Start-up to production. Case of a labor force survey in Mexico

José Alejandro Ruíz Sánchez / Jael Pérez Sánchez

Objectives

- Initial Objectives

  - Reduce the workload of human coders
  - Reduce coding time
  - Maintain or improve encoding quality

What have we done?

- The target variables are Occupation and Economic Activity
- In 2019, We started exploring some algorithms using data from the National Survey of Income and Expenditure (ENIGH 2018)
  - First with 37 thousand register
  - Then with all data base
- In 2020, we use the work developed for ENIGH in the Population and Housing Census, this as an alternative verification mechanism (In total 11.4 million records)

Stages of the NLP Text Classification Process

What kind of algorithms we use?

Vectorization

- TF-IDF ⭐
- W2V
- Fasttext

Classification

- SVM ⭐
- Logistic regression
- Random Forest
- Neural Networks
- XGBoost
- K-NN
- Naive Bayes

First results obtained

These results were obtained using SVM as the classification algorithm

SCIAN

	tf idf		*fasttext*	
	Sin preprocesamiento	Con preprocesamiento	Sin preprocesamiento	Con preprocesamiento
accuacy	86.8%	**87.8%**	82.6%	**83.5%**
f1	63.1%	64.5%	57.5%	58.9%
precision	62.2%	63.4%	54.8%	55.8%
recall	64.9%	67.1%	63.3%	64.7%

SINCO

	tf idf		*fasttext*	
	Sin preprocesamiento	Con preprocesamiento	Sin preprocesamiento	Con preprocesamiento
accuacy	81.6%	**82.0%**	71.4%	**72.6%**
f1	54.4%	55.7%	45.4%	46.5%
precision	52.5%	53.8%	42.3%	42.7%
recall	58.5%	59.9%	53.9%	56.0%

- It is better to use TF-IDF as the vectorization algorithm than others
- Using data preprocessing improves the results slightly

Assembler method

We call "assambler method" the mechanism of choosing the most frequent code when making the classification by the different methods.

Economic Activity				
	Accuracy	**Precision**	**Recall**	**F1**
Assembly with same weights	0.8905	0.6925	0.6149	0.6365
Assembly with differentiated weights	0.8921	0.6767	0.6420	0.6512

Occupation				
	Accuracy	**Precision**	**Recall**	**F1**
Assembly with same weights	0.8447	0.6441	0.5384	0.5639
Assembly with differentiated weights	0.8505	0.6437	0.5637	0.5831

We tested various classification methods: SVM, Random Forest, Neural Networks, XG-Boost, Logistic regression, K-NN, Naive Bayes.

What's the next?

Objetive and about the ENOE

- Second stage objective

  - Adapt Machine Learning algorithms to the coding production process of the National Occupation and Employment Survey (ENOE).

- About ENOE

  - It is a semi-panel type survey
  - It is a survey with national coverage and desegregation by state
  - It is quarterly

Defined stages

1. Define the coding production process for (ENOE) considering Machine Learning (ML) algorithms.

2. Adapt the ML algorithms developed for other projects to the ENOE.

3A. Optimize ML algorithms.

3B. Create a "Ground Truth" database.

5. Adapt the ML algorithms to the traditional ENOE coding system.

6. Conduct a test with the use of ML and code the 3rd quarter of ENOE 2022 in parallel with the traditional process throughout the entire quarter.

7. Carry out an evaluation report on the impact of applying ML.

8. Free up productive.

How to create the ground truth database?

Strategies

1. Encode what is already encoded - First quarter of 2020

Two teams encode the same database independently

Figures involved in the process

Expert encoders (central)

Checkers

Gurú

Main tasks or functions	Job title	Activity to the company	Central Team 1	Central Team 2	"Gurú"
TOMAR HUELLAS DACTILARES A PRESUNTOS DELINCUENTES	CRIMINALISTA	PERSECUCION DEL DELITO	2521	2132	2132
HACER PROGRAMACION DE PROMOCIONES Y OFERTAS	CREADORA DE MARKETING	VENTA DE ABARROTES SERVICIO ELECTRONICOS BEBIDAS ALCOHOLICAS	2112	2511	2511
VERIFICAR QUE LLEVEN MATERIAL CORRECTO Y COMPLETO	EMPLEADO DE LOGISTICA DE EMBARQUE	FABRICACION DE VARILLA	8101	8301	3101

Strategies

2. Code a New Quarter - Third Quarter 2021

Three teams encode the same database independently

Figures involved in the process

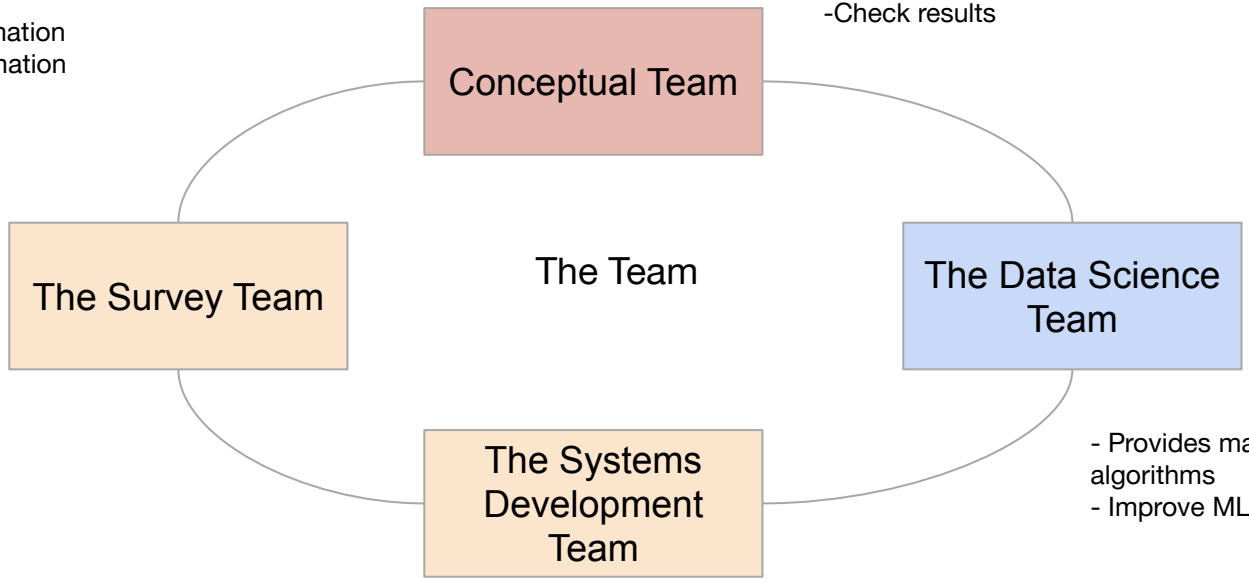
- Expert encoders (Central)
- Expert encoders (States)
- Checkers
- Gurú

Main tasks or functions	Job title	Activity to the company	Central Team 1	Central Team 2	Team in states	"Gurú"
AYUDA A ARRIMAR LAS HERRAMIENTAS PARA SOLDAR LAS VIGAS DE LANAVE INDUSTRIAL	AYUDANTE DE ACOMODADOR DE NAVES INDUSTRIALES	MONTAR E INSTALAR NAVES INDUSTRIALES	9221	9231	9231	9231
MANEJA MAQUINA CORTADORA DE METAL	CORTADOR DE COMALES	FABRICACION DE COMALES ASADORES VAPORERAS DE ACERO	8123	7221	8123	7211
COSE GUANTES DE TELA	COSTURERA	ELABORACIÓN DE GUANTES DE TELA TÉRMICA	7341	7342	7341	7342

The team

-Provide classifiers
-Coding strategies
-Experts Encoders and Gurus
-Encode
-Check results

- Provide the information
- Encode the information
- Check results



- Provides machine learning algorithms
- Improve ML algorithms

- Develop the coding system
- Adapt ML algorithm to the system

Conociendo México

01 800 111 46 34

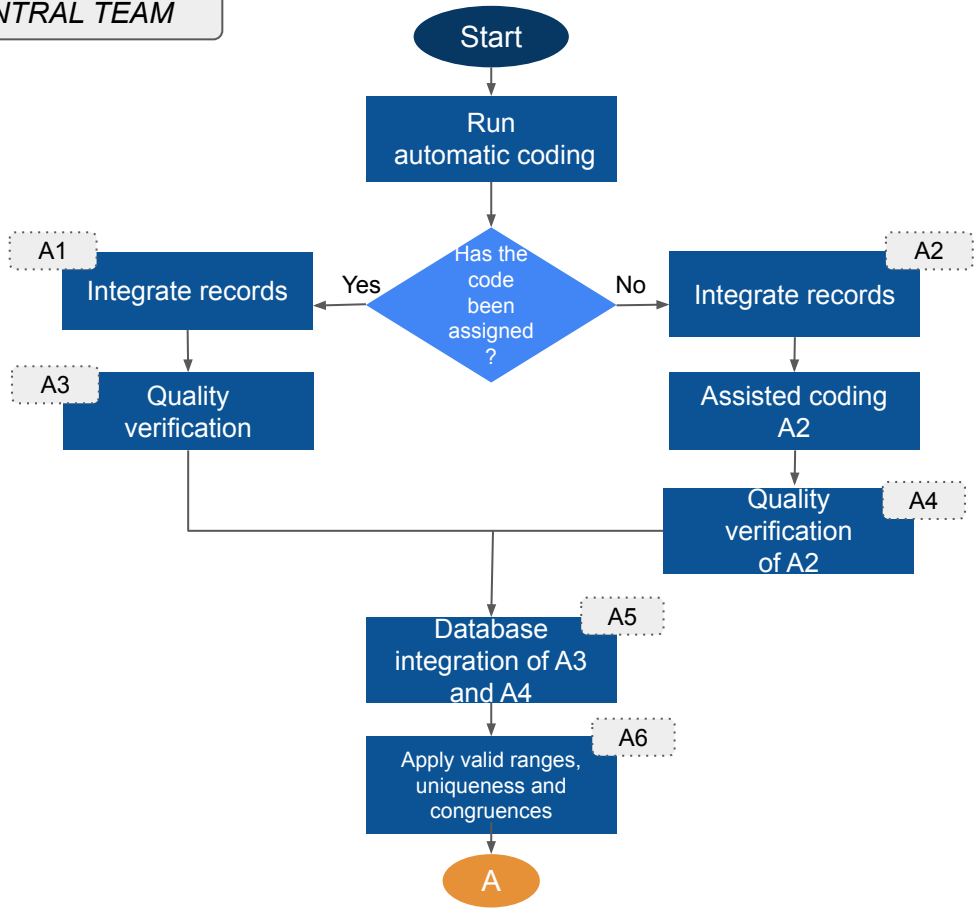
www.inegi.org.mx

atencion.usuarios@inegi.org.mx

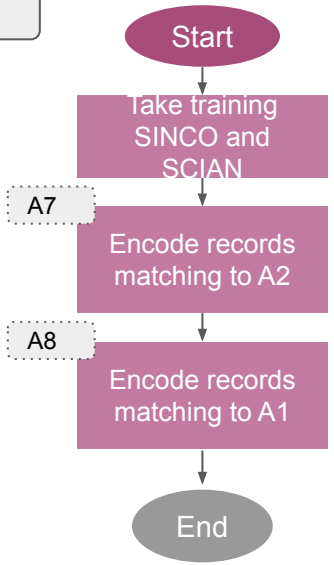
INEGI Informa

Workflow for creating the Ground Truth database

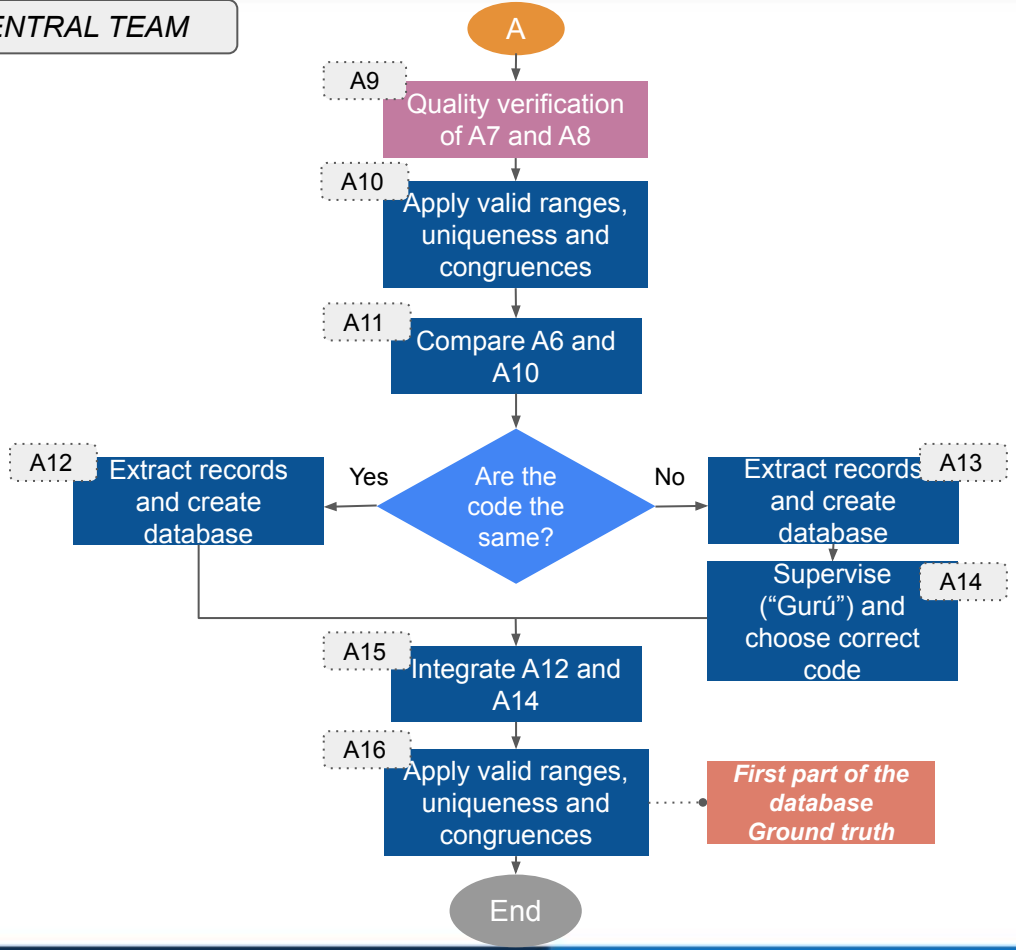
CENTRAL TEAM

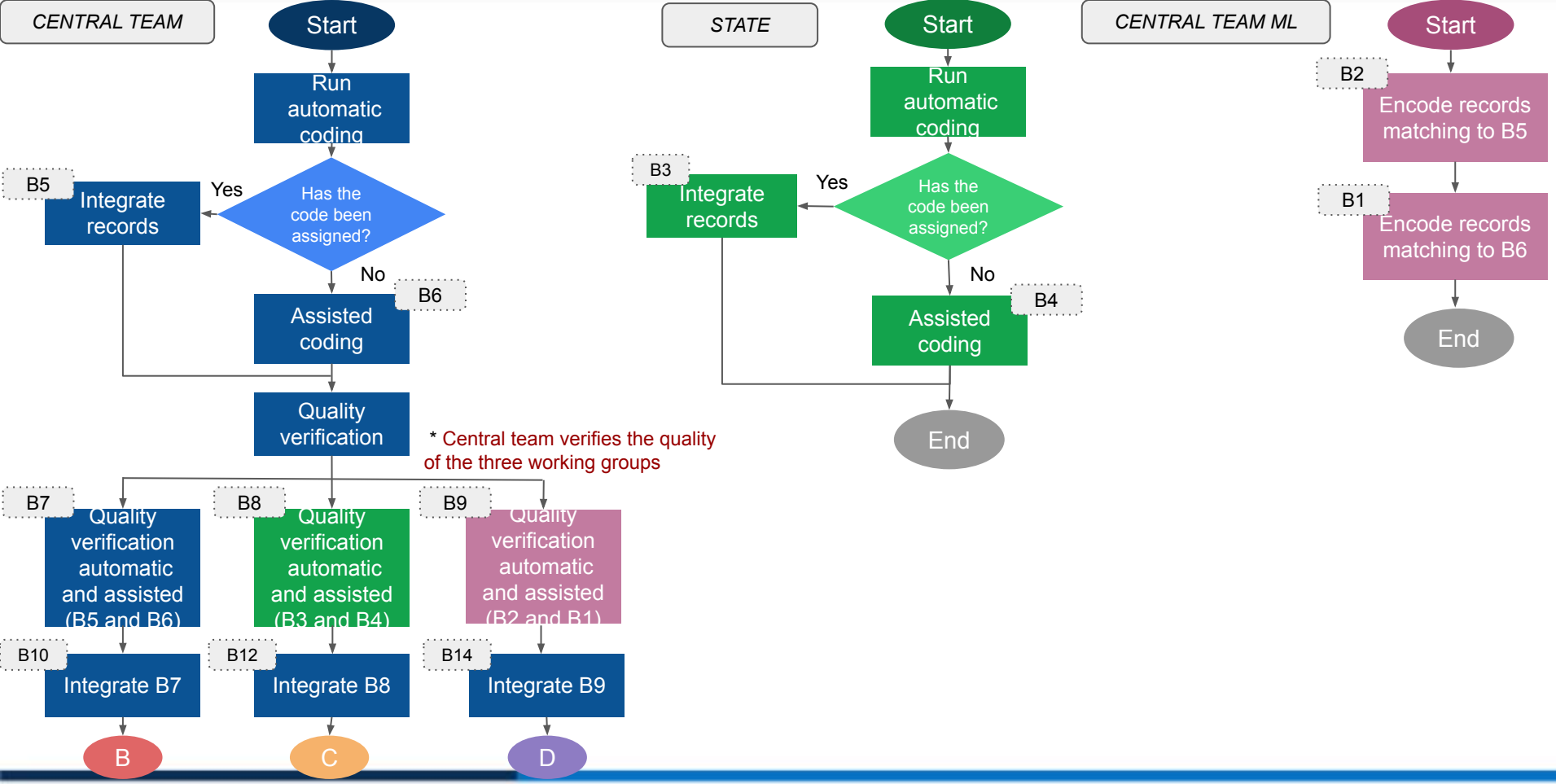


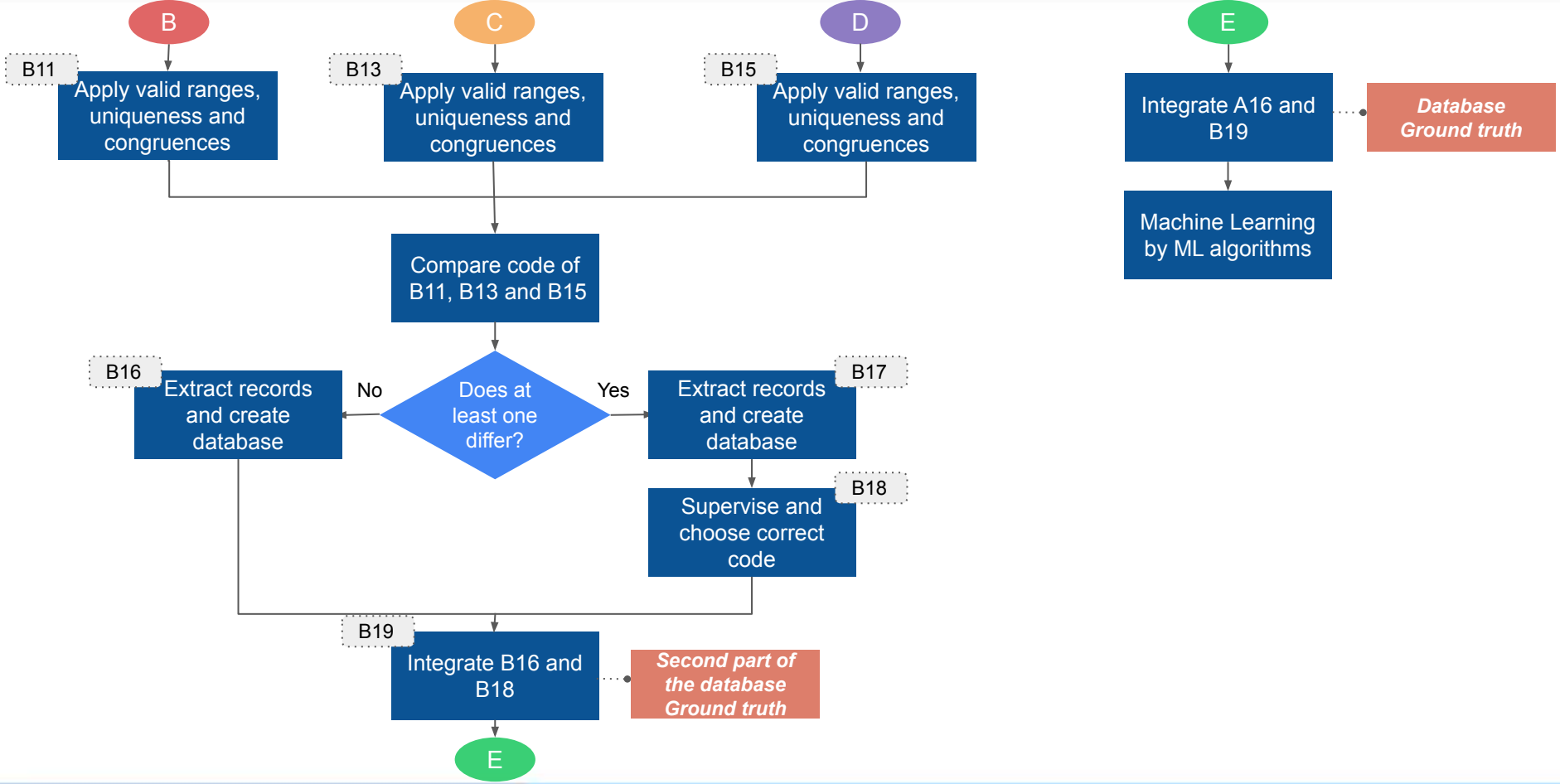
CENTRAL TEAM ML



CENTRAL TEAM




CENTRAL TEAM
Start
Run automatic coding
B5
Integrate records
Yes
Has the code been assigned?
No
B6
Assisted coding
Quality verification
* Central team verifies the quality of the three working groups
B7
Quality verification automatic and assisted (B5 and B6)
B8
Quality verification automatic and assisted (B3 and B4)
B9
Quality verification automatic and assisted (B2 and B1)
B10
Integrate B7
B12
Integrate B8
B14
Integrate B9
B
C
D
STATE
Start
Run automatic coding
B3
Integrate records
Yes
Has the code been assigned?
No
B4
Assisted coding
End
CENTRAL TEAM ML
Start
B2
Encode records matching to B5
B1
Encode records matching to B6
End


B
C
D
E
B11
Apply valid ranges, uniqueness and congruences
B13
Apply valid ranges, uniqueness and congruences
B15
Apply valid ranges, uniqueness and congruences
Compare code of B11, B13 and B15
Does at least one differ?
No
Yes
B16
Extract records and create database
B17
Extract records and create database
B18
Supervise and choose correct code
B19
Integrate B16 and B18
Second part of the database Ground truth
E
Integrate A16 and B19
Database Ground truth
Machine Learning by ML algorithms