Networked Data Management (streams, RFID, sensors)

Γιάννης Θεοδωρίδης

InfoLab, Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιά

http://infolab.cs.unipi.gr

version: Nov.2009

Sources:

- M. Garofalakis, J. Gehrke, R. Rastogi. Querying and Mining Data Streams: You Only Get One Look. VLDB, 2002.
- J. Han. Warehousing and Mining Massive RFID Data Sets. RFDM, 2008.
- R. Motwani. Models and Issues in Data Stream Systems. PODS, 2002.
- D. R. Thompson. Radio Frequency Identification (RFID) Technologies. Tutorial, Univ. Arkansas. http://csce.uark.edu/~drt/rfid.

Outline

- Introduction - Applications
- Data Streams
  - Data, Queries, Synopses, Projects
- RFID data
  - Data management challenges
- Wireless Sensor Networks
  - Architecture, Queries
- Time-series
  - Similarity aspects

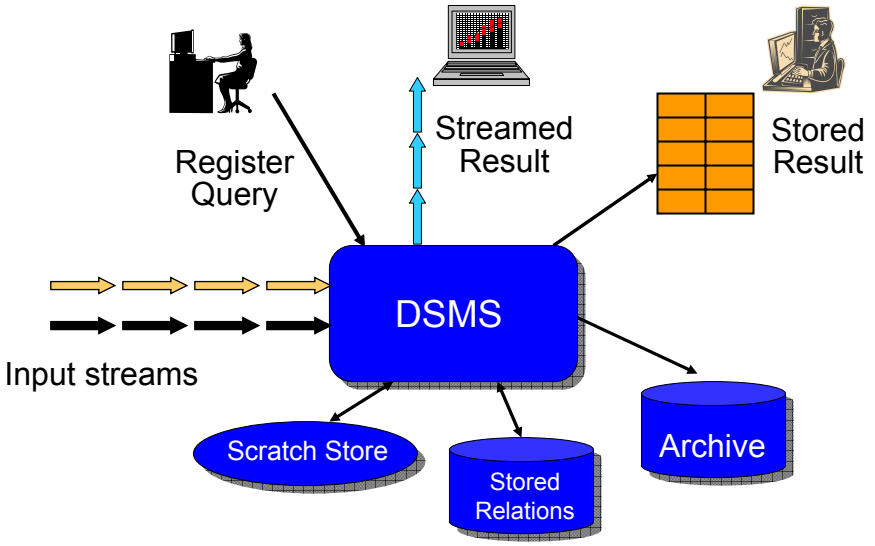
Outline

- Introduction - Applications
- Data Streams
  - Data, Queries, Synopses, Projects
- RFID data
  - Data management challenges
- Wireless Sensor Networks
  - Architecture, Queries
- Time-series
  - Similarity aspects

Introduction

- Data elements in the stream arrive online
- System has no control over order in which data elements to be processed
- Data streams are potentially unbounded in size
- Once an element from a data stream has been processed, it is discarded or archived. It cannot be retrieved easily unless it is stored in memory, which is small relative to the size of data streams

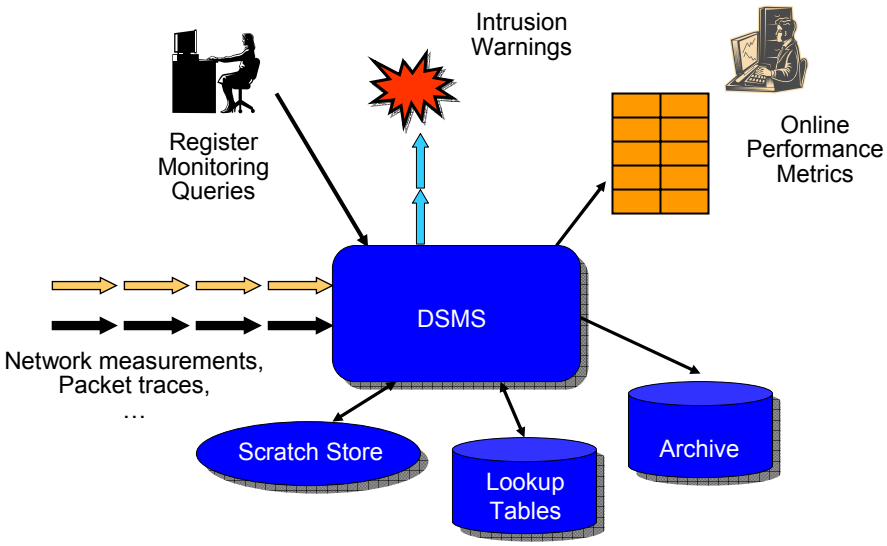
DSMS – big picture



Applications

- New Applications – data input as **continuous, ordered data streams**
  - Network monitoring and traffic engineering
  - Telecom call records
  - Network security
  - Financial applications
  - Sensor networks
  - Logistics, Manufacturing processes (RFID)
  - Web logs and clickstreams

Application: Network Monitoring



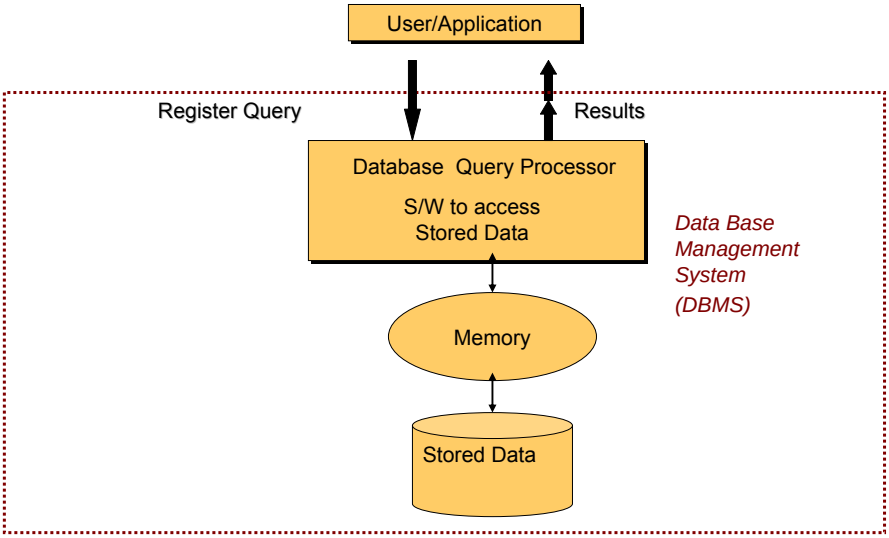
Network Data Processing

- **Traffic estimation**
  - How many bytes were sent between a pair of IP addresses?
  - What fraction network IP addresses are active?
  - List the top 100 IP addresses in terms of traffic
- **Traffic analysis**
  - What is the average duration of an IP session?
  - What is the median of the number of bytes in each IP session?
- **Fraud detection**
  - Identify all sessions whose duration was more than twice the normal
- **Security / Denial of Service**
  - List all IP addresses that have witnessed a sudden spike in traffic
  - Identify IP addresses involved in more than 1000 sessions

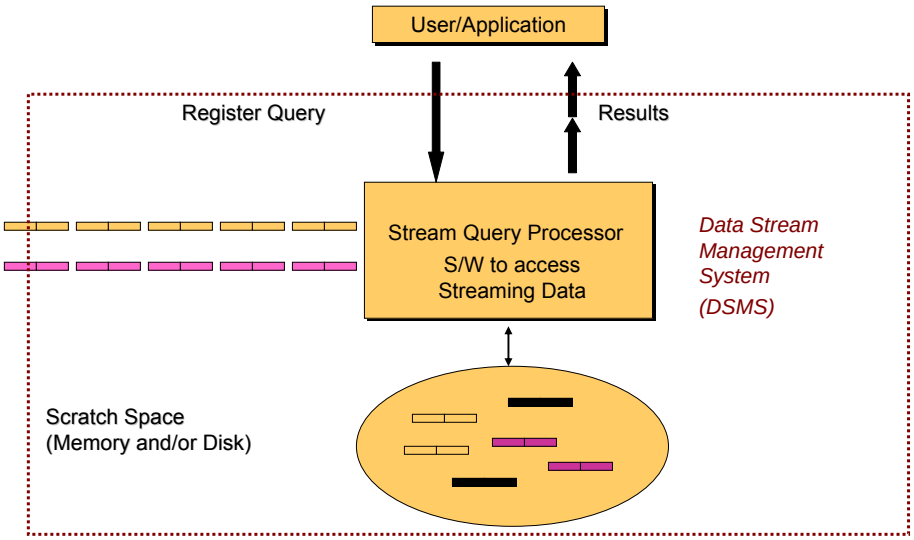
Source	Dest.	Dur.	Bytes	Protocol
10.1.0.2	16.2.3.7	12	20K	http
18.6.7.1	12.4.0.3	16	24K	http
13.9.4.3	11.6.8.2	15	20K	http
15.2.2.9	17.1.2.1	19	40K	http
12.4.3.8	14.8.7.4	26	58K	http
10.5.1.3	13.0.0.1	27	100K	ftp
11.1.0.6	10.3.4.5	32	300K	ftp
19.7.1.2	16.5.5.8	18	80K	ftp

Example IP session data (collected using Cisco NetFlow). AT&T collects 100 GBs of NetFlow data each day!

The Database Model



The Data Stream Model



DBMS vs. DSMS

- Persistent relations
- One-time queries
- Random access
- "Unbounded" disk store
- Only current state matters
- Passive repository
- Relatively low update rate
- No real-time services
- Assume precise data
- Access plan determined by query processor, physical DB design

- Transient streams
- Continuous queries
- Sequential access
- Bounded main memory
- History/arrival-order is critical
- Active stores
- Possibly multi-GB arrival rate
- Real-time requirements
- Data stale/imprecise
- Unpredictable/variable data arrival and characteristics

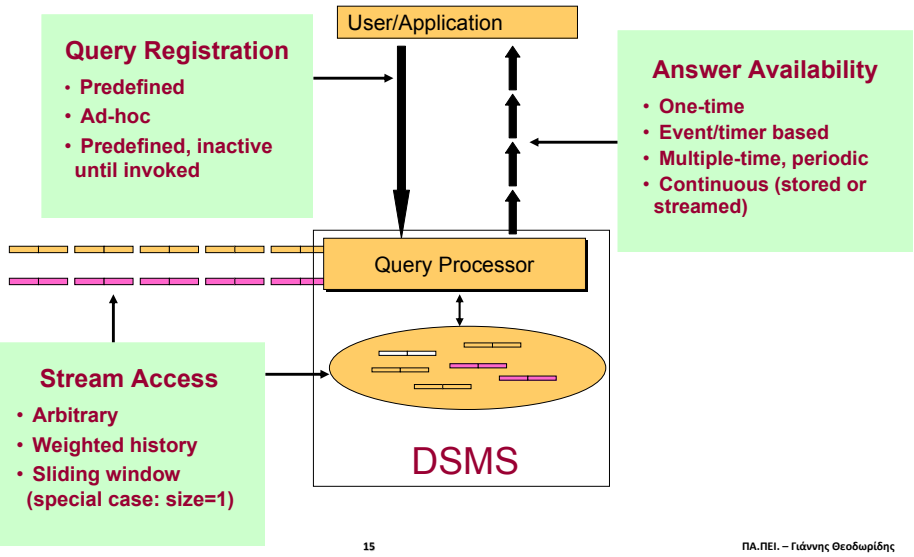
Outline

- Introduction - Applications
- Data Streams
  - Data, Queries, Synopses, Projects
- RFID data
  - Data management challenges
- Wireless Sensor Networks
  - Architecture, Queries
- Time-series
  - Similarity aspects

Data Model

- Append-only
  - Call records
- Updates
  - Infrequently, e.g. stock tickers
- Deletes
  - Infrequently, e.g. in case of disk-resident (transactional) data

Query Model



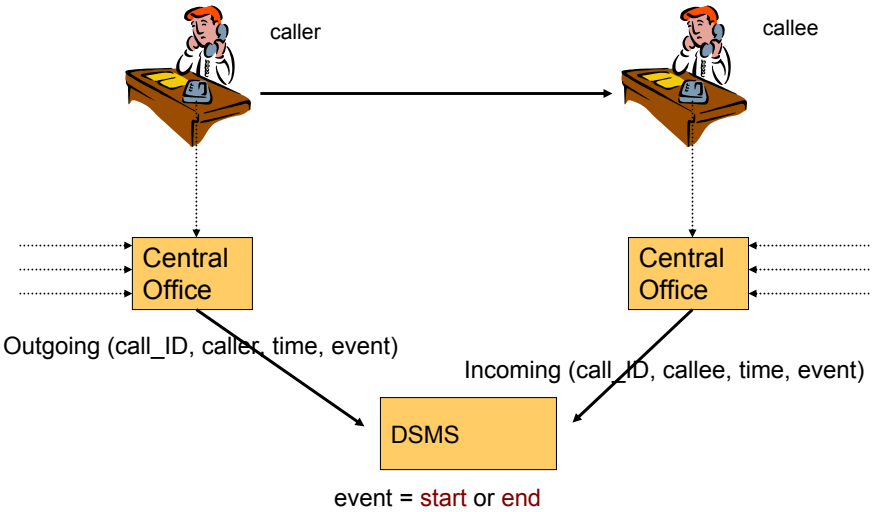
Queries

- One-time queries and Continuous queries
  - One-time queries
    - Evaluated once over a point-in-time snapshot of data set
  - Continuous queries
    - Evaluated continuously as data streams continue to arrive
    - May be stored and updated as new data arrives, or may produce data streams themselves

Queries

- Predefined and Ad hoc queries
  - Predefined queries
    - Supplied to data stream management system before any relevant data has arrived
    - Usually continuous queries
    - Scheduled one-time queries possible
  - Ad hoc queries
    - Can be either one-time or continuous queries
    - Complicates design of data stream management system (DSMS), because they are not known in advance for purposes of query optimization and correctly answering it may require referencing data that may have already arrived on data streams and potentially have already been discarded

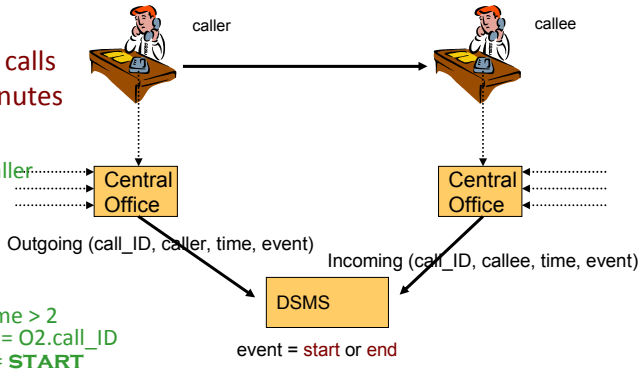
Making Things Concrete



Query 1 (self-join)

- Find all outgoing calls longer than 2 minutes

```
SELECT
    O1.call_ID, O1.caller
FROM
    Outgoing O1,
    Outgoing O2
WHERE
    (O2.time – O1.time > 2
    AND  O1.call_ID = O2.call_ID
    AND  O1.event = START
    AND  O2.event = END)
```

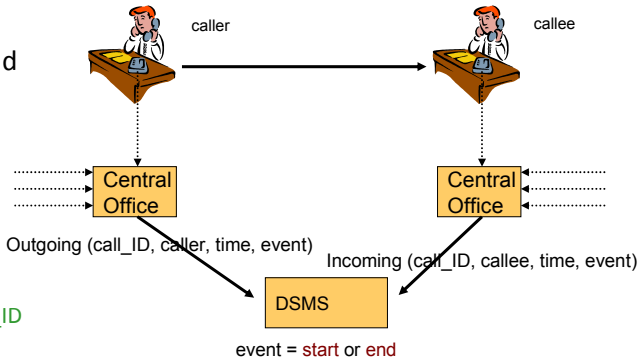


- Result requires unbounded storage
- Can provide result as data stream
- Can output after 2 min, without seeing **END**

Query 2 (join)

- Pair up callers and callees

```
SELECT
    O.caller, I.callee
FROM
    Outgoing O,
    Incoming I
WHERE
    O.call_ID = I.call_ID
```

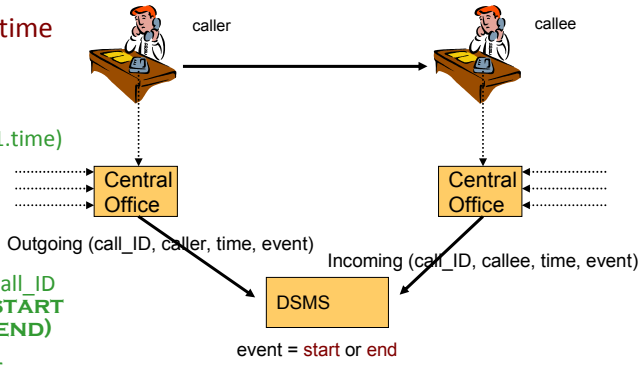


- Can still provide result as data stream
- Requires unbounded temporary storage …
- … unless streams are near-synchronized

Query 3 (group-by aggregation)

- Total connection time for each caller

```
SELECT   O1.caller,
    sum(O2.time – O1.time)
FROM
    Outgoing O1,
    Outgoing O2
WHERE
    (O1.call_ID = O2.call_ID
    AND O1.event = START
    AND  O2.event = END)
GROUP BY    O1.caller
```



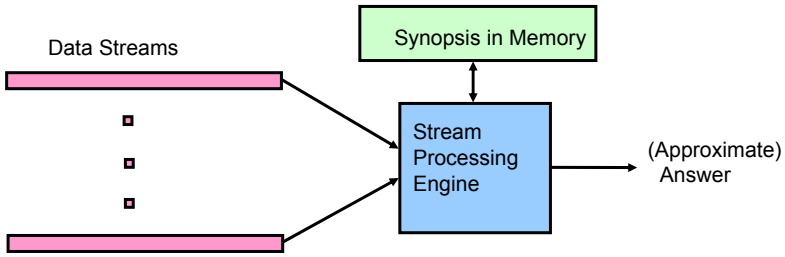
- Cannot provide result in (append-only) stream
  - Output updates?
  - Provide current value on demand?
  - Memory?

Approximate Query Answering

- Why?
  - Handling load
    - streams coming too fast
  - Avoid unbounded storage and computation
    - Data streams are potentially unbounded in size, the amount of storage required to compute exact answer to a query may grow without bound
  - Ad hoc queries need approximate history
- High-quality approximate answers can be an acceptable solution
- How? Sliding windows, synopsis, samples, load-shed

Computation Model

- A data stream is a (massive) sequence of elements

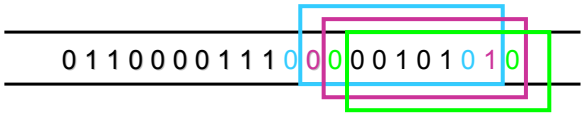


- Stream processing requirements
  - Single pass: Each record is examined at most once
  - Bounded storage: Limited Memory (M) for storing synopsis
  - Real-time: Per record processing time (to maintain synopsis) must be low

Synopses

- Queries may access or aggregate past data
- Need bounded-memory history-approximation
- Synopsis?
  - Succinct summary of old stream tuples
  - Like indexes/materialized-views, but base data is unavailable
- Examples
  - Sliding Windows
  - Samples
  - Sketches
  - Histograms
  - Wavelet representation

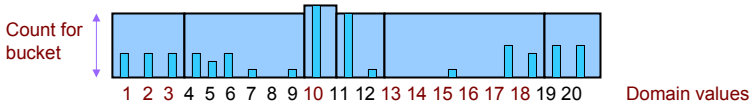
Sliding Window Approximation



- Why?
  - Approximation technique for bounded memory
  - Natural in applications (emphasizes recent data)
  - Well-specified and deterministic semantics
- Issues
  - Extend relational algebra, SQL, query optimization
  - Algorithmic work

Histograms

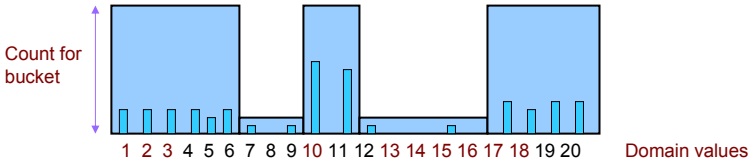
- Equi-Depth Histograms
  - Idea: Select buckets such that counts per bucket are equal



- V-Optimal Histograms

$$\text{minimize} \sum_B \sum_{v \in B} \left(f_v - \frac{C_B}{V_B}\right)^2$$

  - Idea: Select buckets to minimize frequency variance within buckets



Wavelets

- Wavelets: Mathematical tool for hierarchical decomposition of functions/signals
  - Haar wavelets: Simplest wavelet basis, easy to understand and implement
  - *Recursive pairwise averaging and differencing* at different resolutions

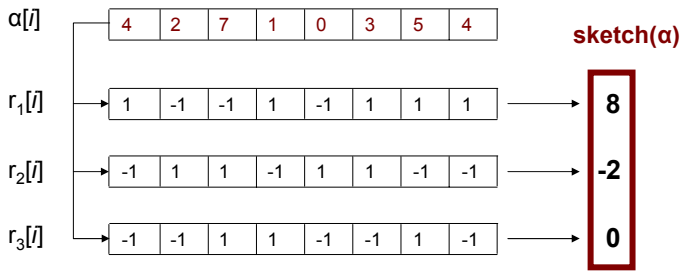
Resolution	Averages	Detail Coefficients
3	[2, 2, 0, 2, 3, 5, 4, 4]	----
2	[2, 1, 4, 4]	[0, -1, -1, 0]
1	[1.5, 4]	[0.5, 0]
0	[2.75]	[-1.25]

Haar wavelet decomposition: [2.75, -1.25, 0.5, 0, 0, -1, -1, 0]

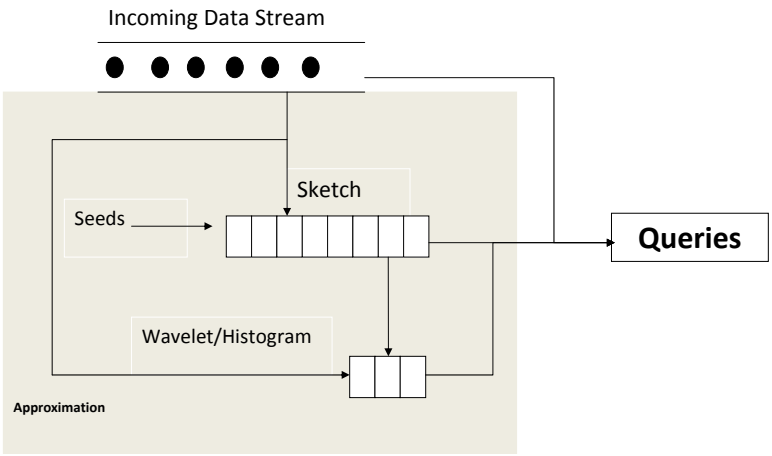
Compression by ignoring small coefficients !!

Sketches

- e.g. internal product of α with $O(\log(N/\delta)/\varepsilon^2)$ pseudorandom {-1,+1} vectors



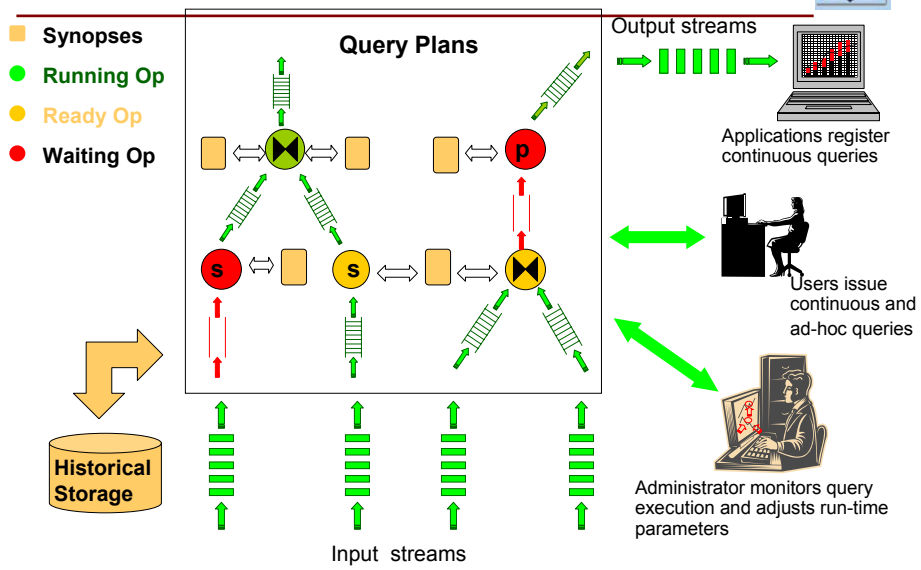
Streaming Model



Stream Projects

- Amazon/Cougar (Cornell) – sensors
- Aurora (Brown/MIT) – sensor monitoring, dataflow
- Hancock (AT&T) – telecom streams
- Niagara (OGI/Wisconsin) – Internet XML databases
- OpenCQ (Georgia) – triggers, incr. view maintenance
- Stream (Stanford) – general-purpose DSMS
- Tapestry (Xerox) – pub/sub content-based filtering
- Telegraph (Berkeley) – adaptive engine for sensors
- Tribeca (Bellcore) – network monitoring

Aurora/STREAM Overview

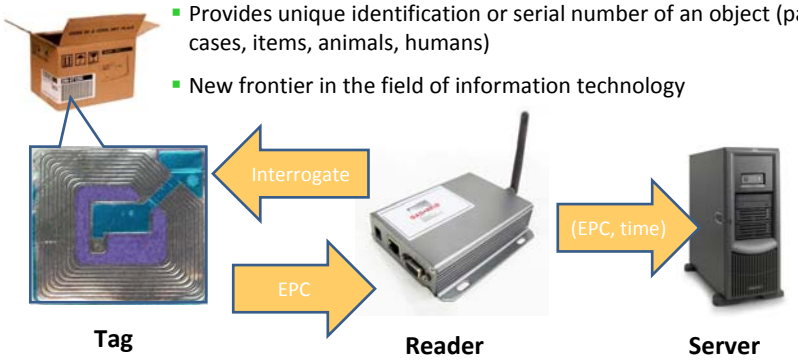


Outline

- Introduction - Applications
- Data Streams
  - Data, Queries, Synopses, Projects
- RFID data
  - Data management challenges
- Wireless Sensor Networks
  - Architecture, Queries
- Time-series
  - Similarity aspects

What is RFID?

- Stands for **R**adio **F**requency **Id**entification
- Technology that allows a sensor (reader) to read, from a distance, and without line of sight, a tag-based unique **electronic product code** (EPC)
  - Provides unique identification or serial number of an object (pallets, cases, items, animals, humans)
  - New frontier in the field of information technology

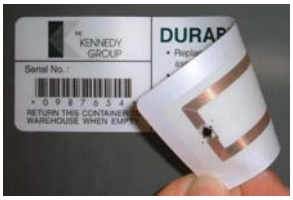


33

ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

EPC vs. UPC (Barcodes)

- Both are forms of Automatic identification technologies
- Universal Product Code (UPC) require line of sight and manual scanning whereas EPC do not
- UPC require optical reader to read whereas EPC reader reads via radio waves
- EPC tags possess a memory and can be written while UPC do not
- EPC tags cost 5 cents, UPC tags cost 1/10 cent



34

ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Applications

- Asset tracking
- E-pass road toll system
- E-passports
- Animal Identification
- Humans' healthcare
- Inventory & supply chain management
- …

Application: Asset Tracking

British Airways loses 20 million bags a year

Application: E-Toll Collection

Illinois: 1 million drivers a day use I-Pass

Application: E-Passports

- (biometric) e-passports since 2007 in US, since 2009 in EU
- ISO 14443 RFID chip in rear cover
- Includes:
  - passport number, name, gender, date and place of birth,
  - dates of passport issuance and expiration,
  - digital image of the bearer's photograph
- Digital photograph is used as biometric identifier
- Anti-skimming material in cover to prevent unauthorized reading when it is closed
- Randomized unique identification (RUID) to prevent tracking
- Information signed with a digital signature

Application: Human healthcare

- Product: VeriChip
- Human implantable RFID tag operating at about 134 KHz because at these frequencies the RF can penetrate mud, blood, and water
- About the size of uncooked grain of rice
- Healthcare applications
  - Implanted medical device identification
  - Emergency access to patient-supplied health information
  - Portable medical records access including insurance information
  - In-hospital patient identification
  - Medical facility connectivity via patient
  - Disease/treatment management of at-risk populations (such as vaccination history)

"… About the size of a grain of rice, the microchip is inserted just under the skin and contains only a unique, 16-digit identifier. The microchip itself does not contain any other data other than this unique electronic ID, nor does it contain any Global Positioning System (GPS) tracking capabilities. And unlike conventional forms of identification, the Health Link cannot be lost, stolen, misplaced, or counterfeited. It is safe, secure, reversible, and always with you."
(source: www.verichipcorp.com)

Application: Inventory Management

How many pens should we reorder?

Application: Supply Chain Management



Factory

Shipping

Warehouse

Checkout

Shelf


Application: Supply Chain Management

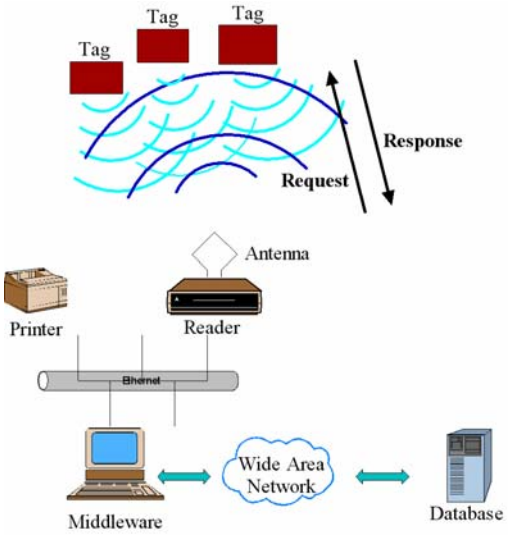
- RFID adds visibility as the items flow through the supply chain from the manufacturer, shippers, distributors, and retailers.
- The added visibility can identify bottlenecks and save money.
- Scope: ~ 6 meters
- **Electronic Product Code (EPC) 96-bit Version**
  - Every product has unique identifier among $2^{96} \cong 8 \times 10^{28}$ different combinations
  - 96 bits can uniquely label all products for the next 1,000 years (!!)

Version	EPC Manager (Manufacturer)	Object Class (Product)	Serial Number
8 bits	28 bits	24 bits	36 bits

Video …

METRO Group
Future Store Initiative

RFID system



RFID reader

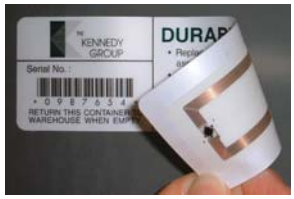
- Also known an interrogator
- **Reader powers passive tags with RF energy**
- Can be handheld or stationary
- Consists of:
  - Transceiver
  - Antenna
  - Microprocessor
  - Network interface

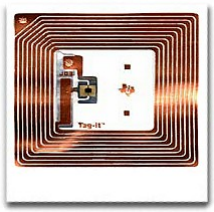
Reader

Antenna

RFID tags

- Tag is a device used to transmit information such as a serial number to the reader in a contact less manner
- Classified as :
  - Passive – energy from reader
  - Active - battery
  - Semi-passive – battery and energy from reader





45

ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Applications, frequencies, and standards

Applications	Frequencies	Standards
Animal Identification (dogs, cats, cattle)	< 135 KHz	ISO 18000–2, ISO 11784, ISO 11785, ISO 14223
Smart cards, Passport, Books at library	13.553 – 13.567 MHz	ISO 18000–3, ISO 7618, ISO 14443, ISO 15693 13.56 MHz ISM Band Class 1
Supply chain for retail	868 – 928 MHz	EPCglobal Class-1 Gen-2 ISO 18000–6

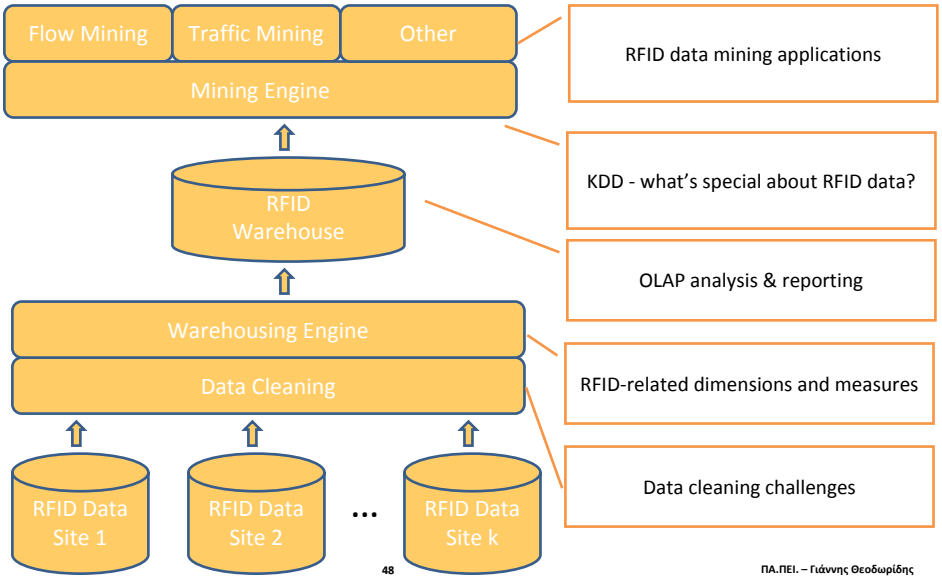
46

ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

RFID Data Management

- Emerging domain!
- Int. Workshop on RFID Data Management
  - Cancun, Mexico, 2008: http://rfid.cs.washington.edu/rfdm08/
- Challenges:
  - Around 3 billion Radio Frequency Identification (RFID) tags were deployed till today.
  - Wal-Mart's in-store implementation will generate about **7 Tbytes of RFID data per day** !!

RFID Data Warehousing and Mining



Example: Trajectory in Supply Chain



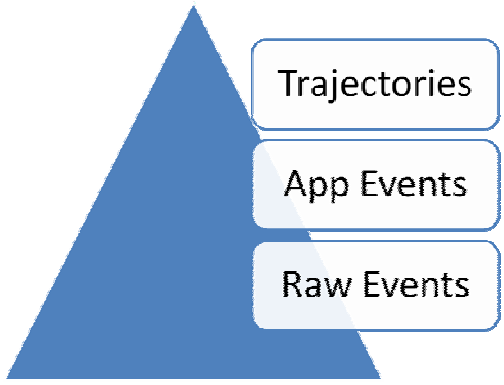
(Factory, T_1,T_2) (Shipping,T_3,T_4) (Warehouse, T_5,T_6)

(Checkout,T_9,T_{10}) (Shelf, T_7,T_8)


49

ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Data Generation



[EPC: $(L_1,T_1)(L_2,T_2)…(L_n,T_n)$]

[EPC, Location, Time_in, Time_out]

[EPC, Location, Time]

50

ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

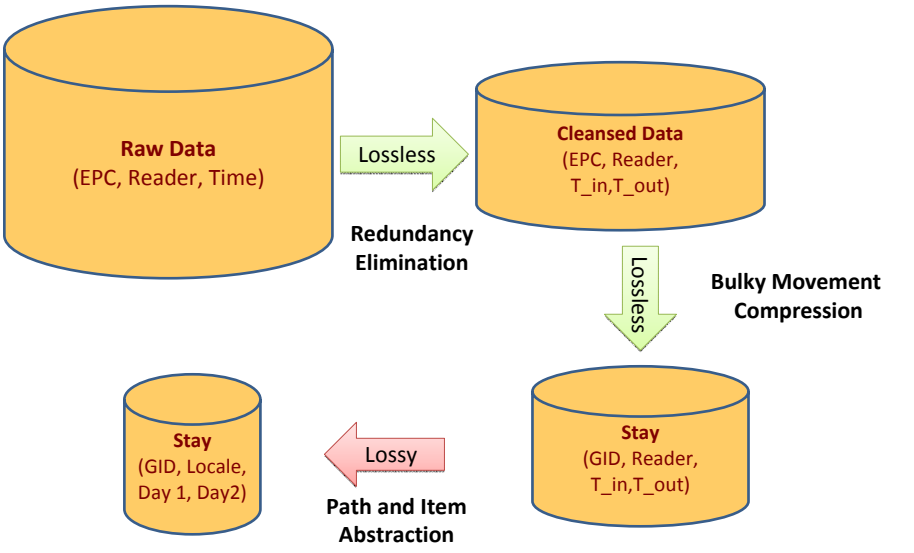
Why RFID-Warehousing?

- Significantly reduce the size of the RFID data set by redundancy removal and grouping objects that move and stay together
  - Lossless compression for bulky movement data !
  - An example: A retailer with 3,000 stores, selling 10,000 items a day per store, with each item being recorded 10 times on average before being sold ➔ Data volume: 300 million tuples per day !!
- Queries:
  - **OLAP**: Avg time for outwear items to move from warehouse to checkout counter in 03/2006?
  - **Mining**: Any correlation between the time spent at transportation and the milk in store S rotten?

Cleaning of RFID Data Records

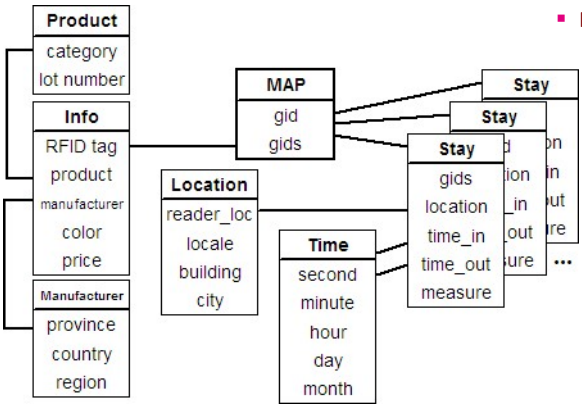
- **Raw Data**: Duplicate records due to multiple readings of a product at the same location
  - (EPC, location, time)
  - Example: (r_1, l_1, t_1) (r_1, l_1, t_2) ... (r_1, l_1, t_{10})
- **Cleansed Data**: Minimal information to store and removal of raw data
  - (EPC, Location, time_in, time_out)
  - Example: (r_1, l_1, t_1, t_{10})
- Warehousing can help fill-up missing records and correct wrongly-registered information

Data Compression



RFID-Cube Architecture

- **Stay Table:** (GIDs, location, time_in, time_out: measures)
  - items that stay together at a given location
- **Map Table:** (GID, <GID_1,..,GID_n>)
  - Links together stages that belong to the same path.



Example RFID Cuboid

Cleansed RFID Database

epc	loc	t_in	t_out
r1	l1	t1	t10
r1	l2	t20	t30
r2	l1	t1	t10
r2	l2	t20	t30
r3	l1	t1	10
r3	l4	t15	t20

Stay Table

Id list	loc	t_in	t_out
g1	l1	t1	t10
g1.1	l2	t20	t30
g1.2	l4	t15	t20

Map Table

gid	gids
g1	g1.1,g1.2
g1.1	r1,r2
g1.2	r3

55 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Outline

- Introduction - Applications
- Data Streams
  - Data, Queries, Synopses, Projects
- RFID data
  - Data management challenges
- Wireless Sensor Networks
  - Architecture, Queries
- Time-series
  - Similarity aspects

56 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

WSNs – An Introduction (1/3)

- The vision:
  - Push connectivity out of the PC and into the real world
  - Billions of sensors and actuators everywhere
  - Zero configuration and administrative cost
  - Build everything out of CMOS so that each device costs pennies
  - Enable new sensing paradigms

New challenges in data stream management

WSNs – An Introduction (2/3)

- Wireless Sensor Networks (WSN) utility:
  - Scatter cheap, tiny motes in an area of interest
  - Perform querying operations
  - Obtain reports of physical quantities under study
  - Support sampling procedures, alert mechanism infrastructures, decision making processes, etc.

WSNs – An Introduction (3/3)

- Mote Features
  - Low Power, Low Power, Low Power...
  - Low processing capabilities
  - Constrained memory capacity
- Network Features
  - Wireless, multi-hop communication using ISM radio zones (433MHz – 2,4GHz)
  - Ad-hoc network topologies

Sensor Net Sample Apps

Habitat Monitoring: Storm petrels on Great Duck island, microclimates on James Reserve.

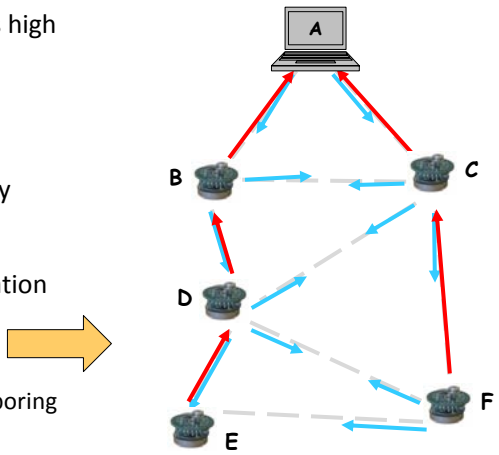
Vehicle detection: sensors along a road, collect data about passing vehicles.

Earthquake monitoring in shake-test sites.

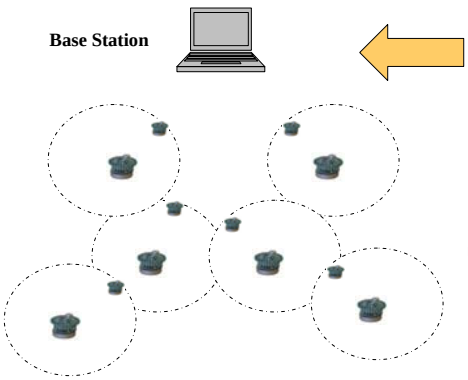
Traditional monitoring apparatus.

Communication in Sensor Nets

- Radio communication has high link-level losses
  - typically about 20% @ 5m
- Ad-hoc neighbor discovery
- Typical TAG (Tiny Aggregation Service) topology
  - 1 **base-station** (root)
  - Many nodes (with neighboring links)



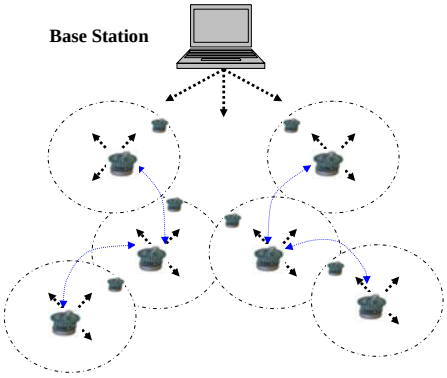
Communication in Sensor Nets



- An alternative TAG topology
  - 1 **base-station** (root)
  - Nodes organized in clusters, with one **clusterhead** and one **forwarder** per cluster
- Example query

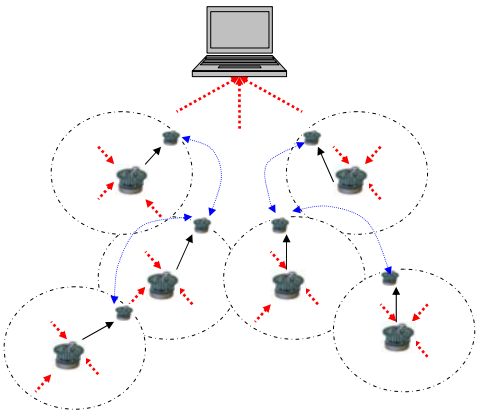
```
SELECT  faggr(attr), attr,…
FROM    sensors
WHERE predicate
OUTPUT ACTION action
SAMPLE PERIOD t sec FOR d sec
```

Communication in Sensor Nets



- Query dissemination:
  - The basestation poses a query of interest
  - The query is received by nearby clusterheads
  - Clusterheads propagate the query to their peers
  - …and to nodes within their cluster

Communication in Sensor Nets



- Query answer:
  - Clusterheads collect mote measurements
  - Send data to forwarders
  - Forwarders propagate the data towards the basestation
  - Answer reaches the basestation

Querying Sensor Nets – Snapshot queries

- Example 1 (output nodes recording light > 400 lux in 1s time intervals):

```
SELECT nodeid, light
   FROM sensors
   WHERE light > 400
   EPOCH DURATION 1s
```

Sensors

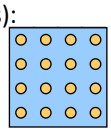
Epoch	Nodeid	Light	Temp	Accel	Sound
0	1	455	x	x	x
0	2	389	x	x	x
1	1	422	x	x	x
1	2	405	x	x	x

Result (streaming):

(1,455)
(1,422),(2,405)
…

Querying Sensor Nets – Aggregation queries

- Example 2 (output average sound in 10s time intervals):

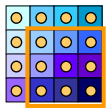


```
SELECT AVG (sound)
   FROM sensors
   EPOCH DURATION 10s
```

Result (streaming):

0: (440)
1: (445)
…

- Example 3 (output average sound per room in 10s time intervals, only when average sound > 200 db):




```
SELECT roomNo,  AVG (sound)
   FROM sensors
   GROUP BY roomNo
   HAVING AVG (sound) > 200
   EPOCH DURATION 10s
```

Result (streaming):

0: (1,360),(2,520)
1: (1,370),(2,520)
…

Querying Sensor Nets – Event-based queries

- Support for events as a mechanism for initiating data collection. Events are generated by either another query or the operation system.
- Query example (when an event „bird-detect“ appears at a location, output the average light and temperature recorded by nearby (i.e., less than 10m distance) sensors in 2s time intervals for the next 30s):

```
ON EVENT bird-detect (loc):
SELECT AVG (light),  AVG (temp), event.loc
FROM  sensors  AS s
WHERE  dist (s.loc, event.loc) < 10m
SAMPLE INTERVAL  2s  FOR  30s
```

- Events allow the system to be dormant until some external conditions occurs, instead of continually polling or blocking on an iterator waiting for some data to arrive.

Outline

- Introduction - Applications
- Data Streams
  - Data, Queries, Synopses, Projects
- RFID data
  - Data management challenges
- Wireless Sensor Networks
  - Architecture, Queries
- Time-series
  - Similarity aspects

Οι διαφάνειες που ακολουθούν βασίστηκαν στις αντίστοιχες του κ. Α. Κοτσιφάκου (ΕΚΠΑ), τον οποίο και ευχαριστούμε θερμά.

Τι είναι χρονοσειρά (time series)

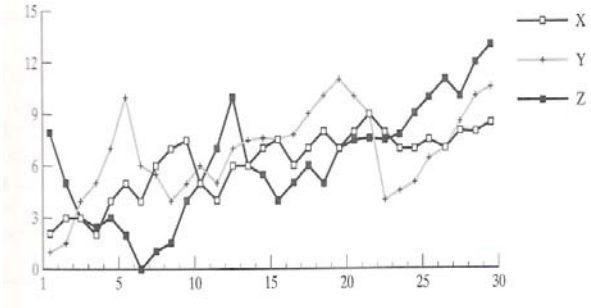
- Ένας τρόπος απεικόνισης πολύπλοκων αντικειμένων δεδομένων.
- Ορισμός: Δοθέντος ενός χαρακτηριστικού, A, χρονοσειρά (time series) είναι ένα σύνολο n τιμών:

$$\{< t_1, a_1 >, < t_2, a_2 >, \ldots, < t_n, a_n >\}$$

- Εδώ υπάρχουν n χρονικές τιμές και κάθε μία αντιστοιχίζεται σε μία τιμή του A. Συχνά οι τιμές αναγνωρίζονται για συγκεκριμένα καλά προσδιορισμένα σημεία στο χρόνο, οπότε μπορούμε να τις δούμε ως ένα διάνυσμα: $< a_1, a_2, \ldots, a_n >$

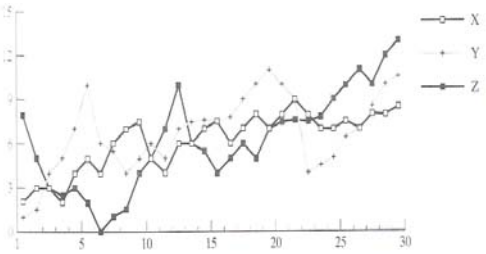
Οπτικοποίηση χρονοσειρών

- Διάγραμμα χρονοσειρών



- Παραδείγματα:
  - Δείκτες αποθεμάτων, ποσότητα πωλήσεων προϊόντων, τηλεπικοινωνιακά δεδομένα, ιατρικά μονοδιάστατα σήματα, ακολουθίες περιβαλλοντικών μετρήσεων.

Ομοιότητα μεταξύ χρονοσειρών



- Ερώτηση: η X είναι πιο όμοια με τη Y ή με τη Z;
- Αλγόριθμοι – μετρικές απόστασης για την ομοιότητα χρονοσειρών:
  - Ευκλείδεια Απόσταση (Euclidean Distance - ED)
  - Δυναμική Χρονική Στρέβλωση (Dynamic Time Warping - DTW)
  - Μακρύτερη Κοινή Υποακολουθία (Longest Common Sub-Sequence - LCSS)
  - …

Ευκλείδεια Απόσταση – ED

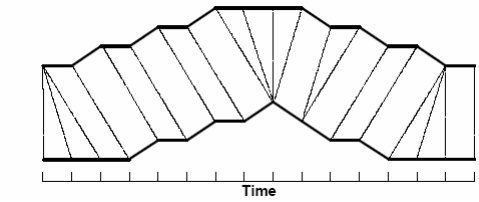
- Έστω X και Y είναι χρονοσειρές μήκους n:

$$X = x_1, x_2 ... x_j, ..., x_n \qquad Y = y_1, y_2 ... y_j, ..., y_n$$

- (απλή) Ευκλείδεια Απόσταση $ED(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- Πλεονέκτημα: απλή μετρική απόστασης
- Μειονεκτήματα
  - Ευαισθησία στις απομακρυσμένες τιμές (outliers)
  - Διαφορετικοί παράγοντες κλίμακας (scale factors)
  - Διαφορετικοί παράγοντες δειγματοληψίας (sampling factors)

Δυναμική Χρονική Στρέβλωση – DTW

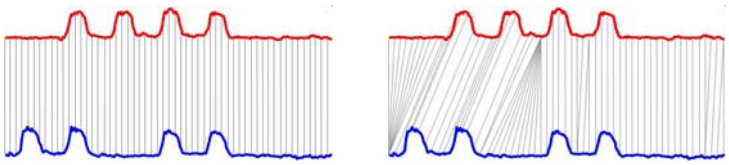
- Τεχνική που βρίσκει την **καλύτερη δυνατή ευθυγράμμιση** μεταξύ δύο χρονοσειρών
- Επιτρέπει τη στρέβλωση μιας χρονοσειράς (επέκταση ή συρρίκνωση κατά μήκος του άξονα χρόνου)
- Παράδειγμα:



73 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Δυναμική Χρονική Στρέβλωση (συν.)

- Οπτική διαφορά ED και DTW:



- Πλεονεκτήματα:
  - Αντιμετωπίζει διαταραχές που εμφανίζονται τοπικά (που οφείλονται σε διαφορετικούς παράγοντες κλίμακας ή δειγματοληψίας)
  - Αναγνωρίζει ως όμοιες δύο χρονοσειρές που η μία είναι παραμορφωμένη εκδοχή της άλλης.
- Εφαρμογές στην αναγνώριση ομιλίας, χειρονομιών, ρομποτική, κατασκευές, ιατρική, video, audio, γραφικά, ταίριασμα εικόνων-σχημάτων.

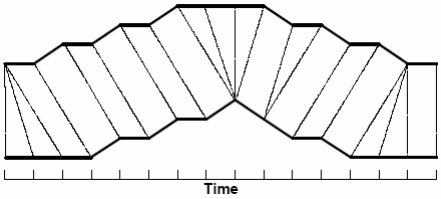
74 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Δυναμική Χρονική Στρέβλωση (συν.)

- Είσοδος: δύο χρονοσειρές

$$X = x_1, x_2, ..., x_i, ..., x_{|X|}$$

$$Y = y_1, y_2, ..., y_i, ..., y_{|Y|}$$



- Έξοδος: το **μονοπάτι στρέβλωσης**

$$P = p_1, p_2, ..., p_t, ... p_T$$

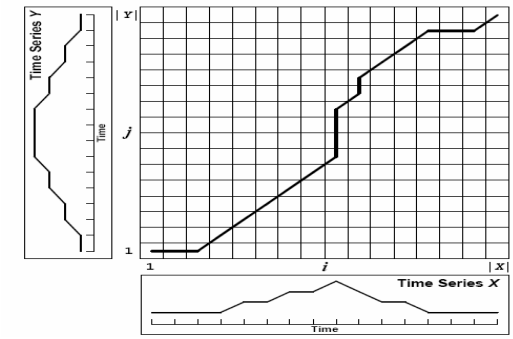
... με $\max(|X|,|Y|) \le T < |X| + |Y|$ $p_t = (i,j)_t$ $p_1 = (1,1)$ $p_T = (|X|,|Y|)$

... τέτοιο ώστε να ελαχιστοποιείται η ολική απόσταση των p_i

$$Dist(P) = \sum_{t=1}^{t=T} Dist(p_{ti}, p_{tj})$$

Δυναμική Χρονική Στρέβλωση (συν.)

- Επίλυση με αλγόριθμο δυναμικού προγραμματισμού
- Πίνακας κόστους:



Μακρύτερη Κοινή Υποακολουθία – LCSS

- Γενική ιδέα: Δύο ακολουθίες (χρονοσειρές) είναι παρόμοιες όταν παρουσιάζουν παρόμοια συμπεριφορά **για ένα μεγάλο μέρος του μήκους τους**.
- Σκοπός: να ξεπεραστεί το πρόβλημα των απομακρυσμένων τιμών (outliers) από το οποίο πάσχει η ED (και σε μικρότερο βαθμό η DTW).
- Απαιτείται μία συνάρτηση ομοιότητας $Sim_{\varepsilon,\delta}(X,Y)$

77 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Μακρύτερη Κοινή Υποακολουθία (συν.)

- Έστω $\delta > 0$ μία ακέραια σταθερά, $0 < \varepsilon < 1$ μία πραγματική σταθερά, και f μία γραμμική συνάρτηση. Δοθέντων δύο ακολουθιών X και Y, έστω ότι

$$X' = (x_{i_1}, \ldots, x_{i_l}) \quad \text{και} \quad Y' = (y_{j_1}, \ldots, y_{j_l})$$

είναι οι μακρύτερες κοινές υποακολουθίες των X και Y, για τις οποίες:

$$|i_k - j_k| \le \delta \qquad y_{j_k}/(1+\varepsilon) \le f(x_{i_k}) \le y_{j_k}(1+\varepsilon) \qquad 1 \le k \le l$$

- Ορίζουμε την **ομοιότητα των δύο ακολουθιών** ως:

$$Sim_{\varepsilon,\delta}(X,Y) = \max_{f \in L}\left\{S_{f,\varepsilon,\delta}(X,Y)\right\}$$

όπου $S_{f,\varepsilon,\delta}(X,Y) = l/n$ η ομοιότητα των δύο ακολουθιών για συγκεκριμένη γραμμική συνάρτηση f, l το μήκος της μακρύτερης κοινής υποακολουθίας και n το μήκος της μεγαλύτερης ακολουθίας.

78 ΠΑ.ΠΕΙ. – Γιάννης Θεοδωρίδης

Πολυπλοκότητες αλγορίθμων

- Ευκλείδεια Απόσταση: $O(n)$
- Δυναμική Χρονική Στρέβλωση: $O(n^2)$
- Μακρύτερη Κοινή Υποακολουθία: $O(\delta n)$

... όπου n είναι το μήκος των χρονοσειρών

Summary

- Novel Data Management issues arise when dealing with data streams
  - Data modeling, Query Processing, etc.
- Exact answers are hard to be given so approximation along with theoretical bounds suffices
- There is a lot of ongoing work that deal with streams
- RFID technology and wireless sensor networks pose new challenges
- Time-series are ubiquitous!

Reading list (streams)

- A.C. Gilbert, Y. Kotidis, S. Muthukrishnan, M. Strauss. Surfing Wavelets on Streams: One Pass Summaries for Approximate Aggregate Queries. Proc. VLDB, 2001.
- B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom. Models and Issues in Data Stream Systems. Proc. ACM SIGMOD/PODS, 2002.
- D. J. Abadi , D. Carney, U. Cetintemel , M. Cherniack , C. Convey C. , S. Lee, M. Stonebraker , N. Tatbul, S. Zdonik. Aurora: a new model and architecture for data stream management. VLDB Journal (2003).
- S. Chandrasekaran, O. Cooper, A. Deshpande, M. J. Franklin, J. M. Hellerstein, W. Hong , S. Krishnamurthy, S. R. Madden, F. Reiss, M. A. Shah. TelegraphCQ: continuous dataflow processing. VLDB Journal (2003).

Reading list (RFID)

- N. Chaudhry, D. R. Thompson, C. Thompson. RFID Technical Tutorial and Threat Modeling. Technical Report, University of Arkansas, 2005. http://csce.uark.edu/~drt/rfid.
- F. Wang, P. Liu. Temporal Management on RFID Data. Proc. VLDB, 2005.
- H. Gonzalez, J. Han, X. Li, D. Klabjan. Warehousing and Analyzing Massive RFID Data Sets. Proc. ICDE, 2006.
- D. Lin, H. G. Elmongui, E. Bertino, B. C. Ooi. Data Management in RFID Applications. Proc. DEXA, 2007.
- R. Derakhshan, N. E. Orlowska, X. Li. RFID Data Management: Challenges and Opportunities. Proc. IEEE Conf. RFID, 2007.
- Q. Z. Sheng, X. Li, S. Zeadally. Enabling Next-Generation RFID Applications: Solutions and Challenges. IEEE Computer, 41(9):21-28, September 2008.

Reading list (sensors)

- Y. Yao, J. Gehrke. The cougar approach to in-network query processing in sensor networks. SIGMOD Record (2002).
- S. R. Madden, M. J. Franklin, J. M. Hellerstein, W. Hong . TAG: a Tiny AGgregation Service for Ad-Hoc Sensor Networks. Proc. OSDI, 2002.
- S. R. Madden, M. J. Franklin, J. M. Hellerstein, W. Hong. TinyDB: an acquisitional query processing system for sensor networks. ACM Trans. Database Syst. (2005).

Reading list (time-series)

- S. Salvador, P. Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. Proc. KDD Workshop Mining Temporal and Sequential Data, 2004.
- B. Bollobas, G. Das, D. Gunopulos, H. Mannila. Time-series similarity problems and well-separated geometric sets. Proc. Computational Geometry Symposium, 1997.