Multitemporal classification without new labels: a solution with optimal transport

D. Tuia[1], R. Flamary[2], A. Rakotomamonjy[3], N. Courty[4]

[1] Uni. Zurich, [2] Uni. Nice Sophia Antipolis
[3] Uni. Rouen, [4] University of South Brittany

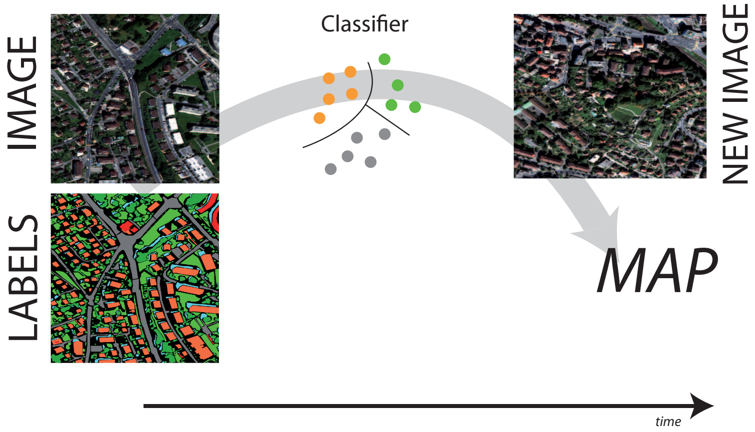
Multitemp 2015 – Annecy

Multitemporal classification w/o new labels

- Revisit time of satellite is shortening
- Frequency of terrestrial campaigns (or photointerpretation) stalks
- We cannot keep the pace of the images!

- We want to
  - adapt classifiers to new acquisitions
  - take advantage of available labeled examples

The problem



The reason

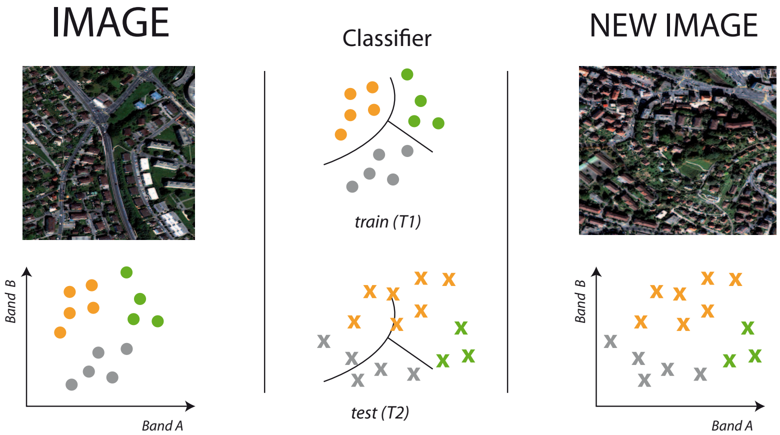
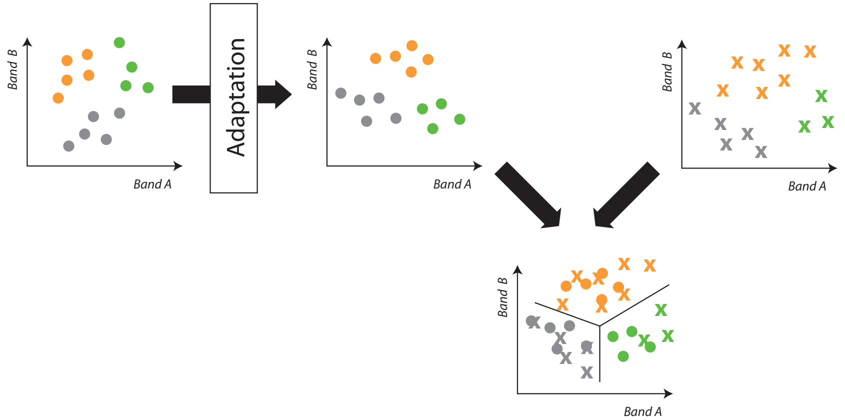


IMAGE
Classifier
NEW IMAGE
train (T1)
test (T2)
Band B
Band A
Band B
Band A

How? Adapting the data representation to match the input spaces



How? Optimal transport

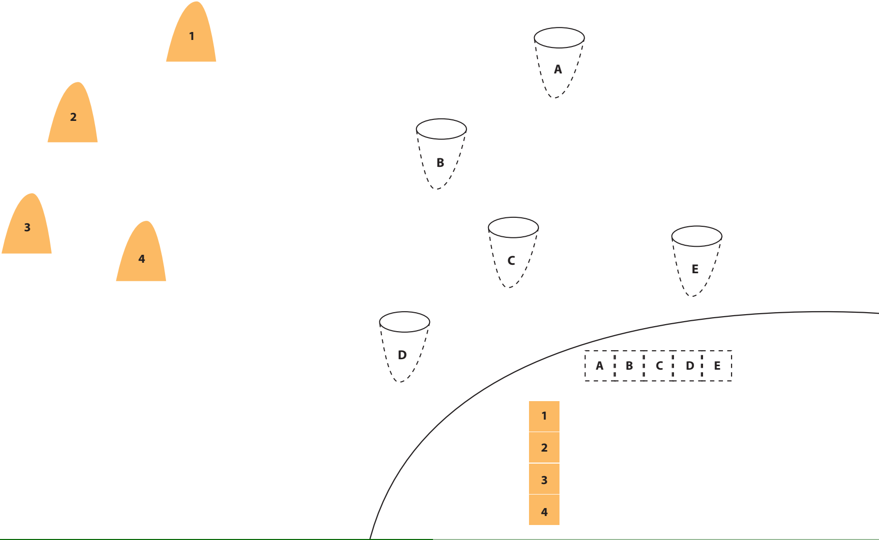
We propose

1. A framework to match data distributions
   $\rightarrow$ `optimal transport (OT)`

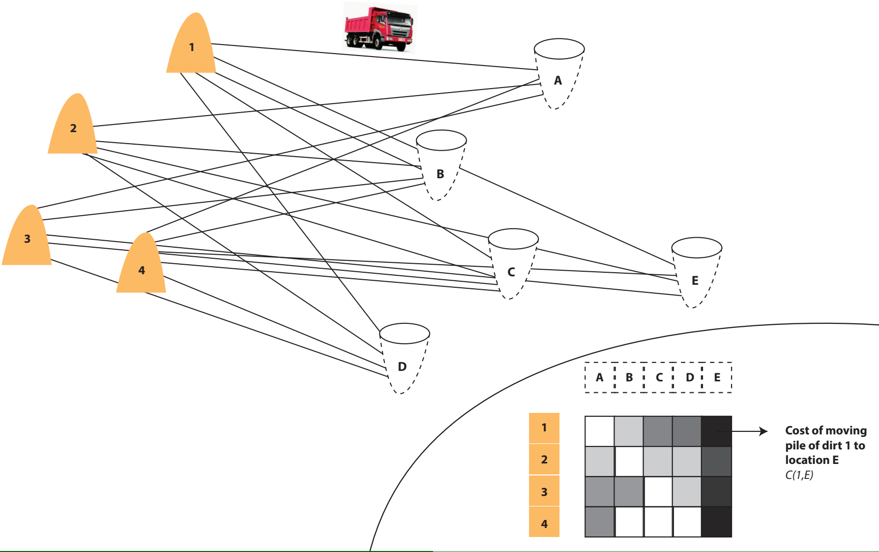
   - No coregistration required
   - Can cope with strong deformations [Courty et al., ECML, 2014]

2. A regularized version using the class labels in the source domain
   $\rightarrow$ `semi-supervised OT without new samples.`

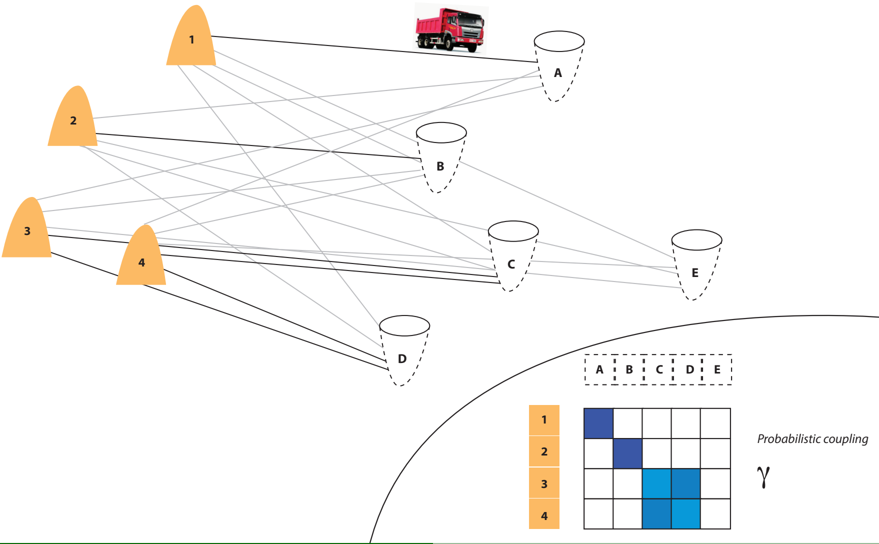
19*th* century: moving dirt in an optimal way



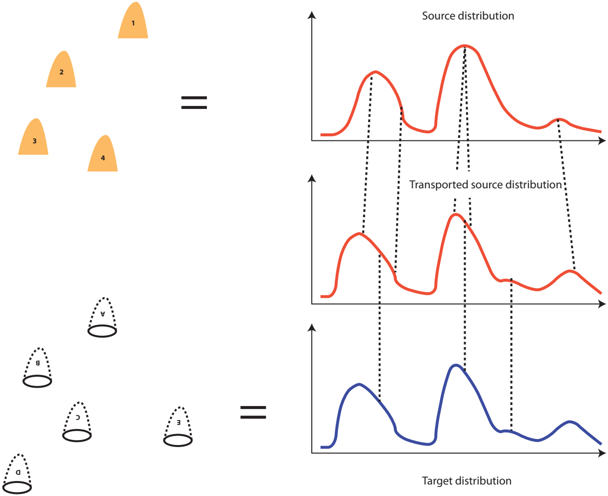
19*th* century: moving dirt in an optimal way



19*th* century: moving dirt in an optimal way



20*th* century: using OT to match distributions



Optimal transport [Kantorovich, 1947]

- All the potential transport matrices $\gamma \in \mathcal{P}$ respect the two marginals

  $\rightarrow$ `no loss of masses`

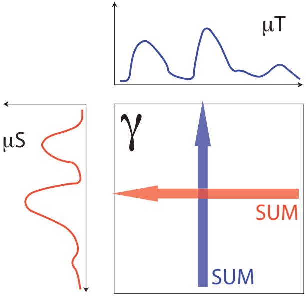
- The optimal one γ_0 is the one minimizing the transportation cost wrt the metric $\mathbf{C}$

  $\rightarrow$ `favors minimal effort`

  $\rightarrow$ `favors sparsity`

- We look for the optimal coupling:

$$\gamma_0 = \underset{\gamma \in \mathcal{P}}{\arg\min}\, \langle \gamma, \mathbf{C} \rangle_F$$



What does that mean?

Once γ_0 is found, we can

- transform the source elements $\mathbf{X}_s$ in a target domain-dependent version $\hat{\mathbf{X}}_s$,
- train any classifier with $\{\hat{\mathbf{X}}_s, Y_s\}$ to predict classes in the target domain.

This way, we can do domain adaptation!

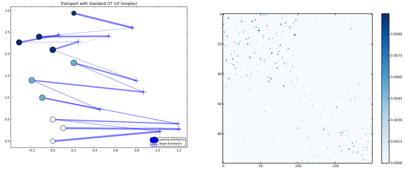
Regularization by entropy

- Maybe this is too sparse
- A wrong assignment it is not recoverable

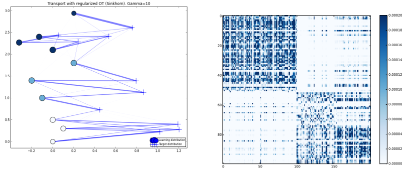
- Solution: de-sparsify

$$\gamma_0^\lambda = \underset{\gamma \in \mathcal{P}}{\arg\min}\, \langle \gamma, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\gamma)$$

where $h(\gamma)$ is the entropy of the γ matrix.



Non-regularized



Entropy-regularized

[Cuturi, NIPS 2013]

Regularization by class

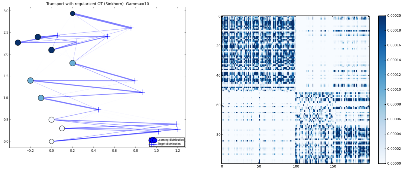
- Enforce that labeled samples of the same class move similarly

- Solution: group-sparse regulariz.

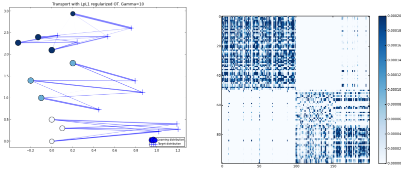
$$\gamma_0^{\lambda_c} = \underset{\gamma \in \mathcal{P}}{\arg\min} \langle \gamma, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\gamma) + \eta \sum_j \sum_c \|\gamma(\mathcal{I}_c, j)\|_q^p,$$

each column of γ is nonzero only for samples of the same (label) class.

[Courty, Flamary, Tuia, ECML 2014]

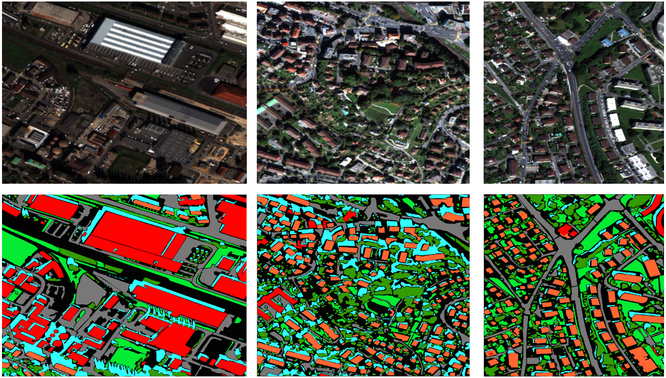


Entropy-regularized



Class-regularized

Data



Malley (`Ma`) (1124 × 1516) 29 . 09 . 2010

Montelly (`Mo`) (1064 × 1248) 29 . 09 . 2010

Prilly (`Pr`) (1040 × 1032) 02 . 08 . 2011

class legend: commercial buildings, residential buildings, meadows, trees, roads, shadows

Setup

- 6 adaptation problems (all couples of domains)

  ($\texttt{Mo} \rightarrow \texttt{Pr}$), ($\texttt{Pr} \rightarrow \texttt{Mo}$), ($\texttt{Pr} \rightarrow \texttt{Ma}$)

  ($\texttt{Ma} \rightarrow \texttt{Mo}$), ($\texttt{Mo} \rightarrow \texttt{Ma}$), ($\texttt{Ma} \rightarrow \texttt{Pr}$)
- 100 labels per class in the source domain
- 600 pixels in the target domain (unlabeled)
- Classification with 1-NN

- Compared to
  - **OT-Sink** vs. KPCA, Graph Matching[1], TCA[2]
  - **OT-labreg** vs. SSTCA[2]

[[1] Tuia et al. IEEE TGRS, 2013]

[[2] Pan et al., IEEE Trans. Knowl. Data Eng., 2010]

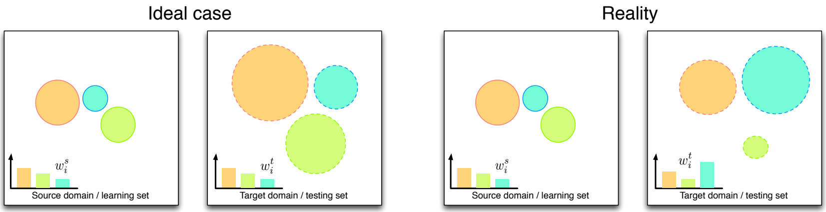
- Without labels in source

# labels X^S	No adapt. -	Training on X^S KPCA 0	TCA 0	GM 0	**OT-Sink** 0
Mo $\rightarrow$ Pr	59.67±2.92	56.21±3.09	49.83±2.22	60.30±1.66	66.81±1.38
Pr $\rightarrow$ Mo	57.12±4.66	53.82±4.25	50.52±2.19	62.07±1.45	71.95±0.97
Ma $\rightarrow$ Mo	45.62±1.79	46.46±3.13	47.47±1.95	43.92±1.70	59.88±1.12
Mo $\rightarrow$ Ma	33.74±1.69	32.61±3.08	31.52±2.50	37.14±2.31	50.75±2.67
Ma $\rightarrow$ Pr	46.84±1.21	45.12±1.41	46.02±1.47	43.87±1.58	57.47±2.33
Pr $\rightarrow$ Ma	26.40±5.27	23.71±4.16	21.49±4.14	38.49±1.89	49.60±1.98

- With labels in source

# labels X^S	No adapt. -	Training on X^S SSTCA 600	**OT-labreg** 600	Training on X^T -
Mo $\rightarrow$ Pr	59.67±2.92	55.88±6.89	65.09±0.82	84.17±0.39
Pr $\rightarrow$ Mo	57.12±4.66	57.36±5.75	72.37±1.08	81.43±0.83
Ma $\rightarrow$ Mo	45.62±1.79	49.27±4.47	70.66±1.98	82.12±0.74
Mo $\rightarrow$ Ma	33.74±1.69	39.58±2.57	55.41±2.38	80.73±0.40
Ma $\rightarrow$ Pr	46.84±1.21	46.12±1.55	66.60±2.69	83.74±0.37
Pr $\rightarrow$ Ma	26.40±5.27	31.41±3.23	54.90±1.47	80.79±0.59

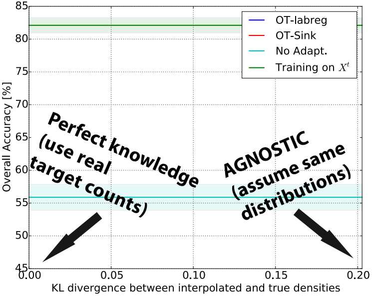
Impact of the changes in w^s vs. w^t ?



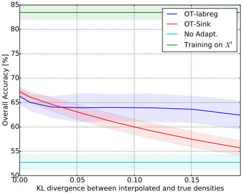
Impact of the changes in w^s vs. w^t ?

- Class distributions might change between domains
- We can either assume they are the same (Agnostic)

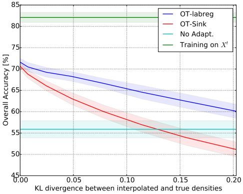
  OR
- Have a (+ / -) good guess of the differences
- Impact of the guess on the KL path



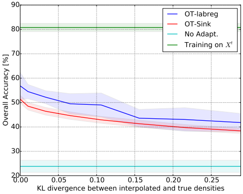
Impact of the changes in w^s vs. w^t ?



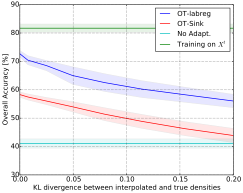
Mo → Pr



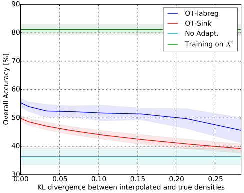
Pr → Mo



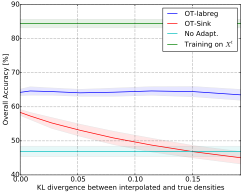
Pr → Ma



Ma → Mo



Mo → Ma



Ma → Pr

Summary

- ✓ Optimal transport is a probabilistic way to adaptation
- ✓ Regularization makes the difference
- ✓ OT can boost performance in very hard scenarios
- × Needs some knowledge of the target nature to be very effective

- And forward!
  a. Multidomain adaptation
  b. Automatic estimation of target class distribution

Thank you!

devis.tuia@geo.uzh.ch
devis.tuia.googlepages.com

... and to these great guys!

Remi

Alain

Nicolas

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION