El papel de la Estadística en las Ciencias Biomédicas

Santander, 14 de mayo de 2008

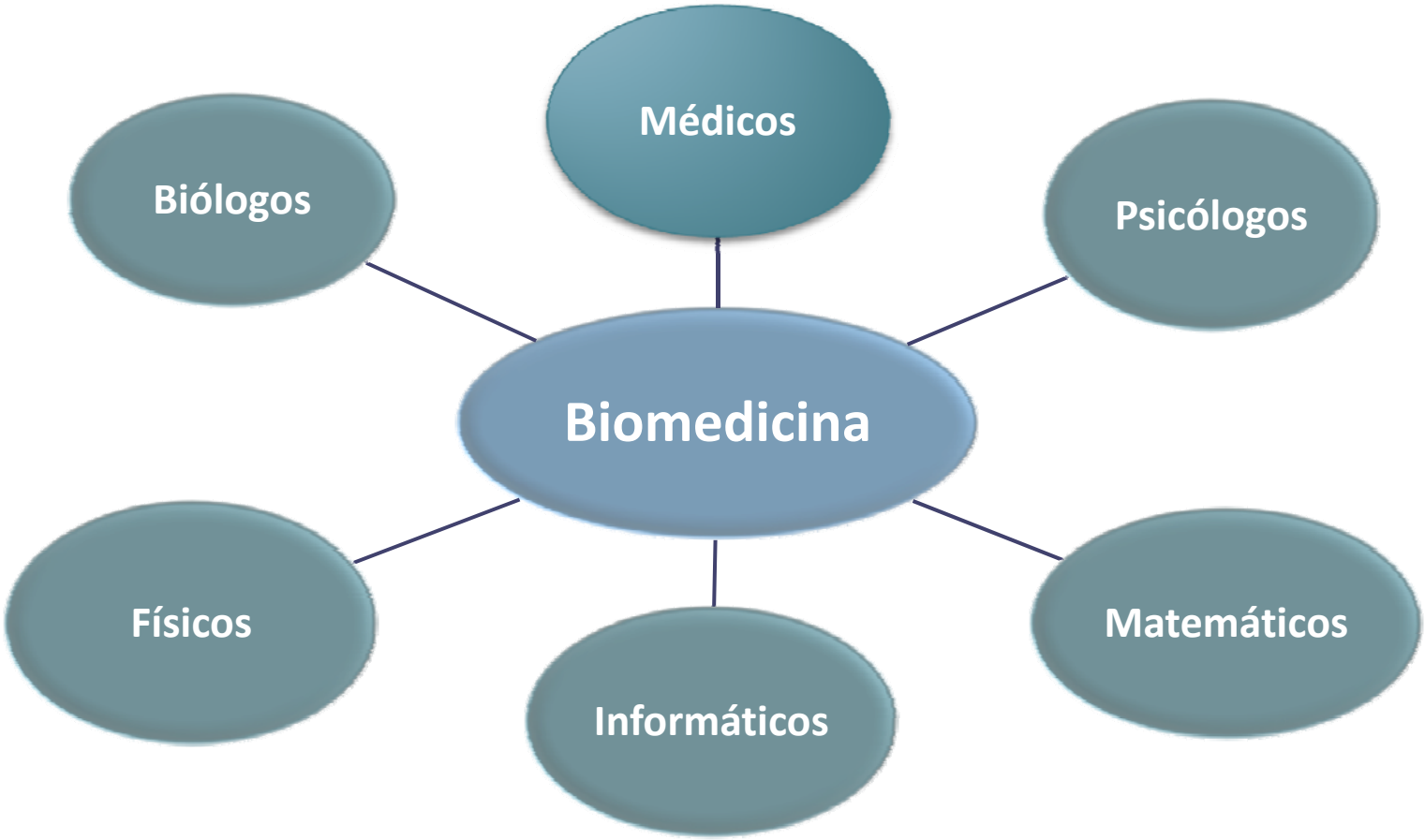
Carmen Cadarso-Suárez

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

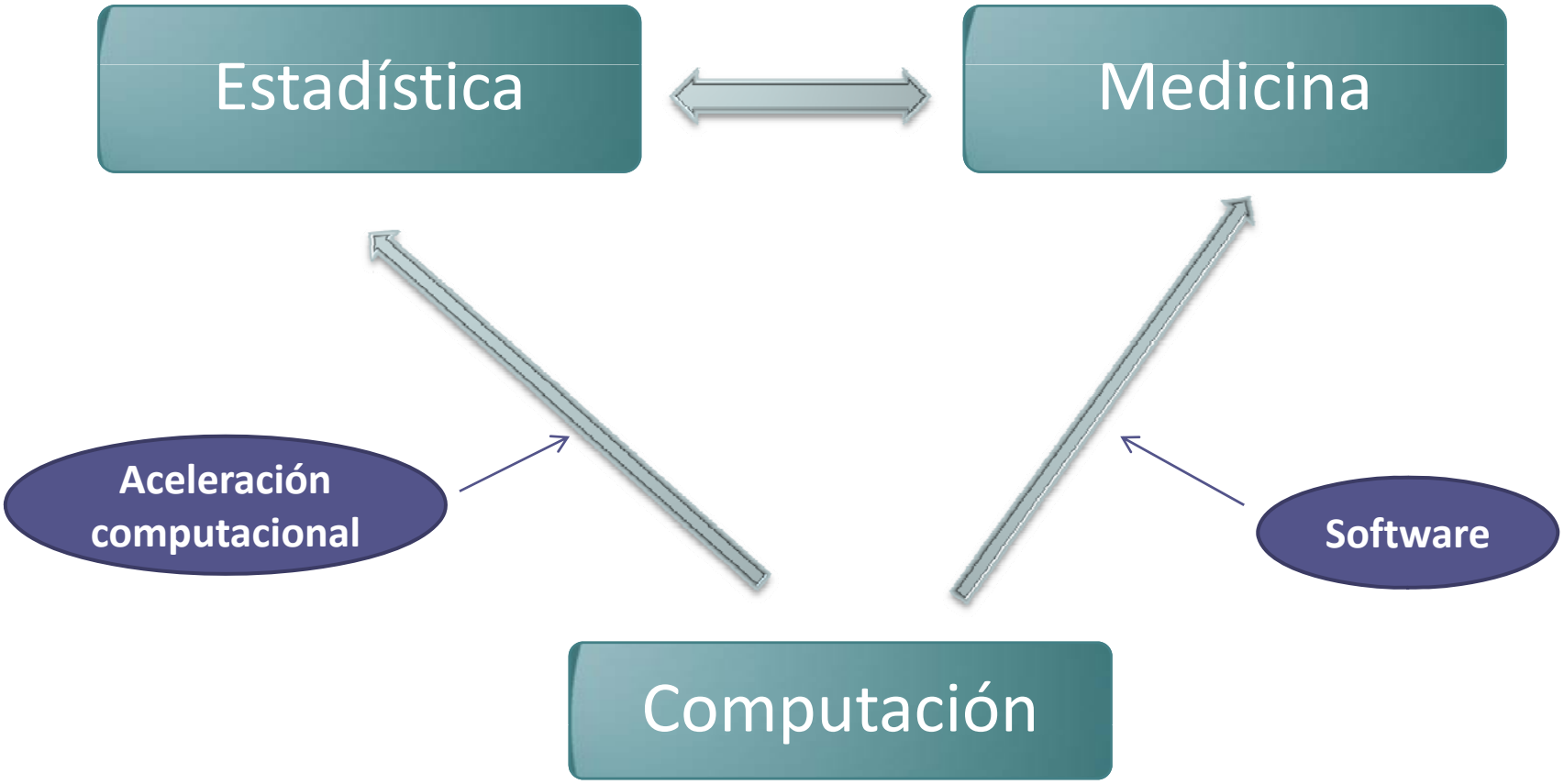
Unidad de Bioestadística
Facultad de Medicina
Departamento de Estadística e IO

La investigación biomédica

Equipos multidisciplinares



Bioestadística actual



Zelen M. (2006). Biostatisticians, Biostatistical Science and the Future. *Stat Med*, **25**, 3409-3414.

Lugares de trabajo

Facultad de Medicina (USC)

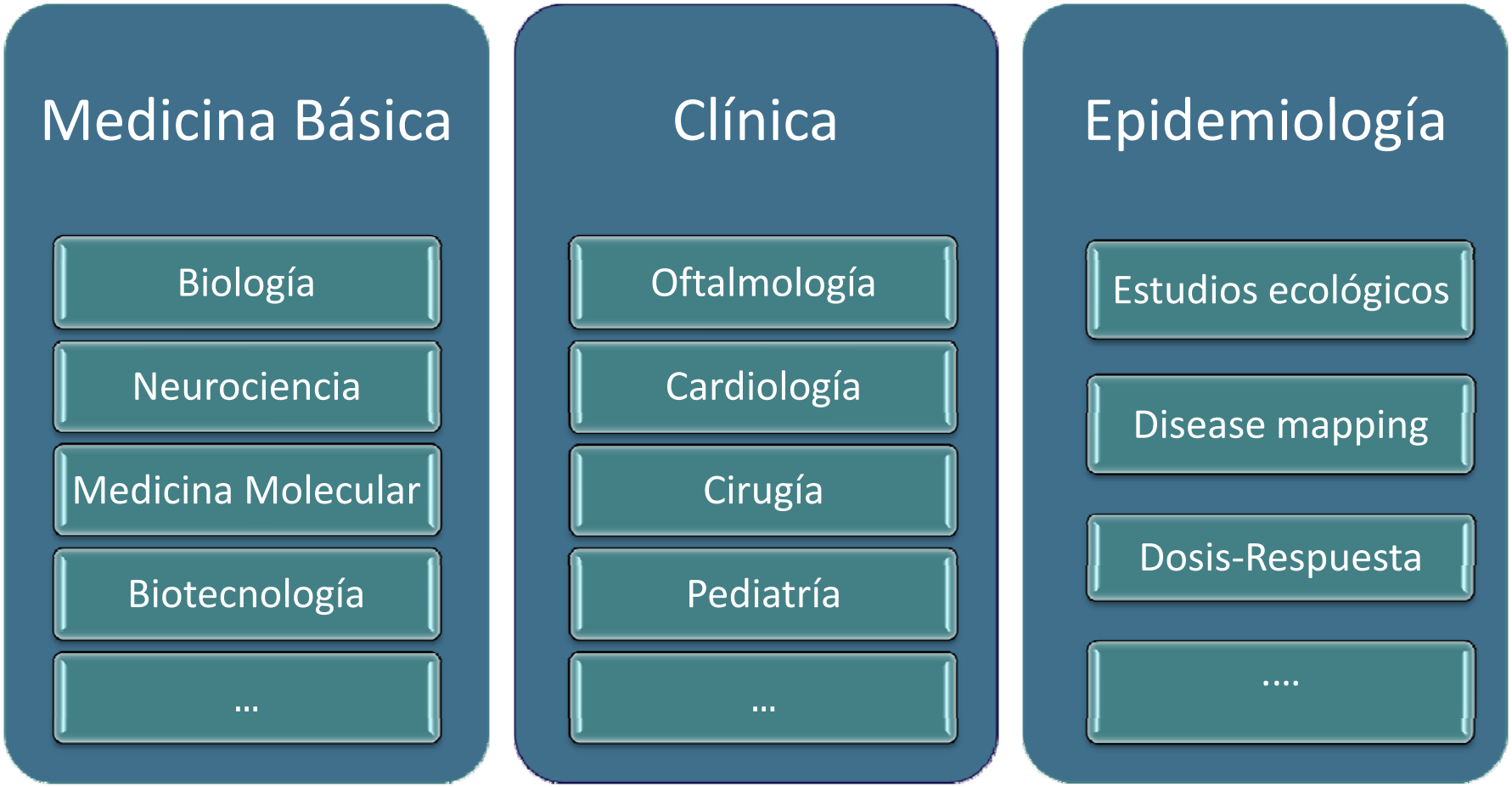
Facultad de Matemáticas (USC)

C. Hospitalario Clínico Universitario

Consellería de Sanidade (Xunta de Galicia)

Medicina:

"Arte de prevenir y curar enfermedades"



Tipos de estudios biomédicos

Asociación

- **Epidemiología y Clínica**
  - Factor de riesgo /enfermedad: Medidas: “OR”, “RR”
  - Factor de riesgo/mortalidad: “Hazard Ratio (HR)”
- **Neurociencia**
  - Asociación entre la actividad neuronal y estímulos

Clasificación

- **Diagnosis clínica**
  - Búsqueda de marcadores de enfermedades
- **Diagnóstico asistido por ordenador**
  - Evaluación y mejora de sistemas CAD

Predicción

- **Prognosis clínica**
  - Pronóstico de enfermedades y supervivencia.
- **Medicina Forense**
  - Predicción de la fecha de muerte en homicidios

EPIDEMIOLOGÍA

“Factores de riesgo y enfermedad”

Investigadores colaboradores: médicos y epidemiólogos

Departamento de Salud Pública (USC)
Complejo Clínico-Universitario de Santiago de Compostela
Dirección Xeral de Saúde Pública (Xunta de Galicia)

Medida de efecto: Riesgo Relativo (RR)

E = enfermedad F =factor de exposición

Estudio de cohortes (incidencia)

$$RR_{F/\bar{F}} = \frac{p\left(E / F\right)}{p\left(E / \bar{F}\right)}$$

$RR = 1$, no asociación entre E y F

$RR > 1$, F factor de riesgo

$RR < 1$, F factor protector

Medida de efecto: Odds-Ratio (OR)

Estudio caso-control

$$Odds(E) = \frac{p(F/E)}{p(\overline{F}/E)} \qquad Odds(\overline{E}) = \frac{p(F/\overline{E})}{p(\overline{F}/\overline{E})}$$

$$OR_{F/\overline{F}} = \frac{Odds(E)}{Odds(\overline{E})}$$

$OR = 1$, no asociación entre E y F

$OR > 1$, F factor de riesgo

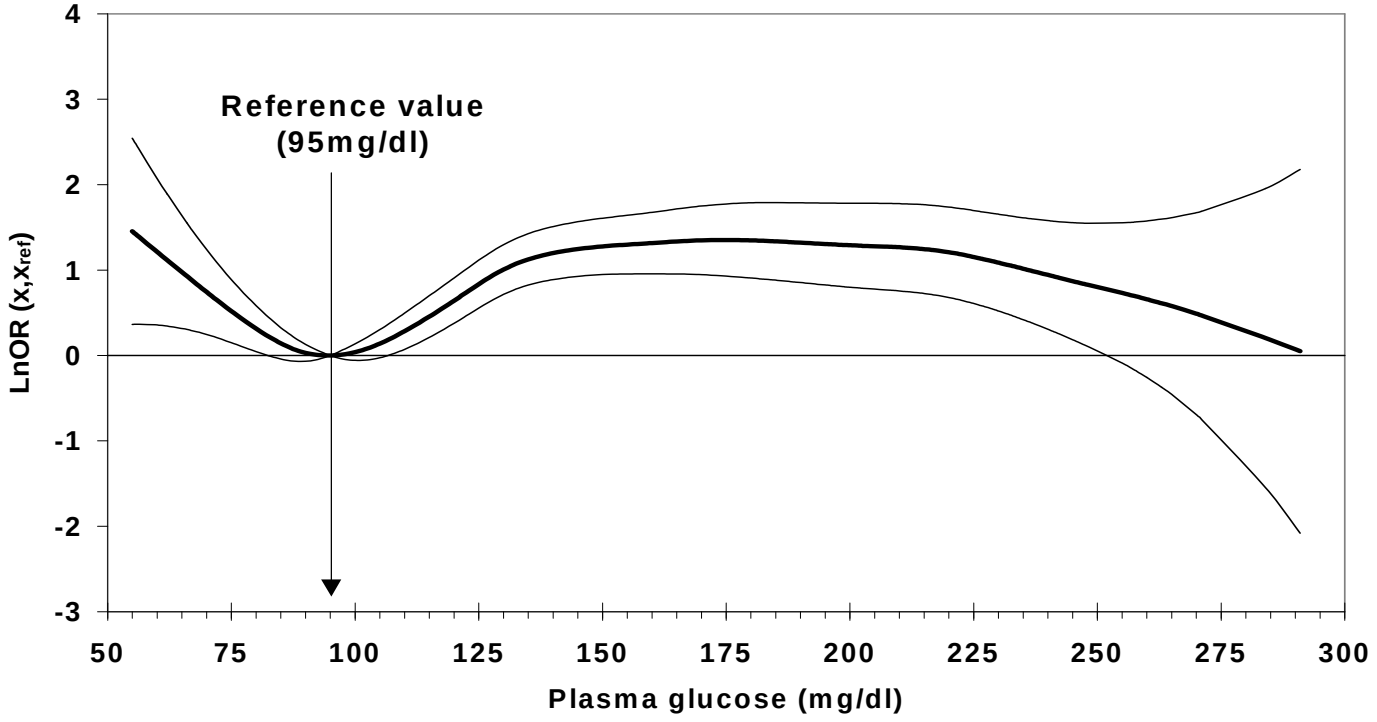
$OR < 1$, F factor protector

Factores de riesgo de Infección post-quirúrgica

La glucosa, ¿es un factor de riesgo para la infección post-quirúrgica?

Modelo Aditivo Generalizado logístico (logistic GAM, Hastie-Tibshirani, 1990)

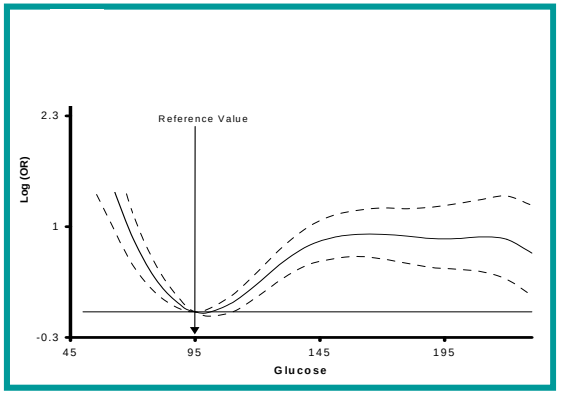
$$log\left(\frac{p\left(INFEC=1/Gluc\right)}{p\left(INFEC=0/Gluc\right)}\right)=logit=\beta_0+f\left(Gluc\right)$$

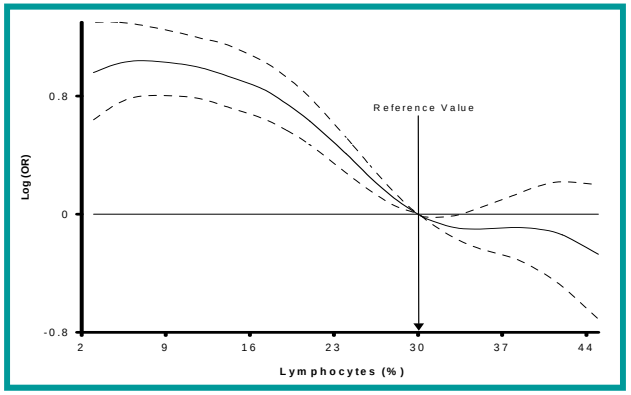


Modelos multivariantes

Efectos principales

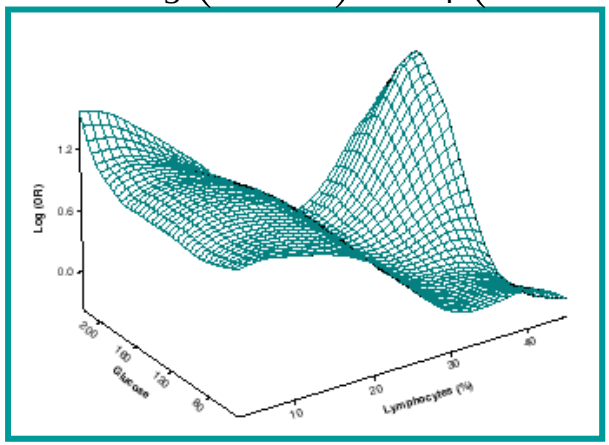
$$logit = \beta_0 + \beta_1 Sex + \beta_2 Age + f_3(Gluc) + f_4(Linf\%)$$





Interacción

$$logit = \beta_0 + \beta_1 Sex + \beta_2 Age + f_3(Gluc) + f_4(Linf\%) + f_{34}(Gluc, Linf\%)$$



NEUROCIENCIA

“Actividad neuronal y conducta”

Investigadores colaboradores: fisiólogos, psicólogos

Departamento de Fisiología(USC)
Laboratorio de Neurociencia Computacional (USC)

Actividad neuronal y conducta

En la Corteza Visual (en primates subhumanos) interesa estudiar:

- Asociación entre estímulos visuales y la respuesta neuronal:
  Tasa de descarga eléctrica
- Disparo simultáneo de varias neuronas:
  Sincronía neuronal
- Poblaciones de neuronas:
  Population-based neuron analysis.

Estos abordajes experimentales nos permiten conocer:

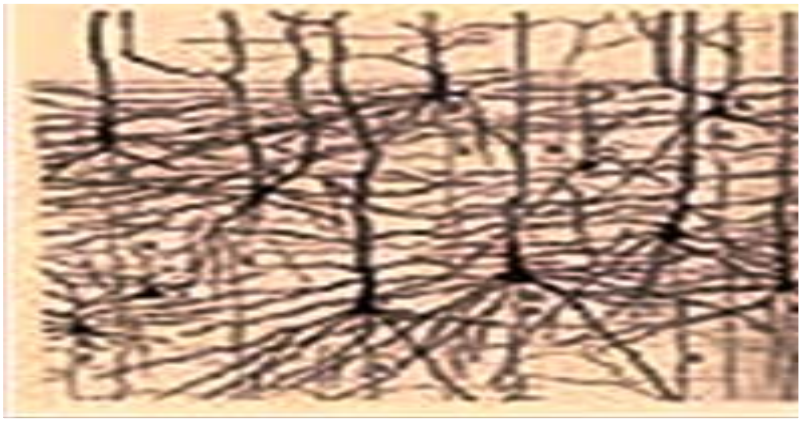
1. Cómo se procesa la **información sensorial** en el cerebro.
2. Cuál es la relación entre la **actividad neuronal** y la **conducta**.

“Doctrina” de la neurona

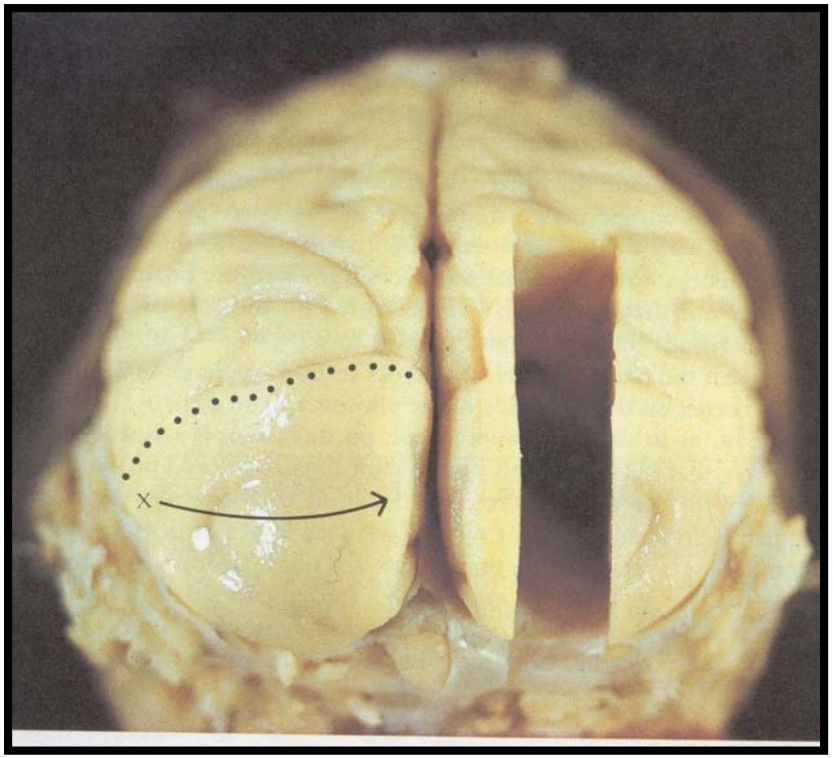
Ramón y Cajal (Premio Nobel, 1906)

La neurona es una entidad discreta que conduce corriente eléctrica.

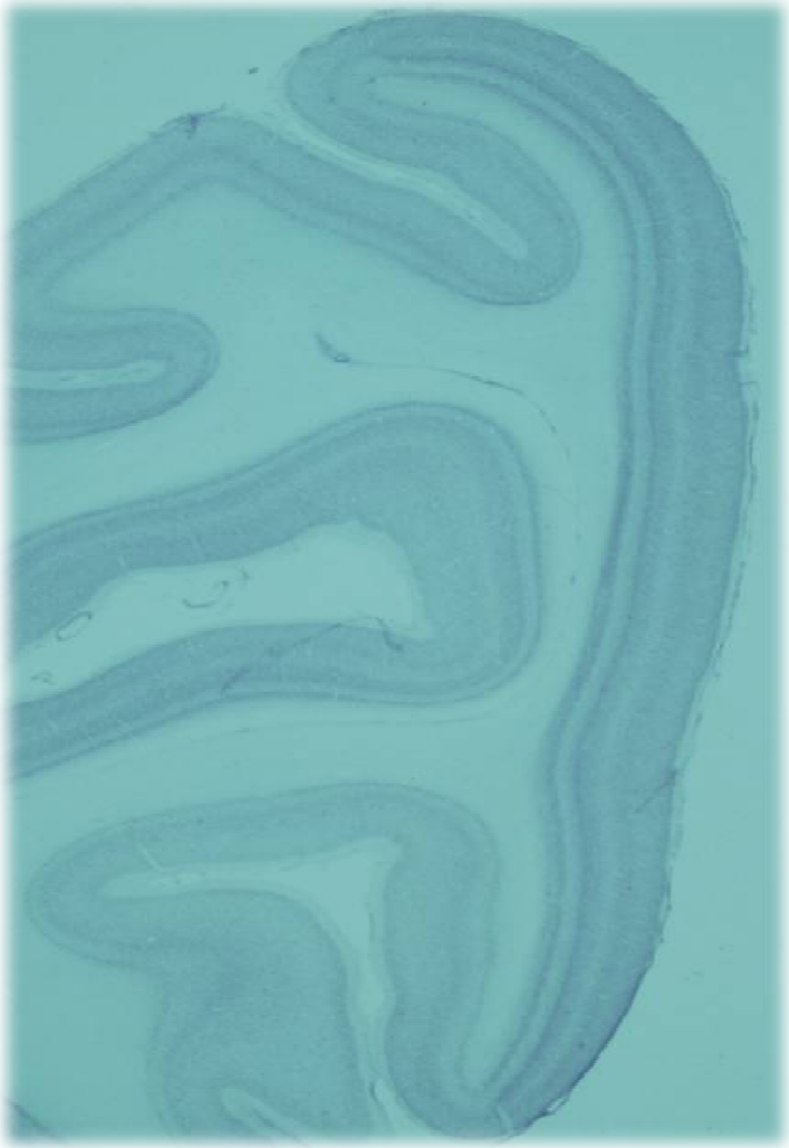
El **SNC está formado por neuronas que conectan entre sí por medio de *sinapsis* produciendo redes neuronales**



Visual Cortex (Área V1)



David H. Hubel (1988) Eye, Brain and Vision. Scientific American Library. New York.



Tarea de decisión en la orientación

¿derecha? ¿izquierda?

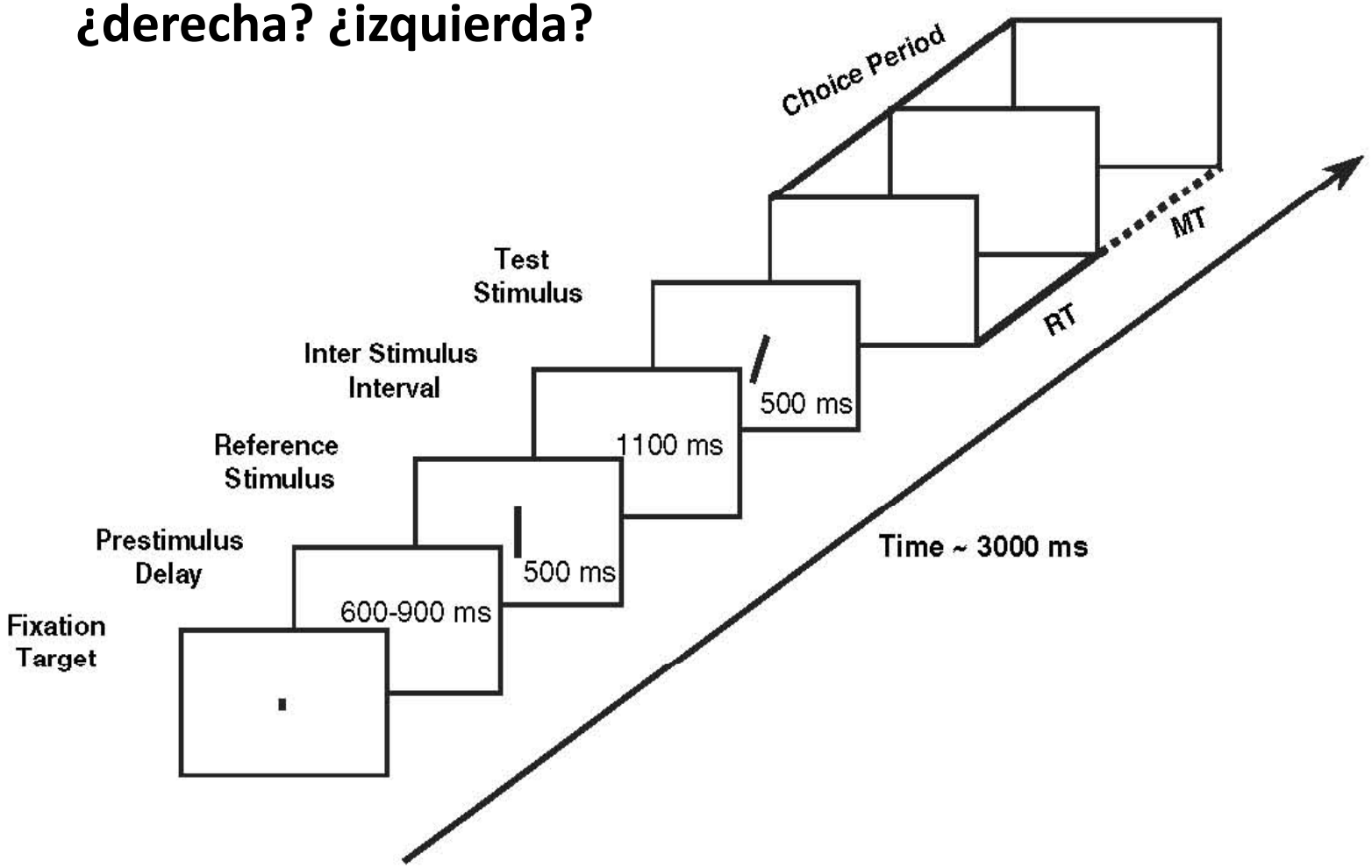
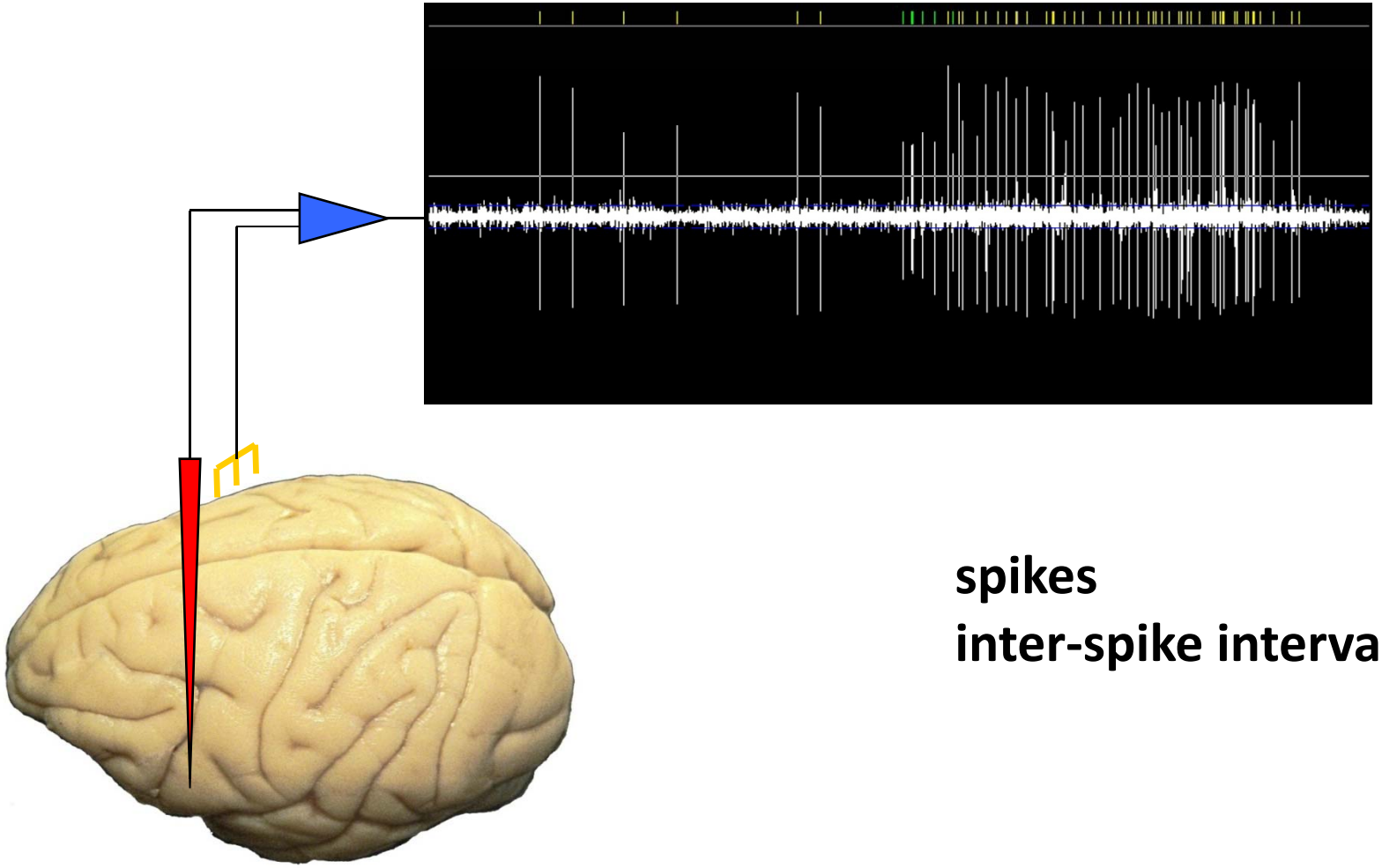


Fig. 1. Outline of the orientation discrimination task described in the text: RT, reaction time; MT, movement time

Codificación de la actividad neuronal

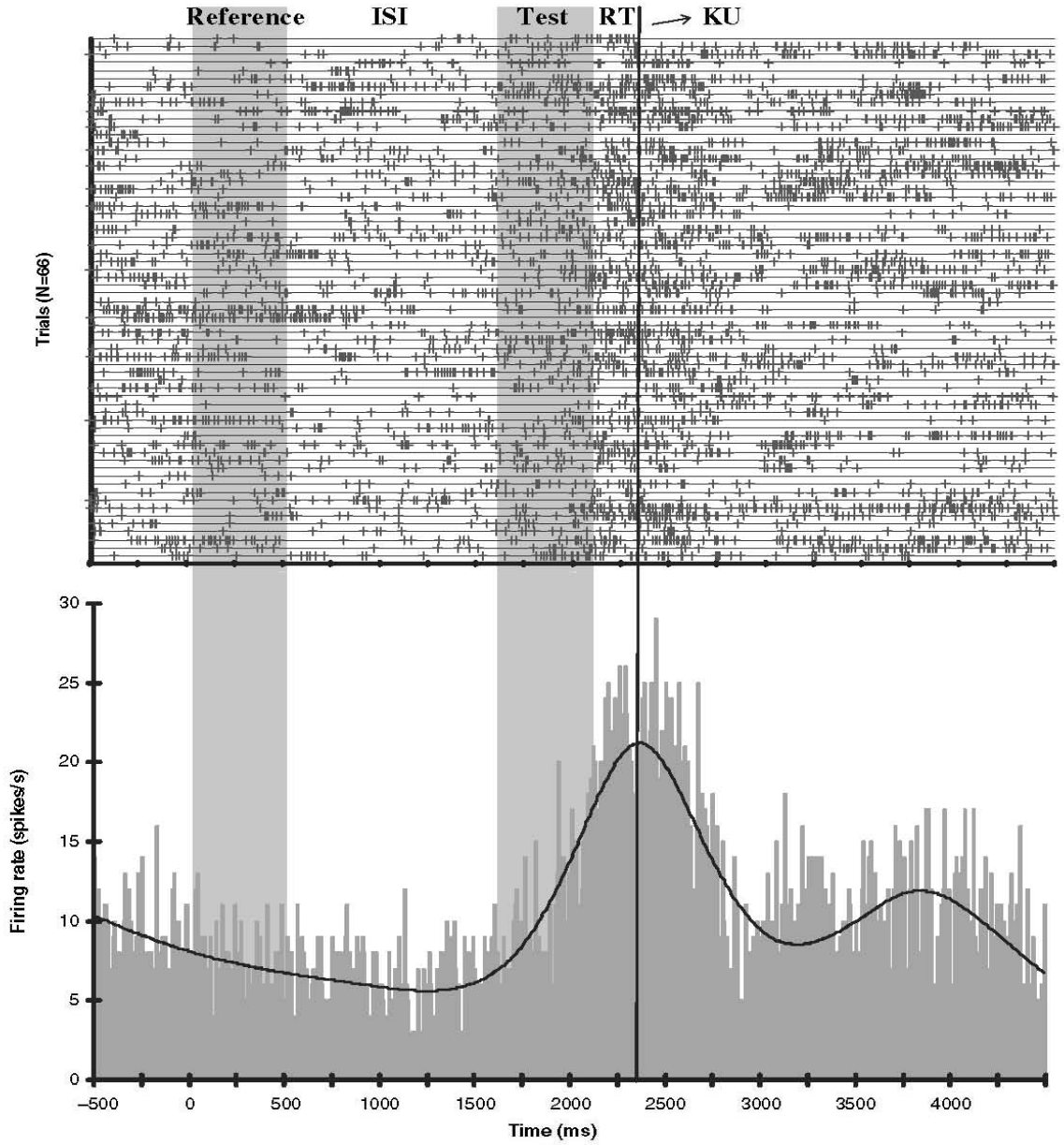


spikes
inter-spike intervals

Tasa de disparo neuronal (1 neurona)

Raster Plot

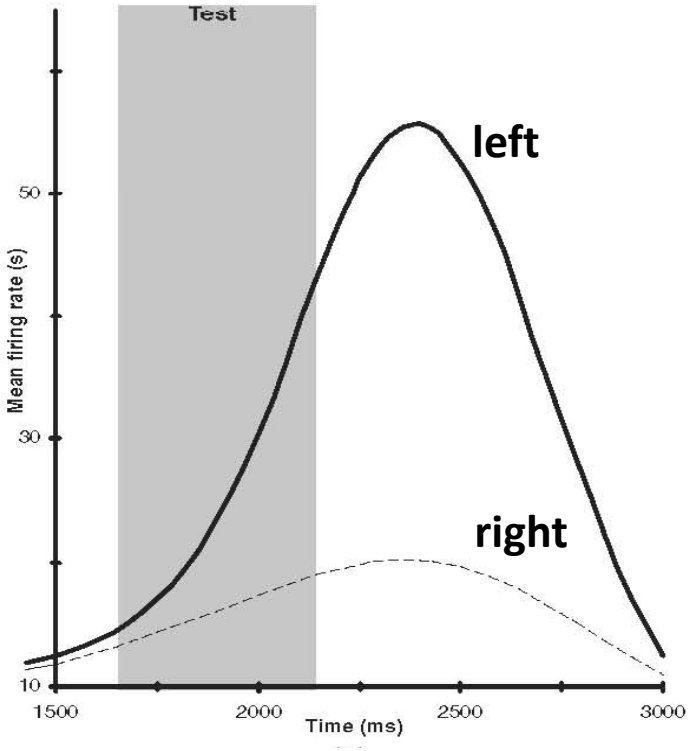
Peri-stimulous Histogram (PSTH)

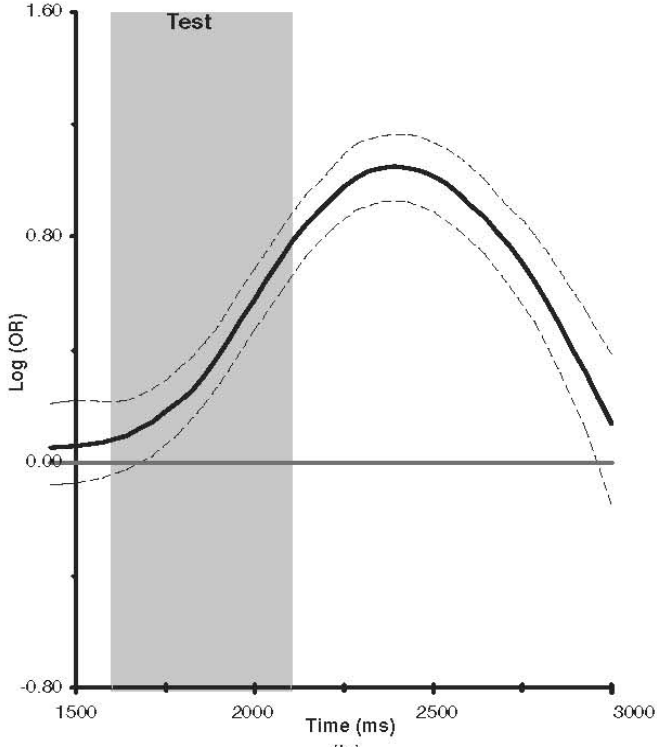


Comparación de tasas de disparo

$$\log\left\{\frac{p(\text{Orien},t)}{1-p(\text{Orien},t)}\right\}=\alpha+f(t)+f_0(t)I_{\{\text{Orien}=0\}}+f_1(t)I_{\{\text{Orien}=1\}}$$

$$\text{OR}(t)=\frac{p(1,t)/\{1-p(1,t)\}}{p(0,t)/\{1-p(0,t)\}}$$

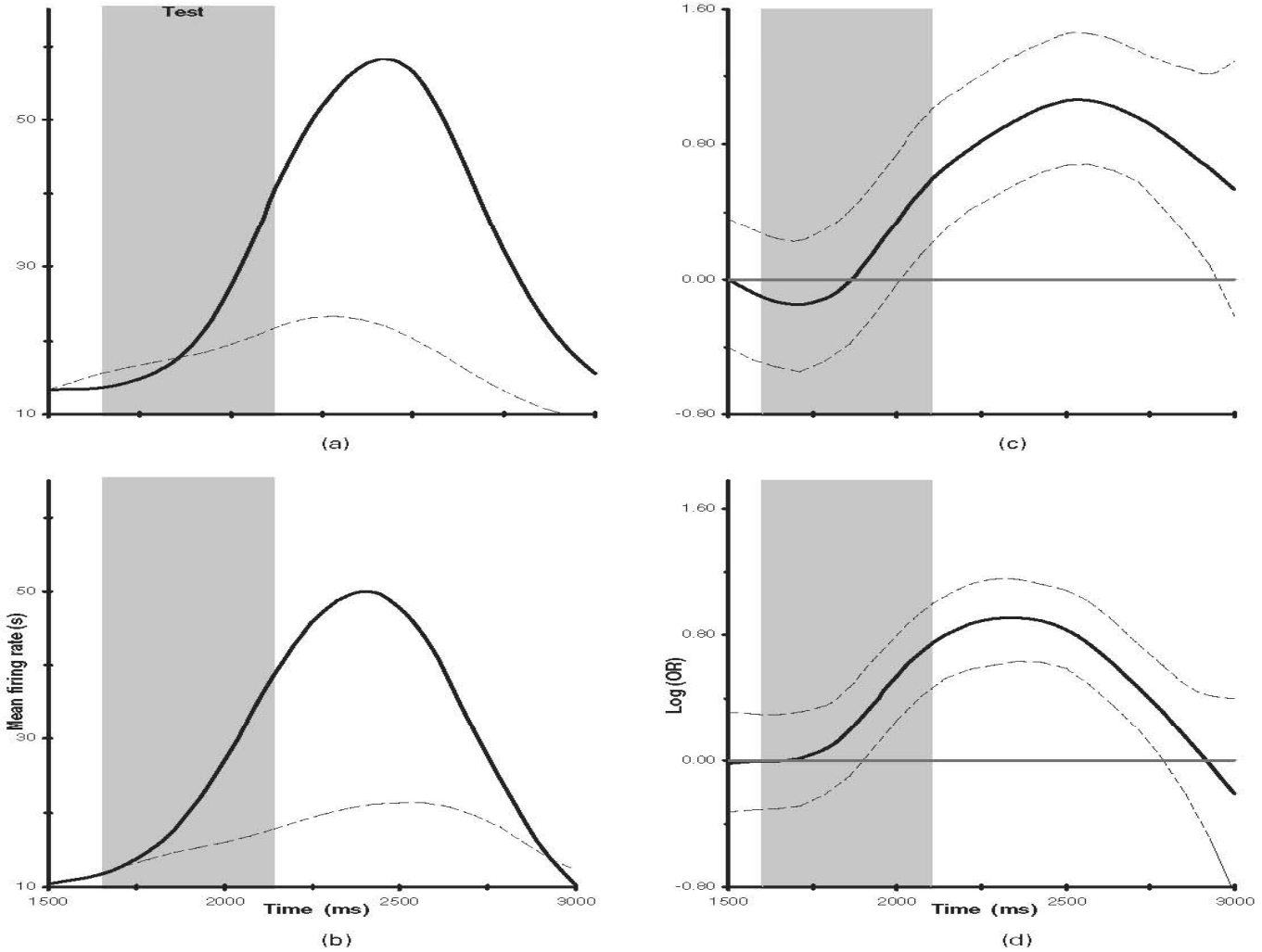




Comparación de tasas de disparo

Easy task

Difficult task



Sincronía neuronal

Tasa de disparo conjunta (2 neuronas) : Joint PSTH

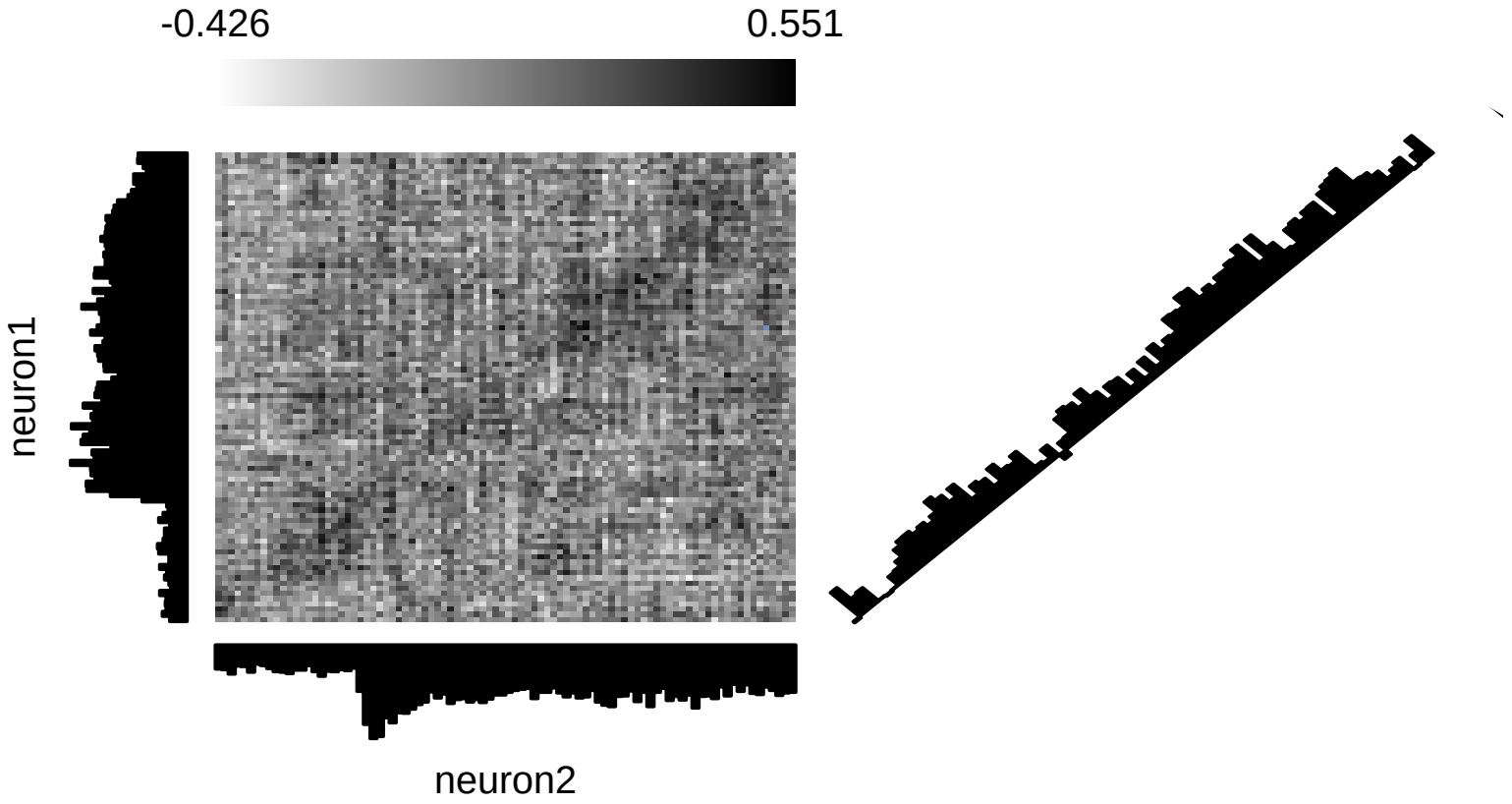
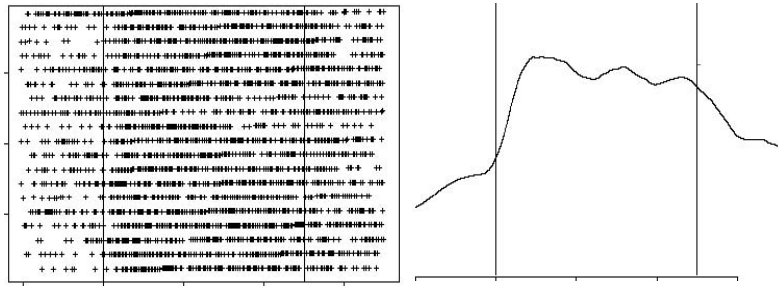


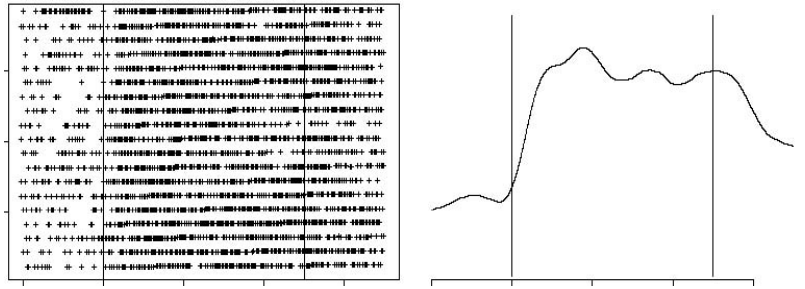
Figure 2: Raster plot (left) of spikes and corresponding peristimulus

Sincronía neuronal

Neuron #1

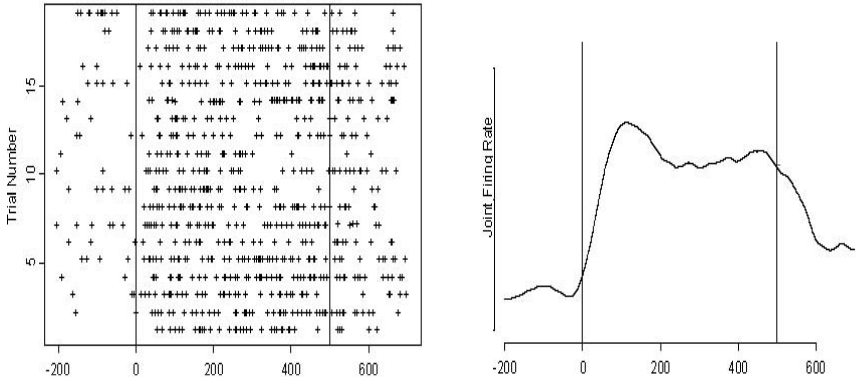


Neuron #2



Spike coincidence

		Neuron k		
		1	0	Totals
Neuron i	1	n_{11}	n_{10}	$n_{11}+n_{10}$
	0	n_{01}	n_{00}	$n_{01}+n_{00}$
Totals		$n_{11}+n_{01}$	$n_{10}+n_{00}$	$N=n_{11}+n_{10}+n_{01}+n_{00}$



Medida de Sincronía

Tasa marginal de la neurona 1: $\pi_{1+}(t) = p(Y_1 = 1/t)$

Tasa marginal de la neurona 2: $\pi_{+1}(t) = p(Y_2 = 1/t)$

Tasa conjunta de disparo: $\pi_{11}(t) = p(Y_1 = 1, Y_2 = 1/t)$

Sincronía Condicional

$$CSM(t) = \frac{\pi_{11}(t)}{\pi_{1+}(t) + \pi_{+1}(t) - \pi_{11}(t)}$$

Modelo Plackett-Dale

Supongamos que interesa estudiar la sincronía bajo ciertas condiciones experimentales del experimento, $\mathbf{x}$.

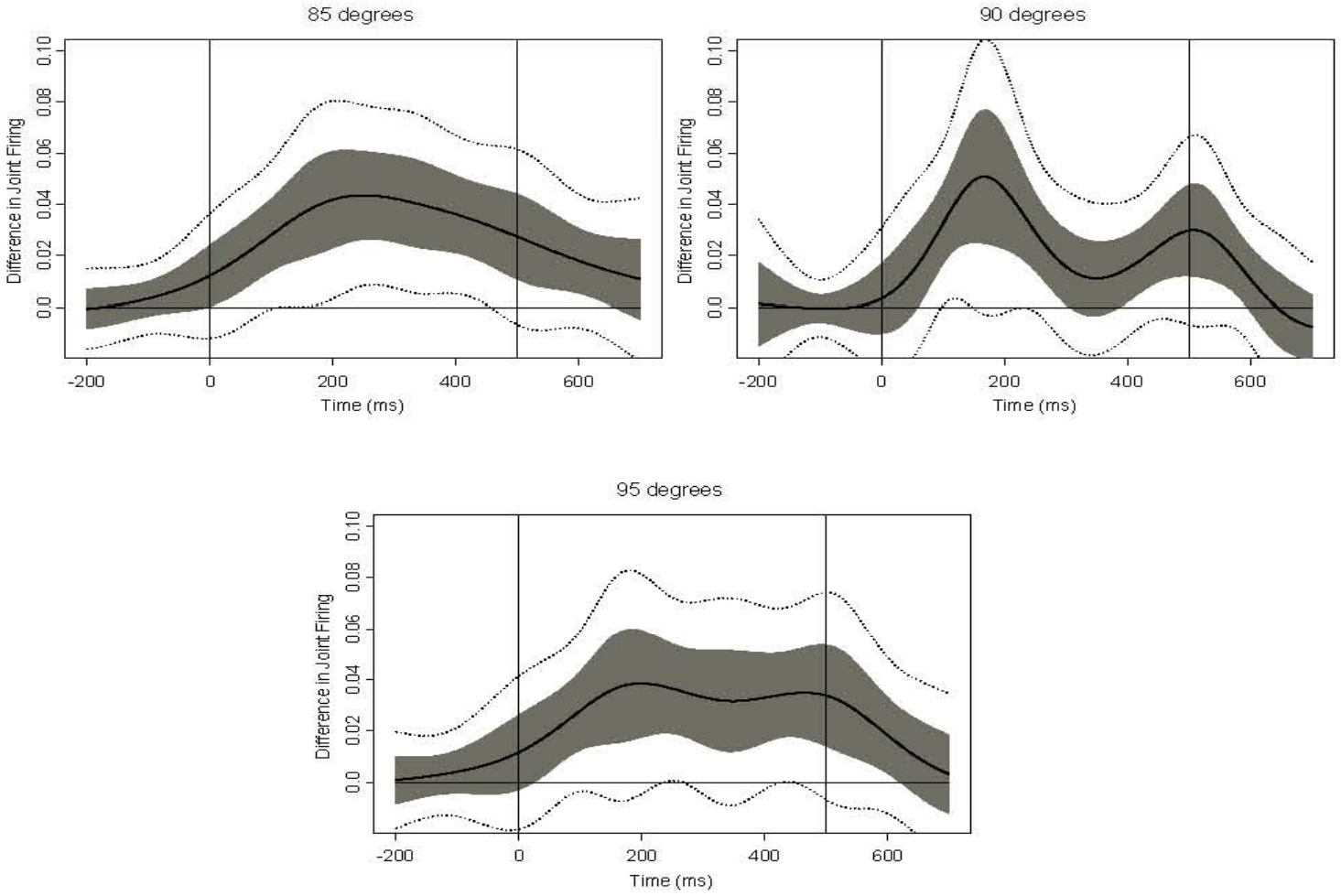
$$\log\frac{\pi_{+1}(t,\mathbf{x})}{1-\pi_{+1}(t,\mathbf{x})} = \alpha_{01} + \beta_1^T \mathbf{x}(t)$$

$$\log\frac{\pi_{1+}(t,\mathbf{x})}{1-\pi_{1+}(t,\mathbf{x})} = \alpha_{02} + \beta_2^T \mathbf{x}(t)$$

$$\log\frac{CSM(t,\mathbf{x})}{1-CSM(t,\mathbf{x})} = \alpha_{03} + \beta_{11}^T \mathbf{x}(t)$$

Sincronía temporal

Orientación (en grados) de la barra = 85, 90, 95



RADIOLOGÍA

“Evaluación de sistemas diagnósticos asistidos por ordenador CAD (Computer-Aided Diagnostic Systems)”

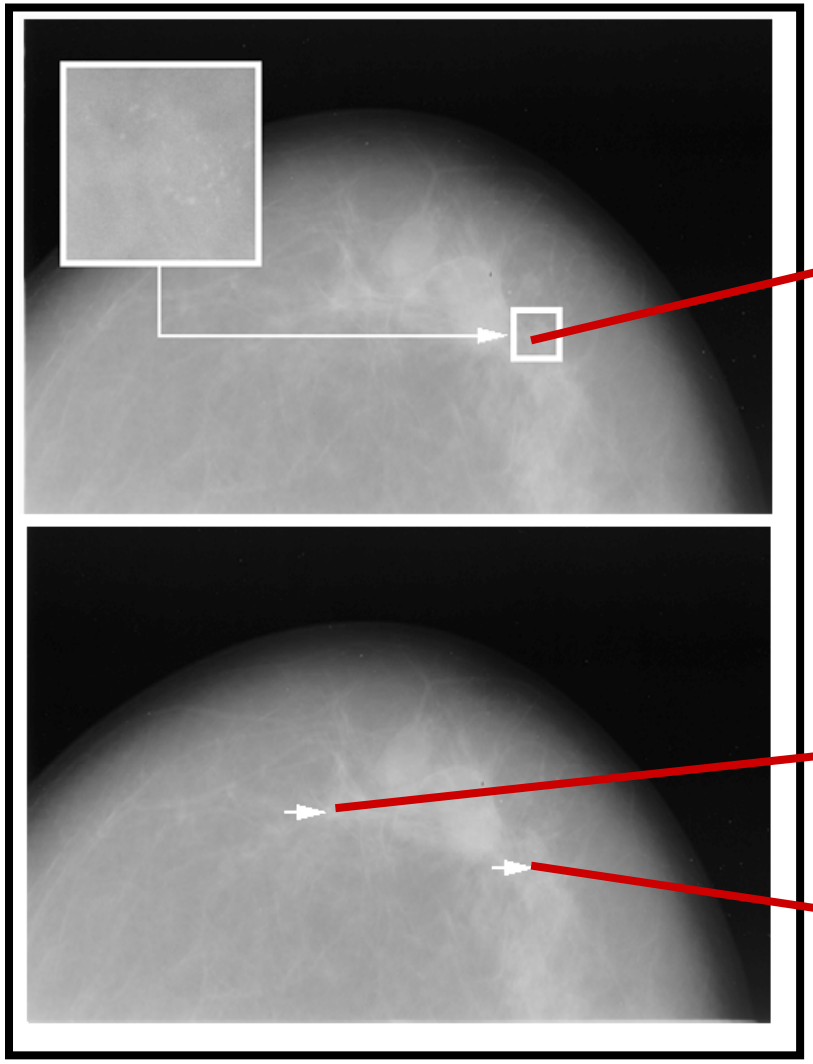
Investigadores colaboradores: físicos, médicos, informáticos

Departamento de Electrónica y Ciencias de la Computación (USC)
Laboratorio de Imagen Radiológica (USC)

Evaluación diagnóstica de sistemas CAD

Radiografía (mama)

CAD Computer-Aided Diagnosis



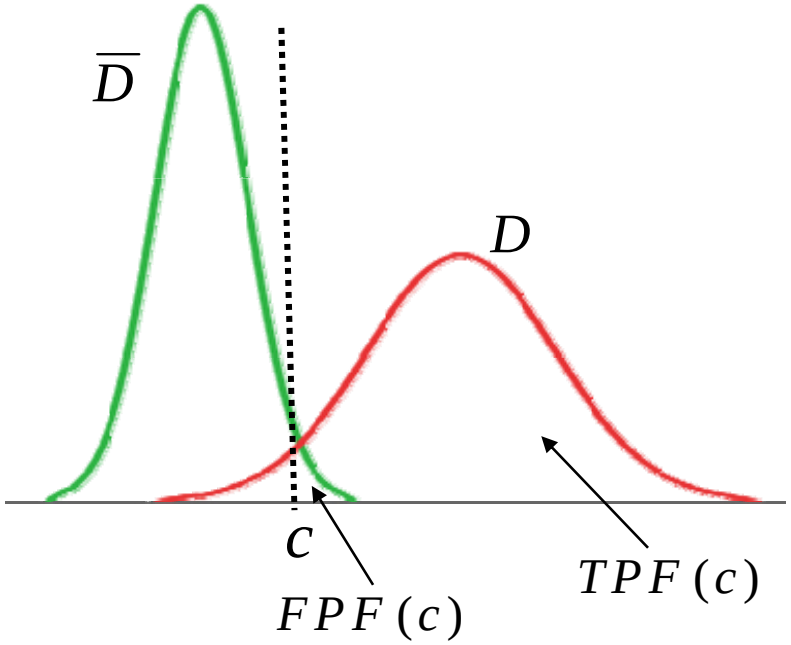
Lesión: Cúmulo (real) de microcalcificaciones

FP (False Positive)

TP (True Positive)

Curva ROC (*Receiver Operating Characteristic*)

- Medida da precisión diagnóstica de tests (marcadores) con respuesta continua, Y
- Proponiendo un punto de corte "c" para clasificar a enfermos y sanos $\left(D, \bar{D}\right)$



$$TPF(c) = P\left[Y \geq c / D\right] = Sensitivity$$

$$FNF(c) = P\left[Y < c / D\right]$$

$$TNF(c) = P\left[Y < c / \bar{D}\right] = Specificity$$

$$FPF(c) = P\left[Y \geq c / \bar{D}\right]$$

Definiciones

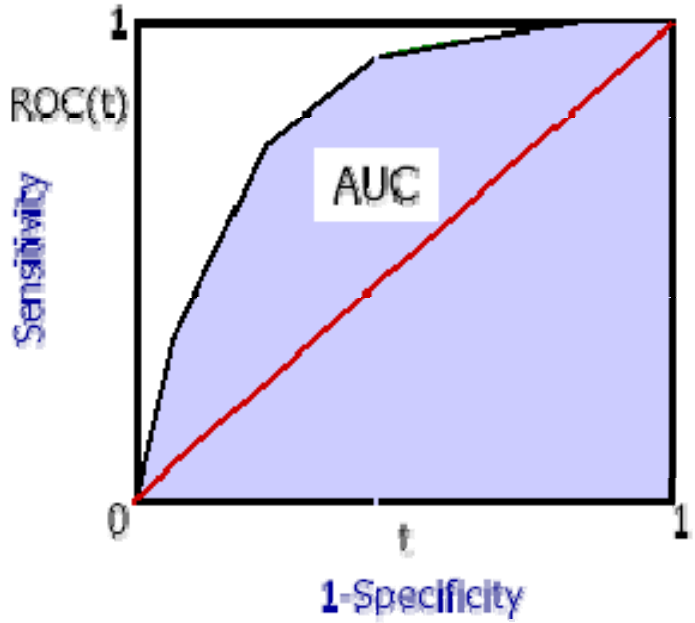
Curva ROC

$$R(\cdot) = \{(FPF(c), TPF(c)),\ c \in (-\infty, \infty)\}$$

$$R(\cdot) - \{(t, ROC(t)),\ t \in (0,1)\} \quad FPF(c) = t \quad ROC(t) = TPF(c)$$

Área bajo la Curva (*Area Under the Curve*, *AUC*)

$$AUC = \int_0^1 ROC(t)dt$$



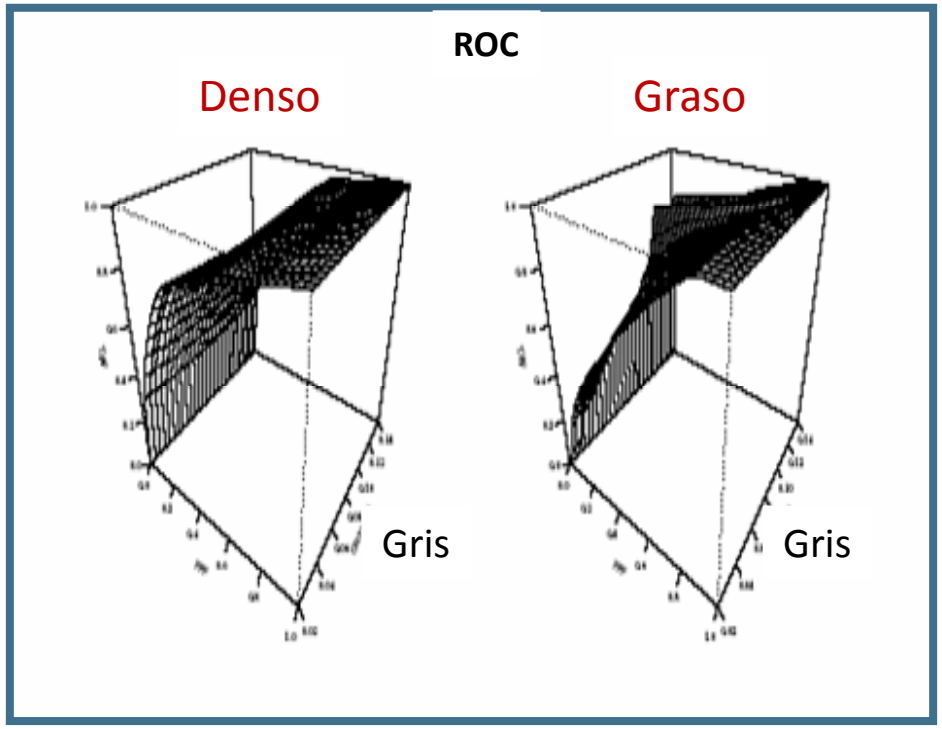
ROC condicional

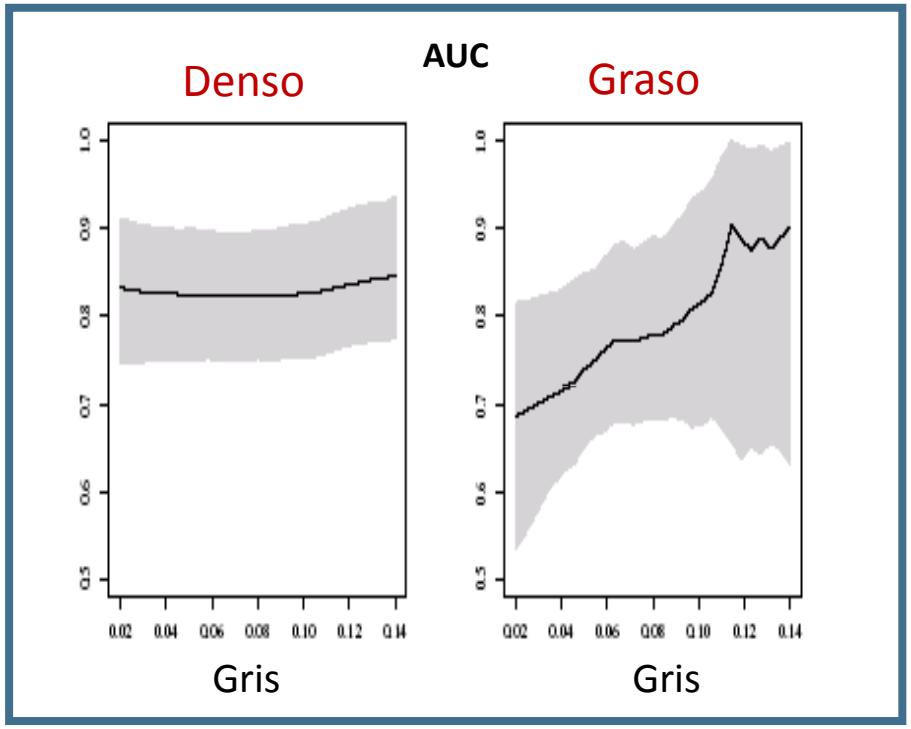
Test diagnóstico: Tamaño del cúmulo.
Status: 1=cúmulo real, 0=no.

Condicionado por:

Nivel de gris de la mamografía digital.

Tipo de tejido de la mama (denso/graso).





Medicina Forense

Colaboración: Instituto de Medicina Legal (USC)

Predicción del intervalo post-mortem

- La estimación del **intervalo post-mortem (PMI)** es un problema fundamental en la Medicina Forense.
- Debido a que la mayoría de las víctimas de homicidio son descubiertas en las primeras horas resulta trascendente disponer de algún método que permita realizar una estima en ese intervalo.
- Métodos tradicionales (temperatura corporal, rigor/livor mortis,…) están siendo sustituidos por determinaciones bioquímicas presentes en diversos fluidos corporales, como el humor vítreo.
- **Sustancias bioquímicas** de interés: Urea [U], Hipoxantina [Hx], Potasio [K+].
- Información adicional: **tipo de homicidio** (ahorcamiento, otros).

$$\log(PMI) = \beta_0 + \beta_1[K^+] + \beta_2[Hx] + \beta_3[U] + \varepsilon$$

Modelos estadísticos de predicción

Modelo lineal (LR)

$$\log(PMI) = \beta_0 + \beta_1[K^+] + \beta_2[Hx] + \beta_3[U] + \varepsilon$$

Modelo Aditivo (AM, Hastie-Tibshirani, 1990)

$$\log(PMI) = \alpha_0 + f_1\left([K^+]\right) + f_2\left([Hx]\right) + f_3\left([U]\right) + \varepsilon$$

Modelo Aditivo con interacciones (AM2)

$$\log(PMI) = \alpha_0 + f_3\left([U]\right) + f_{11}\left([K^+]1_{D=1}\right) + f_{21}\left([Hx]1_{D=1}\right)$$
$$+ f_{12}\left([K^+]1_{D=2}\right) + f_{22}\left([Hx]1_{D=2}\right) + \varepsilon$$

Support Vector Machines(SVM, Statistical Learning Theory, Vapnik 1995)

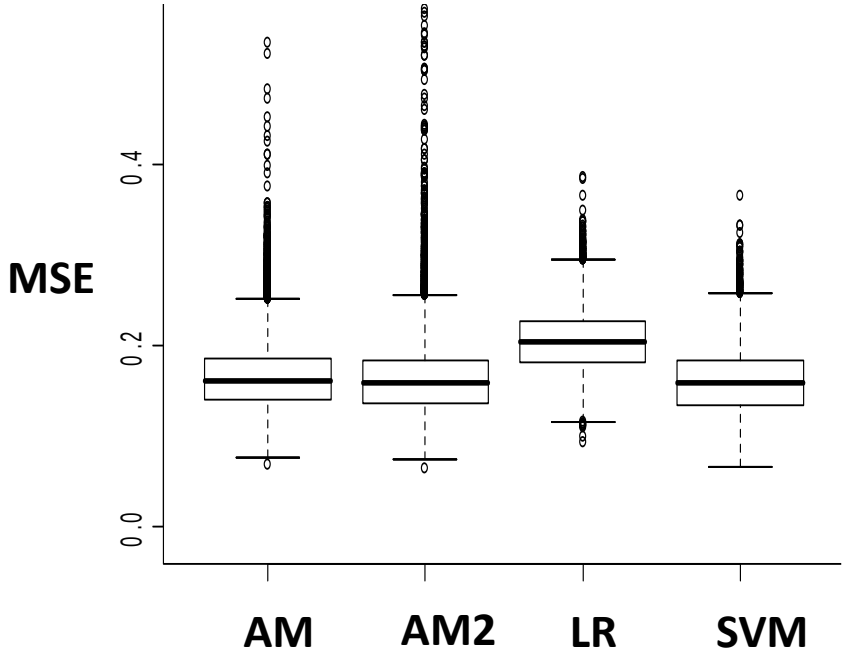
$$\log(PMI) = \sum_{i=1}^{n} \beta_i k(([U],[K^+],[Hx],D),([U],[K^+],[Hx],D)_i) + \varepsilon$$

Validación de los métodos

1) En este estudio se separó la muestra inicial en dos partes:
   70% datos: “training” set.
   30% datos: “validation” set.

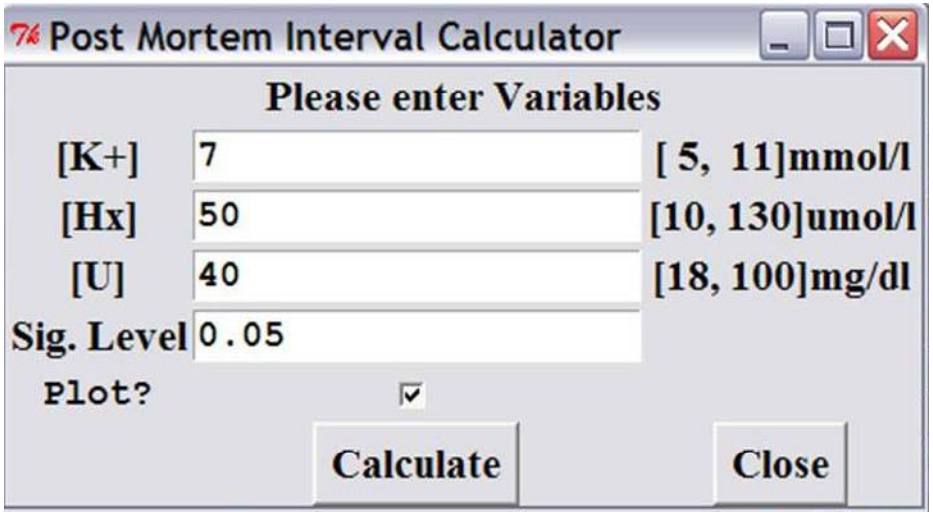
2) Se generaron B=10.000 particiones aleatorias de la muestra inicial.

3) Para cada una de estas particiones se calculó el **error cuadrático medio (MSE)** sobre la muestra de validación, calculando el modelo con la muestra de entrenamiento.

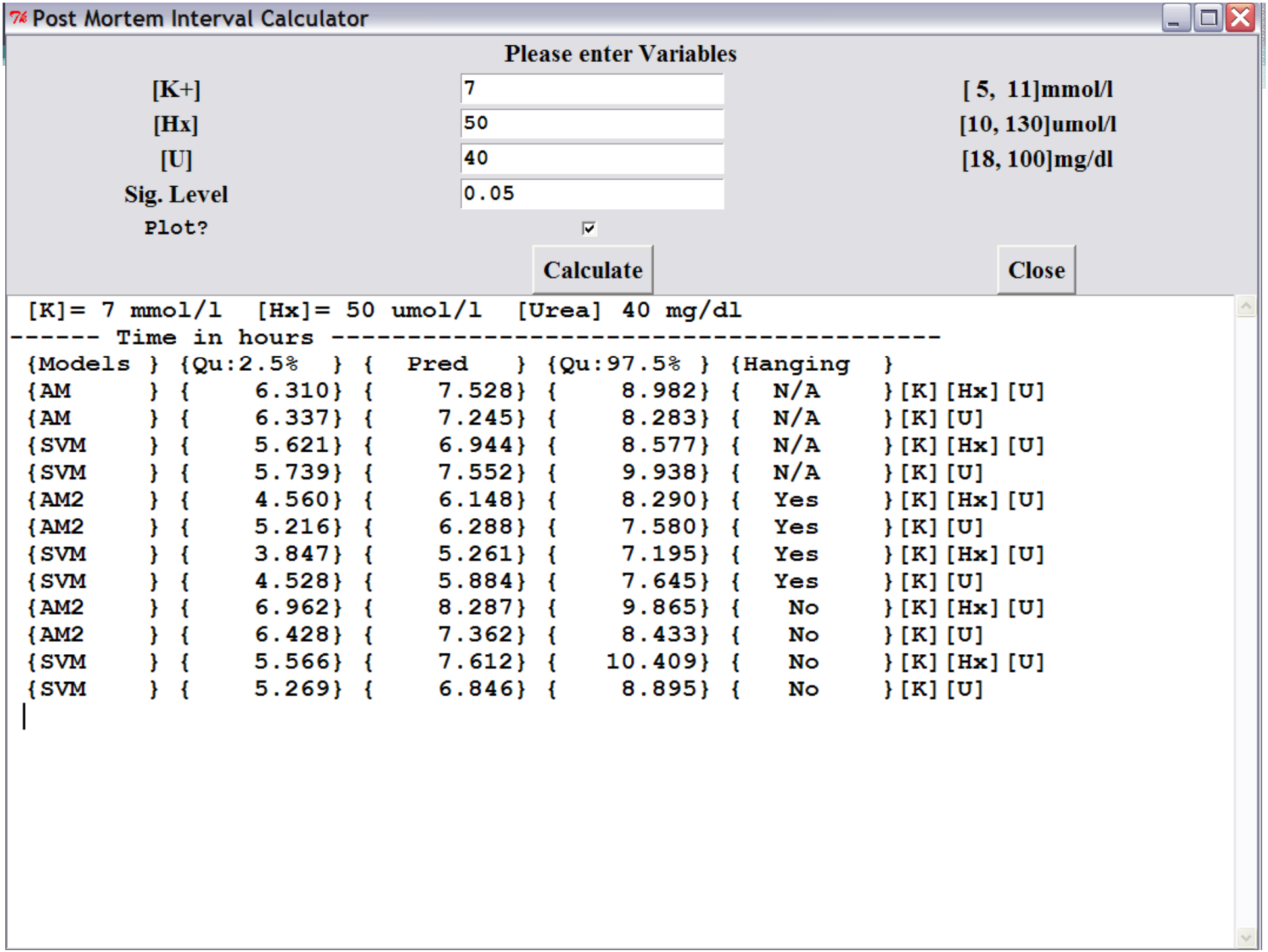


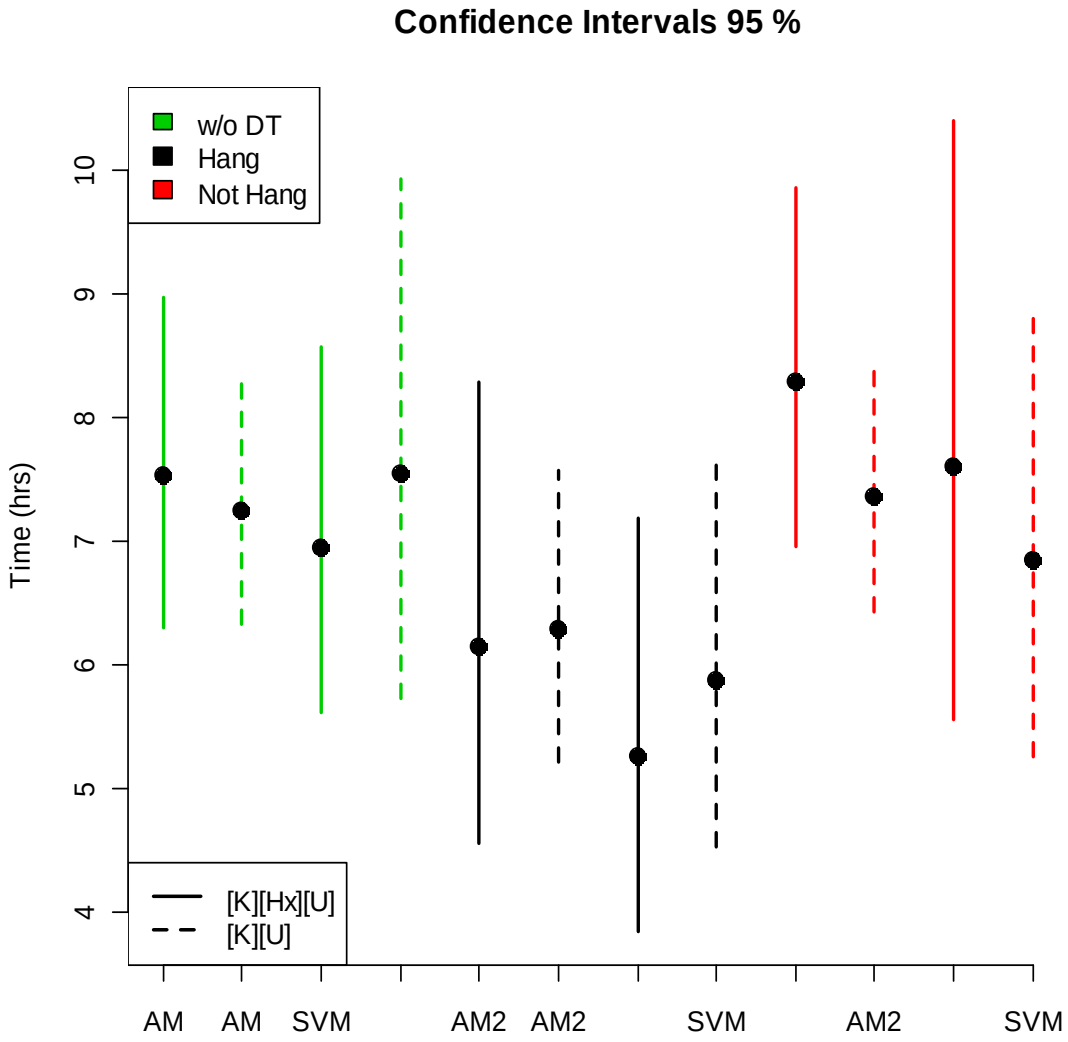
pmicalc ()

Programa en lenguaje R, que permite obtener la predicción del PMI



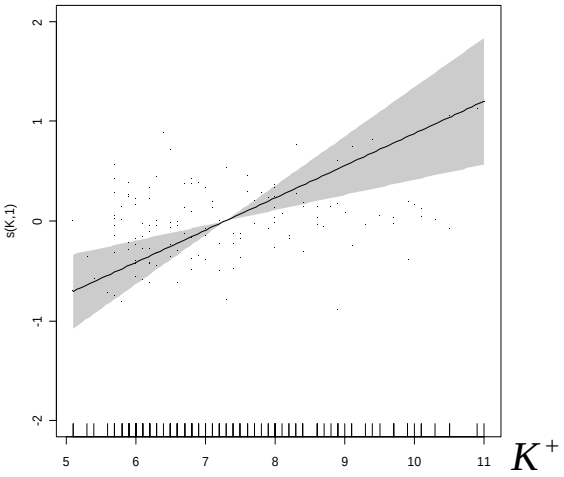
- Salidas numéricas.
- Salidas gráficas.

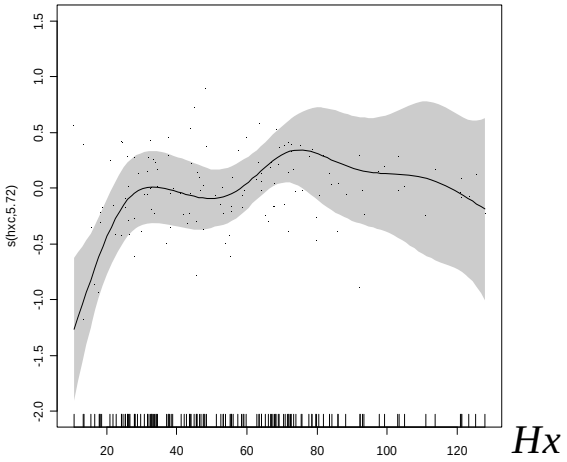

Post Mortem Interval Calculator
Please enter Variables
[K+]
7
[5, 11]mmol/l
[Hx]
50
[10, 130]umol/l
[U]
40
[18, 100]mg/dl
Sig. Level
0.05
Plot?
Calculate
Close
[K]= 7 mmol/l [Hx]= 50 umol/l [Urea] 40 mg/dl
------ Time in hours ----------
{Models } {Qu:2.5% } { Pred } {Qu:97.5% } {Hanging }
{AM } { 6.310} { 7.528} { 8.982} { N/A }[K][Hx][U]
{AM } { 6.337} { 7.245} { 8.283} { N/A }[K][U]
{SVM } { 5.621} { 6.944} { 8.577} { N/A }[K][Hx][U]
{SVM } { 5.739} { 7.552} { 9.938} { N/A }[K][U]
{AM2 } { 4.560} { 6.148} { 8.290} { Yes }[K][Hx][U]
{AM2 } { 5.216} { 6.288} { 7.580} { Yes }[K][U]
{SVM } { 3.847} { 5.261} { 7.195} { Yes }[K][Hx][U]
{SVM } { 4.528} { 5.884} { 7.645} { Yes }[K][U]
{AM2 } { 6.962} { 8.287} { 9.865} { No }[K][Hx][U]
{AM2 } { 6.428} { 7.362} { 8.433} { No }[K][U]
{SVM } { 5.566} { 7.612} { 10.409} { No }[K][Hx][U]
{SVM } { 5.269} { 6.846} { 8.895} { No }[K][U]


Confidence Intervals 95 %
w/o DT
Hang
Not Hang
[K][Hx][U]
[K][U]
Time (hrs)
4
5
6
7
8
9
10
AM
AM
SVM
AM2
AM2
SVM
AM2
SVM

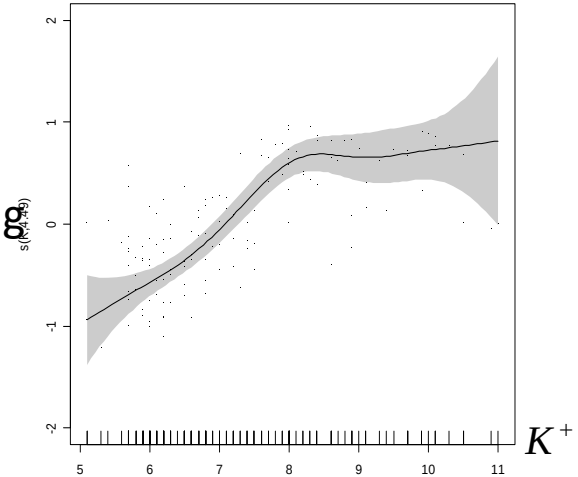
Modelo AM2

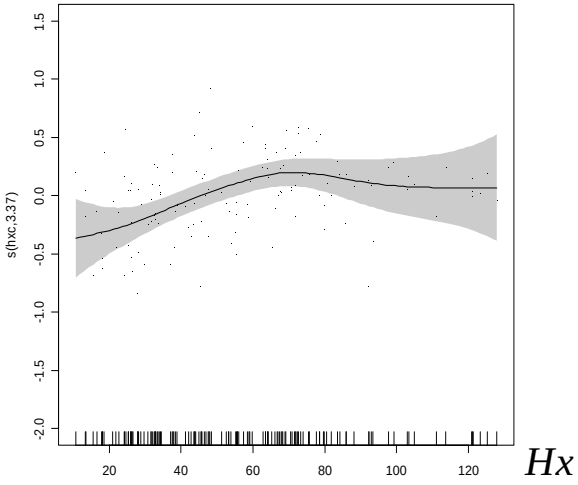
hanging



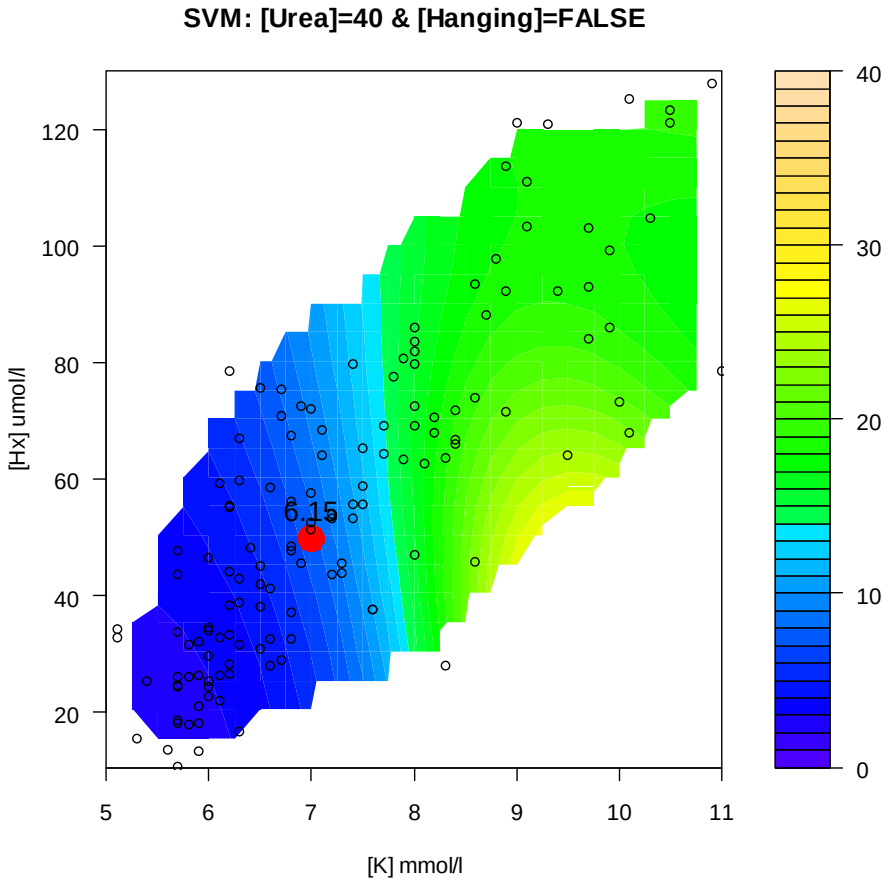


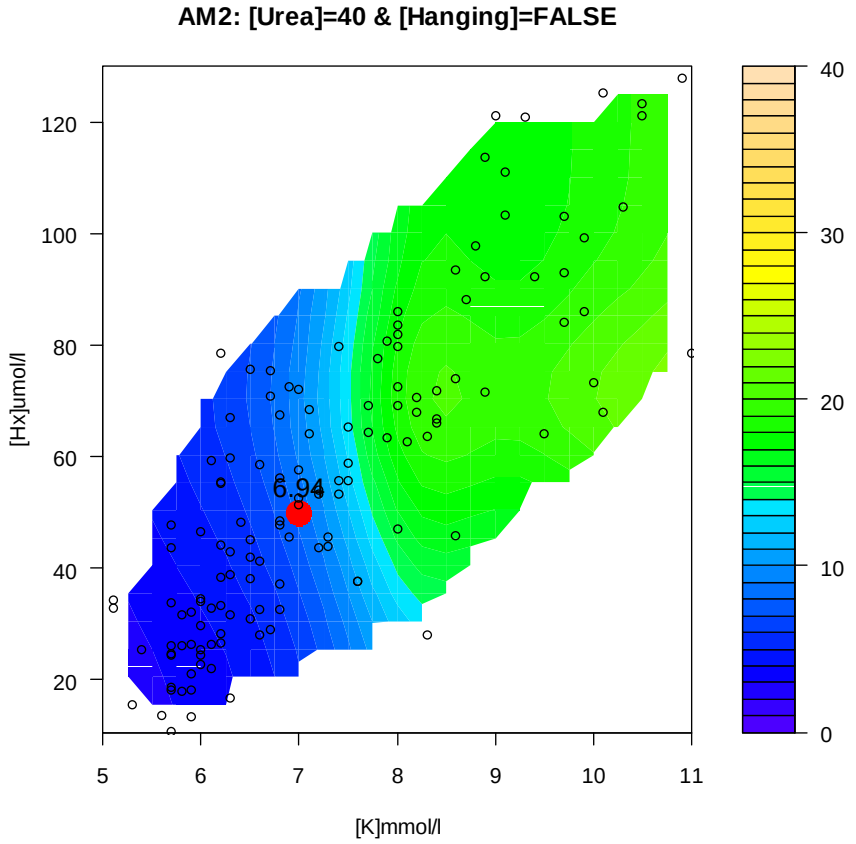
not hanging





Predicción: Modelos SVM y AM2





¿Cómo accede el forense?

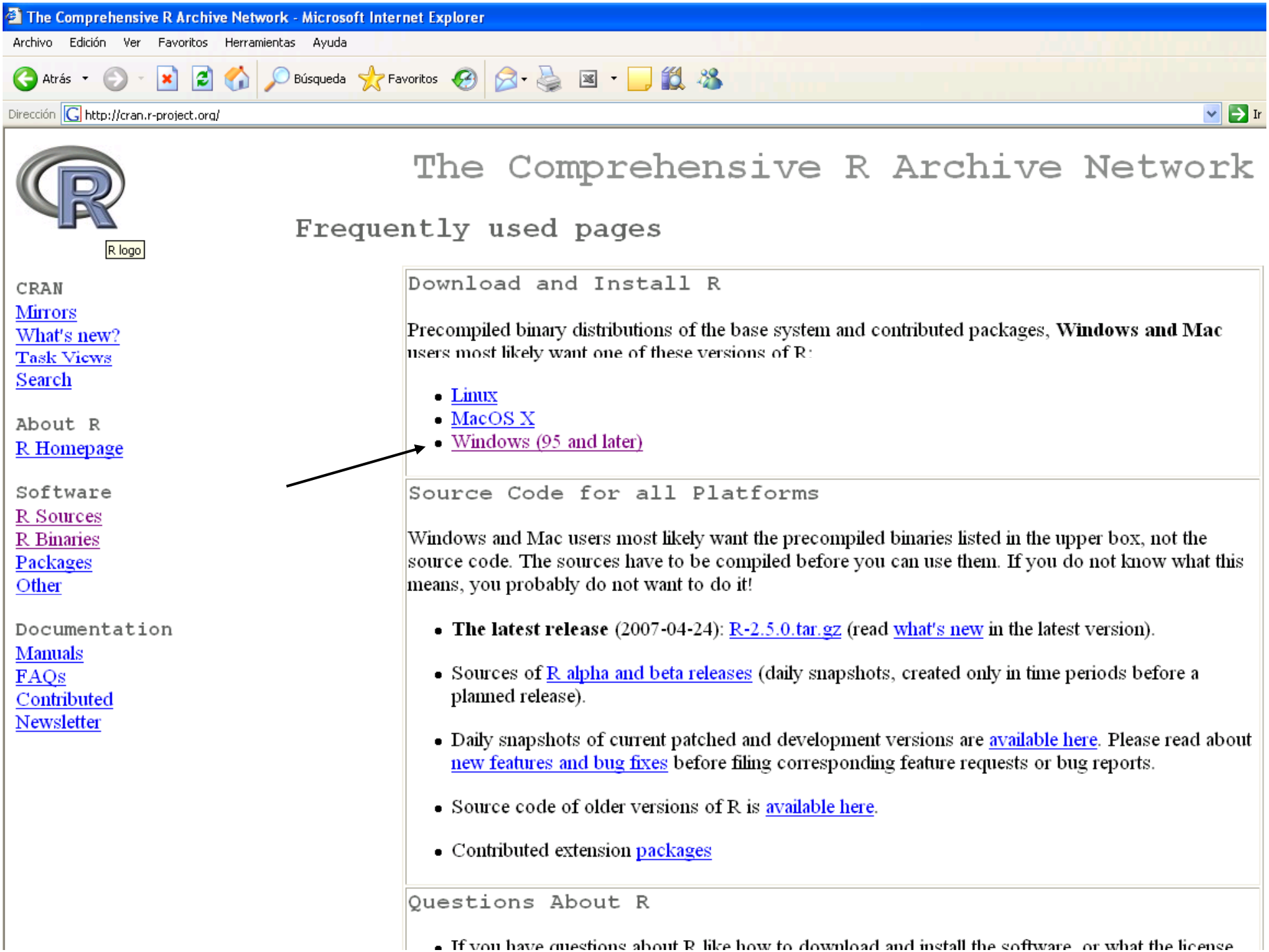
El programa se encontrará en breve en la página web oficial de la **Academia Internacional de Medicina Legal**

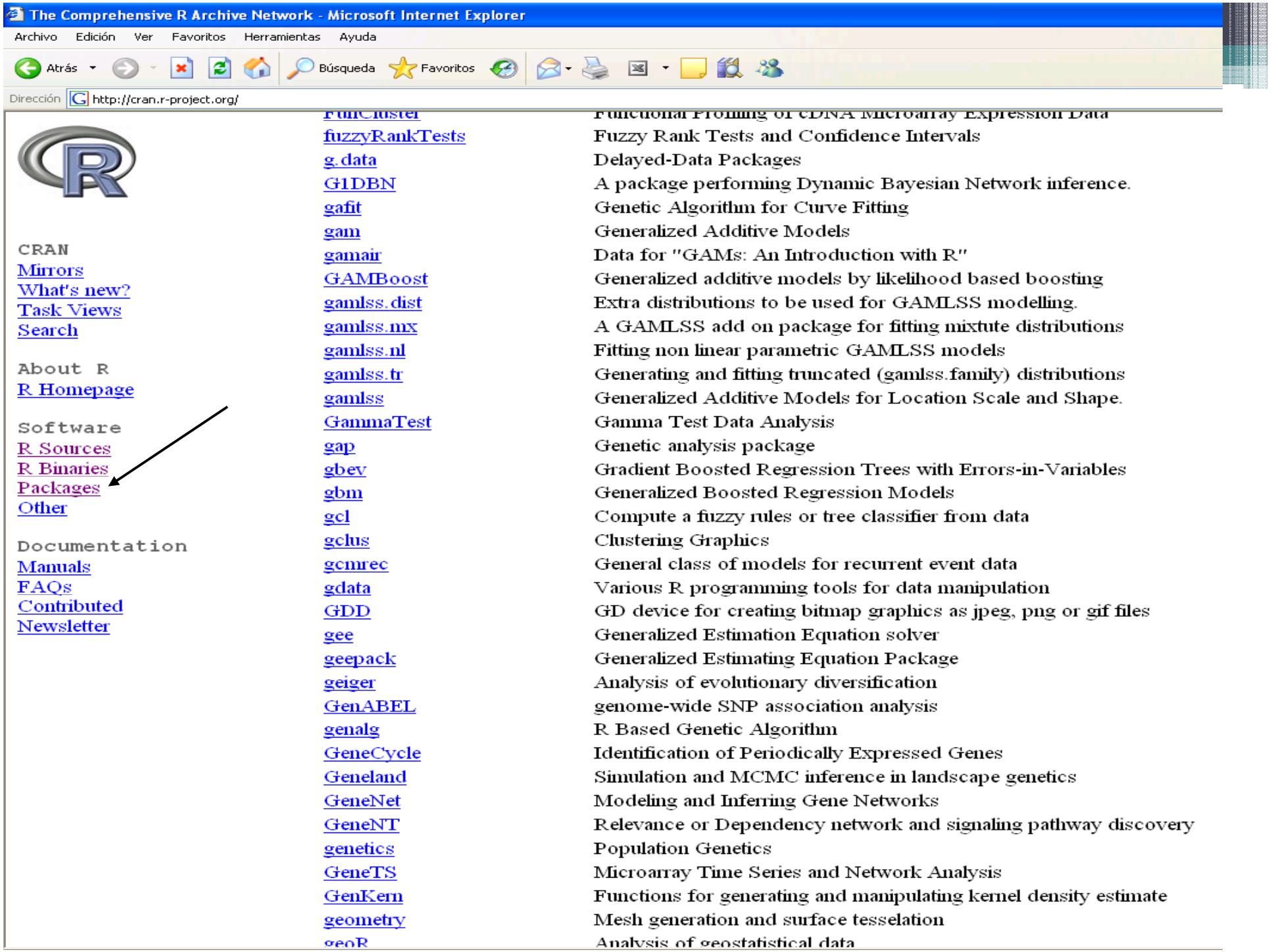
ELABORACIÓN
DE
SOFTWARE

“ Una necesidad para la diseminación
de nuevas metodologías estadísticas y su aplicación práctica”

El entorno R

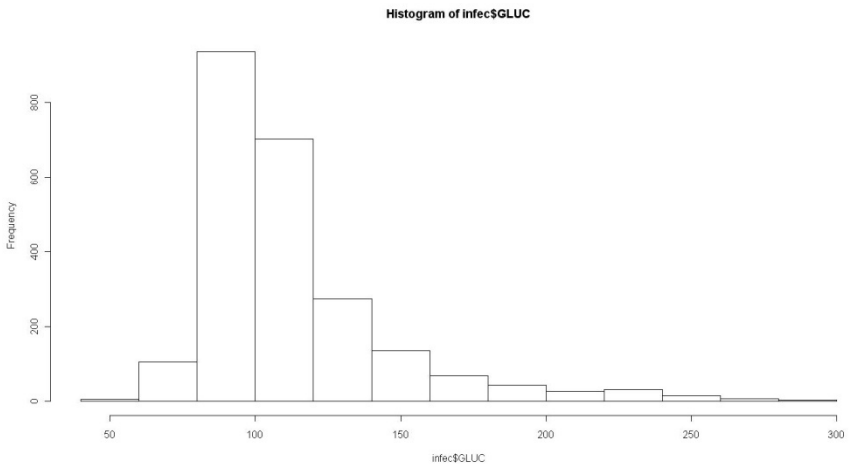
- Lenguaje de programación “orientado a objetos”.
- Permite realizar cálculos matemáticos potentes.
- Potentes herramientas gráficas.
- Distribución libre (**http://cran.r-project.org/).**
- Gran variedad de métodos estadísticos (clásicos y novedosos).
- Librerías estadísticas (Packages).
- Las funciones son de “código abierto”.

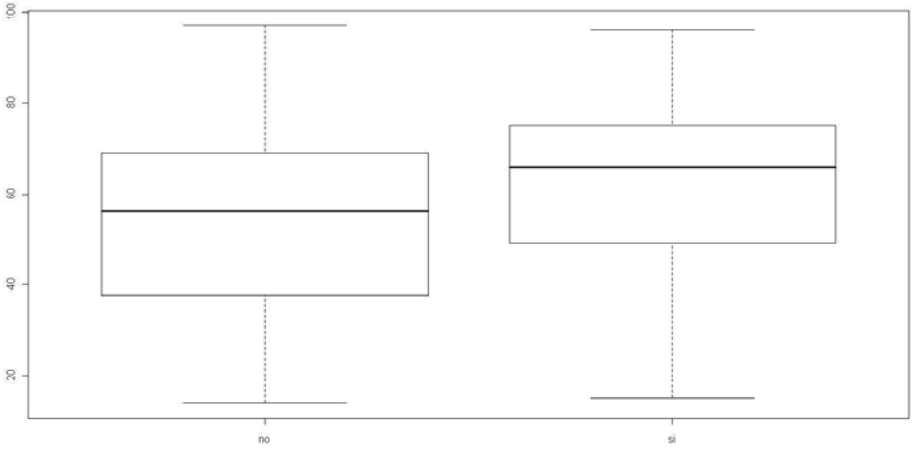

The Comprehensive R Archive Network - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás
Búsqueda
Favoritos
Dirección
http://cran.r-project.org/
Ir
R logo
The Comprehensive R Archive Network
Frequently used pages
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Newsletter
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Linux
MacOS X
Windows (95 and later)
Source Code for all Platforms
Windows and Mac users most likely want the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2007-04-24): R-2.5.0.tar.gz (read what's new in the latest version).
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages
Questions About R


The Comprehensive R Archive Network - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos
Dirección http://cran.r-project.org/
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
Newsletter
FunCluster Functional Profiling of cDNA Microarray Expression Data
fuzzyRankTests Fuzzy Rank Tests and Confidence Intervals
g.data Delayed-Data Packages
G1DBN A package performing Dynamic Bayesian Network inference.
gafit Genetic Algorithm for Curve Fitting
gam Generalized Additive Models
gamair Data for "GAMs: An Introduction with R"
GAMBoost Generalized additive models by likelihood based boosting
gamlss.dist Extra distributions to be used for GAMLSS modelling.
gamlss.mx A GAMLSS add on package for fitting mixtute distributions
gamlss.nl Fitting non linear parametric GAMLSS models
gamlss.tr Generating and fitting truncated (gamlss.family) distributions
gamlss Generalized Additive Models for Location Scale and Shape.
GammaTest Gamma Test Data Analysis
gap Genetic analysis package
gbev Gradient Boosted Regression Trees with Errors-in-Variables
gbm Generalized Boosted Regression Models
gcl Compute a fuzzy rules or tree classifier from data
gclus Clustering Graphics
gcmrec General class of models for recurrent event data
gdata Various R programming tools for data manipulation
GDD GD device for creating bitmap graphics as jpeg, png or gif files
gee Generalized Estimation Equation solver
geepack Generalized Estimating Equation Package
geiger Analysis of evolutionary diversification
GenABEL genome-wide SNP association analysis
genalg R Based Genetic Algorithm
GeneCycle Identification of Periodically Expressed Genes
Geneland Simulation and MCMC inference in landscape genetics
GeneNet Modeling and Inferring Gene Networks
GeneNT Relevance or Dependency network and signaling pathway discovery
genetics Population Genetics
GeneTS Microarray Time Series and Network Analysis
GenKern Functions for generating and manipulating kernel density estimate
geometry Mesh generation and surface tesselation
geoR Analysis of geostatistical data

Ejemplo: Infección post-quirúrgica

```
>library(foreign)

>infec<-read.spss("G:\\infec.sav", to.data.frame=T)

>names(infec)
      [1] "EDAD"     "SEXO"     "LEUCOS"   "GLUC"     "DIABETES" "INFEC"

>hist(infec$GLUC)

>boxplot(infec$EDAD~infec$INFEC)
```





Regresión spline con R

```
>library(splines)

>infec.bs<-glm(INFEC~bs(GLUC,knots=c(96,113)),family=binomial,data=infec)

>summary(infec.bs)
```

Coefficients:

	Estimate	SE	z	Pr(>|z|)	
(Intercept)	0.1426	0.9833	0.145	0.88470	
bs(GLUC, knots = c(96, 113))1	-0.8820	1.3187	-0.669	0.50359	
bs(GLUC, knots = c(96, 113))2	-2.6993	0.9527	-2.833	0.00461	**
bs(GLUC, knots = c(96, 113))3	0.8944	1.1492	0.778	0.43643	
bs(GLUC, knots = c(96, 113))4	-1.6768	1.2508	-1.341	0.18008	
bs(GLUC, knots = c(96, 113))5	-1.5406	1 .4748	-1.045	0.29621	

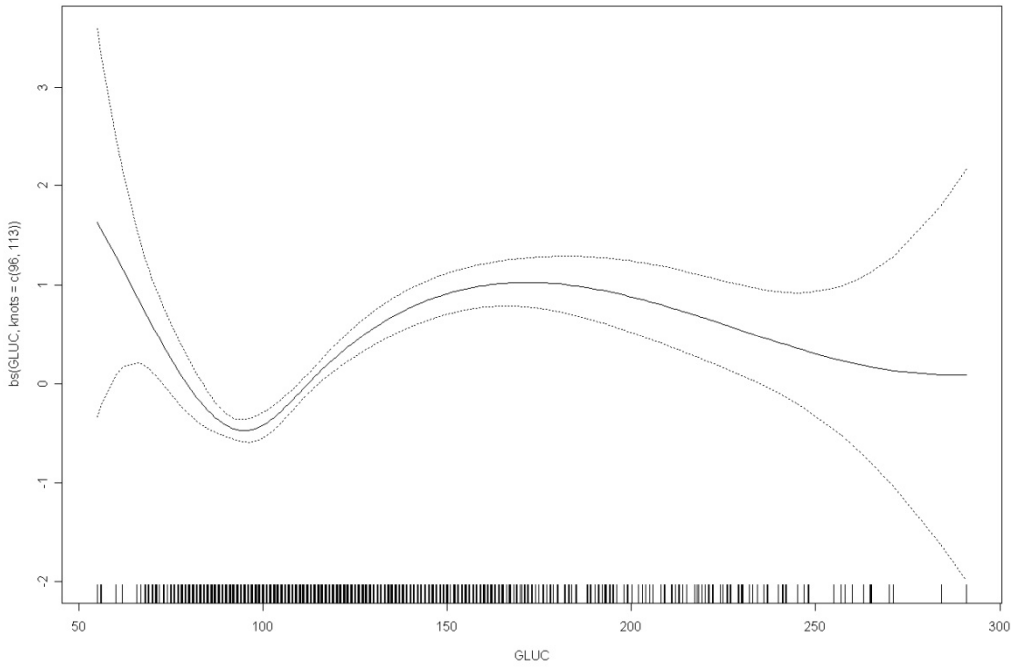
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2281.0 on 2309 degrees of freedom
Residual deviance: 2188.5 on 2304 degrees of freedom
AIC: 2200.5

Regresión spline con R

- Vamos a representar gráficamente el efecto de la glucosa.
- Usaremos la librería **gam** de Hastie-Tibshirani.

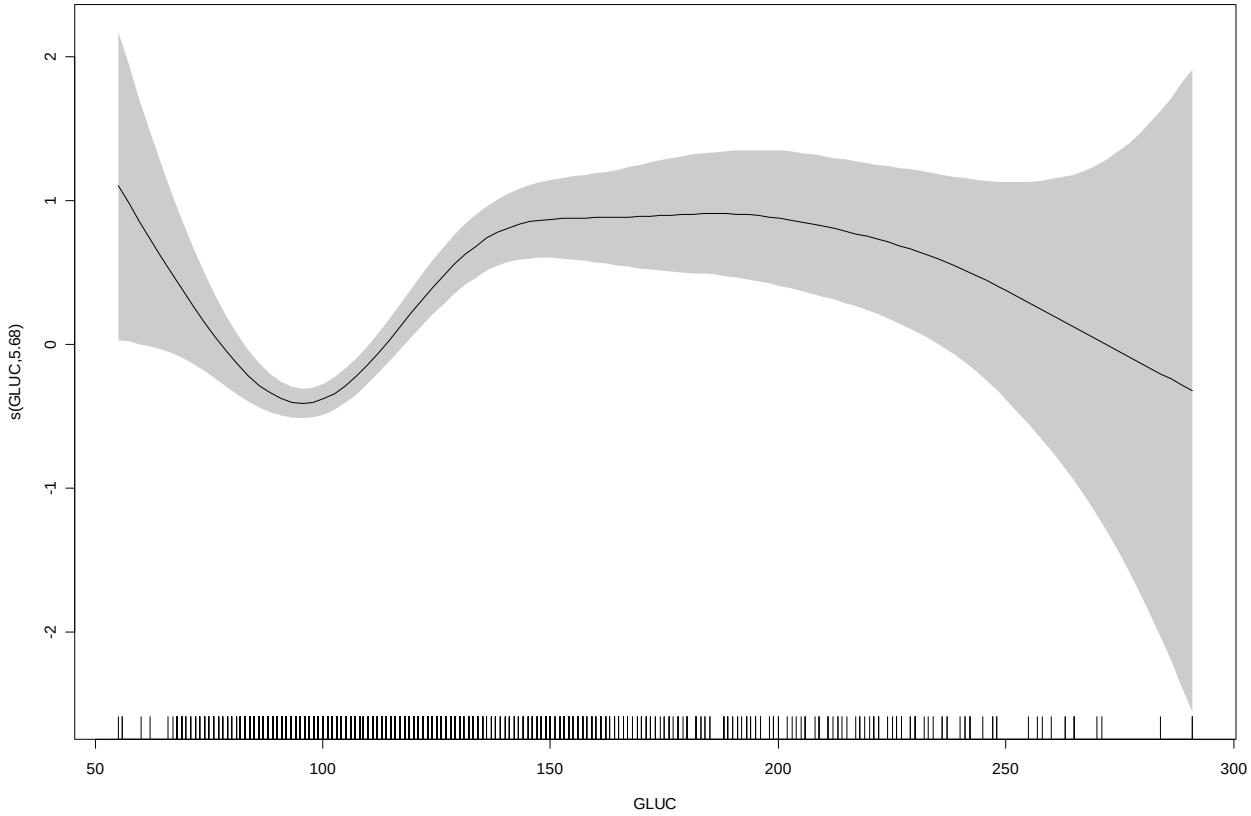
```
>library(gam)

>infec.bs<-gam(INFEC~bs(GLUC,knots=c(96,113)),family=binomial,data=infec)

>plot(infec.bs, se=T)
```



Modelo GAM con R

- S**uavización:** thin regression splines (Wood, 2006).
- **Selección automática de los gl**: Criterio GCV.

```
>library(mgcv)
>infec.gam.gcv<-gam(INFEC~s(GLUC),family=binomial,data=infec)
>plot(infec.gam.gcv,se=T, shade=T)
```



$df_{GCV} \approx 6$

(gl óptimos)

CONSIDERACIONES FINALES

- La Estadística (la Matemática en general) es fundamental en el desarrollo de las Ciencias Biomédicas.

- Puede ayudar en gran manera a la “investigación traslacional en salud humana”.

VII Progama Marco Investigación (UE)

- La **comunidad biomédica** es cada vez más consciente de ello:
  - Demanda de **bioestadísticos**:

    Hospitales.
    Instituciones Públicas.
    Centros de Investigación biomédica.
    Laboratorios farmacéuticos,…
  - Demanda de **cursos** de estadística.

- La **comunidad matemática** también se involucra:
  - Creación del **Instituto Español de Matemáticas** (IeMath).
  - Proyecto Nacional **Ingenio-Mathematica**.

Bioestadística