

Οργάνωση Βάσεων Βιοϊατρικών Δεδομένων Εξόρυξη Γνώσης Βιοϊατρικών Δεδομένων

Σεμινάριο 10: Εξαγωγή πληροφοριών, συστήματα ερωταποκρίσεων

Ευάγγελος Καρκαλέτσης,
Αναστασία Κριθαρά, Γεώργιος Πετάσης

Εργαστήριο Τεχνολογίας Γνώσεων & Λογισμικού,

Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών, Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”

Τηλ.: 210-6503197, Fax: 210-6532175, {vangelis, akrihara, petasis}@iit.demokritos.gr

Ακαδημαϊκό Έτος: 2016 – 2017

Οι διαφάνειες αυτού του μαθήματος βασίζονται εν μέρει σε διαφάνειες του
Ίωνα Ανδρουτσόπουλου για το μάθημα της Γλωσσικής Τεχνολογίας

<http://www.aueb.gr/users/ion/>

Περιεχόμενα

- Εξαγωγή πληροφοριών από κείμενα.
- Βασικές έννοιες ανάκτησης πληροφοριών.
- Συστήματα ερωταποκρίσεων για δομημένα δεδομένα και συλλογές εγγράφων.
 - BioASQ
- Βιοιατρικές οντολογίες & δεδομένα

Εξαγωγή πληροφοριών (information extraction)

3 Δεκ. 2008: Σημαντικά νέα χθες από την General Company Ελλάς, μια από τις μεγαλύτερες ελληνικές κατασκευαστικές εταιρείες. Η GC ανακοίνωσε ότι θα αγοράσει το 42% των μετοχών της Small Company Ltd, μιας Βρετανικής εταιρείας που ειδικεύεται στο χώρο των σιδηρών κατασκευών.



αγοραστής	GCE
αγοραζόμενος	SCL
μερίδιο	0.42
ημερομηνία ανακοίνωσης	2/12/2008

Εξαγωγή πληροφοριών: πρώτα στάδια

<s> <date norm="3/12/2008"> 3 Δεκ. 2008 </date> : </s>

<s> Σημαντικά νέα <date norm="2/12/2008"> χθες </date> από την <company id="GCE"> General Company Ελλάς </company>, μια από τις μεταλύτερες ελληνικές κατασκευαστικές εταιρείες. </s>

<s> Η <company id="GCE"> GC </company> ανακοίνωσε ότι <verb base="αγοράζω"> θα αγοράσει </verb> το <percent norm="0.42"> 42% </percent> των μετοχών της <company id="SCL"> Small Company Ltd </company>, μιας Βρετανικής εταιρείας που ειδικεύεται στο χώρο των σιδηρών κατασκευών. </s>

Εξαγωγή πληροφοριών: πρώτα στάδια

- Προεπεξεργασία:
 - Λεκτικές μονάδες (tokens), χωρισμός σε προτάσεις, χειρισμός ετικετών HTML, XML κλπ.
- Μορφολογική ανάλυση:
 - Μέρη του λόγου, λήμματα (κύριοι τύποι λέξεων) κλπ.
- Αναγνώριση χρονικών, χρηματικών εκφράσεων κλπ.
 - και μετατροπή τους σε κανονικές μορφές.
- Αναγνώριση ονομάτων οντοτήτων (π.χ. προσώπων, εταιρειών, τοποθεσιών, γονιδίων, ασθενειών).
 - Και ταίριασμα διαφορετικών μορφών του ιδίου ονόματος (π.χ. «Ο.Τ.Ε.» και «Οργανισμός Τηλεπικοινωνιών της Ελλάδος», «ο κ. Γεώργιος Μιχαηλίδης» και «ο Μιχαηλίδης»).

Εξαγωγή πληροφοριών: επόμενα στάδια

- Επίλυση **αναφορικών** εκφράσεων.
 - Π.χ. «**Η εταιρεία** ανακοίνωσε επίσης ...»
- Εντοπισμός **σχέσεων** μεταξύ οντοτήτων:
 - Στην πιο απλή περίπτωση χρησιμοποιώντας χειρωνακτικά (ή ημι-αυτόματα) κατασκευασμένους **κανόνες**.
 - Π.χ. BuyEvent(**buyer:C1**, **bought:C2**, **share:P**) → *
Company(id:C1)*Verb(base:αγοράζω/αποκτώ/παίρνω)*
Percent(norm:P)* Company(id:C2)*
- Ή προσπαθούμε να **μάθουμε** να προβλέπουμε τη **σχέση** που συνδέει (αν συνδέει) κάθε **ζευγάρι** κοντινών **ονομάτων οντοτήτων**. Πρόβλημα **κατηγοριοποίησης** ζευγών. Μία **κατηγορία** για κάθε **είδος σχέσης**.

Εξαγωγή σχέσεων με επιβλεπόμενη μάθηση

<s> Η <company id="GCE"> General Company Ελλάς
</company> <verb base="αγοράζω"> αγόρασε </verb> το <percent
norm="0.42"> 42% </percent> των μετοχών της <company
id="SCL"> Small Company Ltd </company>. </s> <s> Η
<company id="LCL"> Large Company Ltd </company> <verb
base="αγοράζω"> είχε επίσης επιχειρήσει ανεπιτυχώς να αγοράσει
</verb> μερίδιο της <company id="SCL"> Small </company>
<date norm="Y2007"> πέρσι </date> . </s>

- Π.χ. για τη σχέση **αγοραστή-αγοραζόμενου**:
 - Εξετάζουμε ζεύγη κοντινών ονομάτων εταιρειών, ανθρώπων.
 - Για κάθε ζεύγος, ένα διάνυσμα ιδιοτήτων.
 - Κατηγορίες: αρνητικό, θετικό (ή γενικότερα ο τύπος της σχέσης, π.χ. εξαγορά, συγχώνευση κλπ).

Εξαγωγή σχέσεων με επιβλεπόμενη μάθηση

- Μπορεί να υπάρχουν **ξεχωριστοί ταξινομητές ανά τύπο σχέσης ονομάτων** (**θετική** ή **αρνητική** περίπτωση) ή **ένας ταξινομητής** που να εντοπίζει **σχέσεις όλων των τύπων** (εξαγορά, συγχώνευση, ..., τίποτα).
- Μπορεί να υπάρχουν **ξεχωριστά στάδια αναγνώρισης σχέσεων** και **κατάταξης** τους σε τύπους.
- Παραδείγματα ιδιοτήτων:
 - **Κατηγορίες** των δύο **ονομάτων**(π.χ. πρόσωπο, οργανισμός).
 - Κύριες λέξεις (heads) των δύο ονομάτων (π.χ. «ΟργανισμόςΤηλεπικοινωνιών της Ελλάδος»).
 - Λέξεις(bag of words) ή ρίζες λέξεων ή *n*-grams ή μέρη του λόγου ή ... **μέσα, ανάμεσα, πριν, μετά** τα ονόματα.
 - **Απόσταση** των δύο ονομάτων (σε λέξεις, μήκος μονοπατιού συντακτικού δέντρου, ή παρεμβαλλόμενα ονόματα).
 - Μονοπάτι συντακτικού δέντρου

Παράδειγμα χρήσης κανόνα

<s> Σημαντικά νέα <date norm="2/12/2008"> χθες </date> από την <company id="GCE"> General Company Ελλάς </company>, μια από τις μεταλύτερες ελληνικές κατασκευαστικές εταιρείες. </s>

<s> Η <company id="GCE"> GC </company> ανακοίνωσε ότι <verb base="αγοράζω"> θα αγοράσει </verb> το <percent norm="0.42"> 42% </percent> των μετοχών της <company id="SCL"> Small Company Ltd </company>, μιας Βρετανικής εταιρείας που ειδικεύεται στο χώρο των σιδηρών κατασκευών. </s>

BuyEvent(buyer:C1, bought:C2, share:P) → * Company(id:C1) *
Verb(base:αγοράζω/αποκτώ/παίρνω) * Percent(norm:P) *
Company(id:C2) *

BuyEvent(buyer:GCE, bought:SCL, share:0.42)

Αναγνώριση χρονικών εκφράσεων

- Είδη χρονικών εκφράσεων:
 - **Απόλυτες χρονικές θέσεις:** π.χ. «στις 18 Ιανουαρίου 2010», «κατά τον 5^ο αιώνα π.Χ.».
 - **Σχετικές χρονικές θέσεις:** «χθες», «δύο εβδομάδες νωρίτερα», «τα επόμενα δύο έτη». Μπορεί να σχετίζονται με την **ημερομηνία δημοσίευσης ή άλλο γεγονός** του κειμένου.
 - **Διάρκειες:** «η μπαταρία κρατάει 8 ώρες», «το δάνειο αποπληρώνεται σε 10 έτη».
- Οι χρονικές εκφράσεις συνήθως περιέχουν **ενδεικτικές λέξεις**(triggers).
 - Π.χ. «ώρα», «εβδομάδα», «έτος», «Κυριακή», «Ιανουάριος».
 - Όμως οι ενδεικτικές λέξεις **εμφανίζονται και αλλού**(π.χ. ταινία «Ποτέ την Κυριακή»).

Τρόποι αναγνώρισης χρονικών εκφράσεων

- **Μέθοδοι επιβλεπόμενης μάθησης** που κατατάσσουν **λεκτικές μονάδες**, όπως και με την αναγνώριση ονομάτων.
- **Χειρωνακτικά** κατασκευασμένες **γραμματικές** για χρονικές εκφράσεις.
 - **Όχι τόσο δύσκολη** η συγγραφή τους όσο η συγγραφή γραμματικών για ολόκληρη τη γλώσσα.
- **Αυτόματα** κατασκευασμένες **γραμματικές**.
 - Π.χ. **Κανονικές** γραμματικές

TimeML (ISO 8601)

Αθήνα, <TIMEX3 id="t1" type="DATE" value="2007-07-02"
functionInDocument="CREATION_TIME"> 2 Ιουλίου
2007</TIMEX3> : Η μείωση εισιτηρίων της ΟΑ <TIMEX3 id="t2"
type="DATE" value="2007-W26" anchorTimeID="t1"> την
προηγούμενη εβδομάδα</TIMEX3> ανάγκασε ...

- TimeML: Καθορίζει **πρότυπο τρόπο επισημείωσης** (σε XML) **χρονικών εκφράσεων, κανονικοποιημένων παραστάσεών τους και γεγονότων**.
 - TimeBank: σώμα κειμένων επισημειωμένο σε TimeML.
 - Χρησιμοποιεί τις **χρονικές σχέσεις διαστημάτων του Allen**.

Αναγνώριση γεγονότων

Αθήνα, 2 Ιουλίου 2007: Η **μείωση** εισιτηρίων της ΟΑ την προηγούμενη εβδομάδα ανάγκασε και τη βασική ανταγωνίστριά της ΑΕ να **ανακοινώσει** ότι θα **μειώσει** όλα τα εισιτήριά της από...

- Αναγνωρίζουμε **λέξεις/φράσεις** που **εισάγουν γεγονότα**.
 - Συνήθως αλλά **όχι πάντα ρήματα** (π.χ. «μειώσει», «μείωση»).
 - Ιδανικά θα θέλαμε και να **διατάσσουμε τα γεγονότα στον άξονα του χρόνου** (δύσκολο, βλ. «tense and aspect theories»).
- «**Γεμίζουμε**» τους **θεματικούς ρόλους** των γεγονότων.
 - Π.χ. ποιος μείωσε, τι μείωσε;
 - Μπορούμε π.χ. να χρησιμοποιήσουμε τους θεματικούς ρόλους που προβλέπει το **FrameNet** ανά είδος γεγονότος.
 - Και **μηχανική μάθηση** για να τους «γεμίσουμε»
- Μπορεί να υπάρχουν και **σχέσεις μεταξύ γεγονότων**.
 - Π.χ. η **μείωση** ανάγκασε την ανταγωνίστριά να **ανακοινώσει**...

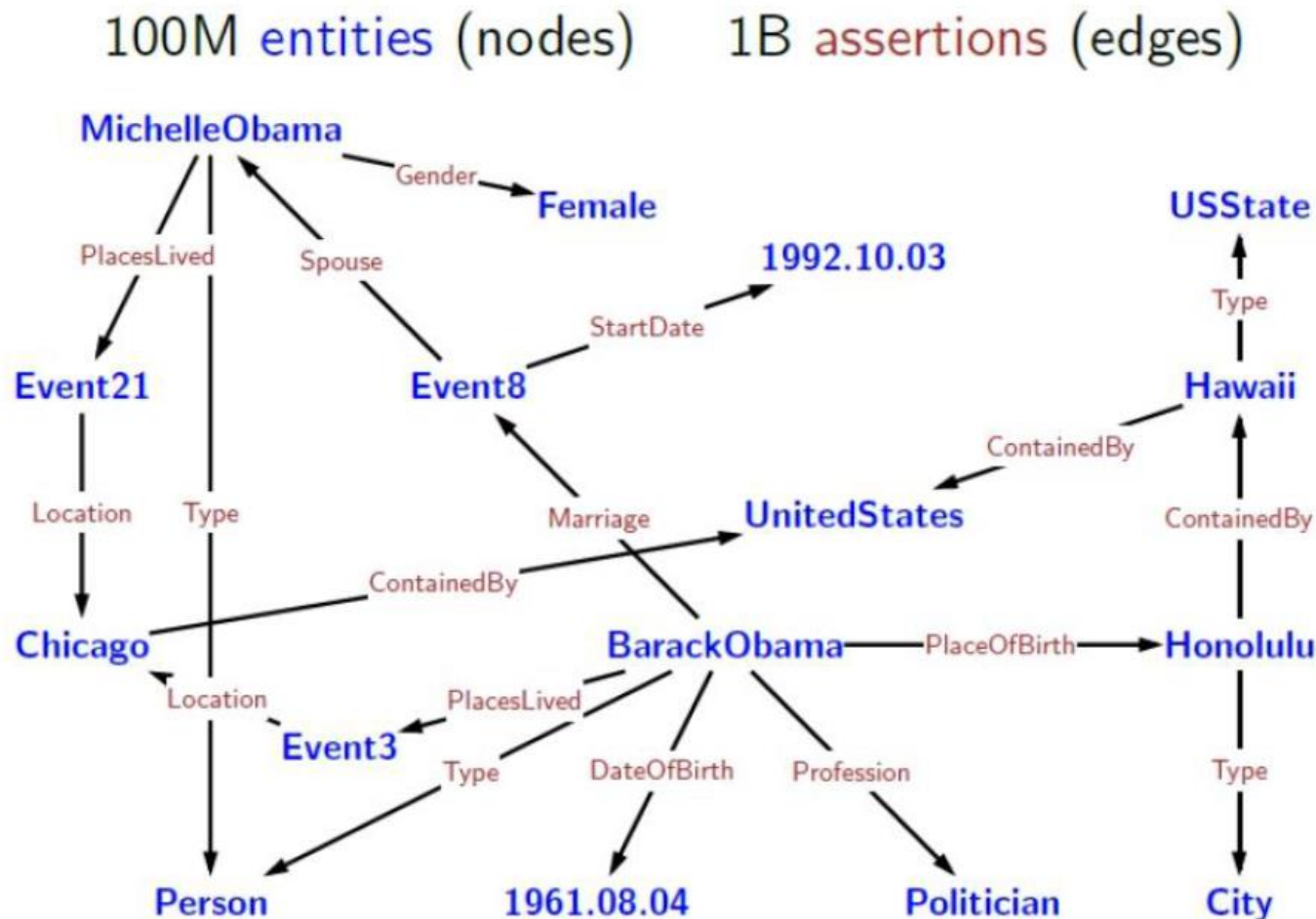
Διεπαφές φυσικής γλώσσας

- **Πλεονεκτήματα και μειονεκτήματα:**
 - Χρήση **φυσικής γλώσσας**, χωρίς εκμάθηση π.χ. SQL.
 - Αλλά **μπορεί να μην είναι προφανές** στο χρήστη **τι μπορεί να καταλάβει το σύστημα**.
 - **Ανάγκη επαναδιατύπωσης** ερωτημάτων, εκνευρισμός...
- Συχνά προτιμότερη η **ελεγχόμενη φυσική γλώσσα**.
 - Συγκεκριμένο **υποσυνόλο μιας ΦΓ**, π.χ. χωρίς δευτερεύουσες προτάσεις, μόνο ενεργητική φωνή, συγκεκριμένο λεξιλόγιο κλπ.
 - Αλλά πρέπει ο χρήστης να την **μάθει** και να την **θυμάται**.
 - Εναλλακτικά εισαγωγή λέξεων/φράσεων μέσω **καταλόγων επιλογών**, ώστε να εισάγονται μόνο εκφράσεις της ελεγχόμενης φυσικής γλώσσας και να βλέπει ο χρήστης τι επιτρέπεται.
 - Ανταγωνισμός από **γραφικές διεπαφές...**

Διεπαφές ΦΓ για ΒΔ (NL interfaces for DBs)

- **Πολλή έρευνα** από τη δεκαετία του 1960.
 - Εξαιρετικά **δημοφιλείς** τη δεκαετία του '80 ως αρχές '90.
 - Το **ενδιαφέρον** έχει πια **στραφεί** περισσότερο στα συστήματα **ερωταποκρίσεων** για **συλλογές εγγράφων** (ή ιστοσελίδων) ή/και **οντολογίες**(π.χ. σε OWL)και **Linked Data**.

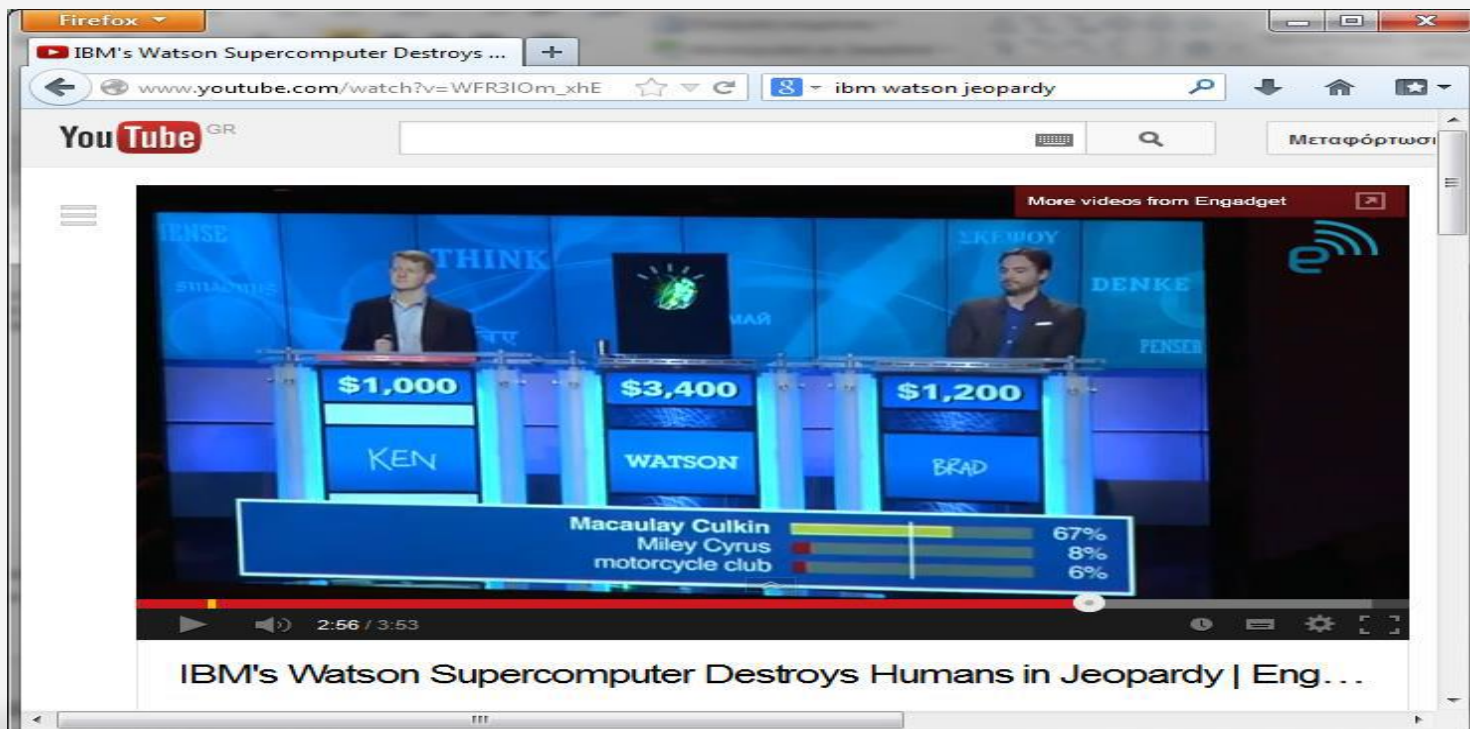
Freebase (<https://www.freebase.com/>)



Πηγή εικόνας: Z. Wang, S. Yan, H. Wang and X. Huang, "An Overview of Microsoft Deep QA System on Stanford WebQuestions Benchmark". Microsoft Research Technical Report MSR-TR-2014-121, 2014.

To IBM Watson στο *Jeopardy*!

- Το σύστημα Watson της IBM κέρδισε τον τελικό του τηλεπαιχνιδιού ερωταποκρίσεων *Jeopardy*!
- Υπάρχουν στον Ιστό πολλά βίντεο για το σύστημα.
- Βλ. π.χ. http://www.youtube.com/watch?v=WFR3IOm_xhE



Ανάκτηση Πληροφοριών (Information Retrieval)

- Ασχολείται (κυρίως) με την **ανάκτηση εγγράφων** που είναι **σχετικά με ερωτήματα χρηστών**.
- Π.χ. ανάκτηση από μια **συλλογή επιστημονικών άρθρων** ή ολόκληρο τον **Παγκόσμιο Ιστό**(μηχανές αναζήτησης).
- Τα «**έγγραφα**» μπορεί να είναι πλέον και **πολυμεσικά** ή μόνο **εικόνες, μουσική** κλπ.
- Ενδέχεται να ανακτώνται **τμήματα** (π.χ. παράγραφοι) εγγράφων.
- Συνήθως τα **ερωτήματα** και τα **έγγραφα** αντιμετωπίζονται ως **σύνολα όρων** (terms).
- Π.χ. καμία διάκριση μεταξύ «σκέφτομαι άρα υπάρχω» και «υπάρχω άρα σκέφτομαι»(**bag of words**).
- Οι **όροι** είναι συνήθως **λέξεις**, αλλά μπορεί να είναι και μεγαλύτερες **φράσεις**(π.χ. «πνευμονικό οίδημα»).

Μοντέλο διανυσματικού χώρου (vector space model)

- Πολύ διαδεδομένη είναι η **διανυσματική παράσταση ερωτημάτων και εγγράφων**.
 - Κάθε **έγγραφο** ή **ερώτημα** παριστάνεται ως ένα **διάνυσμα ιδιοτήτων**.
 - Π.χ. **δυαδικές**(Boolean) **ιδιότητες** που δείχνουν ποιοι **όροι** της συλλογής **εμφανίζονται** ή **όχι** στο έγγραφο ή ερώτημα.
 - Ή πιο συχνά **ιδιότητες** με **τιμές TF-IDF**.
 - Συνήθως **αφαιρούμε** πρώτα **πολύ κοινές λέξεις** (stop-list, π.χ. άρθρα, σύνδεσμοι) και κρατάμε τις **ρίζες** των λέξεων (stemming).
- **Ανακτώνται τα έγγραφα** των οποίων τα **διανύσματα μοιάζουν περισσότερο** με το **διάνυσμα** του **ερωτήματος**.
 - Π.χ. **ομοιότητα συνημιτόνου**.

Αυτόματη βελτίωση ερωτήματος

- Μπορούμε να **επεκτείνουμε** το **ερώτημα** (query expansion).
 - Π.χ. με **συνώνυμα** και (κοντινά) **υπερώνυμα** και **υπώνυμα**.
 - Ή γενικότερα με **σχετικούς όρους** χρησιμοποιώντας ένα **μέτρο σημασιολογικής συνάφειας** ή/και **συσταδοποίηση όρων**.
 - Συνήθως **καλύτερη ανάκληση**, αλλά **χειρότερη ακρίβεια**.
 - Θεωρητικά θα βοηθούσε και η **αποσαφήνιση λέξεων**, αλλά γενικά **αντικρουόμενα αποτελέσματα** σε IR.
- Relevance feedback: Ο χρήστης σημειώνει ***R* σχετικά** και ***S* άσχετα** έγγραφα (π.χ. εξετάζοντας τα πρώτα 10).

$$\vec{q'} = \alpha \cdot \vec{q} + \frac{\beta}{R} \sum_{j=1}^R \vec{r_j} - \frac{\gamma}{S} \sum_{k=1}^S \vec{s_k}$$

- Διαισθητικά, μετατοπίζουμε το **διάνυσμα** του **ερωτήματος** q κατά β **προς το κεντροειδές των σχετικών** και κατά γ **μακριά από το κεντροειδές των άσχετων**.
- Τα α , β , γ καθορίζουν πόσο θέλουμε να επηρεαστεί το ερώτημα από τα σημειωμένα σχετικά και άσχετα έγγραφα.



Αναζήτηση στο Google

☒ Αναζήτηση - Παγκόσμιος Ιστός

☐ Αναζητήστε στις σελίδες της κατηγορίας Ελληνικά

Οι παρακάτω λέξεις είναι πολύ συνηθισμένες και δεν συμπεριλήφθηκαν στην αναζήτησή σας:

 When was the. [[Λεπτομέρειες](#)]

[Παγκόσμιος Ιστός](#)
[Εικόνες](#)
[Ομάδες Συζητήσεων](#)
[Κατάλογος](#)

 Ιστοσελίδες που αναζητήθηκαν με τους όρους **When was the telephone invented?** : 1 έως 10 από περίπου 217,000 αποτελέσματα

Συμβουλή: Στα περισσότερα προγράμματα πλοήγησης μπορείτε απλώς να πατήσετε το πλήκτρο return αντί να επιλέξετε το κουμπί αν

1876 - Telephone Invented

... 1876 - **Telephone Invented**. ... The **telephone**, **invented** in America by Alexander Graham Bell, had more impact on personal and business communication. ...

www.impty.com/parlidex/1876.html - 3k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

World's first practical wireless telephone invented

Dr. Torigata Uichi **invented** the world's first practical wireless **telephone** (the TYK wireless **telephone**) in February 1912 (Meiji 45) at the age of 29. ...

korea.park.org/Japan/NTT/MUSEUM/html_ht/HT912020_e.html - 2k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

Google Search: when was the telephone invented?

... Category: Recreation > Antiques > Telephones and Telegraphs. World's first practical wireless **telephone invented** 1876 - **Telephone Invented**.

ilk.kub.nl/~antalb/abvi/week3/google_results.html - 15k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

Empirical Methods for Natural Language Processing

Είδος αρχείου: Microsoft Powerpoint 97 - [Σε μορφή HTML](#)

... Alexander Graham Bell. When was the **telephone invented**? 1876. (Buchholz & Daelemans, 2001). ... When was the **telephone invented**? Which slots are given? Verb **invented**. ...

www.hit.uib.no/spraaktek02/daelemans2.ppt - [Παρόμοιες σελίδες](#)

Which was invented first -- the telephone or the light bulb?

... Add Ask Yahoo! to My Yahoo! Dear Yahoo! Which was **invented** first -- the **telephone** or the light bulb? Enlighten Me. Dear Enlighten: This is a hot button issue. ...

ask.yahoo.com/ask/20001020.html - 7k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)



[Σύνδεση Αναζήτηση](#) [Προτιμήσεις](#) [Γλωσσικά Εργαλεία](#) [Συμβουλές Αναζήτησης](#)

When did K. Simitis become prime m Αναζήτηση στο Google

☒ Αναζήτηση Παγκόσμιος Ιστός ☐ Αναζητήστε στις σελίδες της κατηγορίας Ελληνικά
 Η λέξη "When" είναι πολύ συνηθισμένη και δεν λήφθηκε υπόψιν στην αναζήτηση σας.
[Λεπτομέρειες](#)

[Παγκόσμιος Ιστός](#) [Εικόνες](#) [Ομάδες Συζητήσεων](#) [Κατάλογος](#)

Ιστοσελίδες που αναζητήθηκαν με τους όρους **When did K. Simitis become prime minister?** : 1 έως 10 από περίπου 490 αποτελέ

Συμβουλή: Στα περισσότερα προγράμματα πλοήγησης μπορείτε απλώς να πατήσετε το πλήκτρο return αντί να επιλέξετε το κουμπί αν

The Greek Prime Minister

... The election of **Simitis** as **Prime Minister** did not discourage A. Tsochatzopoulos who was ... between the electorates of the election of K. **Simitis** as **Prime** ...

www.asteras.aek.net.gr/Communism/Simitis/ - 18k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

1998/02/04 17:52 Two Years of Traditionalist Modernism in ...

... took two full years in power for K. **Simitis**, heralded ... **Minister** s post two years ago, nevertheless, **did** not so ... be able to apply his philosophy, **Simitis** and his ...

www.aimpress.org/dyn/trae/archive/data/199802/80204-019-trae-ath.htm - 17k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

articles

... it took two full years in power for K. **Simitis**, heralded as ... s post two years ago, nevertheless, **did** not so ... To be able to apply his philosophy, **Simitis** and his ...

www.greekhelsinki.gr/english/articles/AIM28-1-98.html - 21k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

Greek-Turkish relations

... the **Prime Minister** of Greece Mr. K. **Simitis**, at ... the beginning of November 1997 **Prime Minister Simitis** and **Prime** ... Although the meeting **did** not produce tangible ...

www.hri.org/MFA/foreign/bilateral/relations.htm - 20k - [Αποθηκευμένη Σελίδα](#) - [Παρόμοιες σελίδες](#)

Σύνδεσμοι Χορηγών

Free Instant Ordination

Be a legally ordained clergy person
Perform weddings, baptisms, etc.

www.SpiritualHumanism.org

Ενδιαφέρον:

[Δείτε το μήνυμά σας εδώ...](#)



Συστήματα ερωταποκρίσεων ΦΓ για τον Ιστό

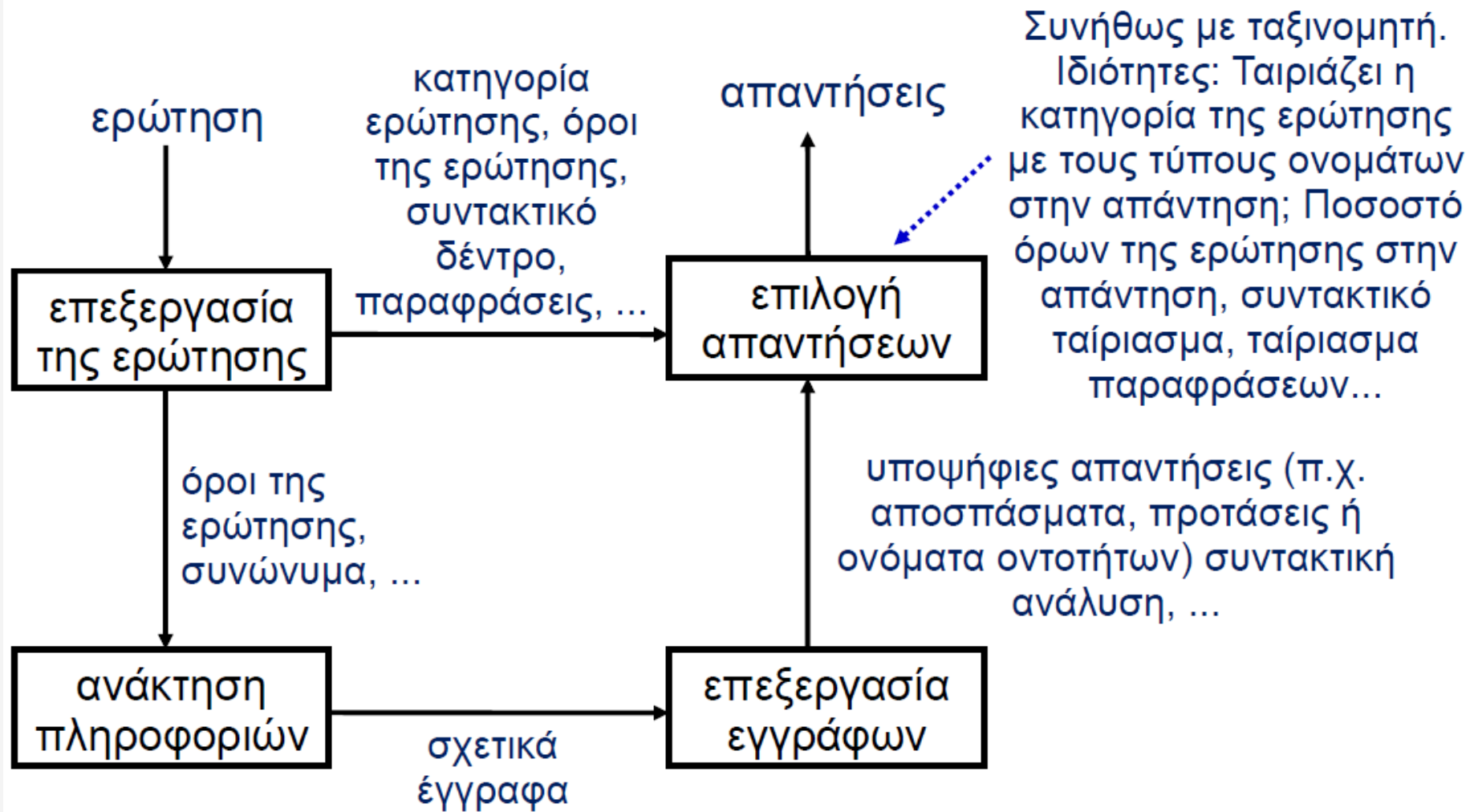
→ *When was the telephone invented?*

- **1876 (25)**  ακριβής απάντηση και βαθμός βεβαιότητας
 - www.clicktron.com/trivia/...: *The telephone was invented* by Alexander Graham Bell **in 1876**.
 - www.komando.com/other/kids/...: They *invented the telephone* **on March 10, 1876**.
 - ...
- **1912 (1)**
 - www...: Dr. Torigata Uichi ***invented the*** word's first practical wireless ***telephone...*** **in 1912**.
 - ...

 Αιτιολόγηση απάντησης

[Βασισμένο σε παράδειγμα των Buchholz & Daelemans (2001).]

Συστήματα ερωταποκρίσεων για συλλογές εγγράφων (ή τον Ιστό)



Επεξεργασία της ερώτησης

- **Κατηγορία της ερώτησης:**
 - Π.χ. απαιτεί συγκεκριμένο όνομα **προσώπου** ή **τοποθεσίας** ή **χρονολογία** ως απάντηση; (factoid questions)
 - Μήπως ζητά **ορισμό**; (Π.χ. «Τι είναι η θαλασσαιμία;») **Πιο δύσκολες ερωτήσεις** και ενδέχεται να χρειάζονται **περίληψη**.
 - Μήπως ζητά **τρόπο εκτέλεσης** μιας διαδικασίας; (Π.χ. «Πώς αλλάζω τη μπαταρία του...;») Επίσης **δύσκολες**.
 - Μπορεί να αντιμετωπιστεί ως **πρόβλημα κατάταξης**. Ιδιότητες: π.χ. **ερωτηματικές λέξεις**, **η-γράμματα**, **κεντρικό ουσιαστικό** («Ποια πόλη....;»), **υπερώνυμα κεντρικού ουσιαστικού**.
- Όροι της ερώτησης, ενδεχομένως και επέκταση, **παράφραση** ή **μετατροπή** της ερώτησης **σε δήλωση**.
 - Π.χ. «Πότε εφευρέθηκε το τηλέφωνο;» → «Το τηλέφωνο εφευρέθηκε [χρονική έκφραση]».
 - Π.χ. «Πώς θεραπεύεται το...» → «Ποια είναι η θεραπεία του...»
- Ενδεχομένως **συντακτική** ή και **σημασιολογική** ανάλυση.

Συστήματα που επιστρέφουν αποσπάσματα

- Τα **αποσπάσματα** συνήθως είναι **προτάσεις** ή **παράγραφοι**.
 - Π.χ. που περιέχουν **όρους** της ερώτησης.
- **Επιλογή** μεταξύ υποψηφίων αποσπασμάτων με **επιβλεπόμενη μάθηση**. Π.χ. με **ιδιότητες**:
 - **Ταιριάζει** η **κατηγορία** της ερώτησης με τις **κατηγορίες ονομάτων οντοτήτων** μέσα στο υποψήφιο **απόσπασμα**;
 - **Ποσοστό όρων** της ερώτησης στο **υποψήφιο απόσπασμα** και **πόσο απέχουν** κατά μέσο όρο μεταξύ τους στο **απόσπασμα**;
- **Ομοιότητα** ερώτησης και αποσπάσματος, στο **λεξιλογικό** (π.χ. cosine similarity), **συντακτικό** ή **σημασιολογικό** επίπεδο.
 - Πόσο ψηλά (rank) επέστρεψε η μηχανή IR το **έγγραφο** από το οποίο προέρχεται το **απόσπασμα**;
 - **Απόσταση** του (διανύσματος του) **συγκεκριμένου υποψηφίου αποσπάσματος** από το **κεντροειδές** όλων των υποψηφίων **αποσπασμάτων** της ερώτησης.

Συστήματα που επιστρέφουν οντότητες (1/2)

- Ενδέχεται να βρίσκουμε πρώτα το **καλύτερο απόσπασμα** και κατόπιν να **ψάχνουμε μέσα του** την καλύτερη οντότητα.
 - Π.χ. ερώτηση προσώπου, το καλύτερο απόσπασμα είναι μια πρόταση με **τρία ονόματα προσώπων**, ποιο διαλέγουμε;
- **Χρήση κανόνων** βασισμένων σε **πρότυπα**:
 - Π.χ. αν ερώτηση προσώπου και το απόσπασμα ταιριάζει με το «[πρόσωπο:X] έγραψε [βιβλίο:Y]» και το [βιβλίο:Y] είναι όρος της ερώτησης, απάντησε με το [πρόσωπο:X].
 - Π.χ. αν ερώτηση ορισμού και το απόσπασμα ταιριάζει με το «X, ένα είδος Y» ή με το «X και άλλα Y» και το X είναι όρος της ερώτησης, απάντησε «Είναι ένα Y».
 - Τα **πρότυπα** μπορεί να έχουν δημιουργηθεί όπως στην **εξαγωγή πληροφοριών**, για σχέσεις που αναμένονται στις ερωτήσεις των χρηστών (π.χ. σχέσεις μεταξύ γονιδίων, ασθενειών, φαρμάκων).

Συστήματα που επιστρέφουν οντότητες (2/2)

- **Αντί για πρότυπα** (ή μαζί με αυτά), μπορούμε να χρησιμοποιήσουμε πάλι **επιβλεπόμενη μάθηση**.
 - Ένα **διάνυσμα** για κάθε **υποψήφια απάντηση** (π.χ. για κάθε όνομα οντότητας μέσα στο απόσπασμα ή για κάθε φράση του τμήματος που ταίριαξε με μεταβλητή ενός προτύπου).
 - **Ιδιότητες παρόμοιες** με εκείνες της **αξιολόγησης αποσπασμάτων**.
- **Αντί να εντοπίζουμε πρώτα το καλύτερο απόσπασμα και μετά την καλύτερη απάντηση** (οντότητα ή φράση) μέσα στο απόσπασμα, ενδέχεται να αξιολογούμε **κατευθείαν υποψήφιας απαντήσεις** (οντότητες, φράσεις).

Αξιολόγηση συστημάτων ερωταποκρίσεων

- Αν το σύστημα επιστρέφει **μία απάντηση ανά ερώτηση**, μπορούμε να μετρήσουμε το **ποσοστό ορθότητας (accuracy)**.
 - Πόσες ερωτήσεις **απάντησε σωστά**.
- Συχνά επιτρέπεται το σύστημα να επιστρέψει **N απαντήσεις ανά ερώτηση**, ταξινομημένες **κατά βαθμό βεβαιότητας**.
 - Επιστρέφει **πιο ψηλά** τις απαντήσεις που θεωρεί **καλύτερες**.
 - Έστω Q οι ερωτήσεις αξιολόγησης και Q' οι ερωτήσεις για τις οποίες το σύστημα επέστρεψε τουλάχιστον μία σωστή απάντηση.
 - Για κάθε ερώτηση $q \in Q'$, $rank q$ είναι η **θέση**(1, 2, ..., N) της **πρώτης** (από πάνω προς τα κάτω) **σωστής απάντησης**.
 - Mean Reciprocal Rank:
$$MRR = \frac{1}{|Q'|} \sum_{q \in Q'} \frac{1}{rank(q)}$$
 - Μεγαλύτερο MRR αν δίνει **σωστές απαντήσεις** και **πιο ψηλά**.

Αυτόματη παραγωγή περιλήψεων

- **Χρήσιμη σε πολλές περιπτώσεις. Π.χ:**
 - Μικρά αποσπάσματα (snippets) **ανακτηθέντων εγγράφων ή ομάδων εγγράφων** στις μηχανές αναζήτησης.
 - Περίληψη της **γνώμης του κοινού** (π.χ. από ιστολόγια, fora) για ένα προϊόν, πρόσωπο, γεγονός κλπ.
 - Περιλήψεις **πρακτικών συναντήσεων ή διαβουλεύσεων.**
 - **Εκτενέστερες απαντήσεις** σε ερωτήσεις π.χ. ορισμού (π.χ. «Ποιος ήταν ο Δαρβίνος;», «Τι είναι η θαλασσαιμία;»).
- **Είδη περιλήψεων:**
 - Από ένα ή πολλά έγγραφα (single vs. multi-document).
 - Περίληψη εστιασμένη σε ερώτημα (query-focused) **ή όχι.**
 - Μόνο **επιλογή αποσπασμάτων** (π.χ. προτάσεων) ή και **επαναδιατύπωση** (extractive vs. abstractive).

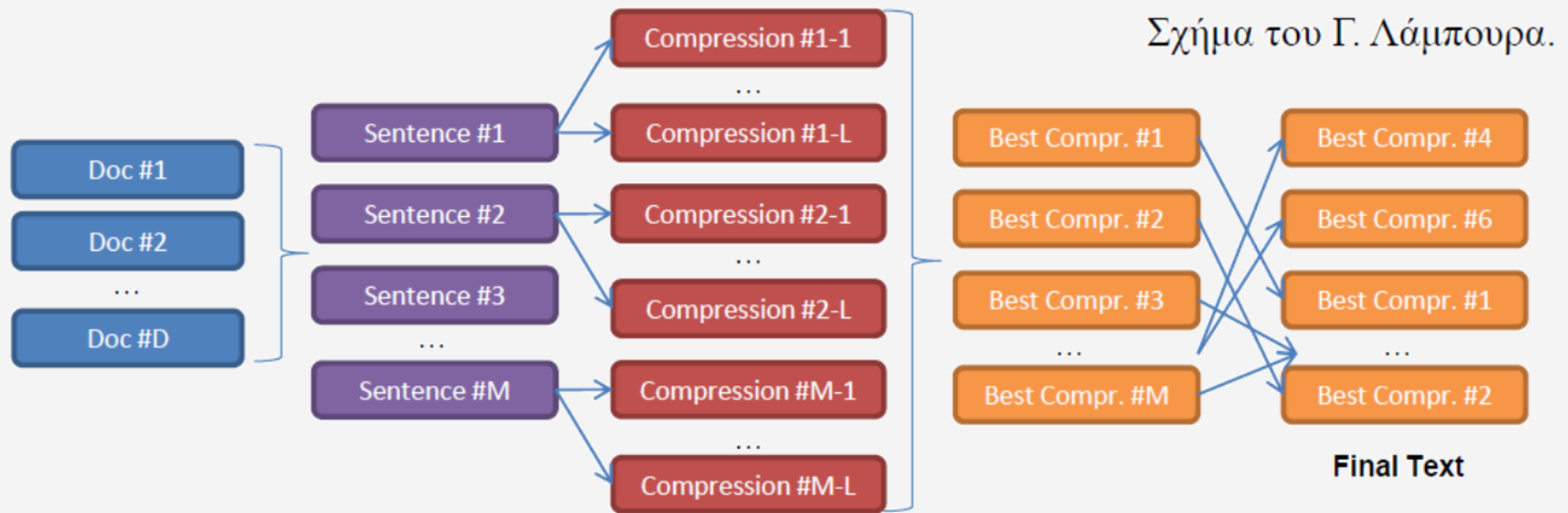
Αυτόματη παραγωγή περιλήψεων –συνέχεια

- Κυριότερα στάδια:
 - **Επιλογή περιεχομένου**, συνήθως προτάσεων, ενδεχομένως και **συμπίεση προτάσεων** (sentence compression).
 - **Σχεδιασμός κειμένου**: συνήθως **διάταξη προτάσεων** (ordering). Ενδεχομένως **αποφάσεις ομαδοποίησης** προτάσεων (aggregation).
 - Επιφανειακή πραγμάτωση (surface realization), π.χ. **πραγμάτωση ομαδοποίησης, επισκευή αναφορικών εκφράσεων**.
- Όταν παράγεται περίληψη **από ένα έγγραφο**, πολλά συστήματα απλά **επιλέγουν προτάσεις**, (ίσως) τις **συντομεύουν** και τις εμφανίζουν με τη **σειρά που είχαν** στο αρχικό έγγραφο.

Επιλογή προτάσεων από ένα έγγραφο

- Συνήθως μέσω **επιβλεπόμενης μηχανικής μάθησης**.
 - Π.χ. **κατάταξη** σε **δύο κατηγορίες** (σημαντική πρόταση ή όχι).
 - Δεδομένα εκπαίδευσης: **κείμενα και προτάσεις που επέλεξαν άνθρωποι** (extractive) ή **προτάσεις περιλήψεων που έγραψαν άνθρωποι, ευθυγραμμισμένες με τις αρχικές** (abstractive).
- **Παραδείγματα** χρήσιμων **ιδιοτήτων**:
 - Μέσο **TF-IDF** των λέξεων της πρότασης (ίσως με stop-list, stemmer).
 - **Απόσταση από το κεντροειδές** όλων των προτάσεων.
 - **Θέση πρότασης** (π.χ. σε τίτλο, αρχή παραγράφου, πρώτη παράγραφο, κορυφαίο 10% του κειμένου) και **μήκος** πρότασης.
 - **Χαρακτηριστικές λέξεις ή εκφράσεις** (π.χ. «Σε αυτό το άρθρο...», «Συνοψίζοντας, ...») και εκφράσεις που σηματοδοτούν **συναίσθημα** (sentiment).

Παραγωγή περίληψης πολλών εγγράφων



- Τώρα πρέπει να **επιλέξουμε προτάσεις** που είναι **σημαντικές**, αλλά και που **δεν λένε τα ίδια** πράγματα.
 - Μπορούμε να χρησιμοποιήσουμε **ταξινομητή** με **ιδιότητες παρόμοιες** της περίπτωσης **περίληψης ενός εγγράφου**.
 - Μπορούμε να **επιλέξουμε επαναληπτικά** την **καλύτερη πρόταση** που απομένει, ως **μέγιστο αριθμό** προτάσεων.
 - **Εξαιρώντας προτάσεις που μοιάζουν πολύ** (π.χ. ομοιότητα συνημιτόνου) με προτάσεις που έχουμε **ήδη επιλέξει**.
 - Ή ομαδοποιούμε (clustering) τις **προτάσεις** και επιλέγουμε **μία** (την καλύτερη) **από κάθε ομάδα** προτάσεων.

Περιλήψεις εστιασμένες σε ερωτήματα

- Το ερώτημα επηρεάζει την επιλογή των εγγράφων, αν χρησιμοποιείται σύστημα **ανάκτησης πληροφοριών**.
- Μπορούμε επίσης να προσθέσουμε στην επιλογή **περιεχομένου** ιδιότητες που εξετάζουν και το ερώτημα.
 - Όπως στην **αξιολόγηση αποσπασμάτων** στα συστήματα QA.
 - Π.χ. **πόσο μοιάζει** η κάθε πρόταση με το ερώτημα.
- Μπορούμε επίσης να κατατάσσουμε τα ερωτήματα σε **κατηγορίες** όπως στα συστήματα QA και να έχουμε **σχεδιάτυπα (templates) απαντήσεων**.
 - Π.χ. αν πρόκειται για **ερώτημα ορισμού προσώπου** (π.χ. «Ποιος ήταν ο Δαρβίνος;»), εντόπισε πληροφορίες για **γέννηση, θάνατο, έργα, επιτεύγματα** κλπ.
 - Και **προκαθορισμένη διάταξη** των πληροφοριών αυτών.
- Ή μπορούμε να **μάθουμε διαφορετικό μοντέλο επιλογής και διάταξης προτάσεων** ανά κατηγορία ερώτησης.

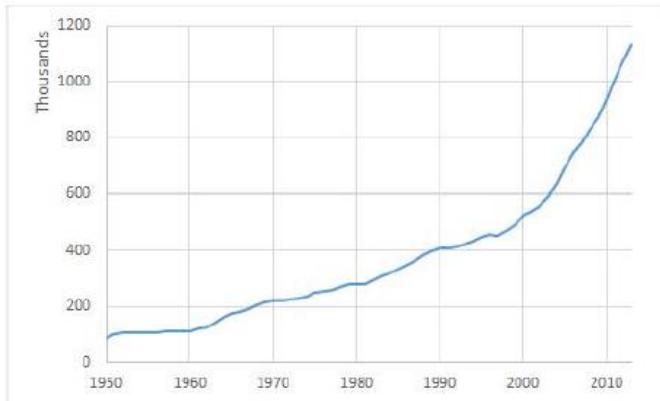
Αξιολόγηση συστημάτων παραγωγής περιλήψεων

- Όπως και στη μηχανική μετάφραση, η πιο αξιόπιστη αξιολόγηση γίνεται με **ανθρώπους-κριτές**.
 - Π.χ. τους ζητάμε να αξιολογήσουν κατά πόσον η περίληψη **διατηρεί τις σημαντικότερες πληροφορίες, την ποιότητα του κειμένου**, το πόσο εύκολα **κατανοητό** είναι κλπ.
 - **Δύσκολο να επαναλαμβάνουμε** αξιολόγηση με ανθρώπους.
- Μέτρα αυτόματης αξιολόγησης.
 - Εξετάζουν **πόσο μοιάζουν οι περιλήψεις του συστήματος με περιλήψεις ανθρώπων**.
 - Κατασκευάζουμε **πολλαπλές περιλήψεις** (π.χ. με διαφορετικούς ανθρώπους) **κάθε κειμένου** αξιολόγησης.
 - Μετράμε π.χ. **πόσα από τα n -γράμματα των περιλήψεων των ανθρώπων υπάρχουν στην αυτόματη περίληψη**.
 - Βλ. J&M για περιγραφή του μέτρου **ROUGE**(recall-based).

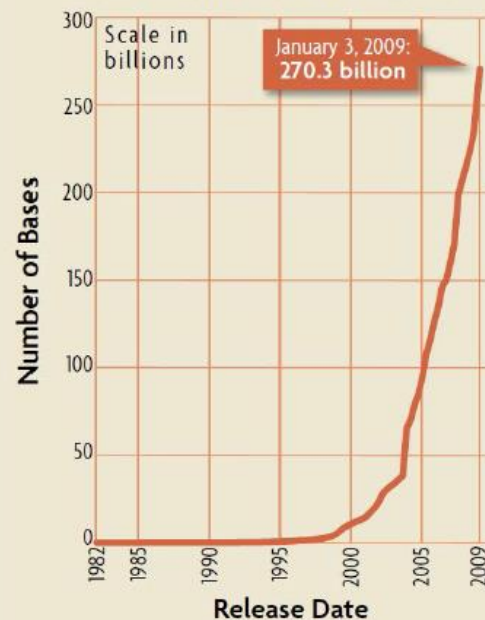
The experience

Data growth is exponential

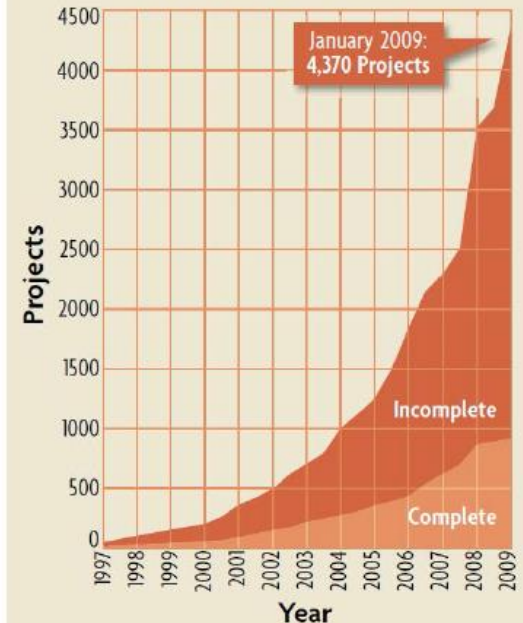
Number of articles indexed by MEDLINE (PUBMED) per year



Growth Rate of EMBL-Bank



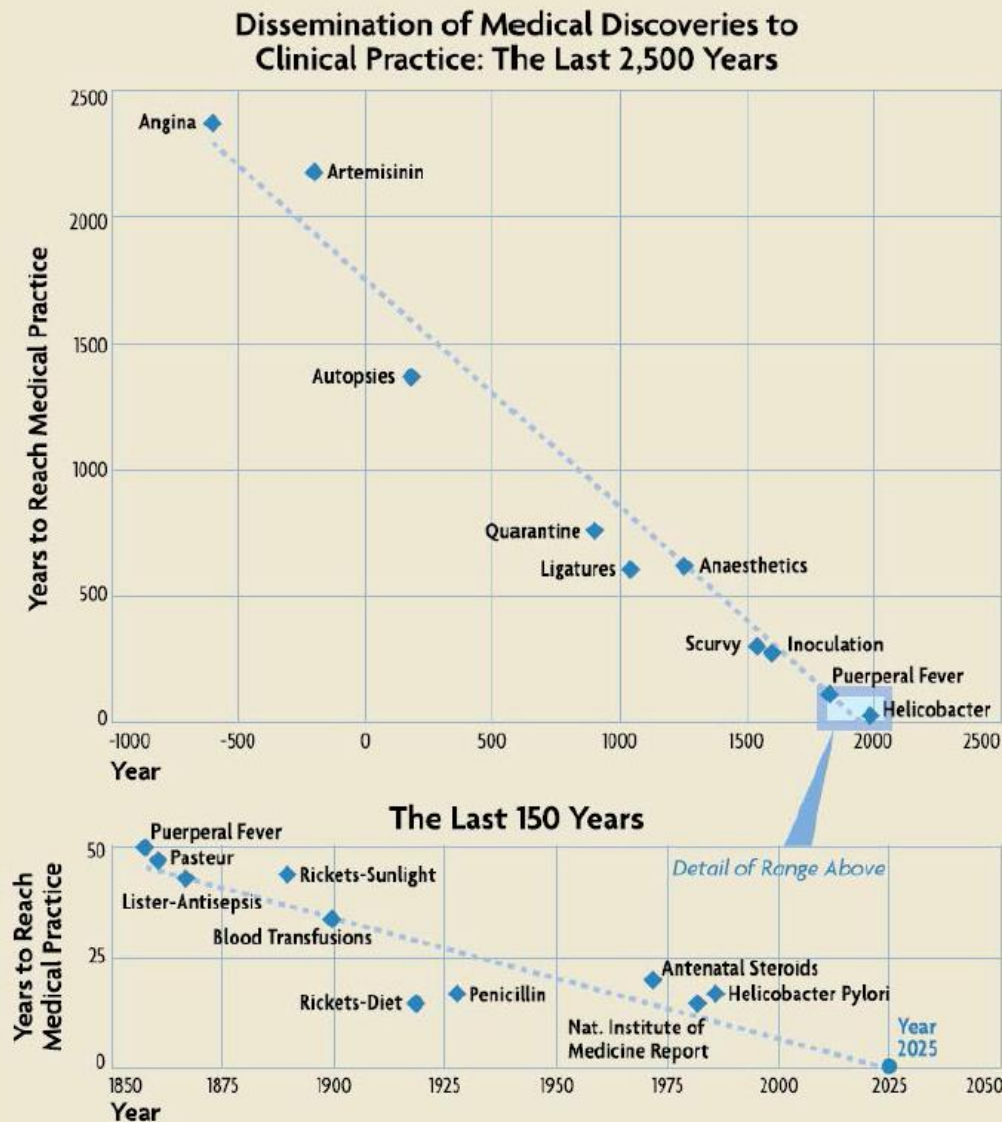
Genome Sequencing Projects on GOLD



<http://dan.corlan.net/medline-trend.html>

Southan and Cameron, *Beyond the tsunami: developing the infrastructure to deal with life sciences data*, The fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corp., 2009.

Data is source of knowledge



Gillam et al., *The healthcare singularity and the age of semantic medicine*, The fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corp., 2009.

One just needs to make the connection

- **The Swanson case: Fish oil and Raynaud's syndrome**
 - Public knowledge since 1975: Raynaud's syndrome is associated with high blood viscosity, platelet aggregability, vasoconstriction.
 - Public knowledge since 1984: Fish oil leads to reductions in blood lipids, platelet aggregability, blood viscosity, and vascular reactivity.
 - Swanson puts the two together in 1986: Can dietary fish oil ameliorate or prevent Raynaud's syndrome? He supports his evidence with relevant literature.
 - DiGiacomo confirms the hypothesis in 1989.
- **Vision:** Create big data machinery that help produce and support more such cases.

BioASQ vision

- 2 articles published in biomedical journals **every minute!**
- Make sure this knowledge is used to the benefit of patients
- Need to make it accessible to biomedical experts
- Search is not effective enough
- Push research in automated answering of questions
- A challenge for such systems can achieve a multiplying effect

What is BioASQ?

A **challenge** funded by the **European Union** (FP7).

Task a: Hierarchical text classification

- Organizers distribute **new unclassified PubMed articles**.
- Participants assign **MeSH terms** to the articles.
- **Evaluation** based on annotations of **PubMed curators**.

Task b: IR, QA, summarization, ...

- Organizers distribute **English biomedical questions**.
- Participants provide: relevant **articles, snippets, concepts, triples, “exact” answers, “ideal” answers**.
- **Evaluation**: both **automatic** (GMAP, MRR, ROUGE etc.) and **manual** (by biomedical experts).

Questions of biomedical experts

Yes/No question

“Are there any DNMT3 proteins present in plants?”

Exact Answer

“Yes”

Ideal Answer

“Yes. The plant DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) is a homolog of the mammalian de novo methyltransferase DNMT3. DRM2 contains a novel arrangement of the motifs required for DNA methyltransferase catalytic activity.”

Questions of biomedical experts

Factoid question

“What is the methyl donor of DNA (cytosine-5)-methyltransferases?”

Exact Answer

“S-adenosyl-L-methionine”

Ideal Answer

“S-adenosyl-L-methionine (AdoMet, SAM) is the methyl donor of DNA (cytosine-5)-methyltransferases. DNA (cytosine-5)-methyltransferases catalyze the transfer of a methyl group from S-adenosyl-L-methionine to the C-5 position of cytosine residues in DNA.”

Questions of biomedical experts (III)

List question

“Which species may be used for the biotechnological production of itaconic acid?”

Exact Answer

“Aspergillus terreus”, “Aspergillus niger”, “Ustilago maydis”

Ideal Answer

“In 1955, the production of itaconic acid was firstly described for Ustilago maydis. Some Aspergillus species, like A. itaconicus and A. terreus, show the ability to synthesize this organic acid and A. terreus can secrete significant amounts to the media. Itaconic acid is mainly supplied by biotechnological processes with the fungus Aspergillus terreus. Cloning of the cadA gene into the citric acid producing fungus A. niger showed that it is possible to produce itaconic acid also in a different host organism.”

Questions of biomedical experts (III)

Summary question

“How do histone methyltransferases cause histone modification?”

Exact Answer

-

Ideal Answer

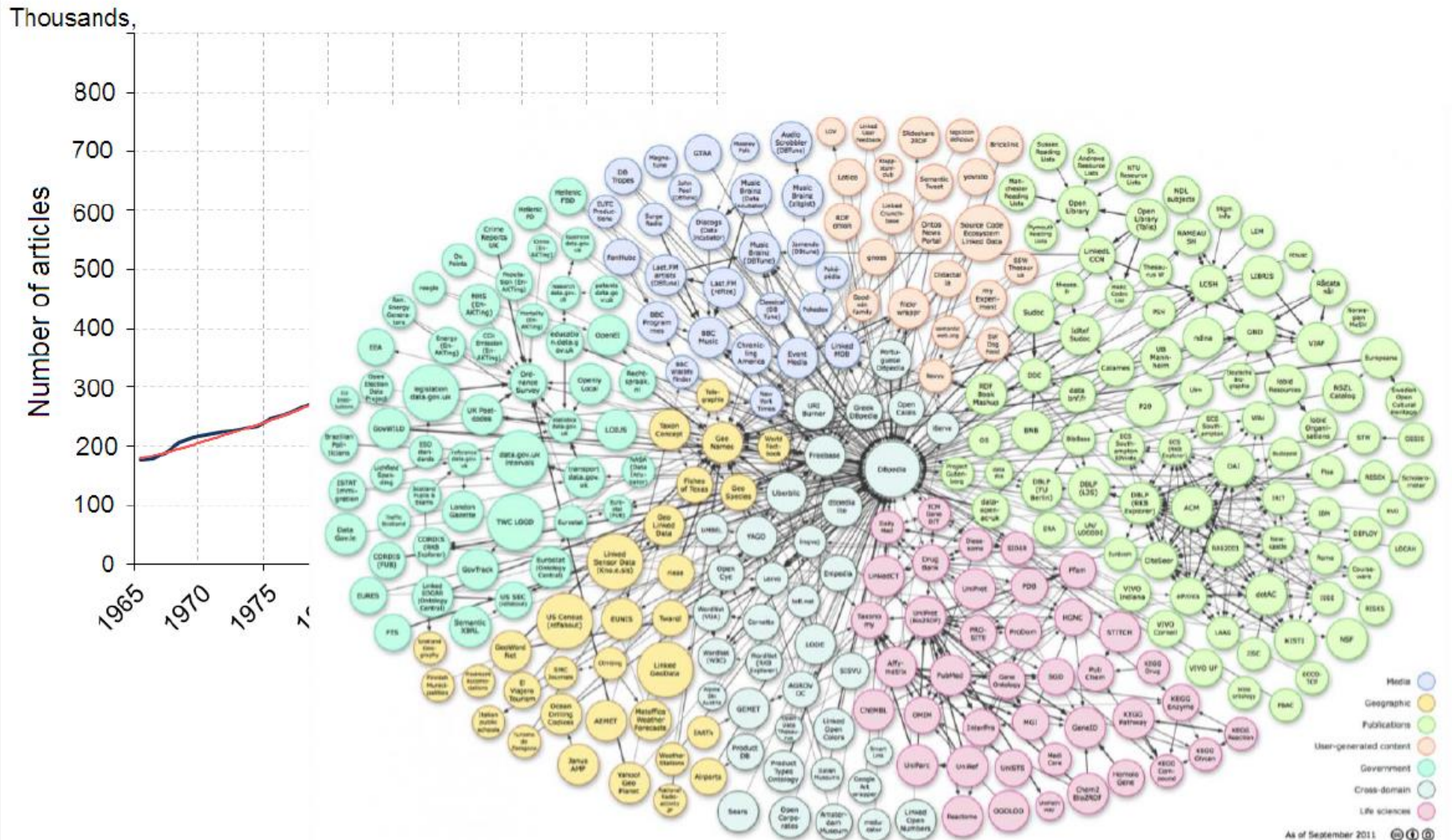
“Histone methyltransferases (HMTs) are responsible for the site-specific addition of covalent modifications on the histone tails, which serve as markers for the recruitment of chromatin organization complexes. There are two major types of HMTs: histone-lysine N-Methyltransferases and histone-arginine N-methyltransferases. The former methylate specific lysine (K) residues such as 4, 9, 27, 36, and 79 on histone H3 and residue 20 on histone H4. The latter methylate arginine (R) residues such as 2, 8, 17, and 26 on histone H3 and residue 3 on histone H4. Depending on what residue is modified and the degree of methylation (mono-, di- and tri-methylation), lysine methylation of histones is linked to either transcriptionally active or silent chromatin.”

Selecting relevant snippets

Release of sunflower seed dormancy by cyanide: cross-talk with ethylene signalling pathway

Introduction Cyanide is a compound known to stimulate germination and to release dormancy of seeds of many species (Taylorson and Hendricks, 1973; Roberts and Smith, 1977; Bogatek and Lewak, 1988; Côme et al., 1988; Bethke et al., 2006). Seed dormancy is defined as the property of a seed that prevents its germination in apparently favourable conditions (Finch-Savage and Leubner-Metzger, 2006). Despite the well-described effects of cyanide on germination and dormancy, the cellular bases of its mechanism are poorly understood, and seem moreover to vary from one species to another. Different hypotheses have been proposed to explain the stimulatory effect of cyanide on germination and dormancy (Côme and Corbineau, 1989). x According to Taylorson and Hendricks (1973), the cyanhydric gas could react with L-cysteine to give the β -L-cyanoalanine necessary for the synthesis of arginine and aspartic acids which could be limiting factors for germination. Hagesawa et al. (1994) suggested that this increase in the amino acid pool might also promote germination by decreasing the water potential in embryonic axis. However, other respiratory inhibitors which are not metabolized, such as NaN_3 or Na_2S , have the same effect as KCN in various species (Roberts and Smith, 1977; Côme and Corbineau, 1989). Some studies proposed that the beneficial effect of cyanide on germination might involve the cyanide-insensitive pathway (Esashi et al., 1979, 1981b; Upadhyaya et al., 1983), the pentose phosphate pathway (Roberts and Smith, 1977; Côme and Corbineau, 1989), the glycolysis (Bogatek, 1995) or the hydrolysis of oligosaccharides and their catabolism (Bogatek and Lewak, 1991; Bogatek et al., 1999). Cyanide is also known to interact with reactive oxygen species (ROS) metabolism; it is an inhibitor of Cu/Zn superoxide dismutase (SOD) (Bowler et al., 1992) and catalase (CAT) (Tejera García et al., 2007) and it has been demonstrated to induce oxidative stress and lipid peroxidation in animals (Johnson et al., 1987; Gunasekar et al., 1998). Oracz et al. (2007) recently demonstrated that cyanide could trigger protein oxidation during sunflower seed dormancy alleviation. At last, cyanide might also interplay with the ethylene signalling pathway. Indeed, hydrogen cyanide is a co-product of ACC oxidase, which converts ACC to ethylene (Peiser et al., 1984), and it has been proposed to stimulate ethylene biosynthesis via a feedback effect (Pirrung and Brauman, 1987). Thus Smith and Arteca (2000) demonstrated that the ACC synthase gene ACS6 was activated by cyanide in Arabidopsis. However, it is actually not known whether ethylene and cya x nide share some molecular components of their downstream transduction pathways. The putative relationship between cyanide and ethylene signalling pathways might be particularly relevant for sunflower seeds, whose dormancy is broken by ethylene (Corbineau et al., 1990). The inability of freshly harvested sunflower seeds to germinate at temperatures below c. 15 °C results from an embryo dormancy which is gradually eliminated during dry storage (Corbineau et al., 1990; Corbineau and Côme, 2003). Embryo dormancy is characterized by the inhibition of radicle extension thus preventing excised embryos to grow (Finch-Savage and Leubner-Metzger, 2006). Little attention has been paid to the

Not only texts: ontologies, linked data, ...



Information from structured data

List question

"Which forms of cancer is the Tpl2 gene associated with?"

Related RDF triple

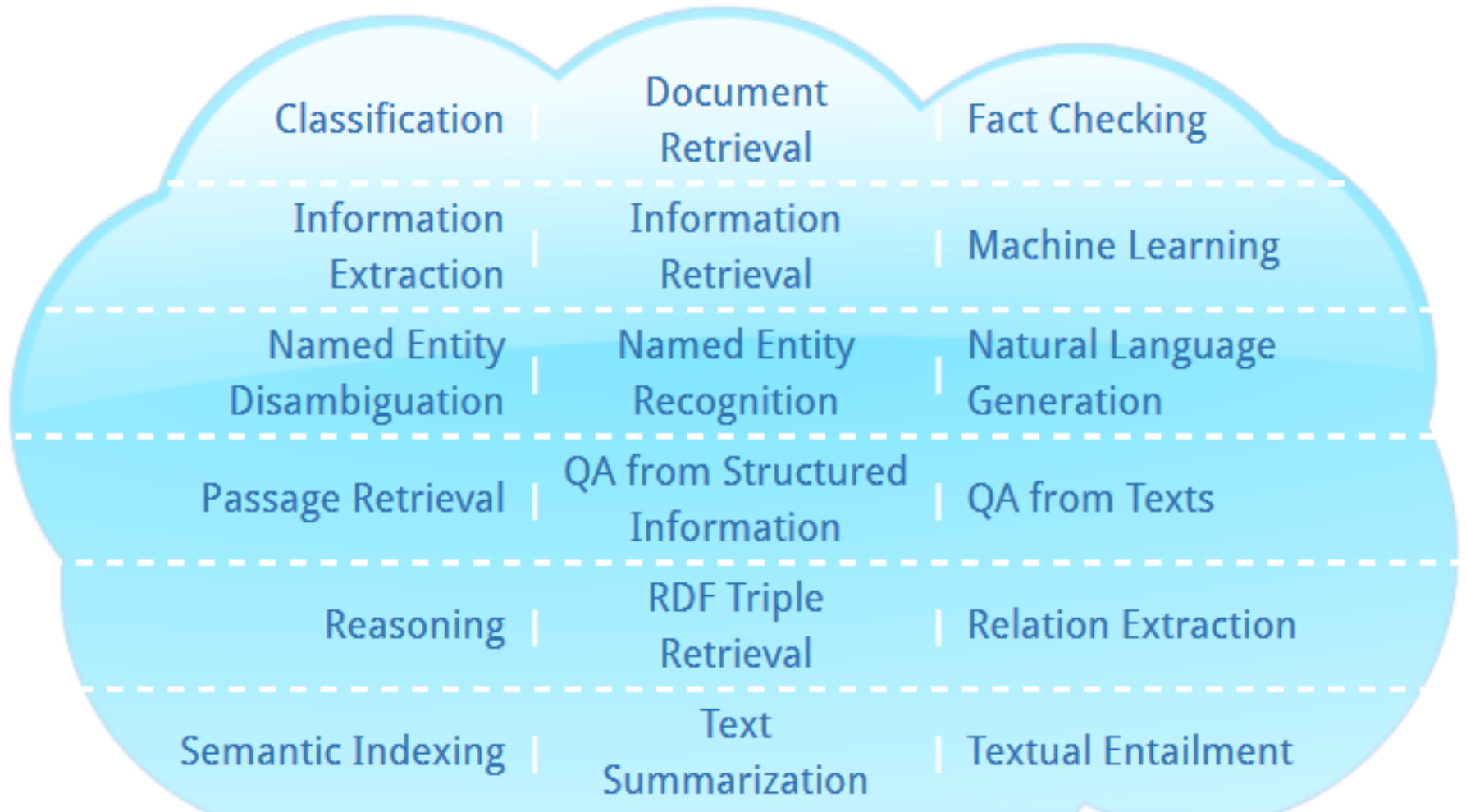
Subject: <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseases/3003> (lung cancer)
Predicate: <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/associatedGene>
Object: <http://www4.wiwiss.fu-berlin.de/diseasome/resource/genes/TPL2>

Related concepts

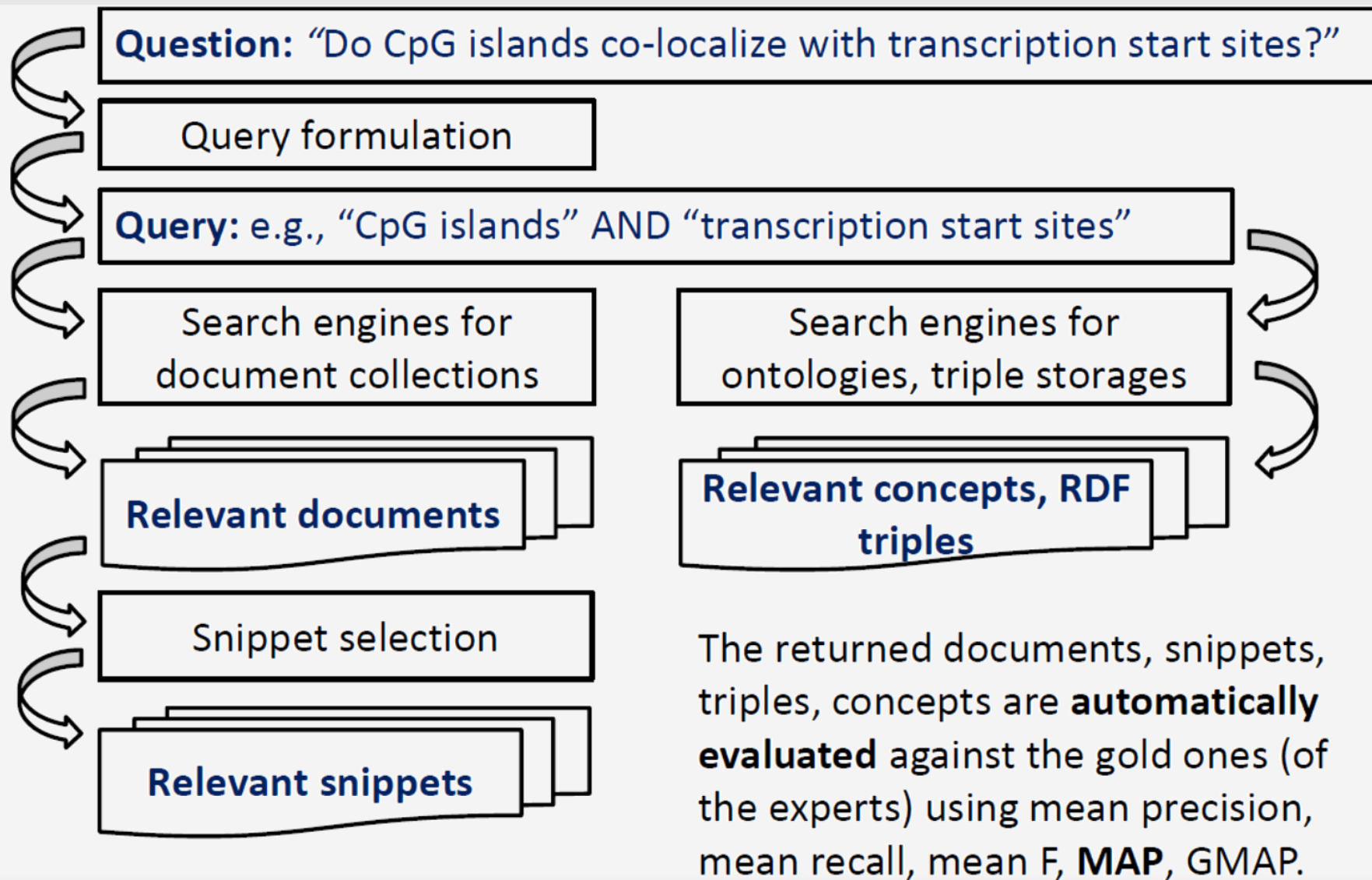
<http://www.disease-ontology.org/api/metadata/DOID:162> (cancer)
http://www.uniprot.org/uniprot/M3K8_RAT (TPL2 synonym)

What is your area of expertise?

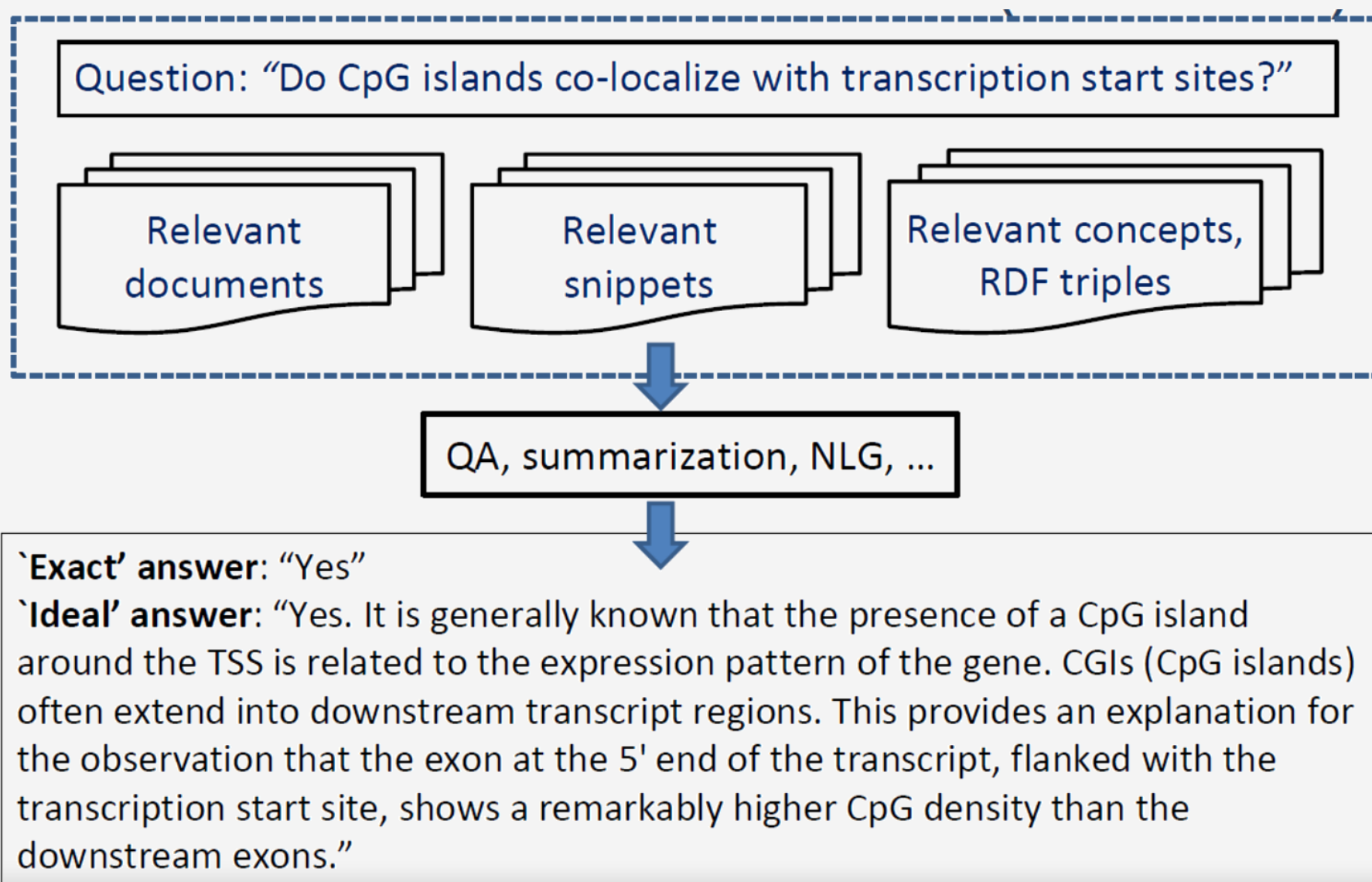
Click on the cloud to find out how your research interests relate to the BioASQ challenges.



BioASQ Task b –Phase A (Retrieval)



BioASQ Task b –Phase B(Answers)



BioASQ ecosystem

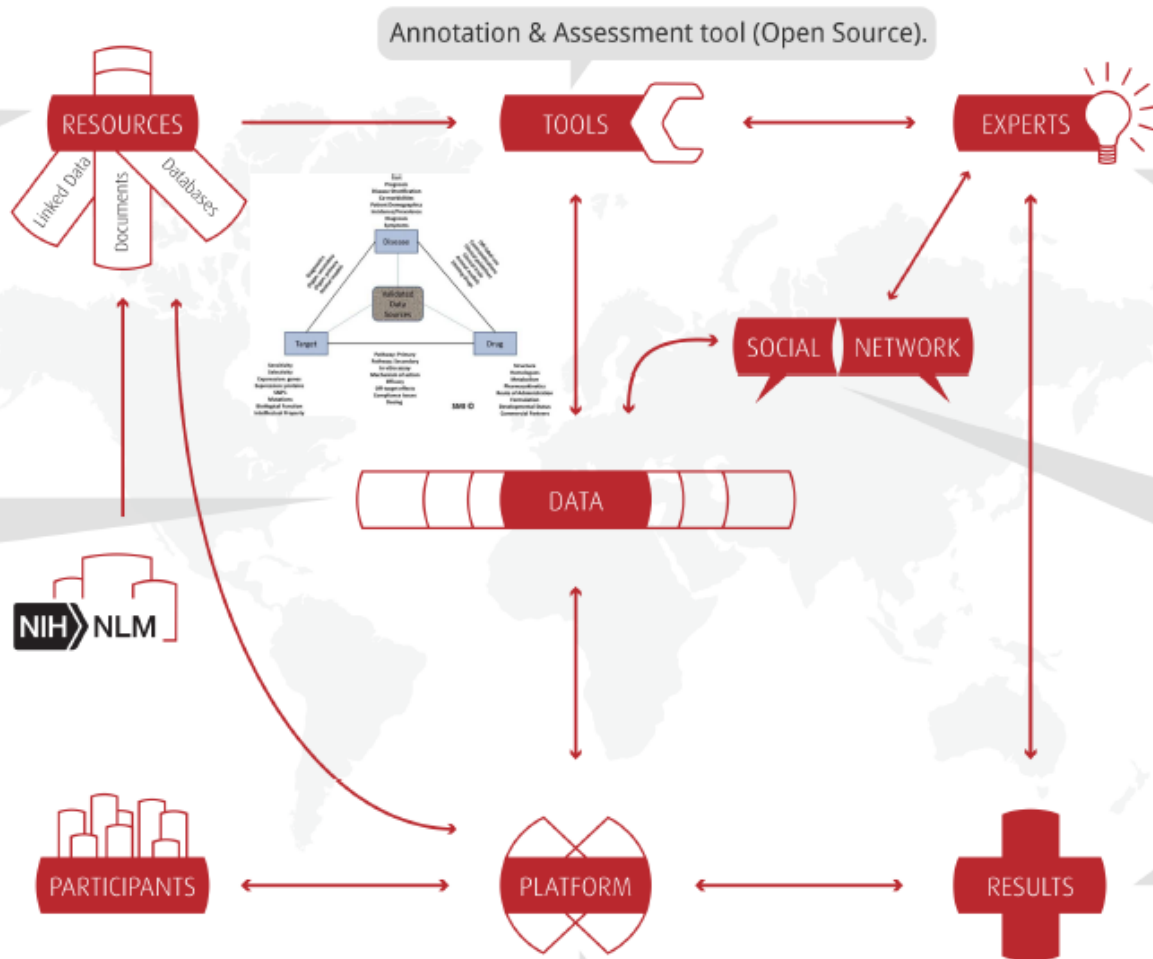
Data sources include both **text** and **structured** info.

- PubMed articles.
- MeSH, Gene Ontology, UniProt, Jochem, Disease Ontology.

Task A: More than 10,000,000 articles with MeSH headings available for training. New test batches available every week.

Task B: More than 1300 English questions, plus gold relevant documents, snippets, concepts, triples, "exact", and "ideal" answers.

Benchmark datasets/oracles available beyond the end of the project.



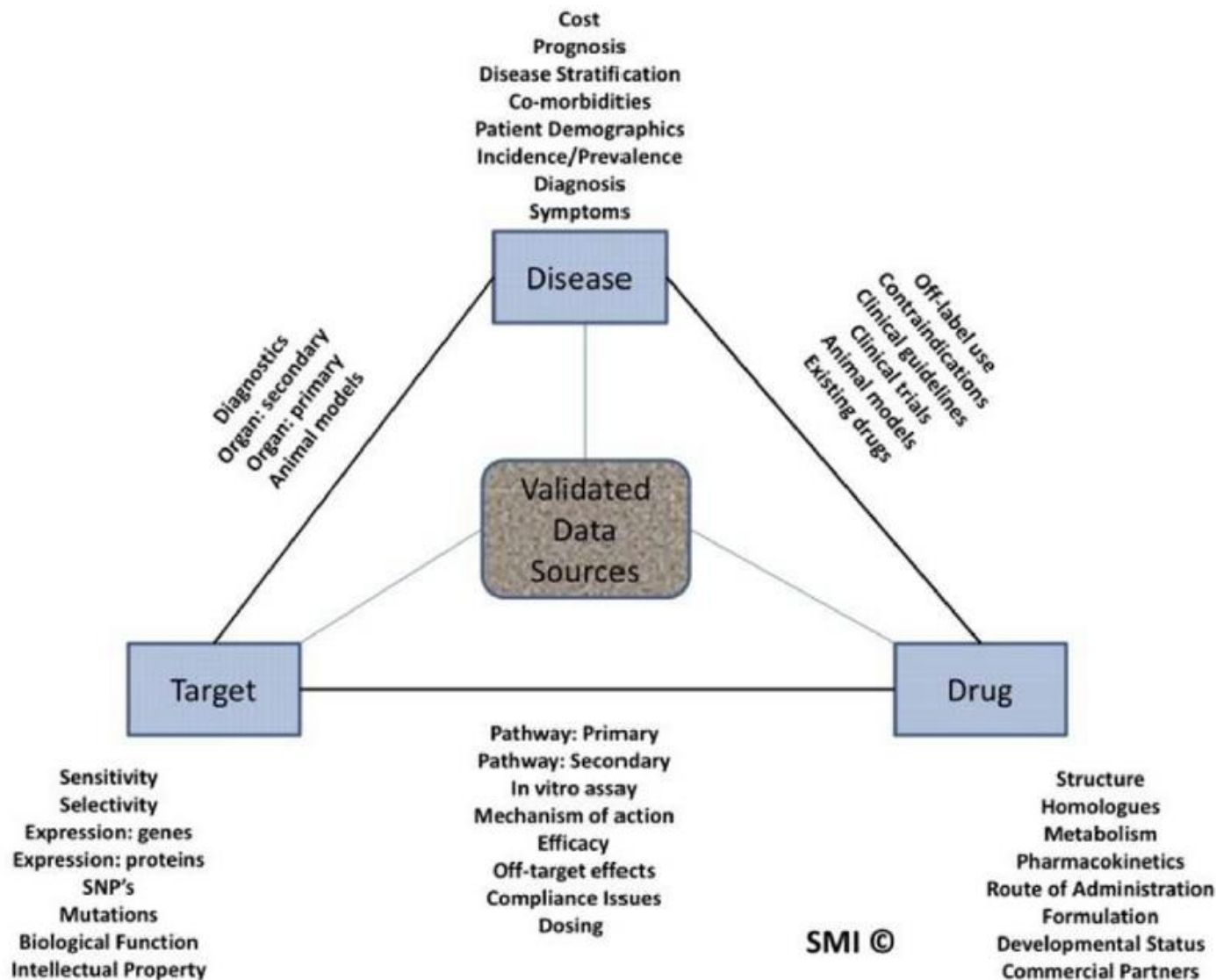
Biomedical experts, responsible for the formulation of questions, the identification of gold relevant documents, snippets, concepts, triples, the composition of gold "exact" and "ideal" answers, and the manual evaluation of system responses.

A social network allows biomedical experts to follow and comment upon benchmark questions and gold answers formulated by their peers.

Automated & manual assessment of system responses. Oracles available for objective retrospective benchmarking beyond the end of the project.

Online platform that enables participants to enter the BioASQ challenge. Infrastructure available beyond the end of the project.

Selection of data sources



Selection of Data Sources

- Ontologies/Biomedical Databases
 - Medical Subject Headings (MeSH)
 - Gene Ontology (GO)
 - Uniprot (Swissprot=reviewed only)
 - Disease Ontology (DO)
 - Joint Chemical Dictionary (Jochem)
 - Linked Life Data (Triples)
- Datasets
 - PubMed titles and abstracts (22M app.)
 - PubMed Central Full Text Articles (800k)

Joint Chemical Dictionary (Jochem)

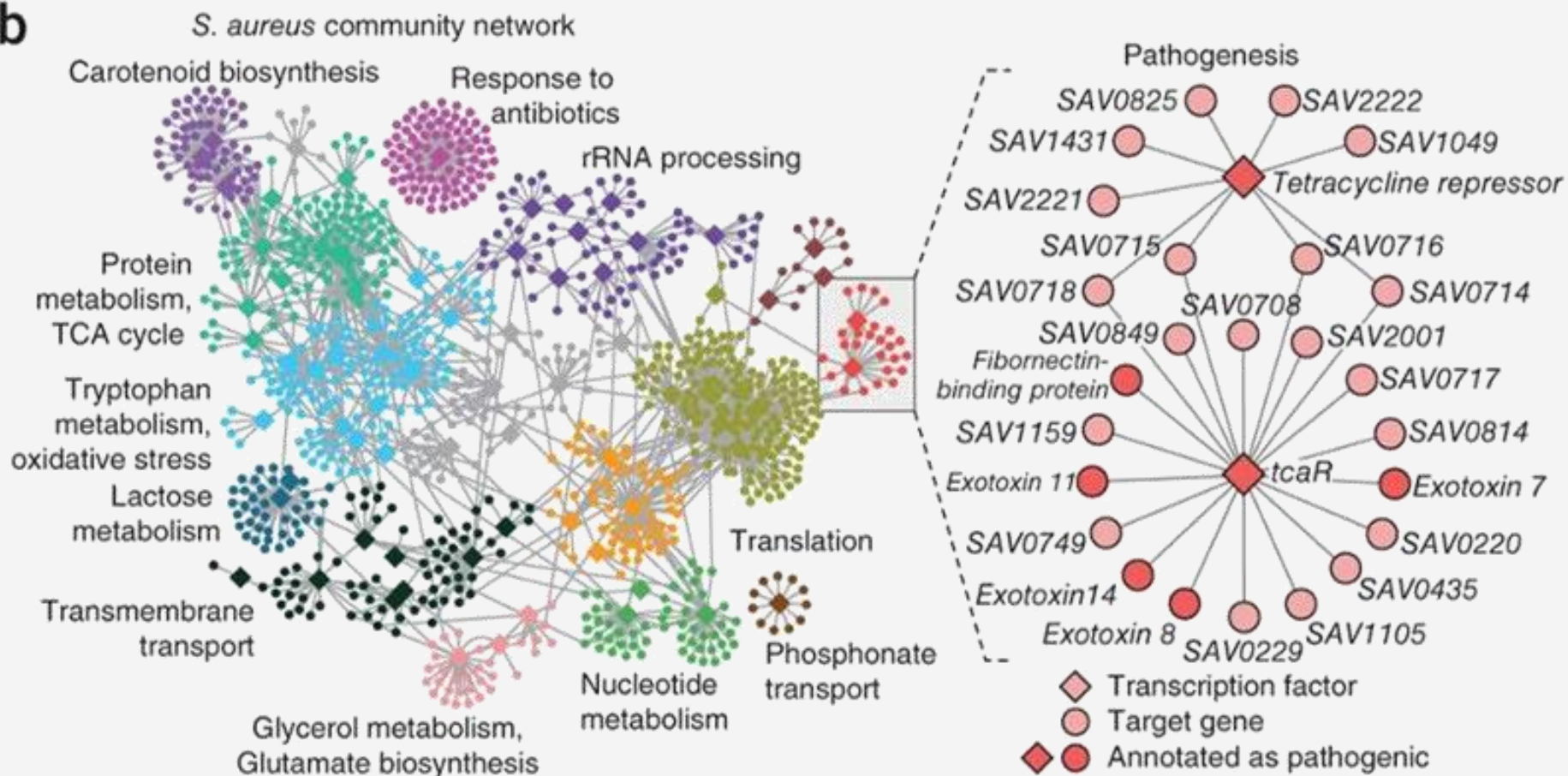
- the Joint Chemical Dictionary, is a dictionary for the identification of small molecules and drugs in text, combining information from UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB, and ChemIDplus.
- One of the largest biomedical resources for drugs and chemicals
- <http://biosemantics.org/index.php/resources/jochem>

Gene Ontology (GO)

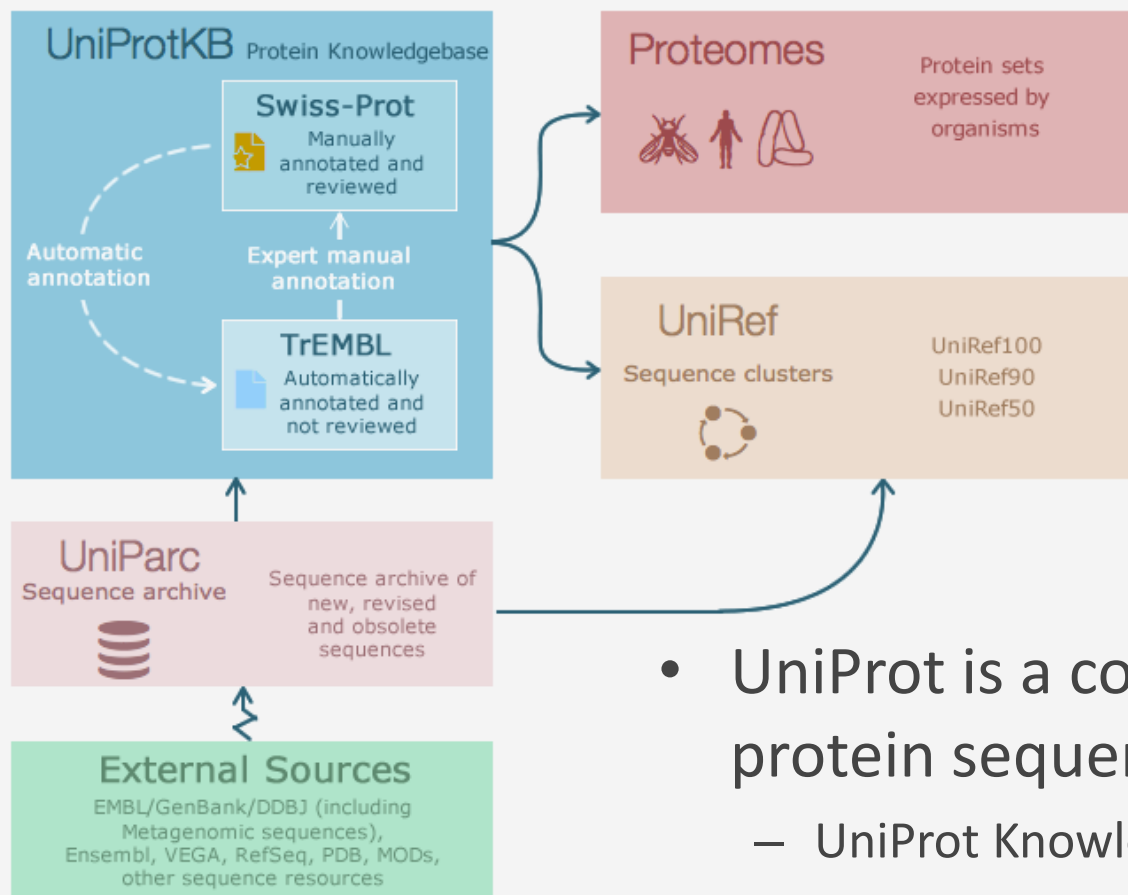
- a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
 - Maintain and develop its controlled vocabulary of gene and gene product attributes;
 - Annotate genes and gene products, and assimilate and disseminate annotation data;
 - Provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO, for example via enrichment analysis.
 - <http://geneontology.org/>

Gene Ontology (GO)

b



Universal Protein Resource (Uniprot)



- UniProt is a comprehensive resource for protein sequence and annotation data.
 - UniProt Knowledgebase (UniProtKB),
 - UniProt Reference Clusters (UniRef)
 - UniProt Archive (UniParc).
- <http://www.uniprot.org/>

Disease Ontology

- The Disease Ontology is a formal open source ontology for the integration of biomedical data that is associated with human disease.
- Approximately 8; 000 inherited, developmental and acquired human diseases are included in the resource.
- <http://disease-ontology.org/>

Linked Life Data (Triples)

- LinkedLifeData is a **data warehouse** that syndicates large volumes of **heterogeneous biomedical knowledge** in a common data model.
- The platform uses an **extension of the RDF model** that is able to track the provenance of each individual fact in the repository and thus update the information.
- It contains currently more than 8 billion statements, with almost 2 billion entities involved.
- The statements are extracted from **26 biomedical resources**, such as PubMed, UMLS, DrugBank, DisGeNET, and Gene Ontology.
- The statements are **publicly available**, and the project provides also a wide list of instance mappings.
- <http://linkedlifedata.com/>

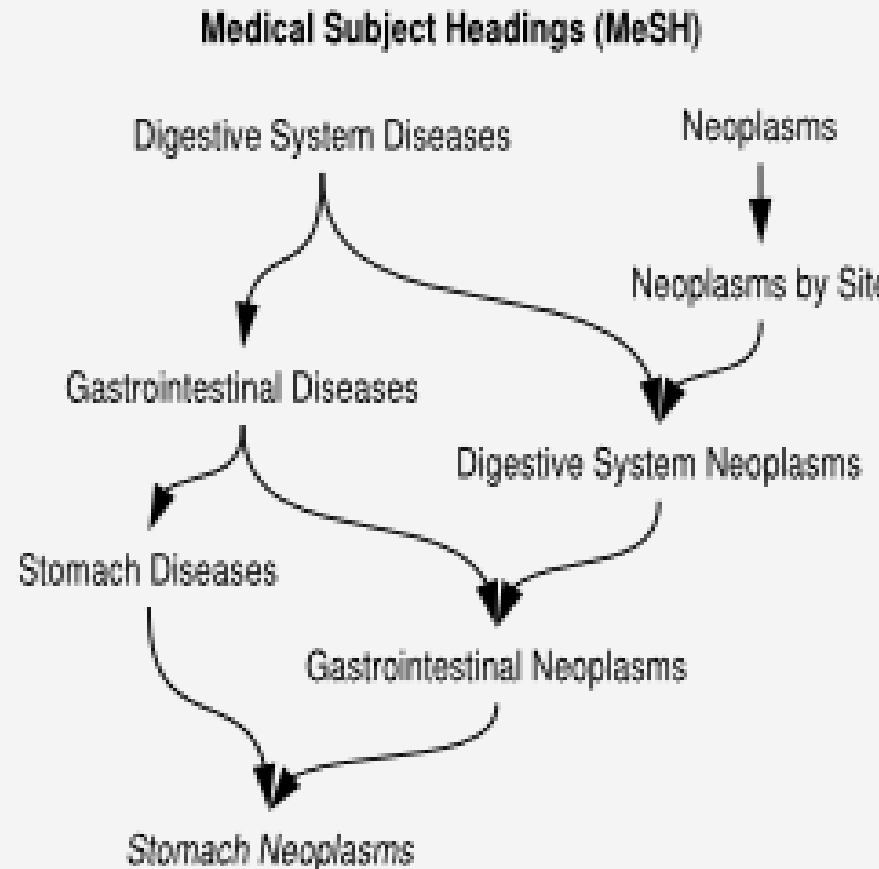
PubMed

- A service provided by the National Library of Medicine (NLM), under the U.S. National Institutes of Health (NIH).
- Contains over 23 million citations from Medline, a bibliographic database (DB) of biomedical literature, and other biomedical and life science journals dating back to the 1950s.
- It is accessible through the National Center for Biotechnology Information (NCBI).
- <http://www.ncbi.nlm.nih.gov/pubmed>

Medical Subject Headings

- **Medical Subject Headings (MeSH)**

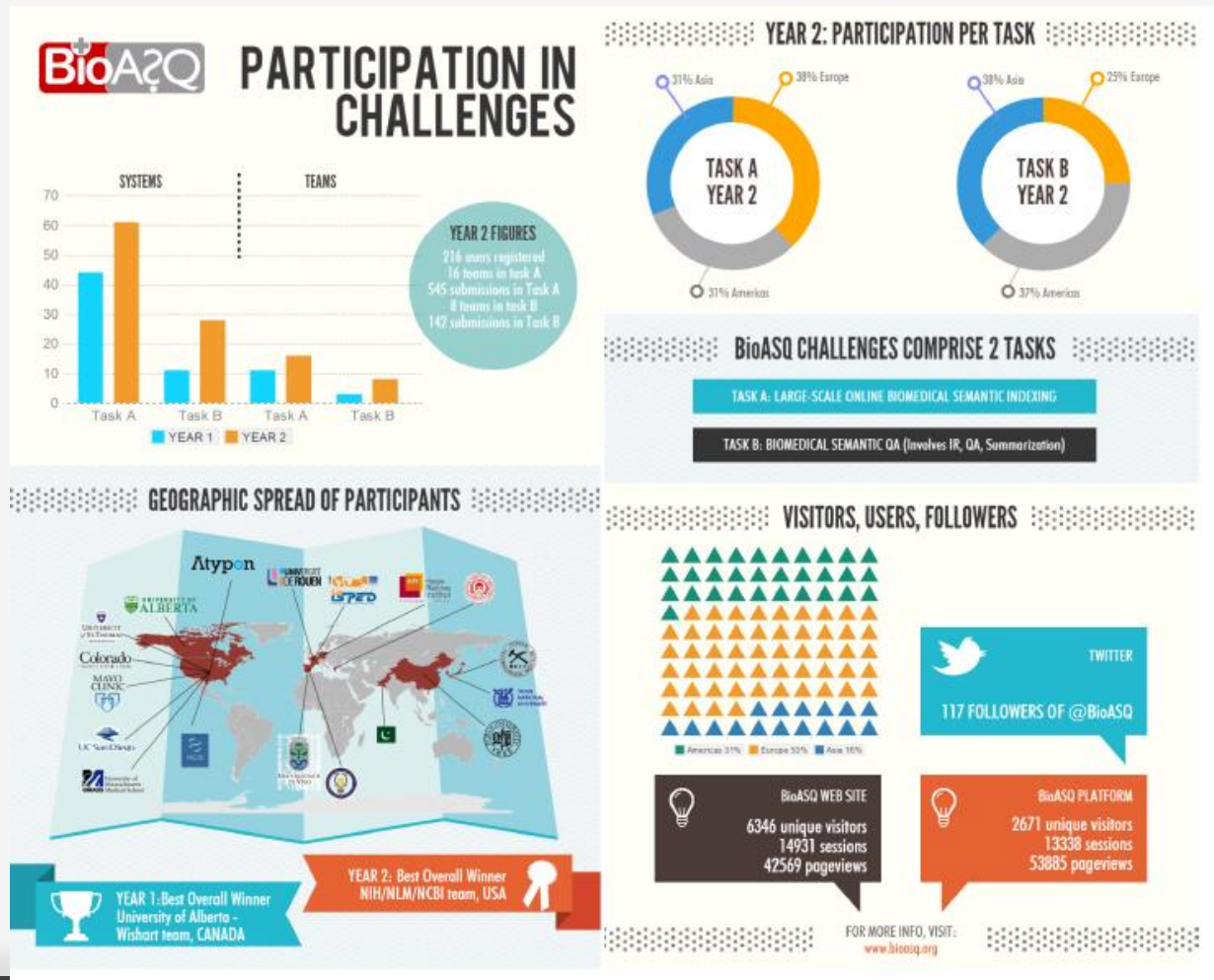
- MeSH is the National Library of Medicine's controlled **vocabulary thesaurus**. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.
- <https://www.nlm.nih.gov/mesh/>



Further Information

www.bioasq.org

Follow [@BioASQ](https://twitter.com/BioASQ)



Σύνοψη

- Εξαγωγή πληροφοριών από κείμενα.
- Βασικές έννοιες ανάκτησης πληροφοριών.
- Συστήματα ερωταποκρίσεων για δομημένα δεδομένα και συλλογές εγγράφων.
 - BioASQ
- Βιοιατρικές οντολογίες & δεδομένα