23384: How AI Revolutionizes Data & Document Management

LEVERTON

Your Data. Your Value.

Florian Kuhlmann (CTO & Co-founder)

LEVERTON - Company overview

100+

Global Corporate Clients

20+

Languages supported

100%

Data Transparency

75+

Employees in offices in
Berlin, London and New York

4+

Years using Deep Learning
Technology

Highest

Data Security – ISO 27001 / 9001

LEVERTON - Company vision

From documents to data

Which problem does LEVERTON solve?

Manual data input

- Inefficient → Use AI to classify documents and extract data
- Non-transparent → Keep link to position in document
- Error prone → Control using efficient 4-eyes principal

Data input with LEVERTON

Three steps from document to data

Upload scanned PDFs

Automated extraction

Review and correct

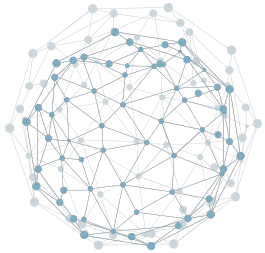
OCR – convert images into machine readable text

Document Classification – categorize documents

Information Extraction – extract relevant data

Train

DATA CORE



Deep Learning

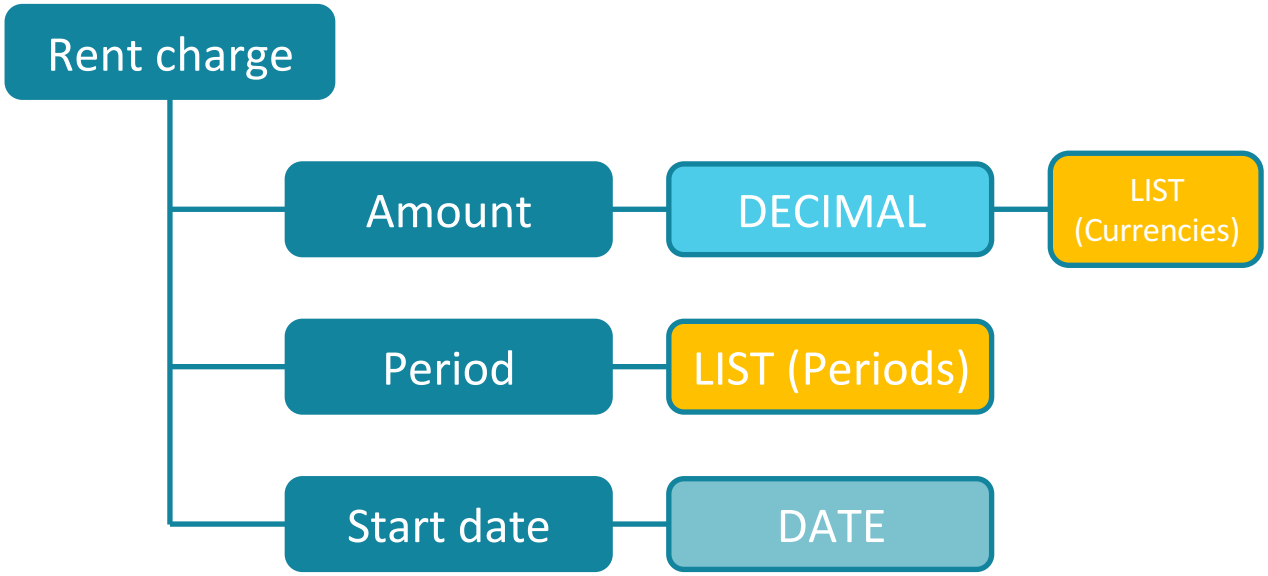
Information Extraction

Automatically extract structured information from un- or semi-structured machine-readable documents
(Wikipedia)

1. Define the **type of information** (e.g. as part of a **data model**)
2. Find **location** (start + end, coordinates) of information **in document**
3. **Extract** the information (parse numbers, dates, …)

Information Extraction: Define Types

Data model as abstraction of real world entities

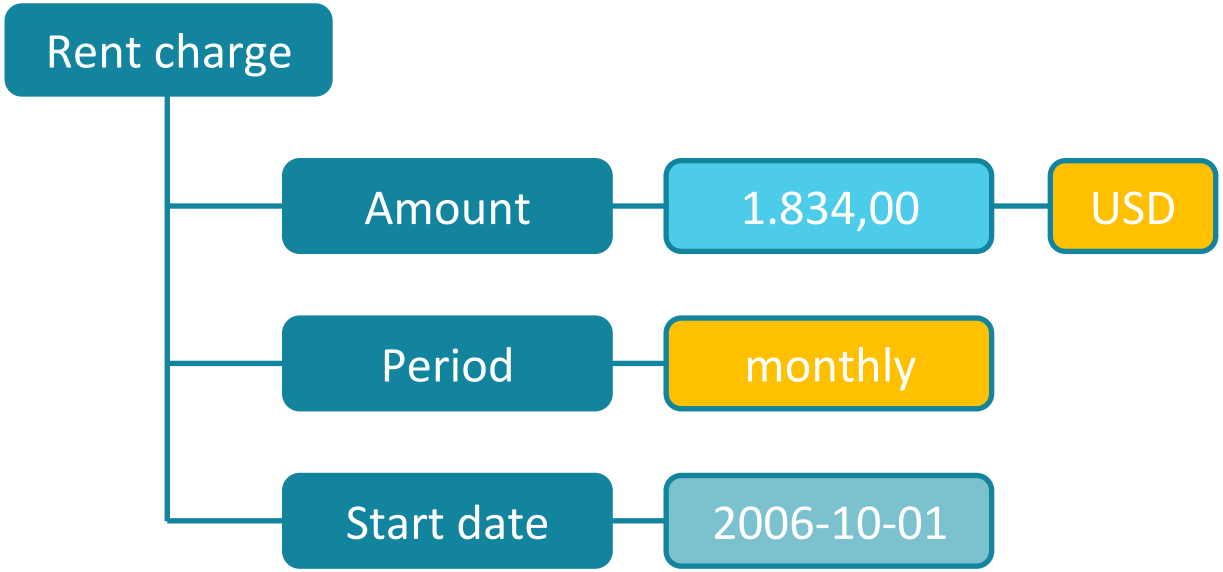


- Example data model for rent charges of a lease -

Information Extraction

Find, classify and extract information

From 1 October 2006 to 31 December 2010: The rent (inclusive of management fees and air-conditioning charges during normal office hours) for the period from the Commencement Date to the Expiration Date of the said term shall be USD ONE THOUSAND EIGHT HUNDRED AND THIRTY FOUR (USD 1,834.00) per calendar month.



Information Extraction

Rule based

Extraction based on hand written rules

1. Apple is bad, if it has brown spots
2. Apple is bad, if it has wrinkles
3. Apple is bad, if it is brown
4. …

Machine & Deep Learning

Learn from examples

Bad apple

Bad apple

Good apple

Rule based Information Extraction

Example: Extract phone or fax number from text

```
Rule: Phone1

        (({Token.string=="Telefon"})
        |
        ({Token.string=="Tel"}({Token.string==".”})?)
        |
        ({Token.string=="Fax"}))
        ({Token.kind==punctuation})?
        ({Token.string=="("})?
        ({Token.numtype==ordinal})[1,3]
        ({Token.string==")"}|{Token.string=="/"})?
        ({Token.numtype==ordinal})[1,6]
        ({Token.string=="-"}{Token.numtype==ordinal})?

:contact

-->

:contact.Phone = {rule = "Phone"}
```

Information Extraction with Deep Learning

Feed annotated examples and wait

Information Extraction

Rule based vs. Deep Learning

Rule based

PROS
- Start without labeled data
- Can be debugged and adjusted easily

CONS
- Rule engineers need domain knowledge
- Rule engineers need language knowledge
- Rules must be adopted to different languages
- Rules can become quite complex and hard to understand
- Rules never cover all possible cases, esp. in complex languages

Machine & Deep Learning

PROS
- Same algorithms for all languages / domains
- No need for engineers to understand the language
- No need for engineers to understand the domain
- Robust, generalizes to different verbalizations
- Fast

CONS
- Labeled data needed
- Cannot be debugged
- Hard to understand and to tune
- Training might take long time…

… So what is the better choice?

Rules vs. Deep Learning @ LEVERTON

What to use?

- Approx. one person year into writing **rules** led to performance (F1 score) of

  → **20%-65% on 20 different data points in 1 language**

- Current **Deep Learning** led to performance (F1 score) of

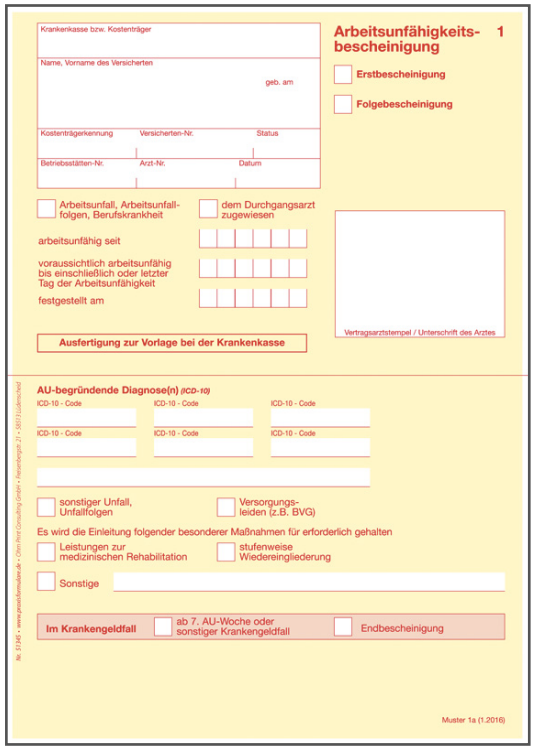
  → **70% to near 100% on 700+ different data points in 4 languages**

  → This sounds good, but how can we get better?

Challenges: Layout

Different layouts might need different extraction strategies

Structured



Krankenkasse bzw. Kostenträger

Name, Vorname des Versicherten

geb. am

Kostenträgerkennung | Versicherten-Nr. | Status

Betriebsstätten-Nr. | Arzt-Nr. | Datum

Arbeitsunfähigkeits-bescheinigung 1

☐ Erstbescheinigung

☐ Folgebescheinigung

☐ Arbeitsunfall, Arbeitsunfall-folgen, Berufskrankheit

☐ dem Durchgangsarzt zugewiesen

arbeitsunfähig seit

voraussichtlich arbeitsunfähig bis einschließlich oder letzter Tag der Arbeitsunfähigkeit

festgestellt am

Vertragsarztstempel / Unterschrift des Arztes

Ausfertigung zur Vorlage bei der Krankenkasse

AU-begründende Diagnose(n) *(ICD-10)*

ICD-10 - Code | ICD-10 - Code | ICD-10 - Code

ICD-10 - Code | ICD-10 - Code | ICD-10 - Code

☐ sonstiger Unfall, Unfallfolgen

☐ Versorgungs-leiden (z.B. BVG)

Es wird die Einleitung folgender besonderer Maßnahmen für erforderlich gehalten

☐ Leistungen zur medizinischen Rehabilitation

☐ stufenweise Wiedereingliederung

☐ Sonstige

Im Krankengeldfall ☐ ab 7. AU-Woche oder sonstiger Krankengeldfall ☐ Endbescheinigung

Muster 1a (1.2016)

Semi structured

Maxi Musterfrau
Nebenstraße 1
12345 Musterhausen

Rechnungsnummer
012010

Datum
12. November 2010

RECHNUNG

Sehr geehrte Frau Musterfrau,

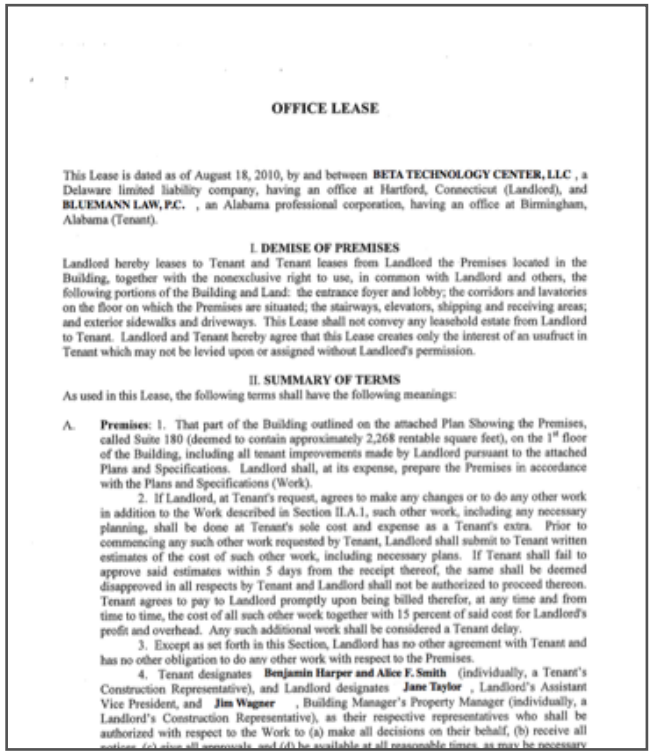
Bitte überweisen Sie den folgenden Rechnungsbeitrag innerhalb von 14 Tagen auf das unten angegebene Konto.

Menge	Einheit	Beschreibung	Einzelpreis (netto)	Gesamtpreis (netto)
10	Std.	Musterbrief erstellt	30,00 EUR	300,00 EUR
Summe ohne MwSt.				300,00 EUR
MwSt. 19%				57,00 EUR
Gesamtsumme inkl. MwSt.				**357,00 EUR**

Mit freundlichen Grüßen,

Max MUSTERMANN

Unstructured



OFFICE LEASE

This Lease is dated as of August 18, 2010, by and between **BETA TECHNOLOGY CENTER, LLC** , a Delaware limited liability company, having an office at Hartford, Connecticut (Landlord), and **BLUEMANN LAW, P.C.** , an Alabama professional corporation, having an office at Birmingham, Alabama (Tenant).

I. DEMISE OF PREMISES

Landlord hereby leases to Tenant and Tenant leases from Landlord the Premises located in the Building, together with the nonexclusive right to use, in common with Landlord and others, the following portions of the Building and Land: the entrance foyer and lobby; the corridors and lavatories on the floor on which the Premises are situated; the stairways, elevators, shipping and receiving areas; and exterior sidewalks and driveways. This Lease shall not convey any leasehold estate from Landlord to Tenant. Landlord and Tenant hereby agree that this Lease creates only the interest of an usufruct in Tenant which may not be levied upon or assigned without Landlord's permission.

II. SUMMARY OF TERMS

As used in this Lease, the following terms shall have the following meanings:

A. **Premises**: 1. That part of the Building outlined on the attached Plan Showing the Premises, called Suite 180 (deemed to contain approximately 2,268 rentable square feet), on the 1st floor of the Building, including all tenant improvements made by Landlord pursuant to the attached Plans and Specifications. Landlord shall, at its expense, prepare the Premises in accordance with the Plans and Specifications (Work).

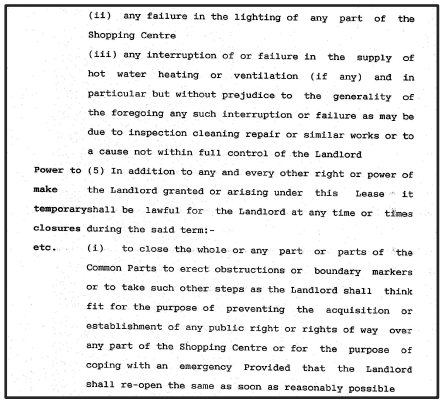
2. If Landlord, at Tenant's request, agrees to make any changes or to do any other work in addition to the Work described in Section II.A.1, such other work, including any necessary planning, shall be done at Tenant's sole cost and expense as a Tenant's extra. Prior to commencing any such other work requested by Tenant, Landlord shall submit to Tenant written estimates of the cost of such other work, including necessary plans. If Tenant shall fail to approve said estimates within 5 days from the receipt thereof, the same shall be deemed disapproved in all respects by Tenant and Landlord shall not be authorized to proceed thereon. Tenant agrees to pay to Landlord promptly upon being billed therefor, at any time and from time to time, the cost of all such other work together with 15 percent of said cost for Landlord's profit and overhead. Any such additional work shall be considered a Tenant delay.

3. Except as set forth in this Section, Landlord has no other agreement with Tenant and has no other obligation to do any other work with respect to the Premises.

4. Tenant designates **Benjamin Harper and Alice F. Smith** (individually, a Tenant's Construction Representative), and Landlord designates **Jane Taylor** , Landlord's Assistant Vice President, and **Jim Wagner** , Building Manager's Property Manager (individually, a Landlord's Construction Representative), as their respective representatives who shall be authorized with respect to the Work to (a) make all decisions on their behalf, (b) receive all

Challenge in Layout:

Determine the correct reading order



(ii) any failure in the lighting of any part of the Shopping Centre

(iii) any interruption of or failure in the supply of hot water heating or ventilation (if any) and in particular but without prejudice to the generality of the foregoing any such interruption or failure as may be due to inspection cleaning repair or similar works or to a cause not within full control of the Landlord

Power to make temporary closures etc.

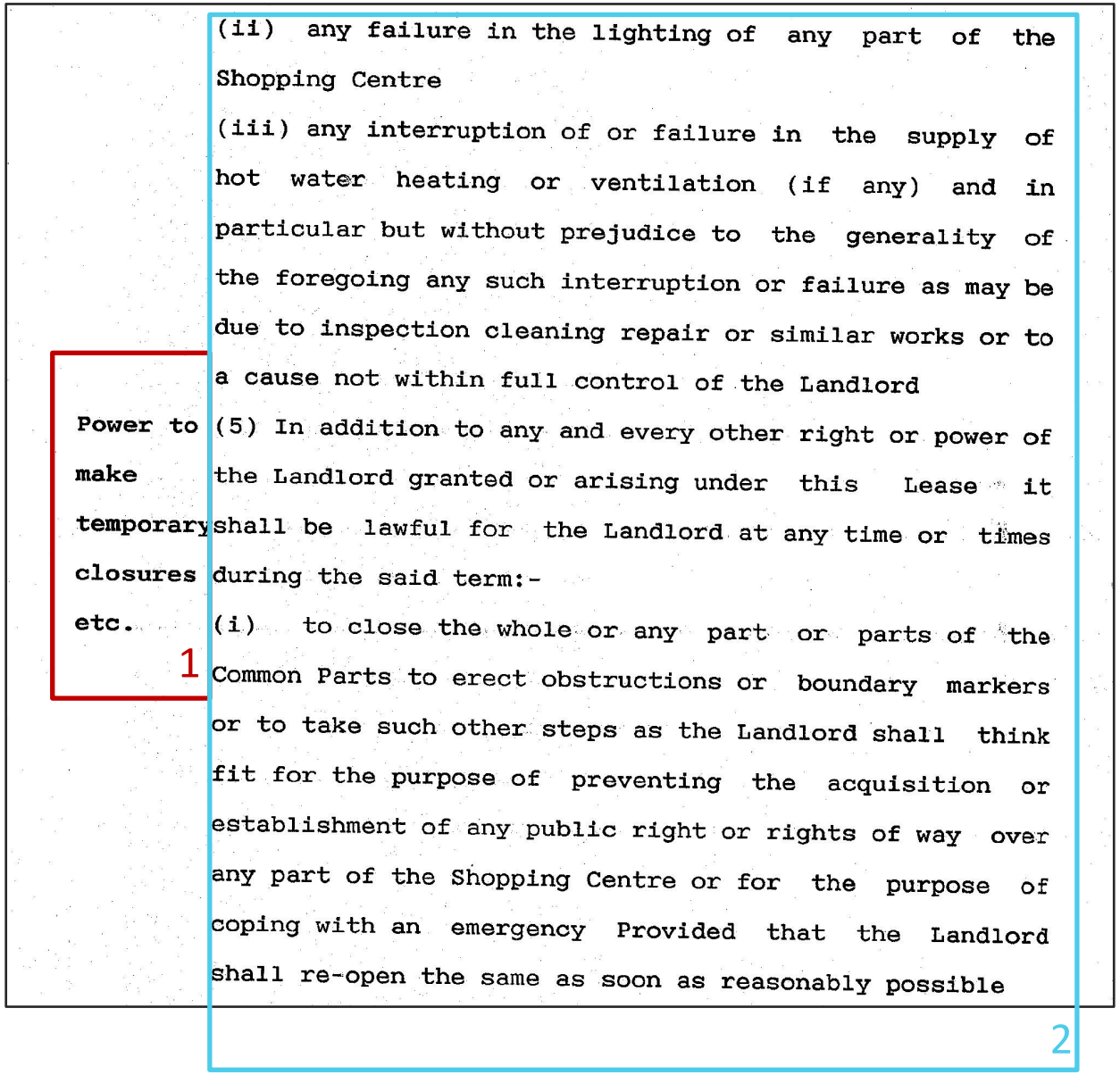
(5) In addition to any and every other right or power of the Landlord granted or arising under this Lease it shall be lawful for the Landlord at any time or times during the said term:-

(i) to close the whole or any part or parts of the Common Parts to erect obstructions or boundary markers or to take such other steps as the Landlord shall think fit for the purpose of preventing the acquisition or establishment of any public right or rights of way over any part of the Shopping Centre or for the purpose of coping with an emergency Provided that the Landlord shall re-open the same as soon as reasonably possible

- **Visual features alone often not sufficient!**
- **Impossible to determine the reading order without reading the text. We need to guess.**

Layout

Determine the correct reading order

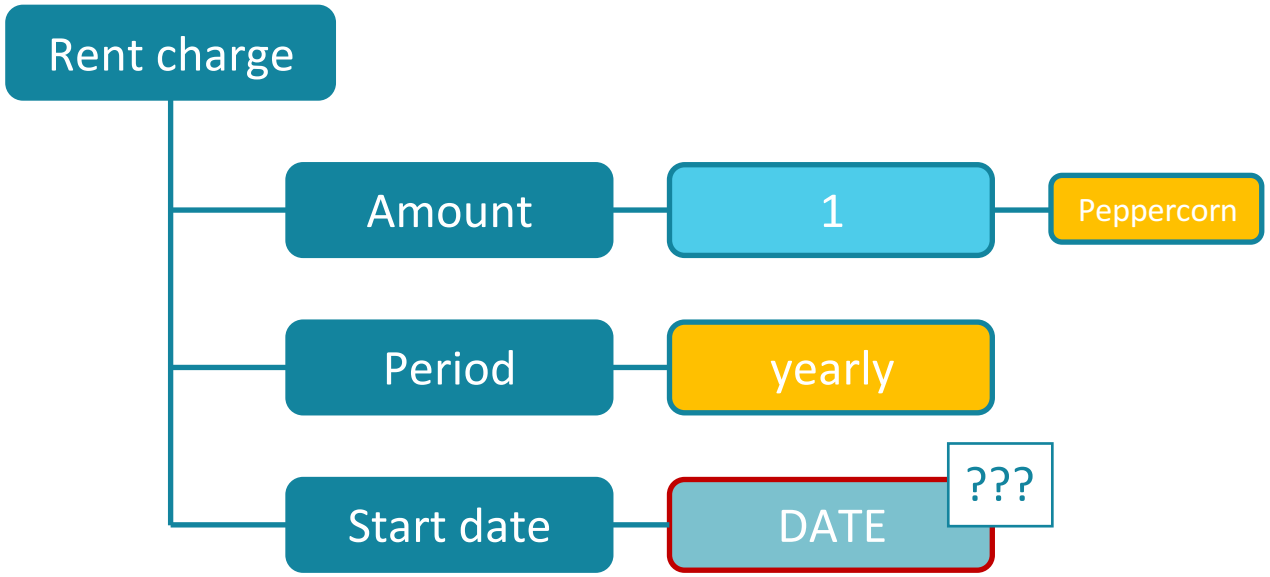


- Combination of visual features and interpretation of extracted text leads to correct reading order
- Human brains are able to combine multiple steps (Visual separation, recognizing of characters and words, interpreting words to form sentences, which leads to correct separation / layout recognition)
- Deep Learning is focused on one problem at a time
- Joint approaches necessary to solve such problems

Challenges: Data model

Abstraction comes with a price

surplus earth PROVIDED that in so doing the Company shall cause as little damage as may be to such piece of land and shall make good and restore the surface thereof TO HOLD the same unto the Company for the term of **ninety-nine years** from and including the date hereof YIELDING AND PAYING therefor yearly during the said term in advance the rent of **one peppercorn** (if demanded) on each anniversary of the date of commencement of the term created by this Lease



More challenges ahead…

How can we improve in the future

- **Co-Referencing** (which information belongs to what)
- Scan quality
- **Scaling** out: Training already takes 6 weeks of one core
- …

… But we are on a good track

Achievements of LEVERTON AI

- **Simple documents: >95% automation** for >50 different data points
- **Complex documents: >70% automation** for >200 different data points
- Available in >20 different languages
- Build **own OCR engine** in < 1 year, competitive to all known OCR engines, more robust on layout

Example Use Case: IFRS16

Changes in balancing standards forces corporates to revisit leasing data

Global consolidation

Extract data

IFRS 16 DATA

- Options
- Valuation
- Indexation
- Payment types

System Integration

LEVERTON can be integrated into ERP systems at customers

Real estate leases

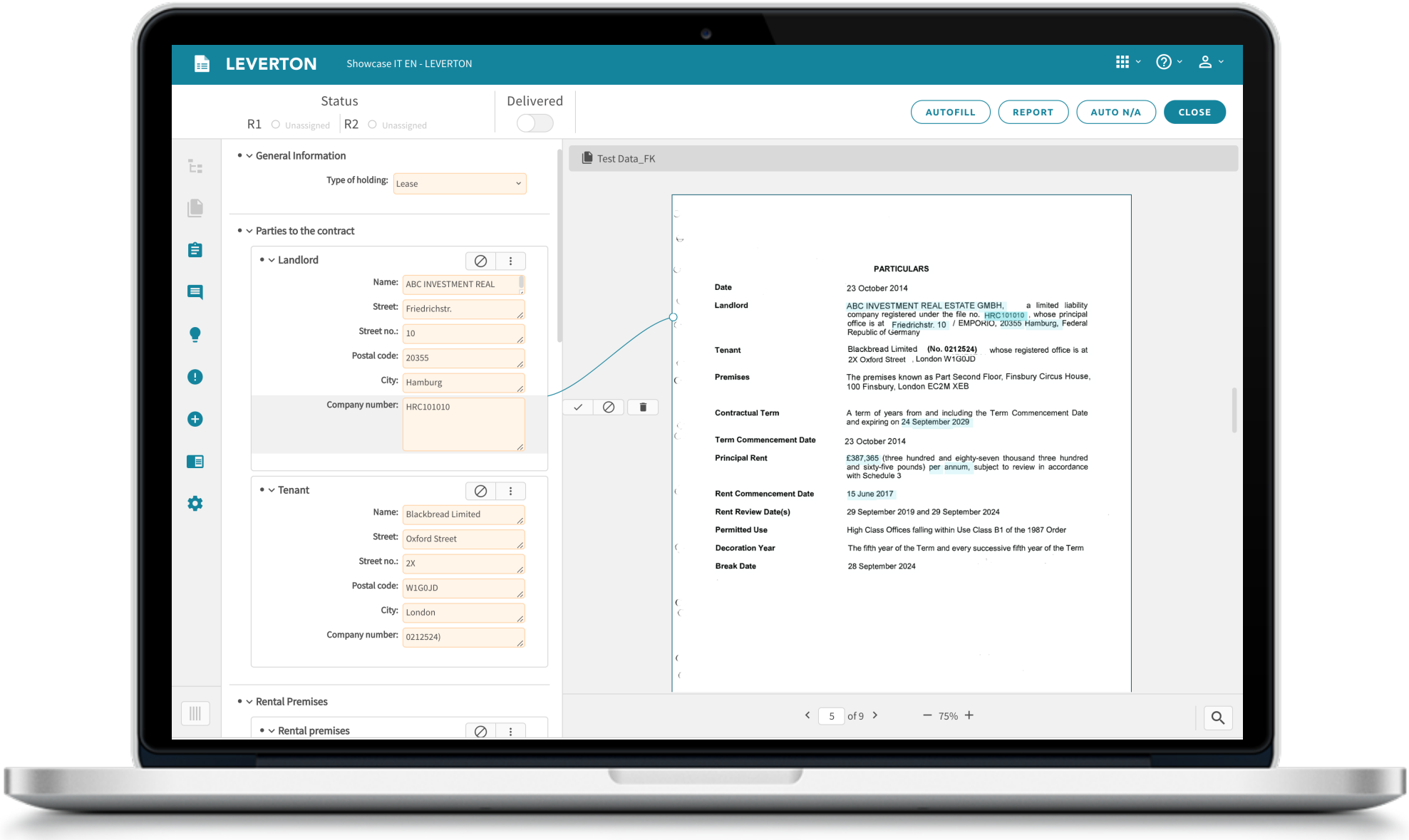
Machinery leases

Car leases
& more

→ **LEVERTON** ↔ SAP

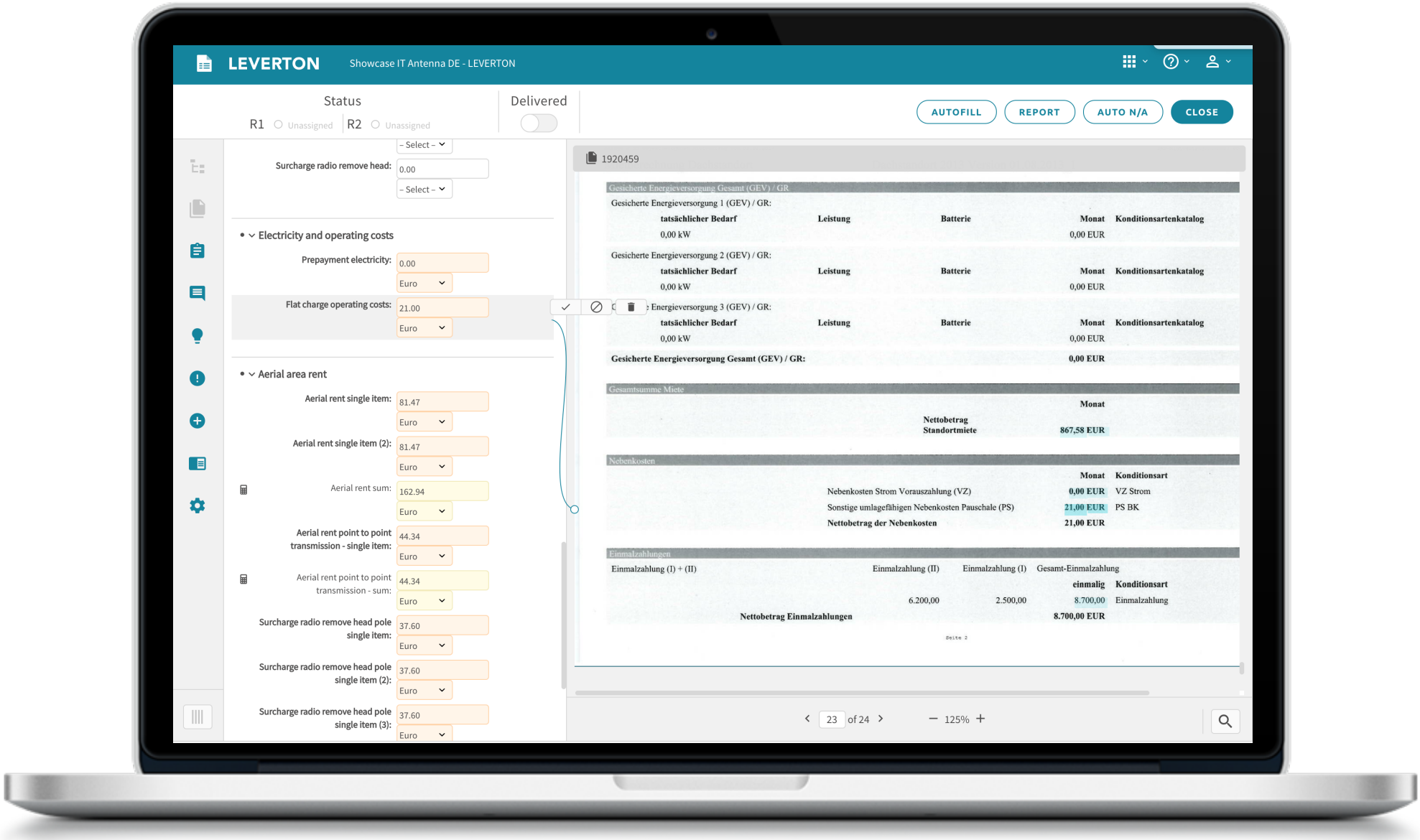
LEVERTON CORE

Example complex document >70% automation



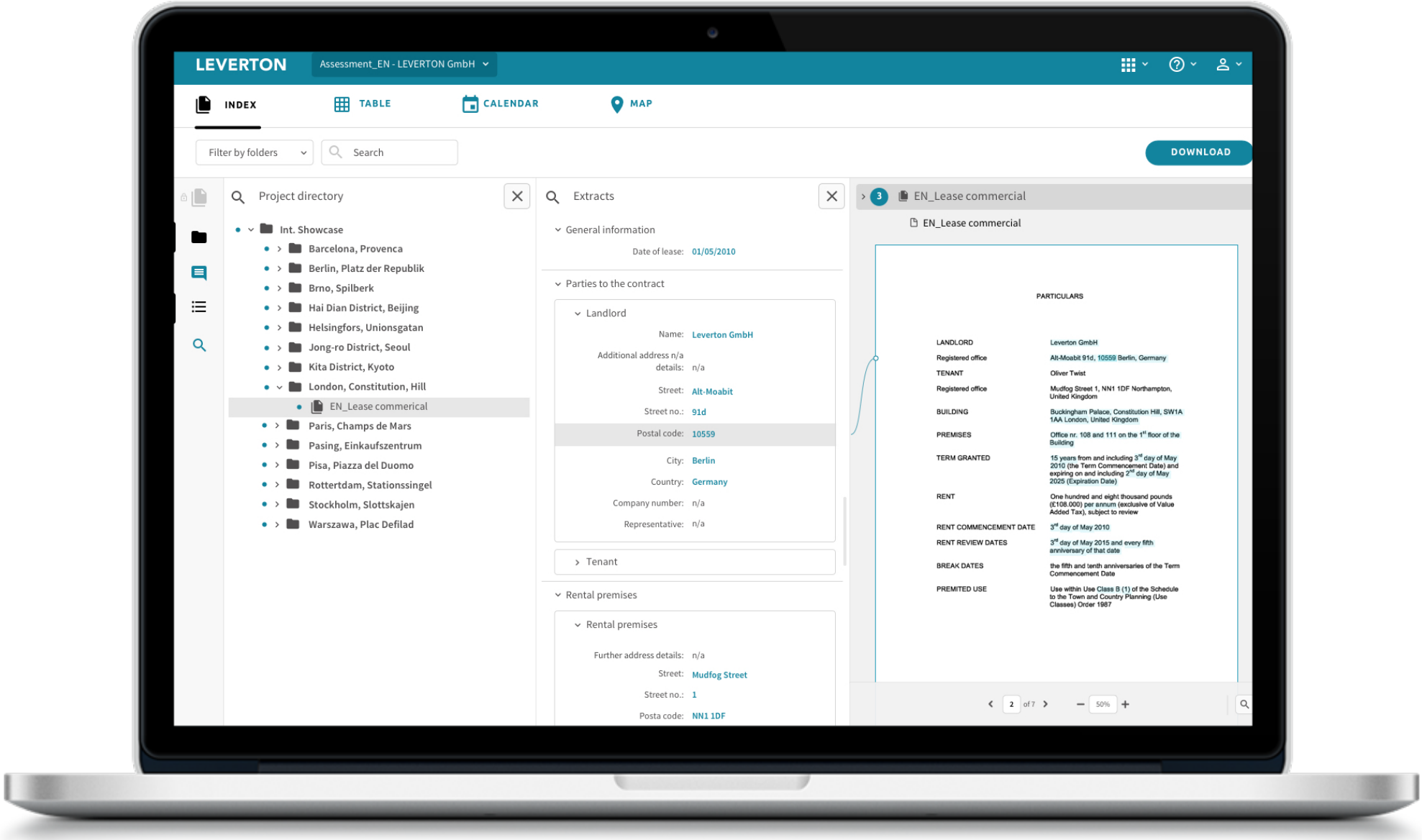
LEVERTON CORE

Example simple document, >95% automation



LEVERTON CORE

Enables decisions using data based on legally binding documents



Recap

How AI Revolutionizes Data & Document Management

- Breakthrough with switch to **Deep Learning**
- More **challenges** for **100% automated** information extraction on **complex docs**
- To achieve this, we probably need to **combine multiple steps** into one large network
- We'll need lots of **computational resources** to do so
- If you have ideas about any of these: **We are hiring**.

LEVERTON

Florian Kuhlmann
CTO & Co-founder
florian.kuhlmann@leverton.ai