

Introducing Kairntech: AI-powered Text Analysis

Stefan Geißler, Kairntech



Oct, 2019

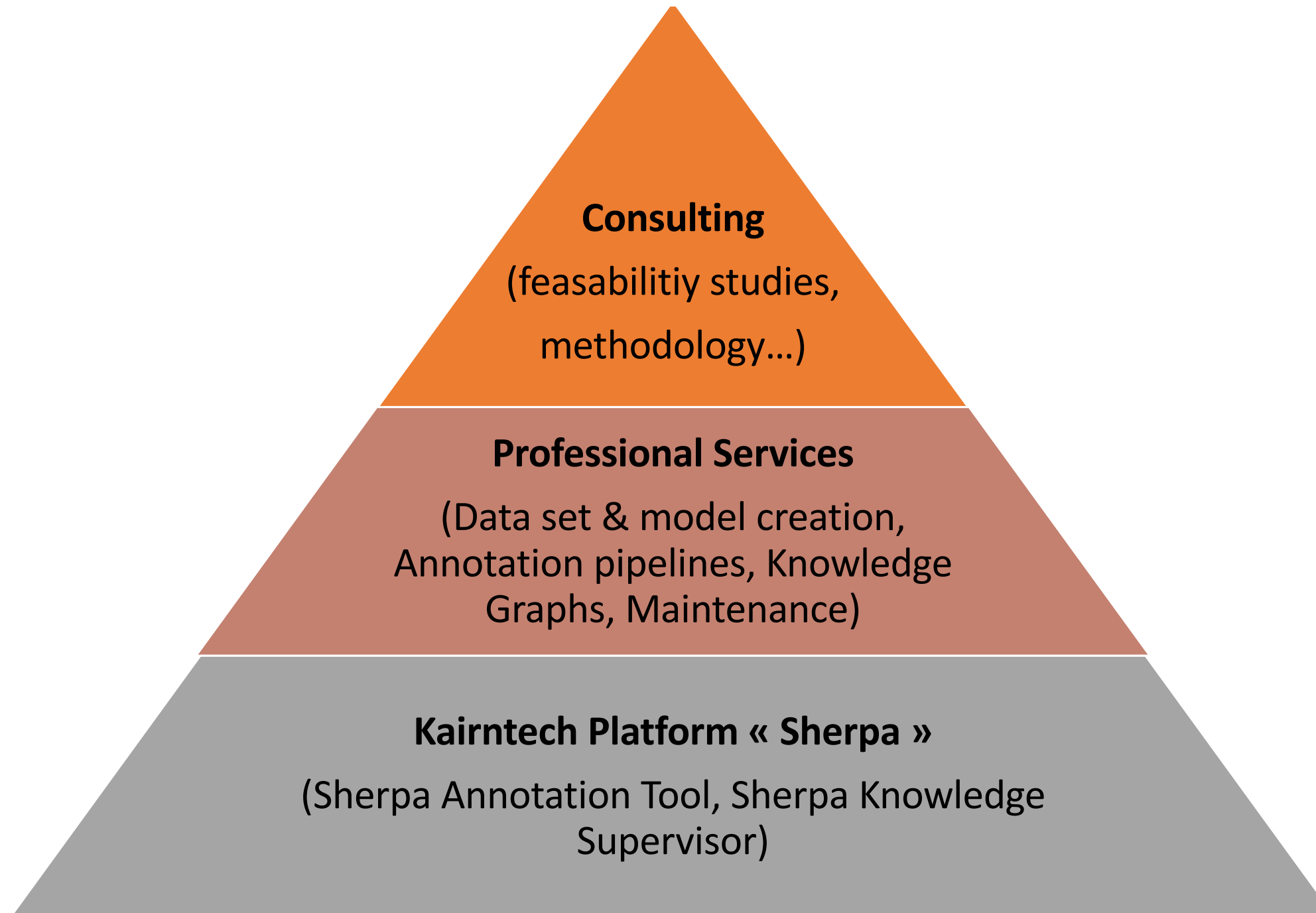
❖ Kairntech: The company

- ❖ Created in dec 2018, 10 partners
- ❖ France (Paris & Grenoble/Meylan), Germany (Heidelberg)

❖ Kairntech: The team

- ❖ Background in Software engineering, Machine Learning, Sales, Management
- ❖ +15 years of experience in NLP development and deployment from Xerox, IBM, TEMIS. Development of components currently in production at CERN, NASA, EPO...





- ❖ Off-the-shelf NLP models often **don't work for specific needs**
- ❖ Implementation is **slowed down** by the need of building specific training dataset
- ❖ AI/NLP services are often require **integration of business glossaries & knowledge graph**
- ❖ Absence of maintenance leads to **quality deviations**

So: We need data, not only algorithms

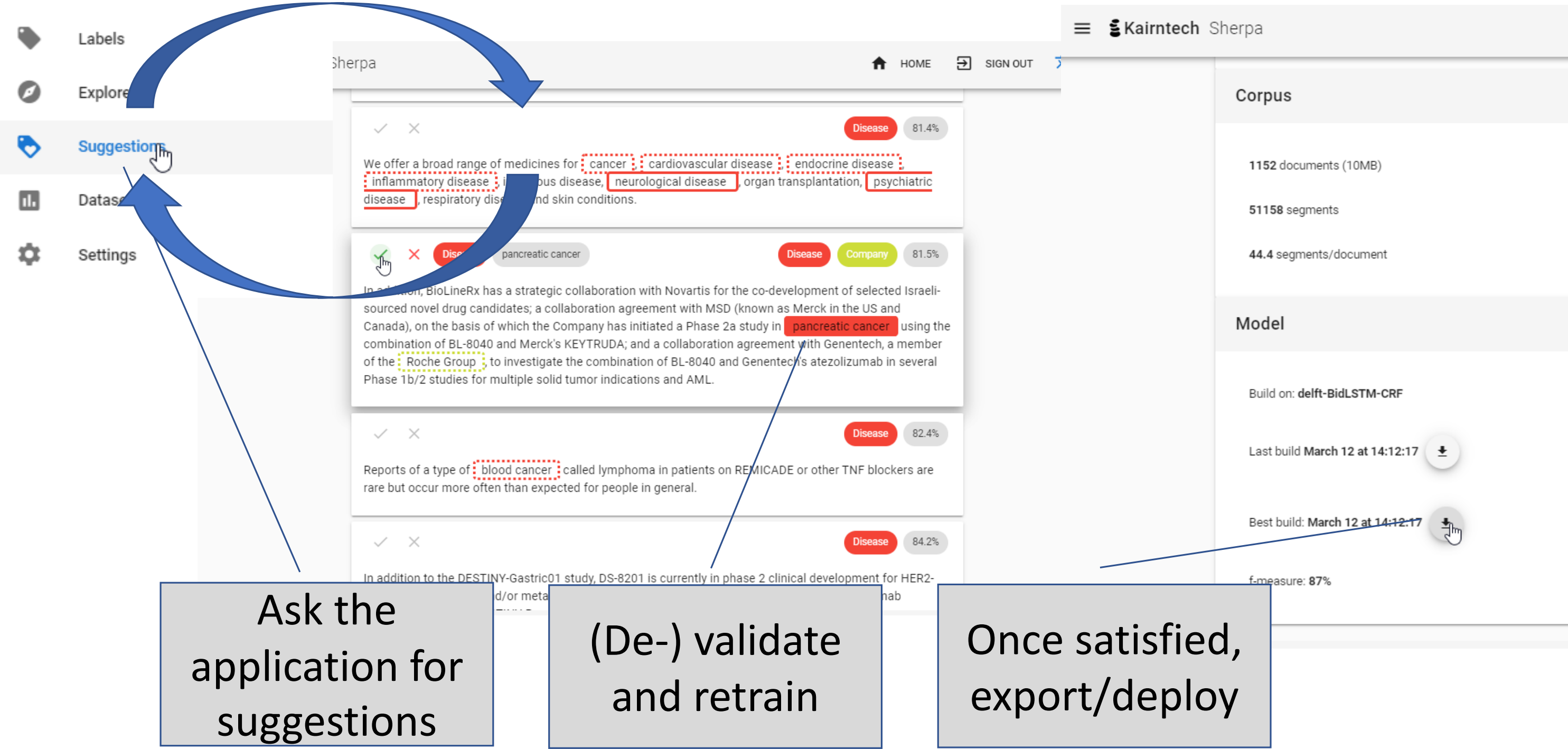


What people think AI is about



The reality

The Sherpa: A non-expert workflow for dataset creation



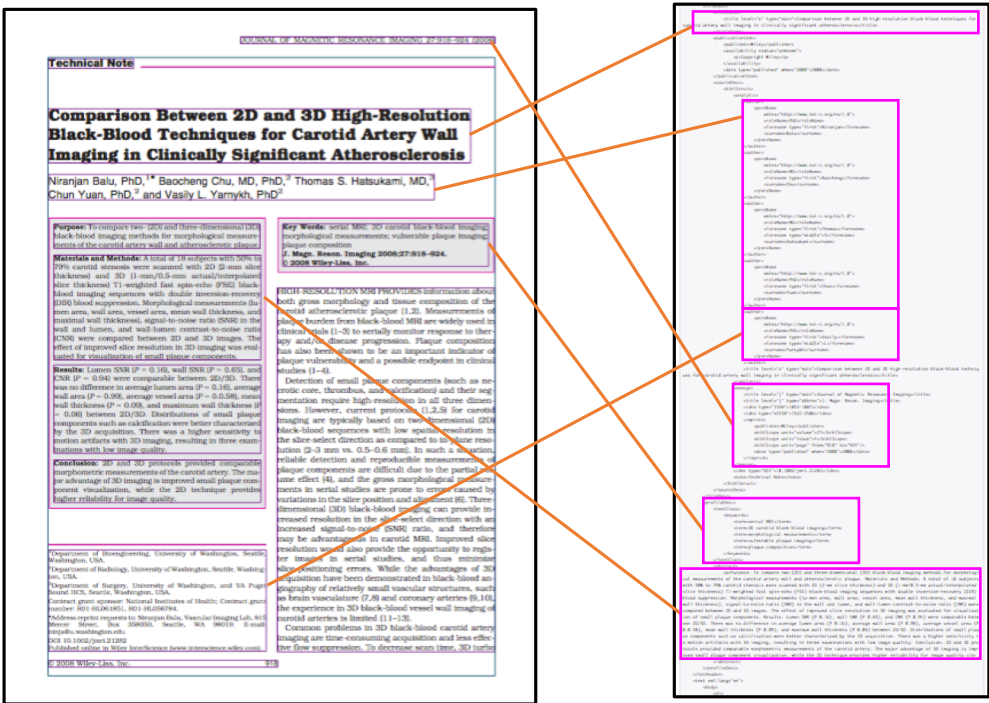
- ❖ **Web-based** (no install), intuitive GUI, usable by domain experts
- ❖ Limit manual annotation efforts: ***Active Learning***
- ❖ **Collaboration** (work in teams, measure inter-annotator agreement)
- ❖ **Not just NER** annotation: Entity typing, document categorization, ...
- ❖ Facilitate **deployment-to-production**

Activities that are not part of a product:

❖ Annotation directly into PDF documents (respecting document structure)

❖ Annotation (normalization, disambiguation, scoring, linking) according to existing vocabularies and thesauri

❖ Integration with Graph Databases



CIGARETTE SMOKE (CS)-induced AIRWAY EPITHELIAL SENESCENCE has been implicated in the PATHOGENESIS of CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) although the underlying mechanisms remain largely unknown. GROWTH DIFFERENTIATION FACTOR 15 (GDF15) is increased in AIRWAY EPITHELIUM of COPD SMOKERS and CS-exposed human AIRWAY EPITHELIAL CELLS, but its role in CS-induced AIRWAY EPITHELIAL SENESCENCE is unclear. In this study, we first analyzed expression of GDF15 and CELLULAR SENESCENCE markers in AIRWAY EPITHELIAL CELLS of current SMOKERS and NONSMOKERS. Second, we determined the role of GDF15 in CS-induced AIRWAY EPITHELIAL SENESCENCE by using the CLUSTERED REGULARLY INTERSPACED SHORT PALINDROMIC REPEATS (CRISPR) CRISPR associated-9 (CAS9) GENOME EDITING approach. Finally, we examined whether EXOGENOUS GDF15 PROTEIN promoted AIRWAY EPITHELIAL SENESCENCE through the ACTIVIN RECEPTOR-LIKE KINASE 1 (ALK1) SMAD1 PATHWAY. GDF15 UP-REGULATION was found in parallel with increased CELLULAR SENESCENCE markers P21, P16 and HIGH MOBILITY GROUP box 1 (HMGB1) in AIRWAY EPITHELIAL CELLS of current SMOKERS compared with NONSMOKERS. Moreover, CS extract (CSE) induced CELLULAR SENESCENCE in cultured human AIRWAY EPITHELIAL CELLS, represented by induced SENESCENCE-ASSOCIATED B-GALACTOSIDASE activity, INHIBITED CELL PROLIFERATION, increased P21 expression, and increased release of HMGB1 and IL-6. Disruption of GDF15 significantly INHIBITED CSE-induced AIRWAY EPITHELIAL SENESCENCE. Lastly, GDF15 PROTEIN BOUND to the ALK1 RECEPTOR and promoted AIRWAY EPITHELIAL SENESCENCE via activation of the SMAD1 PATHWAY. Our findings highlight an important contribution of GDF15 in promoting AIRWAY EPITHELIAL SENESCENCE upon CS exposure. SENESCENT AIRWAY EPITHELIAL CELLS that CHRONICALLY accumulate in CS-exposed LUNGS could contribute substantially to CHRONIC AIRWAY INFLAMMATION in COPD development and progression.

SMAD1

Normalized: Mothers against decapentaplegic homolog 1

Domains: Biology, Engineering

conf: 0.7665

Mothers against decapentaplegic homolog 1 also known as SMAD family member 1 or SMAD1 is a protein that is encoded by the SMAD1 gene.

Entrez Gene ID	4086
HGNC gene symbol	SMAD1
HGNC ID	6767
OMIM ID	601595
subclass of	Q20747295
Ensembl Gene ID	ENSG00000170365
HomoloGene ID	21196
RefSeq RNA ID	XM_011531964
RefSeq RNA ID	XM_011531964
RefSeq RNA ID	XM_011531964

- ❖ **So much data!**

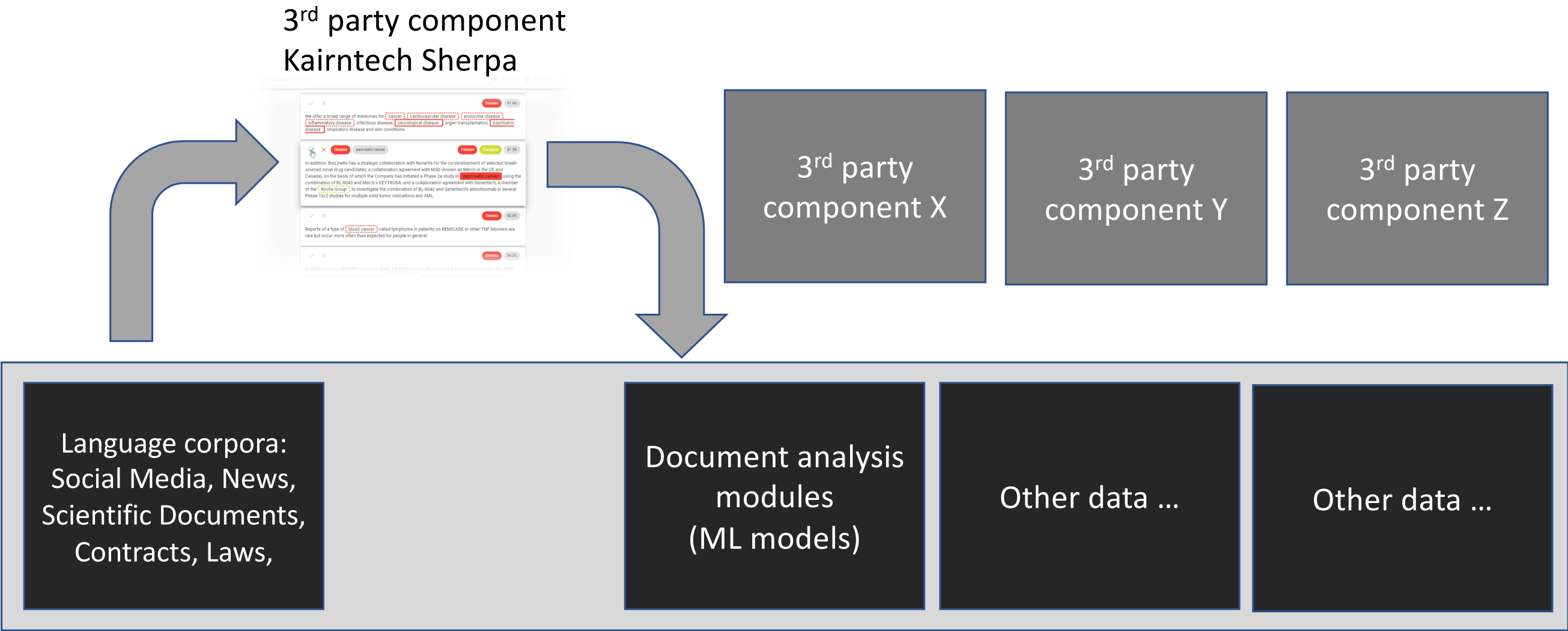
- ❖ But very little of it labelled and useful for supervised learning

- ❖ **So many pretrained models!**

- ❖ But most of the time they do not quite do what you need in your project

- ❖ **So many algorithms!**

- ❖ But a library alone will not allow you to implement the solution you need



European Language Grid: language data, analysis modules, ...

- ❖ **Documented and uniform format for Document corpora**
- ❖ **API to allow third party components to access and process data**
- ❖ **API to feed resulting modules back into the ELG**
- ❖ **License issues:**
 - ❖ Limitations on document corpora and resulting modules
 - ❖ Definition of business model for 3rd party components