

Τα Big Data στο χώρο της υγείας

Σκοπός

Σκοπός της παρούσας εργαστηριακής άσκησης είναι η εισαγωγή του αναγνώστη στο χώρο των δεδομένων μεγάλης κλίμακας (Big Data), η εξοικείωσή του με τη σχετική ορολογία και η κατανόηση από τον αναγνώστη της σύνδεσης του χώρου αυτού με το χώρο της υγείας. Στην άσκηση δίνεται έμφαση στις πολλαπλές και ποικίλες πηγές των Big Data, γίνεται αναφορά στους τρόπους και στις τεχνικές με τις οποίες αναλύονται τα δεδομένα αυτά, καθώς και στα οφέλη που απορρέουν από την αξιοποίησή τους.

Για την εξοικείωση των φοιτητών με την επεξεργασία και την ανάλυση συνόλων δεδομένων, θα γίνει χρήση της προγραμματιστικής γλώσσας R και του ολοκληρωμένου περιβάλλοντος ανάπτυξης RStudio. Μέσα από την ενασχόλησή τους με το εργαστηριακό μέρος, οι φοιτητές θα εξοικειωθούν με τη γλώσσα R. Πιο αναλυτικά, θα είναι σε θέση να διακρίνουν τους διαφορετικούς τύπους μεταβλητών που υποστηρίζει η R, να «φορτώνουν», να επεξεργάζονται και να αναλύουν ποικίλα σύνολα δεδομένων με σκοπό την εξαγωγή χρήσιμων πληροφοριών.

Προσδοκώμενα Αποτελέσματα

Με την ολοκλήρωση της εργαστηριακής άσκησης οι φοιτητές θα είναι σε θέση:

- Να κατανοήσουν την έννοια των Big Data και των χαρακτηριστικών τους.
- Να γνωρίζουν τις σχετιζόμενες τεχνολογίες για αποθήκευση/επεξεργασία των Big Data.
- Να γνωρίζουν τις βασικές πηγές των Big Data στον χώρο της υγείας.
- Να διακρίνουν τα οφέλη που απορρέουν από την αξιοποίηση των Big Data στο χώρο της υγείας.
- Να κατανοήσουν τις βασικές αρχές και τεχνικές ανάλυσης των Big Data.
- Να αναγνωρίζουν τα ζητήματα ασφάλειας, ιδιωτικότητας και ηθικής που εγείρονται από τη χρήση των Big Data.
- Να κατέχουν μία βασική εξοικείωση με τη χρήση του λογισμικού R.

Λέξεις Κλειδιά

Big Data, Υγεία, 5 Vs, Cloud Computing, Ανάλυση Δεδομένων, SQL/NoSQL, Map Reduce, Αξιοποίηση Δεδομένων, Εξόρυξη Δεδομένων

Πίνακας Συντομογραφιών

Πλήρης Όρος	Συντομογραφία
Internet of Things	IoT
Software-as-a-Service	SaaS
Platform-as-a-Service	PaaS
Infrastructure-as-a-Service	IaaS
Ηλεκτρονική Διακυβέρνηση Κοινωνικής Ασφάλισης	Η.ΔΙ.Κ.Α
Εθνικός Οργανισμός Φαρμάκων	Ε.Ο.Φ
Εθνικός Οργανισμός Δημόσιας Υγείας	Ε.Ο.Δ.Υ
Ηλεκτρονικός Φάκελος Υγείας	Η.Φ.Υ.
Ελληνική Στατιστική Αρχή	ΕΛ.ΣΤΑΤ
Εθνικού Οργανισμού Παροχής Υπηρεσιών Υγείας	Ε.Ο.Π.Υ.Υ.

1. Εισαγωγή

Η παρούσα εργαστηριακή άσκηση αποτελείται από δύο διακριτά μέρη: το θεωρητικό και το πρακτικό. Στο θεωρητικό μέρος οι φοιτητές εισάγονται στην έννοια των Big Data, στα ιδιαίτερα χαρακτηριστικά τους και στις τεχνολογίες που σχετίζονται με αυτά. Στο μέρος αυτό δίνεται ιδιαίτερη έμφαση στα Big Data σχετιζόμενα με το χώρο της υγείας, τις διάφορες πηγές τους και στα πλεονεκτήματα που απορρέουν από τη χρήση τους.

Θεωρητικό μέρος

Στο θεωρητικό μέρος, εισάγεται αρχικά η έννοια των δεδομένων μεγάλης κλίμακας, η βασική ορολογία και τα κύρια χαρακτηριστικά τους. Στη συνέχεια, παρουσιάζονται η σχέση τους με τα παραδοσιακά δεδομένα και οι κύριες πηγές προέλευσής τους. Αναλυτική αναφορά γίνεται στα δεδομένα στο χώρο της υγείας και στις πολλές και ετερογενείς οντότητες που τα παράγουν. Ακολούθως, γίνεται λόγος σχετικά με βασικές τεχνικές για τη συλλογή, καταγραφή, καθαρισμό δεδομένων, αναπαράσταση, επεξεργασία ερωτημάτων, μοντελοποίηση δεδομένων, ανάλυση, ερμηνεία και εξαγωγή πληροφοριών. Παρουσιάζονται ακόμη τεχνολογίες που έχουν αναπτυχθεί για την επεξεργασία και ανάλυση των δεδομένων μεγάλης κλίμακας, με έμφαση στις τεχνολογίες NoSQL και MapReduce. Τέλος, σημειώνονται τα πλεονεκτήματα που απορρέουν από την αξιοποίησή τους, αλλά και οι περιορισμοί που υφίσταται.

Πρακτικό μέρος

Στο πρακτικό μέρος της εργαστηριακής άσκησης οι φοιτητές εξερευνούν και εξοικειώνονται με τις δυνατότητες και τις λειτουργίες που παρέχει η γλώσσα προγραμματισμού R μέσω του ολοκληρωμένου περιβάλλοντος ανάπτυξης RStudio. Πιο αναλυτικά παρουσιάζονται στους φοιτητές οι βασικοί τύποι δεδομένων της R αλλά και μερικά βασικά παραδείγματα εντολών. Στη συνέχεια, παρουσιάζεται η φόρτωση ενός συνόλου δεδομένων και τα διάφορα βήματα για την επεξεργασία και ανάλυση των δεδομένων αυτών, με σκοπό να εξαχθούν χρήσιμα μεγέθη και πληροφορίες.

2. Θεωρητικό μέρος

2.1 Η έννοια των Big Data

Τα «Δεδομένα Μεγάλης Κλίμακας» (ή ευρέως γνωστά ως **Big Data**) αποτελούν όρο που χρησιμοποιείται τόσο στο ευρύτερο πεδίο της πληροφορικής όσο και στην επιχειρηματική κοινότητα. Τα **Big Data** εκφράζουν και συνοψίζουν την τάση που βρίσκεται σε εξέλιξη τα τελευταία χρόνια και η οποία δεν είναι άλλη από την εκρηκτική παραγωγή δεδομένων από πλήθος πηγών, τόσο μέσω του διαδικτύου, όσο και μέσω συσκευών που συνδέονται σε αυτό. Η νέα αυτή πραγματικότητα αποτελεί πρόκληση τόσο για τα υπολογιστικά συστήματα και τις εφαρμογές, όσο και τους ανθρώπους που καλούνται να ασχοληθούν με αυτά.

Παρά όμως το γεγονός ότι ο όρος είναι ευρέως διαδεδομένος, δεν υπάρχει κάποιος αυστηρός και κοινά αποδεκτός ορισμός ο οποίος να καλύπτει πλήρως το πνεύμα και τα χαρακτηριστικά τους. Περιεκτικά μπορεί να δοθεί ο εξής ορισμός:

«Τα δεδομένα μεγάλης κλίμακας είναι σύνολα δεδομένων των οποίων η συλλογή, αποθήκευση, διαχείριση και ανάλυση απαιτεί νέες αρχιτεκτονικές, τεχνικές, αλγορίθμους και εφαρμογές, λόγω αυξημένου όγκου ή σημαντικής πολυπλοκότητας που εμπεριέχουν» [1].

Αυτό είναι το βασικό στοιχείο που προσδιορίζει ένα σύνολο δεδομένων ως δεδομένα μεγάλης κλίμακας, δηλαδή **η αδυναμία των παραδοσιακών συστημάτων να τα διαχειριστούν με τρόπο εύχρηστο και επαρκή.**

Ένας άλλος «ορισμός» ο οποίος αξίζει να αναφερθεί, εκφράστηκε από τον George Dyson, και αναφέρει το εξής:

«Big Data ονομάστηκε το φαινόμενο που συνέβη, όταν το κόστος αποθήκευσης των δεδομένων έγινε μικρότερο από το κόστος στην περίπτωση που τα δεδομένα πεταχτούν ή καταστραφούν» [2].

Από την προηγούμενη δήλωση γίνεται εμφανής ο καθοριστικός ρόλος που μπορεί να διαδραματίσουν τα Big Data στην εξαγωγή χρήσιμης πληροφορίας και άρα στη λήψη μιας απόφασης.

1.2 Τα χαρακτηριστικά των Big Data

Ένα σημείο που αξίζει να σημειωθεί είναι ότι ο όρος Big Data δεν έχει σχέση αποκλειστικά με τον **όγκο** των δεδομένων (**Volume**), αν και σαφώς αποτελεί **βασικό χαρακτηριστικό τους**. Δε σημαίνει δηλαδή ότι μια βάση δεδομένων πολύ μεγάλου μεγέθους, αλλά με απλά σχετικά δεδομένα, χωρίς πολύπλοκες συσχετίσεις, εμπίπτει κατ' ανάγκη στο πεδίο. Θα πρέπει να συνυπάρχουν **επιπλέον χαρακτηριστικά**, τα οποία συνολικά είναι:

- Ο μεγάλος **όγκος** δεδομένων (**Volume**)
- Η αυξημένη **ταχύτητα** παραγωγής και κυκλοφορίας των δεδομένων (**Velocity**)
- Η **ποικιλία** τύπων και μορφών (**Variety**), δηλαδή ο μικρός βαθμός δόμησης των δεδομένων και η μη ενιαία μορφή τους
- Η αυξημένη **ακρίβεια** – εγκυρότητα (**Veracity**), δηλαδή η μείωση του βαθμού ασάφειας που παρουσιάζουν εγγενώς τα δεδομένα αυτά

Τα παραπάνω χαρακτηριστικά έχουν καθιερωθεί ως τα τέσσερα βασικά χαρακτηριστικά των Big Data (**4 Vs**). Ένα χαρακτηριστικό το οποίο έχει προστεθεί εκ των υστέρων, αλλά αποδεικνύεται εξαιρετικής σπουδαιότητας, είναι αυτό της **αξίας** (**Value**). Η πρόσβαση στα Big Data είναι ουσιαστικά «άχρηστη» αν από αυτή δεν προκύπτει κάποια χρησιμότητα – αξία για τον αναλυτή. Τα προηγούμενα χαρακτηριστικά αποτελούν τα **πέντε βασικά χαρακτηριστικά των Big Data (5 Vs)** έτσι όπως έχουν επικρατήσει μέχρι σήμερα και απεικονίζονται στην Εικόνα 1.

Τέλος, αξίζει να αναφερθεί πως, σε ορισμένες βιβλιογραφικές πηγές, συναντάται ένα επιπλέον χαρακτηριστικό, η **μεταβλητότητα** (**Variability**) των Big Data, η οποία εκφράζει ουσιαστικά τις συνεχείς μεταβολές των δεδομένων που λαμβάνονται και είναι πιθανόν να δυσχεράνουν την ανάλυση και την αξιοποίησή τους.



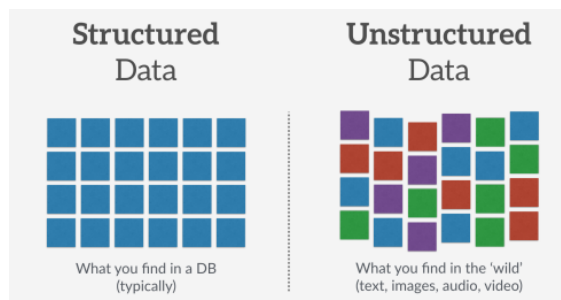
Εικόνα 1: Τα πέντε βασικά χαρακτηριστικά των Big Data (5 Vs)

1.3 Τα Big Data σε σύγκριση με τα «παραδοσιακά» δεδομένα

Τα Big Data διαφέρουν από τα «παραδοσιακά» δεδομένα σε αρκετά σημεία. Όπως αναφέρθηκε προηγουμένως, εκτός από τον όγκο, που αποτελεί βασικό στοιχείο, θα πρέπει να συνυπάρχουν και επιπλέον χαρακτηριστικά (5V's – Volume, Velocity, Variety, Veracity, Value), τα οποία αναλύονται στον πίνακα που ακολουθεί.

	Big Data	Παραδοσιακά δεδομένα
Όγκος	Ο όγκος αναφέρεται στη συσσώρευση υψηλού όγκου δεδομένων με μεγάλο ρυθμό σε μικρά χρονικά διαστήματα. Για παράδειγμα οι πληροφορίες που καταγράφονται και διακινούνται σε κοινωνικά δίκτυα, η καταγραφή των δεδομένων επισκέψεων σε ιστοσελίδες, τα στοιχεία συναλλαγών σε ηλεκτρονικές αγορές, τα στοιχεία συναλλαγών σε χρηματιστηριακές αγορές αποτελούν δείγματα δεδομένων με αυτό το χαρακτηριστικό.	Προφανώς δεν έχει οριστεί κάποιο όριο πέρα από το οποίο τα παραδοσιακά δεδομένα να παύουν να αντιμετωπίζονται ως τέτοια. Γενικά, μπορούμε να αναφέρουμε ότι οι τάξεις μεγέθους από Terabytes και πάνω αναφέρονται σε σχετικά μεγάλους όγκους δεδομένων. Το χαρακτηριστικό αυτό, του τεράστιου όγκου, προκαλεί αδυναμία διαχείρισης από τα παραδοσιακά συστήματα αποθήκευσης και ανάλυσης δεδομένων.
Ποικιλία	Τα Big Data παρουσιάζουν μεγάλη απόκλιση ως προς τον βαθμό δόμησης (unstructured data) και έχουν ποικιλία τύπων, όπως κείμενα, εικόνες, βίντεο, ομιλία που δεν είναι άμεσα και εύκολα διαχειρίσιμα από σχεσιακές βάσεις δεδομένων, όπως παρουσιάζεται στην Εικόνα 2.	Τα παραδοσιακά δεδομένα έχουν κατά κύριο λόγο δομημένη μορφή (structured data) και ανήκουν σε μεγάλο βαθμό σε συγκεκριμένους τύπους, όπως κείμενο ή αριθμητικές τιμές και μπορούν με ευκολία να ενταχθούν σε σχεσιακές βάσεις δεδομένων.
Ταχύτητα	Η ταχύτητα αναφέρεται τόσο στη δημιουργία όσο και κυκλοφορία των δεδομένων, η οποία για τα Big Data είναι αρκετά αυξημένη. Για παράδειγμα, τα δεδομένα που προέρχονται από κοινωνικά δίκτυα ή μηχανές αναζήτησης παράγονται με πολύ υψηλή ταχύτητα καθώς χιλιάδες χρήστες ενεργούν ταυτόχρονα, και κυκλοφορούν επίσης με υψηλή ταχύτητα μεταξύ των χρηστών. Οι ανάγκες δε αρκετές φορές απαιτούν να αναλύονται κατά τη στιγμή της γέννησής τους και πριν αποθηκευτούν.	Η δημιουργία νέων πηγών δεδομένων, που συνεχώς αυξάνεται, είναι αναμενόμενο πως έχει συμβάλει στην παραγωγή ταχύτερων ροών δεδομένων, που δεν υφίστατο στο παρελθόν.

Τα δύο επόμενα στοιχεία, δηλαδή η αξία και η αξιοπιστία έχουν σχέση με την χρήση των δεδομένων και όχι τόσο με τα χαρακτηριστικά τους.		
Αξία	Η αξία αναφέρεται στο γεγονός ότι δεν υπάρχει αξία στα δεδομένα εξ ορισμού, εκτός και αν αυτά αναλυθούν και αξιοποιηθούν για τη λήψη αποφάσεων ή δημιουργία ανταγωνιστικού πλεονεκτήματος. Το σίγουρο είναι όμως πως αποτελεί επιθυμητό στοιχείο αλλά δεν είναι δεδομένο.	Όσο περισσότερα δεδομένα συλλέγει κανείς, τόσο ακριβέστερα αποτελέσματα αναμένεται να έχει. Η αξία των παραδοσιακών δεδομένων, λοιπόν, θα μπορούσε να θεωρηθεί συνολικά μάλλον μικρότερη από αυτή των Big Data αλλά σαφώς πιο δεδομένη.
Αξιοπιστία	Η ασάφεια (έλλειψη αξιοπιστίας) προκύπτει από την ίδια τη φύση των Big Data, καθώς ο υψηλός βαθμός αδόμητης πληροφορίας περιέχει ασάφεια και ατέλεια. Το χαρακτηριστικό αυτό αποτελεί κατά ένα βαθμό τροχοπέδη στην ευρεία χρήση τους για τη λήψη αποφάσεων, καθώς ο βαθμός εμπιστοσύνης σε αυτά δεν είναι ακόμη αρκετά υψηλός, και ο κάθε αναλυτής καλείται να το αντιμετωπίσει.	Αυτό δεν παρατηρείται στα παραδοσιακά δεδομένα καθώς γενικά ο τρόπος παραγωγής και ανάλυσης εγγυόταν την αξιοπιστία των δεδομένων.



Εικόνα 2: Σύγκριση δομημένων και μη δομημένων δεδομένων

Εκτός από τα παραπάνω χαρακτηριστικά, υπάρχουν **διαφορές και στις πηγές των δεδομένων** των δύο τύπων. Καταρχήν, **τα Big Data παράγονται συχνά από μηχανές με κάποιο αυτοματοποιημένο τρόπο**. Στις παραδοσιακές πηγές δεδομένων, αντίθετα, υπάρχει πάντα ένα πρόσωπο που συμμετέχει στη διαδικασία. Για παράδειγμα, οι παραδοσιακές λιανικές συναλλαγές ή οι συναλλαγές με μια τράπεζα απαιτούν κάποιο πρόσωπο που εκτελεί μια εργασία ως τμήμα της δημιουργίας μιας εγγραφής σε ένα αρχείο δεδομένων. Για τα Big Data αυτό δεν ισχύει, καθώς σε πολλές περιπτώσεις πολλές πηγές μεγάλων δεδομένων παράγονται χωρίς καμία ανθρώπινη αλληλεπίδραση. Ένας αισθητήρας για παράδειγμα, στέλνει τα δεδομένα σχετικά με το περιβάλλον του σε πραγματικό χρόνο, ακόμη και αν κανείς δεν τον αγγίζει, ή δεν του ζητά τα δεδομένα.

Δεύτερον, **τα Big Data είναι τυπικά μια εντελώς νέα πηγή δεδομένων**. Δεν είναι απλά μια εκτεταμένη συλλογή υφιστάμενων παραδοσιακών δεδομένων. Για παράδειγμα, με τη χρήση του διαδικτύου, οι πελάτες μπορούν να εκτελέσουν συναλλαγές με τράπεζες ή ηλεκτρονικά καταστήματα σε πραγματικό χρόνο. Οι εργασίες που εκτελούν δε διαφέρουν ως συναλλαγές από τις παραδοσιακές, εκτός του διαφορετικού καναλιού. Ένας οργανισμός μπορεί να συλλέξει αυτές τις συναλλαγές από το διαδίκτυο, που δε διαφέρουν από τα παραδοσιακά δεδομένα, ωστόσο, η πλοήγηση των χρηστών, ή οι νέοι τύποι δεδομένων, δημιουργούν ειδικές συμπεριφορές από τους πελάτες που με τη σειρά τους δημιουργούν ριζικά νέα δεδομένα.

Τρίτον, **πολλές πηγές Big Data δεν έχουν σχεδιαστεί, ώστε να είναι φιλικές και ορισμένες δεν είναι καν προσχεδιασμένες**. Για παράδειγμα, στα κείμενα που συλλέγονται από ιστοχώρους κοινωνικής δικτύωσης δεν υπάρχει κανένας εύκολος τρόπος να επιβληθεί στους χρήστες να ακολουθούν κανόνες γραμματικής ή σωστή σύνταξη ή καθορισμένο λεξιλόγιο. Επομένως, αυτά τα δεδομένα περιέχουν οτιδήποτε καταχωρείται και μπορεί να είναι δύσκολο να εργαστεί κανείς με αυτά. Σε αντίθεση, οι περισσότερες παραδοσιακές πηγές έχουν σχεδιαστεί εκ των προτέρων για να είναι φιλικές και σαφείς. Τα συστήματα που χρησιμοποιούνται για να συλλέξουν στοιχεία συναλλαγών, για παράδειγμα, παρέχουν ειδικές φόρμες για συμπλήρωση από τον χρήστη και τα δεδομένα που συλλέγονται είναι εύκολο να αναλυθούν και να χρησιμοποιηθούν.

Τέλος, **μεγάλα τμήματα ρευμάτων δεδομένων μπορεί να έχουν μικρή αξία ή μπορεί και να είναι άνευ αξίας**. Για παράδειγμα, στα ιστολόγια (blogs) υπάρχουν πληροφορίες που είναι υψηλής αξίας, αλλά υπάρχει επίσης πλήθος πληροφοριών που δεν έχουν αξία. Αντίθετα, οι παραδοσιακές πηγές δεδομένων καθορίζονται εκ των προτέρων, ώστε να είναι 100% σχετικές με τις απαιτήσεις δεδομένων, λόγω των περιορισμών σε επεκτασιμότητα, καθώς υπάρχει κόστος διαχείρισης, αν περιλαμβάνεται σε μία ροή δεδομένων κάτι που δεν είναι σχετικό ή χρήσιμο. Τα παραπάνω δεν ισχύουν υποχρεωτικά για κάθε πηγή Big Data αλλά οι περισσότερες πηγές χαρακτηρίζονται από ορισμένα από αυτά.

1.4 Οι πηγές των Big Data

Οι κύριες πηγές παραγωγής δεδομένων στις μέρες μας είναι οι **μηχανές**, οι **άνθρωποι** και οι **οργανισμοί/διαδικασίες** (Εικόνα 3) [3].



Εικόνα 3: Οι κύριες πηγές παραγωγής Big Data

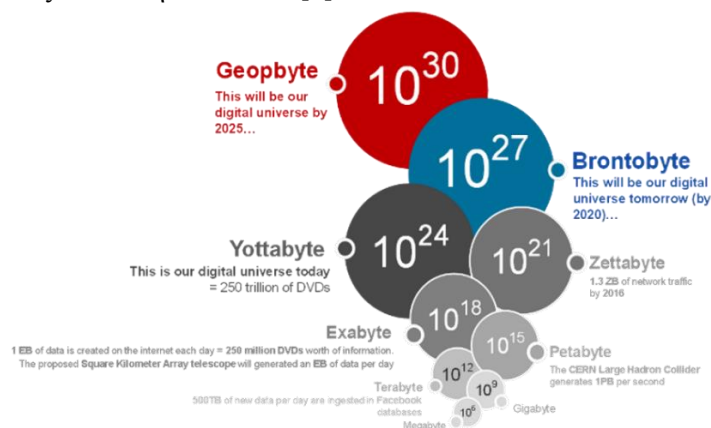
- Με τον όρο «**δεδομένα παραγόμενα από μηχανές**» (machine-generated data) εννοούμε τα δεδομένα που δημιουργούνται, σε πραγματικό χρόνο, χωρίς την ανθρώπινη παρέμβαση, από βιομηχανικό εξοπλισμό σε εργοστάσια ή από αισθητήρες, οι οποίοι μπορεί να είναι περιβαλλοντικοί, βιομηχανικοί, οχημάτων, ιατρικοί κ.α. Επίσης, τα δεδομένα που δημιουργούνται και διαμοιράζονται όταν π.χ. "έξυπνες" οικιακές συσκευές επικοινωνούν μεταξύ τους ή με τους οικιακούς διακομιστές τους, εμπίπτουν στην ίδια κατηγορία. Δημιουργείται έτσι μια άμεση σύνδεση μεταξύ των Big Data και του **Δικτύου των Πραγμάτων** (ή κοινώς **Internet of Things – IoT**) (Εικόνα 4). **Η έννοια του IoT ουσιαστικά ενοποιεί ένα ευρύ φάσμα «πραγμάτων» και τα μετατρέπει σε «έξυπνα» αντικείμενα** (οτιδήποτε από ρολόγια έως ψυγεία, αυτοκίνητα έως σιδηροδρομικές γραμμές κ.α.). Οι συσκευές, που κατά βάση δεν είχαν τη δυνατότητα σύνδεσης στο διαδίκτυο, θα είναι πλέον σε θέση να αποκτήσουν και να επεξεργαστούν δεδομένα, καθώς θα είναι εξοπλισμένα με αισθητήρες και τσιπ υπολογιστών με σκοπό τη συλλογή και ανταλλαγή δεδομένων. Το IoT είναι ουσιαστικά ο τρόπος μέσω του οποίου θα συλλέγονται και θα αποστέλλονται δεδομένα από όλες αυτές τις διασυνδεδεμένες συσκευές. Το IoT φαίνεται πως θα εμφανιστεί σύντομα σε πολλές πτυχές της ζωής μας: στη μεταφορά (π.χ. αυτοκίνητα, έξυπνες σιδηροδρομικές γραμμές και φανάρια), στον

κατασκευαστικό τομέα (π.χ. έξυπνα κτίρια/σπίτια) και, φυσικά, στα καταναλωτικά αγαθά (π.χ. smartphones, φορητές συσκευές, wearables).



Εικόνα 4: Οι διάφορες πηγές του Internet of Things (IoT)

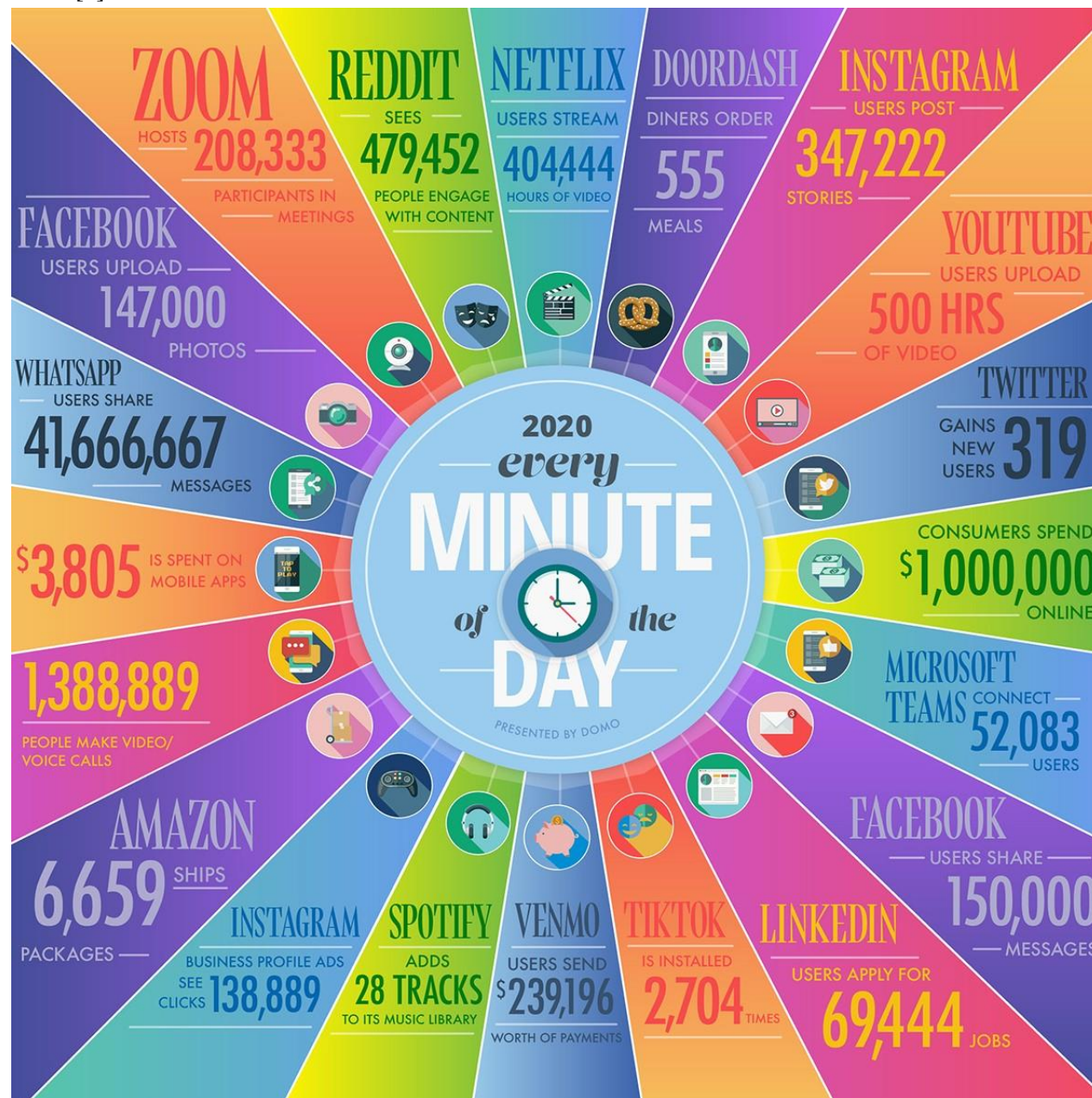
Ως εκ τούτου, η τεχνολογία του IoT απαιτεί νέες υποδομές, συμπεριλαμβανομένων υλικού, λογισμικού καθώς και λειτουργικού συστήματος. Οι άνθρωποι και οι οργανισμοί θα πρέπει να διαχειριστούν την εισροή των παραπάνω δεδομένων και να την αναλύουν σε πραγματικό χρόνο, καθώς αυτή θα αυξάνεται διαρκώς. Στο σημείο αυτό έγκειται η χρήση των Big Data. **Τα εργαλεία ανάλυσης των Big Data είναι σε θέση να χειρίζονται τους μεγάλους όγκους δεδομένων που παράγονται από τις συσκευές του IoT και δημιουργούν μια συνεχή ροή πληροφοριών.** Δηλαδή, οι πληροφορίες από τις συσκευές του IoT βρίσκονται στην πράξη μέσα στα Big Data και μετρούνται μέσω αυτών. Στο άμεσο μέλλον, το μέγεθος που θα χρησιμοποιούν οι αναλυτές για να ποσοτικοποιήσουν τα δεδομένα από τους αισθητήρες και τις συσκευές του IoT θα είναι το Brontobyte (1 Brontobyte = 10^{27} bytes). Οι μονάδες μέτρησης της ποσότητας της πληροφορίας παρουσιάζονται στην Εικόνα 5 [4].



Εικόνα 5: Μονάδες μέτρησης δεδομένων

- Με τον όρο «**δεδομένα παραγόμενα από ανθρώπους**» (human-sourced information) αναφερόμαστε σε δεδομένα που παράγονται από την **καταγραφή των ανθρώπινων δραστηριοτήτων** καθώς από την **αλληλεπίδρασή τους στο διαδίκτυο**, δηλαδή στα μέσα κοινωνικής δικτύωσης (status updates, tweets κτλ.), στην αποστολή άμεσων μηνυμάτων ή/και e-mails, στη λήψη και ανέβασμα φωτογραφιών ή/και βίντεο κ.α. Σήμερα παράγονται δεδομένα οποτεδήποτε κάποιος χρήστης είναι «συνδεδεμένος», από την αναζήτηση στον παγκόσμιο ιστό, την επικοινωνία μέσω άμεσων μηνυμάτων, την αγορά σε ηλεκτρονικά καταστήματα μέχρι τη δημοσίευση σε κοινωνικά δίκτυα. **Ουσιαστικά οποιαδήποτε ψηφιακή κίνηση εκτελείται, αφήνει**

στη συνέχεια ένα ψηφιακό αποτύπωμα και άρα δημιουργεί νέα δεδομένα. Στην Εικόνα 6 απεικονίζονται μερικά ενδεικτικά είδη και μεγέθη δεδομένων, που δημιουργούνται από χρήστες του Internet, σε ένα μόνο λεπτό [5].

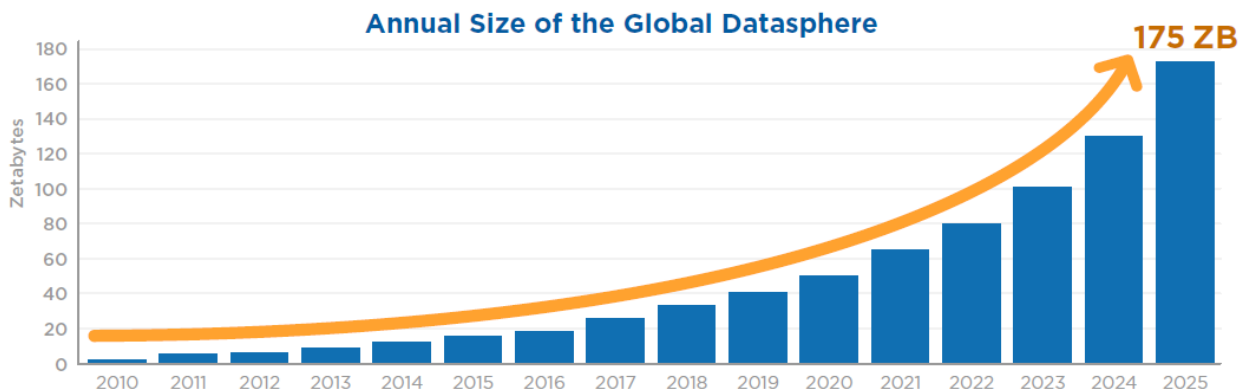


Εικόνα 6: Ενδεικτικά παραδείγματα δεδομένων που παραγόμενα σε ένα λεπτό στο διαδίκτυο

- Τέλος, με τον όρο «δεδομένα παραγόμενα από οργανισμούς ή/και διαδικασίες» (process-mediated data), εννοούμε τα δεδομένα που δημιουργούνται στο περιβάλλον μιας επιχείρησης ή ενός οργανισμού (π.χ. νοσοκομείου) κατά την καταγραφή διαδικασιών και δραστηριοτήτων ενδιαφέροντος. Αυτές μπορεί να αφορούν εισαγωγές ασθενών, εγγραφές πελατών, οικονομικές συναλλαγές, τη λήψη μιας παραγγελίας, τις πληροφορίες διοίκησης συστημάτων (π.χ. logistics), τα δημογραφικά στοιχεία κ.α.

Για να αποκτήσουμε μια αίσθηση του τεράστιου όγκου των παραγόμενων δεδομένων αλλά και του αυξημένου ρυθμού δημιουργίας τους στις μέρες μας, αξίζει να αναφέρουμε ότι πάνω από το 90% των παγκοσμίως

υπαρχόντων δεδομένων έχουν δημιουργεί τα τελευταία 2 χρόνια [6]! Αντιλαμβάνεται κανείς ότι ο ρυθμός παραγωγής νέων δεδομένων είναι πλέον εκθετικός και για την περιγραφή τους χρησιμοποιούνται όλο και μεγαλύτερες μονάδες δεδομένων (βλ. Εικόνα 5). **Ο τρέχων ρυθμός παραγωγής νέων δεδομένων υπολογίζεται σε 16,5 ZB ανά έτος** και υπολογίζεται ότι το 2025 ο συνολικός όγκος των δεδομένων θα ανέλθει στα 175 ZB (Εικόνα 7) [7].

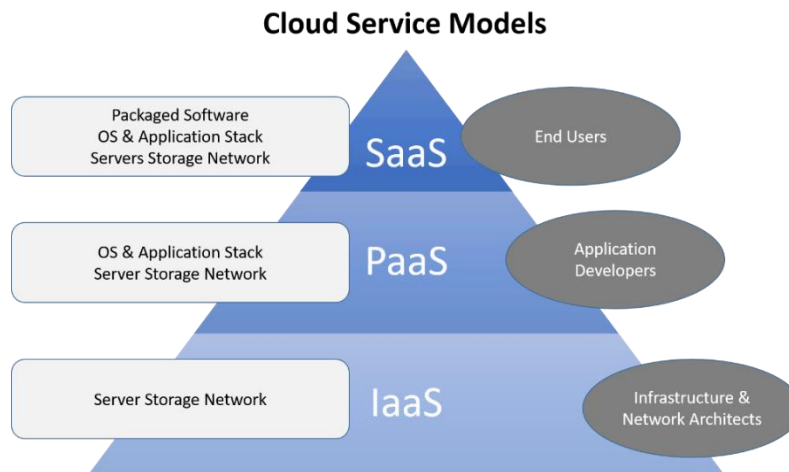


Εικόνα 7: Ενδεικτικός ρυθμός αύξησης της παραγωγής νέων δεδομένων έως το 2025

1.5 Cloud Computing

Ένας εξίσου ενδιαφέρον τεχνολογικός κλάδος είναι αυτός του **Cloud Computing** ή αλλιώς Υπολογιστική Νέφους. Ουσιαστικά αφορά έναν εναλλακτικό τρόπο απόκτησης τεχνολογικών υπηρεσιών, υποδομών και εφαρμογών. **Μέσω του Cloud Computing είναι δυνατή η απόκτηση των επιθυμητών υποδομών (π.χ. εξυπηρετητών και λογισμικού) μέσω του Internet ανάλογα με τη χρήση/κίνηση.** Με πιο απλά λόγια, αφορά το διαμοιρασμό υποδομών (υλικών και λογισμικού), ανάλογα με τις ανάγκες του εκάστοτε χρήστη (π.χ. εταιρίας ή ατόμου). Οι υποδομές αυτές είναι απομακρυσμένα εγκατεστημένες και ο πελάτης μπορεί να τις ενοικιάσει ανάλογα με τις ανάγκες του. Μ' αυτό τον τρόπο εξοικονομεί πόρους, ενώ παράλληλα δεν ασχολείται και με τη συντήρηση και απαρχαίωση του εξοπλισμού, που θα έπρεπε σε διαφορετική περίπτωση να προμηθευτεί, ή την αγορά ακριβού λογισμικού. Ένα ακόμη θετικό στοιχείο είναι η δυνατότητα, ανά πάσα στιγμή, για άμεση κλιμάκωση των υποδομών, όταν αυξάνονται οι ανάγκες του πελάτη. **Ο πελάτης του Cloud Computing μπορεί επομένως να ενοικιάσει υπηρεσίες και υποδομές (εξυπηρετητές, αποθηκευτικό χώρο, κ.ο.κ.), ανάλογα με τις ανάγκες του,** χωρίς να διαθέτει εξειδικευμένο τμήμα πληροφορικής και χωρίς να απασχολείται με τα σχετιζόμενα θέματα. Ο τρόπος αυτός αποδεικνύεται να είναι οικονομικά αποδοτικός, ειδικά όταν οι απαιτήσεις μεγαλώνουν.

Το Cloud Computing έρχεται να δώσει λύση στις ανάγκες των Big Data, μιας και απαιτούνται σημαντικά αυξημένες δυνατότητες και πόροι (π.χ. υπολογιστική ισχύ, αποθηκευτικό χώρο, κ.ά.). Οι απεριόριστοι πόροι που προσφέρονται μέσω του Cloud Computing επιτρέπουν την κατανομή του φόρτου εργασίας, με τρόπο ώστε να είναι δυνατή η σε πραγματικό χρόνο επεξεργασία των τεράστιων όγκων των δεδομένων. Το Cloud Computing επιβάλλει τη χρησιμοποίηση νέων μεθόδων εκτέλεσης των εργασιών επεξεργασίας των δεδομένων ή απόκρισης των συστημάτων σε περίπτωση σφαλμάτων, λόγω του διαμοιρασμού των εργασιών που εκτελούνται. Δεδομένων αυτών των δυνατοτήτων, το Cloud Computing έχει μπει για τα καλά στο πεδίο της ανάλυσης δεδομένων. Οι τεχνολογίες του νέφους παρέχονται με διάφορους τρόπους και μεθόδους, με τα αναμενόμενα οικονομικά οφέλη για τους πελάτες τους. Σε κάθε περίπτωση, θα πρέπει ο πελάτης να γνωρίζει τι ακριβώς θέλει να πετύχει προκειμένου να επιλέξει το ενδεδειγμένο μοντέλο του Cloud Computing. Τα τρία βασικά μοντέλα υπηρεσιών του Cloud Computing (Εικόνα 8) είναι τα ακόλουθα:



***Εικόνα 8:** Τα 3 κύρια μοντέλα υπηρεσιών του Cloud Computing*

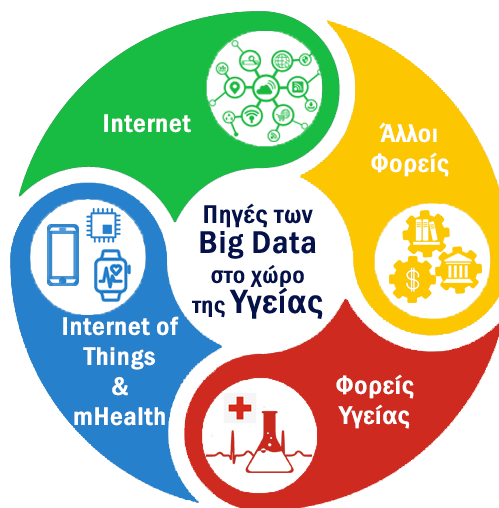
- **Λογισμικό ως Υπηρεσία - Software-as-a-Service (SaaS):** Οι επιθυμητές εφαρμογές παρέχονται με την μορφή υπηρεσίας και δεν απαιτείται η εγκατάστασή τους στους υπολογιστές του πελάτη. Ουσιαστικά πρόκειται για λογισμικό που χρησιμοποιείται μέσω του Internet και «τρέχει» σε κάποιο γνωστό φυλλομετρητή (web browser). Η εφαρμογή τρέχει σε υποδομές που ανήκουν στον πάροχο του SaaS και πωλούνται ως επί το πλείστον σε μηνιαία ή ετήσια βάση.
- **Πλατφόρμα ως Υπηρεσία - Platform-as-a-Service (PaaS):** Αναφέρεται σε μια πλατφόρμα ή περιβάλλον που παρέχεται στον πελάτη ως υπηρεσία, προκειμένου να μπορεί να αναπτύξει και να διαχειριστεί δικές του εφαρμογές. Και σε αυτή την περίπτωση, ο πελάτης έχει πρόσβαση στις συγκεκριμένες εφαρμογές μέσω ενός απλού φυλλομετρητή. Η πλατφόρμα αποτελείται από το απαιτούμενο λογισμικό (π.χ. λειτουργικό σύστημα, άλλες ενδιαμέσες εφαρμογές, πρωτόκολλα επικοινωνίας, κ.ά.) που επιτρέπει στις εφαρμογές να τρέχουν στο νέφος. Τα θέματα ασφάλειας, διαχείρισης, κλιμάκωσης των απαιτήσεων, κ.λπ., σε σχέση με το PaaS, απασχολούν μόνον τον πάροχο της υπηρεσίας και όχι τον πελάτη, ο οποίος πληρώνει μια συνδρομή ανάλογα με τις απαιτήσεις του.
- **Υποδομή ως Υπηρεσία - Infrastructure-as-a-Service (IaaS):** Το συγκεκριμένο μοντέλο παρέχει πρόσβαση σε υποδομές σε ένα εικονικό περιβάλλον μέσω του δικτύου (π.χ. μέσω του Internet). Οι υποδομές αφορούν για παράδειγμα, εξυπηρετητές, άλλο υλικό (π.χ. διάθεση bandwidth, IP διευθύνσεων), κ.ά. Οι πελάτες και πάλι πληρώνουν ανάλογα με τη χρήση και τους πόρους που χρησιμοποιούν ή ενοικιάζουν. Ένα σημαντικό πλεονέκτημα είναι πως υπάρχει η δυνατότητα κλιμάκωσης των υποδομών ανά πάσα στιγμή ανάλογα με τις ανάγκες του πελάτη. Οι υλικές υποδομές (π.χ. εξυπηρετητές) μπορεί να βρίσκονται σε διαφορετικά κέντρα, όπου ο πάροχος είναι υπεύθυνος για τη συντήρηση και την ορθή λειτουργία τους. Με αυτό τον τρόπο δεν απαιτούνται έξοδα συντήρησης και διαχείρισης των υποδομών από τον πελάτη του IaaS.

Όπως γίνεται εύκολα αντιληπτό από τα προαναφερόμενα, οι τεχνολογίες Big Data και Cloud Computing συνεργάζονται άμεσα προκειμένου να επιτύχει κανείς το επιθυμητό αποτέλεσμα.

1.6 Τα Big Data στο χώρο της Υγείας

Ο τομέας της υγείας είναι εκείνος που υιοθετεί πρώτος, αλλά και οδηγεί τις τεχνολογικές εξελίξεις. Επίσης, έχει σημαντικά και ιδιαίτερα χαρακτηριστικά τα οποία τον διακρίνουν, απαιτώντας ιδιαίτερη αντιμετώπιση στις προκλήσεις που θέτει, όπως για παράδειγμα, τα ευαίσθητα προσωπικά δεδομένα, τα θέματα ασφάλειας και διασφάλισης της ποιότητας των δεδομένων, κ.ά. Στη συνέχεια αναλύονται οι βασικότερες **πηγές προέλευσης**

των δεδομένων στο χώρο της υγείας, οι οποίες απεικονίζονται συνοπτικά στην Εικόνα 9 [8] [9] [10] [11]. Τη βασικότερη κατηγορία δεδομένων στο χώρο της υγείας αποτελούν τα αμιγώς βιοϊατρικά – κλινικά δεδομένα προερχόμενα από φορείς υγείας. Στη συνέχεια ακολουθούν τα δεδομένα που προέρχονται από το IoT, την αλληλεπίδραση του ατόμου με το Internet και τέλος τα δεδομένα λοιπών φορέων.



Εικόνα 9: Βασικές πηγές δεδομένων στο χώρο της Υγείας

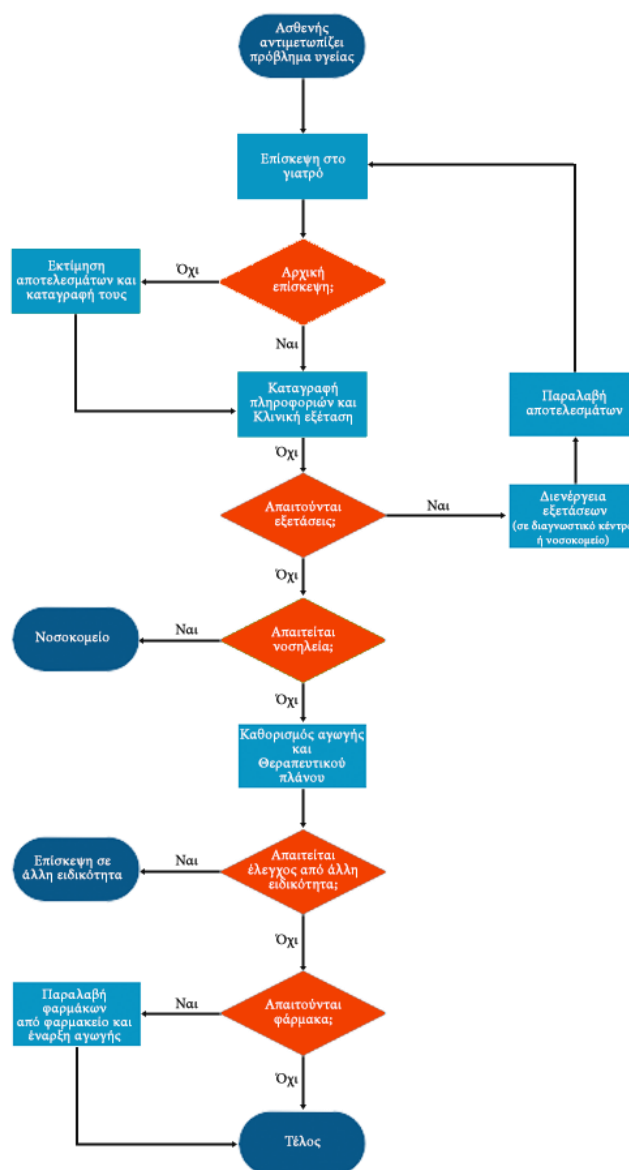
1.6.1 Δεδομένα από Φορείς Υγείας

Οι **φορείς Υγείας** (όπως γιατροί, κέντα υγείας, δημόσια/ιδιωτικά νοσοκομεία, διαγνωστικά κέντρα, κέντρα αποκατάστασης κ.α.) αποτελούν τη βασικότερη και πιο «κλασική» πηγή δεδομένων σχετικά με την υγεία ενός ατόμου. Τα δεδομένα αυτά αποτελούν, κατά βάση, **βιοϊατρικά – κλινικά στοιχεία**. Τέτοια δεδομένα αποτελούν όλα τα στοιχεία από κλινικά ευρήματα, διαγνωστικές εξετάσεις (εργαστηριακές ή/και απεικονιστικές), ιατρικές διαγνώσεις/γνώματαύσεις και ιατρικό ιστορικό (όπως προσωπικά στοιχεία, προβλήματα υγείας, εμβολιασμούς, αλλεργίες, λήψη φαρμάκων κ.α.), τα οποία προέρχονται, όπως αναφέρθηκε, από κάποιο **πάροχο υγείας** (από τον οικογενειακό γιατρό του ασθενούς και το τοπικό φαρμακείο μέχρι κάποια εξειδικευμένη κλινική ή νοσοκομείο). Στην κατηγορία αυτή μπορούν να προστεθούν οι **πληροφορίες από βάσεις βιολογικών δεδομένων**, οι οποίες προκύπτουν από ένα νέο και ταχέως αναπτυσσόμενο κλάδο αυτόν της **γονιδιακής ανάλυσης**. Στο χώρο αυτό δραστηριοποιούνται πλέον εταιρείες που προσφέρουν εμπορικά διαθέσιμο τον πλήρη έλεγχο του γονιδιώματος κάθε ατόμου. Μέσω της ταχείας ανάπτυξης των μεθόδων αλληλούχησης δημιουργήθηκαν τεράστιες ποσότητες γονιδιακών δεδομένων με χαμηλό κόστος, υψηλή ακρίβεια, υψηλή ταχύτητα και ελάχιστη απαίτηση δειγμάτων. Αυτά τα σύνολα δεδομένων αποτελούν νέες πηγές δεδομένων και χρησιμοποιούνται πλέον για την εφαρμογή προσωποποιημένων ιατρικών πρακτικών.

Αναλογιζόμενοι τη λειτουργία του συστήματος υγείας, μπορεί κανείς να διακρίνει επιμέρους χαρακτηριστικά τα οποία παράγουν καθημερινά νέα δεδομένα και αρχικά περνούν απαρατήρητα. Ας πάρουμε για παράδειγμα, μια τυπική περίπτωση ενός πολίτη που αντιμετωπίζει ένα πρόβλημα υγείας [12]. Αρχικά, θα αναζητήσει την συμβουλή από τον οικογενειακό του ιατρό, ο οποίος ανάλογα με τα κλινικά ευρήματα και το ιστορικό του ασθενούς θα του προτείνει μια σειρά διαγνωστικών εξετάσεων ή/και φαρμακευτικής αγωγής. Ο ασθενής θα οδηγηθεί επομένως σε ένα διαγνωστικό κέντρο ή σε ένα νοσοκομείο ή κλινική προκειμένου να εκτελέσει τις συγκεκριμένες εξετάσεις. Λαμβάνοντας τα αποτελέσματα, θα επιστρέψει στον ιατρό του και, αν κριθεί απαραίτητο, θα πρέπει να επισκεφτεί και ιατρό άλλης/ων ειδικότητας/ων με πιθανότητα διενέργειας και άλλων συμπληρωματικών εξετάσεων ή συμπληρωματικής αγωγής ή ακόμη και νοσηλείας σε κάποιο νοσοκομείο. Το πλήρες διάγραμμα ροής μιας τυπικής αντιμετώπισης ενός προβλήματος υγείας αναπαρίσταται στην Εικόνα 10.

Το σενάριο που παρουσιάζεται στην Εικόνα 10 είναι πολύ πιθανόν να συμβεί αρκετές φορές στη ζωή ενός ανθρώπου. Ο ανωτέρω «δρόμος» αντιμετώπισης μιας τυπικής ασθένειας μπορεί να ακολουθήσει ένα οποιοδήποτε μονοπάτι ανάλογα με την περίπτωση του ασθενούς (σοβαρότητα, κλινικά ευρήματα, κ.ο.κ.). Παρατηρώντας τις λεπτομέρειες κάθε βήματος του ασθενούς, ανακύπτουν σημαντικά συμπεράσματα: Κάθε φορέας ή επαγγελματίας υγείας αναζητά, καταγράφει και διατηρεί αρχείο με τα:

- ✓ δημογραφικά στοιχεία του ασθενούς (π.χ. ονοματεπώνυμο, τηλέφωνα επικοινωνίας, κατοικία, κ.ο.κ.),
- ✓ κλινικά ευρήματα (κλινική εξέταση),
- ✓ διαγνωστικά ευρήματα (αποτελέσματα πρόσφατων ή παλαιότερων εξετάσεων),
- ✓ στοιχεία ασφάλισης (π.χ. ιδιωτική ασφάλιση, δημόσια ασφάλιση, ανασφάλιστος, ιδιώτης),
- ✓ ιατρικό ιστορικό,
- ✓ οικογενειακό ιστορικό,
- ✓ κοινωνικό ιστορικό.



Εικόνα 10: Ο συνήθης «δρόμος» αντιμετώπισης ενός προβλήματος υγείας

Όλα τα ανωτέρω συμβαίνουν όμως κάθε φορά που ο ασθενής ακολουθεί ένα από τα μονοπάτια αντιμετώπισης της κατάστασης της υγείας του. Ας αναλογιστούμε τώρα τον χρόνο που δαπανάται από τον εργαζόμενο / επαγγελματία υγείας για την καταγραφή ή αναζήτηση των συγκεκριμένων πληροφοριών από κάθε σημείο με το οποίο έρχεται σε επαφή ο ασθενής. Τέλος, ας προσπαθήσουμε να σκεφτούμε τον όγκο των δεδομένων που πρέπει να αποθηκεύονται προκειμένου να έχουμε πρόσβαση στις επιθυμητές πληροφορίες. Πολλαπλασιάζοντας όλα αυτά επί τον αριθμό των επισκέψεων που πραγματοποιούνται για κάθε έναν πολίτη επί τον όγκο των δεδομένων που παράγονται σε κάθε επίσκεψη για το ίδιο σημείο (π.χ. ιατρός, νοσηλευτικό ίδρυμα, διαγνωστικό κέντρο), προκύπτει ένα μοναδικό και απλό συμπέρασμα: Τα δεδομένα είναι απίστευτα πολλά και αποτελούν ένα ευρύ κομμάτι των Big Data στον τομέα της υγείας.

1.6.2 Δεδομένα υγείας από το IoT

Όπως αναφέρθηκε και σε προηγούμενη ενότητα, το IoT αποτελεί μία βασική πηγή παραγωγής νέων δεδομένων στο χώρο των Big Data. Έτσι, πολλά από αυτά τα δεδομένα σχετίζονται άμεσα ή έμμεσα με την υγεία ενός ατόμου. Πιο συγκεκριμένα, αντικείμενα που εμπίπτουν σε αυτή την κατηγορία είναι τα **smartphones** (λόγω των διαφόρων αισθητήρων που ενσωματώνουν), οι έξυπνες συσκευές που φοριούνται, γνωστά ως **wearables** (Εικόνα 11), και οι φορητές **ιατρικές συσκευές** ή/και **αισθητήρες**. Όλες οι παραπάνω συσκευές απαρτίζουν ουσιαστικά το μεγαλύτερο τμήμα της οντότητας της «κινητής υγείας» ή κοινώς, του **mHealth**. Μέσω όλων αυτών να συσκευών είναι δυνατόν να μετρηθεί μία **πληθώρα παραμέτρων και δεικτών, βιοϊατρικών και μη, τα οποία όμως σχετίζονται όλα με τη συνολική υγεία ενός ατόμου**.

Μερικά παραδείγματα τέτοιων παραμέτρων αποτελούν:

- ο αριθμός των εκτελούμενων βημάτων/ημέρα,
- ο καρδιακός ρυθμός,
- η μεταβλητότητα του καρδιακού ρυθμού,
- η θερμοκρασία, το βάρος,
- η πίεση του αίματος,
- ο κορεσμός οξυγόνου στο αίμα,
- η γλυκόζη του αίματος,
- η αγωγιμότητα του δέρματος,
- τα επίπεδα του άγχους,
- η ποιότητα του ύπνου,
- οι προσλαμβανουσες τροφές (με στόχο τον υπολογισμό θερμίδων και θρεπτικών συστατικών),
- η αναπνευστική λειτουργία.



Εικόνα 11: Συσκευές Wearables

Στις προηγούμενες παραμέτρους, μπορεί να προστεθούν τα δεδομένα και άλλων αισθητήρων, π.χ. περιβαλλοντικών, οι οποίοι μετρούν δείκτες που δύναται να επηρεάσουν την υγεία ενός ατόμου. Η μέτρηση μπορεί να αφορά σε ατμοσφαιρικούς ρύπους, ποσότητες ραδιενέργειας, ρύπανση υδάτων κ.α. Η μέτρηση τέτοιου είδους δεικτών κρίνεται ιδιαίτερη σημαντική σε περιπτώσεις όπου μελετάται η μακροχρόνια επίδραση των περιβαλλοντικών συνθηκών στον άνθρωπο και συνδράμει στη χάραξη κανονισμών και πολιτικών υγείας.

Ακόμη, μέσω των διαφόρων εφαρμογών (**health apps**), **οι ασθενείς είναι σε θέση να καταγράφουν τις εμπειρίες τους σε πραγματικό χρόνο** και αυτό να συνδυάζεται με μια «παθητική συλλογή» δεδομένων μέσω αισθητήρων και κινητών συσκευών. Τα health apps αποτελούν μια πολλά υποσχόμενη επιλογή για να διαχειριστεί κάποιος συστηματικά την υγεία του, βοηθώντας, τόσο τους ασθενείς όσο και τους γιατρούς, στην παρακολούθηση και τον έλεγχο, ακόμα και όταν οι ασθενείς δε βρίσκονται σε κάποιο αυστηρό κλινικό περιβάλλον, όπως ένα ιατρείο ή νοσοκομείο (Εικόνα 12). Με αυτό τον τρόπο, οι πληροφορίες που εξάγονται συνεισφέρουν στο αυξανόμενο σύνολο δεδομένων υγείας και μπορούν να συμβάλουν στην παροχή καλύτερης φροντίδας για κάθε ασθενή καθώς και στη χάραξη πολιτικών υγείας για όλους τους ασθενείς.



Εικόνα 12: Η συμβολή του mHealth στην καταγραφή και παρακολούθηση της υγείας

1.6.3 Δεδομένα υγείας από την αλληλεπίδραση χρηστών με το Internet

Μία νέα πηγή δεδομένων υγείας, η οποία έχει εμφανιστεί τα τελευταία χρόνια και αποδεικνύεται ιδιαίτερα πολύτιμη, είναι οι πληροφορίες που παράγονται κατά την αλληλεπίδραση των χρηστών με το Internet. Αυτή η κατηγορία περιλαμβάνει, μεταξύ άλλων, τις **αναζητήσεις** των χρηστών στις διάφορες **μηχανές αναζήτησης**, το **ιστορικό περιήγησής** τους, τις **ηλεκτρονικές αγορές** τους καθώς και όλη τη δραστηριότητα που διενεργείται στα **μέσα κοινωνικής δικτύωσης** (π.χ. δίκτυο φίλων, αναρτήσεις, τοποθεσίες, φωτογραφίες, ενδιαφέροντα κ.α.). Τα δεδομένα αυτά μπορούν να συλλέγονται είτε με την αυτόματη ανάκτηση δεδομένων, είτε με ερωτήσεις προς τους χρήστες σχετικά με τη συμπεριφορά τους και τις συνήθειες τους ή μέσω ειδικών εφαρμογών. Ένα πολύ χαρακτηριστικό παράδειγμα αξιοποίησης των δεδομένων αυτής της κατηγορίας αποτελεί η παρακολούθηση και ανάλυση της μετάδοσης και εξάπλωσης ασθενειών ή/και επιδημιών (π.χ. γρίπη, HIV), με σκοπό, μεταξύ άλλων, την έγκαιρη λήψη μέτρων [13].

1.6.4 Δεδομένα υγείας από άλλους Φορείς

Τέλος, δεδομένα από άλλους φορείς, όχι άμεσα επιφορτισμένους με την παροχή υπηρεσιών υγείας, μπορούν να αποτελέσουν κι αυτά πηγή των Big Data στο κομμάτι της υγείας. Τέτοιου είδους δεδομένα αποτελούν τα στοιχεία από τους **ασφαλιστικούς οργανισμούς** (δημόσιους ή ιδιωτικούς) και από φορείς συλλογής και παραγωγής σχετιζόμενων πληροφοριών, όπως **κρατικούς φορείς** και **επιδημιολογικά ινστιτούτα**. Τα στοιχεία αυτά δύναται να αφορούν μεταξύ άλλων δημογραφικές πληροφορίες, στατιστικές μελέτες, πολιτικές υγείας, επιδημιολογικά δεδομένα, μητρώα φαρμάκων και φόρμες παραγγελιών. Κάποιοι από τους ελληνικούς οργανισμούς που εμπίπτουν

σε αυτή την κατηγορία είναι το Υπουργείο Υγείας, η Ηλεκτρονική Διακυβέρνηση Κοινωνικής Ασφάλισης (Η.ΔΙ.Κ.Α.), ο Εθνικός Οργανισμός Φαρμάκων (Ε.Ο.Φ.), ο Εθνικός Οργανισμός Δημόσιας Υγείας (Ε.Ο.Δ.Υ.) και η Ελληνική Στατιστική Αρχή (ΕΛ.ΣΤΑΤ.). Ακόμη, τα διοικητικά αρχεία του Εθνικού Οργανισμού Παροχής Υπηρεσιών Υγείας (Ε.Ο.Π.Υ.Υ.) αποτελούν κι αυτά πηγή πολύτιμων πληροφοριών. Στοιχεία όπως ο γάμος, η μετανάστευση, οι γεννήσεις και οι θάνατοι ενσωματώνονται και αυτά με τη σειρά τους σε αυτή την ομάδα δεδομένων.

Τέλος, τα **αποτελέσματα κλινικών μελετών**, οι **επιστημονικές βάσεις δεδομένων**, οι βάσεις δεδομένων σχετικά με ανεπιθύμητες ενέργειες φαρμάκων καθώς και τα διάφορα ερευνητικά αποτελέσματα και μητρώα μελετών ολοκληρώνουν τον τεράστιο όγκο δεδομένων που μπορούν να διαδραματίσουν καταλυτικό ρόλο στην υγεία ενός ατόμου.

1.6.5 Τα Big Data και ο Ηλεκτρονικός Φάκελος Υγείας

Ως **Ηλεκτρονικός Φάκελος Υγείας (ΗΦΥ)** ορίζεται ένα αποθετήριο πληροφοριών, σε επεξεργάσιμη μορφή από υπολογιστή, που σχετίζονται με την υγεία ενός ατόμου, το οποίο αποθηκεύεται και μεταφέρεται με ασφάλεια και είναι εύκολα προσβάσιμο από πολλούς εξουσιοδοτημένους χρήστες. Έχει ένα τυποποιημένο μοντέλο πληροφοριών και σκοπός του είναι η υποστήριξη μιας συνεχούς, ποιοτικής, ολοκληρωμένης υγειονομικής περίθαλψης και η παροχή πληροφοριών [14]. Από τον προηγούμενο ορισμό γίνεται αντιληπτό πως ο ΗΦΥ περιέχει ένα μεγάλο εύρος δεδομένων, συμπεριλαμβανομένων δημογραφικών στοιχείων, φαρμακευτικών αγωγών, κλινικών πληροφοριών/διαγνώσεων, αποτελεσμάτων διαγνωστικών εξετάσεων και ιατρικού ιστορικού. Έτσι, λοιπόν, οι ΗΦΥ των πολιτών αποτελούν σύνολα δεδομένων αυξημένου όγκου, μεγάλης ποικιλομορφίας, μη ενιαίας δόμησης και αναμφίβολα ιδιαίτερα μεγάλης αξίας για την κλινική πράξη, τα οποία αποτελούν και μερικά από τα χαρακτηριστικά των Big Data. Αν συνυπολογίσει κανείς τον αριθμό των ασθενών αθροιστικά, τα καθημερινά περιστατικά υγείας καθώς και συγκεκριμένα περιβάλλοντα που παράγουν συνεχείς ροές δεδομένων (π.χ. μονάδες εντατικής θεραπείας) προκύπτει πως και η ταχύτητα δημιουργίας και απαίτησης ανάλυσης είναι συχνά μεγάλη. Γενικά λοιπόν, ο ΗΦΥ μπορεί να θεωρηθεί ως μία οντότητα Big Data και άρα τα δεδομένα του να αντιμετωπίζονται ως τέτοια [15]. Σαφώς ο ΗΦΥ δεν περιέχει όλα τα δεδομένα που δύναται να συσχετιστούν με την υγεία ενός ατόμου (όπως αυτά που αναφέρθηκαν στις προηγούμενες ενότητες), αλλά ενοποιεί τα περισσότερα από αυτά και εμπλουτίζεται συνεχώς (Εικόνα 13).



***Εικόνα 13:** Ο ΗΦΥ συγκεντρώνει και ενοποιεί μεγάλο μέρος των δεδομένων υγείας ενός ατόμου*

Αντλαμβάνεται λοιπόν κανείς πως **οι πιθανές πηγές Big Data εξελίσσονται συνεχώς** και πως ο όγκος των παραγόμενων δεδομένων είναι τόσο μεγάλος και ο ρυθμός τους εκθετικά αυξανόμενος, που είναι πλέον αδύνατον να συγκεντρωθούν από έναν φορέα και να είναι εύκολα επεξεργάσιμα και διαχειρίσιμα. **Χωρίς αυτά όμως είναι**

αδύνατη η βελτίωση στην παροχή υπηρεσιών υγείας. Χωρίς άμεση και εύκολη πρόσβαση στην επιθυμητή πληροφορία, είναι αδύνατη η επιθυμητή εξέλιξη της επιστήμης, η διεξαγωγή ερευνών, η ανεύρεση νέων θεραπειών και φαρμάκων, κ.ά. Τα δεδομένα είναι αυτά που μας βοηθούν στην περαιτέρω καινοτομία και πρόοδο. Για το λόγο αυτό, απαιτούνται νέες τεχνικές και τεχνολογίες, οι οποίες αναλύονται στη συνέχεια, για να μπορέσουμε να αποκομίσουμε τα μέγιστα από αυτό τον πλούτο πληροφοριών και των εν δυνάμει γνώσεων.

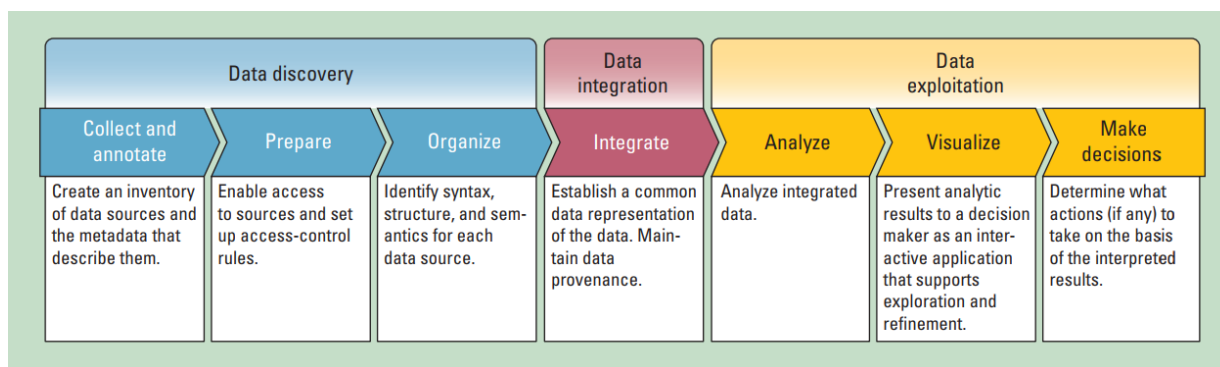
1.7 Επεξεργασία, Ανάλυση και Εξόρυξη Γνώσης

Όπως αναφέρθηκε, η κατοχή μεγάλων όγκων δεδομένων δεν έχει ή δε δημιουργεί αυτόματα κάποιου είδους αξία για ένα οργανισμό. Η αποθήκευση δεδομένων μπορεί να παρομοιαστεί με την αποθήκευση άχρηστων υλικών σε μια αποθήκη τα οποία δεν ανασύρονται ποτέ για χρήση. Αντίθετα, **η αξιοποίησή τους, μέσω της επεξεργασίας, της ανάλυσης και της δημιουργίας εφαρμογών δεδομένων είναι αυτά που προσδίδουν υπεραξία** και ανταγωνιστικό πλεονέκτημα σε επιχειρήσεις και οργανισμούς. Σε αυτό το σημείο αναδεικνύεται η ανάγκη για εργαλεία και μεθόδους ανάλυσης και επεξεργασίας πέραν των παραδοσιακών. Για τον σκοπό αυτό έχουν αναπτυχθεί τεχνολογίες, όπως οι βάσεις δεδομένων τύπου NoSQL και τα κατακεντρωμένα συστήματα αρχείων, τα οποία αποτελούν χώρο έρευνας και διαρκούς ανάπτυξης και αναλύονται στην επόμενη ενότητα.

Η ανάλυση των Big Data περιλαμβάνει αρκετές διακριτές φάσεις και η κάθε μια έχει ιδιαίτερες απαιτήσεις. Οι φάσεις αυτές είναι λίγο έως πολύ κοινές και σε έργα με δεδομένα χαμηλής κλίμακας. Όπως και στη βασική μεθοδολογία της εξόρυξης δεδομένων, βασικά στοιχεία είναι η συλλογή δεδομένων, η επεξεργασία και ο καθαρισμός, η ανάλυση και τέλος η ερμηνεία.

Συνοπτικά, η διαδικασία για την ανάλυση των Big Data και την αξιοποίησή τους μπορεί να αναλυθεί σε τρεις φάσεις (Εικόνα 14) με διακριτές εργασίες:

- Η πρώτη φάση είναι η **διαχείριση των δεδομένων**, η οποία περιλαμβάνει διαδικασίες και υποστηρικτικές τεχνολογίες για τη **λήψη, συλλογή, καταγραφή, καθαρισμό, σύνθεση και αποθήκευση** των δεδομένων. Οι πηγές που σήμερα παράγουν δεδομένα είναι ποικίλες και είναι σε θέση να καταγράφουν τεράστιες ποσότητες δεδομένων. Η πληροφορία που συλλέγεται συχνά δεν είναι σε μορφή που είναι έτοιμη για ανάλυση, οπότε απαιτούνται τεχνικές και τεχνολογίες που θα διασφαλίσουν ότι τα δεδομένα θα λάβουν την κατάλληλη μορφή. Μια μεγάλη πρόκληση είναι η δημιουργία κατάλληλων φίλτρων, ώστε να μη χαθεί πολύτιμη πληροφορία και η αυτόματη δημιουργία σωστών μεταδεδομένων (metadata), τα οποία να περιγράφουν τα δεδομένα και με ποιον τρόπο αυτά καταγράφονται και μετρούνται. Οι διαδικασίες μπορούν να εκτελεστούν με εργαλεία της εξόρυξης δεδομένων ή ειδικά κατά περίπτωση.
- Η ετερογενής προέλευση των δεδομένων απαιτεί τη χρήση **τεχνικών ολοκλήρωσης**, σε μια ενδιάμεση φάση, καθώς οι μέθοδοι ανάλυσης Big Data είναι θεμελιωδώς διαφορετικοί από την παραδοσιακή στατιστική ανάλυση σε μικρά δείγματα. Τα Big Data είναι συχνά δυναμικά, ετερογενή, διασυνδεδεμένα και όχι και πολύ αξιόπιστα. Επίσης δημιουργούν μεγάλα ετερογενή δίκτυα δεδομένων, τα οποία μπορούν να αξιοποιηθούν για την εξαγωγή σχέσεων και μοντέλων καθώς και τη διερεύνηση ελλειπών δεδομένων.
- Η δεύτερη φάση είναι η **αναλυτική** και περιλαμβάνει τη **δημιουργία μοντέλων**, την **ανάλυση και ερμηνεία** τους, που είναι το σημαντικό βήμα, ώστε τα δεδομένα να είναι σε θέση να **παρέχουν αξία** και να συντελέσουν στην ορθή υποβοήθηση για τη λήψη αποφάσεων. Οι αναλυτικές τεχνικές μπορούν πολύ γενικά να διακριθούν σε τεχνικές ανάλυσης κειμένου, εικόνας, ήχου, κοινωνικών δικτύων και προβλέψεις. Τελικά, έπειτα από την ανάλυση και την επεξεργασία τους, φτάνουμε στη γνώση στην οποία έγκειται και η ιδιαίτερα αξία των Big Data.



Εικόνα 14: Η αλυσίδα αξίας (value chain) των Big Data

Η φιλοσοφία των Big Data βασίζεται στην αρχή πως **όσο περισσότερα γνωρίζει κανείς για μία κατάσταση, τόσο πιο βαθιά μπορεί να την κατανοήσει και τόσο πιο αξιόπιστα να προβλέψει το αποτέλεσμά της στο μέλλον**. Τομείς που έχουν το μεγαλύτερο πλεονέκτημα σε αυτό το χώρο είναι ο τομέας της υγείας, ο δημόσιος τομέας, το λιανεμπόριο, καθώς και ο τομέας της βιομηχανικής παραγωγής. Το πλεονέκτημα απορρέει από το γεγονός ότι πρόκειται για τομείς που έχουν τη δυνατότητα συλλογής μεγάλων όγκων δεδομένων από διάφορες πηγές και μπορούν να τα αξιοποιήσουν προς όφελός τους, αυξάνοντας είτε το ανταγωνιστικό τους πλεονέκτημα είτε την ποιότητα των υπηρεσιών προς τους πολίτες. Μερικά ενδεικτικά και ιδιαίτερα υποσχόμενα παραδείγματα της αξιοποίησης αυτής της γνώσης αποτελούν η θεραπεία και πρόληψη ασθενειών (όπως π.χ. του καρκίνου), η εξερεύνηση του διαστήματος (π.χ. νέων πλανητών), η πρόβλεψη και έγκαιρη απόκριση σε μεγάλες φυσικές ή/και ανθρώπινες καταστροφές (π.χ. σεισμοί, πόλεμοι) και η βελτίωση της ασφάλειας (π.χ. πρόληψη της εγκληματικότητας).

1.8 Τεχνικές Ανάλυσης των Big Data

Για την εκτέλεση εργασιών σε Big Data είναι αναγκαία η υποβοήθηση από τεχνικές και κατάλληλες μεθόδους για την επεξεργασία των δεδομένων. Παρόλο που τεχνικές και τεχνολογίες για την ανάλυση και επεξεργασία δεδομένων έχουν διαμορφωθεί και υπάρχουν σε χρήση εδώ και δεκαετίες, διαπιστώνεται ότι εμφανίζουν ελλείψεις σε σημεία όπως η κλιμάκωση των συστημάτων, η διαχείριση της πολυπλοκότητας και του όγκου με αποτέλεσμα να μην είναι επαρκείς για Big Data. Τα τελευταία χρόνια, έχουν αναπτυχθεί πλήθος από νέες τεχνολογίες ενώ ταυτόχρονα οι παραδοσιακές τεχνικές έχουν προσαρμοστεί για την συγκέντρωση, διαχείριση, ανάλυση και οπτικοποίηση των Big Data.

Οι **τεχνικές και οι μέθοδοι** αντλούν θεωρητικά στοιχεία από διάφορους επιστημονικούς τομείς, συμπεριλαμβανομένων της στατιστικής, επιστήμης των υπολογιστών, των εφαρμοσμένων μαθηματικών και της οικονομίας, ενώ οι **τεχνολογίες** συμβαδίζουν με την εξέλιξη των υπολογιστικών συστημάτων. Στη συνέχεια παρουσιάζουμε συνοπτικά ορισμένες από τις τεχνικές και τις τεχνολογίες που είναι σήμερα διαθέσιμες.

1.8.1 Τεχνικές για την ανάλυση δεδομένων

Για την ανάλυση συνόλων δεδομένων έχουν αναπτυχθεί κατά καιρούς αρκετές τεχνικές που αντλούν στοιχεία από διάφορους επιστημονικούς κλάδους, όπως η στατιστική και η επιστήμη των υπολογιστών (κυρίως το πεδίο της μηχανικής μάθησης). Ορισμένες τεχνικές και μεθοδολογίες αναπτύχθηκαν παλαιότερα, όταν υπήρχαν πολύ μικρότερες ποσότητες και ποικιλία στα δεδομένα, αλλά έχουν προσαρμοστεί με επιτυχία, έτσι ώστε να παράγουν αξιόπιστα αποτελέσματα και για τα πολύ μεγάλα και ποικιλόμορφα σύνολα δεδομένων, όπως τα Big Data. Άλλες έχουν αναπτυχθεί πρόσφατα, ειδικά για τα Big Data.

Παρακάτω παρουσιάζονται βασικές κατηγορίες τεχνικών που εφαρμόζονται σε ευρύ φάσμα εφαρμογών. Μπορούμε να ομαδοποιήσουμε τις τεχνικές σε αυτές που προέρχονται από τη **στατιστική** και τα **μαθηματικά** και αυτές που προέρχονται από την **επιστήμη των υπολογιστών**. Ωστόσο, υπάρχουν και ορισμένες τεχνικές που έχουν ευρύτερη αφετηρία.

2.8.1.1. Στατιστικές μέθοδοι για την ανάλυση δεδομένων

Η στατιστική αποτελεί την κατεξοχήν επιστήμη συλλογής, οργάνωσης και ερμηνείας δεδομένων, συμπεριλαμβανομένων του σχεδιασμού ερευνών και πειραμάτων. Οι στατιστικές τεχνικές χρησιμοποιούνται για την παραγωγή κρίσεων για τις σχέσεις μεταξύ μεταβλητών, καθώς και του ποιες σχέσεις μεταξύ των μεταβλητών είναι πιθανό αποτέλεσμα κάποιας υποκείμενης αιτιώδους σχέσης. Ειδικές τεχνικές που εντάσσονται στο ευρύτερο πλαίσιο της στατιστικής και των μαθηματικών, ή έχουν συνάφεια σε κάποιο βαθμό, αναφέρονται παρακάτω:

- Η **ανάλυση ομάδων** (cluster analysis) είναι μια στατιστική μέθοδος για την ταξινόμηση αντικειμένων που διαχωρίζει μια ομάδα σε μικρότερες ομάδες παρόμοιων αντικειμένων, των οποίων τα χαρακτηριστικά της ομοιότητας δεν είναι γνωστά εκ των προτέρων.
- Η **παλινδρόμηση** (regression analysis) προσδιορίζει πώς μεταβάλλεται η τιμή εξαρτημένων μεταβλητών όταν μία ή περισσότερες ανεξάρτητες μεταβλητές τροποποιούνται. Συχνά χρησιμοποιείται για μοντέλα πρόβλεψης.
- Η **ταξινόμηση** (classification) έχει ως αντικείμενο τον προσδιορισμό των κατηγοριών στις οποίες ανήκουν νέα στοιχεία, με βάση ένα σύνολο εκπαίδευσης που περιέχει τα σημεία δεδομένων που έχουν ήδη ταξινομηθεί.
- Η **ανάλυση χρονοσειρών** (time series analysis) είναι ένα σύνολο τεχνικών για την ανάλυση ακολουθιών σημείων δεδομένων, που αντιπροσωπεύουν τιμές σε διαδοχικές χρονικές στιγμές, ώστε να εξαχθούν χρήσιμα συμπεράσματα από τα δεδομένα.
- Η **προγνωστική μοντελοποίηση** (predictive simulation/analytics) κατά την οποία δημιουργείται ή επιλέγεται ένα μαθηματικό μοντέλο, για να προβλέψει καλύτερα την πιθανότητα ενός αποτελέσματος. Η παλινδρόμηση είναι ένα παράδειγμα τεχνικής προγνωστικής μοντελοποίησης.

2.8.1.2. Ανάλυση δεδομένων με σύγχρονες τεχνικές και μεθόδους της επιστήμης υπολογιστών

Η επιστήμη υπολογιστών έχει συμβάλει σε μεγάλο βαθμό στην ανάπτυξη των τεχνικών για τα Big Data, τόσο με τη χρήση αλγορίθμων, όσο και την εξέλιξη των υπολογιστικών συστημάτων. Μερικές τεχνικές που προέρχονται από το πεδίο αυτό είναι οι παρακάτω:

- Η **μηχανική μάθηση** (machine learning) αποτελεί ειδικότητα της επιστήμης των υπολογιστών στο πεδίο της τεχνητής νοημοσύνης (artificial intelligence) και ασχολείται με το σχεδιασμό και ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να εξελίσσουν τις συμπεριφορές τους βασιζόμενοι σε εμπειρικά δεδομένα.
- Η **εποπτευόμενη μάθηση** (supervised learning), ειδικότερα, αποτελεί σύνολο τεχνικών μηχανικής μάθησης που συνάγουν μια λειτουργία ή σχέση από ένα σύνολο δεδομένων εκπαίδευσης.
- Η **εκμάθηση χωρίς επίβλεψη** (unsupervised learning) είναι ένα σύνολο τεχνικών μηχανικής μάθησης που ανακαλύπτουν κρυμμένες δομές σε μη ταξινομημένα δεδομένα. Η ανάλυση ομάδων είναι ένα παράδειγμα εκμάθησης χωρίς επίβλεψη.
- Η **αναγνώριση προτύπων** (pattern recognition) είναι ένα σύνολο τεχνικών μηχανικής μάθησης που αναθέτουν κάποια τιμή εξόδου (ή ετικέτα) σε μια δεδομένη τιμή εισόδου (ή παράδειγμα), σύμφωνα με ένα συγκεκριμένο αλγόριθμο. Οι τεχνικές ταξινόμησης είναι ένα τέτοιο παράδειγμα.
- Η **εξόρυξη δεδομένων** (data mining) είναι μια σειρά από τεχνικές για την εξαγωγή μοτίβων από μεγάλα σύνολα δεδομένων, συνδυάζοντας μεθόδους από τη στατιστική και τη μηχανική μάθηση με τη διαχείριση βάσεων δεδομένων. Περιλαμβάνουν την εκμάθηση κανόνων σχέσεων, την ανάλυση ομάδων, την ταξινόμηση και την παλινδρόμηση.

- Οι **γενετικοί αλγόριθμοι** (genetic algorithms), που είναι εμπνευσμένοι από τη διαδικασία της φυσικής εξέλιξης, χρησιμοποιούνται για βελτιστοποίηση.
- Η **προσομοίωση** (simulation) είναι η μοντελοποίηση της συμπεριφοράς πολύπλοκων συστημάτων, που χρησιμοποιείται συχνά για την πρόβλεψη και τον σχεδιασμό σεναρίων. Για παράδειγμα, οι προσομοιώσεις Monte Carlo είναι μια κατηγορία αλγορίθμων που βασίζονται σε επαναλαμβανόμενες τυχαίες δειγματοληψίες (δηλαδή εκτελούνται χιλιάδες προσομοιώσεις, η καθεμία με βάση διαφορετικές υποθέσεις).
- Τα **νευρωνικά δίκτυα** (neural networks) είναι υπολογιστικά μοντέλα, εμπνευσμένα από τη δομή και λειτουργία των βιολογικών νευρωνικών δικτύων (δηλαδή, τα κύτταρα και τις συνδέσεις εντός του εγκεφάλου), που βρίσκουν μοτίβα σε δεδομένα.
- Η **ανάλυση δικτύου** (network analysis) είναι ένα σύνολο τεχνικών που χρησιμοποιούνται για τον χαρακτηρισμό των σχέσεων μεταξύ διακριτών κόμβων σε ένα γράφημα ή ένα δίκτυο. Στην ανάλυση των κοινωνικών δικτύων αυτό αφορά στις συνδέσεις μεταξύ των ατόμων σε μια κοινότητα ή μια οργάνωση και την ανάλυσή τους (π.χ. πώς οι πληροφορίες ταξιδεύουν, ή ποιος έχει την μεγαλύτερη επιρροή πάνω σε ποιον).

2.8.1.3. Αξιοποίηση των δεδομένων

Εκτός από τις παραπάνω τεχνικές, υπάρχουν και ορισμένες γενικότερες τεχνικές που χρησιμοποιούνται στην ανάλυση δεδομένων με σκοπό την αξιοποίησή τους (data exploitation). Ενδεικτικά αναφέρουμε τις εξής:

- Η **βελτιστοποίηση** (mathematical optimization) αποτελείται από ένα σύνολο αριθμητικών τεχνικών που χρησιμοποιούνται για τον επανασχεδιασμό πολύπλοκων συστημάτων και διαδικασιών με σκοπό τη βελτίωση των επιδόσεών τους, σύμφωνα με μία ή περισσότερες αντικειμενικές μετρήσιμες παραμέτρους.
- Ο **πληθοπορισμός** (crowdsourcing) είναι μια τεχνική για τη συλλογή στοιχείων που υποβλήθηκαν από μια μεγάλη ομάδα ανθρώπων ή κοινότητα, συνήθως μέσω ανοικτής πρόσκλησης και μέσω δικτυωμένων μέσων, όπως το διαδίκτυο.
- Η **οπτικοποίηση** (visualization) περιλαμβάνει τεχνικές που χρησιμοποιούνται για τη δημιουργία εικόνων, διαγραμμάτων, ή κινούμενων σχεδίων για την επικοινωνία, την κατανόηση και τη βελτίωση των αποτελεσμάτων των αναλύσεων δεδομένων μεγάλης κλίμακας.
- Η **ανάλυση συναισθήματος** (sentiment analysis) είναι μια εφαρμογή επεξεργασίας φυσικής γλώσσας και άλλων αναλυτικών τεχνικών για τον εντοπισμό και την εξαγωγή υποκειμενικής πληροφορίας από πηγές κειμένου.

Οι παραπάνω τεχνικές απαιτούν από τον επιστήμονα δεδομένων την κατάλληλη εξοικείωση σε συνδυασμό με την τεχνογνωσία ώστε να είναι σε θέση να αξιοποιήσει τις δυνατότητές τους στο μέγιστο βαθμό.

1.8.2 Ανάλυση Big Data στο χώρο της υγείας

Η ανάλυση των Big Data παρέχει τη δυνατότητα να ριζικής αλλαγής των υφιστάμενων κλινικών μοντέλα για μια έξυπνη και αποτελεσματική παροχή φροντίδας [16]. Τα Big Data Analytics στον τομέα της Υγείας επιτρέπουν την ενσωμάτωση των απο-ταυτοποιημένων πληροφοριών για την υγεία, ώστε να επιτρέπονται οι δευτερεύουσες χρήσεις των δεδομένων. Επιπλέον, μέσω της αναγνώρισης των μοτίβων και της αποκρυπτογράφησης των συσχετισμών, μπορούν να διευκολύνουν την αυτόνομη λήψη αποφάσεων [17]. Στην κλινική πρακτική, τα Big Data Analytics μπορεί να βοηθήσουν στην έγκαιρη ανίχνευση της νόσου, στην ακριβή πρόβλεψη της πορείας της και στον εντοπισμό της απόκλισης από την υγιή κατάσταση, την παρουσίαση επιπλοκών, καθώς επίσης και στην ανίχνευση της απάτης. Παρέχοντας αυτές τις πληροφορίες, βοηθούν τους οργανισμούς υγειονομικής περίθαλψης να εξατομικεύουν τις προβλέψεις, να παράσχουν στοχοθετημένη θεραπεία συνυπολογίζοντας τη σχέση κόστους-αποτελεσματικότητας της περίθαλψης, να μειώσουν τη σπατάλη πόρων, καθώς και να ενθαρρύνουν τα μεμονωμένα άτομα να διατηρήσουν την υγεία τους μέσω της παροχής σχετικών συστάσεων [18]. Τα Big Data

Analytics παρέχουν την ευκαιρία ανίχνευσης γεγονότων σχετικά χαμηλής συχνότητας, τα οποία ωστόσο μπορεί να έχουν σημαντικό κλινικό αντίκτυπο. Πέραν τούτου, η ομογενοποίηση των κλινικών δεδομένων και η αποτελεσματική χρήση τους υποστηρίζουν ένα ευρύ φάσμα εφαρμογών, όπως η παρακολούθηση των ασθενειών, τα συστήματα υποστήριξης κλινικών αποφάσεων, την ατομική διαχείριση της υγειονομικής περίθαλψης, τη βελτίωση της αποτελεσματικότητας και της ποιότητας της υγειονομικής περίθαλψης, και τη μείωση του κόστους της.

2.8.3.1. Συστήματα υποστήριξης κλινικής απόφασης

Η ανίχνευση μιας νόσου από διάφορους παράγοντες ή συμπτώματα είναι ένα πρόβλημα πολλαπλών στρωμάτων που μπορεί επίσης να οδηγήσει σε ψευδείς υποθέσεις με συχνά απρόβλεπτα αποτελέσματα. Ένα σύστημα υποστήριξης κλινικής απόφασης είναι ένα πρόγραμμα υπολογιστή που περιέχει όλες τις σχετικές γνώσεις του ιατρικού τομέα σχετικά με ένα συγκεκριμένο ιατρικό πεδίο και παράγει μια διαφορική διάγνωση με βάση τα επιμέρους ευρήματα ασθενών. Ο μεγάλος όγκος και η ποικιλομορφία των δεδομένων αυτών (απεικονιστικές εξετάσεις, ιστορικό ασθενών κλπ.) καθιστούν τα big data και την ανάλυσή τους απαραίτητα στοιχεία για την ανάπτυξη αυτών των συστημάτων. Ένα σύστημα υποστήριξης κλινικής απόφασης μπορεί να είναι εξαιρετικά χρήσιμο γιατί μπορεί να βελτιώσει την προσβασιμότητα των γνώσεων των ειδικών και των ασθενών, με αποτέλεσμα τη βελτίωση της ποιότητας της διαγνωστικής διαδικασίας, την αύξηση της αποτελεσματικότητας και τη μείωση του κόστους [17].

2.8.3.2. Big data Analytics και εξατομικευμένη ιατρική

Η ανάλυση μεγάλων δεδομένων μπορεί να διαδραματίσει κρίσιμο ρόλο στην ανάπτυξη της εξατομικευμένης υγειονομικής περίθαλψης. Πολλές ασθένειες έχουν προληπτικούς παράγοντες κινδύνου ή δείκτες κινδύνου. Η βαθιά γνώση αυτών των δεδομένων των ασθενειών μπορεί να βοηθήσει στην εξατομικευμένη υγειονομική περίθαλψη και να βοηθήσει στη μείωση της σοβαρότητας των ασθενειών [19]. Ωστόσο, ο πιθανός συνδυασμός παραγόντων κινδύνου είναι τόσο περίπλοκος, είναι αδύνατο για έναν μεμονωμένο ιατρό να το αναλύσει πλήρως (σε πραγματικό χρόνο) τη στιγμή της αλληλεπίδρασης του ασθενούς. Συνεπώς, στην εξατομικευμένη υγειονομική περίθαλψη, απαιτείται υπολογιστικό και αναλυτικό πλαίσιο για να συγκεντρωθούν και να αναλυθούν μεγάλα δεδομένα, παρέχοντας βαθιά γνώση σχετικά με τις ομοιότητες και τις συνδέσεις των ασθενών και να εξαχθούν εξατομικευμένα προφίλ κινδύνου νόσησης και βαρύτητας νόσησης για κάθε μεμονωμένο ασθενή [20].

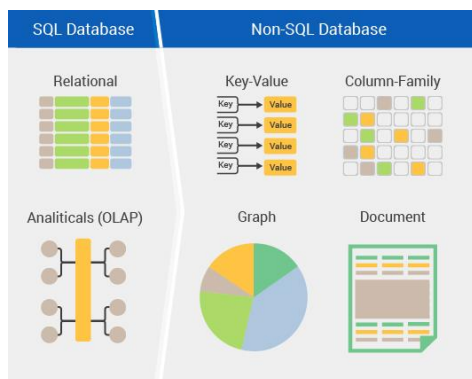
1.9 Τεχνολογίες για Big Data

Για την υποστήριξη των τεχνικών ανάλυσης δεδομένων μεγάλης κλίμακας, έχουν αναπτυχθεί αρκετές τεχνολογίες, και εξακολουθούν να αναπτύσσονται, καθώς ο κλάδος είναι αρκετά δυναμικός και εξελίσσεται με ταχύτητα. Στη συνέχεια παρουσιάζονται ορισμένες από τις τεχνολογίες που είναι χαρακτηριστικές του πεδίου.

1.9.1 Βάσεις δεδομένων NoSQL

Ο όρος NoSQL ή NotOnlySQL χαρακτηρίζει τη σχετικά νέα τεχνολογία στις βάσεις δεδομένων που έχει αναπτυχθεί, εδώ και μερικά χρόνια για την υποστήριξη δεδομένων μεγάλης κλίμακας. Σε αντίθεση με τις σχεσιακές βάσεις δεδομένων και τα ερωτήματα στη γλώσσα SQL, οι βάσεις NoSQL δεν υποστηρίζουν τη γλώσσα SQL και δεν αποθηκεύουν τα δεδομένα σε σχεσιακούς πίνακες με γραμμές και στήλες (Εικόνα 15). Τα συστήματα NoSQL, αν και δεν ακολουθούν στο σύνολό τους τα ίδια πρότυπα, μπορούμε να πούμε πως πρόκειται για μια ευρεία ομάδα συστημάτων διαχείρισης βάσεων δεδομένων που το κύριο χαρακτηριστικό τους είναι η μη τήρηση του σχεσιακού μοντέλου RDBMS (Relational Database Management System). Το βασικό στοιχείο είναι ότι τα

συστήματα NoSQL είναι προσανατολισμένα στην αποθήκευση και ανάκτηση μεγάλων όγκων δεδομένων και όχι τόσο στη δημιουργία συσχετίσεων μεταξύ τους. Είναι δε βελτιστοποιημένα ως προς αυτό και όχι τόσο ως προς τη χρονική απόκριση, καθώς το σημαντικό ζητούμενο είναι οι μεγάλοι όγκοι δεδομένων.

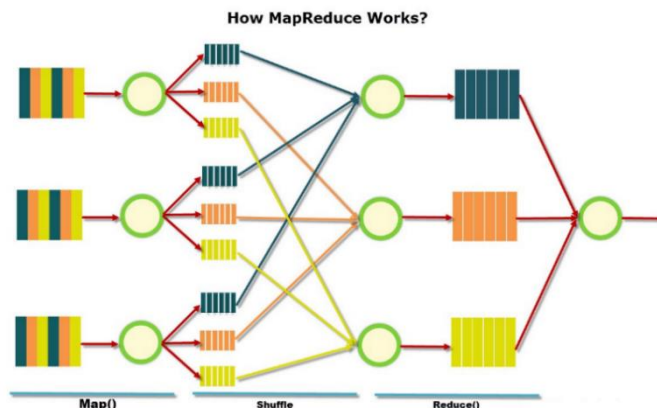


Εικόνα 15: SQL και NoSQL Βάσεις Δεδομένων

Σε αντίθεση με τα σχεσιακά συστήματα βάσεων δεδομένων, τα οποία σε μεγάλο βαθμό εγκαθίστανται σε υποδομές εξυπηρετητών με υψηλές επιδόσεις, τα συστήματα NoSQL μπορούν να υλοποιηθούν, τις περισσότερες φορές, στο υπολογιστικό νέφος (cloud) και η υλοποίησή τους έχει χαμηλότερο κόστος. Αυτό συμβαίνει διότι η αύξηση των επιδόσεων στα RDBMS επιτυγχάνεται με την προσθήκη επιπλέον μνήμης και επεξεργαστών, ενώ στα NoSQL προστίθενται κόμβοι στο δίκτυο για την αύξηση της ισχύος επεξεργασίας. Οι μεταβολές, η διαχείριση του συστήματος, καθώς και ο χρόνος διακοπών (downtime) είναι επίσης μικρότερα και λιγότερο απαιτητικά στα NoSQL από ό,τι στα RDBMS.

1.9.2 Τεχνολογία MapReduce

Στο χώρο των παραδοσιακών βάσεων δεδομένων, το σύνολο της επεξεργασίας λαμβάνει χώρα έπειτα από την πλήρη φόρτωση της πληροφορίας, χρησιμοποιώντας ειδική γλώσσα ερωτημάτων σε δομές δεδομένων, που είναι δομημένες σε υψηλό βαθμό και βελτιστοποιημένες. Σε αντίθεση με αυτό, η προσέγγιση που ξεκίνησε από την Google, και υιοθετήθηκε από αρκετές εταιρίες διαδικτύου, είναι η δημιουργία μιας ροής όπου γίνεται ανάγνωση και εγγραφή σε μορφές αρχείων που δεν είναι αυστηρά καθορισμένες, με τα ενδιάμεσα αποτελέσματα να μεταφέρονται μεταξύ επιπέδων ή βημάτων ως αρχεία, και **ο υπολογισμός να είναι κατανεμημένος μεταξύ διαφορετικών μηχανών** (Εικόνα 16). Γύρω από την προσέγγιση αυτή, που ονομάζεται **MapReduce** έχει δημιουργηθεί ένα νέο οικοσύστημα από εργαλεία και τεχνολογίες τα οποία είναι απαραίτητα για την προσέγγιση των Big Data με αυτό τον τρόπο.



Εικόνα 16: Η μέθοδος MapReduce

1.9.3 Αποθήκευση

Οι εργασίες επεξεργασίας Big Data έχουν πρόσβαση στα δεδομένα με τρόπο που δεν υποστηρίζεται από τα παραδοσιακά συστήματα αρχείων. Τα δεδομένα είναι συνήθως ομαδοποιημένα σε μεγάλα τμήματα πολλών megabytes το καθένα και τόσο η εγγραφή, όσο και η ανάγνωση, γίνονται σε μεγάλα τμήματα. Σε αυτό το περιβάλλον, η αποδοτικότητα είναι βασική προτεραιότητα σε σχέση με την οργάνωση της πληροφορίας με φιλικό τρόπο. Το μεγάλο μέγεθος αρχείων σημαίνει επίσης ότι απαιτείται η αποθήκευση με κατακευματισμένο τρόπο σε πολλές μηχανές. Ως αποτέλεσμα, έχουν εμφανιστεί αρκετές εξειδικευμένες τεχνολογίες, οι οποίες υποστηρίζουν αυτές τις ανάγκες και εξισορροπούν ορισμένα από τα χαρακτηριστικά των συστημάτων αρχείων γενικής χρήσης.

1.9.4 Οπτικοποίηση

Η παρουσίαση των δεδομένων με γραφικό τρόπο βοηθά στην επικοινωνία και κατανόηση του νοήματος που βρίσκεται πίσω από αυτά. Τα τελευταία χρόνια ο τομέας της οπτικοποίησης έχει παρουσιάσει δυναμική εξέλιξη και έχουν αναπτυχθεί αρκετά εργαλεία με πολλές δυνατότητες.

Όλες οι παραπάνω τεχνολογίες καταδεικνύουν τον πλουραλισμό που επικρατεί στον χώρο των Big Data. Ένα σημείο που αξίζει να σημειωθεί είναι ότι υπάρχει πλήθος από τεχνολογίες ανοικτού λογισμικού που υποστηρίζονται από κοινότητες προγραμματιστών και παρέχουν εξαιρετικά ισχυρές δυνατότητες. Ο εκπαιδευόμενος που ενδιαφέρεται για περαιτέρω εμβάθυνση στο τεχνολογικό τμήμα ή σε κάποιο ειδικό εργαλείο, παροτρύνεται να αναζητήσει πληροφορίες από τις αντίστοιχες ιστοσελίδες.

1.10 Οφέλη από τη χρήση των Big Data

Όπως γίνεται αντιληπτό από τα όσα αναλύθηκαν στις προηγούμενες ενότητες, οι τεχνολογίες των Big Data και Cloud Computing εμφανίζουν σημαντικά πλεονεκτήματα και δίνουν λύσεις σε μέχρι σήμερα άλυτα θέματα, ενώ ταυτόχρονα ενισχύουν σημαντικά τις ερευνητικές προσπάθειες των διαφόρων επιστημόνων. Στον κλάδο της υγείας οι εφαρμογές είναι ήδη αρκετές και η ανάγκη για υιοθέτηση τέτοιων τεχνολογιών είναι πραγματικά μεγάλη, μιας και τα δεδομένα είναι υπερβολικά πολλά και μόνον με τέτοιες τεχνολογίες μπορεί κανείς να εξορύξει την επιθυμητή γνώση που εμπεριέχουν.

Οι ίδιες οι τεχνολογίες του Cloud Computing και των Big Data ωθούν από την πλευρά τους τις εξελίξεις μιας και ανοίγουν νέους ορίζοντες για έρευνες με πολλές νέες δυνατότητες. Στον τομέα της υγείας, αναμένονται επενδύσεις σε Cloud Computing της τάξης των 64,7 δις δολαρίων το 2025 από 28,1 δις δολάρια που ανέρχονταν το 2020. Ταυτόχρονα, η **παγκόσμια αγορά για ανάλυση δεδομένων υγείας αναμένεται να φτάσει τα 42,6 δις δολάρια το 2025**, από τα 21,24 δις δολάρια που ήταν το 2019. Παράγοντες όπως η ανάγκη για περαιτέρω μείωση των εξόδων υγείας, η χρήση κινήτρων για ουσιαστική χρήση των πόρων, η τεχνολογία των Big Data και η αύξηση των επενδύσεων στον τομέα, ενισχύουν ακόμη περισσότερο την αγορά της ανάλυσης των δεδομένων υγείας. Ταυτόχρονα όμως, το κενό μεταξύ πληρωτών (π.χ. ασφαλιστικών ταμείων, κ.λπ.) και παρόχων (π.χ. νοσοκομεία, κ.λπ.), τα υψηλά ακόμη κόστη για ανάλυση δεδομένων και η έλλειψη προσωπικού με τις απαιτούμενες δεξιότητες, αντιστέκονται σε αυτή την ανάγκη της συγκεκριμένης αγοράς.

Η χρήση των τεχνικών ανάλυσης των Big Data μπορεί να βοηθήσει πραγματικά τον τομέα της υγείας και πιο συγκεκριμένα:

- **να βελτιώσει την ποιότητα των παρεχόμενων υπηρεσιών υγείας**, με τους εξής τρόπους:
 - πρόσβαση όλων των εμπλεκόμενων (ασθενών και επαγγελματιών υγείας) στην επιθυμητή γνώση, ώστε να λαμβάνουν καλύτερες αποφάσεις
 - έγκαιρη ενέργεια, αλλαγή στάσης των ασθενών και μεταβολή του τρόπου ζωής τους, ώστε να αποφευχθούν σημαντικά προβλήματα υγείας

- ορθότερη λήψη αποφάσεων, μειώνοντας έτσι τις αβλεψίες και τα ιατρικά λάθη
- συλλογή και οπτικοποίηση των πληροφοριών συμβάλλοντας στον εντοπισμό και παρακολούθηση τάσεων υγείας
- συλλογή δεδομένων από διαφορετικές πηγές, με ένα ομοιόμορφο και συγκρίσιμο τρόπο, ώστε να παρέχονται σημαντικές γνώσεις σχετικά με την εξάπλωση των ασθενειών, την πρόβλεψη και την αντιμετώπιση επιδημιών κ.ά.
- ικανότητα εξόρυξης δεδομένων από διαφορετικές πηγές ενισχύοντας την έρευνα της εύρεσης νέων θεραπειών
- εντοπισμός πληθυσμών υψηλού κινδύνου από τους παρόχους υγείας και προσφορά των απαιτούμενων υπηρεσιών πρόληψης
- παροχή αποδοτικότερων θεραπειών και άρα αύξησης της ικανοποίησης των ασθενών
- καθίσταται δυνατή η συλλογή πολλαπλών πληροφοριών και άρα αποδοτικότερη παρακολούθηση των ασθενών (σχετικά με την πορεία ασθένειας, τα σχήματα θεραπείας, τις φαρμακευτικές αγωγές κ.α.)
- δυνατότητα για παροχή εξατομικευμένων υπηρεσιών φροντίδας υγείας
- σημαντική διευκόλυνση στο σχεδιασμό νέων φαρμάκων και στη διεξαγωγή των απαιτούμενων κλινικών δοκιμών
- **να βελτιώσει την αποτελεσματικότητα και την παραγωγικότητα**, με τους εξής τρόπους:
 - μείωση του κόστους της υγειονομικής περίθαλψης, εντοπίζοντας τις πηγές που δημιουργούν ανεξέλεγκτες ή υψηλές δαπάνες
 - μείωση των επανεισαγωγών στα νοσοκομεία, μέσω της παροχής αποτελεσματικότερων θεραπειών
 - παρακολούθηση των πόρων (π.χ. φάρμακα, αναλώσιμα κ.λπ.) που καταναλώνονται
 - σύγκριση της παραγωγικότητας και της αποτελεσματικότητας των επαγγελματιών του τομέα σε σχέση με άλλους συναδέλφους τους
 - συμβολή στον εντοπισμό τυχόν αντιφάσεων στον τρόπο παροχής της περίθαλψης
 - ενίσχυση της ιατρικής βασισμένης σε γεγονότα (evidence-based medicine)
 - υπέρβαση του προβλήματος έλλειψης διαλειτουργικότητας
 - μείωση των εξόδων προμηθειών μέσω της υιοθέτησης της βέλτιστης στρατηγικής
 - βελτιστοποίηση του σχεδιασμού σε επίπεδο πολιτικής υγείας, όπως οι ανάγκες για ανθρώπινο δυναμικό, κ.λπ.

Η αυξανόμενη υιοθέτηση των ηλεκτρονικών φακέλων υγείας, των PACS, των LIS, κ.λπ., από τους παρόχους υπηρεσιών υγείας, προωθεί την ψηφιακή συλλογή των δεδομένων στο χώρο. Οι τεχνικές ανάλυσης των Big Data φαίνεται να μπορούν να βελτιώσουν το χαοτικό περιβάλλον της υγείας. Ταυτόχρονα, η νέα τάση για «ανοικτά δεδομένα» που αναπτύσσεται, ανοίγει νέους ορίζοντες. Ως «**ανοικτά δεδομένα**» ορίζονται τα δεδομένα που είναι διαθέσιμα στο κοινό, χωρίς κόστος, προκειμένου να τα χρησιμοποιήσει όπως νομίζει και να τα αναδημοσιεύσει χωρίς προβλήματα πνευματικής ιδιοκτησίας. Φυσικά, οι προκλήσεις είναι πολλές (νομικές, τεχνικές, κοινωνικές, κ.ά.), τα οφέλη ωστόσο είναι σημαντικά για τους πολίτες, τους ασθενείς και το σύστημα υγείας.

1.11 Ασφάλεια, Ιδιωτικότητα και ζητήματα Ηθικής

Η ασφάλεια και η ιδιωτικότητα των ευαίσθητων δεδομένων, όπως τα δεδομένα σχετικά με την υγεία ενός ατόμου, αποτελεί μία πρόκληση εδώ και δεκαετίες [21].

Η **ασφάλεια** ορίζεται τυπικά ως η προστασία απέναντι σε μη εξουσιοδοτημένη πρόσβαση, με ιδιαίτερη έμφαση στα χαρακτηριστικά της **διαθεσιμότητας** (availability), της **ακεραιότητας** (integrity) και της **εμπιστευτικότητας** (confidentiality) των δεδομένων. Επικεντρώνεται κυρίως στην προστασία των δεδομένων από κακόβουλες επιθέσεις και από κλοπή με σκοπό κάποιο όφελος.

Η **ιδιωτικότητα** ορίζεται συχνά ως η κατάσταση κατά την οποία η πρόσβαση στις ιδιωτικές πληροφορίες ενός συγκεκριμένου ατόμου μπορεί να ελεγχθεί από το ίδιο το άτομο, ακόμα και όταν κάποιο τρίτο μέρος έχει συλλέξει αυτήν την πληροφορία. Επικεντρώνεται στη χρήση και διαχείριση των προσωπικών δεδομένων ενός ατόμου, έτσι ώστε να διασφαλίζεται ότι τα δεδομένα αυτά συλλέγονται, διαμοιράζονται και χρησιμοποιούνται με σωστό τρόπο.

Παρόλο που η ασφάλεια είναι ζωτικής σημασίας για την προστασία των δεδομένων, δεν είναι σίγουρα επαρκής για να εξασφαλίσει την ιδιωτικότητα.

Σήμερα, χρησιμοποιούνται διάφορες τεχνολογίες για την προάσπιση της ασφάλειας και της ιδιωτικότητας των Big Data σχετικών με την υγειονομική περίθαλψη [22] [23].

Οι πιο διαδεδομένες **τεχνολογίες ασφαλείας** ακολουθούν στη συνέχεια:

- **Πιστοποίηση**, ή επαλήθευση ταυτότητας, (authentication) ονομάζεται η διαδικασία της επιβεβαίωσης και του καθορισμού κάποιου ατόμου ή αντικειμένου σαν αυθεντικού, δηλαδή η χρήση διαπιστευτηρίων για την επιβεβαίωση της ταυτότητάς του. Η πιστοποίηση βασίζεται σε έναν ή περισσότερους παράγοντες πιστοποίησης.
- **Κρυπτογράφηση** (encryption) ονομάζεται η διαδικασία μετασχηματισμού ενός μηνύματος σε μία ακατανόητη μορφή, με τη χρήση κάποιου κρυπτογραφικού αλγορίθμου, έτσι ώστε να μην μπορεί να διαβαστεί από κανέναν εκτός του νόμιμου παραλήπτη. Αποτελεί ένα αποτελεσματικό μέσο για την πρόληψη της μη εξουσιοδοτημένης πρόσβασης σε ευαίσθητα δεδομένα.
- Η **αλλοίωση**, ή απόκρυψη, **δεδομένων** (data masking) αντικαθιστά τα ευαίσθητα δεδομένα με μια μη αναγνωρίσιμη τιμή. Στην ουσία, δεν αποτελεί μια τεχνική κρυπτογράφησης καθώς η αρχική τιμή δεν μπορεί να επιστραφεί από την αλλοιωμένη τιμή. Χρησιμοποιεί μια στρατηγική από-ταυτοποίησης ή αλλοίωσης προσωπικών αναγνωριστικών (identifiers) (όπως π.χ. το όνομα και ο αριθμός κοινωνικής ασφάλισης) σε συνδυασμό με την «καταστολή» ή τη γενίκευση των ψευδο-αναγνωριστικών (quasi-identifiers) (όπως π.χ. η ημερομηνία γέννησης και οι ταχυδρομικοί κώδικες). Έτσι, η αλλοίωση δεδομένων είναι μια από τις πιο δημοφιλείς προσεγγίσεις για την ανωνυμοποίηση ζωντανής μετάδοσης δεδομένων.
- Ο **έλεγχος πρόσβασης** (access control), ο οποίος έπεται της πιστοποίησης, αποτελεί μια πολιτική ελέγχου πρόσβασης, που βασίζεται σε δικαιώματα και προνόμια που παρέχονται στον εκάστοτε επαγγελματία υγείας από ασθενείς, ή από επιτρεπόμενα τρίτα μέρη, στους οποίους ανήκει το αρχείο στο οποίο επιχειρείται προσπέλαση.
- Η **συνεχής παρακολούθηση** και ο **συνεχής έλεγχος** (monitoring and auditing). Η παρακολούθηση ασφάλειας συλλέγει και διερευνά περιστατικά ασφαλείας με σκοπό την ανίχνευση εισβολών. Ο έλεγχος πραγματοποιεί καταγραφή, κατά χρονολογική σειρά, όλων των δραστηριοτήτων ενός χρήστη του συστήματος υγειονομικής περίθαλψης. Οι τεχνικές αυτές αποτελούν δύο προαιρετικά μέτρα ασφάλειας για τη μέτρηση και την εξασφάλιση της ασφάλειας ενός συστήματος υγειονομικής περίθαλψης.

Στη συνέχεια παρατίθενται κάποιες ενδεικτικές **μέθοδοι για την προάσπιση της ιδιωτικότητας** κατά τη χρήση των Big Data:

- **Απο-ταυτοποίηση** (de-identification) είναι μια παραδοσιακή μέθοδος για την αποτροπή της αποκάλυψης εμπιστευτικών πληροφοριών, διαγράφοντας οποιαδήποτε πληροφορία μπορεί να ταυτοποιήσει τον ασθενή. Η

μέθοδος αυτή πραγματοποιείται συνήθως με την αφαίρεση συγκεκριμένων αναγνωριστικών του ασθενούς (identifiers). Για καταστεί η μέθοδος πιο αποδοτική, έχουν αναπτυχθεί αποδοτικοί αλγόριθμοι προστασίας της ιδιωτικότητας, με σκοπό την άμβλυνση του κινδύνου επανα-ταυτοποίησης, όπως η k-ανωνυμία (k-anonymity), η l-διαφορετικότητα (l-diversity) και η t-εγγύτητα (t-closeness).

- Το **υβριδικό μοντέλο εκτέλεσης** (HybrEx) είναι ένα μοντέλο για την εμπιστευτικότητα και την ιδιωτικότητα στο cloud computing. Όταν το μοντέλο ενσωματώνεται στο ιδιωτικό σύννεφο ενός οργανισμού, χρησιμοποιεί δημόσια σύννεφα μόνο για ασφαλείς διαδικασίες, δηλαδή χρησιμοποιεί δημόσια σύννεφα μόνο για μη-ευαίσθητα ή δημόσια δεδομένα, ενώ για τα προσωπικά και ευαίσθητα δεδομένα, χρησιμοποιεί το ιδιωτικό σύννεφο. Το μοντέλο λαμβάνει υπόψη το είδος των δεδομένων πριν από την εκτέλεση οποιασδήποτε εργασίας.
- Η **ανωνυμοποίηση βάσει ταυτότητας** (identity based anonymization) αποτελεί ένα είδος εξυγίανσης πληροφοριών, με σκοπό την προστασία της ιδιωτικότητας. Είναι η διαδικασία κρυπτογράφησης ή αφαίρεσης προσωπικών πληροφοριών ταυτοποίησης από σύνολα δεδομένων, έτσι ώστε τα υποκείμενα που περιγράφονται από τα δεδομένα να παραμένουν ανώνυμα. Η βασική δυσκολία που έγκειται σε αυτή την τεχνική αποτελεί ο συνδυασμός της ανωνυμοποίησης, της προστασίας ιδιωτικότητας και των τεχνικών Big Data στην ανάλυση χρήσης δεδομένων με την ταυτόχρονη προστασία των ταυτοτήτων.

Στις μέρες μας, παρά την ταχεία πρόοδο και την ελκυστικότητα την οποία ασκεί η αξιοποίηση των Big Data, η νομική και θεσμική προσαρμογή υστερεί. Η κατάσταση αυτή δημιουργεί προϋποθέσεις πρόκλησης βλάβης στην ιδιωτικότητα και ενδεχομένως στα προσωπικά δεδομένα υγείας του κάθε ατόμου.



Η κάθε χώρα έχει διαφορετικές πολιτικές και νόμους για την προστασία της ιδιωτικότητας. Η **Ευρωπαϊκή Οδηγία για την Επεξεργασία δεδομένων προσωπικού χαρακτήρα** (Data Protection Directive – Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data) εγκρίθηκε το 1995 και ρυθμίζει την επεξεργασία των δεδομένων προσωπικού χαρακτήρα στην Ευρωπαϊκή Ένωση [24]. Αποτελεί σημαντική συνιστώσα του Νόμου της ΕΕ για την ιδιωτικότητα και τα ανθρώπινα δικαιώματα. Ο **Γενικός Κανονισμός για την Προστασία Δεδομένων** (General Data Protection Regulation – Regulation EU 2016/679), που εγκρίθηκε τον Απρίλιο του 2016, θα αντικαταστήσει την παραπάνω οδηγία και θα τεθεί σε εφαρμογή από τις 25 Μαΐου 2018 [25].

Όσον αφορά την Ελλάδα, η επεξεργασία των Big Data πρέπει να διασφαλίζει το **σεβασμό των προσωπικών δεδομένων υγείας**, καθώς συνιστούν ευαίσθητα δεδομένα που απολαμβάνουν αυξημένη προστασία. Σύμφωνα με την ισχύουσα νομοθεσία (ν. 2472/1997), η επεξεργασία τους επιτρέπεται κατ' εξαίρεση όταν είναι αναγκαία για την ιατρική πρόληψη, διάγνωση, περίθαλψη, καθώς και την προστασία της δημόσιας υγείας, με την προϋπόθεση ότι εκτελείται από επαγγελματίες της υγείας, που επιπλέον υπόκειται σε καθήκον εχεμύθειας. Η επεξεργασία επιτρέπεται επίσης για ερευνητικούς και επιστημονικούς σκοπούς, υπό τον όρο ότι **τα δεδομένα ανωνυμοποιούνται** και λαμβάνονται όλα τα απαραίτητα μέτρα για την προστασία των δικαιωμάτων των προσώπων [26].

Η **ανωνυμοποίηση** των δεδομένων υγείας είναι ζήτημα πρωταρχικής σημασίας στη χρήση και επεξεργασία τους, για την προστασία της ιδιωτικότητας του υποκειμένου τους. Στο σημείο αυτό, είναι απαραίτητο να διακριθούν οι όροι «ανωνυμοποιημένα» και «ψευδοανωνυμοποιημένα» δεδομένα:

- Στην **ανωνυμοποίηση** των δεδομένων αφαιρούνται όλα εκείνα τα στοιχεία που οδηγούν σε ταυτοποίηση των υποκειμένων τους, είτε ανεξάρτητα είτε σε συνδυασμό με άλλα δεδομένα.
- Στην **ψευδοανωνυμοποίηση**, αντίθετα, τα στοιχεία ταυτοποίησης των προσώπων αντικαθίστανται με μια τιμή (ή αλλιώς ένα ψευδώνυμο), ωστόσο ο υπεύθυνος επεξεργασίας των δεδομένων έχει το «κλειδί» με το οποίο είναι δυνατόν, ακολουθώντας αντίστροφη διαδικασία, να ταυτοποιήσει τα υποκείμενα των πληροφοριών.

Παράλληλα με την ανωνυμοποίηση, χρησιμοποιούνται διάφορες άλλες μέθοδοι για την προστασία του απορρήτου των πληροφοριών υγείας, όπως αυτές που αναφέρθηκαν προηγουμένως (π.χ. κρυπτογράφηση, αλλοίωση στοιχείων κ.α.).

Τα ζητήματα που τίθενται με τις συλλογές των Big Data, προκύπτουν κυρίως από την **αλλαγή του σκοπού της αρχικής συλλογής και επεξεργασίας**, όταν δηλαδή τα δεδομένα χρησιμοποιούνται περαιτέρω για άλλο σκοπό («δευτερεύουσα χρήση»). Επίσης, καθώς αντιστοιχίζονται μοναδικά, μεγιστοποιούν την πληροφορία, υπερβαίνοντας τον αρχικό σκοπό της επεξεργασίας, τον οποίο καλύπτει η συναίνεση του υποκειμένου.

Η Εθνική Επιτροπή Βιοηθικής εντοπίζει τρία βασικά ηθικά ζητήματα ως προς τη δημιουργία και τη λειτουργία βάσεων Big Data, που πρέπει να αντιμετωπισθούν:

- **Το ζήτημα της συναίνεσης των υποκειμένων**

Ο τρόπος συναίνεσης των υποκειμένων ευαίσθητων δεδομένων υγείας είναι κρίσιμος, κατά το μέτρο που οι συλλογές Big Data είναι «δευτερογενείς», συγκροτούνται δηλαδή από ήδη υπάρχουσες μικρότερες συλλογές. Ακόμη και αν για τη δημιουργία αυτών των τελευταίων έχει δοθεί έγκυρη συναίνεση από τα υποκείμενα των δεδομένων, δεν είναι αυτονόητο ότι η συναίνεση αυτή καλύπτει και τις «δευτερογενείς» συλλογές, τους μελλοντικούς σκοπούς και τις χρήσεις. Απαιτείται, επομένως, είτε νέα ειδική συναίνεση γι' αυτές, είτε εναλλακτικοί τρόποι («γενική» συναίνεση, «εικαζόμενη» συναίνεση), με τον όρο πάντως να μη θίγεται η αυτονομία των προσώπων. Ο Γενικός Κανονισμός για την Προστασία Δεδομένων, που προαναφέρθηκε, κάνει μνεία για τη δυνατότητα γενικής συναίνεσης, όταν μελλοντικοί ερευνητικοί σκοποί συλλογής και επεξεργασίας δεν μπορούν να προκαθορισθούν

- **Το ζήτημα της εμπιστευτικότητας**

Η Επιτροπή επισημαίνει τον αυξημένο κίνδυνο διαρροών, κατά την επεξεργασία ευαίσθητων δεδομένων σε μεγάλες συλλογές, ακριβώς λόγω του όγκου τους. Πρόβλημα, επίσης, αποτελεί η δυνατότητα «βαθιάς εξόρυξης δεδομένων» (deep data mining) (μια διαδικασία που μετατρέπει πρωτογενή δεδομένα σε πληροφορία) που, μέσω του συνδυασμού των διαθέσιμων δεδομένων από διάφορες πηγές, μπορεί να οδηγήσει στην αποκάλυψη της ταυτότητας υποκειμένων, ακόμη και σε ανώνυμα δεδομένα, λόγω της ευχέρειας πλήθους συσχετίσεων.

Επισημαίνεται, ότι ο Γενικός Κανονισμός για την Προστασία Δεδομένων εισάγει το «δικαίωμα στη λήθη», που αποτελεί την έσχατη δυνατότητα προστασίας του προσώπου από αθέμιτη επεξεργασία ευαίσθητων πληροφοριών του. Με βάση το δικαίωμα αυτό, καθένας μπορεί να απαιτήσει την πλήρη απάλειψη όλων των δεδομένων του από μια συλλογή, ανακτώντας τον προσωπικό έλεγχο των ευαίσθητων πληροφοριών του. Δεν πρέπει, πάντως, να παραβλέπεται, ότι η άσκηση του «δικαιώματος στη λήθη» ενδέχεται να στερήσει από το πρόσωπο άλλα σημαντικά δικαιώματα, ιδίως στον τομέα της κοινωνικής ασφάλισης ή της παροχής υπηρεσιών υγείας, όπου η «εμφάνιση» του ατομικού «προφίλ» είναι αναγκαία.

- **Τα «τυχαία ευρήματα»**

Το τρίτο, τέλος, ζήτημα αφορά τον χειρισμό «τυχαίων ευρημάτων» που αφορούν την υγεία του προσώπου. Ο όρος «**τυχαία ευρήματα**» αναφέρεται στα ευρήματα μιας ιατρικής εξέτασης ή μιας έρευνας που έχουν (πιθανή) σημασία για την υγεία ή την αναπαραγωγή του ατόμου και ανακαλύπτονται συμπτωματικά, χωρίς να εμπίπτουν στον αρχικό σκοπό της εξέτασης ή της έρευνας. Τυχαία ευρήματα κατά την επεξεργασία ψευδοανωνυμοποιημένων δεδομένων (όπου είναι δυνατόν να αποκαλυφθεί η ταυτότητα του προσώπου), είναι δυνατόν να προκύπτουν με πολύ μεγαλύτερη συχνότητα από την επεξεργασία δεδομένων σε μεγάλες συλλογές.

Για το λόγο αυτό απαιτείται συγκεκριμένη πολιτική χειρισμού τους. Είναι κρίσιμο, πάντως, να τονισθεί ότι ο χειρισμός τυχαίων ευρημάτων εξαρτάται από την ειδική συναίνεση του προσώπου, που πρέπει να ενημερώνεται προληπτικά για τον σκοπό αυτόν.

Με βάση τα παραπάνω, είναι πολύ σημαντική **η λήψη μέτρων που θα εξασφαλίζουν μια ορθολογική αξιοποίηση των συλλογών Big Data στη χώρα μας**. Τέτοια μέτρα δύναται να αποτελέσουν η γενική συναίνεση του κάθε ατόμου (που προσδιορίζει την περιοχή πιθανών μελλοντικών χρήσεων), η ανωνυμοποίηση των δεδομένων όπου είναι δυνατή, η χρήση όλων των διαθέσιμων τεχνικών προστασίας των δεδομένων (συμπεριλαμβανομένης της πρόβλεψης διαχείρισης διαρροής πληροφοριών). Τέλος, η κατάρτιση κωδικών δεοντολογίας για τη συλλογή, πρόσβαση, διαχείριση και επεξεργασία των Big Data από Επιτροπές Δεοντολογίας της Έρευνας σε φορείς που χρησιμοποιούν Big Data για ερευνητικούς σκοπούς κρίνεται και αυτή απαραίτητη. Είναι προφανές πως η μαζική συλλογή δεδομένων και η ανάλυσή τους μπορεί να δημιουργήσει το αίσθημα της συνεχούς παρακολούθησης και ν' αποθαρρύνει τους ανθρώπους από τη χρήση νέων τεχνολογιών, άρα οι προδιαγραφές ασφαλείας και προάσπισης της ιδιωτικότητας θα πρέπει να είναι αδιαπραγμάτευτες.

1.12 Επίλογος – Συμπεράσματα

Ζούμε στην εποχή των δεδομένων, όπου απίστευτα μεγάλες ποσότητες από αυτά παράγονται κάθε δευτερόλεπτο γύρω μας. Η ανάγκη για αξιοποίηση των διάσπαρτων και συνεχώς παραγόμενων δεδομένων είναι αναντίρρητη. Η δυνατότητα σύγκρισης και ανάλυσής τους, θα μπορούσε να δώσει μεγάλη ώθηση στις καινοτομίες σε κάθε κλάδο. Για το λόγο αυτό, τα τελευταία χρόνια, οι τεχνικές και τεχνολογίες αξιοποίησης των Big Data απασχολούν πολλούς επιστημονικούς κλάδους. Είναι προφανές πως ο χώρος των Big Data είναι δυναμικός και εξελίσσεται συνεχώς με την πάροδο του χρόνου. Έτσι, τα Big Data που προσδιορίζονται έτσι σήμερα δε θα ταυτίζονται με αυτά που θα προσδιορίζονται έτσι έπειτα από μερικά έτη. Επίσης, τα Big Data ενός τομέα δεν περιέχουν αναγκαστικά τα ίδια χαρακτηριστικά με αυτά ενός διαφορετικού πεδίου.

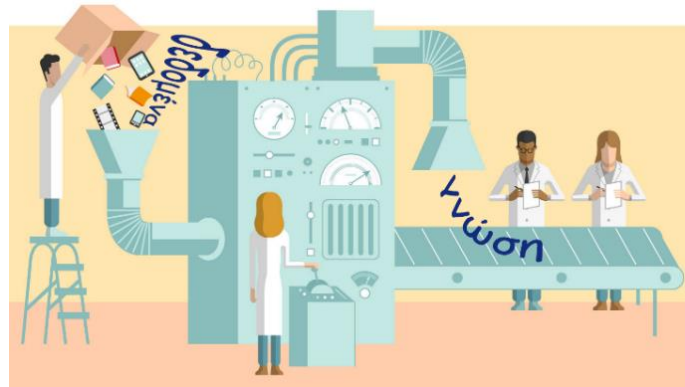
Ειδικότερα, στο χώρο της υγείας, για την αντιμετώπιση των σύγχρονων προκλήσεων που εγείρονται, απαιτούνται ριζικές αλλαγές σε επίπεδο χάραξης στρατηγικών και πολιτικών υγείας, με την ταυτόχρονη υιοθέτηση των τελευταίων τεχνολογικών λύσεων. Τα συστήματα υγειονομικής περίθαλψης πρέπει να αυξήσουν την αποδοτικότητά τους και παράλληλα να μειώσουν δραστικά τα κόστη για να καταστούν βιώσιμα και, κατά το δυνατόν, αποδοτικότερα. Επίσης, η ενδυνάμωση των ασθενών και η συμμετοχή τους στην διατήρηση της υγείας τους είναι κι αυτή εξαιρετικά σημαντική. Είναι επιτακτική πλέον η ανάγκη για τη στροφή προς ένα πιο «ασθενοκεντρικό» σύστημα περίθαλψης. Έτσι, η παρακολούθηση υγείων πολιτών και ασθενών σε πραγματικό χρόνο, μπορεί να επιτρέψει τη βελτίωση των παρεχόμενων υπηρεσιών φροντίδας υγείας, με την ταυτόχρονη μείωση των εξόδων για το ίδιο το σύστημα υγείας.

Οι νέες τεχνολογίες (όπως eHealth και mHealth) επιτρέπουν πλέον την υπερπήδηση του φράγματος του χώρου και του χρόνου και προσφέρουν νέες δυνατότητες. Οι πληροφορίες θα ακολουθούν τον ασθενή όπου και αν βρίσκεται. Η παροχή εξατομικευμένης φροντίδας έγκειται κατά βάση στις διαθέσιμες πληροφορίες, ενώ η δημιουργία νέων θεραπειών και φαρμάκων βασίζονται στη μελέτη των διαθέσιμων, αλλά διάσπαρτων και ετερογενών δεδομένων υγείας. Η εφαρμογή επομένως τεχνικών και τεχνολογιών Big Data, που ενσωματώνουν και αξιοποιούν όλα αυτά τα δεδομένα, μπορεί να αποτελέσει το κλειδί για το επόμενο μεγάλο βήμα. Ο συνδυασμός Big Data και Cloud Computing μπορεί να μας φέρει ένα βήμα πιο κοντά στον επιθυμητό στόχο, με πολλά οφέλη για όλους τους εμπλεκόμενους στο χώρο της υγείας, από ασθενείς και επαγγελματίες υγείας μέχρι ασφαλιστικά ταμεία, αρκεί να υπάρξει ξεκάθαρη στρατηγική.

Προφανώς, εκτός από τα οφέλη, υπάρχουν και σημαντικοί περιορισμοί. Χαρακτηριστικά παραδείγματα αποτελούν τα θέματα ιδιωτικότητας και προστασίας προσωπικών δεδομένων, η διαλειτουργικότητα των

συστημάτων και η ετερογένεια των δεδομένων. Όσο επίσης αυξάνει ο όγκος των δεδομένων, τόσο πιο δύσκολο είναι να τα αντιμετωπίσει κανείς κατάλληλα. Για το λόγο αυτό η ορθή χρήση των δεδομένων με ασφαλή τρόπο, χωρίς διαπραγμάτευση στην ποιότητα των παρεχόμενων υπηρεσιών, αποτελεί σήμερα πραγματική πρόκληση.

Ο επιθυμητός σκοπός είναι να καταφέρουμε ν' αξιοποιούμε στο έπακρο αυτόν τον πλούτο δεδομένων για να εξάγουμε πολύτιμες πληροφορίες και στη συνέχεια γνώσεις, ώστε να λαμβάνουμε κάθε φορά τις βέλτιστες αποφάσεις (Εικόνα 18).



Εικόνα 18: Από τα δεδομένα στη γνώση

3. Εργαστηριακό μέρος

3.1 Η επιστήμη των δεδομένων και η γλώσσα R

Η επεξεργασία, ανάλυση και οπτικοποίηση των Big Data εντάσσεται σε έναν ευρύτερο επιστημονικό τομέα, ο οποίος καλείται **Επιστήμη Δεδομένων (Data Science)**. Η Επιστήμη των Δεδομένων είναι ένα από τους ταχύτερα αναπτυσσόμενους κλάδους της επιστήμης και συνδυάζει αρκετούς επιστημονικούς τομείς, κυρίως, αλλά όχι αποκλειστικά, αυτούς της πληροφορικής και των μαθηματικών. Δημοφιλείς προγραμματιστικές γλώσσες στον τομέα της Επιστήμης Δεδομένων αποτελούν οι Python, R, Julia και Scala. Οι περισσότερες από αυτές είναι χτισμένες για τον χειρισμό μεγάλων ποσοτήτων αριθμητικών δεδομένων, με σταθερά πακέτα που μπορούν να χρησιμοποιηθούν εύκολα και γρήγορα για αναλύσεις. Αυτές οι γλώσσες αυξάνουν την σημαντικότητά τους όσο τα δεδομένα που συλλέγουν οι κυβερνήσεις, οργανισμοί, εταιρείες, επιστήμονες και άλλοι πολλοί φορείς αυξάνονται σε όγκο και σε ποικιλία.

Η **R** είναι μια από τις πιο διαδεδομένες γλώσσες προγραμματισμού στην Επιστήμη Δεδομένων και έχει δημιουργηθεί κυρίως για στατιστική ανάλυση και γραφικά. Τα κύρια πλεονεκτήματα της R είναι το γεγονός ότι η R είναι **ελεύθερο λογισμικό** και ότι υπάρχει πολλή βοήθεια διαθέσιμη στο διαδίκτυο. Αυτός είναι ο λόγος για τον οποίο βρίσκεται στην αιχμή του συγκεκριμένου τομέα. Η λειτουργικότητα της βασικής γλώσσας R μπορεί να επεκταθεί με βιβλιοθήκες που ονομάζονται **πακέτα (packages)**. Ο πιο δημοφιλής χώρος διαμοιρασμού πακέτων είναι το **Comprehensive R Archive Network (CRAN)**. Αυτή τη στιγμή περιέχει πάνω από 17.000 πακέτα. Η γλώσσα R θεωρείται πλέον βασικό προσόν για την αναζήτηση θέσης εργασίας που να σχετίζεται με την επεξεργασία και ανάλυση δεδομένων.

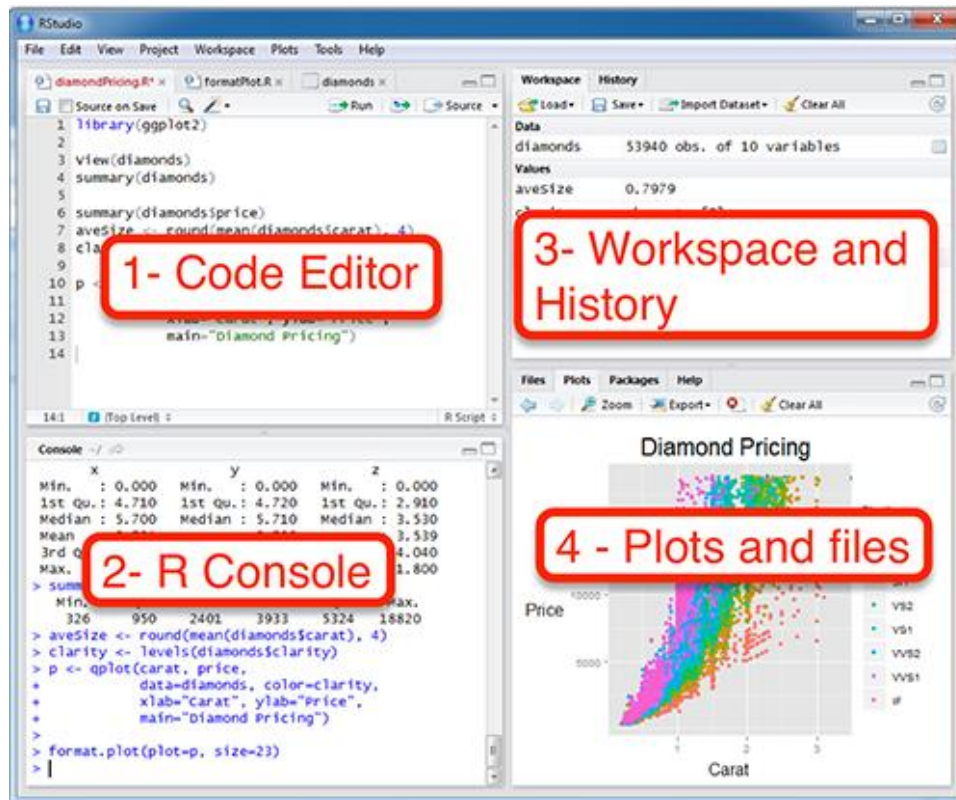
Στόχος της συγκεκριμένης εργαστηριακής άσκησης είναι η εξοικείωση των φοιτητών με τη γλώσσα προγραμματισμού R καθώς και με το ολοκληρωμένο περιβάλλον ανάπτυξης RStudio. Απαραίτητες προϋποθέσεις για την πραγματοποίηση της άσκησης είναι οι εξής:

- Εγκατάσταση της R, από τον ιστότοπο: <https://cran.r-project.org/>
- Εγκατάσταση του RStudio, από τον ιστότοπο: <https://rstudio.com/products/rstudio/download/>

Αφού ολοκληρώσετε τα δύο παραπάνω βήματα, μπορείτε να ανοίξετε το περιβάλλον του RStudio για να ξεκινήσετε να υλοποιείτε τα βήματα των επόμενων ενοτήτων.

3.2 Η διάταξη του RStudio

Η διεπαφή του RStudio αποτελείται από τέσσερα βασικά παράθυρα:



Εικόνα 19: Τα παράθυρα του RStudio

1. **Πάνω αριστερά: παράθυρο επεξεργαστή κειμένου / σεναρίων** (code editor). Εδώ μπορούν να υποστούν επεξεργασία και να σωθούν σύνολα από εντολές (σενάρια). Όταν δεν υπάρχει αυτό το παράθυρο, μπορείτε να το ανοίξετε μέσω της διαδρομής File -> New -> R script. Η απλή πληκτρολόγηση μιας εντολής στο παράθυρο του επεξεργαστή δεν είναι αρκετή, πρέπει επίσης να πάει και στο παράθυρο εντολών πριν η R μπορέσει να εκτελέσει την εντολή αυτή. Εάν θέλετε να τρέξετε μία γραμμή από το παράθυρο σεναρίων (ή και ολόκληρο το σενάριο), μπορείτε να κάνετε κλικ στο Run ή να πατήσετε τα πλήκτρα CTRL+ENTER, ώστε να τη στείλετε στο παράθυρο εντολών.
2. **Κάτω αριστερά: παράθυρο κονσόλας / εντολών** (R console). Εδώ μπορείτε να εισάγετε απλές εντολές μετά το σύμβολο υποβολής ">" και η R στη συνέχεια θα εκτελέσει την εντολή σας. Αυτό είναι το πιο σημαντικό παράθυρο, επειδή στην πραγματικότητα εκεί "τρέχει" η R.
3. **Πάνω δεξιά: χώρος εργασίας / ιστορικό** (workspace and history). Στο παράθυρο του χώρου εργασίας μπορείτε να δείτε ποια δεδομένα και ποιες τιμές έχει η R στη μνήμη της. Μπορείτε να δείτε και να επεξεργαστείτε τις τιμές κάνοντας κλικ πάνω τους. Το παράθυρο του ιστορικού δείχνει το τι έχει πληκτρολογηθεί παλιότερα.
4. **Κάτω δεξιά: αρχεία / γραφικές παραστάσεις / πακέτα / βοήθεια** (plots and files). Από εδώ μπορείτε να ανοίξετε αρχεία, να δείτε γραφικές παραστάσεις (και προηγούμενες γραφικές παραστάσεις, επίσης), να εγκαταστήσετε και να φορτώσετε πακέτα ή να χρησιμοποιήσετε τη λειτουργία της βοήθειας.

Μπορείτε να αλλάξετε το μέγεθος των παραθύρων σέρνοντας τα γκρίζα διαχωριστικά μεταξύ των παραθύρων.

3.3 Κατάλογος εργασίας

Ο **κατάλογος εργασίας** σας (working directory) είναι ο φάκελος του υπολογιστή σας μέσα στον οποίο εργάζεστε. Όταν ζητήσετε από την R να ανοίξει ένα συγκεκριμένο αρχείο, αυτή θα κοιτάξει πρώτα στον κατάλογο εργασίας γι' αυτό το αρχείο, και όταν πείτε στην R να αποθηκεύσει ένα αρχείο δεδομένων ή ένα γράφημα, αυτή θα το αποθηκεύσει στον κατάλογο εργασίας.

Πριν ξεκινήσετε να εργάζεστε, παρακαλούμε θέστε τον κατάλογο εργασίας σας εκεί που έχετε ή εκεί που πρέπει να αποθηκεύονται όλα τα δεδομένα σας και τα αρχεία σεναρίων.

Εισάγετε στο παράθυρο εντολών το εξής: `setwd("directoryname")`. Για παράδειγμα:

```
> setwd("C:/Users/George/Documents/R")
```

Σιγουρευτείτε ότι οι μπάρες είναι πλάγιες μπάρες (/) και ότι δεν έχετε ξεχάσει τις αποστρόφους. Η R διακρίνει τα πεζά από τα κεφαλαία γράμματα, οπότε βεβαιωθείτε ότι γράφετε με κεφαλαία εκεί που απαιτείται.

3.4 Βιβλιοθήκες

Η R μπορεί να κάνει πολλές στατιστικές αναλύσεις και αναλύσεις δεδομένων. Αυτές είναι οργανωμένες στα λεγόμενα **πακέτα ή βιβλιοθήκες**. Με την τυπική εγκατάσταση, εγκαθίστανται και τα περισσότερα συνήθη πακέτα.

Για να δείτε μια λίστα με όλα τα εγκατεστημένα πακέτα, πηγαίνετε στο παράθυρο πακέτων ή πληκτρολογήστε `library()` στο παράθυρο κονσόλας. Εάν το τετραγωνάκι μπροστά από το όνομα του πακέτου είναι σημειωμένο, το πακέτο φορτώνεται (ενεργοποιείται) και τότε μπορεί να χρησιμοποιηθεί.

Υπάρχουν πολλά επιπλέον διαθέσιμα πακέτα στον ιστότοπο της R. Εάν θέλετε να εγκαταστήσετε και να χρησιμοποιήσετε ένα πακέτο (για παράδειγμα, το πακέτο που λέγεται “geometry”), τότε πρέπει να:

- Εγκαταστήσετε το πακέτο: κάντε κλικ στο `install packages` στο παράθυρο πακέτων και πληκτρολογήστε `geometry` ή εισάγετε την εντολή `install.packages("geometry")` στο παράθυρο εντολών.
- Φορτώστε το πακέτο: σημειώστε το κουτάκι μπροστά από το `geometry` ή εισάγετε την εντολή `library("geometry")` στο παράθυρο εντολών.

3.5 Εντολές στην R

Αρχικά, η R μπορεί να χρησιμοποιηθεί ως απλή αριθμομηχανή. Μπορείτε απλά να εισάγετε την εξίσωση που θέλετε στο παράθυρο εντολών μετά από το “>”, π.χ.:

```
> 10 + 2^3
```

και η R θα σας δώσει την απάντηση:

```
[1] 18
```

Μπορείτε επίσης να δώσετε σε αριθμούς ένα όνομα. Κάνοντάς το, αυτοί μετατρέπονται στις λεγόμενες μεταβλητές, οι οποίες μπορούν να χρησιμοποιηθούν αργότερα. Για παράδειγμα, μπορείτε να πληκτρολογήσετε στο παράθυρο εντολών το εξής:

```
> a = 4
```

Μπορείτε να δείτε ότι το `a` εμφανίζεται στο παράθυρο του χώρου εργασίας, κάτι το οποίο σημαίνει ότι η R πλέον “θυμάται” τι είναι το `a`. Μπορείτε επίσης να ζητήσετε από την R να σας “πει” τι είναι το `a` (απλά πατήστε `a` ENTER στο παράθυρο εντολών):

```
> a
```

```
[1] 4
```

ή μπορείτε να κάνετε υπολογισμούς με το `a`:

```
> a * 5
[1] 20
```

Εάν προσδιορίσετε ξανά το `a`, η R θα “ξεχάσει” την τιμή που είχε αυτό πριν. Μπορείτε επίσης να αναθέσετε μια νέα τιμή στο `a` χρησιμοποιώντας την παλιά.

```
> a = a + 10
> a
[1] 14
```

Για να απομακρύνετε όλες τις μεταβλητές από τη μνήμη της R, πληκτρολογήστε:

```
> rm(list=ls())
```

ή κάντε κλικ στο “clear all” στο παράθυρο του χώρου εργασίας. Μπορείτε να δείτε ότι τότε το RStudio αδειάζει το παράθυρο του χώρου εργασίας. Εάν θέλετε να απομακρύνετε μόνο τη μεταβλητή `a`, μπορείτε να πληκτρολογήσετε `rm(a)`.

3.6 Δομές δεδομένων

Όπως και πολλά άλλα προγράμματα, έτσι κι η R οργανώνει τους αριθμούς σε **βαθμωτούς** (απλοί αριθμοί – μηδενικής διάστασης), σε **διανύσματα** (μια ακολουθία αριθμών – μονοδιάστατα), σε **μητρώα** (όπως ένας πίνακας – δισδιάστατα), σε **πλαίσια δεδομένων** ή **λίστες**. Το `a` που ορίσατε πριν ήταν ένας βαθμωτός αριθμός.

3.6.1 Διανύσματα (Vectors)

Για να ορίσετε ένα διάνυσμα με τους αριθμούς 3, 4 και 5, χρειάζεστε τη συνάρτηση `c`, το όνομα της οποίας είναι το αρχικό του ρήματος “concatenate” (στα ελληνικά σημαίνει «συνενώνω»).

```
> b=c(3,4,5)
```

Πιο αναλυτικά, ακολουθούν μερικές ενδεικτικές εντολές και πράξεις με τα διανύσματα, ώστε να γίνει πιο ξεκάθαρη η χρήση τους.

Ένα διάνυσμα `vec1` δημιουργείται ρητά από τη συνάρτηση συνένωσης `c()`, την οποία έχουμε αναφέρει παραπάνω. Με την πληκτρολόγηση του ονόματος του διανύσματος και το πάτημα του ENTER εμφανίζεται ολόκληρο το διάνυσμα.

```
> vec1 = c(1,4,6,8,10)
> vec1
[1] 1 4 6 8 10
```

Τα στοιχεία των διανυσμάτων μπορούν να προσπελαστούν μέσω της πρότυπης ευρετηρίασης `[i]`.

```
> vec1[5]
[1] 10
```

Ένα από τα στοιχεία του διανύσματος μπορεί να αντικατασταθεί με ένα νέο αριθμό.

```
> vec1[3] = 12
> vec1
[1] 1 4 12 8 10
```

Στη συνέχεια, βλέπουμε ακόμα ένα χρήσιμο τρόπο δημιουργίας ενός διανύσματος: τη συνάρτηση `seq()` (sequence).

```
> vec2 = seq(from=0, to=1, by=0.25)
> vec2
[1] 0.00 0.25 0.50 0.75 1.00
```

Εάν προσθέσετε δύο διανύσματα ίδιου μήκους, το πρώτο στοιχείο κάθε διανύσματος αθροίζεται με το άλλο, το ίδιο και το δεύτερο, κ.ο.κ., έχοντας ως αποτέλεσμα ένα νέο διάνυσμα μήκους 5 (ακριβώς όπως και στους κανονικούς υπολογισμούς με διανύσματα). Προσέξτε ότι η συνάρτηση `sum` αθροίζει όλα τα στοιχεία ενός διανύσματος, έχοντας ως αποτέλεσμα έναν αριθμό (ένα βαθμωτό αριθμό).

```
> sum(vec1)
[1] 35
> vec1 + vec2
[1] 1.00 4.25 12.50 8.75 11.00
```

3.6.2 Μητρώα (Matrices)

Τα **μητρώα** δεν είναι τίποτε άλλο παρά δισδιάστατα διανύσματα. Για να ορίσετε ένα μητρώο, χρησιμοποιήστε τη συνάρτηση `matrix`:

```
mat=matrix(data=c(9,2,3,4,5,6),ncol=3)
> mat
      [,1] [,2] [,3]
[1,]    9    3    5
[2,]    2    4    6
```

Το όρισμα `data` καθορίζει ποια νούμερα πρέπει να εισαχθούν στο μητρώο. Χρησιμοποιείτε είτε το `ncol` για να προσδιορίσετε τον αριθμό των στηλών, είτε το `nrow` για να προσδιορίσετε τον αριθμό των γραμμών. Οι πράξεις με μητρώα είναι παρόμοιες με τις πράξεις σε διανύσματα. Τα στοιχεία ενός μητρώου μπορούν να προσπελαστούν με το συνήθη τρόπο: `[row, column]`:

```
> mat[1,2]
[1] 3
```

Όταν θελήσετε να επιλέξετε μια ολόκληρη γραμμή, αφήστε τη θέση για το όρισμα του αριθμού των στηλών κενή (και φυσικά το ανάποδο όταν θελήσετε στήλες).

```
> mat[2,]
[1] 2 4 6
```

Οι συναρτήσεις εξακολουθούν να δουλεύουν όταν έχουν μητρώα ως όρισμα:

```
> mean(mat)
[1] 4.8333
```

3.6.3 Πλαίσια δεδομένων (Data frames)

Ένα **πλαίσιο δεδομένων** είναι ένα μητρώο με ονόματα πάνω από τις στήλες του. Αυτό είναι βολικό, γιατί έτσι μπορείτε να καλέσετε και να χρησιμοποιήσετε όποια από τις στήλες θέλετε χωρίς να γνωρίζετε σε ποια θέση είναι αυτή.

```
> t = data.frame(x = c(11,12,14), y = c(19,20,21), z = c(10,9,7))
> t
   x  y  z
1 11 19 10
2 12 20  9
3 14 21  7
```

```
> mean(t$z)
[1] 8.666667
> mean(t[["z"]])
[1] 8.666667
```

3.6.4 Λίστες (Lists)

Μια άλλη βασική δομή στην R είναι η **λίστα**. Το κύριο πλεονέκτημα των λιστών είναι ότι οι «στήλες» (δεν είναι πλέον διατεταγμένες σε στήλες, αλλά μοιάζουν πιο πολύ με μια συλλογή διανυσμάτων) δεν είναι υποχρεωτικό να έχουν το ίδιο μήκος, αντίθετα με τις περιπτώσεις των μητρώων και των πλαισίων δεδομένων.

Με την παρακάτω εντολή, δημιουργείται μια λίστα μέσω της εισόδου ονομάτων και τιμών:

```
> L = list(one=1, two=c(1,2), five=seq(0, 1, length=5))
```

Η λίστα εμφανίζεται επίσης και στο παράθυρο του χώρου εργασίας.

Εδώ φαίνεται μια τυπική εκτύπωση (μετά από πάτημα των L και ENTER).

```
> L
$one
[1] 1
$two
[1] 1 2
$five
[1] 0.00 0.25 0.50 0.75 1.00
```

Για να δούμε τι υπάρχει μέσα στη λίστα (ονόματα «στηλών»), πληκτρολογούμε:

```
> names(L)
[1] "one" "two" "five"
```

Τέλος, παρουσιάζεται ένας τρόπο χρήσης των αριθμών.

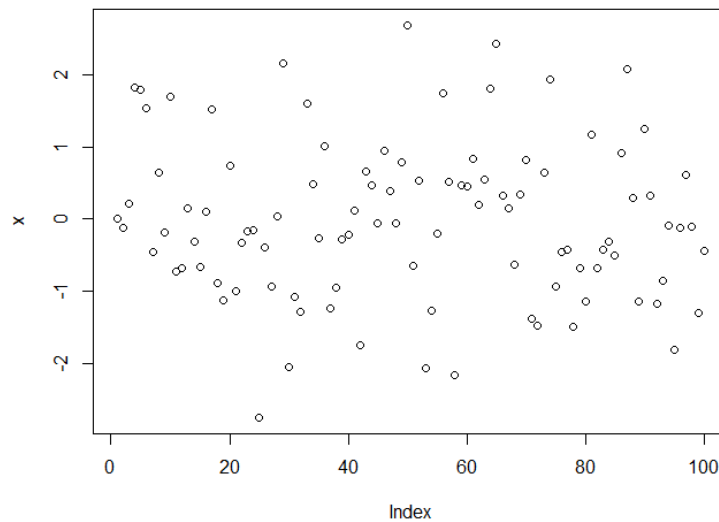
```
> L$five + 10
[1] 10.00 10.25 10.50 10.75 11.00
```

3.7 Γραφικές παραστάσεις

Μία από τις ιδιαιτερότητες της R σε σχέση με άλλες γλώσσες είναι η ευκολία που προσφέρει στη δημιουργία γραφικών παραστάσεων. Το ακόλουθο είναι ένα πολύ απλό παράδειγμα:

```
> x = rnorm(100)
> plot(x)
```

- Στην πρώτη γραμμή, εκχωρούνται 100 τυχαίοι αριθμοί στη μεταβλητή x, η οποία γίνεται διάνυσμα μέσω αυτής της διαδικασίας.
- Στη δεύτερη γραμμή, όλες αυτές οι τιμές σχεδιάζονται στο παράθυρο γραφικών παραστάσεων (Εικόνα 20).

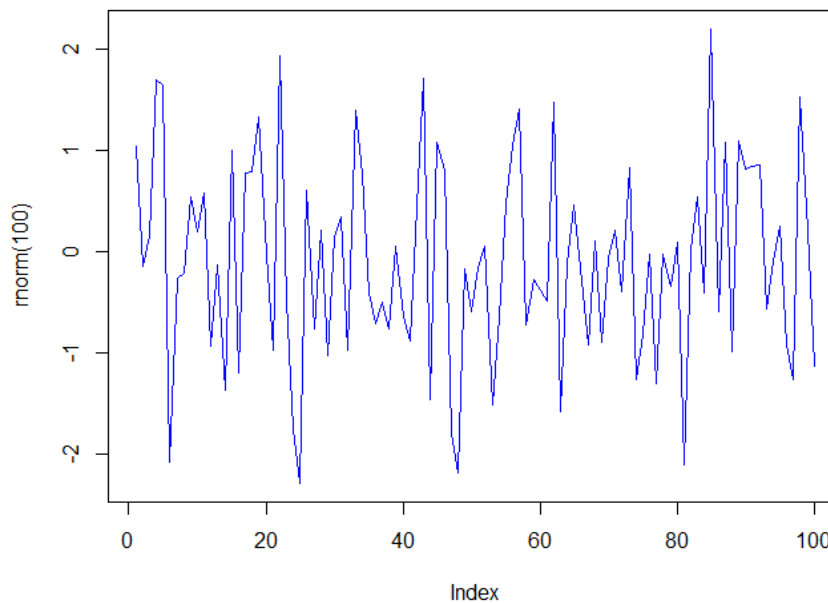


Εικόνα 20: Γράφημα της μεταβλητής x

Η σχεδίαση γραφικών παραστάσεων είναι μια σημαντική στατιστική δραστηριότητα. Οπότε δεν πρέπει να σας εκπλήσσει το γεγονός ότι η R έχει πολλές δυνατότητες σχεδιασμού γραφικών παραστάσεων. Οι ακόλουθες γραμμές εμφανίζουν ένα απλό γράφημα:

```
> plot(rnorm(100), type="l", col="blue")
```

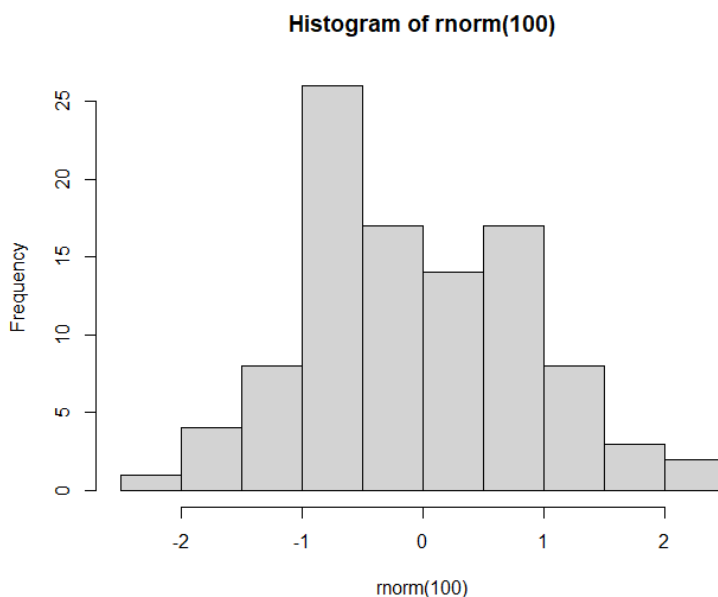
Εκατοντάδες τυχαίοι αριθμοί αναπαρίστανται γραφικά μέσω της σύνδεσης των σημείων με γραμμές (το σύμβολο μέσα σε εισαγωγικά μετά το `type=` είναι το γράμμα `l`, όχι ο αριθμός `1`) σε μπλε χρώμα (Εικόνα 21).



Εικόνα 21: Γραφική παράσταση 100 τυχαίων αριθμών (που ακολουθούν κανονική κατανομή)

Ένα άλλο πολύ απλό παράδειγμα είναι το κλασικό στατιστικό γράφημα του ιστογράμματος, που δημιουργείται από την απλή εντολή:

```
> hist(rnorm(100))
```



Εικόνα 22: Ιστόγραμμα 100 τυχαίων αριθμών (που ακολουθούν κανονική κατανομή)

3.8 Σενάρια

Μπορείτε να αποθηκεύετε τις εντολές σας σε αρχεία, τα λεγόμενα **σενάρια** (scripts). Αυτά τα σενάρια έχουν τυπικά ονόματα αρχείων με την κατάληξη .R, π.χ. foo.R. Μπορείτε να ανοίξετε τον επεξεργαστή κειμένου σε ένα παράθυρο και να επεξεργαστείτε αυτά τα αρχεία κάνοντας κλικ στο File και στο New File ή στο Open file.

Μπορείτε να “τρέξετε” (να στείλετε δηλαδή στο παράθυρο κονσόλας) ένα μέρος του κώδικα, επιλέγοντας γραμμές του και πατώντας CTRL+ENTER ή κάνοντας κλικ στο Run στο παράθυρο του επεξεργαστή κειμένου. Εάν δεν επιλέξετε κάτι, η R θα τρέξει τη γραμμή στην οποία βρίσκεται ο κέρσορας. Μπορείτε πάντα να “τρέξετε” όλο το σενάριο με την εντολή κονσόλας source, κι έτσι π.χ. για το σενάριο στο αρχείο foo.R θα πρέπει να πληκτρολογήσετε:

```
> source("foo.R")
```

Μπορείτε επίσης να κάνετε κλικ στο Run all στο παράθυρο του επεξεργαστή κειμένου ή να πατήσετε τα πλήκτρα CTRL+SHIFT+S ώστε να τρέξετε ολόκληρο το σενάριο με τη μία.

3.9 Διαβάζοντας και γράφοντας αρχεία

Υπάρχουν πολλοί τρόποι για να καταγράψει κανείς δεδομένα σε αρχεία μέσω του περιβάλλοντος της R, και για να διαβάσει δεδομένα από αρχεία. Εδώ θα παρουσιάσουμε έναν τέτοιο τρόπο. Οι ακόλουθες γραμμές παρουσιάζουν τα στοιχειώδη.

Αρχικά, δημιουργείται ένα απλό πλαίσιο δεδομένων ως παράδειγμα και αποθηκεύεται στη μεταβλητή d.

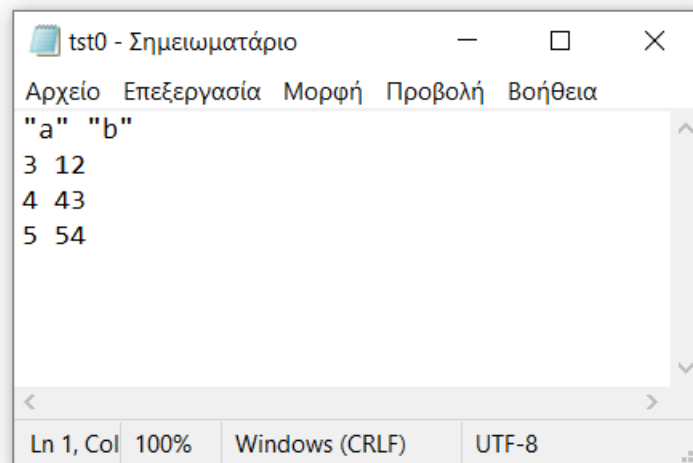
```
> d = data.frame(a = c(3,4,5), b = c(12,43,54))
```

Εδώ φαίνεται το περιεχόμενο αυτού του πλαισίου δεδομένων: δύο στήλες (με όνομα a και b), κάθε μία από τις οποίες περιέχει τρεις αριθμούς.

```
> d
  a  b
1 3 12
2 4 43
3 5 54
```

Στη συνέχεια, καταγράφεται αυτό το πλαίσιο δεδομένων σε ένα αρχείο κειμένου, με όνομα `tst0.txt`. Το όρισμα `row.names=FALSE` αποτρέπει τα ονόματα των γραμμών να καταγραφούν στο αρχείο. Επειδή δεν καθορίζεται κάτι για τα `col.names` (ονόματα των στηλών), επιλέγεται η προκαθορισμένη επιλογή `col.names=TRUE` και τα ονόματα των στηλών καταγράφονται στο αρχείο. Στην Εικόνα 23 φαίνεται το τελικό αρχείο (ανοιγμένο σε έναν επεξεργαστή κειμένου, όπως το Σημειωματάριο), με τα ονόματα των στηλών (a και b) στην πρώτη γραμμή.

```
> write.table(d, file="tst0.txt", row.names=FALSE)
```



Εικόνα 23: Το αρχείο `tst0.txt`

Τέλος, βλέπουμε πως μπορούμε να εισάγουμε ένα αρχείο μέσα σε ένα πλαίσιο δεδομένων. Σημειώστε ότι εισάγονται και τα ονόματα των στηλών. Το πλαίσιο δεδομένων εμφανίζεται επίσης στο παράθυρο του χώρου εργασίας.

```
> d2 = read.table(file="tst0.txt", header=TRUE)
> d2
  a  b
1 3 12
2 4 43
3 5 54
```

3.10 Μη διαθέσιμα δεδομένα

Όταν δουλεύετε με πραγματικά δεδομένα, θα βρεθείτε αντιμέτωποι με τιμές που λείπουν επειδή υπήρξαν αστοχίες στα όργανα μέτρησης. Όταν ένα δεδομένο δεν είναι διαθέσιμο, τότε πρέπει να γράψετε **NA** (τα αρχικά του “Not Available”) αντί κάποιου αριθμού.

```
> j = c(1, 2, NA)
```

Ο υπολογισμός στατιστικών από ημιτελή σύνολα δεδομένων είναι αδύνατος. Κατά συνέπεια, η R θα σας πει ότι δε γνωρίζει ποια είναι η μέγιστη τιμή του `j`:

```
> max(j)
[1] NA
```

Εάν δεν έχετε πρόβλημα με τα δεδομένα που λείπουν και θέλετε να υπολογίσετε τα στατιστικά όπως και να 'χει, μπορείτε να προσθέσετε το όρισμα `na.rm=TRUE` (να απομακρύνω/remove/rm τις τιμές NA).

```
> max(j, na.rm=TRUE)
[1] 2
```

3.11 Κλάσεις

Τα παραδείγματα που είδαμε παραπάνω ήταν σχεδόν όλες με αριθμούς. Μερικές φορές θα χρειαστεί να προσδιορίσετε κάτι το οποίο δεν είναι αριθμός, για παράδειγμα το όνομα ενός σταθμού μετρήσεων ή ενός αρχείου δεδομένων. Σε αυτήν την περίπτωση θέλετε η μεταβλητή να είναι μια ακολουθία χαρακτήρων αντί να είναι αριθμός.

Στις περισσότερες γλώσσες προγραμματισμού υπάρχουν μεταβλητές και τύποι μεταβλητών. Ωστόσο, η R «βλέπει» τα πάντα ως αντικείμενα (object), τα οποία ανήκουν σε μια κλάση (class). Με απλά λόγια, για την R τα αντικείμενα είναι οι μεταβλητές, ενώ η κλάση είναι ο τύπος τους. Στην R δεν απαιτείται η ρητή δήλωση της κλάσης στην οποία ανήκουν τα αντικείμενα. Αυτή καθορίζεται αυτόματα από την τιμή που θα ανατεθεί στο αντικείμενο. Η R έχει **πέντε βασικές ή ατομικές (atomic) κλάσεις αντικειμένων**:

- χαρακτήρας (character)
- αριθμητικός – πραγματικοί αριθμοί (numeric)
- ακέραιος (integer)
- σύνθετος (complex)
- λογικός (logical – True/False)

Η R χρησιμοποιεί, επίσης, βασικές **δομές δεδομένων** ως **κλάσεις αντικειμένων**, τις οποίες είδαμε αναλυτικά στην ενότητα 3.6. Μπορείτε να ρωτήσετε την R ποια είναι η κλάση μιας συγκεκριμένης μεταβλητής πληκτρολογώντας `class(...)`, π.χ.:

```
> class(d)
[1] "data.frame"
```

Επίσης, για να δηλώσετε στην R ότι κάτι είναι ακολουθία χαρακτήρων, πρέπει να πληκτρολογήσετε το κείμενο ανάμεσα σε αποστρόφους, αλλιώς η R θα αρχίσει να ψάχνει για μια καθορισμένη μεταβλητή με το ίδιο όνομα:

```
> m = "apples"
> m
[1] "apples"
> n = pears
Error: object `pears' not found
```

Φυσικά, δεν μπορείτε να κάνετε υπολογισμούς με ακολουθίες χαρακτήρων:

```
> m + 2
Error in m + 2 : non-numeric argument to binary operator
```

3.12 Βοήθεια και τεκμηρίωση

Υπάρχει διαθέσιμο πολύ δωρεάν υλικό τεκμηρίωσης και βοήθειας. Ένα μέρος της βοήθειας εγκαθίσταται αυτόματα. Η πληκτρολόγηση στο παράθυρο της κονσόλας της εντολής:

```
> help(rnorm)
```

θα σας προσφέρει βοήθεια σχετικά με την συνάρτηση `rnorm`. Σας δίνει μια περιγραφή της συνάρτησης, πιθανά ορίσματα και τις τιμές που χρησιμοποιούνται ως προεπιλογή για τα προαιρετικά ορίσματα. Εάν πληκτρολογήσετε:

```
> example(rnorm)
```

θα σας επιστρέψει μερικά παραδείγματα του πώς μπορεί να χρησιμοποιηθεί αυτή η συνάρτηση.

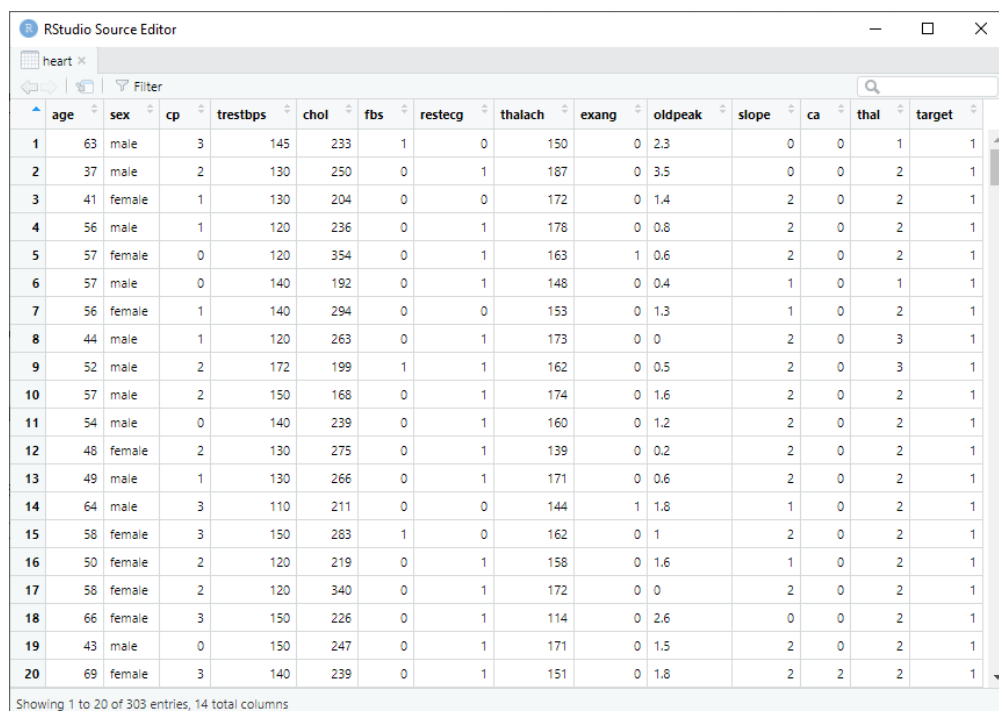
Καθολική βοήθεια βασισμένη σε HTML μπορεί να κληθεί μέσω της εντολής:

```
> help.start()
```

ή με μετάβαση στο παράθυρο βοήθειας.

3.13 Παραδειγματική Άσκηση - Στατιστική ανάλυση σε σύνολο δεδομένων (dataset)

Στη συγκεκριμένη ενότητα θα χρησιμοποιηθεί το σύνολο δεδομένων καρδιαγγειακής νόσου (Heart Disease Data Set) από τη βάση δεδομένων UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>). Πιο αναλυτικά, το συγκεκριμένο σύνολο δεδομένων αποτελεί μία ενοποίηση από τέσσερις βάσεις δεδομένων: Cleveland, Hungary, Switzerland και the VA Long Beach. Περιλαμβάνει συνολικά **76 χαρακτηριστικά**, συμπεριλαμβανομένου του προβλεπόμενου χαρακτηριστικού, αλλά στη συντριπτική πλειοψηφία των δημοσιευμένων μελετών αξιοποιείται ένα υποσύνολο **14 χαρακτηριστικών** από αυτά. Το αρχείο με το σύνολο δεδομένων αποτελεί το *heart.csv* και το αρχείο με την επεξήγηση των χαρακτηριστικών του συνόλου δεδομένων αποτελεί το *heart.pdf*. Ένα δείγμα του συνόλου δεδομένων παρουσιάζεται στην Εικόνα 24. Στο σύνολο αυτό θα πραγματοποιηθούν απλές και βασικές στατιστικές αναλύσεις (π.χ. υπολογισμός ελαχίστων – μεγίστων τιμών, υπολογισμός ΜΟ κτλ.), όπως αυτές που ακολουθούν στη συνέχεια.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	male	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	male	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	female	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	male	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	female	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	male	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	female	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	male	1	120	263	0	1	173	0	0	2	0	3	1
9	52	male	2	172	199	1	1	162	0	0.5	2	0	3	1
10	57	male	2	150	168	0	1	174	0	1.6	2	0	2	1
11	54	male	0	140	239	0	1	160	0	1.2	2	0	2	1
12	48	female	2	130	275	0	1	139	0	0.2	2	0	2	1
13	49	male	1	130	266	0	1	171	0	0.6	2	0	2	1
14	64	male	3	110	211	0	0	144	1	1.8	1	0	2	1
15	58	female	3	150	283	1	0	162	0	1	2	0	2	1
16	50	female	2	120	219	0	1	158	0	1.6	1	0	2	1
17	58	female	2	120	340	0	1	172	0	0	2	0	2	1
18	66	female	3	150	226	0	1	114	0	2.6	0	0	2	1
19	43	male	0	150	247	0	1	171	0	1.5	2	0	2	1
20	69	female	3	140	239	0	1	151	0	1.8	2	2	2	1

Εικόνα 24: Δείγμα του συνόλου δεδομένων *heart*

- Συνολικός αριθμών ασθενών στο σύνολο δεδομένων:

```
> nrow(heart)
```

```
[1] 303
```
- Υπολογισμός μέγιστης τιμής συστολικής πίεσης:

```
> mean(heart$trestbps)
```

```
[1] 131.6238
```
- Υπολογισμός μέγιστης τιμής συστολικής πίεσης:

```
> max(heart$trestbps)
```

```
[1] 200
```
- Εύρεση του φύλου του ατόμου που έχει τη μέγιστη συστολική πίεση:

```
> heart$sex[heart$trestbps == max(heart$trestbps)]
```

```
[1] "female"
```

4. Ερωτήσεις Αυτοαξιολόγησης

1. Ποια είναι τα χαρακτηριστικά των Big Data;
2. Σε τι έγκειται το χαρακτηριστικό της αξίας των Big Data;
3. Αναφέρετε μερικές ενδεικτικές πηγές των Big Data στο χώρο της υγείας.
4. Αναφέρετε ποια είναι τα σημαντικά ζητήματα που θα πρέπει να ληφθούν υπόψη όσον αφορά την αξιοποίηση των Big Data και τα ζητήματα ηθικής, ασφάλειας και ιδιωτικότητας.

Ενδεικτικές Απαντήσεις

Απ.1: Τα χαρακτηριστικά των Big Data είναι ο μεγάλος όγκος δεδομένων (Volume), η αυξημένη ταχύτητα παραγωγής και κυκλοφορίας των δεδομένων (Velocity), η ποικιλία τύπων και μορφών (Variety), δηλαδή ο μικρός βαθμός δόμησης των δεδομένων και η μη ενιαία μορφή τους, η αυξημένη ακρίβεια – εγκυρότητα (Veracity), δηλαδή η μείωση του βαθμού ασάφειας που παρουσιάζουν εγγενώς τα δεδομένα αυτά, και, τέλος, η αξία (Value).

Απ.1: Η κατοχή μεγάλων όγκων δεδομένων δεν έχει ή δε δημιουργεί αυτόματα κάποιου είδους αξία για ένα οργανισμό. Αντίθετα, η αξιοποίησή τους, μέσω της επεξεργασίας, της ανάλυσης και της οπτικοποίησης είναι αυτά που προσδίδουν αξία και ανταγωνιστικό πλεονέκτημα στα υποκείμενα που τα μελετούν. Έτσι, η έννοια της αξίας στα Big Data αναφέρεται στο γεγονός ότι δεν υπάρχει αξία στα δεδομένα εξ ορισμού, αλλά εφόσον αυτά αναλυθούν. Το σίγουρο είναι όμως πως αποτελεί επιθυμητό στοιχείο αλλά δεν είναι δεδομένο. Τέλος, αποτελεί ένα υποκειμενικό χαρακτηριστικό ανάλογα με το ποιος τα μελετά.

Απ.3: Μερικές ενδεικτικές πηγές παραγωγής Big Data στο χώρο της υγείας αποτελούν τα βιοϊατρικά δεδομένα από φορείς υγειονομικής περίθαλψης (ιατρεία, νοσοκομεία, διαγνωστικά κέντρα), τα βιοσήματα που συλλέγονται από φορητούς / φορετούς αισθητήρες, τα στατιστικά δεδομένα από δημόσιους φορείς επιδημιολογικής επιτήρησης (π.χ. ΕΟΔΥ) και τα δεδομένα που παράγονται από τον χρήστη μέσω της αλληλεπίδρασής του με τα κοινωνικά δίκτυα.

Απ.4: Είναι πολύ σημαντική η λήψη μέτρων που θα εξασφαλίζουν μια ορθολογική αξιοποίηση των συλλογών Big Data. Τέτοια μέτρα δύναται να αποτελέσουν η γενική συναίνεση του κάθε ατόμου, η ανωνυμοποίηση των δεδομένων και η χρήση όλων των διαθέσιμων τεχνικών προστασίας τους. Επίσης, η κατάρτιση κωδικών δεοντολογίας για τη συλλογή, πρόσβαση, διαχείριση και επεξεργασία των Big Data από Επιτροπές Δεοντολογίας της Έρευνας σε φορείς που χρησιμοποιούν Big Data για ερευνητικούς σκοπούς κρίνεται και αυτή απαραίτητη. Είναι προφανές πως οι προδιαγραφές ασφαλείας και προάσπισης της ιδιωτικότητας θα πρέπει να είναι αδιαπραγμάτευτες.

5. Βιβλιογραφία

- [1] C. H. Lee and H.-J. Yoon, “Medical big data: promise and challenges,” *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, 2017, doi: 10.23876/j.krcp.2017.36.1.3.
- [2] *Big data now : current perspectives from O'Reilly radar*. O'Reilly Media, 2011.
- [3] UNECE, “Final project proposal: The Role of Big Data in the Modernisation of Statistical Production,” 2013.
- [4] C. Verstraete, “The internet of things.” Hewlett Packard Enterprise, 2015.
- [5] “Data Never Sleeps 8.0 Infographic | Domo.” <https://www.domo.com/learn/data-never-sleeps-8> (accessed Feb. 12, 2021).
- [6] V. Thirani and A. Gupta, “The value of data,” *World Economic Forum*, 2017. .
- [7] “Cloud Security.” <https://www.csa.gov.sg/singcert/publications/cloud-security> (accessed Feb. 12, 2021).
- [8] S. G. Alonso, I. de la Torre Díez, J. J. P. C. Rodrigues, S. Hamrioui, and M. López-Coronado, “A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector,” *J. Med. Syst.*, vol. 41, no. 11, p. 183, 2017, doi: 10.1007/s10916-017-0832-2.
- [9] D. Faggella, “Where Healthcare ’ s Big Data Actually Comes From,” *Techemergence*, 2018.
- [10] K. Hwang, “Sources of Big Data in Medicine,” *Very well health*, 2017.
- [11] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, “Promises and Challenges of Big Data Computing in Health Sciences,” *Big Data Res.*, vol. 2, no. 1, pp. 2–11, Mar. 2015, doi: 10.1016/J.BDR.2015.02.002.
- [12] Ι. Κουμπόπουρος, *Οι Τεχνολογίες Πληροφορίας και Επικοινωνιών στην υγεία*. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015.
- [13] S. D. Young, “A ‘big data’ approach to HIV epidemiology and prevention,” *Prev. Med. (Baltim).*, vol. 70, pp. 17–18, Jan. 2015, doi: 10.1016/j.ypmed.2014.11.002.
- [14] C. A. McGinn *et al.*, “Comparison of user groups’ perspectives of barriers and facilitators to implementing electronic health records: a systematic review,” doi: 10.1186/1741-7015-9-46.
- [15] M. K. Ross and W. Wei, ““ Big Data ’ and the Electronic Health Record,” pp. 97–104, 2014.
- [16] S. R. Sukumar, R. Natarajan, and R. K. Ferrell, “Quality of Big Data in health care,” *Int. J. Health Care Qual. Assur.*, vol. 28, no. 6, pp. 621–634, Jul. 2015, doi: 10.1108/IJHCQA-07-2014-0080.
- [17] A. Dagliati *et al.*, “Big Data as a Driver for Clinical Decision Support Systems: A Learning Health Systems Perspective,” *Front. Digit. Humanit.*, vol. 5, no. 8, p. 8, May 2018, doi: 10.3389/fdigh.2018.00008.
- [18] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: Using analytics to identify and manage high-risk and high-cost patients,” *Health Aff.*, vol. 33, no. 7, pp. 1123–1131, 2014, doi: 10.1377/hlthaff.2014.0041.
- [19] N. V. Chawla and D. A. Davis, “Bringing big data to personalized healthcare: A patient-centered framework,” *Journal of General Internal Medicine*, vol. 28, no. SUPPL.3. Sep. 2013, doi: 10.1007/s11606-013-2455-8.
- [20] M. Panahiazar, V. Taslimitehrani, A. Jadhav, and J. Pathak, “Empowering personalized medicine with big data and semantic web technology: Promises, challenges, and use cases,” in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, Jan. 2015, vol. 2014, pp. 790–795, doi: 10.1109/BigData.2014.7004307.
- [21] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, “Big data security and privacy in healthcare: A Review,” *Procedia Comput. Sci.*, vol. 113, pp. 73–80, Jan. 2017, doi: 10.1016/J.PROCS.2017.08.292.
- [22] K. Abouelmehdi, A. Beni-hessane, and H. Khaloufi, “Big healthcare data: preserving security and privacy,” doi: 10.1186/s40537-017-0110-7.

- [23] P. Jain, M. Gyanchandani, and N. K. Background, “Big data privacy: a technological perspective and review,” *J. Big Data*, doi: 10.1186/s40537-016-0059-y.
- [24] Ευρωπαϊκό Κοινοβούλιο και Συμβούλιο, “Επεξεργασία δεδομένων προσωπικού χαρακτήρα.” .
- [25] Ευρωπαϊκό Κοινοβούλιο και Συμβούλιο, “ΚΑΝΟΝΙΣΜΟΣ (ΕΕ) 2016/679 ΤΟΥ ΕΥΡΩΠΑΪΚΟΥ ΚΟΙΝΟΒΟΥΛΙΟΥ ΚΑΙ ΤΟΥ ΣΥΜΒΟΥΛΙΟΥ της 27ης Απριλίου 2016 για την προστασία των φυσικών προσώπων έναντι της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα και για την ελεύθερη κυκλοφορία των δεδομένων αυτών και την .” .
- [26] Ε. Ε. Βιοηθικής, “Μεγάλα Δεδομένα (Big Data) στη Υγεία,” Αθήνα, 2017.