Open data & AI

Ανοικτά Δεδομένα και Τεχνητή Νοημοσύνη

Θοδωρής Παπαδόπουλος

Ερευνητής / Υποψήφιος Διδάκτορας

Πανεπιστήμιο Αιγαίου

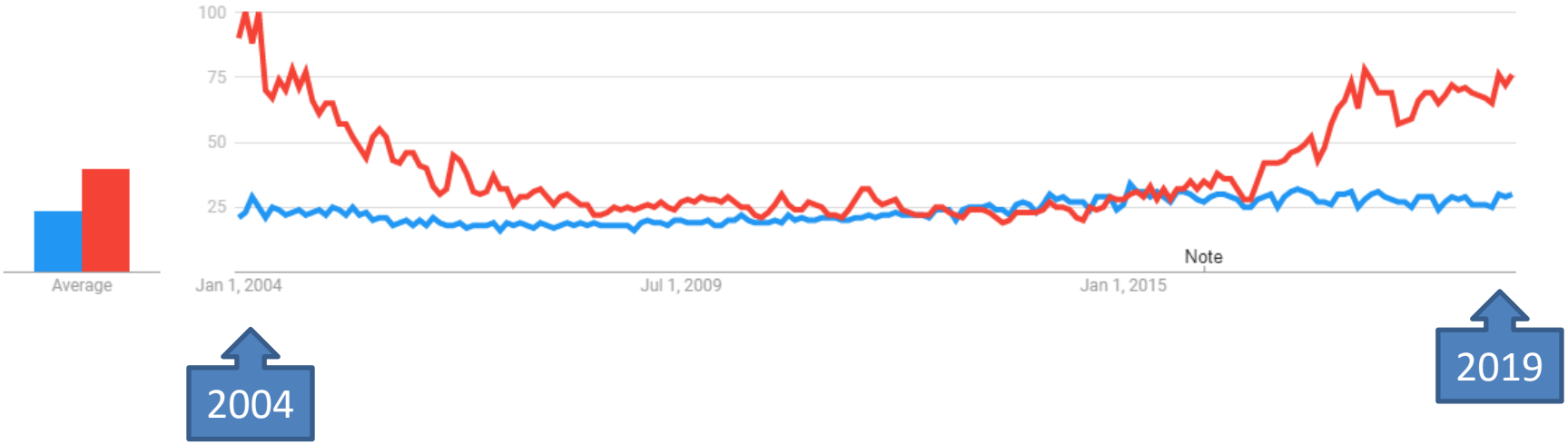
ΙΝ.ΕΠ.

ΗΜΕΡΙΔΑ ΑΝΟΙΚΤΩΝ ΔΕΔΟΜΕΝΩΝ

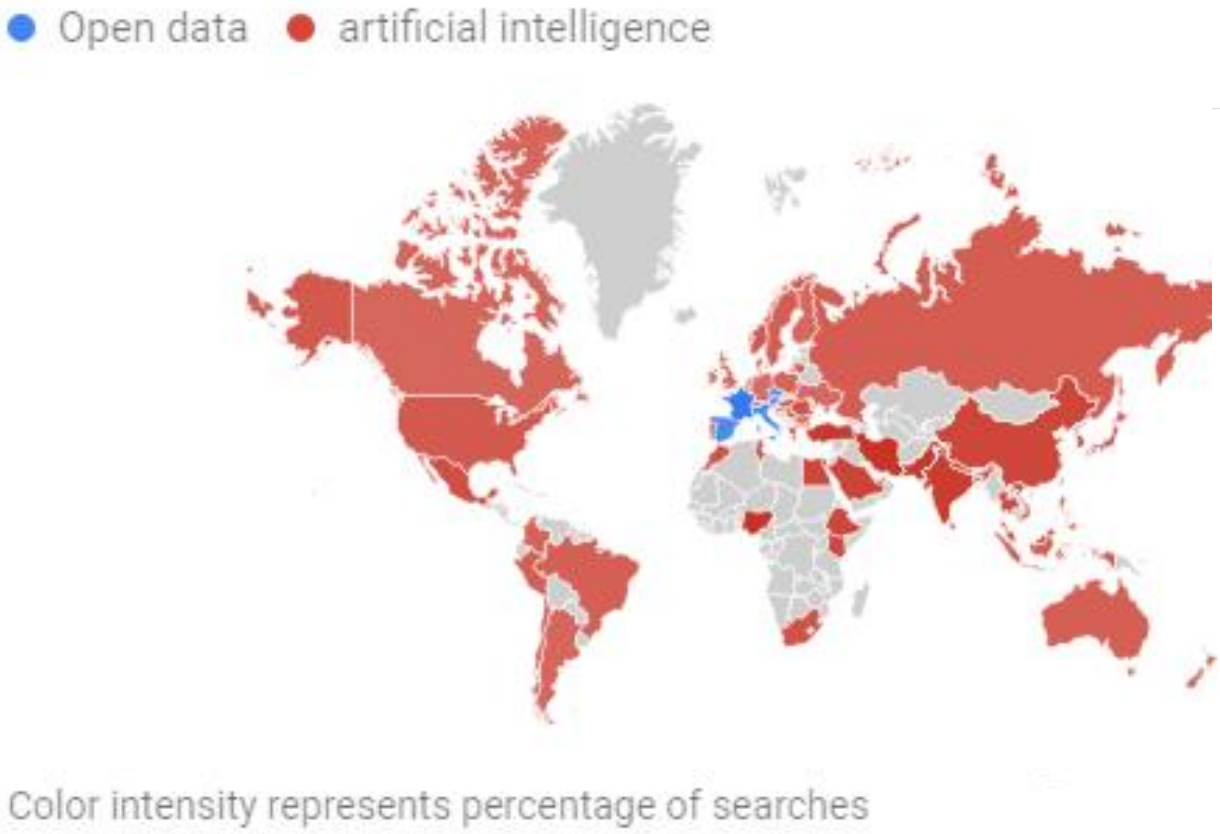
Περιεχόμενα

- Η Σχέση Ανοικτών Δεδομένων και Τεχνητής Νοημοσύνης (ΤΝ)
- Προκλήσεις για την Ιδιωτικότητα και τα Προσωπικά Δεδομένα
- Αξιοποίηση Ανοικτών Δεδομένων σε συστήματα ΤΝ

Open data vs AI (in time)



Open data vs AI (in space)



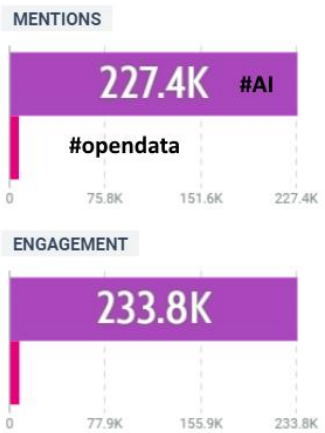
Sort: Interest for Open data

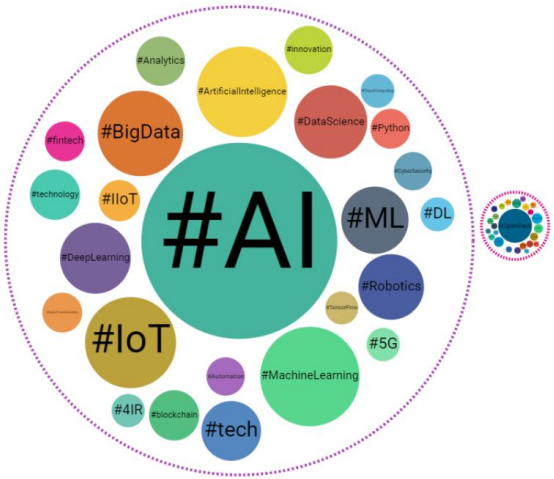
1	France	
2	Italy	
3	Taiwan	
4	Czechia	
5	Spain	

Sort: Interest for artificial intelligence

1	Nigeria	
2	Iran	
3	Pakistan	
4	Bangladesh	
5	Turkey	

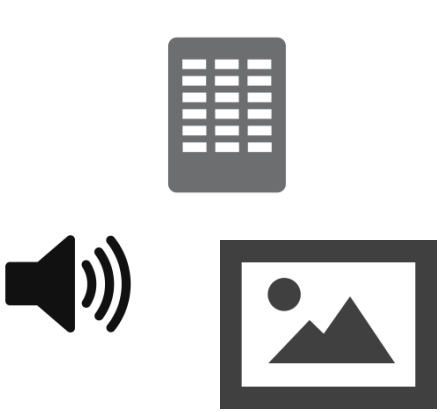
Open data vs AI (in twitter)

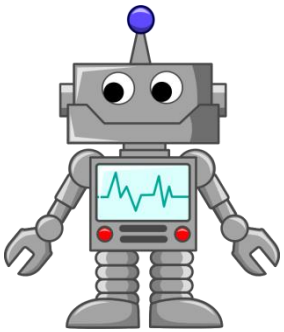




Labels	#opendata	#ai
#BigData	**77**	**5365**
#openscience	66	0
#data	54	0
#OpenSource	52	0
#DatosAbiertos	40	0
#opengov	29	0

Τι είναι η Τεχνητή Νοημοσύνη ;



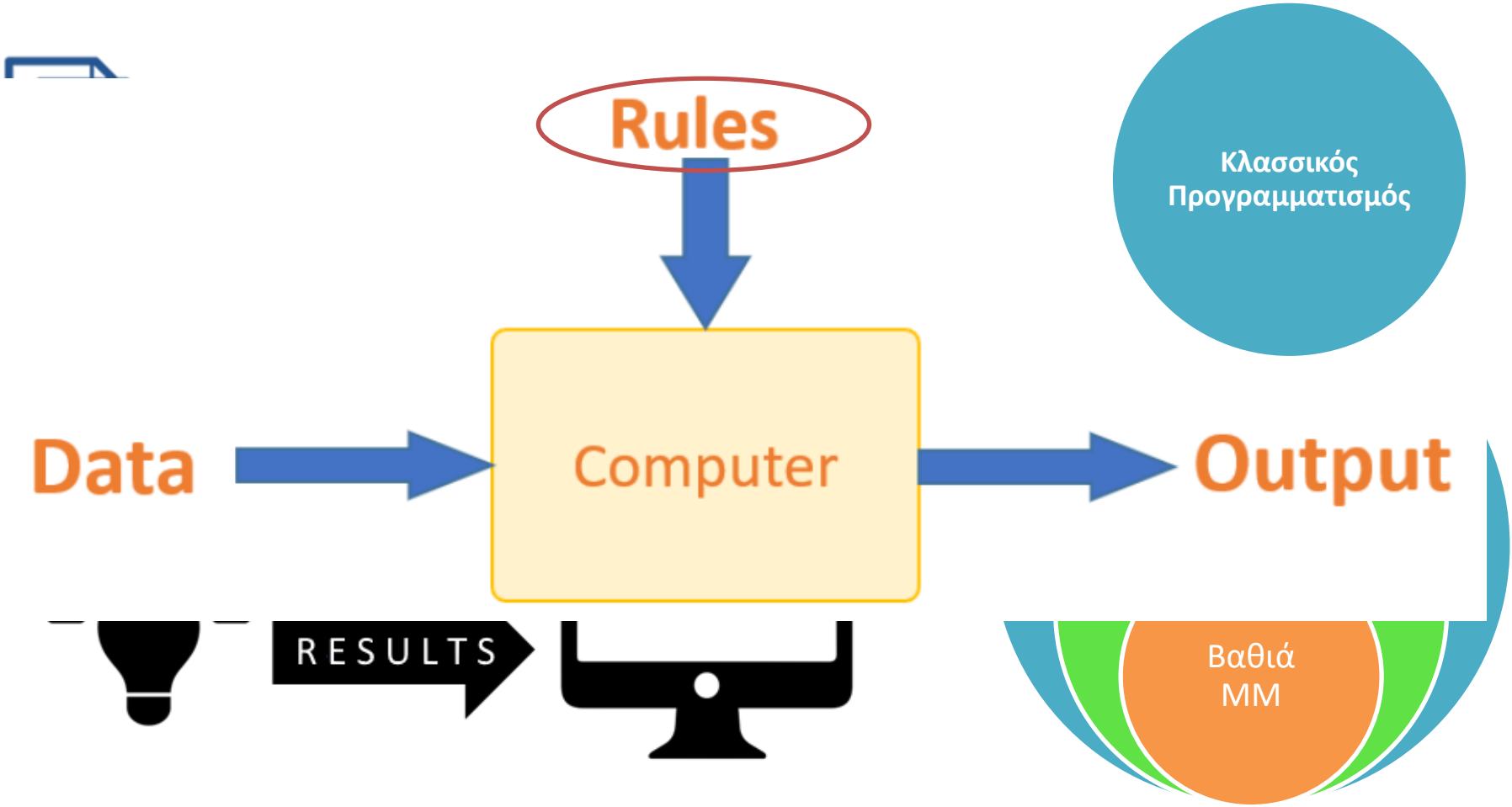


<u>Εισόδος:</u>
Δεδομένα
Αισθητήρες
Εικόνες / Ήχος

Τεχνητή νοημοσύνη

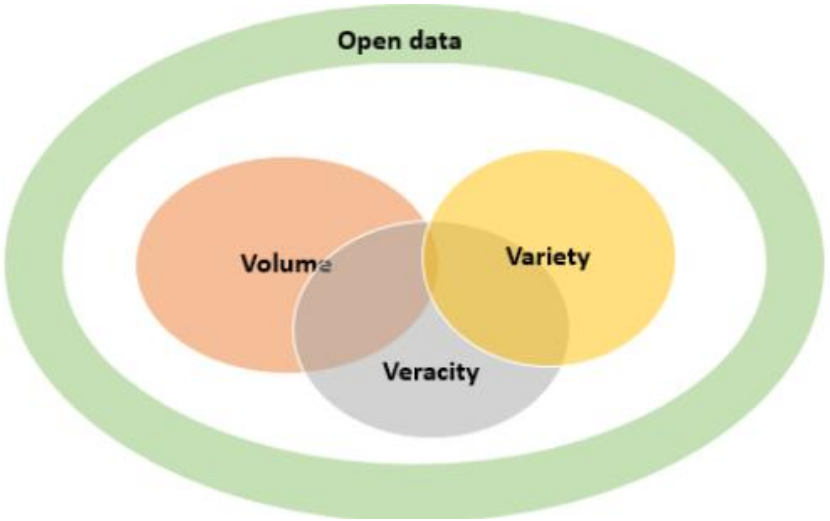
<u>Έξοδος:</u>
Δράση
Κίνηση
Κείμενο

Τι είναι η Τεχνητή Νοημοσύνη ;



Η σημασία των Ανοικτών Δεδομένων για την ΤΝ

- Η πρόσφατη αύξηση της ανάπτυξης της τεχνητής νοημοσύνης προκαλείται από έναν συνδυασμό μεγαλύτερης διαθεσιμότητας δεδομένων και πιο ισχυρών υπολογιστών
- 90% όλων των δεδομένων στον κόσμο δημιουργήθηκαν τα τελευταία δύο χρόνια
- Τα Ανοικτά Δεδομένα είναι ένας κρίσιμος πόρος για τους προγραμματιστές τεχνητής νοημοσύνης που αξιοποιούνται για την εκπαίδευση των συστημάτων τους
- Σημαντικές απαιτήσεις :
  1. Όγκος δεδομένων (Volume),
  2. Ποικιλία δεδομένων (Variety),
  3. Αλήθεια των δεδομένων (Veracity)



Η Τεχνητή Νοημοσύνη ως μέσο για τη βελτίωση ανοικτών δεδομένων

- Καθώς οι κυβερνήσεις υιοθετούν αλγόριθμους και συστήματα τεχνητής νοημοσύνης για τη βελτίωση της παροχής υπηρεσιών, πρέπει να διασφαλίζουν ότι τα συστήματα αυτά έχουν εκπαιδευτεί σε ανοικτά σύνολα δεδομένων που παράγουν δίκαια αποτελέσματα, καθώς και υπηρεσίες υψηλότερης ποιότητας.

- Η μηχανική μάθηση επιτρέπει να αυτοματοποιηθεί
  - Η μετατροπή μη δομημένων δεδομένων σε δομημένα δεδομένα.
    - Εξαγωγή πληροφορίας από κείμενο / κατηγοριοποίηση κειμένων (NLP)
    - Αναγνώριση εικόνας και video (π.χ. αριθμός διελεύσεων από διόδια , εντοπισμός σπιτιών με πισίνα κ.α.)
  - Ο εμπλουτισμός συνόλων δεδομένων με εξόρυξη γνώση που προκύπτει από τον εντοπισμό κρυμμένων μοτίβων σε υπάρχοντά σύνολα δεδομένων

Προκλήσεις : Ιδιωτικότητα και προσωπικά δεδομένα

- Η ικανότητα της μηχανικής μάθησης να αποκαλύπτει συνδέσεις μεταξύ συνόλων δεδομένων αυξάνει την ανησυχία για την ιδιωτικότητα και εγείρει νέα ηθικά ζητήματα σχετικά με τη δημοσίευση δεδομένων
- Οι κινητές συσκευές και οι έξυπνες συσκευές αποτελούν πλέον την βασική πηγή αυτοματοποιημένης παραγωγής προσωπικών δεδομένων χωρίς τη γνώση & συναίνεση του υποκειμένου αυτών των δεδομένων.
- The General Data Protection Regulation (GDPR)
  - Όρια στη λήψη αποφάσεων με βάση αποκλειστικά την αυτοματοποιημένη επεξεργασία και την **κατάρτιση προφίλ** (άρθρο 22).
    - Για παράδειγμα, ένα μοντέλο τεχνητής νοημοσύνης δεν μπορεί να είναι το μόνο βήμα για να αποφασιστεί εάν ένας δανειολήπτης είναι επιλέξιμος για να δικαιούται δάνειο.
  - Δικαίωμα να παρέχονται ουσιαστικές πληροφορίες σχετικά με τη λογική βάσει της οποίας λήφθηκε μια αυτοματοποιημένη απόφαση από ένα σύστημα ΤΝ (άρθρο 13 παράγραφος 2 στοιχείο στ. και άρθρο 15 παράγραφος 1 στοιχείο η.)

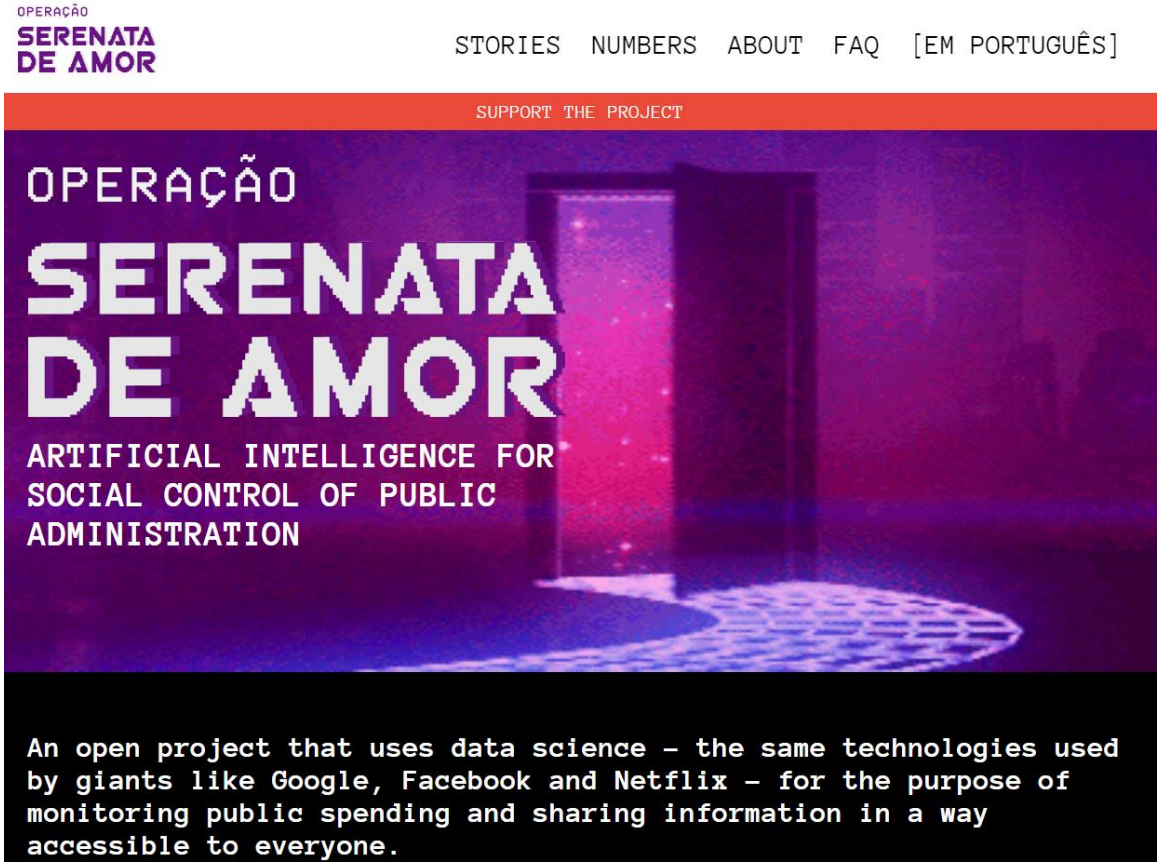
Προκλήσεις : Μεροληψία και διακρίσεις

- **Η ΤΝ είναι τόσο καλή όσο τα δεδομένα πάνω στα οποία εκπαιδεύεται**
  - «κακά δεδομένα» μπορεί να περιέχουν έμμεσες φυλετικές, έμφυλες ή ιδεολογικές προκαταλήψεις
  - ένας αλγόριθμος ποινικής δικαιοσύνης που χρησιμοποιείται στην Φλόριντα, χαρακτήριζε εσφαλμένα Αφρο-Αμερικάνους κατηγορουμένους ως "υψηλού κινδύνου" σε ρυθμό διπλάσιο από τον πραγματικό .
  - Ένα σύστημα ΤΝ της amazon για προσλήψεις προσωπικού εμφάνισε μεροληψία υπερ των ανδρών, γιατί στο σύνολο δεδομένων που εκπαιδεύτηκε οι γυναίκες υπο-εκπροσωπούνταν
- Η **επεξηγησιμότητα** της ΤΝ διαδραματίζει σημαντικό ρόλο στην επίτευξη δικαιοσύνης, λογοδοσίας και διαφάνειας στη μηχανική μάθηση. Βασίζεται στην ιδέα ότι όλες οι αυτοματοποιημένες αποφάσεις που ελήφθησαν θα πρέπει να μπορούν να εξηγηθούν.
- Συστήματα ΤΝ τα οποία μπορούν να περιγράψουν το σκοπό, το σκεπτικό και τη διαδικασία λήψης αποφάσεων με τρόπο που να μπορεί να γίνει κατανοητός από τον μέσο άνθρωπο.

Προκλήσεις : Ιδιόκτητα σύνολα δεδομένων με δημόσιο ενδιαφέρον

- Μεγάλες επιχειρήσεις έχουν τόσο τους υπολογιστικούς πόρους όσο και την πρόσβαση σε ιδιόκτητα σύνολα δεδομένων για να συνδυαστούν με ανοικτά δεδομένα, είναι πιθανό να διατηρήσουν ανταγωνιστικό πλεονέκτημα (Facebook / Google / airbnb / skroutz)
- Νέες ρυθμιστικές απαντήσεις
- ***Η δημόσια πολιτική θα πρέπει να ενθαρρύνει την ευρύτερη διαθεσιμότητα ιδιωτικών δεδομένων, διασφαλίζοντας παράλληλα τον πλήρη σεβασμό της νομοθεσίας για την προστασία των δεδομένων προσωπικού χαρακτήρα. Η Επιτροπή καλεί τις εταιρείες να αναγνωρίσουν τη σημασία της επαναχρησιμοποίησης των μη προσωπικών δεδομένων, για την εκπαίδευση Τεχνητής Νοημοσύνης.***
  - *«Ανακοίνωση της Επιτροπής προς το Ευρωπαϊκό Κοινοβούλιο, το Ευρωπαϊκό Συμβούλιο, το Συμβούλιο, την Ευρωπαϊκή Οικονομική και Κοινωνική Επιτροπή και την Επιτροπή των Περιφερειών — Τεχνητή νοημοσύνη για την Ευρώπη» [COM(2018) 237 final])*

Εφαρμογές που συνδυάζουν Ανοικτά Δεδομένα και ΤΝ



Operacao SERENATA DE AMOR

Το σύστημα χρησιμοποιεί ένα δημοφιλές κιτ εργαλείων μηχανικής μάθησης για να αναλύσει εγγραφές δημοσίων δαπανών και να εντοπίσει «υπερβολικά ακριβές» πληρωμές που επισημαίνει για ανθρώπινη έρευνα. Μέχρι σήμερα, το εργαλείο έχει εντοπίσει πάνω από 8000 ύποπτες πληρωμές, συνολικής αξίας 3,6 εκατομμυρίων.
Twitter bot

79% accuracy using an artificial intelligence (AI) method developed by researchers at UCL, the University of Sheffield and the University of Pennsylvania.

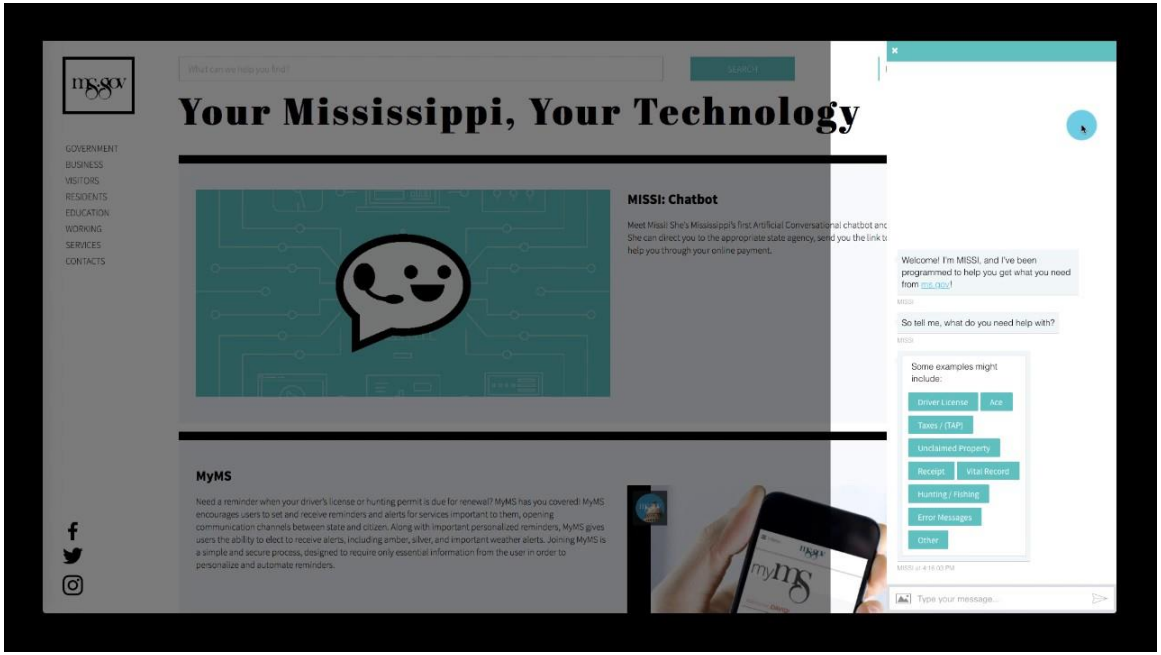
Προγνωστική δικαιοσύνη – Ευρωπαϊκό Δικαστήριο ΔΑ

Λογισμικό που προβλέπει δικαστικές αποφάσεις με βάση την ανάλυση μεγάλης ποσότητας νομολογία. Μελέτη του Πανεπιστημιακού Κολλεγίου του Λονδίνου βάσει 584 αποφάσεων του ΕΔΔΑ.

80% ποσοστό επιτυχίας

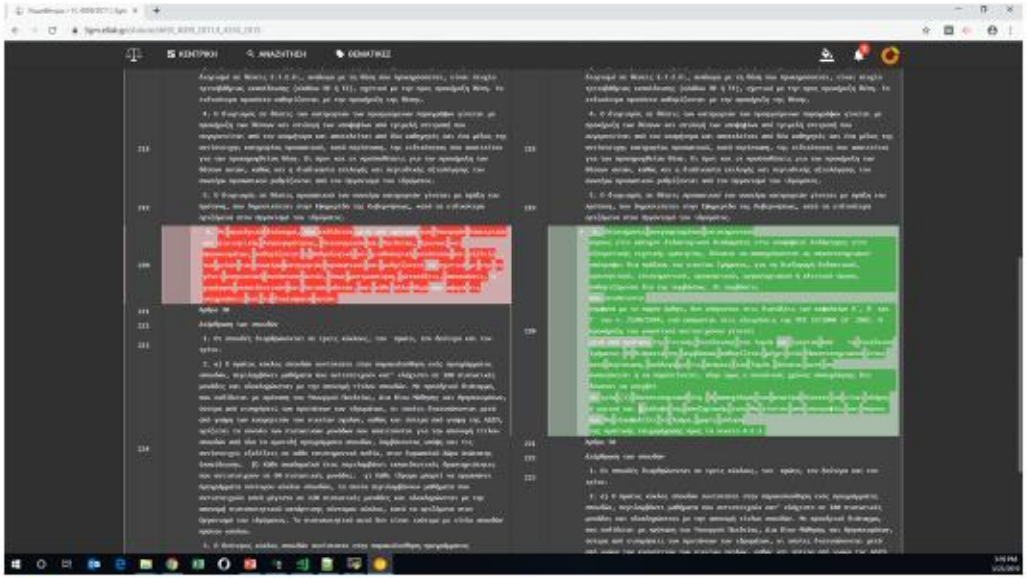
Μέριλαντ – Εντοπισμός Φορολογικής απάτης

Το Μέριλαντ χρησιμοποιεί τεχνικές μηχανικής μάθησης για την ανίχνευση απάτης στην απόδοση της φορολογίας. Με τη χρήση μεγάλου όγκου δεδομένων και αλγορίθμων για την πρόβλεψη μοτίβων οι ελεγκτές μπορούν να βρουν φορολογικές απάτες και να λάβουν τα κατάλληλα μέτρα. Αύξηση των φορολογικών δηλώσεων στο κράτος από περίπου 15 εκατομμύρια ετησίως σε 40 εκατομμύρια



MISSI – chatbot

Το MISSI είναι η εφαρμογή chatbot της πολιτείας του Μισισιπή και έχει σχεδιαστεί για να βοηθήσει τους κατοίκους και τους επισκέπτες να μάθουν πληροφορίες σχετικά με τις συναλλαγές τους με το κράτος. Η εφαρμογή chatbot δίνει τη δυνατότητα διάδρασης είτε μέσω ενός παραθύρου μηνυμάτων συνομιλίας ή με φωνητική είσοδο μέσω της φωνής βοηθού Αλέξα της Amazon.



OPENLAWS

Εκτελεί την εξαγωγή κειμένων των ΦΕΚ από το Εθνικό Τυπογραφείο (ΕΤ), τα διασυνδέει μεταξύ τους και, τέλος, προσδιορίζει και εφαρμόζει τροποποιήσεις του νομικού κειμένου μέσω της αυτόματης κωδικοποίησης της ελληνικής νομοθεσίας με μεθόδους και τεχνικές επεξεργασίας φυσικής γλώσσας.

Openlaws - Στοιχεία Έργου

- Αναπτύχθηκε υπό την αιγίδα του προγράμματος **Google Summer of Code 2018**
- Εκτελεί την εξαγωγή κειμένων των ΦΕΚ από το Εθνικό Τυπογραφείο (ΕΤ), τα διασυνδέει μεταξύ τους και, τέλος, προσδιορίζει και εφαρμόζει τροποποιήσεις του νομικού κειμένου μέσω της αυτόματης κωδικοποίησης της ελληνικής νομοθεσίας με μεθόδους και τεχνικές επεξεργασίας φυσικής γλώσσας (NLP).
- https://openlaws.ellak.gr/home
- 3-4 ανθρωπομήνες
- development state
- Ανοιχτός κώδικας διαθέσιμος στο github

Ευχαριστώ !

Θοδωρής Παπαδόπουλος ,
t.papadopoulos@aegean.gr