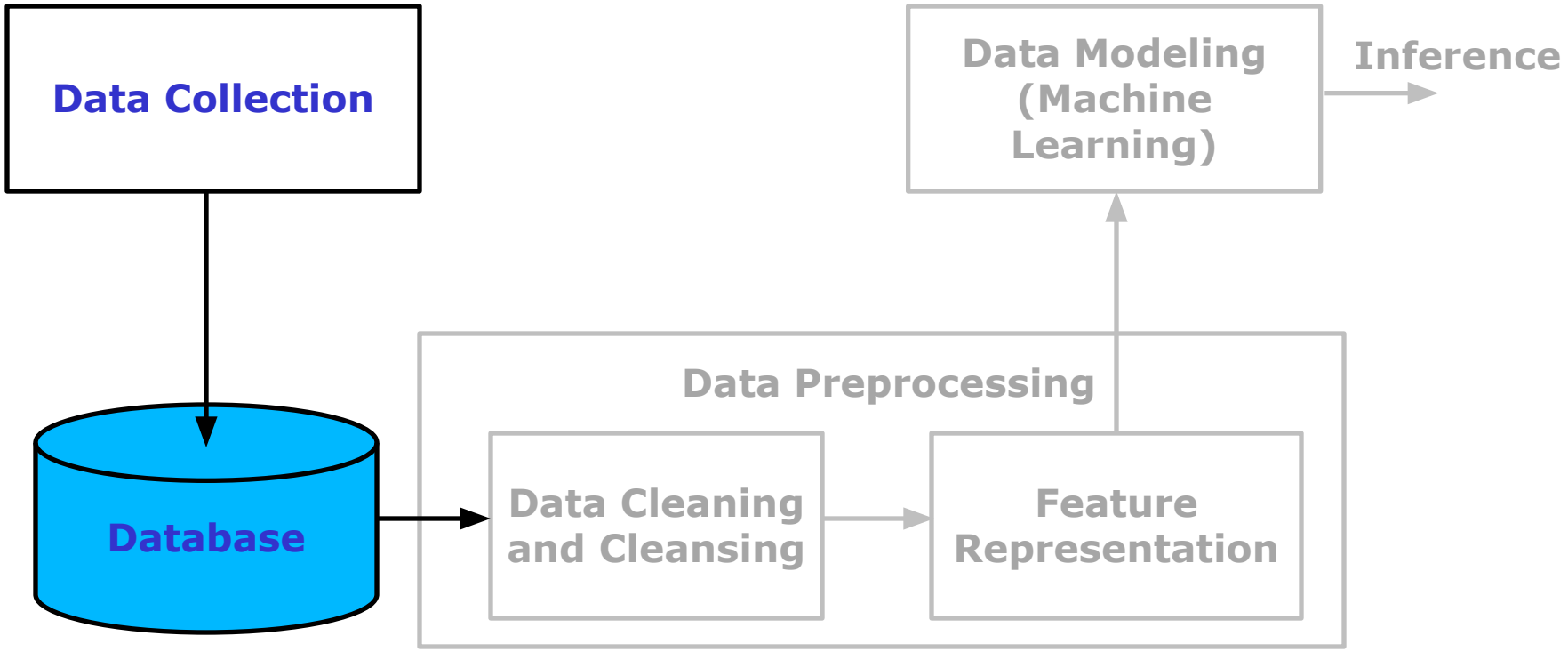
Data, Types of Data and Data Collection using Sensors

Need for Data Preprocessing

Summary of Previous Class:

- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data
- Machine learning uses data to extract knowledge



Summary of Previous Class:

Types of Data: Based on Organization

1. Unstructured data:
2. Structured data:
   - It is a tabular data (rows and columns), which are very well defined
   - Each row is finite ordered list (sequence) of elements, where each element in a column is belonging to an attribute of specific type
   - Example: Spreadsheets [Comma Separated Value (CSV) format]
3. Semi-structured data:

Summary of Previous Class:

Type of Data: Based on Variables (Value) found in Data

- Mainly in Structured Data:

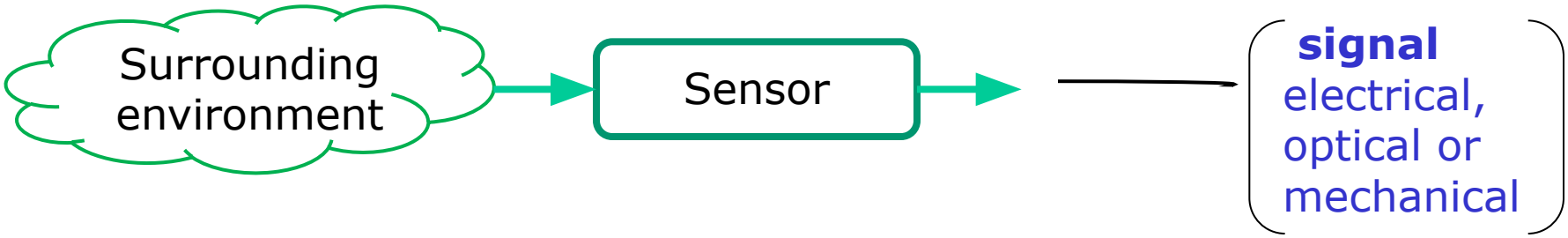
1. Numerical data:
   - Two types based on the values taken:
     - Continuous valued data:
     - Discrete valued data:
2. Categorical data:
   - Three types values they hold:
     - Ordinal values:
     - Nominal values:
     - Binary values:
3. Time series data:

Data Collection

- Data manifests itself in many different forms
- Different forms of data require different ways to collect them and different storage solutions
- Collection of data may consists of sending out surveys, polls or doing other experiments
- Data based on the way it is collected:
  - Data that comes from surveys
    - Usually textual form of data or mixed
  - Data entered in a database as system entry
    - E.g. Student information entered on academic automation system etc.
  - **Data in the form of signals (comes from sensors)**
    - Speech/Audio, Images and videos, Temperature readings, Humidity, Seismic data, EEG (all bio-type signals) etc.
- According to the objective of the task, the way the data is collected will change

Data Collection from Sensors

- Sensors are the devices that respond to the environment around it and convert the physical parameters into a signal (e.g., optical, electrical, mechanical) suitable for processing



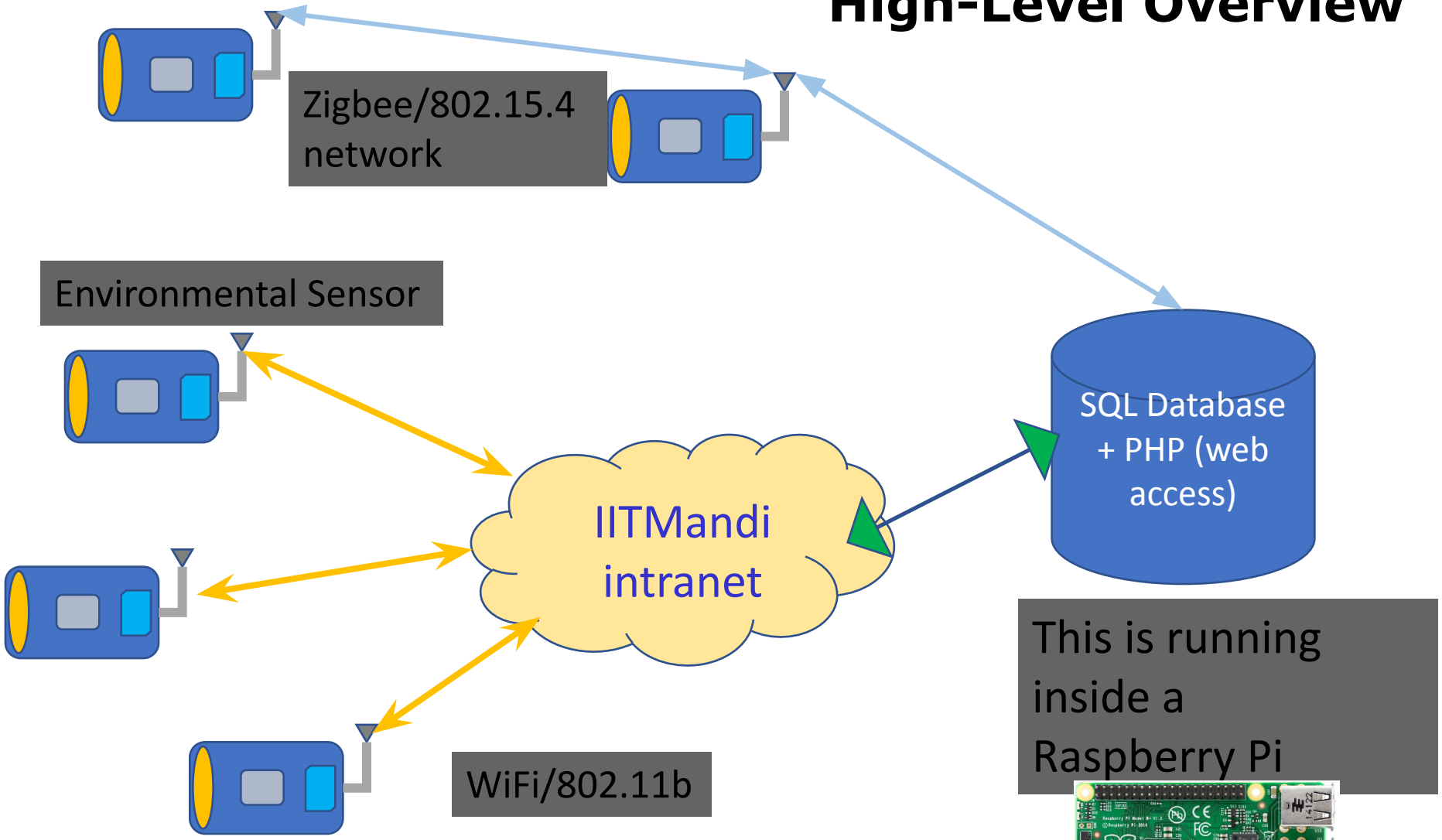
- ***Example***: a temperature sensor outputs an electrical signal whose voltage or current can be used to identify the temperature around it
- Sensors can be an electrical/mechanical component, a module or a subsystem

Different Types of Sensors

- Acoustic, sound sensors (e.g., microphone)
- Visual sensors (e.g. cameras)
- Environmental sensors (e.g., temperature, humidity, pressure etc.)
- Chemical sensors (e.g., Diesel Nitrogen Oxide (Nox) sensors to measure engine-out NOx gas concentration)
- Flow sensors (e.g., water flow sensors)
- Motion sensors (e.g., gyroscope)
- Proximity or presence sensor (e.g., Passive Infrared (PIR))
- Biosensors (e.g., glucose monitor)
- And many more …

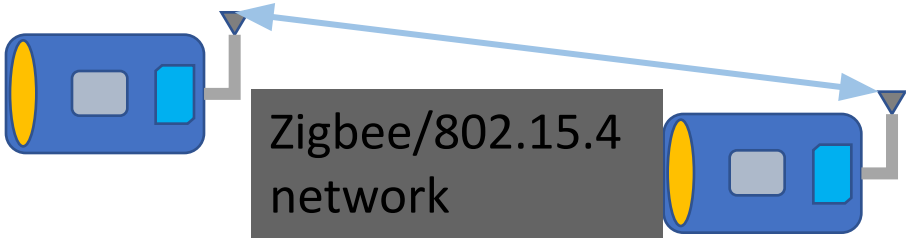
IIT Mandi Weather Station: Environmental Data (Temperature, Humidity, Pressure etc) Collection

High-Level Overview

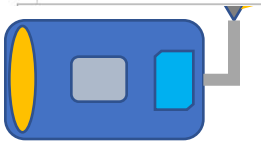


Source: Dr. Siddhartha Sarma

High-Level Overview: Environmental Data (Temperature, Humidity, Pressure etc) Collection

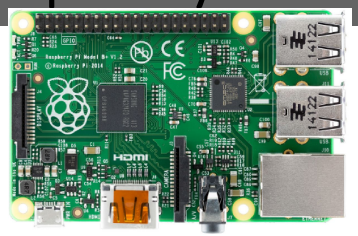


	datestamp	nodeaddr	nodePktId	nodeRSSI	nodeLQI	nodeVolt	tempVal1	tempVal2	tempVal3	humVal	presVal
2											
3	03-11-2017 07:33	fc:c2:3d:00:00:10:ab:fa	1	-53	23	3.027	16.37		16	64	905
4											
5	03-11-2017 07:33	fc:c2:3d:00:00:10:ab:35	2	-84	24	2.905	17.62	17.9794	17	63	904
6											
7	03-11-2017 07:38	fc:c2:3d:00:00:10:ab:fa	3	-54	18	3.027	16.62		16	64	905
8											
9	03-11-2017 07:38	fc:c2:3d:00:00:10:ab:35	4	-84	20	2.905	17.62	17.9794	17	63	904
10											
11	03-11-2017 07:43	fc:c2:3d:00:00:10:ab:fa	5	-50	27	3.027	16.37		16	64	905
12											
13	03-11-2017 07:43	fc:c2:3d:00:00:10:ab:35	6	-86	15	2.905	17.62	18.0789	17	63	904
14											
15	03-11-2017 07:48	fc:c2:3d:00:00:10:ab:fa	7	-52	22	3.027	16.25		16	65	905
16											



WiFi/802.11b

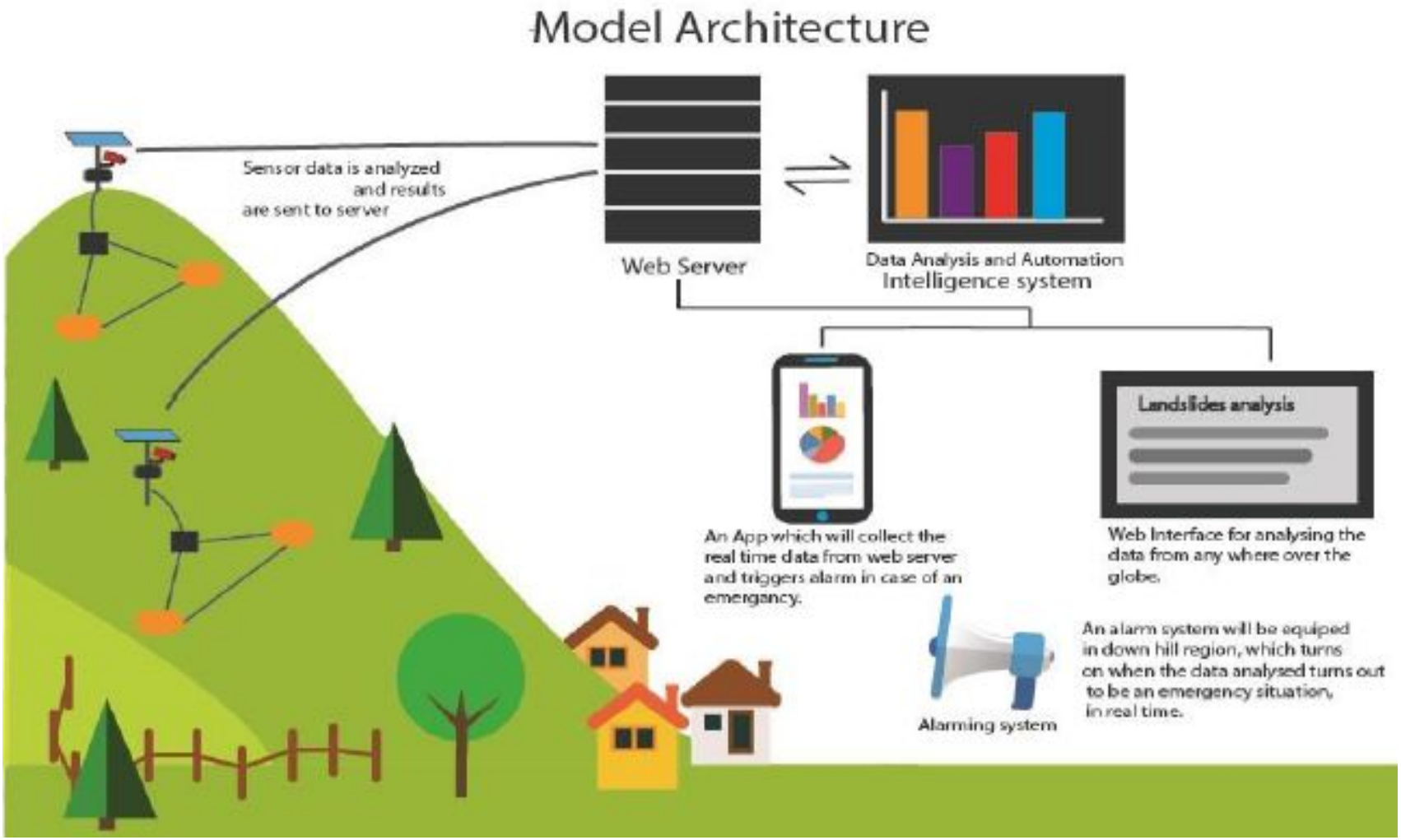
Raspberry Pi



Source: Dr. Siddhartha Sarma

Land Slide Monitoring System (LMS)

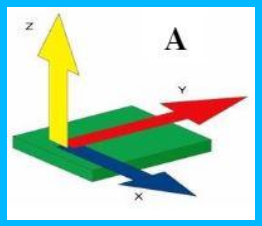
- LMSs that rely on Internet of Things (IoT) and low-cost Micro-Electro-Mechanical Systems (MEMS) sensors



Source: Dr. Varun Dutt (SCEE) and Dr. Uday Kala (SMME)

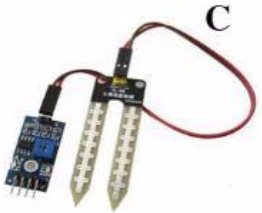
Components of LMS

- The LMS monitors a number of weather and soil parameters via sensors on deployment location



GY 61 Accelerometer Sensor

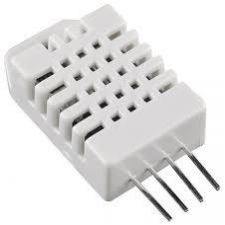
Pin Diagram of GY-61



YL 69 Soil Moisture Sensor

SIM 900A GSM Module

E
Force Sensor



F
Humidity Sensor DHT 22

G
Light Sensor BH-1750

H
Temperature and Pressure Sensor BMP-180

I
Tipping Rain Gauge

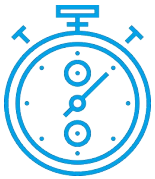
Source: Dr. Varun Dutt (SCEE) and Dr. Uday Kala (SMME)

Architecture and Features of LMS

- The LMS monitors a number of weather and soil parameters via sensors on deployment location

Temperature & Humidity

(-40 C to +80 C & 0-100 %)



Barometric Pressure

(300-1100 mb)

Rainfall Intensity

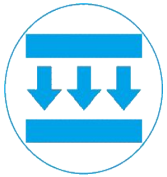
(in mm)

Light Intensity

(0 - 65535 Lux)

Soil movement

(±2000°/sec rotational & ±16g gravitational acceleration)



Soil force

(0-100N)

Soil moisture

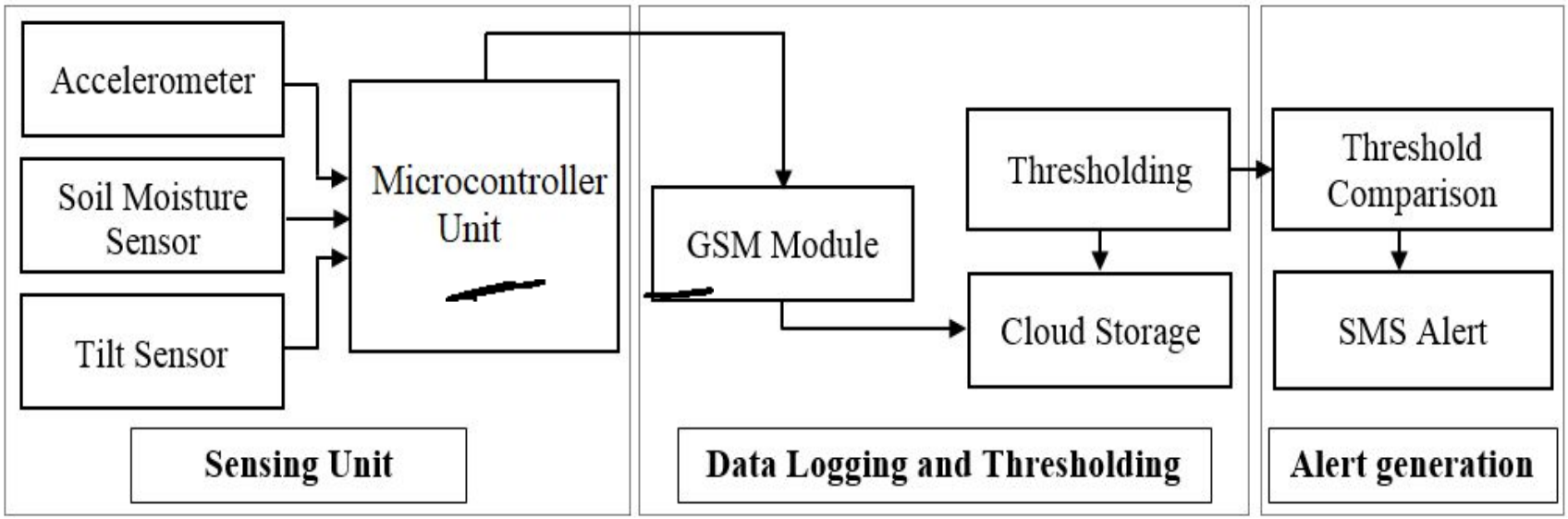
(0-100 %)

Source: Dr. Varun Dutt (SCEE) and Dr. Uday Kala (SMME)

Architecture and Features of LMS

- The LMS monitors a number of weather and soil parameters via sensors on deployment location

Architecture diagram of LMS



The LMS will alert people via traffic lights, SMSs, or smart-apps on mobile phones about the danger of impending landslides

Source: Dr. Varun Dutt (SCEE) and Dr. Uday Kala (SMME)

Architecture and Features of LMS

- The LMS monitors a number of weather and soil parameters via sensors on deployment location

Date/ Time	Temperature (C)/ Humidity (%)	Pressure (Pa)	Rain (inches)	Light intensity (lux)	Accelerations (g)	Force (N)	Moisture (%)
2017-09-06 18:44:32	23.00,56.00	617.64	0.01	3	0.52,0.31,-0.80,0.00,0.00,0.00,31.36,-159.01	0.02	81.00
2017-09-06 18:33:32	24.00,58.00	619.47	0.01	12	0.52,0.30,-0.79,0.00,0.00,0.00,31.45,-159.12	0.02	82.00
2017-09-06 18:22:39	24.00,58.00	623.37	0.00	71	0.52,0.31,-0.80,0.00,0.00,0.00,31.35,-158.88	0.02	83.00
2017-09-06 18:11:31	25.00,60.00	627.02	0.05	194	0.51,0.31,-0.80,0.00,0.00,0.00,30.80,-159.00	0.02	81.00

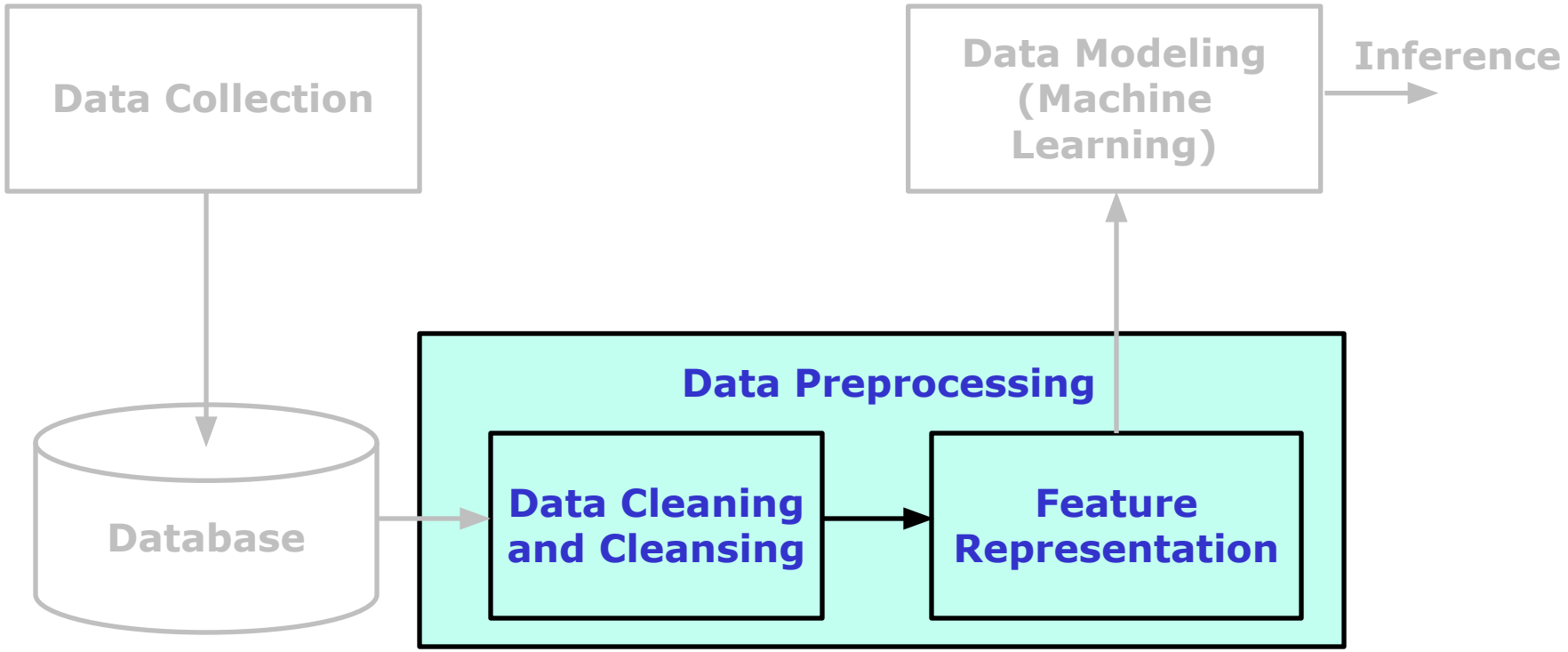
The LMS will alert people via traffic lights, SMSs, or smart-apps on mobile phones about the danger of impending landslides

Source: Dr. Varun Dutt (SCEE) and Dr. Uday Kala (SMME)

Data Preprocessing

Data Science

- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data
- Machine learning uses data to extract knowledge



Need for Data Preprocessing

- Real world data are tend to be incomplete, noisy and inconsistent due to their huge size and their likely origin from multiple heterogeneous sources
- Preprocessing is important to clean the data
- Low quality data will lead to low quality of analysis results
- If the users believe the data is of low quality (dirty), they are unlikely to trust the results of any data analytics that has been applied to
- Low quality data can cause confusion for analytic procedure using machine learning techniques, resulting in unreliable output
- Incomplete, noisy and inconsistent data are common properties of large real world databases

Tuple (Record) in Structured Data

- A tuple (record) is finite ordered list (sequence) of elements, where each element is belonging to an attribute

Date/ Time	Temperature (C)/ Humidity (%)	Pressure (Pa)	Rain (inches)	Light intensity (lux)	Accelerations (g)	Force (N)	Moisture (%)
2017-09-06 18:44:32	23.00,56.00	617.64	0.01	3	0.52,0.31,-0.80,0.00,0.00,0.00,31.36,-159.01	0.02	81.00
2017-09-06 18:33:32	24.00,58.00	619.47	0.01	12	0.52,0.30,-0.79,0.00,0.00,0.00,31.45,-159.12	0.02	82.00
2017-09-06 18:22:39	24.00,58.00	623.37	0.00	71	0.52,0.31,-0.80,0.00,0.00,0.00,31.35,-158.88	0.02	83.00
2017-09-06 18:11:31	25.00,60.00	627.02	0.05	194	0.51,0.31,-0.80,0.00,0.00,0.00,30.80,-159.00	0.02	81.00

Tuple (record)

- Each row is a tuple

Incomplete Data

- Many tuple (records) have no recorded value for several attributes
- Example:

	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018	—	—	83.14912	—
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	24.29851	87.68657	963
6	08-07-2018	t11			
7	09-07-2018	t11	26.8494	61.10241	15
8	10-07-2018	t11	27.88806	75.07463	13583.25
9	11-07-2018	t11	27.35915	76.02113	19768.5
10	23-07-2018	t12	24.39024	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.75
12	25-07-2018				
13	26-07-2018	t12	22.19718	99	864

Incomplete Data

- Many tuple (records) have no recorded value for several attributes
- Reasons for incomplete data:
  - User forgot to fill in a field
  - User chose not to fill out the field as it was not considered important at the time of the entry
  - Relevant data may not be recorded due to malfunctioning of equipment
  - Data might have lost while transferring from recorded place
  - Data may not be recorded due to programming error
  - Data might not be recorded due to technology limitations like limited memory

Noisy Data

- Many tuple (records) have incorrect value for several attributes
- Reasons for noisy data:
  - There may be human or computer error occurring in data entry
  - The data collection instruments used may be faulty
  - Error in data transmission
  - There may be technology limitation such as limited buffer size for coordinating synchronised data transfer and consumption

Inconsistent Data

- Data containing discrepancies in stored values for some attributes
- Reasons for inconsistent data:
  - It may result from inconsistencies in
    - name conventions or
      - Example: "Dept_ID", "Department_ID"
        "Roll_No", "Registation_No"
    - data codes used (mismatch in writing values) or
      - Example: For department – "SCEE", "School of Computing and EE"
    - inconsistent formats of input fields such as date
      - Example: "dd-mm-yy", "dd-mm-yyyy", "mm/dd/yyyy"
  - Inconsistency in name convention or formats of input fields while integrating
    - Example: While Integrating temperature records from different locations, if the name conventions are different
  - Inconsistent data may be due to human or computer error occurring in data entry