



HAL
open science

Deep Learning in Medical Imaging: What's Needed for Training Data?

Francesca Galassi

► **To cite this version:**

Francesca Galassi. Deep Learning in Medical Imaging: What's Needed for Training Data?. Biogenouest days 2023, Oct 2023, Rennes (Campus Villejean), France. hal-04247143

HAL Id: hal-04247143

<https://inria.hal.science/hal-04247143v1>

Submitted on 18 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Inria



Deep Learning in Medical Imaging: What's Needed for Training Data?

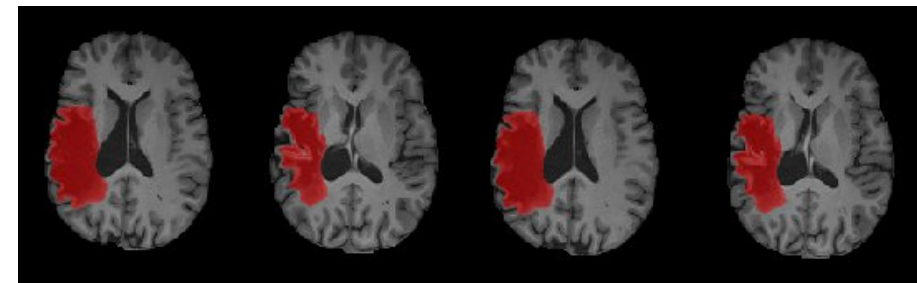
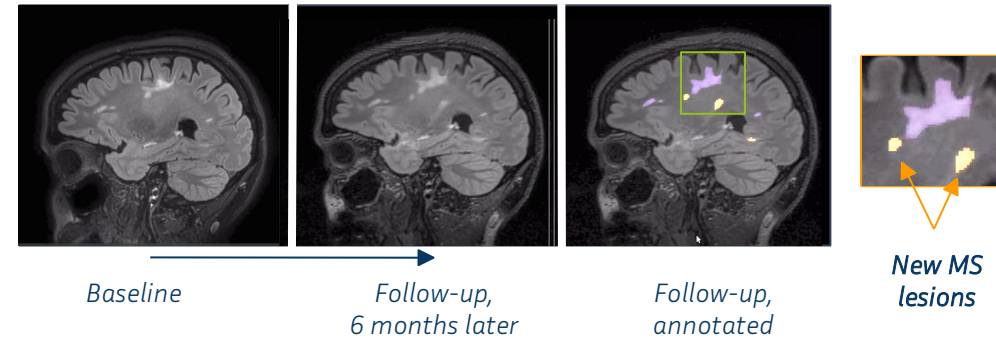
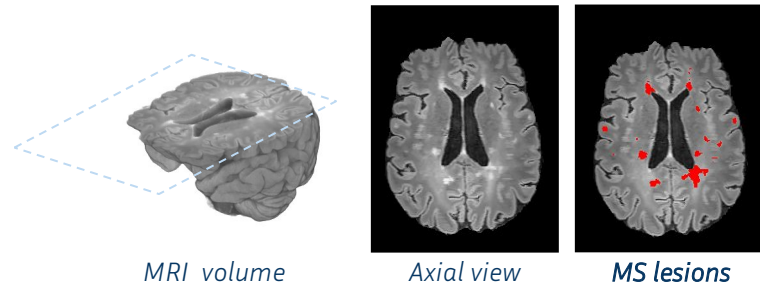
Galassi Francesca, MCF, Empenn lab

Inria Rennes

francesca.galassi@irisa.fr



Deep Learning in Medical Imaging



Stroke lesion

Hussein, B. R., Meurée, C., Gaubert, M., Masson, A., Kerbrat, A., Combès, B., & Galassi, F. (2023, April). A study on loss functions and decision thresholds for the segmentation of multiple sclerosis lesions on spinal cord MRI. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.

Masson, A., Le Bon, B., Kerbrat, A., Edan, G., Galassi, F., & Combes, B. (2021). A nnUnet implementation of new lesions segmentation from serial FLAIR images of MS patients. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, 5. MICCAI MSSEG-2 Challenge.

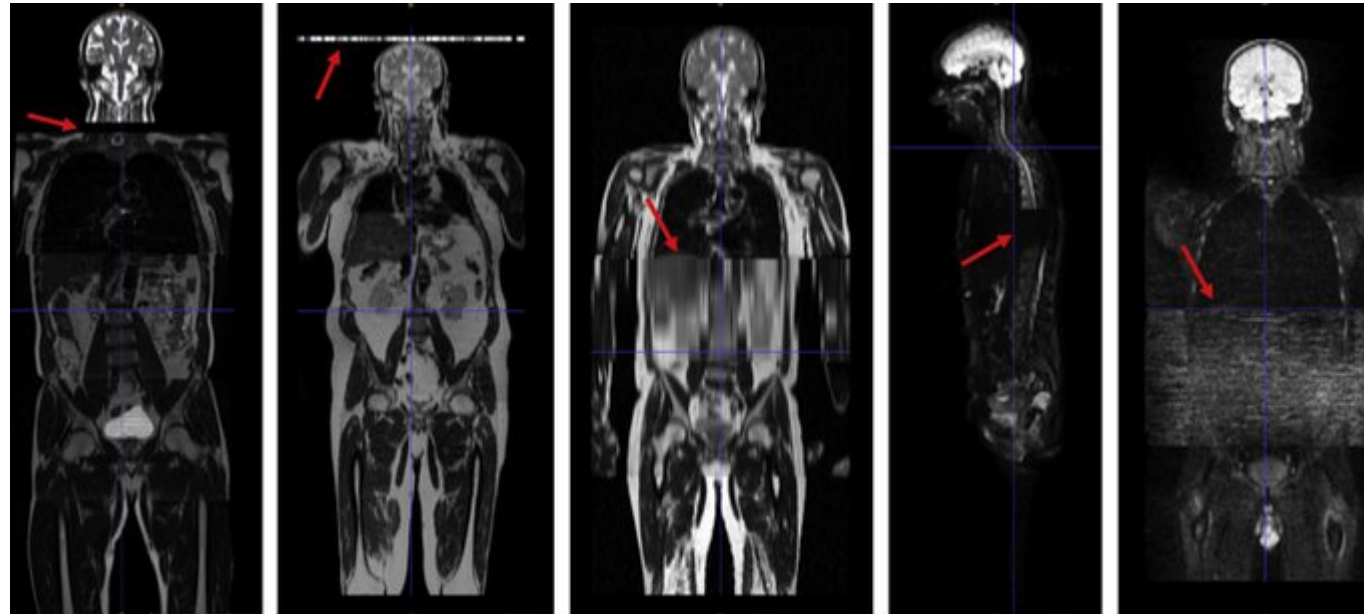
Good vs. bad

"Data, data everywhere and not a set to process."

Neil Lawrence

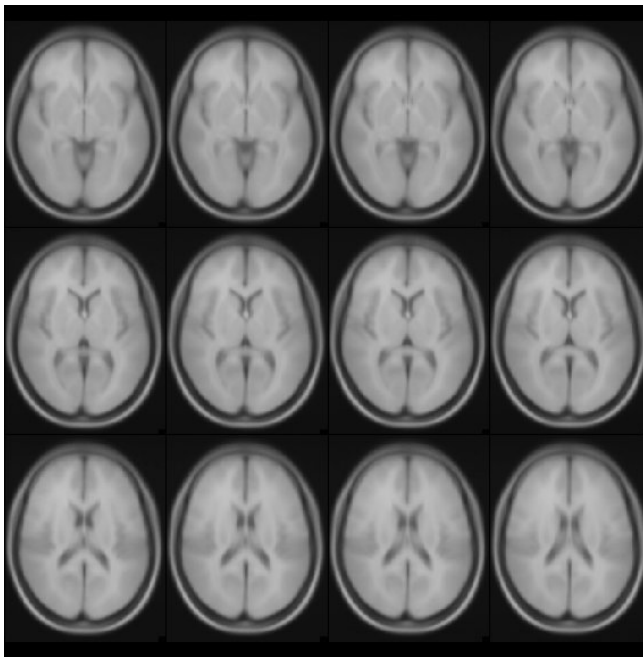
Data Quality

- **Artifacts, noise, and corruption**
- Considerations for **robustness** and **quality control**

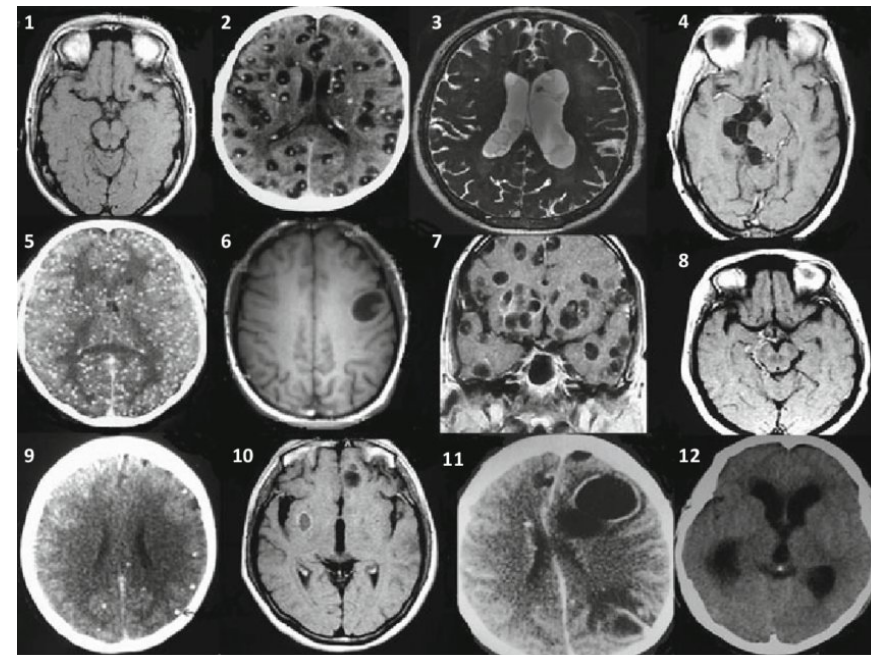


Data Variety

- Training on homogenous data may not **generalize** well to **heterogeneous** test data.
- Consider adopting a **versatile** or **narrowly focused** approach.



MNI atlas



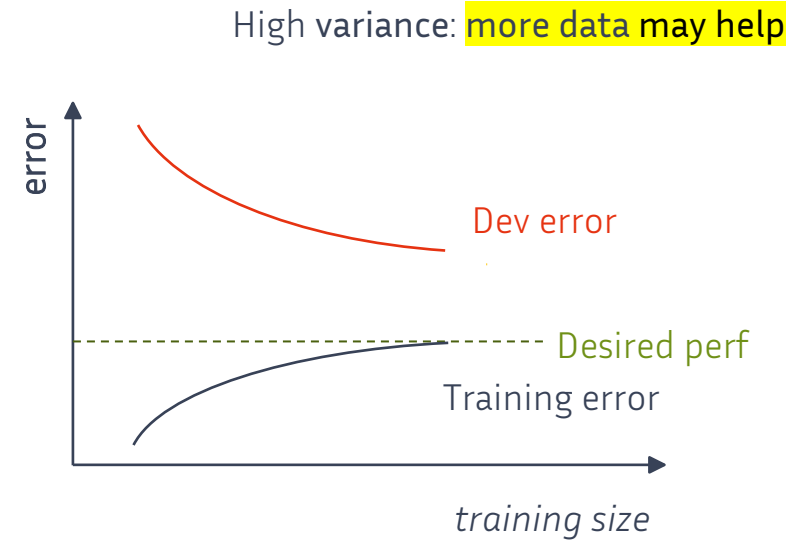
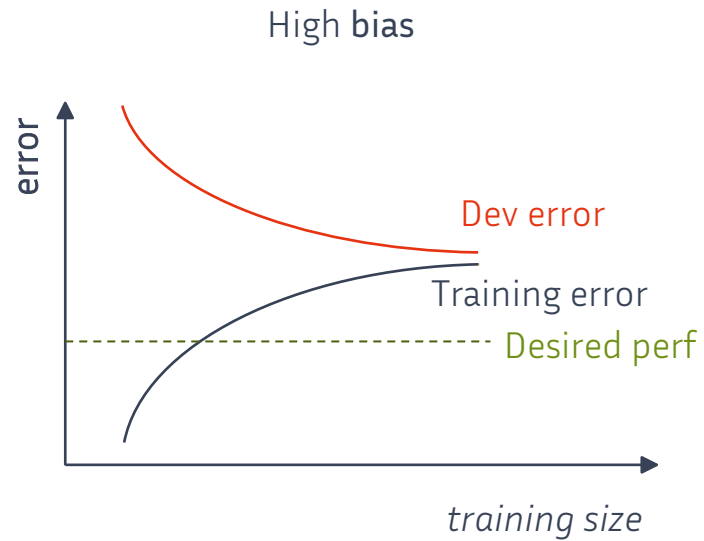
Data Volume

How much data do you need? - Clinician

How much data do you have? - DL Researcher

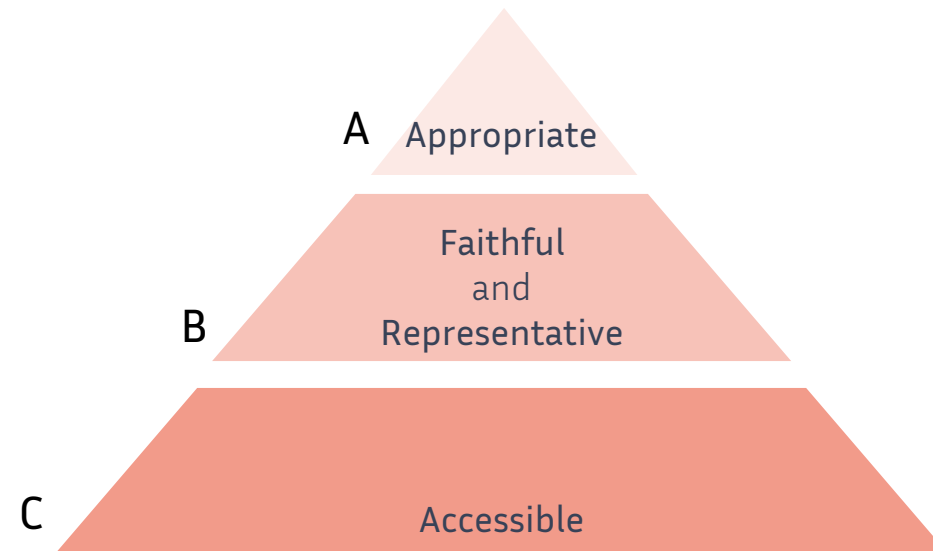
- **Traditional statistical rules don't** always **fit** the machine learning paradigm
- DL has **mechanisms** to determine whether more **data** could help

Data Volume



Data Readiness

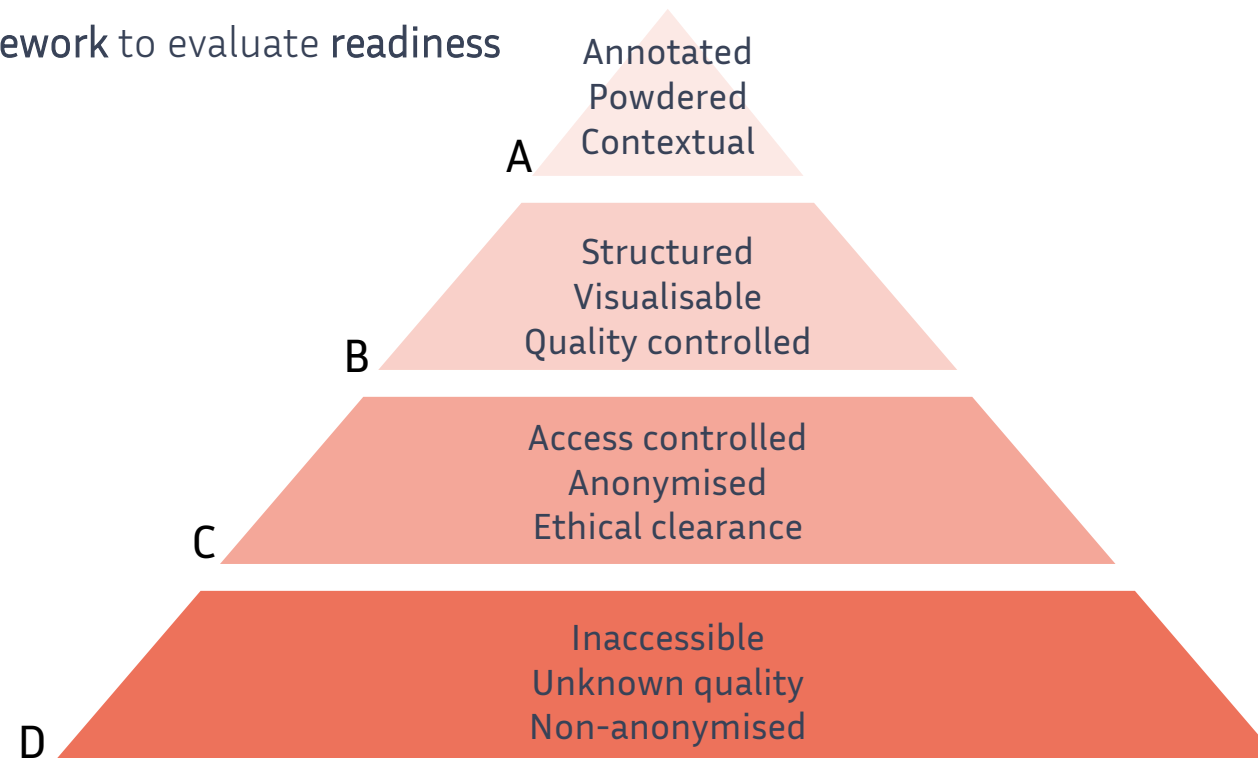
- Are data **ready** for a DL process?
- Consider using a **framework** to evaluate **readiness**



Lawrence, N. D. (2017). Data readiness levels. *arXiv preprint arXiv:1705.02245*.

Data Readiness

- Are data **ready** for a DL process?
- Consider using a **framework** to evaluate **readiness**



Harvey, H., & Glocker, B. (2019). *A standardised approach for preparing imaging data for machine learning tasks in radiology*. Artificial intelligence in medical imaging: opportunities, applications and risks, 61-72.

Dataset shift

“Dataset shift is any situation in which the training and test data distributions disagree due to exogenous factors, e.g. dissimilar cohorts or inconsistent acquisition processes.”

Daniel C. Castro

Predictive Modeling

- ❖ Given an image X , train a model to predict some annotation Y

$$P(Y|X)$$

Assumptions:

- ❖ Availability of sufficient training data (X, Y) – *data volume*
- ❖ Consistency in training and testing data distribution – *data variety*

Challenges in DL for medical imaging: *Data Scarcity & Data Mismatch*

Image and Annotation – *a causal perspective*

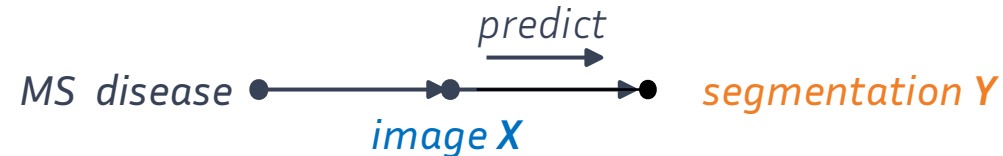
What is the relationship between image X and annotation Y ?

cause X \longrightarrow Y *effect*
causal
predict effect from cause

Example: MS Lesion segmentation

X - structural brain MRI scan

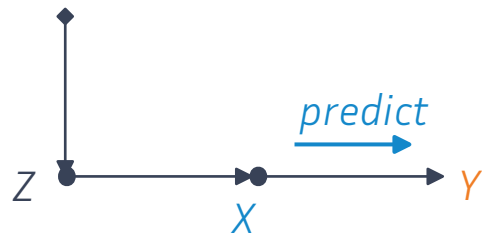
Y - manual segmentation masks



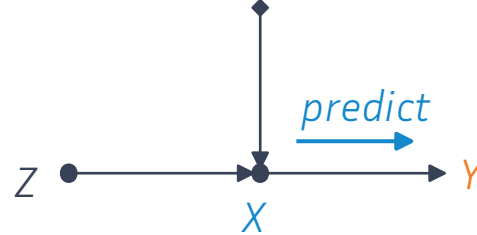
Dataset Shift – *data mismatch*

❖ Training and test data distributions may **disagree** due to ...

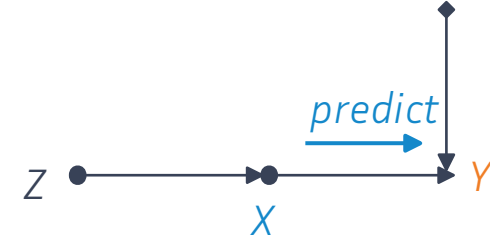
Population shift



Acquisition shift



Annotation shift



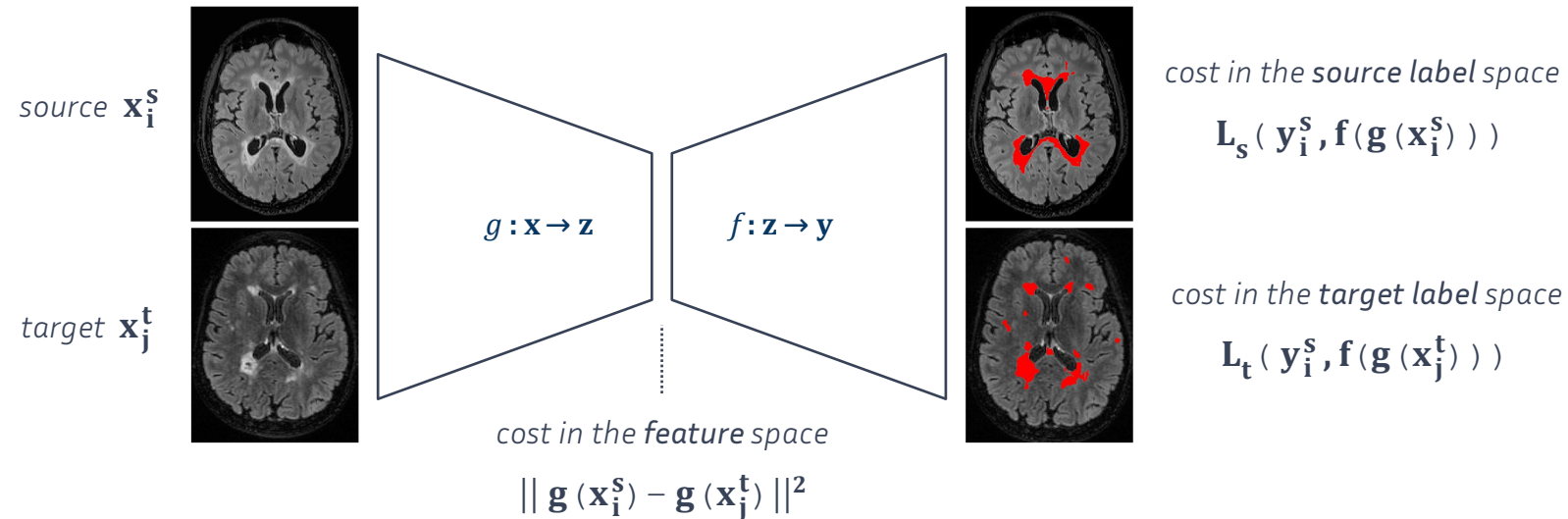
Z: unobserved true anatomy

X: image

Y: annotation

Domain Adaptation – a case study

- Unsupervised DA – *unlabeled target data accessible at training time*

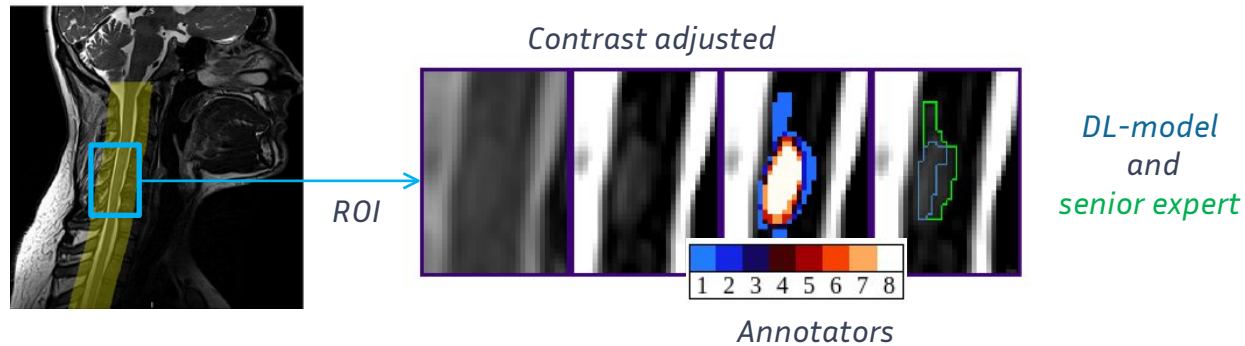


- Optimization problem:

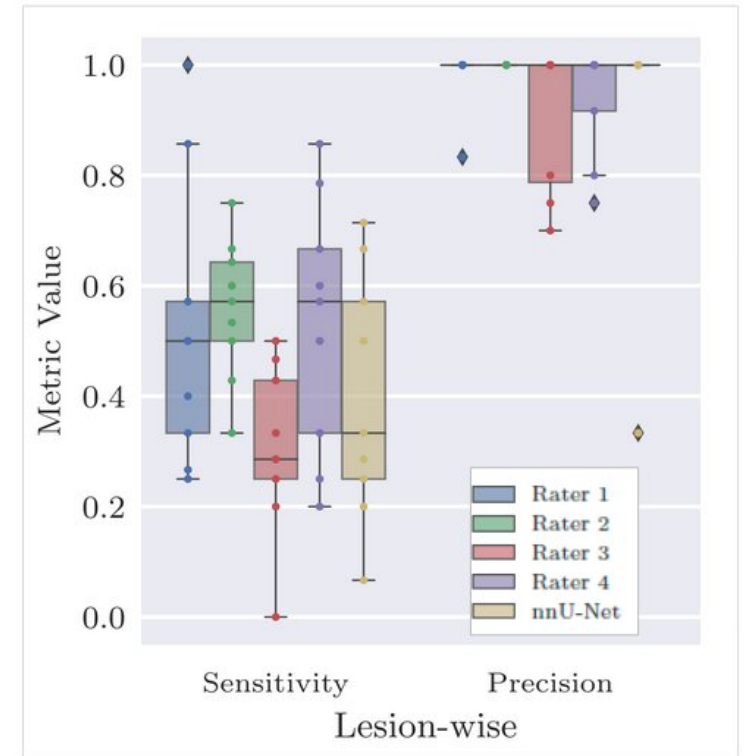
$$\min_{\gamma, \mathbf{f}, \mathbf{g}} \sum_i L_s(\mathbf{y}_i^s, \mathbf{f}(\mathbf{g}(\mathbf{x}_i^s))) + \sum_{i,j} \gamma_{i,j} (\alpha \|\mathbf{g}(\mathbf{x}_i^s) - \mathbf{g}(\mathbf{x}_j^t)\|^2 + \beta L_t(\mathbf{y}_i^s, \mathbf{f}(\mathbf{g}(\mathbf{x}_j^t))))$$

the coupling matrix

Annotation Shift – a case study



- Considerable inter- and intra-rater variability observed
- High precision and low sensitivity of individual experts



Key Takeaways

- Data **quality, volume** and **variety** matter
- **Error analysis** guides model improvements and data collection
 - A **data readiness framework** can help assessing data suitability for a DL process
- Dataset shifts arise from factors like **population, acquisition, and annotation variations**
 - **Causal diagram** can help describing the shift in the specific project
- Consider **strategies** for addressing these challenges, i.e., label noise reduction and domain adaptation