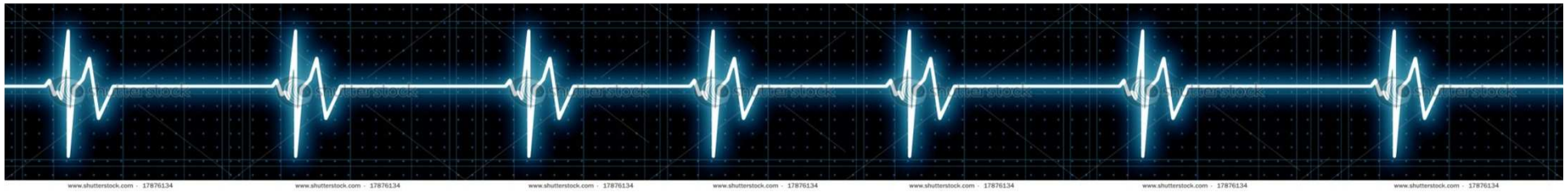
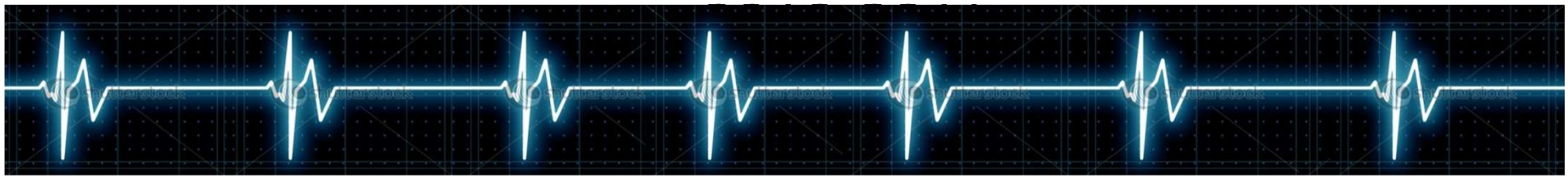


# Metodi di classificazione automatica in medicina



# Problema

Dati:

- un evento x
- un numero M di classi $w_1$, $w_2$….$w_M$

assegnare l'evento ad una di queste classi

Esempi:

Dal segnale ECG misurato in un soggetto, definire se il soggetto è normale, oppure ha ipertrofia ventricolare oppure è in corso un infarto

Da misure attraverso spettrometria di massa del proteoma di un soggetto, definire se il soggetto è normale o ha una certa forma di tumore

Da una immagine al microscopio riconoscere e contare cellule con caratteristiche diverse per quantificare emocromo e formula leucocitaria

# Definizione delle classi

E’ uno degli  aspetti principali di questo problema.
In alcuni (rari) casi le classi sono definite in modo preciso e la classificazione non è ambigua
Es. classificare le figure piane in triangoli (3 lati/3 angoli) quadrilateri (4lati/4 angoli), ecc.

Più in generale non esiste una definizione “esatta” di classi (vd esempi precedenti). Considereremo nel seguito la situazione in cui le classi sono caratterizzate attraverso un  processo di apprendimento, basato sull’osservazione di un numero di esempi di cui è nota con precisione l’appartenenza ad una specifica classe. Tale insieme di esempi viene detto training set e con riferimento a queste situazione si parla di classificazione supervised.

Ci sono poi delle situazioni in cui  si dispone di un numero di esempi di cui però non è nota la classe di appartenenza. Un primo obiettivo (cluster analysis) è allora di riconoscere la presenza in questo insieme di gruppi di dati con caratteristiche simili, ognuno dei quali viene ricondotto ad una classe. Si parla allora di classificazione unsupervised.

# Classificazione supervised

Esistono due  approcci diversi al problema:

- geometrico (o matematico)

  esistono vari metodi: statistici,  reti neurali, alberi decisionali, support vector machines ecc.

- sintattico (o linguistico)

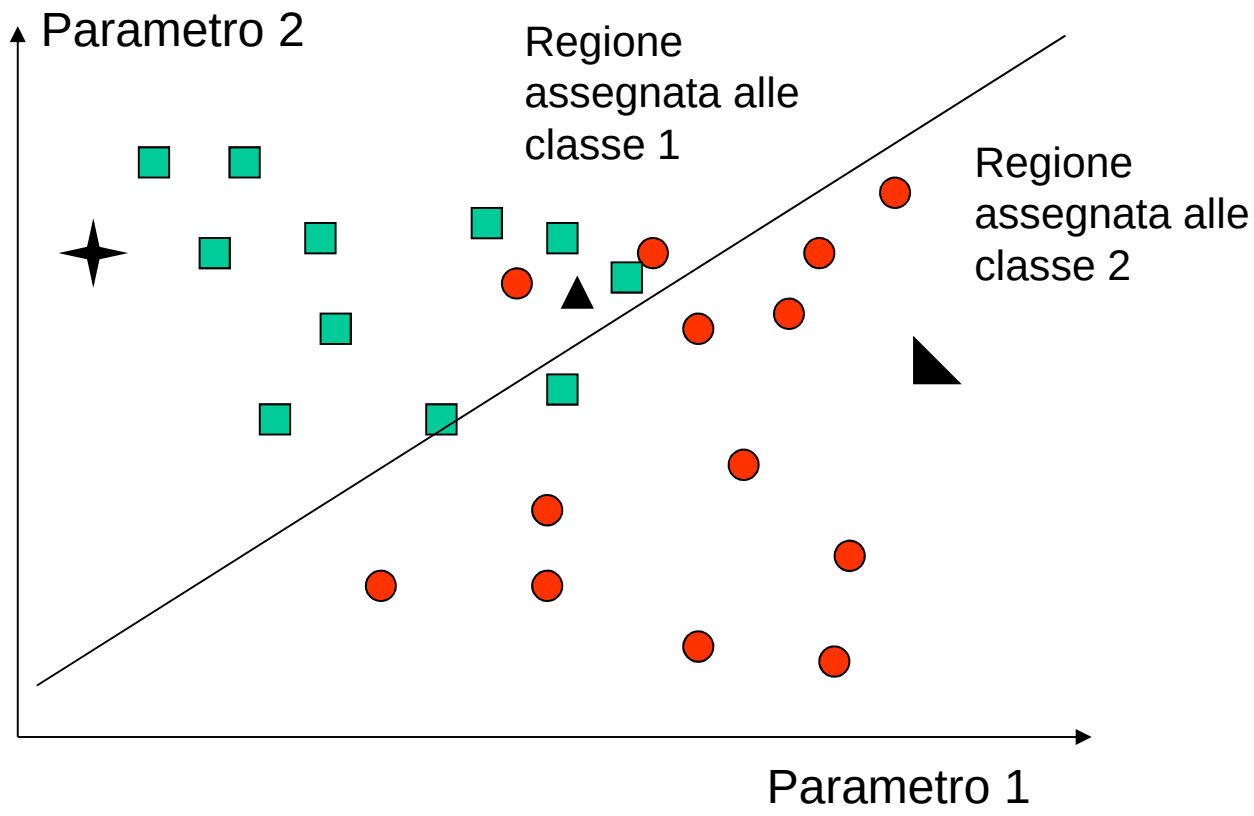
  Si basa sull'assunzione che eventi complessi possono essere descritti come composizione ricorsiva di eventi più semplici, così come in un linguaggio un periodo è composto da più frasi, a loro volta composte da più parole. Le regole di composizione (grammatica), auspicabilmente diverse in ogni classe,  vengono apprese attraverso l'osservazione del training set.

  L'elemento da classificare  viene scomposto e la sua grammatica  esplorata. L'elemento viene assegnato alla classe che ha una grammatica più simile

  E' un approccio attraente ma difficile da tradurre in pratica, infatti non sono molte le applicazioni in cui l'approccio è stato utilizzato con success

# Approccio geometrico

Ogni elemento da classificare è rappresentato con un vettore di parametri (features) Questi valori vengono definiscono un punto nello spazio dei parametri. Se il vettore dei parametri è scelto in modo adeguato, gli elementi del training set permettono di ripartire lo spazio dei parametri, attraverso opportuni criteri (vedremo tecniche bayesiane) in regioni assegnati alle varie classi.

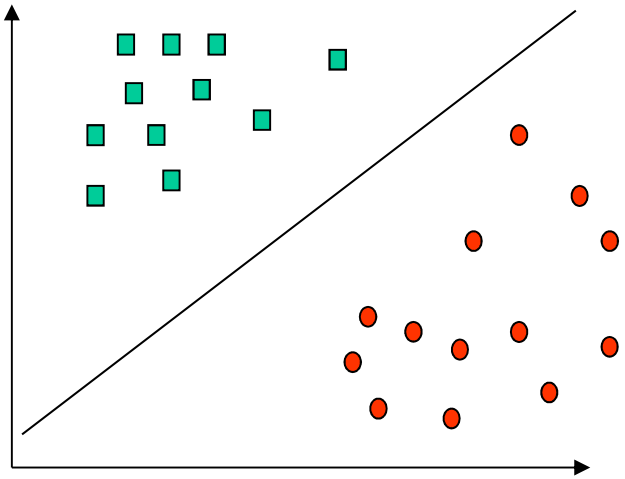



Training set:
- Classe 1
- Classe 2

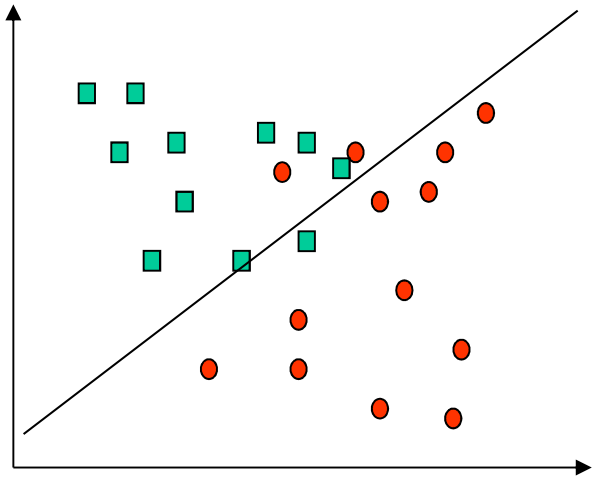
Elementi da classificare
- Classe 1
- Classe 1
- Classe 2

# Limiti del classificatore

Spesso le classi non sono separate in modo netto. Quindi non si può escludere che il classificatore commetta degli errori



Situazione ideale



Situazione realistica

E' importante quantificare o almeno stimare l'errore di classificazione  (validazione)

## Progetto del classificatore

Dal training set, viene individuata la ripartizione dello spazio di parametri (ad es. nella figura precedente si è usato un classificatore lineare) sulla base di opportuni criteri (es. di tipo statistico, o basati su misure di distanza.

## Validazione del classificatore

Si testa la capacità del classificatore di dare una risposta corretta, quantificando le sue prestazioni su un insieme di dati di cui è nota la classe di appartenenza (testing set)

## Uso del classificatore

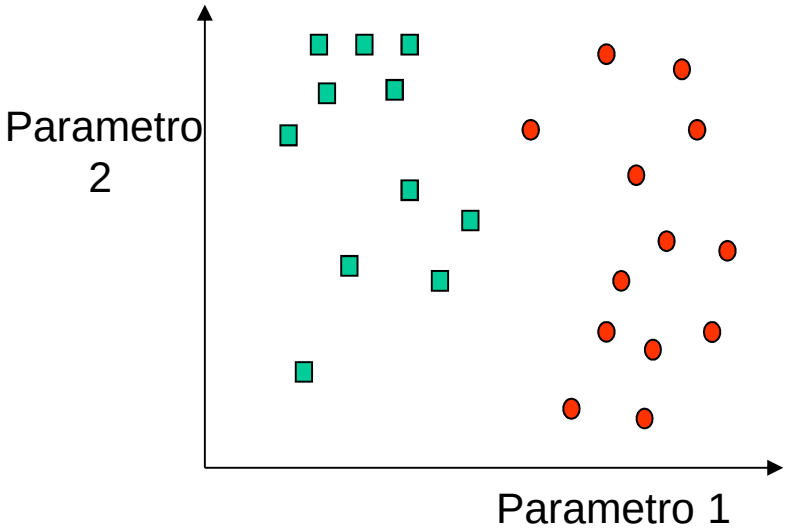
Si applica il classificatore all'evento x da classificare

# Definizione dei parametri

Il vettore dei parametri deve essere scelto in modo da ottimizzare la separazione tra le classi. Quindi la scelta dei parametri è critica, e non ci sono indicazioni generali su come scegliere i parametri. Può essere allora vantaggioso usare un numero elevato di features.

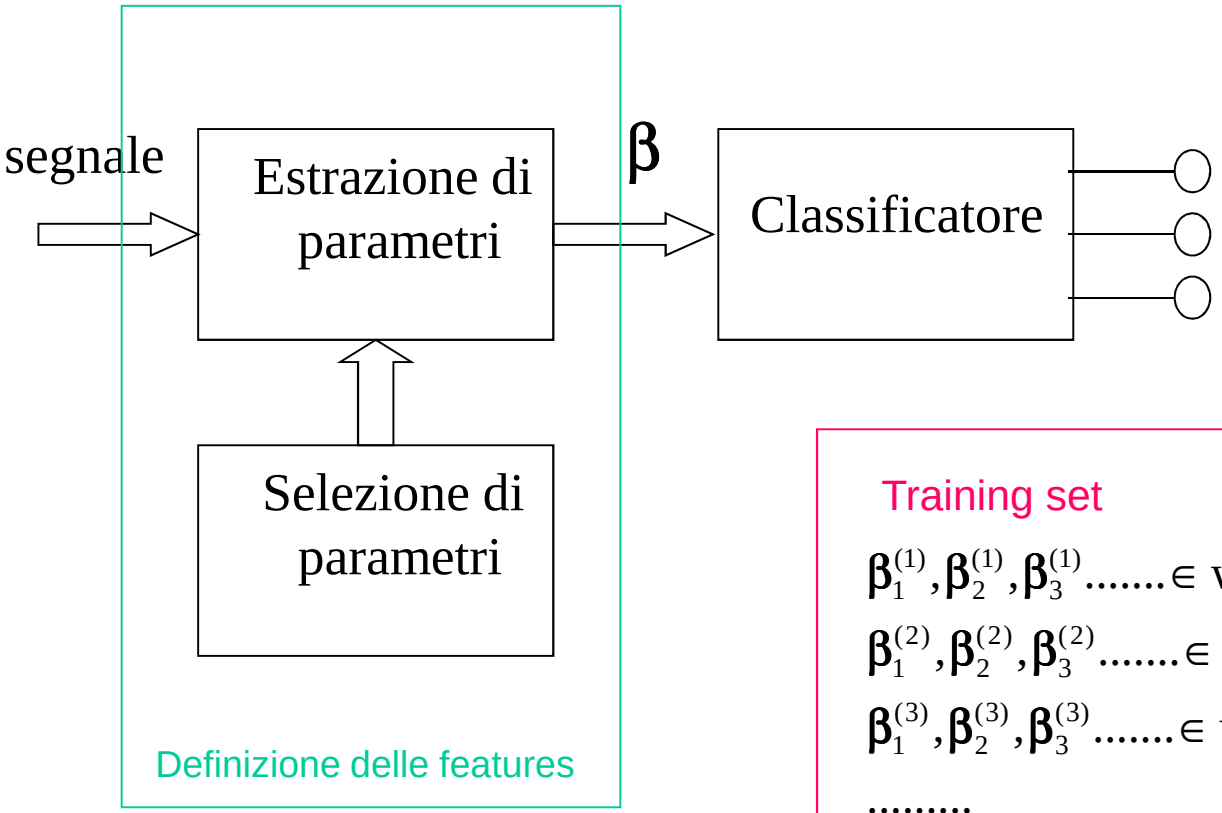
D'altra parte, più sono numerose le features, più complessa è la fase di progetto del classificatore e più numeroso dovrebbe essere il training set. Inoltre la prestazione del classificatore si degrada se le features sono ridondanti

Allora si parte definendo un gran numero di features, e poi si lascia ad una fase successiva detta di selezione (esistono varie tecniche allo scopo) il compito di isolare le poche features, o combinazioni di features, in grado di discriminare tra le varie classi




Il parametro 2 non discrimina, si può classificare altrettanto bene usando solo il parametro 1

# Schema del progetto di un classificatore




Training set

$\beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)} \ldots\ldots \in w_1$

$\beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)} \ldots\ldots \in w_2$

$\beta_1^{(3)}, \beta_2^{(3)}, \beta_3^{(3)} \ldots\ldots \in w_3$

.....

$\beta_1^{(M)}, \beta_2^{(M)}, \beta_3^{(M)} \ldots\ldots \in w_M$

Nel seguito : 1. estrazione di parametri (slides)

2. progetto di un classificatore bayesiano

3. selezione → fotocopie

4. validazione

# 1. Estrazione di parametri

Supponiamo che l'evento da classificare sia un segnale x(n), n=0,1….Ntot-1 i cui campioni sono contenuti nel vettore **y** di dimensione Ntot

L'estrazione del vettore $\beta$ che contiene le N features può essere visto come la funzione

$$\beta = F(\mathbf{y})$$

con F funzione in genere non lineare.

Si distinguono due tipi di parametri

Specifici : ampiezza, durata, pendenza,,,,

vantaggi: conservano il significato fisiopatologico

svantaggi:

- Relazione non lineare con il segnale
- Difficile misurarli con uguale precisione
- Tra loro dipendenti (ridondanti)
- Difficile quantificare la loro informazione utile

Quindi arbitraria la scelta del tipo e del numero

Non Specifici : da trasformazione lineari dei dati (es. PCA, Fourier, ,,,,
vantaggi & svantaggi : complementari ai precedenti

Spesso ai parametri estratti da un  segnale vengono aggiunti parametri relativi al soggetto, ad es, età, fattori di rischio, esiti di altri esami, trattamenti ricevuti. Alcune di queste variabili sono continue (es, età) altre sono discrete (il soggetto ha o non ha ricevuto un trattamento).

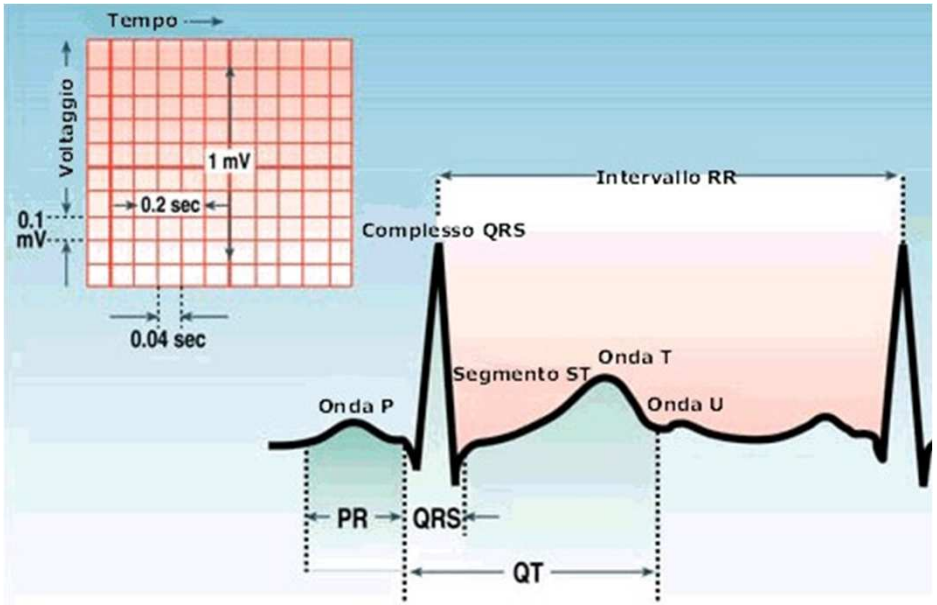
Inoltre le varie features possono assumere valori diversi in termini di ordine di grandezza, anche in relazione alle unità di misura adottate.

Soluzione: **Normalizzazione dei parametri**

Tipiche funzioni di normalizzazione:

- Divisione  per il massimo valore assunto nel training set
- Divisione per il valore medio assunto nel training set

# Esempio1: ECG




Parametri specifici:

Ampiezza, durata, area sottesa per le onde P, QRS, T, U, durata e pendenza dei segmenti PR e ST ecc.

Si può arrivare a definire fino a 300 parametri morfologici

Alcuni problemi: che periodo scegliere?

mediare alcuni periodi e poi estrarre i parametri?

estrarre i parametri da più complessi e poi mediare?

Esistono dei margini di soggettività nel modo in cui i parametri sono definiti
Importanza di standardizzare il modo in cui i parametri vengono estratti!!!

# Esempio 2: EMG

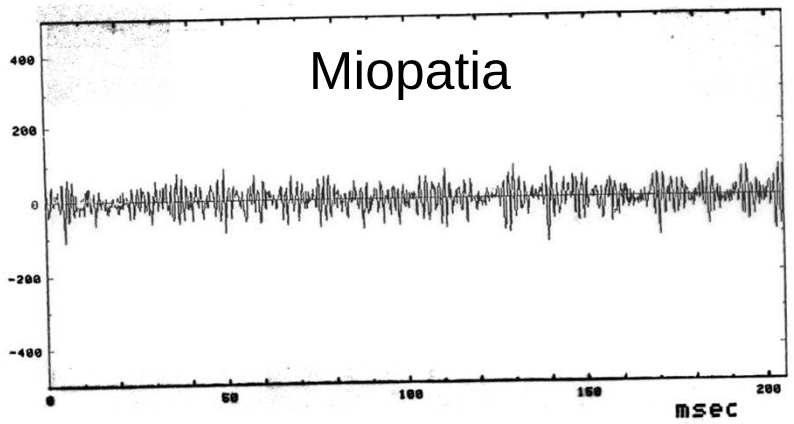



Fig. 10.4 Segnale EMG miopatico

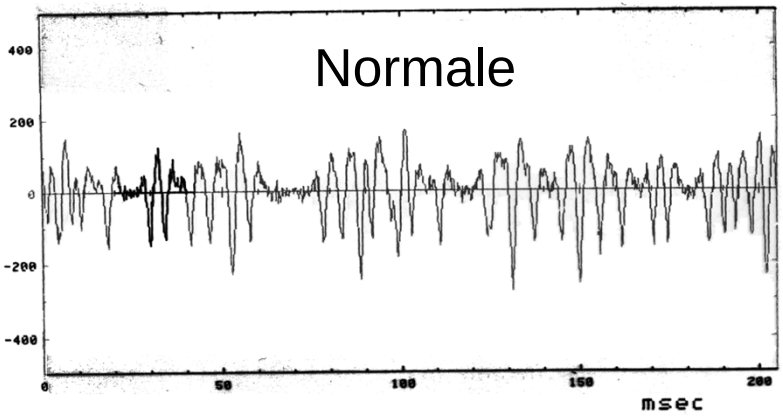



Fig. 10.5 Segnale EMG normale

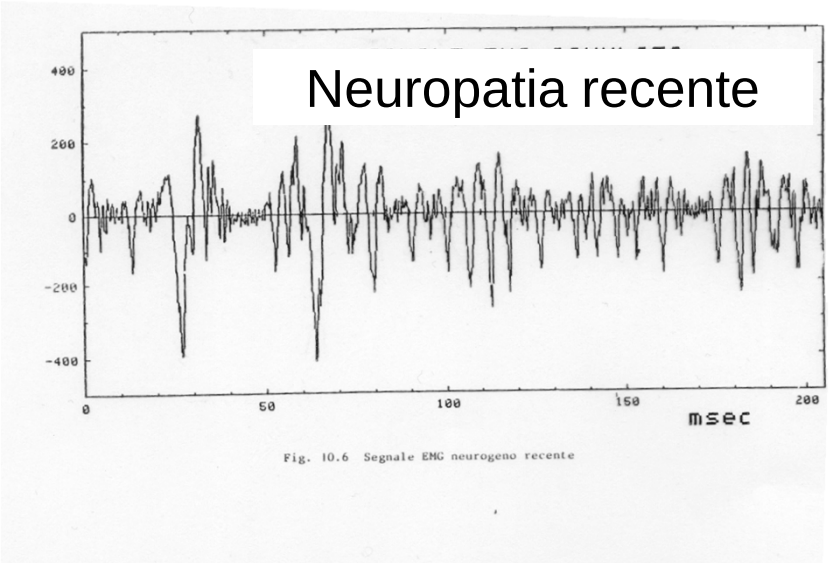



Fig. 10.6 Segnale EMG neurogeno recente

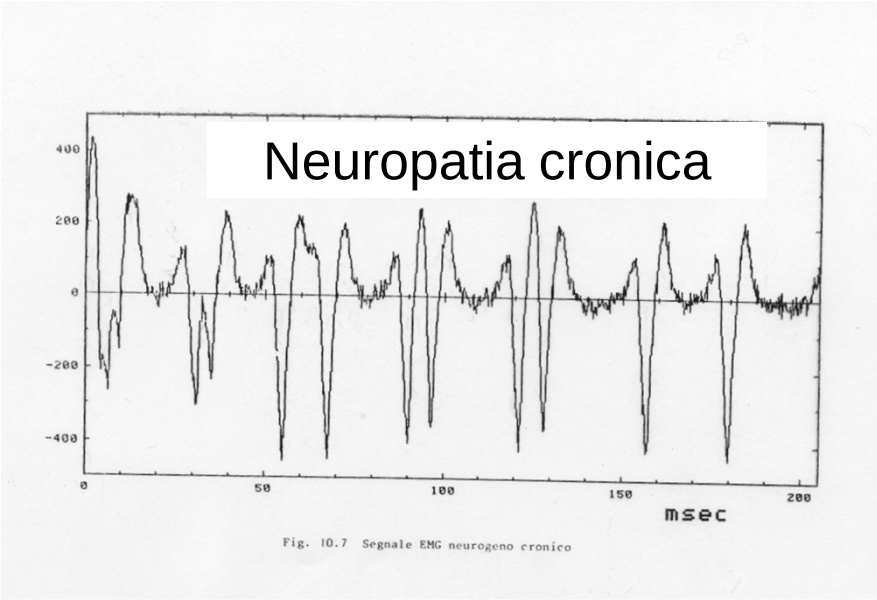



Fig. 10.7 Segnale EMG neurogeno cronico

# EMG: parametri (specifici) nel dominio del tempo

| | | |
|---|---|---|
| Ampiezza media | MIOPATIA ↓ | NEUROPATIA ↑ |
| Numero di massimi positivi sopra una certa soglia | MIOPATIA ↑ | NEUROPATIA ↓ |
| Numero di attraversamenti della linea di base | MIOPATIA ↑ | NEUROPATIA ↓ |
| Rapporto tra numero di max positivi e ampiezza media | MIOPATIA ↑ | NEUROPATIA ↓ |

# Esempio 2: spettro dell'EMG

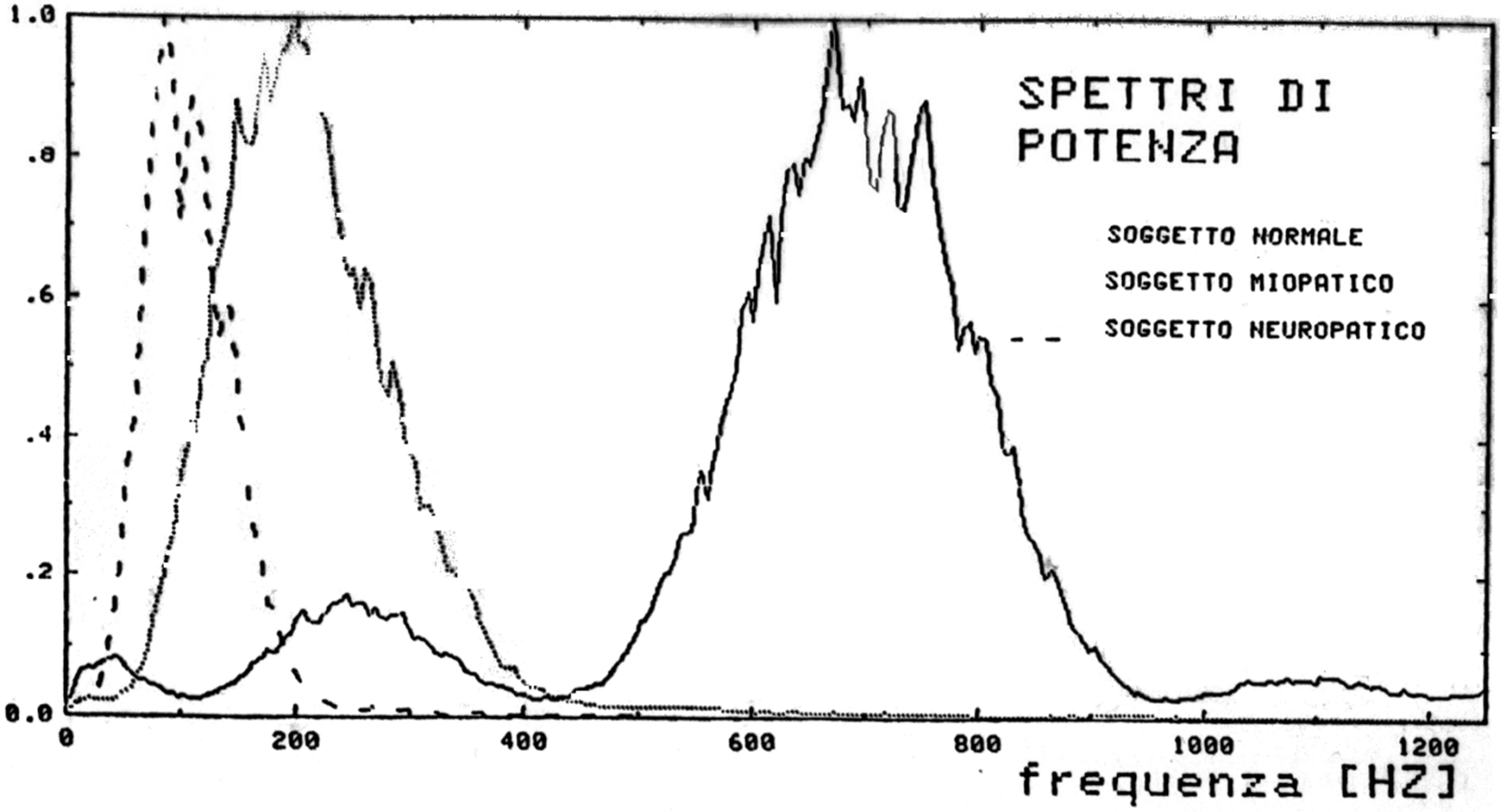



Fig. 10.15

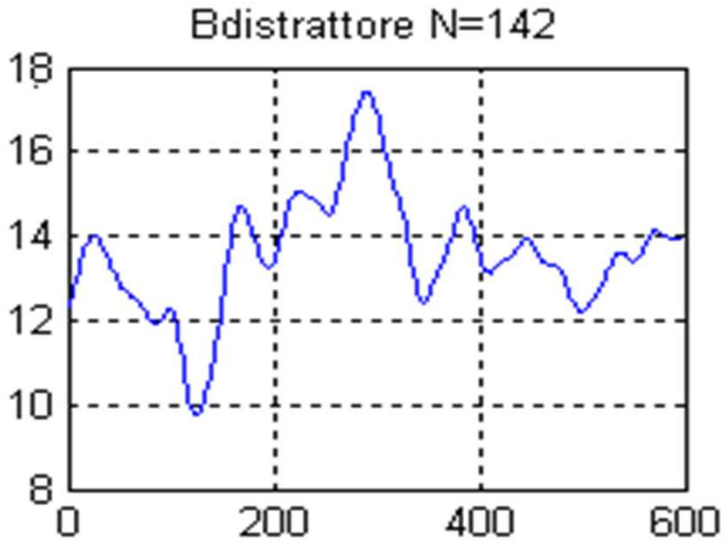
# EMG: parametri (specifici) nel dominio della frequenza

| | | |
|---|---|---|
| Baricentro di frequenza | MIOPATIA ⇩ | NEUROPATIA ⇧ |
| Frequenza mediana | MIOPATIA ⇩ | NEUROPATIA ⇧ |
| Frequenza di massimo | MIOPATIA ⇩ | NEUROPATIA ⇕ |

# EMG: parametri di un modello AR

Si usa lo stesso ordine (medio) per tutti i dati. I parametri forniscono una descrizione sintetica del modello

# Esempio3: EP



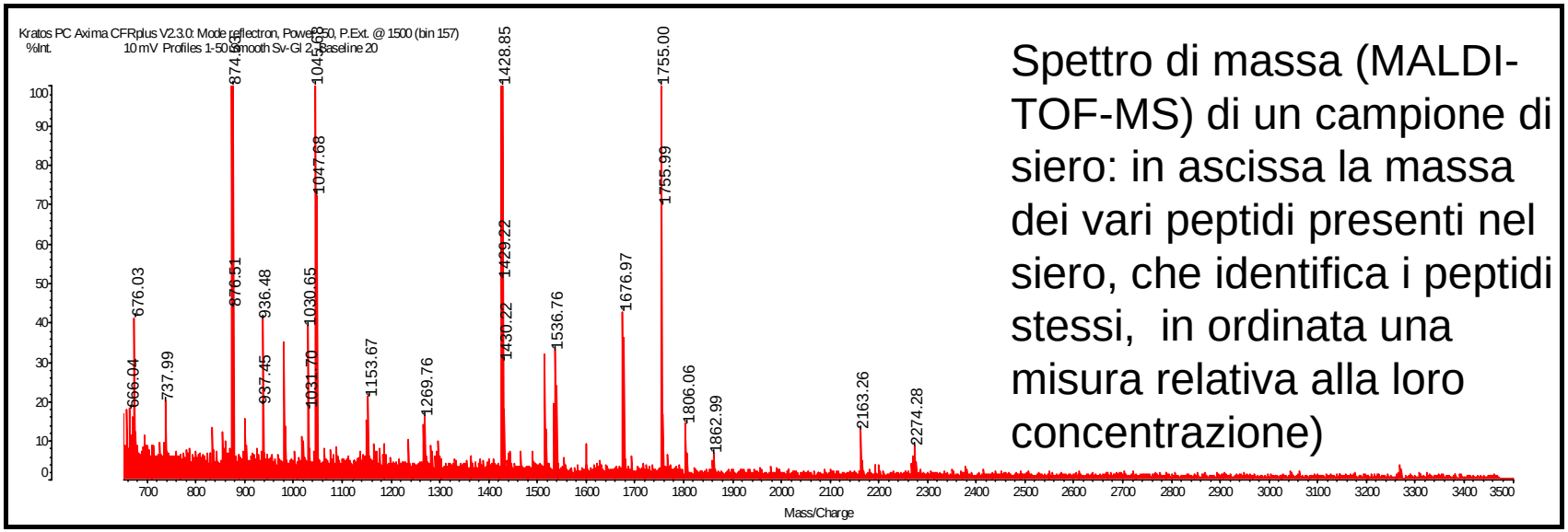

Parametri specifici:

Ampiezza e latenza dei vari picchi

In alternativa, dato che il segnale è a durata limitata e il numero di campioni è relativamente basso (circa 70 se frequenza = 128hz e durata 600msec) si può pensare di usare i campioni stessi come features, eventualmente sottocampionando

# Esempio 4: Classificazione in proteomica

- Asse x : m/z (Dalton = 1/12 massa atomo $^{12}C$)
- Asse y : concentrazione (assoluta o %)
- Range:  700:20,000 Da




Spettro di massa (MALDI-TOF-MS) di un campione di siero: in ascissa la massa dei vari peptidi presenti nel siero, che identifica i peptidi stessi,  in ordinata una misura relativa alla loro concentrazione)

- Fine: determinare peptidi e/o proteine (biomarker) in grado di diagnosticare particolari condizioni patologiche (es. tumori)
- Dati: 20,000-370,000 picchi per spettro
- Generalmente 2 classi : normale/patologia
- Si usano test statistici per preselezionare i picchi significativamente espressi nelle due classi

# The dataset

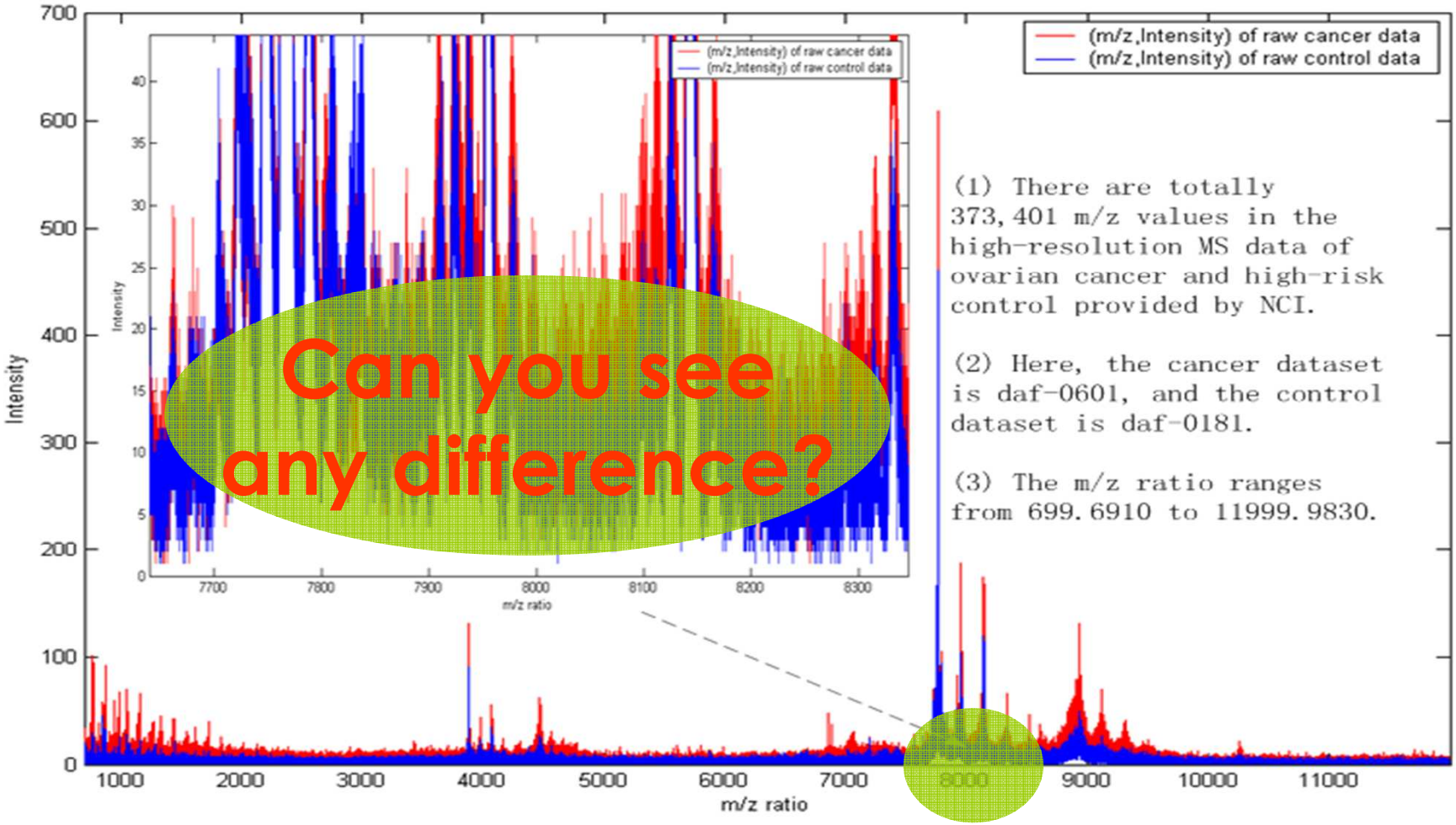
Dept. of Urology, University Hospital, Innsbruck
Inst. for Analytical Chemistry and Radio-chemistry, Innsbruck University.

Prostate Cancer
MELDI-TOF-MS

166 controls
133 cancers
96,040 pps
3 replicates for
each subjects

- Very high data dimensionality
- Noisy and redundant data
- Limited number of samples per diagnostic category

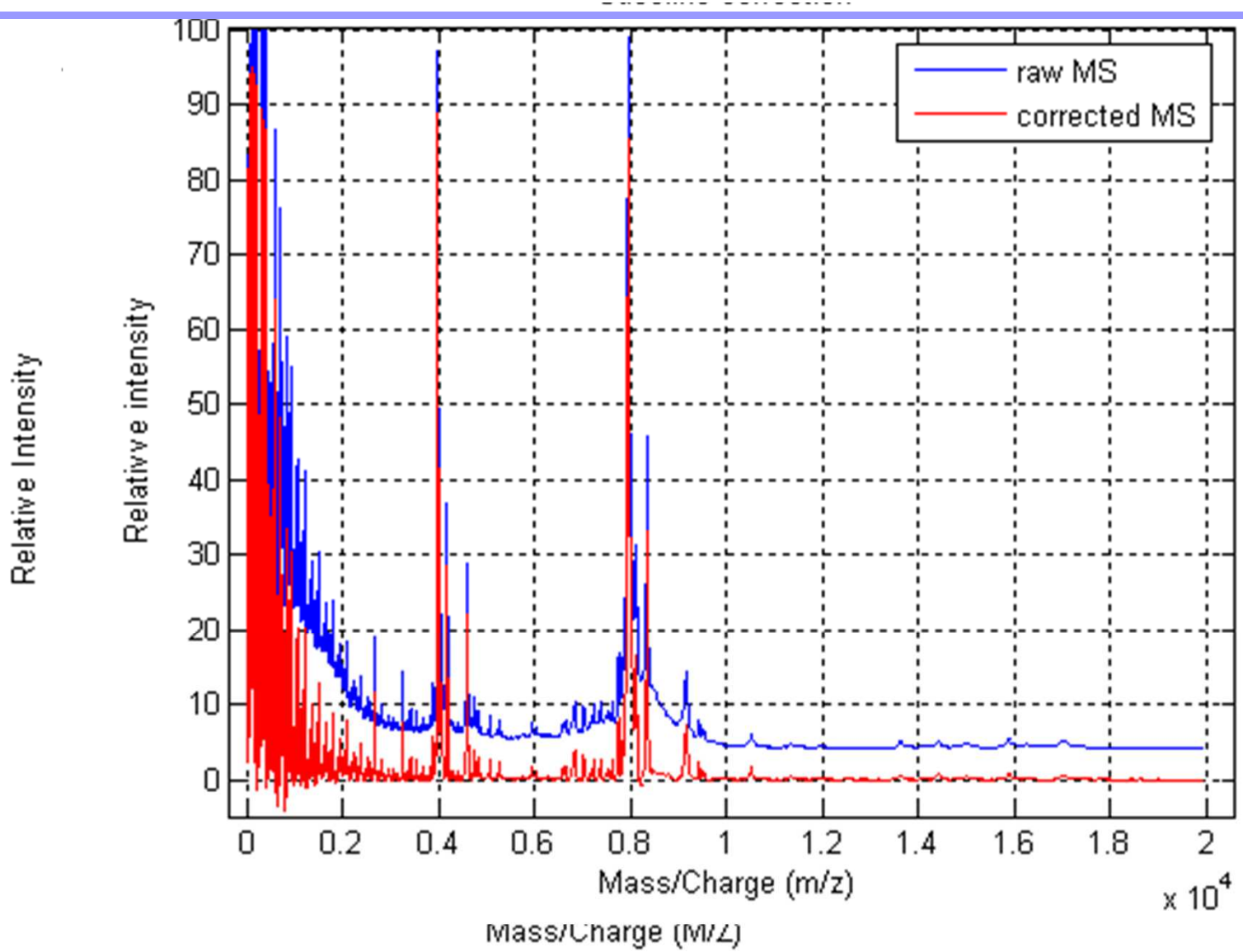
# Example of raw MS spectra



# Feature extraction

- Baseline correction
- Alignment
- Normalization
- Binning
- Feature selection via statistical tests

# Baseline correction



# Alignment to internal standard

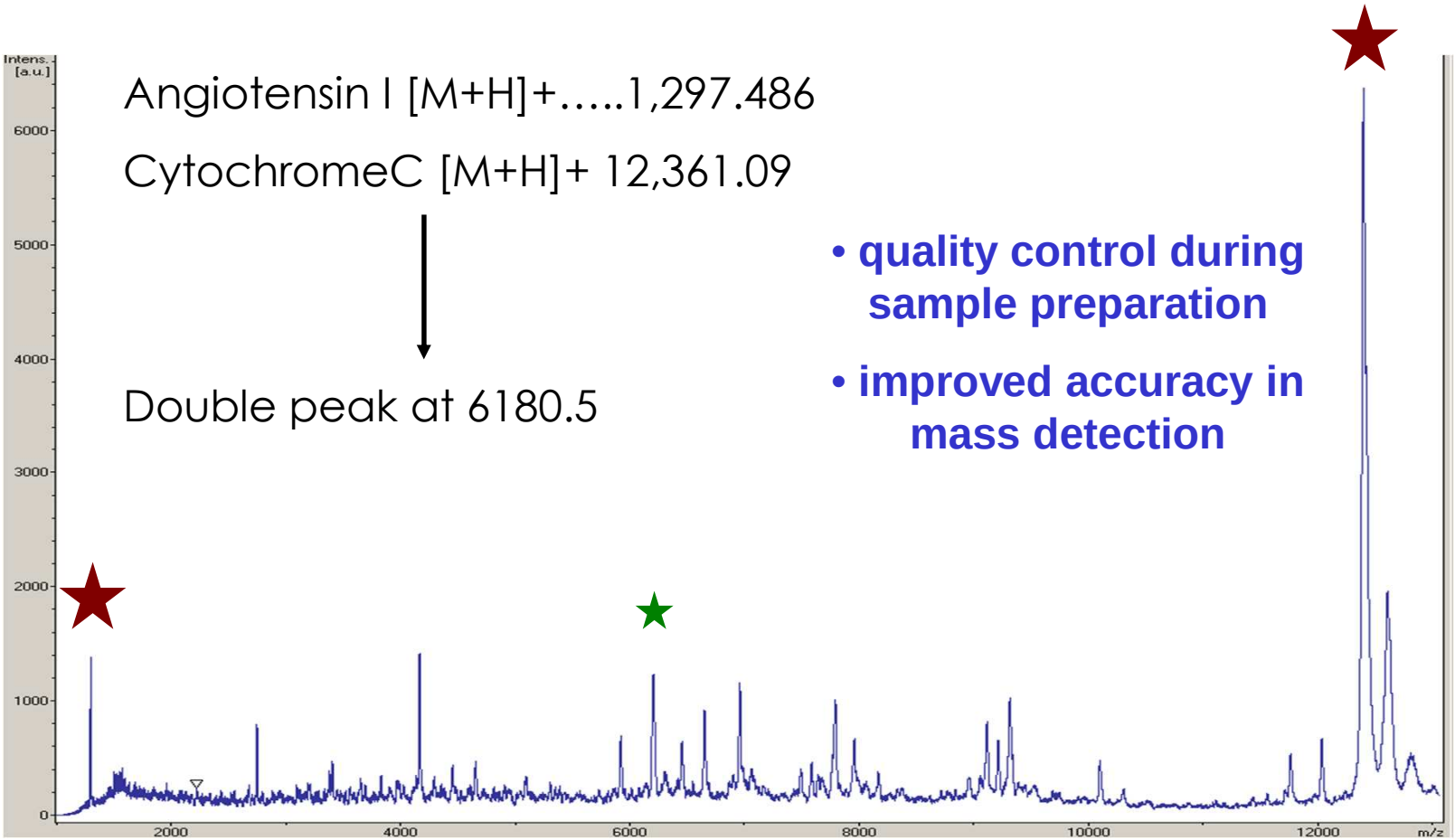
**internal standards**

Angiotensin I [M+H]+…..1,297.486

CytochromeC [M+H]+ 12,361.09

Double peak at 6180.5

- **quality control during sample preparation**
- **improved accuracy in mass detection**



# Normalization

- Each spectrum is scaled according to its Area Under Curve (AUC) → relative measures
- 3 measures for each subject → average spectrum

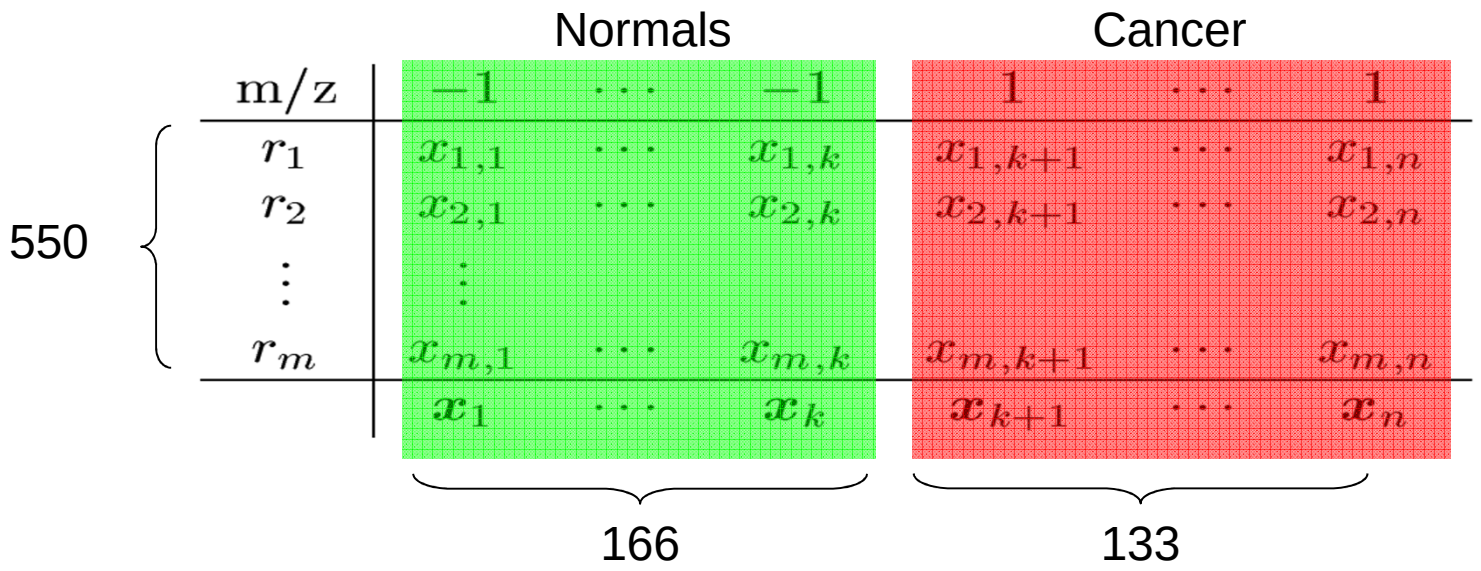
# Binning

Each bin is an interval of m/z values

The intensity of a bin is the sum of all intensities at the m/z in the interval

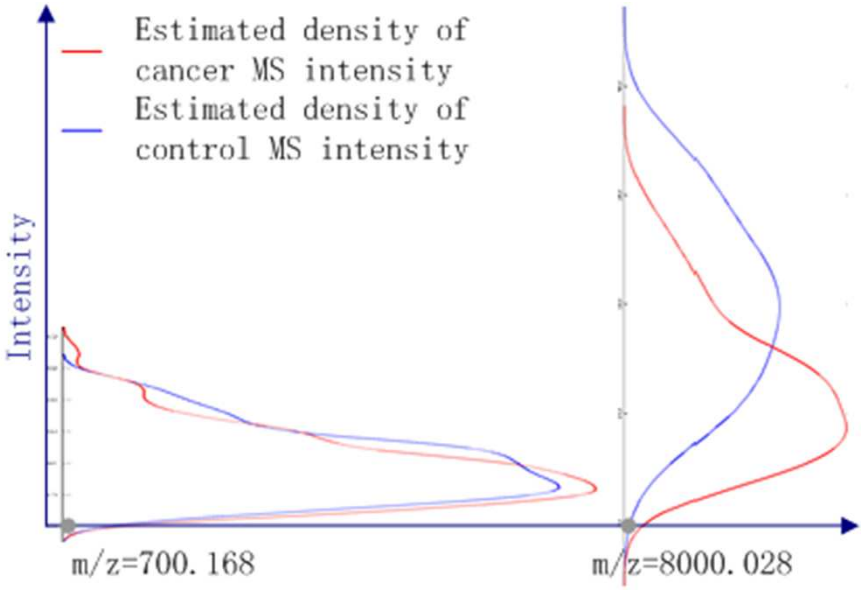
96,430 → ~550

After binning:



# Feature selection

- **Statistical tests** (e.g. t-test, Wilcoxon test, Kolmogorov-Smirnov) to select statistically different binned m/z




~550 → ~40