

# Information abstraction from IoT streaming Greenhouse data

**Niketh S A**  
**(1BM16SCS11)**

Under the guidance of  
**Dr. KAYARVIZHY N**

Associate professor  
Dept. of CSE  
BMSCE

BMS College of Engineering  
Bull Temple Road, Basavanagudi, Bangalore 560019

# Abstract

- *Internet of Things* (IoT) is the **interaction** and **communication** of billions of devices that produce and exchange data, which leads to **tremendous volume of highly variable streaming data**.
- They produce **huge volume of real world streaming data**.



- To get **insight** of data and to get some **actionable information**, An effective mechanism is required.



- Aim of the work is to represent the data produced by the **greenhouse IoT devices** from device point of view to user-centric point of view using Data Analytics and find the **correlation** among them using **Latent Dirichlet Allocation(LDA)**.

# INTRODUCTION

- Growing crops in open field has several disadvantages like *pests attacks*, *extreme high/low temperature*, *affect from radiations*, *wind*, *hailstorm*.



Attack of Yellow striped armyworm (Caterpillars) and Black frass pellets  
(near the top) to the tomato fruit.

- In order to overcome these problems, Greenhouse were developed.



# Greenhouse

- Greenhouse are structures made of transparent materials where the user use to grow the plants with the required climatic conditions.



**Fig: A simple greenhouse**



**Fig: Inside tomato greenhouse**

## ➤ ADVANTAGES

- The yield can be 10-12 times more
- off-season production
- Protected from insects and diseases from them
- Less requirement of water
- In-house parameters can be controlled



# Data Abstraction

- Data abstraction is the process of reducing multiple data to simplified version without altering the meaning.
- There are many method like
  - **Signal Preprocessing** in which *Low pass filter & High pass filter* cuts the current signal with the cut-off frequency.
  - **Mathematical/Statistical Preprocessing** in which based on mean median, **peaks** of the data are removed. **Min and Max cutoff.**



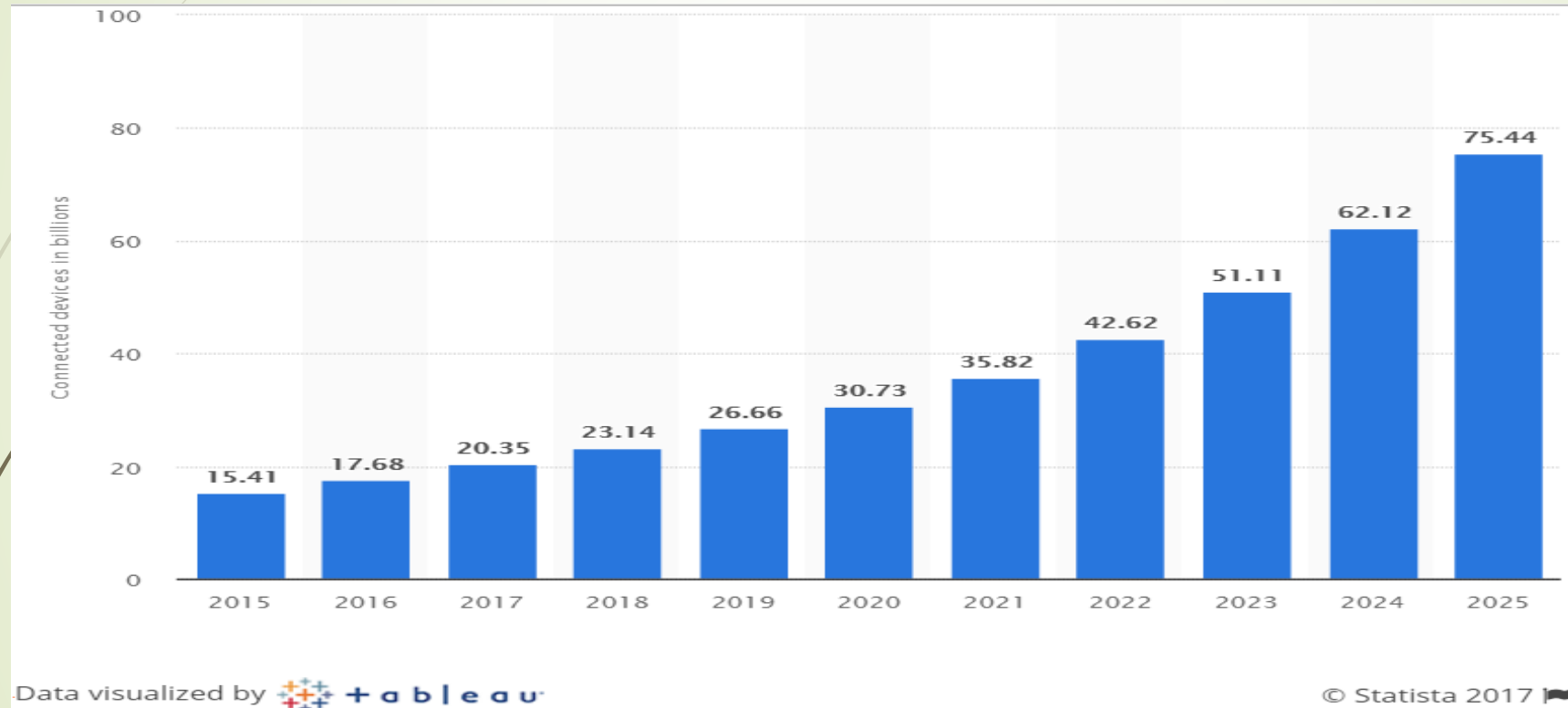
# Why do we need?

- “Duplicate and dirty data costs the healthcare industry over *\$300 billion* every year” – Joe Fusaro (Marketing Data analyst)
- “Inaccurate data has a direct impact. The average company losing 12% of its revenue” – Ben Davis (Econsultancy)
- The data generation will be **high** in future.
- Processing of Such huge data **takes time**.
- Majority of the data are **unused** later.
- The semantic data Occupies **less memory**.
- Can be used for **further processing** or **further analytics**.





➤ According to *Gartner forecast*, by **2020** nearly **21 billion** devices will be connected to internet.



**Fig :** This statistic shows the number of connected devices (Internet of Things; IoT) worldwide from 2015 to 2025. (Source: Gartner )

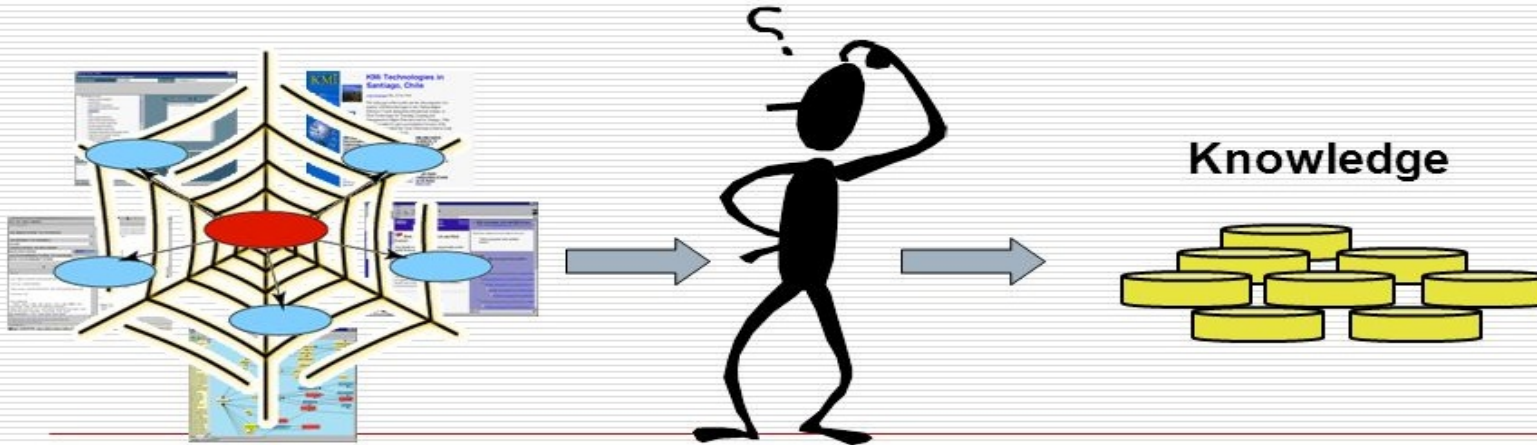
# What we do

- A huge amount of greenhouse data are produced.



The image shows a large table with multiple columns and rows of numerical data. The numbers are arranged in a grid-like pattern, typical of a data spreadsheet. The text 'wiseGEEK' is visible in the bottom right corner of the table area.

## What is it?



- **Interesting information** has to be extracted from the Raw data and **correlation** among the parameters has to be known.

# Literature Survey

| Authors  | Title  | Technology used   | Output   |
|--|--|---|--|
| Frieder Ganz,<br>Daniel Puschmann,<br>Payam Barnaghi<br>(2015) | A Practical Evaluation of<br>Information Processing<br>and Abstraction<br>Techniques for the Internet<br>of Things | Information Abstraction<br>Processes(Pre-processing,<br>Dimensional Reduction,<br>Feature Extraction) | Transformation of raw<br>sensor data to human<br>readable format |
| Daniel Puschmann et. Al<br>(2016)                              | Adaptive Clustering for<br>Dynamic IoT Data<br>Streams   | Adaptive clustering<br>method(turning point)  | Semantic data  |
| Frieder Ganz<br>(2013)   | Information Abstraction<br>for Heterogeneous<br>Real World Internet Data   | Symbolic Aggregate<br>Approximation   | Transmit of abstract data<br>to the end device                   |



# Literature Survey(Cont..)

| Authors   | Title   | Technology used  | Output   |
|---|---|--|--|
| Adnan Akbar, Francois Carrez, and Klaus Moessner et. al. (2015) | Context-Aware Stream Processing for Distributed IoT Applications                        | Micro Complex Event Processing and Adaptive Clustering   | Extract high-level knowledge from data         |
| Altti Ilari Maarala et al(2014)                                 | Semantic data provisioning and Resoning for the Internet of Things                      | Aggregating and Reasoning engine   | Delivering semantic data from IoT nodes        |
| Daniel Puschmann, Payam Barnaghi et al. (2018)                  | Using LDA to Uncover the Underlying Structures and Relations in Smart City Data Streams | Piecewise aggregate approximation<br>Symbolic Aggregate Approximation<br>Latent Dirichlet Allocation | Correlation among traffic and temperature data |

# Literature Survey(Cont..)

| Authors  | Title   | Technology used  | Output   |
|--|---|--|--|
| Daniel Puschmann, Payam Barnaghi et al. (2018) | Using LDA to Uncover the Underlying Structures and Relations in Smart City Data Streams | Piecewise aggregate approximation<br>Symbolic Aggregate Approximation<br>Latent Dirichlet Allocation | Correlation among traffic and temperature data |
| D M Blei et al. (2013)                         | Latent Dirichlet Allocation   | Latent Dirichlet Allocation  | Short representations of the discrete data     |

# Greenhouse Sensors

- Sensors **sense the change** in the environment of greenhouse and sends the information to other electronic devices like computer.
- There are 2 types
  - Analog (0-1024)
  - Digital
- Here, In order to work with the real time data, *Virtual sensor programs* are written which acts like the greenhouse Sensors based on the sensors used in monitoring greenhouse parameters.



**Table 1.1: Sensors used in monitoring parameters of the greenhouse.**

| Name         | Operating Range                 | Supply Voltage(V) | Interface | Accuracy          |
|--------------|---------------------------------|-------------------|-----------|-------------------|
| LM35         | -55°C to +155°C                 | 4 to 30 V         | Analog    | 0.5° C            |
| SHT75        | -40°C to +125°C<br>0 - 100% RH  | 2.4 to 5.5V       | Digital   | ±3° C<br>±1.8% RH |
| MQ5          | 200-10000 ppm                   | 4.9 to 5.1V       | Analog    | -                 |
| BH1750       | 1 – 65535 lx                    | 3.3 to 5 V        | Digital   | -                 |
| SHT11        | 0 - 100% RH<br>-40 °C to +125°C | 2.4 to 5.5 V      | Digital   | ±3%<br>±0.4°C     |
| TSL2561      | 0 – 40000 Lux                   | 2.7 to 3.6 V      | Digital   | -                 |
| TGS4161      | 350 to 10000 ppm                | 5V                | Analog    | ±20% ppm          |
| 18B20        | -55°C to +125°C                 | 3.0 to 5.5 V      | Digital   | ±0.5° C           |
| SEN<br>13322 | 0% - 100%                       | 5V                | Digital   | ±0.5%             |

**Sensors considered for Virtual sensor programs**

# How it is done ?



Raw data

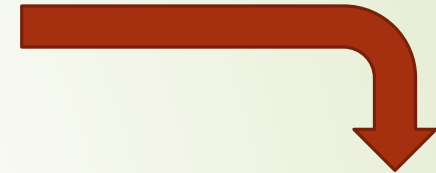


Numerical  
Data



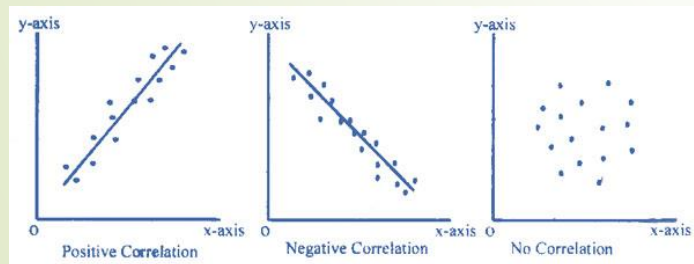
Equivalent Text  
Format

Abstraction

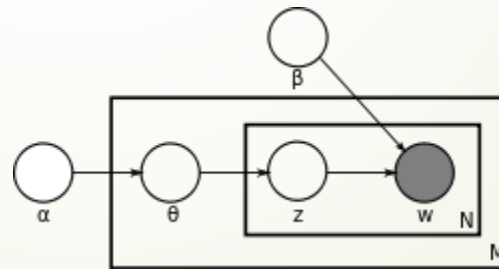


Tokenizing  
Stop Words  
Stemming

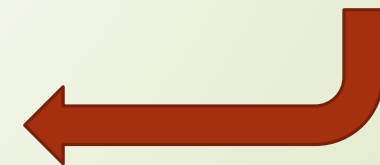
Data  
Preprocessing



Correlation



LDA

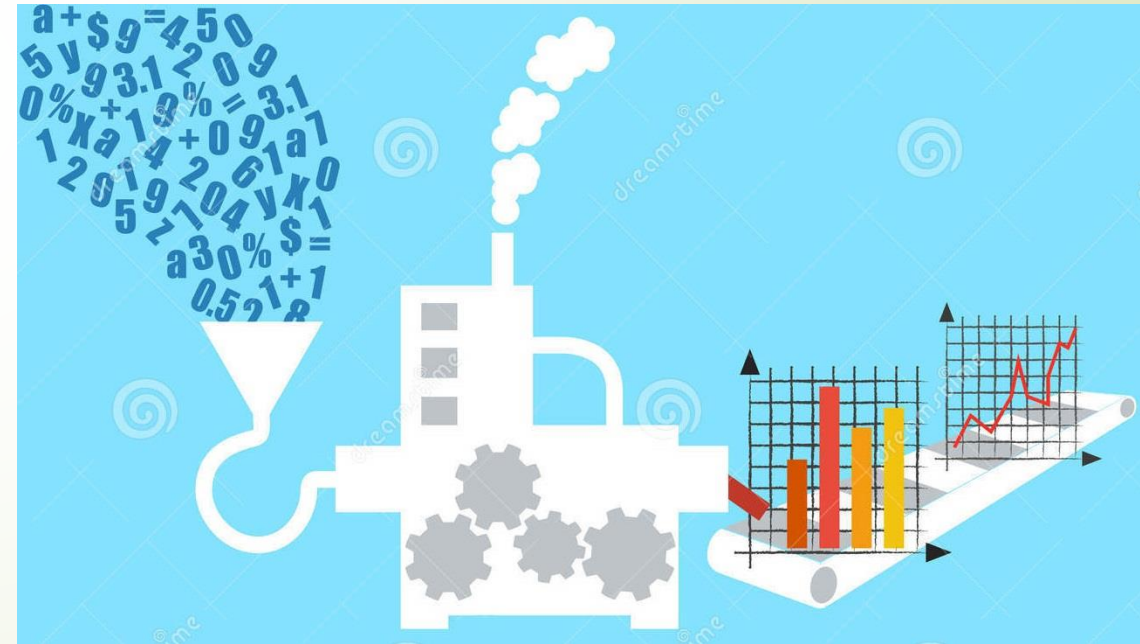


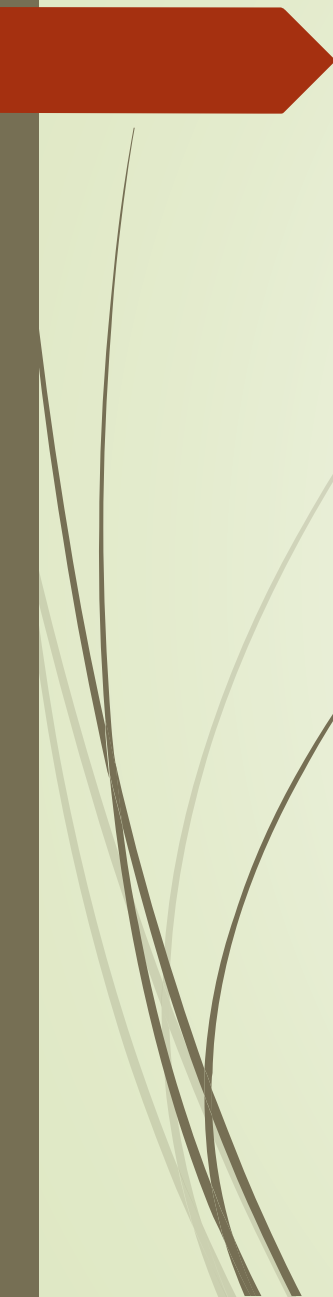




# ❖ Data Abstraction

- **Data abstraction** is the reduction of a particular body of **data** to a simplified representation of the whole. To gain more information about the data and infer knowledge.
- Taking data input within a **certain frequency**.
- Transformation of *numerical or alphabetical* digital information into a corrected, ordered, and simplified form.





| Parameters          | Low                   | Normal                      | High                   |
|---------------------|-----------------------|-----------------------------|------------------------|
| CO2                 | <700 ppm              | 700 – 1000 ppm              | >2000ppm               |
| Inside Humidity     | <40                   | 40 – 45 %                   | >45%                   |
| Outside Humidity    | <50 %                 | 50 – 55 %                   | >55 %                  |
| Luminosity          | <1000 lux             | 1000 – 2000 lux             | 2000 lux               |
| Radiation           | <977 W/m <sup>2</sup> | 977 - 1000 W/m <sup>2</sup> | <1000 W/m <sup>2</sup> |
| Soil Moisture       | <20%                  | 20 – 30 %                   | >30 %                  |
| Soil Temperature    | <20 °C                | 20 – 25 °C                  | >25 °C                 |
| Temperature Inside  | <20 °C                | 20 – 25 °C                  | >25 °C                 |
| Temperature Outside | <25 °C                | 25 – 30 °C                  | >30 °C                 |

**Fig: Conditions considered for abstraction**

# Preprocessing

- **Tokenizer:** Tokenization is the act of breaking down the set of strings into pieces. It may contain words, symbols, phrases, keywords, etc. The abstracted data is given as the input to this model. It breaks the strings to form words.

eg: 'soil moisture is normal' -> 'soil' 'moisture' 'is' 'normal'





## ➡ Stop Words:

Stop words are words such as “the”, “a”, “is”, “an”, “in” etc., and the words that appear twice which can be removed without altering the content's meaning.

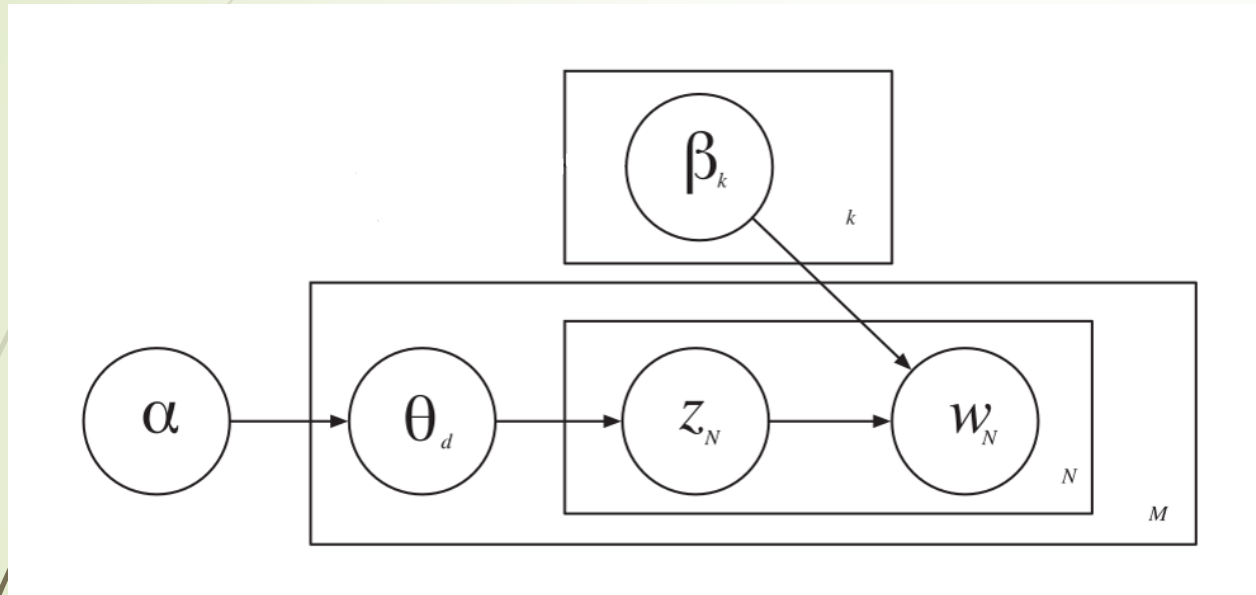
eg: “ 'soil\_moisture\_is\_normal', 'radiation\_is\_normal', 'soil\_moisture\_is\_normal'.”  
-> 'soil\_moisture\_is\_normal', 'radiation\_is\_normal',

## ➡ Stemming:

Stemming is the process of removing the words which carry the same meaning.

eg: outliers -> outlier  
-> following -> follow

# Latent Dirichlet Allocation



**Fig: Plate model Representation of LDA**

- $M$  is the total document set
- $N$  is the collection of the words represented by vector  $W_N$
- $w$  is the particular word in the document
- $z_{ij}$  is the topic that is most likely to have generated  $w_{ij}$ .
- $k$  represents the number of topics which is fixed at the initial
- $\theta$  is topic distribution
- $\alpha$  is per document topic distribution
- $\beta$  is the multinomial distribution of words which represents the topics
- $j$  is the word count and  $i$  is document count.

# Latent Dirichlet Allocation

- ▶ LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution.
  - ▶ Determine the number of words in a document. For example, let us assume the document has 6 words.
  - ▶ Determine the mixture of topics in that document. For example, the document might contain  $1/2$  the topic “health” and  $1/2$  the topic “vegetables.”
  - ▶ Using each topic’s multinomial distribution, output words to fill the document’s word slots. In our example, the “health” topic is  $1/2$  our document, or 3 words. The “health” topic might have the word “diet” at 20% probability or “exercise” at 15%, so it will fill the document word slots based on those probabilities.

## Software Requirements:

- Ubuntu Operating System
- Python3
- Libraries
  - Numpy
  - Nltk and Nltk.tokenize
  - Stop words
  - Nltk.stem.porter
  - Gensim


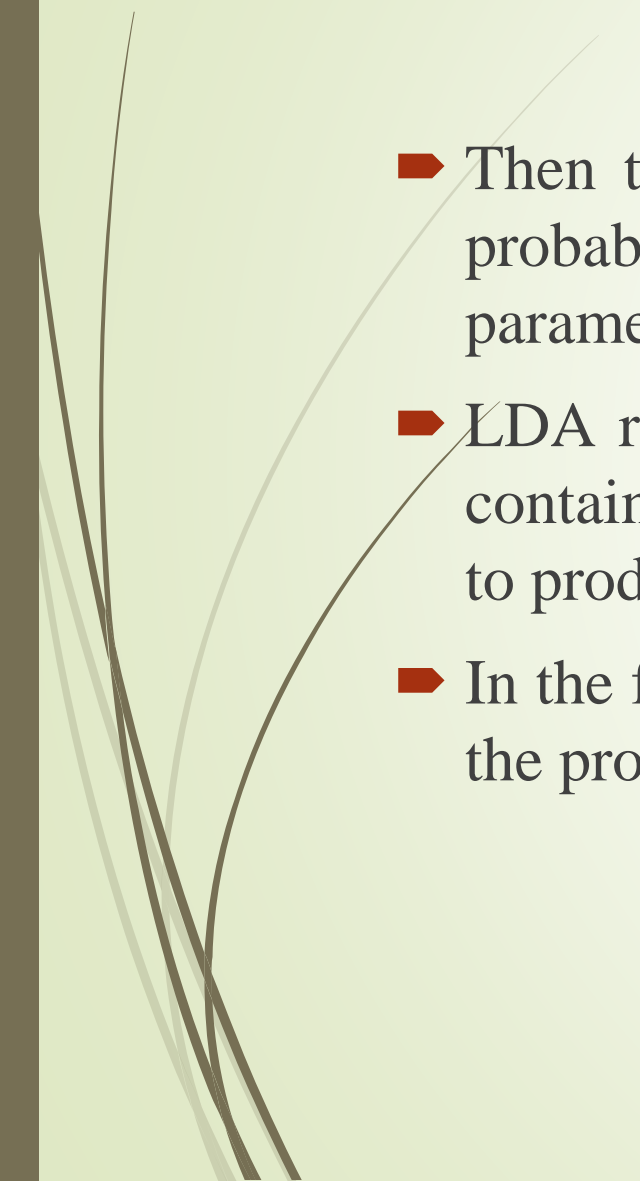
## Hardware Requirements:


- RAM 2GB or Higher
- Storage 100GB or Higher
- Intel i5 with 4CPU cores and operating frequency of CPU at 2.60GHz is used.



# Results

- Greenhouse parameters *temperature, humidity, co2 concentration, luminosity, radiation, soil moisture* and *soil temperature* are considered.
- Virtual sensor programs are written to get the real-time data.
- The data are stored in the files with respect to the sensor.
- The raw sensor values are converted to respective parameter unit according to the sensor conversion formula and checked with the respective conditions provided.

- 
- 
- Then the text file is read by the LDA module which calculates the probability of occurrences of every phrases. The correlation among the parameters of the greenhouse is found by it.
  - LDA reads 10 text files at a time in which each individual text file contains 10 samples of abstracted data. The LDA algorithm was tuned to produce 1 topics on each run with various data sets.
  - In the first test, all the parameters are kept to normal and checked with the probability.



```
[(0, '0.111*"co2_is_normal" + 0.111*"humidity_inside_is_normal"
+ 0.111*"humidity_outside_is_normal" + 0.111*"luminosity_is_normal"
+ 0.111*"radiation_is_normal" + 0.111*"soil_moisture_is_normal"
+ 0.111*"soil_temperature_is_normal" + 0.111*"temperature_inside_is_normal"
+ 0.111*"temperature_outside_is_normal"')]
```

**Fig: LDA results when all the parameters are kept normal**

- In second test, the parameters like *inside temperature*, *outside temperature*, *luminosity*, *radiation* and *co2* of some files are changed from normal to high. The result obtained are shown in figure.
- **Inside temperature** is high when the **outside temperature** is high which has got the same probability and are related. **Radiation** and **luminosity** has got the same probability and are related to each other.

```
[(0, '0.111*"humidity_inside_is_normal" + 0.111*"humidity_outside_is_normal"
+0.111*"soil_moisture_is_normal" + 0.111*"soil_temperature_is_normal"
+ 0.067*"co2_is_high" + 0.067*"luminosity_is_high" + 0.067*"radiation_is_high"
+ 0.067*"temperature_inside_is_high" + 0.067*"temperature_outside_is_high"
+ 0.045*"co2_is_normal" + 0.045*"luminosity_is_normal" + 0.045*"radiation_is_normal"
+ 0.045*"temperature_inside_is_normal" + 0.045*"temperature_outside_is_normal"')]
```

**Fig: LDA results with variable parameters1**

- In the third test, *inside humidity* and *outside humidity* are changed from *normal* to *low* in some of the files and obtained same probability which is shown in figure.

```
[(0, '0.111*"co2_is_normal" + 0.111*"luminosity_is_normal" + 0.111*"radiation_is_normal"  
+ 0.111*"soil_moisture_is_normal" + 0.111*"soil_temperature_is_normal"  
+ 0.111*"temperature_inside_is_normal" + 0.111*"temperature_outside_is_normal"  
+ 0.078*"humidity_inside_is_low" + 0.078*"humidity_outside_is_low"  
+ 0.034*"humidity_inside_is_normal" + 0.034*"humidity_outside_is_normal"')]
```

**Fig: LDA results with variable parameters2**

- In the fourth test, *inside temperature*, *outside temperature*, *co2*, *luminosity* and *radiation* are changed from *normal* to *high*. *Inside humidity* and *outside humidity* are changed from *normal* to *low* and noted that the parameters like *inside temperature*, *co2*, *radiation* and *luminosity* changes when the *outside temperature* is changed. Inside humidity is affected by the outside humidity.

```
[(0, '0.110*"soil_moisture_is_normal" + 0.110*"soil_temperature_is_normal"  
+ 0.078*"co2_is_high" + 0.078*"humidity_inside_is_low" + 0.078*"humidity_outside_is_low"  
+ 0.078*"luminosity_is_high" + 0.078*"radiation_is_high" + 0.078*"temperature_inside_is_high"  
+ 0.078*"temperature_outside_is_high" + 0.034*"co2_is_normal" + 0.034*"humidity_inside_is_normal"  
+ 0.034*"humidity_outside_is_normal" + 0.034*"luminosity_is_normal" + 0.034*"radiation_is_normal"  
+ 0.034*"temperature_inside_is_normal" + 0.034*"temperature_outside_is_normal"')]
```

**Fig: LDA results with variable parameters3**





► The following correlations were found out.

1. *Outside temperature is high* is highly correlated with *inside temperature being high*. This is also related to both *inside* and *outside humidity* as high and other factors like radiation and CO2 both being high.
2. *Humidity inside being low* is related to *humidity outside being low*, *normal luminosity* and *normal soil moisture* and *CO2* being *normal*.
3. Also, *temperature is normal*, *co2* and *humidity normal* is related to all other attributes being *normal*.

# Conclusion

- In this work, a novel approach to extract the Information hidden in the raw IoT data from Greenhouse and to find the correlation among the data.
- The process involves collecting the raw data from the sensory devices in a particular frequency and representing the equivalent text form for respective sensors data. The semantic form of the data is further preprocessed with the techniques tokenization, removal of stop words and stemming.
- Further the abstracted data is given to LDA, a topic modelling method to find the correlation among different parameters of the Greenhouse. Hence, LDA can be used to find the correlation among the greenhouse parameters

# References

- [1] Frieder Ganz, *Student Member, IEEE*, Payam Barnaghi, *Senior Member, IEEE*, and Francois Carrez. “Information Abstraction for Heterogeneous Real World Internet Data,” *IEEE SENSORS JOURNAL*, VOL. 13, NO. 10, OCTOBER 2013.
- [2] A. Akbar, F. Carrez, K. Moessner, J. Sancho, and J. Rico, “Context-aware stream processing for distributed IoT applications,” in *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*, 2015, pp. 663–668.
- [3] E. Masciari, “A framework for outlier mining in RFID data,” in *Proc. International Database Engineering and Applications Symposium*, 2007, pp. 263–267.
- [4] Frieder Ganz, Daniel Puschmann, Payam Barnaghi, “A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things”, 10.1109/JIOT.2015.2411227, *IEEE Internet of Things Journal*.

# References

- [6]Frieder Ganz, Daniel Puschmann, Payam Barnaghi, “A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things”, 2327-4662 (c) 2015 IEEE.
- [7]Sefki Kolozali , Daniel Puschmann, Athanasios Karapntelakis Et al. “Real-Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications”, GRANT AGREEMENT No 609035 FP7-SMARTCITIES-2013
- [8]Daniel Puschmann, Payam Barnaghi, “Adaptive Clustering for Dynamic IoT Data Streams”, 2327-4662 (c) 2016 IEEE.
- [9]Altti Ilari Maarala et al.Semantic Data Provisioning and Reasoning for the Internet of Things



# PO Mapping / Reflections

- **Research skills:** Based on the literature survey done, various advantages of block chain technology were known. Along with this defects of older technology for solving the problem mentioned were known.
- **Usage of modern tools:** For execution of this project, tools like solidity and mint browser are used.
- **Problem solving and critical thinking:** To identify the pain point and map it to the technology used to arrive at the desired solution.



THANK YOU