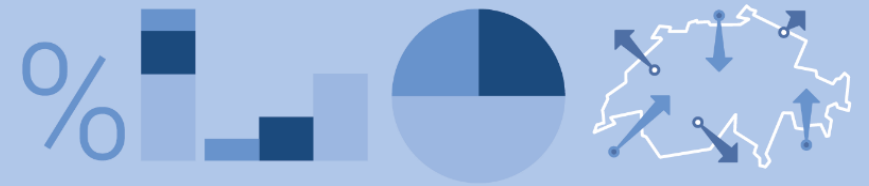




Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Uffizi federal da statistica UST



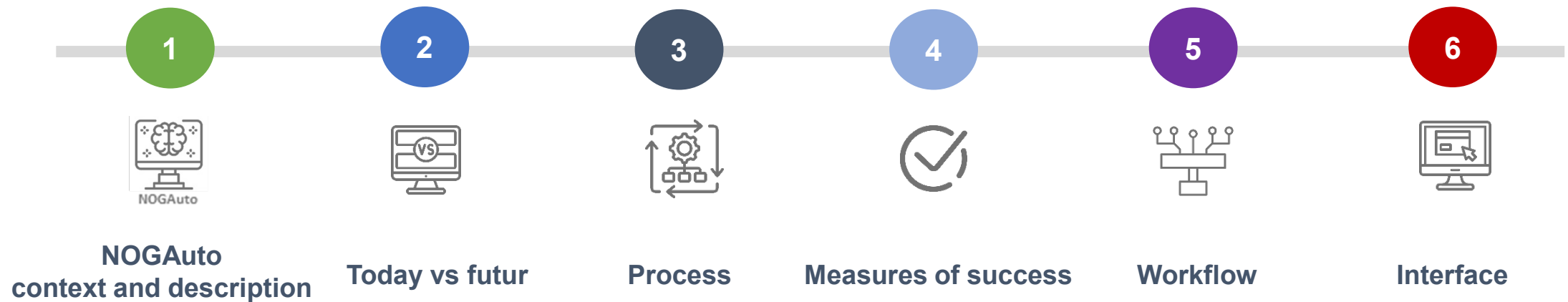
The Data Innovation Project NOGAuto



Lorenz Helbling. Cindia Duc Sfez
Group of Experts on Business Registers
September 28, 2022

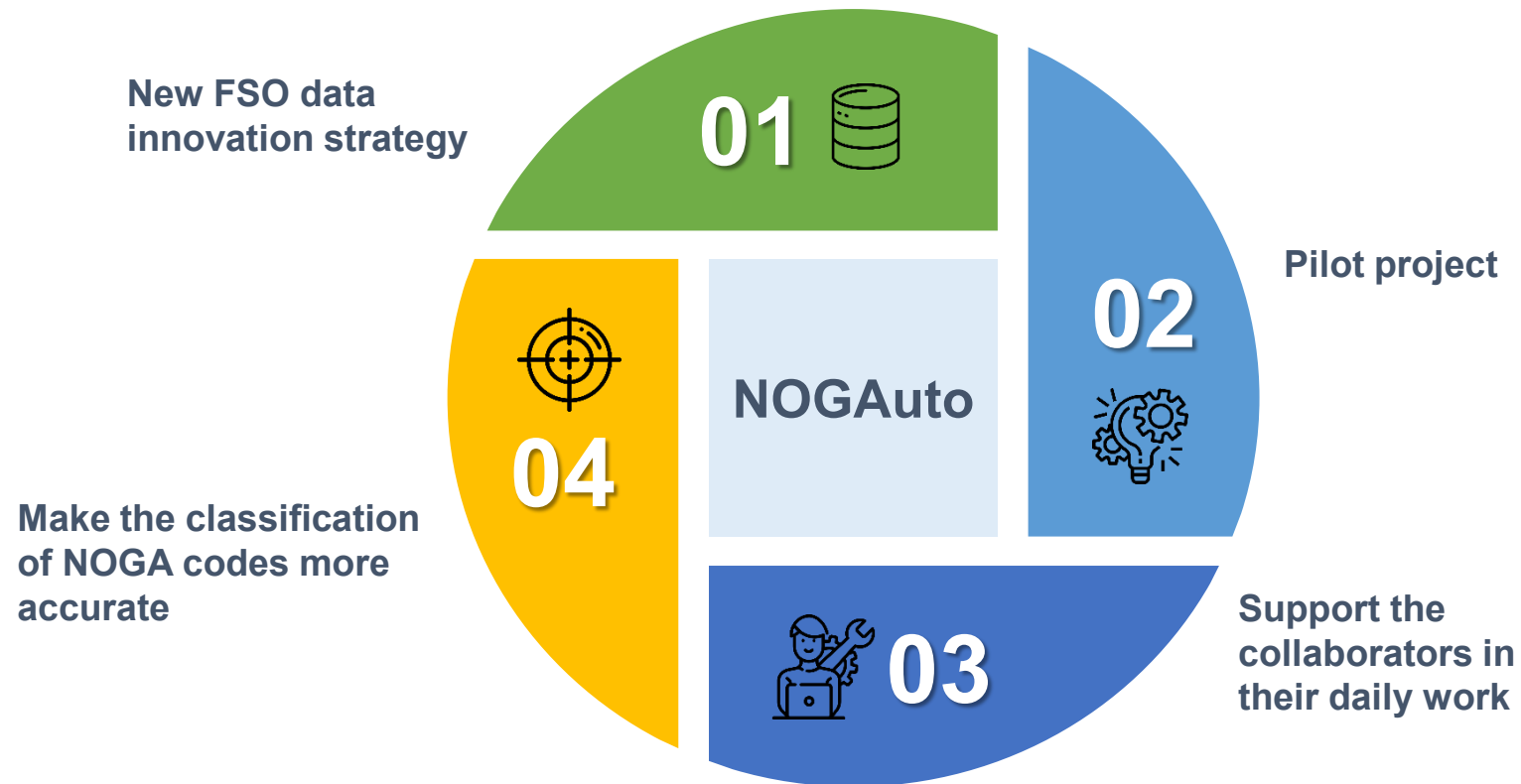


Agenda





NOGAuto Context

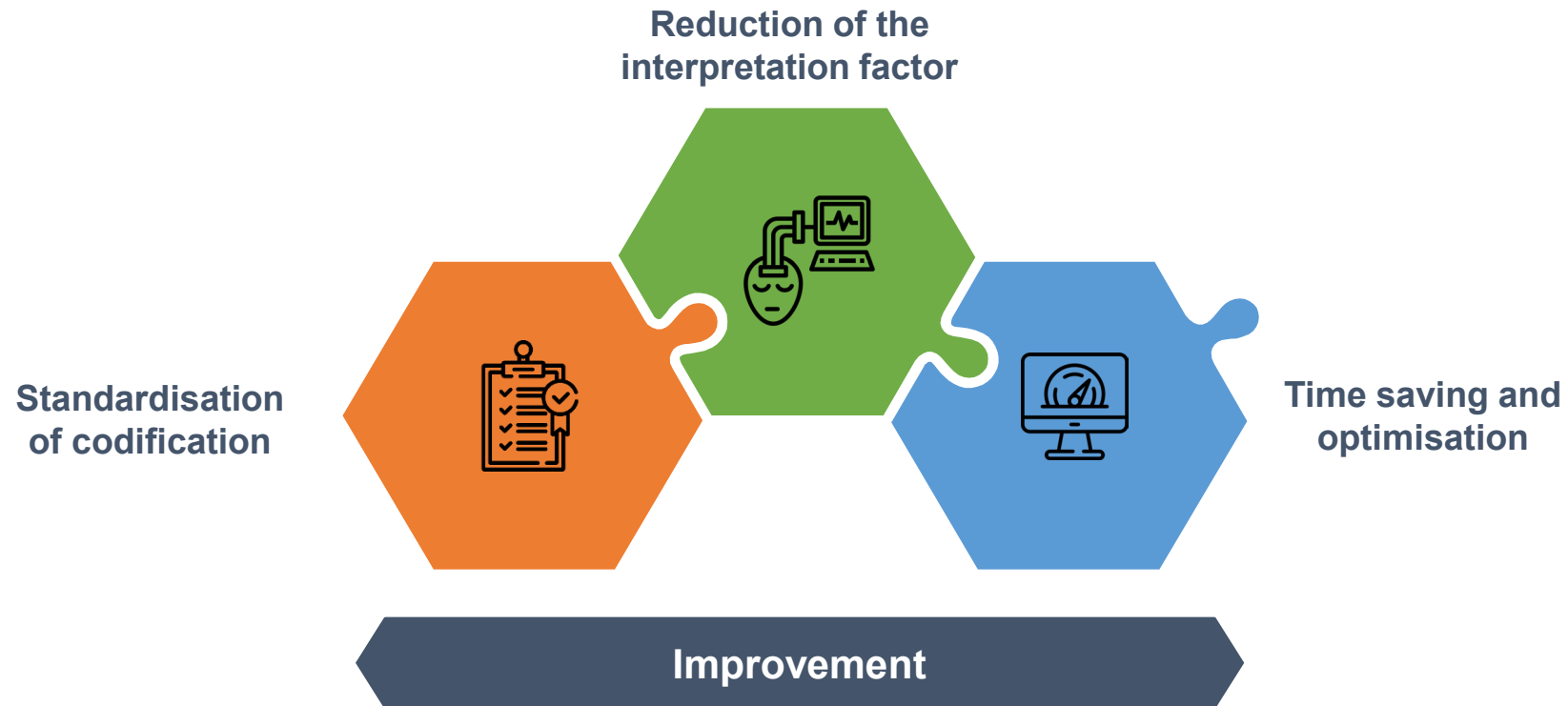


Leading consequently to an overall quality improvement of all enterprise statistics.



Description of the NOGAuto project

The creation of a program that can automatically predict companies' NOGA codes using “Machine Learning” (ML) techniques with a coding quality that is equal or higher than the manual coding currently performed.





Today vs. Futur

Input: Activity description
of a company

Coding



Quality control by
sampling (~ 30%)



Processing of requests



Futur

Today



Input: Activity description
of a company

Coding



+



Quality control by
sampling (~ 99.9%)



Processing of requests



+

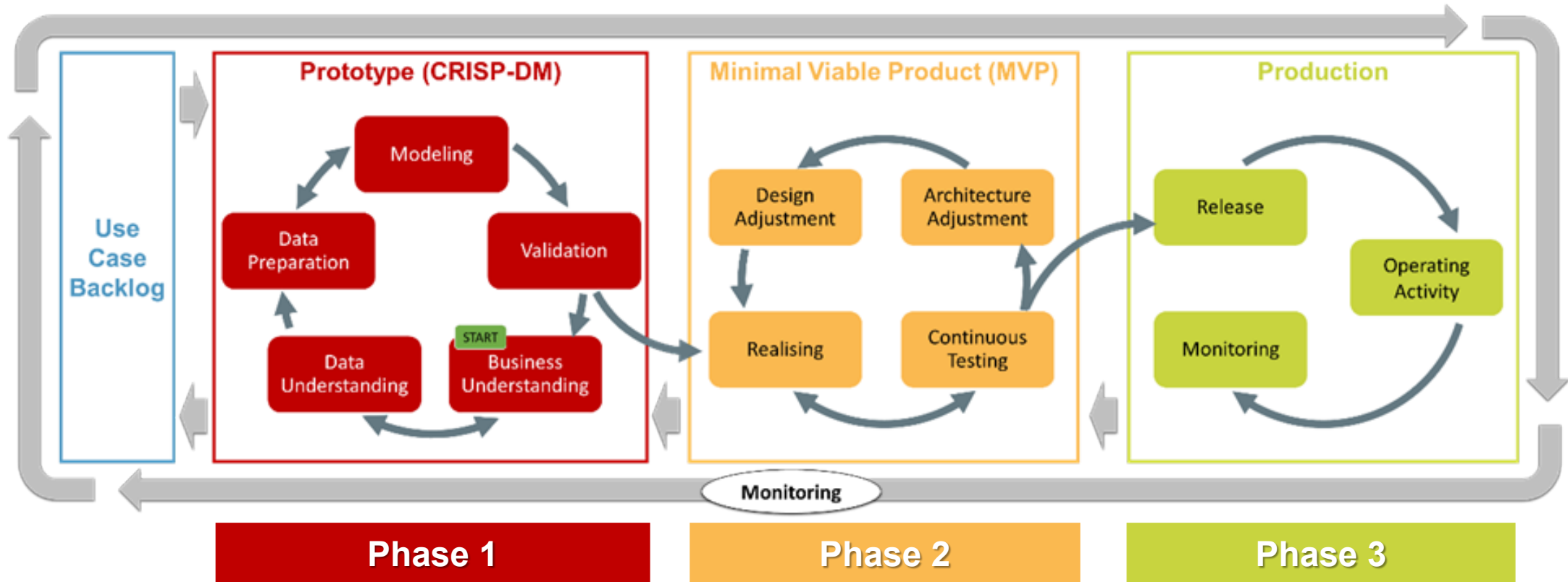


&



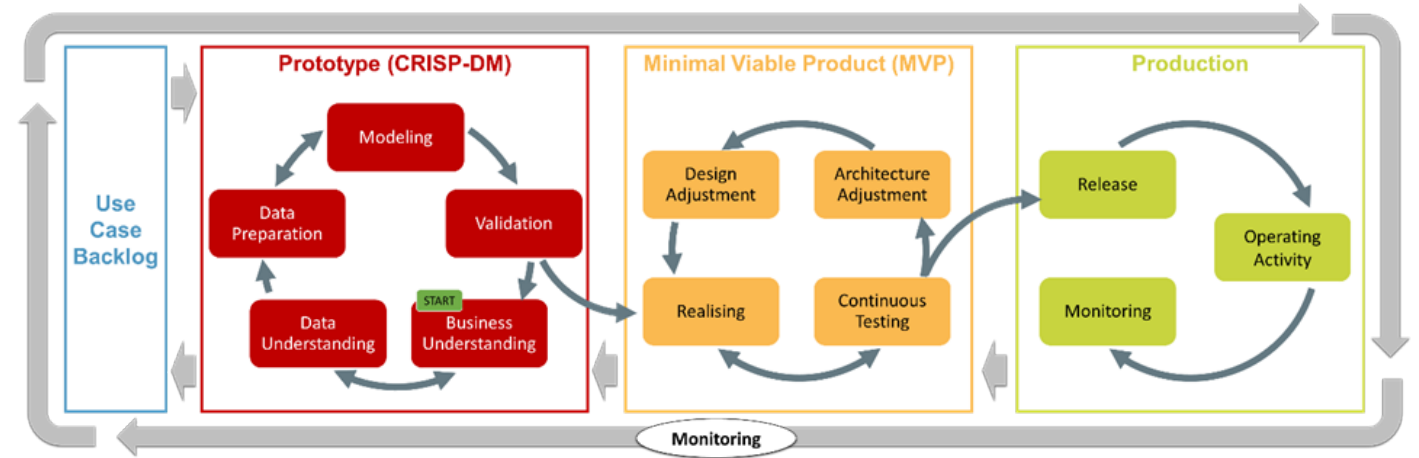


Data innovation as a process





Measures of success



		Phase 1	Phase 2	Phase 3
Prediction ML system-centered	Predictive accuracy (1 – generalisation error)	75 %	85 %	95 %
	Documentation	✓	✓	✓
	Monitoring	✓	✓	✓
Domain user-centered	Interface		✓	✓
	Performance	✓	✓	✓
	Daily use		✓	✓
	Added value for the user	✓	✓	✓

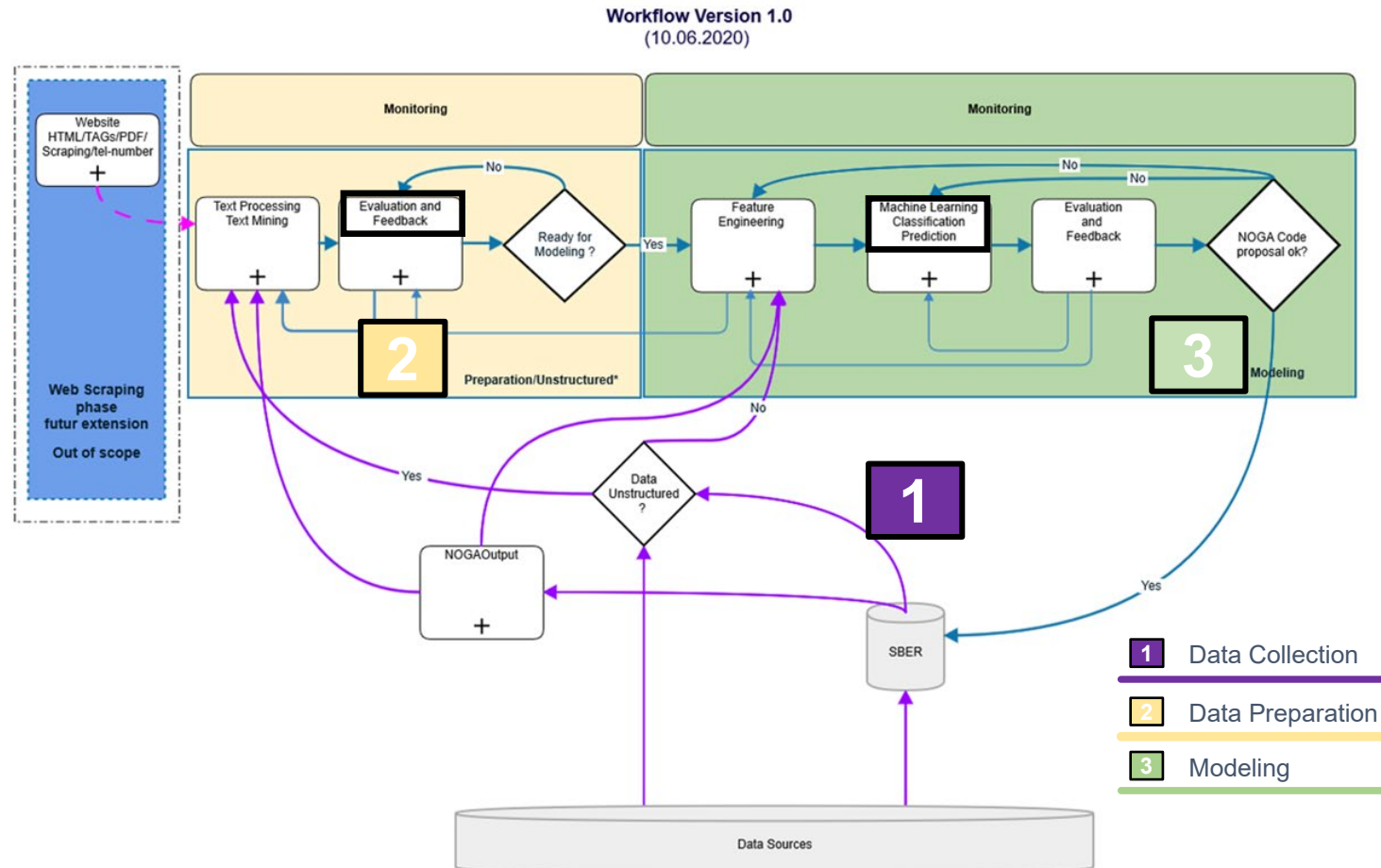


Software used



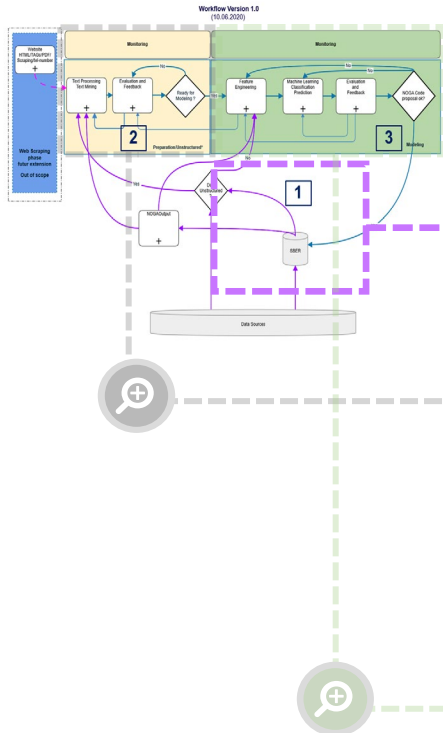


Workflow of the «NOGAuto ML system»





NOGAuto Data Collection



1. Data Collection

- Internal Database
- Business Enterprise Register BER (commercial registers, VAT, customer transactions (borders))
- Over 1'000'000 observations with 56 variables each

2. Data Preparation

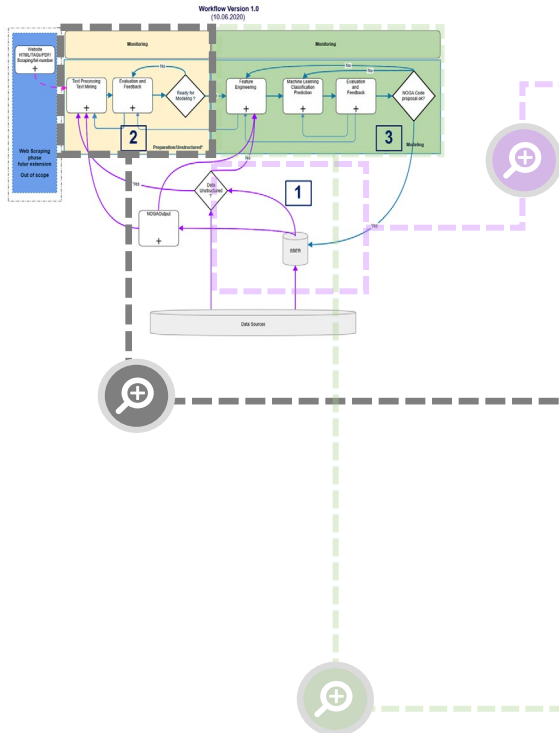
- Language detection (e.g. naive Bayes, deep learning, N-gram, root cause analysis)
- Text mining & natural language processing (NLP) (e.g. one-hot encoding, Text2vec, Word2vec, t-SNE, word clouds)

3. Modeling

- GBM: Gradient Boosting Machine
- Three different models trained according to the language
- Cross-validation



NOGAuto Data Preparation



1. Data Collection

- Internal Database
- Business Enterprise Register BER (commercial registers, VAT, customer transactions (borders))
- Over 1'000'000 observations with 56 variables each

2. Data Preparation

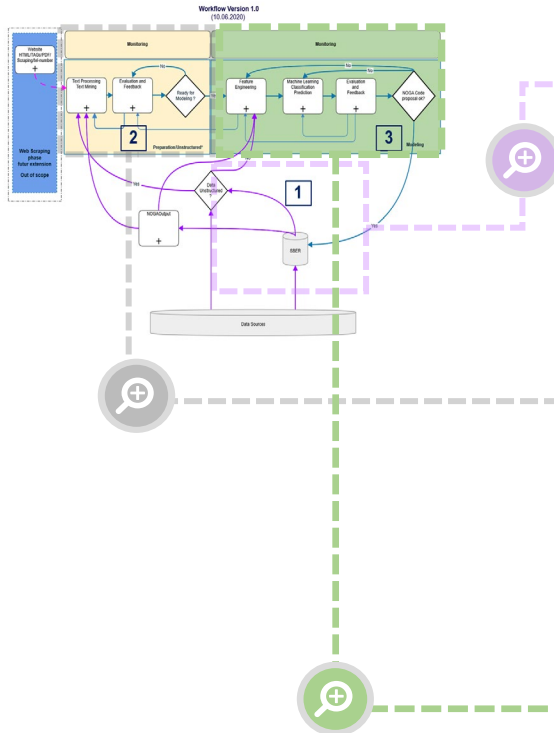
- Language detection (e.g. naive Bayes, deep learning, N-gram, root cause analysis)
- Text mining & natural language processing (NLP) (e.g. one-hot encoding, Text2vec, Word2vec, t-SNE, word clouds)

3. Modeling

- GBM: Gradient Boosting Machine
- Three different models trained according to the language
- Cross-validation



NOGAuto Data Modeling



1. Data Collection

- Internal Database
- Business Enterprise Register BER (commercial registers, VAT, customer transactions (borders))
- Over 1'000'000 observations with 56 variables each

2. Data Preparation

- Language detection (e.g. naive Bayes, deep learning, N-gram, root cause analysis)
- Text mining & natural language processing (NLP) (e.g. one-hot encoding, Text2vec, Word2vec, t-SNE, word clouds)

3. Modeling

- GBM: Gradient Boosting Machine
- Three different models trained according to the language
- Cross-validation



NOGAuto Modeling – 2 Digits predictions

2 Digits Prediction							
Real NOGA_2D	1st Prediction	In %	2nd Prediction	In %	3rd Prediction	In %	Match (=1) vs No Match(=0)
01	01	99.395%	03	0.281%	02	0.114%	1
47	47	86.32%	10	8.718%	46	1.96%	1
96	96	99.87%	47	0.064%	56	0.014%	1
01	86	78.79%	01	19.602%	47	0.191%	1
86	86	41.00%	74	14.32%	70	13.70%	1
47	47	93.15%	70	3.717%	96	1.048%	1
49	96	16.37%	49	15.68%	69	7.51%	1
01	01	99.896%	03	0.037%	96	0.007%	1
01	01	99.988%	03	0.002%	02	0.001%	1
01	01	99.893%	03	0.035%	86	0.009%	1
01	01	99.366%	03	0.515%	96	0.013%	1
49	49	98.74%	88	0.387%	85	0.133%	1
85	85	85.35%	90	10.38%	47	1.044%	1
01	01	99.896%	03	0.037%	96	0.007%	1

Dataset	Predictive accuracy (top 3)
French	~ 99.12%
German	~ 96.16%
Italian	~ 98.25%



NOGAuto Modeling – 4 Digits predictions

4 Digits Prediction						
Real NOGA Code	1st Prediction	In %	2nd Prediction	In %	3rd Prediction	In %
8412	8891	53.85%	8899	7.112%	8810	4.48%
0143	0143	13.81%	0150	13.62%	0121	8.602%
8621	8621	24.05%	8622	19.03%	8690	4.455%
0141	0141	10.54%	0150	10.42%	0121	7.594%
8690	8690	52.246%	9602	5.707%	9609	5.373%
8621	8621	51.30%	8622	4.221%	8559	3.22%
0149	0149	31.91%	0150	14.55%	0143	1.639%
2341	2341	24.078%	8552	19.55%	8559	6.423%
0150	0141	19.24%	0111	7.834%	0150	6.713%
8622	8621	58.545%	8622	8.796%	9900	8.482%

Dataset	Predictive accuracy (top 3)
French	~ 88%
German	~ 40%
Italian	~ 95%



Example of activity descriptions

1

FRENCH

La société a pour but les conseils et formations en matière de finance, marketing et développement commercial; elle peut acquérir et administrer des participations dans d'autres entreprises ainsi qu'exercer des activités de support du management (pour but complet cf. statuts).

2

GERMAN

Ausführung aller Arbeiten im Fachgebiet Schreinerarbeiten, sowie den Handel mit Produkten der Holzindustrie.

3

ITALIAN

La produzione e la vendita di vini e distillati, aceto e olio d'oliva sia in Svizzera che all'estero così come ogni altra attività atta a conseguire lo scopo sociale. La società potrà inoltre acquistare, vendere e amministrare beni immobili.



UI: User Interface

FSO's Classification System NOGAAuto Other classification

Einfügen der benötigten Variablen

Geben Sie die Beschreibung der Unternehmensaktivitäten ein

exploitation d'un atelier d'ébénisterie-menuiserie, fabrication et pose de tous produits relevant des domaines de l'ameublement et de la construction. Achat, exploitation, mise en valeur et vente de tous immeubles, bâtis ou non, et de tous droits immobiliers.

Erkannte Sprache:
Französisch

Wenn die Sprache nicht korrekt ist, wählen Sie bitte:
 Französisch Deutsch Italienisch

Search

Erweiterte Suche

Wählen Sie die Rechtsform des Unternehmens

Geben Sie die Anzahl der Mitarbeiter im Unternehmen an

NOGA Code Voraussagen

Erste Voraussage

1623

27.74%

0 100

Select code

Zweite Voraussage

1629

19.52%

0 100

Select code

Dritte Voraussage

1622

17.61%

0 100

Select code

NOGA Code	Code Beschreibung	Activity description
1623	Herstellung von sonstigen Konstruktionsteilen, Fertigbauteilen, Ausbauelementen und Fertigteilbauten aus Holz	Initial sentence exploitation d'un atelier d'ébénisterie-menuiserie, fabrication et pose de tous produits relevant des domaines de l'ameublement et de la construction. Achat, exploitation, mise en valeur et vente de tous immeubles, bâtis ou non, et de tous droits immobiliers. Cleaned sentence ateli ebenister menuiser fabriqu pos produit relev domain ameubl construct achat mis vent immeubl bat droit immobili
1629	Herstellung von Holzwaren a. n. g. Kork-, Flecht- und Korbwaren (ohne Möbel)	
1622	Herstellung von Parketttafeln	

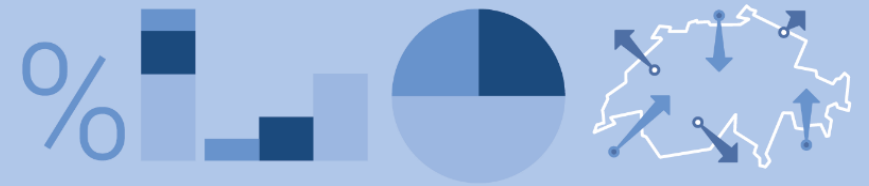
Feedback

Schreiben Sie unten einen Feedback

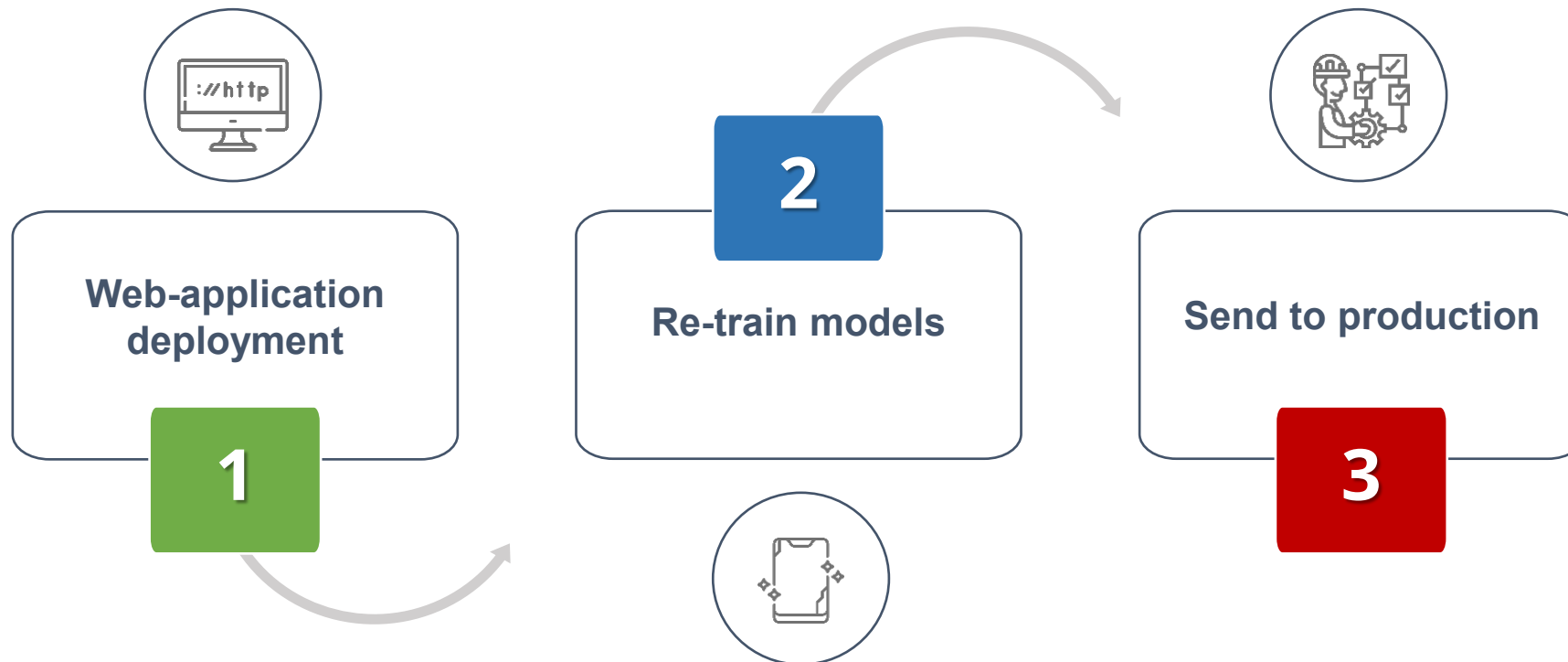
Feedback senden

Powered by





Next steps





Other potential deployment



NACE Revision (2024-2025)

- Potential supporting tool in the double codification for the collaborators for the complex cases namely companies with 1:n relations.
- The system will use only the observations of the dataset which are in the n categories and will train a ML model for the company being recoded.



Application on other nomenclatures

- Use the same methodology on different nomenclatures such as the profession ISCO nomenclature.
- First results are promising



Thank you for your attention Questions?

