IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

Data Handling and Analytics – Part II

Data is Precious

Dr. Sudip Misra
Associate Professor
Department of Computer Science and Engineering
IIT KHARAGPUR
Email: smisra@sit.iitkgp.ernet.in
Website: http://cse.iitkgp.ac.in/~smisra/

What is Data Analytics

- ✓ *"Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses."*

  [An admin's guide to AWS data management]

Types of Data Analysis

- ✓ Two types of analysis
  - ✓ Qualitative Analysis
    - ✓ Deals with the analysis of data that is categorical in nature
  - ✓ Quantitative Analysis
    - ✓ Quantitative analysis refers to the process by which numerical data is analyzed

Qualitative Analysis

- ✓ Data is not described through numerical values
- ✓ Described by some sort of descriptive context such as text
- ✓ Data can be gathered by many methods such as interviews, videos and audio recordings, field notes
- ✓ Data needs to be interpreted
- ✓ The grouping of data into identifiable themes
- ✓ Qualitative analysis can be summarized by three basic principles (Seidel, 1998):
  - ✓ Notice things
  - ✓ Collect things
  - ✓ Think about things

Quantitative Analysis

- ✓ Quantitative analysis refers to the process by which numerical data is analyzed
- ✓ Involves descriptive statistics such as mean, media, standard deviation
- ✓ The following are often involved with quantitative analysis:

- ✓ Statistical models
- ✓ Analysis of variables
- ✓ Data dispersion
- ✓ Analysis of relationships between variables
- ✓ Contingence and correlation
- ✓ Regression analysis
- ✓ Statistical significance
- ✓ Precision
- ✓ Error limits

Comparison

Qualitative Data	Quantitative Data
Data is observed	Data is measured
Involves descriptions	Involves numbers
Emphasis is on quality	Emphasis is on quantity
Examples are color, smell, taste, etc.	Examples are volume, weight, etc.

Advantages

- ✓ Allows for the identification of important (and often mission-critical) trends
- ✓ Helps businesses identify performance problems that require some sort of action
- ✓ Can be viewed in a visual manner, which leads to faster and better decisions
- ✓ Better awareness regarding the habits of potential customers
- ✓ It can provide a company with an edge over their competitors

Statistical models

- ✓ The **statistical model** is defined as the mathematical equation that are formulated in the form of relationships between variables.
- ✓ A statistical model illustrates how a set of random variables is related to another set of random variables.
- ✓ A statistical model is represented as the ordered pair (X , P)
  - ✓ X denotes the set of all possible observations
  - ✓ P refers to the set of probability distributions on X

Statistical models (Contd.)

- ✓ Statistical models are broadly categorized as
  - ✓ Complete models
  - ✓ Incomplete models

- ✓ Complete model does have the number of variables equal to the number of equations
- ✓ An incomplete model does not have the same number of variables as the number of equations

Statistical models (Contd.)

- ✓ In order to build a statistical model
  - ✓ Data Gathering
  - ✓ Descriptive Methods
  - ✓ Thinking about Predictors
  - ✓ Building of model
  - ✓ Interpreting the Results

Analysis of variance

- ✓ Analysis of Variance (ANOVA) is a parametric statistical technique used to compare datasets.
- ✓ ANOVA is best applied where more than 2 populations or samples are meant to be compared.
- ✓ To perform an ANOVA, we must have a continuous response variable and at least one categorical factor (e.g. age, gender) with two or more levels (e.g. Locations 1, 2)
- ✓ ANOVAs require data from approximately normally distributed populations

Analysis of variance (Contd.)

- ✓ Properties to perform ANOVA –
  - ✓ Independence of case
    - ✓ The sample should be selected randomly
    - ✓ There should not be any pattern in the selection of the sample
  - ✓ Normality
    - ✓ Distribution of each group should be normal
  - ✓ Homogeneity
    - ✓ Variance between the groups should be the same (e.g. should not compare data from cities with those from slums)

Analysis of variance (Contd.)

- ✓ Analysis of variance (ANOVA) has three types:
- ✓ One way analysis
  - ✓ One fixed factor (levels set by investigator). Factors: age, gender, etc.
- ✓ Two way analysis
  - ✓ Factor variables are more than two
- ✓ K-way analysis
  - ✓ Factor variables are k

Analysis of variance (Contd.)

- ✓ **Total Sum of square**
  - ✓ In statistical data analysis, the total sum of squares (TSS or SST) is a quantity that appears as part of a standard way of presenting results of such analyses. It is defined as being the sum, over all observations, of the squared differences of each observation from the overall mean.
- ✓ **F –ratio**
  - ✓ Helps to understand the ratio of variance between two data sets
  - ✓ The F ratio is approximately 1.0 when the null hypothesis is true and is greater than 1.0 when the null hypothesis is false.
- ✓ **Degree of freedom**
  - ✓ Factors which have no effect on the variance
  - ✓ The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

Data dispersion

- ✓ A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data becomes more diverse.

- ✓ Examples of dispersion measures:
  - ✓ Range
  - ✓ Average absolute deviation
  - ✓ Variance and Standard deviation

Data dispersion (Contd.)

- ✓ **Range**
  - ✓ The range is calculated by simply taking the difference between the maximum and minimum values in the data set.
- ✓ **Average absolute deviation**
  - ✓ The average absolute deviation (or mean absolute deviation) of a data set is the average of the absolute deviations from the mean.
- ✓ **Variance**
  - ✓ Variance is the expectation of the squared deviation of a random variable from its mean
- ✓ **Standard deviation**
  - ✓ Standard deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values

Contingence and correlation

- ✓ In statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables.
- ✓ Provides a basic picture of the interrelation between two variables
- ✓ A crucial problem of multivariate statistics is finding (direct-)dependence structure underlying the variables contained in high-dimensional contingency tables

Contingence and correlation (Contd.)

- ✓ Correlation is a technique for investigating the relationship between two quantitative, continuous variables
- ✓ Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.
- ✓ Correlations are useful because they can indicate a predictive relationship that can be exploited in practice

Regression analysis

- ✓ In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables
- ✓ Focuses on the relationship between a dependent variable and one or more independent variables
- ✓ Regression analysis estimates the conditional expectation of the dependent variable given the independent variables

Regression analysis (Contd.)

- ✓ The estimation target is a function of the independent variables called the regression function
- ✓ Characterize the variation of the dependent variable around the regression function which can be described by a probability distribution
- ✓ Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning
- ✓ Regression analysis is also used to understand which among the independent variables are related to the dependent variable

Statistical significance

- ✓ Statistical significance is the likelihood that the difference in conversion rates between a given variation and the baseline is not due to random chance
- ✓ Statistical significance level reflects the risk tolerance and confidence level
- ✓ There are two key variables that go into determining statistical significance:
  - ✓ Sample size
  - ✓ Effect size

Statistical significance (Contd.)

- ✓ Sample size refers to the sample size of the experiment
- ✓ The larger your sample size, the more confident you can be in the result of the experiment (assuming that it is a randomized sample)
- ✓ The effect size is just the standardized mean difference between the two groups
- ✓ If a particular experiment replicated, the different effect size estimates from each study can easily be combined to give an overall best estimate of the effect size

Precision and Error limits

- ✓ Precision refers to how close estimates from different samples are to each other
- ✓ The standard error is a measure of precision
- ✓ When the standard error is small, estimates from different samples will be close in value and vice versa
- ✓ Precision is inversely related to standard error

Precision and Error limits (Contd.)

- ✓ The limits of error are the maximum overestimate and the maximum underestimate from the combination of the sampling and the non-sampling errors

- ✓ The margin of error is defined as –
  - ✓ Limit of error = Critical value x Standard deviation of the statistic
  - ✓ Critical value: Determines the tolerance level of error.

References

- ✓ Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. (1995). Fast discovery of association rules, Advances in Knowledge Discovery and Data Mining , AAAI/MIT Press, Cambridge, MA.
- ✓ Agresti, A. (1996). An Introduction to Categorical Data Analysis , Wiley, New York.
- ✓ Agresti, A. (2002). Categorical Data Analysis (2nd Ed.), Wiley, New York
- ✓ Anderson, T. (2003). An Introduction to Multivariate Statistical Analysis, 3rd ed., Wiley, New York.
- ✓ Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components, Journal of the American Statistical Association, 101: 119–137.
- ✓ Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoid function, IEEE Transactions on Information Theory, 39: 930–945.
- ✓ Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society Series B. 85: 289–300.
- ✓ Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion), Journal of the Royal Statistical Society, Series B, Methodo logical, 45: 311–354.

Thank You!!

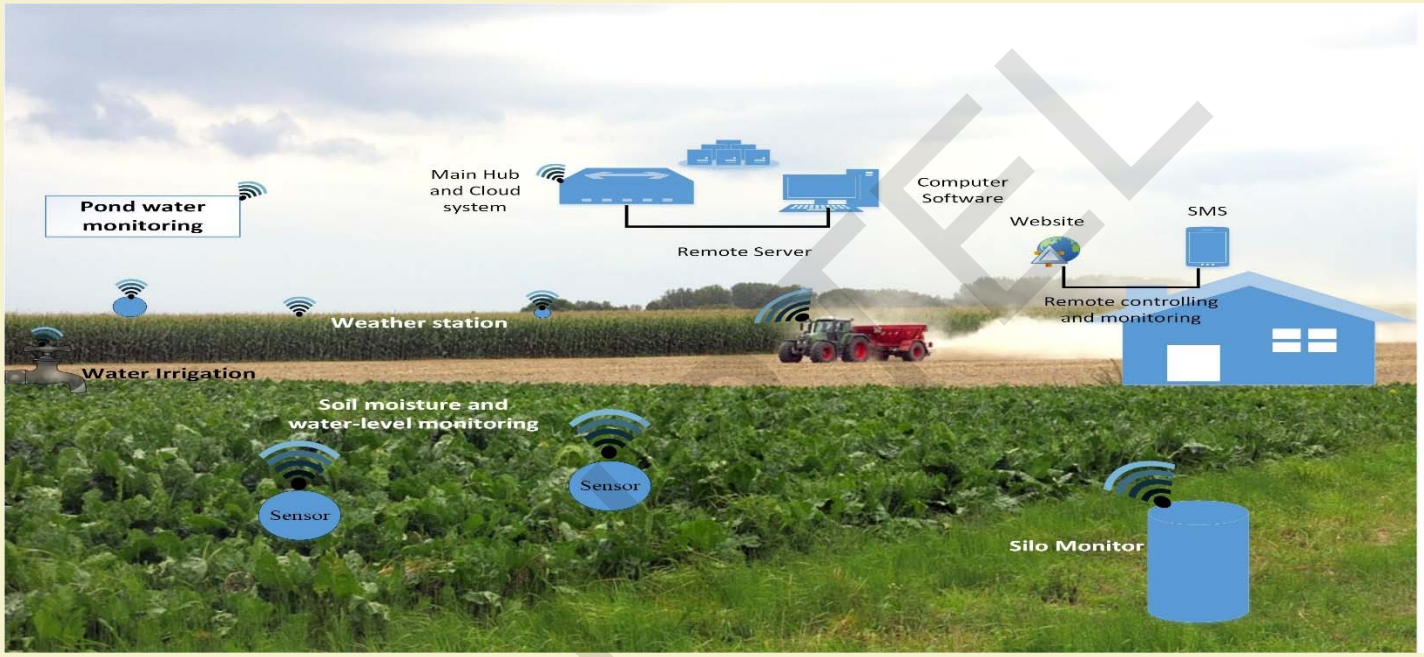
IIT KHARAGPUR

NPTEL ONLINE
CERTIFICATION COURSES

Case Study: Agriculture

Dr. Sudip Misra
Associate Professor
Department of Computer Science and Engineering
IIT KHARAGPUR
Email: smisra@sit.iitkgp.ernet.in
Website: http://www.cse.iitkgp.ac.in/~smisra/

Future of IoT application in agriculture



- ✓ Soil moisture and water level monitoring
- ✓ Automated irrigation system
- ✓ Automation in Recycling of Organic Waste and Vermicomposting
- ✓ Automated sowing and weeding system

Image template source: https://pixabay.com/p-747175/?no_redirect

Case study on
Smart Water Management Using IoT

AgriSens: Smart Water Management using IoT

- ✓ Objectives
  - More yields with less water
  - Save limited water resource in a country
  - Automatic irrigation
  - Dynamic irrigation treatments in the different phases of a crop's life cycle
  - Remote monitoring and controlling

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

- ✓ Proposed architecture
  - Sensing and actuating layer
  - Processing, storage, and service layer
  - Application layer

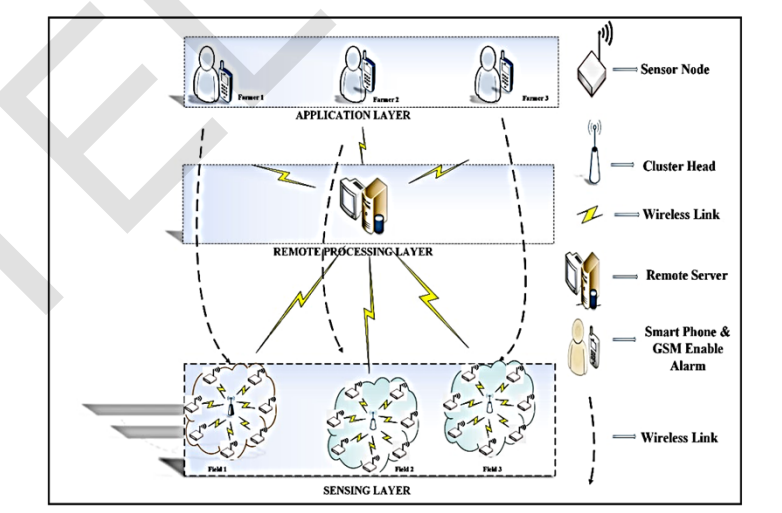


Fig 1: The proposed architecture of AgriSens

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

- ✓ Design
  - Integrated design for sensors
  - Integrated design for sensor node
  - Integrated design for remote server

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

✓ Integrated design for sensors

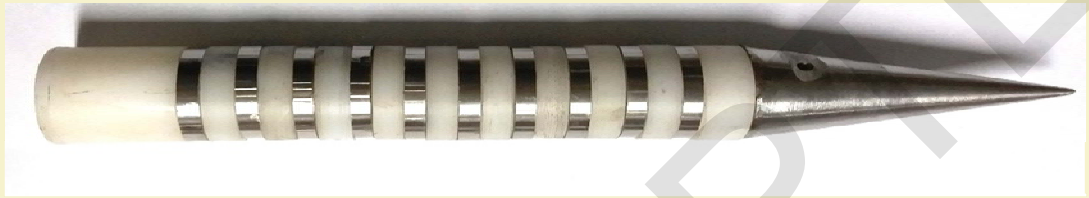


Fig 4: Designed water-level sensor

Fig 5: EC-05 soil moisture sensor

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

✓ Integrated design for sensor node

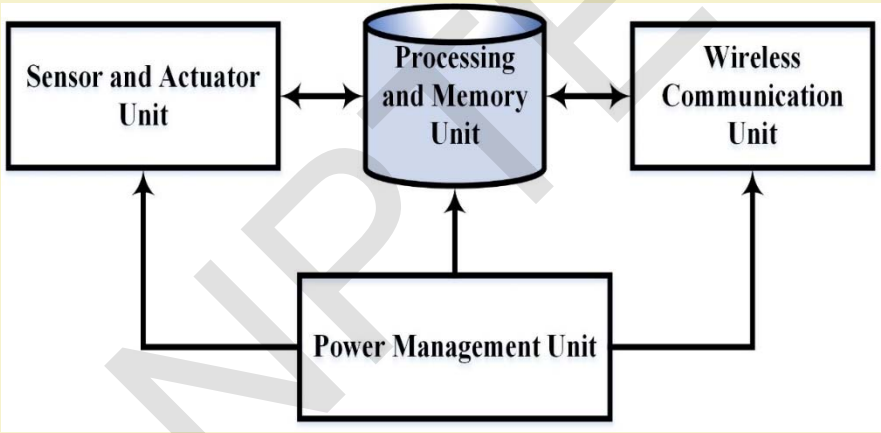


Fig 2: The block diagram of a sensor node

AgriSens: Smart Water Management using IoT (Contd.)

✓ Integrated design for sensor node

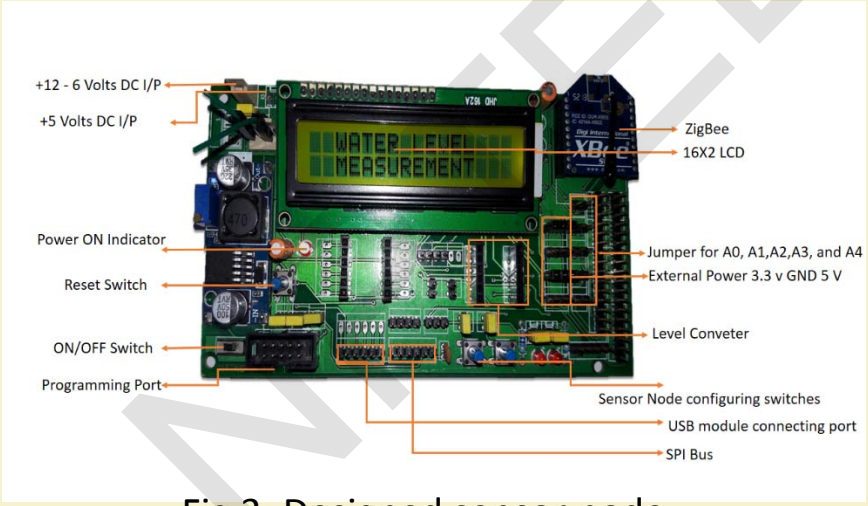


Fig 3: Designed sensor node

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

- ✓ Integrated design for remote server
  - Repository data server: Communicates with the deployed IoT gateway in the field by using GPRS technology
  - Web server: To access field data remotely
  - Multi users server: Sends field information to farmer's cell using SMS technology and also executes farmer's query and controlling messages

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

- ✓ Implementation
  - Field demo
  - Website demo
  - Project details from website

AgriSens: Smart Water Management using IoT (Contd.)

✓ Results

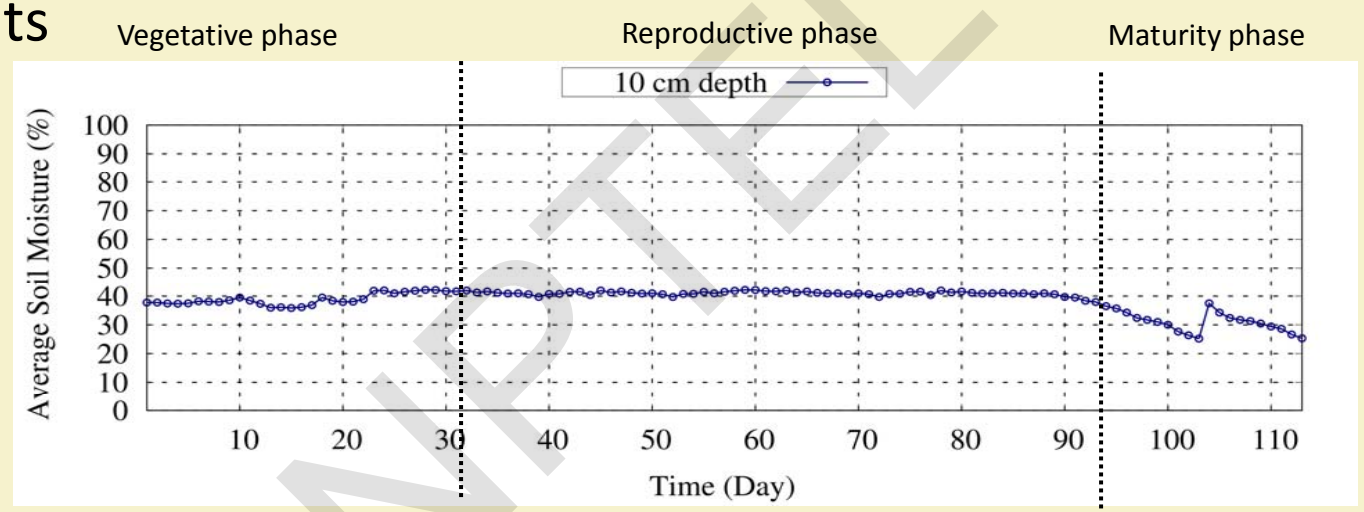


Fig. 6: Average soil moisture

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

✓ Results

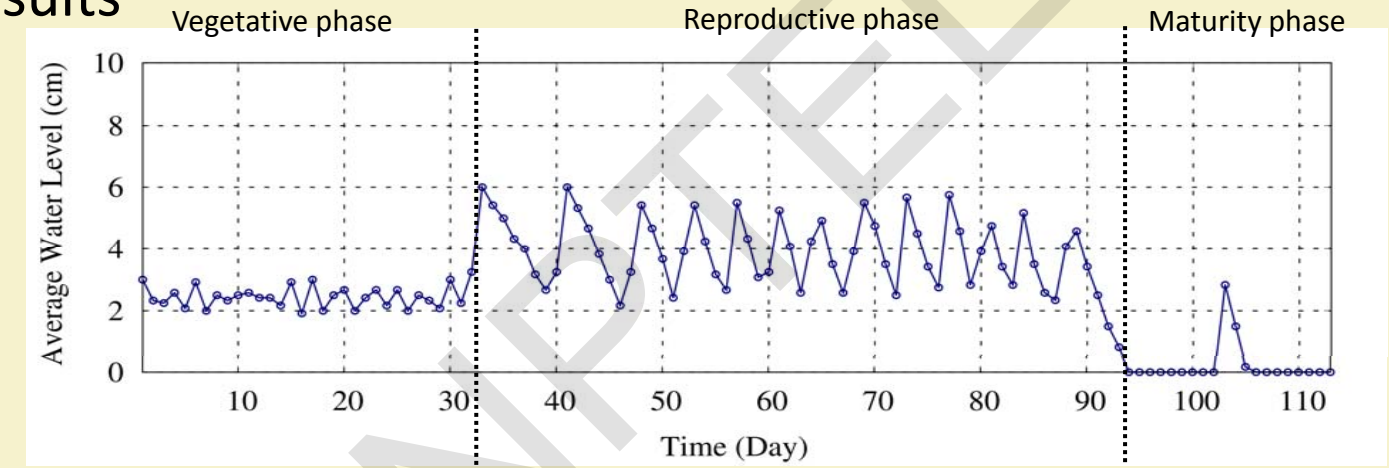


Fig. 7: Average water level

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

AgriSens: Smart Water Management using IoT (Contd.)

✓ Results

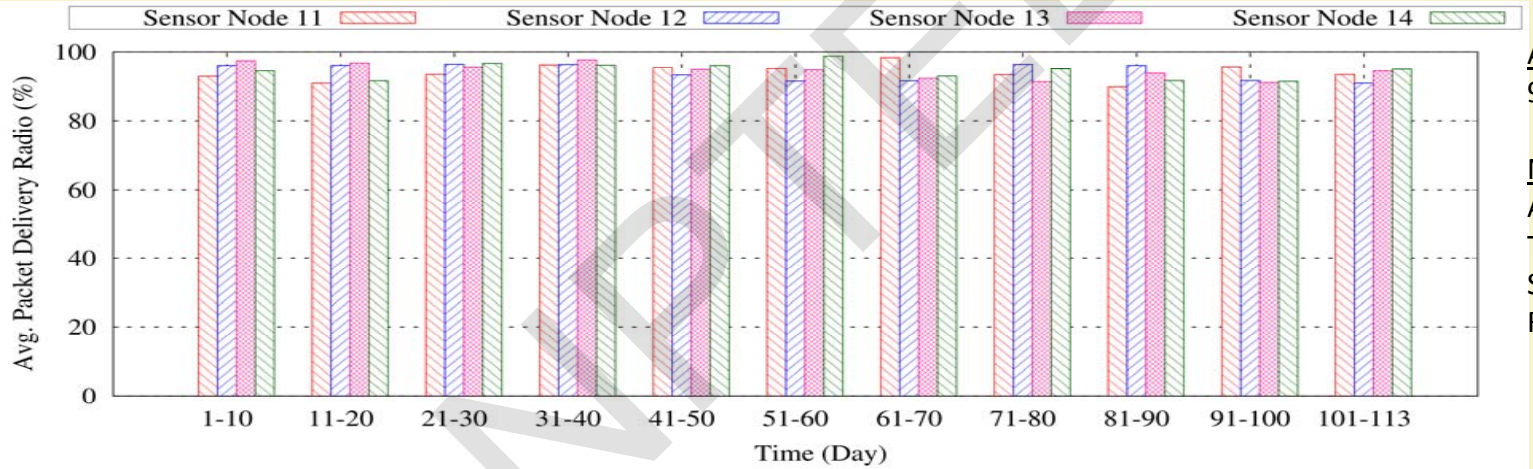


Fig. 8: Average packet delivery ratio

Avg. PDR:
98.75 – 89.75%

Noises:
Air flow,
Temperature,
Solar radiation,
Rain

Source: Project name: Development of a Sensor based Networking System for Improved Water Management for Irrigated Crops, funded by MHRD, Govt. of India

Thank You!!

IIT KHARAGPUR

NPTEL ONLINE
CERTIFICATION COURSES

Case study: Healthcare

Dr. Sudip Misra
Associate Professor
Department of Computer Science and Technology
IIT KHARAGPUR
Email: smisra@sit.iitkgp.ernet.in
Website: http://www.cse.iitkgp.ac.in/~smisra/

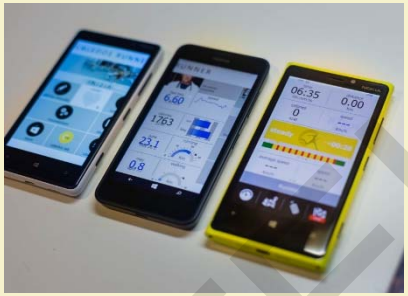
Emergence of IoT Healthcare

- ✓ Advances in sensor and connectivity
  - Collect patient data over time
  - Enable preventive care
  - Understanding of effects of therapy on a patient
- ✓ Ability of devices to collect data on their own
  - Automatically obtain data when and where needed by doctors
  - Automation reduces risk of error
  - Lower error implies increased efficiency and reduced cost

Components of IoT Healthcare

- ✓ Components of IoT is organized in 4 layers
  - Sensing layer: Consists of all sensor, RFIDs and wireless sensor networks (WSN). E.g: Google glass, Fitbit tracker
  - Aggregated layer: Consists of different types of aggregators based on the sensors of sensing layer. E.g: Smartphones, Tablets
  - Processing layer: It consists of servers for processing information coming from aggregated layer.
  - Cloud platform: All processed data are uploaded in cloud platform, which can be accessed by large no. of users

Sensing & Measurement



Data Aggregation

Cloud storage & Analytics

IoT in Healthcare : Directions

IoT Healthcare : Remote Healthcare

- Many people without ready access to effective healthcare
- Wireless IoT driven solutions bring healthcare to patients rather than bring patients to healthcare
- Securely capture a variety of medical data through IoT based sensors, analyze data with smart algorithms
- Wirelessly share data with health professionals for appropriate health recommendations

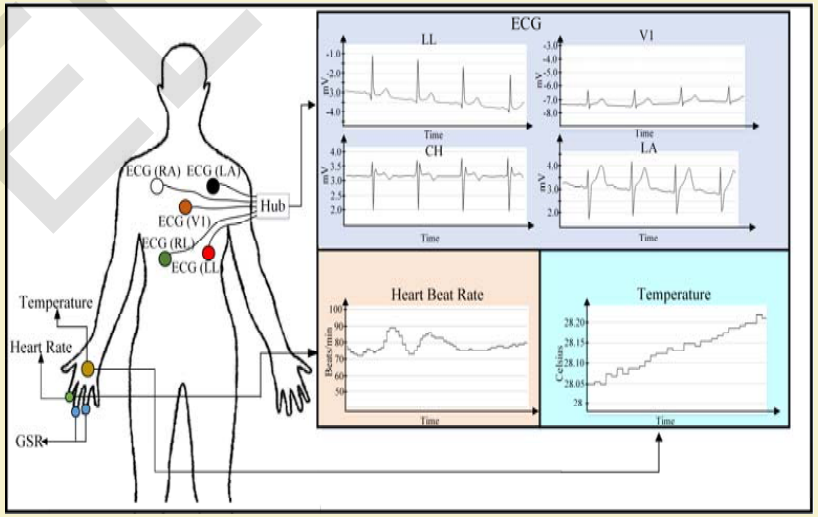
Withings BP Monitor*

Shimmer Temperature Monitor^

*http://www.withings.com/ ^http://www.shimmersensing.com/

IoT Healthcare : Real-time Monitoring

- IoT-driven non-invasive monitoring
- Sensors to collect comprehensive physiological information
- Gateways and cloud-based analytics and storage of data
- Wirelessly send data to caregivers
- Lowers cost of healthcare



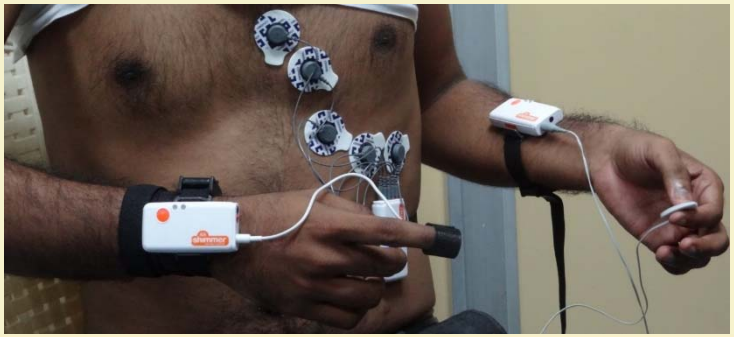
IoT Healthcare : Preventive care

- Fall detection for seniors
- Emergency situation detection and alert to family members
- Machine learning for health trend tracking and early anomaly detection

AmbuSens: Use-case of Healthcare system using IoT

Problem Definition & its Scope

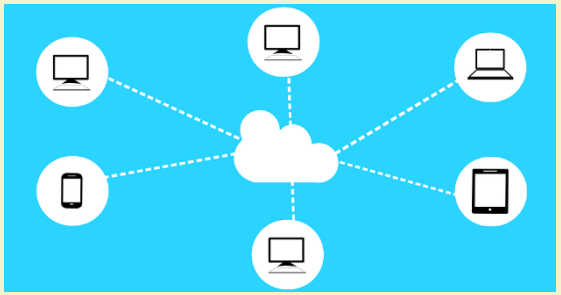
- ✓ **Telemedicine and Remote Healthcare:**
  - Problem - Physical presence necessary
  - Solution - Wireless sensors
- ✓ **Emergency Response Time**:
  - Problem – Not equipped to deal with complications.
  - Solution
    - Instant remote monitoring
    - Feedback by the skilled medical professionals



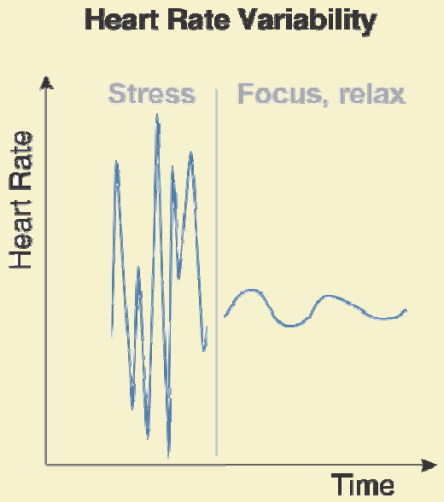
Problem Definition & its Scope (cont.)

- ✓ **Real Time Patient Status Monitoring:**
  - Problem – Lack of collaboration.
  - Solution - Real-time monitoring.
- ✓ **Digitized Medical History:**
  - Problem
    - Inconsistent
    - Physical records vulnerable to wear and tear and loss.
  - Solution - Consistent cloud-based digital record-keeping system

HashID	SensorValue_GSR_kOhms	SensorValue_HeartRate	Cur_Date	Cur_Time
9662SHR23y	97.78	73	Mon 14/11/2016	12:39:06:914 IST
e0c1SHR23y	97.97	73	Mon 14/11/2016	12:39:06:975 IST
7354SHR23y	98.19	75	Mon 14/11/2016	12:39:07:072 IST
7228SHR23y	97.83	75	Mon 14/11/2016	12:39:07:169 IST
ee32SHR23y	98.31	75	Mon 14/11/2016	12:39:07:243 IST
39c2SHR23y	98.48	73	Mon 14/11/2016	12:39:07:300 IST
350cSHR23y	98.99	73	Mon 14/11/2016	12:39:07:328 IST
f002SHR23y	100.36	75	Mon 14/11/2016	12:39:07:400 IST
d228SHR23y	101.01	75	Mon 14/11/2016	12:39:07:491 IST
bed0SHR23y	101.59	77	Mon 14/11/2016	12:39:07:537 IST
a58bSHR23y	100.7	77	Mon 14/11/2016	12:39:07:590 IST
4869SHR23y	100.65	77	Mon 14/11/2016	12:39:07:650 IST

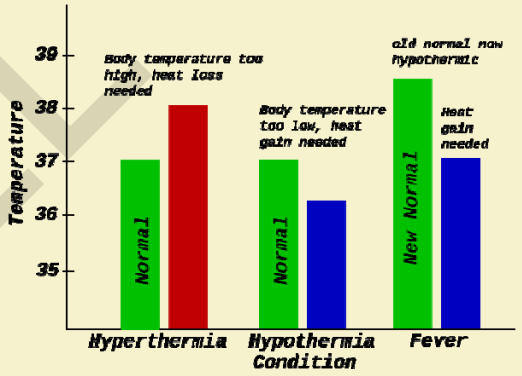


AmbuSens: Physiological Parameters

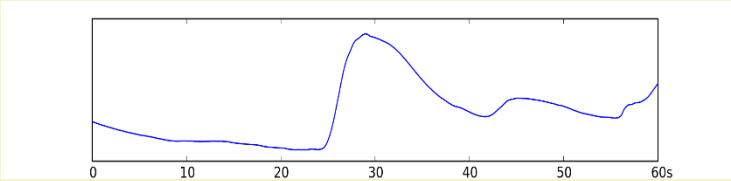


Heart Rate

Electrocardiogram (ECG)



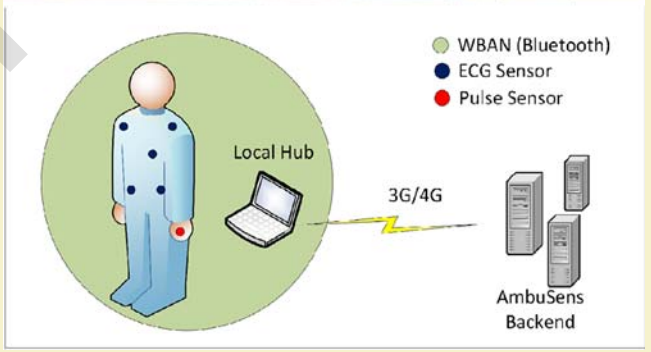
Temperature



Galvanic Skin Response (GSR)

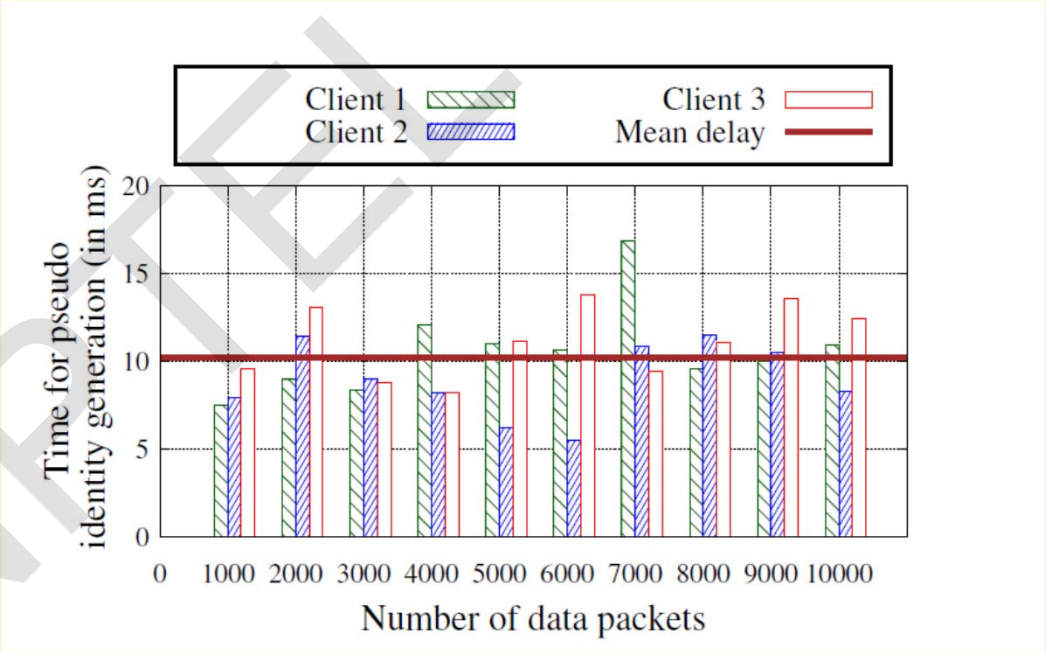
AmbuSens: Development of WBAN

- ✓ Single hop ***wireless body area network*** (WBAN)
- ✓ Communication protocol used is ***Bluetooth*** i.e. IEEE 802.15.1
- ✓ ***Power management*** and data-rate tuning
- ✓ ***Calibration*** of data
- ✓ ***Filtering*** and noise removal



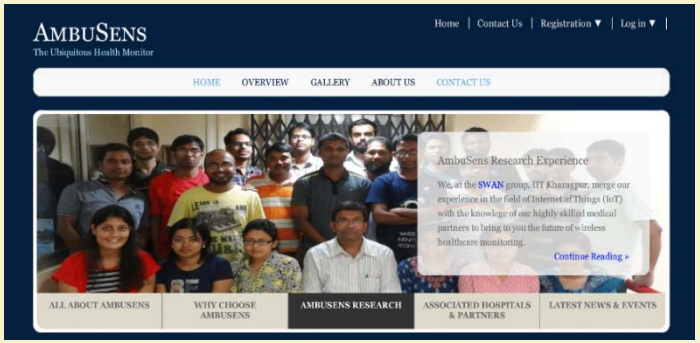
AmbuSens: Development of Cloud Framework

- ✓ ***Health-cloud*** framework
- ✓ The developed system is strictly ***privacy-aware***
- ✓ ***Patient-identity masking*** involves hashing and reverse hashing of patient ID
- ✓ ***Scalable*** architecture

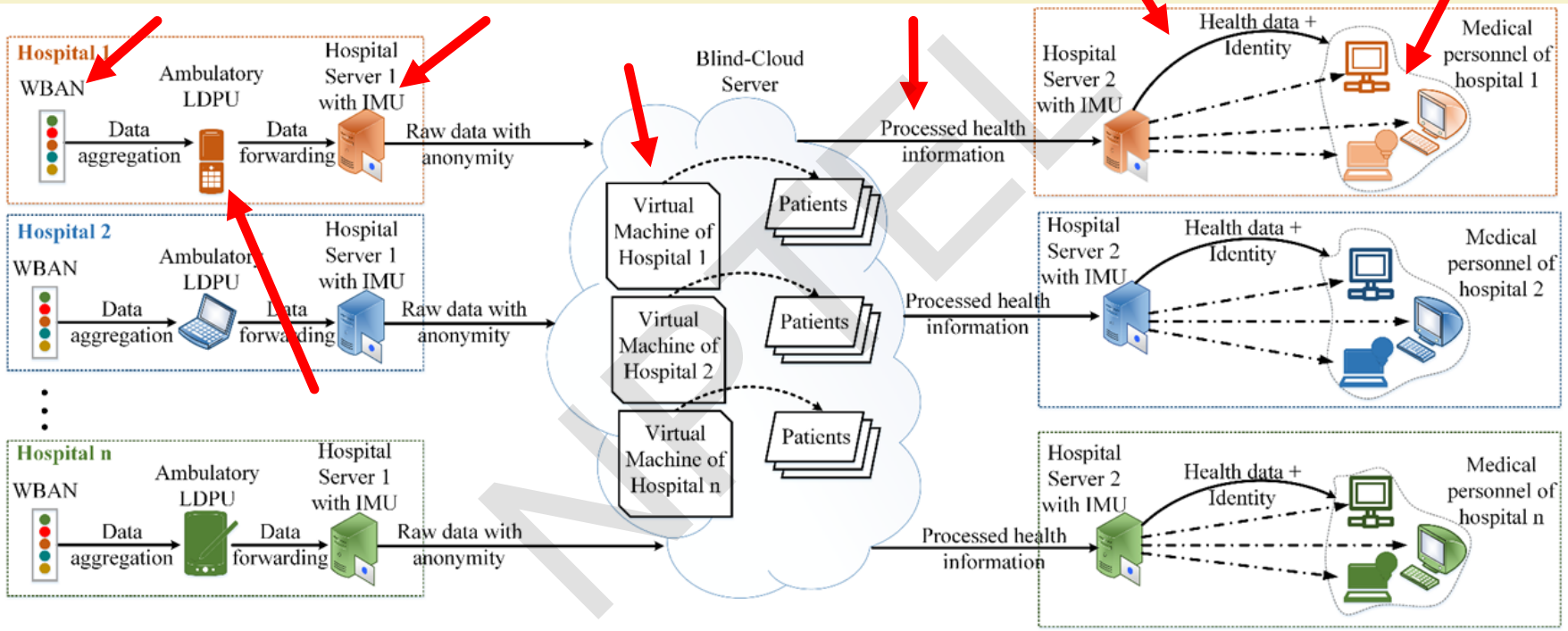


AmbuSens: Web Interface

- ✓ **URL**: ambusens.iitkgp.ac.in
- ✓ ***Paramedic*** and ***Doctor*** portals for ease of use.
- ✓ Provision for recording ***medical history*** and sending ***feedback.***
- ✓ Allows sensor ***initialization*** and ***data streaming.***
- ✓ Includes data ***visualization*** tools for better understanding.

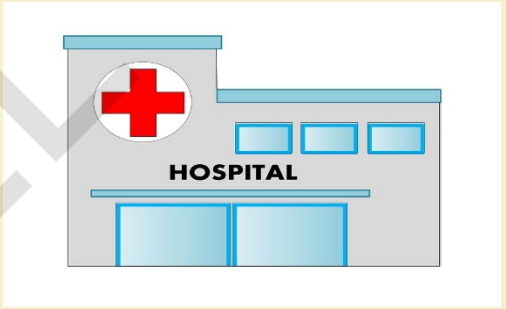


AmbuSens: System Architecture



AmbuSens: Implementation

- ✓ AmbuSens Implementation demo
  - Field demo animation
    - Part 1
      - AmbuSens in the ***Hospital***
      - Brief description of the ***sensors***
    - Part 2
      - ***Ambulatory*** Healthcare



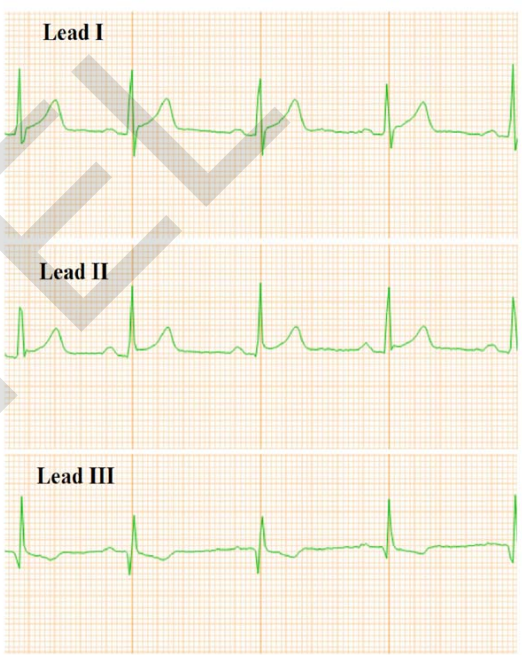
AmbuSens: System Trials

Figure 1: Hospital system trials

Figure 2: Ambulatory system trials

AmbuSens: Results (Comparison of ECG tracing)

ECG tracing from manual system



Real-time ECG tracing from AmbuSens

Thank You

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

Activity Monitoring - Part 1

Dr. Sudip Misra
Associate Professor
Department of Computer Science and Engineering
IIT KHARAGPUR
Email: smisra@sit.iitkgp.ernet.in
Website: http://cse.iitkgp.ac.in/~smisra/

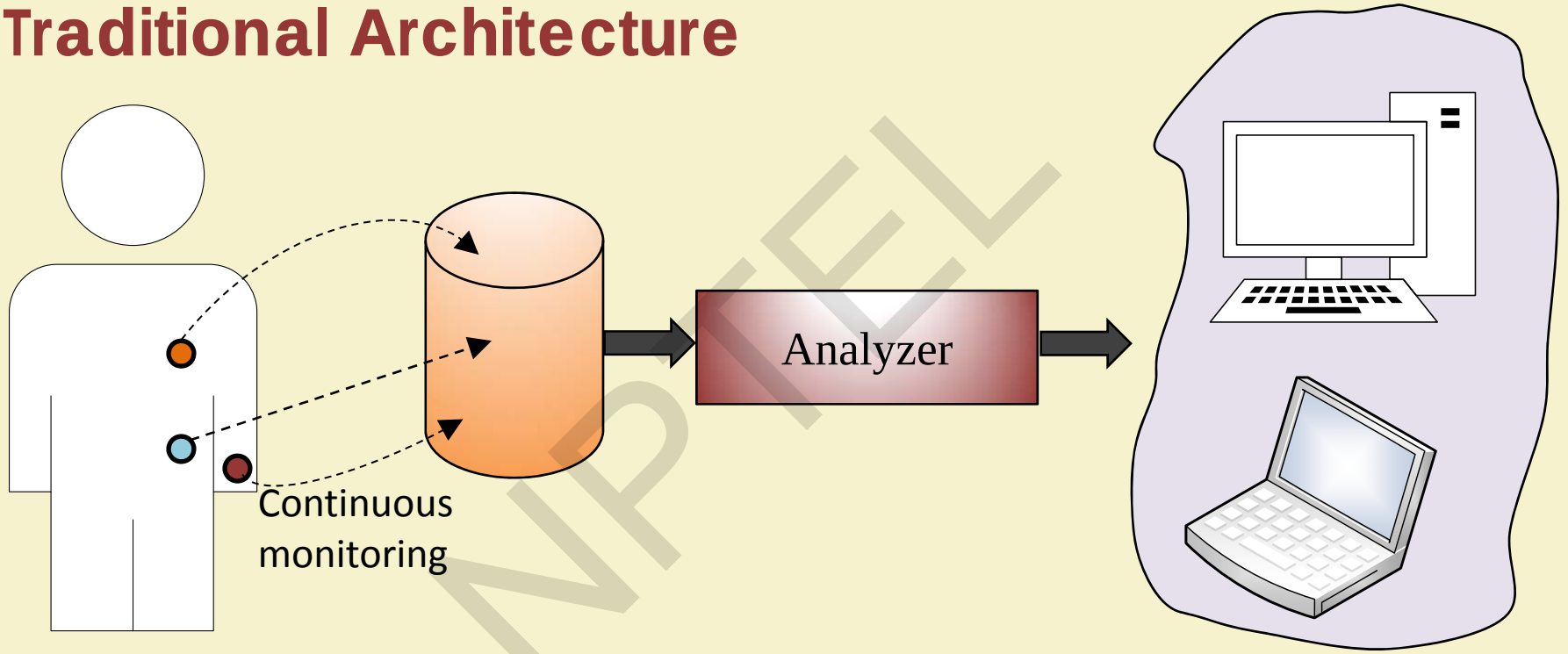
Introduction

- ✓ Wearable sensors have become very popular for different purposes such as:
  - Medical
  - Child-care
  - Elderly-care
  - Entertainment
  - Security
- ✓ These sensors help in monitoring the physical activities of humans

Introduction (Contd.)

- ✓ Particularly in IoT scenarios, activity monitoring plays an important role for providing better quality of life and safe guarding humans.
- ✓ Provides information accurately in a reliable manner
- ✓ Provides continuous monitoring support.

Traditional Architecture



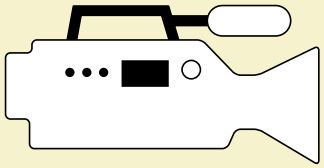
Advantages

- ✓ Continuous monitoring of activity results in daily observation of human behavior and repetitive patterns in their activities.
- ✓ Easy integration and fast equipping
- ✓ Long term monitoring
- ✓ Utilization of sensors of handheld devices
  - Accelerometer
  - Gyroscope
  - GPS
  - Others

Important Human Activities

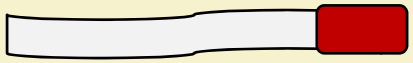
Actions	• Running • Jumping
Gesture	• Folding legs • Moving hand

Types of Sensors



Camera

Smart Phone



Activity Tracker Band

Data Analysis Tools

- ✓ Statistical
  - Sensor data
- ✓ Machine Learning Based
  - Sensor data
- ✓ Deep Learning Based
  - Sensor data
  - Images
  - Videos

Approaches

- ✓ In-place
  - On the device
  - Power intensive
  - No network connection required
- ✓ Network Based
  - Larger and processing intensive methods can be applied
  - Group based analytics possible
  - Low power consumption
  - Average to good network connection

Thank You!!