

L'analisi testuale automatica di note cliniche: *approcci, applicazioni e opportunità*

Francesco Ronzano

Integrative Biomedical Informatics Group

Hospital del Mar Medical Research Institute & Universitat Pompeu Fabra
Barcelona, Spain

Recenti esperienze di ricerca in NLP



**Universitat
Pompeu Fabra**
Barcelona

TALN
Natural Language Processing
Research Group

- Estrazione di informazione e creazione di riassunti di letteratura scientifica
- Analisi del sentimento e del linguaggio figurativo in social media



Institut Hospital del Mar
d'Investigacions Mèdiques



IBI Group

Integrative Biomedical
Informatics Group

- Estrazione di informazione da note cliniche
- Analisi di letteratura scientifica biomedica

Sommario

- Fonti di informazione medica: le note cliniche
- L'analisi del linguaggio delle note cliniche
 - Identificazione di abbreviazioni e acronimi
 - Caratterizzazione di negazioni
 - Identificazione e normalizzazione di concetti
 - Analisi di informazione temporale
- Esempi di applicazioni
 - Fenotipizzazione dei pazienti
 - Supporto alle decisioni cliniche
 - Studio di comorbidità
- Privacy e anonimizzazione
- Risorse e framework
- Conclusioni

Fonti di informazione medica: le note cliniche



Letteratura scientifica
(bio)medica



Social Media



Note cliniche

Elaborazione del Linguaggio nel campo medico

**Medicina
personalizzata**

**Monitoraggio
socio-sanitario**

**Identificazione
di effetti avversi**

**Supporto alle
decisioni cliniche**

**Ottimizzazione
dei costi**

**Fenotipizzazione
di pazienti**

**Monitoraggio di
pratiche cliniche**

**Prevenzione del
rischio**

Fonti di informazione medica: le note cliniche

DIGITALIZZAZIONE DEI SERVIZI SANITARI



**OTTIMIZZAZIONE GESTIONALE E
AMMINISTRATIVA PER IL SISTEMA
SANITARIO E PER IL PAZIENTE**

Spesa per sanità digitale

Italia : 21 euro per abitante
Francia: 40 euro per abitante
Inghilterra : 60 euro per abitante
Danimarca: 70 euro per abitante



USI SECONDARI

Identificazione
di e

Fenotipizzazione

Supporto alle
decisioni

Monitoraggio di
pratiche cliniche

Spesa per sanità digitale

- Cartella Clinica Elettornica
- Fascicolo Sanitario Elettronico

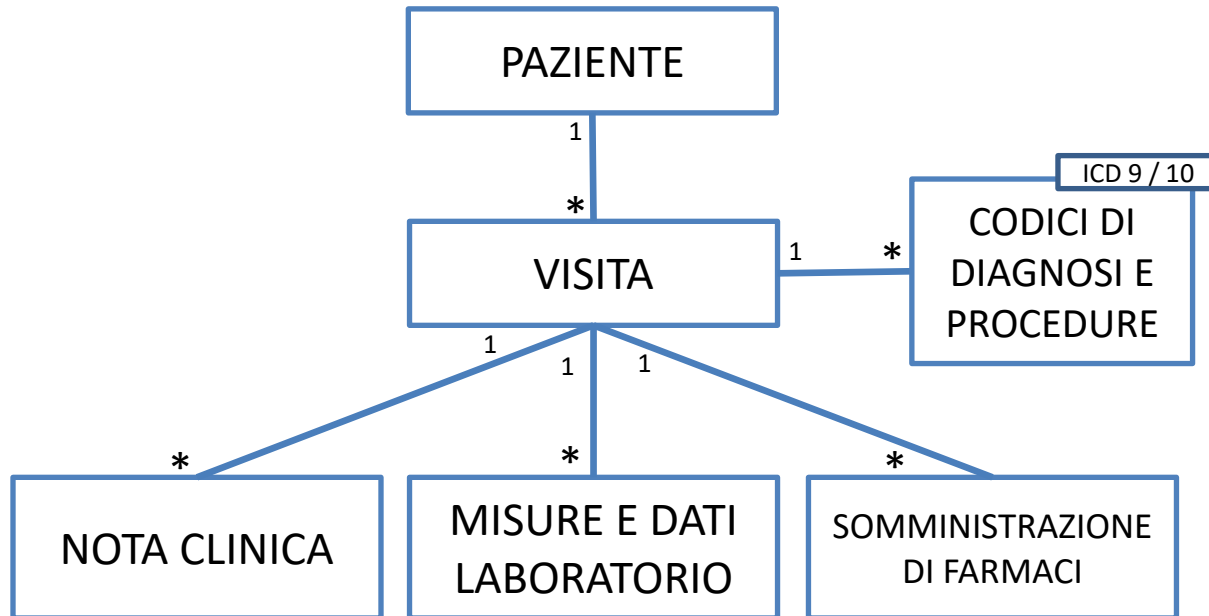
Fonti di informazione medica: le note cliniche

Il database ospedaliero anonimizzato IMASIS



Institut Hospital del Mar
d'Investigacions Mèdiques

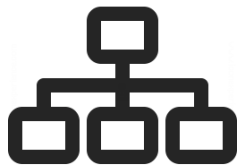
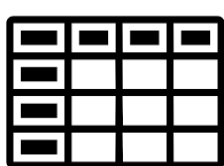
- HOSPITAL DEL MAR
- HOSPITAL DE L'ESPERANÇA
- CENTRE FÒRUM
- CUAP PERACAMPS
- CENTRE DR. EMILI MIRA



- 1,5 milioni di pazienti
- 15,8 milioni di visite
- 7 milioni di note cliniche testuali

Fonti di informazione medica: le note cliniche

Informazione strutturata



Informazione non strutturata (testo libero)



Una parte essenziale dell'informazione clinica digitalizzata viene codificata per mezzo di **campi di testo libero** detti **note cliniche**

L'uso del testo libero:

- permette **descrizioni più dettagliate / personalizzate**
- si adatta meglio al **ritmo spesso serrato della compilazione dei referti clinici**
- evita le **situazioni di mancanza di espressività** dovute alla strutturazione dell'informazione clinica
- rende più agile **l'interazione medico-paziente**



Le tecniche di Elaborazione del Linguaggio Naturale costituiscono strumenti essenziali per processare efficacemente l'informazione clinica

Fonti di informazione medica: le note cliniche

NOTA CLINICA

documento **testuale**

informazioni circa il
**percorso diagnostico-
terapeutico**
di un **paziente**, al fine di
determinare
le cure da somministrare

Rogonfiamento recente della regione parotidea sinistra.
Lindagine radiologica rileva in corrispondenza del polo caudale della parotide una formazione rotondeggiante ipoecogena. La ghiandola sottomandibolare di sn. Appare regolare con linfonodo vicino al polo dorsale. Linfonodo infiammato parotideo inferiore a sinistra con possibile evoluzione verso la colliquazione. Effettuare ulteriore controllo.

componente essenziale
dell'attività di **refertazione
medica**

redatto da **personale socio-
sanitario**

Una nota clinica può formar parte di:

- Cartelle cliniche ospedaliere
- Visite e referti specialistici
- Medico di famiglia
- Operatori socio-sanitari

*2,500 note cliniche al giorno
nel Parc de Salut Mar
di Barcellona
(1,000 / minuto in Spagna)*

L'analisi del linguaggio delle note cliniche

Estrarre efficacemente informazione strutturata dalle note cliniche costituisce **un passo previo essenziale per il loro uso sistematico**

AP: fumatore 2pt/g, HTA, AMI nel 2010, pancreat. nel 2011, no abuso alcolici
Padre e fratello diabetici. Padre iperteso.

Dolori addominali persistono. Richiesta valutazione eco.

Abomil: 1 cps. ogni 12 ore

Programmata visita di controllo.

ANTECEDENTI PATOLOGICI

Tabagismo

2 pacchetti
/giorno

Ipertensione arteriale

Infarto acuto miocardio

2010

Pancreatite

2011

ANTECEDENTI FAMILIARI

Ipertensione a.

PADRE

Diabete

FRATELLO

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa criticità e peculiarità**

AP: **fumatorre** 2pt/g, HTA, AMI nel 2010,
pancreat. nel 2011, no abuso alcolici
Padre e fratello diabetici.

Errori ortografici

Afebrile, BP 80-130
Descrive **dolorabilita** addominale diffusa.
Sospetto intossicazione alimentare.
Somministro SF. Richiedo eco addominale.

Pracetamololo 1g / 8h per 5 gg.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo fra 2 settimane.

La ghiandola sottomandibolare di sn. appare
regolare con linfonodo vicino al p. dorsale. Si
scarta processo infiammatorio del ganglio.

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa criticità e peculiarità**

AP: **fumatorre** 2pt/g, HTA, AMI nel 2010,
pancreat. nel 2011, **no** abuso alcolici
Padre e fratello diabetici.

Errori ortografici

Negazione

Afebrile, BP 80-130
Descrive **dolorabilità** addominale diffusa.
Sospetto intossicazione alimentare.
Somministro SF. Richiedo eco addominale.

Pracetamololo 1g / 8h per 5 gg.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo fra 2 settimane.

La ghiandola sottomandibolare di sn. appare
regolare con linfonodo vicino al p. dorsale. **Si**
scarta processo infiammatorio del ganglio.

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa peculiarità**

AP: **fumatore** 2pt/g, HTA, AMI nel 2010,
pancreat. nel 2011, **no** abuso alcolici
Padre e fratello diabetici.

Afebrile, BP 80-130
Descrive **dolorabilità** addominale diffusa.
Sospetta intossicazione alimentare.
Somministra SF. Richiedo **eco** addominale.

Pracetamol 1g / 8h per 5 gg.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo fra 2 settimane.

La ghiandola sottomandibolare di **sn.** appare
regolare con linfonodo vicino al **p.** dorsale. **Si**
scarta processo infiammatorio del ganglio.

Errori ortografici

Negazione

Abbreviazioni

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa peculiarità**

AP: **fumatore** 2pt/g, **HTA**, **AMI** nel 2010,
pancreat. nel 2011, **no** abuso alcolici
Padre e fratello diabetici.

Afebrile, **BP** 80-130
Descrive **dolorabilità** addominale diffusa.
Sospetto intossicazione alimentare.
Somministro **SF**. Richiedo **eco** addominale.

Pracetamololo 1g / 8h per 5 gg.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo fra 2 settimane.

La ghiandola sottomandibolare di **sn.** appare
regolare con **linfonodo** vicino al **p.** dorsale. **Si**
scarta processo infiammatorio del **ganglio**.

Errori ortografici

Negazione

Abbreviazioni

Sinonimia

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa criticità e peculiarità**

AP: **fumatore** 2pt/g, HTA, AMI nel 2010,
pancreat. nel 2011, **no** abuso alcolici
Padre e fratello diabetici.

Afebrile, BP 80-130
Descrive **dolorabilità** addominale diffusa.
Sospetto intossicazione alimentare.
Somministro SF. Richiedo **eco** addominale.

Pracetamololo 1g / 8h per 5 gg.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo fra 2 settimane.

La ghiandola sottomandibolare di **sn.** appare
regolare con **linfonodo** vicino al **p.** dorsale. **Si**
scarta processo infiammatorio del **ganglio**.

Errori ortografici

Negazione

Abbreviazioni

Sinonimia

Espressioni di incertezza

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa peculiarità**

AP: **fumatore** 2pt/g, **HTA**, **AMI** nel 2010,
pancreat. nel 2011, **no** abuso alcolici
Padre e fratello diabetici.

Afebrile, **BP** 80-130
Descrive **dolorabilità** addominale diffusa.
Sospetto intossicazione alimentare.
Somministro **SF**. Richiedo **eco** addominale.

Pracetamololo 1g / 8h **per** 5 gg.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo **fra** 2 settimane.

La ghiandola sottomandibolare di **sn.** appare
regolare con **linfonodo** vicino al **p.** dorsale. **Si**
scarta processo infiammatorio del **ganglio**.

Errori ortografici

Negazione

Abbreviazioni

Sinonimia

Espressioni di incertezza

Espressioni temporali

L'analisi del linguaggio delle note cliniche

L'analisi automatica del linguaggio delle note cliniche
presenta **numerosa peculiarità**

AP: **fumatore** 2pt/g, HTA, AMI nel 2010,
pancreat. nel 2011, **no** abuso alcolici
Padre e fratello diabetici.

Afebrile, BP 80-130
Descrive **dolorabilità** addominale diffusa.
Sospetto intossicazione alimentare.
Somministro **SF**. Richiedo **eco** addominale.

Pracetamololo 1g / 8h **per 5 gg**.
Pom. oft. TININ / 12 h fino a regressione
gonfiore. Controllo **fra 2 settimane**.

La ghiandola sottomandibolare di **sn.** appare
regolare con **linfonodo** vicino al **p.** dorsale. **Si**
scarta processo infiammatorio del **ganglio**.

Errori ortografici

Negazione

Abbreviazioni

Sinonimia

Espressioni di incertezza

Espressioni temporali

Informazioni di terzi

...dosaggio medicinali, mix di lingue

Abbreviazioni

Le note cliniche sono **uno dei tipi di testo che contengono più abbreviazioni e acronimi**, spesso creati arbitrariamente e di difficile interpretazione

AP: fumatore 2pt/g, HTA, AMI nel 2010,
pancreat. nel 2011, no abuso alcolici
Padre e fratello diabetici.

La caratterizzazione delle abbreviazioni si compone di due fasi:

- Identificazione dell'abbreviazione

AP: fumatore 2pt/g, HTA, AMI nel 2010...

- Disambiguazione

AP: fumatore 2pt/g, HTA, AMI nel 2010...

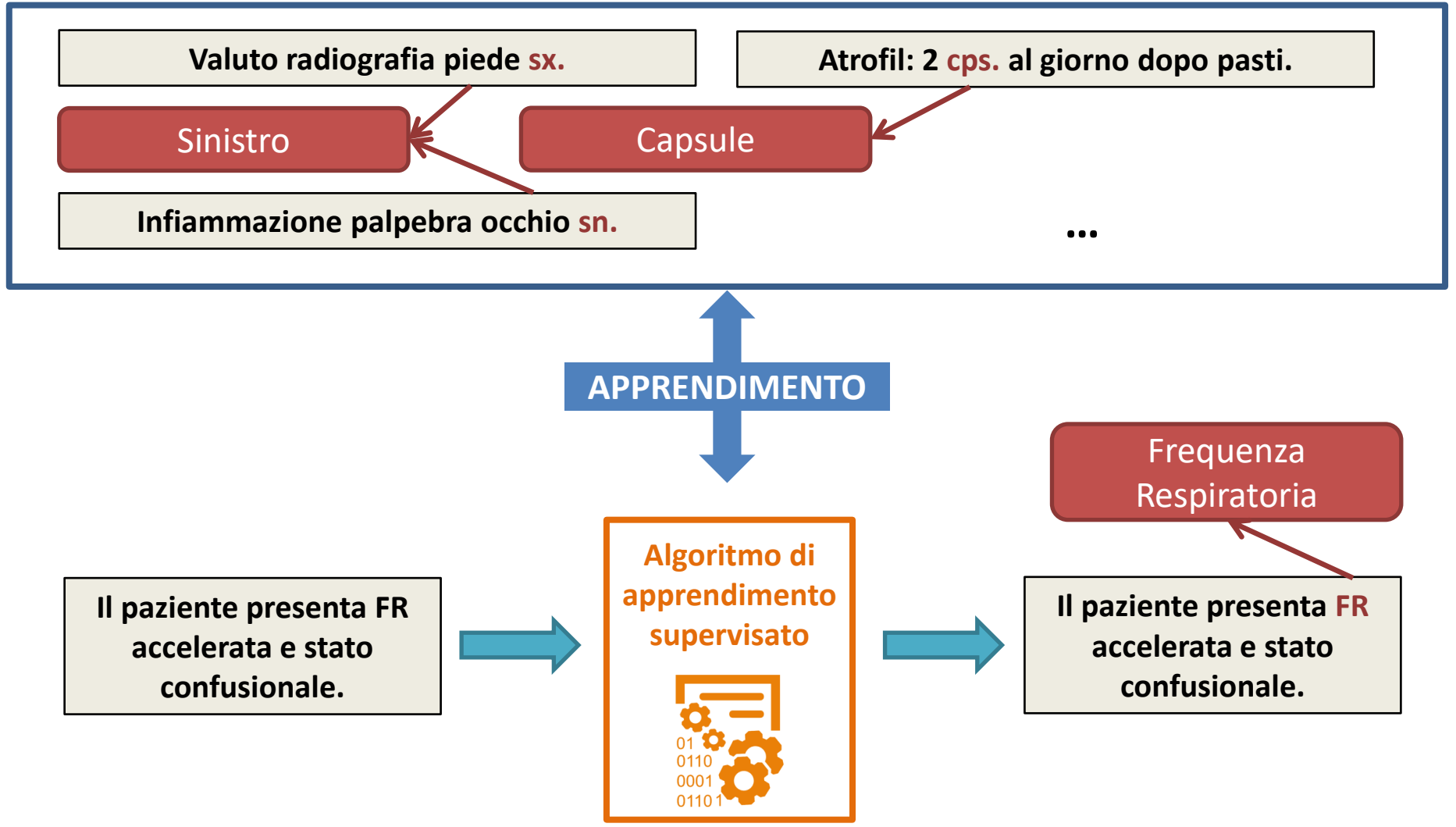
Antecedenti
Patologici

Potenziale d'Azione

Antero Posteriore

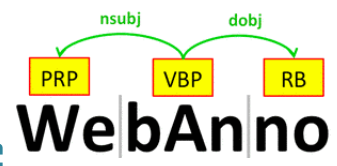
Abbreviazioni

Apprendimento supervisionato → Schema generale



Abbreviazioni

Apprendimento supervisionato → 1) Annotazione corpus di abbreviazioni



Corpus

Valuto radiografia piede **sx**.

Atrofil: 2 **cps**. al giorno dopo pasti.

Sinistro

Capsule

Risonanza Magnetica Nucleare

Infiammazione palpebra occhio **sn**.

Il paziente consegna il referto della **RMN**.

Paziente indirizzato da **CO** regionale.

Esame della **CO** rivela difficoltà di...

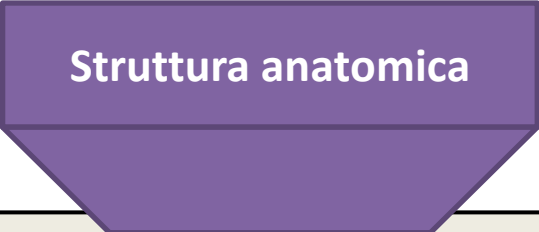
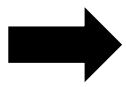
Centrale Operativa

Conduzione Ossea

...

Abbreviazioni

Apprendimento supervisionato → 2) Analisi automatica del testo



Il paziente consegna il referto della RMN. Evidenza di lesione della cartilagine del ginocchio sn. ...

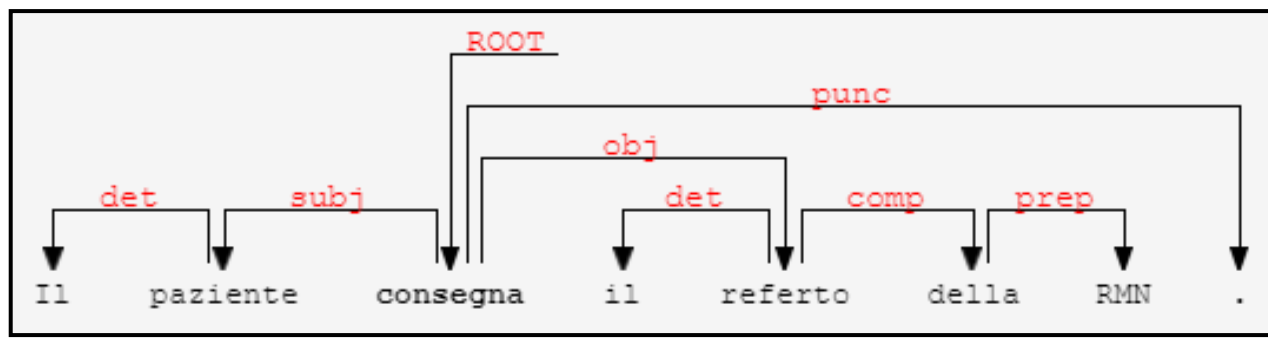
FRASE 1

FRASE 2

...

Il paziente consegna il referto della RMN .

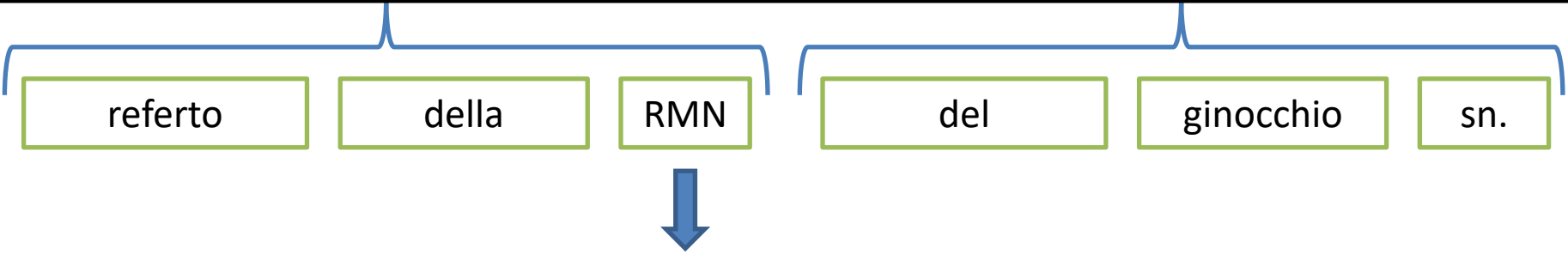
Articolo Sostantivo Verbo Articolo Sostantivo Preposizione Sostantivo Punt.



Abbreviazioni

Apprendimento supervisionato → 3) Generazione vettori di caratteristiche

Il paziente consegna il **referto della RMN**. Evidenza di lesione della cartilagine **del ginocchio sn.** ...



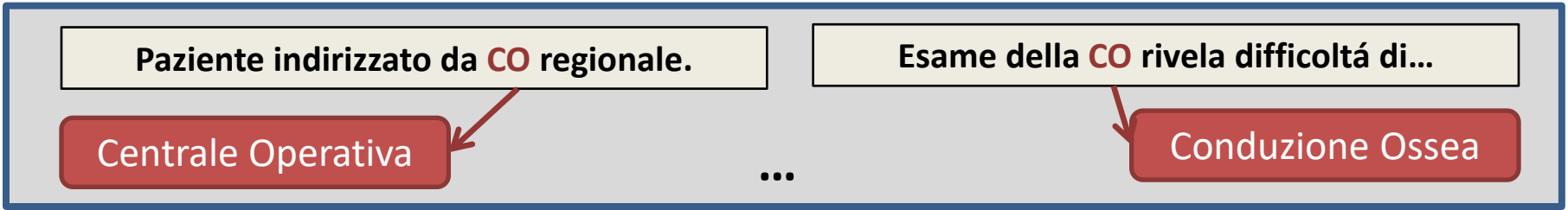
Numero caratteri	3
% caratteri maiuscoli	1
% caratteri numerici	0
% caratteri punteggiatura	0
Primo / ultimo carattere maiuscolo?	Sí/ Sí
Ultimo carattere punteggiatura?	No
Numero di ripetizioni della parola nel testo	4
Presente nel dizionario medico?	Sí
Parte del discorso	Sostantivo
Parte del discorso dell'head	Preposizione

Vettore di
caratteristiche

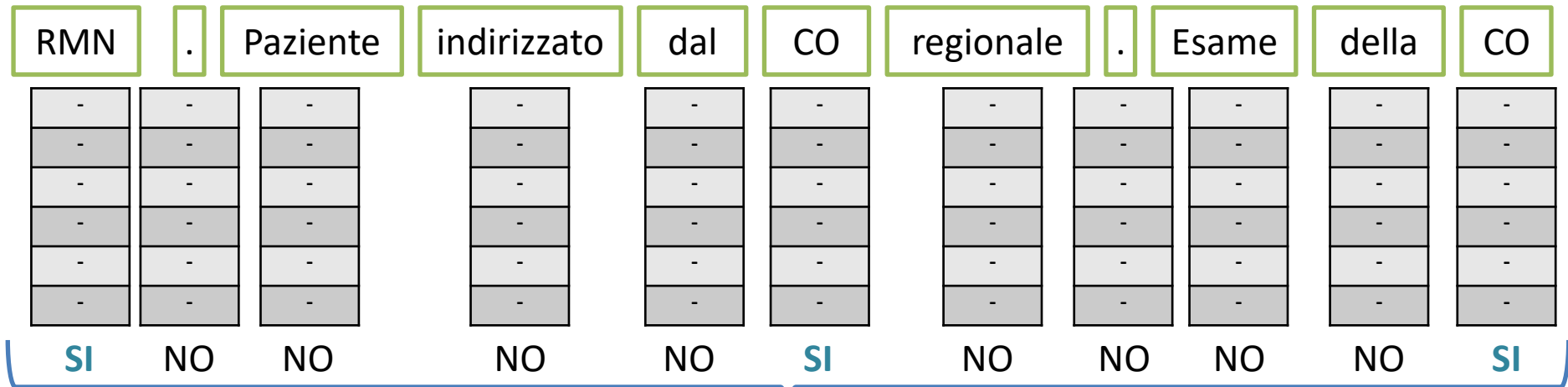
Abbreviazioni

Apprendimento supervisionato → 4) Training algoritmo

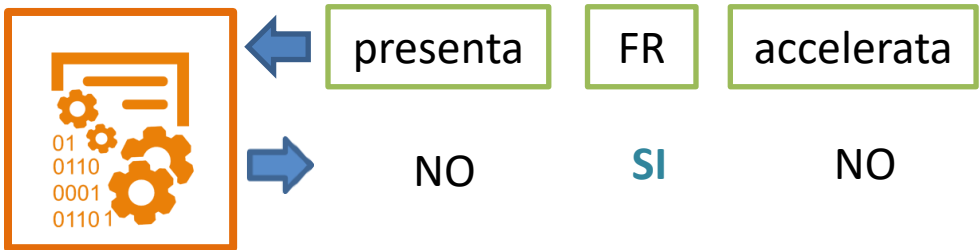
Corpus



Esempi di apprendimento



- Support Vector Machine
- Conditional Random Field
- Neural Networks (LSTM)



Abbreviazioni

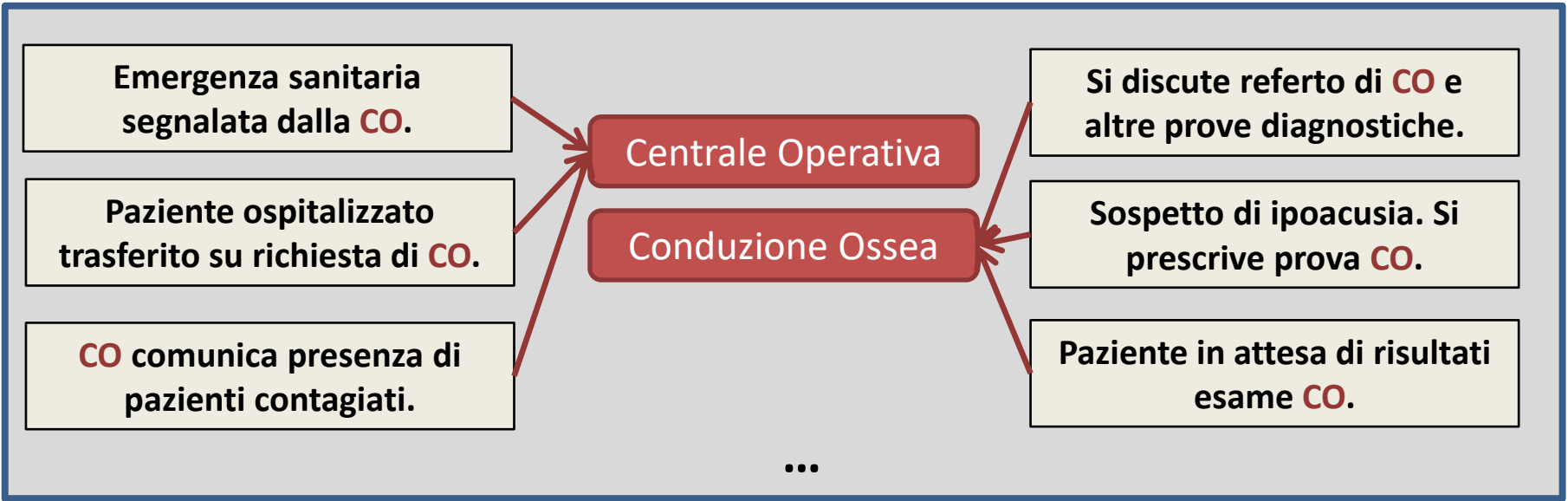
Apprendimento supervisionato → 5) Applicazione algoritmo e disambiguazione



A cosa si riferisce la abbreviazione **CO**?

Corpus

Disambiguazione basata sulla similarità del contesto



Negazioni

Nelle note cliniche quasi la metà dei sintomi sono negati

AP: fumatore 2pt/g, HTA, AMI nel 2010, pancreat. nel 2011, **no** abuso alcolici

Afebrile, BP 80-130
Si scarta processo infiammatorio ganglio.

NegEx / ContTex:

regole applicate al contesto delle espressioni di cui si vuole determinare se sono negate

Contesto della negazione

Il paziente **nega** tosse, **ma** riferisce **forte irritazione della gola**

Trigger di
negazione

Terminatore
di negazione

Proposti miglioramenti grazie all'uso di:

- Relazioni grammaticali
- Word embeddings

Identificazione e normalizzazione di concetti

L'identificazione e normalizzazione di concetti rappresenta un passo essenziale nella normalizzazione di informazione clinica

Riferisce miglioramento del **dolore addominale**.

Dolore della zona toracica

Dolore addominale

Lombalgia

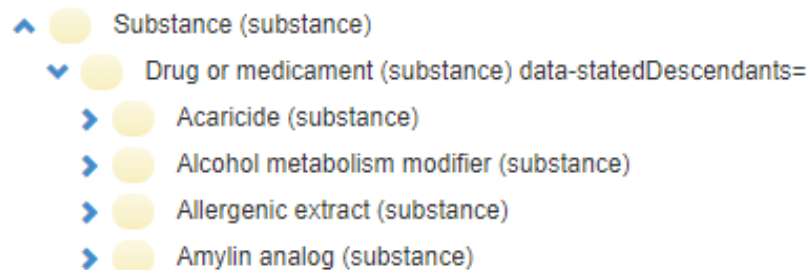
Terminologie

SNOMED CT

- + 300,00 concetti di terminologia clinica
- organizzati gerarchicamente
- disponibili estensioni in varie lingue
- dal 2007, mantenuta da organizzazione no-profit IHTSDO

RxNORM

- normalizzazione nomi di farmaci
- specifica degli USA
- mantenuta dalla NLM



'Naproxen 250 MG Oral Tablet'

'Naproxen' / 'Naproxen 250 MG' / 'Naproxen Oral Tablet'
'Naproxen Oral Products' / 'Naproxen Pills'

Identificazione e normalizzazione di concetti

L'identificazione e normalizzazione di concetti rappresenta un passo essenziale nella normalizzazione di informazione clinica

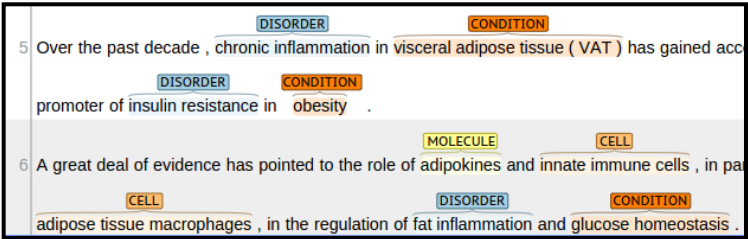
Riferisce miglioramento del dolore addominale.

2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records

- 871 note cliniche annotate
- **Classi di concetti:** problemi medici, trattamenti, prove cliniche
- **Caratterizzazione di concetti:** presente, assente, possibilmente presente, presente in persone differenti dal paziente, etc.
- **Relazioni tra concetti:** trattamento amministrato per curare problema medico, trattamento causa problema medico, etc.

Algoritmi di apprendimento supervisionato o ibridi





- **Identificazione classi di conc.:** .88-.92 F-score
- **Caratterizzazione di conc.:** .91-.93 F-score
- **Relazioni tra conc.:** .65-.73 F-score

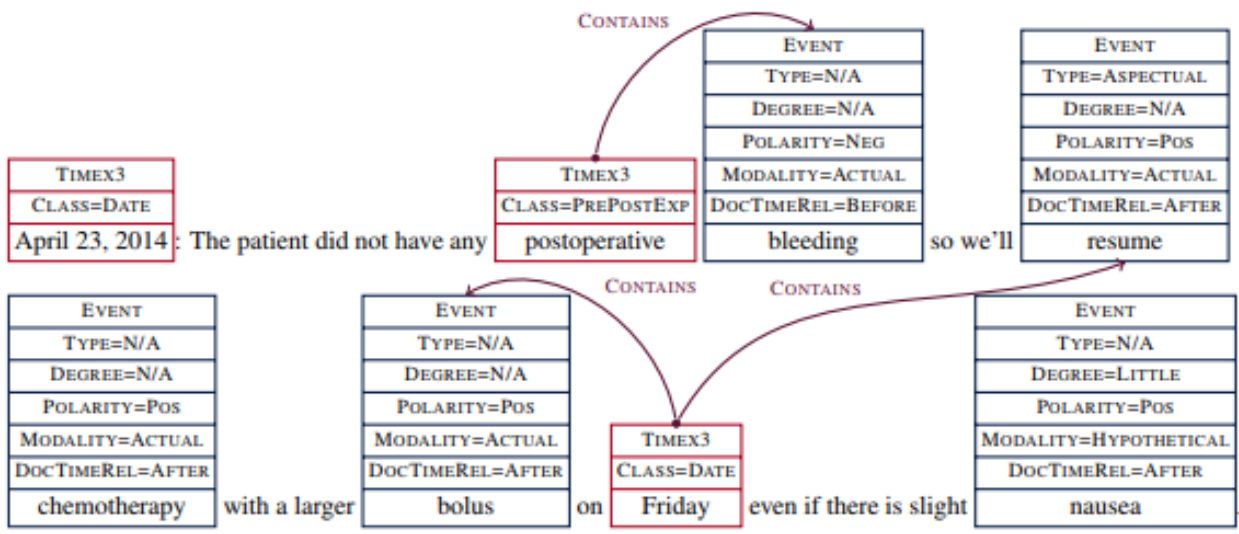
Informazione temporale

La corretta identificazione e interpretazione delle espressioni temporali é essenziale per contestualizzare il contenuto delle note cliniche

Clinical TempEval Challenge @ SemEval 2017

600 note cliniche di pazienti con cancro

- **Tipi di espressione temporale:** DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP, or SET
- **Eventi caratterizzati da:** grado di certezza, polarità e tipo
- **Relazioni:** “espressioni temporali – data nota clinica” ed “eventi – espressioni temporali”



- **Identificazione espressioni temporali:** .31-.59 F-score
- **Caratterizzazione di eventi:** .62-.72 F-score
- **Relazioni tra evento e espressione temporale:** .14-.33 F-score

Esempi di applicazioni

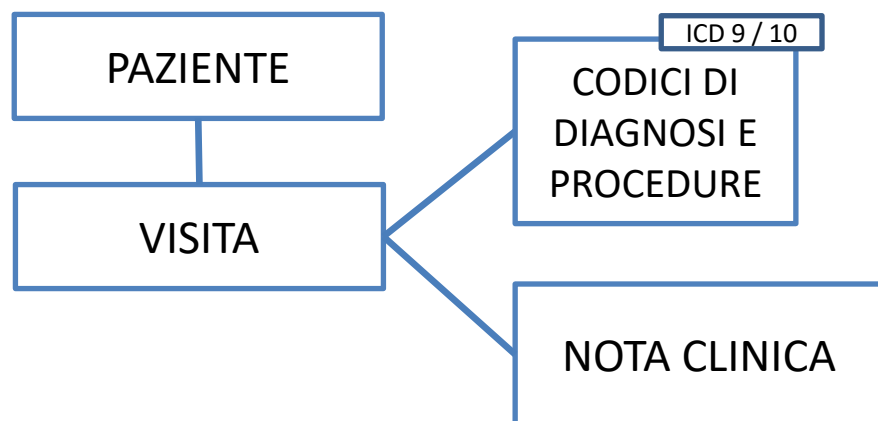
Numerosi tipi di uso secondario dell'informazione testuale delle note cliniche sono stati proposti negli ultimi anni

- Fenotipizzazione dei pazienti
- Supporto alle decisioni cliniche
- Studio di comorbidità
- Assegnazione di codici ICD-9 / 10
- Previsione del rischio di riammissione
- Identificazione di tossicità ed effetti collaterali di farmaci
- Studio di aderenza alle linee guida cliniche
- Creazione di riassunti visuali di storie cliniche
- etc.

Fenotipizzazione dei pazienti

Determinare se un paziente presenta una condizione clinica specifica

- Selezione pazienti per studi clinici
- Analisi di dati medici circoscritte a insiemi specifici di pazienti



- 250 Diabete mellito
- 402 Cardiopatia ipertensiva

...controllo. Paziente diabetico (DM II) e iperteso. Si richiede analisi sangue per valutazione...

Complementare l'informazione clinica strutturata mediante l'analisi del testo delle note cliniche permette di migliorare l'accuratezza della fenotipizzazione di pazienti

Fenotipizzazione dei pazienti

Determinare se un paziente presenta una specifica condizione clinica

MIMIC-III - note cliniche di dimissione

1,610 note cliniche di dimissione paziente annotate manualmente con uno di 10 fenotipi

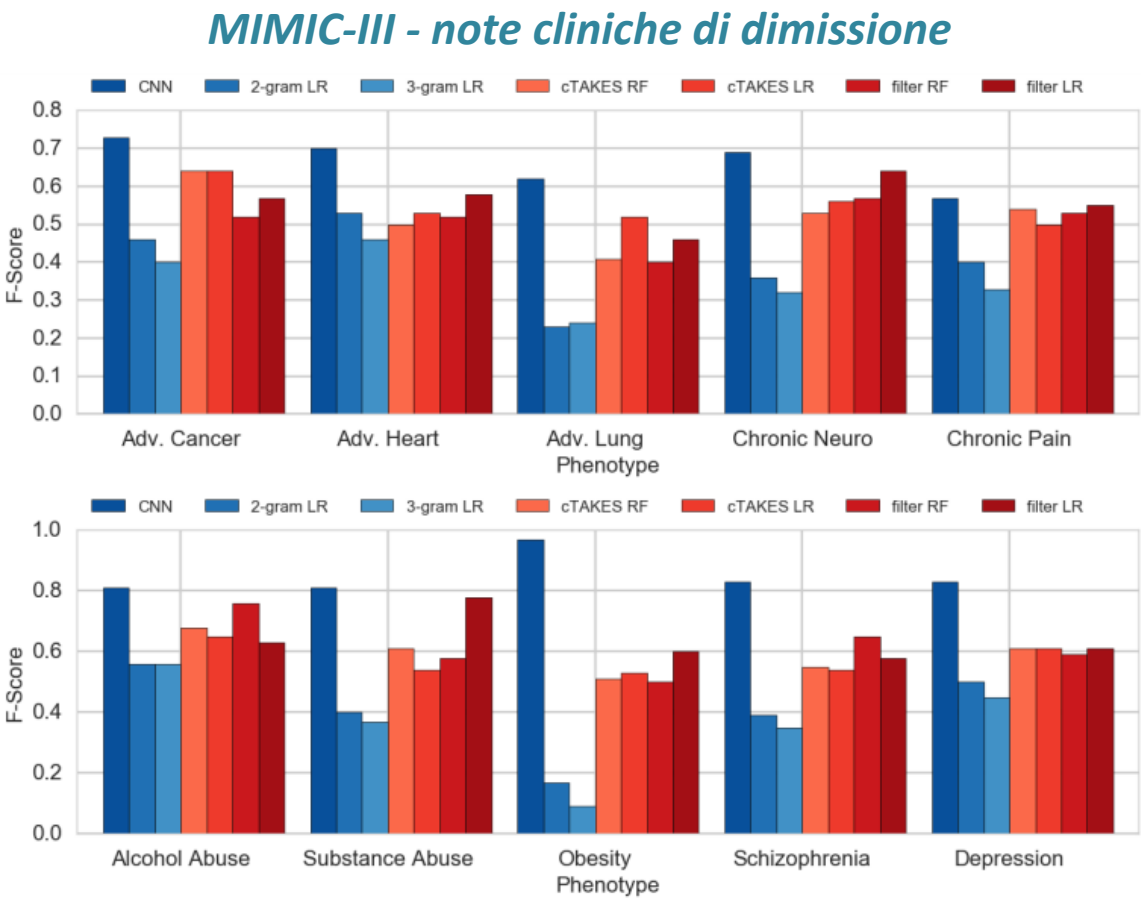
Phenotype	#positive
Adv. / Metastatic Cancer	161
Adv. Heart Disease	275
Adv. Lung Disease	167

Phenotype	#positive
Chronic Neurological Dystrophies	368
Chronic Pain	321
Alcohol Abuse	196
Substance Abuse	155

Phenotype	#positive
Obesity	126
Psychiatric Disorders	295
Depression	460

Fenotipizzazione dei pazienti

Determinare se un paziente presenta una specifica condizione clinica

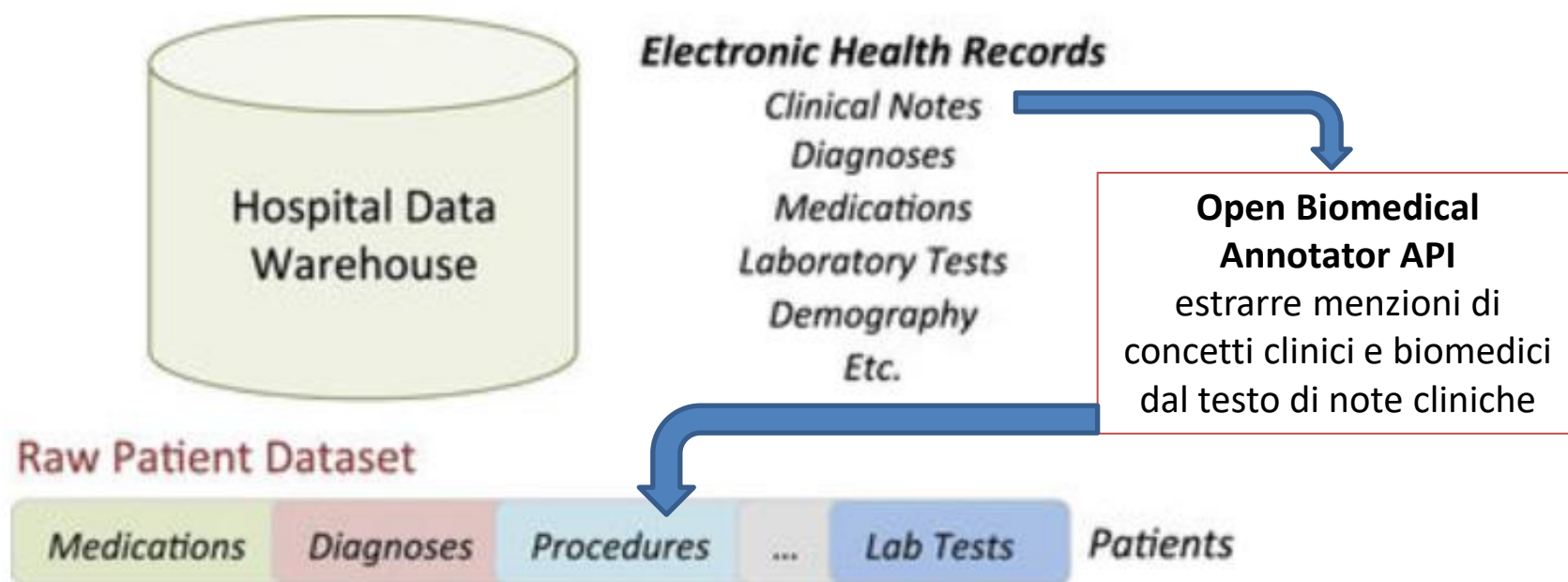


Supporto alle decisioni cliniche

Determinare il rischio o prevedere l'evoluzione clinica di un paziente o gruppo di pazienti

- **Identificare pazienti a rischio da monitorare più strettamente**
- **Modellare l'evoluzione delle patologie**

- Deep patient -

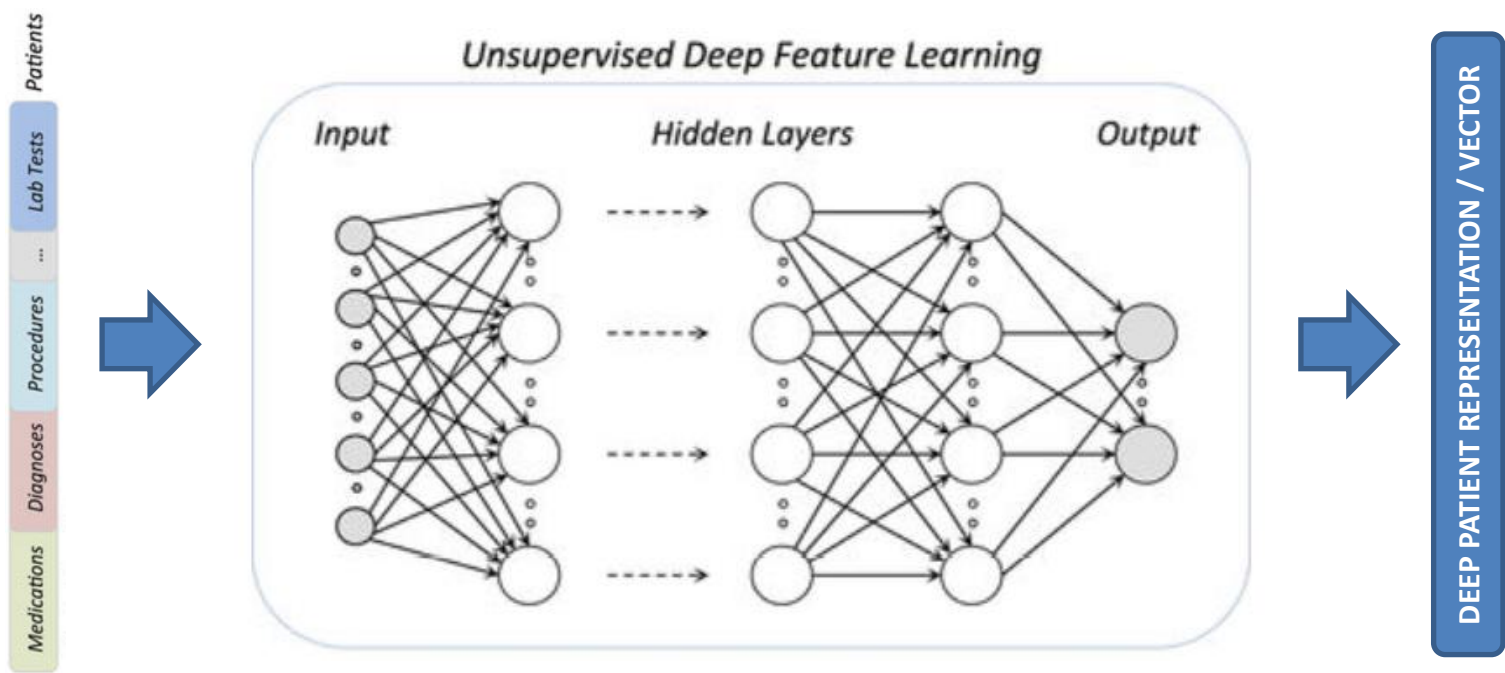


Supporto alle decisioni cliniche

Determinare il rischio o prevedere l'evoluzione clinica di un paziente o gruppo di pazienti

- **Identificare pazienti a rischio da monitorare più strettamente**
- **Modellare l'evoluzione delle patologie**

- Deep patient -

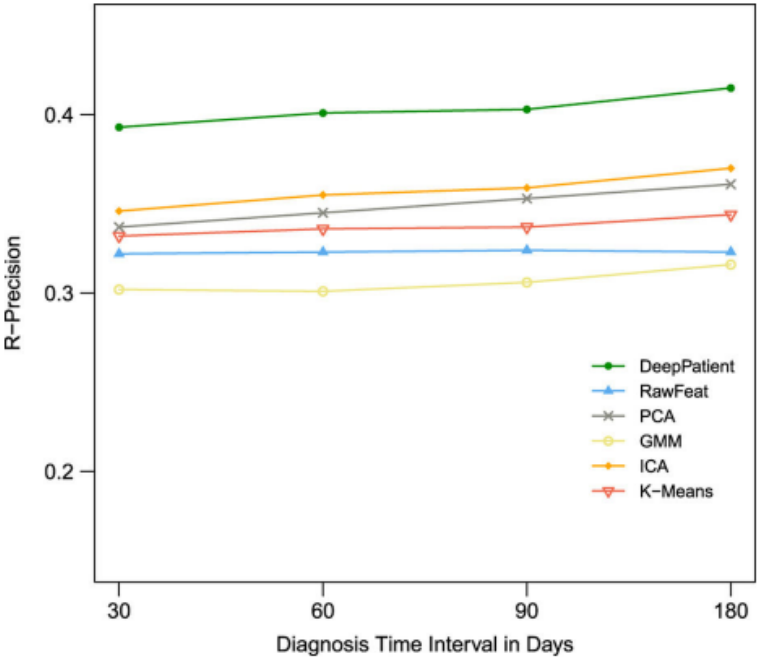
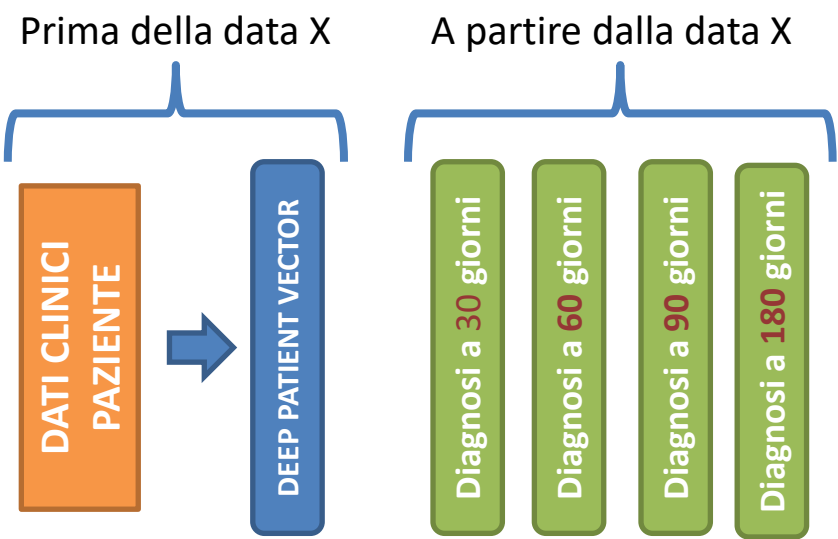


Supporto alle decisioni cliniche

Determinare il rischio o prevedere l'evoluzione clinica di un paziente o gruppo di pazienti

- **Identificare pazienti a rischio da monitorare più strettamente**
- **Modellare l'evoluzione delle patologie**

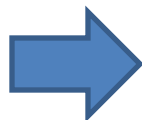
- Deep patient -



Studio di comorbidità

Studio su una popolazione di pazienti di coesistenze significative di patologie differenti in uno stesso individuo

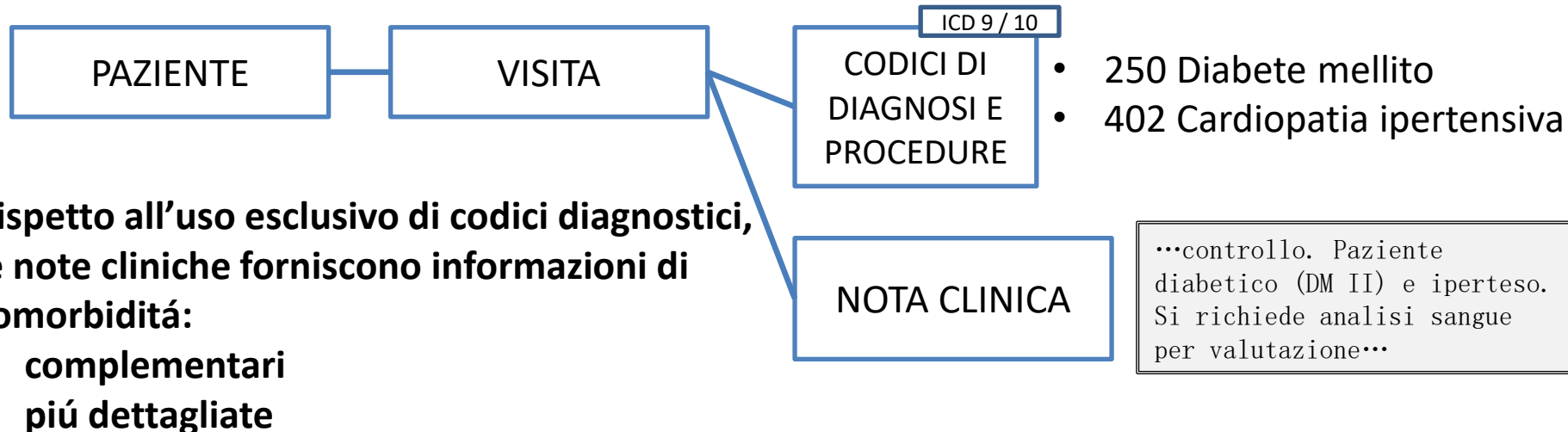
- Valutazione rapporto costi/benefici nel trattamento di patologie
- Modellare il rischio di sviluppo di patologie



DIABETE



IPERTENSIONE



Privacy e anonimizzazione

Differenti approcci sono stati proposti per identificare automaticamente informazione sensibile nelle note cliniche

2014 i2b2/UTHealth corpus Patient deidentification challenge

- Corpus con 1,304 note cliniche di 396 pazienti
- Deidentificate sostituendo valori semanticamente consistenti ma inventati

- NAME
 - PATIENT, DOCTOR, USERNAME
- PROFESSION
- LOCATION
 - ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER
- AGE
 - over 90, under 90
- DATE
- CONTACT
 - PHONE, FAX, EMAIL, URL, IPADDRESS
- IDs
 - SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER
- OTHER

Nota clinica deidentificata

Record date: <DATE>2074-04-05</DATE>

<PHI TYPE="HOSPITAL">EMMANUEL HOME</HOSPITAL> EMERGENCY DEPT VISIT

<PATIENT>JACOB,LARRY</PATIENT>

<MEDICALRECORD>910-66-83-7</MEDICALRECORD>

VISIT DATE: <DATE>04/05/74</DATE>

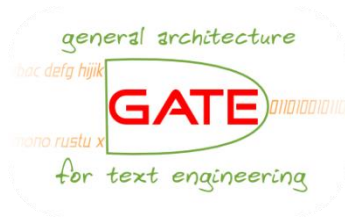
This patient was seen, interviewed and examined by myself as well as Dr. <DOCTOR>Naylor</DOCTOR> whose I have reviewed and whose findings I have confirmed.

HISTORY OF PRESENTING COMPLAINT: This is a <AGE>53</AGE>-year-old male who

➤ **Average performance:
.90 F-score**

Risorse e framework

Framework di Elaborazione del Linguaggio Naturale



Framework di Apprendimento Automatico



Framework di Salvataggio / Indicizzazione / Visualizzazione Dati



Conclusioni

- L'analisi del linguaggio delle note cliniche presenta **numerosa peculiarità** da affrontare per estrarre informazione affidabile
- **Numerosi sono le possibili applicazioni** dell'informazione estratta
- L'informazione sensibile contenuta nelle note cliniche rende **più lenta la sperimentazione di metodi di analisi**
- Nonostante i tentativi di strutturazione, **le note cliniche in testo libero restano una componente essenziale della refertazione medica**
- La maggior parte delle risorse / esperimenti per l'elaborazione del linguaggio delle note cliniche riguarda testi in **l'inglese**



Institut Hospital del Mar
d'Investigacions Mèdiques

Grazie per l'attenzione!

L'analisi testuale automatica di note cliniche: *approcci, applicazioni e opportunità*

Francesco Ronzano

Integrative Biomedical Informatics Group

Hospital del Mar Medical Research Institute & Universitat Pompeu Fabra

Barcelona, Spain

Lingua e Computer – Università degli Studi della Basilicata – 29 Marzo 2018