Challenges in Human-centric Sensor Networks

Invited Speaker: Professor Tarek Abdelzaher

University of Illinois at Urbana Champaign

http://www.artist-embedded.org/

Challenges in Human-centric Sensor Networks

Acknowledgements: This work was done in collaboration with IBM Research and Microsoft, and Sponsored in part by NSF, ARL, and IBM.

A Little About Me

- M.Sc. In Automatic Control, Ain Shams Univ., Egypt, 1994
- Ph.D. in Computer Science, University of Michigan, 1999
    - QoS Adaptation in Real-time Systems (Advisor: Kang Shin)
- University of Virginia (Dept. of CS), 1999-2005
- University of Illinois at Urbana Champaign, 2005-now
    - Cyber-physical Computing Group
    - ~ 10 Ph.D. students
    - ~ 1-2 Postdocs and R&D Associates
    - Part of the Embedded Systems Labs (Lead: Lui Sha, ~ 30 students)
- Research interests: Computing systems that interact with the physical world (and with people).

An Evolving Research Landscape

- Embedded systems (avionics, robotics, automotive, factory automation, …)

  to:

- Cyber-physical systems (large embedded systems of systems: coupling, complexity, composition challenges, …)

  to:

- Cyber-physical systems *in social spaces* (CPS that massively interact with humans: quality of information, uncertainty, robustness, non-determinism, “network science”)

The Rise of Social Sensing

People

Sensors

Analytics

Data

Future Applications

Social Sensing:
A Confluence of Three Trends

Mass Dissemination Media

twitter

flickr

facebook

YouTube

twitpic
share photos on twitter

Connectivity

wimax

Game Consoles on Internet

Cars on Internet

Cell-phones

Sensors

Glucose monitor

GPS

Sportsware

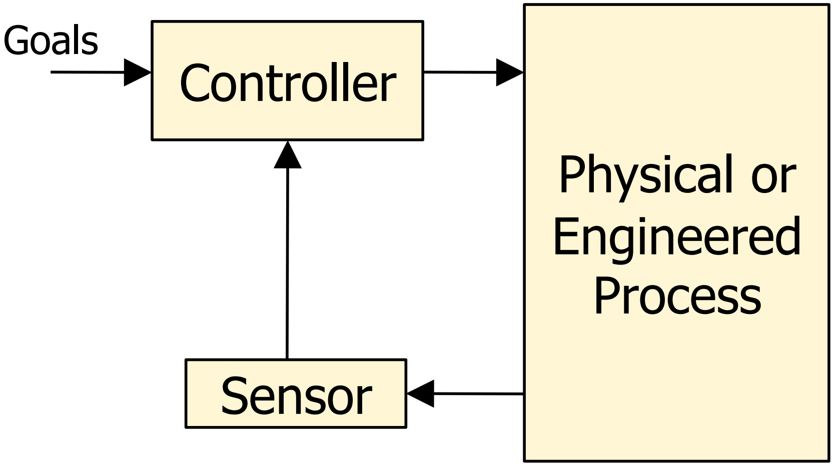
Pulse oximeter

Smart Meter

Cyber-Physical Computing:
From Engineered to Social Systems

Classical

Control, automation



Goals

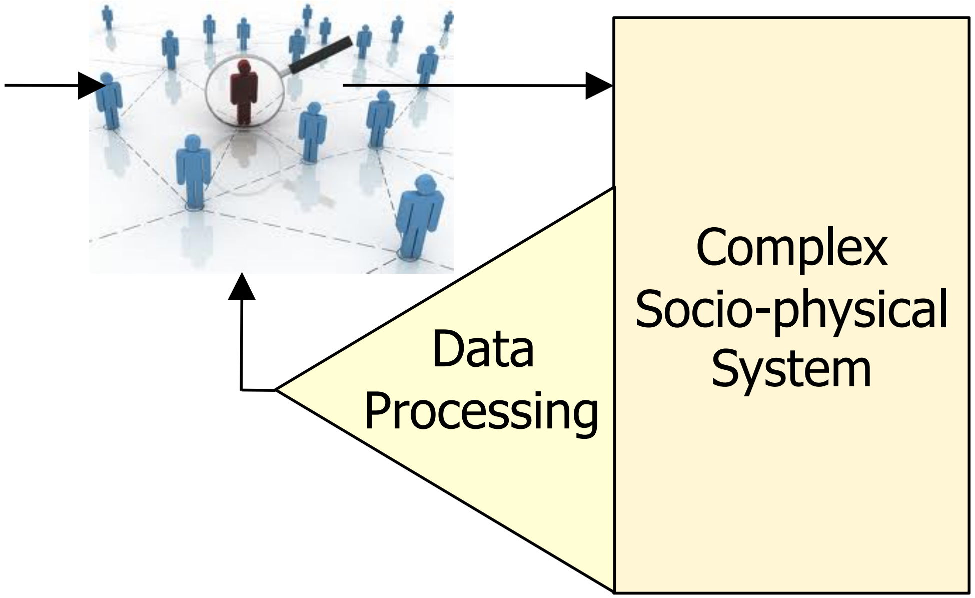
Controller

Physical or Engineered Process

Sensor

Emerging

Data processing, decision making



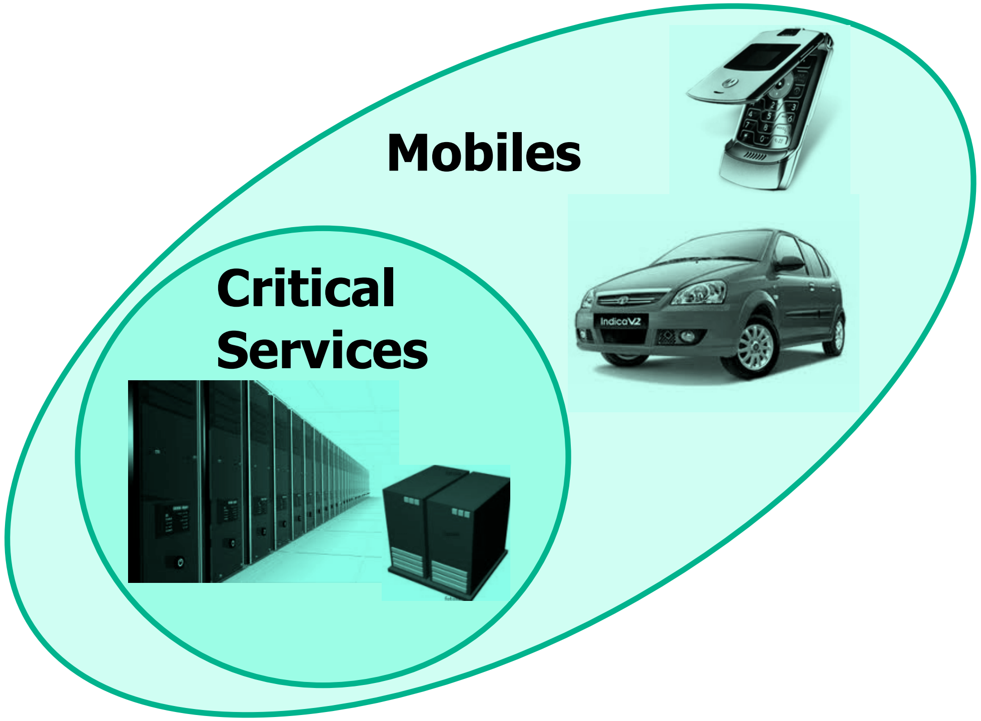
Complex Socio-physical System

Data Processing

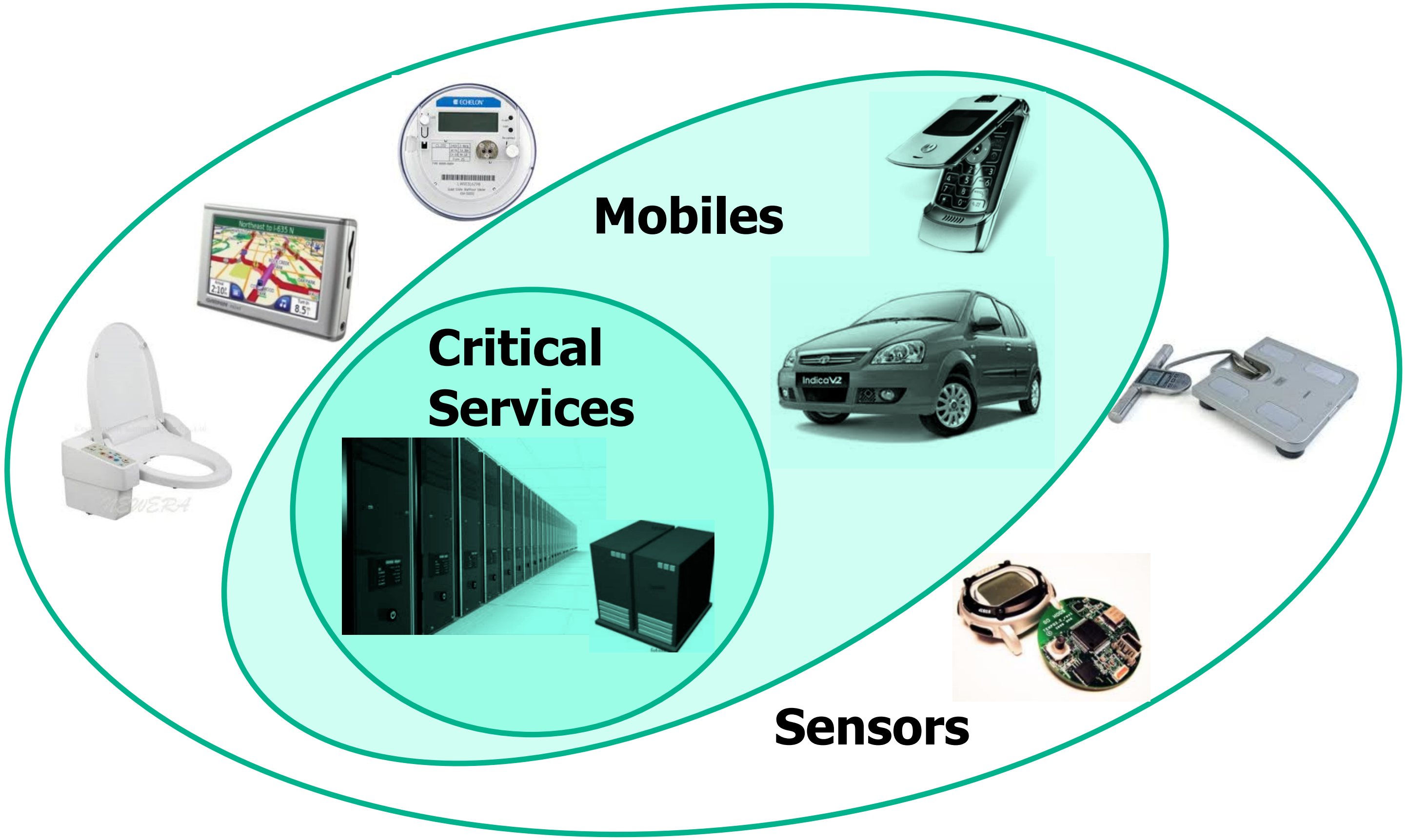
An Architecture for Social Cyber-Physical Applications

Critical
Services

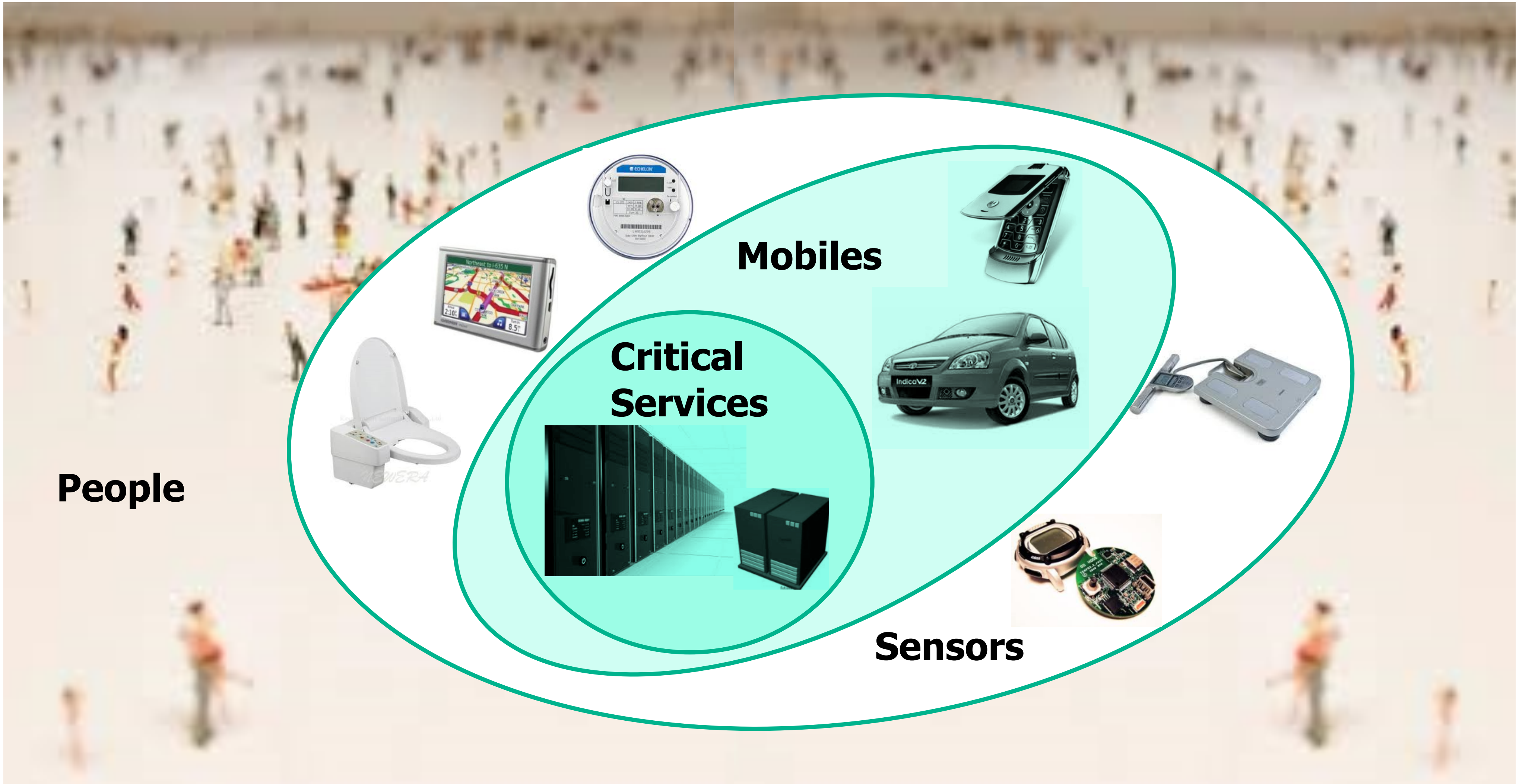
An Architecture for Social Cyber-Physical Applications



An Architecture for Social Cyber-Physical Applications

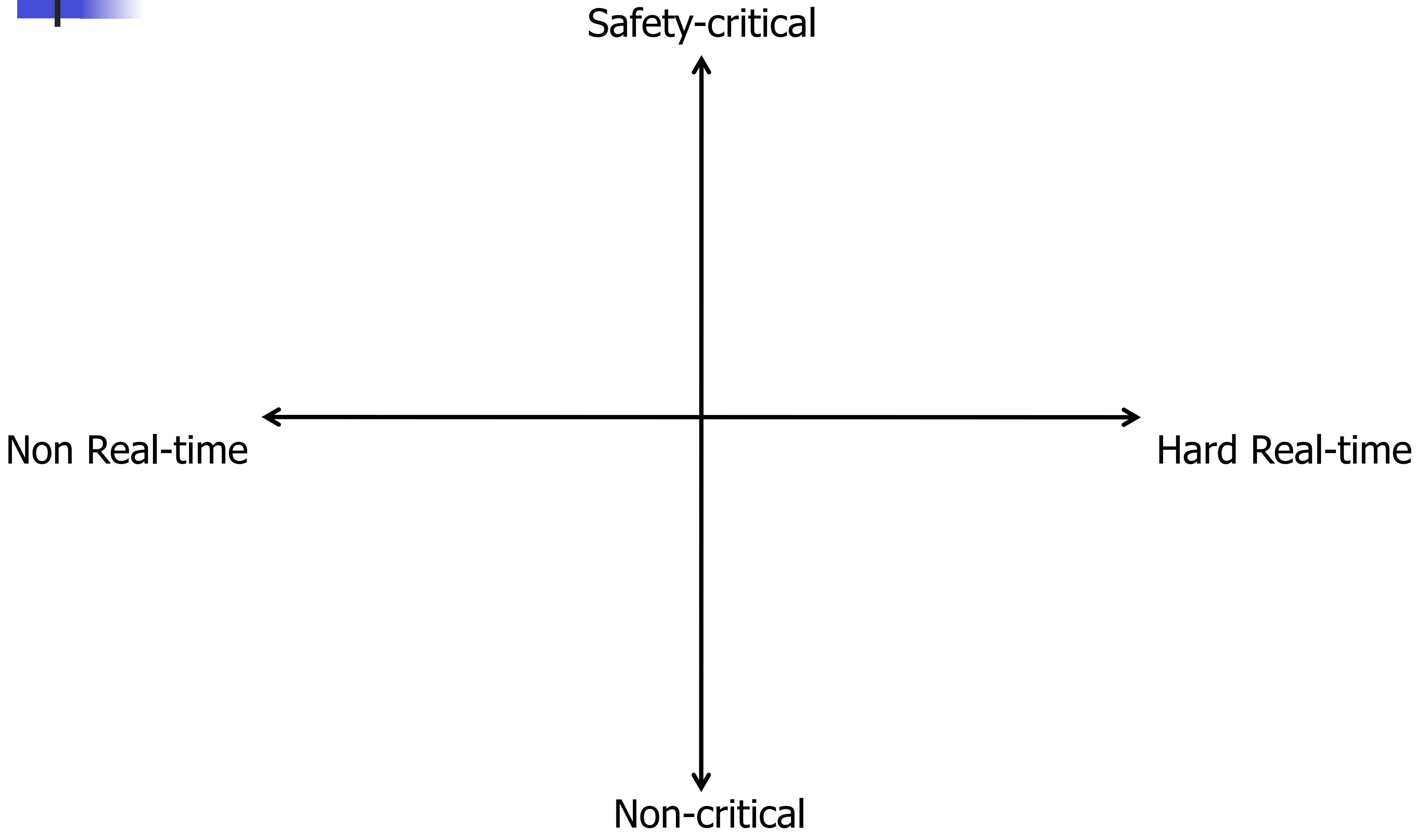


An Architecture for Social Cyber-Physical Applications



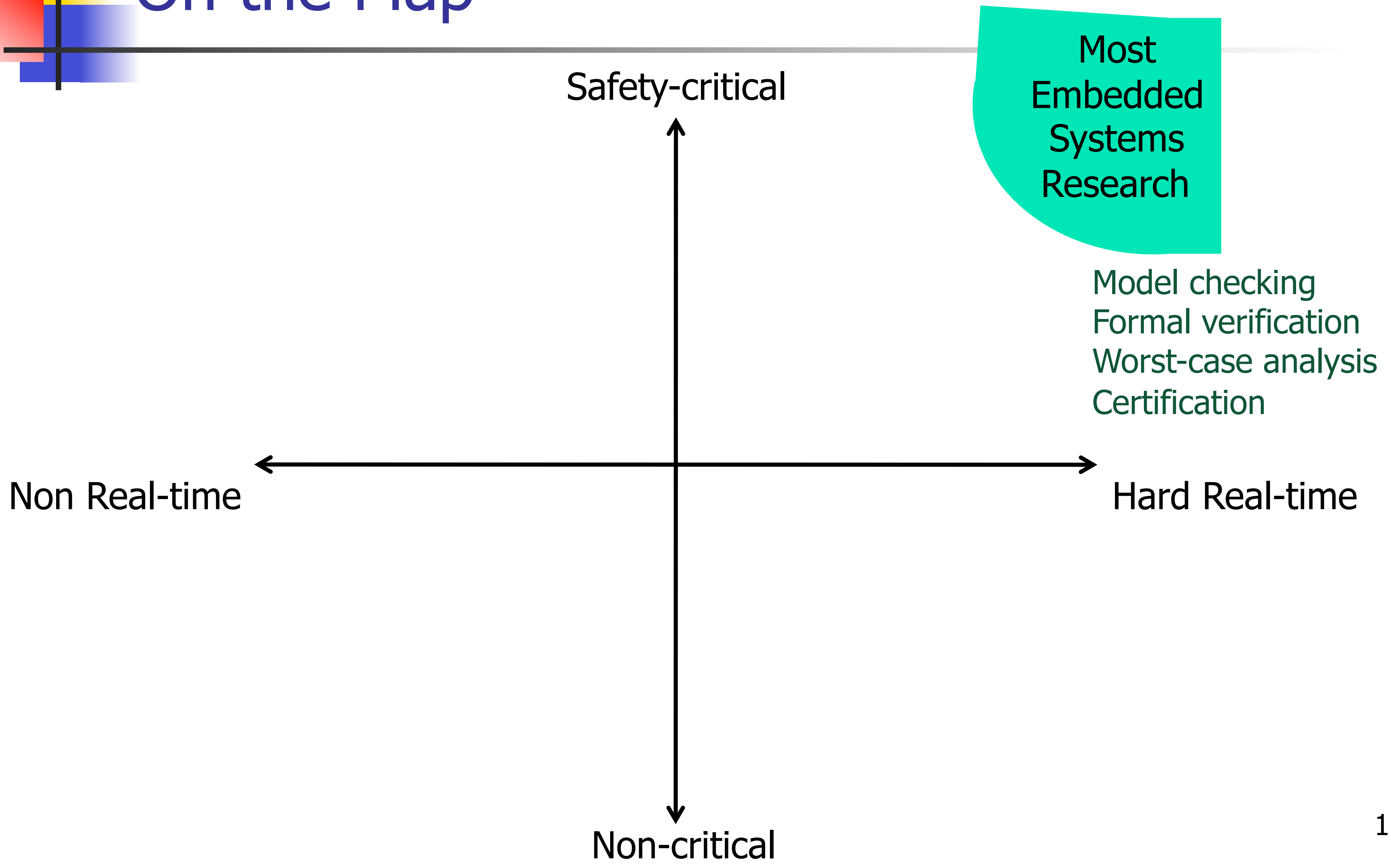
Social Cyber-Physical Systems

On the Map



Social Cyber-Physical Systems

On the Map



Safety-critical

Most Embedded Systems Research

Model checking
Formal verification
Worst-case analysis
Certification

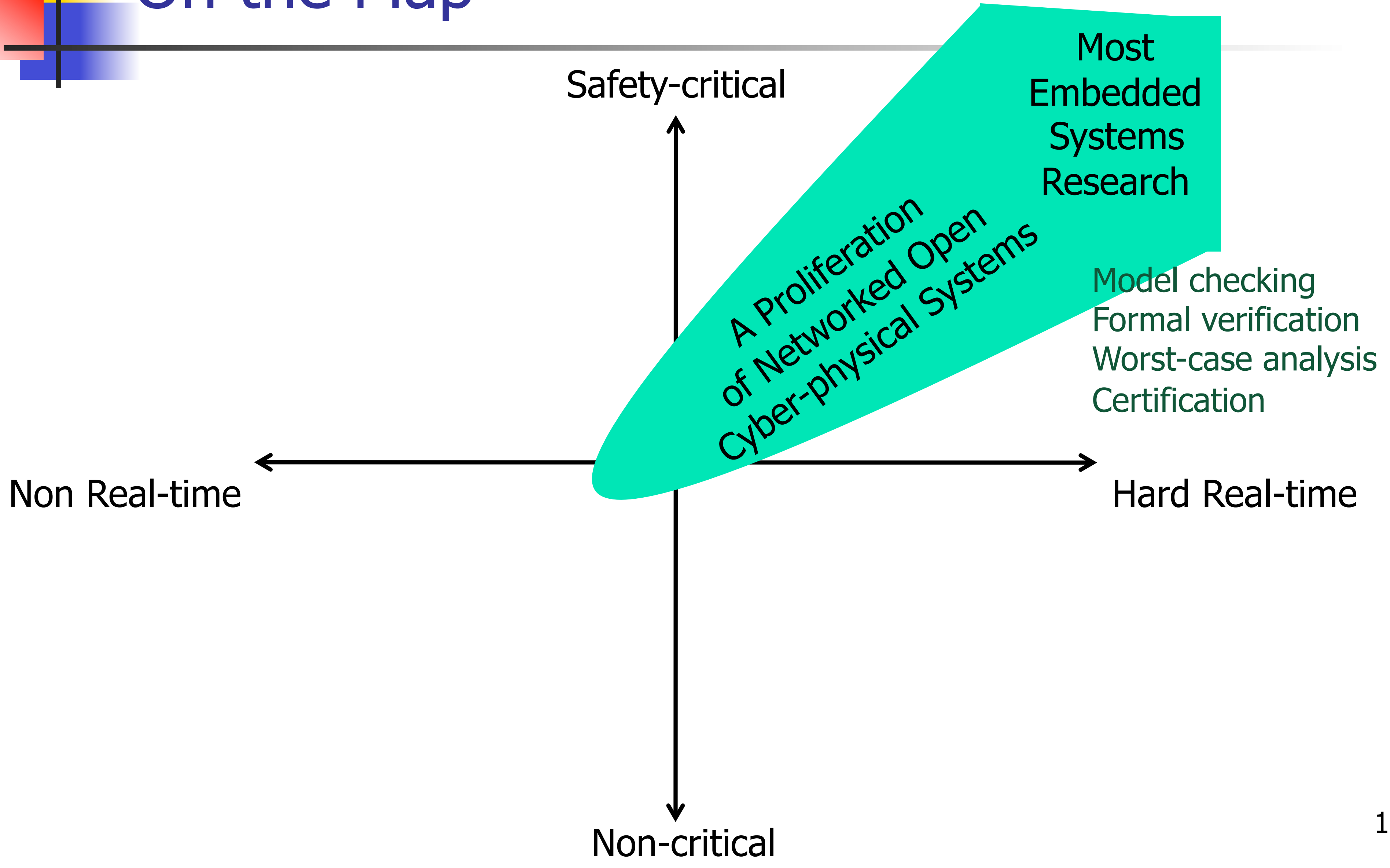
Non Real-time

Hard Real-time

Non-critical

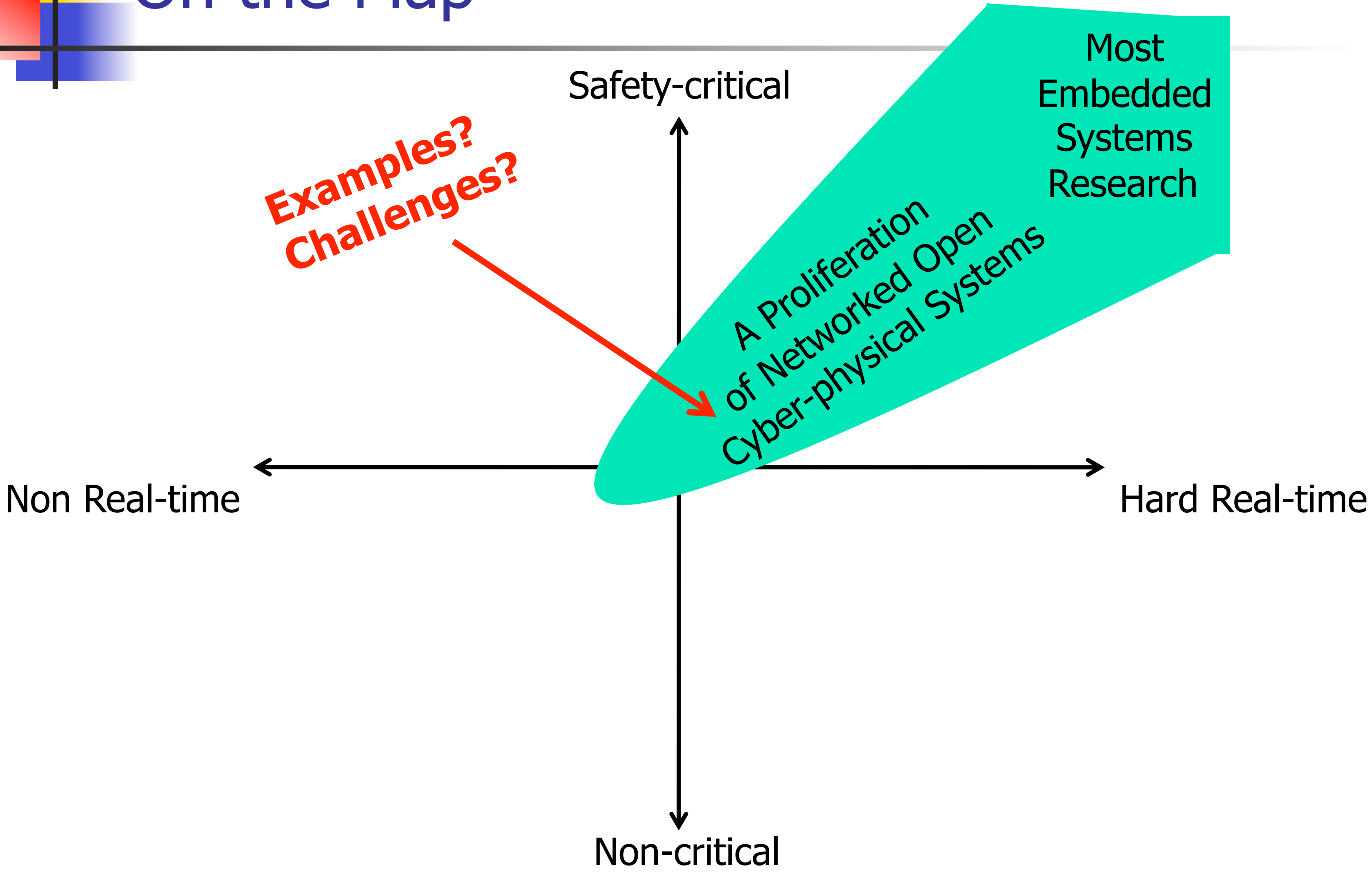
Social Cyber-Physical Systems

On the Map



Social Cyber-Physical Systems

On the Map


Safety-critical
Examples?
Challenges?
Most Embedded Systems Research
A Proliferation of Networked Open Cyber-physical Systems
Non Real-time
Hard Real-time
Non-critical

Social Sensing

Powered by Proliferation of Common Sensors

- WWW → a gathering place around topics of mutual interest
- Social sensing web → a gathering place around mutually interesting data pools (and derived info)
  - Feng Zhao: MSR Sensor Map
  - Dave Clark: The future Internet will link more sensors and embedded devices that traditional hosts
  - Van Jacobson: Named-data networking paradigm (we use the Internet as an information source not a communication medium)

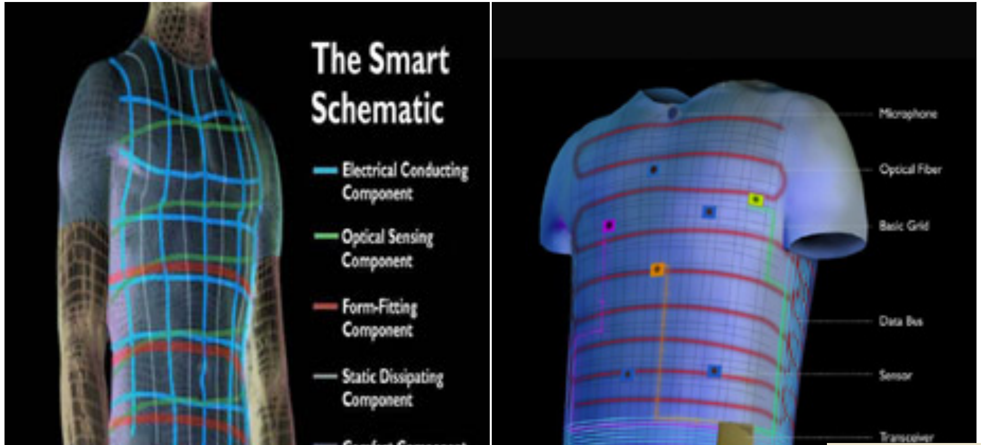
Personal Sensing

Human Activity Monitoring

Smart Jacket for Human Activity Monitoring

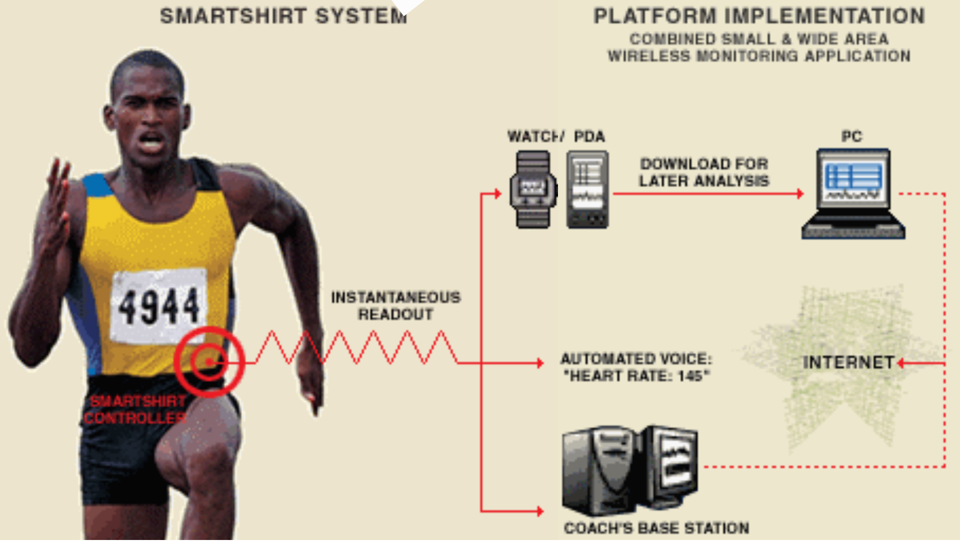
Personal Sensing

Sports and Entertainment



GPS

Nike -iPod



http://www.sensatex.com

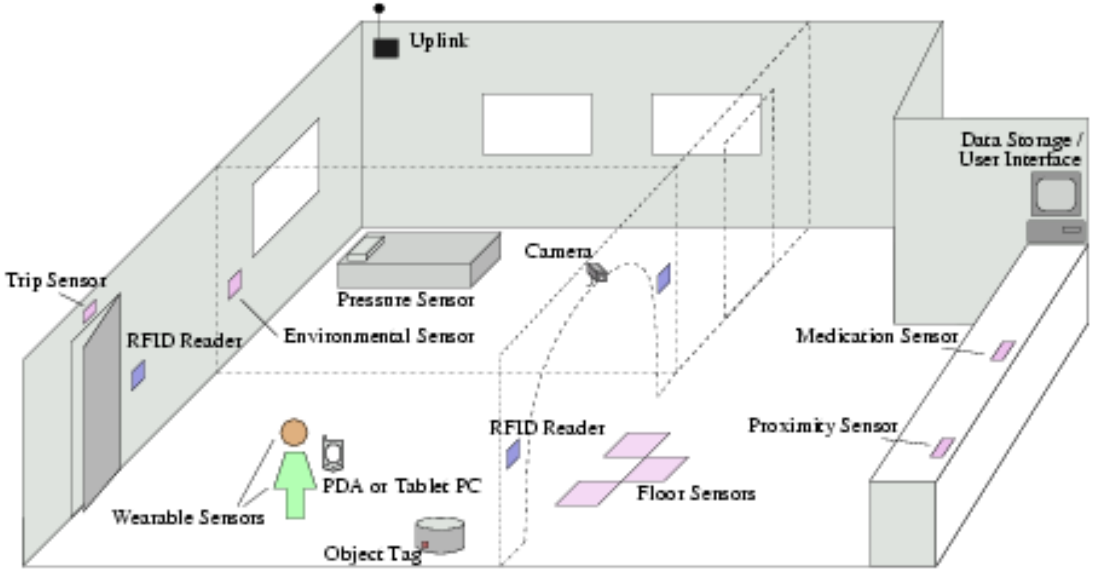
Spot

Wii

Personal Sensing

Smart Spaces

- Instrumented spaces for "aging in place"
  - Reduce cost of long-term care by facilitating independent living



AlarmNet testbed at the University of Virginia

Personal Sensing

Health and Wellness (HealthVault)

- HealthVault (Microsoft): Fitness and biometric monitoring devices automatically upload data to a central repository for safekeeping and analysis
  - A significant number of medical device vendors announced devices compatible with healthVault

Precision weight scale

Glucose monitor

Blood pressure monitor

Pedometer

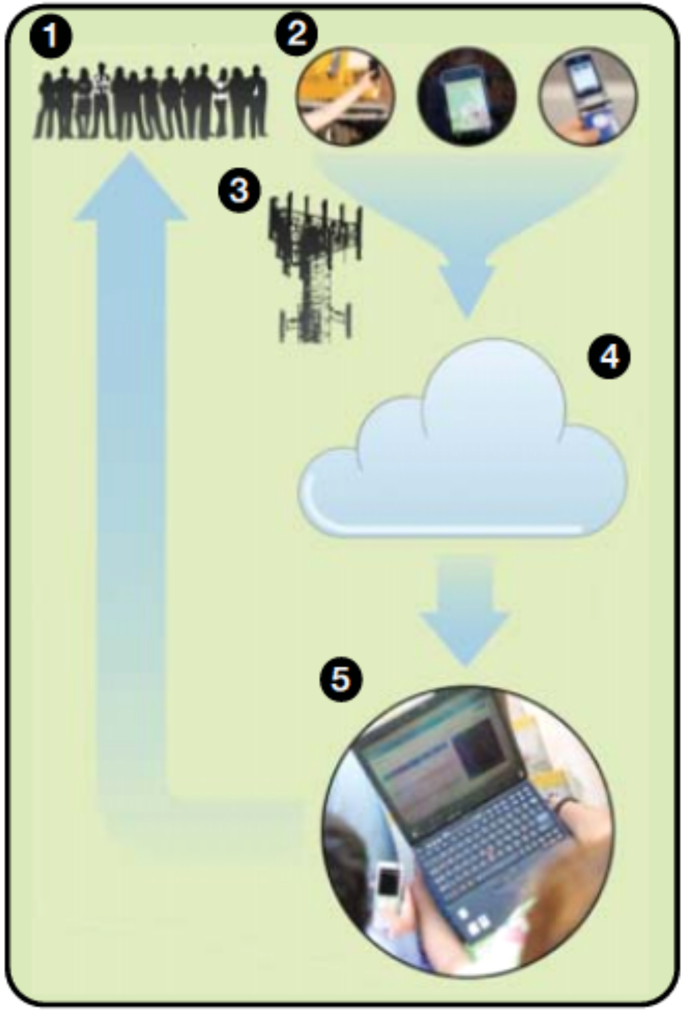
Pulse oximeter

Heart rate monitor

Social Sensing

Geo-tagging the World

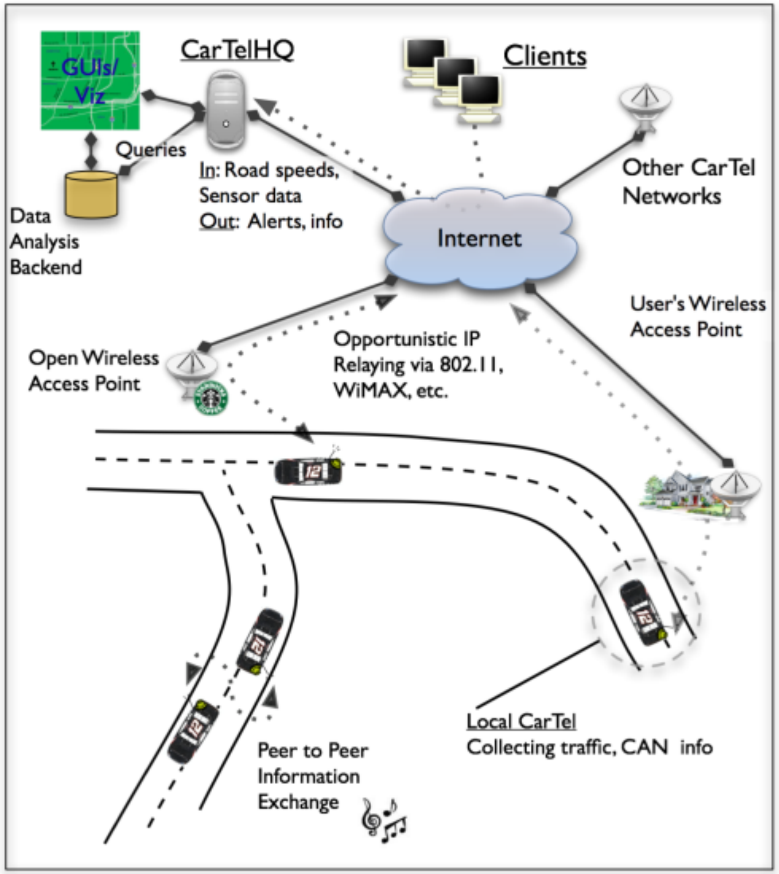
- Phone-based geo-tagging of events of interest (UCLA)
  - Crowds/pollution on beach
  - Invasive species (weeds)
  - Trucks in residential neighborhoods
  - Drinking fountains



Reprinted from UCLA/CENS

Social Sensing

Street Statistics: CarTel, BikeNet, …


GUIs/ Viz
CarTelHQ
Clients
Queries
In: Road speeds, Sensor data
Out: Alerts, info
Other CarTel Networks
Data Analysis Backend
Internet
User's Wireless Access Point
Open Wireless Access Point
Opportunistic IP Relaying via 802.11, WiMAX, etc.
Peer to Peer Information Exchange
Local CarTel
Collecting traffic, CAN info

Reprinted from http://cartel.csail.mit.edu/overview.html

- CarTel (MIT): An ad hoc network of vehicles with sensors
  - Measures road congestion
  - Generates annotated maps
- Bikenet (Dartmouth College): A self-selected community of biking enthusiasts
  - Shares bike route statistics

An Example Application:
Transportation Energy Efficiency

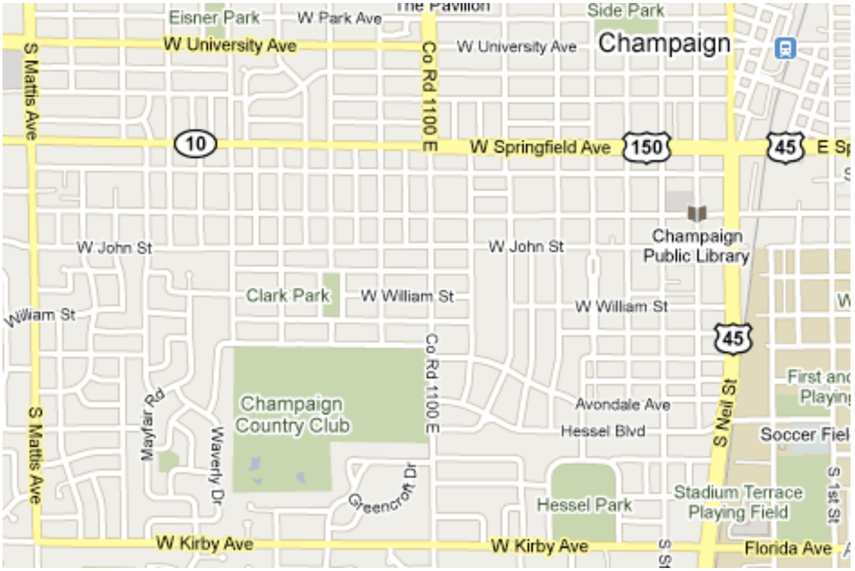
In the US:

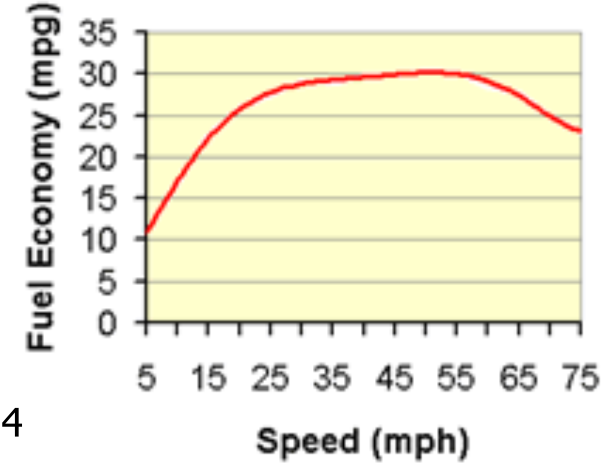
- 200 million light vehicles on the streets
- Each driven 12000 miles annually on average
- Average MPG is 20.3 miles/gallon
- 118 Billion Gallons of Fuel per year!
- **Savings of 1% = One Billion Gallons**

Source: US EPA

GreenGPS: Fuel Efficient Routing

- Individuals share fuel consumption values on various streets at different times of the day
- Models of fuel efficient routes are computed
- They differ from shortest or fastest routes
  - Congestion → shortest may not be fuel efficient
  - MPG lower at higher speeds → fastest may not be fuel efficient

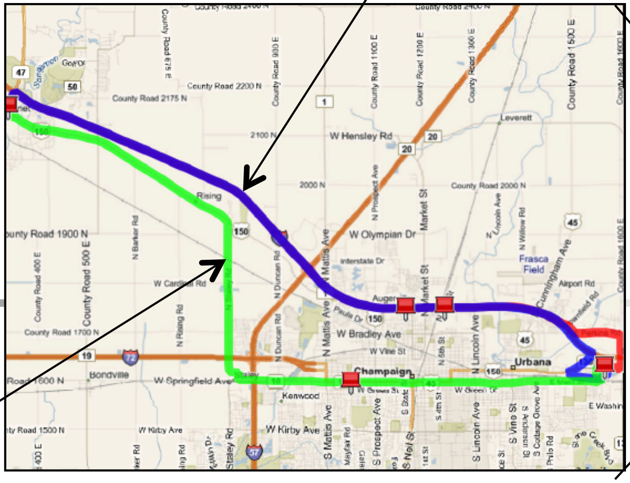




Source: US EPA

Green GPS

Shortest and fastest



Most fuel-efficient

Green GPS
The fuel
efficient option

Saves 6% over shortest path
and 13% over fastest path

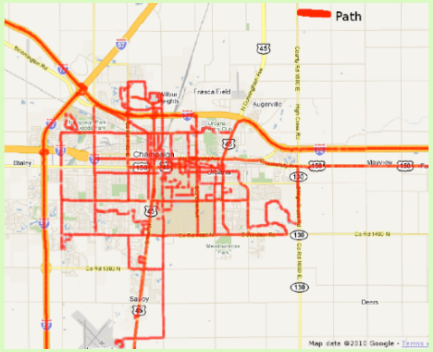
Subscribers

OBDII-WiFi
Adaptor ($50)

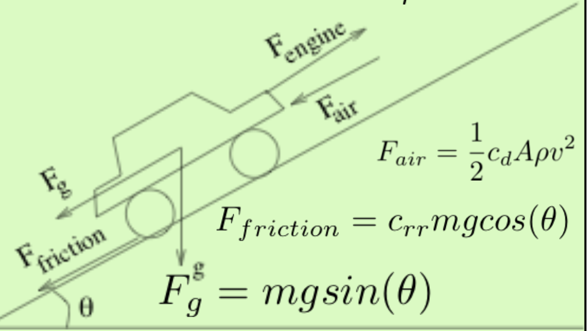
+

GPS Phone

Fuel Data + Physical Models



$$F_{engine} = \frac{\Gamma(\omega)Gg_k}{r}$$



$$F_{air} = \frac{1}{2}c_d A\rho v^2$$

$$F_{friction} = c_{rr}mgcos(\theta)$$

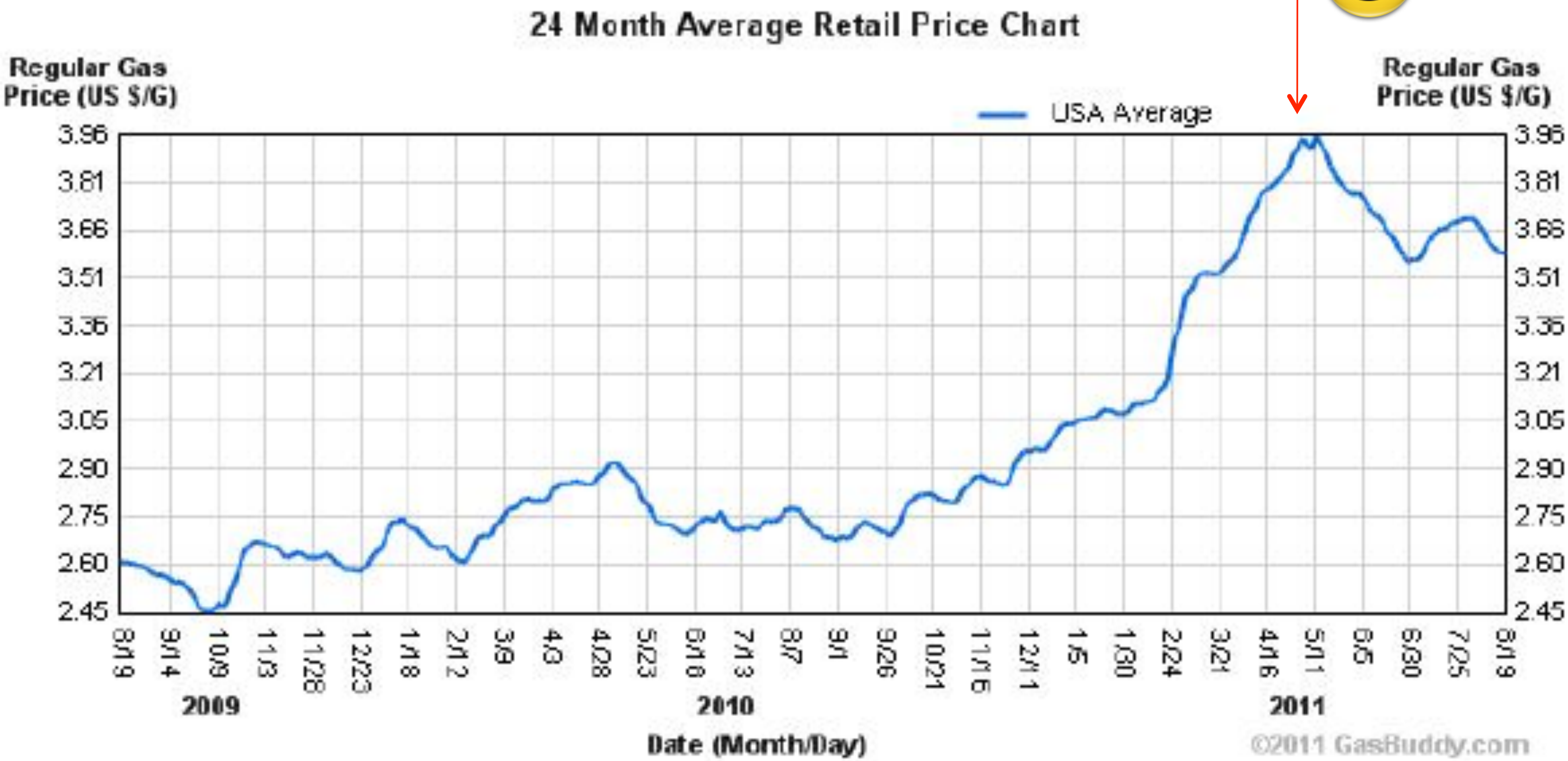
$$F_g = mgsin(\theta)$$

Server

$$F_{car} = F_{engine} - F_{friction} - F_{air} - F_g$$

Gas Prices in the US

May 3rd: Story
“GreenGPS Saves”



Faster? Shorter? Try Cheaper, Greener

Program Gives Drivers the Most Fuel-Efficient Route

Tracy Cozzens

Most GPS devices in cars today give the driver two choices: shortest route or fastest route. GreenGPS provides a third option: most fuel-efficient route.

With gas prices skyrocketing, many drivers would be happy to spend a few more minutes on the road, or take

the engine's fuel efficiency and customizes navigation advice to the particular vehicle, Abdelzaher explained.

The best route computed by Green

may c

other.

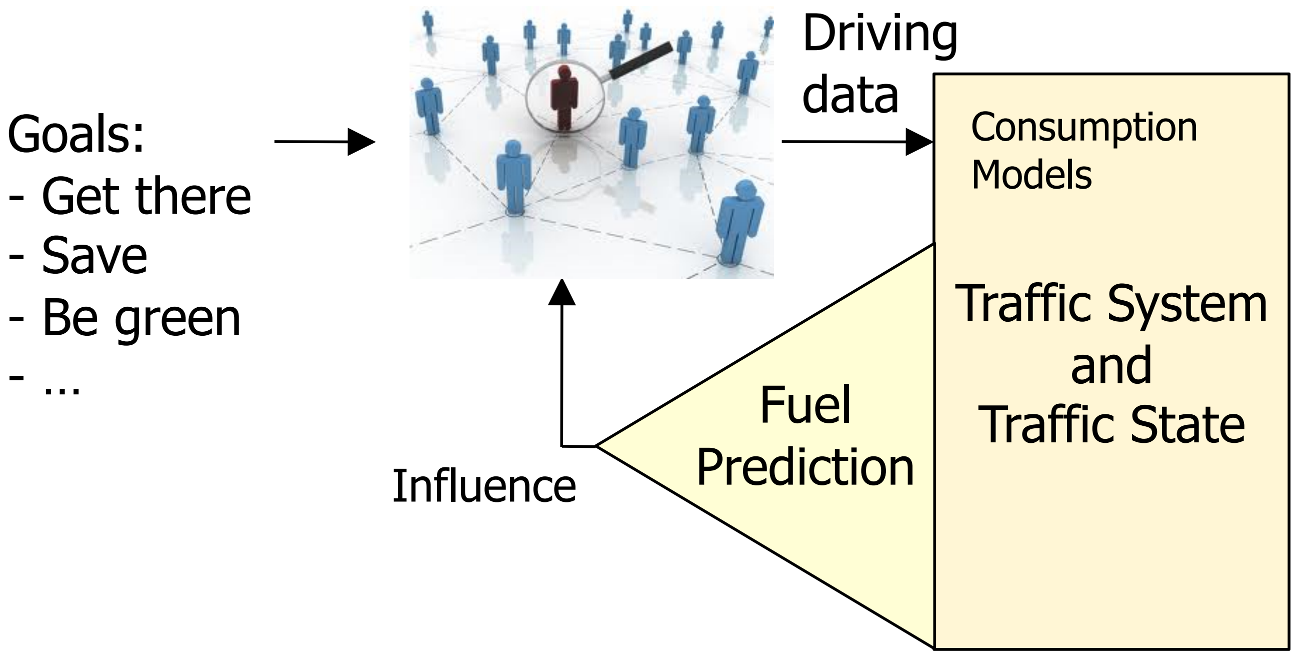
about

loped by University

ponent

Social Sensing Components

- Break-down the challenges
  - Sensing → Modeling → Control



Goals:
- Get there
- Save
- Be green
- …

Driving data

Consumption Models

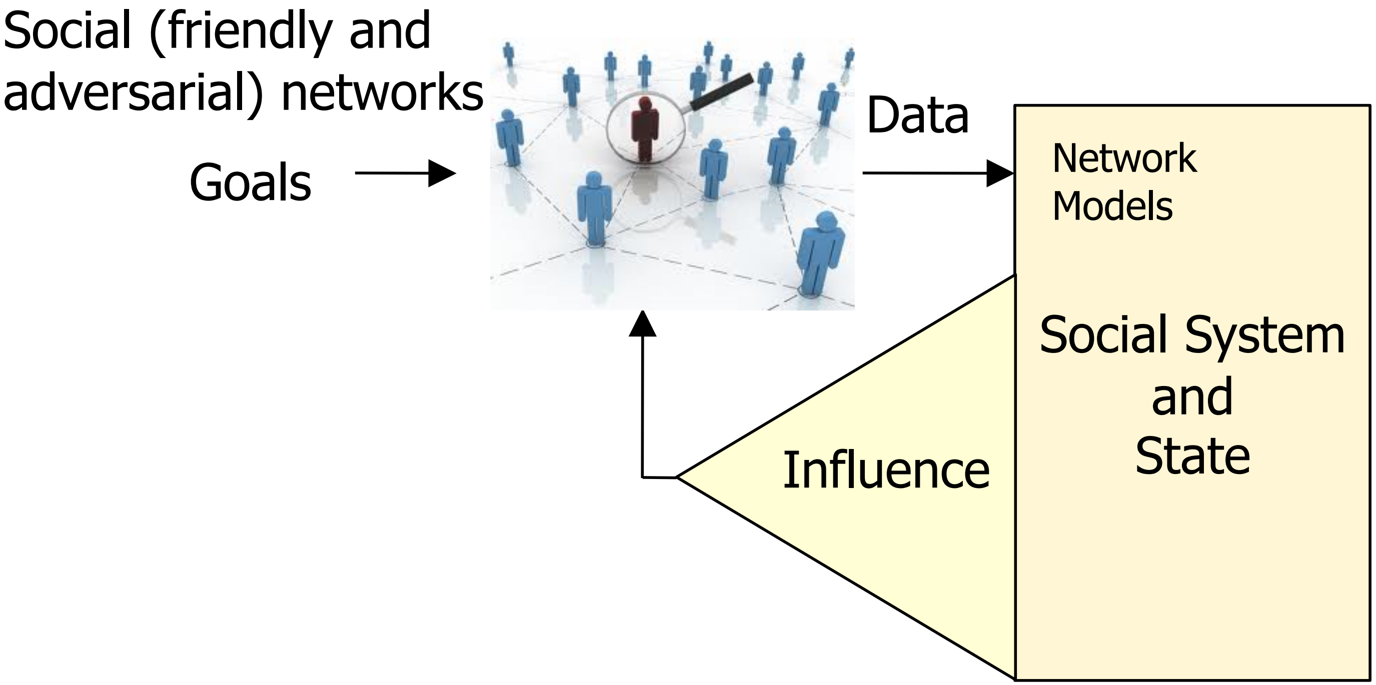
Traffic System and Traffic State

Fuel Prediction

Influence

Counter-insurgency (ARL): A Motivating Application

- Break-down the challenges
  - Sensing → Modeling → Control



Social (friendly and adversarial) networks

Goals

Data

Network Models

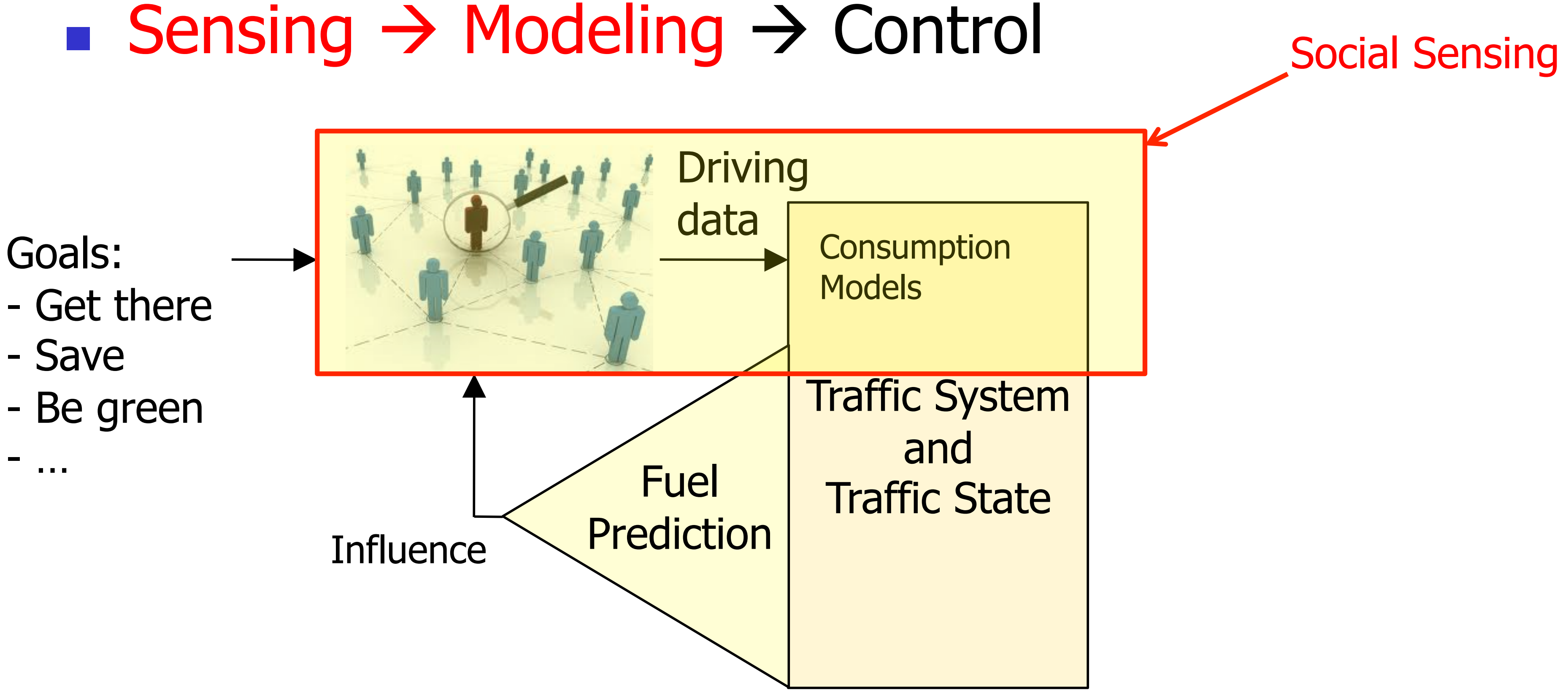
Social System and State

Influence

Tahrir Square, Cairo during Egypt Unrest

GreenGPS: A Motivating Application

- Break-down the challenges
  - Sensing → Modeling → Control



Sensing Challenges in Social Cyber-Physical Systems

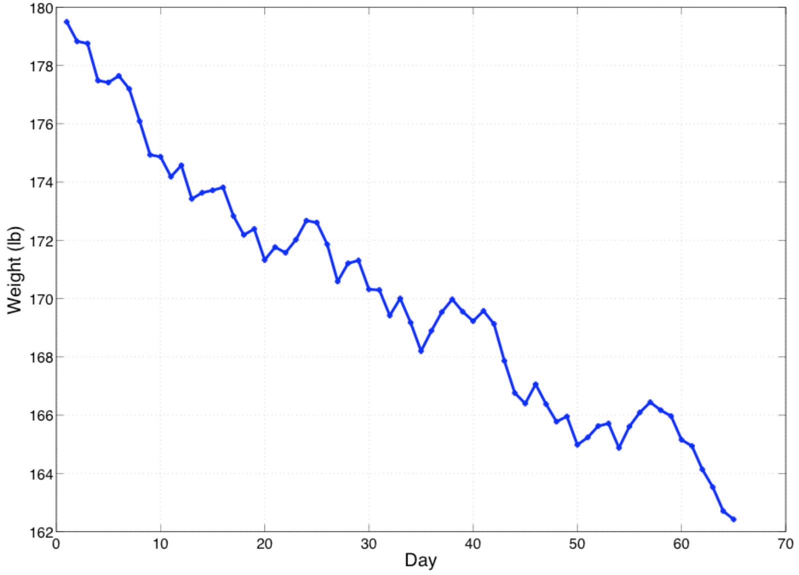
- Privacy
  - How to enable people to share data without violating their privacy?
- “Fact finding” (from noisy data)
  - How to determine reliability of data and sources?
- Modeling and prediction
  - How to efficiently generalize from incomplete data?
- Control (future work)

Social Sensing Challenge #1: Data Source Privacy

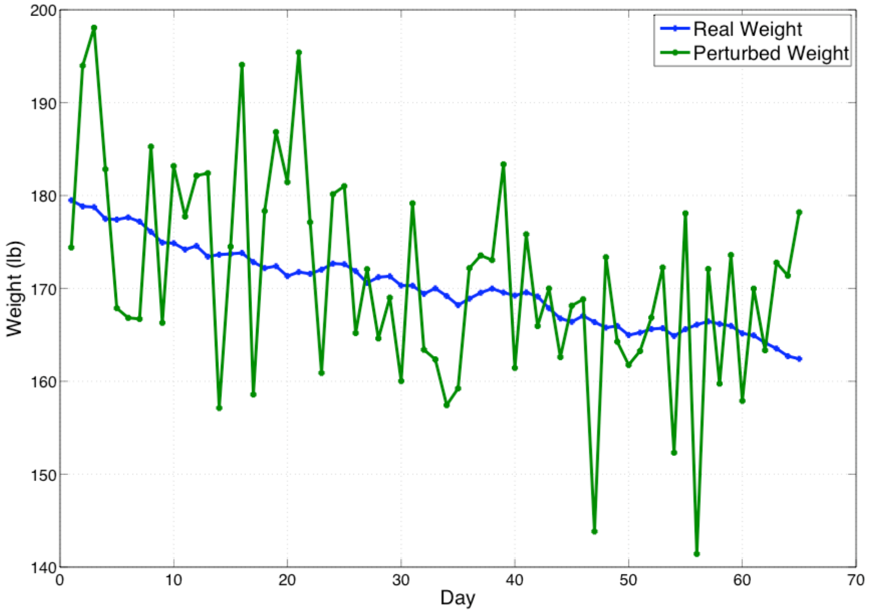
- Clients do not necessarily wish to share their data with the service
  - “Who my cell phone spent the night with is *my* business”
- Data, even if anonymised, can reveal identify of source
- Develop perturbation that preserves privacy of individuals
  - Cannot infer individuals’ data without large error
  - Reconstruction of community distribution can be achieved within proven accuracy bounds

An Example

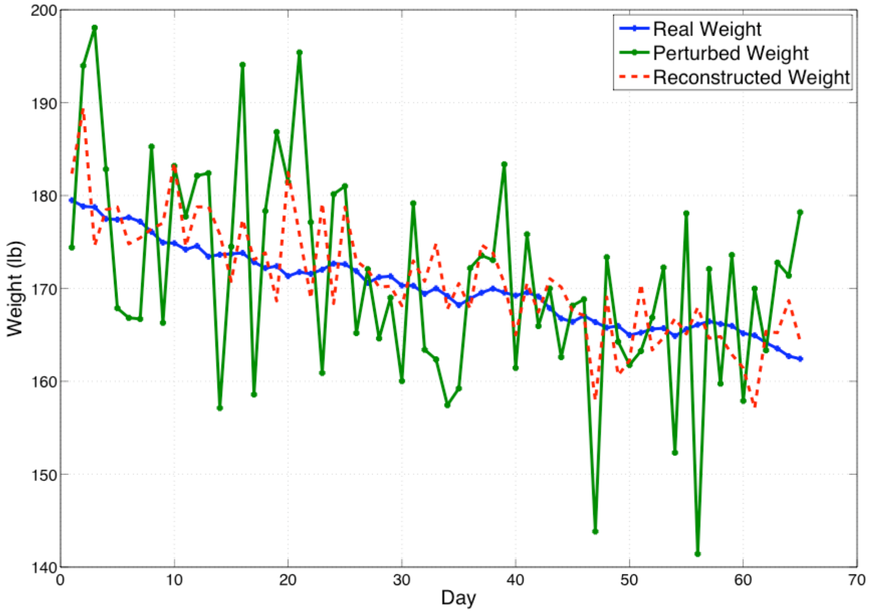
- Dieters want to share weight information to find efficacy of the given diet, without revealing their true weight, average, trend (loss or gain of weight), etc…



Perturb data? Add Noise?

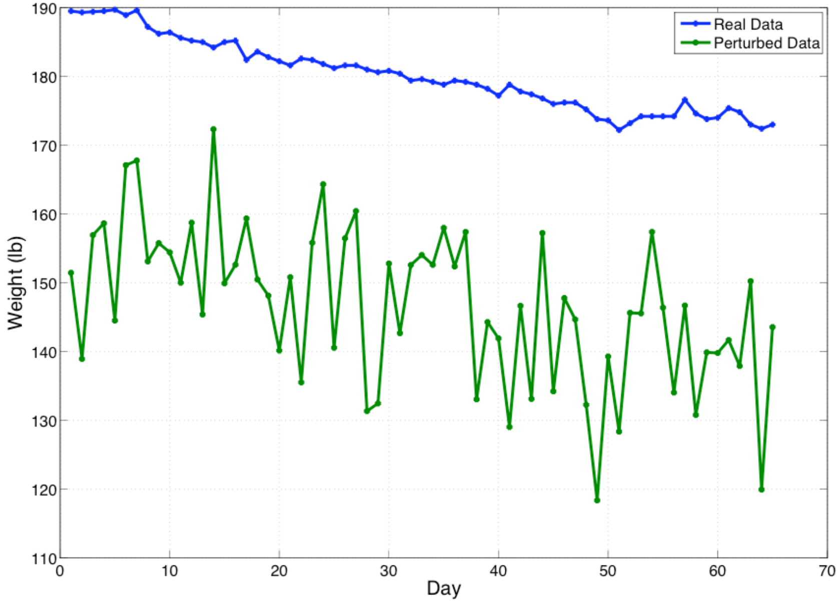


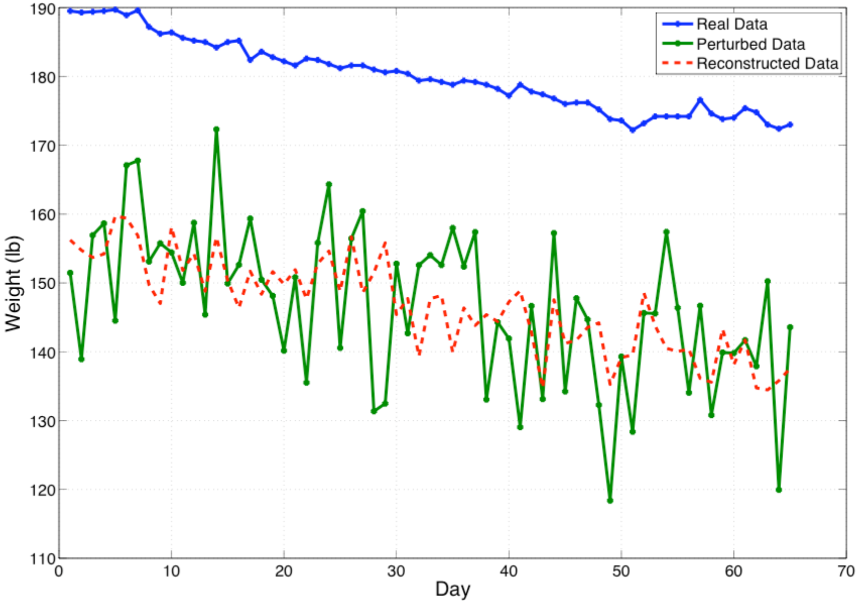
Weight curve perturbed by adding independent random noise



Estimation using PCA to breach privacy of user

Add Noise and Random Offset?





Weight curve perturbed by adding independent random noise and a random offset

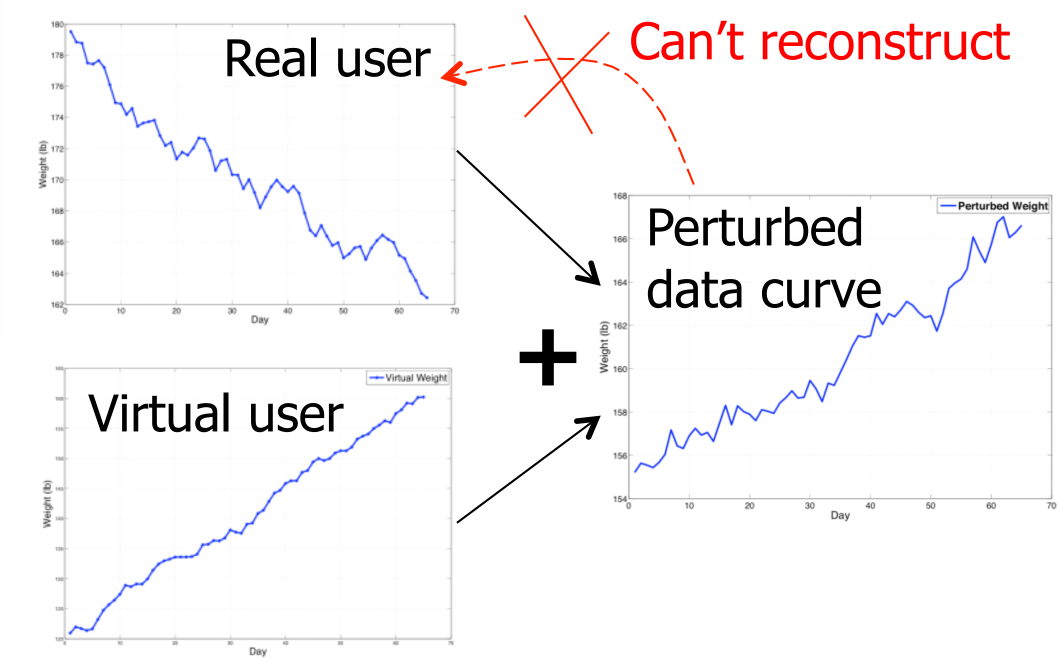
Estimation using PCA to estimate the data of the user

Problem Statement

- Develop perturbation that preserves privacy of individuals
  - Cannot infer individuals’ data without large error
  - Reconstruction of community distribution can be achieved within proven accuracy bounds
  - Perturbation can be applied by non-expert users

Intuitive Approach

- Client adds noise time-series with co-variance that largely mimics covariance of actual data (overlap in frequency domain)



Real user

Virtual user

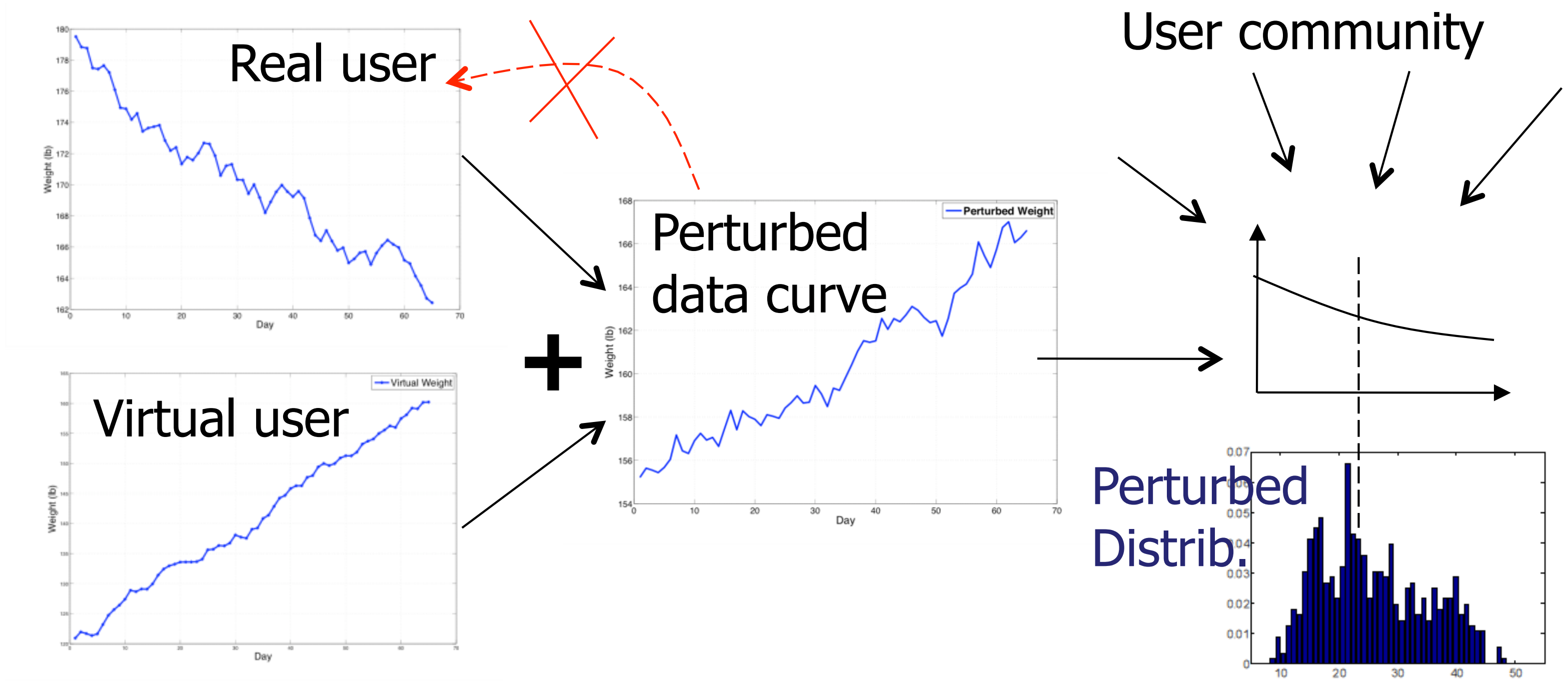
+

Perturbed data curve

Can’t reconstruct

Intuitive Approach

- Client adds noise time-series with co-variance that largely mimics covariance of actual data (overlap in frequency domain)
- Users send their perturbed data to aggregation server



Real user

Virtual user

+

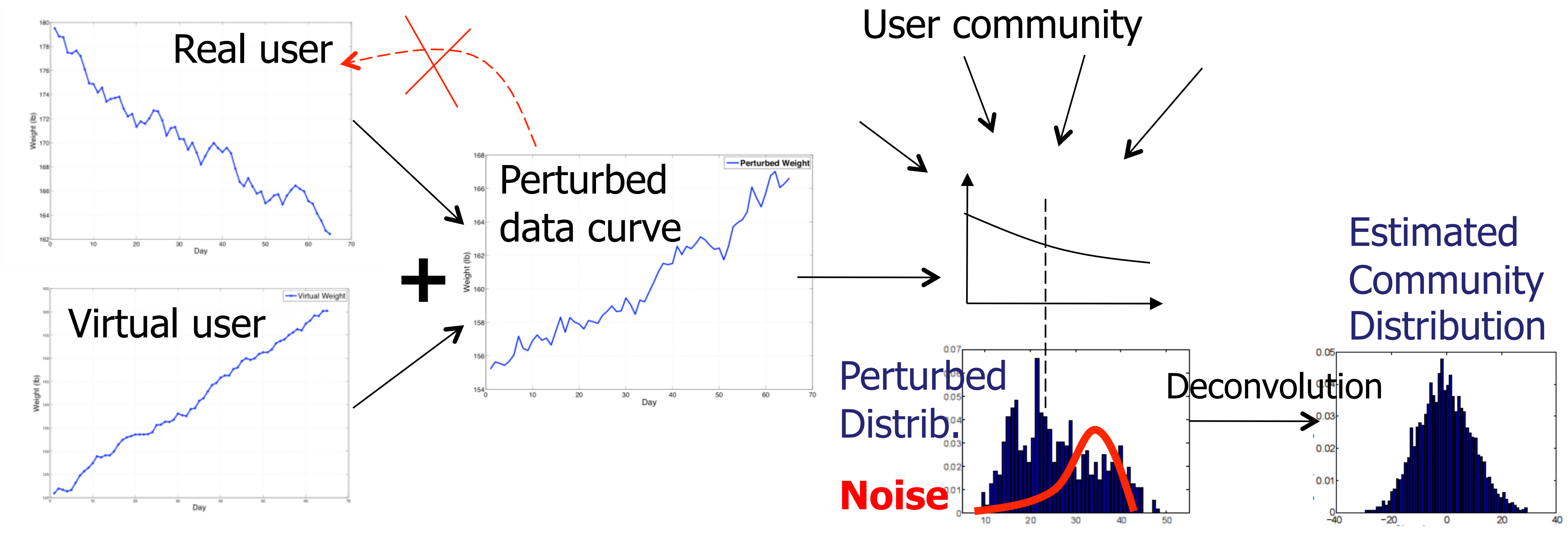
Perturbed data curve

User community

Perturbed Distrib.

Intuitive Approach

- Client adds noise time-series with co-variance that largely mimics covariance of actual data (overlap in frequency domain)
- Users send their perturbed data to aggregation server
- Given perturbed community distribution and noise, server uses de-convolution to reconstruct original data distribution at any point in time



Real user

Virtual user

+

Perturbed data curve

User community

Estimated Community Distribution

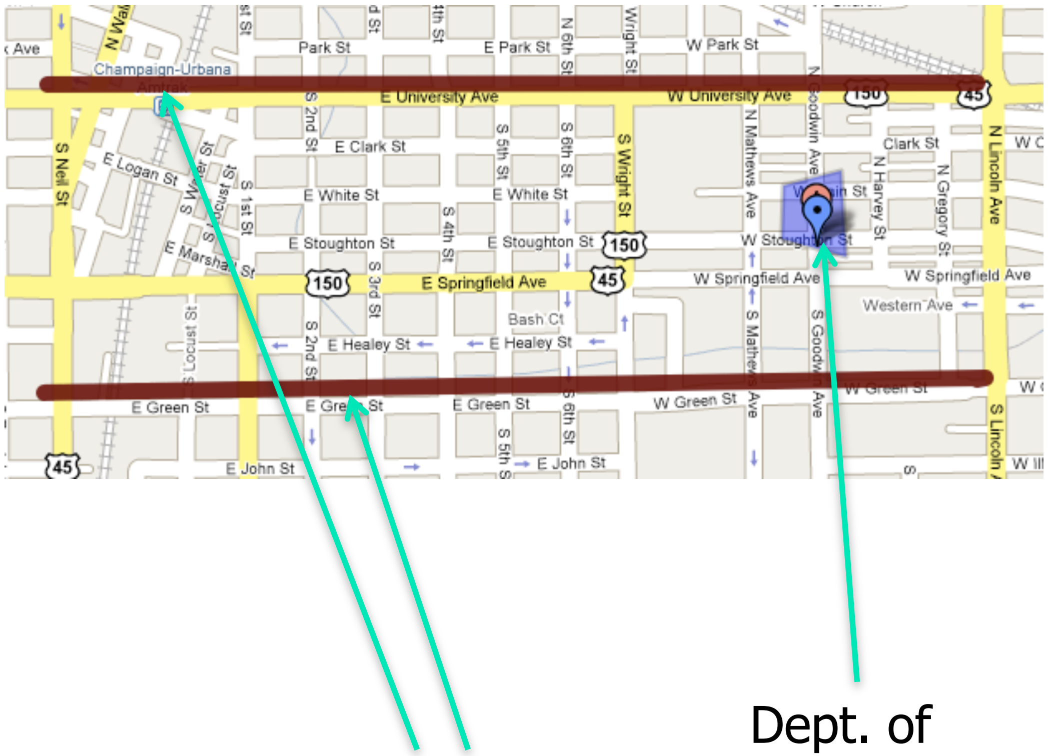
Perturbed Distrib.

Noise

Deconvolution

Traffic Analyzer

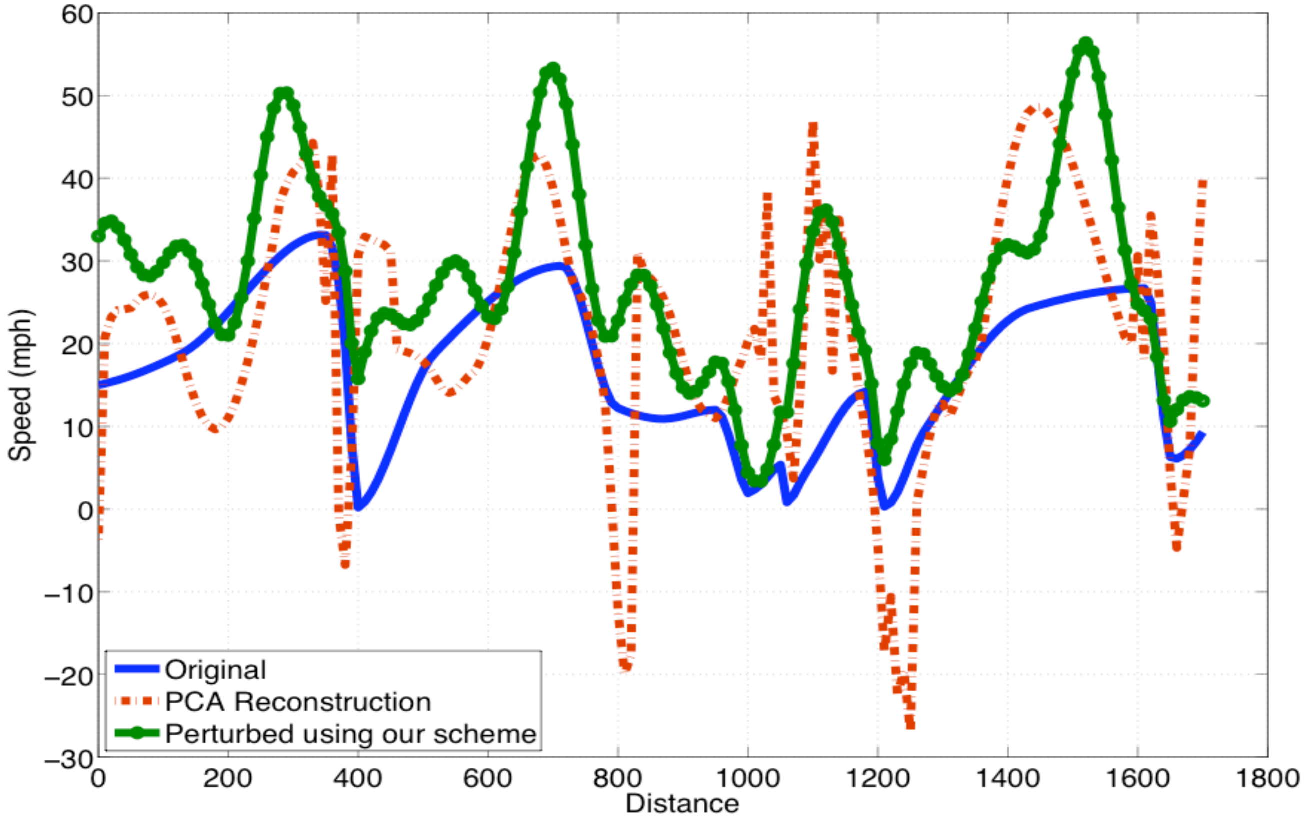
- Users share perturbed speed data with aggregation server
- Server combines perturbed speed data and uses de-convolution with noise model to compute original speed distribution
- Garmin GPS used for data collection
- Results are from real data collection in Urbana-Champaign in 2008



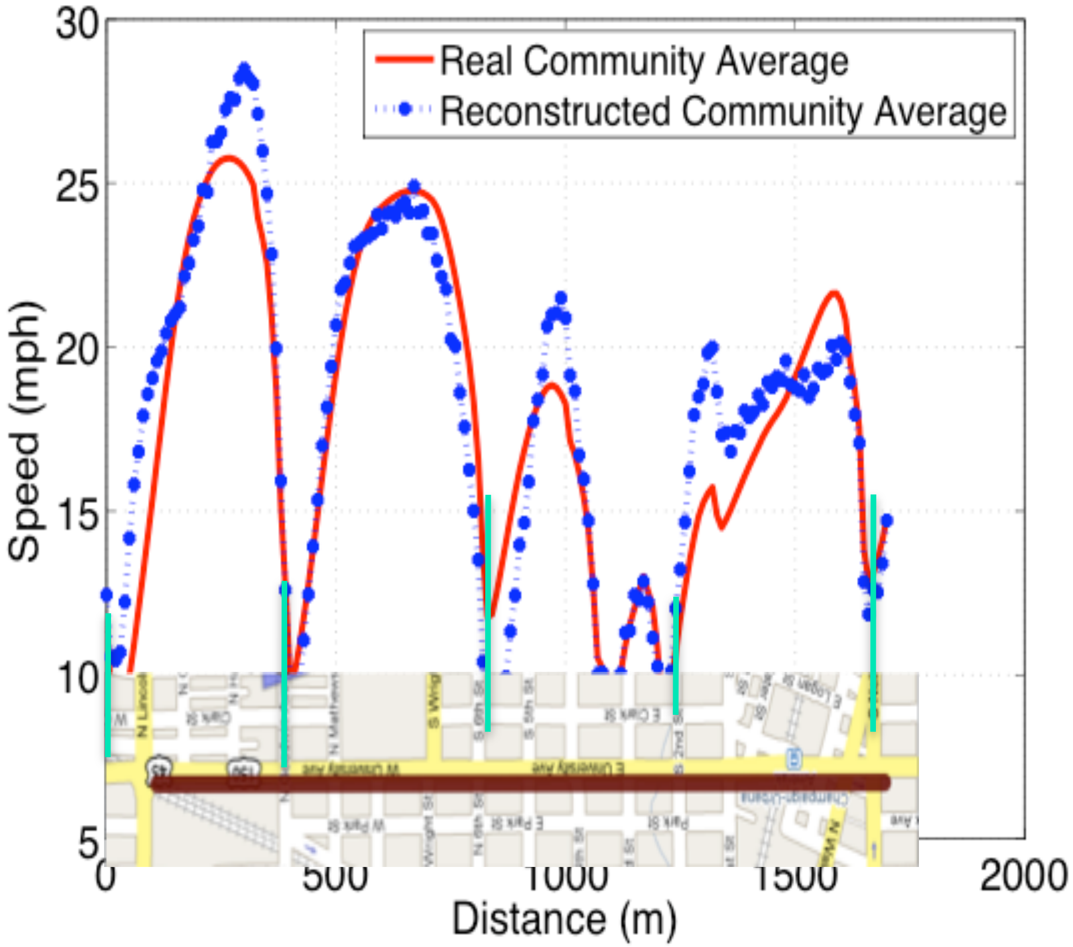
Roads for which we want to estimate average speed

Dept. of Computer Science

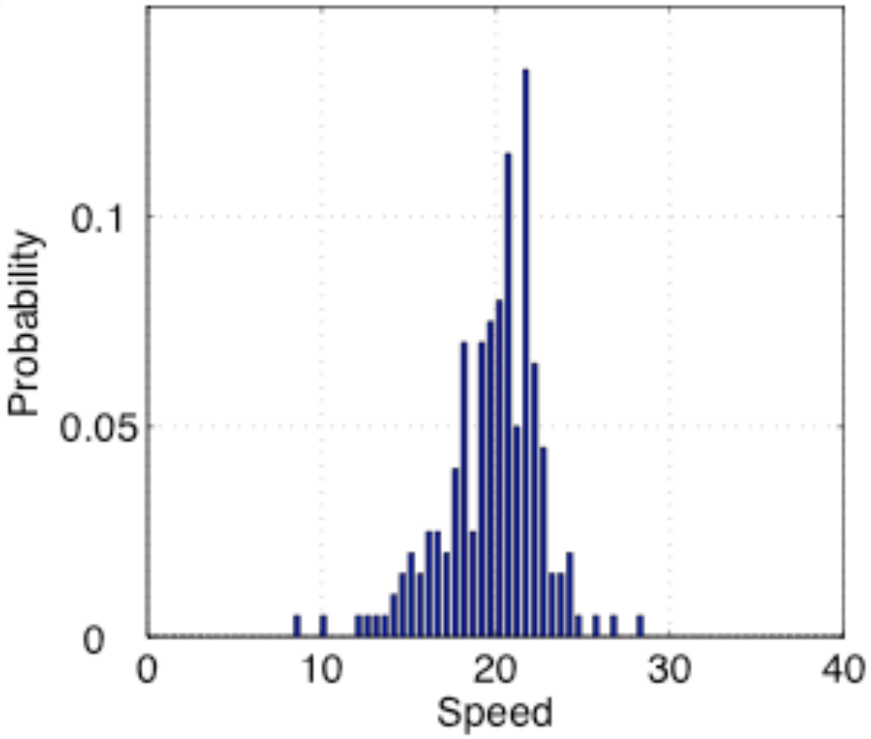
Perturbing Speed



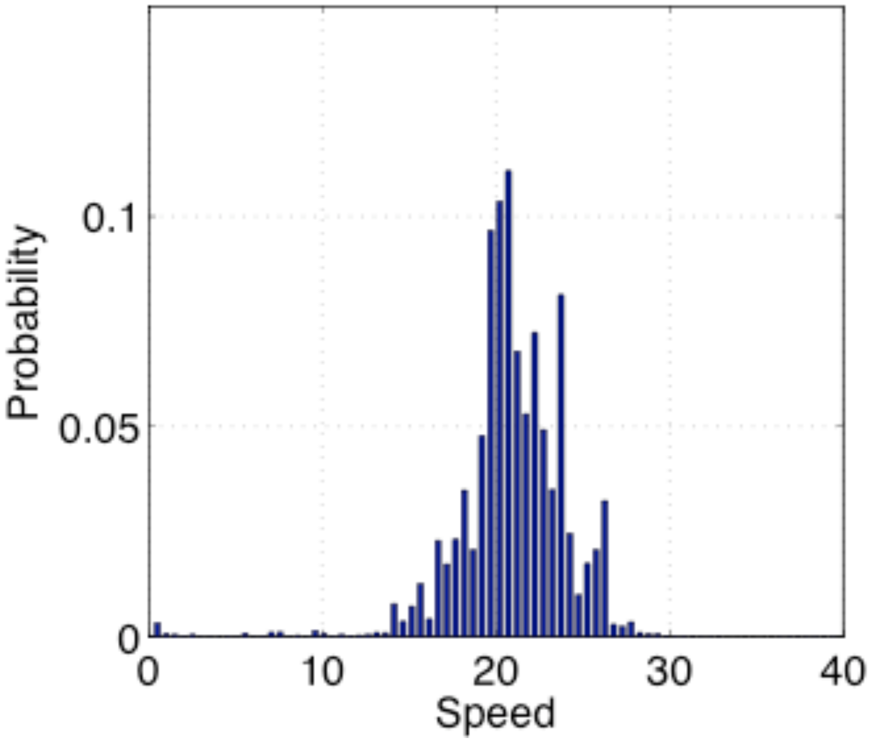
Reconstruction of Average Speed



Reconstruction of Community Speed Distribution



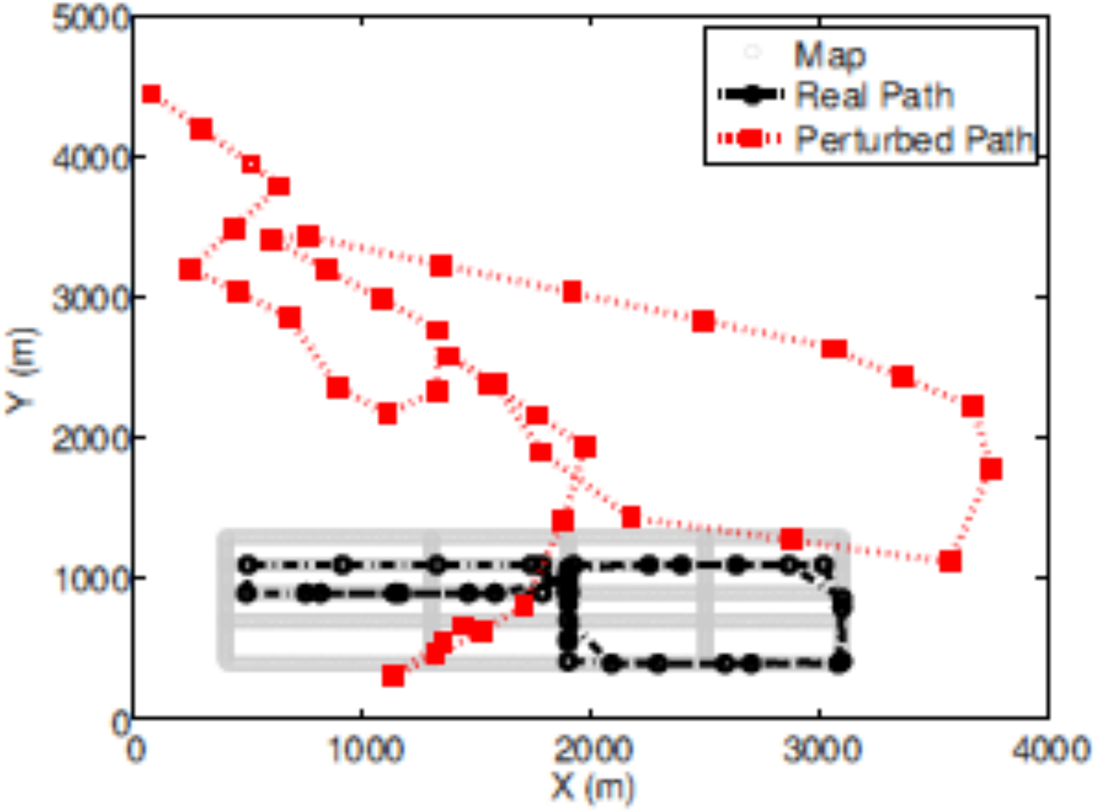
Real community distribution of speed



Reconstructed community distribution of speed

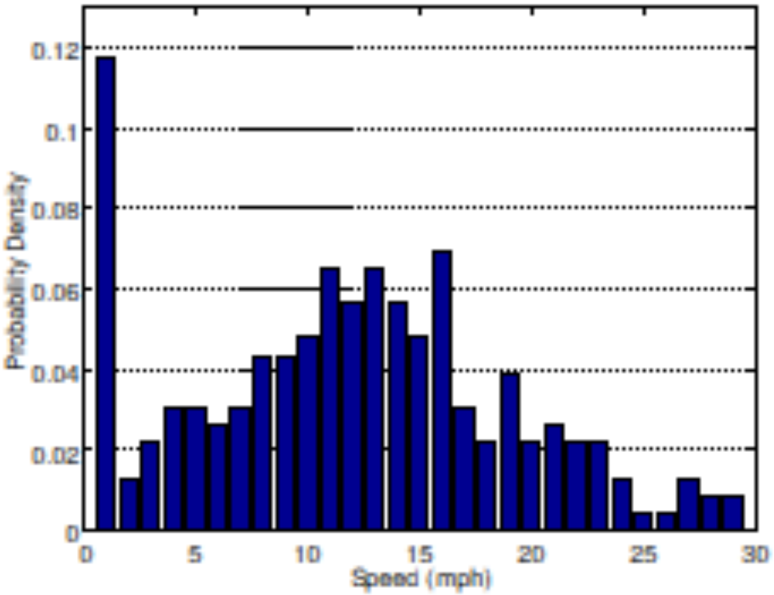
Perturbing Speed and Location

- Clients lie about both their location and speed

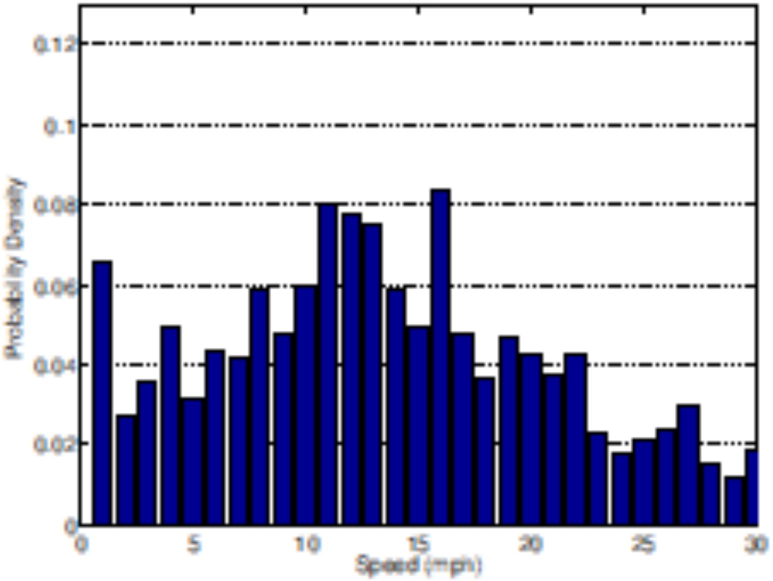


Reconstruction Accuracy

- Real versus reconstructed speed



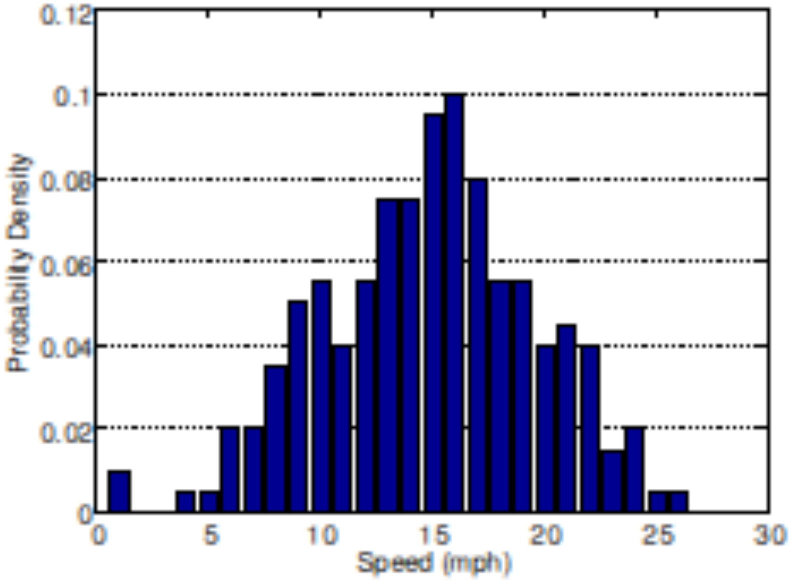
Real community distribution of speed



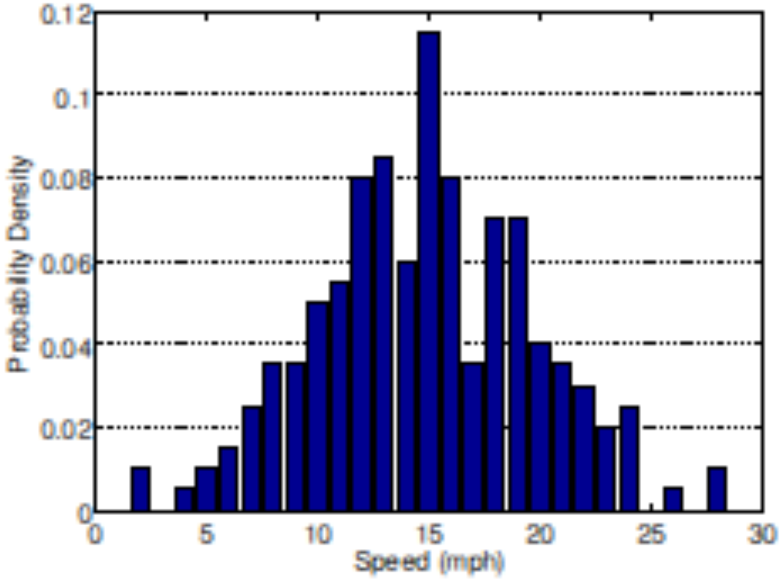
Reconstructed community distribution of speed

More on Reconstruction Accuracy

- Real versus reconstructed speed on Washington St., Champaign



Real community distribution of speed



Reconstructed community distribution of speed

How Many are Speeding?

- Real versus estimated percentage of speeding vehicles on different streets (from data of users who “lie” about both speed and location)

Street	Real % Speeding	Estimated % Speeding
University Ave	15.6%	17.8%
Neil Street	21.4%	23.7%
Washington Street	0.5%	0.15%
Elm Street	6.9%	8.6%

Privacy and Optimal Perturbation

- Is the an optimal perturbation scheme?
- What is the measure the privacy?
- How can we generate the optimal perturbation?

Privacy Measure

- We use the *mutual information* $I(X;Y)$ to measure the information about X contained in Y
- *Minimal information leak* under noise power constraint

$$\mathcal{P}_X^o = \min_Z I(X, X+Z)$$
$$\text{subject to } P_Z \leq P_0$$

- X is the original data
- Y is the perturbed data
- Z is the noise
- P_Z is the power of Z

Upper Bound on Privacy

- Lemma (Ihara, 78)
- The noise that minimizes the upper bound on information leak is a Gaussian noise

$$I(X;Y) \leq I(X_G, X_G + Z_G) = \frac{1}{2} \frac{\det(K_X + K_Z)}{\det(K_Z)}$$

Covariance of signal

Mutual Information (Leak)

Covariance of noise

Finding the Optimal Noise

- Solving for the optimal noise’s covariance matrix

$$K_Z^* = \underset{K_Z}{argmin} \frac{1}{2n} \log \frac{det(K_X + K_Z)}{det(K_Z)}$$

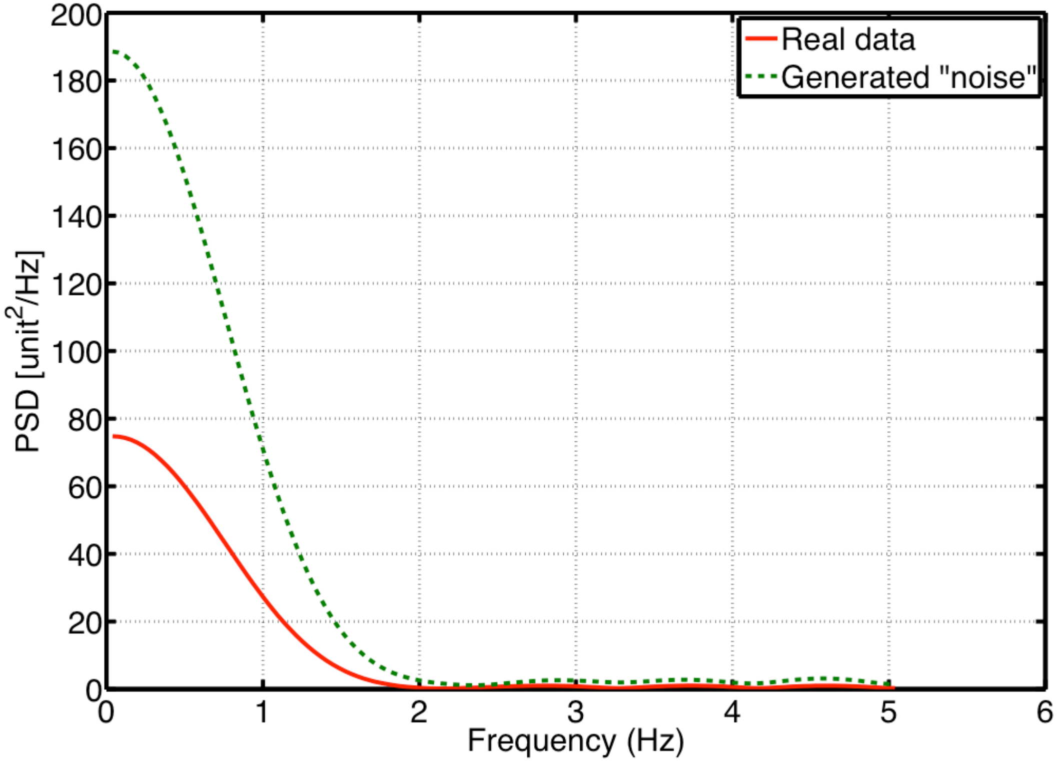
$$\text{subject to}$$

$$\frac{1}{n} trace(K_Z) \leq P_0$$

$$K_Z \succ 0$$

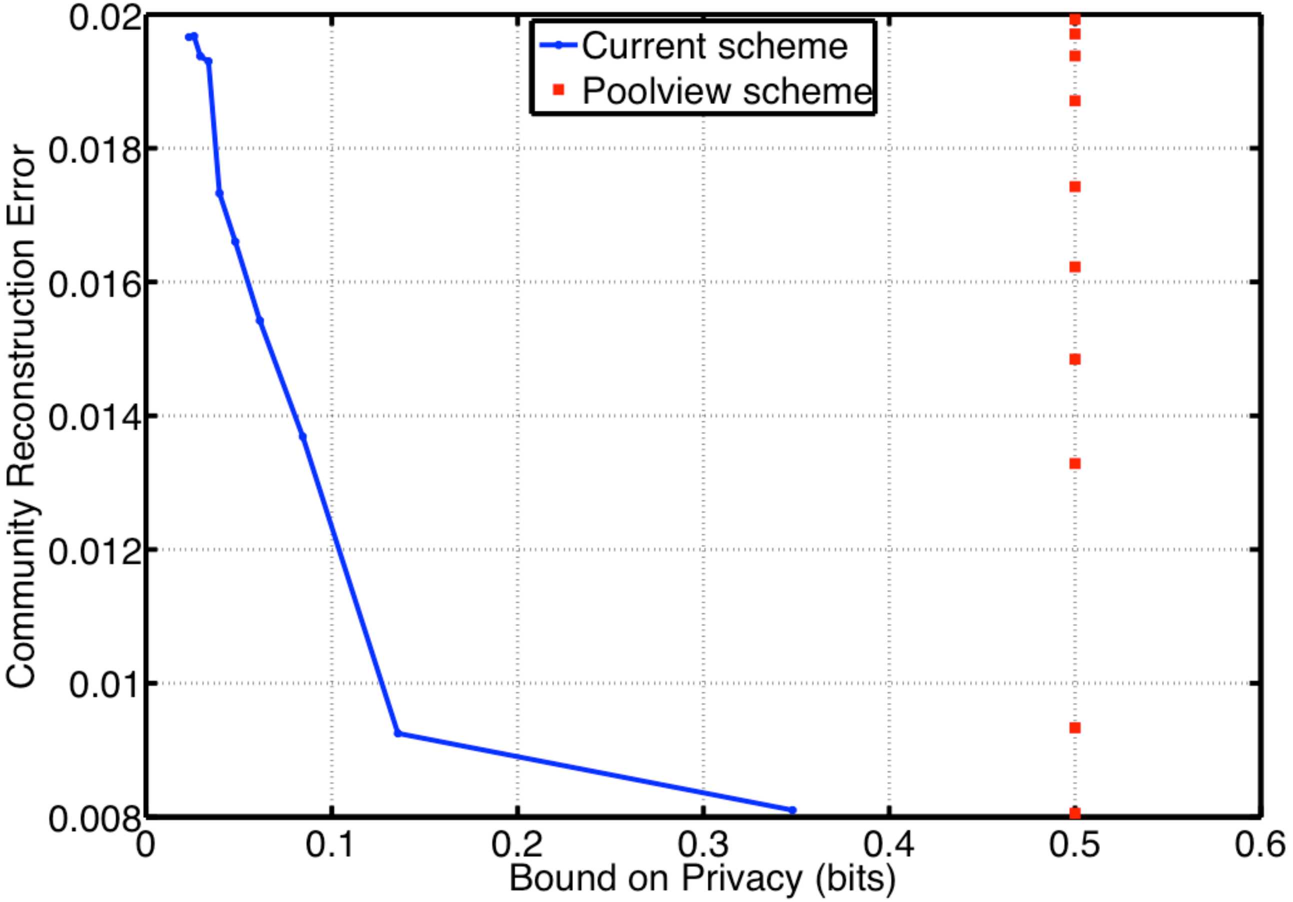
$$K_Z \text{ is Symmetric Toeplitz}$$

Optimal Noise



- The noise generation method can be seen as the optimal allocation of noise energy in the frequency domain

Utility vs. Privacy Trade-off

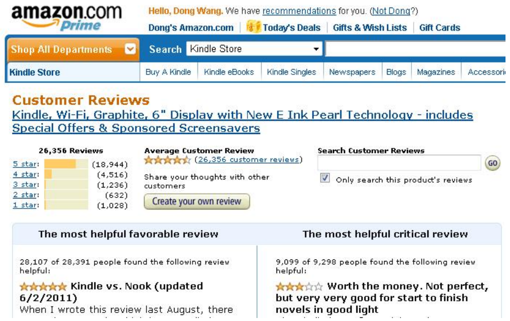


Social Sensing Challenge #2: Fact-finding from Noisy Data

- In social sensing applications, participants may not be known or vetted *a priori*
- Some data may be incorrect and some sources unreliable
- Non-numeric data: Human text, images, etc.
- How to tell good from bad sources?

The Problem

Human are involved in the sensing and data fusion loop



amazon.com Prime

Hello, Dong Wang. We have recommendations for you. (Not Dong?)

Dong's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Shop All Departments | Search Kindle Store

Kindle Store | Buy A Kindle | Kindle eBooks | Kindle Singles | Newspapers | Blogs | Magazines | Accessori

Customer Reviews

Kindle, Wi-Fi, Graphite, 6" Display with New E Ink Pearl Technology - includes Special Offers & Sponsored Screensavers

26,356 Reviews

5 star:	(18,944)
4 star:	(4,516)
3 star:	(1,236)
2 star:	(632)
1 star:	(1,028)

Average Customer Review

(26,356 customer reviews)

Share your thoughts with other customers

Create your own review

Search Customer Reviews

GO

☑ Only search this product's reviews

The most helpful favorable review

28,107 of 28,391 people found the following review helpful:

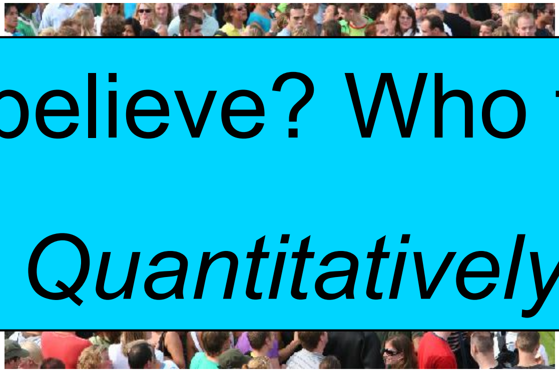
Kindle vs. Nook (updated 6/2/2011)

When I wrote this review last August, there

The most helpful critical review

9,099 of 9,298 people found the following review helpful:

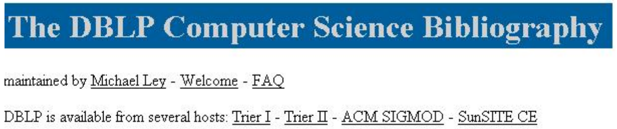
Worth the money. Not perfect, but very very good for start to finish novels in good light



What to believe? Who to believe?

Quantitatively?

Detailed prior knowledge on source reliability is *unknown.*



The DBLP Computer Science Bibliography

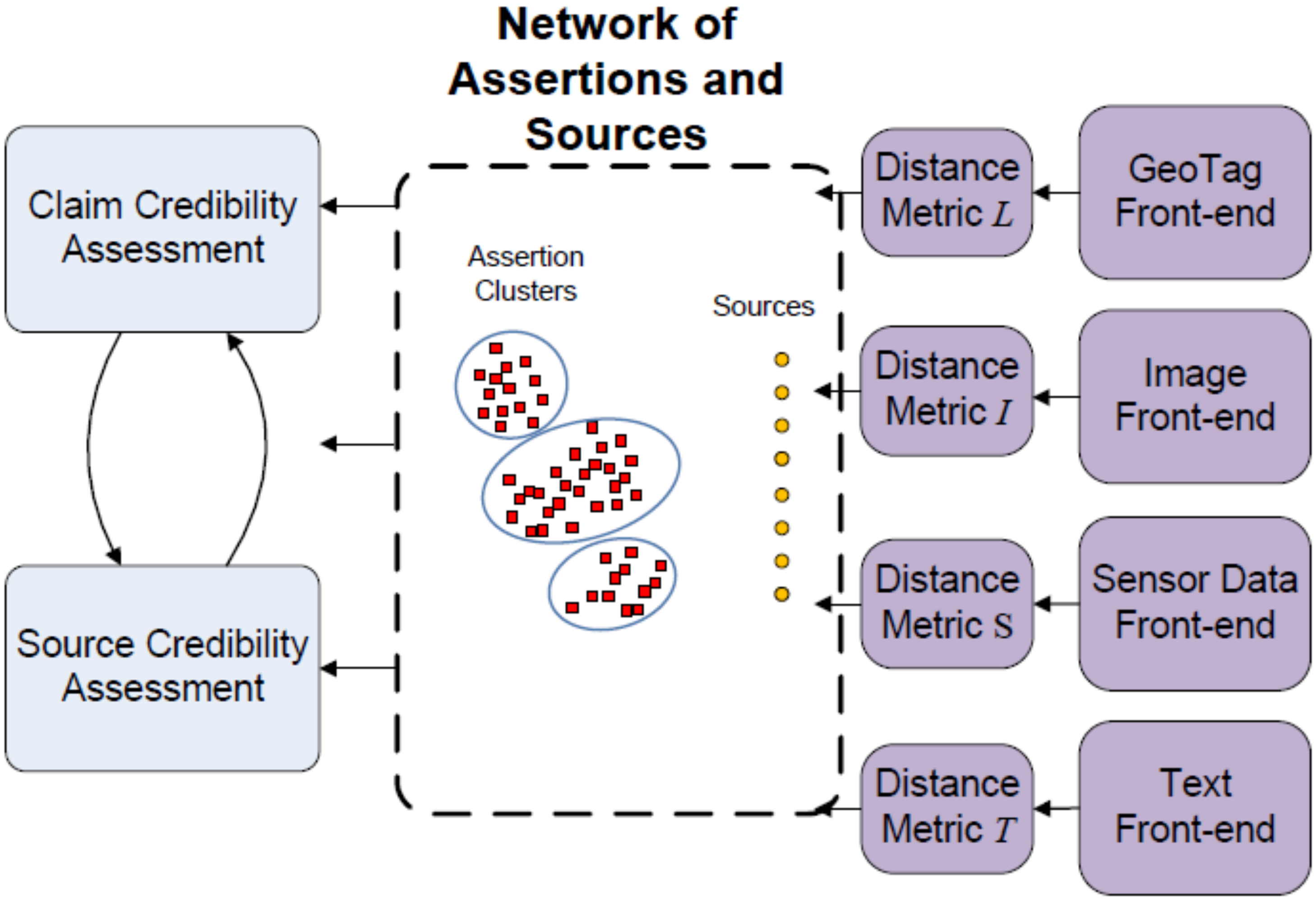
maintained by Michael Ley - Welcome - FAQ

DBLP is available from several hosts: Trier I - Trier II - ACM SIGMOD - SunSITE CE

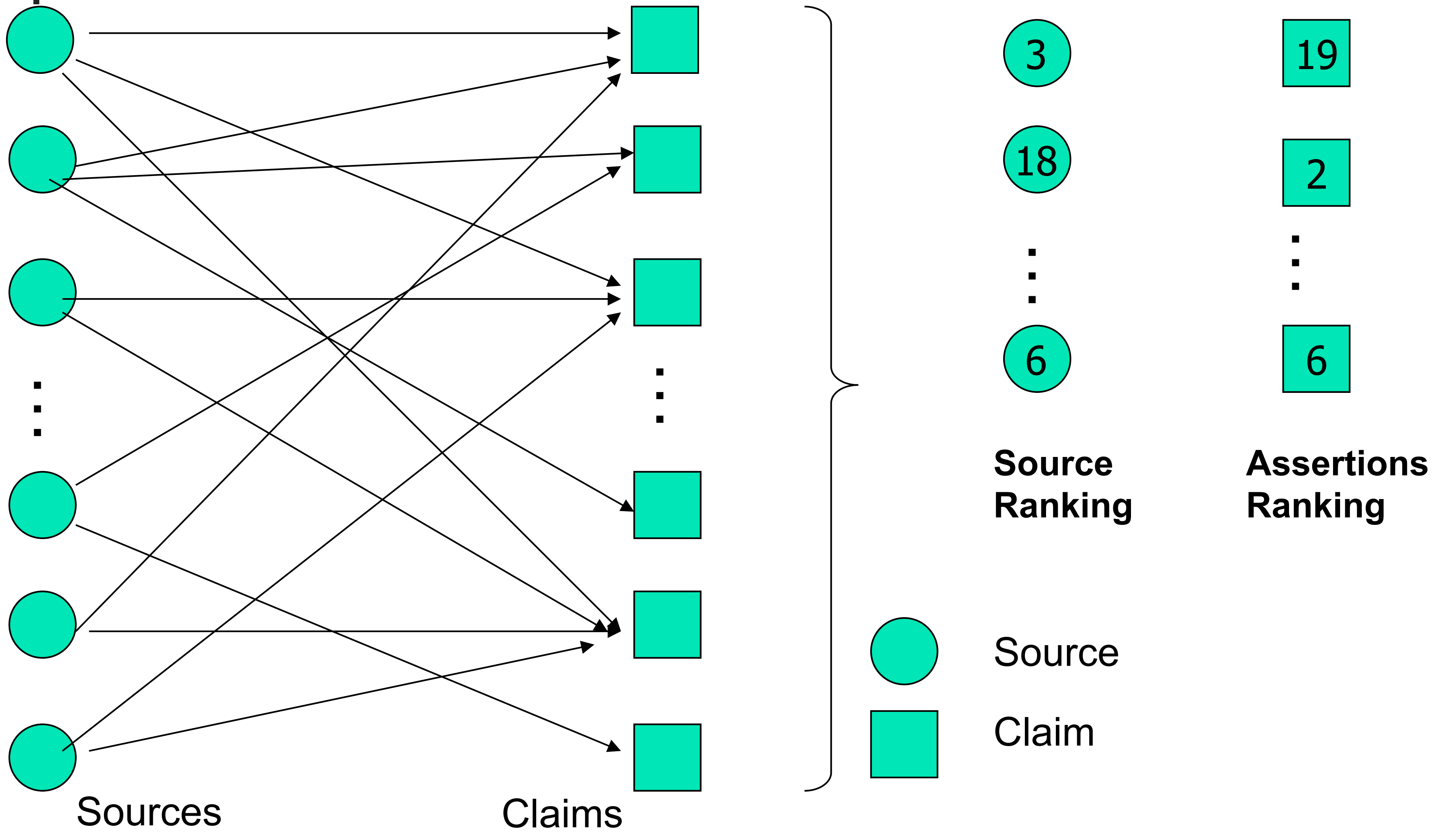
Apollo: A General Fact-finding Service for Human-centric Sensing

- Human-centric sensing applications
  - Use potentially unreliable or unverified sources
  - May be plagued by noisy and incorrect data, especially in large deployments with un-vetted participants
- Apollo:
  - A “generic tool” for data cleaning and fact-finding
  - Does not rely on application-specific methods for distilling sensor data
  - Works with a wide range of applications involving data types ranging from time-series of sensor readings and GPS location tags to image and text

High-level Architecture



Fact-Finding
Optimal Assignment of Truth Values to Sources and "Claims"



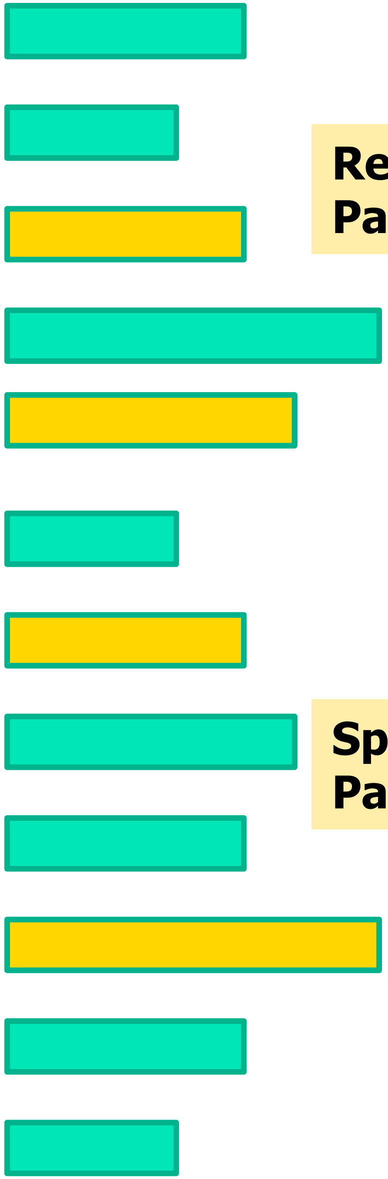
The Apollo Analytic Engine

- Formulates the fact-finding problem as one of maximum likelihood estimation
- Solves it using the *Expectation Maximization* (EM) algorithm
- Computes a bound on estimation accuracy (using the Cramer Rao Bound)

Math Formulation

🟩 True Assertions

🟨 False Assertions



Reliability of Participant *i* = 🟩 *i* / (🟩 *i* + 🟨 *i*)

Participant Reliability

$$t_i = P(C_j^t \mid S_iC_j)$$

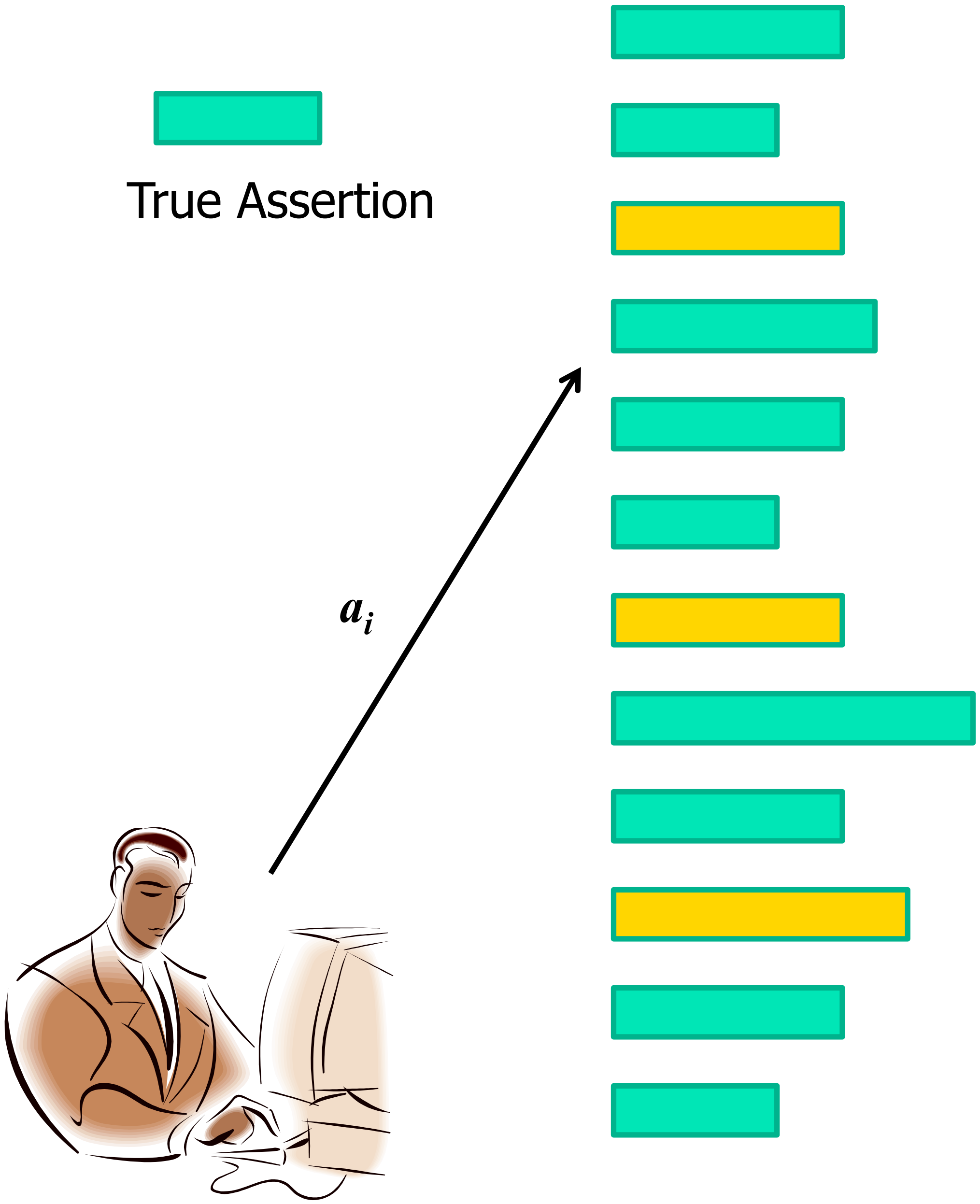
S_iC_j : participant i claims assertion j

Speak Rate of Participant *i* ∝ (🟩 *i* + 🟨 *i*) / (🟩 *All* + 🟨 *All*)

Participant i speak with rate s_i

$$s_i = P(S_iC_j)$$

Math Formulation

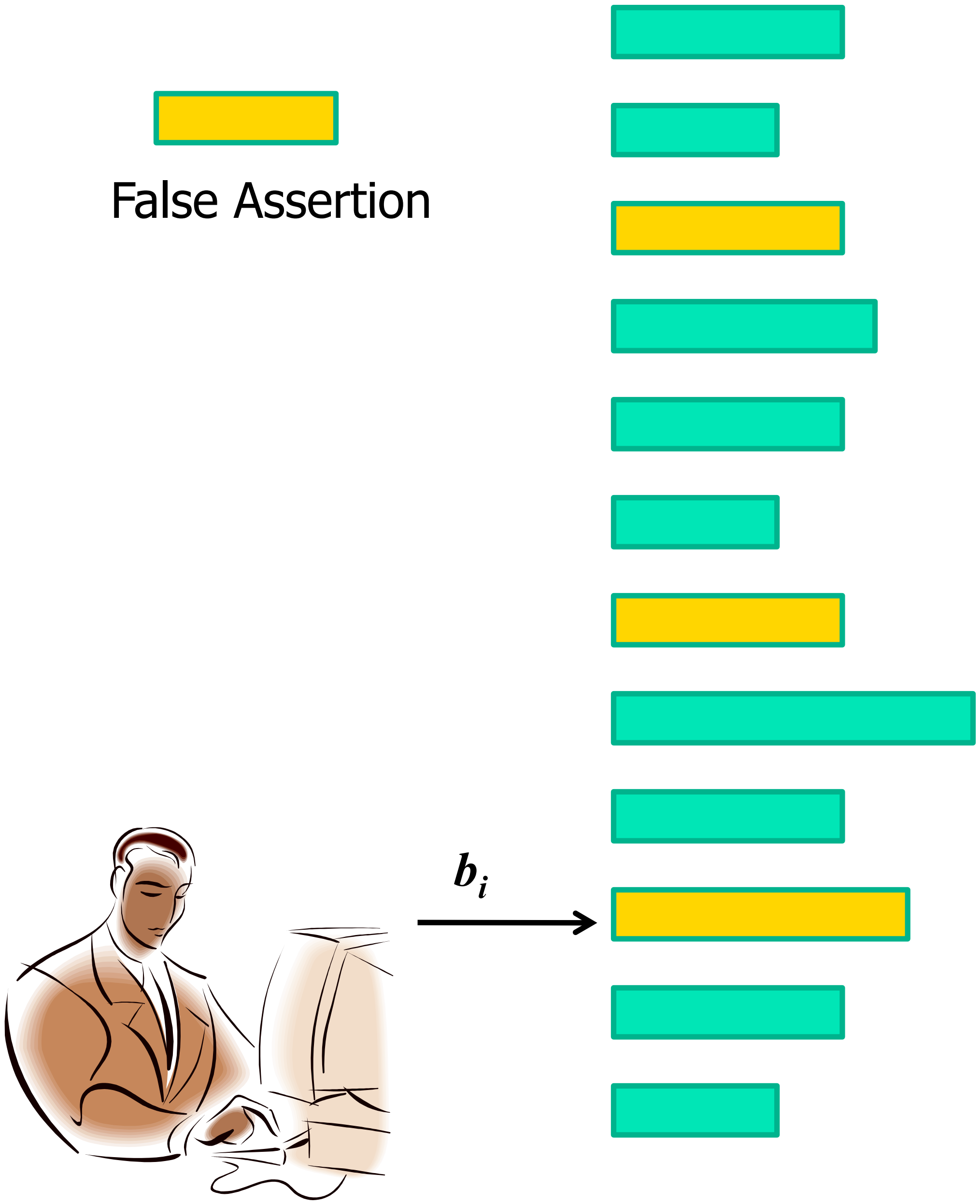


$$a_i = P(S_i C_j \mid C_j^t)$$

Using Bayesian Theorem: $a_i = \dfrac{t_i \times s_i}{d}$

where d is the overal prior that a randomnly chozen assertion is true

Math Formulation



🟨 False Assertion

$$b_i = P(S_i C_j \mid C_j^f)$$

Using Bayesian Theorem: $b_i = \dfrac{(1-t_i)\times s_i}{1-d}$

where d is the overal prior that a randomnly chozen assertion is true

Math Formulation

Log-likelihood Function of EM Scheme:

$$l_{em}(x;\theta)=\sum_{j=1}^{N}\left\{\begin{array}{l}\boxed{z_j}\times\left[\sum_{i=1}^{M}\left(S_iC_j\log a_i+(1-S_iC_j)\log(1-a_i)+\log d\right)\right]\\+(1-z_j)\times\left[\sum_{i=1}^{M}\left(S_iC_j\log b_i+(1-S_iC_j)\log(1-b_i)+\log(1-d)\right)\right]\end{array}\right\}$$

where $z_j=1$ when measured variable j is true and 0 otherwise

Math Formulation

Our Goal is to find the Confidence Interval of Particpant Reliability MLE:

Confidence Interval !

$$\boxed{\left(t_i^{MLE} - c_p^{lower}, t_i^{MLE} + c_p^{upper}\right) \text{ with } \quad c\%}$$

where $c\%$ is the confidence level of the estimation interval, c_p^{lower} and c_p^{upper} represents the lower and upper bound on the estimation deviation from MLE t_i^{MLE}

Derivation of Confidence

Estimation and Statistic Background

Fisher information is defined as

$$\boxed{I(\theta) = E_X\left[\varphi(x;\theta)\ \varphi(x;\theta)^T\right]} \longrightarrow \text{Fisher Information}$$

Score vector $\varphi(x;\theta)$ for a $k \times 1$ estimation vector $\theta = [\theta_1, \theta_2, ..., \theta_k]^T$

$$\varphi(x;\theta) = \left[\frac{\partial l(x;\theta)}{\partial \theta_1}, \frac{\partial l(x;\theta)}{\partial \theta_2}, ..., \frac{\partial l(x;\theta)}{\partial \theta_k}\right]^T$$

Fisher Information Matrix can rewritten as (under regularity condition of EM) :

$$(I(\theta))_{i,j} = -E_X\left[\frac{\partial^2 l(x;\theta)}{\partial \theta_i \partial \theta_j}\right]$$

Cramer-Rao Bound (CRB) is defined as the inverse of Fisher information

$$\boxed{CRB = I^{-1}(\theta)} \longrightarrow \text{Cramer-Rao Bound}$$

Derivation

EM outputs the maximum likelihood estimation (MLE) of participant reliability

$$\hat{a}_i^{MLE} = \frac{\sum_{j \in SJ_i} Z_j^c}{\sum_{j=1}^{N} Z_j^c} \qquad \hat{b}_i^{MLE} = \frac{K_i - \sum_{j \in SJ_i} Z_j^c}{N - \sum_{j=1}^{N} Z_j^c}$$

where SJ_i is the set of measured variables reported by participant S_i, , Z_j^c is the converged value of $Z(t, j)$ (i.e., $p(z_j = 1 | X_j, \theta^{(t)})$) and K_i is the number of observations from participant S_i.

Derivation

Plugging $l_{em}(x;\theta)$ into the Fisher Information Matrix:

$$(I(\hat{\theta}_{MLE}))_{i,j} = \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{1}{N}\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{1}{N}\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

The inverse of above matrix is:

Diagonal Matrix!

$$(I^{-1}(\hat{\theta}_{MLE}))_{i,j} = \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{N}{\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{N}{\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

Derivation

Substituing $I^{-1}(\hat{\theta}_{MLE})$ into the normal distribution, the covariance matrix $Cov(\hat{\theta}_{MLE})$ for MLE of EM scheme is:

$$
\begin{aligned}
&(Cov(\hat{\theta}_{MLE}))_{i,j} \\
&= \begin{cases}
0 & i \neq j \\
\frac{\hat{a}_i^{MLE} \times (1 - \hat{a}_i^{MLE})}{N \times d} & i = j \in [1, M] \\
\frac{\hat{b}_i^{MLE} \times (1 - \hat{b}_i^{MLE})}{N \times (1-d)} & i = j \in (M, 2M]
\end{cases}
\end{aligned}
$$

Variance of MLE from EM!

Confidence Interval Derived

Given $a_i = \frac{t_i \times s_i}{d}$, $\left(\hat{t}_i^{MLE} - t_i^0\right)$ follows a norm distribution with 0 mean and variance given by:

$$Var\left(\hat{t}_i^{MLE}\right) = \left(\frac{d}{s_i}\right)^2 Var\left(\hat{a}_i^{MLE}\right)$$

Confidence Interval of reliability estimation of participant S_i $\left(\text{i.e, } \hat{t}_i^{MLE}\right)$:

Desired Confidence Interval!

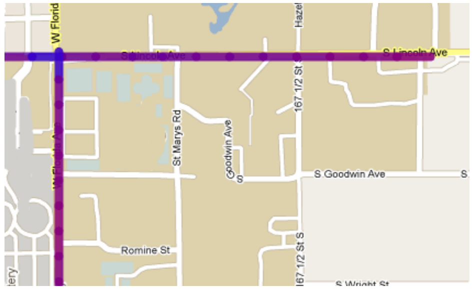
$$\boxed{\left(\hat{t}_i^{MLE} - c_p\sqrt{Var\left(t_i^{MLE}\right)},\ \ t_i^{MLE} + c_p\sqrt{Var\left(t_i^{MLE}\right)}\right)}$$

where is the standard score (z-score) of confidence level p

Example Applications

- Humans *operate* sensors: PictureMe

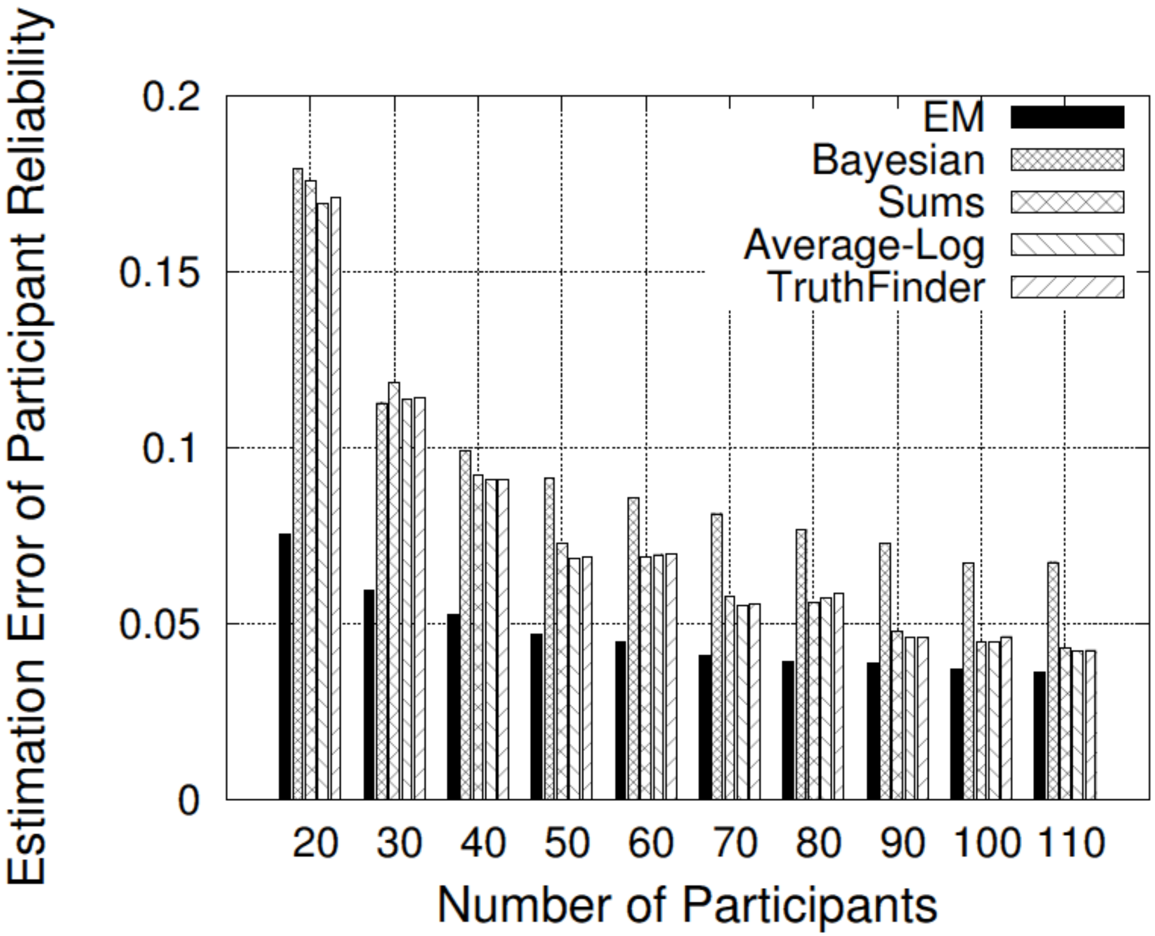
- Humans *carry* sensors: Speed Mapping



- Humans *are* the sensors: Event and timeline reconstruction from Tweets

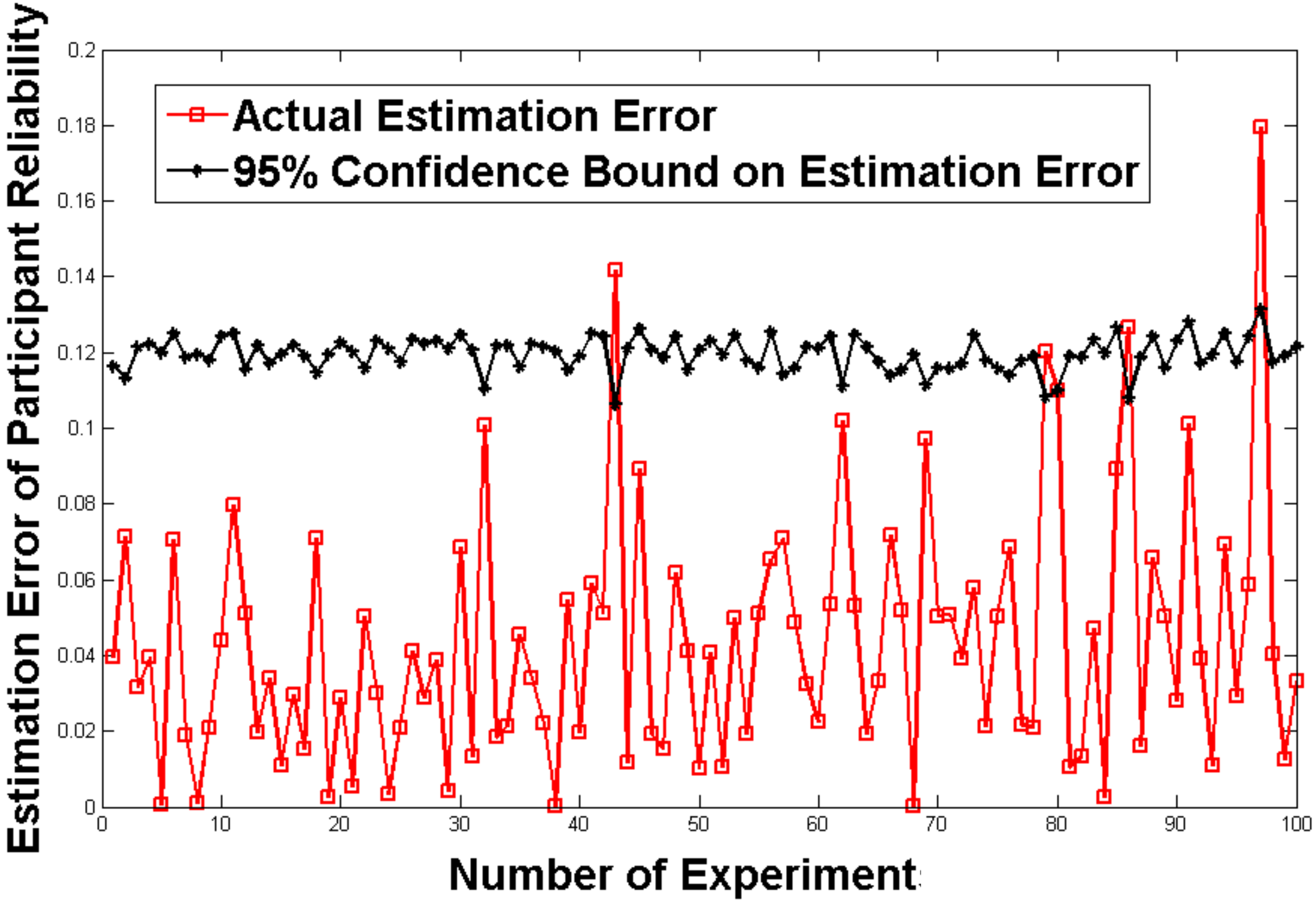
Evaluation
Estimation Error

- More accurate than state of the art fact-finders



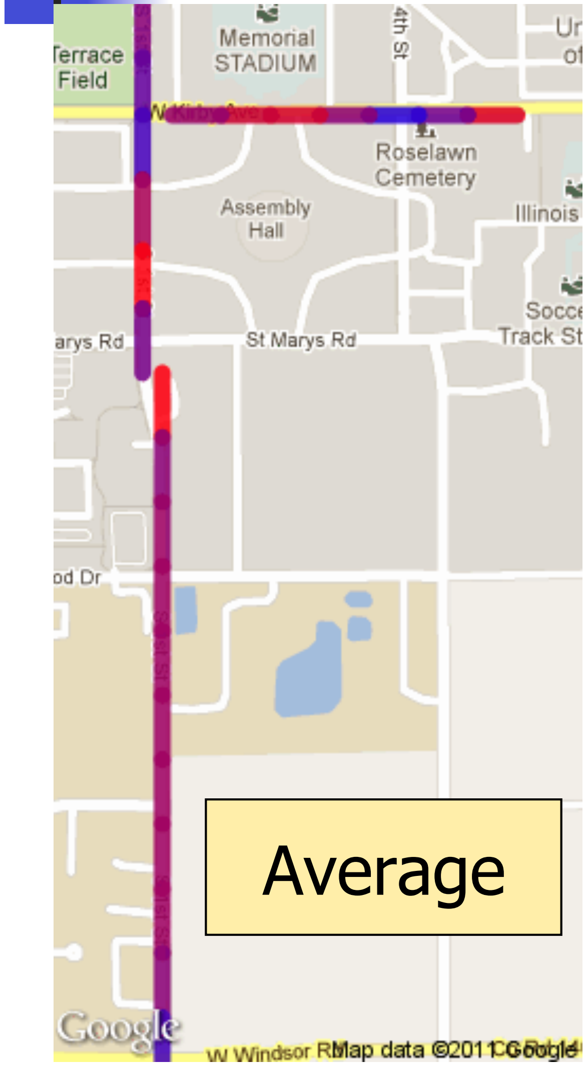
Evaluation
Error Bound

- Empirical data suggests the confidence interval is accurate

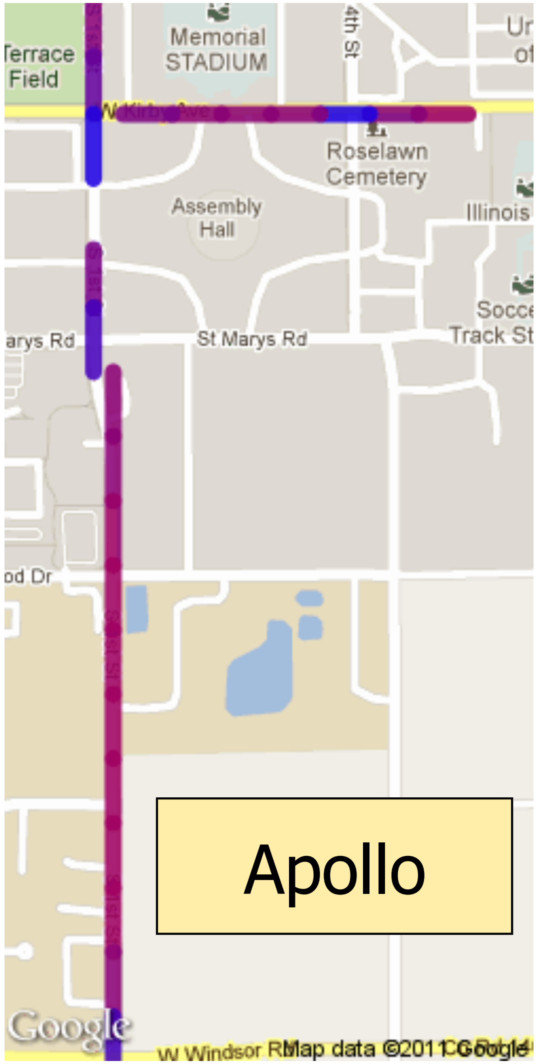


Apollo

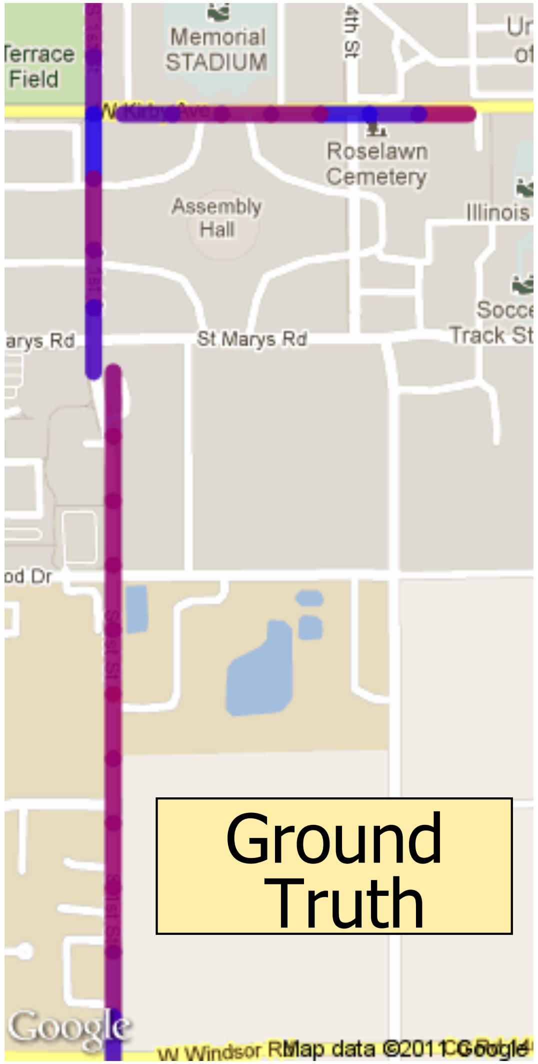
Cleaning Noisy Speed Data



Average

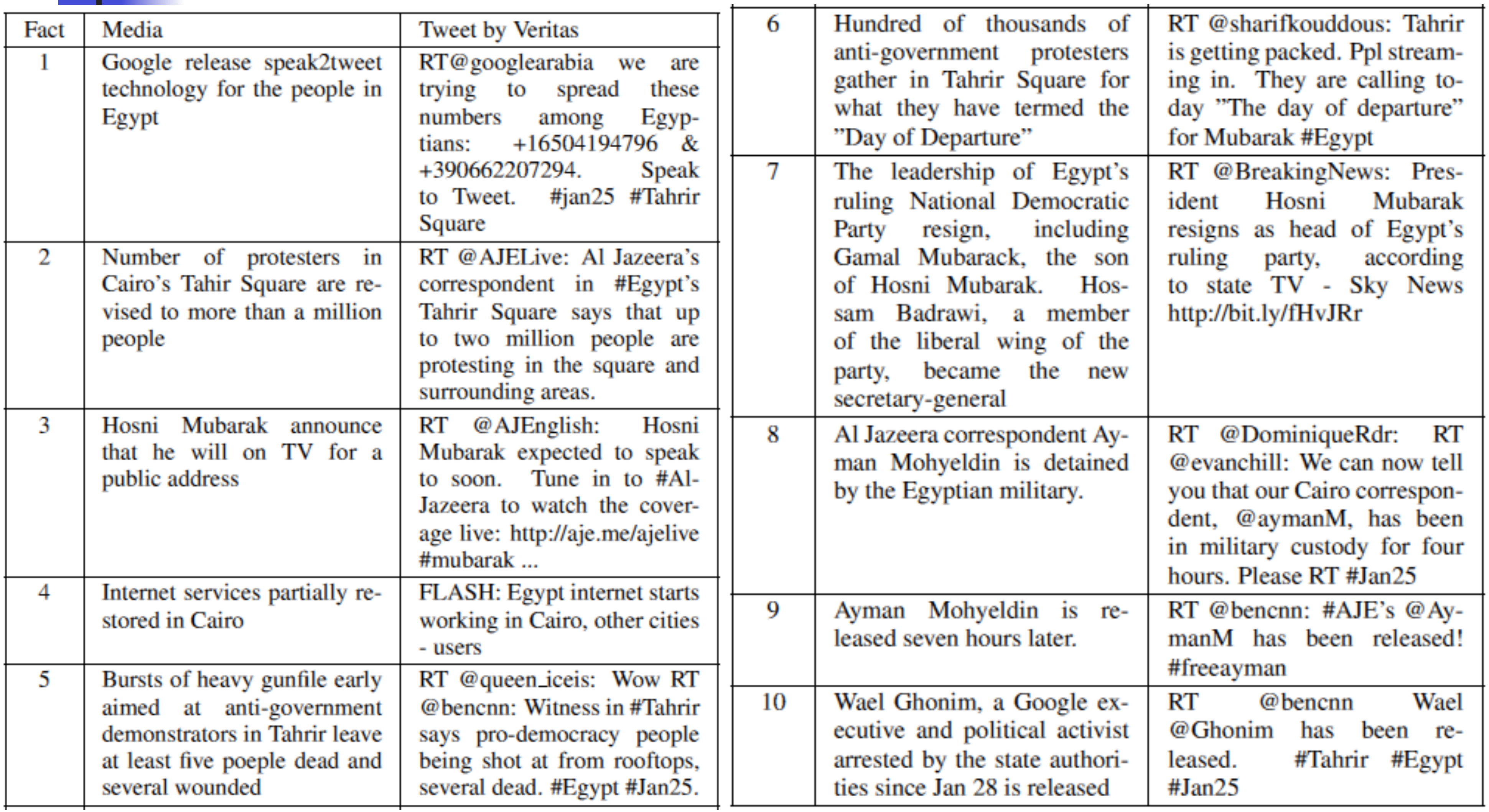


Apollo



Ground Truth

Apollo
Cleaning Noisy Twitter Data



Fact	Media	Tweet by Veritas
1	Google release speak2tweet technology for the people in Egypt	RT@googlearabia we are trying to spread these numbers among Egyptians: +16504194796 & +390662207294. Speak to Tweet. #jan25 #Tahrir Square
2	Number of protesters in Cairo's Tahir Square are revised to more than a million people	RT @AJELive: Al Jazeera's correspondent in #Egypt's Tahrir Square says that up to two million people are protesting in the square and surrounding areas.
3	Hosni Mubarak announce that he will on TV for a public address	RT @AJEnglish: Hosni Mubarak expected to speak to soon. Tune in to #Al-Jazeera to watch the coverage live: http://aje.me/ajelive #mubarak ...
4	Internet services partially restored in Cairo	FLASH: Egypt internet starts working in Cairo, other cities - users
5	Bursts of heavy gunfile early aimed at anti-government demonstrators in Tahrir leave at least five poeple dead and several wounded	RT @queen_iceis: Wow RT @bencnn: Witness in #Tahrir says pro-democracy people being shot at from rooftops, several dead. #Egypt #Jan25.
6	Hundred of thousands of anti-government protesters gather in Tahrir Square for what they have termed the "Day of Departure"	RT @sharifkouddous: Tahrir is getting packed. Ppl streaming in. They are calling today "The day of departure" for Mubarak #Egypt
7	The leadership of Egypt's ruling National Democratic Party resign, including Gamal Mubarack, the son of Hosni Mubarak. Hossam Badrawi, a member of the liberal wing of the party, became the new secretary-general	RT @BreakingNews: President Hosni Mubarak resigns as head of Egypt's ruling party, according to state TV - Sky News http://bit.ly/fHvJRr
8	Al Jazeera correspondent Ayman Mohyeldin is detained by the Egyptian military.	RT @DominiqueRdr: RT @evanchill: We can now tell you that our Cairo correspondent, @aymanM, has been in military custody for four hours. Please RT #Jan25
9	Ayman Mohyeldin is released seven hours later.	RT @bencnn: #AJE's @AymanM has been released! #freeayman
10	Wael Ghonim, a Google executive and political activist arrested by the state authorities since Jan 28 is released	RT @bencnn Wael @Ghonim has been released. #Tahrir #Egypt #Jan25

Social Sensing Challenge #3 Modeling: One Size Does Not Fit All

- Regression modeling:
  - Problem: one size does not fit all. Who says that Fords and Toyotas have the same fuel consumption model?
- Regression model per car?
  - Problem: How to use data collected by some cars to predict fuel consumption of others?
- Challenge: Must jointly determine both (i) regression models and (ii) their scope of applicability, to cover the whole data space with acceptable modeling error.

Generalization and Modeling

- Complex general system models with a large number of parameters are hard to train (need a lot of training data) and have a high inference cost (need a lot of inputs)
  - Poor cost/quality trade-off
- ***Main idea:*** Break-up complex general models into trees of simpler (but more specialized models)
  - Model has fewer parameters

    → less run-time data collection cost
  - Model may fit special case better

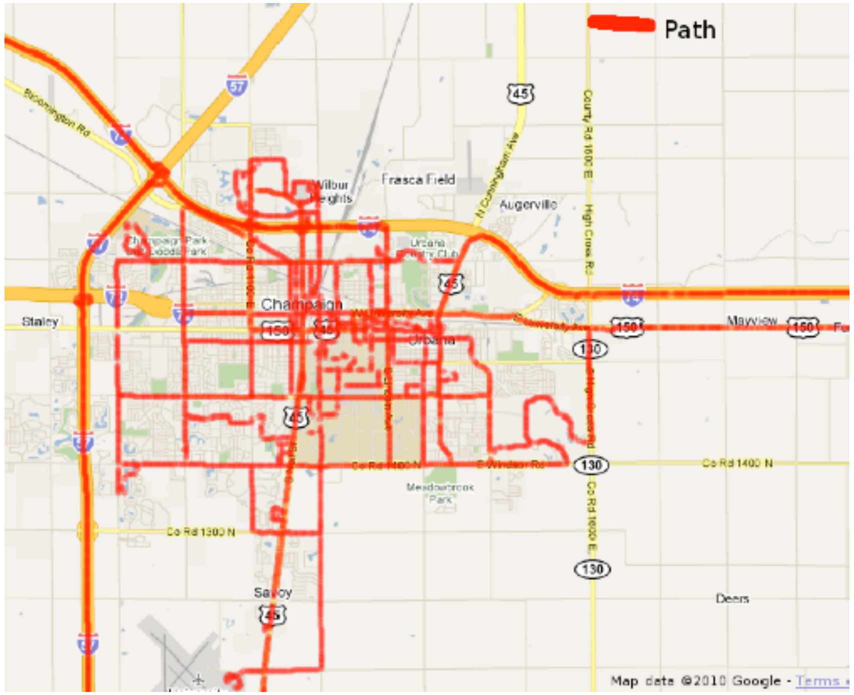
    → higher accuracy

  → ***Improved cost/quality trade-off!***

The Participant Data Modeling Challenge

- A *phenomenon is sampled* by participants in spatial and temporal dimensions
- Sampling is *sparse* (at least in conditions of partial adoption)
- The phenomenon is *high-dimensional*
- Question: *how to generalize* models obtained from the limited samples to cover the high-dimensional phenomenon space*?*

Sampling Regression Modeling Framework



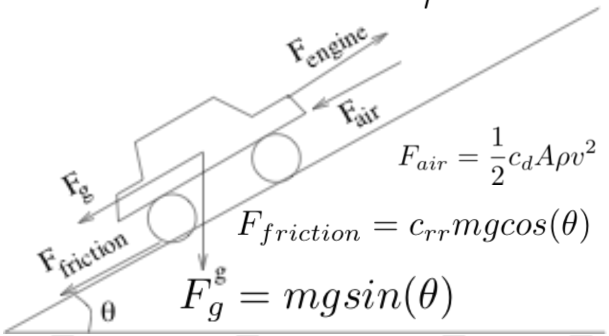
Fuel consumption of 16 cars driven on a few roads → Predict fuel consumption of any car on any road

Fuel Consumption Model

- Simple model for fuel consumption derived from physics principles
- Approximate based on easily measurable parameters (e.g. stop signs, speed limits)

$$gpm = k_1 m\bar{v}^2 \frac{ST + \nu TL}{\Delta d} + k_2 m \frac{\bar{v}^2}{\Delta d} + k_3 mcos(\theta) + k_4 A\bar{v}^2 + k_5 msin(\theta)$$

$$F_{engine} = \frac{\Gamma(\omega)Gg_k}{r}$$



$$F_{air} = \frac{1}{2}c_d A\rho v^2$$

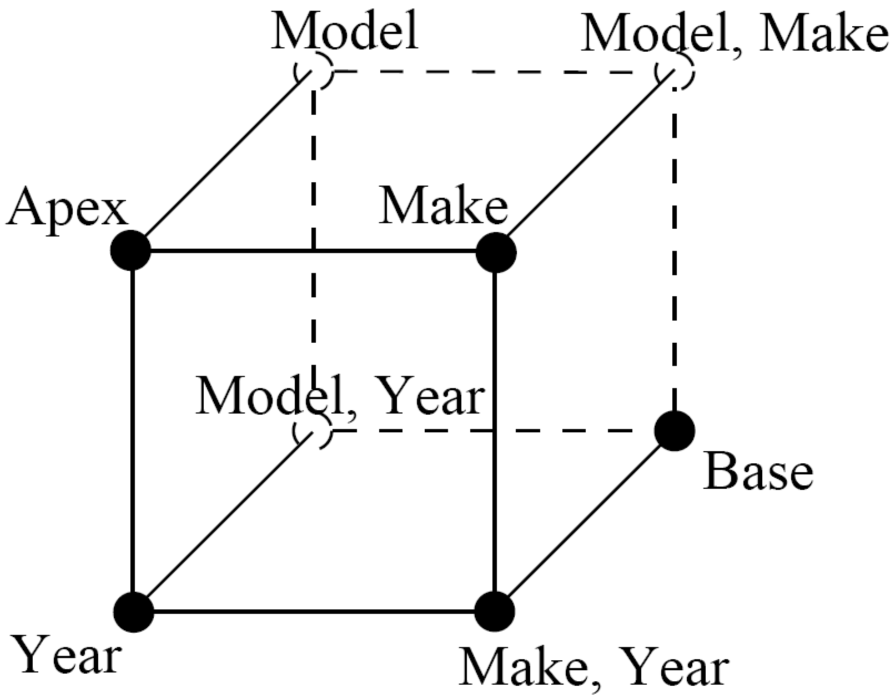
$$F_{friction} = c_{rr} mgcos(\theta)$$

$$F_g = mgsin(\theta)$$

$$F_{car} = F_{engine} - F_{friction} - F_{air} - F_g$$

A Generalization Hierarchy

- Goal: predict fuel consumption
  - Group by make, model, or year



Regression Cubes

- Data cells correspond to models derived from corresponding data subsets. In each cell, c:
  - Model output $Y_c = \{y_i\}$
  - Model inputs $x_{i1}, \ldots, x_{ik}, X_c = \{x_{ij}\}$

$$\hat{Y}_c = X_c \hat{\eta}_c$$

  - Regression modeling error:

$$Err_c = (Y_c - X_c \hat{\eta}_c)^T (Y_c - X_c \hat{\eta}_c)$$

The Challenge of Regression Cubes

- Main challenge: compute cuboid measures, the model and error, recursively (without reprocessing raw data)
- Model parameters and estimation error at cell c
  - Not distributive

$$\hat{Y}_c = X_c \hat{\eta}_c$$

$$Err_c = (Y_c - X_c\hat{\eta}_c)^T (Y_c - X_c\hat{\eta}_c)$$

Efficient Representation

- Compressed representation of a cell c:
  - $\rho_c = Y_c^T Y_c$: scalar value
  - $\nu_c = X_c^T Y_c$: vector of size k (number of inputs)
  - $\Theta_c = X_c^T X_c$: k by k matrix
  - n_c : number of samples
- A cell c may be the union of several smaller cells (e.g., all Toyota cars):

$$\rho_c = \sum_{i=1}^{m} \rho_i \qquad \nu_c = \sum_{i=1}^{m} \nu_i \qquad \Theta_c = \sum_{i=1}^{m} \Theta_i \qquad n_c = \sum_{i=1}^{m} n_{c_i}$$

Efficient Model Parameter and Error Computation

$$\rho_c = \sum_{i=1}^{m} \rho_i \qquad \nu_c = \sum_{i=1}^{m} \nu_i \qquad \Theta_c = \sum_{i=1}^{m} \Theta_i \qquad n_c = \sum_{i=1}^{m} n_{c_i}$$

- Model coefficients:

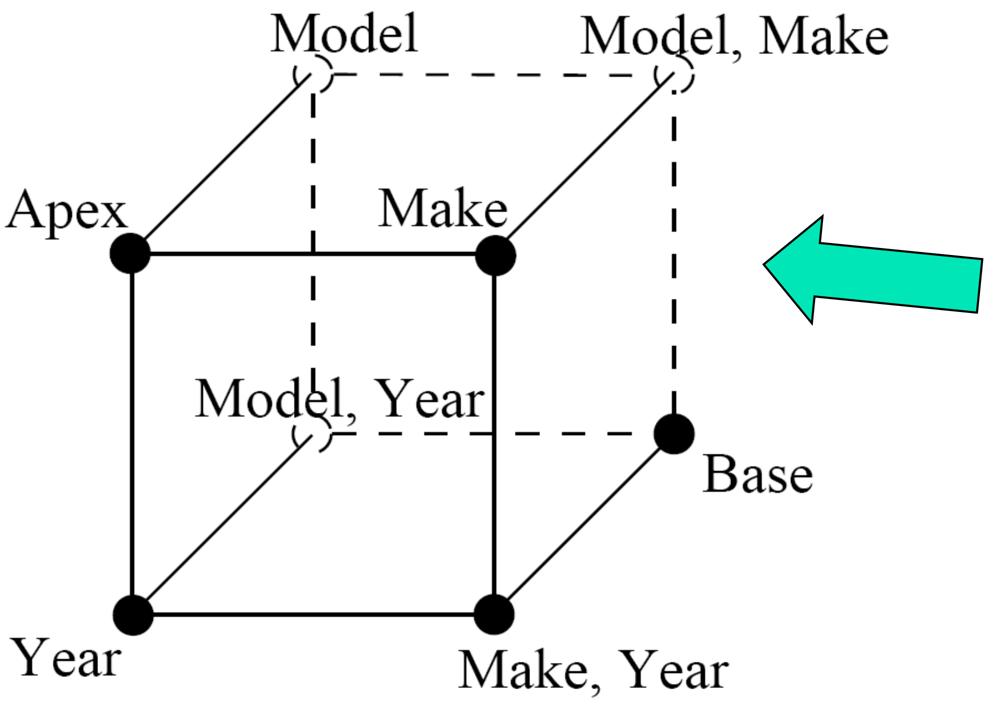
$$\hat{\eta}_c = (X_c^T X_c)^{-1} X_c^T Y_c = \Theta_c^{-1} \nu_c$$

- Error:

$$\begin{aligned} Err_c = (Y_c - X_c\hat{\eta}_c)^T (Y_c - X_c\hat{\eta}_c) = \\ Y_c^T Y_c - (X_c\hat{\eta}_c)^T Y_c - Y_c^T X_c\hat{\eta}_c + (X_c\hat{\eta}_c)^T X_c\hat{\eta}_c = \\ \rho_c - \hat{\eta}_c^T \nu_c - \nu_c^T \hat{\eta}_c + \hat{\eta}_c^T \Theta_c \hat{\eta}_c \end{aligned}$$

GreenGPS Regression Cubes

- Goal: predict fuel consumption



Model and modeling error are efficiently computed for each possible generalization.

Model Reduction

- Independently find *a subset of attributes* for each cell, such that:
  - The cell is reliable
  - Corresponding error is minimized
  - Exponential number of possible subsets
- Our heuristic:

	Error	Reliable
L = {v}	0.031	yes
L = {m}	0.152	yes
L = {A}	0.043	yes
L = {S}	0.056	yes

Attributes
Velocity (v)
Mass (m)
Frontal area (A)
Stop signs (S)

Model Reduction

Attributes
Velocity (v)
Mass (m)
Frontal area (A)
Stop signs (S)

	Error	Reliable
L = {v}	0.031	yes
L = {m}	0.152	yes
L = {A}	0.043	yes
L = {S}	0.056	yes

Model Reduction

Attributes
Velocity (v)
Mass (m)
Frontal area (A)
Stop signs (S)

L = {v}
L = {m}
L = {A}
L = {S}

→

	Error	Reliable
L = {v, m}	0.021	no
L = {v, A}	0.030	yes
L = {v, S}	0.028	yes

Model Reduction

Attributes
Velocity (v) Mass (m) Frontal area (A) Stop signs (S)

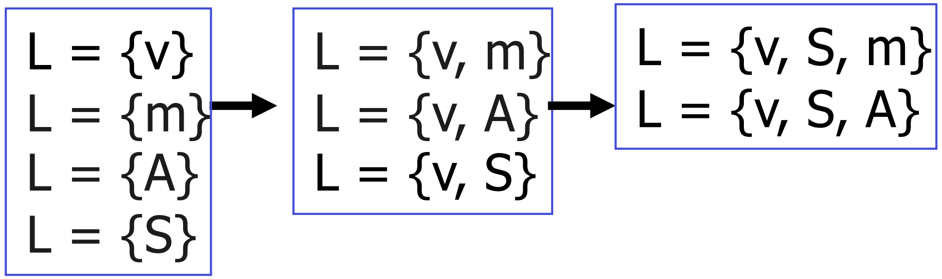
L = {v}
L = {m}
L = {A}
L = {S}

→

	Error	Reliable
L = {v, m}	0.021	no
L = {v, A}	0.030	yes
L = {v, S}	0.028	yes

Model Reduction

Attributes
Velocity (v)
Mass (m)
Frontal area (A)
Stop signs (S)



L = {v}
L = {m}
L = {A}
L = {S}

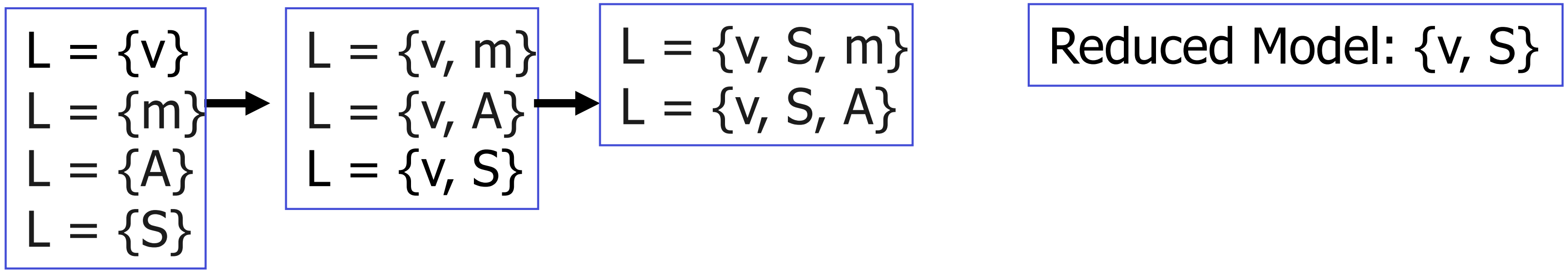
L = {v, m}
L = {v, A}
L = {v, S}

L = {v, S, m}
L = {v, S, A}

Error	Reliable
0.024	no
0.026	no

Model Reduction

Attributes
Velocity (v) Mass (m) Frontal area (A) Stop signs (S)



L = {v}
L = {m}
L = {A}
L = {S}

L = {v, m}
L = {v, A}
L = {v, S}

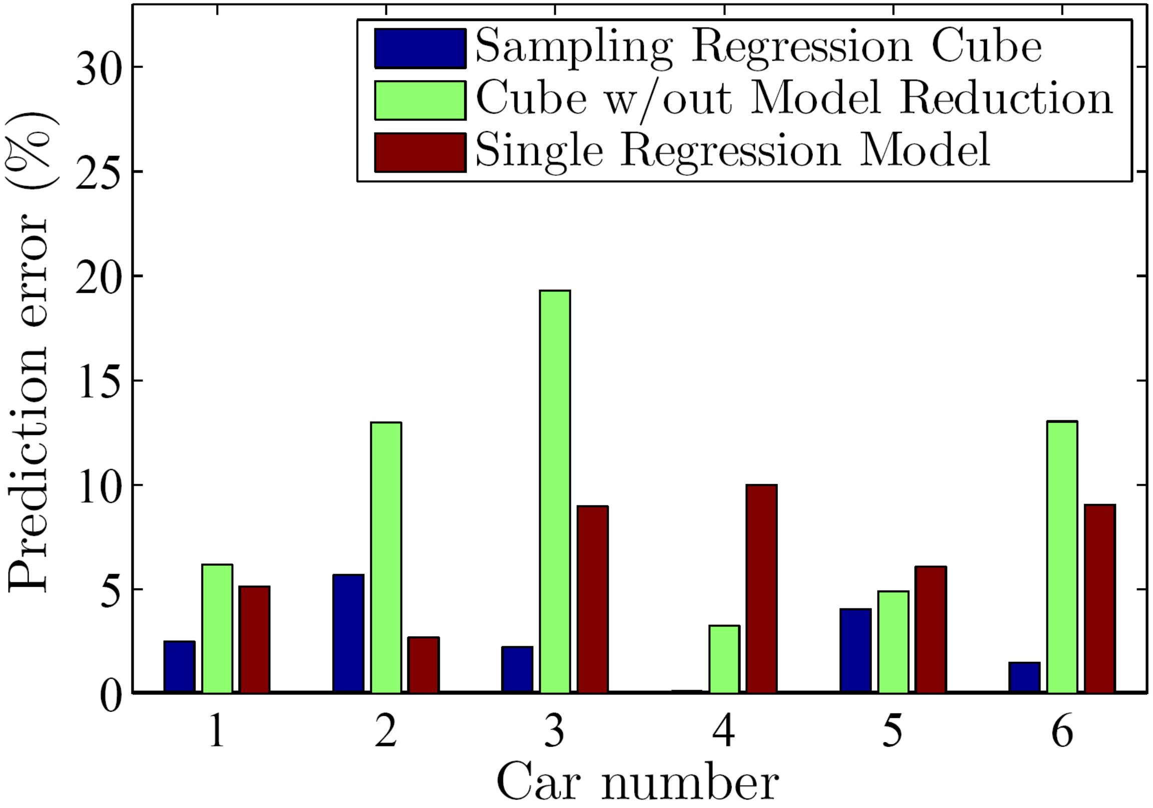
L = {v, S, m}
L = {v, S, A}

Reduced Model: {v, S}

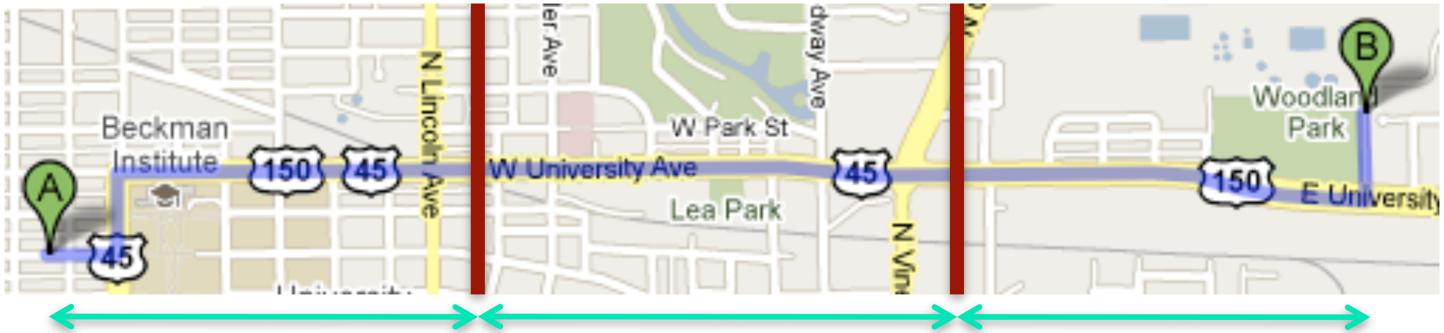
Accuracy Results

- The sampling regression cube improves prediction accuracy significantly

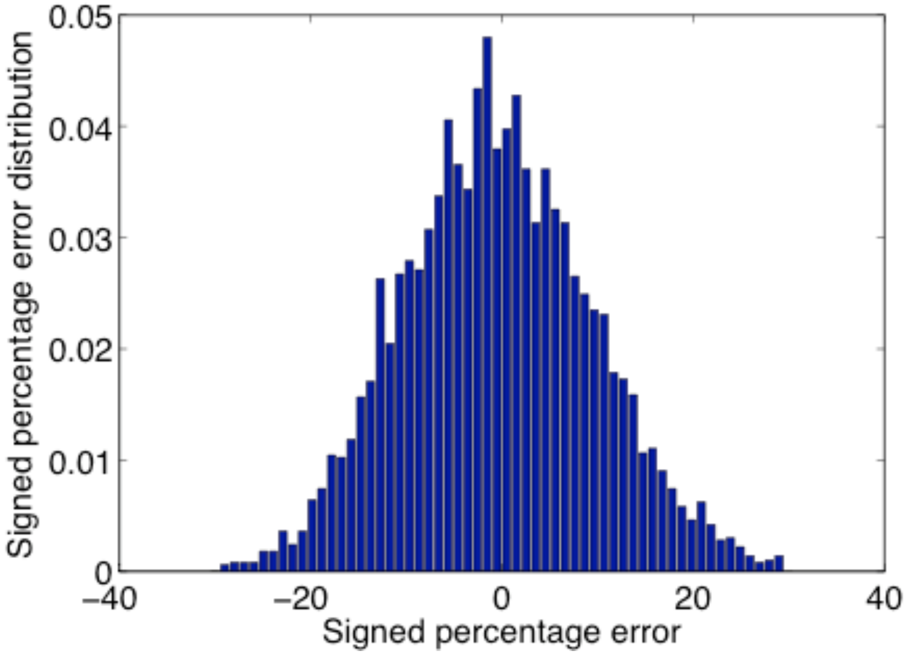
Sparse sampling challenge: A regression cube without model reduction is worse than a single "one-size fits-all" model!



Model Performance

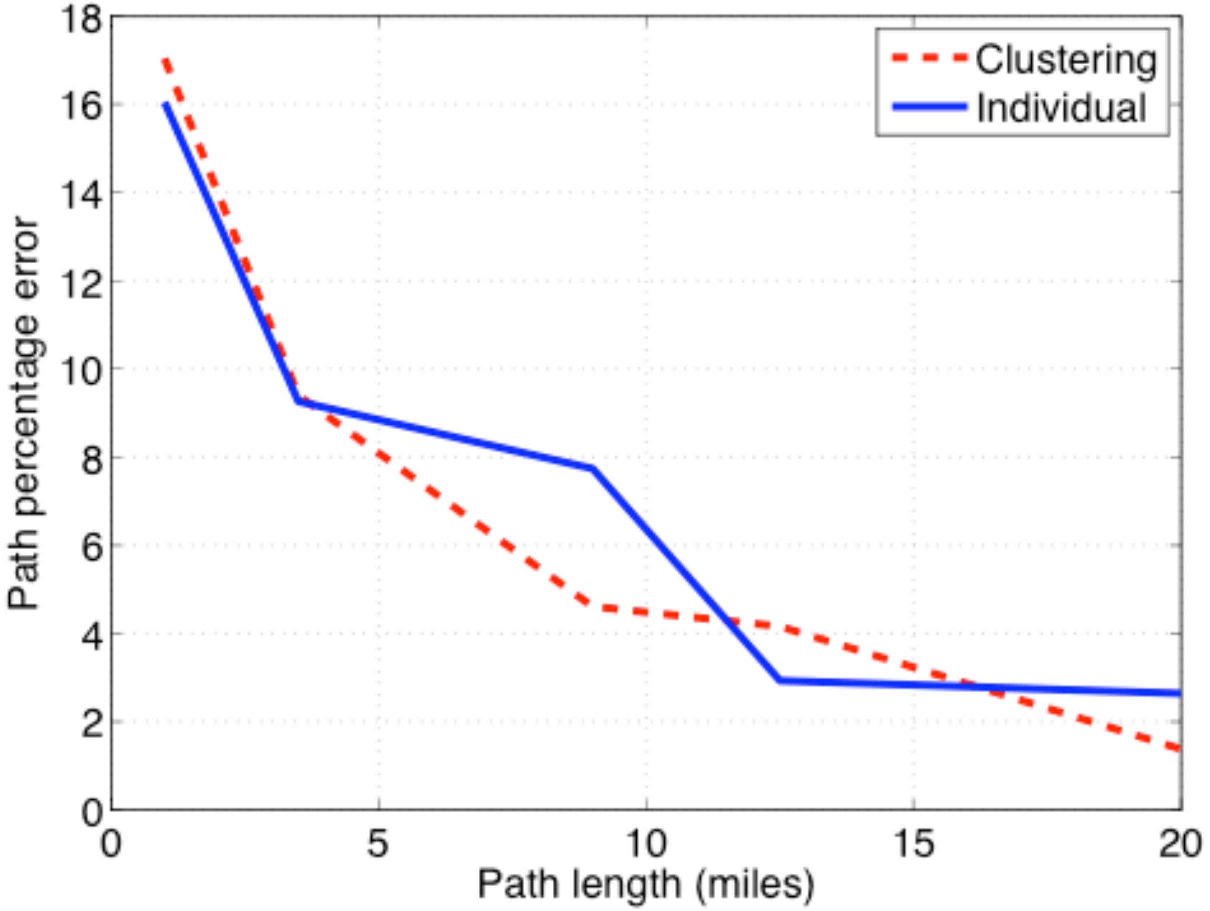


Segment 1 Segment 2 Segment 3



All driven paths are split into smaller *segments* to capture variations in fuel consumption on individual streets

Long Path Error



Reduction in cumulative error with increasing path length

Fuel Savings Evaluation

- Experiment:
  - Given shortest and fastest routes, GreenGPS predicts best route.
  - Driver drives both routes repeatedly and compares average fuel consumption of the two.

Car Details	Landmarks	Route	Savings %
Honda Accord 2001	H1 to Mall	Shortest	31.4
	H1 to Gym	Shortest	19.7
Ford Taurus 2001	H2 to Restaurant	Shortest	26
Toyota Celica 2001	H2 to Work	Fastest	10.1
Nissan Sentra 2009	H3 to CUPHD	Fastest	8.4
Honda Civic 2002	Grad to Work	Fastest	18.7

Comment #1: Privacy - Revisited

- Can we offer privacy without data perturbation (or encryption)?
- The Problem: It is desired to derive a model (e.g., fuel-efficiency of a car) from inputs and outputs that are private
  - The model itself is not private
  - The inputs and outputs are!

Reminder: Efficient Representation

- Compressed representation of a cell c:
  - $\rho_c = Y_c^T Y_c$: scalar value
  - $\nu_c = X_c^T Y_c$: vector of size k (number of inputs)
  - $\Theta_c = X_c^T X_c$: k by k matrix
  - n_c : number of samples
- A cell c may be the union of several smaller cells (e.g., all Toyota cars):

$$\rho_c = \sum_{i=1}^{m} \rho_i \qquad \nu_c = \sum_{i=1}^{m} \nu_i \qquad \Theta_c = \sum_{i=1}^{m} \Theta_i \qquad n_c = \sum_{i=1}^{m} n_{c_i}$$

Reminder: Model Parameter and Error Computation

$$\rho_c = \sum_{i=1}^{m} \rho_i \qquad \nu_c = \sum_{i=1}^{m} \nu_i \qquad \Theta_c = \sum_{i=1}^{m} \Theta_i \qquad n_c = \sum_{i=1}^{m} n_{c_i}$$

- Model coefficients:

$$\hat{\eta}_c = (X_c^T X_c)^{-1} X_c^T Y_c = \Theta_c^{-1} \nu_c$$

- Error:

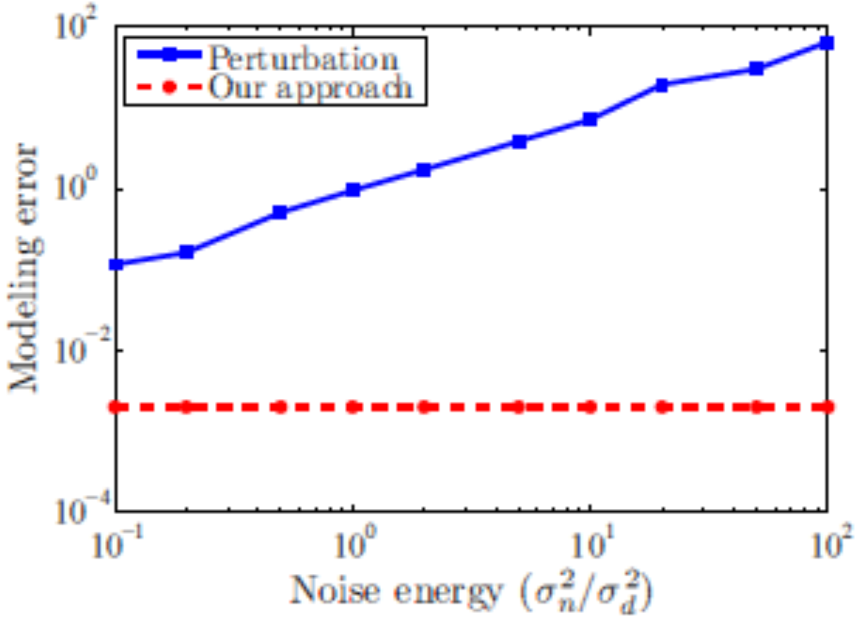
$$Err_c = (Y_c - X_c\hat{\eta}_c)^T (Y_c - X_c\hat{\eta}_c) =$$

$$Y_c^T Y_c - (X_c\hat{\eta}_c)^T Y_c - Y_c^T X_c\hat{\eta}_c + (X_c\hat{\eta}_c)^T X_c\hat{\eta}_c =$$

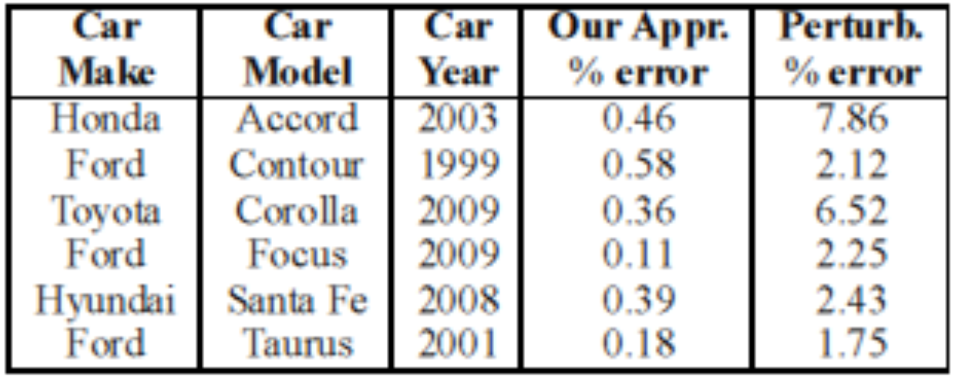
$$\rho_c - \hat{\eta}_c^T \nu_c - \nu_c^T \hat{\eta}_c + \hat{\eta}_c^T \Theta_c \hat{\eta}_c$$

Evaluation (Privacy-preserving Regression vs. Perturbation)

- No additional error is introduced into modeling



Lower total modeling error

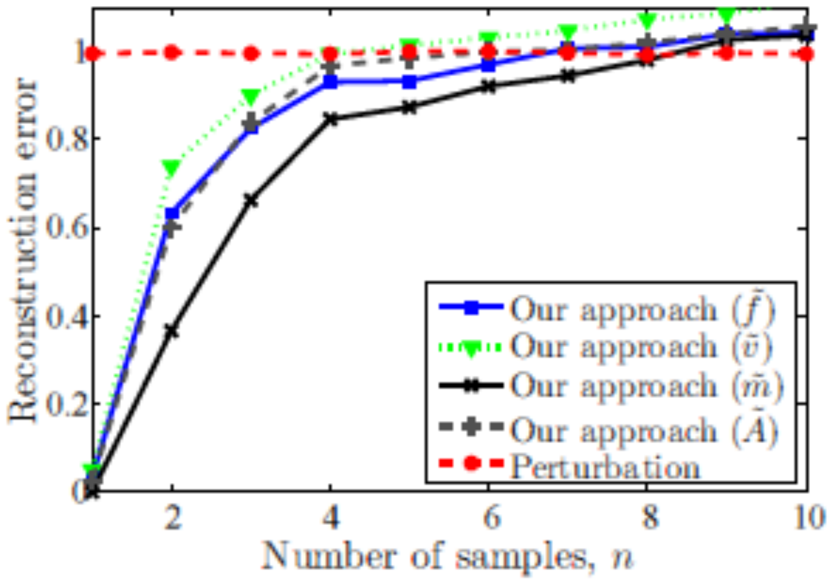


Car Make	Car Model	Car Year	Our Appr. % error	Perturb. % error
Honda	Accord	2003	0.46	7.86
Ford	Contour	1999	0.58	2.12
Toyota	Corolla	2009	0.36	6.52
Ford	Focus	2009	0.11	2.25
Hyundai	Santa Fe	2008	0.39	2.43
Ford	Taurus	2001	0.18	1.75

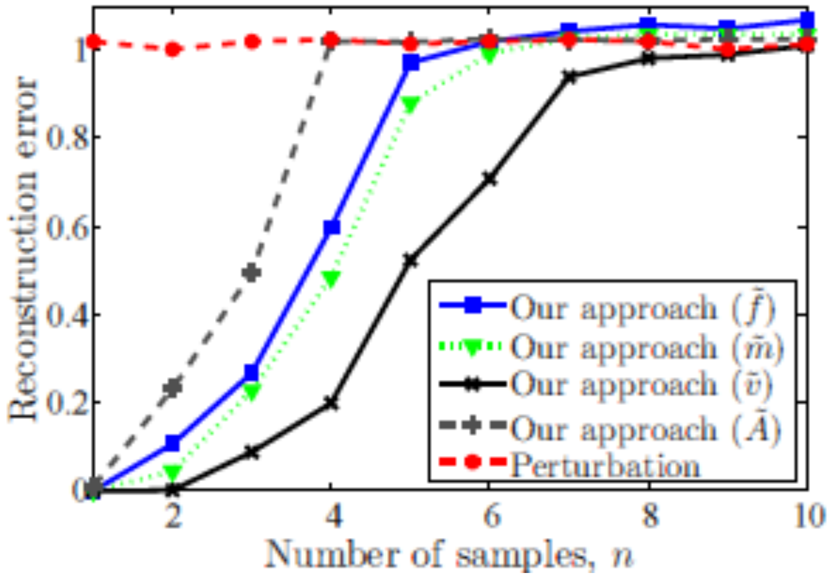
Better prediction of gas consumption for individual vehicles

Evaluation (Privacy-preserving Regression vs. Perturbation)

- Single-stream reconstruction accuracy



3-parameter model

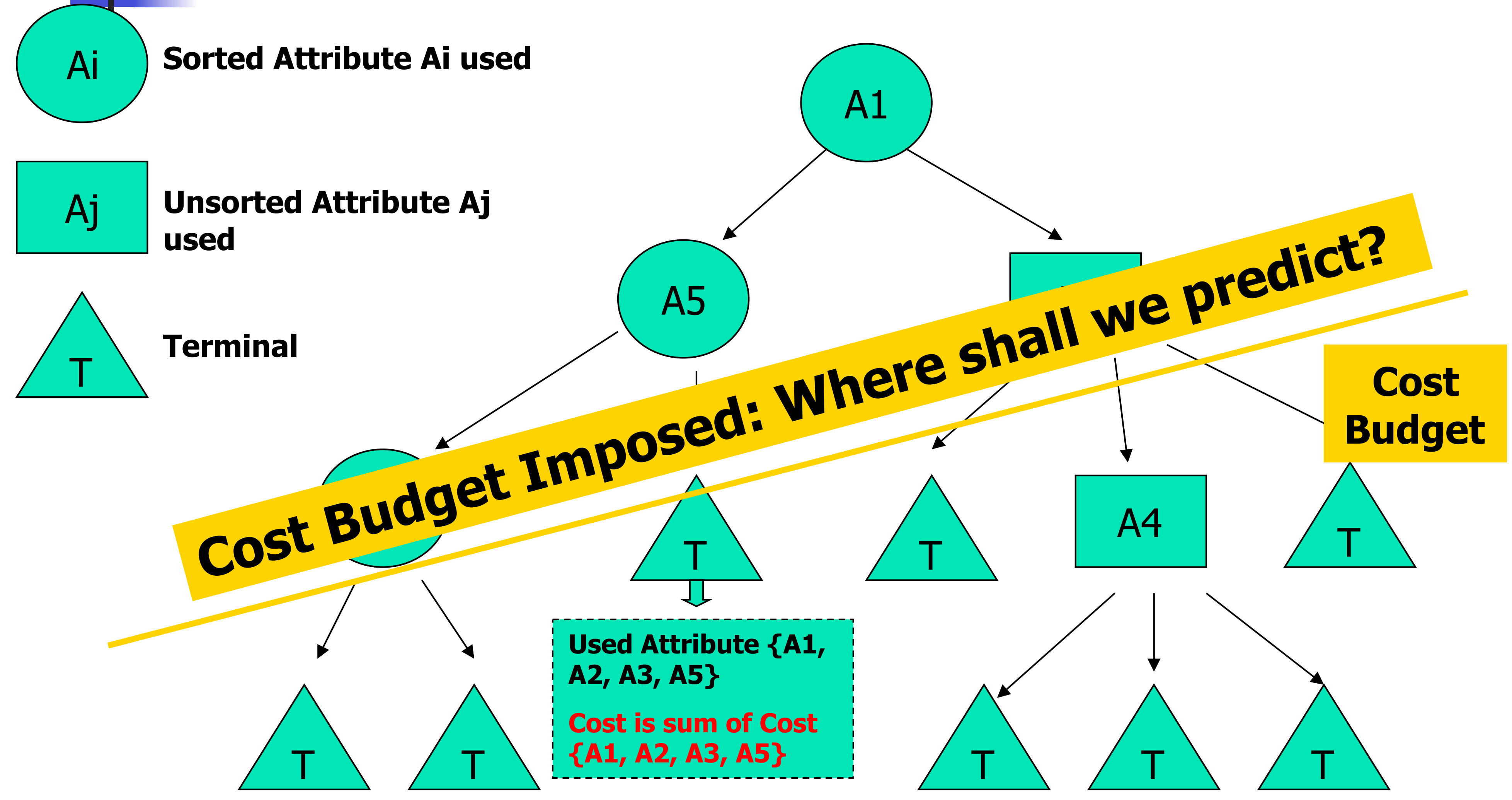


4-parameter model

Comment #2: Cost-sensitive Regression

- What if data collection had costs? Is it possible to derive models that are cost sensitive?

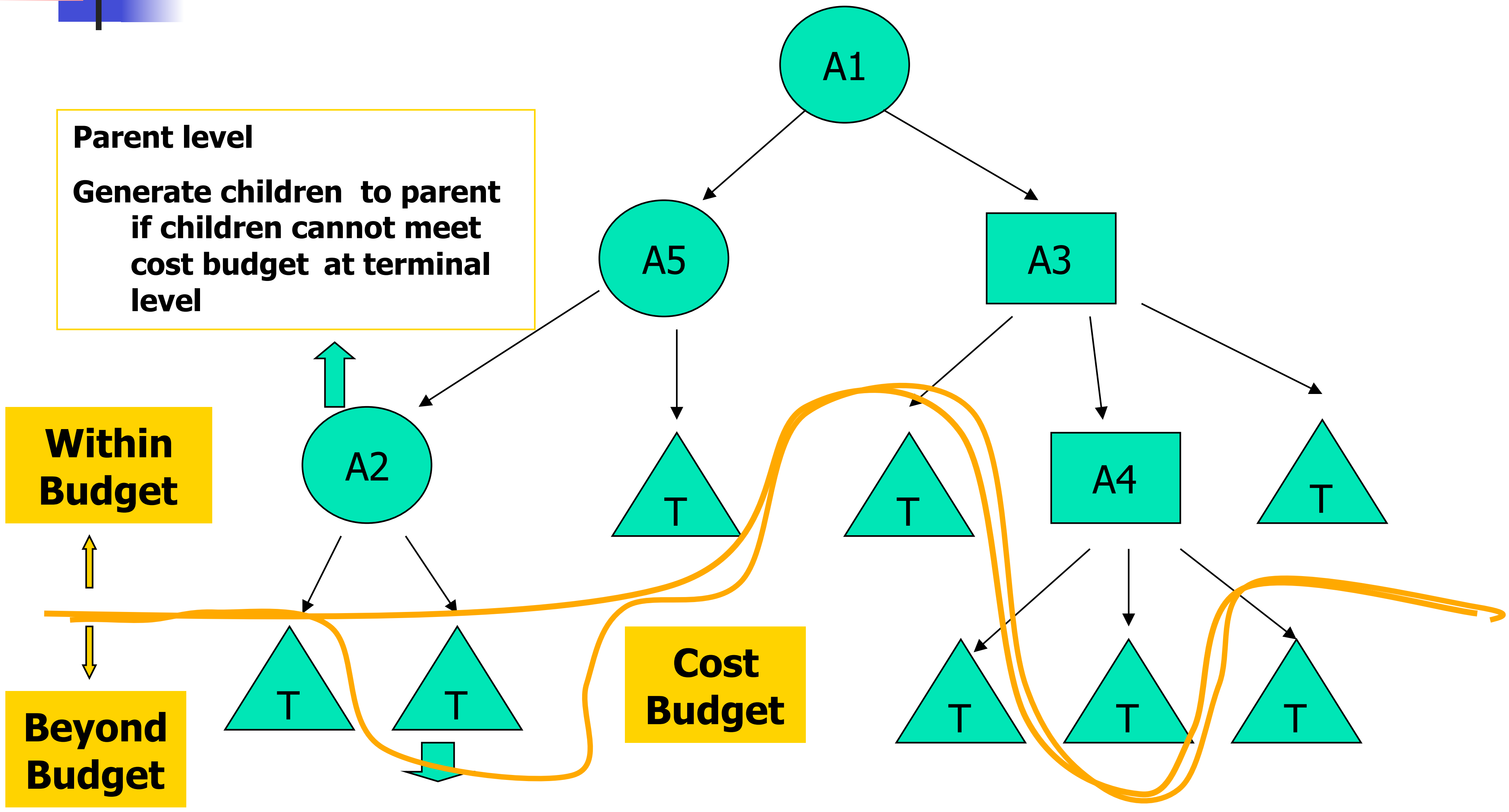
Cost-sensitive Regression


Ai
Sorted Attribute Ai used
Aj
Unsorted Attribute Aj used
T
Terminal
A1
A5
A4
T
T
T
T
T
T
T
T
Cost Budget
Cost Budget Imposed: Where shall we predict?

Used Attribute {A1, A2, A3, A5}

Cost is sum of Cost {A1, A2, A3, A5}

Cost-sensitive Regression



Parent level

Generate children to parent if children cannot meet cost budget at terminal level

Evaluation on GreenGPS

Method Used	Prediction Error (%)	Cost
Single Model* (Cost-insensitive)	34.39%	35
Cube Model (Cost-insensitive)	21.25%	33
Cost-insensitive Hybrid Regress Tree	19.47%	34
Cost-sensitive Hybrid Regress Tree	18.88%	23

*Single Model: use all data (without splitting into subspaces) to build a single regression model to predict

Conclusions

- Social sensing systems are becoming ubiquitous
- Some problems become more important
  - Privacy, fact-finding (data cleaning), quality of information, modeling, robustness, …
- Needed:
  - Analytic results for collection and use of social sensing data (accuracy estimation, privacy-preserving perturbation, modeling, control, …)
  - A tool set to embody the analytic results (obfuscation tools, fact-finders, modeling libraries, …)
- Planned deployment: GreenGPS on 100 cars

Publications (1/4)

Green GPS

- Raghu Ganti, Nam Pham, Hossein Ahmadi, Saurabh Nangia, Tarek Abdelzaher, "GreenGPS: A Participatory Sensing Fuel-Efficient Maps Application," *Mobisys*, San Francisco, CA, June 2010.
- Tarek Abdelzaher, "Green GPS-assisted Vehicular Navigation," *Handbook of Energy-Aware and Green Computing,* Chapman & Hall/CRC, expected in 2011.

Publications (2/4)

Privacy

- Hossein Ahmadi, Nam Pham, Raghu Ganti, Tarek Abdelzaher, Suman Nath, Jiawei Han, "Privacy-aware Regression Modeling of Participatory Sensing Data," *Sensys*, Zurich, Switzerland, November 2010.
- Nam Pham, Tarek Abdelzaher, Suman Nath, "On Bounding Data Stream Privacy in Distributed Cyber-physical Systems," *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (IEEE SUTC)*, Newport Beach, CA, June, 2010. (Invited)
- Nam Pham, Raghu Ganti, Md. Yusuf Uddin, Suman Nath, Tarek Abdelzaher, "Privacy-Preserving Reconstruction of Multidimensional Data Maps in Vehicular Participatory Sensing," *European Conference on Wireless Sensor Networks (EWSN)*, Coimbra, Portugal, February, 2010.
- Raghu Ganti, Nam Pham, Yu-En Tsai, Tarek Abdelzaher "PoolView: Stream Privacy for Grassroots Participatory Sensing," *Sensys*, Raleigh, NC, November 2008.

Publications (3/4)

Data Cleaning

- Dong Wang, Tarek Abdelzaher, Hossein Ahmadi, Jeff Pasternack, Dan Roth, Manish Gupta, Jiawei Han, Omid Fatemieh, Hieu Le, Charu Aggrawal, "On Bayesian Interpretation of Fact-finding in Information Networks," in Proc *14th International Conference on Information Fusion (Fusion '11)*, Chicago, IL, July 2011.
- Dong Wang, Tarek Abdelzaher, Lance Kaplan, Charu Aggarwal, "On Quantifying the Accuracy of Maximum Likelihood Estimation of Participant Reliability in Social Sensing," 7th International Workshop on Data Management for Sensor Networks, 2012, August 2011

Publications (4/4)

Modeling

- Dong Wang, Hossein Ahmadi, Tarek Abdelzaher, Harsha Chenji, Radu Stoleru, Charu Aggarwal, "Optimizing Quality-of-Information in Cost-sensitive Sensor Data Fusion," *IEEE DCoSS*, Barcelona, Spain, June 2011.
- Hossein Ahmadi, Tarek Abdelzaher, Jiawei Han, Raghu Ganti and Nam Pham, "On Reliable Modeling of Open Cyber-physical Systems and its Application to Green Transportation," ICCPS, Chicago, IL, April 2011.