



Universität Augsburg  
Fakultät für Angewandte  
Informatik

# Calibrated and Uncertainty- aware Multimodal Emotion Recognition

Nicolas Kolbenschlager

Supervisor: Lukas Stappen, M.Sc.

Bachelor thesis, 2021

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)

# Agenda

---

- 1** Uncertainty in Deep Learning (Background)
- 2** Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3** Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4** Evaluation and Discussion (Conclusion)

# Uncertainty in Deep Learning

## Introduction to Uncertainty



Fig.: [https://s3-prod.autonews.com/MOBILITY\\_171039940\\_V2\\_-1\\_TUWUZFIKOQB.jpg](https://s3-prod.autonews.com/MOBILITY_171039940_V2_-1_TUWUZFIKOQB.jpg)

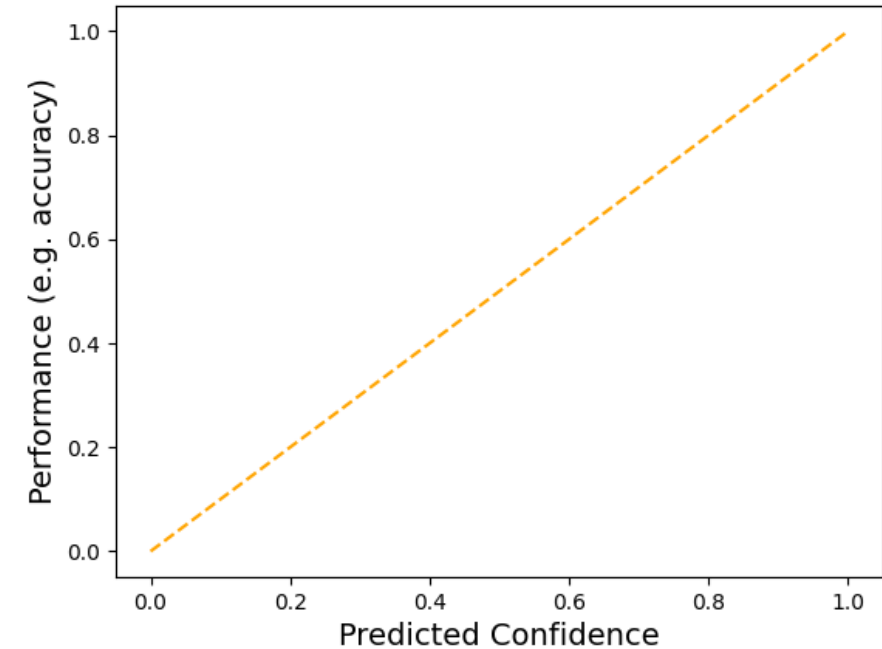


Fig.: *Well-calibrated* uncertainty measurements: the predicted confidence exactly matches the prediction error (here, represented by accuracy).

Fig.: [https://cdn.meine-vm.de/jNQral5caPE3wbl9LnbdFH\\_MO0=/480x240/smart/819%2F23096%2F23096806%2F49963577.jpg](https://cdn.meine-vm.de/jNQral5caPE3wbl9LnbdFH_MO0=/480x240/smart/819%2F23096%2F23096806%2F49963577.jpg)



# Uncertainty in Deep Learning

---

## Types and Sources of Uncertainty

- **Aleatory** (*Data*) Uncertainty:
  - Caused by noise in the sample.
  - Image of a dog and a cat next to each other fed into a cat-vs.-dog classifier.
- **Epistemic** (*Model/Knowledge*) Uncertainty:
  - Appears if the model lacks knowledge on a sample.
  - E.g., the image of a ship, fed into a cat-vs.-dog-classifier.

# Uncertainty in Deep Learning

## Measuring Predictive Uncertainty

- Classification: softmax
- Bayesian modelling:
  - Obtain multiple forecasts for  $\hat{y}$  (with possibly different outcomes).
  - Use the mean of them as final prediction and the **variance** among them **as measurement for uncertainty**.
  - Approaches: Monte Carlo Dropout, Ensemble Averaging, ...
- Distributional parameter estimation:
  - Output parameters for probability distribution over  $\hat{y}$  instead of  $\hat{y}$ .

# Uncertainty in Deep Learning

## Calibration

- Often predicted uncertainty is not well-calibrated initially.
- *(Re-) calibration*: make non well-calibrated uncertainties well-calibrated.
  1. Uncalibrated measurements of uncertainty  $\hat{U}$  from a hold-out validation set.
  2. Calculate true uncertainty  $U$  of the validation set.
  3. Train an auxiliary model, the *calibrator*, to map  $\hat{U}$  to  $U$ .
  4. Calibrator can be applied on uncalibrated uncertainty quantification obtained from the test set.
- Often: Isotonic Regression

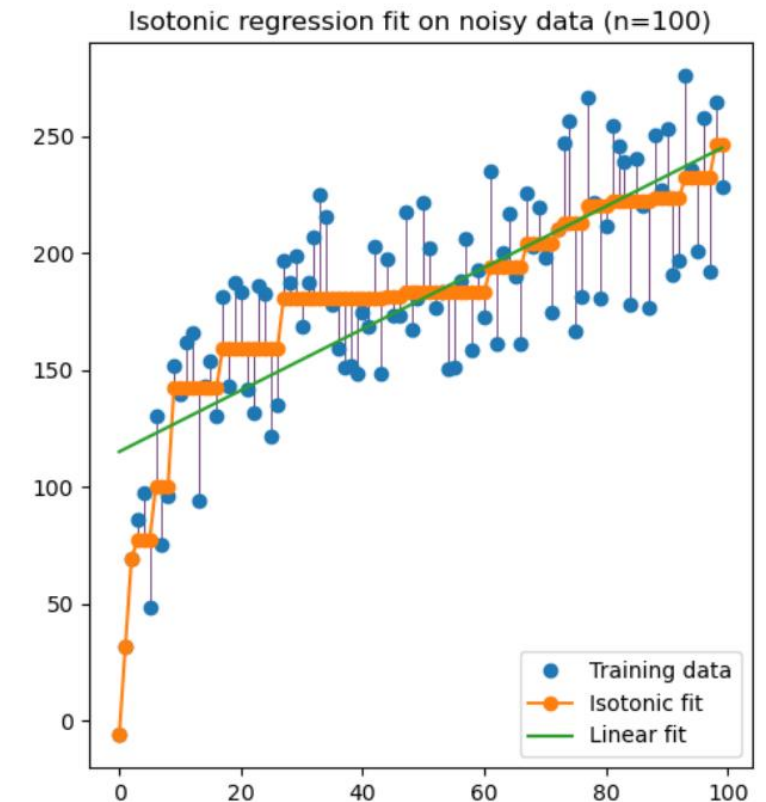


Fig.: Isotonic Regression. Image taken from: [https://scikit-learn.org/stable/auto\\_examples/miscellaneous/plot\\_isotonic\\_regression.html](https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_isotonic_regression.html)

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)



# Obtaining Estimates of Predictive Uncertainty

## Research Questions

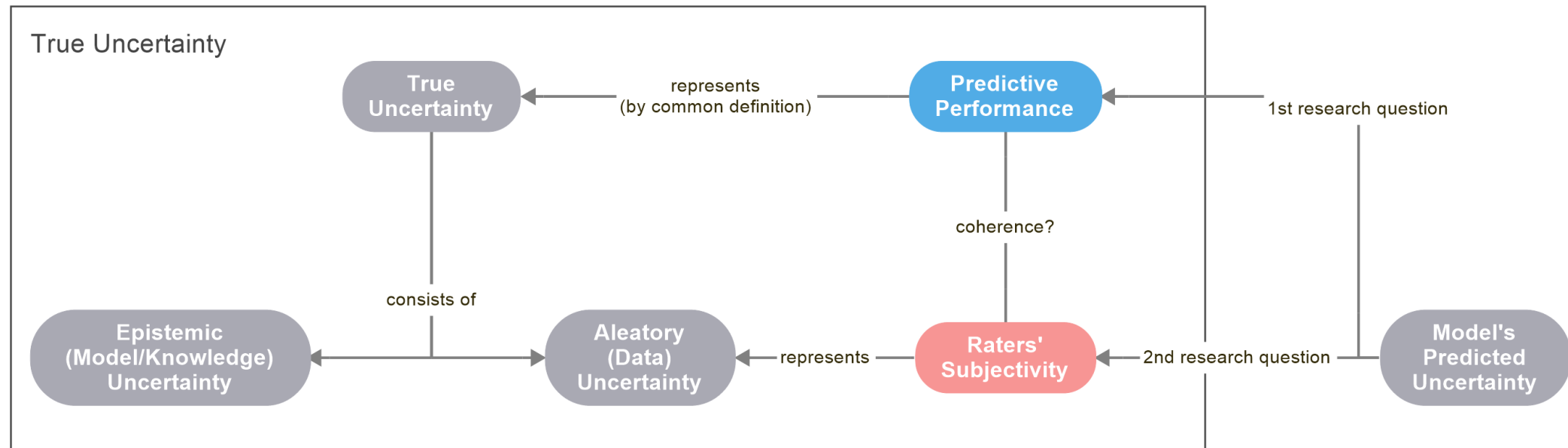


Fig.: Our research questions visualized

# Predictive Uncertainty, Subjectivity and Emotion Recognition

## The MuSe-CaR dataset

- Contains YouTube videos of car reviews contiguously (one label per time step) annotated by valence and arousal.
- For each sample, there are available **5 annotations by different human annotators**.
- The rates used joysticks for annotation while watching the videos.
- Further, there exists a gold-standard, calculated from the multiple annotations by fusion techniques.

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)

# Obtaining Estimates of Predictive Uncertainty

## Underlying Model and Modality

- Predict either valence or arousal
- Modality and used feature sets:
  - Valence: BERT
  - Arousal: VGGish
- The model:
  - Attention  $\rightarrow$  LSTM  $\rightarrow$  Fully Connected

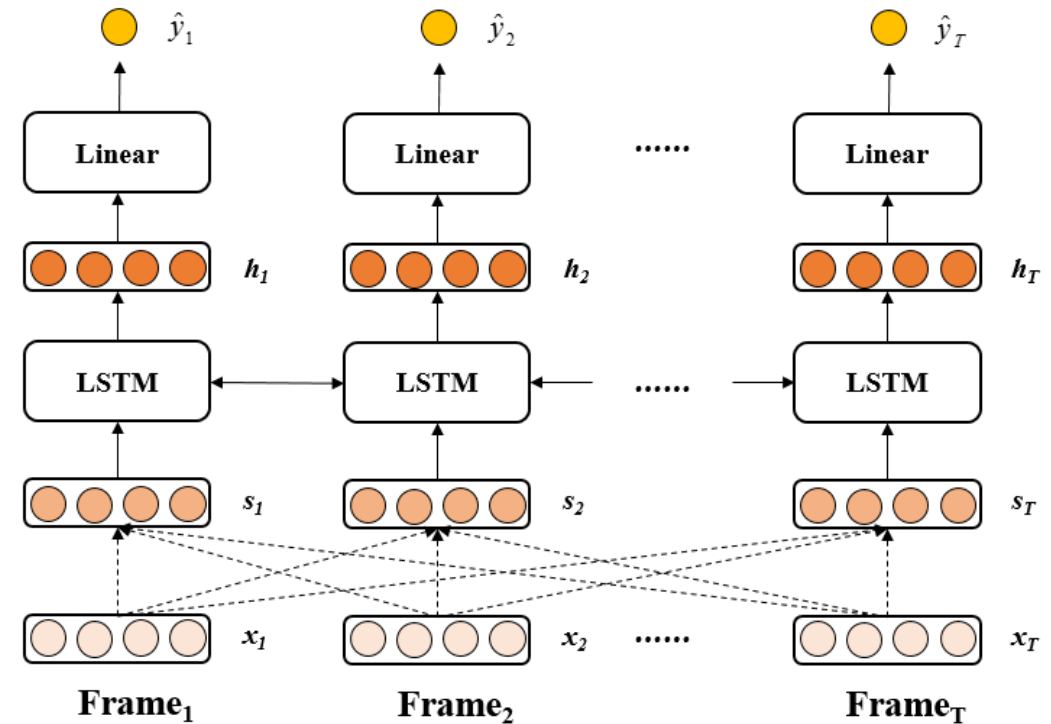


Fig.: The underlying model, which we equipped with techniques for confidence quantification

# Obtaining Estimates of Predictive Uncertainty

## Before we start: Replacing Variance by Correlation

- Bayesian Modelling: variance to obtain confidence.
- But we predict steps in time of a **contiguous** emotional state.
- Each point in time  $t$  is rated based on the evaluation of the signal at  $t$  and the annotation of the latest time step  $t - 1$ .
- Now: ~~variances~~ → correlations:
  - Obtain multiple (differing) predictions, as usual.
  - Compute the **average correlation** (PCC or CCC) **among each pair** of them.
  - The higher the correlations, the higher the model's confidence.

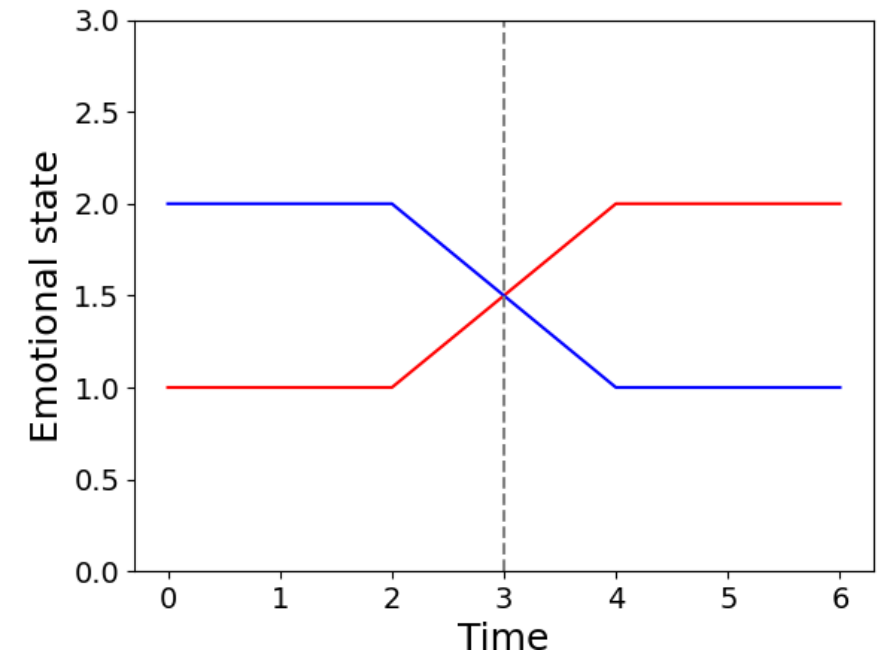


Fig.: At  $x = 3$  both lines have the same value. But when it comes to measuring the disagreement of them, the correlation at this point is much more expressive, especially because the shown values at each time step come from one contiguous measurement. Each point in time is not a standalone prediction.

# Obtaining Estimates of Predictive Uncertainty

## Overview over Concepts

- Two main conceptional approaches:
  - Local uncertainty quantification: for now, anything happens **per time step**, or *locally*.
  - Global uncertainty quantification: measurements are done for **multiple time steps** (sub-samples, respectively) **together**, or *globally*.

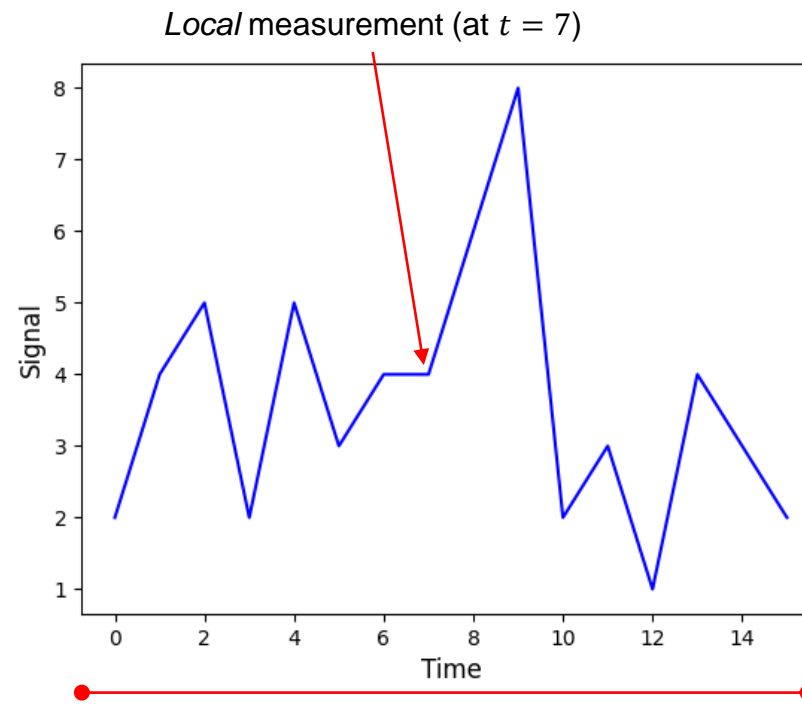


Fig.: *Global* and *local* measurements

# Obtaining Estimates of Predictive Uncertainty

## Local Uncertainty Quantification: Definitions of True Uncertainty

### Predictive Performance

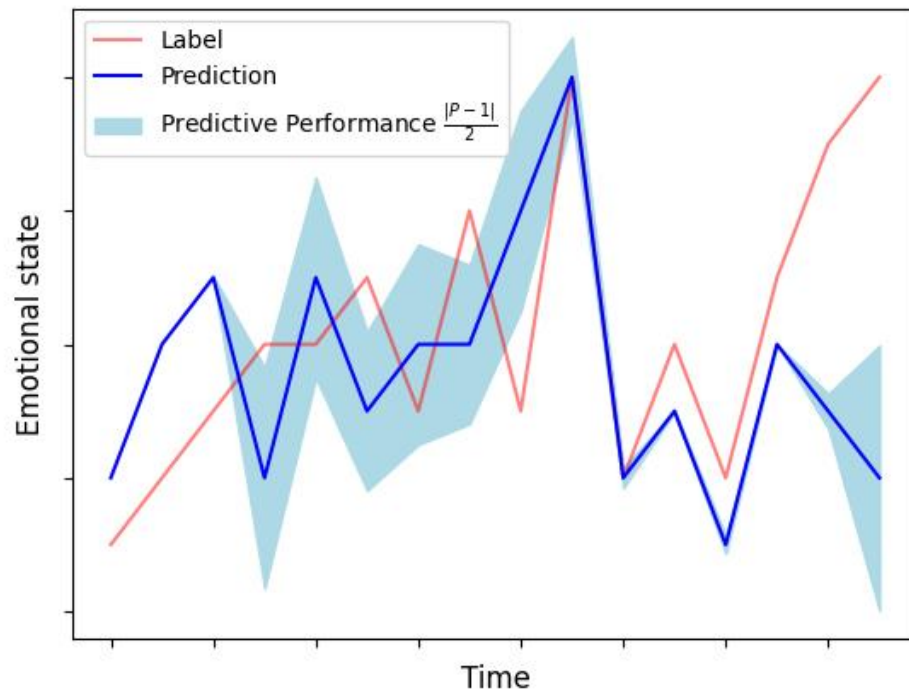


Fig.: Predictive Performance of a predictions in terms of the label

### Subjectivity among raters

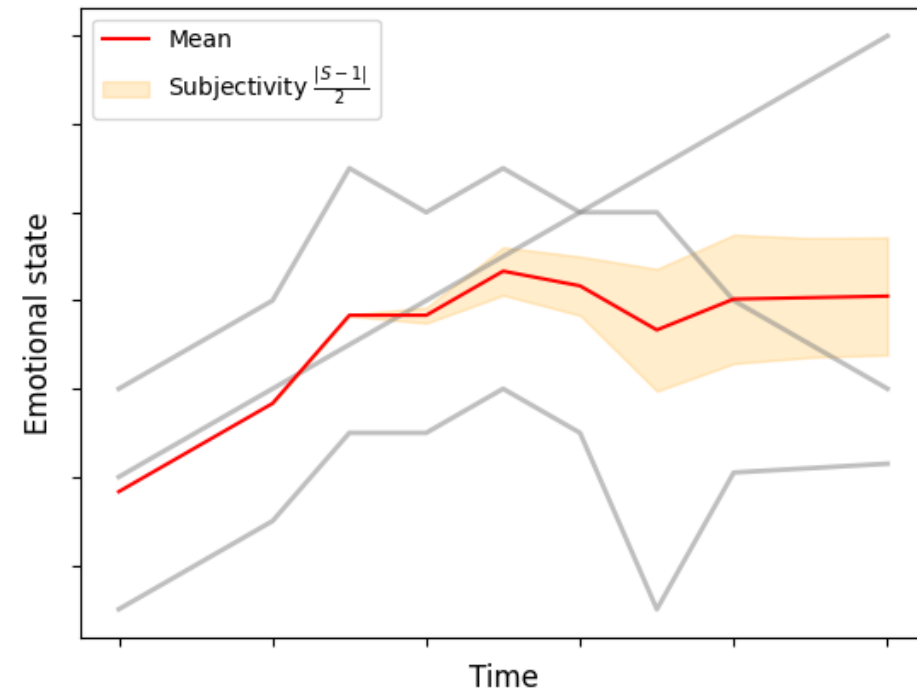


Fig.: Subjectivity among multiple (gray) raters

# Obtaining Estimates of Predictive Uncertainty

## Local Uncertainty Quantification: Definitions of Predicted Uncertainty

- Monte Carlo Dropout
- Proposing *tilted CCC (tCCC)*:
  - Mid node: usual CCC loss criterion
  - Outer nodes:  $RCE_3$  and  $RCE_{10}$  as criterion ( $n$ : sample length).
- Creates mini-ensemble with 2 nodes, focusing on different errors.

$$RCE_w(\hat{y}) = \frac{1}{n} \sum_{j=w}^n (1 - PCC(\hat{y}_{j-w+1} \dots \hat{y}_j, y_{j-w+1} \dots y_j))$$

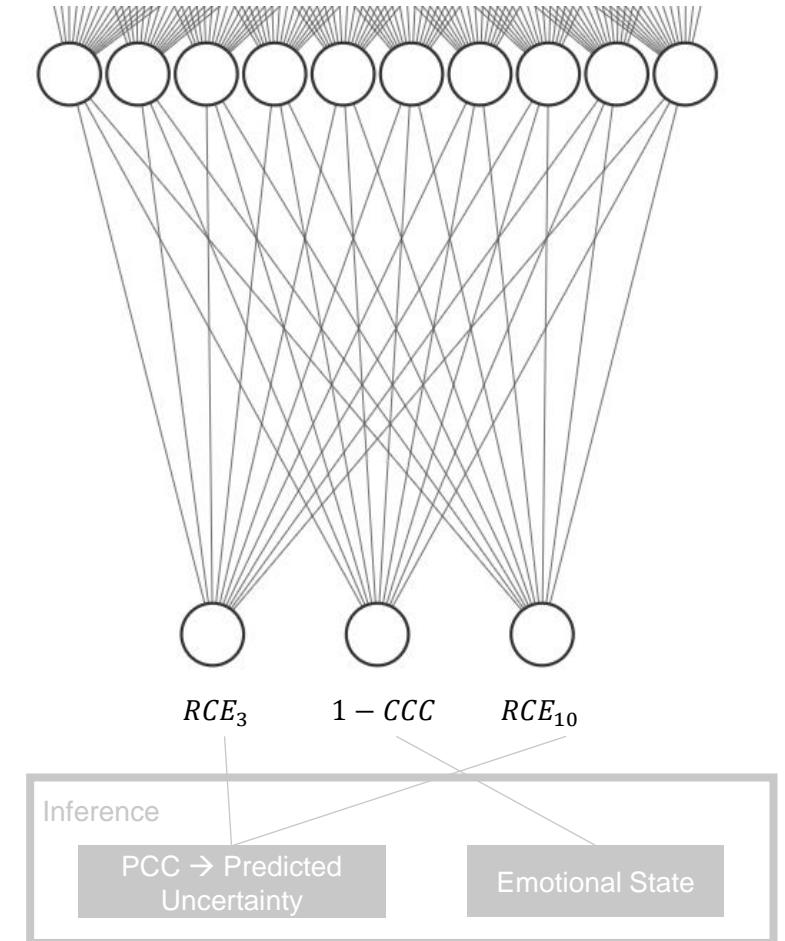


Fig.: Losses and inference of tCCC visualized



# Obtaining Estimates of Predictive Uncertainty

## Local Uncertainty Quantification: Results

- Quantified uncertainty did not match true uncertainty, whether Subjectivity nor Predictive Performance.
- Conceptual issue:
  - Monte Carlo Dropout: model optimized using CCC loss → evaluates the complete sample, not per time step.
  - During training with tCCC, the outer nodes did not converge significantly.
  - → Emotion recognition (here) works more globally.
  - **(Local) forecasts prone to be wrong (so meaningless) → they inherently lead to non-reliable confidence estimates.**

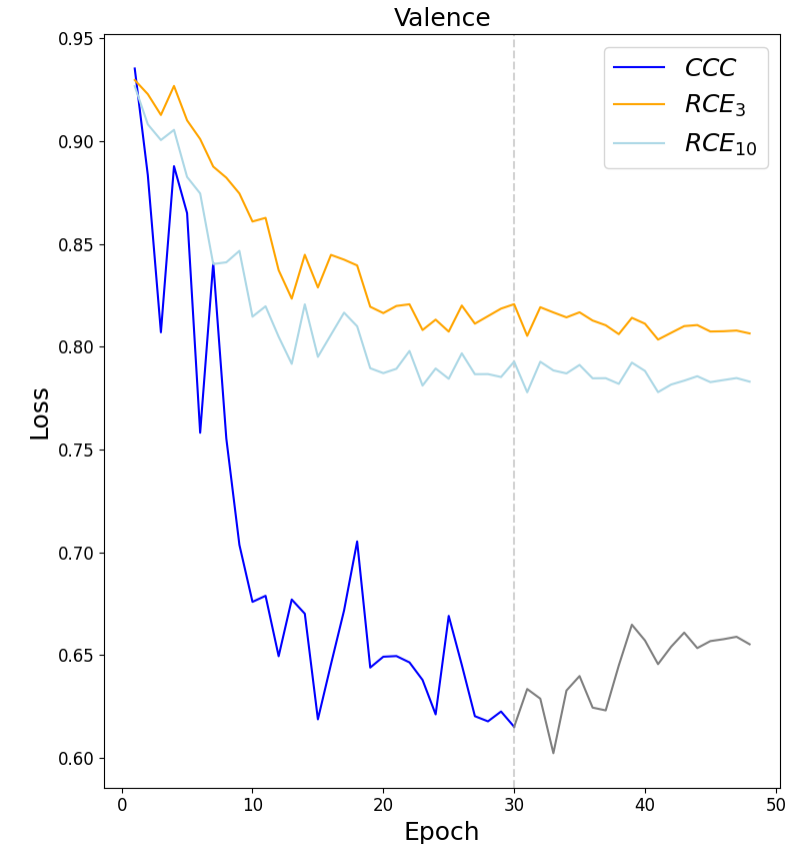


Fig.: tCCC node convergence

# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Definitions

- Instead of measuring one value for uncertainty per time step, we now define uncertainty for complete samples.
- As we expect uncertainty to vary over time, we split up samples into non-overlapping sub-samples of length 20 (5 seconds).

### True Uncertainty (for one sub-sample)

- **Predictive Performance**

$$P = CCC(y, \hat{y})$$

- With label  $y$  and prediction  $\hat{y}$ .

- **Subjectivity among raters**

$$S = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 CCC(x_i, x_k)$$

- With  $x_i$ , the annotation of the  $i$ -th rater.

### Predicted Uncertainty (for one sub-sample)

- Monte Carlo Dropout
- Ensemble Averaging
- → Correlation among differing prediction as measurement for model's confidence.

# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Results

True Uncertainty	Emotional Dimension	Method	CCC devel / test [uncalibrated]	CCC devel / test [Isotonic Regression]
Predictive Performance	Valence	Ensemble Averaging	<b>0,18 / 0,27</b>	0,13 / 0,16
		Monte Carlo Dropout	0,18 / 0,24	0,16 / 0,17
	Arousal	Ensemble Averaging	<b>0,23 / 0,12</b>	0,22 / 0,11
		Monte Carlo Dropout	0,20 / 0,05	0,27 / 0,08
Raters' Subjectivity	Valence	Ensemble Averaging	0,03 / 0,04	0,02 / 0,03
		Monte Carlo Dropout	0,03 / 0,04	0,02 / 0,03
	Arousal	Ensemble Averaging	0,08 / 0,04	<b>0,13 / 0,06</b>
		Monte Carlo Dropout	0,04 / 0,01	0,06 / 0,02

Verification: Predictive Performance and subjectivity correlate.

# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Results

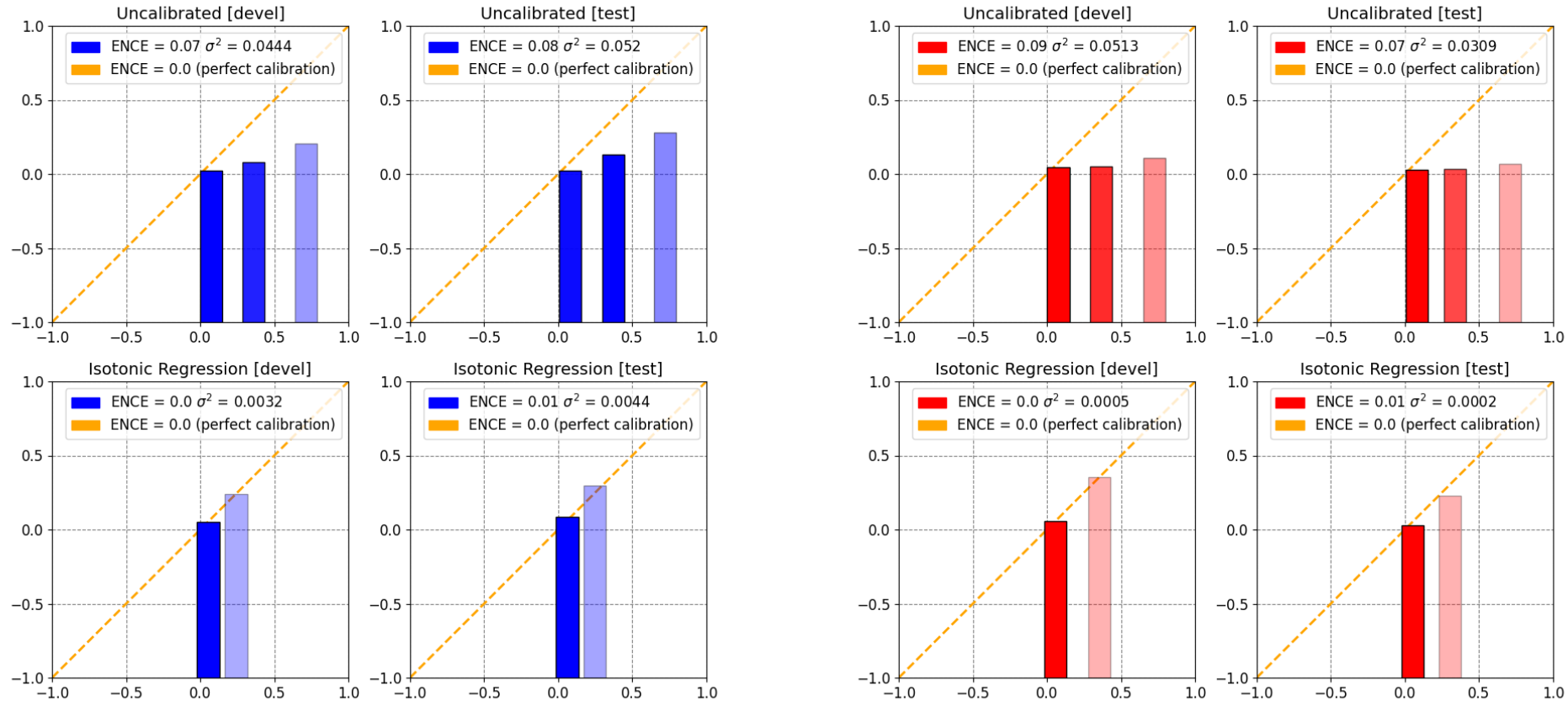


Fig.: Some results visualized with Reliability Diagrams. Predictive Performance as true confidence with Ensemble Averaging for valence (left) and Subjectivity as true confidence with Ensemble Averaging for arousal (right). As more transparent a bin, the less (sub-) samples fall into it.

# Agenda

---

- 1 Uncertainty in Deep Learning (Background)
- 2 Predictive Uncertainty, Subjectivity and Emotion Recognition (Research Goals)
- 3 Obtaining Estimates of Predictive Uncertainty (Approach and Results)
- 4 Evaluation and Discussion (Conclusion)

# Evaluation and Discussion

## Summary

- We proposed techniques to transfer the procedure of uncertainty quantification to contiguous forecasting, by replacing variance with correlations.
- We defined **Subjectivity among multiple (human) raters** as additional comparative value for a model's confidence, next to the **Predictive Performance** of the model.
- Initial assumption: *Prediction Error*  $\sim^{(1)}$  *Predicted Uncertainty*  $\sim^{(2)}$  *Raters' Subjectivity*
- As experiments in local uncertainty failed because of conceptual issues, we restrict further thoughts on global uncertainty measurements.

# Evaluation and Discussion

---

## Conclusion: first research question

- Observed coherence between **Predictive Performance** and predicted uncertainty.
- The correlation of about 0.2
  - Predictive performance unites each source of uncertainty
  - No matter what underlying reasons cause the model's uncertainty, they should reflect on its predictive performance → correlation too weak?!
- **Uncertainty isn't an explicit forecast → implicit measurement → anything better than approximation would be weird.**

# Evaluation and Discussion

## Conclusion: second research question

- As for the coherence between **Subjectivity among annotations** and predicted uncertainty, it is, if at all, barely available.
- Subjectivity is only one source of uncertainty, among others.
- Subjectivity among raters indicates a noisy sample (as features did not transport an unambiguous signal), so aleatory uncertainty.
- So, what about epistemic uncertainty?
- Epistemic uncertainty is caused by missing knowledge of the model about the fed in data, which is not contained in raters' subjectivity.
- The weaker correlations between Subjectivity among annotations and predicted uncertainty, than those between predictive performance and predicted uncertainty, might be explained by the disregard of epistemic uncertainty in this definition of true uncertainty.



# Evaluation and Discussion

---

## Future Work

1. Vary the length of sub-samples (not just 20)
  - First experiments with lengths 10 and 50 do **not** indicate a significant improvement.
2. Use advanced fusion technique to obtain final prediction from Monte Carlo Dropout and Ensemble Averaging:
  - This would reflect on Predictive Performance and make it, as a definition of true confidence, more reliable.
3. Epistemic Uncertainty:
  - Furthermore, we could think about manually quantifying the lack of knowledge for a sub-sample and use this as a definition for true uncertainty, as we did with subjectivity among annotations.

# Evaluation and Discussion

Raters' subjectivity is independent of the feature set!

## Future Work: Incorporating Epistemic Uncertainty

- Simpler feature embeddings (valence on fastText, instead of BERT): try to observe, if this increases predicted uncertainty and decreases the proportion of aleatory uncertainty relative to the overall uncertainty.
- Less expressive features → expect **larger general model's uncertainty** and relatively more epistemic uncertainty, so **relatively less aleatory uncertainty**, as overall uncertainty decomposes into aleatory and epistemic uncertainty.

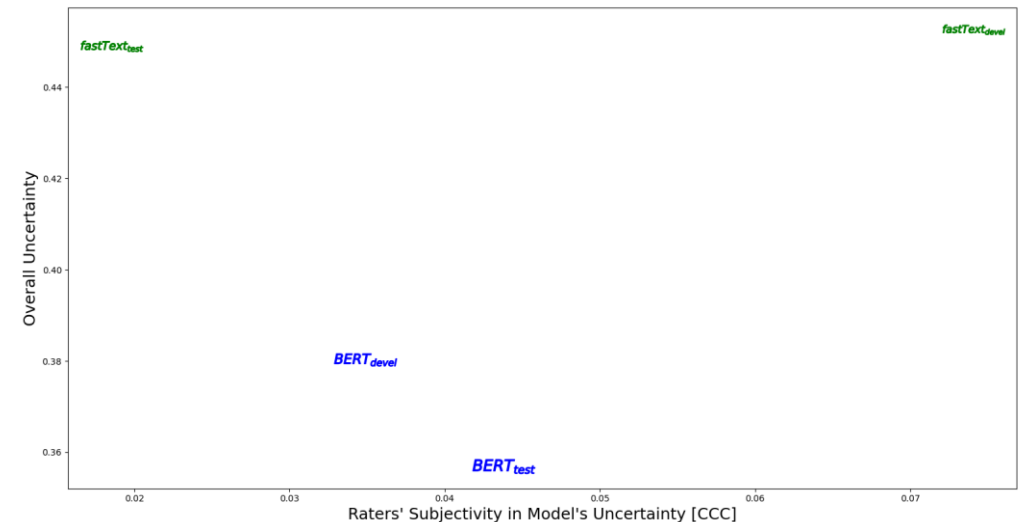
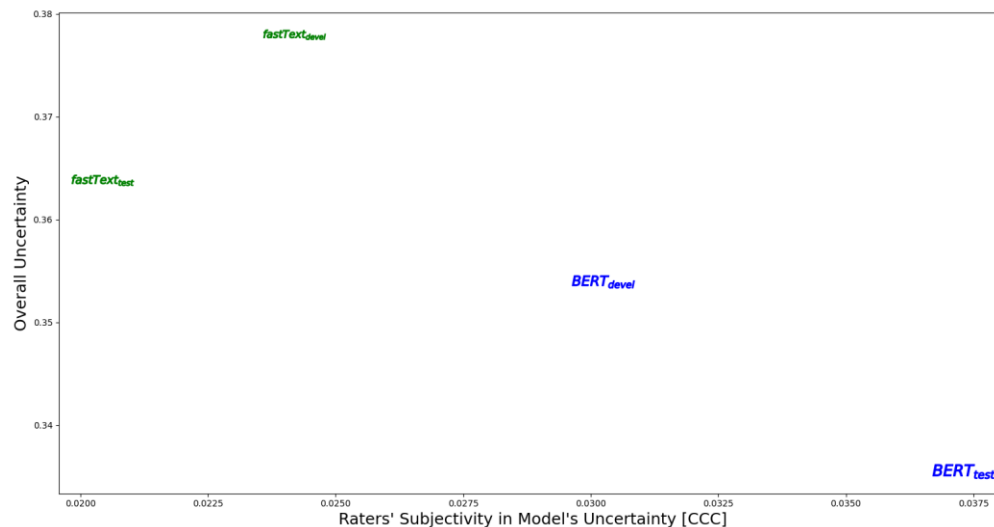


Fig.: Comparing predicted uncertainty with different features (left: Monte Carlo Dropout, right: Ensemble Averaging).



*Thank you for your attention!*

*Any questions?*



Nicolas Kolbenschlager  
Universität Augsburg  
nicolas.kolbenschlager@student.uni-augsburg.de  
www.uni-augsburg.de

# Backlog

# Predictive Uncertainty, Subjectivity and Emotion Recognition

## Uncertainty-aware Emotion Recognition

- Aims of this work:
  1. Measure predictive uncertainty for the task of emotion recognition.
  2. Investigate subjectivity among annotations as source of predictive uncertainty.
- **Emotion Recognition:** prediction of the emotional state, here described by valence and arousal.
- Motivation:
  - We **want uncertainty** to appear where the **prediction error** is large (common definition).
  - But why should it indeed appear there? Or, why actually is there prediction error, if it is?
  - Model can't know prediction error during inference → underlying sources that cause uncertainty.
  - We investigate **subjectivity among annotations** as indicator for where we **expect uncertainty**.

# Obtaining Estimates of Predictive Uncertainty

## Local Uncertainty Quantification: Formalization

### True Uncertainty (at time step j)

- **Predictive Performance** (left figure)

$$P_j = PCC(y_{j-t} \dots y_j, \hat{y}_{j-t} \dots \hat{y}_j) \text{ with } t = 2$$

- With the prediction for the emotion  $\hat{y}_i$  and the label  $y$ .

- **Subjectivity among raters** (right figure)

$$S_j = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 PCC(x_{i,j-t} \dots x_{i,j}, x_{k,j-t} \dots x_{k,j-t}) \text{ with } t = 2$$

- With  $x_i$ , the annotation of the i-th rater.

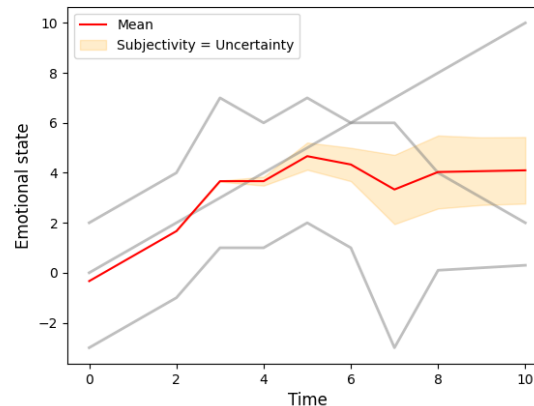
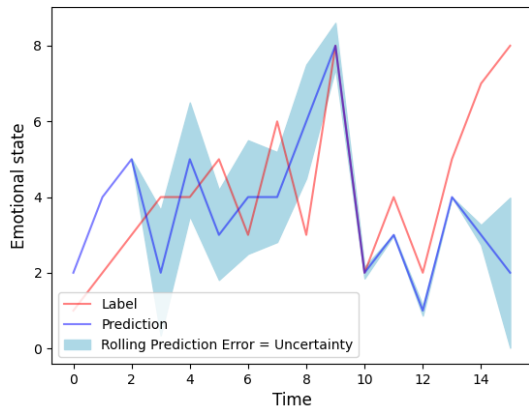


Fig.: Predictive Performance (left) and subjectivity among raters (right).

### Predicted Uncertainty (at time step j)

- Monte Carlo Dropout

$$MCD_j = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 PCC(\hat{y}_{i,j-t} \dots \hat{y}_{i,j}, \hat{y}_{i,k-t} \dots \hat{y}_{k,j}) \text{ with } t = 2$$

- With the model's i-th prediction for the emotional state  $\hat{y}_i$ .

- Proposing *tilted* CCC (tCCC)

- Network has three output nodes:

- Mid node: usual CCC loss criterion (used for inference of emotional state).
- Outer nodes:  $RCE_3$  and  $RCE_{10}$  as criterion ( $n$ : sample length).

$$RCE_w(\hat{y}) = \frac{1}{n} \sum_{j=w}^n (1 - PCC(\hat{y}_{j-w+1} \dots \hat{y}_j, y_{j-w+1} \dots y_j))$$

- Creates mini-ensemble with 2 nodes, focusing on different errors (rolling correlation error over the 3 and 10 latest time steps, respectively).
- Used to obtain uncertainty quantification, by computing correlation between them (similar to  $P_j$ ).

# Obtaining Estimates of Predictive Uncertainty

## Global Uncertainty Quantification: Formalization

- Instead of measuring one value for uncertainty per time step, we now define uncertainty for complete samples.
- But, as we expect uncertainty to vary over time, we split up samples into non-overlapping sub-samples of length 20 (5 seconds).

### True Uncertainty (for one sub-sample)

- **Predictive Performance**

$$P = CCC(y, \hat{y})$$

- With label  $y$  and prediction  $\hat{y}$ .

- **Subjectivity among raters**

$$S = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 CCC(x_i, x_k)$$

- With  $x_i$ , the annotation of the  $i$ -th rater.

### Predicted Uncertainty (for one sub-sample)

- **Monte Carlo Dropout**

$$MCD = \frac{1}{10} \sum_{i=1}^5 \sum_{k=i+1}^5 CCC(\hat{y}_i, \hat{y}_k)$$

- With the model's  $i$ -th prediction for the emotional state  $\hat{y}_i$ .

- **Ensemble Averaging**

- Exact the same, as Monte Carlo Dropout, only with forecasts  $\hat{y}_i$  from five different models/seeds, instead of five forward runs.

# Correlation Metrics

- Pearson Correlation Coefficient (PCC):

$$PCC(x, y) = \frac{Cov(x, y)}{\sqrt{\sigma^2(x)}\sqrt{\sigma^2(y)}}$$

- Concordance Correlation Coefficient (CCC):

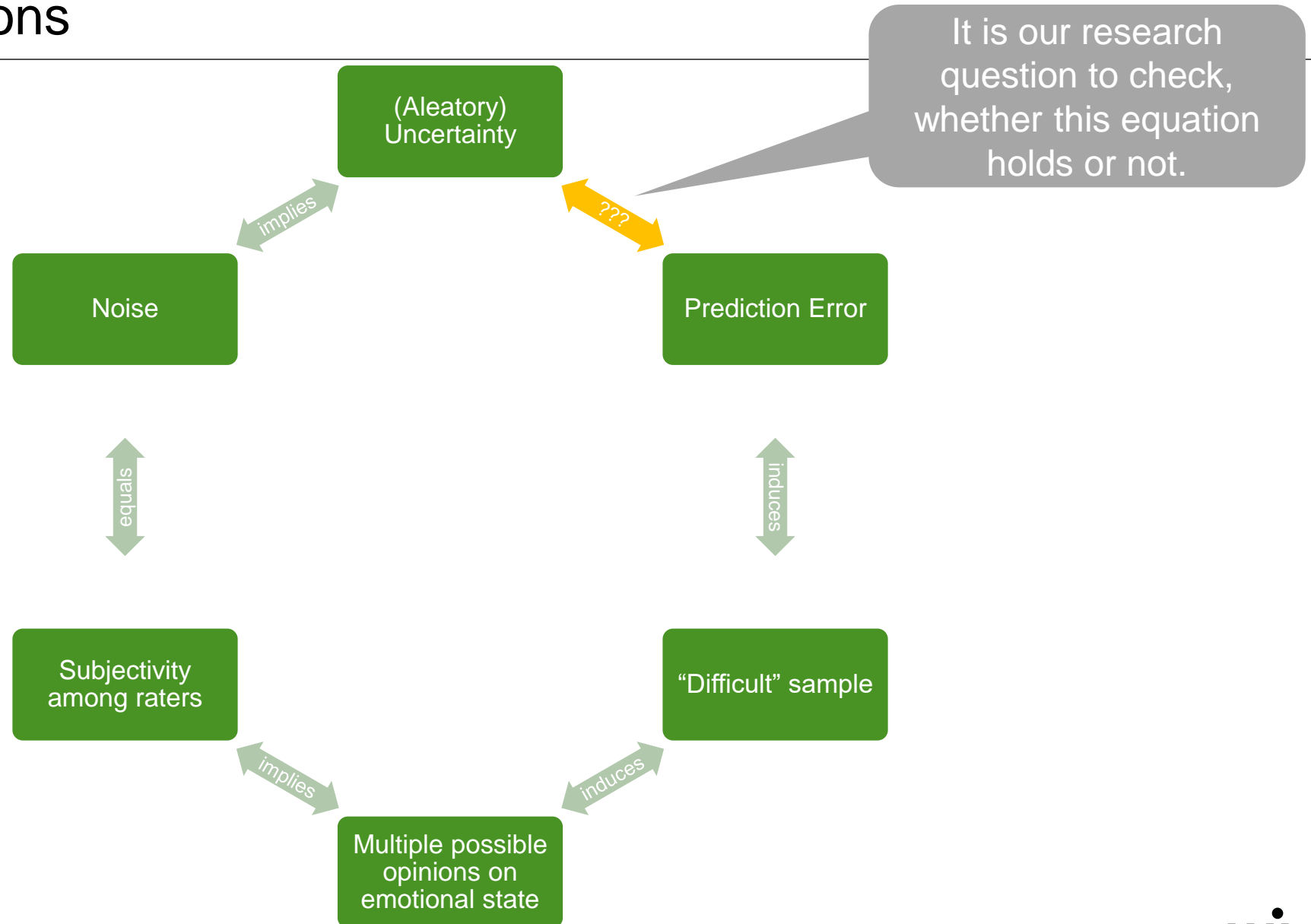
$$CCC(x, y) = \frac{2Cov(x, y)}{\sigma^2(x) + \sigma^2(y) + (\mu(x) - \mu(y))^2}$$

Notation: the mean  $\mu$ , variance  $\sigma^2$  and  $Cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu(x))(y_i - \mu(y))$

Both are limited to  $[-1, 1]$  with -1 indicating total negative correlation and +1 perfect correlation.



# Research Questions



# tCCC Node Convergence

