Multi-Source Domain Adaptation through Dataset Dictionary Learning in Wasserstein Space

Eduardo Montesuma [1] Fred Ngolè Mboula [1] Antoine Souloumiac [1]

[1]Université Paris-Saclay, CEA, List, F-91120 Palaiseau France

Abstract

We seek to solve **Multi-Source Domain Adaptation (MSDA)**, which aims to mitigate data distribution shifts when transferring knowledge from multiple labeled source domains to an unlabeled target domain. We propose a novel MSDA framework based on dictionary learning and optimal transport. We interpret each domain in MSDA as an empirical distribution. As such, we express each domain as a Wasserstein barycenter of dictionary atoms, which are empirical distributions. We propose a novel algorithm, **Dataset Dictionary Learning (DaDiL)**, for learning via mini-batches: (i) atom distributions; (ii) a matrix of barycentric coordinates. Based on our dictionary, we propose two novel methods for MSDA: **DaDiL-R**, based on the reconstruction of labeled samples in the target domain, and **DaDiL-E**, based on the ensembling of classifiers learned on atom distributions. We evaluate our methods in 3 benchmarks: **Caltech-Office**, **Refurbished-Office 31**, and **CRWU**, where we improved previous state-of-the-art by **3.15%**, **2.29%**, and **7.71%** in classification performance. Finally, we show that interpolations in the Wasserstein hull of learned atoms provide data that can generalize to the target domain.

Methodology

Wasserstein Barycenters of Labeled Distributions

When calculating Optimal Transport between labeled distributions, one needs to integrate labels in the ground-cost. Let $\mathbf{y}_i^{(P)} \in \Delta_{n_c}$ denote the soft-labels of sample $\mathbf{x}_i$. We use,

$$C_{i,j} = \|\mathbf{x}_i^{(P)} - \mathbf{x}_j^{(Q)}\|_2^2 + \beta\|\mathbf{y}_i^{(P)} - \mathbf{y}_j^{(Q)}\|_2^2, \quad (1)$$

where $\beta > 0$ controls the importance of label discrepancy. While simple, this choice allows us to motivate the barycentric projection of [1], and the label propagation of [2] as first-order optimality conditions of $W_c(\hat{P}, \hat{Q})$,

$$\begin{cases} \hat{\mathbf{x}}_i^{(P)} = T_\pi(\mathbf{x}_i^{(P)}) = n_P \sum_{j=1}^{n_Q} \pi_{i,j}\mathbf{x}_j^{(Q)}, \\ \hat{\mathbf{y}}_i^{(P)} = T_\pi(\mathbf{y}_i^{(P)}) = n_P \sum_{j=1}^{n_Q} \pi_{i,j}\mathbf{y}_j^{(Q)}. \end{cases} \quad (2)$$

As a consequence, we can interpolate between two point clouds, since $\hat{\mathbf{y}}_i^{(P)}$ corresponds to a soft-label (i.e., probabilities). We use equations 1 and 2 for proposing a new barycenter strategy between labeled point clouds, shown in algorithm 1.

Algorithm 1 Labeled Wasserstein Barycenter

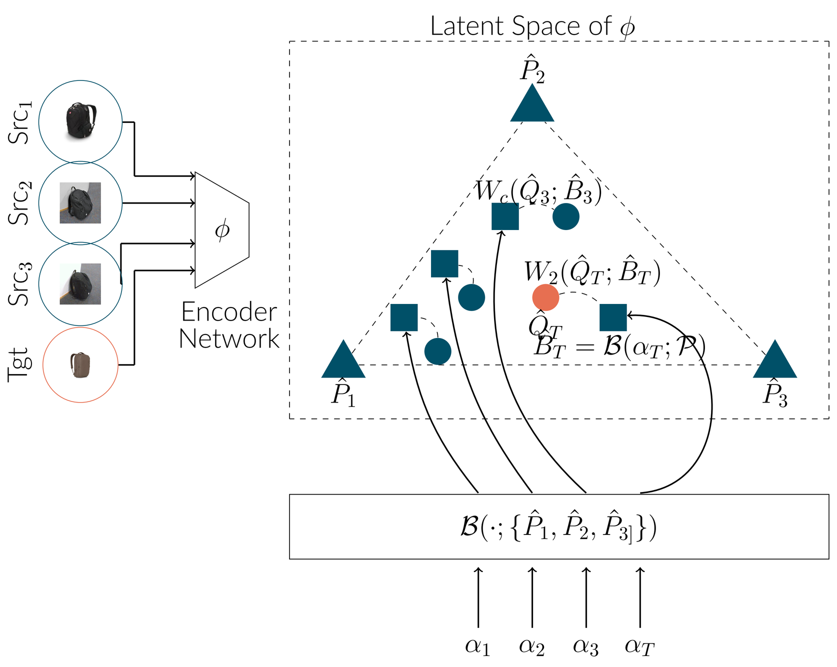
Input: $\{\mathbf{X}^{(P_k)}, \mathbf{Y}^{(P_k)}\}_{k=1}^K$, $\alpha \in \Delta_K$, $\tau > 0$, N_{itb}.
```
1:  for i = 1, ..., n_B do
2:      x_i^(B) ~ N(0, I_d), y_i^(B) = randint(n_c)
3:  end for
4:  while |J_it - J_it-1| >= tau and it <= N_itb do
5:      for k = 1, ... K do
6:          pi^(k,it) = OT((X^(P_k), Y^(P_k)); (X_it^(B), Y_it^(B)))
7:      end for
8:      J_it = sum_{k=1}^K alpha_k <pi^(k,it), C^(k)>_F
9:      X_{it+1}^(B) = sum_{k=1}^K alpha_k T_{pi^(k,it)}(X_it^(B))
10:     Y_{it+1}^(B) = sum_{k=1}^K alpha_k T_{pi^(k,it)}(Y_it^(B))
11: end while
```
Output: Labeled barycenter support $(\mathbf{X}^{(B)}, \mathbf{Y}^{(B)})$.

Dataset Dictionary Learning (DaDiL)

Let $\mathcal{Q} = \{\hat{Q}_{S_\ell}\}_{\ell=1}^{N_S} \cup \{\hat{Q}_T\}$ correspond to N_S labeled sources and an unlabeled target. Let $\mathcal{A} = [\alpha_1, \cdots, \alpha_{N_S}, \alpha_{N_S+1}]$, and $\mathcal{P} = \{\hat{P}_k\}_{k=1}^K$. The $\hat{P}_k$'s are an empirical approximation of the point clouds that interpolate distributional shift and $\alpha_T := \alpha_{N_S+1}$. For $N = N_S + 1$, DaDiL consists on minimizing,

$$(\mathcal{P}^\star, \mathcal{A}^\star) = \underset{\mathcal{P}, \mathcal{A} \in (\Delta_K)^N}{\text{argmin}} \frac{1}{N}\sum_{\ell=1}^{N} \mathcal{L}(\hat{Q}_\ell, \mathcal{B}(\alpha_\ell; \mathcal{P})),$$

where, $\mathcal{L}(\hat{Q}_\ell, \hat{B}_\ell) = W_c(\hat{Q}_\ell, \hat{B}_\ell)$ for the sources, and $\mathcal{L}(\hat{Q}_T, \hat{B}_T) = W_2(\hat{Q}_T, \hat{B}_T)$, for the target.



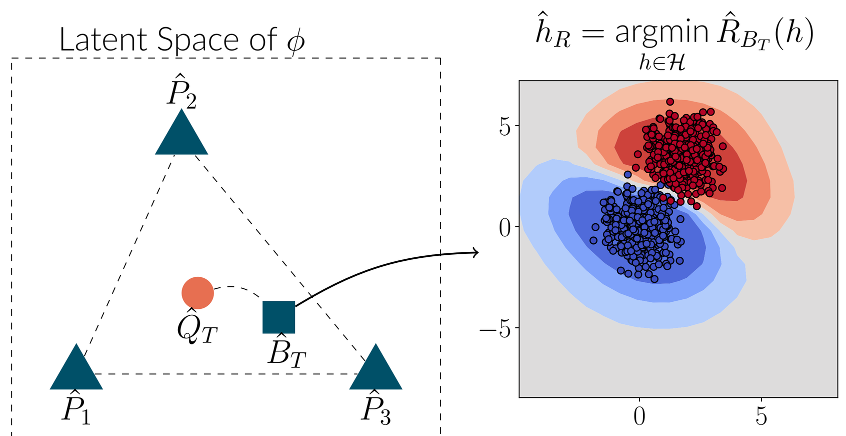
Multi-Source Domain Adaptation Strategies

DaDiL-Reconstruction. Relies on the reconstruction of distributions through Wasserstein barycenters.

$$\mathbf{X}^{(B_T)} = \sum_{k=1}^{K} \alpha_{T,k}\pi^{(k)}\mathbf{X}^{(P_k)}$$

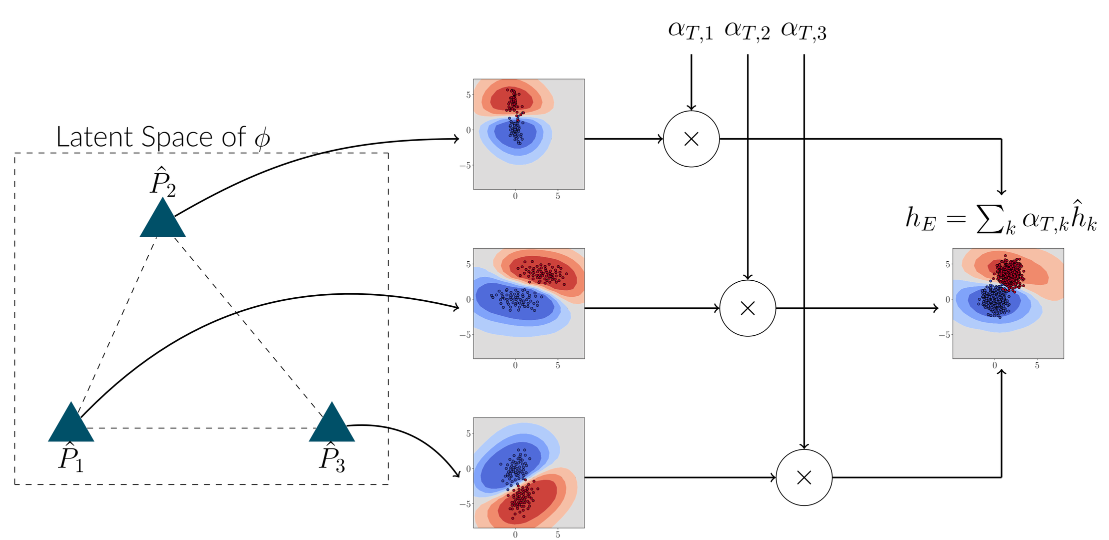
$$\mathbf{Y}^{(B_T)} = \sum_{k=1}^{K} \alpha_{T,k}\pi^{(k)}\mathbf{Y}^{(P_k)}$$

$$\hat{h}_R = \underset{h \in \mathcal{H}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h(\mathbf{x}_i^{(B_T)}), y_i^{(B_T)})$$



$$\mathcal{R}_{Q_T}(h) \leq \mathcal{R}_{B_T}(h) + \underbrace{W_2(\hat{Q}_T, \hat{B}_T)}_{\text{Reconstruction Error}} + \underbrace{\sqrt{2(\log 1/\delta)/\xi'}\left(\sqrt{1/n_P} + \sqrt{1/n_Q}\right)}_{\text{Sample Complexity } \mathcal{O}(n^{-1/2})} + \underbrace{\min_{h \in \mathcal{H}} \mathcal{R}_Q(h) + \mathcal{R}_P(h)}_{\text{Adaptation Complexity}},$$

DaDiL-Ensembling: relies on the ensembling of classifiers fit on atom data.

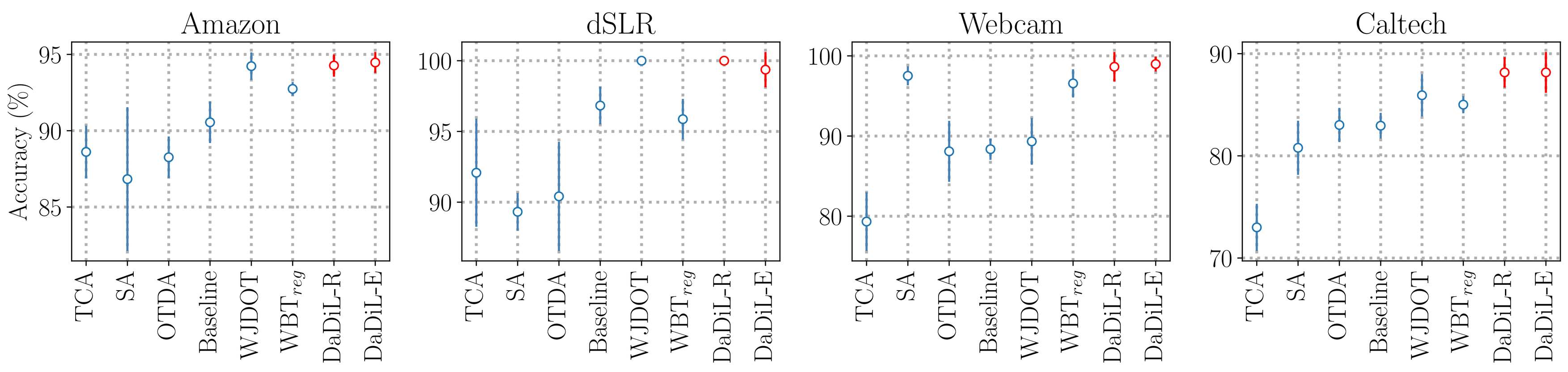


$$\mathcal{R}_{Q_T}(\hat{h}_\alpha) \leq \mathcal{R}_\alpha(\hat{h}_\alpha) + \underbrace{\mathrm{W}_2(\mathcal{B}(\alpha; \mathcal{P}), \hat{Q}_T)}_{\text{Reconstruction Error}} + \underbrace{\sum_{k=1}^{K} \alpha_k \mathrm{W}_2(\hat{P}_k, \mathcal{B}(\alpha; \mathcal{P}))}_{\text{Dictionary Geometry}}$$

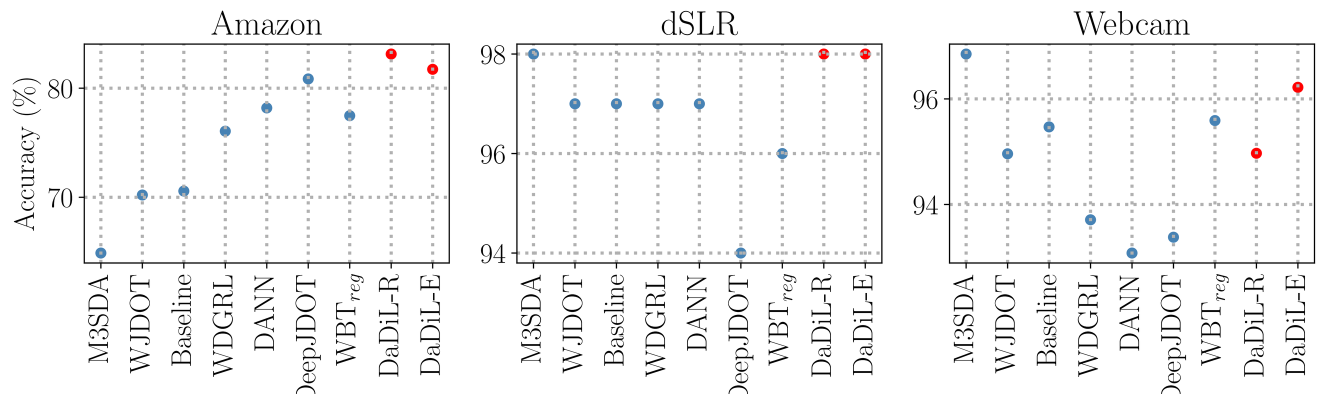
$$+ \underbrace{\sum_{k=1}^{K} \alpha_k \sqrt{2\log 1/\delta/\xi'}\left(\sqrt{1/n_k} + \sqrt{1/n_T}\right)}_{\text{Sample Complexity}} + \underbrace{\sum_{k=1}^{K} \alpha_k \left(\min_{h \in \mathcal{H}} \mathcal{R}_{P_k}(h) + \mathcal{R}_{Q_T}(h)\right)}_{\text{Adaptation Complexity}},$$

Empirical Results

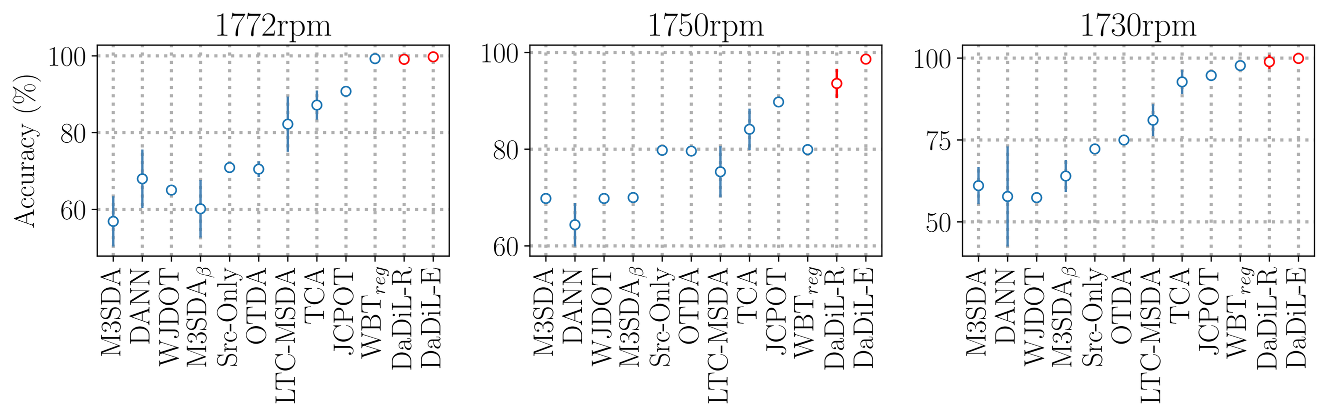
Caltech-Office 10



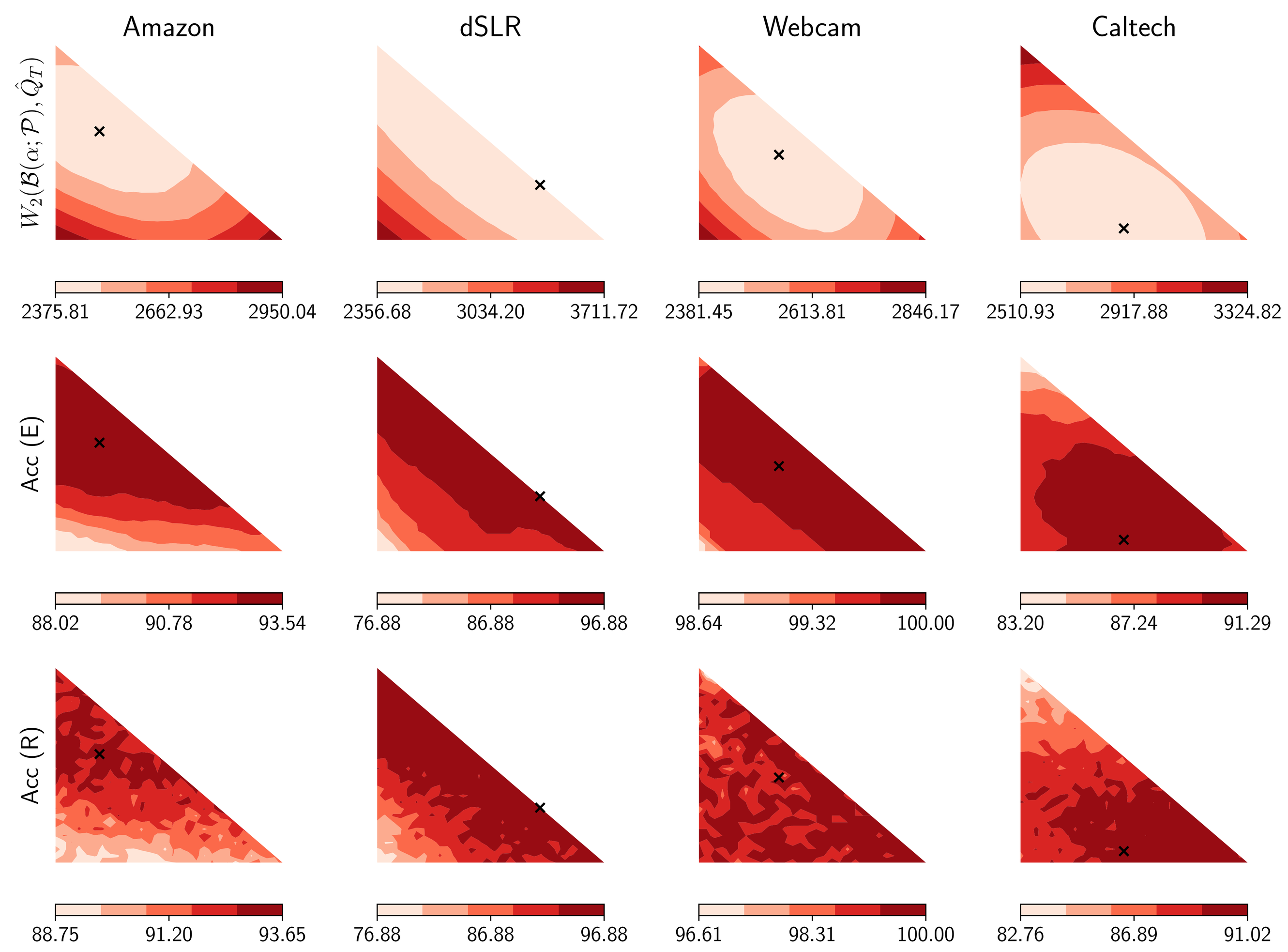
Refurbished Office 31



CWRU



Atom Interpolations



Conclusions

- We propose a **novel dictionary learning** method, called **DaDiL**
- **DaDiL learns to model distributional shift** between distributions
- DaDiL has **state-of-the-art performance** on various **domain adaptation** benchmarks.
- Besides optimal choices given by our algorithm, **DaDiL** defines a rich **interpolation space** between atoms.

Future Works

Federated Learning [4] Dataset Distillation [5] Cross-Domain Fault Diagnosis [6].

References

[1] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 39(9):1853–1865, 2016.

[2] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Bennani. Advances in domain adaptation theory. Elsevier, 2019.

[3] Eduardo Fernandes Montesuma and Fred Ngole Mboula. Wasserstein barycenter for multi-source domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) June 2021.

[4] Eduardo Fernandes Montesuma, Michela Mulas, Fred Ngolè Mboula, Francesco Corona, and Antoine Souloumiac. Multi-Source Domain Adaptation for Cross-Domain Fault Diagnosis of Chemical Processes. arXiv:2308.11247. August 2023

[5] Eduardo Fernandes Montesuma, Michela Mulas, Fred Ngolè Mboula, Francesco Corona, and Antoine Souloumiac. Multi-Source Domain Adaptation for Cross-Domain Fault Diagnosis of Chemical Processes. arXiv:2308.11247. August 2023

[6] Eduardo Fernandes Montesuma, Michela Mulas, Fred Ngolè Mboula, Francesco Corona, and Antoine Souloumiac. Multi-Source Domain Adaptation for Cross-Domain Fault Diagnosis of Chemical Processes. arXiv:2308.11247. August 2023

Learn more about DaDiL!

DaDiL Paper DaDiL Demo Portfolio