Data Mining for Sustainable Data Centers

Manish Marwah
Senior Research Scientist
Sustainable Ecosystem Research Group
Hewlett Packard Laboratories
manish.marwah@hp.com

Motivation

Industry challenge:

Create technologies, IT infrastructure and business models for the low-carbon economy

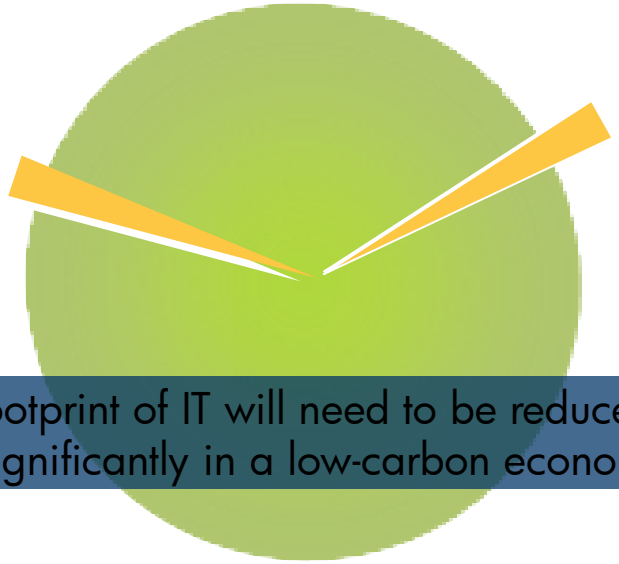
IT industry

2%

Total carbon emissions

Aviation

2%



The footprint of IT will need to be reduced quite significantly in a low-carbon economy.

Motivation

Industry challenge:

Create technologies, IT infrastructure and business models for the low-carbon economy

Total carbon emissions

IT industry

2%

The rest of the global economy

98%

IT must play a central role in addressing the global sustainability challenge.

Sustainability

"sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs"

the Brundtland Commission of the United Nations, 1987

Sustainability

What do I mean by "sustainability"?

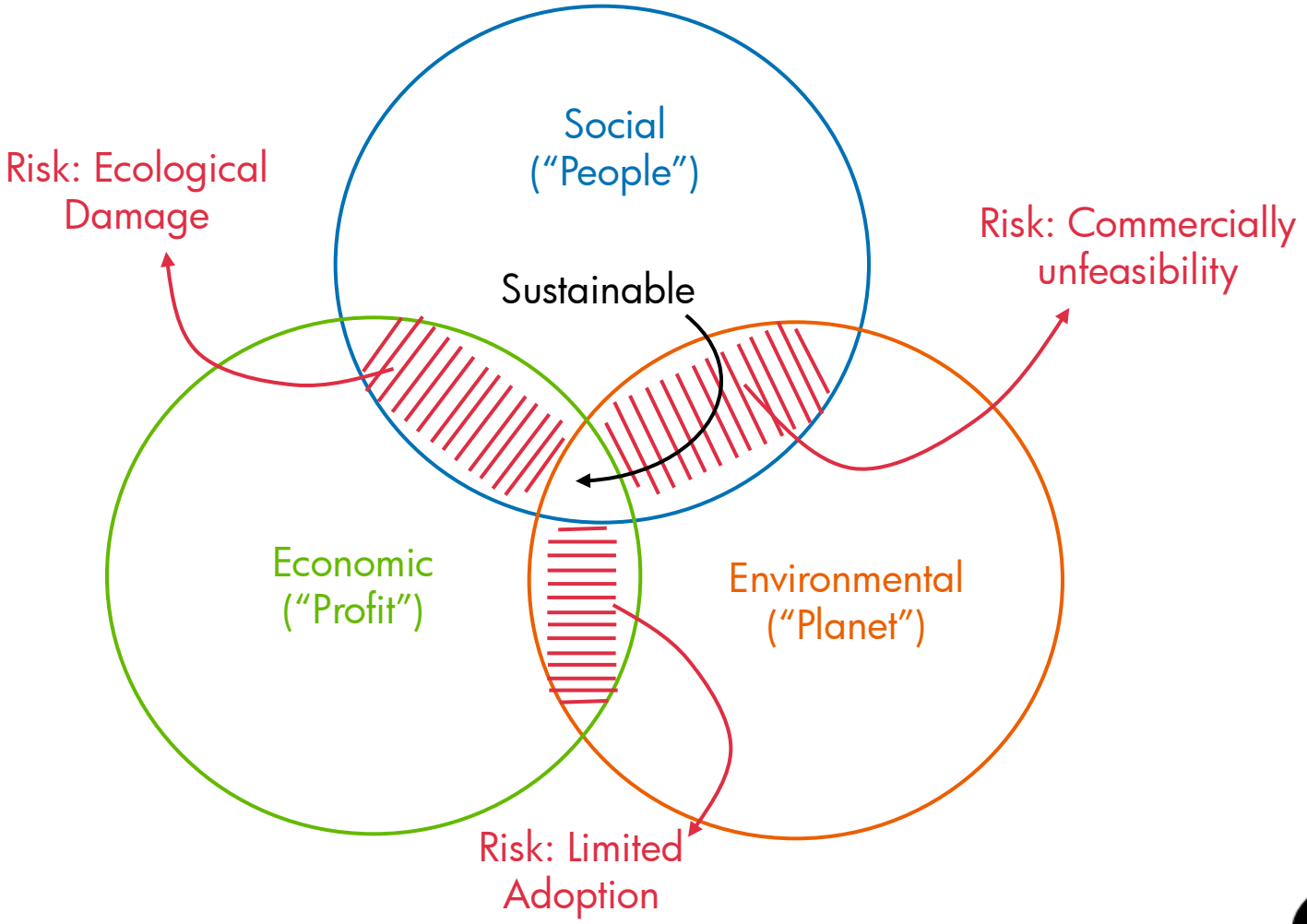
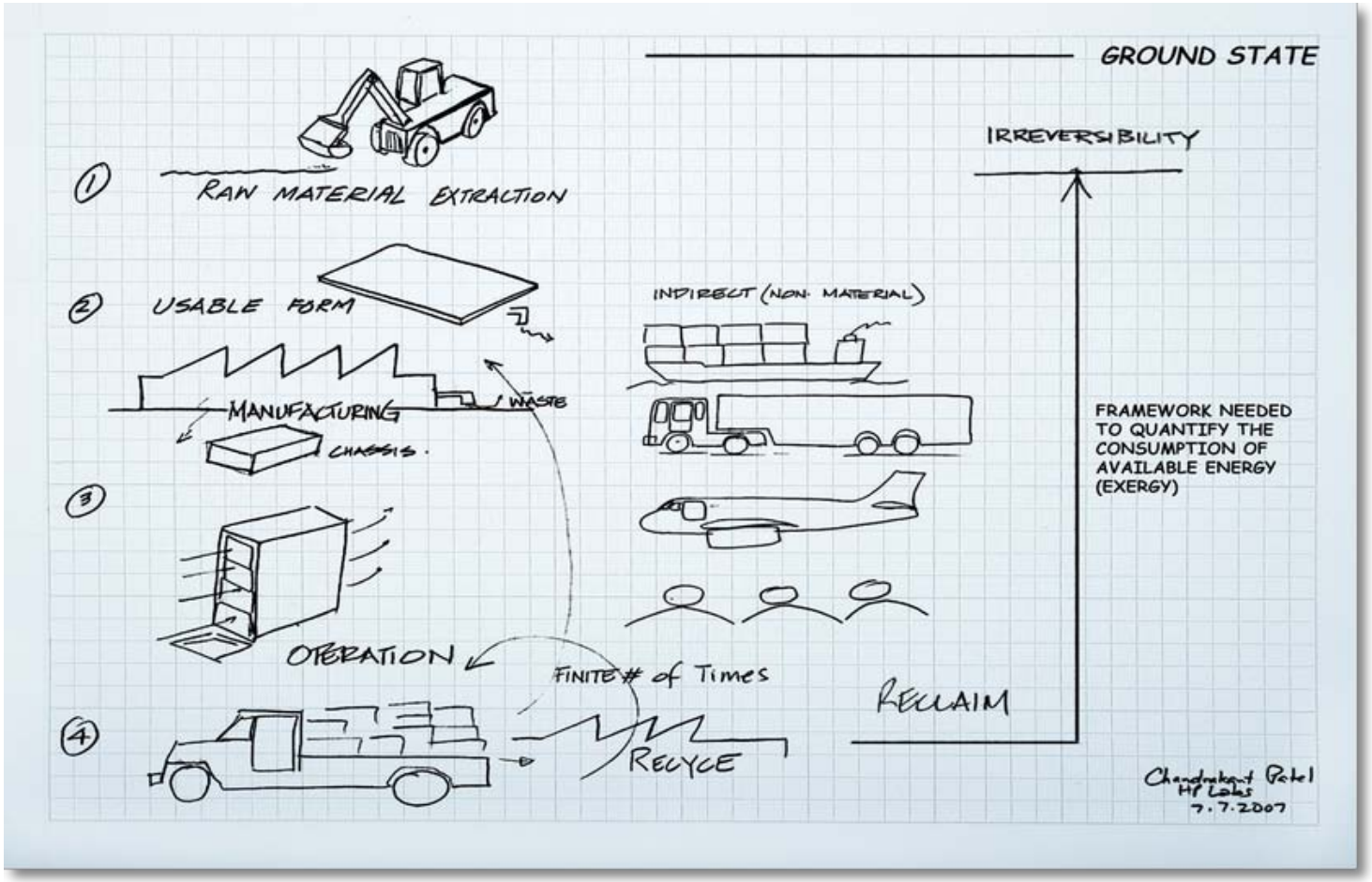


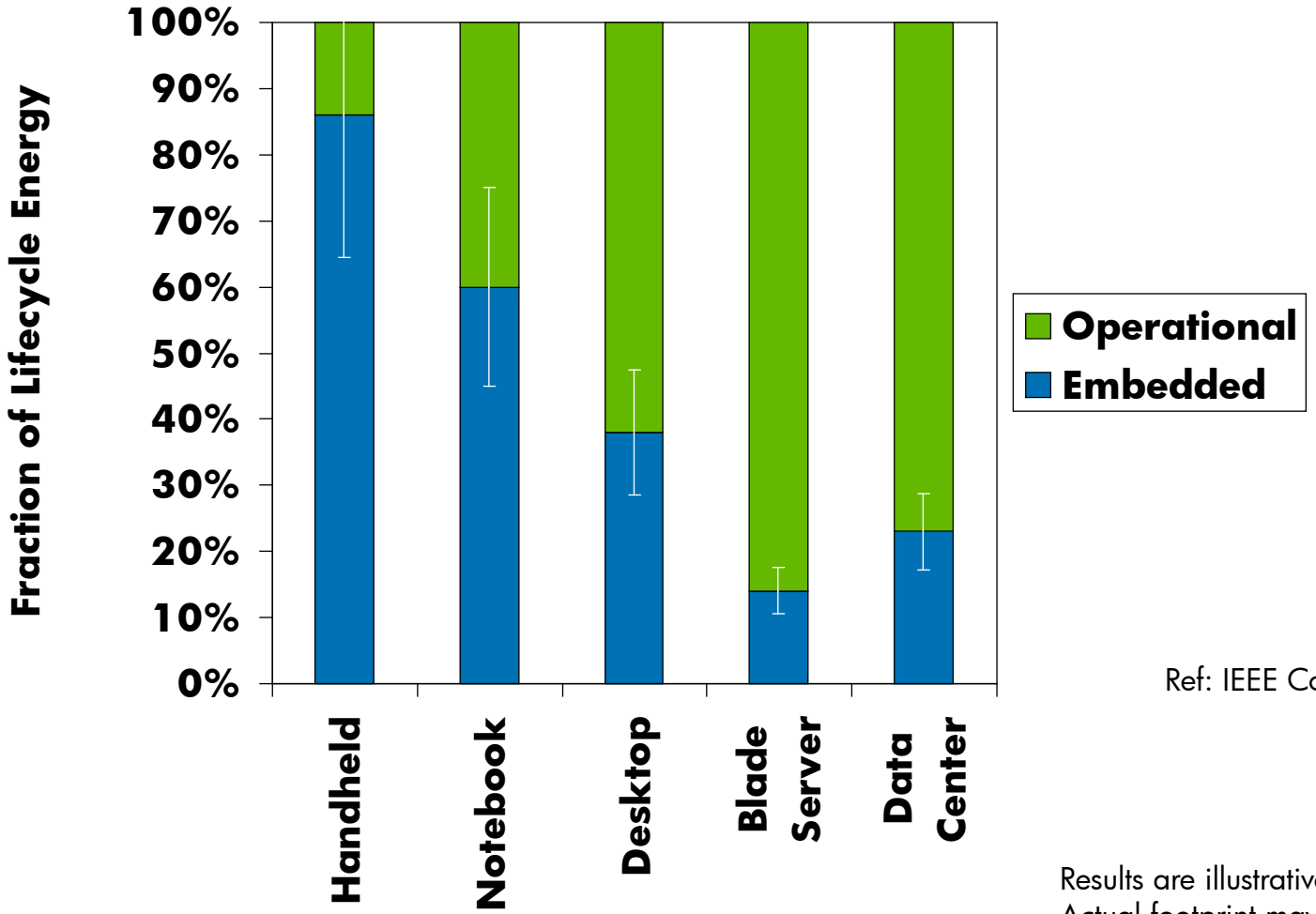
Figure Credit: A. Agogino, UC Berkeley

Environmental Sustainability

- Life Cycle View



Sustainable Data Centers
Lifecycle Assessment

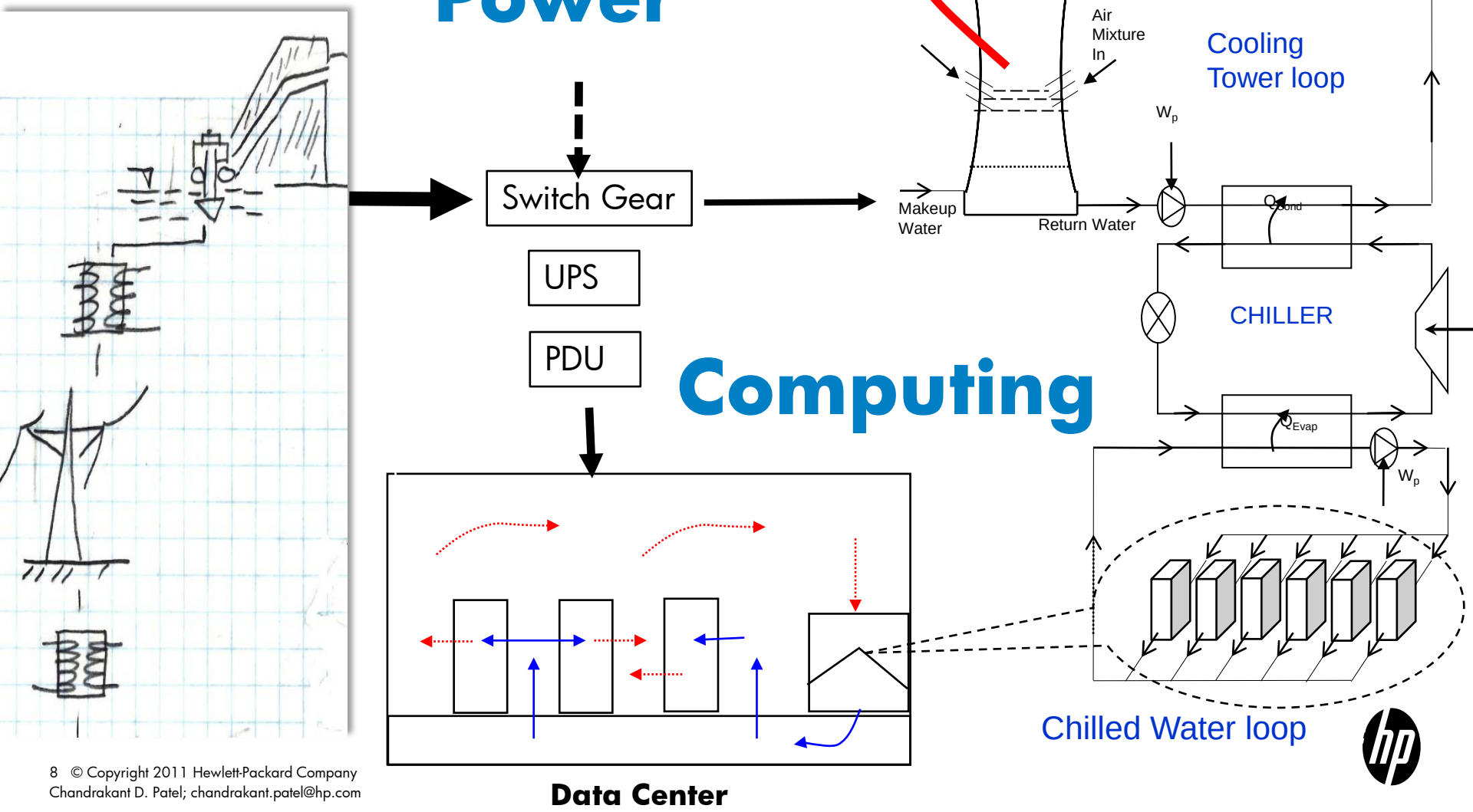


Ref: IEEE Computer 2009

Results are illustrative only.
Actual footprint may differ.

Cloud Data Center

Supply and Demand Side



Chandrakant D. Patel; chandrakant.patel@hp.com

Sustainable Ecosystem Research Group
HP Labs

- Sustainable Data Center
  - Integrated management of IT, power and cooling towards a net-zero data center

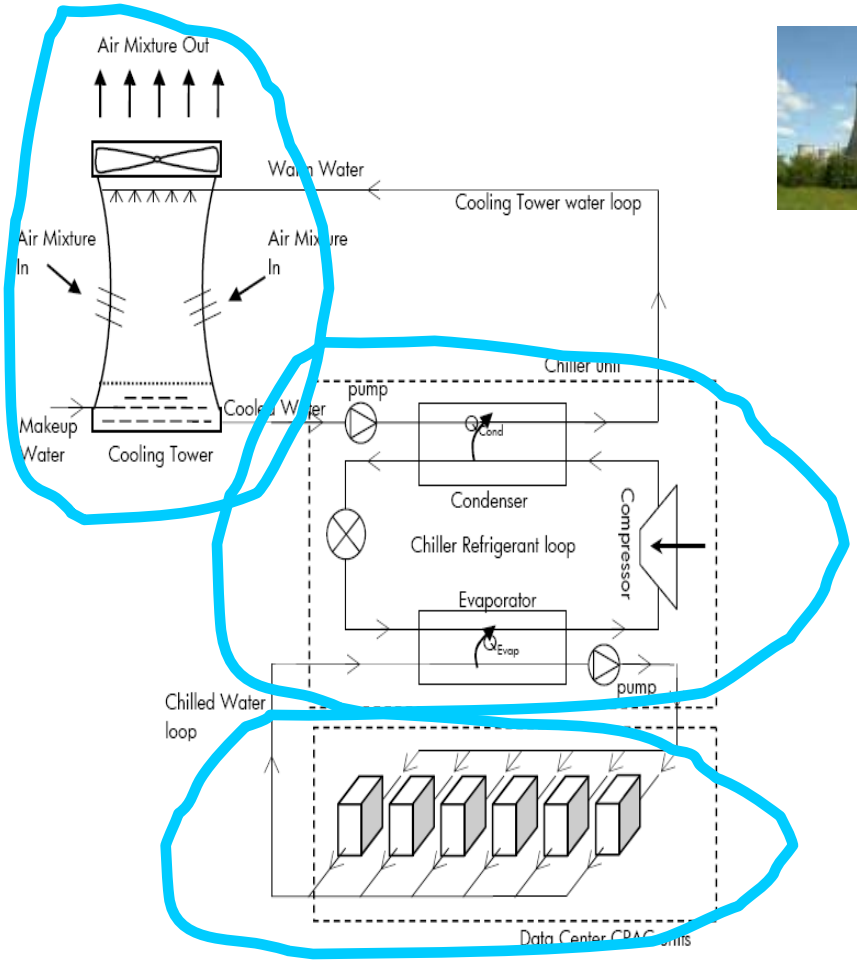
- Resource Management as a Service
  - Improve sustainability of urban infrastructure, e.g. power, water.

Sustainable Operation and Management of Chillers using Temporal Data Mining (KDD '09)

- Data Centers
  - Cooling Infrastructure
- Problem Statement
- Prior Work
- Our Approach
  - Symbolic representation
  - Event encoding
  - Motif mining
  - Sustainability characterization
- Experimental Results
- Summary

Data Center Cooling Infrastructure

Consumes from 1/3 up to 1/2 of total power consumption



Cooling Towers

Chiller Unit

Water Return (T_{in}) →  → Water Supply (T_{out})

Computer room air-conditioner (CRAC)

Ensemble of Chillers

- Challenging to operate efficiently
  - Complex physical system
    - Dynamic
    - Heterogeneous
    - Inter-dependencies
    - Many constraints
  - Accurate models not available
  - Rapid cycles undesirable – reduce lifespan
- Domain experts determine settings based on heuristics
- Can it be automated through a data-driven approach?

Chiller Ensemble

- Which unit to turn ON/OFF?
- At what utilization?
- How to handle increase/decrease in cooling load?

Problem Statement

- Given the following chiller time series
  - utilization levels
  - power consumption
  - cooling loads
- Is it possible to determine which operational settings are more energy efficient?
- And then use this information to advise data center facility operators

Some Terminology

- IT cooling load
- Chiller utilization
- Chiller power consumption
- Coefficient of performance (COP)

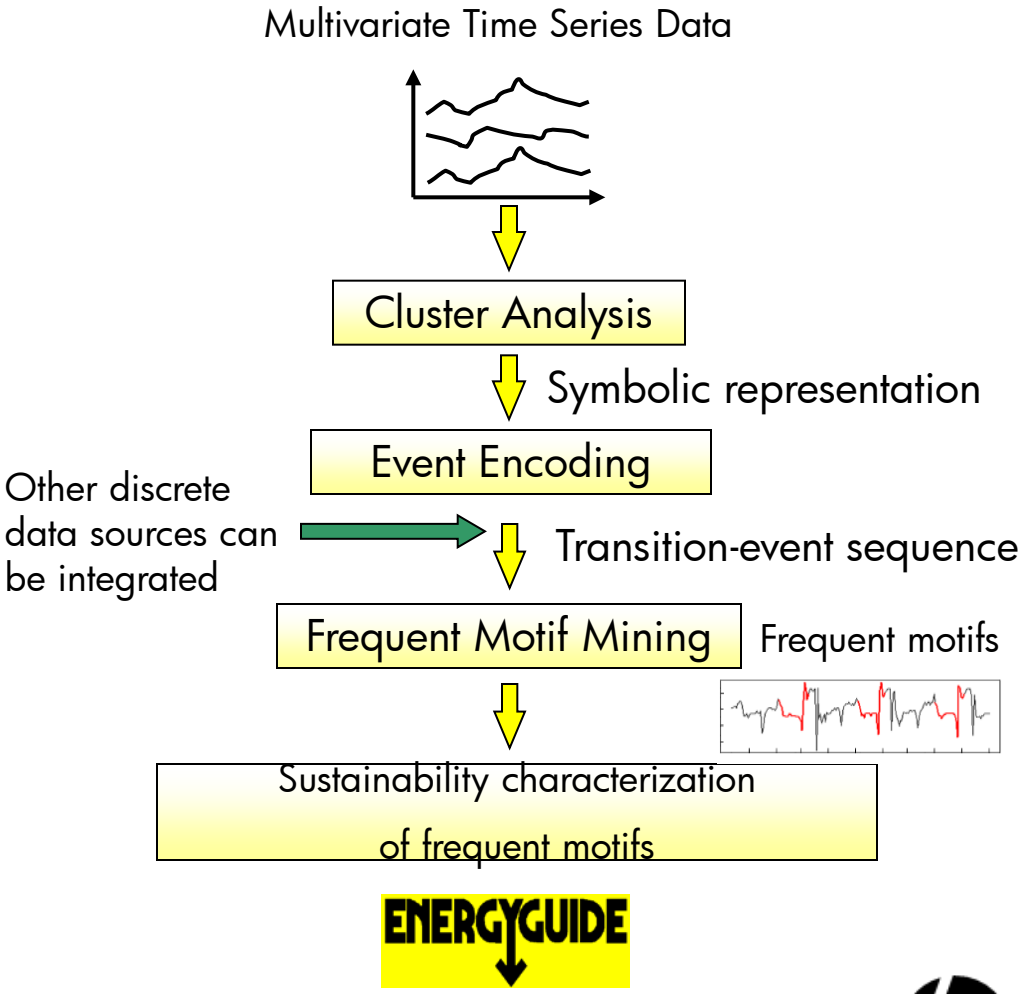
$$\frac{\text{Cooling Load}}{\text{Power consumption}}$$

Prior Work

- Classical approaches to model time series data
  - Principal component analysis
  - Discrete Fourier transforms
- Discrete representations: SAX [Keogh et al.]
- Motifs: Repeating subsequences [Yankov et al.]

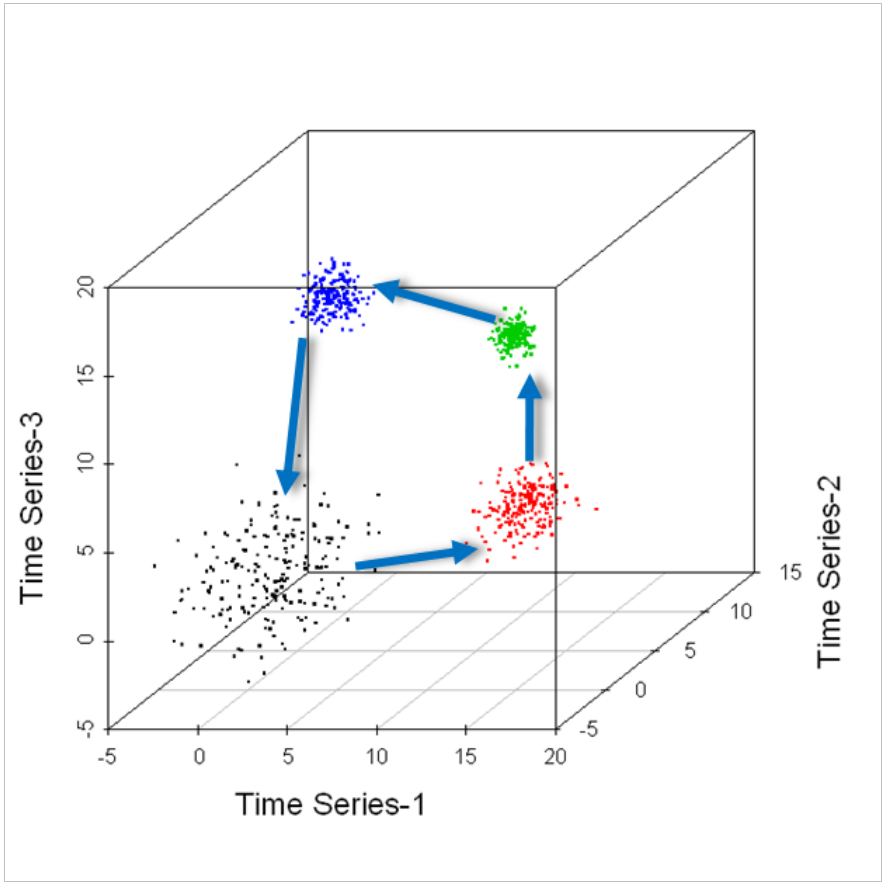
Our approach

- Goal: Sustainability characterization of multi- variate time series data
  - Chiller utilization data
- Four Main Steps
  - Symbolic representation
  - Event encoding
  - Motif mining
  - Sustainability Characterization



Clustering

- Individual vector: Utilization across all chiller units
- Raw Data: Sequence of such vectors
- Perform k-means clustering
- Use cluster labels to encode multi-variate time series



Event Encoding and Motif Mining

- Event sequences
- Motif mining
  - Episode Framework
  - Non-overlapped occurrences
  - Inter-event gap constraint

Some Definitions

- Event Sequence

$$\langle (E_1, t_1), (E_2, t_2), \ldots, (E_N, t_N) \rangle$$

E_i = Event type t_i = Time of occurrence

$$\langle (A,1), (B,3), (D,4), (C,6), (A,12), (E,14), (B,15), (D,17), (C,20), (A,21) \rangle$$

- Episode
  - Ordered collection of events occurring together

$$(A \rightarrow B \rightarrow C)$$

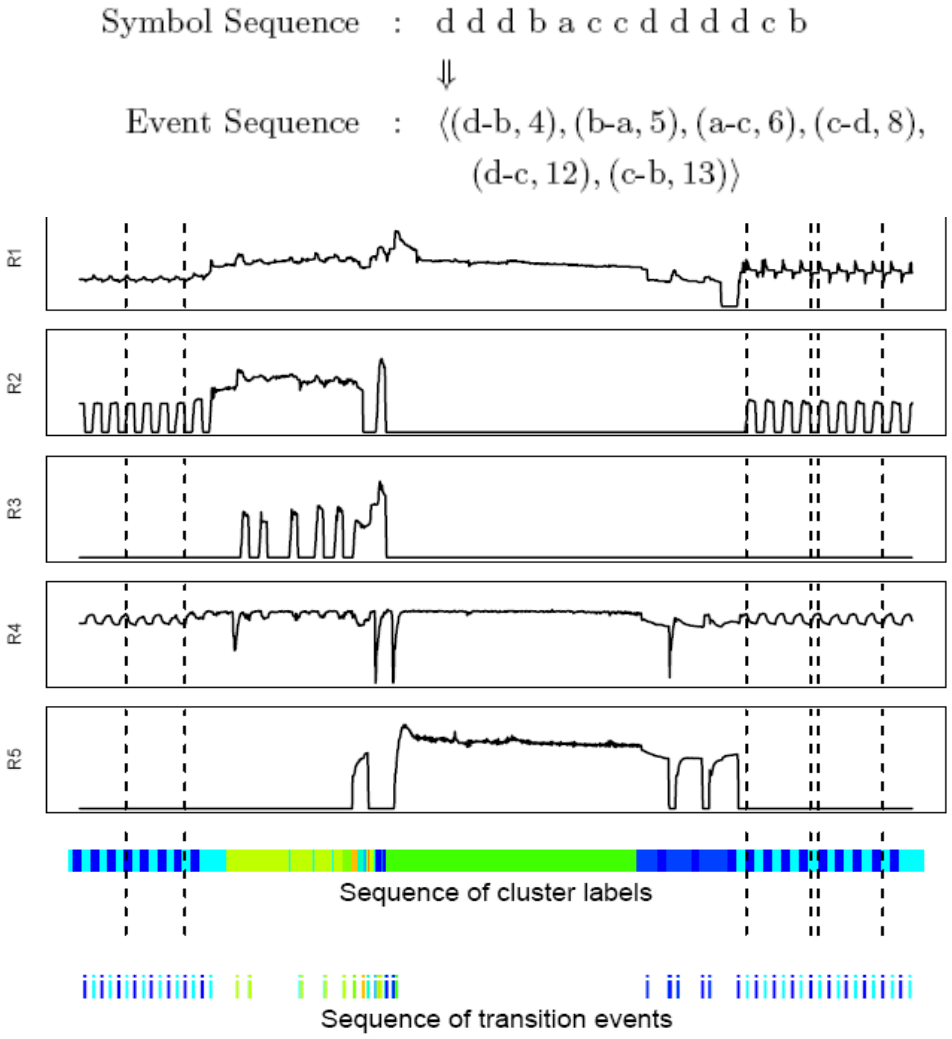
- Episode occurrence
  - Events same ordering as episode in the **data**.

<**(A,1)**, **(B,3)**, (D,4), **(C,6)**, (E,12), **(A,14)**, **(B,15)**, **(C,17)**>

- Motifs
  - Frequently occurring episodes

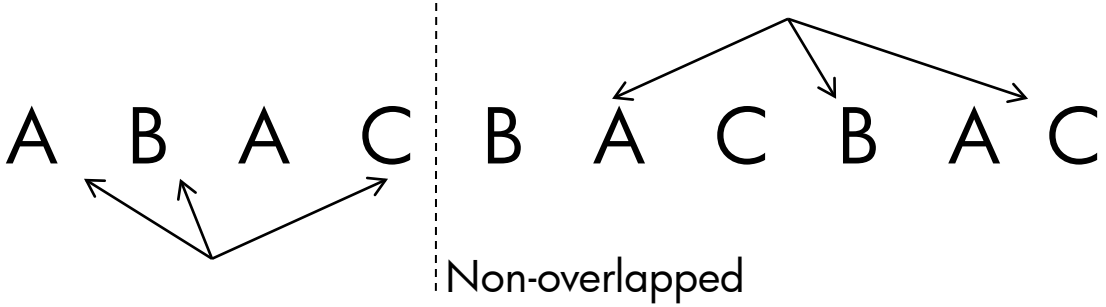
Redescribing time series data

- Perform run-length encoding:
  - Note transitions from one symbol to another
- Higher level of abstraction
  - Transition events



Motif mining

- Frequency counting: Non-overlapped occurrences



- Level-wise (Apriori-style) episode mining

Itemset Mining/Association rule mining

- Example: Market Basket Analysis
- Items frequently purchased together:

Bread ⇒PeanutButter

- Uses:
  - Placement
  - Advertising
  - Sales
  - Coupons

Apriori Algorithm

- ***Frequent Itemset Property:***

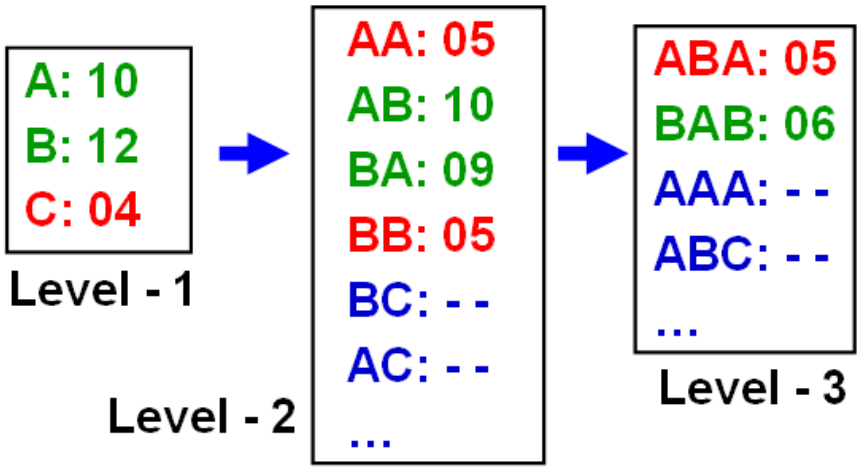
Any subset of a frequent itemset is frequent.

- Contrapositive:

If an itemset is not frequent,

none of its supersets are frequent.

Level-wise (Apriori-based) motif mining

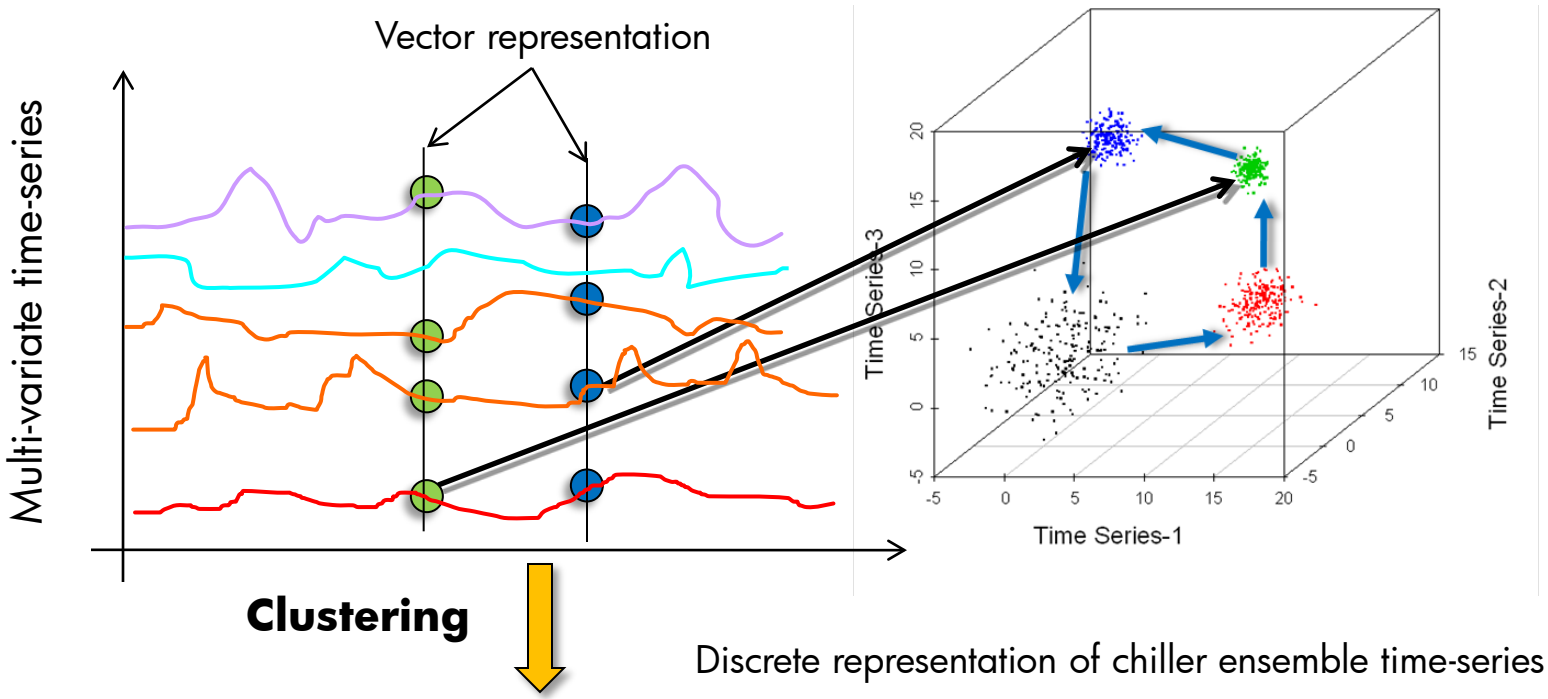


Candidate generation followed by counting

Episode Counting

- Finite state automata based counting algorithm
- Support = |largest set of non-overlapped occurrences of transition-event episodes|
- Count allows gaps or intervening junk symbols

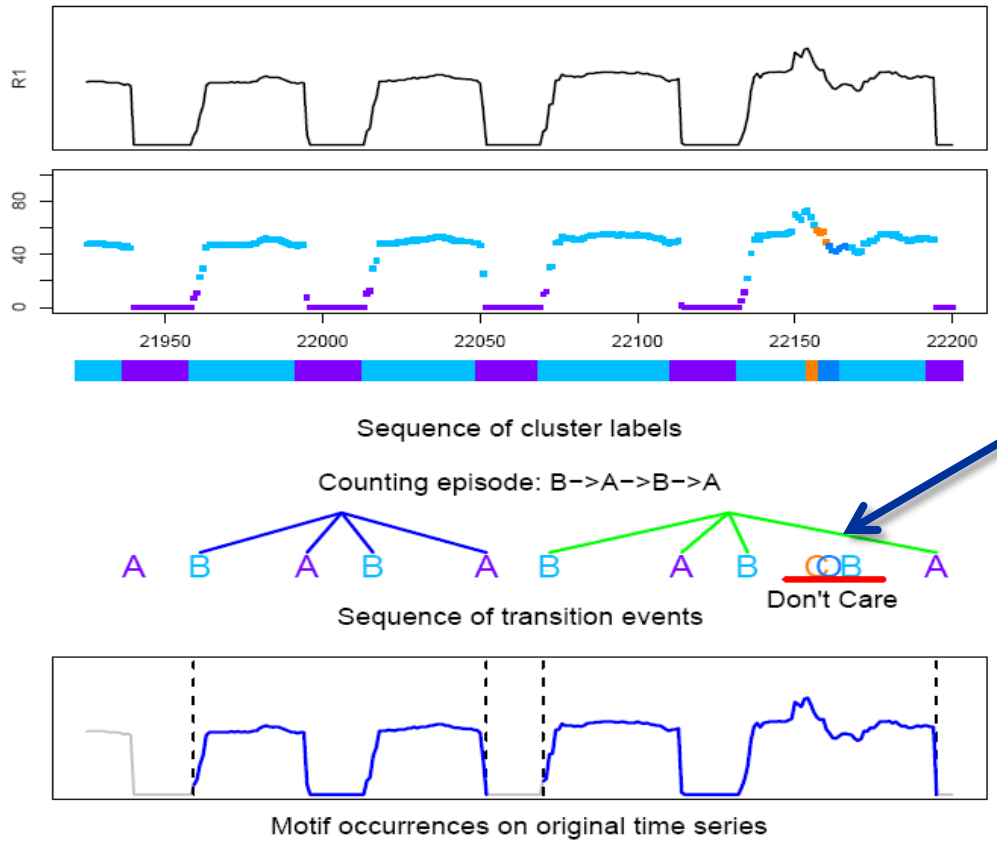
Methodology Summary



Advantages of our approach

- We model transitions from one state to another
  - States correspond to clusters
- We allow don't cares between state transitions in a more expressive way
  - Provides robustness to clustering
- Result of mining is a set of occurrences of a motif
  - Motifs must repeat at least N times to be considered frequent
  - Lowers the likelihood of finding false positives

Robustness of motif occurrences



Don't care transition events in the encoded sequence

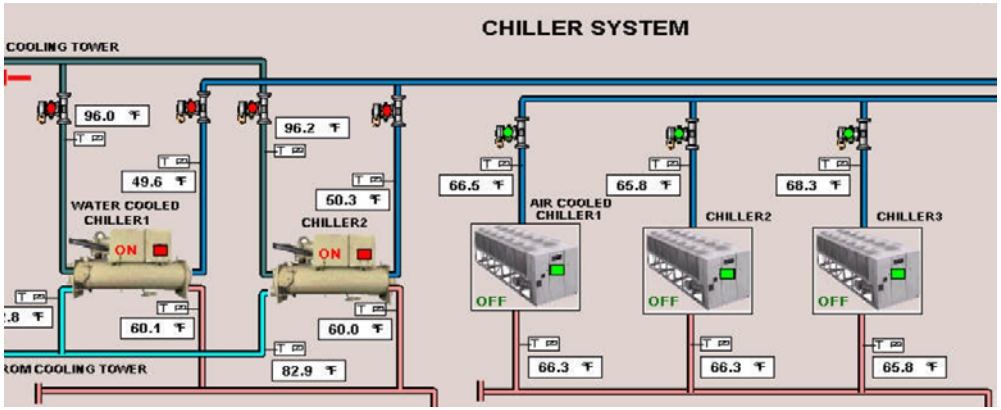
Matches approximately similar patterns

Sustainability characterization of Motifs

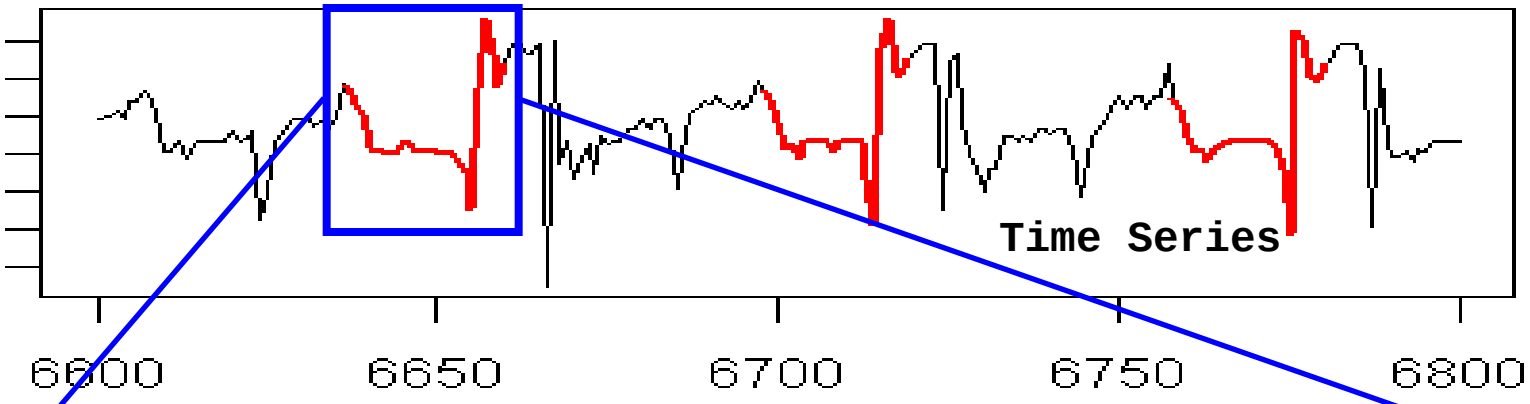
- Average motif COP (coefficient of performance)
  - Indicates cooling efficiency of a chiller unit
    - $\text{COP} = \dfrac{\text{IT Cooling Load}}{\text{Power consumed}}$
- Frequency of oscillations of a motif
  - Impacts chiller lifespan
  - Normalized number of mean-crossings

Experimental Results

- Data
  - From HP R&D data center in Bangalore
    - 70,000 sq ft
    - 2000 racks of IT equipments
  - Ensemble of five chiller units
    - 3 air cooled chillers
    - 2 water cooled chillers
  - 480 hours of data
    - July 2 – 7, Nov 27 – 30, Dec 16 – 26, 2008
- 22 motifs found in the data



A Motif – Detailed Example (1/3)



Symbol seq: 11,11,11,8,8,8,8,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,13,14,12,12,11,11,11

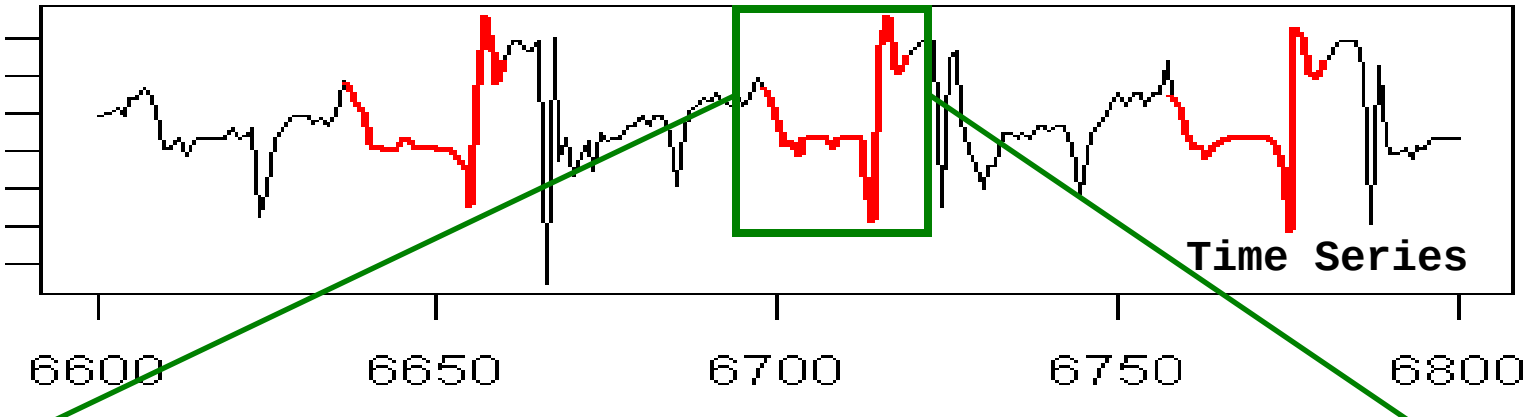
Encoded seq: 11-8,6637 8-10,6641 10-13,6656 13-14,6657 14-12,6658 12-11,6660

4 15 1 1 2

Transition Motif: [11-8 , 8-10 , 10-13 , 13-14 , 14-12, 12-11]

Inter-transition gap constraint = 20 min

A Motif – Detailed Example (2/3)



Symbol seq: 11,11,11,8,8,8,10,10,10,10,10,10,10,10,10,10,10,10,13,13,14,12,12,11,11,11

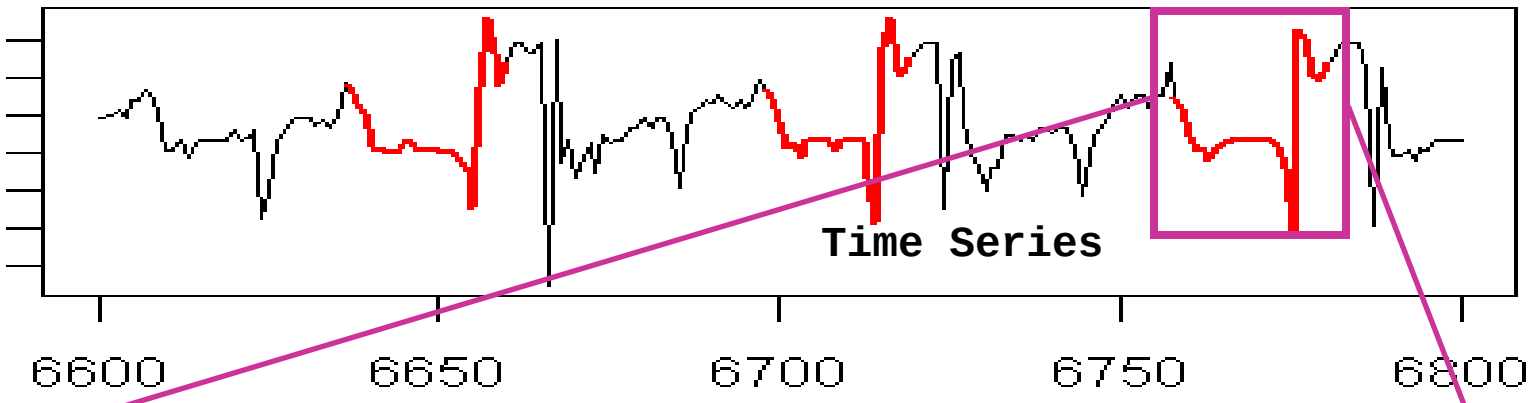
Encoded seq: 11-8,6698 8-10,6701 10-13,6714 13-14,6716 14-12,6717 12-11,6719

3 13 2 1 2

Transition Motif: [11-8 , 8-10 , 10-13 , 13-14 , 14-12, 12-11]

Inter-transition gap constraint = 20 min

A Motif – Detailed Example (3/3)



Symbol seq: 11,11,11,8,8,8,10,10,10,10,10,10,10,10,10,10,10,10,10,13,14,12,12,12,11,11,11

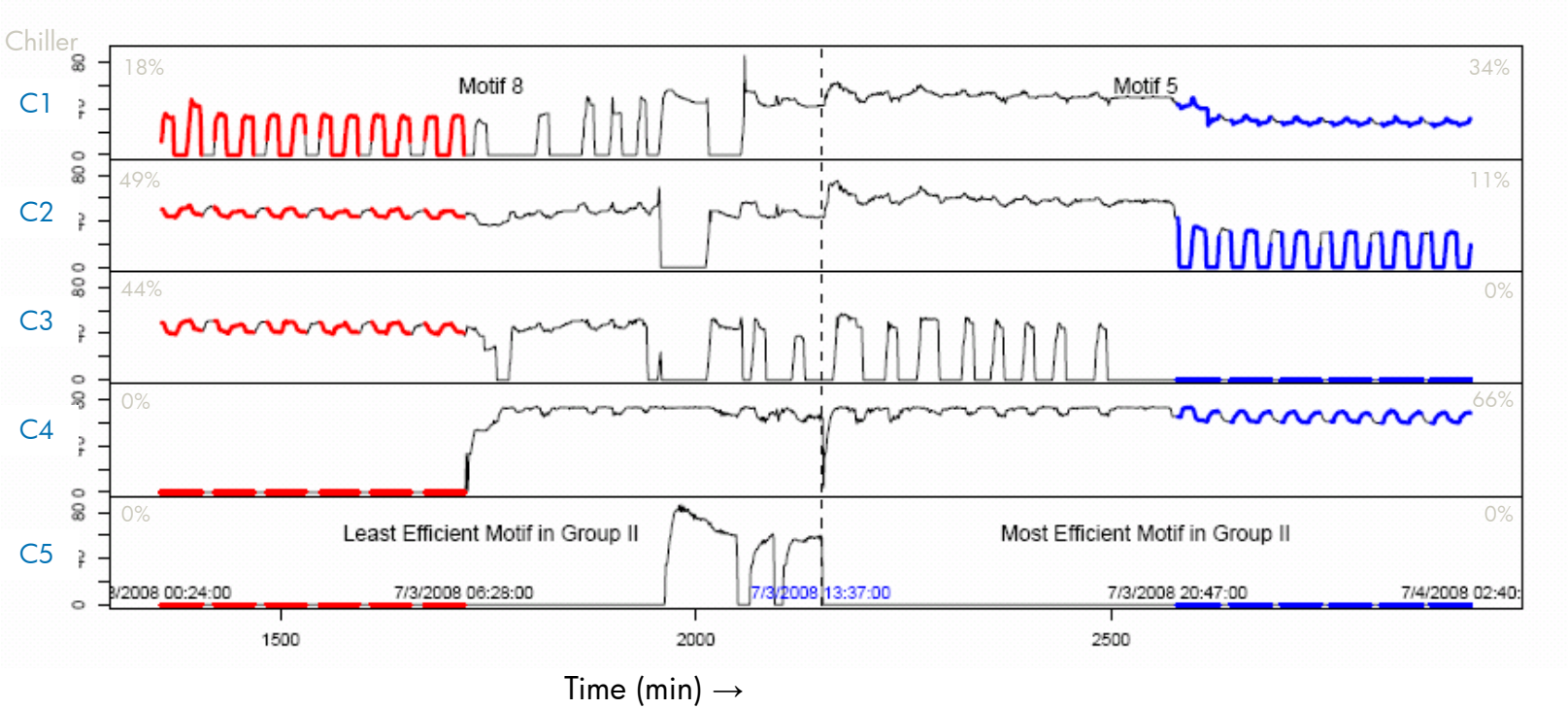
Encoded seq: 11-8,6758 8-10,6761 10-13,6775 13-14,6776 14-12,6777 12-11,6780

3 14 1 1 3

Transition Motif: [11-8 , 8-10 , 10-13 , 13-14 , 14-12, 12-11]

Inter-transition gap constraint = 20 min

Two Interesting Motifs



C1, C2, C3 → Air cooled

C4, C5 → Water cooled

Motif 8

Motif 5

	Motif 8	Motif 5
COP	4.87	5.40
Units operating	3 air-cooled	2 air-cooled, 1 water cooled

Potential Savings

	Load (KW)		Most Efficient Motif	Least Efficient Motif	Potential Power Savings	
	Ave.	Std			KW	%
Group II	2089	35	5	8	41	9.83%

- Annual saving from operating in Motif 5 instead of Motif 8
  - Cost savings = $40,000 (~10%)
  - Carbon footprint savings = 287,328 kg of CO_2

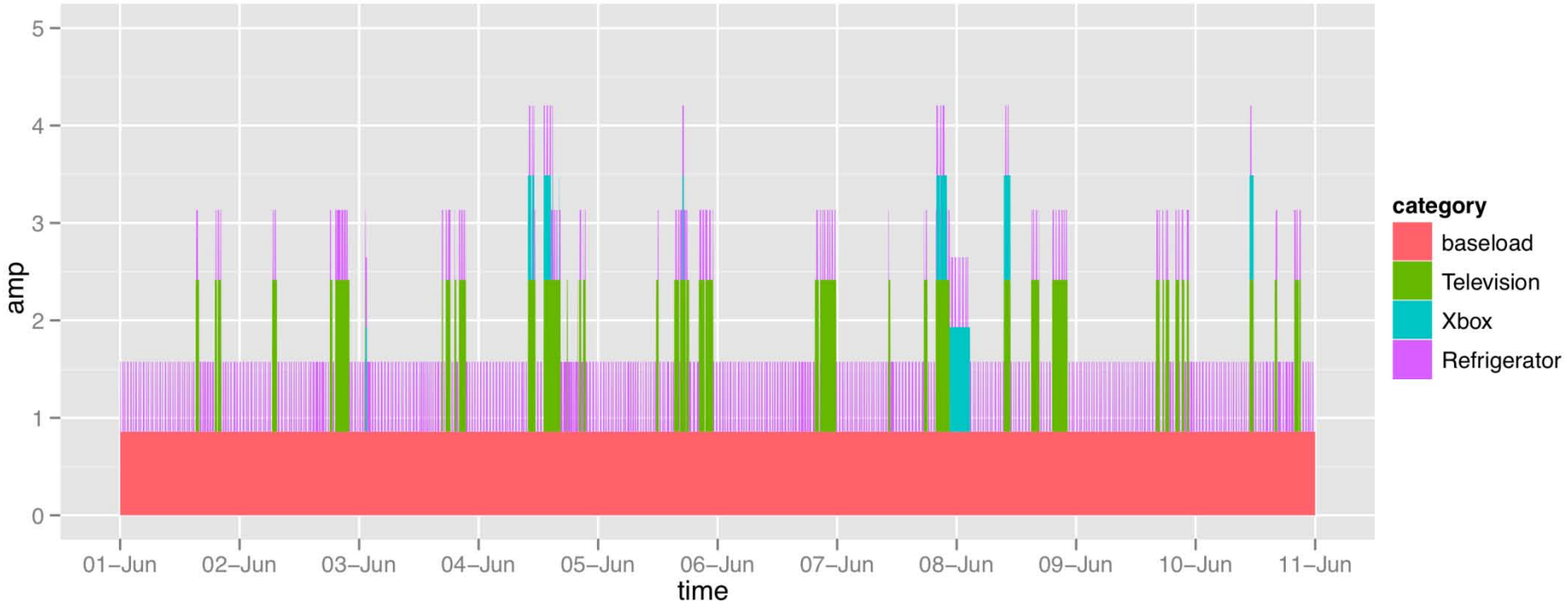
Summary

- Data centers chillers consume substantial power
  - Ensemble of chillers – part of data center cooling infrastructure – are challenging to operate energy efficiently
- Mine and characterize motifs
  - Symbolic representation
  - Event encoding
  - Motif mining
  - Sustainability characterization
- Demonstrated our approach on data from a real data center – indicates significant potential energy savings

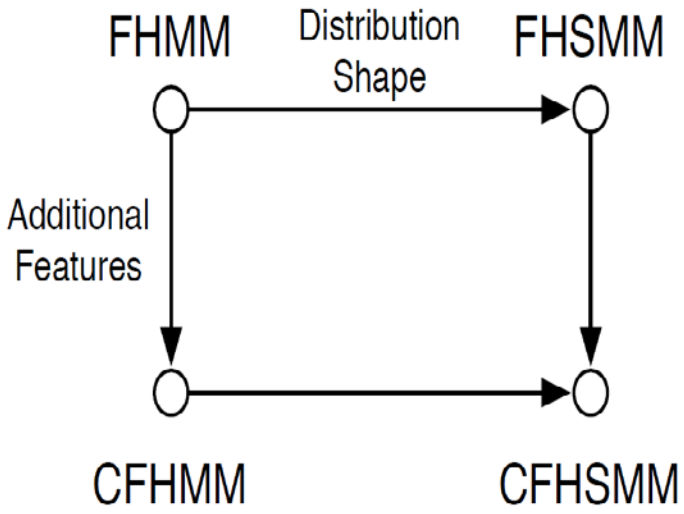
Some other projects

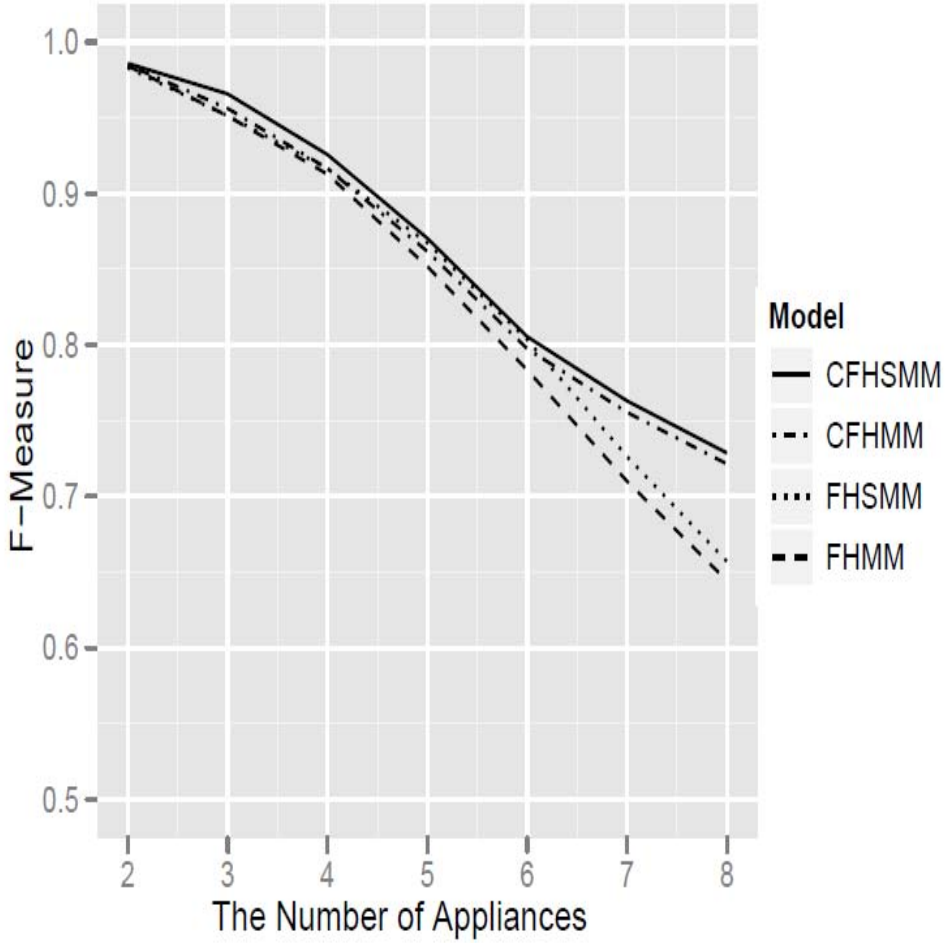
- Anomaly detection (SensorKDD 2010)
- Energy Disaggregation (SDM 2011)
- Automating Life Cycle Assessment (IEEE Computer 2011)
- Fine-grained PV output prediction (AAAI 2012)
- Building Energy Management (BuildSys 2011)

Energy Disaggregation

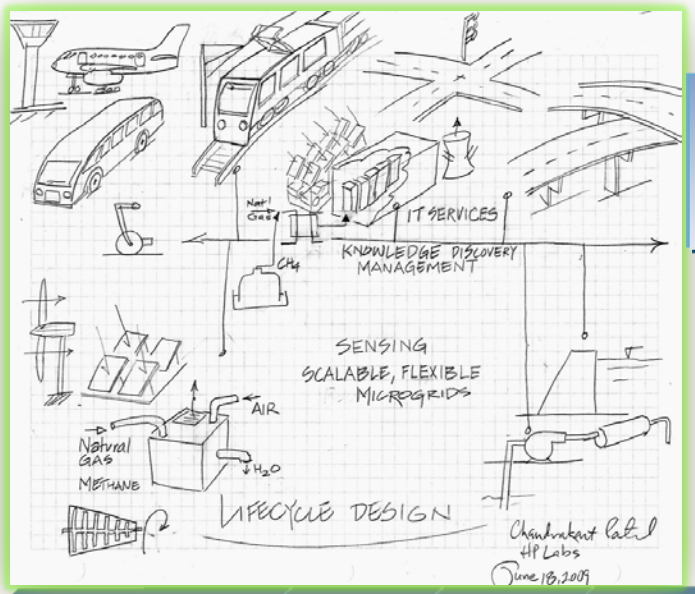


Proposed Variant of Factorial HMM's (SDM 2011)





Data Analytics for Urban Infrastructure



Streaming Data

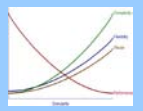
Analytics Engine

Policy
Sustainability Metrics

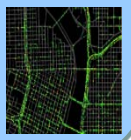
Probabilistic Models

Simulation

Data Fusion



Optimization



Waste | Power | Transport | Water | ...

Actions

References

• P. Chakraborty, M. Marwah, M. Arlitt, and N. Ramakrishnan. Fine-grained Photovoltaic Output Prediction using a Bayesian Ensemble, in *Proceedings of the 26th Conference on Artificial Intelligence (AAAI'12)*, Toronto, Canada, 7 pages, July 2012, To appear.

• Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, C. Hyser, "Renewable and Cooling Aware Workload Management for Sustainable Data Centers", ACM SIGMETRICS/Performance, June 11-15 2012, London, UK, To appear.

• Manish Marwah, Amip Shah, Cullen Bash, Chandrakant Patel, Naren Ramakrishnan, "Using Data Mining to Help Design Sustainable Products," IEEE Computer, August 2011

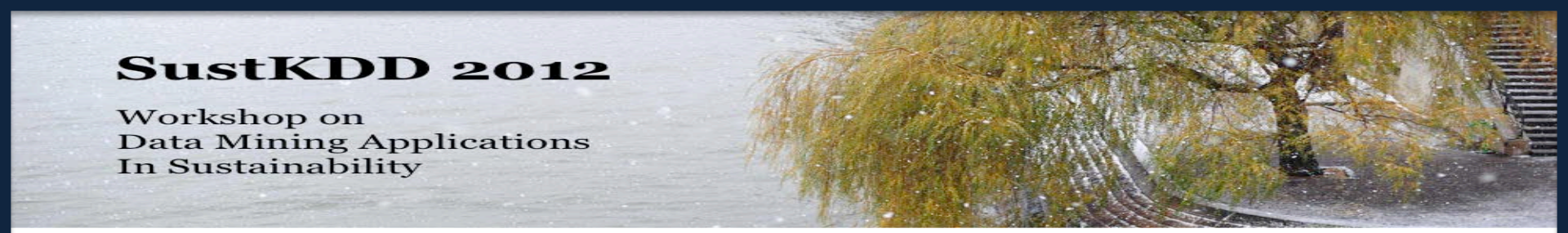
• Hyungsul Kim, Manish Marwah, Martin Arlitt, Geoff Lyon and Jiawei Han, "Unsupervised Disaggregation of Low Frequency Power Measurements", SIAM International Conference on Data Mining (SDM 11), Mesa, Arizona, April 28-30, 2011.

• Gowtham Bellala, Manish Marwah, Martin Arlitt, Geoff Lyon, Cullen Bash, "Towards an understanding of campus-scale power consumption." In ACM BuildSys, November 1, 2011, Seattle, WA.

• Manish Marwah, Ratnesh Sharma, Wilfredo Lugo, Lola Bautista, "Anomalous Thermal Behavior Detection in Data Centers using Hierarchical PCA," in SensorKDD in conjunction with KDD 2010.

• D. Patnaik, M. Marwah, Sharma, Ramakrishna, "Sustainable Operation and Management of Data Center Chillers using Temporal Data Mining," In ACM KDD, June 27 - July 1, 2009, Paris, France.

• Amip Shah, Tom Christian, Chandrakant D. Patel, Cullen Bash, Ratnesh K. Sharma: Assessing ICT's Environmental Impact. IEEE Computer 42(7): 91-93, July 2009.



2nd KDD Workshop on Data Mining Applications In Sustainability

Date: August 12, 2012 Location: Beijing, China

Objective

The goals of this KDD workshop are:

- to bring together researchers working on applications of KDD to sustainability in diverse areas, especially in infrastructures such as IT, Smart Grids, water, and transportation.
- to familiarize the mainstream KDD community with diverse application areas within sustainability.
- to serve as a meeting ground and launchpad to galvanize and foster the development of this budding sub-community.

Organizing Committee Chairs

- Naren Ramakrishnan, Virginia Tech (co-chair)
- Manish Marwah, HP Labs (co-chair)
- Mario Berges, CMU (co-chair)
- Zico Kolter, MIT (co-chair)

Paper Submission

Two types of papers in ACM SIGKDD format are encouraged: long papers with a maximum of 8 pages describing completed work on data mining problems in sustainability and short papers of 4-6 pages describing ongoing research or preliminary results. We also invite a 1-2 pages extended abstract for early-stage work to be presented as posters.

Important Dates:

Submission: May 23, 2012
Notification: June 4, 2012
Camera-ready Versions: June 8, 2012
Workshop: August 12, 2012

For More Information:
http://marioberges.com/SustKDD12