RETREAT GRBIO 2024

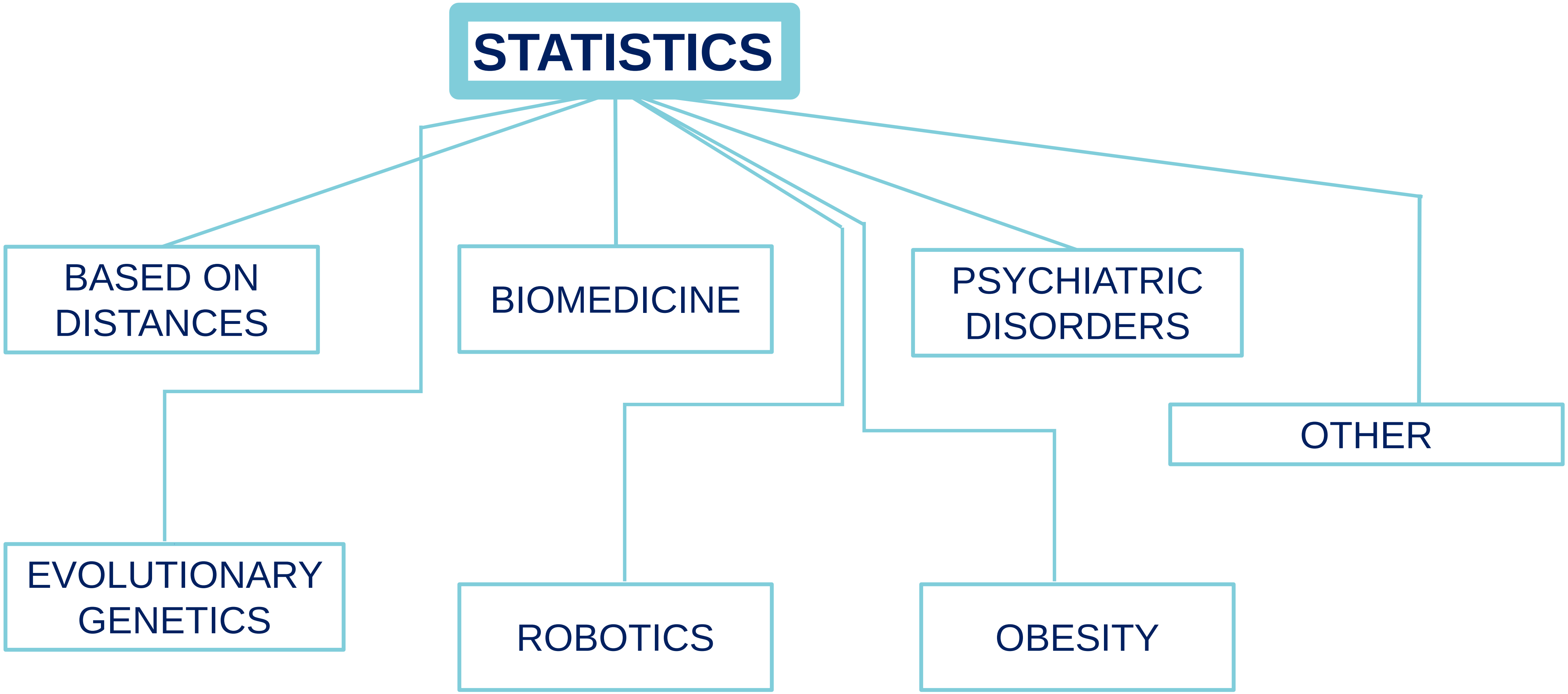
WORKING IN STATISTICS WITH FLIES, ROBOTS AND HUMANS

CONXITA ARENAS

WORKING IN....

My research activity in statistics is grouped into different blocks

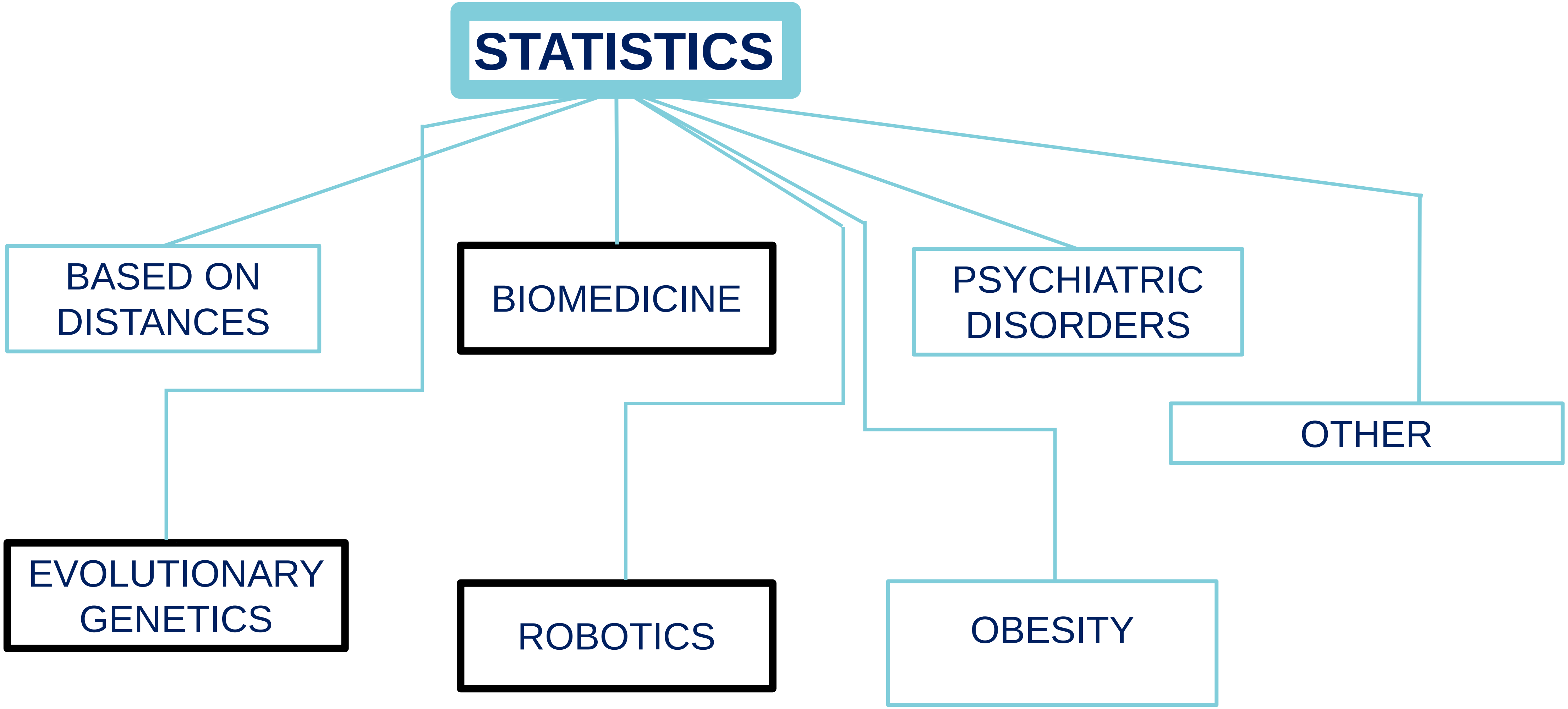
WORKING IN....


STATISTICS
BASED ON DISTANCES
BIOMEDICINE
PSYCHIATRIC DISORDERS
OTHER
EVOLUTIONARY GENETICS
ROBOTICS
OBESITY

WORKING IN....

and I will comment on three of them

WORKING IN....


STATISTICS
BASED ON DISTANCES
BIOMEDICINE
PSYCHIATRIC DISORDERS
OTHER
EVOLUTIONARY GENETICS
ROBOTICS
OBESITY

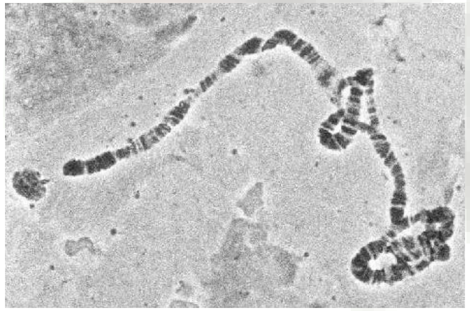
STATISTICS IN EVOLUTIONARY GENETICS

At the end of the 90s I began the collaboration, still active, with Dr. Mestres of the Department of Genetics, Microbiology and Statistics

This collaboration focuses on three lines of research

STATISTICS IN EVOLUTIONARY GENETICS

Drosophila subobscura

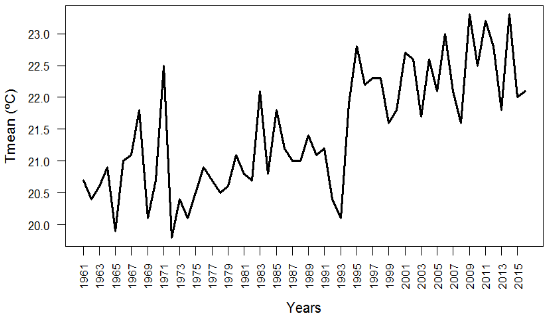


Genotype:	O_GO_G	O_GO_5	O_5O_5
Relative fitness:	$w_1 = 1 - s$	$w_2 = 1$	$w_3 = 0.$

Using the model species *Drosophila subobscura* to understand different aspects of the origin of the colonization of America. A model was developed that allowed to quantify the effect of natural selection

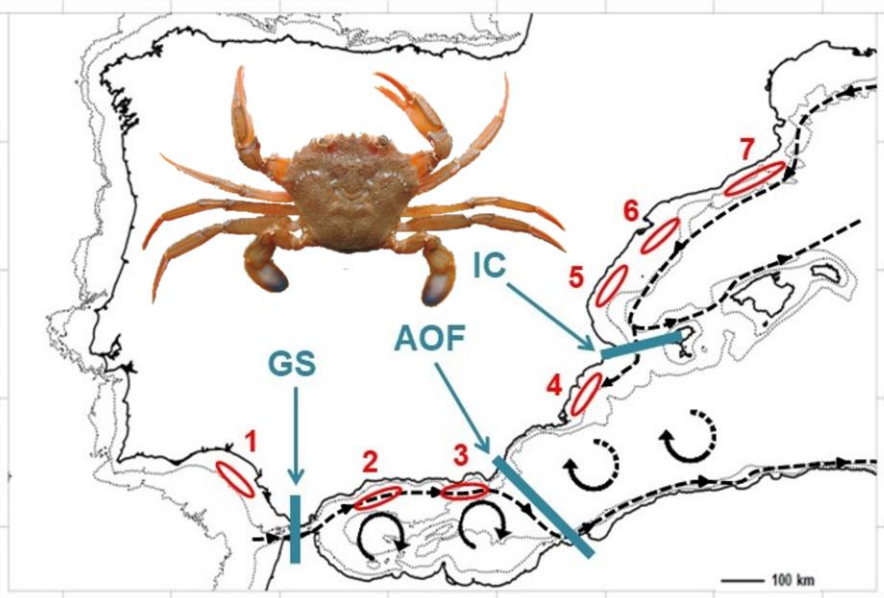
STATISTICS IN EVOLUTIONARY GENETICS

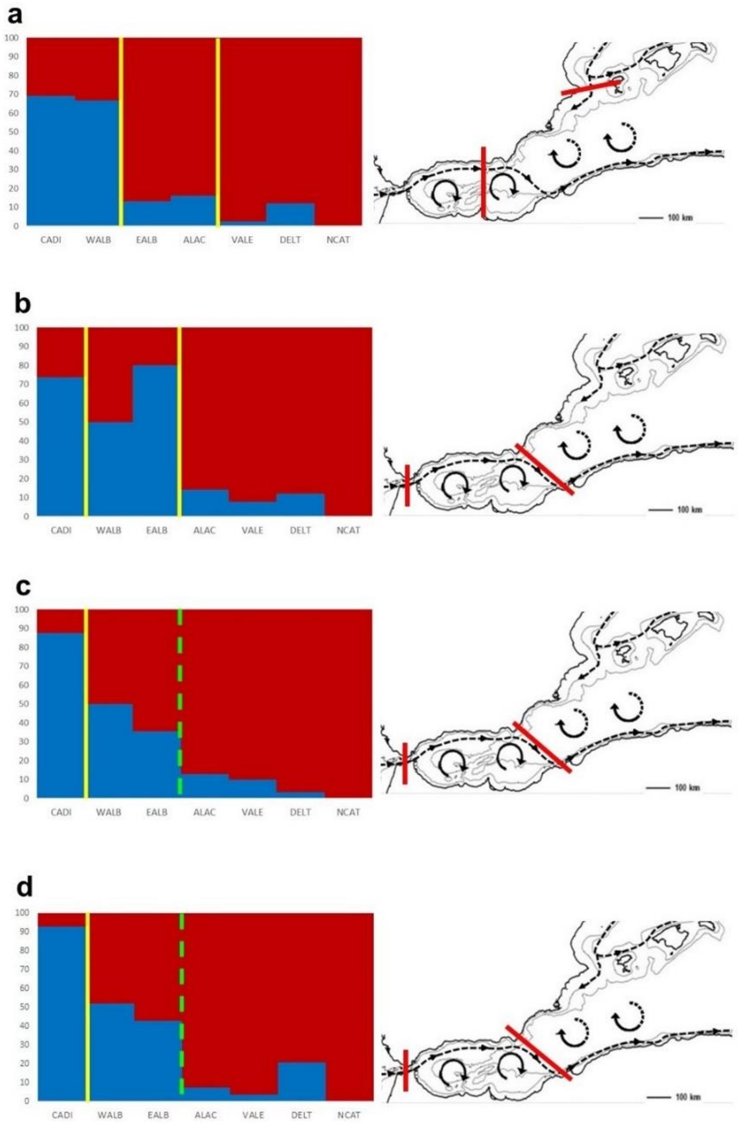
To quantify the adaptive effect of chromosomal inversions of *D. subobscura* with respect to climate change



$$\mathrm{CTI} = \frac{W - C}{W + C} = \frac{W - C}{\mathrm{TA}}$$

STATISTICS IN EVOLUTIONARY GENETICS





Using the marine crab *Liocarcinus depurator*, we can study whether ocean currents act or not as a barrier to gene exchange between populations. This is of interest both, to define marine protected areas and develop fishing policies

STATISTICS IN EVOLUTIONARY GENETICS

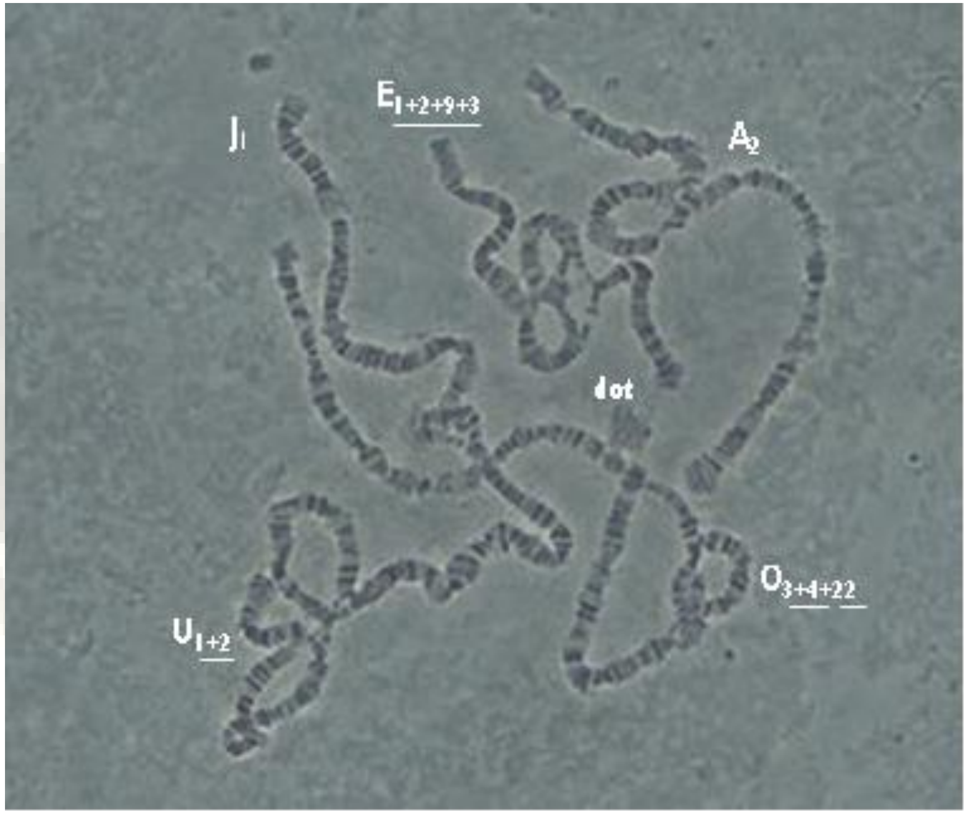
Drosophila subobscura

To quantify the adaptive effect of chromosomal inversions of *D. subobscura* with respect to climate change

STATISTICS IN EVOLUTIONARY GENETICS

Adaptations are features of an organism's design that allow it to survive and reproduce. To be useful, they must be heritable

An **inversion** is a chromosome fragment facing the opposite direction to normal. It does not represent any problem for the specimens



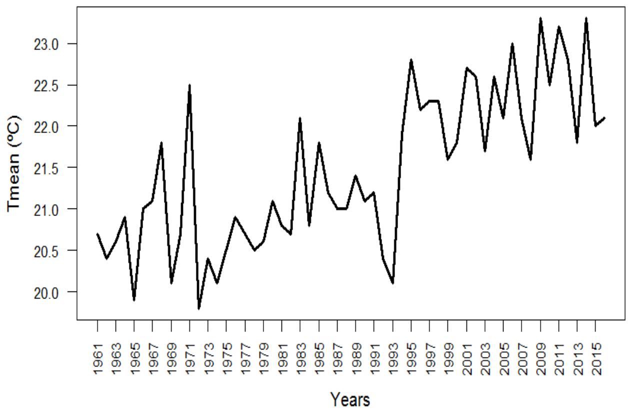
It is well known that chromosomal **inversions** in natural populations of the species *Drosophila subobscura* are **adaptations** to changes in the environment

STATISTICS IN EVOLUTIONARY GENETICS

If the gene combination inside an inversion is adaptive and allows it to survive and reproduce better to particular environment conditions, it will be favored by natural selection and will increase in frequency over generations

STATISTICS IN EVOLUTIONARY GENETICS

CTI index (chromosomal thermal index), that allows to quantify the thermally adaptation. With it, we can compared different years or populations



UNIVERSIDADE da MADEIRA

UNIVERSITY OF BELGRADE

Sampling different populations of Serbia, Madeira, Iran, Catalonia and America (North and South), using CTI and climatic variables, it can be observed that *warm* inversions are significantly increasing as the temperature increases and *cold* ones are decreasing

Now we are analyzing a Serbia's sample from 2023

WORKING IN....

STATISTICS

BIOMEDICINE

PSYCHIATRIC DISORDERS

ROBOTICS

OBESITY

STATISTICS IN BIOMEDICINE

This line of research is my own line that I started in 2003

During these years, methodological contributions have been made, generally motivated by a biomedical situation

STATISTICS IN BIOMEDICINE

- The development of cluster methods based on the concept of geometric variability
- Make a solution to the problem of typicality for any distance and number of groups

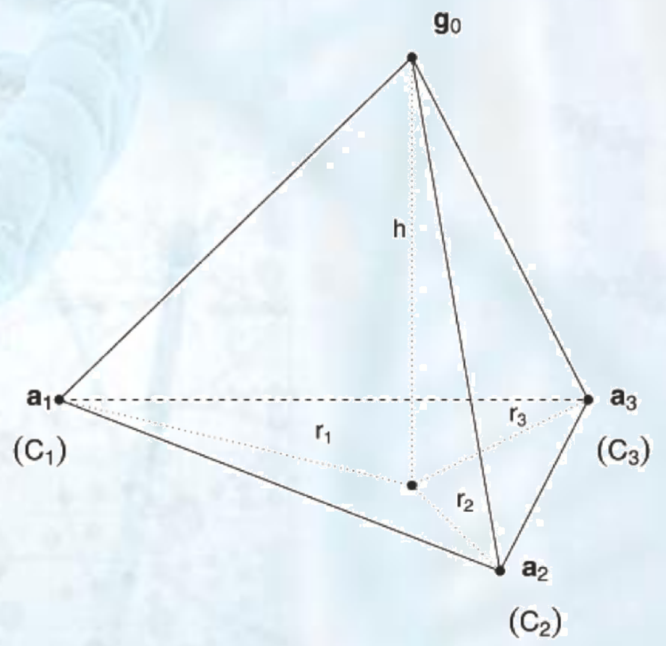
statistic, INCA statistic, that generalizes Rao's statistic [11] is defined as follows:

$$W(\mathbf{y}_0) = \min_{\alpha_i}\{L(\mathbf{y}_0)\}, \quad \sum_{i=1}^{k} \alpha_i = 1 \tag{3}$$

where

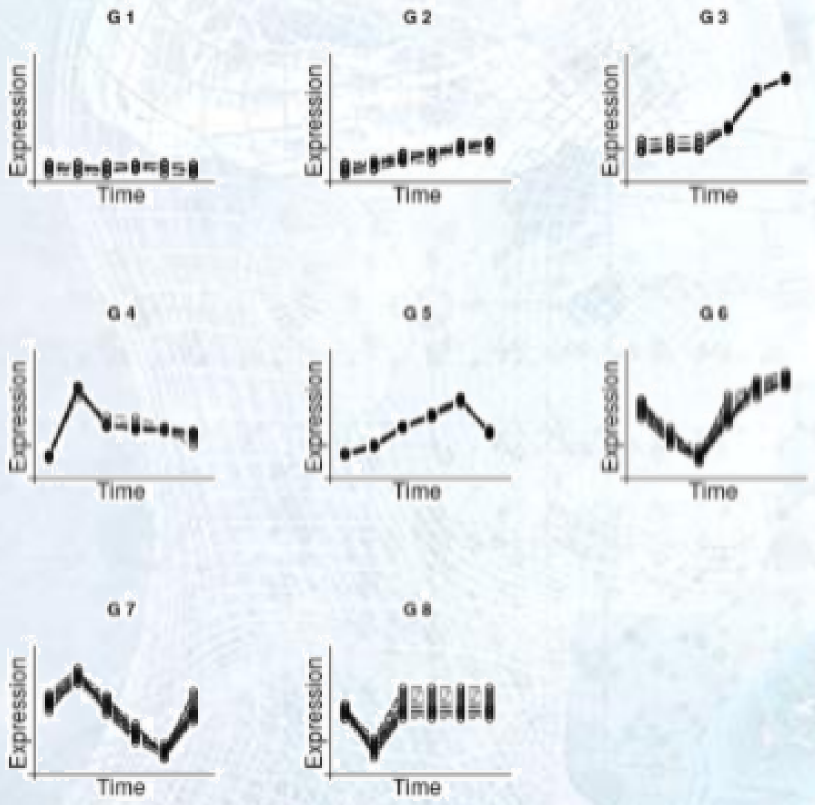
$$L(\mathbf{y}_0) = \sum_{i=1}^{k} \alpha_i \phi_i^2(\mathbf{y}_0) - \sum_{1 \leqslant i < j \leqslant k} \alpha_i \alpha_j \Delta_{ij}^2$$

$\phi^2_{(\mathbf{y}_0)}$ is the proximity function of $\mathbf{y}_0$ to C_i and Δ^2_{ij} is the squared distance between C_i and C_j. The INCA statistic $W(\mathbf{y}_0) = \min_{\alpha_i} L(\mathbf{y}_0)$ trades off between minimizing the weighted sum



STATISTICS IN BIOMEDICINE

- Construction of an adequate distance for microarray clustering
- Detection of gene expression patterns over time



STATISTICS IN BIOMEDICINE

- Construction of a depth function to detect the most characteristic units of clusters
- Method to detect differentially expressed genes

$$I = I(\mathbf{z}_i, C) = \left[1 + \frac{\|\mathbf{z}_i - E(\mathbf{Z})\|^2}{E(\|\mathbf{Z} - E(\mathbf{Z})\|)}\right]^{-1}, \qquad (2)$$

function (1) takes values in [0,1], and assigns to any unit a degree of centrality with respect to the data cloud. Thus,

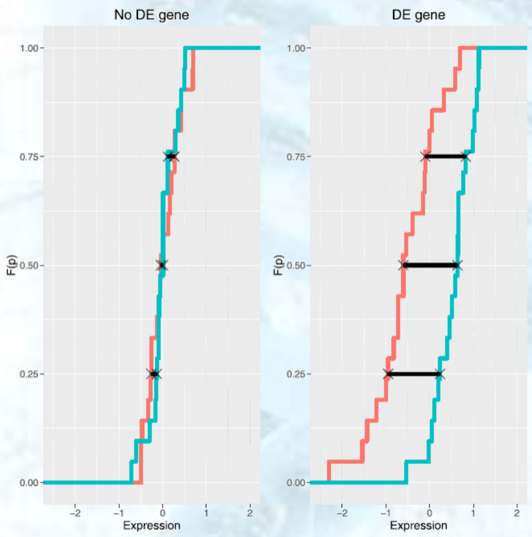
For each class C_k we weight the discriminant score δ_k^1 by $1 - I_k(\mathbf{y}^*)$, that is, given a new unit $\mathbf{y}^*$, we define a new discriminant score for class k by:

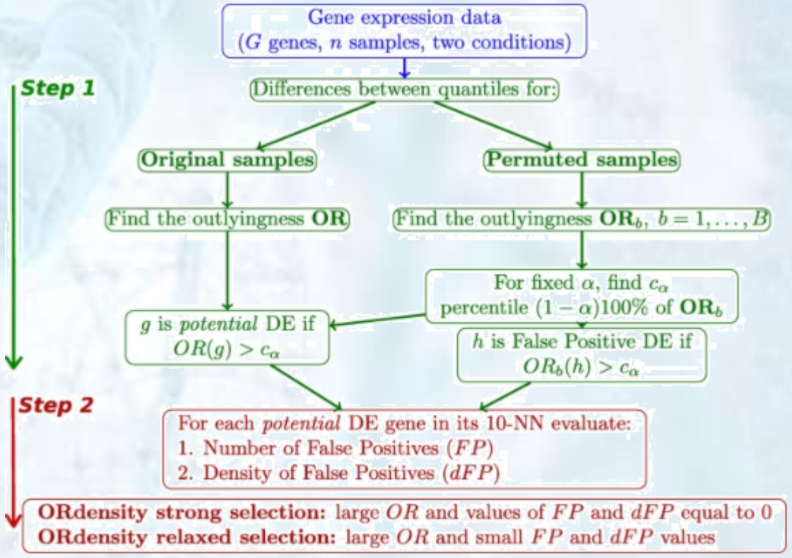
$$\delta_k^2(\mathbf{y}^*) = \delta_k^1(1 - I_k(\mathbf{y}^*)) = \phi^2(\mathbf{y}^*, C_k)(1 - I_k(\mathbf{y}^*)). \qquad (5)$$

The shrinkage we use, reduces the proximity values, this reduction being greater for deeper units. Thus, this new classification rule,

$$C_{WDB}(\mathbf{y}^*) = l \quad \text{where} \quad \delta_l^2(\mathbf{y}^*) = \min_{k=1,\dots,K}\left\{\delta_k^2(\mathbf{y}^*)\right\}, \qquad (6)$$

allocates a new unit $\mathbf{y}^*$ to the class which has the minimal proximity and maximal depth values.





General outline of the proposed ORdensity approach. In green the first step of the method and in red the second step of the method

STATISTICS IN BIOMEDICINE

- Fuzzy classification methodology, both supervised and unsupervised

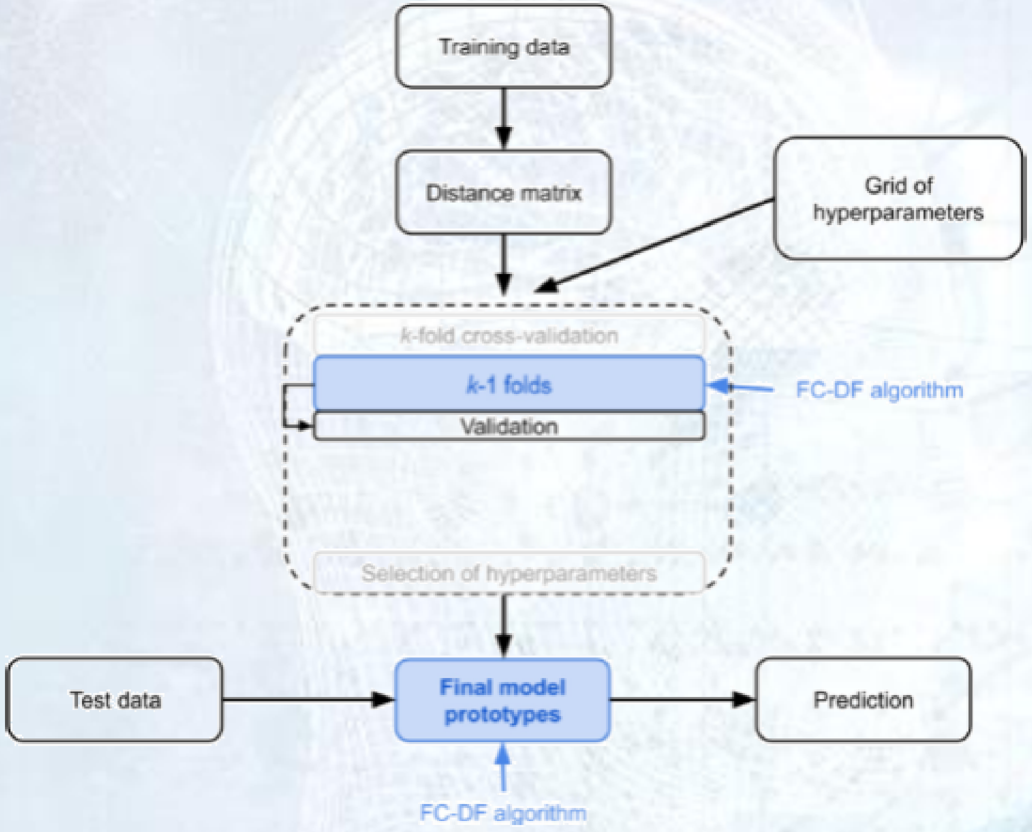


Fig. 1. Workflow of the general process followed to build the classifier.

$$\min_{u_{ik}, \mathbf{a}_k, \mathbf{z}_k} \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \delta^2(\mathbf{x}_i, \mathbf{a}_k) + \alpha \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} l(y_i, \mathbf{z}_k) + \gamma \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \log(u_{ik}),$$

subject to

$$\sum_{k=1}^{K} u_{ik} = 1, \quad u_{ik} \geq 0, \quad i = 1, \ldots, n, \; k = 1, \ldots, K,$$

The method

Fuzzy classification with distance-based prototypes

Irigoien, I.[1], Arenas, C.[2]

[1] Computation Sciences and Artificial Intelligence
University of the Basque Country UPV/EHU
[2] Statistics Section. Department of Genetics, Microbiology and Statistics.
Universitat de Barcelona

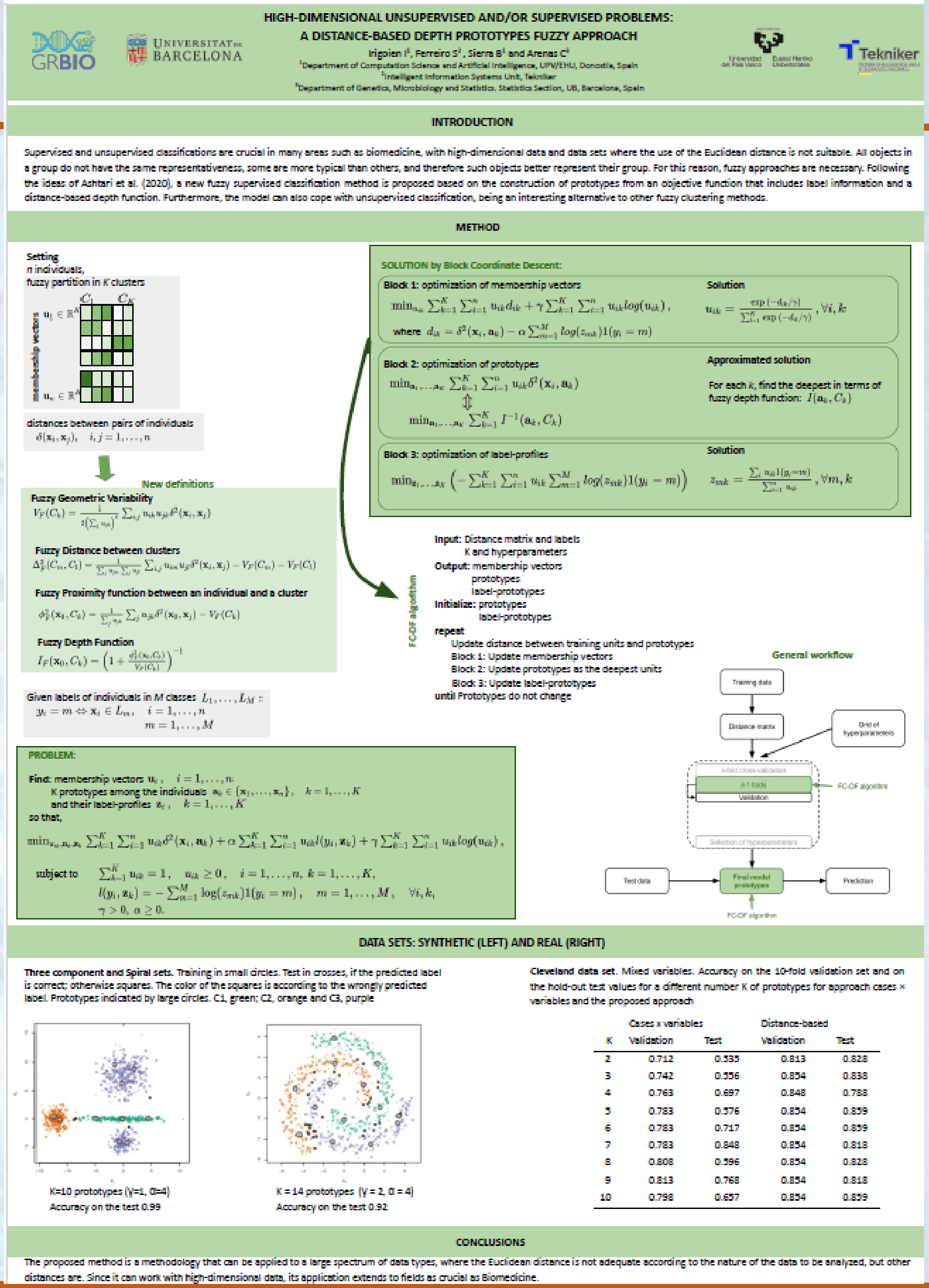
COMPSTAT 2022
Bologna

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

STATISTICS IN BIOMEDICINE

The complete method was presented in



HIGH-DIMENSIONAL UNSUPERVISED AND/OR SUPERVISED PROBLEMS:
A DISTANCE-BASED DEPTH PROTOTYPES FUZZY APPROACH

Irigoien I[1], Ferreiro S[2], Sierra B[1] and Arenas C[3]
[1]Department of Computation Science and Artificial Intelligence, UPV/EHU, Donostia, Spain
[2]Intelligent Information Systems Unit, Tekniker
[3]Department of Genetics, Microbiology and Statistics. Statistics Section, UB, Barcelona, Spain

INTRODUCTION

Supervised and unsupervised classifications are crucial in many areas such as biomedicine, with high-dimensional data and data sets where the use of the Euclidean distance is not suitable. All objects in a group do not have the same representativeness, some are more typical than others, and therefore such objects better represent their group. For this reason, fuzzy approaches are necessary. Following the ideas of Ashtari et al. (2020), a new fuzzy supervised classification method is proposed based on the construction of prototypes from an objective function that includes label information and a distance-based depth function. Furthermore, the model can also cope with unsupervised classification, being an interesting alternative to other fuzzy clustering methods.

METHOD

Setting
n individuals,
fuzzy partition in K clusters

distances between pairs of individuals
$\delta(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \dots, n$

New definitions

Fuzzy Geometric Variability
$V_F(C_k) = \frac{1}{2(\sum_i u_{ik})^2} \sum_{i,j} u_{ik} u_{jk} \delta^2(\mathbf{x}_i, \mathbf{x}_j)$

Fuzzy Distance between clusters
$\Delta_F^2(C_m, C_l) = \frac{1}{\sum_i u_{im} \sum_j u_{jl}} \sum_{i,j} u_{im} u_{jl} \delta^2(\mathbf{x}_i, \mathbf{x}_j) - V_F(C_m) - V_F(C_l)$

Fuzzy Proximity function between an individual and a cluster
$\phi_F^2(\mathbf{x}_0, C_k) = \frac{1}{\sum_j u_{jk}} \sum_j u_{jk} \delta^2(\mathbf{x}_0, \mathbf{x}_j) - V_F(C_k)$

Fuzzy Depth Function
$I_F(\mathbf{x}_0, C_k) = \left(1 + \frac{\phi_F^2(\mathbf{x}_0, C_k)}{V_F(C_k)}\right)^{-1}$

Given labels of individuals in M classes $L_1, \dots, L_M$:
$y_i = m \Leftrightarrow \mathbf{x}_i \in L_m, \quad i = 1, \dots, n$
$m = 1, \dots, M$

PROBLEM:

Find: membership vectors $\mathbf{u}_i, \quad i = 1, \dots, n$;
K prototypes among the individuals $\mathbf{a}_k \in \{\mathbf{x}_1, \dots, \mathbf{x}_n\}, \quad k = 1, \dots, K$
and their label-profiles $\mathbf{z}_k, \quad k = 1, \dots, K$
so that,

$$\min_{\mathbf{u}_i, \mathbf{a}_k, \mathbf{z}_k} \sum_{k=1}^K \sum_{i=1}^n u_{ik} \delta^2(\mathbf{x}_i, \mathbf{a}_k) + \alpha \sum_{k=1}^K \sum_{i=1}^n u_{ik} l(y_i, \mathbf{z}_k) + \gamma \sum_{k=1}^K \sum_{i=1}^n u_{ik} log(u_{ik}),$$

subject to $\sum_{k=1}^K u_{ik} = 1, \quad u_{ik} \geq 0, \quad i = 1, \dots, n, \ k = 1, \dots, K,$
$l(y_i, \mathbf{z}_k) = -\sum_{m=1}^M \log(z_{mk}) 1(y_i = m), \quad m = 1, \dots, M, \quad \forall i, k,$
$\gamma > 0, \ \alpha \geq 0.$

SOLUTION by Block Coordinate Descent:

Block 1: optimization of membership vectors
$\min_{\mathbf{u}_i} \sum_{k=1}^K \sum_{i=1}^n u_{ik} d_{ik} + \gamma \sum_{k=1}^K \sum_{i=1}^n u_{ik} log(u_{ik}),$
where $d_{ik} = \delta^2(\mathbf{x}_i, \mathbf{a}_k) - \alpha \sum_{m=1}^M log(z_{mk}) 1(y_i = m)$

Solution
$u_{ik} = \frac{\exp(-d_{ik}/\gamma)}{\sum_{l=1}^K \exp(-d_{il}/\gamma)}, \forall i, k$

Block 2: optimization of prototypes
$\min_{\mathbf{a}_1, \dots, \mathbf{a}_K} \sum_{k=1}^K \sum_{i=1}^n u_{ik} \delta^2(\mathbf{x}_i, \mathbf{a}_k)$
$\Updownarrow$
$\min_{\mathbf{a}_1, \dots, \mathbf{a}_K} \sum_{k=1}^K I^{-1}(\mathbf{a}_k, C_k)$

Approximated solution
For each k, find the deepest in terms of fuzzy depth function: $I(\mathbf{a}_k, C_k)$

Block 3: optimization of label-profiles
$\min_{\mathbf{z}_1, \dots, \mathbf{z}_K} \left(-\sum_{k=1}^K \sum_{i=1}^n u_{ik} \sum_{m=1}^M log(z_{mk}) 1(y_i = m)\right)$

Solution
$z_{mk} = \frac{\sum_i u_{ik} 1(y_i = m)}{\sum_{i=1}^n u_{ik}}, \forall m, k$

FC-DF algorithm

Input: Distance matrix and labels
K and hyperparameters
Output: membership vectors
prototypes
label-prototypes
Initialize: prototypes
label-prototypes
repeat
Update distance between training units and prototypes
Block 1: Update membership vectors
Block 2: Update prototypes as the deepest units
Block 3: Update label-prototypes
until Prototypes do not change

General workflow

Training data
Distance matrix
Grid of hyperparameters
k-fold cross-validation
k-1 folds
Validation
FC-DF algorithm
Selection of hyperparameters
Test data
Final model prototypes
Prediction
FC-DF algorithm

DATA SETS: SYNTHETIC (LEFT) AND REAL (RIGHT)

Three component and Spiral sets. Training in small circles. Test in crosses, if the predicted label is correct; otherwise squares. The color of the squares is according to the wrongly predicted label. Prototypes indicated by large circles. C1, green; C2, orange and C3, purple

K=10 prototypes (γ=1, α=4)
Accuracy on the test 0.99

K = 14 prototypes (γ = 2, α = 4)
Accuracy on the test 0.92

Cleveland data set. Mixed variables. Accuracy on the 10-fold validation set and on the hold-out test values for a different number K of prototypes for approach cases × variables and the proposed approach

	Cases x variables		Distance-based	
K	Validation	Test	Validation	Test
2	0.712	0.535	0.813	0.828
3	0.742	0.556	0.854	0.838
4	0.763	0.697	0.848	0.788
5	0.783	0.576	0.854	0.859
6	0.783	0.717	0.854	0.859
7	0.783	0.848	0.854	0.818
8	0.808	0.596	0.854	0.828
9	0.813	0.768	0.854	0.818
10	0.798	0.657	0.854	0.859

CONCLUSIONS

The proposed method is a methodology that can be applied to a large spectrum of data types, where the Euclidean distance is not adequate according to the nature of the data to be analyzed, but other distances are. Since it can work with high-dimensional data, its application extends to fields as crucial as Biomedicine.

STATISTICS IN BIOMEDICINE

We are working on the program to improve the calculation time, to be able to work with large data sets, such as those generated in GWAS studies, that present the high dimensionality problem

WORKING IN....

STATISTICS

PSYCHIATRIC DISORDERS

ROBOTICS

OBESITY

WORKING IN....

STATISTICS

PSYCHIATRIC DISORDERS

This line of research started at 2011, collaboration still active with Dr. Cormand

(Dept. Genetics, Microbiology and Statistics)

To identify genes associated with psychiatric disorders such as ADHD, autism

or substance addiction

WORKING IN....

STATISTICS

This line of research started at 2019, collaboration still active with Dr. Rosa (Anthropology Section)

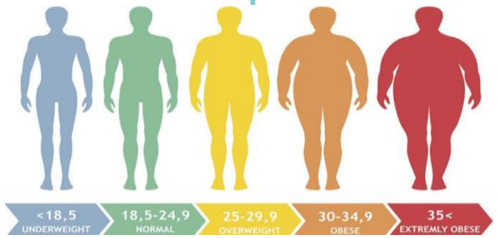
To identify genes associated with obesity

OBESITY

WORKING IN....

STATISTICS

ROBOTICS



STATISTICS IN ROBOTICS

In 2011, I began the collaboration, still active, with Dr. Sierra of the Basque Country University

During these years, methodological contributions have been made developing statistical methods in robotics

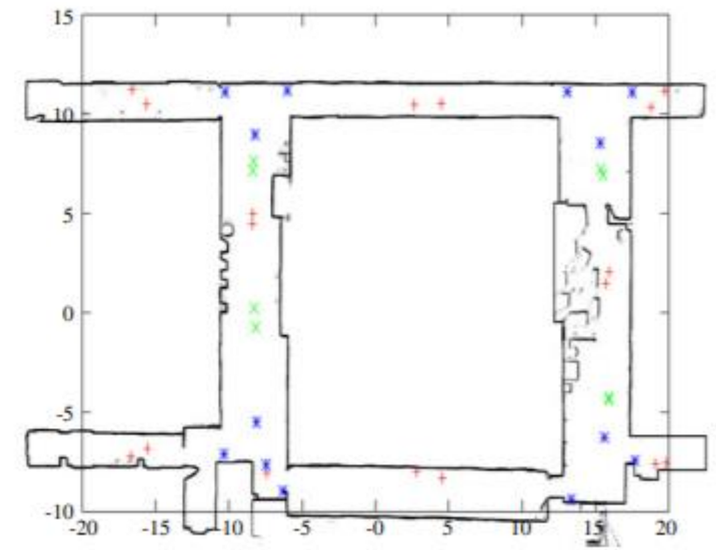
STATISTICS IN ROBOTICS




Robots assist elderly people, lost people, people recovering from injuries or people with mobility problems

STATISTICS IN ROBOTICS

In mobile robotics, one-class classification approaches can be applied to robot mapping, that is, to automatically learn the structure of its environment



The locations distribution over the map. Corridors: +; Halls: ×; Crossings: ∗.

Tartalo is the robot used in the real experiments.

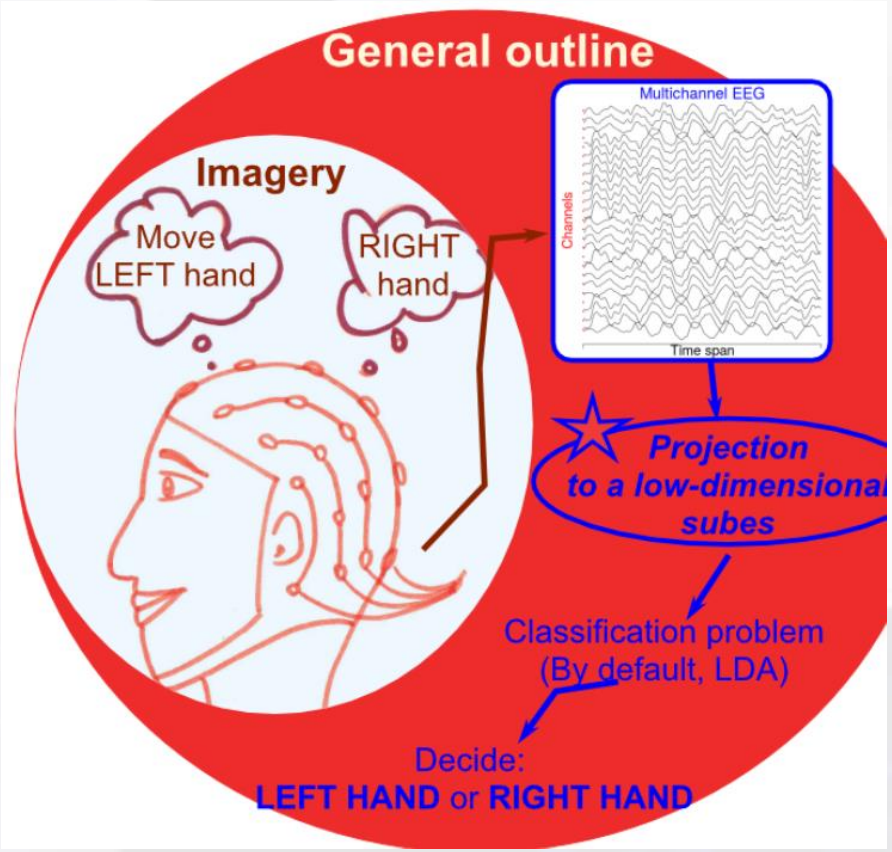
STATISTICS IN ROBOTICS

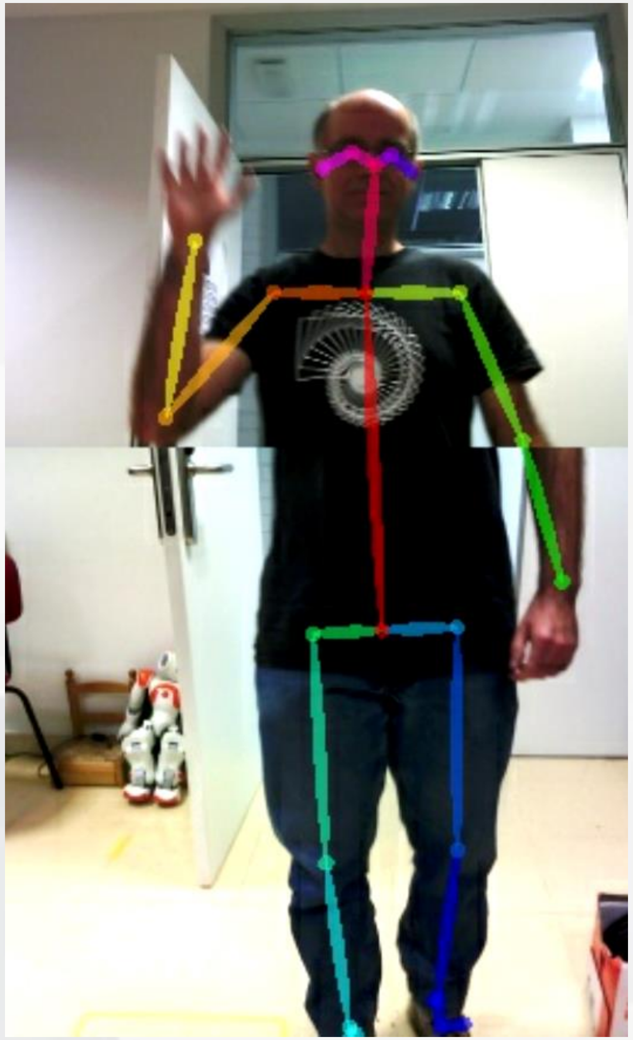
Re-identification of images from videos, specifically re-identification of people, using classification

Figure 1. Deputy captures of the Canary Islands Parliament. These images show different problematic situations where correct (green) and incorrect (red) intervener matches are presented.

STATISTICS IN ROBOTICS

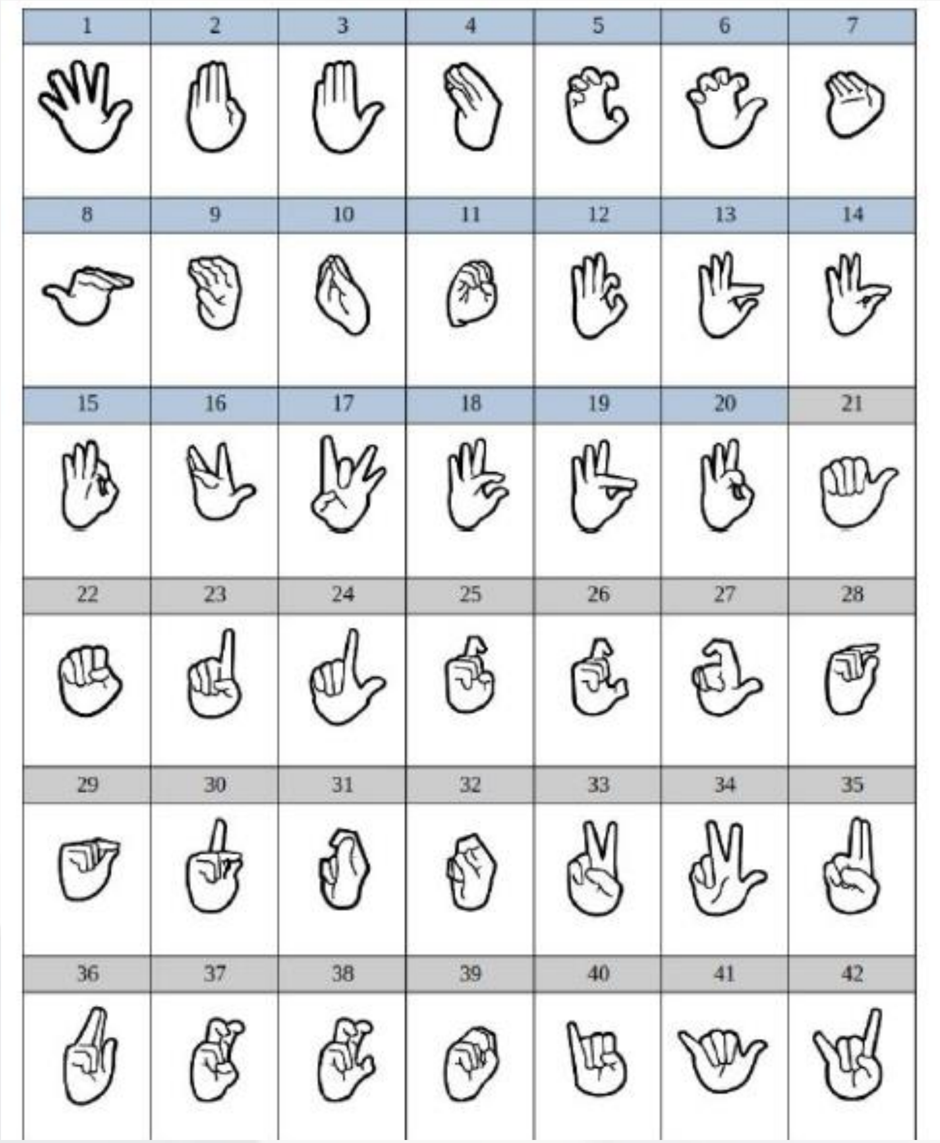
Classification of EEGs and Movement Identification, generalization based on distances of the Common Spatial Patterns method



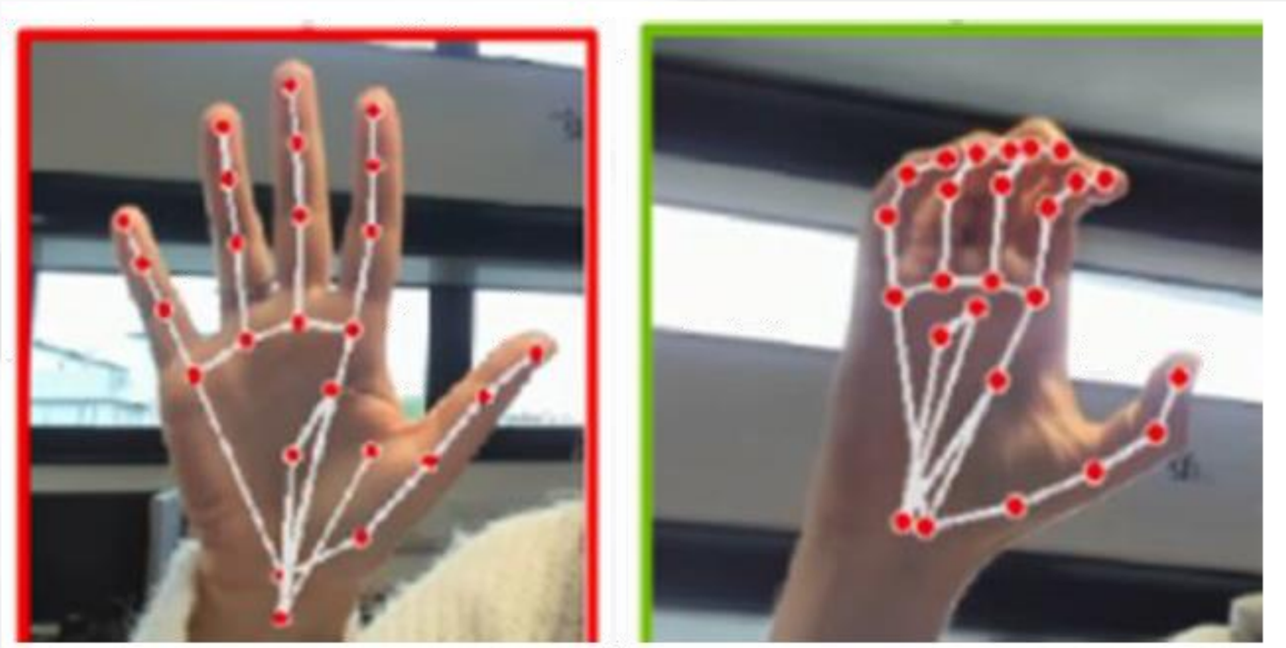


STATISTICS IN ROBOTICS

Be able to communicate using Spanish sign language

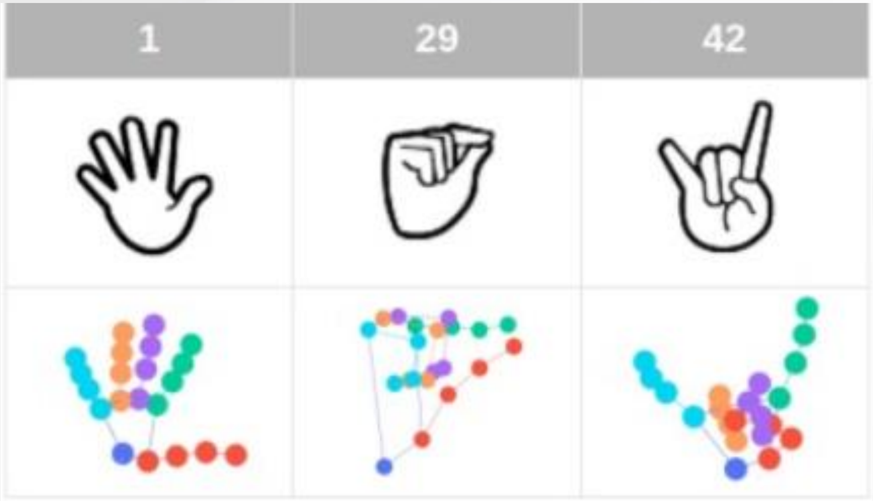


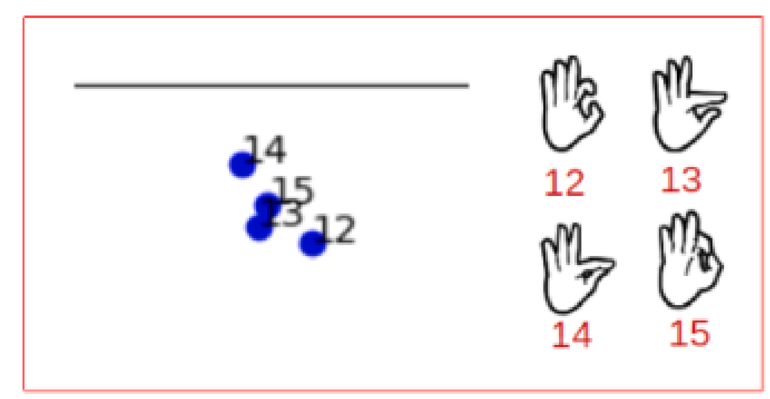
STATISTICS IN ROBOTICS

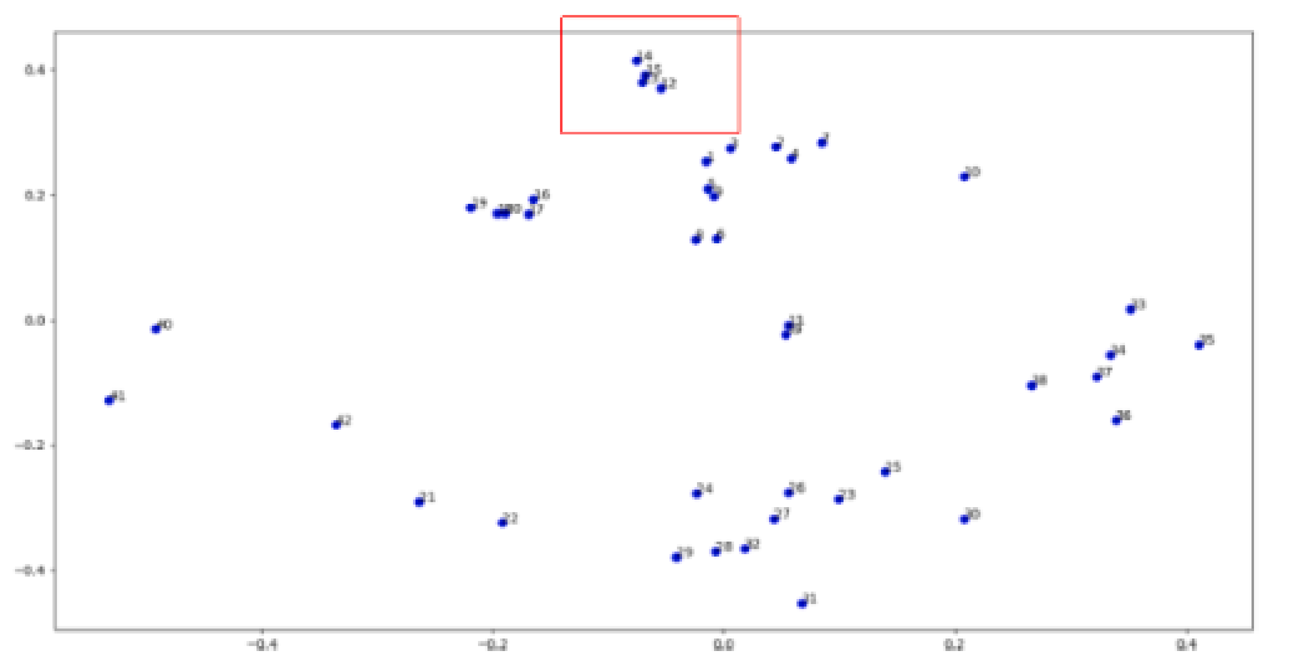


Using MediaPipe, 21 key points (CPs) are extracted from the hand. Each hand is characterized by a 21x2 (or 21x3) matrix whose rows contain the coordinates in the plane (space) of each CP. The third coordinate represents the depth with respect to the wrist.

STATISTICS IN ROBOTICS



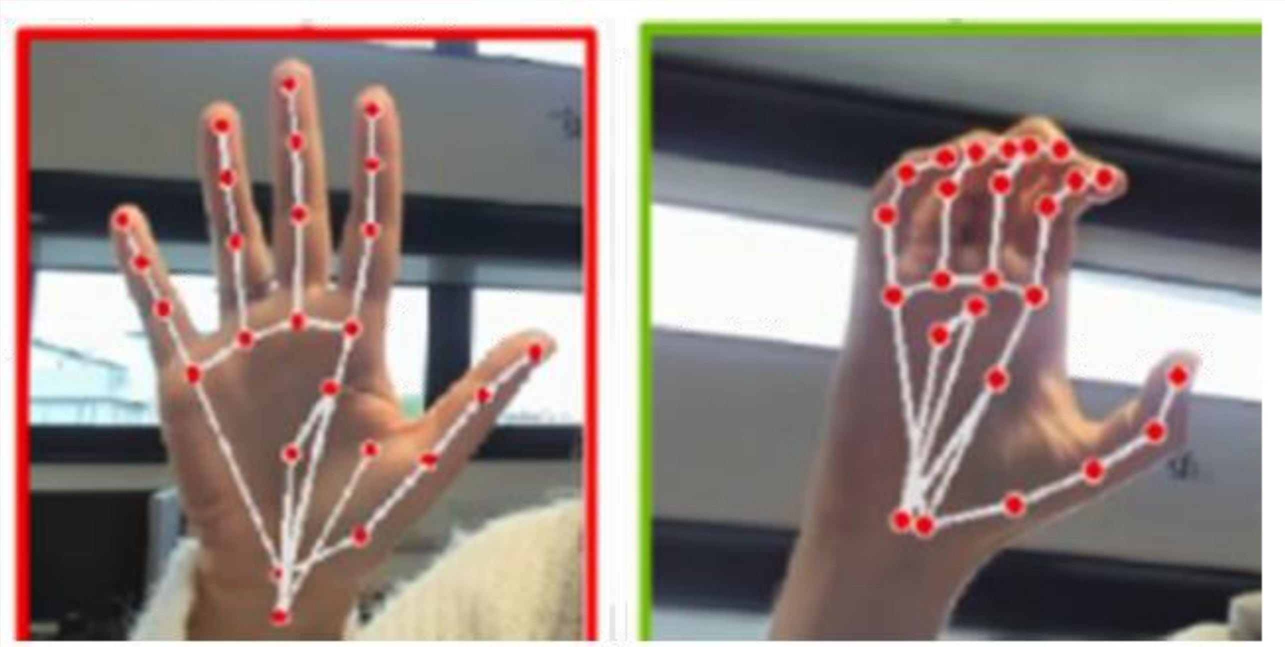


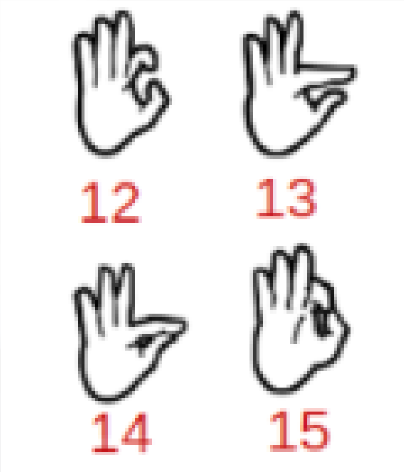


Using Procrustes distance we defined a rule of classification

STATISTICS IN ROBOTICS

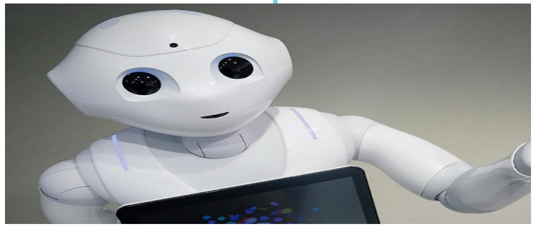
We are working on introducing more interdistances between key points and identifying those that are significant for each sign

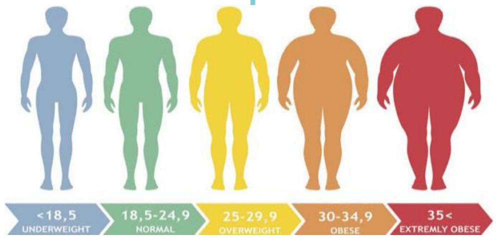




WORKING IN....

STATISTICS





WORKING IN STATISTICS DURING 2023-2024

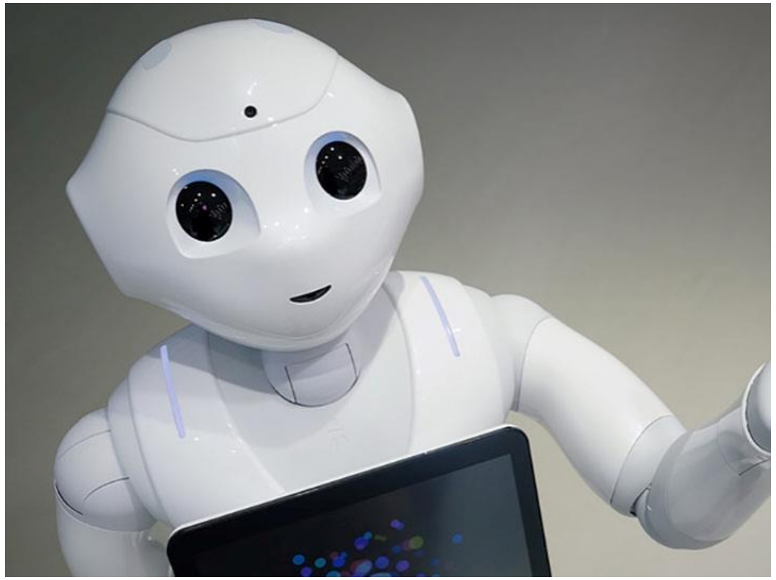
6 ARTICLES

5 CONGRESS

Currently: 2 projects Plan Nacional, 1 project Generalitat de Catalunya, Marató TV3 project and member of Redes de Investigación project

UAB
Universitat Autònoma de Barcelona
SJD
Sant Joan de Déu
Barcelona · Hospital
Vall d'Hebron
Hospital
UNIVERSITY OF WASHINGTON
1861
WILLIAM & MARY
CHARTERED 1693
UNIVERSITY OF BELGRADE
UNIVERSIDADE da MADEIRA
UPC
INSTITUT de BIOLOGIA EVOLUTIVA
IBE
CSIC
upf.
KING'S College LONDON
Parc Taulí
Hospital Universitari
Universidad del País Vasco
Euskal Herriko Unibertsitatea

WORKING IN STATISTICS WITH FLIES, ROBOTS AND HUMANS



THANKS!