Oregon State University OSU

Toward Explainable Uncertainty

Alan Fern and Tom Dietterich

School of Electrical Engineering and Computer Science
Oregon State University

Goal 1: Uncertainty Aware ML Systems

- Design machine learning systems that "know what they know"
  [Li, Littman, Walsh ICML'08]
  - Provide guarantees on predictions
  - Allow systems to abstain and/or produce ambiguous predictions

- Achieve this in:
  - Closed Worlds = Known Unknowns
  - Open Worlds = Unknown Unknowns

- Why?
  - Safe and Trustworthy AI
  - End User Acceptability
  - Computational Efficiency – use more complex model if simpler model is uncertain

Goal 2: Transparent Uncertainty in ML Systems

- Design machine learning systems that can "explain their uncertainty"
  - Give insight into why they abstained or produced ambiguous answer

- Achieve these goals in:
  - Closed Worlds = Known Unknowns
  - Open Worlds = Unknown Unknowns

- Why?
  - Basis for feedback to learning systems
  - Basis for investigating anomalies
  - Mechanism for building trust

Outline

- Conformal Prediction for Uncertainty Aware Classification
  - Empirical performance in closed worlds
  - Empirical performance in open worlds
  - Not effective in open worlds → Suggests integrating with anomaly detection

- Explanations for Anomaly Detection
  - What is an anomaly explanation?
  - How to compute explanations?
  - How to evaluate explanations?

Standard Classification

Classes: {Human, Animal, Hybrid}

Predictor

Animal

Human

Human

Uncertainty in the system is not made explicit in the predictions.

Conformal Prediction [Vovk et al., 2005]

Most basic type of explanation of uncertainty

Classes: {Human, Animal, Hybrid}

Conformal Predictor

{**Animal**, Hybrid}

{Animal, Human, **Hybrid**}

{**Human**}

- Conformal predictors output sets of labels.
- Label set is correct if it contains true label.

Conformal Prediction: Accuracy

Classes: {Human, Animal, Hybrid}

Conformal Predictor

{**Animal**, Hybrid}

{Animal, Human, **Hybrid**}

{**Human**}

Accuracy Constraint (e.g. 95%)

Requirement: must return label sets that achieve accuracy constraint

Conformal Prediction: Accuracy

But we can get 100% accuracy by always returning all labels.

Classes: {Human, Animal, Hybrid}

Conformal Predictor

{**Animal**, Human, Hybrid}

{Animal, Human, **Hybrid**}

{Animal, **Human**, Hybrid}

Accuracy Constraint (e.g. 95%)

Requirement: must return label sets that achieve accuracy constraint

Conformal Prediction: Ambiguity

Want to minimize ambiguity of returned label sets.

Classes: {Human, Animal, Hybrid}

Conformal Predictor

{**Animal**, Hybrid}
ambiguity = 1/2

{Animal, Human, **Hybrid**}
ambiguity = 1

{**Human**}
ambiguity = 0

Accuracy Constraint (e.g. 95%)

Ambiguity = $\frac{\#\ of\ predicted\ labels - 1}{\#\ of\ labels - 1} \in [\mathbf{0},\mathbf{1}]$

Conformal Prediction: Constrained Objective

Minimize: expected ambiguity

Subject to: accuracy constraint

Classes: {Human, Animal, Hybrid}

Conformal Predictor

{**Animal**, Hybrid}

{Animal, Human, **Hybrid**}

{**Human**}

Accuracy Constraint (e.g. 95%)

Inside Conformal Prediction

Conformal prediction is a wrapper around any predictor that produces "non-conformity scores" over the classes relative to training data

Predictor quality influences ambiguity of conformal prediction.

Non-conformity Scores
[human,animal,hybrid]

Your Favorite Predictor

[6, 0.5, 0.1]

[0.1, 0.5, 0.15]

[0.01, 10, 7]

Conformal Prediction

{**Animal**, Hybrid}

{Animal, Human, **Hybrid**}

{**Human**}

Training Data

Accuracy Constraint (e.g. 95%)

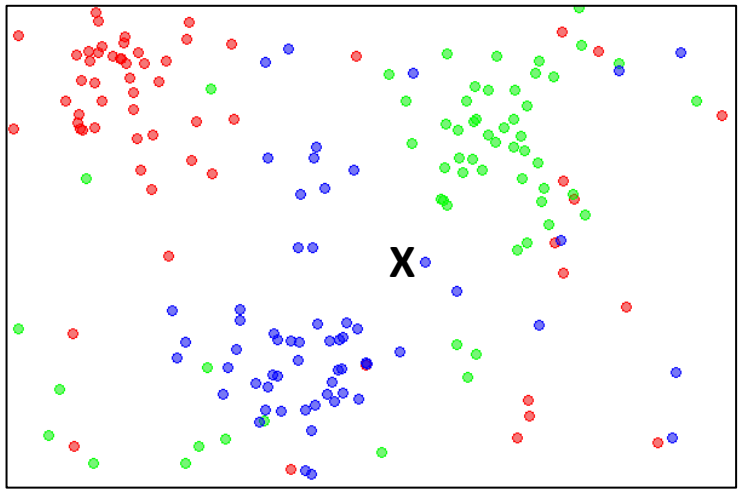
Non-conformity Scores: Nearest Neighbor

X

Nearest Neighbor

NonConformity(X, Y) = *distance to closest Y/distance to closest non Y*

$Y \in \{human, animal, hybrid\}$

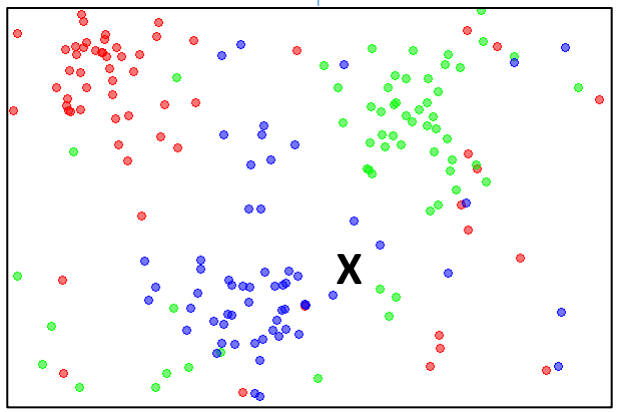


Training Data

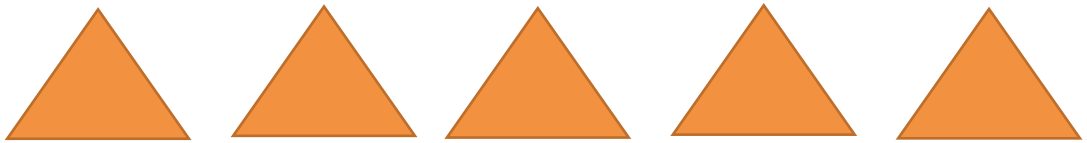
Non-conformity Scores: Random Forest

X

Random Forest



NonConformity(X, Y) = *# of trees not predicting Y/# of trees*



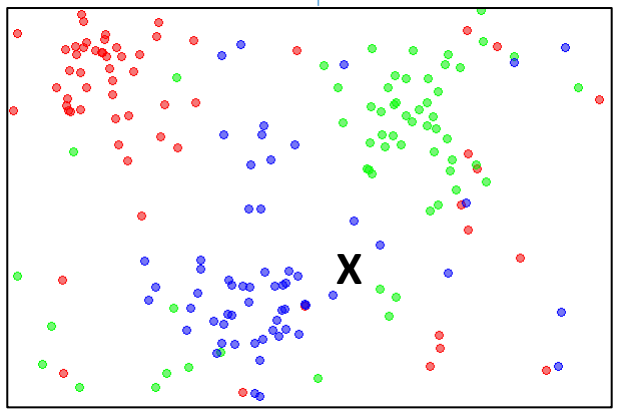
animal hybrid human hybrid human

NonConformity(X, animal) = 4/5
NonConformity(X, human) = 3/5
NonConformity(X, hybrid) = 3/5

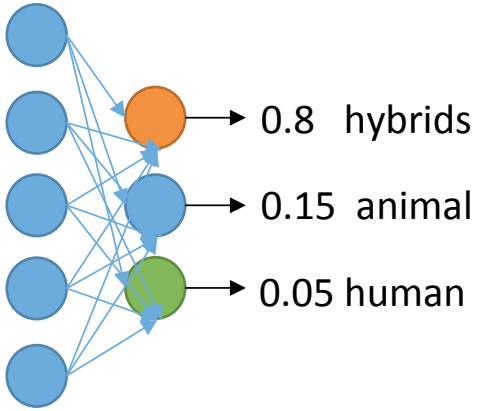
Conformal Prediction: Neural Networks

X

Neural Networks



NonConformity(X, Y) = *max output for not Y /output for Y*



NonConformity(X, animal) = 0.8/0.15 = 5.3
NonConformity(X, human) = 0.8/0.05 = 16
NonConformity(X, hybrid) = 0.15/0.8 = 0.19

Inside Conformal Prediction

Your Favorite Predictor

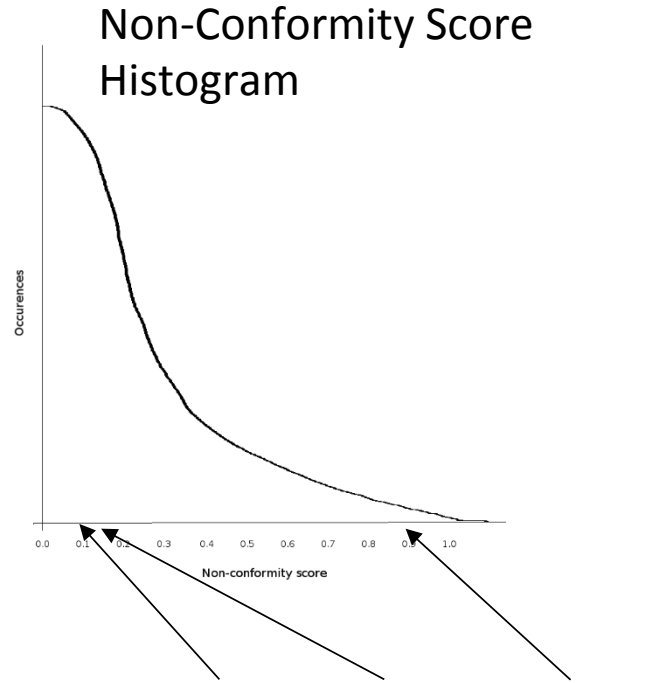
Non-conformity Scores [0.05, 0.9, 0.1]

Conformal Prediction

Calibration Data

Training Data

Accuracy Constraint (e.g. 95%)



Are at least 5% of the calibration scores weirder than this label with this example?

Inside Conformal Prediction

Your Favorite Predictor

Non-conformity Scores [0.05, 0.9, 0.1]

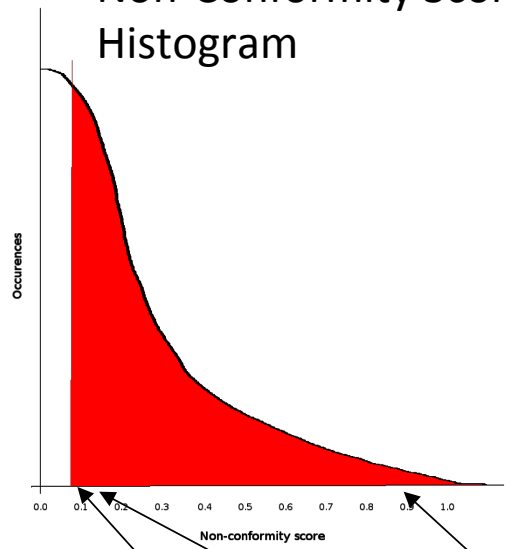
Conformal Prediction

Calibration Data

Training Data

Accuracy Constraint (e.g. 95%)

Non-Conformity Score Histogram



Animal Hybrid ~~Human~~

Are at least 5% of the calibration scores weirder than this label with this example?

Conformal Prediction: Empirical Evaluation

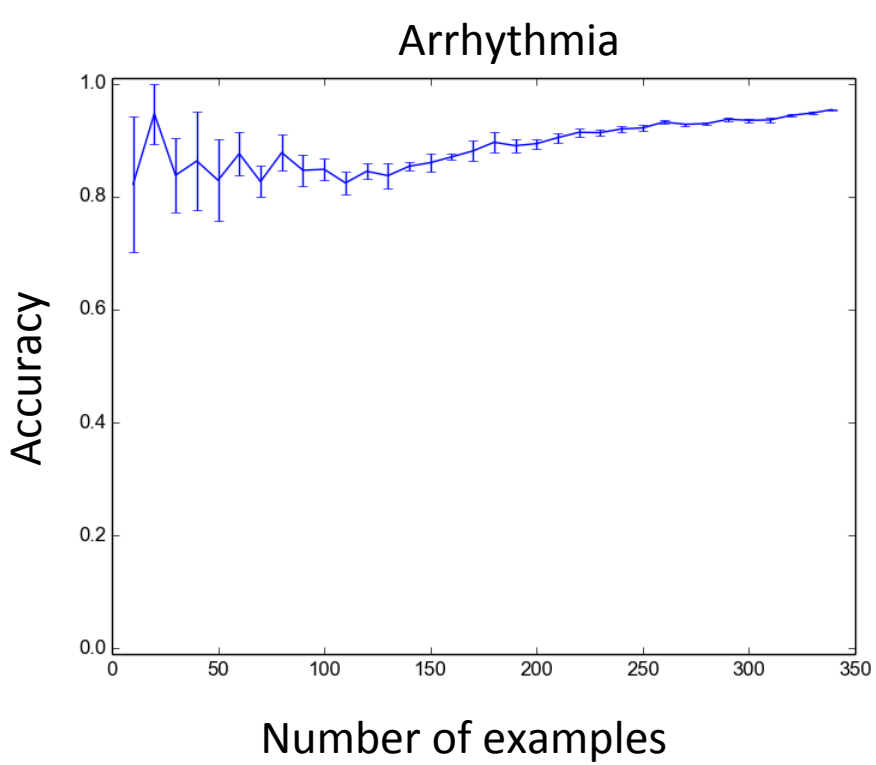
- Very few empirical evaluations of conformal prediction
  - Rarely look at ambiguity
- Most results for Nearest Neighbor -- often yields large ambiguity in our experience
- How does ambiguity vary with amount of training data in closed worlds?
- How does conformal prediction perform in open worlds?

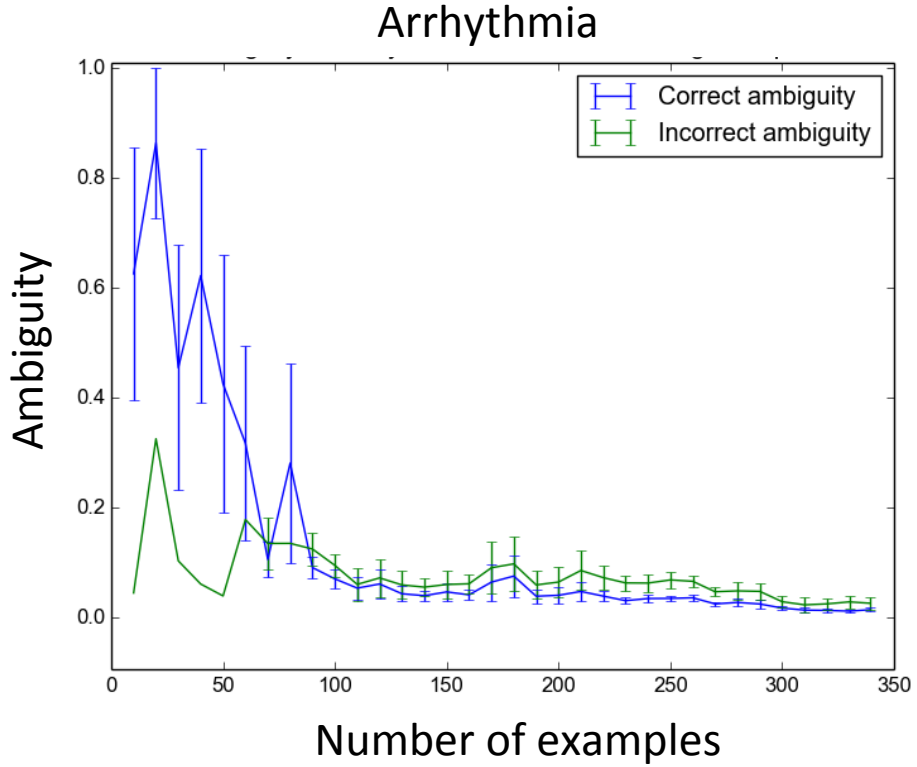
Closed World: Random Forest Results

- Arrhythmia: 452 data points, 13 labels, major class imbalances
- Cardiotocography: 2126 instances, 10 labels, balanced classes
- Image Segmentation: 2310 instances, 8 labels, balanced classes
- Iris: 150 instances, 3 labels, balanced classes

Closed World: Random Forest

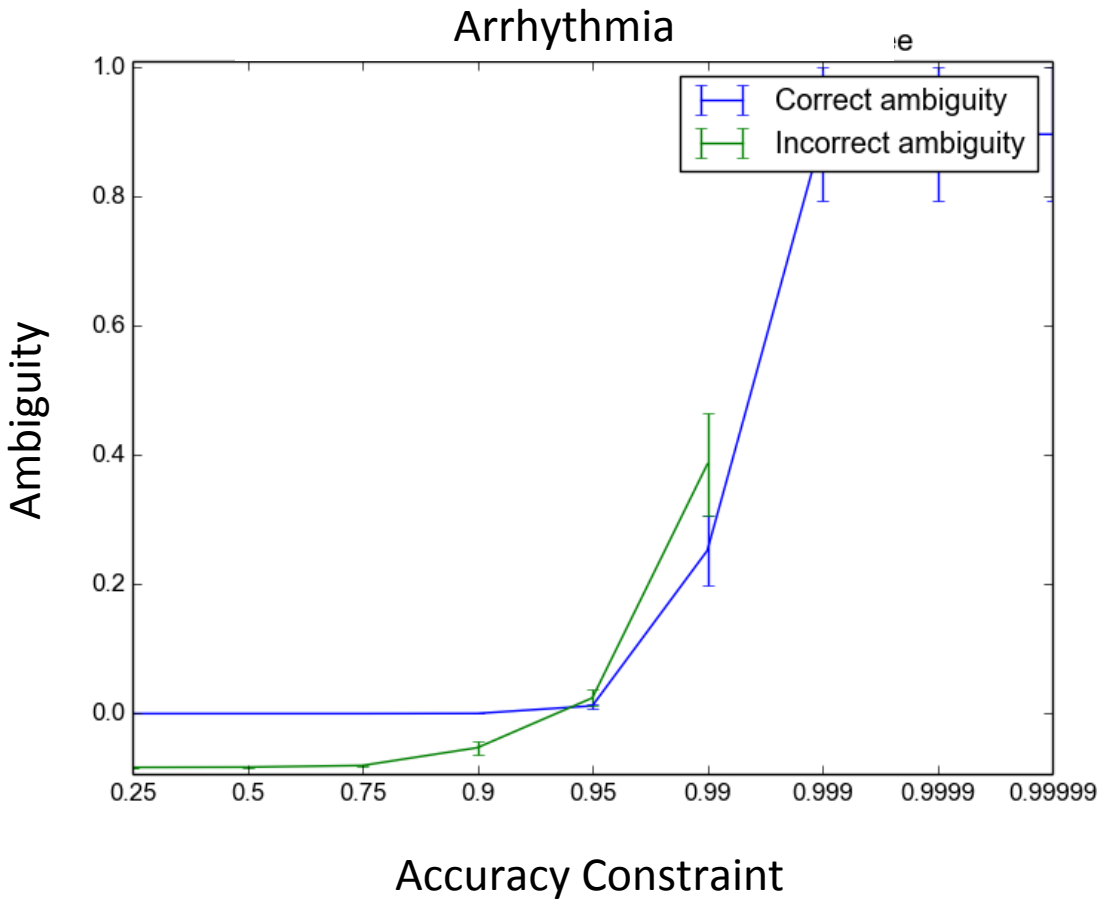
Accuracy Constraint = 95%

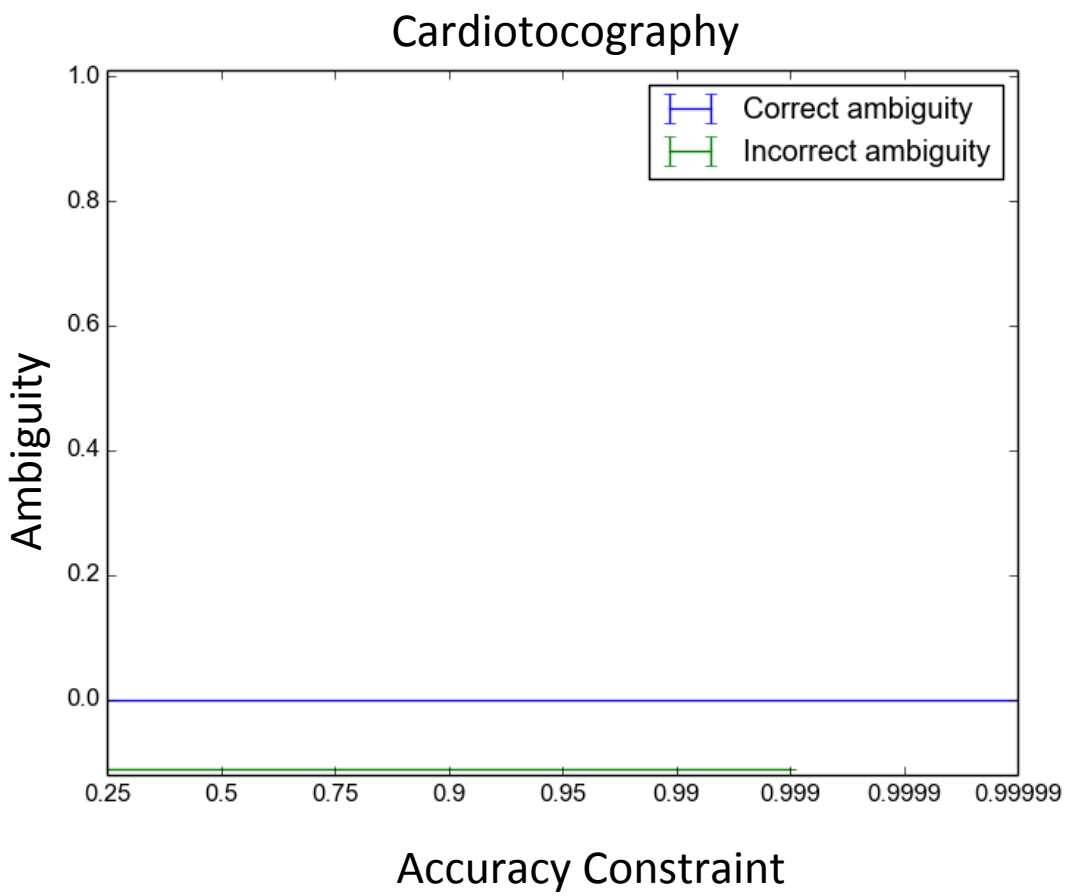




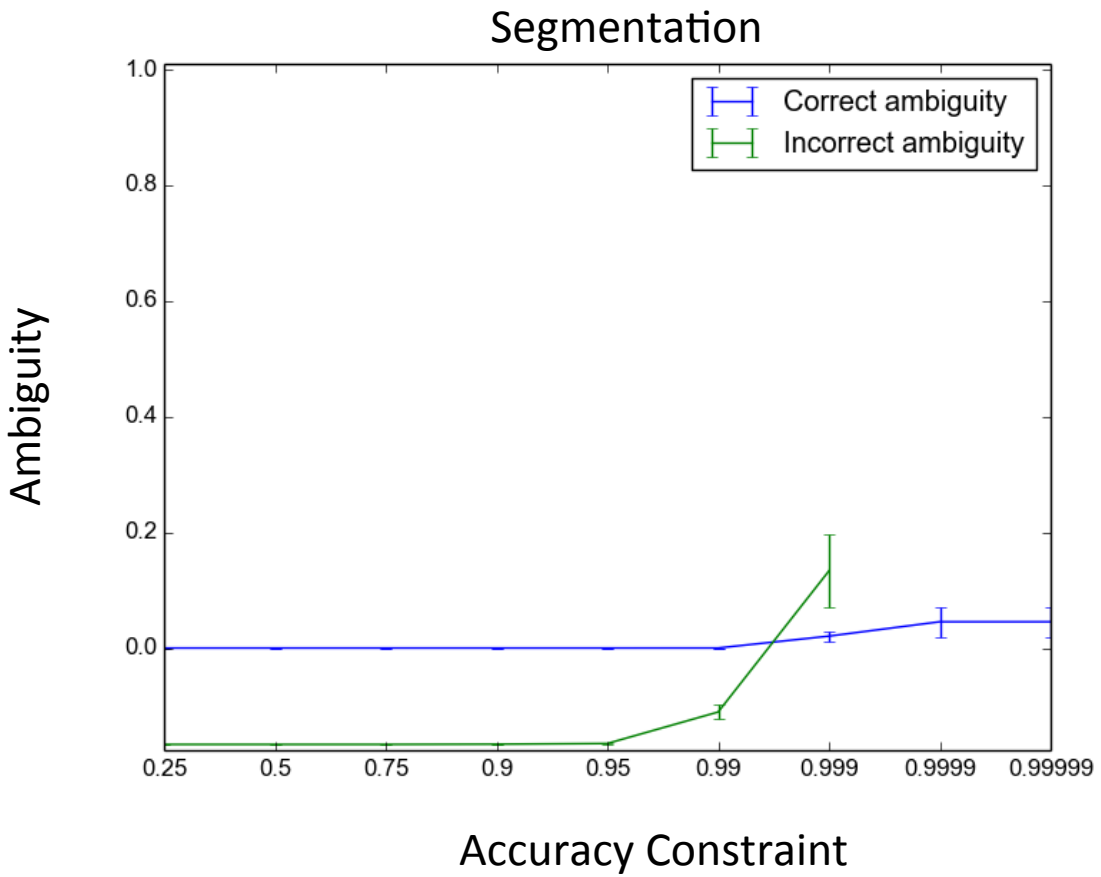
Other data sets are qualitatively similar.

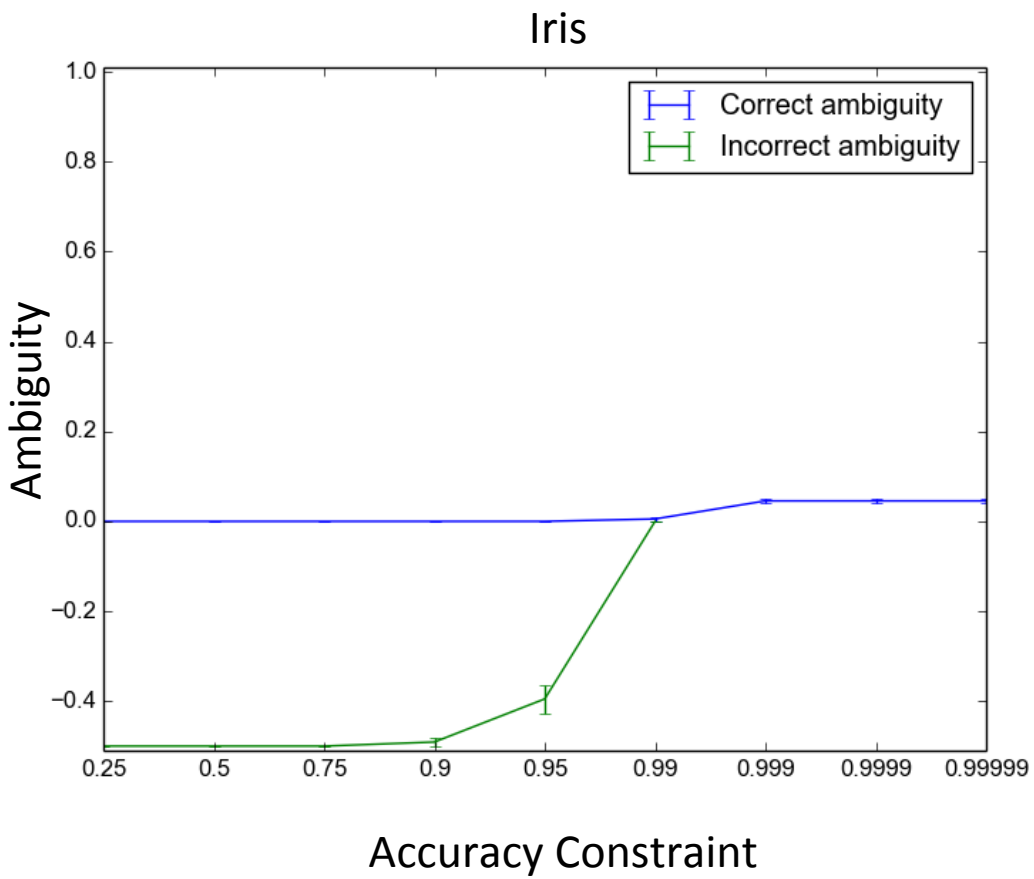
Closed World: Random Forest





Closed World: Random Forest

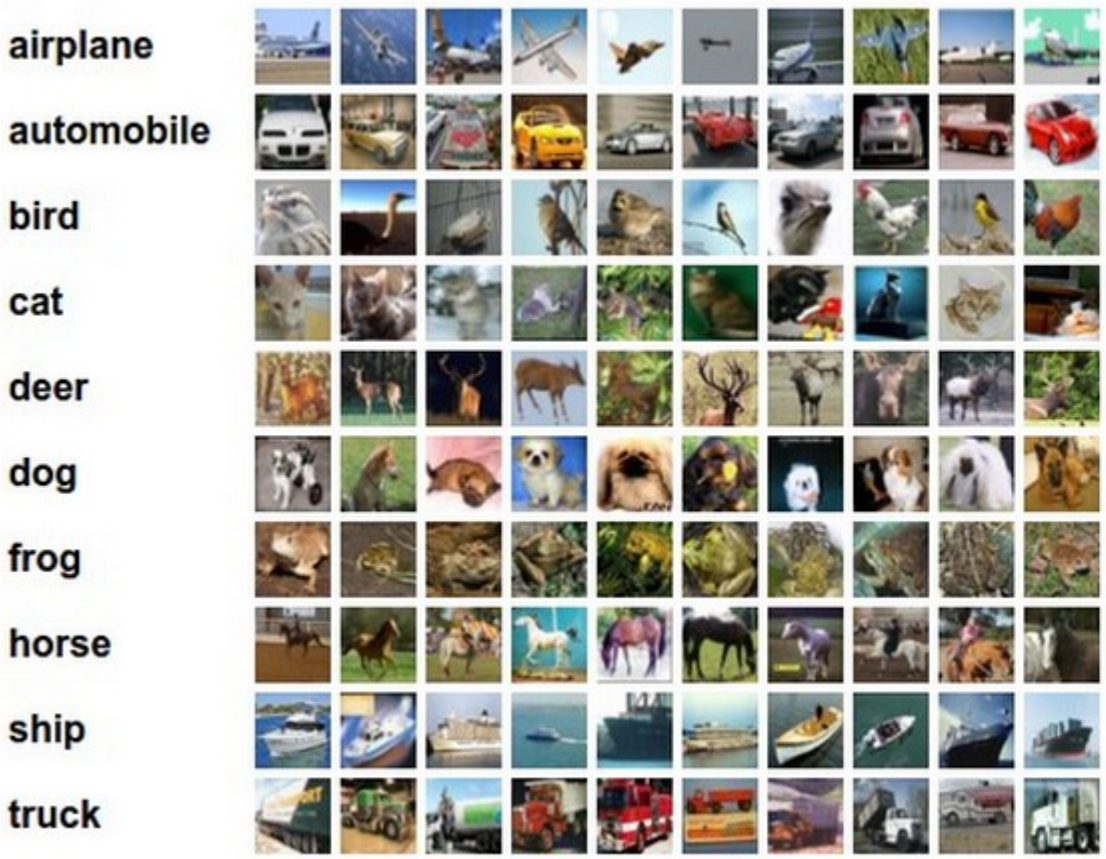


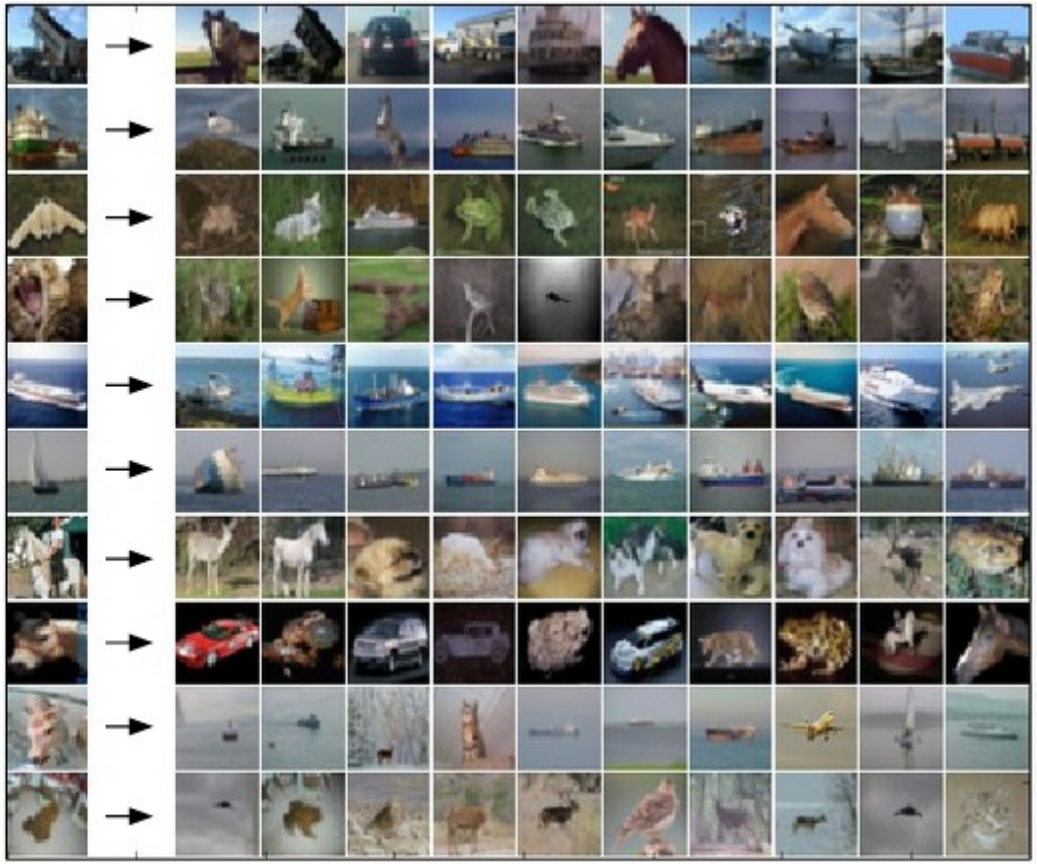


Closed World: Random Forest

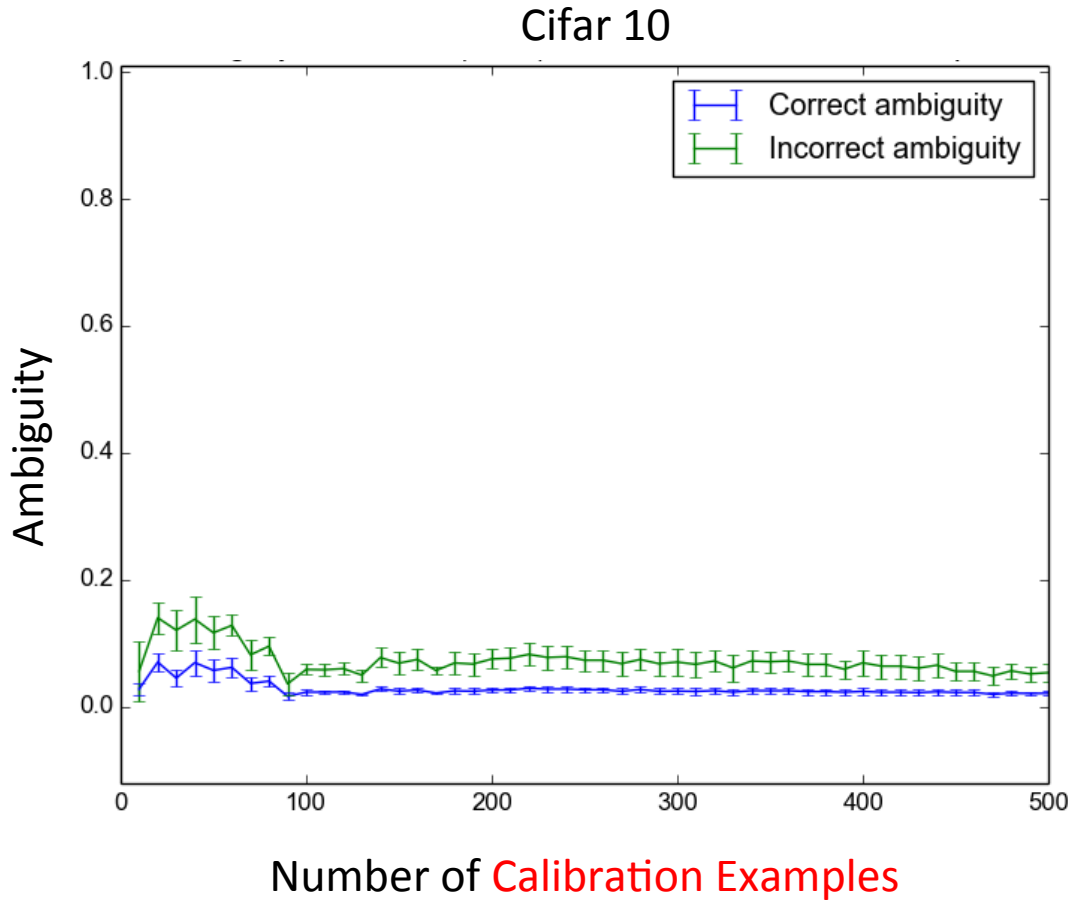
- Overall, we see “ideal behavior” on these data sets.
- Close to 0 ambiguity with small amount of data.

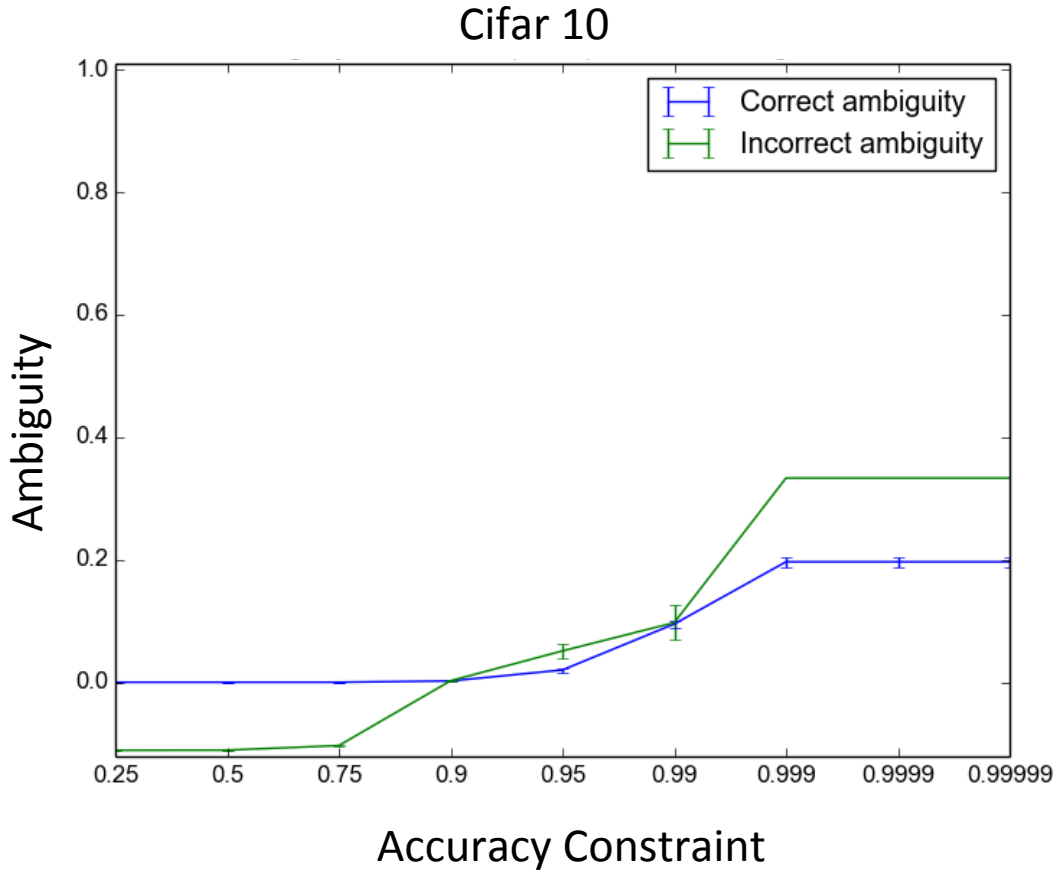
Closed World: Convolutional Network: Cifar 10





Closed World: Deep Net





Closed World Observations

- Overall, we see “ideal behavior” on these data sets for both random forest and convolutional network.
- Close to 0 ambiguity with small amount of data.
- Neural network very rarely abstains (negative ambiguity) compared to the random forest

Open World: Conformal Prediction

Training Labels = {Human, Animal, Hybrid}

Predictor

Conformal Prediction

{Animal}

Ø

{Human, Animal}

We don’t have a label for “monster”!

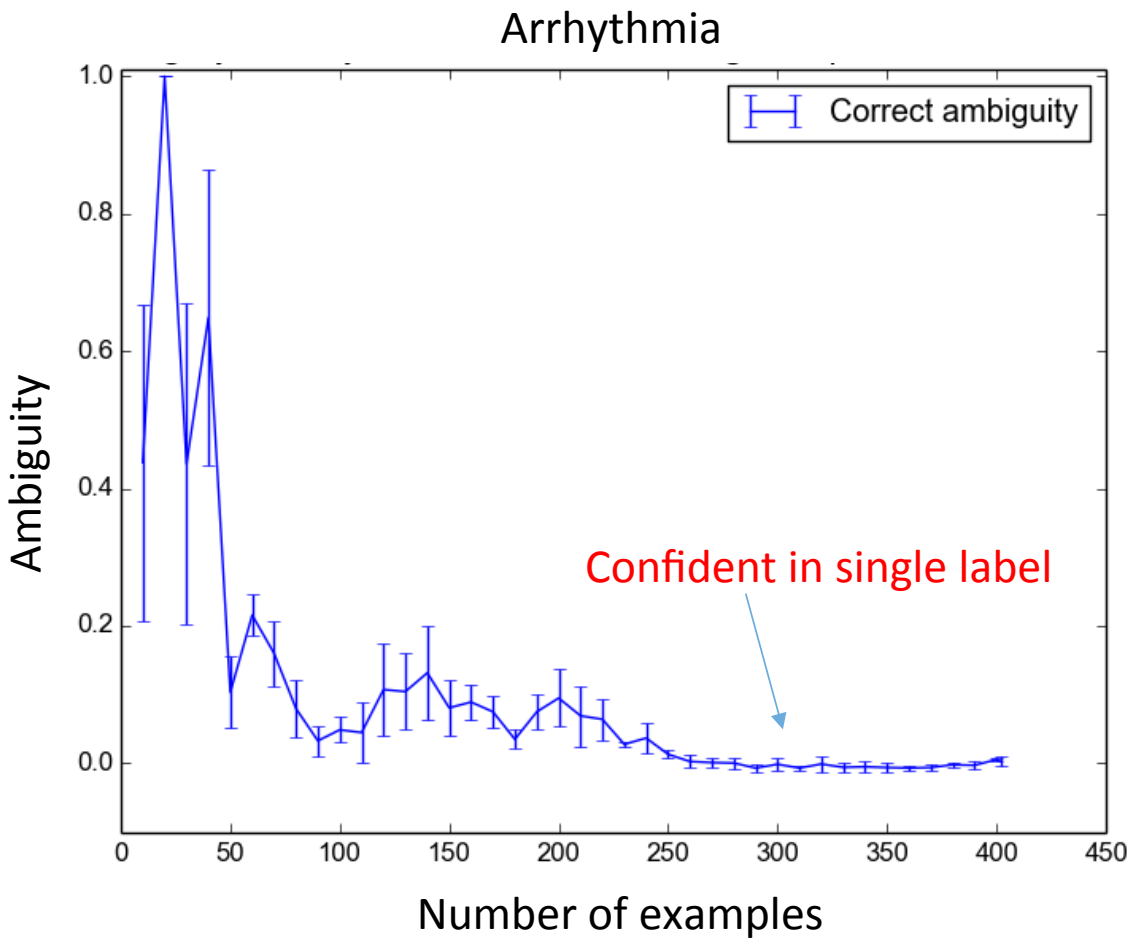
Only reasonable output is to abstain (empty set)

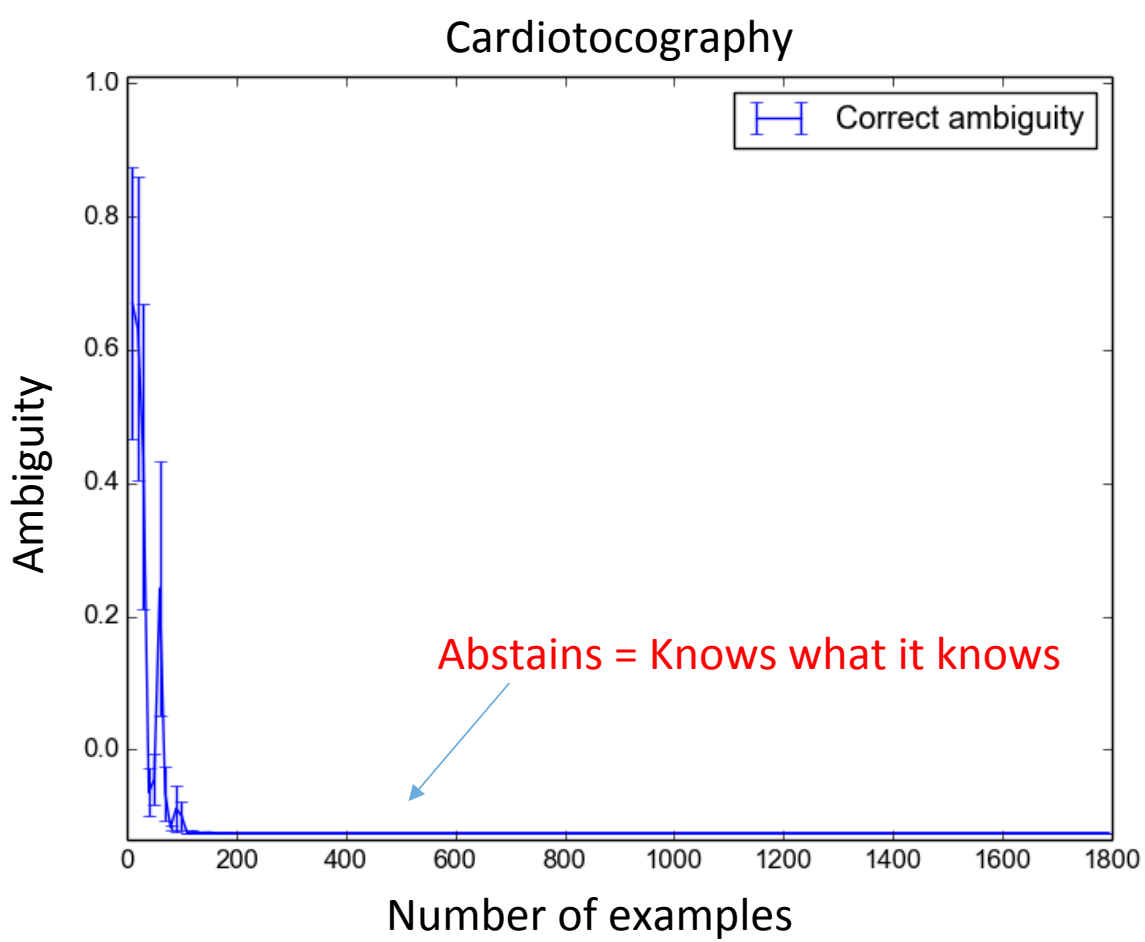
Open World Experiments

- Feed novel classes to conformal predictor
- **Random Forest** : withheld a label from each training set
- **Convolutional Network** : feed it images that have nothing to do with Cifar 10

Open World: Random Forest

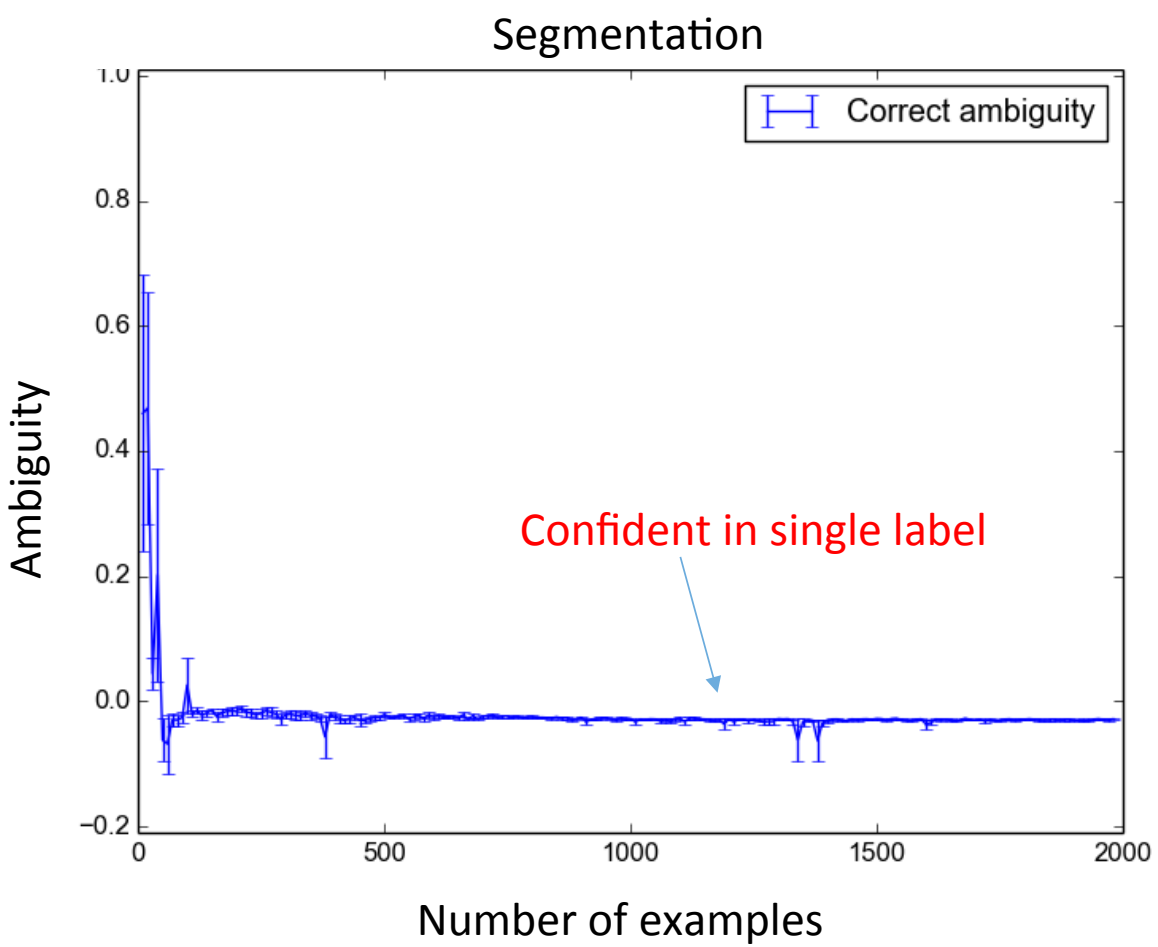
Ambiguity for just novel classes

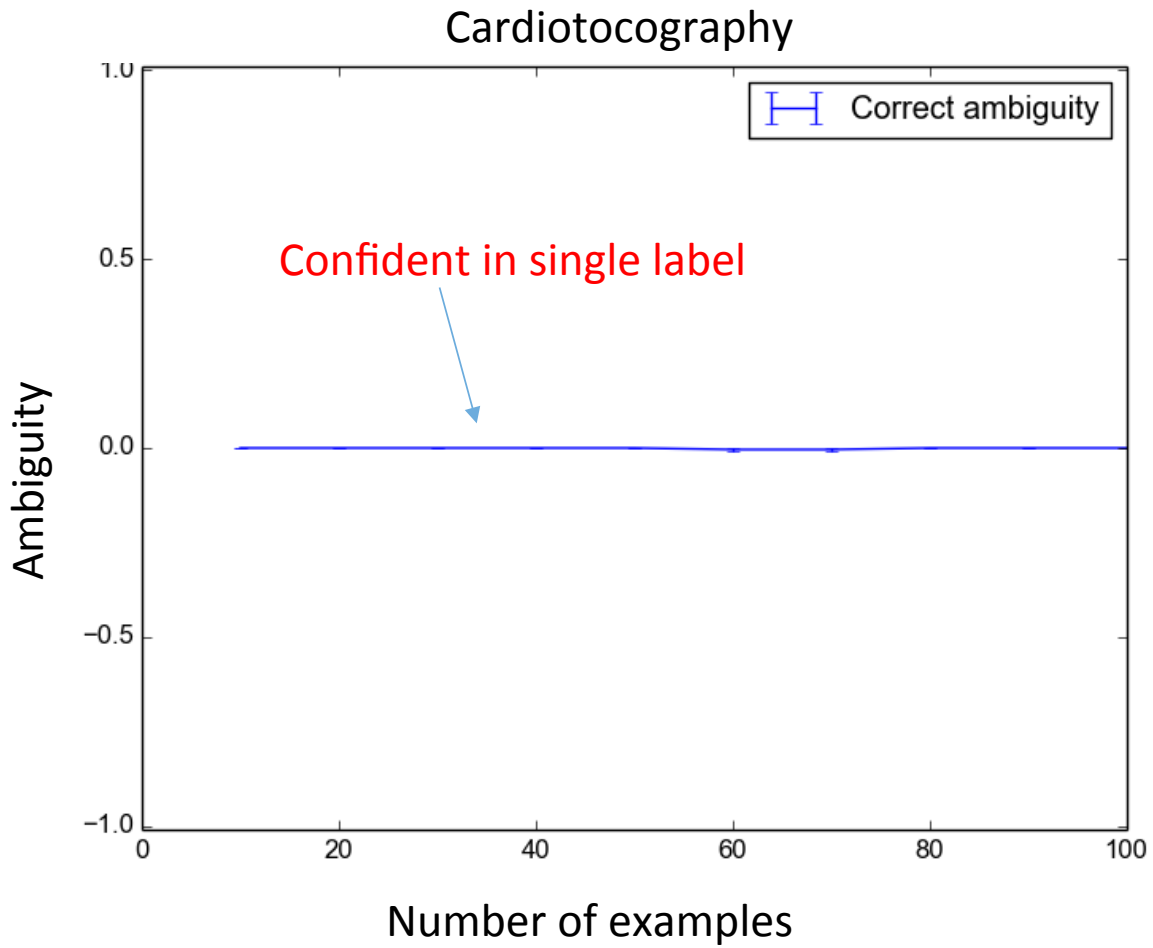




Open World: Random Forest

Ambiguity for just novel classes



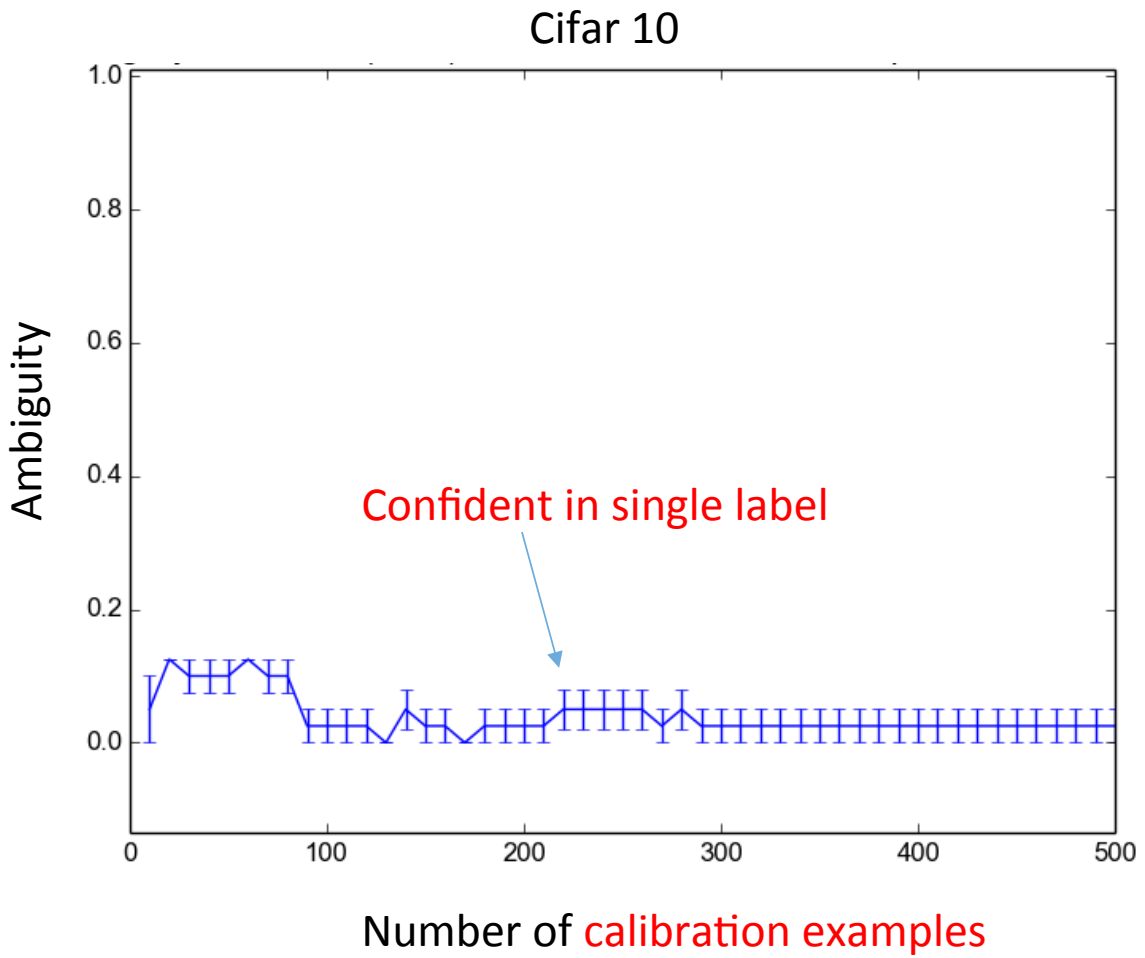


Open World: Convolutional Network

Nethack sprite sheet images

Open World: Convolutional Network

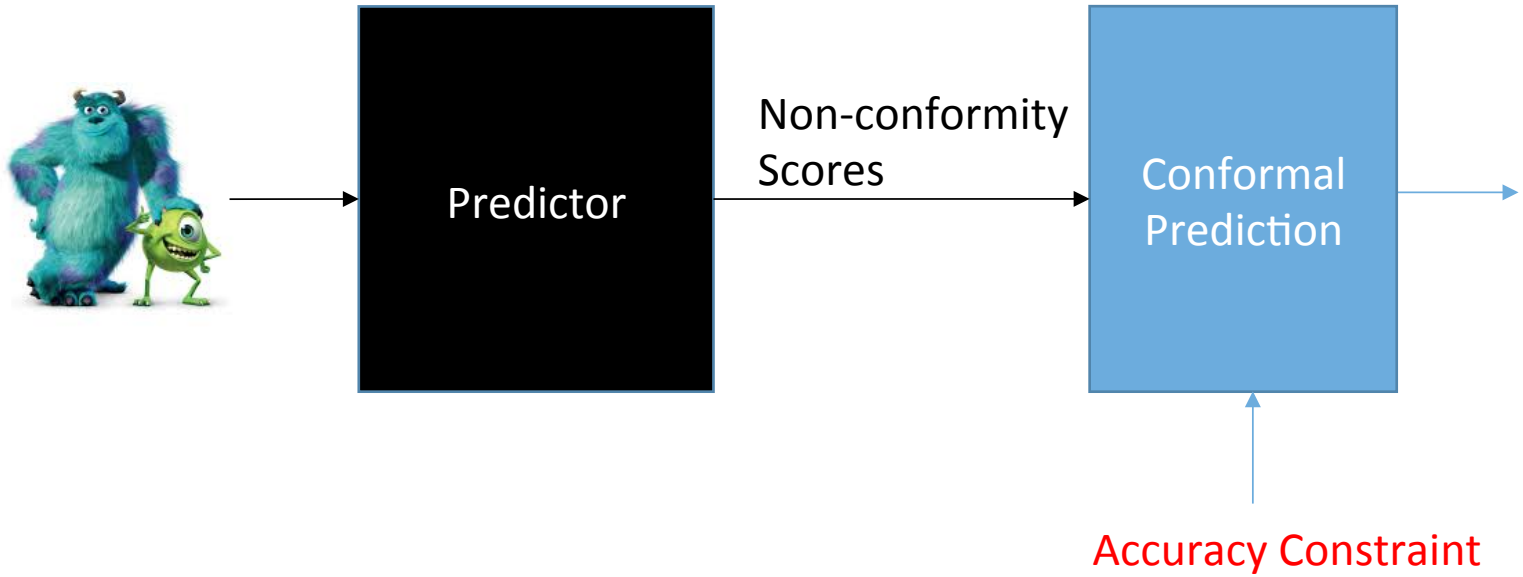
Ambiguity for just novel Nethack images



Open World Observations

- In all but one case there was practically no abstention
- The theory of conformal prediction does not address the issue of open worlds
- Appears that standard conformal prediction on its own is not sufficient for open worlds

Next Steps for Open Worlds



New algorithms for training predictors.

Goal: yield reliable abstention for novel classes.

Next Step for Open Worlds

Anomaly Detector

normal data

Predictor

Non-conformity Scores

Conformal Prediction

anomalous data

Anomaly Handler

Accuracy Constraint

How to select anomaly threshold?

Can we provide any guarantees in open worlds?