Από τα Δεδομένα στις Γνώσεις και στις Αποφάσεις

Fundamentals of Machine Learning for Predictive Data Analytics

Αθανάσιος Σάκκας, ΟΠΑ

1. **Μετατροπή Επιχειρηματικών Προβλημάτων σε Αναλυτικές Λύσεις**
   - Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου
2. **Αξιολόγηση Εφικτότητας**
   - Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου
3. **Σχεδιασμός του Analytics Base Table**
   - Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου
4. **Σχεδιασμός & Υλοποίηση Χαρακτηριστικών**
   - Διαφορετικοί τύποι δεδομένων
   - Διαφορετικοί τύποι χαρακτηριστικών
   - Χειρισμός του χρόνου
   - Νομικά ζητήματα
   - Υλοποίηση χαρακτηριστικών
   - Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου
5. **Σύνοψη**

Μετατροπή Επιχειρηματικών Προβλημάτων σε Αναλυτικές Λύσεις

- Η μετατροπή ενός επιχειρηματικού προβλήματος σε μια αναλυτική λύση περιλαμβάνει την απάντηση στα ακόλουθα βασικά ερωτήματα:
  1. Ποιο είναι το επιχειρηματικό πρόβλημα;
  2. Ποιοι είναι οι στόχοι που θέλει να πετύχει η επιχείρηση;
  3. Πώς λειτουργεί σήμερα η επιχείρηση;
  4. Με ποιους τρόπους θα μπορούσε ένα μοντέλο προγνωστικής αναλυτικής να βοηθήσει στην αντιμετώπιση του επιχειρηματικού προβλήματος;

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

Παρά το γεγονός ότι υπάρχει ομάδα διερεύνησης απάτης που ερευνά έως και το 30% όλων των απαιτήσεων αποζημίωσης, μια ασφαλιστική εταιρεία αυτοκινήτου εξακολουθεί να χάνει πάρα πολλά χρήματα λόγω δόλιων απαιτήσεων.

- Ποιες λύσεις προγνωστικής αναλυτικής θα μπορούσαν να προταθούν για να βοηθήσουν στην αντιμετώπιση αυτού του επιχειρηματικού προβλήματος;

- Πιθανές αναλυτικές λύσεις περιλαμβάνουν:
  - Πρόβλεψη απαίτησης (Claim prediction)
  - Πρόβλεψη πελάτη/μέλους (Member prediction)
  - Πρόβλεψη αίτησης (Application prediction)
  - Πρόβλεψη πληρωμής (Payment prediction)

Αξιολόγηση Εφικτότητας

- Η αξιολόγηση της εφικτότητας μιας προτεινόμενης αναλυτικής λύσης περιλαμβάνει τα εξής ερωτήματα:
  1. Είναι διαθέσιμα τα δεδομένα που απαιτούνται από τη λύση ή θα μπορούσαν να γίνουν διαθέσιμα;
  2. Ποια είναι η ικανότητα της επιχείρησης να αξιοποιήσει τις γνώσεις/ευρήματα που θα παρέχει η αναλυτική λύση;

- Ποιες είναι οι απαιτήσεις σε δεδομένα και «ικανότητα αξιοποίησης» (capacity) για την προτεινόμενη λύση **Πρόβλεψης Απαίτησης** στο σενάριο απάτης στην ασφάλιση αυτοκινήτου;

- Ποιες είναι οι απαιτήσεις σε δεδομένα και «ικανότητα αξιοποίησης» (capacity) για την προτεινόμενη λύση **Πρόβλεψης Απαίτησης** στο σενάριο απάτης στην ασφάλιση αυτοκινήτου;

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

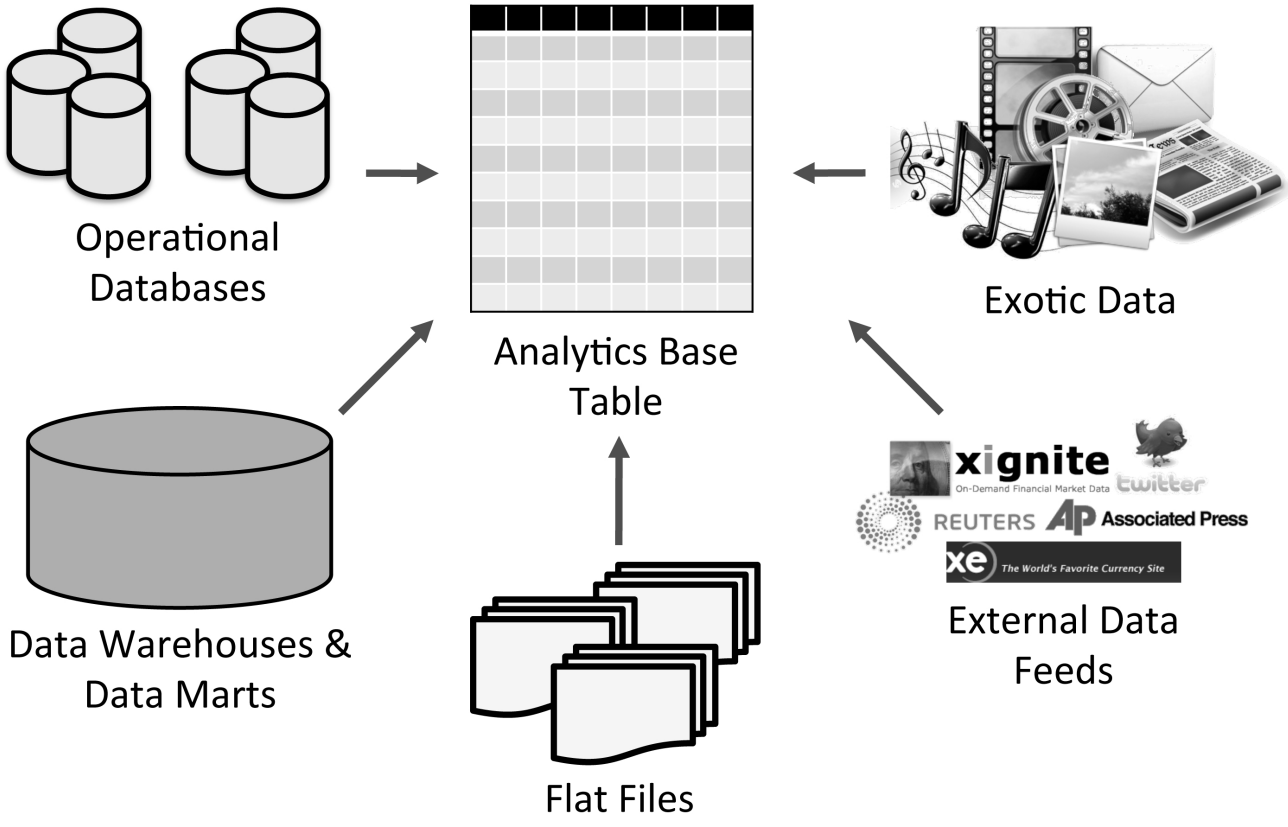
[Πρόβλεψη απαίτησης]
Απαιτήσεις δεδομένων: Μια μεγάλη συλλογή ιστορικών απαιτήσεων με σήμανση *'δόλια'* και *'μη δόλια'*. Επίσης, πρέπει να είναι διαθέσιμες οι λεπτομέρειες κάθε απαίτησης, του σχετικού συμβολαίου και του αιτούντος/ασφαλισμένου.
Απαιτήσεις ικανότητας (capacity): Η κύρια απαίτηση είναι να υπάρχει μηχανισμός που να ενημερώνει τους ερευνητές/πραγματογνώμονες ότι ορισμένες απαιτήσεις έχουν προτεραιότητα έναντι άλλων. Αυτό απαιτεί επίσης η πληροφορία για τις απαιτήσεις να γίνεται διαθέσιμη έγκαιρα, ώστε η διαδικασία διερεύνησης να μην καθυστερεί λόγω του μοντέλου.

Σχεδιασμός του Analytics Base Table

- Η βασική δομή με την οποία αποτυπώνουμε ιστορικά σύνολα δεδομένων είναι ο **analytics base table** (**ABT**).

Descriptive Features						Target Feature
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----

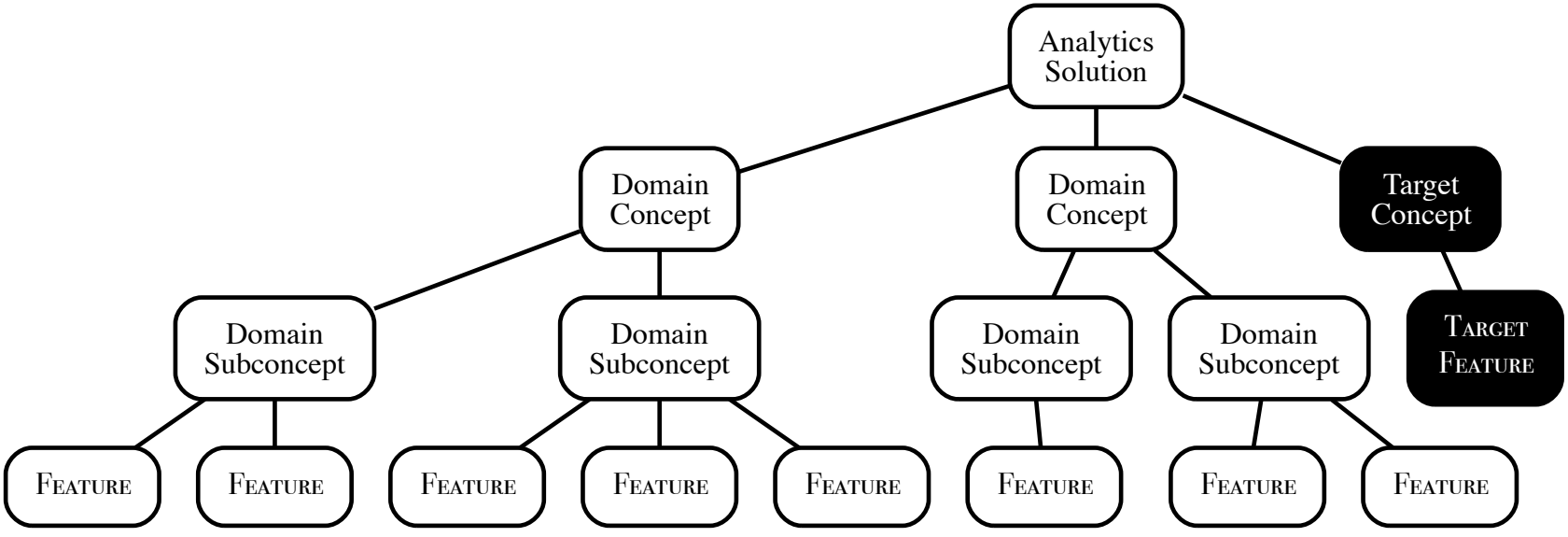
Σχήμα: Η γενική δομή ενός **analytics base table**---περιγραφικά χαρακτηριστικά και ένα χαρακτηριστικό-στόχος.



Σχήμα: Οι διαφορετικές πηγές δεδομένων που συνήθως συνδυάζονται για να δημιουργηθεί ένα ABT.

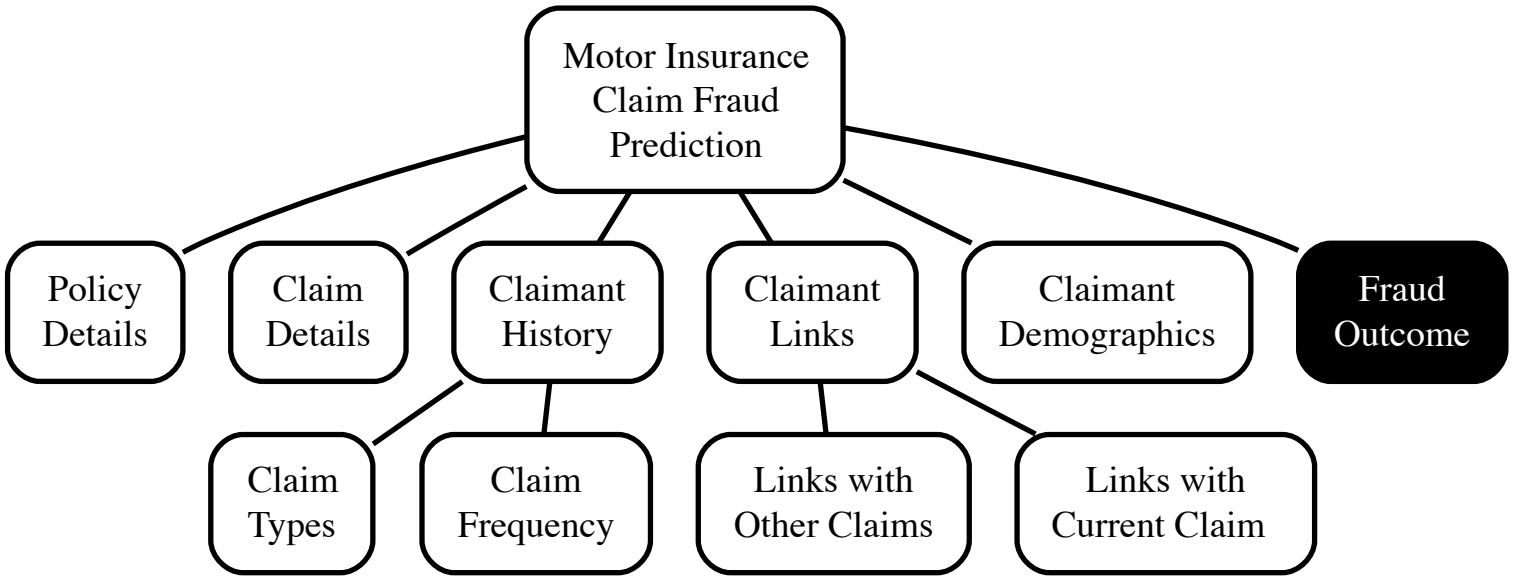
- Το **αντικείμενο πρόβλεψης** (prediction subject) ορίζει το βασικό επίπεδο στο οποίο γίνονται οι προβλέψεις και κάθε γραμμή στο ABT αντιστοιχεί σε μία οντότητα του αντικειμένου πρόβλεψης---συχνά χρησιμοποιείται η φράση **μία-γραμμή-ανά-οντότητα** (one-row-per-subject).
- Κάθε γραμμή σε ένα ABT αποτελείται από ένα σύνολο περιγραφικών χαρακτηριστικών και ένα χαρακτηριστικό-στόχο.
- Ο ορισμός χαρακτηριστικών μπορεί να είναι δύσκολος!

- Ένας καλός τρόπος για να ορίσουμε χαρακτηριστικά είναι να εντοπίσουμε τις βασικές **έννοιες πεδίου** (domain concepts) και στη συνέχεια να βασίσουμε τα χαρακτηριστικά σε αυτές τις έννοιες.



Σχήμα: Η ιεραρχική σχέση μεταξύ μιας αναλυτικής λύσης, των εννοιών πεδίου και των περιγραφικών χαρακτηριστικών.

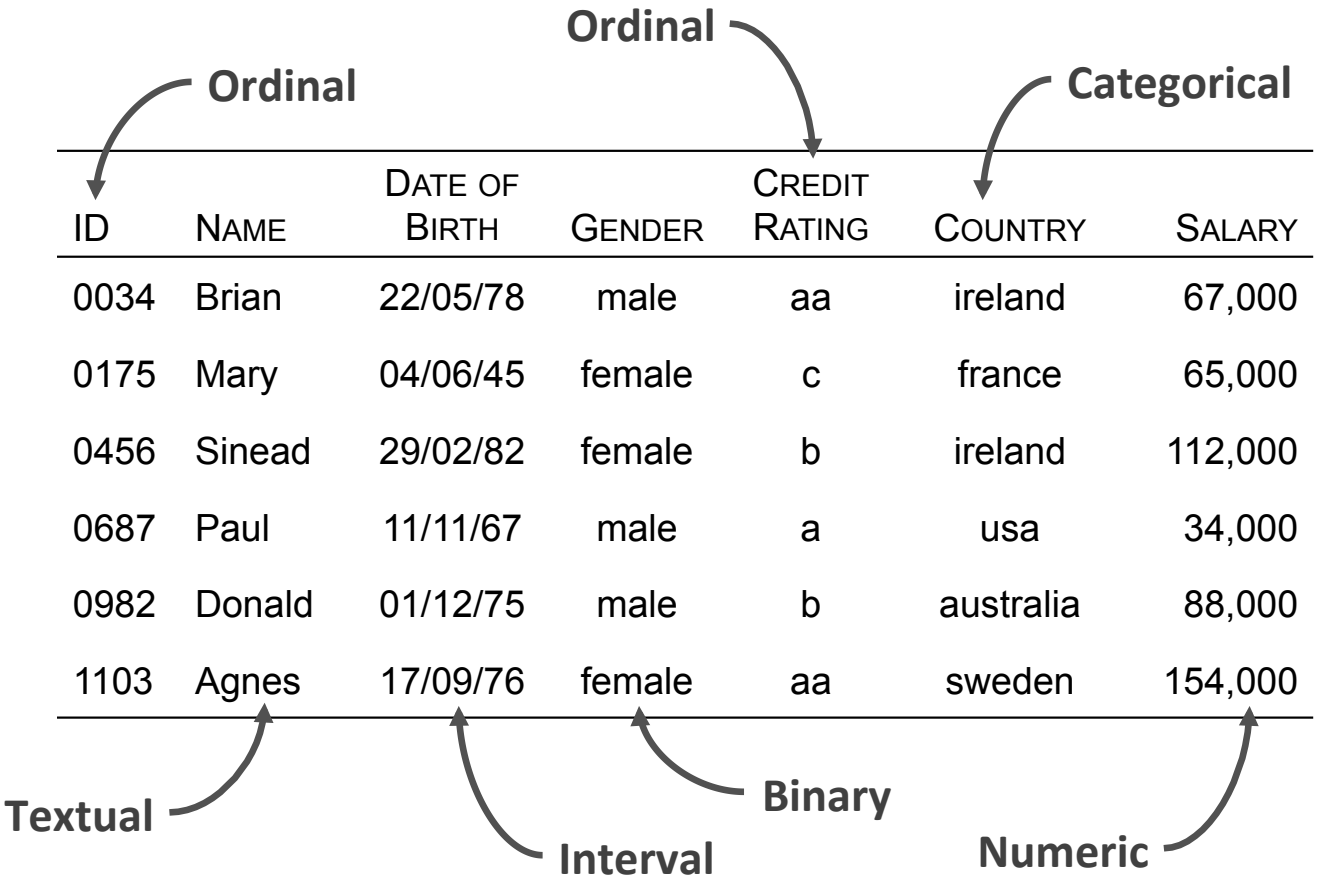
- Υπάρχουν ορισμένες γενικές έννοιες πεδίου που συχνά είναι χρήσιμες:
  - Λεπτομέρειες αντικειμένου πρόβλεψης
  - Δημογραφικά
  - Χρήση
  - Μεταβολές στη χρήση
  - Ειδική χρήση
  - Φάση κύκλου ζωής
  - Συνδέσεις δικτύου (network links)



Σχήμα: Ενδεικτικές έννοιες πεδίου για μια λύση πρόβλεψης δόλιων απαιτήσεων στην ασφάλιση αυτοκινήτου.

Σχεδιασμός & Υλοποίηση Χαρακτηριστικών

- Τρεις βασικές παράμετροι των δεδομένων είναι ιδιαίτερα σημαντικές όταν σχεδιάζουμε χαρακτηριστικά.
  - **Διαθεσιμότητα δεδομένων**
  - **Χρονισμός**
  - **Διάρκεια/διαχρονικότητα** (longevity)

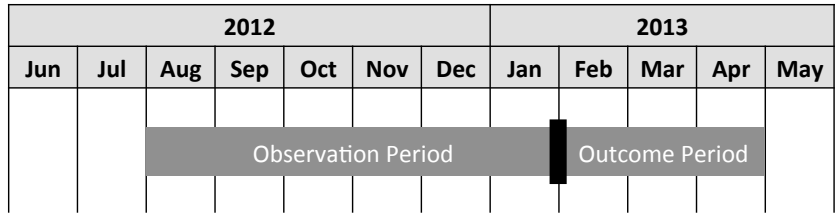


Σχήμα: Δείγμα περιγραφικών δεδομένων που αναδεικνύει αριθμητικούς, δυαδικούς, διατακτικούς, διαστημικούς, κατηγορικούς και κειμενικούς τύπους.

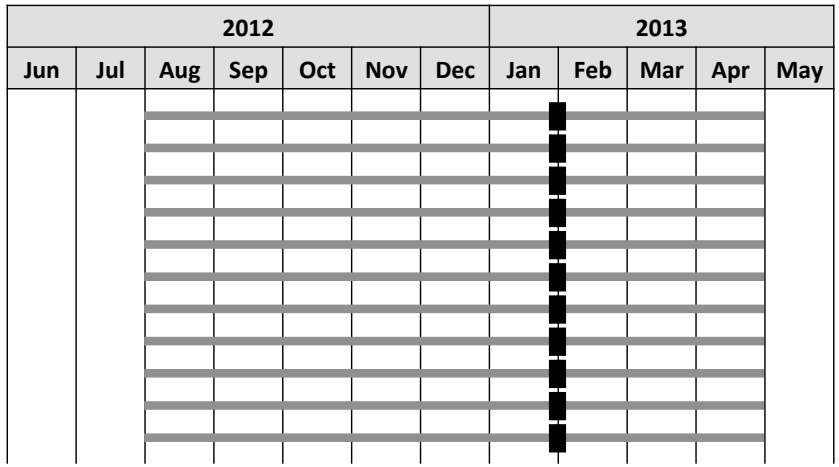
- Τα χαρακτηριστικά σε ένα ABT μπορούν να είναι δύο τύπων:
  - **πρωτογενή (raw) χαρακτηριστικά**
  - **παράγωγα (derived) χαρακτηριστικά**
- Υπάρχουν αρκετοί συνηθισμένοι τύποι παράγωγων χαρακτηριστικών:
  - **Συναθροίσεις** (Aggregates)
  - **Δείκτες/Σημαίες** (Flags)
  - **Λόγοι** (Ratios)
  - **Αντιστοιχίσεις** (Mappings)

- Πολλά από τα προγνωστικά μοντέλα που κατασκευάζουμε είναι **μοντέλα προδιάθεσης** (propensity models), τα οποία έχουν εγγενώς χρονικό στοιχείο.
- Για **μοντελοποίηση προδιάθεσης** υπάρχουν δύο βασικές περίοδοι:
  - η **περίοδος παρατήρησης** (observation period)
  - η **περίοδος αποτελέσματος** (outcome period)

- Σε ορισμένες περιπτώσεις, η περίοδος παρατήρησης και η περίοδος αποτελέσματος μετρώνται στις ίδιες ημερομηνίες για όλα τα αντικείμενα πρόβλεψης.

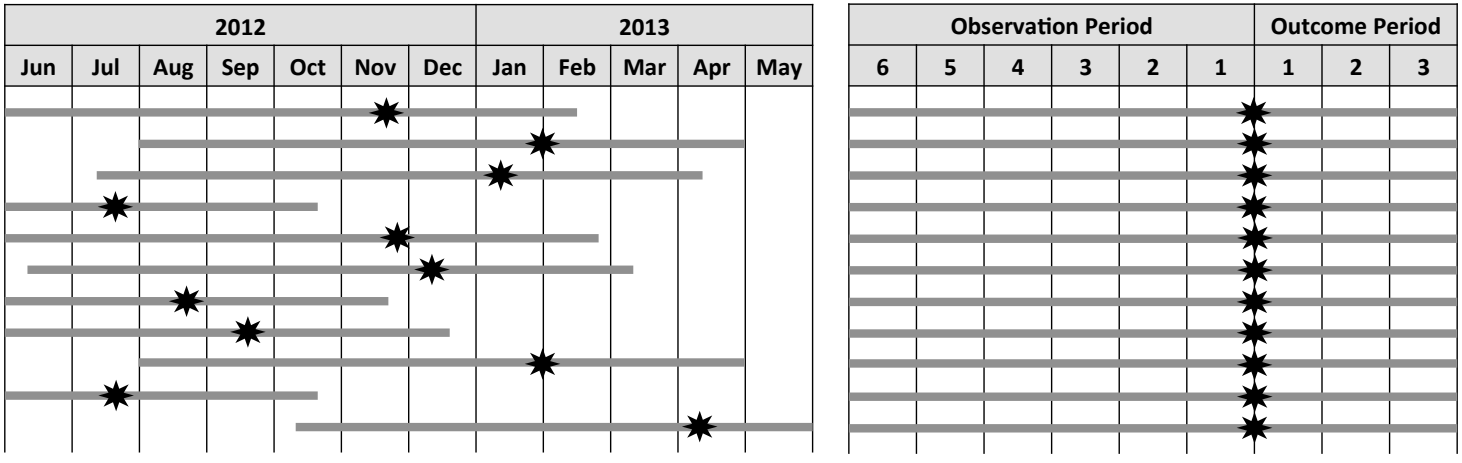


(α□) Περίοδος παρατήρησης και περί-οδος αποτελέσματος



(β□) Περίοδος παρατήρη-σης/αποτελέσματος για πολλούς πελάτες (κάθε γραμμή αντιστοιχεί σε πελάτη)

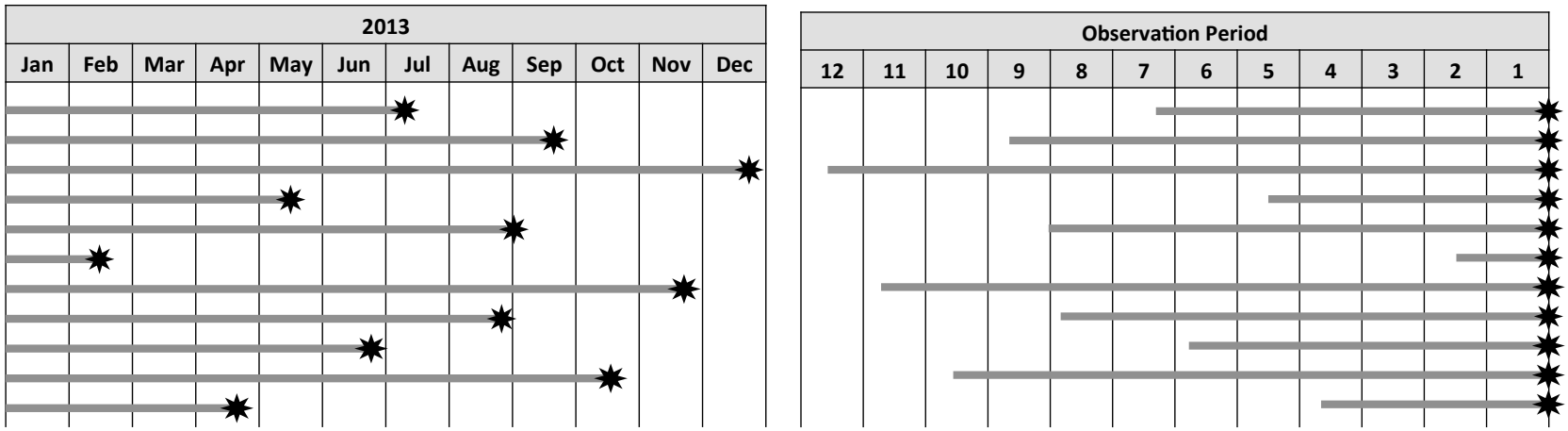
- Συχνά, η περίοδος παρατήρησης και η περίοδος αποτελέσματος μετρώνται σε διαφορετικές ημερομηνίες για κάθε αντικείμενο πρόβλεψης.



(α□) Πραγματικό (β□) Ευθυγραμμισμένο

Σχήμα: Περίοδοι παρατήρησης και αποτελέσματος ορισμένες από ένα γεγονός, και όχι από σταθερό χρονικό σημείο (κάθε γραμμή είναι ένα αντικείμενο πρόβλεψης και τα αστέρια συμβολίζουν γεγονότα).

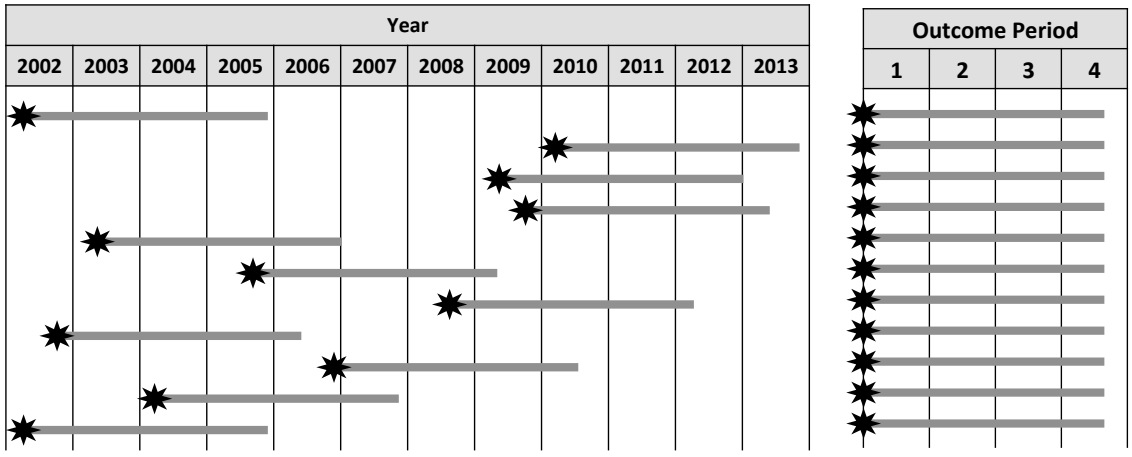
- Σε ορισμένες περιπτώσεις, μόνο τα περιγραφικά χαρακτηριστικά έχουν χρονικό στοιχείο, ενώ το χαρακτηριστικό-στόχος είναι χρονικά ανεξάρτητο.



(α□) Πραγματικό (β□) Ευθυγραμμισμένο

Σχήμα: Μοντελοποίηση χρονικών «σημείων» για σενάριο χωρίς πραγματική περίοδο αποτελέσματος (κάθε γραμμή αντιστοιχεί σε πελάτη, και τα αστέρια συμβολίζουν γεγονότα).

- Αντίστροφα, το χαρακτηριστικό-στόχος μπορεί να έχει χρονικό στοιχείο ενώ τα περιγραφικά χαρακτηριστικά να μην έχουν.



(α□) Πραγματικό (β□) Ευθυγραμμισμένο

Σχήμα: Μοντελοποίηση χρονικών «σημείων» για σενάριο χωρίς πραγματική περίοδο παρατήρησης (κάθε γραμμή αντιστοιχεί σε πελάτη, και τα αστέρια συμβολίζουν γεγονότα).

- Οι επαγγελματίες της αναλυτικής δεδομένων συχνά απογοητεύονται από νομοθεσίες που τους εμποδίζουν να συμπεριλάβουν χαρακτηριστικά τα οποία φαίνονται ιδιαίτερα κατάλληλα σε ένα ABT.
- Υπάρχουν σημαντικές διαφορές στη νομοθεσία ανά δικαιοδοσία, όμως μερικές βασικές αρχές σχεδόν πάντα ισχύουν:
  1. **Νομοθεσία κατά των διακρίσεων**
  2. **Νομοθεσία προστασίας δεδομένων**

- Αν και η νομοθεσία προστασίας δεδομένων αλλάζει σημαντικά μεταξύ διαφορετικών δικαιοδοσιών, υπάρχουν ορισμένες κοινές αρχές που επηρεάζουν τον σχεδιασμό ABT:
  - **Αρχή περιορισμού της συλλογής** (collection limitation)
  - **Αρχή καθορισμού σκοπού** (purpose specification)
  - **Αρχή περιορισμού χρήσης** (use limitation)

- Η υλοποίηση ενός **παράγωγου χαρακτηριστικού**, ωστόσο, απαιτεί τον συνδυασμό δεδομένων από πολλαπλές πηγές σε μία ενιαία τιμή χαρακτηριστικού.
- Μερικές βασικές **λειτουργίες χειρισμού δεδομένων** χρησιμοποιούνται συχνά για τον υπολογισμό παράγωγων χαρακτηριστικών:
  - συνένωση πηγών δεδομένων (join)
  - φιλτράρισμα γραμμών (rows)
  - φιλτράρισμα πεδίων (fields)
  - παραγωγή νέων χαρακτηριστικών με συνδυασμό ή μετασχηματισμό υπαρχόντων
  - συναθροίσεις/ομαδοποιήσεις (aggregation)

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

- Ποιες είναι η περίοδος παρατήρησης και η περίοδος αποτελέσματος στο σενάριο πρόβλεψης δόλιων απαιτήσεων;

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

- Ποιες είναι η περίοδος παρατήρησης και η περίοδος αποτελέσματος στο σενάριο πρόβλεψης δόλιων απαιτήσεων;
- Η περίοδος παρατήρησης και η περίοδος αποτελέσματος μετρώνται σε διαφορετικές ημερομηνίες για κάθε απαίτηση, οριζόμενες σε σχέση με την ημερομηνία της συγκεκριμένης απαίτησης.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

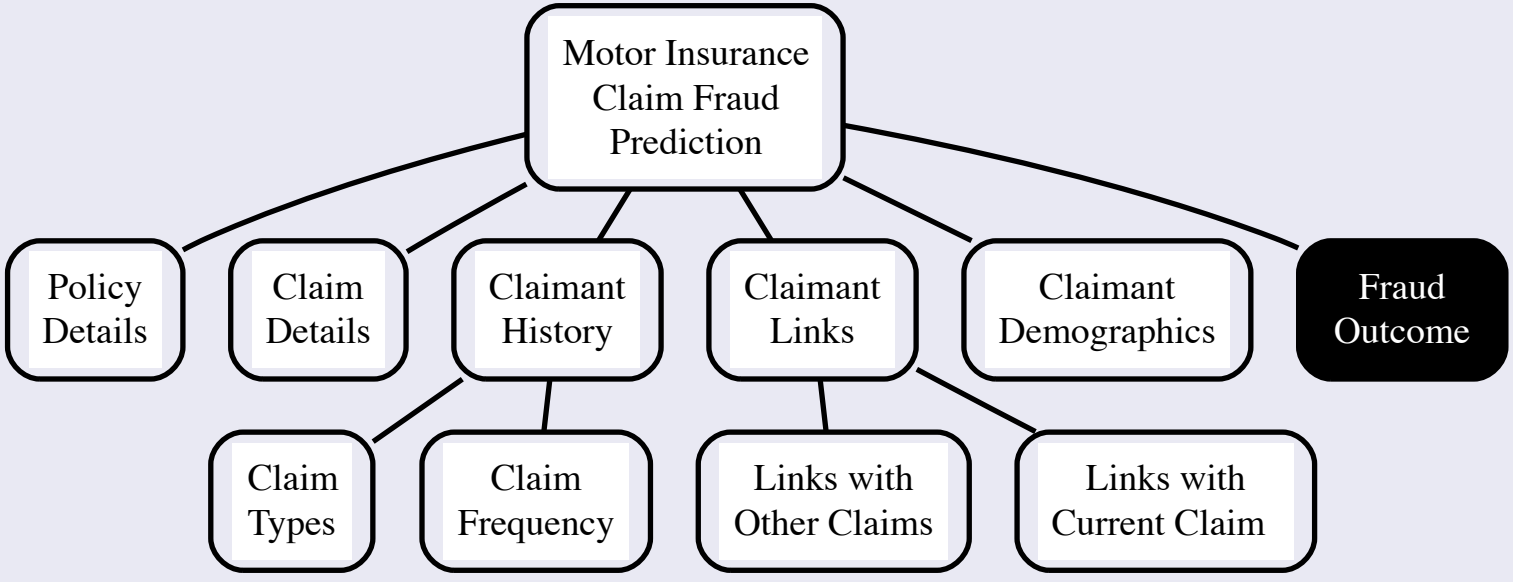
- Ποιες είναι η περίοδος παρατήρησης και η περίοδος αποτελέσματος στο σενάριο πρόβλεψης δόλιων απαιτήσεων;
- Η περίοδος παρατήρησης και η περίοδος αποτελέσματος μετρώνται σε διαφορετικές ημερομηνίες για κάθε απαίτηση, οριζόμενες σε σχέση με την ημερομηνία της συγκεκριμένης απαίτησης.
- Η περίοδος παρατήρησης είναι ο χρόνος πριν από το γεγονός της απαίτησης, στον οποίο υπολογίζονται τα περιγραφικά χαρακτηριστικά που αποτυπώνουν τη συμπεριφορά του αιτούντος/ασφαλισμένου.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

- Ποιες είναι η περίοδος παρατήρησης και η περίοδος αποτελέσματος στο σενάριο πρόβλεψης δόλιων απαιτήσεων;
- Η περίοδος παρατήρησης και η περίοδος αποτελέσματος μετρώνται σε διαφορετικές ημερομηνίες για κάθε απαίτηση, οριζόμενες σε σχέση με την ημερομηνία της συγκεκριμένης απαίτησης.
- Η περίοδος παρατήρησης είναι ο χρόνος πριν από το γεγονός της απαίτησης, στον οποίο υπολογίζονται τα περιγραφικά χαρακτηριστικά που αποτυπώνουν τη συμπεριφορά του αιτούντος/ασφαλισμένου.
- Η περίοδος αποτελέσματος είναι ο χρόνος αμέσως μετά το γεγονός της απαίτησης, κατά τον οποίο θα προκύψει αν η απαίτηση είναι δόλια ή γνήσια.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

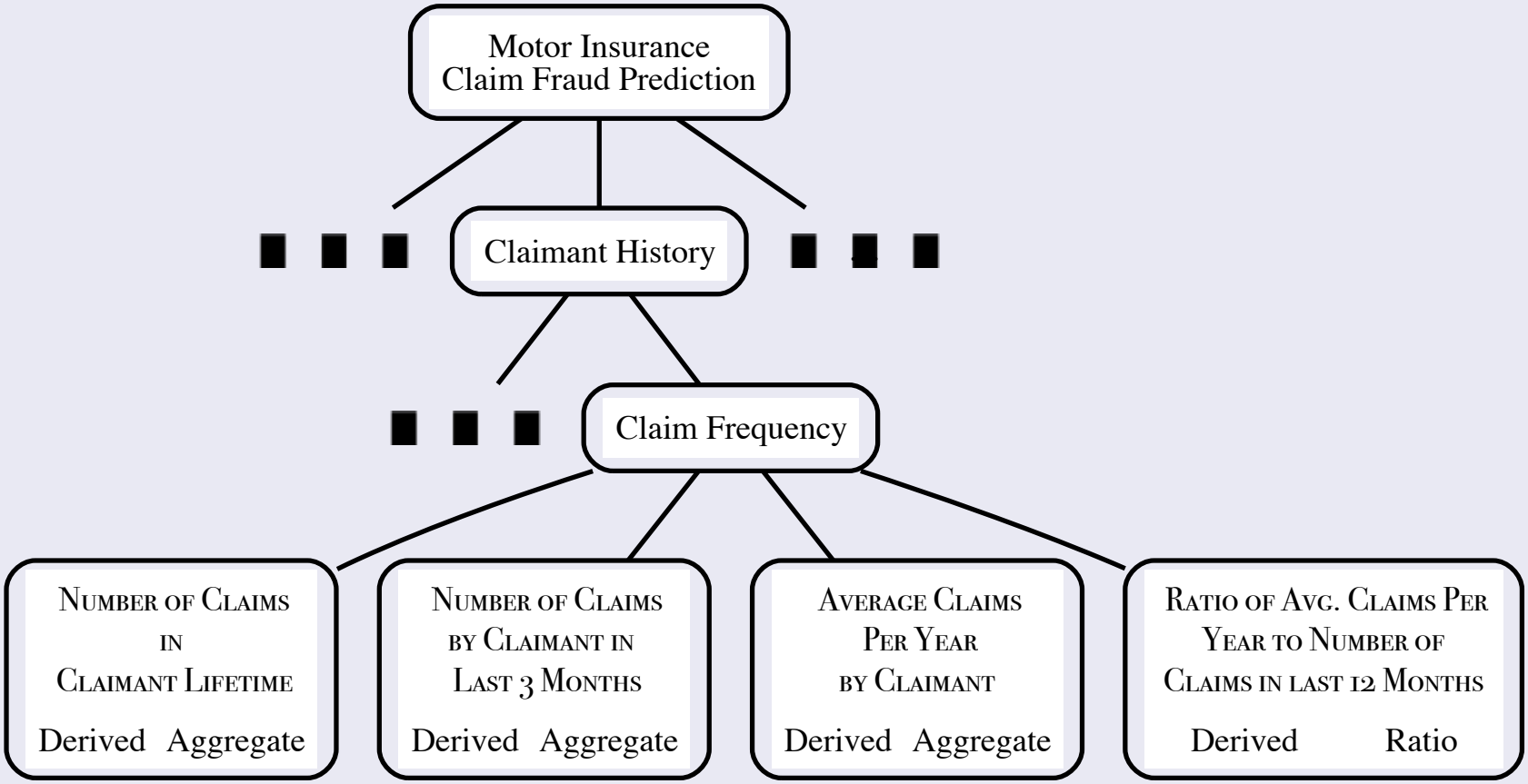
Ποια χαρακτηριστικά θα μπορούσατε να χρησιμοποιήσετε για να αποτυπώσετε την έννοια πεδίου *Συχνότητα Απαιτήσεων (Claim Frequency)*;



Σχήμα: Ενδεικτικές έννοιες πεδίου για λύση πρόβλεψης απάτης στην ασφάλιση αυτοκινήτου.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

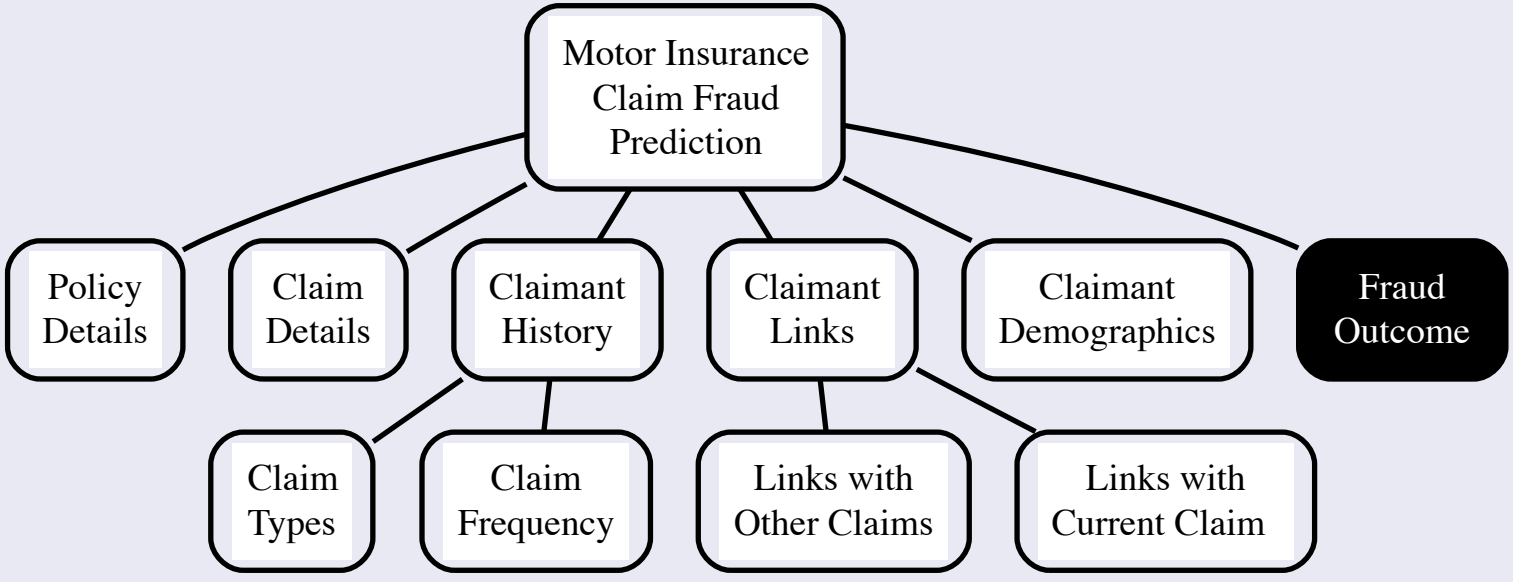
Ποια χαρακτηριστικά θα μπορούσατε να χρησιμοποιήσετε για να αποτυπώσετε την έννοια πεδίου *Συχνότητα Απαιτήσεων (Claim Frequency)*;



Σχήμα: Υποσύνολο των εννοιών πεδίου και των σχετικών χαρακτηριστικών για λύση πρόβλεψης απάτης στην ασφάλιση αυτοκινήτου

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

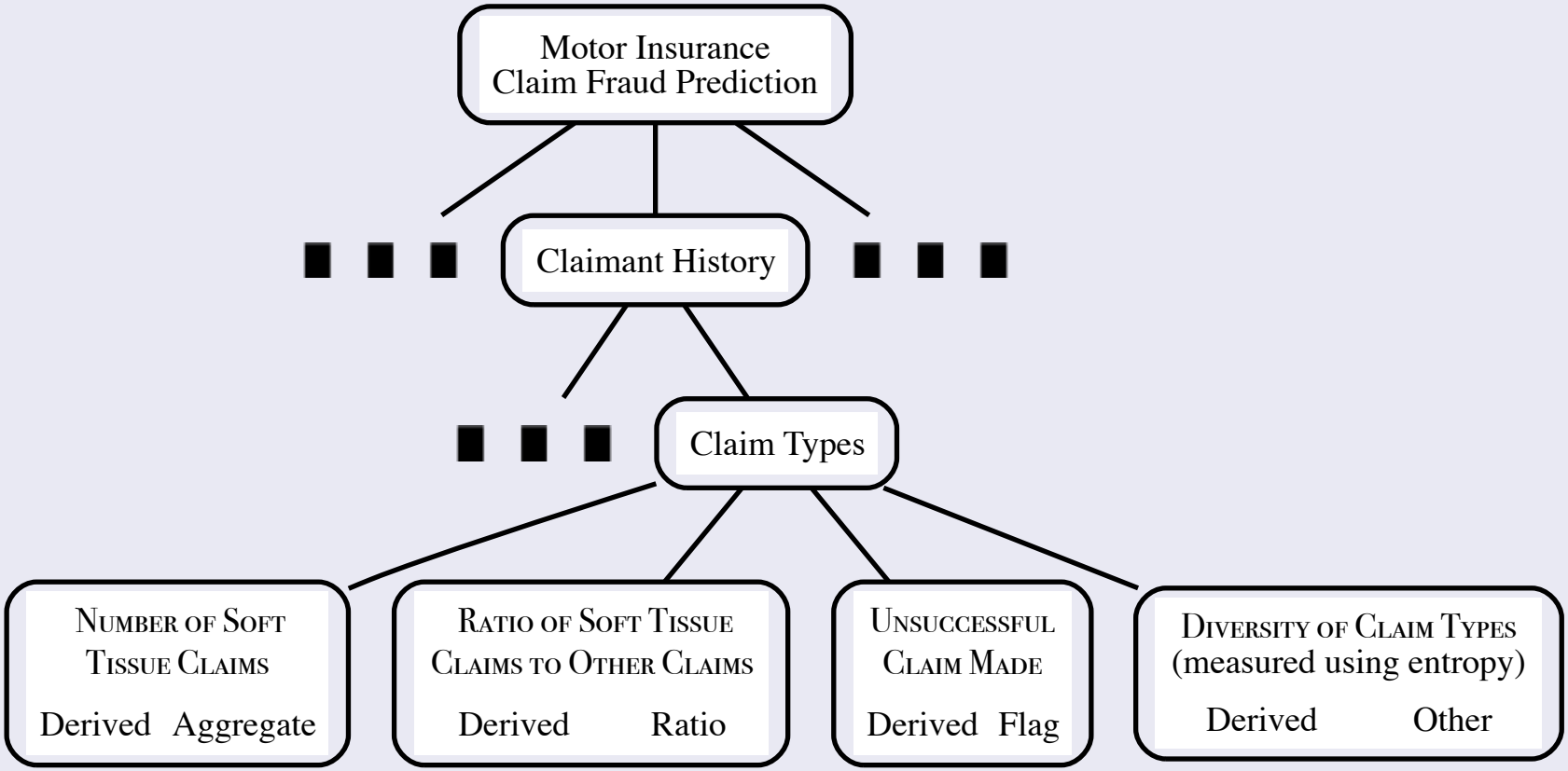
Ποια χαρακτηριστικά θα μπορούσατε να χρησιμοποιήσετε για να αποτυπώσετε την έννοια πεδίου *Τύποι Απαιτήσεων (Claim Types)*;



Σχήμα: Ενδεικτικές έννοιες πεδίου για λύση πρόβλεψης απάτης στην ασφάλιση αυτοκινήτου.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

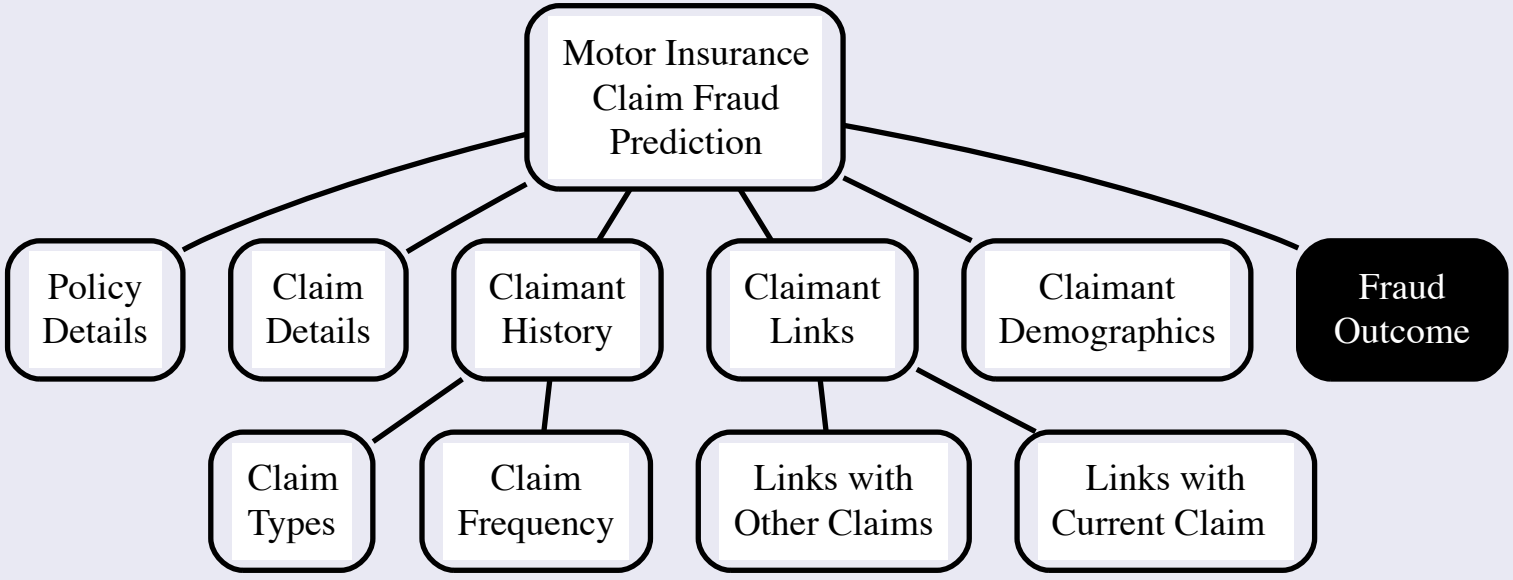
Ποια χαρακτηριστικά θα μπορούσατε να χρησιμοποιήσετε για να αποτυπώσετε την έννοια πεδίου *Τύποι Απαιτήσεων (Claim Types)*;



Σχήμα: Υποσύνολο των εννοιών πεδίου και των σχετικών χαρακτηριστικών για λύση πρόβλεψης απάτης στην ασφάλιση αυτοκινήτου.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

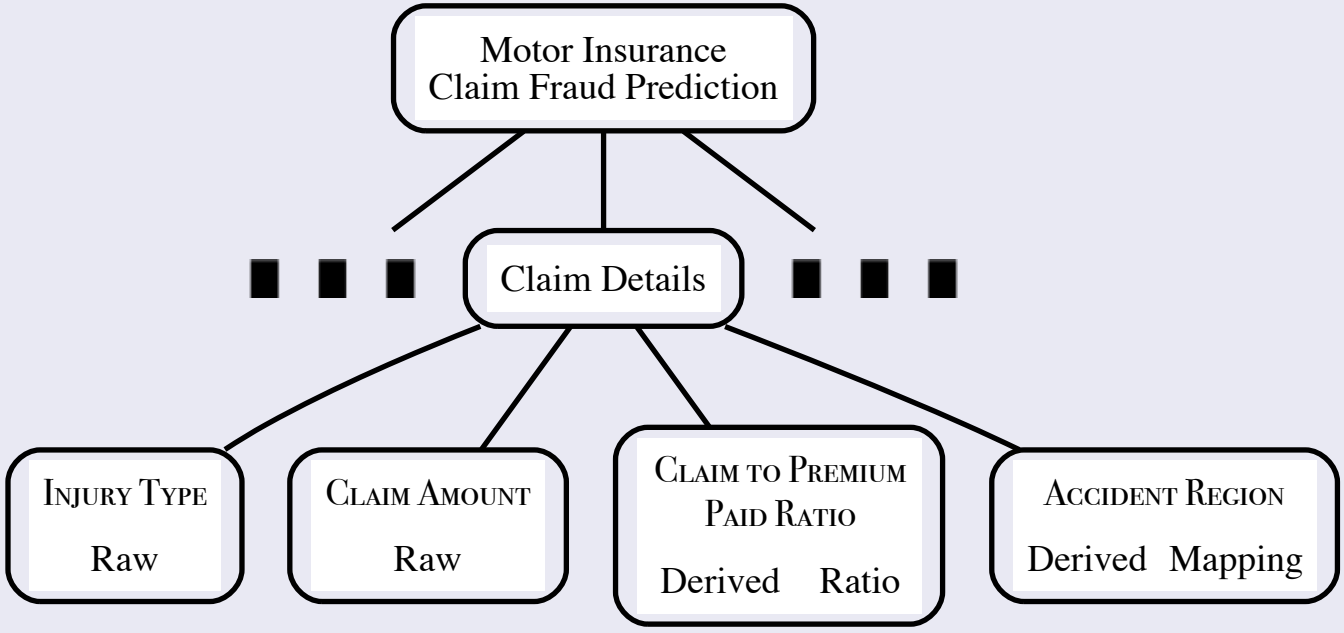
Ποια χαρακτηριστικά θα μπορούσατε να χρησιμοποιήσετε για να αποτυπώσετε την έννοια πεδίου *Λεπτομέρειες Απαίτησης (Claim Details)*;



Σχήμα: Ενδεικτικές έννοιες πεδίου για λύση πρόβλεψης απάτης στην ασφάλιση αυτοκινήτου.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

Ποια χαρακτηριστικά θα μπορούσατε να χρησιμοποιήσετε για να αποτυπώσετε την έννοια πεδίου *Λεπτομέρειες Απαίτησης (Claim Details)*;



Σχήμα: Υποσύνολο των εννοιών πεδίου και των σχετικών χαρακτηριστικών για λύση πρόβλεψης απάτης στην ασφάλιση αυτοκινήτου.

Μελέτη περίπτωσης: Απάτη στην ασφάλιση αυτοκινήτου

- Ο ακόλουθος πίνακας απεικονίζει τη δομή του τελικού ΑΒΤ που σχεδιάστηκε για τη λύση ανίχνευσης απάτης σε απαιτήσεις αποζημίωσης.
- Ο πίνακας περιλαμβάνει περισσότερα περιγραφικά χαρακτηριστικά από αυτά που συζητήσαμε.
- Εμφανίζει επίσης τις πρώτες τέσσερις περιπτώσεις.
- Αν τον εξετάσουμε προσεκτικά, βλέπουμε αρκετές περίεργες τιμές (π.χ. $-9\,999$) και αρκετές ελλείπουσες τιμές—θα επανέλθουμε σε αυτά στο Κεφάλαιο 3.

Πίνακας: Το ABT για τη λύση ανίχνευσης απάτης σε απαιτήσεις αποζημίωσης ασφάλισης αυτοκινήτου.

ID	Type	Inc.	Marital Status	Num. Clmnts.	Injury Type	Hospital Stay	Claim Amt.
1	CI	0		2	Soft Tissue	No	1 625
2	CI	0		2	Back	Yes	15 028
3	CI	54 613	Married	1	Broken Limb	No	-9 999
4	CI	0		3	Serious	Yes	270 200
		⋮				⋮	

ID	Total Claimed	Num. Claims	Num. Claims 3 Months	Avg. Claims Per Year	Avg. Claims Ratio	Num. Soft Tissue	% Soft Tissue
1	3 250	2	0	1	1	2	1
2	60 112	1	0	1	1	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
		⋮			⋮		

ID	Unsucc. Claims	Claim Amt. Rec.	Claim Div.	Claim to Prem.	Region	Fraud Flag
1	2	0	0	32.5	MN	1
2	0	15 028	0	57.14	DL	0
3	0	572	0	-89.27	WAT	0
4	0	270 200	0	30.186	DL	0
		⋮			⋮	

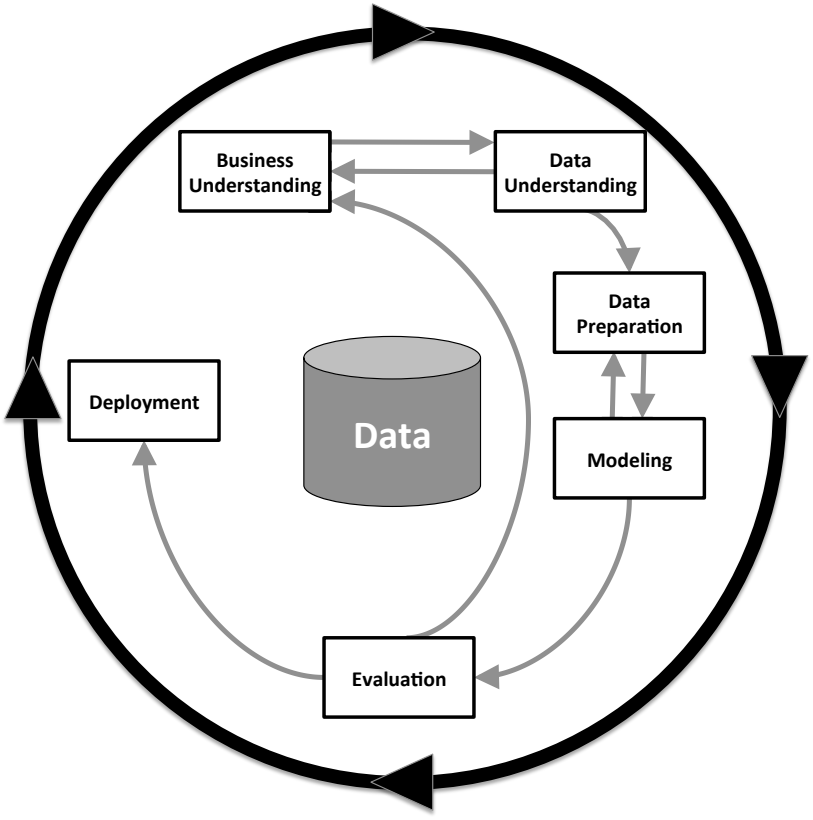
Σύνοψη

- Τα μοντέλα προγνωστικής αναλυτικής που κατασκευάζονται με τεχνικές μηχανικής μάθησης είναι εργαλεία που μπορούμε να χρησιμοποιήσουμε για να λαμβάνουμε καλύτερες αποφάσεις μέσα σε έναν οργανισμό — δεν είναι αυτοσκοπός.
- Είναι σημαντικό να κατανοούμε πλήρως το επιχειρηματικό πρόβλημα που καλείται να αντιμετωπίσει ένα μοντέλο — αυτός είναι ο σκοπός πίσω από τη *μετατροπή επιχειρηματικών προβλημάτων σε αναλυτικές λύσεις*.

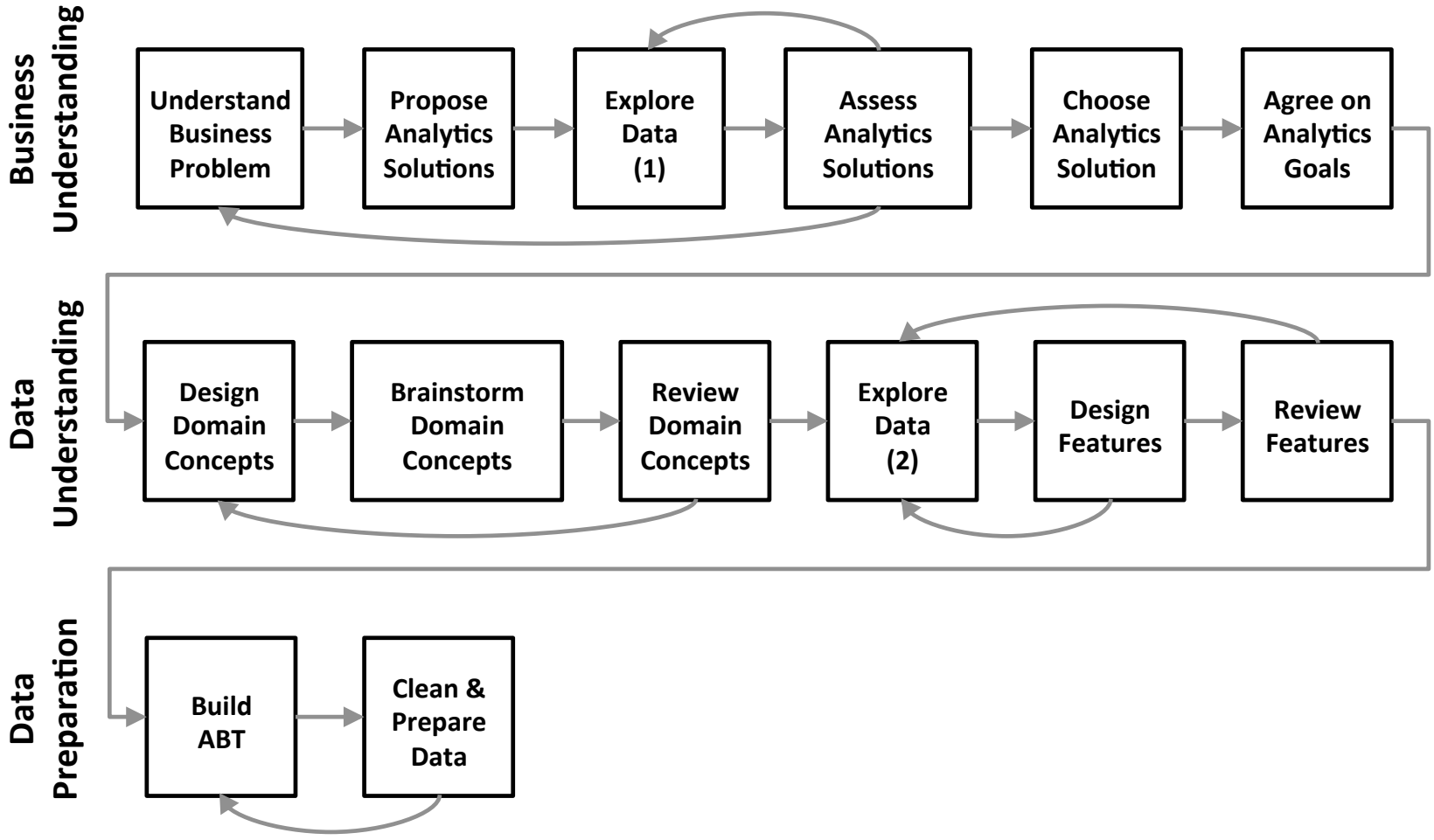
- Τα μοντέλα προγνωστικής αναλυτικής εξαρτώνται από τα δεδομένα που χρησιμοποιούνται για την κατασκευή τους — τον **analytics base table** (**ABT**).
- Το πρώτο βήμα στον σχεδιασμό ενός ABT είναι να αποφασίσουμε το **αντικείμενο πρόβλεψης**.
- Ένας αποτελεσματικός τρόπος σχεδιασμού ABT είναι να ξεκινάμε ορίζοντας, σε συνεργασία με την επιχείρηση, ένα σύνολο **εννοιών πεδίου** και στη συνέχεια να σχεδιάζουμε **χαρακτηριστικά** που εκφράζουν αυτές τις έννοιες ώστε να σχηματίσουμε τον ABT.

- Τα χαρακτηριστικά (τόσο τα περιγραφικά όσο και το χαρακτηριστικό-στόχος) είναι συγκεκριμένες αριθμητικές ή συμβολικές αναπαραστάσεις των εννοιών πεδίου.
- Είναι χρήσιμο να διακρίνουμε μεταξύ **πρωτογενών χαρακτηριστικών** που προκύπτουν άμεσα από υπάρχουσες πηγές δεδομένων και **παράγωγων χαρακτηριστικών** που κατασκευάζονται με χειρισμούς τιμών από υπάρχουσες πηγές.
- Συνηθισμένοι χειρισμοί: συναθροίσεις, δείκτες/σημαίες, λόγοι και αντιστοιχίσεις — αν και κάθε χειρισμός είναι έγκυρος.

- Οι τεχνικές που περιγράφηκαν εδώ καλύπτουν τις φάσεις **Business Understanding**, **Data Understanding** και (εν μέρει) **Data Preparation** της διαδικασίας **CRISP-DM**.



Σχήμα: Ένα διάγραμμα της διαδικασίας CRISP-DM.



Σχήμα: Σύνοψη των εργασιών στις φάσεις Business Understanding, Data Understanding και Data Preparation της διαδικασίας **CRISP-DM**.