Interreg
Ελλάδα-Κύπρος
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΥΡΩΠΑΪΚΗ ΕΝΩΣΗ
SmartWater2020

ΔΕΣΜΟΙ ΑΝΑΠΤΥΞΗΣ

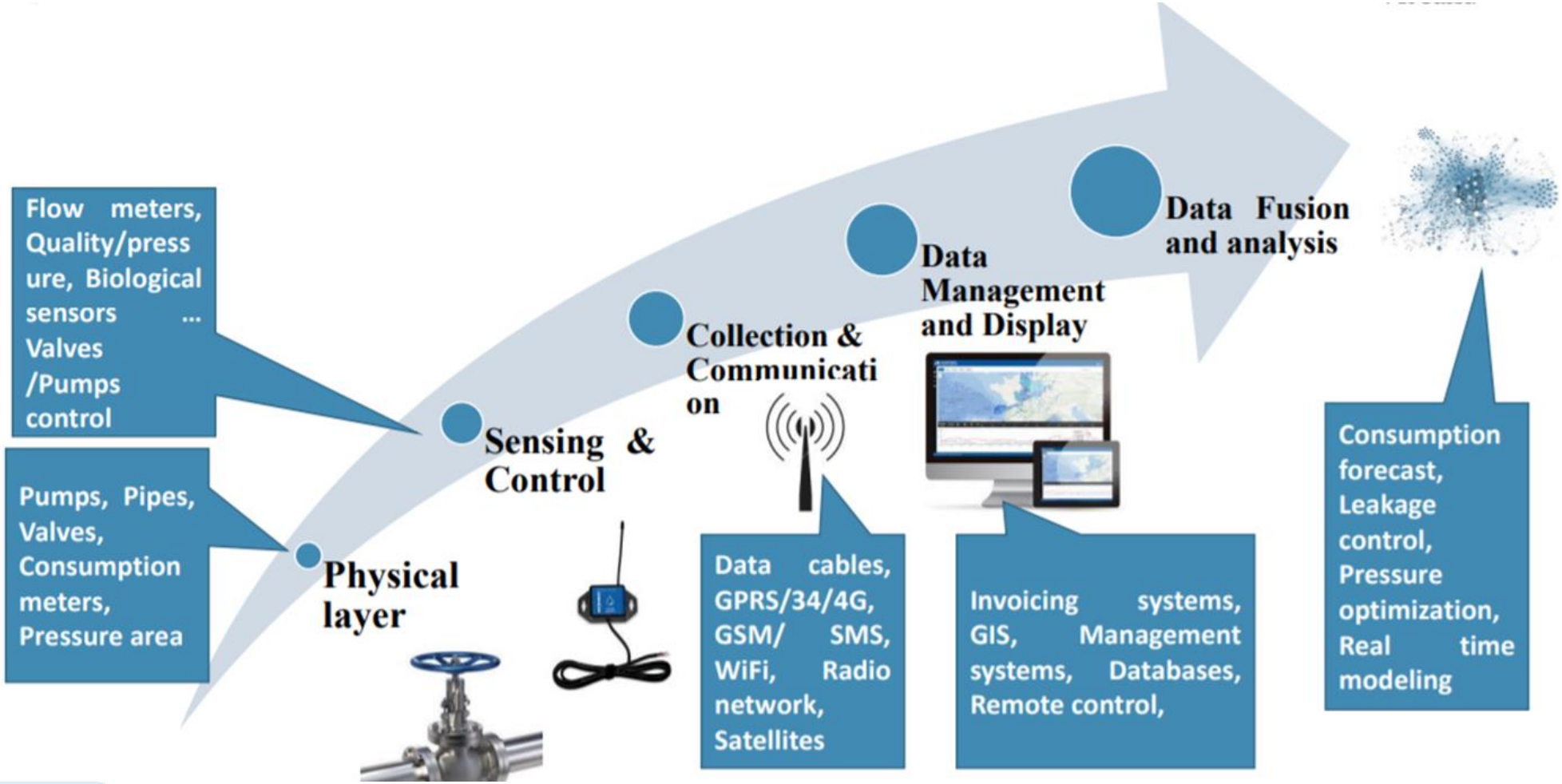
Ηράκλειο, 30/11/2020

Καλές Πρακτικές Ανάλυσης Δεδομένων –Το Μοντέλο CRISP-DM

Ρουμπάκης Στέλιος, Μηχανικός Λογισμικού, ΙΤΕ-ΙΠ

Εισαγωγή σε καλές πρακτικές ανάλυσης δεδομένων: Το μοντέλο CRISP-DM

Ανάλυση Δεδομένων


Flow meters, Quality/pressure, Biological sensors ... Valves /Pumps control
Pumps, Pipes, Valves, Consumption meters, Pressure area
Physical layer
Sensing & Control
Collection & Communication
Data Management and Display
Data Fusion and analysis
Data cables, GPRS/34/4G, GSM/ SMS, WiFi, Radio network, Satellites
Invoicing systems, GIS, Management systems, Databases, Remote control,
Consumption forecast, Leakage control, Pressure optimization, Real time modeling

Κύκλος Ζωής Δεδομένων

Διαδικασία Εξόρυξης Δεδομένων (**CRISP-DM**):

Μια διαδικασία που στοχεύει στην αύξηση της χρήσης δεδομένων σε μια μεγάλη ποικιλία επιχειρηματικών και βιομηχανικών εφαρμογών

Ορισμός Στόχων

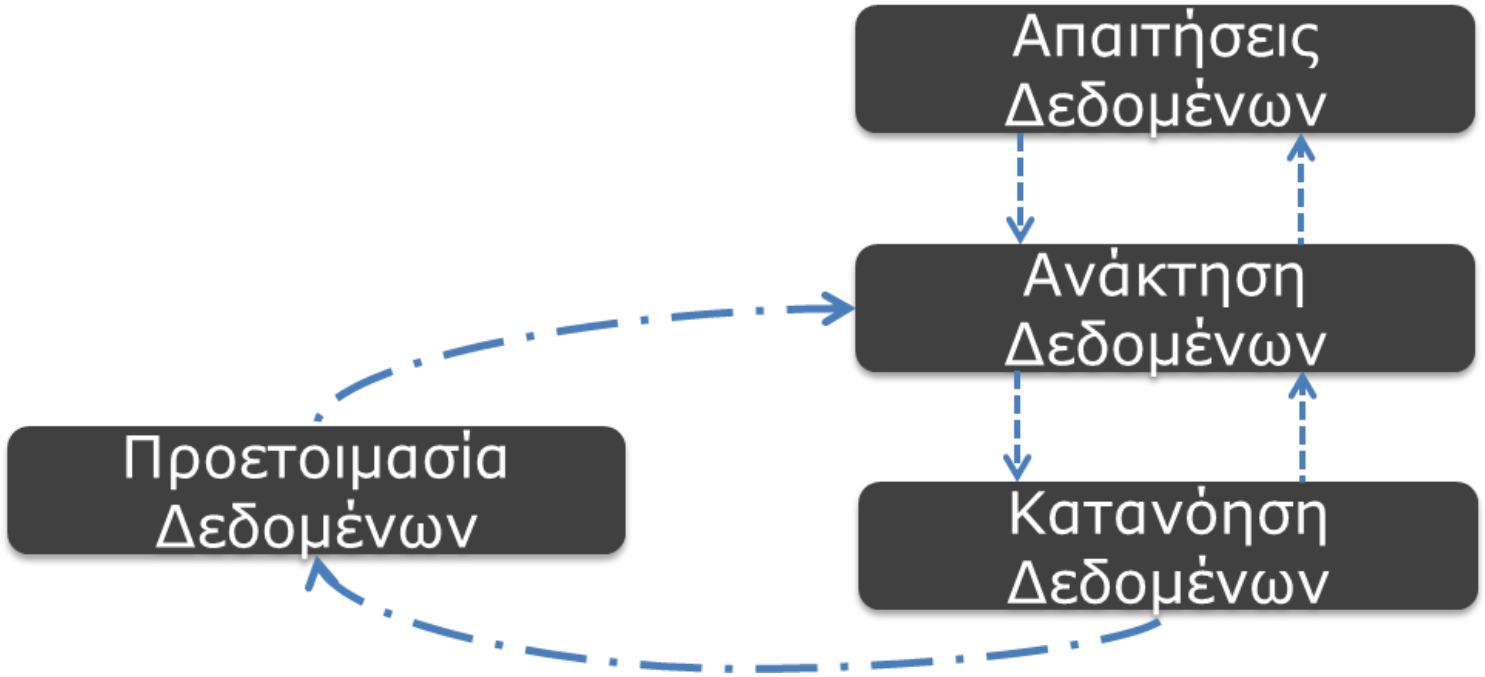
- Τι προσπαθούμε να κάνουμε - Ποιος είναι ο στόχος του έργου?
- Οι χορηγοί του εγχειρήματος παίζουν τον πιο κρίσιμο ρόλο
- Πώς ορίζεται η επιτυχία και πώς μπορούμε να τη μετρήσουμε?

Κατανόηση Εγχειρήματος

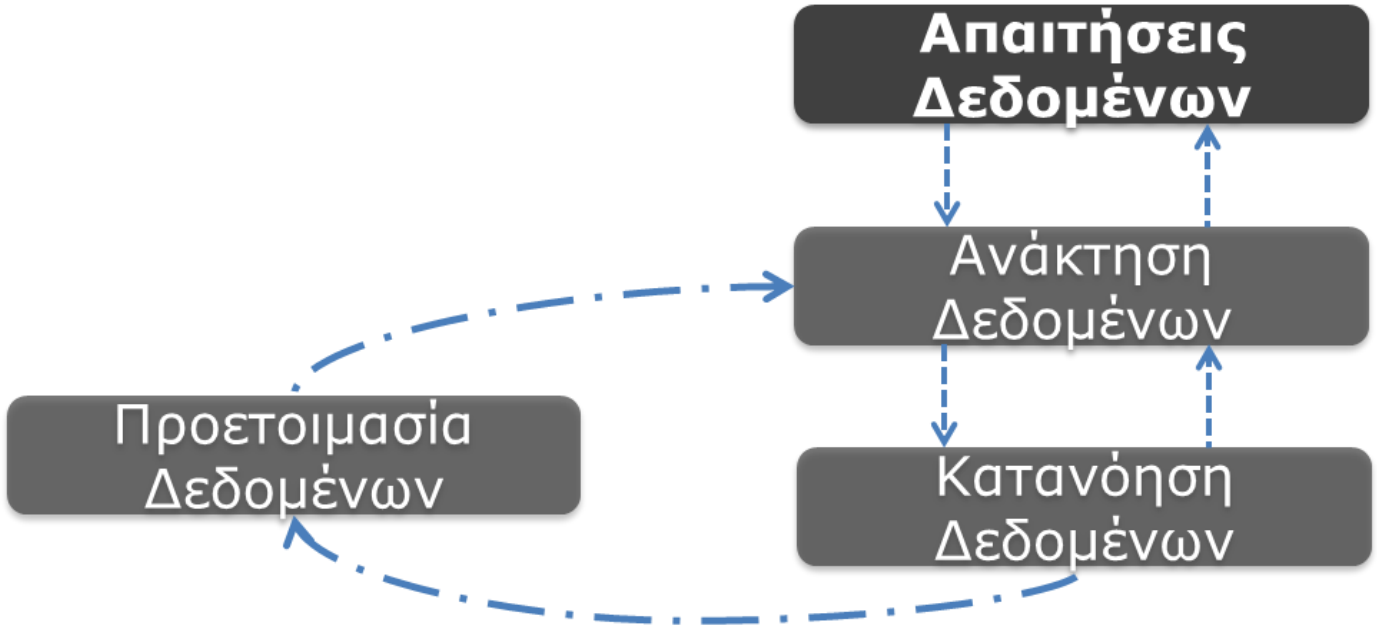
Εκφράζουμε το πρόβλημα στο πλαίσιο των τεχνικών στατιστικής και μηχανικής μάθησης

- **Regression**: "Πρόβλεψη εσόδων το επόμενο τρίμηνο?"
- **Classification**: "Η ποιότητα του νερού είναι κατηγορίας Α ή κατηγορίας Β?"
- **Clustering**: "Υπάρχουν ομάδες καταναλωτές που έχουν παρόμοια συμπεριφορά?"
- **Recommendation/Personalization**: "Πώς μπορώ να προσφέρω στοχευμένες εκπτώσεις σε συγκεκριμένους πελάτες?"
- **Outlier Detection:** "Πώς μπορώ να εντοπίσω μια διαρροή? "

Συλλογή Δεδομένων



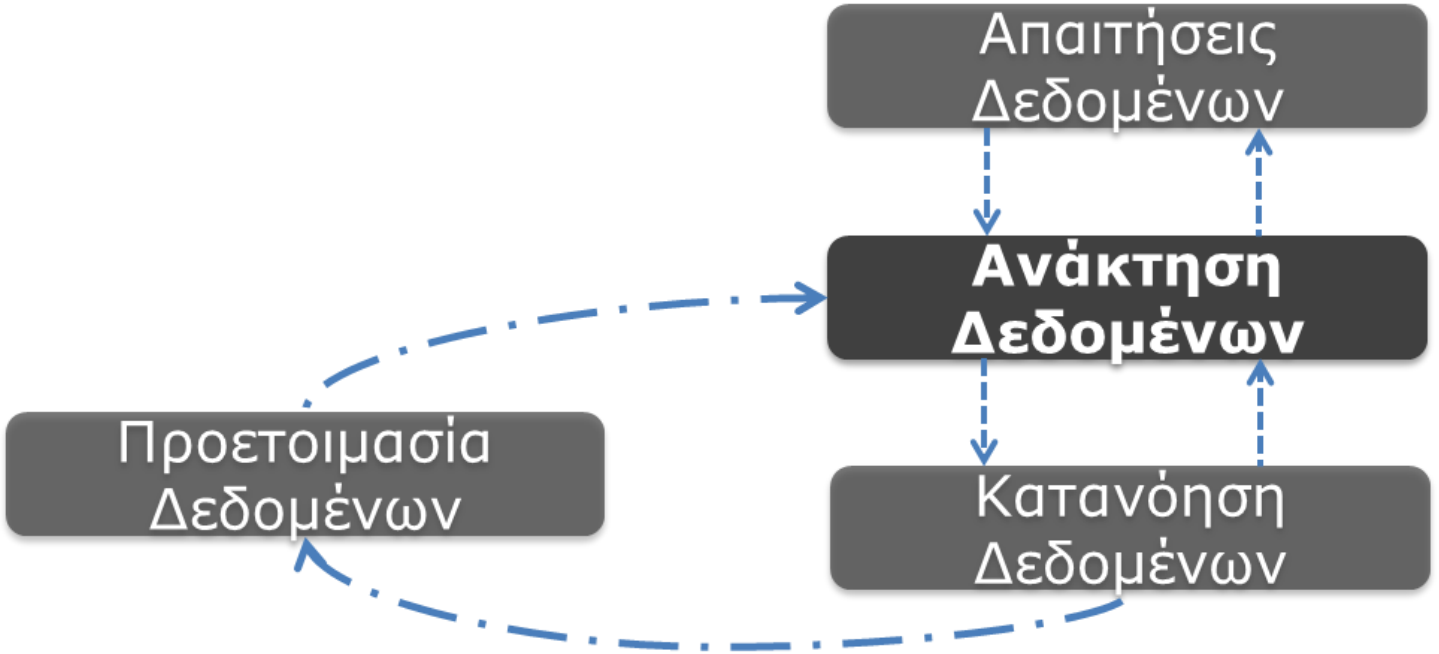
Συλλογή Δεδομένων



Η **Αναλυτική Προσέγγιση** καθορίζει τις **Απαιτήσεις Δεδομένων** και συγκεκριμένα πληροφορίες για:

- Το περιεχόμενο
- Τη δομή
- Και το τρόπο αναπαράστασης

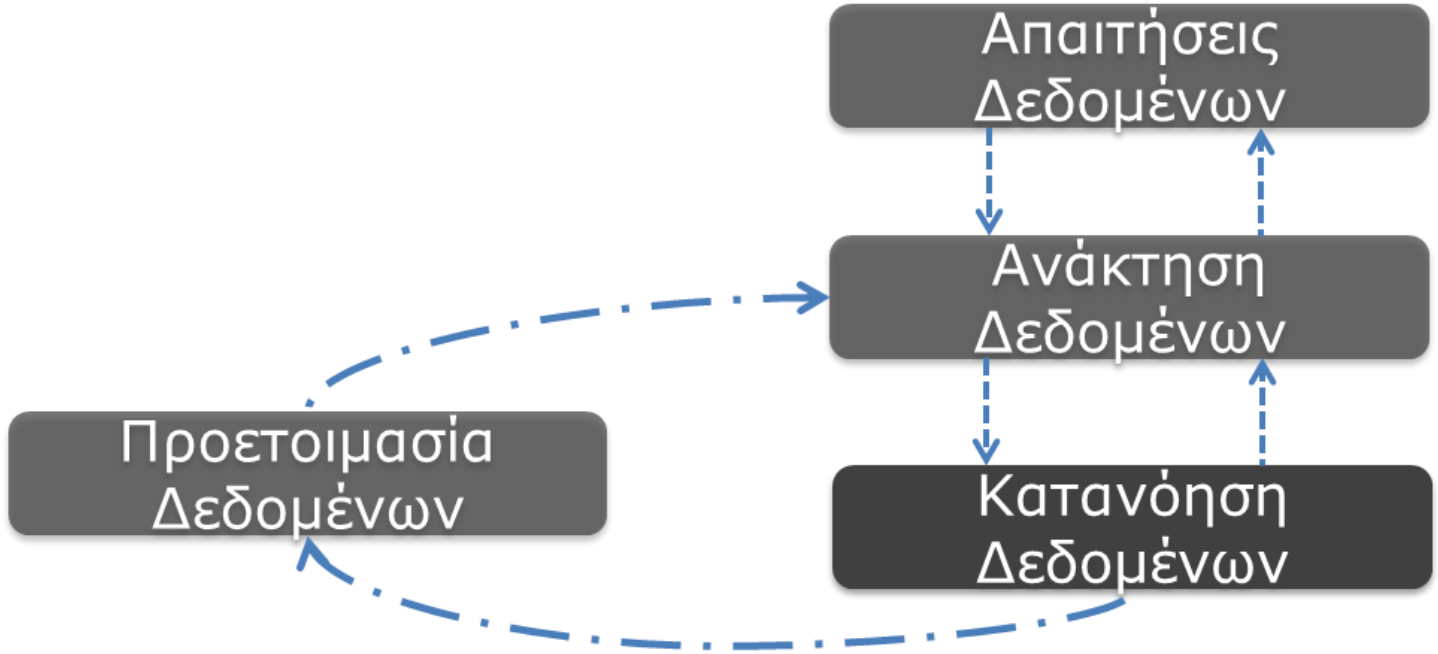
Συλλογή Δεδομένων



Αρχική **Ανάκτηση Δεδομένων**

- *Διαθεσιμότητα Δεδομένων?*
- *Μέθοδοι ανάκτησης δεδομένων?*
- *Αναθεώρηση των απαιτήσεων των δεδομένων*

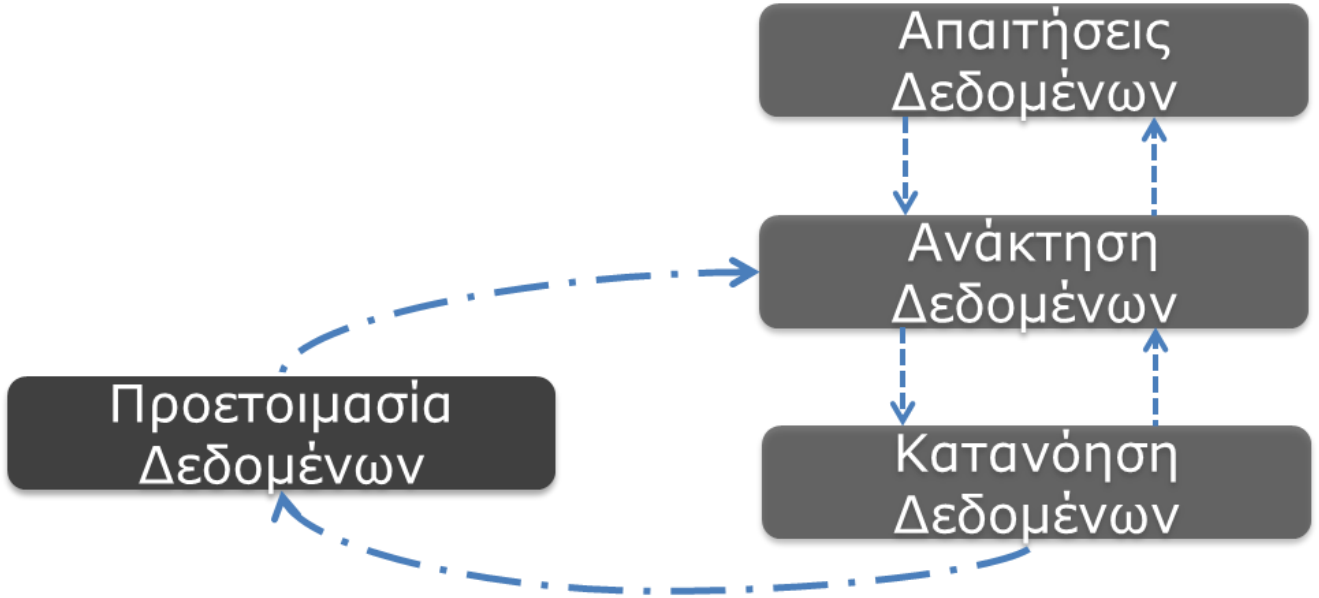
Συλλογή Δεδομένων



Έπειτα η **Κατανόηση Δεδομένων** προσφέρει:

- ❑ Αρχική εκτίμηση για τη φύση των δεδομένων
- ❑ Περιγραφικά στατιστικά στοιχεία και οπτικοποίηση
- ❑ Πρόσθετη συλλογή δεδομένων για την κάλυψη κενών, εάν χρειάζεται

Συλλογή Δεδομένων



*Η **Προετοιμασία Δεδομένων** περιλαμβάνει όλες τις ενέργειες καθαρισμού και δόμησης των δεδομένων*

Καθαρισμός Δεδομένων

- *Χαμένες ή μη έγκυρες μετρήσεις*
- *Εξάλειψη διπλών σειρών*
- *Σωστή μορφοποίηση*

Συνδυασμός πολλαπλών πηγών δεδομένων

Συλλογή Δεδομένων

- Ανάπτυξη προγνωστικών ή περιγραφικών μοντέλων
- Μπορούν να δοκιμαστούν πολλαπλοί διαφορετικοί αλγόριθμοι
- Συνεχώς επαναληπτική διαδικασία

Μοντελοποιήση

Η **Αξιολόγηση** του μοντέλου πραγματοποιείται κατά τη διάρκεια της ανάπτυξης του και πριν την **Εφαρμογή**

- Κατανόηση της ποιότητας του μοντέλου
- Έλεγχος επίτευξης των λειτουργικών στόχων που έχουν τεθεί
- Διαγνωστικές μετρικές
    - *Κατάλληλες για τη τεχνική μοντελοποίησης που χρησιμοποιείται*
    - *Διαχωρισμός δεδομένων σε Εκπαίδευση/Επαλήθευση*
    - *Αναπροσαρμογή του μοντέλου βάσει αποτελεσμάτων*
- Tests στατιστικής σημασίας

Ανάπτυξη / Εφαρμογή

Μόλις οριστικοποιηθεί, το μοντέλο **εφαρμόζεται** σε περιβάλλον παραγωγής

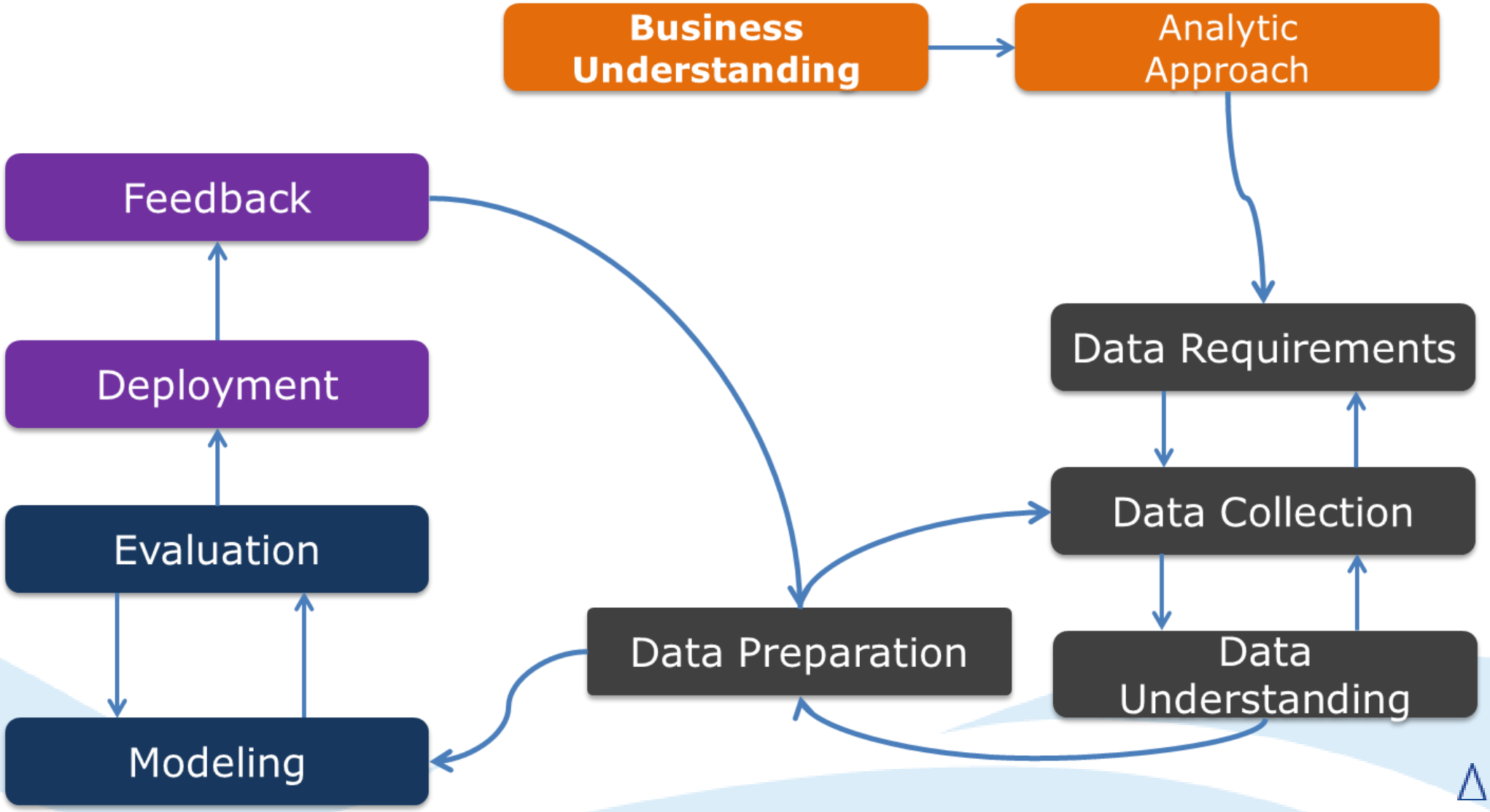
- Παρόλα αυτά ξεκινάει σε ένα περιορισμένο περιβάλλον δοκιμών
- Στην **εφαρμογή** ενδέχεται να συμμετέχουν:
  - *Χορηγοί*
  - *Τμήμα Marketing*
  - *Προγραμματιστές εφαρμογών*
  - *Διαχειριστές των συστημάτων*

Ανάπτυξη / Εφαρμογή

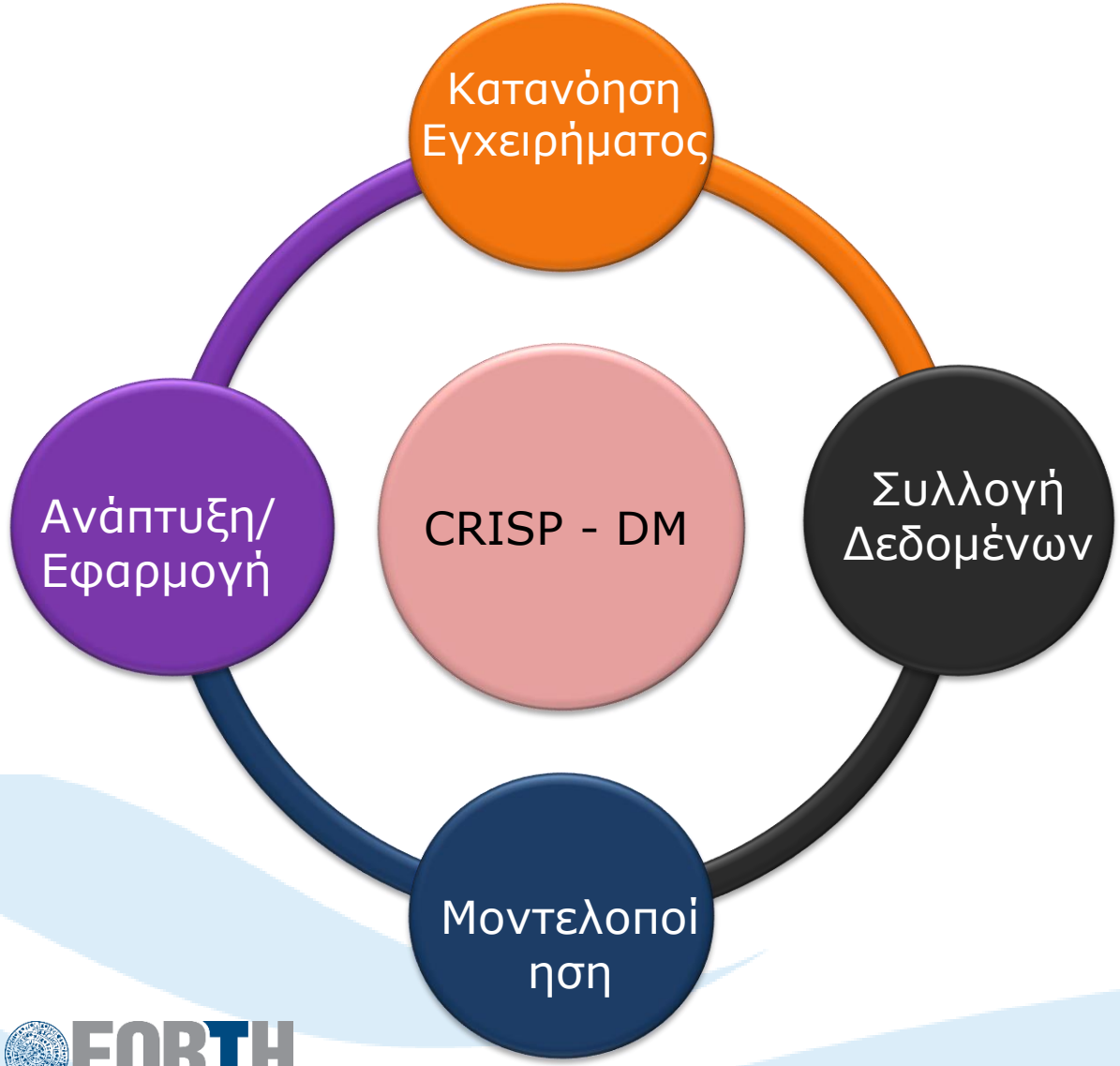
Ανάδραση:

- Πόσο καλή είναι η απόδοση του μοντέλου?
- Επαναληπτική διαδικασία για τη βελτίωση και επανεφαρμογή του μοντέλου
- A/B testing

CRISP-DM


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback

CRISP-DM



- **Συμβούλιο:**
- **Data Scientist:**
- **Μηχανικοί-Τεχνικοί:**
- **Τμήμα Οικονομικών:**

CRISP-DM

Το CRISP-DM είναι ένα εξαιρετικά ευέλικτο και επαναληπτικό μοντέλο

- ✓ **Ευελιξία:** απαιτείται σε κάθε βήμα, σε συνδυασμό με την επικοινωνία ούτως ώστε να διατηρηθεί ο έλεγχος του έργου. Ίσως χρειαστεί να επανεξετάσουμε κάποιο προηγούμενο βήμα και να γίνουν αλλαγές

- ✓ **Επαναληψιμότητα**: ακόμα και μετά το πέρας του εγχειρήματος ενδέχεται να αναπροσδιοριστούν οι στόχοι του για την εξασφάλιση της βιωσιμότητας του έργου