# INTRODUCTION TO DATA-CENTRIC AI

## Lecture 3 - Dataset Creation and Curation

https://dcai.csail.mit.edu

# After looking through the entire dataset, we have:

| $C_{\tilde{y},y^*}$ | $y^*=dog$ | $y^*=fox$ | $y^*=cow$ |
|---|---|---|---|
| $\tilde{y}=dog$ | 100 | 40 | 20 |
| $\tilde{y}=fox$ | 56 | 60 | 0 |
| $\tilde{y}=cow$ | 32 | 12 | 80 |

# From $C_{\tilde{y},y^*}$ we obtain the joint distribution of label noise

Estimated

| $\hat{p}(\tilde{y}, y^*)$ | $y^*=dog$ | $y^*=fox$ | $y^*=cow$ |
|---|---|---|---|
| $\tilde{y}=dog$ | 0.25 | 0.1 | 0.05 |
| $\tilde{y}=fox$ | 0.14 | 0.15 | 0 |
| $\tilde{y}=cow$ | 0.08 | 0.03 | 0.2 |

# Dataset Curation: ImageNet Train Set

The largest off-diagonals of $\boldsymbol{C}(\tilde{y}, y^*)$ reveal ontological issues.

Note the **(is a)** and **(has a)** relationships

Does this also work for val/test sets?

# Dataset Curation: ImageNet Val Set

| | | | |
|---|---|---|---|
| 26 | n02979186 | cassette_player | n04392985 tape_player |
| 23 | n03773504 | missile | n04008634 projectile |
| 23 | n03642806 | laptop | n03832673 notebook |
| 23 | n02808440 | bathtub | n04493381 tub |
| 23 | n13133613 | ear | n12144580 corn |
| 22 | n03710721 | maillot | n03710637 maillot |
| 22 | n01682714 | American_chameleon | n01693334 green_lizard |
| 21 | n02895154 | breastplate | n03146219 cuirass |
| 20 | n02412080 | ram | n02415577 bighorn |
| 19 | n04008634 | projectile | n03773504 missile |
| 18 | n01753488 | horned_viper | n01756291 sidewinder |
| 18 | n02107908 | Appenzeller | n02107574 Greater_Swiss_Mountain_dog |
| 18 | n12144580 | corn | n13133613 ear |
| 17 | n03146219 | cuirass | n02895154 breastplate |
| 17 | n02113624 | toy_poodle | n02113712 miniature_poodle |
| 16 | n03710637 | maillot | n03710721 maillot |

# There are indistinguishable examples in these classes


## Appenzeller Sennenhund

Dog breed

The Appenzeller Sennenhund is a medium-size breed of dog, one of the four regional breeds of Sennenhund-type dogs from the Swiss Alps. The name Sennenhund refers to people called Senn, herders in the Appenzell region of Switzerland. Wikipedia


## Greater Swiss Mountain Dog

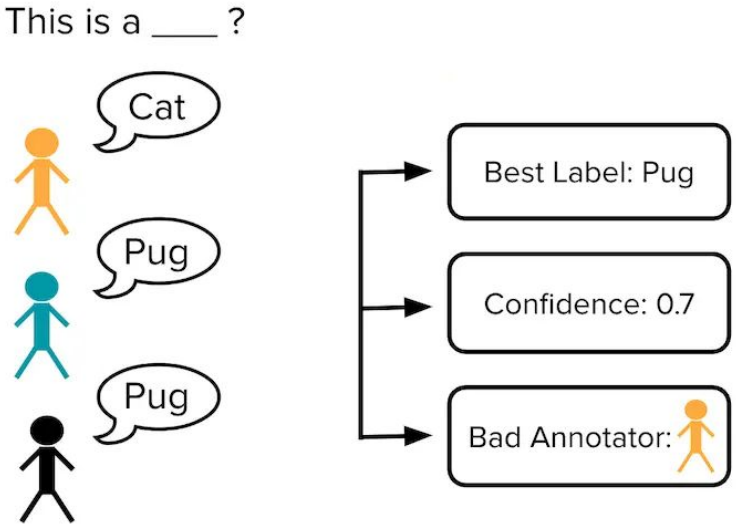
Dog breed

The Greater Swiss Mountain Dog is a dog breed which was developed in the Swiss Alps. The name Sennenhund refers to people called Senn or Senner, dairymen and herders in the Swiss Alps. Wikipedia

Can you spot the difference?

# Rest of Lecture

- Concerns when sourcing the data
  - selection bias
  - confounding
  - distribution shift




- Concerns when sourcing the labels
  - how to work with multiple data annotators and assess quality

Good Textbook: **Human-in-the-loop Machine Learning** by R. Munro, R. Monarch

# Key questions when sourcing training data

1. How will the resulting ML model be used?

    - On what population will model be making predictions and when

2. Hypothetical edge cases where we need model to make the right prediction?

    - High stakes scenarios, rare events

# Trained image classifier predicts the left image contains a cow, but this model fails to make same prediction for the right image

Beery et al. 2018

This slide intentionally left blank