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

Francesca Galassi

Deep Learning in Medical Imaging:

What's Needed for Training Data?

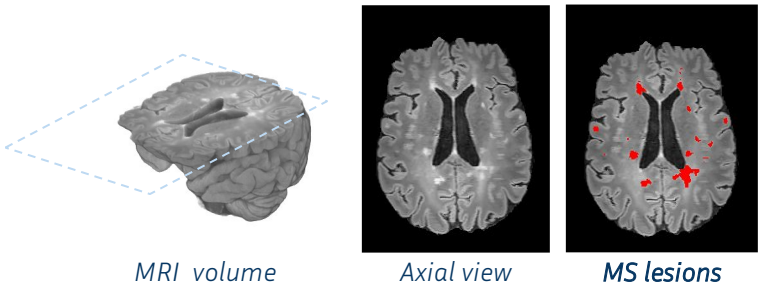
Galassi Francesca, MCF, Empenn lab

Inria Rennes

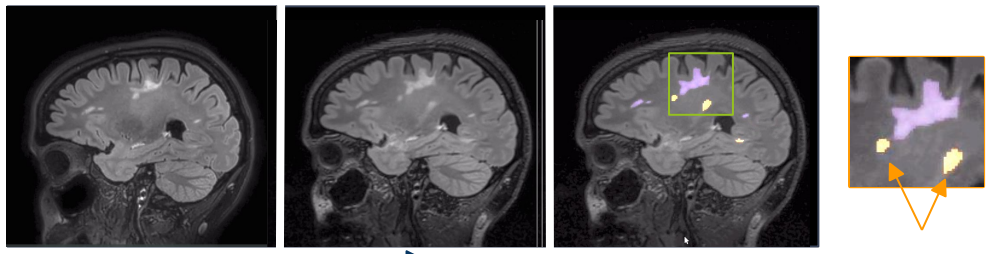
francesca.galassi@irisa.fr

Deep Learning in Medical Imaging

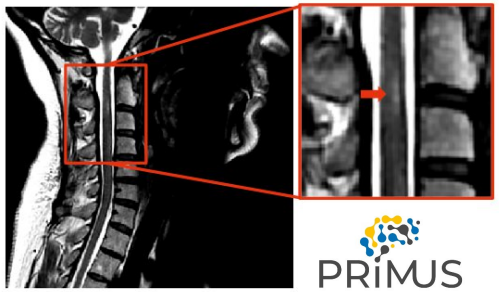
EMPENN

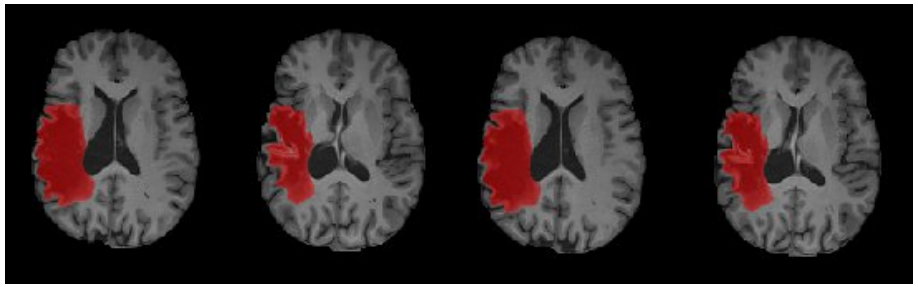


MRI volume *Axial view* ***MS lesions***



Baseline *Follow-up, 6 months later* *Follow-up, annotated* ***New MS lesions***





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.

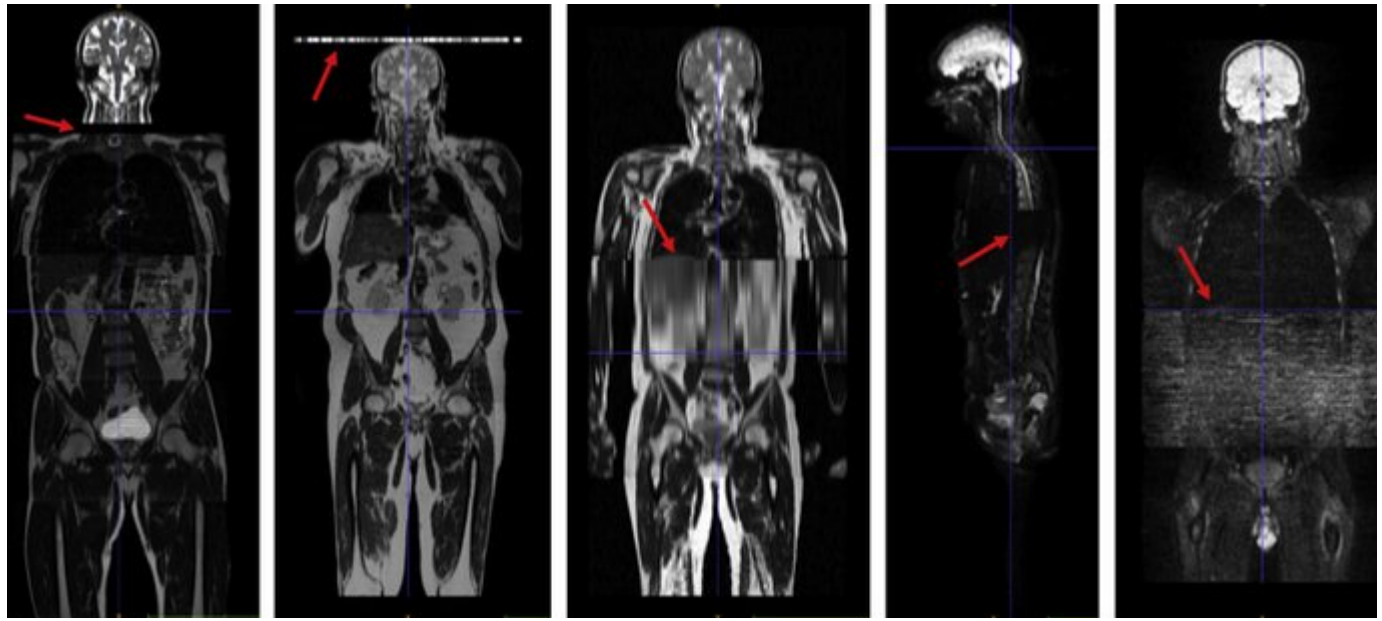
Good vs. bad

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

Neil Lawrence

Data Quality

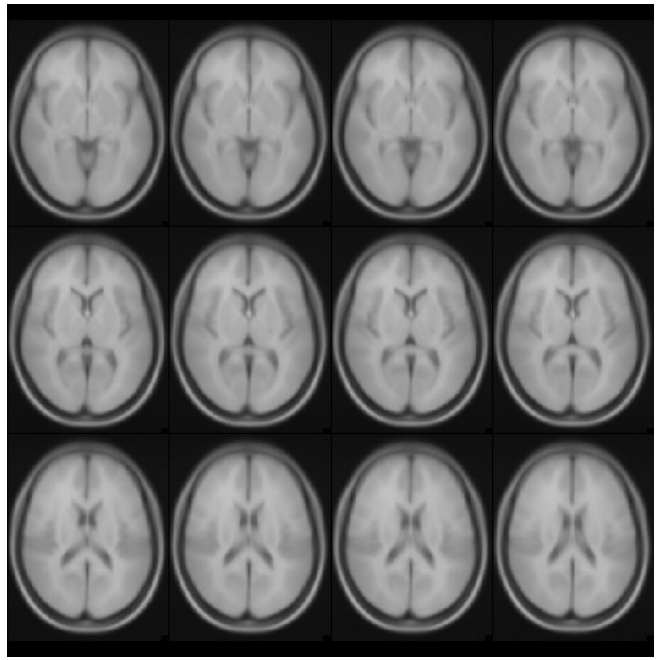
- **Artifacts**, **noise**, and **corruption**
- Considerations for **robustness** and **quality control**



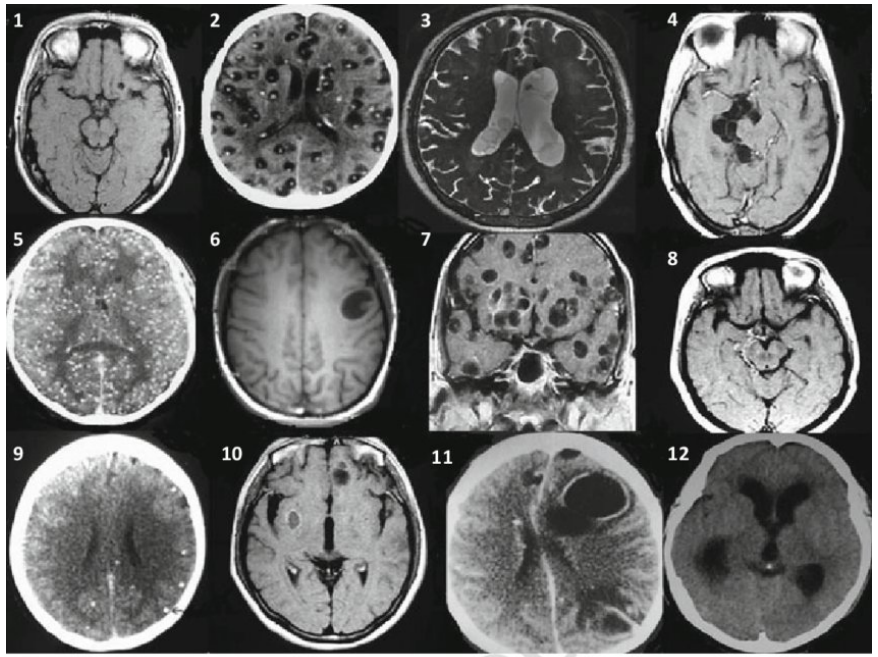
Lavdas, I., Glocker, B., Rueckert, D., Taylor, S. A., Aboagye, E. O., & Rockall, A. G. (2019). Machine learning in whole-body MRI: experiences and challenges from an applied study using multicentre data. Clinical radiology, 74(5), 346-356.

Data Variety

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



MNI atlas



Helminth Infections and their Impact on Global Public Health (pp.180-207). Edition: 1, Chapter: 7. Publisher: Life Sciences Springer Wien. Editors: Fabrizio Bruschi

Data Volume

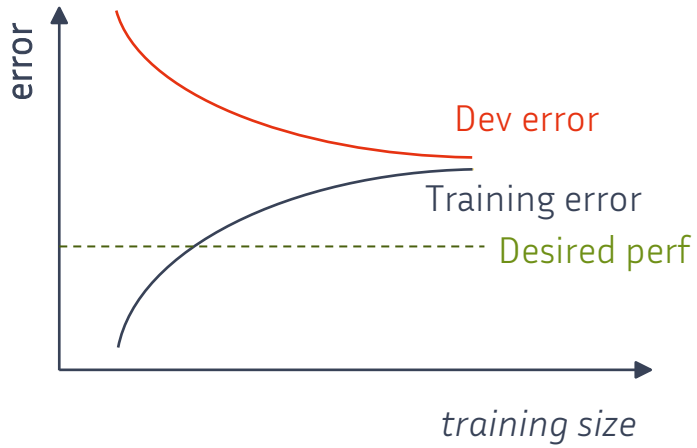
How much data do you need? - Clinician

How much data do you have ? - DL Researcher

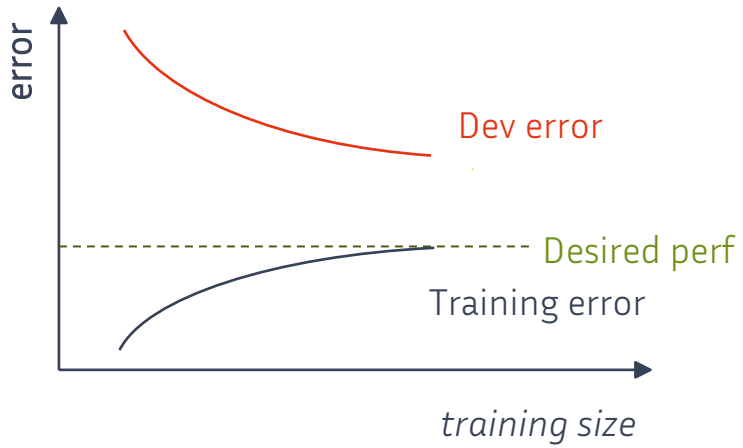
- **Traditional statistical rules don't** always **fit** the machine learning paradigm
- DL has **mechanisms** to **determine** whether more **data** could help

Data Volume

High **bias**



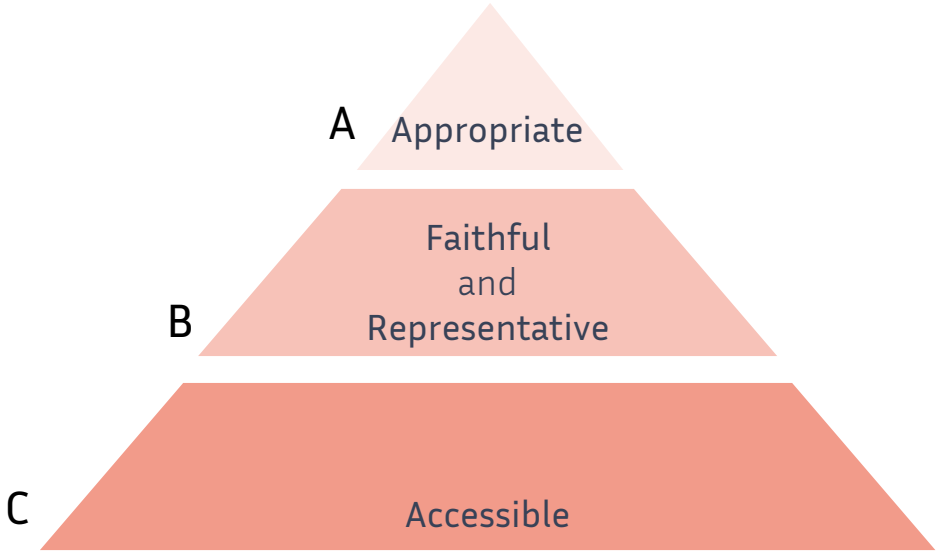
High **variance**: more data may help



Source of inspiration: *Andrew NG machine learning course*

Data Readiness

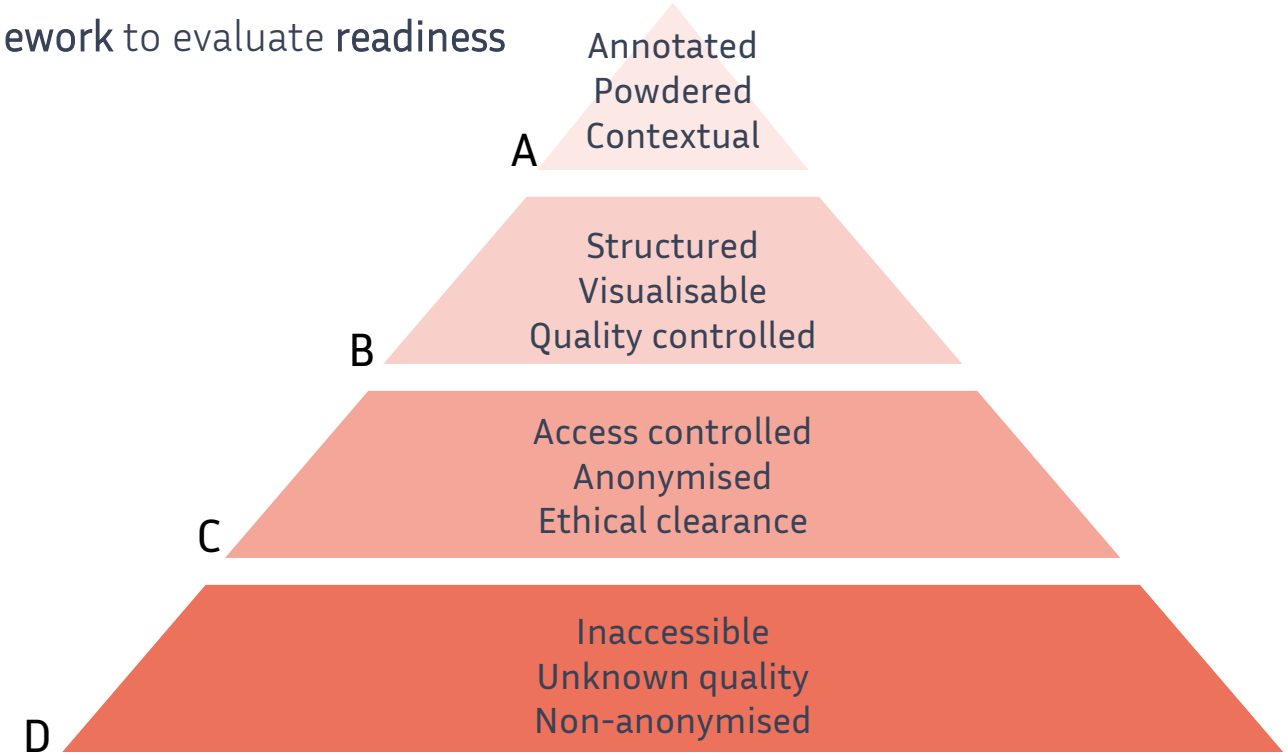
- Are data **ready** for a DL process?
- Consider using a **framework** to evaluate **readiness**



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***

$$P(Y \mid X)$$

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 ⟶ ***Y effect***

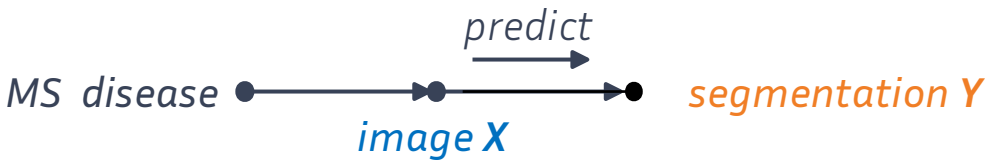
causal

predict effect from cause

Example: MS Lesion segmentation

X - structural brain MRI scan

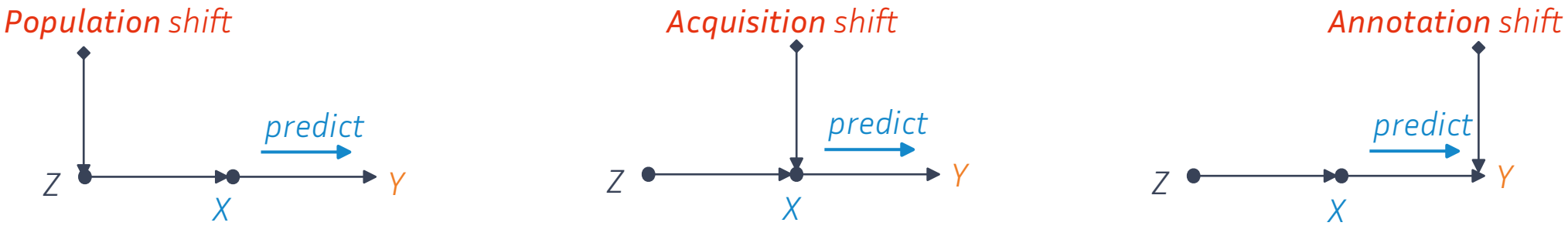
Y - manual segmentation masks



Castro, D. C., Walker, I., & Glocker, B. (2020). *Causality matters in medical imaging.* Nature Communications, 11(1), 3673.

Dataset Shift – *data mismatch*

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

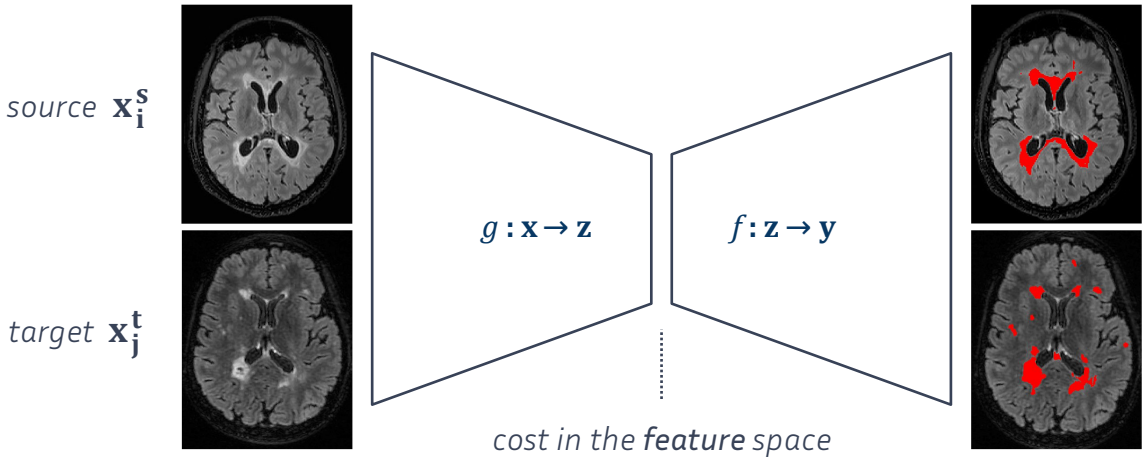


Z : unobserved true anatomy *X* : image *Y* : annotation

Castro, D. C., Walker, I., & Glocker, B. (2020). *Causality matters in medical imaging.* Nature Communications, 11(1), 3673.

Domain Adaptation – *a case study*

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



cost in the **source label** *space*

$$\mathbf{L_s}\left(\mathbf{y_i^s}, \mathbf{f}\left(\mathbf{g}\left(\mathbf{x_i^s}\right)\right)\right)$$

cost in the **target label** *space*

$$\mathbf{L_t}\left(\mathbf{y_i^s}, \mathbf{f}\left(\mathbf{g}\left(\mathbf{x_j^t}\right)\right)\right)$$

cost in the **feature** *space*

$$\left|\left|\mathbf{g}\left(\mathbf{x_i^s}\right) - \mathbf{g}\left(\mathbf{x_j^t}\right)\right|\right|^2$$

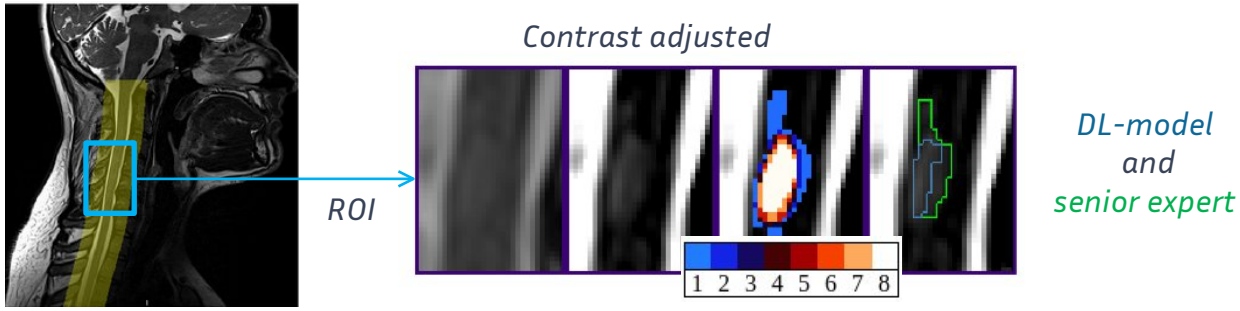
- Optimization problem:

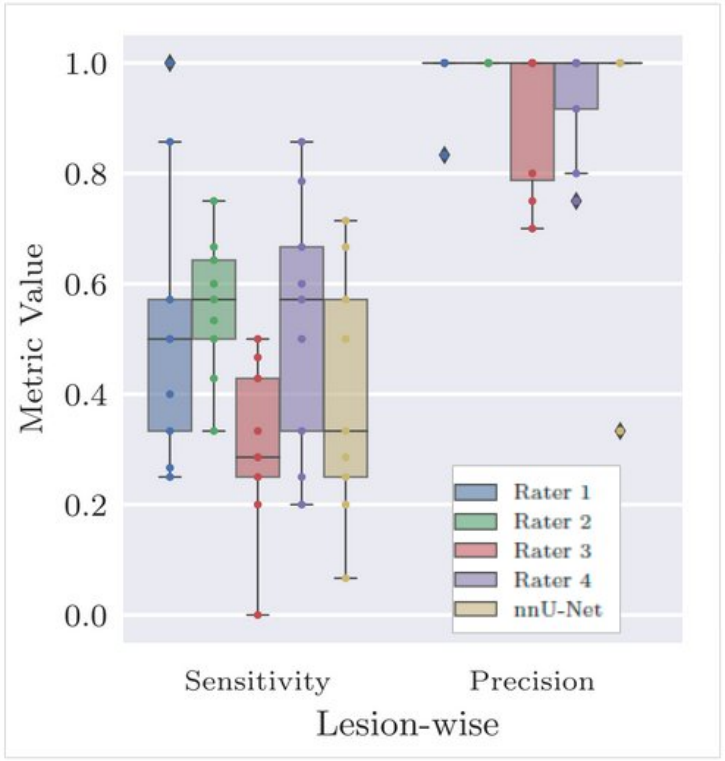
$$\min_{\gamma,f,g} \sum_i \mathbf{L_s}\left(\mathbf{y_i^s}, \mathbf{f}\left(\mathbf{g}\left(\mathbf{x_i^s}\right)\right)\right) + \sum_{i,j} \gamma_{i,j} \left(\alpha \left|\left|\mathbf{g}\left(\mathbf{x_i^s}\right) - \mathbf{g}\left(\mathbf{x_j^t}\right)\right|\right|^2 + \beta \mathbf{L_t}\left(\mathbf{y_i^s}, \mathbf{f}\left(\mathbf{g}\left(\mathbf{x_j^t}\right)\right)\right)\right)$$

the coupling matrix

Ackaouy, A., Courty, N., Vallée, E., Commowick, O., Barillot, C., & Galassi, F. (2020). *Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from MRI data*. Frontiers in computational neuroscience, 14, 19.

Annotation Shift – *a case study*





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

Walsh, Ricky, et al. *Expert Variability and Deep Learning Performance in Spinal Cord Lesion Segmentation for Multiple Sclerosis Patients*. IEEE CBMS 2023

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