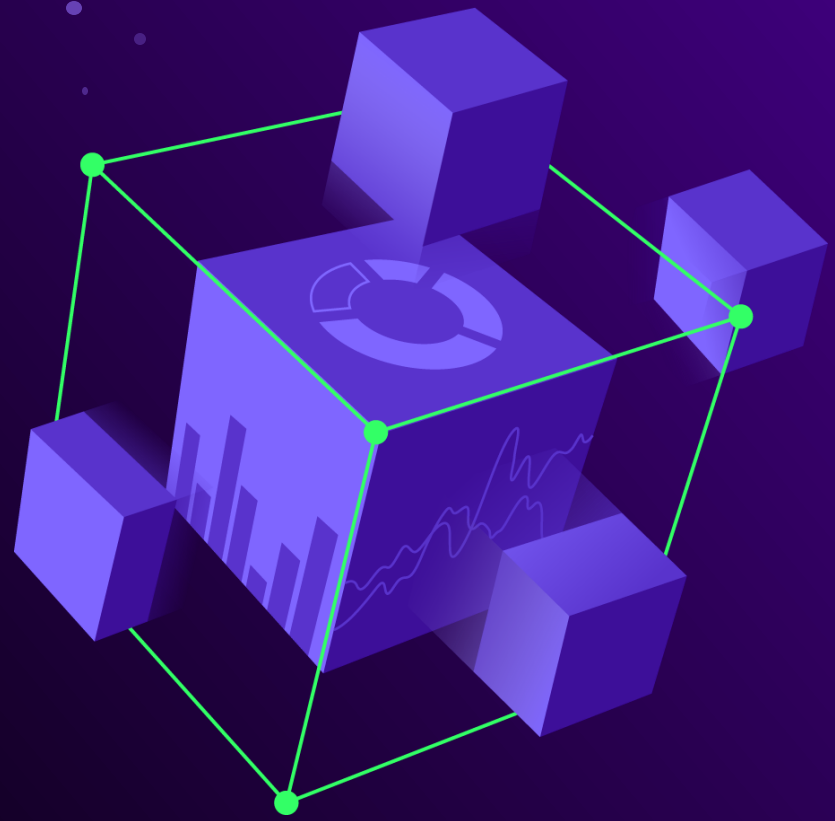
DATA SCIENCE SUMMIT



NLP for automatic analysis of contracts

Nikolai Karelin, nikolai.karelin@silkdata.ai
Head of AI, Silk Data

www.dssconf.pl 03 grudnia 2020 PGE Narodowy + Online

ORGANIZATORZY: Academic Partners Fundacja Data Science Warsaw Wydział Matematyki i Nauk Informacyjnych Politechnika Warszawska

Outline

- Topic and team
- Contract classification
- Clause classification
- User interface
  - Web
  - MS Word extension
- Conclusion, further steps

The Problem: Level 3 Contract Review

Costly and slow process

- USD 500+/hour

Prone to mistakes

- Junior attorneys
- High time pressure

Boring work

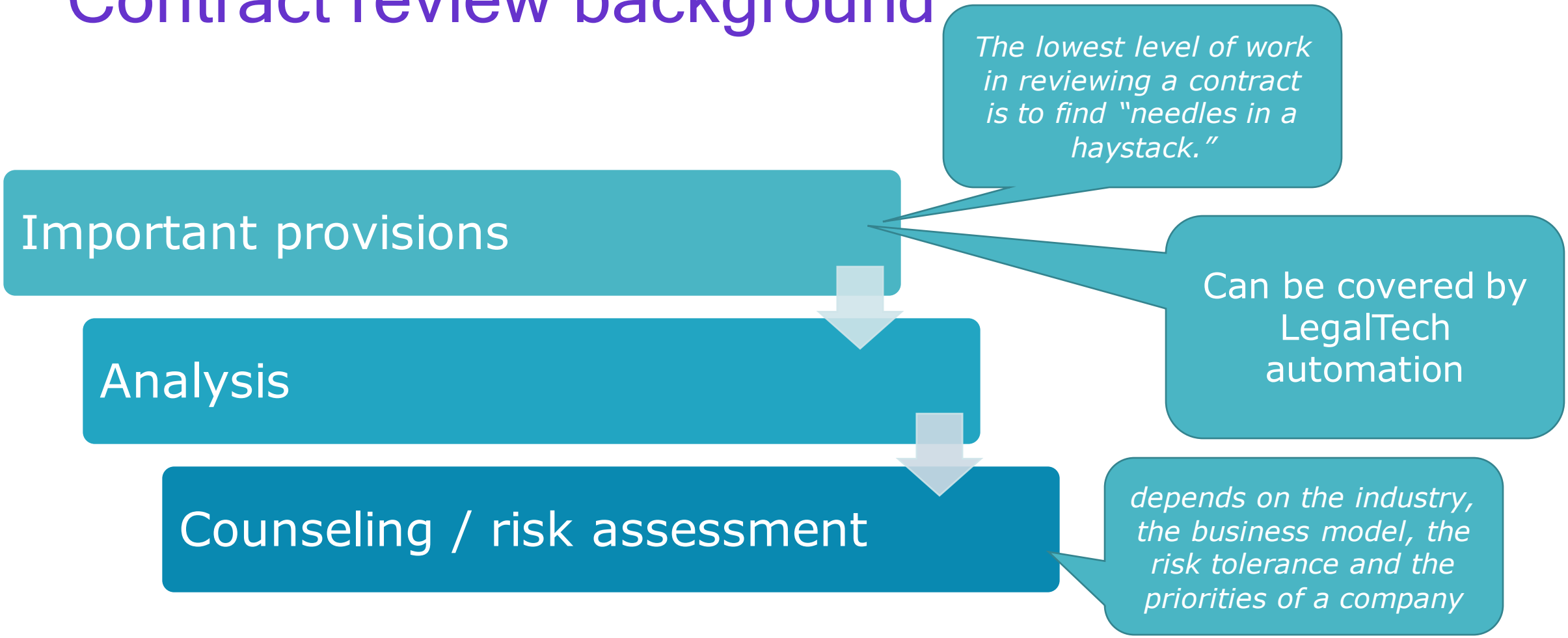
- Low motivation
- Hard employee retention

*"… contracts review, which **costs** humans substantial time, money, and attention. Many law firms spend approximately 50% of their time reviewing contracts. Due to the specialized training necessary…, the billing rates for lawyers at large law firms are typically around **$500-$900 per hour in the US**.*

*As a result, many transactions cost companies **hundreds of thousands of dollars** just so that lawyers can verify that there are no problematic obligations or requirements included in the contracts. Contract review can be a source of drudgery and, in comparison to other legal tasks, is widely considered to be **especially boring**."*

Source here and below: https://arxiv.org/abs/2103.06268

Contract review background



The team

ЮрСпектр
Правовые информационные технологии
ilex PODPIS.BY

silk data

Expertise

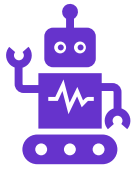
- iLex.by – the biggest legal info service provider in BY
  - >100K legal and analytical documents
  - Human supervision for AI

Advisory

Sergey Sinkevich
Sr. Director @ EPAM
Business Angel @ AngelsBand.by

Technology

Silk Data – software engineering company, focused on NLP, AI consulting & development, Machine Learning Engineering, Data Science
https://www.linkedin.com/company/silkdata/



AI consulting & development

Machine learning engineering

Data Science & Analytics

Web development

Mobile development

IT consulting & prototyping

Goals

- Create automated paralegal ('spell-checker' for legal expert)
- Fast onboarding
  - Ready solution for small/medium-sized business
- Flexibility
  - Support of other languages and legislatures

Contract classification

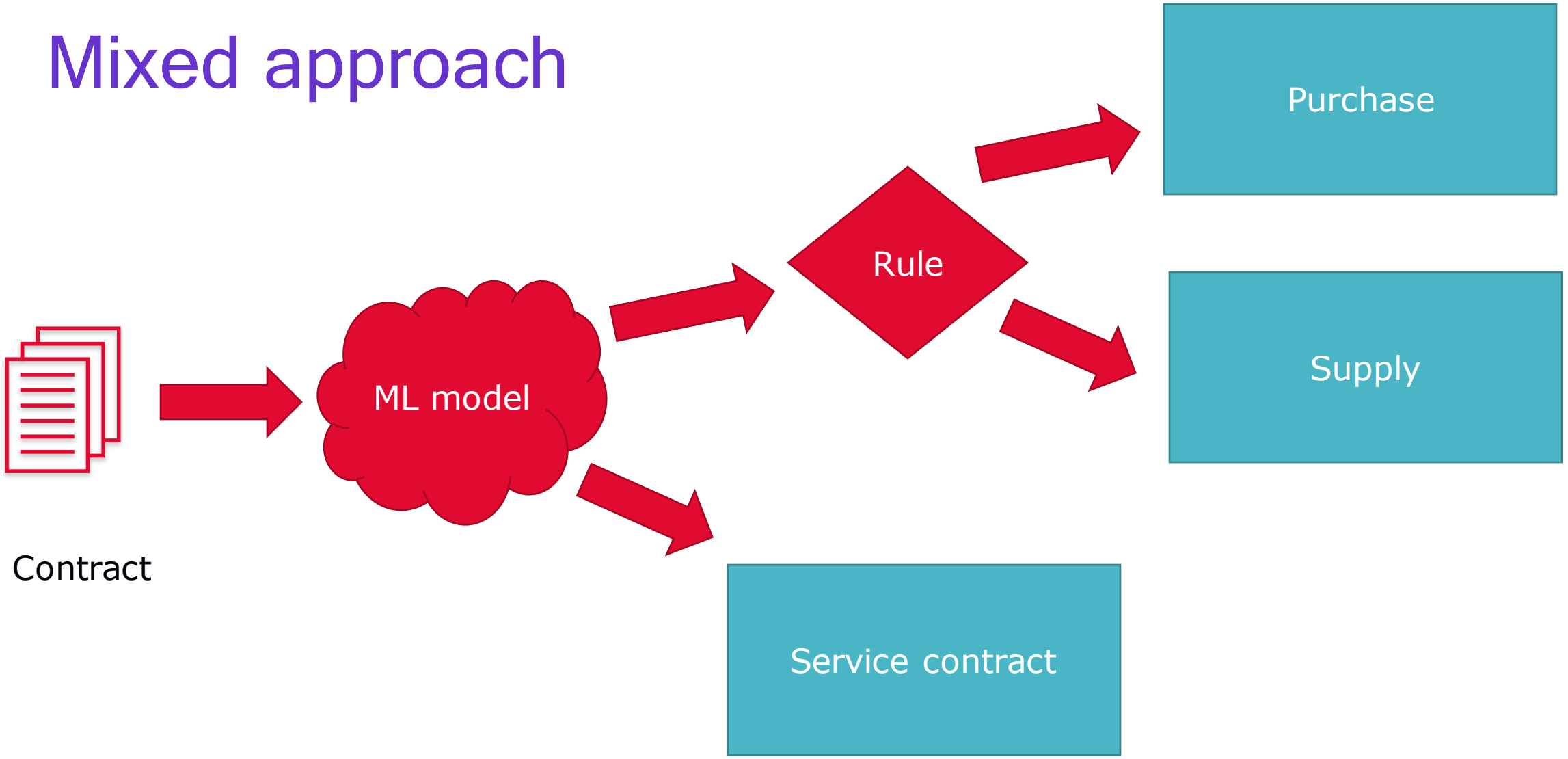
- Why?

- The contract type
  - defines applicable law provisions, or
  - contract negotiation playbook use

- Approach: mixed
  - ML model
  - Rules

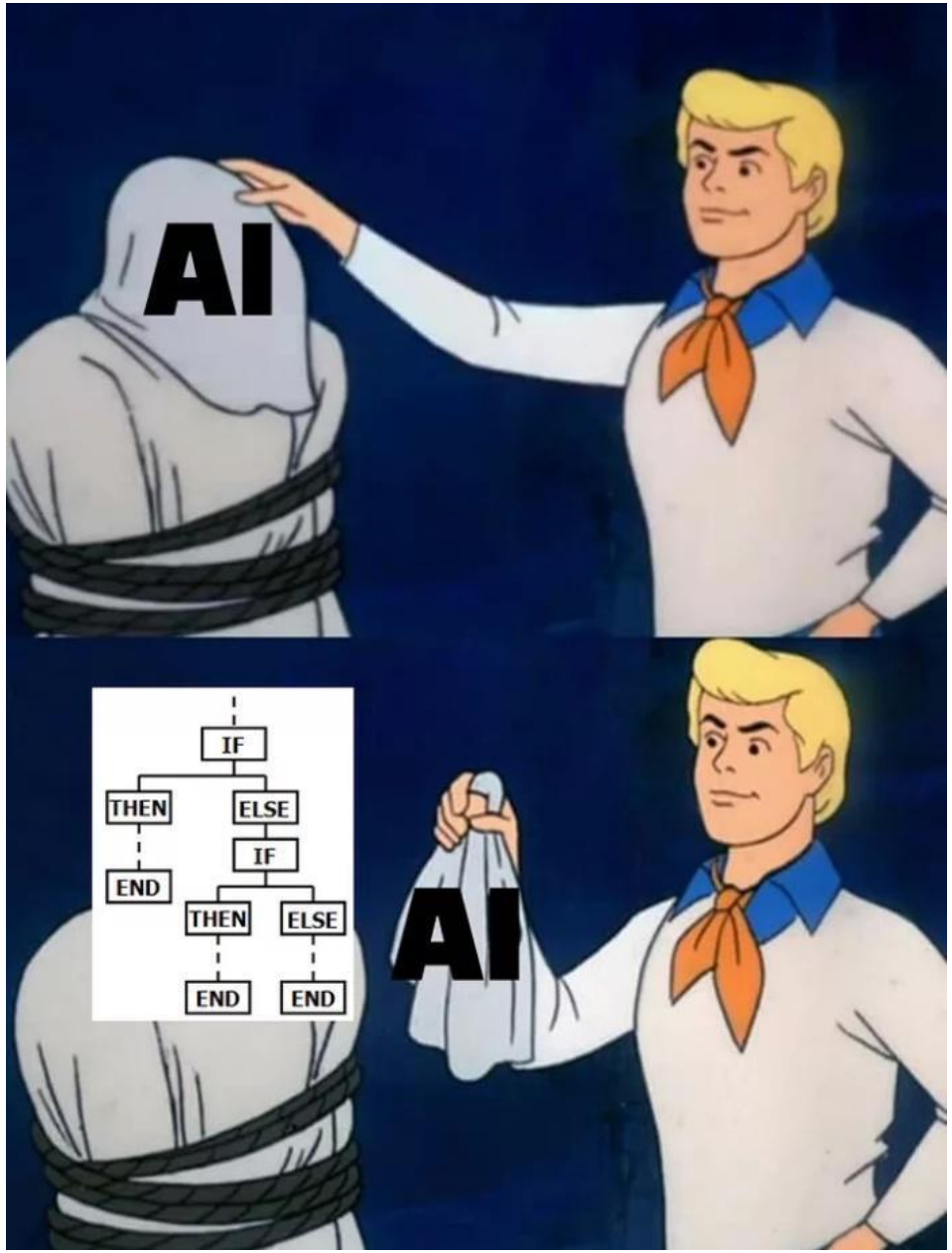
- Supported types:
  - Purchase contracts
  - Consignment (supply) contracts
  - Service contracts

Mixed approach



Classification accuracy: ~90%

Mixed approach



Source: https://www.linkedin.com/feed/update/urn:li:activity:6537313774251106304

Data (EN): Atticus project

- A corpus of 13,000+ labels in 510 commercial legal contracts
- Manually labeled under the supervision of experienced lawyers to identify 41 types of legal clauses (considered important in contract review in connection with a corporate transaction, including mergers & acquisitions, corporate finance,…)
- 25 contract types
- https://www.atticusprojectai.org/ (non-profit organization)
- Paper: https://arxiv.org/abs/2103.06268
- Code: https://github.com/TheAtticusProject/cuad

Data (EN): Atticus project

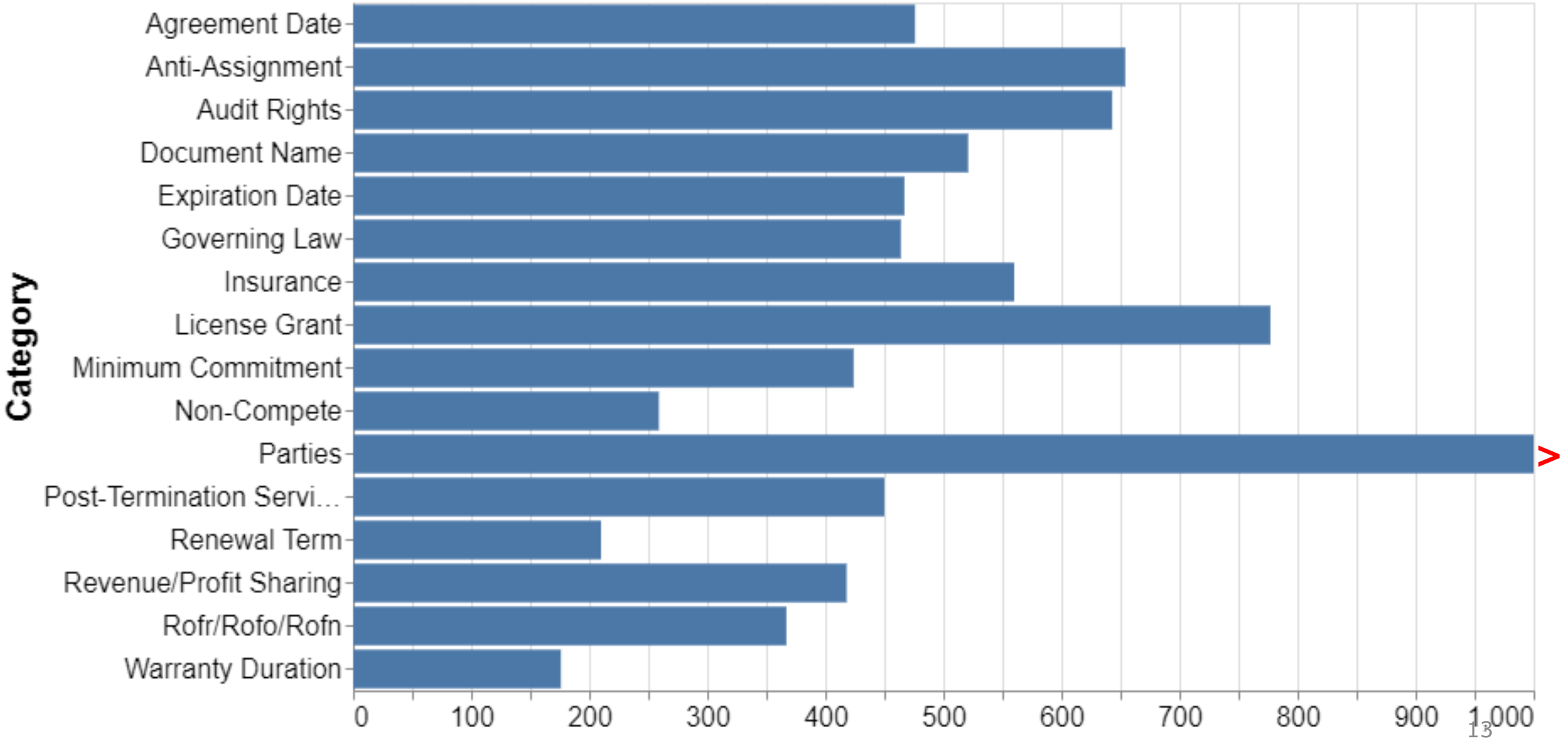
- "As a result of this effort, a conservative estimate of the pecuniary value of CUAD of is over **$2 million** (each of the **9283 pages** were reviewed at least **4 times**, each page requiring **5-10 minutes**, assuming a rate of **$500 per hour**)."
- License: CC-BY 4.0, https://creativecommons.org/licenses/by/4.0/ (attribution required)

Data (EN)

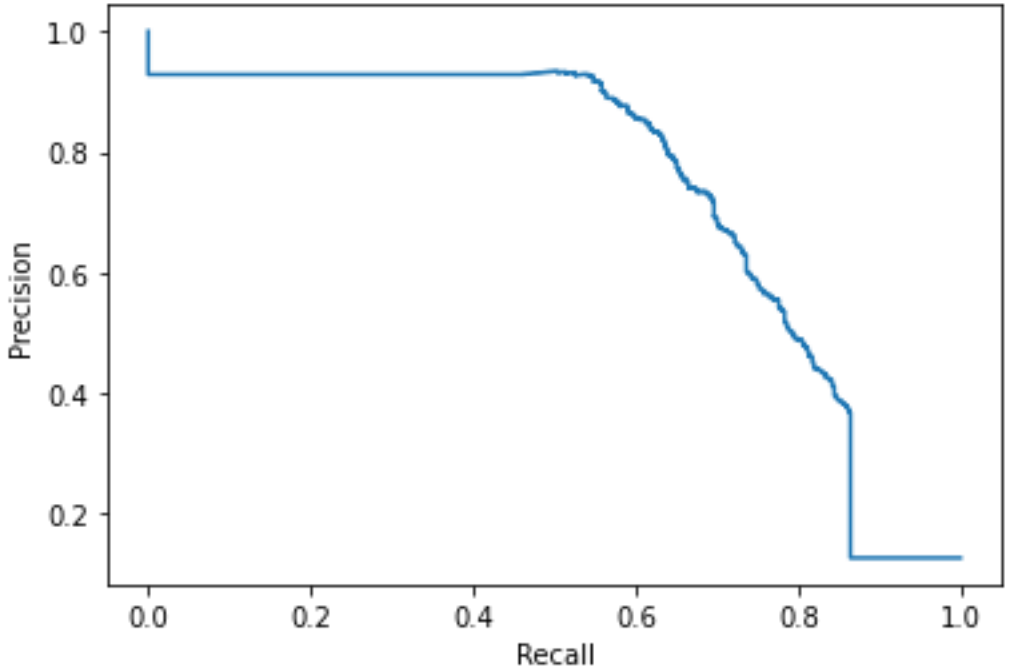
- 500 labeled contracts
  - 'Needle in a haystack' problem: about 0.25% clauses per category

- Unlabeled legal texts:
  - EU legislation, https://eur-lex.europa.eu/
  - UK legislation, https://legislation.gov.uk/
  - US case law (free part), https://case.law/bulk/download/

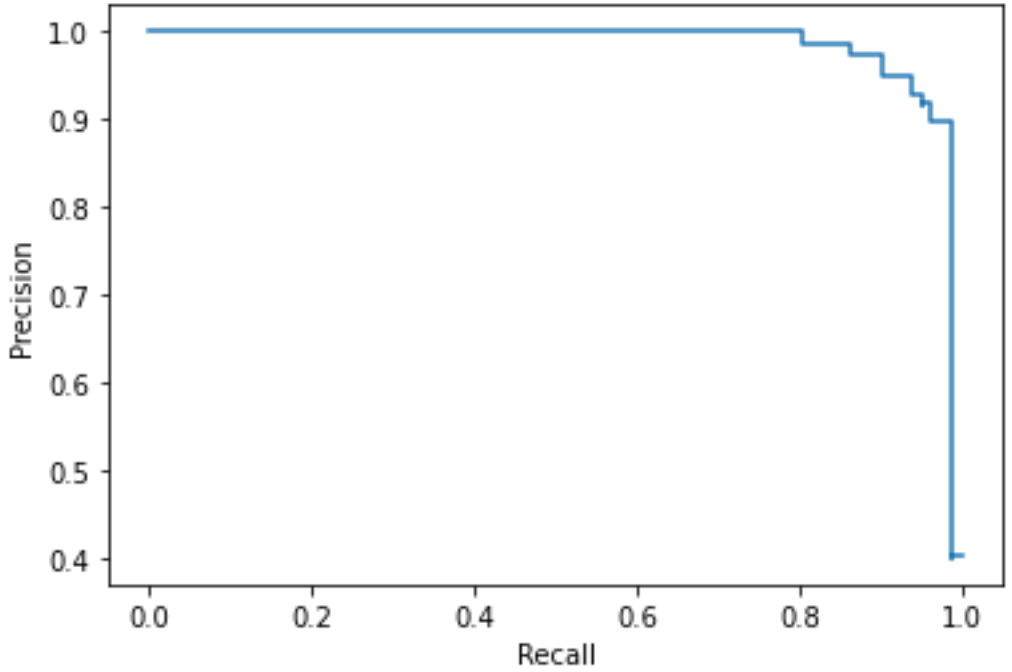
Provisions categories: statistics



Provision classification quality



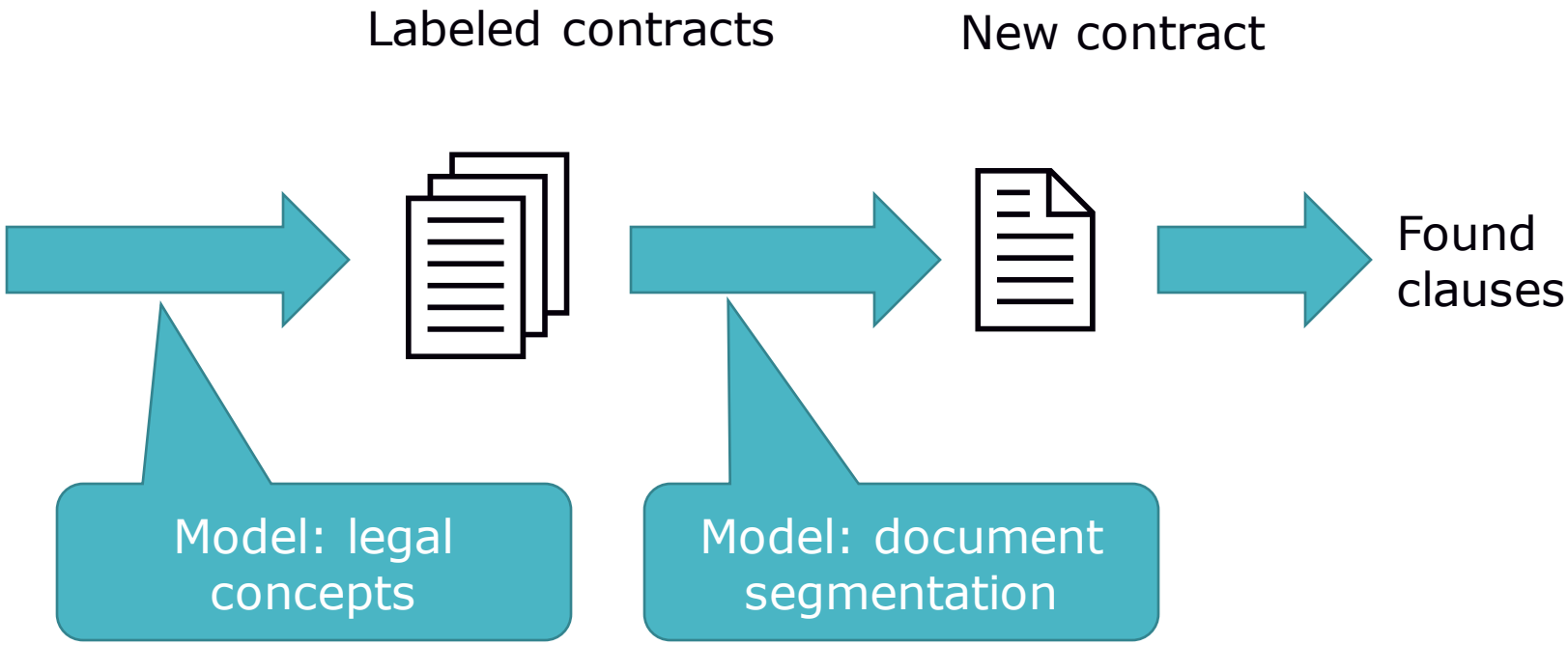
- Precision/recall curve for important clause identification
- Averaged F1=0.83



- Precision/recall curve for clause classification
- Averaged F1=0.82

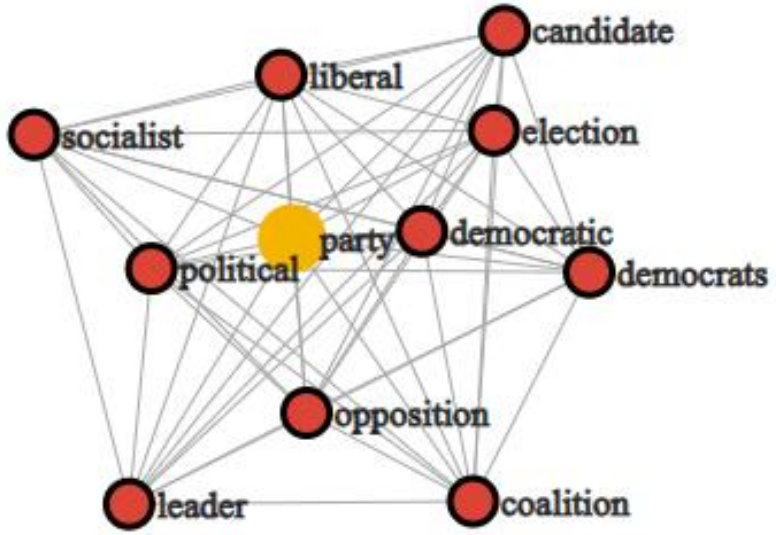
Principle: simplest ‘transfer learning’

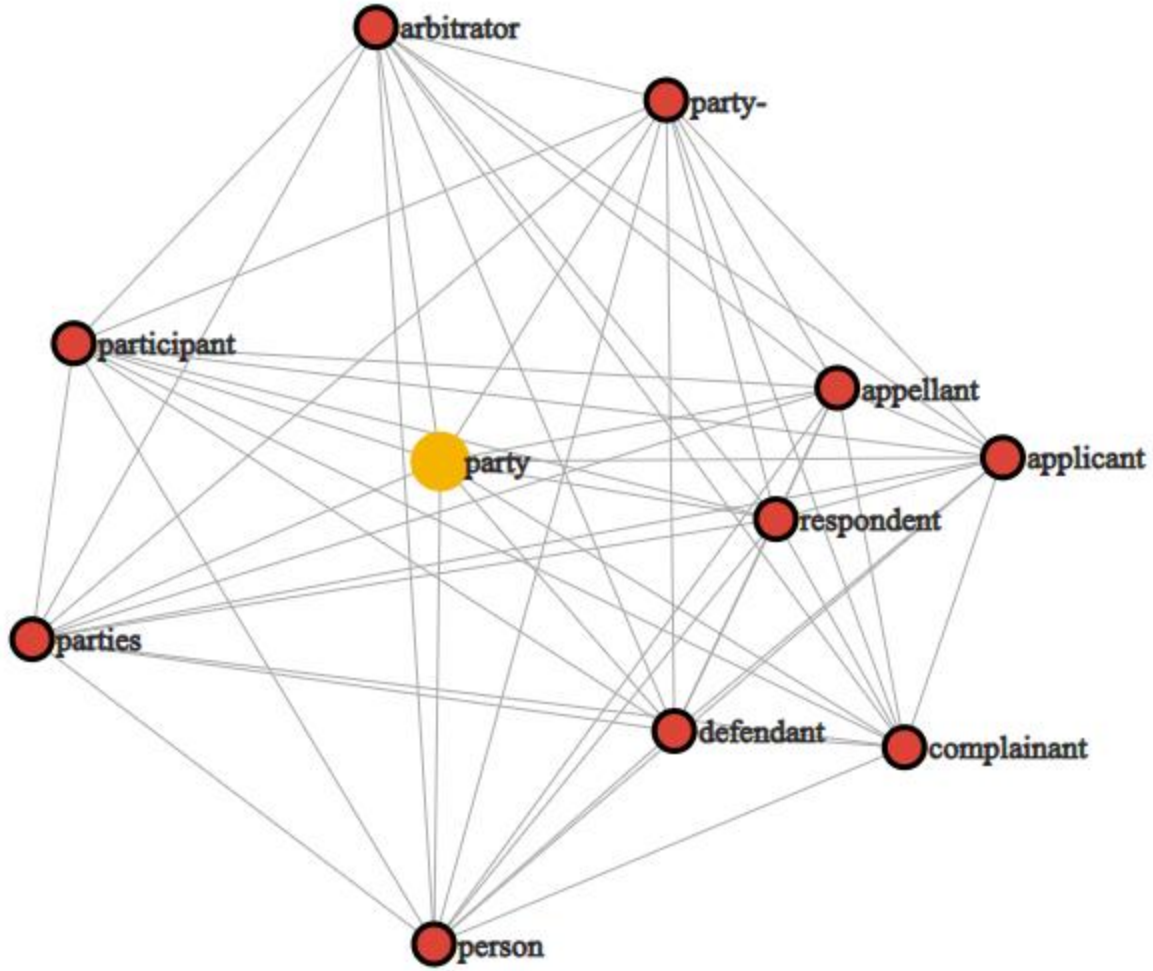
Large collection of legal texts (laws, court cases, books, …)



Structure of concepts

Wikipedia-based





Legal data-based

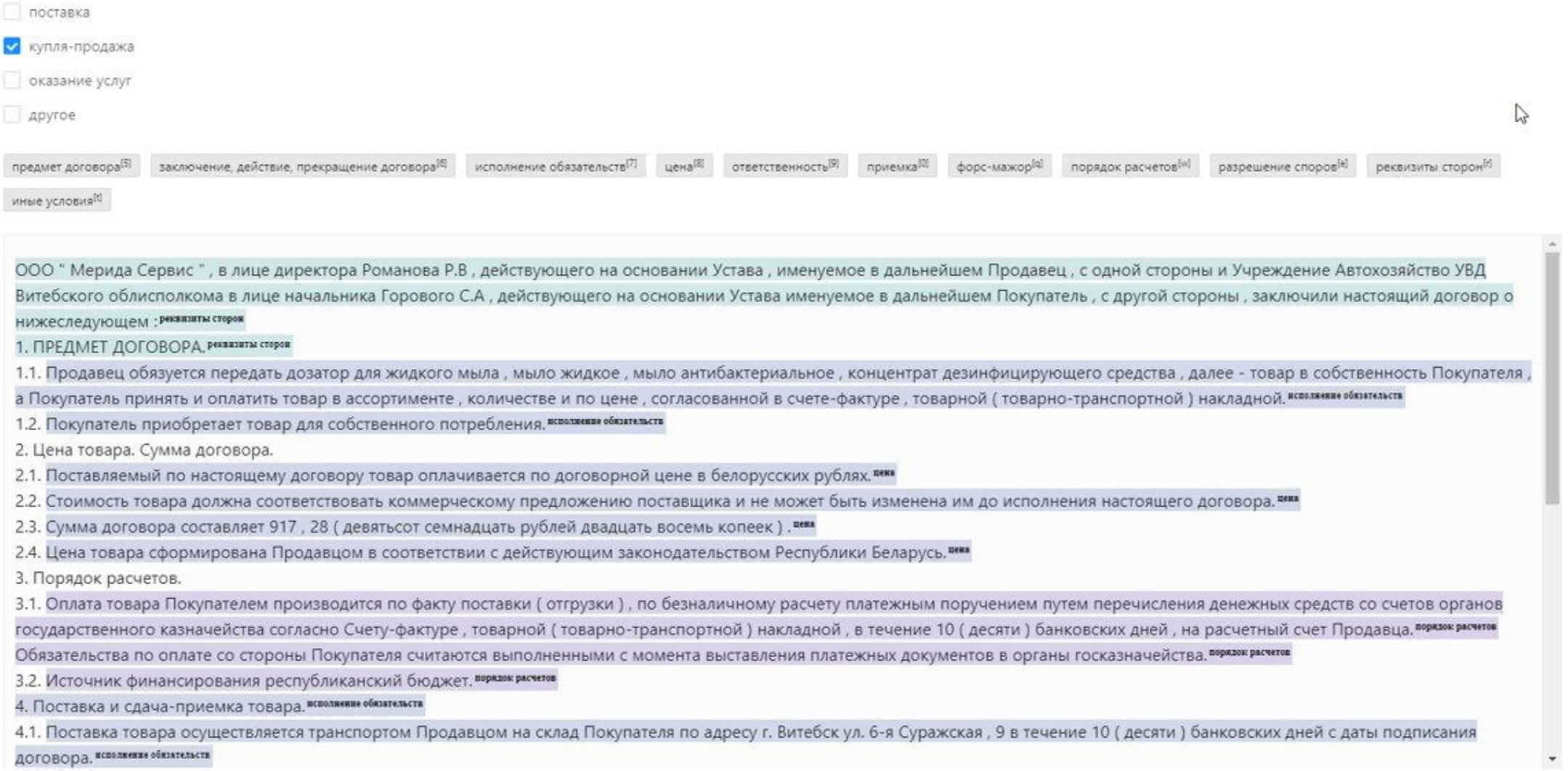
Data (RU)

- Large non-labeled legal corpus
- Contracts for classification, ca. 2500
- About 100 contracts with lawyer-labeled provisions (shorter that English variants)

- Accuracy: ~70-80%

- Further data collection:
  - User's documents
  - Active learning

Active learning



Why classical approaches

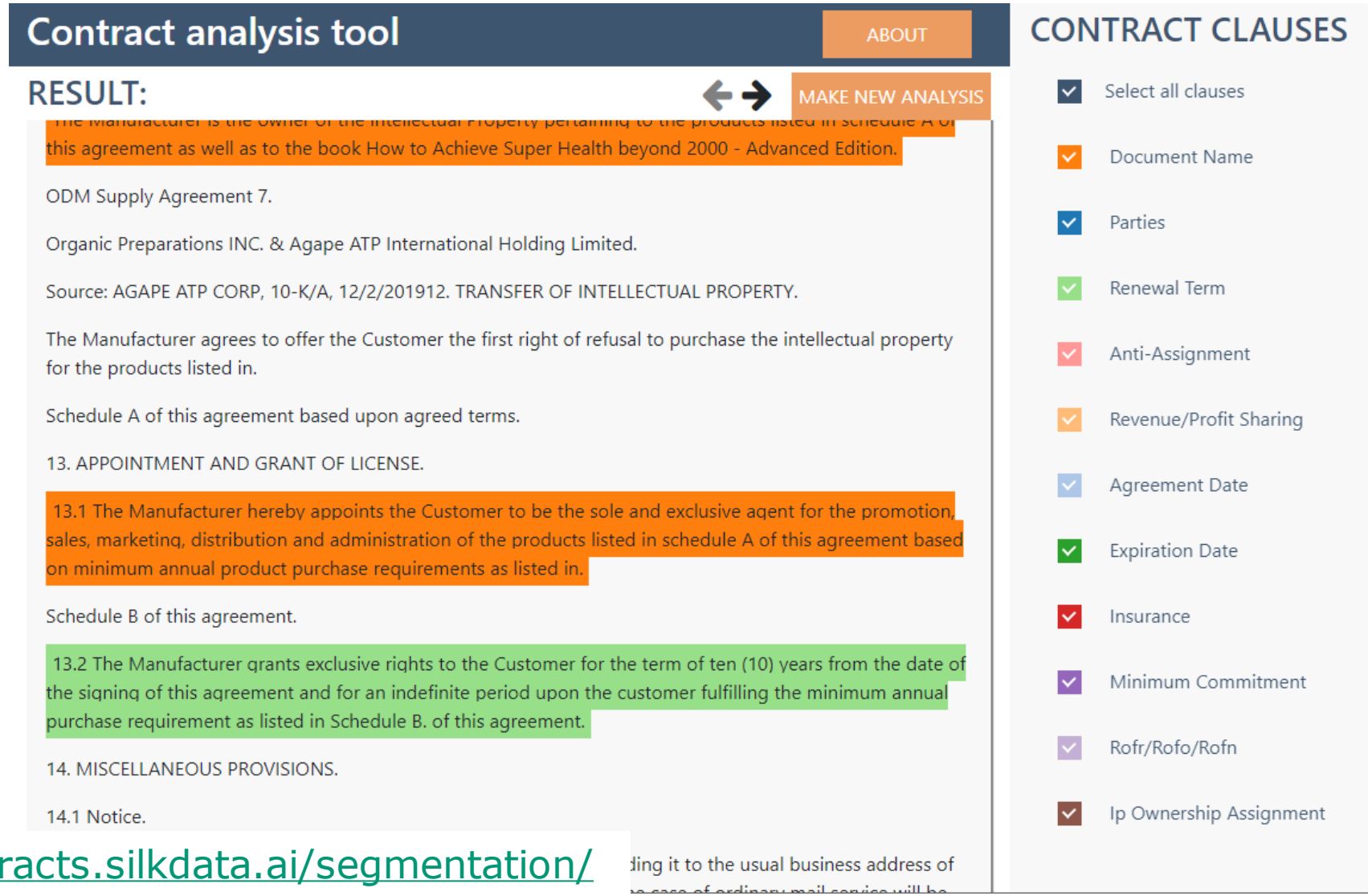
- Performance is good enough
- Easy interpretation
  - Asked by many customers
- Works for limited-resource languages
- Flexibility

MVP implementation

- Web interface
  - Easiest start
  - Text copy-paste
  - File processing (DOCX, PDF)

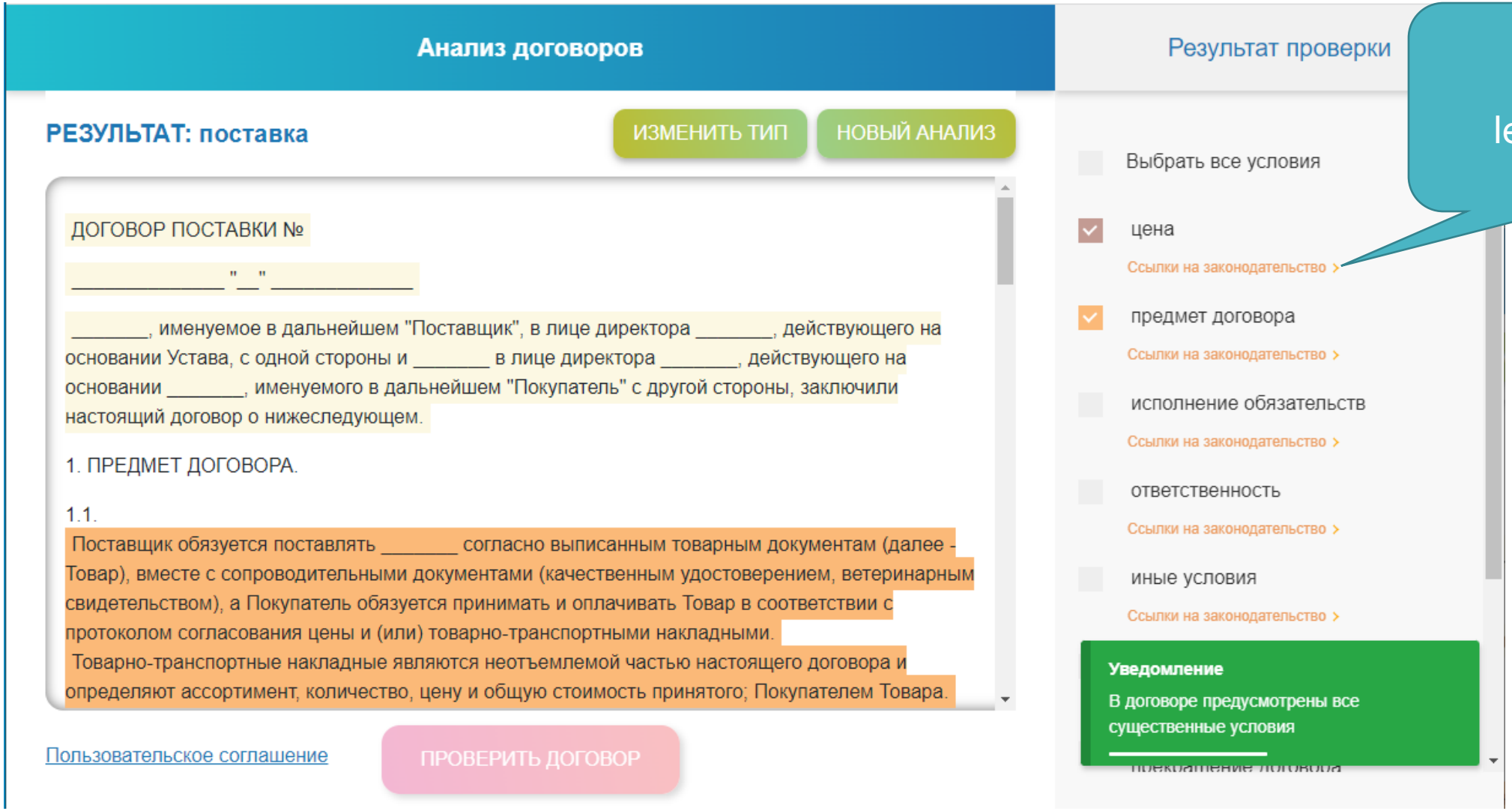
- Word add-on (in development)
  - Where the users work on documents
  - Minimal intrusion: text selection only

Contract Analysis (EN): web-version



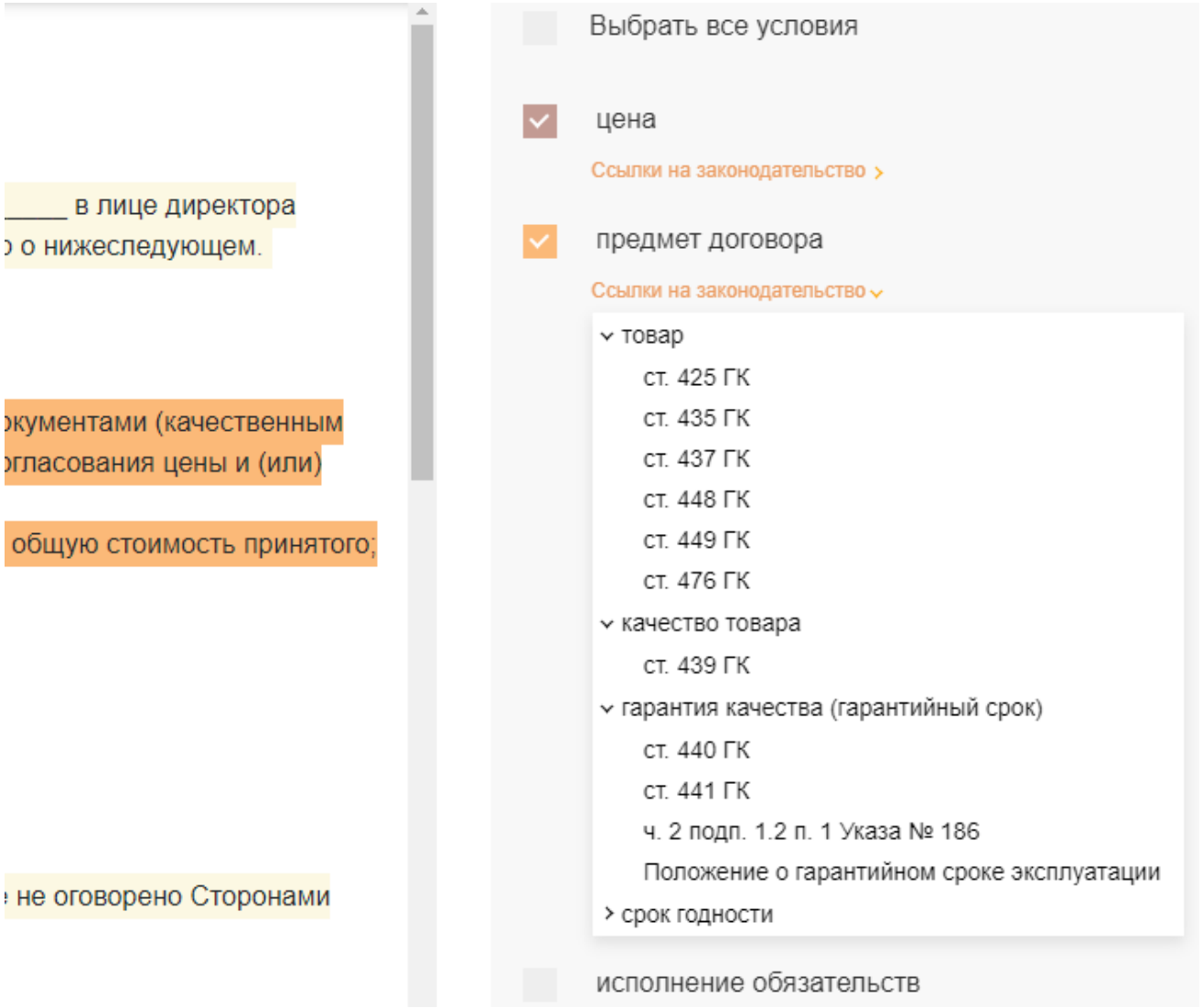
https://contracts.silkdata.ai/segmentation/

Contract Analysis (RU): Web



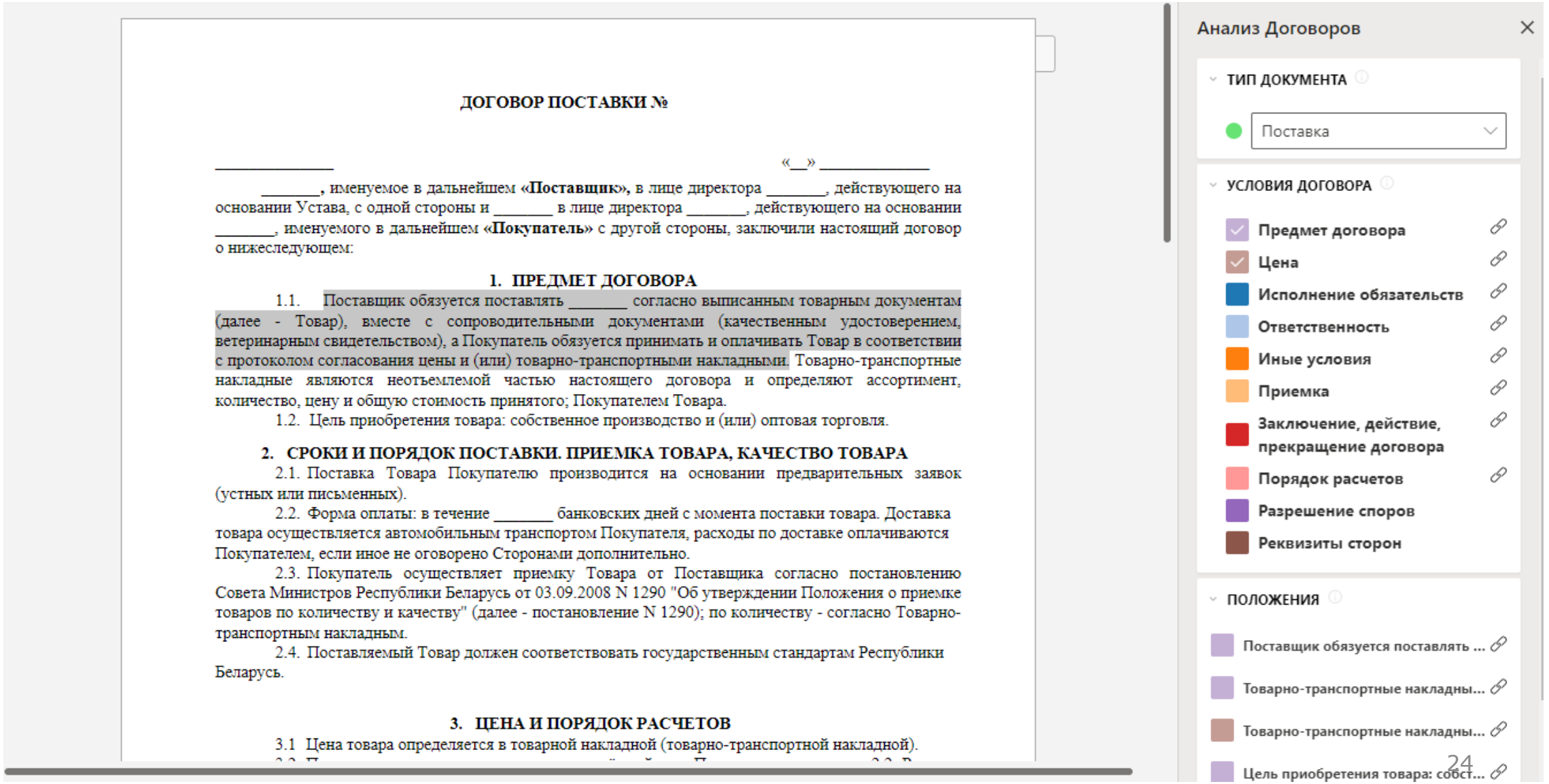
https://contract.by/app/

Links to legislation



- Currently:
  - Expert-provided links
- In future:
  - Automated suggestion of law provisions and edits

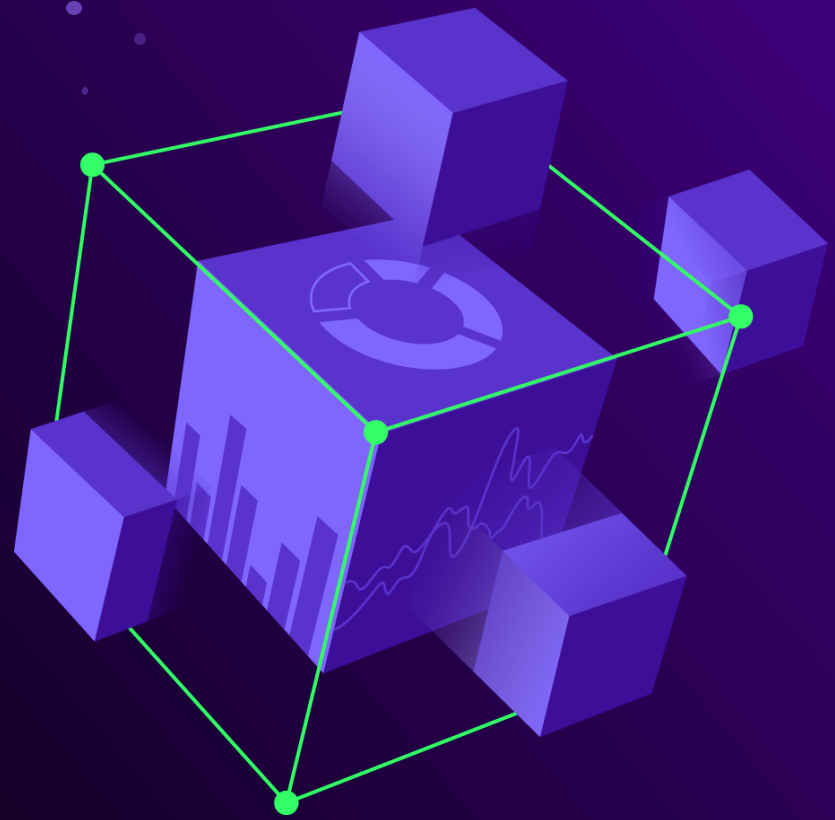
Contract Analysis: Word add-on



Conclusion: current state

- Finalized MVP
- Actively used by pilot users
  - Ca. 300 contracts checked a week
  - About 10% recurring users
  - Positive feedback
- Analysis of user activity and suggestions
- Looking for partnership on other markets

DATA SCIENCE SUMMIT



Thank you for watching!

Remember to leave your **questions** and **rate** the presentation in the screen below.

nikolai.karelin@silkdata.ai

www.dssconf.pl 03 grudnia 2020 PGE Narodowy + Online