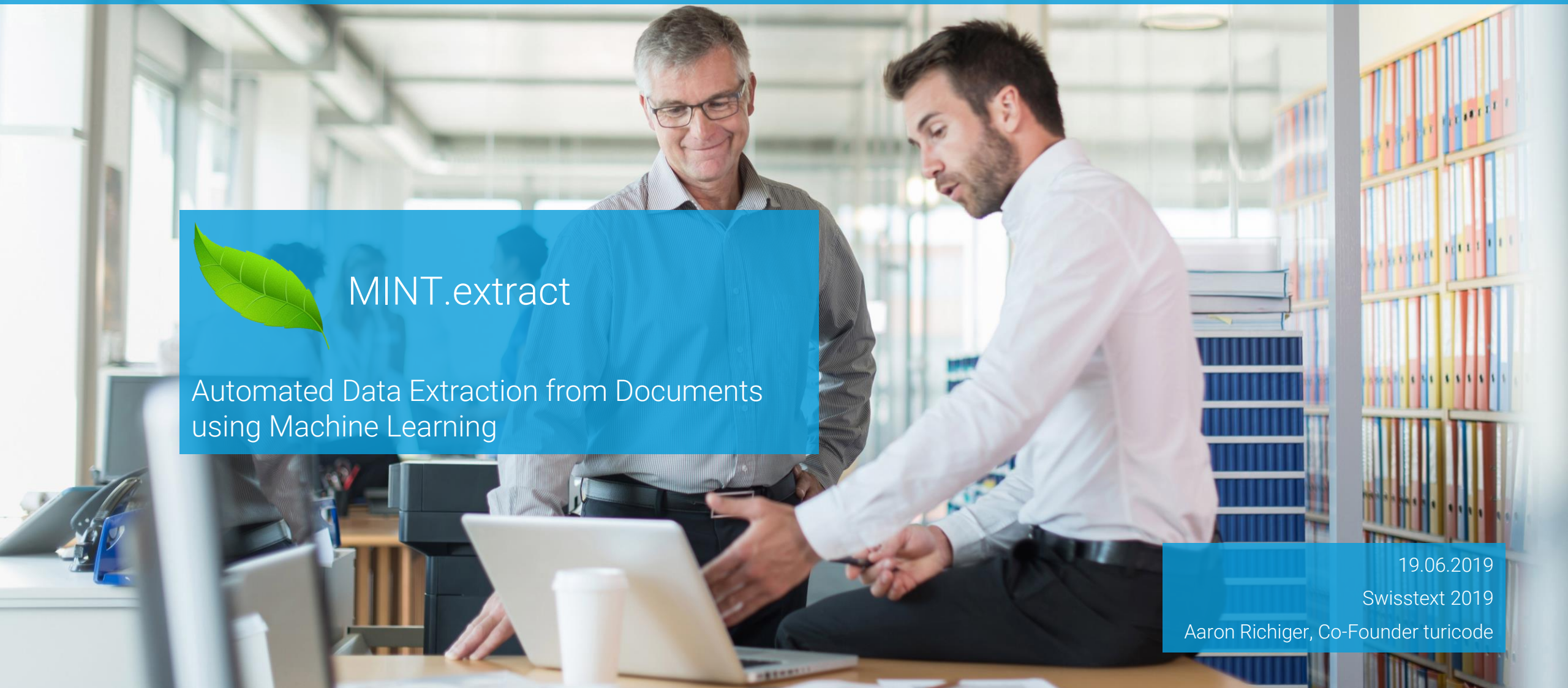
# Truly refreshing document analysis



## MINT.extract

Automated Data Extraction from Documents using Machine Learning

19.06.2019

Swisstext 2019

Aaron Richiger, Co-Founder turicode

# Agenda

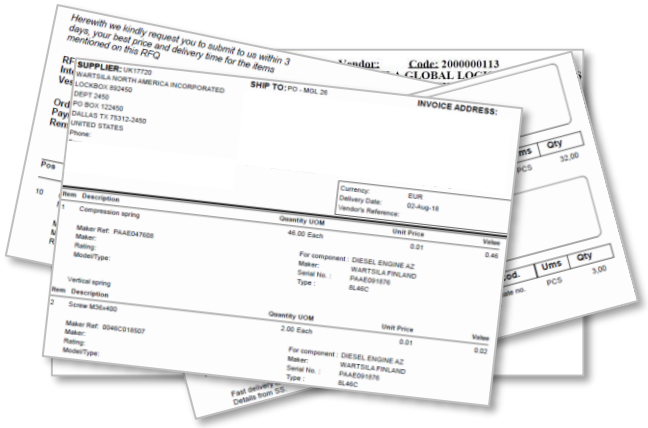
MINT.extract

- What it is
- How it works
- Results
- How to use it

# Problem Statement

MINT.extract

## Input



Input:

- Scans, PDFs, Emails, Excel/CSV, etc.
- Different structures and layouts
- High volumes and/or frequencies

## Software as a Service

Our Focus:

- Classification
- Extraction
- Semantic analysis
- Validation

MINT.extract

## Output

RESTful API

JSON

XML

Output:

- Machine-readable data
- Web or mobile apps
- Simple integration into existing systems

# How MINT.extract structures your documents

MINT.extract

Input | Pre-Processing | Classification | Data - Extraction | Data - Analysis | Output

OCR / Text Recognition

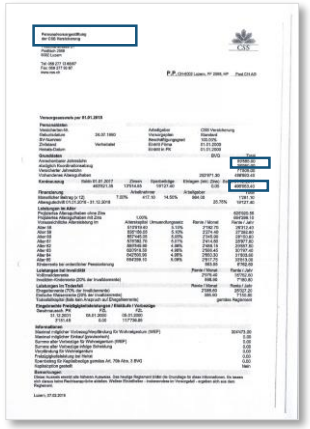
PDF

Pension Statements

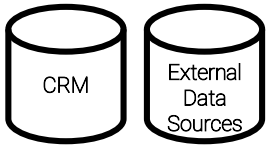
Tax Statements

Bank Statements

Insurance Policies



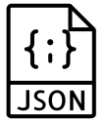
Extraction of text, images etc. with Machine Learning




Automated Validation and Enrichment


Web-Application for visual inspection

RESTful API



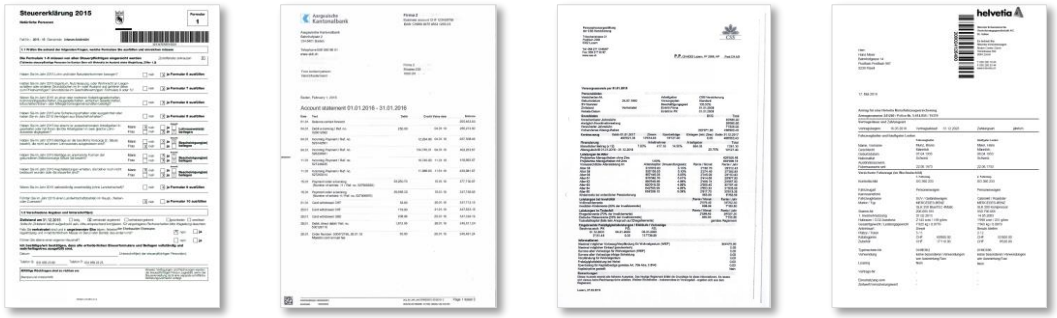



```
{
"Pensionfund":
"Personalvorsorgestiftung
der CSS Versicherung ",
"Salary": 99585 ,
"Saldo": 496863
}
```

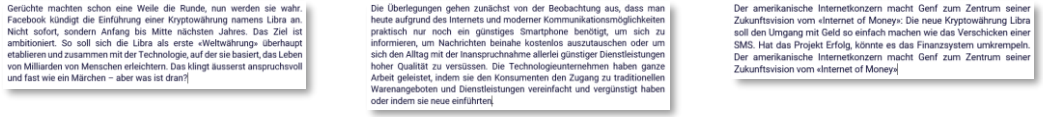
# How it works: Classification

MINT.extract

## Image Based Classification



## Text Based Classification



Gerüchte machten schon eine Weile die Runde, nun werden sie wahr. Facebook kündigt die Einführung einer Kryptowährung namens Libra an. Nicht sofort, sondern Anfang bis Mitte nächsten Jahres. Das Ziel ist ambitioniert. So soll sich die Libra als erste «Weltwährung» überhaupt etablieren und zusammen mit der Technologie, auf der sie basiert, das Leben von Milliarden von Menschen erleichtern. Das klingt äusserst anspruchsvoll und fast wie ein Märchen – aber was ist dran?

Die Überlegungen gehen zunächst von der Beobachtung aus, dass man heute aufgrund des Internets und moderner Kommunikationsmöglichkeiten praktisch nur noch ein günstiges Smartphone benötigt, um sich zu informieren, um Nachrichten beinahe kostenlos auszutauschen oder um sich den Alltag mit der Inanspruchnahme allerlei günstiger Dienstleistungen hoher Qualität zu versüssen. Die Technologieunternehmen haben ganze Arbeit geleistet, indem sie den Konsumenten den Zugang zu traditionellen Warenangeboten und Dienstleistungen vereinfacht und vergünstigt haben oder indem sie neue einführten.

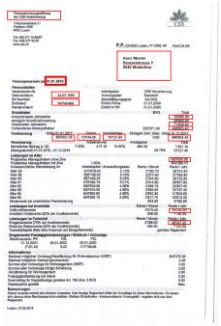
Der amerikanische Internetkonzern macht Genf zum Zentrum seiner Zukunftsvision vom «Internet of Money»: Die neue Kryptowährung Libra soll den Umgang mit Geld so einfach machen wie das Verschicken einer SMS. Hat das Projekt Erfolg, könnte es das Finanzsystem umkrempeln. Der amerikanische Internetkonzern macht Genf zum Zentrum seiner Zukunftsvision vom «Internet of Money»

## Combination of Text and Image Based Classification

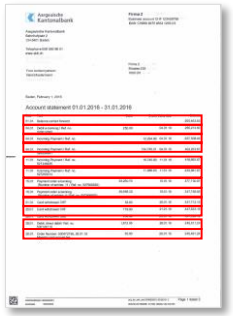
- Showed the best results
- Importance of text or image input depends on the documents

# How it works: Data Extraction

MINT.extract



1. Word level classification based on
   - Text value
   - Textual context
   - Layout information

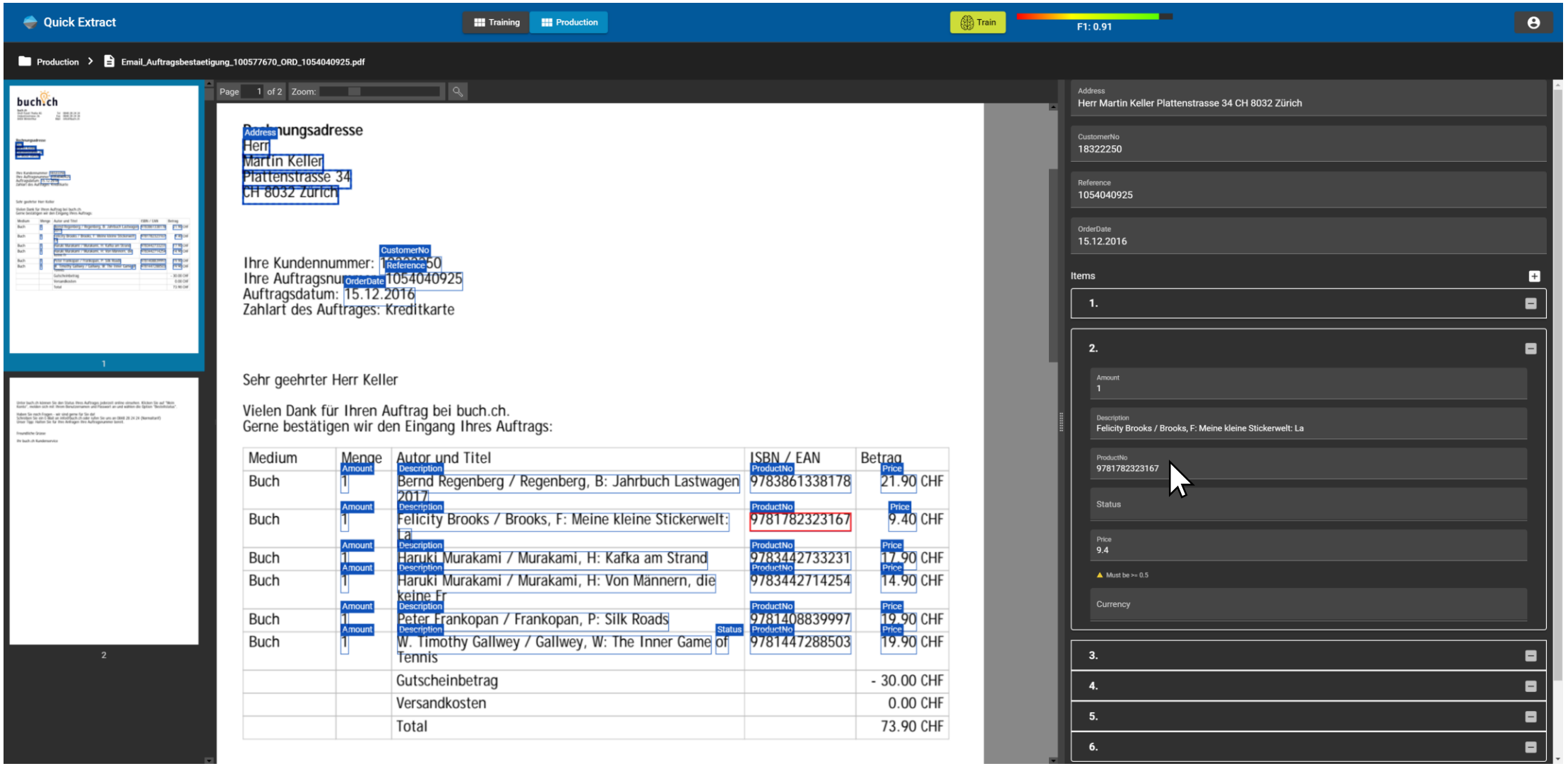


2. Grouping datapoints to structured entities


3. Character level classification

# How it works: Data Extraction

MINT.extract



# How it works: Analytics & Validation

MINT.extract

Based on the extracted data, we build project-specific ML systems for further analytics and validation:

## Data Analytics with ML

- NLP (e.g. NER, topic detection, etc.)
- Predictive Maintenance
- Trigger a specific workflow

## Validation with ML

- Anomaly detection
- Validation classification (critical vs. non-critical validation violations)

# Results: Classification

MINT.extract

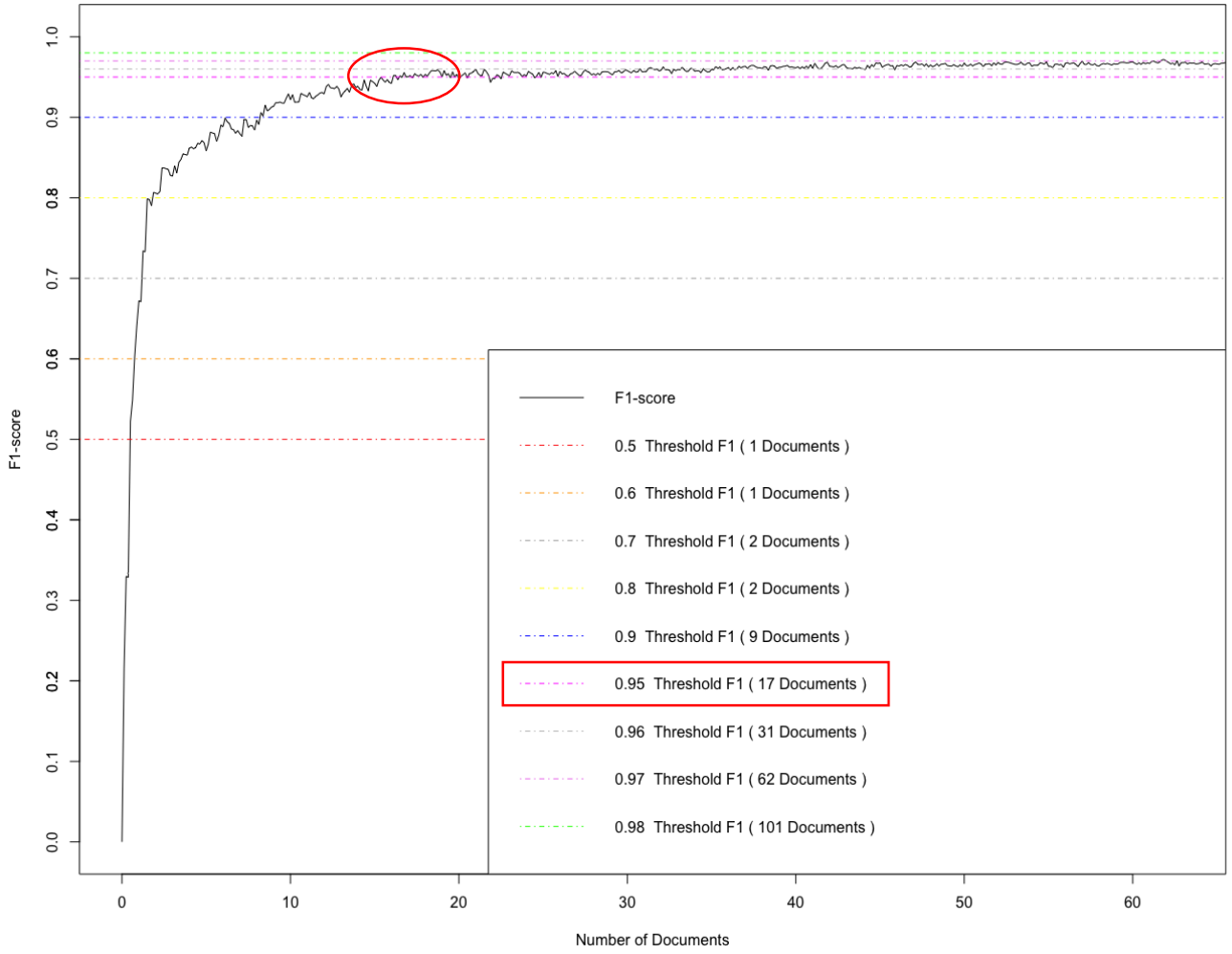
Documents: Financial documents
Case: Incoming mail classification
Data: 17k sample documents for 4 classes
Approach: Combined text- and image-based classification (80/20 split with CV)

| Class | Precision | Recall | F1 score |
|---|---|---|---|
| Official Documents (Amtliche Dokumente) | 0.98 | 0.99 | 0.99 |
| Card Application (Kartenanträge) | 0.99 | 0.99 | 0.99 |
| Direct Debiting Application (LSV) | 0.99 | 0.99 | 0.99 |
| Payment Order (Zahlungsauftrag) | 0.96 | 0.98 | 0.97 |
| **Weighted Average** | **0.99** | **0.99** | **0.99** |

# Results: Data Extraction

MINT.extract

| | |
|---|---|
| Documents: | Purchase Orders |
| Case: | Data extraction and validation |
| Data: | 70 documents from 3 customers with 8 data points to extract |
| Approach: | Word level classification with 80/20 split and CV |



# How to use MINT.extract

MINT.extract

## Option 1: Service provided by turicode

- You send us the documents and the spec
- We provide you with:
  - REST API (for ongoing services)
  - Excel, JSON, XML, TXT (for projects)

## Option 2: Try and train it yourself

- Request access to MINT.extract on our website [turicode.com](turicode.com)
- We provide you access to your own MINT.extract web application
- Upload, annotate and process your documents

# Excerpt from our References

MINT.extract

Kantonspolizei Zürich

MIGROSBANK

NOVARTIS

GSK SHAS SSAS Gesellschaft für Schweizerische Kunstgeschichte

Vontobel

ETH zürich

METTLER TOLEDO

LEGARTIS

SRF

Stadt Zürich

Universität Zürich[UZH]

wüestpartner

Supported by:

swiss made software

Co-funded by the Horizon 2020 programme of the European Union

# Contact


turi
code
turicode AG
Technoparkstrasse 2
CH-8406 Winterthur
Aaron Richiger
aaron.richiger@turicode.ch
+41 79 613 71 78
www.turicode.com
swiss made software