

Conformal Prediction for Bone Surgery

10th IEEE Swiss Conference on Data Science

23.06.2023

About Us

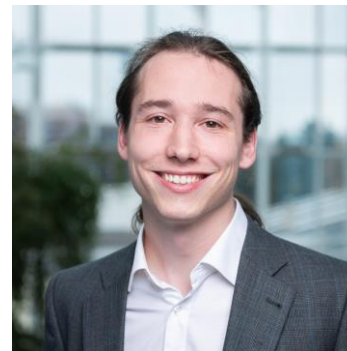
Advanced Osteotomy Tools

Simon Pezold

Senior Machine Learning Engineer

<https://www.linkedin.com/in/spezold>

<https://github.com/spezold>



Silvan Melchior

Expert Data Scientist

<https://www.linkedin.com/in/silvan-melchior/>

<https://github.com/silvanmelchior>

Zühlke

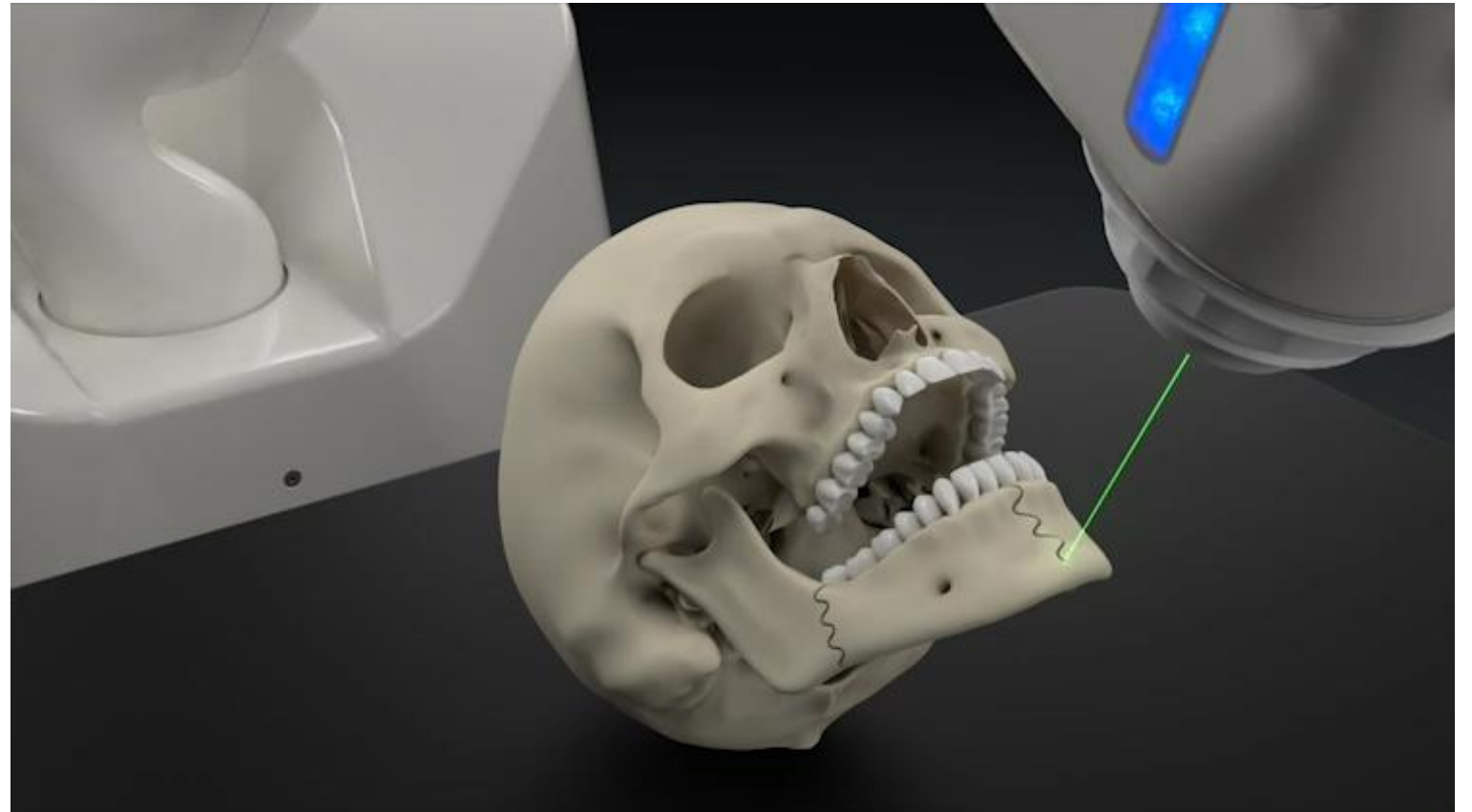
CARLO

Cut bones with laser

- Autonomous, precise intervention
- Limitless cut geometries
- Faster healing process
- Contact-free, sterile laser
- Digital workflow

Facts and figures

- World's first CE-certified laser osteotome
- 100+ patients successfully treated
- Er:YAG laser with 10 pulses/second
- 20+ mm achievable depth



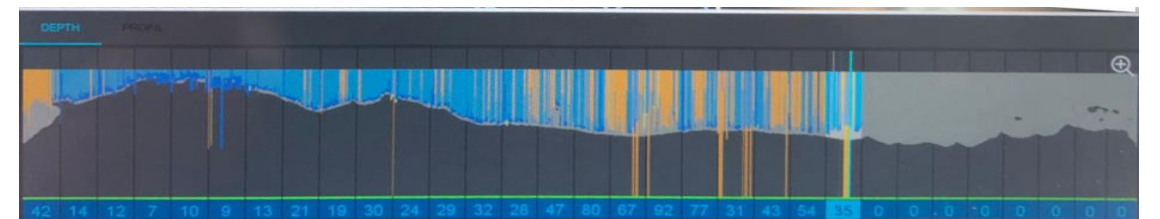
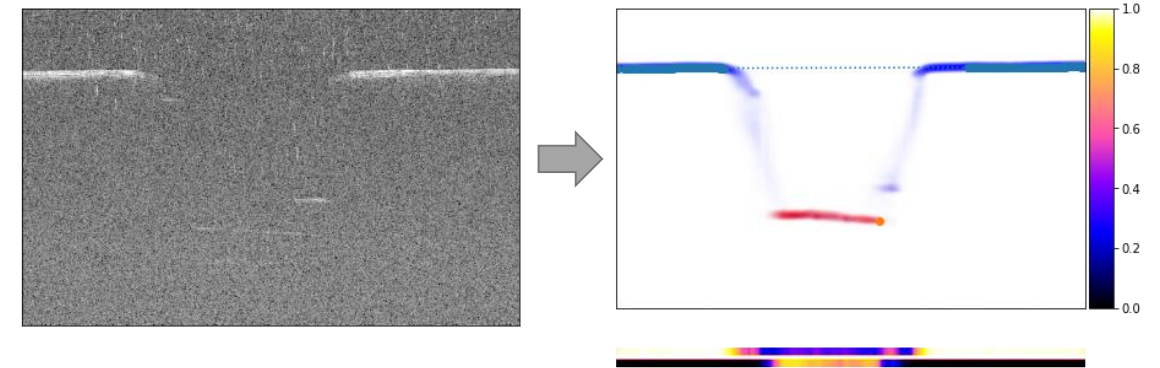
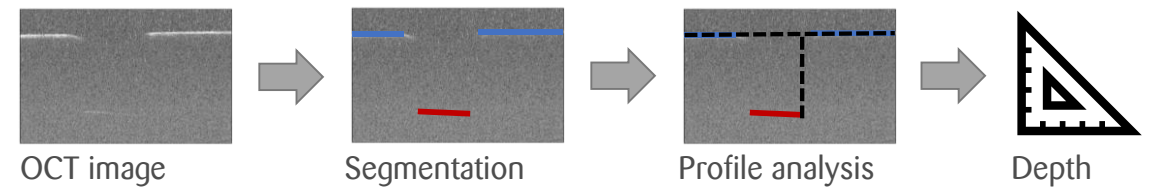
Computer Vision based Depth Estimate

Depth from OCT

- OCT image: 2D cut profile per laser pulse
 - Segmentation: surface extraction with U-Net
 - Profile analysis: rule-based algorithm
- New depth estimate for pulsed location on bone

Depth visualization during surgery

- 1D depth profile along cut path
- Real-time update with new depth estimates
- Color coding: **measured depth** vs. **no depth**





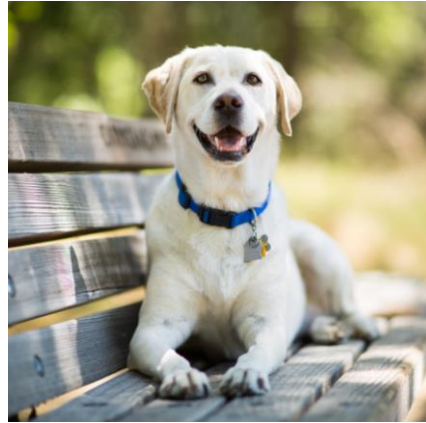
Conformal Prediction

Uncertainty Quantification

Goal: Get uncertainty of prediction out of model



Cat: 99%
Dog: 1%



Cat: 1%
Dog: 99%



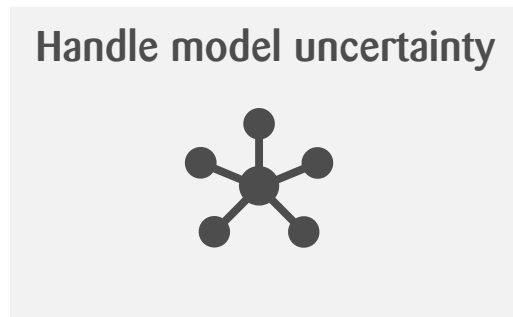
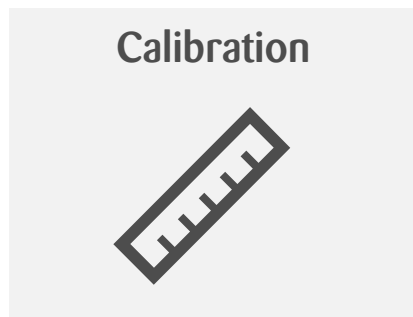
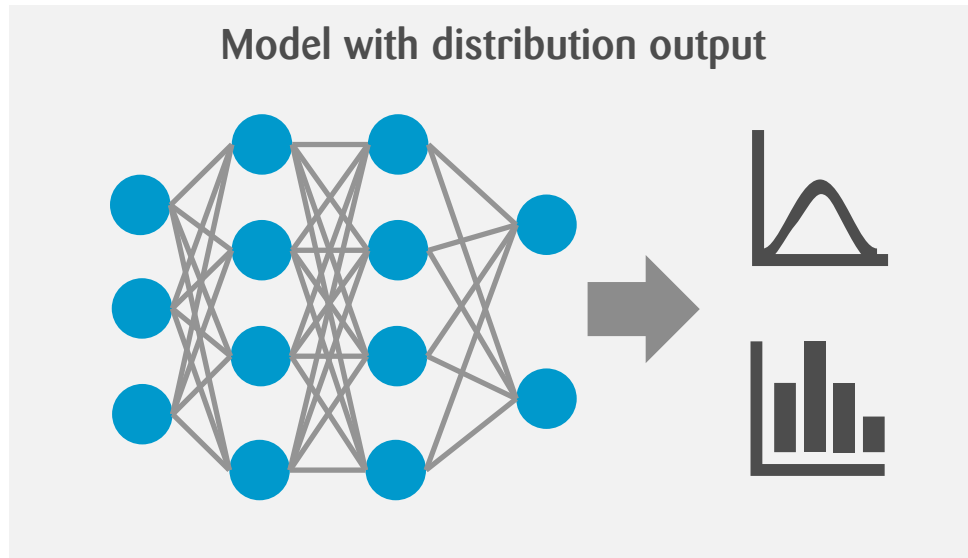
Cat: 40%
Dog: 60%




???

image credits cat-dog: Gấu Mèo Bắc Mỹ

Usual Approaches



 **Problem: Assumptions**

- Distribution (e.g. Gaussian)
- Independence
- Scaling-induced assumptions

➔ Wrong uncertainty estimates

Conformal Prediction

A bit a different approach to uncertainty quantification...

Start with **desired guarantee**

I want my prediction to be “correct” in 95% of all cases

Then design mechanism which takes prediction & turns it into a **prediction set** with the given guarantee



fox squirrel



conformalize



{fox squirrel, gray fox, bucket, rain barrel}



1'100'000 \$



conformalize



900'000 – 1'300'000 \$

correct = “ground truth is within prediction set”

Conformal Prediction Properties

Distribution-free

- Output is a prediction set
- Only assumption: Data points are exchangeable

Model-agnostic

- Works on top of any model
- Works for classification, regression, segmentation, outlier detection, ...

Coverage guarantee

- Statistical guarantees
- Desired uncertainty set in advance

Simple

- Principles easy to use
- Good libraries available

WE



CONFORMAL PREDICTION



Recipes

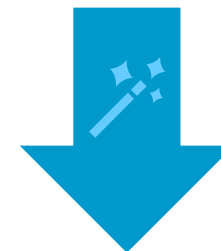
Classification Example



Color: yellow
Size: 10cm
Weight: 0.2kg
...



banana: ~~30%~~
lemon: ~~25%~~
pepper: ~~15%~~
...

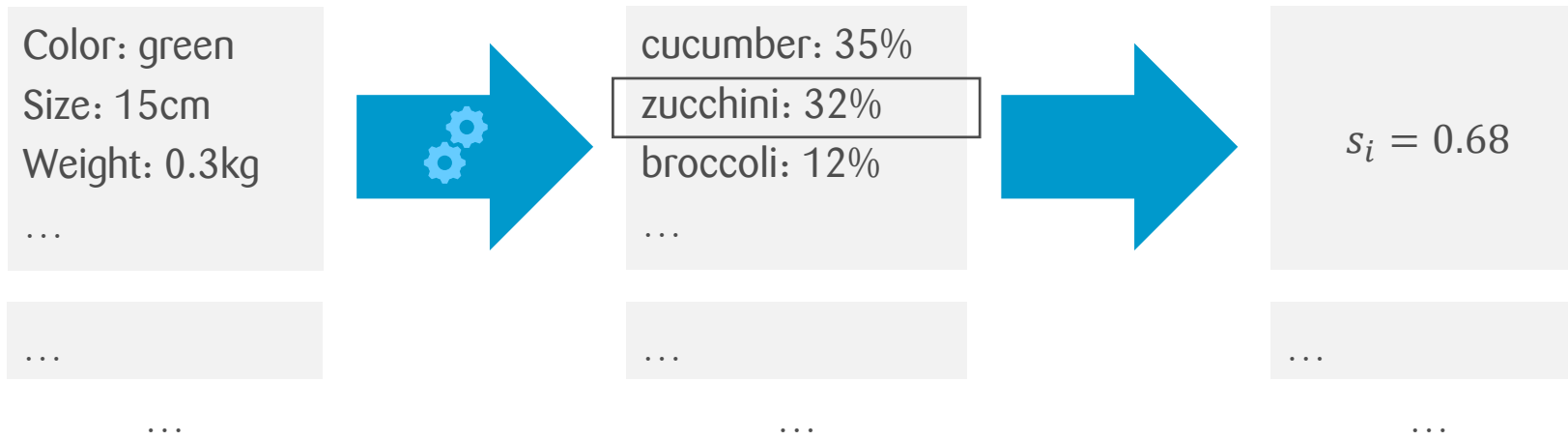


It is one of
{banana, lemon}
with 90%
probability ✓

Conformalized Classification (LABEL method)

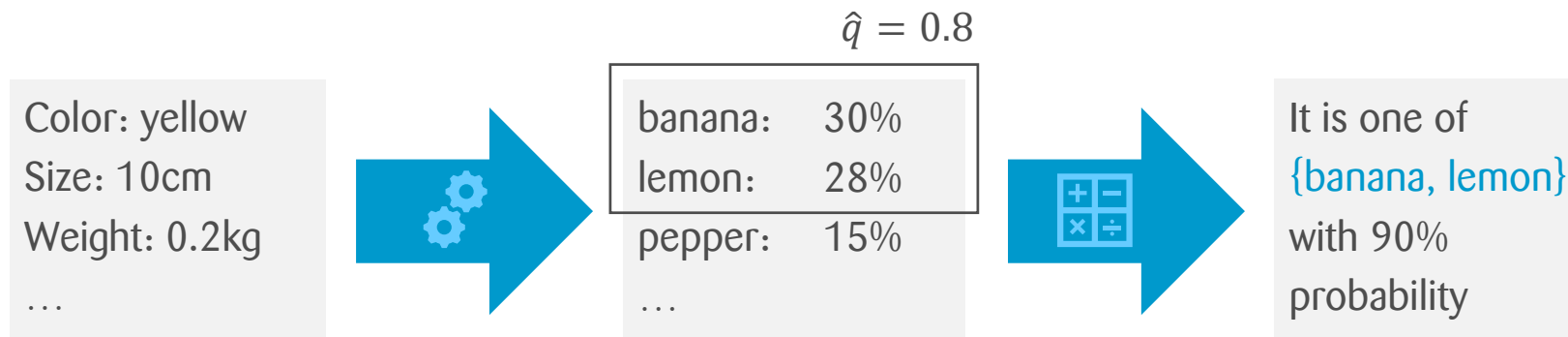
- 1 **Build a calibration set**
 - Representative & left out from training

- 2 **Calculate “non-conformity scores”**
 - For each data point (x_i, y_i) , calculate $s_i = 1 - f(x_i)[y_i]$
 - “1 minus score of true class”



Conformalized Classification (LABEL method)

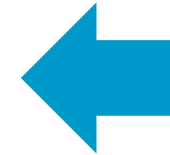
- 3 Calculate 90%-quantile \hat{q} of scores**
 - We now know that for prediction uncertainties s_i below \hat{q} , the class is correct with 90% probability
- 4 For every prediction now, build prediction set**
 - Includes all classes, whose conformal score is below \hat{q}



General Recipe

For a general input x and output y with a chosen error rate α

- Define a score function $s(x, y)$
 - Larger scores should indicate worse agreement between x and y
 - Under the hood uses model
- Compute \hat{q} as $(1 - \alpha)$ quantile of calibration scores $s(x_1, y_1), \dots, s(x_n, y_n)$
- Use quantile to form prediction sets $C(x_{test}) = \{y \mid s(x_{test}, y) \leq \hat{q}\}$ for new examples



This choice determines most properties of the prediction sets

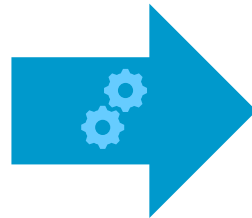
Following this recipe, with i.i.d. $(x_1, y_1), \dots, (x_n, y_n), (x_{test}, y_{test})$ we have $P(y_{test} \in C(x_{test})) \geq 1 - \alpha$

one detail left out: finite sample correction required slight adjustment of quantile

Regression: Conformalized Quantile Regression

Quantile Regression

- Train a model to predict quantiles
- Can for example train two models (5% and 95%) to get interval predictions for 90%



$$q_{05} = 0.05 \text{ kg}$$

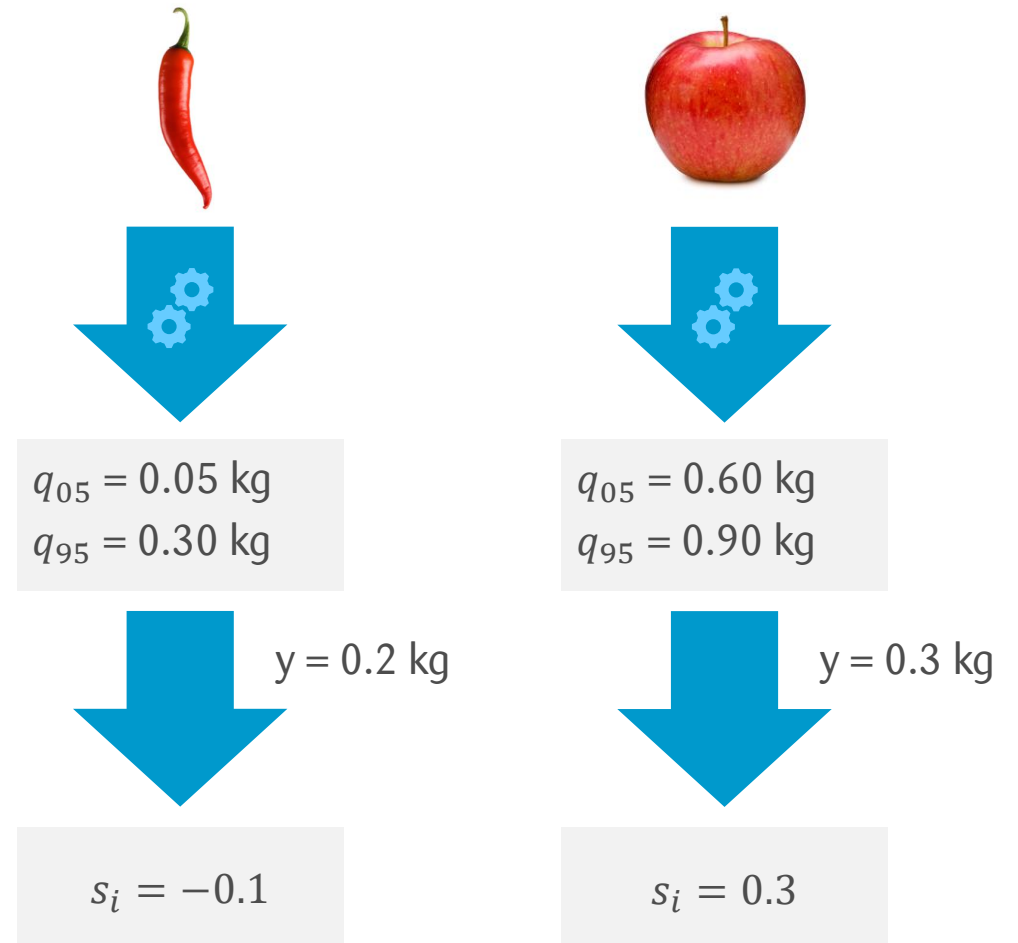
$$q_{95} = 0.30 \text{ kg}$$



Regression: Conformalized Quantile Regression

Conformalized Quantile Regression

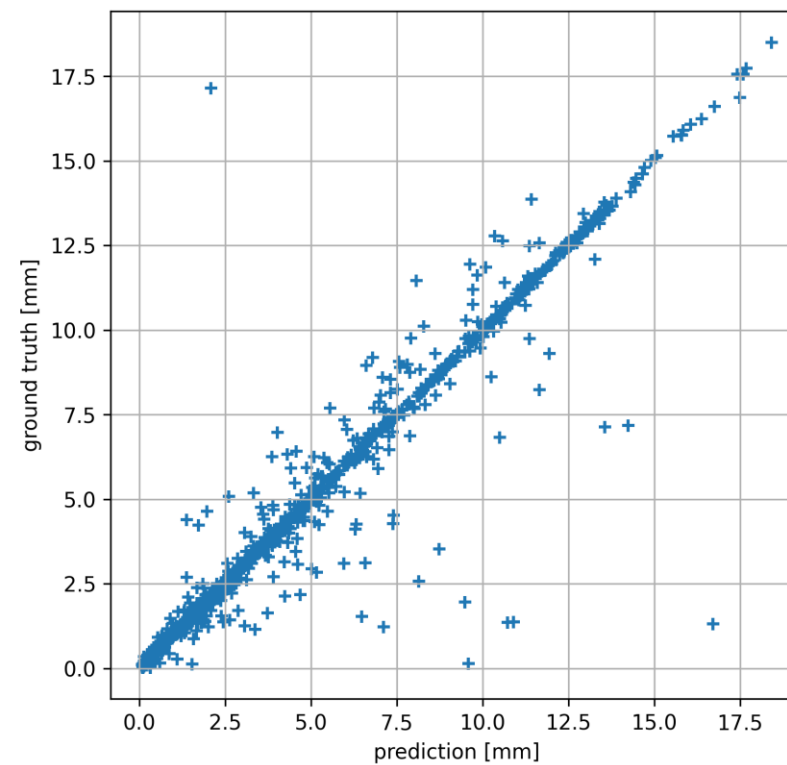
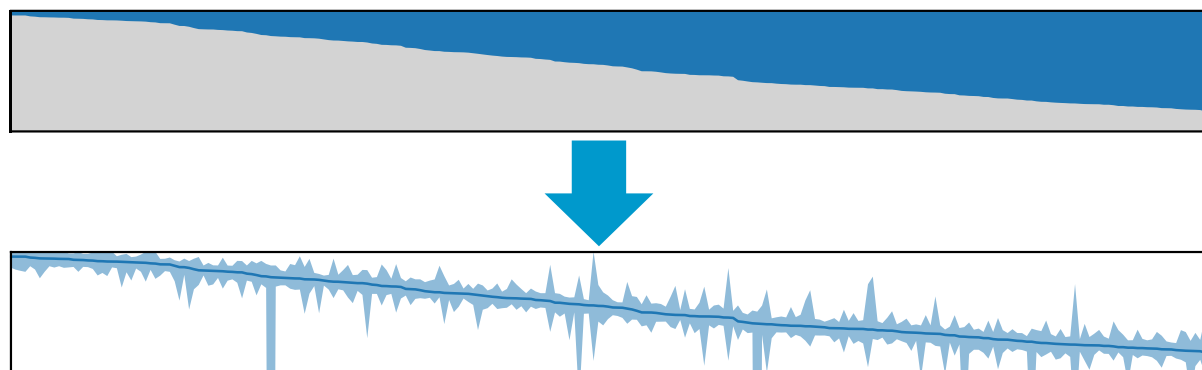
- Score: $s(x, y) = \max\{f_5(x) - y, y - f_{95}(x)\}$
 - Distance to nearest model output (positive if outside, negative if inside)
- Prediction set: $C(x) = [f_5(x) - \hat{q}, f_{95}(x) + \hat{q}]$
 - Just grow or shrink outputs



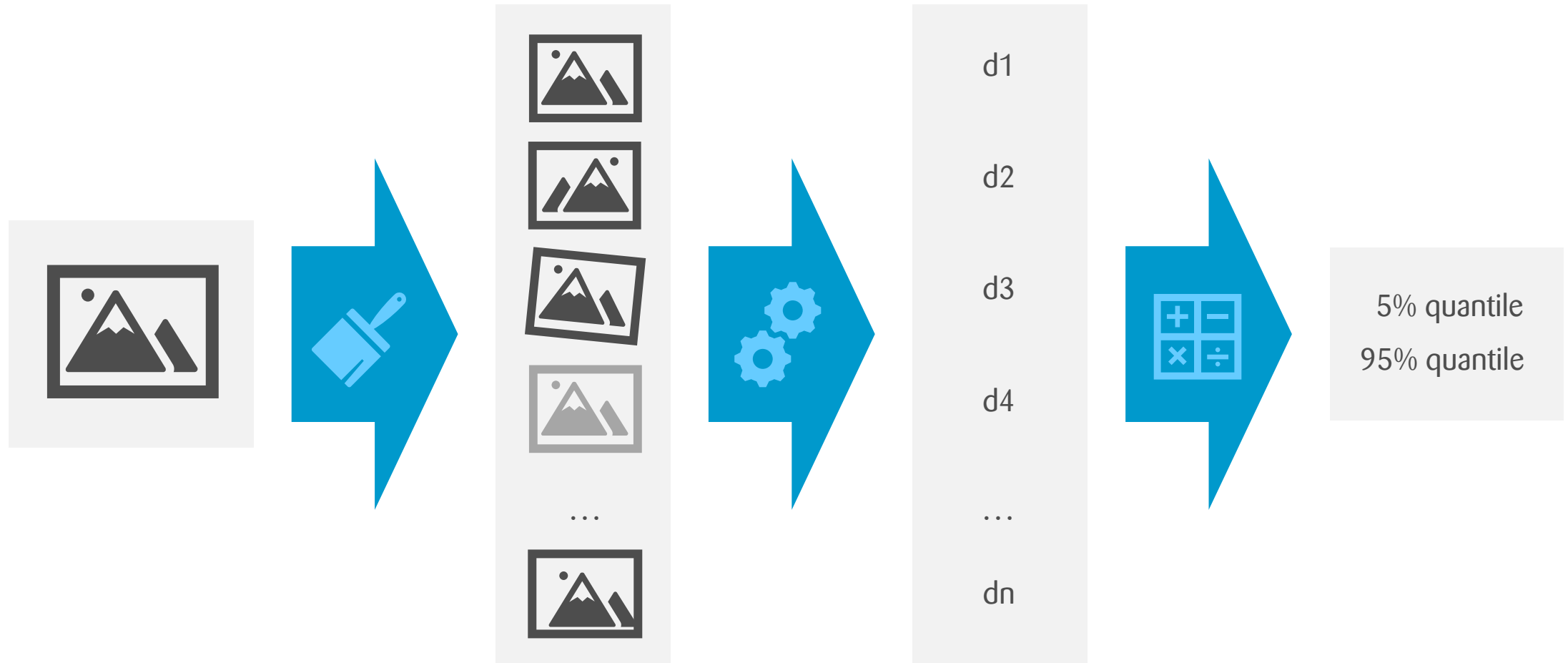


Conformalized CARLO

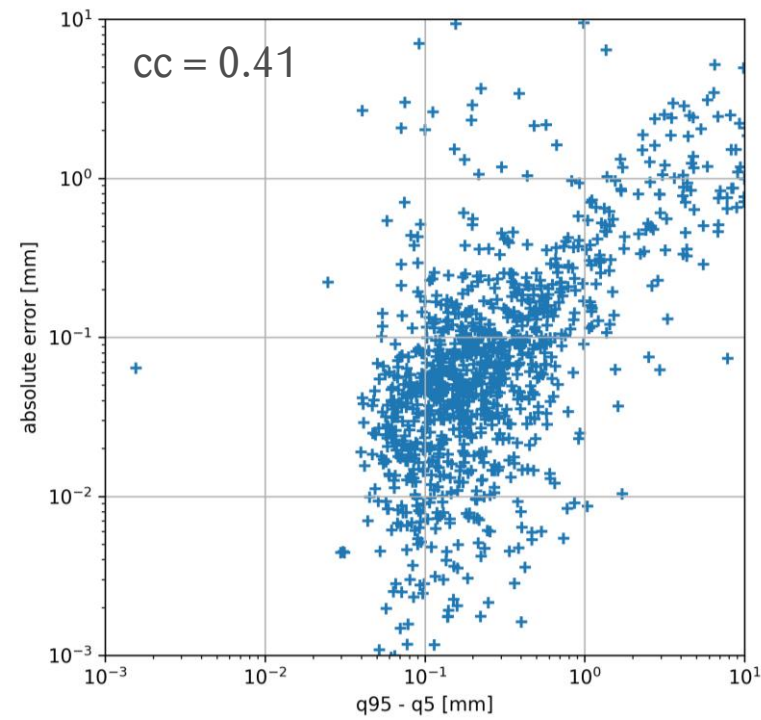
CARLO



Test-time Data Augmentation

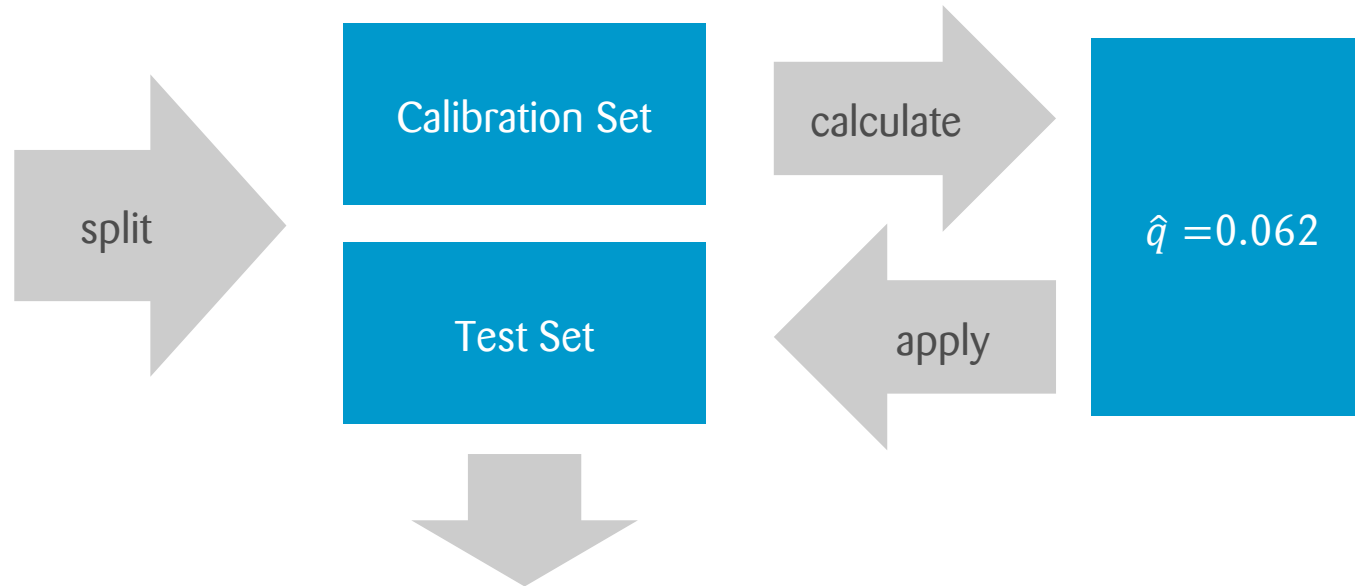
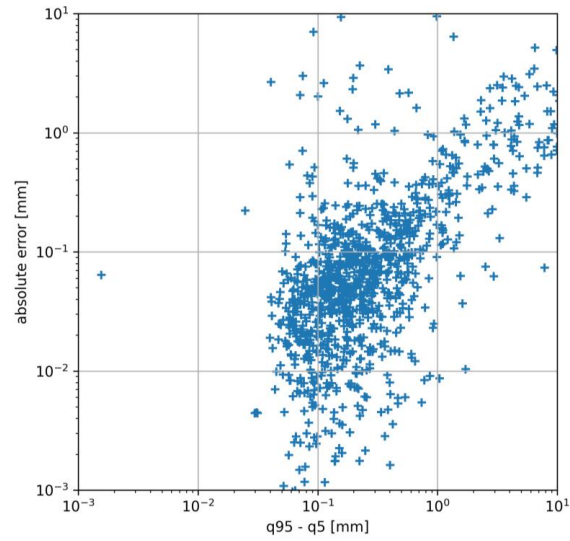


Data Augmentation Results



 Coverage of quantile interval: 70.8%

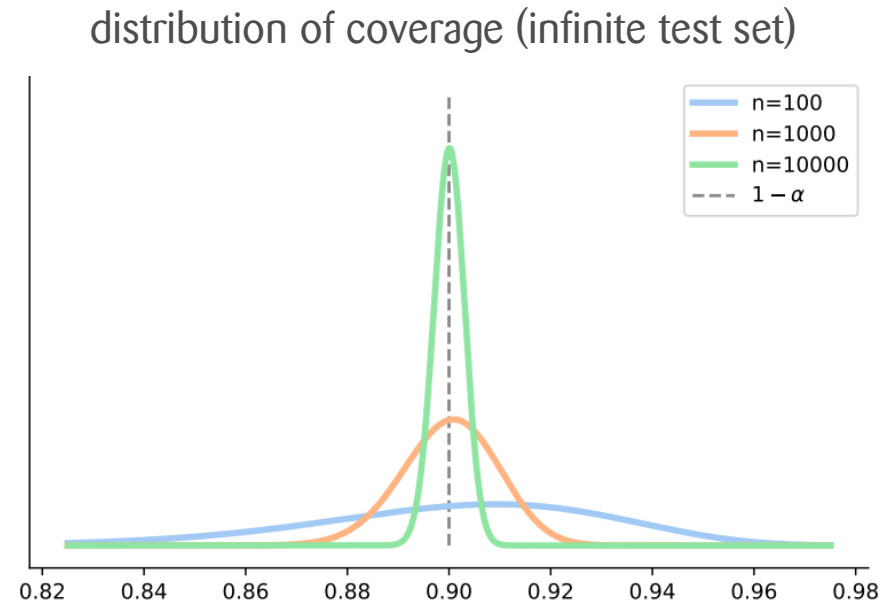
Results



✓ Coverage of conf. quantile interval: 91.0%

Why 91%?

- Conformal procedures are a stochastic process in the choice of the calibration set
- The coverage guarantee only holds “on average”
- There is an analytic solution to the coverage distribution



Plot from Angelopoulos et al., 2022

Prediction Set Properties

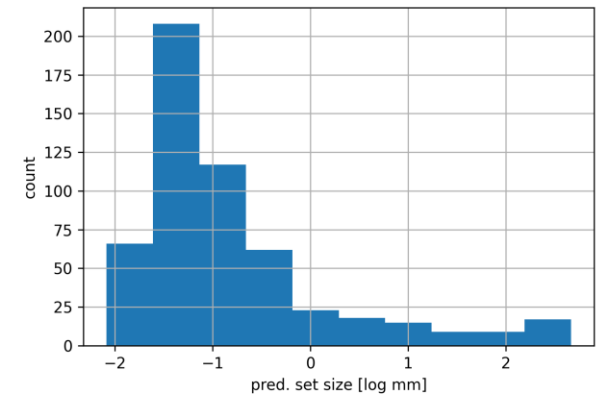
Average prediction set size

- Shows how well score is aligned with underlying model
- However: Procedure with smallest avg. size is not necessarily the best (!)

Adaptivity

- Prediction set size should faithfully represent model uncertainty
- Easy examples should have small sets, hard examples large sets
- Histogram of set sizes should be spread wide

0.99 mm

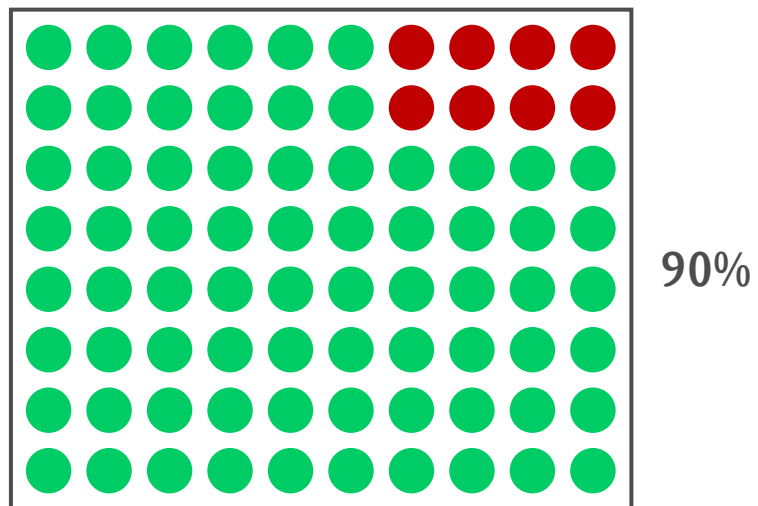




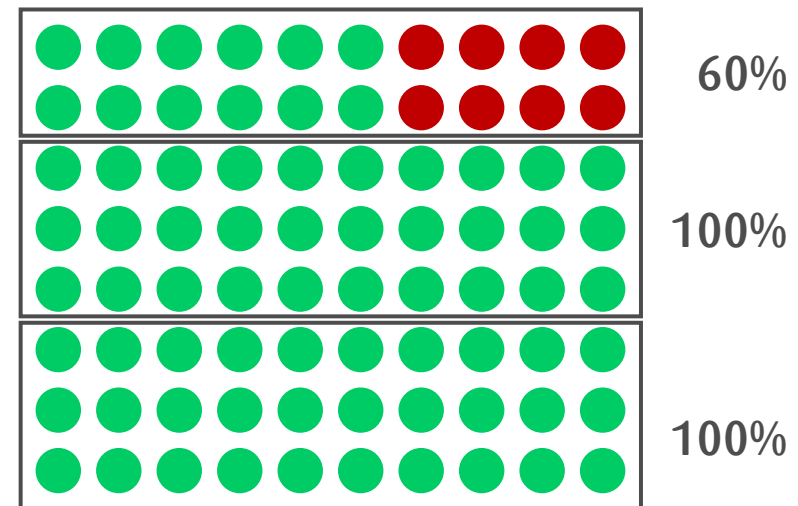
Conditional Coverage

The Problem with Marginal Coverage

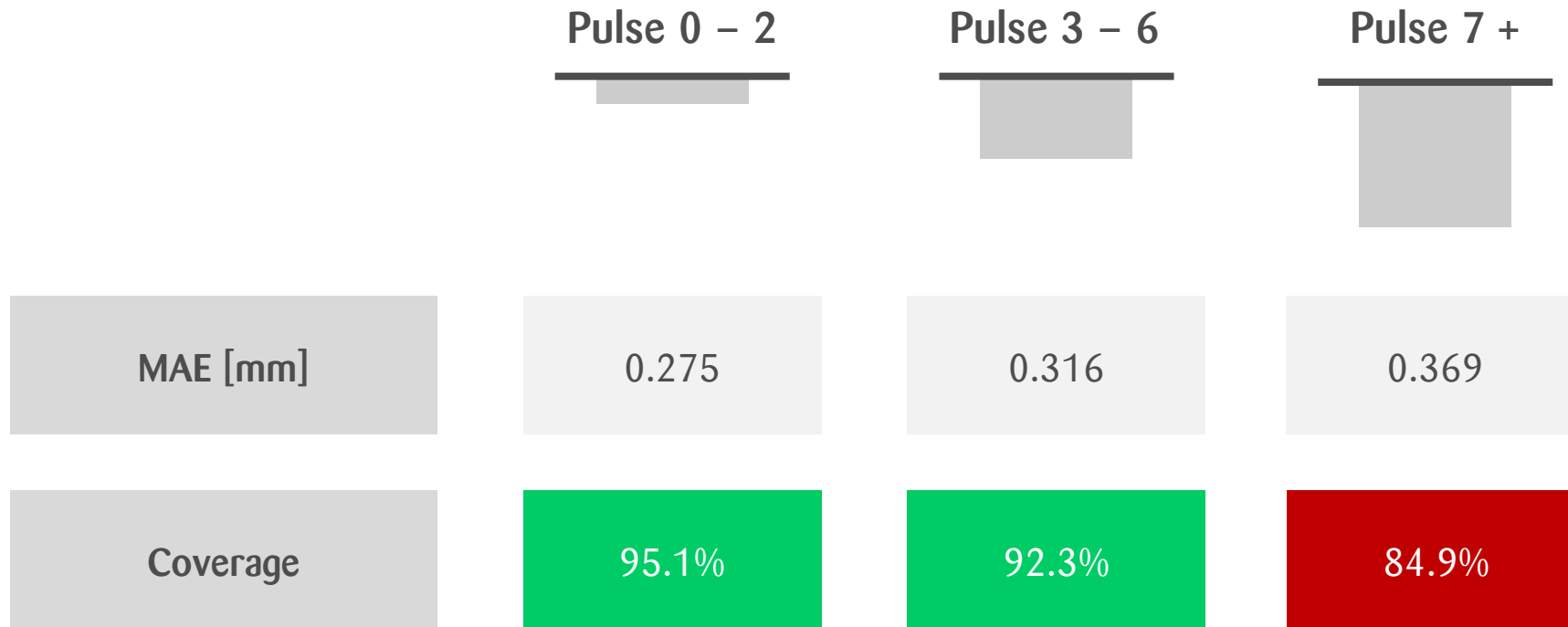
Marginal Coverage



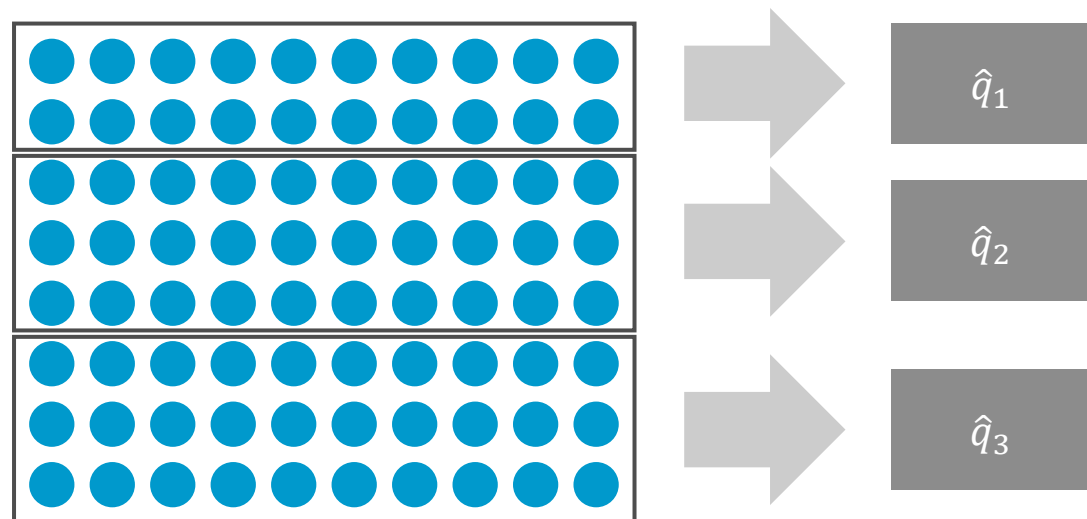
Conditional Coverage



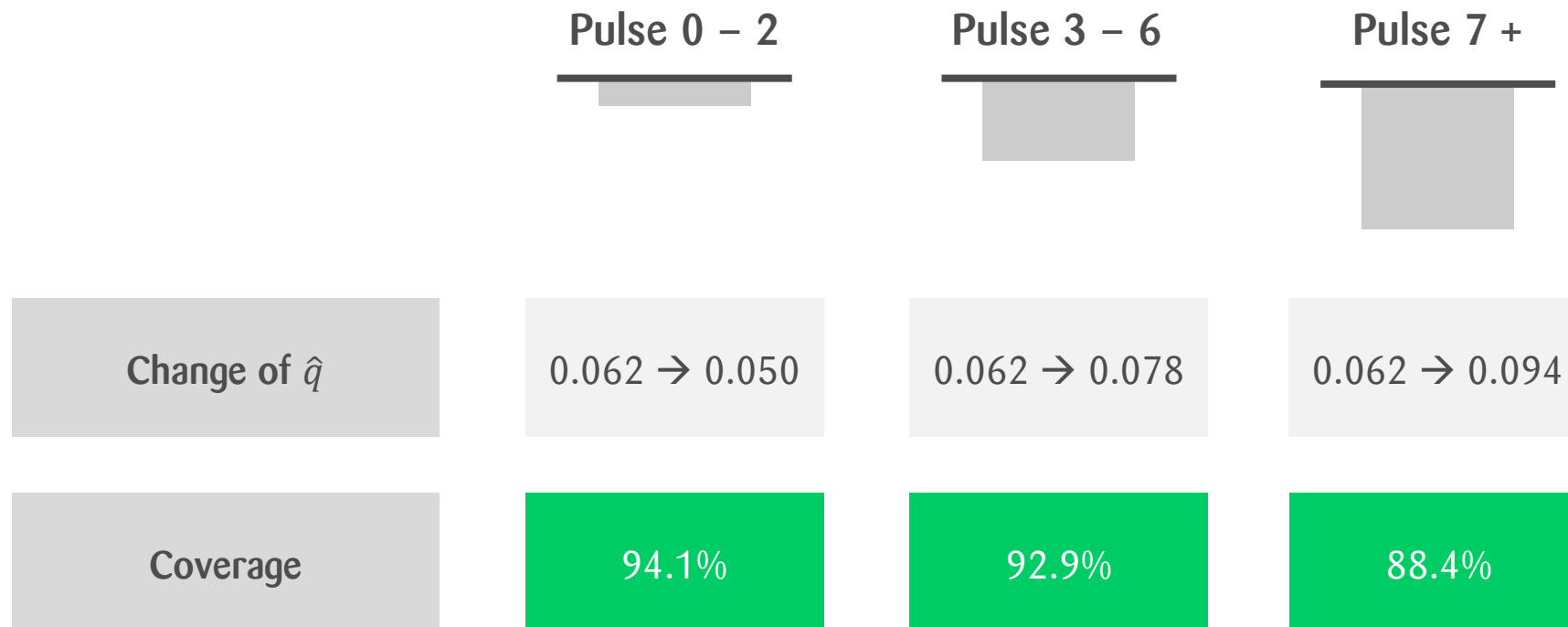
Results



Group-Balanced Conformal Prediction



Results





Conclusion

Conclusion

WE



CONFORMAL PREDICTION

Conformal predictions are highly flexible

- Distribution free
- Model & task agnostic

Conformal predictions are easy to use

- Simple recipes
- Great libraries

But they are no silver bullet

- Marginal coverage is not everything
- Avg. prediction set size is not everything
- Calibration set is critical

Thanks!

Conformal Prediction for Bone Surgery

23.06.2023



Appendix

There is much more

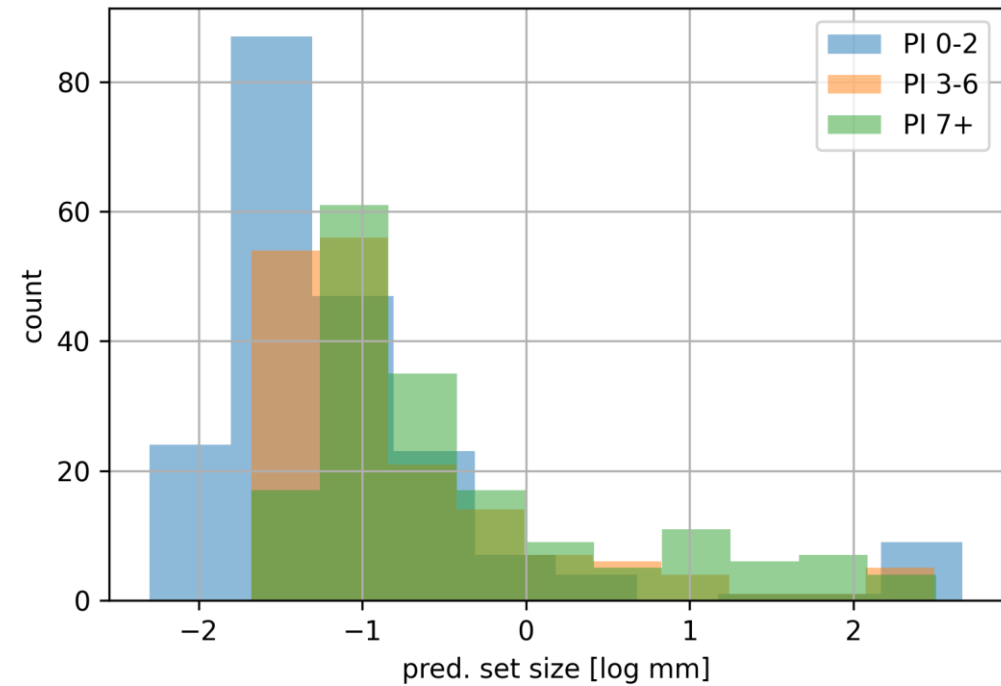
- Many more recipes (Regression, Classification, Bayes)
- Full CP (instead of split CP)
- Conformal Risk Control

- Great libraries, e.g. MAPIE

- Link to intro paper
 - Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint arXiv:2107.07511 (2021).

Results Group-Balanced Conf. Prep.

	avg. pred. set size [mm]
PI 0 – 2	0.89
PI 3 – 6	0.87
PI 7 +	1.30



Conformalized Mean Variance Estimation

Mean Variance Estimation (MVE)

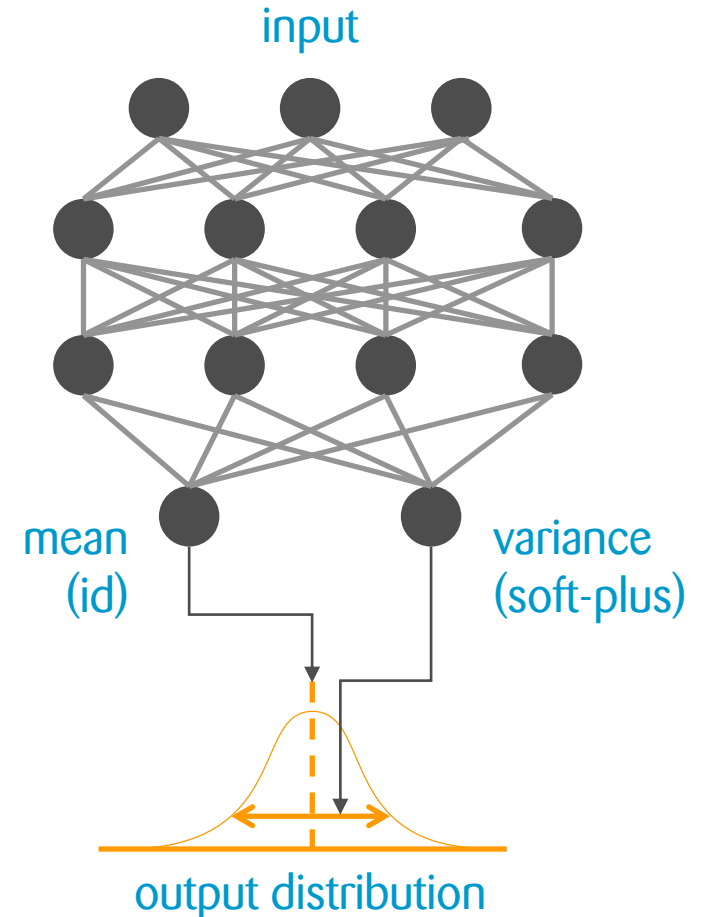
- Model output as Gaussian, predict mean and variance
- Usually trained by minimizing negative log-likelihood

Conformalized Mean Variance Estimation

- Score: $s = \frac{|y - f_\mu(x)|}{f_\sigma(x)}$
- Prediction set: $C(x) = [f_\mu(x) - \hat{q}f_\sigma(x), f_\mu(x) + \hat{q}f_\sigma(x)]$

Discussion

- MVE often not ideal, because Gaussian assumption violated
- Can do this for any kind of uncertainty estimate (ensembles, MC dropout, ...)
- However, quantiles (from quantile regression) in general a better approximation for prediction intervals than distribution-based methods



Adaptive Prediction Sets (APS)

Previous (warm-up) Example: “LABEL” method

- $s(x, y) = 1 - f(x)[y]$
- Smallest avg. prediction set size
- However, only uses softmax output of true class
 - Ignores a lot of information from the model
 - Tends to be not very “adaptive” (will see definition and examples later)

Adaptive Prediction Sets

- Score: Sum of softmax outputs from highest to correct class
- Prediction: Add classes, starting from highest confidence, until sum reaches threshold
- Intuition: If prob. of model were perfect, just greedily add classes until reach $1 - \alpha$
- (will revisit later)

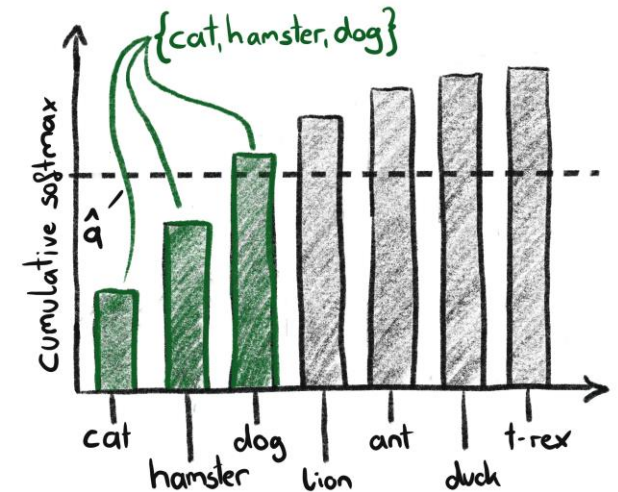
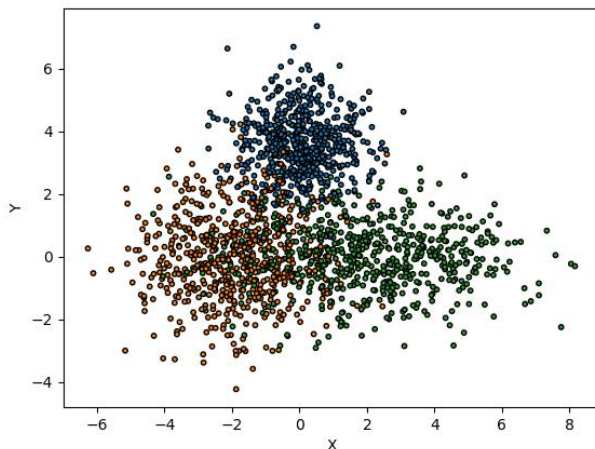


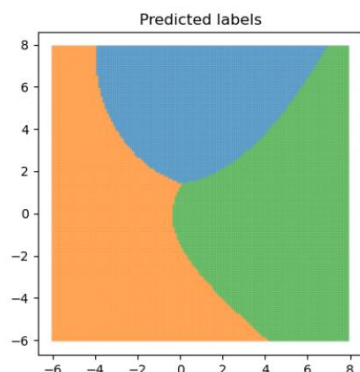
image credits: Christoph Molnar

Example: LABEL vs. APS

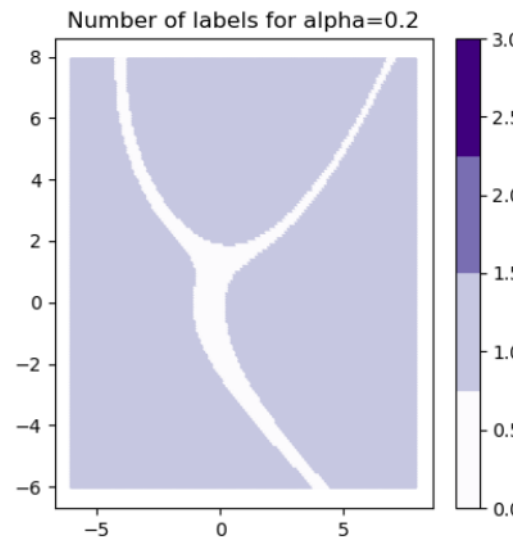
Data



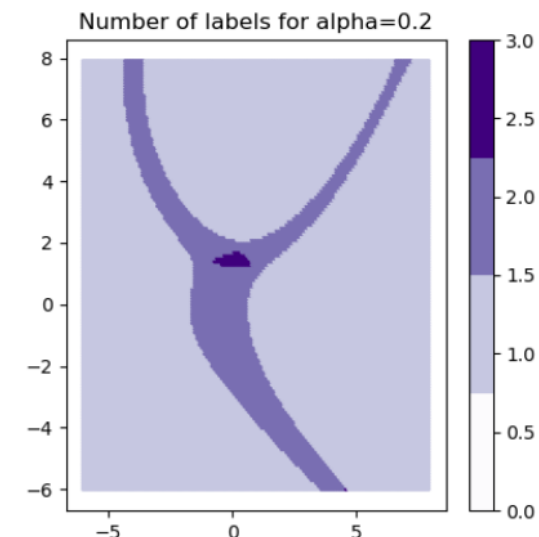
Model predictions



LABEL



APS



low avg. pred. set size



higher avg. pred. set size

set empty if ambiguity too high



set size > 1 if ambiguity too high

LABEL looks at true class only and decides to ignore if too uncertain. APS looks at all classes and sees that all of them uncertain

image credits: MAPIE docs

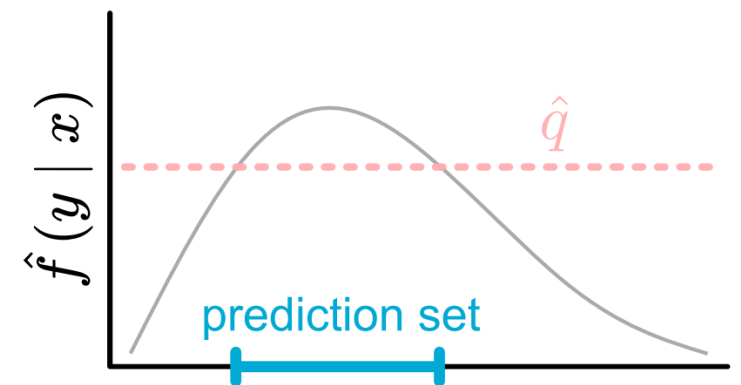
Conformalized Bayes

Bayes Modeling

- Use Bayes rule to estimate posterior predictive density $\hat{f}(y | x)$
- Usually first calculate distribution over model weights and then integrate over it for prediction
- In theory, already gives correct predictions
- However, in practice assumptions often violated, yielding wrong probabilities

Conformalized Bayes Modeling

- Score: $s(x, y) = -\hat{f}(y | x)$
- Prediction set: $C(x) = \{y | \hat{f}(y | x) > -\hat{q}\}$



Plot from Angelopoulos et al., 2022

Full Conformal Prediction

Split Conformal Prediction

- So far, we did “split conformal prediction”: split data in train and calibrate
- Non-optimal use of data
- Higher variance of prediction sets
- Ignores model refit variance
- But: Cheap (only train one model)

Full Conformal Prediction

- Train on all (!) the data
- To do so, train for every prediction multiple model on all the data + the new data point with every possible label
- Then modify recipe a bit to consider scores of all the trained models
- Very expensive, rarely used in practice

Other Data Splits

Tradeoffs

- Can do cross-validation or Jackknife
- Calculate scores on data-points which were left out in current training run

Result Aggregation

- “Standard”: Train a final model on all data
 - So single model for prediction only
 - But could lose coverage guarantee if model refit variance too high
- “Plus”: Use result of each individual model
 - Guarantee of $1 - 2\alpha$ if individual models have $1 - \alpha$
- “Minmax”: Use min and max of all models
 - Guarantee of $1 - \alpha$ if individual models have $1 - \alpha$

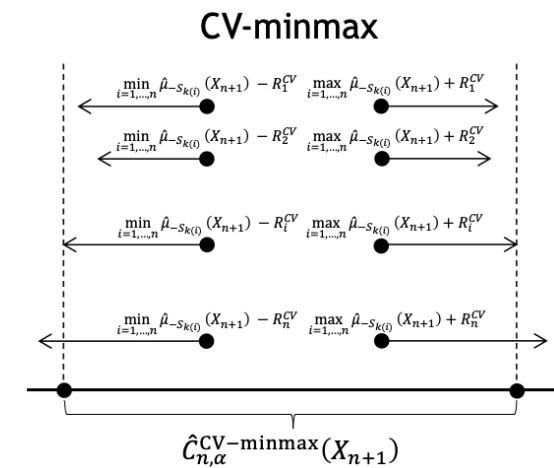
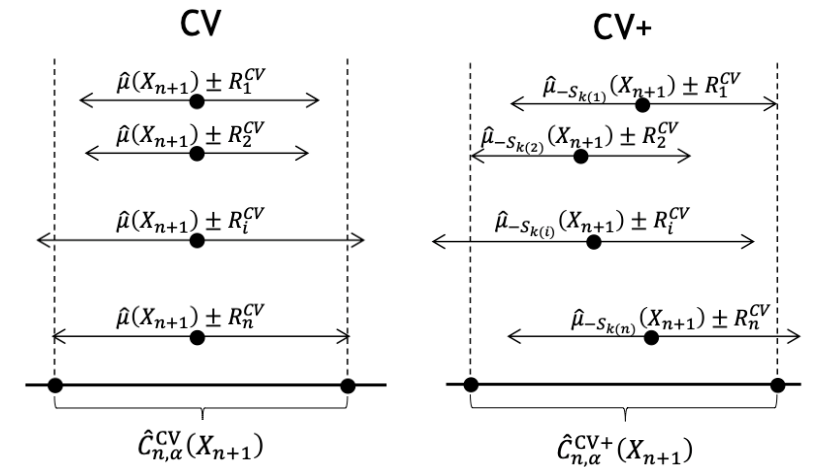


image credits: MAPIE docs

Conformal Risk Control

Current Focus: (Non-) Coverage Guarantee

$$P(y_{test} \notin C(x_{test})) < \alpha$$

Generalization

$$E[l(C(x_{test}), y_{test})] < \alpha$$

for any bounded loss function l

(miss-coverage loss $l(C(x_{test}), y_{test}) = I\{y_{test} \notin C(x_{test})\}$)

Examples

Multilabel classification

- Coverage does not really make sense
- For example, would like to give guarantee that large fraction of true classes contained in prediction set

Segmentation

- Could want to give guarantee on area recall

Outlier Detection

- Could want to give guarantee on false positive rate

Generalized Recipe

Need a post-processing of the model outputs

- Given model predictions, construct prediction set $C_\lambda(\cdot)$
- Loss function needs to be non-increasing in λ
 - The larger λ , the smaller the loss
 - Multilabel classification: Include all classes larger than $1 - \lambda$
(prediction set get larger, so fraction of covered true classes as well)
 - Segmentation: Include pixels larger than $1 - \lambda$
(prediction set gets larger, so area recall as well)
 - ...

Now do nearly the same as before

- Find value of λ which fulfills risk guarantee on calibration set
- Use this λ to construct prediction sets for test points