

AI Applications in POWER Engineering

Established in 1909



Gurunath Gurrala, gurunath@iisc.ac.in

Associate Professor,

Department of Electrical Engineering

Indian Institute of Science

**Convener, POWERGRID Center of
Excellence in Cyber Security (PGCoE)**

18 September 2026



To provide for **advanced instruction**
and to conduct **original investigations**
in all branches of knowledge as are
likely to **promote the material and**
industrial welfare of India

Question Answering over Power Grid Documents

Knowledge discovery

The following websites documents have lot of documents related to power system operation. We want to build a repository called PowerGPT

1. Welcome to SRPC Ministry of Power

<https://www.srpc.kar.nic.in/html/xml-search/meetings.html>

2. Western Regional Power Committee

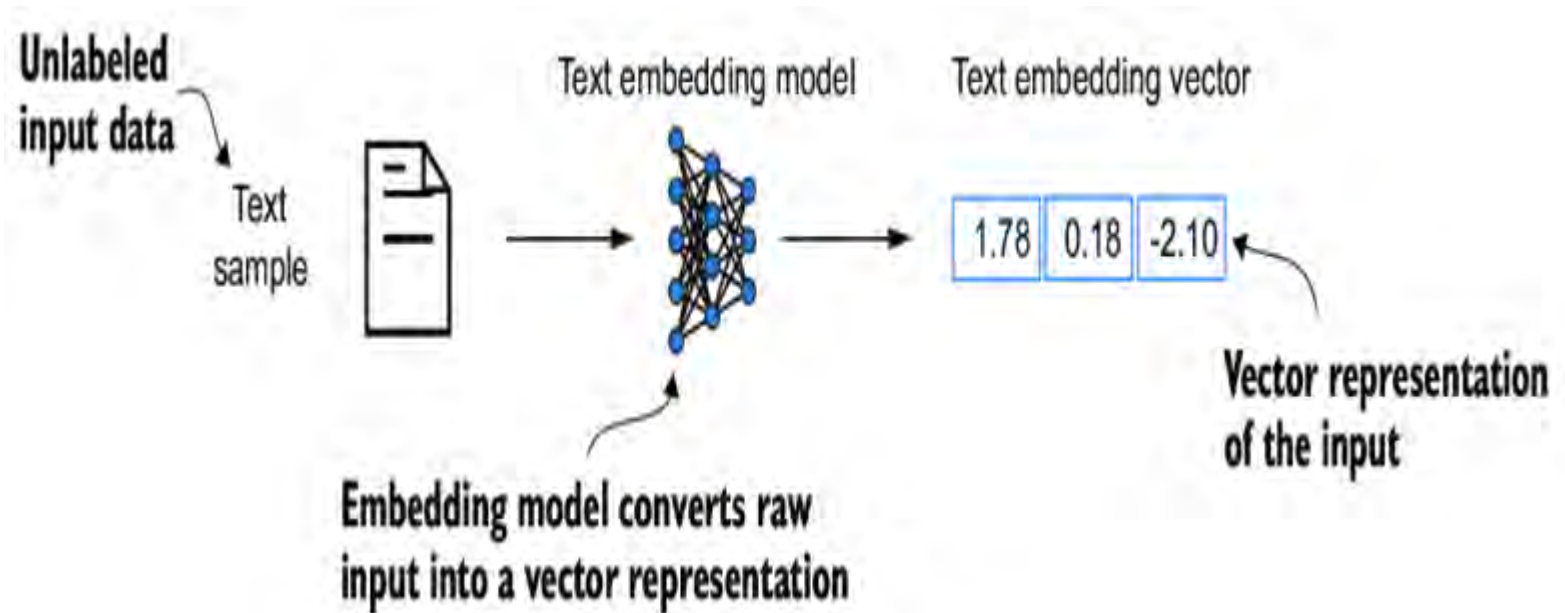
https://wrpc.gov.in/menu/Minutes_459

3. Welcome to Eastern Regional Power Committee :: » PCC

<https://erpc.gov.in/meeting-2/pcc/>

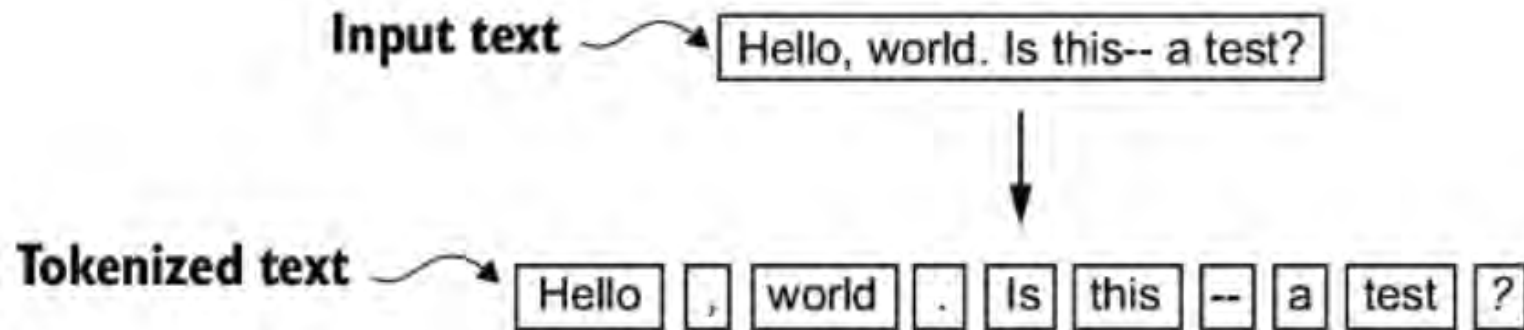
Working with Text Data

- ❑ Deep neural network models, including LLMs, cannot process raw text directly. Therefore, we need a way to represent words as continuous-valued vectors.
- ❑ The concept of converting data into a vector format is often referred to as embedding.
- ❑ Using a specific neural network layer or another pretrained neural network model, we can embed text into vectors.
- ❑ The primary purpose of embeddings is to convert nonnumeric data into a format that neural networks can process.



Tokenization

- ❑ Before embedding we need to some preprocessing on text data.
- ❑ **Tokenization** or dividing the text into tokens is one of the preprocessing step.
- ❑ Tokenization splits text into Words, subwords, punctuation, special symbols, etc.
- ❑ Then we given **token id** to each word to convert the word into numerical data.



Techniques of Tokenization

❑ **Whitespace & Regex Tokenization :**

- Splits on spaces or punctuation using regular expressions.
- Simple and fast, but may not handle rare/compound words well
- E.g. "Hello world" → "Hello", "world"

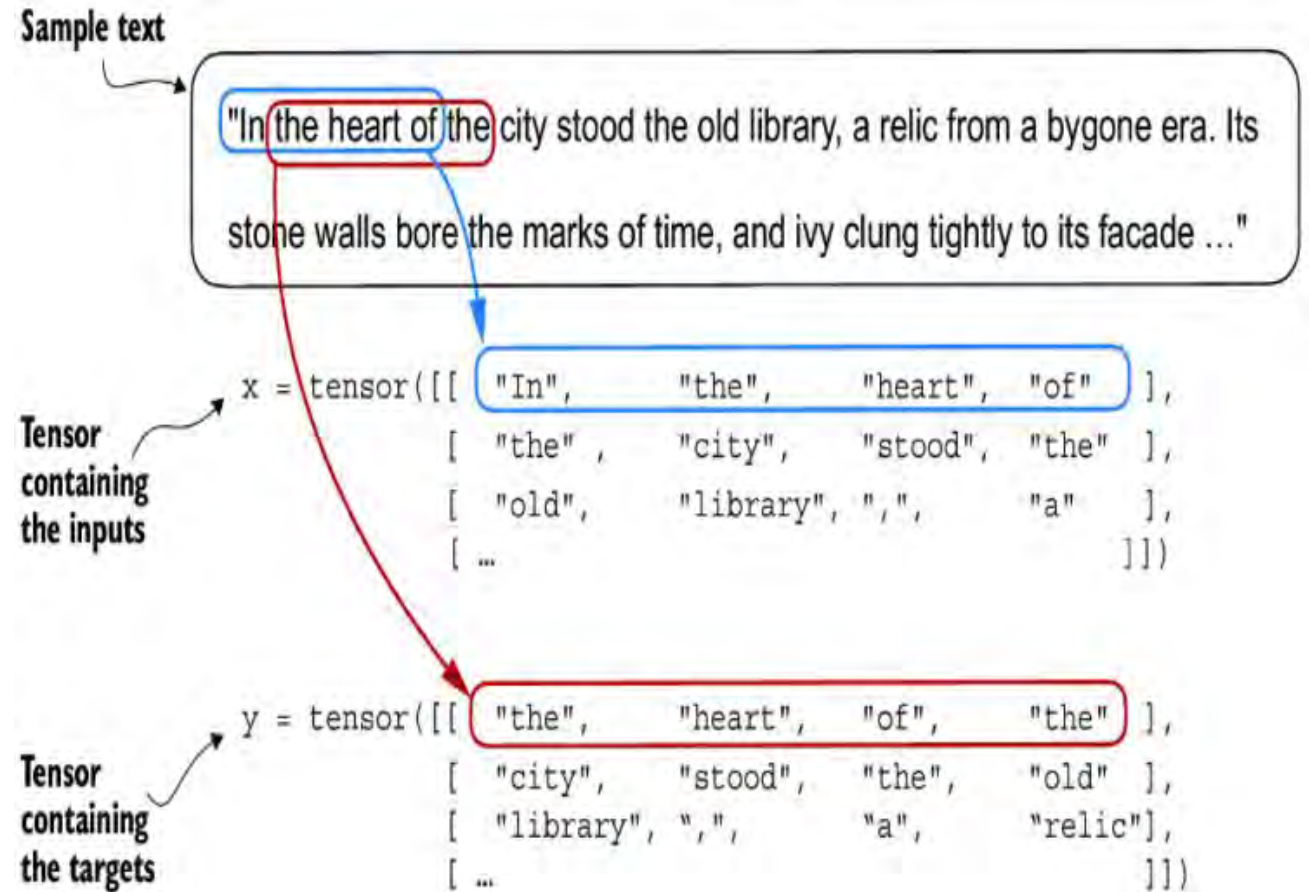
❑ **Byte Pair Encoding (BPE) :**

- Starts with characters, merges most frequent pairs.
- Efficient vocabulary size, fewer special tokens.
- E.g. "someunknownPlace" → "some", "unknown", "Place"

- ❑ During the tokenization we also add some special tokens like start of sentence **[BOS]**, end of sentence **[EOS]** unknown tokens **[UNK]**, etc for smooth tokenization, handle unknown words and unambiguity.

Data sampling with a sliding window

- ❑ The next step in creating the embeddings for the LLM is to generate the input–target pairs required for training an LLM.
- ❑ Use a **sliding window** to create training data.
- ❑ It enables model to learn **next-token prediction**



Creating token embeddings

- ❑ The last step in preparing the input text for LLM training is to convert the token IDs into embedding vectors
- ❑ Token IDs are mapped to **dense vectors** using matrix multiplication and one hot encoding
- ❑ Embedding layer contains learnable weights.
- ❑ Since our model is a LLM so we need to combine the input with **positional embedding** before getting the embedded token.

Encoding Word Positions

□ We can use two broad categories of position aware embeddings:

- **Absolute positional embeddings.**

Absolute positional embeddings are directly associated with specific positions in a sequence. For each position in the input sequence, a unique embedding is added to the token's embedding to convey its exact location.

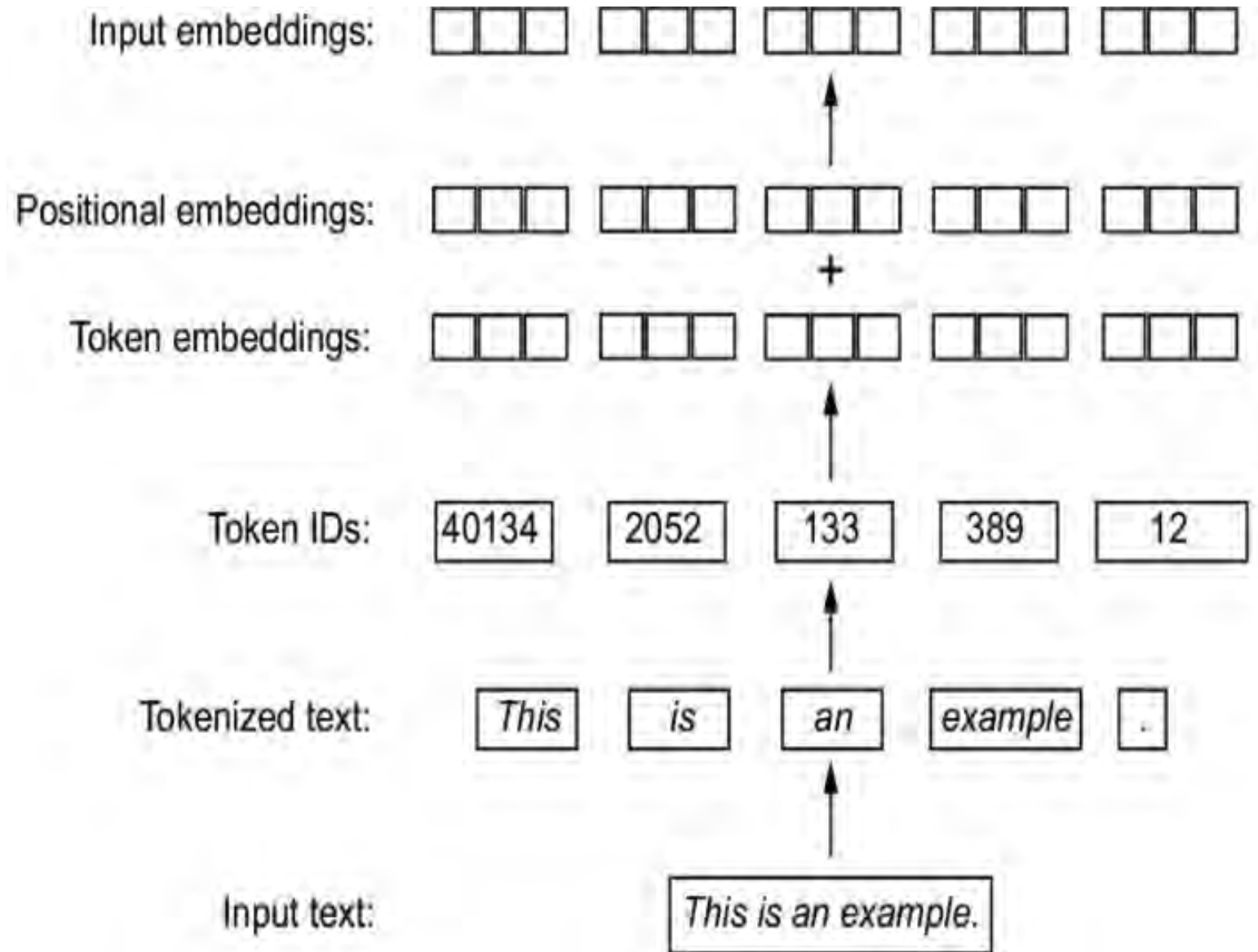
e.g. some fixed number for each position.

- **Relative positional embeddings**

the emphasis of relative positional embeddings is on the relative position or distance between tokens. This means the model learns the relationships in terms of “how far apart” rather than “at which exact position”.

e.g. sinusoidal positional encoding.

Summary



Sample PowerGPT Questions and Answers

Answer from PowerGPT:

```
answer = rag_chain.invoke("Due to hudhud cyclone, which transmission lines were effected ? ")  
print(answer)
```

✓ 0.8s

The transmission lines affected due to Hudhud cyclone were:

1. CTU Lines:

- SIMHADRI - GAZUWAKA-II at 08:51:00 Hrs
- SIMHADRI - GAZUWAKA-I at 09:59:00 Hrs
- VEMAGIRI - SIMHADRI-I at 10:06:00 Hrs
- VEMAGIRI - SIMHADRI-II at 10:37:00 Hrs
- JEYPORE - GAZUWAKA-I at 10:16:00 Hrs
- JEYPORE - GAZUWAKA-II at 10:30:00 Hrs
- HVDC GAJUWAKA POLE-I at 10:30:00 Hrs
- HVDC GAJUWAKA POLE-II at 10:30:00 Hrs

2. State Owned Lines:

- KALPAKKA - GAZUWAKA-I at 10:30:00 Hrs
- KALPAKKA - GAZUWAKA-II at 08:37:00 Hrs
- KALPAKKA - VEMAGIRI-I & II at 10:06:00 Hrs
- KHAMMAM - KALPAKKA-I & II at 10:11:00 Hrs
- SIMHADRI - KALPAKKA-II at 10:22:00 Hrs
- SIMHADRI - KALPAKKA-III at 10:06:00 Hrs
- SIMHADRI - KALPAKKA-IV at 10:06:00 Hrs

From DOC:

(Page 3 of 10)

- On 12th October 2014 due to 'Hudhud' cyclone effect, the following main 400 kV transmission lines had tripped:

CTU Lines:

- ✓ SIMHADRI - GAZUWAKA-II at 08:51:00 Hrs
- ✓ SIMHADRI - GAZUWAKA-I at 09:59:00 Hrs.
- ✓ VEMAGIRI - SIMHADRI-I at 10:06:00 Hrs.
- ✓ VEMAGIRI - SIMHADRI-II at 10:37:00 Hrs.
- ✓ JEYPORE - GAZUWAKA-I at 10:16:00 Hrs.
- ✓ JEYPORE - GAZUWAKA-II at 10:30:00 Hrs.
- ✓ HVDC GAJUWAKA POLE-I at 10:30:00 Hrs
- ✓ HVDC GAJUWAKA POLE-II at 10:30:00 Hrs.

State Owned Lines:

- ✓ KALPAKKA - GAZUWAKA-I at 10:30:00 Hrs.
- ✓ KALPAKKA - GAZUWAKA-II at 08:37:00 Hrs.
- ✓ KALPAKKA - VEMAGIRI-I & II at 10:06:00 Hrs.
- ✓ KHAMMAM - KALPAKKA-I & II at 10:11:00 Hrs.
- ✓ SIMHADRI - KALPAKKA-II at 10:22:00 Hrs.
- ✓ SIMHADRI - KALPAKKA-III at 10:06:00 Hrs.
- ✓ SIMHADRI - KALPAKKA-IV at 10:06:00 Hrs.

Sample PowerGPT Questions and Answers

Answer from PowerGPT:

```
answer = rag_chain.invoke("Due to hudhud cyclone, which transmission lines were effected ? ")  
print(answer)
```

✓ 0.8s

The transmission lines affected due to Hudhud cyclone were:

1. CTU Lines:

- SIMHADRI - GAZUWAKA-II at 08:51:00 Hrs
- SIMHADRI - GAZUWAKA-I at 09:59:00 Hrs
- VEMAGIRI - SIMHADRI-I at 10:06:00 Hrs
- VEMAGIRI - SIMHADRI-II at 10:37:00 Hrs
- JEYPORE - GAZUWAKA-I at 10:16:00 Hrs
- JEYPORE - GAZUWAKA-II at 10:30:00 Hrs
- HVDC GAJUWAKA POLE-I at 10:30:00 Hrs
- HVDC GAJUWAKA POLE-II at 10:30:00 Hrs

2. State Owned Lines:

- KALPAKKA - GAZUWAKA-I at 10:30:00 Hrs
- KALPAKKA - GAZUWAKA-II at 08:37:00 Hrs
- KALPAKKA - VEMAGIRI-I & II at 10:06:00 Hrs
- KHAMMAM - KALPAKKA-I & II at 10:11:00 Hrs
- SIMHADRI - KALPAKKA-II at 10:22:00 Hrs
- SIMHADRI - KALPAKKA-III at 10:06:00 Hrs
- SIMHADRI - KALPAKKA-IV at 10:06:00 Hrs

From DOC:

(Page 3 of 10)

- On 12th October 2014 due to 'Hudhud' cyclone effect, the following main 400 kV transmission lines had tripped:

CTU Lines:

- ✓ SIMHADRI - GAZUWAKA-II at 08:51:00 Hrs
- ✓ SIMHADRI - GAZUWAKA-I at 09:59:00 Hrs.
- ✓ VEMAGIRI - SIMHADRI-I at 10:06:00 Hrs.
- ✓ VEMAGIRI - SIMHADRI-II at 10:37:00 Hrs.
- ✓ JEYPORE - GAZUWAKA-I at 10:16:00 Hrs.
- ✓ JEYPORE - GAZUWAKA-II at 10:30:00 Hrs.
- ✓ HVDC GAJUWAKA POLE-I at 10:30:00 Hrs
- ✓ HVDC GAJUWAKA POLE-II at 10:30:00 Hrs.

State Owned Lines:

- ✓ KALPAKKA - GAZUWAKA-I at 10:30:00 Hrs.
- ✓ KALPAKKA - GAZUWAKA-II at 08:37:00 Hrs.
- ✓ KALPAKKA - VEMAGIRI-I & II at 10:06:00 Hrs.
- ✓ KHAMMAM - KALPAKKA-I & II at 10:11:00 Hrs.
- ✓ SIMHADRI - KALPAKKA-II at 10:22:00 Hrs.
- ✓ SIMHADRI - KALPAKKA-III at 10:06:00 Hrs.
- ✓ SIMHADRI - KALPAKKA-IV at 10:06:00 Hrs.

Sample PowerGPT Questions and Answers

Here we can see that it can give correct answer for a small context inside a paragraph also

Answer from PowerGPT:

```
answer = rag_chain.invoke("Reason for NLC one day delay in UPI payments?")  
print(answer)
```

✓ 0.3s

The reason for NLC one day delay in UPI payments was due to the 'Thane Cyclone' effect.

From DOC:

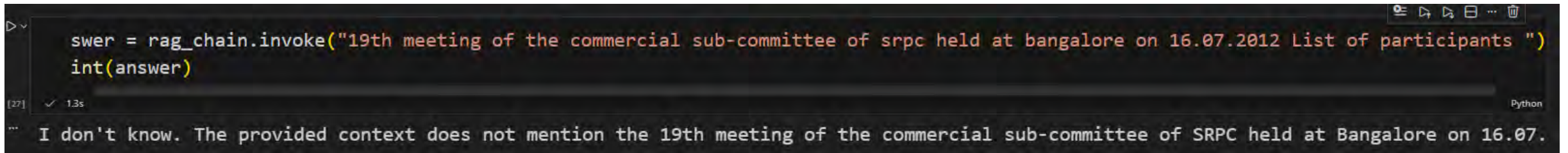
SEPL : ₹9.18 Lakhs

- 9.2 In the meeting NLC explained that the one day delay in making UI payment by NLC was due to the 'Thane Cyclone' effect. NLC had subsequently paid the UI charges and also the applicable interest for the delay. NLC had been very prompt in making UI payment to the pool and had defaulted only once out of the 88 occasions. It was also stated that NLC was no longer drawing power for commissioning activities under UI mechanism as they had started drawing power from TN as HT consumer. NLC therefore requested to exempt from opening LC.
- 9.3 MS I/c, SRPC said that in view of the track record of NLC and as an one time measure NLC TPS-II Expansion could be exempted from opening LC. The

Some of the results :

However it also fails for some of the queries

Answer from RAG :



```
> ✓ 1.3s Python  
[27]: swer = rag_chain.invoke("19th meeting of the commercial sub-committee of srpc held at bangalore on 16.07.2012 List of participants ")  
      int(answer)  
      ""  
      I don't know. The provided context does not mention the 19th meeting of the commercial sub-committee of SRPC held at Bangalore on 16.07.
```

In order to address this issue, this week I am planning to go through different vector DB techniques and their implementations.

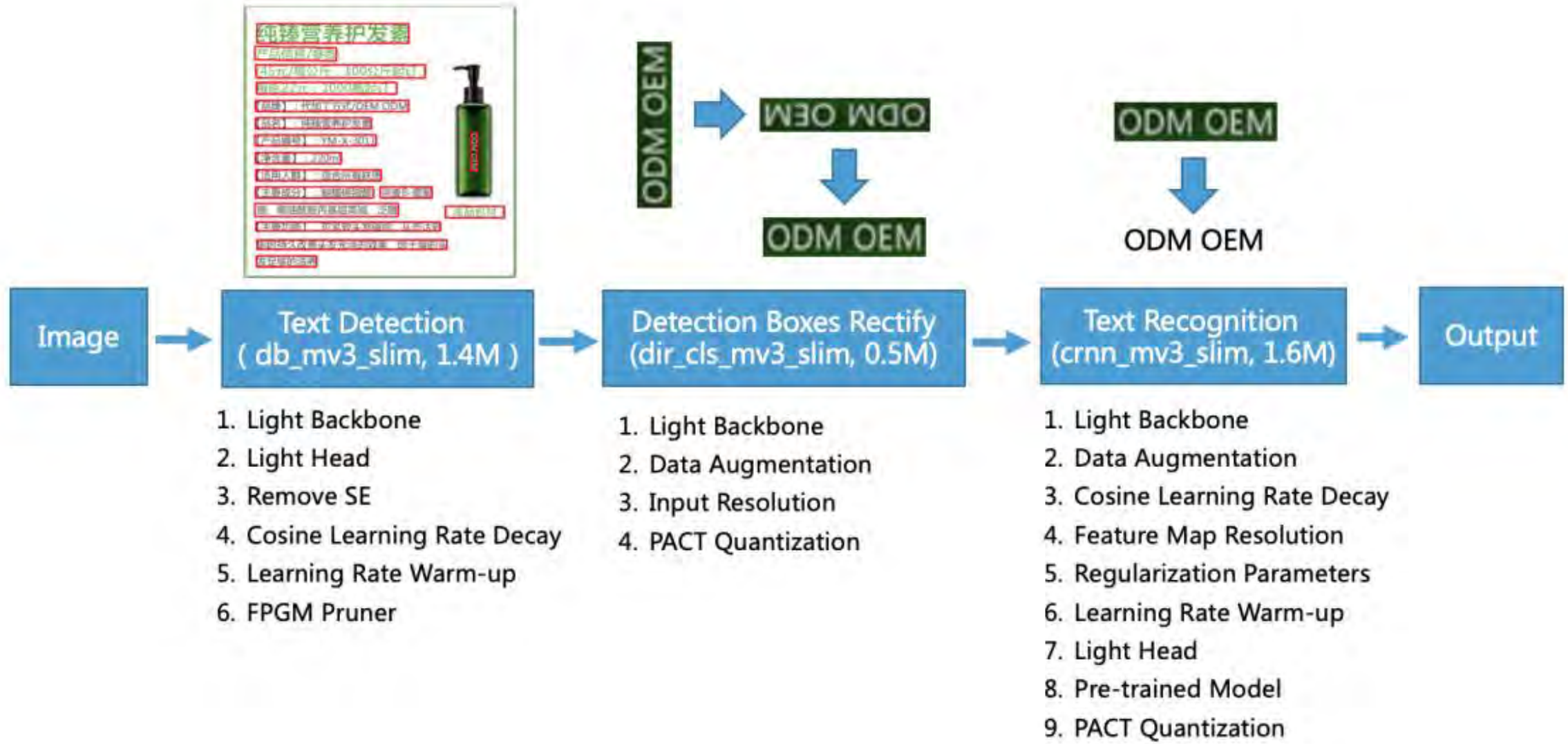
PowerGPT development Approach – Step-1



Text Recognition

- ❑ In PP-OCR, we use **CRNN (Shi, Bai, and Yao 2016)** as text recognizer, which is widely used and practical for text recognition.
- ❑ CRNN integrates feature extraction and sequence modeling. It adopts the **Connectionist Temporal Classification(CTC)** loss to avoid the inconsistency between prediction and label. Before recognizing the detected text, the text box needs to be transformed into a **horizontal rectangle box** for subsequent text recognition, which is easy to be achieved by geometric transformation as the detection frame is composed of four points.
- ❑ However, the rectified boxes may be **reversed**. Thus, a classifier is needed to determine the text direction. If a box is determined reversed, further **flipping** is required.
- ❑ To enhance the model ability and reduce the model size of a text recognizer, the following nine strategies are used:
 - **light backbone,**
 - **cosine learning rate decay,**
 - **regularization parameters,**
 - **light head,**
 - **PACT quantization.**
 - **data augmentation,**
 - **feature map/input resolution,**
 - **learning rate warm-up,**
 - **pre-trained model and**
 - **remove SE module,**
 - **FPGM pruner.**
- ❑ Finally, the model size of the text direction classifier is **500KB**.

Text Recognition Pipeline



Implementation Pipeline

- ❑ Convert pdf into images using **pd2image**
- ❑ OCR per page using **PaddleOCR**
- ❑ Store the result in **.txt** files
- ❑ Installed CUDA 11.8
- ❑ Manually exported paths for cuDNN, cuBLAS
- ❑ Performance 5-10x faster vs CPU

ANNEXURE-II
(2/2)

Agenda Notes in the matter of availing HT power supply from respective DISCOM by NTPC/NLC for supplying to power existing colonies in the premises of inter-state generating station

The Hon'ble Central Electricity Regulatory Commission had notified the regulations on terms & conditions of tariff related matters, 2014 for Inter-State Generating Companies as well as Inter-State Transmission Utilities. The new regulations are effective from 1.4.2014 and are valid for period of 5 years i.e., up to 31.3.2019. As per the prevailing regulations, 2014, the colony consumption as well as construction power are not part of auxiliary consumption. Due to this, NTPC/NLC has to obtain the HT supply from respective DISCOM for supplying the power to employee colonies existing in the respective thermal stations. In case AP, NTPC shall have to obtain the HT supply from APEPDCL for supplying the power to colonies of Simhadri Power Project.

PP-OCR Generated Text

PAGE

02

01/01/2011

Annexure-II.

(Page2of2)

Agenda Notes in the matter of availing HT power supply from repective DiScOM by.NTPc/NLC for supplying to power oxisting.cglonies_in the premises of inter-state gienrating station.

The Hon'ble Central Electricity Regulatory Commission had notified the regulations on terms & conditions of tariff reiaied matters, 2014 for Inter-State Generating Companies an well as Inter-State Transmission Utilities. The new regutations are offective from 1.4.2014 and are valld for period of 5 years i.e., up to 31.3.2019. As per the prevaing regulatione, 2014, the colony consumption as well as construction power are.not part of auxillary consumption. Due t this, NTPC/NLC has to obtain :he HT supply from respective DiscOM for supplying the!power to einployee colonies existing In the respective thermal stations.I case AP, NTPC shall have to obtaln the HT supply from APEPDCL for supplying the!power to colonies of Sirnhadri Power Projecl.

Tesseract Generated Text

02:43 BAIL aG23317E22

1

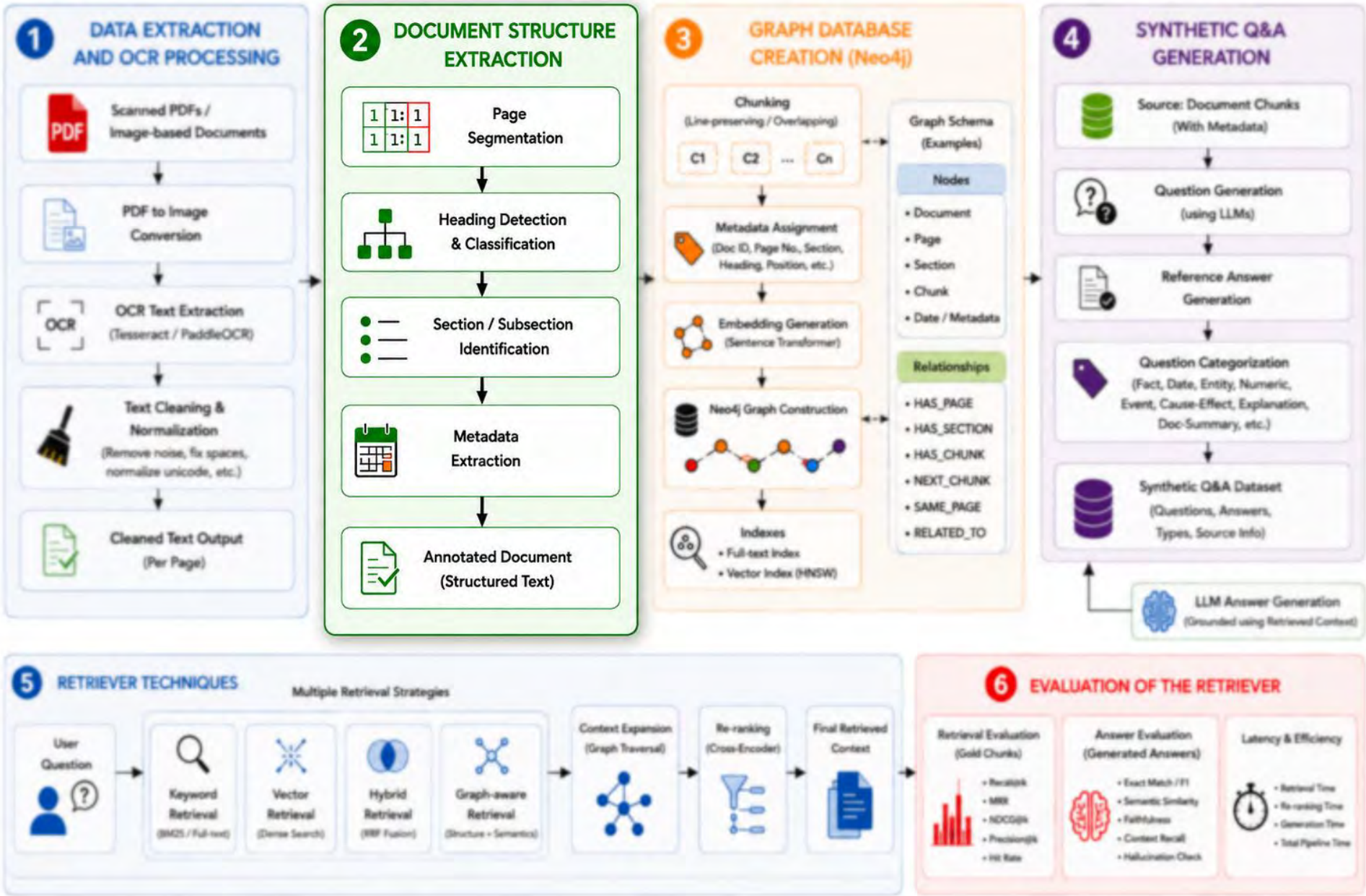
AG enda Notes in the mat ter of avalina, at pewer supply from reanective

BISCOM by NTPGI/NLC for suppl ying to power existing colonies in the
premi sas ofi inter-state fenerating station 4

The Hombie Centrat Electricity Regulatory Commission had notified the
regulations on terms & conditions of tariff retaiect matters, 2014 for Inter State

Generating Companies as well as Inter-State. Transmission Utilities. The new
regulations are effective from 1.4.2014 and are valid for period of 8 years i.e., up
to 3.2019. As per the prevailing regulations, 2014, the colony consumption as
well as construction power are. not part of auxiliary consumption. Due to this,
NTPC/NLC has to obtain the PT supply from respective DISCOM for supplying
the power to employee colonies existing in the respective thermal stations. If
in case AP, NTPC shall have to obtain the HT supply from APEPDCL for supplying
the power to colonies of Simhadri Power Project.

PowerGPT development Approach – Step-2



Document Structure Extraction Process

- ❑ We tried to build relationship between the chunks based on the different kinds of headings.
- ❑ First we find the candidate headings based on some rules.
- ❑ Then we use pass these candidate headings, doc txt and the page image to a llama-18B multimodal instruct llm to extract anymore candidate heading which we might have missed.
- ❑ Then we got the final list of candidate headings.
- ❑ Now we need to classify these headings into various categories like – Page heading, section heading, subsection heading and non headings, making a hierarchy between them.
- ❑ To classify the heading, we passed the candidate headings, doc txt and page image into the same llama-18B model.

Candidate headings based on the rules

```
{
  "source_file": "../OCR_BENCH_OUTPUT/ppocr/filtered/agenda_anx_extracted_f
  "num_pages": 48,
  "pages": [
    {
      "page_num": 1,
      "page_image": "page_001.png",
      "candidate_headings": [
        {
          "line_index": 0,
          "text": "Annexure-I"
        },
        {
          "line_index": 2,
          "text": "FAX/SPEEDPOST"
        },
        {
          "line_index": 4,
          "text": "Government of India"
        },
        {
          "line_index": 6,
          "text": "Central Electricity Authority"
        }
      ]
    }
  ]
}
```

Doc info

Page number of corresponding doc

Candidate headings

Candidate headings based on the rules and LLM extracted

```
{
  "source_file": "../OCR_BENCH_OUTPUT/ppocr/filtered/agenda_anx_extracted_f
  "num_pages": 48,
  "pages": [
    {
      "page_num": 1,
      "page_image": "page_001.png",
      "candidate_headings": [
        {
          "line_index": 0,
          "text": "Annexure-I",
          "anchor_text": "Annexure-I"
        },
        {
          "line_index": 2,
          "text": "FAX/SPEEDPOST",
          "anchor_text": "FAX/SPEEDPOST"
        },
        {
          "line_index": 4,
          "text": "Government of India",
          "anchor_text": "Government of India"
        },
        {
          "line_index": 6
```

Doc info

Page number of corresponding doc

Candidate headings

Anchor Text from the doc

Candidate headings based
on the rules and LLM
extracted

```
"pages": [  
  {  
    "page_num": 1,  
    "page_image": "page_001.png",  
    "classified_headings": [  
      {  
        "line_index": 0,  
        "text": "Annexure-I",  
        "anchor_text": "Annexure-I",  
        "label": "page_heading",  
        "level": 0  
      },  
      {  
        "line_index": 2,  
        "text": "FAX/SPEEDPOST",  
        "anchor_text": "FAX/SPEEDPOST",  
        "label": "section_heading",  
        "level": 1  
      },  
      {  
        "line_index": 4,
```

Page number of corresponding doc

Candidate headings

Anchor Text from the doc

Classified Heading

PowerGPT development Approach – Step-3



Neo4j Graph Structure

□ The graph models a document as a hierarchy of nodes:

1. Document
2. Page
3. Headings
4. Chunk.

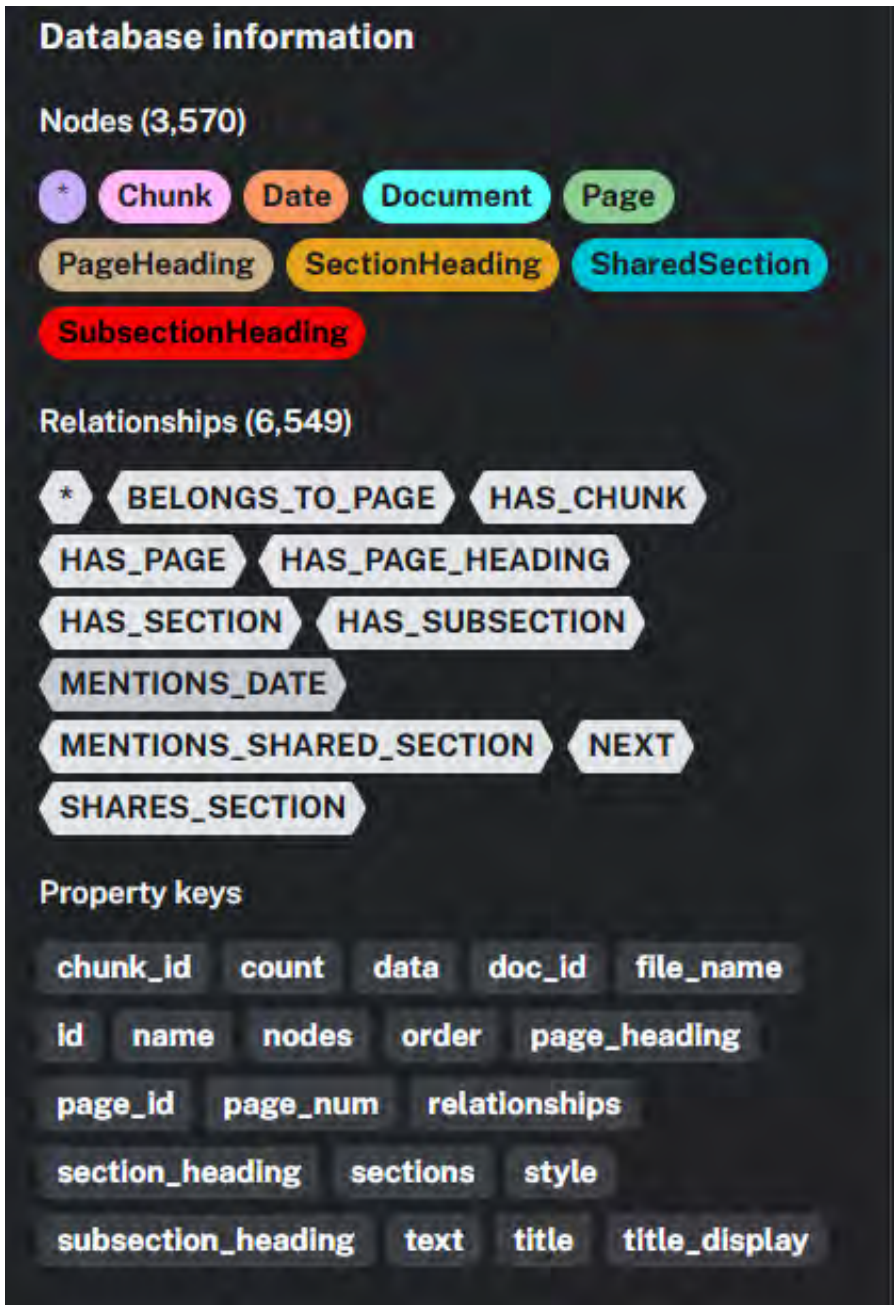
□ Main node types used are:

- Document,
- Page,
- SectionHeading,
- PageHeading,
- SubsectionHeading,
- Chunk, and
- Date.

Relationships in the Graph

- ❑ **(Document)-[:HAS_PAGE]->(Page)** : Connects each document to its pages.
- ❑ **(Page)-[:HAS_PAGE_HEADING]->(PageHeading)** : Links sections under a page heading.
- ❑ **(SectionHeading)-[:HAS_SUBSECTION]->(SubsectionHeading)** : Links headings
- ❑ **(:Page / :PageHeading / :SectionHeading / :SubsectionHeading)-[:HAS_CHUNK]->(Chunk)** :
Attaches text chunks to the most specific parent.
- ❑ **(Chunk)-[:BELONGS_TO_PAGE]->(Page)** : Keeps page reference for every chunk.
- ❑ **(Chunk)-[:NEXT]->(Chunk)** : Maintains reading order across the document.
- ❑ **(Chunk)-[:MENTIONS_DATE]->(Date)** : Extracts and links dates found in text.

DataBase Schema Information



Document Node

neo4j\$ MATCH (n:Document) RETURN n LIMIT 25;

Graph Table RAW

minutes_extract...

minutes_and_ex...

agenda_and_ag...

agenda_extract...

Node details

Document

Key	Value
<id>	4:9b20e604-b39b-4e17-87db-1f1a4d68ccf4:174
doc_id	"doc_2"
file_name	"agenda_anx_extracted_filtered.txt"

Started streaming 4 records after 25 ms and completed after 25 ms.

Page Node

neo4j\$ MATCH (n:Page) RETURN n LIMIT 25;

GraphTableRAW

doc_1

doc_1

doc_1

doc_1

doc_1

Node details

Page

Key	Value
<id>	4:9b20e604-b39b-4e17-87db-1f1a4d68ccf4:54
doc_id	"doc_1"
page_id	"doc_1_page_6"
page_num	6

PageHeading Node

neo4j\$ MATCH (n:PageHeading) RETURN n LIMIT 25;

GraphTableRAW

Page 2/17

Page 2/17

Minutes of 19th...

Node details

PageHeading

Key	Value
<id>	4:9b20e604-b39b-4e17-87db-1f1a4d68ccf4:38
doc_id	"doc_1"
id	"doc_1_ph_5"
order	5
page_num	3
title	"Minutes of 19th Meeting of Commercial Sub-Committee of SRPC"

Started streaming 25 records after 35 ms and completed after 38 ms.

SectionHeading Node

neo4j\$ MATCH (n:SectionHeading) RETURN n LIMIT 25;

GraphTableRAW

2002 CTU had b...

TNEB even af...

SRLDC vide let

Node details

SectionHeading

Key	Value
<id>	4:9b20e604-b39b-4e17-87db-1f1a4d68ccf4:30
doc_id	"doc_1"
id	"doc_1_sh_9"
order	9
page_num	3
title	"2002 CTU had been advised f or installing SEMs at remote e nd of"

Started streaming 25 records after 35 ms and completed after 36 ms.

SubsectionHeading Node

neo4j\$ MATCH (n:SubsectionHeading) RETURN n LIMIT 25;

GraphTableRAW

MEPL

Member Secreta...

AGM, SRLDC

Manage- r,...

Node details

SubsectionHeading

Key	Value
<id>	4:9b20e604-b39b-4e17-87db-1f1a4d68ccf4:284
doc_id	"doc_1"
id	"doc_1_ssh_6"
order	6
page_num	20
title	"Member Secretary I/c"

Started streaming 25 records after 31 ms and completed after 33 ms.

Chunk Node

neo4j\$ MATCH (n:Chunk) RETURN n LIMIT 25;

GraphTableRAW

agenda_extract...

agenda_extract...

agenda_extract...

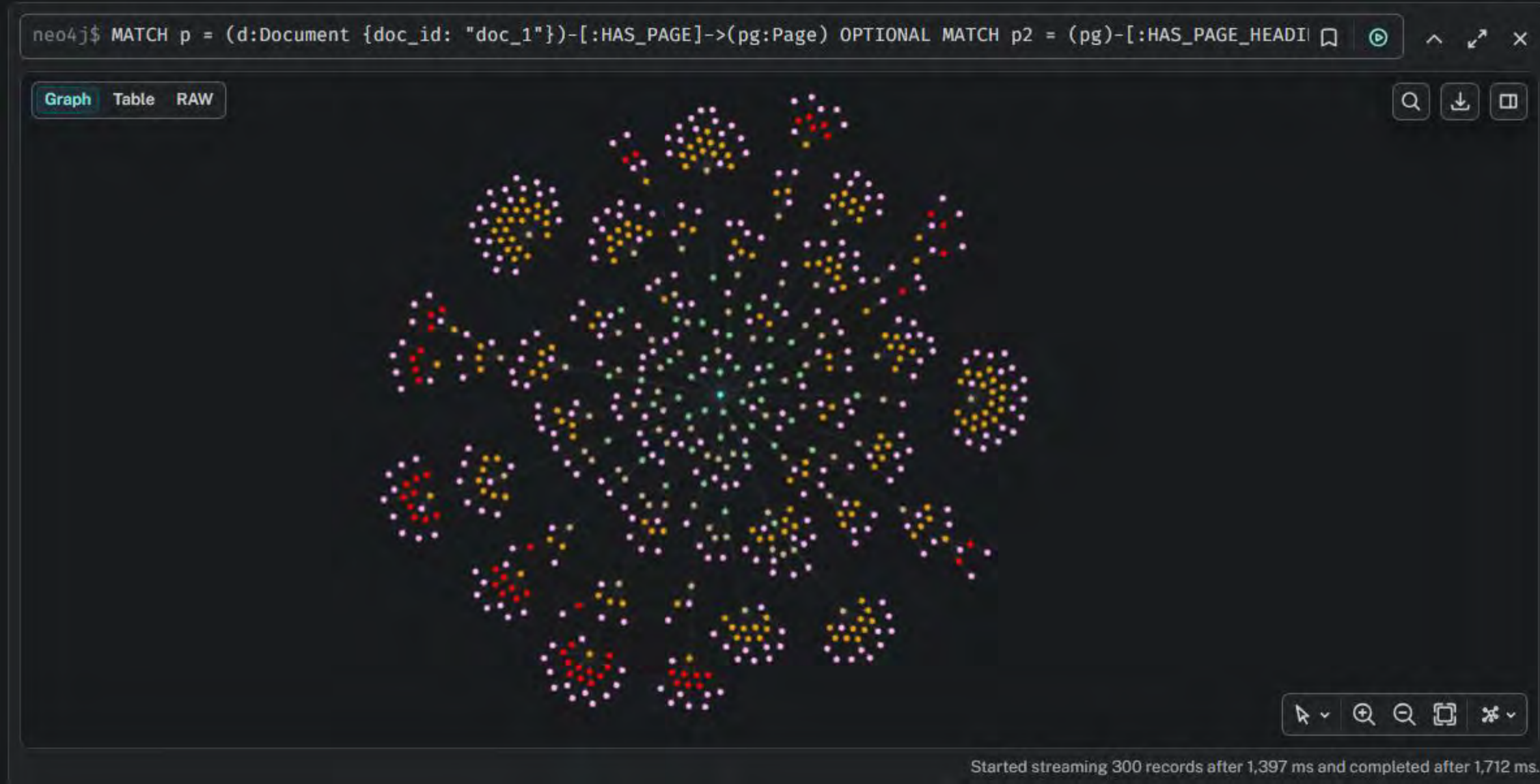
Node details

Chunk

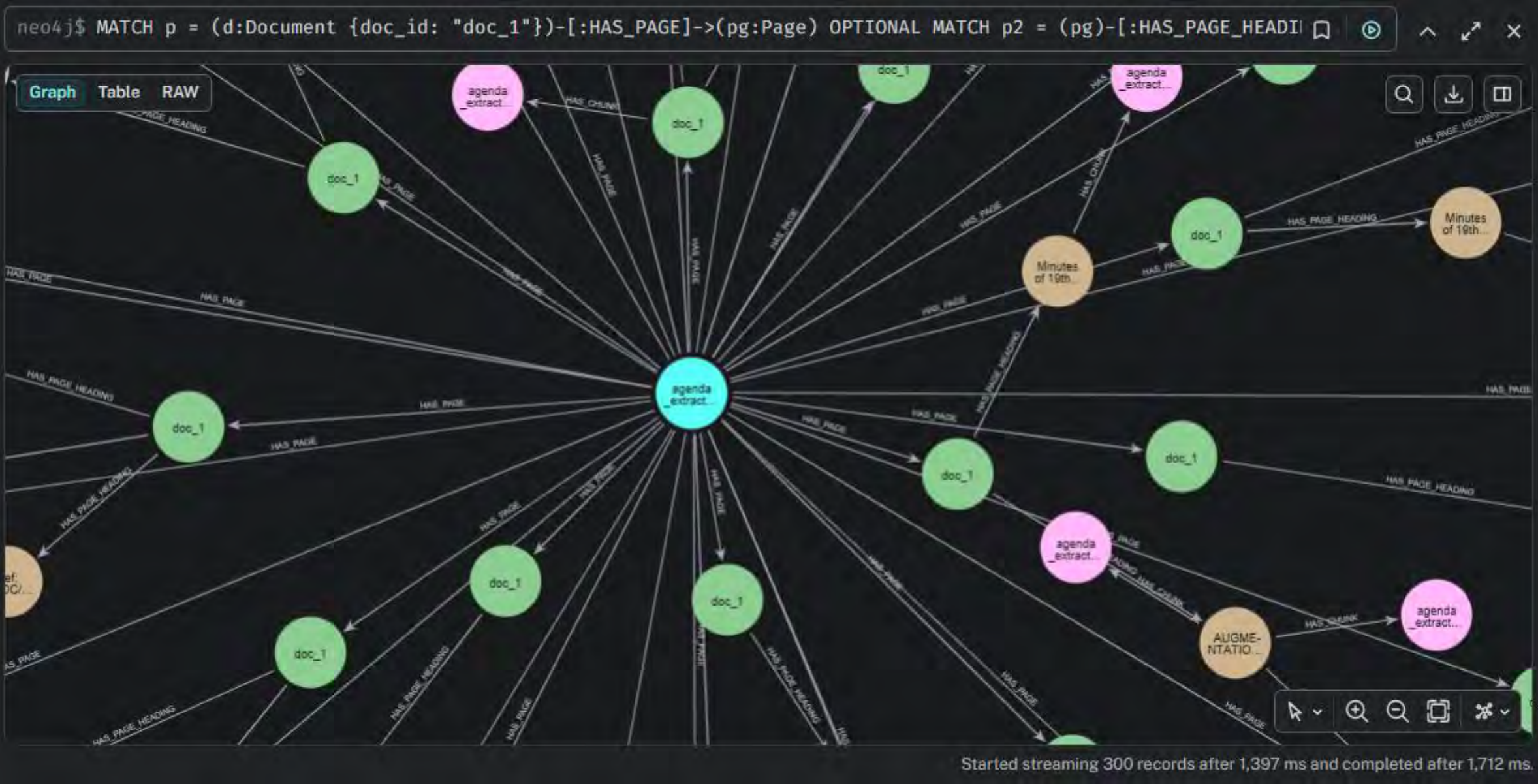
Key	Value
<id>	4:9b20e604-b39b-4e17-87db-1f1a4d68ccf4:12
chunk_id	"doc_1_chunk_6"
doc_id	"doc_1"
file_name	"agenda_extracted_filtered.txt"
order	6
page_heading	"SOUTHERN REGIONAL POWER COMMITTEE"
page_num	2
section_heading	"CONFIRMATION OF THE MINUTES OF THE 18th MEETING"

Started streaming 25 records after 28 ms and completed after 28 ms

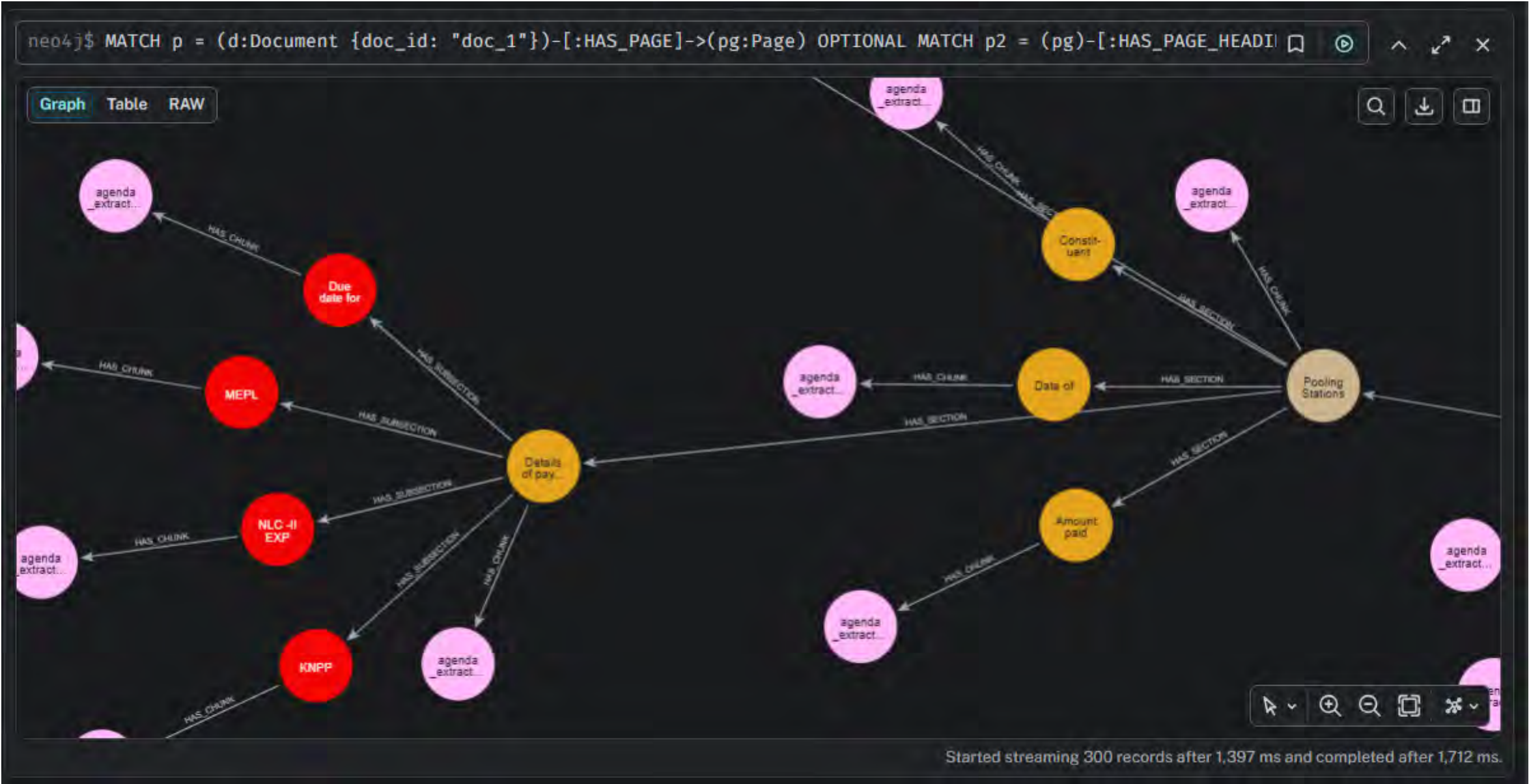
Single Doc Relationship



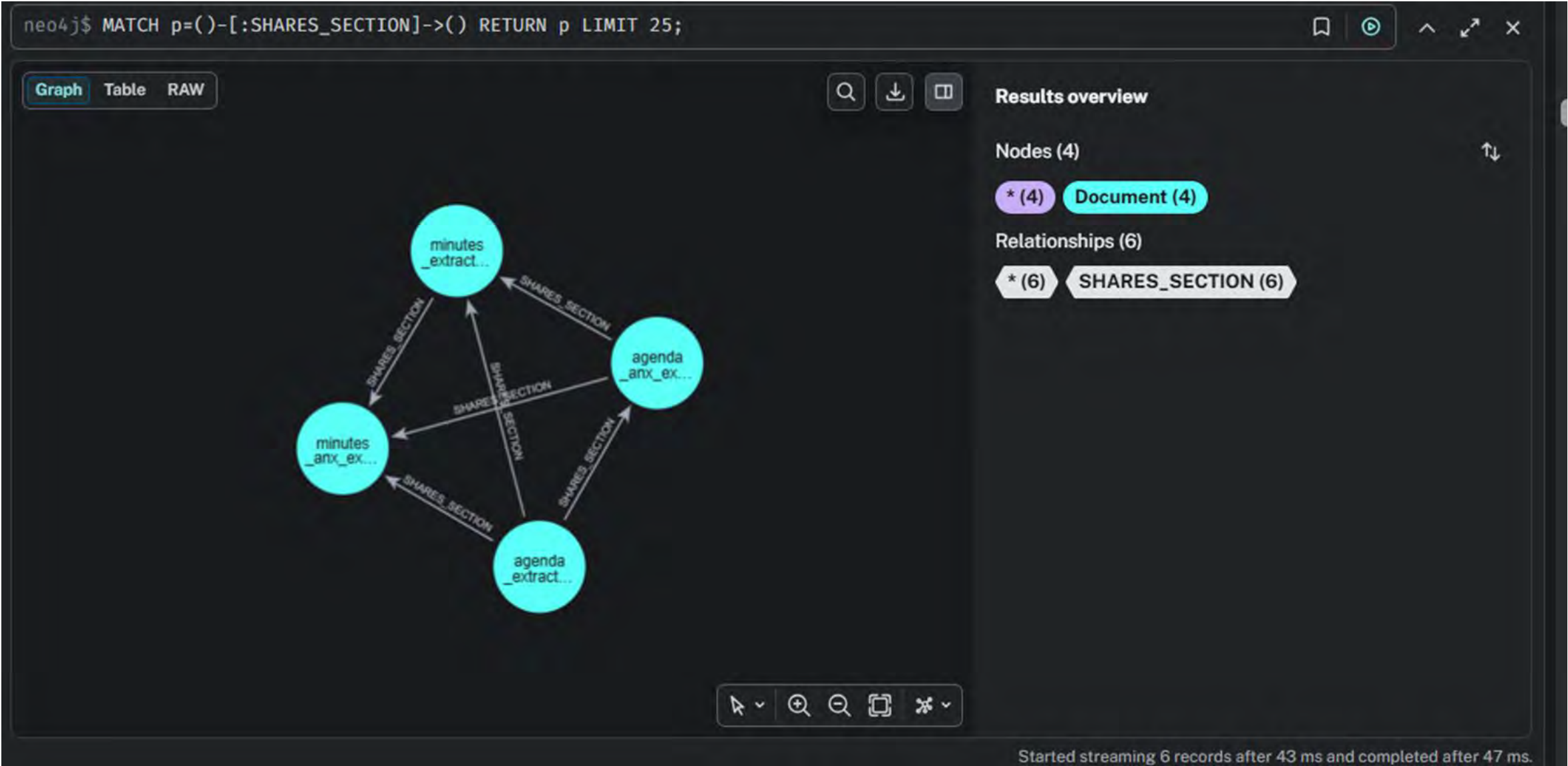
Single Doc Relationship (Zoomed)



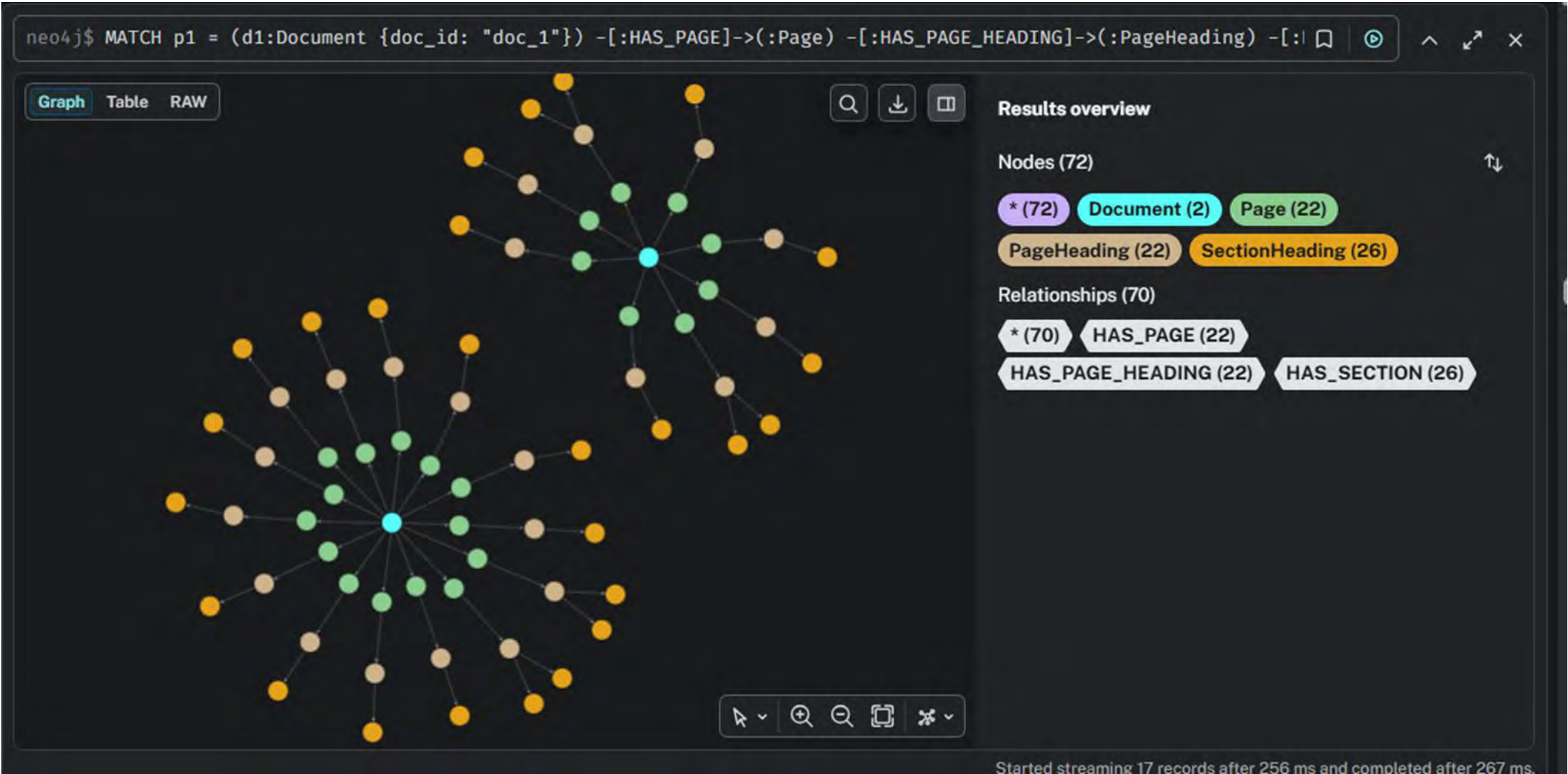
Single Doc Relationship (Zoomed)



Showing relationship between the docs with common headings



Docs with common headings



PowerGPT development Approach – Step-4



Types of Q&A Generated

❑ Factual :

```
{
  "qid": "agenda_anx_extracted_filtered_fact_69cfcd0d",
  "question": "What is the date of the SRPC meeting about Simhadri STPS schedules?",
  "answer": "07.11.2014",
  "type": "fact",
  "gold_doc_id": "agenda_anx_extracted_filtered",
  "gold_file_name": "agenda_anx_extracted_filtered.txt",
  "gold_chunk_ids": [
    "agenda_anx_extracted_filtered_chunk_1",
    "agenda_anx_extracted_filtered_chunk_2",
    "agenda_anx_extracted_filtered_chunk_7"
  ],
  "gold_page_nums": [
    1,
  ]
}
```

□ Date and Time :

```
{
  "qid": "agenda_anx_extracted_filtered_date_or_time_84e78a30",
  "question": "What was the period from which Simhadri STPS schedules were discussed?",
  "answer": "12.10.2014 to 25.10.2014",
  "type": "date_or_time",
  "gold_doc_id": "agenda_anx_extracted_filtered",
  "gold_file_name": "agenda_anx_extracted_filtered.txt",
  "gold_chunk_ids": [
    "agenda_anx_extracted_filtered_chunk_1",
    "agenda_anx_extracted_filtered_chunk_2",
    "agenda_anx_extracted_filtered_chunk_7"
  ],
  "gold_page_nums": [
    1,
    2
  ]
}
```

□ Entity-based :

```
'qid': "agenda_anx_extracted_filtered_entity_e83329a1",
'question': "Which organizations participated in the SRPC meeting on Simhadri STPS schedules?",
'answer': "APSLDC, TSSLDC, SRLDC, NTPC, SRPC",
'type': "entity",
'gold_doc_id': "agenda_anx_extracted_filtered",
'gold_file_name': "agenda_anx_extracted_filtered.txt",
'gold_chunk_ids': [
    "agenda_anx_extracted_filtered_chunk_1",
    "agenda_anx_extracted_filtered_chunk_2",
    "agenda_anx_extracted_filtered_chunk_7"
],
'gold_page_nums': [
    1,
    2
```

❑ Numeric :

```
},  
{  
  "qid": "agenda_extracted_filtered_numeric_f02edbf4",  
  "question": "How many members are listed in the meeting of the Commercial Sub-Committee of SRPC?",  
  "answer": "10",  
  "type": "numeric",  
  "gold_doc_id": "agenda_extracted_filtered",  
  "gold_file_name": "agenda_extracted_filtered.txt",  
  "gold_chunk_ids": [  
    "agenda_extracted_filtered_chunk_1",  
    "agenda_extracted_filtered_chunk_2",  
    "agenda_extracted_filtered_chunk_3"  
  ],  
  "gold_page_nums": [  
    1  
  ],  
}
```

□ Event-based :

```
IS_MULTI_DOC : false
},
{
  "qid": "agenda_extracted_filtered_event_bfb75b1c",
  "question": "What was proposed by PGCL vide letter dated 13th July 2012?",
  "answer": "To discuss the issue regarding sharing of cost (approx. € 1.95 Crores) in respect",
  "type": "event",
  "gold_doc_id": "agenda_extracted_filtered",
  "gold_file_name": "agenda_extracted_filtered.txt",
  "gold_chunk_ids": [
    "agenda_extracted_filtered_chunk_119",
    "agenda_extracted_filtered_chunk_123",
    "agenda extracted filtered chunk 124"
```

☐ Cause-effect :

```
'qid': "agenda_anx_extracted_filtered_cause_effect_45d7f325",
'question': "Why has the issue of frequency linked penalty/incentive for over/under drawal of sha
'answer': "Because APTRANSCO, KPTCL, and KSEB failed to indicate a convenient date for conducting
'type': "cause_effect",
'gold_doc_id': "agenda_anx_extracted_filtered",
'gold_file_name': "agenda_anx_extracted_filtered.txt",
'gold_chunk_ids': [
    "agenda_anx_extracted_filtered_chunk_170",
    "agenda_anx_extracted_filtered_chunk_171",
    "agenda_anx_extracted_filtered_chunk_172"
],
'gold_page_nums': [
```

□ Explanation :

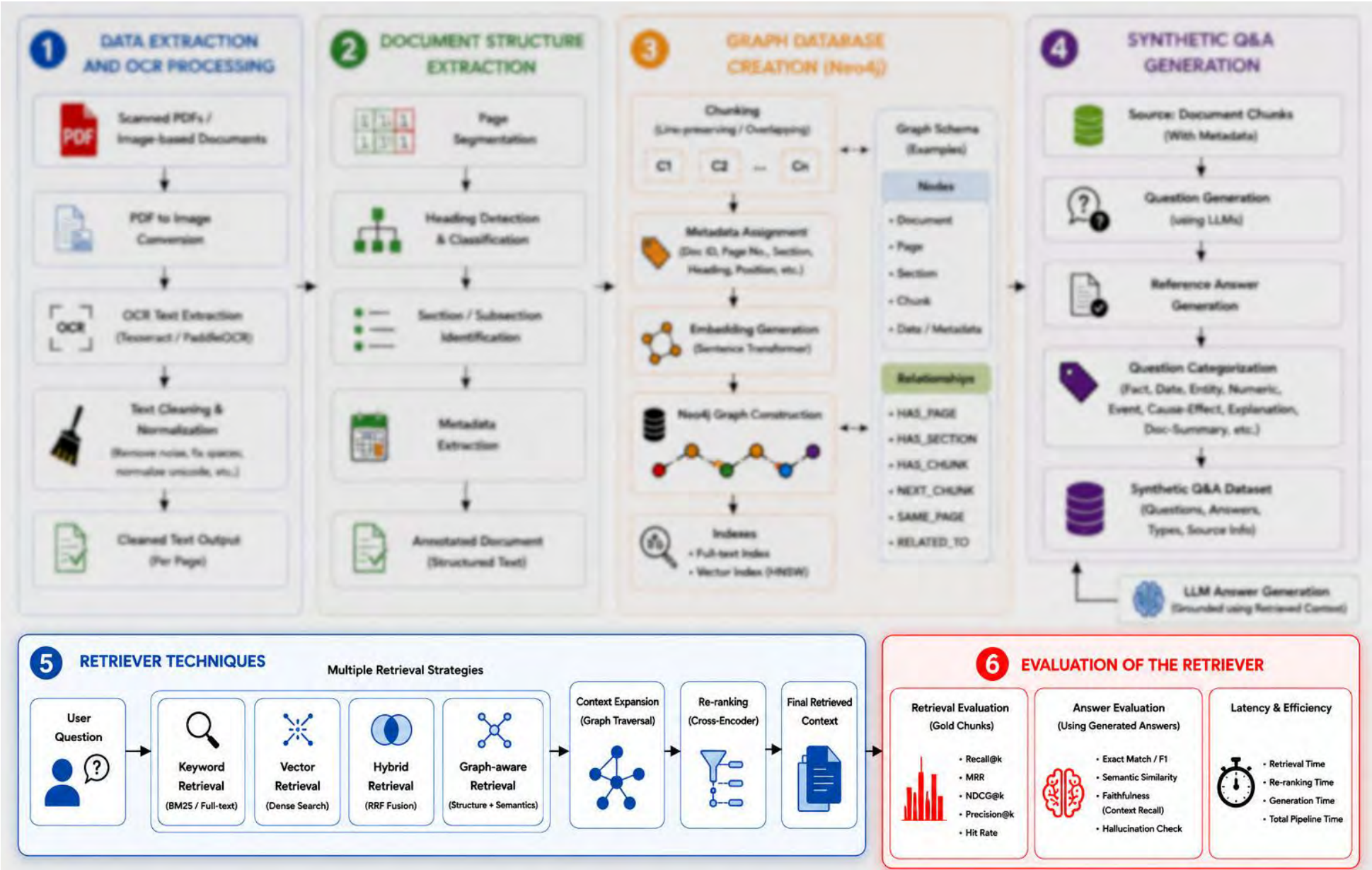
```
'qid': "agenda_anx_extracted_filtered_explanation_18bfc38c",
'question': "Explain why SRLDC is disowning themselves from the issue of frequency linked penalty",
'answer': "SRLDC is disowning themselves from the issue because they are the nodal agency for the",
'type': "explanation",
'gold_doc_id': "agenda_anx_extracted_filtered",
'gold_file_name': "agenda_anx_extracted_filtered.txt",
'gold_chunk_ids': [
    "agenda_anx_extracted_filtered_chunk_170",
    "agenda_anx_extracted_filtered_chunk_171",
    "agenda_anx_extracted_filtered_chunk_172",
    "agenda_anx_extracted_filtered_chunk_174",
    "agenda_anx_extracted_filtered_chunk_177"
],
'gold_page_nums': [
    27,
    28
```

□ Document summary :

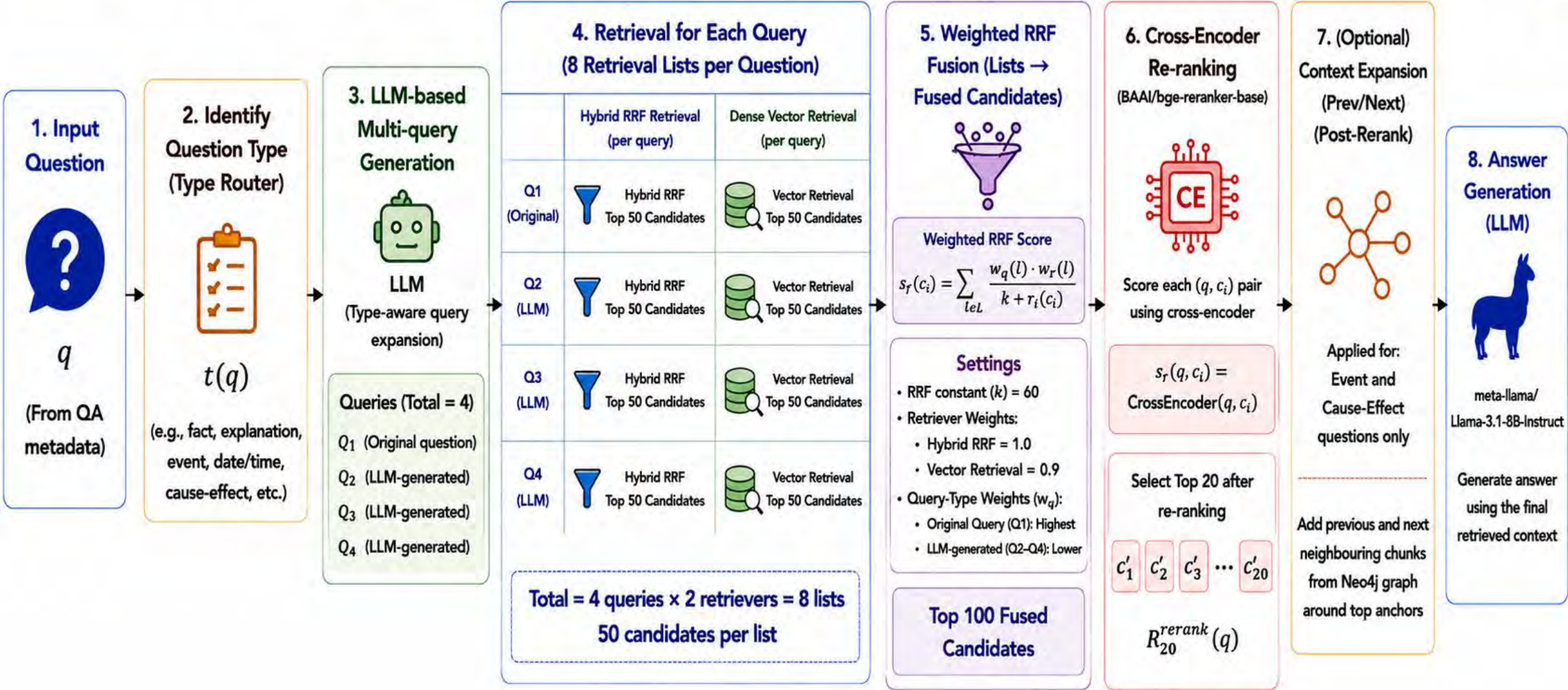
```
ti_doc": false

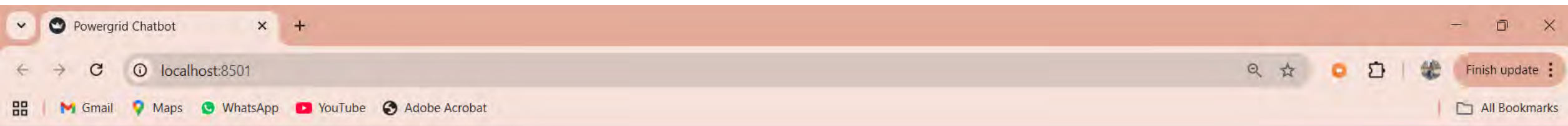
"agenda_anx_extracted_filtered_document_summary_0494a108",
on": "What are the main issues and decisions discussed in the document regarding the Simhadri STPS and related matters?",
": "The main issues discussed include scheduling issues of NTPC, Simhadri STPS, the effect of 'Hudhud' cyclone on Simhadri
"document_summary",
oc_id": "agenda_anx_extracted_filtered",
ile_name": "agenda_anx_extracted_filtered.txt",
hunk_ids": [
da_anx_extracted_filtered_chunk_1",
da_anx_extracted_filtered_chunk_2",
da_anx_extracted_filtered_chunk_7",
da_anx_extracted_filtered_chunk_9",
```

PowerGPT development Approach – Step-5 and Step-6



LLM-BASED MULTI-QUERY FUSION PIPELINE (Type-Aware Multi-Query Retrieval + Fusion + Re-ranking)





Powergrid Chatbot



Reason for NLC one day delay in UPI payments?



The reason for NLC one day delay in UPI payments was due to the 'Thane Cyclone' effect.

Some of the results :

Here we can see that the RAG model is giving the details of the meeting.

Answer from RAG :

```
answer = rag_chain.invoke("When and where 19th meeting of the commercial sub-committee of srpc held")  
print(answer)
```

Pyth

The 19th meeting of the commercial sub-committee of SRPC was held on 16th July 2012 at Bangalore.



When and where 19th meeting of the commercial sub-committee of srpc held?



The 19th meeting of the commercial sub-committee of SRPC was held on 16th July 2012 at Bangalore.

Powergrid Chatbot

localhost:8501

Gmail

Maps

WhatsApp

YouTube

Adobe Acrobat

Finish update

All Bookmarks

Deploy

Due to hudhud cyclone, which transmission lines were effected ?

Due to Hudhud cyclone, the following transmission lines were affected:

1. CTU Lines:

SIMHADRI - GAZUWAKA-I at 08:51:00 Hrs

SIMHADRI - GAZUWAKA-I at 09:59:00 Hrs

VEMAGIRI - SIMHADRI-I at 10:06:00 Hrs

VEMAGIRI - SIMHADRI-II at 10:37:00 Hrs

JEYPORE - GAZUWAKA-I at 10:16:00 Hrs

JEYPORE - GAZUWAKA-II at 10:30:00 Hrs

HVDC GAJUWAKA POLE-I at 10:30:00 Hrs

HVDC GAJUWAKA POLE-II at 10:30:00 Hrs

2. State Owned Lines:

KALPAKKA - GAZUWAKA-I at 10:30:00 Hrs

KALPAKKA - GAZUWAKA-II at 08:37:00 Hrs

KALPAKKA - VEMAGIRI-I & II at 10:06:00 Hrs

KHAMMAM - KALPAKKA-I & II at 10:11:00 Hrs

SIMHADRI - KALPAKKA-II at 10:22:00 Hrs

SIMHADRI - KALPAKKA-III at 10:06:00 Hrs

SIMHADRI - KALPAKKA-IV at 10:06:00 Hrs

Give me the list of participants 19th meeting of the commercial sub-committee of srpc held at bangalore on 16.07.2012

Ask a question...



Give me the list of the name of participants 19th meeting of the commercial sub-committee of srpc held at bangalore on 16.07.2012



The list of participants of the 19th meeting of the commercial sub-committee of SRPC held at Bangalore on 16.07.2012 is not available in the given context. However, the list of participants of the 23rd meeting of the commercial sub-committee of SRPC held at Bangalore on 04.12.2013 is available in the context.

The list of participants of the 23rd meeting of the commercial sub-committee of SRPC held at Bangalore on 04.12.2013 is as follows:

1. ANDHRA PRADESH

- Shri B.Bhanu Prasad Divisional Engineer (Comml.), APPCC

2. KARNATAKA

- Shri R.V.Dilip Kumar Executive Engineer (E), SLDC, KPTCL
- Shri M.Appaji Controller of Accounts, TBC, KPTCL
- Shri V.Bharatheesha Rao Deputy Director, PCKL

3. KERALA STATE ELECTRICITY BOARD

- Shri George V.James Executive Engineer, SLDC
- Shri Satish Kumar G Executive Engineer (Comml.)

4. TANTRANSCO Ltd.

- Shri N.Mahendiran Financial Controller
- Shri P.Raja Gunanidhi Executive Engineer (Grid)
- Shri P.Murugavelan Asst. Executive Engineer (Grid)

5. ELECTRICITY DEPARTMENT, PUDUCHERRY

Power system stability assessment: Machine Learning Applications

Gurunath Gurrula
Ankita Gupta, P.S.Sastry



Multi-Swing Instability - Prediction

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

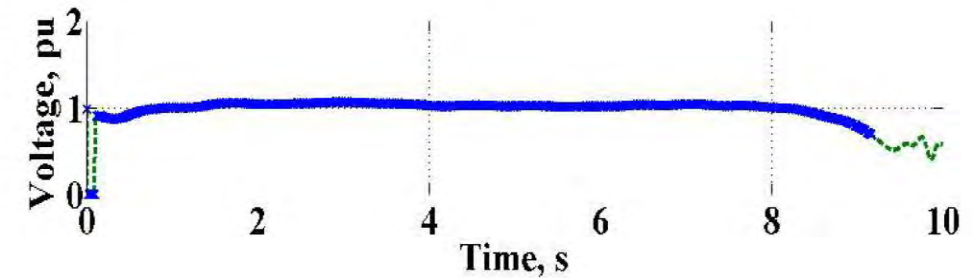
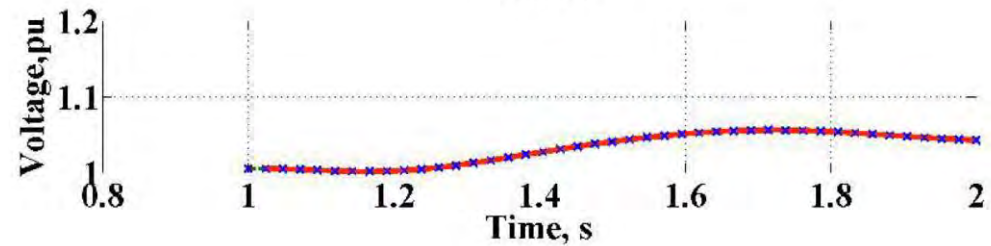
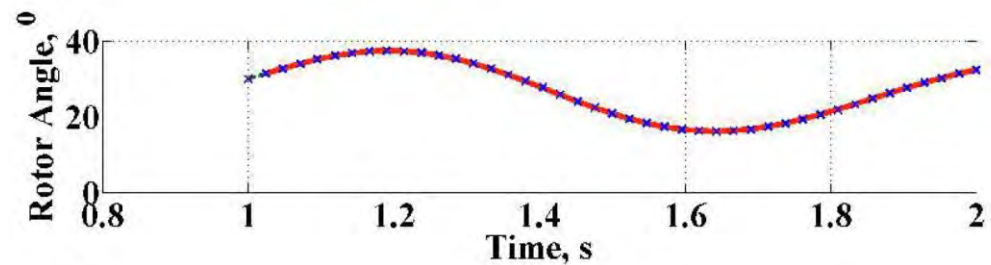
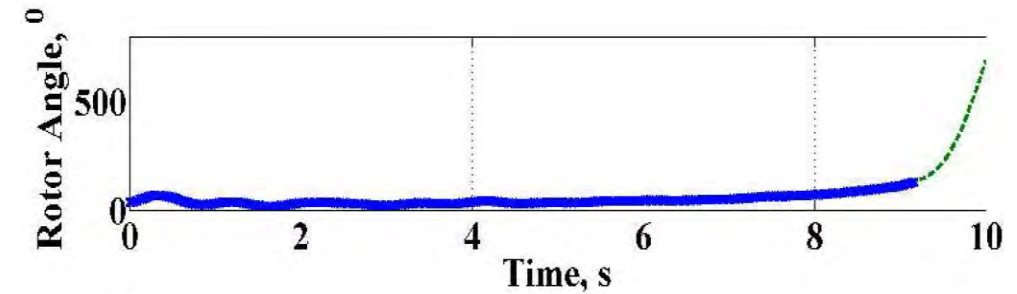
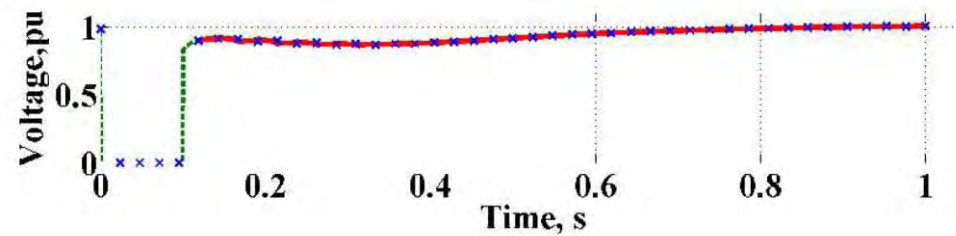
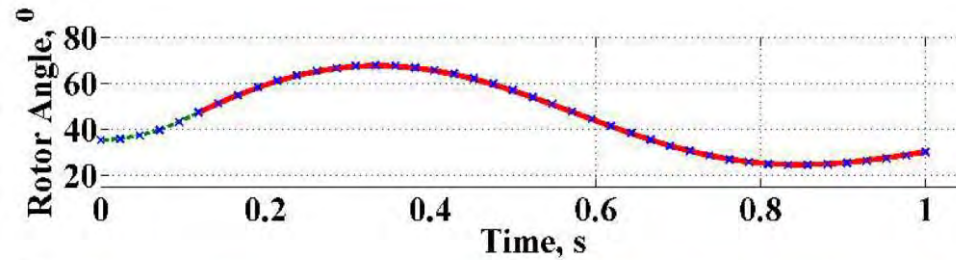
Data Generation

Assessment

Results

Conclusion

References



We aim at **Early Prediction of such Multi-Swing Instability** which can lead to cascading effects and a possible blackout.

Available Methods: Time Domain Simulations

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

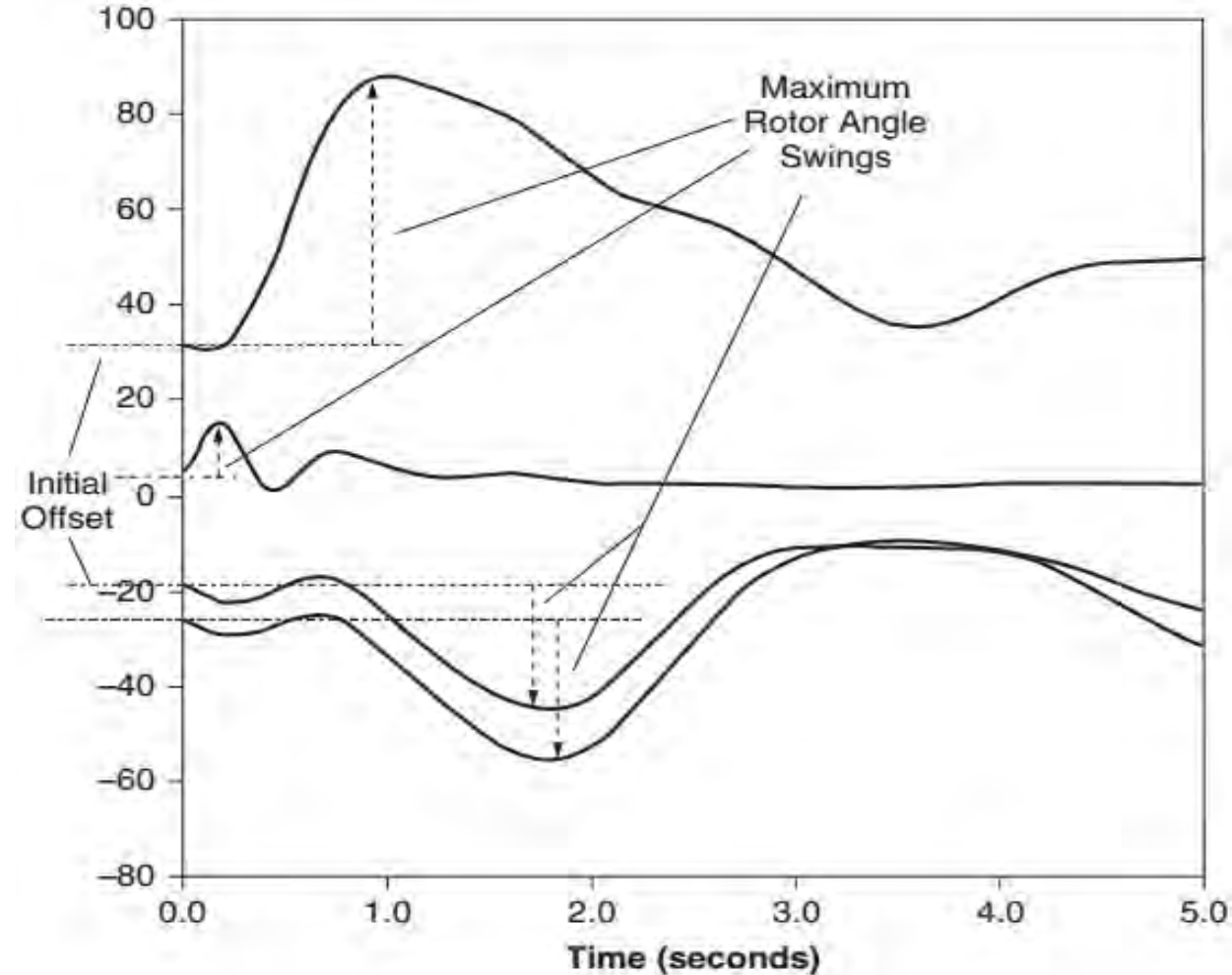
Data Generation

Assessment

Results

Conclusion

References

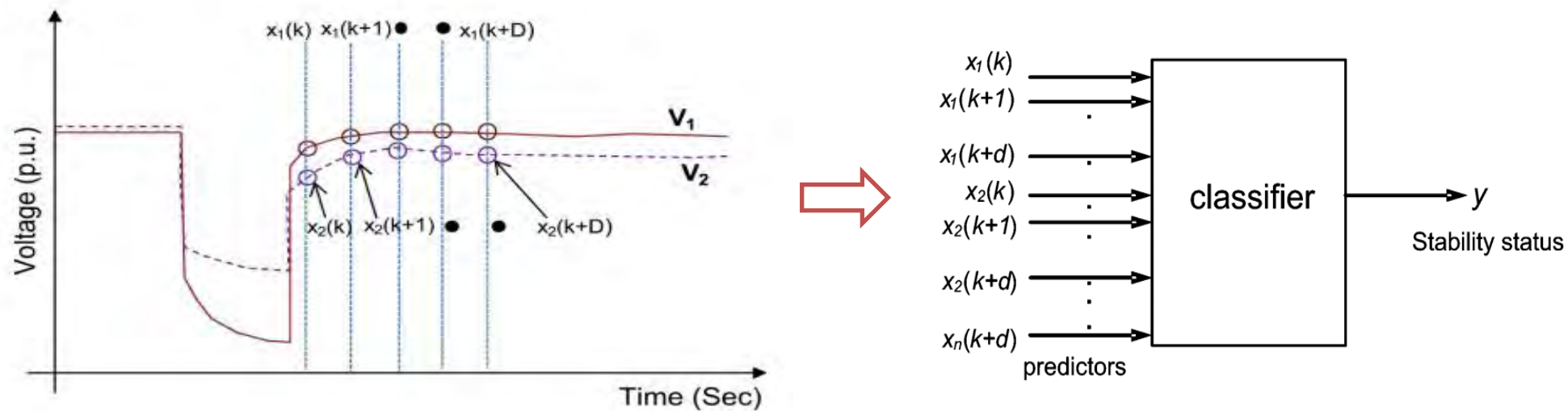


- × Demands accurate information about network configuration
- × Computationally intensive with high time complexity

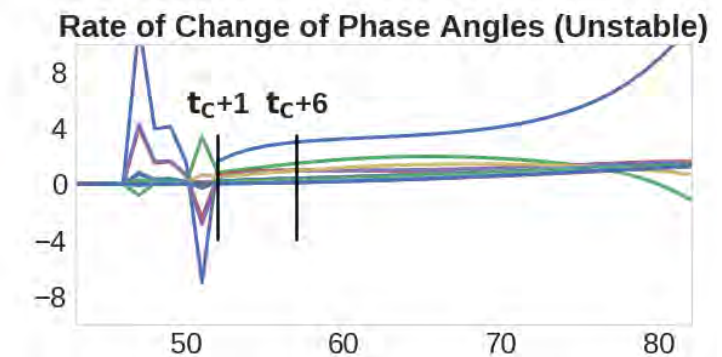
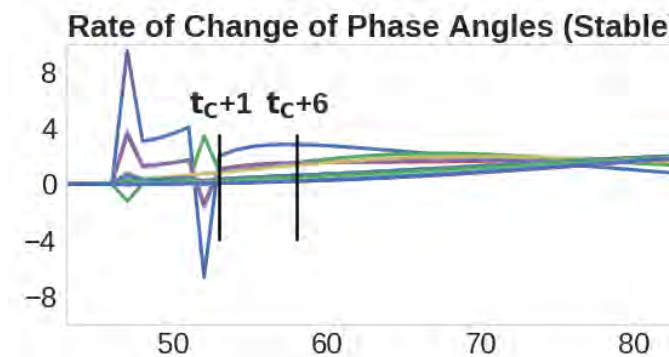
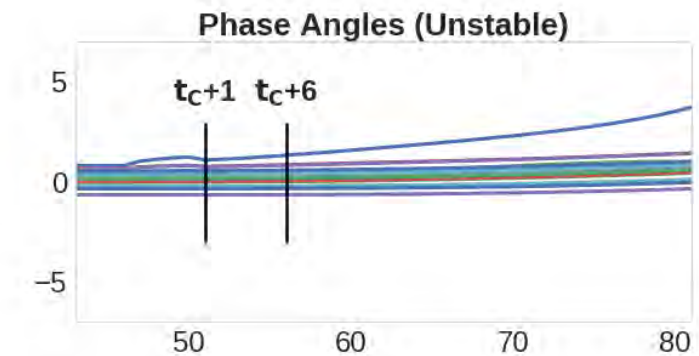
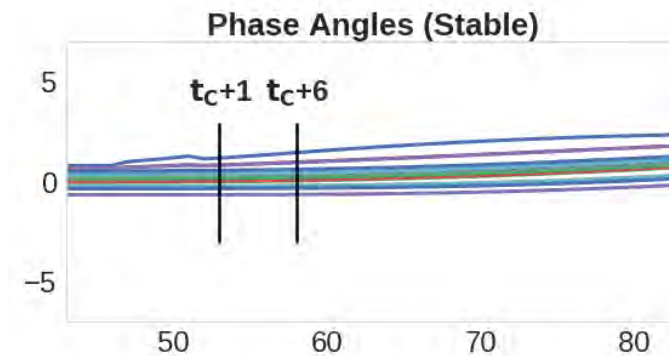
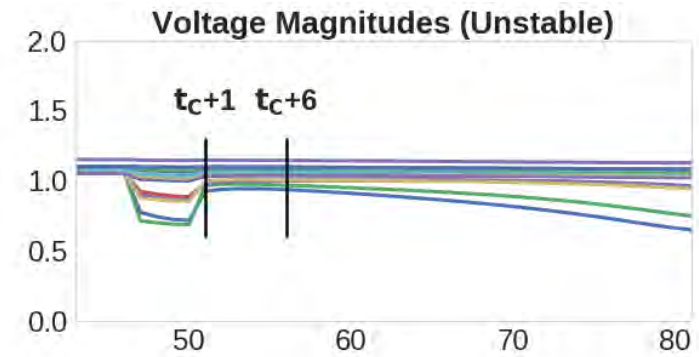
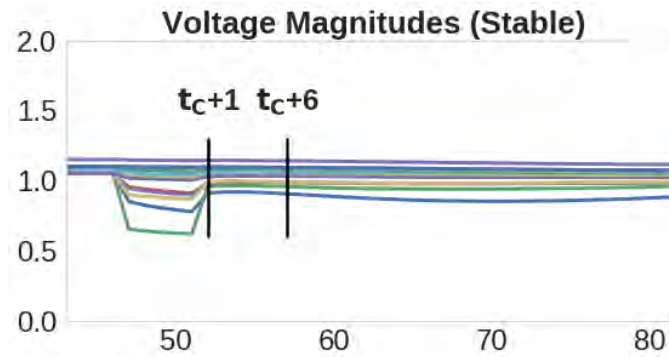
Available Methods: Machine Learning Based

- Neural Networks [Jensen *et al.*, 2001]
- Decision Trees [Kamwa *et al.*, 2009]
- Support Vector Machines [Rajapakse *et al.*, 2010]
- Ensemble Methods [Zhou *et al.*, 2016] SVMs outperform all other existing ML Techniques.

Synchronized Measurements available using Phasor Measurement Units (PMU)



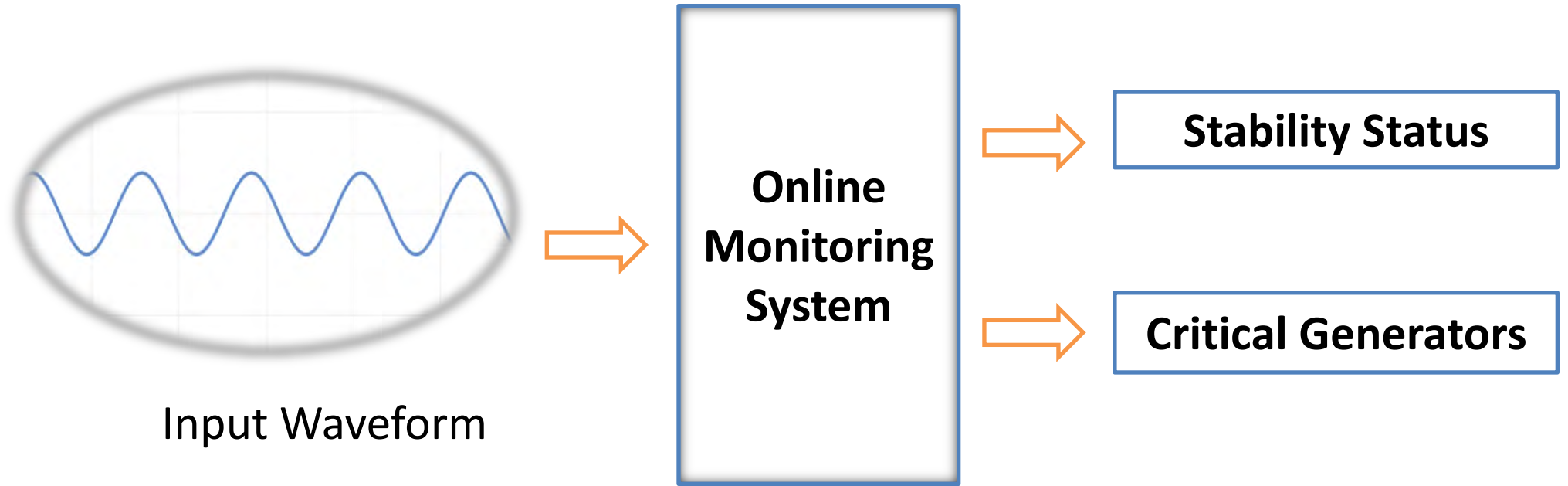
× **Accurate Fault Information is Presumed.**



While the instability is easy to see if we wait long enough, predicting within the interval shown is what is challenging.

Problem Statement

- Introduction
- Literature Review
- Problem Statement
- Online Monitoring System
- Classifiers
CNN
- Critical Generator
PCA
Y-Net
- Data Generation
- Assessment
- Results
- Conclusion
- References



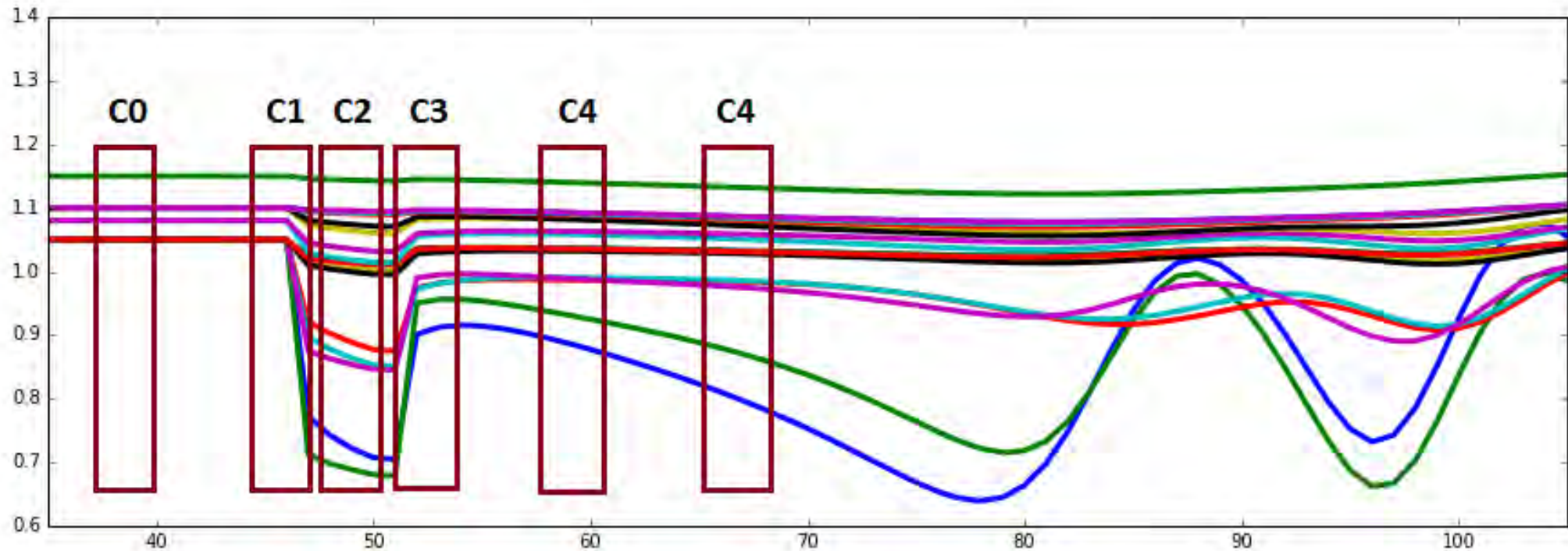
- Evaluates Power System Stability Status.
- Generates an **Alarm** to be used as an early warning to the Operator, in case of Impending Transient Instability.
- Identifies the group of **Critical Generators**.

Accurate fault information may not be feasible in practical scenario.
This System doesn't need any such information.

Online Monitoring System

Input wave is divided into windows.

Each window corresponds to one of the five identified states of Power System.



C0 : No Disturbance/ Stable State

C1 : Fault Occurrence

C2 : Fault Duration

C3 : Fault Clearance

C4 : Alert/ Unstable State

Architecture

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

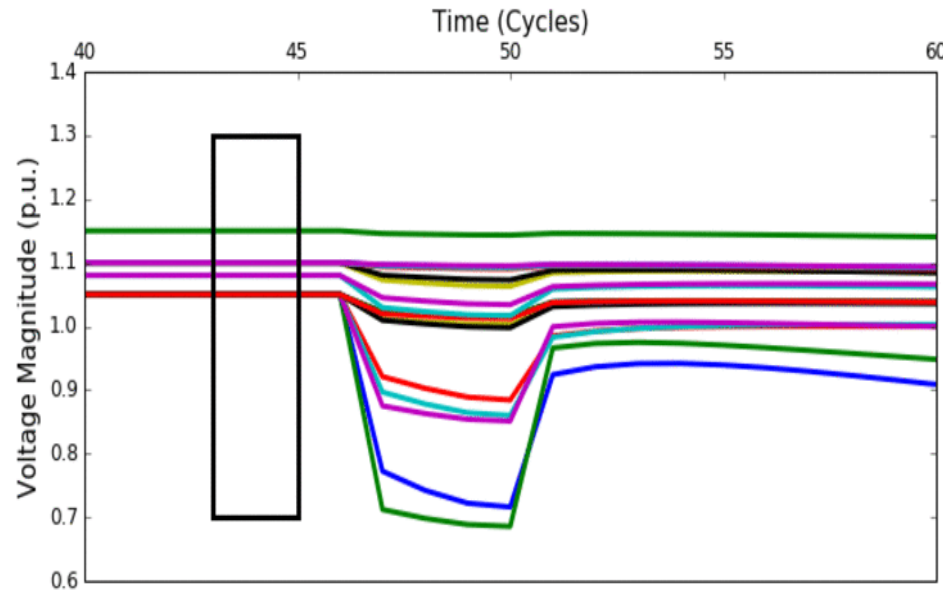
Data Generation

Assessment

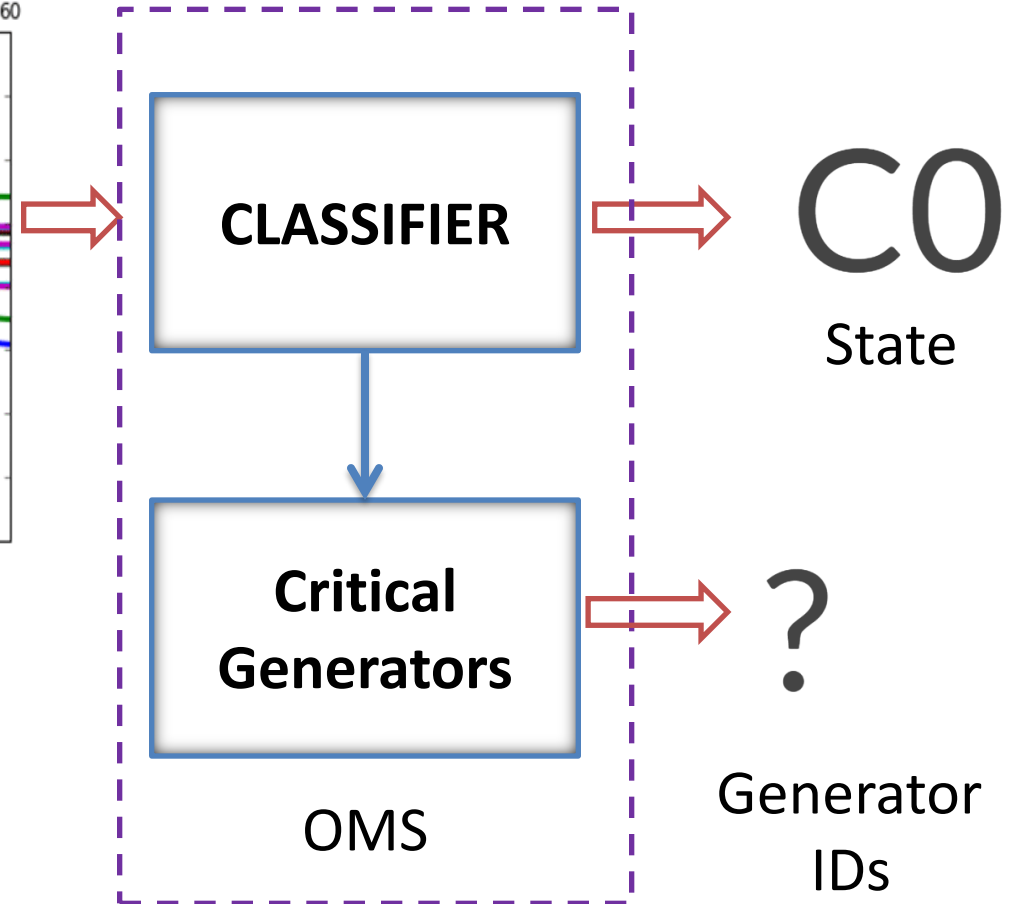
Results

Conclusion

References

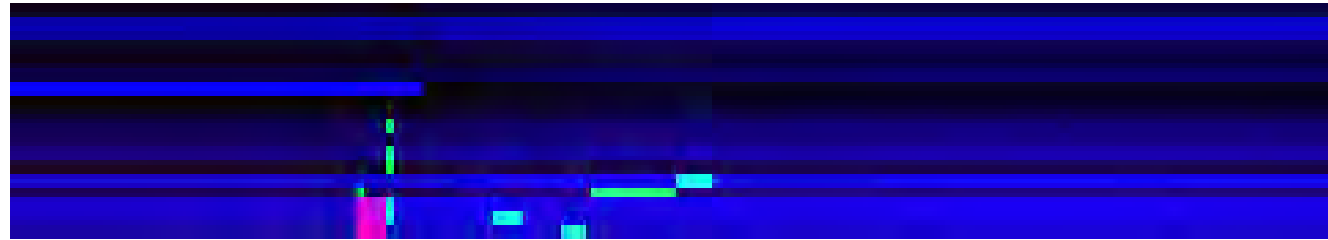


Sliding window

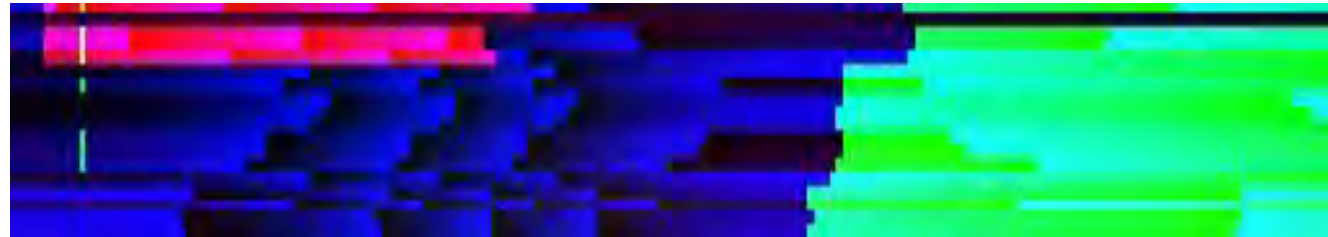


Motivation for Convolutional Neural Networks

Heatmap : Three network variables are viewed as R-G-B channels of an image. These images reveal Interesting Patterns.



Heatmap for a Stable Case



Heatmap for an Unstable Case



Characteristic pattern present only in Unstable Cases.

Convolutional Neural Networks

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

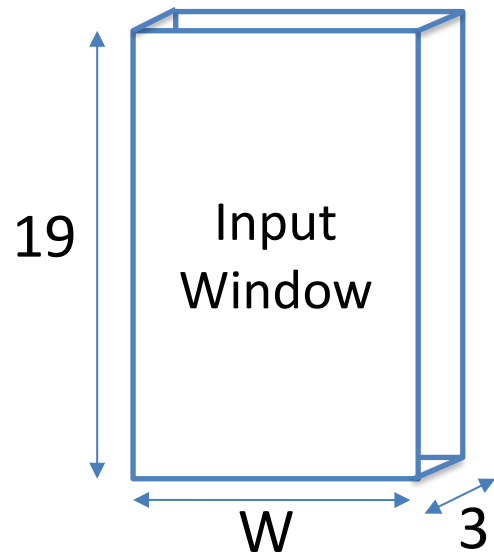
Assessment

Results

Conclusion

References

Input Data



Heatmap of
Input Window

Details of Network

Input Size 19x5x3

2D Convolutional Layer 32 Kernels of size 19 x 3 x 3

1D Convolutional Layer 16 Kernels of size 3 x 3

Fully Connected Layer 256

Fully Connected Layer 32

Fully Connected Layer 5 (Softmax)

Loss Cross Entropy

- Heatmap captures the variations in network variables more effectively.
- CNN extracts the features which help assess transient stability in a more efficient way.

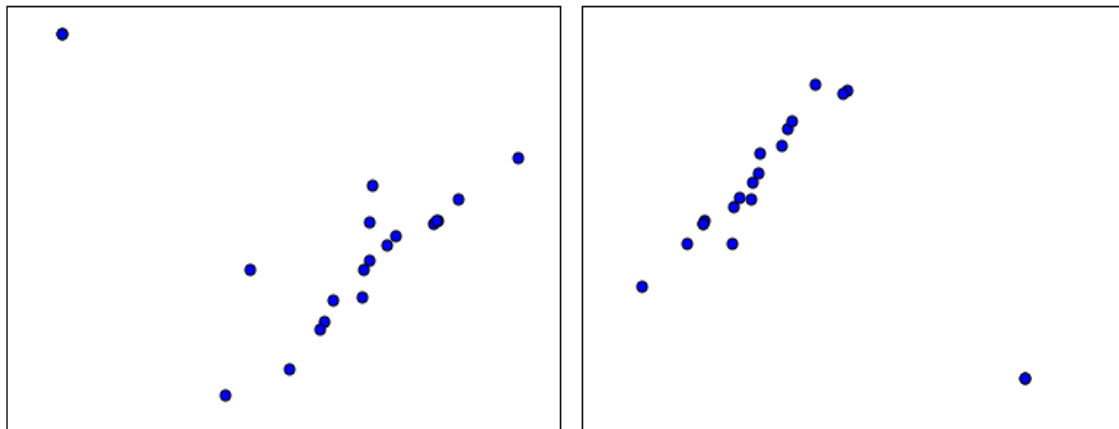
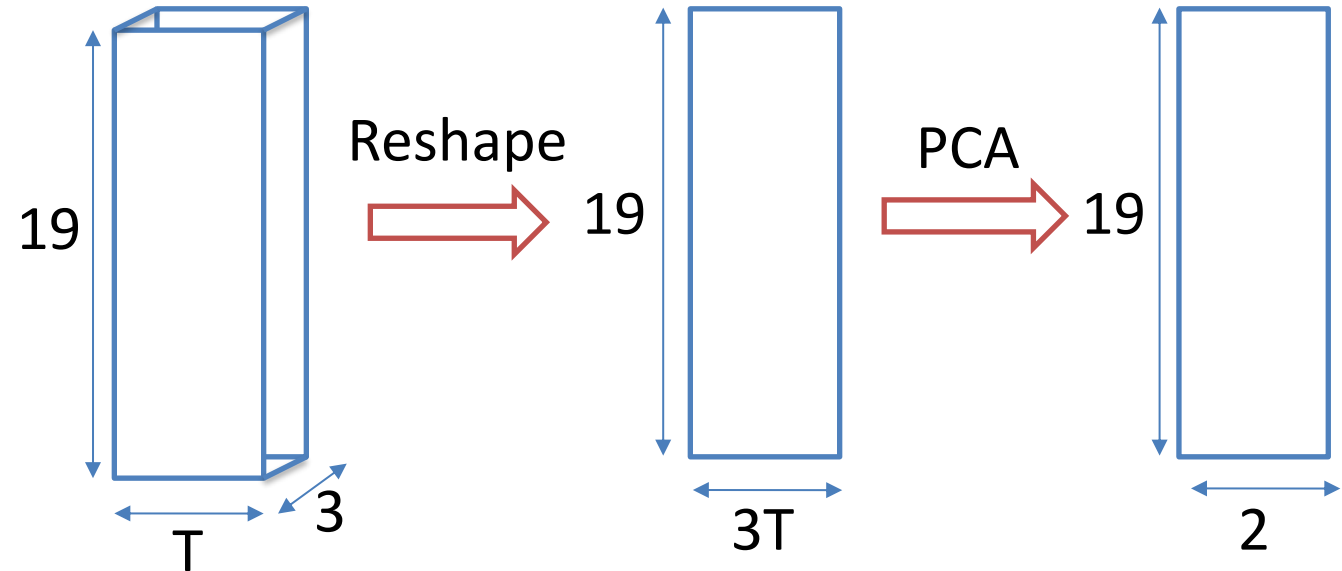
Identification of Most Critical Generator

Projection of measurements after fault detection onto subspace spanned by first two principal components segregates the critical generators from the rest.

T_F : Instance of Fault Detection

T_A : Instance of Alarm

T : $T_A - T_F$



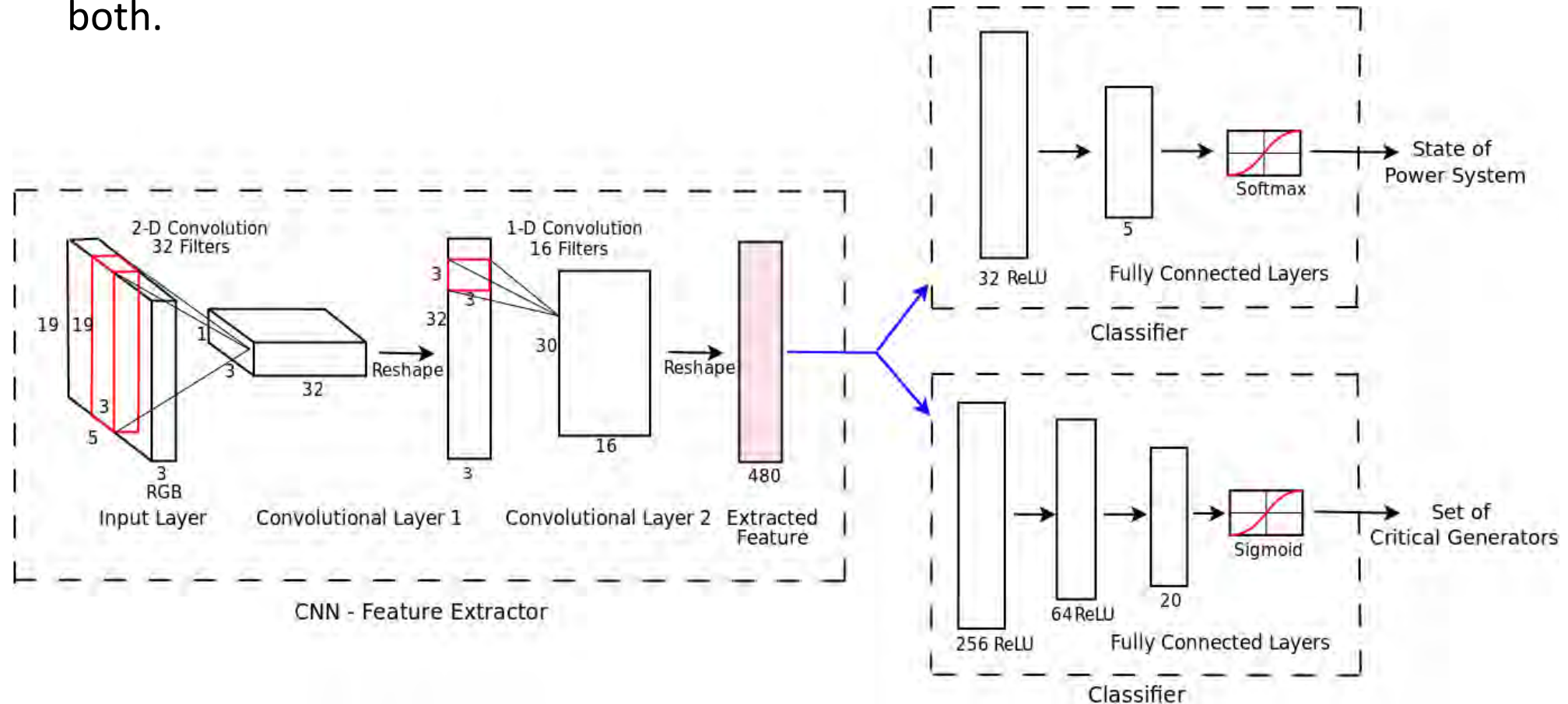
Outlier Detection

- ✓ Identify Outlier by fitting a gaussian density.
- ✓ Parameters : Sample Mean and Minimum Covariance Determinant
- ✓ Point with minimum likelihood is an outlier.

Identification of Group of Critical Generators

Y-Net

- An Integrated framework.
- Detects transient instability in the power system.
- Identifies all the generators that are responsible for such instability.
- The idea is that the learnt representation by CNN should be informative enough for both.



Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

Data Generation

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

– Simulator :

- Generated Temporal Profiles of Power System Variables.

– Network : IEEE 118-bus system, IEEE 145 Bus System

– Contingencies :

- Symmetrical fault on each bus.
- Three locations (25%, 50%, and 75% of the length) on each transmission line.
- Three loading levels (base load + 5%, 7%, 10%).
- Random fault instance with Fault durations from 5 to 8 cycles.

– Measured Features :

- Voltage Magnitudes at Generator Buses
- Voltage Phase Angles $\approx$ Rotor Angles

- Derivative of Phase Angles with respect to time $\approx$ Frequency

– Datasets :

- Time Step to solve ODE (RK-4) : 5 ms
- IEEE 118 Bus
 - ❖ 26,000 Temporal Profiles
 - ❖ 2309 Unstable, 23691 Stable
- IEEE 145 Bus
 - ❖ 33073 Temporal Profiles
 - ❖ 10099 Unstable, 22974 Stable
- The dataset of each system is randomly split into training and test sets 10 times and all results shown are averages over these trials.

Labelling Temporal Profiles

We label each profile using Transient Stability Index.

$$\text{Transient Stability Index, } \eta = \frac{360^\circ - |\Delta\delta|_{max}}{360^\circ + |\Delta\delta|_{max}}$$

$\eta < 0$: Temporal Profile is Unstable.

$|\Delta\delta|_{max}$: Absolute value of maximum angle of separation between two generators post fault .

Critical Generators : The generators having separation more than 360° from rest of the generators (S_i).

Most Critical Generator : The generator that separates at the earliest among the critical generators.

Data Sample : $(X_i, Y_i, S_i) ; i = 1, 2, \dots, N$
 X_i is a 1000 point time series with each element a $3 \times n_g$ dimensional vector (corresponding to magnitude, phase angle and derivative of phase angle of voltages of n_g generators) .
 Y_i is 0/1 depending on sign of η .

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

Labeling Windows

- ❖ Each profile, X_i is converted into a sequence of overlapping windows.
- ❖ We label each window based on what is happening in the system during that window.

Window : $w = [t_s, t_E]$

t_s : Start index of window

t_E : End index of window

t_F : Instance of Fault Occurrence

t_C : Instance of Fault Clearance

Label of window

$$y_w = \begin{cases} 0 & \text{if } (t_E \leq t_F) \text{ or } (t_s \geq t_C \text{ and } \eta > 0) \\ 1 & \text{if } (t_s \leq t_F \leq t_E) \\ 2 & \text{if } (t_F \leq t_s \leq t_E \leq t_C) \\ 3 & \text{if } (t_F \leq t_s \leq t_C \leq t_E) \\ 4 & \text{if } (t_s \geq t_C \text{ and } \eta < 0) \end{cases}$$

0 : No Disturbance/ Stable State

1 : Fault Occurrence

2 : Fault Duration

3 : Fault Clearance

4 : Alert/ Unstable State

An alert will be issued if Class 4 is predicted with >0.9 probability for 15 cycles wait period

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

Assessment of OMS Performance

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

How effectively OMS predicts Instability ?

False Alarm Rate

Missed Detection Rate

How early OMS predicts Instability ?

The time when the OMS predicts class-4 with 0.9 probability for the first time in an unstable profile relative to the fault clearance is also used as a measure,

$$\text{Latency of Prediction} = \Upsilon = \frac{1}{N} \sum_{i=1}^N (T_{\eta}^i - T_A^i)$$

N : Total Number of Temporal Profiles in Test Set

T_{η}^i : The instant at which the transient stability index, η , becomes negative.

T_A^i : The instant at which OMS predicts instability.

How accurately OMS predicts set of Critical Generators ?

$$\text{Jaccard Similarity} = J(S_{actual}, S_{predicted}) = \frac{|S_{actual} \cap S_{predicted}|}{|S_{actual} \cup S_{predicted}|}$$

S_{actual} : The actual set of Critical Generators.

$S_{predicted}$: The set of Critical Generators predicted by OMS.

Results: IEEE 118 Bus System

Y-NET BASED OMS PERFORMANCE

Case	Noise Level (Type I)	FA Rate %	MD Rate %	Υ (cycles)	ζ (cycles)	JS acc. %
IEEE-118 Bus System						
Base Topology	None	0.41	0.22	30.12	4.67	85.31
	$\pm 1\%$	0.82	0.51	25.58	6.88	86.13
	$\pm 2\%$	1.97	0.51	26.12	6.98	85.46
	$\pm 3\%$	2.48	0.57	26.88	7.37	85.69
Topology Change	None	0.95	0.3	20.43	4.96	81.81
	$\pm 1\%$	1.01	0.55	19.42	5.39	79.45
	$\pm 2\%$	2.39	0.60	18.58	6.26	78.94
	$\pm 3\%$	2.69	0.80	18.37	6.44	78.29
Generator Parameters Change	None	0.96	0.51	15.36	4.75	81.21
	$\pm 1\%$	1.04	1.01	14.71	5.19	81.01
	$\pm 2\%$	1.35	1.11	14.68	5.77	77.67
	$\pm 3\%$	1.94	1.31	14.65	5.81	77.27

Results: IEEE 145 Bus System

Introduction

Literature Review

Problem Statement

Online Monitoring System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

Y-NET BASED OMS PERFORMANCE

Case	Noise Level (Type I)	FA Rate %	MD Rate %	Υ (cycles)	ζ (cycles)	JS acc. %
IEEE-145 Bus System						
Base Topology	None	1.12	0.04	120.59	7.37	89.22
	$\pm 1\%$	1.85	0.15	118.64	8.42	86.49
	$\pm 2\%$	1.89	0.18	119.48	8.86	86.37
	$\pm 3\%$	1.93	0.18	112.02	9.02	86.17
Topology Change	None	1.14	0.35	100.38	7.53	77.38
	$\pm 1\%$	1.15	0.36	93.08	7.98	73.21
	$\pm 2\%$	1.18	0.48	92.23	8.48	69.64
	$\pm 3\%$	1.44	0.48	91.09	8.99	66.31
Generator Parameters Change	None	1.56	0.04	103.83	7.53	81.83
	$\pm 1\%$	2.82	0.15	102.18	7.95	78.40
	$\pm 2\%$	3.03	1.01	101.19	8.45	76.45
	$\pm 3\%$	3.72	1.18	100.05	8.88	73.58

Results: Missing Measurements

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

- (i). measurements for random 10 % of generators (145 Bus System : 5 Gens and 118 Bus System : 2 Gens) are missing with 0.1 probability at every time instance;
- (ii). measurements for any one of the stable generators (whether the case is actually stable/unstable) is missing at all time instances;
- (iii). measurement for any one of the unstable generators is missing at all time instances.

Type of Generator Missing	FA Rate %	MD Rate %	Υ (cycles)	JS Accuracy %
IEEE-118 Bus System				
Random 10%	1.41	0.33	44.56	83.73
Stable	1.75	0.60	42.21	82.33
Unstable	1.85	0.93	42.06	68.93
IEEE-145 Bus System				
Random 10%	1.22	0.09	138.62	89.06
Stable	2.64	0.31	137.09	87.15
Unstable	2.74	0.64	129.67	83.57

Impact of Confidence level and Wait Period

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

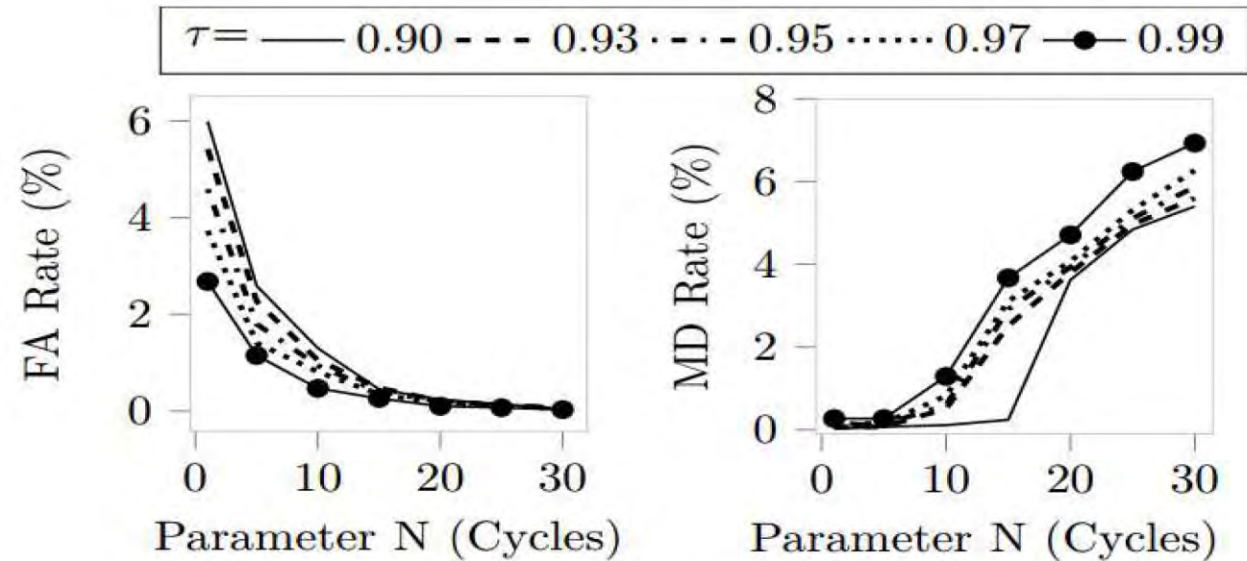
Data Generation

Assessment

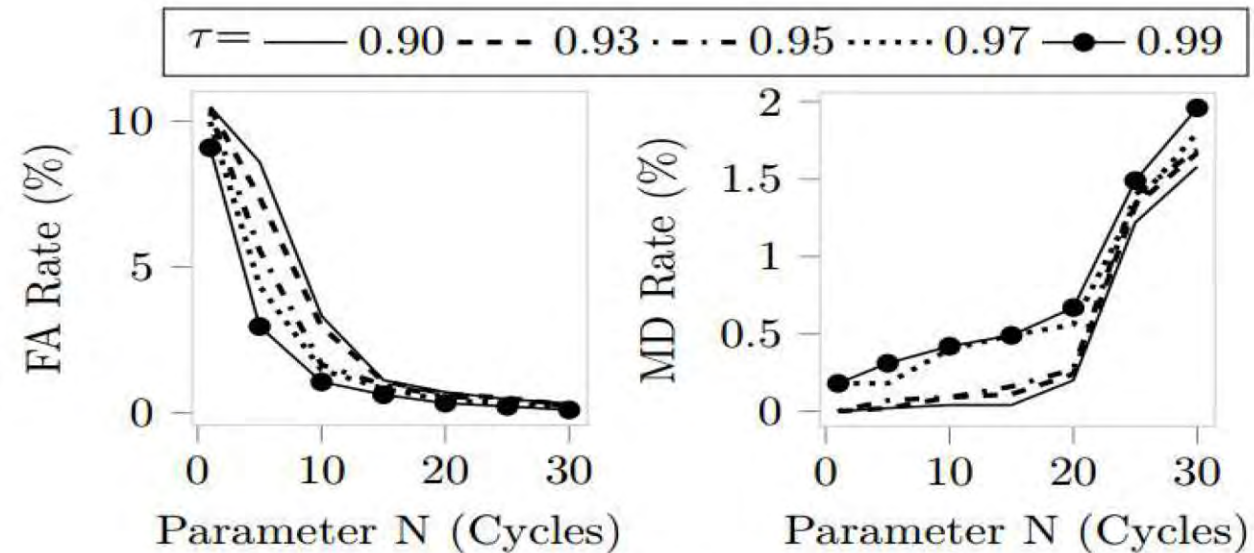
Results

Conclusion

References



(a) 118 Bus System



(b) 145 Bus System

Identification of Most Critical Generator using PCA & Y-Net

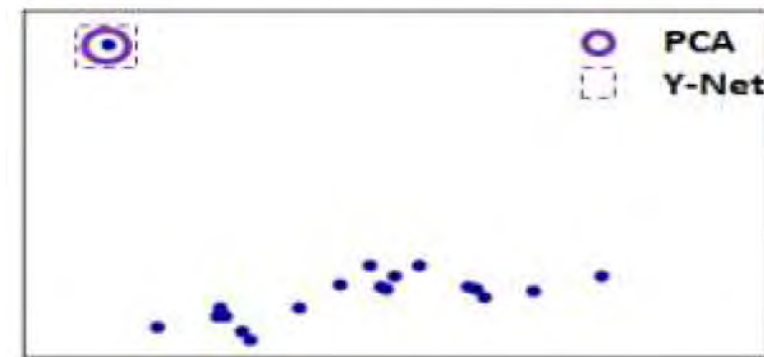
Projection of measurements after fault detection onto subspace spanned by first two Principal Components segregates the critical generators from the rest



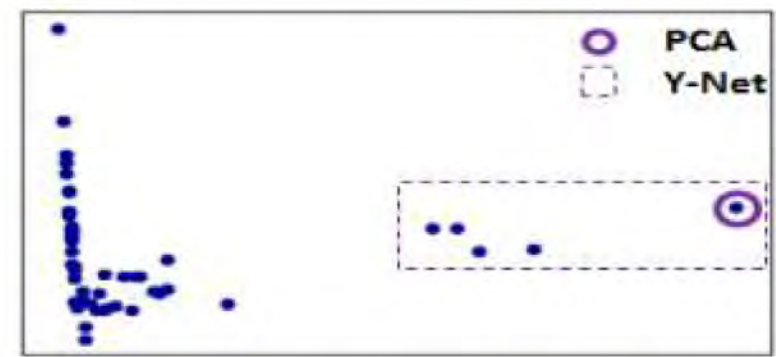
(a) 118 Bus : Unstable Case 1



(b) 118 Bus : Unstable Case 2



(c) 145 Bus : Unstable Case 1



(d) 145 Bus : Unstable Case 2

Strengths and Limitations of OMS

Strengths

- Requires **No Fault Information**.
- **Robust** to Noise, Topology Changes and Generator Parameter Changes.
- **Detects all critical generators** ($JS = 1$) in 86% of unstable cases even under Noise.
- **Predicts instability ahead of its manifestation** in system, allowing operator to take control actions.

Limitations

- Assumes **measurements** from all generators are **available at a central location** synchronously. Can be addressed by availability of PMUs.
- The **architecture of OMS** classifier and computational cost incurred is **System Specific**. For example, input size for 118-Bus (19 Generators) system is $19 \times 5 \times 3$ and for 145-Bus system (50 Generators) it is $50 \times 5 \times 3$.

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

References

Conclusion

Introduction

Literature Review

Problem Statement

Online Monitoring
System

Classifiers
CNN

Critical Generator
PCA
Y-Net

Data Generation

Assessment

Results

Conclusion

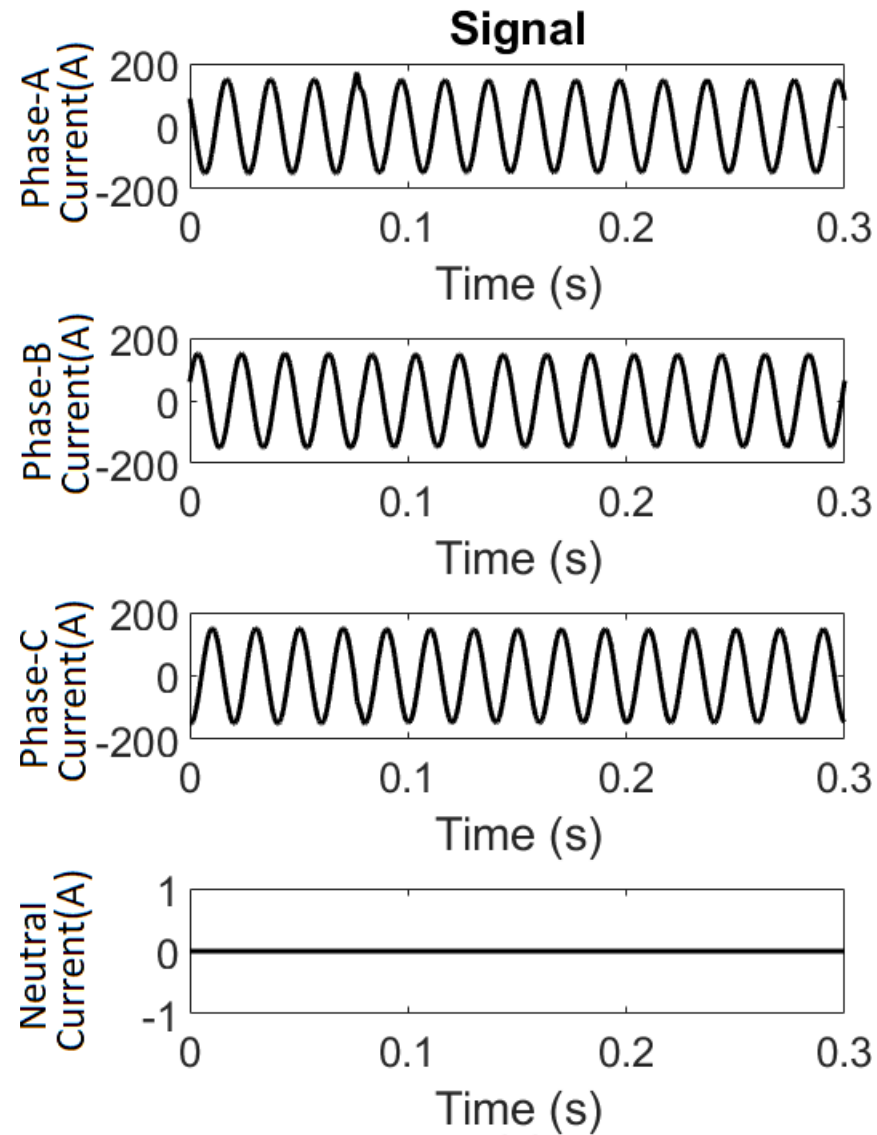
References

- Developed **Deep-architectures** for **predicting** imminent **transient instability** in a power system.
- Demonstrated how **existing techniques** like **SVM fail** when exact fault information is not known.
- The proposed OMS requires **NO Fault Information** for Instability Prediction (First step in this direction to the best of our knowledge).
- The proposed architectures are **robust** to noisy measurements, topology changes and generator parameter changes.
- **PCA based visualization** is effective in detecting most critical generator in 85% of unstable cases.
- **Y-Net** can identify entire group of critical generators accurately in about 86% of unstable cases. It is also robust to all variations possible.

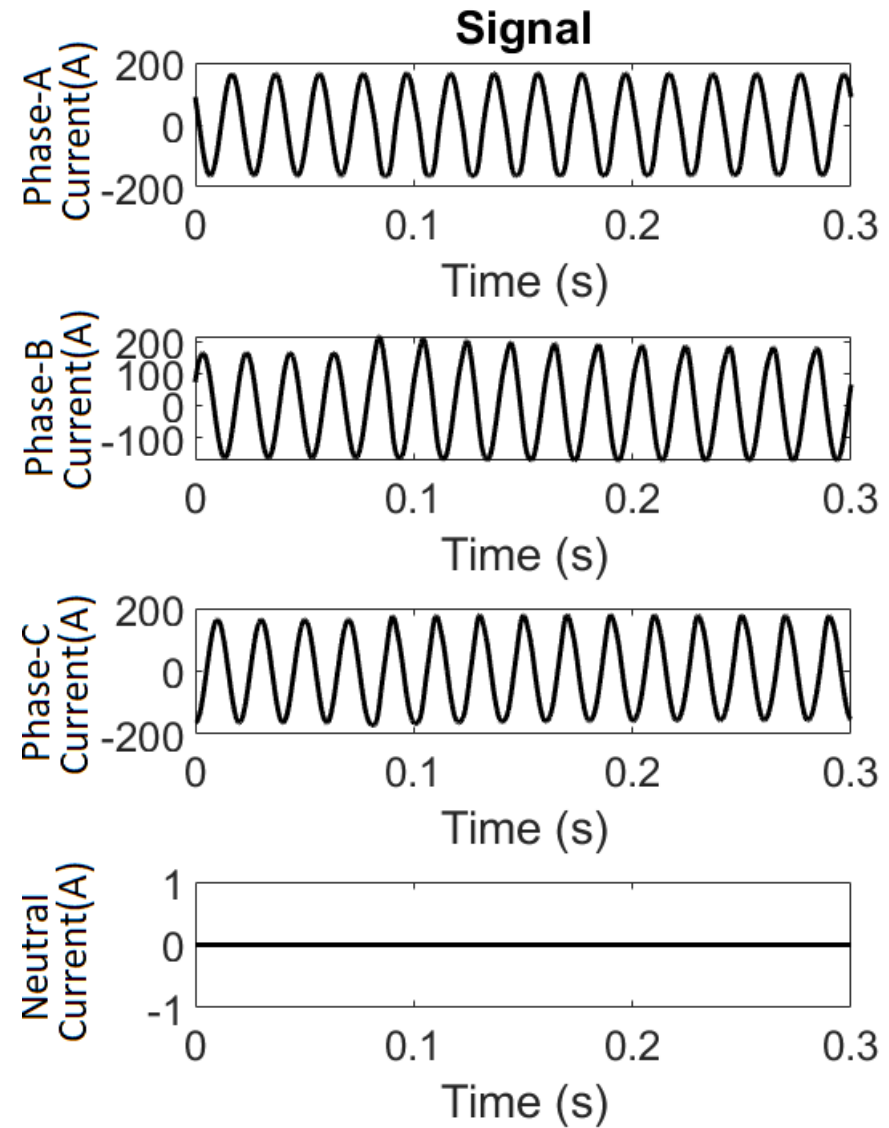
Ankita Gupta, Gurunath Gurrula and P.S. Sastry, "An Online Power System Stability Monitoring System Using Convolutional Neural Networks," in *IEEE Transactions on Power Systems*, October 2018

Symbolic dynamic filtering-based Transient Event Detection and Classification

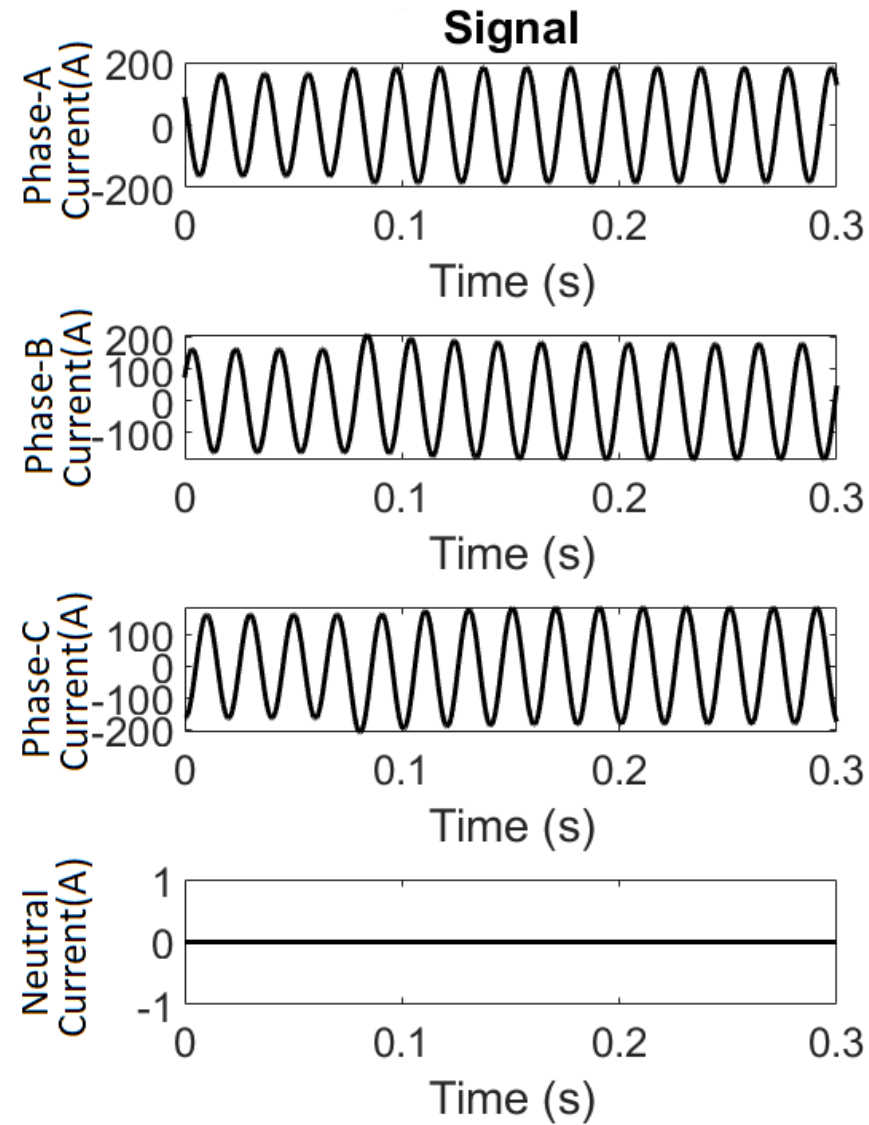
Event Detection: 100 kVAR Capacitor bank switching



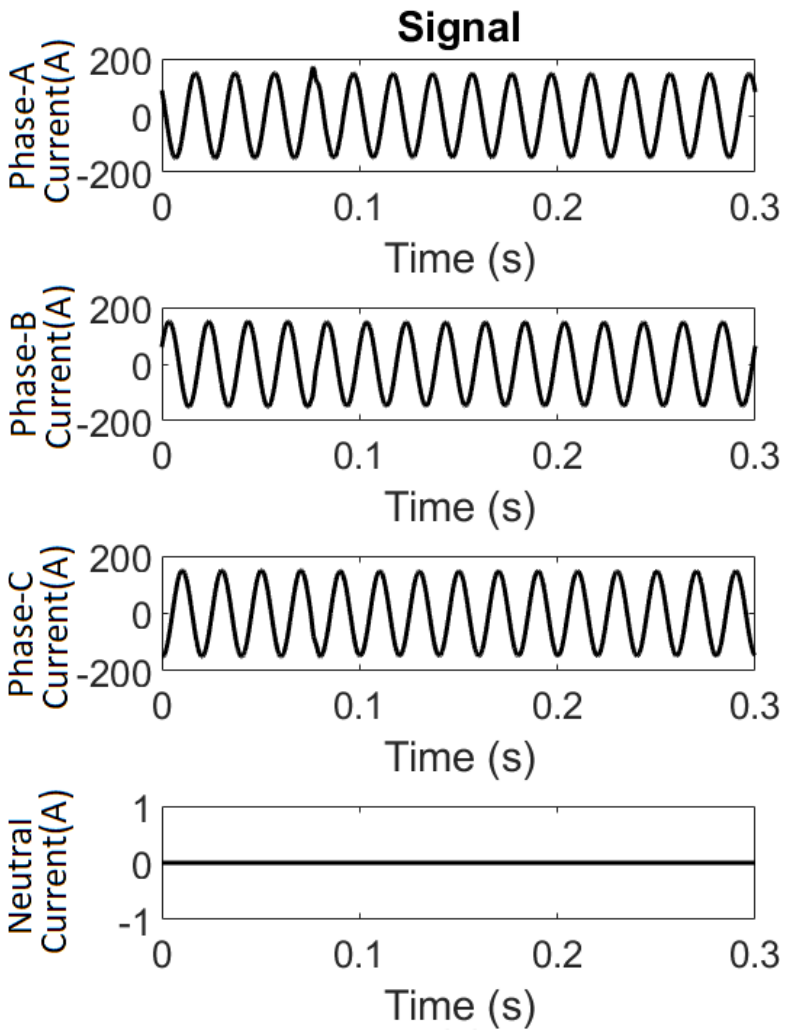
Transformer switching disturbance



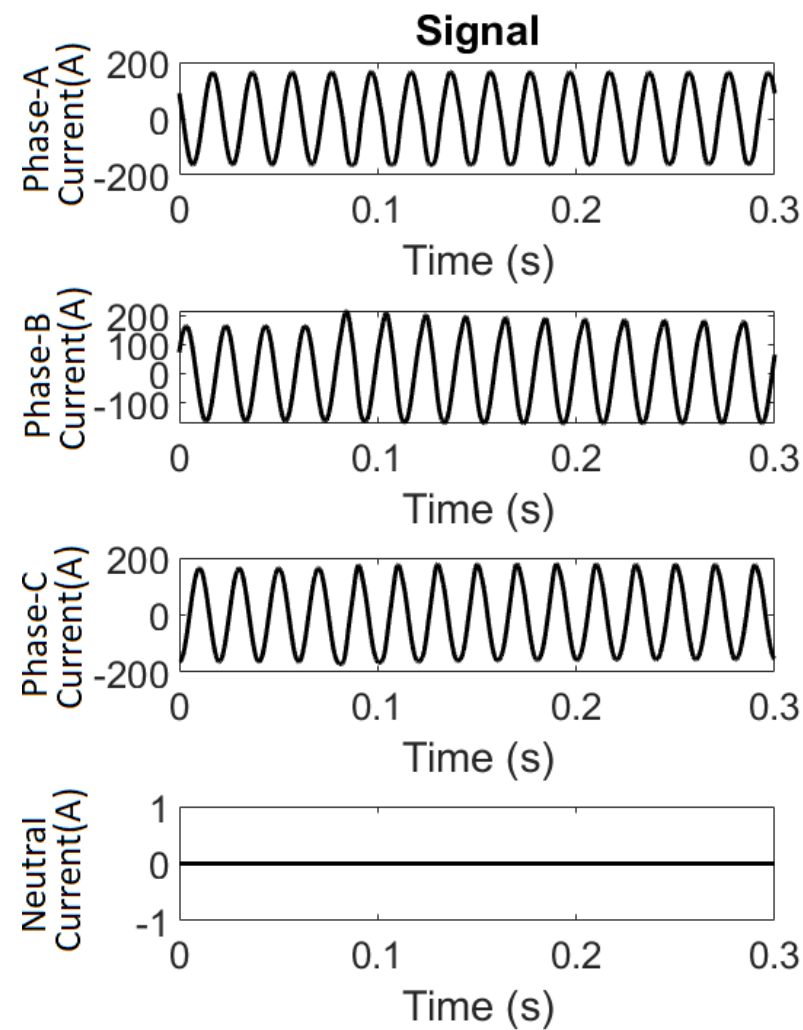
19 kW Electric Arc Furnace Switching



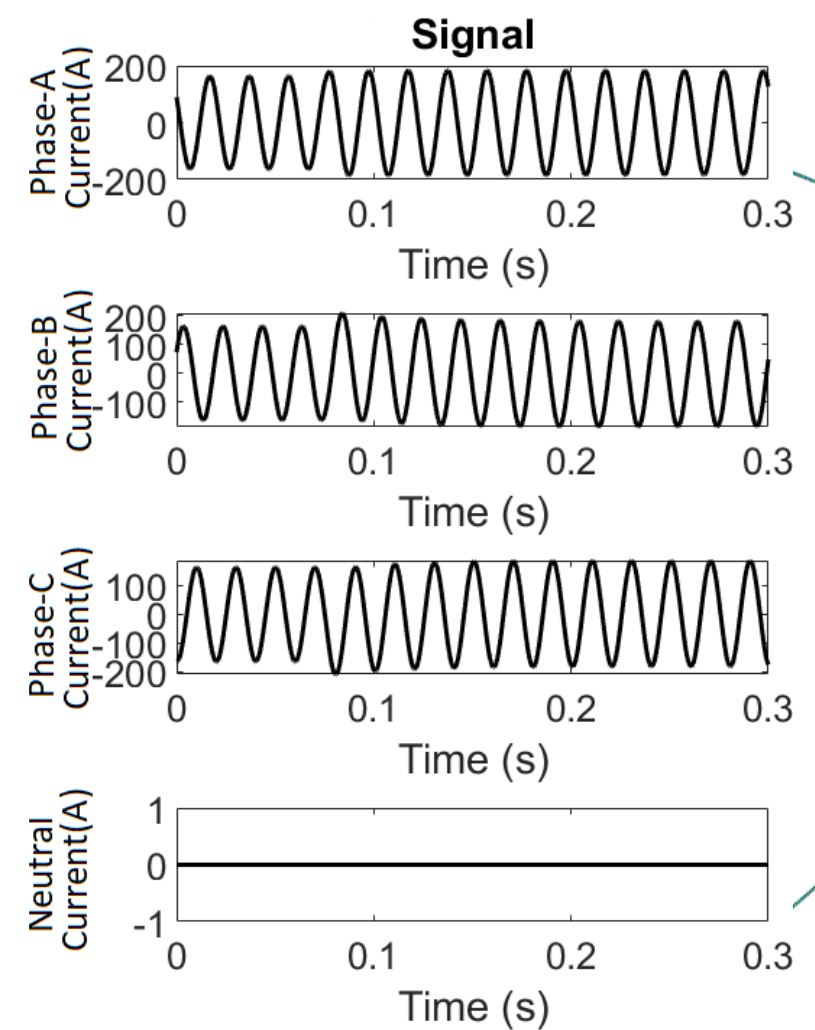
Visual Comparison



Capacitor



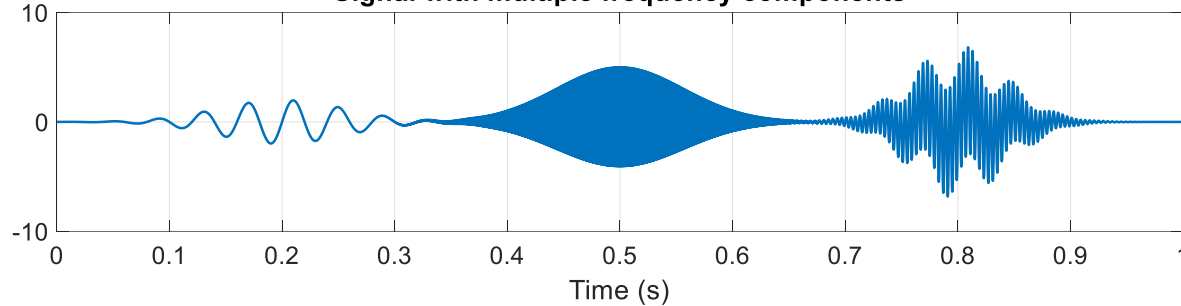
Transformer



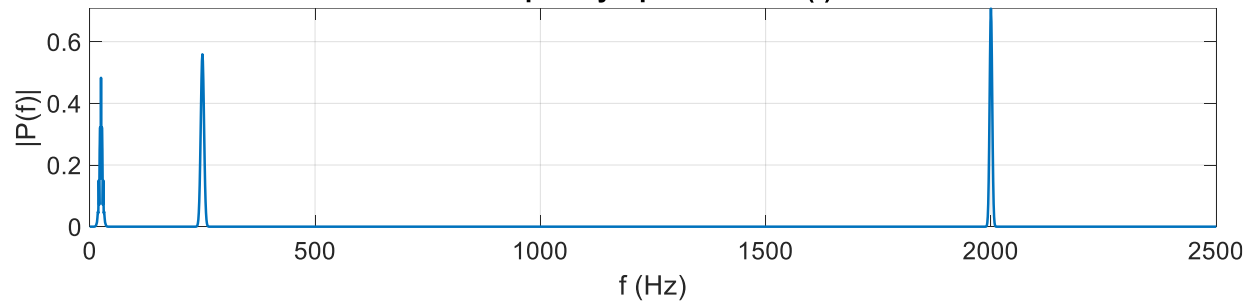
Arc Furnace

Introduction to Wavelets

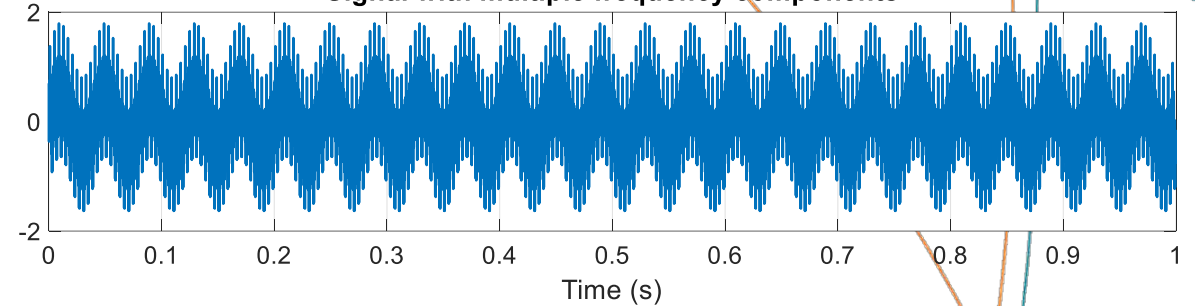
Signal with multiple frequency components



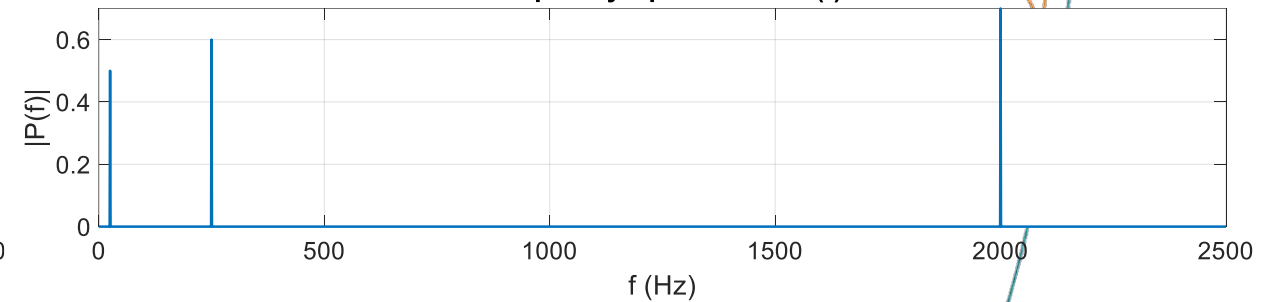
Frequency Spectrum of $x(t)$



Signal with multiple frequency components

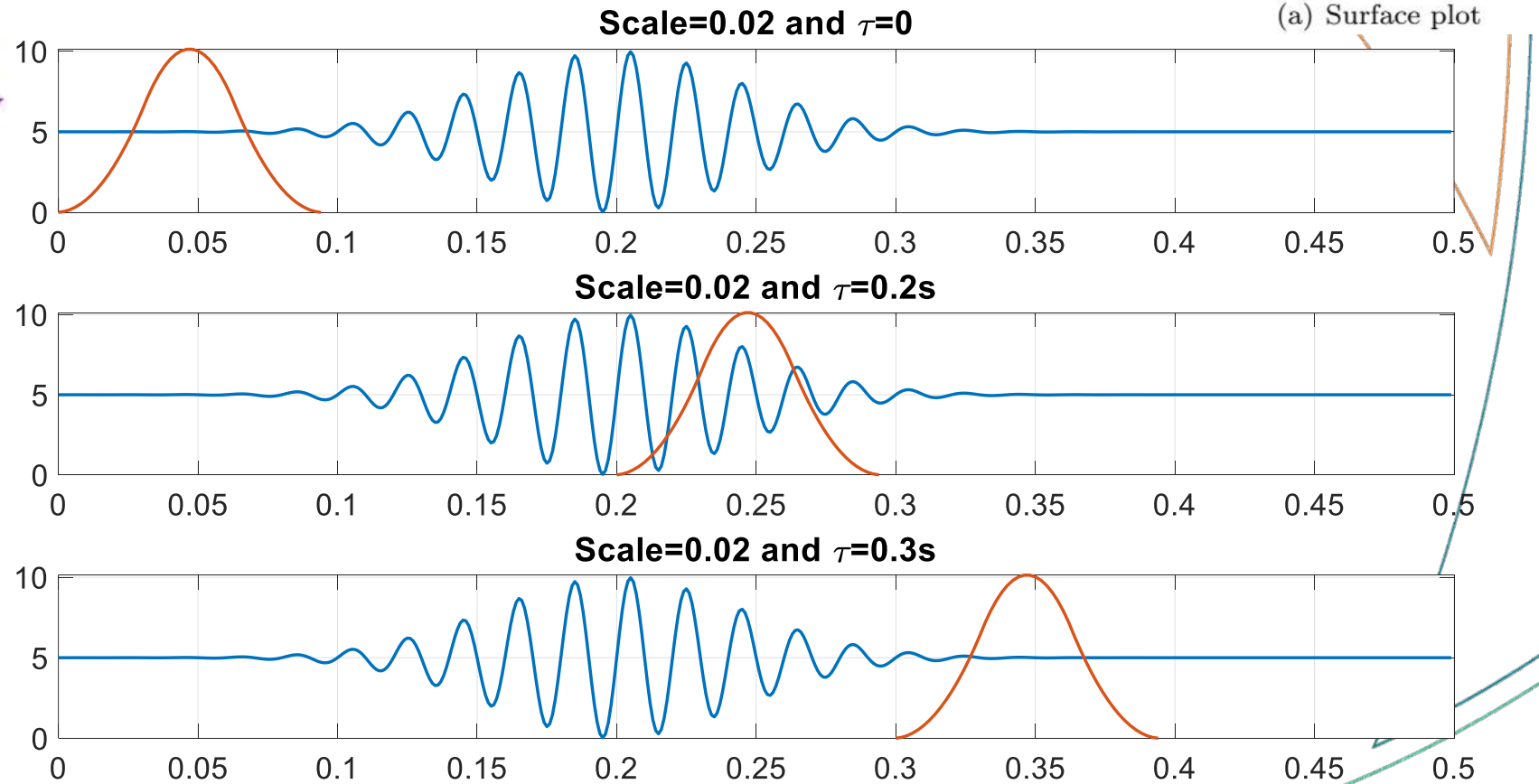
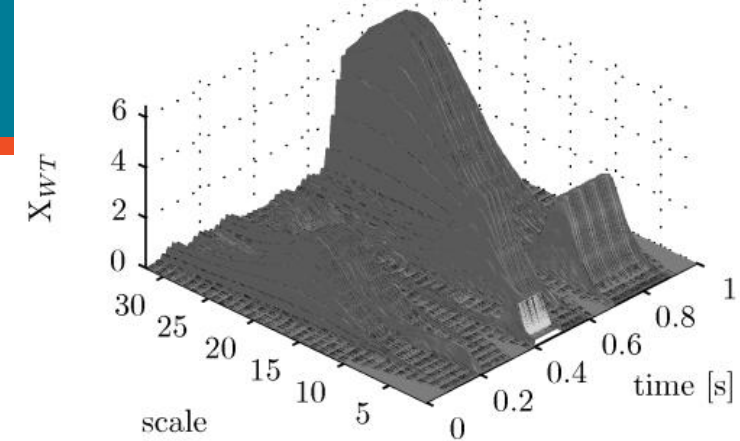
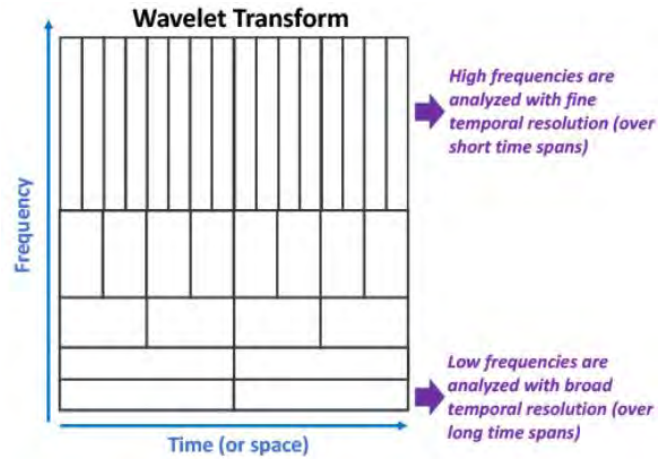


Frequency Spectrum of $x(t)$

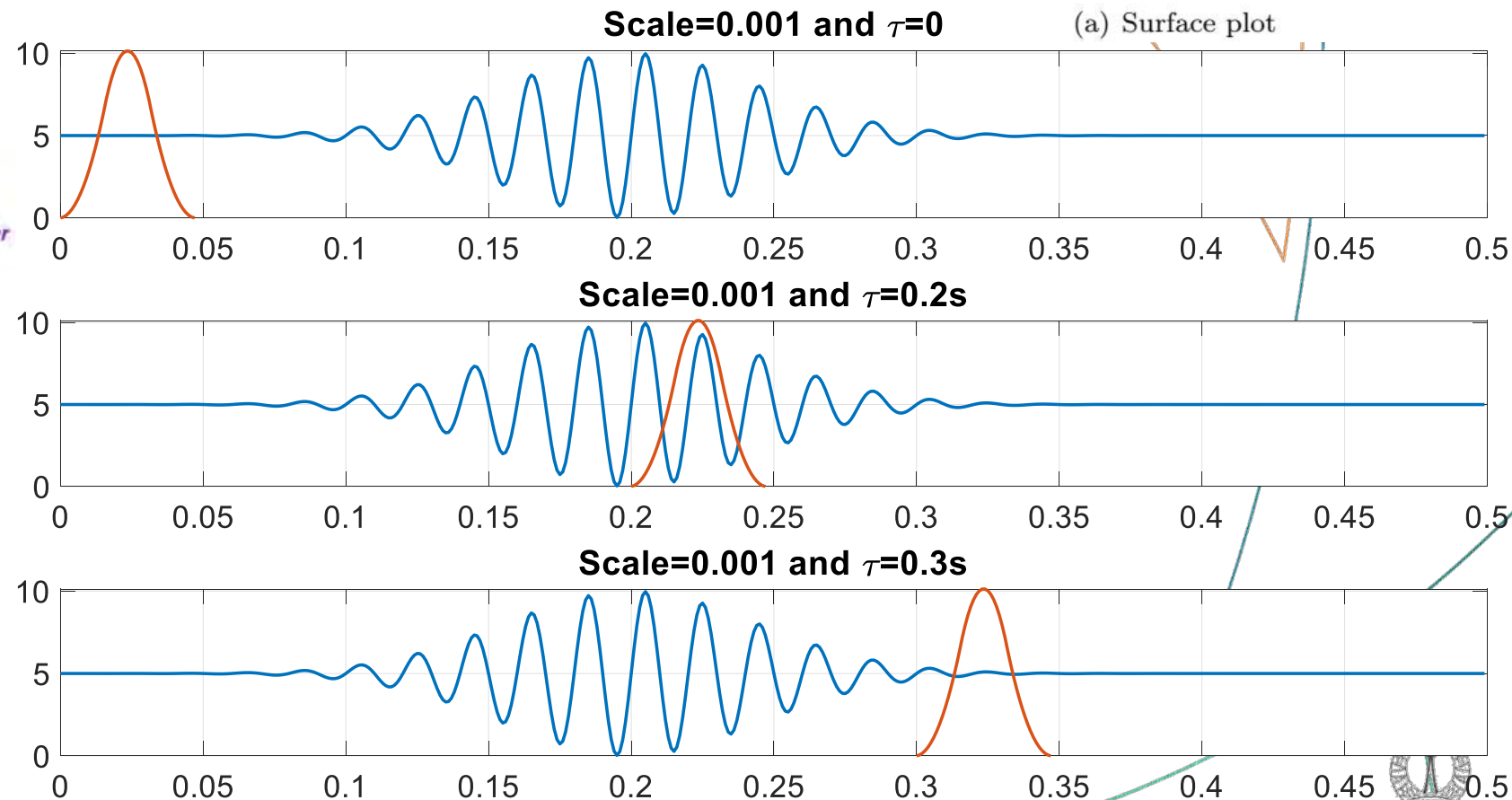
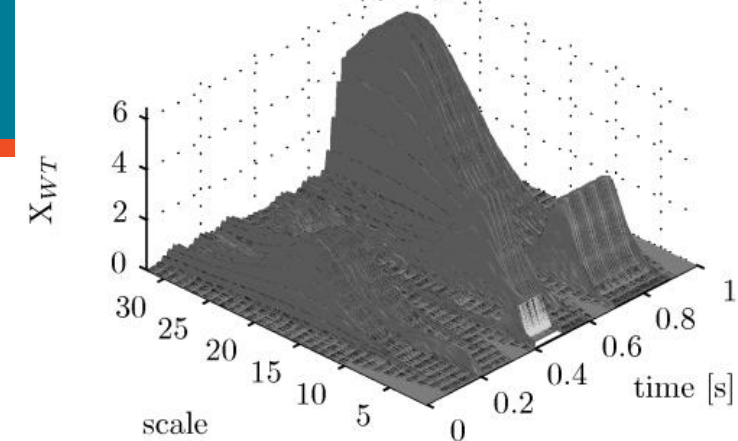
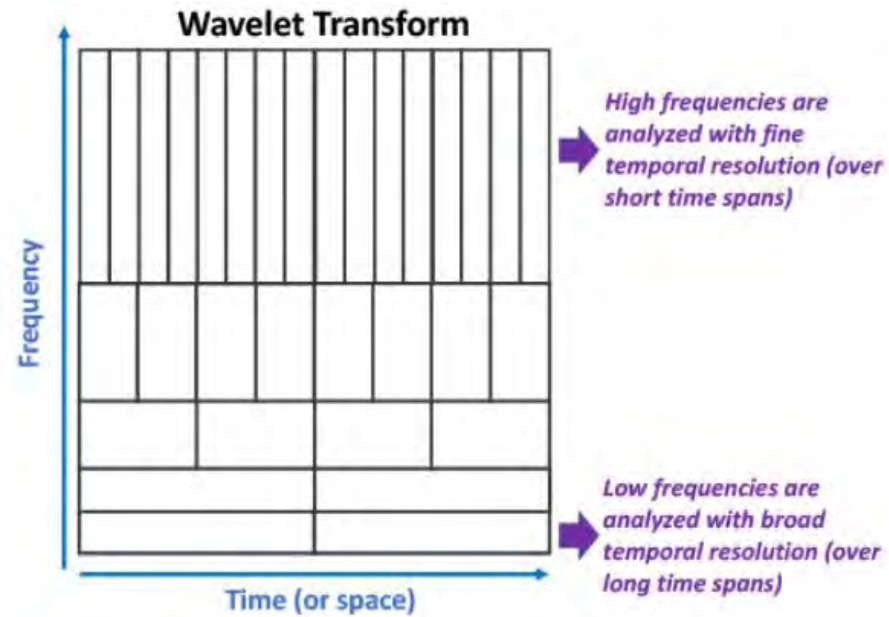


- The FT gives what frequency components (spectral components) exist in the signal.
- When the **time localization** of the spectral components are needed, a transform giving the **TIME-FREQUENCY REPRESENTATION** of the signal is needed.

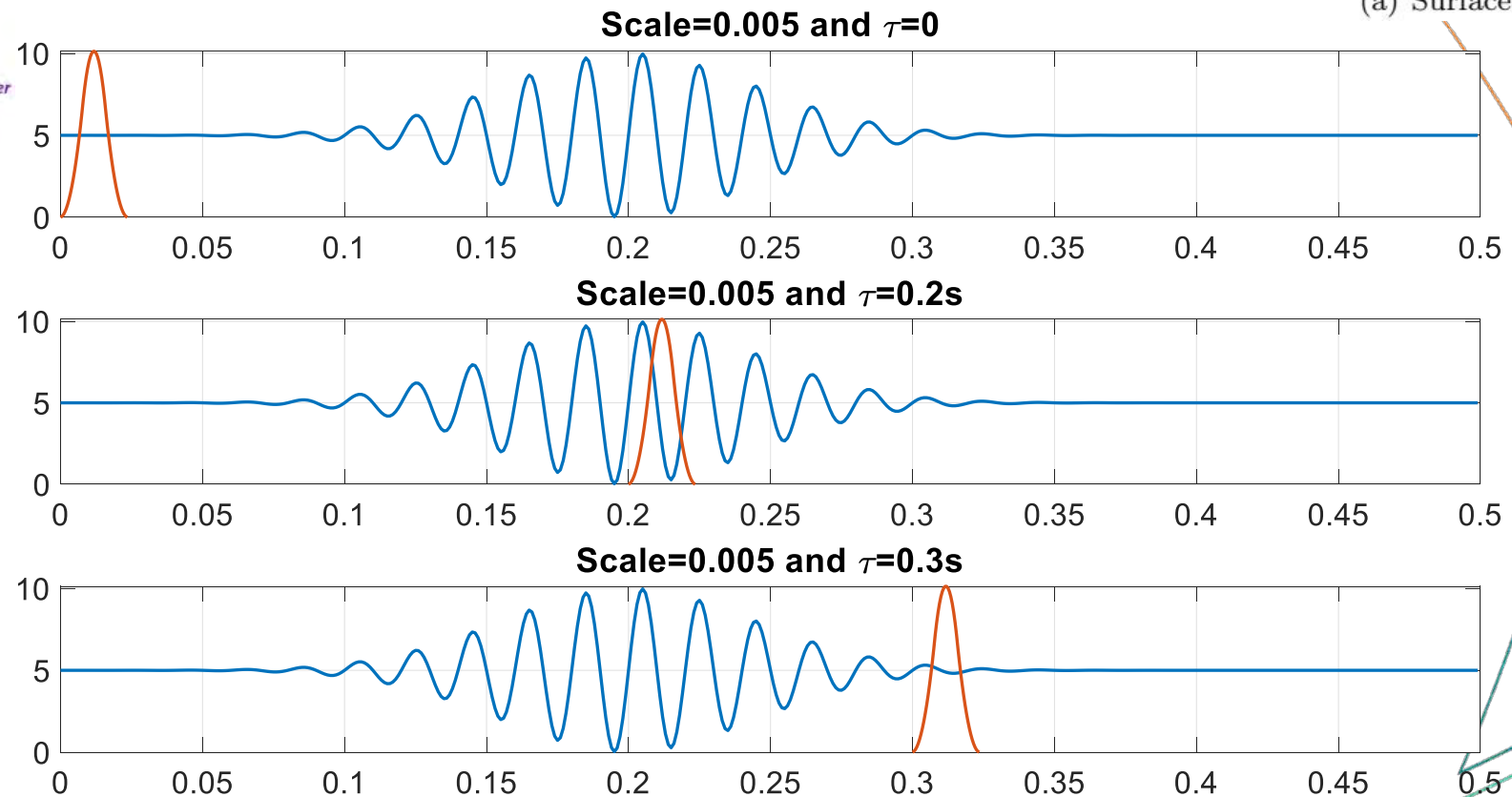
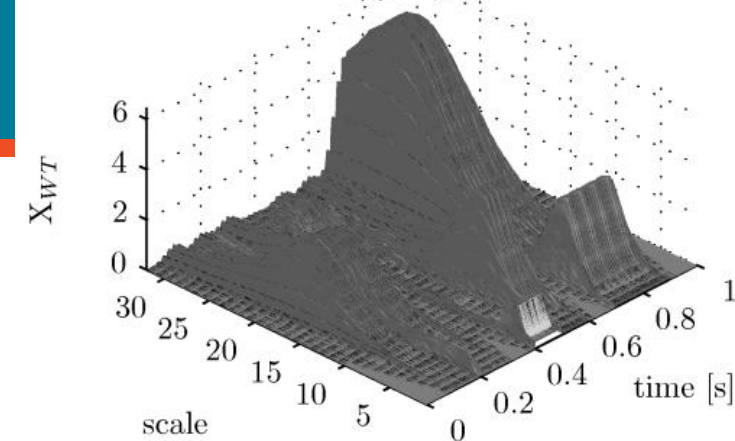
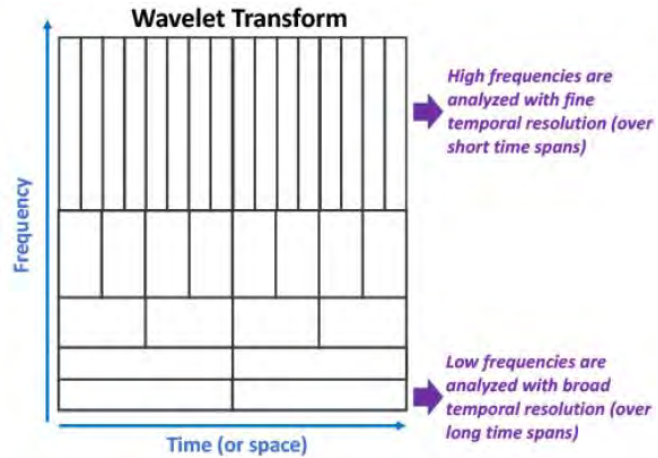
Introduction to Wavelets



Introduction to Wavelets



Introduction to Wavelets



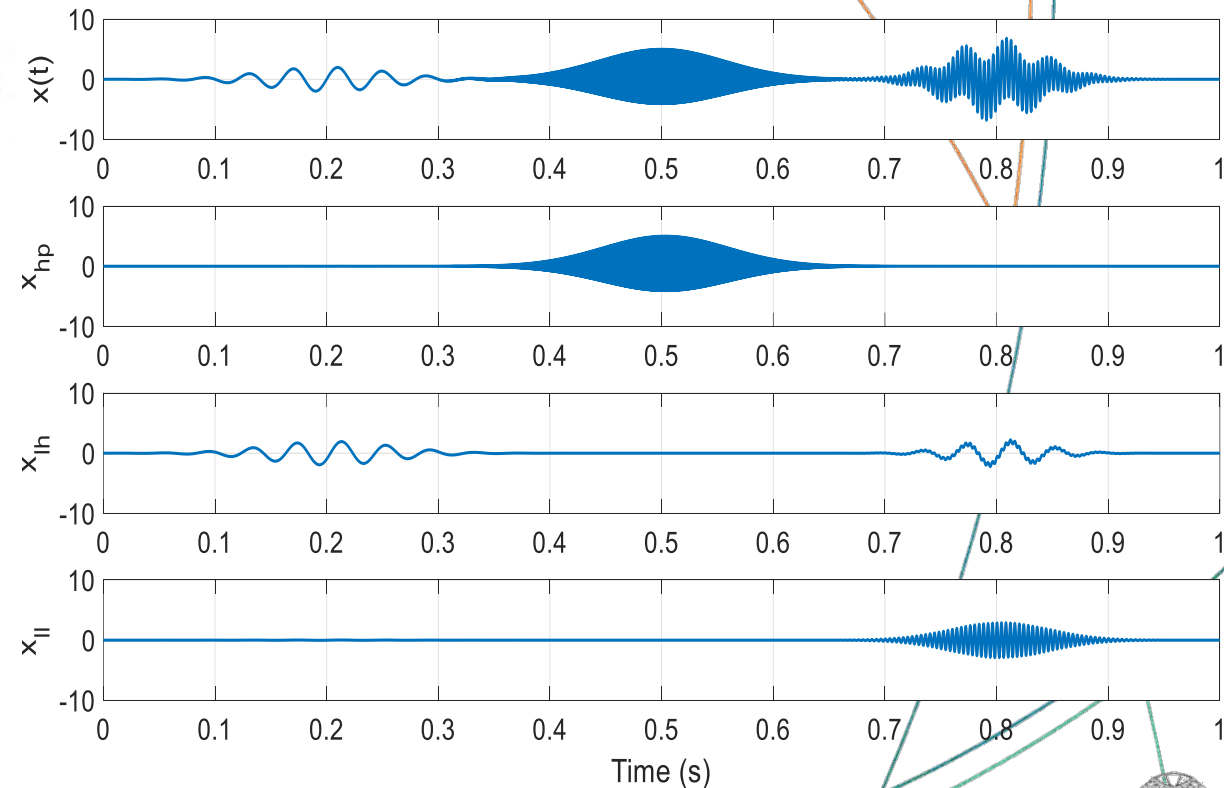
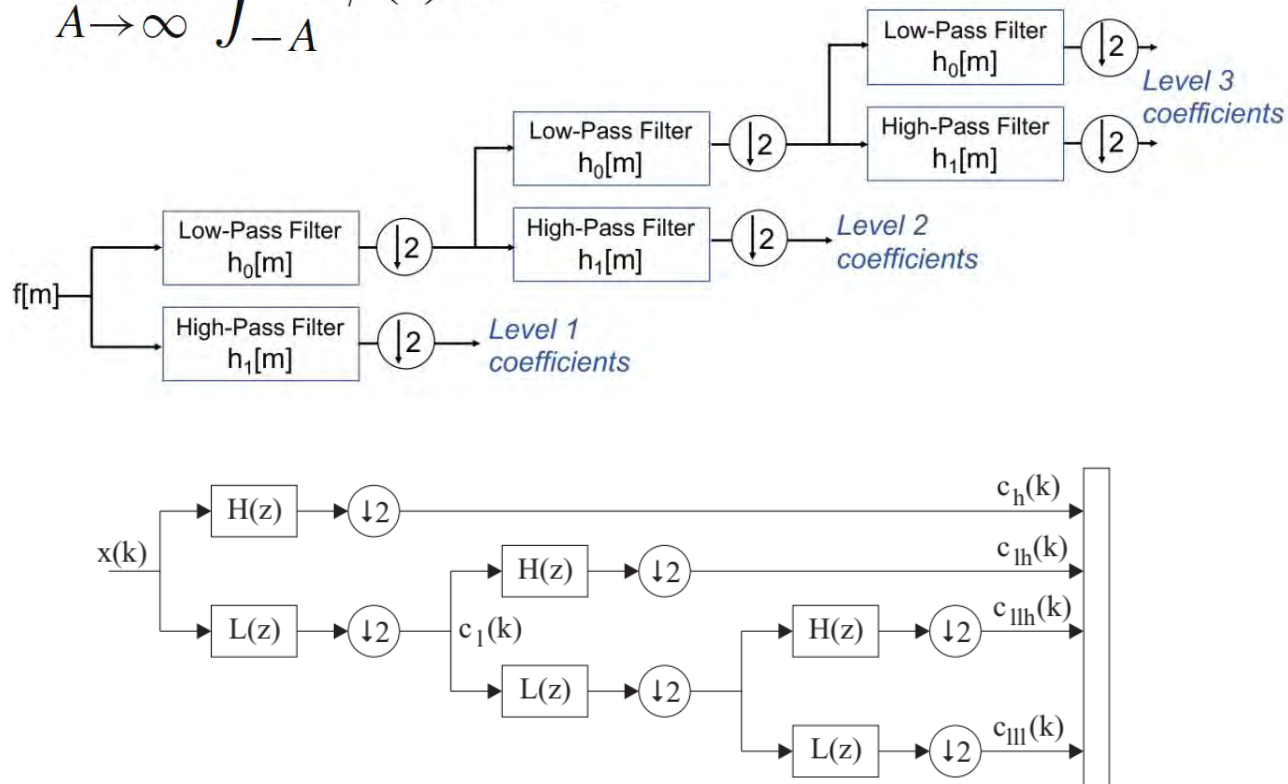
(a) Surface plot

Introduction to Wavelets

$$(W_\psi f)(b, a) = \langle f, \psi_{b,a} \rangle = \frac{1}{a} \int_{-\infty}^{\infty} f(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt$$

$$\lim_{A \rightarrow \infty} \int_{-A}^A \psi(t) dt = 0$$

$$\psi_{b,a}(t) = \frac{1}{a} \psi\left(\frac{t-b}{a}\right)$$



Example: SDF on a time series data with 24 samples

In this example, one cycle of an alternating current with 24 samples is considered

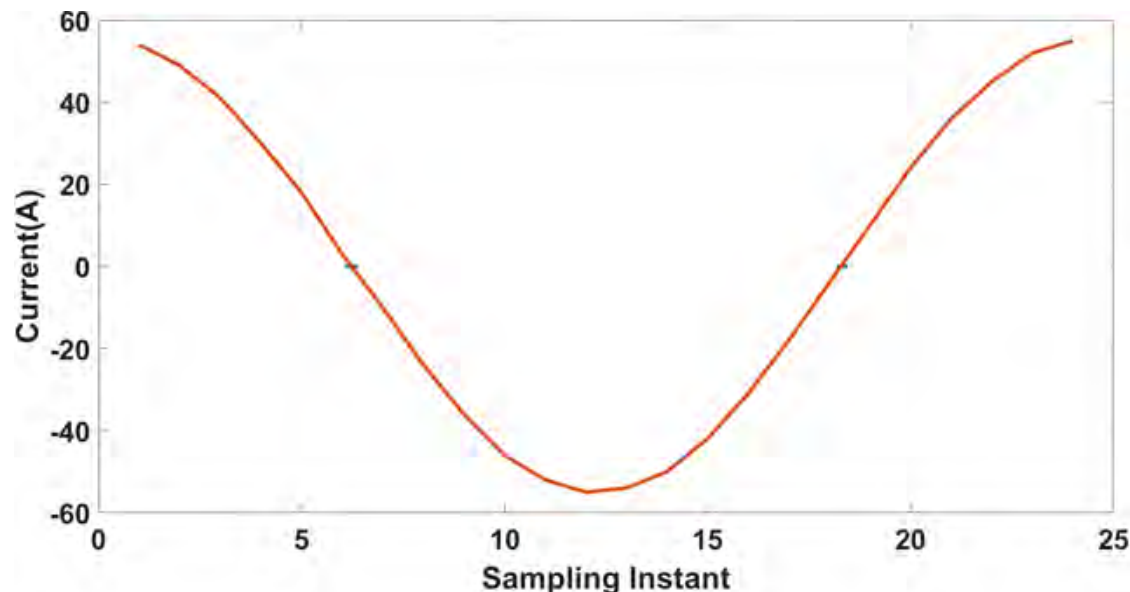
Step-1: Perform continuous wavelet transform of time series data.

Matlab Code for Step 1:

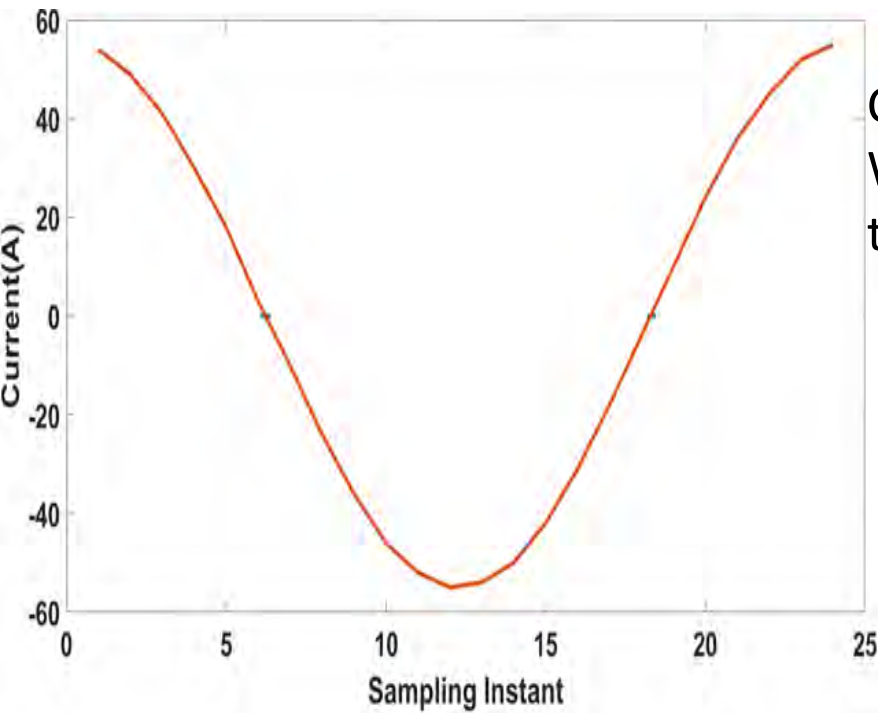
`coeffs=cwt(time_series);` gives wavelet coefficient matrix

Time series data=[54 49 41 30 18 3 -10 -24 -36 -46 -52 -55 -54 -50 -42 -31 -18 -4 10 24 36 45 52 55]

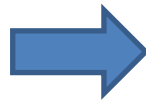
Plot of time series data having 24 samples



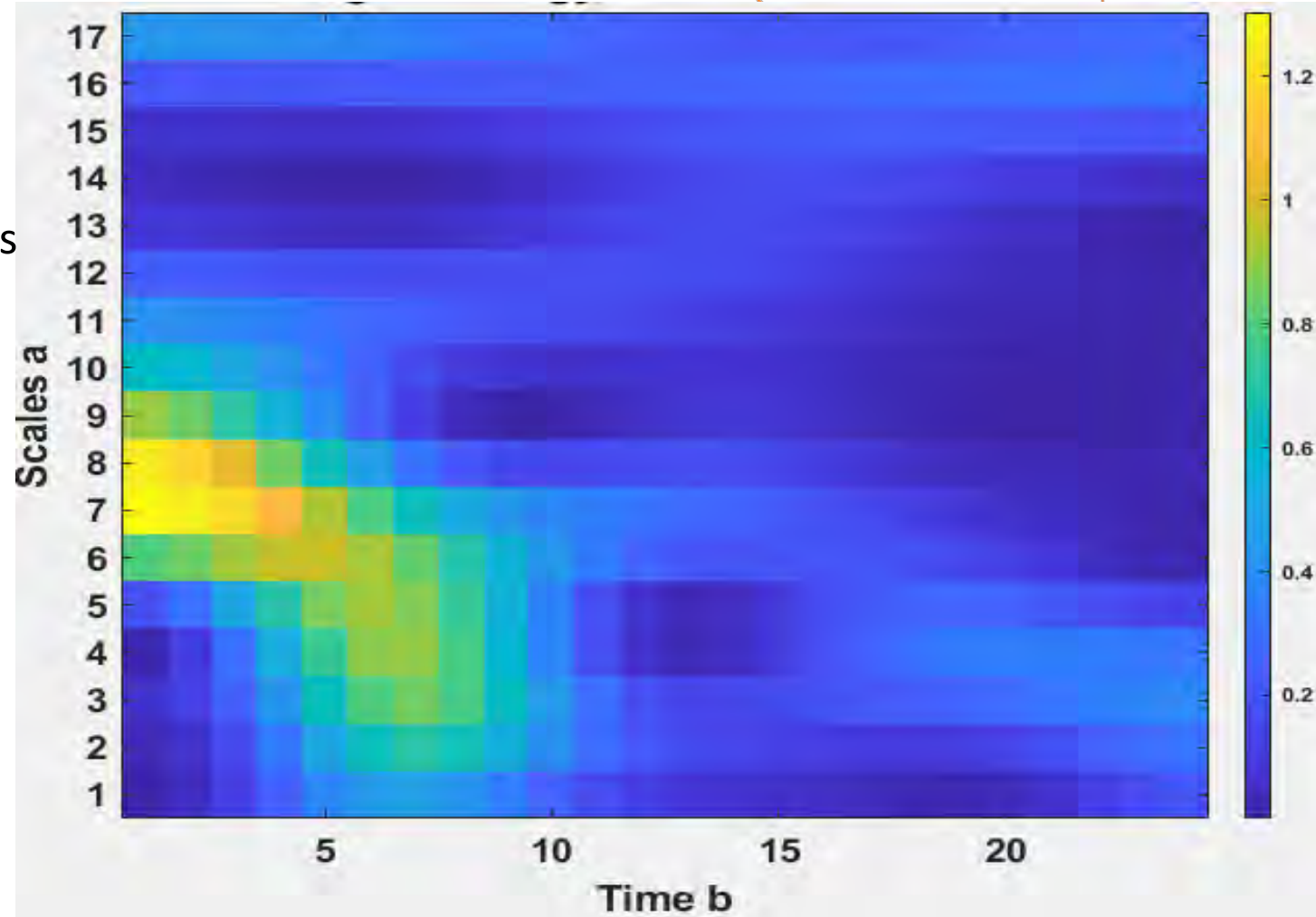
Step-1: Wavelet coefficient matrix-17x24 matrix



Continuous
Wavelet
transform



Plot of Wavelet coefficients matrix-Scalogram



Step-2 : The wavelet domain (wavelet 2-D image) is converted into 1-D scale

- Wavelet coefficients are stacked at selected time-shift positions, starting with the smallest value of scale and ending with its largest value and then back from the largest value to the smallest value of the scale at the next instant of time shift.
- Scale series data structurally similar to time series data in the phase space.
- For symbol generation, the scale series data can be handled in a similar way as time series data.

Example of a Coefficient matrix and Scale series data

Time shift→	1	2	3
Scales			
1	5	4	1
2	6	5	3
3	2	8	6

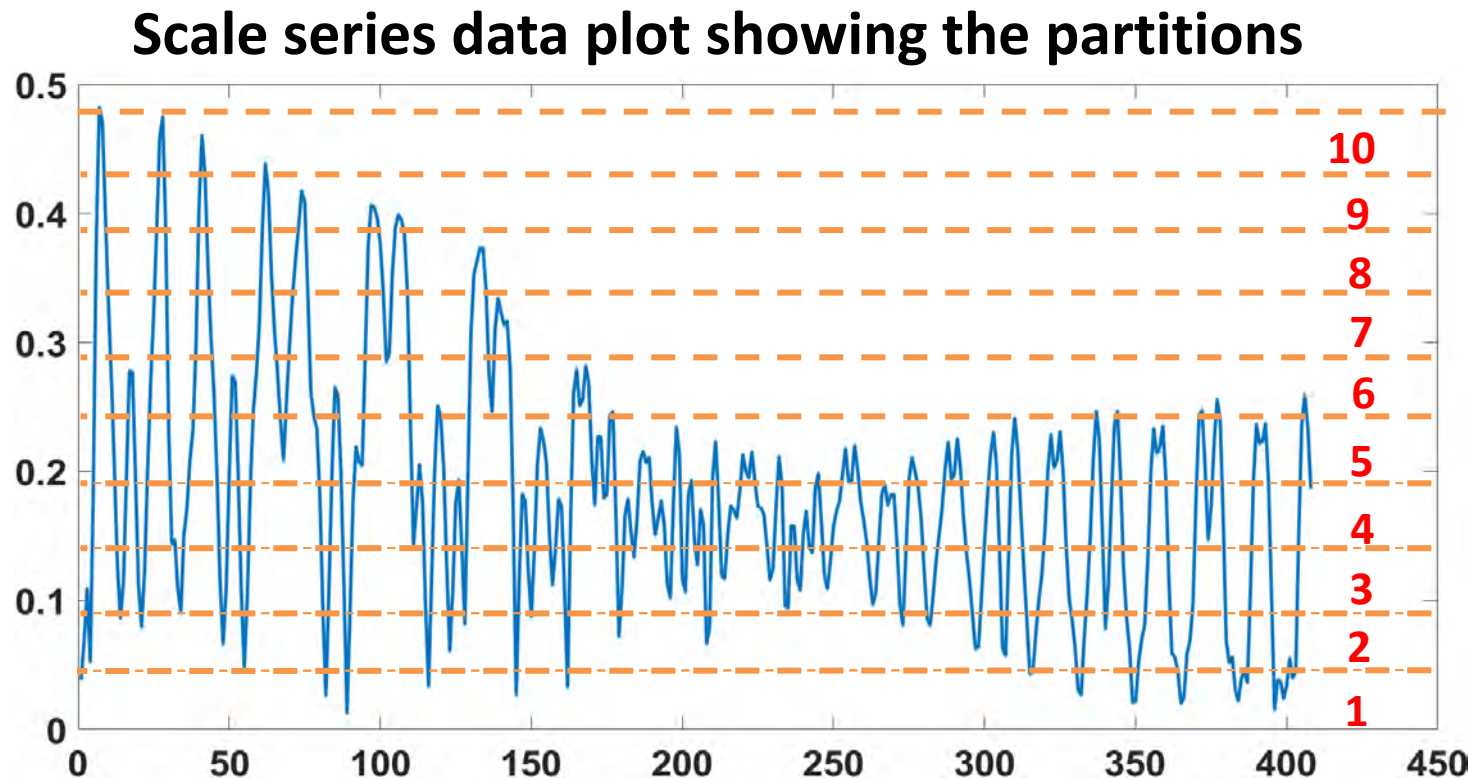
Scale series data formed: 5 6 2 8 5 4 1 3 6

Step-3 Partitioning the Scale series data to form symbolic series

There are different types of Partitioning

1. Uniform partitioning:

- In this scale series data is divided into equal regions and a unique symbol is assigned to each regions.
- In the current example scale series data is divided into 10 regions and each region is assigned with numbers one to ten as shown in figure.



Step-3 -Symbolic series

The scale series data is assigned with symbols based on the regions they belong, and symbolic series data is formed

Symbolic series data

1 2 3 1 4 8 10 10 9 7 6 5 3 2 3 4 6 6 4 3 2 3 5 6 7 8 10 10 8 5 3 3 3 2 3 4 5 5 7 9 10 9 8 7 6 4 3 2 3 4 6 6 4 3 1 2 4 6 6 7 9
10 9 8 7 6 5 5 6 7 7 8 9 9 9 7 6 5 5 4 2 1 3 4 6 6 4 3 1 2 4 5 5 5 6 8 9 9 9 8 7 6 6 8 9 9 9 8 7 5 3 4 5 4 2 1 3 5 6 5 5 3 2 3 4
4 3 2 4 7 8 8 8 8 7 6 5 7 7 7 7 7 6 4 1 3 4 4 3 2 3 5 5 5 5 3 3 3 4 4 3 1 4 6 6 6 6 6 6 5 4 5 5 4 4 5 5 4 2 3 4 4 3 3 4 5 5 5 5
4 3 4 4 4 3 2 4 5 5 3 3 4 4 3 3 4 4 2 2 4 5 4 3 3 3 4 4 4 4 5 4 4 5 4 4 4 4 3 3 3 4 5 4 2 2 4 4 3 3 3 4 3 3 4 4 4 3 3 3 4 4 4 4
5 4 4 5 5 4 4 3 3 2 2 3 4 4 4 4 4 3 2 2 3 4 5 4 4 4 3 2 2 2 3 3 4 5 5 4 4 5 5 4 3 3 2 2 2 2 3 4 5 5 5 3 2 1 3 5 5 5 4 3 2 1 1 2
2 3 4 5 5 5 5 5 4 3 2 2 2 1 1 2 2 3 5 5 5 4 2 3 4 5 5 4 3 2 2 1 1 1 2 2 3 4 5 5 5 5 4 3 1 1 1 1 1 1 2 2 4 5 6 5 3 4 5 6 5 4 2 1
1 1 1 1 1 1 3 4 5 5 5 5 4 2 1 1 1 1 1 1 1 1 1 3 5 6 5 4

Step 4. State Assignment and State Transition Matrix

Step 4.1

State assignment. A group of symbols or a symbol is assigned with a state. In the current example each symbol in the symbolic series is assigned with a state and sequence of state transition is found. Here state numbered one to ten is assigned to each symbol

Sequence of state transitions (same as the sequence of symbols in the example considered):

1 2 3 1 4 8 10 10 9 7 6 5 3 2 3 4 6 6 4 3 2 3 5 6 7 8 10 10 8 5 3 3 3 2 3 4 5 5 7 9 10 9 8 7 6 4 3 2 3 4 6 6 4 3 1 2 4 6 6 7 9
10 9 8 7 6 5 5 6 7 7 8 9 9 9 7 6 5 5 4 2 1 3 4 6 6 4 3 1 2 4 5 5 5 6 8 9 9 9 8 7 6 6 8 9 9 9 8 7 5 3 4 5 4 2 1 3 5 6 5 5 3 2 3 4
4 3 2 4 7 8 8 8 8 7 6 5 7 7 7 7 7 6 4 1 3 4 4 3 2 3 5 5 5 5 3 3 3 4 4 3 1 4 6 6 6 6 6 6 5 4 5 5 4 4 5 5 4 2 3 4 4 3 3 4 5 5 5 5
4 3 4 4 4 3 2 4 5 5 3 3 4 4 3 3 4 4 2 2 4 5 4 3 3 3 4 4 4 4 5 4 4 5 4 4 4 3 3 3 4 5 4 2 2 4 4 3 3 3 4 3 3 4 4 4 3 3 3 4 4 4 4
5 4 4 5 5 4 4 3 3 2 2 3 4 4 4 4 4 3 2 2 3 4 5 4 4 4 3 2 2 2 3 3 4 5 5 4 4 5 5 4 3 3 2 2 2 2 3 4 5 5 5 3 2 1 3 5 5 5 4 3 2 1 1 2
2 3 4 5 5 5 5 5 4 3 2 2 2 1 1 2 2 3 5 5 5 4 2 3 4 5 5 4 3 2 2 1 1 1 2 2 3 4 5 5 5 5 4 3 1 1 1 1 1 1 2 2 4 5 6 5 3 4 5 6 5 4 2 1
1 1 1 1 1 1 3 4 5 5 5 5 4 2 1 1 1 1 1 1 1 1 1 1 3 5 6 5 4

Step 4. State Assignment and State Transition Matrix

Step 4.2

To evaluate the probability of state transitions, the number of present state and number of present state to next state transition is found.

Step 4.3

Finding the state transition probability matrix (A).

$$P[q_k|q_j] \approx N(s1s2)/N(s1)$$

Example of a state transition probability matrix(A) with four states.

$$A = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & p_{0,3} \\ p_{1,0} & p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,0} & p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,0} & p_{3,1} & p_{3,2} & p_{3,3} \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/2 & 1/3 & 0 \\ 0 & 1/3 & 1/2 & 1/6 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

- An $n \times n$ matrix is called a Markov matrix if all entries are nonnegative, and the sum of each column vector is equal to 1
- Transpose of state transition probability matrix (A^T) is a Markov matrix.
- The stable equilibrium distribution of A^T is the eigenvector corresponding to eigen value of unity

Step-4 –State transition probability matrix(A)

➤ State transition probability matrix(A)

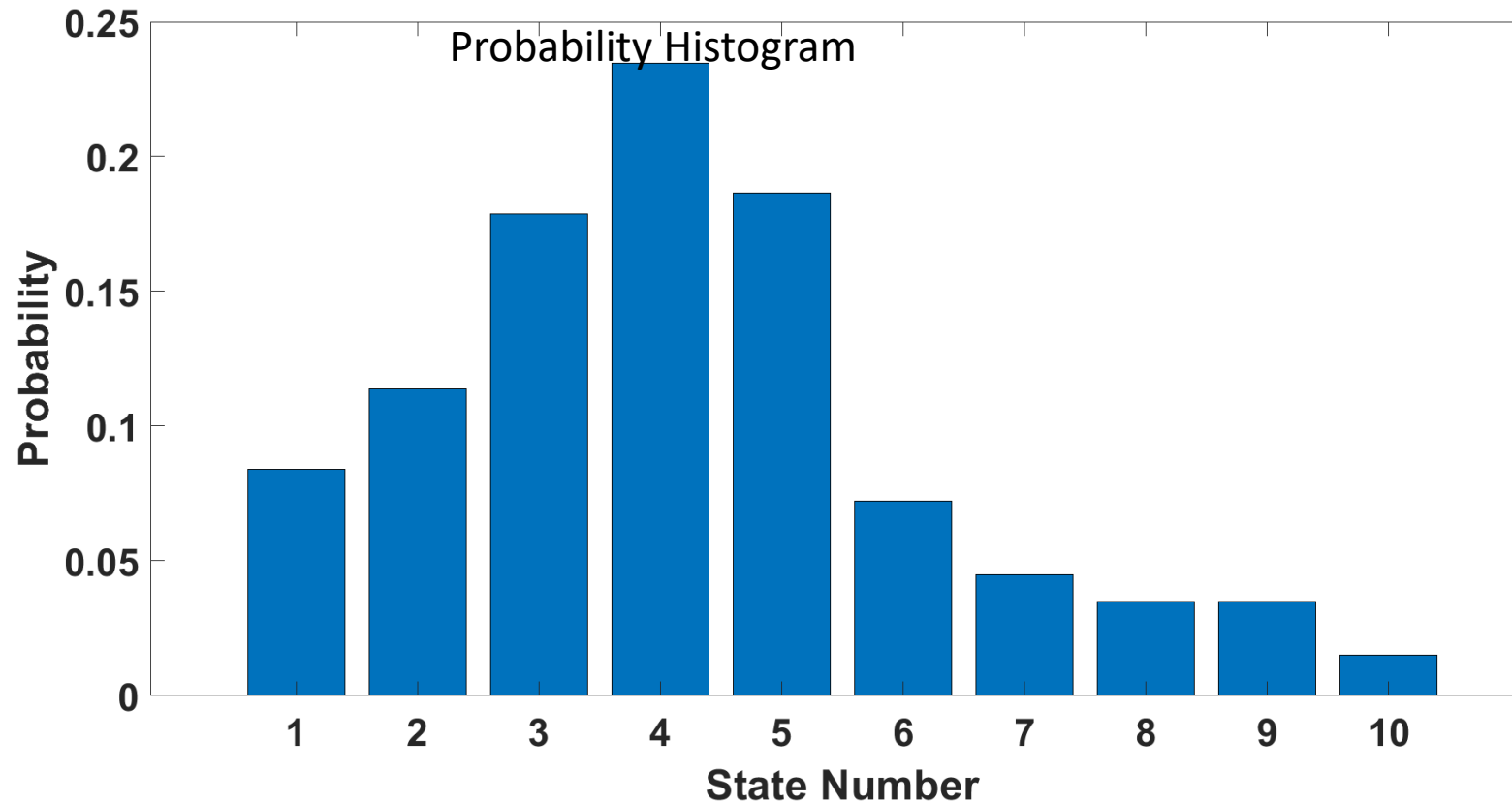
State transition probability matrix

0.59459	0.18919	0.16216	0.054054	0	0	0	0	0	0
0.17021	0.34043	0.34043	0.14894	0	0	0	0	0	0
0.068493	0.21918	0.26027	0.36986	0.082192	0	0	0	0	0
0.010638	0.085106	0.25532	0.32979	0.24468	0.053191	0.010638	0.010638	0	0
0	0	0.10667	0.30667	0.46667	0.093333	0.026667	0	0	0
0	0	0	0.17241	0.31034	0.34483	0.10345	0.068966	0	0
0	0	0	0	0.055556	0.38889	0.27778	0.16667	0.11111	0
0	0	0	0	0.071429	0	0.35714	0.21429	0.21429	0.14286
0	0	0	0	0	0	0.14286	0.28571	0.42857	0.14286
0	0	0	0	0	0	0	0.16667	0.5	0.33333

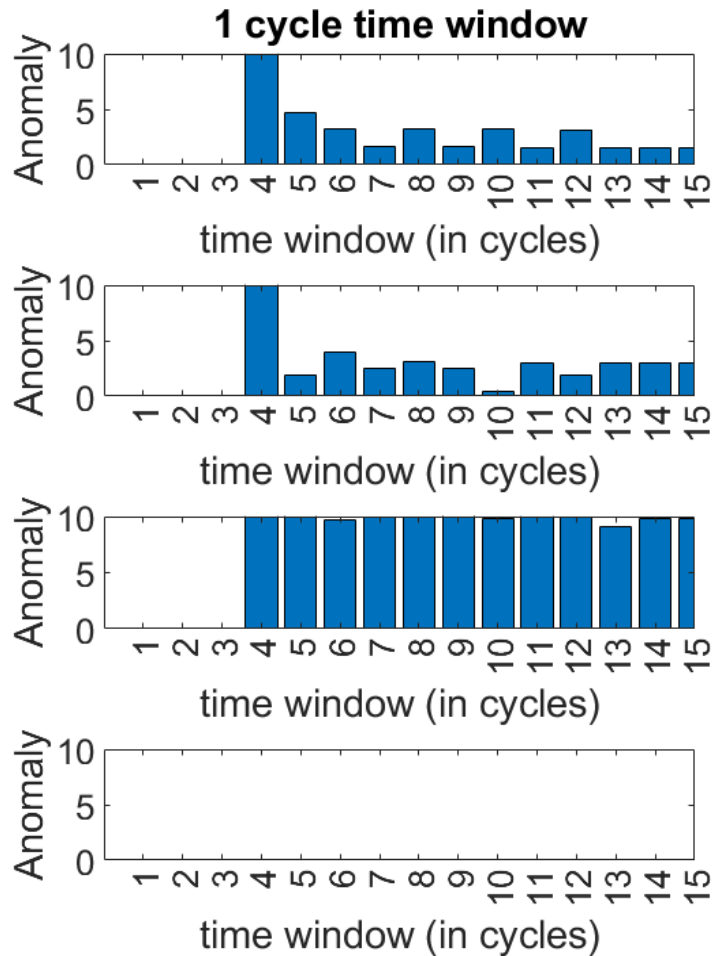
Step-5 –Probability of each state.

The stable equilibrium distribution of A^T is the feature vector

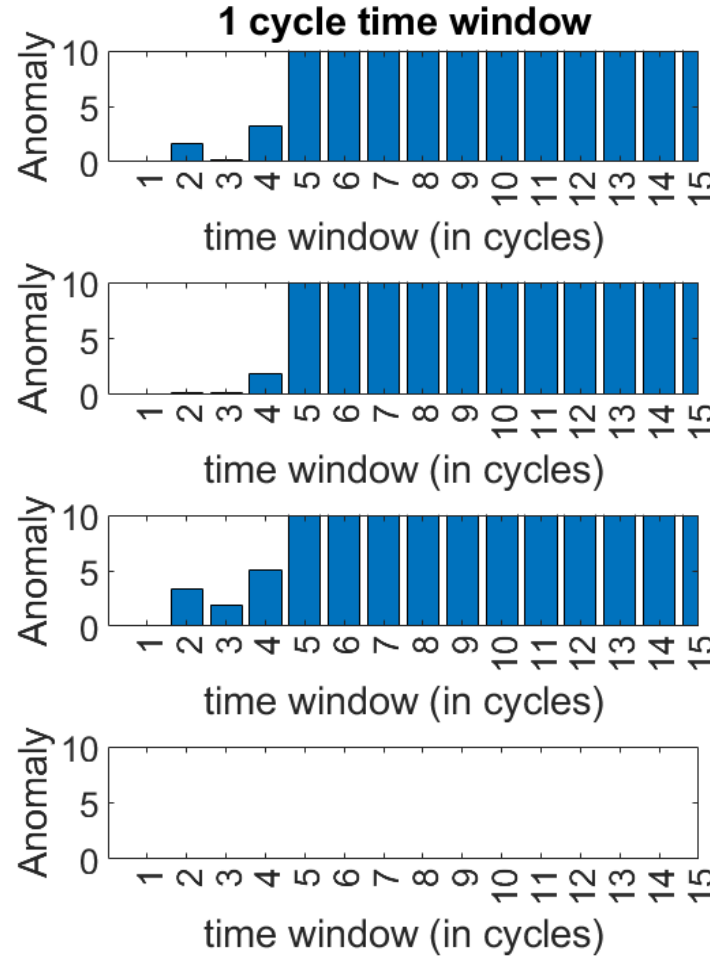
The stable equilibrium distribution of A^T = 0.0842 0.1138 0.1788 0.2347 0.1866 0.0723 0.0449 0.0349
0.0349 0.0150



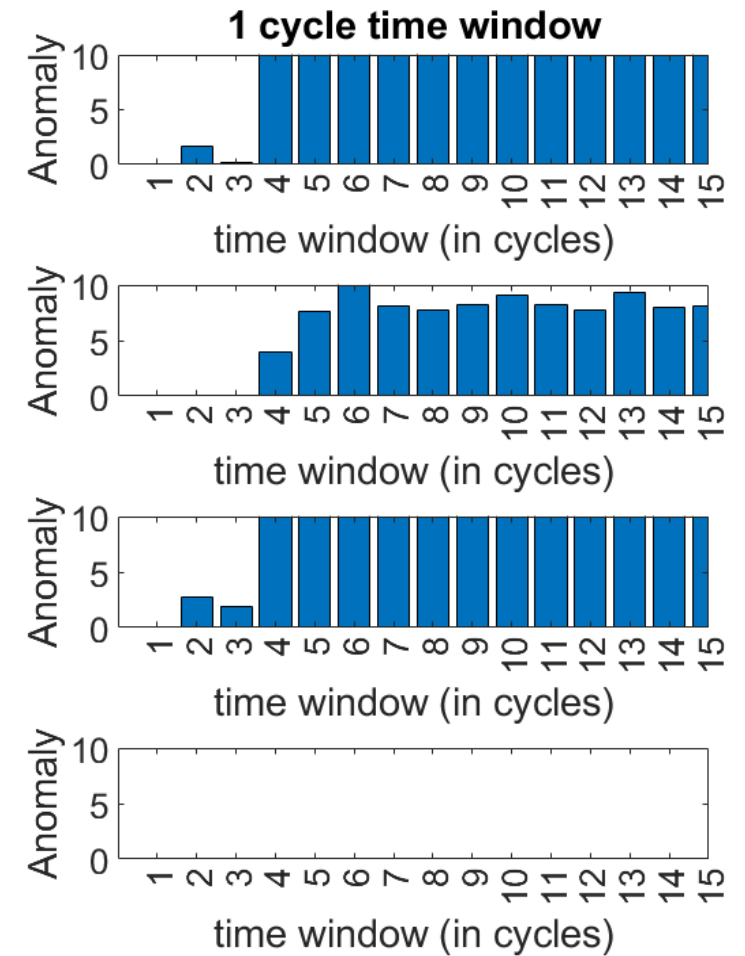
Event Detection: Apply SDF



Capacitor

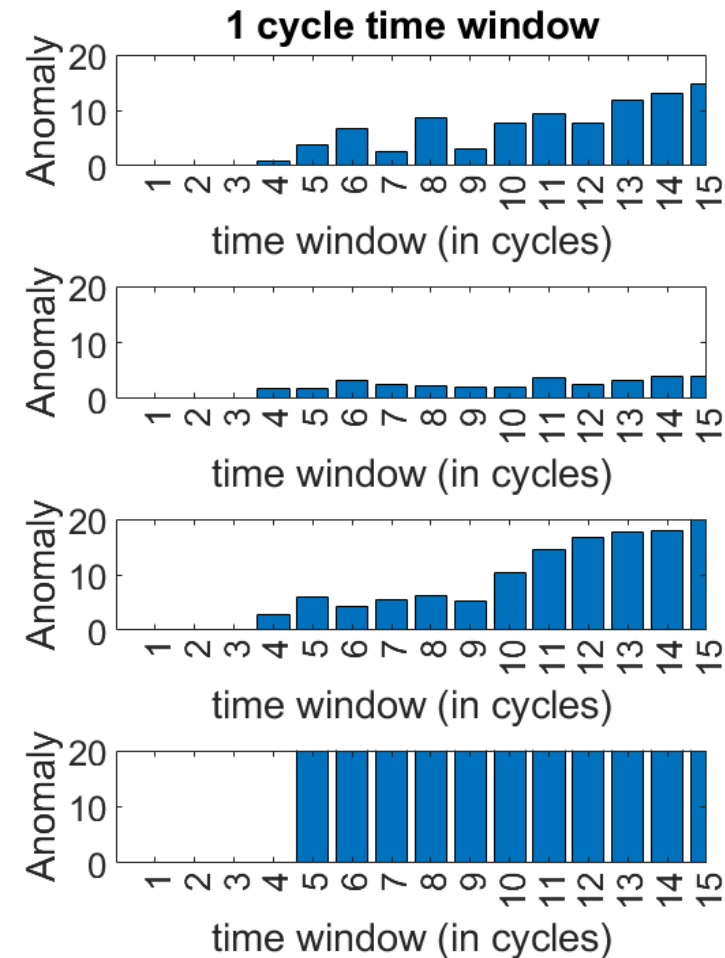
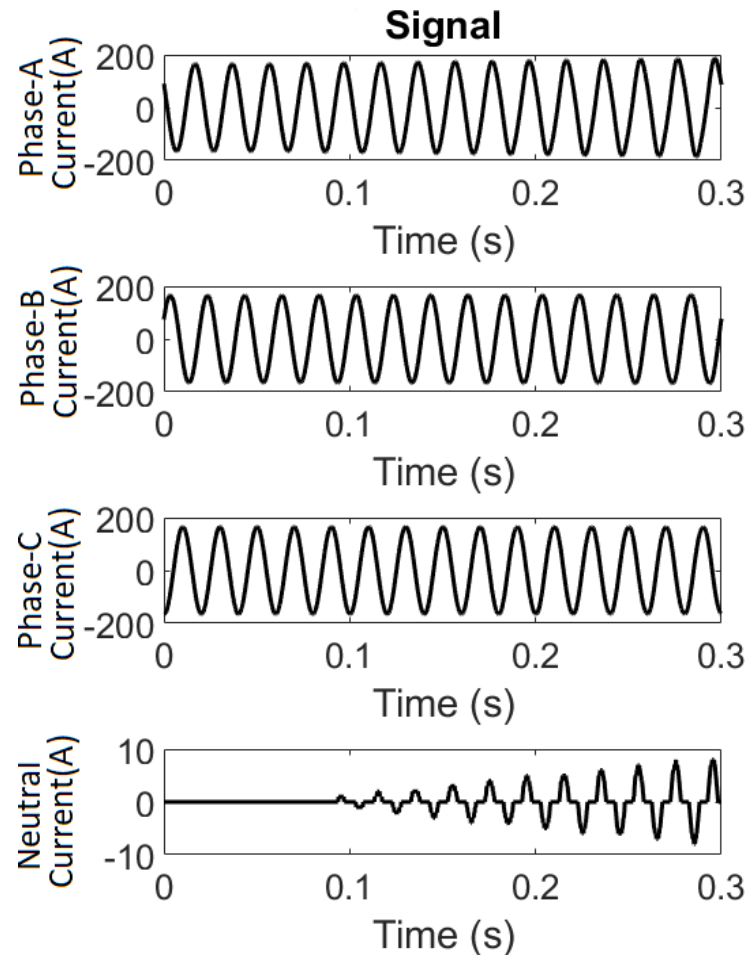


Transformer

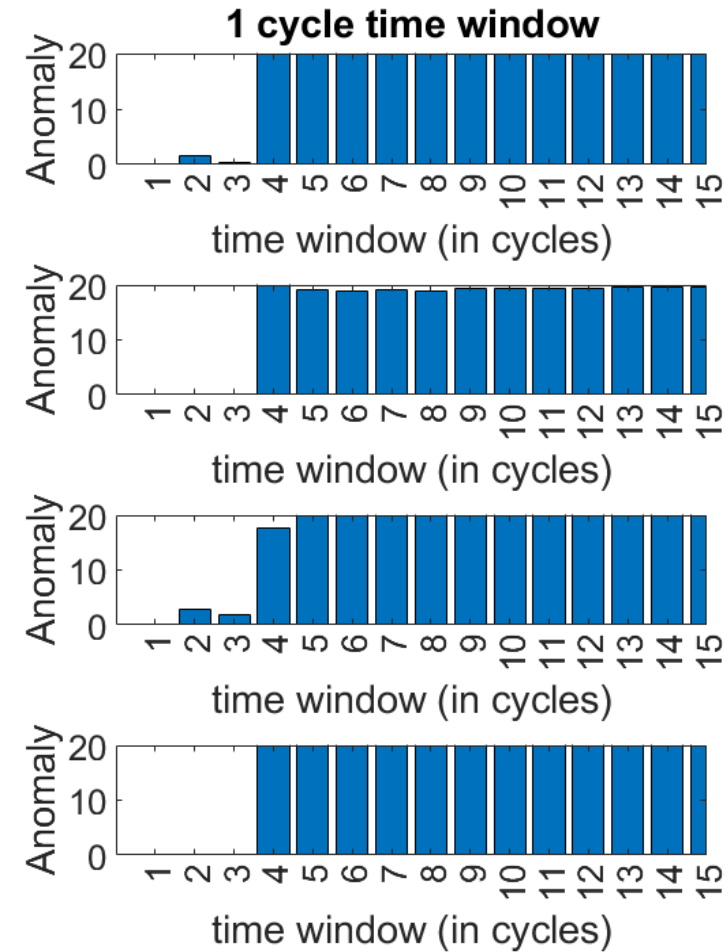
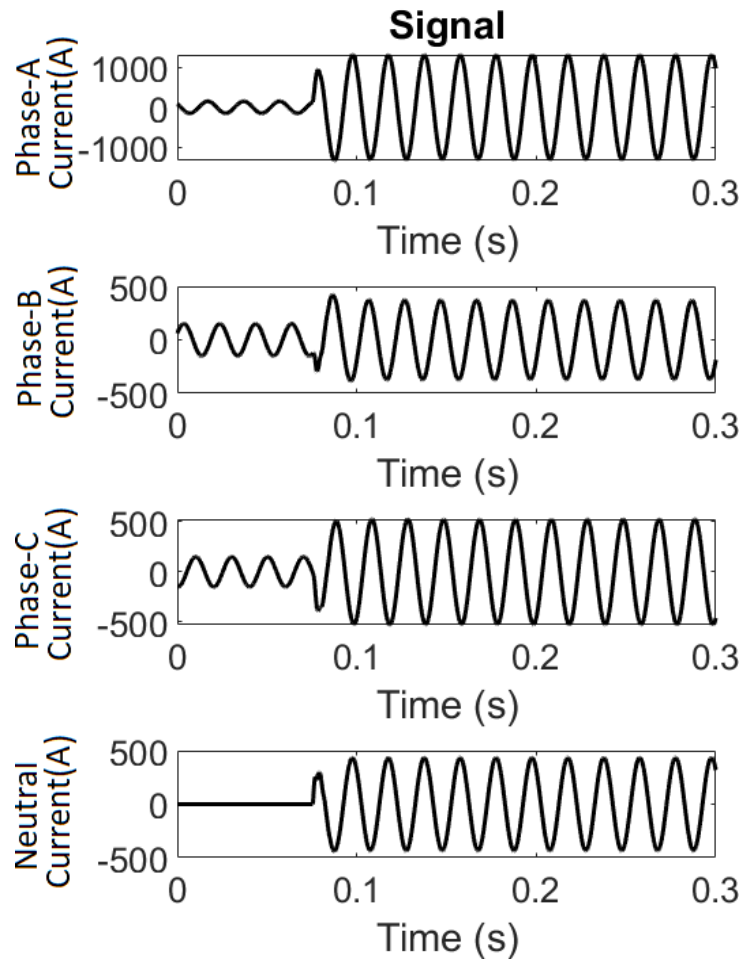


Arc Furnace

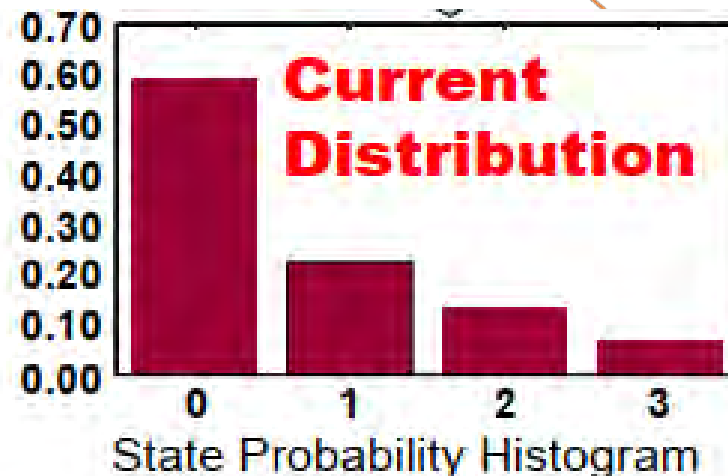
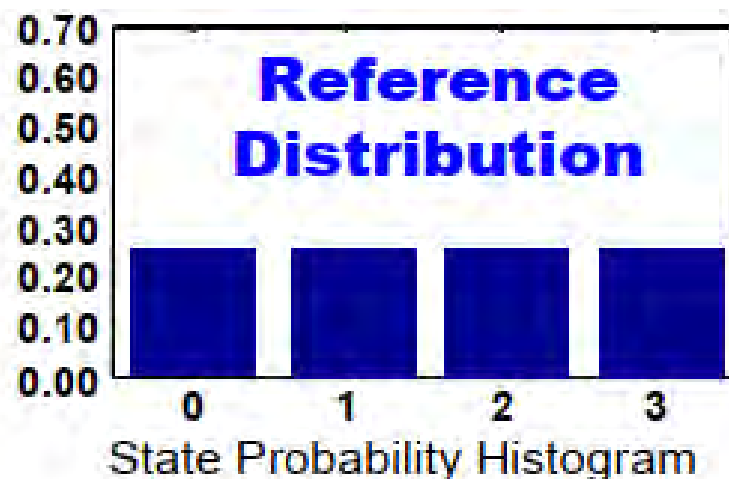
Event Detection: AG High Impedance Fault



Event Detection: AG Low Impedance Fault



How do you detect



Using some anomaly measure

$$\psi^k = \arccos \left(\frac{\langle P^k, P^0 \rangle}{\|P^k\|_2 \|P^0\|_2} \right)$$

where $\langle P^k, P^0 \rangle$ is the dot product between the vectors

Privacy Preserving Load Forecasting

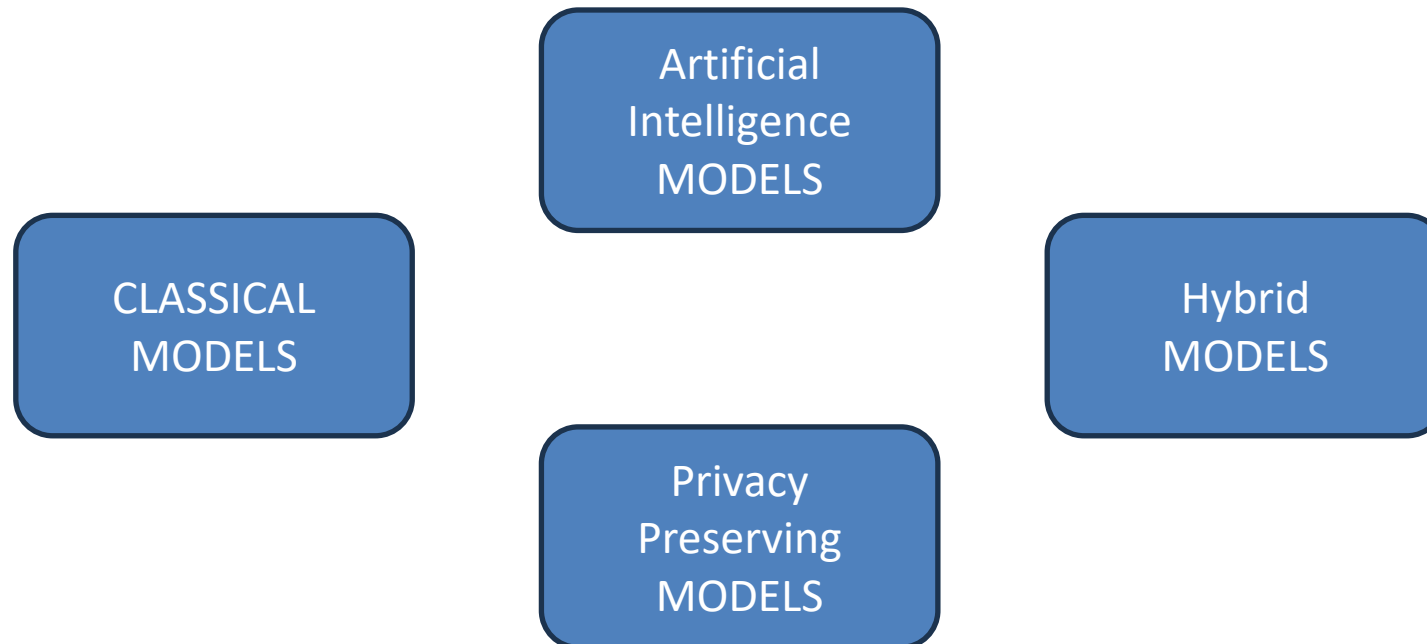
Load Forecasting

Short Term Load Forecasting

≤ 24 hours; 15 min to 1 hour for markets

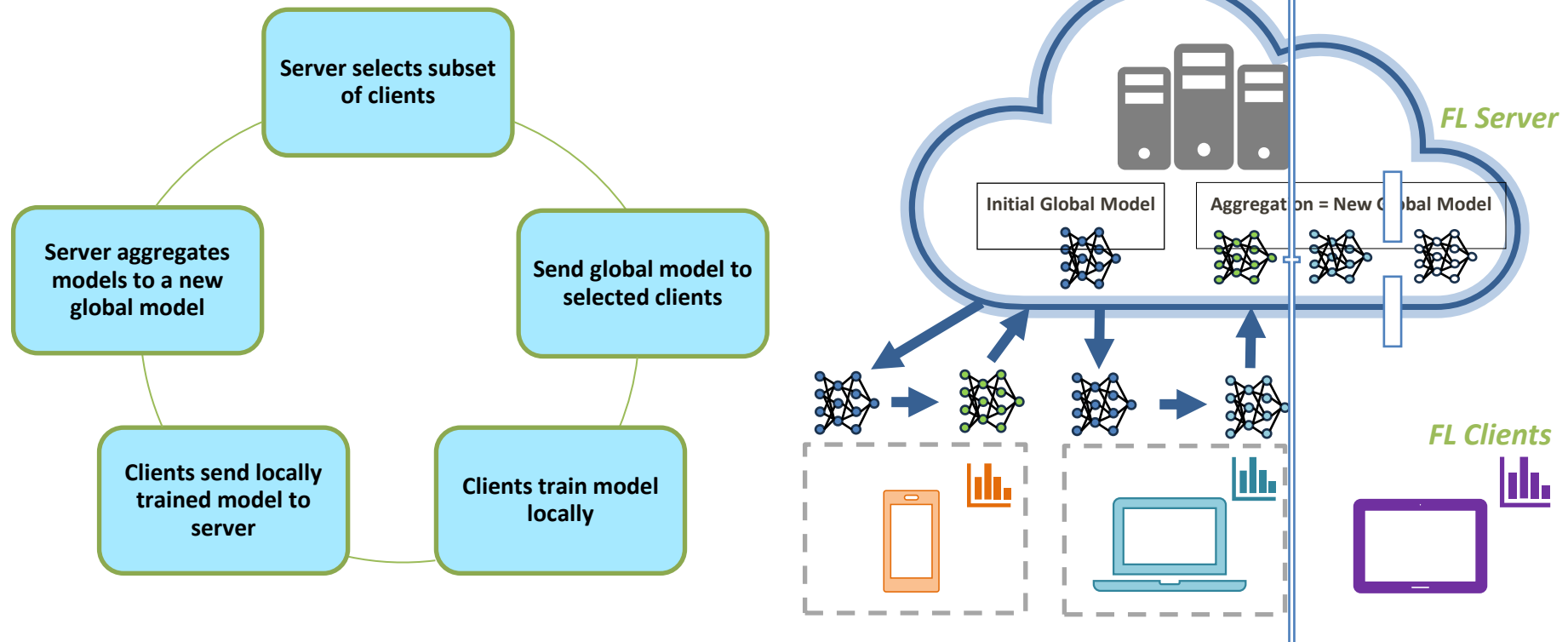
Long Term Load Forecasting

> 24 hours



Federated Learning

- IoT and edge devices are a **rich data source** – eg. Traffic cameras, Smart Meters etc.
- Data cannot be moved to a central location – **huge network costs, privacy concerns**
- Solution? **Federated Learning!**



Federated Learning

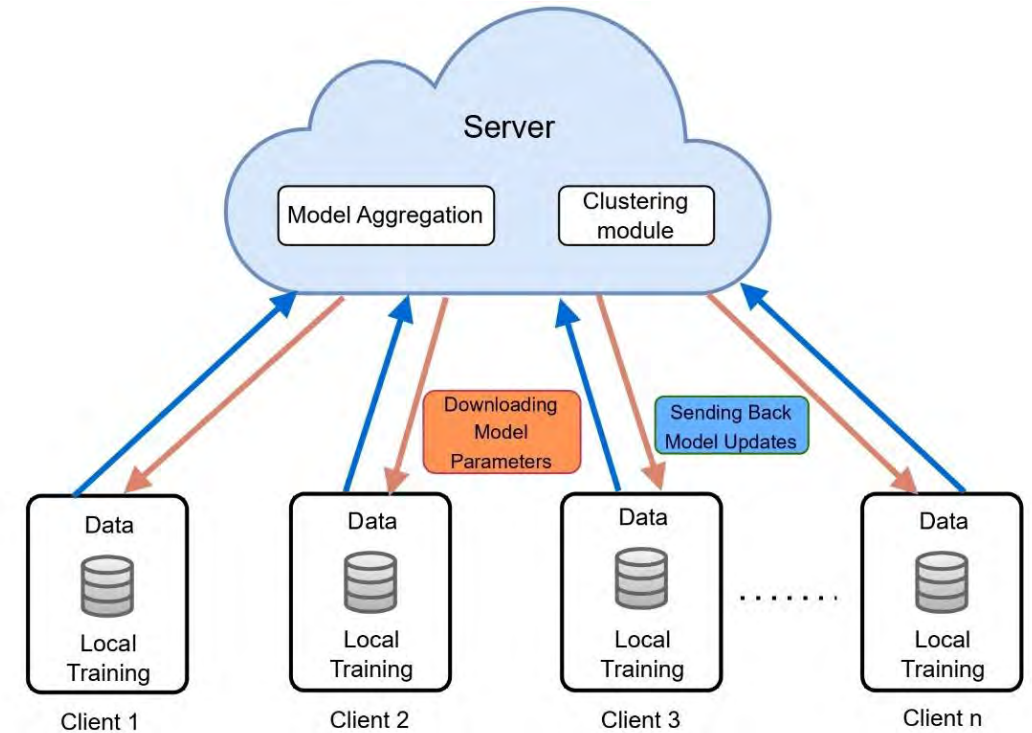
- Decentralized machine learning approach
- Offers privacy-preserving benefits by ensuring that raw data remains on user devices and is not shared externally
- Local model training on individual devices
- Exchange of model updates instead of raw data
- Centralized aggregation of model updates
- Used FedAvg and FedProx algorithm

Here are some more examples of use cases for federated learning:

1. Mobile application:-build model for user behavior
2. Healthcare:- train model on patient data without violating privacy regulations.
3. Transportations:- self-driving cars

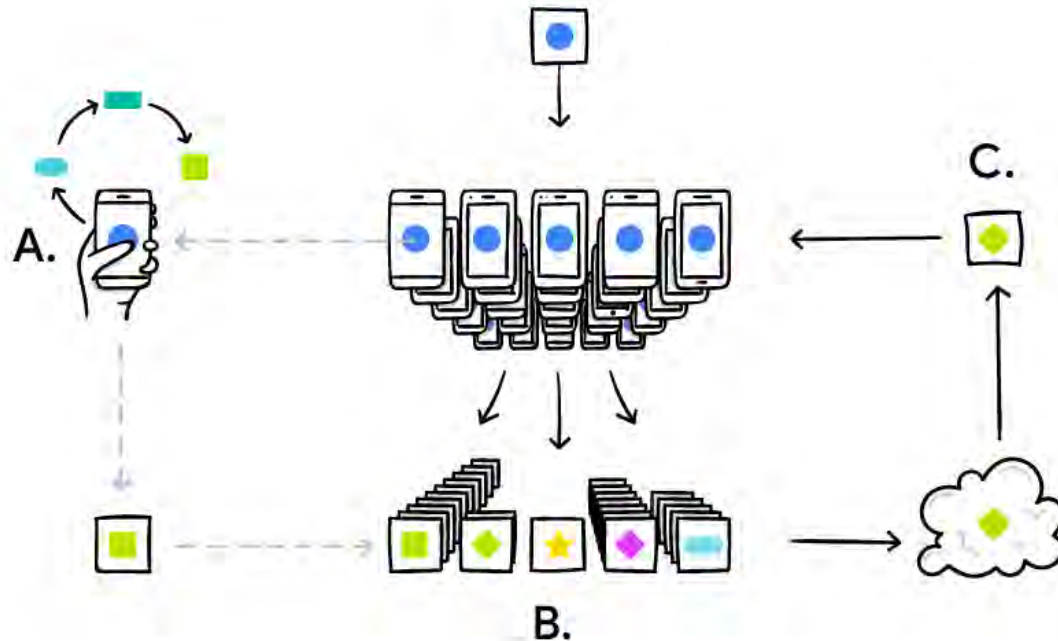
References:

- Federated Learning for Mobile Keyboard Prediction - Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Franc, oise Beaufays, Sean Augenstein, Hubert Eichner, Chlo'e Kiddon, Daniel Ramage
- Vertical Federated Learning - Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-QinZhang, Qiang Yang
- Federated Learning: Collaborative Machine Learning without Centralized Training Data (Google research blog) [Federated Learning: Collaborative Machine Learning without Centralized Training](#)



Federated Learning

- Federated learning was introduced in Google AI's 2016 blog.
- The standard federated learning process as follows:
 - Device download the current model and improve
 - Many devices updates, aggregates and send back
 - Updated shared model send back and repeat until model is fully trained



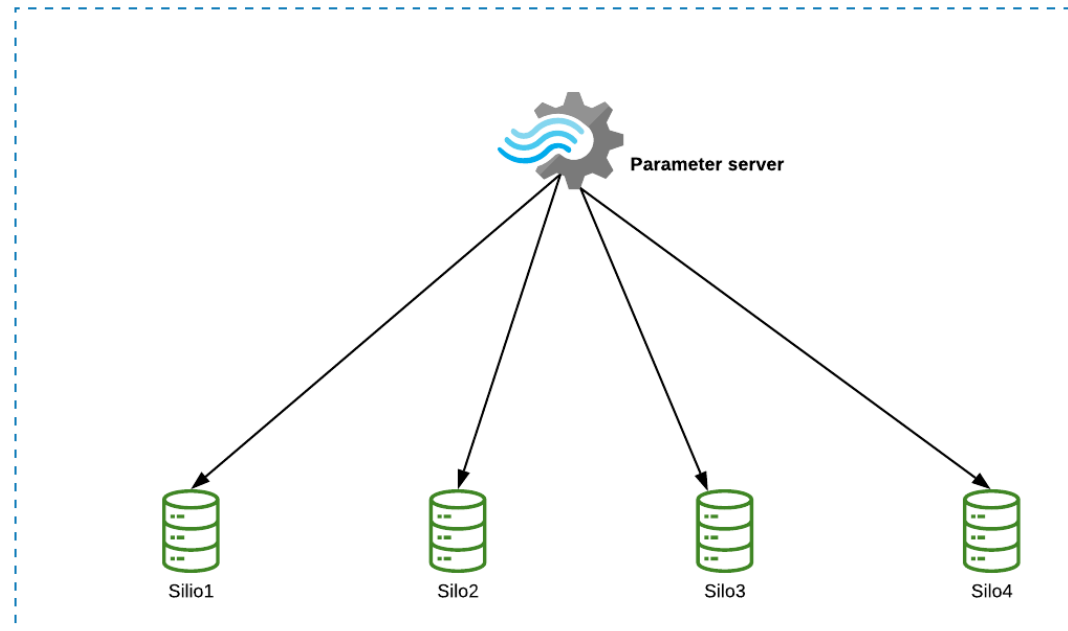
Federated Learning

Federated ML can be categorized into two base type:

1) Model-centric (Based on participating clients and the model training scale)

➤ **Cross-silo FLM :**

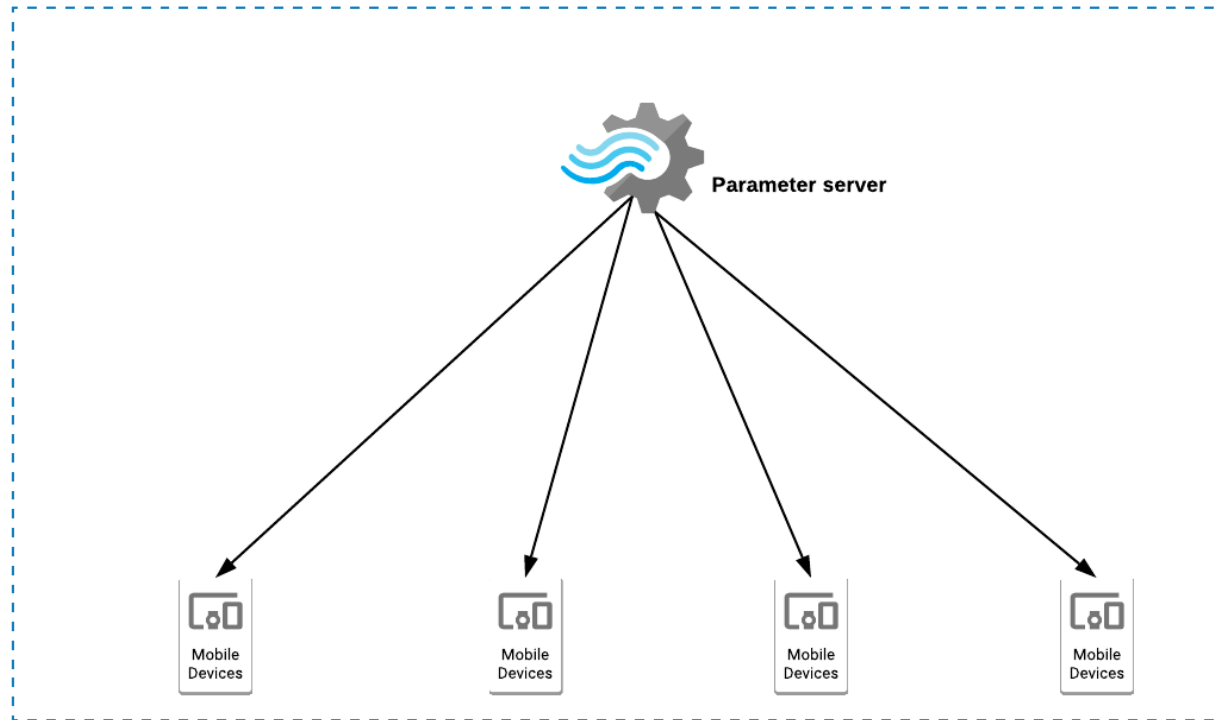
- o Silo is a segregated data storage contains unstructured, raw data
- o Involves learning across database for many users
- o Clients are organizations/companies here



Federated Learning

➤ Cross-device FLM:

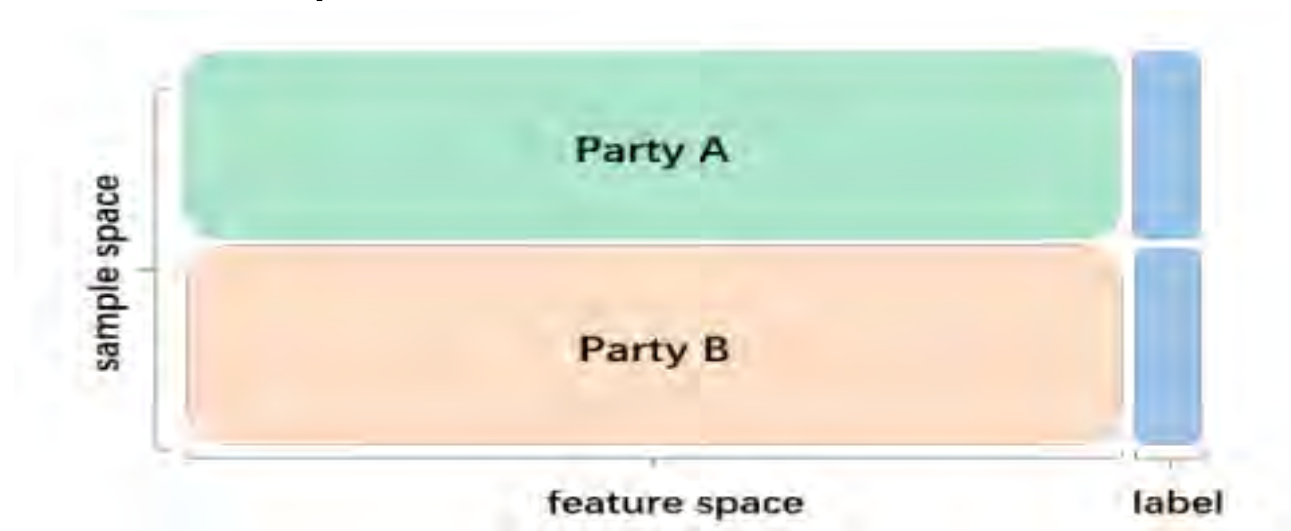
- o Deals with the data from individual devices within a network
- o End-user devices like laptop, mobile act as a source of data
- o Each user device generates a locally trained model
- o Many data points improve its prediction accuracy



Federated Learning

1) Data-centric (Based on data partitioned in sample and feature space)

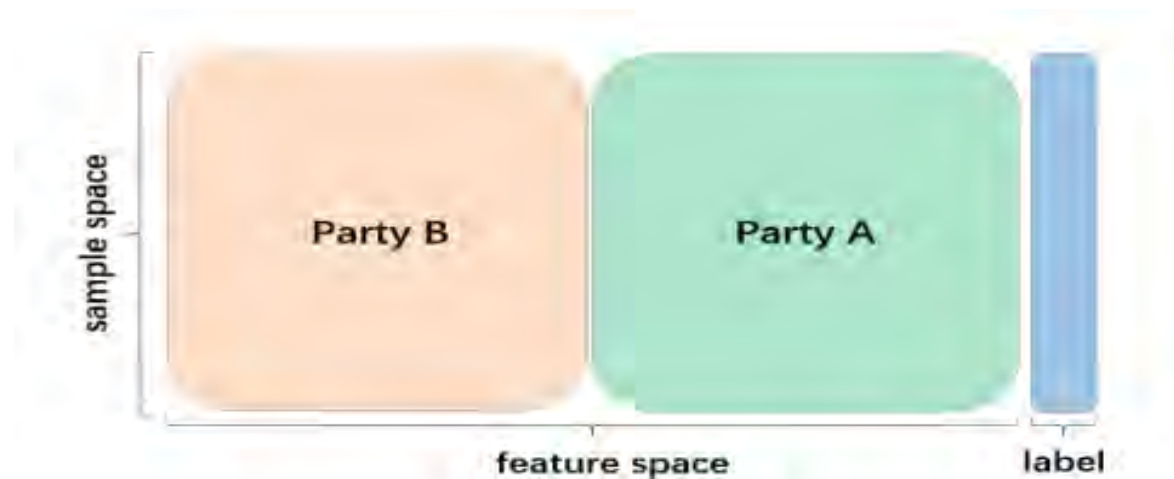
- **Horizontal FL** :- Where participants share the same feature space while holding different samples.
- Eg: Google uses next-word prediction model



(a) Horizontal Federated Learning

Federated Learning

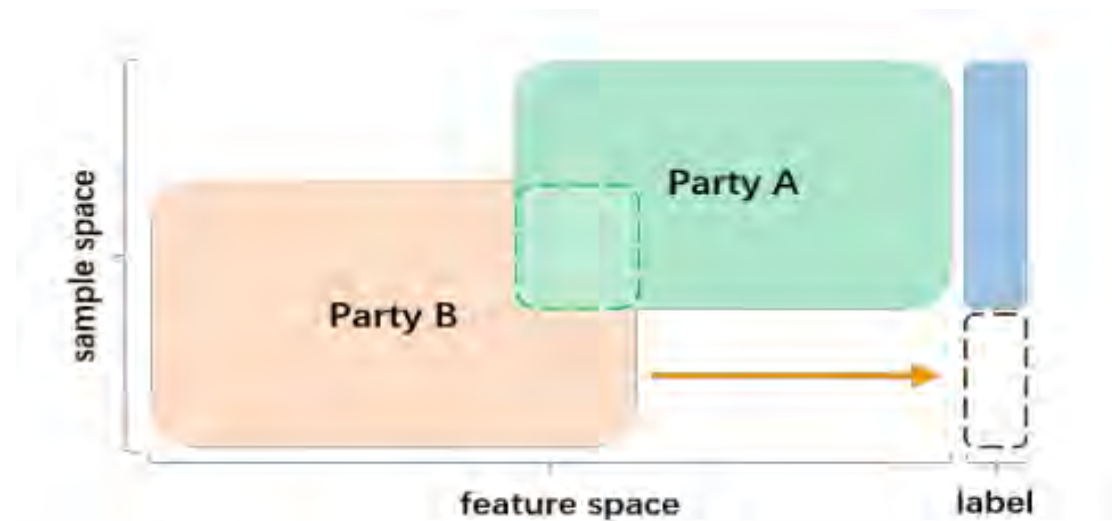
- **Vertical FL:-** Where datasets share same samples while holding different features
- Eg: A bank and an e-commerce company in same city. Their user sets are likely to contains most of the resident of the area.
 - Suppose we want both parties to have a prediction model for product purchased based on user and product info.



(b) Vertical Federated Learning

Federated Learning

- **Federated transfer learning:-** Where datasets differ in both feature and sample space with limited overlaps
- Eg: A group of university wanting to build a FL system to measure student performance.
- Movie recommendation for the user's past browsing behavior



(c) Federated Transfer Learning

Frameworks for FL

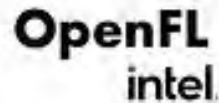
The logo for FATE (Federated Analytics Trusted Environment) features the word "FATE" in a bold, yellow, sans-serif font.The logo for TensorFlow Federated consists of an orange icon resembling a stylized 'TF' above the text "TensorFlow" in a black sans-serif font, with "federated" in a smaller, italicized font below it.The logo for NVIDIA FLARE features a green icon of a stylized eye or leaf above the text "NVIDIA FLARE" in a black sans-serif font.The logo for Flower (Federated Learning over eXtensible Workflows) features a black icon of a flower with a gear-like center above the word "Flower" in a black sans-serif font.The logo for OpenFL intel features the text "OpenFL" in a bold black sans-serif font above the word "intel" in a smaller, lowercase black sans-serif font.The logo for PySyft features a colorful, multi-segmented circular icon to the left of the text "PySyft" in a black sans-serif font.The logo for FedML (Federated Machine Learning) features a colorful icon of four overlapping circles (red, green, blue, yellow) to the left of the text "FedML" in a black sans-serif font.The logo for Federated Learning features the text "Federated Learning" in a white sans-serif font on a blue rectangular background, with a white icon of two hands shaking to the right.

Image taken from:

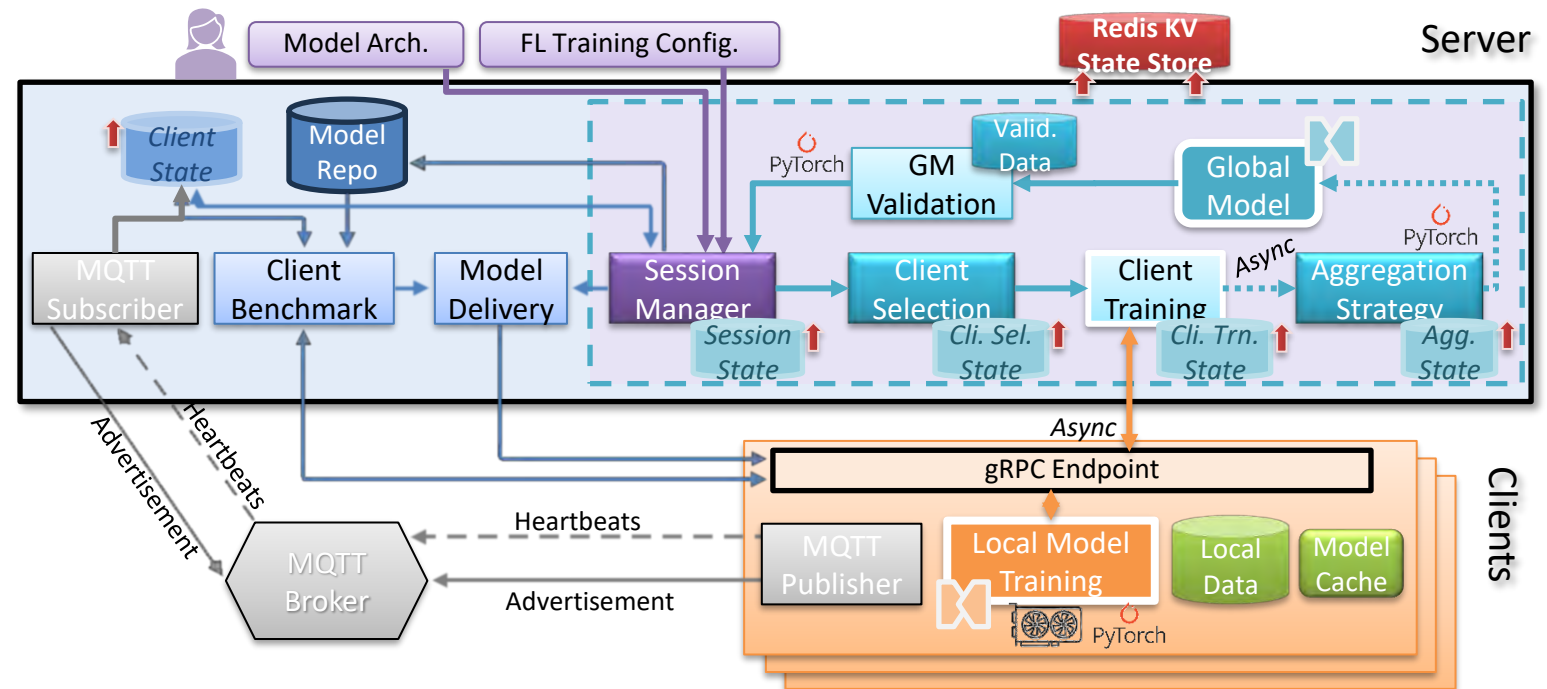
[Flower, FATE, PySyft & Co. — Federated Learning Frameworks in Python | by Alex Braungardt | ELCA IT | Medium](#)

Flotilla – Opensource Framework

- **Client and Server Discovery**

- Well known MQTT pub-sub broker
- Clients advertise their availability for training to a topic to which server subscribes
- Clients publish details about themselves such as their gRPC endpoint, etc
- Server advertise their availability to a separate topic, to which all clients subscribe

[GitHub - dream-lab/flotilla](https://github.com/dream-lab/flotilla): Flotilla: A Federated Learning Framework for Heterogeneous Resources
<https://github.com/dream-lab/flotilla>



Flotilla Features

Framework	Modularity					Resilience		Deployment					
	Custom Client Selection	Custom Model Aggregation		ML Framework		Server Reliability	Client Failure Handling	On-Device	Simulation	Containerization	Automatic Client Bootstrapping	Intelligent Client Selection for Session	Model Delivery
		Sync	Async	PyTorch	Tensor Flow								
FedML AI	✓	✓	X ¹	✓	✓	X	X	✓	✓	✓	X	X	X
FedScale	X ²	X ²	X ²	✓	✓	X	✓	✓	✓	✓	X	X	X ³
Flower	✓	✓	X	✓	✓	X ⁴	✓	✓	✓	✓	X	X	X
Leaf	X	X	X	X	✓	X	X	X	✓	X	X	X	X
OpenFL (Intel)	X	✓	X	✓	✓	X	X	✓	✓	✓	X	X	X
PySyft	✓	✓	X ⁵	✓	✓	X	✓	✓	✓	✓	X	X	X
TFF	✓	✓	X	X	✓	X	X	X	✓	X	X	X	X
Flotilla	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

1: Baselines for Asynchronous Strategies provided with the framework, supported in simulation.

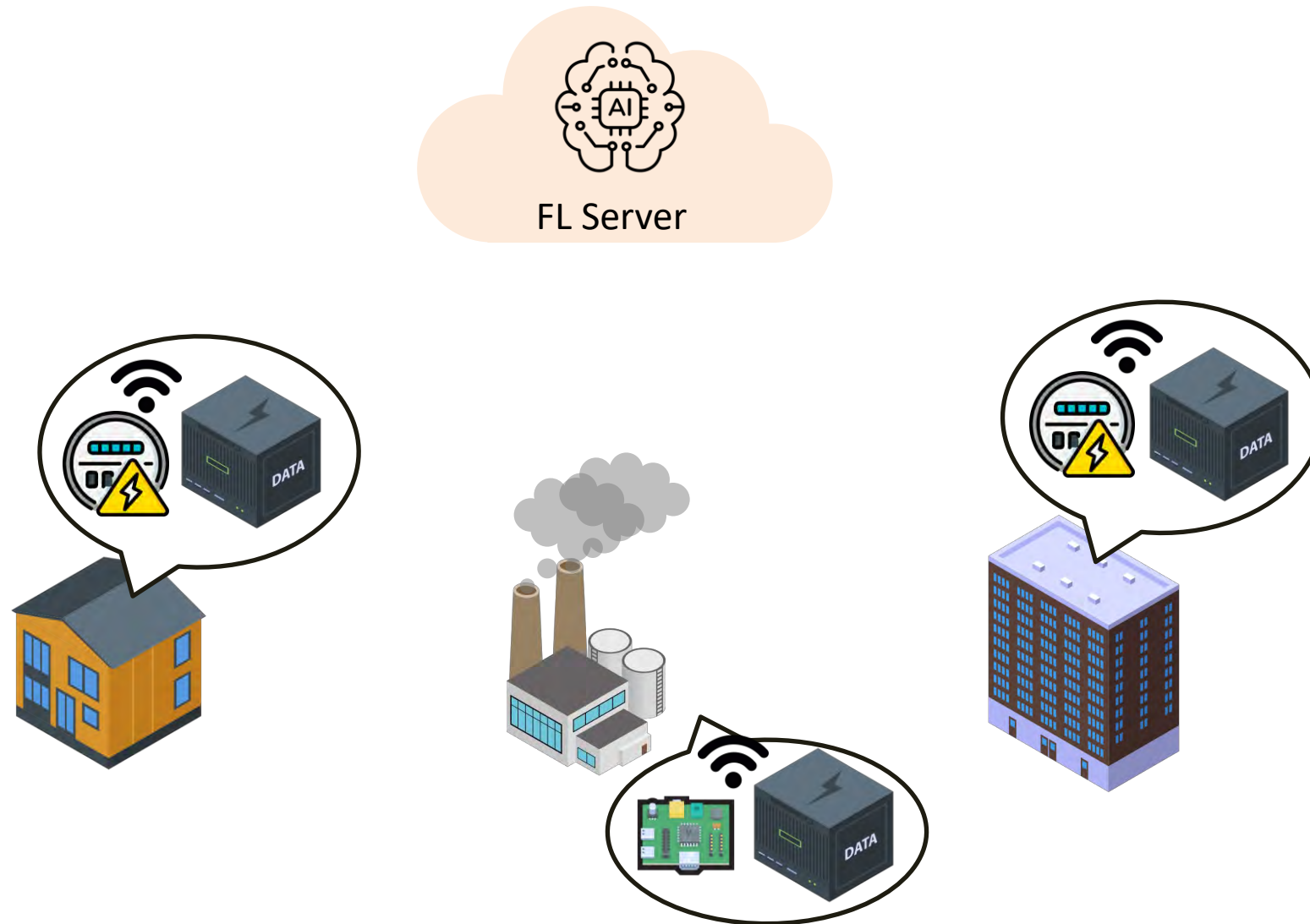
2: Authors claim implementation of custom strategy possible by modifying a module in the framework code.

3: Some update to the model layers possible using a config file.

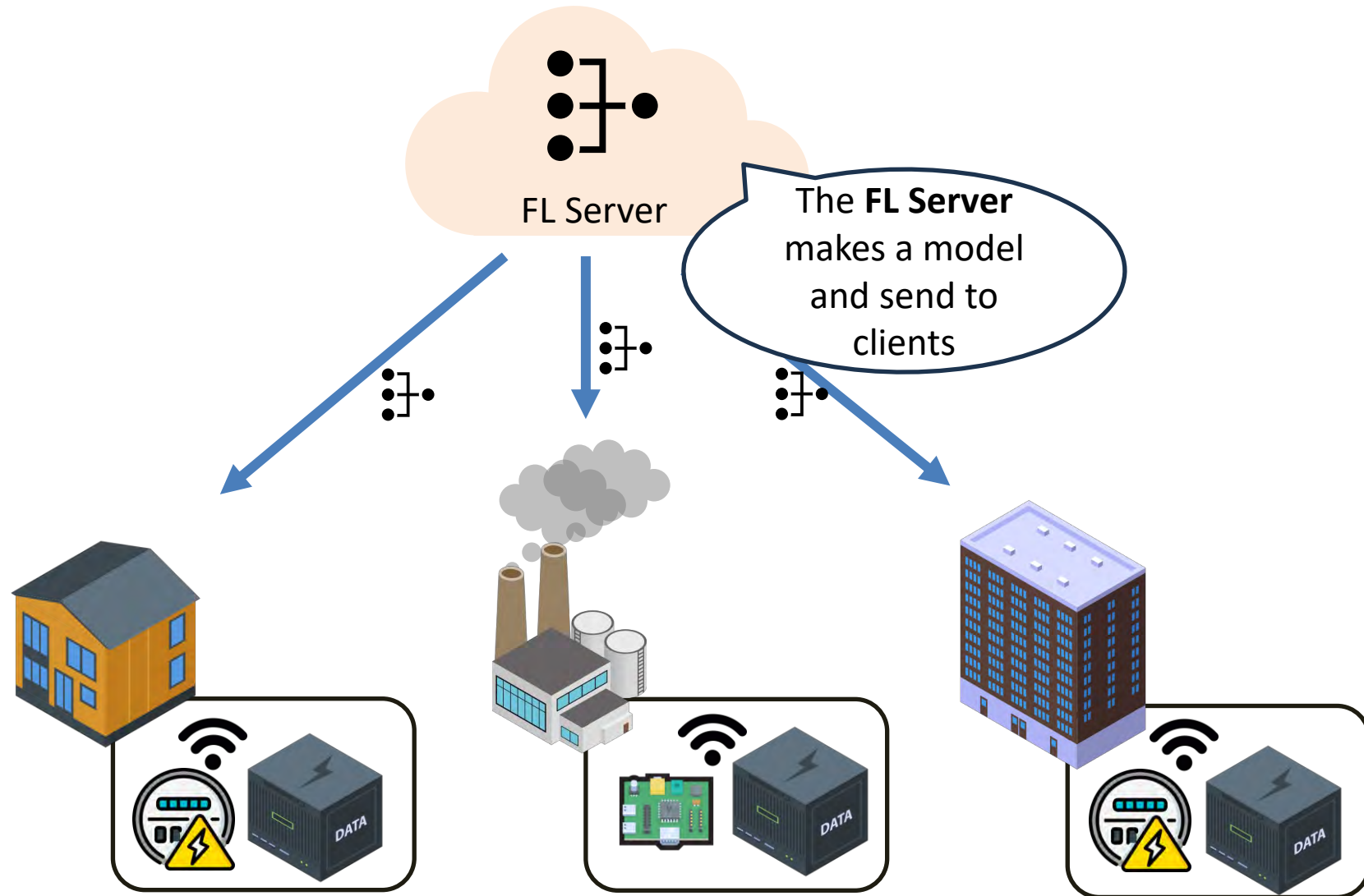
4: Experimental feature, allowing saving state to a SQLite Database. Were not able to validate.

5: Multiple papers have used PySyft to implement Asynchronous Strategies. No description of Async on PySyft documentation, which itself is very vague and not very helpful.

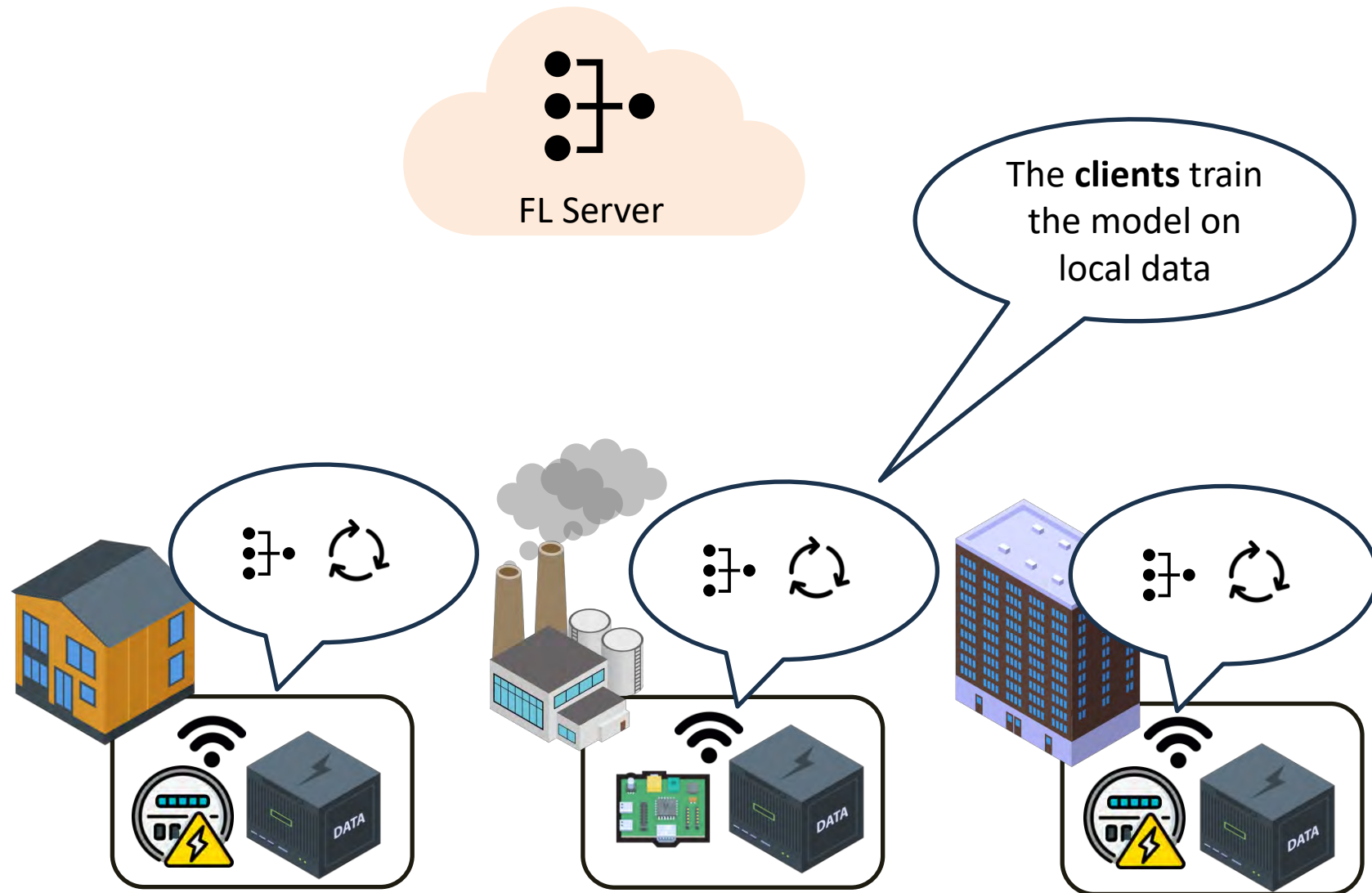
Federated Learning on Smart Meters



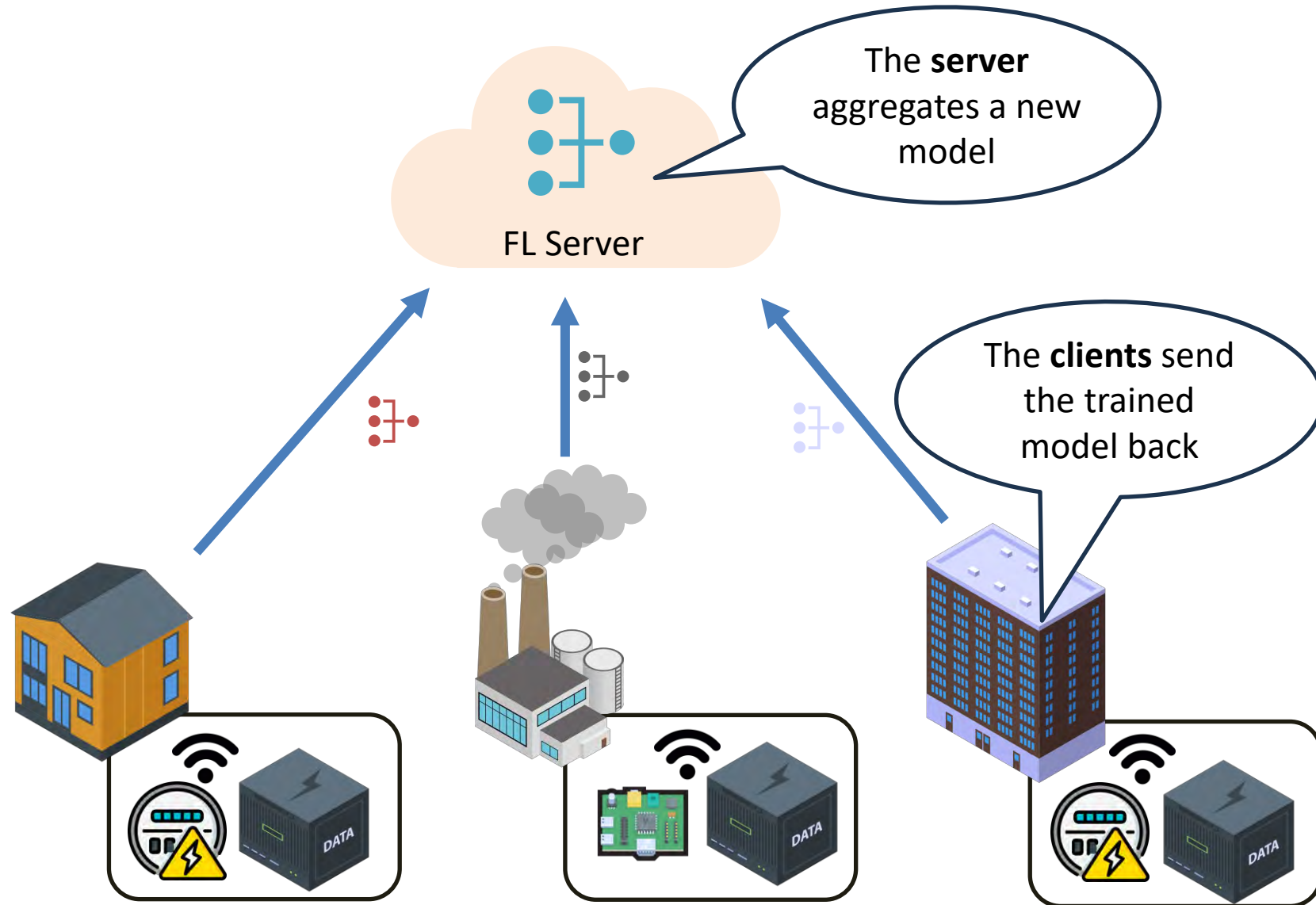
Federated Learning on Smart Meters



Federated Learning on Smart Meters



Federated Learning on Smart Meters

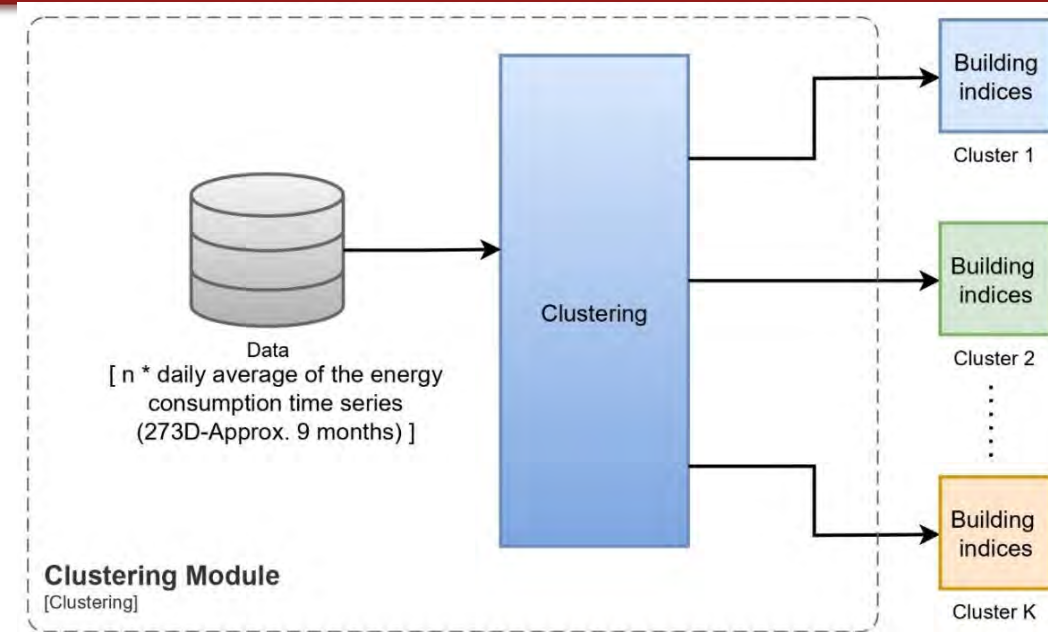


Key challenges in FL

- **Data Heterogeneity (Non-IID):**
 - Clients in a microgrid have widely diverse energy usage patterns due to differences in building types, operational hours, and appliances.
 - Introduces skews in training and can prevent a single global model from converging to high accuracy
- **Limited Resources:**
 - Smart meters have very low compute onboard – the model must be very light
- **Scalability:**
 - Training individual models for thousands of consumers is computationally expensive and unsustainable
 - Must train using data from a small, representative subset of clients
- **Prediction accuracy:**
 - Model's accuracy must be stable even for longer-ranged forecasts (60mins)

Methodology

- **Clustering Techniques:**
 - To address data heterogeneity, clients with similar consumption patterns are grouped into clusters using **K-means clustering** on daily averaged energy consumption data.
 - This allows for training **separate, specialized federated models** for each cluster, which improves forecasting accuracy compared to a single model trained on all clients.
- **Forecasting Models:**
 - The primary models evaluated are Deep Neural Networks (DNNs), specifically **Long Short-Term Memory (LSTM)** and **Gated Recurrent Unit (GRU)** models, which are well-suited for time-series forecasting.
- **Custom Loss Function:**
 - An **Exponentially Weighted Mean Squared Error (EW-MSE)** loss function is used for training.
 - This function assigns exponentially increasing weights to prediction errors at farther time steps, penalizing inaccuracies in harder-to-predict future values.



Clustering for using Data Summaries

- **Problem:** Client-level data heterogeneity can degrade the performance of a global federated learning model
- **Solution:**
 - Clustering groups consumers with similar time-series consumption patterns to address this data heterogeneity.
 - Training separate models for each cluster improves forecasting accuracy by tailoring the model to the specific characteristics of that group.
- **Methodology:**
 - **Data Representation:** A vector is created for each building by calculating the **daily average** of its energy consumption (in kWh) over a period of 9 months.
 - **Clustering Algorithm: K-means clustering** is used to partition consumers into distinct clusters based on the similarity of these vectors.
 - **Model Training:** Following clustering, a separate forecasting model is trained for each cluster.
- **Benefits:** This approach allows for the creation of more generalized and robust forecasting models by incorporating the unique characteristics of each consumer segment.

Forecast quality metrics

❑ Mean Absolute Error:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

❑ Mean Absolute Percentage Error:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \cdot 100\%$$

Accuracy = Complement of MAPE

❑ Root Mean Square Error:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

❑ R squared Score:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

$\hat{y}$ - predicted value of y

$\bar{y}$ - mean value of y

y_i - actual value of y

Models and Datasets

- RNN models – **LSTM**
- Baseline - **SArima**
- Dataset: **OpenEIA dataset** - Time series data of energy usage in kilowatt-hours (kWh) at **15-min granularity** for residential and commercial buildings.
- We choose data for commercial buildings in the states of California (CA), Florida (FLO) and Rhode Island (RI).
- We use Min–Max scaling over the entire data.
- We use **look-back window of 8 steps (2 hours)** and a **look-ahead horizon of 4 steps (1 hour)**.
- The training and testing sets are a 75:25 ratio

Exponentially Weighted Mean Squared Error (EW-MSE)

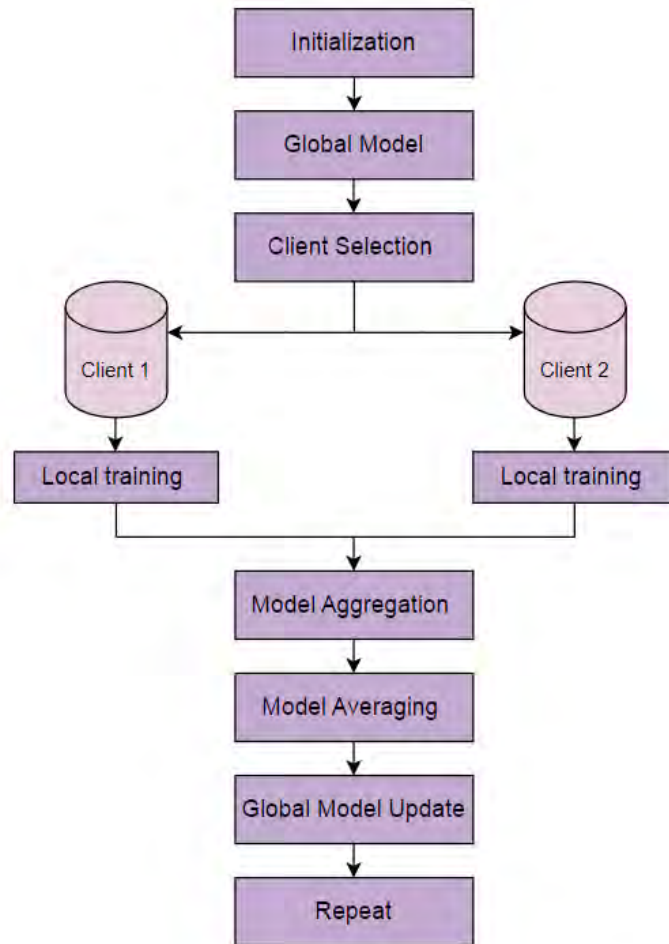
$$\text{EW-MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N \beta^{i-1} \cdot (y_i - \hat{y}_i)^2$$

$\beta > 1$ is a tunable hyperparameter that controls the degree of emphasis placed on future predictions.

When $\beta = 1$, EW-MSE reduces to the standard MSE.

By amplifying the penalty for errors further ahead in the prediction horizon, EWMSE encourages the model to more carefully learn the dynamics governing long-range behavior

Federated Averaging workflow - FedAvg



Algorithm 1 FederatedAveraging. The K clients are indexed by k ; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:

```
initialize  $w_0$ 
for each round  $t = 1, 2, \dots$  do
   $m \leftarrow \max(C \cdot K, 1)$ 
   $S_t \leftarrow$  (random set of  $m$  clients)
  for each client  $k \in S_t$  in parallel do
     $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$ 
   $w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$ 
```

ClientUpdate(k, w): // Run on client k
 $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
for each local epoch i from 1 to E do
 for batch $b \in \mathcal{B}$ do
 $w \leftarrow w - \eta \nabla \ell(w; b)$
return w to server

- K - Number of clients
- C - Fraction of clients participating in that round
- S_t - Random set of clients
- n_k - No. of data points held by client k
- n – Total no. of data points held by all clients
- W_{t+1}^k - Model weights on communication round on client k
- B - Local minibatch size used for client update
- E - Local rounds
- P_k - Set of datapoints on client k

Federated Proximal work flow - FedProx

- Federated Proximal is a way of making sure computer models learn better together.
- It adds a special term called regularization term that helps the models on different devices work well together.
- This term makes sure that the individual models on each device aren't too different from the overall model.
- This is helpful when the data on different devices isn't exactly the same.
- FedProx is good in situations where devices have different amounts of data or when the data is varied.

FedProx algorithm (Proposed Framework):

1. **Input:** $K, T, \mu, \gamma, w^0, N, p_k, k = 1, \dots, N$
2. for $t = 0, \dots, T - 1$ do
3. Server selects a subset S_t of K devices at random (each
4. device k is chosen with probability p_k)
5. Server sends w^t to all chosen devices
5. Each chosen device $k \in S_t$ finds a w_k^{t+1} which is
6. a γ_k^t -inexact minimizer of: $w_k^{t+1} \approx$
7. $\arg \min_w h_k(w; w^t) = F_k(w) + \frac{\mu}{2} \|w - w^t\|^2$
8. Each device $k \in S_t$ sends w_k^{t+1} back to the server
9. Server aggregates the w' 's as $w^{t+1} = \frac{1}{K} \sum_{k \in S_t} w_k^{t+1}$
10. **End for**

where,

μ - hyperparameter controlling the strength of the proximal term

$F_k(w)$ - objective function used for training the model on client k with the current weights w

Large Dataset for Load Forecasting

- Part of the Open Energy Data Initiative (**Open EIA**) from US Department of Energy)
- Contains energy consumption of the US residential and commercial buildings
- 15-minute spaced data like “total.energy_consumption”, “fans.energy_consumption”, “water_systems.energy_consumption” .
- We have experimented with “out.site_energy.total.energy_consumption” for buildings in the state of California, Florida and Rhode Island for the year 2018.
- The dataset spans just one year.

Source- ¹ oedi-data-lake → nrel-pds-building-stock → end-use-load-profiles-for-us-building-stock → 2023 → comstock_amy2018_release_1 → timeseries_individual_buildings → by_state → upgrade%3D10 → state%3DCA

https://data.openei.org/s3_viewer?bucket=oedi-data-lake&prefix=nrel-pds-building-stock%2Fend-use-load-profiles-for-us-building-stock%2F2023%2Fcomstock_amy2018_release_1%2Ftimeseries_individual_buildings%2Fby_state%2Fupgrade%3D10%2Fstate%3DCA%2F

Open EAI Data Set - Statistics

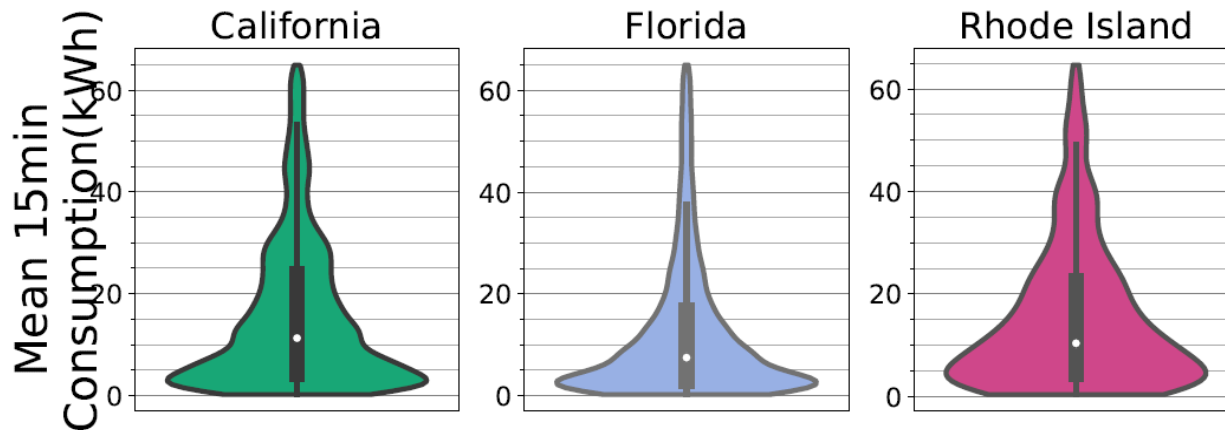
State	# of buildings	# of train buildings	# of test buildings	Data Year	Samples per Building
California (CA)	39,391	100	39,291	2018	35,040
Florida (FLO)	24,444	100	24,344	2018	35,040
Rhode Island (RI)	1376	100	1376	2018	35,040

The data is reported at 15 min granularity, and 1 year of data is available.

data for commercial buildings in the states of California (CA), Florida (FLO) and Rhode Island (RI).

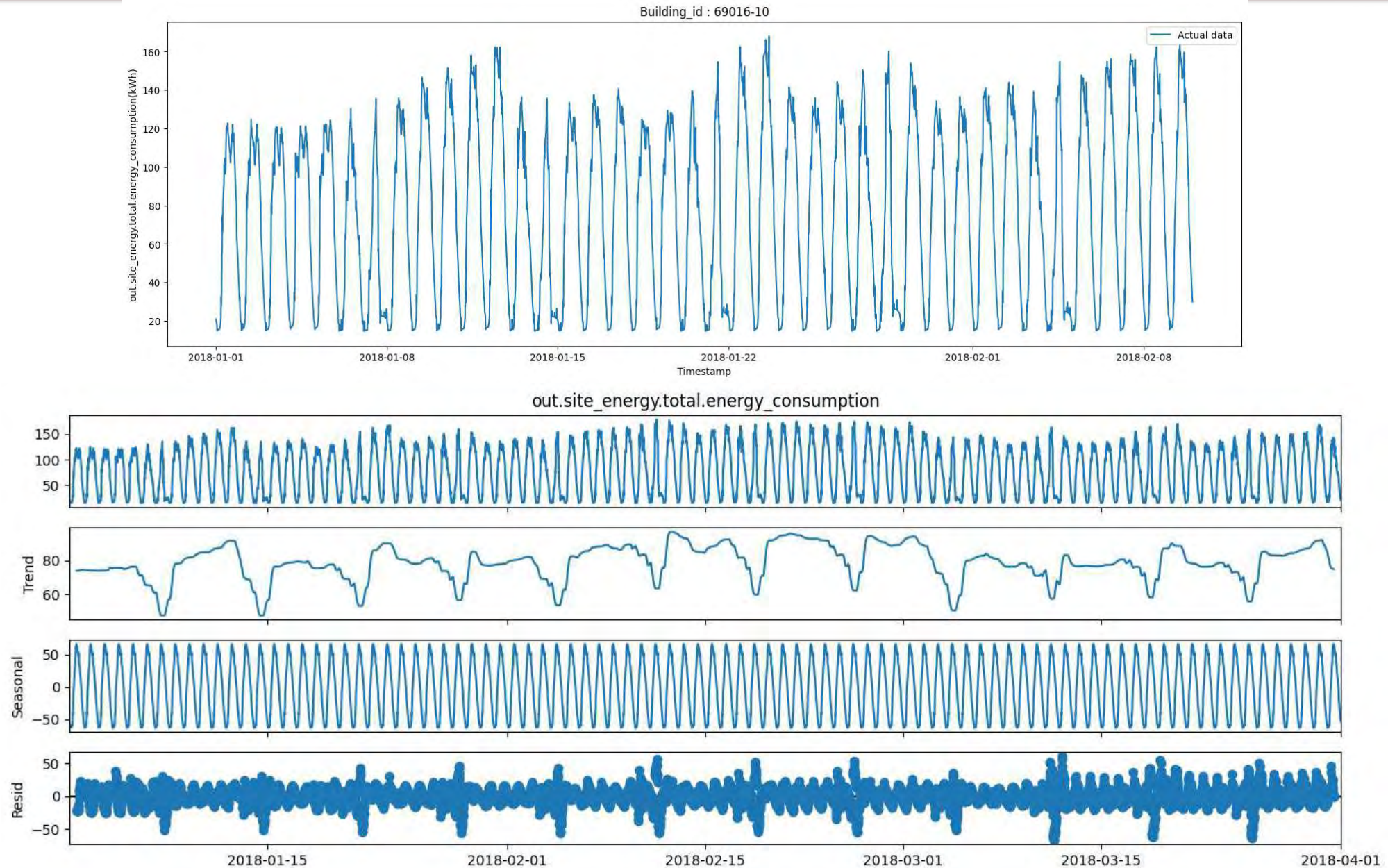
The minimum consumption is at 0.16 kWh, the maximum at 63.8 kWh, median at 12.7 kWh, Q1 and Q3 are at 4.7 kWh and 28.4 kWh.

Around 3200 buildings have mean consumption greater than 63.8 kWh.



Violin Plots for mean consumption over 15 minute kWh data for commercial buildings in CA, FLO and RI

Sample Data



Last 40 days data plot of building id- 69016

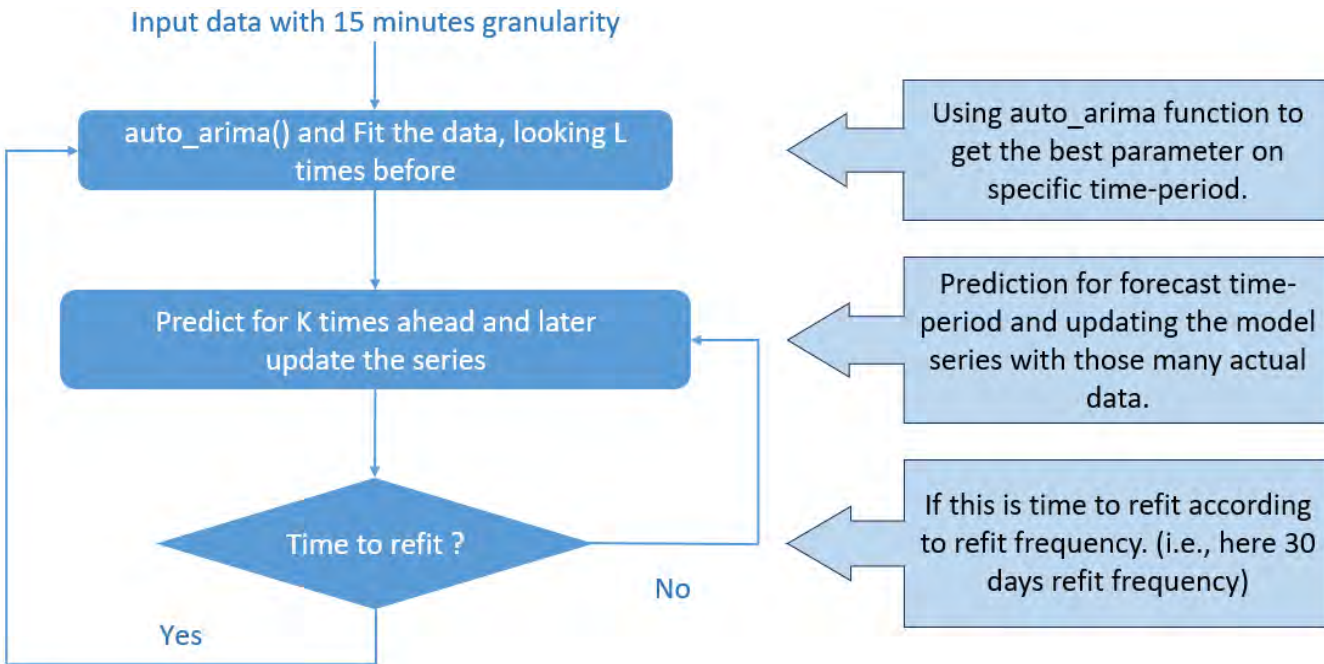
Data Preparation and System Configuration

- Each building functions as an independent FL client, resulting in a total of 100 clients whose consumption data remain private to each participant.
- Each client trains a local LSTM model on its own 15-minute interval usage data, which is first normalized to the $[0, 1]$ range using Min–Max scaling over the entire year.
- We frame the forecasting task using a look-back window of 8/16 steps (2/4 hours) and a look-ahead horizon of 4 steps (1 hour).
- The resulting time series for each client is partitioned into training and testing sets in a 75:25 ratio – approximately the first nine months of data for training and the final three months for evaluation.
- The federated training proceeds for 500 rounds, with each client performing local updates before the server aggregates model parameters according to our proposed FL scheme.

K-means Clustering Hyperparameters

- We set $k = 4$ clusters, as observed to be the optimal number of clusters from an elbow plot, and sample 25 buildings
- per cluster for training across 100 buildings per state.
- The client consumption vector used for clustering t_p is the 273-dimensional daily average consumption (≈ 9 months) per building.
- We train 5 different flavors of models.
- While F_A is the FedAvg FL model trained on all 100 buildings,
- F_{C1} , F_{C2} , F_{C3} and F_{C4} are the models for each of the 4 clusters, each trained on 25 buildings.
- We validate against a large held-out set per state, and on a smaller set of 30 buildings on a per-cluster basis.

BASELINE - SARIMA - Load Forecasting



Training prediction loop method

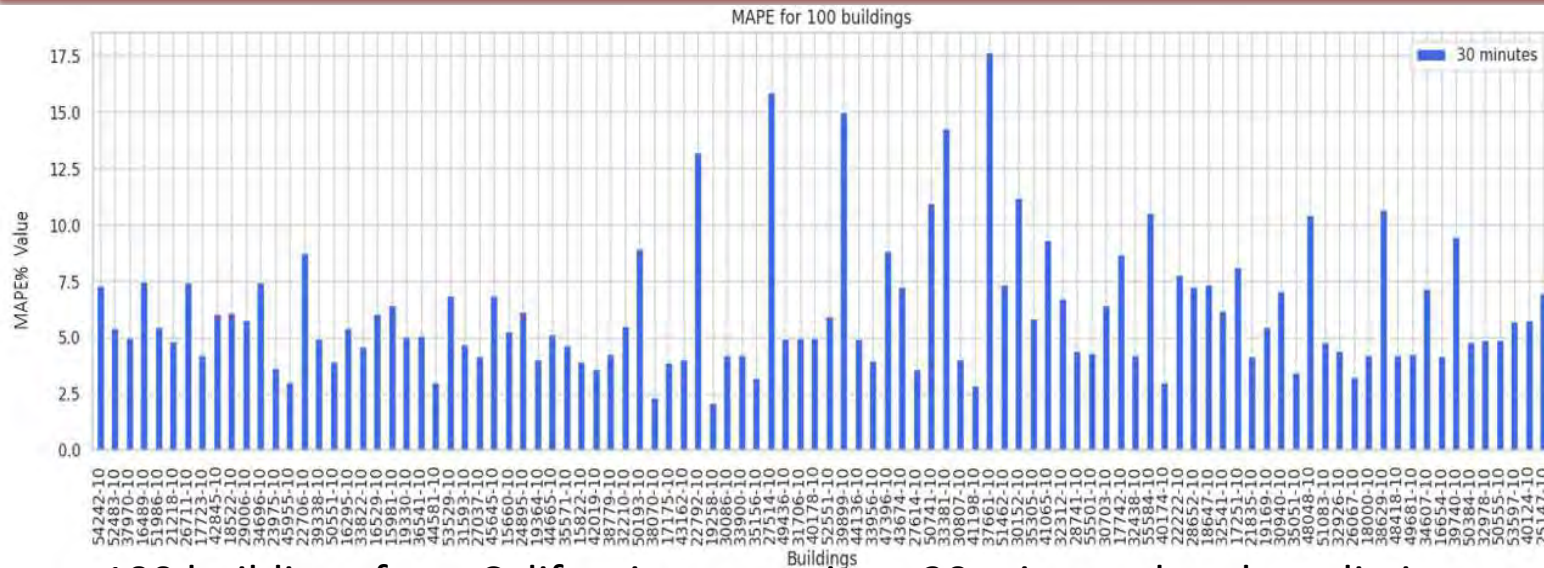
- Initially train SARIMA model over a 30 days time period
- After making predictions for a specific duration (i.e 30 minutes and 1 hour), periodically recalibrate the model's parameters using the auto arima() function
- This iterative approach involves re-estimating the model's parameters at regular 'refit' intervals (i.e here 30 days),

Quartile-based clustering is a method that involves grouping time series data based on their positions within quartiles, essentially dividing the dataset into four equal parts. This approach assigns data points to clusters according to their quartile memberships, facilitating the identification of unique patterns and behaviors associated with each quartile

Building id	Mean (kWh/15 min)	Standard Deviation
48605 (min)	0.16	0.11
22561	2.33	0.77
20752 (Q1)	4.70	1.49
15228	6.64	4.51
55173 (Q2/median)	12.71	8.5
34637	18.18	6.10
44941 (Q3)	28.36	11.91
15221	56.45	17.94
54003 (max)	1075.16	388.49

Distribution of 9 selected buildings from 4 different Quartiles on California dataset

SARIMA - Load Forecasting – Results

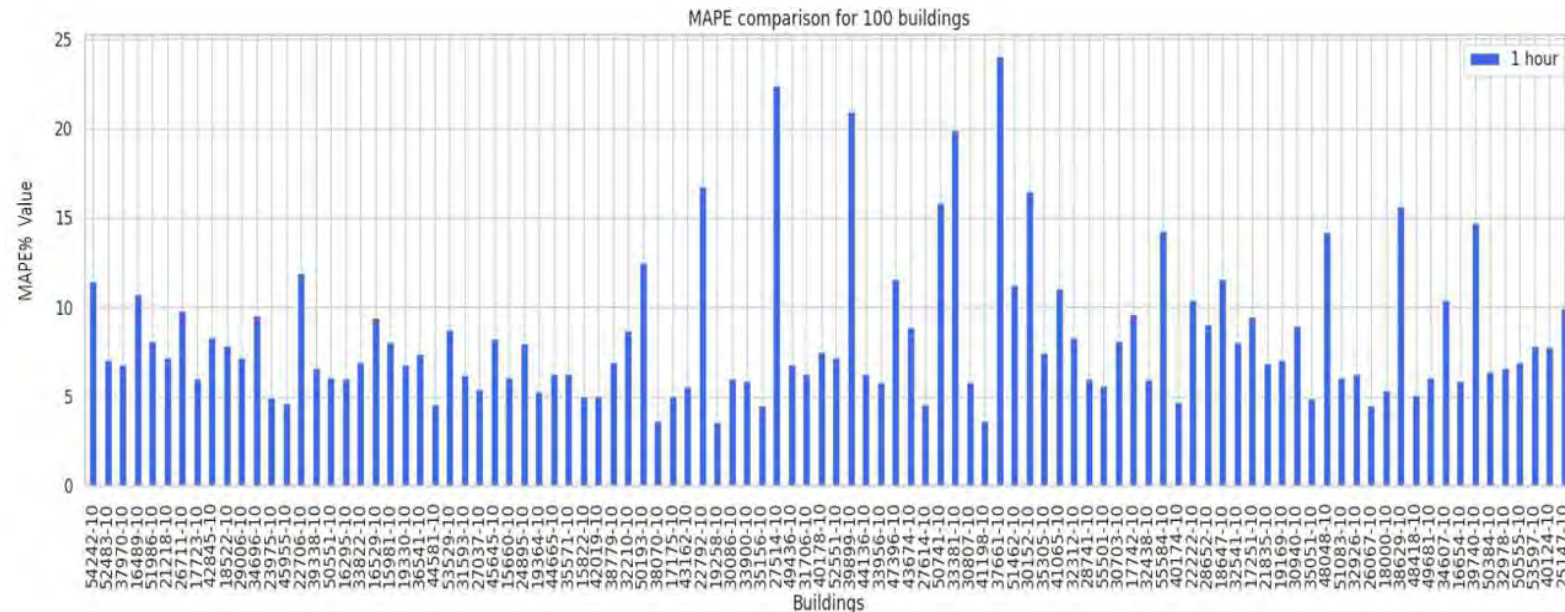
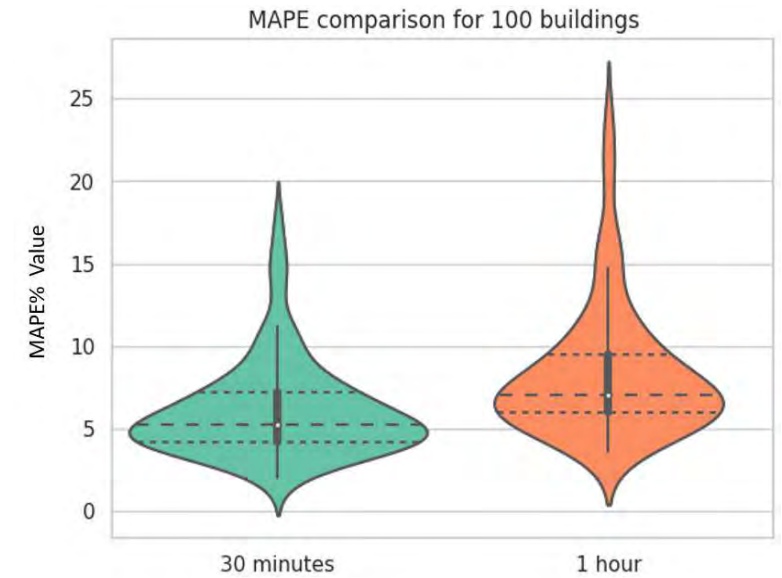


100 buildings from California state using - 30 minute ahead predictions

we sample 25 buildings for training from each quartile

For 30 minutes predictions, more than 50 buildings having less than 5% MAPE value and average MAPE for 30 minute forward prediction is 5.35 for 100 buildings

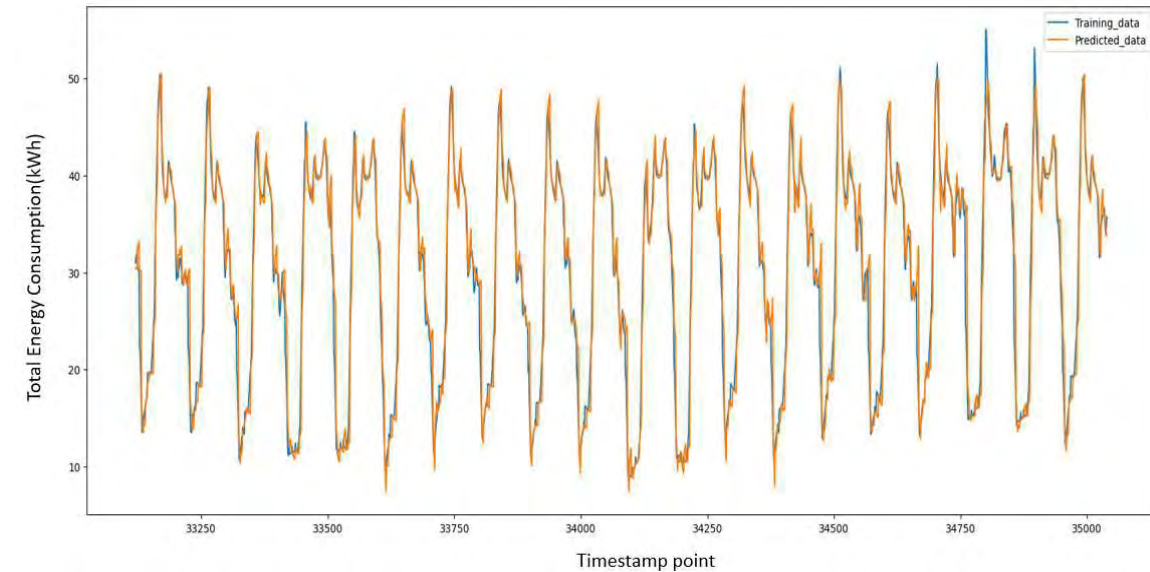
For 1 hour predictions, more than 75 buildings having less than 9% MAPE value and average MAPE for 1 hour ahead prediction is 7.85 for 100 buildings.



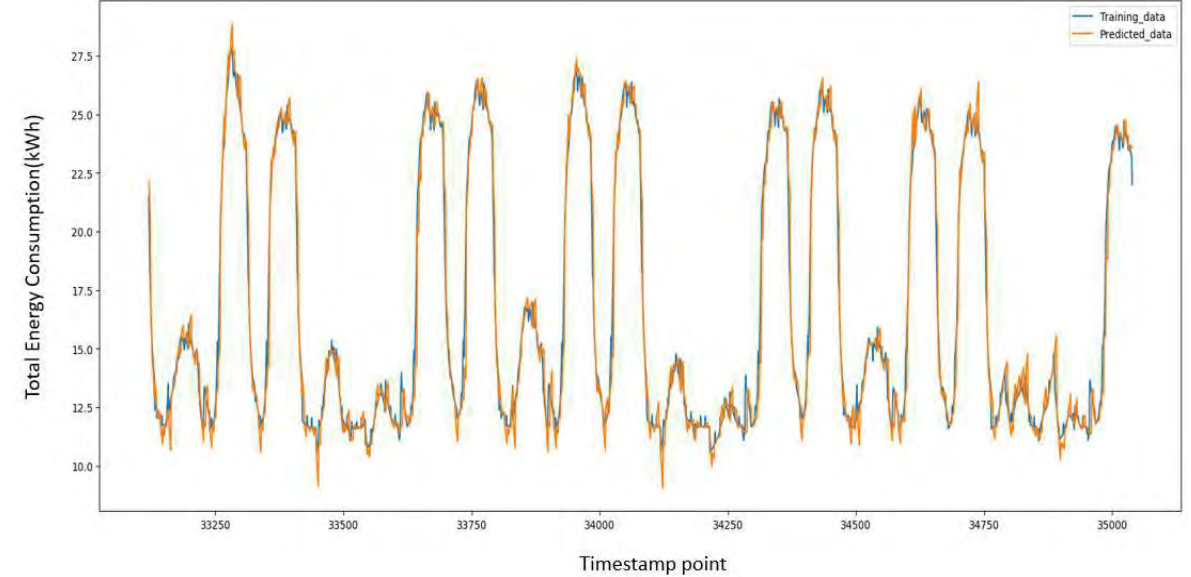
100 buildings from California state using – 1 hour ahead predictions

SARIMA - Load Forecasting Results - Quartile Cluster

Initially, 9 buildings are selected for preliminary evaluation such that their mean is distributed across all four quartiles across all commercial buildings in CA for the year 2018 towards ensuring diversity in consumption



December month predictions for 1 hour for building-id 44941



December month predictions for 1 hour for building-id 34637

Building id	MAPE%		SMAPE		MAE		MRSE	
	30 min	1 Hour	30 min	1 Hour	30 min	1 Hour	30 min	1 Hour
48605 (min)	21.28	31.08	13.42	20.15	0.01	0.02	0.03	0.04
22561	3.22	5.04	3.21	5.05	0.07	0.11	0.12	0.16
20752	3.14	4.14	3.15	4.13	0.14	0.18	0.20	0.27
15228	3.72	6.22	3.73	6.28	0.21	0.366	0.36	0.59
55173	9.45	12.85	7.78	10.74	1.07	1.47	3.73	4.23
34637	2.14	3.23	2.15	3.27	0.36	0.55	0.57	0.89
44941	5.10	6.49	5.19	6.41	1.09	1.41	1.64	2.19
15221	4.17	5.92	4.13	5.89	1.87	2.63	3.25	4.57
54003 (max)	2.20	3.28	2.18	3.27	22.25	33.32	38.25	52.19

4 hours lookback and 60 min ahead predictions on (SARIMA)

K-means clustering

California state data

	15min	30min	45min	60min	MAPE	SD_MAPE	SMAPE	SD_SMAPE	RMSE	SD_RMSE
Cluster1	96.43	95.74	93.86	93.20	6.80	3.59	6.36	2.51	4.70	7.08
Cluster2	95.93	95.08	93.70	92.61	7.39	3.29	7.39	3.29	6.22	15.80
Cluster3	94.73	93.64	90.66	89.16	10.84	6.82	9.10	4.00	5.14	5.52
Cluster4	95.76	94.63	92.90	91.86	8.14	4.90	7.79	3.55	2.27	3.04
Average	95.71	94.77	92.78	91.70	8.29	4.65	7.66	3.34	4.59	7.86

We sample 25 buildings per cluster for training and 30 buildings per cluster for testing i.e. 120 held out set (30 from each cluster)

	cluster1	cluster2	cluster3	cluster4
Cluster1	93.20	91.91	86.73	91.08
Cluster2	93.29	92.61	87.78	91.83
Cluster3	93.21	93.19	89.16	92.60
Cluster4	93.42	92.76	88.28	91.86

4 hours lookback and 60 min ahead predictions (SARIMA)

MAPE

Florida dataset results:

	cluster1	cluster2	cluster3	cluster4
Cluster1	87.66	87.10	88.64	90.86
Cluster2	84.97	83.90	86.78	88.34
Cluster3	87.77	87.17	88.70	90.79
Cluster4	87.26	86.47	88.51	90.29

Rhode Island dataset results:

	cluster1	cluster2	cluster3	cluster4
Cluster1	85.14	88.61	81.46	89.40
Cluster2	83.94	87.84	79.64	88.14
Cluster3	84.17	88.11	81.77	88.77
Cluster4	84.81	88.38	81.58	89.33

Results for 39k California Buildings (SARIMA model)

California	15_min	30_min	45_min	60_min	Overall accuracy
Average Accuracy	95.93	92.68	90.11	87.86	91.64
SD	2.29	4.40	5.62	6.72	4.75

- Results on 39k test building with individual models trained for each 15-minute prediction with model built on 100 train buildings using FedAVG

	15min accu	30min accu	45min accu	60min accu	AVG Accu
F_100 (100/50)	91.94	89.63	87.36	85.13	88.515

Results for 39k California Buildings (FedAvg) with 2-hour lookback, 1-hour

Average Accuracy

	F_{100}	F_{C1}	F_{C2}	F_{C3}	F_{C4}	
Cluster1	93.06					
Cluster2	94.22		93.66			
Cluster3	91.19					
Cluster4	94.33					
Average	93.20					

Average Std. Deviation

	F_{100}	F_{C1}	F_{C2}	F_{C3}	F_{C4}	
Cluster1	4.87					
Cluster2	3.55		4.50			
Cluster3	7.24					
Cluster4	3.85					
Average	4.87					

F_{100} - Model trained on 100 sampled buildings using FedAVG.

F_{C1} - Model trained on 25 sampled buildings in cluster1 using FedAVG.

F_{C2} - Model trained on 25 sampled buildings in cluster2 using FedAVG.

F_{C3} - Model trained on 25 sampled buildings in cluster3 using FedAVG.

F_{C4} - Model trained on 25 sampled buildings in cluster4 using FedAVG.

Results for 24k Florida Buildings (SARIMA model)

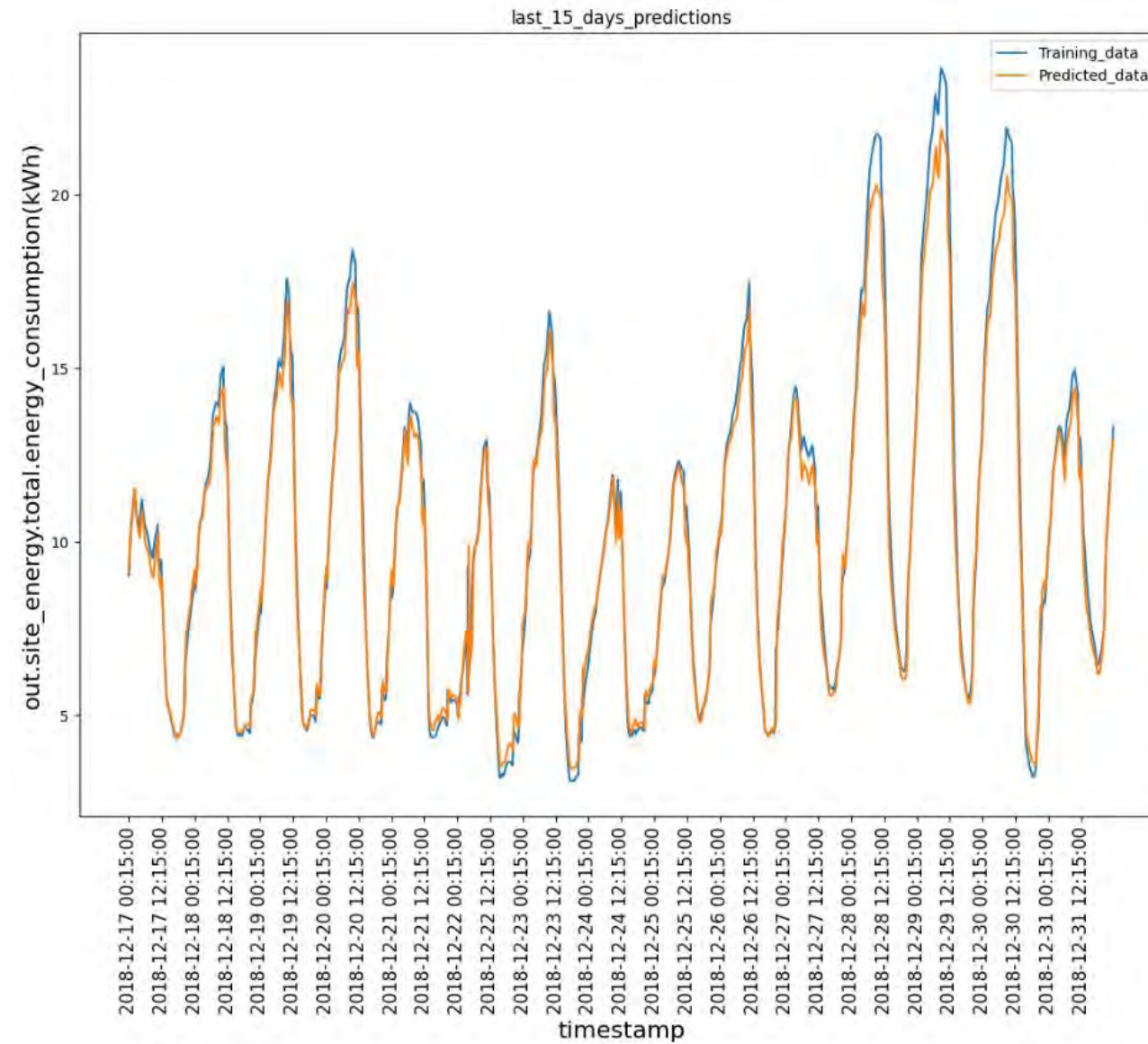
Florida	15_min	30_min	45_min	60_min	Overall accuracy	MAPE
Average Accuracy	92.98	88.24	84.53	81.36	86.77	
	Overall Accuracy	15min	30min - noncumulative	45min - noncumulative	60min - noncumulative	
F_100(100/50)	82.62	88.58	84.75	80.72	76.62	
F_100(500/5)	82.46	88.55	84.48	80.45	76	

Results for 1276 Rhode Island Buildings (SARIMA model)

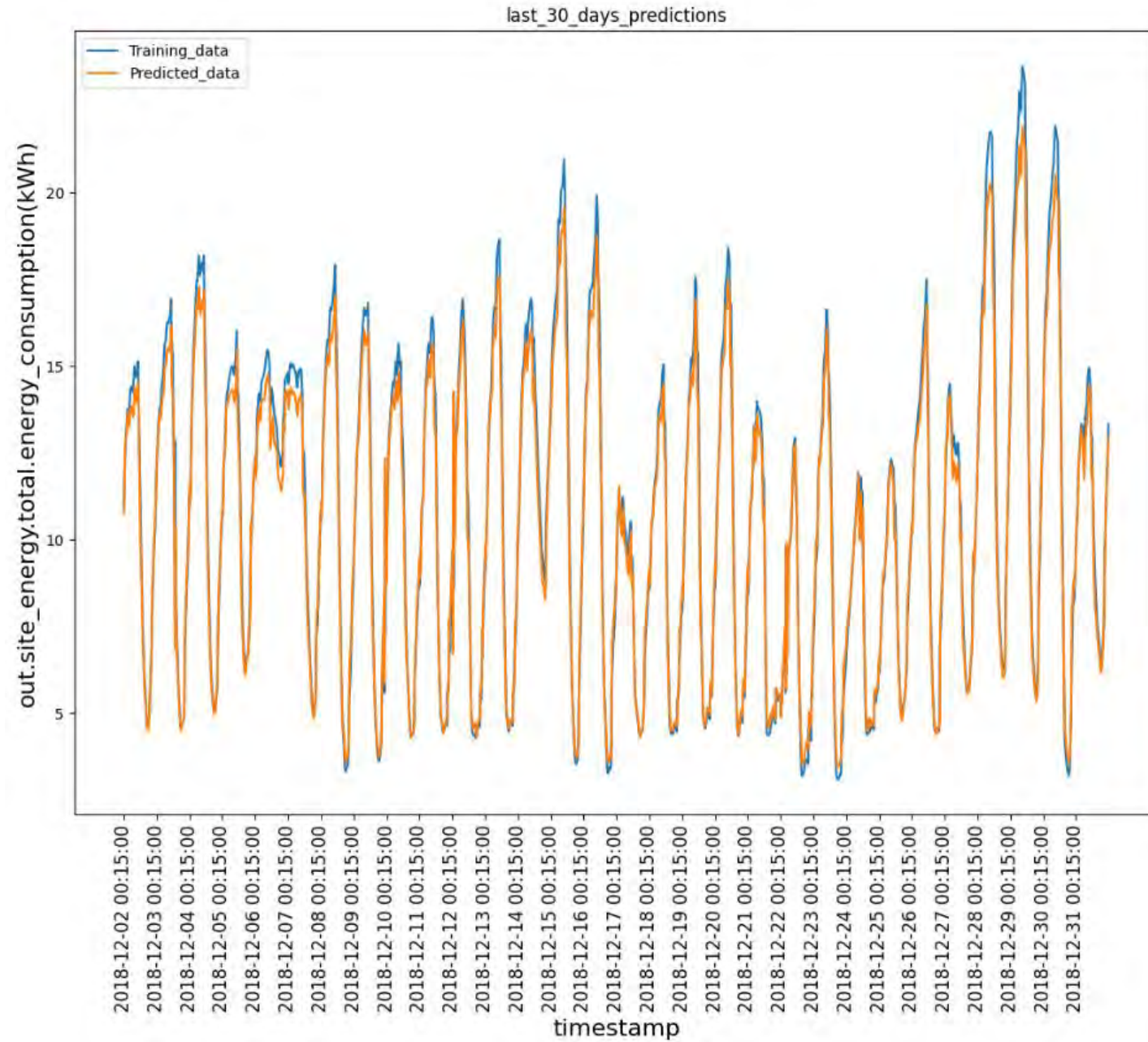
Rhode Island	15_min	30_min	45_min	60_min	Overall accuracy
Average Accuracy	88.96	84.13	81.58	74.84	82.37

	Overall Accuracy	15min Accu	30min Accu	45min Accu	60min Accu	mean RMSE	SD RMSE	Accuracy sd
F_100 (100/50)	82.14	88.04	84.19	80.27	76	4.37	5.84	10.18
F_100 with weather	73.88	79.44	75.8	72.65	68.41	4.55	5.96	14.69

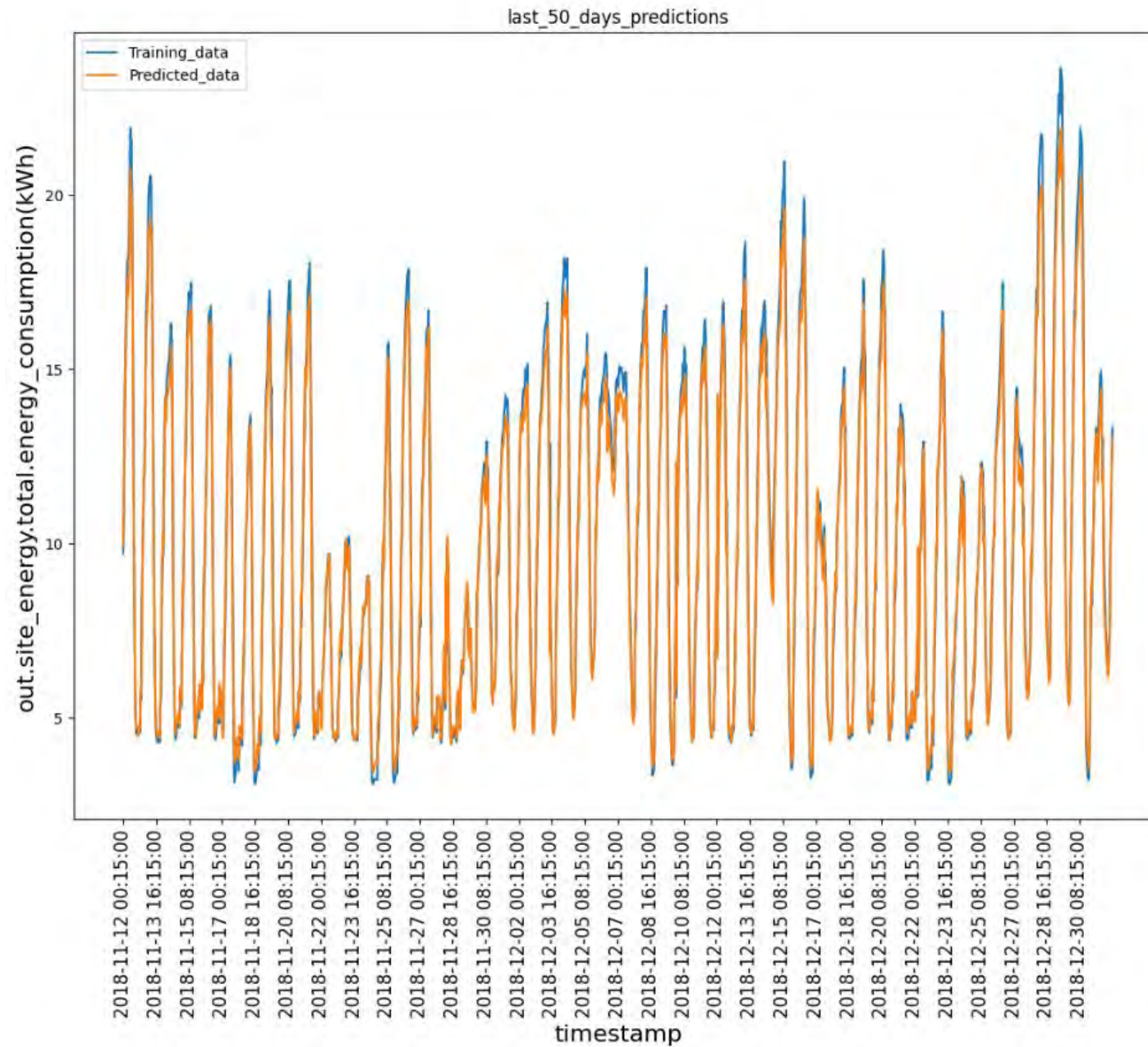
Plots result for FedAvg:



Plots result for FedAvg:



Plots result for FedAvg:



Loss Functions for LSTM Training

MSE Loss

The MSE loss used to train LSTM is given in equation 2 , where y_i is the actual value , $\hat{y}_i$ is the predicted value N is the number for forward prediction.

$$\text{MSE}(y, \hat{y}) = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}$$

As evident for equation 2 all the forward prediction $\hat{y}_1$ to $\hat{y}_N$ are equally weighed but its easier for the model to predict $\hat{y}_1$ in comparison to $\hat{y}_N$,so we introduce a exponentially weighted MSE.

Exponential Weighted MSE Loss

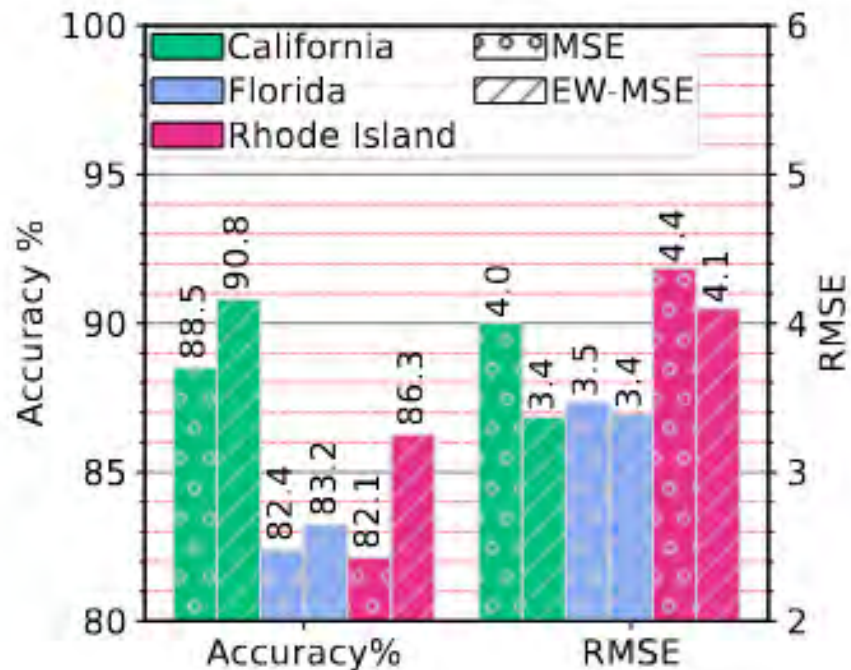
The introduced exponentially weighted MSE given in equation 3 ,where β is a hyperparameter.Here more weight is given the farther prediction ,as these predictions are harder to learn hence we penalised them more.Here value of β must be greater than 1

$$\text{EW-MSE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} \beta^i * (y_i - \hat{y}_i)^2}{N}$$

Benefits of EW-MSE Loss

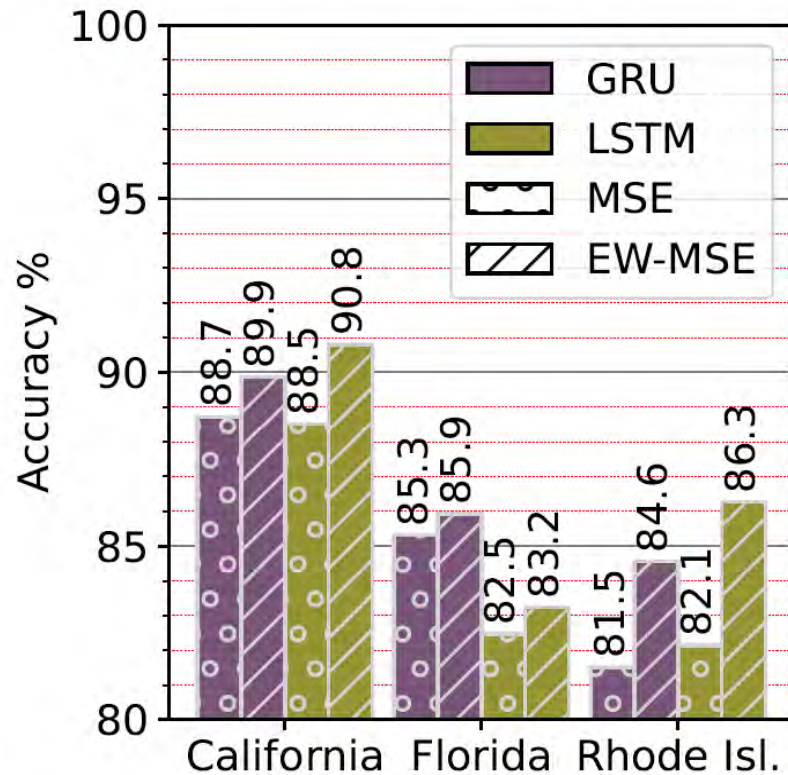
Accuracy for single FedAvg (FA) model trained using MSE loss vs. EW-MSE loss, for every 15th-min of a 1 h forward prediction with LSTM model.

	15 th min non-cumulative accuracy		30 th min non-cumulative accuracy		45 th min non-cumulative accuracy		60 th min non-cumulative accuracy	
	MSE	EW MSE	MSE	EW MSE	MSE	EW MSE	MSE	EW MSE
CA	91.9	93.6	89.6	91.9	87.4	89.9	85.1	87.9
FLO	88.5	89.4	84.4	85.7	80.4	81.3	76.0	76.7
RI	88.0	90.1	84.2	87.8	80.3	85.5	76.0	81.9

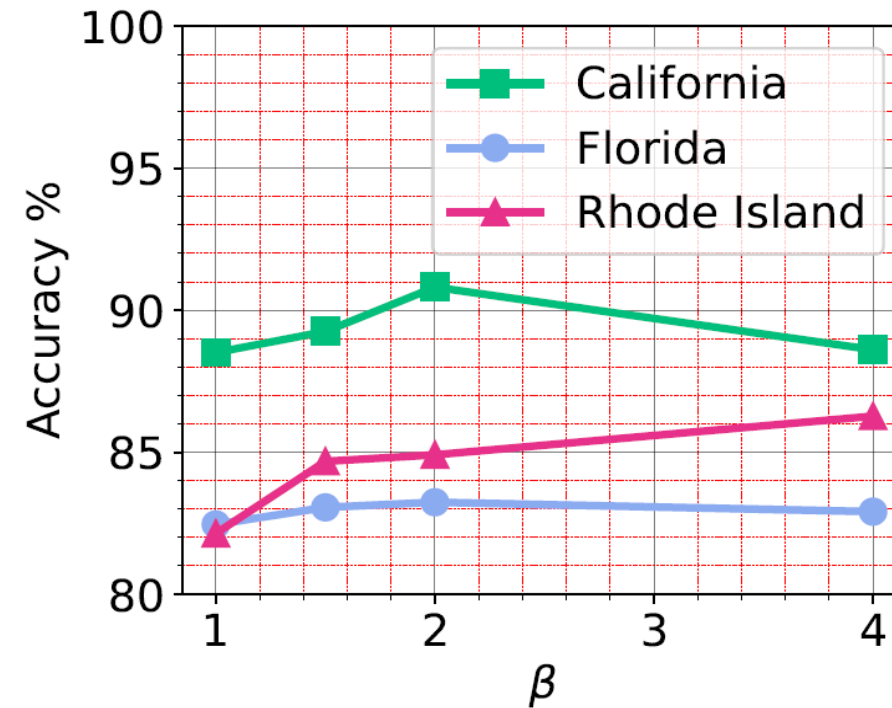


Avg. Accuracy and RMSE for models using MSE vs. EW-MSE loss with LSTM model, without clustering.

Benefits of EW-MSE Loss

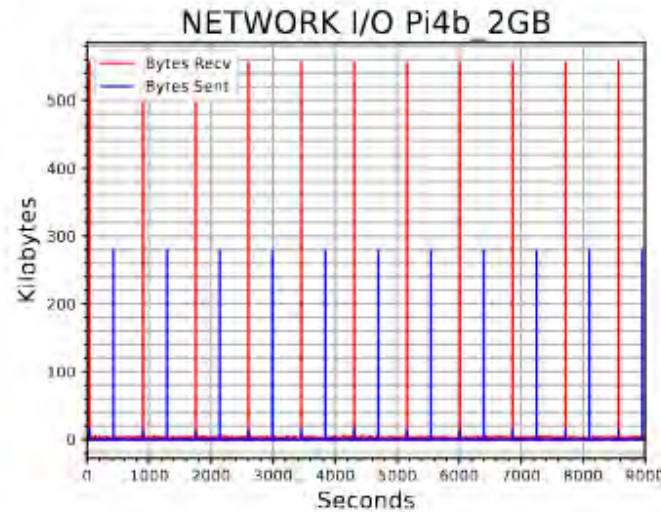
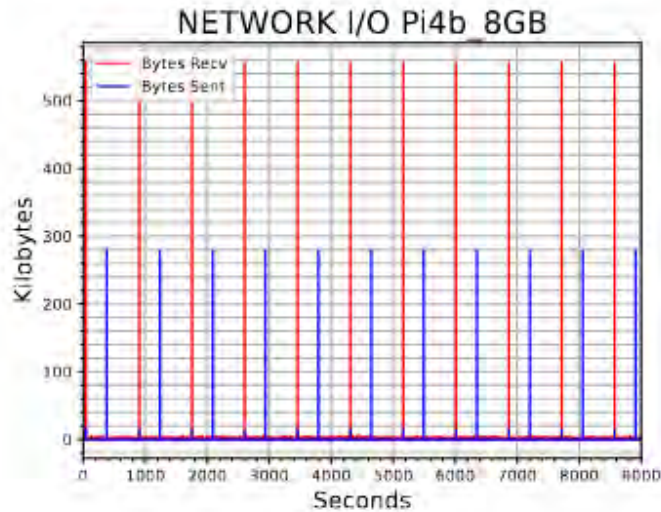
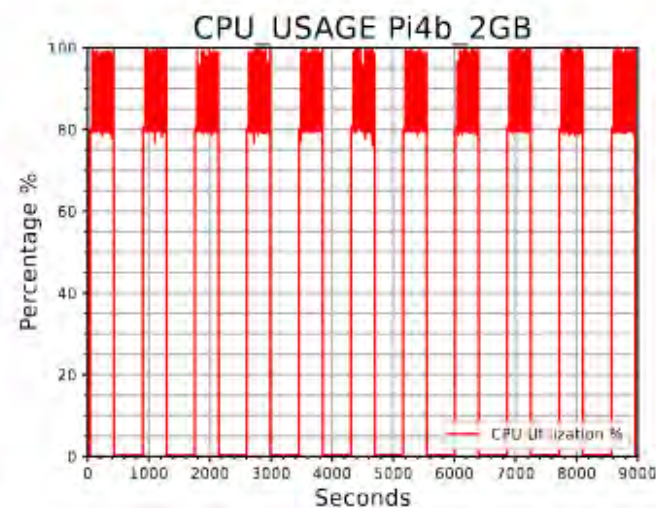
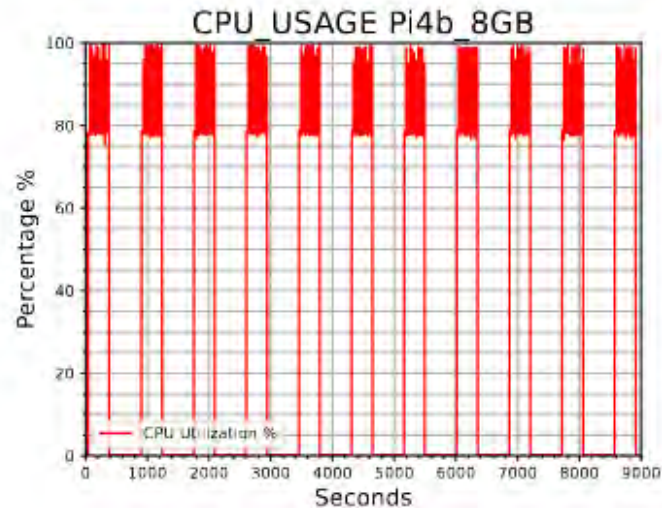


Avg. Accuracy of LSTM vs. GRU models using MSE and EWMSE losses, without clustering



Avg. Accuracy with different β in EW-MSE loss with LSTM model. Here clustering is not applied

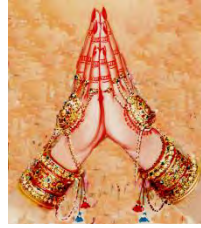
Resource Utilization on Raspberry Pis



Cluster of 50 Raspberry Pis.

Key Takeaways

- **Privacy Preservation:** FL allows for collaborative training of models on private, consumer-level data without it ever leaving the local device, thus protecting privacy.
- **Data Heterogeneity:** The clustering approach takes care of the data-heterogeneity
- **Scalability:** The methodology demonstrates that a model trained on a small, representative sample of clients can be effectively deployed to thousands of unseen clients, making the approach highly scalable and cost-effective.
- **Model Choice:** LSTMs are found to be more robust than GRUs.
- **Computational Efficiency:** The system is designed to be lightweight enough to run on resource-constrained edge devices like Raspberry Pis, as confirmed by a practical validation.
- **Clustering Impact:** Implementing K-means clustering and training separate FL models for each cluster leads to a small but consistent improvement in forecasting accuracy compared to training a single model on all clients.
- **EW-MSE Loss Function:** The **Exponentially Weighted MSE (EW-MSE)** loss function significantly improves forecasting accuracy for 60-minute horizons, with accuracy boosts of up to 4% in some cases.



Thank You

