

ΥΠΟΛΟΓΙΣΤΙΚΕΣ  
ΕΠΙΣΤΗΜΕΣ



VIRTUAL EVENT

# Ημέρες Δράσης

23-24-25  
ΙΟΥΝΙΟΥ  
2021



**ΑΧΑΝΑ'**  
Έρευνα & Καινοτομία  
Τεχνολογίες Πληροφορίας



ΕΥΦΥΗ  
ΒΙΟΜΗΧΑΝΙΚΑ  
ΣΥΣΤΗΜΑΤΑ



ΤΕΧΝΟΛΟΓΙΕΣ  
ΔΕΔΟΜΕΝΩΝ  
ΚΑΙ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ

# Ο Κόσμος της Επιστήμης Μεγάλων Δεδομένων

## ΜΕΔΑ.2 – Βιβλιογραφία και Τεχνολογία Αιχμής



**Χάρης Νάκος**

Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών  
**Ινστιτούτο Πληροφοριακών Συστημάτων**



**ΑΘΗΝΑ**

Έρευνα & Καινοτομία  
Τεχνολογίες Πληροφορίας

Τα προαναφερθέντα έργα υλοποιήθηκαν στο πλαίσιο της Πρόσκλησης με κωδικό 031 με τίτλο «ΔΡΑΣΗ ΣΤΡΑΤΗΓΙΚΗΣ ΑΝΑΠΤΥΞΗΣ ΕΡΕΥΝΗΤΙΚΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΩΝ ΦΟΡΕΩΝ», που χρηματοδοτείται από το Επιχειρησιακό Πρόγραμμα «Ανταγωνιστικότητα, Επιχειρηματικότητα και Καινοτομία», ΕΣΠΑ 2014-2020 ΕΣΠΑ 2014-2020, με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης (Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης). Αναλυτικά: "Υπολογιστικές Επιστήμες και Τεχνολογίες Δεδομένων, Περιεχομένου και Αλληλεπίδρασης" (MIS 5002437) με 3 υποέργα: Υποέργο 1 «Γλωσσικές Τεχνολογίες Ανάλυσης Περιεχομένου και Αλληλεπίδρασης», Υποέργο 2 «Τεχνολογίες Ανάλυσης Περιεχομένου στον Πολιτισμό», Υποέργο 3 «ΜΕΔΑ: Μεγάλα Δεδομένα – Προκλήσεις, Μέθοδοι και Τεχνικές Αποδοτικής Διαχείρισης» και "Καινοτόμος εφαρμογή του βιομηχανικού διαδικτύου των πραγμάτων σε ευφυή περιβάλλοντα" (MIS 5002434) με Υποέργο 1: "Καινοτόμος εφαρμογή του βιομηχανικού διαδικτύου των πραγμάτων σε ευφυή περιβάλλοντα".



Ευρωπαϊκή Ένωση  
Ευρωπαϊκό Ταμείο  
Περιφερειακής Ανάπτυξης



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ  
ΥΠΟΥΡΓΕΙΟ  
ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΣΕΩΝ  
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΔΙΑΡΘΡΩΤΙΚΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ  
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΙΧΕΙΡΗΣΙΑΚΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ



Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



## ΙΠΣΥ, Μεγάλα Δεδομένα, ΜΕΔΑ.2

Το ΙΠΣΥ μέσα από τα ερευνητικά πεδία του ασχολείται με τα Μεγάλα Δεδομένα

ΜΕΔΑ.2: Βηματισμός πέρα από το state-of-the-art





# Αρχικοί στόχοι έργου

Εξοικείωση με προχωρημένα εργαλεία και βιβλιοθήκες της επιστήμης δεδομένων

- Επεξεργασία δεδομένων
- Μηχανική μάθηση
- Εξόρυξη γνώσης
- Ενσωμάτωση και ομογενοποίηση ετερογενών δεδομένων
- Στατιστική και οπτική ανάλυση

Επισκόπηση ερευνητικής βιβλιογραφίας

- Τεχνικές επιστήμης δεδομένων που αξιοποιούν μεγάλους όγκους δεδομένων
- Μελέτη πραγματικών συστημάτων και μελέτες περίπτωσης

Τα ερευνητικά πεδία του ΙΠΣΥ που αφορούν μεγάλα δεδομένα αποτελούν το πρίσμα μέσα από το οποίο γίνεται η εξέταση των εργαλείων της επιστήμης δεδομένων

Τα συμπεράσματα τροφοδοτούν το επόμενο στάδιο του έργου για την πρακτική εξοικείωση με τα εργαλεία που κρίνονται καταλληλότερα για την εξυπηρέτηση των ερευνητικών στόχων





## Τεχνολογική Στάθμη Ερευνητικών Πεδίων

Ερευνητικές κατευθύνσεις που εξετάστηκαν στο πλαίσιο του έργου

- Ανάλυση δεδομένων κοινωνικών δικτύων για τον αυτόματο εντοπισμό ψευδών ειδήσεων
- Επεξεργασία ερωτημάτων σε δεδομένα μορφής γράφου
- Χωρική ευρετηρίαση και ανάλυση σε μεγάλα δεδομένα
- Ερωτήματα συνένωσης σε γεωχωρικά δεδομένα
- Ανάλυση γονιδιωματικών δεδομένων
- Ανάλυση βιβλιογραφικών δεδομένων
- Ανωνυμοποίηση σε δεδομένα υγείας
- RDF δεδομένα





# Παραδείγματα Ερευνητικών Κατευθύνσεων του ΙΠΣΥ





## Ανάλυση Δεδομένων Κοινωνικών Δικτύων για τον Αυτόματο Εντοπισμό Ψευδών Ειδήσεων

Στόχοι και ερευνητική κατεύθυνση

- Κοινωνικά δίκτυα: αποτελεσματικοί δίαυλοι για ακούσια και εκούσια παραπληροφόρηση
- Ανάπτυξη αλγοριθμικών μεθόδων για τον αυτόματο εντοπισμό παραπλανητικού περιεχομένου στα κοινωνικά δίκτυα

Τεχνολογική στάθμη

- Διασταύρωση γνώσης με χρήση γράφων γνώσης
- Επεξεργασία φυσικής γλώσσας για εξαγωγή γλωσσικών χαρακτηριστικών που αποτελούν ενδείξεις εξαπάτησης
- Εξέταση χαρακτηριστικών διεργασίας διάχυσης είδησης στο δίκτυο





# Ανάλυση Γονιδιωματικών Δεδομένων

Στόχοι και ερευνητική κατεύθυνση

- Ανάλυση ακολουθιών RNA και DNA οργανισμών
- Ανακάλυψη ρόλου γονιδίων, μορίων miRNA, και άλλων βιομορίων σε βιολογικούς μηχανισμούς

Τεχνολογική στάθμη

- Αλγόριθμοι και υπολογιστικές τεχνικές για την πρόβλεψη ομάδων γονιδίων που συμμετέχουν σε βιολογικές διαδικασίες ενδιαφέροντος
- Στατιστικές αναλύσεις γονιδιακών δεδομένων





# Ανωνυμοποίηση σε Δεδομένα Υγείας

Στόχοι και ερευνητική κατεύθυνση

- Επεξεργασία ιστορικού υγείας ασθενή
- Προστασία προσωπικών δεδομένων χωρίς σημαντική απώλεια πληροφορίας

Τεχνολογική στάθμη

- Τεχνικές ομοιότητας δεδομένων
- Τα δεδομένα υγείας ασθενών ομαδοποιούνται σε γενικότερες κλάσεις, ώστε να μην υπάρχει η δυνατότητα ταυτοποίησης κάποιου ασθενή μέσα από τα επεξεργασμένα δεδομένα





# Εργαλεία της Επιστήμης Δεδομένων

Κατηγορίες εργαλείων

- Μαζική επεξεργασία δεδομένων (Processing)
- Μηχανές αποθήκευσης (Storage Engines)
- Επεξεργασία ροών δεδομένων (Streaming)
- Μηχανική μάθηση γενικού σκοπού (General-Purpose Machine Learning)
- Βαθιά μάθηση (Deep Learning)
- Οπτικοποίηση και αναλυτική δεδομένων (Data Visualization / Analytics)
- Διαχειριστές πόρων και δρομολογητές εργασιών (Resource Managers / Schedulers)
- Μηχανές ερωτημάτων και αναλυτική (Query Engines / Analytics)
- Μηχανές χωρικής αναλυτικής (Spatial Analytics Engines)





# Μεθοδολογία Σύγκρισης Εργαλείων

Διαστάσεις σύγκρισης

- Ημερομηνία κυκλοφορίας αρχικής έκδοσης
- Ημερομηνία κυκλοφορίας τρέχουσας έκδοσης
- Άδεια
- Γλώσσα υλοποίησης
- Υποστηριζόμενα λειτουργικά συστήματα
- Προγραμματιστικές διεπαφές / μέθοδοι πρόσβασης
- Υποστηριζόμενες γλώσσες προγραμματισμού
- Υποστηριζόμενοι αλγόριθμοι
- Προγραμματιστικά πρότυπα
- Μοντέλα αποθήκευσης δεδομένων
- Υποστηριζόμενοι τύποι ευρετηρίων
- Υποστήριξη αντιγράφων
- Λοιπές εξειδικευμένες διαστάσεις ανάλογα με την κατηγορία των εργαλείων





# Παραδείγματα Σύγκρισης Εργαλείων



# Εργαλεία Επεξεργασίας Ροών Δεδομένων

| Name                                    | Kafka                                                             | Storm                                                           | Apache Samza                                                    | Apache Flink                                                      | Spark Streaming                                                                       | NAIAD                                                                                                                                                                                                                                          |
|-----------------------------------------|-------------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------------------|-------------------------------------------------------------------|---------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Ιστοσελίδα                              | <a href="https://kafka.apache.org/">https://kafka.apache.org/</a> | <a href="http://storm.apache.org/">http://storm.apache.org/</a> | <a href="http://samza.apache.org/">http://samza.apache.org/</a> | <a href="https://flink.apache.org/">https://flink.apache.org/</a> | <a href="https://spark.apache.org/streaming/">https://spark.apache.org/streaming/</a> | <a href="https://github.com/frankmcsherry/timely-dataflow">https://github.com/frankmcsherry/timely-dataflow</a><br><a href="https://github.com/frankmcsherry/differential-dataflow">https://github.com/frankmcsherry/differential-dataflow</a> |
| Πρώτη έκδοση                            | 2011                                                              | 2011                                                            | 2014                                                            | 2017                                                              | 2014                                                                                  | 2013                                                                                                                                                                                                                                           |
| Τρέχουσα έκδοση                         | 1.0 / Νοέμβριος 2017                                              | 1.2.1 / Φεβρουάριος 2018                                        | 0.13.1 / Αύγουστος 2017                                         | 1.6.2 / Οκτώβριος 2018                                            | Φεβρουάριος 2018                                                                      | Νοέμβριος 2018                                                                                                                                                                                                                                 |
| License                                 | Apache License 2.0                                                | Apache License 2.0                                              | Apache License 2.0                                              | Apache License 2.0                                                | Apache License 2.0                                                                    | MIT                                                                                                                                                                                                                                            |
| Γλώσσα υλοποίησης                       | Scala, Java                                                       | Clojure, Java                                                   | Scala, Java                                                     | Scala, Java                                                       | Scala                                                                                 | Rust                                                                                                                                                                                                                                           |
| APIs / μέθοδοι πρόσβασης                | Producer API, Consumer API, Streams API, Connector API            | -                                                               | -                                                               | Dataset API, Table API, SQL                                       | APIs σε Python, Java, Scala και SQL, και πολλές ενσωματωμένες βιβλιοθήκες             | APIs σε Rust                                                                                                                                                                                                                                   |
| Υποστηριζόμενες γλώσσες προγραμματισμού | Οποιαδήποτε                                                       | Οποιαδήποτε                                                     | Java, Scala, γλώσσες JVM                                        | Java, Scala, Python, SQL                                          | Java, Scala, Python, R                                                                | Rust                                                                                                                                                                                                                                           |
| Batch Processing                        | όχι                                                               | όχι                                                             | όχι                                                             | ναι                                                               | ναι                                                                                   | ναι                                                                                                                                                                                                                                            |



# Εργαλεία Οπτικοποίησης και Αναλυτικής Δεδομένων

| Name                            | Tableau                                                         | TIBCO Spotfire                                                        | DataWrangler                                                                                                                                       | D3                                                                     | IBM SPSS Analytic Server                                                                                                                                                                                                                  | Apache Superset                                                                                  | Redash                                                                                                   | Metabase                                                                   | Sisense                                                                                                                 |
|---------------------------------|-----------------------------------------------------------------|-----------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| Description                     | Business Intelligence and Analytics                             | Data Visualization & Analytics Software                               | Wrangler is an interactive tool for data cleaning and transformation                                                                               | D3.js is a JavaScript library for manipulating documents based on data | IBM SPSS Analytic Server enables IBM SPSS Modeler to use big data as a source for predictive modeling. Together they can provide an integrated predictive analytics platform using data from Hadoop distributions and Spark applications. | Apache Superset (incubating) is a modern, enterprise-ready business intelligence web application | Connect and query your data sources, build dashboards to visualize data and share them with your company | Metabase is the easy, open source way to ask questions and learn from data | Sisense's BI software makes it easy to instantly reveal business insights from complex data – any data source, any size |
| Website                         | <a href="https://www.tableau.com/">https://www.tableau.com/</a> | <a href="https://spotfire.tibco.com/">https://spotfire.tibco.com/</a> | <a href="http://vis.stanford.edu/wrangler/">http://vis.stanford.edu/wrangler/</a>                                                                  | <a href="https://d3js.org/">https://d3js.org/</a>                      | <a href="https://www.ibm.com/us-en/marketplace/spss-analytic-server">https://www.ibm.com/us-en/marketplace/spss-analytic-server</a>                                                                                                       | <a href="https://superset.incubator.apache.org/">https://superset.incubator.apache.org/</a>      | <a href="https://redash.io/">https://redash.io/</a>                                                      | <a href="https://metabase.com/">https://metabase.com/</a>                  | <a href="https://www.sisense.com/">https://www.sisense.com/</a>                                                         |
| Initial release                 | 2003                                                            | 1996                                                                  | 2011                                                                                                                                               | 2011                                                                   | 2013                                                                                                                                                                                                                                      | 2016                                                                                             | 2013                                                                                                     | 2015                                                                       | 2013 (as Prism)                                                                                                         |
| Current release                 | 10.5.2 / March 2018                                             | 7.12 / March 2018                                                     | 2011                                                                                                                                               | 5.0.0 / January 2018                                                   | 3.1 / June 2017                                                                                                                                                                                                                           | 0.24.0 / March 2018                                                                              | 3.0.0 / December 2017                                                                                    | 0.28.4 / March 2018                                                        | 7.1 / April 2018                                                                                                        |
| License                         | Proprietary                                                     | Proprietary                                                           | Open Source, <a href="https://github.com/StanfordHCI/wrangler/blob/master/LICENSE">https://github.com/StanfordHCI/wrangler/blob/master/LICENSE</a> | BSD                                                                    | Proprietary                                                                                                                                                                                                                               | Apache License 2.0                                                                               | BSD 2-Clause                                                                                             | AGPL                                                                       | Proprietary                                                                                                             |
| Implementation language         | C++ (Desktop), Java/JavaScript/C++ (Server)                     | -                                                                     | JavaScript                                                                                                                                         | JavaScript                                                             | -                                                                                                                                                                                                                                         | JavaScript, Python                                                                               | Python, JavaScript                                                                                       | Clojure, JavaScript                                                        | -                                                                                                                       |
| APIs / other access methods     | -                                                               | URL API                                                               | -                                                                                                                                                  | -                                                                      | Hadoop                                                                                                                                                                                                                                    | REST                                                                                             | URL API                                                                                                  | REST                                                                       | REST, Hadoop, Hive                                                                                                      |
| Supported programming languages | Does not offer interface for programming a custom library       | C#, Java                                                              | Wrangle                                                                                                                                            | JavaScript                                                             | -                                                                                                                                                                                                                                         | -                                                                                                | Native query language of data source                                                                     | MetabaseQL (MBQL)                                                          | SQL-like language                                                                                                       |



# Βιβλιογραφική Επισκόπηση και Εργαλεία της Επιστήμης Δεδομένων

Τεχνική αναφορά «Αποτελέσματα Βιβλιογραφικής Επισκόπησης και Εξοικείωσης με τα Εργαλεία της Επιστήμης Δεδομένων»

- Τεχνολογική στάθμη των ερευνητικών πεδίων

- Για κάθε ερευνητικό πεδίο
  - Στόχοι και ερευνητική κατεύθυνση
  - Τεχνολογική στάθμη
  - Βιβλιογραφία

- Εργαλεία της επιστήμης δεδομένων

- Για κάθε κατηγορία εργαλείων
  - Γενικά
  - Περιγραφή εργαλείων
  - Σύγκριση εργαλείων
  - Συμπεράσματα

Συγκεντρωτικά στοιχεία

- 8 ερευνητικές κατευθύνσεις με 199 σχετικές αναφορές
- 9 κατηγορίες εργαλείων της επιστήμης δεδομένων που περιλαμβάνουν 70 εργαλεία της επιστήμης δεδομένων
- 9 συγκριτικοί πίνακες εργαλείων της επιστήμης δεδομένων



# Βιβλιογραφική Επισκόπηση και Εργαλεία της Επιστήμης Δεδομένων

## Τεχνική αναφορά

- 152 σελίδες
- Διαθέσιμη στην ερευνητική κοινότητα

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ «Ανταγωνιστικότητα Επιχειρηματικότητα και Καινοτομία»

ΑΣΟΝΑΣ ΠΡΟΤΕΡΑΙΟΤΗΤΑΣ 03 «Ανάπτυξη μηχανισμών στήριξης της επιχειρηματικότητας»

ΠΡΟΣΚΛΗΣΗ «ΔΡΑΣΗ ΣΤΡΑΤΗΓΙΚΗΣ ΑΝΑΠΤΥΞΗΣ ΕΡΕΥΝΗΤΙΚΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΩΝ ΦΟΡΕΩΝ»

ΠΡΑΣΗ «Υπολογιστικές Επιστήμες και Τεχνολογίες Δεδομένων, Περιχόμενων και Αλληλεπίδρασης», Κωδικός ΟΠΣ «5002437»

Υποέργο 3 «Από τα Μεγάλα Δεδομένα στην Επιστήμη Δεδομένων»

### ΤΙΤΛΟΣ ΠΑΡΑΔΟΤΕΟΥ

Π1.1: «Αποτελέσματα Βιβλιογραφικής Επισκόπησης και Εξοικείωσης με τα Εργαλεία της Επιστήμης Δεδομένων»

«Αθηνά» Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών και της Γνώσης

Ινστιτούτο Πληροφοριακών Συστημάτων

Αθήνα

Δεκέμβριος 2018





VIRTUAL EVENT

# Ημέρες Δράσης

23-24-25  
ΙΟΥΝΙΟΥ  
2021



Thank you!