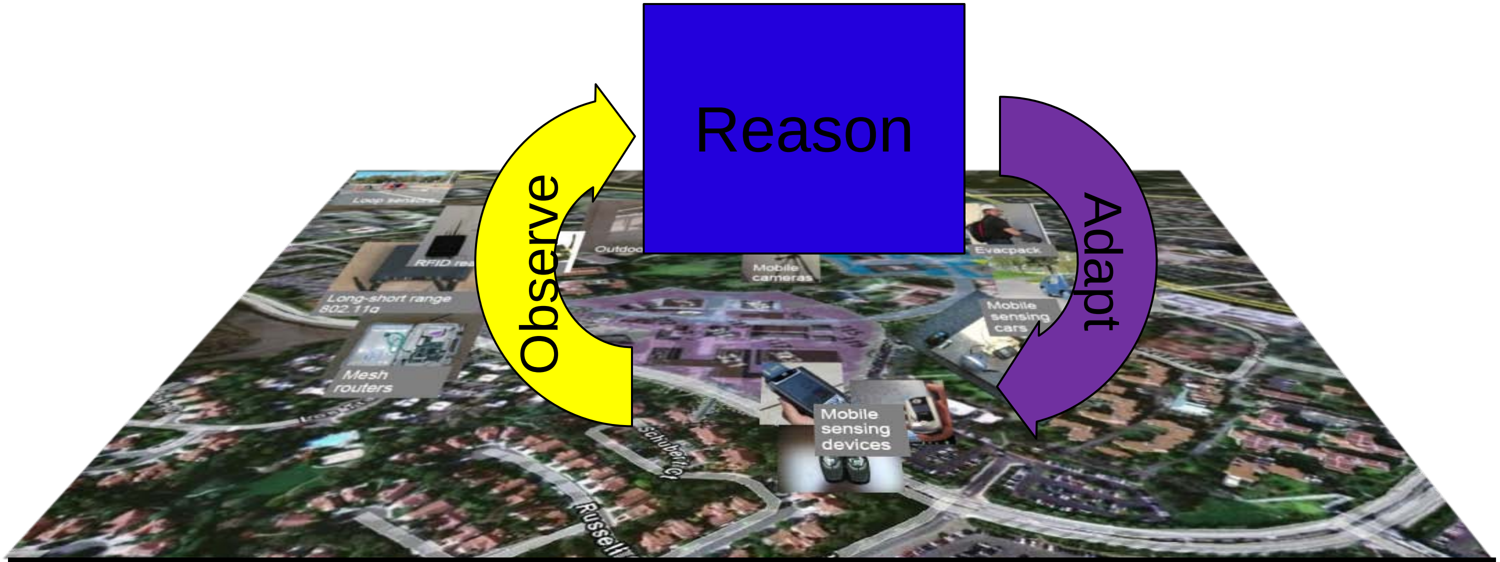
Middleware for Pervasive Spaces: Balancing Privacy and Utility

D. Massaguer, B. Hore, M. H. Diallo,
S. Mehrotra, and N. Venkatasubramanian

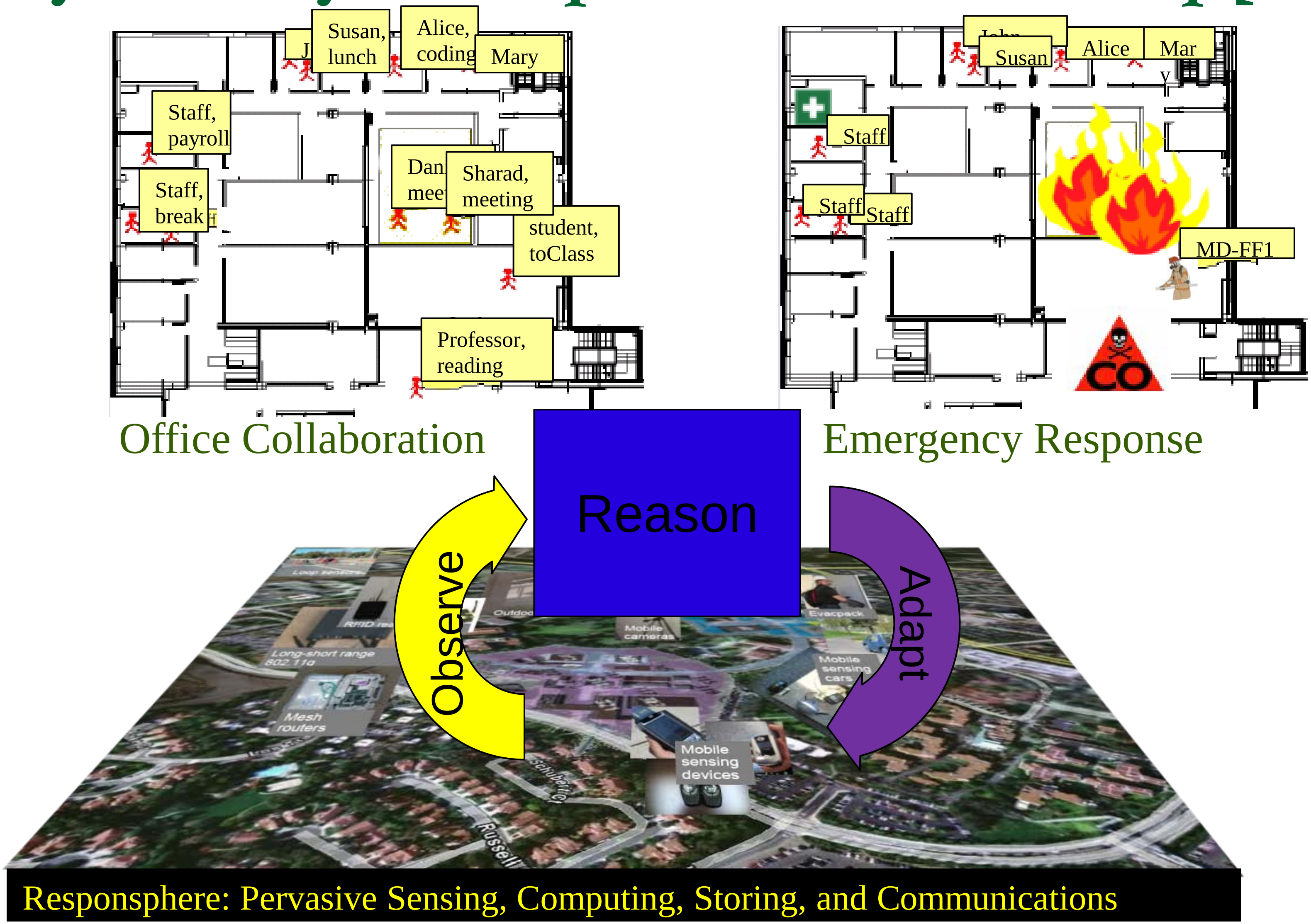
Presenter: **Daniel Massaguer**
PhD candidate
dani.massaguer@gmail.com

Cyber-Physical Spaces Control Loop[

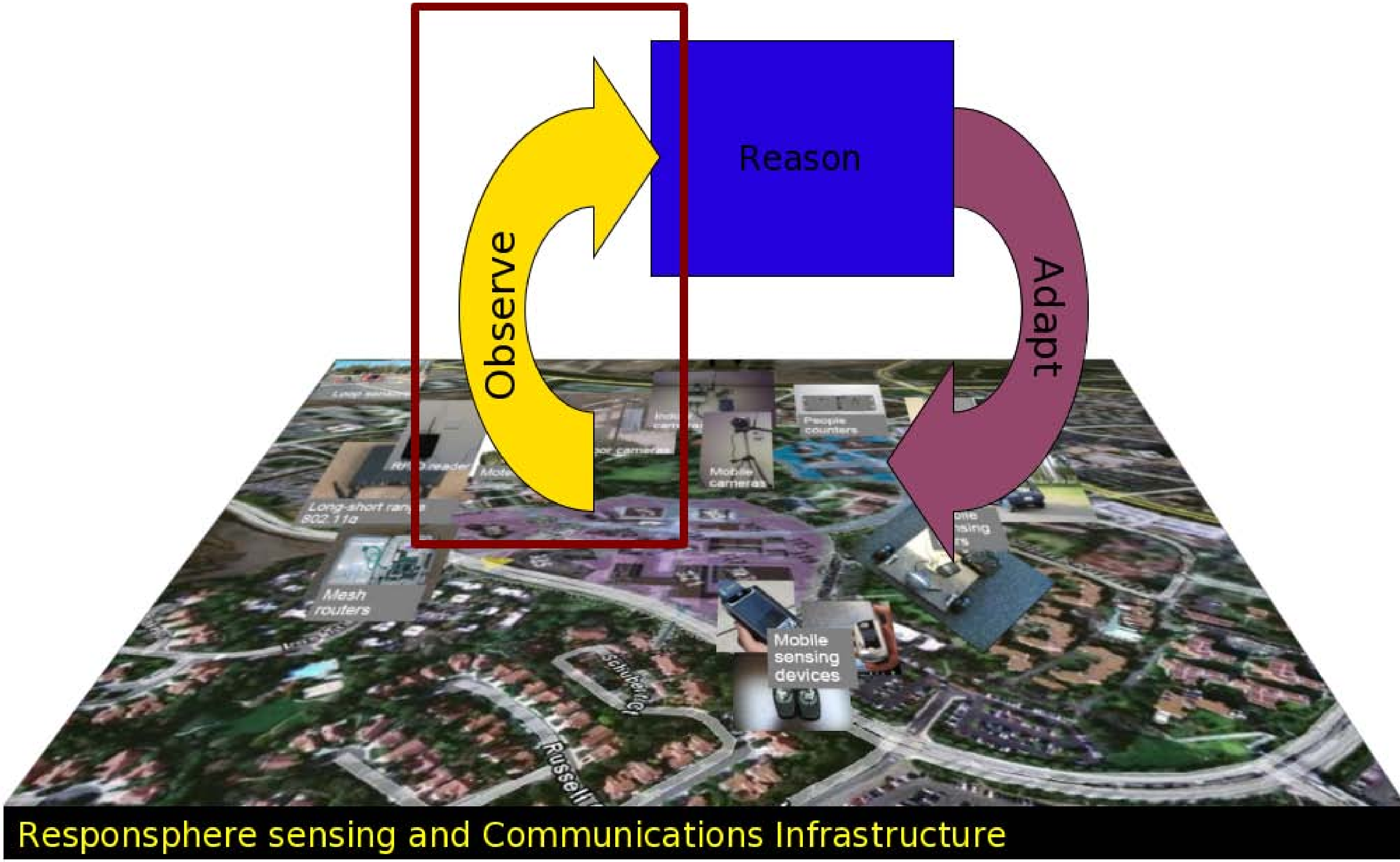


Responsphere: Pervasive Sensing, Computing, Storing, and Communications

Cyber-Physical Spaces Control Loop[

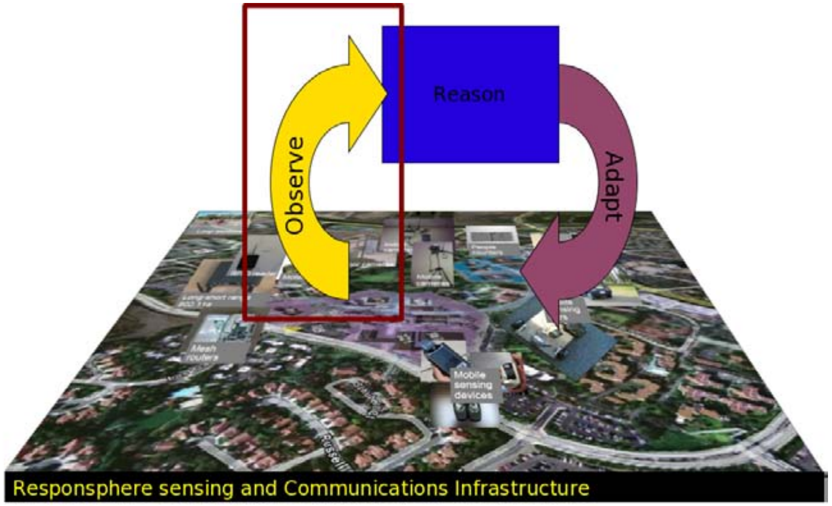


Sentient Spaces



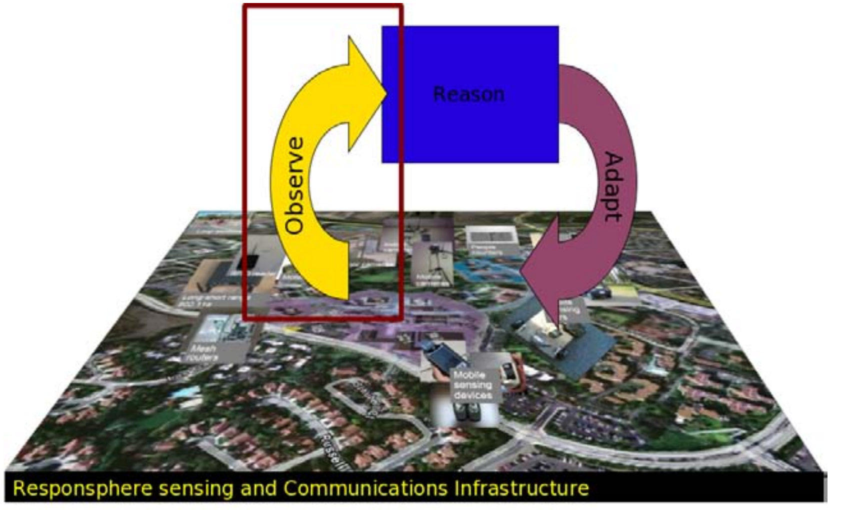
Challenges

- **Programming complexity**
  - Due to heterogeneity of
    - Sensors, computers, networks, complex event detection algorithms.
- **Shared cyber-physical infrastructure**
  - Used by several applications
  - Shared by people and their activities
  - Real-world changes non-functional requirements of observations

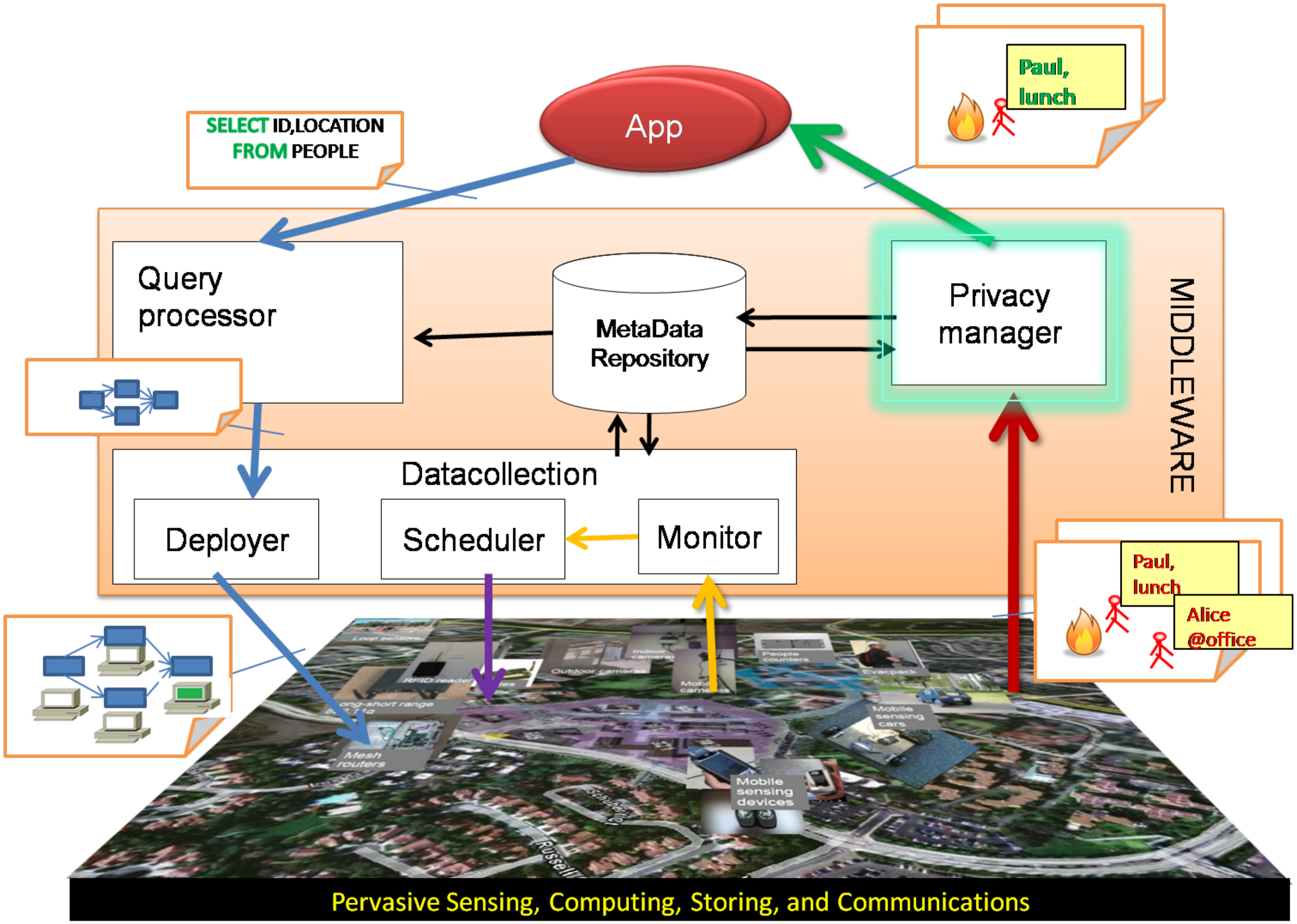


This talk

- Mechanisms to be able to release observations while protecting **privacy** of the people in the space
  [Middleware09]



Distributed and Stream Architecture



ODB.Base

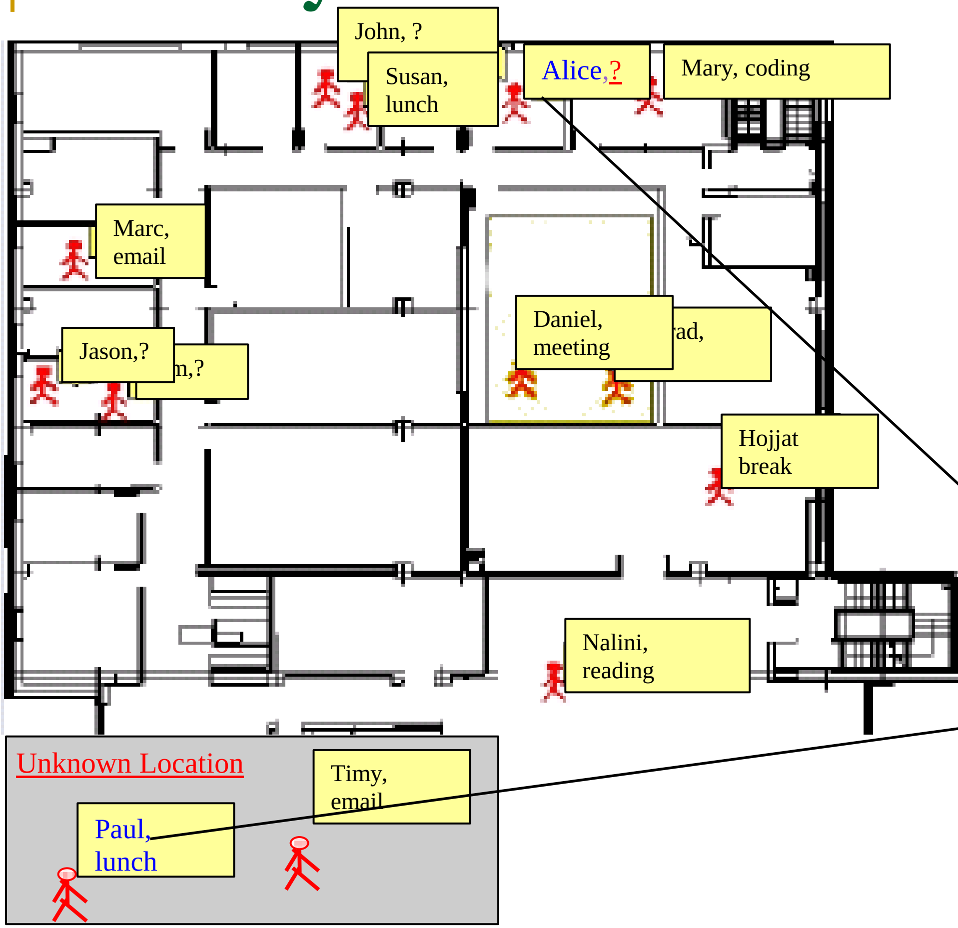
A Semantic View of the Space for Applications

ODB.Base			
ObjectId	AttName	AttValue	Time
Alice	*Location*	*Kitchen1*	10:12:50 03/04/09
Alice	*HeartRate*	*60*	10:12:54 03/04/09
John	*Location*	*ConfRoom1*	10:12:40 03/04/09
FireTeam	*Location*	*Kitchen1*	10:12:50 03/04/09
FireTeam	*Location*	*Kitchen2*	10:12:51 03/04/09

A *virtual* table that *would contain* the latest values observed

SELECT *
FROM ***ODB.Base***
WHERE ObjectId = 'Peter'
AND AttName = 'Location';

Privacy



Office monitor

Privacy challenges:

1.- Inference

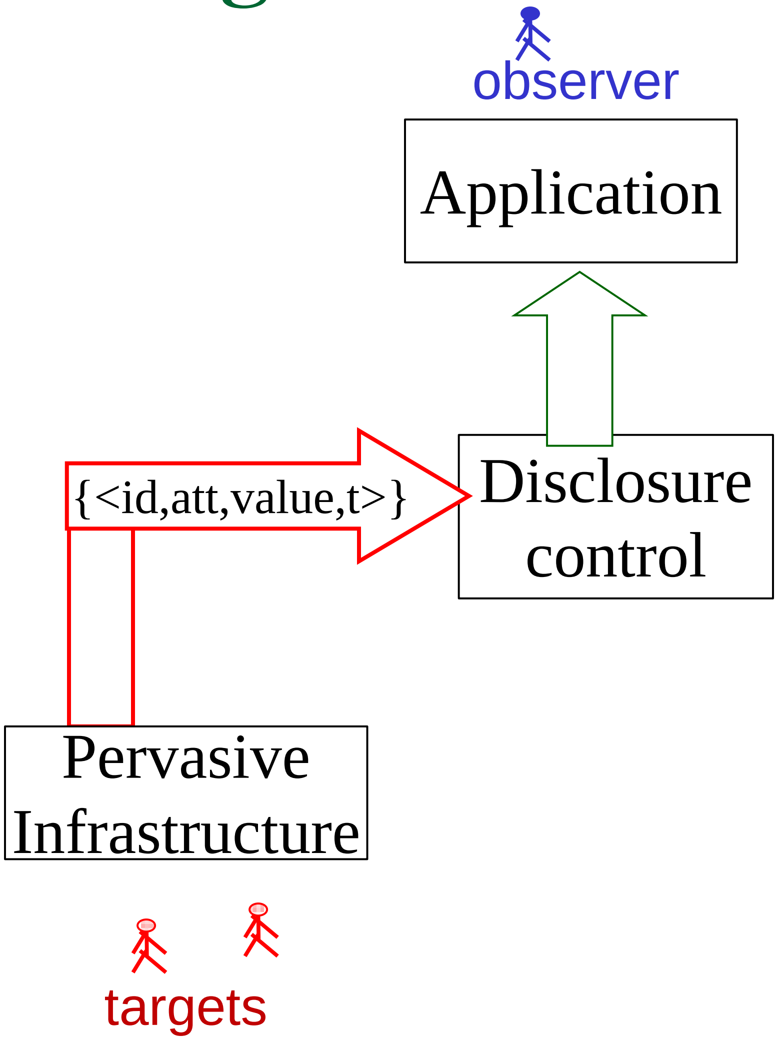
Public knowledge:
"Alice and Paul always have lunch together."

+

→ *Alice is having lunch*
→ *Paul is at Alice's office*

2.- What is privacy and how do users express it?

Our Setting

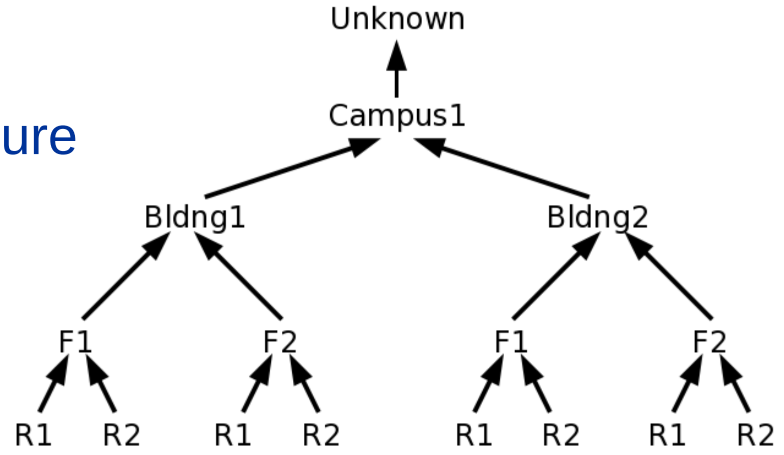

observer
Application
{<id,att,value,t>}
Disclosure control
Pervasive Infrastructure
targets

Our Approach

- Utility-based framework
  - Model privacy as negative utility of query targets
  - Model information requirements as positive utility of observers
  - Utility dynamically specified with policies and utility-elicitation mechanisms

- Compute Inferable Data
  - Total Privacy is Impossible → Closed-world approach
  - Represent background knowledge with *pDatalog* KB

- Generalize Data to reduce risk of disclosure



Privacy as Negative Utility

Intuition:

1.- “some information is more private than other
e.g., my location if I am closer to a deadline”

2.- *privateness* of information depends on consequences of *misusage*
e.g., being interrupted

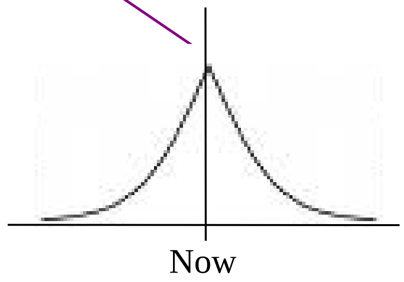
Privacy as Negative Utility

$$EU_T(y) = \underbrace{\qquad\qquad\qquad\qquad}\ * \underbrace{\qquad\qquad\qquad}\ * \ \omega(y.t)$$

Pr info being (mis)used (e.g., being interrupted)

How (un)happy if info is (mis)used

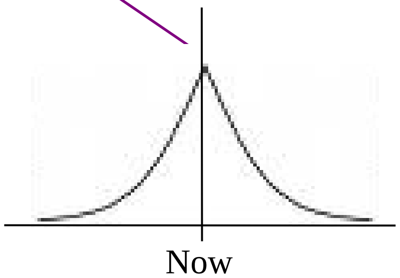
$$EU_O(y) = \overbrace{\qquad\qquad\qquad\qquad}\ * \overbrace{\qquad\qquad\qquad}$$



Privacy as Negative Utility

$$EU_T(y) = \underbrace{Pr(y \mid Y_{rel} \wedge BK) * P(y)}_{\textit{Pr info being (mis)used (e.g., being interrupted)}} * \underbrace{neg\_utility(y) * \omega(y.t)}_{\textit{How (un)happy if info is (mis)used}}$$

$$EU_O(y) = \overbrace{Pr(y \mid Y_{rel} \wedge GH) * P(y)} * \overbrace{pos\_utility(y)}$$



Y_{rel} : information released

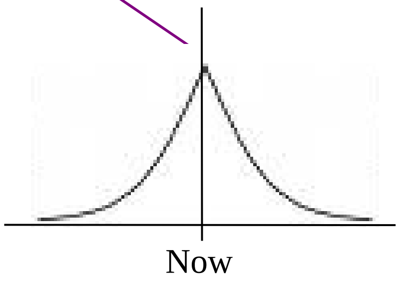
Y_{req}: information before disclosure control

Y^{i}_{rel} : independent partition in Y_{rel}

Privacy as Negative Utility

$$EU_T(y) = \underbrace{Pr(y \mid Y_{rel} \wedge BK) * P(y)}_{\textit{Pr info being (mis)used (e.g., being interrupted)}} * \underbrace{neg\_utility(y) * \omega(y.t)}_{\textit{How (un)happy if info is (mis)used}}$$

$$EU_O(y) = \overbrace{Pr(y \mid Y_{rel} \wedge GH) * P(y)} * \overbrace{pos\_utility(y)}$$



Y_{rel} : information released

Y_{req}: information before disclosure control

Y^{i}_{rel} : independent partition in Y_{rel}

$$\max_{Y^i_{rel}} EU_O(Y^i_{rel})$$

s.t.

$$min\_EU_T(Y^i_{rel}) + max\_EU_O(Y^i_{rel}) \geq 0.0$$

$$Y^i_{rel} \preceq Y^i_{req}$$

Background Knowledge

- pDatalog Knowledge Base (association rules):
  - Tuple(Alice, Location, l, t) : $p * 0.8 \leftarrow$ Tuple(Mary, Location, l, t) : p
- Feasible approach
  - Populated by admins (intended space usage) +
  - learned by system (calibration + rule mining)

Background Knowledge

- pDatalog Knowledge Base (association rules):
  - Tuple(Alice, Location, l, t) : $p * 0.8 \leftarrow$ Tuple(Mary, Location, l, t) : p
- Feasible approach
  - Populated by admins (intended space usage) +
  - learned by system (calibration + rule mining)

- Identical facts combined with MAX (i.e., worst-case inference)

Background Knowledge

- pDatalog Knowledge Base (association rules):
  - Tuple(Alice, Location, l, t) : p $*$ 0.8 $\leftarrow$ Tuple(Mary, Location, l, t) : p
- Feasible approach
  - Populated by admins (intended space usage) +
  - learned by system (calibration + rule mining)

- Identical facts combined with MAX (i.e., worst-case inference)
- Uncertainty functions (e.g., p*0.8) adhere "natural restrictions" *[pD]*
  - monotonicity ($f(x1, \ldots, xn) \leq f(y1, \ldots, yn)\ \forall i \in [1..n]\ xi \leq yi$),
  - boundedness ($f(x1, \ldots, xn) \leq xi \forall i \in [1..n]$), and
  - continuity w.r.t its arguments

➔ Inference (Rete) finishes with polynomial time *[pD][AI]*

Optimization Problem

$$\max_{Y^i_{rel}} EU_O(Y^i_{rel})$$

s.t.

$$min\_EU_T(Y^i_{rel}) + max\_EU_O(Y^i_{rel}) \geq 0.0$$

$$Y^i_{rel} \preceq Y^i_{req}$$

search space is <u>exponential = O(m^N) !</u>

Distr. Parallel Simulated Annealing

- Optimization problem is inherently parallel:
  - Independent partitions
- Execution environment is inherently distributed and parallel
  - Pool of multi-core PCs
- Need fast solution
  - Stochastic optimization

Y_{rel} =findMinIndPartitions(Y_{req}, BK_{obs})
for each ($Y^i_{rel} \in Y_{req}$)
do *n* times in parallel
    SimulatedAnnealing(Y^i_{rel})
enddo
endfor

Distr. Parallel Simulated Annealing

Configuration

accept(s, T) *=exp(−ΔE/T)*

$$E(Y^j_{rel}) = \rho\left(\frac{\sum_{y_r \in Y^j_{rel}} EU_O(y_r)}{|Y^j_{rel}|}\right) +$$

$$\frac{1}{\rho}(Nat(-\max_{y_r \in Y^j_{rel}} (EU_O(y_r) * \omega(y_r.t)) -$$

$$\min_{y_d \in Y^j_{derived})} (EU_T(y_d) * \omega(y_d.t))$$

$\rho = 10^{-r}$, with $r \geq 1$

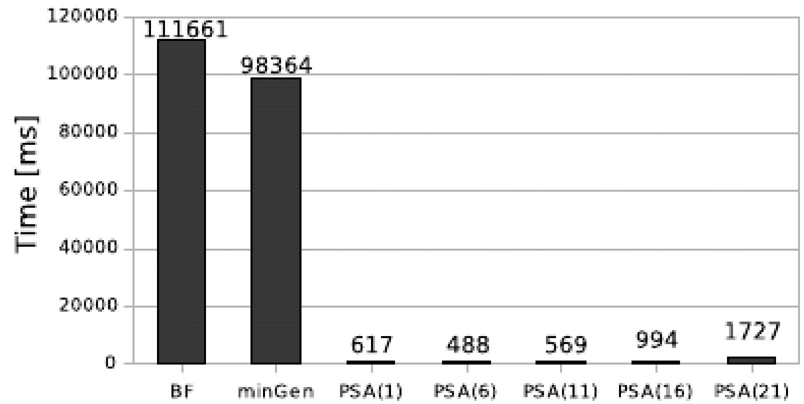
T(0) *= 1/ ρ*
Temperature Schedule: *T(k) = δ * T(k-1)*
Same temperature: *N*max(m)/2* iterations
Termination:
E == 0.0, *T(i)== δ*, or Feasible Solution.
δ = ρ = 0.1

Time complexity is polynomial

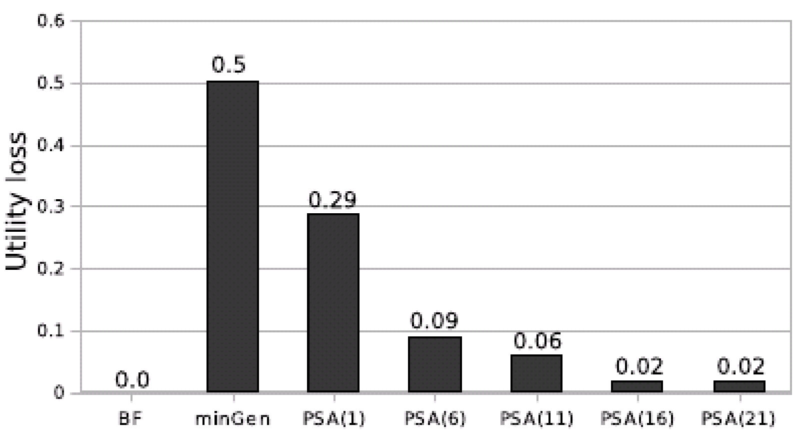
$$O((rf^c + r^2 f) + (Nrf^c + N^2))$$

r: rules in knowledge base
f: facts in knowledge base
N: queries

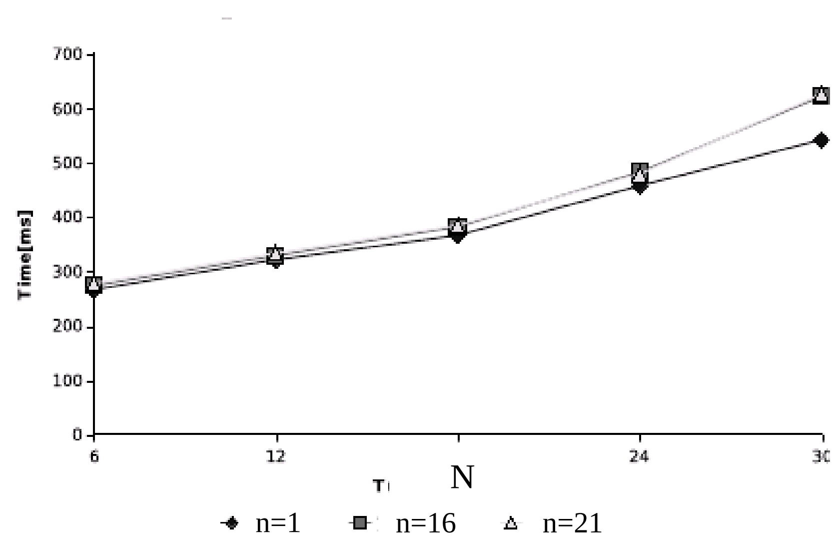
Results



Good time



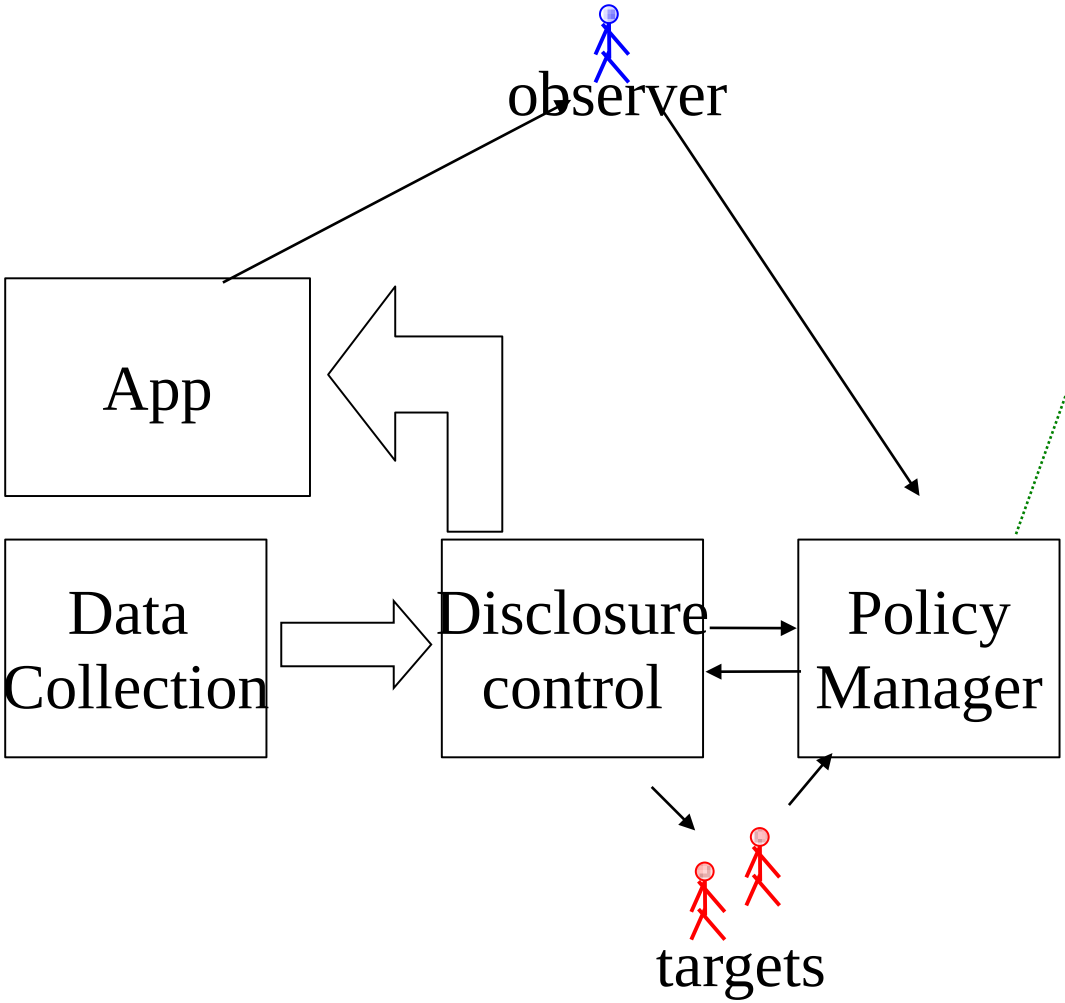
Good answer



Polynomial w.r.t. N

Specifying Privacy and Utility

A Control Loop



Target	
Labels	**Utilities**
Extremely Sensitive	-1.00
Very Sensitive	-0.75
Sensitive	-0.50
Somewhat Sensitive	-0.25
Not Sensitive	0.00
Observer	
Labels	**Utilities**
Don't Care	0.00
Information Curiosity	0.25
Information Useful	0.50
Information Needed	0.75
Always Needed	0.99

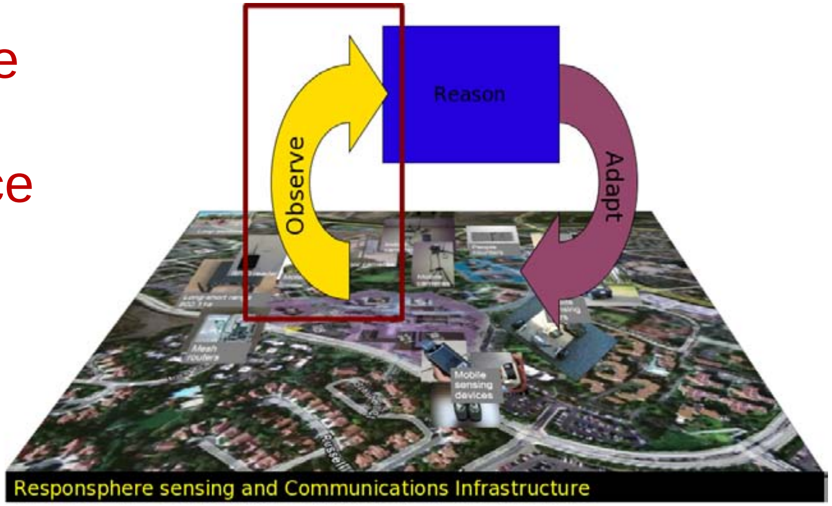
+

Preference network

[COPnet]

Summary and Future Work

- Summary of Contributions
  - Mechanisms to be able to release observations while protecting **privacy** of the people in the space
- Future work
  - Generalization of entity
  - Efficient storage of background knowledge



Acknowledgments

Thanks to

SATware research group and my PhD committee for their valuable input and coauthoring papers and code, which contributed in making my research come to live.

Special thanks to Roberto Gamboni, Jay Lickfett, Jonathan Cristoforetti, Alessandro Ghigi, Francisco Servant, Ronen Vaisenberg, Shengyue Ji, Hojjat Jafarpour, Minyoung Kim, Jooyoung Park, Kyoungwoo Lee, Mamadou Diallo, Bijit Hore, Haynes Mathew George, Chris Davison, Jon Hutchins, Utz Westermann, Gloria Mark, Ramesh Jain, Sharad Mehrotra, Don Patterson, and Nalini Venkatasubramanian.

Thanks also to all the anonymous reviewers of the papers in which the work here presented was first explained.

This work has been partially supported by the NSF under award Numbers 0331707, 0331690, and 0403433.

Thank you

dani.massaguer@gmail.com

Q&A

Extra Slides

Privacy is impossible

Maximum disclosure risk for sentient spaces:(adapted from data publishing [Martin07][skyline]):

$$\max_{y \in \text{Private},\ \forall BK^k \in \text{PL-Horn}} \Pr(y \mid Y_{rel} \wedge BK^k)$$

Privacy is impossible

Maximum disclosure risk for sentient spaces:(adapted from data publishing [Martin07][skyline]):

$$\max_{y \in Private,\ \forall BK^k \in PL\text{-}Horn} \Pr(y \mid Y_{rel} \wedge BK^k) = \mathbf{1.0} \quad \mathbf{k>0}$$

That is, privacy-preservation cannot be guaranteed.

Privacy is impossible

Maximum disclosure risk for sentient spaces:(adapted from data publishing [Martin07][skyline]):

$$\max_{y \in \text{Private},\ \forall BK^k \in \text{PL-Horn}} \Pr(y \mid Y_{rel} \wedge BK^k) = \mathbf{1.0} \quad \mathbf{k>0}$$

That is, privacy-preservation cannot be guaranteed.

PROOF:

Since $\exists\, y_0 : 1.0 \in Y_{rel}$

in the worst-case, the adversarial BK has the rule

$y_0 \rightarrow y$

$\rightarrow \Pr(y \mid Y_{rel} \wedge BK) = 1.0$. *QED.*

→ We need to explicitly represent realistic rules in a knowledge base (KB).

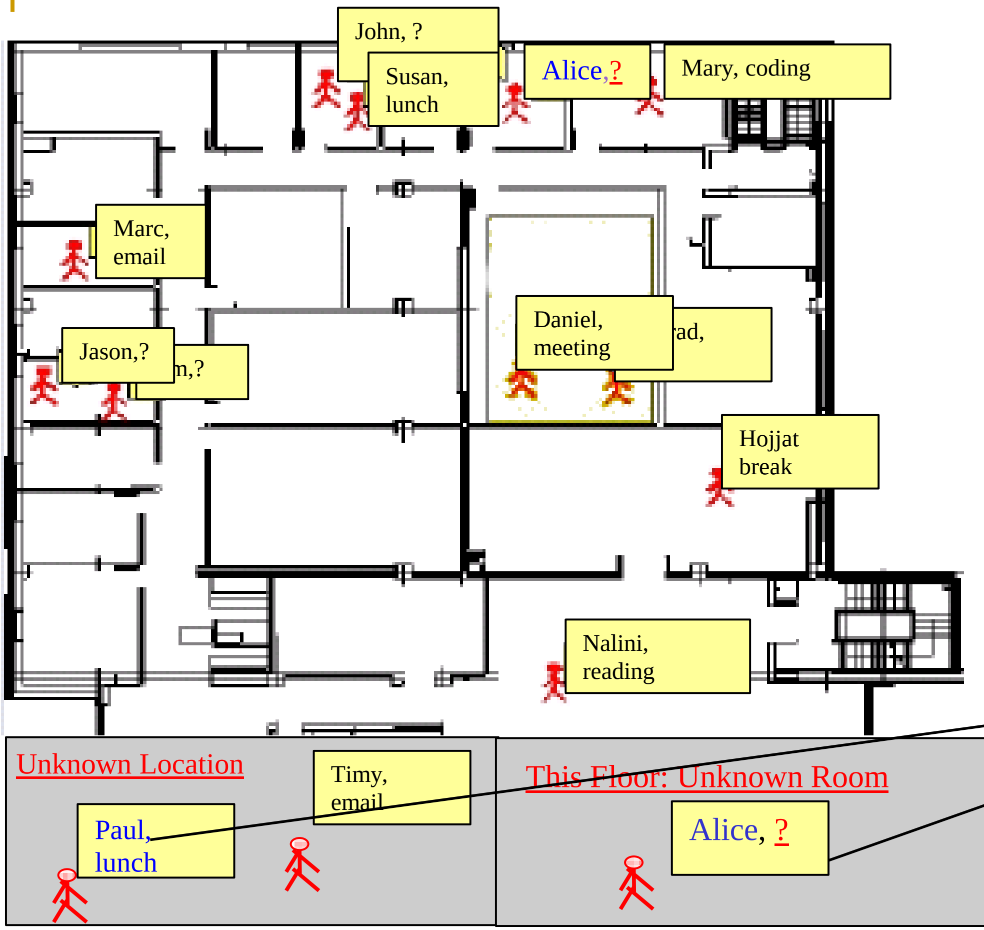
KB can be learned (e.g., traditional rule mining) [Middleware09]

context

$$ctxt(u) = \{y =< id, att, v, t > | id = u \text{ or } id = benignObj\}$$

```
Function SimulatedAnnealing(Y irel)
Y jrel = Y irel.neighbor()
Y rel = max(Y jrel,Y irel)
T = T(0)
While(!terminate)
  if(accept(Y jrel, T))
    if(Y jrel.energy < Y rel.energy)
       Y rel = Y jrel
    endif
  endif
  if(!change temperature)
    Y jrel = Y jrel.neighbor()
  else
    T.decrease();
   if(!terminate)
       Y jrel = Y jrel.neighbor()
   endif
  endif
Endwhile
Return Y
rel
endfunctionx
```

Our Approach: Exploit Generalization Hierarchies



Office monitor

"Alice and Paul always have lunch together."

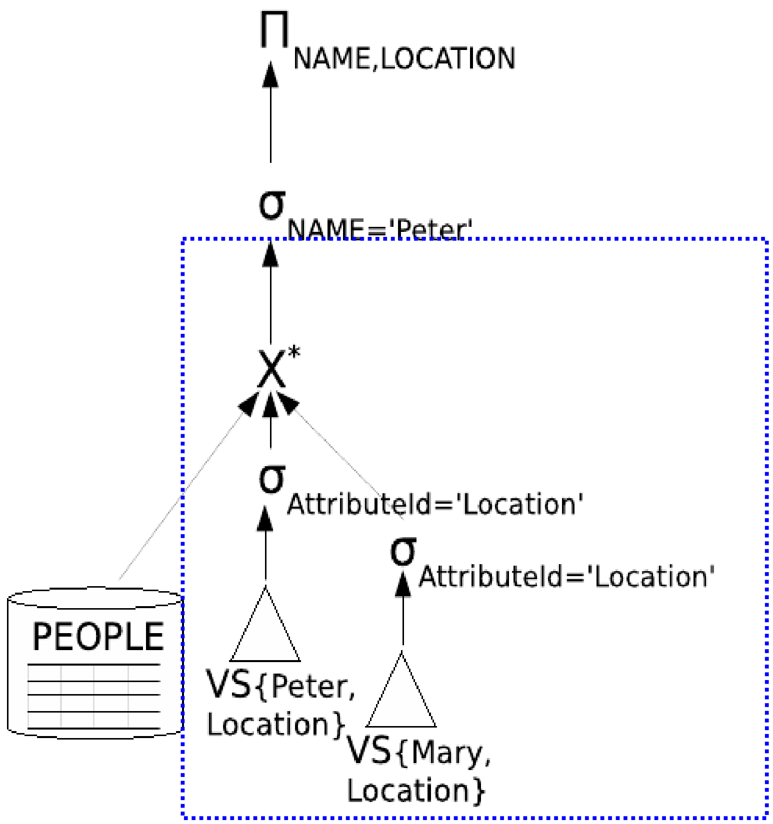
+

→ Alice is having lunch
→ Paul is at Alice's office

Query Plan

Abstraction

SELECT NAME,LOCATION
FROM PEOPLE
WHERE NAME=’PETER’;



Privacy: Existing Work

Traditional access control

Summary: Access is denied or granted according to policies [P3P][Rei][PaWS]
Specific Limitations: Inference not taken into account.

Pervasive/ Ubicomp

Summary: Not trusting other devices: hop-to-hop anonymous routing [MIST-Gaia], each device computes its location [Cricket][PlaceLab]
Specific Limitations: Data is assumed not useful beyond the client's device, data recipient is not another user.

Data publishing

Summary: Focus is on anonymization of statistical databases [k-anonymization][l-diversity][worst-case-bk].
Specific Limitations: Mechanisms are for aggregated static data. With concrete data (i.e., with prob=1.0), analyses w/o explicit background knowledge representation are not applicable. Privacy is defined as a binary concept: data is either public or private

Defining privacy

Summary: Privacy is subjective, ever-changing [Altman][Dourish], depends on observer, target, context and purpose,
Information (**mis)use** is a primary concern [PAL],