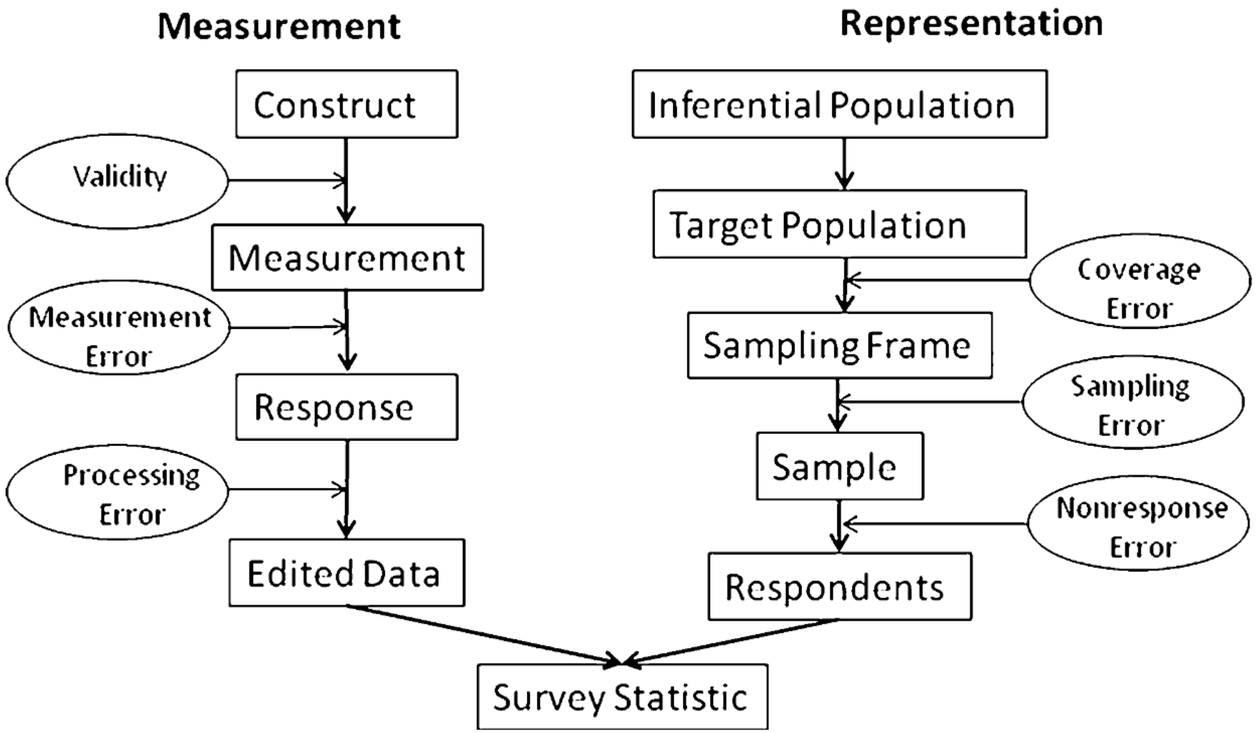
Training



Total Survey Error model

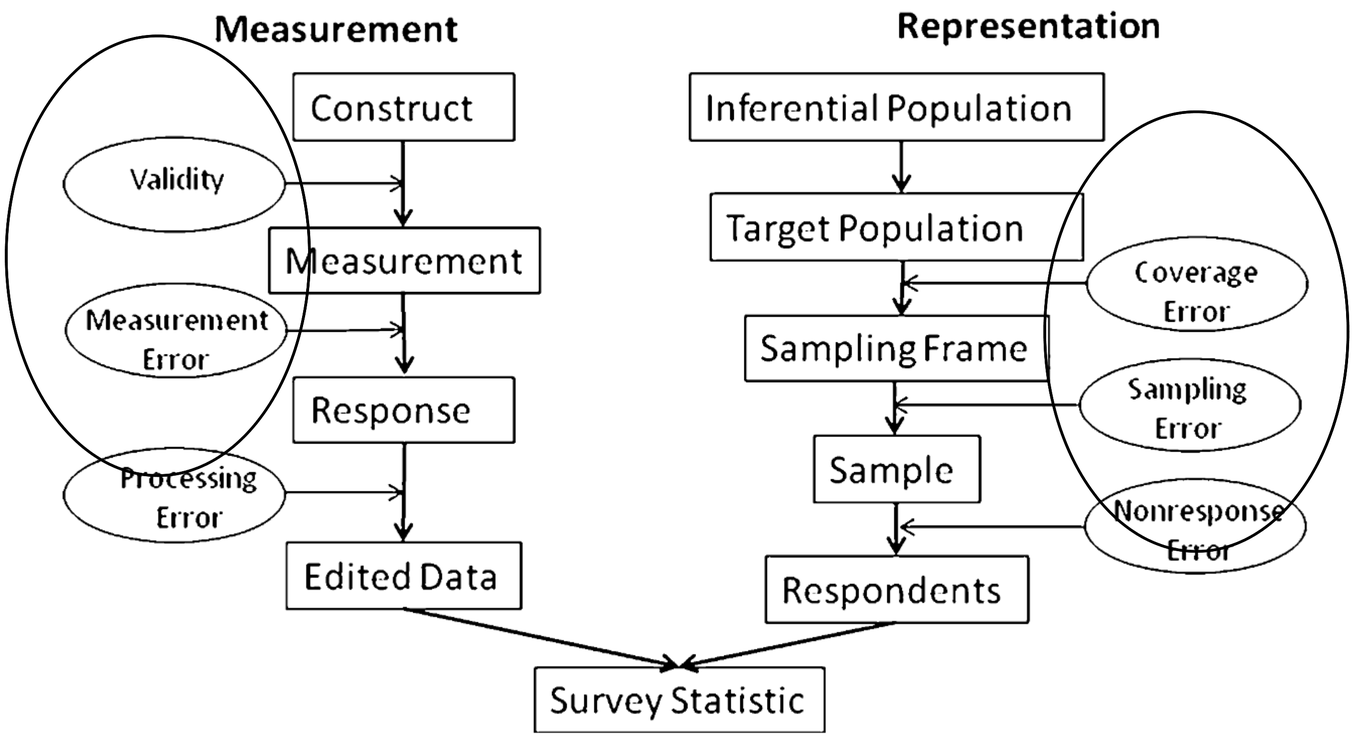
Training

Prediction

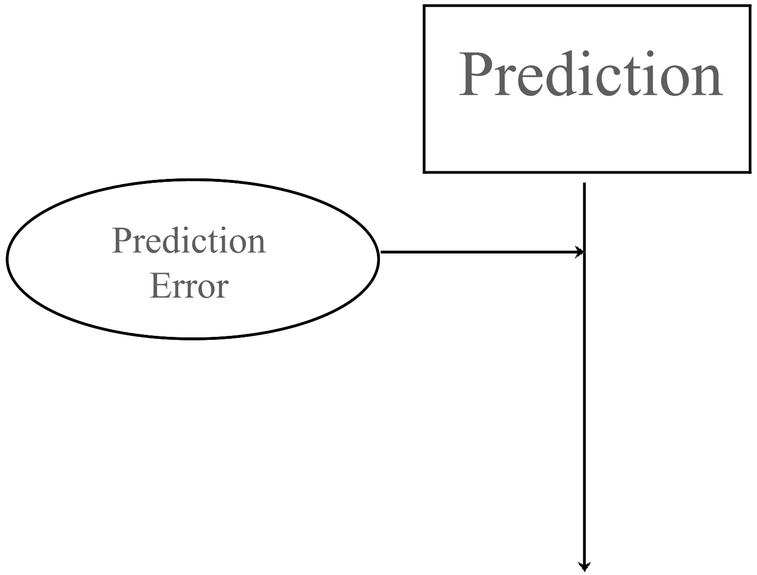


Total Survey Error model

Training

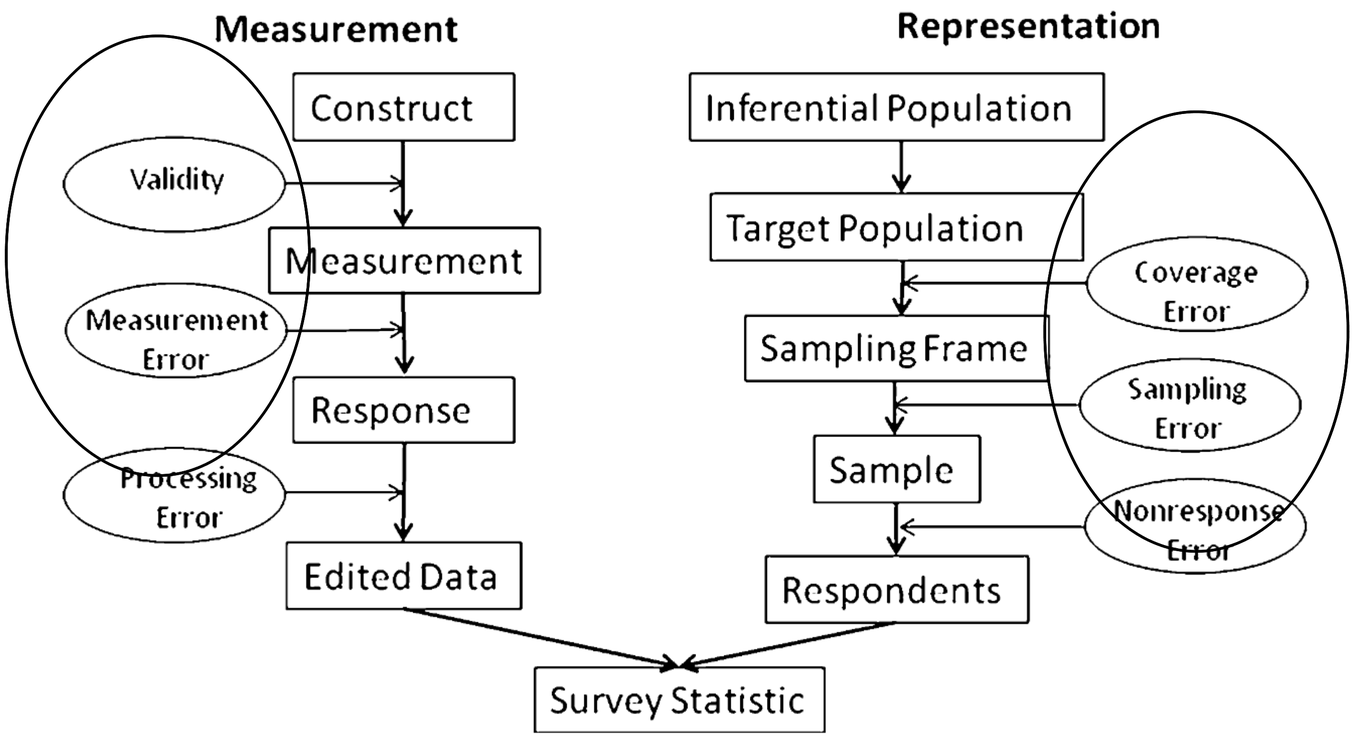


Prediction

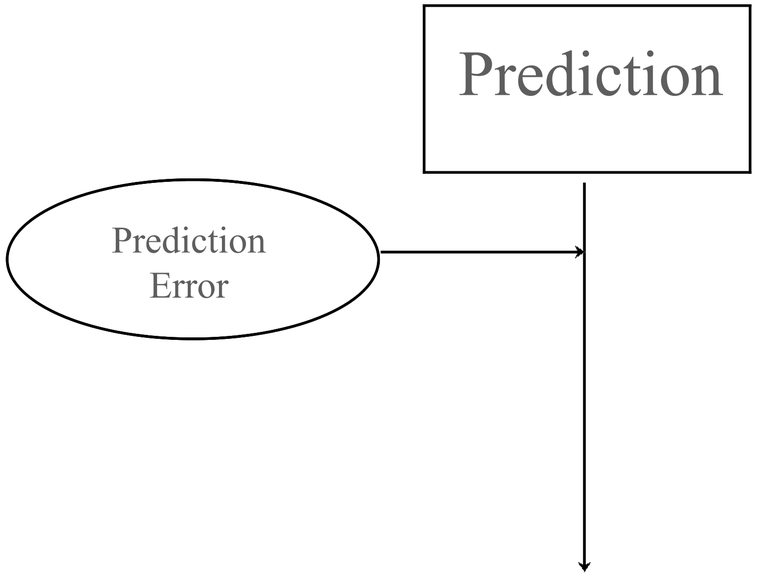


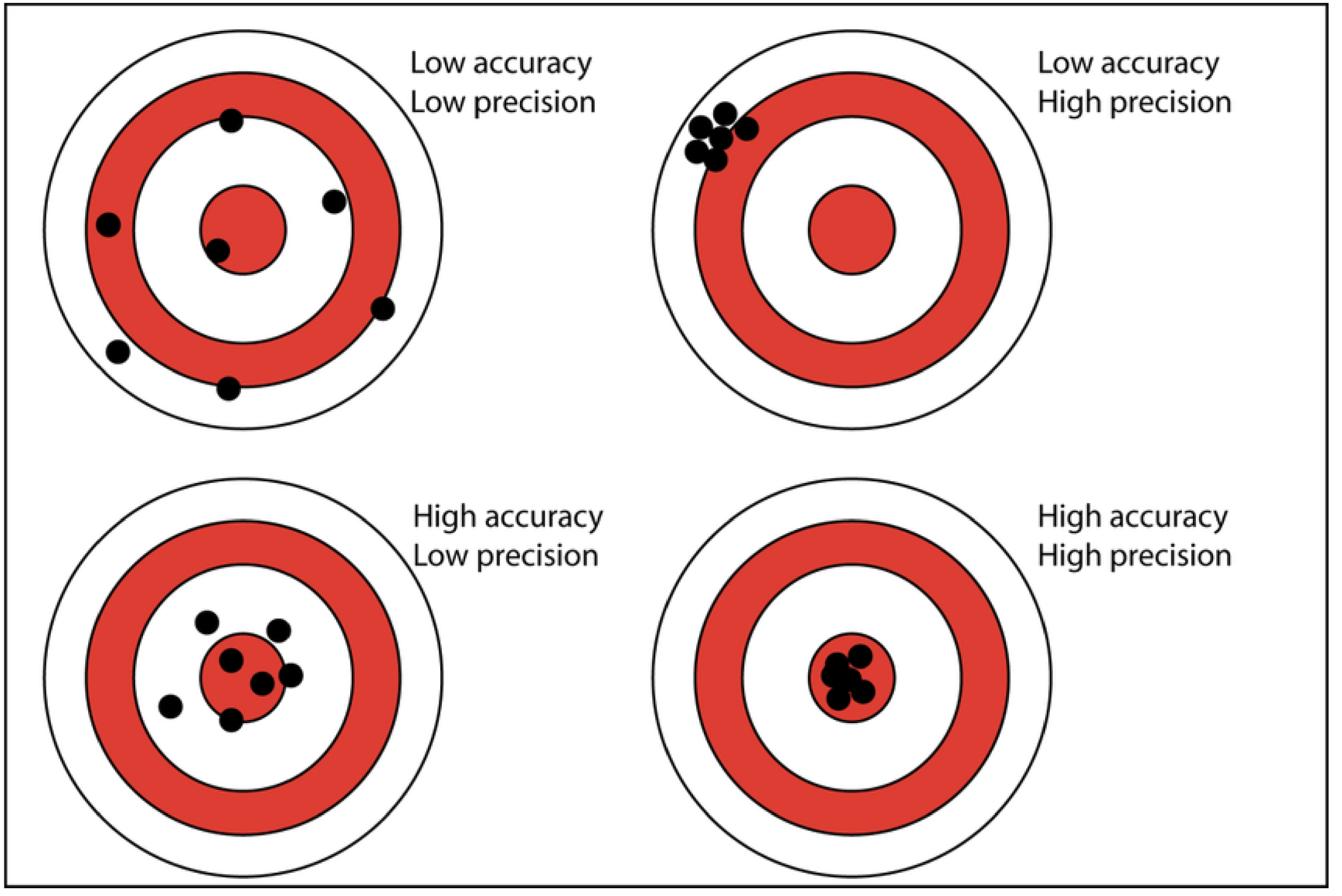
Total Survey Error model

Training



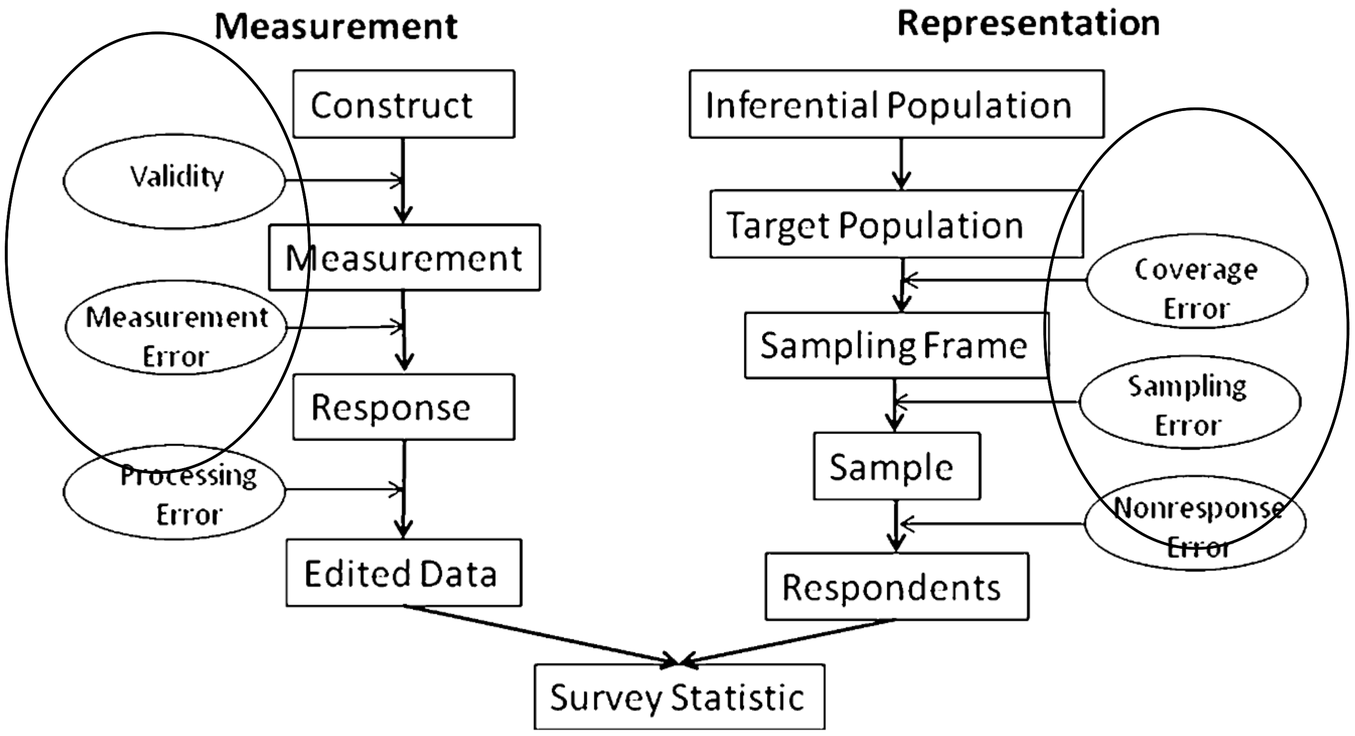
Prediction




Low accuracy
Low precision
Low accuracy
High precision
High accuracy
Low precision
High accuracy
High precision

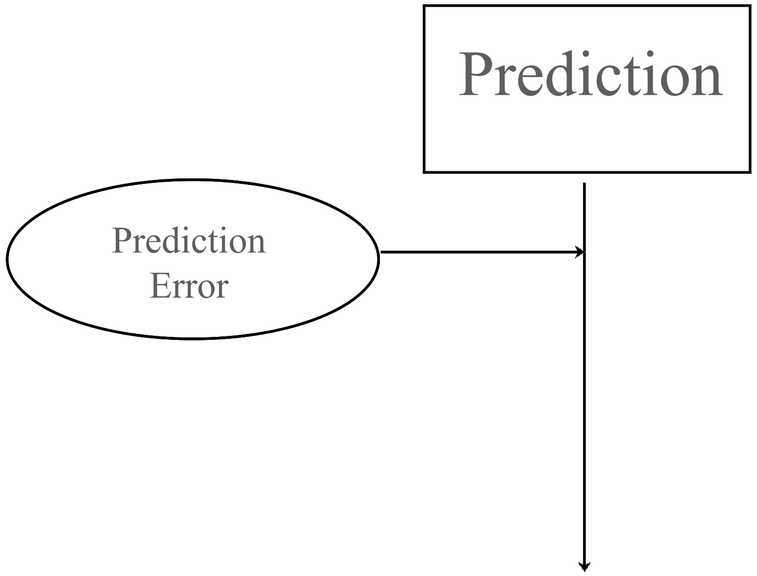
Total Survey Error model

Training

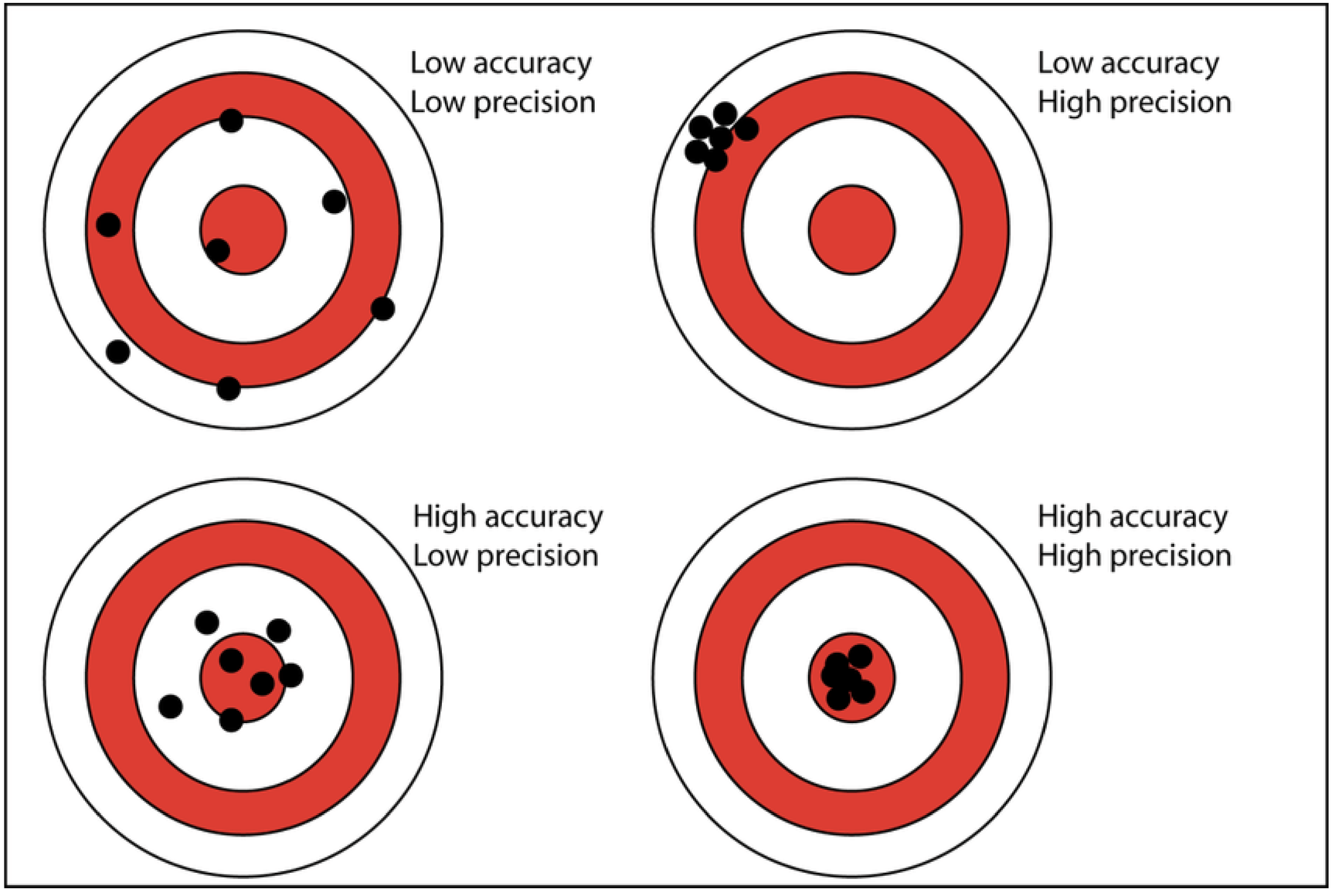


Accuracy + Precision

Prediction



Accuracy


Low accuracy
Low precision
Low accuracy
High precision
High accuracy
Low precision
High accuracy
High precision

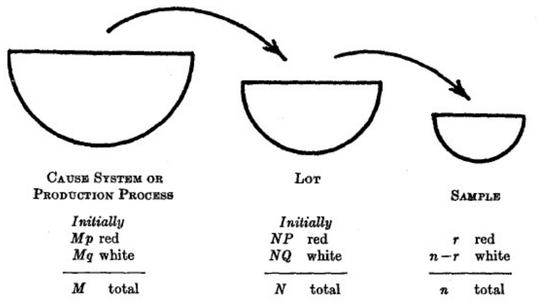
Model based and design based inference

- Designed based inference:
  - Finite Population
  - Non-exchangeable
  - Sample is a set theory concept
- Model based inference:
  - Infinite population
  - Exchangeability
  - Sample is realization of a random variable

Survey methodology

ML

Finite Population methodology



Conclusion

- Survey Methodology important when dealing with a sample as training set.
- Total survey error model can be used.
- Most errors in the sample used as a training set result in a bias.
- In case of human annotation: monitor the annotation process
- What are we modelling? the infinite or finite population?