MATTER

Domain adaptation from theory to practice

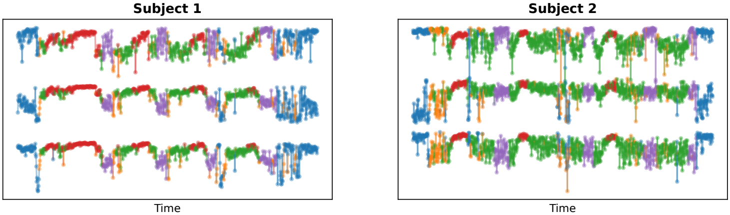
PI : **Rémi Flamary**, Professor, École Polytechnique

ERC-2022-COG Interview, November 15, 2022

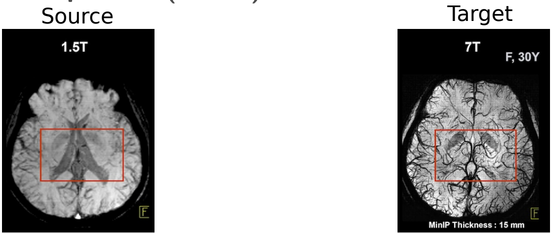
Adapation in Machine Learning and AI

Data shift in real life

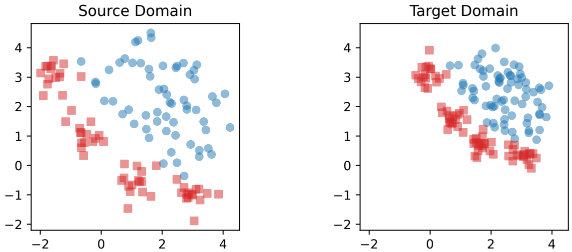
Shift between subjects and sessions (Sleep Stage Classification)



Shift in acquisition process (sensor)



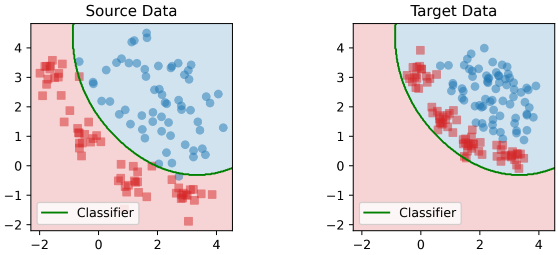
Data shift for prediction



Data shift in supervised learning

- $\mathcal{P}^s$ is the training distribution (Source domain)
- $\mathcal{P}^t$ is the test distribution (Target domain)
- **Data shift:** $\mathcal{P}^s \neq \mathcal{P}^t$
- $\Rightarrow$ A predictor f trained on source $\mathcal{P}^s$ won't work well on target $\mathcal{P}^t$.

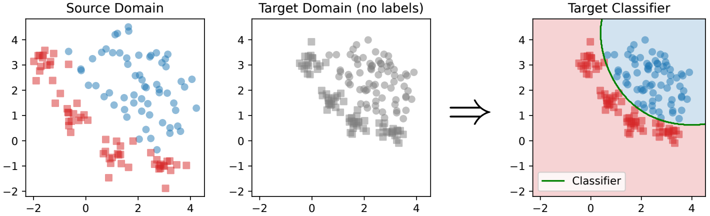
Data shift for prediction



Data shift in supervised learning

- $\mathcal{P}^s$ is the training distribution (Source domain)
- $\mathcal{P}^t$ is the test distribution (Target domain)
- **Data shift:** $\mathcal{P}^s \neq \mathcal{P}^t$
- $\Rightarrow$ A predictor f trained on source $\mathcal{P}^s$ won't work well on target $\mathcal{P}^t$.

Domain Adaptation (DA)



The domain adaptation challenge

- Domain adaptation data:
  - $\{\mathbf{x}_i^s, y_i^s\}$ source labeled data
  - $\{\mathbf{x}_j^t\}$ unlabeled target data (no labels available on target)
- Objective: learn a label predictor f **on target** using data above.
- Many existing approaches : re-weighting, invariant features, mapping.

Why isn't DA used all the time?

Key bottlenecks for DA

Trust and practical validation for DA (WP1)

- Classical generalization bounds and impossibility theorems:

$$\underbrace{\Delta^-(\mathcal{P}^s,\mathcal{P}^t)}_{\text{Imp. thm (if }\nearrow)} \leq \underbrace{E_{\mathbf{x},y\sim\mathcal{P}^t}[L(f(\mathbf{x}),y)]}_{\text{Error on target}} \leq \underbrace{E_{\mathbf{x},y\sim\mathcal{P}^s}[L(f(\mathbf{x}),y)]+\Delta^+(\mathcal{P}^s,\mathcal{P}^t)}_{\text{Generalization bound (if }\searrow)}$$

- Generalization bounds are pessimistic (can diverge with the shift).
- Validation for DA still open question (hard because no target labels).

Data and problems are evolving (WP2 + WP3)

- Multiple domains (with or without labels).
- Heterogeneous datasets or structures/architectures.

Lack of software and Benchmark (WP4)

- No reference software implementation for most DA methods.
- Lack of benchmark on different type of data (not only CV).

Project MATTER organization

WP1. Bridging theory and practical DA

- Shift estimation and performance guarantees
- Practical validation and Automatic DA

WP2. Multiple Domain Adaptation

- Multiple shifts with domain manifold learning
- Online and distributed MDA optimization
- Flagship challenge : Sleep Stage Classification

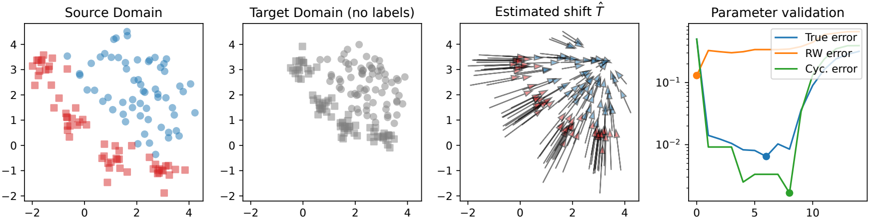
WP3. Learning across spaces and structures

- Heterogeneous DA with Optimal Transport
- Learning across graphs and neural architectures

WP 4. Repr. research and dissemination

- SKADA Toolbox and benchmark

WP 1. Bridging theory and practical DA



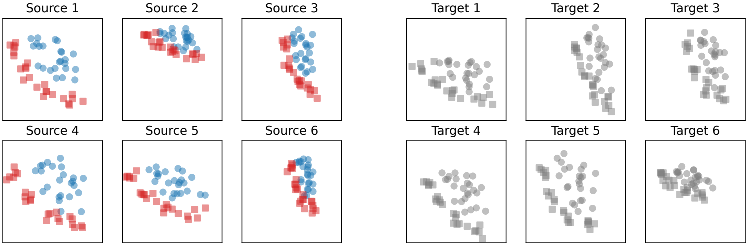
Task 1: Possibility theorems for DA with shift modeling

- Possibility theorems for DA with data shift modeling ($\mathcal{P}^t \approx T\#\mathcal{P}^s$).
- DA with partially observed Schrödinger bridge on joint feature/label space.
- Convergence and generalization for DA with Wasserstein gradient flow.

Task 2: Road to Automatic DA (AutoDA)

- Validation for DA : Cyclical/circular and performance score.
- Leave-One-Domain-Out validation for multi-source DA.
- Flagship objective: Automatic Domain Adaptation.

WP 2. Multiple Domain Adaptation (Multi-DA)



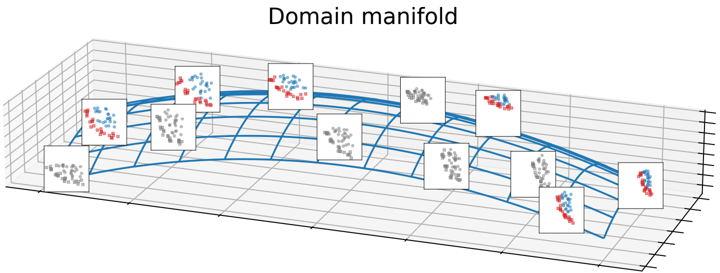
Task 1: Learning Domain Adaptation manifolds

- Learning manifolds on labeled and unlabeled domains with OT geodesics.
- Multi-DA as a semi-supervised meta learning problem.

Task 2: Online and Distributed MDA

- Online and incremental update for Multi-DA.
- Solving MDA with federated learning.

WP 2. Multiple Domain Adaptation (Multi-DA)



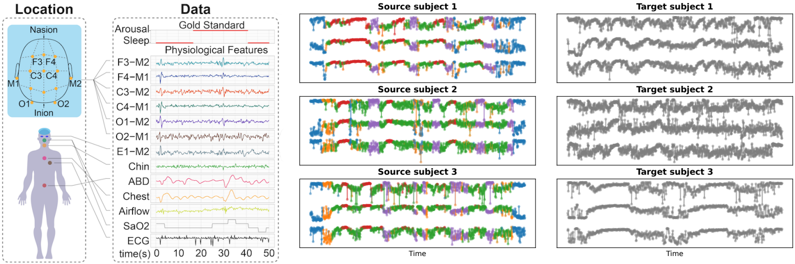
Task 1: Learning Domain Adaptation manifolds

- Learning manifolds on labeled and unlabeled domains with OT geodesics.
- Multi-DA as a semi-supervised meta learning problem.

Task 2: Online and Distributed MDA

- Online and incremental update for Multi-DA.
- Solving MDA with federated learning.

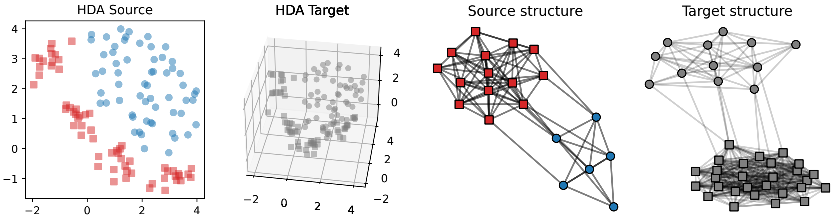
WP 2. Multiple Domain Adaptation (Multi-DA)



Task 3 Flagship application : Sleep Stage Classification

- Adaptive and interpretable sleep stage classification of EEG.
- Privacy preserving federated DA for sleep stage classification.
- With partners at INRIA Saclay expert on this type of data.

WP 3. Learning across spaces and structures



Task 1: Heterogeneous Domain Adaptation with Optimal Transport (OT)

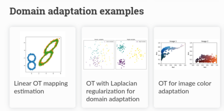
- HDA with feature and sample selection using Unbalanced OT.
- OT across spaces : convex formulation, theory and numerical optimization.
- Multi-source HDA modeling with generalized OT barycenters.

Task 2: Learning across spaces and structures with OT

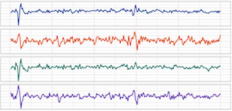
- Graph representation and attention models with GNN.
- Transfer learning and pre-training across architectures.

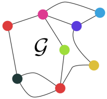
WP 4. Reproducible research and dissemination

Open source toolbox



Open DA benchmark





Dissemination

Task 1 : Open source, reproducible research and dissemination

- Open source Python Domain Adaptation toolbox.
- Open benchmark with realistic validation and multiple data types.
- Dissemination, education, documentation, tutorials and workshops.
- Full time engineer during the project.

Ability to deliver

PI: Rémi Flamary, Professor, École Polytechnique

- 4000+ citations ($\approx$ 2000 in 2021-2022), h-index 29.
- Creator and maintainer of POT Python Optimal Transport:
  - 800k downloads, 370 citations.
  - Published in JMLR open source software track.

Project MATTER's team

- Recruited :
  - 3 PhD and 3 PostDoc (2 years) on WP[1-3].
  - 1 Engineer 5 years on WP4.
- Collaborators with expertise in
  - Statistics, statistical physics, optimization, federated learning.
  - Biomedical applications, sleep stage classification, benchmarks.

Project MATTER in a nutshell

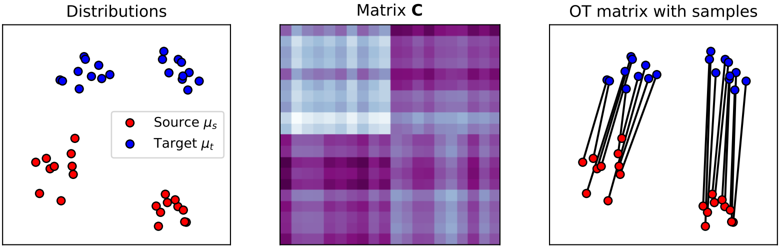
Domain adaptation from theory to practice

- Make DA a standard approach of machine learning.
- Model the data shifts for better theoretical performance and interpretability.
- Propose novel DA methods for modern learning (AutoDA, MDA, HDA).
- Provide easy-to-use open software and benchmark for DA.

Potential impact

- Adapting models without labels or with weak labels.
- Shift modeling for bias and algorithmic fairness.
- Dynamical systems, causal modeling and discovery,

Optimal transport between discrete distributions



OT Linear Program

When $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

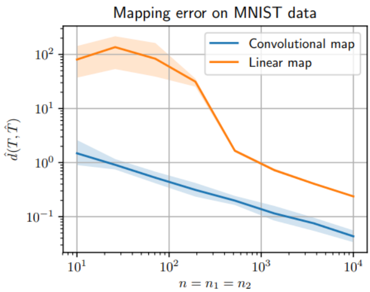
$$\mathbf{T}_0 = \underset{\mathbf{T} \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \quad \left\{ \langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

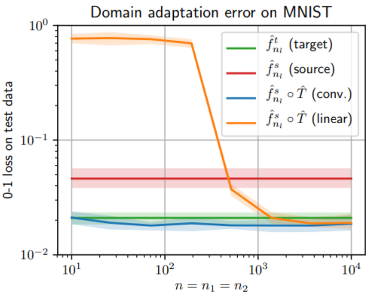
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

$$\Pi(\mu_s, \mu_t) = \left\{ \mathbf{T} \in (\mathbb{R}^+)^{n_s \times n_t} \middle| \, \mathbf{T}\mathbf{1}_{n_t} = \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_s} = \mathbf{b} \right\}$$

Linear program with $n_s n_t$ variables and $n_s + n_t$ constraints.

Possibility theorems for DA





Preliminary results for mapping estimation [Flamary et al., 2021]

$$E_{\mathbf{x},y\sim\mathcal{P}^t}[L(f \circ \hat{m}^{-1}(\mathbf{x}), y)] \leq E_{\mathbf{x},y\sim\mathcal{P}^s}[L(f(\mathbf{x}), y)] + Kd(m, \hat{m}) \qquad (1)$$

- Target error converges for adapted predictor converges to source error.
- Importance of the estimation of the mapping (d is divergence).
- Extension to other mappings (supervised mapping estimation, stochastic entropic OT)

Reality check for DA

Unsupervised Domain Adaptation: A Reality Check

Kevin Musgrave, Cornell Tech
Serge Belongie, University of Copenhagen
Ser-Nam Lim, Meta AI

Paper : [Musgrave et al., 2021]

- Meta Analysis from papers: Performance gain, Validation procedure.
- Comparison of numerous DA methods with realistic validation (several DA CV scores compared).
- Comparison between reproduced performance (with proper validation) and from paper.

Multiple DA

Domain manifold learning

$$\min_{\mathfrak{M}} \quad \sum_i W(\hat{\mathcal{P}}_i^s, P_{\mathfrak{M}}(\hat{\mathcal{P}}_i^s)) + \sum_j \min_{f_j} W(\hat{\mathcal{P}}_{j,f_j}^t, P_{\mathfrak{M}}(\hat{\mathcal{P}}_{j,f_j}^t)), \tag{2}$$

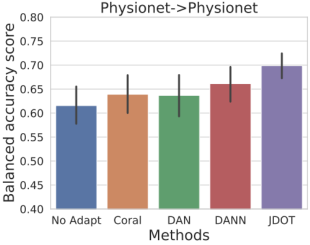
- $\mathfrak{M}$ is the manifold of joint distributions and W the Wasserstein distance.
- $P_{\mathfrak{M}}(\mu) = \arg\min_{\nu \in \mathfrak{M}} W(\mu, \nu)$ is a projection operator on the manifold.
- f_j are the prediction functions on target domains.
- Simplified proof of concept in [Turrisi et al., 2022] with source reweighting.

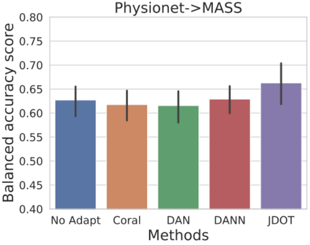
Meta learning for Multi-DA

- Training dataset (source domains) : $\{\hat{\mathcal{P}}_{i,\mathcal{X}}^s, \hat{f}_i^s\}_i$
- Target data (semi supervised) : $\{\hat{\mathcal{P}}_{j,\mathcal{X}}^t\}_j$
- Learned with structured output prediction and Hypernetwork models.

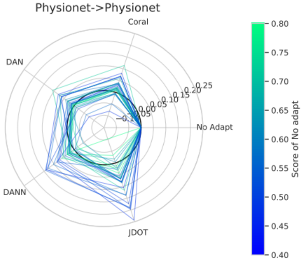
Sleep Stage Classification preliminary results

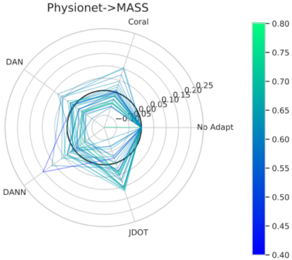
Performance of DA on sleep stage classification



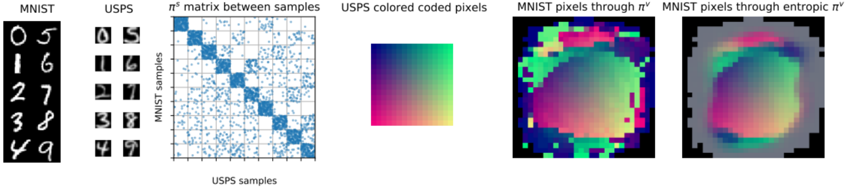


Performance of DA on sleep stage classification subjects





Optimal Transport across spaces



Co-Optimal Transport [Redko et al., 2020]

$$\min_{\substack{\boldsymbol{T}^s \in \Pi(\mathbf{w}, \mathbf{w}') \\ \boldsymbol{T}^v \in \Pi(\mathbf{v}, \mathbf{v}')}} \sum_{i,j,k,l} L(X_{i,k}, X'_{j,l}) \boldsymbol{T}^s_{i,j} \boldsymbol{T}^v_{k,l} \tag{3}$$

- $L(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}^+$ is the similarity measure.
- $\boldsymbol{T}^s$ is the OT matrix between samples, $\boldsymbol{T}^v$ is the OT matrix between features/variables.

References I

[Flamary et al., 2021] Flamary, R., Lounici, K., and Ferrari, A. (2021).
Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation.

[Musgrave et al., 2021] Musgrave, K., Belongie, S., and Lim, S.-N. (2021).
Unsupervised domain adaptation: A reality check.
arXiv preprint arXiv:2111.15672.

[Redko et al., 2020] Redko, I., Vayer, T., Flamary, R., and Courty, N. (2020).
Co-optimal transport.
In *Neural Information Processing Systems (NeurIPS)*.

[Turrisi et al., 2022] Turrisi, R., Flamary, R., Rakotomamonjy, A., and Pontil, M. (2022).
Multi-source domain adaptation via weighted joint distributions optimal transport.
In *Uncertainty in Artificial Intelligence*.