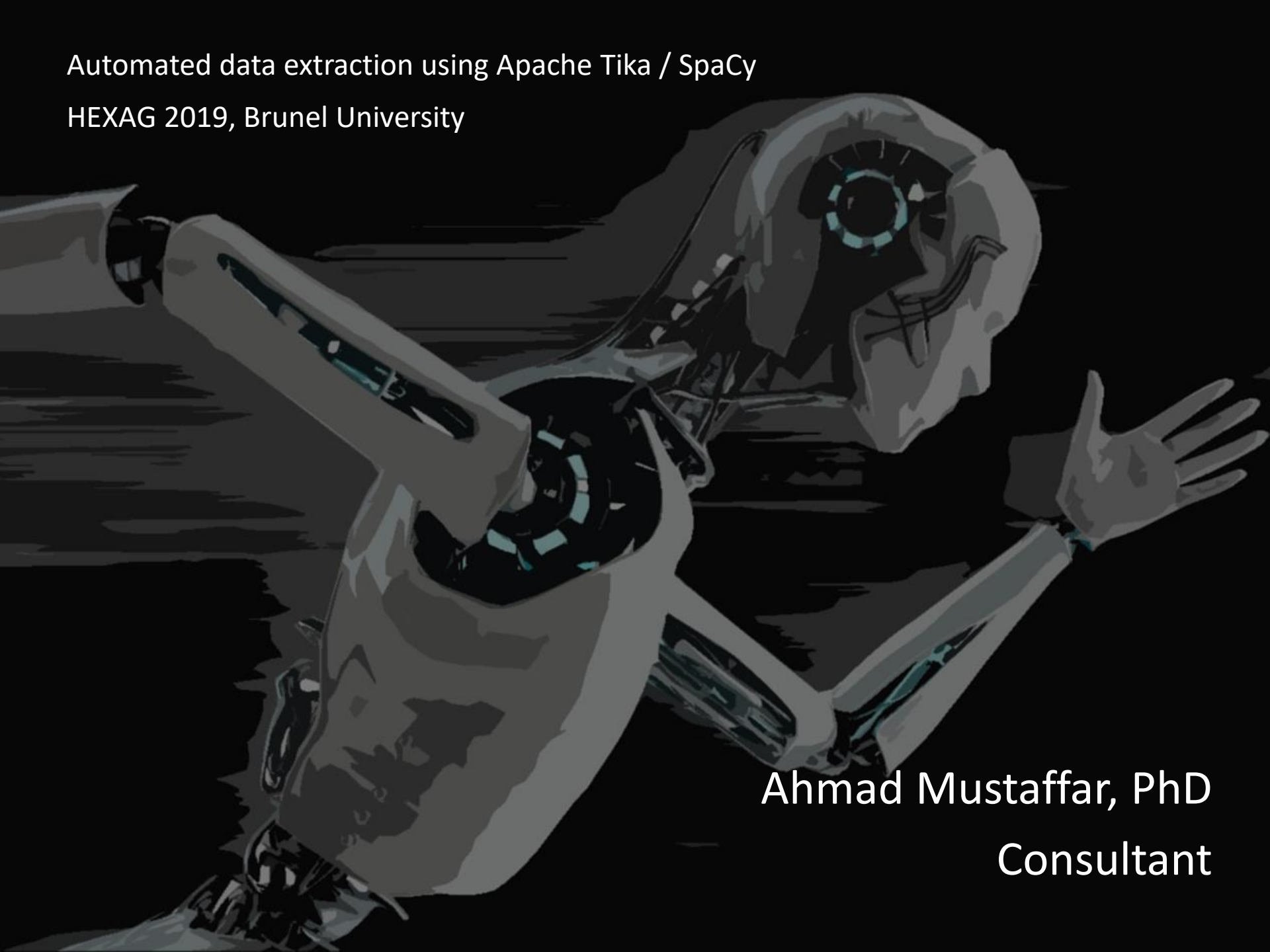
Automated data extraction using Apache Tika / SpaCy

HEXAG 2019, Brunel University



Ahmad Mustaffar, PhD

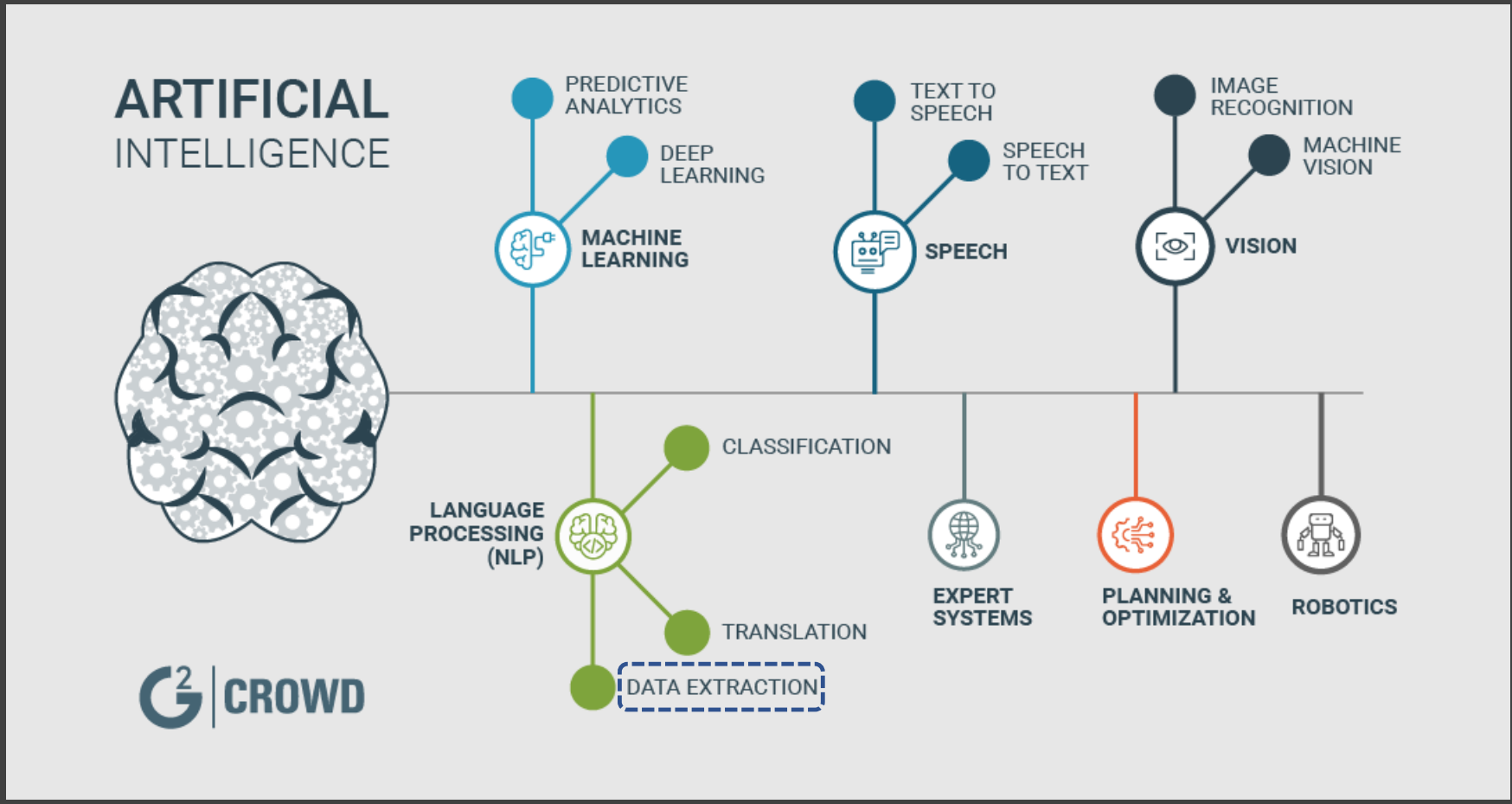
Consultant

Background

- In this presentation, we will:
  - Mine a certain data set from a collection of pdfs (100 heat transfer articles)
  - For demonstration, we will mine:
    - Number of “pcm” acronyms in a file
    - The numerical data of xxx kWh
    - Readability score (Flesch reading ease score)
  - Create a script to automate the data mining and output into a pandas data frame for further analysis.

ML common algorithms:

- Regression
- Classification
- Decision trees
- Random forest
- Artificial neural networks



Today's presentation

Ease of date extraction vs. file size

X 1

Easy

X 10

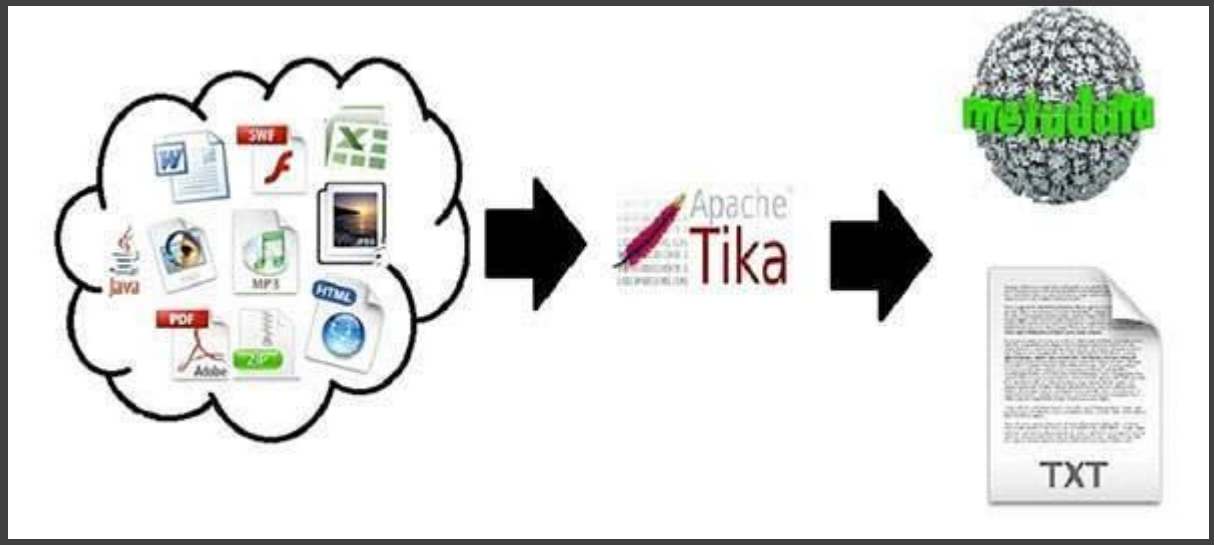
Fairly easy

X 10^6

Good luck!

…big data requires automation

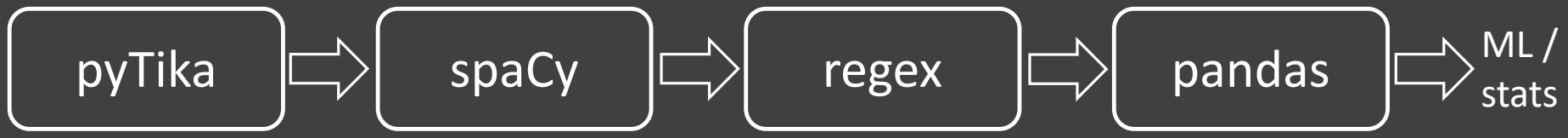
Engine of data extraction: Apache Tika parser



1000+ file types → Metadata

→ Content

Pipeline



- Apache Tika parser

- Tokeniser
- NER
- POS tag
- Stopwords
- Etc.

- Search pattern

- Data frame

Let’s mine something

- Downloaded last night:
  - 100 heat transfer research articles
  - Science direct
  - For 2020 release

- Mine (as a quick example):
  - ”PCM” or “pcm” - acronyms
  - “xx kwh or kWh” – numerical data
  - Readability score (Flesch Reading Ease) – metric

Score	School level	Notes
100.00-90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–0.0	College graduate	Very difficult to read. Best understood by university graduates.

Search pattern - REGEX

- PCM or pcm →

  'pcm', re.IGNORECASE

- 123 kwh or 1.23 kWh or 12.3 kwH →

  '(\d*\.?\d+?\s?)kwh', re.IGNORECASE

Code

- Short and efficient python code
- Can run on 1, 10, or even millions of any types of docs: returns dataframe

```
from tika import parser
import pandas as pd
import json
import glob
import spacy
import re
import textstat

nlp = spacy.load('en')

def process_file(file):
    parsedPDF = parser.from_file(file)
    doc = nlp(parsedPDF["content"])

    find_kwh_res = find_kwh(doc)
    find_pcm_res = find_pcm(doc)
    flesch_ease_res = flesch_ease(doc)

    return {"File" : file, "kWh list": find_kwh_res, "Num.PCM": find_pcm_res, "Flesch metric": flesch_ease_res}

def find_kwh(doc):
    regex_kwh = "\d*\.?\d+?\s?kwh"
    list_kwh = re.findall(regex_kwh, str(doc), re.IGNORECASE)
    return list_kwh

def find_pcm(doc):
    words = list(filter(lambda w: not w.is_stop, doc))
    num_pcm = re.findall('pcm', str(words), re.IGNORECASE)
    return len(num_pcm)

def flesch_ease(doc):
    return textstat.flesch_reading_ease(str(doc))

files = glob.glob("*.pdf")
data = list(map(process_file, files))
```

```
dF = pd.DataFrame(data=data)
dF
```

Data frame

- Can export to csv, xlsx, etc. for further analysis

	File	Flesch metric	Num.PCM	kWh list
0	Combustion-process-of-a-Korean-wood-pellet-at-...	46.00	0	[]
1	Conservation-of-Moroccan-truffle--Terfezia-bou...	47.93	0	[]
2	CFD-modeling-and-evaluation-the-performance-of...	44.07	232	[2.8 kWh, 5.4 kWh, 7.5 kWh, 4.2\nkWh, 5.47 kWh]
3	Reduced-order-modeling-approach-for-parametriz...	49.38	0	[]
4	Experimental-study-on-double-pipe-PCM-floor-he...	40.28	131	[6.8 kWh, 6.9 kWh, 10.9 kWh, 15 kWh]
5	Evaluation-of-neutron-radiation-damage-in-zirc...	52.43	0	[]
6	Thermal-performance-of-a-solar-fa-ade-system-f...	46.00	3	[0.693 kWh, 0.591 kWh, 0.693 kWh, 0.5 kWh, 0.5...
7	Effects-of-saturated-soil-on-the-lengths-of-a-...	42.85	0	[10113.6 kWh, 7080 kWh, 3033.6 kWh, 10113.6 kW...
8	Effect-of-inner-pipe-type-on-the-heat-transfer...	40.31	0	[]
9	3D-porous-V2O5-architectures-for-high-rate-lit...	13.51	0	[]
10	A-study-on-electrochemical-hydrogen-storage-pe...	40.18	0	[]
11	Modeling-the-short-term-and-long-term-behaviou...	46.24	0	[]
12	Performance-evaluation-of-7-2-kWp-standalone-b...	55.00	0	[8927.1 kWh, 1000 kWh, 637 kWh, 7.2 kWh, 3.7 k...
13	The-role-of-pole-and-molecular-geometry-of-fat...	56.79	0	[]
14	Optimal-design-for-solar-greenhouses-based-on-...	30.91	1	[]
15	Micro-cracks-distribution-and-power-degradatio...	38.39	0	[]
16	Performance-improvement-of-a-flat-plate-solar-...	45.90	0	[]
17	Experimental-and-numerical-investigations-of-a...	50.20	0	[]
18	Full-scale-numerical-study-on-the-flow-charact...	44.78	0	[]
19	A-preliminary-sensitivity-study-of-Planetary-B...	40.08	0	[]
20	Field-investigation-of-a-photonic-multi-layere...	34.09	0	[]

Conclusions

- Automation:
  - Save time
  - Efficient
  - Accurate
  - Highly customisable search pattern
- Great for literature review, survey research, analysis

Code for today is available at:

https://github.com/amustaffar/data_extraction_heat_transfer_articles

Thanks!