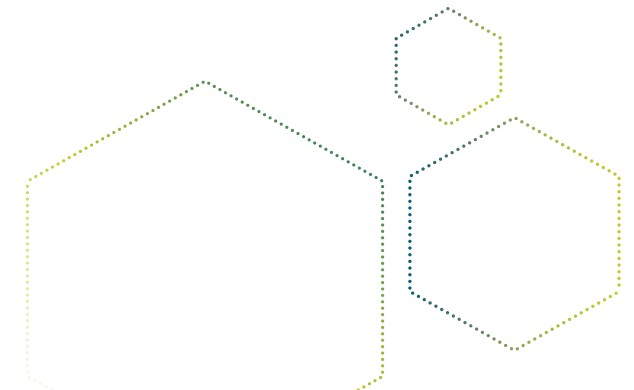# AI-based tools for patent classification

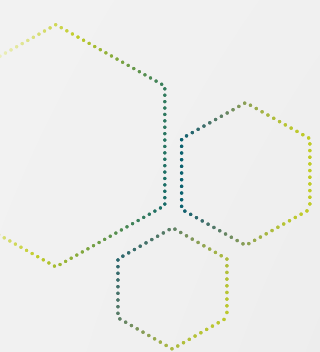WIPO SCP/33 – Session on the use of AI for patents – 07st December 2021

Houda MOUZOUN – Data Scientist

inpi

SUMMARY

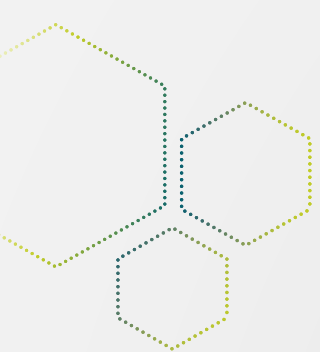AI FOR...

- Patent Pre-classification

- Patent Classification

- Other Applications
    – Project Portfolio

inpi

PATENT PRECLASSIFICATION
A time consuming task

Patent Applications

- ▶ 300 applications per week

- ▶ Distribution among 3 services and then 10 examination teams

- ▶ Significant analysis work by service managers and team leads

- ▶ Error prone even when done by humans

S1   S2   S3

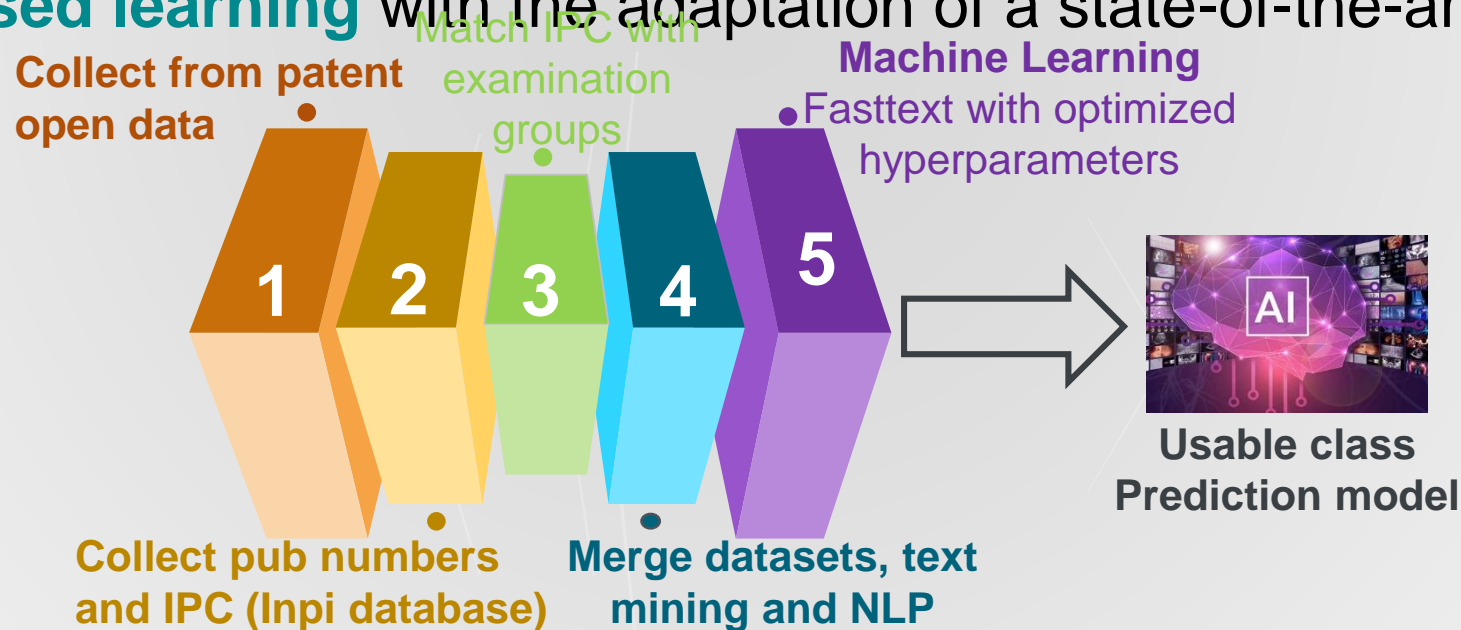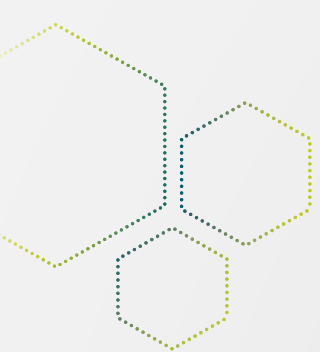P11 P12 P13   P21 P22 P23 P24   P31 P32 P33

Examiners

INPI

PATENT PRECLASSIFICATION
Model development using AI

▶ **Data collection** with 150 000 patents published by INPI over a 1O years period

▶ **Natural language processing** to prepare and pre-process data

▶ **Supervised learning** with the adaptation of a state-of-the-art model

Collect from patent open data

Match IPC with examination groups

Machine Learning
Fasttext with optimized hyperparameters

1 **2** **3** **4** **5**

Collect pub numbers and IPC (Inpi database)

Merge datasets, text mining and NLP

AI

Usable class Prediction model

inpi

PATENT PRE-CLASSIFICATION
Successful automation

► **In production** since 2019 with use by patent department replacing manual dispatching

  ► **Significant time savings** for service managers estimated at 10h per week

  ► **Good accuracy**, around 80%, similar to manual dispatching performance

► **On going Model upgrade** by further adapting word representations to patent vocab

|							
P1	86,06%	3,86%	2,47%	0,48%	2,79%	0,63%	0,00%	3,17%
P2	4,78%	76,53%	3,91%	0,48%	3,14%	6,62%	3,07%	3,57%
P3	2,39%	4,82%	85,39%	1,45%	6,97%	2,52%	0,61%	1,59%
P4	1,59%	0,96%	3,91%	90,80%	5,92%	1,89%	2,76%	0,40%
P5	2,79%	2,57%	2,06%	3,62%	74,22%	4,42%	0,61%	1,98%
P6	0,80%	6,11%	1,85%	1,45%	1,74%	75,39%	3,07%	1,98%
P7	0,80%	2,25%	0,41%	0,48%	2,79%	3,47%	86,50%	2,78%
P8	0,80%	2,89%	0,00%	1,21%	2,44%	5,05%	3,37%	84,52%

Correct class: 83%

Ambiguous class: 8%

Clear error: 9%

Confusion matrix

In production model accuracy (initial setup with 8 groups)

INPI
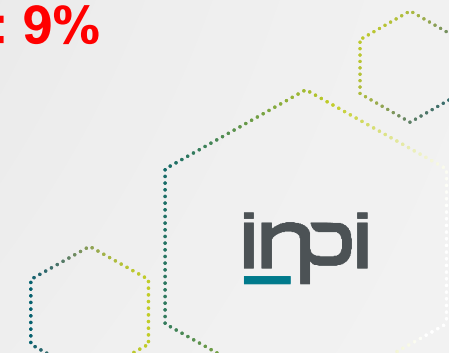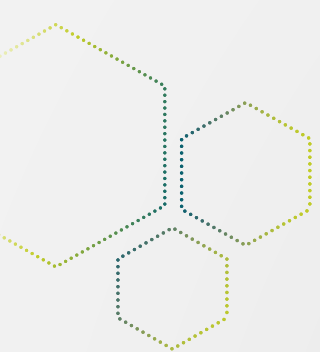
PATENT CLASSIFICATION
A complex task …

IPC: A 5-level classification hierarchy

G   Physics

G06   Computing; Calculating; Counting

G06F   Electric digital data processing

G06F 8/00   Arrangements for software engineering

G06F 8/20   Software design

► An **I**nternational **P**atent **C**lassification to characterize the addressed technological domain

► A multi-label classification with one primary code and several secondary codes assigned for each patent

► A fine-grained classification with more than 60 000 subdivisions at lower level!

inpi
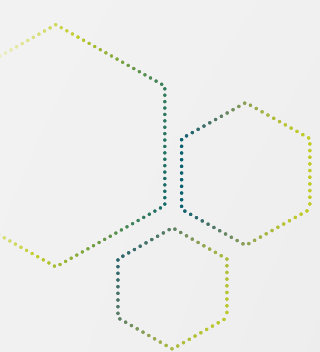
PATENT CLASSIFICATION
… requiring sophisticated models

► **Focus** on extreme multi-label text classification models to move toward automation

► **Feasibility** study by evaluating state-of-the art approaches on IPC4 and IPC8 classification

► **Collaboration** with a research institute on this cutting-edge technology

XML-CNN

A family of convolutional neural networks that is tailored to extreme multi-label text classification.

http://nyc.lti.cs.cmu.edu/yiming/Publications/jliu-sigir17.pdf

Parabel

Partitioned Label Trees for extreme multi-label learning and classification.

http://manikvarma.org/code/Parabel/download.html

X-BERT

Finetuning of BERT models on extreme Multi-label Text Classification.

https://github.com/guoqunabc/X-BERT
https://github.com/google-research/bert

inpi
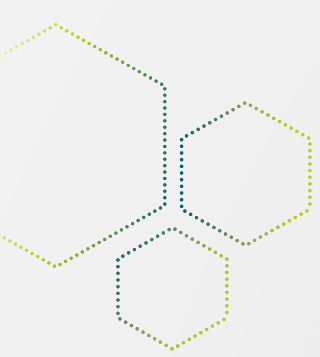
AN EXPANDING PROJECT PORTFOLIO

**Periodically assessed and updated roadmap,
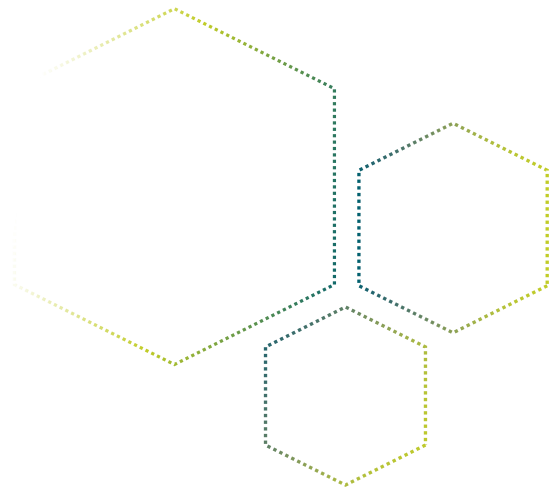with already 4 deployed AI-based tools :**

► Patent (pre-)classification

► Vienna classification of trademarks

► Automation for recordals

► Document anonymization

… and new projects at various stages of development :

► Customer data repository consolidation with deduplication

► A voicebot following the extension of our chatbot