

Computer  
Assisted  
Diagnosis  
(CAD)

&

Distributed  
Computing  
Infrastructure  
(GRID)

# Tecnologie informatiche nella diagnostica per immagini



Piergiorgio Cerello (INFN-Torino)  
cerello@to.infn.it



# Sintesi

- La diagnostica per immagini
  - ◆ Le immagini digitali (2d e 3d)
- Intelligenza Artificiale in fisica medica
  - ◆ Pre processamento
  - ◆ Reti Neurali
- Un esempio di applicazione:
  - ◆ il progetto CALMA: Computer Assisted Library in Mammography
- Sistemi di gestione di risorse geograficamente distribuite (GRID)
  - ◆ Network
  - ◆ Computing
  - ◆ Storage
  - ◆ Virtual Organisation
- Un esempio di applicazione:
  - ◆ il progetto GPCALMA: Grid Platform for Computer Assisted Library in Mammography

# Le immagini digitali

## ➤ Analogiche

- ◆ Lastre, digitizzabili con scanner

## ➤ Digitali

### ◆ Formato DICOM

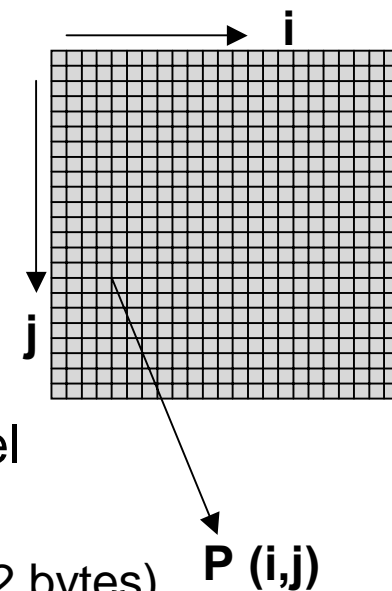
- ✓ Header: informazioni generali

- ◆ Paziente
- ◆ Condizioni di acquisizione

- ✓ Core: lista di valori associati ai diversi pixel

- ✓ Dimensione:


- ◆  $N\_bytes = (size\_i/pitch\_i) * (size\_j/pitch\_j) * (2 \text{ bytes})$
- ◆ Livello di Saturazione a 12 bits  $S = 2^{12} = 4096$
- ◆ Normalizzazione:  $P'(i,j) = P(i,j)/S$        $0 < P'(i,j) < 1$

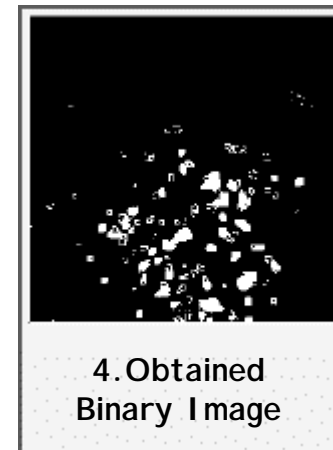
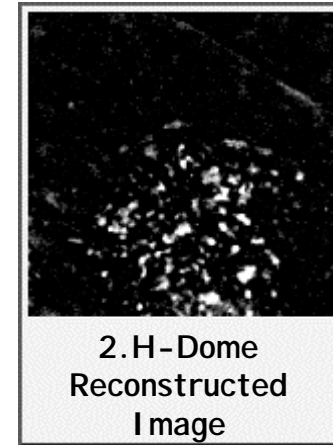
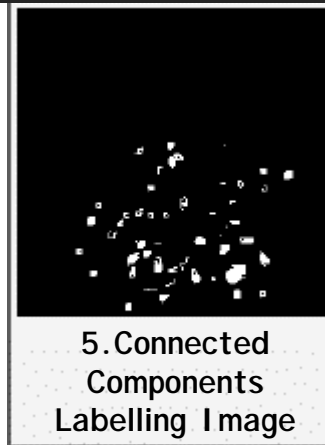
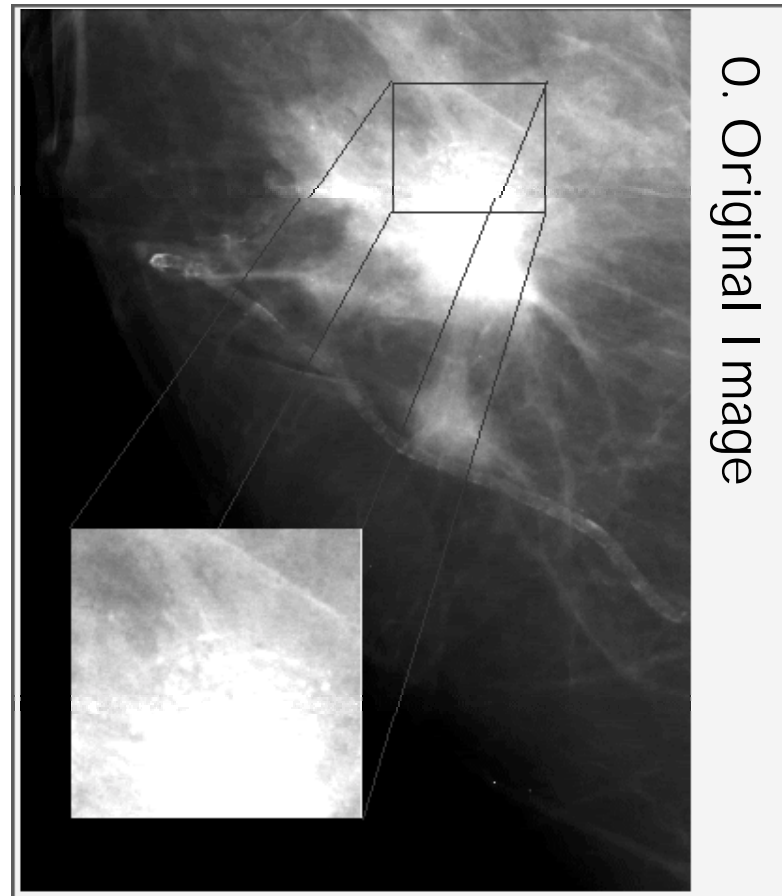
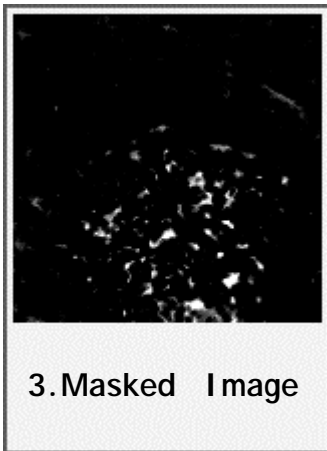


# Algoritmi di analisi

- Ogni immagine contiene: segnale e fondo
- 2 fasi di analisi dell'immagine
  - ◆ Pre-processamento
    - ✓ eliminazione (riduzione) del fondo  
SENZA perdere SEGNALE
  - ◆ Classificazione con reti neurali
    - ✓ estrazione dei parametri caratteristici
    - ✓ classificazione
  - ◆ Analisi con:
    - ✓ Pixel Vector Methods
    - ✓ Active Connection Matrices

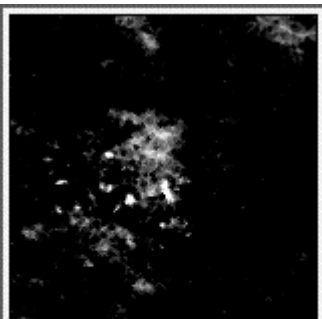
# Pre-processing

- 1. Original Image H-Dome Transformation  
 Image with Peaks
  - 2. Disk Shaped Structure Element Filter (Morphological Erosion Mask)
  - 3. Binarization of the Masked Image
  - 4. Connected Component Labeling Operation - Size constraint
  - 5. Connected Component Labeling Operation - Cluster Constraint
  - Feature Extraction
- 
- Backpropagation Learning and Classification

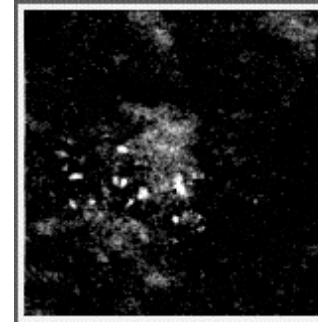




1. Original  
Extracted Image



3. Masked Image



2. H-Dome  
Reconstructed  
Image



4. Obtained  
Binary Image

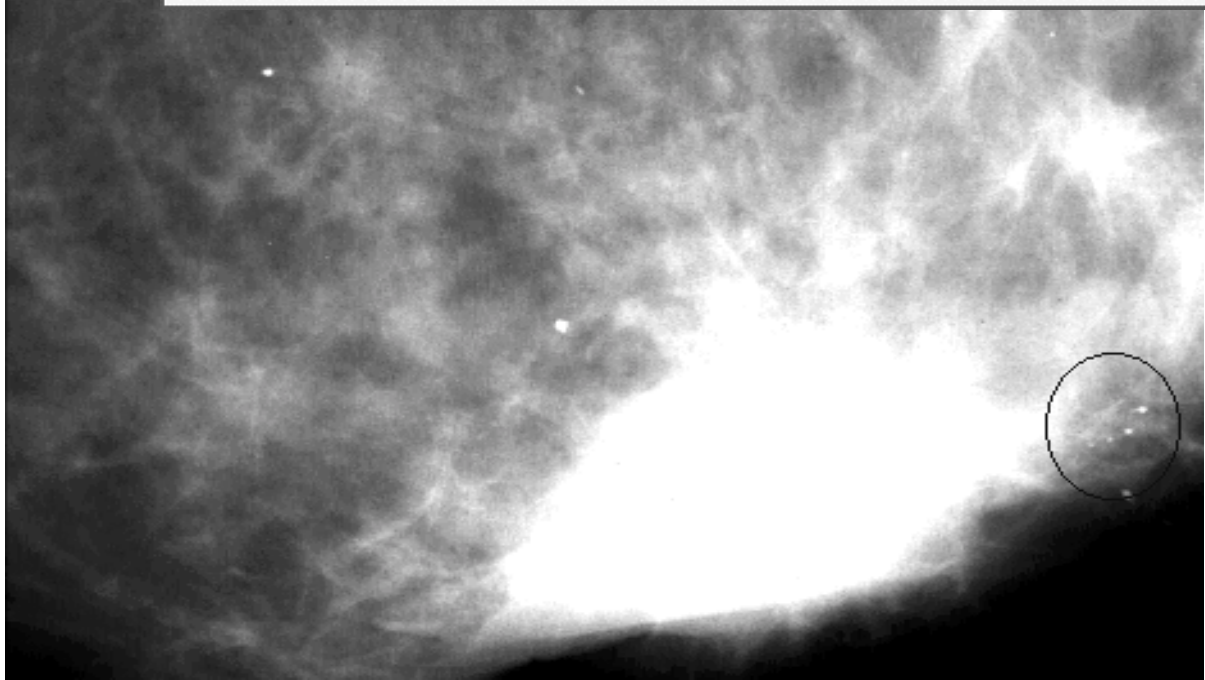


5. Connected  
Components  
Labelling Image

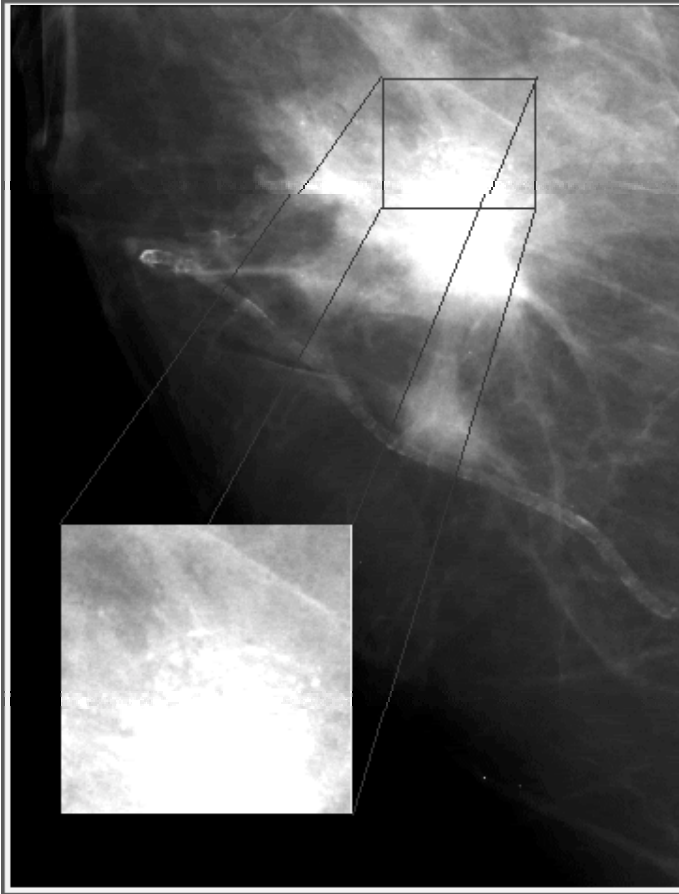


# cult example

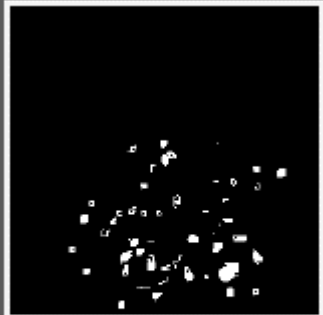
## 4. Connected Components Labelling Image



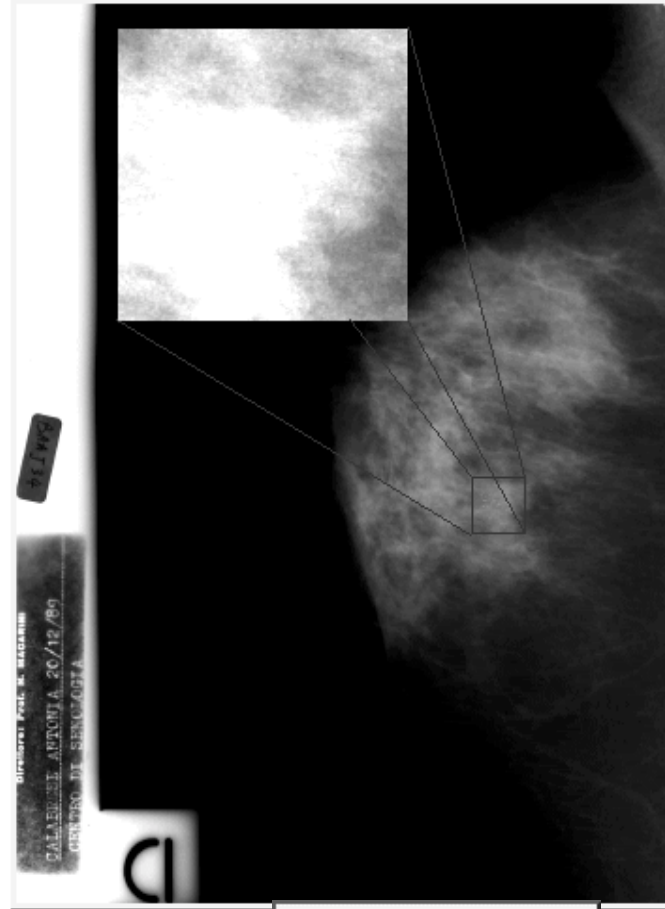




0. Original Image



5. Connected Components Labelling Image



0. Original Image



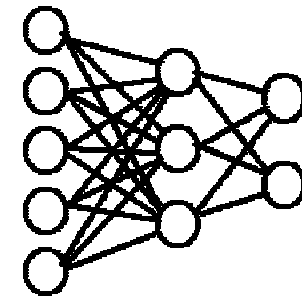
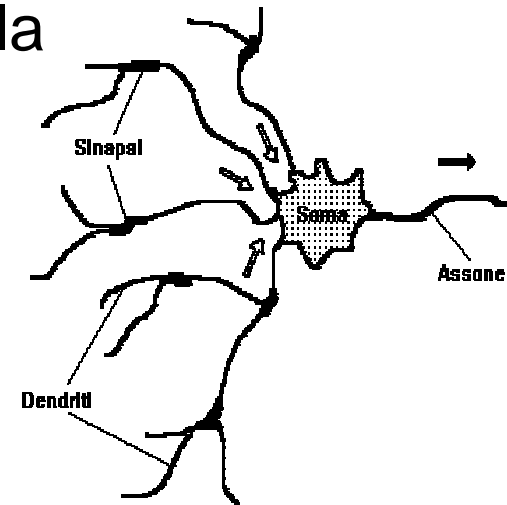
5. Connected Components Labelling Image

# Introduzione alle Reti Neurali

Architettura  
Procedure di Apprendimento  
Applicazioni

# Artificial Neural Networks (ANN)

- Scopo: simulazione ed emulazione della struttura e delle funzioni del cervello
- Le reti neurali (biologiche e artificiali) sono composte da N unita' elementari, interconnesse tra loro
- Ogni unita' elementare, il neurone
  - ◆ elabora n segnali di ingresso (input) e
  - ◆ genera 1 segnale di uscita (output) che trasmette a tutte le unita' connesse
- La procedura avviene tramite
  - ◆ processi elettro-chimici nei sistemi biologici
  - ◆ Emulazione da parte di componenti elettronici (algoritmi) nelle implementazioni hardware (software)

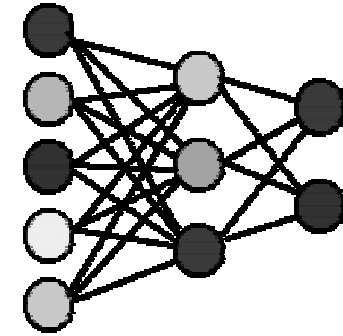


# Confronto tra cervello e ANN

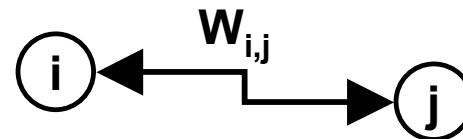
## Cervello                      ANN

- |                                    |                       |                       |
|------------------------------------|-----------------------|-----------------------|
| ➤ Frequenza di processamento:      | $\sim 10^3 \text{Hz}$ | $\sim 10^9 \text{Hz}$ |
| ➤ Numero di neuroni:               | $10^{11}$             | $10^{2-3}$            |
| ➤ Numero di sinapsi (connessioni): | $10^{13}$             | $10^{2-3}$            |
| ➤ Numero sinapsi/neurone:          | $10^{2-4}$            | $10^2$                |
- Il cervello e' estremamente efficiente, molto piu' veloce del piu' veloce processore
  - La potenza di calcolo del cervello e' stimata intorno a  $10^{15}$  operazioni/s, con una memoria di circa  $10^8$  Mbytes.
  - ANN sono simili al cervello in quanto:
    - ◆ La competenza viene acquisita attraverso un processo di apprendimento (learning o training)
    - ◆ La "forza" delle connessioni tra neuroni (definita da un parametro detto "peso sinaptico") viene usata per conservare la competenza (vedremo tra poco come)

# Il neurone



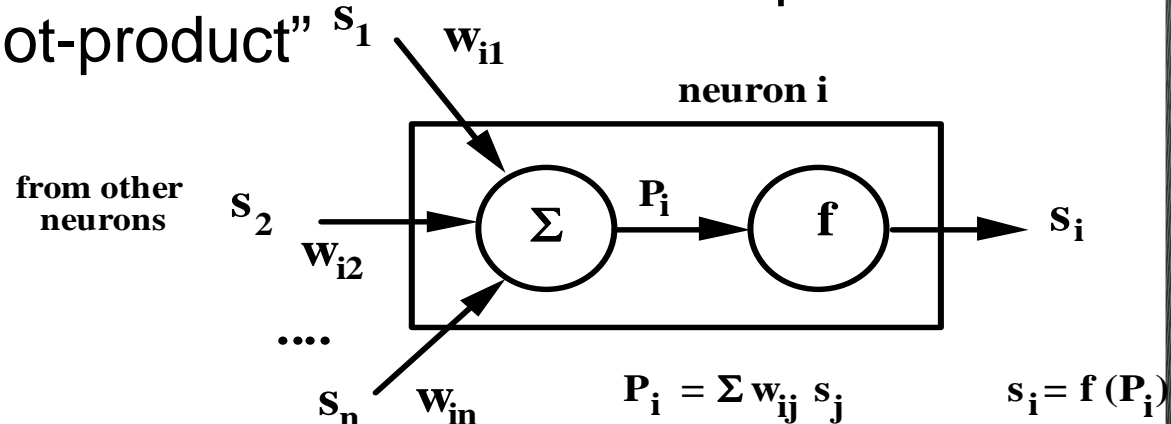
- Un ANN e' un insieme di unita' di processamento, chiamate neuroni e connesse tra loro.
- Ogni connessione e' caratterizzata da una "coupling constant" tra i due neuroni connessi. Per i neuroni  $i$  e  $j$ , l'accoppiamento e' definito da  $w_{i,j}$ , detto peso.



- Ogni neurone e' caratterizzato dallo stato  $s_i$  e dalla funzione di attivazione  $f$ .

# Il neurone “Dot-product”

- Un esempio di modello che descrive il comportamento di un neurone: il “dot-product”



- un neurone  $i$  calcola la somma pesata delle uscite dei neuroni ad esso connessi:

$$P_i = \sum_j w_{ij} \cdot s_j$$

- L'operazione consiste nel prodotto scalare di 2 vettori: quello dei pesi e quello degli stati di attivazione dei neuroni connessi al neurone  $i$ :

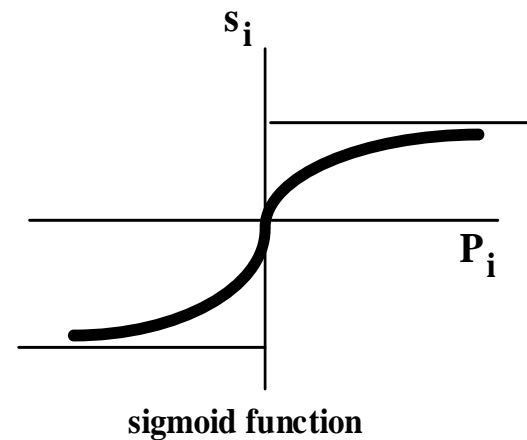
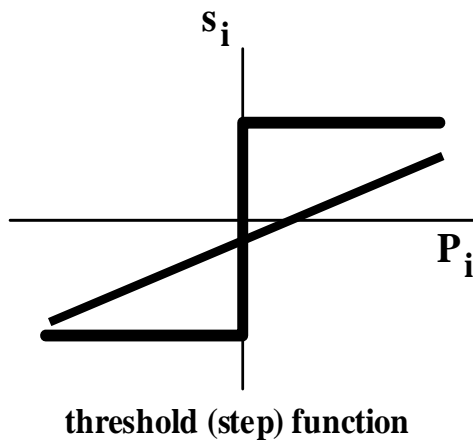
$$\vec{W}_i = \{w_{ij}\}; \vec{S} = \{s_j\}; \vec{P}_i = \vec{W}_i \cdot \vec{S}$$

- Il neurone agisce sulla somma pesata e genera un output (il suo stato) secondo la sua funzione di attivazione:  $s_i = f(P_i)$



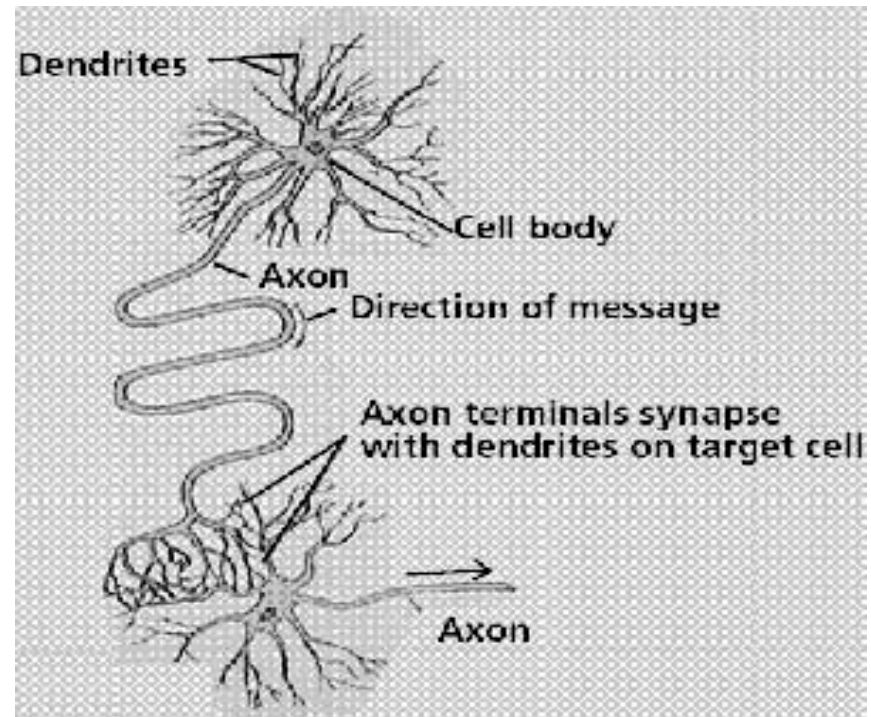
# Funzioni di Attivazione

- Esistono varie funzioni di attivazione
  - ◆ Generalmente, consistono nella applicazione di una soglia e/o in una amplificazione
- Le funzioni piu' utilizzate sono:
  - ◆ Una semplice funzione a 2 stati, in funzione di una soglia
  - ◆ Una funzione lineare
  - ◆ Una funzione non-lineare a forma di sigmoide



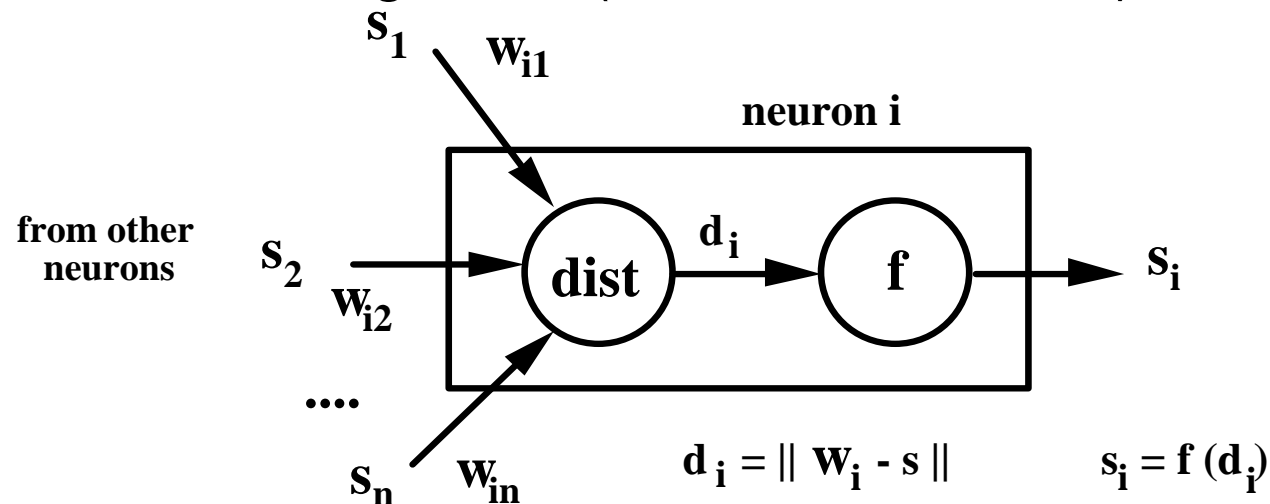
# Analogia con la biologia

The “dot-product” model of neuron is directly derived from biology: the synapsis plays the role of the connection between the dendritis and the neuron body: so the "weight" is in the synapsis, while the axons carry the information from a neuron to another.

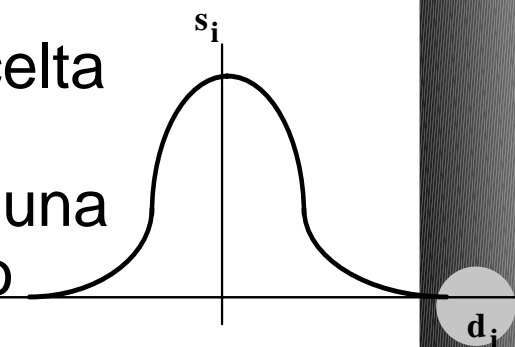


# Un altro esempio: “Distance neuron”

- Questo neurone calcola la “distanza” tra il vettore dei pesi ed il vettore degli stati (la norma euclidea):



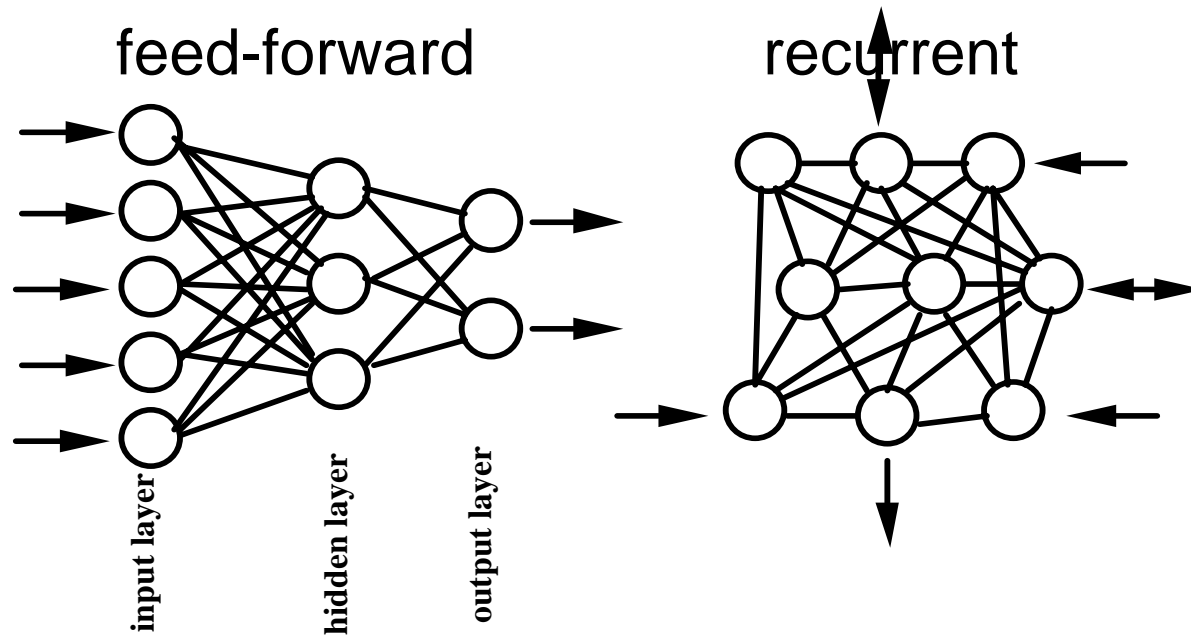
- In molte applicazioni, la funzione di attivazione applicata alla distanza viene scelta in modo tale da fornire valori tanto più alti quanto più i vettori confrontati sono simili; una funzione tipica è una campana con il picco centrato a distanza nulla.



bell-shaped function

# Architetture di Reti Neurali

- Il modo in cui i neuroni di un ANN sono connessi e' strettamente legato all'algoritmo di apprendimento usato nella fase di training
- I due ANN piu' comuni in applicazioni di Fisica sono le configurazioni:



Feed-Forward Network

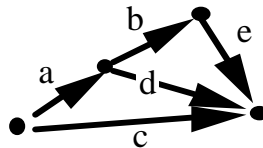
Recurrent Network

# A) Local Information Storage

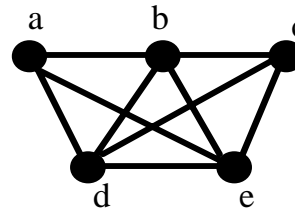
- Le informazioni sono univocamente associate allo stato dei neuroni
- I pesi corrispondono alla relazione tra due unita' di informazione determinata, in un approccio algoritmico convenzionale, dalla procedura computazionale

## A: LOCAL INFORMATION STORAGE IN NEURAL NET

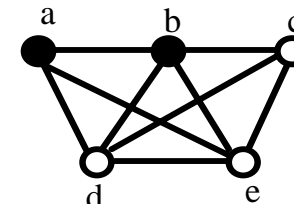
Each neuron  $\Leftrightarrow$  a physical instance  
e.g. neuron "a" active  $\Rightarrow$  points (X1,Y1) and (X2,Y2)  
belong to the track



Recurrent  
Neural Net



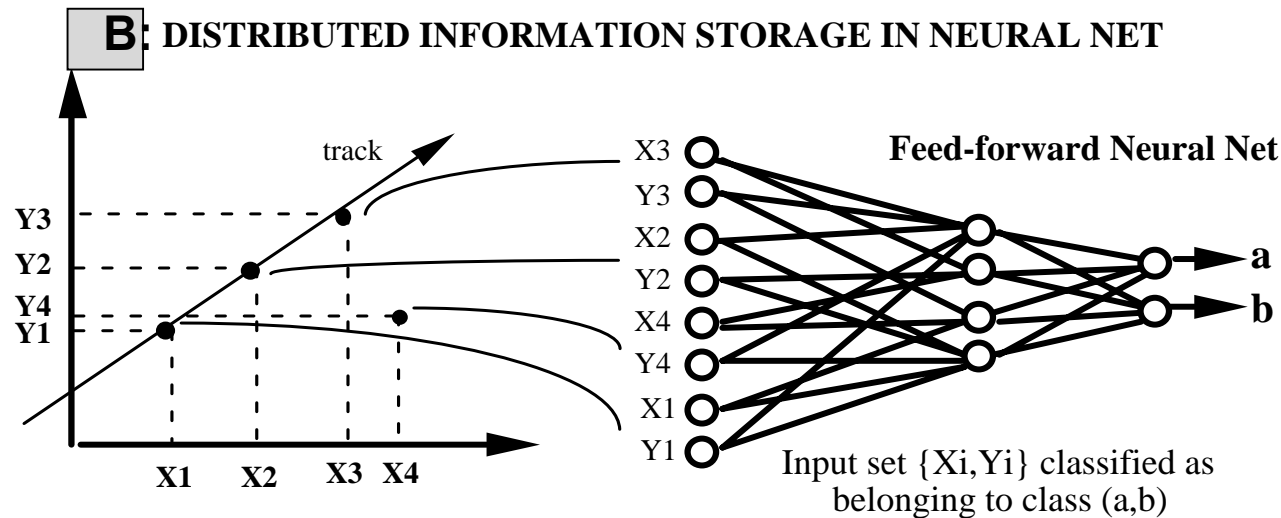
all neurons active  
at startup



only neurons a and b  
active at the end

# B) Distributed Information Storage

- I dati sono condivisi tra due o piu' neuroni
- I neuroni di input ricevono dati "reali" (es. Coordinate X,Y), i neuroni nascosti "mescolano" l'informazione
- In questo tipo di NN si puo' "tracciare" all'indietro il ruolo di ogni neurone
- Il calcolo dei valori ottimali dei pesi e' definito attraverso una procedura di apprendimento ("learning" o "training")





# Proprieta' degli ANNs

## 1) Distributed associative memory

- I valori dei pesi rappresentano lo stato di “conoscenza” dell' ANN
- se all'ANN viene fornito un input parziale, il risultato e' la scelta dell'input memorizzato piu' simile a quello parzialmente noto e la generazione dell'output  
mancanza o cattiva qualita' di parti dell'immagine
- generalizzazione: gli ANN possono gestire dati incompleti, rumorosi o mai visti prima

# Proprieta' degli ANNs

## 2) Fault tolerance

- la distruzione o alterazione di uno o piu' elementi di processamento causa (solo) una degradazione delle prestazioni: e' una conseguenza dello stoccaggio distribuito dell'informazione
- gli ANN sono particolarmente adatti ad applicazioni in cui non e' accettabile una "assenza" di controllo (es. centrali nucleari)

# Proprieta' degli ANNs

## 3) Pattern recognition

- il riconoscimento di strutture richiede capacita' di gestire velocemente grandi quantita' di input e di generare un risultato generalizzato
- un ANN e' un dispositivo computazionale intrinsecamente parallelo (e' noto che la capacita' del cervello di riconoscere rapidamente strutture complesse e' legata alla sua struttura parallela)

ricerca di lesioni in immagini mammografiche

# Come si usa una rete neurale?

- Apprendimento (learning)
  - ◆ definizione di tutti i parametri caratteristici a partire da un insieme di dati di input a cui corrispondono risultati noti
- Validazione (testing)
  - ◆ valutazione delle prestazioni del sistema a partire da un insieme di dati di input indipendente, di cui e' noto (ma non dichiarato al NN) l'output

# Procedure di apprendimento

- Ovvero, come si definiscono I valori dei parametri che descrivono le proprietà di una rete neurale, data la sua architettura
- Nei sistemi convenzionali la conoscenza viene espressa in termini di “regole” generali(zzate). L’apprendimento consiste nella individuazione delle regole...
- Negli ANN, la conoscenza e’ espressa in termini di un insieme di esempi il cui significato (risultato) e’ noto
- L’apprendimento (Learning) e’ un processo con cui i parametri liberi di un NN vengono adattati ampliando progressivamente la quantità di informazione nota
- Il tipo di apprendimento e’ definito dal modo in cui I parametri vengono modificati
- La validazione avviene “a posteriori”

# Apprendimento Supervisionato

- Per ogni insieme di valori di input, l'output desiderato viene dichiarato
- I pesi vengono adattati con una procedura iterativa, in modo da ottenere la corrispondenza ideale tra input e output
- Durante la fase di apprendimento, i pesi variano in funzione del tempo (del numero di iterazioni):

$$\mathbf{w}_{ij}(t) \rightarrow \mathbf{w}_{ij}(t+1) = \mathbf{w}_{ij}(t) + \Delta \mathbf{w}_{ij}$$

- Le regole di variazione dei pesi piu' importanti sono:
  - ◆ Hebbian learning
  - ◆ Competitive learning
  - ◆ Delta rule learning.



# Hebbian Learning

- Il peso associato alla connessione tra due neuroni cresce proporzionalmente al prodotto degli stati di attivazione:

$$\Delta w_{ij} = \eta s_i(t) s_j(t)$$

$\eta$ : learning rate

- Ovvero, in termini biologici, piu' una connessione viene utilizzata, piu' viene rafforzata

# Competitive Learning

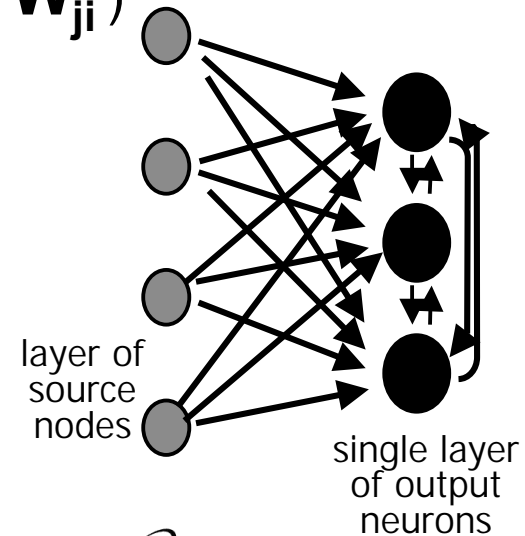
- Gli elementi di processamento alternativi “competono” tra loro; quello che da’ la risposta piu’ elevata – il vincitore – modifica il suo peso:

$$\Delta w_{ji} = \begin{cases} \eta (\mathbf{X}_i - \mathbf{W}_{ji}) & \text{(if the neuron } j \text{ wins the competition)} \\ 0 & \text{(if the neuron } j \text{ loses the competition)} \end{cases}$$

where  $\mathbf{X}_i$  = input pattern (a vector)

$\mathbf{W}_{ji}$  = weights from input  $i$  to unit  $j$   
(a row of the matrix  $\mathbf{W}_{ji}$ )

- Ogni neurone si specializza nel riconoscimento di “input patterns” a lui simili

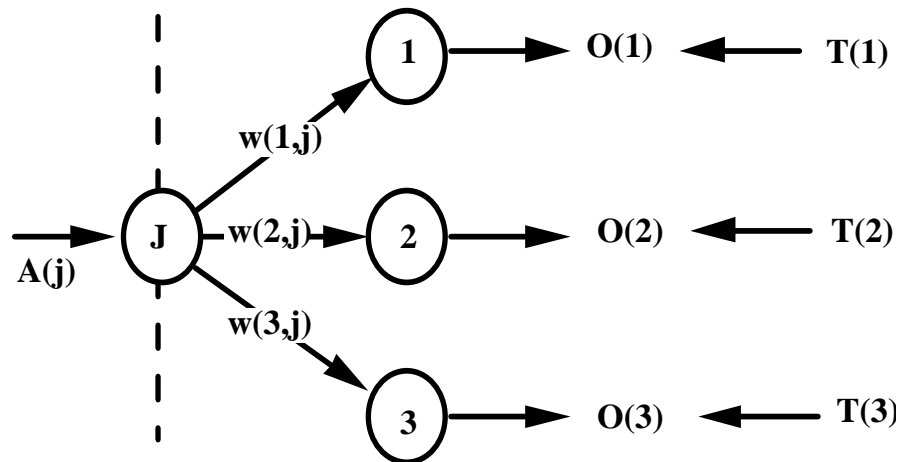


# Delta rule learning (1)

- E' basata sulla riduzione dell'errore tra l'output effettivo di un neurone e quello desiderato

Esempio: Two-layer FF-ANN.

- Assumendo di presentare un insieme di input patterns  $\{A_p\}$  allo strato iniziale del NN
- Per ogni input pattern, il NN produce un vettore di output  $O_p$ , da confrontare con il risultato desiderato  $T_p$



# Delta rule learning (2)

- La differenza  $O_p(i) - T_p(i)$  puo' essere calcolata per ogni componente del vettore  $O_p$ , e la funzione di errore puo' essere cosi' definita:

$$E_p = 1/2 ||O_p - T_p||^2 = 1/2 \sum [O_p(i) - T_p(i)]^2$$

- La Delta rule consiste nella modifica dei pesi proporzionalmente alla diminuzione del gradiente dell'errore rispetto a quel peso:

$$\Delta w(i,j) = - \eta \partial E / \partial w(i,j)$$

# Generalised Delta Rule

## Learning: Backpropagation

- In un NN multistrato, la delta rule puo' essere applicata ad ogni strato intermedio, utilizzando l'output di uno strato come input del successivo
- Questo metodo e' noto come "back-propagation", in quanto la propagazione parte dall'ultimo strato, di cui e' noto l'output desiderato
- Con una Generalized Delta Rule con back-propagation, la funzione di errore puo' avere vari minimi relativi, in cui si cerca di non restare intrappolati modificando il *learning parameter*  $\eta$ , anche durante la fase di apprendimento.
- Inoltre, per velocizzare la discesa, in particolare durante le iterazioni iniziali, si usa il parametro **a** (*momentum*), che tiene conto del fatto che all'iterazione precedente la variazione del peso era grande e consente alla funzione E di variare piu' rapidamente:

$$\Delta_n \mathbf{w}(i,j) = - \eta \partial E_n / \partial \mathbf{w}(i,j) + \mathbf{a} \Delta_{n-1} \mathbf{w}(i,j)$$

- I valori iniziali dei pesi e delle soglie di attivazione vengono scelti in modo casuale
- Una iterazione dell'intero insieme di vettori di input e' detta *epoca*

# Learning and testing (1)

- Innanzitutto, occorre dividere l'insieme delle configurazioni di input (e gli output corrispondenti) in due gruppi:
  - ◆ **training sample**
  - ◆ **testing sample**
- La prima fase dell'apprendimento consiste nella presentazione delle configurazioni del training sample e nell'applicazione dell'algoritmo scelto per l'adattamento iterativo dei pesi del NN. La procedura viene ripetuta finché l'errore (es. La somma dei quadrati delle differenze tra i valori di output attesi e quelli ottenuti dal NN) raggiunge un minimo
- La convalida dei valori dei pesi così ottenuti viene effettuata utilizzando il testing sample per misurare le prestazioni (la sensibilità/efficienza) del NN



# Learning and testing (2)

- Non ci sono regole che definiscono in modo univoco il numero di configurazioni di input da utilizzare nella fase di learning
- E' opportuno che il numero di input patterns sia almeno alcune volte piu' grande del numero di parametri da ottimizzare.
  - ◆ Quindi, quando non e' disponibile un campione numeroso di input patterns, e' necessario che l'architettura del NN sia semplice (ovvero il numero di neuroni deve essere piccolo)
- Il numero di unita' nascoste e' in generale determinato in modo empirico, come il minimo numero che consente buone prestazioni. Infatti, se le unita' nascoste sono troppe, si rischia un "overtraining"
  - ◆ il NN memorizza gli input pattern in modo eccellente con il campione di "learning", ma poi funziona male con il campione di "testing": il NN non sa piu' generalizzare

# Riepilogando:

## Vantaggi degli ANN:

- Gli ANN possono risolvere problemi estremamente complessi, non affrontabili con algoritmi tradizionali
- imparano
- Lavorano in “tempo reale”
- Consentono di gestire con risultati ragionevoli dati incompleti o di cattiva qualità

## Limiti degli ANN:

- Gli ANN sono “black-boxes”: I problemi vengono risolti ma non capiti (“modellizzati”)
- Grandi quantità di dati sono necessari come “training set”

# Applicazioni degli ANN

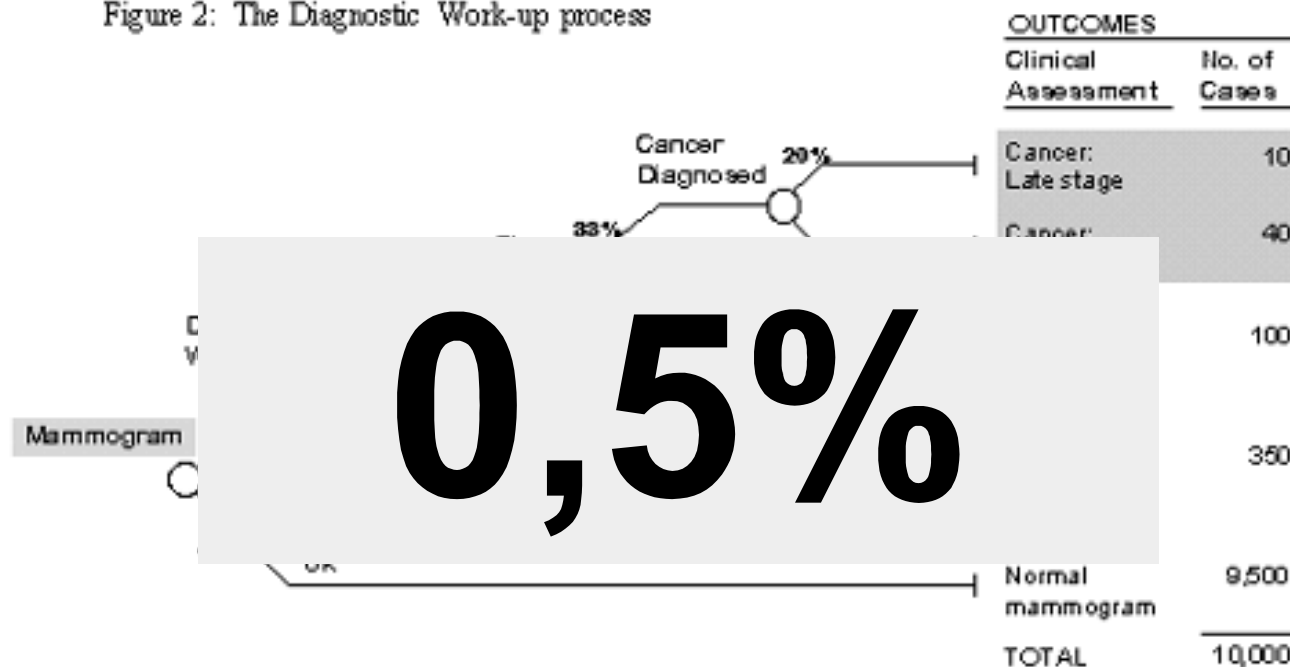
- Handwritten character recognition
- Phonetic typewriter
- Sonar signal recognition
- Electronic nose
- Image compression
- Noise reduction
- Self-driving devices
- Credit assignment
- Monitoring the condition of machinery
- Travelling Salesman's Problem (TSP)
- Stock Market predictions
- Computer-Assisted-Detection (CAD) systems in medical imaging

# Considerazioni generali

Obiettivo:

- Alta sensibilità per rilevare tutte le possibili lesioni
- Alta specificità per evitare inutili approfondimenti diagnostici

Figure 2: The Diagnostic Work-up process



American Cancer Society  
[www.r2tech.com](http://www.r2tech.com)

5 marzo 2004

Piergiorgio Cerello (cerello@to.infn.it)



36

# Chi si ammala di cancro alla mammella?

*American Cancer Society Research, 2001*

Probabilità di sviluppare il cancro in funzione dell'età:

<b>Età</b>	<b>Probabilità di sviluppare il tumore alla mammella nei prossimi 10 anni</b>	<b>e cioè 1 su...</b>
20	0.05%	2044
30	0.40%	249
40	1.49%	67
50	2.77%	36
60	3.45%	29
70	4.16%	24

# Prevenzione del tumore alla mammella

## 1. Rivelazione precoce

Riduzione della mortalità grazie alla maggiore efficacia del trattamento terapeutico applicato

## 2. Controllo periodico

Nella fase iniziale la malattia è asintomatica

E' opportuno che le donne si controllino regolarmente entro l'intervallo di tempo di sviluppo della malattia

## 3. Doppia lettura

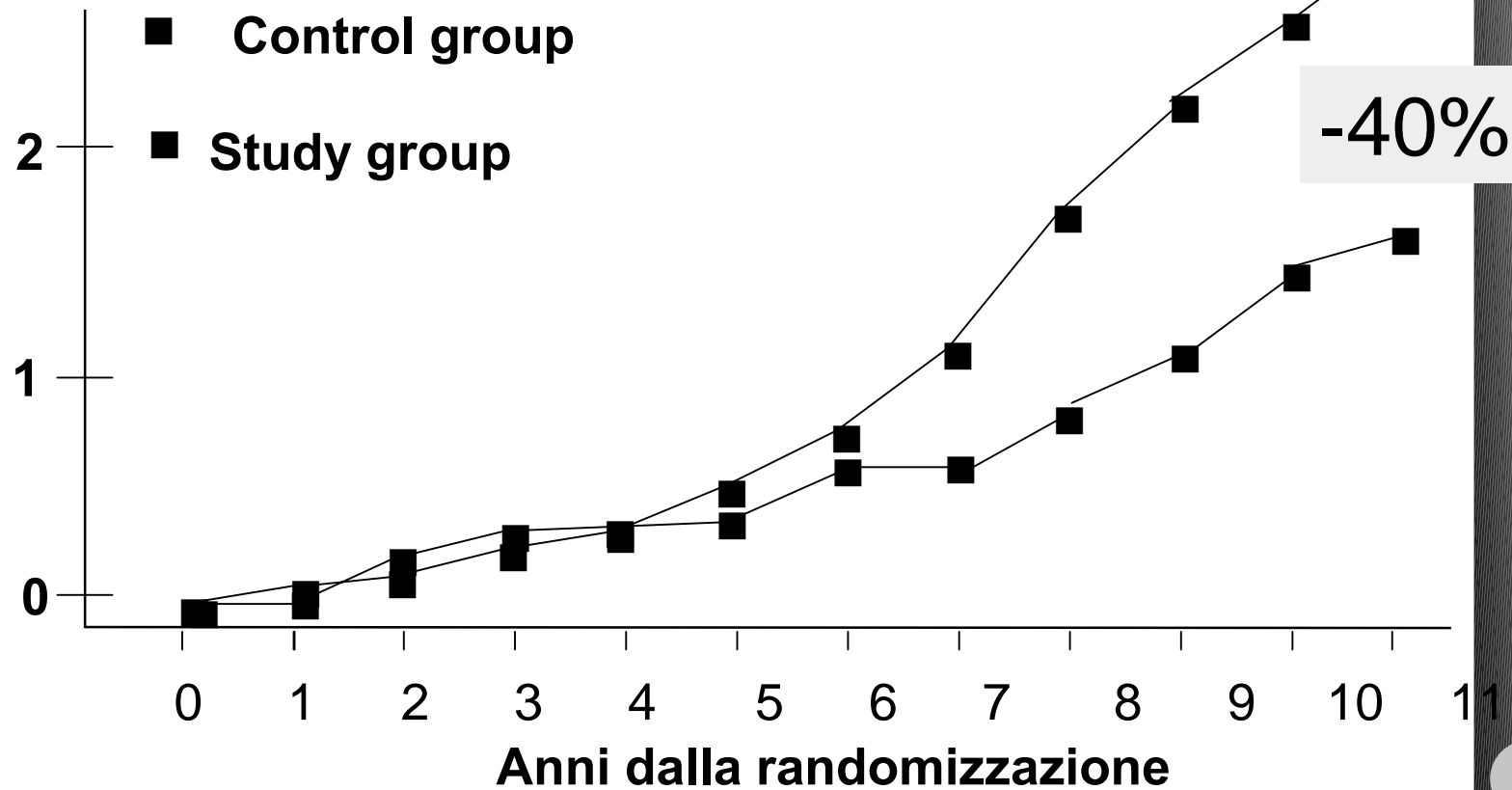
E' dimostrato che con una doppia lettura indipendente delle mammografie si ha un incremento di sensibilità fino al 15%

# Screening

Gothenburg screening

Popolazione: 39-49

MORTALITA' CUMULATIVA PER  
CARCINOMA MAMMARIO (per mille)



*Bjurstam et al., Cancer, 1997*

5 marzo 2004

Piergiorgio Cerello (cerello@to.infn.it)



39



# Breast Cancer Screening

## ➤ Breast Cancer Screening

- ◆ Increased survival rate

Sensitivity  
(positives/true positives) → 73% - 88%

Specificity  
(negatives/true negatives) → 83% - 92%

2% - 10% increase with *double reading*

- ◆ Problems: costs and manpower
- ◆ Computer Assisted Detection

# Costi

	COSTO MEDIO (\$)	N° PAZIENTI	COSTO TOTALE (\$)
Mammografia	90	10900	981000
Approfondimento diagnostico	500	500	250000
Biopsia	2000	160	320000
TRATTAMENTO			
Stadio iniziale	11000	40	440000
Stadio avanzato	140000	13	1820000

*American Cancer Society*  
*www.r2tech.com*

5 marzo 2004

Piergiorgio Cerello (cerello@to.infn.it)



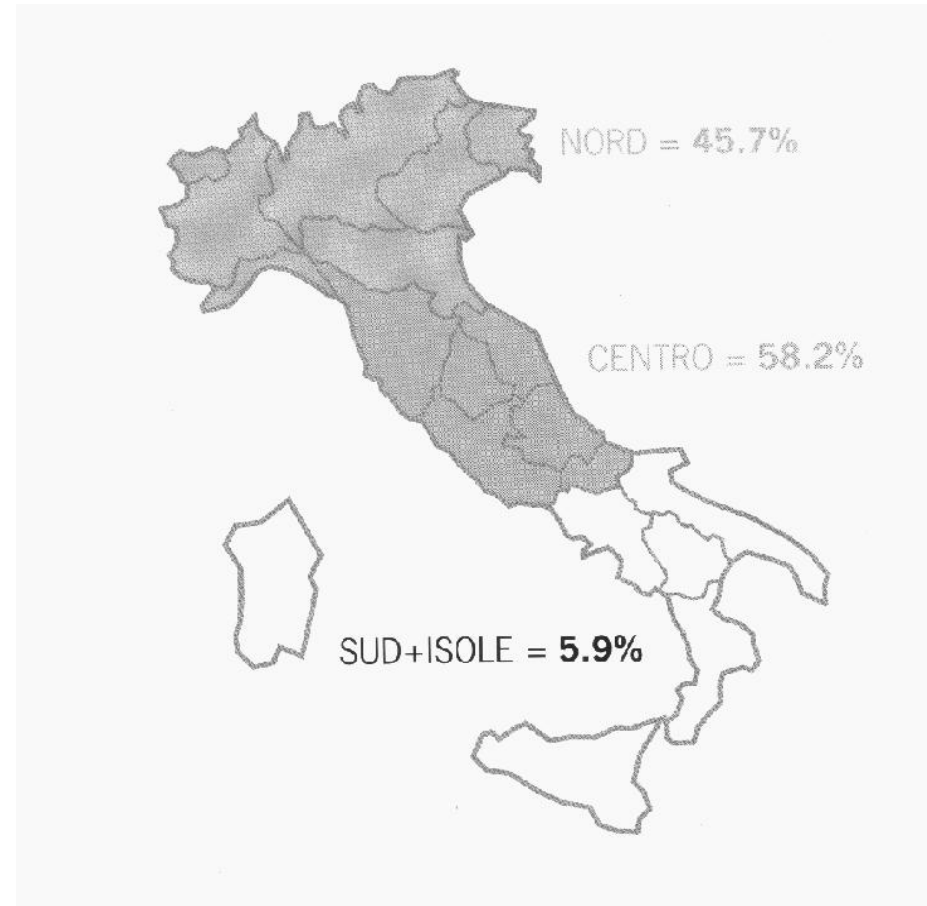
41

# Screening

Dati GISMA\* 2000 sullo screening

CONTROLLO  
PERIODICO  
+  
DOPPIA LETTURA

Sensibilita'  
+ 15%

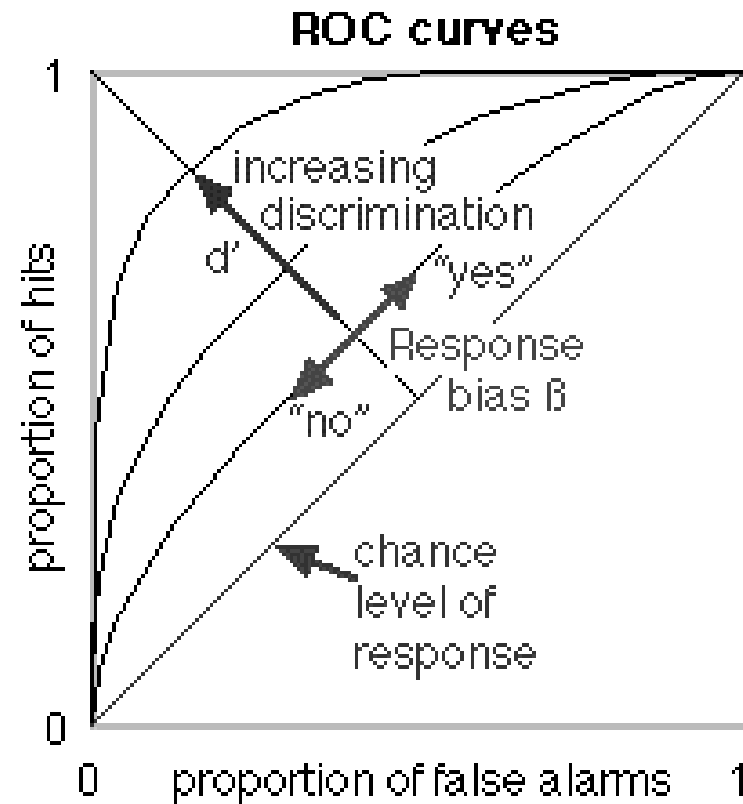


# Computer Assisted Detection

- Analisi computerizzata delle immagini che fornisce una *lettura alternativa* al medico radiologo
- Indica le regioni di interesse potenzialmente sospette (ROI), associandovi una *probabilità*'
- Lascia al radiologo la decisione finale
- Utilizzato sia in mammografia clinica che in programmi di screening

# Computer Assisted Detection

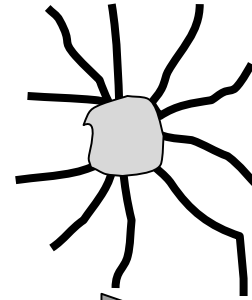
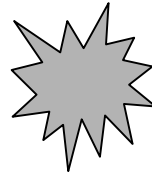
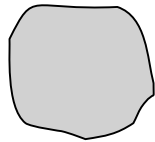
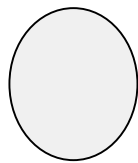
- Sensibilita' (efficienza):
  - ◆ Positivi/(Totale positivi)
- Specificita':
  - ◆ Negativi/(Totale negativi)
- Percentuale di Falsi negativi:
  - ◆  $1 - \text{Sensibilita'}$
- Percentuale di Falsi positivi:
  - ◆  $1 - \text{Specificita'}$



STUDIO	OBIETTIVO	$A_z$	
		senza CAD	con CAD
Getty et al, 1988	Differenziazione tra lesioni benigne e maligne nei mammogrammi	<b>0.832</b>	<b>0.880</b>
Chan et al.,1990	Rivelazione di cluster di microcalcificazioni in mammogrammi	<b>0.924</b>	<b>0.953</b>
Kegelmeyer et al.,1994	Rivelazione di lesioni massive in mammogrammi (sens/spec)	sens: <b>0.806</b> spec: <b>0.954</b>	sens: <b>0.903</b> spec: <b>0.969</b>
Jiang et al.,1999	Differenziazione tra cluster di microcalcificazioni benigne e maligne nei mammogrammi	<b>0.614</b>	<b>0.755</b>
Kobayashi et al., 1996	Rivelazione di noduli polmonari in radiografie del torace	<b>0.894</b>	<b>0.940</b>
Difazio et al., 1997	Rivelazione di variazioni polmonari nel tempo	<b>0.887</b>	<b>0.984</b>
Monnier et al., 1998	Rivelazione di infiltrazioni interstiziali nelle radiografie del torace	<b>0.948</b>	<b>0.970</b>

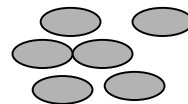
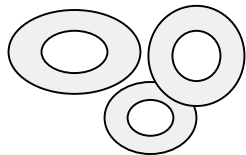
# Ricerca di Lesioni

Esistono due tipi di lesioni  
su cui concentrare la ricerca:



MASSE

**Grado di malignità**

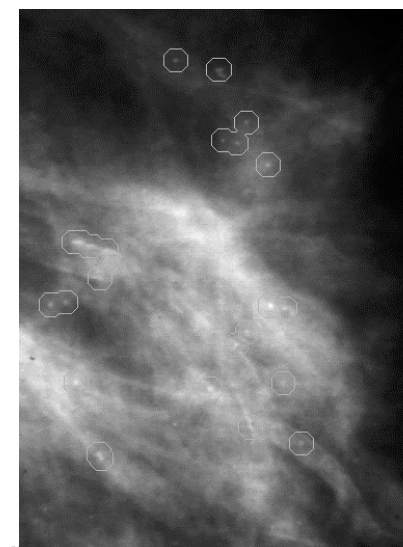
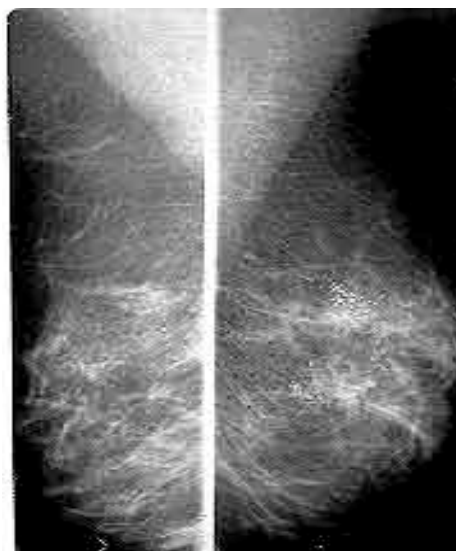
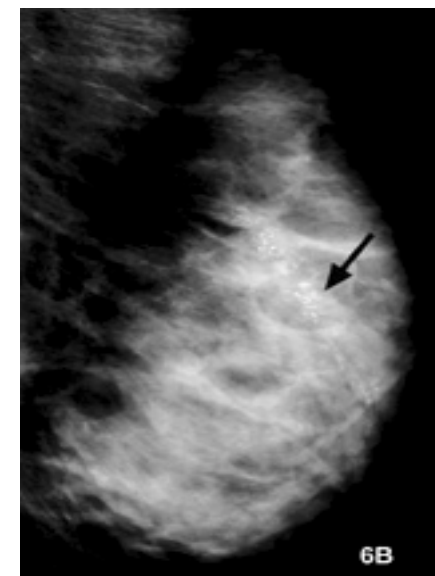
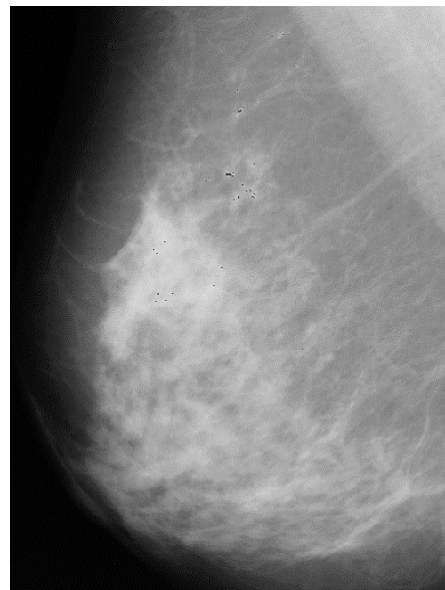
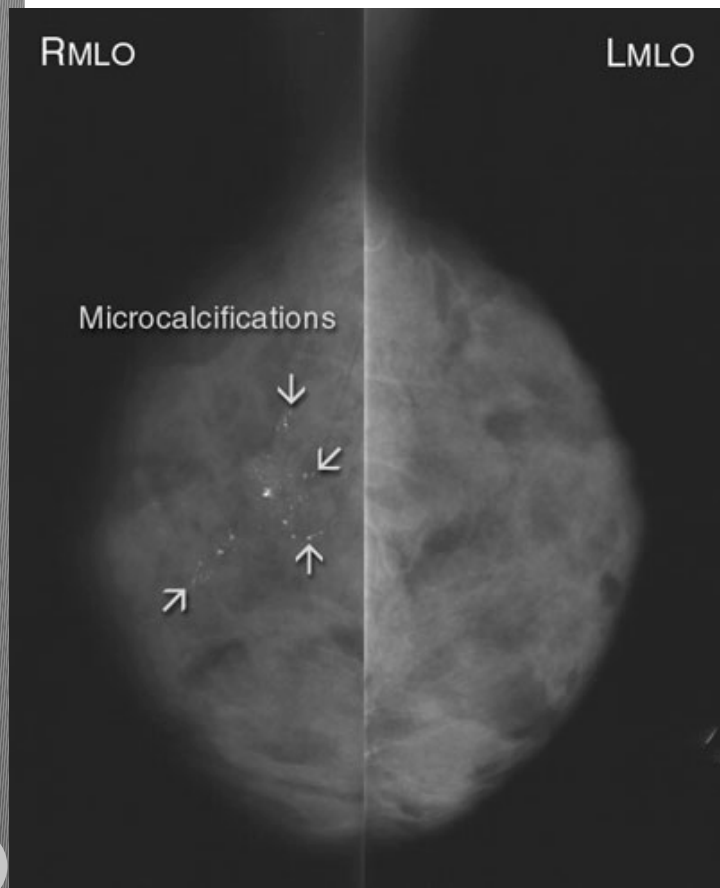
A large, grey, right-pointing arrow with a black outline, positioned horizontally across the middle of the slide. The text 'Grado di malignità' is written inside the arrow.

MICROCALCIFICAZIONI



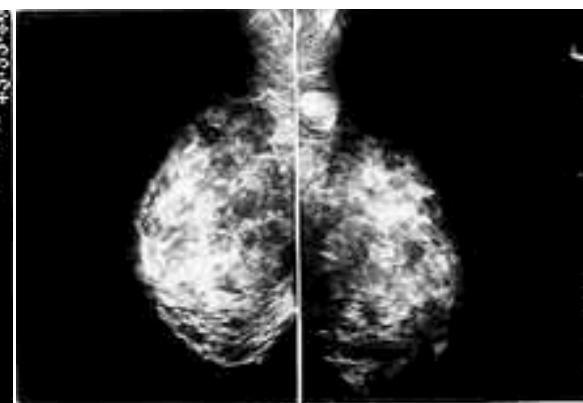
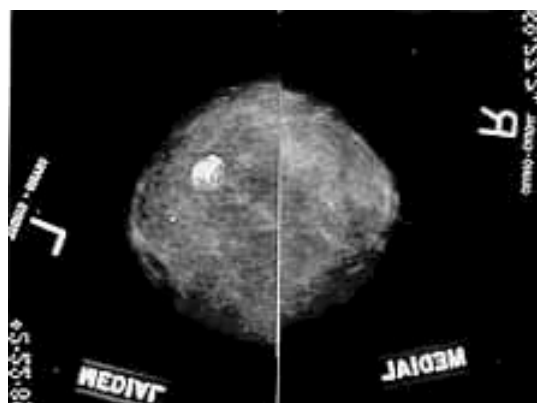
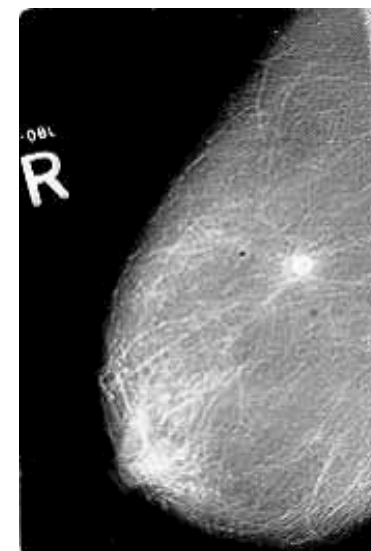
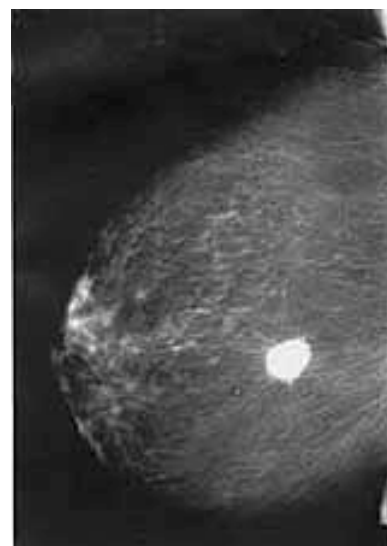
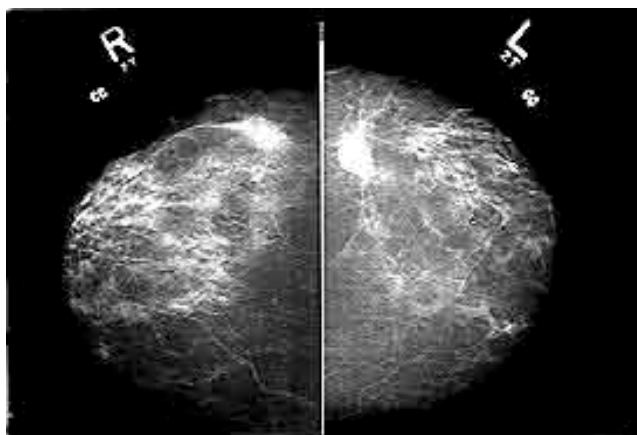
# Ricerca di lesioni

## Microcalcificazioni



# Ricerca di lesioni

Lesioni massive



5 marzo 2004

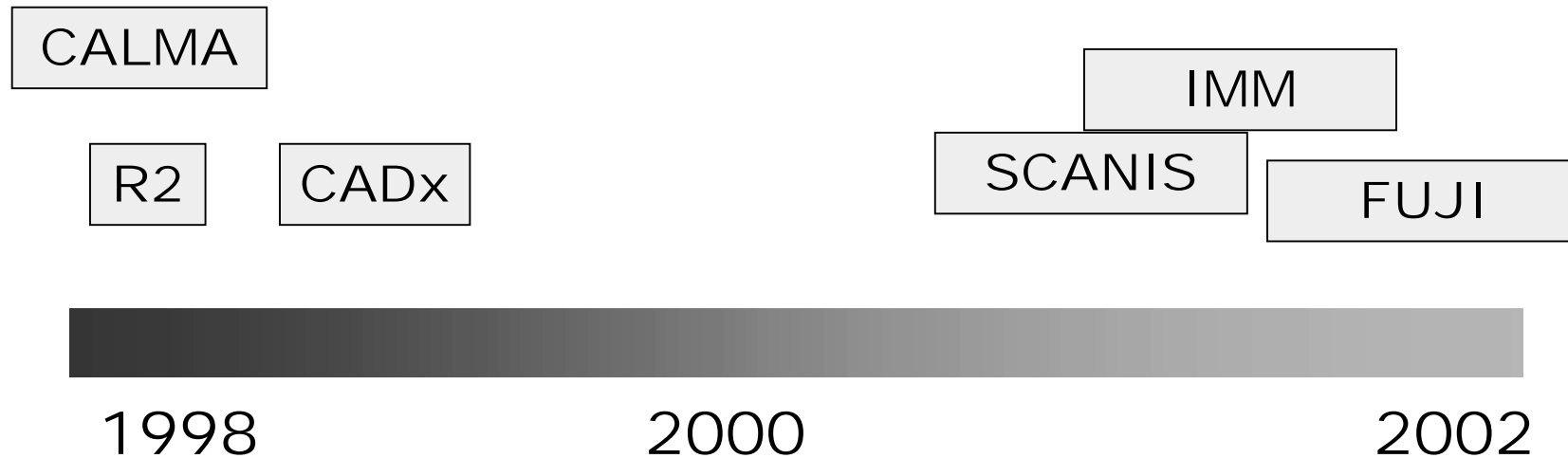
Piergiorgio Cerello (cerello@to.infn.it)



48

# CAD esistenti

Esistono diversi sistemi CAD in commercio e nel mondo della ricerca:



# CALMA

## ALGORITMI DI RICERCA

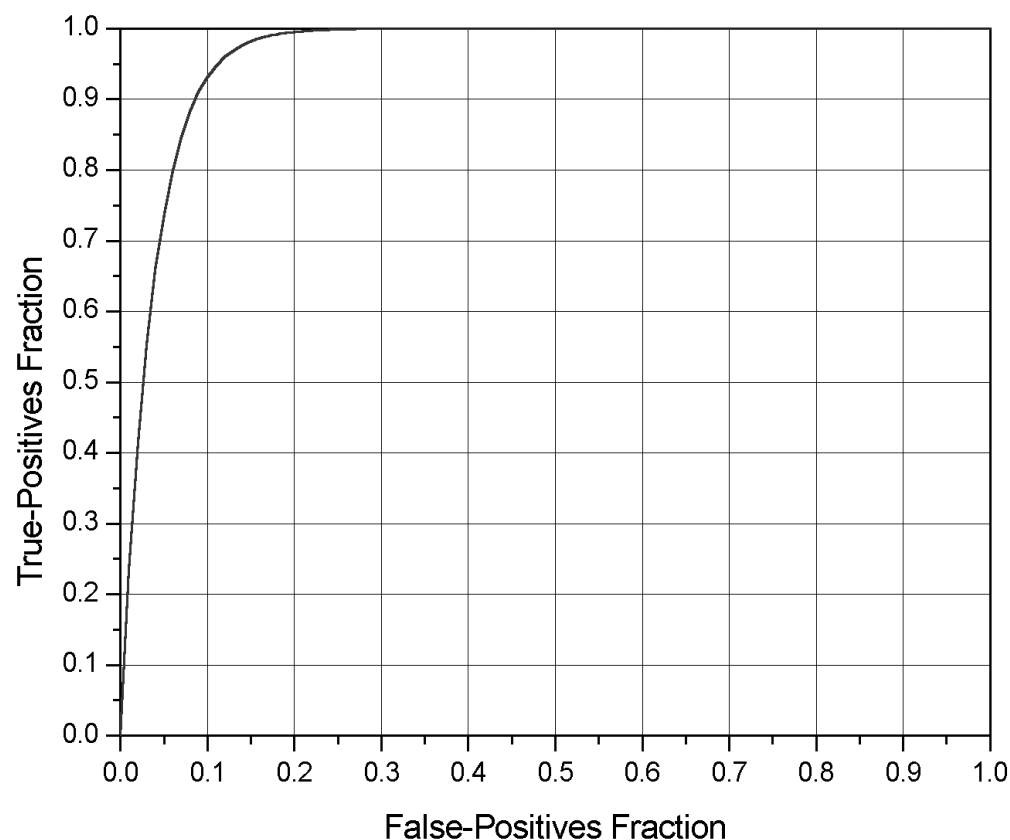
- Per le  
Sang  
con S
- Per le  
imple

## VALIDAZI

- 5000
- Circa
- ROC

5 marzo 2004

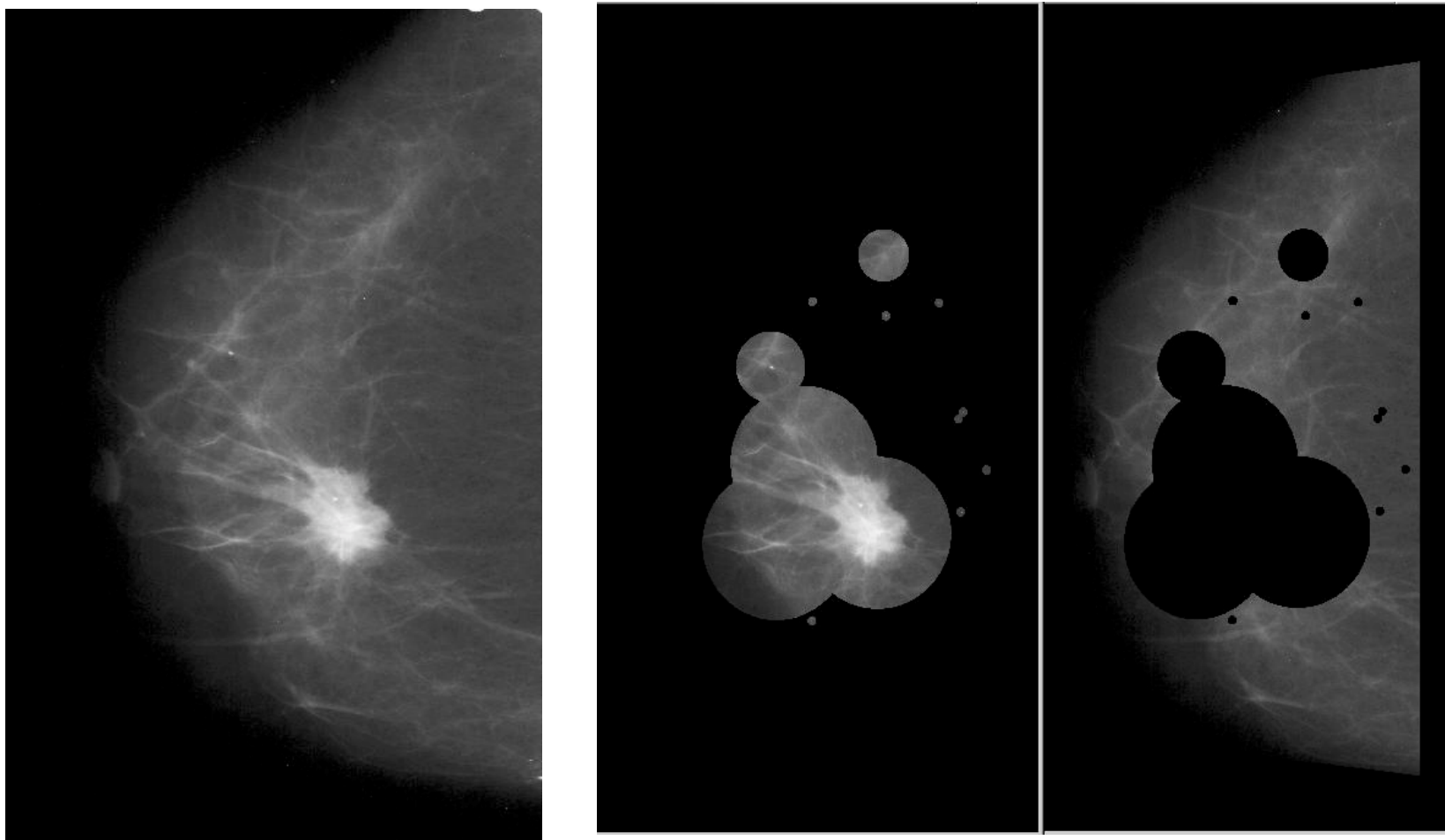
865 mammograms (370 with 495 without microcalcification clusters)



tipo  
e

# Ricerca di lesioni massive

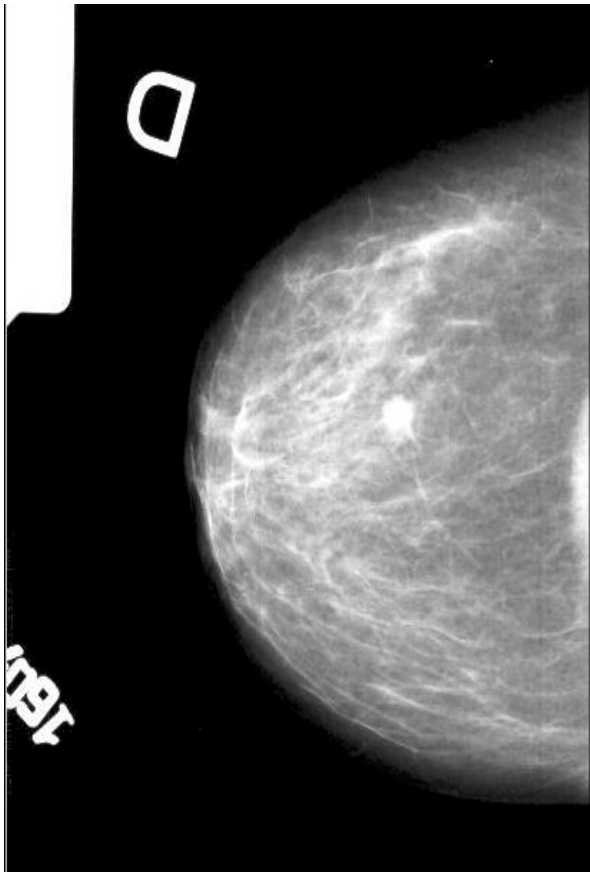
*Estrazione delle ROI*



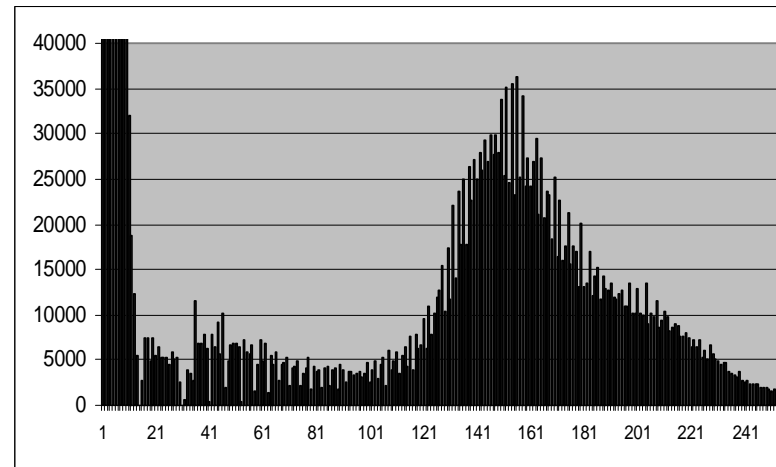


# Ricerca di lesioni massive

## *Image processing*



Distribuzione dell' intensità  
dei pixel ( livelli di grigio )



## *Estrazione di caratteristiche*

- l'intensità media dei pixel
- la varianza
- la asimmetria

# Ricerca di lesioni massive

## *Reti Neurali*

- **Feed-Forward back-propagation con momentum**

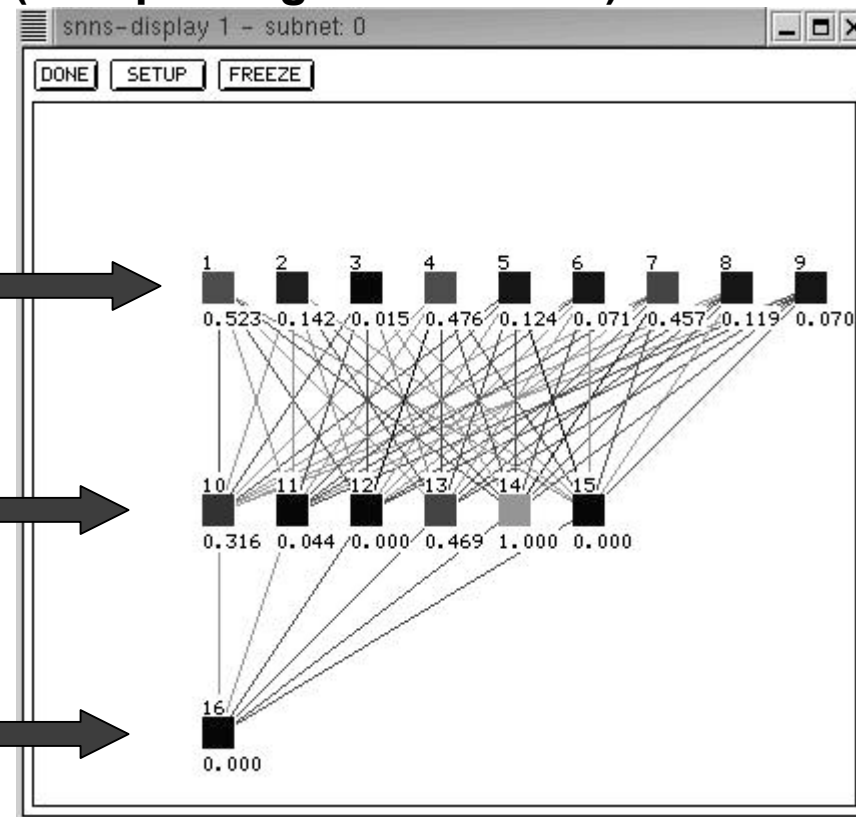
**Training set: 515 pattern ( 102 patologici e 413 sani).**

**Testing set: 515 pattern ( 102 patologici e 413 sani).**

- **9 neuroni di ingresso** →

- **6 neuroni nascosti** →

- **1 neurone d'uscita** →

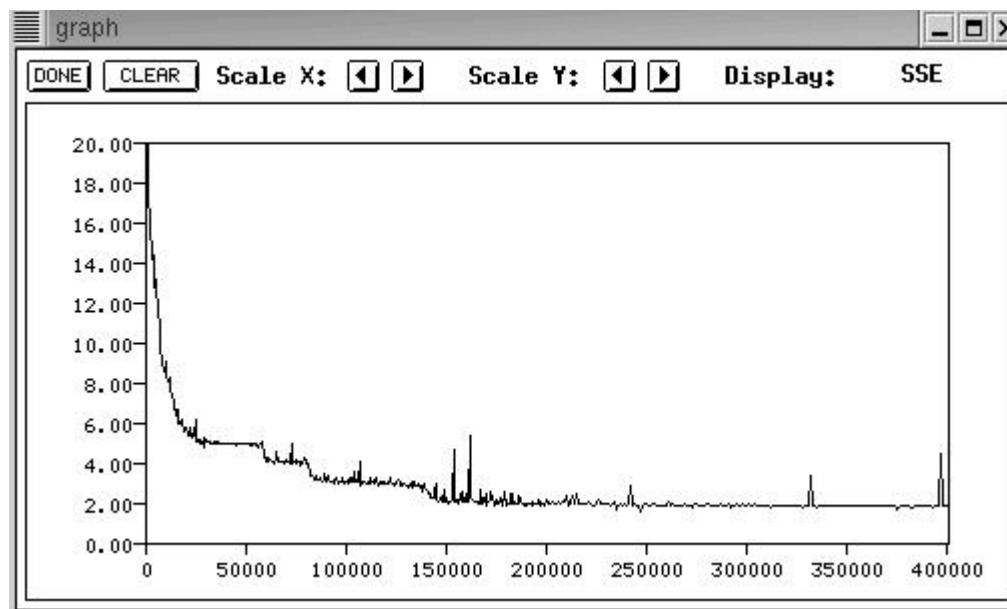




# Ricerca di lesioni massive

## *Reti Neurali : Apprendimento*

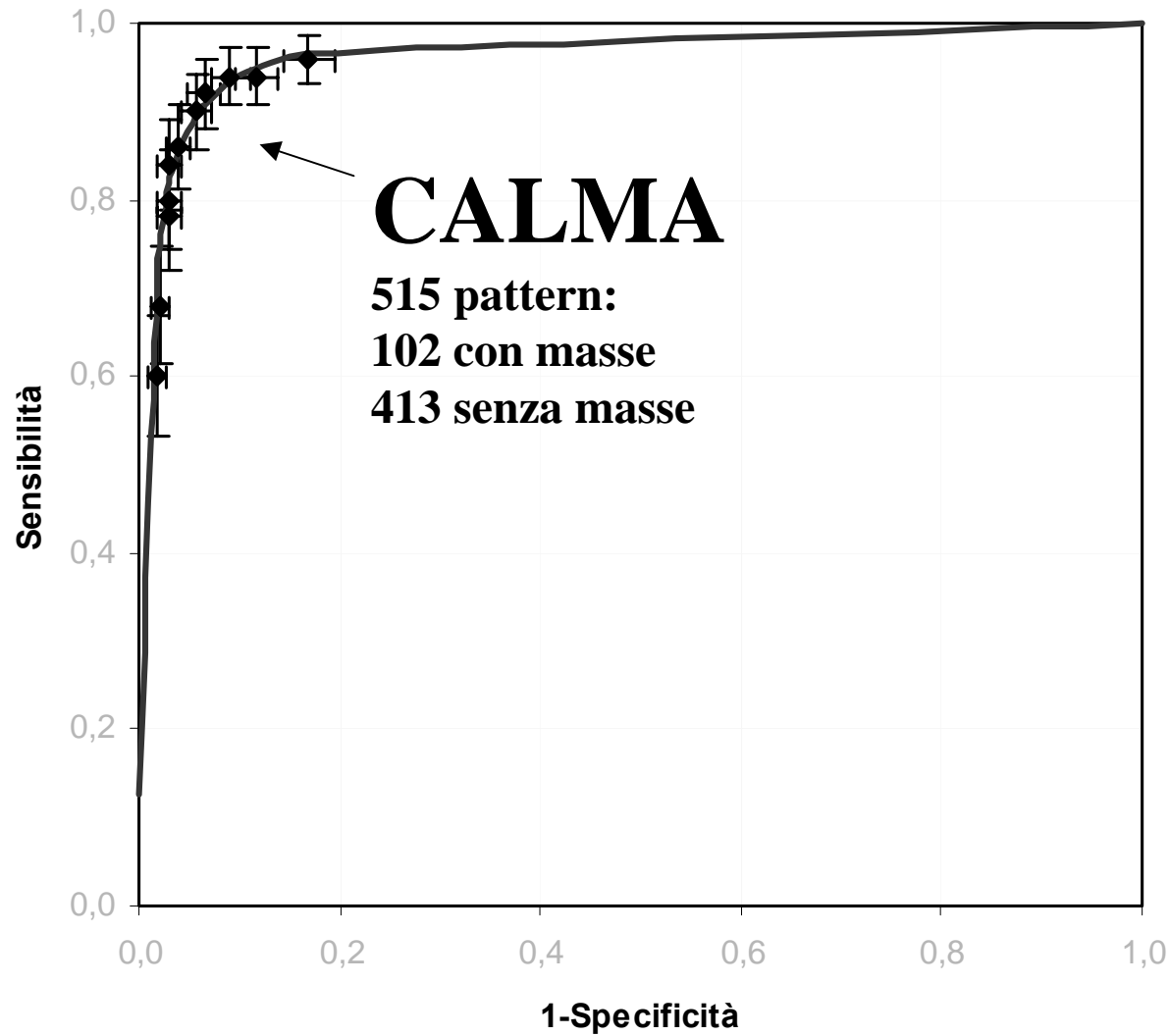
400.000 epoche



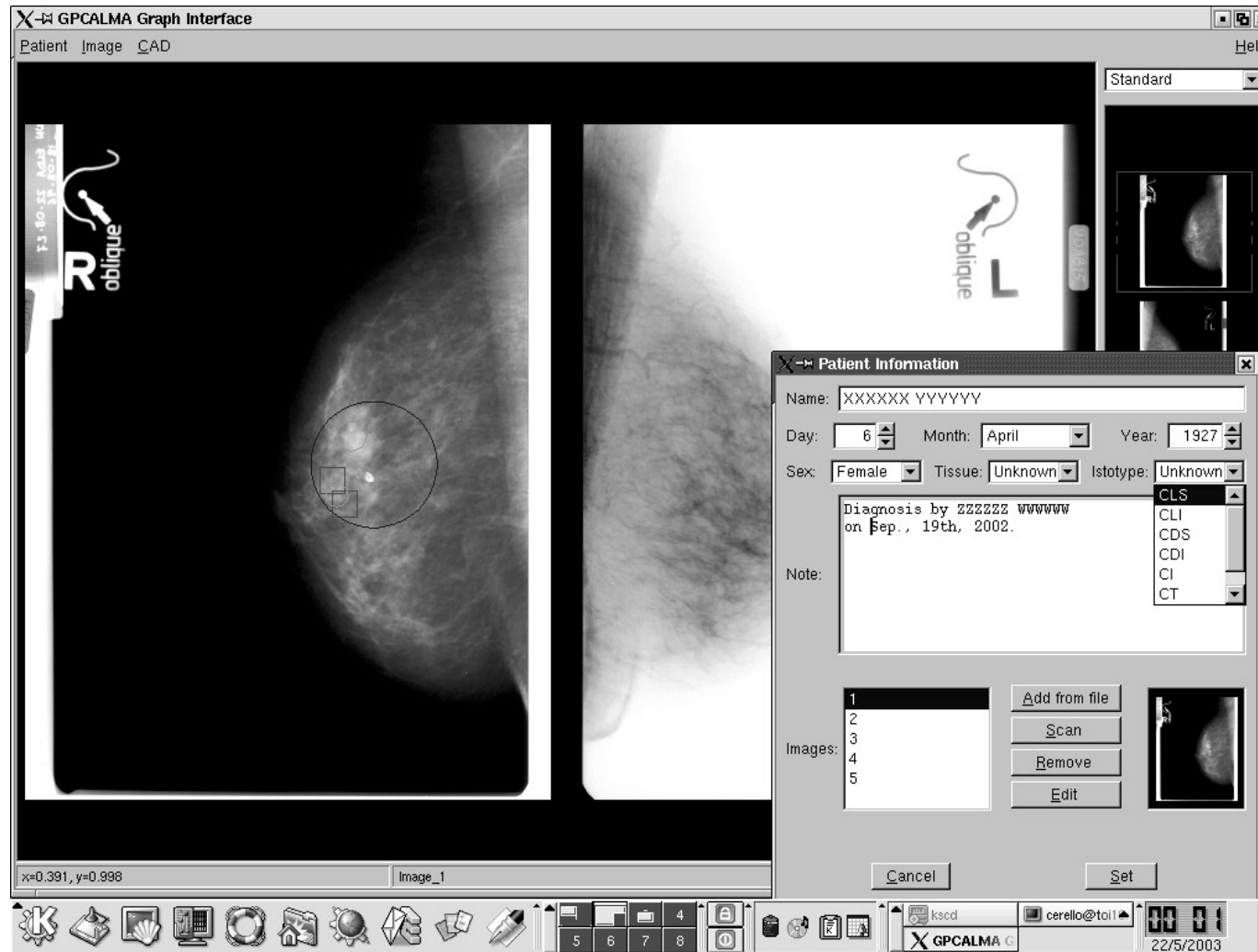
## *Reti Neurali : Test*

Ogni nuovo pattern di ingresso viene elaborato  
Il risultato e' **il valore della probabilità**  
**che la ROI contenga una lesione**

# Risultati: lesioni massive



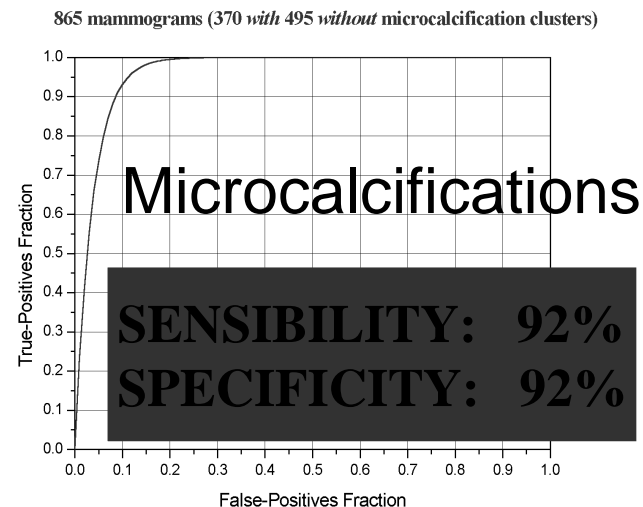
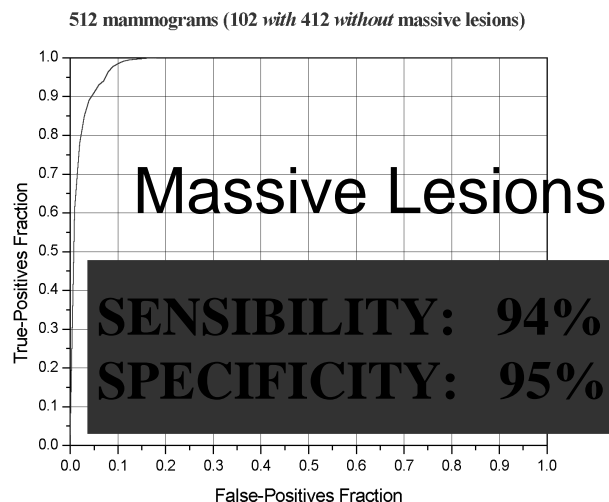
# The GPCALMA Graphic User Interface



In use:  
Bari  
Napoli  
Sassari  
Torino

# CALMA Results

- Largest Database of digitised mammograms ( > 5000)
- ROC (Receiver Operating Characteristic) Curve



Improved Sensitivity & Reduced Specificity

	Radiologist	+ CAD X	+ CALMA		Radiologist	+ CAD X	+ CALMA
A	82.8 (4.5)	94.3 (2.8)	94.3 (2.8)	A	87.5 (3.0)	84.2 (3.3)	87.5 (3.0)
B	80.0 (4.8)	88.2 (3.8)	90.0 (3.6)	B	91.7 (2.6.)	85.9 (3.2)	88.4 (2.9)
C	71.5 (5.4)	82.9 (4.5)	87.1 (4.0)	C	74.2 (4.0)	70.8 (4.2)	70.9 (4.1)

# CALMA Open Issues

## ➤ Virtually unlimited Database size

- ◆ Intrinsically distributed Database – many sources
- ◆ Poor network connections
- ◆ Access required to all the images

## The “GRID philosophy” in mammographic CAD

## ➤ Use Cases

- ◆ Large Scale Screening
- ◆ Teleradiology: diagnosis & training

Example: Italy  
4 mammograms/exam  
(60 MB)/exam  
6.7 Mpeople, 1 exam/2y  
3.35 Mexams/year  
about 200 TB/year

**1 PB/year on the  
European scale  
Huge amount of  
distributed data**

# GRID

- Distributed System (hardware infrastructure and software to manage it) which allows a Community of Users to share Computing Resources and access Distributed Data
- No standard yet – under construction in GGF
- Many projects in EU & US
- Major Components
  - ◆ **Virtual Organisation (users community)**
  - ◆ **Huge amounts of Distributed Data (storage)**
  - ◆ **Distributed Computing Resources (computing)**
  - ◆ **Network connections (infrastructure)**

**The System complexity**  
**must be hidden**

# GRID Services

## 1: "Quasi - static" Services: framework related

**Computing FARM Configuration**

**Storage System Maintenance**

**User Authorization**

**Monitoring Configuration:**

**Network, Computing & Storage**

**Application Software**

## 2: "Dynamic" Services: job related

**User Authentication**

**Data Management:**

**Registration, Access & Replication**

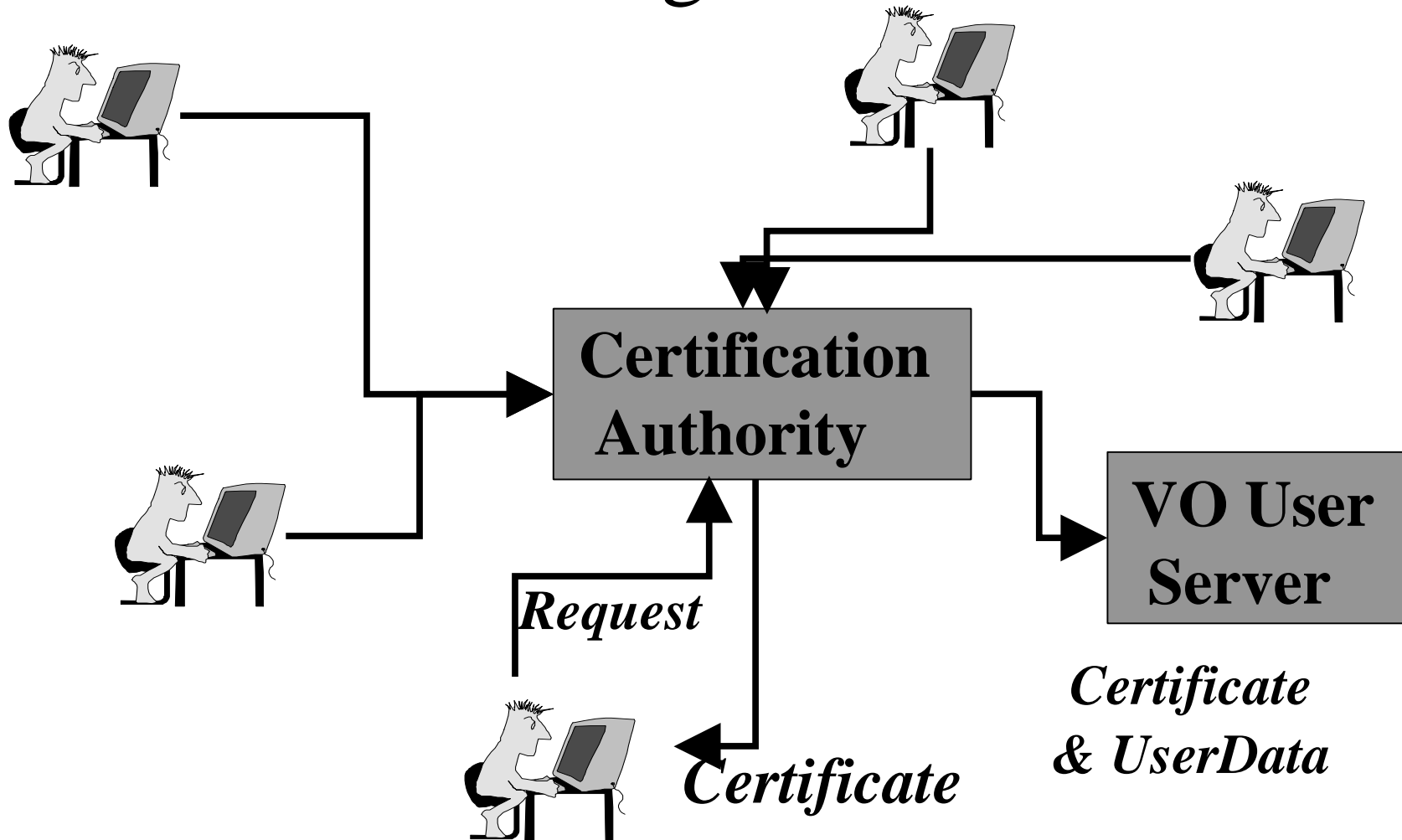
**Job Management**

**Monitoring**



# GRID Security & Management

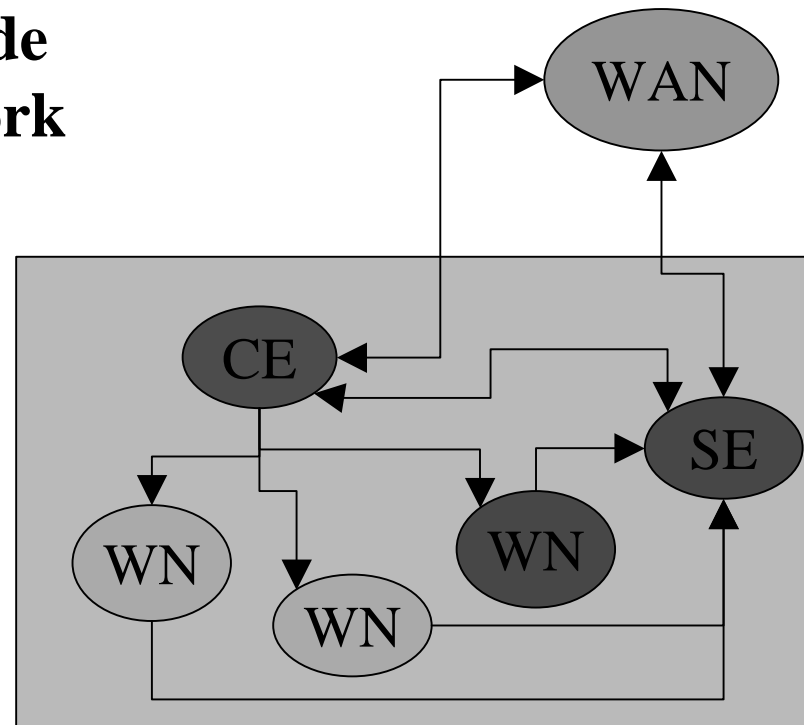
User



# GRID Computing Farm Configuration

**WAN: Wide Area Network**

**CE: Compute Element**

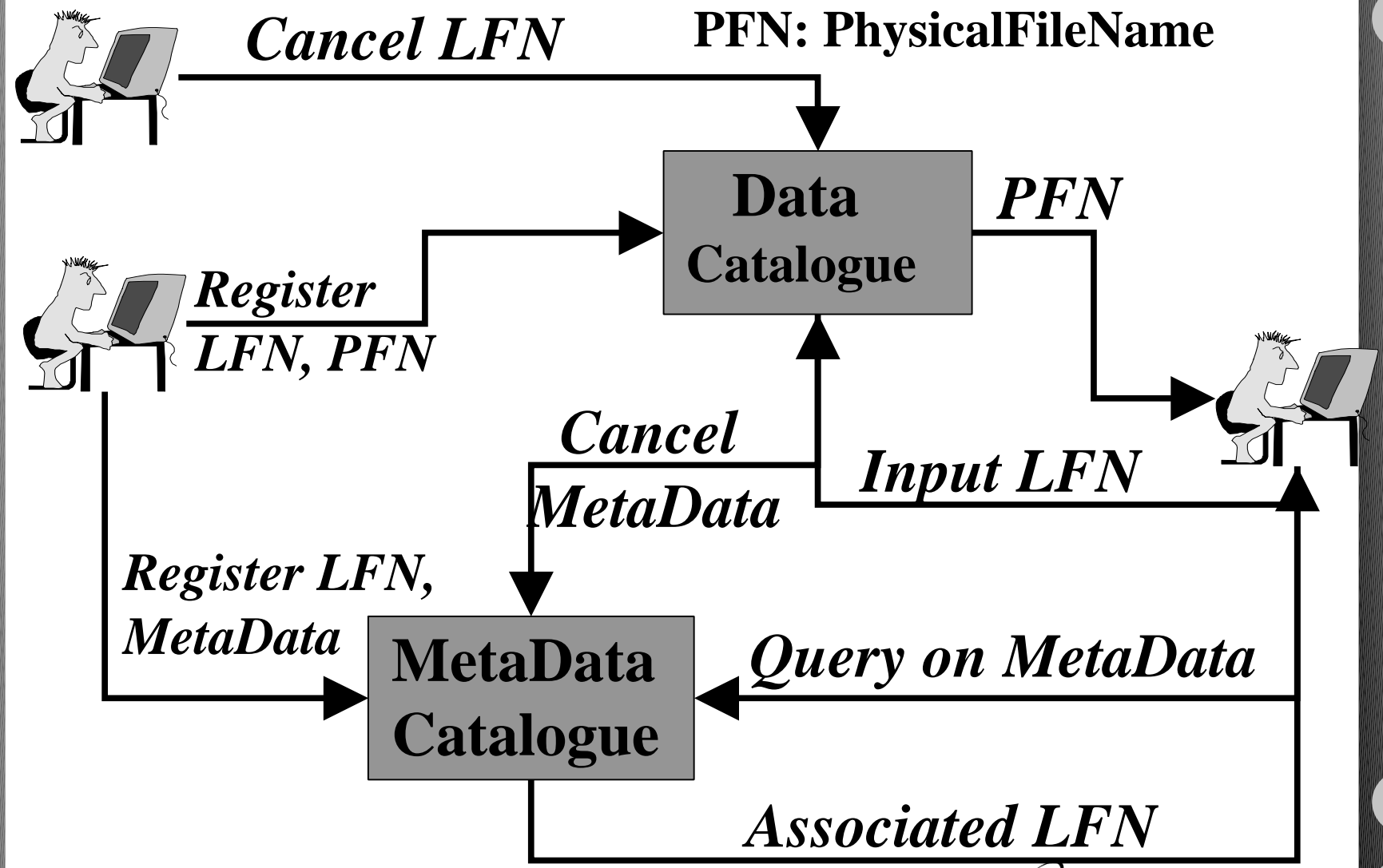


**SE: Storage Element**

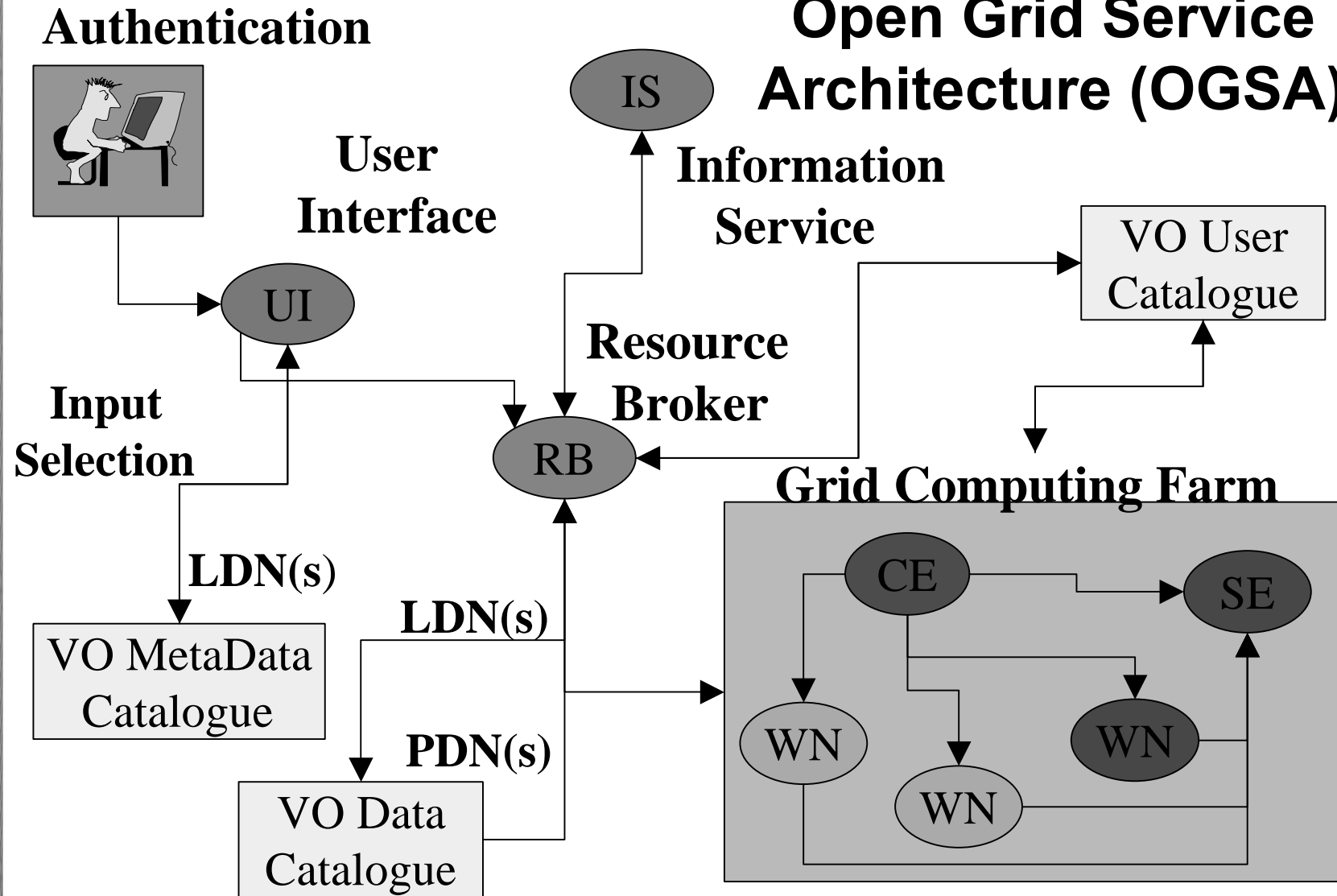
**WN: WorkerNode**

# GRID Data Management

LFN: LogicalFileName  
PFN: PhysicalFileName



# Open Grid Service Architecture (OGSA)

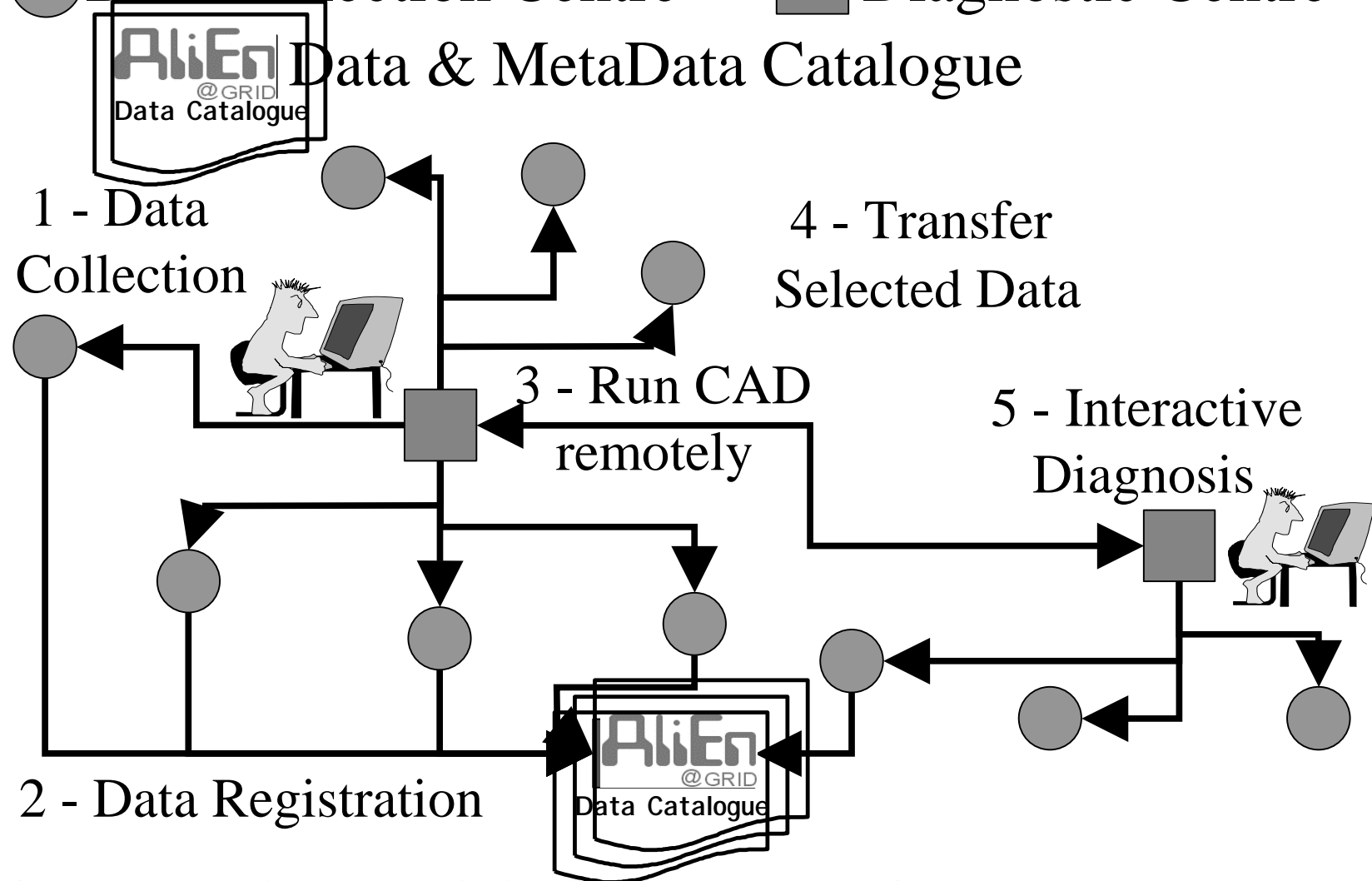


**LDN: LogicalDataSetName**  
**PDN: PhysicalDataSetName**

**CE: ComputeElement**  
**SE: StorageElement**  
**WN: WorkerNode**

# GPCALMA Screening Use Case

● Data Collection Centre      ■ Diagnostic Centre



**CAD selection to minimize data transfers**

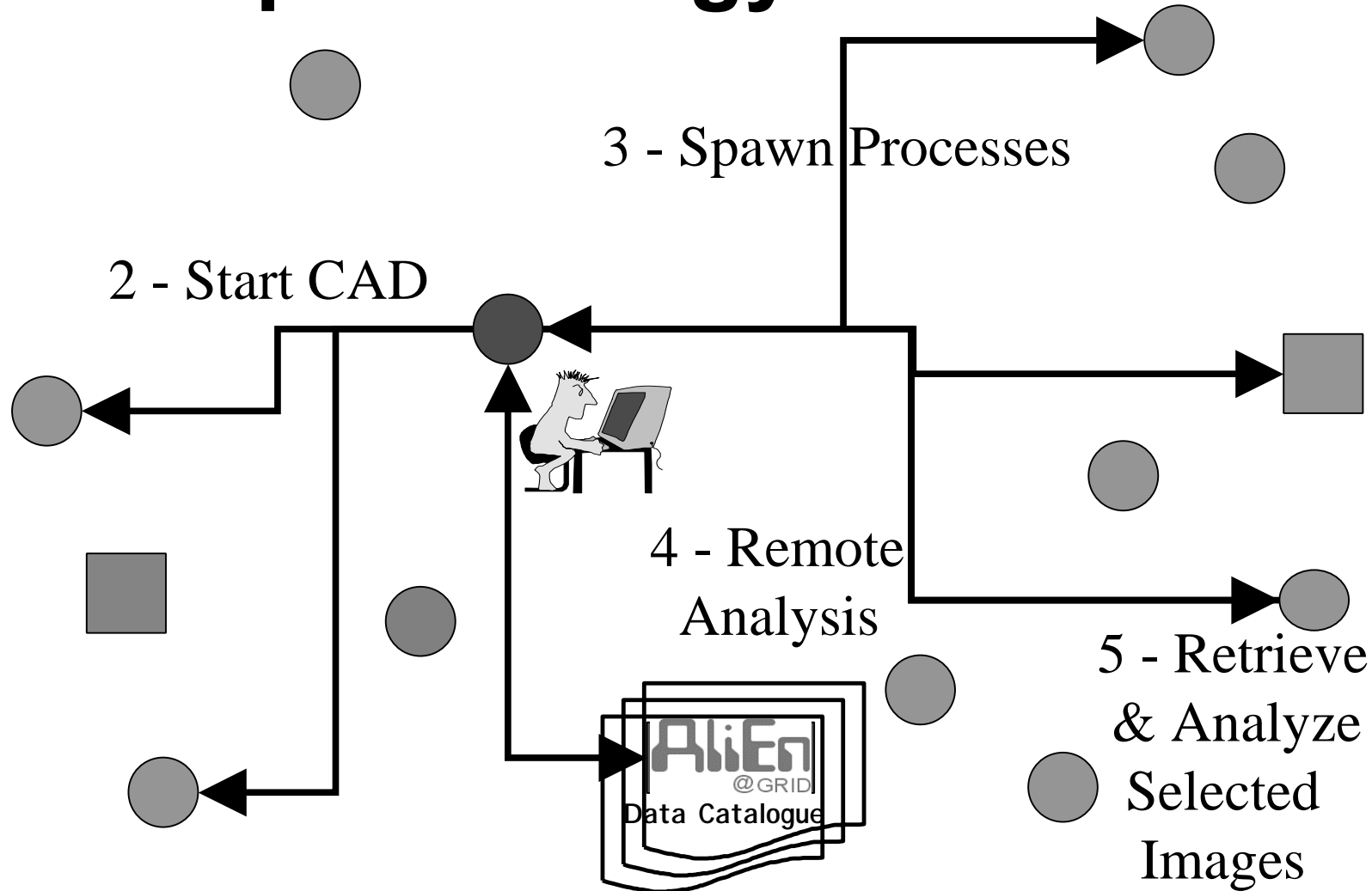
5 marzo 2004

Piergiorgio Cerello (cerello@to.infn.it)



65

# GPCALMA Tele-training & Epidemiology Use Case



# GPCALMA

## How to implement the above described Use Cases?

- **Move code rather than data**
- **Share the images without moving them**



*PROOF ( [http:// root.cern.ch](http://root.cern.ch) )*

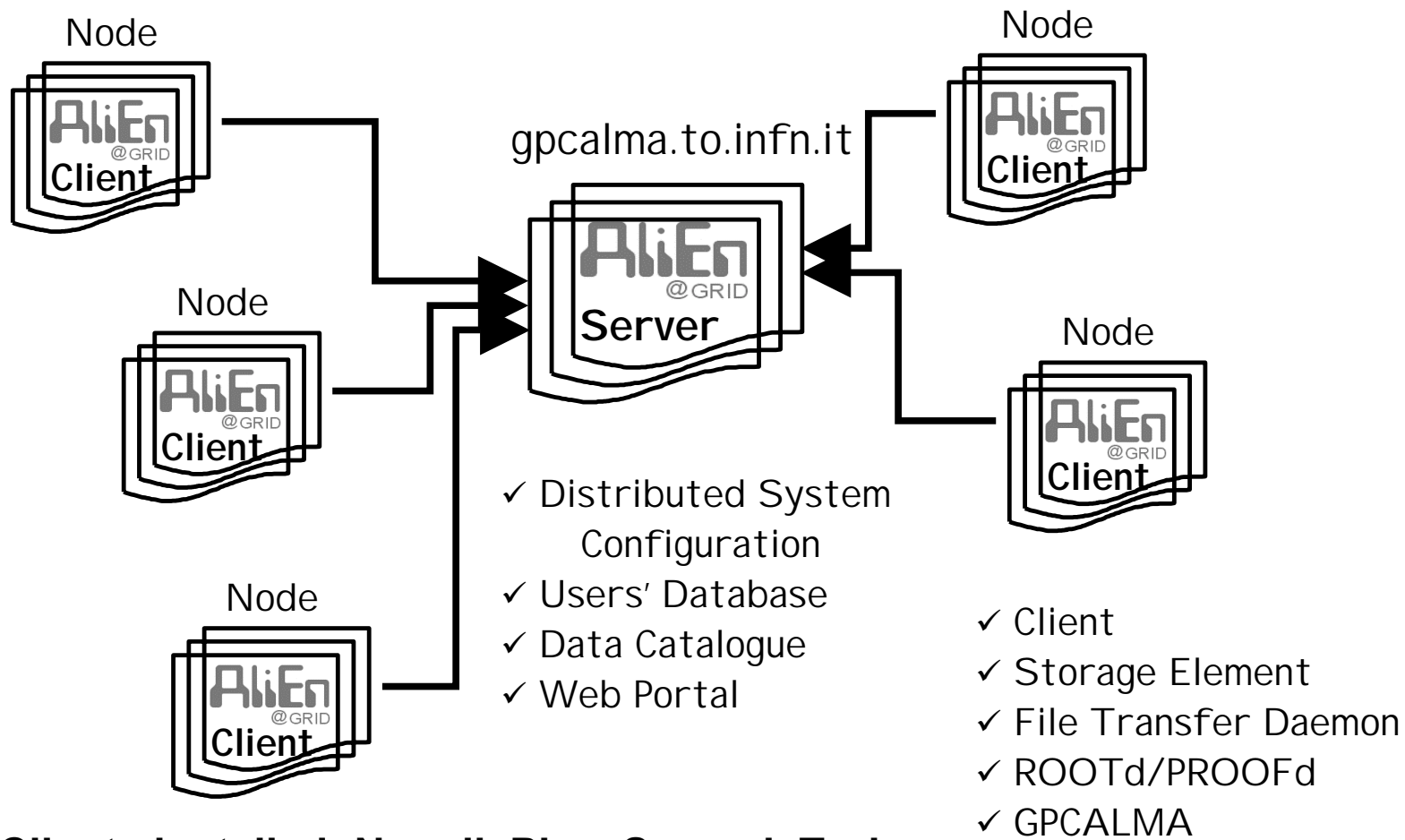
- **Secure Access**
- **Distributed Data Management**
- **Scheduling of Computing Resources**



*AliEn ( [http:// alien.cern.ch](http://alien.cern.ch) )*



# The GPCALMA distributed system configuration



✓ **Clients installed: Napoli, Pisa, Sassari, Torino**

# The AliEn-GPCALMA Web Portal ( <http://gpcalma.to.infn.it> )

The screenshot shows the AliEn-GPCALMA web portal in a Netscape browser window. The browser title is "Netscape: Alien Controls". The address bar shows the URL "http://toftxl.to.infn.it/". The page content includes a navigation menu on the left, a statistics table, and a task queue table.

**Navigation Menu:**

- ALIEN
  - Home
  - About
  - Project
  - Help
    - Installation
    - Site Installation
    - Certificates
    - API
    - Tutorial
    - Man pages
    - FAQ
  - Status
    - Sites
    - Active Clusters
    - Task Queue
    - Transfer Queue

**[Statistics]**

DONE	ERROR_E	KILLED
22	3	7

**[Tasks][Validate]**

ID	USER	EXEC	STATUS	COMMAND	START	TIME	MARK
32	bagnasco	toftxl.to.infn.it	DONE	/bin/HelloWorld-v2.sh	Mon Apr 28 16:09	00:00:02	┘
31	bagnasco	toftxl.to.infn.it	ERROR_E	/bin/HelloWorld.sh	Mon Apr 28 15:58	524:41:01	┘
30	bagnasco	toftxl.to.infn.it	ERROR_E	/bin/HelloWorld.sh	Mon Apr 28 15:58	524:41:04	┘
29	bagnasco	toftxl.to.infn.it	KILLED	/bin/HelloWorld.sh	Mon Apr 28 15:40	00:16:32	┘
28	bagnasco	toftxl.to.infn.it	KILLED	/bin/HelloWorld.sh	Mon Apr 28 14:39	01:18:06	┘

CERN © 2001 - European Organisation for Nuclear Research - ALICE Experiment

# Catalogue query

GPCALMA Available files

Images found for patient Antigone Cuorcontento

	SE	Hospital	Date	Filename	PFN
0	GPCALMA::Torino::policarpo	testHospital	2004-0-26	testImage1	file://policarpo.to.infn.it/home/bagnasco/GPCALMA/S
1	GPCALMA::Torino::tof0xl	testHospital	2004-0-26	testimage2	file://tof0xl.to.infn.it/home/gpcalma/SEData/00001/0
2	GPCALMA::Torino::tof0xl	testHospital	2004-0-26	testimage3	file://tof0xl.to.infn.it/home/gpcalma/SEData/00001/0

Submit Cancel

GPCALMA Query for available images

Patient first name Antigone

Patient last name Cuorcontento

Patient birth place Sacramento

Patient birth date 2001-01-01

Submit Cancel

debug  dry  verbose

## Patient creation

GPCALMA Create a new patient

Patient first name Gastroenterico

Patient last name Cuorcontento

Patient birth place Sacramento

Patient birth date 2001-01-01

Patient ID TESTPATIENT2677 Generate

Submit Cancel

debug  dry  verbose

GPCALMA Main


GPCALMA Data Catalogue

Create new patient

Register new exam

Find existing images

Exit



## Image registration

GPCALMA Register images

Site Torino

Hospital Valdese

Radiologist E.Zanor

Date 2004-0-26

Patient ID TESTPATIENT16200 Find

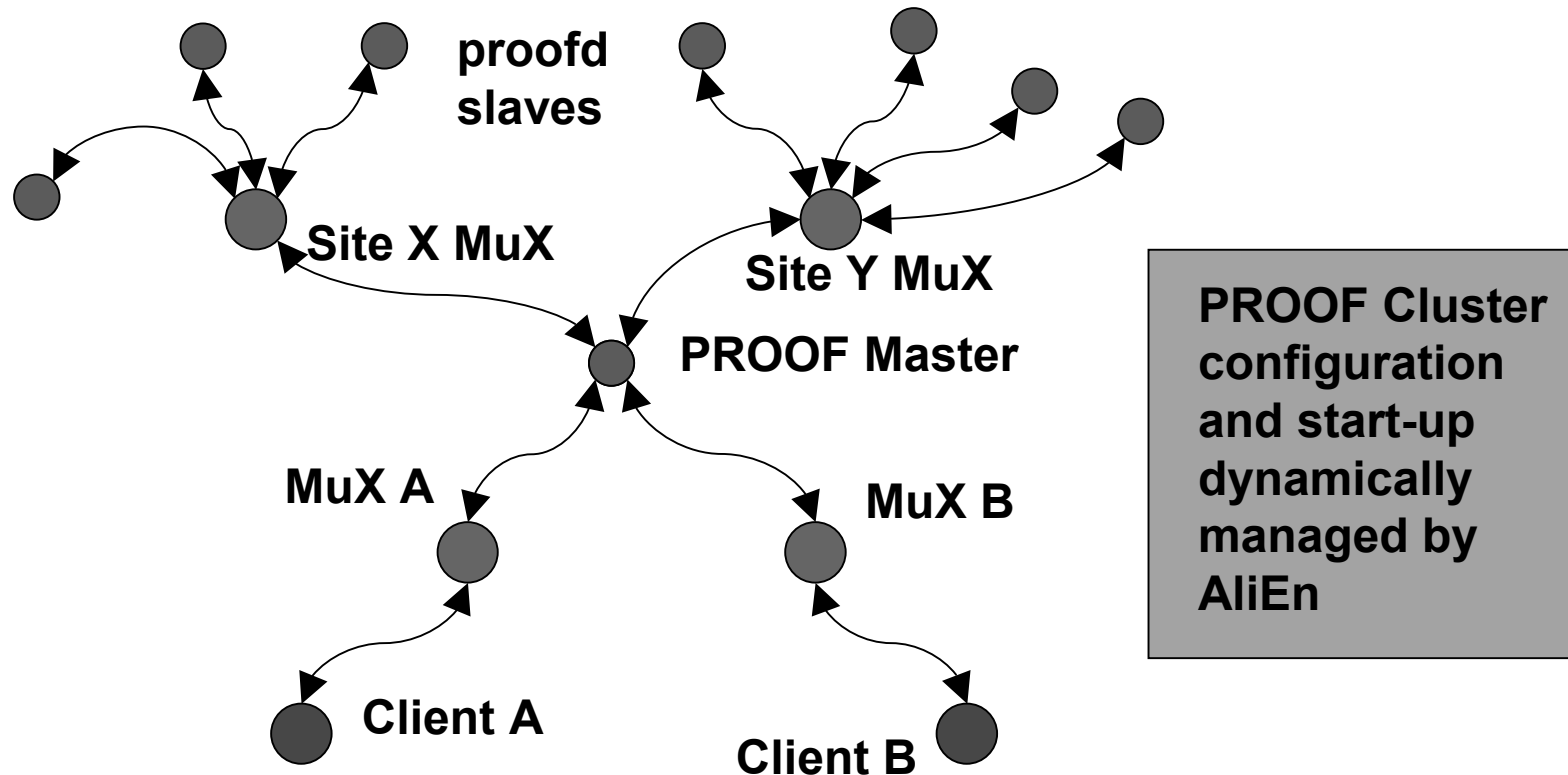
Submit Select files Clear list Cancel

debug  dry  verbose

/home/bagnasco/GPCALMA/gpcalma/ALIEN/testImage1  
/home/bagnasco/GPCALMA/gpcalma/ALIEN/testImage2

# AliEn-PROOF Interface

- Access to the (Meta)Data Catalogue from the ROOT shell!!!!
- Under development, prototype available and installed in Torino



# Achievements

- ✓ CALMA rewritten in C++, based on ROOT
- ✓ New GUI, with functionality to manipulate the images
- ✓ AliEn server and clients operational
- ✓ PROOF cluster configured
- ✓ 1<sup>st</sup> mammogram remotely analysed in March
- ✓ data/metadata structure being (re)defined
- ✓ re-organisation of the CALMA Database

2  
0  
0  
2

2  
0  
0  
3

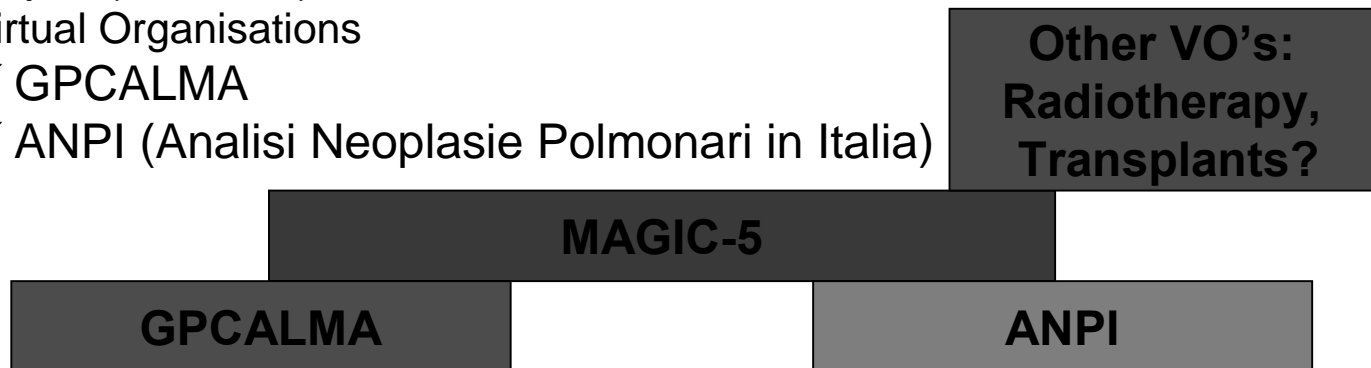
**The basic functionality is available and tested**

## Ongoing tasks:

- CALMA-DICOM format conversion
- C++ ROOT-AliEn API for Input Data Selection
- Upgrade GUI for Access to the AliEn Catalogue
- optimise the implementation of data and metadata
- set up a prototype in the participating hospitals

# Why MAGIC-5?

- INFN expertise and leadership in:
  - ◆ CAD development
  - ◆ Grid Middleware
- Does any other Medical field but mammography require a similar approach?
  - ◆ Contact by Medical Doctors who knew about (GP)CALMA developments (Milano)
- CAD for Lung Cancer detection... it's on time – like CALMA!
  - ◆ 3D CT images
  - ◆ search for different patterns
  - ◆ same Grid approach
    - ✓ AliEn is the best available Grid implementation in terms of functionality, stability and scalability
- New separate project (ANPI) or integration with GPCALMA?
  - ◆ 1 project (MAGIC-5)
  - ◆ 2 Virtual Organisations
    - ✓ GPCALMA
    - ✓ ANPI (Analisi Neoplasie Polmonari in Italia)





# CAD for Lung Cancer?

- 5 years survival rate for lung cancer: 14% (US), 10-15% (EU)
- no improvement in the past 20 years

- ◆ Low dose CT: 6 times more efficient than Chest X-Ray (CXR) in the detection of state I malignant nodules

- ◆ CAD methods are being explored

- ◆ Gurcan et al., Med. Phys. 29(11), Nov. 2002, 2552:

“...computerized detection for lung nodules in helical CT images is promising...large variations in performance, indicating that the computer vision techniques in this area have not been fully developed. Continued effort will be required to bring the performances of these computerized detection systems to a level acceptable for clinical implementation.”

**About 43 images/patient**  
**About 0.5 MB/image**

Number of cases	Sensitivity (%)	FP/image	Authors
17	95.7	0.3	Fiebich
17	72	4.6	Armato
26	30	6.3	Fiebich
43	71	1.5	Armato
16	86	2.3	Ko
34	84	1.74	Gurcan

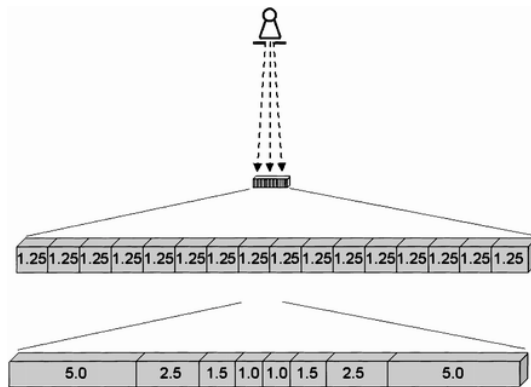


# Spiral CT imaging principles

- Linear patient motion through the gantry
- Beam rotation
  - ◆ spiral pattern of data acquisition
  - ◆ one continuous set of volume data
  - ◆ Reconstruction options
    - ✓ (Slice reconstruction increment)
    - ✓ (Interpolation algorithm)
    - ✓ (Effective slice thickness)

**Best available trade-off  
between sensitivity for the  
detection of nodules and  
absorbed dose**

- Single(Multi)-slice: 1(1) tube + 1(N) detector array(s) with 500-900 elements + 1(4) DAQ channel: 1(2)D curved array, shorter scan time



- $N \geq 4$  detector arrays
  - ◆ (A)symmetric detector arrays
  - ◆ Detector elements or arrays can be combined to obtain different thickness and/or width
  - ◆ Collimators can also be used

# Multi-slice vs. Single-slice

- Volume Coverage:

$$\frac{N \times P \times S \times T}{R}$$

N= number of DAQ channels = 4

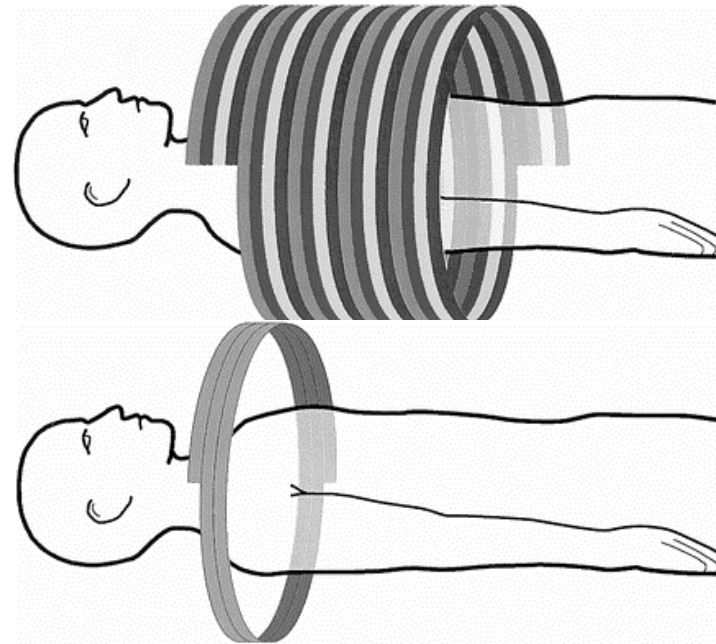
P= pitch

(linear movement in T/beam collimation)

S= detector width (mm)

T= execution time (s)

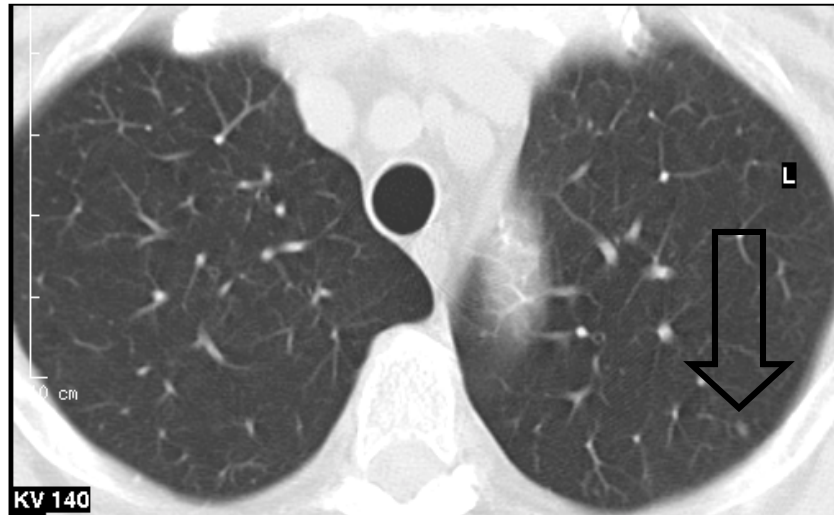
R= rotation time (s) = 0.5 s



	mAs	kV	Collimation (mm)	Pitch	T (s)	Step (mm)
SSCT	43	140	3-5	2:1	1	1
MSCT	20	120	1(x4)	7:1	0.5	2-5

# Images: an example

5 mm  
140 KV  
120 mAs



1 mm 120 KV 20 mAs

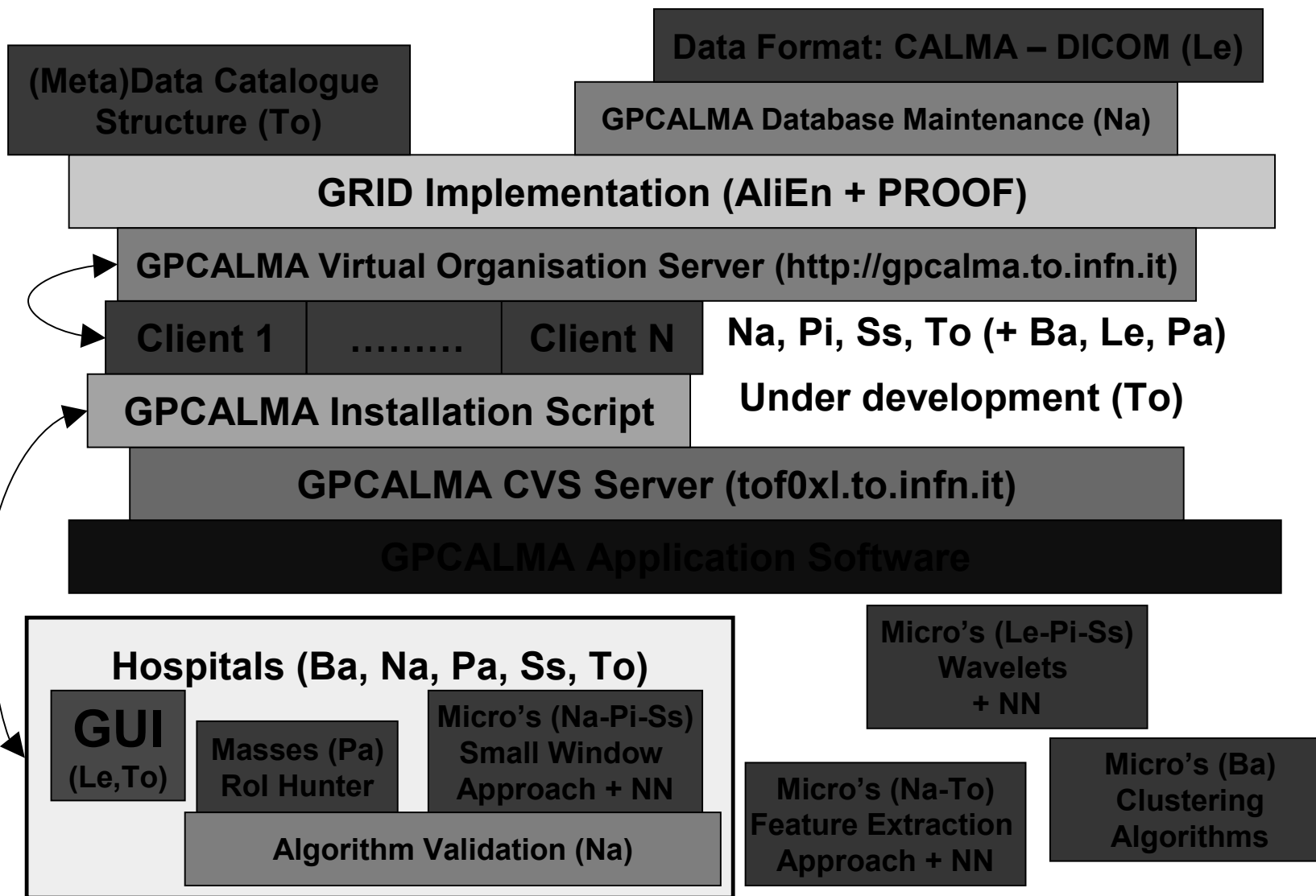


Ric. 5 mm 120 KV 20 mAs

# Screening in Italy & EU-US

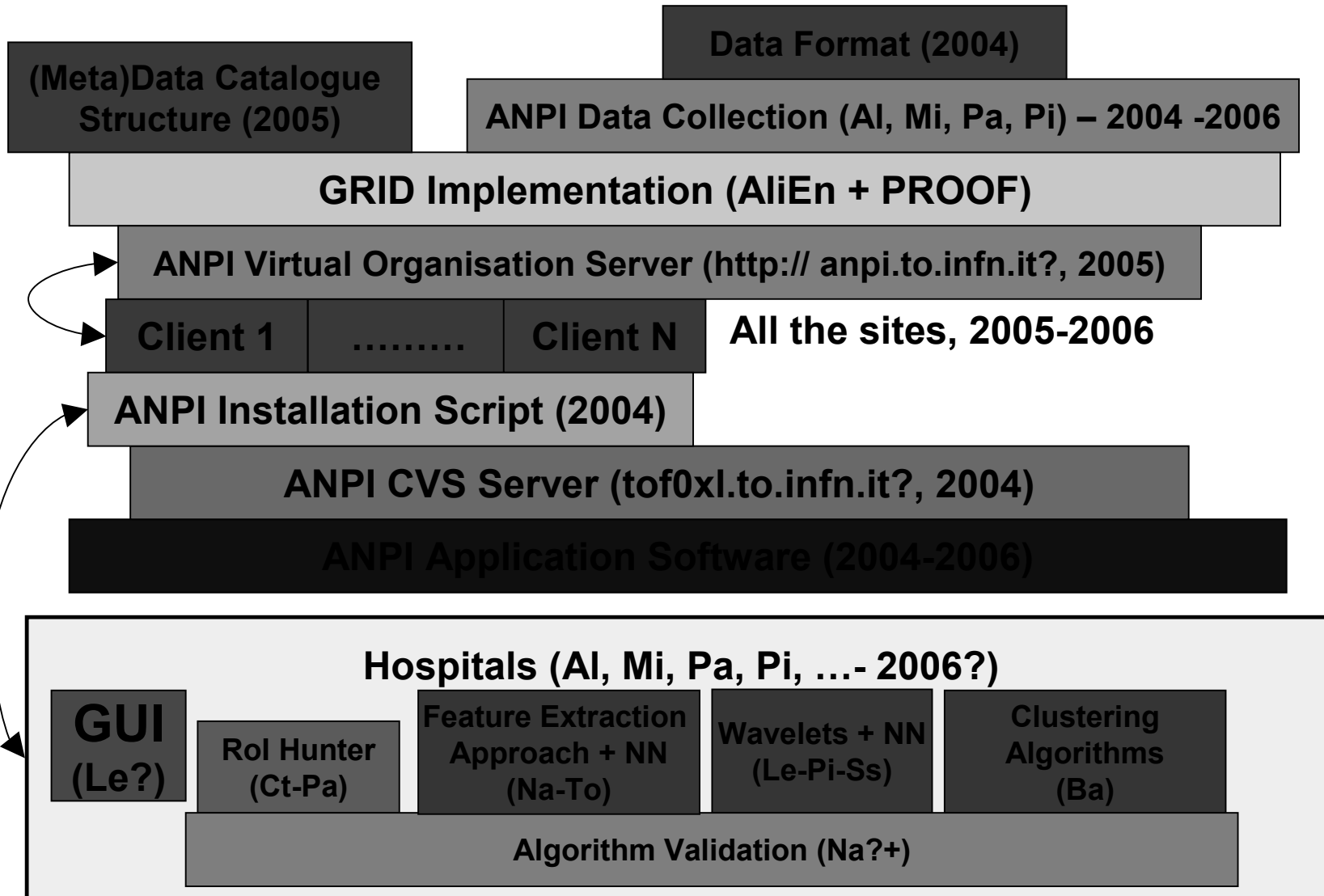
- **Main goal**
  - ◆ reduce the death rate caused by lung cancer
  - ◆ The sample
    - ◆ 55-69y
    - ◆ >20 (packs/day) \* y
    - ◆ Smokers (or ex-smokers < 10 y)
    - ◆ Agreement
    - ◆ No previous cancer
- **Italy**
  - ◆ Ongoing programs: Genova, Milano, Torino
  - ◆ Starting phase: Regione Toscana – Emilia-Romagna
    - ✓ About 7000 exams in 4 years
- **EU – US**
  - ◆ Collaborative Spiral CT-group
  - ◆ I-ELCAP: International Early Lung Cancer Action Project
  - ◆ EU ELCDG: EU Early Lung Cancer Detection Group
  - ◆ US: National Lung Screening Trial (50,000 people)

# MAGIC-5/GPCALMA





# MAGIC-5/ANPI



# Summary and Conclusions

- Breast and Lung Cancer Detection in Screening Programs: good examples of e-health application that would benefit from the use of GRID Services
- The AliEn/PROOF based approach allows
  - ◆ **Minimisation of data transfers**
  - ◆ **Secure management of a distributed Virtual Organisation**
- The success will depend on:
  - ◆ **the reliability and stability of AliEn GRID Services**
  - ◆ **the performance of CAD algorithms**
  - ◆ **the quality of the GUI**
- GPCALMA Virtual Organisation in the participating Hospitals
- ANPI
  - ◆ **prototype CAD algorithms for Lung Cancer detection**
  - ◆ **performance evaluation**