ΔΙΑΧΕΙΡΙΣΗ WEB ΠΕΡΙΕΧΟΜΕΝΟΥ & ΓΛΩΣΣΙΚΑ ΕΡΓΑΛΕΙΑ

Information Extraction(Εξαγωγή Πληροφορίας)

Πληροφορία κειμένου

- User Generated Content: μη δομημένη πληροφορία κειμένου(UGC)
- Ανάγκη εξαγωγής δομημένης πληροφορίας για επεξεργασία, διαχείριση και εξόρυξη δεδομένων
- **Εξαγωγή Πληροφορίας**

Εξαγωγή Πληροφορίας

Στόχος: **Συμπλήρωση πεδίων ΒΔ από μέρη του κειμένου**

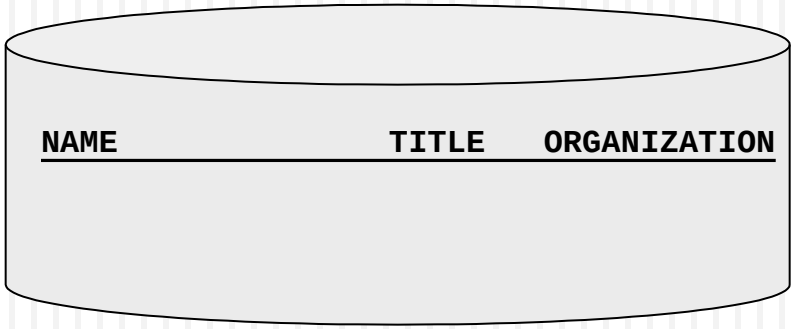
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



Εξαγωγή Πληροφορίας

Στόχος: **Συμπλήρωση πεδίων ΒΔ από μέρη του κειμένου**

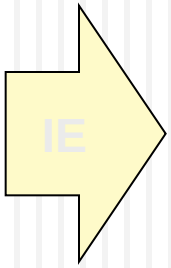
October 14, 2002, 4:00 a.m. PT

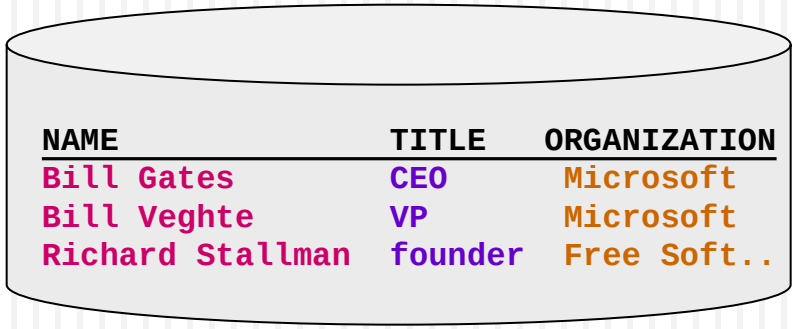
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…





NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Εξαγωγή Πληροφορίας

Τεχνικές:

Εξαγωγή Πληροφορίας =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Εξαγωγή ονοματικών οντοτήτων

Εξαγωγή Πληροφορίας

Τεχνικές:

Εξαγωγή Πληροφορίας =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Εξαγωγή Πληροφορίας

Τεχνικές:

Εξαγωγή Πληροφορίας =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates
Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

Εξαγωγή Πληροφορίας

Τεχνικές:

Εξαγωγή Πληροφορίας =
segmentation + classification + association + clustering

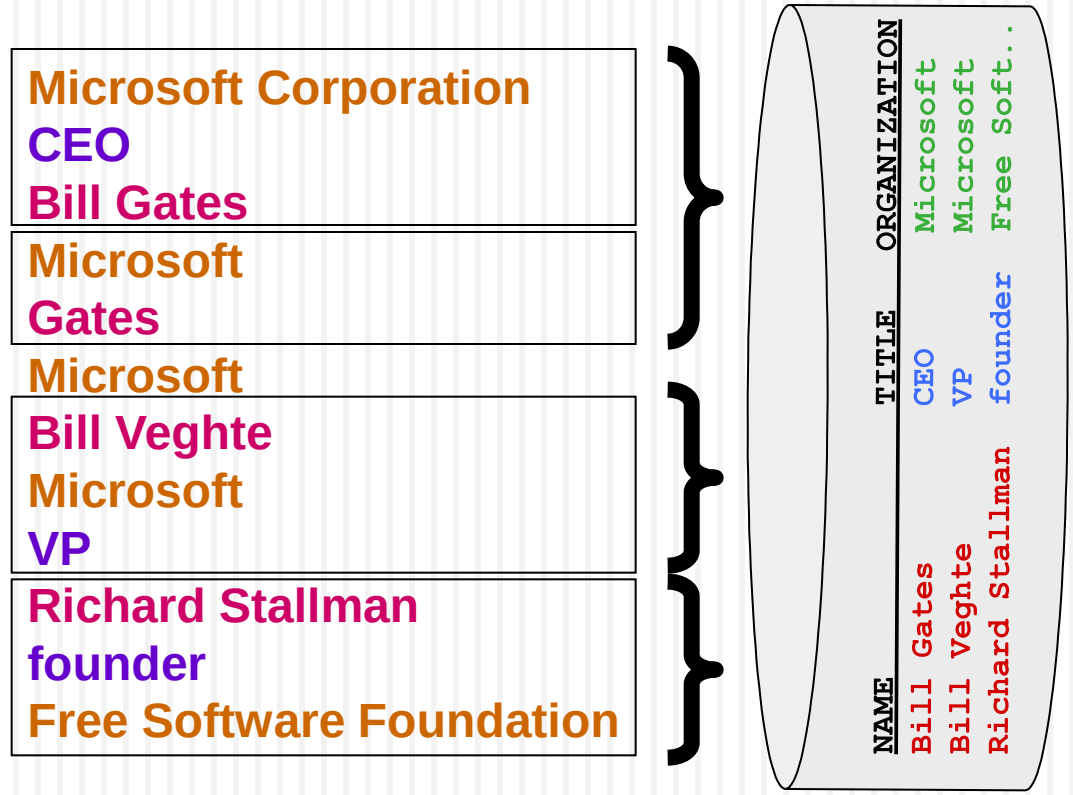
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

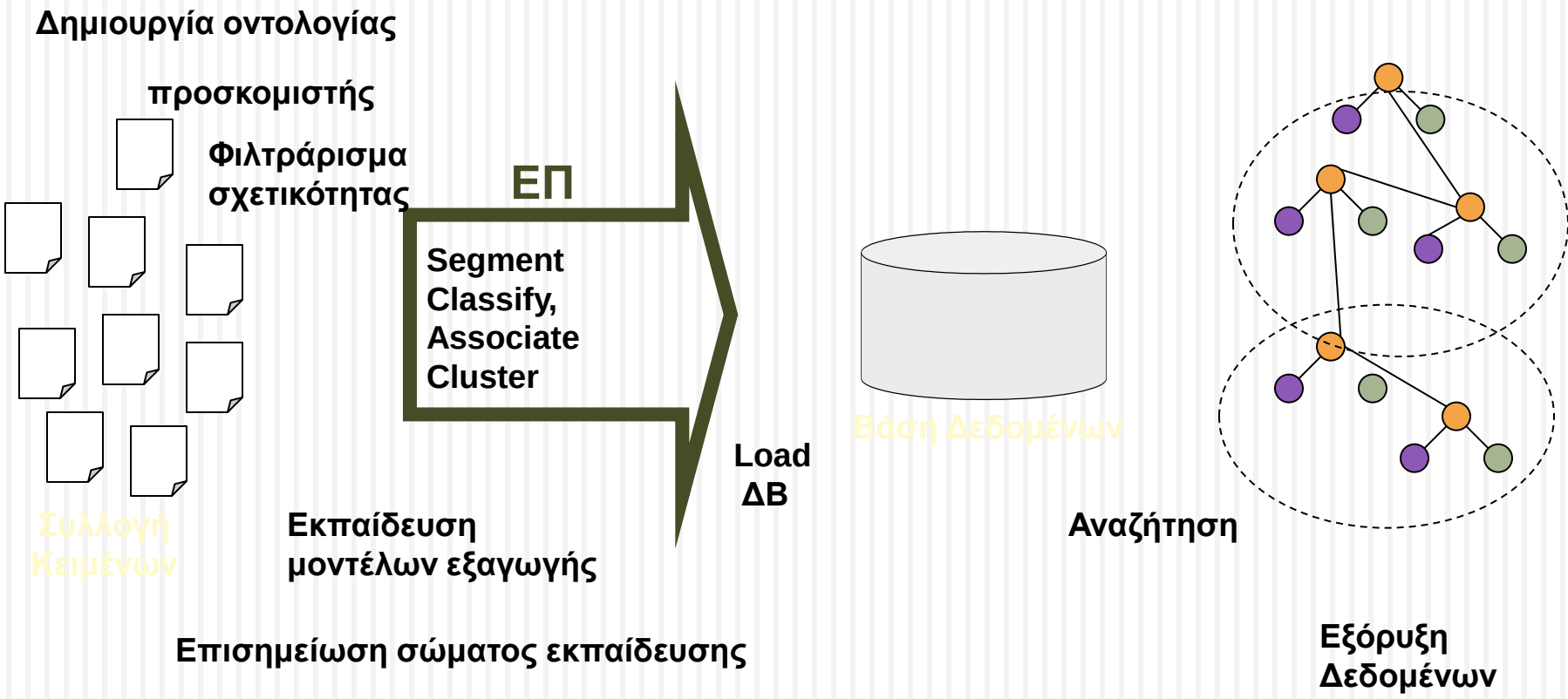
Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



Εξαγωγή Πληροφορίας από περιεχόμενο



Επισκόπηση θεμάτων

- Βήματα για την εξαγωγή πληροφορίας
  - Επισημείωση οντοτήτων
  - Εξαγωγή συσχετίσεων
  - Εξαγωγή γεγονότων
- Κλιμάκωση εξαγωγής πληροφορίας
  - Κλιμάκωση για δεδομένα μεγάλου όγκου
  - Ανοιχτά ζητήματα κλιμάκωσης

Βήματα Εξαγωγής Πληροφορίας

- Εξαγωγή οντοτήτων και συσχετίσεων
  - **Οντότητες**: ονοματικές και γενικές
  - **Συσχετίσεις**: σύνδεση οντοτήτων
  - **Γεγονότα**: αποτελούνται από πλειάδες πολλών σχέσεων

- Βήματα εξαγωγής:
  - **Προεπεξεργασία**: διαχωρισμός προτάσεων, συντατική ανάλυση
  - Δημιουργία **κανόνων** ή **εξαγωγή προτύπων**: χειρωνακτικά, μηχανική μάθηση, υβριδικά
  - **Εφαρμογή** προτύπων ή κανόνων για **εξαγωγή** νέας πληροφορίας
  - **Υστερο-επεξεργασία** και **ενσωμάτωση** πληροφορίας

Επισημείωση Οντοτήτων

- Ανίχνευση αναφορών σε οντότητες στο κείμενο (π.χ. Ονόματα ανθρώπων, τοποθεσίες,κτλ)
- Χειρωνακτικά vs. Τεχνικές μηχανικής μάθησης
- Η βέλτιστη προσέγγιση εξαρτάται από τον τύπο οντοτήτων και το πεδίο
  - **Κλειστού τύπου** (π.χ., γεωγραφικές περιοχές, ονόματα ασθενειών): χειρωνακτικά+ λεξικά
  - **Συντακτική** (π.χ., τηλεφωνικοί αριθμοί, ταχυδρομικοί κώδικες): regular expressions
  - **Σημασιολογική** (π.χ., ονόματα ανθρώπων): συνδυασμός περιεχομένου, συντακτικών γνωρισμάτων, λεξικών, κα.

Παράδειγμα εξαγωγής οντοτήτων

Citation

Ronald Fagin, Combining Fuzzy Information from Multiple Systems, Proc. of ACM SIGMOD, 2002

Segment(s_i)	Sequence	Label(s_i)
S_1	Ronald Fagin	Author
S_2	Combining Fuzzy Information from Multiple Systems	Title
S_3	Proc. of ACM SIGMOD	Conference
S_4	2002	Year

Χειρωνακτικές μέθοδοι

- Αποδοτικές για ορισμένες περιπτώσεις (π.χ αναγνώριση τιμών, ταχυδρομικού κώδικα, ονόματα συνεδρίων, κτλ)

ContactPattern ← RegularExpression(Email.body,”can be reached at”)
[IBM Avatar]

- Ζητήματα κλιμάκωσης:
  - Κοπιαστική εργασία
  - Domain-specific
  - Corpus-specific
  - Ακριβό το ταίριασμα των κανόνων

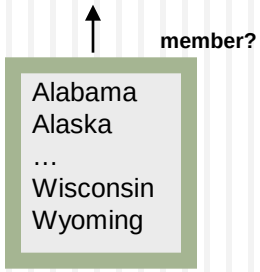
Τεχνικές Μηχανικής Μάθησης

- Αποτελεσματικές όταν υπάρχουν μεγάλα σώματα εκπαίδευσης
- Αποτυπώνουν σύνθετα πρότυπα που είναι δύκολο να κωδικοποιηθούν χειρωνακτικά
  - Αν μια αξιολόγηση είναι θετική ή αρνητική
  - Χωρίς τοπικές εξαρτήσεις

Μοντέλα αναπαράστασης [Cohen and McCallum, 2003]

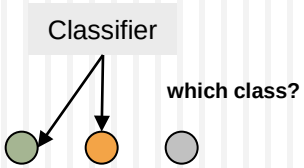
Lexicons

Abraham Lincoln was born in Kentucky.



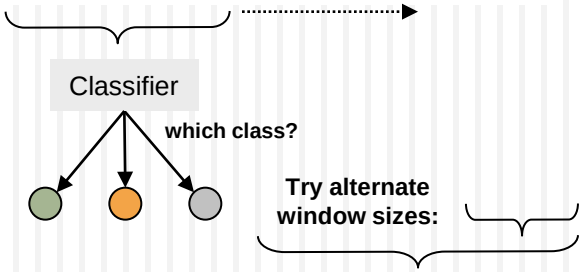
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

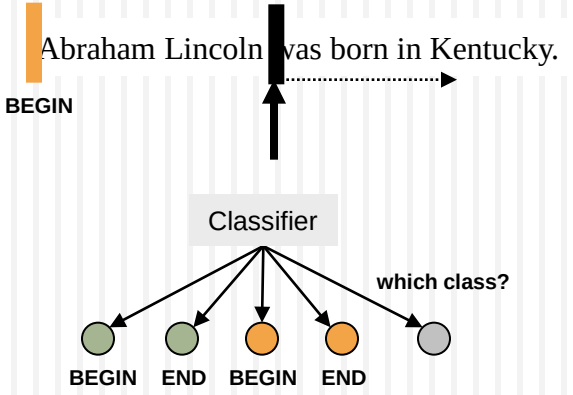


Sliding Window

Abraham Lincoln was born in Kentucky.

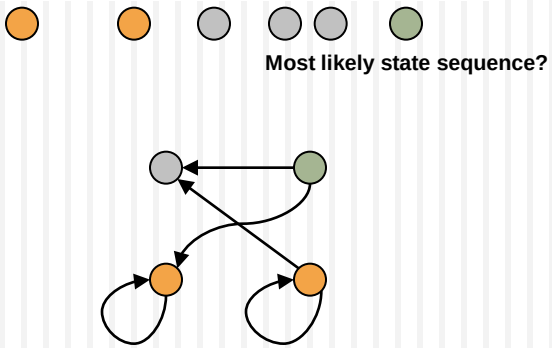


Boundary Models



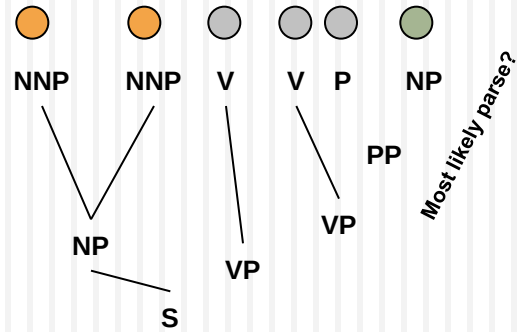
Finite State Machines

Abraham Lincoln was born in Kentucky.



Context Free Grammars

Abraham Lincoln was born in Kentucky.



…and beyond

Εξαγωγή Συσχετίσεων

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly **Ebola** epidemic in **Zaire**, is finding itself hard pressed to cope with the crisis...

Relation Extraction

Disease Outbreaks relation

Date	*Disease Name*	*Location*
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

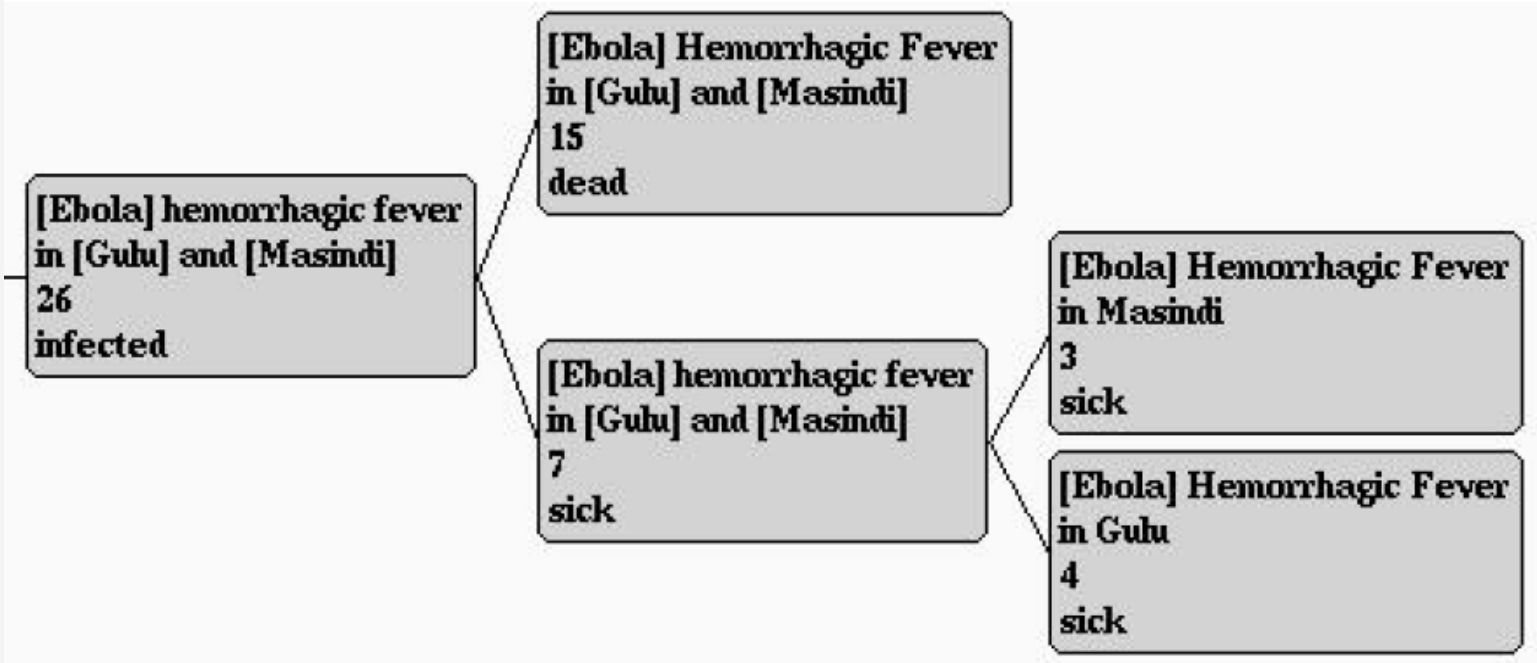
Τεχνικές εξαγωγής συσχετίσεων

Μηχανική Μάθηση

- **Εποπτευόμενη**: εκπαίδευση συστήματος σε χειρωνακτικά επισημειωμένα δεδομένα
- **Ημι-εποπτευόμενη**: εκπαίδευση συστήματος με bootstrapping από "seed" παραδείγματα
- **Υβριδικά** ή **διαδραστικά** συστήματα:
  - Ειδικοί αλληλεπιδρούν με αλγορίθμους μηχανικής μάθησης για να διορθώσουν επαναλληπτικά κανόνες και πρότυπα
  - Οι διαδράσεις περικλείουν παραδείγματα επισημείωσης, κανόνες τροποποίησης ή συνδυασμούς

Εξαγωγή Γεγονότων

- Παρόμοια με την εξαγωγή συσχετίσεων, αλλά:
  - Τα γεγονότα μπορεί να είναι εμφωλευμένα
  - Μεγαλύτερη πολυπλοκότητα
  - Συχνά απαιτείται επίλυση συν-αναφορών, αποσαφήνιση και συμπερασμός
- Π.χ.integrated disease outbreak event



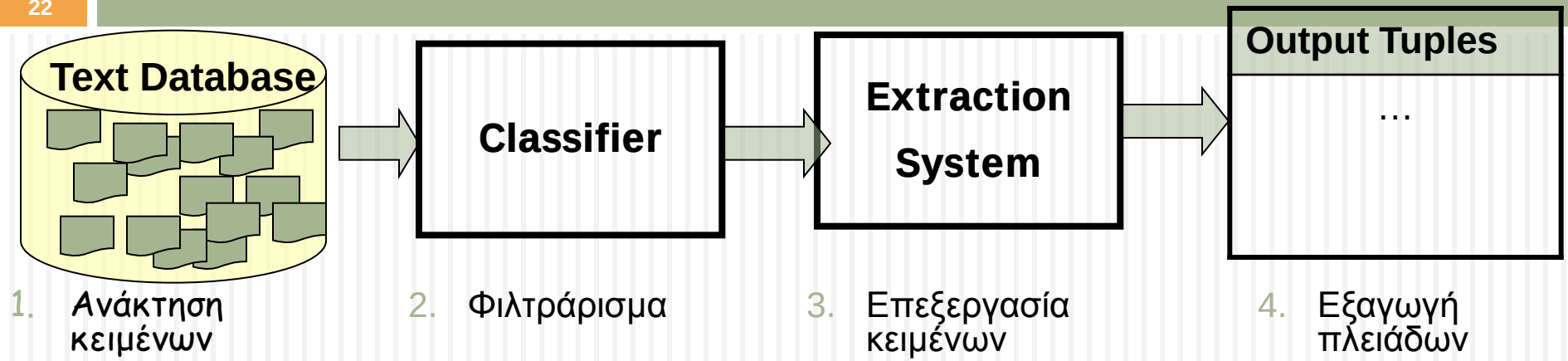
Προκλήσεις στην εξαγωγή γεγονότων

- Η πληροφορία βρίσκεται σε πολλά κείμενα
  - Απούσες ή λανθασμένες τιμές
  - Συνδυασμός πλειάδων για σύνθετα γεγονότα
  - Απουσία μοναδικού κλειδιού για την ομαδοποίηση διπλότυπων κατά το διαχωρισμό παρόμοιων αλλά διαφορετικών οντοτήτων
  - Ασάφεια: διαφορετικές οντότητες με κοινό όνομα (Kennedy)

Κλιμάκωση εξαγωγής πληροφορίας

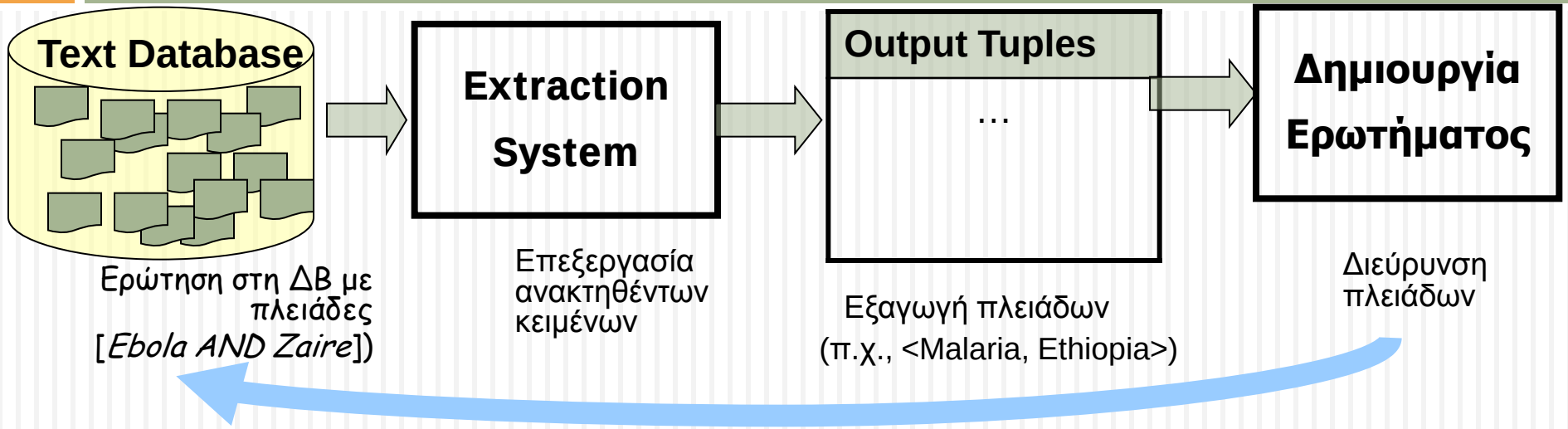
- Διαστάσεις κλιμάκωσης
  - Μέγεθος δεδομένων
    - Ακριβή η εφαρμογή κανόνων/προτύπων
    - Αποτελεσματικοί τρόποι επιλογής σχετικών κειμένων
  - Προσβασιμότητα κειμένων
    - Αόρατος ιστός: πρόσβαση μέσω διεπαφής
    - Δυναμικά δεδομένα
  - Ετερογένεια πηγής
    - Ακριβή η εκμάθηση προτύπων για κάθε πηγή
    - Απαιτούνται πολλοί κανόνες
  - Διαφοροποίηση πεδίου
    - Εξαγωγή πληροφορίας από κάθε πεδίο

Αποδοτική Εξαγωγή Πληροφορίας



- Κανόνας 80/20: λίγοι και απλοί κανόνες για την εξαγωγή των περισσότερων στιγμιοτύπων
- Εκπαίδευση κατηγοριοποιητή για απαλοιφή των μη σχετικών κειμένων χωρίς εξέταση
- Διαμοιρασμός κοινών επισημειώσεων (ετικέτες οντοτήτων) για πολλαπλές εργασίες εξαγωγής

Επαναληπτική διεύρυνση συνόλου



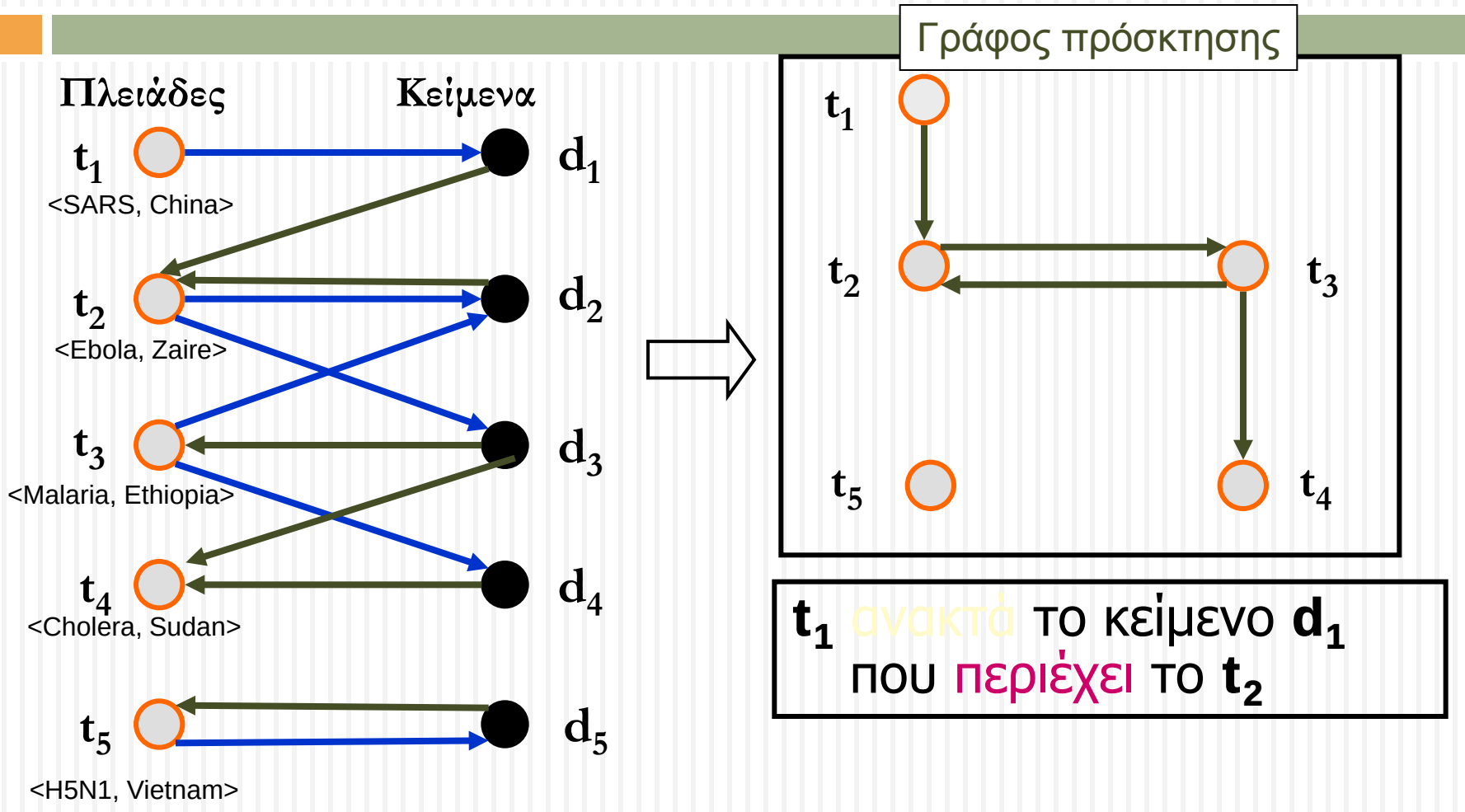
Execution time = |Retrieved Docs| * (R + P) + |Queries| * Q

Χρόνος ανάκτησης κειμένου

Χρόνος επεξεργασίας κειμένου

Χρόνος απάντησης ερωτήματος

Πρόσκτηση μέσω ερωτημάτων



Άνω όριο ανάκλησης: καθορίζεται από το μέγεθος του πιο συνδεδεμένου στοιχείου

Όρια πρόσκτησης

QXtract

1. Πρόσκτηση δείγματος κειμένου με “μάλλον αρνητικά” και “μάλλον θετικά” παραδείγματα
2. Επισημείωση δειγμάτων κειμένων χρησιμοποιώντας το σύστημα εξαγωγής ως “μαντείο”
3. Εκπαίδευσης κατηγοριοποιητών για την “αναγνώριση” χρήσιμων κειμένων
4. Δημιουργία ερωτημάτων από τους κανόνες κατηγοριοποίησης