



ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«Διοίκηση στη Ναυτική Επιστήμη και Τεχνολογία»



**Διαχείριση Μεγάλου Όγκου Δεδομένων
στη Ναυτιλία και το Θαλάσσιο Περιβάλλον**

Ακαδημαϊκό Έτος 2025 – 2026

Εισαγωγή

- ◆ Τι είναι τα Big Data;
- Τα μεγάλα δεδομένα είναι ένα πεδίο που αντιμετωπίζει τους τρόπους ανάλυσης, συστηματικής απόσπασης πληροφοριών ή άλλης αντιμετώπισης με σύνολα δεδομένων που είναι πολύ μεγάλα ή περίπλοκα για να αντιμετωπιστούν από το παραδοσιακό λογισμικό εφαρμογών επεξεργασίας δεδομένων.
- Με απλά λόγια, τα μεγάλα δεδομένα είναι μεγαλύτερα, πιο σύνθετα σύνολα δεδομένων από διαφορετικές πηγές δεδομένων.

Εισαγωγή

- Αν και η έννοια των μεγάλων δεδομένων (Big Data) είναι σχετικά νέα, η προέλευση των μεγάλων δεδομένων αρχίζει από τη δεκαετία του 1960 - 1970, καθώς ο κόσμος των μεγάλων δεδομένων μόλις ξεκίνησε, με τις πρώτες βάσεις δεδομένων και την ανάπτυξη της σχεσιακής βάσης δεδομένων.
- Ωστόσο στην αρχή του 2005, οι άνθρωποι άρχισαν να συνειδητοποιούν το μέγεθος των δεδομένων.
- Με πρώτους τους χρήστες του Facebook, του YouTube και άλλων online πλατφορμών να χρησιμοποιούν τεράστιο όγκο δεδομένων.
- Με τη δημιουργία του Hadoop (δημιουργήθηκε ένα πλαίσιο ανοιχτού κώδικα ειδικά για την αποθήκευση και την ανάλυση μεγάλων συνόλων δεδομένων) την ίδια χρονιά
- Το NoSQL άρχισε επίσης να κερδίζει δημοτικότητα κατά τη διάρκεια αυτής της περιόδου.
- Η ανάπτυξη πλαισίων ανοιχτού κώδικα, όπως το Hadoop (και πιο πρόσφατα η Spark), ήταν ουσιαστικής σημασίας για την ανάπτυξη μεγάλων δεδομένων επειδή καθιστούν τα δεδομένα μεγάλης χωρητικότητας πιο εύχρηστα και φθηνότερα στην αποθήκευση, όπως θα δούμε παρακάτω.

Εισαγωγή

- Με την εμφάνιση του Ίντερνετ των πραγμάτων (IoT), περισσότερα αντικείμενα και συσκευές συνδέονται στο διαδίκτυο, συγκεντρώνοντας δεδομένα σε πραγματικό χρόνο.
- Η εμφάνιση της μηχανικής μάθησης έχει παίξει επίσης σημαντικό ρόλο στην κατανόηση και ανάλυση της πληροφορίας.
- Το Cloud computing έχει επεκτείνει περισσότερο τις δυνατότητες των μεγάλων δεδομένων.

Εισαγωγή

Σήμερα, το πλήθος των διαθέσιμων δεδομένων είναι τεράστιο και αυξάνεται εκθετικά κάθε μέρα.

Η μείωση στο κόστος συλλογής και της δυσκολίας στη συλλογή και αποθήκευση των δεδομένων συνετέλεσε σημαντικά στην ανάπτυξη του πεδίου αυτού.

Ο τεράστιος όγκος δεδομένων, που συσσωρεύεται στις βάσεις και στις αποθήκες δεν μπορεί να αξιοποιηθεί όπως είναι.

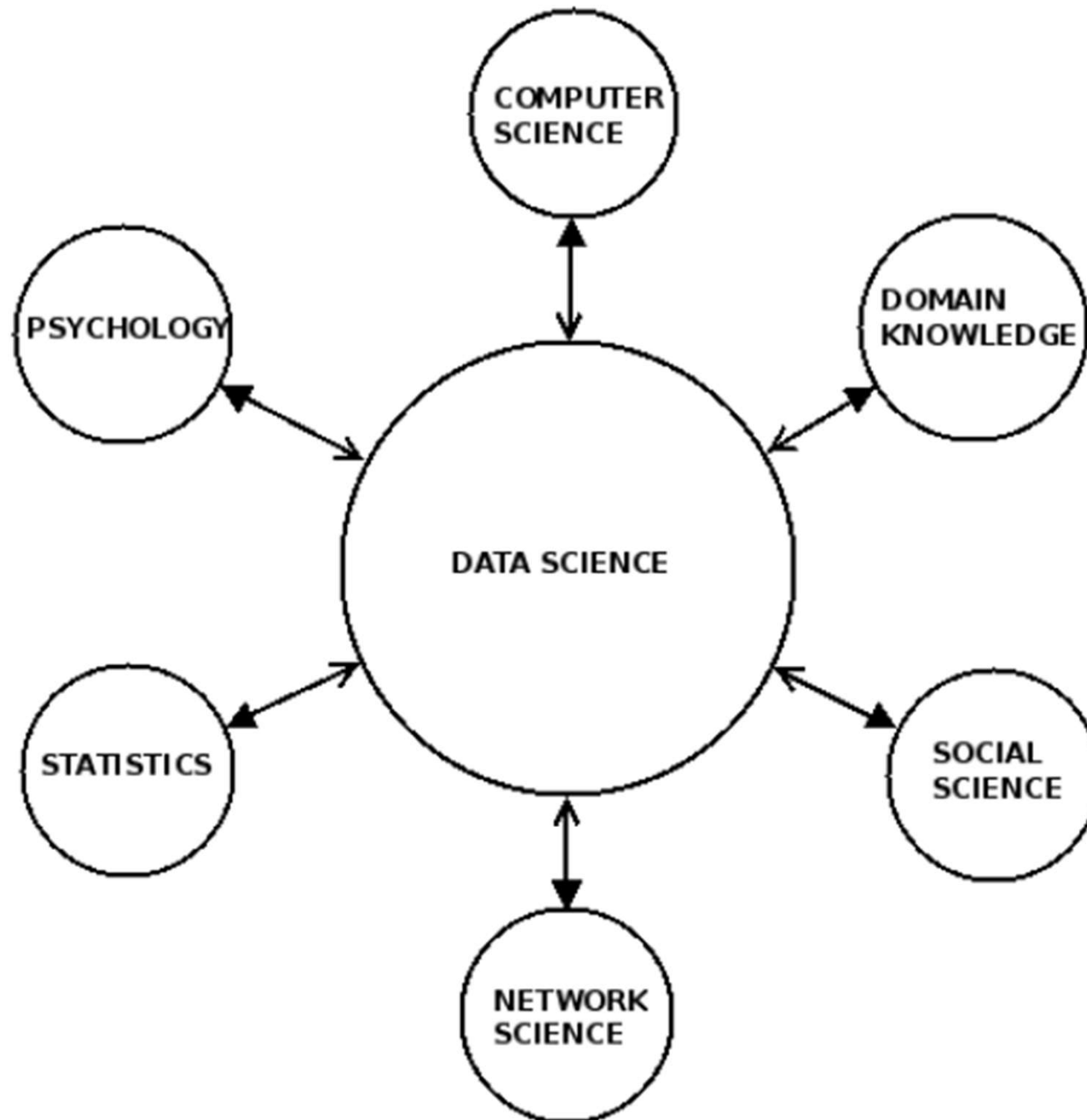
Πρέπει αρχικά να γίνουν κάποιες ενέργειες για να δομηθούν κατάλληλα τα δεδομένα, ώστε στη συνέχεια να μπορούν να αξιοποιηθούν.

Πηγές Δεδομένων

Table 1. Different Sources of Data

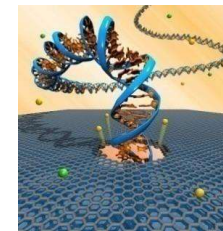
Sector	Data Produced	Use
Astronomy	Movement of stars, satellites, etc.	To monitor the activities of asteroid bodies and satellites
Financial	News content via video, audio, twitter and news report	To make trading decisions
Healthcare	Electronic medical records and images	To aid in short-term public health monitoring and long-term epidemiological research programs
Internet of Things (IoT)	Sensor data	To monitor various activities in smart cities
Life Sciences	Gene sequences	To analyze genetic variations and potential treatment effectiveness
Media/Entertainment	Content and user viewing behavior	To capture more viewers
Social Media	Blog posts, tweets, social networking sites, log details	To analyze the customer behavior pattern
Telecommunications	Call Detail Records (CDR)	Customer churn management
Transportation, Logistics, Retail, Utilities	Sensor data generated from fleet transceivers, RFID tag readers and smart meters	To optimize operations
Video Surveillance	Recordings from CCTV to IPTV cameras and recording system	To analyze behavioral patterns for service enhancement and security

Διεπιστημονικότητα



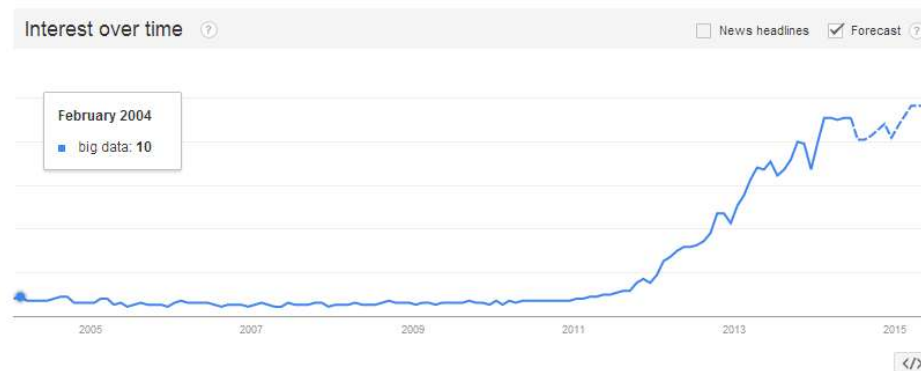
Big Data Facts (1)

- 90% των σημερινών δεδομένων δημιουργήθηκαν τα τελευταία 2 χρόνια
- Νόμος του Moore: Διπλασιασμός δεδομένων κάθε 18 μήνες
- YouTube: 13 εκατ. ώρες και 700 δις αναπαραγωγές το 2010, 48h/min -> 8years/day upload
- Facebook: 20TB/ημέρα zipped, 1B shares/day
- CERN/LHC: 40TB/μέρα (15PB/έτος)



Big Data Facts (2)

- By 2015, 4.4M new Big Data IT jobs
- By 2027, a shortage of 250K big-data experts
- Bad/poor data costs US 600B \$/year
- For a Fortune 1000 company, a 10% increase on data usability -> 2B USD increase



Εισαγωγή

◆ Τι είναι τα Big Data;

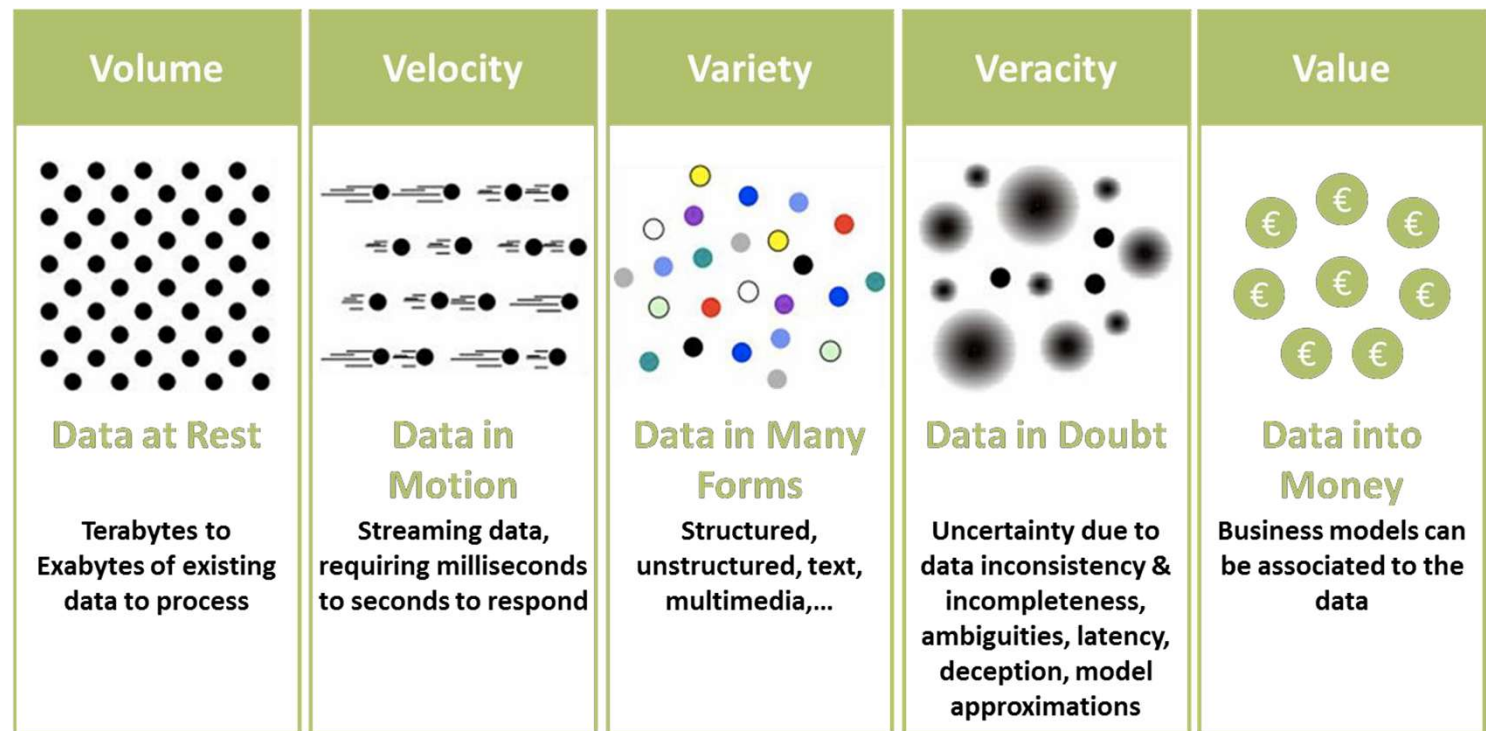
- Ορισμός: Εξαιρετικά μεγάλα σύνολα δεδομένων που μπορούν να αναλυθούν υπολογιστικά για να αποκαλύψουν μοτίβα, τάσεις και συσχετίσεις.

- Χαρακτηριστικά:

“5 Vs”
○ Όγκος (Volume),
○ Ποικιλομορφία (Variety),
○ Ταχύτητα (Velocity),
○ Αυθεντικότητα (Veracity)
○ Τιμές (Value)

Μεγάλα Δεδομένα (Big Data)

- 5 Vs:
- Volume
- Velocity
- Variety
- Veracity
- Value



Adapted by a post of Michael Walker on 28 November 2012

“5 Vs”

- Όγκος (Volume),
- Ποικιλομορφία (Variety),
- Ταχύτητα (Velocity),
- Αυθεντικότητα (Veracity)
- Τιμές (Value)

Ο ίδιος ο όρος Big Data σχετίζεται με ένα μέγεθος που είναι μεγάλο.

Το μέγεθος των δεδομένων παίζει πολύ σημαντικό ρόλο στον προσδιορισμό της αξίας των δεδομένων.

Επίσης, εάν ένα σύνολο δεδομένων μπορεί πραγματικά να θεωρηθεί ως Big Data ή όχι, εξαρτάται από τον όγκο τους.

“5 Vs”

- Όγκος (Volume),
- **Ποικιλομορφία (Variety),**
- Ταχύτητα (Velocity),
- Αυθεντικότητα (Veracity)
- Τιμές (Value)

Η διαφορετικές μορφές δεδομένων αναφέρονται σε ετερογενείς πηγές και τη φύση των δεδομένων, τόσο δομημένων όσο και μη.

Κατά τη διάρκεια των προηγούμενων ετών, τα υπολογιστικά φύλλα και οι βάσεις δεδομένων ήταν οι μόνες πηγές δεδομένων που έλαβαν υπόψη οι περισσότερες εφαρμογές.

Στις μέρες μας, στις εφαρμογές ανάλυσης εξετάζονται και δεδομένα με τη μορφή email, φωτογραφιών, βίντεο, συσκευών παρακολούθησης, PDF, ήχου κ.λπ.

Αυτή η ποικιλία μη δομημένων δεδομένων θέτει ορισμένα ζητήματα για την αποθήκευση, την εξόρυξη και την ανάλυση τους.

Variety

- ◈ Ποικίλοι τύποι δεδομένων:
 - ◈ Δομημένα δεδομένα (structured data)
 - ◈ Σχεσιακές ΒΔ
 - ◈ Ημι-δομημένα δεδομένα (semi-structured data)
 - ◈ XML, JSON
 - ◈ Σχεδόν δομημένα δεδομένα (quasi-structured data)
 - ◈ Ροές δεδομένων
 - ◈ Αδόμητα Δεδομένα (unstructured data)
 - ◈ e-mails, chat messages, webpages, word documents, pdf files, εικόνες, βίντεο, αρχεία ήχου
- ◈ Big Data: μίγμα τύπων δεδομένων
- ◈ Μια εφαρμογή μπορεί να παράξει/συλλέξει πολλούς τύπους δεδομένων

Variety

Ποικίλες Πηγές δεδομένων:

- ◊ Web Δεδομένα
 - ◊ Log files για καταγραφή των κινήσεων των χρηστών/πελατών
- ◊ Δεδομένα από έξυπνα δίκτυα και αισθητήρες
 - ◊ Δεδομένα GPS (π.χ. τοποθεσία χρήστη)
 - ◊ Δεδομένα σεισμικών, γεωφυσικών, περιβαλλοντικών αισθητήρων
- ◊ Δεδομένα κοινωνικών δικτύων
 - ◊ Δεδομένα για καταναλωτικές συνήθειες, προφίλ χρήστη, ...
- ◊ Ανοικτά δεδομένα
 - ◊ Διεθνείς οργανισμοί και θεσμοί, υπουργεία, επιστημονικοί οργανισμοί,...
- ◊ Επιστημονικά δεδομένα
 - ◊ Έξοδοι ιατρικών μηχανημάτων
 - ◊ Ακολουθία νουκλεοτιδίων στο DNA
- ◊ Επιχειρησιακά Δεδομένα
 - ◊ Αξιοποιούνται στα συστήματα ERP, CRM, κ.α.

“5 Vs”

- Όγκος (Volume),
- Ποικιλομορφία (Variety),
- Ταχύτητα (Velocity),
- Αυθεντικότητα (Veracity)
- Τιμές (Value)

Ο όρος «ταχύτητα» αναφέρεται στην ταχύτητα παραγωγής και επεξεργασία των δεδομένων.

Το πόσο γρήγορα παράγονται και επεξεργάζονται τα δεδομένα για να ανταποκριθούν στις απαιτήσεις, καθορίζει τις πραγματικές δυνατότητες τους.

Το Big Data Velocity ασχολείται με την ταχύτητα με την οποία ρέουν δεδομένα από πηγές όπως επιχειρηματικές διαδικασίες, αρχεία καταγραφής εφαρμογών, δίκτυα και ιστότοποι μέσω κοινωνικής δικτύωσης, αισθητήρες, φορητές συσκευές **αλλά και κίνηση πλοίων, δεδομένα μηχανών, δεδομένα θαλάσσιου περιβάλλοντος, κ.α.**

Η ροή δεδομένων σε ορισμένες από αυτές της πηγές είναι τεράστια και συνεχής.

“5 Vs”

- Όγκος (Volume),
- Ποικιλομορφία (Variety),
- Ταχύτητα (Velocity),
- **Αυθεντικότητα (Veracity)**
- Τιμές (Value)

Μπορεί να εγγυηθεί την δυνατότητα αξιόπιστης χρήσης των δεδομένων

“5 Vs”

- Όγκος (Volume),
- Ποικιλομορφία (Variety),
- Ταχύτητα (Velocity),
- Αυθεντικότητα (Veracity)
- **Τιμές (Value)**

Το νόημα που εξάγεται από την ανάλυση των δεδομένων πρέπει να έχει πραγματική αξία εξυπηρετώντας έναν συγκεκριμένο στόχο.

Περισσότερα V's

- ◆ **Volatility:** μεταβλητότητα δεδομένων που αφορά το χρονικό διάστημα εγκυρότητας και διατήρησης των δεδομένων
- ◆ **Viscosity:** ο χρόνος καθυστέρησης (lag time) των δεδομένων που σχετίζονται με το συμβάν που περιγράφεται (μπορεί να σχετιστεί με το Velocity)
- ◆ **Virality:** ο ρυθμός διάδοσης των δεδομένων, δηλαδή πόσο συχνά επιλέγονται και αναπαράγονται από άλλους χρήστες ή συμβάντα
- ◆ **Variability:** αναφέρεται στα δεδομένα των οποίων η σημασία σταθερά αλλάζει, π.χ. όταν η συλλογή των δεδομένων εξαρτάται από τη γλωσσική επεξεργασία.

Εφαρμογές μεγάλων δεδομένων
στη καθημερινότητα

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

- Ανάπτυξη προϊόντων: Εταιρείες χρησιμοποιούν μεγάλα δεδομένα για να προβλέψουν τη ζήτηση και τις ανάγκες των πελατών.
- Επιπλέον, εταιρείες χρησιμοποιούν δεδομένα και αναλύσεις από κοινωνικά δίκτυα και μηχανές αναζήτησης για να προγραμματίσουν, να δημιουργήσουν και να προωθήσουν στην αγορά νέα προϊόντα.
- Συντήρηση: Παράγοντες που μπορούν να προβλέψουν μηχανικές βλάβες μπορεί να είναι βαθιά θαμμένοι σε δομημένα δεδομένα, όπως το έτος του εξοπλισμού, ημερομηνία κατασκευής και μοντέλο μιας μηχανής.
- Καθώς και σε μη δομημένα δεδομένα που καλύπτουν εκατομμύρια καταχωρήσεις καταγραφής, δεδομένα αισθητήρων, σφάλματα μηνυμάτων και τη θερμοκρασία του εξοπλισμού.
- Με την ανάλυση των μεγάλων δεδομένων παίρνουμε ενδείξεις πιθανών ζητημάτων-σφαλμάτων πριν συμβούν προβλήματα-βλάβες, οι οργανώσεις μπορούν να αναπτύξουν πιο αποτελεσματικά τη συντήρηση και να μεγιστοποιήσουν το χρόνο λειτουργίας των εξαρτημάτων και εξοπλισμού.

Εφαρμογές μεγάλων δεδομένων στη καθημερινότητα

- Εμπειρία πελατών: Με την ύπαρξη των Μεγάλων δεδομένων (Big Data) οι εταιρείες έχουν μια πιο σαφή εικόνα για την ικανοποίηση των πελατών.
- Μια σαφέστερη εικόνα της εμπειρίας των πελατών είναι πιο δυνατή τώρα από ποτέ.
- Με τα μεγάλα δεδομένα τώρα υπάρχει η δυνατότητα της συλλογής δεδομένων από κοινωνικά δίκτυα, επισκέψεις σε ιστοσελίδες (cookies), αρχεία καταγραφής κλήσεων και άλλες πηγές δεδομένων για τη βελτίωση της εμπειρίας του πελάτη.

Μεγάλα Δεδομένα και Πληροφορική

- Τα μεγάλα δεδομένα συνδυάζονται πρακτικά με το τομέα της πληροφορικής.
- Παραδείγματα περιλαμβάνουν: την κατανόηση του παγκόσμιου ιστού και τις αναζητήσεις του μέσου χρήστη.
- Με τη συλλογή των πληροφοριών αυτών μπορεί να εξαχθούν δεδομένα με τις ηλεκτρονικές αγορές, την αλληλεπίδραση με τα μέσα κοινωνικής δικτύωσης.
- Τα αναλυτικά στοιχεία Big Data είναι πράγματι μια επανάσταση στον τομέα της Πληροφορικής.
- Η χρήση της ανάλυσης δεδομένων από τις εταιρείες ενισχύεται κάθε χρόνο καθώς οι ανάγκες των εταιρειών αυξάνονται.
- Τα οφέλη σε πραγματικό χρόνο του Big Data Analytic είναι πολλά.

Examples of Big Data

- The New York Stock Exchange is an example of Big Data that generates about *one terabyte* of new trade data per day.



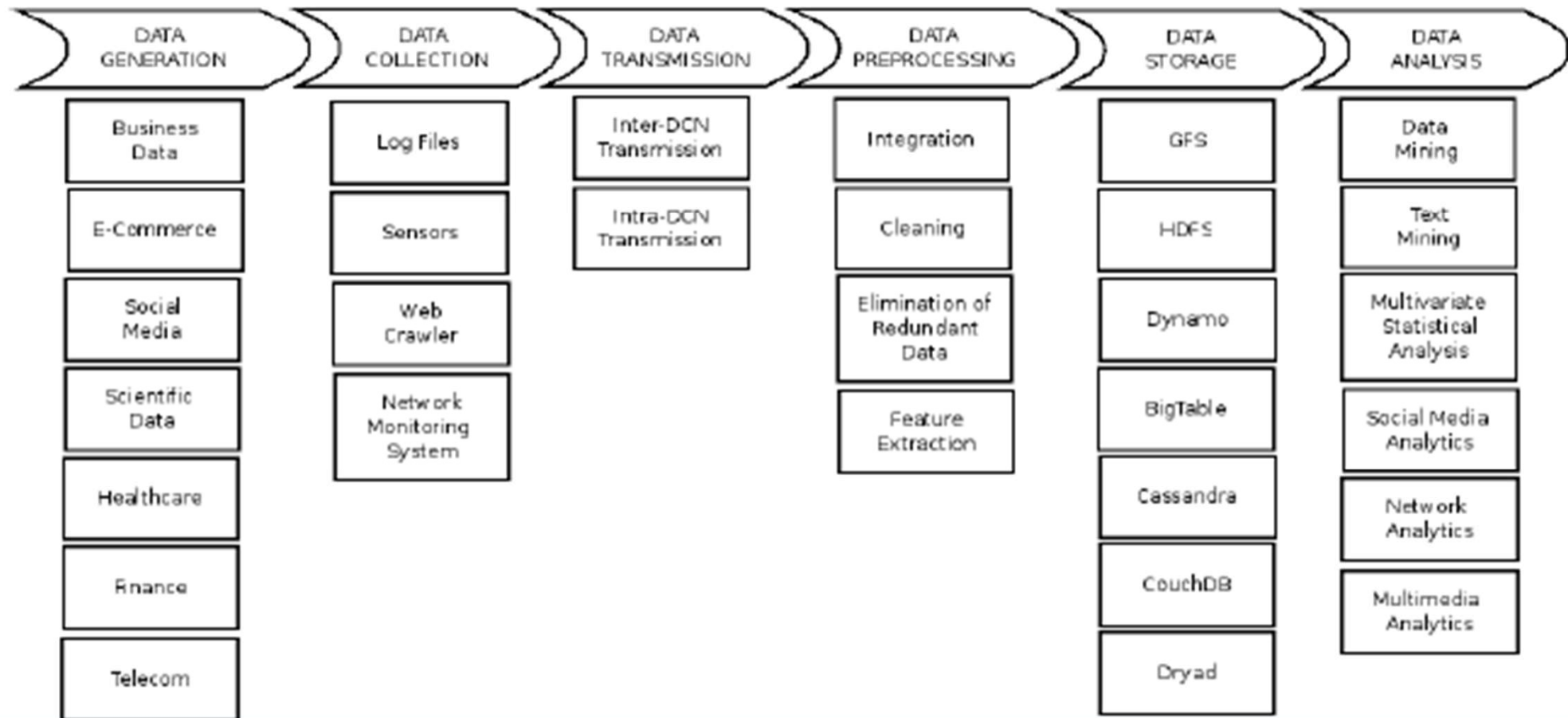
➤ Social Media

The statistic shows that +500 terabytes of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

Σημαντικότητα

- ◊ Γιατί είναι Σημαντικά – Μερικοί Λόγοι μόνο:
 - ◊ Ενισχύουν/Υποστηρίζουν τις διαδικασίες λήψης αποφάσεων
 - ◊ Προσδιορίζουν τις τάσεις της αγοράς και τις προτιμήσεις των πελατών
 - ◊ Βελτιώνουν διαδικασίες και λειτουργικότητα και μειώνουν το κόστος
 - ◊ Προωθούν την καινοτομία και την ανάπτυξη νέων προϊόντων

Big Data Value Chain



Η Επιστήμη των Δεδομένων

Η Επιστήμη των Δεδομένων (Data Science), είναι ένας καινούριος όρος, ο οποίος ήρθε να αντικαταστήσει προγενέστερους όρους, όπως **Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Data-base)** ή **Εξόρυξη Δεδομένων (Data Mining)**.

Οι τρεις αυτοί όροι χρησιμοποιούνται σχεδόν εναλλακτικά, για να περιγράψουν μία **ημι-αυτοματοποιημένη διαδικασία**, σκοπός της οποίας είναι **να αναλύσει έναν μεγάλο όγκο δεδομένων** που αφορούν ένα συγκεκριμένο πρόβλημα, συνήθως εμπορικού ή επιστημονικού ενδιαφέροντος, για **την παραγωγή προτύπων (patterns)**, όπως συνηθίζεται να λέμε σε ορισμένους τομείς, όπως

- 1 στη Στατιστική,
- 2 στη Μηχανική Μάθηση (Machine Learning) και
- 3 στην Αναγνώριση Προτύπων (Pattern Recognition).

Η Επιστήμη των Δεδομένων

Τα πρότυπα αυτά, τα οποία συναντάμε σε διάφορες μορφές, όπως **συσχετίσεις**, **ανωμαλίες**, **συστάδες**, **κλάσεις** κ.λπ., αποτελούν δομές ή περιστατικά, που εμφανίζονται στα δεδομένα και έχουν κάποια ιδιαίτερη σημασία από στατιστικής πλευράς.

Ένα από τα σημαντικότερα θέματα από πλευράς ουσίας στην Επιστήμη των Δεδομένων αφορά τόσο την εύρεση ή αλλιώς ανακάλυψη (η λέξη ανακάλυψη εμπεριέχει σε μεγάλο βαθμό το γεγονός ότι τα πρότυπα αυτά δεν ήταν αναμενόμενα εκ των προτέρων), όσο και τον χαρακτηρισμό αυτών των προτύπων.

Πιο συγκεκριμένα, ένα πρότυπο αναφέρεται και ως μία διευθέτηση ή αλλιώς διάταξη, στην οποία θεωρείται ότι υπάρχει κάποιου είδους οργάνωση της υποκείμενης δομής.

Big Data Key Technologies

- ◇ Storage
- ◇ Data Processing and Analysis
- ◇ Databases
- ◇ Modeling & Simulation
 - ◇ Neural Networks
 - ◇ Numerical Modeling
 - ◇ Statistical Modeling
- ◇ Machine Learning

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

- Δομημένα Δεδομένα: Ο όρος δομημένα δεδομένα αναφέρεται σε δεδομένα που έχουν καθορισμένο μήκος και μορφή.
- Παραδείγματα δομημένων δεδομένων περιλαμβάνουν αριθμούς, ημερομηνίες και ομάδες λέξεων και αριθμών που ονομάζονται strings.
- Οι πηγές των δομημένων δεδομένων είναι συνήθως 2:
 1. Δημιουργία από υπολογιστή ή μηχανή: Τα δεδομένα που παράγονται από μηχανή γενικά αναφέρονται σε δεδομένα που δημιουργούνται από μηχάνημα χωρίς ανθρώπινη παρέμβαση.
 2. Ανθρώπινα: Αυτά είναι δεδομένα που παρέχουν οι άνθρωποι, σε αλληλεπίδραση με τους υπολογιστές.

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

- Μη δομημένα δεδομένα: Τα μη δομημένα δεδομένα είναι δεδομένα που δεν ακολουθούν μια καθορισμένη μορφή.
- Τα μη δομημένα δεδομένα είναι πραγματικά τα περισσότερα από τα δεδομένα που υπάρχουν.
- Μέχρι πρόσφατα, όμως, η τεχνολογία δεν μπορούσε να κάνει πολλά μαζί με τα δομημένα αυτά, εκτός από την αποθήκευση ή την ανάλυσή τους.

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

- Τα μη δομημένα δεδομένα είναι παντού.
- Στην πραγματικότητα, τα περισσότερα άτομα και οργανισμοί περιτριγυρίζονται γύρω από μη δομημένα δεδομένα.
- Όπως συμβαίνει με τα δομημένα δεδομένα, τα μη δομημένα δεδομένα είτε είναι μηχανικά είτε παράγονται από ανθρώπους.

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

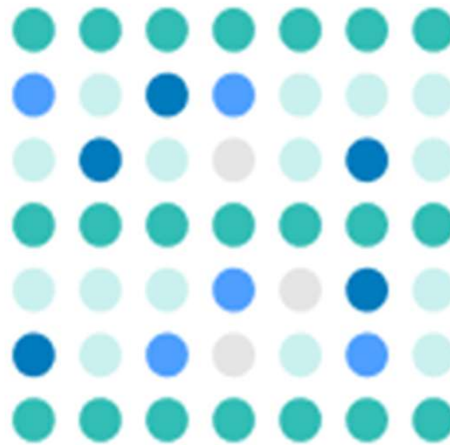
- Τα ημι-δομημένα δεδομένα είναι ένα είδος δεδομένων που εμπίπτει μεταξύ δομημένων και μη δομημένων δεδομένων.
- Τα ημι-δομημένα δεδομένα δεν συμμορφώνονται υποχρεωτικά με ένα καθορισμένο σχήμα, δηλαδή δομή, αλλά μπορεί να έχουν την δική τους αξία.
- Παραδείγματα ημιδομημένων δεδομένων περιλαμβάνουν τα EDI, SWIFT και XML.
- Μπορεί να θεωρηθούν ως είδος βοήθειας για την επεξεργασία , πολυσύνθετων και πολύπλοκων δεδομένων.

Κατηγορίες μεγάλων δεδομένων

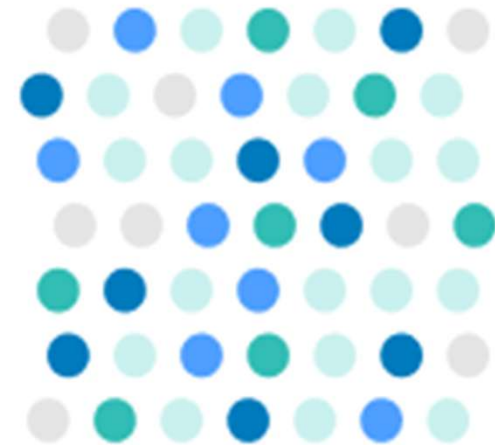
- Δομημένα Δεδομένα (Structured data)
- Μη δομημένα δεδομένα (Unstructured data)
- Ημι-δομημένα δεδομένα (Semi-Structured data)



Structured Data



Semi-Structured Data



Unstructured Data

Δομημένα Δεδομένα (Structured data)

- Τα δομημένα δεδομένα είναι πληροφορίες που έχουν μορφοποιηθεί και μετατραπεί σε ένα καλά καθορισμένο μοντέλο δεδομένων

Τα δομημένα δεδομένα είναι ποσοτικά δεδομένα. Αυτό σημαίνει ότι περιλαμβάνουν μετρήσιμες αριθμητικές τιμές, όπως αριθμούς, ημερομηνίες και ώρες.

Παράδειγμα δομημένων δεδομένων

- Excel files.
- SQL databases.
- Point-of-sale data.
- Web form results.
- Search engine optimization
- Product directories.
- Inventory control.
- Reservation systems.

μόνο περίπου το 20% των δεδομένων των επιχειρήσεων είναι σε δομημένη μορφή.

Μη Δομημένα Δεδομένα (Un-Structured data)

Τα μη δομημένα δεδομένα (συχνά αποκαλούμενα «μεγάλα δεδομένα» ή «ακατέργαστα δεδομένα») είναι δεδομένα που δεν έχουν προκαθορισμένη μορφή ή μοντέλο.

Επειδή τα μη δομημένα δεδομένα είναι συνήθως πλούσια σε κείμενο και μη αριθμητικά, θεωρούνται ποιοτικά.

Αν και μπορεί να περιλαμβάνουν αριθμούς, αυτοί χρησιμοποιούνται γενικά για την κατηγοριοποίηση πληροφοριών. Περίπου το 80% των δεδομένων των επιχειρήσεων είναι σε μη δομημένη μορφή

Παράδειγμα μη δομημένων δεδομένων:

- Books & PDF
- Email
- Camera Recordings
- Machine and Sensor Data
- Excel sheets
- Φωτογραφίες

Ημι Δομημένα Δεδομένα (semi Structured data)

Τα ημι-δομημένα δεδομένα αποτελούν μια μορφή δομημένων δεδομένων που δεν ακολουθούν την δομή των μοντέλων δεδομένων που σχετίζονται με σχεσιακές βάσεις δεδομένων ή άλλες μορφές πινάκων δεδομένων.

Παράδειγμα ημι-δομημένων δεδομένων:

- Email Different Data bases

Ημιδομημένα Δεδομένα (Semi-Structured Data)

- ◆ Είναι τα δεδομένα που δεν ακολουθούν πλήρως τη μορφή προκαθορισμένων σχημάτων, αλλά έχουν μια οργανωμένη ιεραρχική δομή που τα καθιστά ευκολότερα στην ανάλυση σε σχέση με τα μη δομημένα δεδομένα.
- ◆ Τα Χαρακτηριστικά τους είναι:
 - α. Ευέλικτη Δομή: Τα δεδομένα μπορεί να έχουν μεταβλητό αριθμό πεδίων ή μορφών.
 - β. Χρήση Μεταδεδομένων: Περιέχουν ετικέτες (tags) και ιδιότητες που τα περιγράφουν (π.χ., XML, JSON).
- ◆ Παραδείγματα: Αρχεία XML, JSON, e-mails, δεδομένα κοινωνικών δικτύων.
- ◆ Εφαρμογές στα Big Data:
 - α. Ανάλυση μέσω πλατφορμών NoSQL (π.χ., MongoDB, Cassandra).
 - β. Διαχείριση αρχείων και δεδομένων ιστού.

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

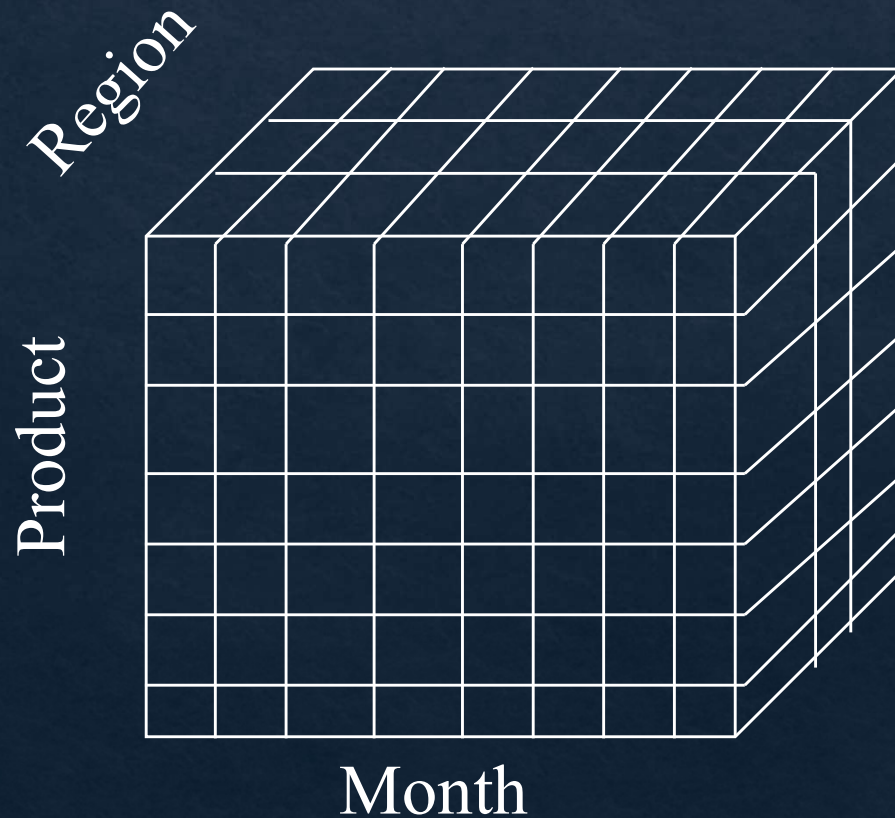
- Μεταδεδομένα (Metadata): Ένα κρίσιμο στοιχείο για την ενσωμάτωση όλων αυτών των δεδομένων είναι τα μεταδεδομένα.
- Τα μεταδεδομένα είναι οι ορισμοί, αντιστοιχίσεις και άλλα χαρακτηριστικά που χρησιμοποιούνται για να περιγράψουν τον τρόπο εύρεσης, πρόσβασης και χρήσης των στοιχείων δεδομένων.
- Ένα παράδειγμα μεταδεδομένων είναι δεδομένα σχετικά με έναν αριθμό λογαριασμού.

Τύποι – Κατηγορίες Μεγάλων Δεδομένων

- Αυτό μπορεί να περιλαμβάνει τον αριθμό, την περιγραφή, τον τύπο δεδομένων, το όνομα, τη διεύθυνση, τον αριθμό τηλεφώνου και το επίπεδο προστασίας προσωπικών δεδομένων.
- Τα μεταδεδομένα μπορούν να χρησιμοποιηθούν για να βοηθήσουν να οργανώσουν τα δεδομένα και την οργάνωση των νέων και μεταβαλλόμενων πηγών δεδομένων.
- Παρόλο που η ιδέα των μεταδεδομένων δεν είναι νέα, αλλάζει και εξελίσσεται στο πλαίσιο των μεγάλων δεδομένων.

Πολυδιάστατα δεδομένα Multidimensional Data

- ◈ Όγκος πωλήσεων ως συνάρτηση product, month, region



**Διαστάσεις (Dimensions):
Product, Location, Time**

Ελλιπείς τιμές (missing values)

1	Gender	Race	Birth_Ye...	Marital_...	Years_o...	Hours_P...	Preferre...	Preferre...	Preferre...	Read_N...	Online_...	Online_...	Facel
2	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y
3	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y
4	F	African A...	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y
5	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N
6	M	White	1954	M	2	3	Internet ...	Bing	Hotmail	Y	Y	N	Y
7	M	African A...	1982	D	15	4	Internet ...	Google	Yahoo	Y	N	Y	N
8	M	African A...	1981	D	11	2	Firefox	Google	Yahoo		Y		Y
9	M	White	1977	S	3	3	Internet ...	Yahoo	Yahoo	Y			Y
10	F	African A...	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N
11	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y
12	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y

Ελλιπείς
τιμές

Ελλιπείς τιμές (missing values)

Προκειμένου να διορθώσουμε τις εγγραφές που παρουσιάζουν ελλιπείς τιμές μπορούμε να εφαρμόσουμε μια από τις παρακάτω μεθόδους

- ◆ Αντικατάσταση τιμής βάσει
 - ◆ του ΜΟ των υπολοίπων τιμών
 - ◆ Της μικρότερης τιμής
 - ◆ Της μεγαλύτερης τιμής
 - ◆ Να βάλουμε μια δική μας τιμή
 - ◆ Να βάλουμε το μηδέν.

Δεδομένα με ακραίες τιμές (outliers)

Μέσω της διαδικασίας «Outlier detection» μπορούμε να εντοπίσουμε αντικείμενα τα οποία παρουσιάζουν συμπεριφορά διαφορετική από την αναμενόμενη.

Σε αυτά τα αντικείμενα συνήθως εστιάζει ένας αναλυτής για περαιτέρω ανάλυση γιατί ενδεχομένως

- ◊ Να δείχνουν μια οικονομική απάτη
- ◊ Να ανιχνεύουν κάποιο φαινόμενο σε μια ανάλυση ιατρικών δεδομένων
- ◊ Να ανιχνεύουν παραβιάσεις στην ασφάλεια ενός δικτύου.

Βέβαια υπάρχουν και περιπτώσεις που μπορεί απλά να προστέθηκαν από λάθος στο δείγμα και να πρέπει να απομακρυνθούν για διορθωθεί τα δεδομένα.