Research Topics in Information Systems and Services

Lipyeow Lim
Assistant Professor
Information and Computer Sciences

Agenda

- Information Systems & Services Focus Area
  - Overview, Faculty & Resources
  - Secondary Exam
  - Primary Exam
- Research topics
  - Optimizing Content Freshness of Relations Extracted From the Web Using Keyword Search
  - Mining Workflows for Data Integration Patterns
  - Optimizing Sensor Data Acquisition for Energy-Efficient Smartphone-based Continuous Event Processing

Information Systems & Services

*All aspects of organizing, retaining and retrieving **information***

- Storage and indexing methods
  - Storage technologies and storage format
  - Indexing structures and algorithms
- Information retrieval and search strategies
  - Search and query models for structured, semi-structured, unstructured data
  - Citation indexes and impact factors
  - Relevance ranking algorithms
- Personalized information systems
  - User modeling and user profiles
  - Information filtering systems.

Faculty and Resources

- http://discourse.ics.hawaii.edu/workspace/144/note/547
- Peter Jasco's Topics
  - http://www2.hawaii.edu/~jacso/iss-faq/
- Luz Quiroga's Topics
  - http://www2.hawaii.edu/~lquiroga/service/ISRsecAreasLMQ.htm
  - http://www2.hawaii.edu/~lquiroga/service/ISRprimAreasLMQ.htm
- Lipyeow Lim's Topics
  - http://www2.hawaii.edu/~lipyeow/cisiss.html

Secondary Exam

- **Data storage**
  - storage devices (disk, flash etc), data structures and organization.
- **Data models**
  - relational, XML, RDF, unstructured text, multi-media.
- **Query languages**
  - Keyword search, SQL, SPARQL, XPath, XQuery, XSLT, streamSQL.
- **Indexing**
  - B+tree, inverted indexes, R-trees, XML indexes, RDF indexes
- **Query optimization**
  - access path selection, statistics, cost models, plan enumeration, search space pruning.

Courses: ICS 321 Data Storage and Retrieval

Primary Exam

- **Distributed/parallel data management**
  - Parallel database paradigm, Map-Reduce paradigm, parallel programming paradigms (eg. MPI)
  - Consistency requirements of distributed data processing.
- **Cloud computing and data management**
  - Virtualization technology
- **Scientific data management**
  - cf traditional business data
  - Challenges posed by scientific applications in astronomy, bioinformatics, genomics, environmental sensors etc
- **Semantic web technologies**
  - Leveraging RDF and OWL for data processing, data integration, and knowledge management
  - Challenges and opportunities

Courses : ICS421 Database Systems & ICS624 Advanced Data Management

Data Management

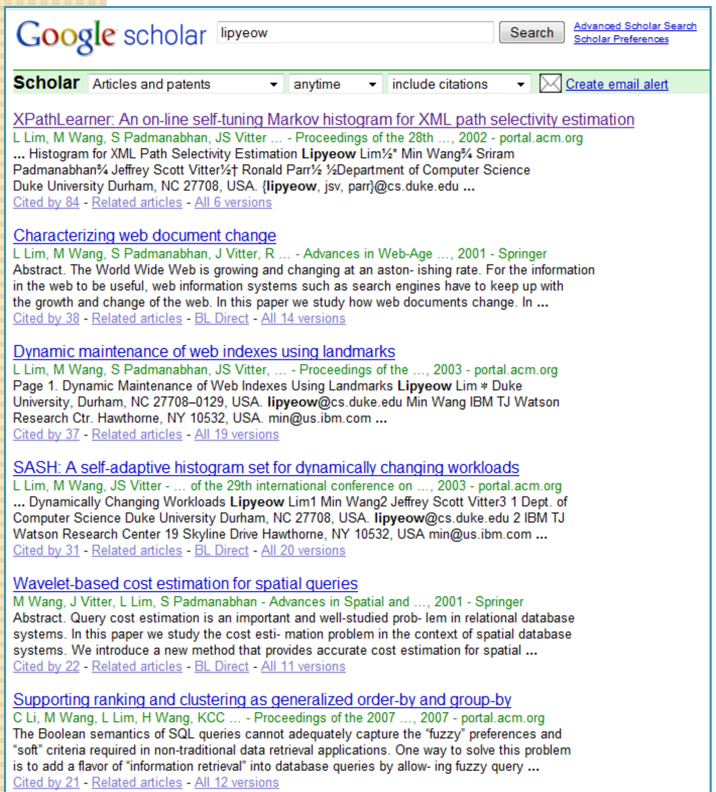
- Motivation:
  - http://www.youtube.com/watch?v=EWL312zbEKg
  - http://www.youtube.com/watch?v=DsQ9UxVALSs
  - http://www.youtube.com/watch?v=jbkSRLYSojo
- Core questions:
  - What is the best way to store data ?
  - How do we query and/or update the data ?
  - How to speed up queries ?
- Research Drivers:
  - New applications.
  - New data types. New query types

Optimizing Content Freshness of Relations Extracted From the Web Using Keyword Search

Mohan Yang (Shanghai Jiao Tong University),
Haixun Wang (Microsoft Research Asia),
Lipyeow Lim (UHM)
Min Wang (HP Labs China)

Motivating Application

- Management at a prominent research institute wanted to analyze the impact of the publications of its researchers ...


Google scholar lipyeow Search Advanced Scholar Search Scholar Preferences

Scholar Articles and patents anytime include citations Create email alert

XPathLearner: An on-line self-tuning Markov histogram for XML path selectivity estimation
L Lim, M Wang, S Padmanabhan, JS Vitter ... - Proceedings of the 28th ..., 2002 - portal.acm.org
... Histogram for XML Path Selectivity Estimation **Lipyeow** Lim½* Min Wang¾ Sriram Padmanabhan¾ Jeffrey Scott Vitter½† Ronald Parr½ ½Department of Computer Science Duke University Durham, NC 27708, USA. {**lipyeow**, jsv, parr}@cs.duke.edu ...
Cited by 84 - Related articles - All 6 versions

Characterizing web document change
L Lim, M Wang, S Padmanabhan, J Vitter, R ... - Advances in Web-Age ..., 2001 - Springer
Abstract. The World Wide Web is growing and changing at an aston- ishing rate. For the information in the web to be useful, web information systems such as search engines have to keep up with the growth and change of the web. In this paper we study how web documents change. In ...
Cited by 38 - Related articles - BL Direct - All 14 versions

Dynamic maintenance of web indexes using landmarks
L Lim, M Wang, S Padmanabhan, JS Vitter, ... - Proceedings of the ..., 2003 - portal.acm.org
Page 1. Dynamic Maintenance of Web Indexes Using Landmarks **Lipyeow** Lim ∗ Duke University, Durham, NC 27708–0129, USA. **lipyeow**@cs.duke.edu Min Wang IBM TJ Watson Research Ctr. Hawthorne, NY 10532, USA. min@us.ibm.com ...
Cited by 37 - Related articles - All 19 versions

SASH: A self-adaptive histogram set for dynamically changing workloads
L Lim, M Wang, JS Vitter - ... of the 29th international conference on ..., 2003 - portal.acm.org
... Dynamically Changing Workloads **Lipyeow** Lim1 Min Wang2 Jeffrey Scott Vitter3 1 Dept. of Computer Science Duke University Durham, NC 27708, USA. **lipyeow**@cs.duke.edu 2 IBM TJ Watson Research Center 19 Skyline Drive Hawthorne, NY 10532, USA min@us.ibm.com ...
Cited by 31 - Related articles - BL Direct - All 20 versions

Wavelet-based cost estimation for spatial queries
M Wang, J Vitter, L Lim, S Padmanabhan - Advances in Spatial and ..., 2001 - Springer
Abstract. Query cost estimation is an important and well-studied prob- lem in relational database systems. In this paper we study the cost esti- mation problem in the context of spatial database systems. We introduce a new method that provides accurate cost estimation for spatial ...
Cited by 22 - Related articles - BL Direct - All 11 versions

Supporting ranking and clustering as generalized order-by and group-by
C Li, M Wang, L Lim, H Wang, KCC ... - Proceedings of the 2007 ..., 2007 - portal.acm.org
The Boolean semantics of SQL queries cannot adequately capture the "fuzzy" preferences and "soft" criteria required in non-traditional data retrieval applications. One way to solve this problem is to add a flavor of "information retrieval" into database queries by allow- ing fuzzy query ...
Cited by 21 - Related articles - All 12 versions

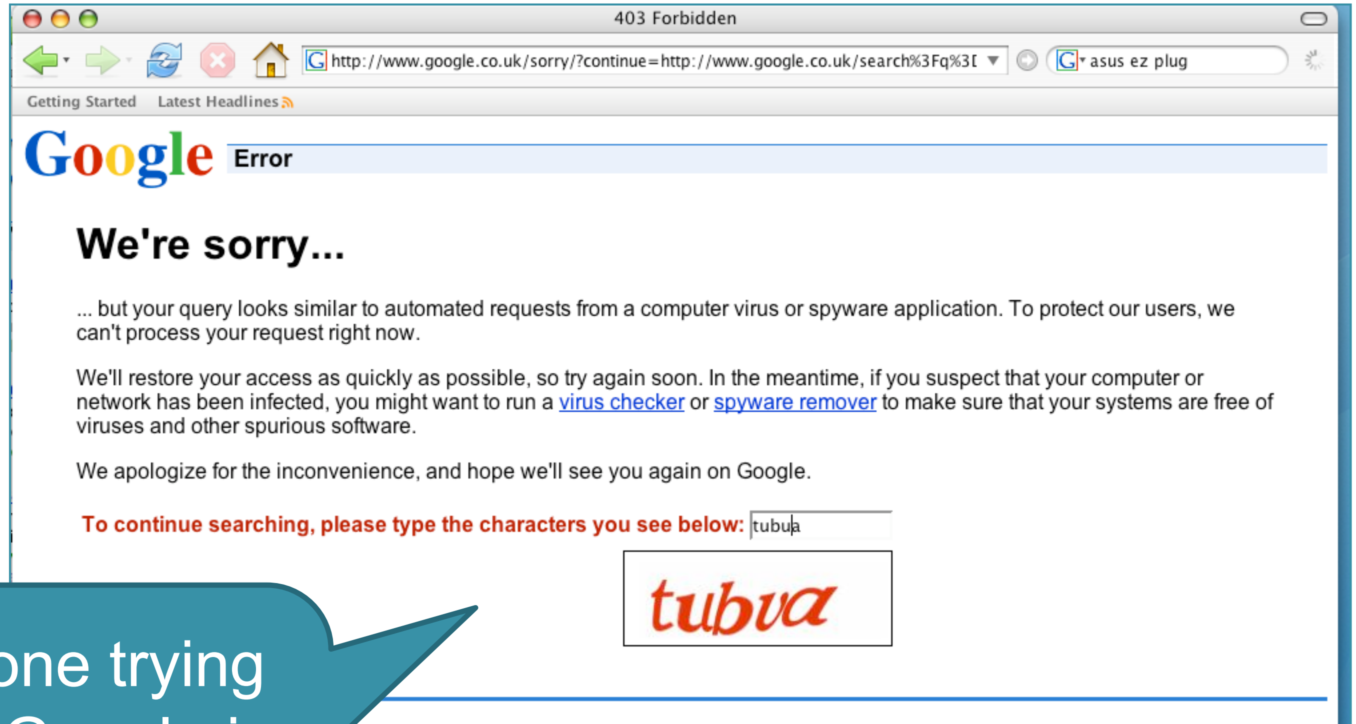
Employee	Publication	Citation
Lipyeow	XPathLearner ...	84
Lipyeow	Characterizing...	38
Haixun	Clustering by ...	308
Haixun	Mining concept ...	424
...	...	...

The Simple Solution

```
Loop
    Q = set of keyword queries
    Foreach q in Q
        Send q to Google Scholar
        Scrape the first few pages into tuples
        Update local relation using scraped tuples
    Sleep for t seconds
End Loop
```

- Query Google Scholar using researcher's name and/or publication title to get
  - new publications and
  - updated citation counts

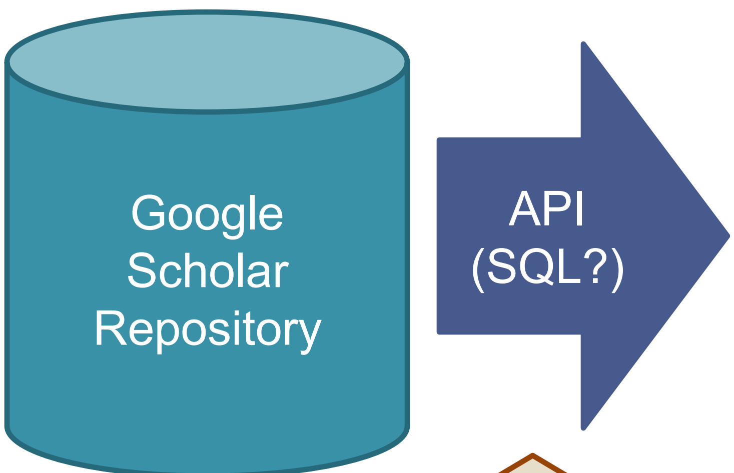
Problem with the Simple Solution



Everyone trying to use Google in the building got this screen !

The Elegant Solution

- All this hacking (including the solution I am about to present) could be avoided if there was an API to get structured relations from Google Scholar.



Employee	Publication	Citation
Lipyeow	XPathLearner ...	84
Lipyeow	Characterizing...	38
Haixun	Clustering by ...	308
Haixun	Mining concept ...	424
...	...	...

Linking Open Data effort might address this issue...

But ...

- Such API's don't exist (yet?)
- And ...

I need those citation counts by next week!

Problem Statement

- Local database periodically synchronizes its data subset with the data source
- Data source supports keyword query API only
- Extract relations from the top k results (ie first few result pages) to update local database

> At each synchronization,
> find a set of queries that will maximize the "content freshness" of the local database.

- Only relevant keywords are used in the queries
- Keywords cover the local relation
- Number of queries should be minimized
- Result size should be minimized

NP-Hard by reduction to Set Cover

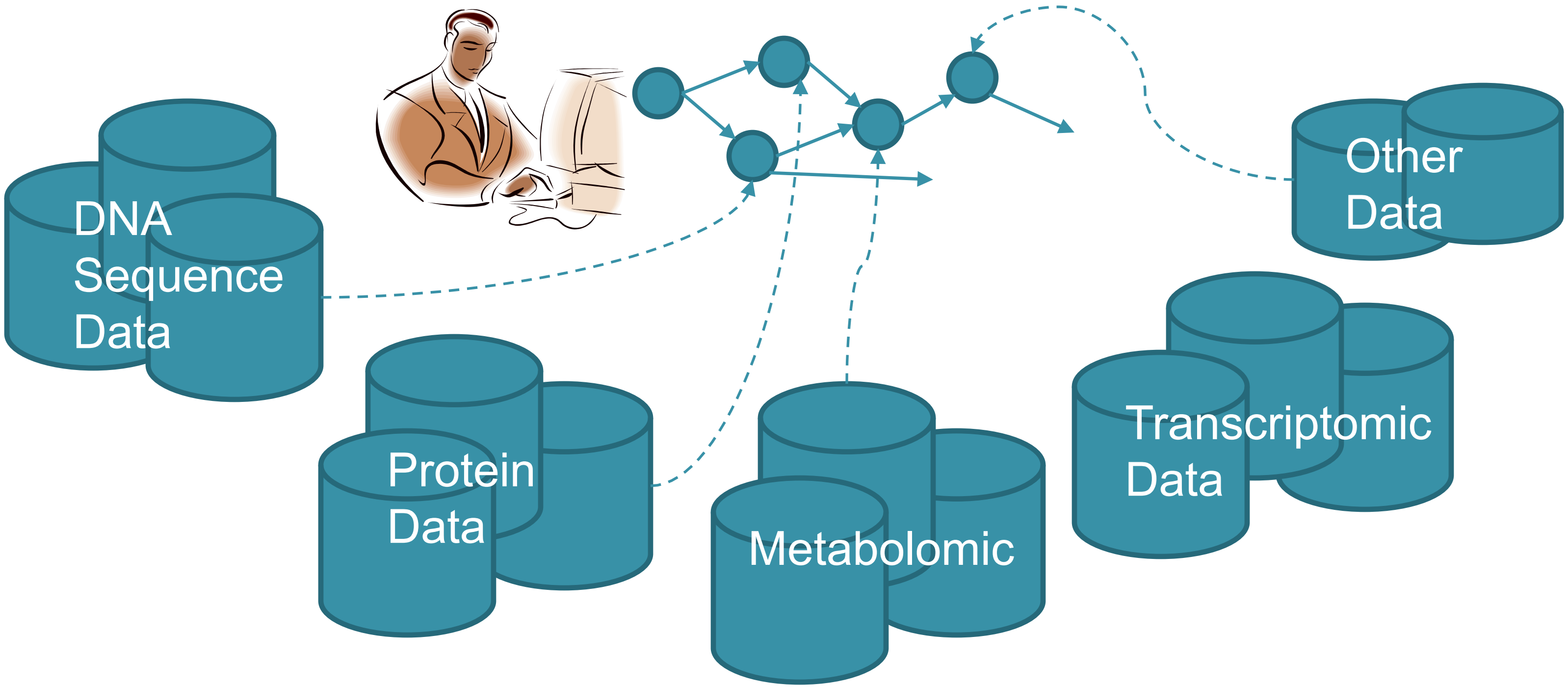
Picking the Right Queries ...

```
Loop
    Q = set of keyword queries
    Foreach q in Q
        Send q to Google Scholar
        Scrape the first few pages into tuples
        Update local relation using scraped tuples
    Sleep for t seconds
End Loop
```

- The simple algorithm is fine, we just need to pick the right queries...
  - Not all tuples are equal – some don't get updated at all, some are updated all the time
  - Some updates are too small to be significant

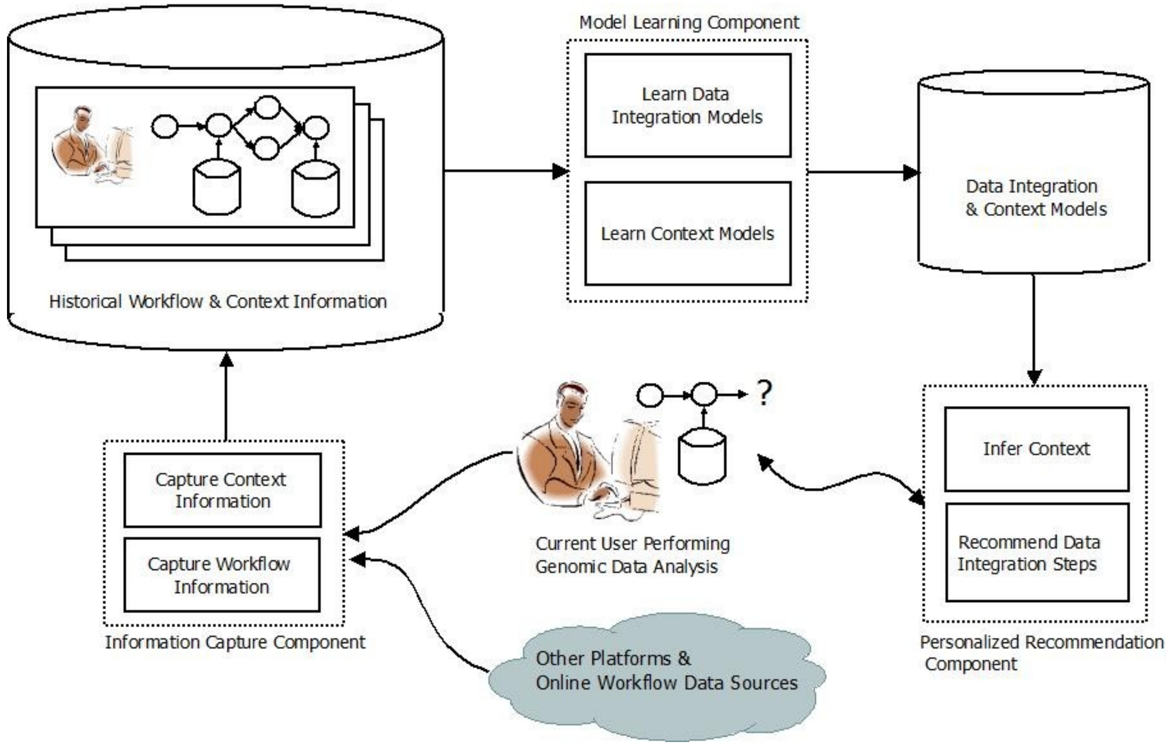
Mining Workflows for Data Integration Patterns

Bio-Informatics Scenario



- Each category has many online data sources
- Each data source may have multiple API and data formats
- Workflow is like a program or a script
  - A connected graph of operations

A Data Integration Recommender



- Data integration patterns
  - Generalize on key-foreign key relationships
  - Associations between schema elements of data and/or processes
- Analyze historical workflows to extract data integration patterns
- Make personalized recommendations to users as they create workflows

Problems & Tasks

- What are the different types of data integration patterns we can extract from workflows ?
- How do we model these patterns ?
- How do we mine workflows for these patterns ?
- How do we model context ?
- How do we make recommendations ?

Optimizing Sensor Data Acquisition for Energy-Efficient Smartphone-based Continuous Event Processing

Lipyeow Lim
University of Hawai`i at Mānoa

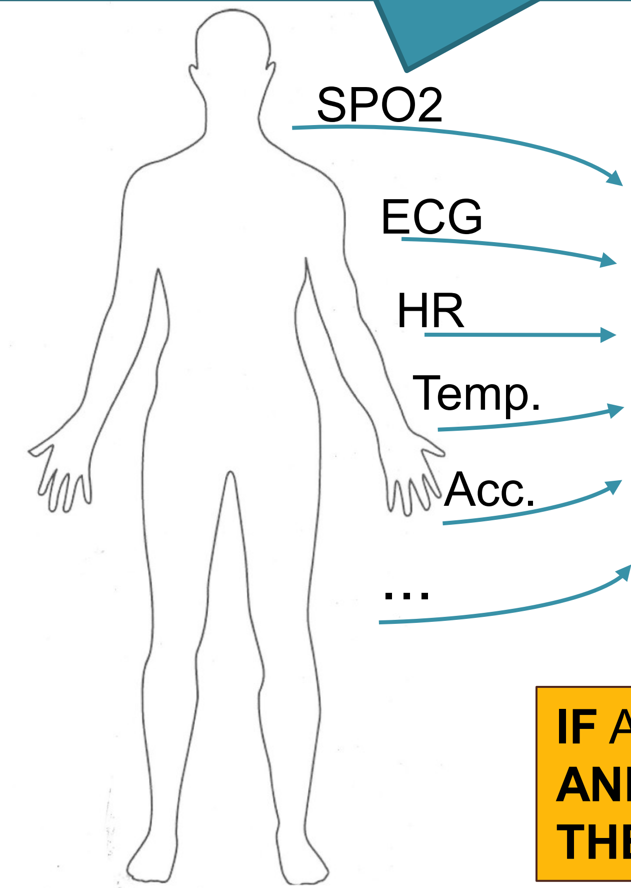
Archan Misra
Singapore Management University

Telehealth Scenario

Wearable sensors transmit vitals to cell phone via wireless (eg. bluetooth)

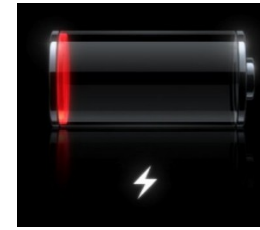
Phone runs a complex event processing (CEP) engine with rules for alerts

Alerts can notify emergency services or caregiver



IF Avg(Window(HR)) > 100
AND Avg(Window(Acc)) < 2
THEN SMS(doctor)

Energy Efficiency



- Energy consumption of processing
  - Sensors: transmission of sensor data to CEP engine
  - Phone: acquisition of sensor data
  - Phone: processing of queries at CEP engine

This Talk

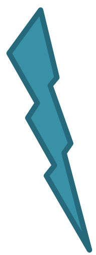
- Optimization objectives
  - Minimize energy consumption at phone
  - Maximize operational lifetime of the system.

This Talk

Sensor Data Acquisition

3D acc.
ECG,
EMG,
GSR

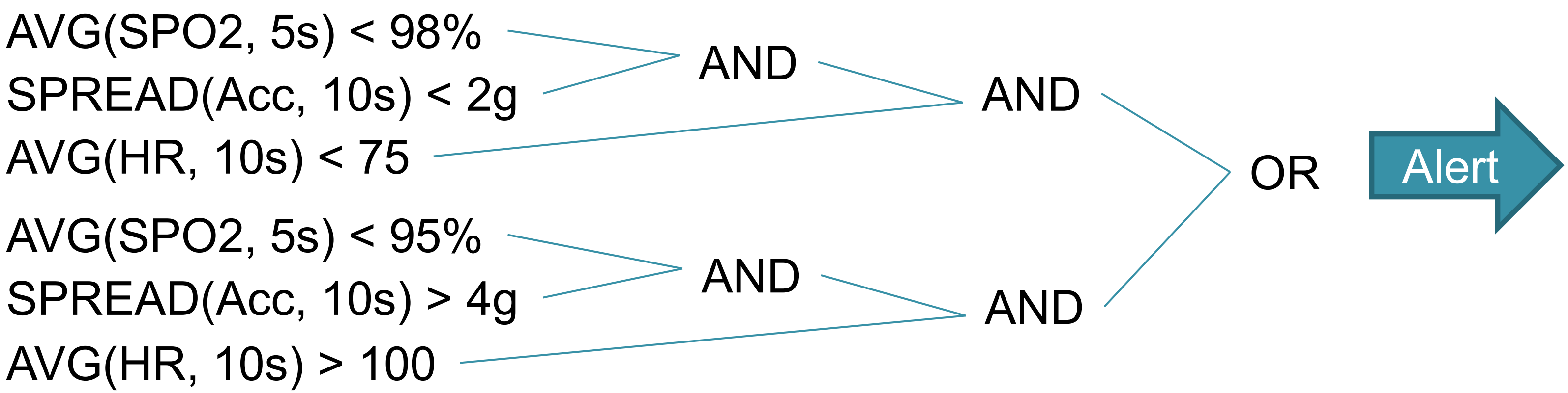
Bluetooth
Or 802.11
Or 802.15



- Constant sampling rate
- 802.11 (wifi) uses 2 power modes: active, idle
- Bluetooth has 3 modes: active, idle, sleep (not relevant).
- Time needed to switch modes
- Energy expended to switch

Sensor Type	Bits/ sensor channel	Channels/ device	Typical sampling frequency (Hz)
GPS	1408	1	1 Hz
SpO2	3000	1	3 Hz
ECG (cardiac)	12	6	256 Hz
Accelero-meter	64	3	100 Hz
Temperature	20	1	256 Hz

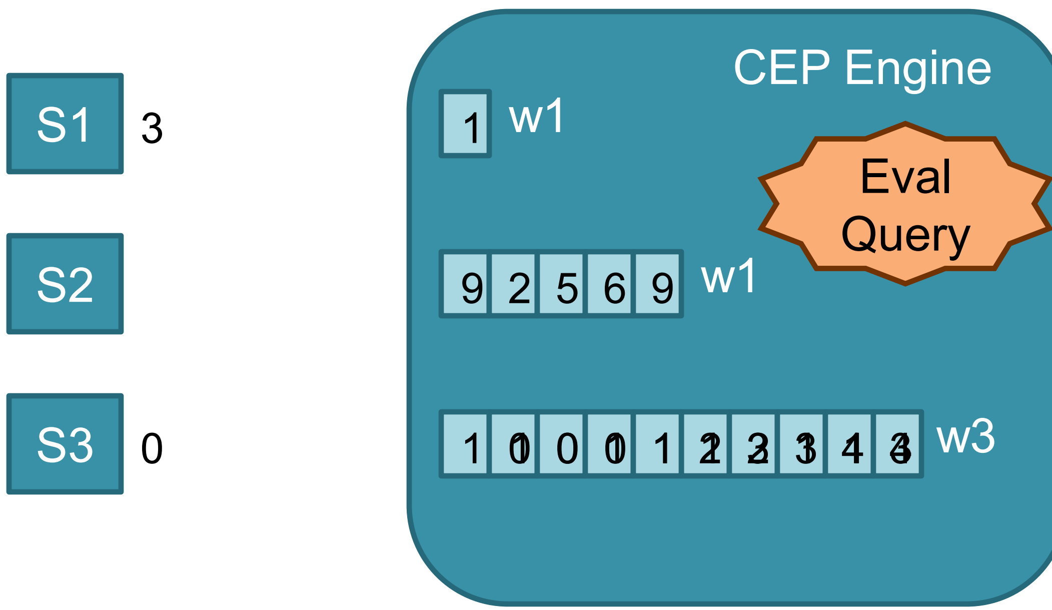
Query Model



- A query is a boolean combination of predicates
- Predicates
  - Aggregation functions over a time-based window of sensor data
- Traditional push model
  - A given query is evaluated whenever a new sensor reading arrives

Continuous Evaluation

if Avg(S2, 5)>20 AND S1<10 AND Max(S3,10)<4 then email(doctor).



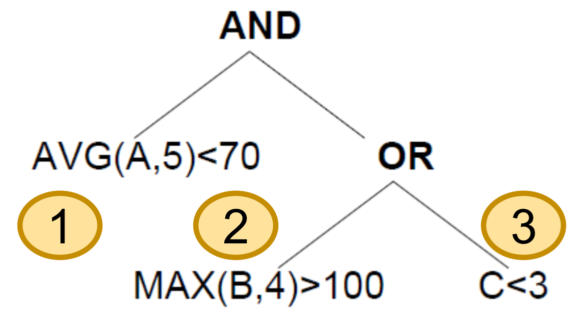
Push
When t_i of Si arrives
  Enqueue t_i into W_i
  If Q is true,
  Then output alert

Pull
Loop
  Acquire t_i for Si
  Enqueue t_i into W_i
  If Q is true,
  Then output alert
End loop

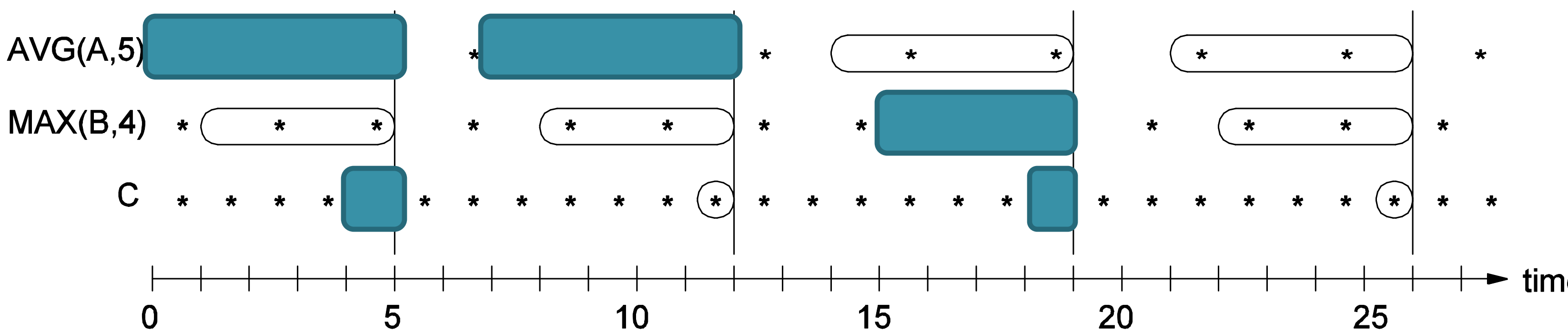
Key Ideas

- Pull model
  - Evaluate a query every ω seconds
  - Acquire only data that is needed
- Evaluation order of predicates matter!
  - Shortcircuiting can avoid data acquisition
- Batching

Example: $\omega=7$



- Time 5: eval order is 3,1,2
- Time 12: eval order is 1,2,3
- Time 19: eval order is 2,3,1



Evaluation Order

if Avg(S2, 5)>20 AND S1<10 AND Max(S3,10)<4 then email(doctor).

Predicate	Avg(S2, 5)>20	S1<10	Max(S3,10)<4
Acquisition	5 * .02 = 0.1 nJ	0.2 nJ	10 * .01 = 0.1 nJ
Pr(false)	0.95	0.5	0.8
Acq./Pr(f)	0.1/0.95	0.2/0.5	0.1/0.8

- Evaluate predicates with lowest energy consumption first
- Evaluate predicates with highest false probability first
- Hence, evaluate predicate with lowest normalized acquisition cost first.

A Lot More Work Needed

- Improve simulator
  - Disjunctive normal form query representation
  - More realistic data generators
- Trade-off between semantics of the query with energy
- Estimation algorithms for P(pred=true)
  - Condition on context
- Batching: wait say 3 ω before query evaluation
  - Design and implement the algorithm
  - Evaluation via simulation
- End-to-end evaluation on Android phone
  - Maximize operational lifetime of phone+sensors

Other projects

- Cloud-based SQL Processor for Scientific Applications
  - Benchmarking work
  - Query optimization for parallel SQL processing
  - Elastic & dynamic parallelization
- Develop a journal version of: *Optimizing Access Across Multiple Hierarchies in Data Warehouses*
- Data compression of Join Query Result Sets