Tecnologie e applicazioni del Riconoscimento Automatico del Parlato

Fabio Brugnara

brugnara@itc.it

ITC-irst, Centro per la ricerca scientifica e tecnologica

Trento

Sommario

- La comunicazione verbale
- Promesse e problemi del Riconoscimento Automatico del Parlato
- Tecnologie per il RAP: modello acustico e linguistico
- Un sistema completo di base
- Aspetti legati a trascrizione con grandi vocabolari
- Dimostrazioni

Vantaggi della Comunicazione Verbale

Efficiente: ad alto contenuto informativo

Facilmente diffondibile: con telefono, radio, televisione...

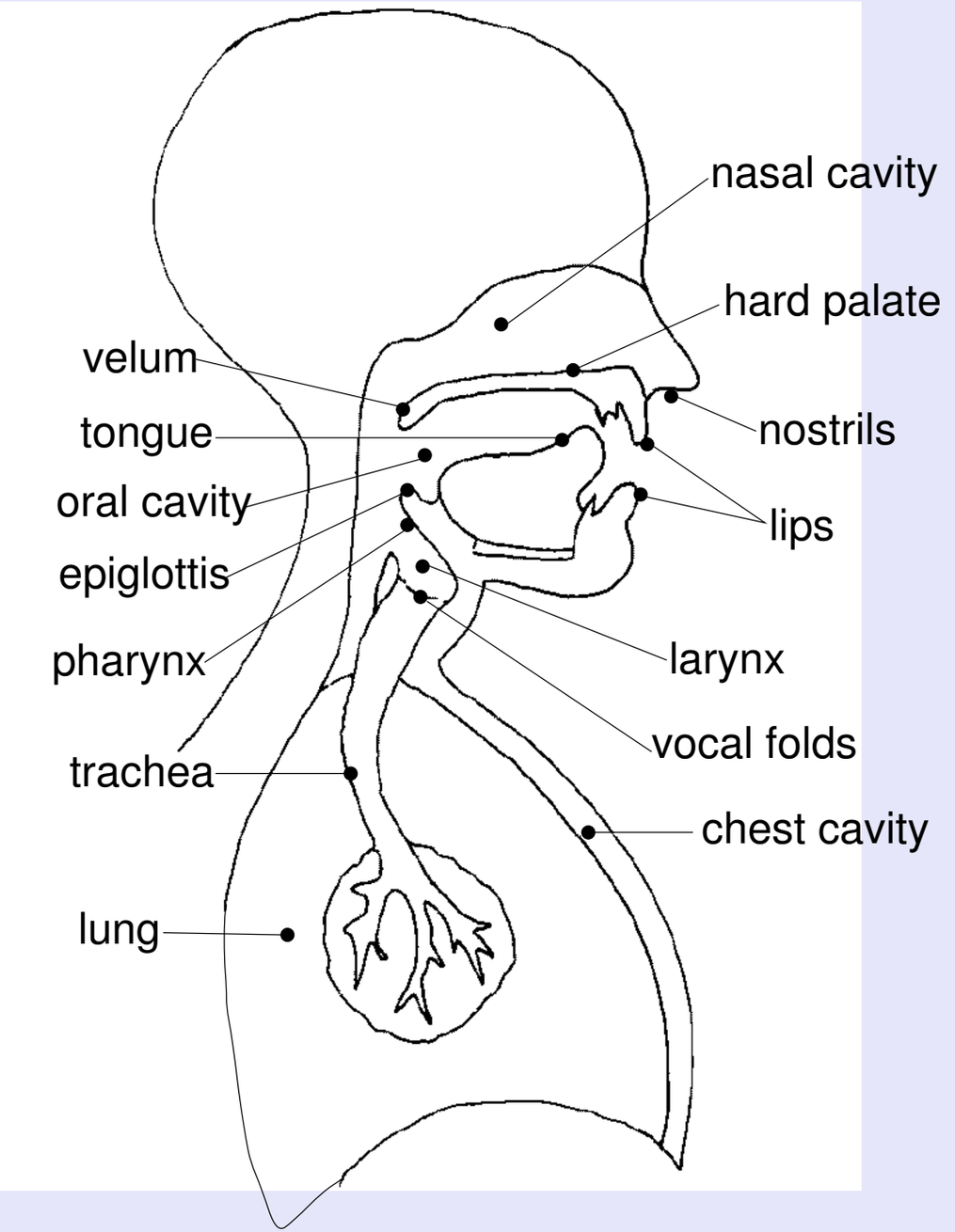
Inoltre, per l'interazione uomo-macchina:

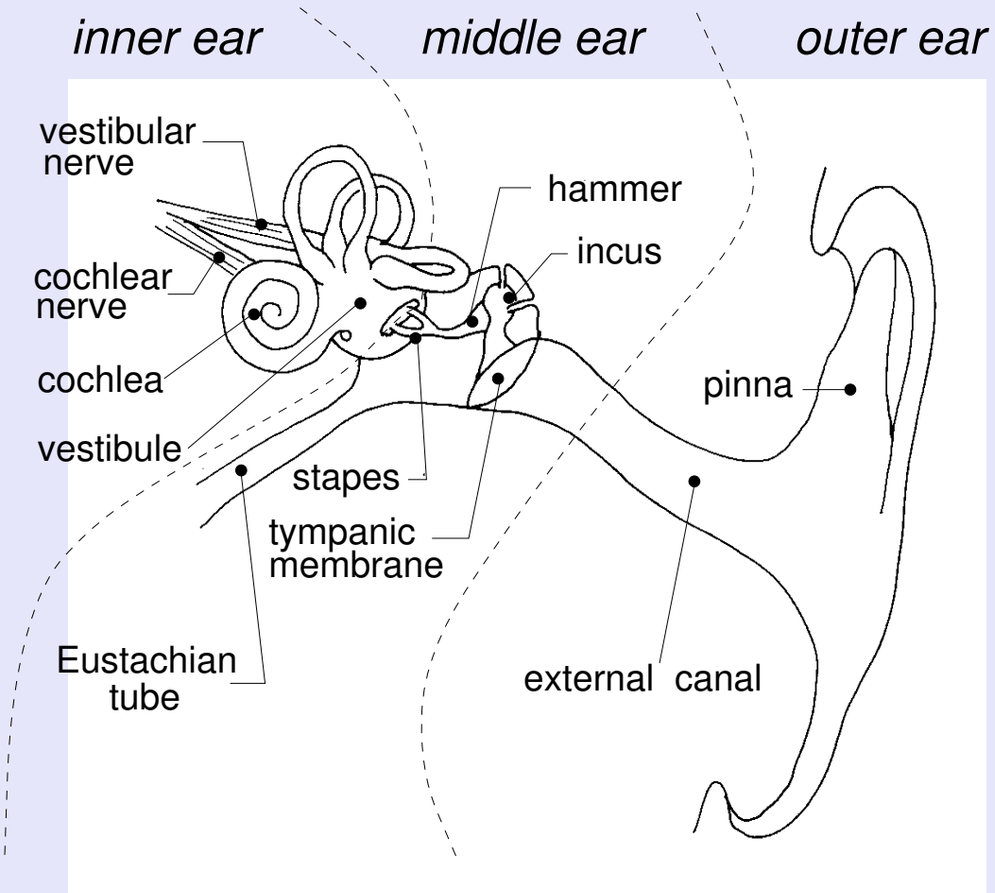
- Naturale: non ha bisogno di un addestramento particolare
- Flessibile: lascia le mani e gli occhi liberi

Le interfacce vocali sono appropriate per l'accesso all'informazione quando:

- Lo spazio informativo è rappresentato in forma lasca e complessa
- È disponibile soltanto il telefono
- Mani ed occhi debbono restare liberi

Produzione del messaggio vocale




inner ear
middle ear
outer ear
vestibular nerve
cochlear nerve
cochlea
vestibule
Eustachian tube
hammer
incus
stapes
tympanic membrane
pinna
external canal

Aree di Applicazione delle Interfacce Vocali

- Input vocale (richiede il riconoscimento)
  - Dettatura automatica
  - Inserimento di dati
  - “Control and Command” semplice
  - Miglioramento delle competenze orali
- Conversazione interattiva (richiede la comprensione)
  - Chiosco informativo
  - Gestione di transazioni
- Gestione di informazione
  - Trascrizione di documenti audio(visivi)
  - Estrazione di informazione

Problemi della Comunicazione Verbale

- L'informazione linguistica e semantica contenuta nel segnale vocale sono mescolate a molti fattori irrilevanti
- Si deve passare da un dominio continuo ad un dominio discreto. Per le parole scritte esiste una precisa identità, per le parole pronunciate no
- Si è "obbligati" a tracciare confini dove i confini non ci sono
- C'è dell'ambiguità ineludibile: p.e.

  "ciascuno a casa propria" = "ciascuno ha casa propria"

  "malanno preso" = "ma l'hanno preso"

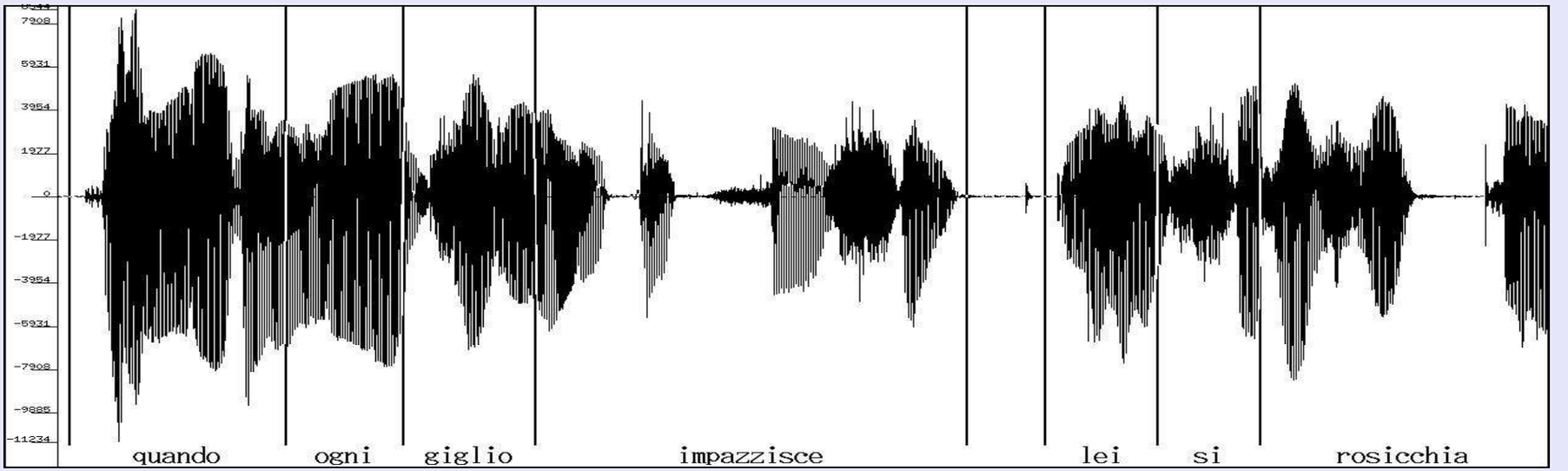
  "vado per recarlo a Trento" = "vado per re Carlo a Trento"

  sono omofoni, e anche plausibili linguisticamente

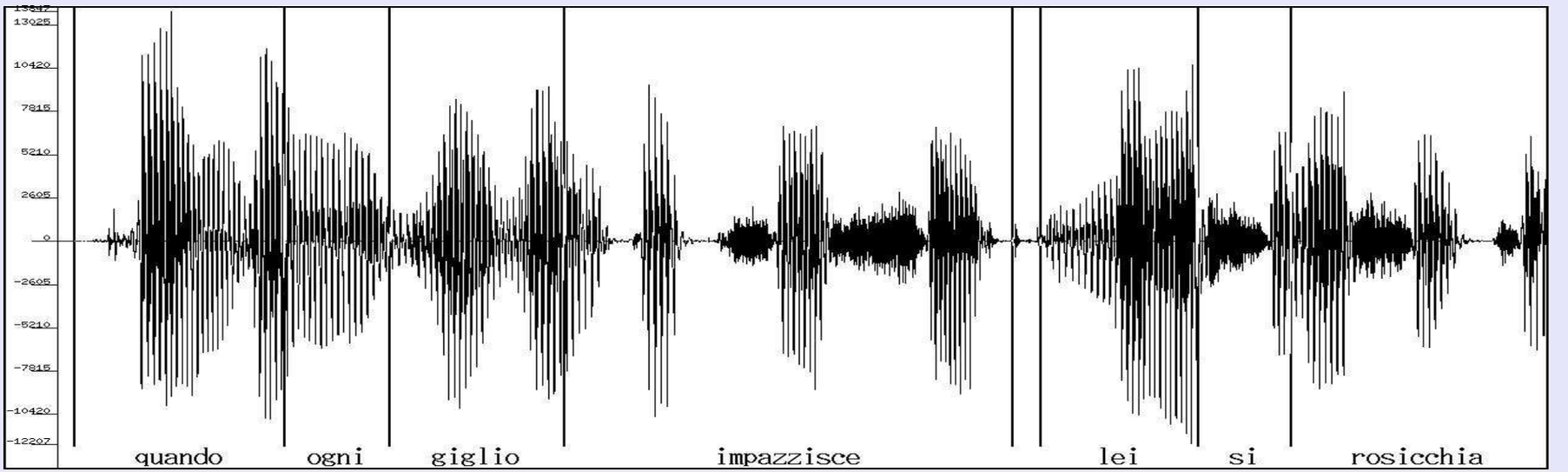
Principali Difficoltà nel Riconoscimento Automatico del Parlato

- Coarticolazione. La realizzazione acustica dei suoni elementari dipende fortemente dal contesto in cui occorre
- Dipendenza dal parlatore. Le caratteristiche della voce variano da parlatore a parlatore
- Parlato spontaneo. È caratterizzato da vari fattori come disfluenze, parole fuori vocabolario, false partenze, etc. che sono difficili da trattare
- Rumore. Le condizioni ambientali di acquisizione possono influire notevolmente sulla accuratezza del riconoscimento

Il punto di partenza: la forma d'onda



Voce femminile

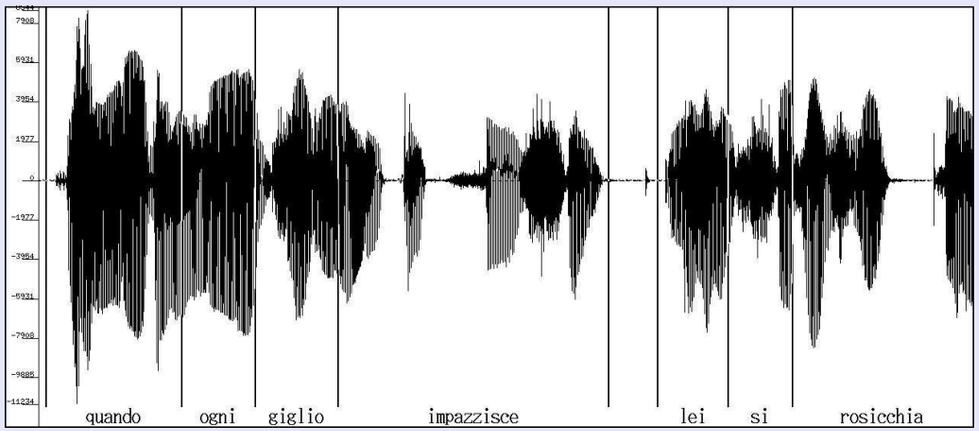


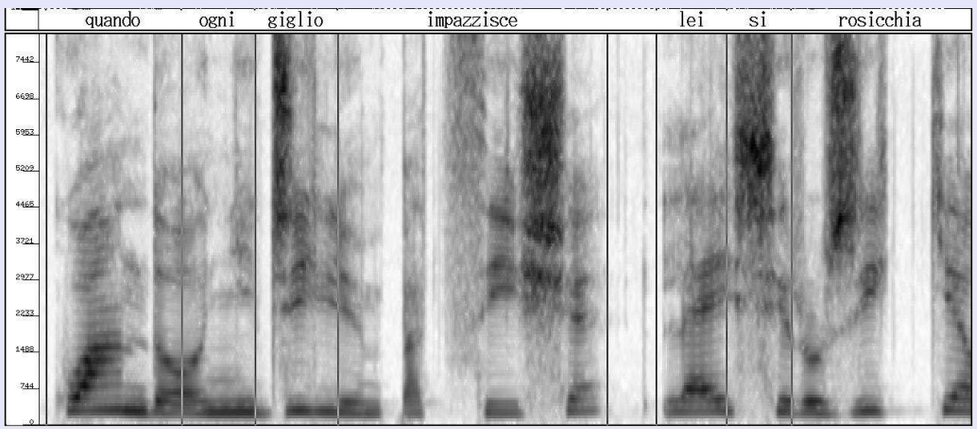
Voce maschile

Fondamenti della rappresentazione acustica

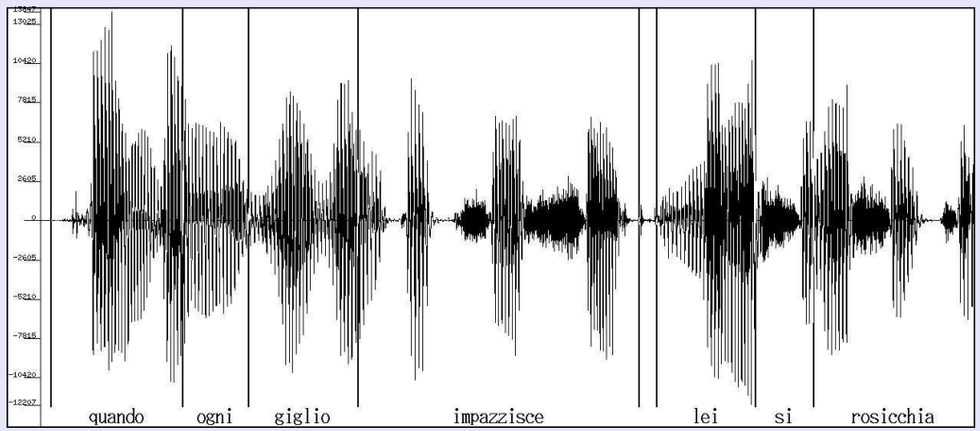
- La forma d'onda è molto ridondante ai fini del contenuto linguistico
- Sappiamo che l'orecchio effettua una sorta di analisi in frequenza, con una risoluzione che diminuisce al crescere della frequenza
- Già nel 1937 è stata proposta la "scala MEL" che riflette la variazione di accuratezza nel discriminare fra frequenze diverse.
- Negli anni '40 viene introdotto lo spettrogramma:
  - Il segnale viene "spezzettato" in brevi segmenti (p.e. 20*ms*), detti frames, in cui si assume stazionario
  - Su questi si effettua una trasformata di Fourier discreta per trovare l'intensità di ciascuna frequenza elementare
  - Il segmento viene rappresentato con il vettore di intensità

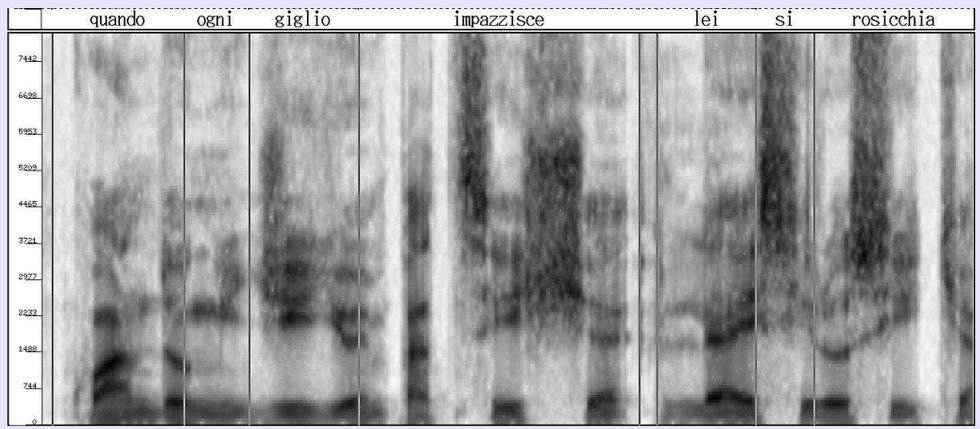
Spettrogramma





Voce femminile





Voce maschile

Primo approccio: regole da “esperto”

- Con dovuta preparazione, uno spettrogramma si può “leggere”. Ci sono formanti, armoniche, ecc...
- I primi tentativi cercavano di “meccanizzare” questa abilità.
- Purtroppo, funziona solo in un contesto molto circoscritto.
- Disillusione, inizia un lungo cammino. Che avesse ragione Zenone?
- Ma rimane la elaborazione frame based e spectrum oriented
- Le “features acustiche” d’ora in poi saranno dei vettori di reali che rappresentano un breve segmento di segnale, con lunghezza attorno ai 10ms

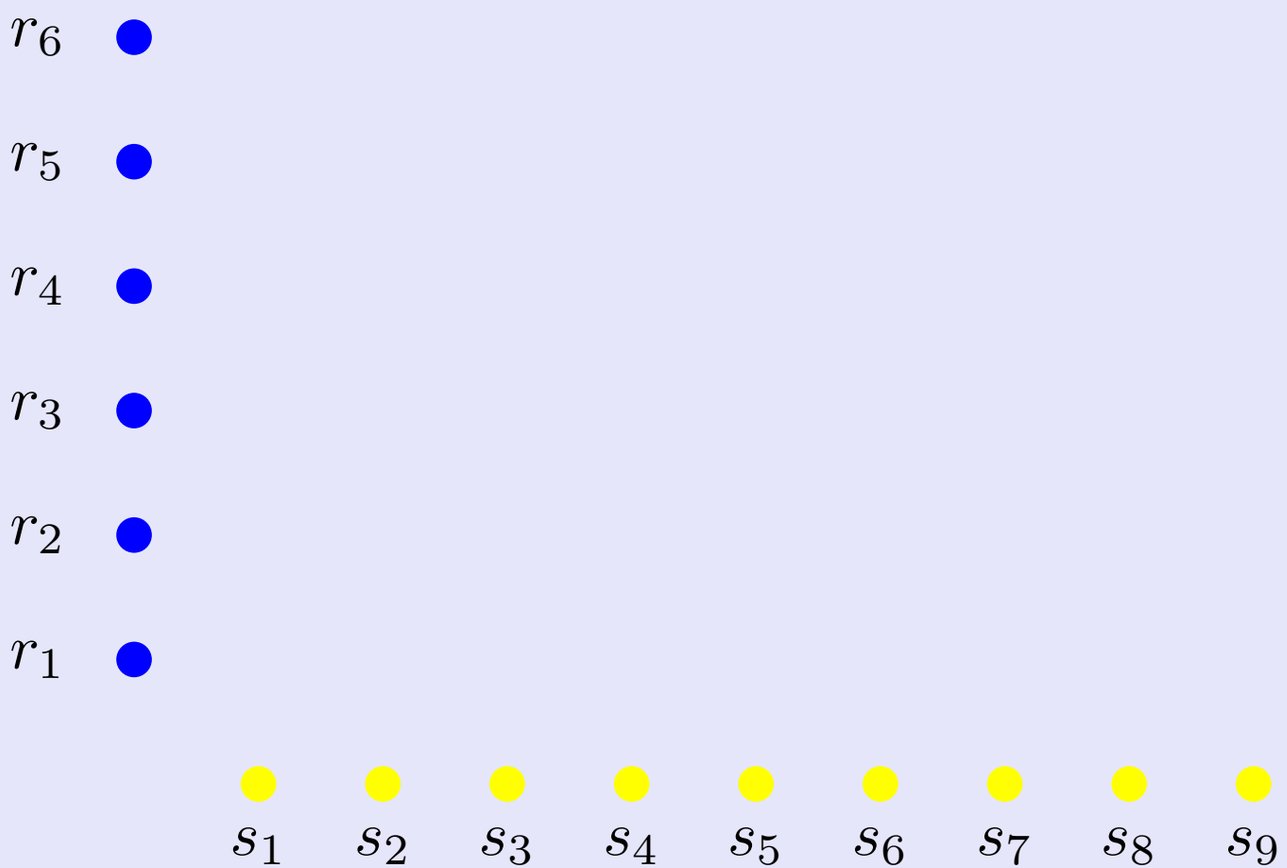
Approccio basato su esempi

- La conoscenza viene estratta dai dati invece che inserita a priori
- Si seleziona un esempio significativo per ciascun elemento da riconoscere
- In riconoscimento, si valuta la distanza dei dati in ingresso da ciascun esempio, e si sceglie il più vicino.
- Si caratterizza in base a:
  - scelta dell'esempio per ciascun suono
  - scelta della misura di distanza
- Le sequenze da confrontare non sono di lunghezza fissa..

Dynamic Time Warping

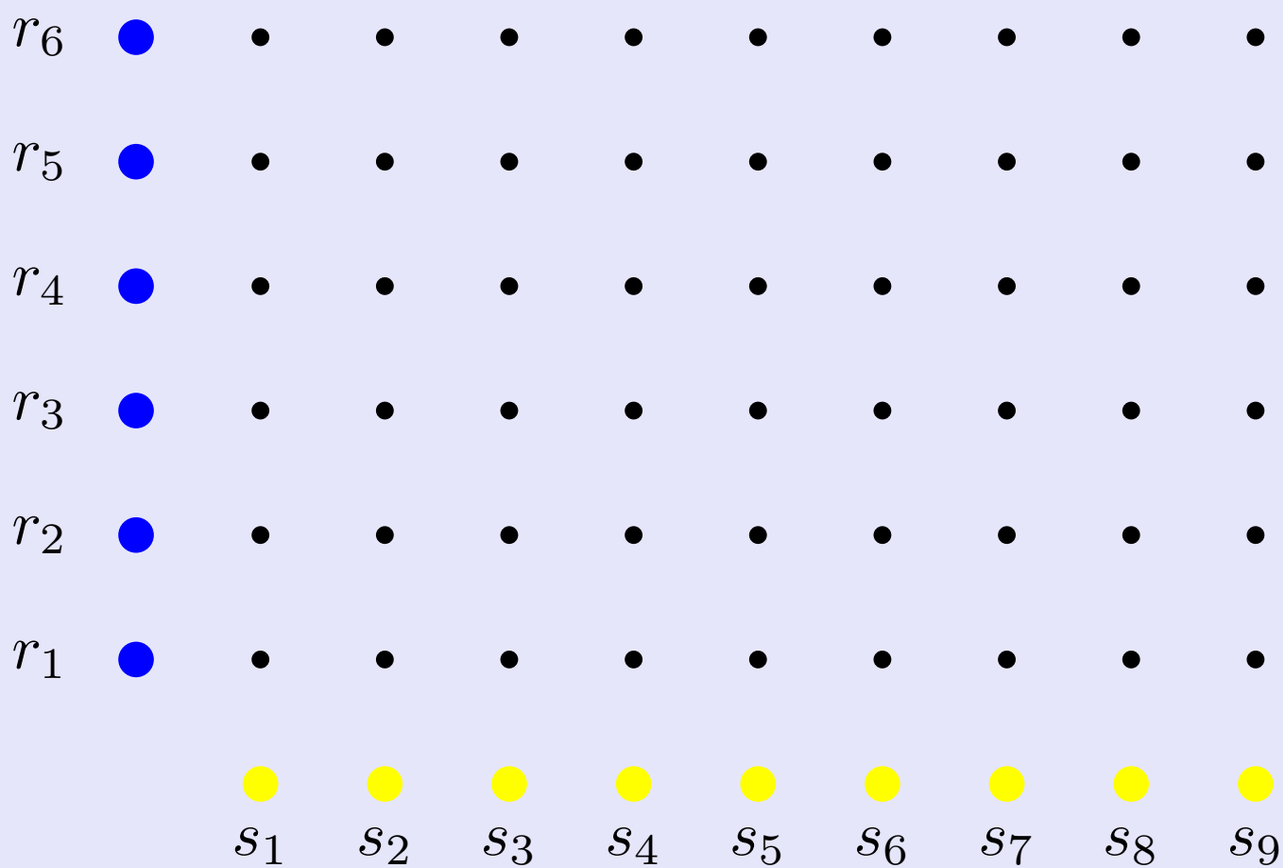
Confronta sequenze di features di lunghezza diversa in maniera efficiente grazie alla programmazione dinamica.

Si cerca un allineamento (vincolato) che minimizzi la distorsione totale, somma delle distanze locali.



Confronta sequenze di features di lunghezza diversa in maniera efficiente grazie alla programmazione dinamica.

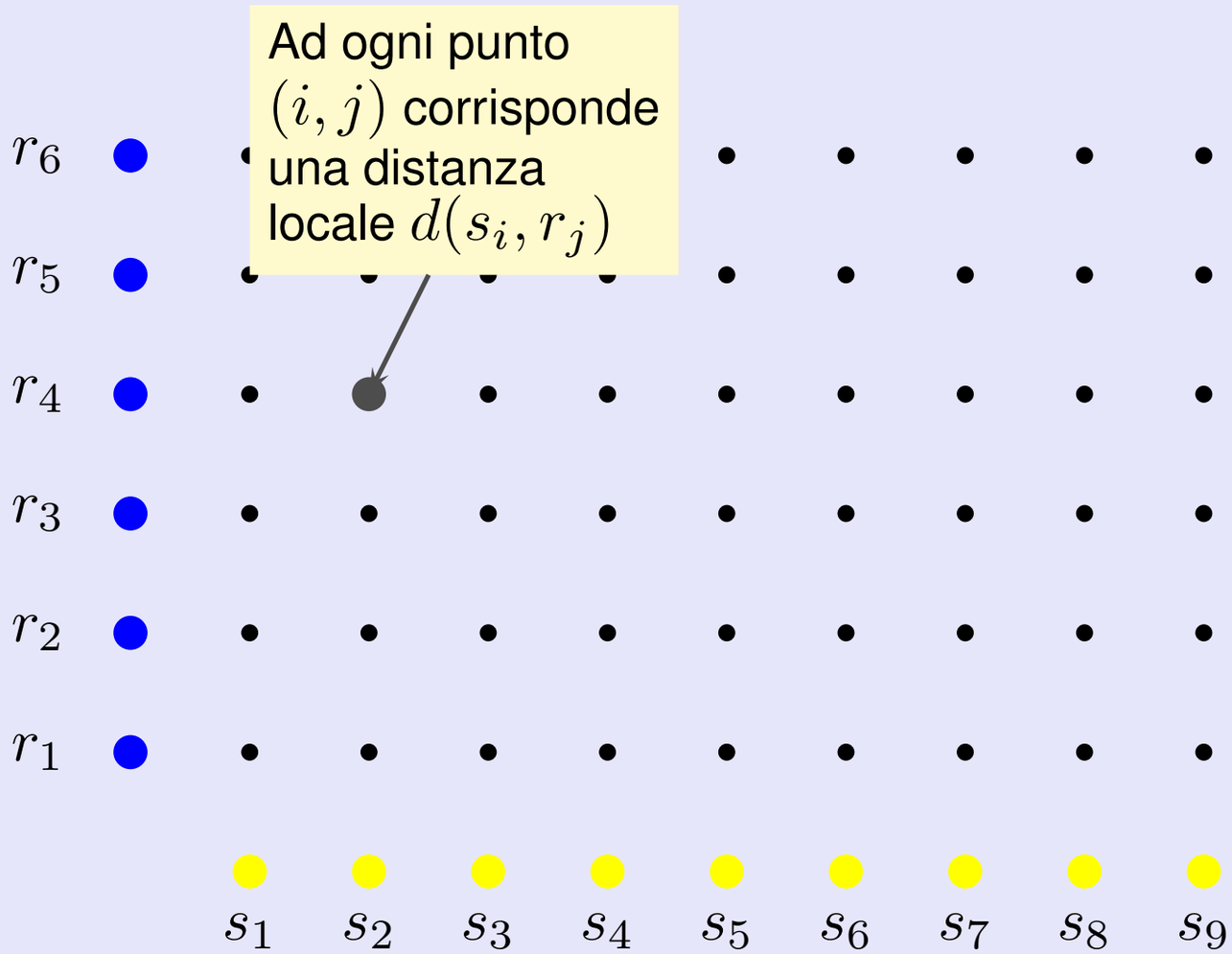
Si cerca un allineamento (vincolato) che minimizzi la distorsione totale, somma delle distanze locali.



Dynamic Time Warping

Confronta sequenze di features di lunghezza diversa in maniera efficiente grazie alla programmazione dinamica.

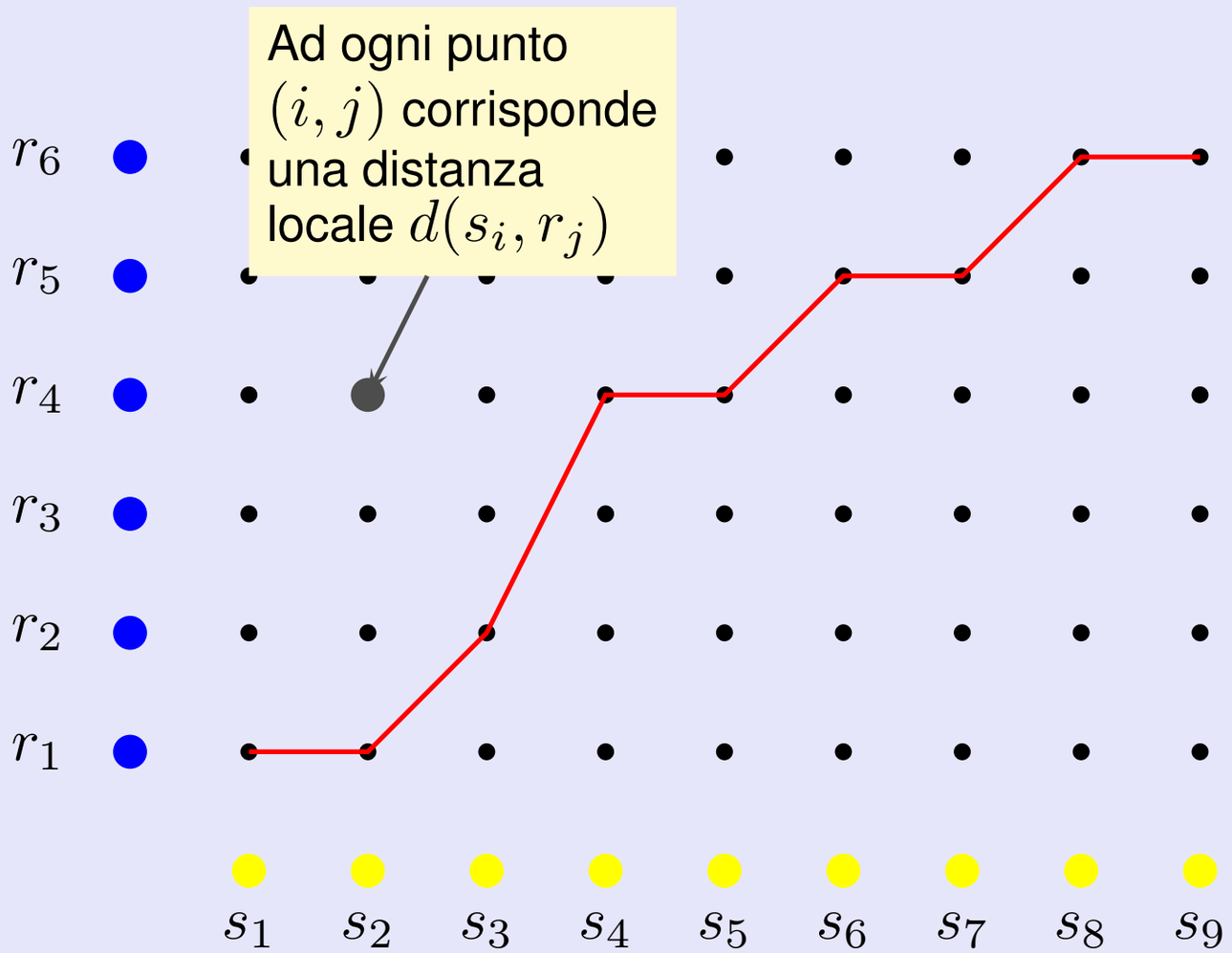
Si cerca un allineamento (vincolato) che minimizzi la distorsione totale, somma delle distanze locali.



Dynamic Time Warping

Confronta sequenze di features di lunghezza diversa in maniera efficiente grazie alla programmazione dinamica.

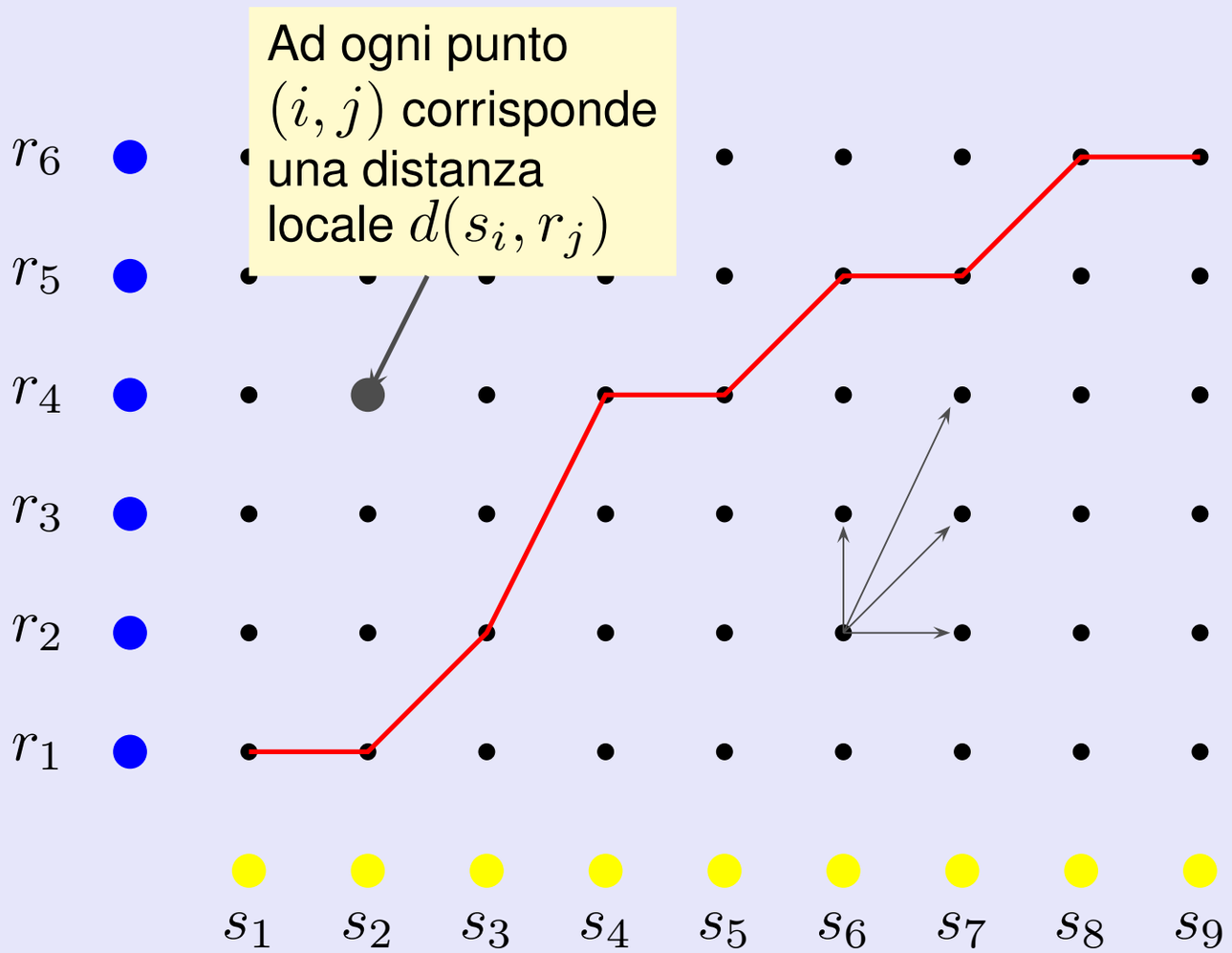
Si cerca un allineamento (vincolato) che minimizzi la distorsione totale, somma delle distanze locali.



Dynamic Time Warping

Confronta sequenze di features di lunghezza diversa in maniera efficiente grazie alla programmazione dinamica.

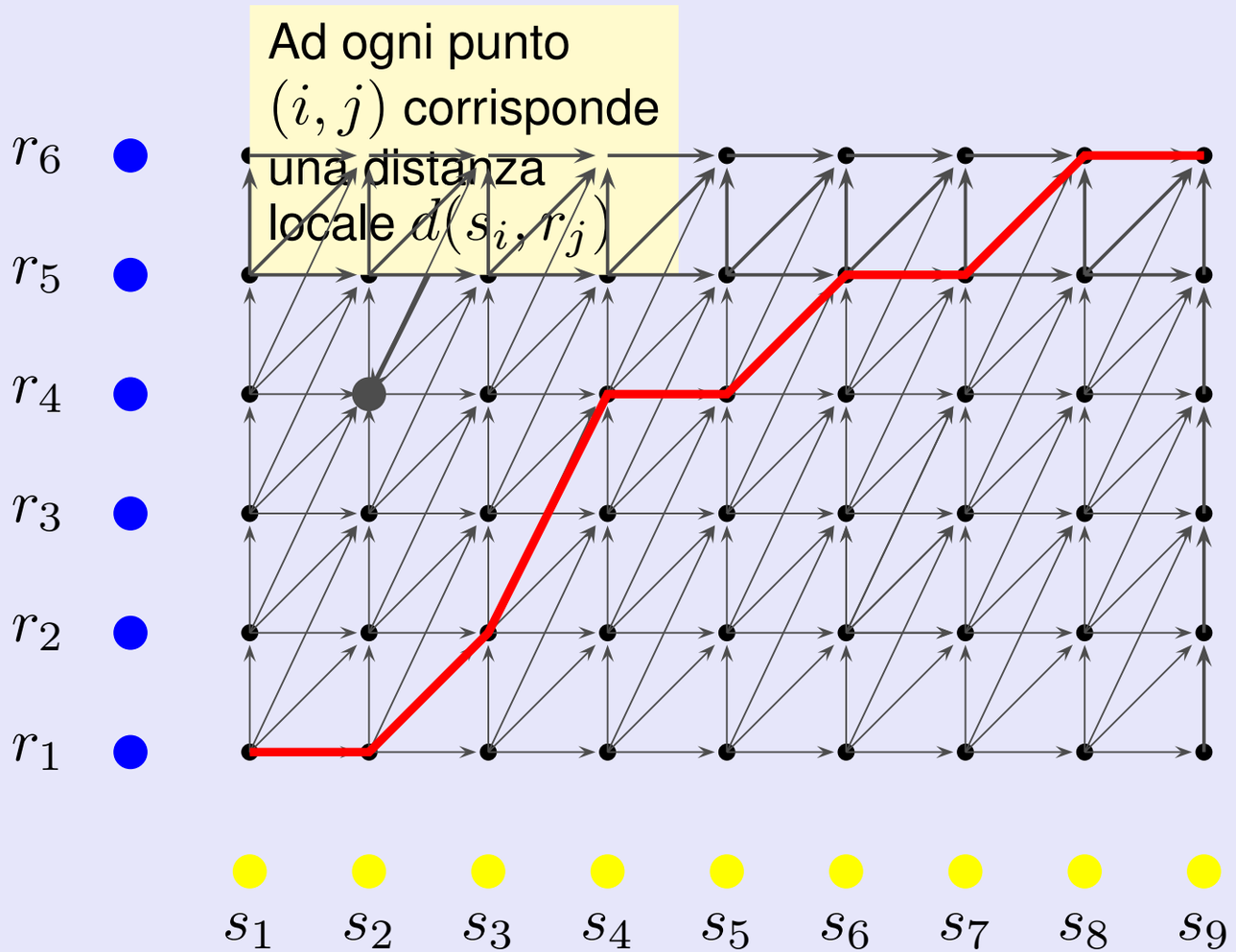
Si cerca un allineamento (vincolato) che minimizzi la distorsione totale, somma delle distanze locali.



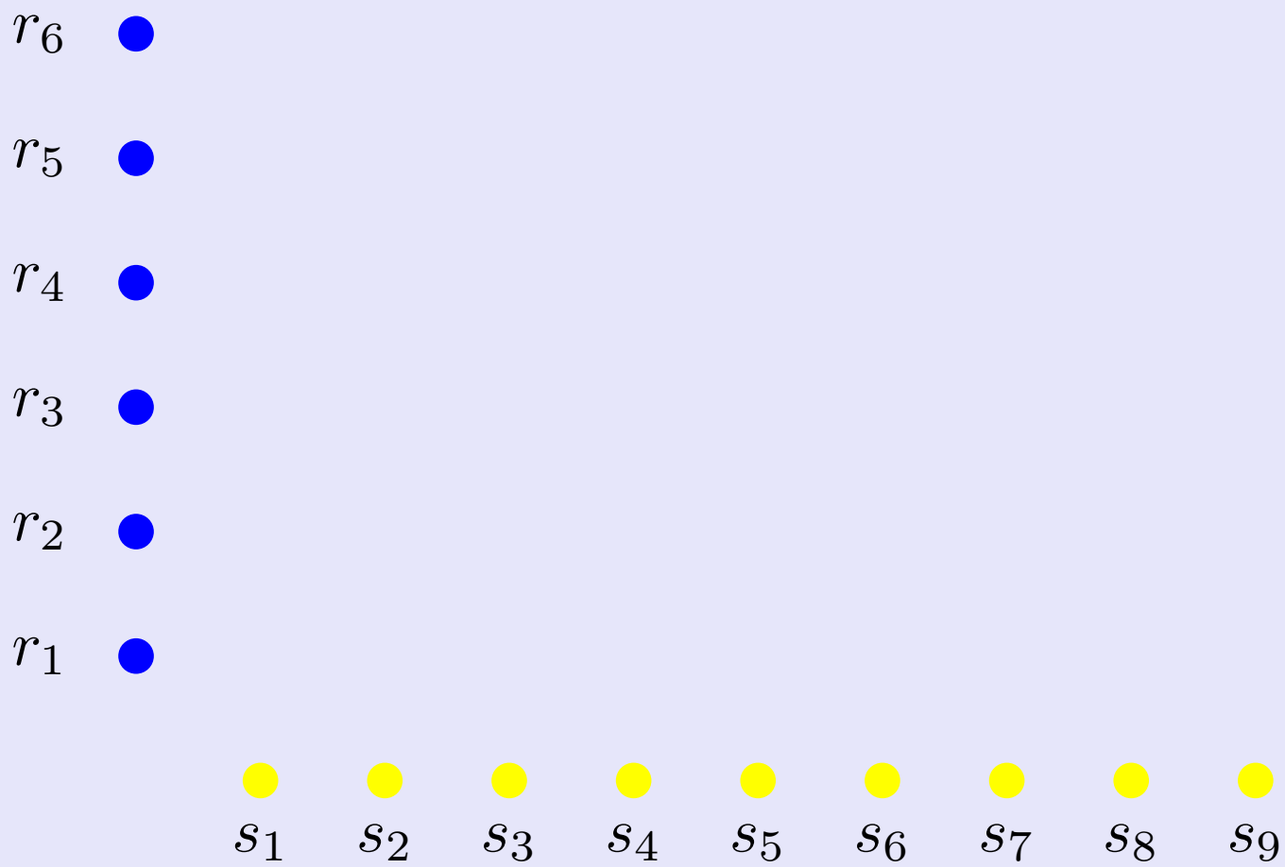
Dynamic Time Warping

Confronta sequenze di features di lunghezza diversa in maniera efficiente grazie alla programmazione dinamica.

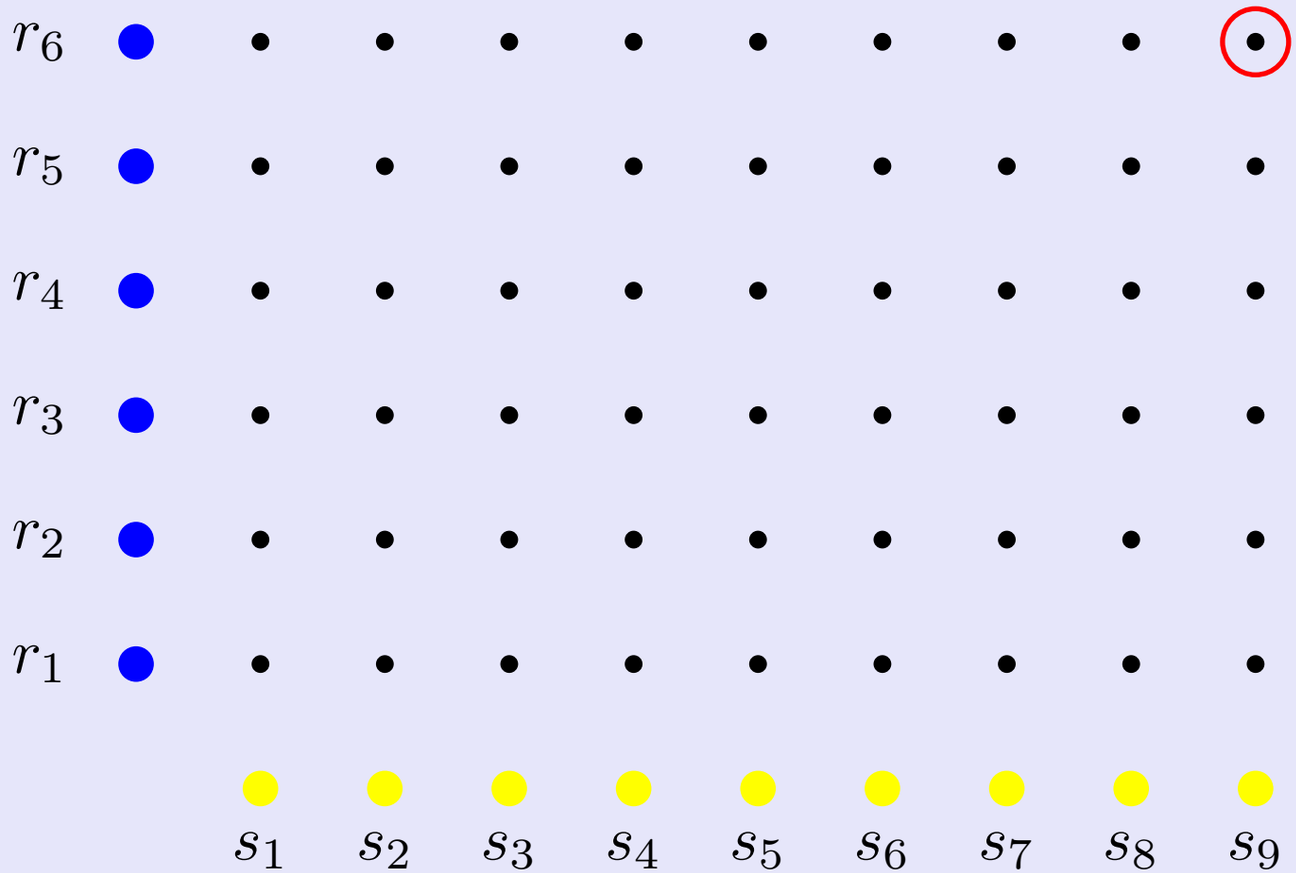
Si cerca un allineamento (vincolato) che minimizzi la distorsione totale, somma delle distanze locali.



Programmazione Dinamica per DTW

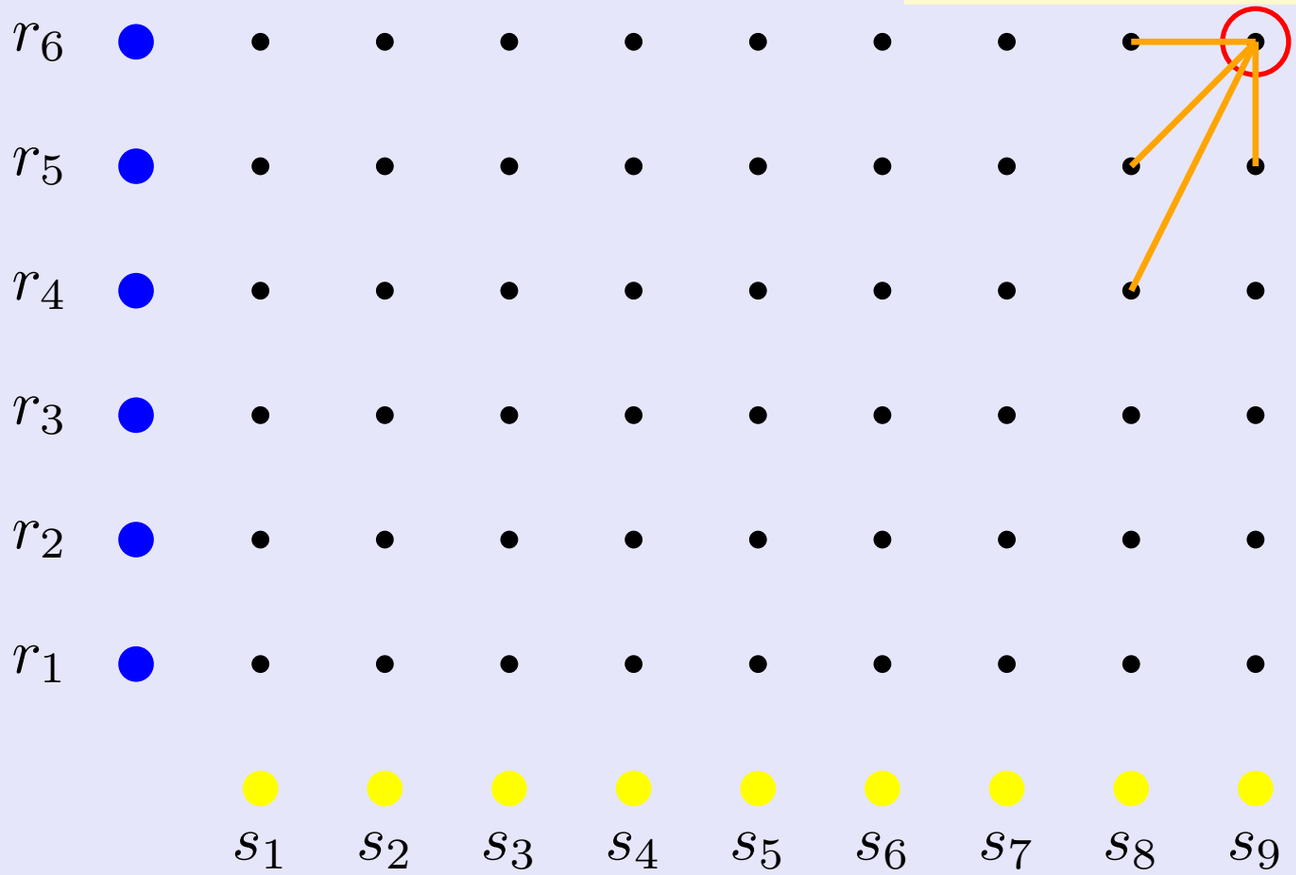


Programmazione Dinamica per DTW



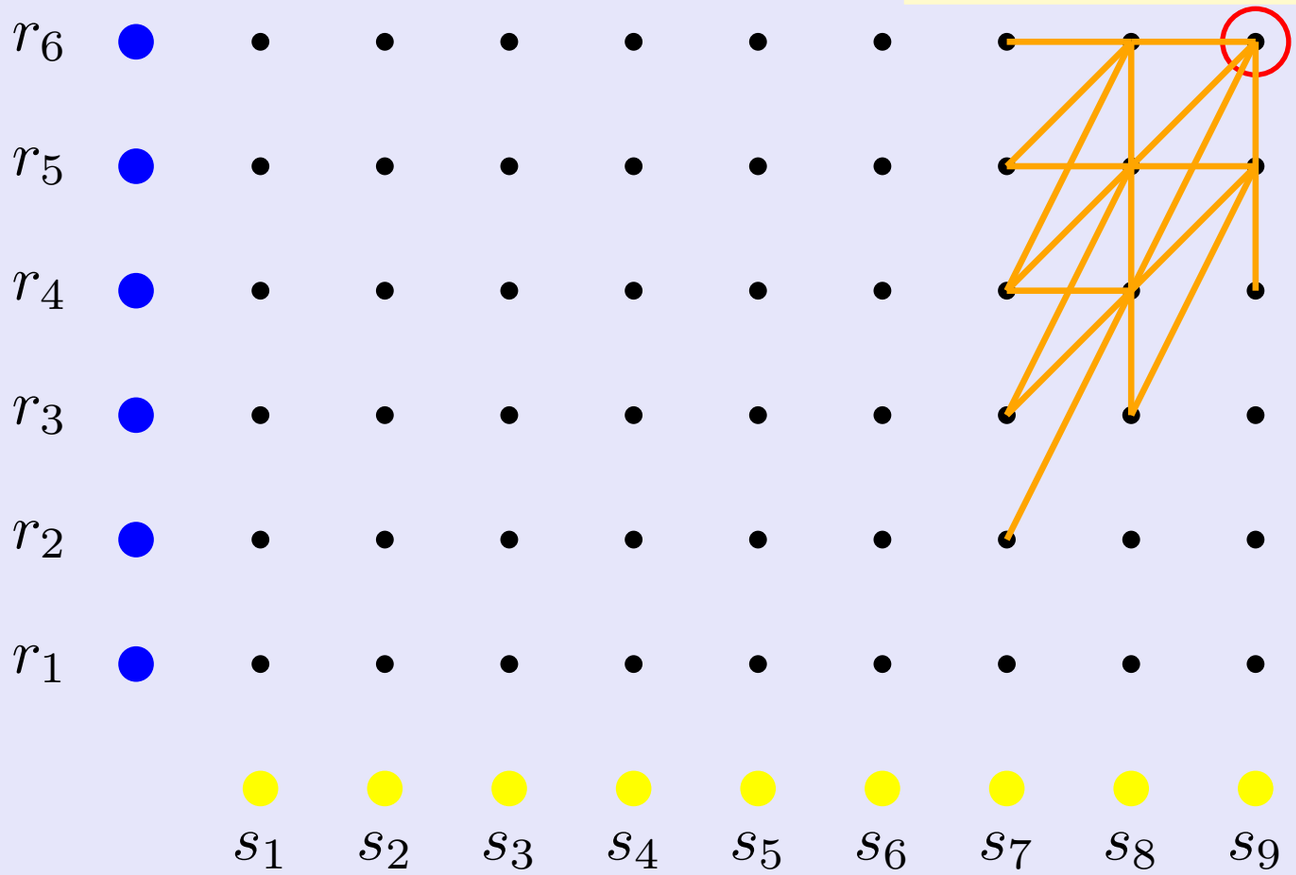
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



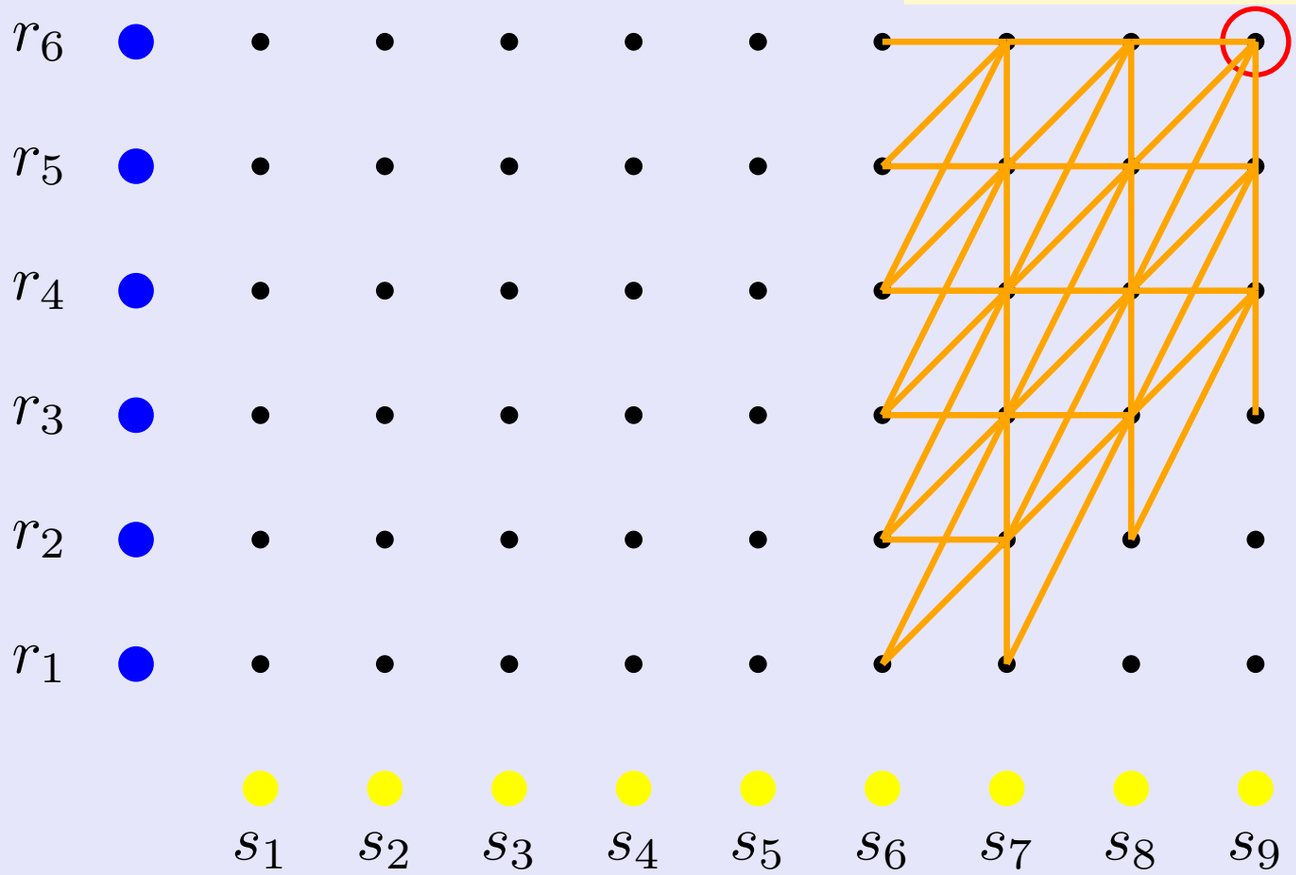
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



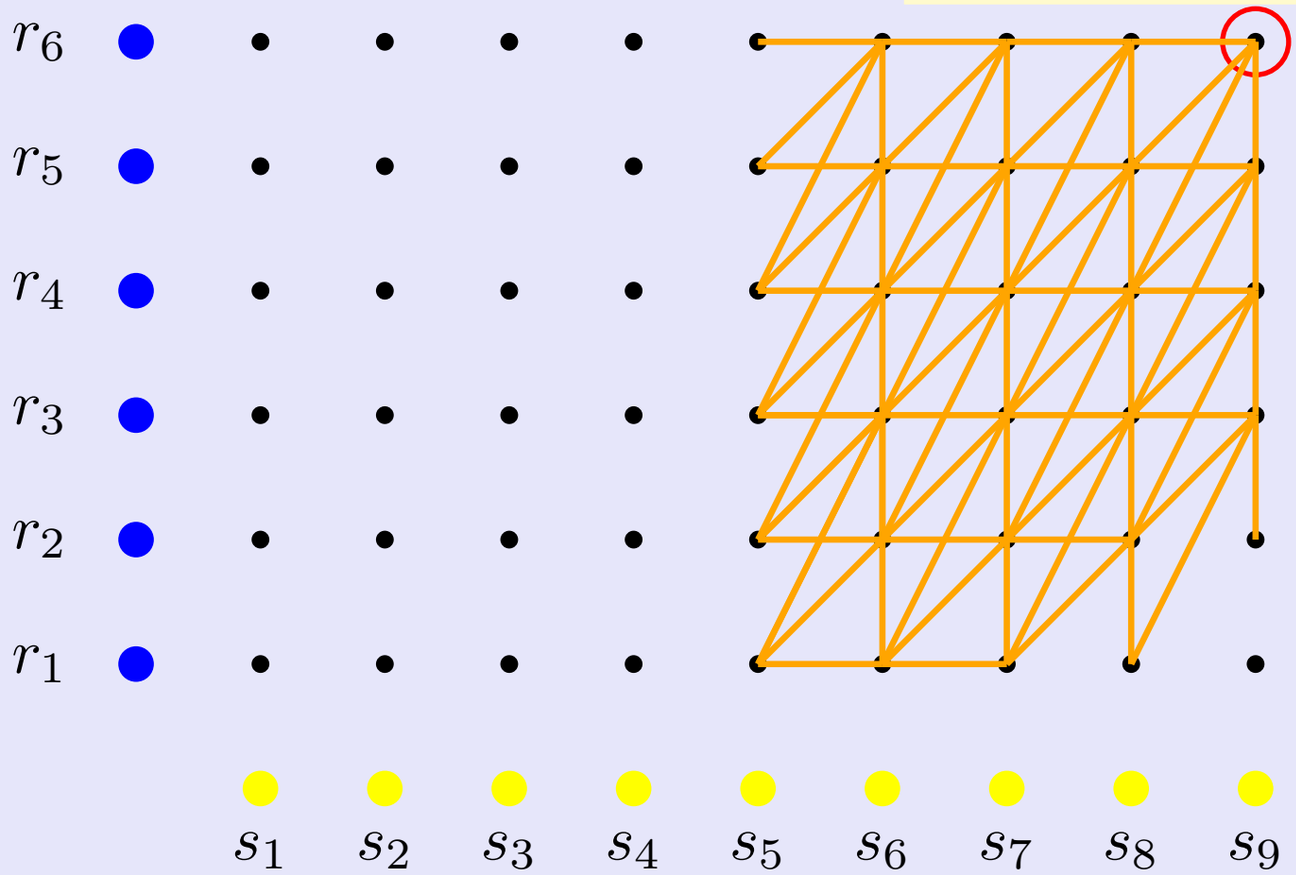
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



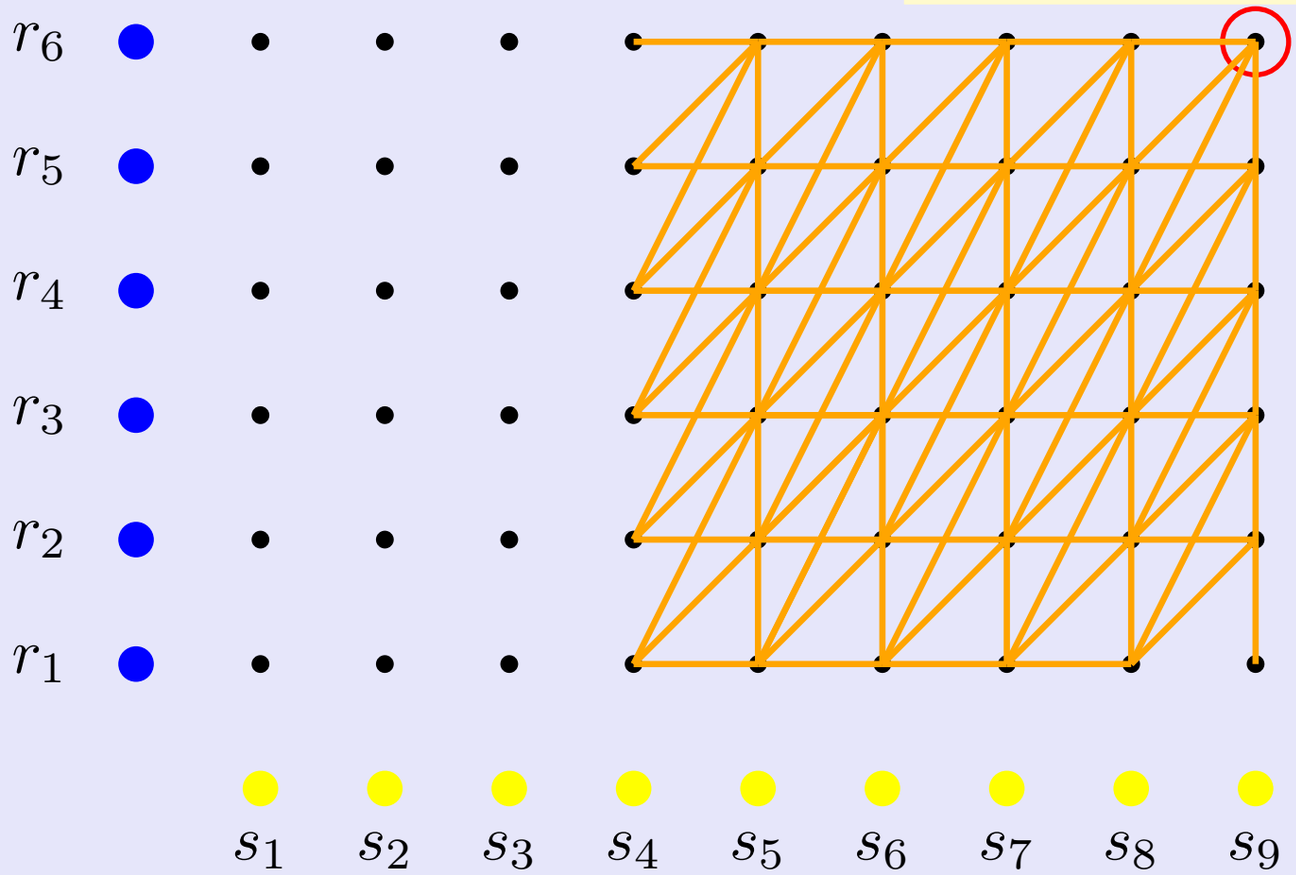
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



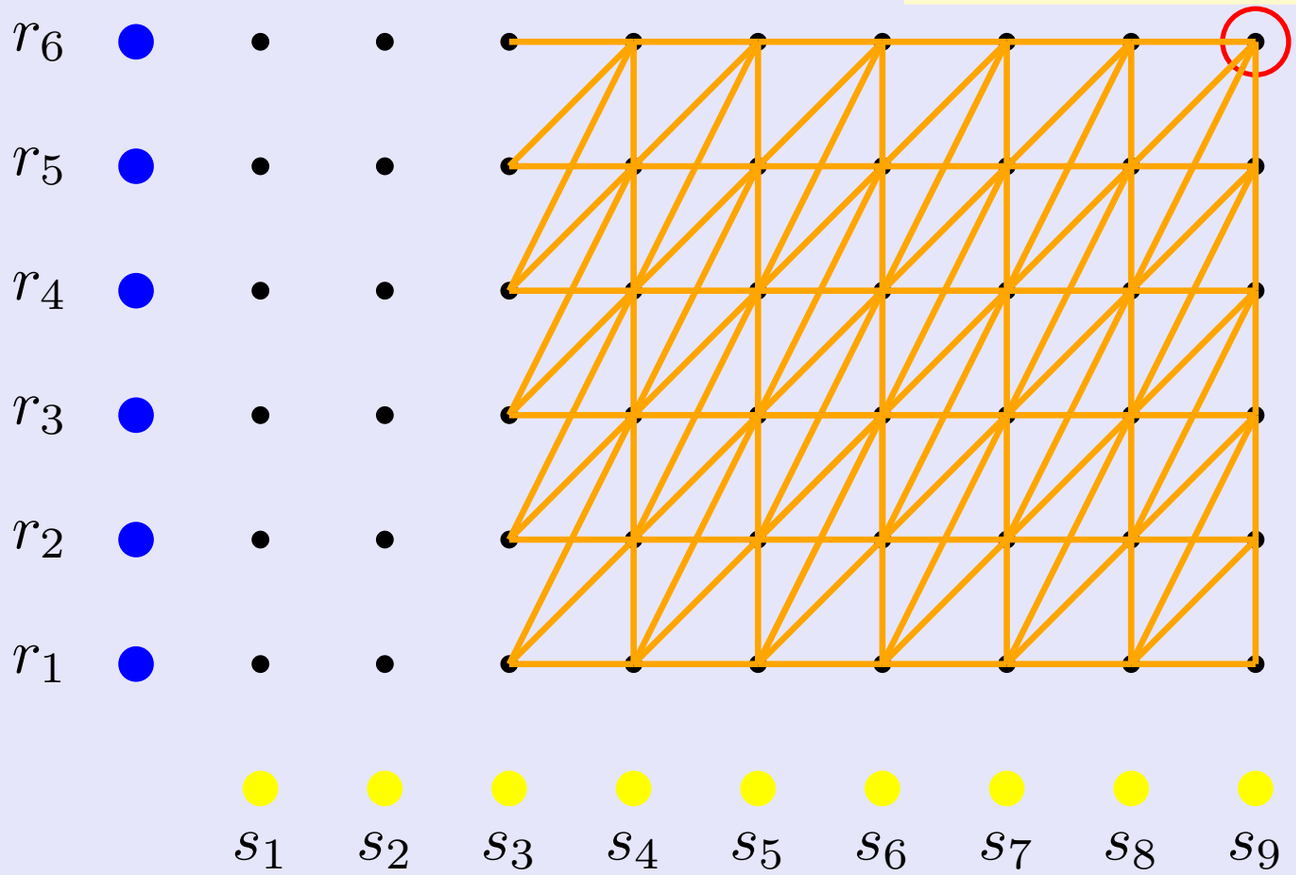
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



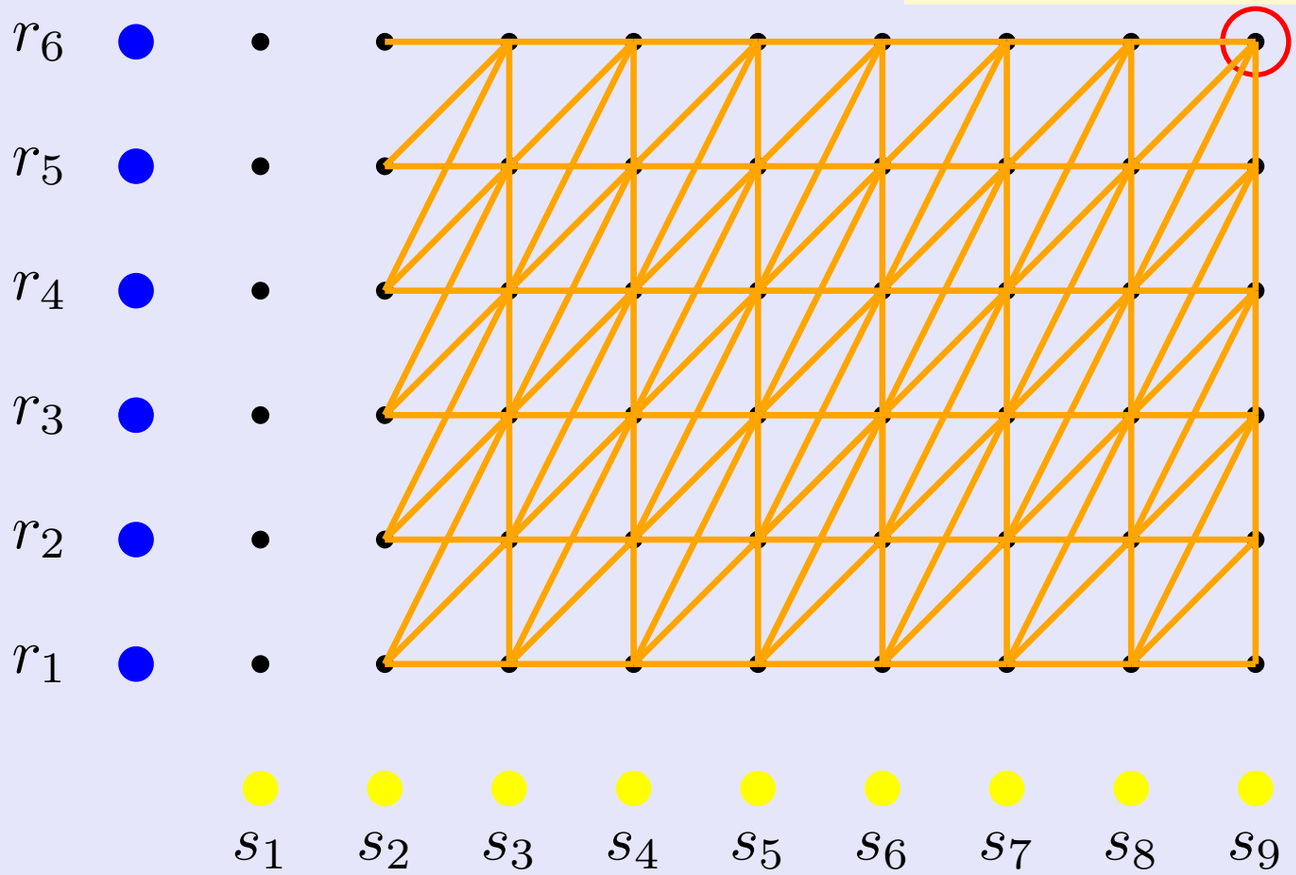
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



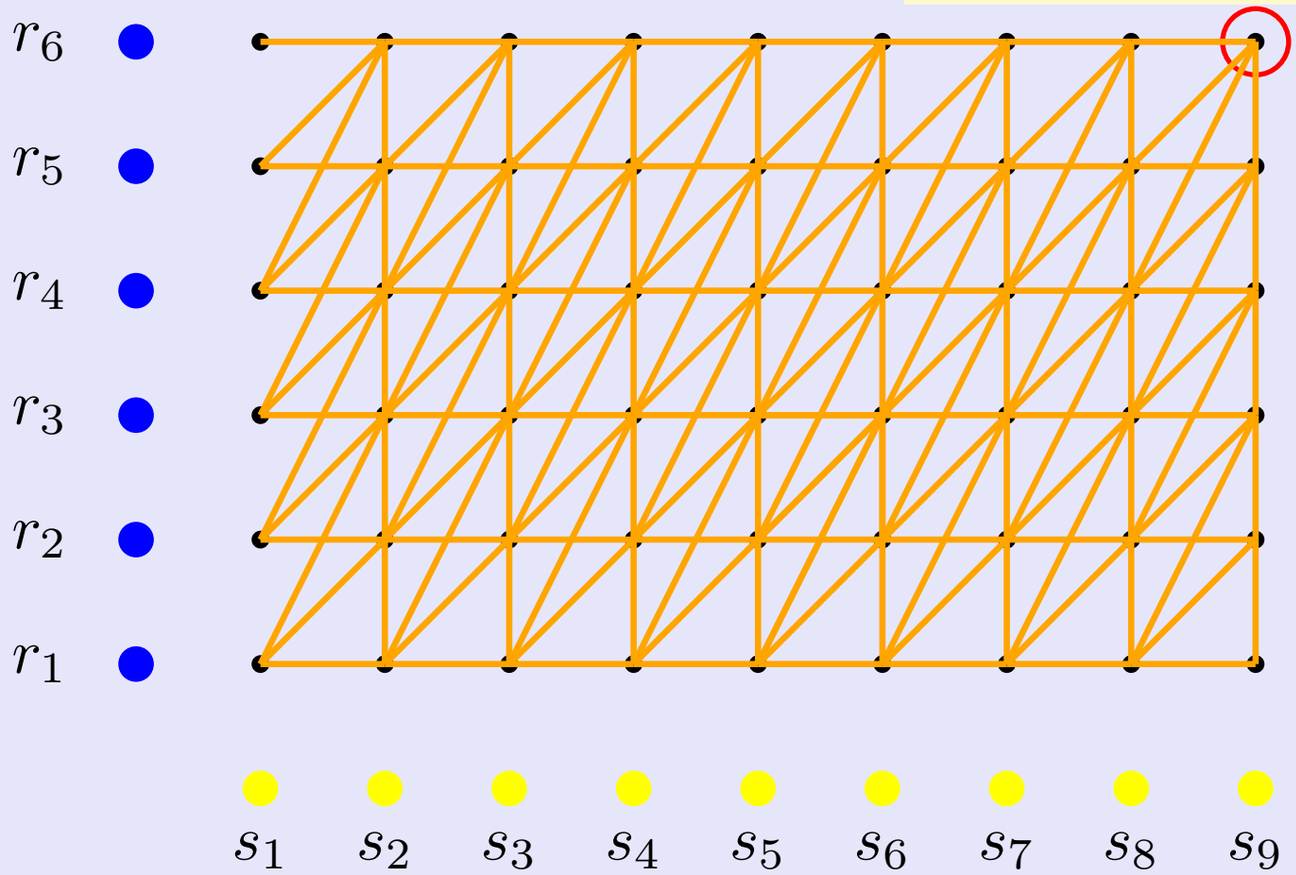
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



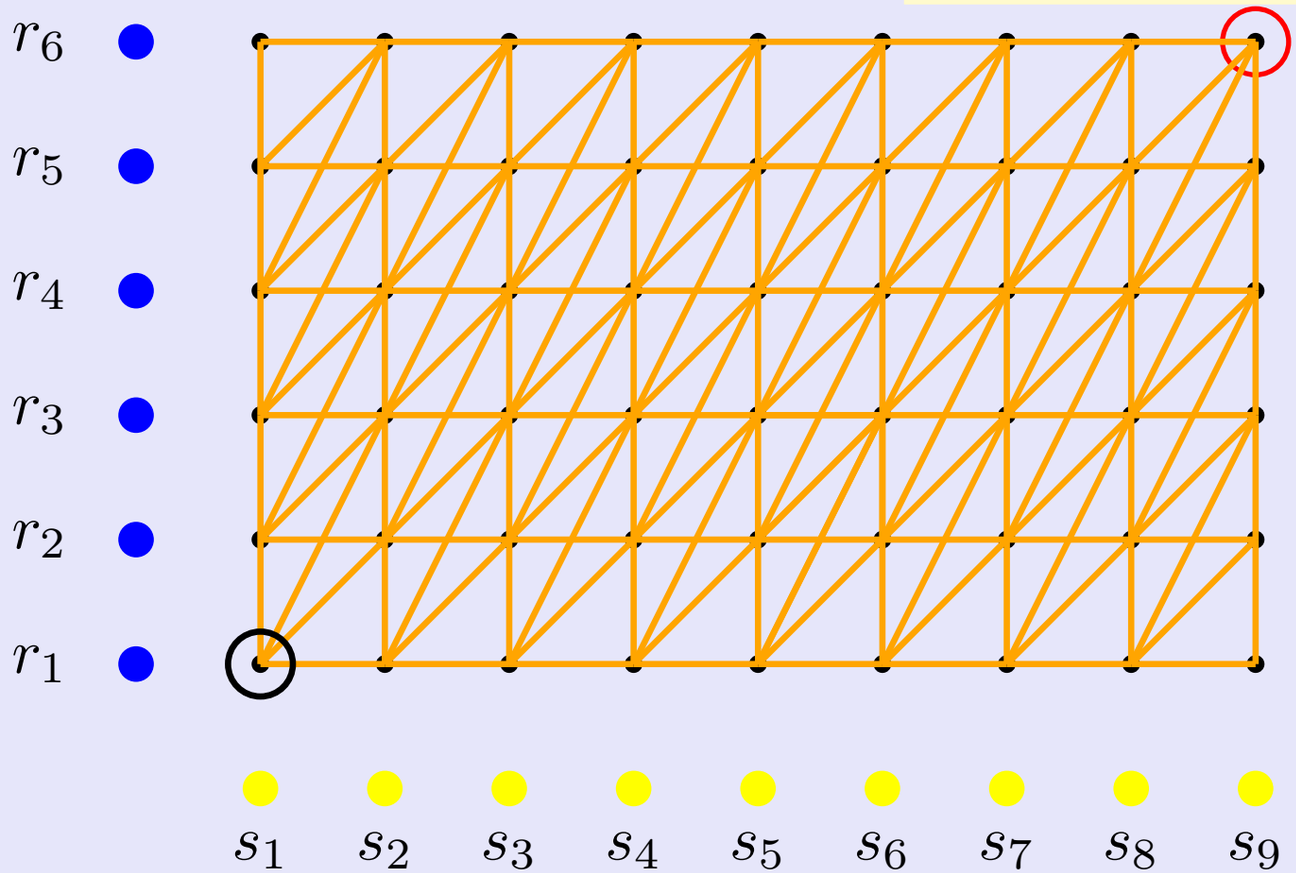
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



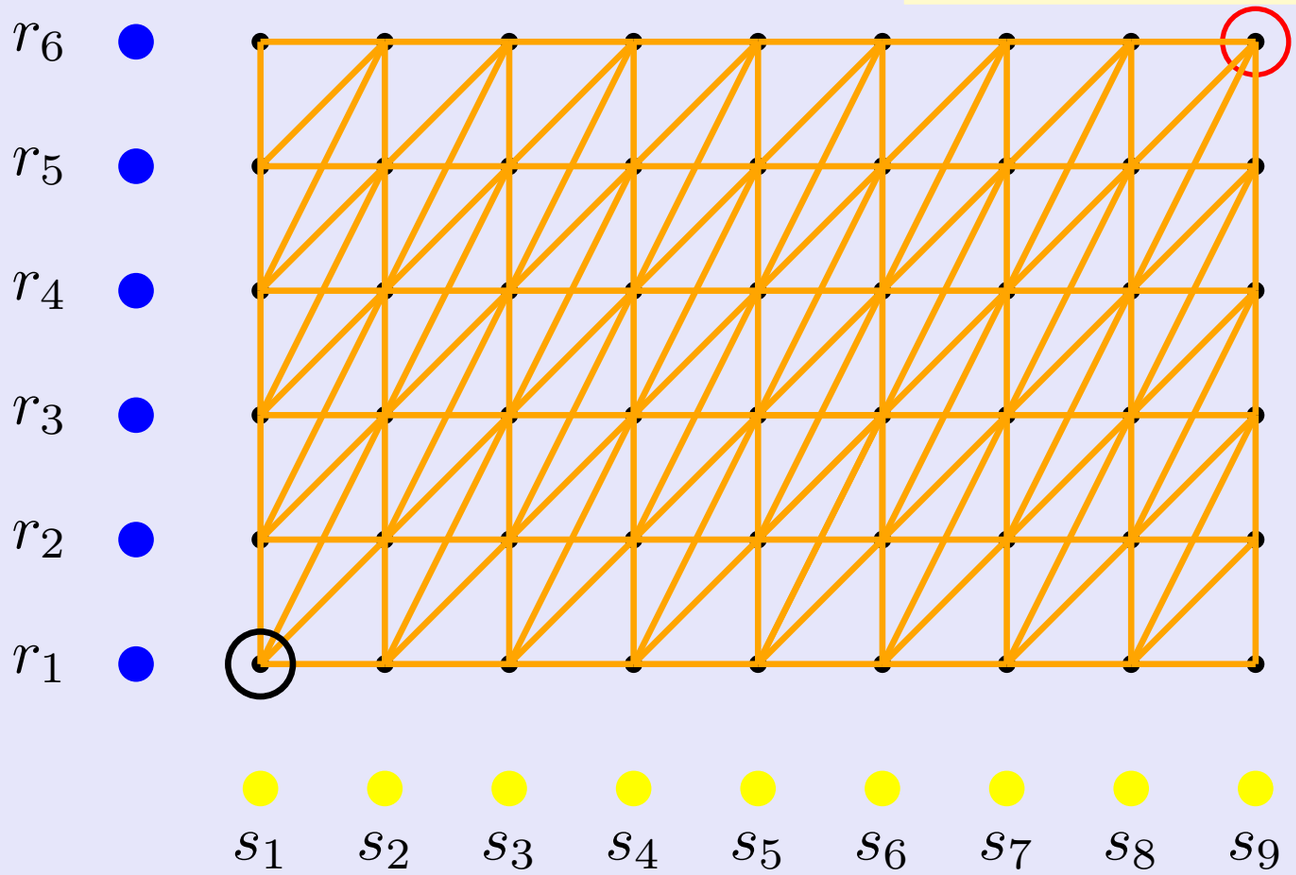
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.



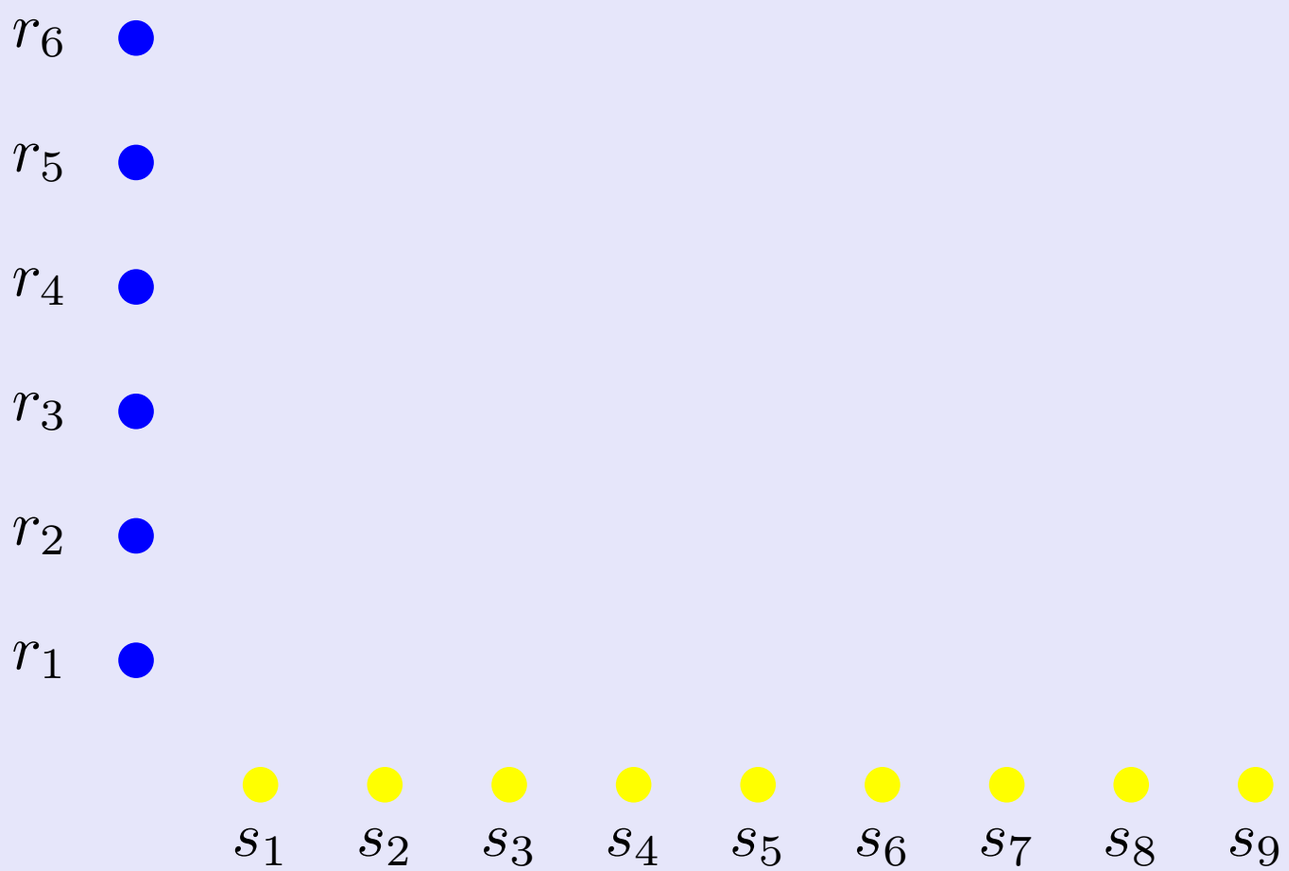
Programmazione Dinamica per DTW

I cammini possono entrare solo da alcuni stati. Se ho risolto quelli posso facilmente scegliere il migliore.

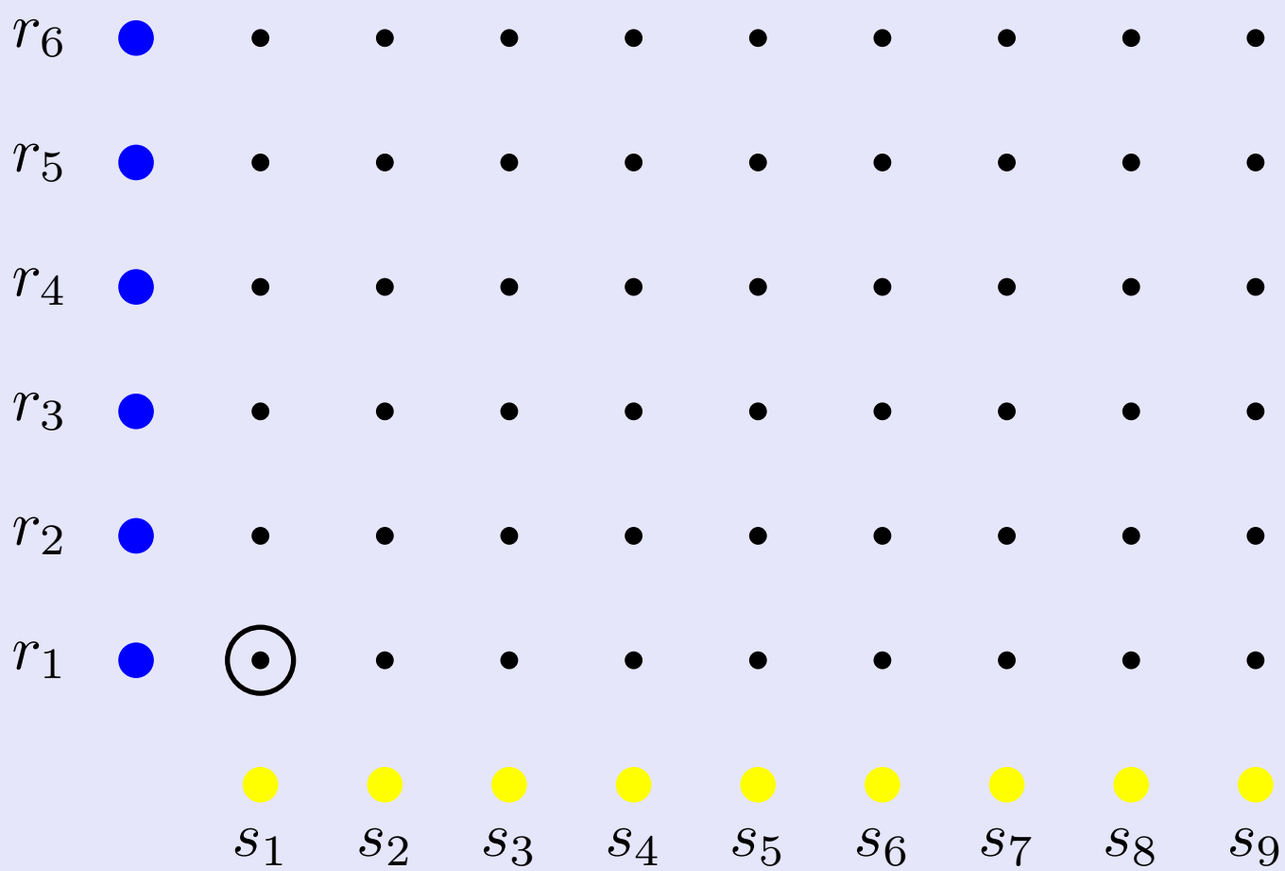


Il problema è ricondotto ad un problema elementare. Basterà quindi eseguire i calcoli nell'ordine opportuno.

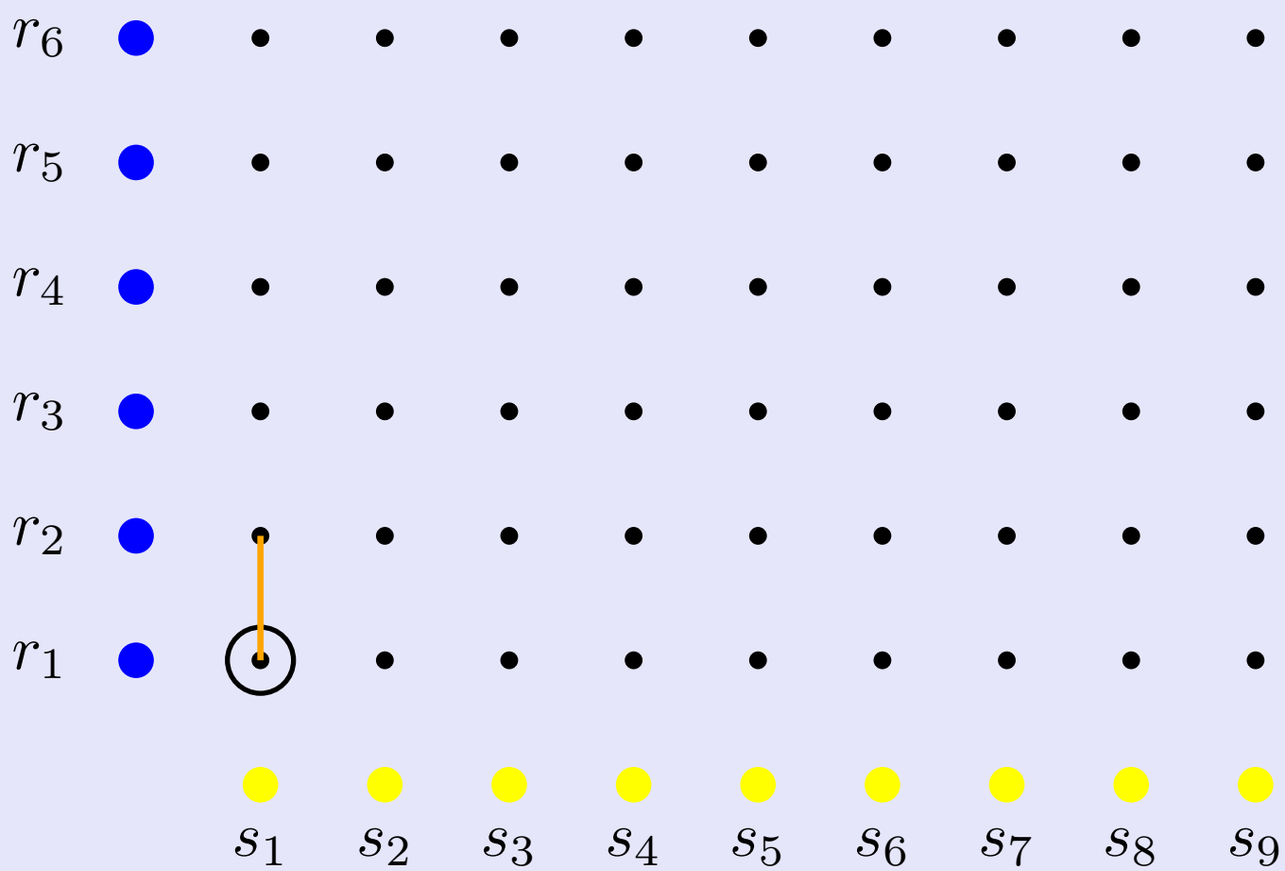
Programmazione Dinamica per DTW



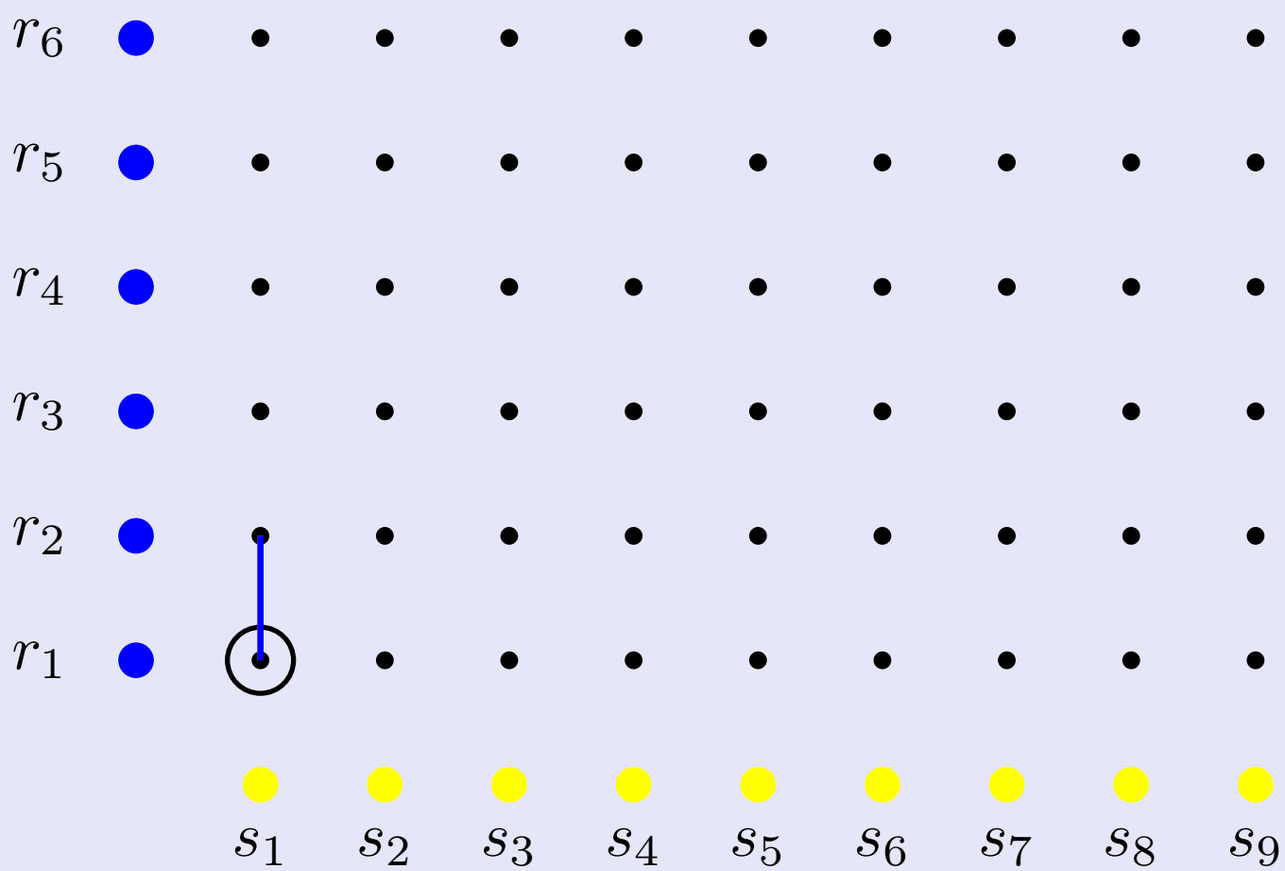
Programmazione Dinamica per DTW



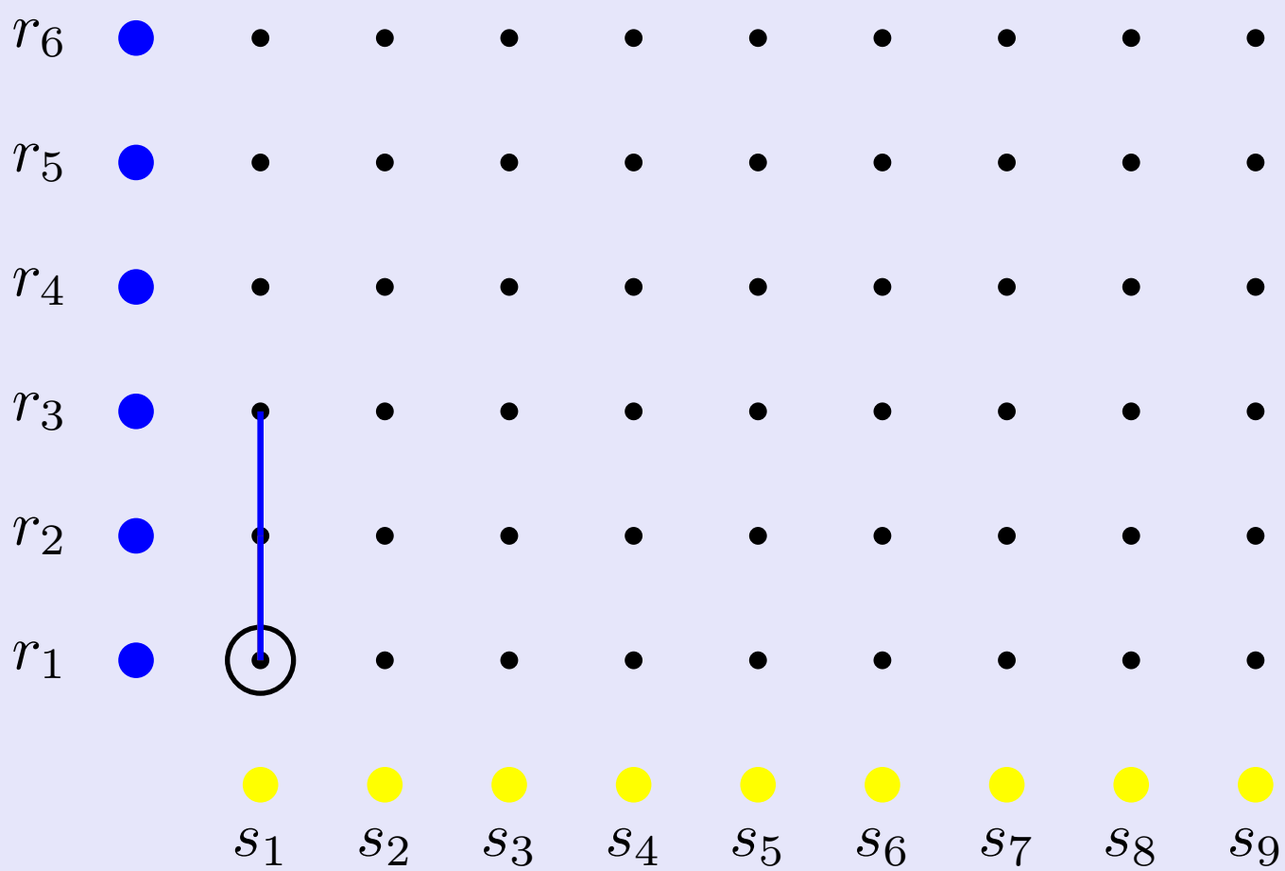
Programmazione Dinamica per DTW



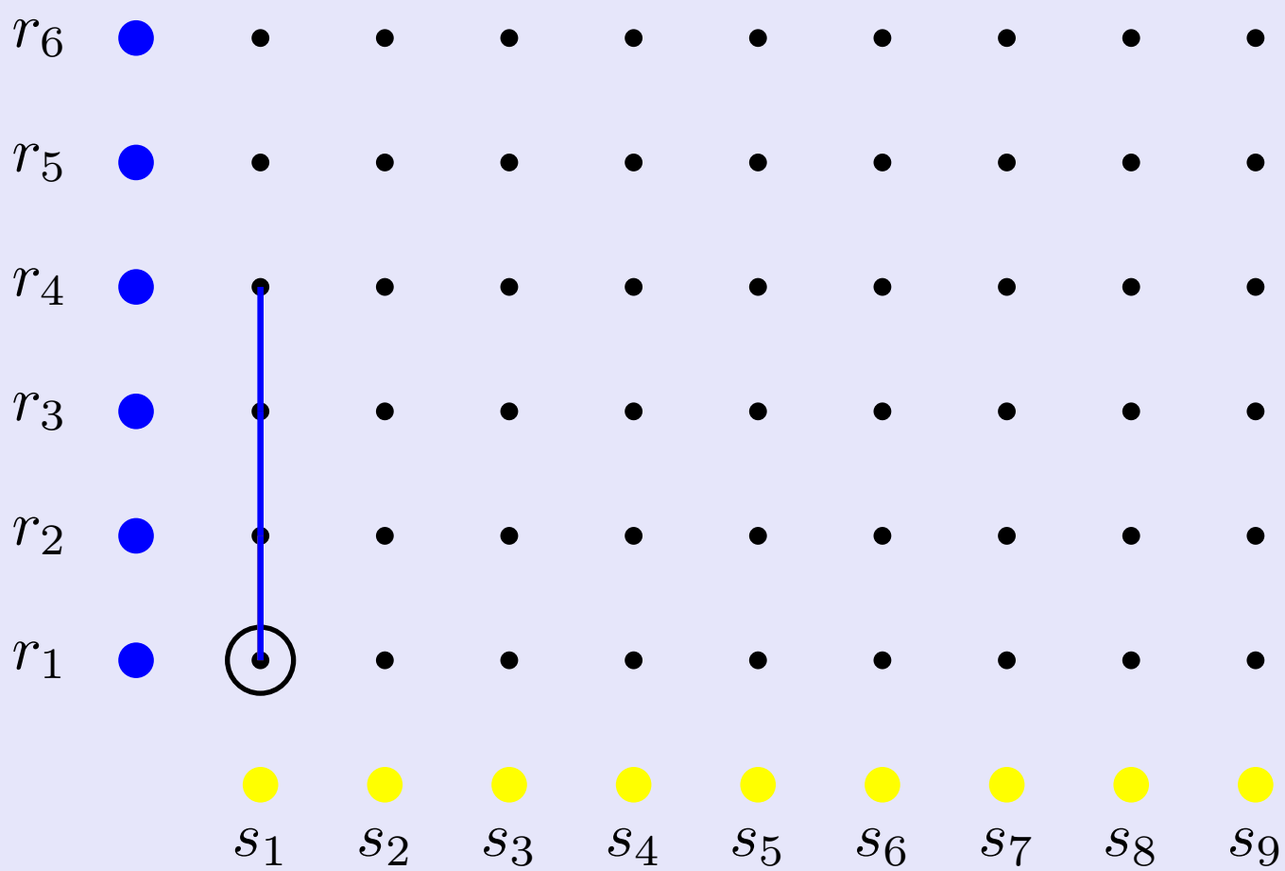
Programmazione Dinamica per DTW



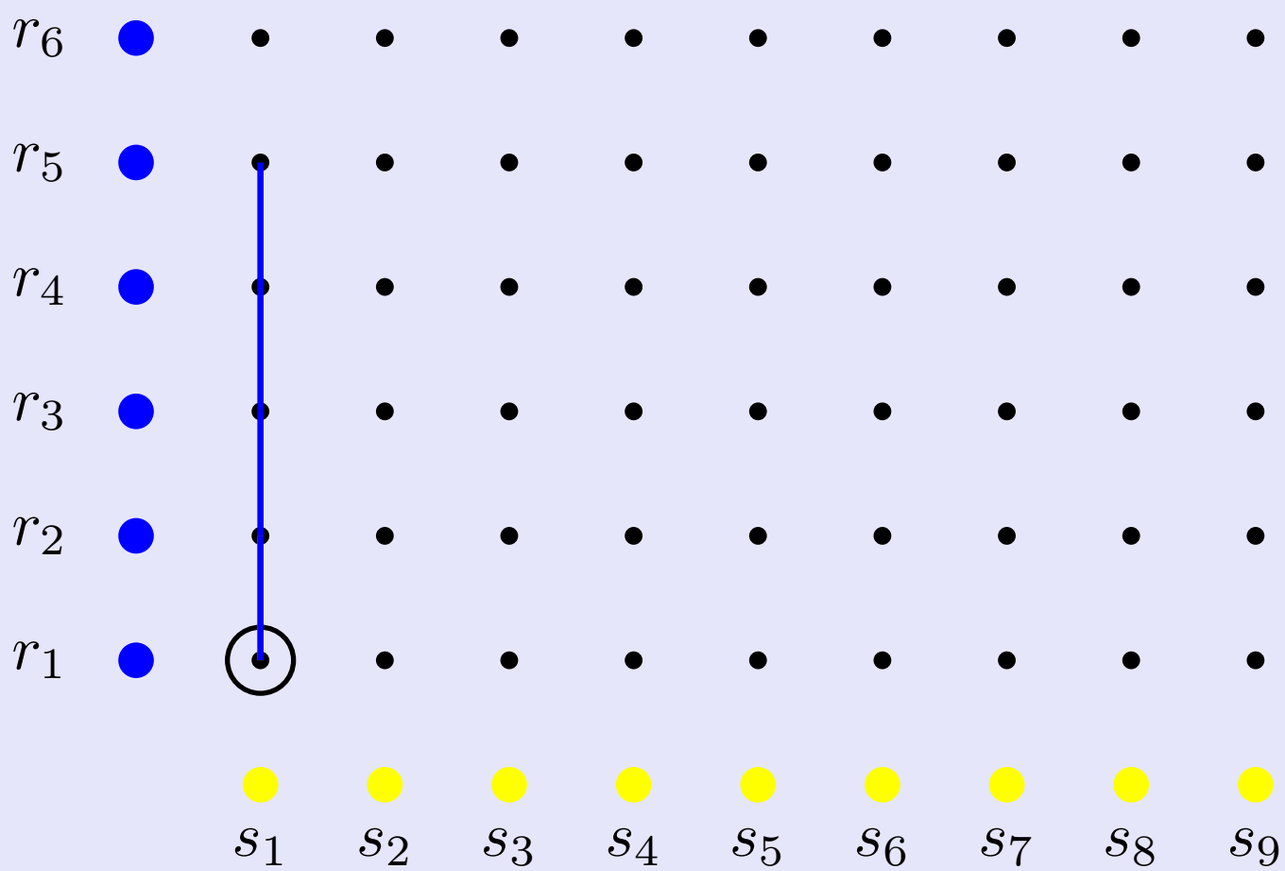
Programmazione Dinamica per DTW



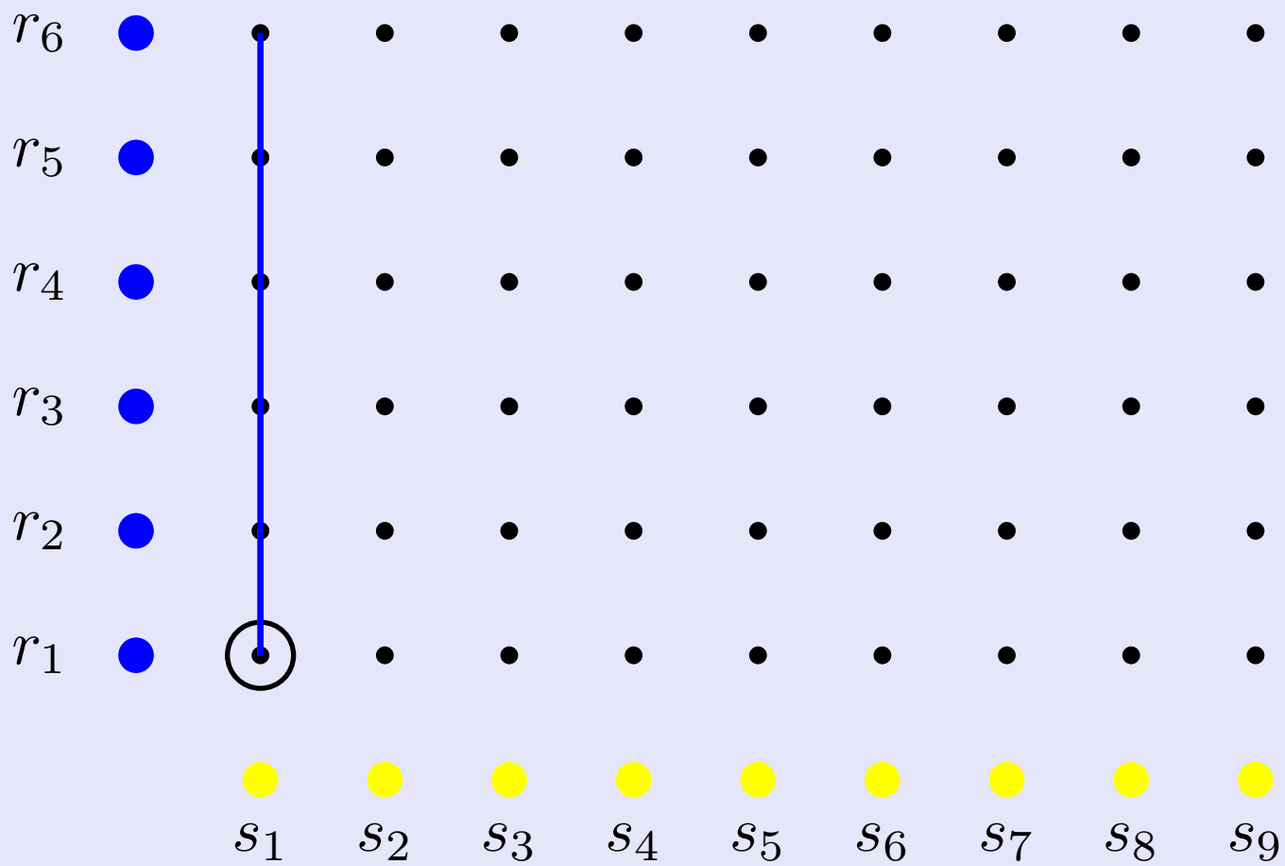
Programmazione Dinamica per DTW



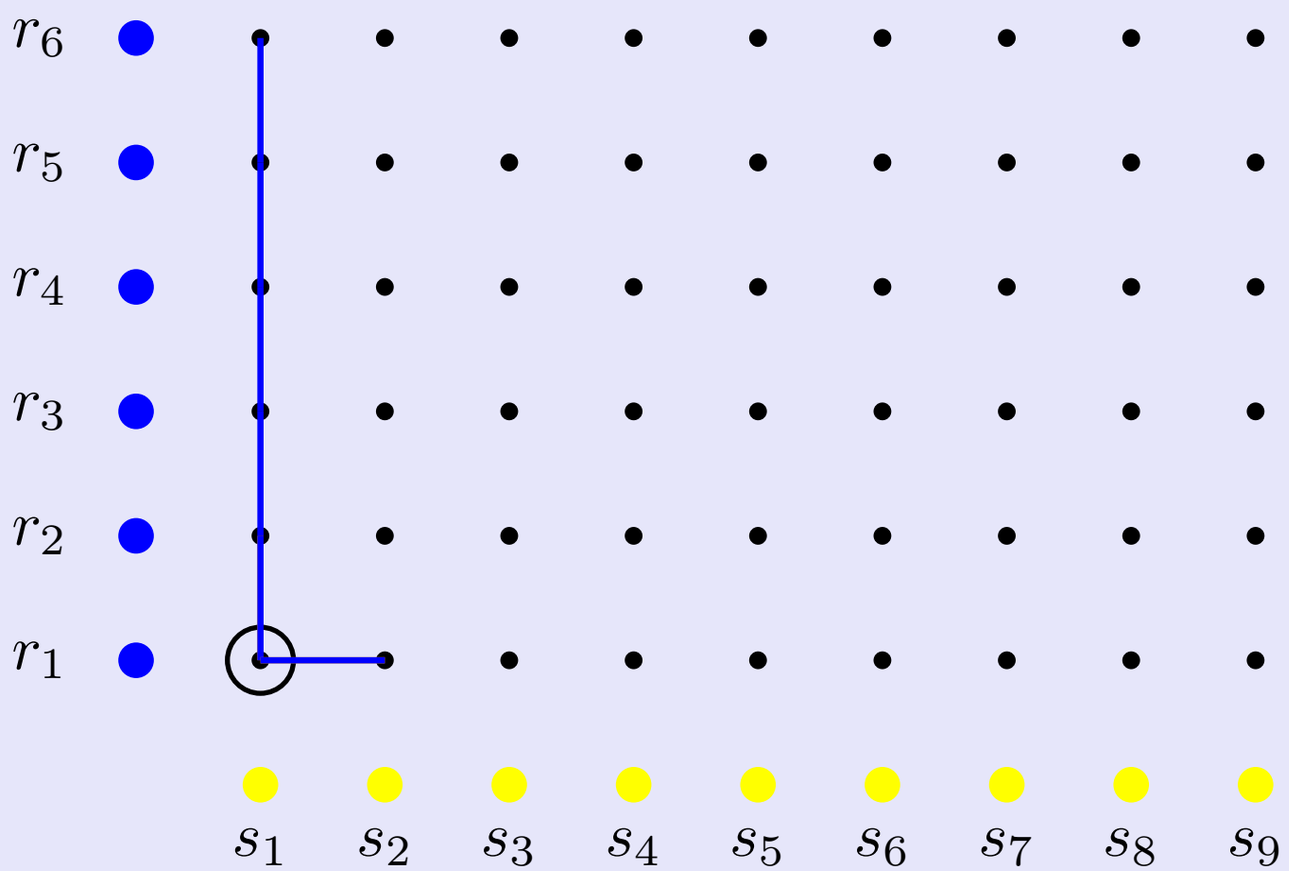
Programmazione Dinamica per DTW



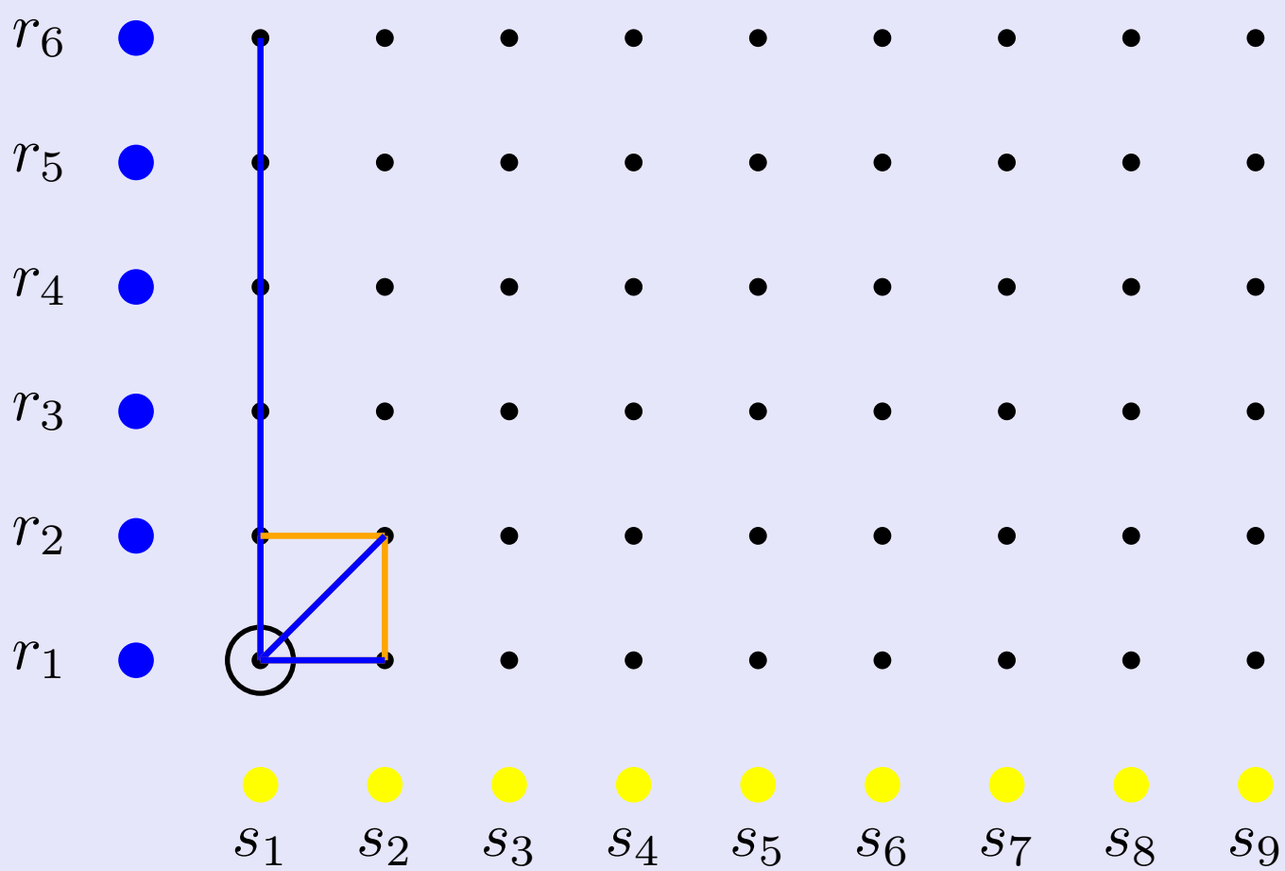
Programmazione Dinamica per DTW



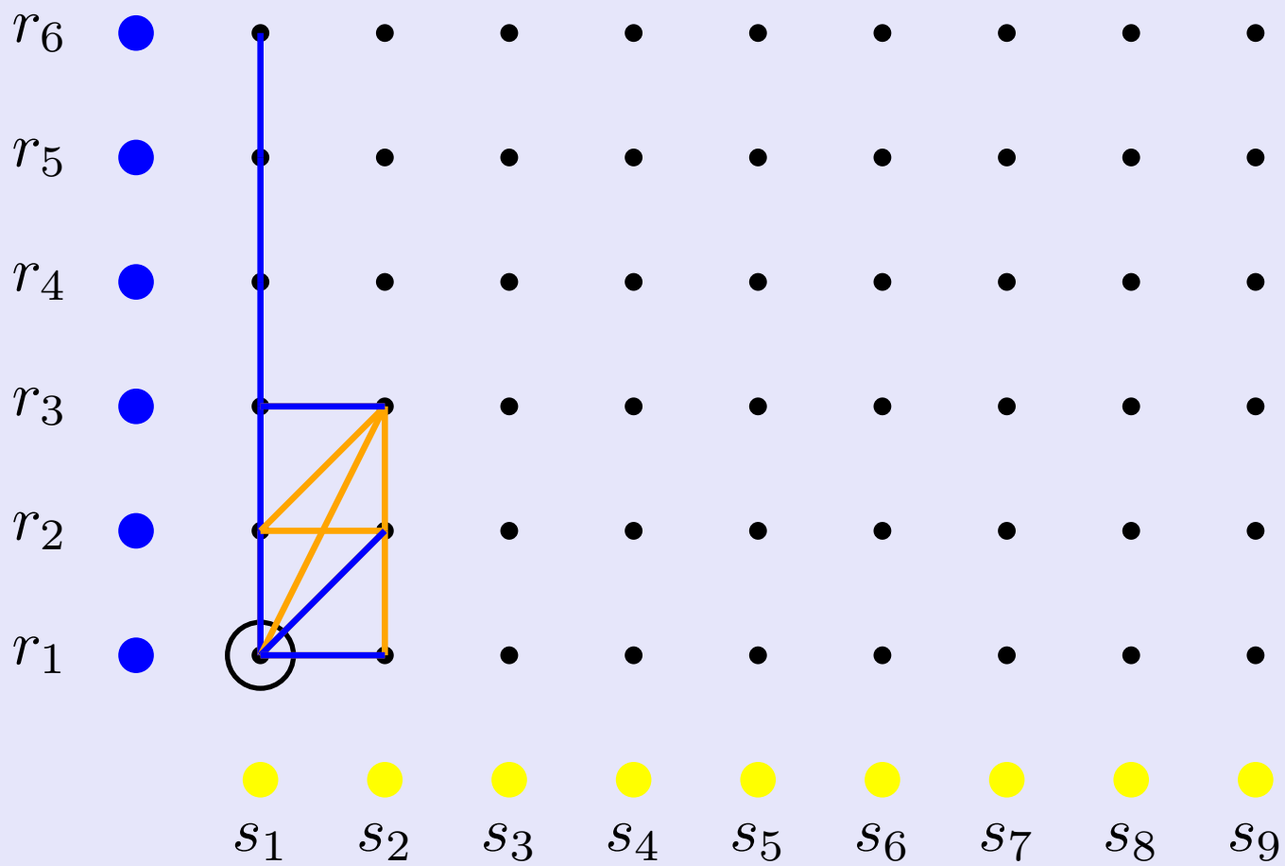
Programmazione Dinamica per DTW



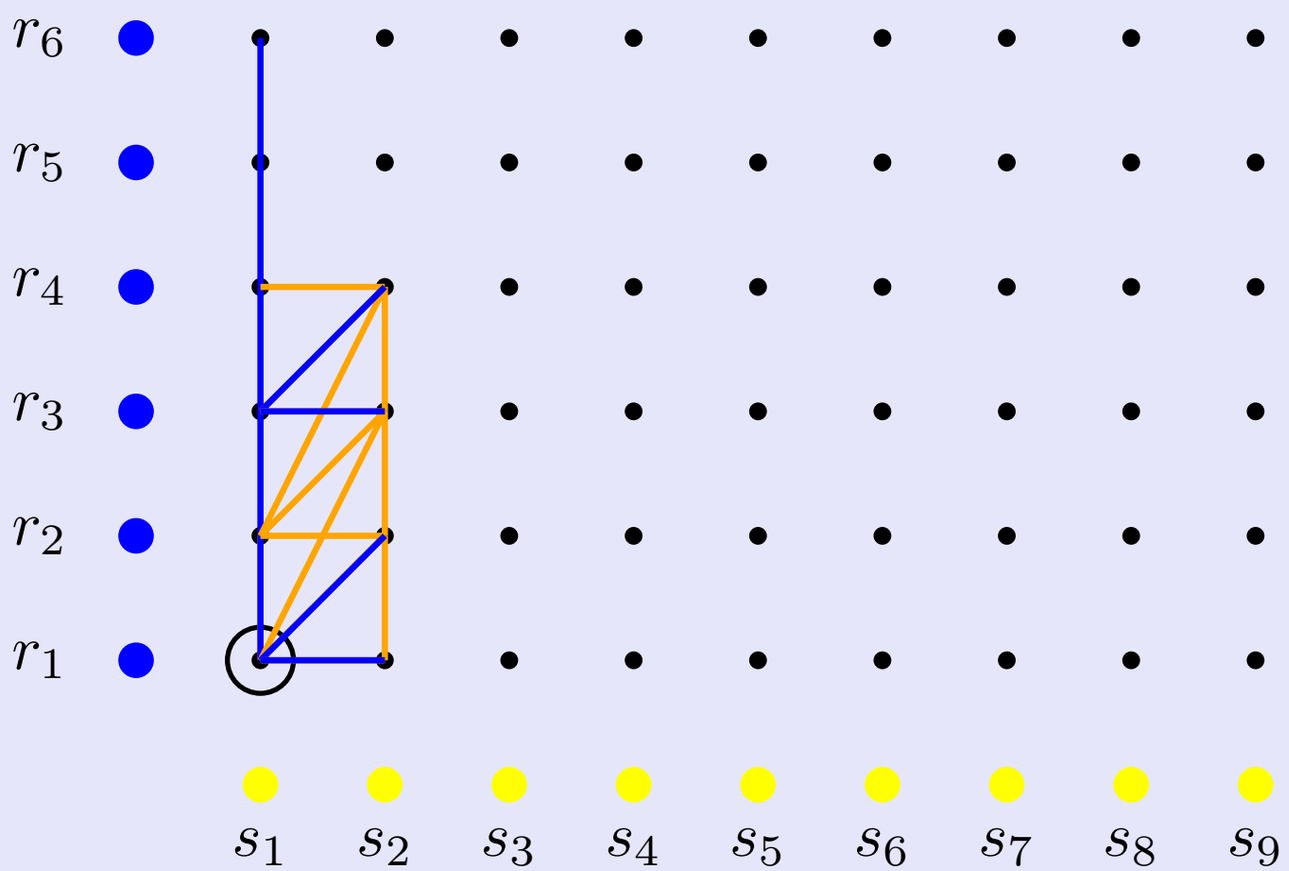
Programmazione Dinamica per DTW



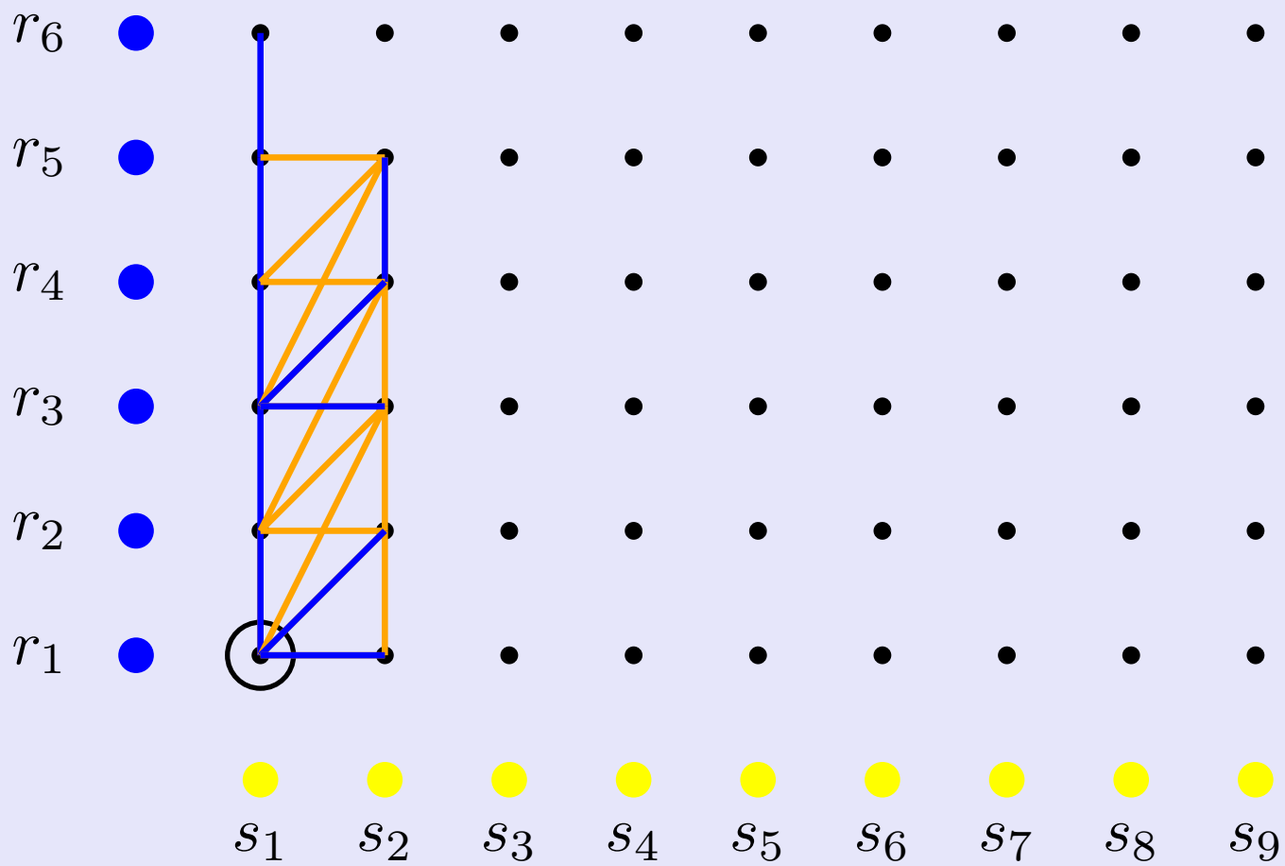
Programmazione Dinamica per DTW



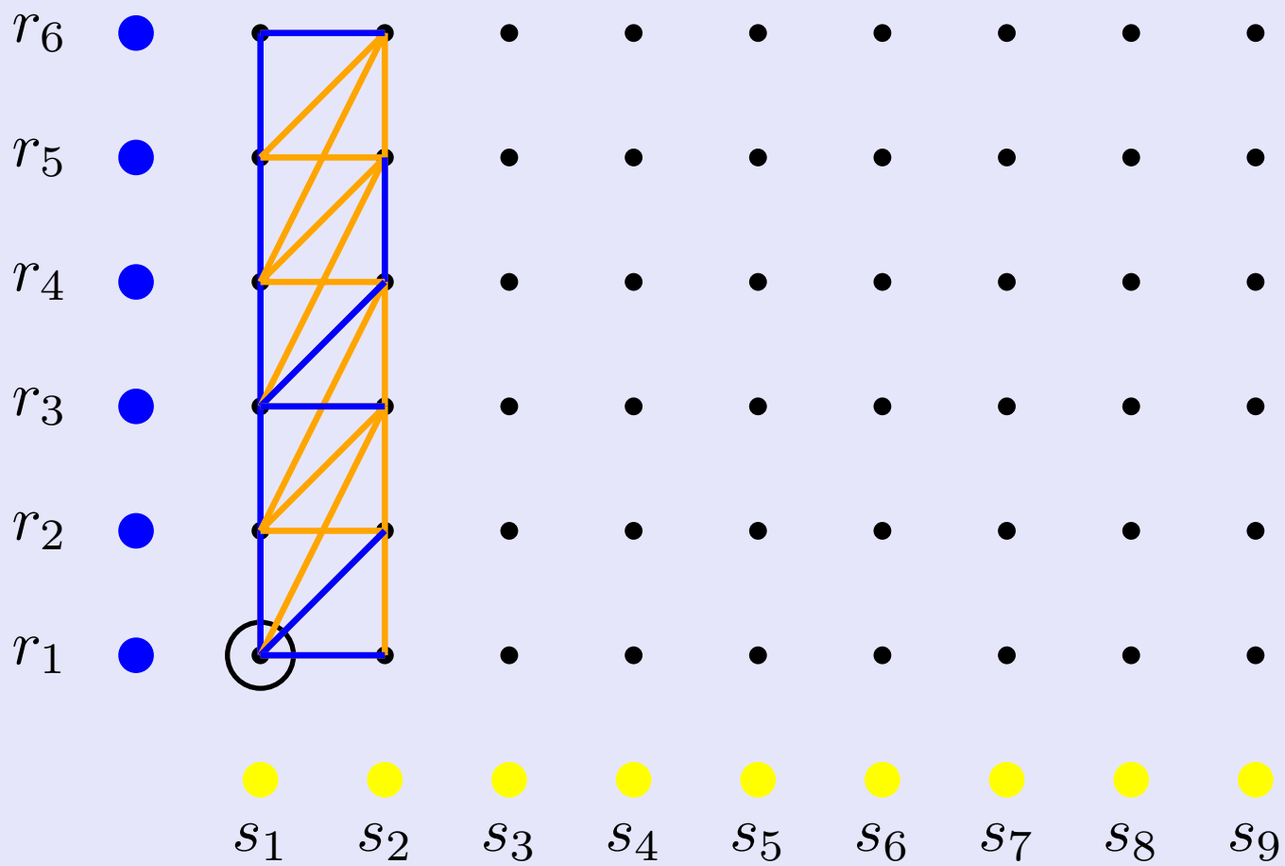
Programmazione Dinamica per DTW



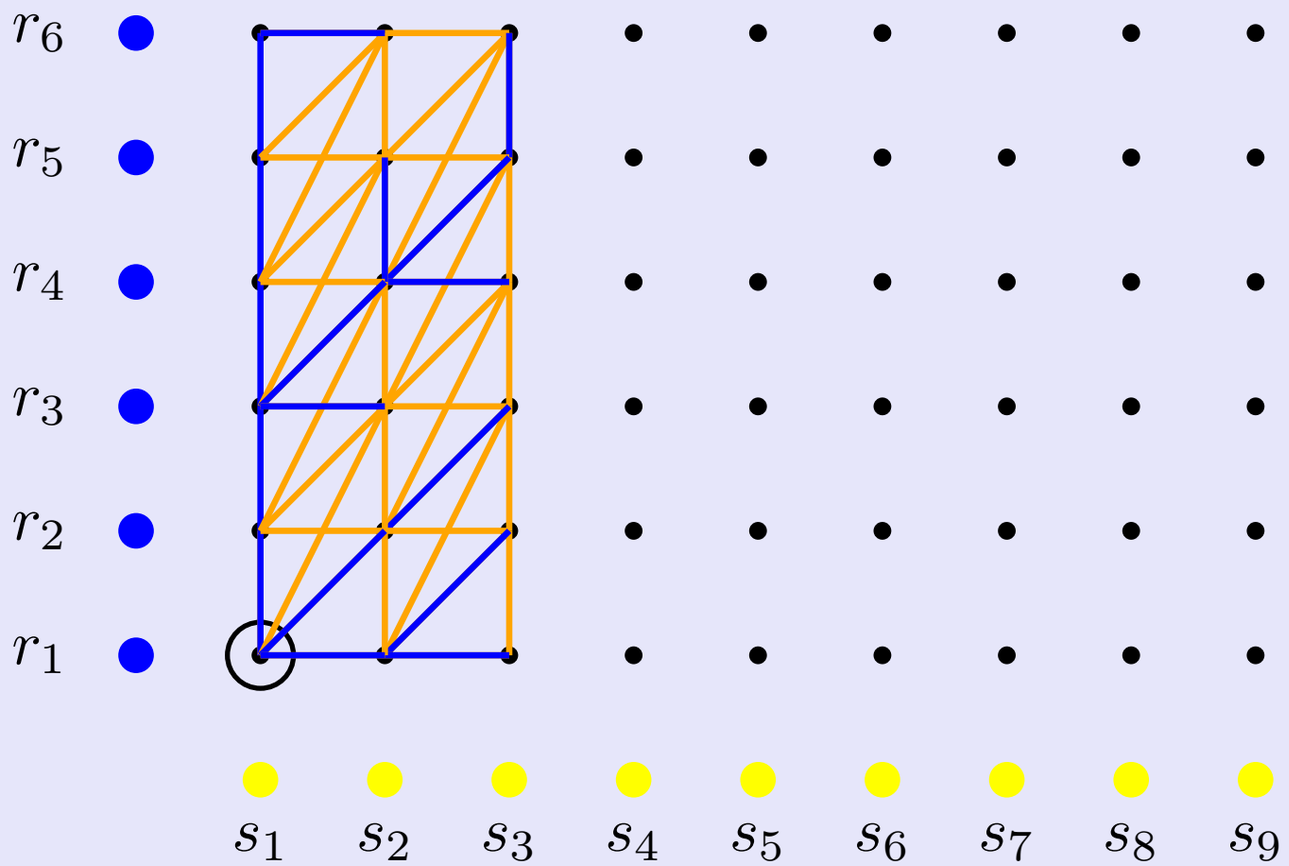
Programmazione Dinamica per DTW



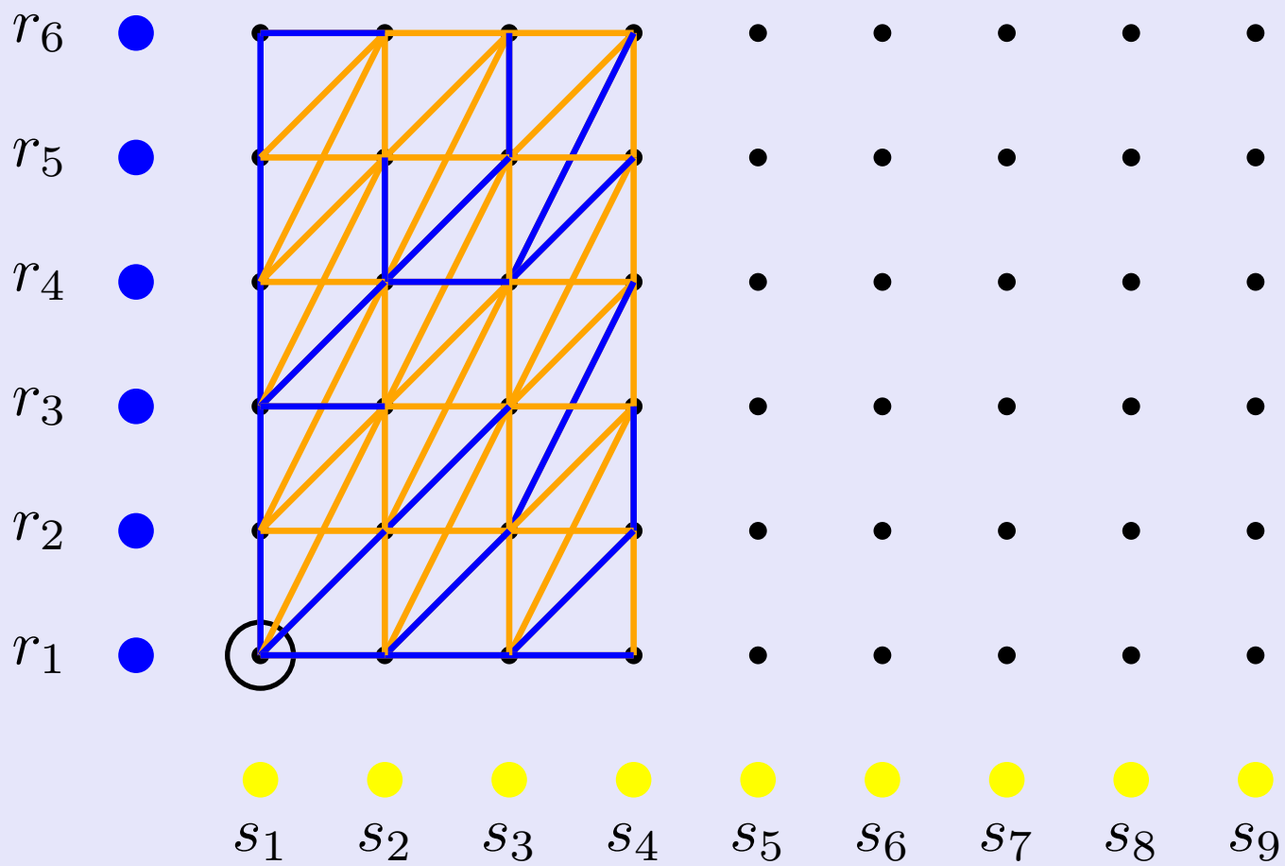
Programmazione Dinamica per DTW



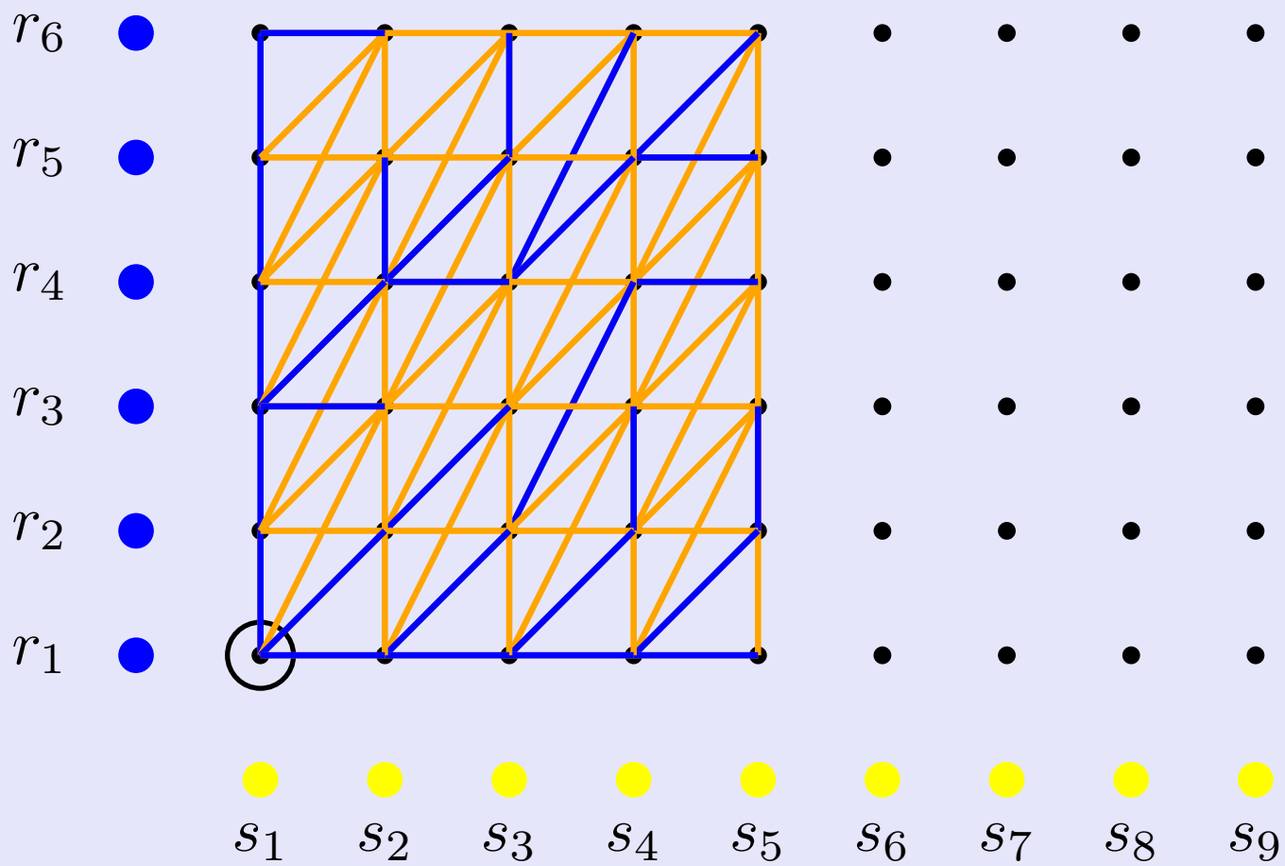
Programmazione Dinamica per DTW



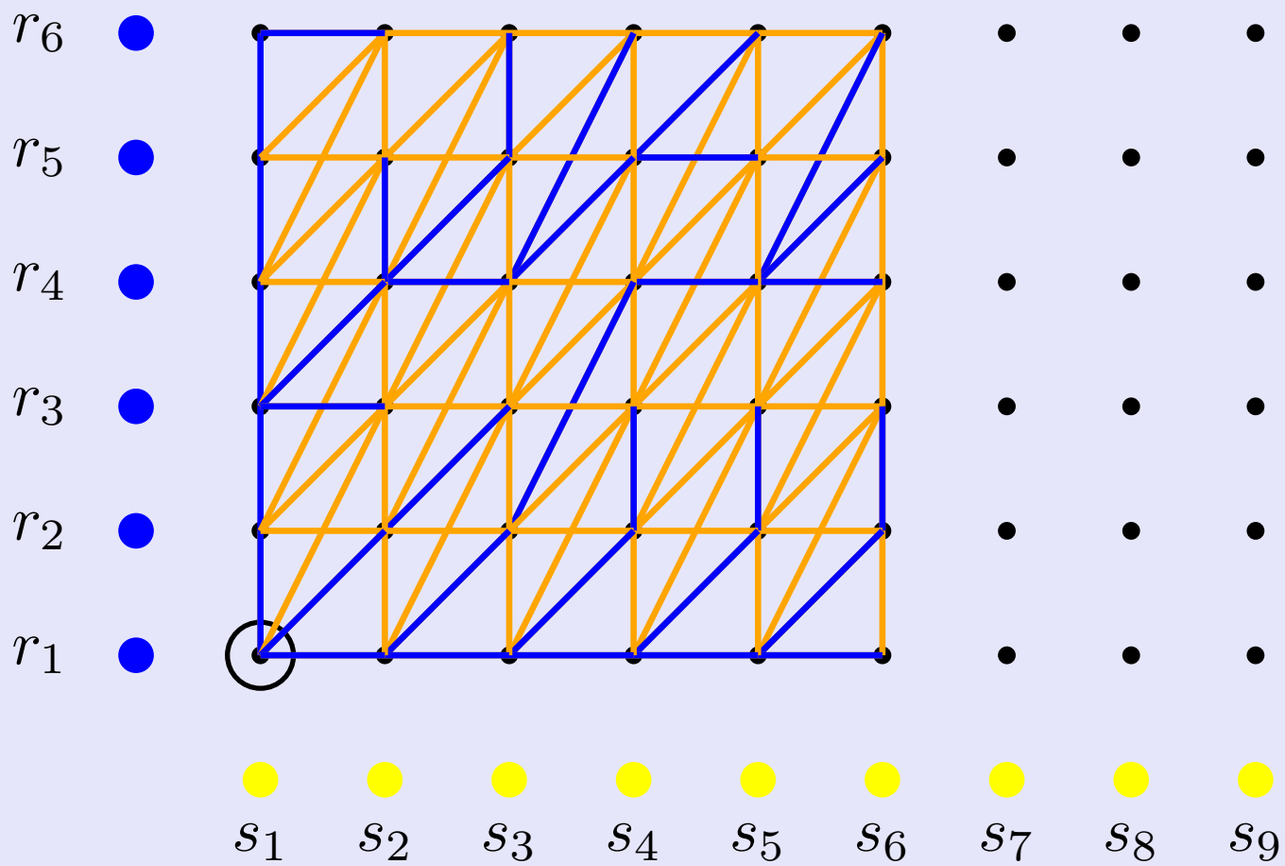
Programmazione Dinamica per DTW



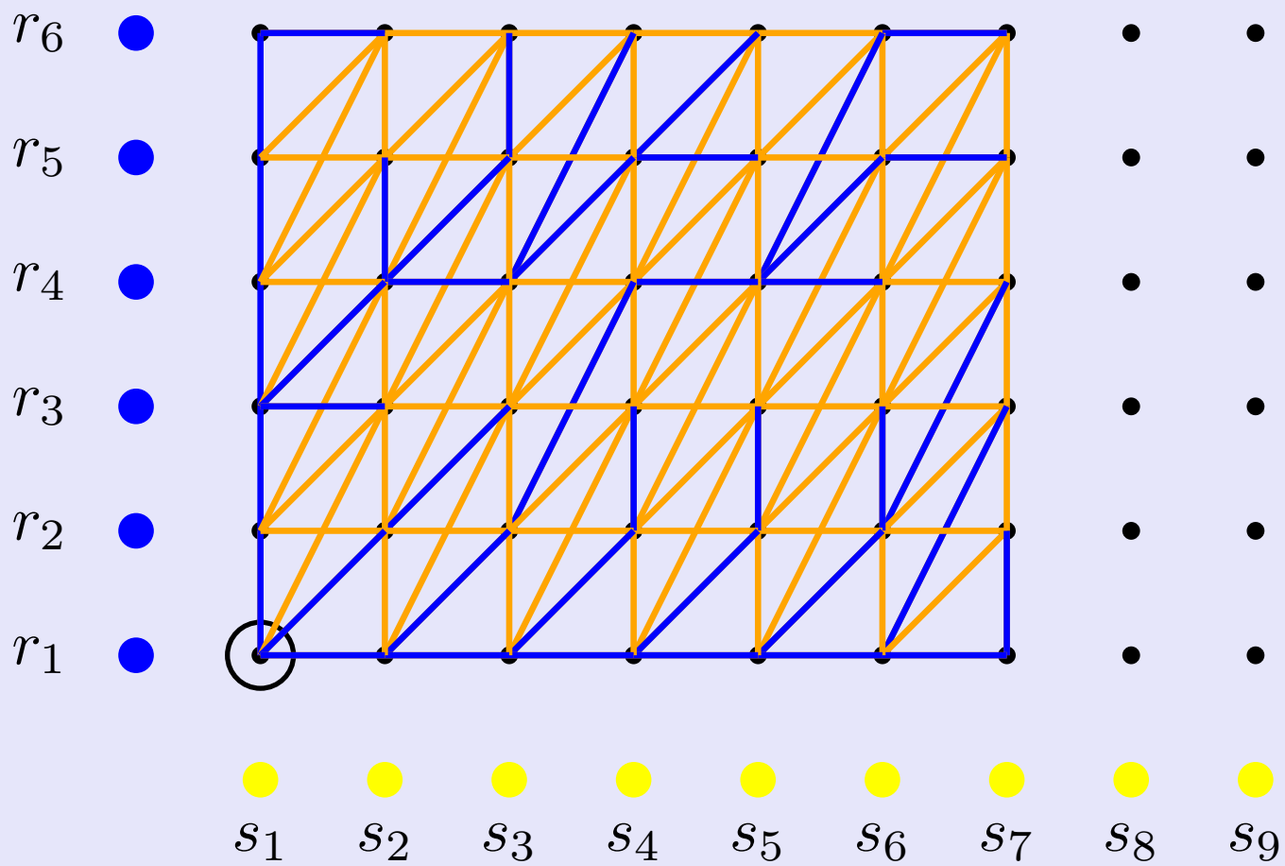
Programmazione Dinamica per DTW



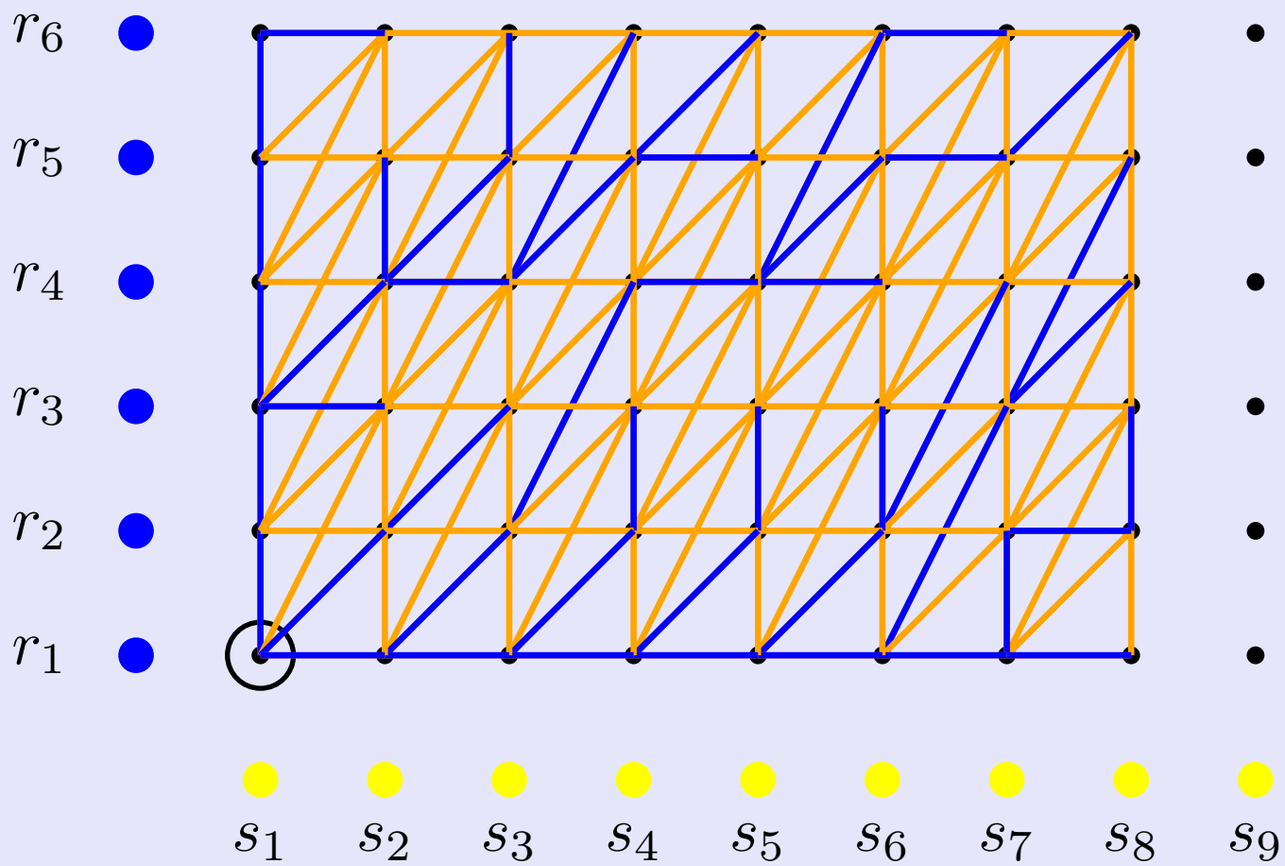
Programmazione Dinamica per DTW



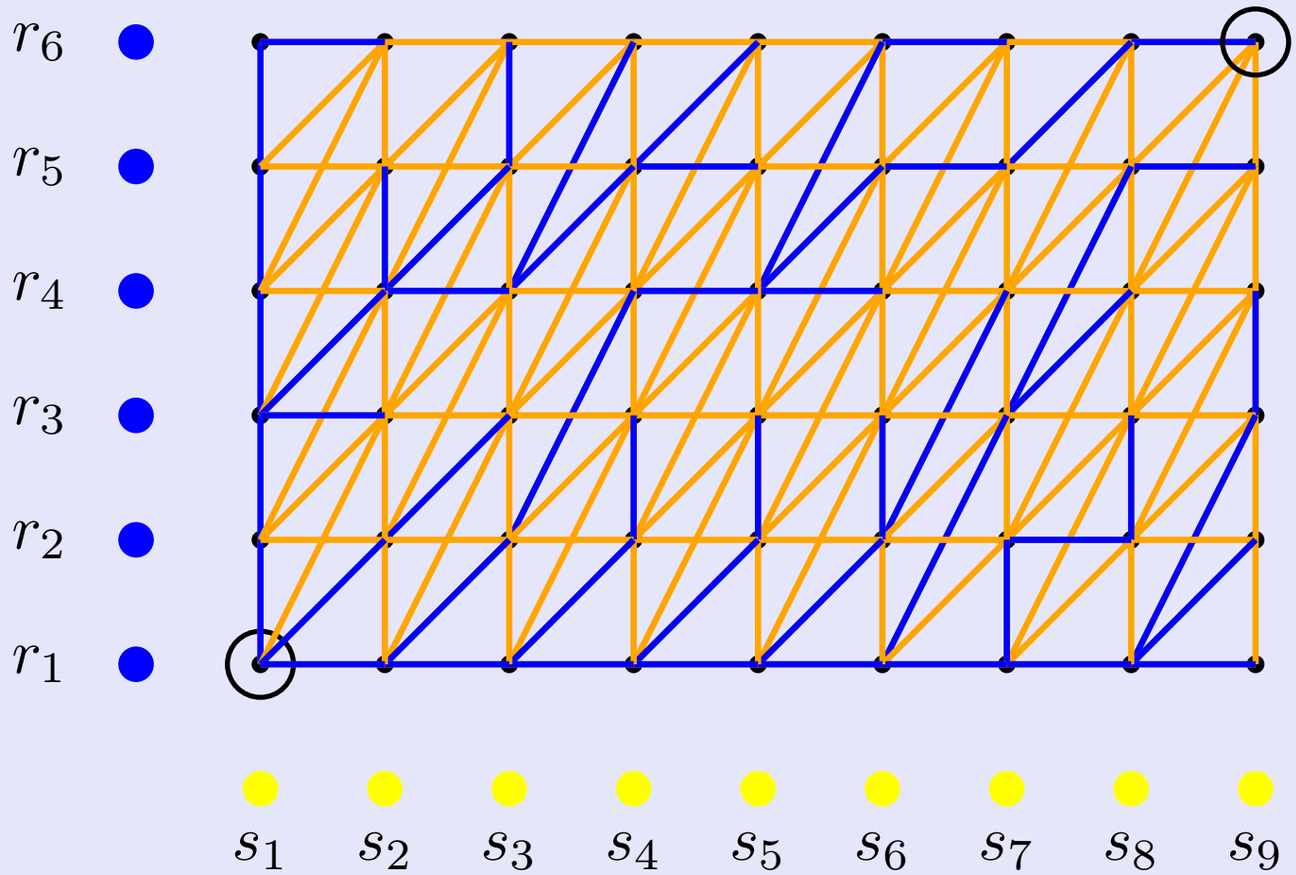
Programmazione Dinamica per DTW



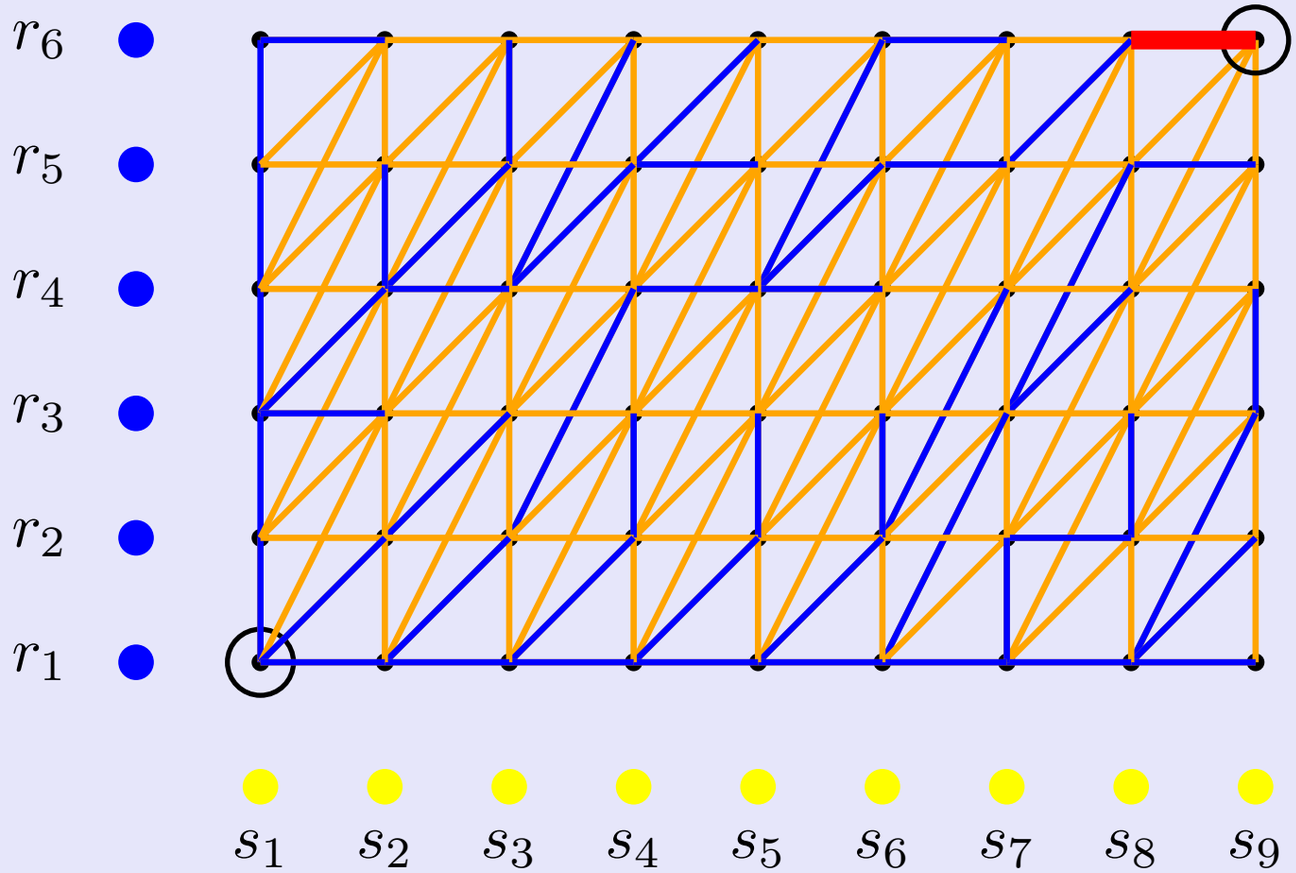
Programmazione Dinamica per DTW



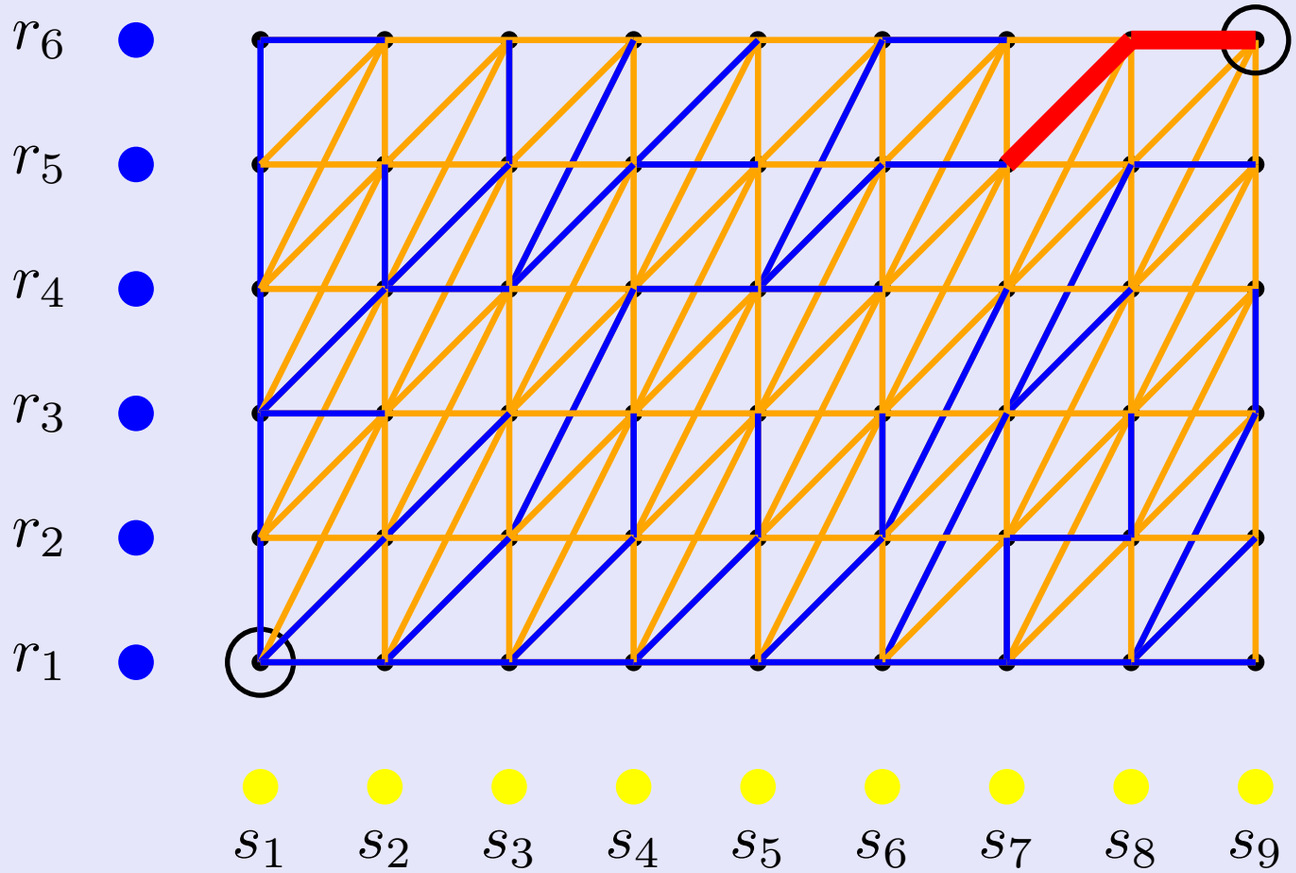
Programmazione Dinamica per DTW



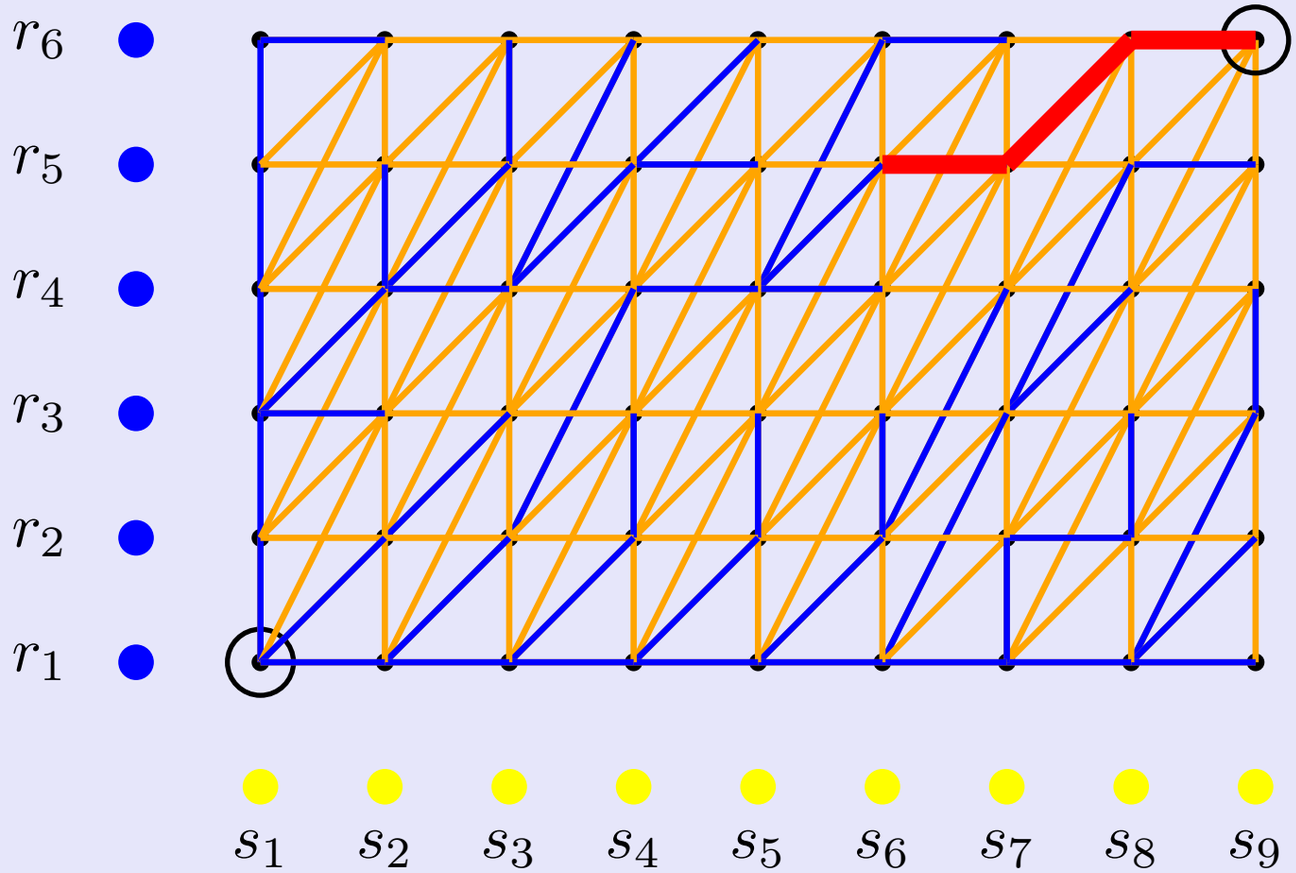
Programmazione Dinamica per DTW



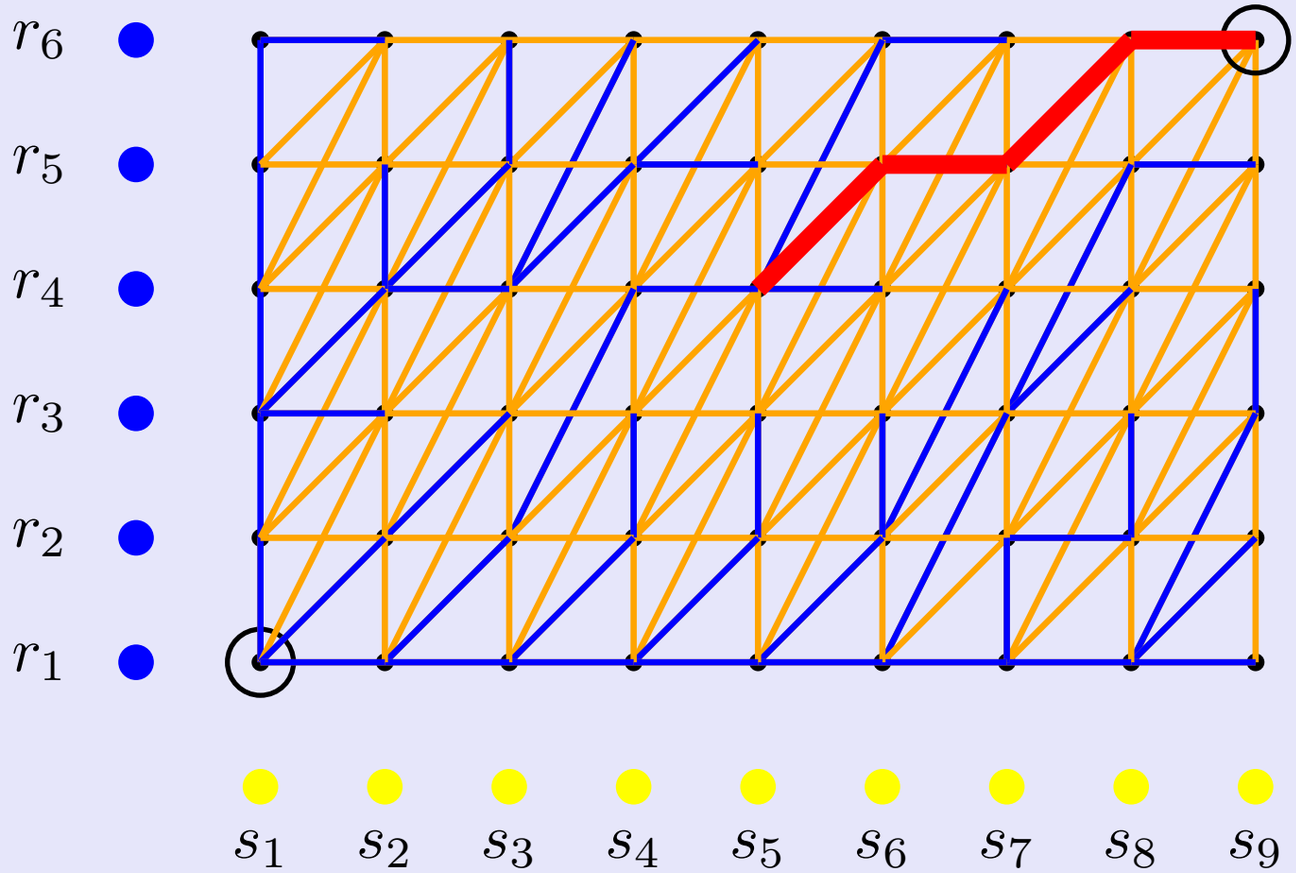
Programmazione Dinamica per DTW



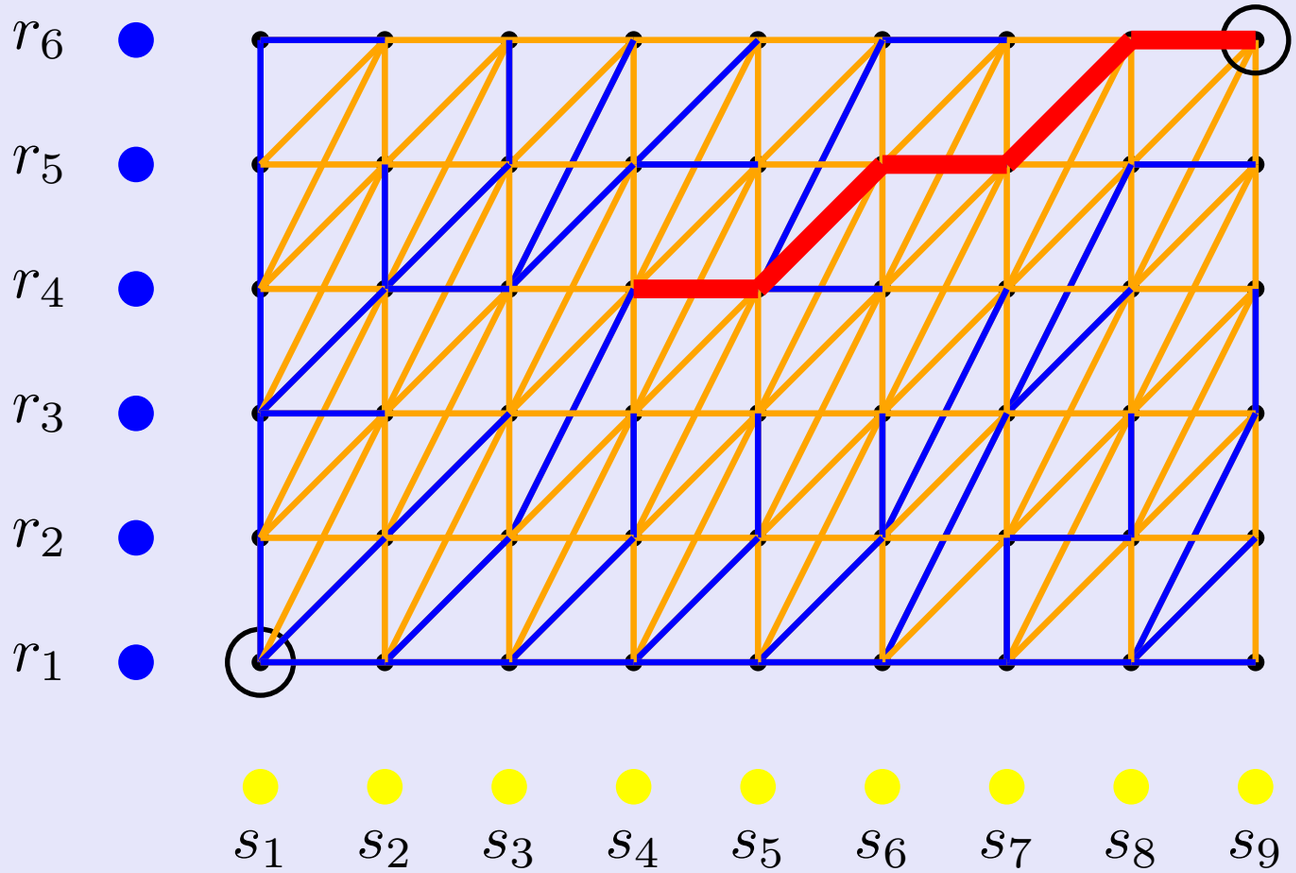
Programmazione Dinamica per DTW



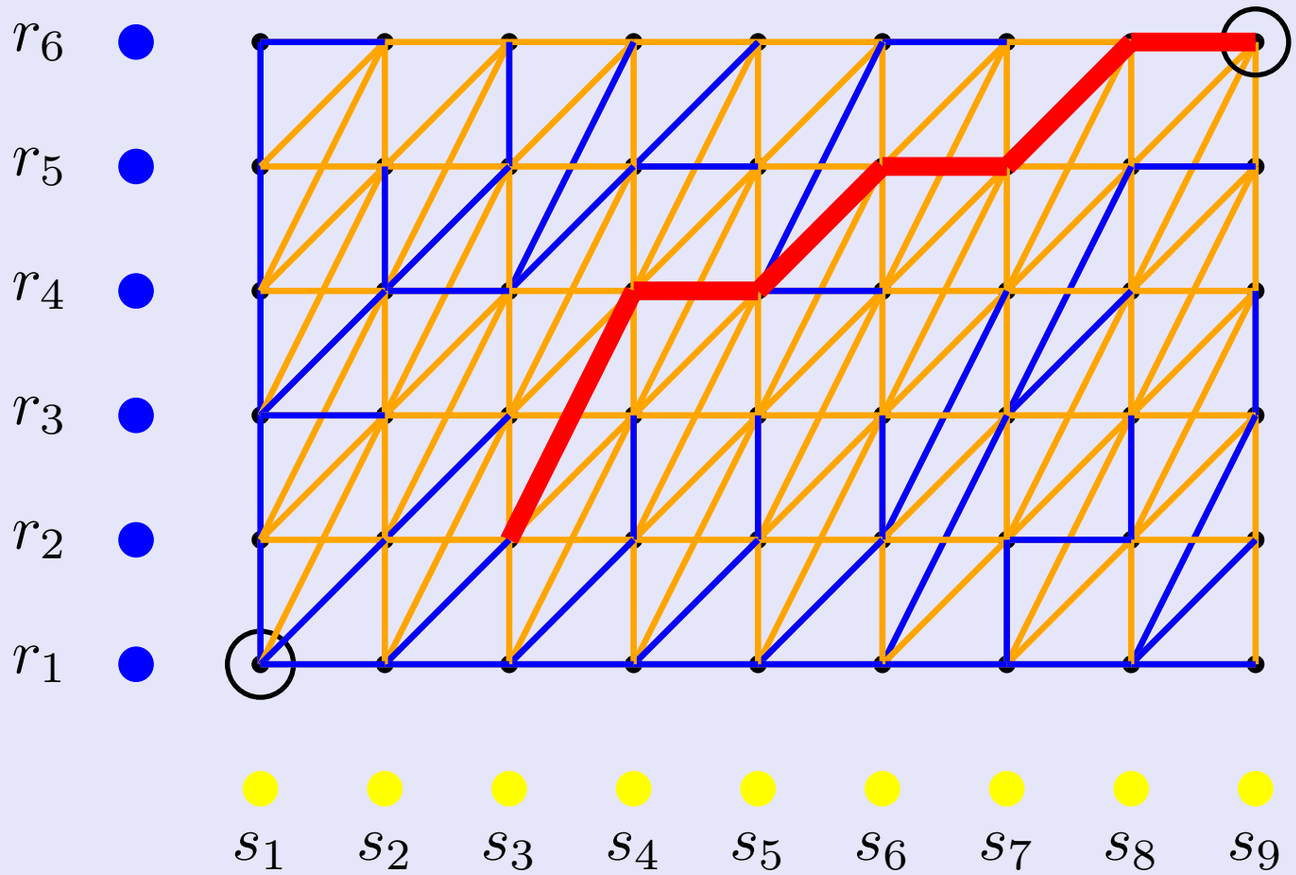
Programmazione Dinamica per DTW



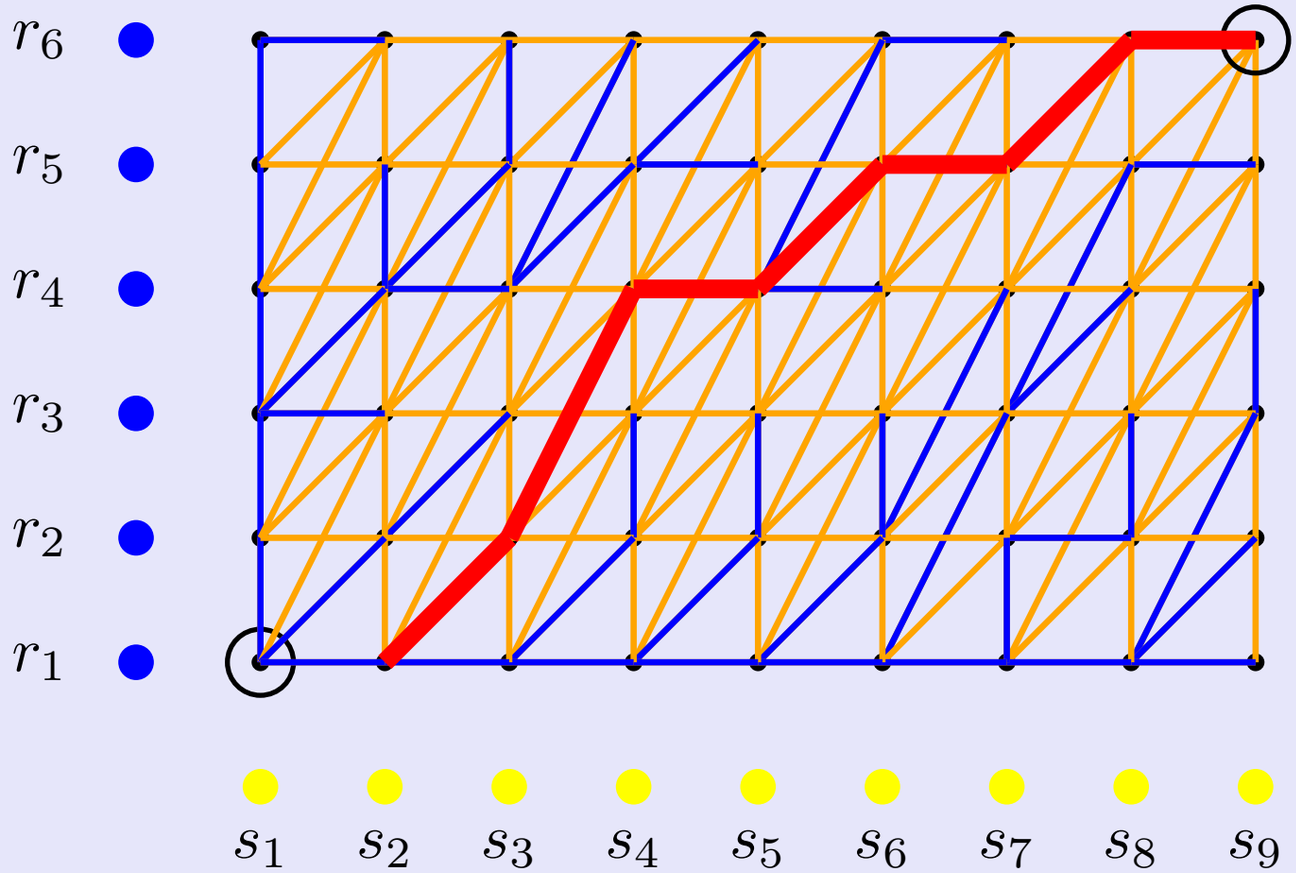
Programmazione Dinamica per DTW



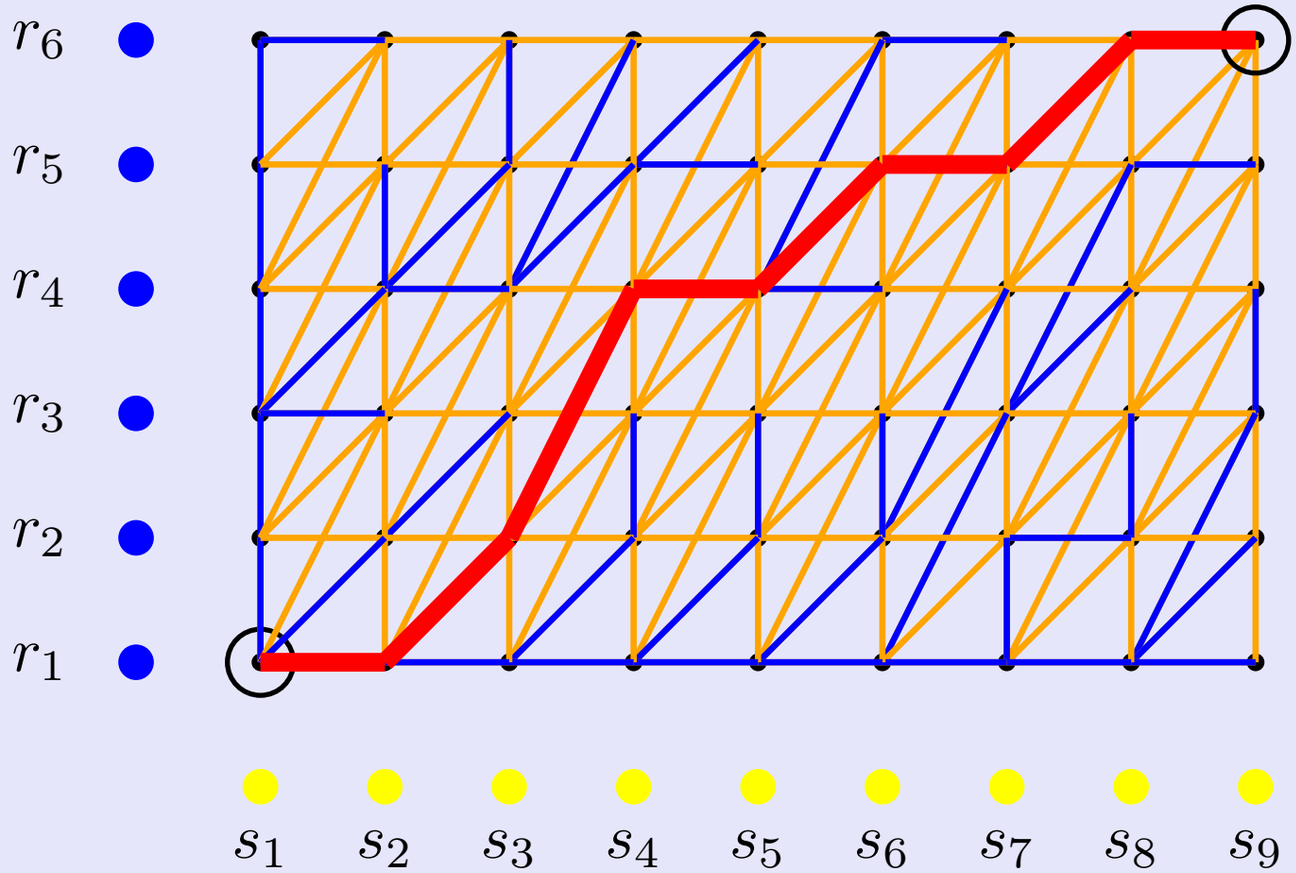
Programmazione Dinamica per DTW



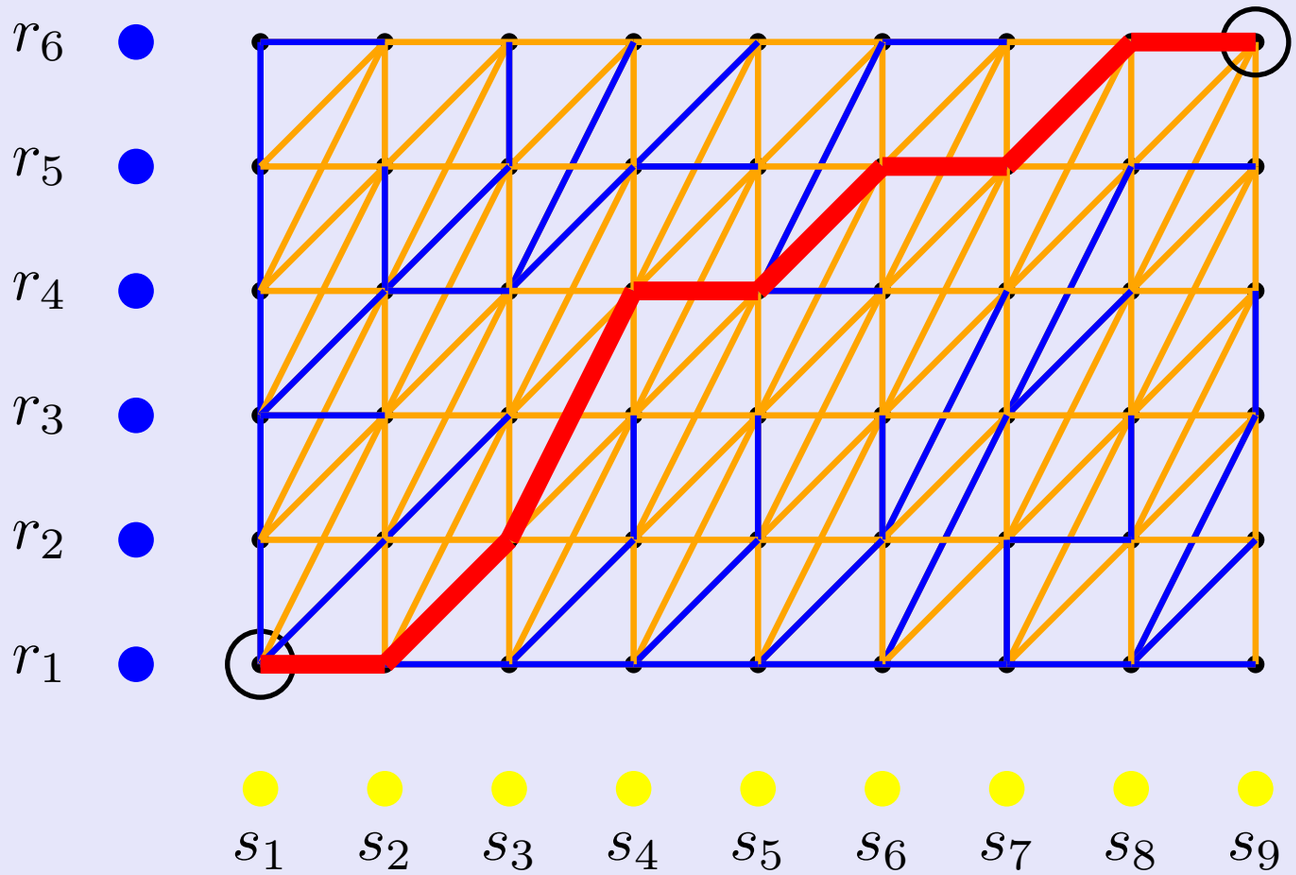
Programmazione Dinamica per DTW



Programmazione Dinamica per DTW

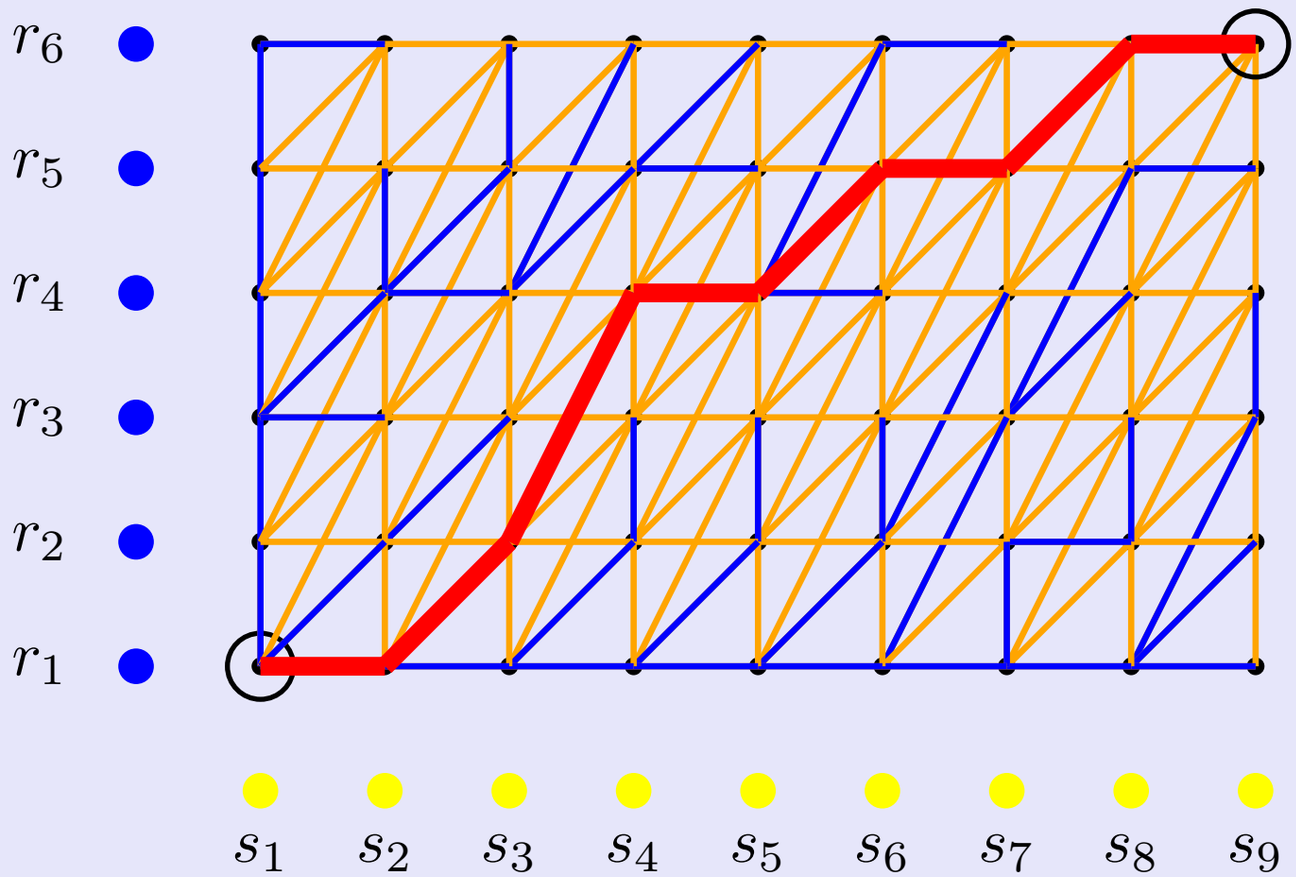


Programmazione Dinamica per DTW



Tempo di esecuzione: proporzionale al prodotto delle lunghezze

Programmazione Dinamica per DTW



Tempo di esecuzione: proporzionale al prodotto delle lunghezze

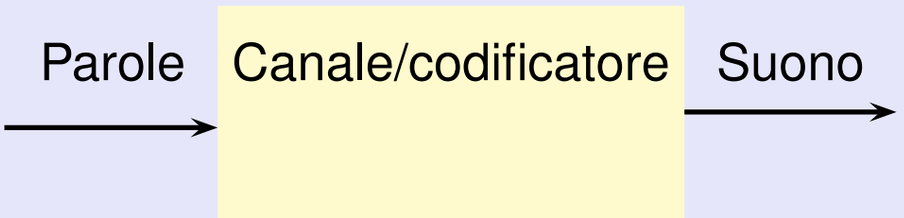
Time-synchronous: da sinistra a destra, Level-building: dal basso verso l'alto

Da DTW a modello statistico

- Il calcolo della distanza è una forma di ricerca su grafo
- Può essere generalizzato per riconoscere il parlato connesso
- Viene applicato con discreto successo p.e. alle sequenze di cifre
- Ma un esempio è troppo limitativo, se ne possono scegliere alcuni...
- Oppure costruirne uno che sia una sintesi di molti
- Ma la formalizzazione più consistente di una sintesi parametrica dai dati è un modello statistico

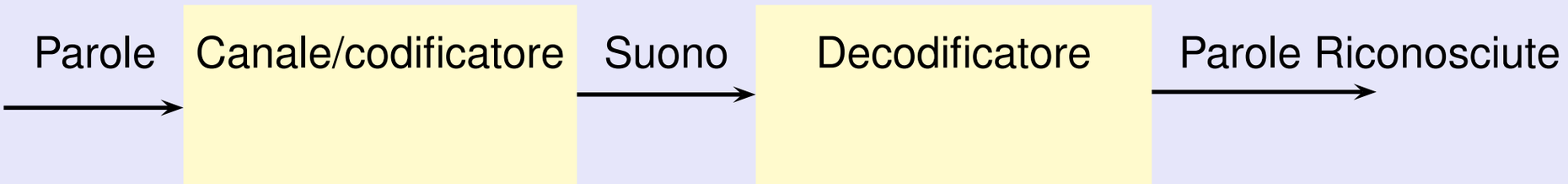
Approccio statistico: modello del canale rumoroso

Il messaggio originale passa attraverso un “canale” che lo trasforma in maniera aleatoria, ma non arbitraria



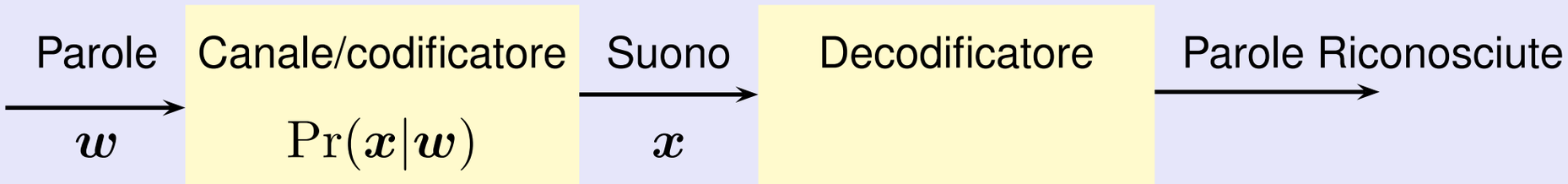
Approccio statistico: modello del canale rumoroso

Il messaggio originale passa attraverso un “canale” che lo trasforma in maniera aleatoria, ma non arbitraria



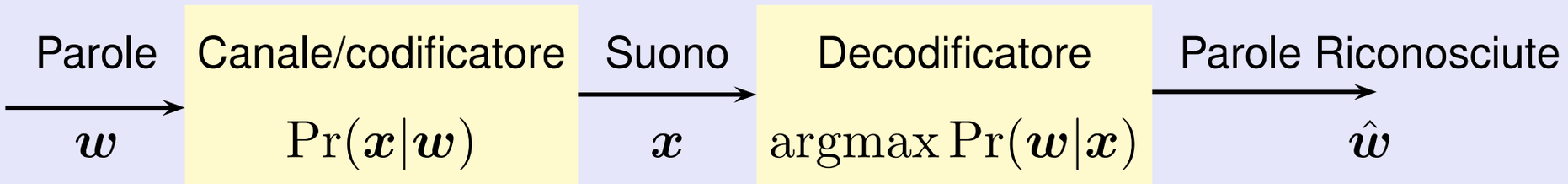
Approccio statistico: modello del canale rumoroso

Il messaggio originale passa attraverso un “canale” che lo trasforma in maniera aleatoria, ma non arbitraria



Approccio statistico: modello del canale rumoroso

Il messaggio originale passa attraverso un “canale” che lo trasforma in maniera aleatoria, ma non arbitraria



Bayes decision rule:

$$\hat{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{w}} \Pr(\boldsymbol{w}|\boldsymbol{x})$$

è ottimale rispetto al rischio di errore nell’ipotesi di avere modelli statistici esatti (situazione ovviamente ideale).

Approccio statistico al RAP

Data la rappresentazione parametrica del segnale, $\boldsymbol{x} = x_1 x_2 ... x_T$, la probabilità che sia stata pronunciata la sequenza di parole $\boldsymbol{w} = w_1 w_2 ... w_N$ viene calcolata come

$$\Pr[\boldsymbol{w}|\boldsymbol{x}] = \frac{\Pr[\boldsymbol{w}, \boldsymbol{x}]}{\Pr[\boldsymbol{x}]}$$

Approccio statistico al RAP

Data la rappresentazione parametrica del segnale, $\boldsymbol{x} = x_1 x_2 ... x_T$, la probabilità che sia stata pronunciata la sequenza di parole $\boldsymbol{w} = w_1 w_2 ... w_N$ viene calcolata come

$$\Pr[\boldsymbol{w}|\boldsymbol{x}] = \frac{\Pr[\boldsymbol{w}, \boldsymbol{x}]}{\Pr[\boldsymbol{x}]} \longleftarrow \text{Indipendente da } \boldsymbol{w}$$

Approccio statistico al RAP

Data la rappresentazione parametrica del segnale, $\boldsymbol{x} = x_1 x_2 ... x_T$, la probabilità che sia stata pronunciata la sequenza di parole $\boldsymbol{w} = w_1 w_2 ... w_N$ viene calcolata come

$$\Pr[\boldsymbol{w}|\boldsymbol{x}] = \frac{\Pr[\boldsymbol{w}, \boldsymbol{x}]}{\Pr[\boldsymbol{x}]} \longleftarrow \text{Indipendente da } \boldsymbol{w}$$

Scopo del riconoscitore è quindi quello di trovare la sequenza di parole $\hat{\boldsymbol{w}}$ in modo che $\Pr[\hat{\boldsymbol{w}}, \boldsymbol{x}]$ sia massima:

$$\begin{aligned} \hat{\boldsymbol{w}} &\stackrel{\Delta}{=} \underset{\boldsymbol{w}}{\operatorname{argmax}} \Pr[\boldsymbol{w}, \boldsymbol{x}] \\ &= \underset{\boldsymbol{w}}{\operatorname{argmax}} \Pr[\boldsymbol{w}] \Pr[\boldsymbol{x}|\boldsymbol{w}] \end{aligned}$$

Data la rappresentazione parametrica del segnale, $\boldsymbol{x} = x_1 x_2 ... x_T$, la probabilità che sia stata pronunciata la sequenza di parole $\boldsymbol{w} = w_1 w_2 ... w_N$ viene calcolata come

$$\Pr[\boldsymbol{w}|\boldsymbol{x}] = \frac{\Pr[\boldsymbol{w}, \boldsymbol{x}]}{\Pr[\boldsymbol{x}]}$$

Indipendente da $\boldsymbol{w}$

Scopo del riconoscitore è quindi quello di trovare la sequenza di parole $\hat{\boldsymbol{w}}$ in modo che $\Pr[\hat{\boldsymbol{w}}, \boldsymbol{x}]$ sia massima:

$$\begin{aligned} \hat{\boldsymbol{w}} &\stackrel{\Delta}{=} \operatorname*{argmax}_{\boldsymbol{w}} \Pr[\boldsymbol{w}, \boldsymbol{x}] \\ &= \operatorname*{argmax}_{\boldsymbol{w}} \Pr[\boldsymbol{w}] \Pr[\boldsymbol{x}|\boldsymbol{w}] \end{aligned}$$

Modello del linguaggio

Modello acustico

Approccio Statistico al RAP

- Un modello parametrico ad una sequenza assegna una probabilità, non una distanza
- Deve permettere di integrare efficacemente i modelli acustico e linguistico
- Il modello acustico deve avere flessibilità sia sullo spazio acustico che nel tempo
- ... e deve catturare dipendenze temporali
- Entrambi devono poter essere in grado di raccogliere informazione da grandi quantità di esempi, anche eterogenei

Il Modello Acustico: Hidden Markov Model

- Introdotto negli anni '70, largamente prevalente dagli anni '80
- Un automa probabilistico, composto da stati e transizioni, che genera osservazioni evolvendo nel tempo
- Ad ogni istante di tempo t:
  - Transita da uno stato i ad uno stato j seguendo una *p.d.f.* che dipende solo da i
  - Emette un simbolo x con una *p.d.f.* che dipende solo dalla transizione $i \rightarrow j$
- Ad ogni diversa sequenza di stati (o cammino) corrisponde quindi una diversa *p.d.f.* congiunta sulla sequenza di simboli osservabili
- Le sequenze osservabili sono solo i simboli emessi

Il Modello Acustico: Hidden Markov Model

Formalmente, un HMM è composto da una coppia $(\boldsymbol{S}, \boldsymbol{X})$ di processi aleatori dove:

$\boldsymbol{S}$ ha valori in un insieme finito $\mathcal{S}$ di *stati*, e non è osservabile

$\boldsymbol{X}$ ha valori nello spazio delle features acustiche $\mathcal{X}$, discreto o continuo.

Vincoli

$$i, j \in \mathcal{S}, x \in \mathcal{X}, t, T \in \mathbb{N}, t <= T$$

Markov:
$$\Pr[\boldsymbol{S}_t = i | \boldsymbol{S}_0^{t-1} = \boldsymbol{s}_0^{t-1}] = \Pr[\boldsymbol{S}_t = i | \boldsymbol{S}_{t-1} = s_{t-1}]$$

Output Independence:
$$\Pr[\boldsymbol{X}_t = x | \boldsymbol{X}_0^{t-1} = \boldsymbol{x}_0^{t-1}, \boldsymbol{S}_0^T = \boldsymbol{s}_0^T] = \Pr[\boldsymbol{X}_t = x | \boldsymbol{S}_{t-1}^t = \boldsymbol{s}_{t-1}^t]$$

Parametri

$$\begin{aligned} a_{ij} &\triangleq \Pr[\boldsymbol{S}_t = j | \boldsymbol{S}_{t-1} = i] \\ b_{ij}(x) &\triangleq \Pr[\boldsymbol{X}_t = x | \boldsymbol{S}_{t-1} = i, \boldsymbol{S}_t = j] \end{aligned}$$

Il Modello Acustico: Hidden Markov Model

Formalmente, un HMM è composto da una coppia $(\boldsymbol{S}, \boldsymbol{X})$ di processi aleatori dove:

$\boldsymbol{S}$ ha valori in un insieme finito $\mathcal{S}$ di *stati*, e non è osservabile

$\boldsymbol{X}$ ha valori nello spazio delle features acustiche $\mathcal{X}$, discreto o continuo.

Vincoli

$$i, j \in \mathcal{S}, x \in \mathcal{X}, t, T \in \mathbb{N}, t <= T$$

Markov:
$$\Pr[\boldsymbol{S}_t = i | \boldsymbol{S}_0^{t-1} = \boldsymbol{s}_0^{t-1}] = \Pr[\boldsymbol{S}_t = i | \boldsymbol{S}_{t-1} = s_{t-1}]$$

Output Independence:
$$\Pr[\boldsymbol{X}_t = x | \boldsymbol{X}_0^{t-1} = \boldsymbol{x}_0^{t-1}, \boldsymbol{S}_0^T = \boldsymbol{s}_0^T] = \Pr[\boldsymbol{X}_t = x | \boldsymbol{S}_{t-1}^t = \boldsymbol{s}_{t-1}^t]$$

Parametri

Probabilità di transizione

Probabilità di emissione

$$a_{ij} \triangleq \Pr[\boldsymbol{S}_t = j | \boldsymbol{S}_{t-1} = i]$$

$$b_{ij}(x) \triangleq \Pr[\boldsymbol{X}_t = x | \boldsymbol{S}_{t-1} = i, \boldsymbol{S}_t = j]$$

Tipi di densità sulle osservazioni

Discreta:

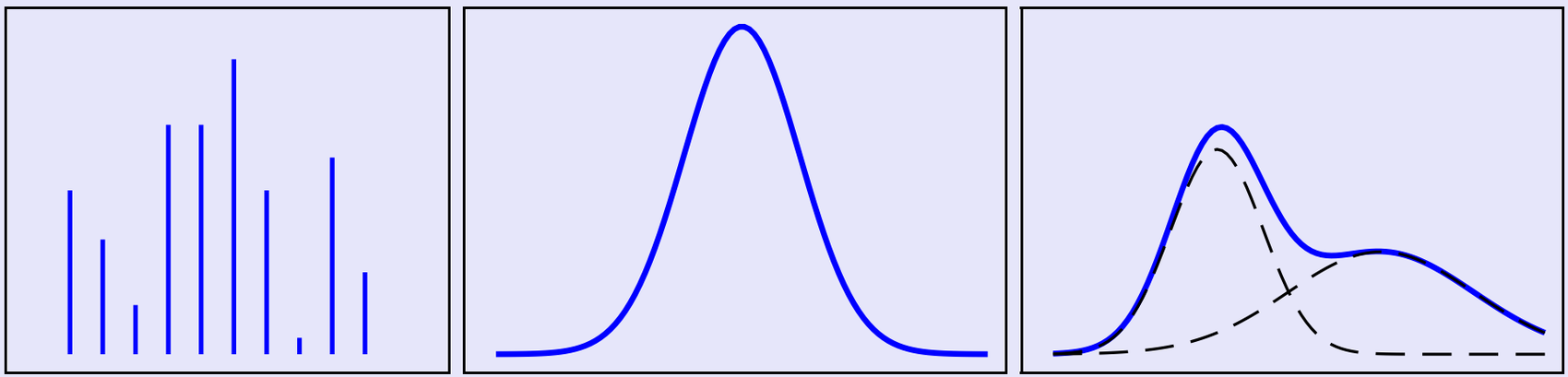
$x \in \mathbb{N}_Q,\ b \in {\mathbb{R}_+}^Q,\ \sum_{i=1}^{Q} b_i = 1, \qquad b(x) = b_x$

Gaussiana:

$x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+, \qquad b(x) = \mathcal{N}(\mu, \sigma; x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

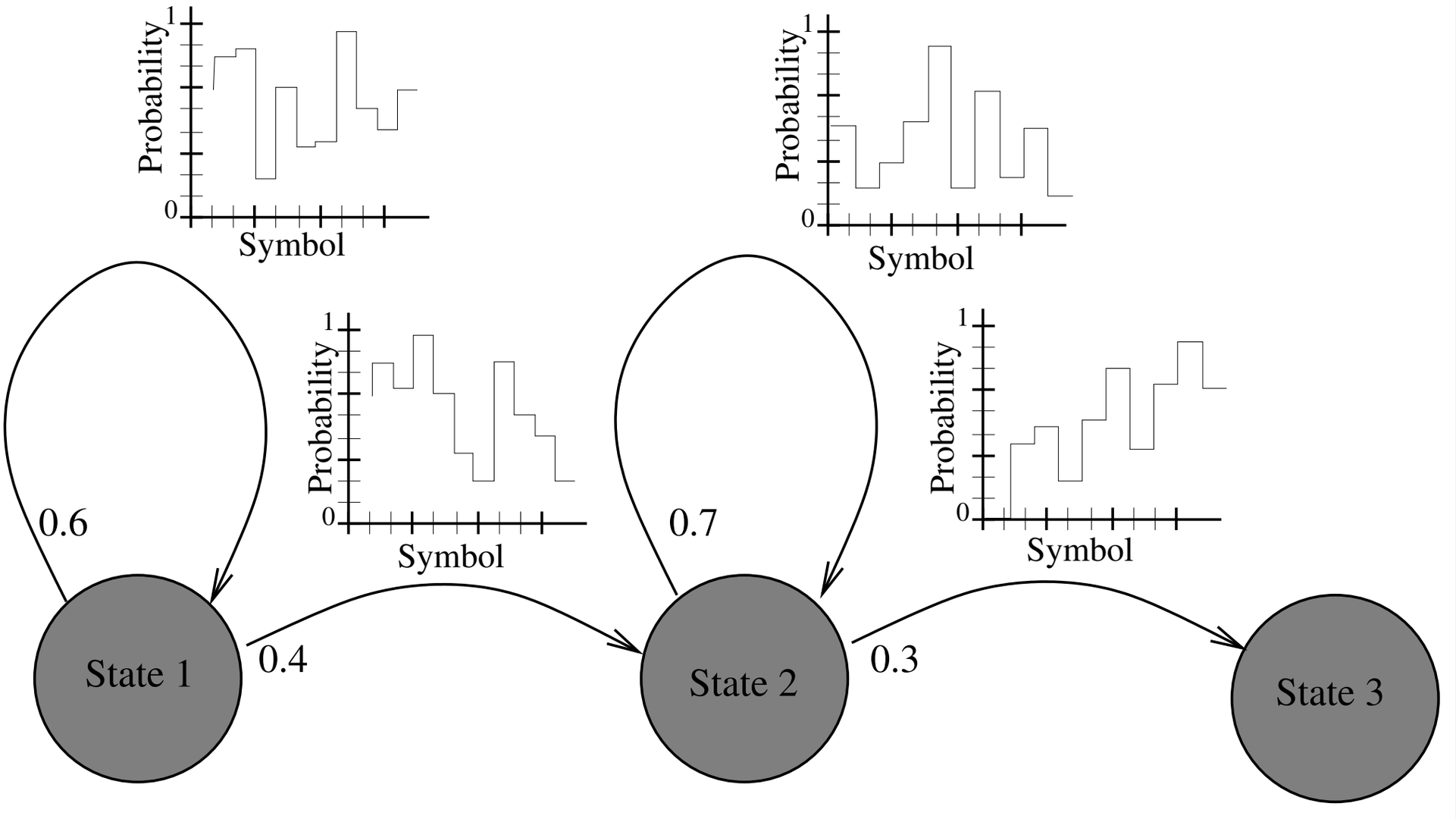
La più usata, *Mistura di Gaussiane*:

$x \in \mathbb{R}, \mu_k \in \mathbb{R}, \sigma_k \in \mathbb{R}_+, \sum_{k=1}^{K} w_k = 1, \quad b(x) = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \sigma_k; x)$



In realtà si usano le analoghe multidimensionali (alcune decine di componenti)

Rappresentazione a grafo di un HMM



HMMs: i tre problemi tipici

- ▷ Il calcolo della probabilità: Come calcolare $\Pr[\boldsymbol{x}_1^T]$
- ▷ La decodifica: Come determinare il cammino nascosto ottimo
- ▷ La stima: Come stimare i parametri del modello in modo da rappresentare efficacemente un congruo numero di esempi

L'esistenza di algoritmi efficienti per questi problemi è uno dei motivi del successo degli HMM.

HMM: calcolo di probabilità

Grazie alle ipotesi di independenza:

Probabilità di un cammino:
$$\Pr(\boldsymbol{s}_0^T) = \prod_{t=1}^{T} a_{s_{t-1}s_t}$$

Probabilità di emissione lungo un cammino:
$$\Pr(\boldsymbol{x}_1^T|\boldsymbol{s}_0^T) = \prod_{t=1}^{T} b_{s_{t-1}s_t}(x_t)$$

Probabilità totale di emissione:
$$\Pr(\boldsymbol{x}_1^T) = \sum_{\boldsymbol{s}_0^T \in \mathcal{S}^{T+1}} \prod_{t=1}^{T} a_{i_{t-1}i_t} b_{i_{t-1}i_t}(x_t)$$

La probabilità totale di emissione è data dalla somma delle probabilità di emissione lungo tutti i cammini possibili

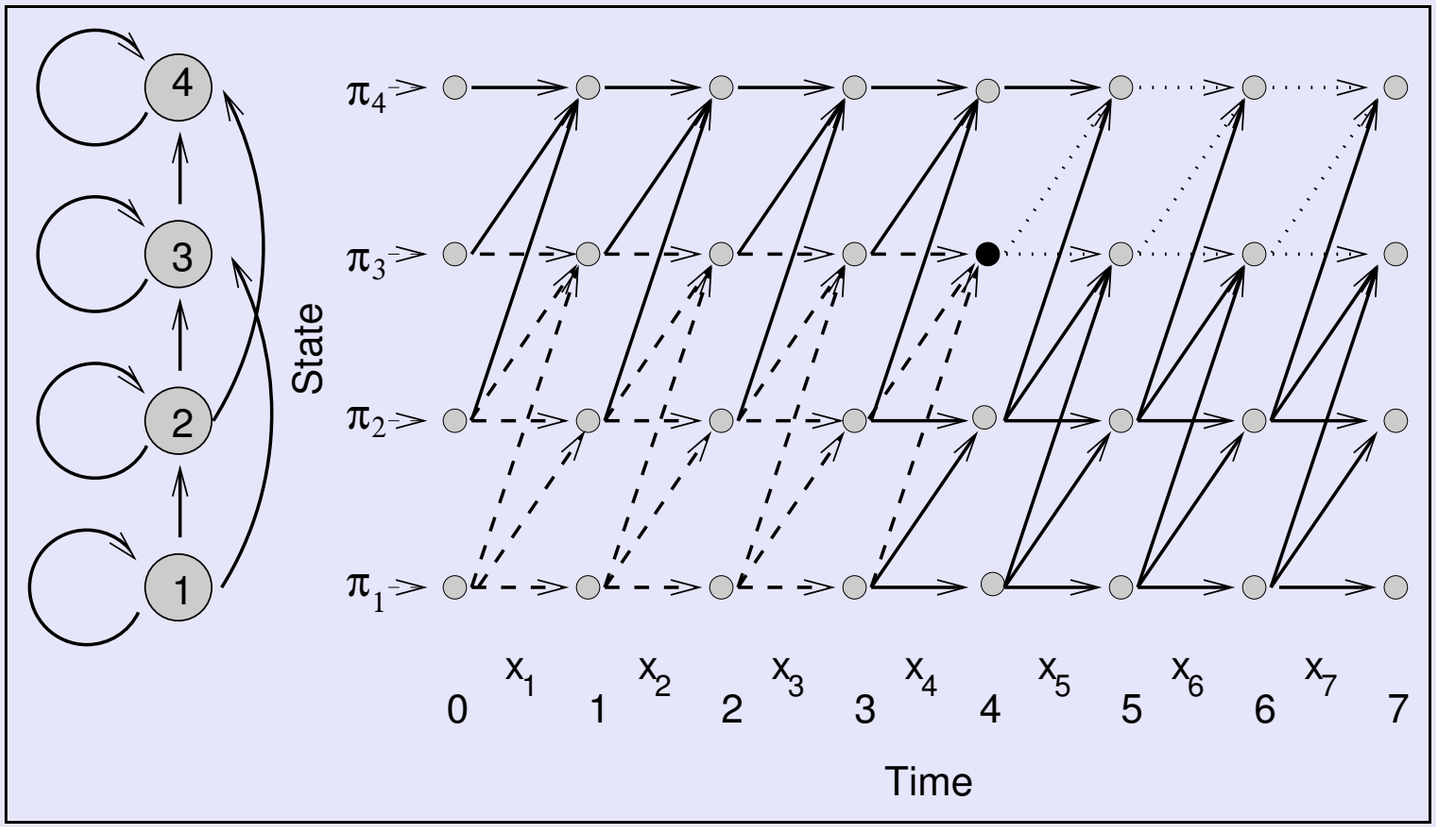
Il calcolo diretto ha complessità esponenziale, ma...

Struttura di calcolo: il trellis

Un grafo ottenuto con lo ”sviluppo” dell’automa nel tempo

Ciascun arco $(i, t) \rightarrow (j, t + 1)$ ha un peso dato da $a_{i,j} b_{i,j}(x_t)$

Ciascun cammino da $(0, 0)$ a $(7, 4)$ rappresenta una particolare evoluzione che può aver emesso la sequenza osservata.

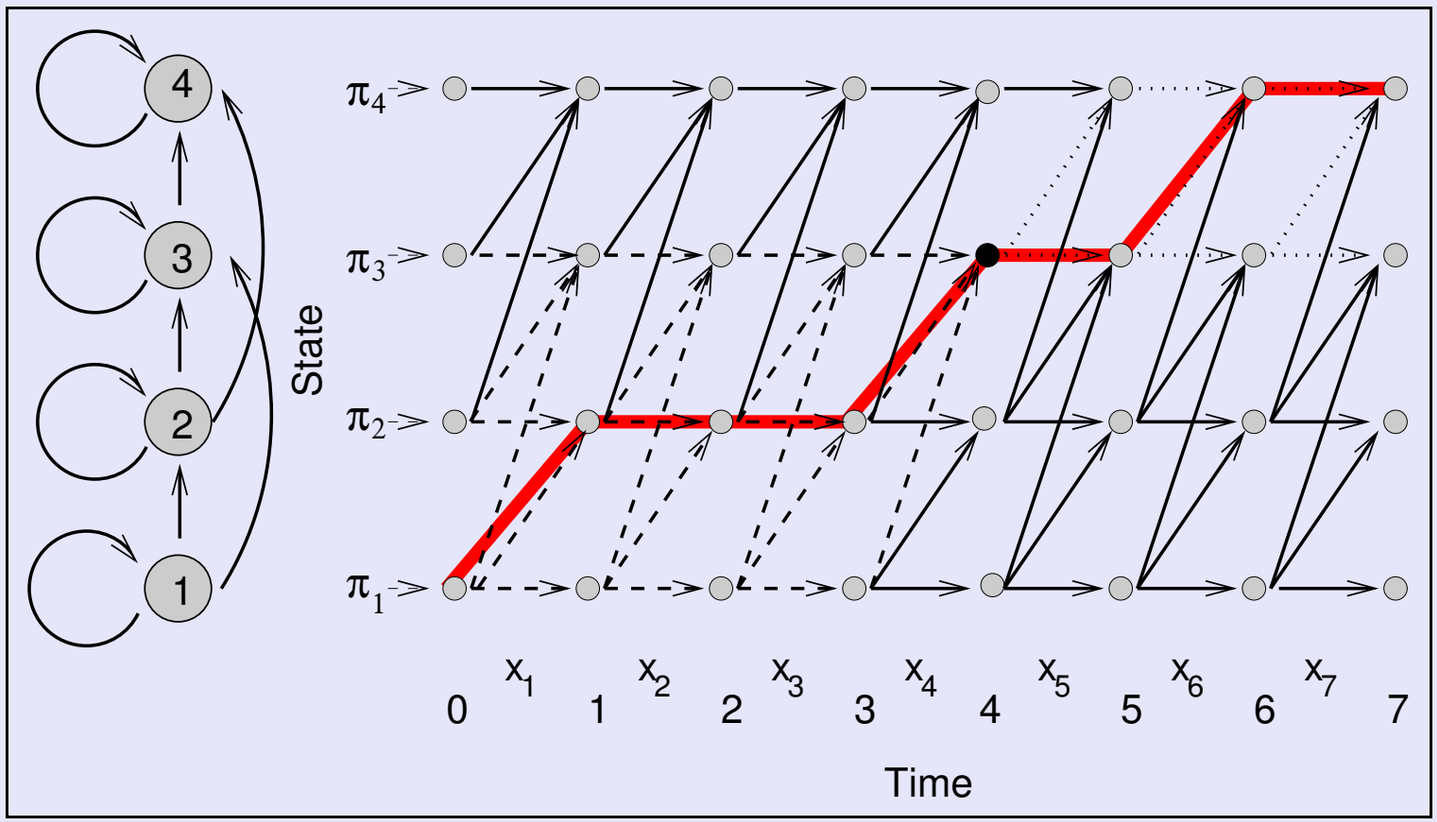


Programmazione Dinamica $\Rightarrow$ calcoli “per colonna” in maniera efficiente.

Struttura di calcolo: il trellis

Per il calcolo della probabilità di emissione, si sommano i pesi di tutti i cammini (algoritmo forward-backward)

Per la decodifica, si cerca quello che dà probabilità massima (algoritmo di Viterbi)



HMM: riconoscimento elementare

Per il caso di parole isolate:

- ▷ Si addestra un modello per ciascuna delle parole da riconoscere
- ▷ Data una sequenza in input, se ne calcola la probabilità di emissione con ciascuno dei modelli, eventualmente moltiplicandola per la probabilità a priori.
- ▷ Si sceglie la parola il cui modello dà probabilità massima

Per il parlato continuo:

- ▷ Si sfrutta integrazione fra AM e LM (in seguito)

HMM: addestramento

- Lo schema più usato si basa su Maximum Likelihood: cerca i valori dei parametri che massimizzano la probabilità di emissione dei dati di training
- Soluzione diretta impossibile, per la presenza della parte “nascosta”
- Usa un metodo iterativo (Baum-Welch, Expectation-Maximization).
- Non si possono osservare le statistiche necessarie alla stima diretta dei parametri, quindi:
  - Calcola i valori attesi delle statistiche in base alle osservazioni ed ai parametri attuali
  - Calcola i valori dei parametri che massimizzano la likelihood in base ai valori attesi
  - e ripete ...
- Necessita di molti esempi per ciascun elemento da riconoscere
- Sono usati anche criteri alternativi quali Maximum Mutual Information o Minimum Phone Error, più complessi.

Unità acustiche

- Abbiamo a disposizione dei modelli adeguati, ma quali modelli possiamo costruire?
- Nel caso di grandi vocabolari, non è possibile usare parole
- Siamo ”costretti” a scegliere un insieme ridotto di eventi fondamentali
- Fonemi: un numero limitato di suoni elementari con cui si possono comporre tutte le parole di una lingua.
- Sono pochi e ben rappresentati.
- Purtroppo, sono astrazioni e mal definiti acusticamente.
- Per esempio, la loro realizzazione è molto influenzata dal contesto (suoni adiacenti) a causa della coarticolazione

È una limitazione: si assume che la realizzazione di una parola sia la realizzazione di una catena di eventi. È proprio cosi‘?

Unità acustiche

- Per una data lingua i fonemi sono solitamente qualche decina

  `casa` $\rightarrow$ `/k/ /a/ /z/ /a/`

- Gli allofoni invece dipendendono dal contesto e sono decine di migliaia

  `casa` $\rightarrow$ `/?_k_a/ /k_a_z/ /a_z_a/ /z_a_?/`

- Potenzialmente, N^3 per i trifoni con N fonemi, ma ci sono vincoli fonotattici che ne riducono il numero

- È necessario produrre la trascrizione fonetica:
  - Manualmente, precisa ma costosa
  - Automaticamente, economica ma soggetta ad errori
  - Più o meno difficile a seconda della lingua

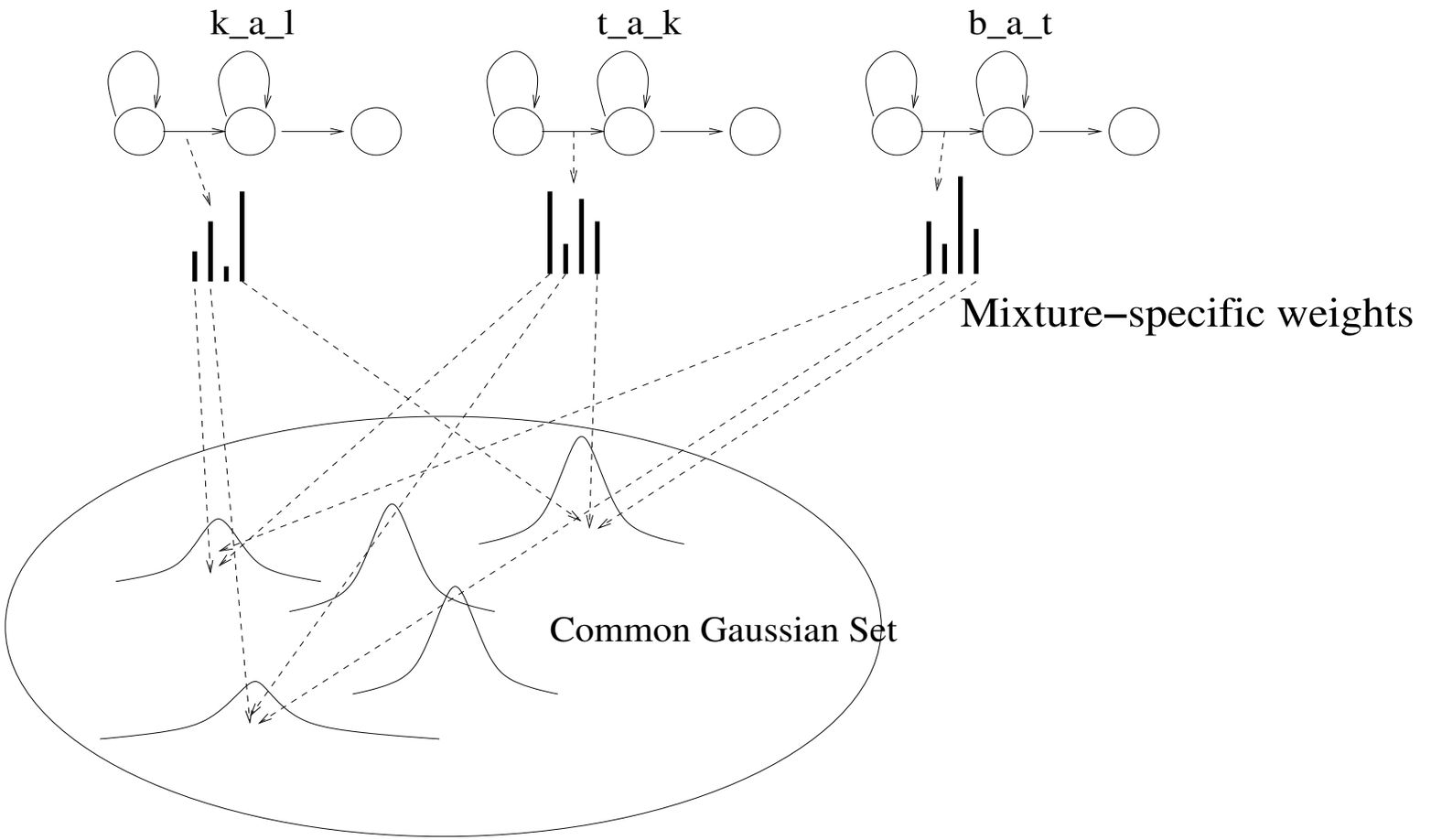
- È spesso necessario considerare più varianti di pronuncia di una stessa parola

Sparsità dei dati: Parameter Tying

L'utilizzo di unità *context-dependent* pone nuovamente problemi di sparsità dei dati in addestramento, ma...

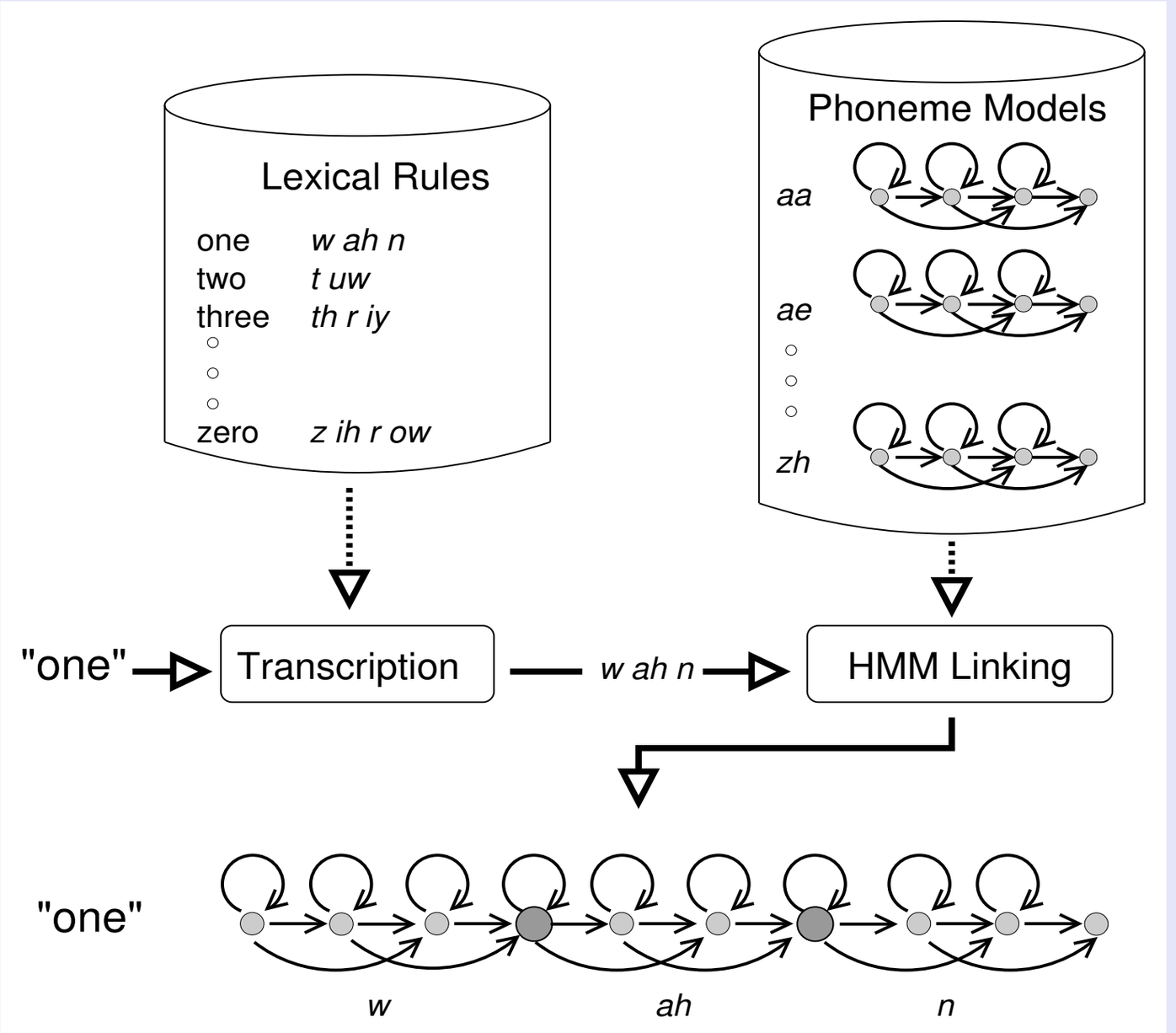
molte unità sono relativamente simili, possono condividere parte dei parametri.

Esempio: *Phonetically Tied Mixtures*



Composizione di modelli

Gli HMM si prestano bene a costruire modelli composti a partire da modelli elementari:



Un modello alternativo: Reti Neurali

- Sono strutture di calcolo basate su nodi e connessioni, ispirate dalla struttura dei neuroni
- Derivano da un approccio diverso, in cui si pone l'accento sulla discriminazione fra eventi
- Idealmente, possono imparare una qualunque mappa di classificazione, poiché non fanno assunzioni sulle caratteristiche dell'input (a differenza dei modelli statistici)
- Nate come classificatori di pattern statici, sono state estese a classificatori di sequenze usando reti ricorrenti
- Sono state applicate anche al RAP, con risultati simili ad HMMs, ma non sui task più grandi
- Dei maggiori sistemi odierni, pochi le usano, e questi sono sistemi ibridi, in cui le NN sono integrate in HMMs sostituendo le misture di Gaussiane per il calcolo della probabilità di emissione.
- In quest'ultima architettura, sono vantaggiose in termini di efficienza

Modelli del linguaggio

Sono necessari per compensare l'imprecisione del modello acustico.

Vincoli rigidi o elastici, quindi modelli:

- A regole (grammatiche):
  - Accettano solo un ben definito insieme di sequenze.
  - Utili per comandi, menu, espressioni comuni (date, numeri, ecc..).
  - In genere grammatiche *regolari*, al più *context-free*.
- Statistici:
  - Consentono di riconoscere qualunque sequenza componibile con le parole del dizionario, ma assegnano probabilità diverse
  - Utili per dettatura di testi, giornali, notiziari, ecc..

Esempio di grammatica a regole: Numeri da 1 a 1000

```
DIGIT2  ::=  “due” | ...| “nove”
 DIGIT  ::=  “uno” | DIGIT2
  TEEN  ::=  “dieci” | “undici” | ...| “diciannove”
   TEN  ::=  “venti” | “trenta” | ...| “novanta”

  NUM2  ::=  DIGIT | TEEN | TEN
  NUM2  ::=  TEN DIGIT

  NUM3  ::=  NUM2
  NUM3  ::=  “cento” | DIGIT2 “cento”
  NUM3  ::=  “cento” NUM2 | DIGIT2 “cento” NUM2
  NUM3  ::=  “mille”
```

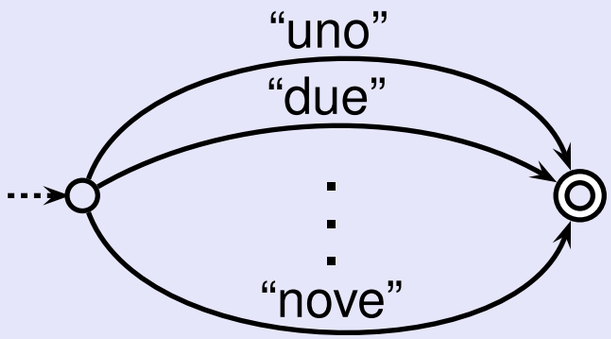
Dizionario: 29 parole, Linguaggio: 1000 parole composte

Grammatiche e automi

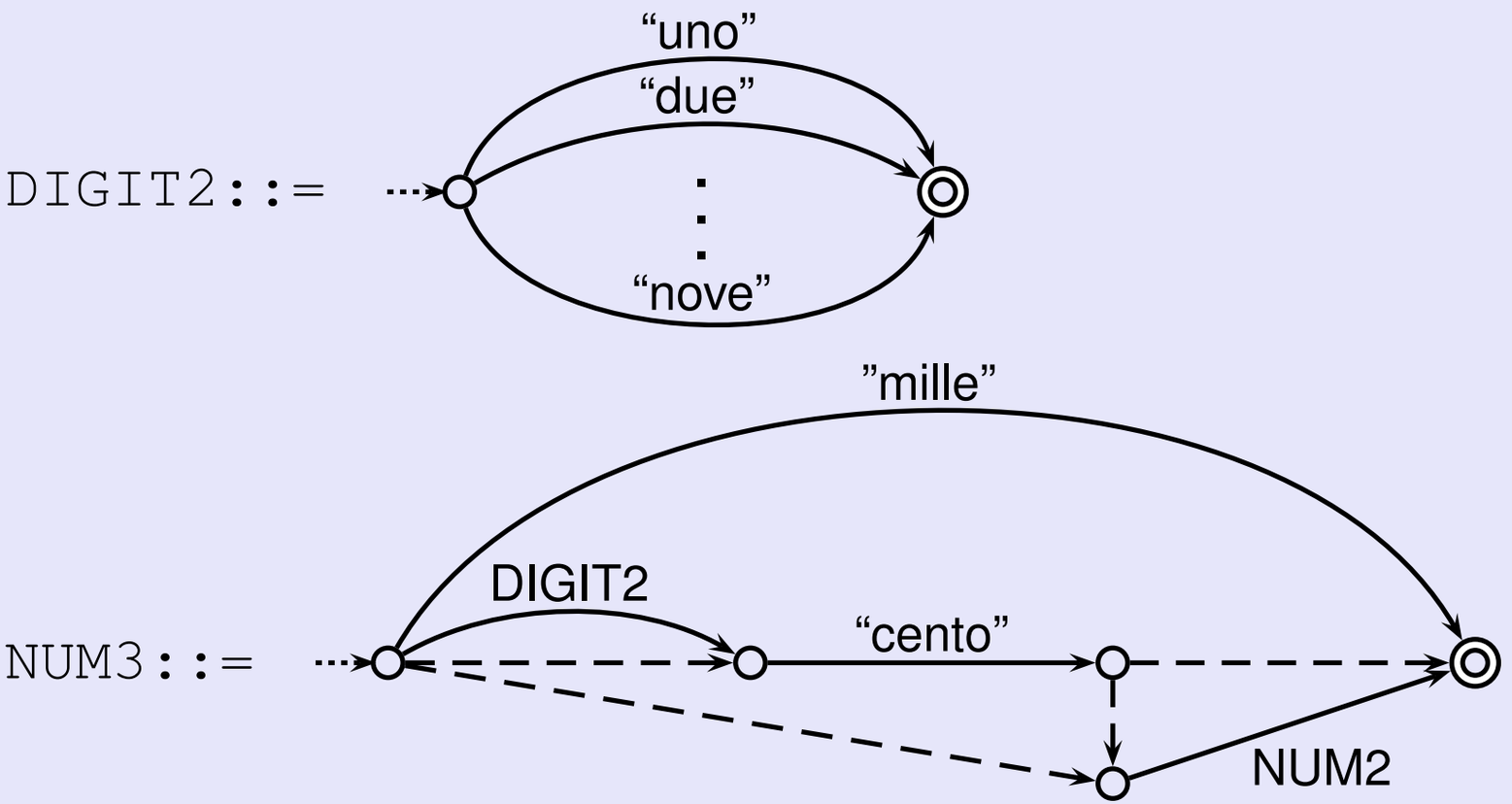
Grammatiche regolari o context-free si possono rappresentare con automi (ricorsivi per context-free).

Grammatiche regolari o context-free si possono rappresentare con automi (ricorsivi per context-free).

DIGIT2::=



Grammatiche regolari o context-free si possono rappresentare con automi (ricorsivi per context-free).



Modello del linguaggio statistico

- Non ha funzione discriminativa: non ci sono modelli diversi in competizione e non è legato all'evidenza acustica
- Racchiude conoscenza sul dominio del linguaggio
- È essenziale per assegnare delle probabilità a priori alle sequenze di parole
- Indirizza la decodifica acustica a scartare le ipotesi linguisticamente scorrette anche se acusticamente plausibili (p.e. *"il con vento"* vs. *"il convento"*)

Il più comune modello del linguaggio statistico: n-grammi

Predice la probabilità di una parola in base alle $n-1$ precedenti, ovvero:

è una catena di Markov di ordine $n-1$.

$$\Pr[\boldsymbol{w}_1^T] = \prod_{t=1}^{T} \Pr[w_t | w_1 \dots w_{t-1}]$$

$$\Pr[\boldsymbol{w}_1^T] \approx \prod_{t=1}^{T} \Pr[w_t | w_{t-n+1} \dots w_{t-1}]$$

$n = 2 \rightarrow$ bigrammi

$$\Pr[\boldsymbol{w}_1^T] \approx \prod_{t=1}^{T} \Pr[w_t | w_{t-1}]$$

$n = 3 \rightarrow$ trigrammi

$$\Pr[\boldsymbol{w}_1^T] \approx \prod_{t=1}^{T} \Pr[w_t | w_{t-2} w_{t-1}]$$

In pratica si arriva fino a $n = 5$

Numero potenziale di parametri proibitivo, p.e.: 10.000 parole $\Rightarrow$ mille miliardi di trigrammi

Modello a n-grammi: addestramento

- ▷ Si usa anche qui il criterio Maximum Likelihood, e stavolta non ci sono parametri "nascosti"
- ▷ Il problema maggiore è la sparsità dei dati, rispetto al grande numero di parametri
- ▷ Degli n-grammi possibili, solo una frazione può apparire, anche in insiemi grandi di dati
- ▷ Si usano tecniche di smoothing e backoff, che mantengono il numero di parametri in limiti ragionevoli
- ▷ Si distinguono nel modo in cui assegnano probabilità ad eventi mai visti

Esempio: smoothing per interpolazione

La stima ML assegnerebbe al trigramma $\Pr(w_3|w_1, w_2)$ la sua frequenza relativa

$$f(w_3|w_1, w_2) = \frac{c(w_1, w_2, w_3)}{c(w_2, w_3)}$$

Invece, si utilizza un frequenza scontata (discounted)

$$f^*(w_3|w_1, w_2) \leq f^*(w_3|w_1, w_2)$$

e si interpola con un ML di ordine inferiore:

$$\Pr(w_3|w_1, w_2) = f^*(w_3|w_1, w_2) + \lambda(w_1, w_2) \Pr(w_3|w_2)$$

Tramite discounting si riserva una certa “massa” di probabilità ad eventi mai visti.

Esempio: smoothing per interpolazione

$$\Pr(w_3|w_1,w_2) = f^*(w_3|w_1,w_2) + \lambda(w_1,w_2)\Pr(w_3|w_2)$$

λ deve essere tale da far sì che le probabilità sommino a 1:

$$\lambda(w_1,w_2) = 1 - \sum_w f^*(w|w_1,w_2)$$

e quindi dipende dal metodo di discounting applicato.

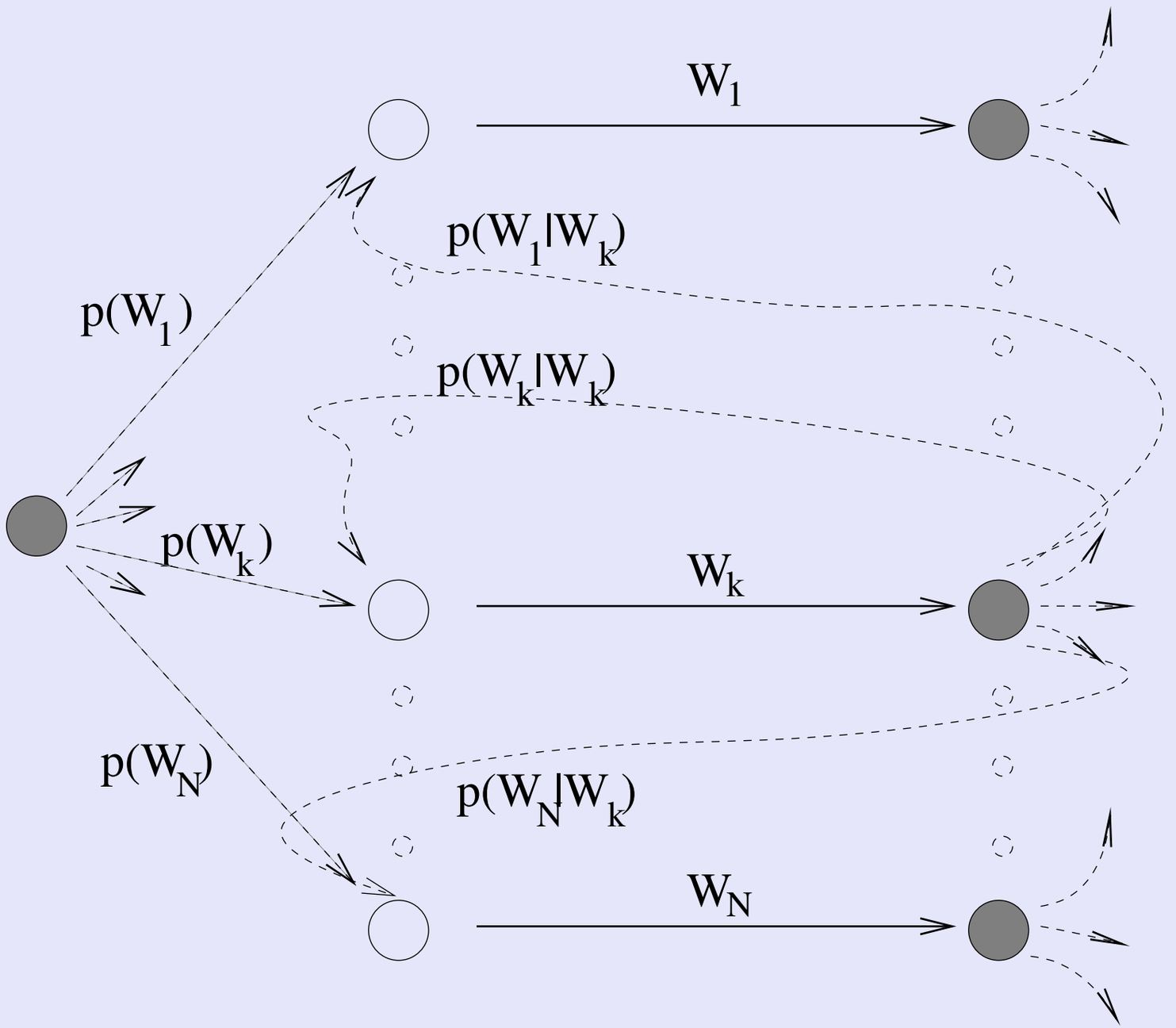
Per esempio Shift-1 discounting sottrae 1 al conteggio delle occorrenze:

$$f^*(w_3|w_1,w_2) = \max\left\{\frac{c(w_1,w_2,w_3)-1}{c(w_1,w_2)}, 0\right\}$$

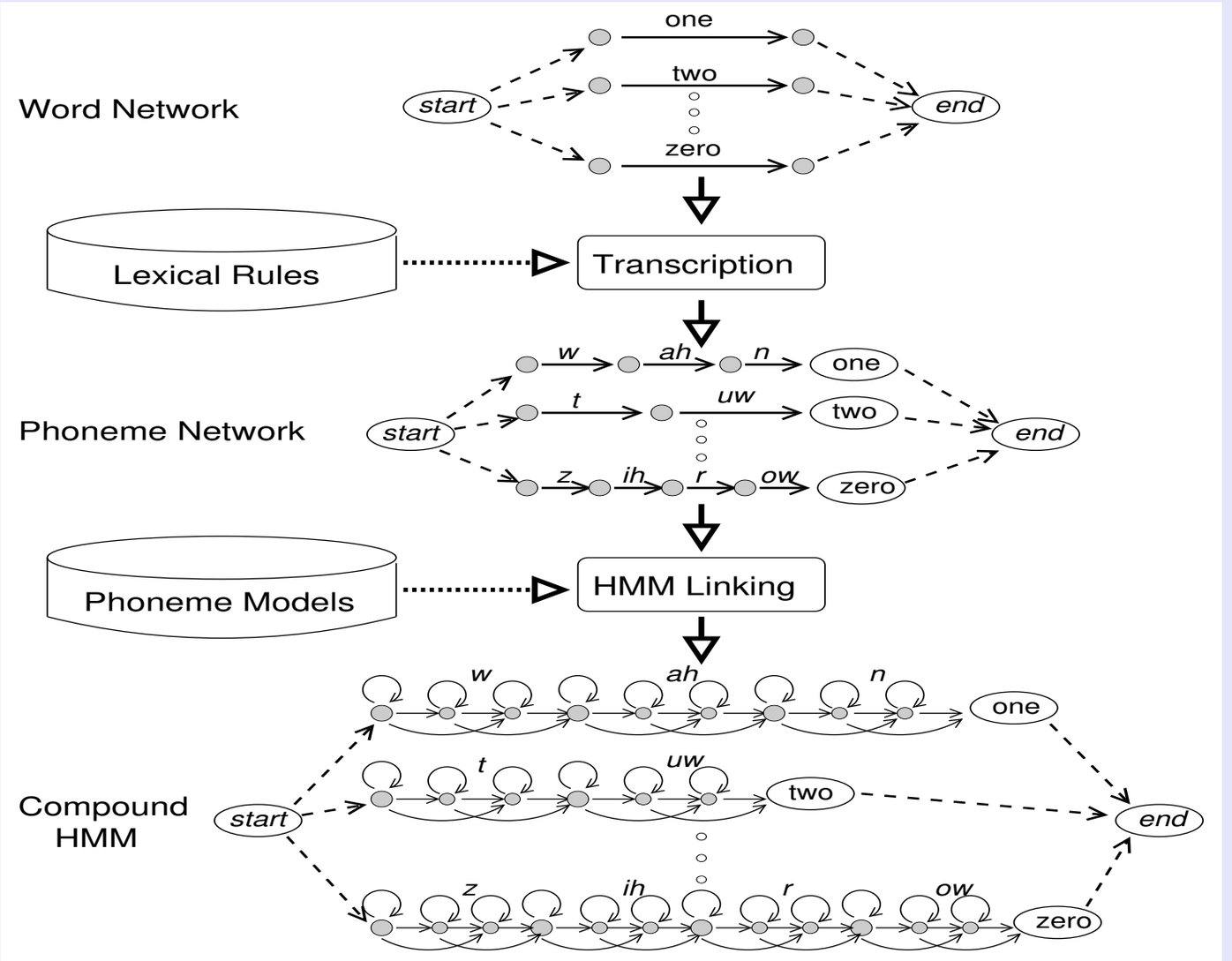
$$\lambda(w_1,w_2) = \frac{|\mathrm{succ}(w_1,w_2)|}{c(w_1,w_2)}$$

Ci sono altri metodi più sofisticati, che non “cancellano” contatori.

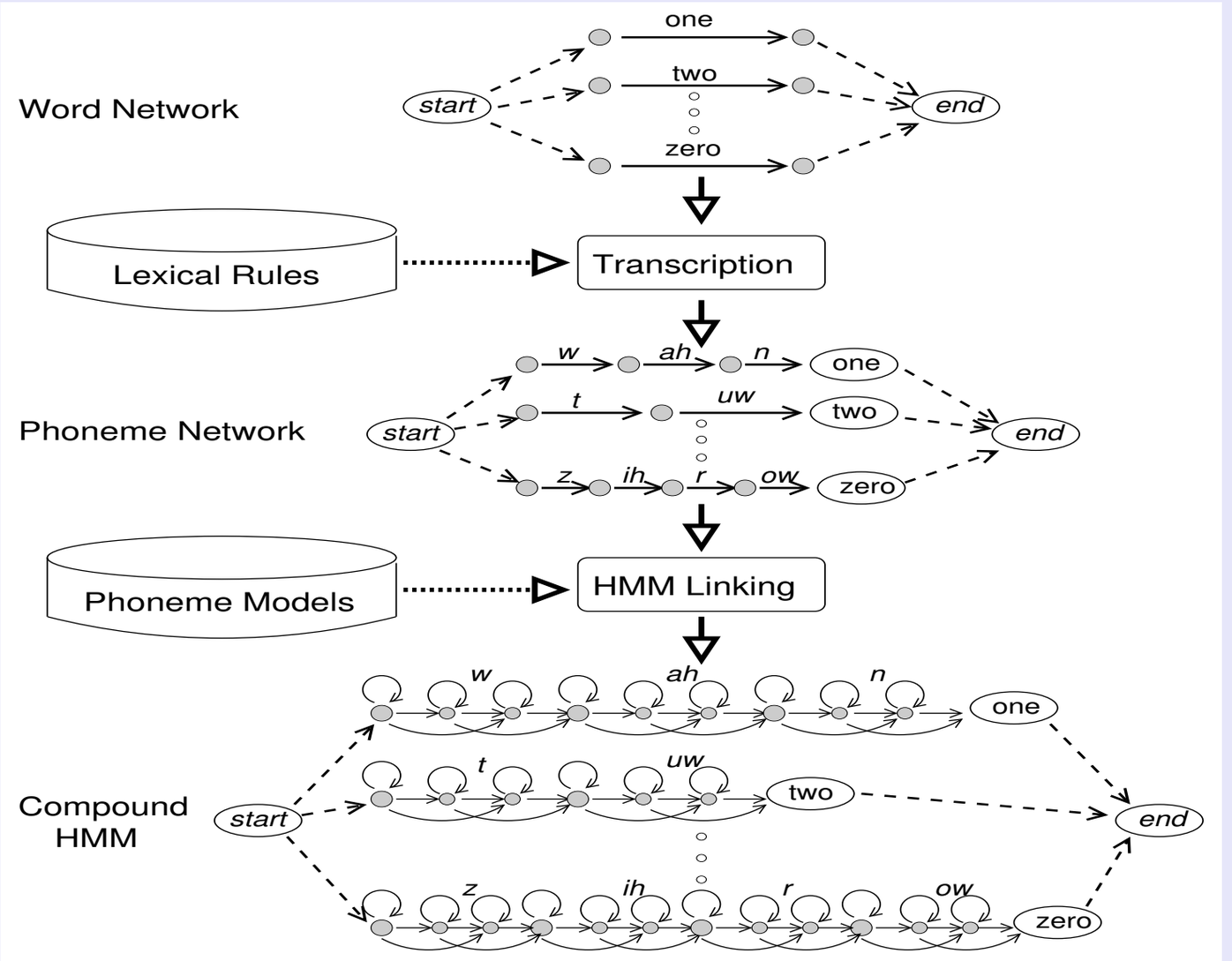
Rappresentazione a grafo di un ML a bigrammi



Composizione di reti integrate



Composizione di reti integrate



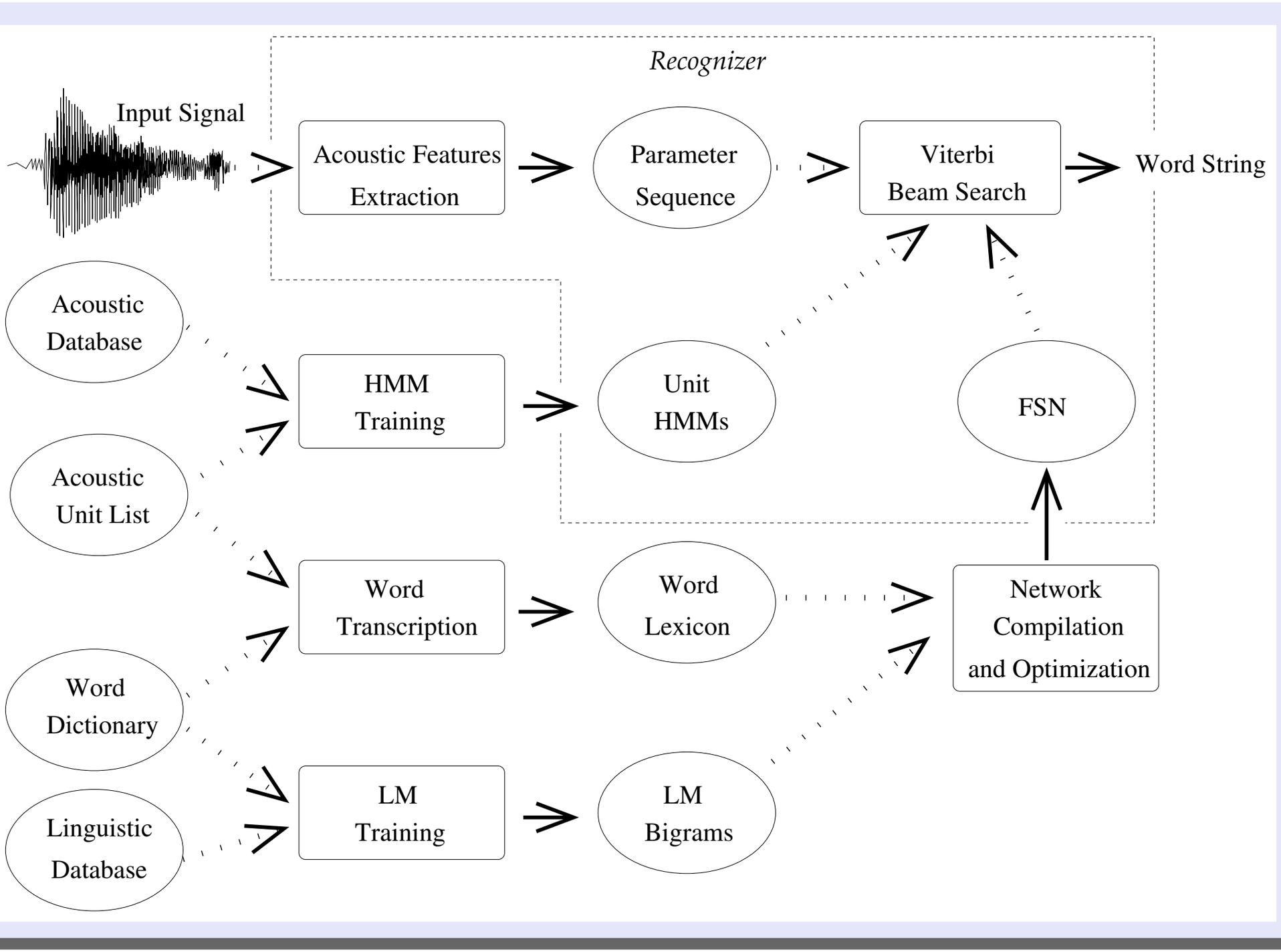
A questo punto, il problema del riconoscimento diventa la ricerca di un cammino ottimale in un grande Modello di Markov Nascosto.

Riassumendo...

Abbiamo:

- ▷ Estrazione di features da forma d'onda basata su short-term spectral analysis
- ▷ Un modello statistico per calcolare probabilità di sequenze di features
- ▷ Un modello statistico per assegnare probabilità a sequenze di parole, e quindi rappresentare il dominio linguistico
- ▷ Dei metodi per stimare i parametri che possono elaborare grandi quantità di dati, e modulare opportunamente la complessità del modello
- ▷ Un formalismo che permette di rappresentare in forma omogenea le due fonti di conoscenza
- ▷ Un algoritmo efficiente per ritrovare la sequenza di stati che con massima probabilità ha emesso una sequenza osservata

Schema di base di un sistema RAP completo



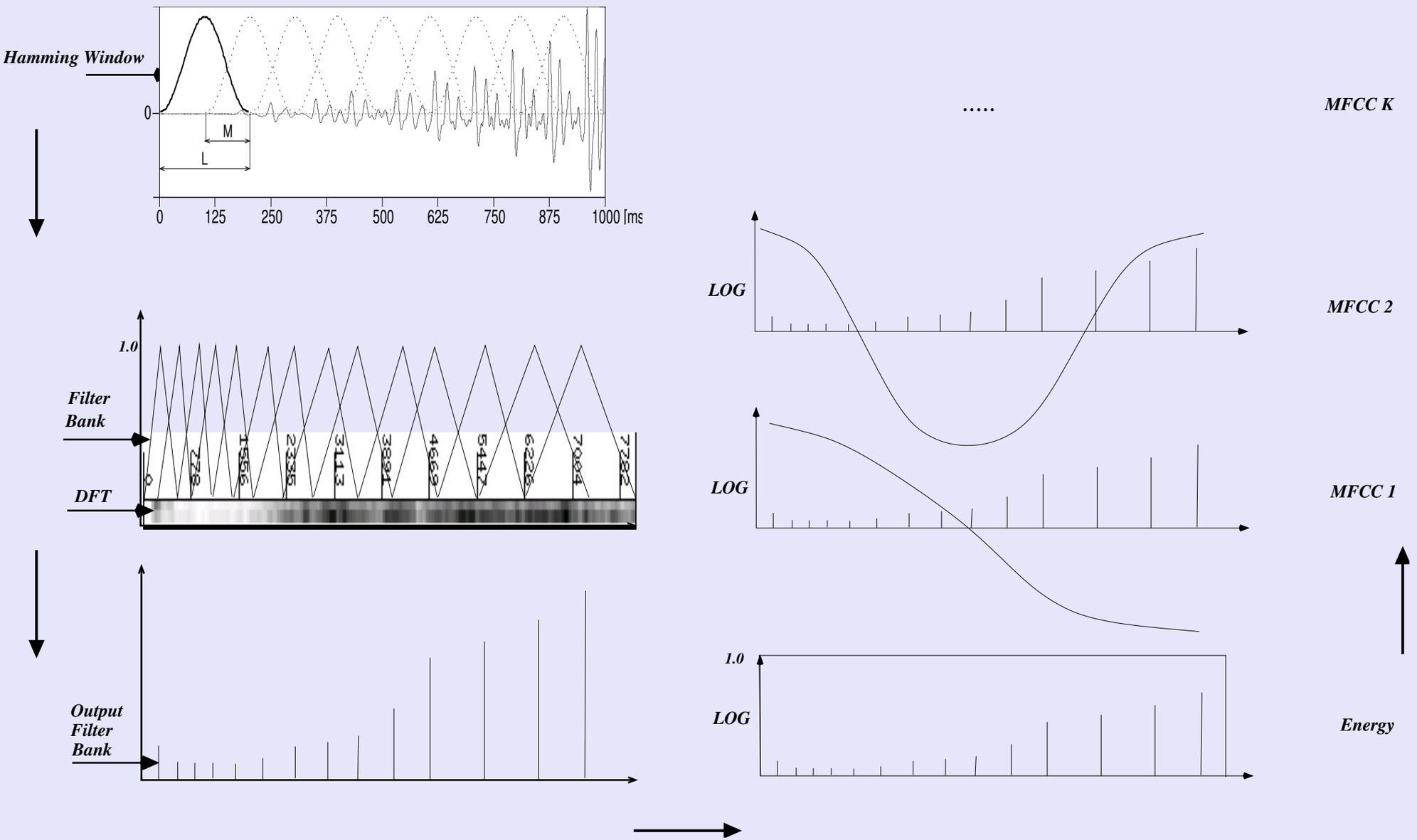
Calcolo Features Acustiche

- Ne esistono diverse, ma anche qui ci sono scelte prevalenti.
- Le rappresentazioni parametriche più comuni sono
  - MFCC: *MEL scaled cepstral coefficients*.
  - PLP: *Perceptual Linear Prediction*
- Hanno basi percettive lasche (variazione della risoluzione in frequenza su scala percettiva)
- Tentativi di modellazione più accurata dell'orecchio non hanno portato a vantaggi significativi

Seguono uno schema simile:

1. Blocking
2. Finestratura
3. Analisi spettrale
4. Banco di filtri
5. Trasformazione coefficienti
6. Aggiunta parametri dinamici

Calcolo Features Acustiche: MFCC



Addestramento di un sistema

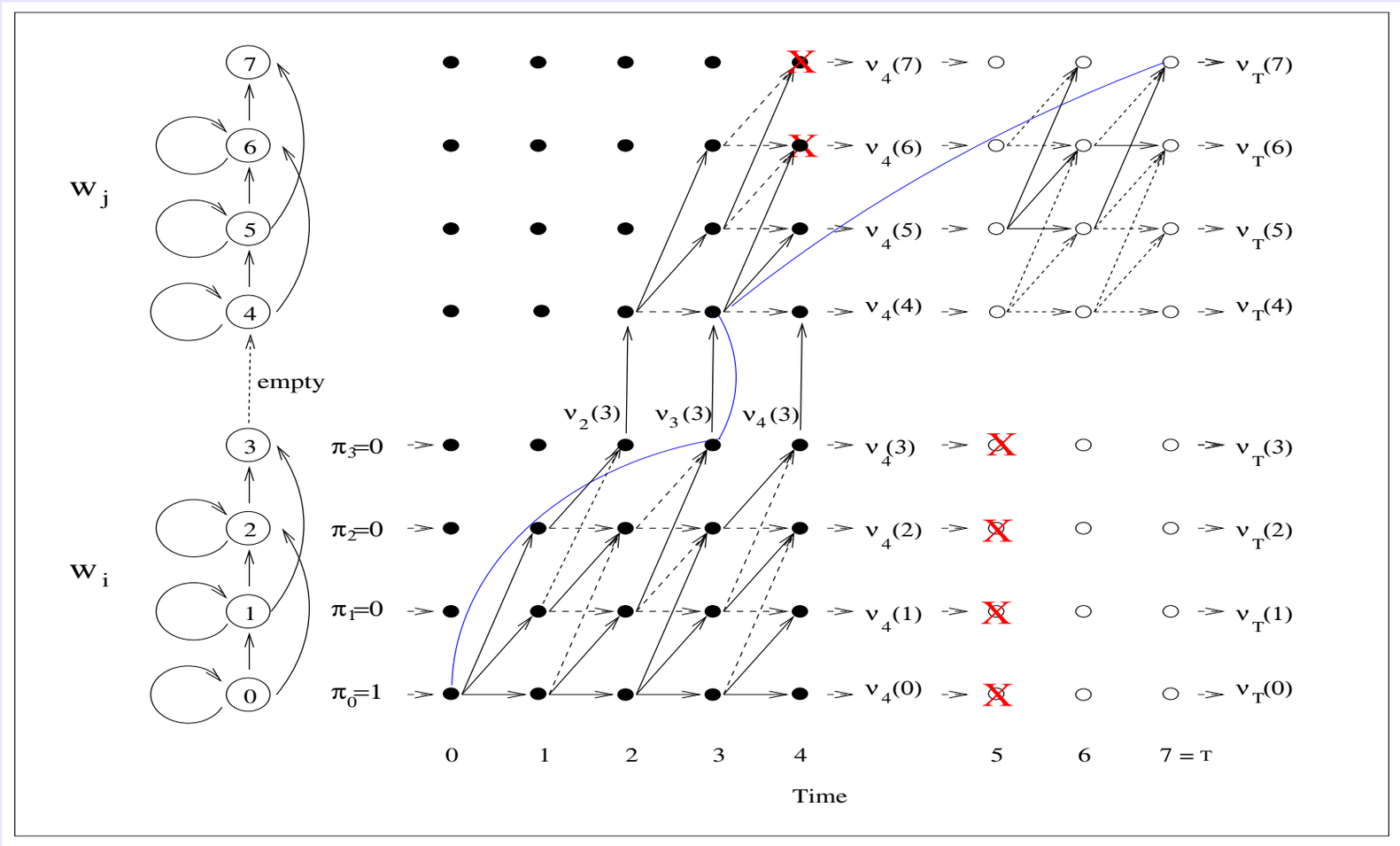
- Sono necessarie cospicue quantità di materiale acustico e linguistico
- Ordine di grandezza per i sistemi più grandi:
  - centinaia di ore per AM
  - centinaia di milioni di parole per LM
- Il materiale acustico “dovrebbe” essere trascritto accuratamente, per servire da esempio, ma ciò è costoso
- Ora sta diventando comune usare materiale acustico trascritto automaticamente

Riconoscimento con grandi vocabolari

Ricerca di un cammino a massima probabilità su un grafo: il trellis corrispondente ad un HMM composto.

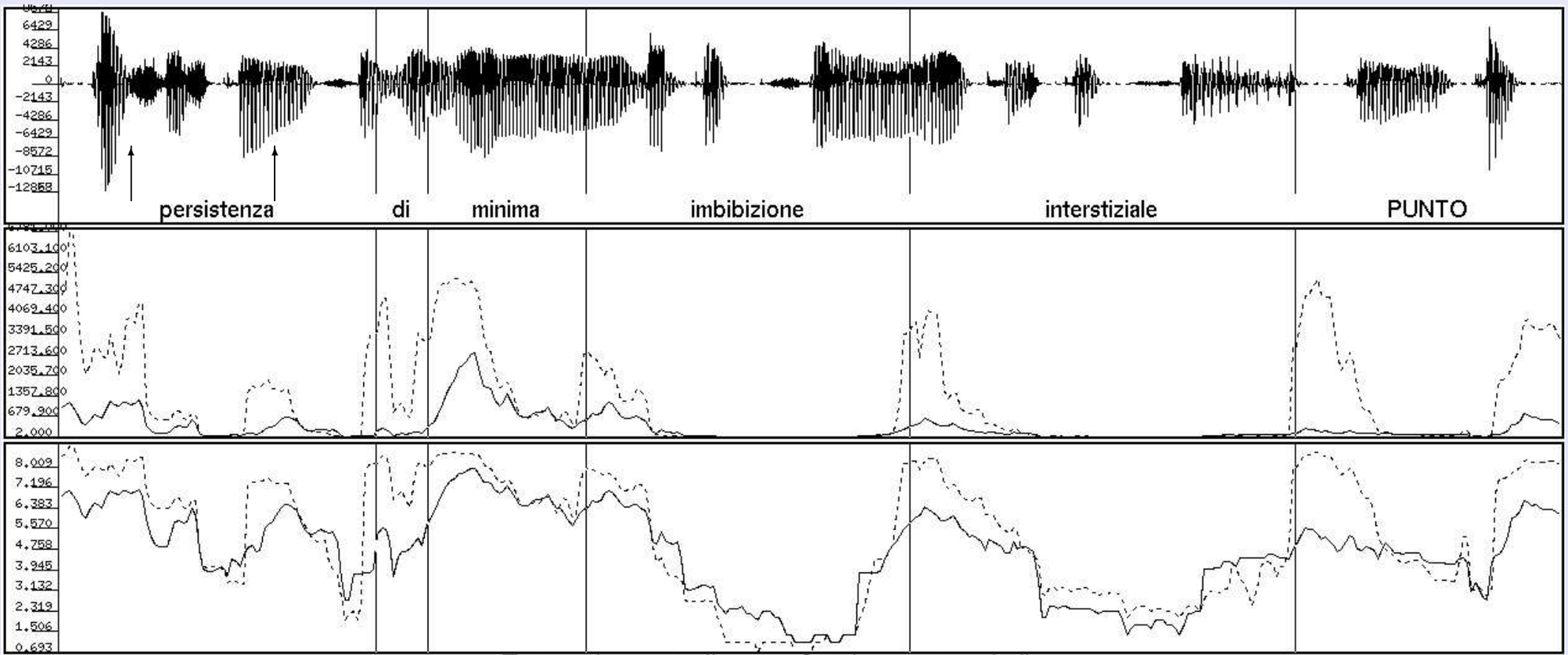
- ▷ L'algoritmo di base è l'algoritmo di Viterbi.
- ▷ Si mantiene però una distinzione fra stati interni ai modelli e stati della rete di unità, per poter ricostruire solo le parti significative del cammino
- ▷ Lo spazio di ricerca può essere molto grande, anche miliardi di stati ed archi.
- ▷ Una esplorazione esaustiva è generalmente impossibile.
- ▷ Si utilizzano tecniche di semplificazione, quali Beam-Search, che evita di considerare ipotesi poco promettenti
- ▷ Una efficiente implementazione è fondamentale per poter usare grandi vocabolari
- ▷ La rete può essere interamente compilata a priori, oppure espansa dinamicamente.

Viterbi Beam Search



Ad ogni istante di tempo, scarta gli stati la cui probabilità differisce più di una certa soglia dal migliore.

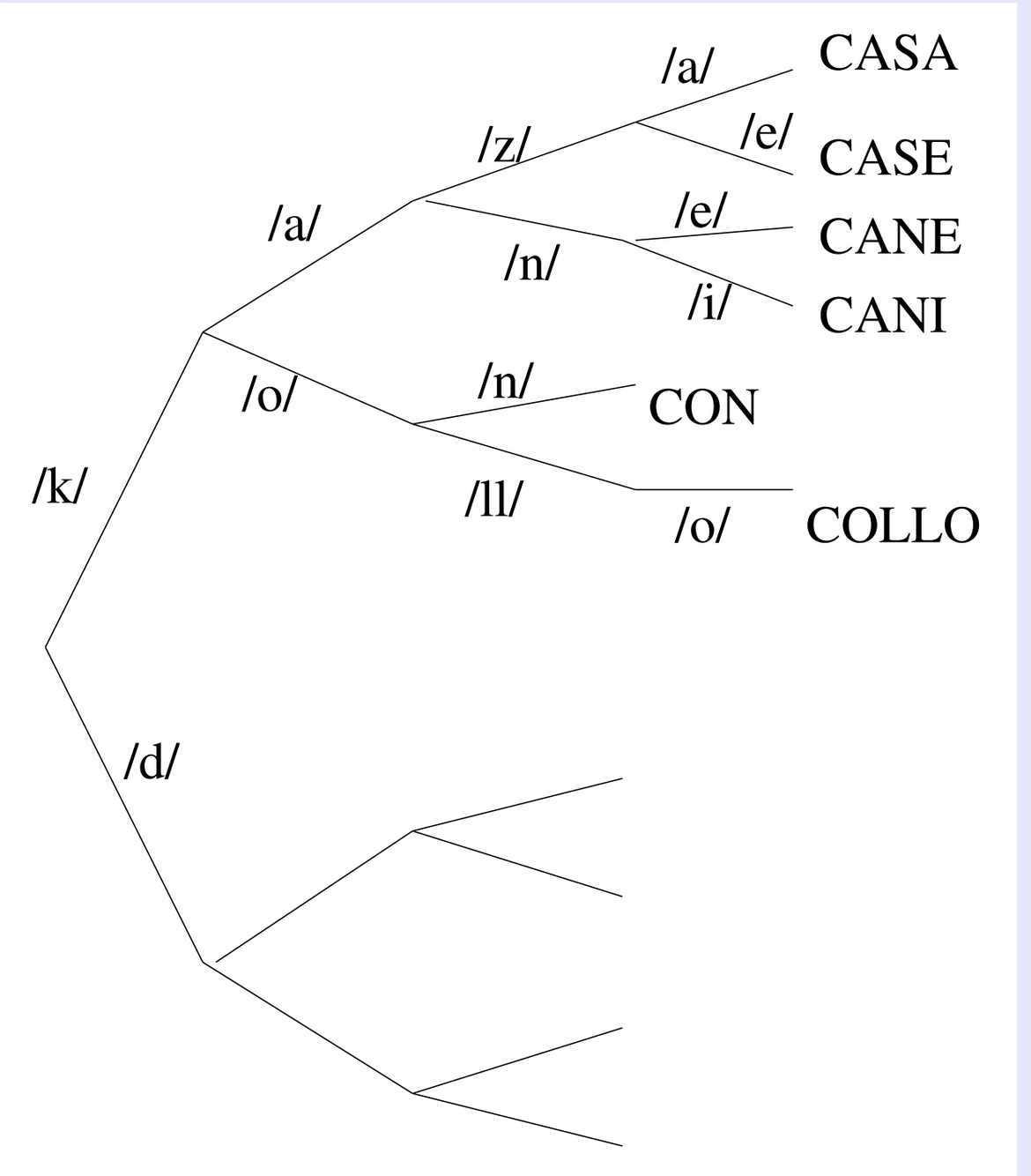
Regioni critiche in decodifica



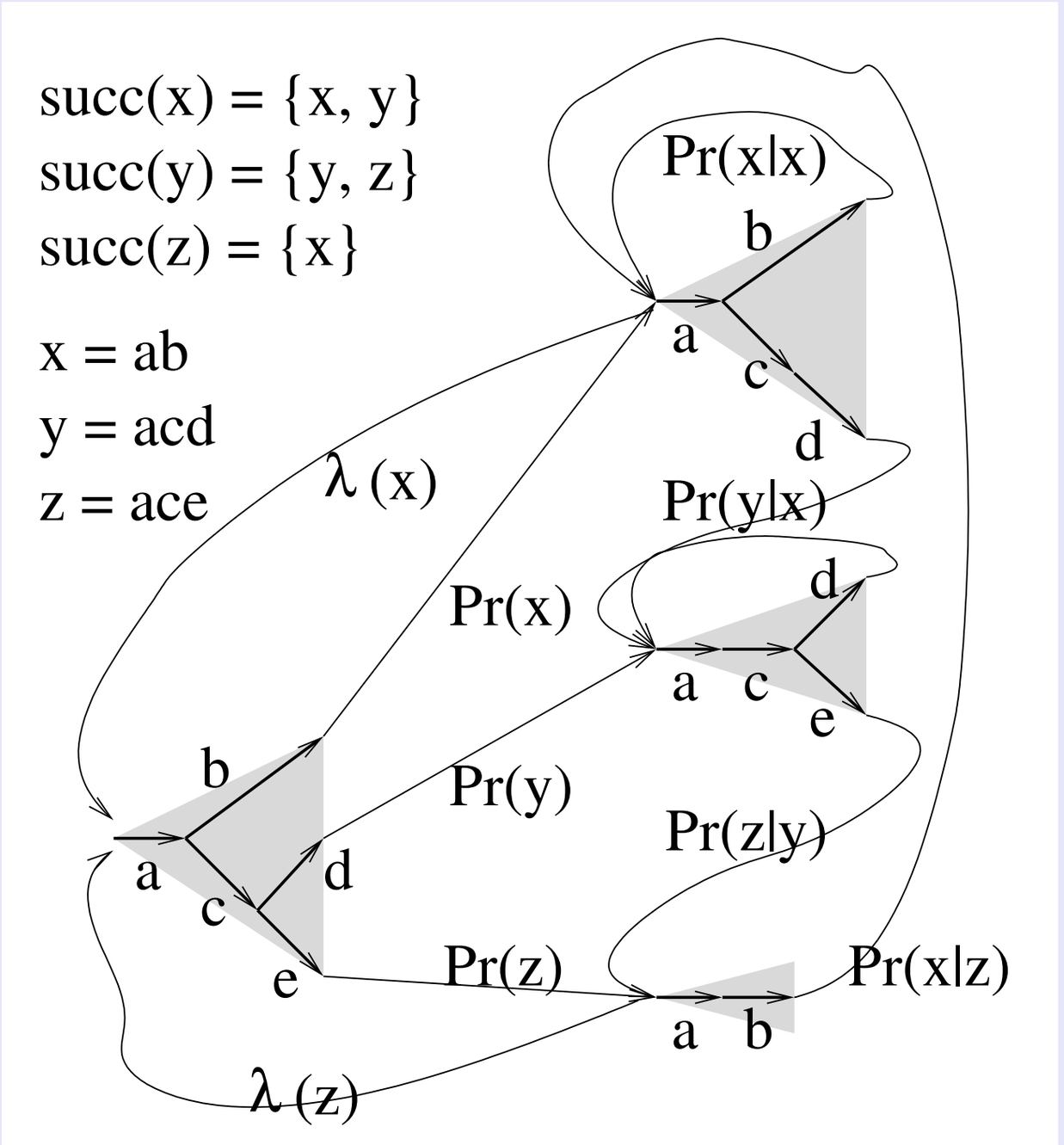
Tratteggiata: rete lineare, Continua: rete ad albero

Variazione del numero di stati attivi durante la decodifica di una frase

Lessico ad albero



Rappresentazione ad albero di bigrammi



È sufficiente?

- Solo per task non troppo complicati, purtroppo
- Quando si ha a che fare con tanti parlatori e tante condizioni diverse, un modello generale non basta
- Bisogna focalizzare i parametri sui singoli parlatori

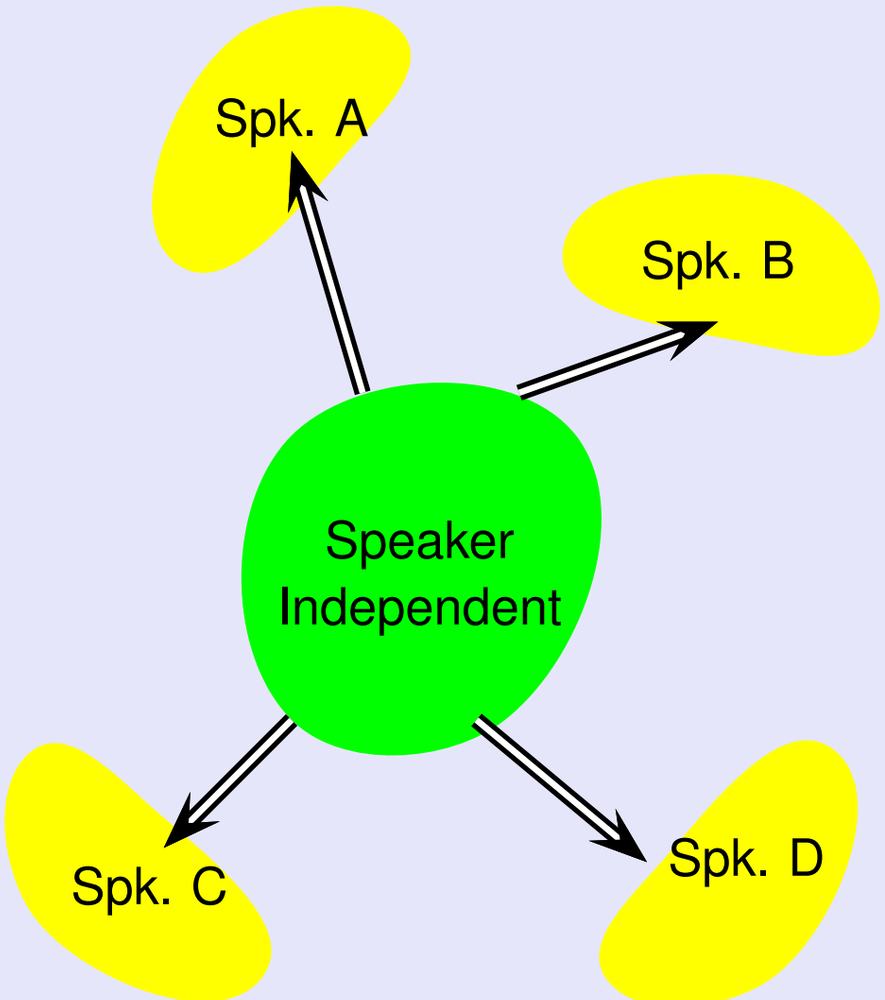
Adattamento

- Data una quantità non grande di dati trascritti di un certo parlatore, non si possono stimare i parametri di un modello
- Però si possono alterare i parametri di un modello generale in modo da meglio rifletterne le caratteristiche.
- Esistono diverse tecniche, alcune agiscono sui modelli, altre sulle features

(L'adattamento esiste anche per il modello del linguaggio, in questo caso si adatta a domini diversi)

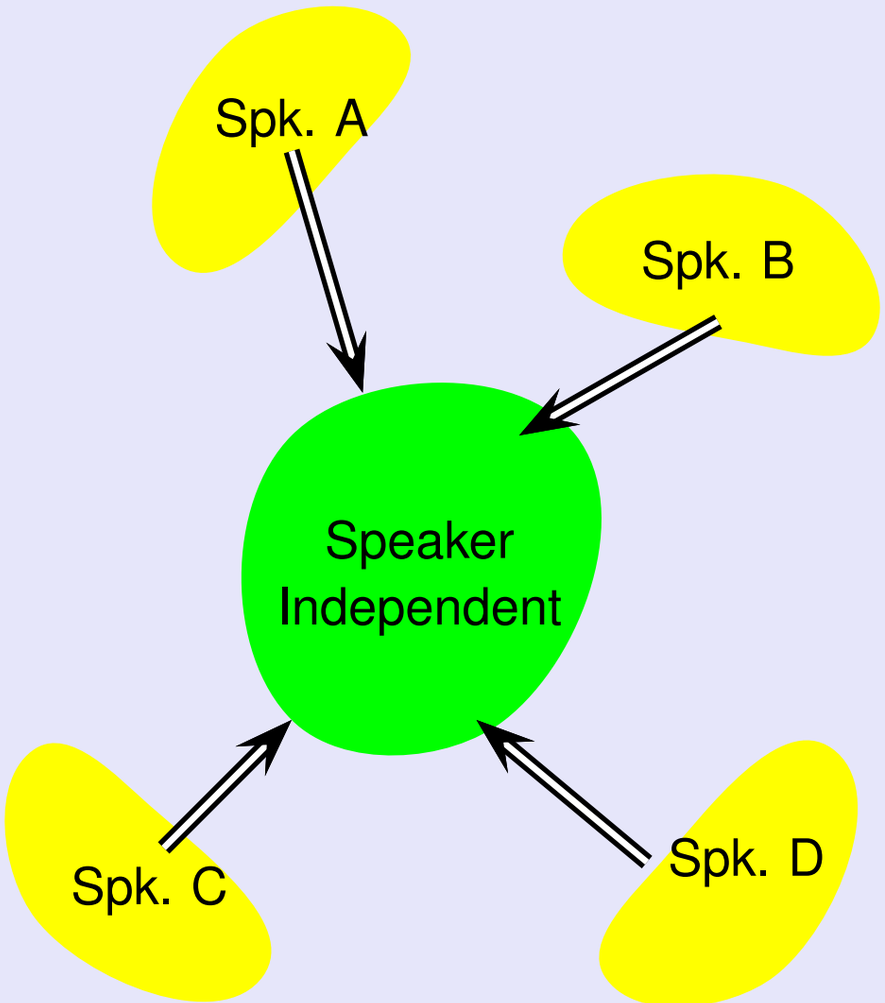
Adattamento di Modelli

Lo spazio dei parametri di un modello speaker-independent viene alterato per avvicinarsi a quello di ciascun parlatore



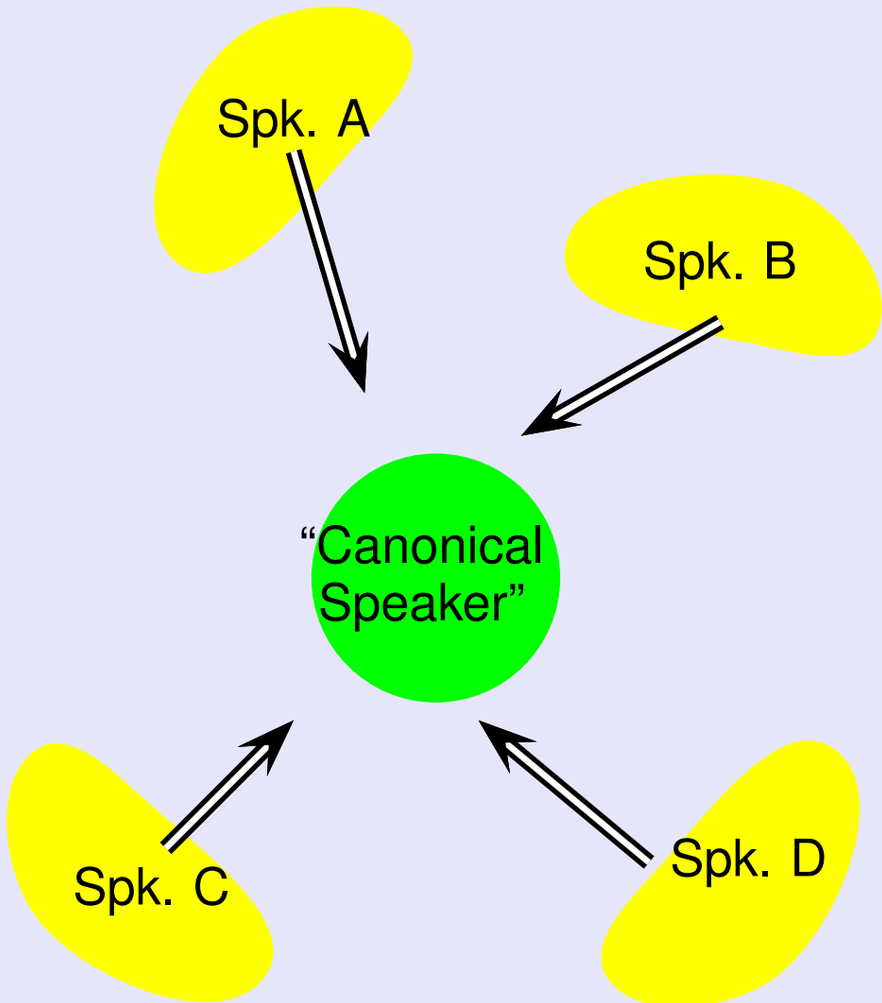
Normalizzazione del parlatore

Lo spazio delle features di ciascun parlatore viene alterato per avvicinarsi a quello del sistema speaker independent



Normalizzazione del parlatore in addestramento adattivo

Lo spazio delle features di ciascun parlatore viene alterato per avvicinarsi a quello di un modello “canonico”, anche in addestramento. I modelli lavorano solo su dati normalizzati.



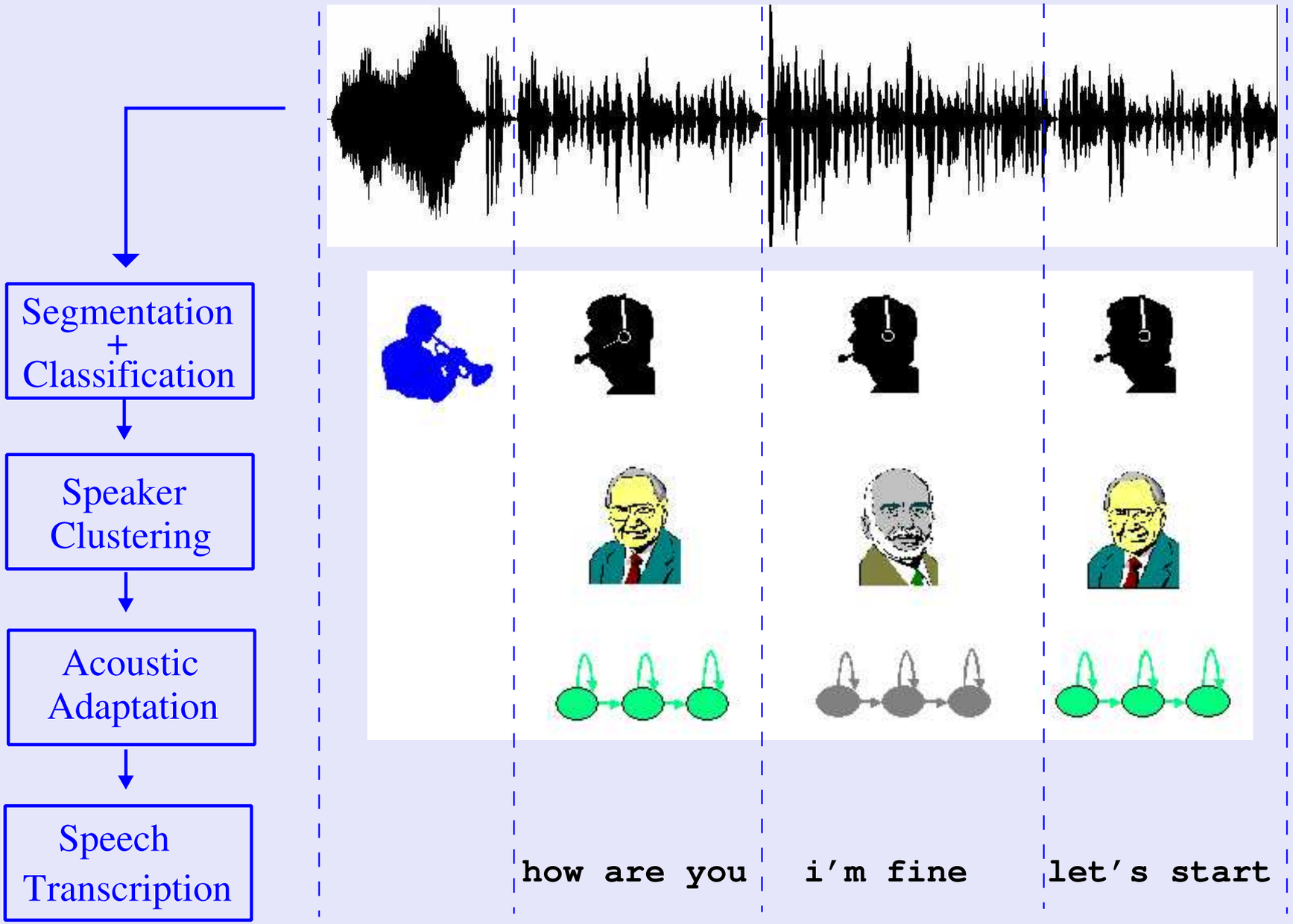
Riconoscimento a più passi

L'adattamento è efficace, ma bisogna diversificare i parlatori, e avere dati di adattamento trascritti. Come si fa in fase di riconoscimento?

Sono necessari:

- ▷ Un nuovo modulo di segmentazione e classificazione, che separa porzioni diverse dell'audio, e raggruppa porzioni omogenee, in modo da poterle considerare obiettivi per l'adattamento
- ▷ Una trascrizione preliminare dell'input, da usare come se fosse esatta per l'adattamento

Un sistema di trascrizione



Riferimenti Bibliografici

L.R. Rabiner and B.W. Juang, *Fundamentals of Speech Recognition*,
Prentice-Hall, 1993, ISBN: 0-13-015157-2

F. Jelinek, *Statistical Methods for Speech Recognition*,
MIT Press, 1998, ISBN: 0-262-10066-5

R. De Mori, (Ed.), *Spoken Dialogues with Computers*,
Academic Press, 1998, ISBN: 0-12-209055-1

S. Furui, *Digital Speech Processing, Synthesis, and Recognition*,
Marcel Dekker, 2000, ISBN: 0-8247-0452-5

X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing*,
Prentice Hall, 2001, ISBN: 0-13-022616-5

Demo

- Trascrizione di sedute della Camera dei Deputati
- Trascrizione (e traduzione) di sessioni plenarie del Parlamento Europeo

Demo: Camera dei Deputati

Modello Acustico:

- Addestramento adattivo con tecnica CMLSN
- 2 passi di riconoscimento: uno con modelli SI e uno con modelli normalizzati
- Entrambi gli insiemi di modelli hanno:
  - $\sim$ 18.000 unità context-dependent
  - $\sim$ 50.000 Gaussiane, condivise da $\sim$ 11.000 misture con tying
- addestrati su $\sim$110 ore di parlato, per la maggior parte trascritto automaticamente

Modello del linguaggio:

- Dizionario di $\sim$ 64000 parole
- Quadrigrammi, stimati sui resoconti di quattro anni di legislatura: $\sim$ 40 milioni di parole

Prestazioni: circa 11% Word Error Rate

Demo: EPPS

Modello Acustico:

- Addestramento adattivo con tecnica CMLSN, unsupervised per il primo passo, supervised per i seguenti
- Proiezione discriminativa nello spazio delle features (HLDA)
- 3 passi di riconoscimento: uno con modelli SI e due con modelli normalizzati.
- Entrambi gli insiemi di modelli hanno:
  - $\sim$ 24.000 unità context-dependent, con $\sim$ 8000 stati condivisi
  - $\sim$ 250.000 Gaussiane,
- addestrati su $\sim$150 ore di parlato, per circa metà trascritto automaticamente

Modello del linguaggio:

- Dizionario di $\sim$ 50000 parole
- Trigrammi per i primi due passi, quadrigrammi per il terzo
- stimati su $\sim$ 200 milioni di parole

Prestazioni: circa 13% Word Error Rate

Grazie per l'attenzione