Basics of machine learning

Nicolas Duchateau, Rémi Emonet, Carole Lartizien

Deep Learning for Medical Imaging

Lyon (FR) - 15/04/2019

Program

~30min	**1. Introduction** Carole Lartizien
~75min	**2. Supervised learning** Rémi Emonet + Carole Lartizien
~75min	**3. Unsupervised learning** Nicolas Duchateau + Rémi Emonet
~30min	**4. Methods evaluation** Carole Lartizien + Rémi Emonet + Nicolas Duchateau
~30min	**5. Conclusion**

---

4. Methods evaluation

Supervised learning in a nutshell

- **Split the sample dataset** into three parts : a training, a validation and a test dataset
- **Choose a parameterized model function** with parameters Θ_1 and hyperparameters Θ_2 from an hypothesis space H
- **Fit the model parameters Θ_1** to the training dataset for a fixed value of Θ_2
  - **Choose an error function** that measures the misfit between the decision function D(f(xi)) and the class yi of all training data points (xi, yi)
  - **Minimize the error function**
- **Evaluate the performance** of your model on the validation dataset
- Retrain your model with another hyperparameter set Θ_2
- Select the best parameter set
- **Evaluate the performance of your best model** on the test dataset

---

4. Methods evaluation

Supervised learning in a nutshell

- **Split the sample dataset** into three parts : a training, a validation and a test dataset
- **Choose a parameterized model function** with parameters Θ_1 and hyperparameters Θ_2 from an hypothesis space H
- **Fit the model parameters** Θ_1 to the training dataset for a fixed value of Θ_2
  - **Choose an error function** that measures the misfit between the decision function D(f(xi)) and the class yi of all training data points (xi, yi)
  - **Minimize the error function**
- **Evaluate the performance** of your model on the validation dataset
- Retrain your model with another hyperparameter set Θ_2
- Select the best parameter set
- **Evaluate the performance of your best model** on the test dataset

4. Methods evaluation

Supervised learning

Objective and definitions

- To estimate the intrinsic performance of a decision model
- **The empirical error** on the training dataset is optimistically biased
- **The generalization error** is the error achieved on new dataset representative of the data distribution P, ie data different from the training database.
- How to estimate the generalization error from samples drawn from the data distribution?

---

4. Methods evaluation

Supervised learning

Objective and definitions

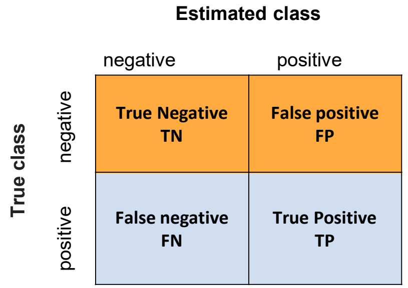
Performance evaluation on a dataset of N samples in a representation space of dimension d:

- ○ What performance metric?
- ○ What method to best approximate the generalization performance?

---

4. Methods evaluation

Metrics for the evaluation of supervised classification models



		Estimated class	
		negative	positive
True class	negative	**True Negative** **TN**	**False positive** **FP**
	positive	**False negative** **FN**	**True Positive** **TP**

Confusion matrix

---

4. Methods evaluation

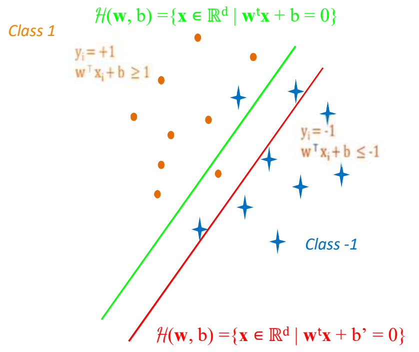
Metrics for the evaluation of supervised classification models

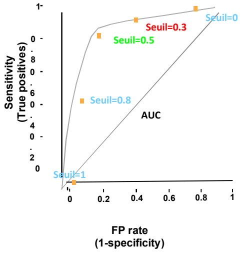
- **Sensitivity** = Fraction of true positives = TP/(TP+FN)
- **Specificity** = Fraction of true negatives = TN/(TN+FP)
- **Accuracy** = Fraction of correct answers =(TN+TP)/(TN+TP+FN+FP)

4. Methods evaluation

Metrics for the evaluation of supervised classification models

AUC = Area under the ROC curve



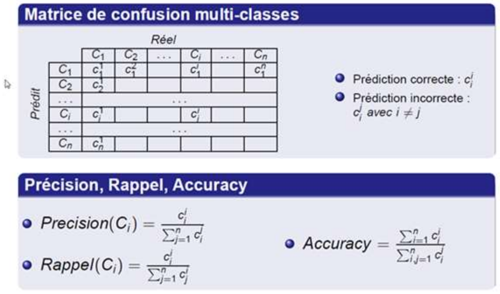


---

4. Methods evaluation

Supervised learning

Extension to the multi-class case



Matrice de confusion multi-classes

Réel

Prédit

	C_1	C_2	…	C_j	…	C_n
C_1	c_1^1	c_1^2		c_1^j		c_1^n
C_2	c_2^1					
…	…					
C_i	c_i^1			c_i^j		
…	…					
C_n	c_n^1					

- Prédiction correcte : c_i^i
- Prédiction incorrecte : c_i^j avec $i \neq j$

Précision, Rappel, Accuracy

- $Precision(C_i) = \frac{c_i^i}{\sum_{j=1}^n c_i^j}$
- $Rappel(C_i) = \frac{c_i^i}{\sum_{j=1}^n c_j^i}$
- $Accuracy = \frac{\sum_{i=1}^n c_i^i}{\sum_{i,j=1}^n c_i^j}$

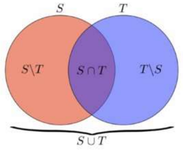
---

4. Methods evaluation

Supervised learning

Metrics for the evaluation of segmentation tasks

$$Dice(S,T) = \frac{2|S \cap T|}{|S| + |T|}$$



$$D_H(S,T) = \max\left\{\sup_{s \in S} \inf_{t \in T} d(s,t), \sup_{t \in T} \inf_{s \in S} d(s,t)\right\}$$

---

4. Methods evaluation

Supervised learning

Hold-out

- **Principle :**
    - Split the dataset into one learning and one validation dataset
    - Learn a model on the training dataset and evaluate its performance on the validation dataset
- **Advantages :**
    - Easy to implement
    - Low computational cost
- **Bottlenecks:**
    - Requires a lot of data
    - Sensitive to the database splitting
    - Robustness of the results

4. Methods evaluation

K-folds cross-validation

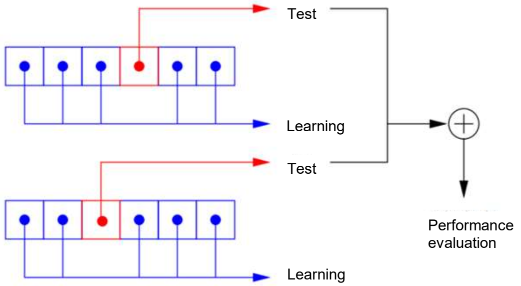
- Choose a number of folds k
- Split the dataset into k subsets $D_1, \dots, D_k$
- For each subset k:
    - Learn the model on the union of D_j subsets with j ≠ i
    - Evaluate the performances on fold D_i
    - Combine performances achieved on all folds D_j

if k = N (the number of samples) → **leave-one-out** (LOO)

---

4. Methods evaluation

Cross-validation



[source : Cours de F. Rossi, Telecom ParisTech]

---

4. Methods evaluation

Cross-validation

- Advantages :
    - Easy to implement
    - Make use of all available data to estimate the model performance
- Limits :
    - Sensitive to the splitting strategy
    - High computational time
    - Do not provide a model

---

4. Methods evaluation

Bootstrap

2. Draw one bootstrap dataset $\mathbf{x}^{*b}$ by randomling sampling N samples with replacement from the original dataset $\mathbf{x}$
3. Estimate the performance metric by hold-out on the sample $\mathbf{x}^{*b}$
4. Estimate the performance metric of the model trained $\mathbf{x}^{*b}$ and tested on $\mathbf{x}$
5. Compute an estimate of the bias of the hold-out method
6. Evaluate the performance metric by subtracting the estimated bias w from the hold-out performance estimated on the original dataset $\mathbf{x}$

```
for $b = 1$ to $B$ do
    $\mathbf{x}^{*b} \leftarrow \mathrm{bootstrap}(\mathbf{x})$
    $w^b \leftarrow AUC(\mathbf{x}^{*b}, \mathbf{x}^{*b}) - AUC(\mathbf{x}^{*b}, \mathbf{x})$
end for
$w \leftarrow \mathrm{moy}_{b=1..B}(w^b)$
$AUC_{\mathrm{ORD}} \leftarrow AUC(\mathbf{x}, \mathbf{x}) - w$
```

4. Methods evaluation

Bootstrap

- Advantages :
  - Easy to implement
  - Make use of all available data to estimate the model performance
  - Provide confidence intervals
- Limits :
  - High computational time
  - Do not provide a model

---

4. Methods evaluation

Unsupervised…

no “ground truth” labels = not so easy?

➔ Some pratical examples

(but highly depends on the **data + clinical question**)

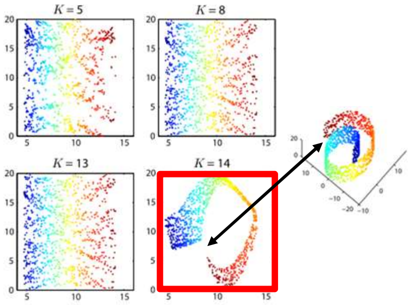
---

4. Methods evaluation

Unsupervised

1) Neighborhood graph: short-circuit?

Synthetic data = easy?



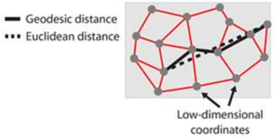
Estimated embeddings

Duchateau et al. *Med Image Anal* 2012

---

4. Methods evaluation

Unsupervised

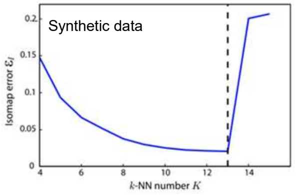


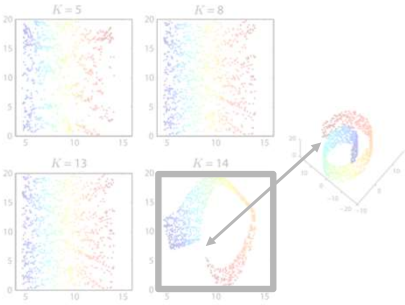
1) Neighborhood graph: short-circuit?

Synthetic data = easy?

$$\epsilon(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{d_{kNN}(\mathbf{x}_i, \mathbf{x}_j)}$$

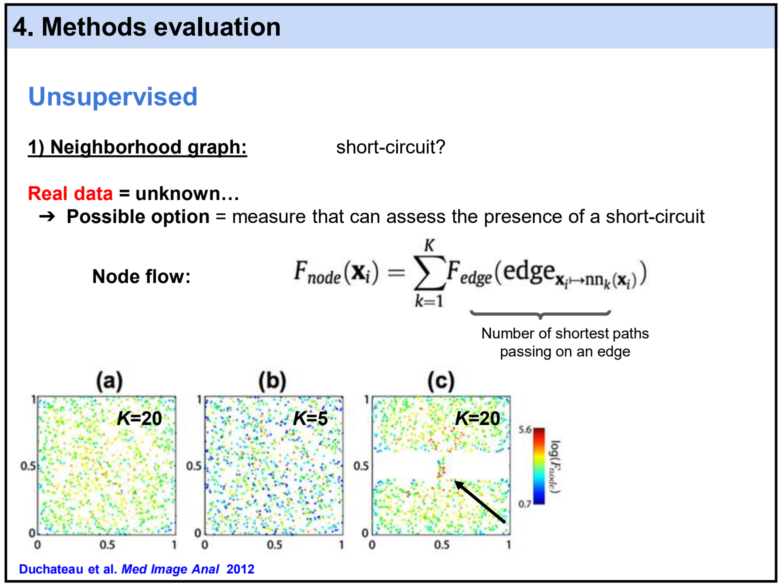
Error vs. geodesic distance





Estimated embeddings

Duchateau et al. *Med Image Anal* 2012



4. Methods evaluation

Unsupervised

1) Neighborhood graph: short-circuit?

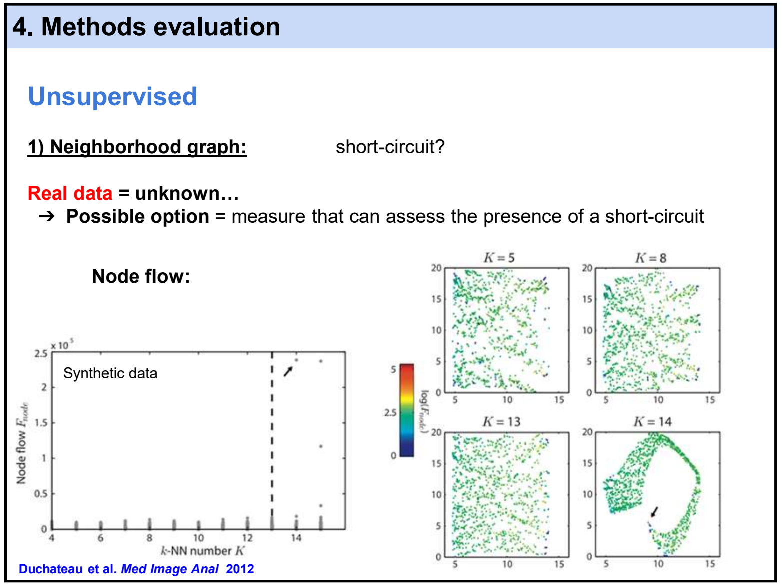
Real data = unknown…

➔ **Possible option** = measure that can assess the presence of a short-circuit

Node flow:

$$F_{node}(\mathbf{x}_i) = \sum_{k=1}^{K} \underbrace{F_{edge}(\text{edge}_{\mathbf{x}_i \mapsto \text{nn}_k(\mathbf{x}_i)})}_{\text{Number of shortest paths passing on an edge}}$$

Duchateau et al. *Med Image Anal* 2012



4. Methods evaluation

Unsupervised

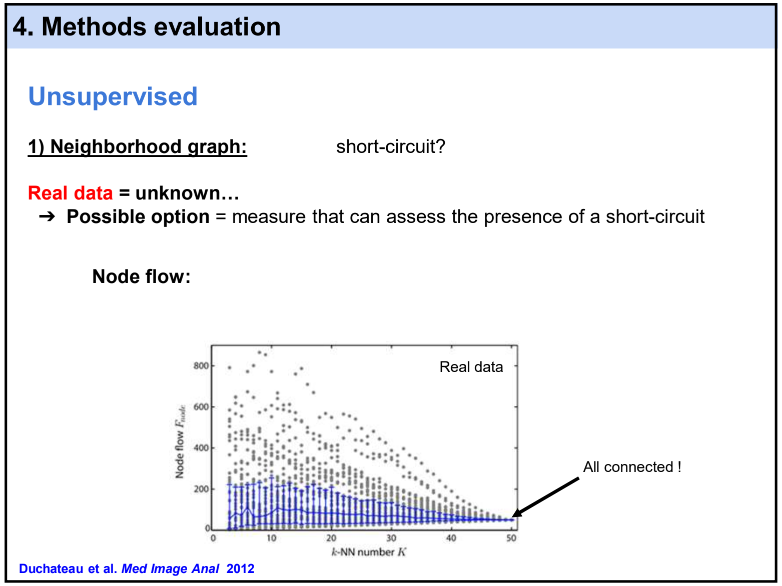
1) Neighborhood graph: short-circuit?

Real data = unknown…

➔ **Possible option** = measure that can assess the presence of a short-circuit

Node flow:

Duchateau et al. *Med Image Anal* 2012



4. Methods evaluation

Unsupervised

1) Neighborhood graph: short-circuit?

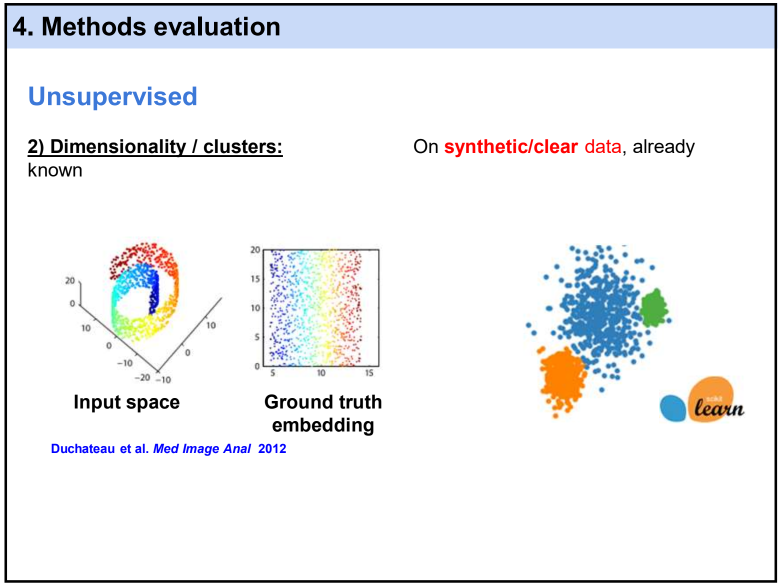
Real data = unknown…

➔ **Possible option** = measure that can assess the presence of a short-circuit

Node flow:

All connected !

Duchateau et al. *Med Image Anal* 2012



4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **synthetic/clear** data, already known

Input space

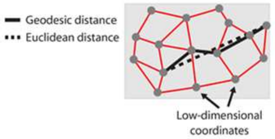
Ground truth embedding

Duchateau et al. *Med Image Anal* 2012

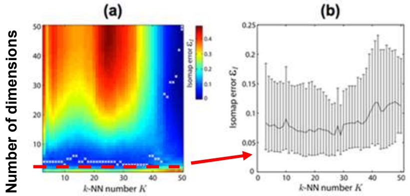
4. Methods evaluation

Unsupervised

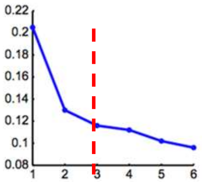
2) Dimensionality / clusters: On **real data = ?**



Theoretical error
ex: error vs. geodesic distance



Duchateau et al. *Med Image Anal* 2012



Gerber et al. *Med Image Anal* 2010

---

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**

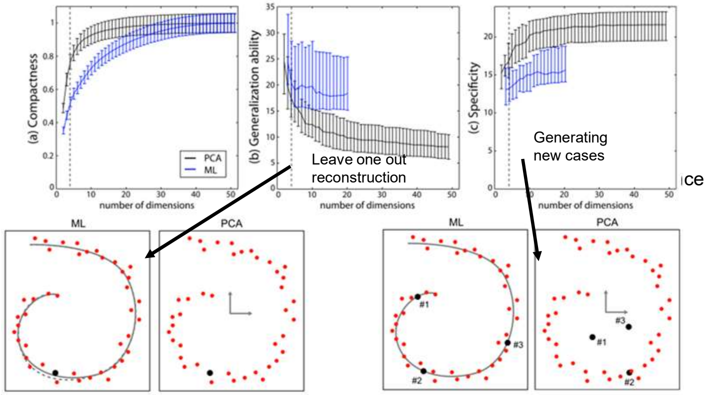
- **Compactness** = data explained by few dimensions?
- **Generalization ability** = error in reconstructing known cases?
- **Specificity** = consistency of generated new data?

---

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**



Leave one out reconstruction

Generating new cases

ce

Duchateau et al. *Med Image Anal* 2012

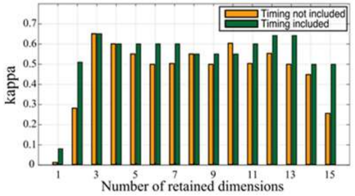
---

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**

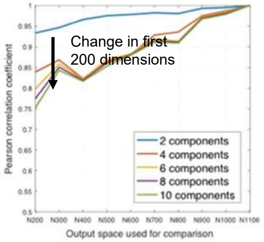
Error vs. existing (arguable) clinical labels:



Sanchez-Martinez et al. *Med Image Anal* 2017

Stability of low-dimensional coordinates:

Reference = 200 subjects of the whole OS



Change in first 200 dimensions

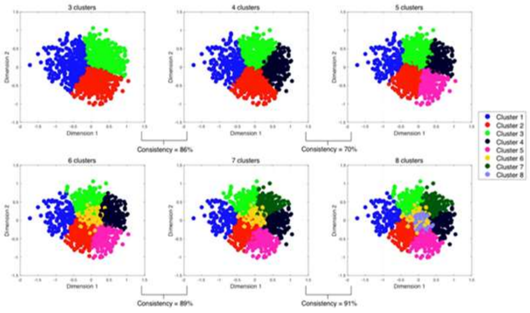
Cikes et al. *Eur J Heart Fail* 2019

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**

Consistency of clusters



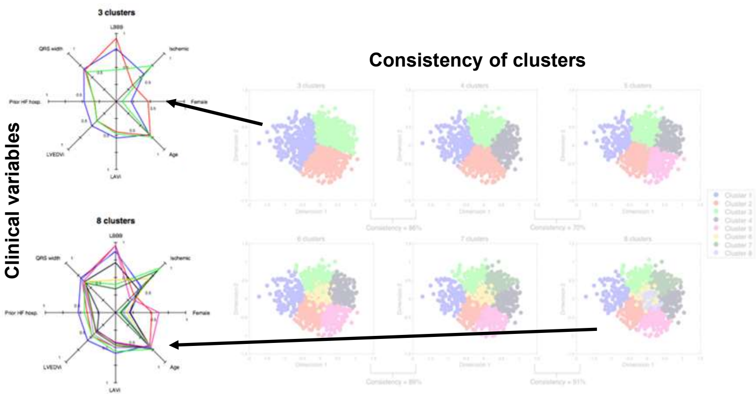
Cikes et al. *Eur J Heart Fail* 2019

---

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**



Cikes et al. *Eur J Heart Fail* 2019

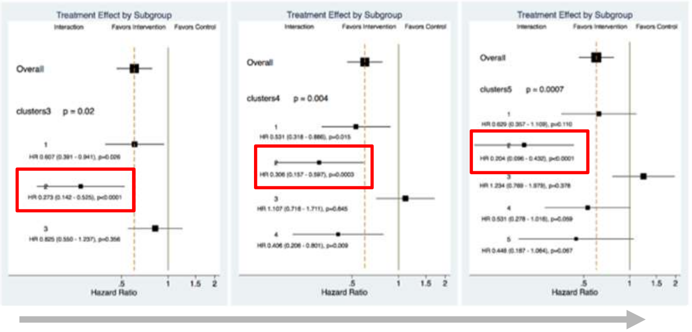
---

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**

vs. risk stratification:



Cikes et al. *Eur J Heart Fail* 2019

---

4. Methods evaluation

Unsupervised

2) Dimensionality / clusters: On **real data = ?**

vs. risk stratification:

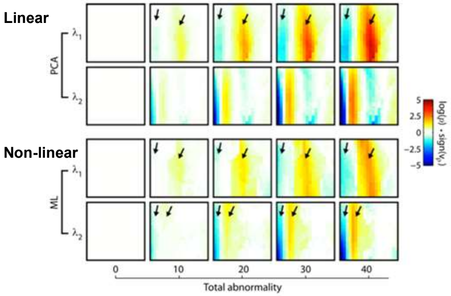
Cikes et al. *Eur J Heart Fail* 2019

4. Methods evaluation

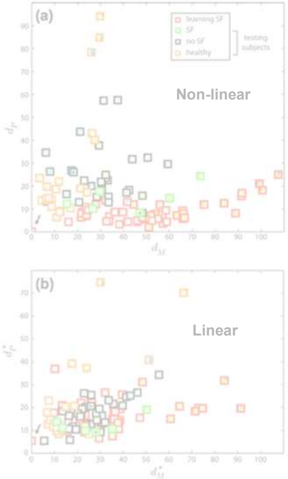
Unsupervised

3) Linear / non-linear:

vs. physiological consistency:



vs. explanatory power:



Duchateau et al. *Med Image Anal* 2012

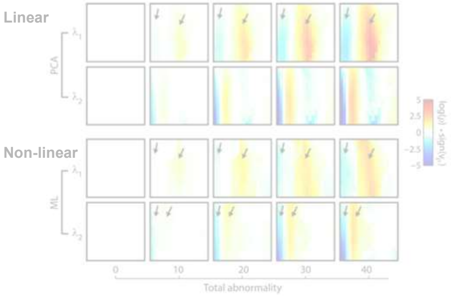
---

4. Methods evaluation

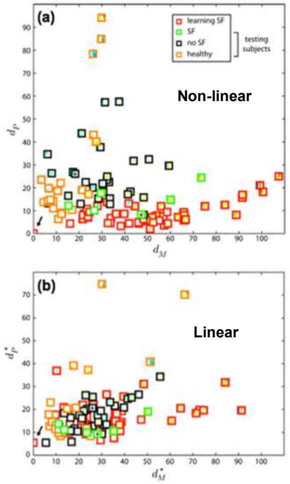
Unsupervised

3) Linear / non-linear:

vs. physiological consistency:



vs. explanatory power:



Duchateau et al. *Med Image Anal* 2012

Program

~30min	**1. Introduction** Carole Lartizien
~75min	**2. Supervised learning** Rémi Emonet + Carole Lartizien
~75min	**3. Unsupervised learning** Nicolas Duchateau + Rémi Emonet
~30min	**4. Methods evaluation** Carole Lartizien + Rémi Emonet + Nicolas Duchateau
	5. Conclusions / to go further

---

5. Conclusions

Take-home message

➔ **Supervised / unsupervised**
- Can handle complex data
- Proper definition of your problem vs. application
- Proper evaluation is required
- Transparency = open code, open data

➔ **For medical imaging applications:**
- “Standard” learning = already powerful to go beyond simple statistics
- Specificities of **medical data** vs. computer vision or data science
  - High dimension
  - Specific properties (ex: tensor)

5.Conclusions

Take-home message… and challenges

- ➔ **Supervised / unsupervised / semi-supervised**
    - Can handle complex data
    - Proper definition of your problem vs. application
    - Proper evaluation is required
    - Transparency = open code, open data
    - Mixing **multiple heterogeneous** descriptors?
- ➔ **For medical imaging applications:**
    - “Standard” learning = already powerful to go beyond simple statistics
    - Specificities of **medical data** vs. computer vision or data science
        - High dimension
        - Specific properties (ex: tensor)
        - **Real life** data = quality / completeness / amount
        - **Physiological** prior
        - Consequences of **uncertainties**?

5.Conclusions

Take-home message… and challenges

- ➔ **Supervised / unsupervised / semi-supervised**
    - Can handle complex data
    - Proper definition of your problem vs. application
    - Proper evaluation is required
    - Transparency = open code, open data
    - Mixing **multiple heterogeneous** descriptors?
- ➔ **For medical imaging applications:**
    - “Standard” learning = already powerful to go beyond simple statistics
    - Specificities of **medical data** vs. computer vision or data science
        - High dimension
        - Specific properties (ex: tensor)
        - **Real life** data = quality / completeness / amount
        - **Physiological** prior
        - Consequences of **uncertainties**?

Thanks !... Any questions?