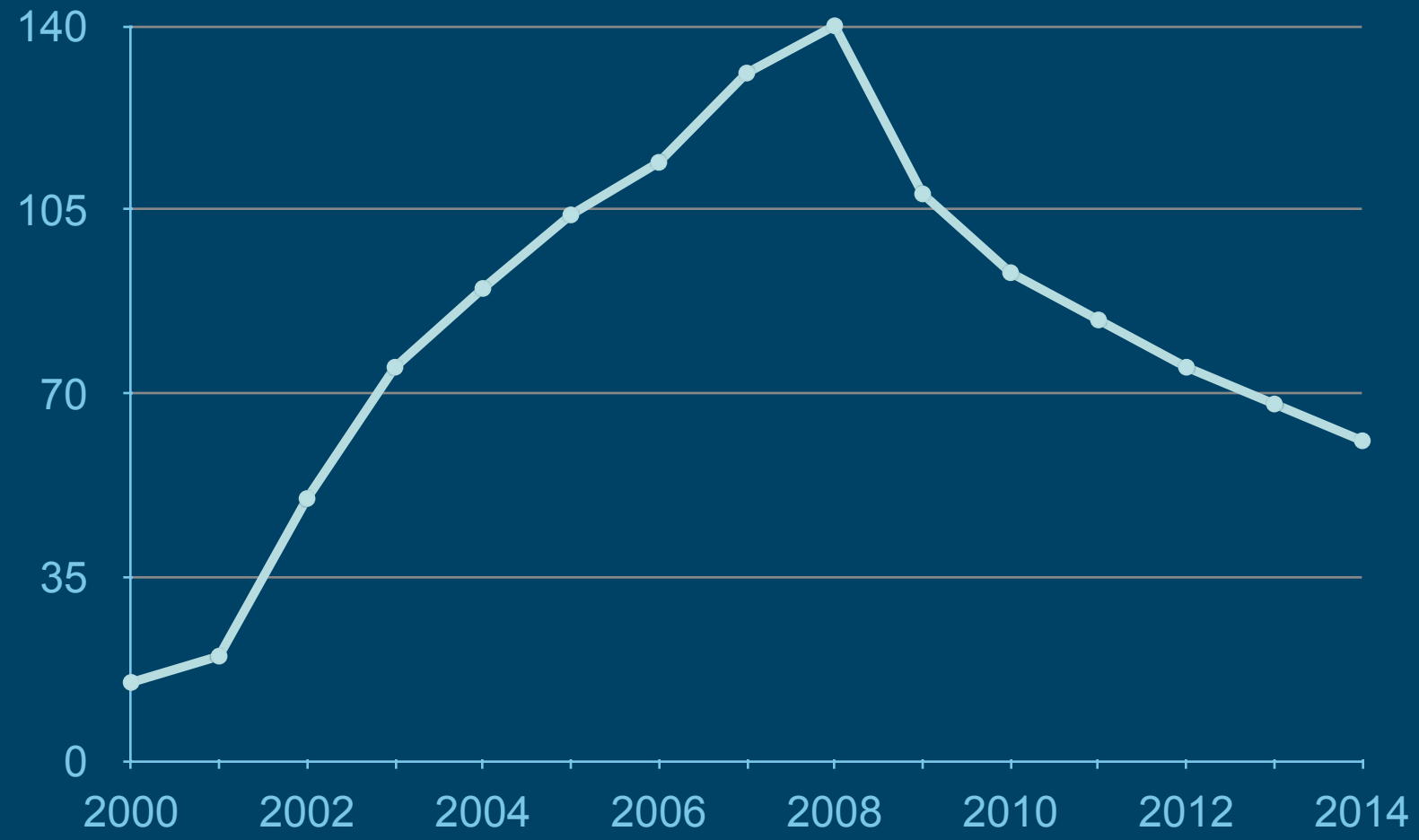
Analisi ed Estrazione di Informazioni da un Account Email a beneficio delle Imprese

Matteo Renzi

Relatore: Prof.ssa Sonia Bergamaschi
Correlatore: Jim Spohrer

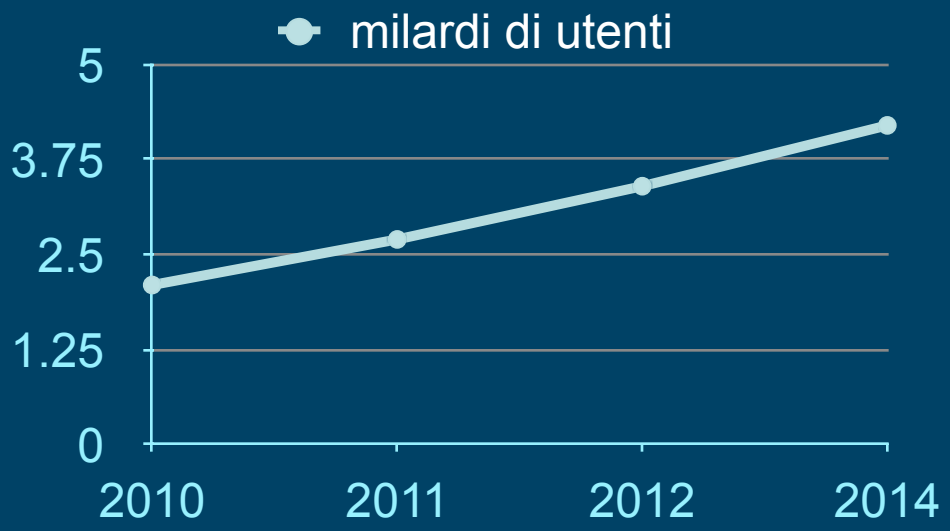
Problema

numero medio di email scambiate per giorno nel settore Business in calo



Motivazioni

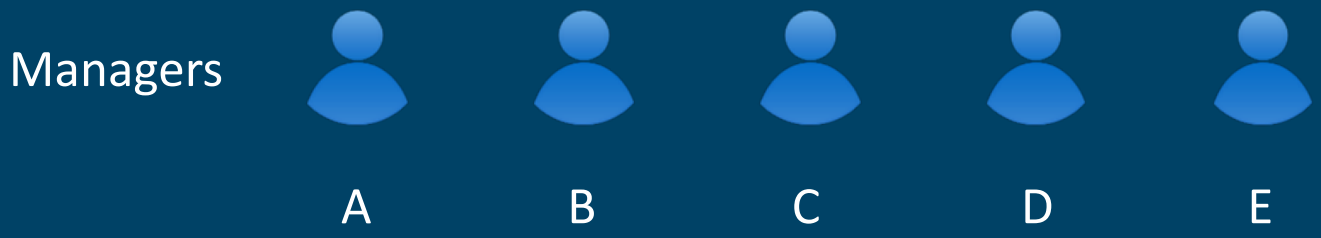
- Crescita sistemi di **Social Networking** per comunicazione sia tra i consumatori che gli utenti business



- **Confusione**:

  ci si *"perde"* nelle proprie email

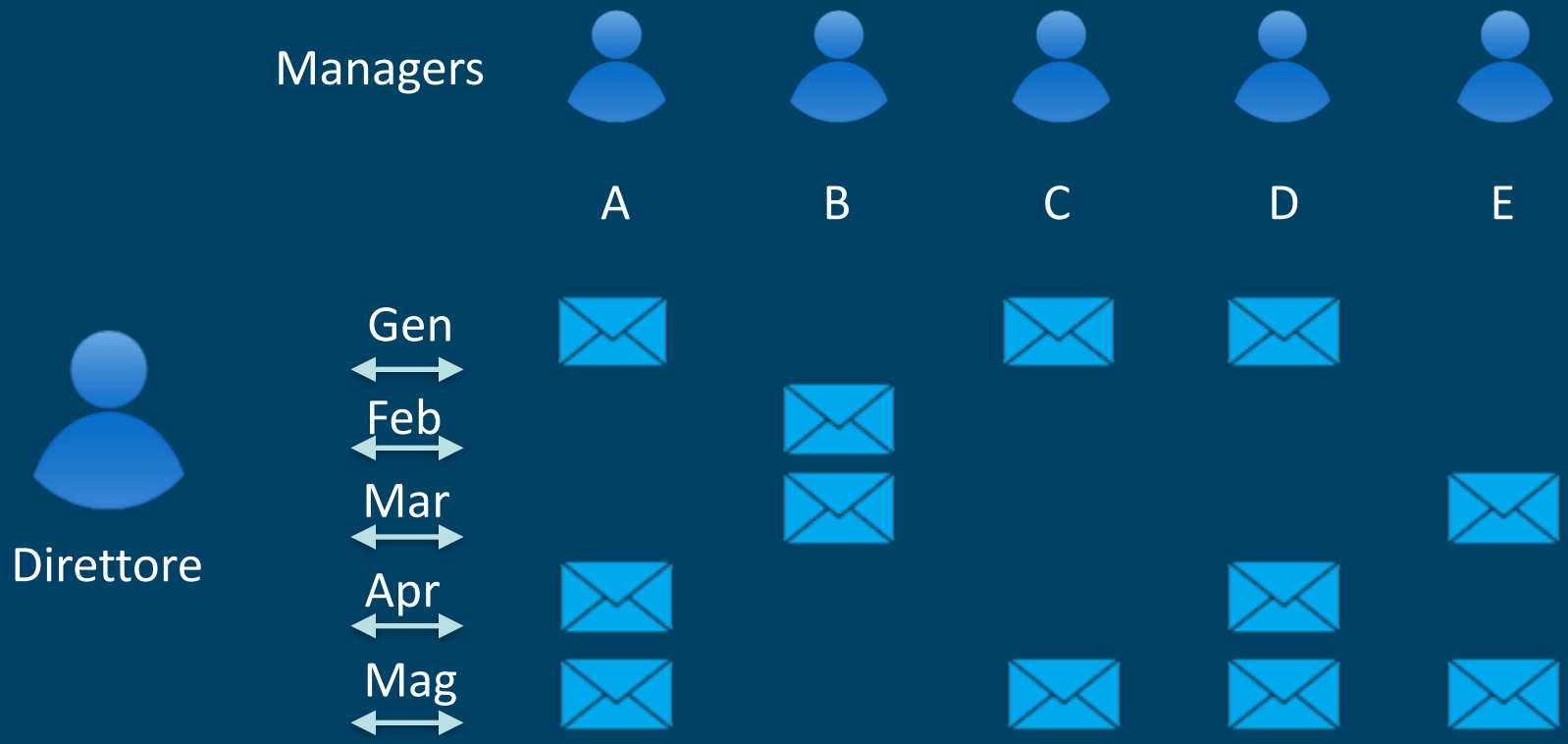
Esempio Pratico



Direttore

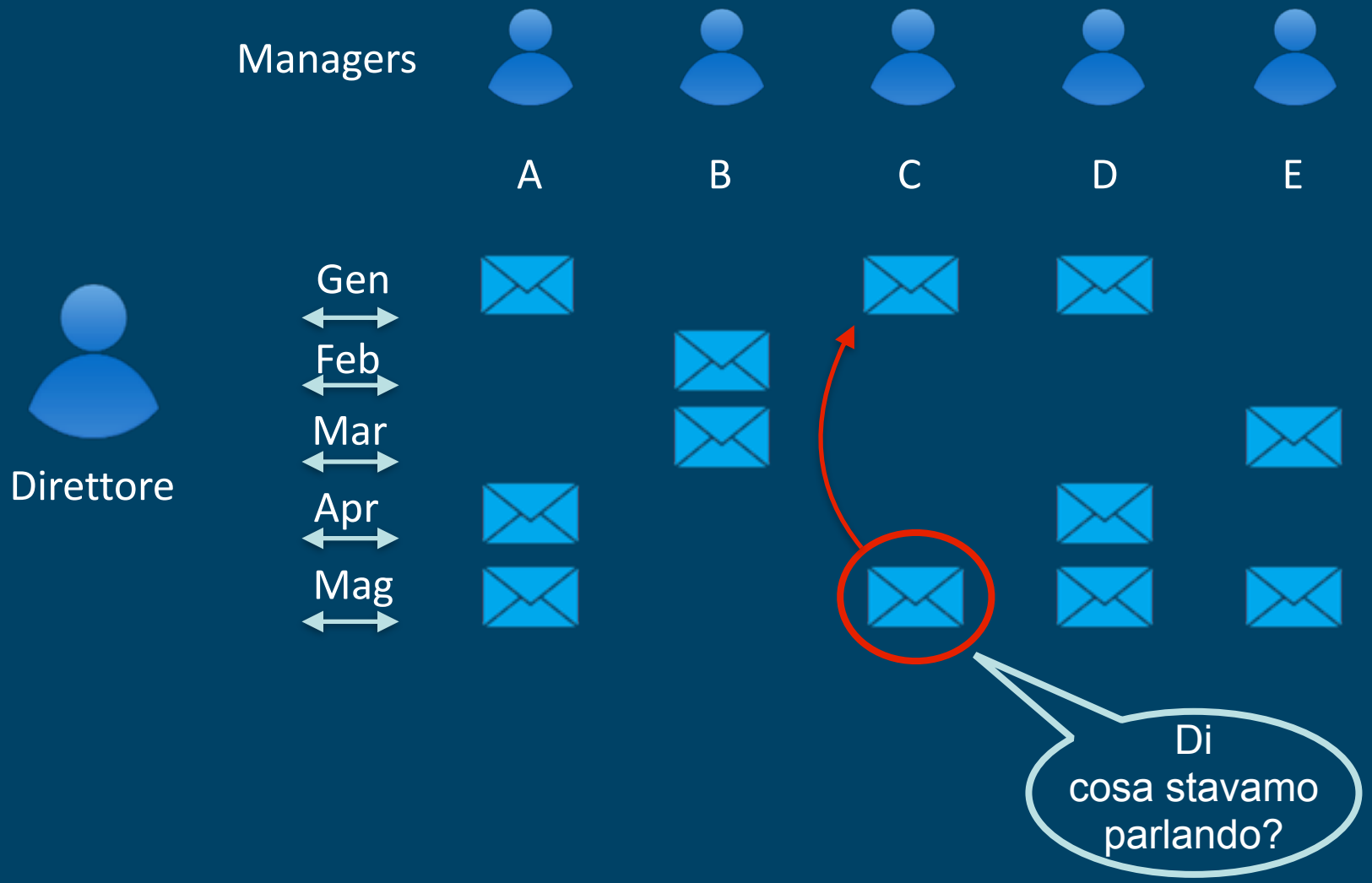
Moltiplichiamo questo per almeno 30 contatti in parallelo…

Esempio Pratico



Moltiplichiamo questo per almeno 30 contatti in parallelo…

Esempio Pratico



Moltiplichiamo questo per almeno 30 contatti in parallelo…

Soluzione

MyCogs

- destinatari —> Top Manager di IBM
- sviluppato in collaborazione con IBM **Watson Group** presso il centro di ricerca IBM Almaden, San Jose (CA)

Cosa mi può aiutare durante i meeting?

Soluzione

MyCogs

- destinatari —> Top Manager di IBM
- sviluppato in collaborazione con IBM **Watson Group** presso il centro di ricerca IBM Almaden, San Jose (CA)

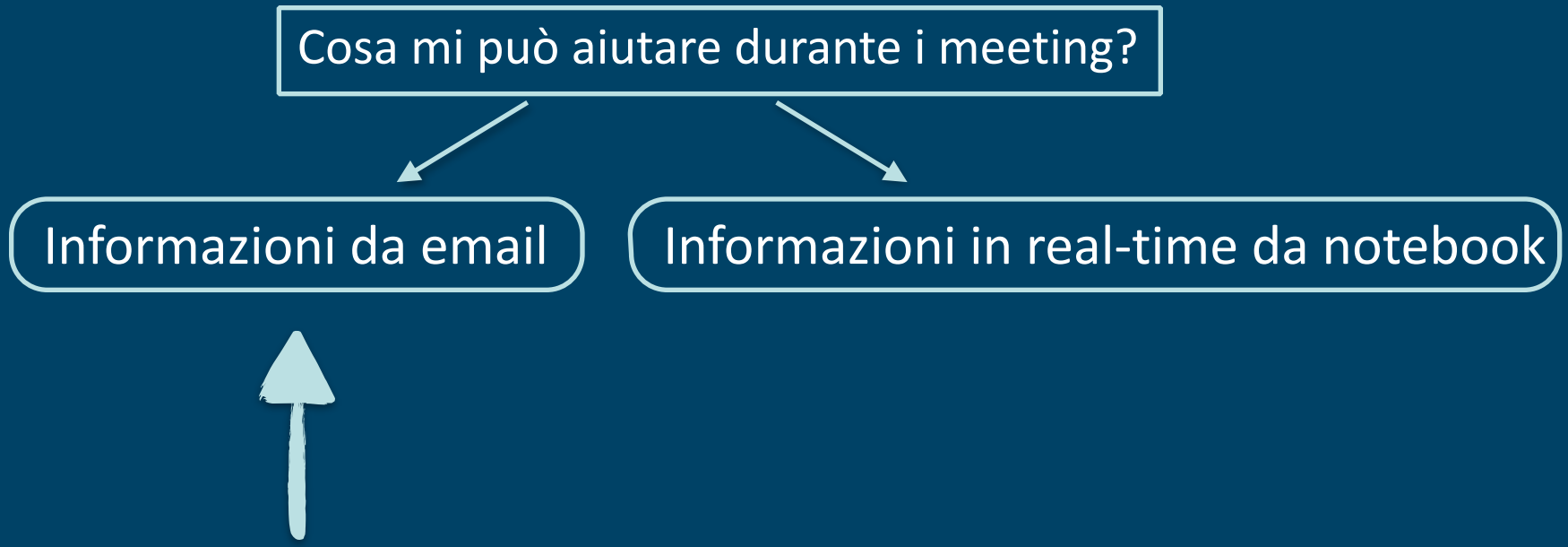
Cosa mi può aiutare durante i meeting?

Informazioni da email

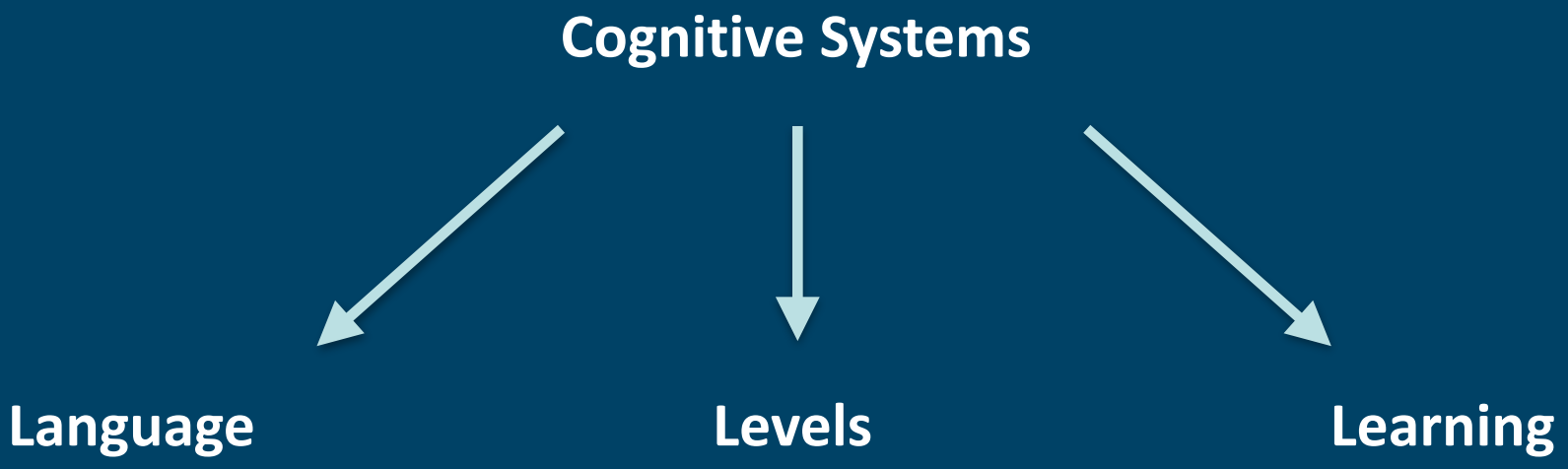
Soluzione

MyCogs

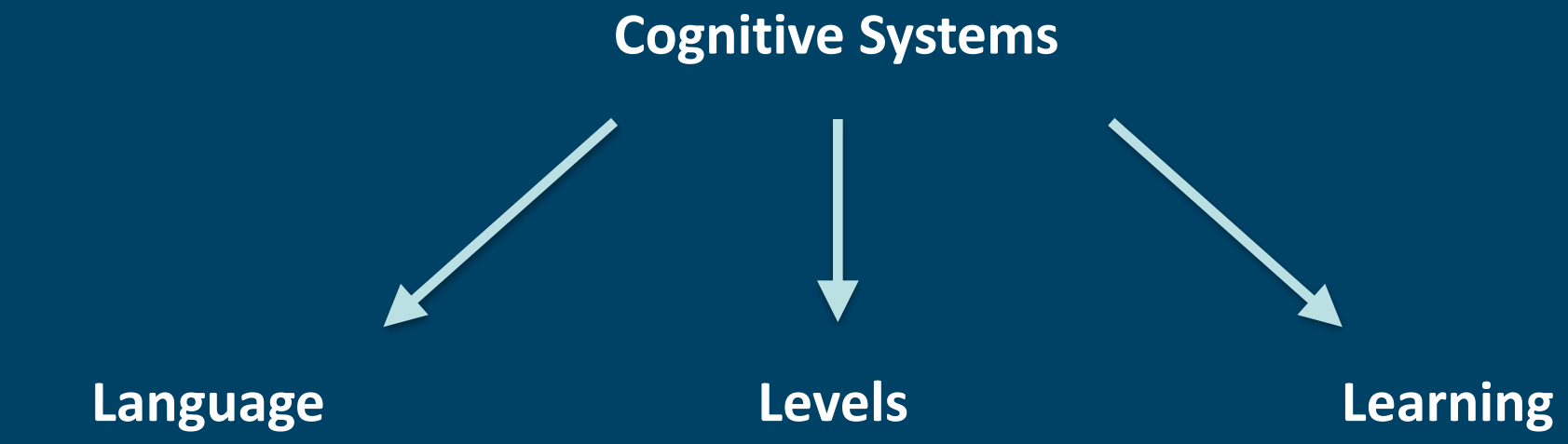
- destinatari —> Top Manager di IBM
- sviluppato in collaborazione con IBM **Watson Group** presso il centro di ricerca IBM Almaden, San Jose (CA)



Background

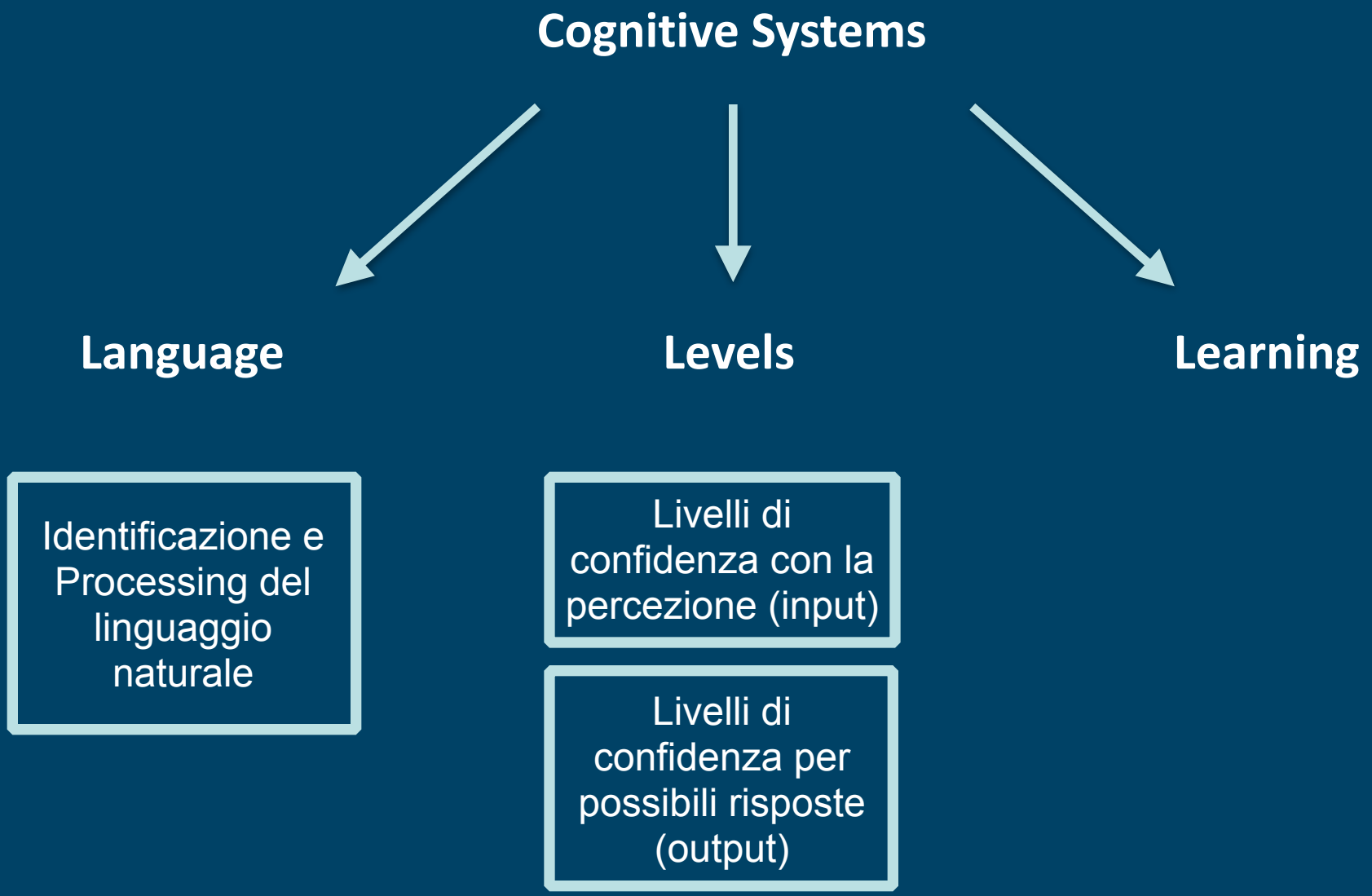


Background

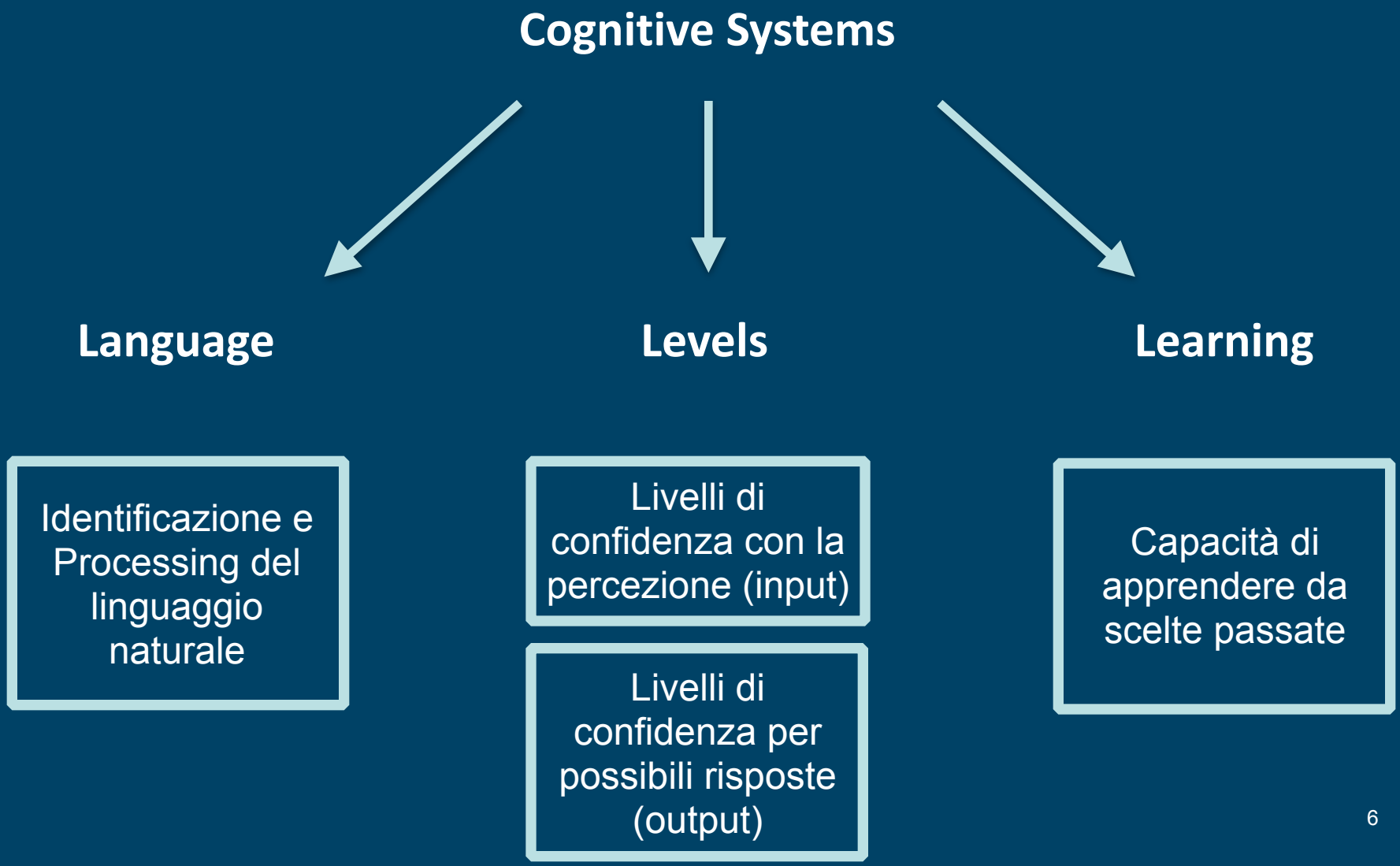


Identificazione e Processing del linguaggio naturale

Background


Cognitive Systems
Language
Levels
Learning
Identificazione e Processing del linguaggio naturale
Livelli di confidenza con la percezione (input)
Livelli di confidenza per possibili risposte (output)

Background


Cognitive Systems
Language
Levels
Learning
Identificazione e Processing del linguaggio naturale
Livelli di confidenza con la percezione (input)
Livelli di confidenza per possibili risposte (output)
Capacità di apprendere da scelte passate

Funzionamento Generale

Funzionamento Generale

ACCOUNT

From: Antonio <antonio@gmail.com>
o
Fwd: Mauro <mauro@gmail.com>

CONTACT LIST

Funzionamento Generale

ACCOUNT

From: Antonio <antonio@gmail.com>
o
Fwd: Mauro <mauro@gmail.com>

CONTACT LIST

OTTIENI TESTO

APRI EMAIL

Funzionamento Generale

ACCOUNT

From: Antonio <antonio@gmail.com>
o
Fwd: Mauro <mauro@gmail.com>

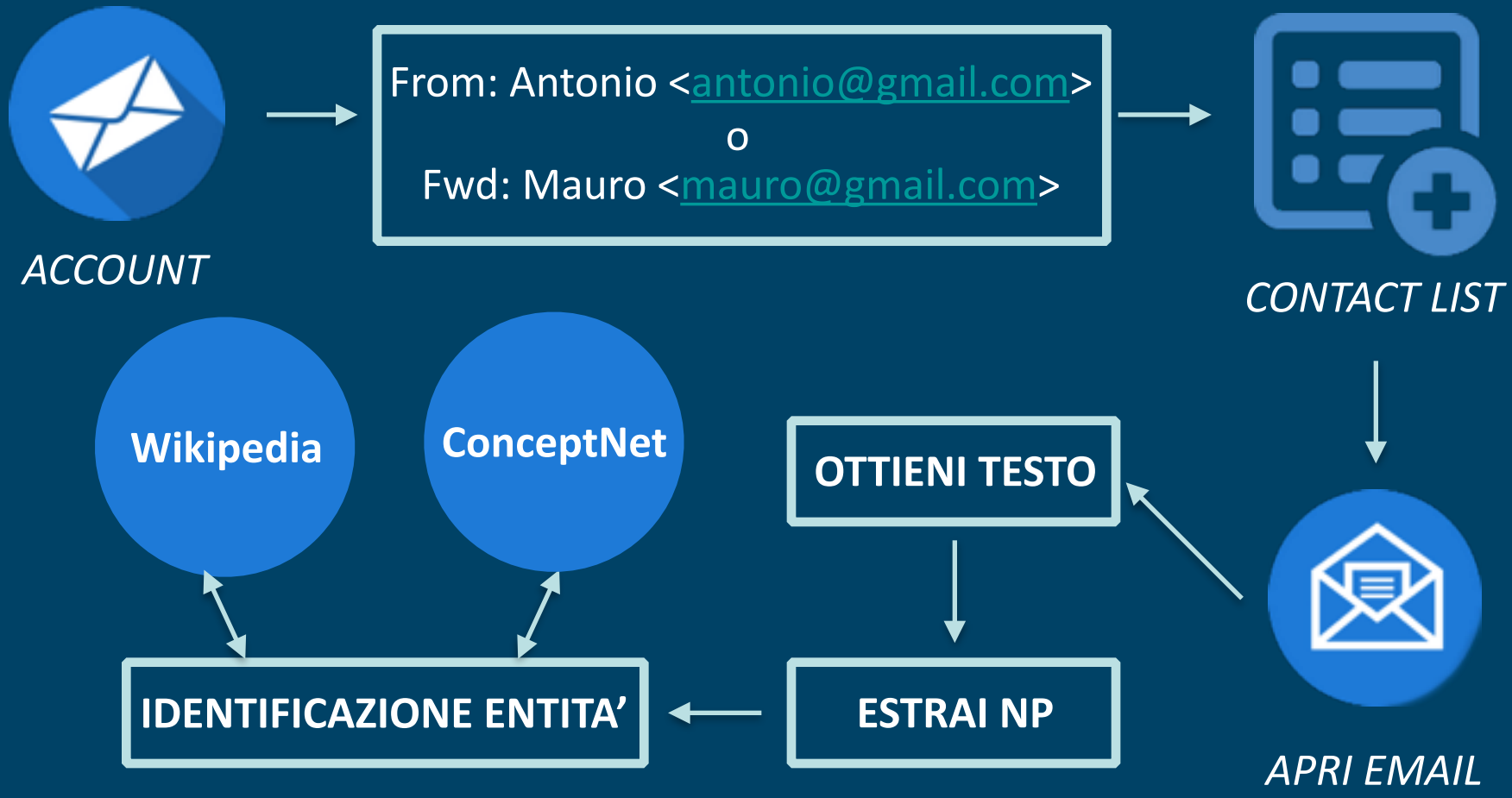
CONTACT LIST

OTTIENI TESTO

ESTRAI NP

APRI EMAIL

Funzionamento Generale



Funzionamento Generale

ACCOUNT

From: Antonio <antonio@gmail.com>
o
Fwd: Mauro <mauro@gmail.com>

CONTACT LIST

Wikipedia

ConceptNet

OTTIENI TESTO

APRI EMAIL

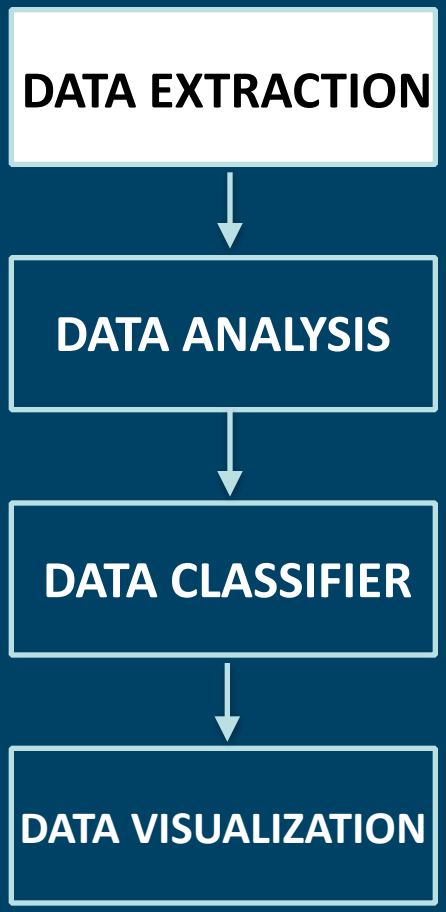
IDENTIFICAZIONE ENTITA'

ESTRAI NP

GENERAZIONE CONOSCENZA

VISUALIZZAZIONE
MODIFICA

Sviluppo - 4 moduli distinti



Modulo di Estrazione delle informazioni (1/2)

- basato su **protocollo IMAP**
  - interrogazioni HEADER direttamente sul server
  - mailbox directory
  - ricerca senza download
  - download di singola sezione MIME

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary=frontier
This is a message with multiple parts in MIME format.
--frontier
Content-Type: text/plain
This is the body of the message.
--frontier
Content-Type: application/octet-stream
Content-Transfer-Encoding: base64
PGh0bWw+CiAgPGhlYWQ+CiAgPC9oZWFkPgogIDxib2R5PgogICAgPHA+
   VGhpcyBpcyB0aGUg
Ym9keSBvZiB0aGUgbWVzc2FnZS48L3A+CiAgPC9ib2R5Pgo8L2h0bWw+Cg==
--frontier--
```

Modulo di Estrazione delle informazioni (1/2)

- basato su **protocollo IMAP**
  - interrogazioni HEADER direttamente sul server
  - mailbox directory
  - ricerca senza download
  - download di singola sezione MIME

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary=frontier
This is a message with multiple parts in MIME format.
--frontier
Content-Type: text/plain
This is the body of the message.
--frontier
Content-Type: application/octet-stream
Content-Transfer-Encoding: base64
PGh0bWw+CiAgPGhlYWQ+CiAgPC9oZWFkPgogIDxib2R5PgogICAgPHA+
   VGhpcyBpcyB0aGUg
Ym9keSBvZiB0aGUgbWVzc2FnZS48L3A+CiAgPC9ib2R5Pgo8L2h0bWw+Cg==
--frontier--
```

Modulo di Estrazione delle informazioni (1/2)

- basato su **protocollo IMAP**
  - interrogazioni HEADER direttamente sul server
  - mailbox directory
  - ricerca senza download
  - download di singola sezione MIME

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary=frontier
This is a message with multiple parts in MIME format.
--frontier
Content-Type: text/plain
This is the body of the message.
--frontier
Content-Type: application/octet-stream
Content-Transfer-Encoding: base64
PGh0bWw+CiAgPGhlYWQ+CiAgPC9oZWFkPgogIDxib2R5PgogICAgPHA+
   VGhpcyBpcyB0aGUg
Ym9keSBvZiB0aGUgbWVzc2FnZS48L3A+CiAgPC9ib2R5Pgo8L2h0bWw+Cg==
--frontier--
```

Modulo di Estrazione delle informazioni (1/2)

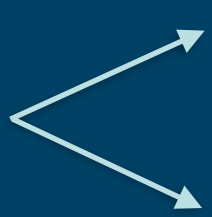
- basato su **protocollo IMAP**
  - interrogazioni HEADER direttamente sul server
  - mailbox directory
  - ricerca senza download
  - download di singola sezione MIME

This is the body of the message.

Modulo di Estrazione delle Informazioni (2/2)

2 OBBIETTIVI

1) Generazione della lista dei mittenti (contact list)



MITTENTI DIRETTI

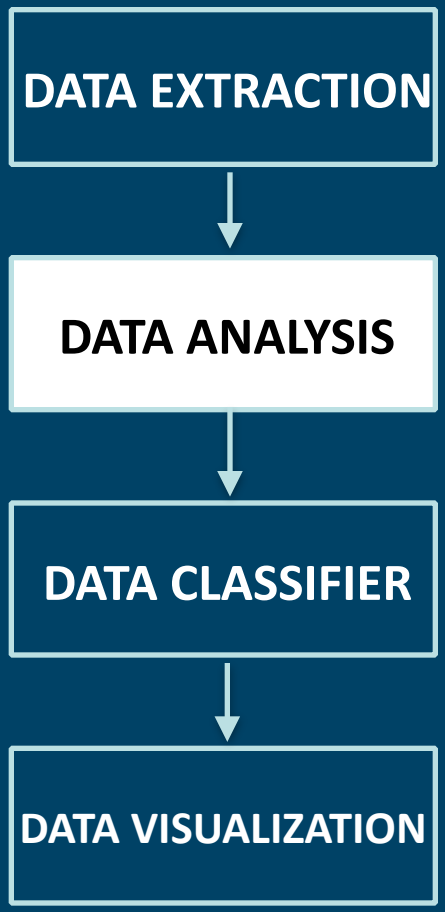
MITTENTI REALI *identificazione di email inoltrate*

→ *identificazione dell'identità della persona*

"Mario Rossi" per mario.rossi@gmail.com

2) Estrazione del testo da una singola email

identifico sezione "text/plain" → estraggo e pulisco → codifico in UTF-8


DATA EXTRACTION
DATA ANALYSIS
DATA CLASSIFIER
DATA VISUALIZATION

Modulo di Analisi delle informazioni (1/2)

Estrazione delle **frasi nominali (NP)**

Punti chiave:

- gestione di **“speciali” parole** frequenti in una mail: url, nomi di persona e luoghi, date, indirizzi e numeri di telefono
- utilizzo dell’algoritmo **TnT tagger** per l’analisi logica della frase
- attraversamento albero in modalità TOP-DOWN per recupero delle NPs

Modulo di Analisi delle informazioni (1/2)

Estrazione delle **frasi nominali (NP)**

We	saw	the	yellow	dog	.	He	barked	at	the	cat

Punti chiave:

- gestione di **“speciali” parole** frequenti in una mail: url, nomi di persona e luoghi, date, indirizzi e numeri di telefono
- utilizzo dell’algoritmo **TnT tagger** per l’analisi logica della frase
- attraversamento albero in modalità TOP-DOWN per recupero delle NPs

Modulo di Analisi delle informazioni (1/2)

Estrazione delle **frasi nominali (NP)**

We	saw	the	yellow	dog		.	He	barked	at	the	cat

S S

Punti chiave:

- gestione di **“speciali” parole** frequenti in una mail: url, nomi di persona e luoghi, date, indirizzi e numeri di telefono
- utilizzo dell’algoritmo **TnT tagger** per l’analisi logica della frase
- attraversamento albero in modalità TOP-DOWN per recupero delle NPs

Modulo di Analisi delle informazioni (1/2)

Estrazione delle **frasi nominali (NP)**

We	saw	the	yellow	dog		.	He	barked	at	the	cat
PRP	VBD	DT	JJ	NN			PRP	VBD	IN	DT	NN

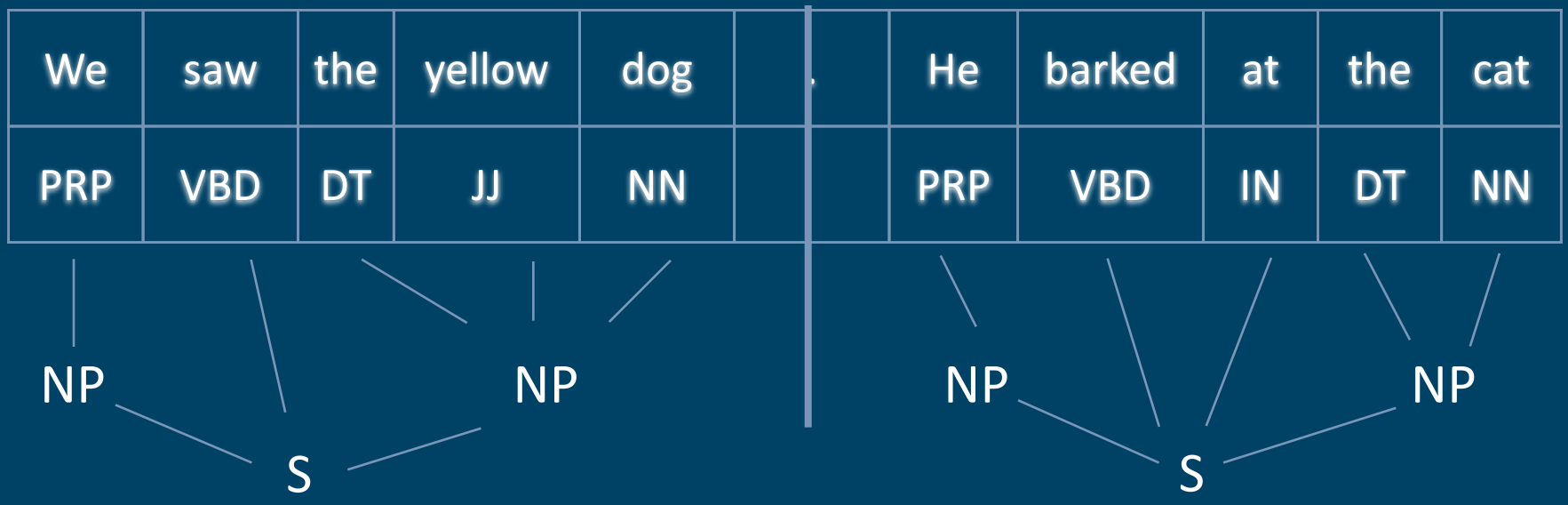
S S

Punti chiave:

- gestione di **“speciali” parole** frequenti in una mail: url, nomi di persona e luoghi, date, indirizzi e numeri di telefono
- utilizzo dell’algoritmo **TnT tagger** per l’analisi logica della frase
- attraversamento albero in modalità TOP-DOWN per recupero delle NPs

Modulo di Analisi delle informazioni (1/2)

Estrazione delle **frasi nominali (NP)**



Punti chiave:

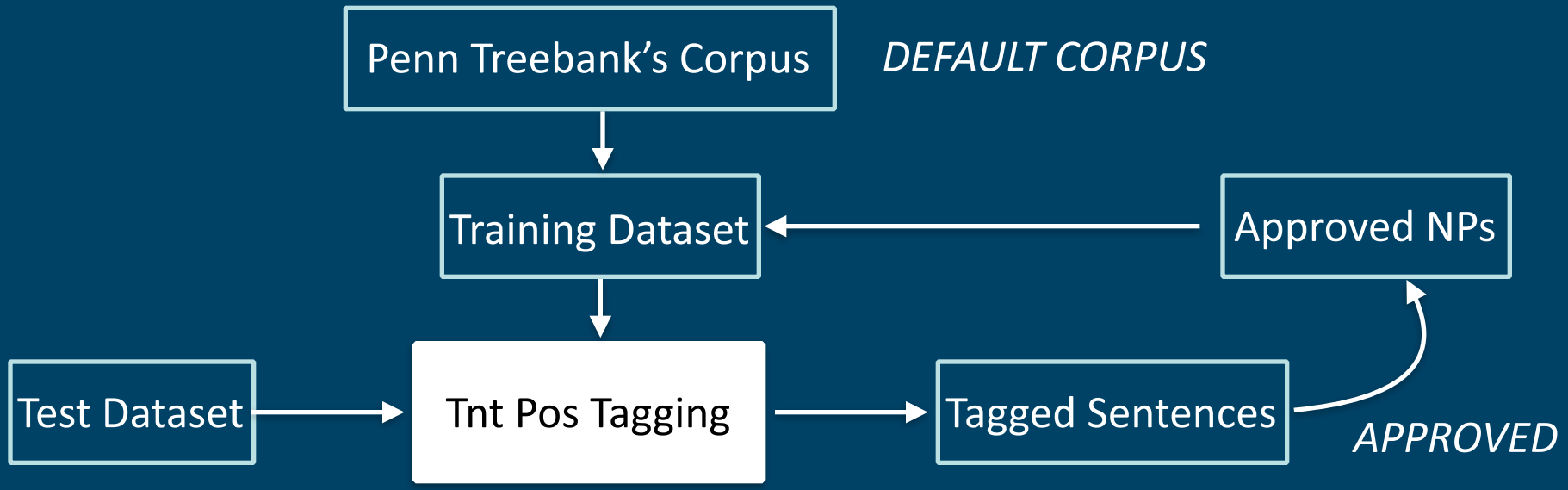
- gestione di **“speciali” parole** frequenti in una mail: url, nomi di persona e luoghi, date, indirizzi e numeri di telefono
- utilizzo dell’algoritmo **TnT tagger** per l’analisi logica della frase
- attraversamento albero in modalità TOP-DOWN per recupero delle NPs

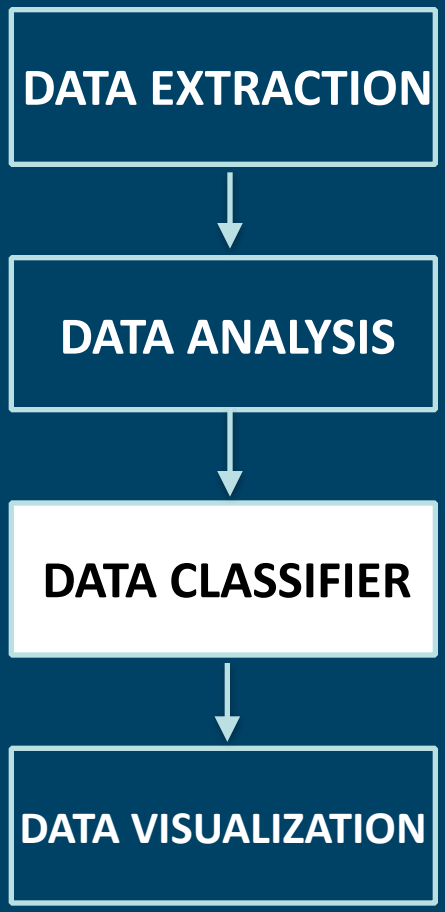
Modulo di Analisi delle Informazioni (2/2)

Perche' Tnt Tagger?

- basato sull'implementazione dell'algoritmo di Viterbi per il modello di Markov del secondo ordine

- ottimizzato per essere **trainable** su una grande varietà di corpus




DATA EXTRACTION
DATA ANALYSIS
DATA CLASSIFIER
DATA VISUALIZATION

Modulo di Classificazione delle informazioni (1/2)

MACHINE READING

corretta identificazione delle entità menzionate nel testo tramite recupero di informazioni da datasets quali Wikipedia e ConceptNet5

"IBM" → Wikipedia → ok

"Pineapple juice" → Wikipedia → non trovo la corrispondente pagina

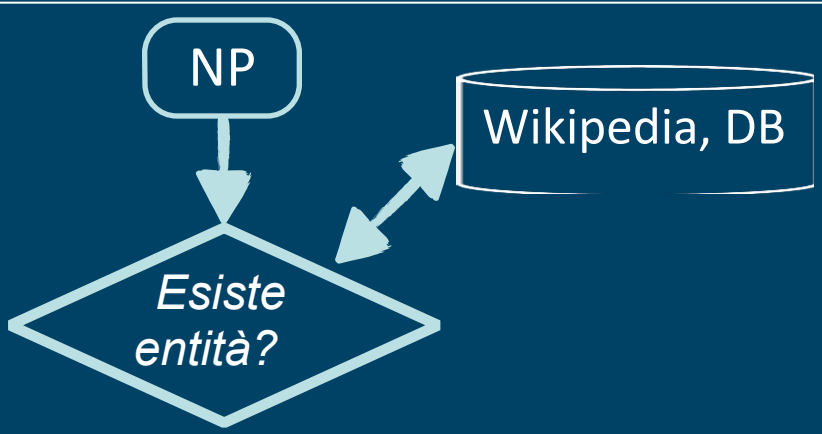
Unlinkable noun phrase problem

"Pineapple juice" → e' un tipo di *"Juice"*

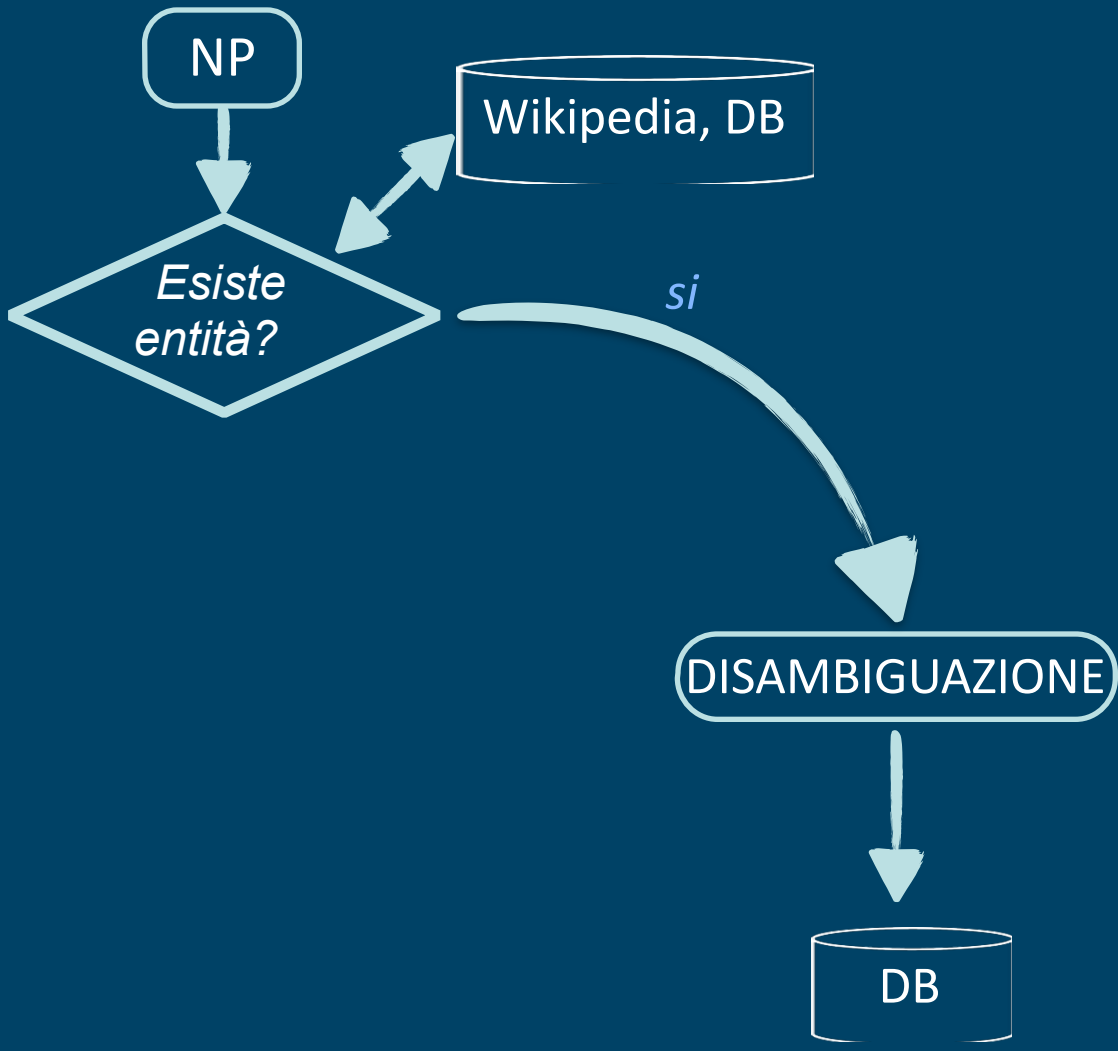
"Facebook image processing" → va suddivisa in *"Facebook"* e *"image processing"*

Modulo di Classificazione delle informazioni (2/2)

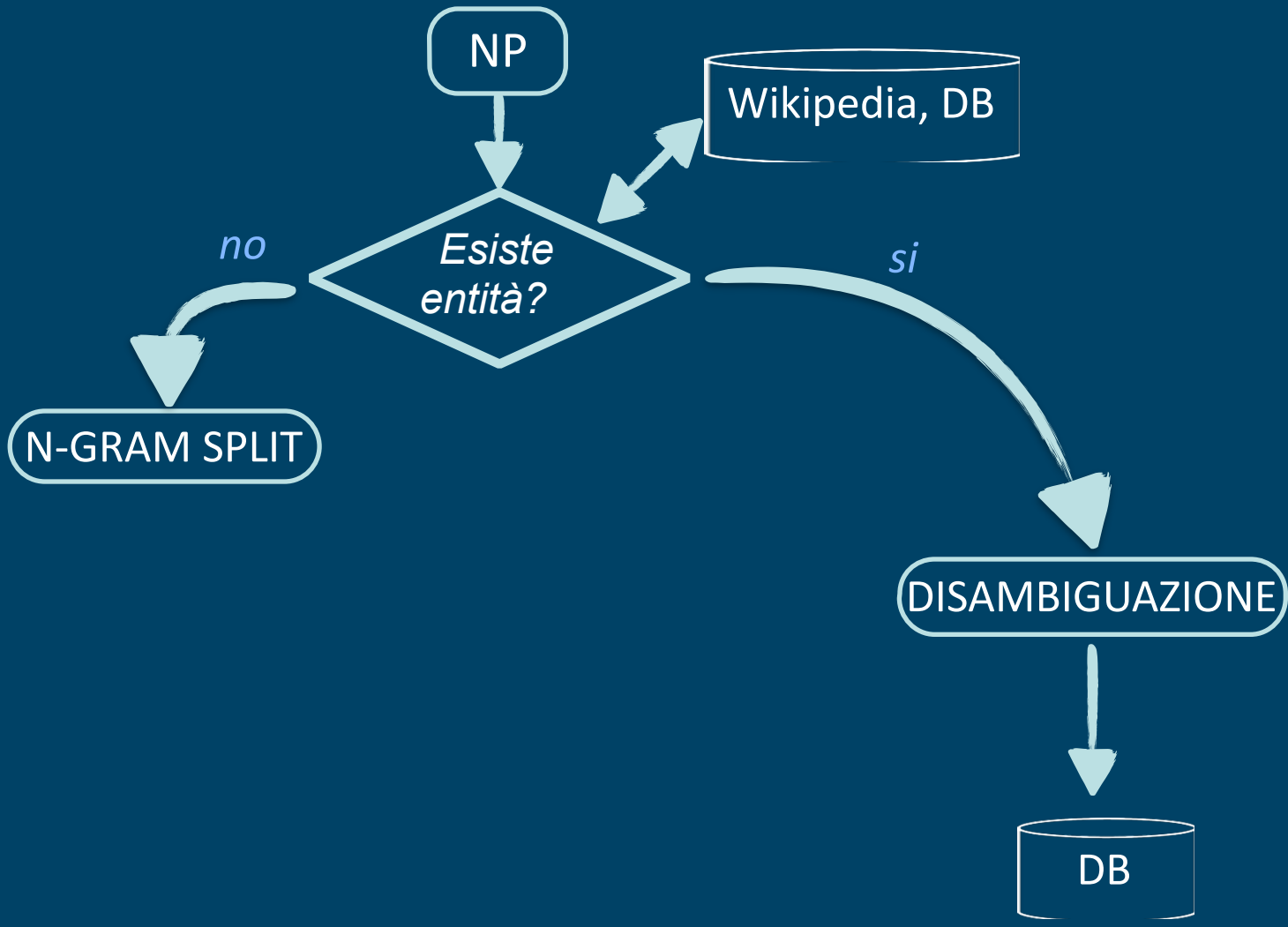
Modulo di Classificazione delle informazioni (2/2)



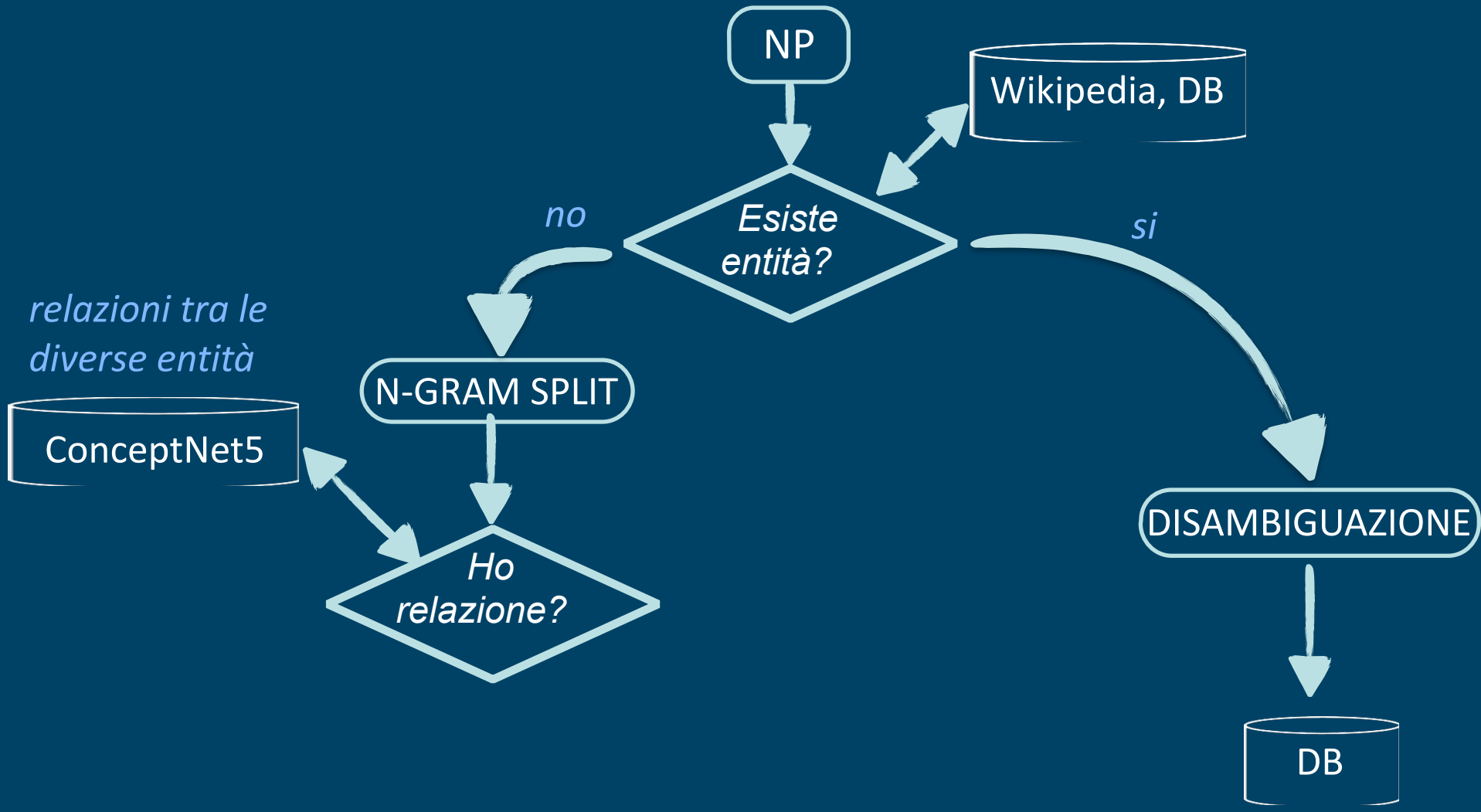
Modulo di Classificazione delle informazioni (2/2)



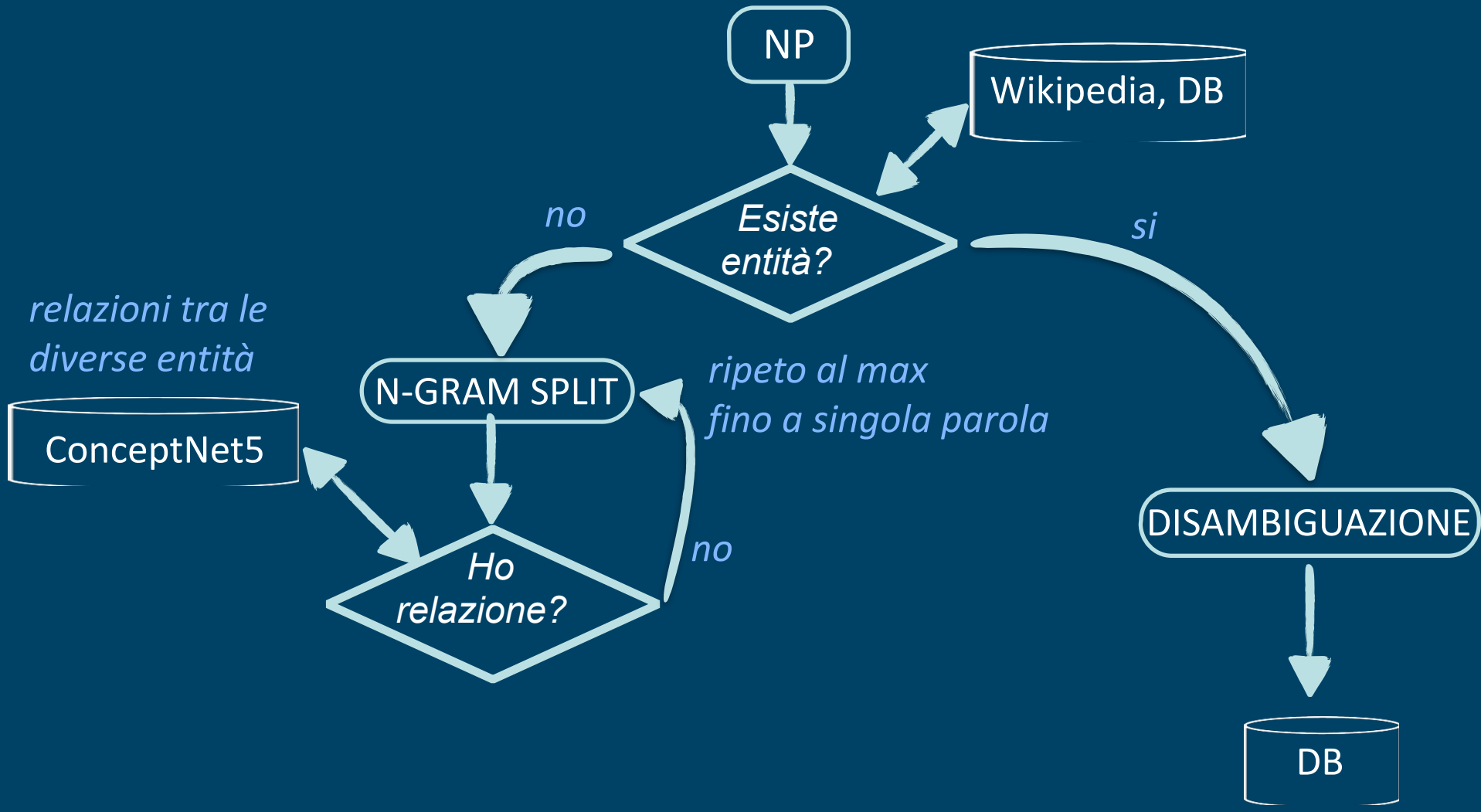
Modulo di Classificazione delle informazioni (2/2)



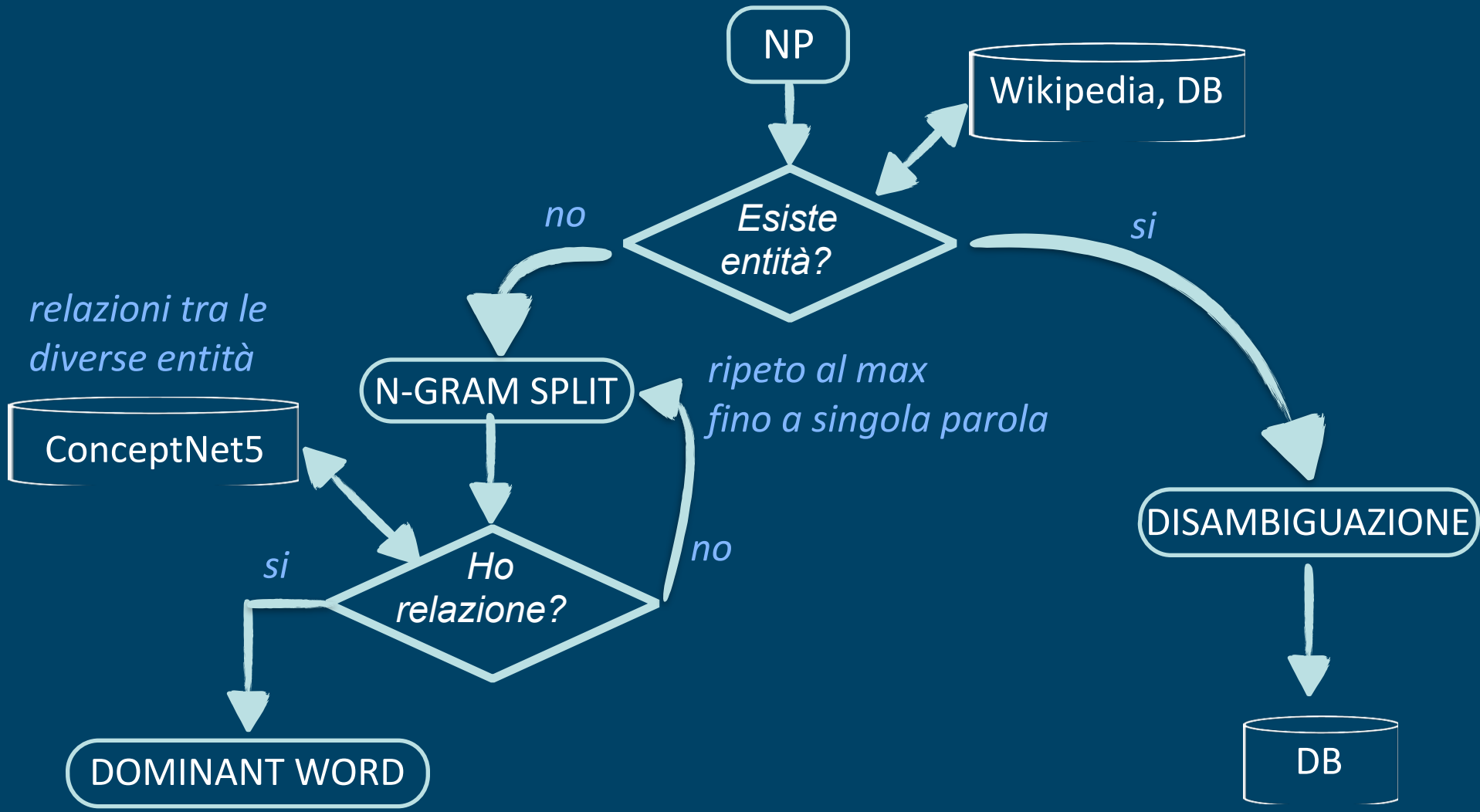
Modulo di Classificazione delle informazioni (2/2)



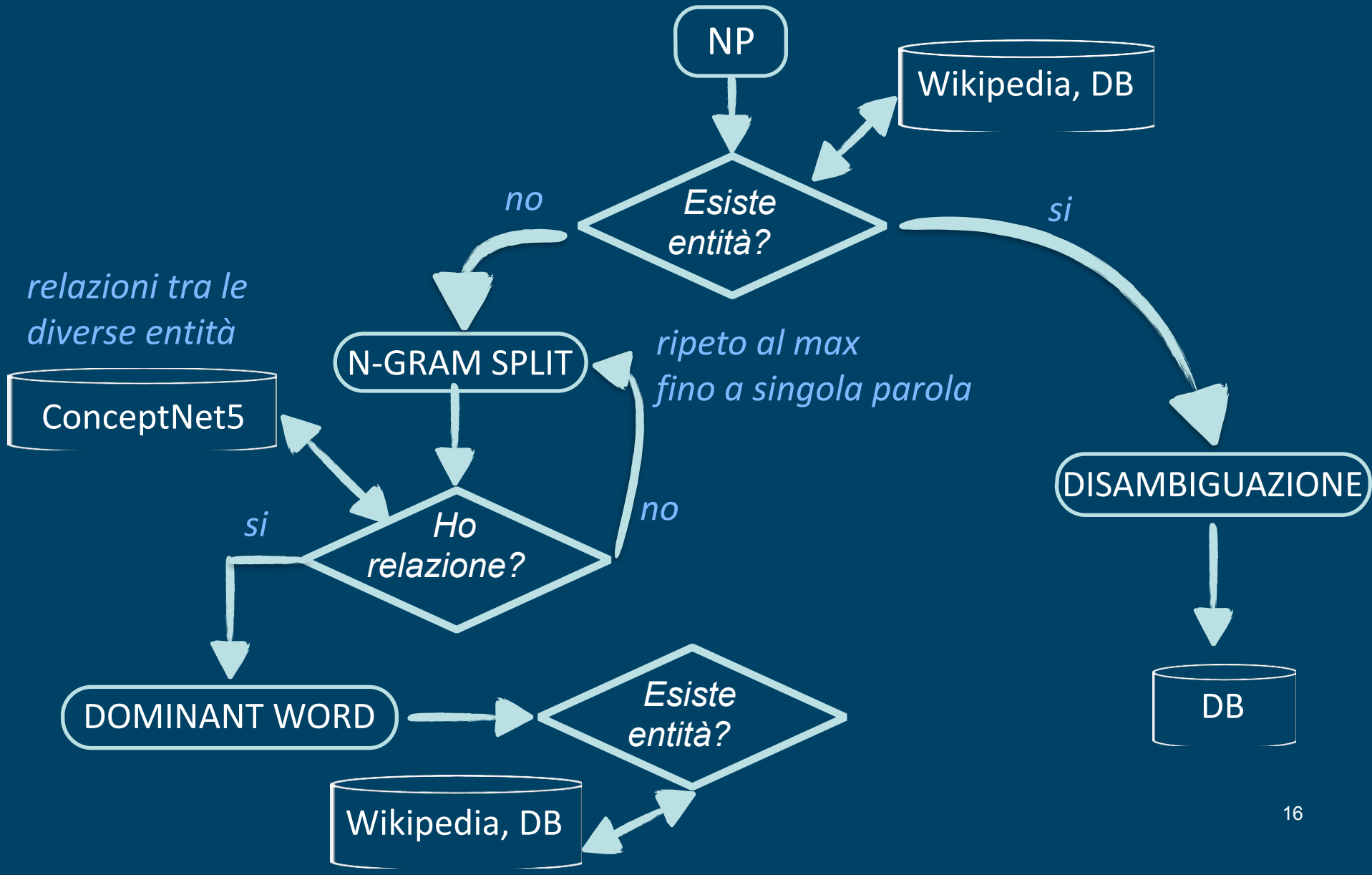
Modulo di Classificazione delle informazioni (2/2)



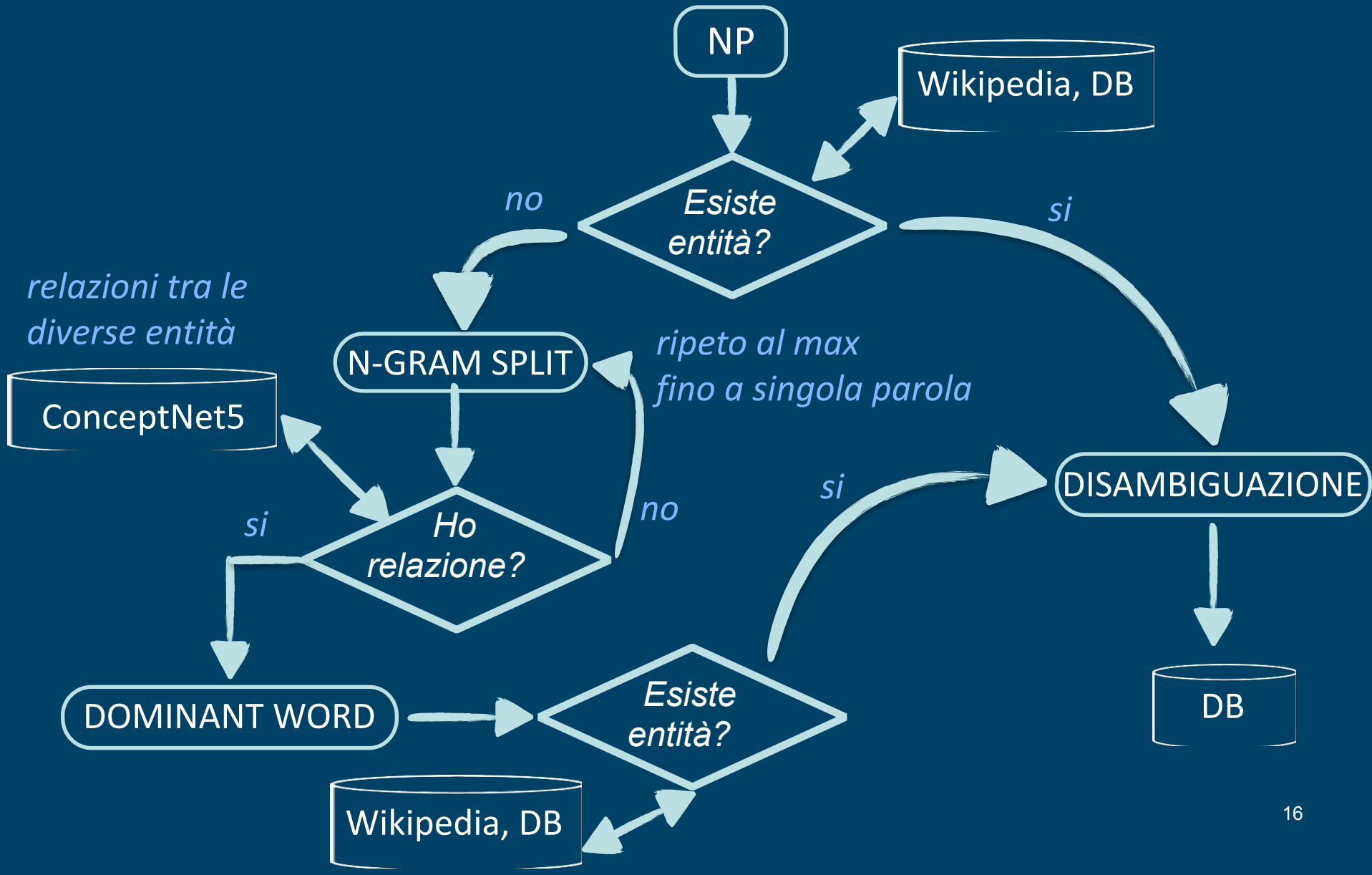
Modulo di Classificazione delle informazioni (2/2)



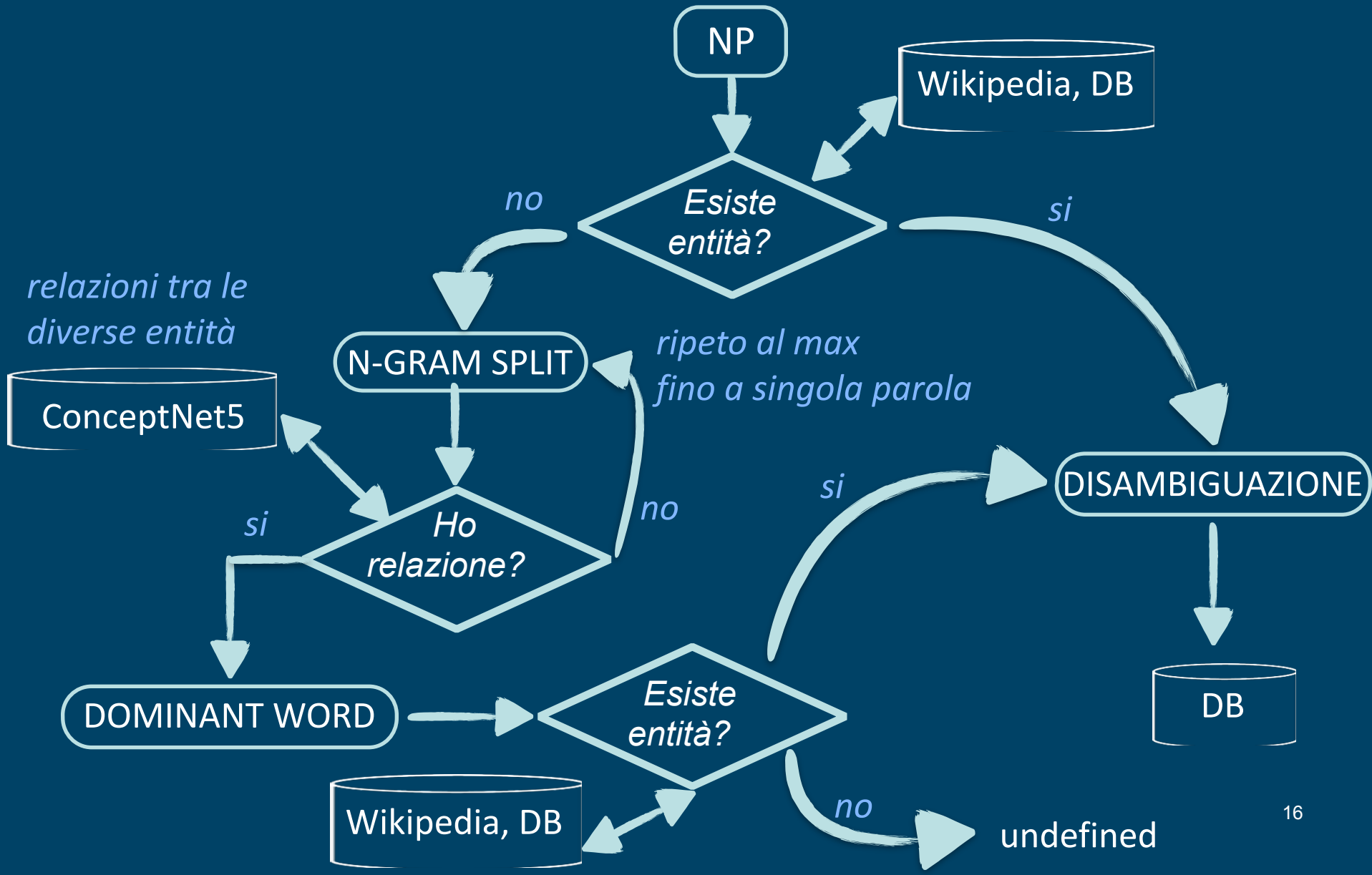
Modulo di Classificazione delle informazioni (2/2)



Modulo di Classificazione delle informazioni (2/2)

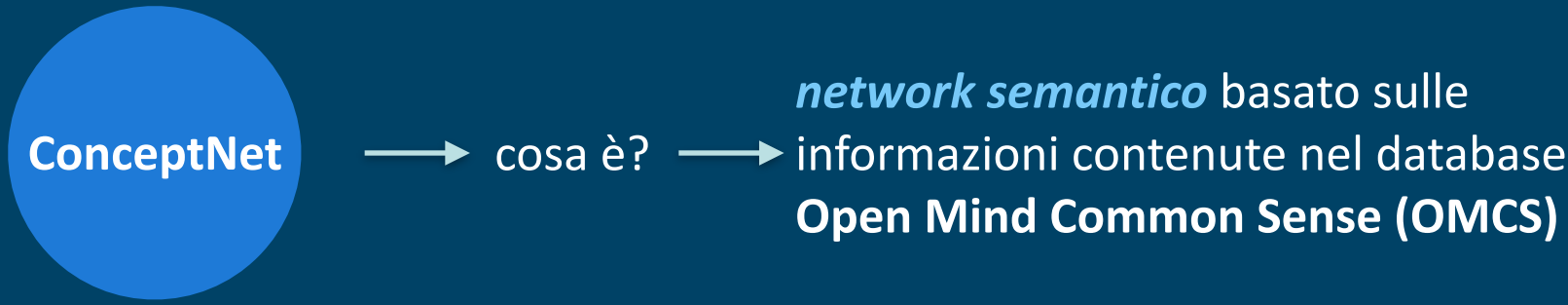


Modulo di Classificazione delle informazioni (2/2)



Ricerca di relazione

Ricerca di relazione



Ricerca di relazione

→ cosa è? → ***network semantico*** basato sulle informazioni contenute nel database **Open Mind Common Sense (OMCS)**

↓

progetto di AI creato nei laboratori MIT che ha come scopo la creazione e utilizzo di una vasta conoscenza “comune” generata dal contributo di migliaia di persone attraverso il Web

Ricerca di relazione

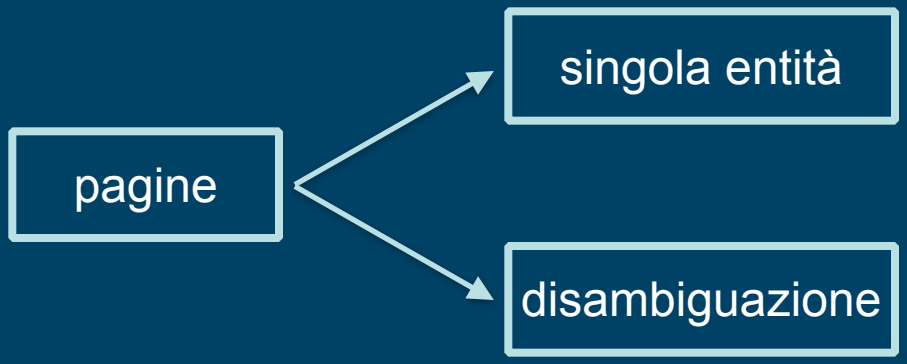
ConceptNet → cosa è? → ***network semantico*** basato sulle informazioni contenute nel database **Open Mind Common Sense (OMCS)**

progetto di AI creato nei laboratori MIT che ha come scopo la creazione e utilizzo di una vasta conoscenza "comune" generata dal contributo di migliaia di persone attraverso il Web

- **nodi** che rappresentano parole o frasi brevi
- 23 differenti tipi di **relazioni** trattate
- informazioni scambiate in formato JSON

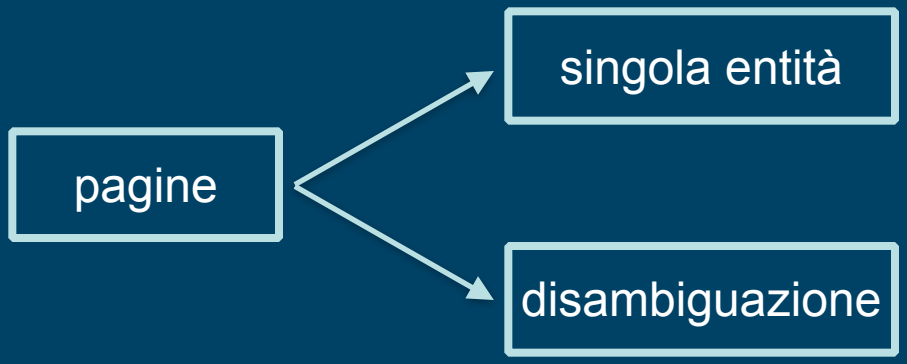
Disambiguazione

Ho trovato l'entità corrispondente in Wikipedia, perché necessario un altro controllo?

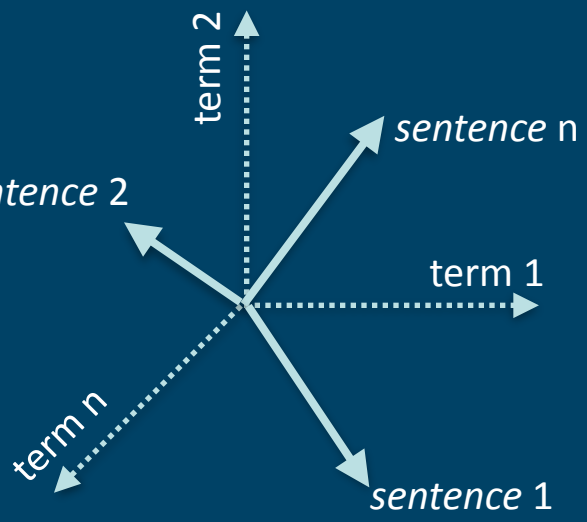


Disambiguazione

Ho trovato l'entità corrispondente in Wikipedia, perché necessario un altro controllo?

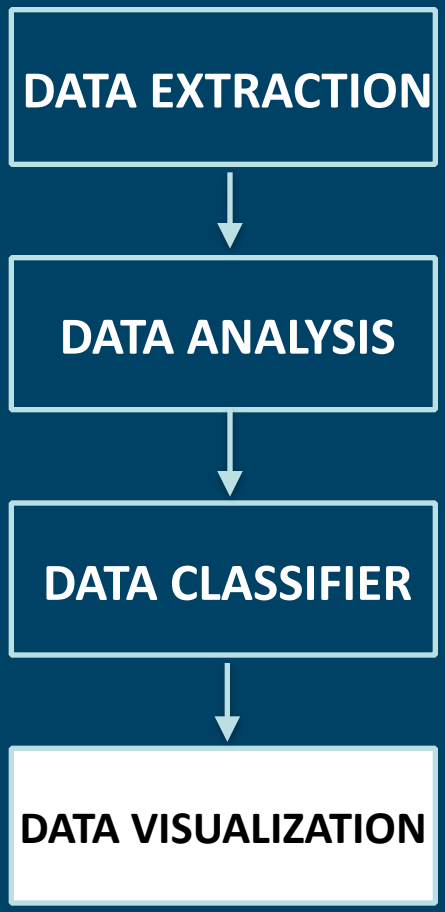


utilizzo della Similarità del Coseno nel Modello di Spazio Vettoriale



elementi di confronto:
- entità grezza (così come estratta)
- tutte le possibili soluzioni proposte da Wikipedia

→ ranking delle possibili soluzioni


DATA EXTRACTION
DATA ANALYSIS
DATA CLASSIFIER
DATA VISUALIZATION

Modulo di Visualizzazione delle informazioni

Costruito sul framework grafico KIVY (MIT):

- deployment multipiattaforma
- open source
- veloce → sviluppo dell'applicazione → *"kv language"* sintassi python
- veloce → esecuzione dell'applicazione → funzioni critiche in C
  - algoritmi intelligenti
  - sfrutta la GPU (se presente)

- supporto interazione touch e particolari eventi multi-touch

Specifiche Tecniche

- Sviluppato in **Python**
- Supporto **multi-piattaforma** (iOS, Android, OSX, Windows, Linux)
- Librerie esterne utilizzate:
  - **Kivy** Framework,
  - **NLTK**,
  - Numpy,
  - BeautifulSoup 4,
  - LXML

Linguaggio	*files*	*vuote*	*commento*	*codice*
Python	*116*	*3572*	*4223*	*13479*
XML	*7*	*5*	*0*	*843*
.make	*1*	*9*	*0*	*14*
TOTALE	*124*	*3586*	*4223*	*14336*

Sviluppi futuri

- strutturare l'applicativo come sistema Enterprise con funzionalità di Web Service

  → porre attenzione alla sensibilità delle informazioni

  → elimino restrizioni dovute all'ambiente mobile

- potenziare la struttura del sistema semantico ottenuto

  → implementare una rete neurale

  → possibilità di interrogare il sistema in linguaggio naturale

IBM Watson API

da una ricerca basata su parole chiave

↓

a una ricerca basata su "conversazione"

GRAZIE PER L’ ATTENZIONE