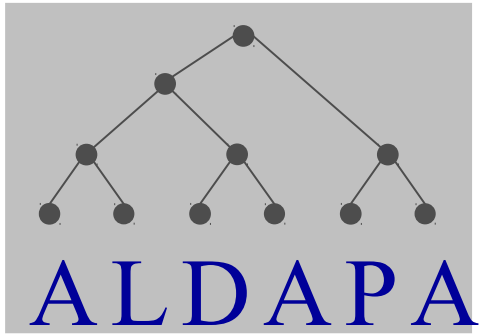
Algorithms Data Mining and Parallelism



Javier Muguerza

www.sc.ehu.es/aldapa

ALDAPA - Algorithms, Data Mining and Parallelism

- 12 researchers
  - 6 lecturers (PhD)
    - Olatz Arbelaitz
    - Agustin Arruabarrena
    - Ibai Gurrutxaga
    - José Ignacio Martín
    - Javier Muguerza
    - Jesús M. Pérez
  - 1 researcher (PhD)
    - Jose María Martínez
  - 4 PhD students
    - Joseba Alberdi
    - Igor Ibarguren
    - Aizea Lojo
    - Iñigo Perona

ALDAPA - Algorithms, Data Mining and Parallelism

- Arises in 2001 from the High Performance Solutions team (1990)
- Research activity:
  - **machine learning**, data mining
    - supervised and unsupervised learning, prediction, optimisation
  - solving **real problems**
  - **efficient** solutions
    - prallelism and high performance computing

- Latest colaborations
  - **S21sec** (computer security)
  - **Ikerlan** (object movement modelling)
  - **Tecnalia** (prediction – robotics)
  - **IK4 – Vicomtech – Bidasoa Turismo** (web user modelling - tourism)
  - **Nano-bio Spectroscopy Group** (HPC, GP-GPU)
  - **LIPCNE - egokituz** (user modelling – web mining and social networks, BCI) *[integration process]*

- Current research projects
  - UPV/EHU – Research group
    - All research lines
  - Basque Government - Saiotek
    - Web mining, Social web mining, Adaptive systems
  - Spanish government
    - Web mining for users with special needs

ALDAPA - Algorithms, Data Mining and Parallelism

Previous lines

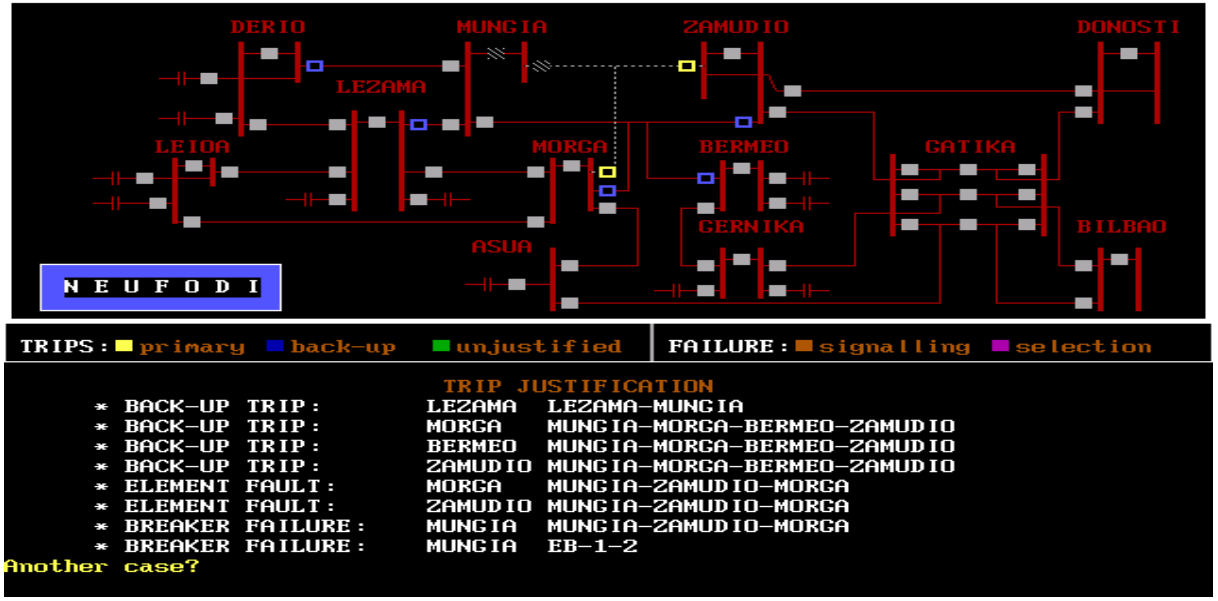
>> **NEUFODI:** ANN for detection and diagnosis
>> **OCR – FORM-LESS:** automatic character recognition
>> **Optimization:** VRPTW, PET
>> **MALBEC:** malware behaviour modelling
>> **CAMIKER:** object movement modelling

Current lines

>> **CTC – HARITZA:** comprehensibility, class imbalance
>> **MODELACCESS:** web user modelling, social web mining
>> **Physiological CS:** BCI, ECG

>> **HPC – Parallelism:** material physics, gp-gpu

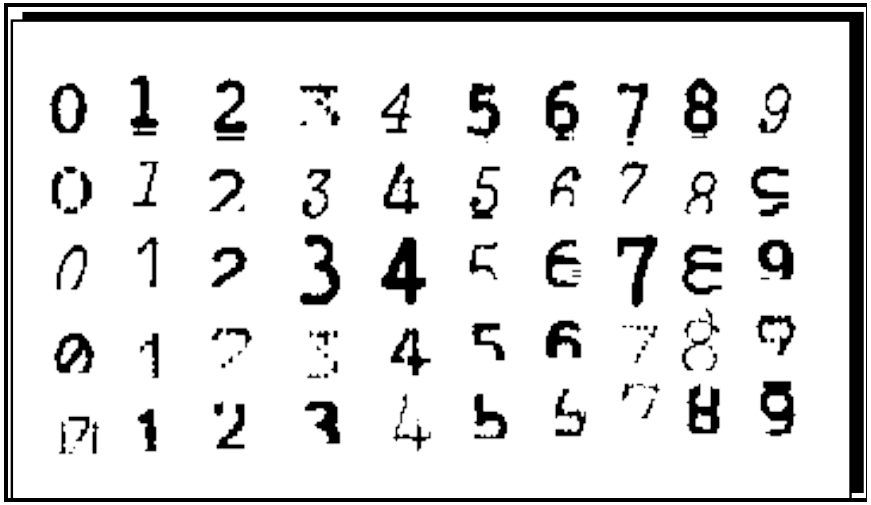
NEUFODI: ANN for detection and diagnosis



- Iberdrola – electric power transportation
- flood of alarms

- Techniques:
  - Artificial Neural Networks (MLP) + Parallelism

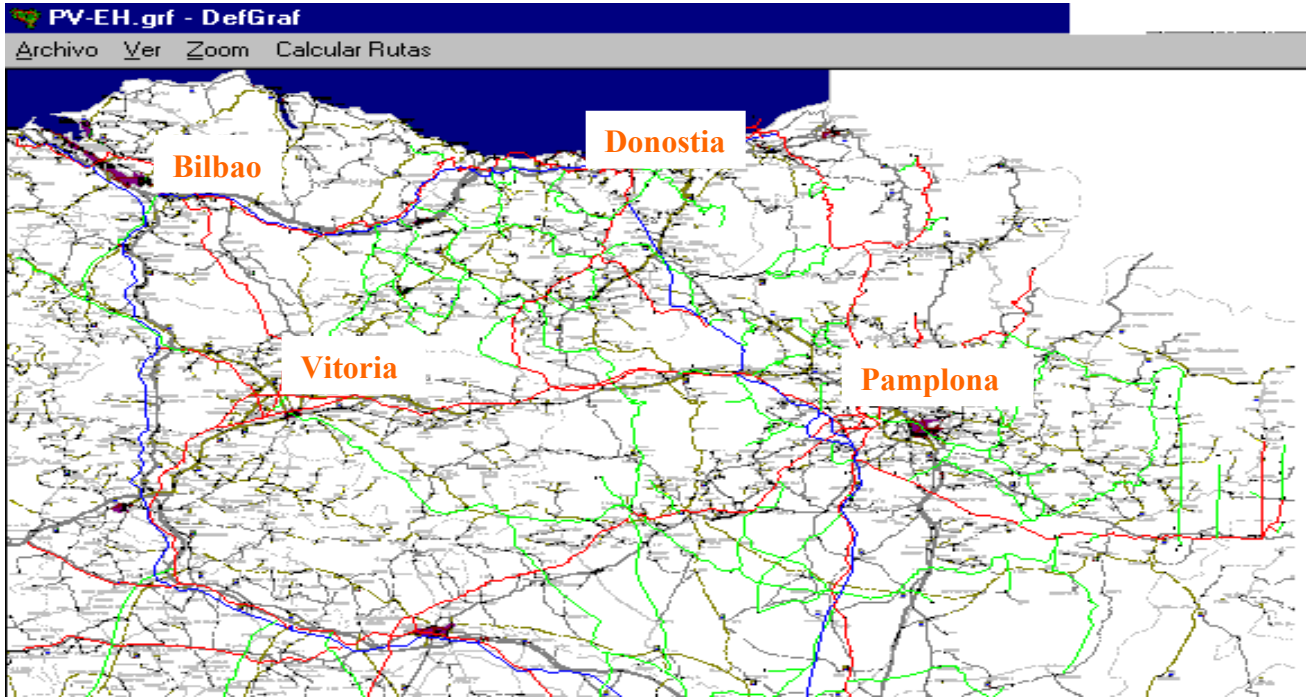
OCR – FORM-LESS: automatic character recognition



- Typewritten text (250K)
- High variability and poor quality
- error 2% → 3 months
- High accuraccy

- Techniques:
  - Supervised Classification (*k*-NN,...)
  - Hierarchical systems
  - Unsupervised learning
  - Parallelism

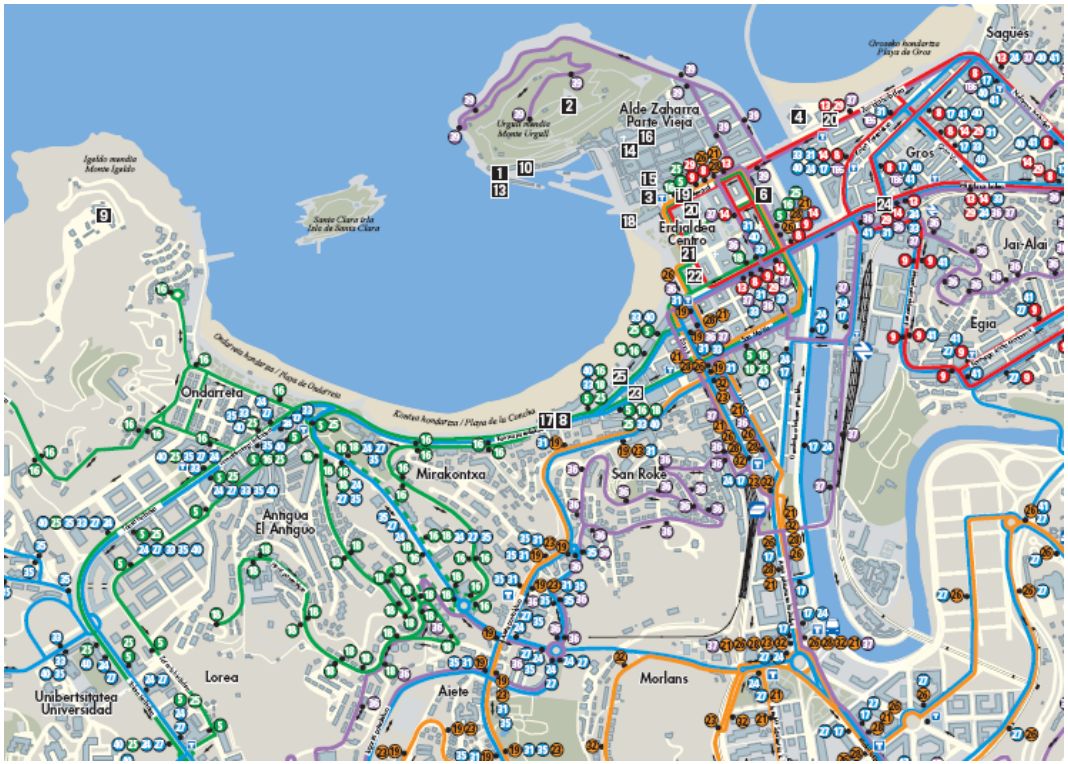
Optimization: Vehicle Routing Problem with TW



- Delimited time
- Lots of constraints
- Route
- Minimizing costs

- Techniques:
  - Simulated Annealing
  - Genetic Algorithms
  - Greedy – Search

Optimization: Personalized Electronic Tourist guide



- MCTOPTW
- Real time
- Adaptive system
- Maximizing User experience

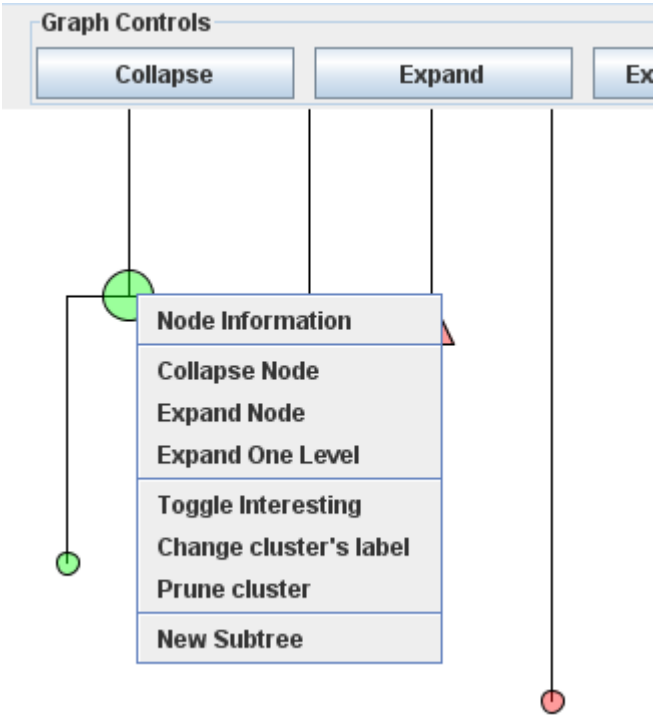
- Techniques:
  - Iterative local search
  - Time dependent algorithms

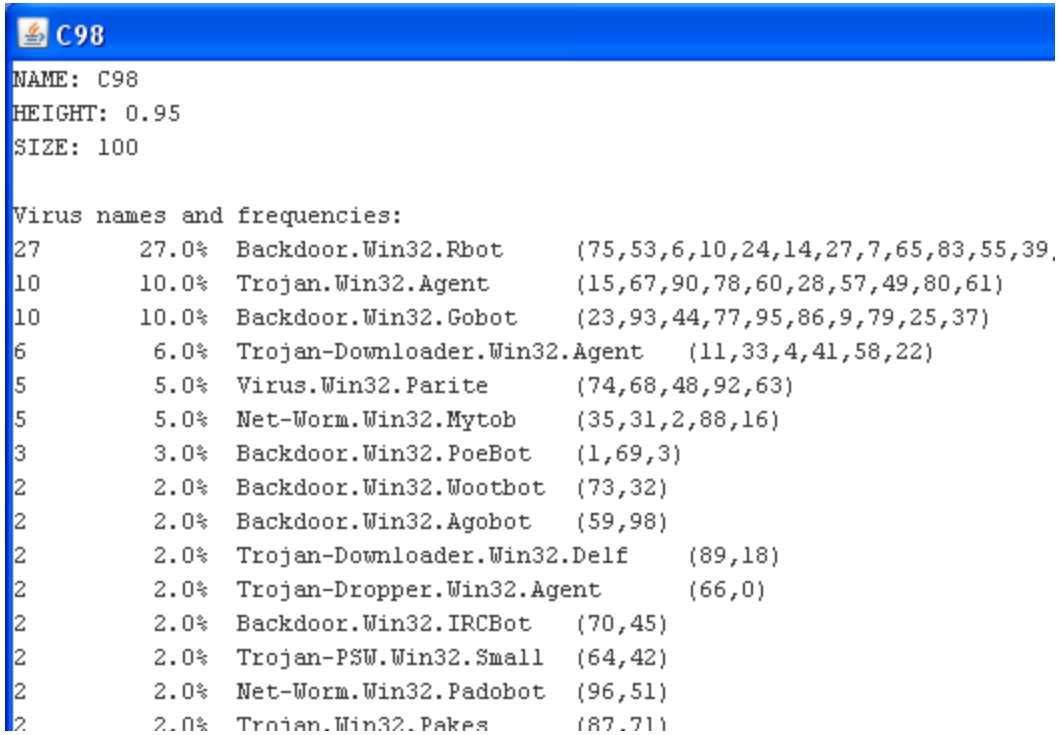
MALBEC: malware behaviour modelling

- Computer security
  - Intrusion detection systems (Unsupervised anomaly detection)
  - Malware Classification (Clustering) --> MALBEC

MALBEC: malware behaviour modelling

- MALBEC platform



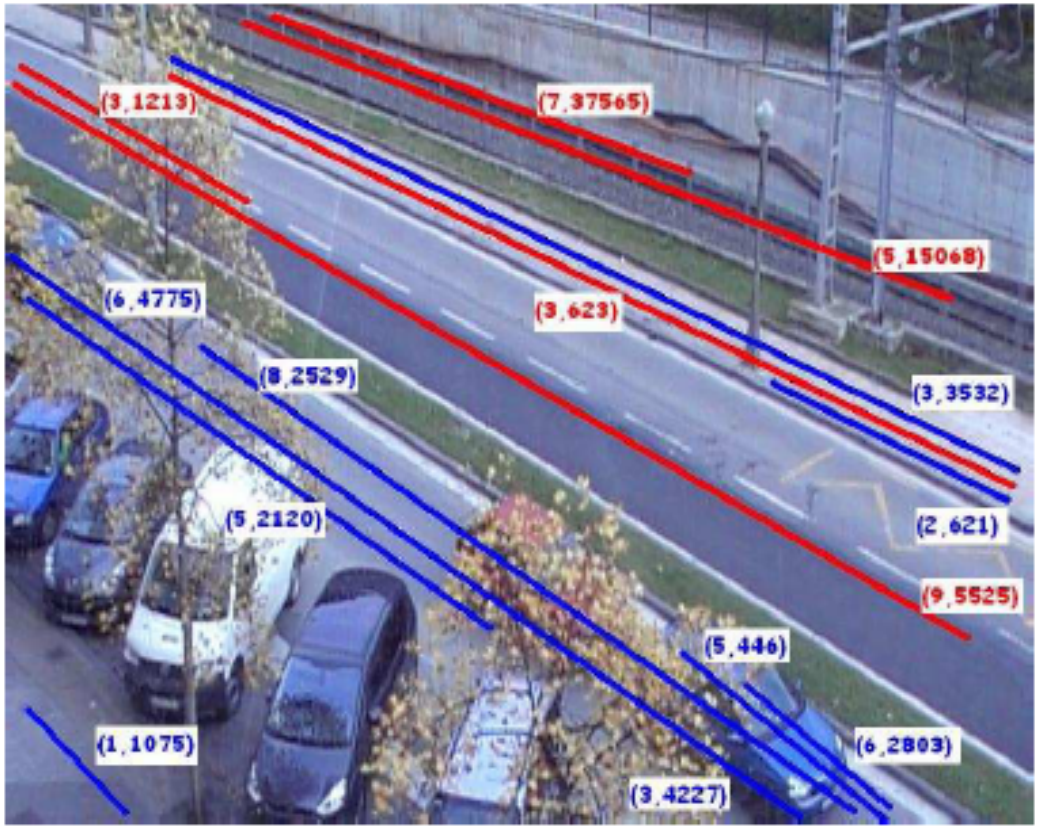


C98

```
NAME: C98
HEIGHT: 0.95
SIZE: 100

Virus names and frequencies:
27        27.0%  Backdoor.Win32.Rbot      (75,53,6,10,24,14,27,7,65,83,55,39,
10        10.0%  Trojan.Win32.Agent       (15,67,90,78,60,28,57,49,80,61)
10        10.0%  Backdoor.Win32.Gobot     (23,93,44,77,95,86,9,79,25,37)
6          6.0%  Trojan-Downloader.Win32.Agent   (11,33,4,41,58,22)
5          5.0%  Virus.Win32.Parite       (74,68,48,92,63)
5          5.0%  Net-Worm.Win32.Mytob     (35,31,2,88,16)
3          3.0%  Backdoor.Win32.PoeBot    (1,69,3)
2          2.0%  Backdoor.Win32.Wootbot   (73,32)
2          2.0%  Backdoor.Win32.Agobot    (59,98)
2          2.0%  Trojan-Downloader.Win32.Delf    (89,18)
2          2.0%  Trojan-Dropper.Win32.Agent      (66,0)
2          2.0%  Backdoor.Win32.IRCBot    (70,45)
2          2.0%  Trojan-PSW.Win32.Small   (64,42)
2          2.0%  Net-Worm.Win32.Padobot   (96,51)
2          2.0%  Trojan.Win32.Pakes       (87,71)
```

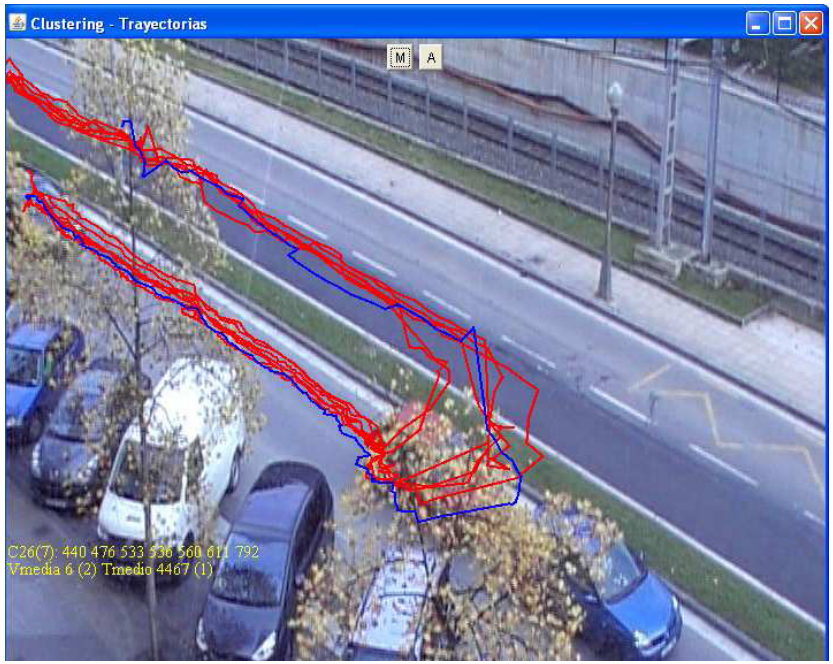
CAMIKER: object movement modelling

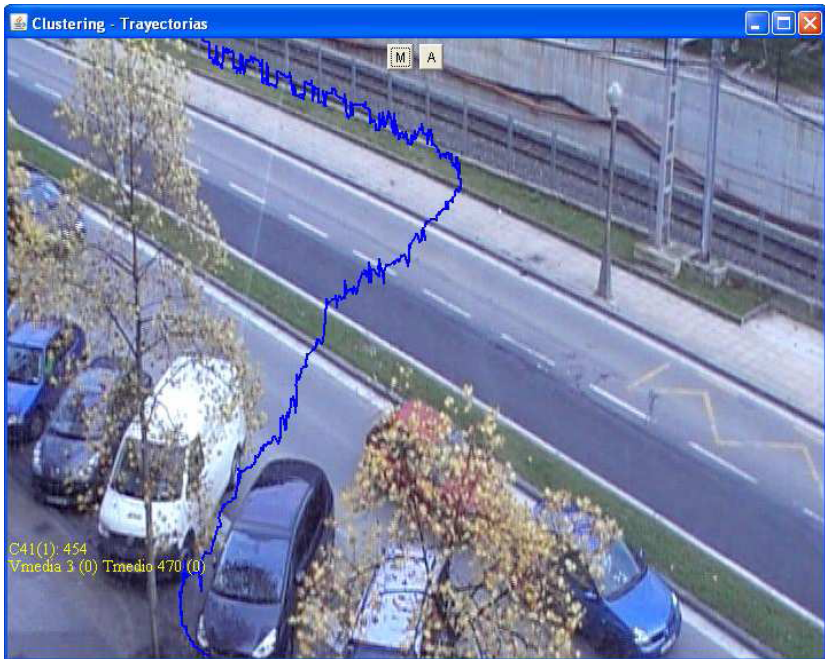


- Movement modelling
- Anomaly detection

- Techniques:
  - Clustering
  - Time-series

CAMIKER: object movement modelling





- Techniques:
  - Clustering
  - Time-series

ALDAPA - Algorithms, Data Mining and Parallelism

Previous lines

>> **NEUFODI:** ANN for detection and diagnosis
>> **OCR – FORM-LESS:** automatic character recognition
>> **Optimization:** VRPTW, PET
>> **MALBEC:** malware behaviour modelling
>> **CAMIKER:** object movement modelling

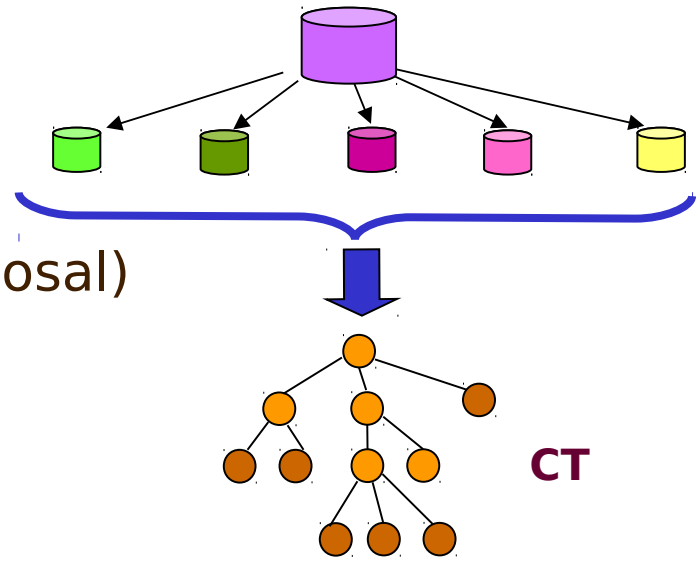
Current lines

>> **CTC – HARITZA:** comprehensibility, class imbalance
>> **MODELACCESS:** web user modelling, social web mining
>> **Physiological CS:** BCI, ECG

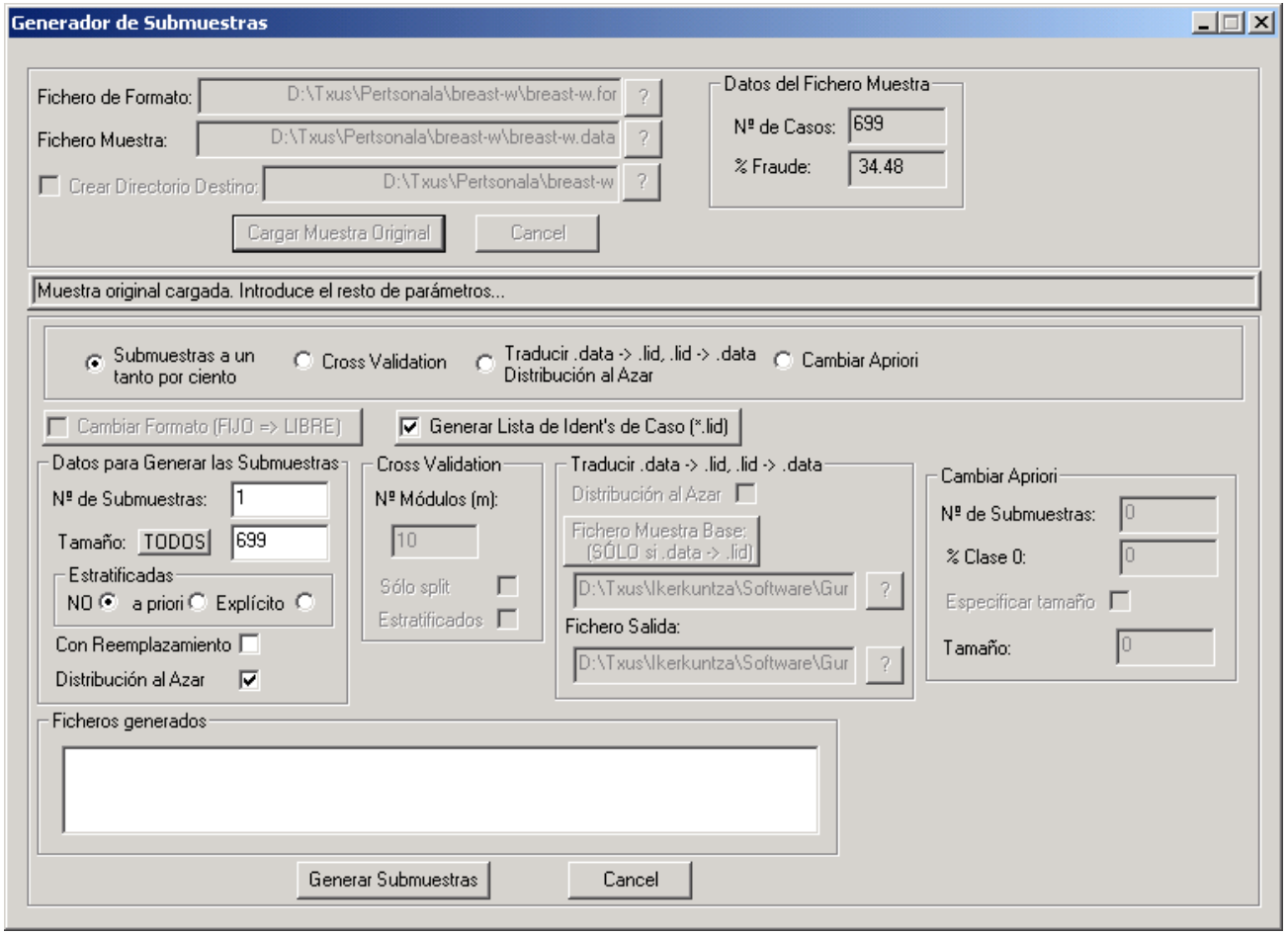
>> **HPC – Parallelism:** material physics, gp-gpu

CTC – HARITZA: comprehensibility, class imbalance

- Need of machine learning techniques that
  - Produce a **comprehensible** models
  - Able to deal with **class imbalance problems**
- Supervised techniques:
  - Classification trees (DT)
  - Rule induction, ...
  - **Consolidated trees: CT** (own proposal)
    - Benefit of combining **multiple subsamples** but without loss of **explaining** capacity (single tree)



Haritza Platform – GUI



Haritza Platform – Scripts

2.3.CTC Boot.scp - Bloc de notas

Archivo Edición Formato Ayuda

```
SCRIPT_COM

##########################################################################################
#                                        M000
#========================================================================================
#                           Generación CT con 30 Submuestras Bootstrap
#----------------------------------------------------------------------------------------
EXPERIMENTO_COM

DIRECTORIO_TRABAJO: I:\breast-w\M000\

GENERADOR_SUBMUESTRAS_COM
        Crear_Directorio:       Boot\
        FICHERO_DATOS:          breast-w+Sin000+0.data
        FICHERO_FORMATO:        ..\breast-w.for

        Tipo_generador:         Muestras
        Numero_muestras:        30
        Tam_muestras:           EL_MISMO
        Estratificacion:        NO

        Modo_Extraccion:        CON_REEMPLAZO
        Distribuir_al_Azar:     NO

        Modo_nombre_fichero:    automatico
        nombre_fichero:         Boot\breast-w+*+0.data
        num_ficheros:           30
        Crear_directorios:      NO
GENERADOR_SUBMUESTRAS_FIN

CARGAR_MUESTRA_COM
        Tipo_Arbol:             CONSOLIDADO
        Num_Ficheros:           30
        FICHERO_DATOS:          Boot\breast-w+*+0.data
        FICHERO_FORMATO:        ..\breast-w.for
CARGAR_MUESTRA_FIN

CONSTRUIR_ARBOL_CONSOLIDADO_COM
        Guardar sin_datos CT_30_Boot.arb
        Fich_Training_Base: breast-w+Sin000+0.data
CONSTRUIR_ARBOL_CONSOLIDADO_FIN

PODAR_ARBOL_COM
        Nom_Fic_data: breast-w+Sin000+0.data
        Nom_Fic_Arb_Podado: CT_30_Boot_pruned.arb
PODAR_ARBOL_FIN

EXPERIMENTO_FIN

##########################################################################################
#                                        M001
#========================================================================================
#                           Generación CT con 30 Submuestras Bootstrap
```

MODELACCESS: web user modelling

- **Objective** in any context:
  - To **adapt web** pages to the need of the **users**
- Adaptation becomes especially **critical** when the users have **special needs**
- Contexts:
  - Tourism website – Bidasoa Turismo
  - Discapnet - ONCE

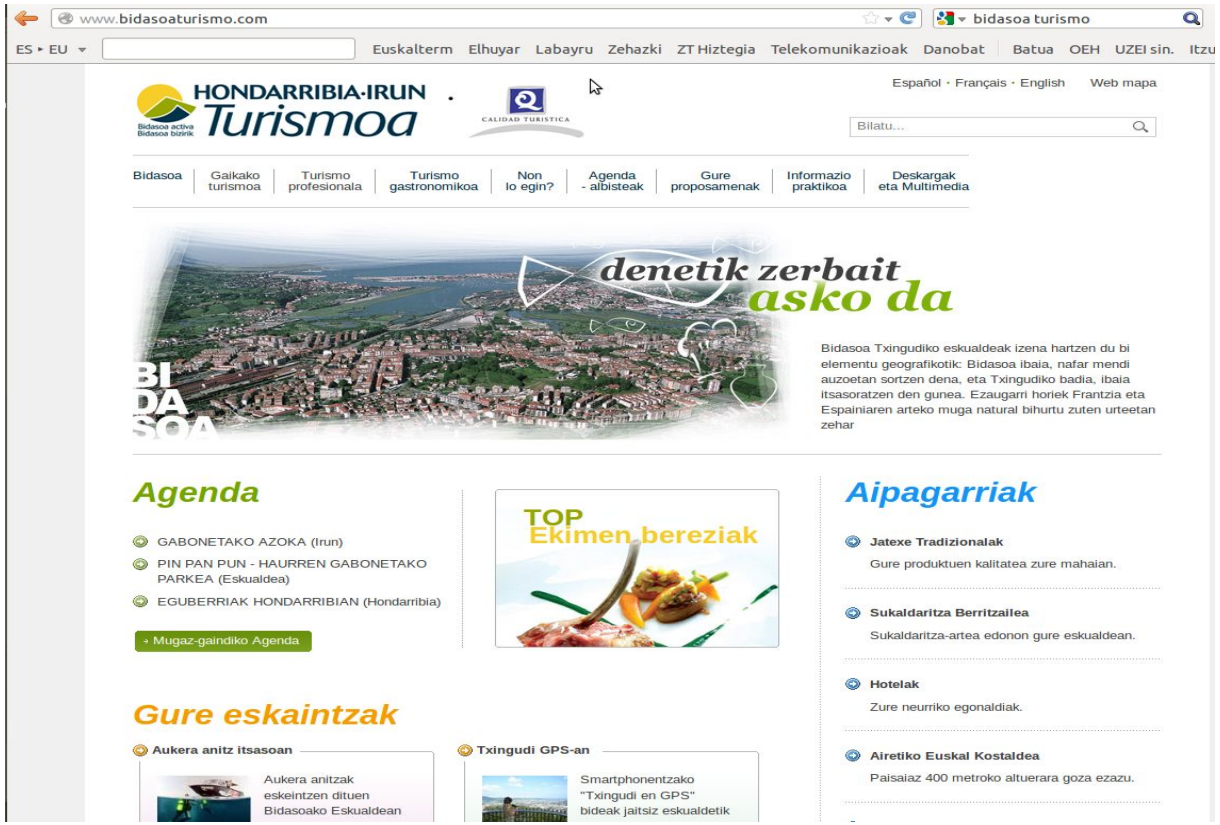
MODELACCESS: web user modelling

- **Web mining:** the application of data mining techniques to the Web data
  - **Usage** mining
  - **Content** mining
  - **Structure** mining

- Phases:
  - Data acquisition and **preprocessing**.
  - **Pattern discovery** and analysis: *Machine Learning*.
  - **Exploitation**.

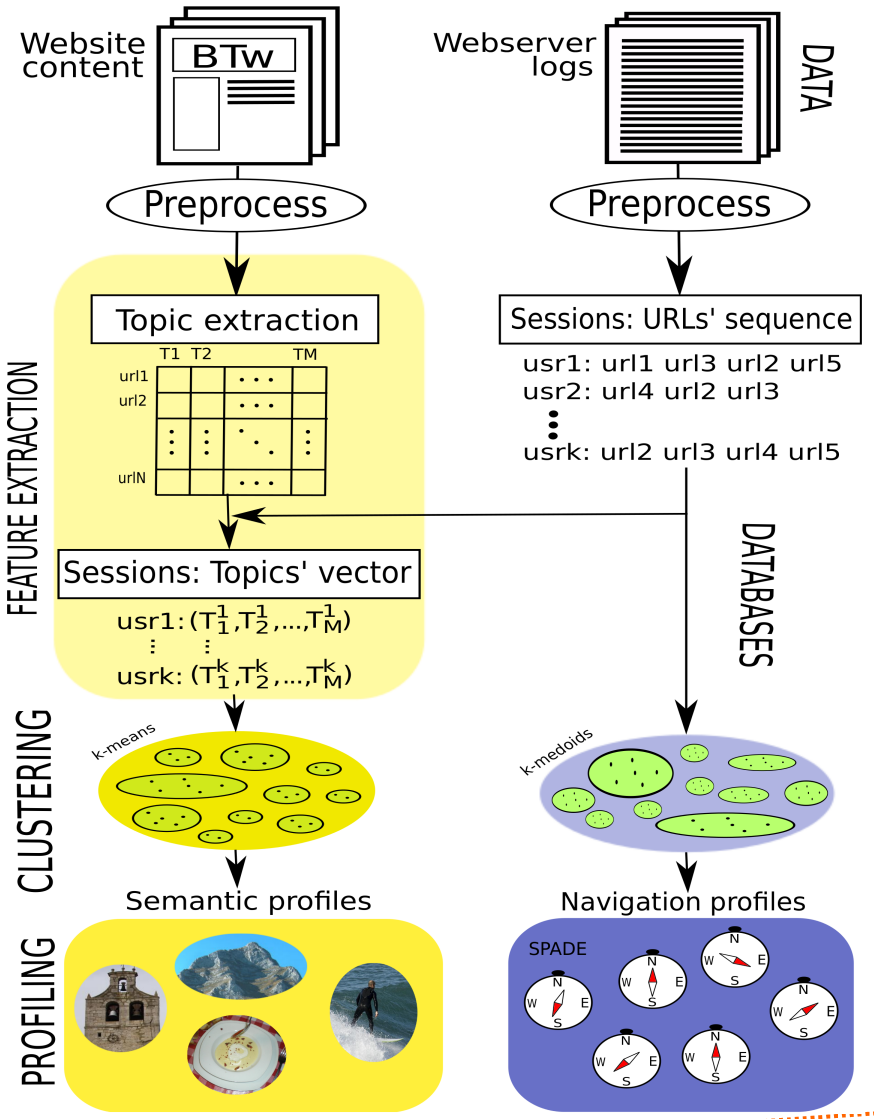
MODELACCESS: web user modelling

- **bidasoa turismo**



- Personalization of web navigation
- Knowledge about interest of users

MODELACCESS: web user modelling



- Techniques:
  - Clustering
  - Frequent Pattern Mining
  - Topic Modelling

MODELACCESS: web user modelling

- **discapnet**

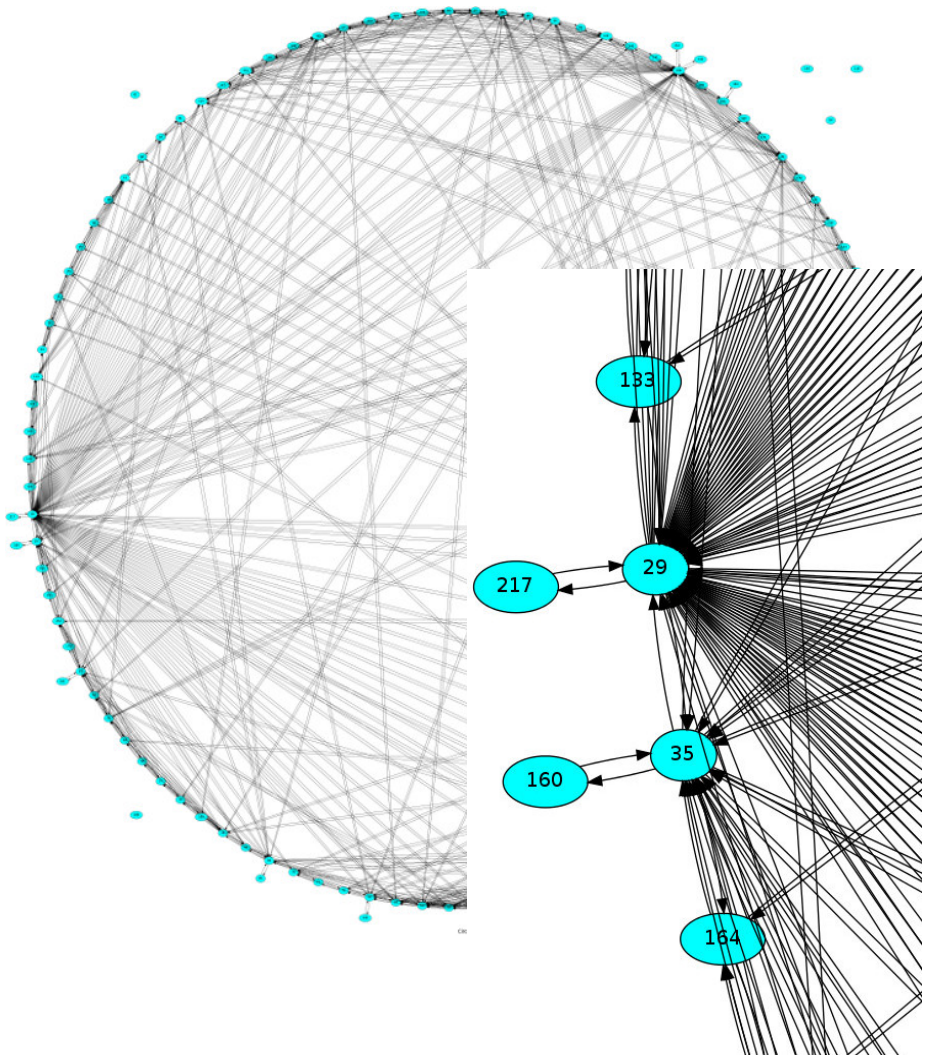
- Problem detection (user, web)
- Web personalization

- Techniques:
  - Supervised and unsupervised algorithms

MODELACCESS: social web mining

elegune

- **guremintza** – GUREAK
- social network adapted for users with special needs
- expansion stage
- social network mining



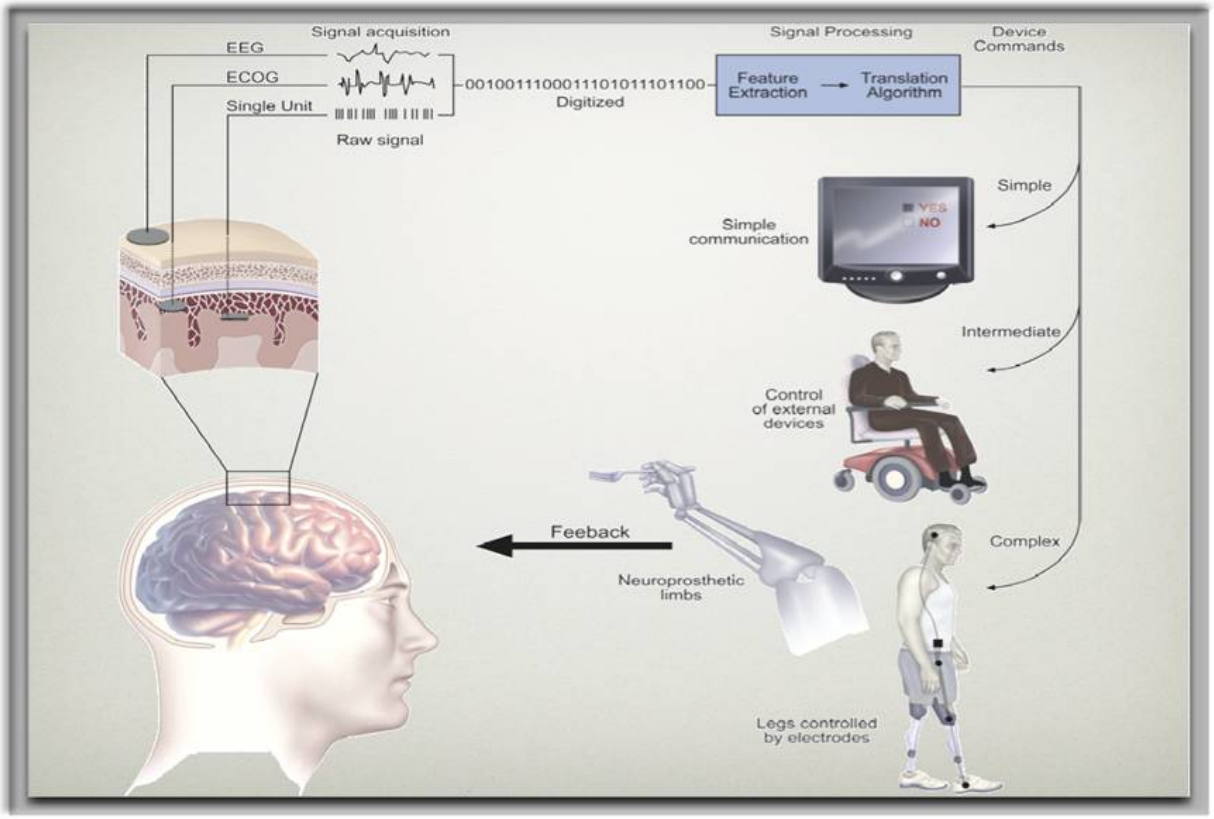
MODELACCESS: social web mining

elegune

- guremintza – GUREAK

Physiological Computing Systems: BCI, ECG

- **Brain Computer Interface**
  - From electroencephalography signal (EEG) to command devices



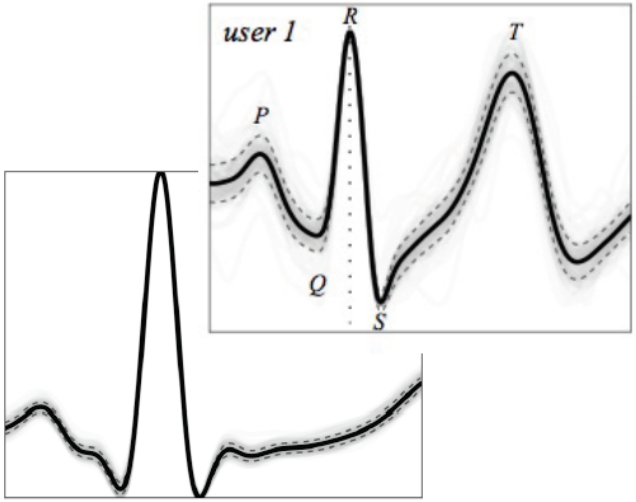
From https://www.etsu.edu/cas/bcilab/

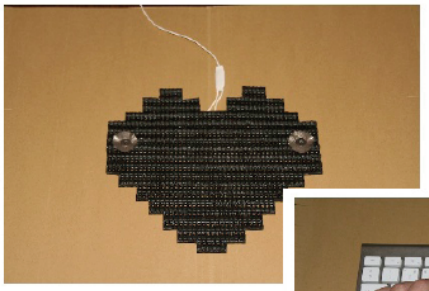
Physiological Computing Systems: ECG

Behavioral Biometrics (ii)

Next generation one-lead setup

- Hand **palms/fingers** sensor
- **Two electrodes** (no gel)
- Integratable in everyday items





Recognition rates (identification)

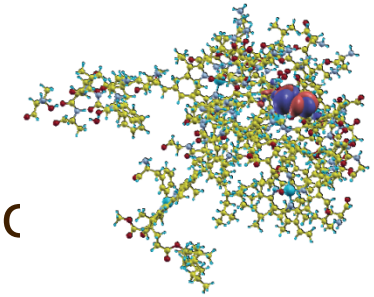
- 60 subjects
- >94±0.9%, with 4s or less

Hugo Silva, Instituto de Telecomunicações

HPC – Parallelism: material physics, gp-gpu

- **High Performance Computing** (HPC)

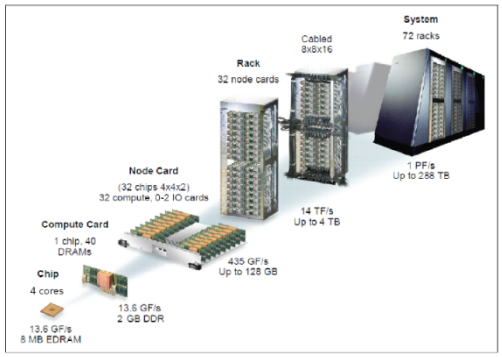
  Collaboration with Nano-Bio Spectroscopy group (physicist)

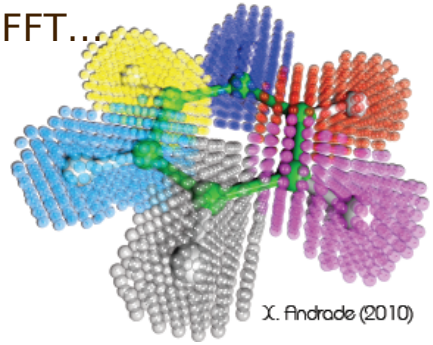


  - Reach **petaflop** computing with a scientific code → analysis of performance and scalability
    - Simulate photosynthesis of the light in chlorophyll (**OCTOPUS** code: http://www.tddft.org/programs/octopus/)
  - Machines

    Vargas, Mare Nostrum, Juguene…
  - Multi-level parallelism (MPI, OpenMP, **GP-GPU**)

    Scientific libraries: BLAS, LAPACK, FFT…





Eskerrik asko
Muchas gracias
Thanks for your attention

http://www.sc.ehu.es/aldapa