Smart Data > Big Data

TOWARDS A COST-OPTIMIZED LABELING STRATEGY

Jennifer Prendki, PhD

Data Day Conference – Austin, TX

January 25, 2020

ABOUT ALECTIO

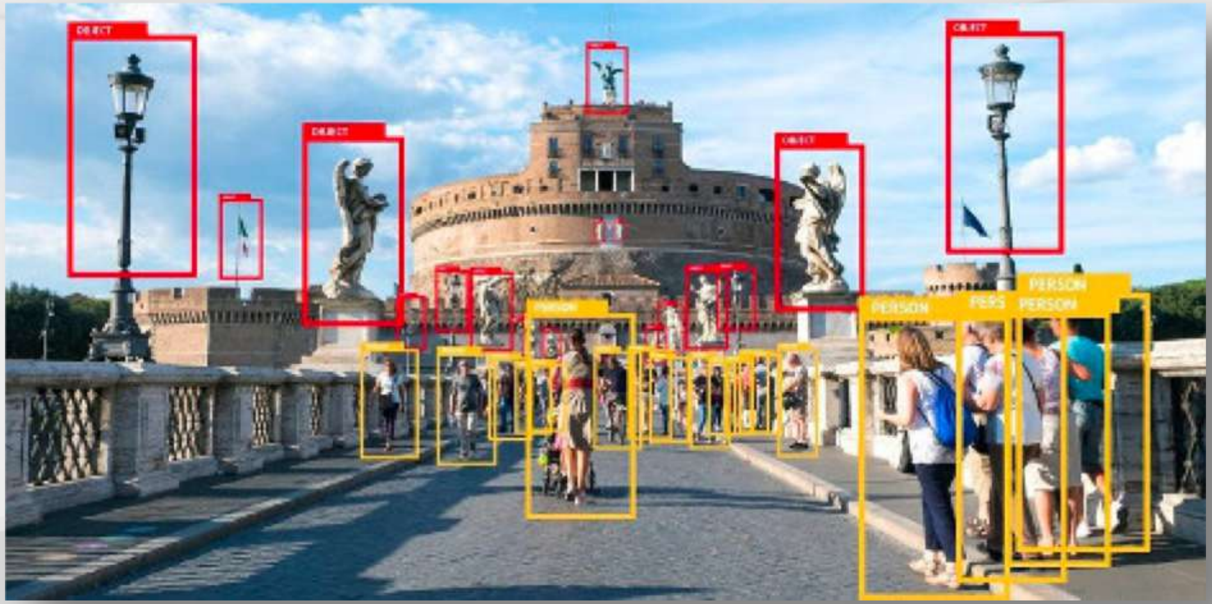
- The first ML company dedicated to Data Curation
- Founded in 2019
- Mission:

Empower ML experts to build, train and retrain models with less data, and hence less resources.

OUTLINE

- The Big Data Labeling Crisis
- Understanding Class Separation
- Not All Data is Created Equal
- How to Best Spend your Labeling Budget
- Results and Conclusions

BIG DATA LABELING CRISIS



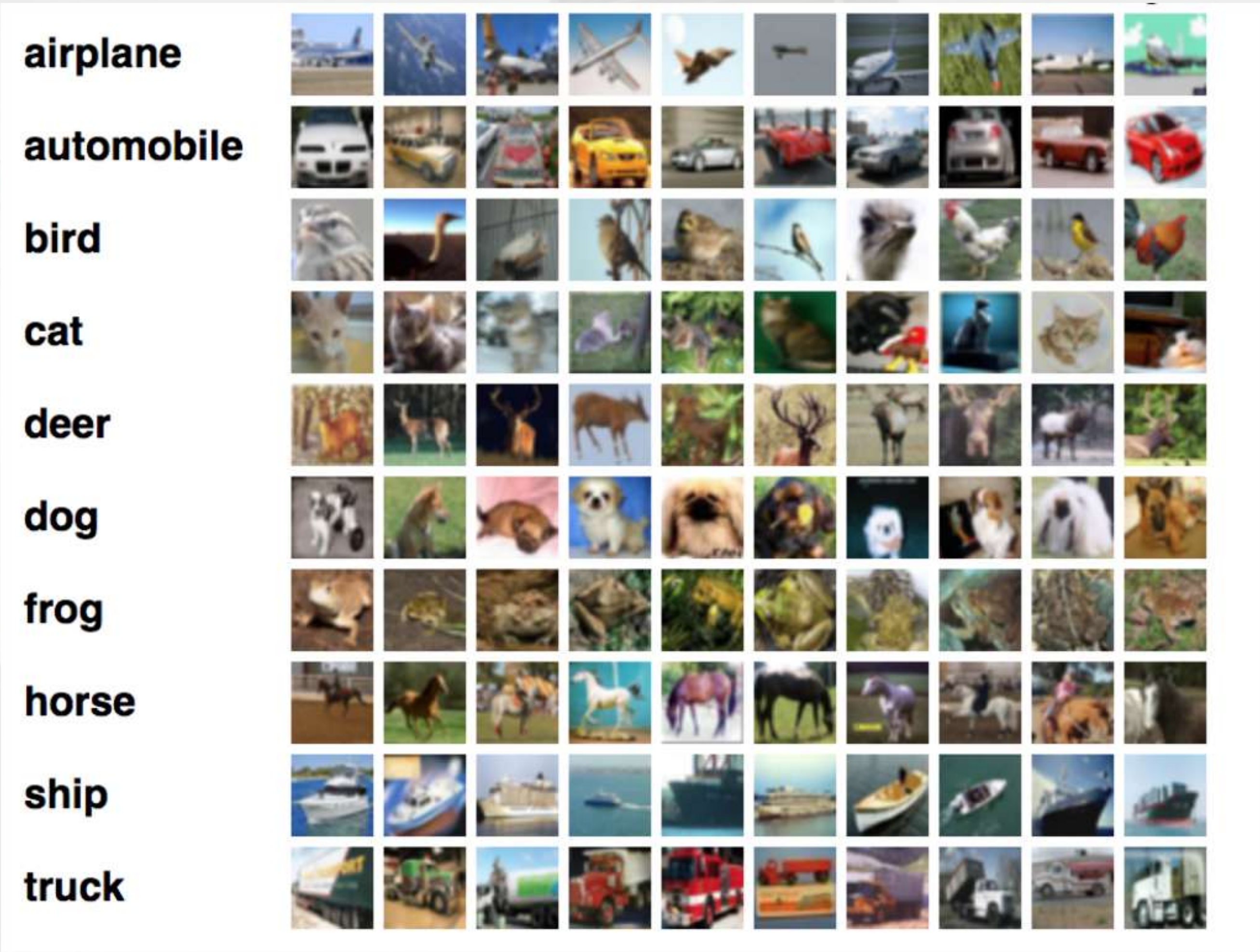
OUR 'TOY' CASE STUDY: CIFAR-10

The Data

- CIFAR-10
- 10 classes of everyday "objects"
- 50,000 training samples
- 10,000 testing samples

The Model

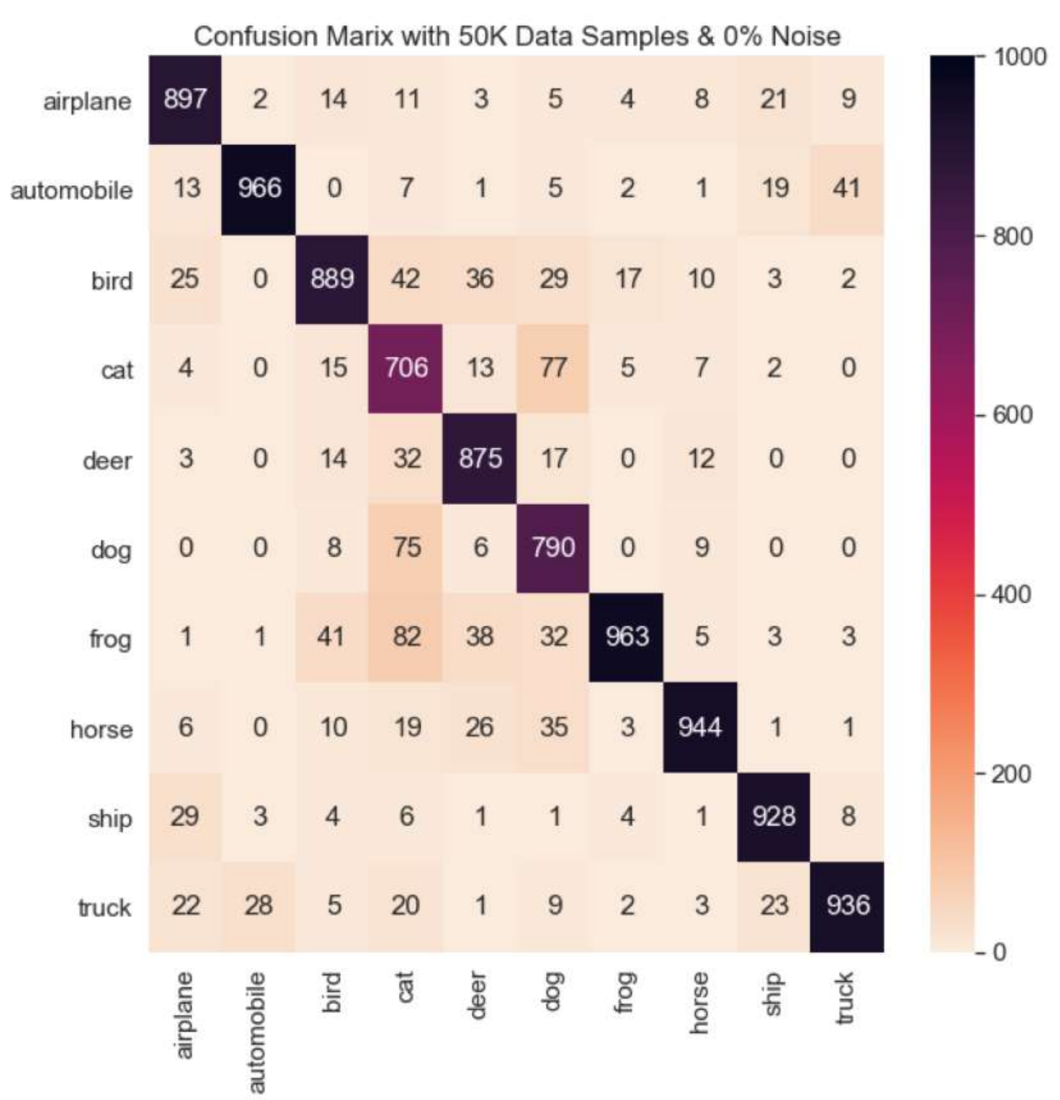
- Small CNN
  - 7 layers
  - 309,290 total parameters
  - 308,394 trainable parameters
  - 896 non-trainable parameters



BASELINE RESULTS

Results

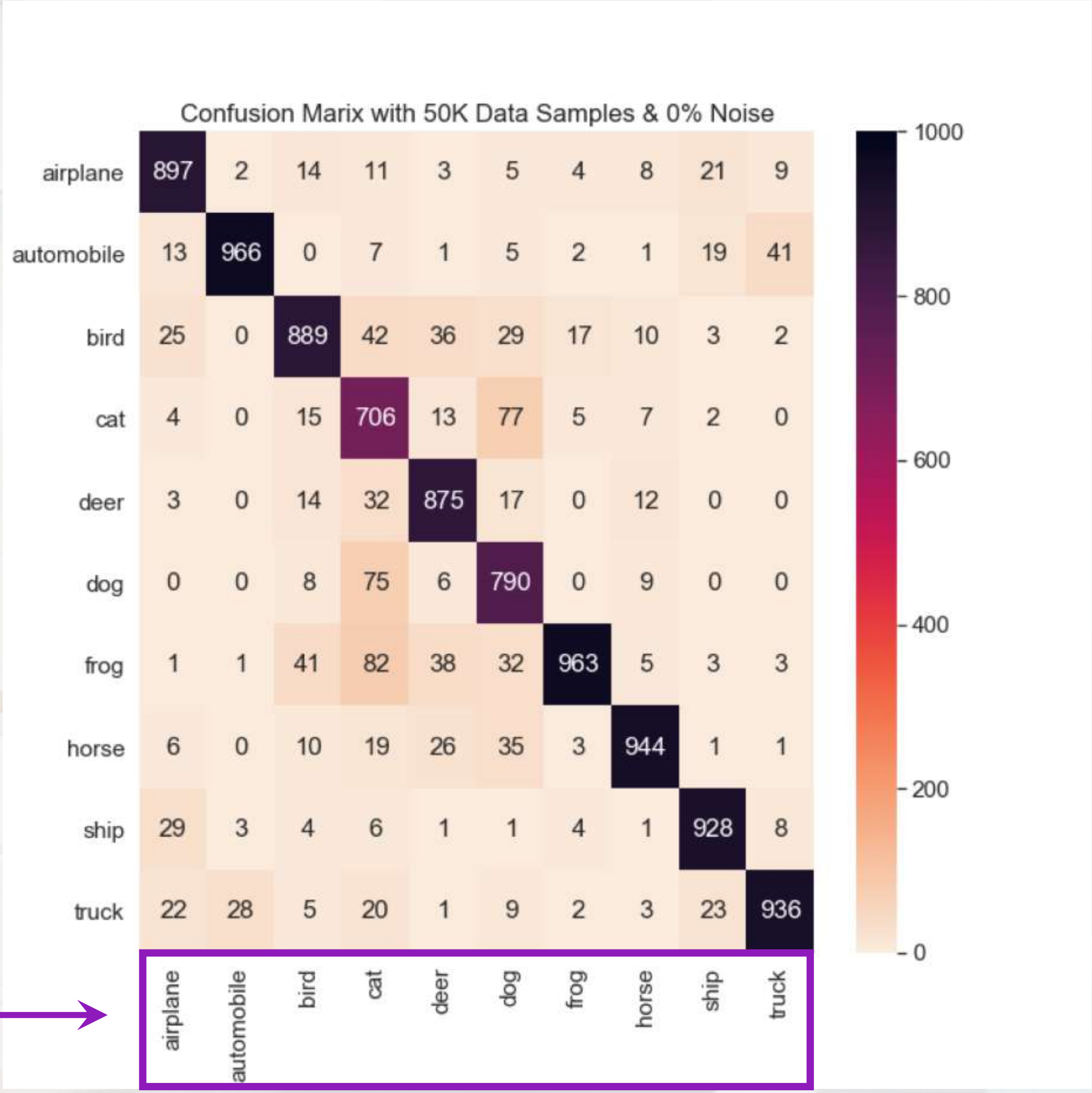
- **Baseline accuracy**: 89% (across all classes)



BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**

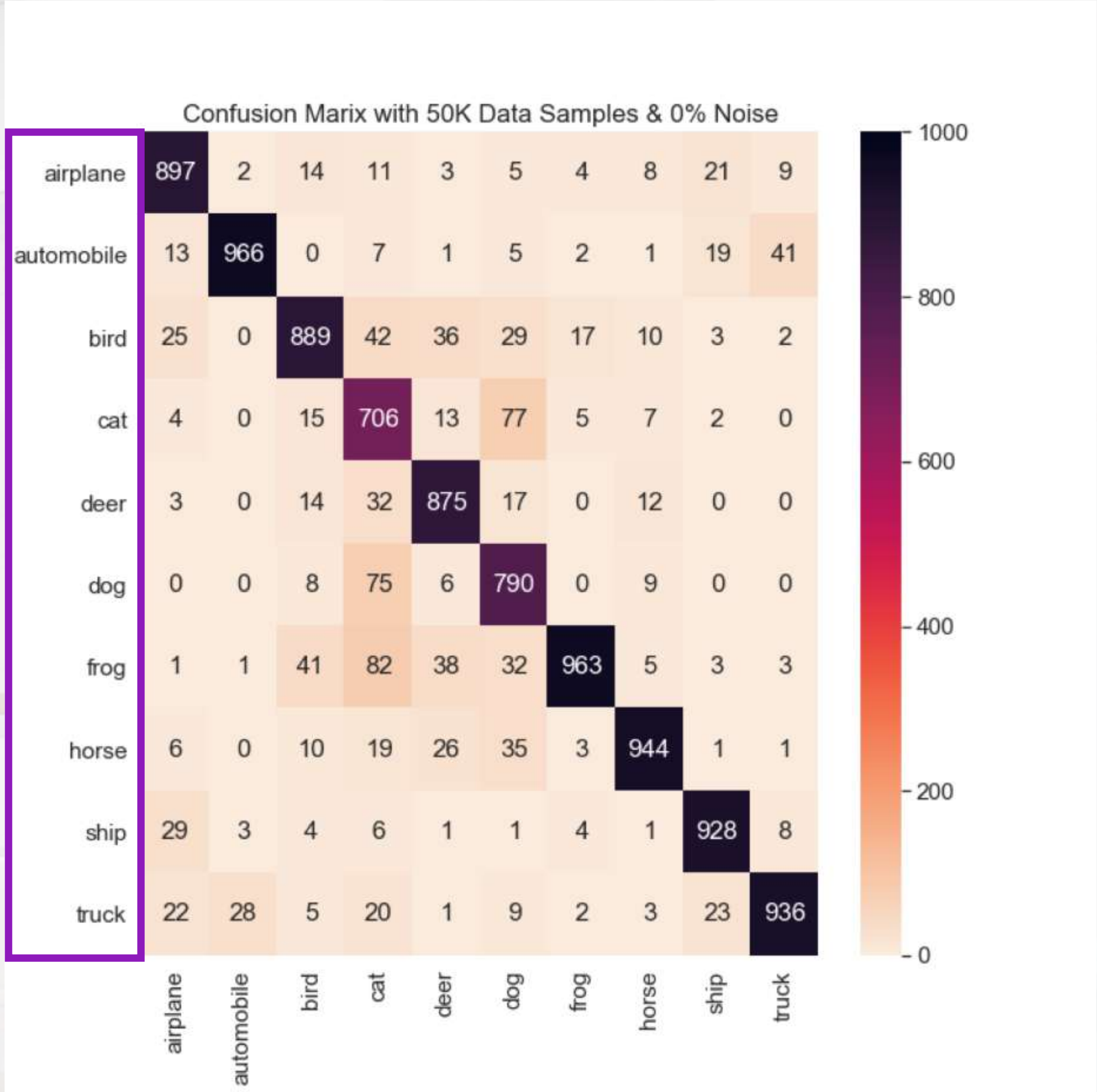


Ground truth →

BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**



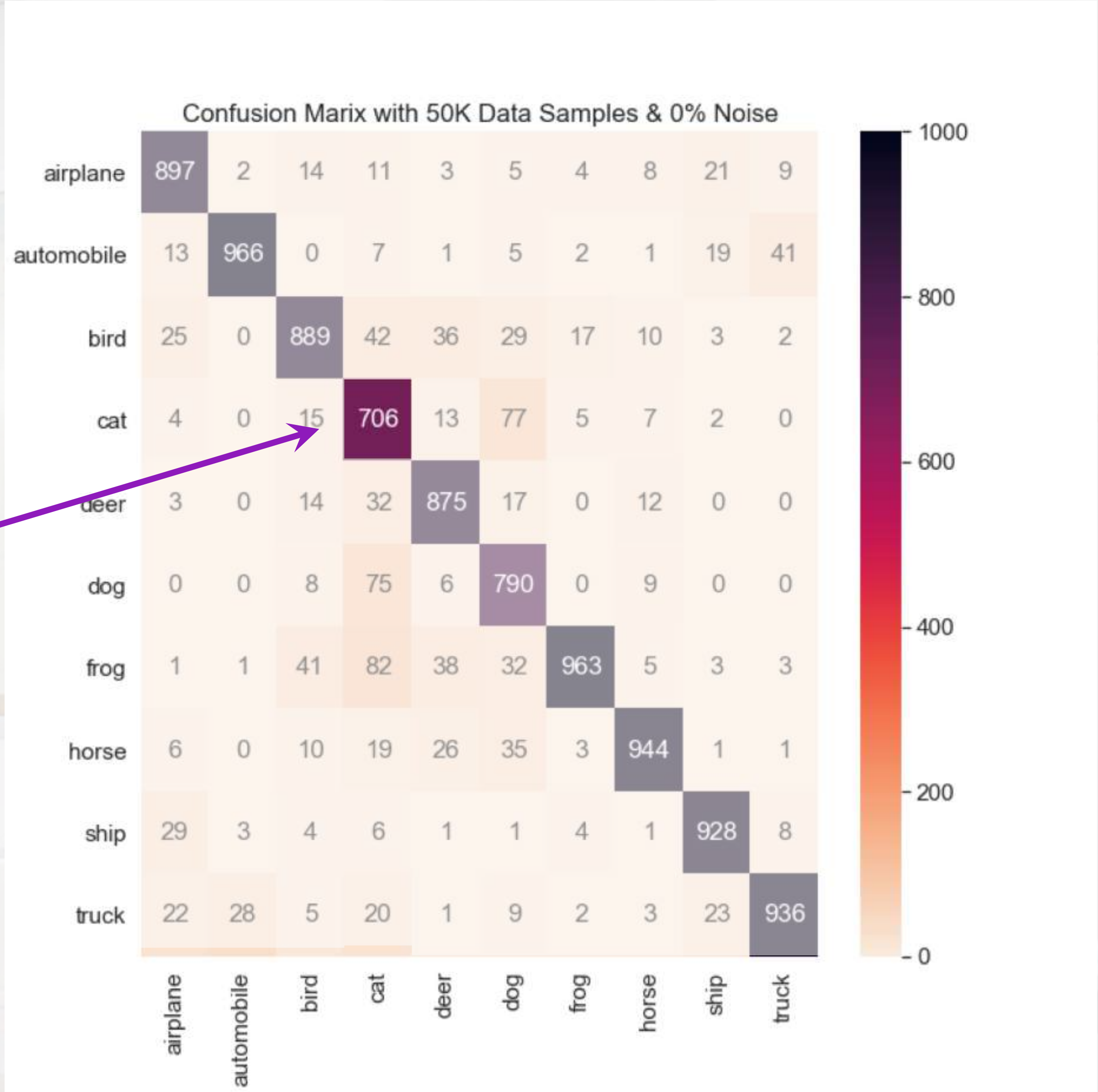
Predictions

BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**

True Positive Rate
(x 1000)

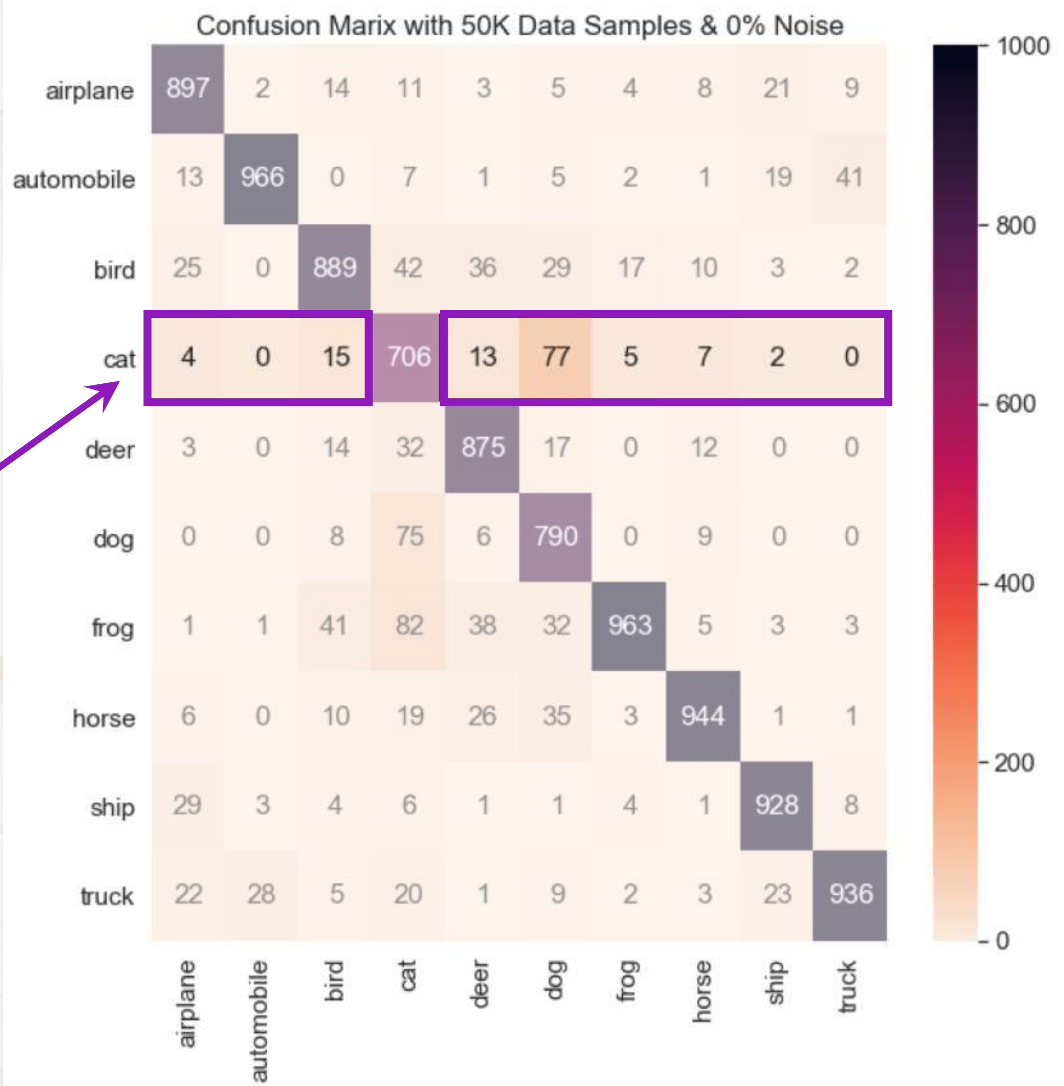


BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**

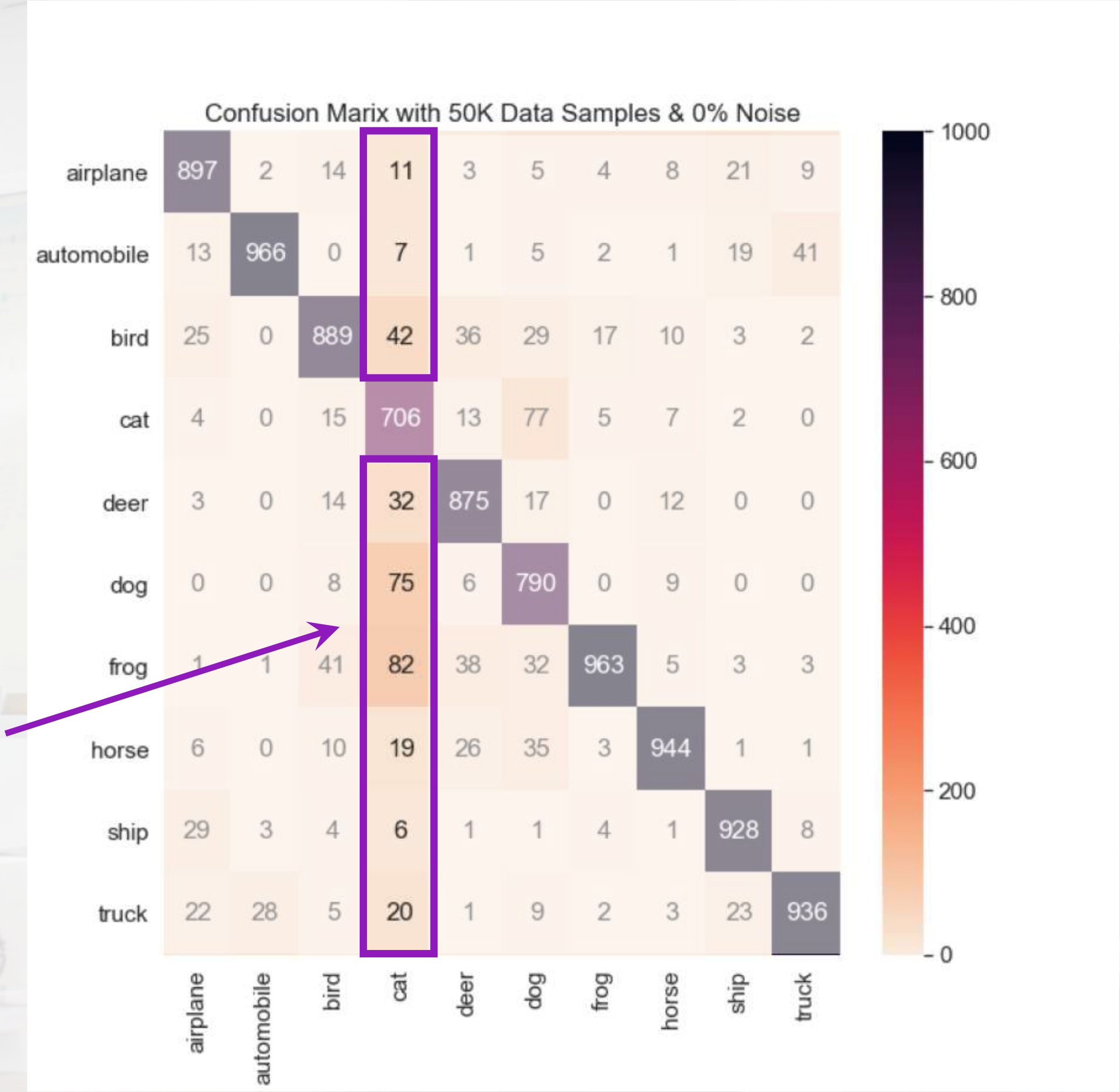
False Positive Rate



BASELINE RESULTS

Results

- **Baseline accuracy:** 89% (across all classes)

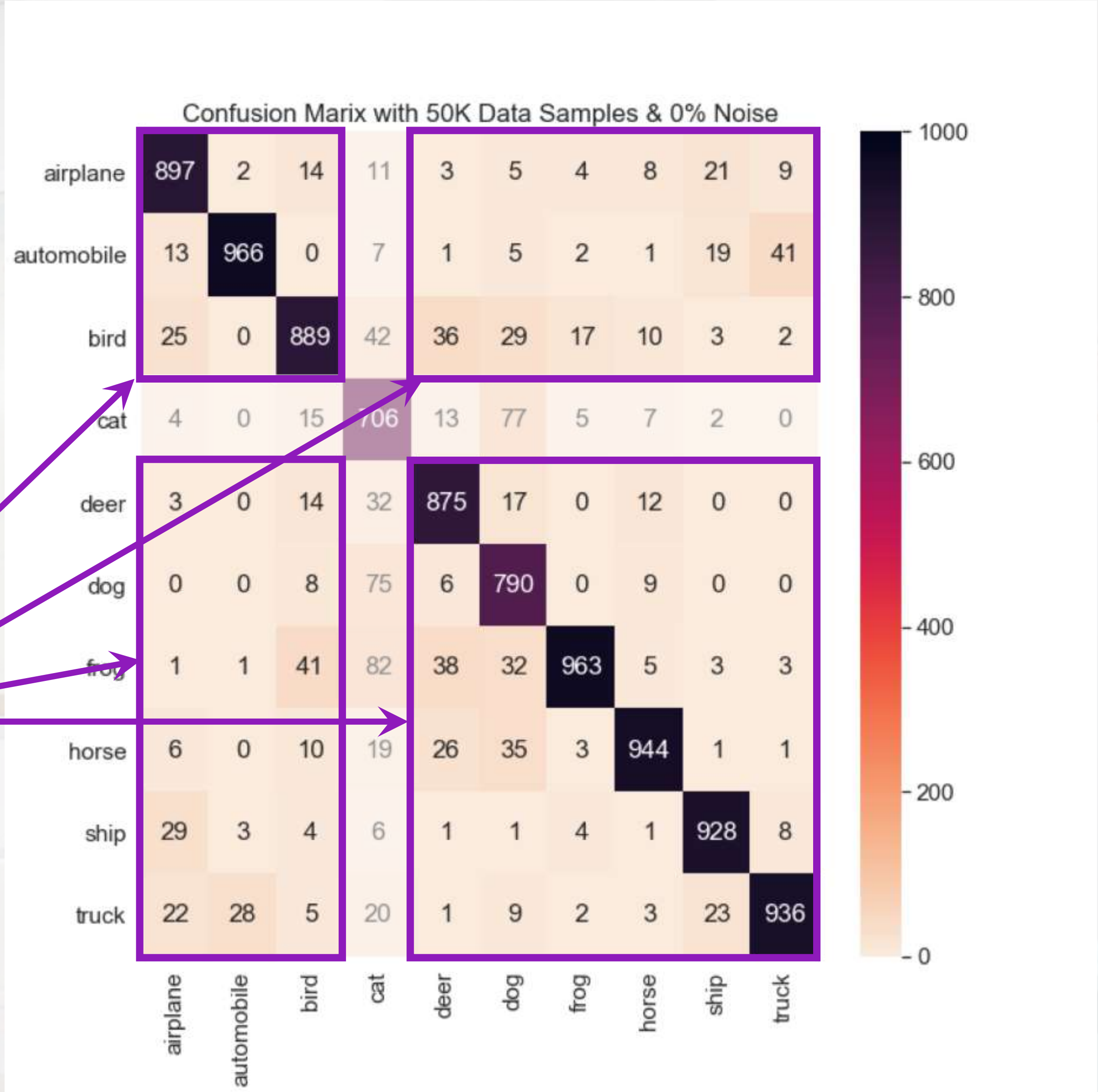


False Negative Rate

BASELINE RESULTS

Results

- **Baseline accuracy:** 89% (across all classes)

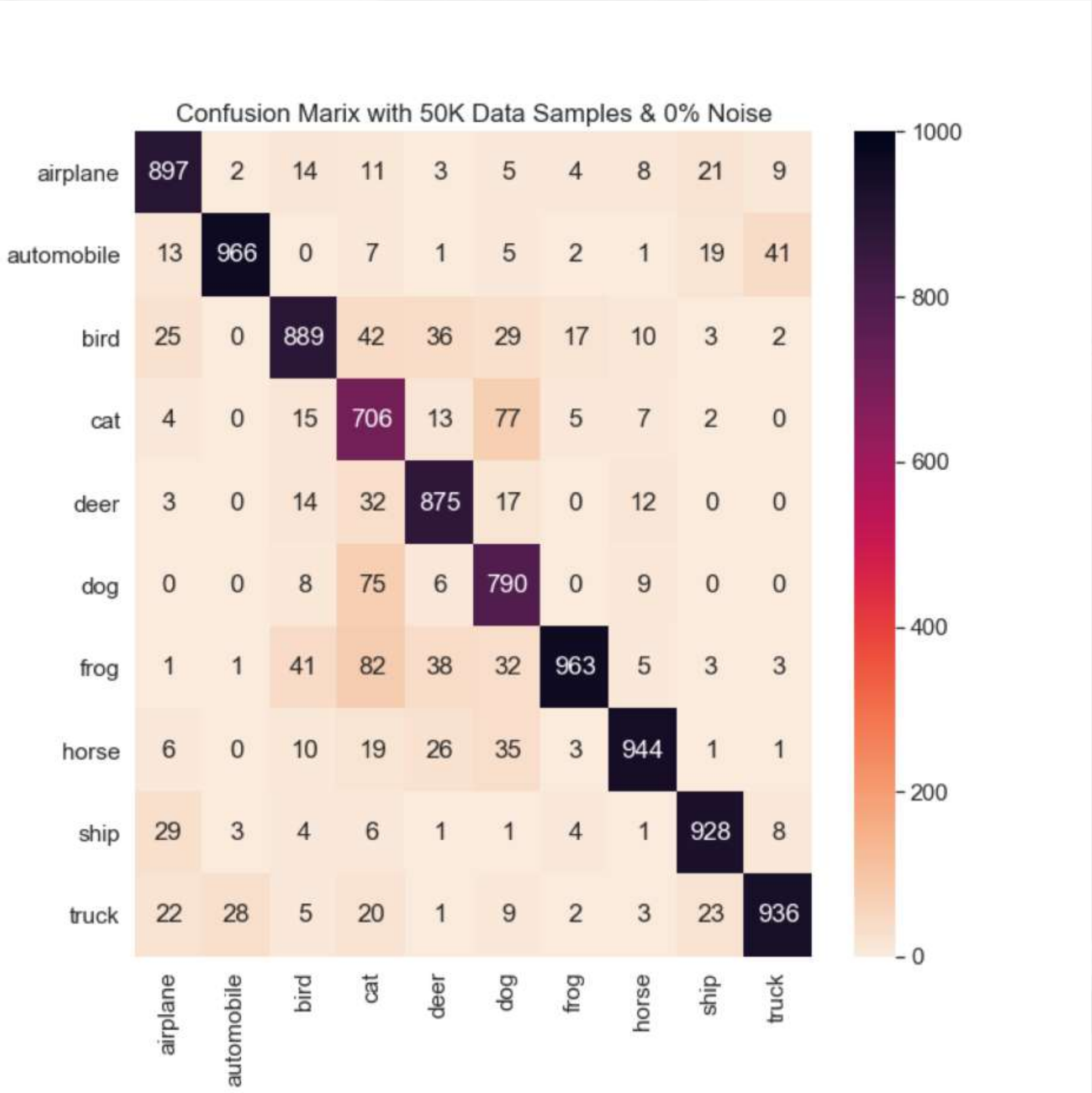


True Negative Rate

BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**
- **Accuracy varies dramatically across classes**



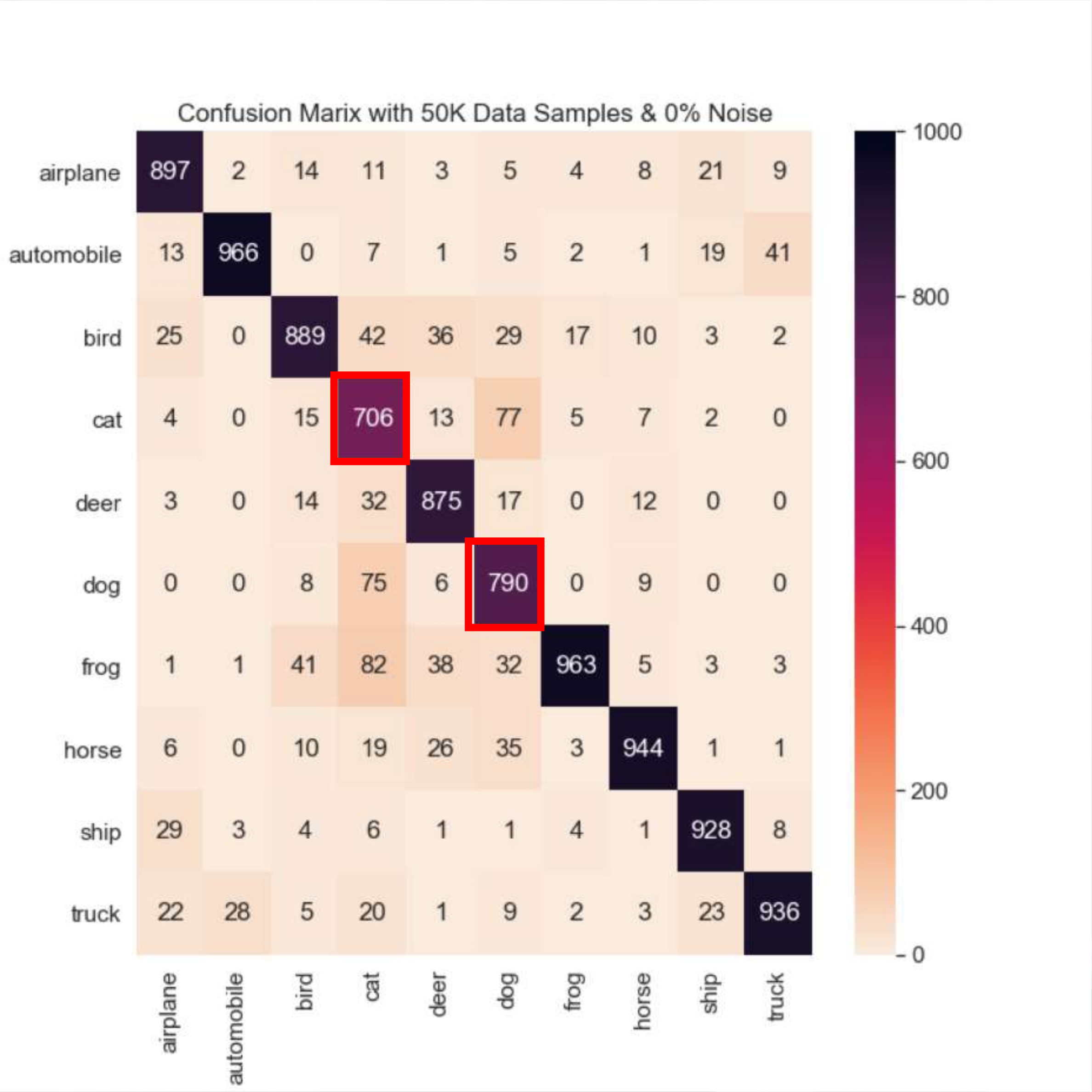
BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**
- **Accuracy varies dramatically across classes**

More details…

- **Lowest accuracy for class 'cat' and 'dog'**



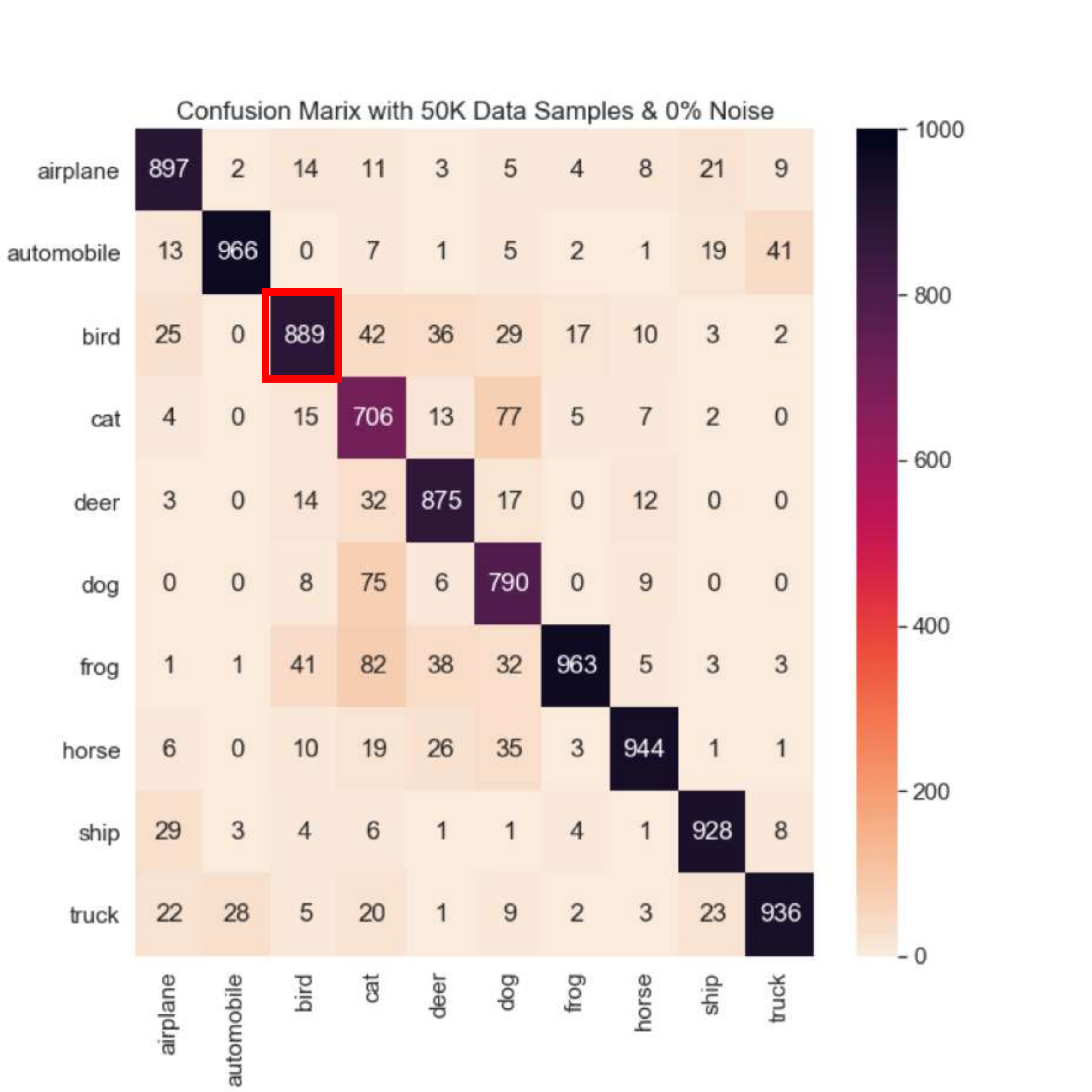
BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**
- **Accuracy varies dramatically across classes**

More details…

- **Lowest accuracy for class 'cat' and 'dog'**
- **Class 'bird' has a fairly high accuracy**



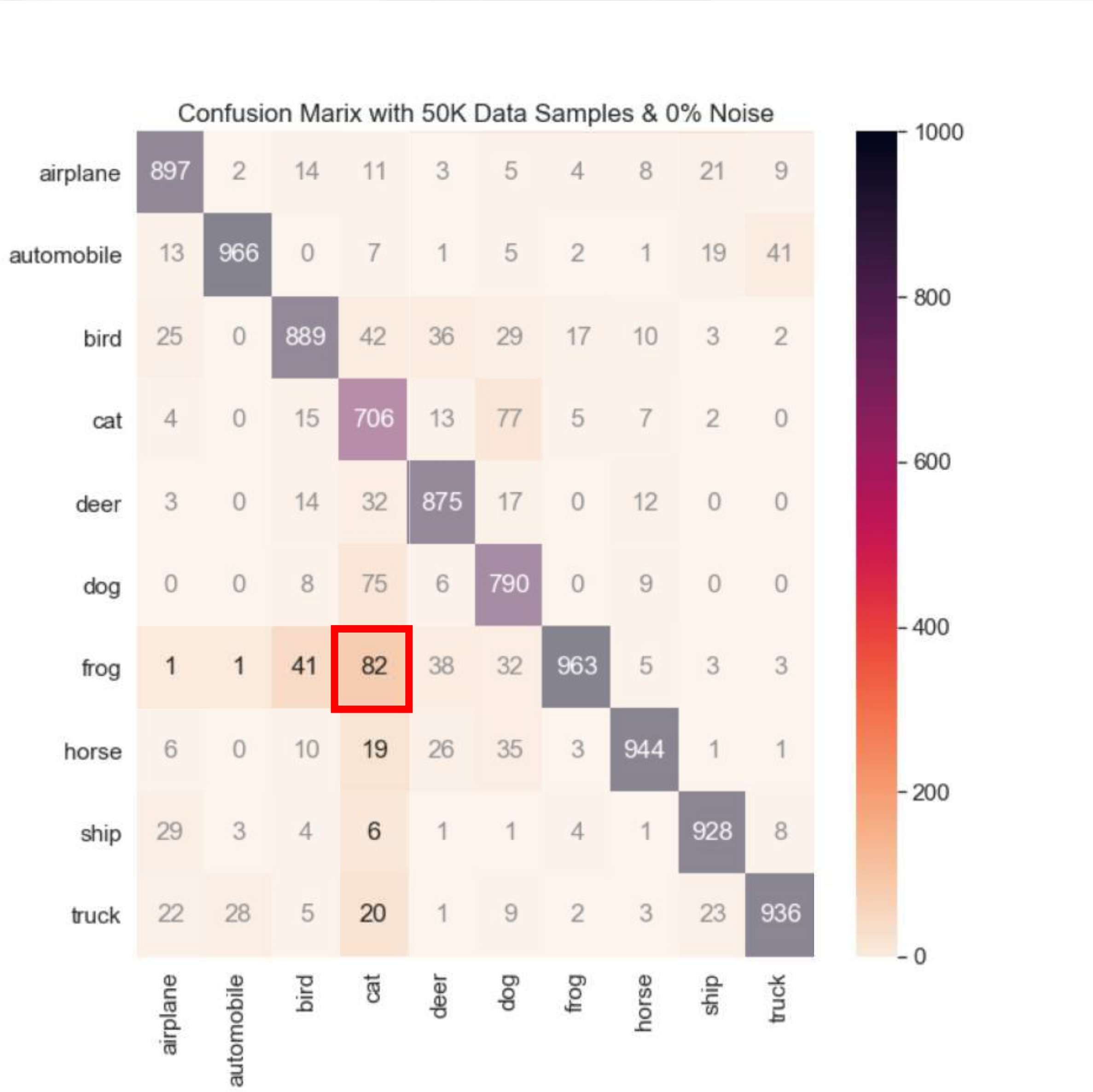
BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**
- **Accuracy varies dramatically across classes**

More details…

- **Lowest accuracy for class 'cat' and 'dog'**
- **Class 'bird' has a fairly high accuracy**
- **Higher confusion for 'cat' → 'frog' and for 'cat' → 'dog'**



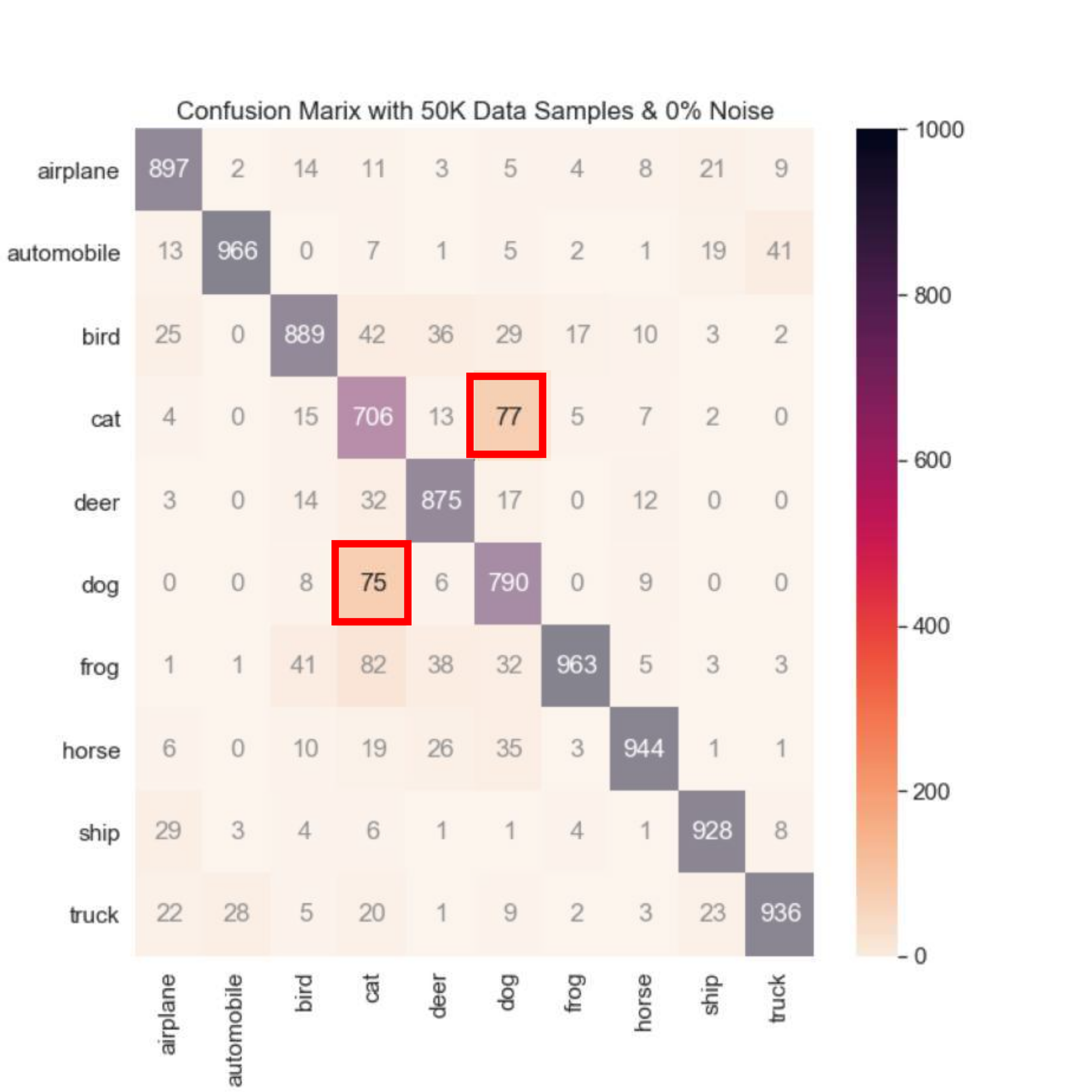
BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**
- **Accuracy varies dramatically across classes**

More details…

- **Lowest accuracy for class 'cat' and 'dog'**
- **Class 'bird' has a fairly high accuracy**
- **Higher confusion for 'cat' ➔ 'frog' and for 'cat' ➔ 'dog'**
- **As easy to mistake a cat for a dog, than a dog for a cat**



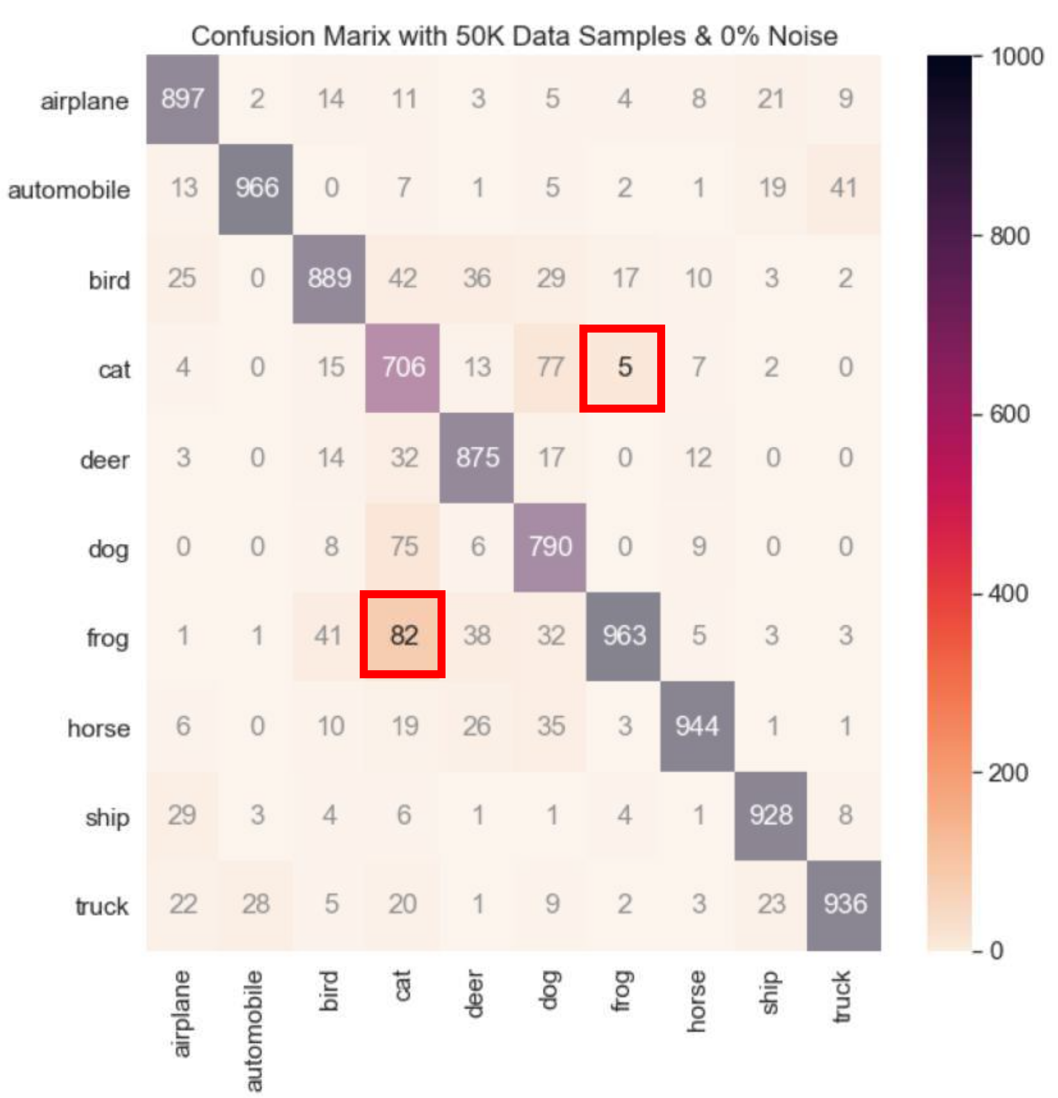
BASELINE RESULTS

Results

- **Baseline accuracy: 89% (across all classes)**
- **Accuracy varies dramatically across classes**

More details…

- **Lowest accuracy for class 'cat' and 'dog'**
- **Class 'bird' has a fairly high accuracy**
- **Higher confusion for 'cat' → 'frog' and for 'cat' → 'dog'**
- **As easy to mistake a cat for a dog, than a dog for a cat**
- **Easier to mistake a cat for a frog, than a frog for a cat**
- **Confusion is NOT SYMMETRICAL across classes**



EXPERIMENT #1: LABELING POLLUTION

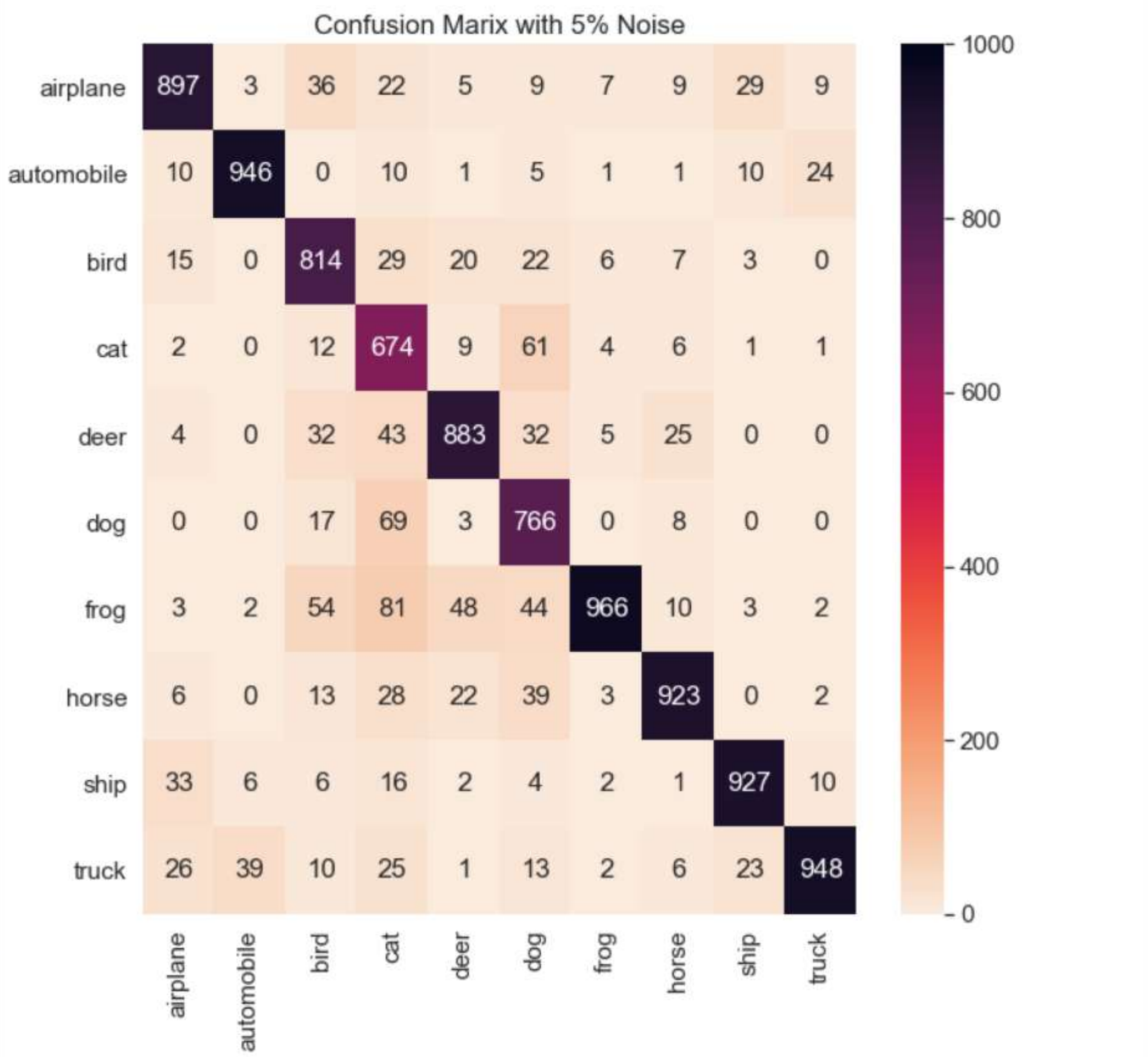
Goal:

Study impact of noise in labeling process on model performance

Protocol:

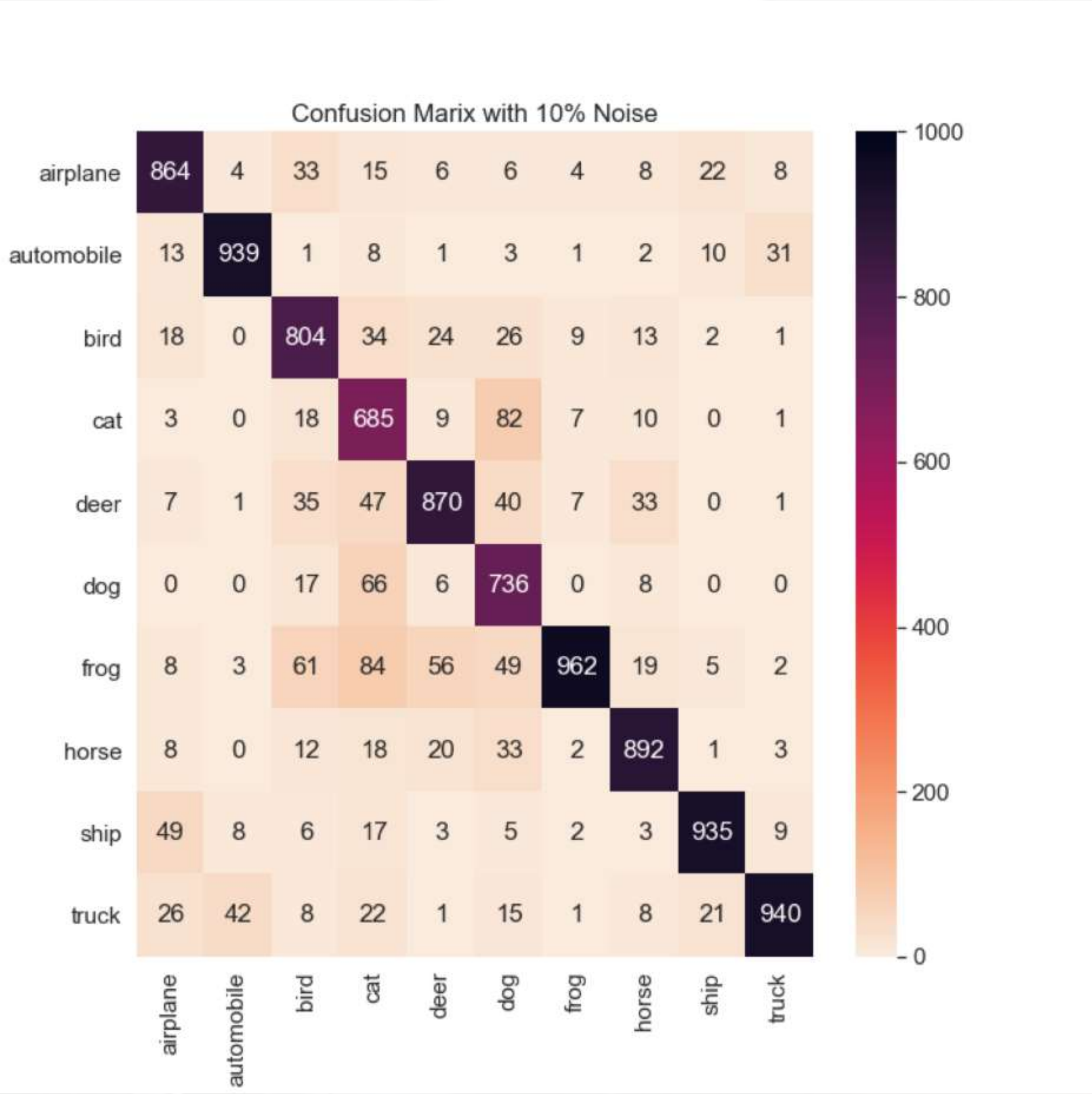
- We randomly shuffle the labels within the selected subset
- We select n% of the 50,000 records (full dataset)
  - Those records are chosen randomly, with no distinction of the class
- We repeat the same experiment 5 times for each amount to eliminate noisy results
  - Different levels of noise of data might lead to different results
  - We chose 5 times because of compute power limitations
- We observe the accuracy and the confusion matrix

EXPERIMENT #1: LABELING POLLUTION



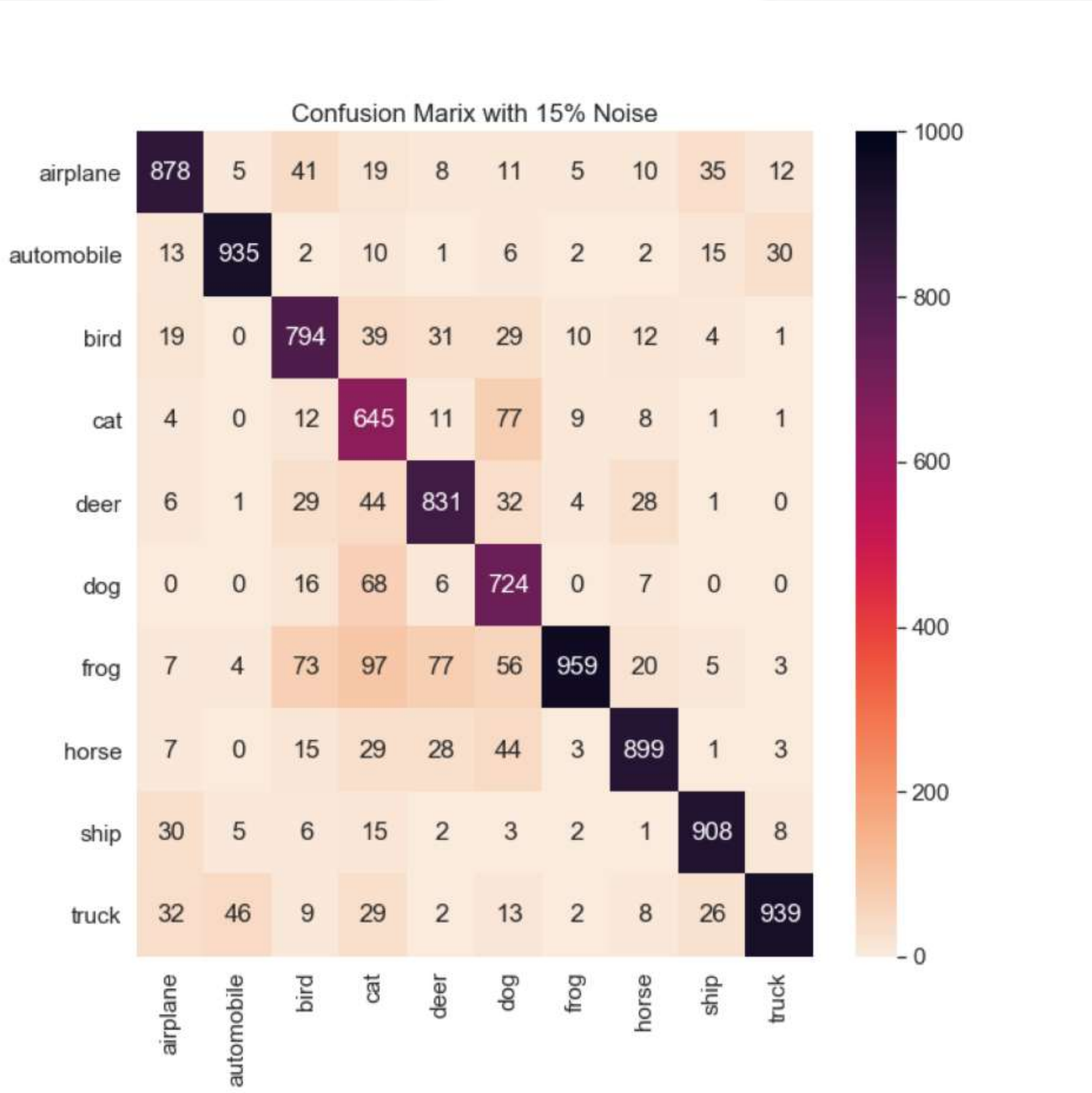
Average Confusion Matrix with 5% noisy labels

EXPERIMENT #1: LABELING POLLUTION



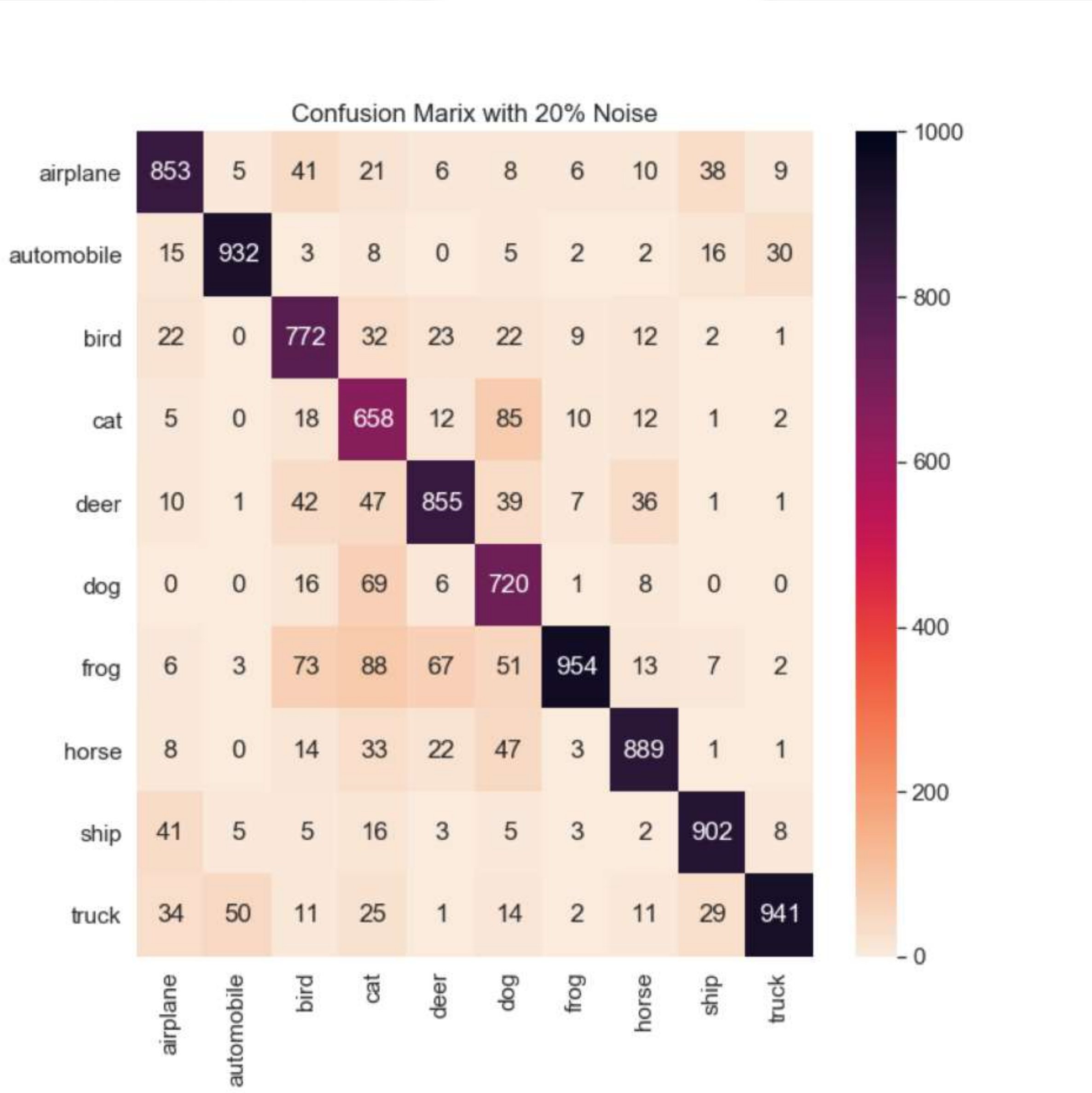
Average Confusion Matrix with 10% noisy labels

EXPERIMENT #1: LABELING POLLUTION



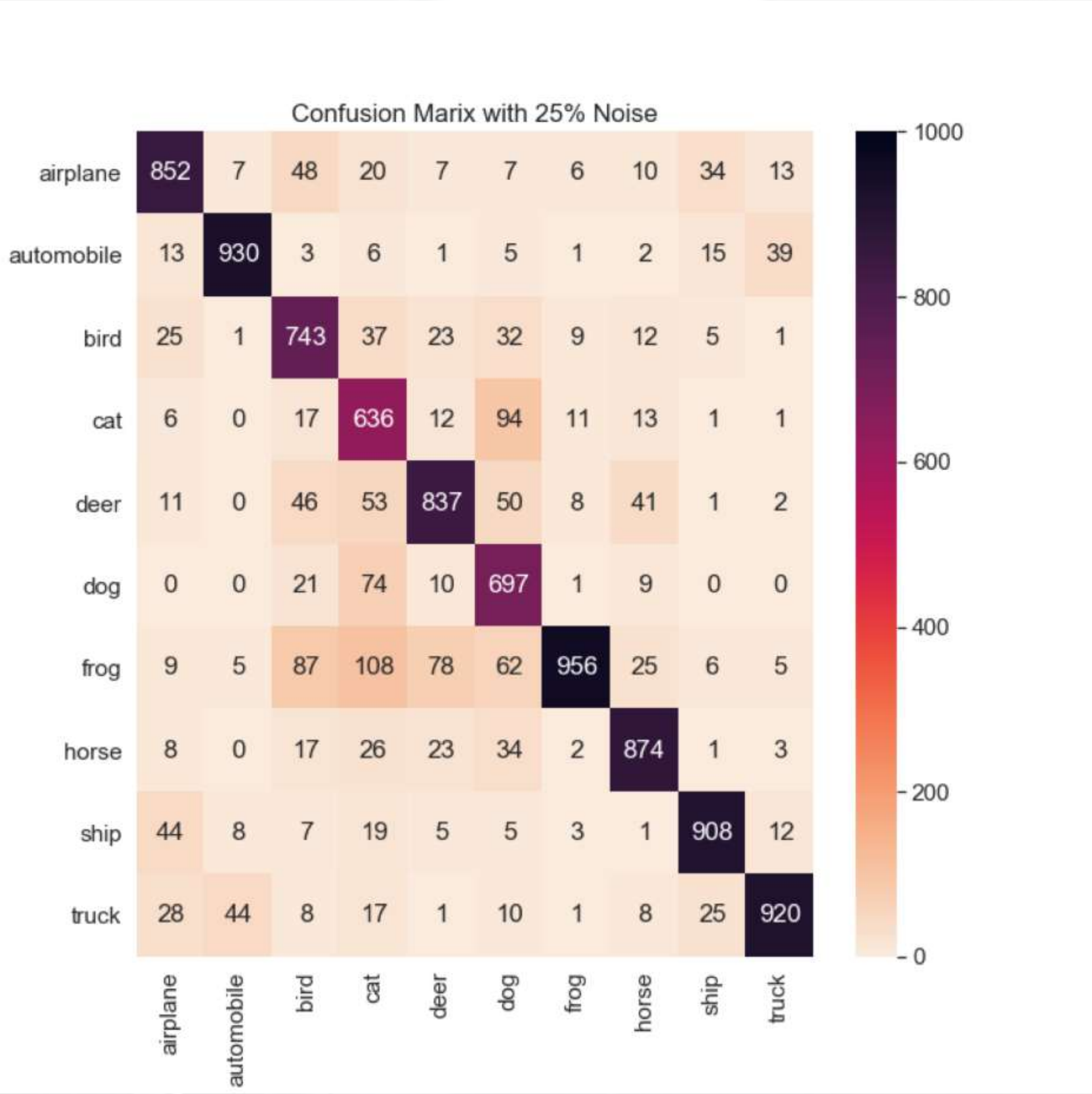
Average Confusion Matrix with 15% noisy labels

EXPERIMENT #1: LABELING POLLUTION



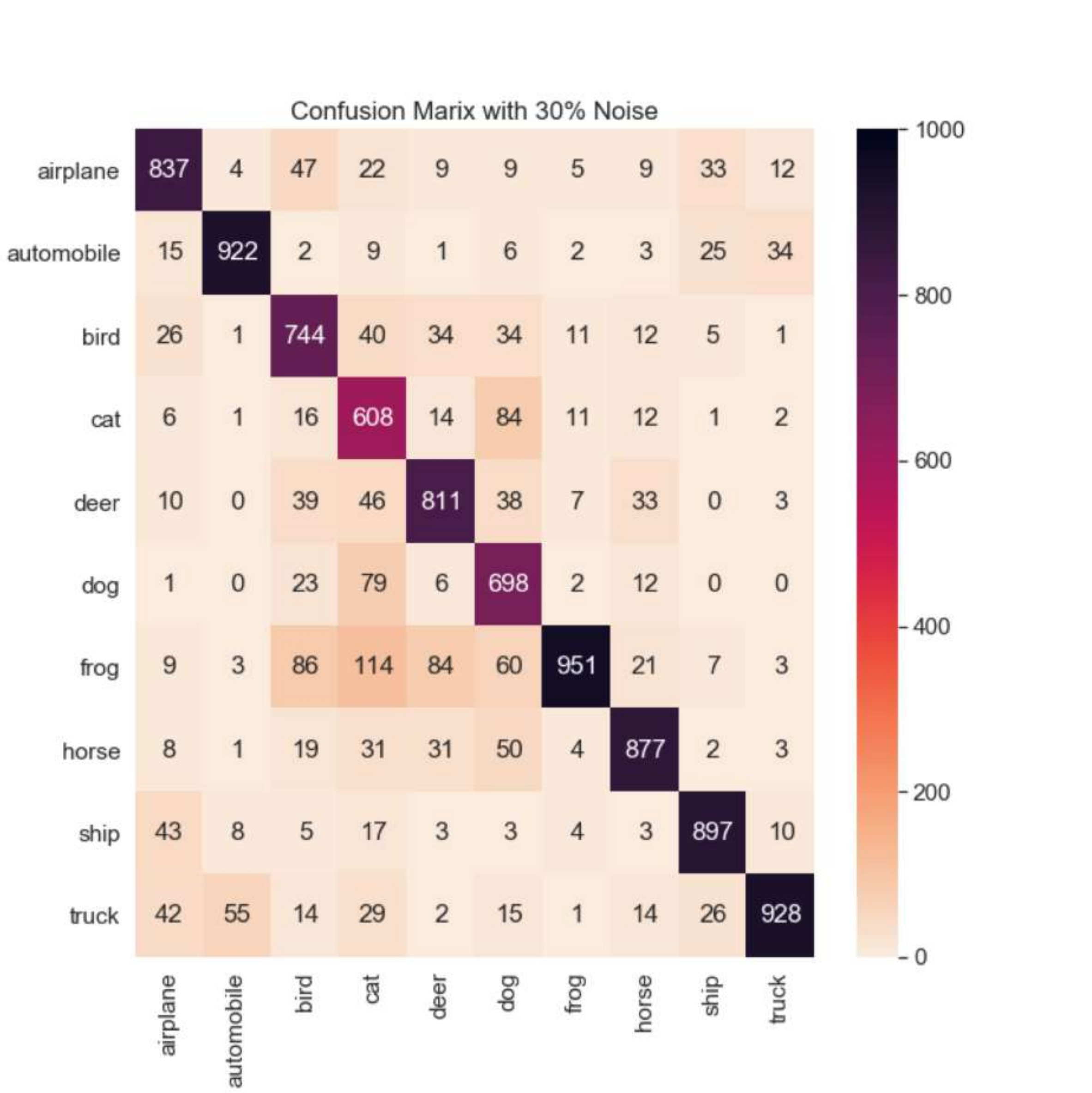
Average Confusion Matrix with 20% noisy labels

EXPERIMENT #1: LABELING POLLUTION



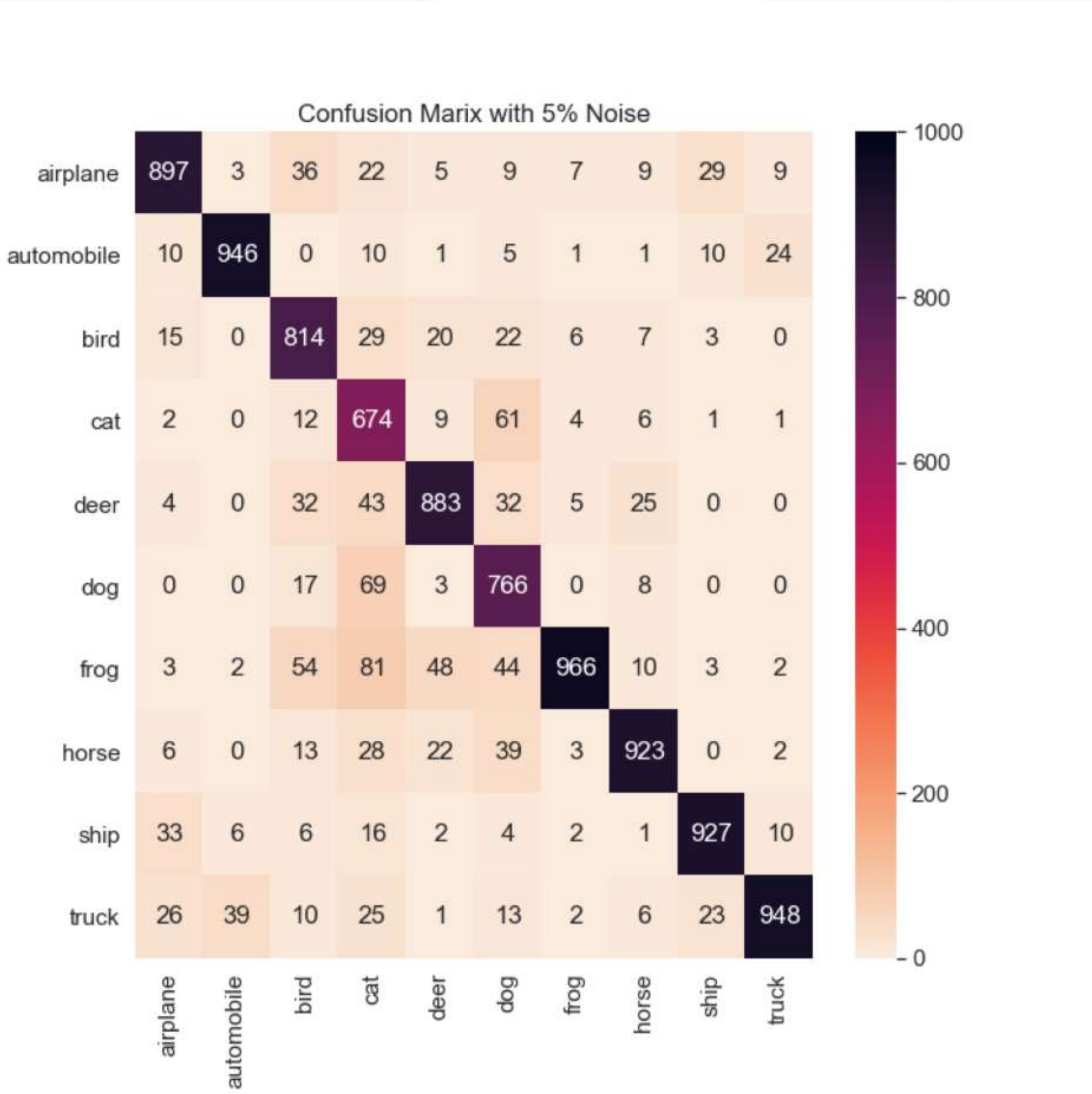
Average Confusion Matrix with 25% noisy labels

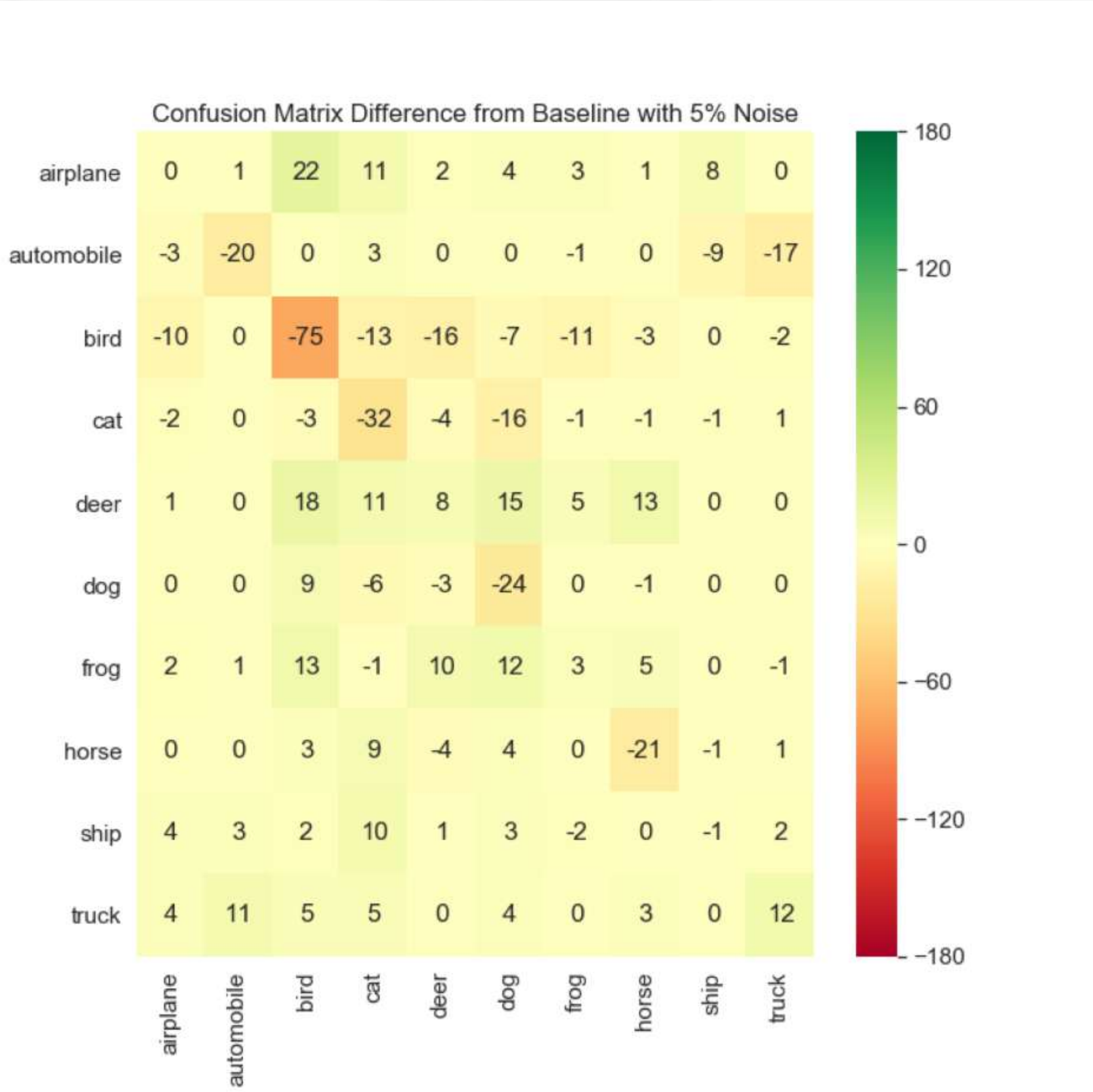
EXPERIMENT #1: LABELING POLLUTION



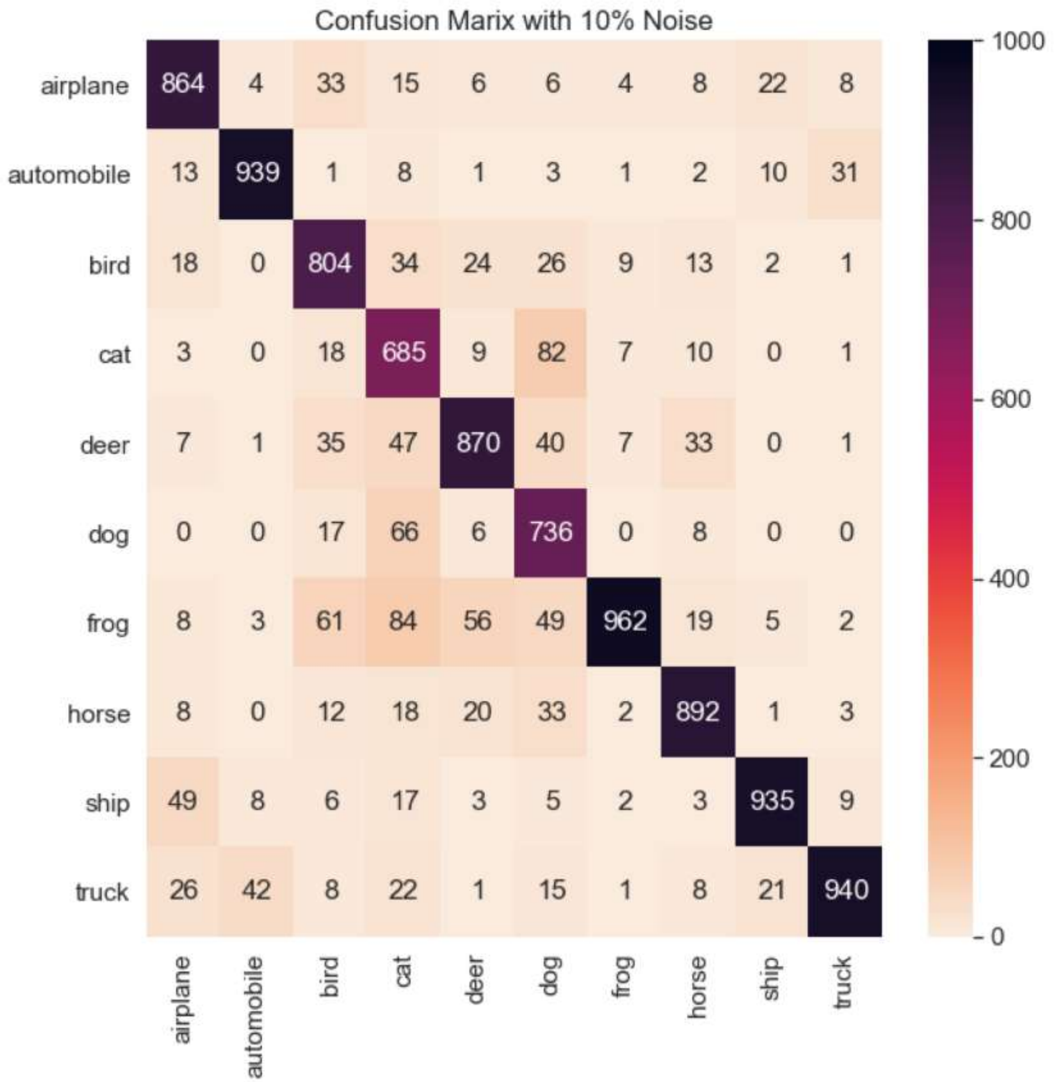
Average Confusion Matrix with 30% noisy labels

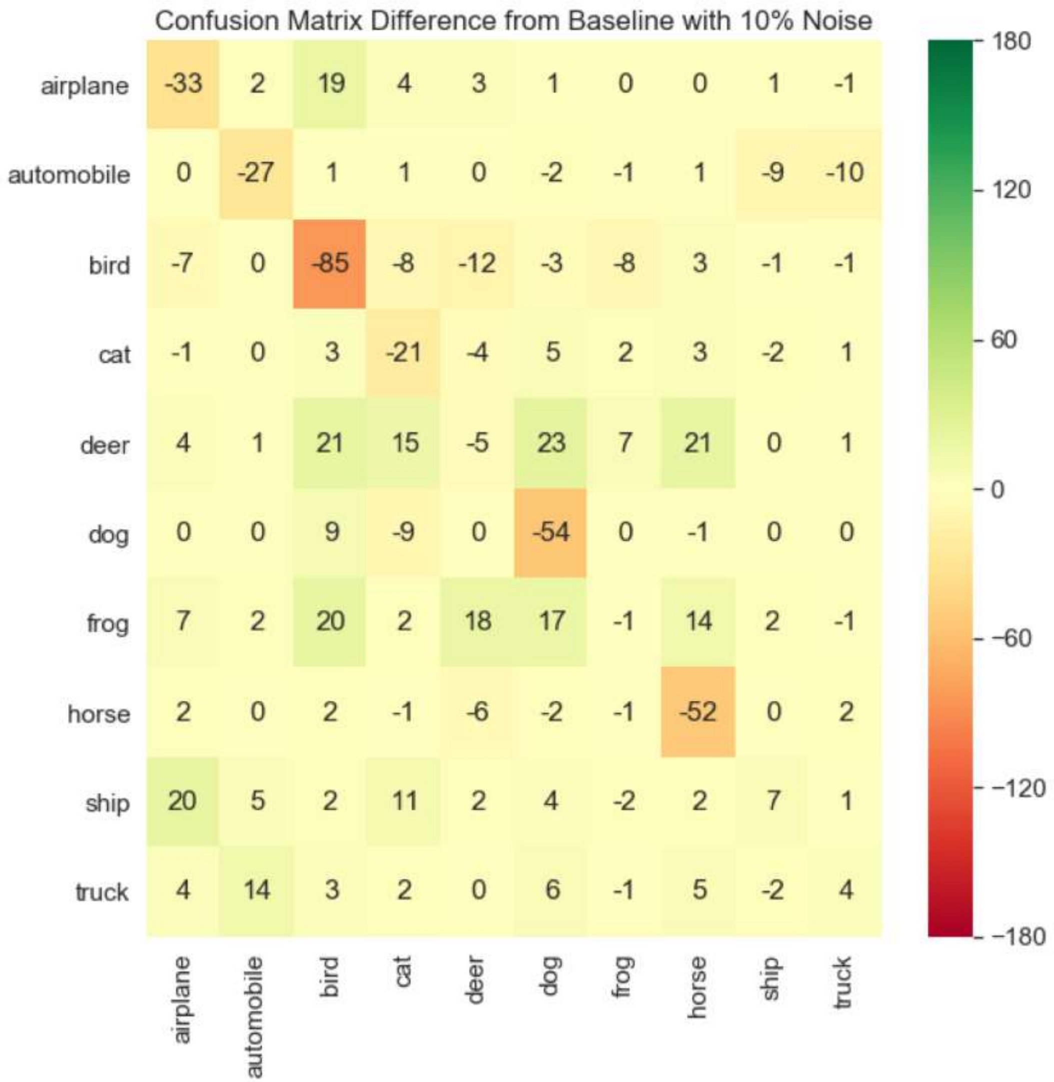
EXPERIMENT #1: LABELING POLLUTION



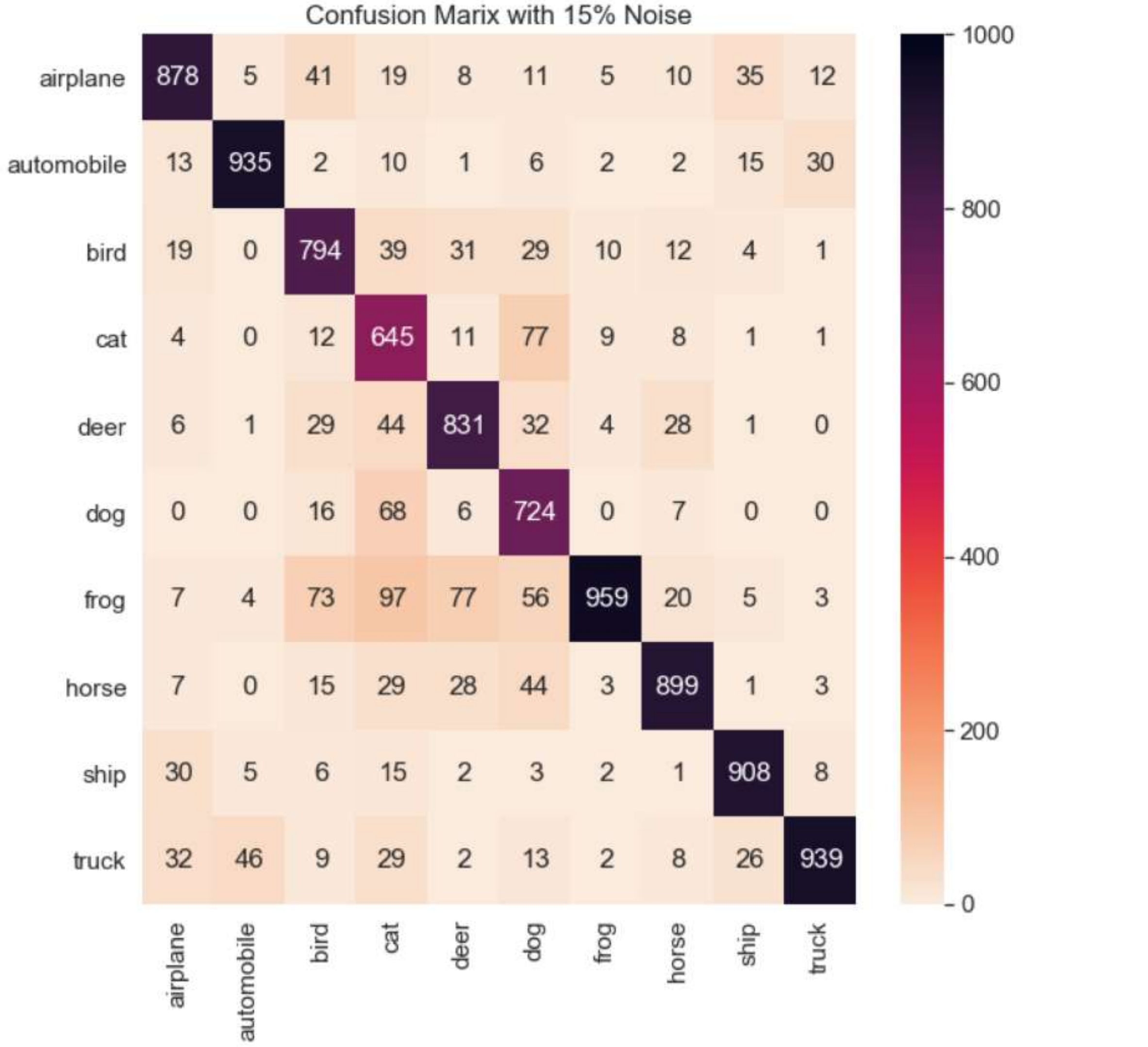


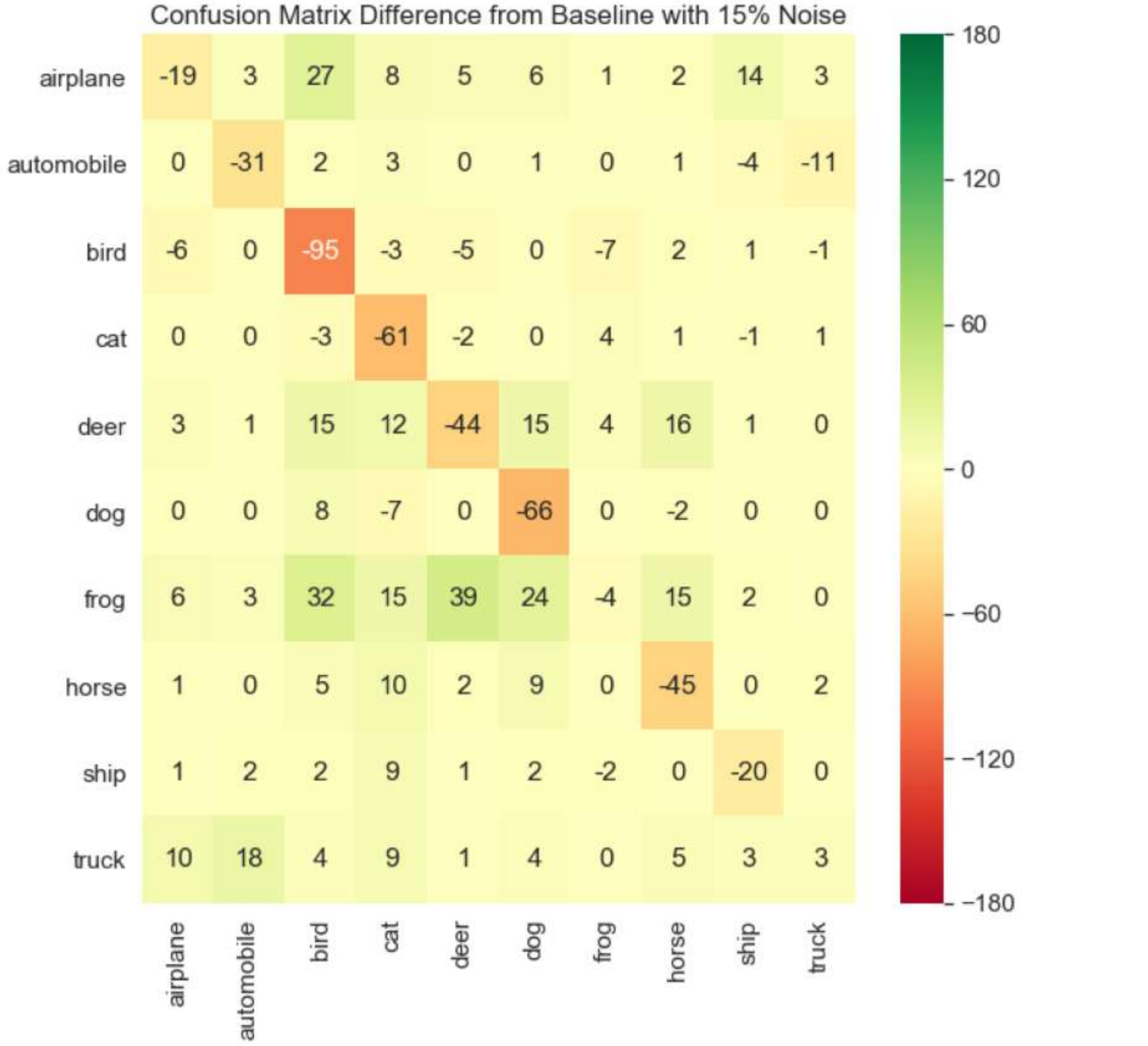
EXPERIMENT #1: LABELING POLLUTION



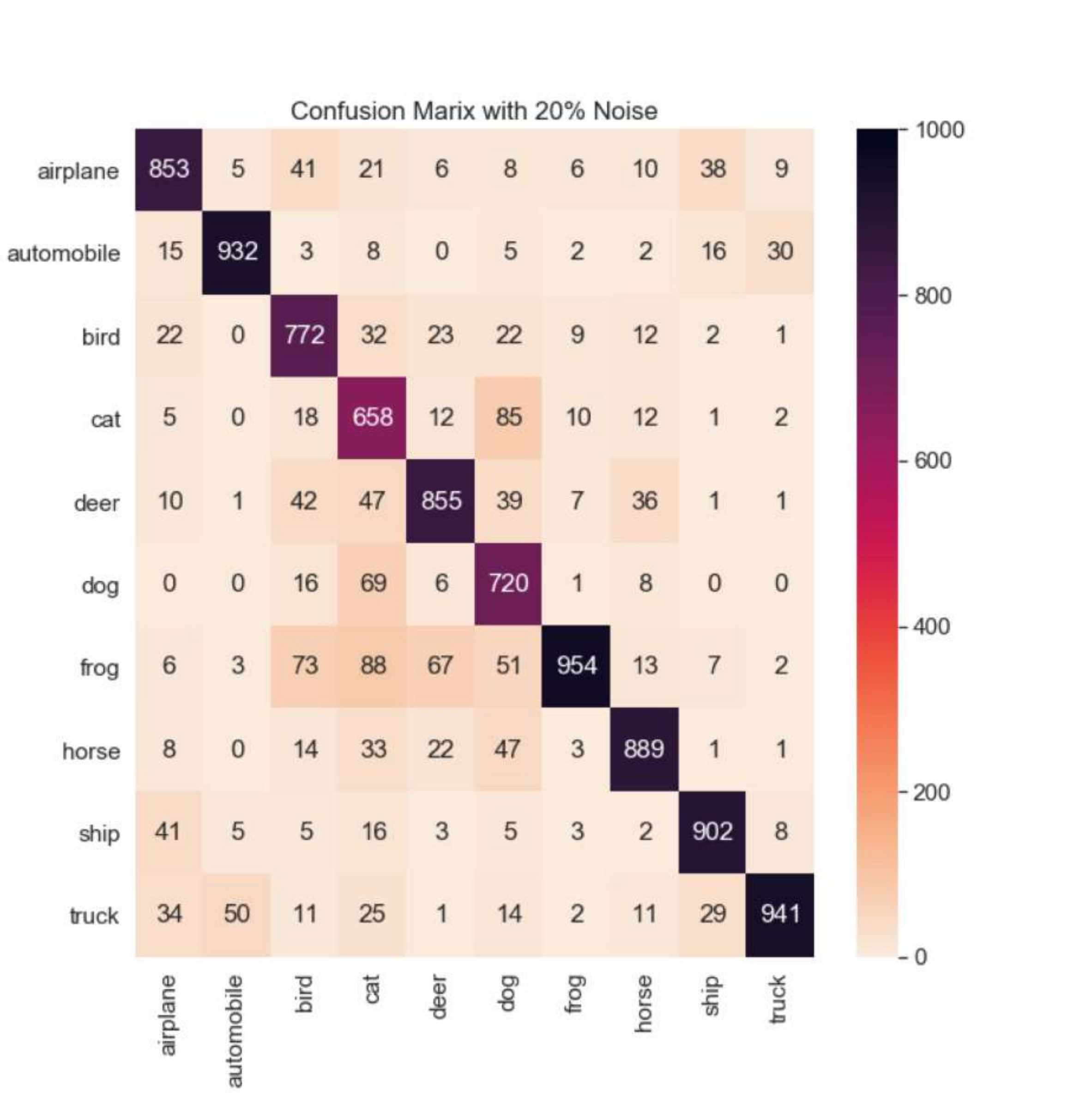


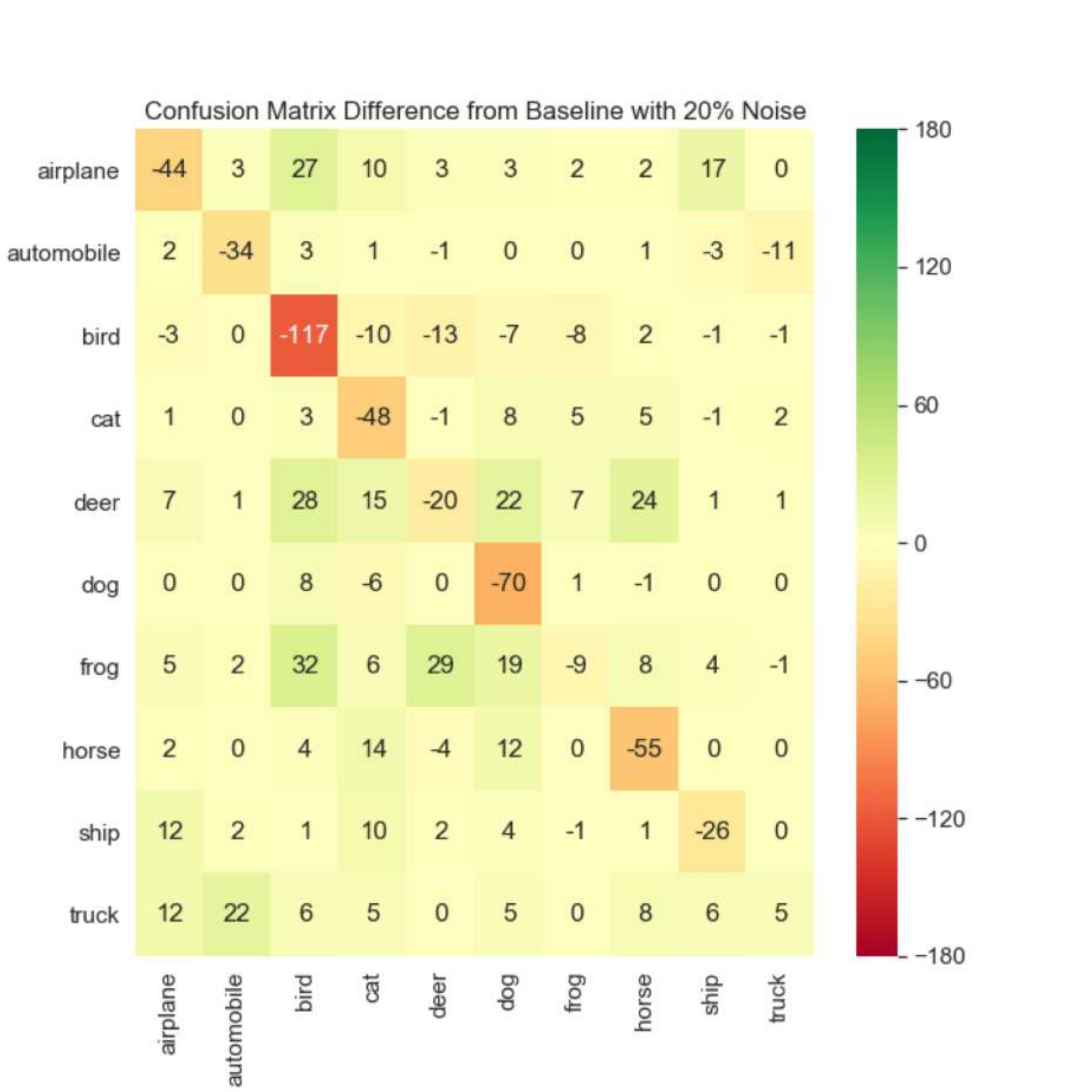
EXPERIMENT #1: LABELING POLLUTION



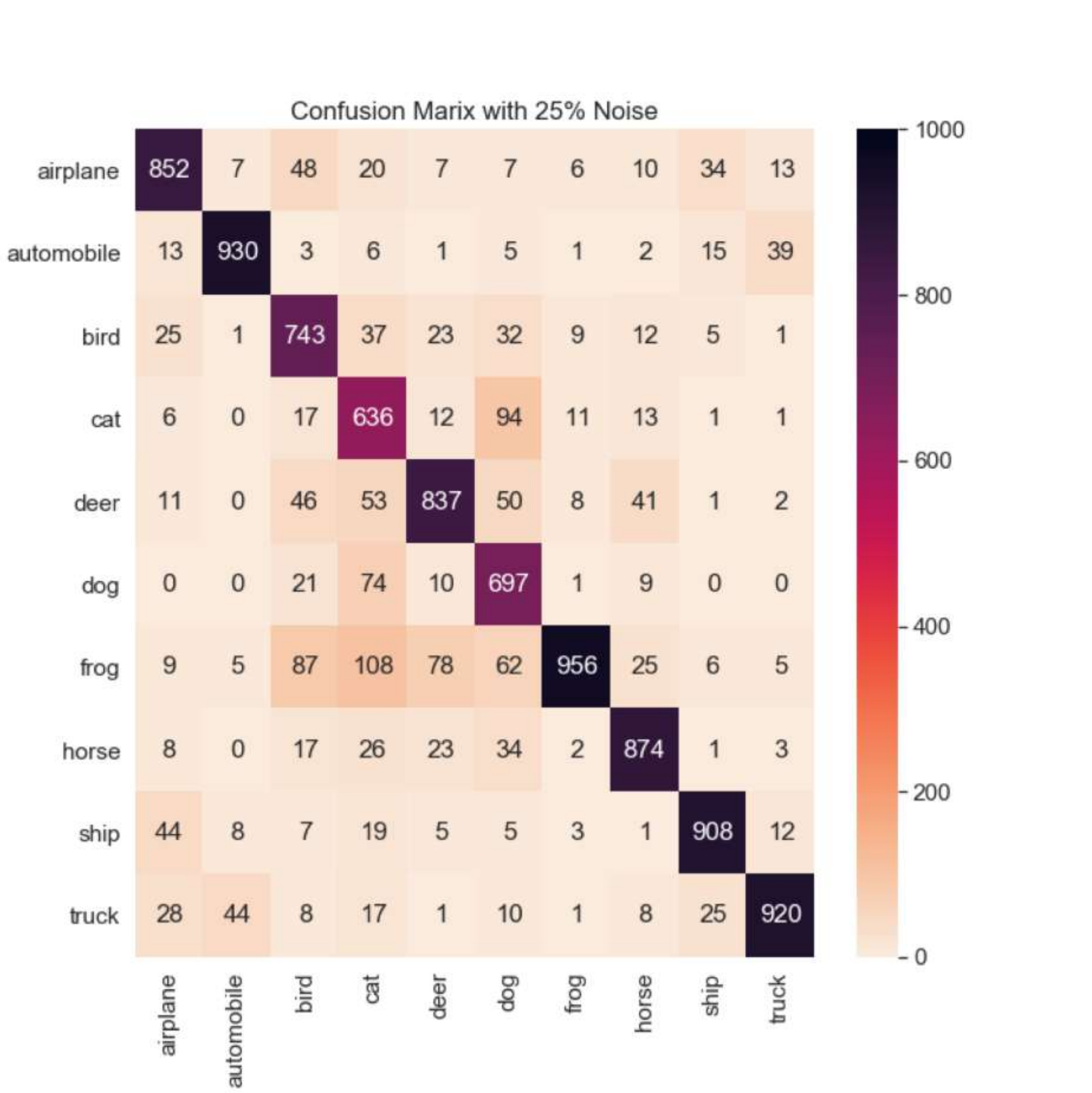


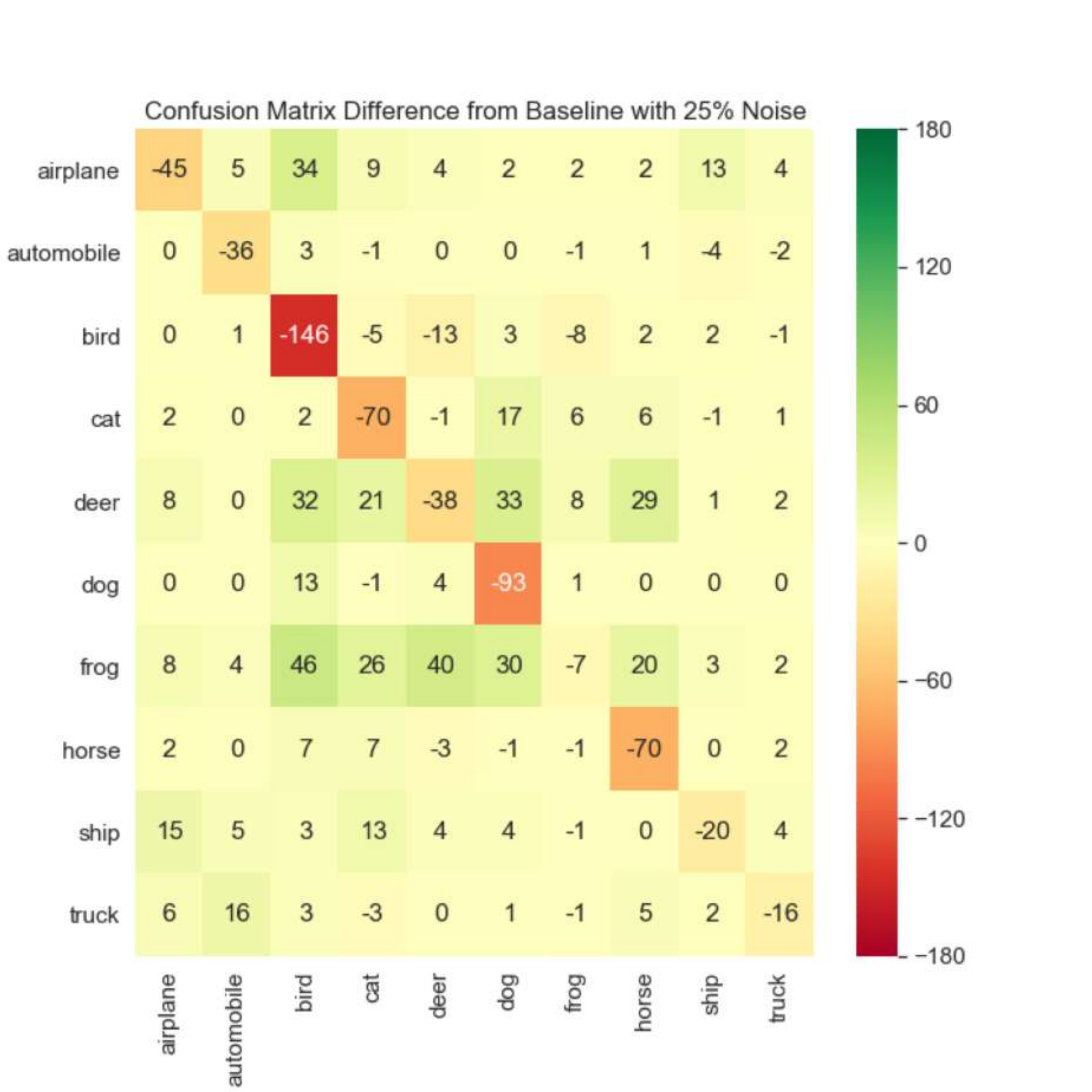
EXPERIMENT #1: LABELING POLLUTION



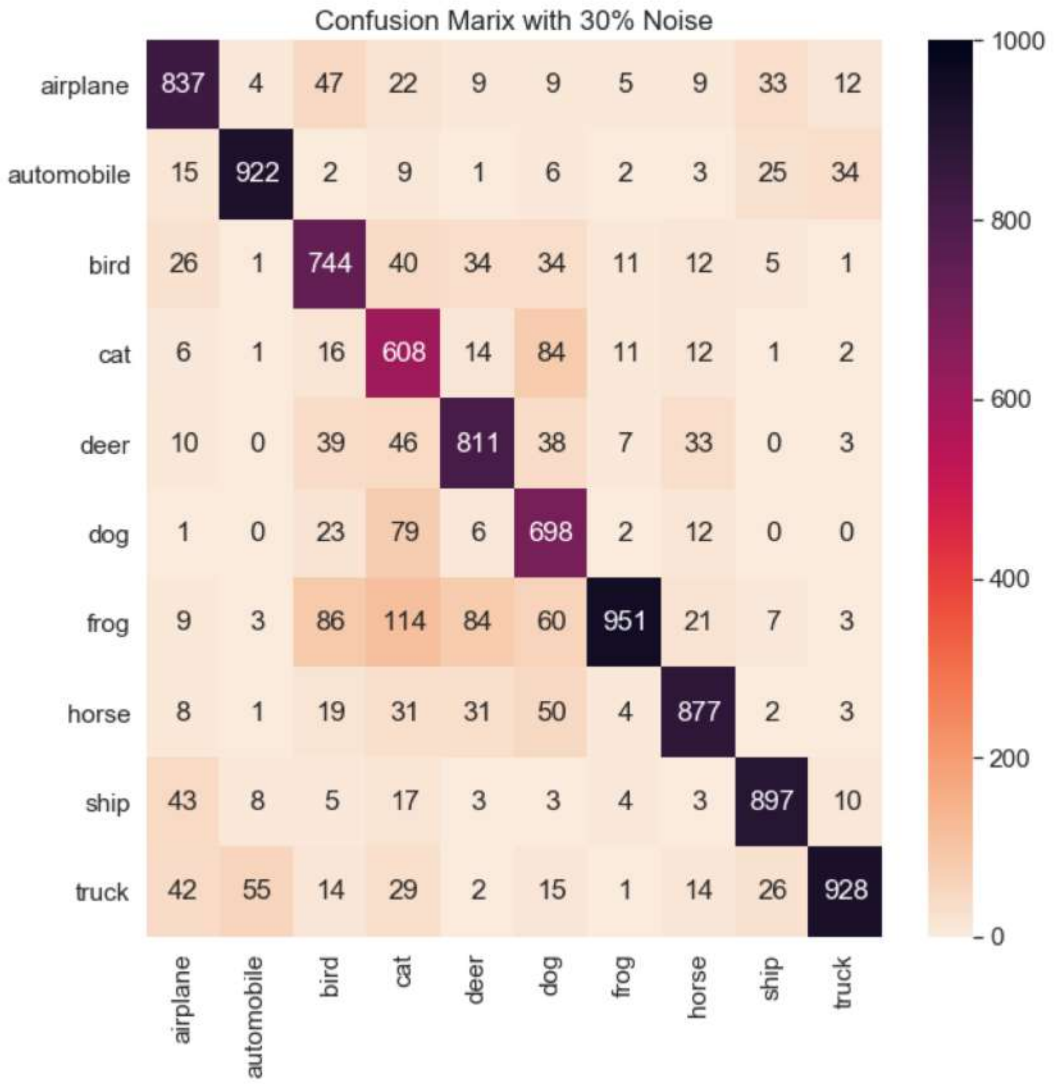


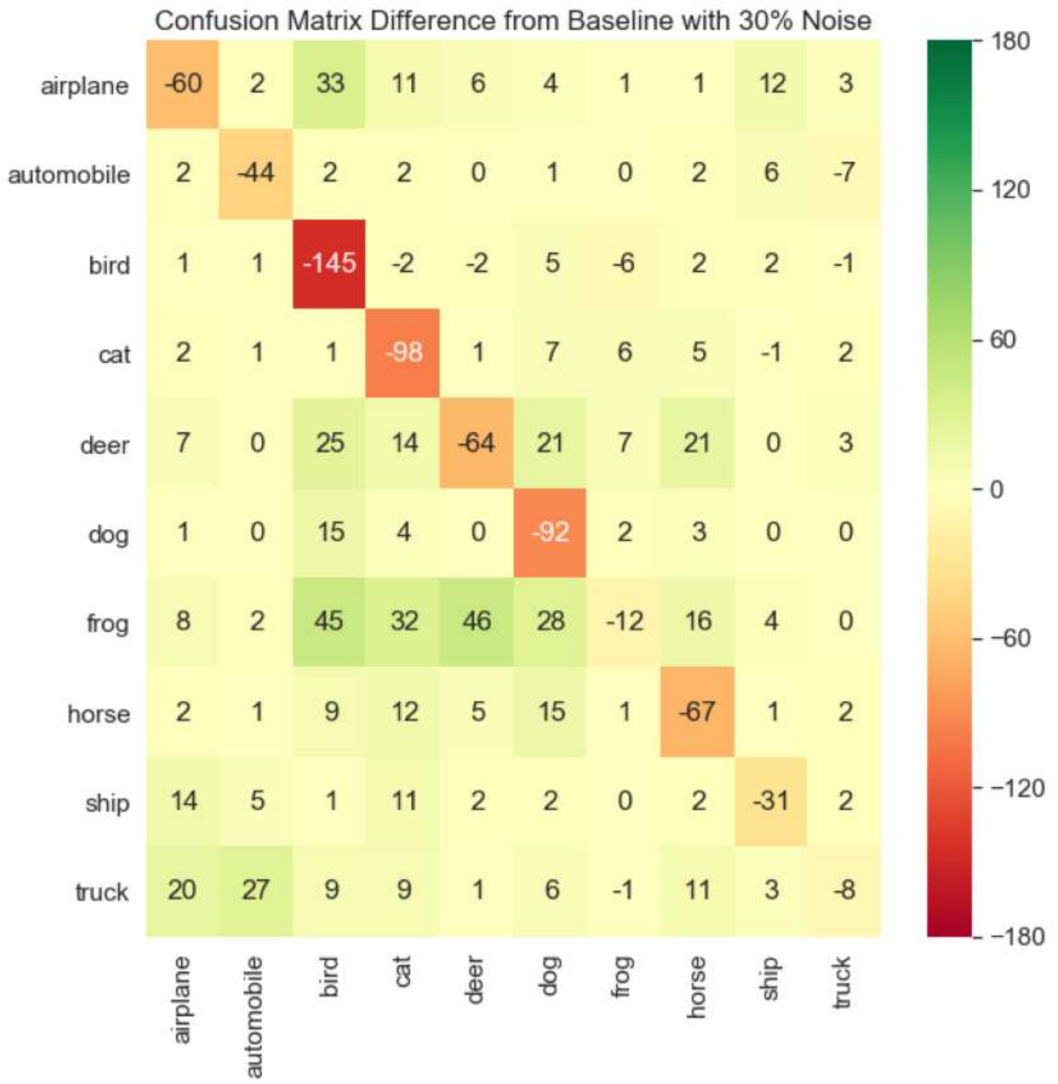
EXPERIMENT #1: LABELING POLLUTION



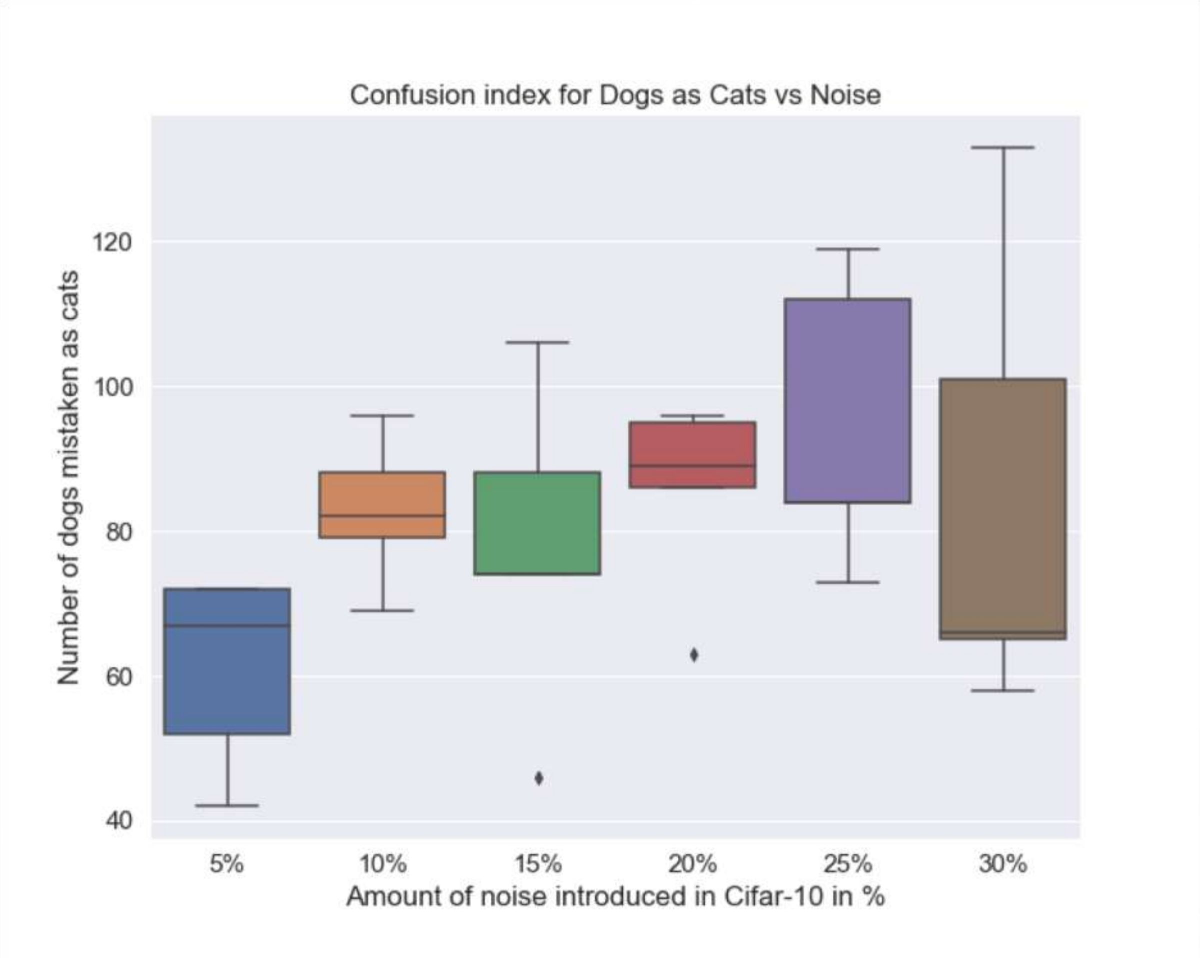


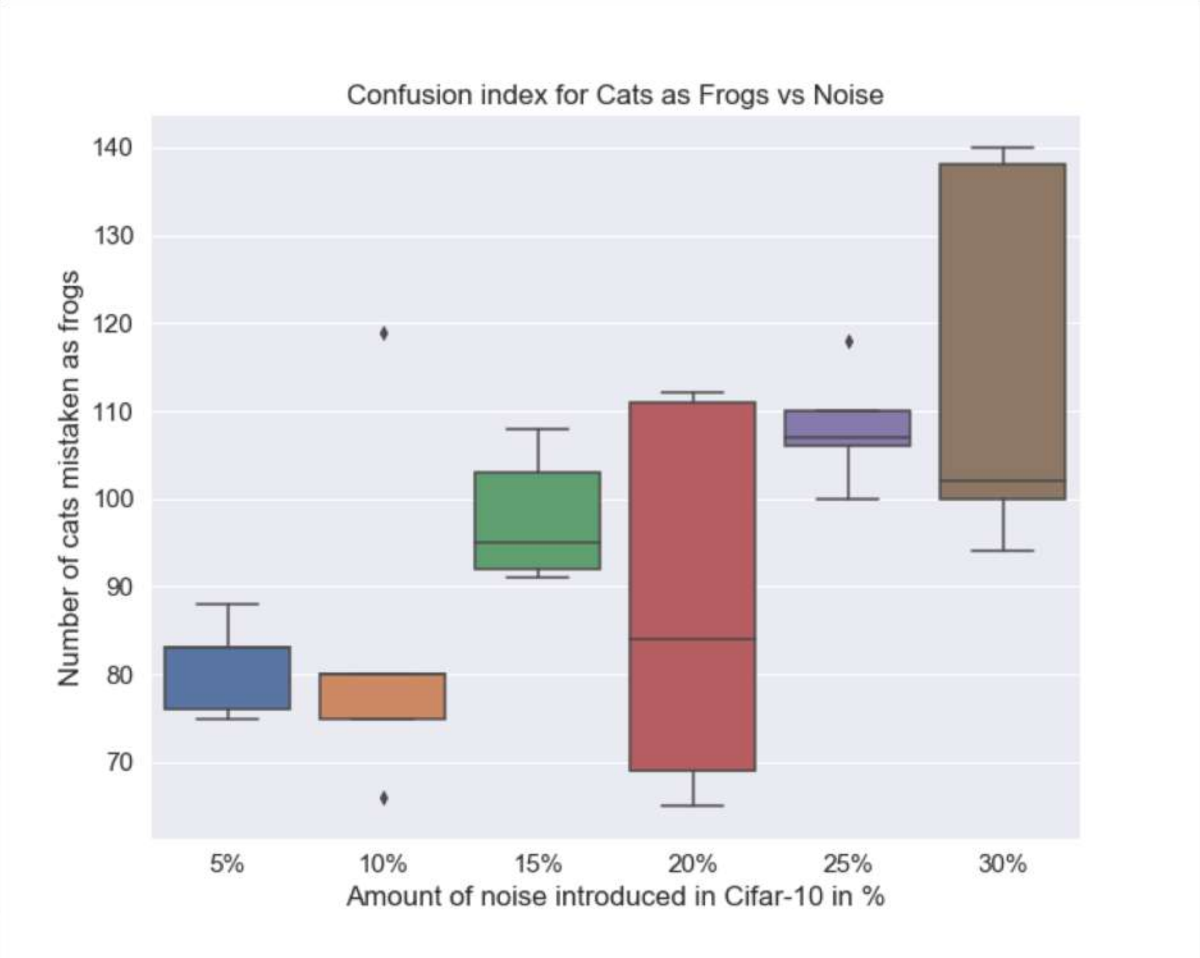
EXPERIMENT #1: LABELING POLLUTION





EXPERIMENT #1: LABELING POLLUTION





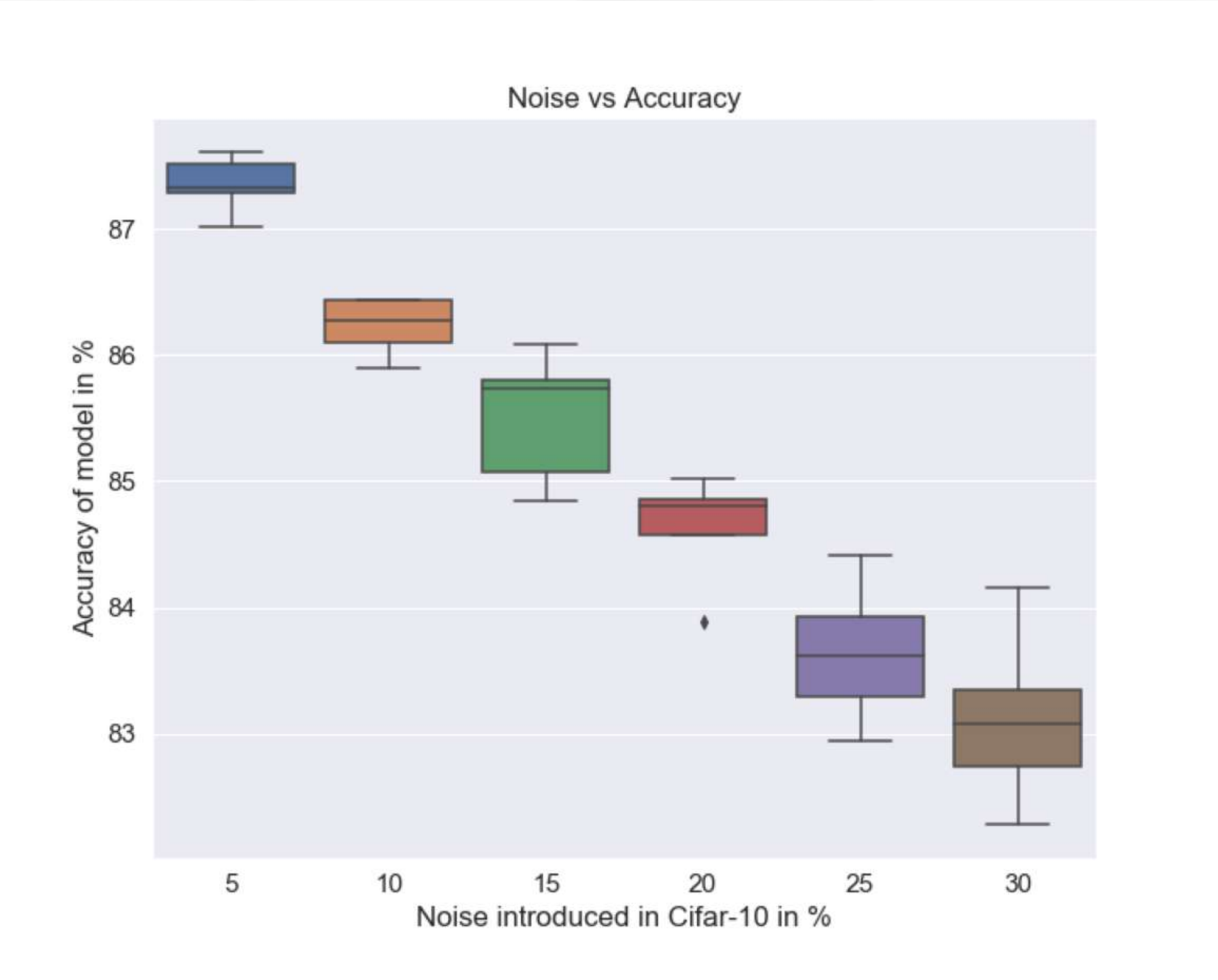
Confusion {dog → cat} vs. labeling noise level

Confusion {cat → frog} vs. labeling noise level

EXPERIMENT #1: LABELING POLLUTION

Results

- **Accuracy seems to drop linearly with the amount of noise in the labels**



EXPERIMENT #2: DATA VOLUME REDUCTION

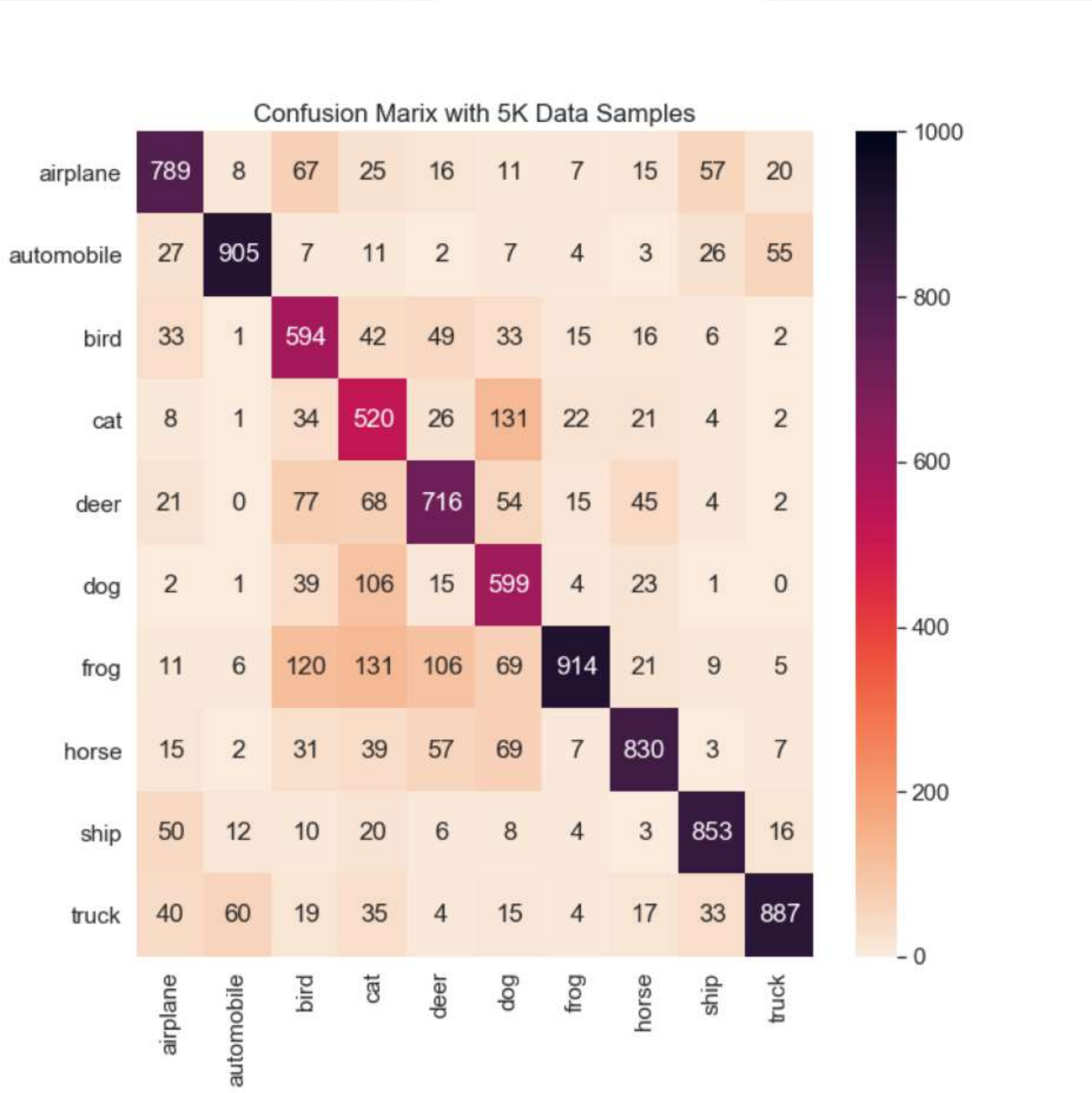
Goal:

Study impact of size of training set on model performance

Protocol:

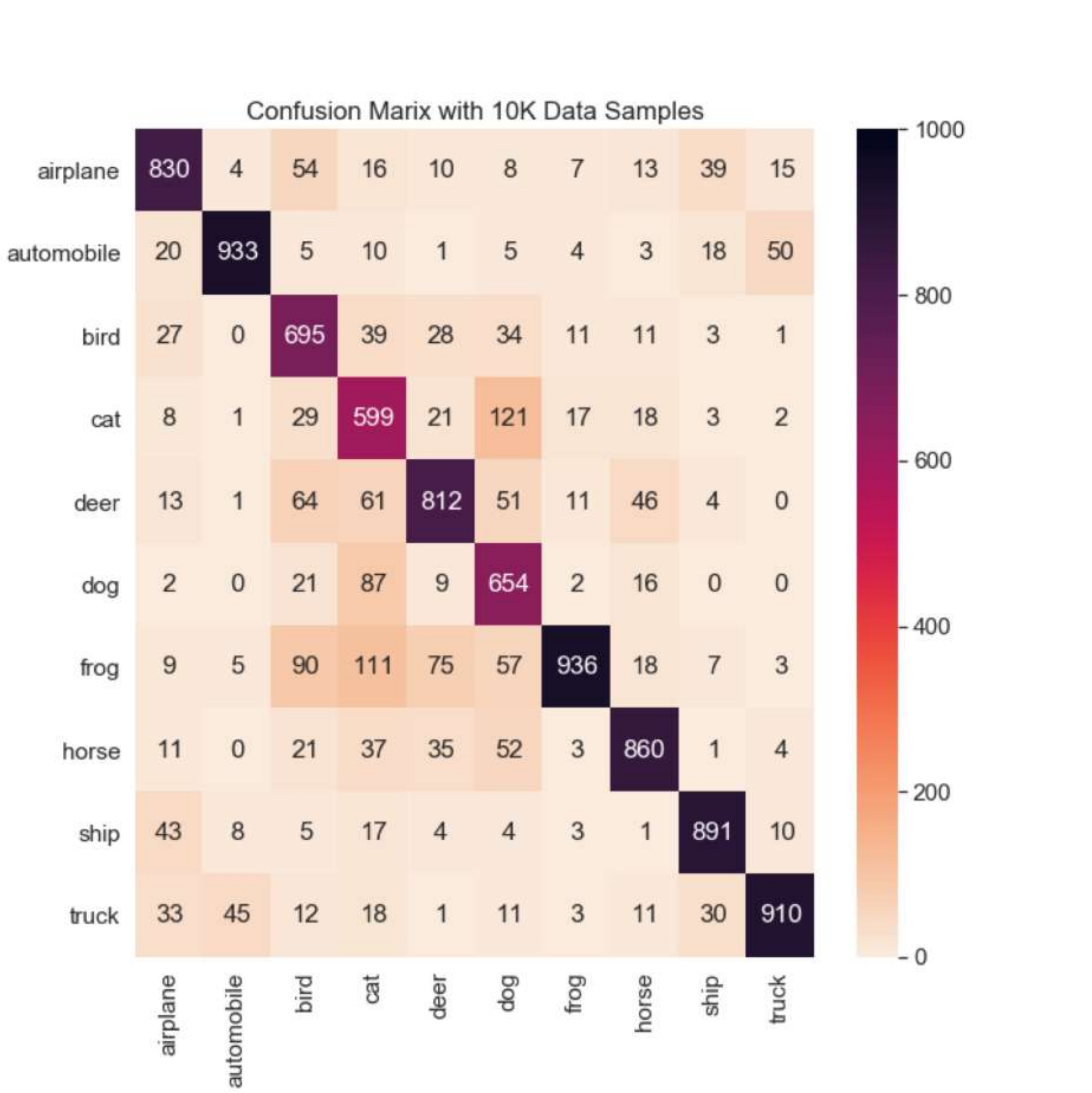
- We increase the size of the training set from 5,000 records (10%) to 50,000 records (full dataset)
  - Those records are chosen randomly
- We repeat the same experiment 5 times for each amount to eliminate noisy results
  - Different subsets of data might lead to different results
  - We chose 5 times because of compute power limitations
- We report the accuracy and the confusion matrix

EXPERIMENT #2: DATA VOLUME REDUCTION



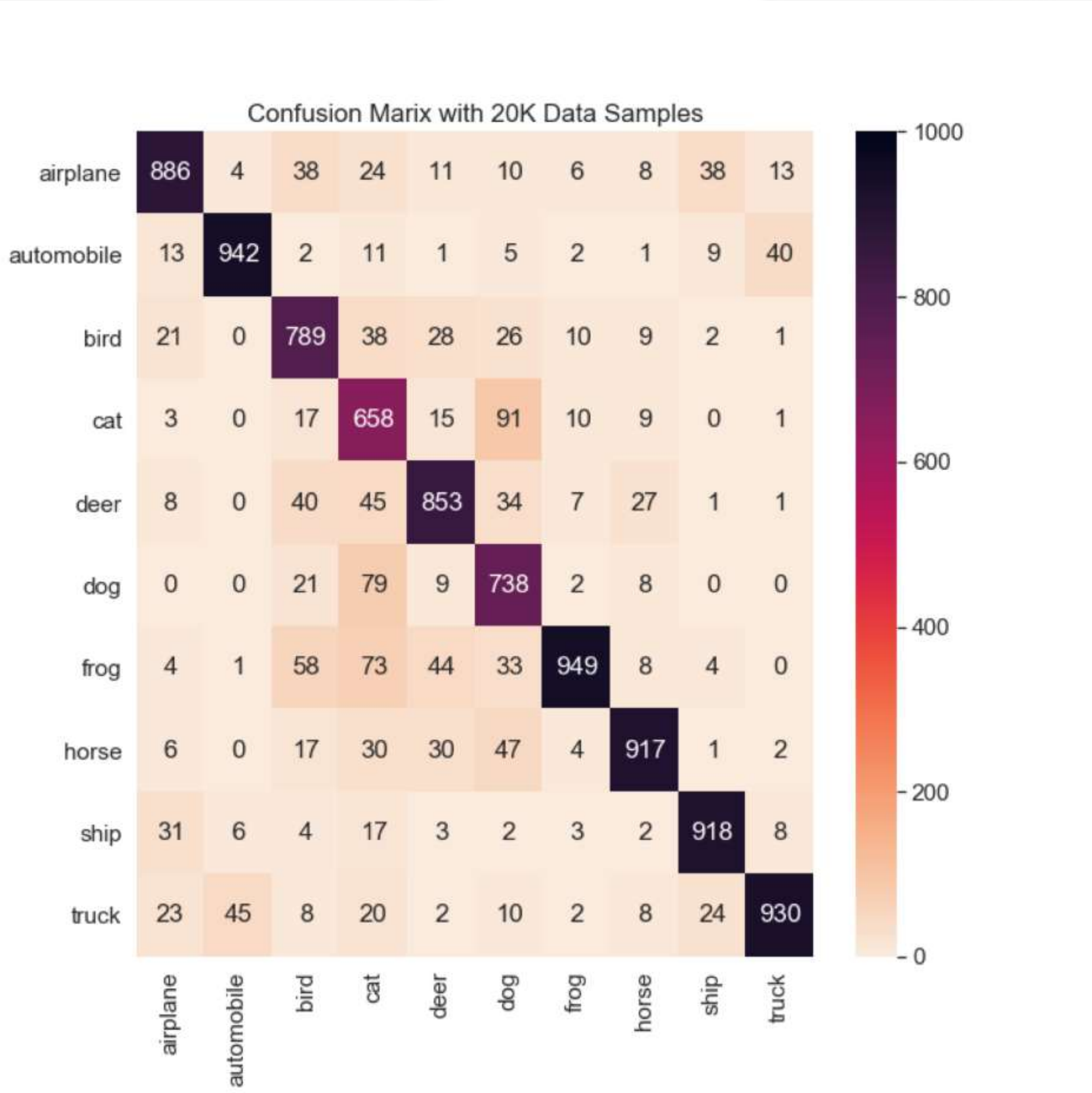
Average Confusion Matrix with size 5k samples

EXPERIMENT #2: DATA VOLUME REDUCTION



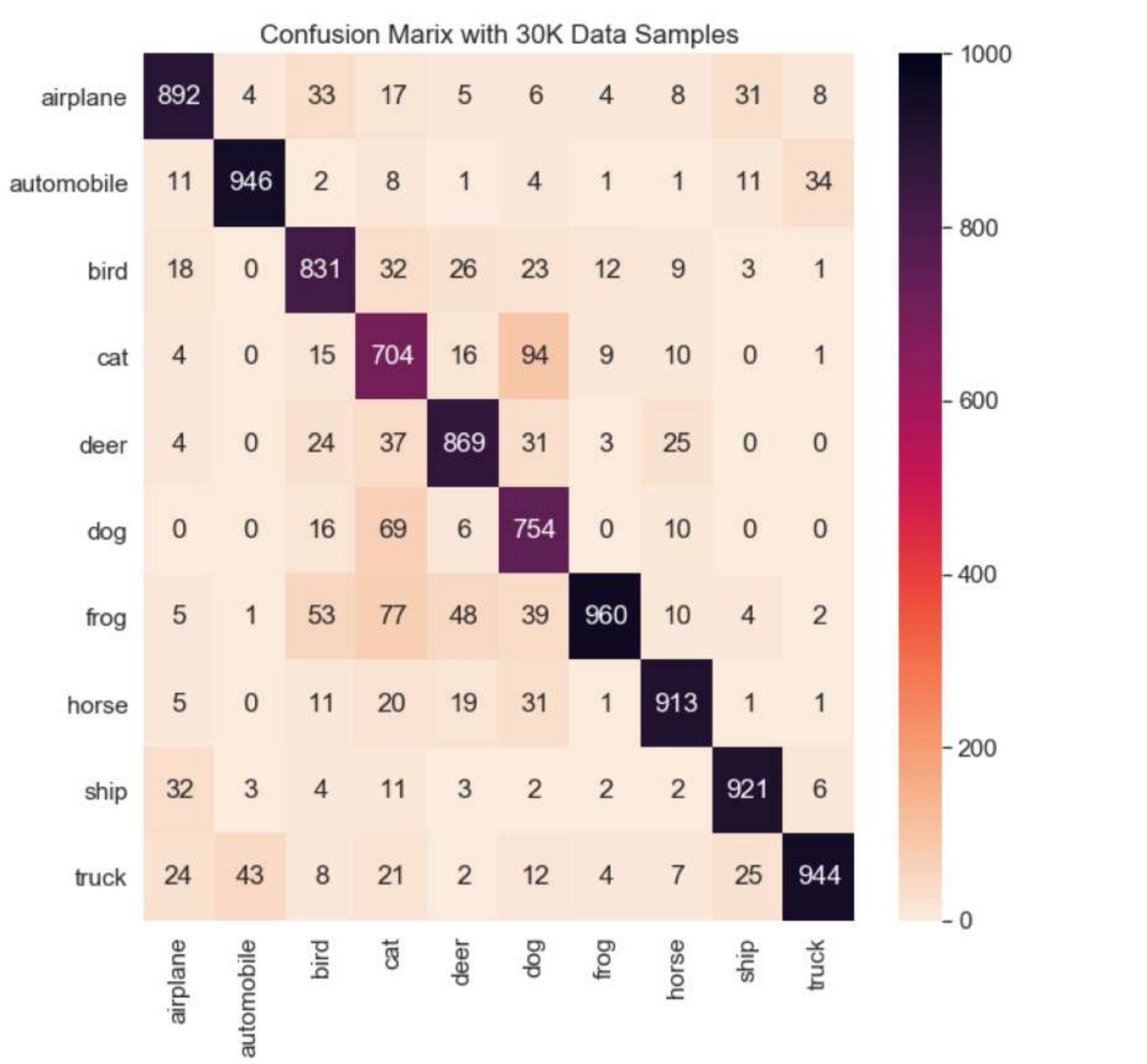
Average Confusion Matrix with size 10k samples

EXPERIMENT #2: DATA VOLUME REDUCTION



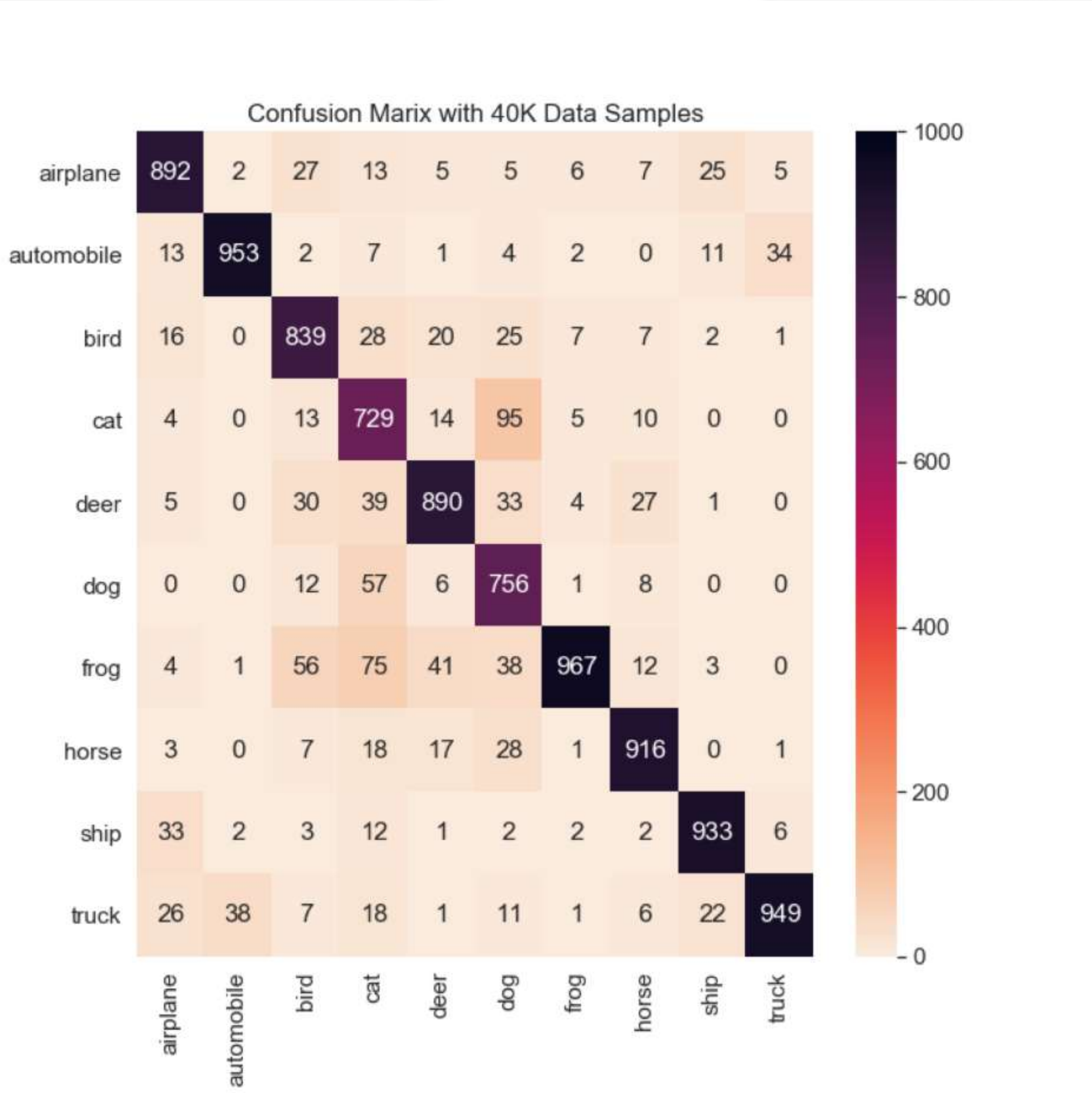
Average Confusion Matrix with size 20k samples

EXPERIMENT #2: DATA VOLUME REDUCTION



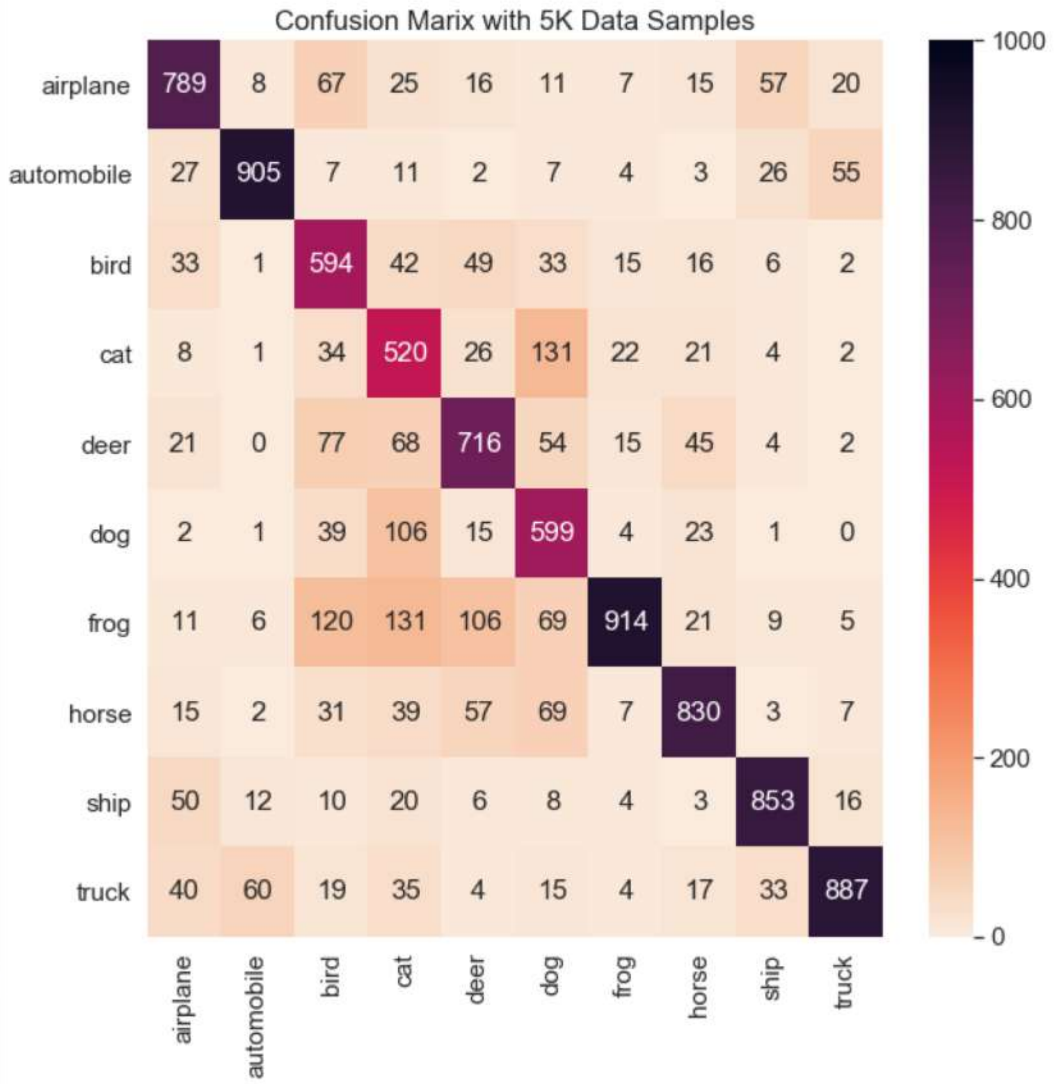
Average Confusion Matrix with size 30k samples

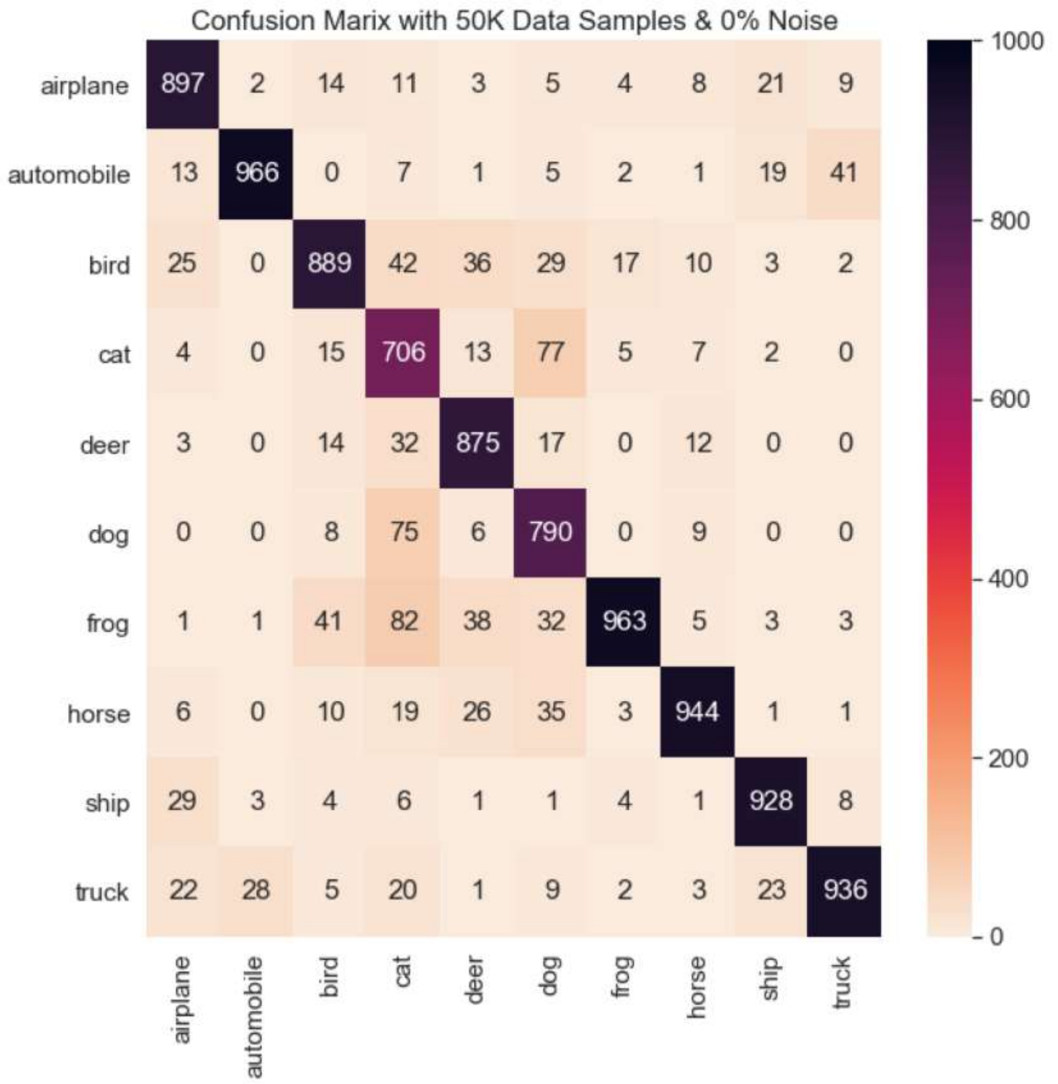
EXPERIMENT #2: DATA VOLUME REDUCTION



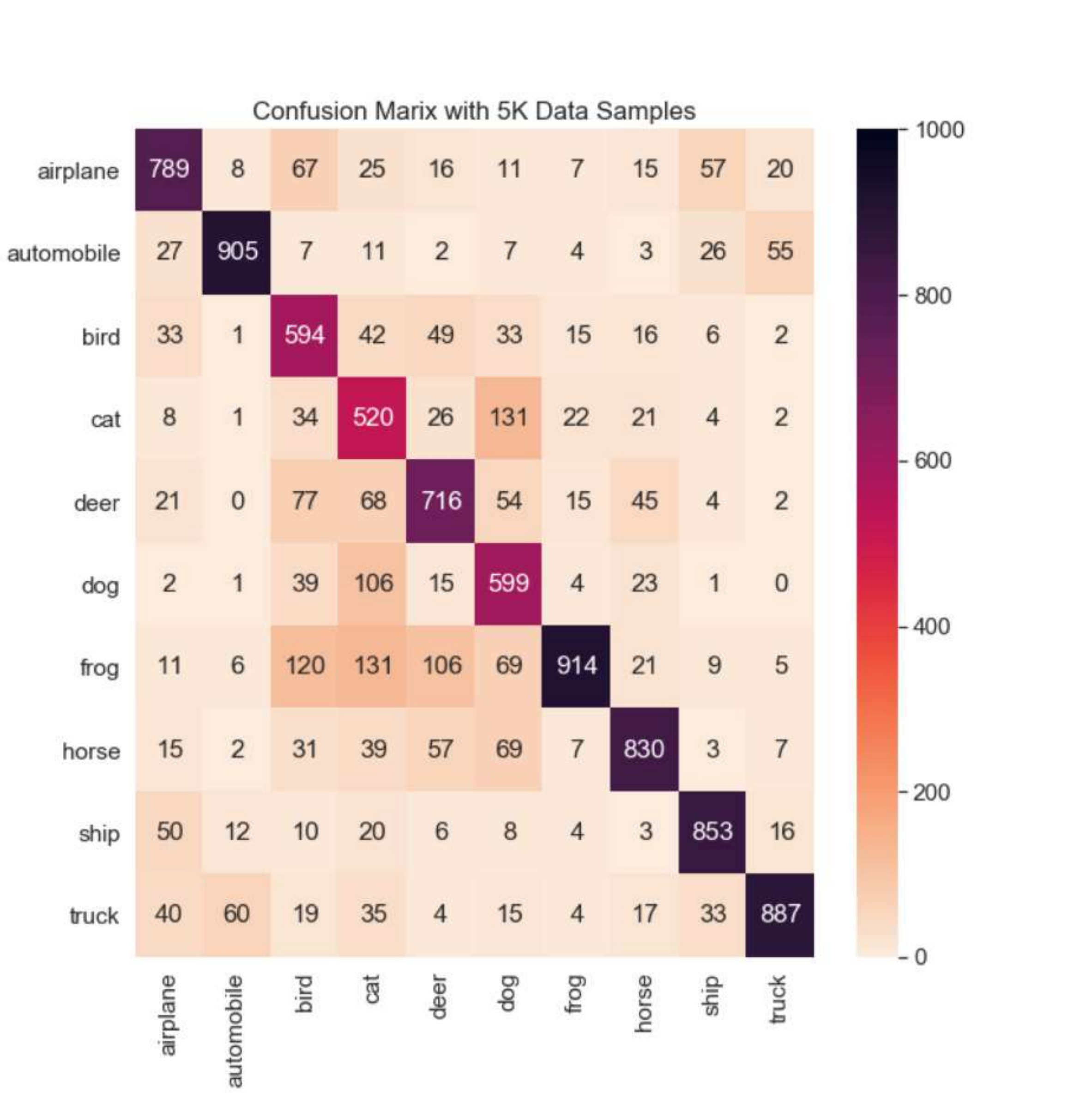
Average Confusion Matrix with size 40k samples

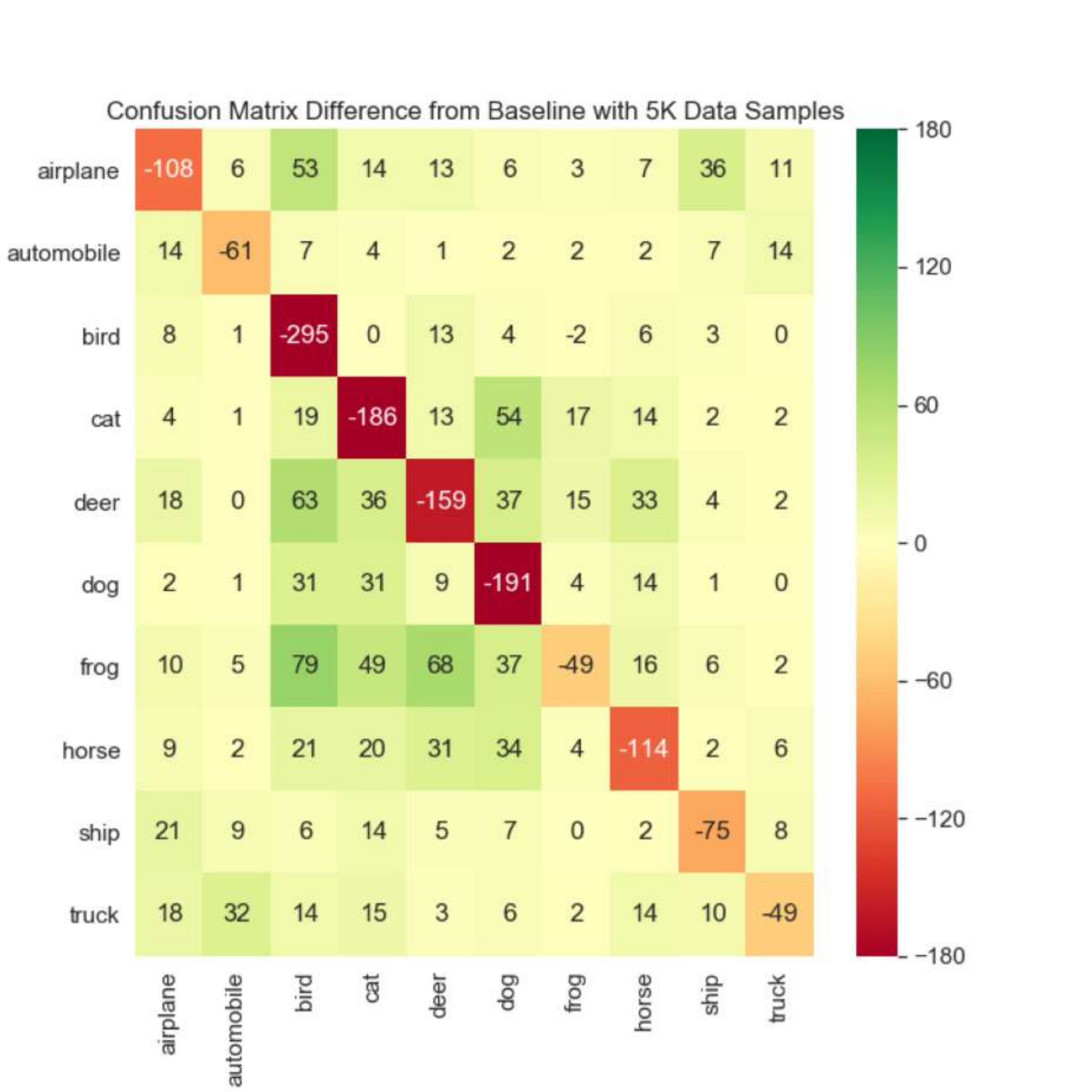
EXPERIMENT #2: DATA VOLUME REDUCTION



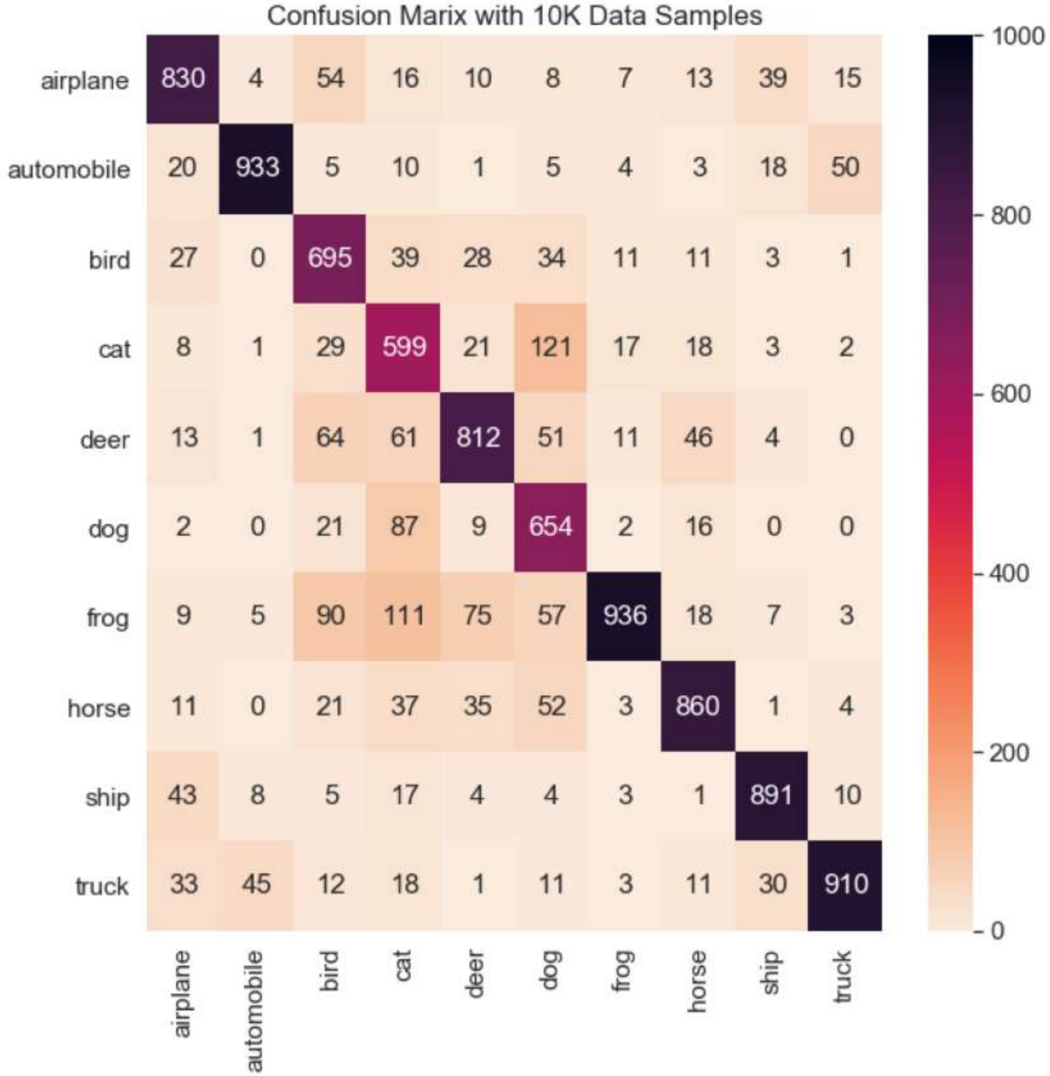


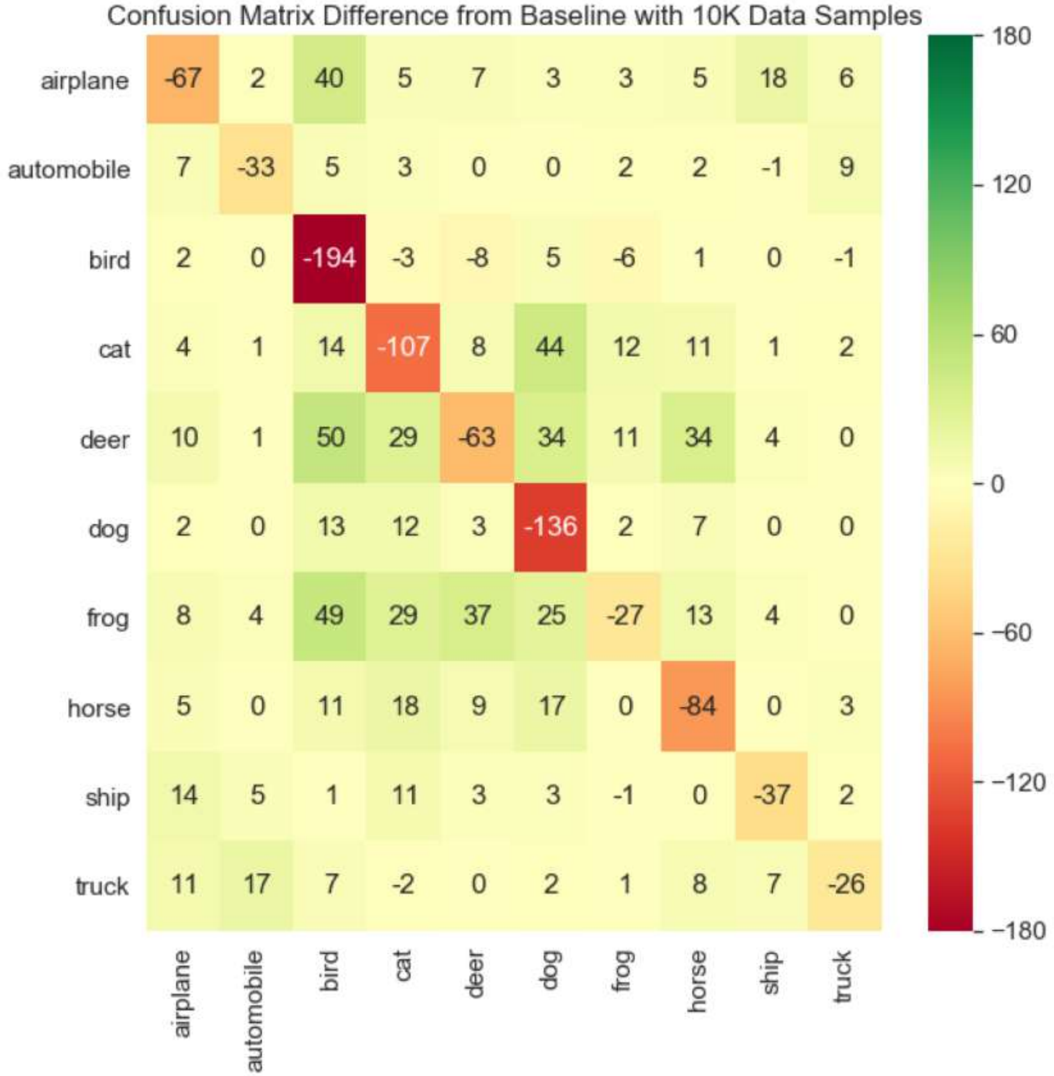
EXPERIMENT #2: DATA VOLUME REDUCTION



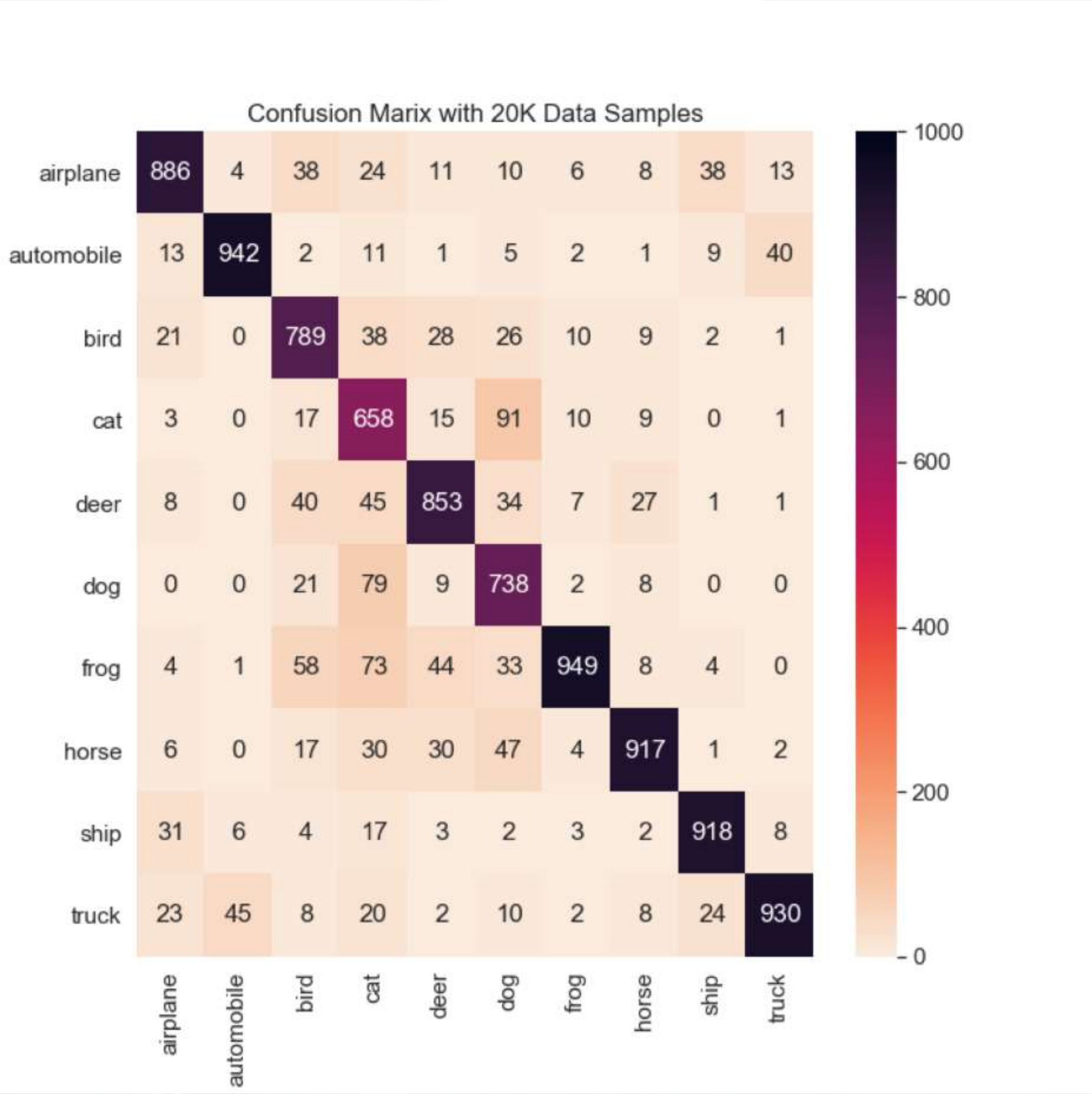


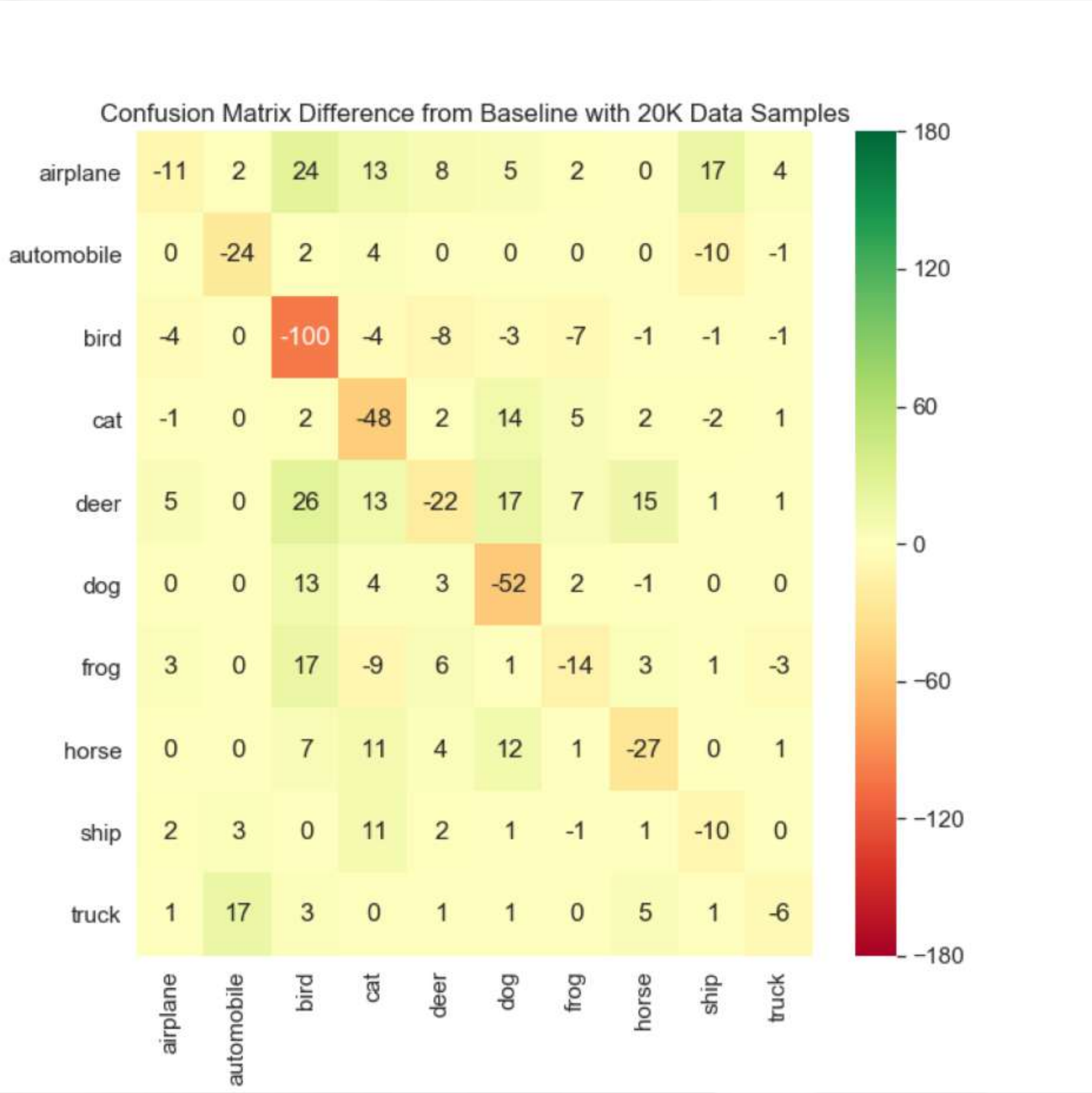
EXPERIMENT #2: DATA VOLUME REDUCTION



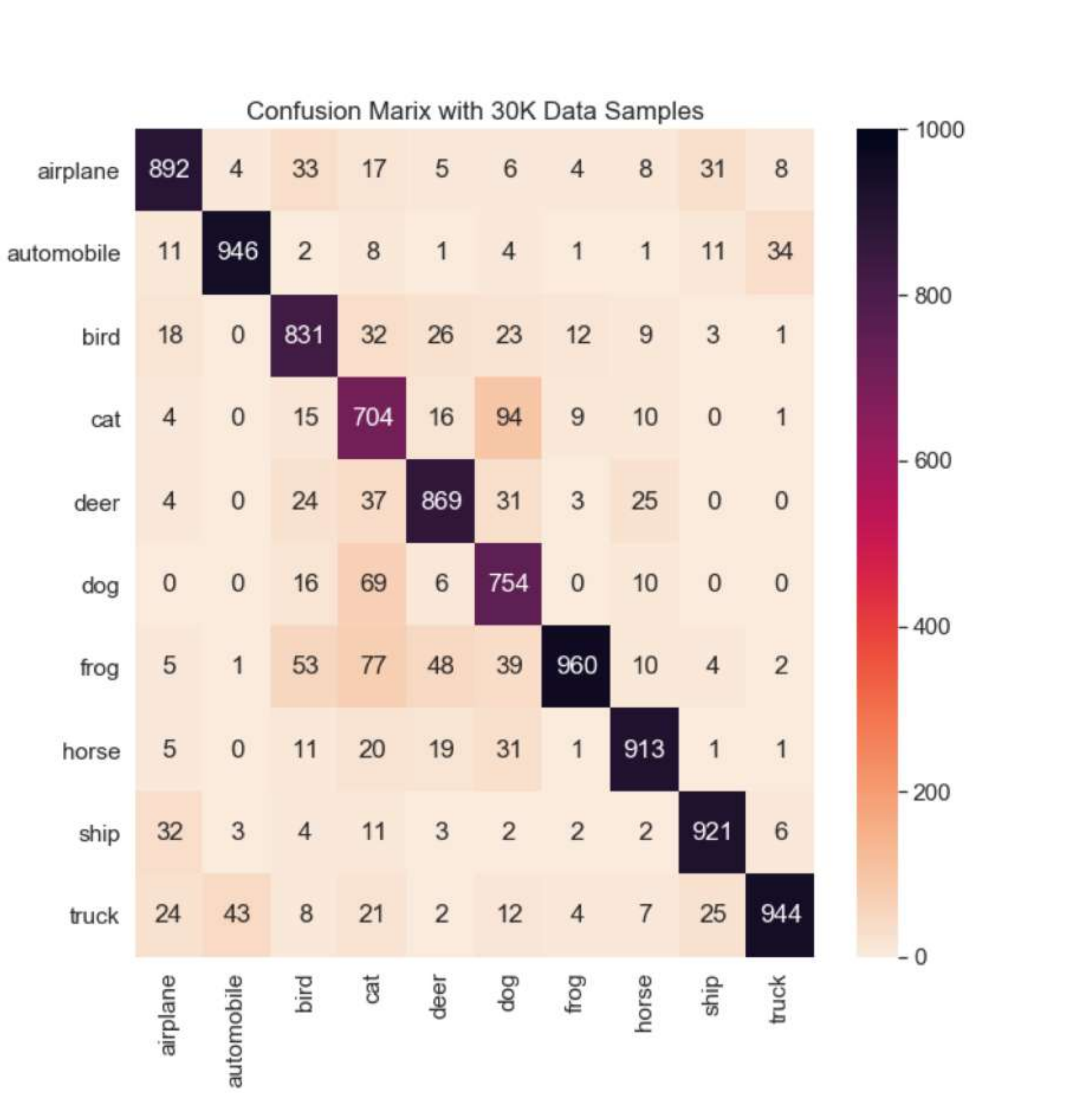


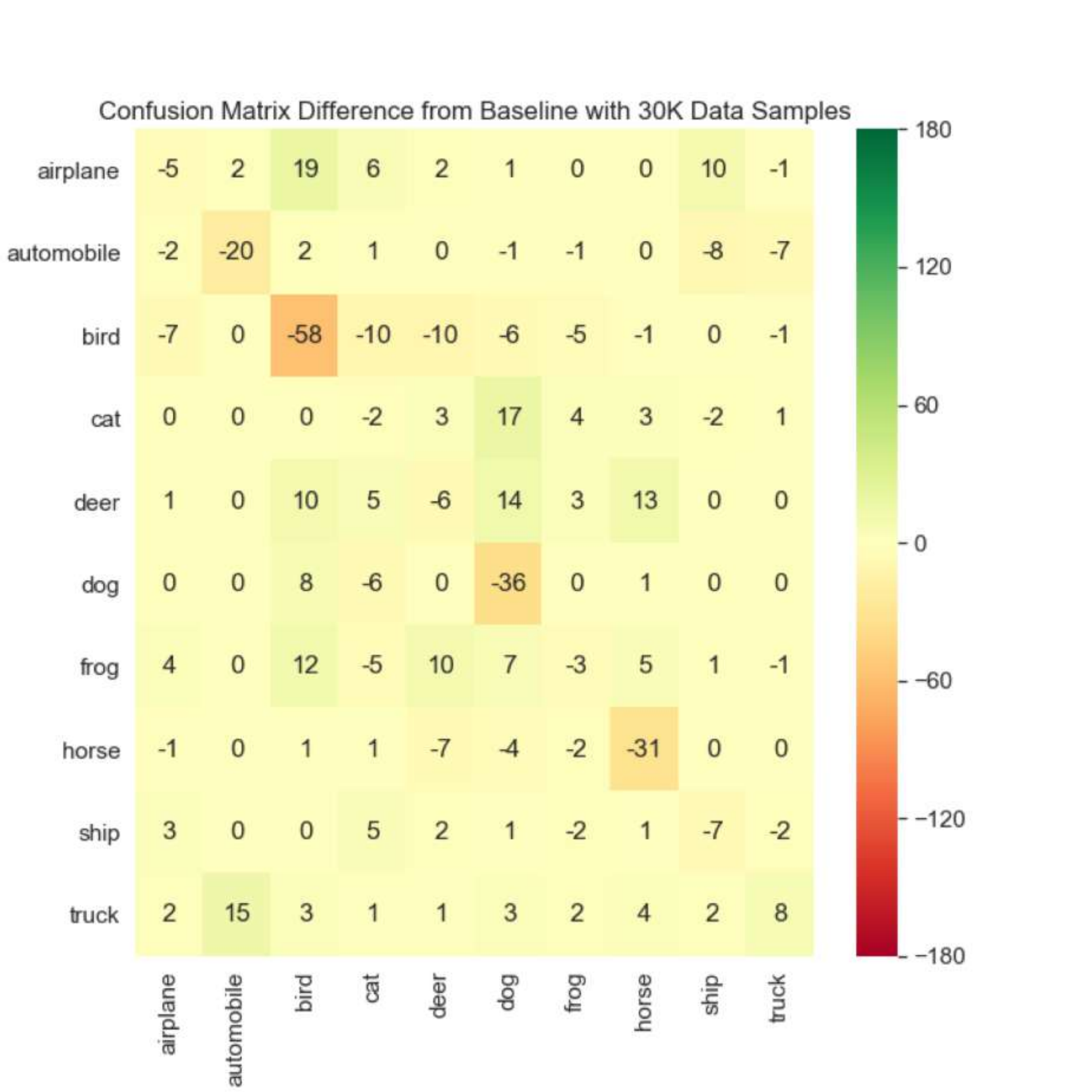
EXPERIMENT #2: DATA VOLUME REDUCTION



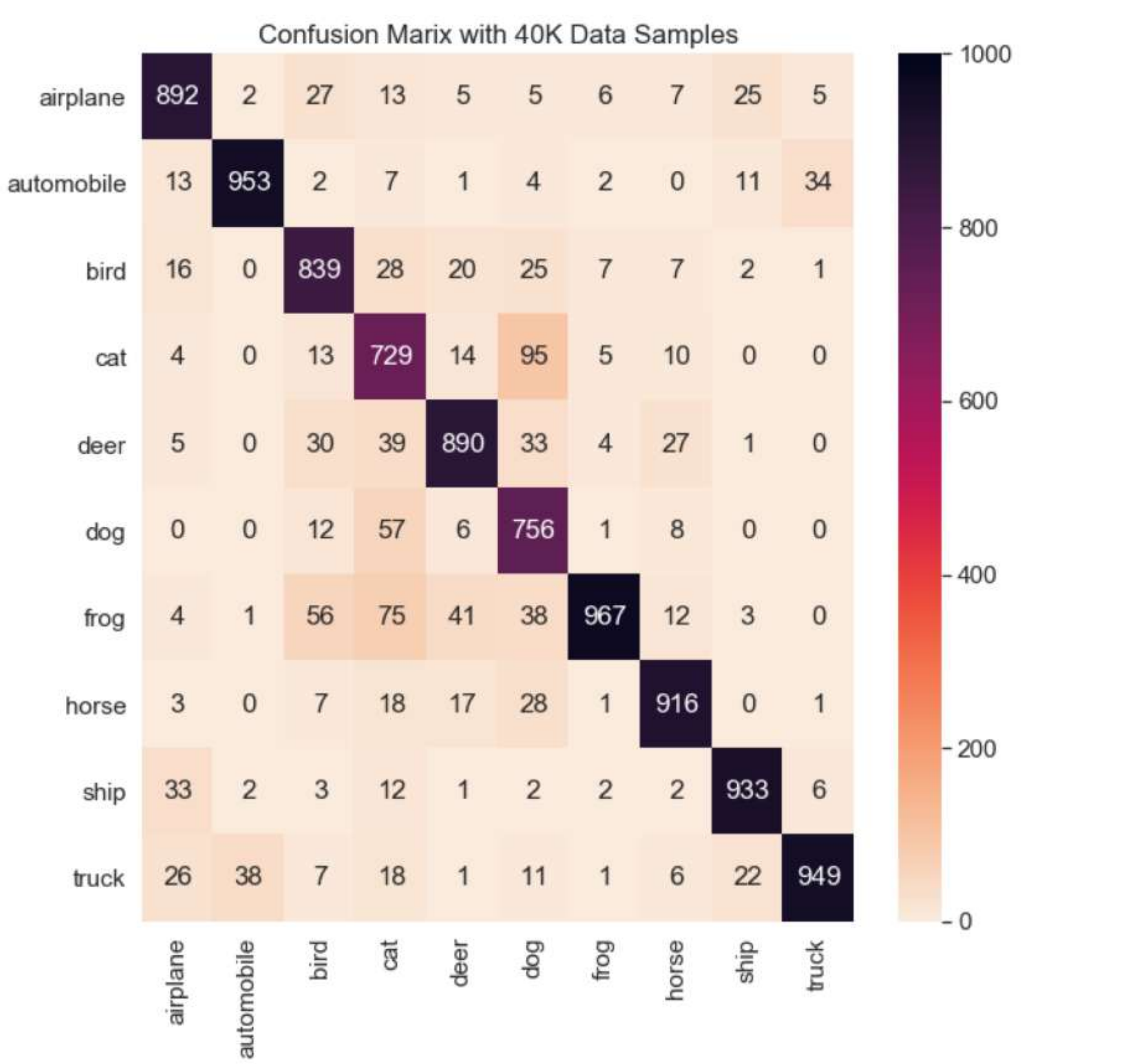


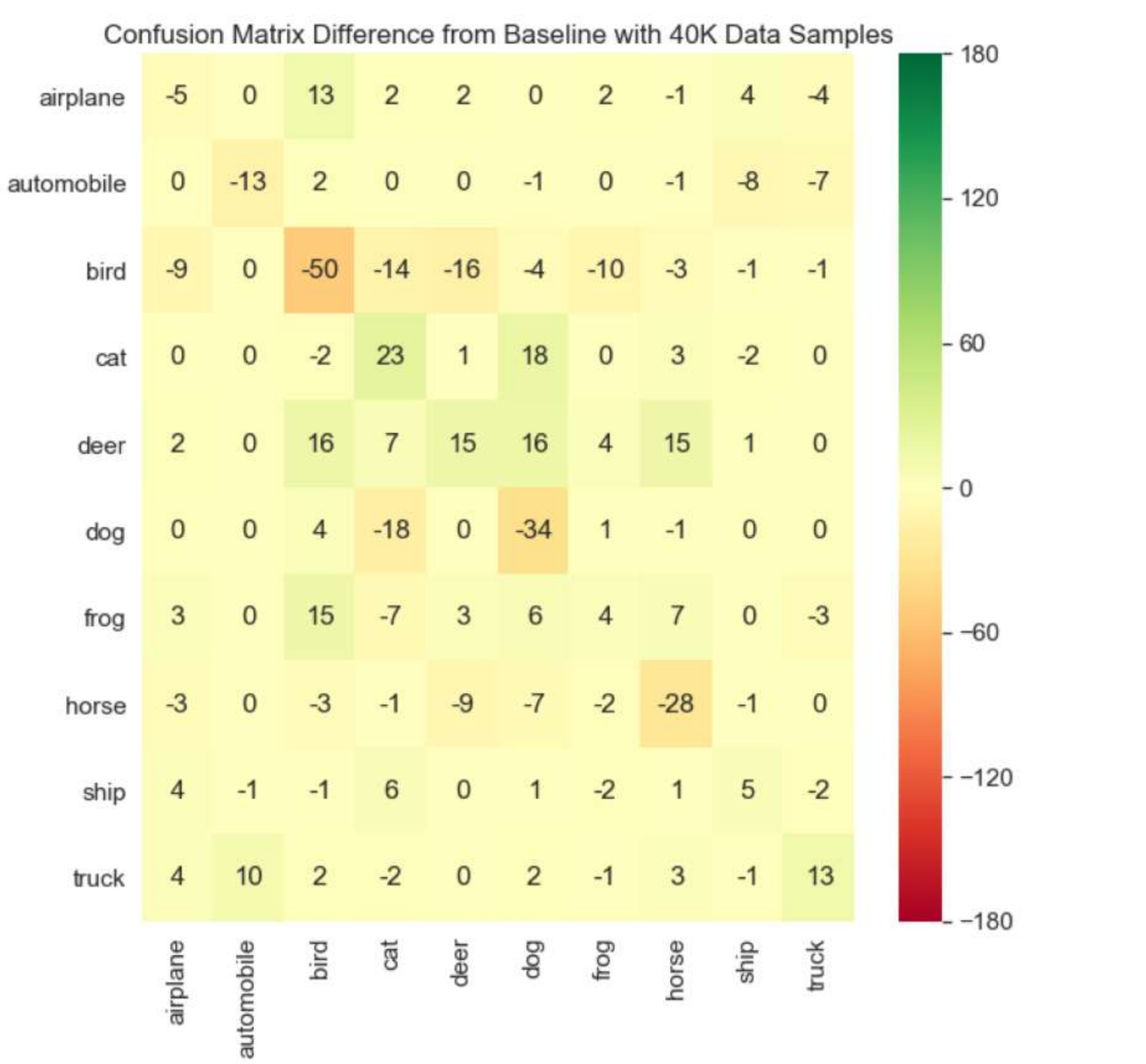
EXPERIMENT #2: DATA VOLUME REDUCTION



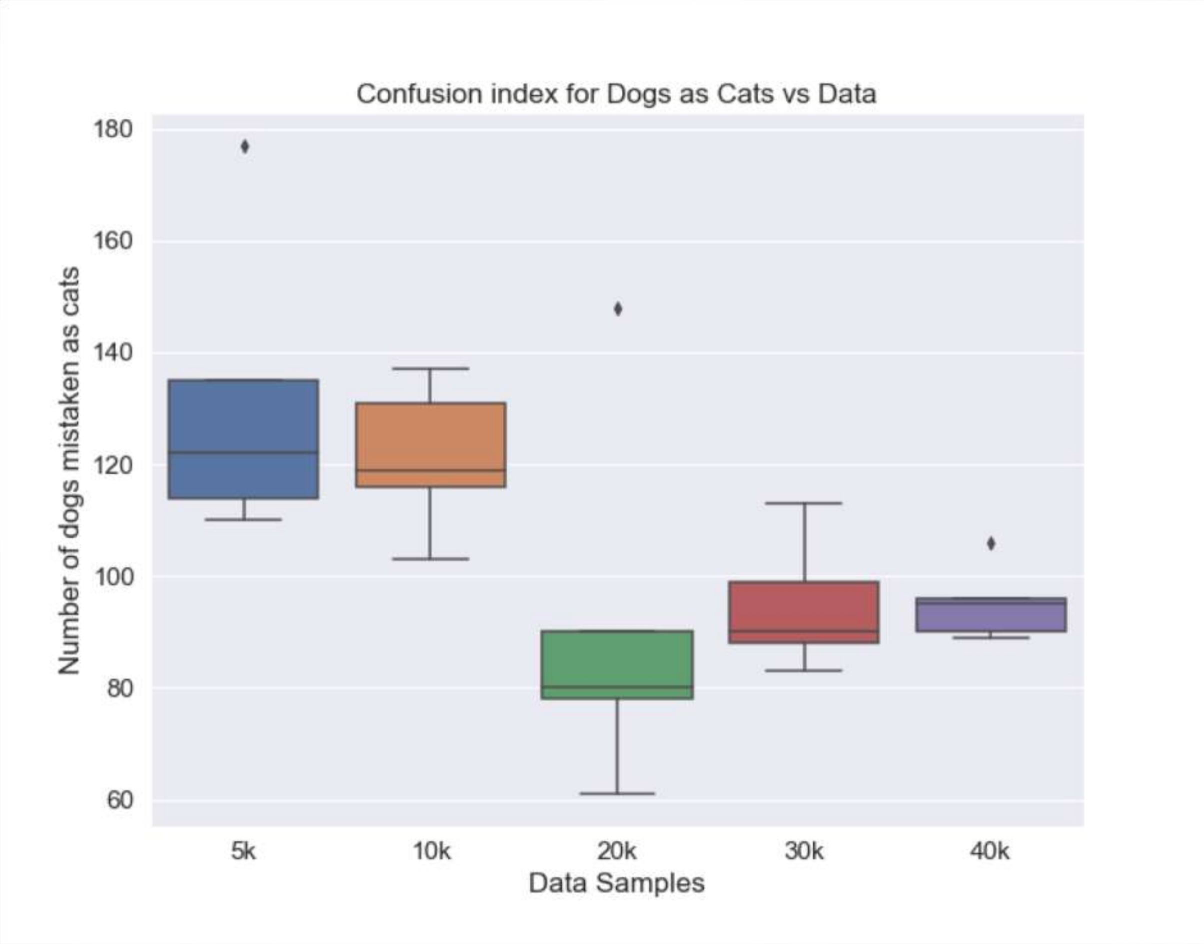


EXPERIMENT #2: DATA VOLUME REDUCTION

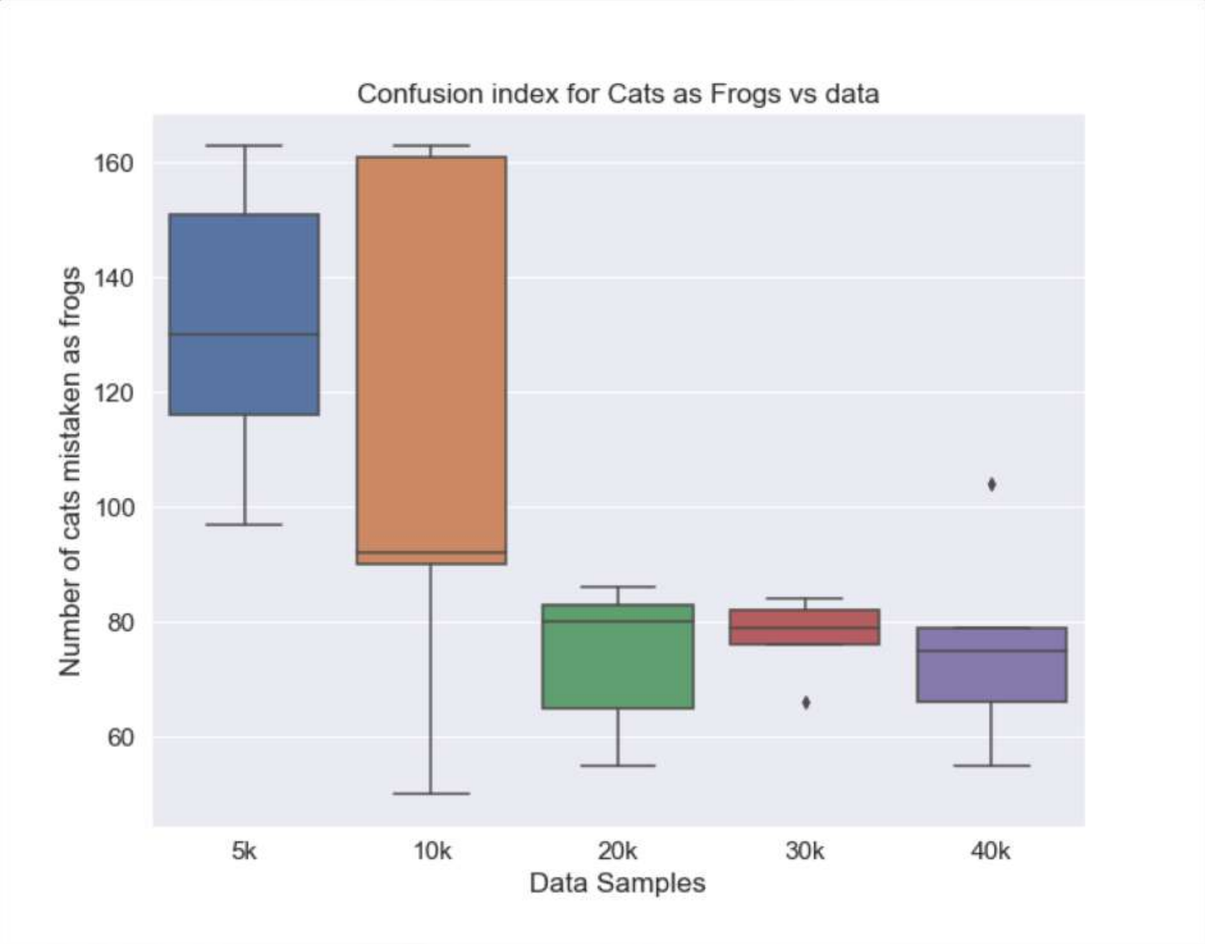




EXPERIMENT #2: DATA VOLUME REDUCTION



Confusion {dog → cat}
vs. volume of training data

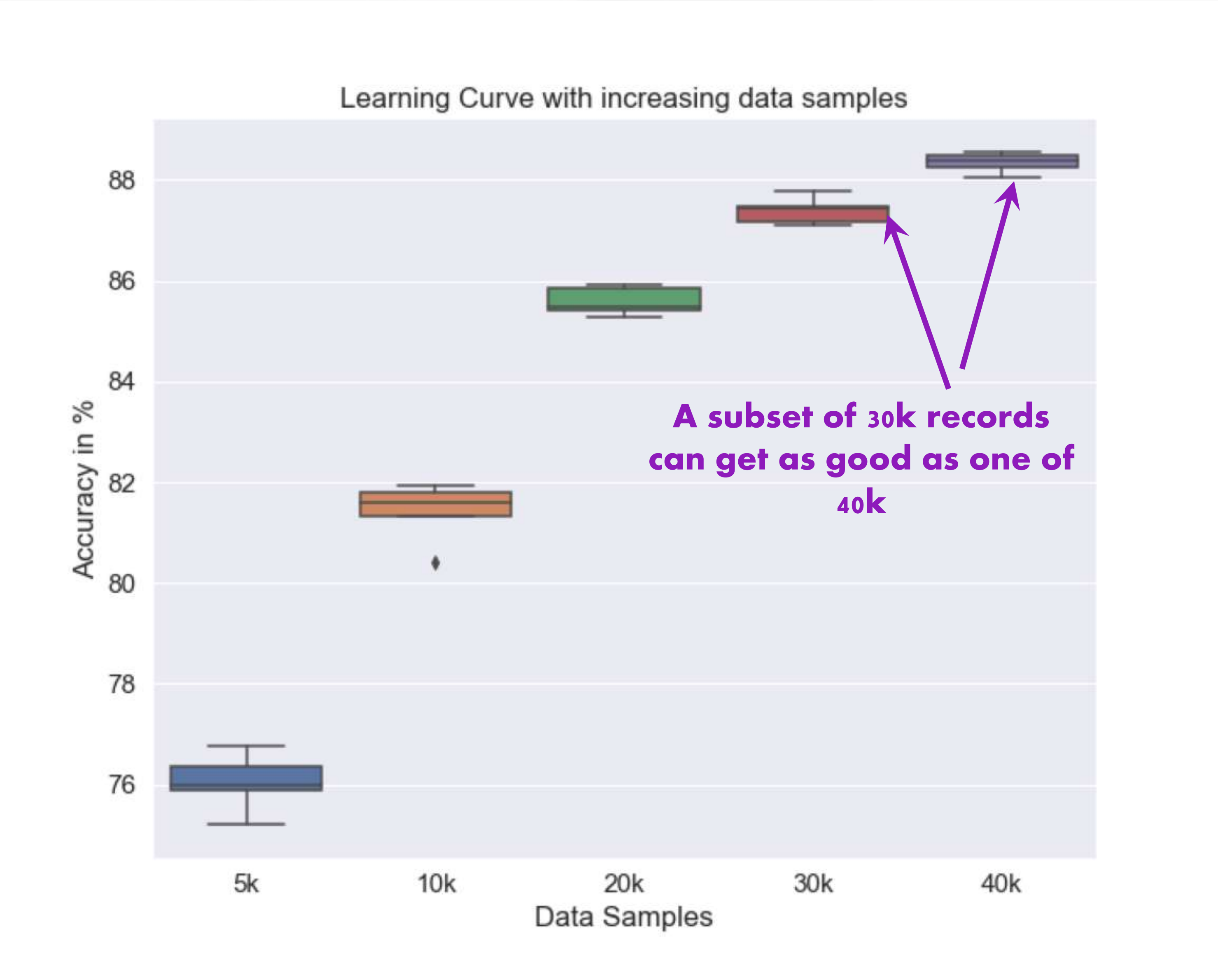


Confusion {cat → frog}
vs. volume of training data

EXPERIMENT #2: DATA VOLUME REDUCTION

Results

- Accuracy gets asymptotically better with more data
- 10k gets us more than 90% of the way there
- 20k (less than half) gets us 95+% of the way there
- The best sample of size 30k gets similar accuracy to the worse one with size 40k



TIME TO DRAW SOME (REAL) CONCLUSIONS

DISCUSSION: ARE ALL CLASSES EQUALLY IMPACTED?

A FEW CONCLUSIONS...

Lowest accuracy

Highest accuracy

Baseline		30% Labeling Noise		5K Data Samples	
Cat	706	Cat	608	Cat	520
Dog	790	Dog	698	Bird	594
Deer	875	Bird	744	Dog	599
Bird	889	Deer	811	Deer	716
Airplane	897	Airplane	837	Airplane	789
Ship	928	Horse	877	Horse	830
Truck	936	Ship	897	Ship	853
Horse	944	Automobile	922	Truck	887
Frog	963	Truck	928	Automobile	905
Automobile	966	Frog	951	Frog	914

DISCUSSION: ARE ALL CLASSES EQUALLY IMPACTED?

A FEW CONCLUSIONS…

- **'Cat'** is the least accurate class even with labeling noise and data quantity

Lowest accuracy → Highest accuracy

Baseline		30% Labeling Noise		5K Data Samples	
Cat	706	Cat	608	Cat	520
Dog	790	Dog	698	Bird	594
Deer	875	Bird	744	Dog	599
Bird	889	Deer	811	Deer	716
Airplane	897	Airplane	837	Airplane	789
Ship	928	Horse	877	Horse	830
Truck	936	Ship	897	Ship	853
Horse	944	Automobile	922	Truck	887
Frog	963	Truck	928	Automobile	905
Automobile	966	Frog	951	Frog	914

DISCUSSION: ARE ALL CLASSES EQUALLY IMPACTED?

A FEW CONCLUSIONS…

- **'Cat'** is the least accurate class even with labeling noise and data quantity
- **'Bird'** class relative performance decreases with labeling noise and volume reduction

Lowest accuracy → Highest accuracy

Baseline		30% Labeling Noise		5K Data Samples	
Cat	706	Cat	608	Cat	520
Dog	790	Dog	698	Bird	594
Deer	875	Bird	744	Dog	599
Bird	889	Deer	811	Deer	716
Airplane	897	Airplane	837	Airplane	789
Ship	928	Horse	877	Horse	830
Truck	936	Ship	897	Ship	853
Horse	944	Automobile	922	Truck	887
Frog	963	Truck	928	Automobile	905
Automobile	966	Frog	951	Frog	914

DISCUSSION: ARE ALL CLASSES EQUALLY IMPACTED?

A FEW CONCLUSIONS...

- **'Cat'** is the least accurate class even with labeling noise and data quantity
- **'Bird'** class relative performance decreases with labeling noise and volume reduction
- **'Frog'** class stays stable with noise induction as well as data volume reduction

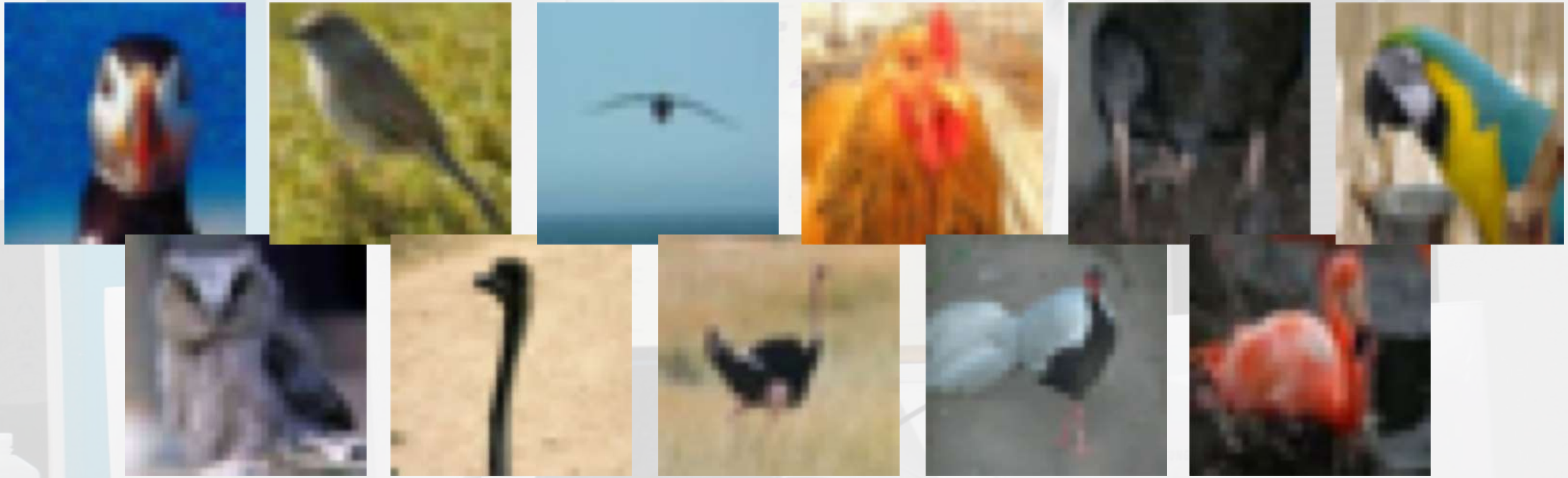
Lowest accuracy

Highest accuracy

Baseline		30% Labeling Noise		5K Data Samples	
Cat	706	Cat	608	Cat	520
Dog	790	Dog	698	Bird	594
Deer	875	Bird	744	Dog	599
Bird	889	Deer	811	Deer	716
Airplane	897	Airplane	837	Airplane	789
Ship	928	Horse	877	Horse	830
Truck	936	Ship	897	Ship	853
Horse	944	Automobile	922	Truck	887
Frog	963	Truck	928	Automobile	905
Automobile	966	Frog	951	Frog	914

DISCUSSION: ARE ALL CLASSES EQUALLY IMPACTED?

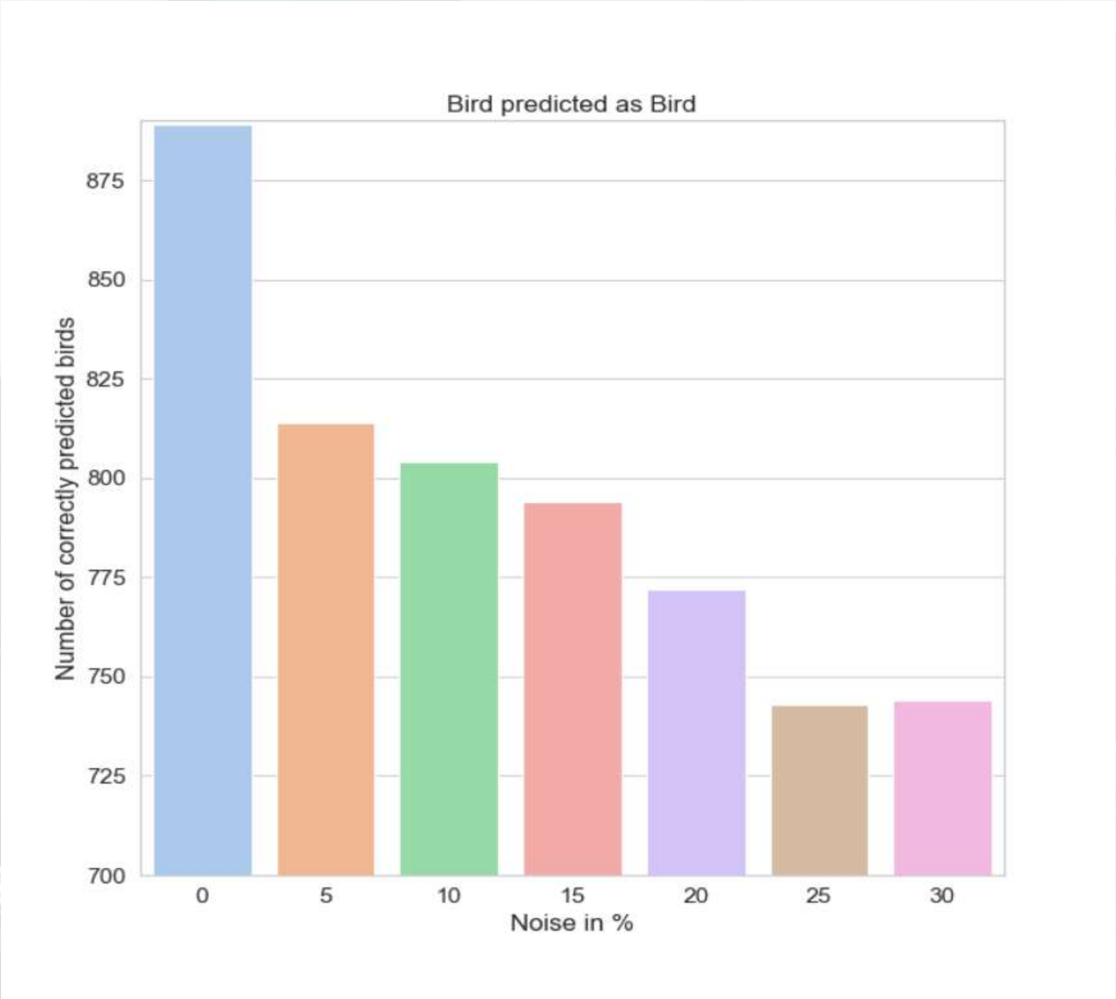
INTUITION: ‘BIRD’ CLASS VARIANCE



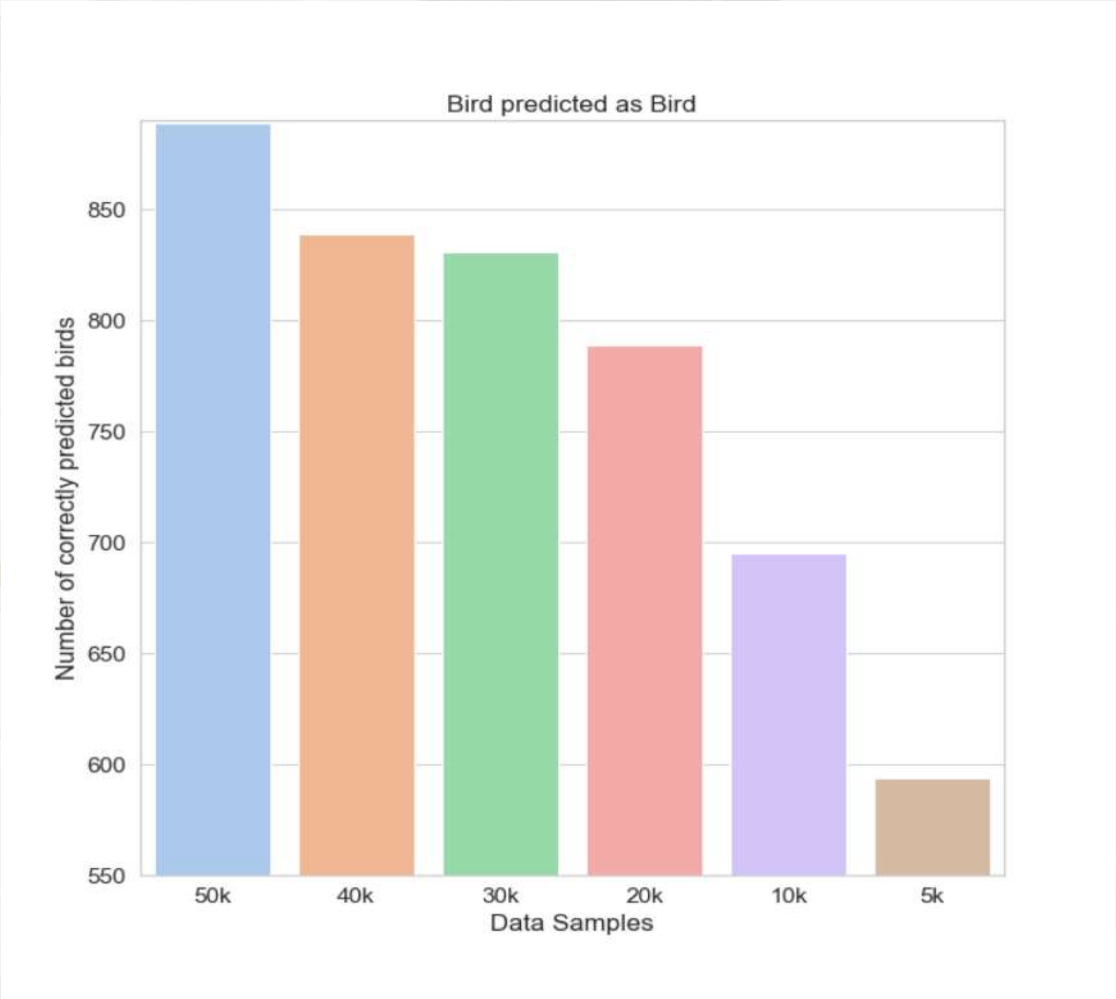
DISCUSSION: ARE ALL CLASSES EQUALLY IMPACTED?

Most Sensitive Class – 'Bird'

Results with labeling pollution



Results with data volume reduction

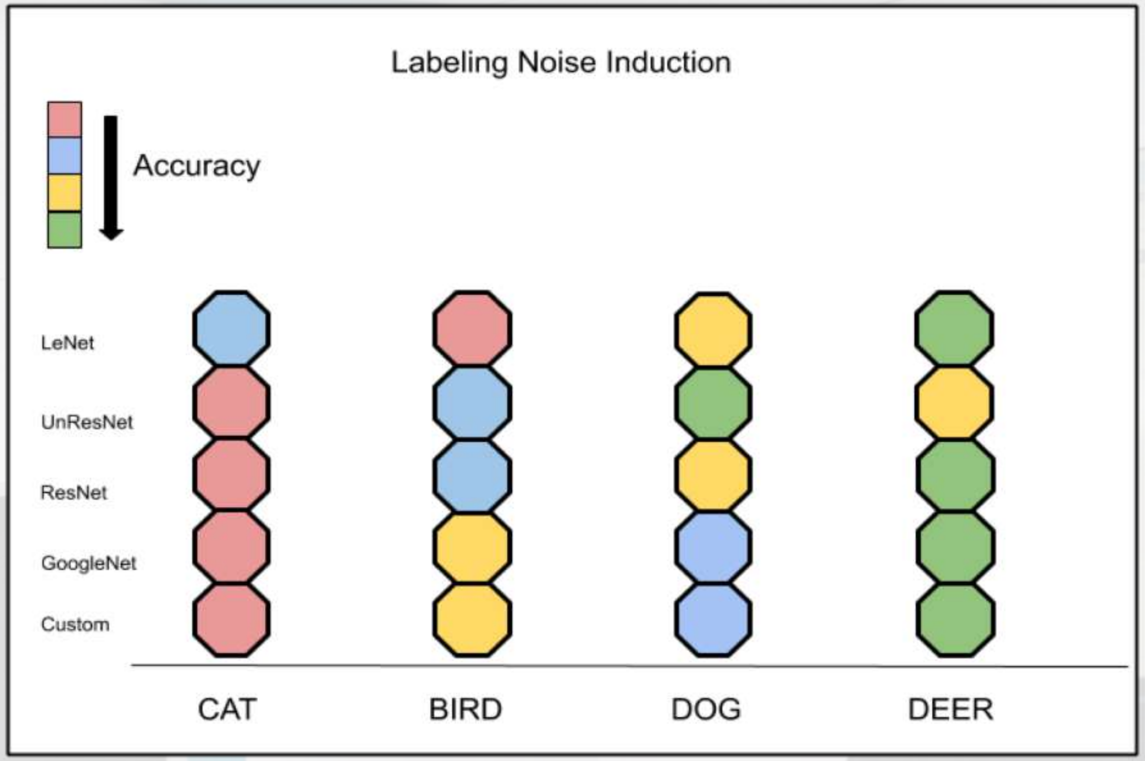


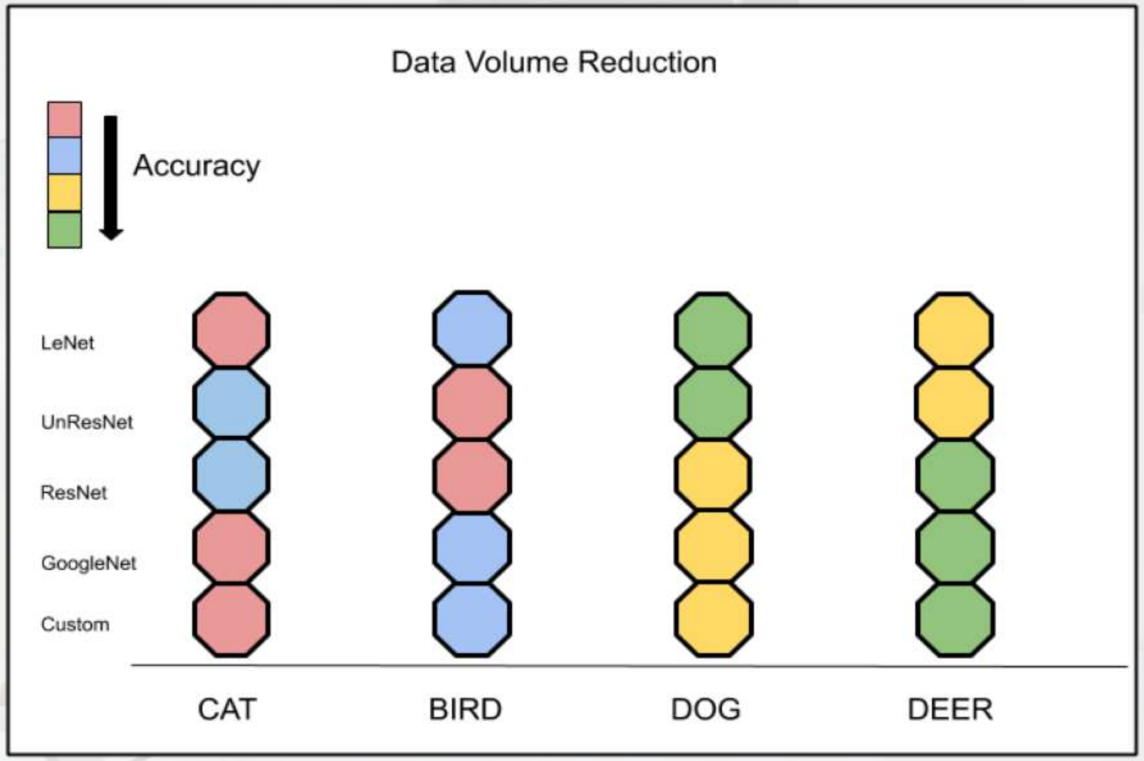
DISCUSSION: IS IT THE MODEL OR THE DATA?

Model	Epochs	Batch Size	Accuracy
Custom (Keras with TF backend)	125	64	88.94
LeNet (Pytorch)	125	64	66.6
ResNet18 (Pytorch)	25	64	88.29
UnResNet18 (Pytorch)	25	64	85.77
GoogLeNet (Pytorch)	25	64	88.6

Same Experiments, Different Models

DISCUSSION: IS IT THE MODEL OR THE DATA?

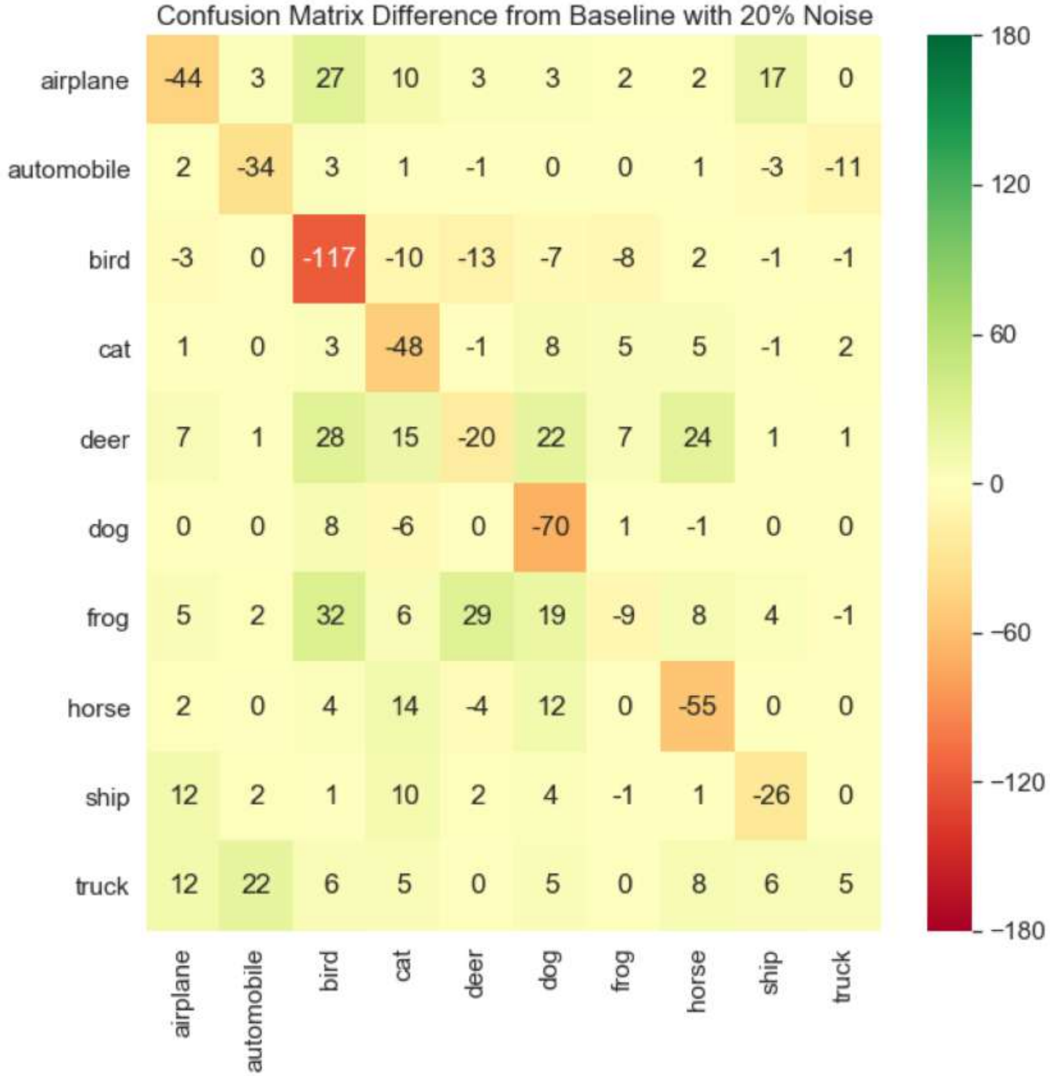




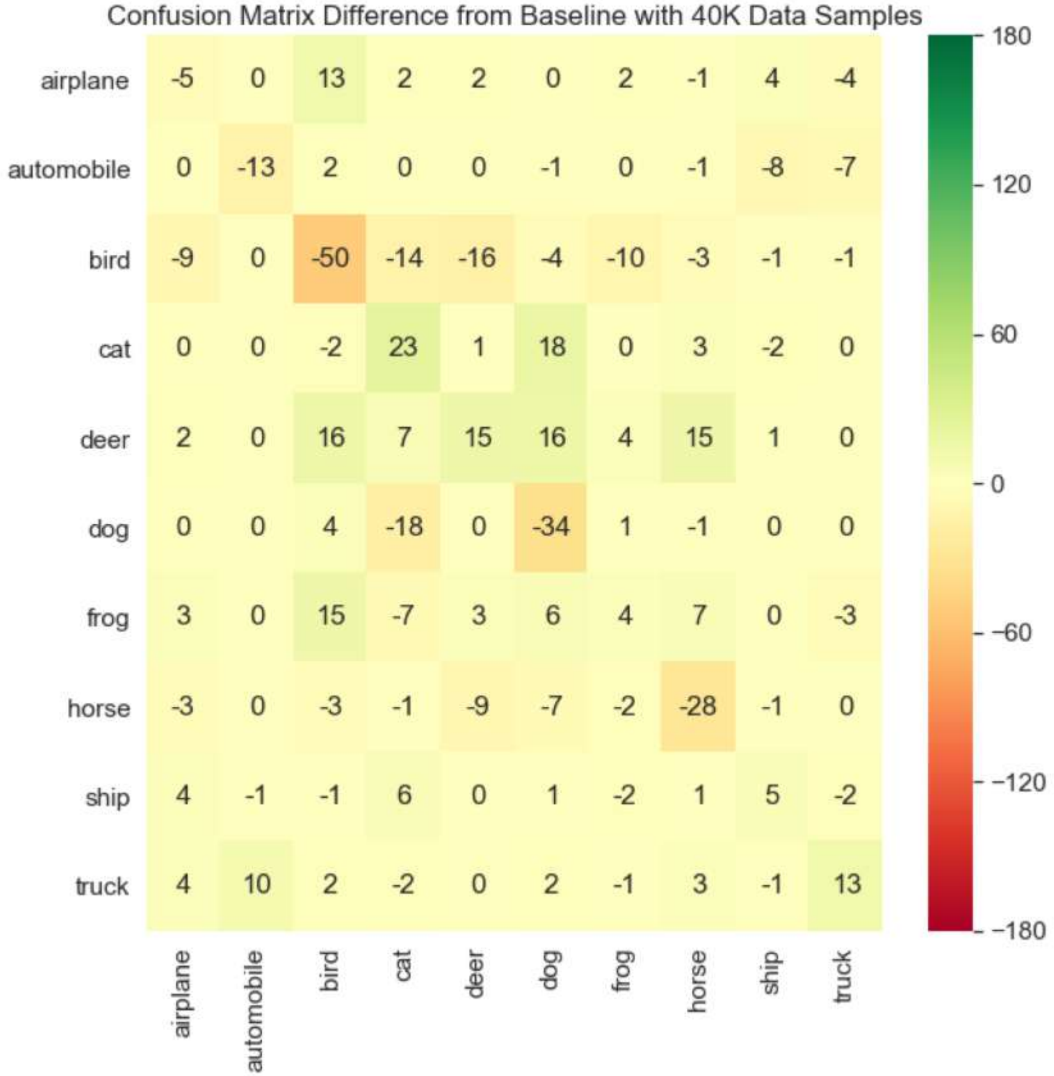
Same Experiments, Different Models

DISCUSSION: VOLUME REDUCTION VS. LABELING NOISE

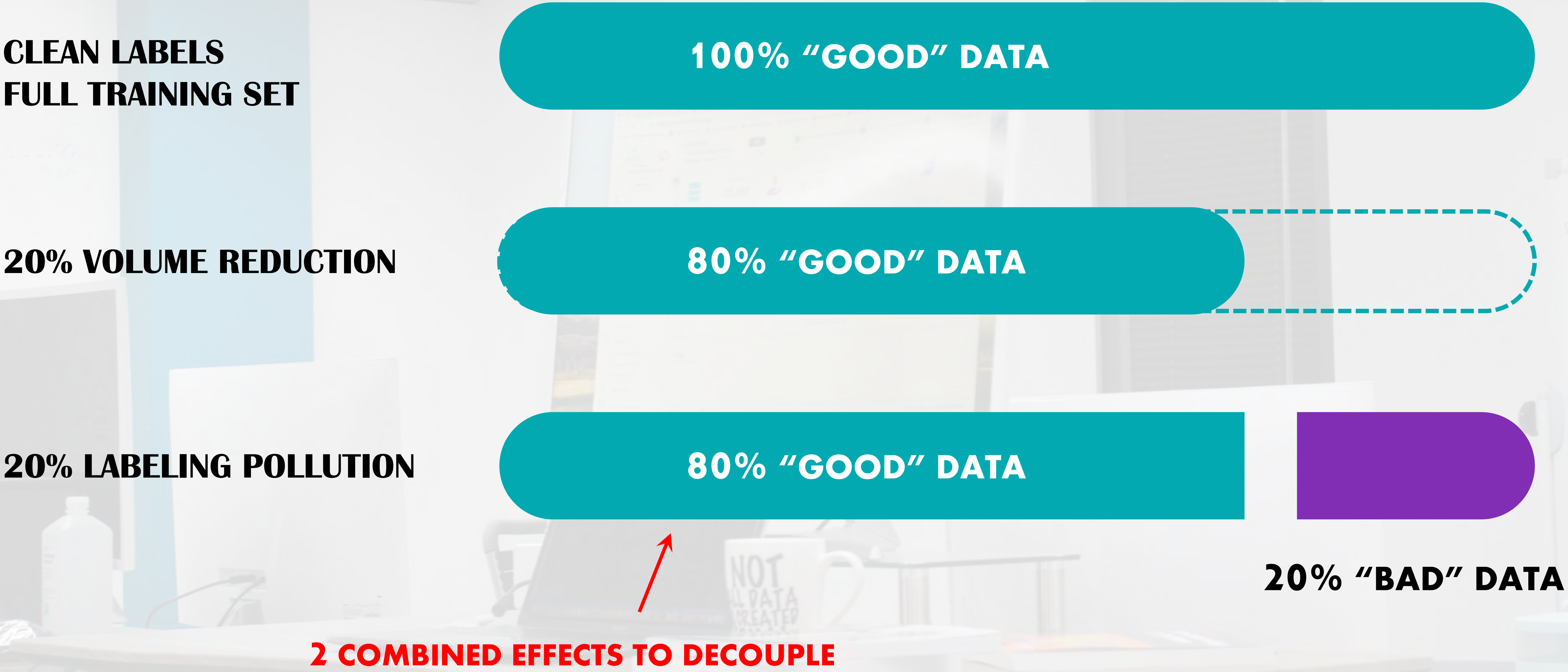
20% Labeling Noise Induction



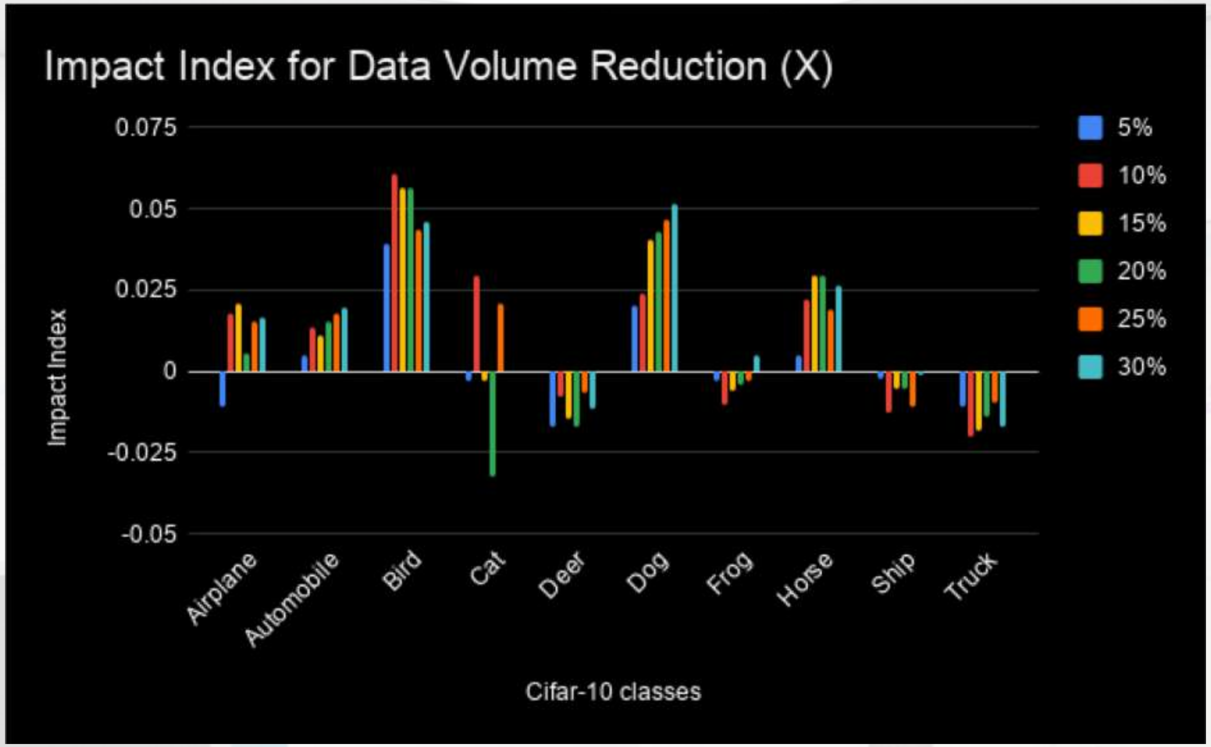
20% Data Volume Reduction

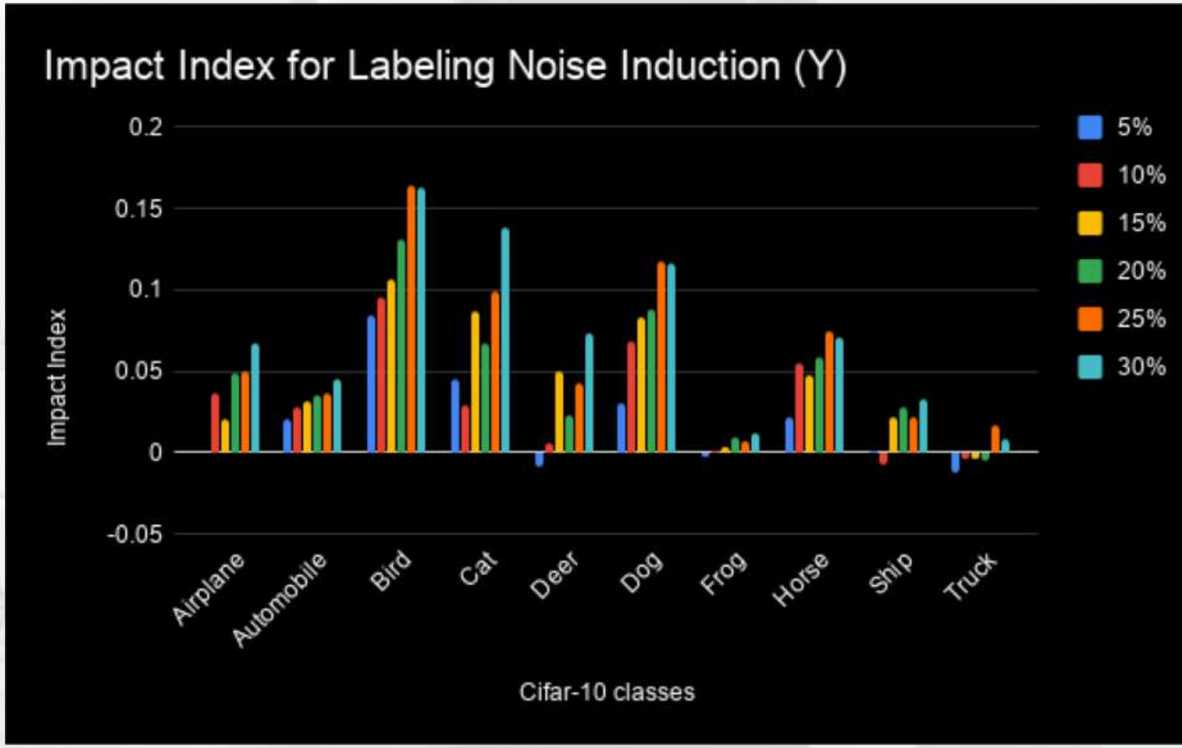


DISCUSSION: VOLUME REDUCTION VS. LABELING NOISE



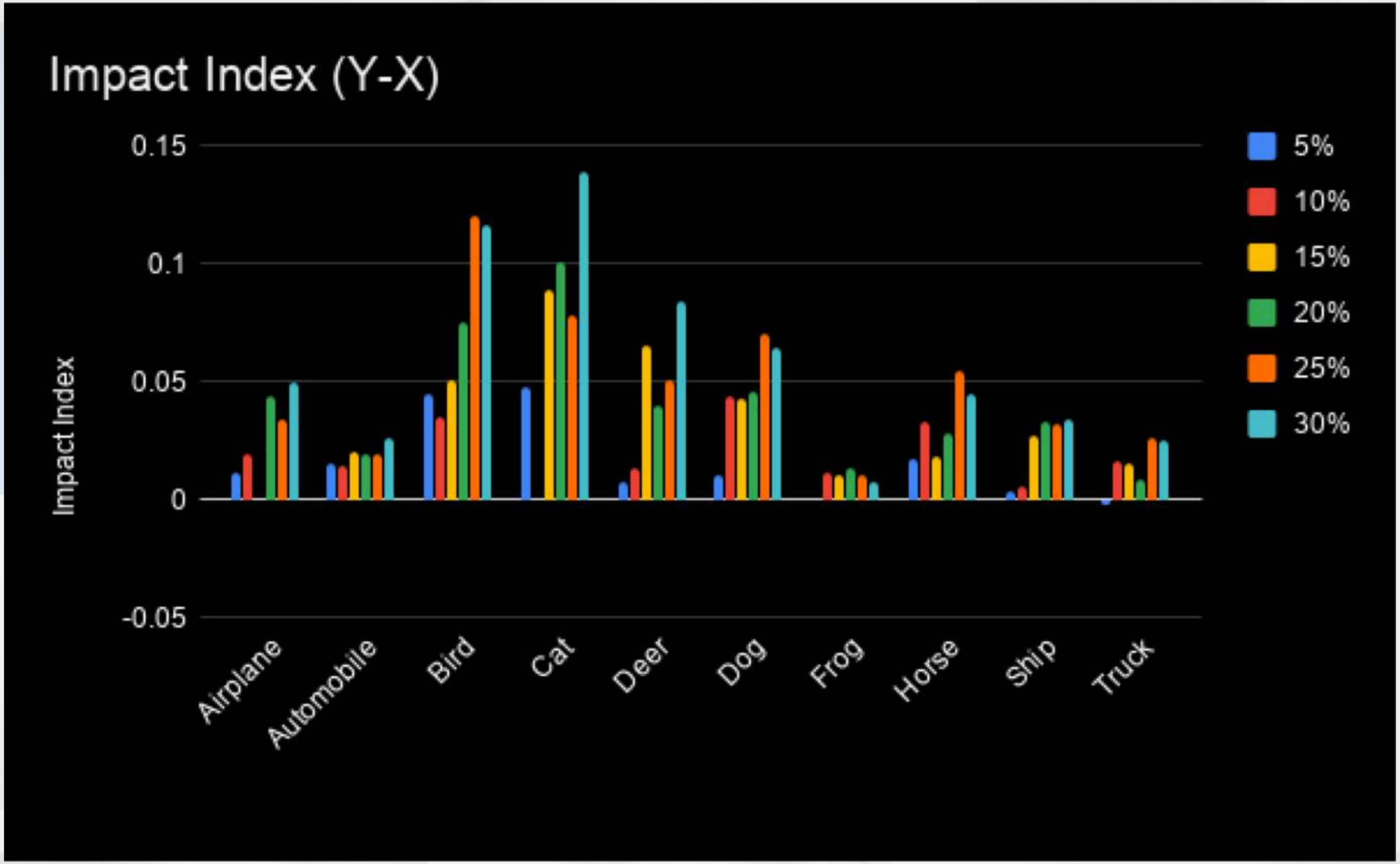
DISCUSSION: VOLUME REDUCTION VS. LABELING NOISE





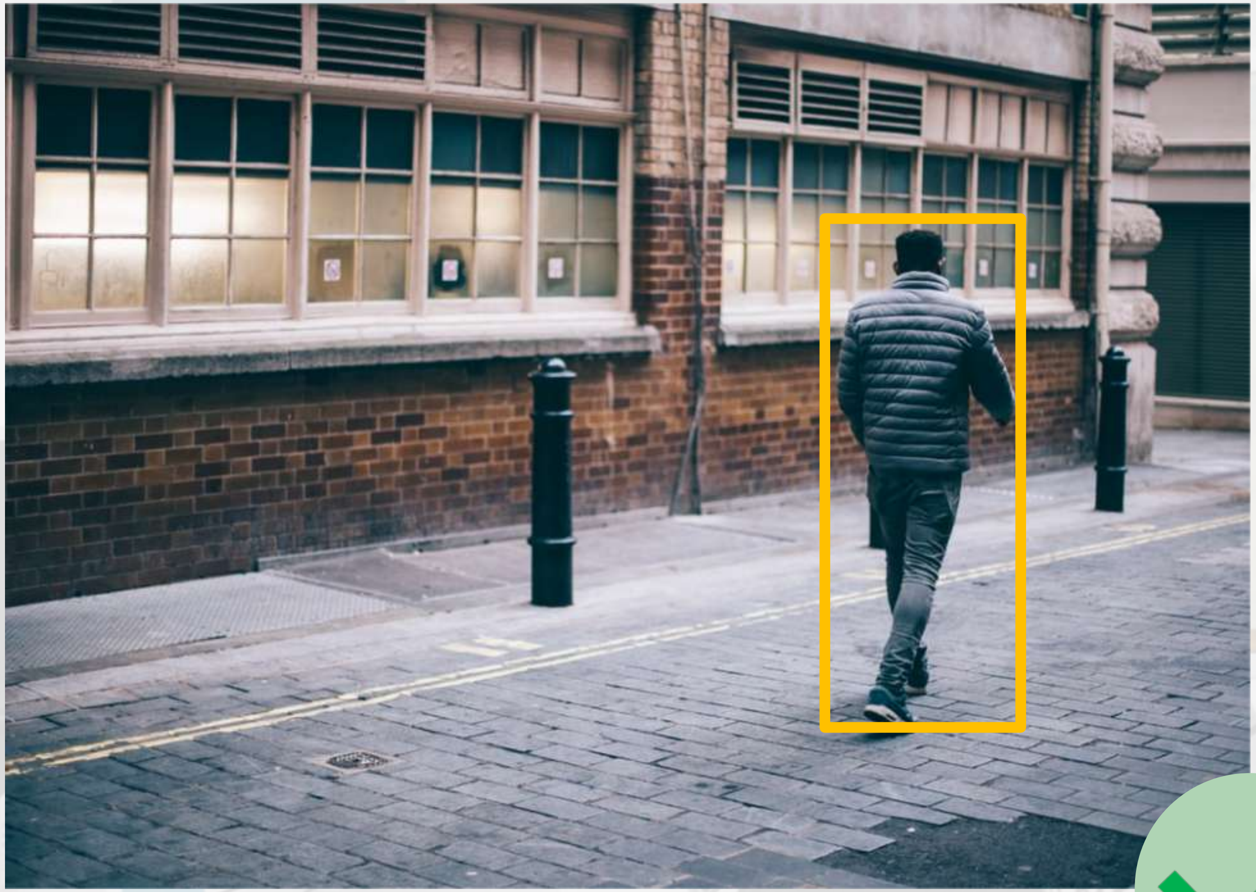
DISCUSSION: VOLUME REDUCTION VS. LABELING NOISE

Harder to compensate for bad quality with higher volume

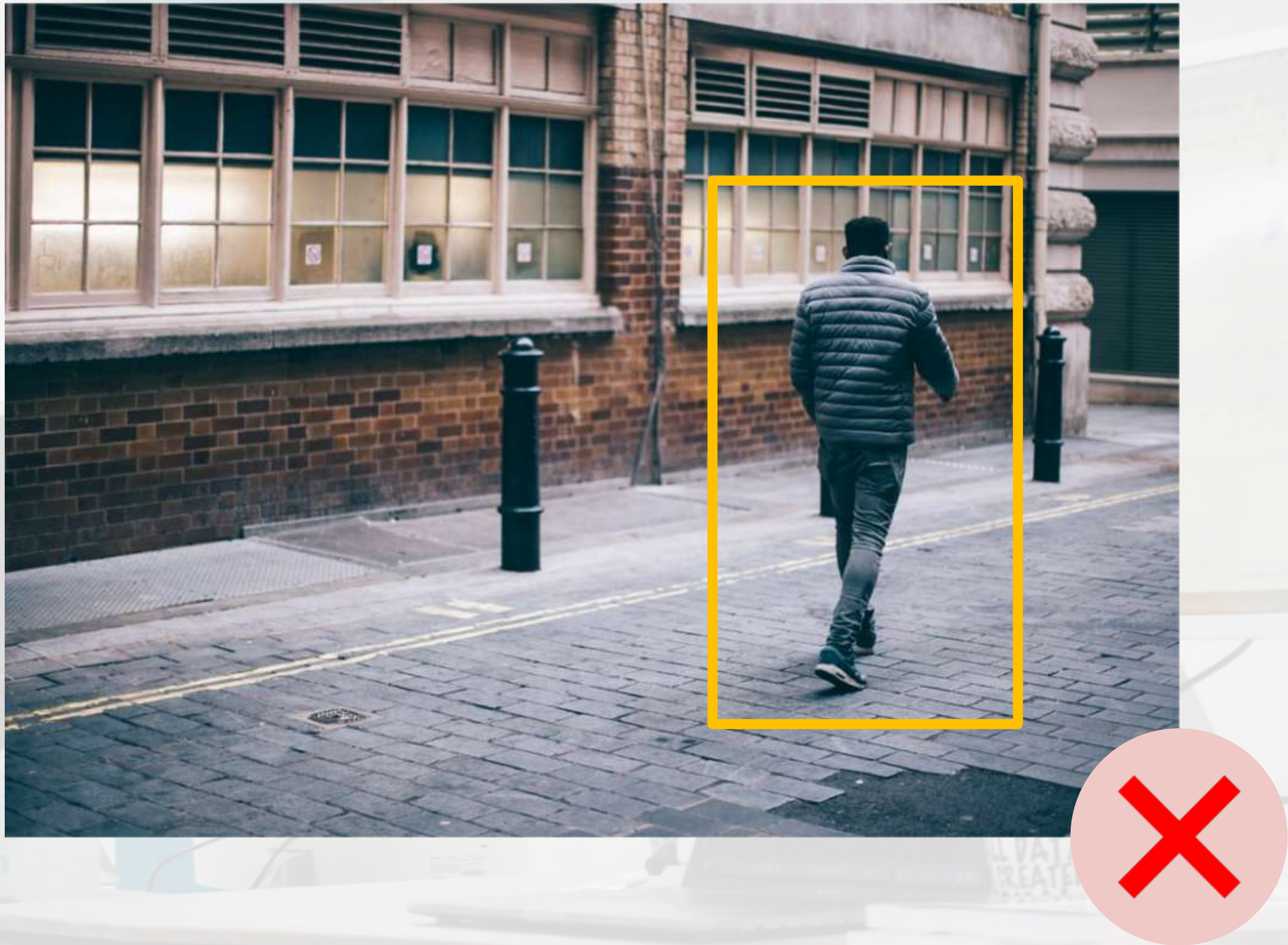


LET'S SAVE SOME MONEY!

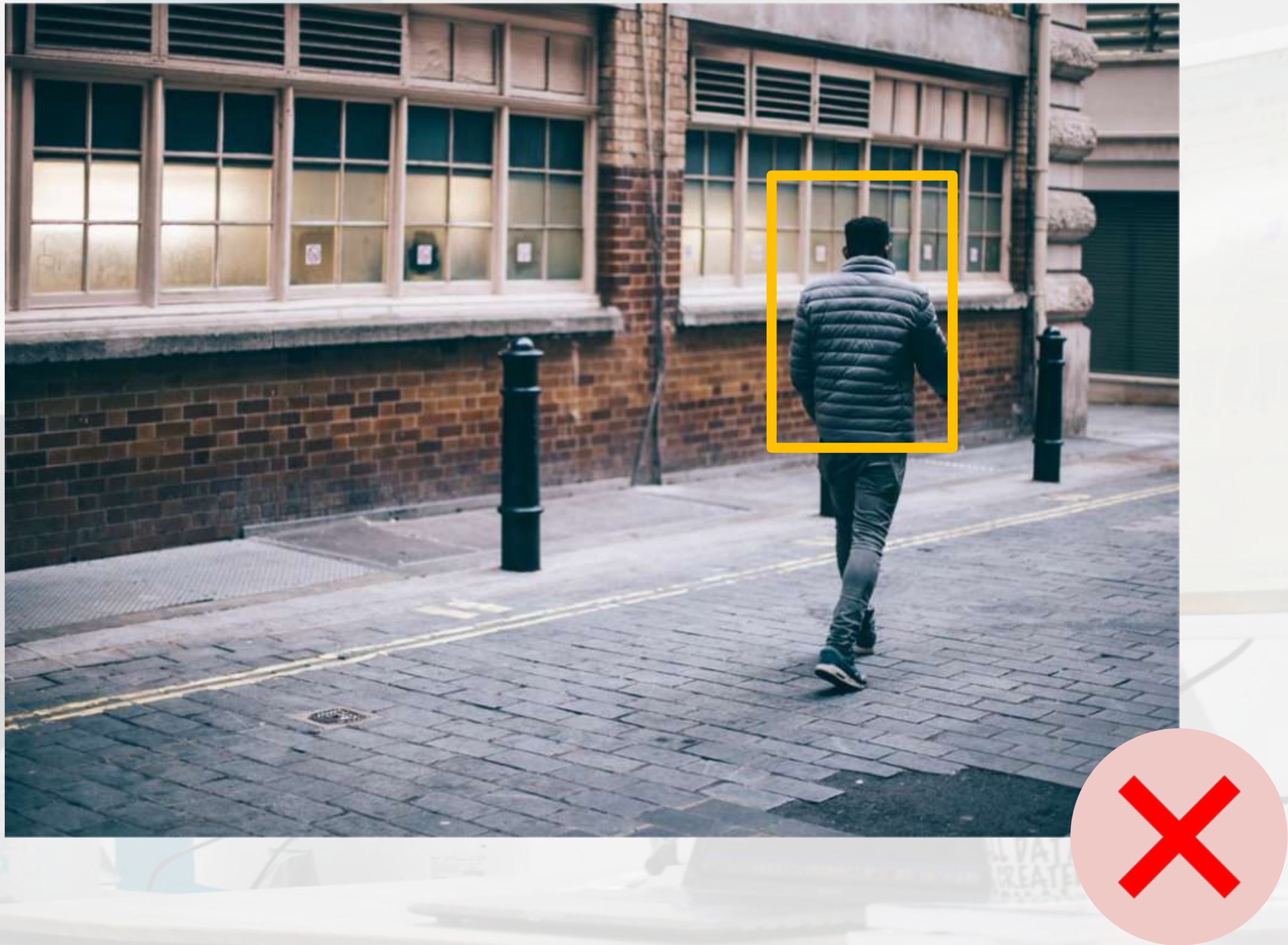
TOWARDS A SMART LABELING STRATEGY



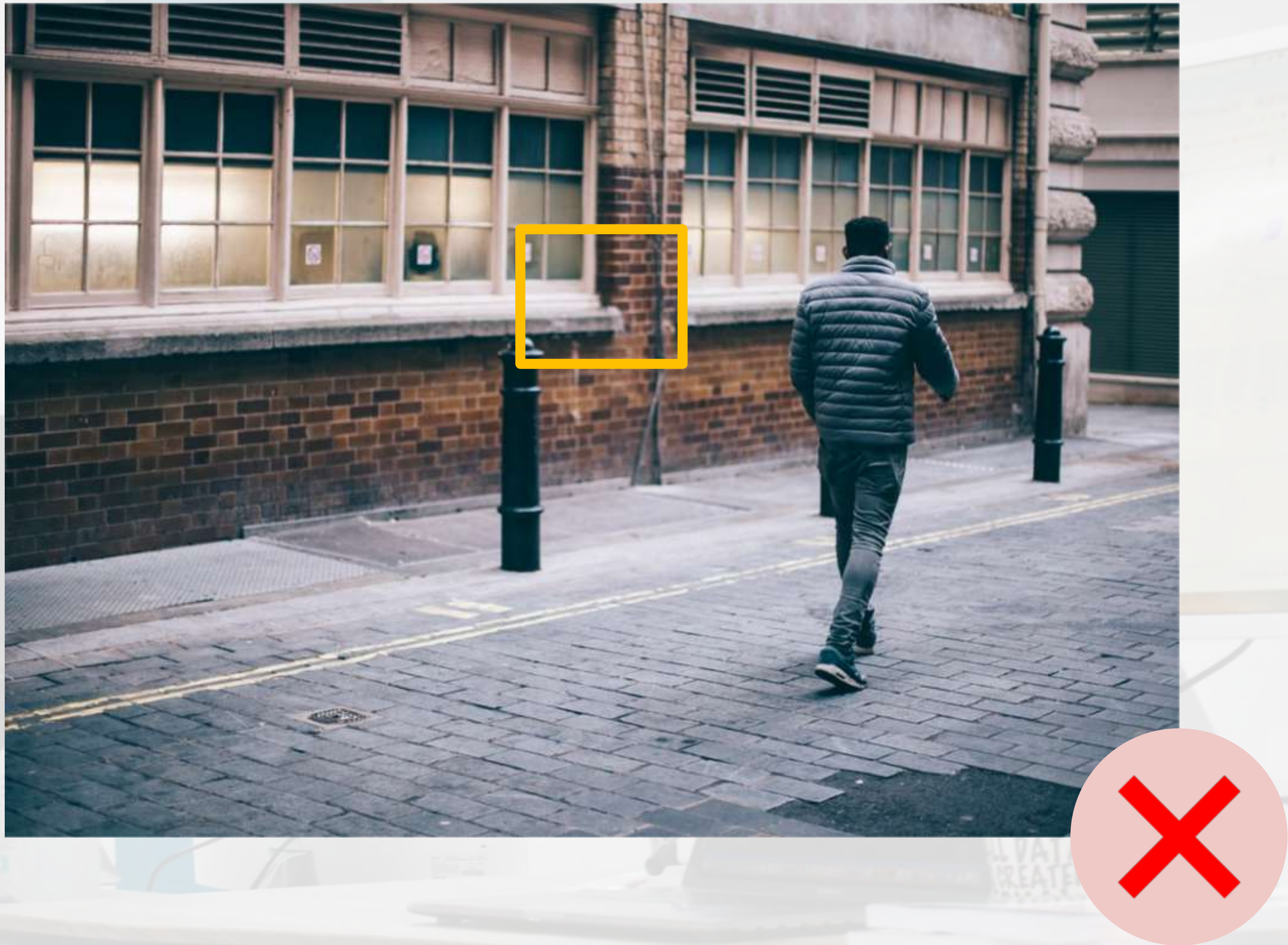
TOWARDS A SMART LABELING STRATEGY



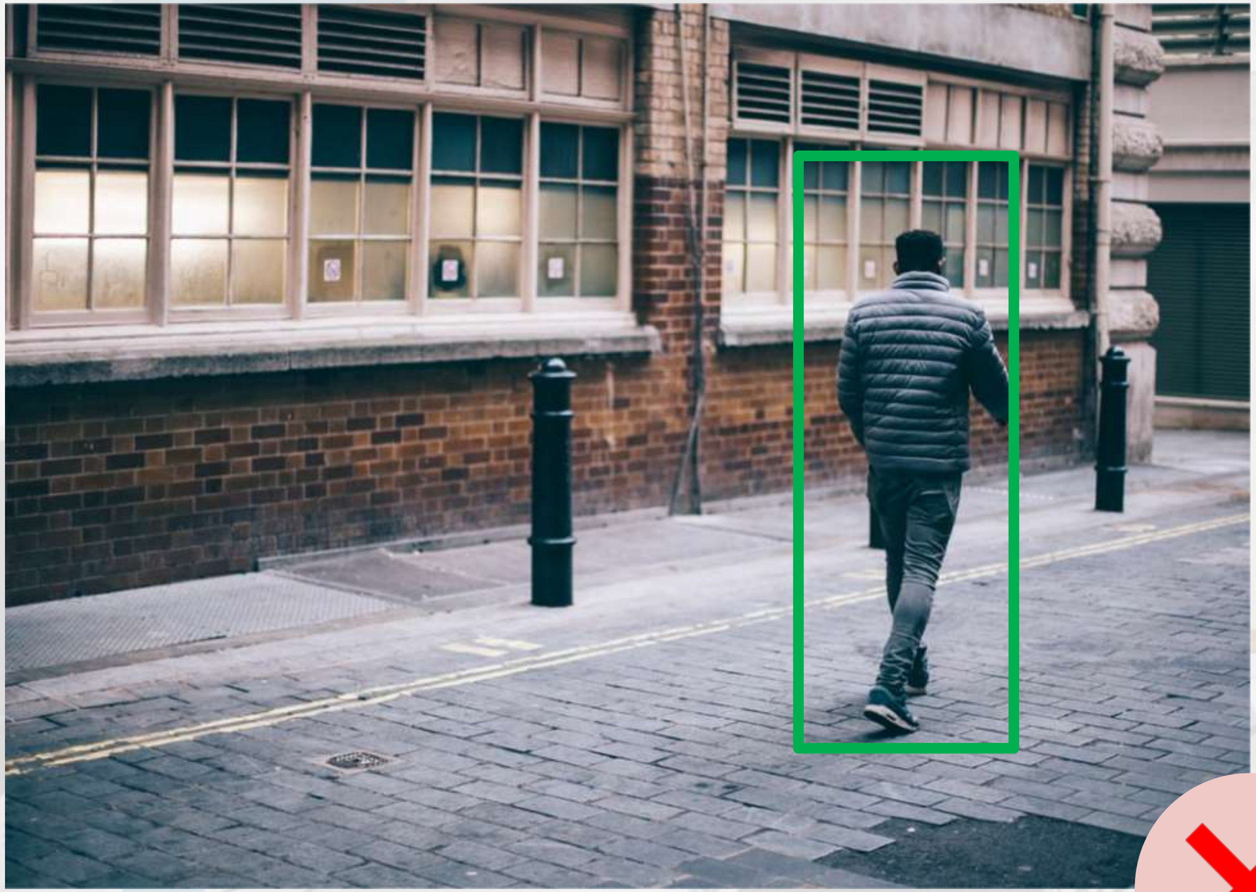
TOWARDS A SMART LABELING STRATEGY



TOWARDS A SMART LABELING STRATEGY

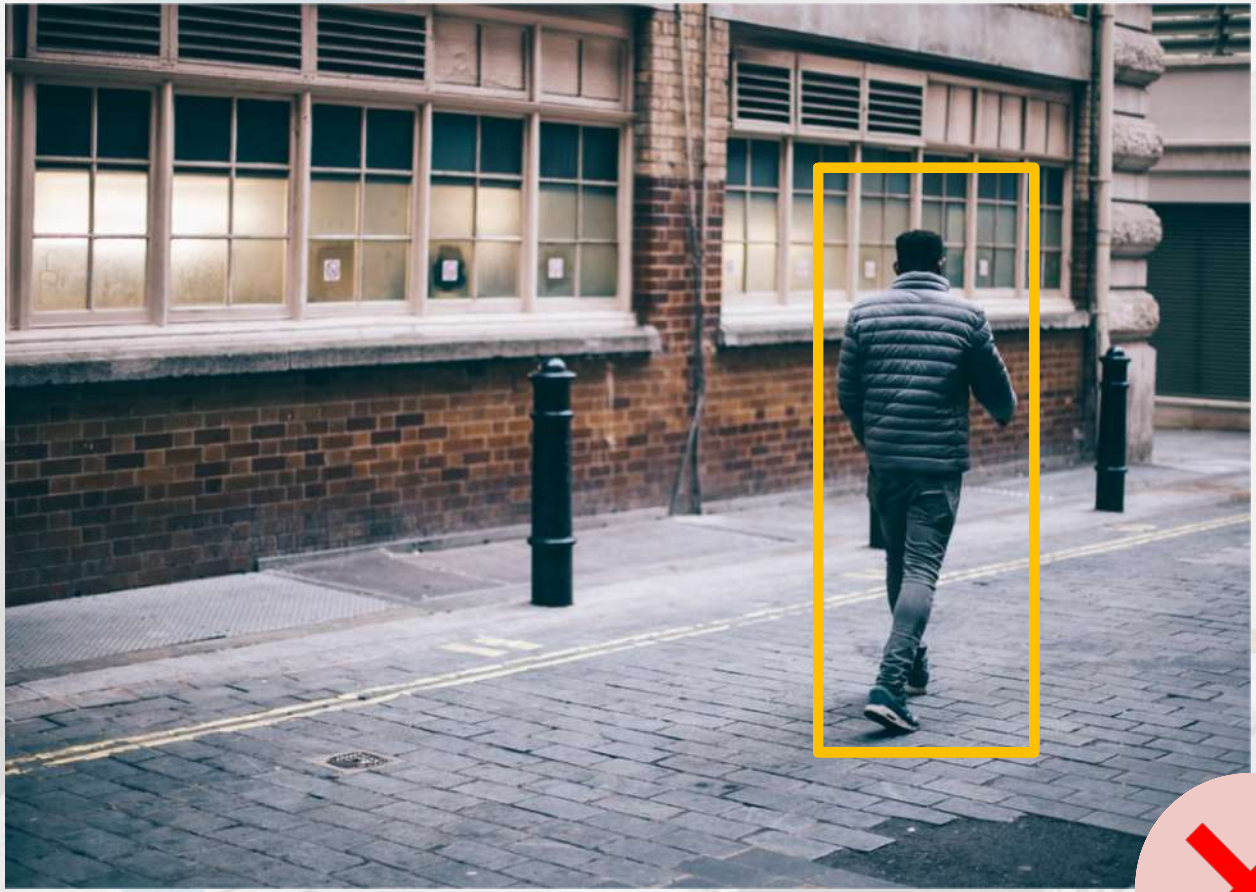


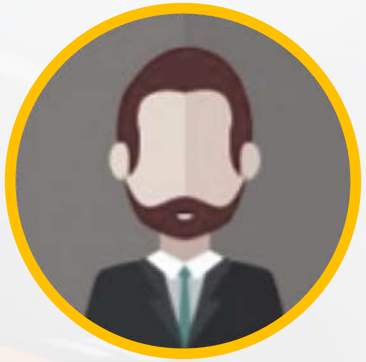
TOWARDS A SMART LABELING STRATEGY



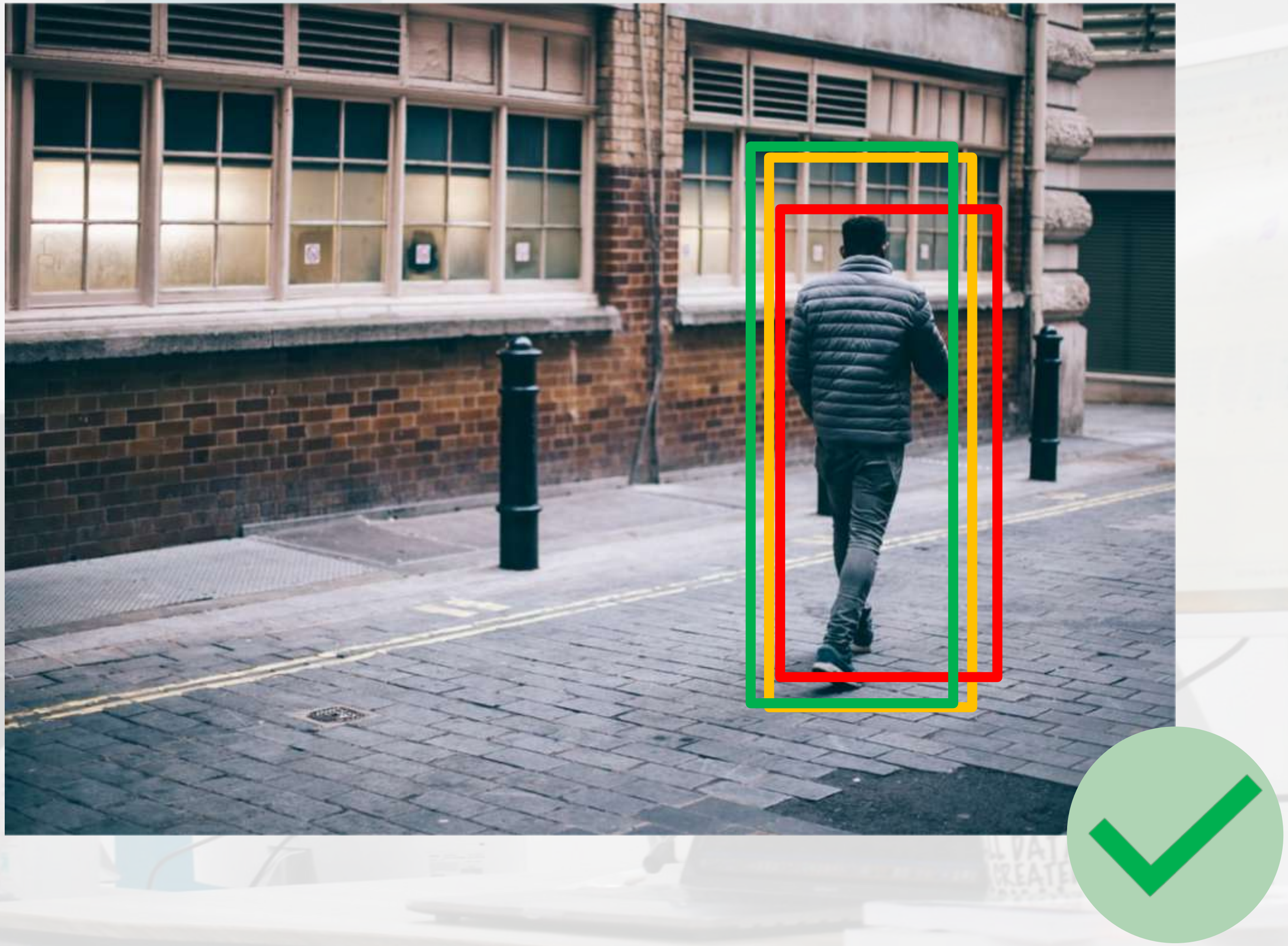
TOWARDS A SMART LABELING STRATEGY

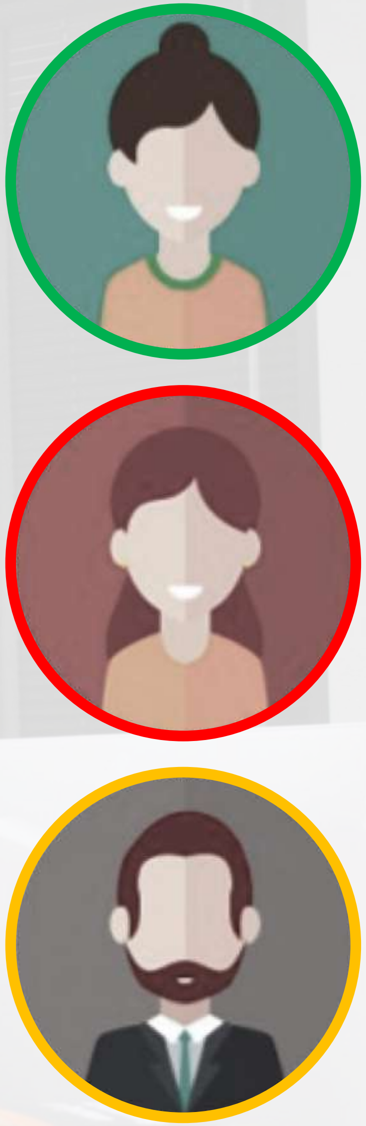
TOWARDS A SMART LABELING STRATEGY





TOWARDS A SMART LABELING STRATEGY





TOWARDS A SMART LABELING STRATEGY

SUPERVISED LEARNING

- All data is labeled
- No. of annotations is predetermined
- No. of annotations is arbitrary

TOWARDS A SMART LABELING STRATEGY

ACTIVE LEARNING

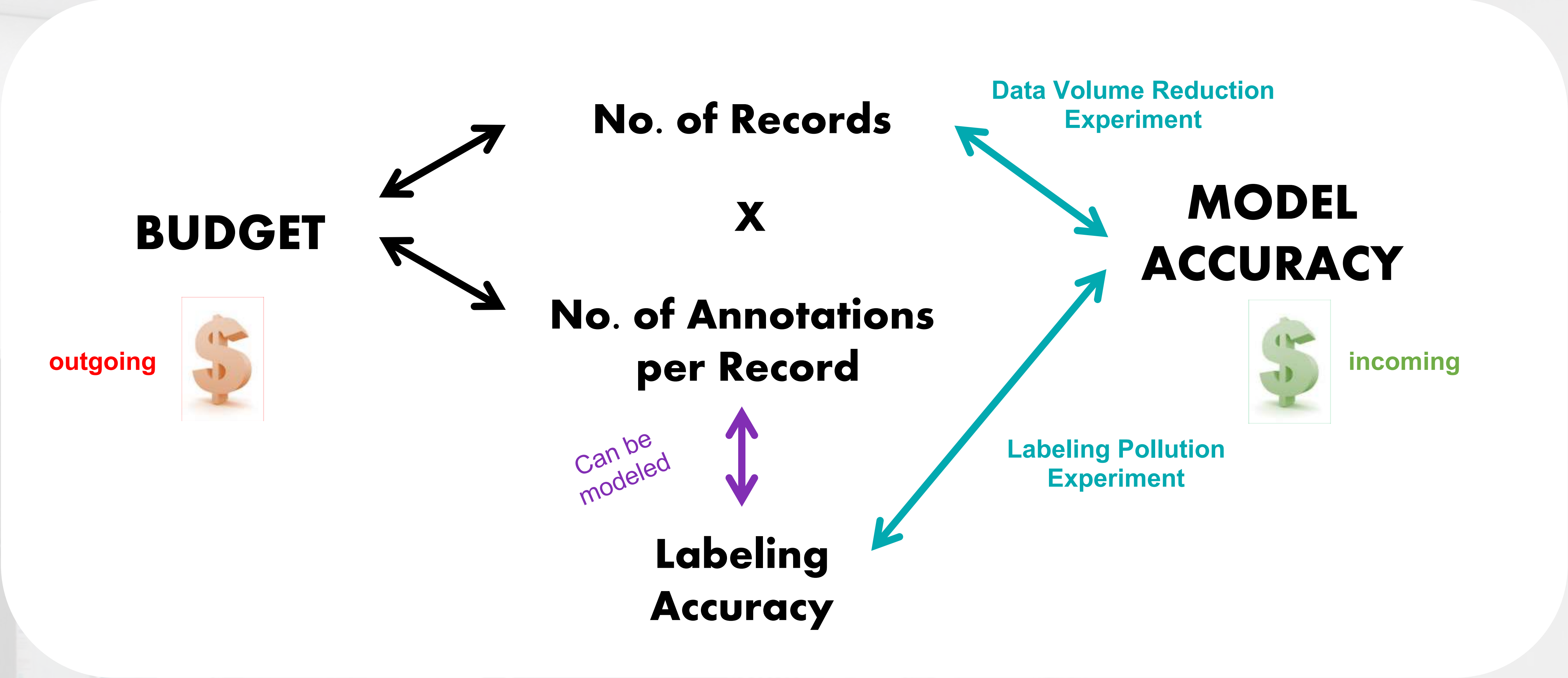
TOWARDS A SMART LABELING STRATEGY

ACTIVE LEARNING

vs.

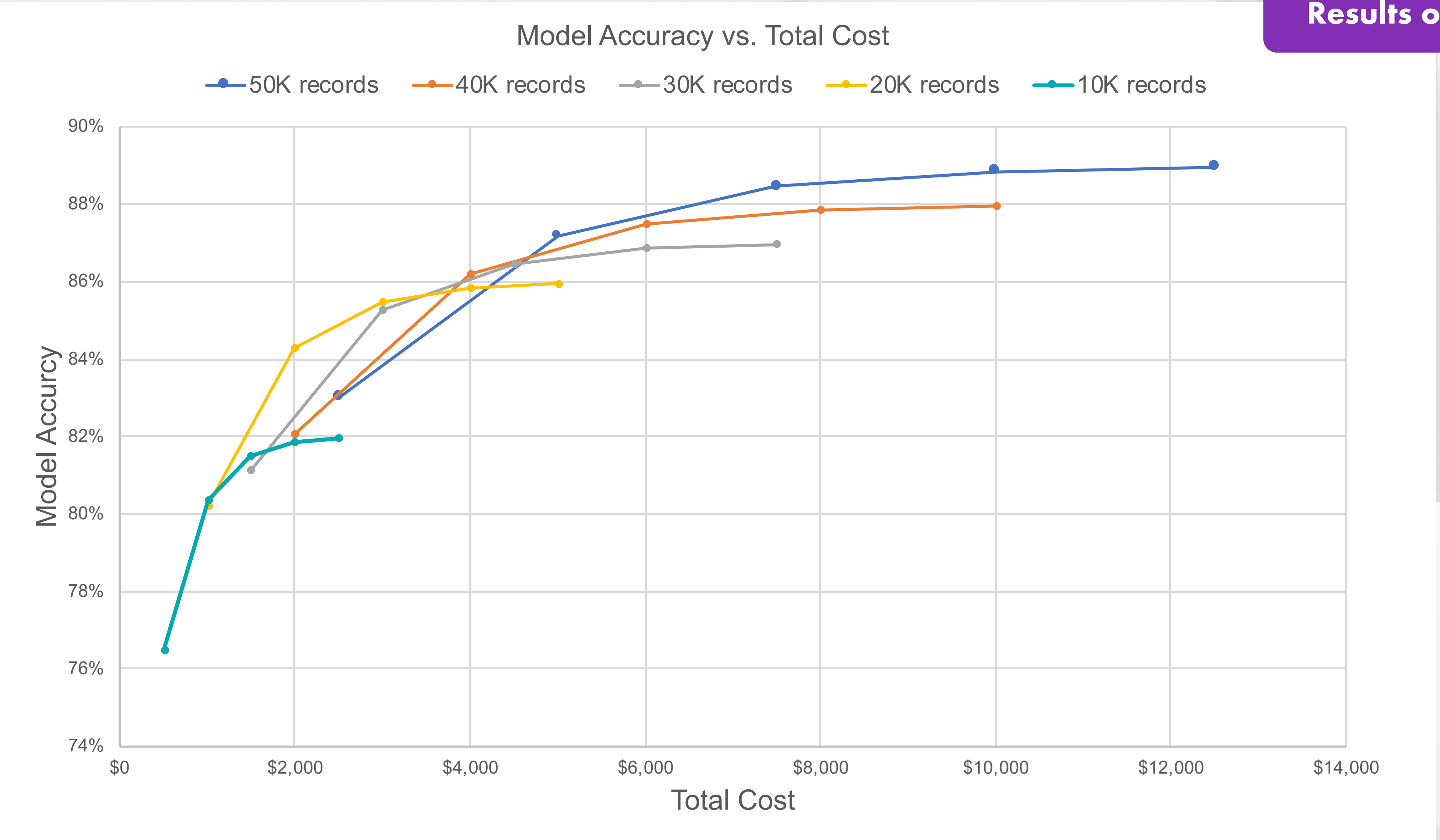
SMART LABELING STRATEGY

TOWARDS A SMART LABELING STRATEGY



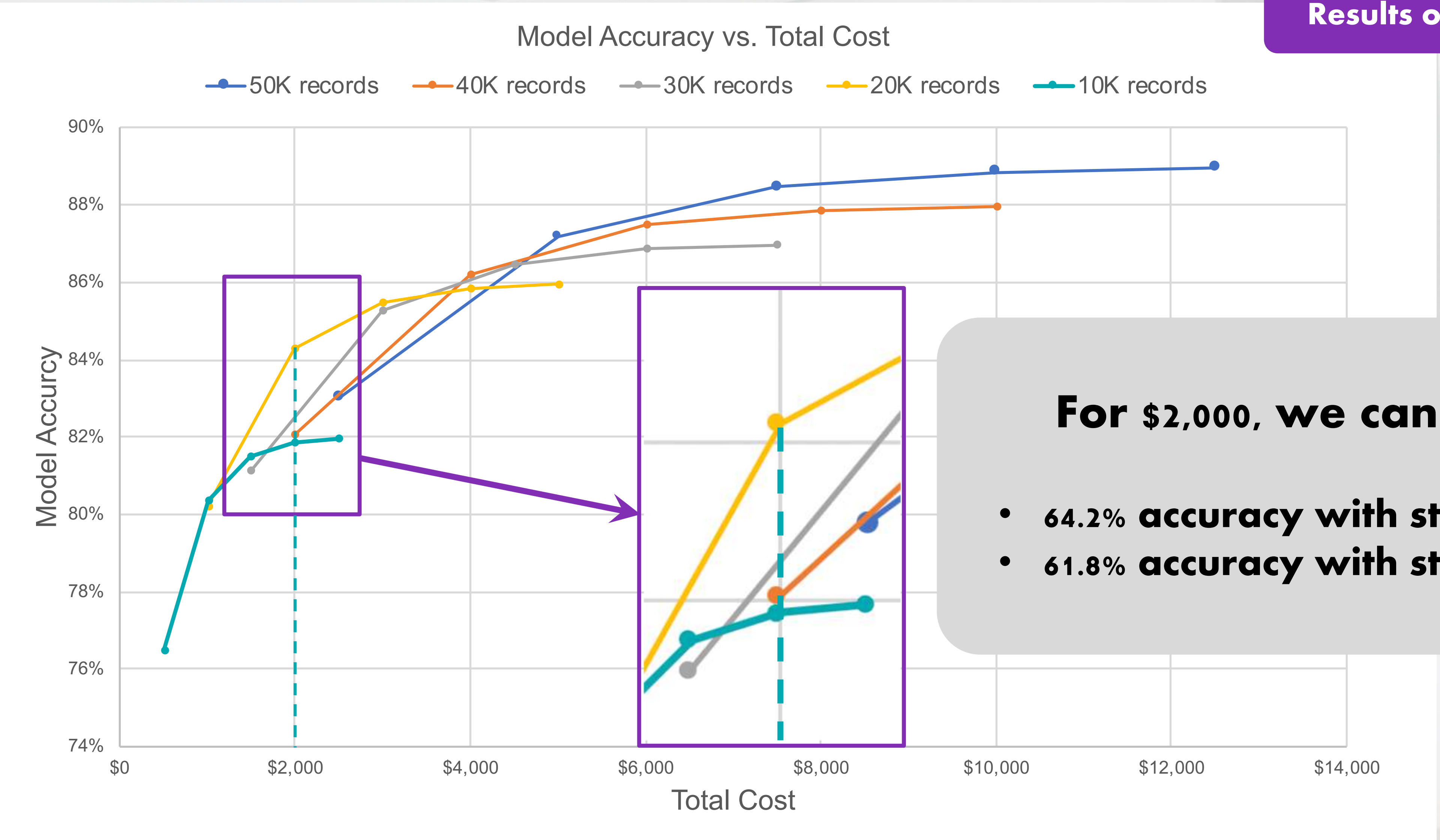
TOWARDS A SMART LABELING STRATEGY

Results on CIFAR-10 Study



TOWARDS A SMART LABELING STRATEGY

Results on CIFAR-10 Study

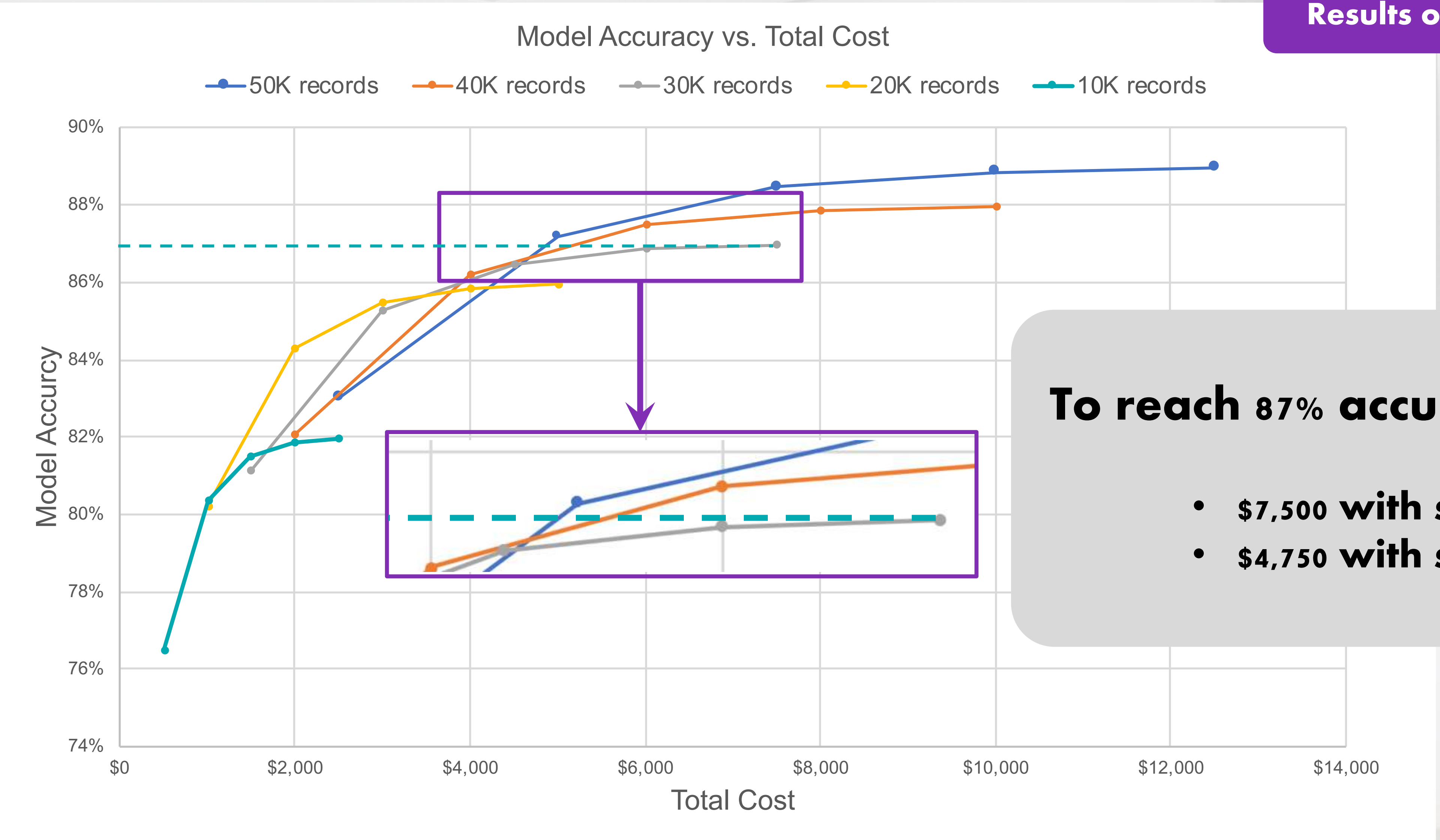


For $2,000, we can get:

- **64.2% accuracy with strategy 4**
- **61.8% accuracy with strategy 1**

TOWARDS A SMART LABELING STRATEGY

Results on CIFAR-10 Study

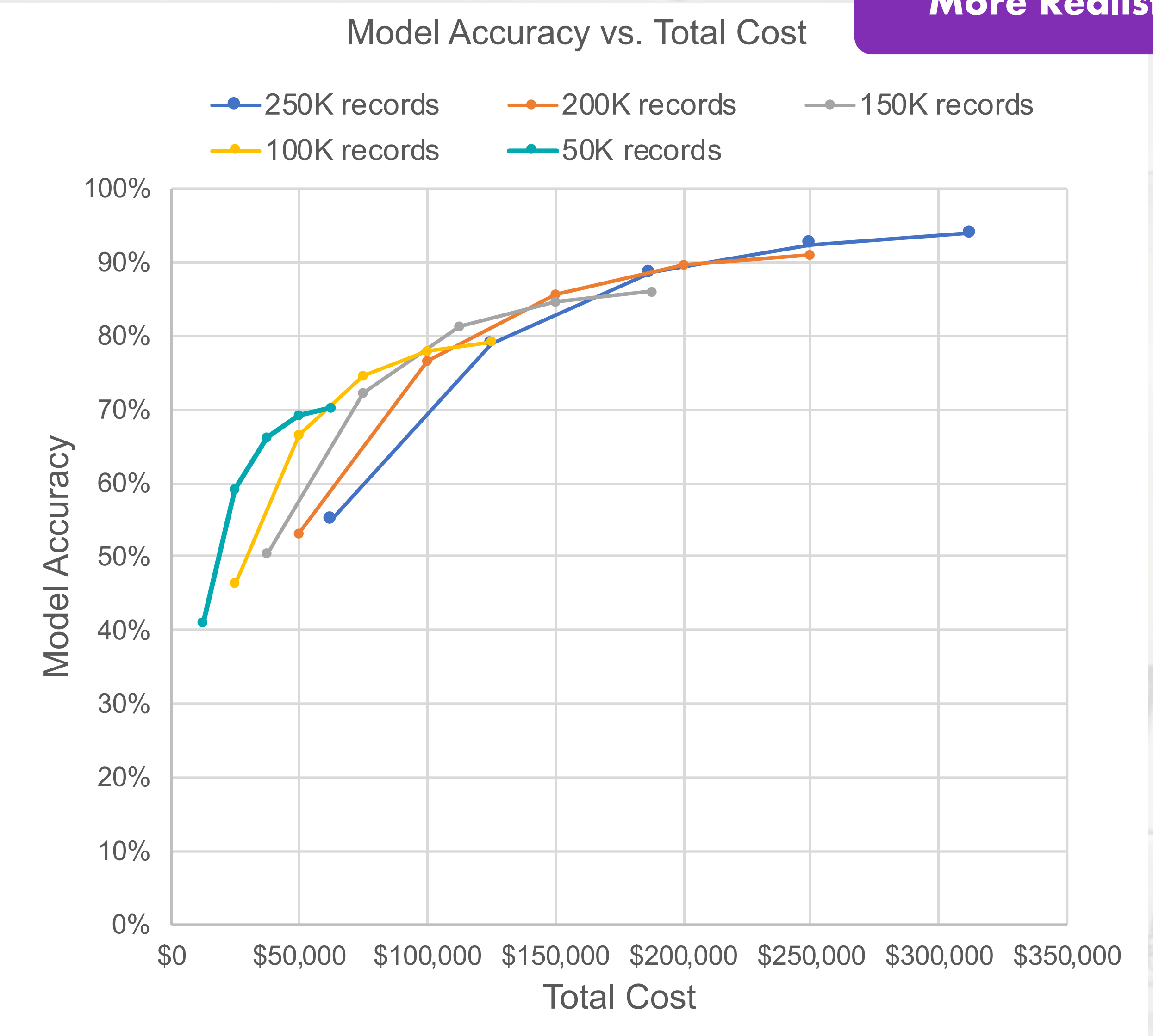


To reach 87% accuracy we need:

- **$7,500 with strategy 3**
- **$4,750 with strategy 1**

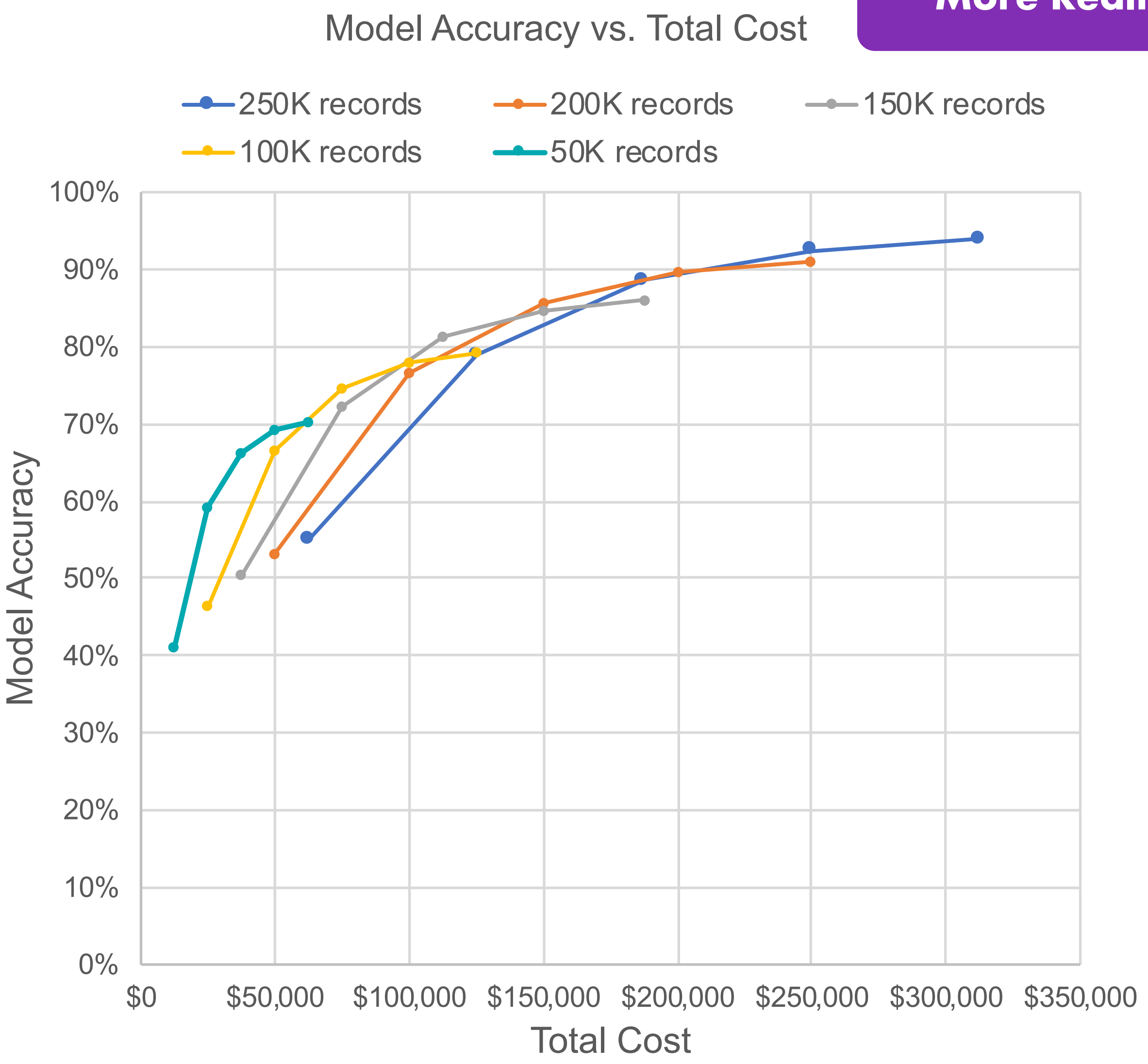
TOWARDS A SMART LABELING STRATEGY

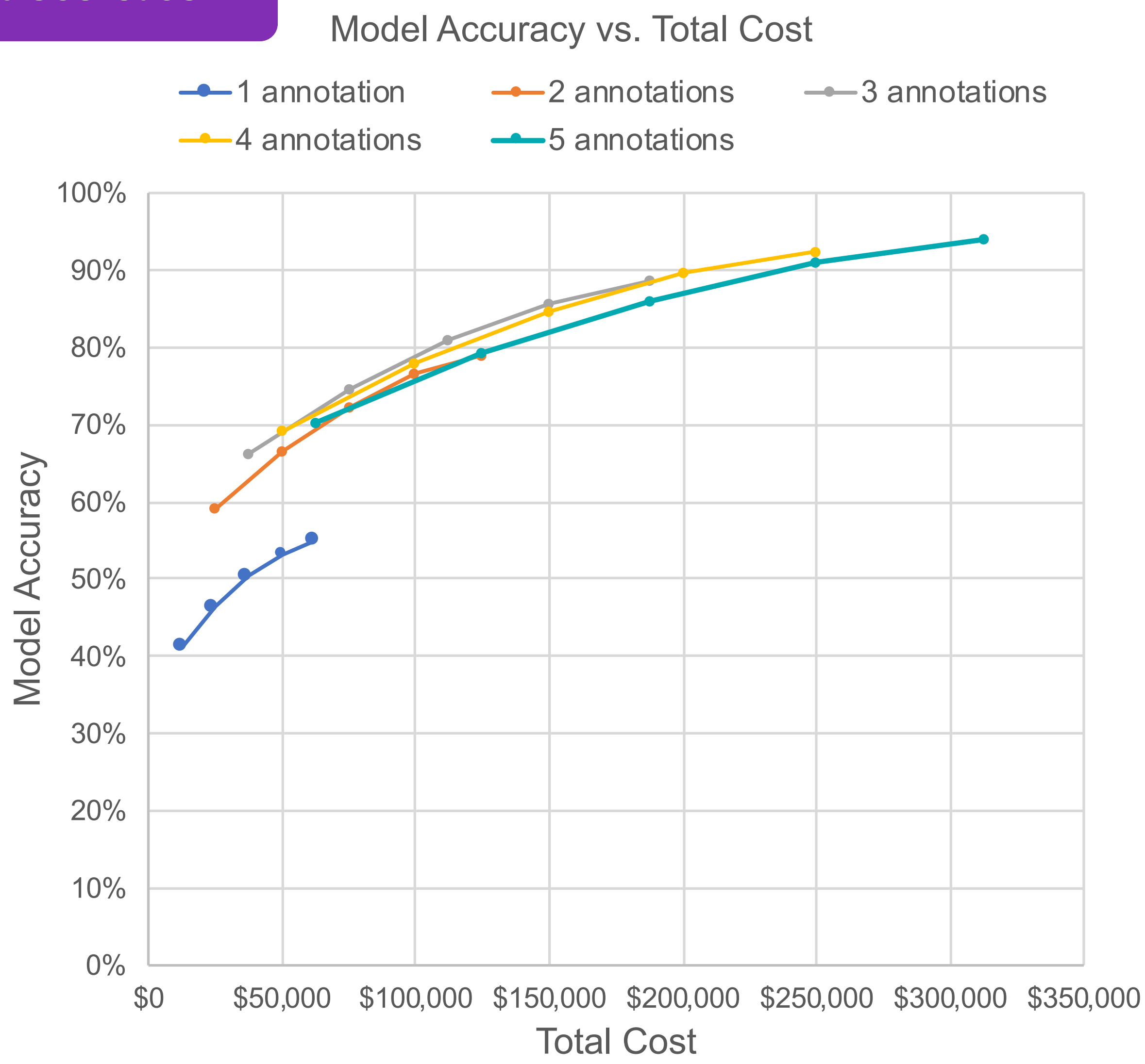
More Realistic Use Case



TOWARDS A SMART LABELING STRATEGY

More Realistic Use Case





TOWARDS A SMART LABELING STRATEGY

NOT COVERED IN THIS TALK:

- Sensitivity by cluster (instead of class)
- Combining data usefulness with difficulty to label
- Combining with AL: “non-binary” Active Learning

CONCLUSIONS

- Class sensitivity is inerrant to the data
  - Not all data requires as much labeling care
  - Better models can't solve everything…
- "Compensating" for bad labels
  - Is more or less difficult depending on the class
  - Might not be possible as all
- Smarter labeling strategies are needed
  - Saving $$ on labeling doesn't necessarily imply labeling less data
  - Local optimization is coming (record level labeling recommendations)
  - Bring the area of non-binary Active Learning

Smart Data > Big Data

THANK YOU!

www.alectio.com

Follow us on LinkedIn & Medium