Data Mining for Internet of Things: A Survey

Cristian Consonni

Data Mining for Internet of Things: A Survey

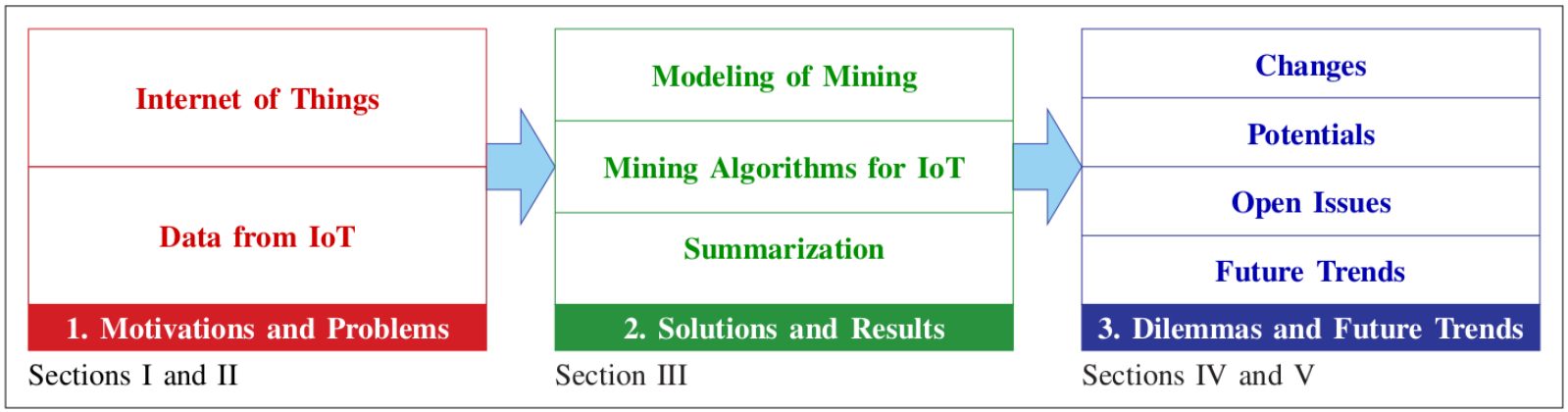
C.-W. Tsai, C.-F. Lai, M.-C. Chang, and L. T. Yang

Communications Surveys & Tutorials, IEEE 16.1 (2014): 77-97.

url: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6674155.

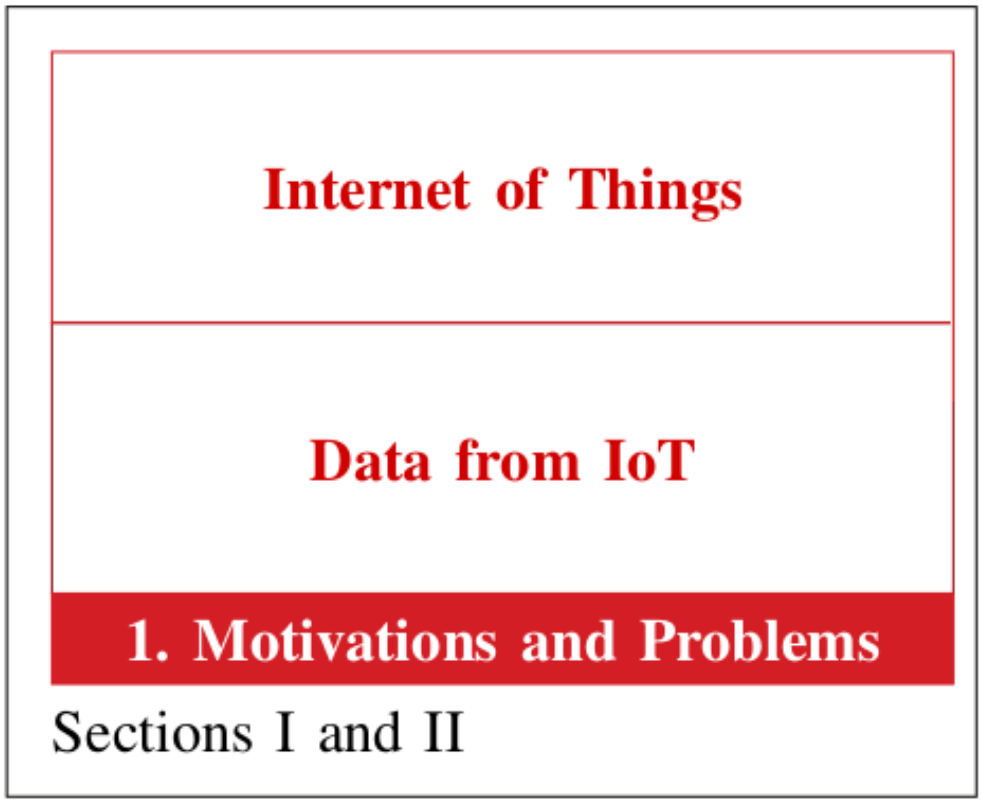
Outline

Roadmap


Internet of Things
Data from IoT
1. Motivations and Problems
Sections I and II
Modeling of Mining
Mining Algorithms for IoT
Summarization
2. Solutions and Results
Section III
Changes
Potentials
Open Issues
Future Trends
3. Dilemmas and Future Trends
Sections IV and V

Outline for section 1

Motivation and Problems



Motivation

1. What is the **Internet of Things**
2. Internet of Things $\rightarrow$ Big Data (data from IoT, data for IoT)
3. we need mining technologies
4. Data generated or captured by IoT are considered having highly useful and valuable information
5. Changes, Potentials, Open Issues, and Future Trends

Internet of Things: Definition (I)

Definition

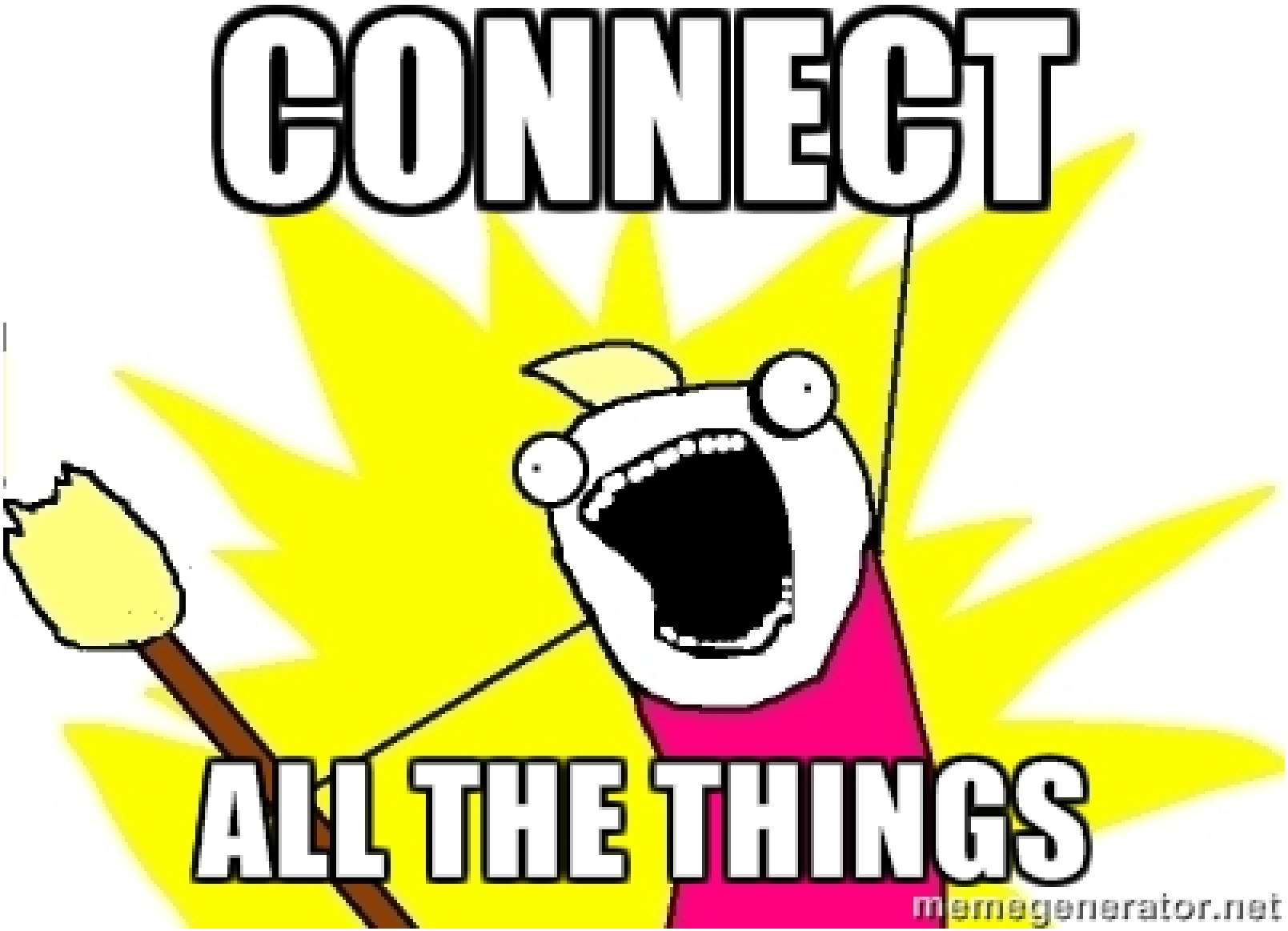
Technology for **seamlessly** integrating classical networks and networked objects

Internet of Things: Motivation and Context

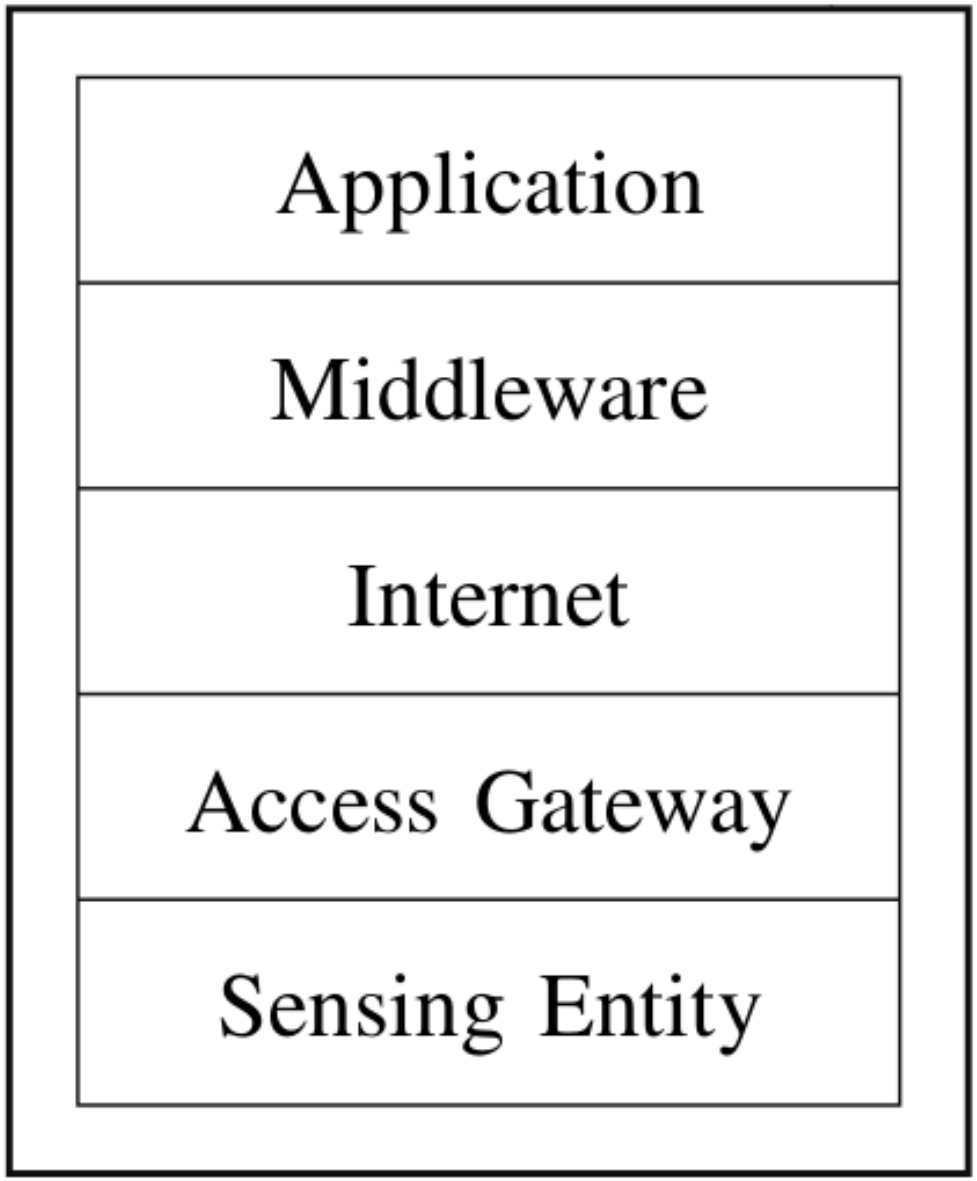
Internet of Everything:

url: `http://www.cisco.com/web/about/ac79/docs/innov/IoE-Value-Index_External.pdf`

Internet of Things: Definition (II)



Internet of Things: Architecture



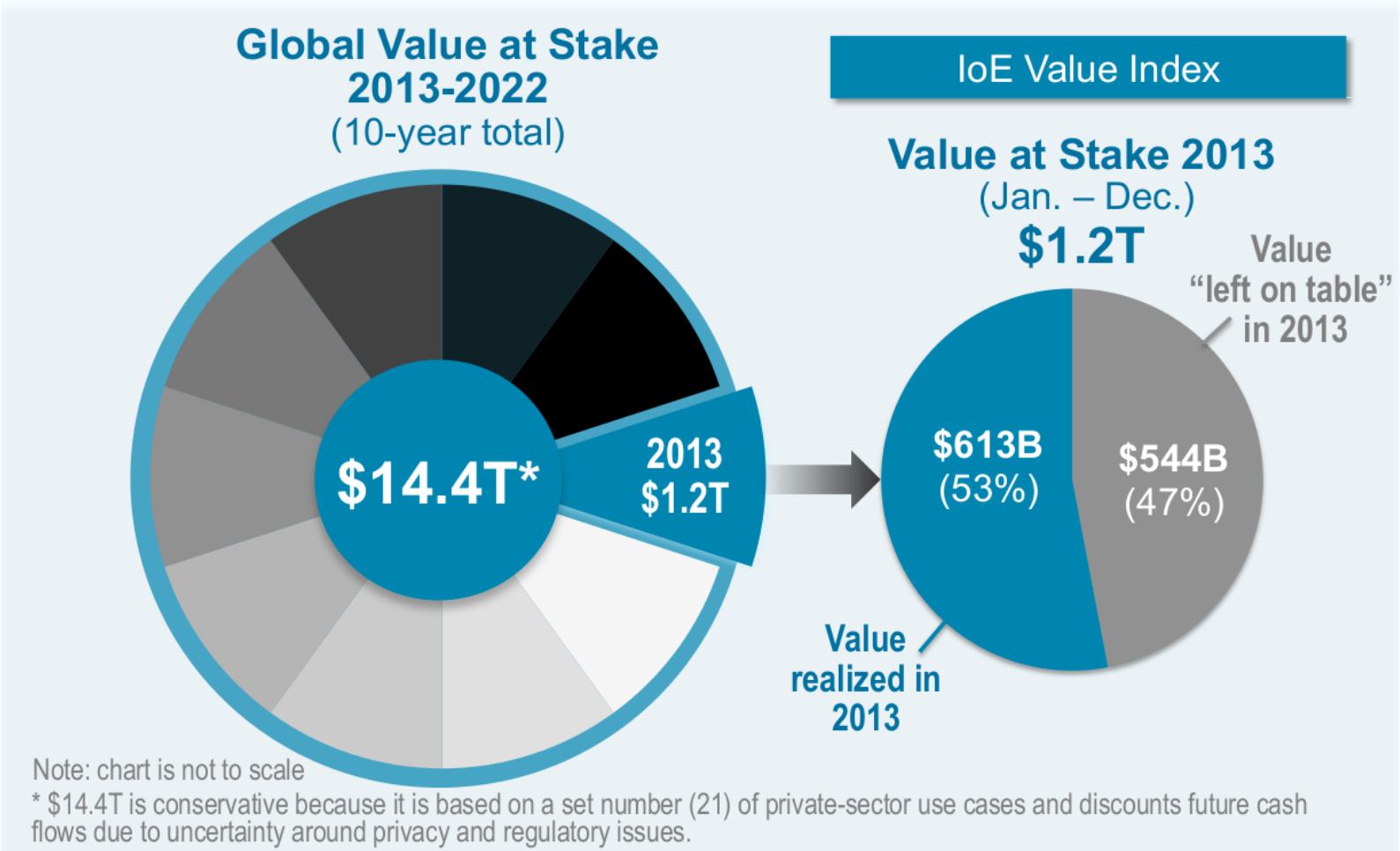
Internet of Things: Characteristics

Things in IoT are "supposed to have intelligence":

- Capable of being identified
- Capable of sensing events
- Capable of interacting with humans, other systems or the environment
- Capable of making decisions by themselves

Internet of Things: Motivation (II)

High economic value 44B USD (2011) $\rightarrow$ 290B USD (2017)



url: `http://www.cisco.com/web/about/ac79/docs/innov/IoE-Value-Index_External.pdf`

Internet of Things: Problem Statement

Problem

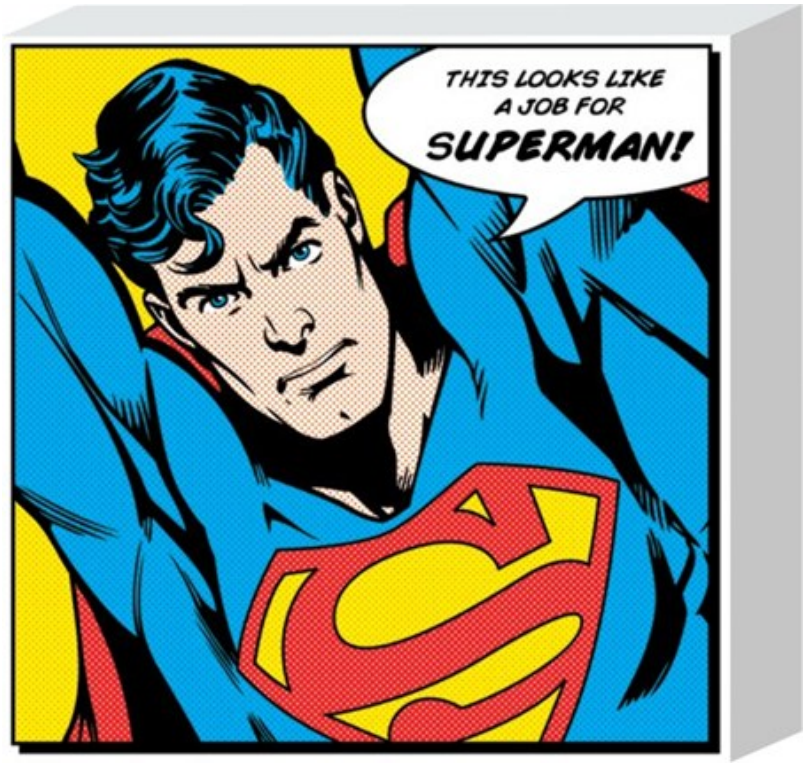
How do we trasform data collected by IoT in knowledge for people?

Big Data and IoT Data are interesting for companies because they provide a **competitive advantage**

Internet of Things: Problem Statement

Solution

With **knowledge Discovery in Databases** (KDD) and **Data Mining**



Internet of Things: Motivation and Context (III)

We want to find out **hidden information in the data of IoT** to:

- enhance the performance of the system
- improve the quality of the services provided

Outline for section 2

Data Deluge From IoT

⇒ Bottleneck of data processing shift from sensors to data prerocessing, communication, storage

A Taxonomy for Data From IoT

Simple taxonomy for **data from IoT**:

- *"data about things"*: data that describe the things themselves (state, location, identity, ...)
  - $\rightarrow$ can be used to optimize the performance of the structure/system
- *"data generated by things"*: data captured or sensed by the things as the result of the interaction between humans, things and humans, things and things, ...
  - $\rightarrow$ can be used to enhance the service provided by IoT

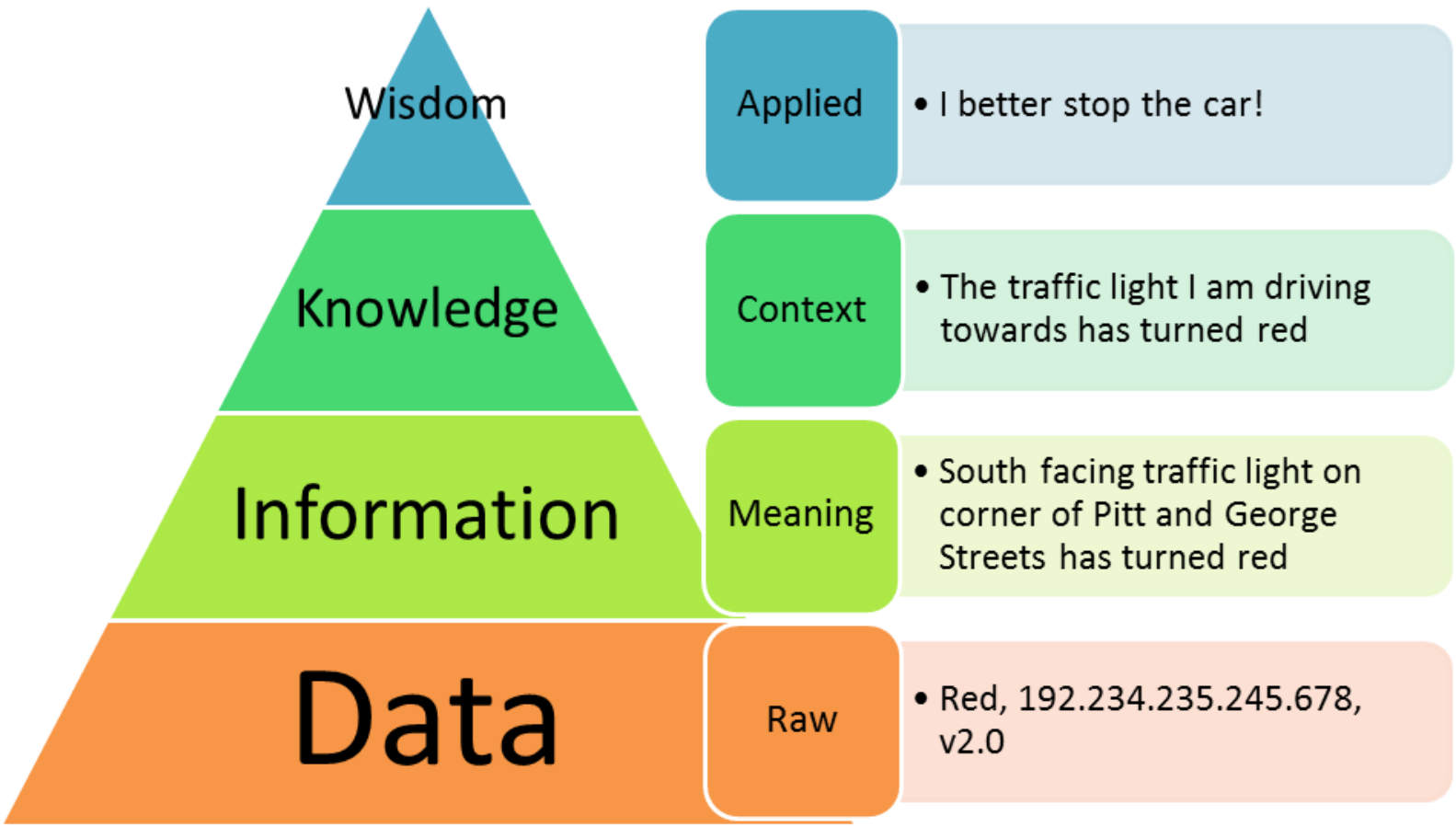
How To Face the Data Deluge

There are several approaches to handle Big Data from IoT:

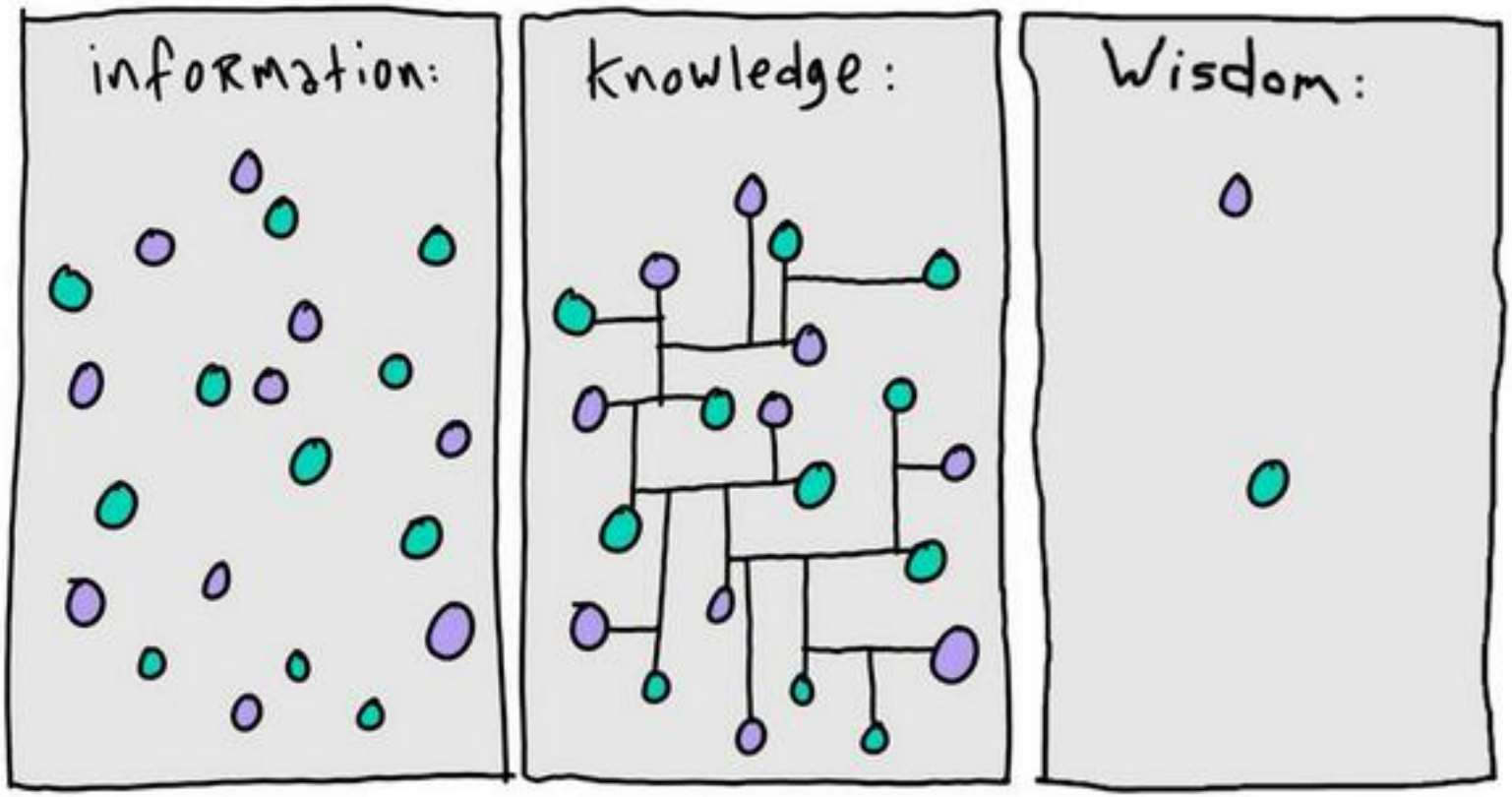
- **Reduce the size of data**: random sampling, acquire only the interesting data (pre-processing on the sensor using PCA or other dimensionality reduction), data condensation, feature selection
- **Reduce the computational needs**: divide and conquer, incremental learning
- **Scale Out Computing Power**: distributed computing, cloud computing

Decision Making Chain

Goal of Data Mining: Wisdom Knowledge Information (I)



Goal of Data Mining: Wisdom Knowledge Information (II)



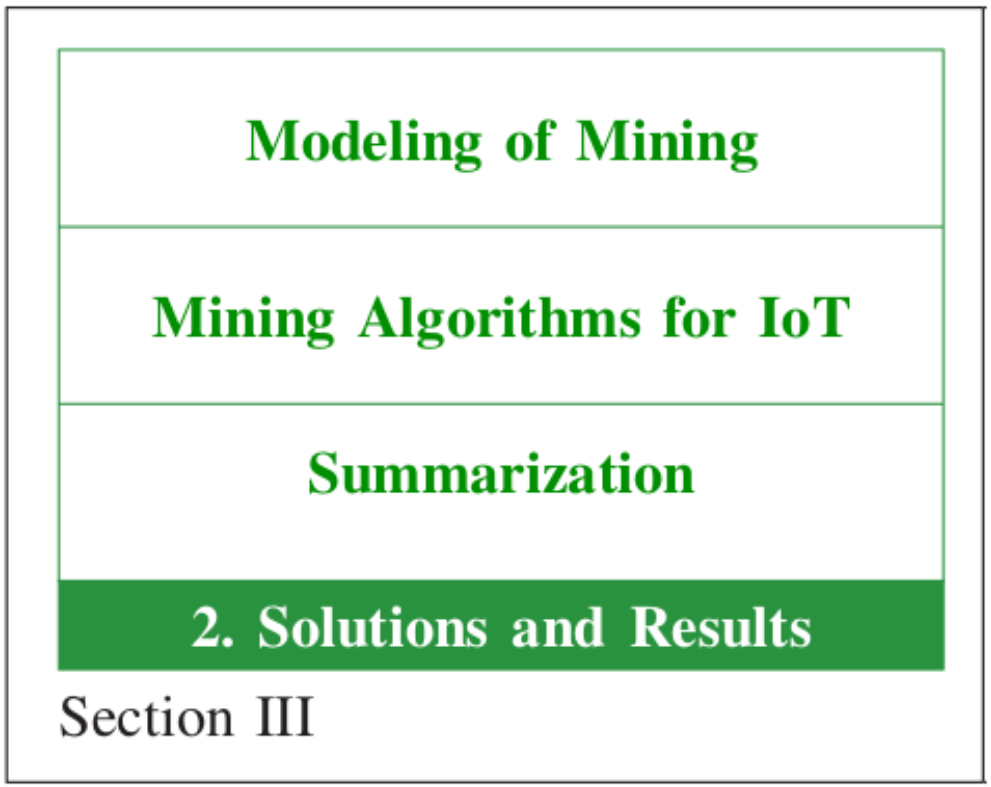
Caveats about Data Mining for Knowledge Discovery

- data fusion, large scale data, data transmission, and decentralized computing issues may have a stronger impact on the system 0performance than KDD or data mining algorithms
- more data win over better algorithms:
  1. "More data usually beats better algorithms" (`http://bit.ly/datawocky-1`)
  2. "More data usually beats better algorithms, Part 2" (`http://bit.ly/datawocky-2`)
  3. "More data beats better algorithm at predicting Google earnings" (`http://bit.ly/datawocky-3`)

  → by Stanford University professor Anand Rajaraman

Outline for section 3

Solutions and Results



Problem Formalization

Most mining technologies are designed to work on a single system, traditional algorithm cannot be applied to Big Data.
Considerations when choosing the applicable mining technology:

- **Objective** (O): problem definition, assumption, limitations
- **Data** (D): characteristics of data (size, distribution, representation)
- **Mining algorithm** (A)

$\rightarrow$ wireless sensor network have to count for the load of clustering algorithm, which is usually ignored

Unified Data Mining Framework

Algorithm 1 Unified Data Mining Framework

1	Input data D	
2	Initialize candidate solutions r	
3	**While** the termination criterion is not met	
4	$d = Scan(D)$ [Optional]	**S**
5	$v = Construct(d, r, o)$	**C**
6	$r = Update(v)$	**U**
7	**End**	
8	Output rules r	

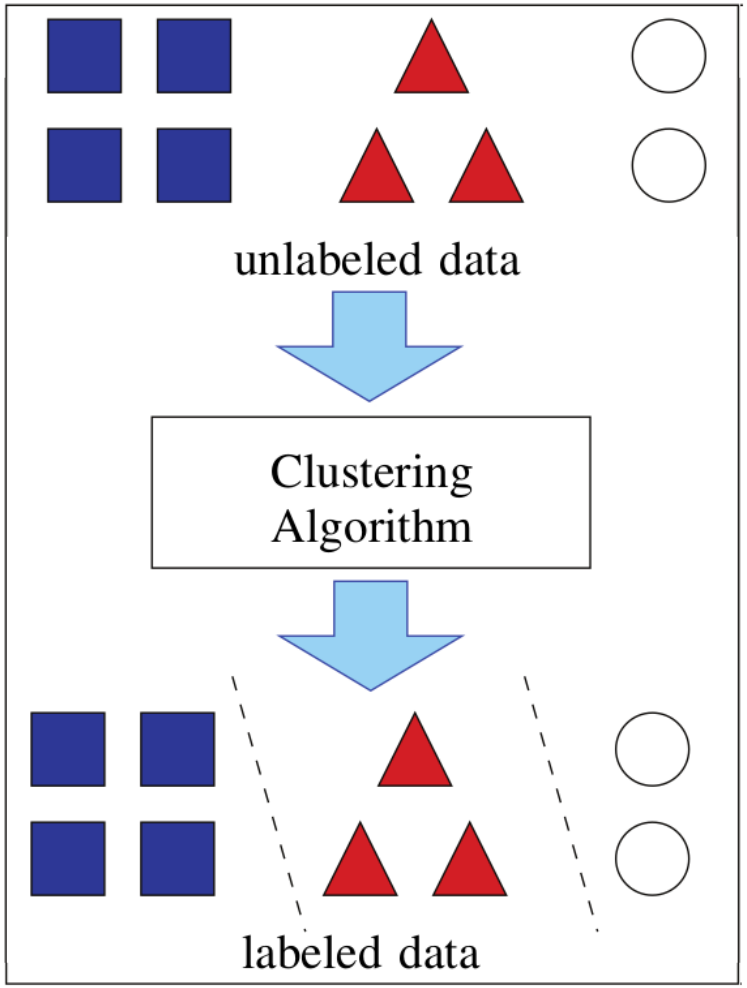
→ this framework can be used to describe metaheuristic algorithms

Clustering

Characteristics of Clustering

- Clustering is an *unsupervised learning* algorithm:
  - *input*: unlabeled patterns $X = \{x_1, x_2, ...x_n\}$ (d-dimensional space);
  - *goal*: k clusters $\Pi = \{\pi_1, \pi_2, ..., \pi_k\}$ based on a similarity metric;
  - *output*: set of k centroids $C = \{c_1, c_2, ..., c_k\}$;
- Quality can be measured using SSE or PSNR (it depends on the application)

Depiction of Clustering



k-means

k-means examples:

1. `http://bit.ly/kmeans-1`
2. `http://bit.ly/kmeans-2`

Clustering for Infrastructure of IoT

Clustering can be applied to WSN to:

- enhance the performance of an IoT system on the integration of identification, sensing and actuation;
- reduce energy consumption (LEACH algorithm);
- speed-up data transmission (TPC algorithm);
- speed-up information exchange between nodes (dynamic clustering);

Clustering for Services of IoT

Clustering can be applied to services provided by IoT to help devices make decision by themselves or provide better services.

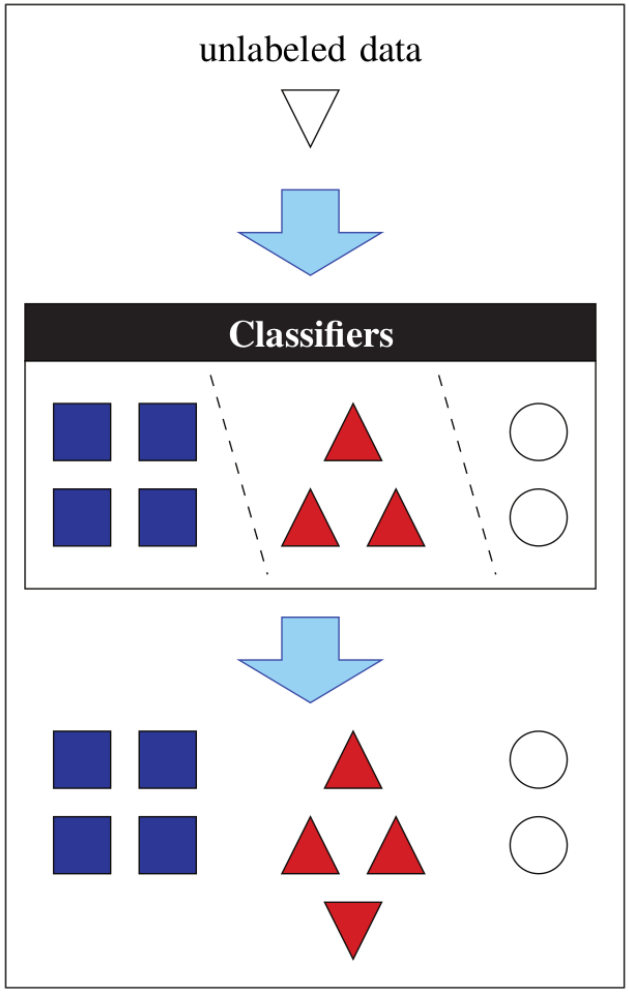
- smart home mining frequent patterns, tracking user behavior in a smart home;
- improve location tracking from RFIDs (DBSCAN algorithm);
- improve agricultural productivity;
- studies on social networks using smartphones, personal digital assistants, ecc.;

Classification

Characteristics of Classification

- Clustering is an *supervised learning* algorithm:
  - *input*: set of labeled data L and unlabeled daata U;
  - *output*: Labeled data are used to train a set of classifiers;
  - *goal*: label unlabeled data;
- Quality can be measured using accuracy Rate (AR), Precision (P), recal; (R) and F-measure (F)

Depiction of Classification (II)



Classification Algorithms

Several algorithms can be used as classifiers

- Decision trees:
- k-nearest neighbor:
- Naive Bayes classification:
- adaboost
- Support Vector Machine (SVM) (note: non-linear):

Classification for Infrastructure of Iot

Classification algorithms are applied to the problem of *identification* of devices:

- Unique Item Identifier (UII):
- Several standards have been idenpendently developed:
- Iots of UII queries (more than current DNS queries):

Classification for Infrastructure of Iot

Classification can be applied to:

- *outdoor* tasks:
  - traffic jam problem: traffic forecasting services, accident detection, detecting traffic information from social netwrks;
  - parking problem: passive infra-red sensor;
- *indoor* tasks:
  - healtcare and smart home systems:
    - infrared presence sensors, microphones, wearable kinematic sensors
    - definition of activities: bathing, dressing, etc.
    - detect posture and falling event
    - avoid suddend death events
    - replace ECG in hospital with home systems
  - other tasks
    - object identification
    - face recognition
    - facial expression classification

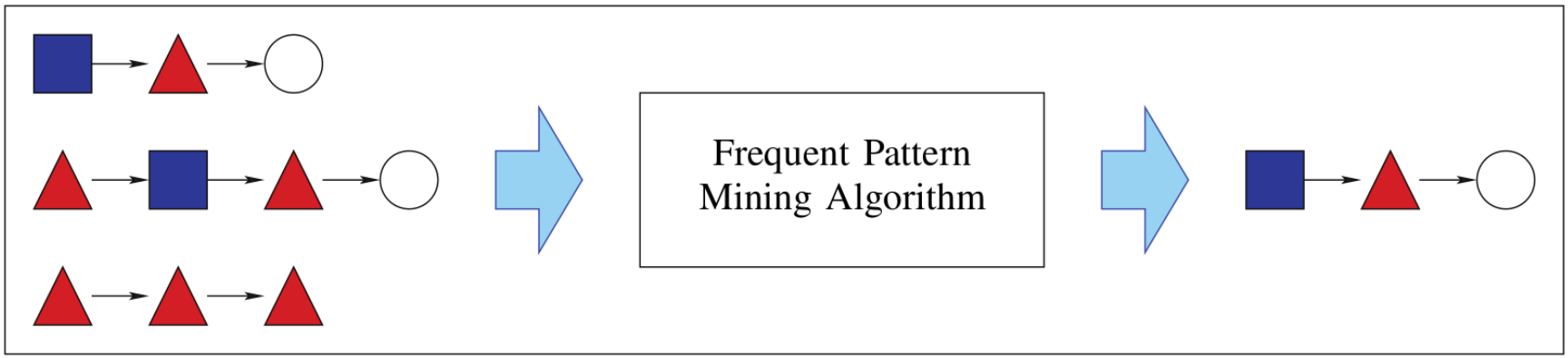
Note: especially indoor system raise important privacy concerns.

Frequent Pattern Mining

Characteristics of Frequent Pattern Mining

- Frequent Pattern Mining is an *unsupervised learning* algorithm:
  - *input*: set of sequeces;
  - *output*: most common sequences;
  - *goal*: find "interesting patterns" from a large set of items;
- we introduce the notion of *support* and *confidence*

Depiction of Frequent Pattern Mining



Frequent Pattern Mining for Infrastructure of IoT

Several application of Frequent Pattern Mining for Infrastructure of IoT:

- temporal data mining, privacy preserving on IoT, web usage mining, bionformatics, intrusion detection:
- 90% of the data crreated today is born in electronic form
- manage RFID events, spatial collocation mining in GIS

Frequent Pattern for Services of IoT

Frequent Pattern Mining can be used to provide better services to IoT users:

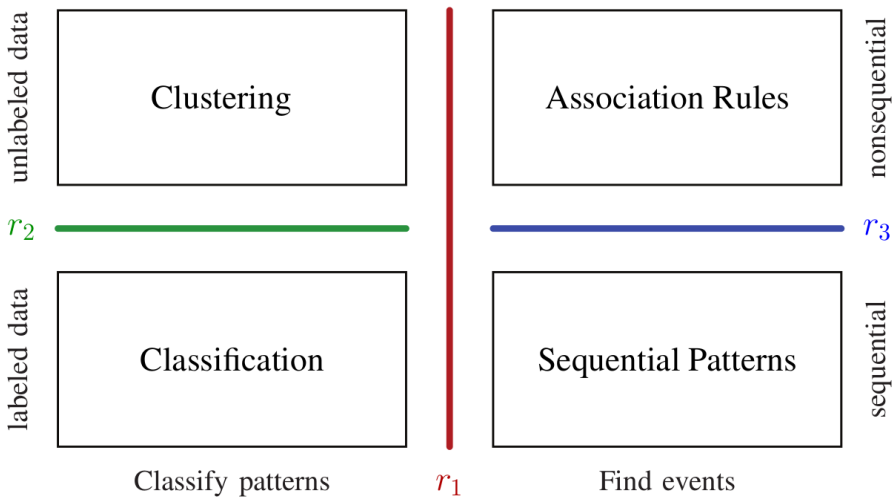
- analyze purchase behavior is classic but still active:
- help customer find products they are looking for quickly (customer specific rules, category-based rules, association rules)

Summary of Mining Technologies

Overall Comparison

Mining Algorithm	Goal	Data Source	References
Clustering	Network performance enhancement	wireless sensor	[55], [105], [56], [57], [58], [51], [106], [59], [60], [61]
	Inhabitant action prediction	X10 lamp and home appliances	[107]
	Provisioning of the needed services	Raw location tracking data	[63]
	Housekeeping	Vacuum sensor	[108]
	Managing the plant zones	GPS and sensor for agriculture	[64], [65]
	Relationships in a social network	RFID, smart phone, PDA, and so on	[66], [67], [68]
Classification	Device recognition	RFID	[74]
	Traffic event detection	GPS, smart phone, and vehicle sensor	[75], [76], [77]
	Parking lot management	Passive infrared sensor	[78]
	Inhabitant action prediction	RFID, sensor, video camera, microphone, wearable kinematic sensor, and so on	[79], [109], [110], [80], [81], [82], [89] [83], [84]
	Inhabitant action prediction	Video camera	[87], [88]
	Inhabitant action prediction	microphone	[90]
	physiology signal analysis	wireless ECG sensor	[91], [92]
Frequent Pattern	RFID tag management	RFID	[100], [101]
	Spatial colocation pattern analysis	GPS and sensor	[102], [103], [104]
	Purchase behavior analysis	RFID and sensor	[111]
	Inhabitant action prediction	RFID and sensor	[112], [113], [114], [115], [116], [117]
Hybrid	Inhabitant action prediction	RFID and sensor	[62], [118], [119], [120], [121], [122]

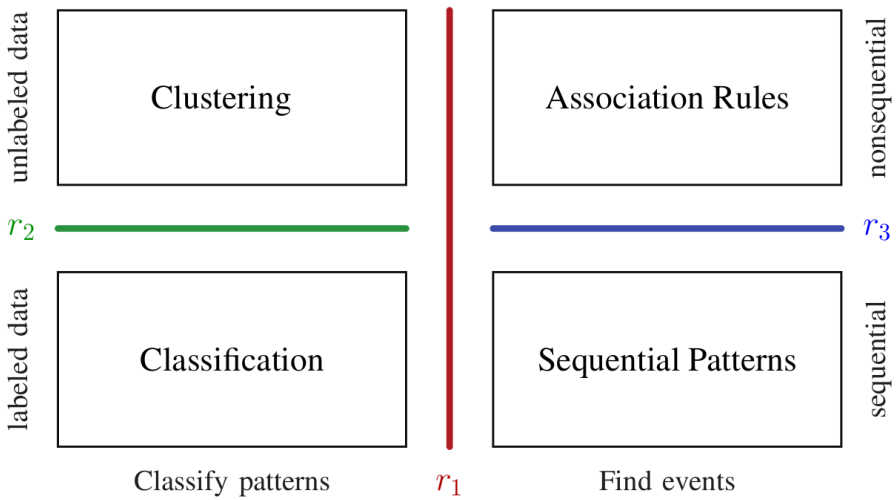
Technology Comparison



Three rules:

- r_1: divide or classify patterns vs discovering patterns
- r_2: clustering vs classification
- r_3: associative rules vs sequential patterns

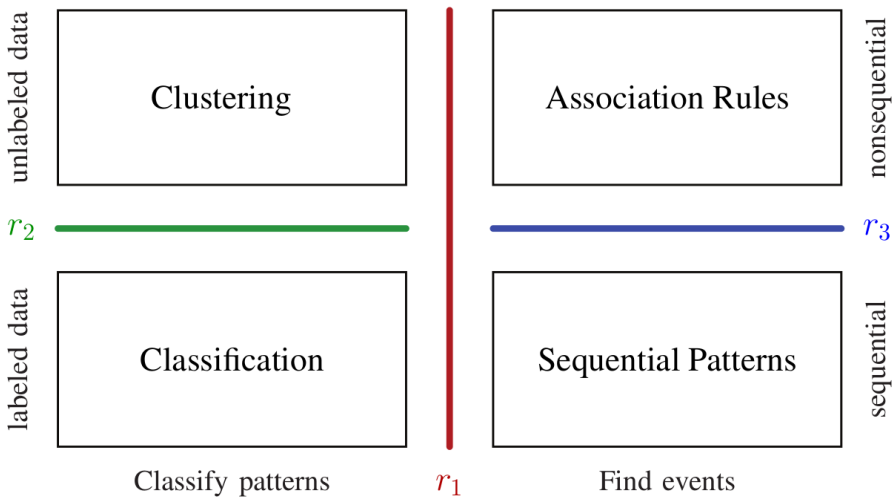
Technology Comparison



Three rules:

- r_1: divide or classify patterns vs discovering patterns
- r_2: clustering vs classification
- r_3: associative rules vs sequential patterns

Technology Comparison



Three rules:

- r_1: divide or classify patterns vs discovering patterns
- r_2: clustering vs classification
- r_3: associative rules vs sequential patterns

How to Choose your Mining Technology (I)

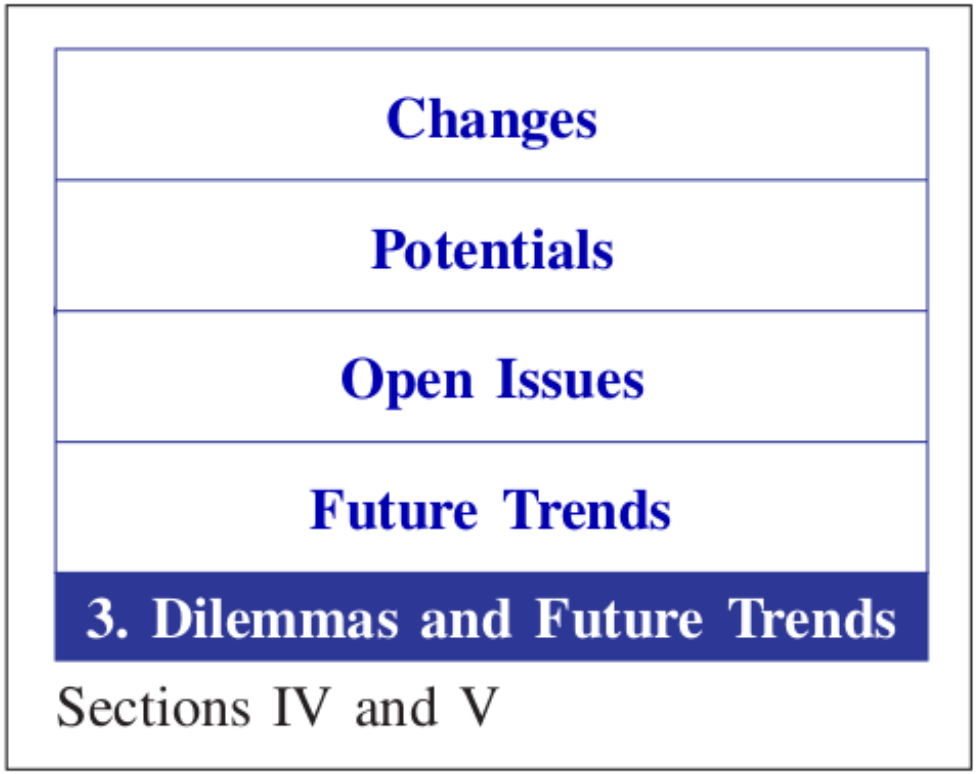
- Based on use case:
  - *clustering*: classify patterns all of which are unlabeled
  - *classification*: classify patterns some of which are unlabeled and some labeled
  - *association rules*: find events in no particular order
  - *sequential patterns*: find events in some particular order

How to Choose your Mining Technology (II)

- Combination of technologies:
  - clustering $\rightarrow$ classification: unsupervised classification system, clustering creates a set of classifiers then classify incoming patterns
  - classification $\rightarrow$ clustering: semi-supervised learning system or incremental classification system: classifiers from labeled patterns, clustering is used to incrementally add new classifiers enable handling of new patterns
  - clustering $\rightarrow$ classification: unsupervised classification system, clustering creates a set of classifiers then classify incoming patterns
  - clustering $\rightarrow$ classification $\rightarrow$ frequent pattern, classification $\rightarrow$ clustering $\rightarrow$ frequent pattern clustering creates a set of classifiers then classify incoming patterns
  - iterative systems where a solution is applied several times

Outline for section 4

Dilemmas and Future Trends



Changes Caused by IoT

Changes Caused by IoT

Three kinds of change:

1. *thing*-oriented: new devices
2. *internet*-oriented: network stress issues
3. *semantics*-oriented: definition of smart object

Potentials of Using IoT

Potentials of Using IoT

IoT can be applied to several things

1. *people*-oriented: recommendations, support for decision making
2. *self*-oriented: enhance performace, automatic filtering of redundant data
3. *things*-oriented: enhance machine-2-machine exchange of data

Open Issues of IoT

Infrastructure Perspective

Two main issues:

1. decentralization
2. heterogeneity

Data Perspective

Three main issues:

1. velocity
2. variety
3. velocity

$\rightarrow$ determiny what needs to remain on the sensor is hard

Algorithm Perspective

Two main issues:

1. Flexibility
2. Dynamicity

* additional issue of how to combine mining technologies, becaise we need smarter systems

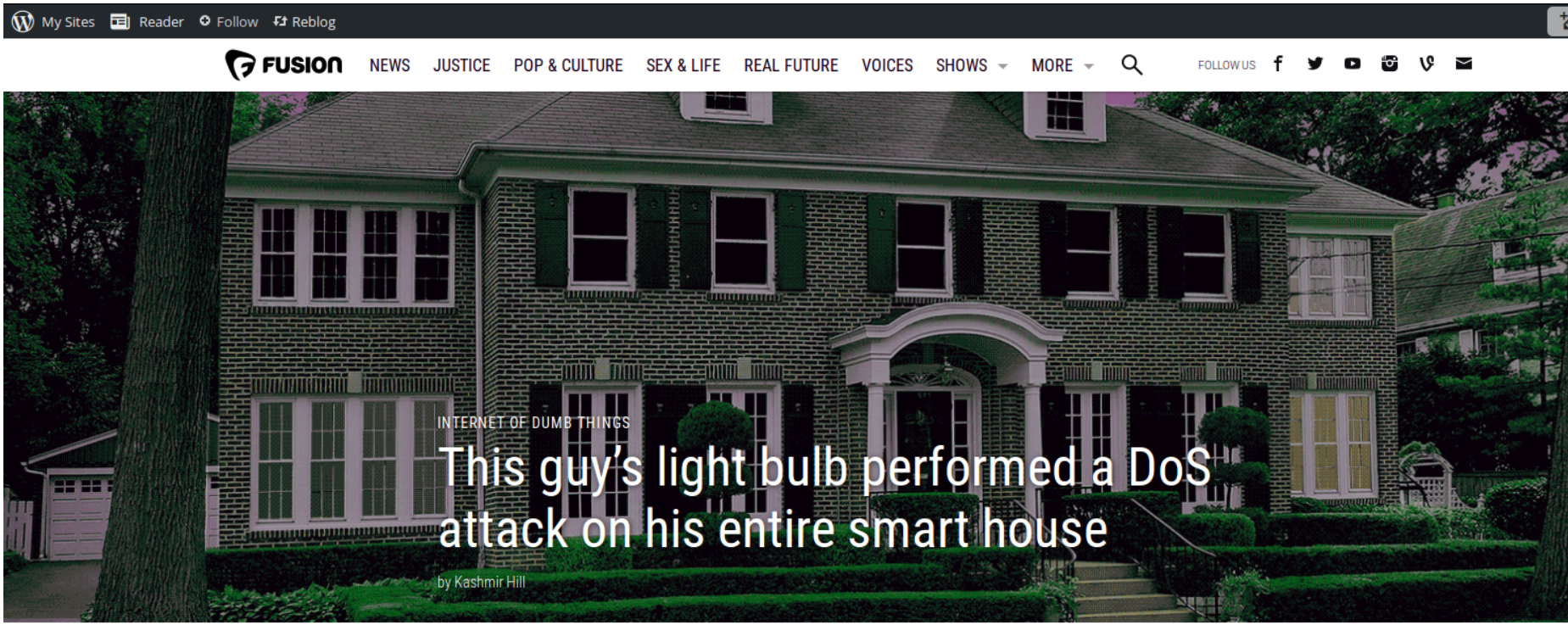
Open Issues on Privacy and Security

ioT introducces many challenges about privacy and secutity:

1. sensors collect data that can be sensitive (e.g. video images, heealth data)
2. concentration of data in big companies
3. how to protect access to sensors and data

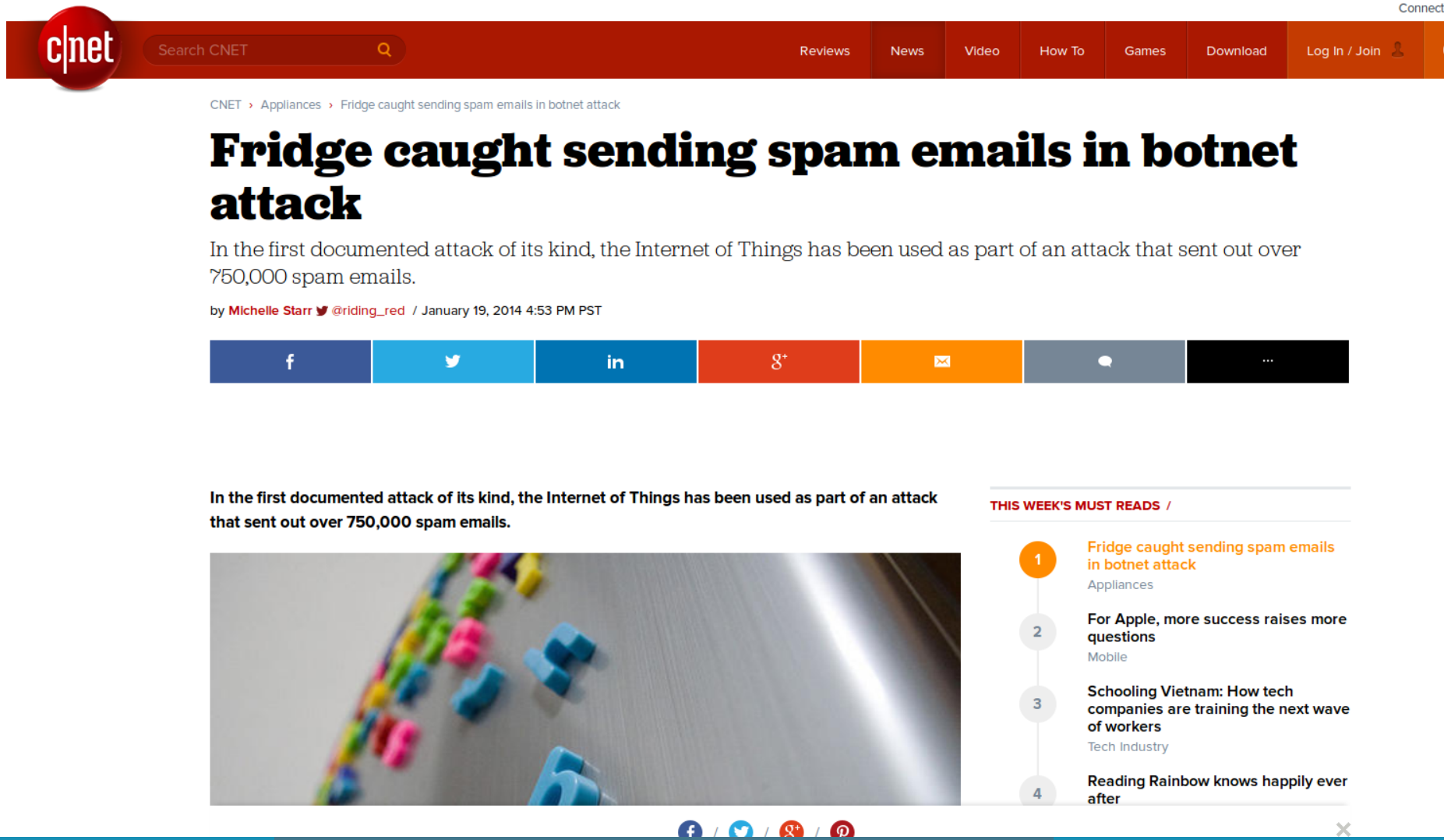
Open Issues on Privacy and Security: Examples (I)

`http://internetofshish.tumblr.com/`



Open Issues on Privacy and Security: Examples (II)

http://internetofshish.tumblr.com/



Open Issues on Privacy and Security: Examples (III)

http://internetofshish.tumblr.com/

Cristian Consonni ▾ subscribe search jobs dating more ▾ UK edition ▾

theguardian

UK world politics sport football opinion culture business lifestyle fashion environment tech travel browse all sections

home › world europe US americas asia australia africa middle east cities development

South Korea

South Korean woman's hair 'eaten' by robot vacuum cleaner as she slept

The woman was sleeping on the floor of her home when the robotic cleaner ingested her hair leaving her in agony

Justin McCurry in Tokyo

Monday 9 February 2015 04.53 GMT

Shares 138,050 Comments 340

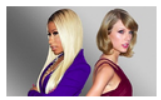
Article saved

Firefighters try to rescue a woman at her house in Changwon, southeast South Korea after her hair was sucked into a robot vacuum cleaner. She lost about 10 strands of hair but was not injured. Photograph: Yonhap/AAPIMAGE

When a South Korean woman invested in a robot vacuum cleaner, the idea was to leave her trustworthy gadget to do its work while she took a break from

Most popular

Taylor Swift v Nicki Minaj row: Katy Perry joins in... and brings Rihanna with her



The Nicki Minaj debate is bigger than Taylor Swift's ego

Chelsea stunned 4-2 by New York Red Bulls - as it happened

Teenager finds photos of first Nirvana gig tucked inside her dad's books

Outline for section 5

Conclusions

1. IoT and Big Data are trend of the future years;
2. Ontology and semantics can help solve the issue of dealing with Big Data;
3. IoT, Big Data, Smart Grids, Cloud computing will impact our life at the same time $\Rightarrow$ we need to approach them together;

Thank you

Thank you!

Backup

Activities and Tasks

Example of activities and tasks:

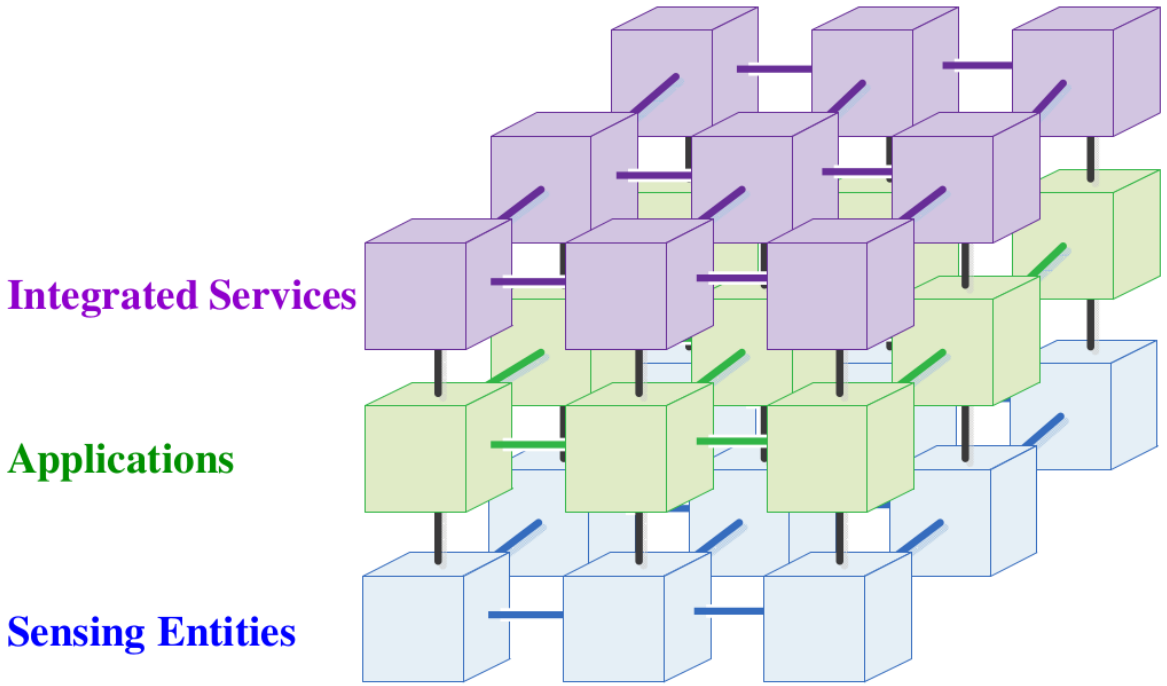
- tacking human health
- smart home
- tracking human activities
- smart home vacuum system
- smart city (power system, public transportation, security, emergency, ecc.)
- plug-in hybrid electric vehicle

Activities and Tasks

Example of activities and tasks:

- tacking human health
- smart home
- tracking human activities
- smart home vacuum system
- smart city (power system, public transportation, security, emergency, ecc.)
- plug-in hybrid electric vehicle

Smart City HyperCube



Credits

- Superman image:
  `https://dgeiu3fz282x5.cloudfront.net/g/l/lgCV95054.jpg`