Maurizio Gallotti

ECM Client Technical Professional

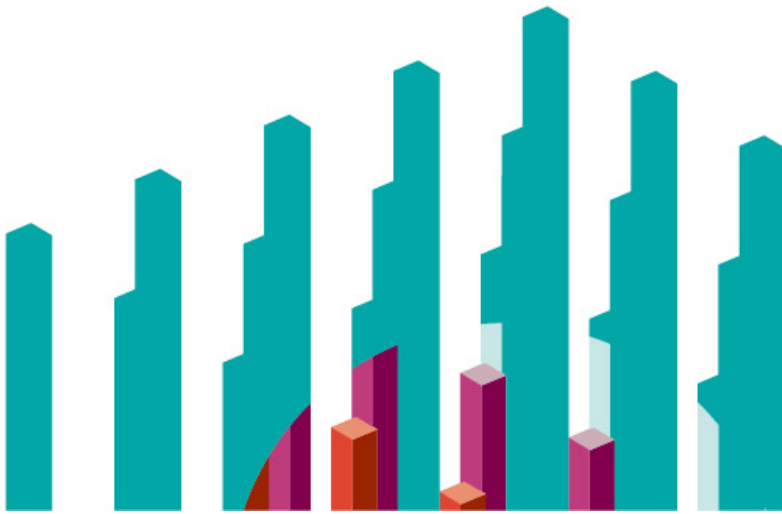
IBM Software | Industry Solutions | Enterprise Content Management

Content Analytics

Acquisire conoscenza dalle informazioni strutturate e non strutturate

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti



IBM Content Analytics
l'esplosione delle informazioni

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

- Nella ricerca full-text i criteri sono espressi mediante termini o espressioni regolari
- La Business Intelligence mostra dati e informazioni ricavate dai metadati disponibili dalle sorgenti dati strutturate
- **Content Analytics** consente di:

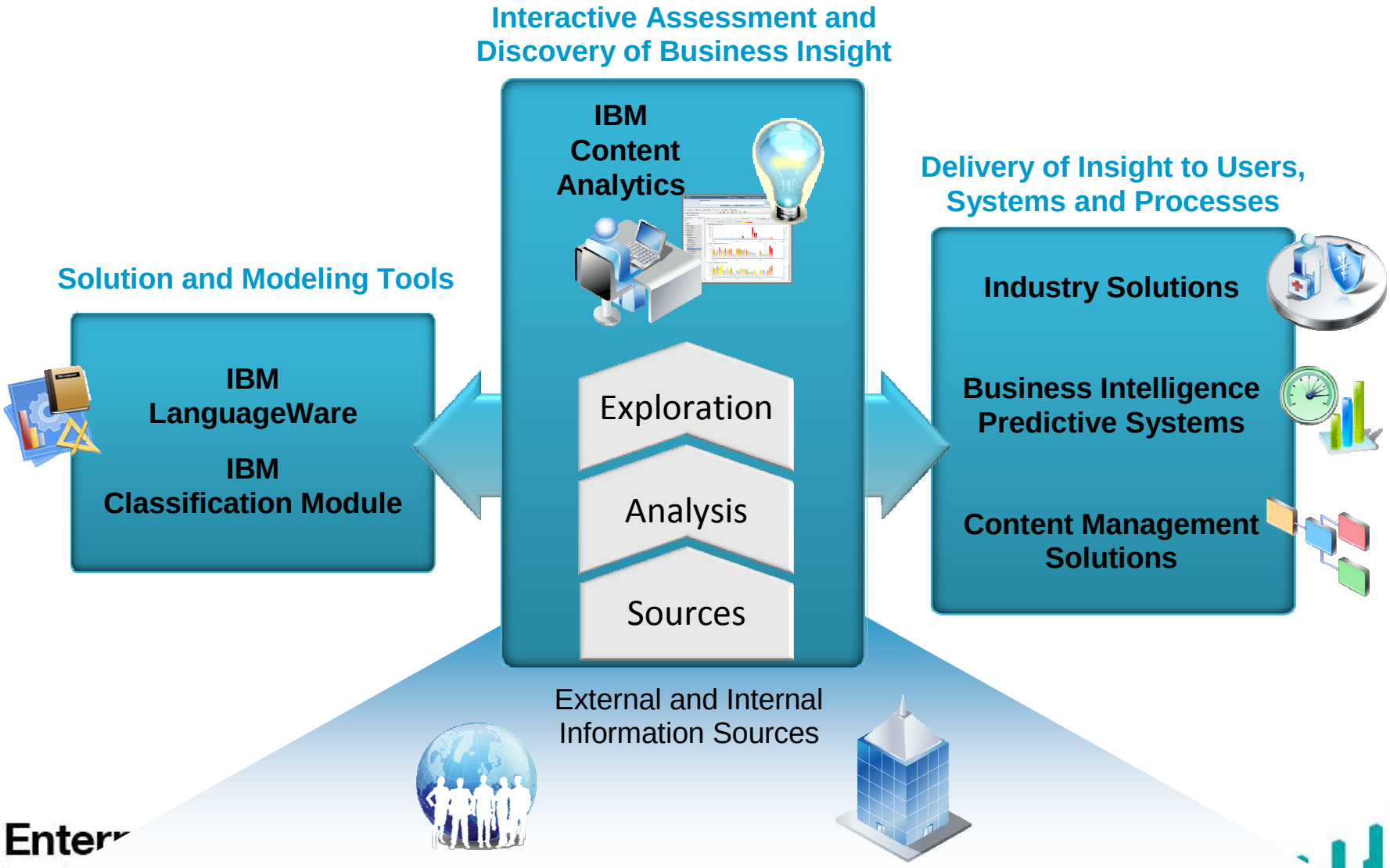
• estrarre, ordinare ed analizzare informazioni di valore contenute nei documenti
• semplificare la condivisione del patrimonio informativo aziendale
• facilitare e rendere efficace l'analisi di grandi quantità di documenti

IBM Content Analytics
caratteristiche

Trova ed estrae informazioni, effettua sui dati non strutturati le stesse analisi che vengono attualmente fatte sui dati strutturati

- Combina dati strutturati e non strutturati rendendo l'analisi agevole e fluida
- Permette di analizzare in modo semplice e veloce i contenuti esplorandoli da diverse prospettive
- Identifica ed evidenzia automaticamente qualsiasi relazione inusuale tra i contenuti
- Si integra facilmente con gli strumenti di business intelligence Cognos
- Accede a tutte le tipologie di documento e a numerosissime fonti documentali
- Offre la possibilità di espandere le capacità di analisi, compresa quella semantica, attraverso tecnologie aperte

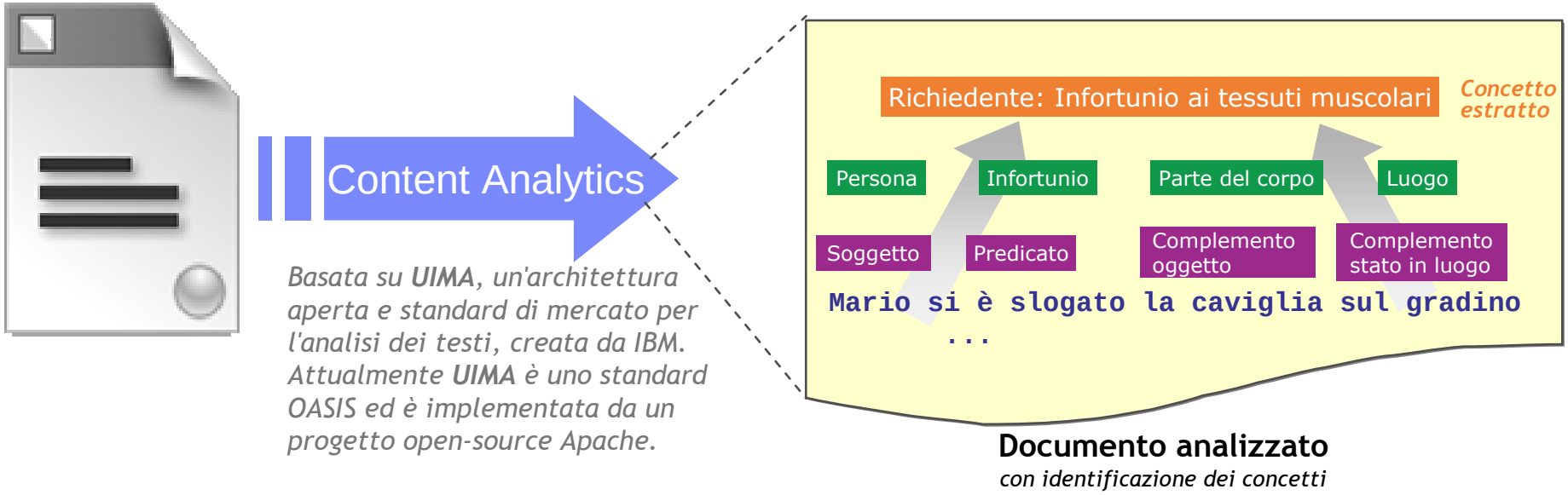
IBM Content Analytics
caratteristiche



IBM Enter
Content Management
Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics

La chiave di volta tecnologica: l'analisi dei contenuti

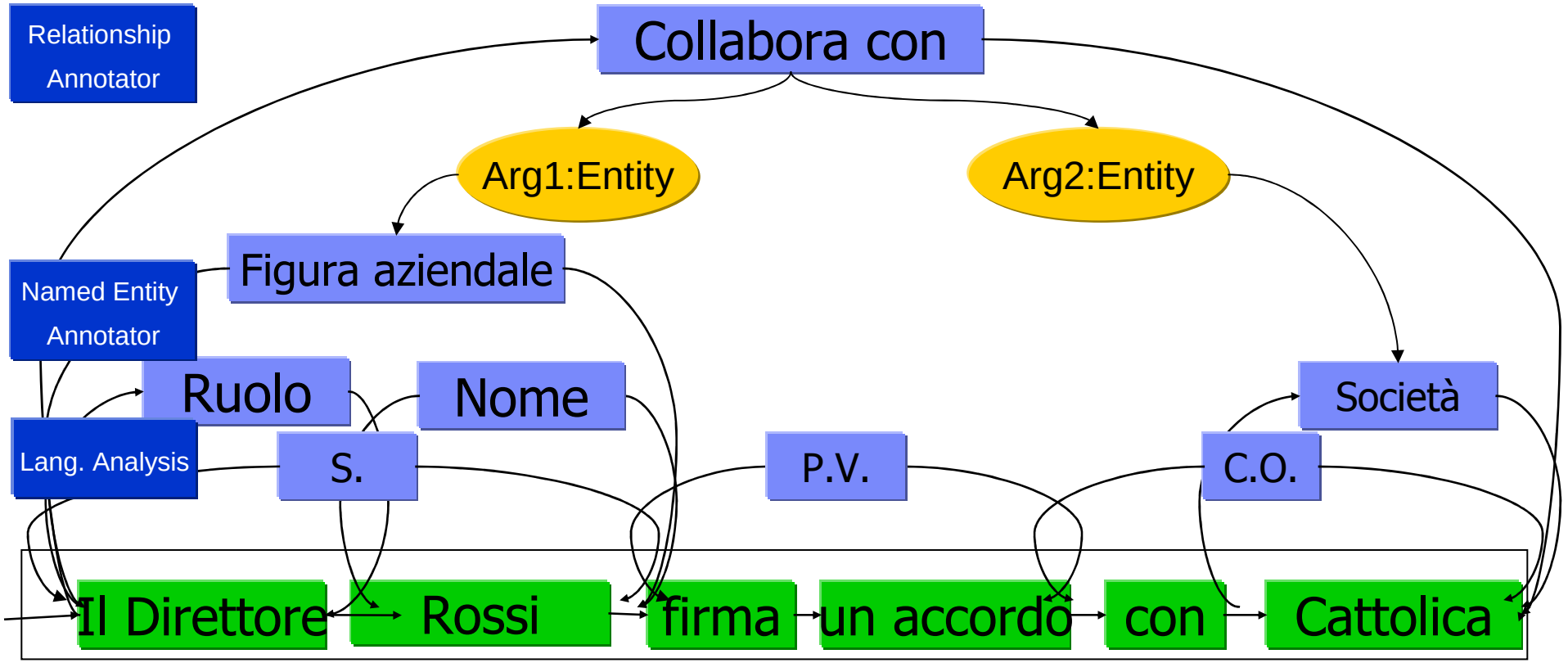


*Basata su **UIMA**, un'architettura aperta e standard di mercato per l'analisi dei testi, creata da IBM. Attualmente **UIMA** è uno standard OASIS ed è implementata da un progetto open-source Apache.*

Documento analizzato
con identificazione dei concetti

- Content Analytics comprende la struttura di una frase e crea degli indici che facilitano l'esplorazione delle "informazioni" contenute nei documenti
- Estrae:
  - **Entità**, identificando item individuali come nomi, compagnie, prodotti, date, persone, luoghi, prezzi, ..
  - **Fatti**, definiti come collezioni di entità che identificano eventi, ruoli, circostanze, azioni, incidenti, ...
  - **Concetti**, definiti come collezioni di fatti che identificano tendenze, comportamenti, (es. sentimenti, reputazioni, frodi, affinità..)

IBM Content Analytics
Analisi dei contenuti



<FiguraAziendale><Ruolo>Direttore</Ruolo><Persona>Rossi</Persona></FiguraAziendale> firma un accordo <Società>Cattolica</Società>

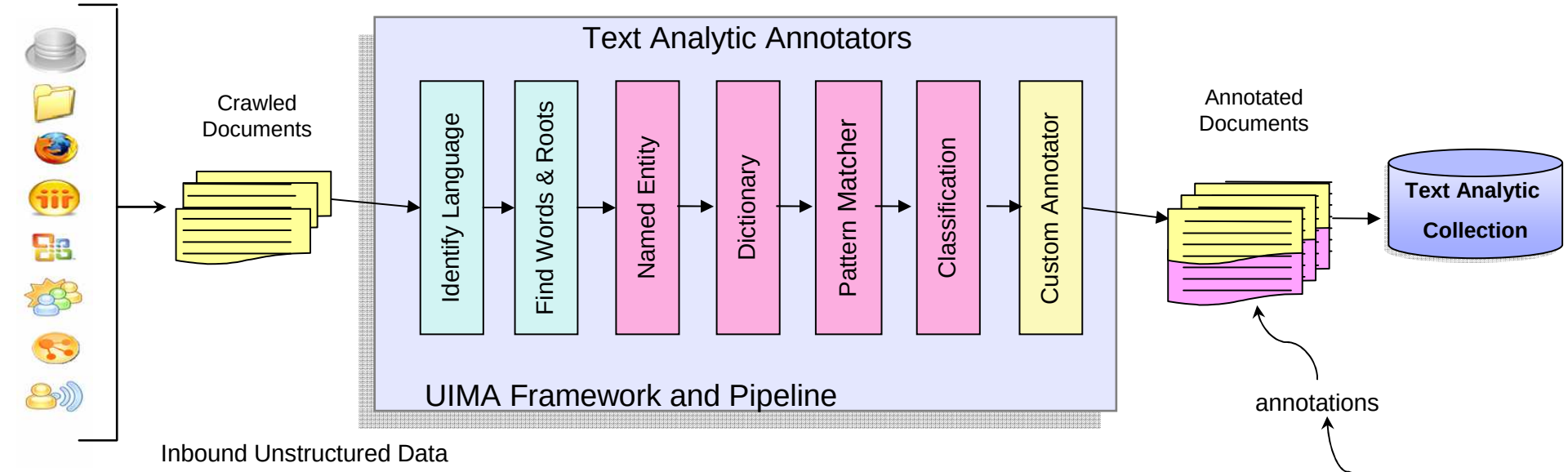
IBM Content Analytics
Unstructured Information Management Architecture (UIMA)

UIMA Background

- Created by IBM in the late 1990s
- Accepted into the Apache Incubator in 2006
- Approved as OASIS Standard in 2009

ICA Implementation

- Multiple languages
- Multiple best of breed analytic technologies
- Open & customizable text analytics pipeline



Device Malfunction Description:

It was reported that during a gastric bypass roux-en-y procedure, on the 3rd firing with a blue load on the stomach there was bleeding. Not sure if staples were formed properly. They over sewed the staple line. There was no PT consequence.

Extracted / Derived Information

Involved Body Part	stomach
Type of Injury	bleeding
Procedure Performed	gastric bypass surgery

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
UIMA and Watson

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
UIMA and Watson

EVEN A BROKEN
ONE OF THESE
ON YOUR WALL
IS RIGHT
TWICE A DAY

**ANCHE SE ROTTO
UNO DI QUESTI
SUL TUO MURO
DICE IL GIUSTO
DUE VOLTE AL GIORNO**

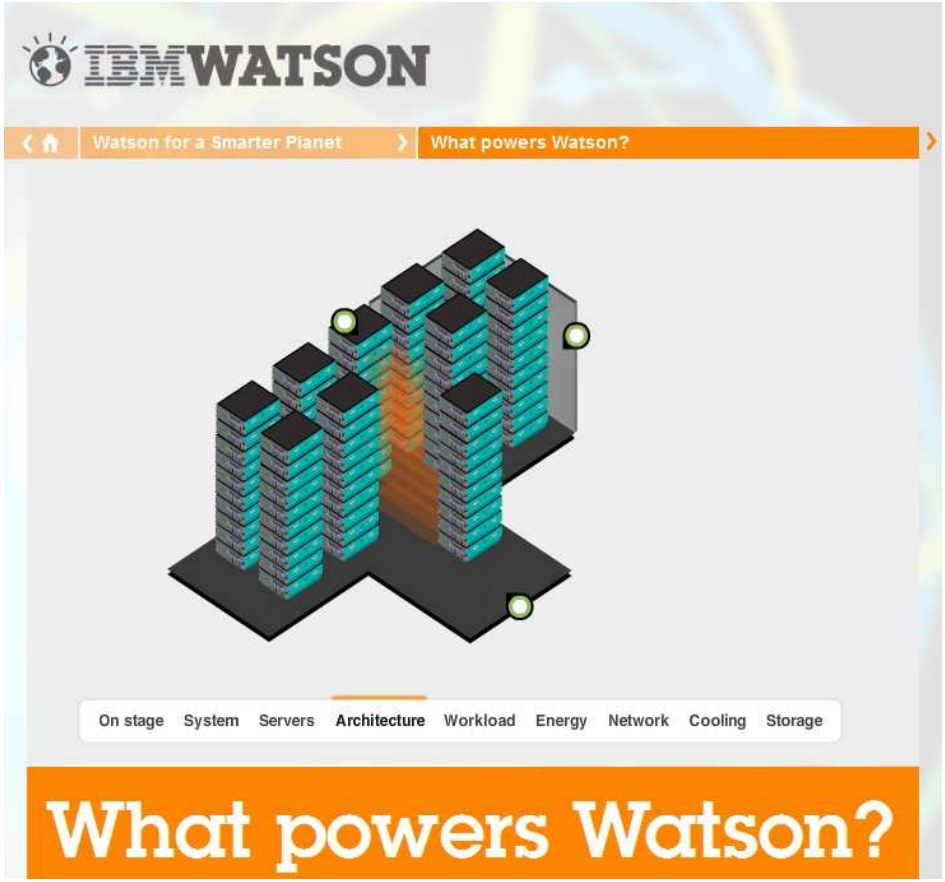
Il supercomputer che partecipa al quiz tv e sbaraglia i concorrenti umani

A «Jeopardy», sulla Nbc, dalle risposte bisogna risalire alle domande. Battuti i due campioni dello scorso anno

la sfida a Jeopardy

MILANO - Qualcuno l'ha definito «il discendente più brillante di Hal 9000», il computer dell'astronave di *2001: Odissea nello spazio*, perché come la macchina presente nel film di Stanley Kubrick ha una singolare «intelligenza artificiale». Si chiama Watson, è un supercomputer ideato dall'Ibm e in questi giorni è il protagonista assoluto di *Jeopardy*, uno dei giochi a quiz più famosi d'America, che va in onda sulla Nbc dal 1964. Nonostante gareggi con Brad Rutter e Ken Jennings, i due più forti campioni dell'anno passato, Watson sta dimostrando di essere il numero uno e nei primi due giorni di gara ha letteralmente sbaragliato gli avversari.

IBM Content Analytics
UIMA and Watson



- Watson è disegnato secondo lo standard UIMA. Questa architettura software è lo standard per lo sviluppo di programmi che analizzano informazioni non strutturate come testo, audio ed immagini.

- Watson utilizza Apache UIMA per sfruttare il “natural language processing” in parallelo sui processori POWER7, consentendo di effettuare migliaia di computazioni analitiche simultaneamente sul cluster di server per rispondere ad ogni domanda nel più breve tempo possibile.

Link utili per approfondimenti :
- http://www-03.ibm.com/innovation/us/watson/?S_CMP=swnews211
- http://www-03.ibm.com/innovation/us/watson/watson-for-a-smarter-planet/watso

IBM Content Analytics
Funzionalità

Scopre:

Identifica ed etichetta automaticamente gli attributi e le entità fondamentali all'interno dei contenuti,

Analizza tutte le sorgenti di contenuti, estraendo le parole chiave ed elementi delle frasi

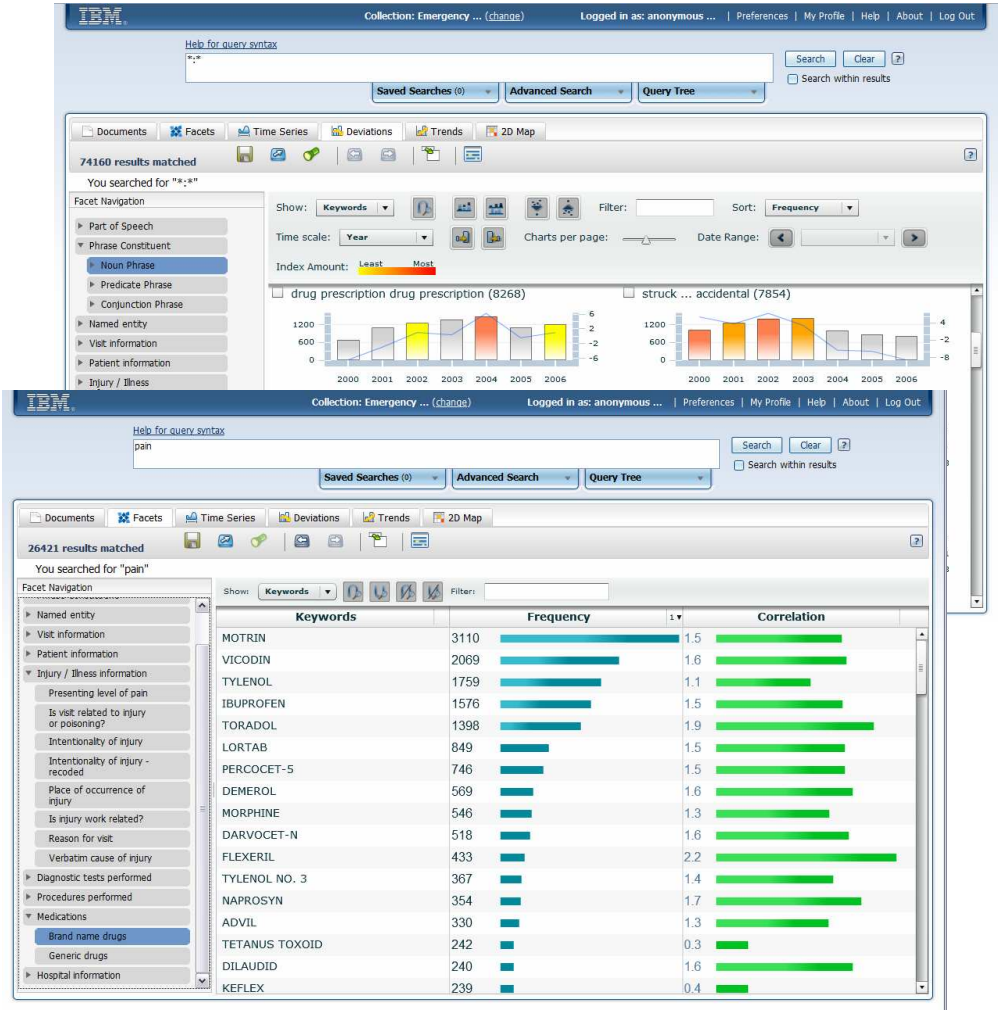
Raffina:

Consente la navigazione ed il drill-down sulla base degli attributi, delle entità e delle dimensioni estratte

Visualizza:

Utilizza una modalità avanzata di visualizzazione che consente il mining.

Evidenzia deviazioni ed anomalie in modo da consentire di prendere decisioni ed intraprendere azioni correttive

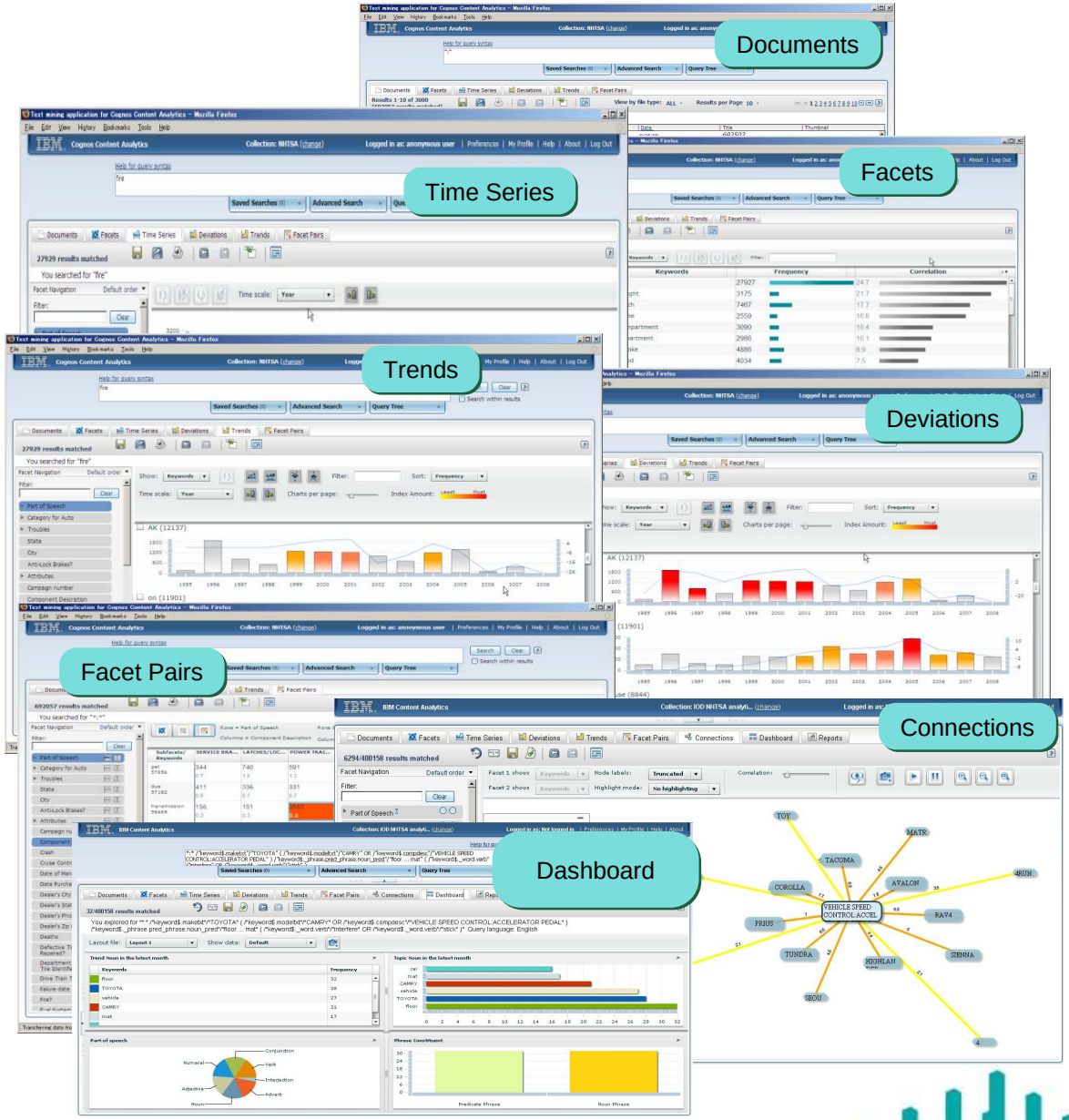


IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining

- **Diverse viste dei dati**
  - Documenti
  - Facets
  - Serie temporale
  - Deviazioni
  - Trend
  - Facet Pairs
  - Connections
  - Dashboard

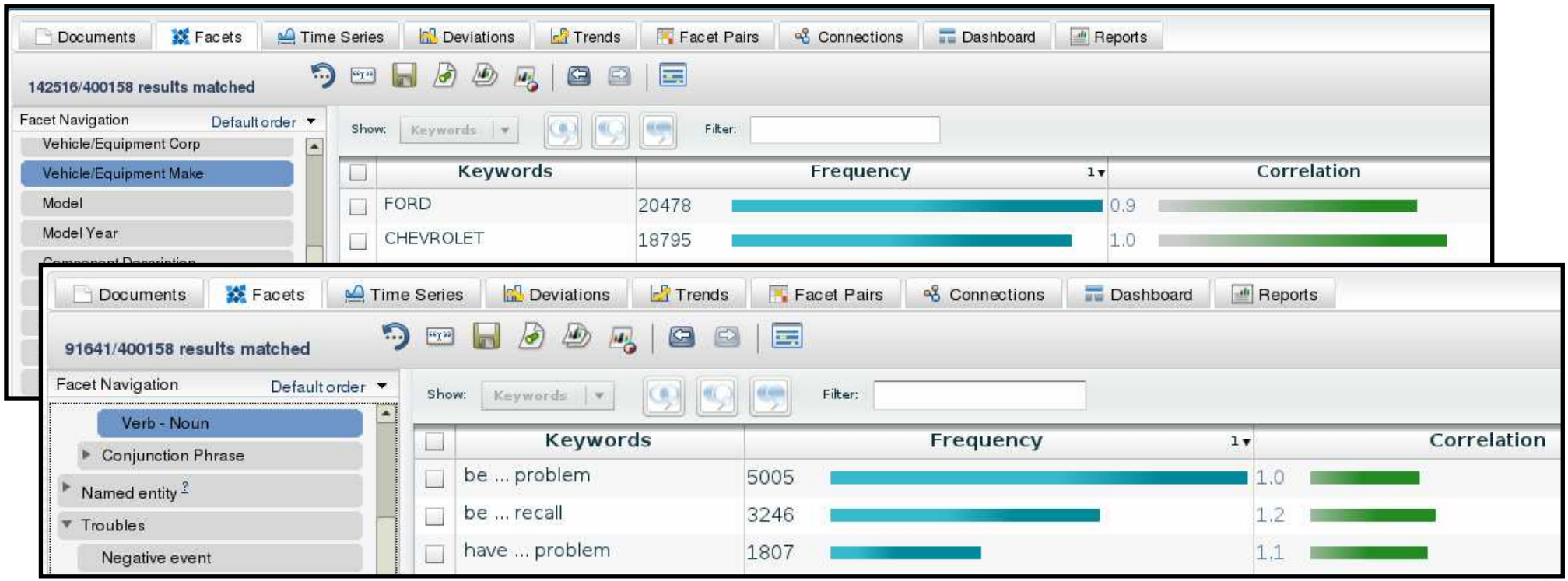


IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining: FACETs

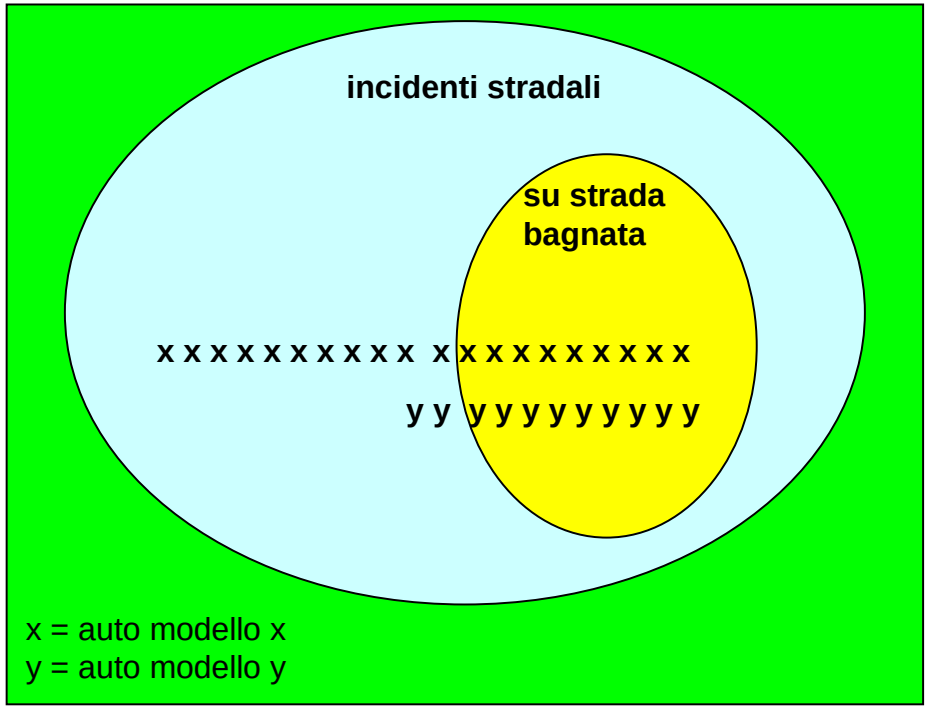
- Attraverso le facets è possibile esplorare i contenuti sia attraverso I dati struturati disponibili che le informazioni estratte dai documenti
- Consente di visualizzare i contenuti aggregati per keyword
- Mostra frequenza ed indici di correlazione tra facets
- Drill-down tramite l'aggiunta della parola chiave selezionata alla condizione di ricerca



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining: ***Frequenza e Correlazione***



Frequenza auto x = 20
Frequenza auto y = 11

SU STRADA BAGNATA
frequenza auto x = 10
frequenza auto y = 9

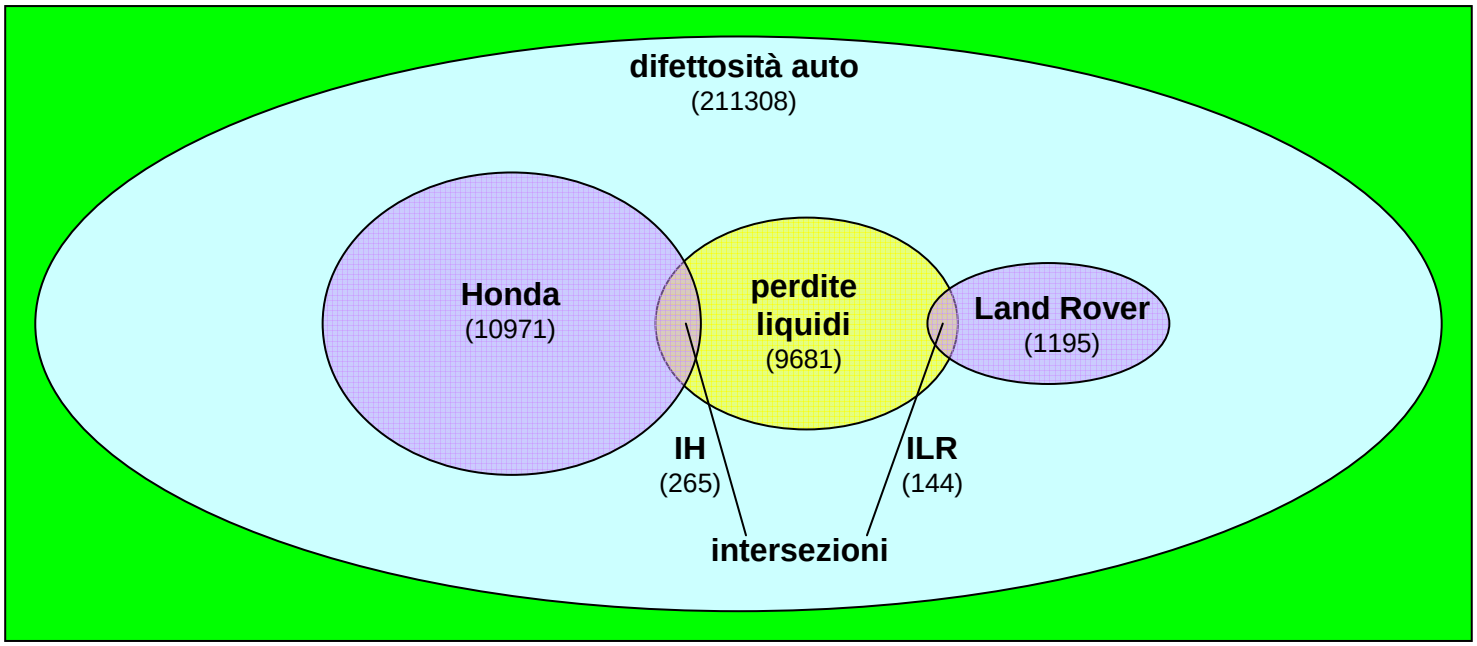
NEL CONTESTO
incidenti stradali su strada bagnata

VALORE CORRELAZIONE
PIU' ELEVATO PER AUTO y

- La frequenza esprime il numero di documenti che contengono la keyword (evento, oggetto, ecc.)
- La correlazione indica la rilevanza reciproca di due sottoinsiemi all'interno dell'aggregato preso in considerazione.

IBM Content Analytics

Funzionalità di Text Mining: ***Frequenza e Correlazione***



densità = occorrenze sottoinsieme / totale

correlazione = $\frac{\text{densità intersezione}}{\text{prodotto densità sottoinsiemi interessati}}$ * indice di correlazione

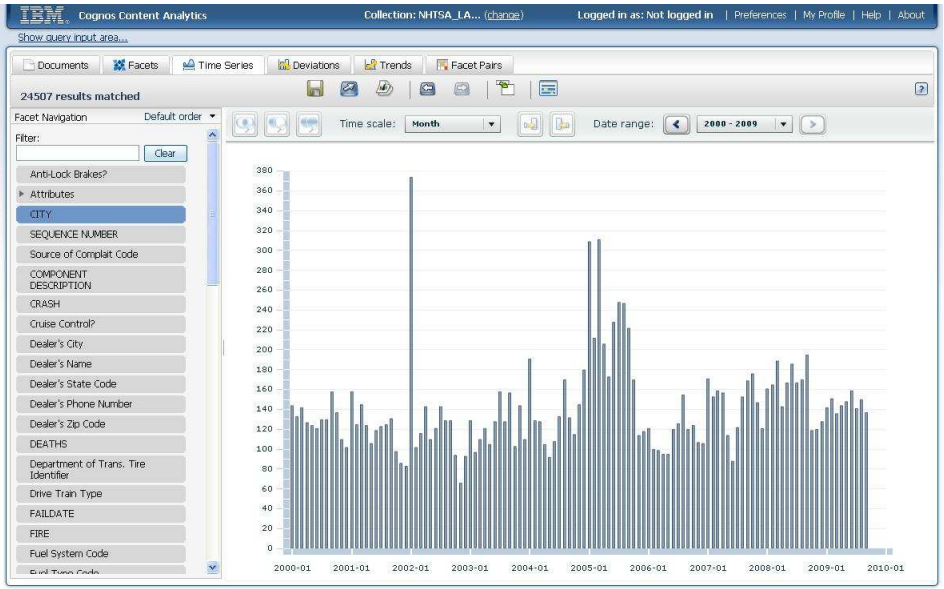
	occorrenze	densità
Honda	10971	0,05190
Land Rover	1195	0,00565
perdite liquidi	9681	0,04580
IH	265	0,00125
ILR	144	0,00068

correlazione Honda - perdite = $\frac{0,00125}{0,05190 * 0,04580}$ * 0,77 = 0,4

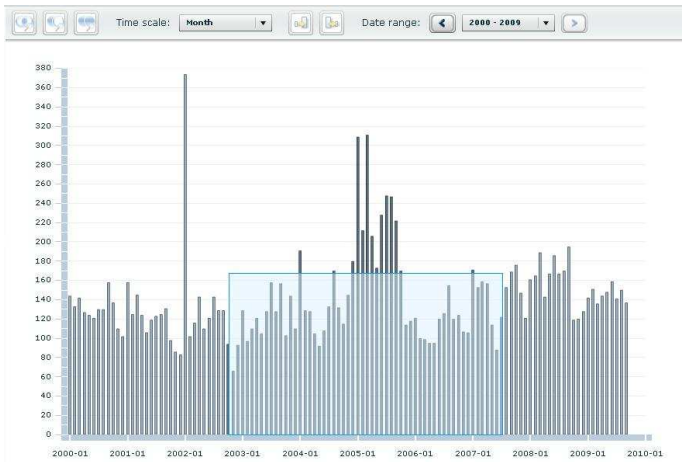
correlazione Land Rover perdite = $\frac{0,00068}{0,00565 * 0,04580}$ * 0,80 = 2,1

IBM Content Analytics

Funzionalità di Text Mining: ***Serie temporali***

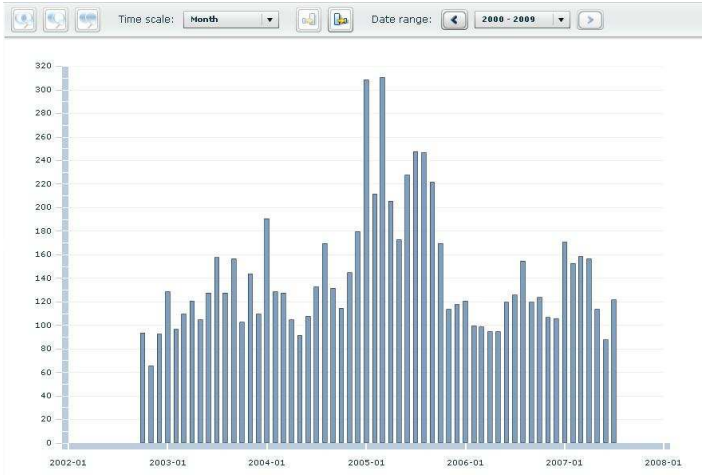


- Mostra quanto spesso i documenti soddisfano la condizione di ricerca in un determinato periodo di tempo
- Consente di raffinare la ricerca tramite la selezione di una o più vincoli di date



Zoom in

Zoom out

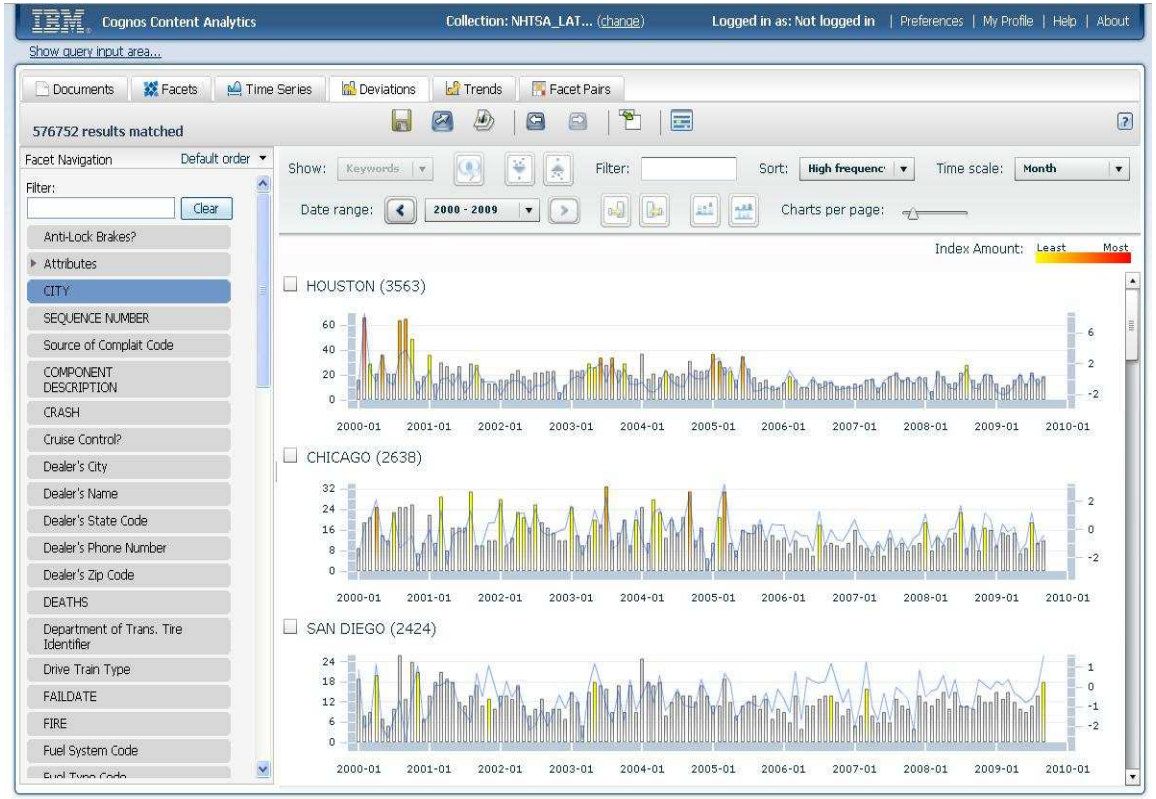


IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics

Funzionalità di Text Mining: Deviazioni e Trend



- **Mostra la frequenza delle parole chiave ed i valori calcolati**
- Deviazioni
  - Mostra incrementi inaspettati della frequenza di una parola chiave nel tempo
- Trend
  - Mostra incrementi inaspettati della frequenza di un insieme di parole chiave (facet) nel tempo

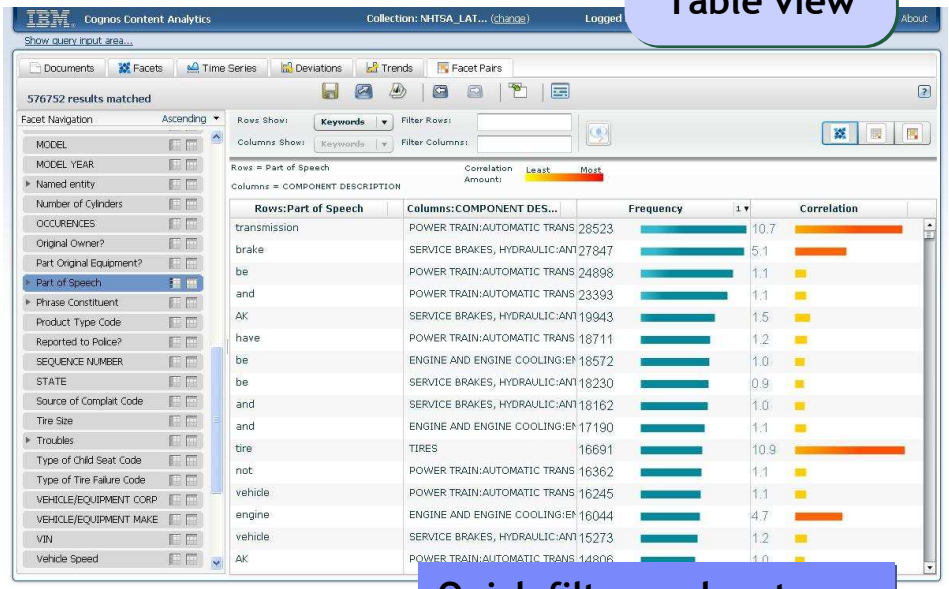
- Permette di raffinare la ricerca usando il valore della facet e la data
- Ordina, filtra, zoom-in, zoom-out

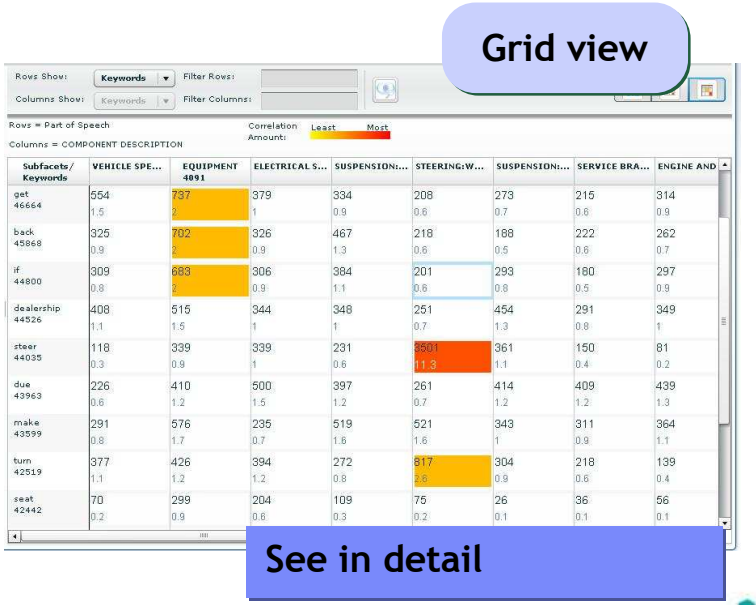
IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining: ***Facet Pairs***

- Mostra la correlazione tra parole chiave appartenenti a due facet diverse
- 3 modalità di visualizzazione





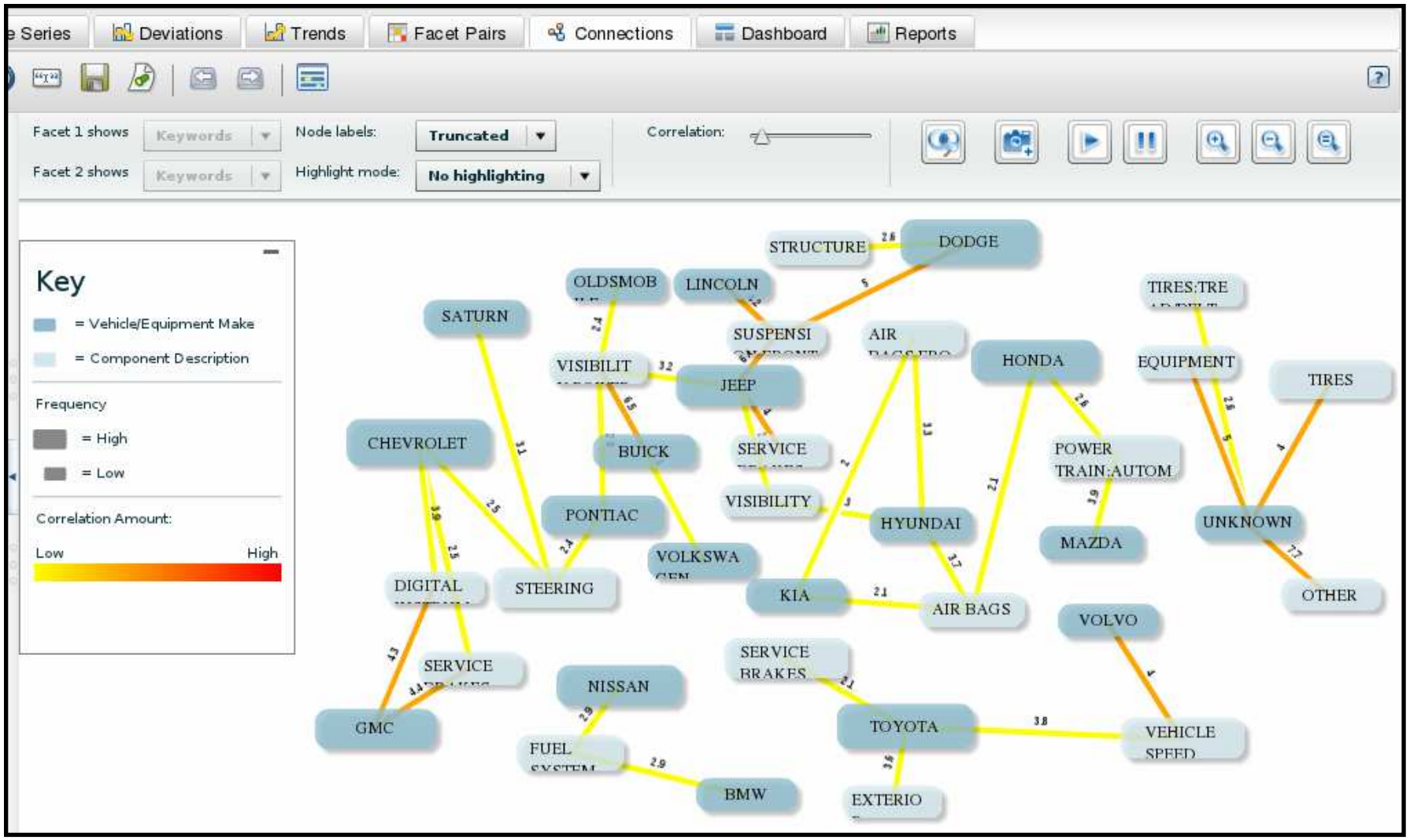
IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics

Funzionalità di Text Mining: **Connections**

La vista **Connections** mostra il grafico delle relazioni delle keyword relative alle facet pairs selezionate



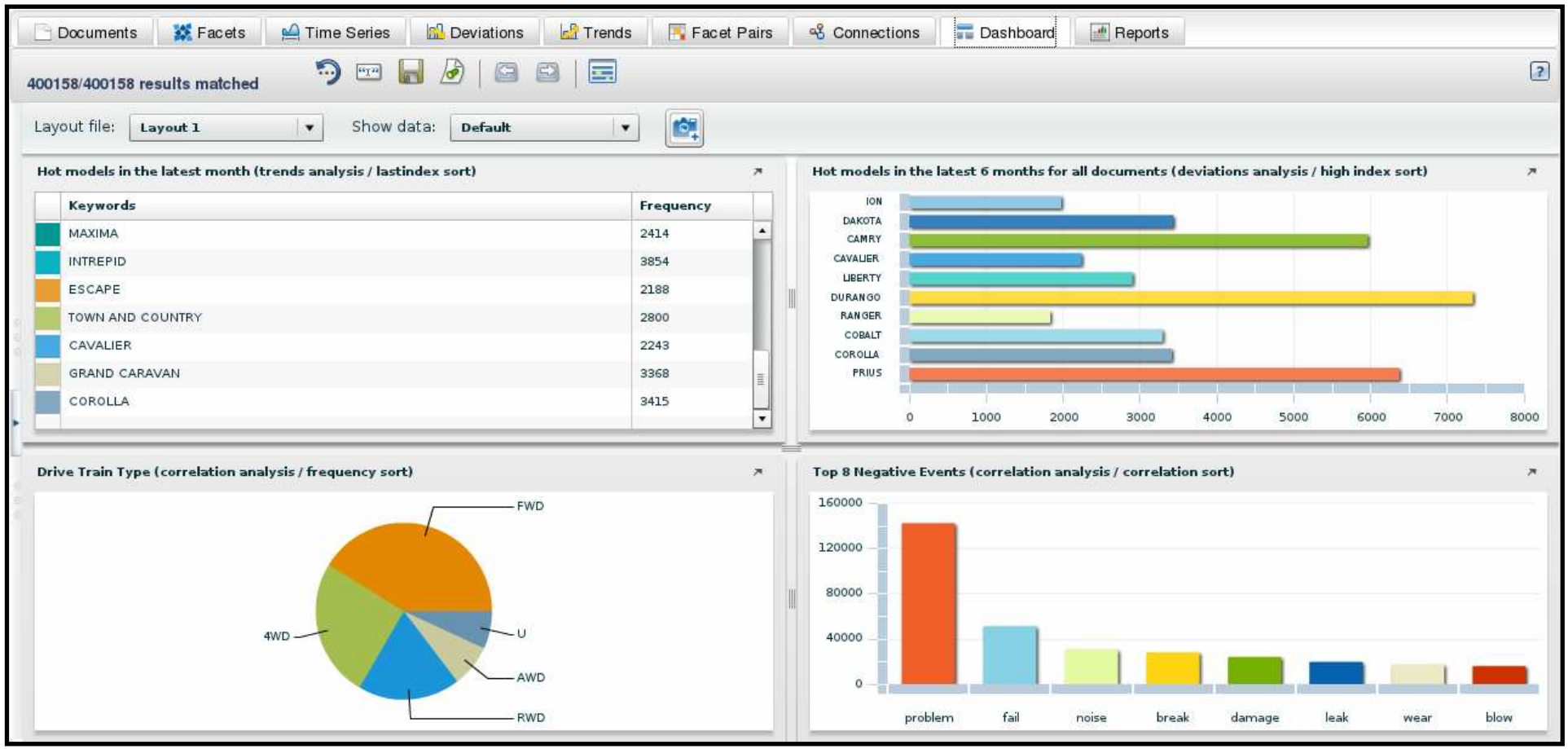
IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics

Funzionalità di Text Mining: *Dashboard*

La vista **Dashboard** mostra varie tavole predefinite in vista unica, per visualizzare e correlare vari aspetti dei dati analizzati



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Fonti Documentali analizzabili

Web

- HTTP
- HTTPS
- WebSphere Portal Web pages
- WebSphere Portal Document Manager
- IBM Workplace Web Content Management
- Newsgroup (NNTP)

Collaboration

- Lotus Notes databases
- Domino.doc
- QuickPlace
- Quickr
- Lotus WCM
- Lotus Connections
- MS Exchange public folders
- Microsoft SharePoint Services (2003 & 2007)
- Windows file systems
- UNIX file systems

ECM

FileNet

documentum a division of EMC

Hummingbird

OpenText Livelink

IBM CM & CMOD

Database

DB2

Sybase

Oracle

Informix

Microsoft

MySQL

IBM Websphere MQ with Event Publisher

Mainframe: VSAM, IMS, CA-Datacom, Software AG Adabas

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

Native security support

Domande ?