ε.δε.μ²
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

infostrag

Προκλήσεις & Ευθύνες της Ακαδημαϊκής Κοινότητας μπροστά στην Επέλαση της Τεχνητής Νοημοσύνης

Βασίλης Μάγκλαρης

Ομότιμος Καθηγητής Σχολής Ηλεκτρολόγων Μηχ. & Μηχ. Υπολογιστών Ε.Μ.Π.

maglaris@netmode.ntua.gr www.netmode.ntua.gr

19ο Σεμινάριο της Ερμούπολης για την Κοινωνία της Πληροφορίας & την Οικονομία της Γνώσης

Συνεδριακή Αίθουσα Επιμελητηρίου Κυκλάδων
Παρασκευή 12/7/2024

Ορισμοί

Τεχνητή Νοημοσύνη (Artificial Intelligence - AI):
Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind

IBM: https://www.ibm.com/topics/artificial-intelligence

Μηχανική Μάθηση (Machine Learning - ML):
Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy

IBM: https://www.ibm.com/topics/machine-learning

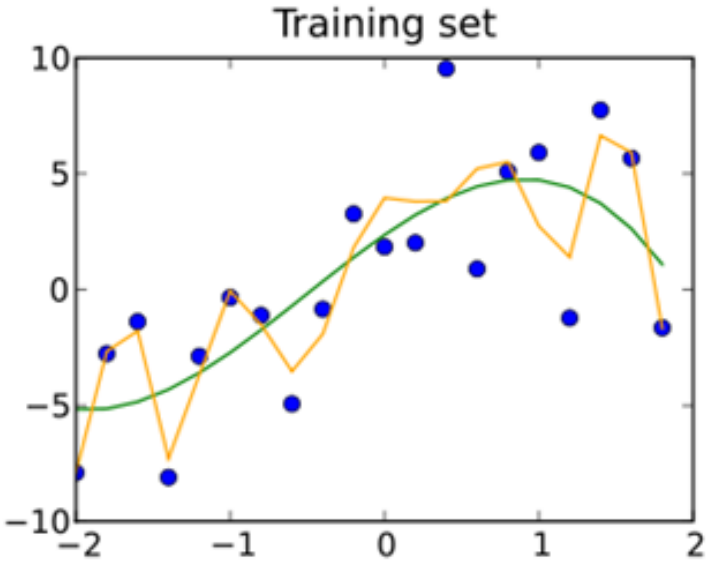
Διασχολικό Μεταπτυχιακό Πρόγραμμα Σπουδών Ε.Μ.Π.
Επιστήμη Δεδομένων – Μηχανική Μάθηση, Data Science – Machine Learning
https://dsml.ece.ntua.gr/

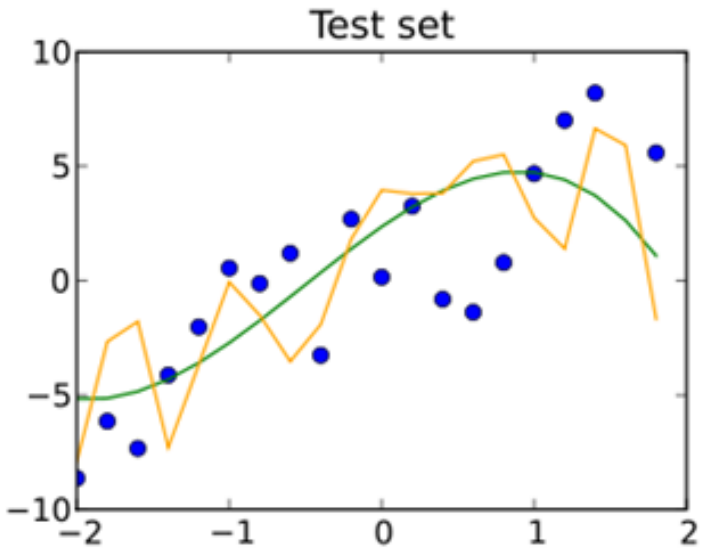
Λόγοι Ανάπτυξης Μηχανικής Μάθησης

- Η κατακλυσμιαία ανάπτυξη υπολογιστικών υποδομών αποθήκευσης και επεξεργασίας δεδομένων, επιτρέπει σήμερα την υλοποίηση αλγορίθμων στατιστικής ανάλυσης και στοχαστικής βελτιστοποίησης με βάση ιστορικά στοιχεία δείγματος μάθησης
- Η αλματώδης συσσώρευση τεράστιου όγκου πολυδιάστατων δεδομένων (*big data*) με πολλά χαρακτηριστικά, απαιτεί την ανάπτυξη ευφυών αλγορίθμων εξόρυξης εκτιμήσεων, προβλέψεων και ταξινόμησης νεοεμφανιζόμενων δειγματικών στοιχείων
- Η κατανόηση μεθόδων μάθησης σε βιολογικά συστήματα οδηγεί σε αλγορίθμους τεχνητής νοημοσύνης για συμπλήρωση και ελεγχόμενη πρόβλεψη (ή/και δημιουργία) δειγματικών στοιχείων, συμπεριλαμβανομένων ακολουθιών και χρονοσειρών (στοχαστικών διαδικασιών, *stochastic processes*) με βάση παρεμφερή στατιστικά χαρακτηριστικά αποθηκευμένου δείγματος μάθησης
- Η επεξεργασία, ταξινόμηση και αναζήτηση σε δημόσια & ιδιωτικά αρχεία, με την επιταχυνόμενη ψηφιοποίηση τους και τον αυτοματισμό εξαγωγής συνόψεων και μεταδεδομένων (*metadata*), συνδυάζεται με την διαθεσιμότητα προηγμένων τεχνολογιών (*Natural Language Processing* - *NLP*, *Optical Character Recognition* - *OCR*, *Big-data Analytics*, *Multimedia User Interfaces*….), ανοικτών εργαλείων & ευρυζωνικών υποδομών

Σύνολα Δεδομένων (Datasets)

https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets





Training Dataset (***μπλε σημεία μάθησης***)

- Λεπτομερής **κίτρινη** καμπύλη εκτίμησης με απόκλιση MSE=4
- Απλή **πράσινη** καμπύλη με απόκλιση MSE=9

Test Dataset (***μπλε σημεία γενίκευσης***)

- Απόκλιση από **κίτρινη** καμπύλη MSE=15 (από 4) ***OVERFITTING***
- Απόκλιση από **πράσινη** καμπύλη MSE=13 (από 9)

Μοντέλα Μηχανικής Μάθησης

Διακριτικά Μοντέλα (**Discriminative Models**):

Μέθοδοι ταξινόμησης (*classification*) ή εκτίμησης (παλινδρόμηση, *regression*) δειγματικών στοιχείων (*data elements*) μέσω υπό συνθήκη πιθανότητας (*conditional density*) εξόδου (*label*) βάσει χαρακτηριστικών (*features*) του, όπως αυτές προσεγγίστηκαν σε στοιχεία δείγματος μάθησης (*training sample*) για γενίκευση σε *test datasets* (*generalization*)

Ενδεικτικές Εφαρμογές:

- *Ταξινόμηση δειγματικών στοιχείων* με βάση συνάρτηση χαρακτηριστικών τους
- *Αναγνώριση προτύπων* με βάση κύρια χαρακτηριστικά τους (*pattern recognition*)
- *Εκτίμηση εξόδου* συμβατή με διαθέσιμα ζεύγη εισόδου - στόχου (*regression*)

Μοντέλα Μηχανικής Μάθησης

Διακριτικά Μοντέλα (**Discriminative Models**):
Μέθοδοι ταξινόμησης (*classification*) ή εκτίμησης (παλινδρόμηση, *regression*) δειγματικών στοιχείων (*data elements*) μέσω υπό συνθήκη πιθανότητας (*conditional density*) εξόδου (*label*) βάσει χαρακτηριστικών (*features*) του, όπως αυτές προσεγγίστηκαν σε στοιχεία δείγματος μάθησης (*training sample*) για γενίκευση σε *test datasets* (*generalization*)

Ενδεικτικές Εφαρμογές:

- *Ταξινόμηση δειγματικών στοιχείων* με βάση συνάρτηση χαρακτηριστικών τους
- *Αναγνώριση προτύπων* με βάση κύρια χαρακτηριστικά τους (*pattern recognition*)
- *Εκτίμηση εξόδου* συμβατή με διαθέσιμα ζεύγη εισόδου - στόχου (*regression*)

Παραγωγικά Μοντέλα (**Generative Models**):
Μέθοδοι εκτίμησης τρόπων παραγωγής (*generation*) δειγματικών στοιχείων, στατιστικά συμβατών με ιδιότητες του δείγματος μάθησης (*training sample*) μέσω συνδυασμένων πιθανοτήτων (*joint probabilities*) εξόδου (*output*) και χαρακτηριστικών (*features*) εισόδου, όπως υπολογίστηκαν στα στοιχεία του δείγματος μάθησης (*training ample elements*)

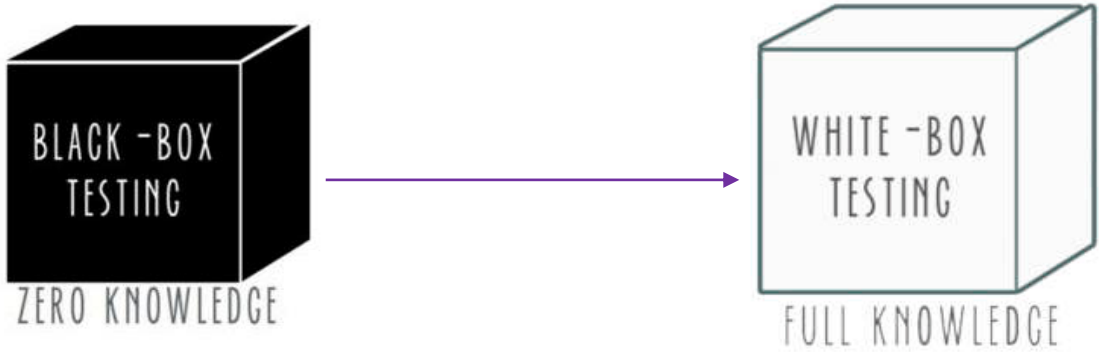
Ενδεικτικές Εφαρμογές:

- *Δημιουργία προσομοιωμένων στοιχείων*: κειμένων (συμβατών με αποδεκτά μοντέλα *Natural Language processing - NLP*), εικόνων, κινούμενων σχεδίων, ιδεατών τοπίων…
- *Εμπλουτισμός Μηχανών Αναζήτησης* (*Google*, *MS Bing* + *OpenAI Chat Generative Pre-trained Transformer - ChatGPT*)
- *Επικράτηση αληθοφανών εναλλακτικών εκτιμήσεων* σε συνέργεια με εργαλεία θεωρίας παιγνίων (*Generative Adversarial Networks - GAN*)

Εμπόδια για τη Διάδοση Τεχνολογιών Έξυπνων Συστημάτων Αποφάσεων

- Χαμηλό υπόβαθρο θεωρητικών θεμελίων: ***Μαθηματική Στατιστική*** & ***Αρχές Στατιστικής Μηχανικής***, Λογική, Διακριτά Μαθηματικά, Αναγνώριση Προτύπων, Φυσικές Γλώσσες, Διαχείριση Μεγάλων Βάσεων Δεδομένων, Θεωρία Αλγορίθμων…
- Ελλειμματική τεχνογνωσία χρηστών - δεν αρκεί ευκολία προγραμματισμού (***Python***)
- Δυσπιστία για αποτελεσματικότητα μεθόδων Τεχνητής Νοημοσύνης (***Artificial Intelligence*** - ***AI***) & αλγορίθμων Μηχανικής Μάθησης (***Machine Learning*** - ***ML***)
- Περιβαντολλογική επιβάρυνση
- Νομικά ζητήματα προστασίας ιδιωτικότητας προσωπικών δεδομένων (***GDPR*** κλπ.), ρυθμιστικές παρεμβάσεις (Κρατικές, Κοινοτικές, Αυτορρύθμιση…),
- Φόβοι οργανισμών οικονομικής, πολιτικής & κοινωνικής δραστηριότητας από έκθεση δεδομένων τους σε αντιπάλους, ***hackers***…
- Μεγάλος ανταγωνισμός εμπορικών συμφερόντων παραγωγής λογισμικού, δαιμονοποίηση ανοικτού λογισμικού και ανοικτών δεδομένων
- Υπερεκτίμηση δυνατοτήτων & κινδύνων αυτοματοποίησης διαδικασιών αποφάσεων με βάση τις (απεριόριστες;) προόδους Τεχνολογιών Πληροφορικής & Επικοινωνιών (***ICT***)
- Κραυγαλέες αποτυχίες μεθόδων ***ML***, ***overfitting***, παραισθήσεις (***hallucinations***) που προκύπτουν από πολυδιαφημισμένες παραγωγικές μεθόδους ***AI*** (***GenAI***, ***ChatGPT***)
- Ανάγκη Διαφάνειας & Επεξηγησιμότητας (***Explainability***) μεθόδων ***AI***

Αναγκαιότητα Μεθόδων eXplainable Artificial Intelligence (XAI)



- Η ακρίβεια (***accuracy***) δεν είναι αρκετή για την επιλογή μοντέλου ***Machine Learning*** (***ML***)
- Δυσπιστία χρηστών - σχεδιαστών - αναλυτών - ρυθμιστών πολιτικής σε μη ερμηνεύσιμα (***uninterpretable***) συστήματα αποφάσεων τύπου ***black-box***
- Αναγκαιότητα μεθόδων ***eXplainable Artificial Intelligence*** (***XAI***) ως προς τα κριτήρια λήψης αποφάσεων & ρύθμισης μοντέλου προς σχεδιαστές - χρήστες συστημάτων ***ML***
- Δικαιολόγηση αποφάσεων συστημάτων ***ML*** σε ερωτήσεις χρηστών - πελατών για προσωπικά ζητήματα που τους αφορούν (***local interpretations***)
- Ανάγκη για ***συγκριτική αξιολόγηση features*** και δικαιολόγηση σχεδιαστικών ***επιλογών*** μοντέλου, παραμέτρων/υπερπαραμέτρων κλπ.
- Ανάπτυξη εργαλείων αποτύπωσης σημασίας (***importance***) και συσχετίσεων (***correlations***) χαρακτηριστικών (***features***) σε γραφικό περιβάλλον, φιλικό προς τους χρήστες

Κίνδυνοι & Ανατροπές Εργασιακών Σχέσεων

- Παραπληροφόρηση, ψευδή/παραποιημένα στοιχεία
- Ανεξέλεγκτη επιβάρυνση ενεργειακού αποτυπώματος
- Παραβιάσεις ιδιωτικότητας προσωπικών δεδομένων
  - Αναγνώριση προσώπων, αποτύπωση συνηθειών, επίδραση σε αγοραστικές πρακτικές…
  - Ζητήματα ηθικής & αξιοπιστίας για καταγραφή - αποθήκευση - διάδοση πληροφορίων
- Κατακλυσμός πληροφοριών
  - Αδυναμία επεξεργασίας & ελέγχου χωρίς δυσνόητα (και κοστοβόρα;) υπολογιστικά εργαλεία → άνοιγμα ψηφιακού & ηλικιακού χάσματος (George Orwell 1984;)
- Ανατροπή εργασιακών σχέσεων, κατάργηση επαγγελμάτων
  - Υπερεξειδίκευση, απαξίωση δεξιοτήτων ταξινόμησης εγγράφων, περίληψης κειμένων, εξαγωγής μεταδεδομένων…
- Ημιμάθεια
  - Υπερτονισμός εμπειρίας προγραμματισμού χωρίς θεωρητικό υπόβαθρο (στατιστική, ανάλυση αλγορίθμων, διακριτά μαθηματικά...)
- Ορατοί κίνδυνοι για Ειρήνη, Δημοκρατία, Ισονομία, Ανθρώπινα Δικαιώματα…

Έχουν γνώση οι φύλακες ;;;

Προκλήσεις Ακαδημαϊκής Κοινότητας

- Εκπαίδευση στελεχών σε μεθόδους Τεχνητής Νοημοσύνης (ανθρωποκεντρικά με έμφαση σε κινδύνους & ρυθμιστικές παρεμβάσεις…)
- Έρευνα & Πρωτοπορειακές εφαρμογές μεθόδων *ML - AI* έξω από εμπορικές σκοπιμότητες
- Ώσμωση πολλαπλών γνωστικών περιοχών (*σχεδόν όλων*) με συνιστώσες αβεβαιότητας & απαιτήσεις ευρείας γνωσιακής εμπειρίας
  - Προαπαιτούμενα: ***Στατιστική***, Στοχαστική Βελτιστοποίηση, Γραμμική Άλγεβρα, Διακριτά Μαθηματικά, Λογική
  - Επιστήμες Μηχανικού & Υπολογιστών (Προγραμματιστικές Τεχνικές, Επεξεργασία Βάσεων Δεδομένων, Αλγόριθμοι & Πολυπλοκότητα, Φυσικές Γλώσσες, Επεξεργασία Εικόνων, Όραση Η/Υ, Αναγνώριση Προτύπων, Αυτόματος Έλεγχος & Ρομποτική, Θεωρία Παιγνίων…)
  - Βιοφυσική (Βιολογικά μαθησιακά μοντέλα, Φυσική → Τεχνητή Νοημοσύνη)
- Αξιολόγηση Ευφυών Οικονομικών/Χρηματιστηριακών Μοντέλων
- Μελέτη Νομικού/Ρυθμιστικού Πλαισίου
- Κυβερνοασφάλεια (π.χ. συνεργατική αντιμετώπιση επιθέσεων βάσει καταγραμμένης εμπειρίας και (;) συντονιστή Τρίτη Έμπιστη Οντότητα)
- Ανάπτυξη εργαλείων εξατομικευμένης μάθησης & αξιολόγησης Μαθητών/Σπουδαστών…
- Εξειδίκευση Καινοτόμων Μηχανών Αναζήτησης περιεχόμενου & εφαρμογών ψυχαγωγίας
- …

Ευθύνες Ακαδημαϊκής Κοινότητας

- Επισήμανση δυνατοτήτων *ML - AI* αλλά και προβληματισμοί σχετικά με αδυναμίες τους
- Προστασία ατομικών δικαιωμάτων, καταγγελία παράνομων – ανήθικων εφαρμογών, αποφυγή συμμετοχής σε πολεμικά παιχνίδια (εμπορικά, κρατικά…)
- Μελέτη περιβαντολλογικών διαστάσεων, ενεργειακού αποτυπώματος – *data centers*…
- Συμμετοχή σε διαμόρφωση ρυθμιστικού περιβάλλοντος για προστασία προσωπικών δεδομένων και ελεύθερης πρόσβασης σε (ανώνυμα) αρχεία περιπτώσεων
- Οι μέθοδοι *ML – AI* βασίζονται σε στατιστικά μοντέλα με υπαρκτές πιθανότητες αστοχιών:
  - Τα αποτελέσματα απαιτούν συνεχείς παρεμβάσεις & ελέγχους για επαλήθευση από ανθρώπινη νοημοσύνη
  - Οι αυτοματισμοί δεν υποκαθιστούν την εμπειρία και τεράστιες δυνατότητες επεξεργασίας του ανθρώπινου νου (για το προβλεπτό μέλλον!)
- Αλλά μη πέσουμε στο άκρο του Λουδισμού (*Luddism*) της Βιομηχανικής Επανάστασης!!!

Ευθύνες Ακαδημαϊκής Κοινότητας

- Επισήμανση δυνατοτήτων *ML - AI* αλλά και προβληματισμοί σχετικά με αδυναμίες τους
- Προστασία ατομικών δικαιωμάτων, καταγγελία παράνομων – ανήθικων εφαρμογών, αποφυγή συμμετοχής σε πολεμικά παιχνίδια (εμπορικά, κρατικά…)
- Μελέτη περιβαντολλογικών διαστάσεων, ενεργειακού αποτυπώματος – *data centers*…
- Συμμετοχή σε διαμόρφωση ρυθμιστικού περιβάλλοντος για προστασία προσωπικών δεδομένων και ελεύθερης πρόσβασης σε (ανώνυμα) αρχεία περιπτώσεων
- Οι μέθοδοι *ML – AI* βασίζονται σε στατιστικά μοντέλα με υπαρκτές πιθανότητες αστοχιών:
  - Τα αποτελέσματα απαιτούν συνεχείς παρεμβάσεις & ελέγχους για επαλήθευση από ανθρώπινη νοημοσύνη
  - Οι αυτοματισμοί δεν υποκαθιστούν την εμπειρία και τεράστιες δυνατότητες επεξεργασίας του ανθρώπινου νου (για το προβλεπτό μέλλον!)
- Αλλά μη πέσουμε στο άκρο του Λουδισμού (*Luddism*) της Βιομηχανικής Επανάστασης!!!

Διαχρονικά τα Πανεπιστήμια ήταν χώροι επιστημονικής προόδου & ανοικτής καινοτομίας, χωρίς περιορισμούς εμπορικών συμφερόντων, καθώς και φορείς κριτικής μετάδοσης γνώσης στις επερχόμενες γενιές

Ας το κρατήσουμε και στην εποχή της Τεχνητής Νοημοσύνης

Οι Κίνδυνοι της Ανεξέλεγκτης Ανάπτυξης Τεχνητής Νοημοσύνης - AI

Geoffrey Hinton: Βρετανό-Καναδός, γεννήθηκε στο Λονδίνο το 1947

- 1977: Διδακτορικό, University of Edinburgh, Σκωτία
- Ακαδημαϊκή καριέρα σε Μ. Βρετανία, ΗΠΑ, Καναδά
- Ερευνητής Νευρωνικών Δικτύων (Μηχανές Boltzmann, Deep Belief Networks, Generative AI…)
- 2018: Turing Award
- 2013-2023: Σύμβουλος της Google & Καθηγητής στο University of Toronto
- Μάϊος 2023: Παραιτήθηκε από την Google για να αποδεσμεύεται στη διατύπωση επερχόμενων κινδύνων από την ανεξέλεγκτη ανάπτυξη της AI

Οι Κίνδυνοι της Ανεξέλεγκτης Ανάπτυξης Τεχνητής Νοημοσύνης - AI

Geoffrey Hinton: Βρετανό-Καναδός, γεννήθηκε στο Λονδίνο το 1947

- 1977: Διδακτορικό, University of Edinburgh, Σκωτία
- Ακαδημαϊκή καριέρα σε Μ. Βρετανία, ΗΠΑ, Καναδά
- Ερευνητής Νευρωνικών Δικτύων (Μηχανές Boltzmann, Deep Belief Networks, Generative AI…)
- 2018: Turing Award
- 2013-2023: Σύμβουλος της Google & Καθηγητής στο University of Toronto
- Μάϊος 2023: Παραιτήθηκε από την Google για να αποδεσμεύεται στη διατύπωση επερχόμενων κινδύνων από την ανεξέλεγκτη ανάπτυξη της AI

In early May 2023, Hinton revealed in an interview with BBC that AI might soon surpass the information capacity of the human brain. He described some of the risks posed by these chatbots as "quite scary". Hinton explained that chatbots have the ability to learn independently and share knowledge. This means that whenever one copy acquires new information, it is automatically disseminated to the entire group. This allows AI chatbots to have the capability to accumulate knowledge far beyond the capacity of any individual.
https://en.wikipedia.org/wiki/Geoffrey_Hinton

"The Godfather of A.I." Leaves Google and Warns of Danger Ahead:
Generative A.I. can already be a tool for misinformation. Soon, it could be a risk to jobs. Somewhere down the line, tech's biggest worriers say, it could be a risk to humanity
https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

Πρόσφατες Σχετικές Δημοσιεύσεις Ερευνητικής Ομάδας

NetMan @ netmode.ece.ntua.gr

- "**Orchestrating DDoS Mitigation via Blockchain-based Network Provider Collaborations**", **Adam Pavlidis**, **Marinos Dimolianis**, **Kostas Giotis**, **Loukas Anagnostou**, **Nikolaos Kostopoulos**, **Theocharis Tsigkritis**, **Ilias Kotivas**, **Dimitrios Kalogeras** & **Vasilis Maglaris**, *The Knowledge Engineering Review*, Volume 35, pp. 1-17, April 2020

- "**Signature-Based Traffic Classification and Mitigation for DDoS Attacks Using Programmable Network Data Planes**", **Marinos Dimolianis**, **Adam Pavlidis** & **Vasilis Maglaris**, *IEEE Access*, Volume 9, 2021, pp. 113061-113076

- "**DDoS Attack Detection via Privacy-aware Federated Learning in Multi-domain Cyber Infrastructures**", **Marinos Dimolianis**, **Dimitrios Kalogeras**, **Nikos Kostopoulos** & **Vasilis Maglaris**, *11th IEEE International Conference on Cloud Networking* (CloudNet), Paris, France, November 2022, pp. 118-125

- "**SHAP Interpretations of Tree and Neural Network DNS Classifiers for Analyzing DGA Family Characteristics**", **Nikos Kostopoulos**, **Dimitris Kalogeras**, **Dimitris Pantazatos**, **Mary Grammatikou** & **Vasilis Maglaris**, *IEEE Access*, Volume 11, 2023, pp. 61144-61160

- **"Enhancing Soft Skills in Network Management Education: A Study on the Impact of GenAI-Based Virtual Assistants"**, **Dimitris Pantazatos**, **Mary Grammatikou** & **Vasilis Maglaris**, *IEEE Global Engineering Education Conference* (EDUCON), Kos Island, Greece, May 2024, pp. 1-5