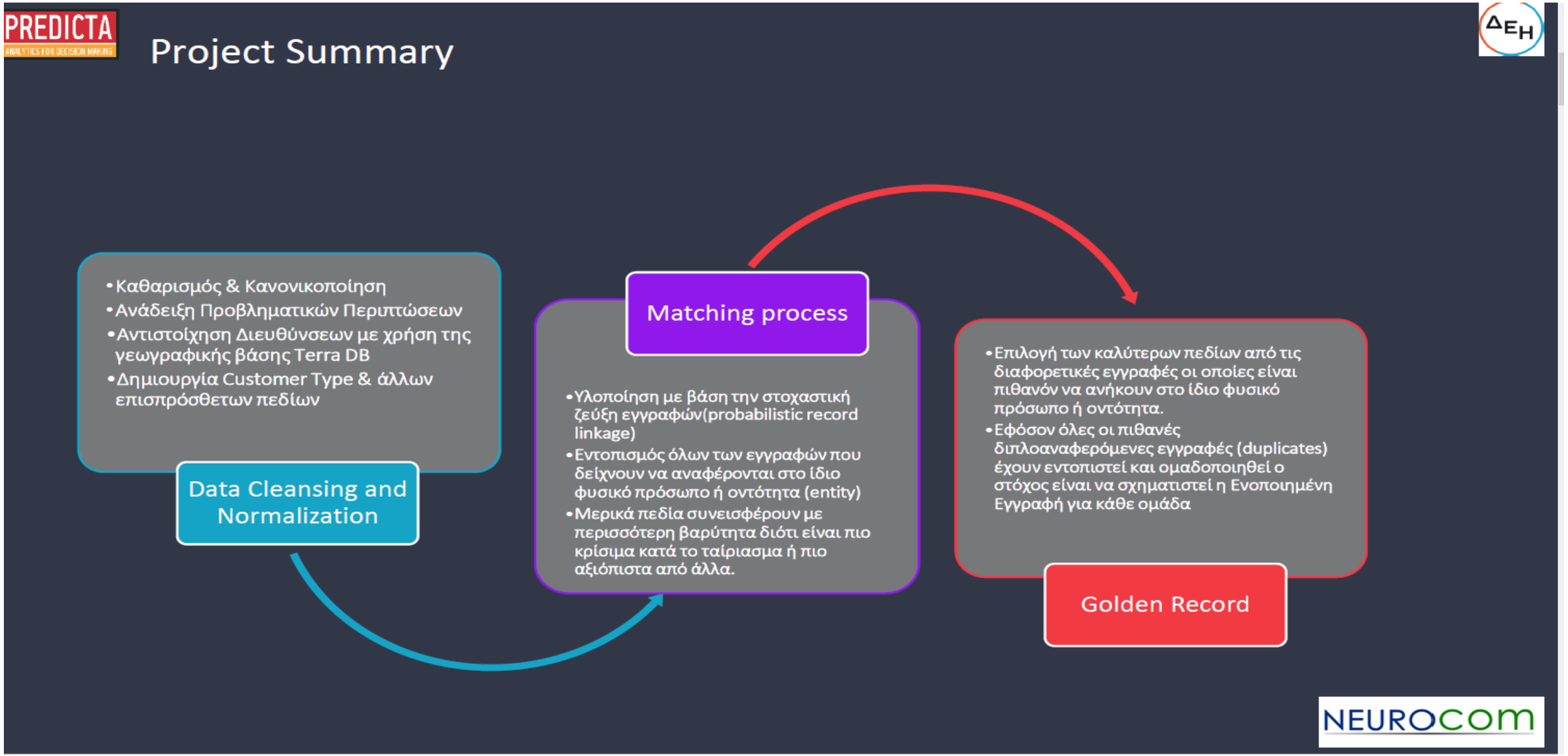
Data Cleansing & Golden Record

May 2023 – Projects & Customer Experience BU

Data Cleansing – *Summary*



Scope Deliverables

PPC Data Cleansing

Κύριος Στόχος: Παρουσίαση αποτελεσμάτων για παραδοτέα του έργου που αφορά τα παρακάτω θέματα:

ΠΑΡΑΔΟΤΕΟ 1

a) Τεχνολογική πλατφόρμα εγκατεστημένη

ΠΑΡΑΔΟΤΕΟ 2

a) Καθαρισμός και κανονικοποίηση πεδίου τύπου ονοματεπωνύμου
b) Καθαρισμός και κανονικοποίηση πεδίου τύπου διεύθυνσης
c) Ανάδειξη (flagging) προβληματικών records στοιχείων επικοινωνίας πελάτη (τηλ και emails) και ΑΦΜ
d) Εύρεση διπλών, τριπλών,…εγγραφών, ταίριασμα/matching, και υλοποίηση ενιαίων και καλά καθορισμένων οντοτήτων (δημιουργία "Golden Record")

ΠΑΡΑΔΟΤΕΟ 3

a) Τελική καθαρισμένη βάση, με επιτυχία 95%

ΠΑΡΑΔΟΤΕΟ 4

a) Εκπαίδευση της εσωτερικής εμπλεκόμενης ομάδας της ΔΕΗ πάνω στις παραχθείσες κατά τη διάρκεια του έργου διαδικασίες (transfer of knowledge on processes, streams, flows, …), καθώς και παράδοση πλήρους σχετικού documentation

Scope Deliverables

PPC Data Cleansing

Επιπλέον διαδικασίες που έχουν γίνει στη βάση

1. Καθαρισμός και κανονικοποίηση όνομα πατέρα
2. Δημιουργία πεδίου Customer Type
3. Εκτός από την ανάδειξη (flagging) προβληματικών records στοιχείων επικοινωνίας (τηλ και emails) δημιουργήθηκαν νέα πεδία για σταθερά και κινητά τηλέφωνα ανά πηγή
4. Εκτός από την ανάδειξη (flagging) προβληματικών records για ΑΦΜ δημιουργήθηκαν νέα πεδία για χαρακτηρισμό του ΑΦΜ (έγκυρου, suspicious VAT_Format_ADDED_ZERO, Invalid).
5. Ανάδειξη (flagging) προβληματικών εγγραφών για Αριθμούς Αστυνομικής Ταυτότητας και Διαβατηρίου
6. Δημιουργία νέων πεδίων με έγκυρο ελληνικό format για Αριθμούς Αστυνομικής Ταυτότητας και Διαβατηρίου

Fields

Cleaning and Normalization process

Καθαρισμός και κανονικοποίηση ανά πηγή (SAP, E-value, E-bill & Faidra)

Σύνολο μοναδικών Contract Accounts = 15.073.720

Σύνολο εγγραφών = 17.851.869

Cleaning & Normalization (Names)

- ✓First_Name
- ✓Last_Name
- ✓Father_Name

Correction and Flag

- ✓ VAT_ID
- ✓ DOY
- ✓ AT
- ✓ Passport
- ✓ Cellphone_Number
- ✓ Fixed_Phone_Number
- ✓ Email

Cleaning Normalization & Address Mapping using Terra DB (Addresses)

- ✓Municipality_Name
- ✓Street
- ✓House_Number
- ✓Postal_Code
- ✓Region
- ✓Billing_Municipality_Name
- ✓Billing_Street
- ✓Billing_House_Number
- ✓Billing_Postal_Code
- ✓Billing_Address_Region
- ✓PoD_Municipality_Name
- ✓PoD_Street
- ✓PoD_House_Number
- ✓PoD_Postal_Code
- ✓PoD_REGION
- ✓B2B_Municipality_Name
- ✓B2B_Street
- ✓B2B_House_Number
- ✓B2B_Postal_Code
- ✓B2B_Address_Region

Cleaning and Normalization and Flagging process

SUMMARY Aggregated KPIs

- 19,8 περισσότερες φορές έχει αναγνωριστεί το Non-Person (από 106,1K σε 2,1M)
- 2,5 περισσότερες φορές έχει γίνει αναγνώρισή και συμπλήρωση του First Name (από 5,04M σε 12,88M)
- Το ποσοστό αλλαγών που έγινε στο Last Name είναι 60,2%
- 2,01 περισσότερες φορές έχει γίνει αναγνώρισή κα συμπλήρωση του Father Name (από 4,8M σε 9,6M)
- Το σύνολο των διευθύνσεων που επεξεργάστηκαν είναι 44,8M. Το ποσοστό αναγνώρισης εύρεσης στη TERRA DB, (ύπαρξη συντεταγμένων) είναι >96%.
- 15,07M CAs που υπήρχαν στον Input πίνακα αναγνωρίστηκαν με έγκυρα VAT το 80,2%
- 5,82M Active CAs που υπήρχαν στον Input πίνακα αναγνωρίστηκαν με έγκυρα VAT το 79.2%
- Για τα 19,9M τηλέφωνα που επεξεργάστηκαν (ανεξαρτήτου πηγής) βρέθηκαν έγκυρα τα 18,9M (95.1%), όπου τα 7,3M (36.5%) είναι σταθερά τηλέφωνα και τα 11,7M (58,6%) είναι κινητά τηλέφωνα,
- 73% αναγνωρίστηκαν να έχουν ΑΔΤ ελληνικό Format και 0,7% αναγνωρίστηκαν να έχουν Passport Format
- Ο συνολικός αριθμός των Emails που επεξεργάστηκαν και έχουν έγκυρο format είναι 3,17M (ανεξαρτήτου πηγής).

Sample Test

Cleaning and Normalization and Flagging process

Sample Test Names –Addresses

Σε τυχαίο δείγμα που δόθηκε από την ΔΕΗ τα ποσοστά καθαρισμού σωστού ονόματος, διεύθυνσης και Customer Type είναι μεγαλύτερα του 95%

- Το ποσοστό αναγνώρισης σωστού Company Type είναι 99,31%
- Το ποσοστό σωστού ονόματος και επιθέτου είναι 99,6%
- Το ποσοστό σωστού ονόματος επιθέτου και όνομα πατέρα είναι 98,2%
- Ποσοστό αναγνώρισης διεύθυνσης 95,59%

SAP Addresses	PREDICTA/ NEUROCOM
Cases Not Recognized	23
Null Not Recognized	74
Total Cases	2200
% Not Recognized (including Null)	4,41%

Test Cases

Golden Record Groups Results

- Σε 52 test cases που δόθηκαν από την ΔΕΗ που είχαν αναγνωριστεί σαν λάθος στο προηγούμενο run τα 46 έχουν διορθωθεί, 4 έχουν χωριστεί σε περισσότερα από 1 γκρουπ και σε δύο έχει γίνει λάθος.

- Σε επίπεδο Golden Record έχουν δημιουργηθεί 90 Golden Record Groups με τα παρακάτω αποτελέσματα:

	Total Number of GR	**CORRECT GR**	**Partial Correct (more than one GR)**	**False Positive GR**
Counts	90	76	11	3
%	100%	84,44%	12,22%	3,33%

- Σε επίπεδο Contracts Account υπήρχαν 3.239 Contract Accounts που ομαδοποιήθηκαν σε 90 Golden Record Groups με τα παρακάτω αποτελέσματα:

	TOTAL CAs	Correct	Partial Correct	False Positive GR
Counts	3.239	3.213	19	7
%	100%	99,20%	0,59%	0,22%

- Αν από αυτά αφαιρέσουμε τα Golden Groups που έχουν μεγάλο όγκο Contracts Account (Vodafone&ΥΠΕΘΑ) τότε τα αποτελέσματα σε επίπεδο Contract Account είναι τα παρακάτω:

	TOTAL CAs	**Correct**	**Partial Correct**	**False Positive GR**
Counts	303	290	6	7
%	100%	95,7%	2,0%	2,3%

Golden Record – *Matching Process*

PREDICTA
ANALYTICS FOR DECISION MAKING

Matching Process

Main Steps

Η διαδικασία συσχέτισης έχει υλοποιηθεί βασισμένη επάνω στην στοχαστική ζεύξη των υπο διερεύνηση εγγραφών (probabilistic record linkage)

	Βασικά Βήματα (Passes)	Σχόλια
1	B2B NAME-VAT	Ίδιο B2B NAME μέσα στην ίδια ομάδα ενοποιημένων εγγραφών (Golden Record Group)
2	'VAT-CUSTOMER TYPE – SIMILAR NAME'	Βήμα επεξεργασίας με το μεγαλύτερο επίπεδο σπουδαιότητας
3	'ΑΔΤ-CUSTOMER TYPE – SIMILAR NAME'	Αύξηση της βαθμολογίας όταν υπάρχουν ίδιες διευθύνσεις ή άλλες λεπτομέρειες επικοινωνίας. Συνδυασμός των ομάδων με το προηγούμενο βήμα. Βήματα επεξεργασίας με το μεγαλύτερο επίπεδο σπουδαιότητας
4	Customer_Type & Company_Name	Ως επί το πλείστον για Εταιρείες χωρίς ΑΦΜ (Null VAT). Επιπρόσθετη βαθμολογία όταν υπάρχουν ίδιες διευθύνσεις ή άλλες λεπτομέρειες επικοινωνίας. Σύγκριση αποτελεσμάτων με αυτά από τα προηγούμενα βήματα.
	ΚΟΙΝΟΧΡΗΤΑ-SAME BILLING ADDRESS	ΚΟΙΝΟΧΡΗΣΤΑ ΜΕ ΙΔΙΟ BILLING ADDRESS → Ίδια ομάδα ενοποιημένων εγγραφών (SAME Golden Record Group)
5	'Fix Phones or Cell Phones '	Ίδια ονόματα και ίδιος τύπος πελάτη, επιπρόσθετη βαθμολογία όταν υπάρχουν ίδιες διευθύνσεις ή άλλες λεπτομέρειες επικοινωνίας. Σύγκριση αποτελεσμάτων με αυτά από τα προηγούμενα βήματα.
6	Addresses	Ίδια ονόματα και ίδιος τύπος πελάτη, επιπρόσθετη βαθμολογία όταν υπάρχουν ίδιες διευθύνσεις ή άλλες λεπτομέρειες επικοινωνίας. Σύγκριση αποτελεσμάτων με αυτά από τα προηγούμενα βήματα.
7	Ίδια Contract_ID από διαφορετικές πηγές	Ίδια ονόματα και άλλες λεπτομέρειες επικοινωνίας.

Μερικά πεδία συνεισφέρουν με περισσότερη βαρύτητα διότι είναι πιο κρίσιμα κατά το ταίριασμα ή πιο αξιόπιστα από άλλα.
Οι στατιστικές ιδιότητες των πεδίων της πληροφορίας μαζί με ρυθμιστικές παραμέτρους καθορίζουν την συνεισφορά σε κάθε συσχετιζόμενη ομάδα εγγραφών.

ΔΕΗ

NEUROCOM

Golden Record – *Summary KPIs*

PREDICTA
ANALYTICS FOR DECISION MAKING

Golden Record Process

SUMMARY Aggregated KPIs – Type of CA

ΔΕΗ

- Ο χαρακτηρισμός του κάθε Contract Account σε σχέση με το Golden Record Group που ανήκει βασίζεται στο Score Weight που δημιουργείται ανά Βήμα (Pass) και ανήκει σε ένα από τους παρακάτω τύπους:
  - Master Record, 24.2%
  - Duplicate Record, 37,9%
  - Clerical, 2%
  - Residual, 35,9%

	Master	Duplicate	Clerical	Residual	Grand Total
Grand Total	3.644.276	5.712.782	299.141	5.417.521	15.073.720
%	24,2%	37,9%	2,0%	35,9%	100%

Η διαδικασία αυτή δημιουργεί αριθμητικές τιμές (weights) για κάθε σύγκριση συγκεκριμένων πεδίων. Οι τιμές αυτές μετρούν την συνεισφορά καθενός από αυτά στο συνολικό βάρος κάθε ομαδοποιημένου συνόλου εγγραφών (matching group).

NEUROCOM

Golden Record – *Summary KPIs*

PREDICTA
ANALYTICS FOR DECISION MAKING

Golden Record Groups

SUMMARY Aggregated KPIs – Additional Info

ΔΕΗ

Contracts Accounts που *δεν είχαν VAT* αλλά με την ένταξή τους σε Golden Record Group συμπληρώθηκαν από το VAT του Golden Record Group	558,2K
Contracts Accounts που *δεν είχαν Fix Phone (SAP)* αλλά με την ένταξή τους σε Golden Record Group συμπληρώθηκαν από το FixPhone του Golden Record Group	2,04M
Contracts Accounts που *δεν είχαν Cell Phone (SAP)* αλλά με την ένταξή τους σε Golden Record Group συμπληρώθηκαν από το Cell Phone του Golden Record Group	1,77M
Contracts Accounts που *δεν είχαν Email* αλλά με την ένταξή τους σε Golden Record Group συμπληρώθηκαν από το Email του Golden Record Group	2,23M

- Δεν υπάρχει Golden Record Group που να έχει διαφορετικά VAT στο ίδιο Group (Golden Rule). Εκτός από τις περιπτώσεις που έχει γίνει τυπογραφικό λάθος.

- Τα κάθε Golden Record Group έχει το ίδιο Customer Type. Δεν υπάρχουν Golden Record Groups που να έχουν διαφορετικό Customer Type στο ίδιο Group (Golden Rule)

NEUROCOM

Thank you