



# Entity Recognition and Accounting Type Prediction: Simplifying Accounting on the Basis of Machine Learning

Hasham Munir and David Bläsi

Sevenit GmbH



sevDesk



# About us

- David Bläsi
  - M.Sc. in Mathematics (2017) with focus on probability theory and medical statistics
  - Data Scientist / Machine Learning Engineer at Sevenit GmbH since September 2017: focus on automated invoice processing
- Hasham Munir
  - M.Sc. Communication & Media Engineering 2015 with focus on Database systems & Technologies
  - Data Scientist / Machine Learning Engineer at Sevenit GmbH since Oct.2016



# Agenda

1. Introduction
2. Entity Recognition for invoice data
  - Task description
  - Workflow
  - Model comparison
3. Accounting type prediction
  - Idea of using accounting type
  - sevCleaner
  - sevProto
  - sevAuto
4. Live demo



# Introduction: sevDesk and its functionalities



Digitize receipts



Cancellation bills / credits



Write bills



Write reminders



Online banking



VAT return



order processing



Net income method



and many more...





# Entity Recognition for invoice data

David Bläsi - Data Scientist / Machine Learning Engineer



# Task description - motivation

- Idea: facilitate process of digitalizing relevant information from invoices and receipts
  - Our users:
    - have between 10 and 150 incoming invoices each month
    - come from a wide variety of domains
- Goal: build universal invoice recognition system

# Task description - entity recognition

- Goal: recognize and classify parts of natural language texts
- Named entity: e.g. persons, organizations, dates, etc.

Organisation Time Location  
Sevenit GmbH was founded in October 2013 and is based in Offenburg.

- Supervised learning algorithms have proven suitable for this task





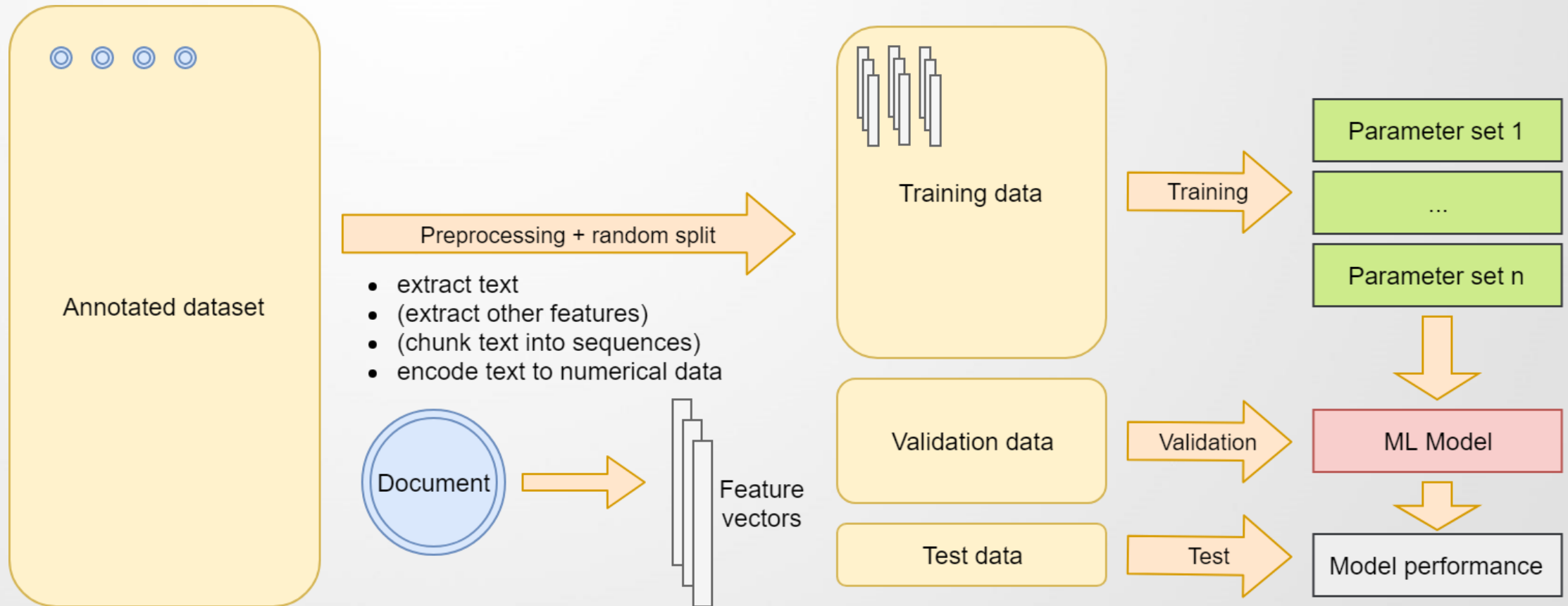
# Workflow - data preparation

- We use data provided by our users to train our models
- Raw data has to be annotated to create a dataset for supervised learning

The screenshot displays a mobile application interface for document viewing. On the left, a sidebar shows the user's name 'David Bläsi', search and filter options, and a list of documents, including one titled 'as' from '2018-03-12 16:02:05'. The main area shows a document titled 'Rechnung Für den Zeitraum 01.12.-31.12.2016' from 'Drillisch Online AG'. The document content includes a 'winSIM' logo, a date of '31.12.2016', and a tariff of 'LTE All 3 GB'. The right sidebar lists various metadata fields such as 'IBAN (iban)', 'BIC (bic)', 'Logo (l)', 'CreditorName (cn)', 'InvoiceNumber (in)', 'InvoiceDate (id)', 'NetAmount (net)', 'TaxRate (tax)', 'Amount (a)', 'TaxId (tid)', and 'VATId (vid)'. The document is marked as 'Seite 1 von 1'.



# Workflow - training ML models



# Model comparison

## Conditional random fields (CRF)

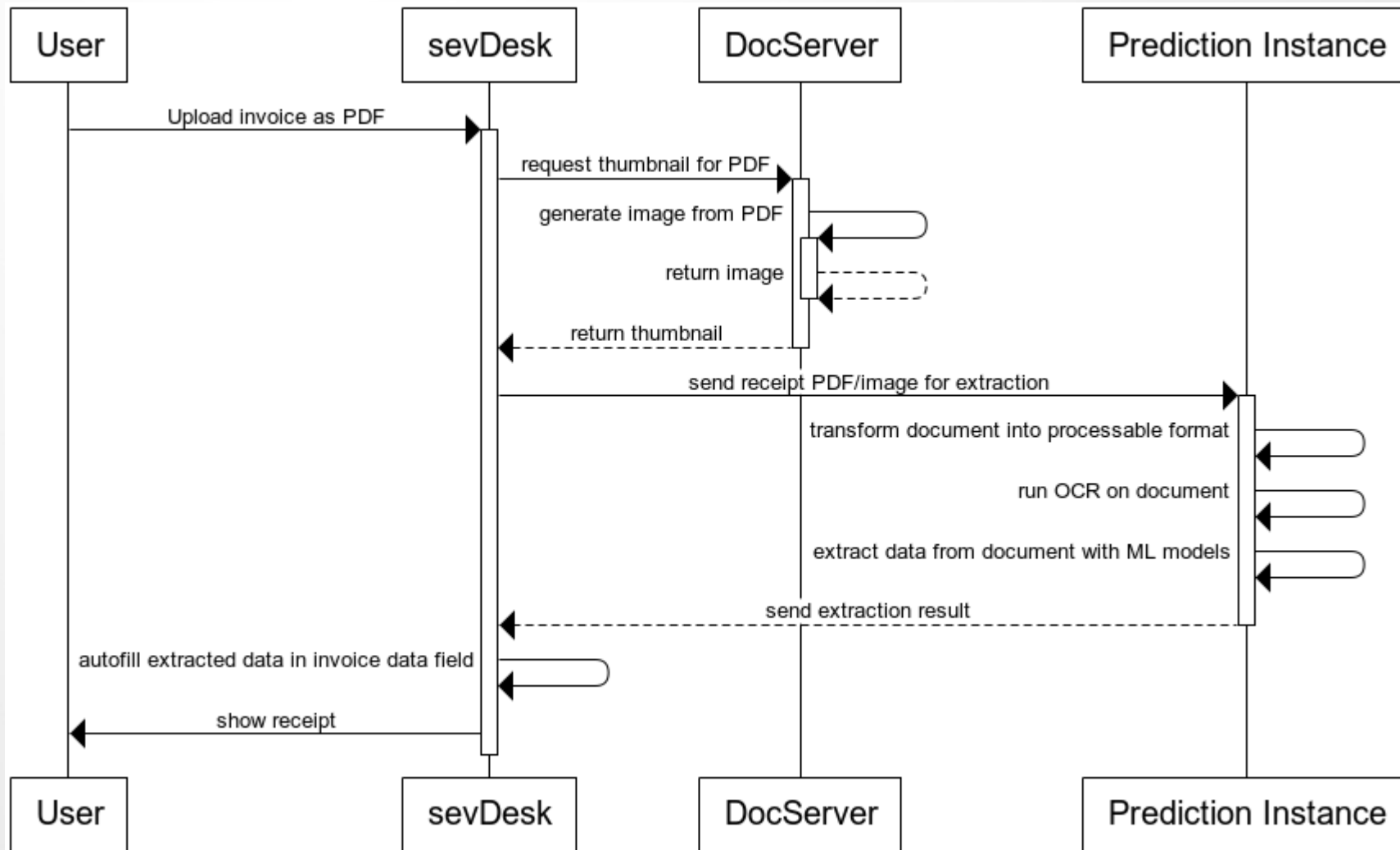
- relies on hand-crafted features
- works well with small amount of data
- flexible document length
- comparatively few open source frameworks / libraries

## Recurrent neural networks (RNN)

- no hand-crafted features
- scales well to large amounts of data
- expects rigid data format (document length)
- large choice of frameworks and active community



# The model in production





# Accounting Type Prediction

Hasham Munir - Data Scientist / Machine Learning Engineer



# Accounting Type

- Accounting type helps to store the VAT
- Public accounting types
- Private accounting types

### Buchungskonto

**Name**

**Kontoart**

Einnahmen

**SKR03 Konto**

**Automatikkonto**

**Zuordnung Zahlungskorrektur**

OK

### Kategorie auswählen

Eigene Buchungskonten	>	Darlehen & Tilgung	640
Privat	>	Aufnahme und Rückzahlung eines Darlehens	
<b>Banken / Finanzen</b>	>	Geldtransit	1360
Büro	>	Für die Geldbewegung zwischen zwei betrieblichen Konten	
Dienstleistung / Beratung	>	Girokonto Zinsen	2110
Fahrzeug	>	z. B. für einen Dispo-Kredit	
Maschine / Gebäude	>	Kontoführung / Kartengebühren	4970
Material / Waren	>	z. B. Kontoführungsgebühren und Gebühren für Kreditkarten	
Personal	>	Kreditgebühren	2120
Raumkosten	>	Gebühren für einen aufgenommenen Kredit	
Reisen / Verpflegung	>	Kreditzinsen	2120
Sonstiges	>	Zu zahlende Zinsen für einen Kredit	
Versicherungen / Beiträge	>		
Werbung	>		
Steuer	>		
Verbindlichkeiten	>		
Sonstige Erträge	>		
Umsätze	>		

■ Kosten ■ Erträge

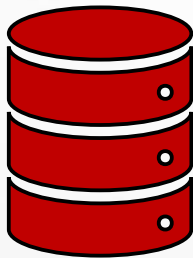


# sevCleaner

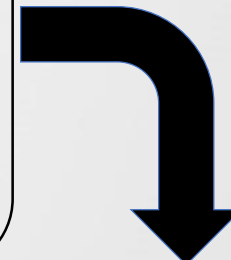
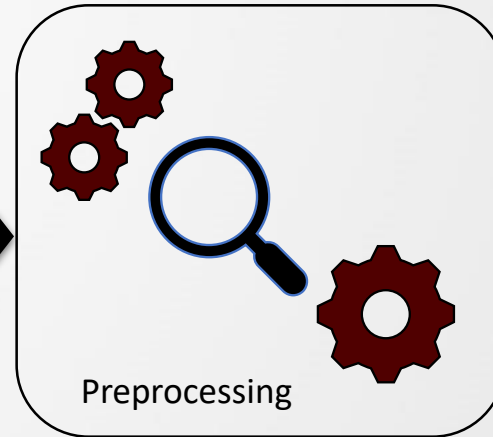
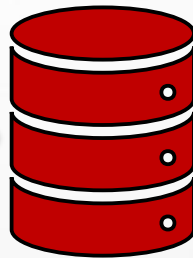
W.W Customers



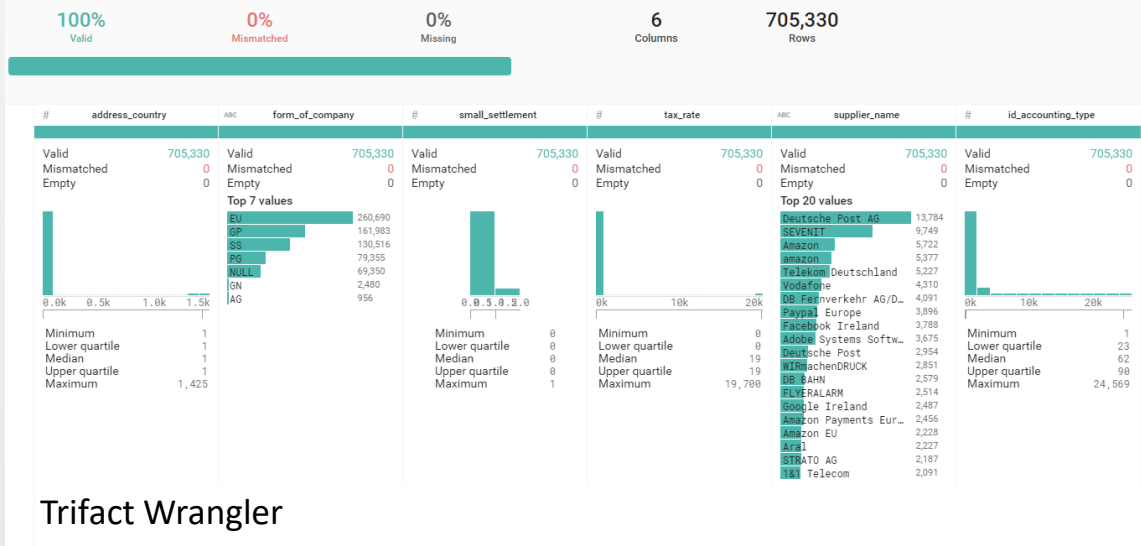
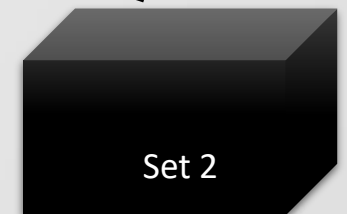
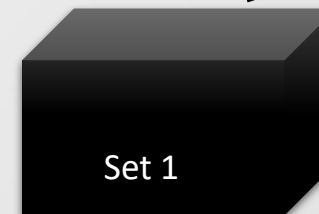
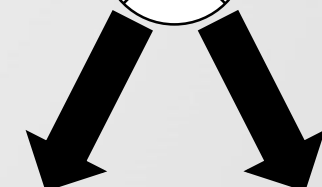
Live Servers



Test Server

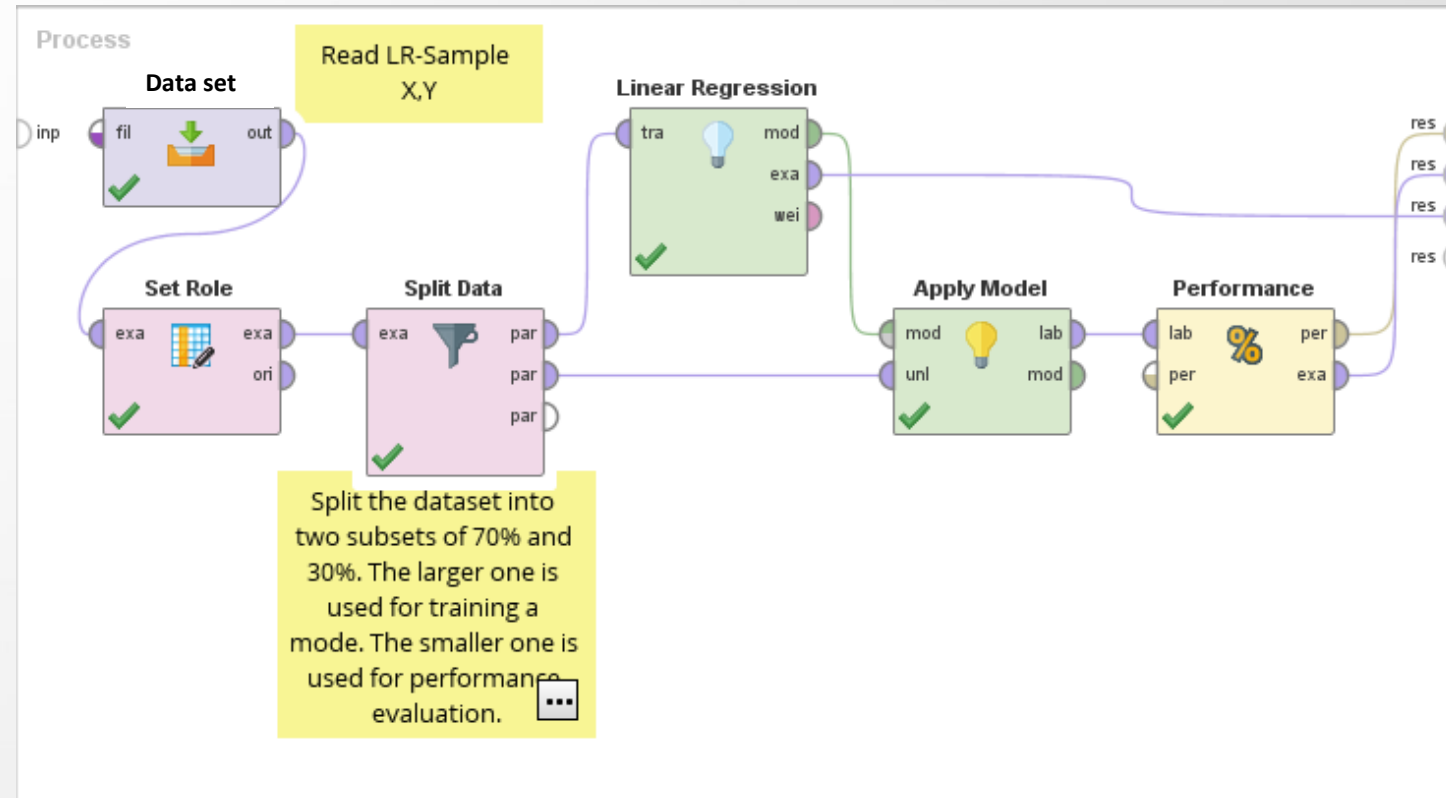


Data Splitter



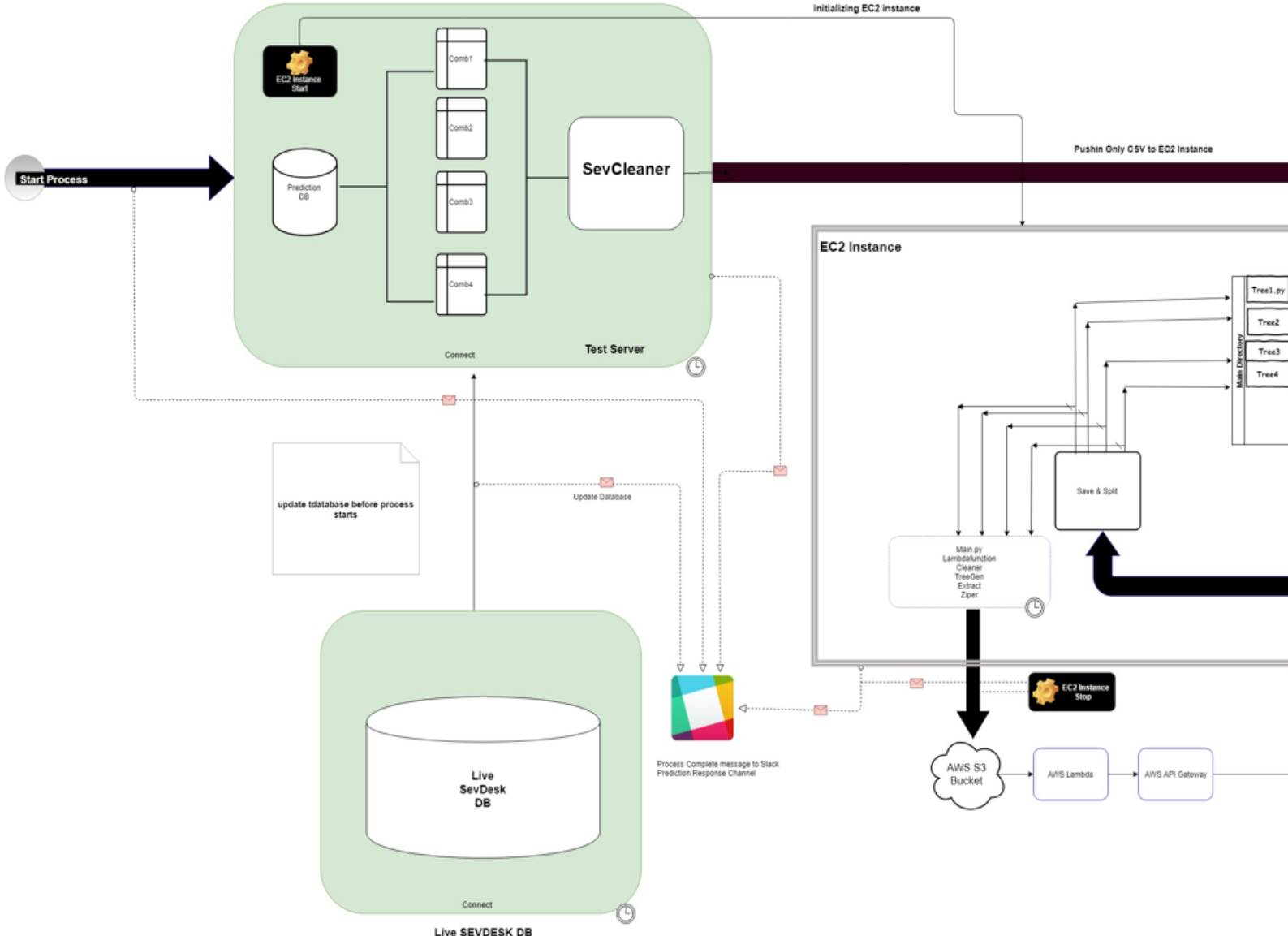
# sevProto

- Select Dataset
- Using Rapidminer
- Model Selection
- Model with high accuracy
- Testing with actual data





# Prediction Automation with EC2 Instance



## #pebe\_prediction

☆ | 👤 4 | 🗨️ 0 | ➕ Add a topic

Initializing EC2 Instance.....  
 .....  
 .....  
 EC2 instance started.....  
 Database update initialized...  
 process Data update start  
 Database updated  
 removing null accounting\_types from Database  
 Database connection closed  
 Database Connection start  
 Tree1 Transfer process ON  
 Transfer complete  
 SUCCESS: All Folders Updated in EC2 INSTANCE  
 Tree Generation will start in EC2

Tree Generation Process Start

Reading from csv file

CSV Loaded  
 Dataframe ready  
 Supplier Dictionary ready  
 FOC Dictionary ready  
 Train and Target ready

| KNN Score 75.82417582417582  
 KNN file stored with joblib protocol 2  
 Cross Validation Process  
 Cross Validation complete  
 RF Score 76.55677655677655


| Bagging with Bagging\_knn file stored  
 | Bagging with Bagging\_DT file stored  
 | RF file stored  
 | zip complete

Automation successfull

Boto connect ON  
 | S3 Bucket Pebe-rel-2 updated  
 EC2 Instance Shut Down

# Live Demonstration





Thanks for your attention!

Questions?

