

## Challenges in Social Sensing

Tarek Abdelzaher  
University of Illinois at Urbana Champaign

## The Rise of Social Sensing

People

Sensors

Data

Analytics

Future Applications

## Social Sensing: A Confluence of Three Trends

**Mass Dissemination Media**

twitter flickr facebook youtube twitpic wimax

**Connectivity**

Game Consoles on Internet

Cell-phones

Cars on Internet

Pulse oximeter

Smart Meter

**Sensors**

Glucose monitor

GPS

Sportsware

## An Architecture for Social Sensing

People

Mobiles

Sensors

Critical Services

## Our Goal

- Build a software platform to support applications of social sensing
- Approach
  - Develop a set of representative apps
  - Investigate services common to that set
    - On the front end (phones)
    - In the back-end (the cloud)

Complex Socio-physical System

Data Processing

## The UIUC Application: Transportation Energy Efficiency

- 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

**Server**

Fuel Data + Physical Models

$$F_{\text{fuel}} = F_{\text{engine}} + F_{\text{aerodynamic}} + F_{\text{rolling}} + F_{\text{braking}}$$

$$F_{\text{aerodynamic}} = \frac{1}{2} \rho v^2 C_d A$$

$$F_{\text{rolling}} = c_r m g \sin(\theta)$$

$$F_{\text{braking}} = F_{\text{engine}} - F_{\text{aerodynamic}} - F_{\text{rolling}} - F_g$$

## 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 GreenGPS may differ from the shortest or fastest route.

Lark Abdelzaher  
Ph.D. Professor

shortest  
fastest  
fuelOptim

## Challenges in Developing Social Sensing Applications

- Privacy
  - How to enable people to share data without violating their privacy?
- Cleaning
  - How to determine reliability of data and sources?
- Modeling and prediction
  - How to generalize from incomplete data?

## The Privacy Challenge

- 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

## Intuitive Approach

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

Real user + Perturbed data curve = Virtual user

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

### 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 deconvolution to reconstruct original data distribution at any point in time

### Perturbing Speed and Location

- Clients lie about both their location and speed

### Reconstruction Accuracy

- Real versus reconstructed speed

### More on Reconstruction Accuracy

- Real versus reconstructed speed on Washington St., Champaign

### 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%

## The Data Cleaning Challenge

- In social sensing applications, participants may not be known or vetted *a priori*
- Some data may be incorrect and some sources unreliable
- How to tell good from bad sources?

## The Problem

Human are involved in the sensing and data fusion loop

What to believe and Who shall we believe

Quantitatively?

Detailed prior knowledge on source reliability is unknown.

## 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

## Apollo Analytic Contributions

- Formulation of the fact-finding problem as one of maximum likelihood estimation
- Solution using the *Expectation Maximization* (EM) algorithm
- Computing a bound on estimation accuracy (using the Cramer Rao Bound)

## Example Applications

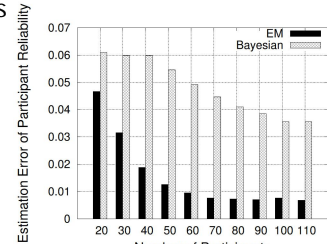
- Humans *operate* sensors: Geo-tagging and PictureMe
- Humans *carry* sensors: Speed Mapping
- Humans *are* the sensors: Event and timeline reconstruction from Tweets



25

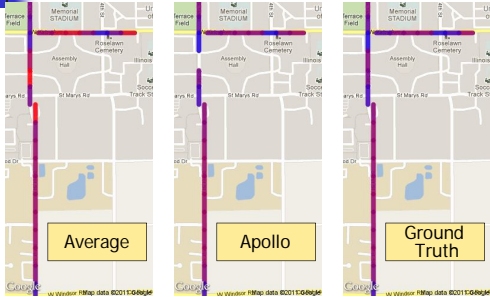
## Evaluation

- More accurate than state of the art fact-finders



26

## Apollo Cleaning Noisy Speed Data



27

## 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 & +39662707294. Speak to Tweet #Jan25 #Tahrir Square	6	Hundred of thousands of anti-government protesters gather in Tahrir Square for what they have termed the "Day of Departure" for Mubarak #Egypt
2	Number of protesters in Cairo's Tahrir 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.	7	The leadership of Egypt's ruling National Democratic Party resign, including Gamal Mubarak, the son of Hosni Mubarak. Hosni Mubarak, a member of the liberal wing of the party, became the new secretary-general.
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 #AlJazeera to watch the coverage live: <a href="http://ajp.me/ajelive/mubarak">http://ajp.me/ajelive/mubarak</a>	8	Al Jazeera correspondent Ayman Mohyeldin is detained by the Egyptian military.
4	Internet services partially restored in Cairo	FLASH: Egypt internet starts working in Cairo, other cities users	9	Ayman Mohyeldin is released seven hours later.
5	Bursts of heavy gunfire early aimed at anti-government demonstrators in Tahrir leave at least five people dead and several wounded	RT @queen.acts: Wow RT @bessem: Witness in Tahrir says pro-democracy people being shot at from rooftops, several dead. #Egypt #Jan25	10	Wael Ghonim, a Google executive and political activist arrested by the state authorities since Jan. 28 is released.

28

## Generalization and Modeling

- Regression modeling:
  - Problem: one size does not fit all. Who says that Fords and Toyotas have the same regression 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.

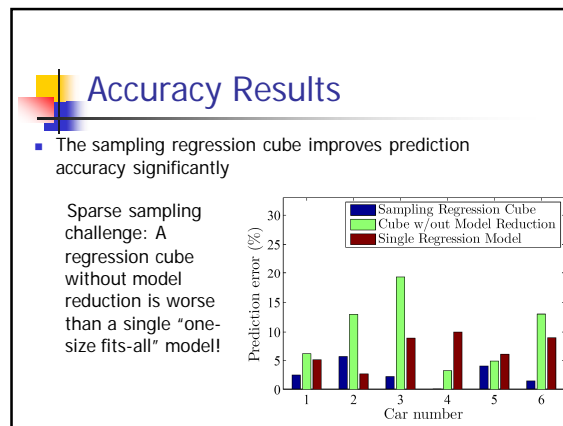
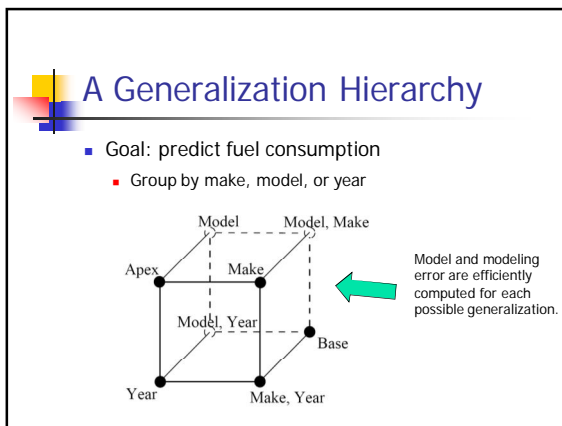
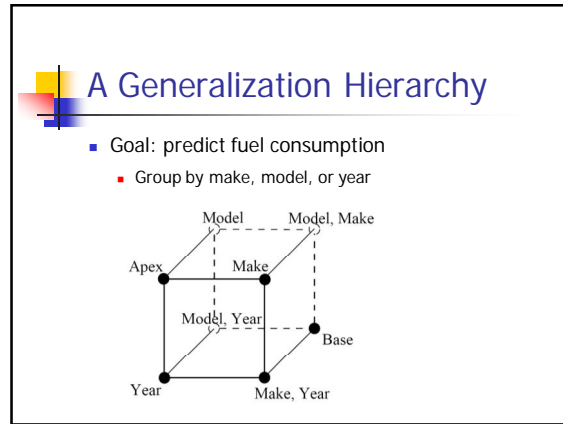
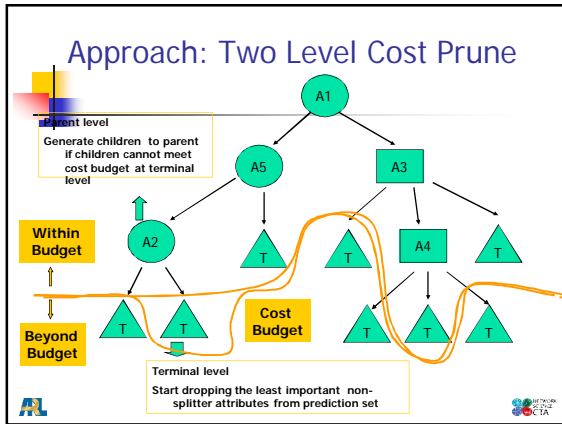
29

## 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!**

30



- ### Conclusion
- Social sensing systems are becoming ubiquitous
  - Some problems become more important
    - Privacy, data cleaning, quality of information, modeling, data analytics, inference robustness, ...
  - Needed:
    - New theory and analytic results for social sensing data management
    - A tool set and a driving demo application to embody the analytic results (e.g., combining data mining, information theory, control, social modeling, ...)

- ### 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 Fatemeh, Hieu Le, Charu Aggarwal, "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.