Data Reliability I:
A Fundamental Challenge in Social Sensing

With Humans as Sensors

CSE 40437/60437-Spring 2015

Prof. Dong Wang

Cyber-Physical System Challenges

- Functional Correctness Guarantees
- Temporal Correctness Guarantees
- Data Correctness Guarantees

Cyber-Physical System Challenges

- Functional Correctness Guarantees
- Temporal Correctness Guarantees
- Data Correctness Guarantees

Outline

- Introduction:
  - Using Humans as Sensors
- Analytical Foundation:
  - Maximum Likelihood Estimation
- Performance:
  - Simulation and Emulation
  - Real world case studies based on Twitter

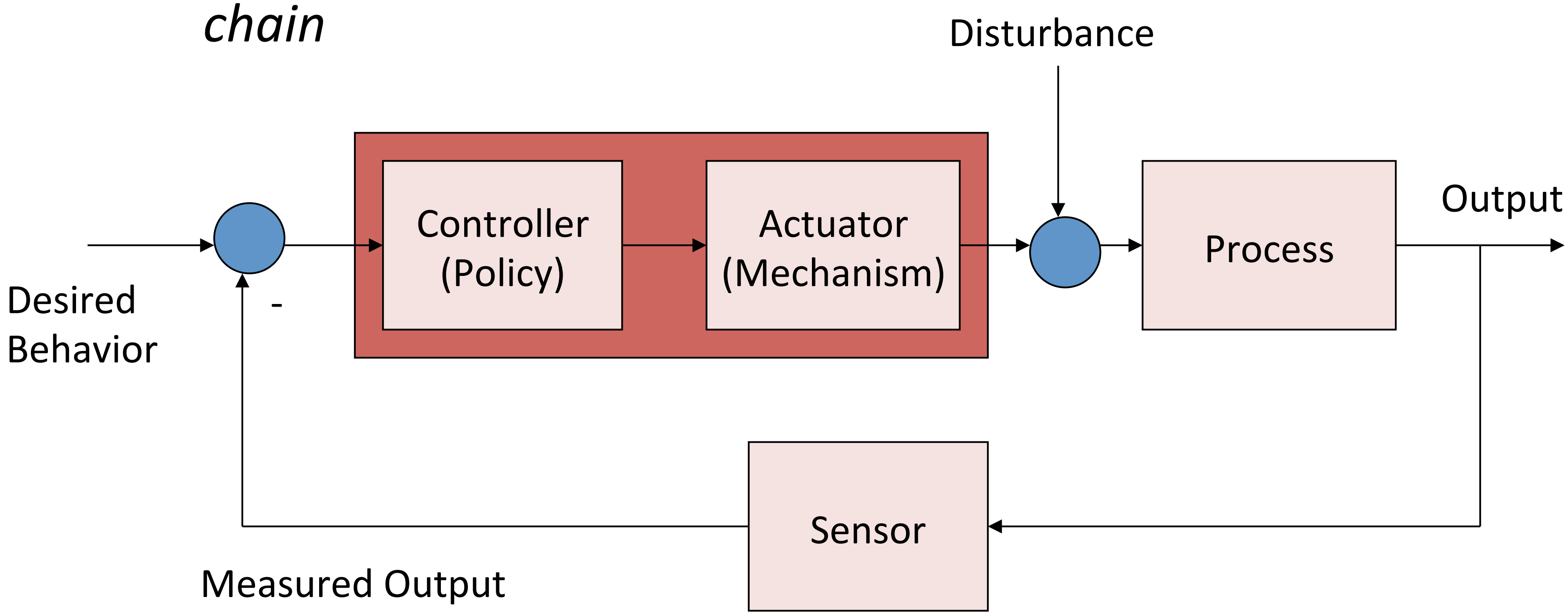
Outline

- **Introduction:**
  - **Using Humans as Sensors**
- Analytical Foundation:
  - Maximum Likelihood Estimation
- Performance:
  - Simulation and Emulation
  - Real world case studies based on Twitter

Role of Humans in the Loop?

- The CPS "Control Loop"

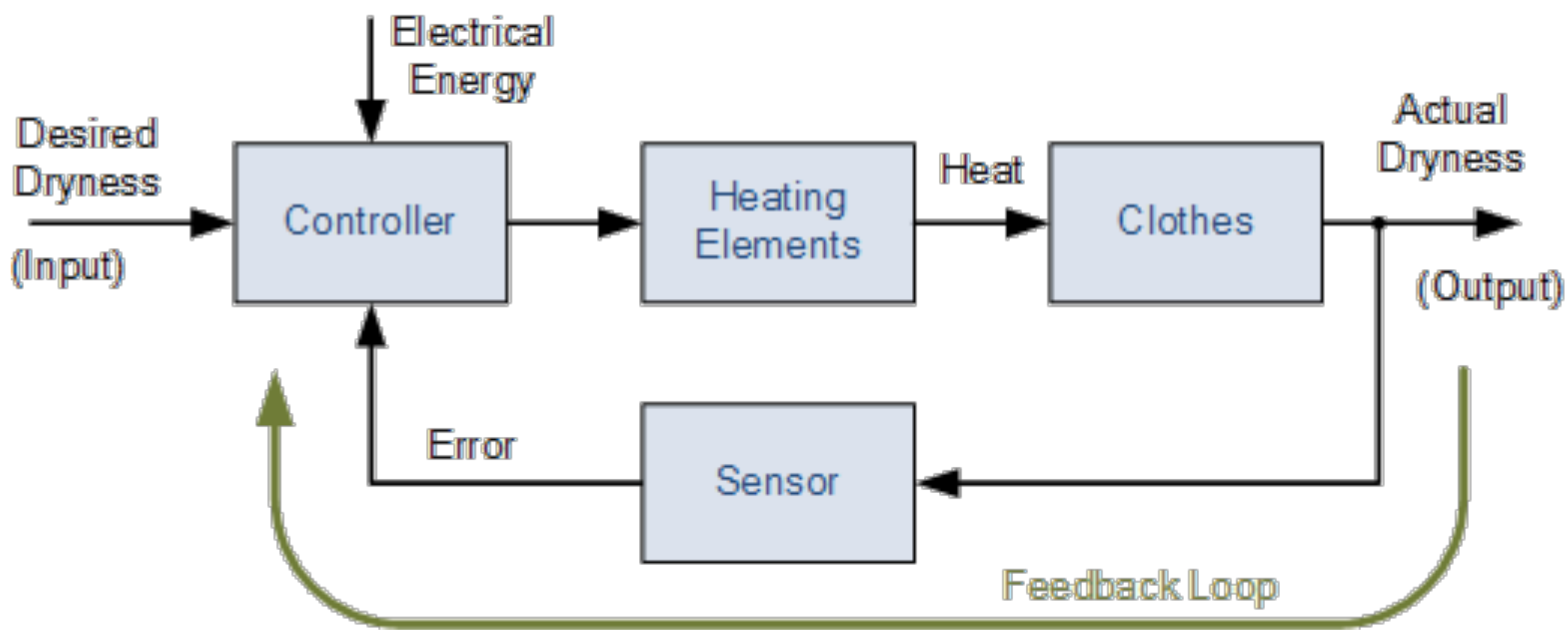
The loop represents a *causality chain*



Role of Humans in the Loop?

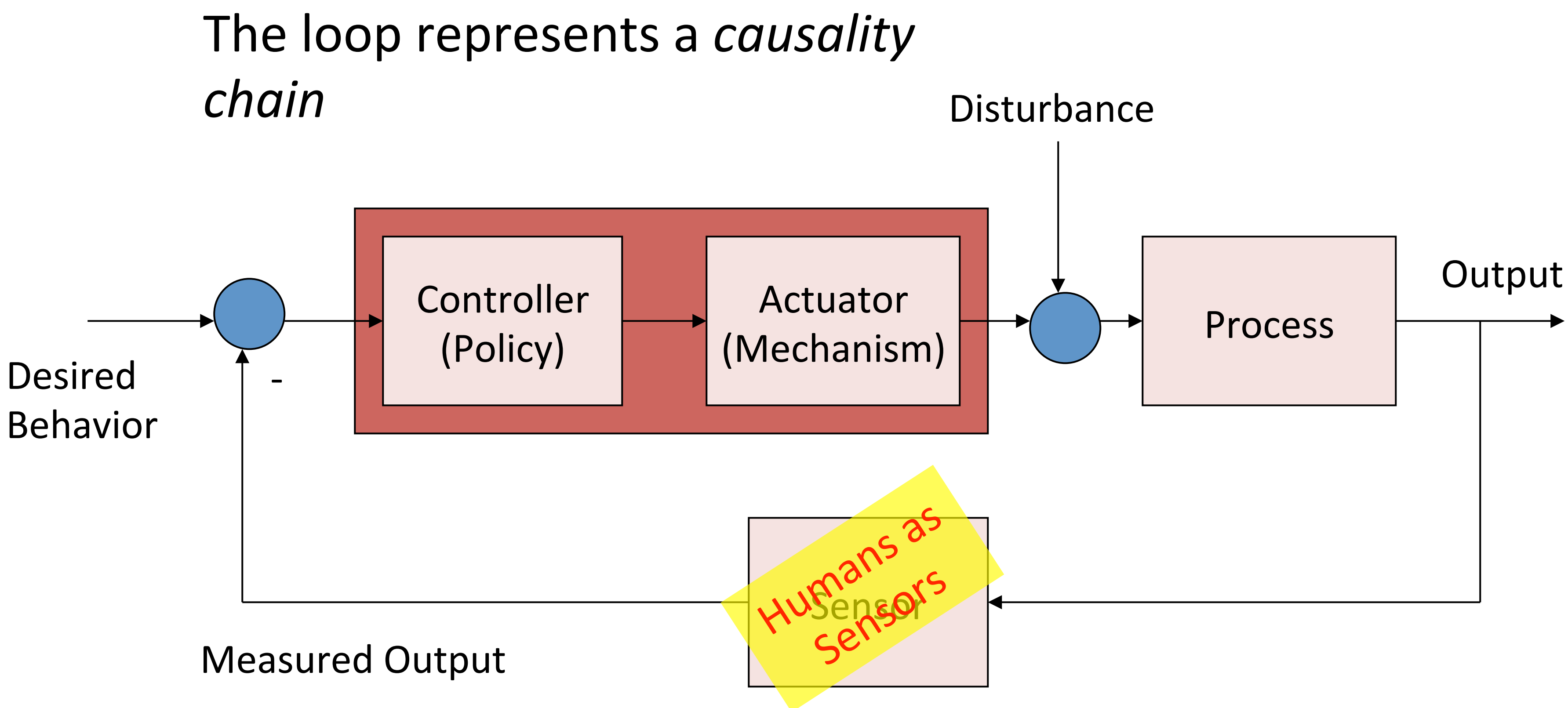
- The CPS "Control Loop"

A Drying Machine Example



Role of Humans in the Loop?

- The CPS "Control Loop"



Humans as Sensors

Social Networks

Sensor Networks

Human

Sensor

Overview

- Following problems are not well addressed/defined in **traditional** sensor network application:

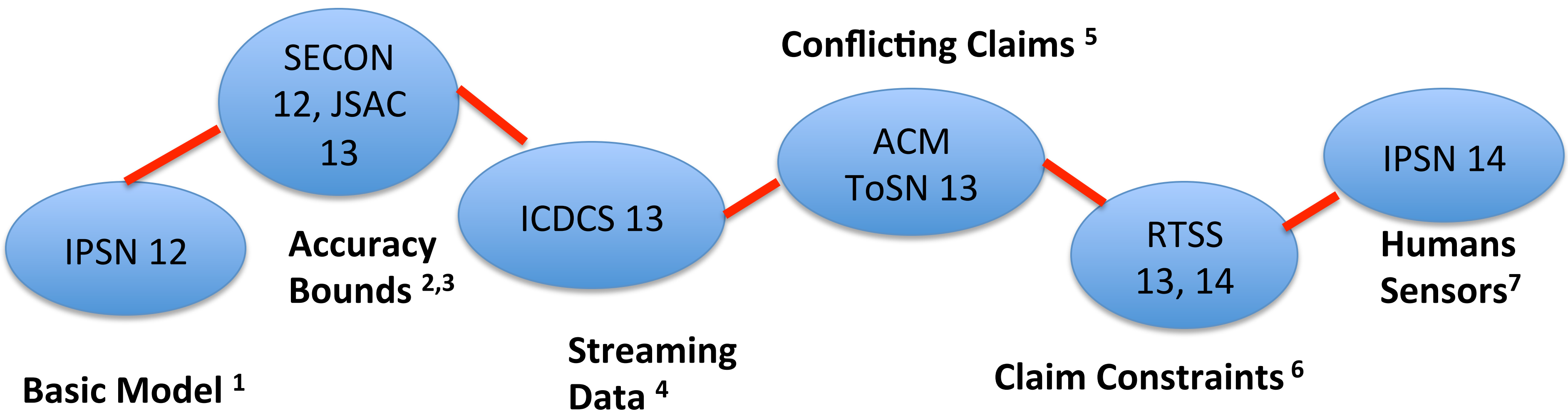
  Q1: What would happen if "sensors" are **not known** to the application a priori?

  Q2: How to model a **person as a "sensor"**

  Q3: How to assess the quality of the results **without independent ways** of verifying the reliability of sources and correctness of their measurements?

- The work discussed in this lecture made efforts to address the above problems emerging in social sensing !

Related Work
on Data Reliability in Social Sensing



1. Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach." The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12). Beijing, China April 2012.

2. Dong Wang , Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Scalability and Robustness Limitations of Real and Asymptotic Confidence Bounds in Social Sensing." The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12), Seoul, Korea, June, 2012.

3. Dong Wang, Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Credibility Tradeoffs in Assured Social Sensing." IEEE JSAC special issue on Network Science, June, Vol. 31, No. 6, 2013.

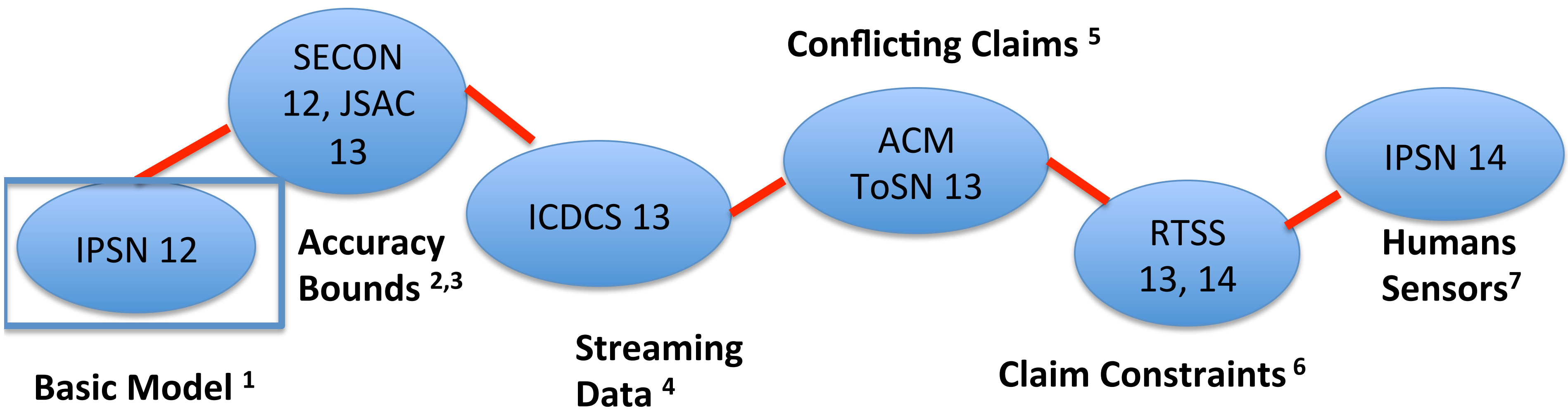
4. Dong Wang, Tarek Abdelzaher, Lance Kaplan and Charu C. Aggarwal. "Recursive Fact-finding: A Streaming Approach to Truth Estimation in Crowdsourcing Applications.", 33rd International Conference on Distributed Computing Systems (ICDCS 13) Philadelphia, PA, July 2013.

5. Dong Wang, Lance Kaplan and Tarek Abdelzaher. "On Truth Discovery in Social Sensing with Conflicting Observations: A Maximum Likelihood Estimation Approach." ACM Transaction on Sensor Network s (TOSN), in press, 2013

6. Dong Wang, Tarek Abdelzaher, Lance Kaplan and Raghu Ganti. "Exploitation of Physical Constraints for Reliable Social Sensing," IEEE 34th Real-Time Systems Symposium (RTSS'13) Vancouver, Canada, December, 2013.

7. Dong Wang , Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu Aggrawal, Raghu Ganti, XinLei Wang, Prasant Mohapatra, Boleslaw Szymanski, Hieu Le, "Humans as Sensors: An Estimation Theoretic Perspective," The 13th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 14), Berlin, Germany, April, 2014.

Related Work
on Data Reliability in Social Sensing



1. Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach." The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12). Beijing, China April 2012.

2. Dong Wang , Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Scalability and Robustness Limitations of Real and Asymptotic Confidence Bounds in Social Sensing." The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12), Seoul, Korea, June, 2012.

3. Dong Wang, Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Credibility Tradeoffs in Assured Social Sensing." IEEE JSAC special issue on Network Science, June, Vol. 31, No. 6, 2013.

4. Dong Wang, Tarek Abdelzaher, Lance Kaplan and Charu C. Aggarwal. "Recursive Fact-finding: A Streaming Approach to Truth Estimation in Crowdsourcing Applications.", 33rd International Conference on Distributed Computing Systems (ICDCS 13) Philadelphia, PA, July 2013.

5. Dong Wang, Lance Kaplan and Tarek Abdelzaher. "On Truth Discovery in Social Sensing with Conflicting Observations: A Maximum Likelihood Estimation Approach." ACM Transaction on Sensor Network s (TOSN), in press, 2013

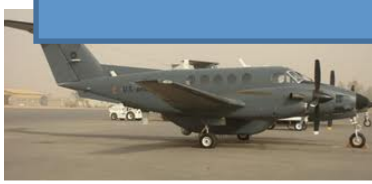
6. Dong Wang, Tarek Abdelzaher, Lance Kaplan and Raghu Ganti. "Exploitation of Physical Constraints for Reliable Social Sensing," IEEE 34th Real-Time Systems Symposium (RTSS'13) Vancouver, Canada, December, 2013.

7. Dong Wang , Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu Aggrawal, Raghu Ganti, XinLei Wang, Prasant Mohapatra, Boleslaw Szymanski, Hieu Le, "Humans as Sensors: An Estimation Theoretic Perspective," The 13th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 14), Berlin, Germany, April, 2014.

Sensing is Evolving

Platform

Sensors are increasingly used by everyday people

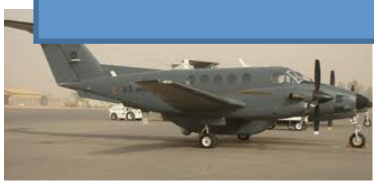


Smart Phone

Sensing is Evolving

Platform

Sensors are increasingly used by everyday people



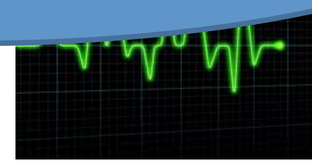
Smart Phone

Social (Human-Centric) Sensing is Emerging!

Application

Environment Monitoring

Target Tracking



Health Monitoring

Humans are getting into the Loop of Sensing.

Smart House

follow us on twitter

Social Sensing

Examples of Social Sensing

Participatory Sensing

BikeNet

Geotagging

Opportunistic Sensing

CabSense

CenceMe

Human's Role in Social Sensing

Human are sensor carriers

Human are sensor operators

Human are sensors themselves!

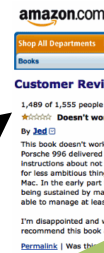
Data Reliability Problem in Social Sensing

Who to believe?

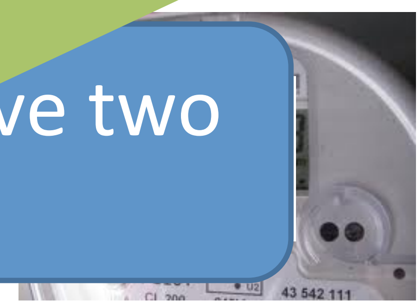
People

Smart Devices

Sources



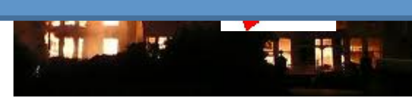
Wh ve?



1. Ho bove two ons?

Guaranteed Data Correctness!

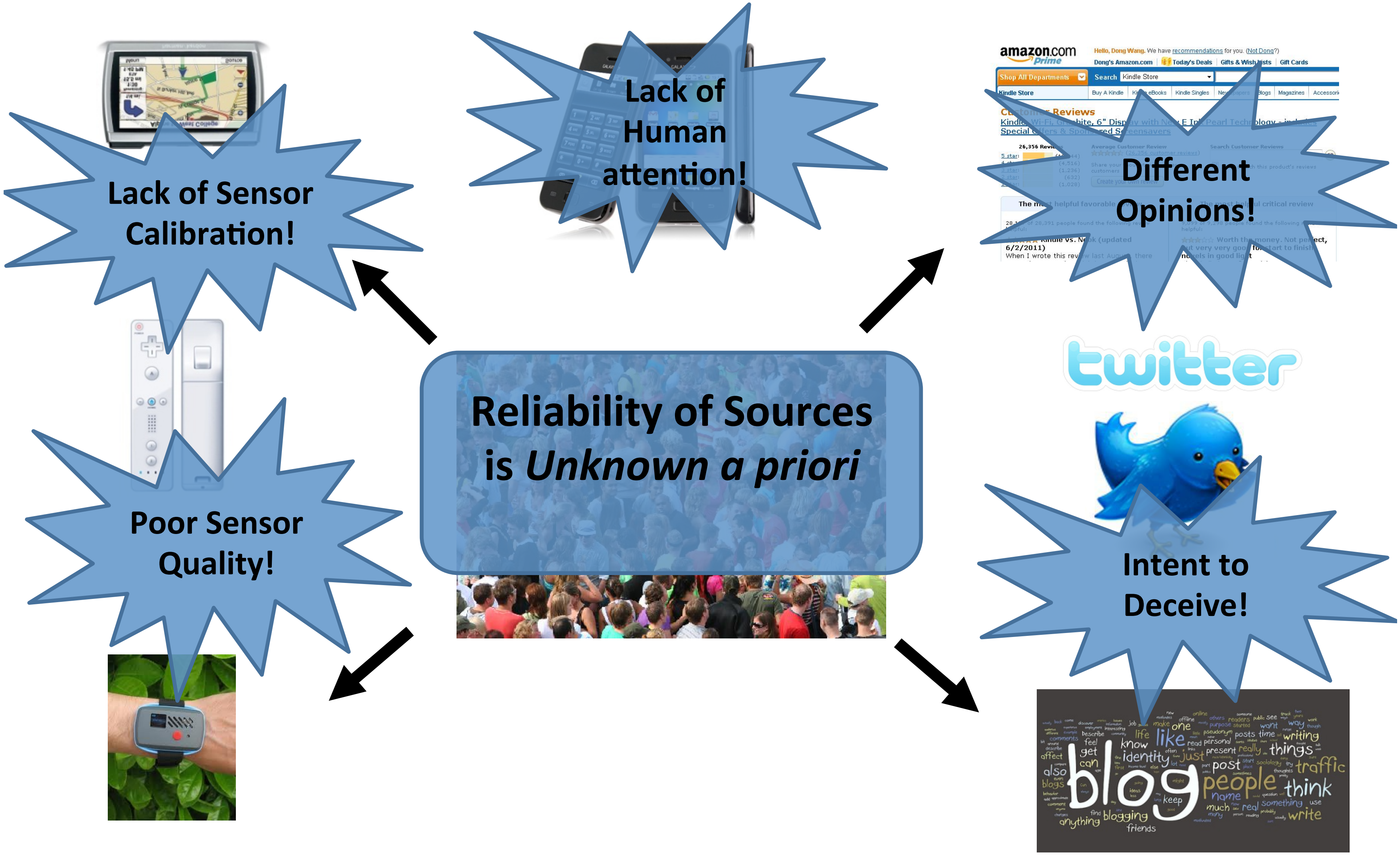
w to ***Assess the Quality*** of our answers ?



Images

Measurements

Challenges in Reliable Social Sensing



Bridging Data Mining and Estimation Theory

Estimation Theory

~~Knowledge Gap~~

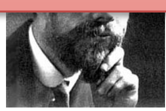
Data Mining

1. What does a source rank # 3 really mean?

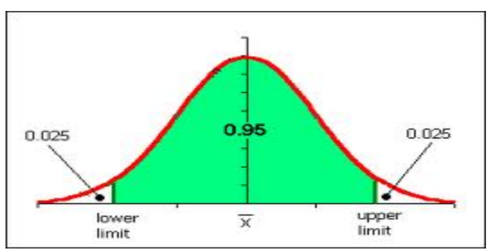
Source Rank	Ranking Score
1	128
2	110
3	20

ssured Information Distillation

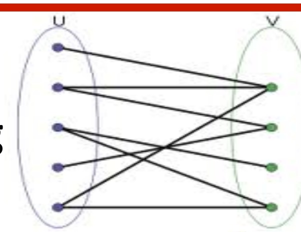
Confidence?



Source Rank	Ranking Confidence
1	?
2	?
3	?

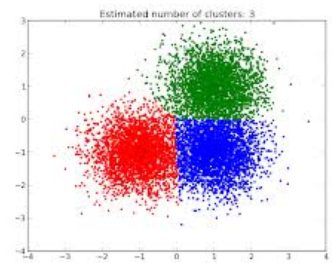


Fact-finding

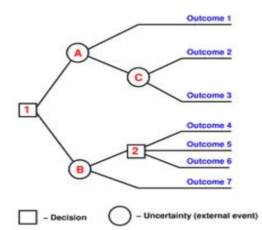


Ranking

Clustering



Classification



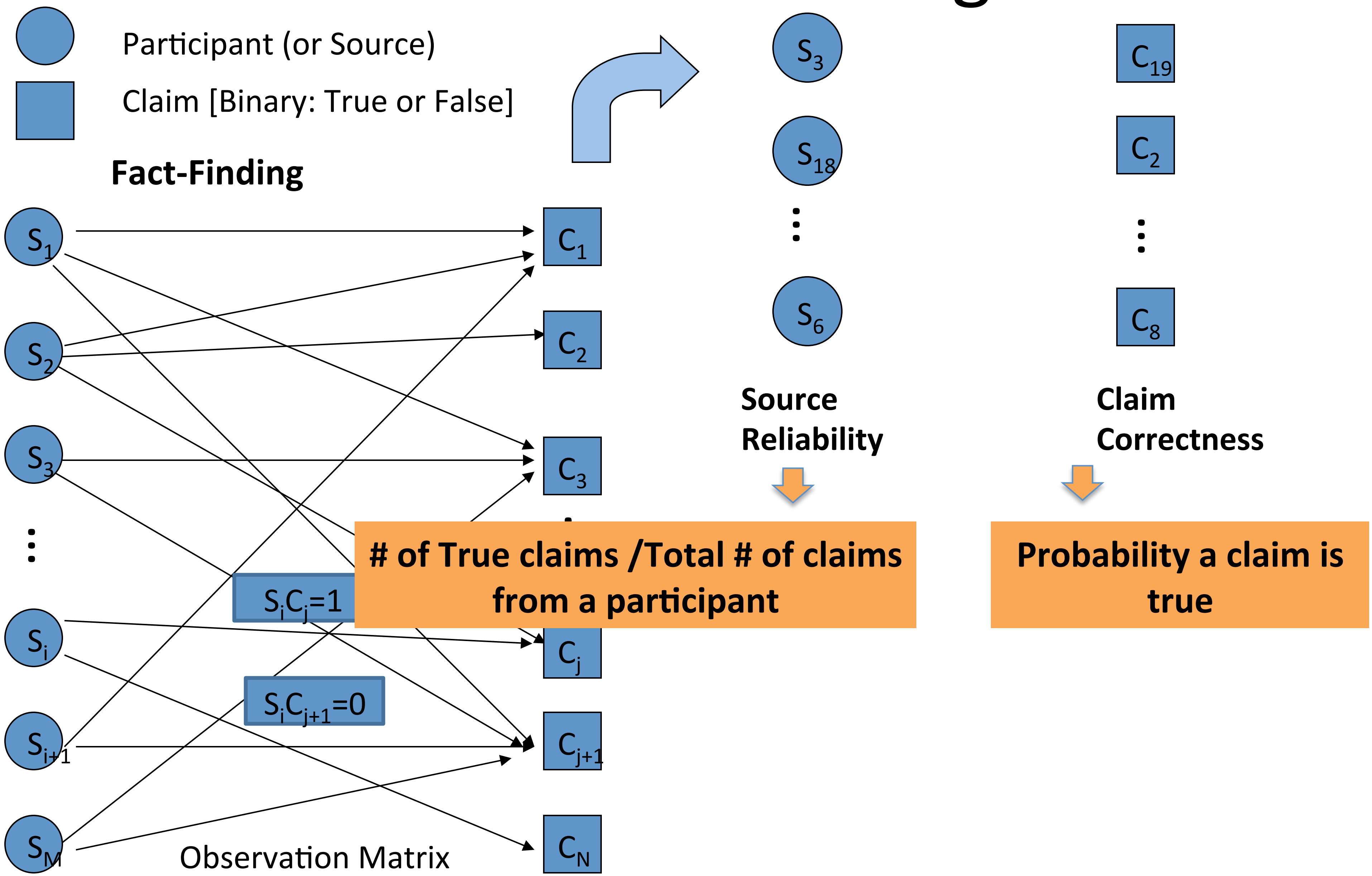
Outline

- Introduction:
  - Using Humans as Sensors
- **Analytical Foundation:**
  - **Maximum Likelihood Estimation**
- Performance :
  - Simulation and Emulation
  - Real world case studies based on Twitter

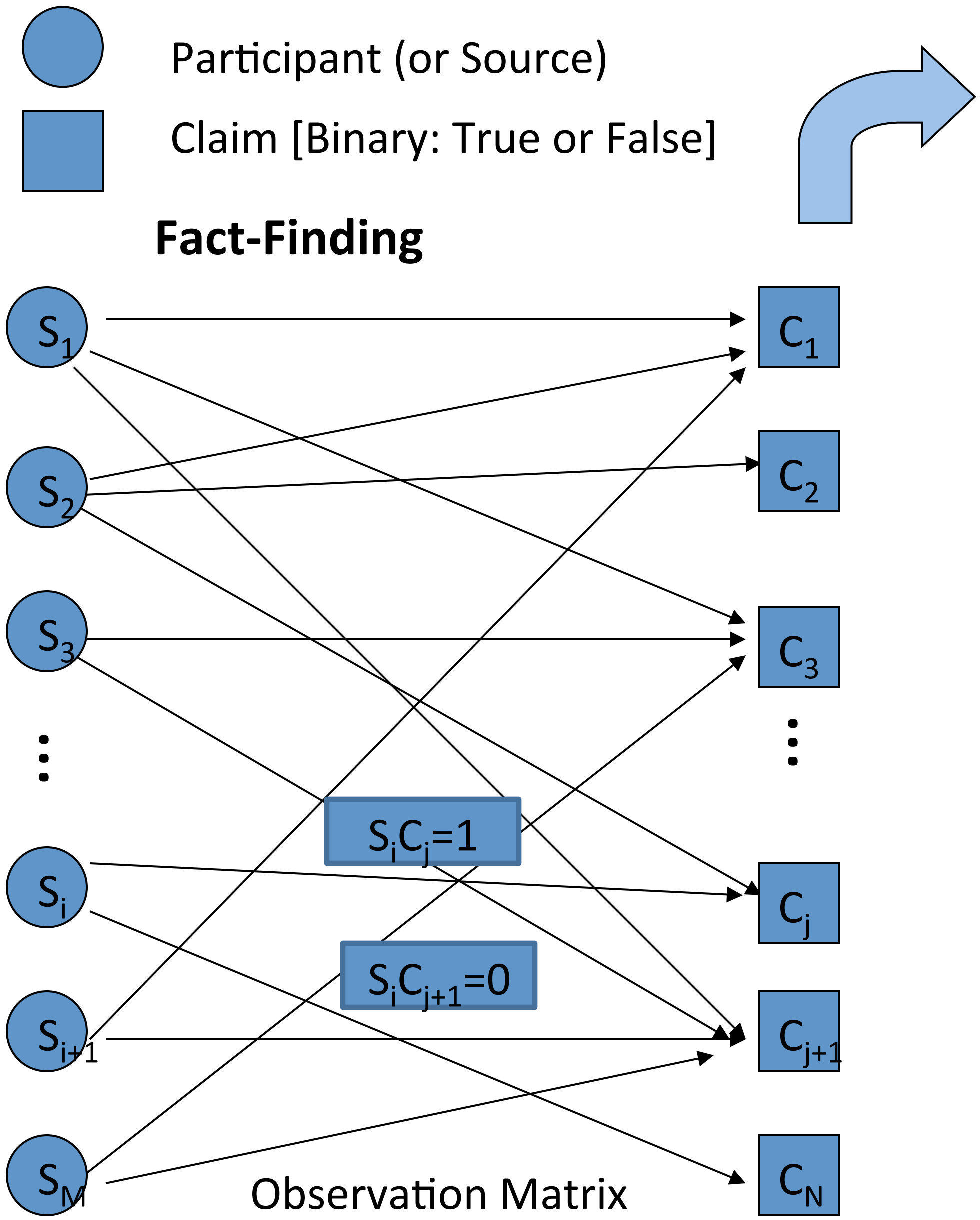
State of the Art

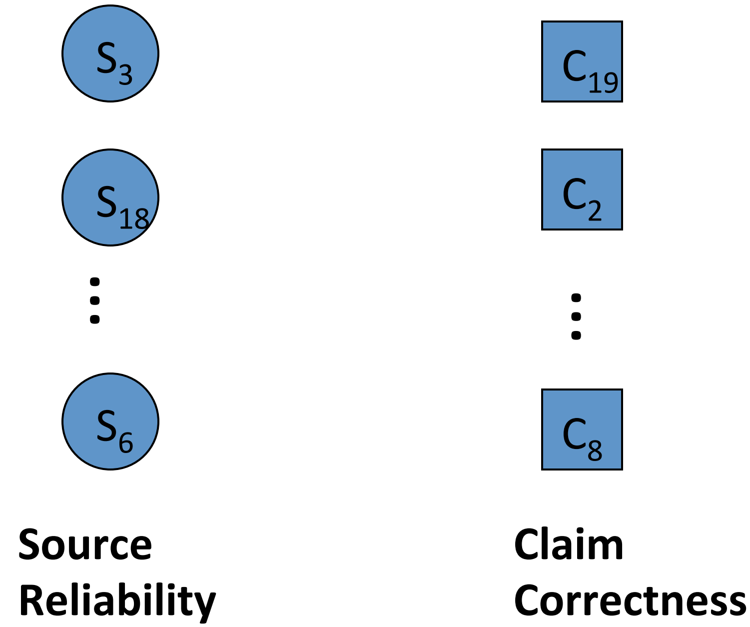
- **Fact-finders**
  - Rank both sources and assertions
  - Heuristic based approach
- **Data Cleaning and Outlier Analysis**
  - Data Mining: Binning, Regression, Clustering, Statistical-based, etc.
  - Statistics: Kalman Filter, Particle Filter, etc.
- **Reputation Systems**
  - Simple counting: eBay
  - Complex heuristic: similar as PageRank

Basic Fact-finding



Basic Fact-finding





Voting

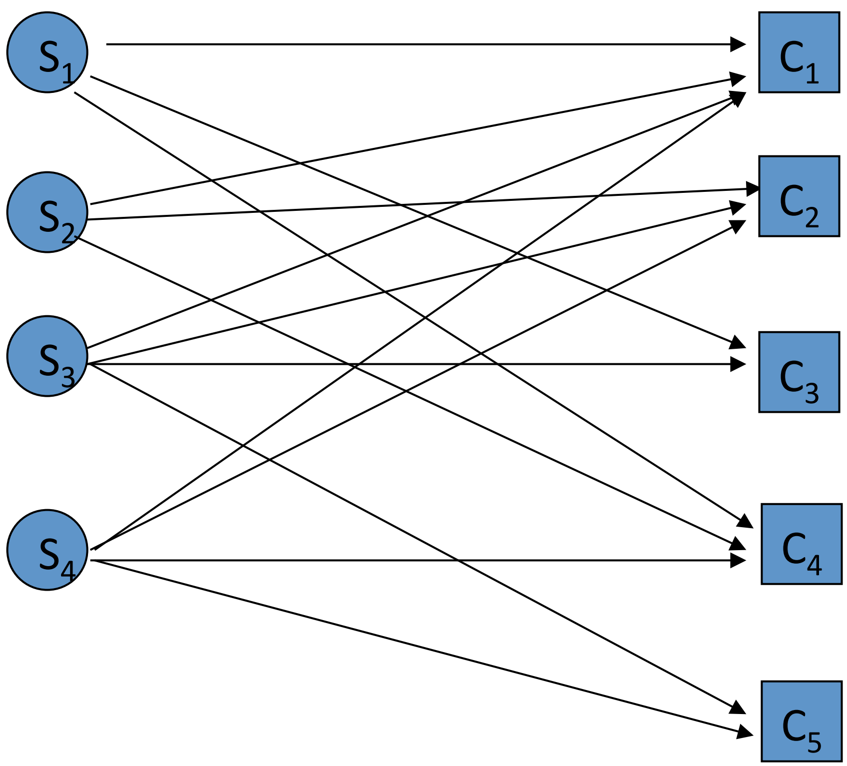
What is the problem with Voting?

Basic Fact-finding

Participant (or Source)

Claim [Binary: True or False]

Fact-Finding



Voting

Claim ID	Votes
C1	4
C2	3
C4	3
C3	2
C5	2

True

False

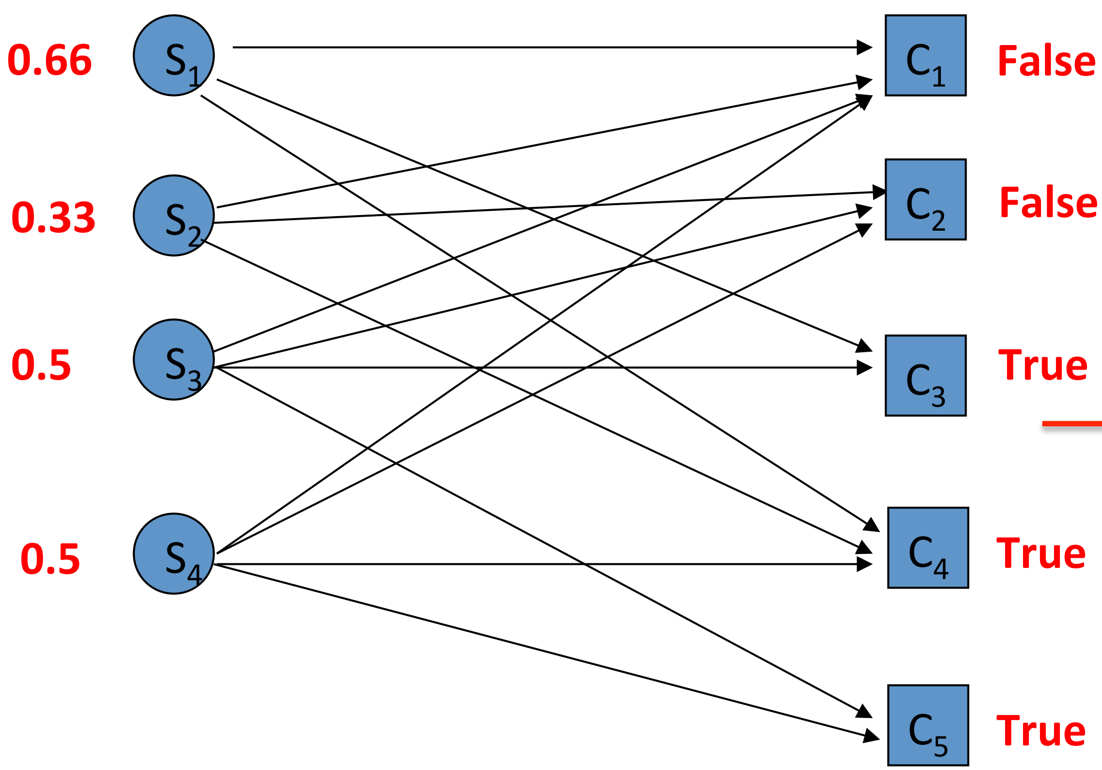
Prior: 3 true claims and 2 false claims

Basic Fact-finding

Participant (or Source)

Claim [Binary: True or False]

Fact-Finding



Prior: 3 true claims and 2 false claims

Voting

Claim ID	Votes
C1	4
C2	3
C4	3
C3	2
C5	2

Not all sources are equally reliable!

Examples: Some Fact-Finding Algorithms

1. Sum (Hubs and Authorities)

$$S^{i}_{cred}(s) = \sum_{c \in C_s} C^{i-1}_{cred}(c)$$

$$C^{i}_{cred}(c) = \sum_{s \in S_c} S^{i}_{cred}(s)$$

where C_s is the set of claims made by source s

and S_c is the set of sources that make claim c

2. Average-Log

$$S^{i}_{cred}(s) = \log |C_s| \frac{\sum_{c \in C_s} C^{i-1}_{cred}(c)}{|C_s|}$$

avoid source from obtaining high credibility by simply making many trival claims

3. Investment

$$S^{i}_{cred}(s) = \sum_{c \in C_s} C^{i-1}_{cred}(c) \frac{S^{i-1}_{cred}(s)}{|C_S| \sum_{r \in S_C} \frac{S^{i-1}_{cred}(r)}{|C_r|}}$$

incoporate the trust the source previously invested into each claim

Difference:

The specific way to compute source or assertion credibility in iterations.

Similarity:

Only output ranking, not posterior probability desired .

Basics of Maximum Likelihood Estimation

A Simple Example:

- A random number generator G(T):
  - It can generate a random integer in [1,T] with a *uniform probability* distribution
- Question:
  - If T only has two possible values: **10** and **20**, we run G(T) once, the generate number is **5**. **What is the most likely value of T?**

Basics of Maximum Likelihood Estimation

A Simple Example:

- A random number generator G(T):
  - It can generate a random integer in [1,T] with a *uniform probability* distribution
- Question:
  - If T can be **any integer value**, we run G(T) once, the generate number is still **5**. **What is the most likely value of T?**

MLE: Make the guess of the estimated parameters for which the observed data is least surprising!

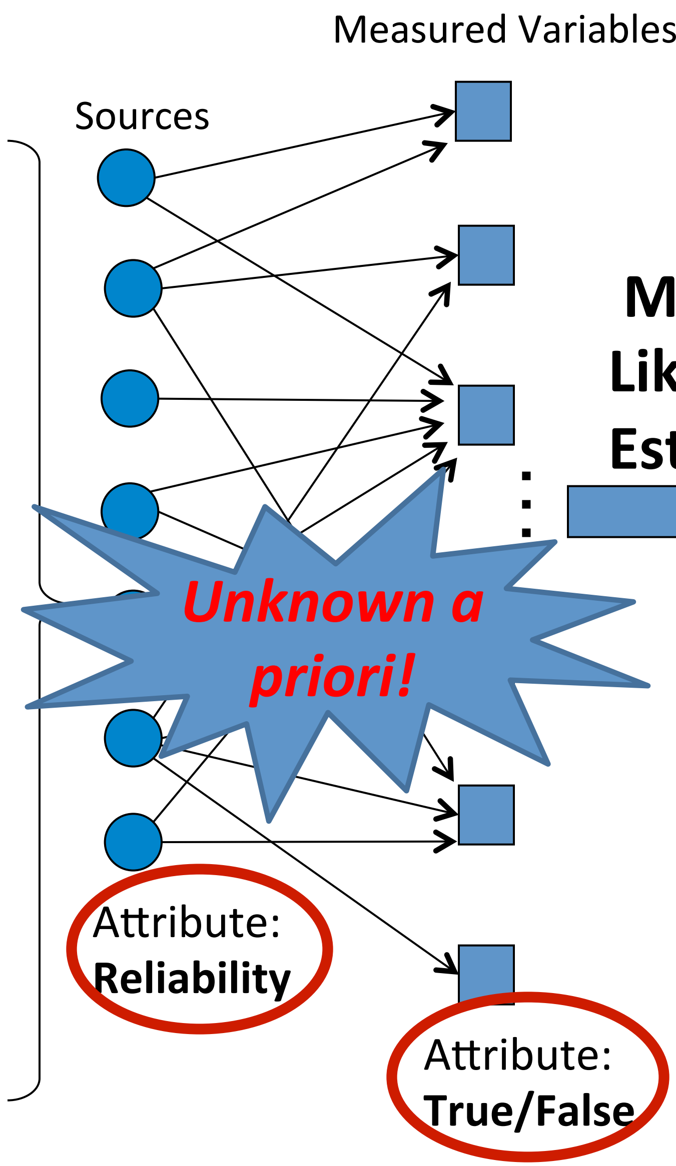
Maximum Likelihood Estimation

Events

Hurricane Sandy

Boston Marathon Explosion

Egypt President Arrest



Maximum Likelihood Estimation

# of True variables /Total # of variables a source reports

-Reliability of sources
-Correctness of variables

Probability a measured variable is true

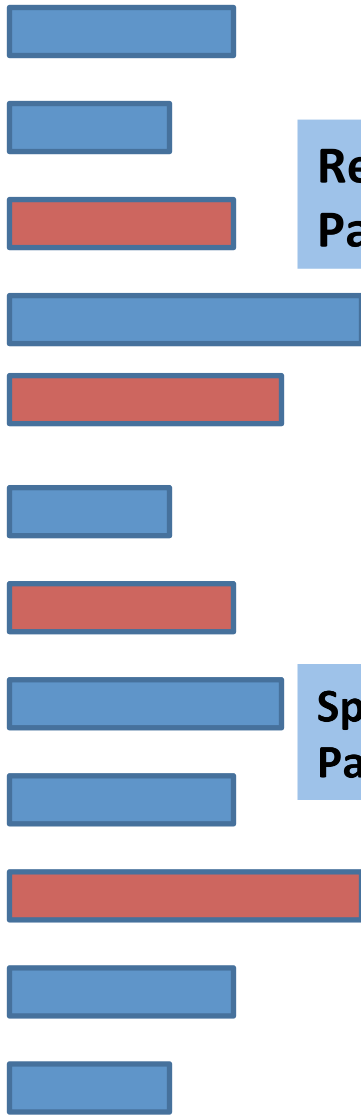
D. Wang, et al., IPSN, 2012

Maximum Likelihood Estimation

Basic Definition

True Measured Variable

False Measured Variable



Reliability of Participant *i* = $\dfrac{i}{i + i}$

Participant Reliability

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

$S_i C_j$: participant i reports measured variable j

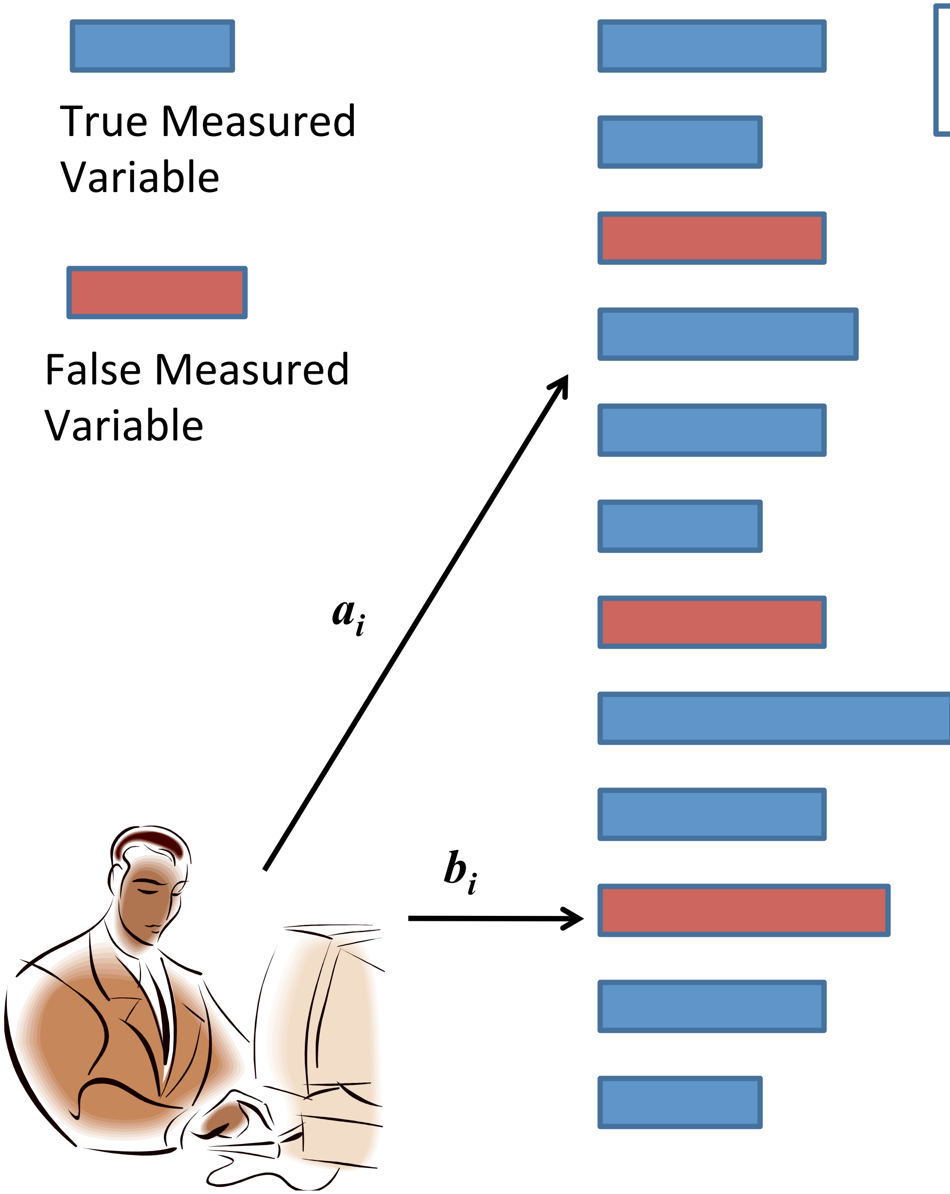
Speak Rate of Participant *i* $\propto \dfrac{i + i}{All + All}$

Participant i speak with rate s_i

$s_i = P(S_i C_j)$

Maximum Likelihood Estimation

Basic Definition



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

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

where d is the overal prior that a randomly chozen measured variable is true

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

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

where d is the overal prior that a randomly chozen measured variable is true

Expectation Maximization

Background and Problem Formulation

Expectation Maximization

$$L(\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta)$$

Estimation parameter

Observed data

Hidden Variable

Expectation Step (E-step)

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$$

Maximization Step (M-step)

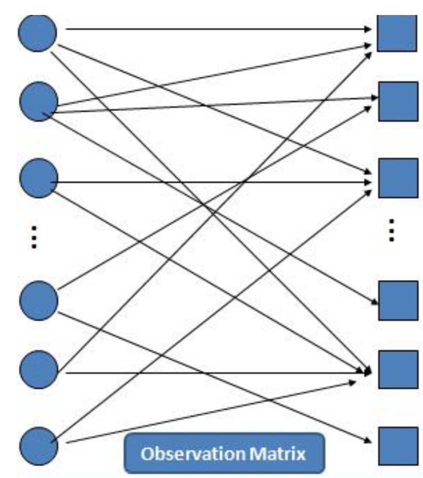
$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(t)}\right)$$

Apply EM

$Z=\{z_1, z_2, \ldots z_N\}$ where z_j =1 when assertion C_j is true and 0 otherwise

X

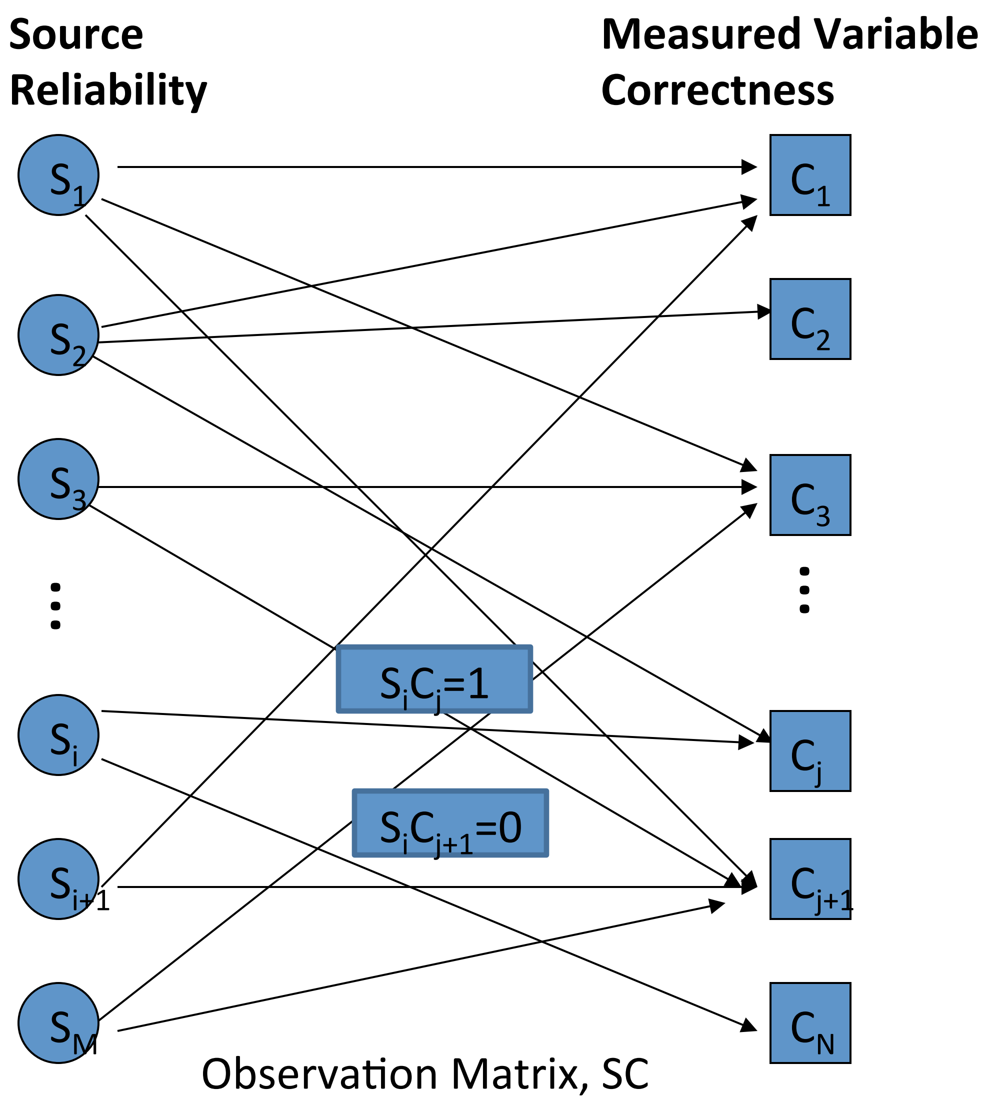
Observation Matrix



$$\theta = (a_1, a_2, \ldots a_M; b_1, b_2, \ldots b_M)$$

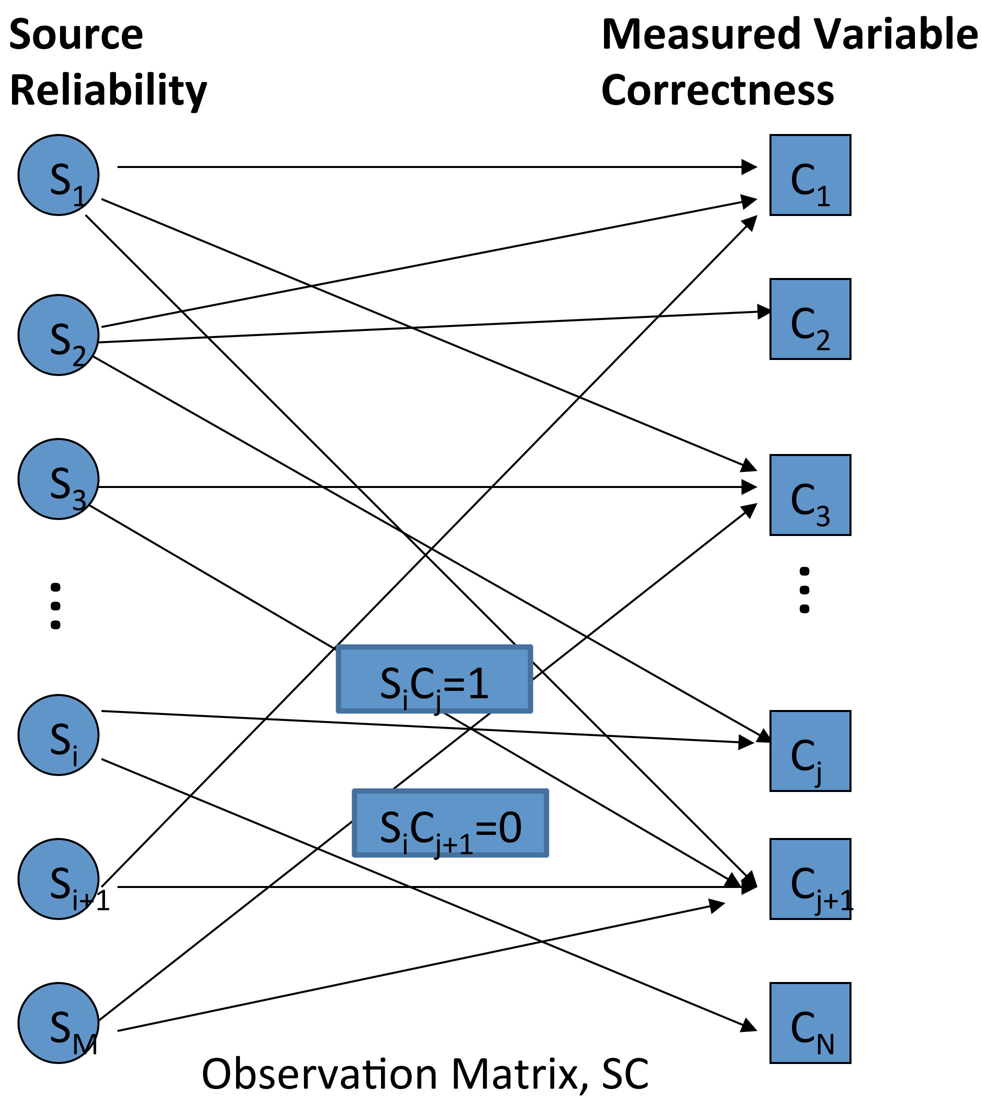
Find MLE of estimation parameter and values of hidden variables

Maximum Likelihood Estimation



Find the "unknown" values of variables, θ, that maximize the probability of observations

Maximum Likelihood Estimation

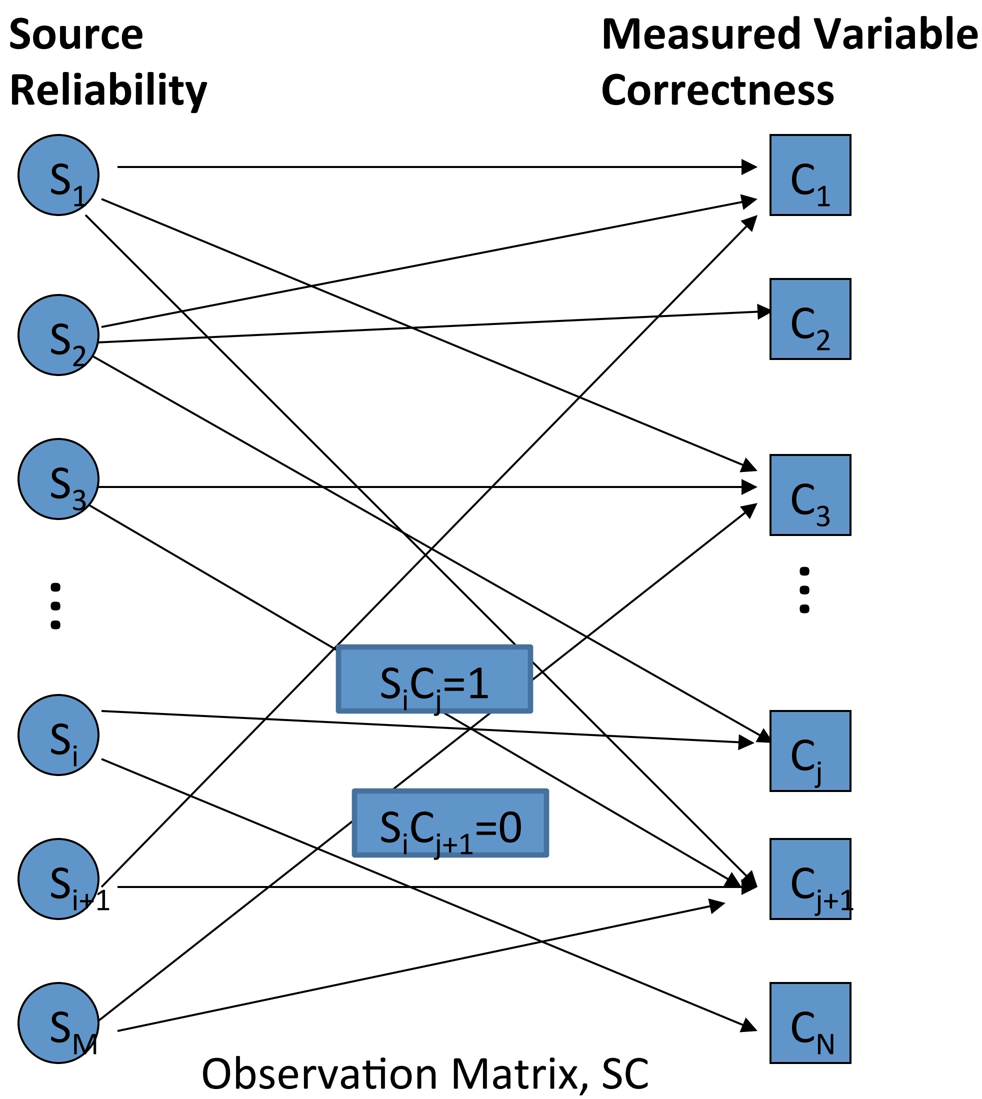


Find the "unknown" values of variables, θ, that maximize the probability of observations

Maximize:

$$P(SC|\theta)$$

Maximum Likelihood Estimation



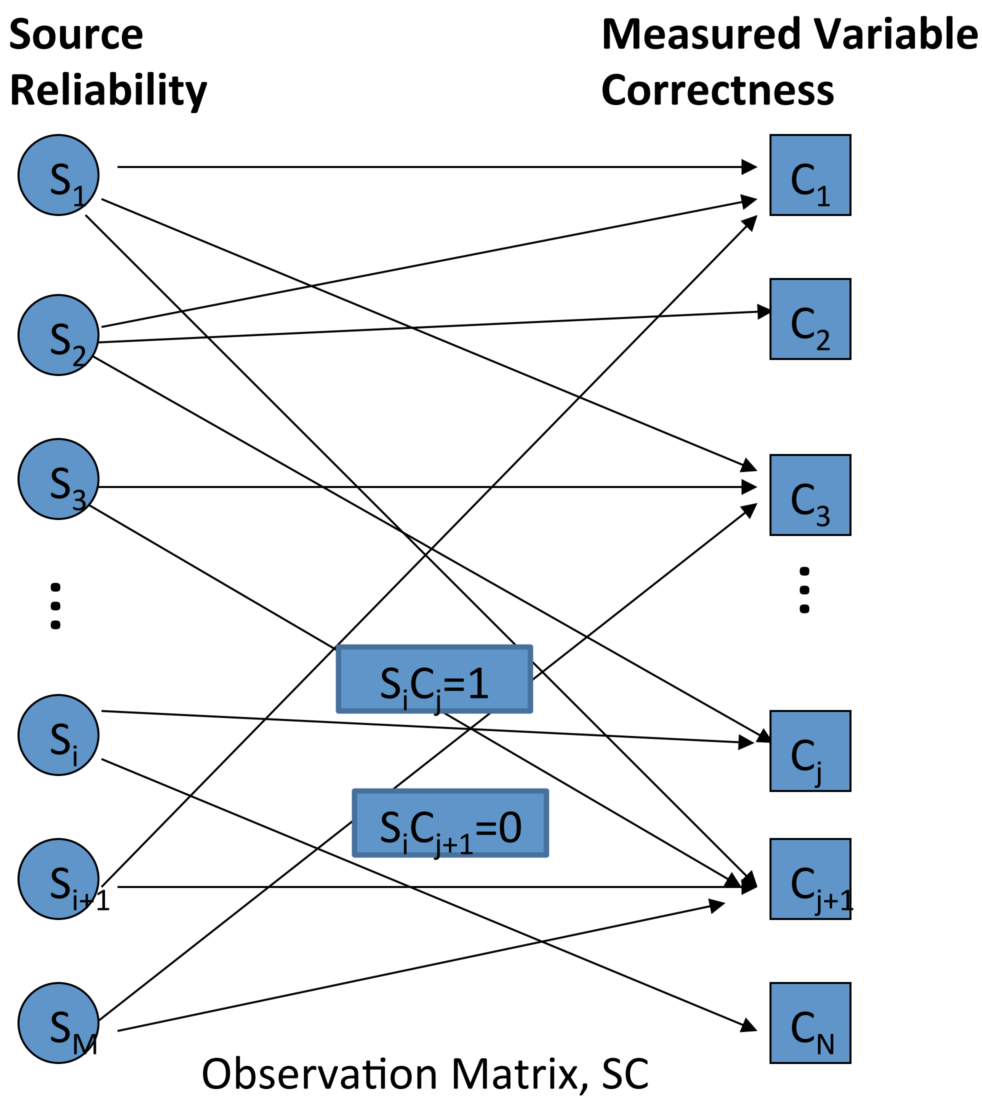
Find the "unknown" values of variables, θ, that maximize the probability of observations

Maximize:

$$P(SC|\theta)$$

Continuous unknowns that depend on discrete unknowns, z?

Maximum Likelihood Estimation



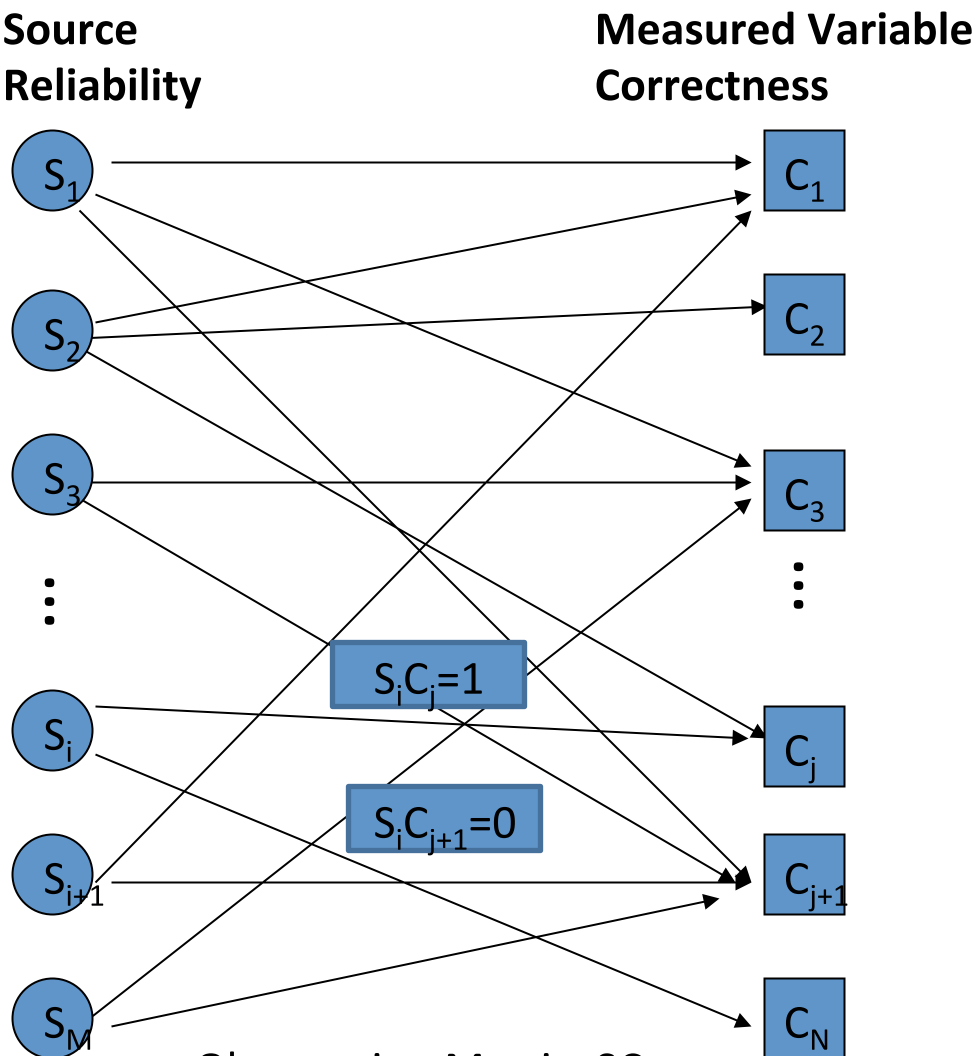
Find the "unknown" values of variables, θ, that maximize the probability of observations

Maximize:

$$P(SC|\theta) = \sum_z P(SC, z|\theta)$$

Continuous unknowns that depend on discrete unknowns, z?

Maximum Likelihood Estimation



Observation Matrix, SC

Find the "unknown" values of variables, θ, that maximize the probability of observations

Source reliability

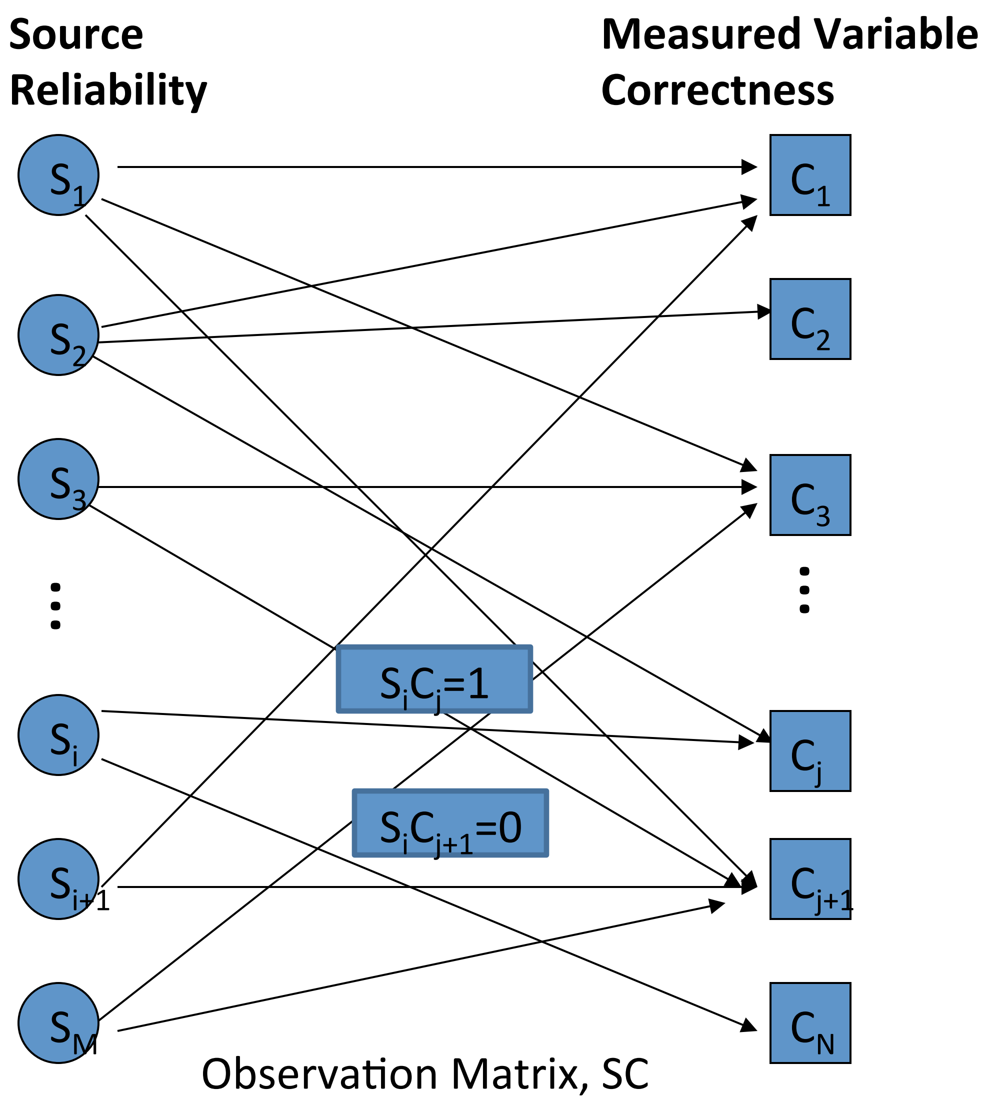
Variable correctness

Maximize:

$$P(SC|\theta) = \sum_z P(SC, z|\theta)$$

Continuous unknowns that depend on discrete unknowns, z?

Maximum Likelihood Estimation



Source reliability

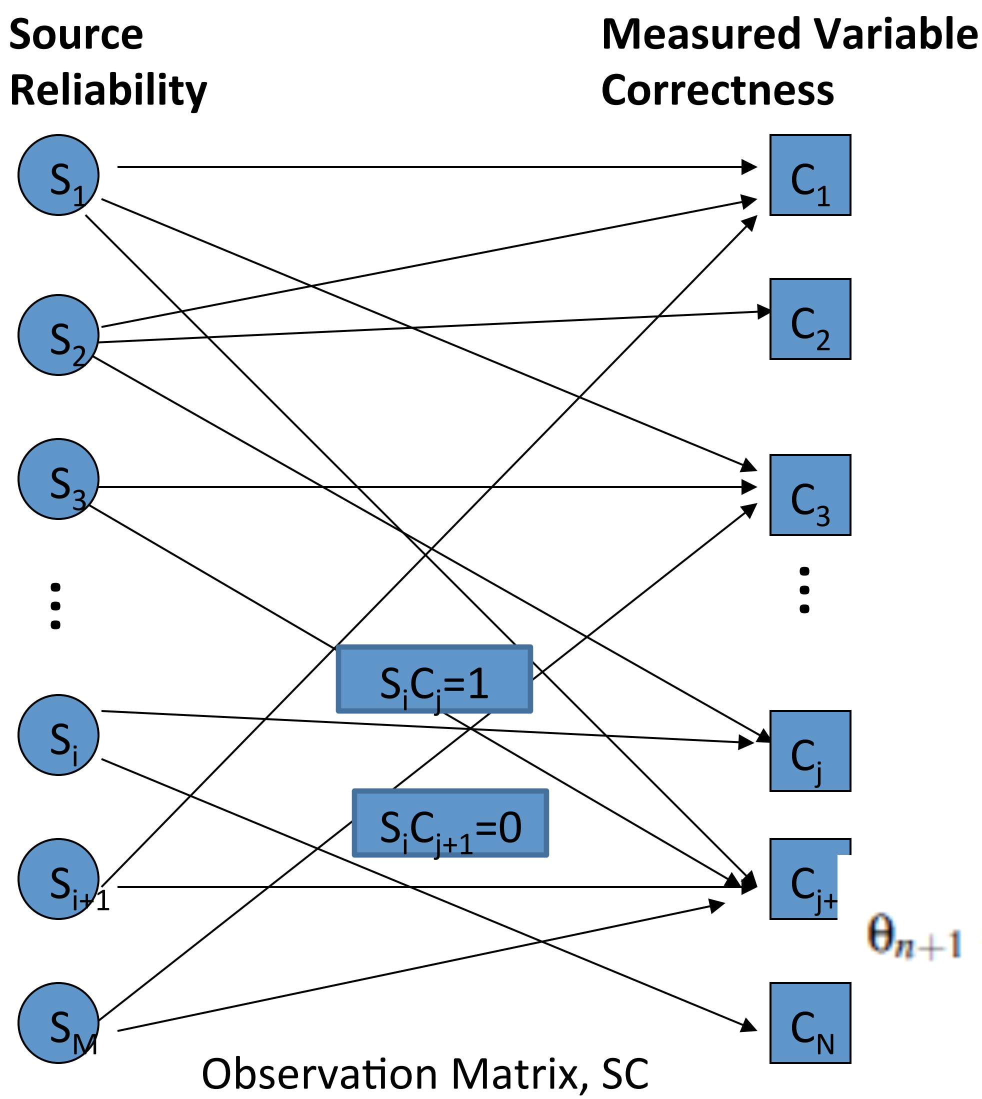
Variable correctness

Maximize:

$$P(SC|\theta) = \sum_z P(SC, z|\theta)$$

- Take log of both sides
- Observe that log of sum is higher than sum of log
- Maximize sum of log

Maximum Likelihood Estimation



Source reliability
Variable correctness

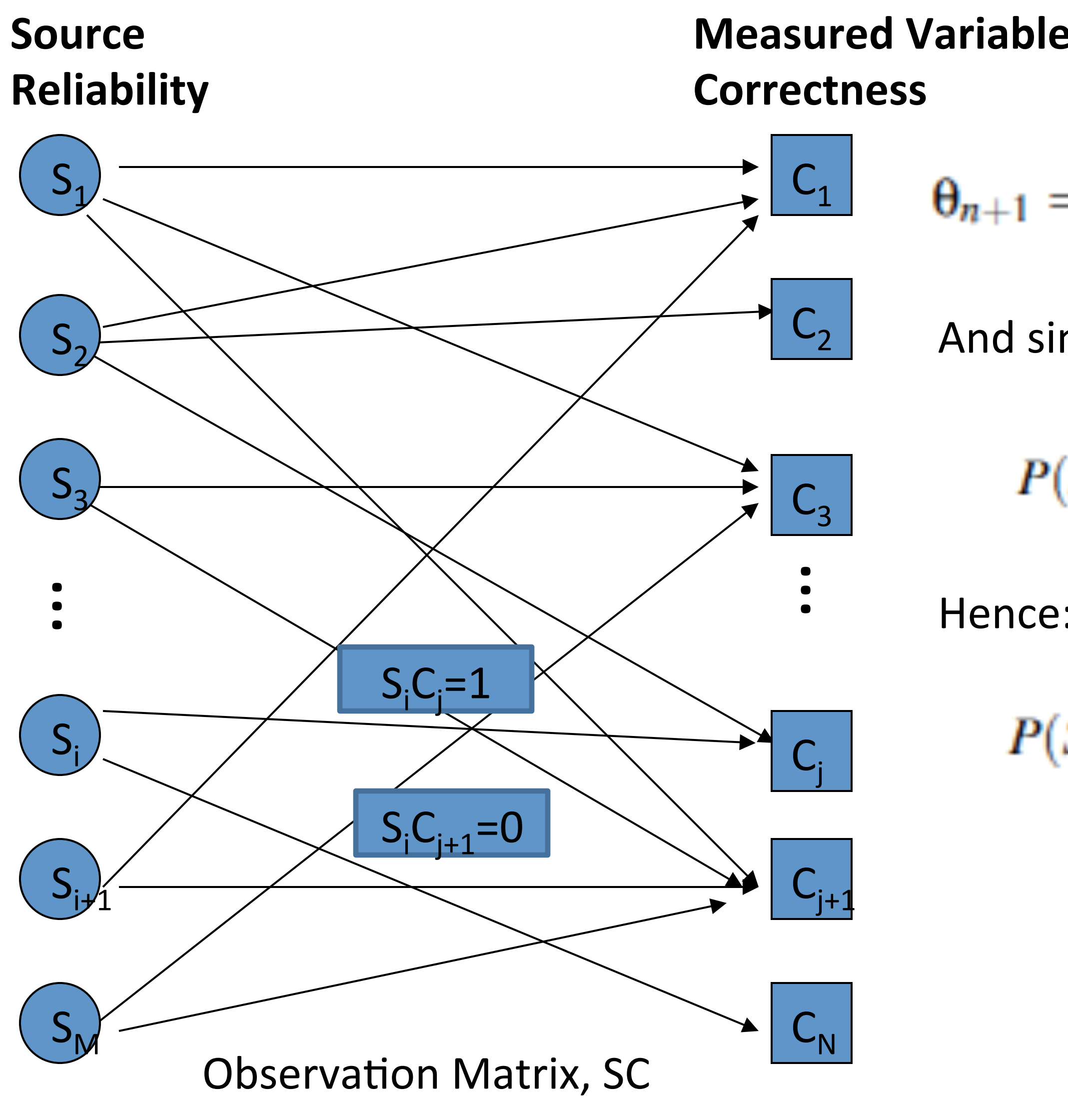
Maximize:

$$P(SC|\theta) = \sum_{z} P(SC, z|\theta)$$

- Take log of both sides
- Observe that log of sum is higher than sum of log
- Maximize sum of log

$$\theta_{n+1} = argmax_{\theta}\{E_{z|SC,\theta_n}\{lnP(SC, z|\theta)\}\}$$

Maximum Likelihood Estimation



$$\theta_{n+1} = argmax_{\theta}\{E_{z|SC,\theta_n}\{lnP(SC,z|\theta)\}\}$$

And since variables are independent:

$$P(SC,z|\theta) = \prod_{j=1}^{N} P(SC_j, z_j|\theta)$$

Hence:

$$P(SC,z|\theta) = \prod_{j=1}^{N} P(SC_j|\theta, z_j)P(z_j)$$

Expectation Maximization

Solution

Likelihood function of EM

$$L(\theta; X, Z) = p(X, Z|\theta)$$

$$= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{S_i C_j} (1 - a_i)^{(1 - S_i C_j)} \times d \times z_j + \prod_{i=1}^{M} b_i^{S_i C_j} (1 - b_i)^{(1 - S_i C_j)} \times (1 - d) \times (1 - z_j) \right\}$$

Expectation Step (E-Step)

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$$

$$Z(t, j) = f(a^{(t)}, b^{(t)}, d^{(t)} j)$$

$$= \sum_{j=1}^{N} \left\{ p(z_j = 1|X_j, \theta^{(t)}) \times \left[\sum_{i=1}^{M} (S_i C_j \log a_i + (1 - S_i C_j) \log(1 - a_i) + \log d) \right] \right.$$

$$\left. + p(z_j = 0|X_j, \theta^{(t)}) \times \left[\sum_{i=1}^{M} (S_i C_j \log b_i + (1 - S_i C_j) \log(1 - b_i) + \log(1 - d)) \right] \right\}$$

Maximization Step (M-Step)

$$a_i^{(t+1)} = a_i^* = \frac{\sum_{j \in SJ_i} Z(t, j)}{\sum_{j=1}^{N} Z(t, j)}$$

$$b_i^{(t+1)} = b_i^* = \frac{K_i - \sum_{j \in SJ_i} Z(t, j)}{N - \sum_{j=1}^{N} Z(t, j)}$$

$$d_i^{(t+1)} = d_i^* = \frac{\sum_{j=1}^{N} Z(t, j)}{N}$$

Iterate

Expectation Maximization

Solution-Optimal Iteration

EM's Iteration

E-Step

$$\boxed{Z(t,j) =}$$

$$\frac{\prod_{i=1}^{M} a_i^{(t)S_iC_j}(1-a_i^{(t)})^{(1-S_iC_j)} \times d}{\prod_{i=1}^{M} a_i^{(t)S_iC_j}(1-a_i^{(t)})^{(1-S_iC_j)} \times d + \prod_{i=1}^{M} b_i^{(t)S_iC_j}(1-b_i^{(t)})^{(1-S_iC_j)} \times (1-d)}$$

M-Step **Optimal Non-linear**

$$\boxed{a_i^{(t+1)}} = \frac{\sum_{j \in SJ_i} Z(t,j)}{\sum_{j=1}^{N} Z(t,j)} \qquad \boxed{b_i^{(t+1)}} = \frac{K_i - \sum_{j \in SJ_i} Z(t,j)}{N - \sum_{j=1}^{N} Z(t,j)}$$

Fact-finder's Iteration

$$\boxed{\bar{C}^{est}_{cred}} = [CS]\bar{S}_{cred}$$

Approximate Linear

Iterate

$$\boxed{\bar{S}_{cred}} = [CS]^T \bar{C}_{cred}$$

Outline

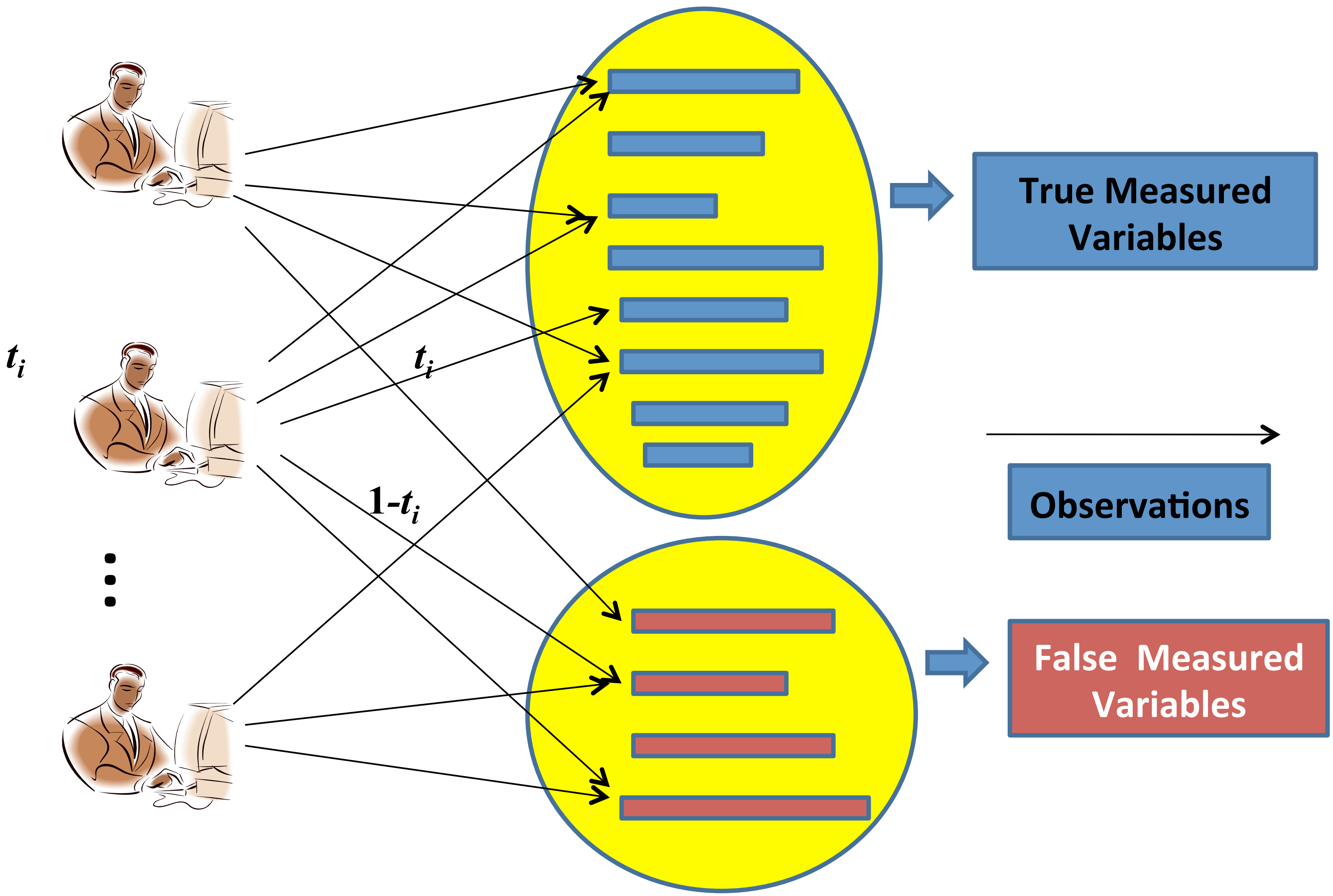
- Introduction:
  - Using Humans as Sensors
- Analytical Foundation:
  - Maximum Likelihood Estimation
- **Performance :**
  - **Simulation and Emulation**
  - **Real world case studies based on Twitter**

Expectation Maximization Evaluation

- Simulations:
  - Source Reliability Estimation Error
  - Measured Variable Classification:
    - False Positives: False variables misclassified as true/ False variables
    - False Negatives: True variables misclassified as false/ True variables
- Evaluate EM through emulated and real social sensing applications

Evaluation

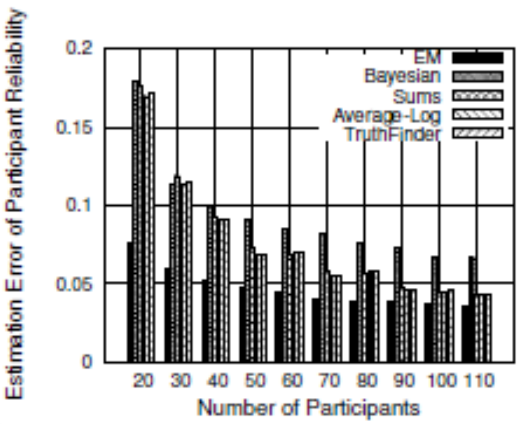
Simulation: Sensing Topology Generation



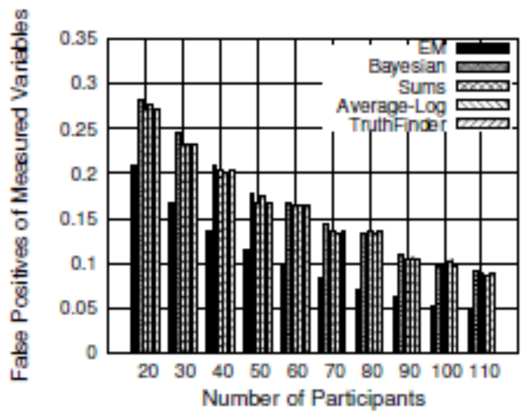
Expectation Maximization

Simulation

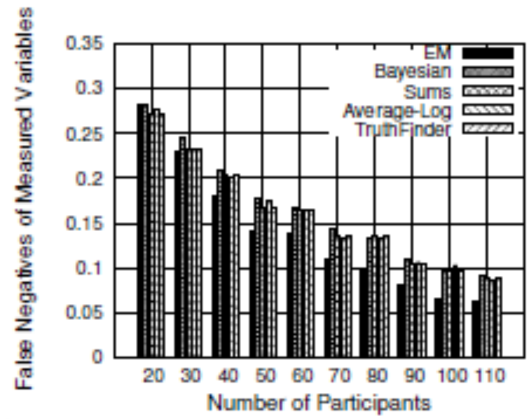
Estimation Accuracy of EM vs Baselines with Varying # of Participants



(a) Participant Reliability Estimation Accuracy



(b) Measured Variable Estimation: False Positives



(c) Measured Variable Estimation: False Negatives

EM outperforms state-of-art heuristics

Parameters:

Number of Participants: 20-110

Number of True Measured Variables: 1000

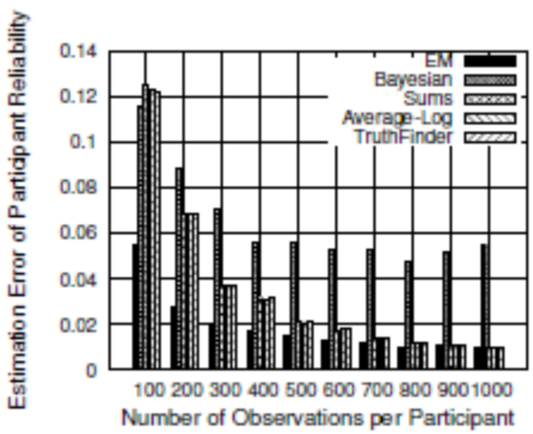
Number of False Measured Variables: 1000

Average Number of Observations per Participant: 100

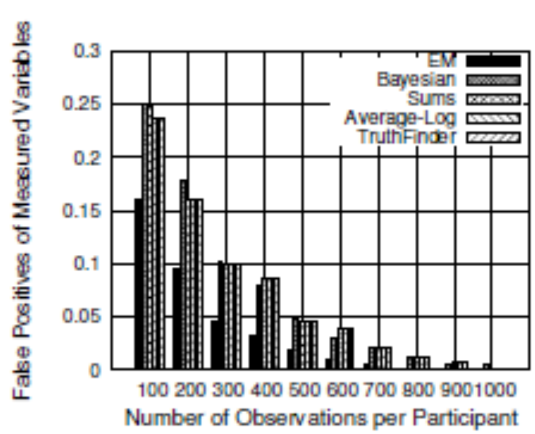
Expectation Maximization

Simulation

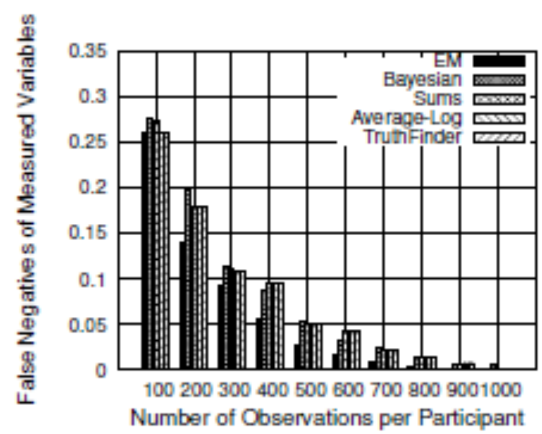
Estimation Accuracy of EM vs Baselines with <u>Varying # of Observations per Participants</u>



(a) Participant Reliability Estimation Accuracy



(b) Measured Variable Estimation: False Positives



(c) Measured Variable Estimation: False Negatives

EM outperforms state-of-art heuristics

Parameters:

Number of Participants: 30

Number of True Measured Variables: 1000

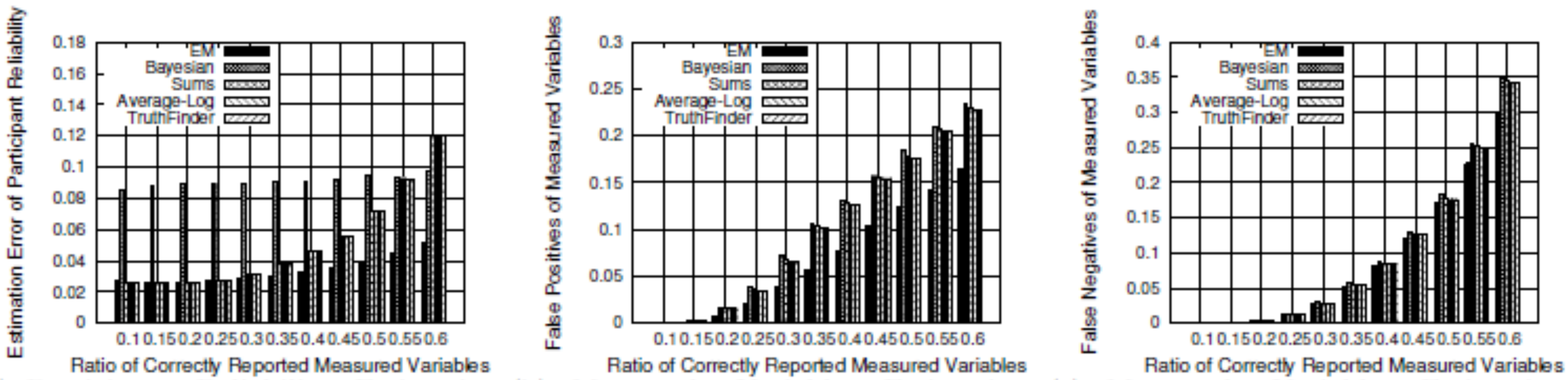
Number of False Measured Variables: 1000

Average Number of Claims per Participant: 100-1000

Expectation Maximization

Simulation

Estimation Accuracy of EM vs Baselines with <u>Varying Ratio of True Measured Variables</u>



(a) Participant Reliability Estimation Accuracy

(b) Measured Variable Estimation: False Positives

(c) Measured Variable Estimation: False Negatives

EM outperforms state-of-art heuristics

Parameters:

Number of Participants: 30

Number of Total Measured Variables: 2000

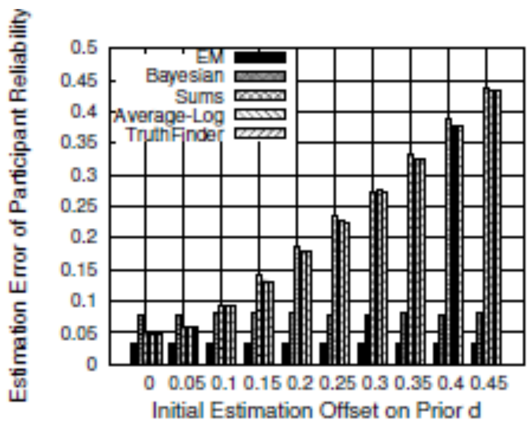
Average Number of Observations per Participant: 150

Ratio of true Measured Variables: 0.1-0.6

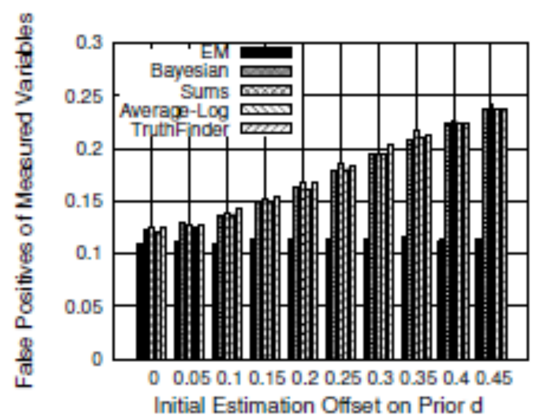
Expectation Maximization

Simulation

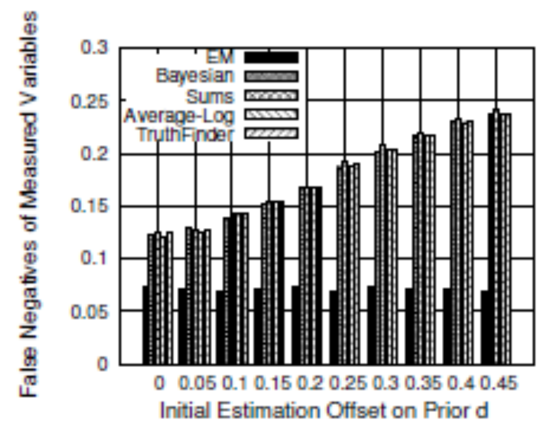
Estimation Accuracy of EM vs Baselines with <u>Initial Estimation offset on Claim Prior (d)</u>



(a) Participant Reliability Estimation Accuracy



(b) Measured Variable Estimation: False Positives



(c) Measured Variable Estimation: False Negatives

EM outperforms state-of-art heuristics

Parameters:

Number of Participants: 30

Number of True Measured Variables: 1000

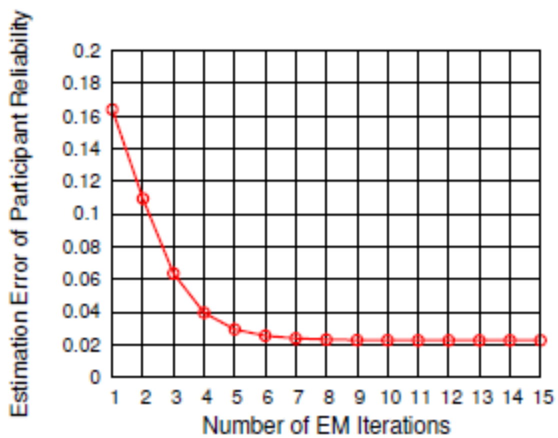
Number of False Measured Variables: 1000

Average Number of Observations per Participant: 150

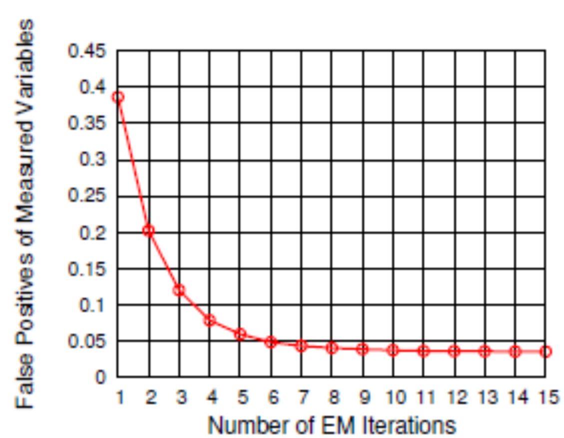
Initial Estimation Offset on Prior d: 0.1-0.45

Evaluation

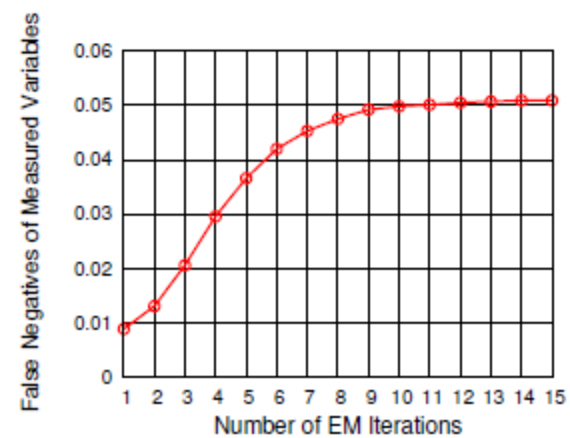
Estimation Convergence of EM



(a) Participant Reliability Estimation Accuracy



(b) Measured Variable Estimation: False Positives



(c) Measured Variable Estimation: False Negatives

EM Converges Quickly

Parameters:

Number of Participants: 50

Number of Measured Variables: 1000

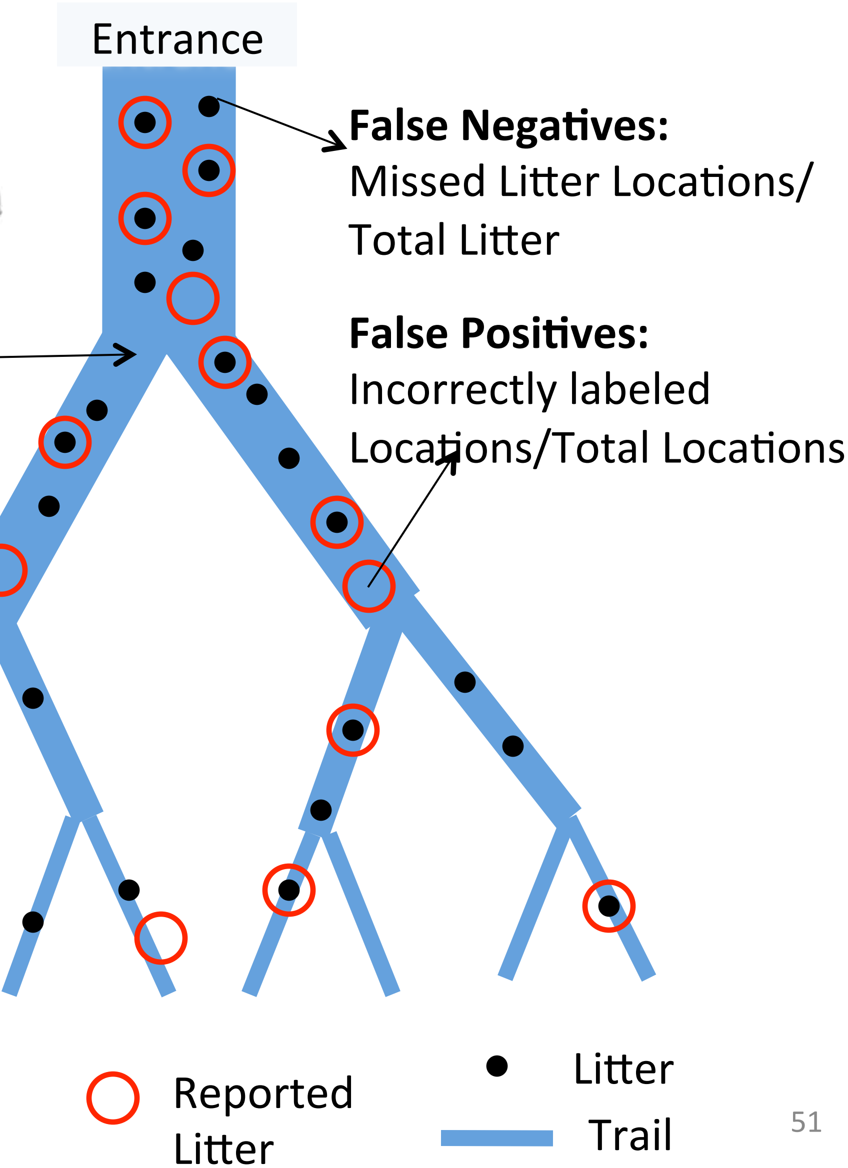
Number of False Measured Variables: 1000

Average Number of Observations per Participant: 250

Initial Estimation offset on d: 0.3

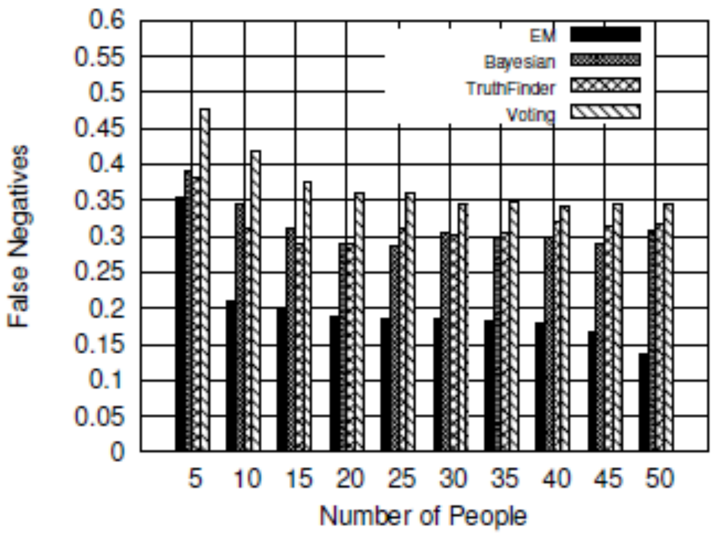
Expectation Maximization

Emulated Geotagging Application

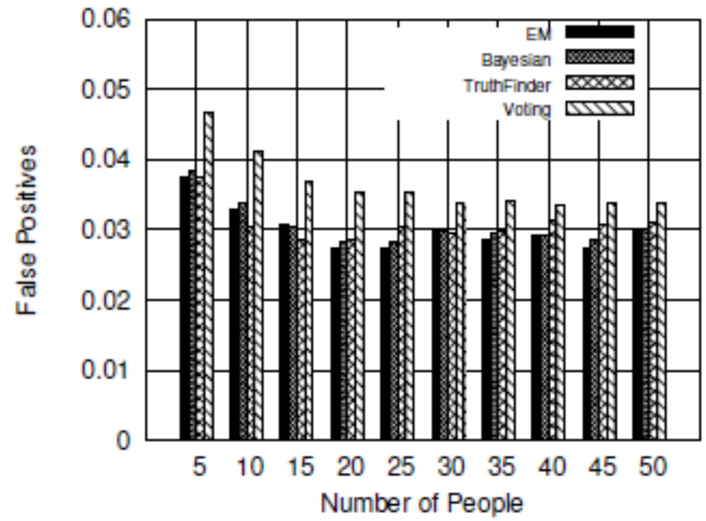


Expectation Maximization

Emulated Geotagging Application



(a) False Negatives (missed/total litter)



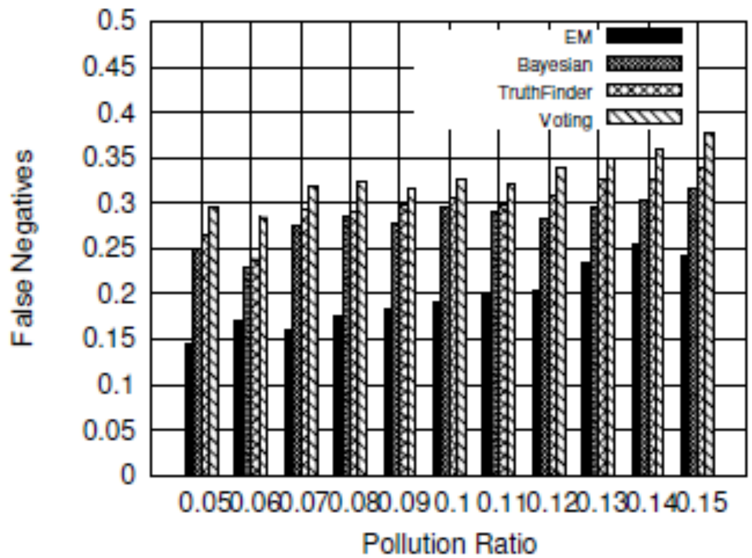
(b) False Positives (false/total locations)

Litter Geotagging Accuracy <u>versus Number of People</u>

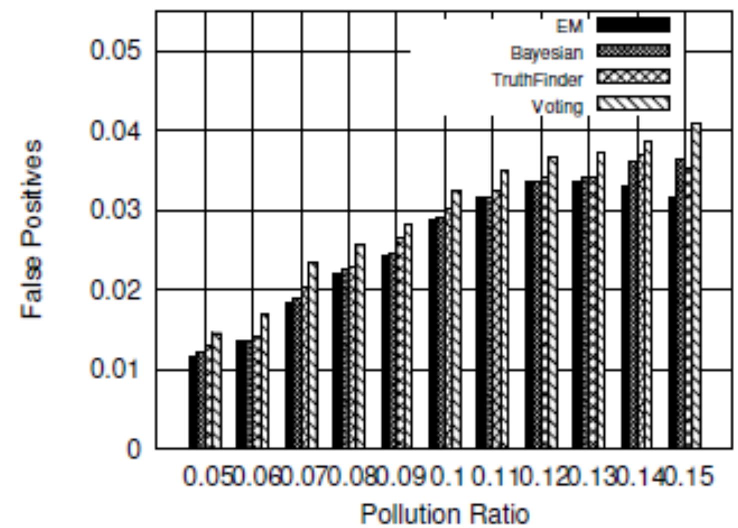
EM is better at detecting correct litter locations than baselines

Expectation Maximization

Emulated Geotagging Application



(a) False Negatives (missed/total litter)



(b) False Positives (false/total locations)

Litter Geotagging Accuracy <u>versus Park Pollution Ratio</u>

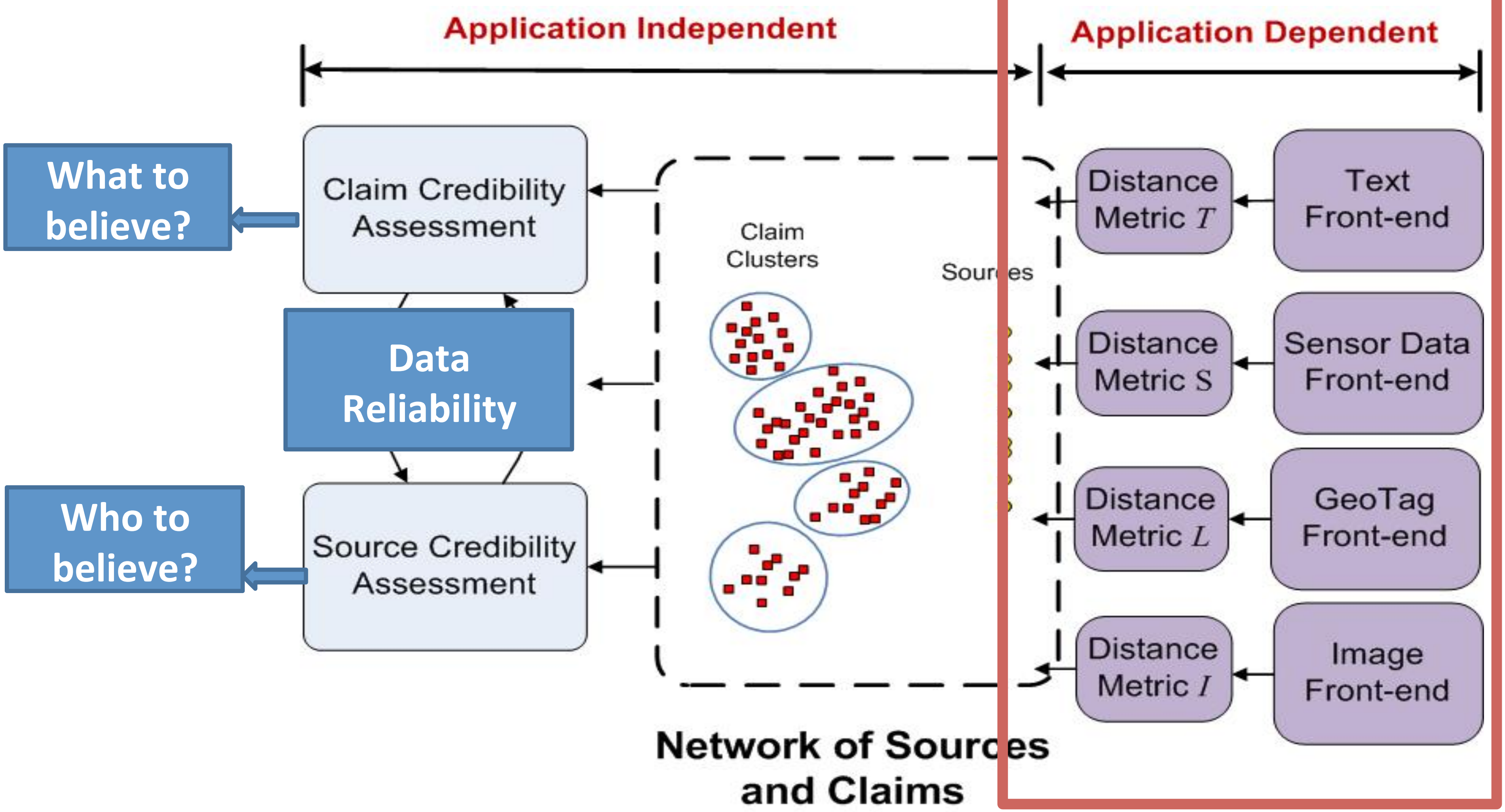
EM is better at detecting correct litter locations than baselines

The Apollo Fact-finder

http://apollo.cse.nd.edu/

The Apollo Fact-finder

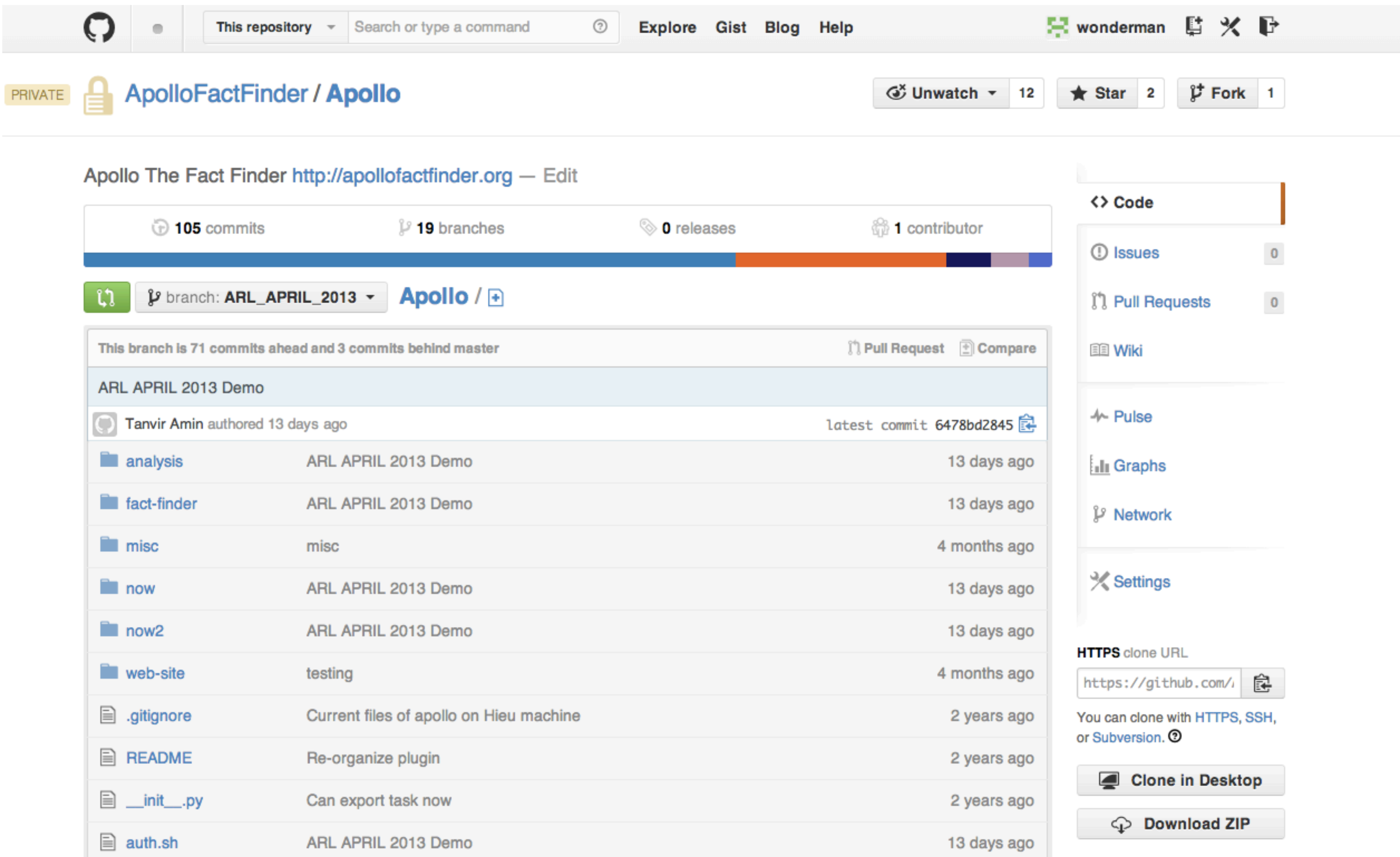
http://apollo.cse.nd.edu/



The Architecture of Apollo

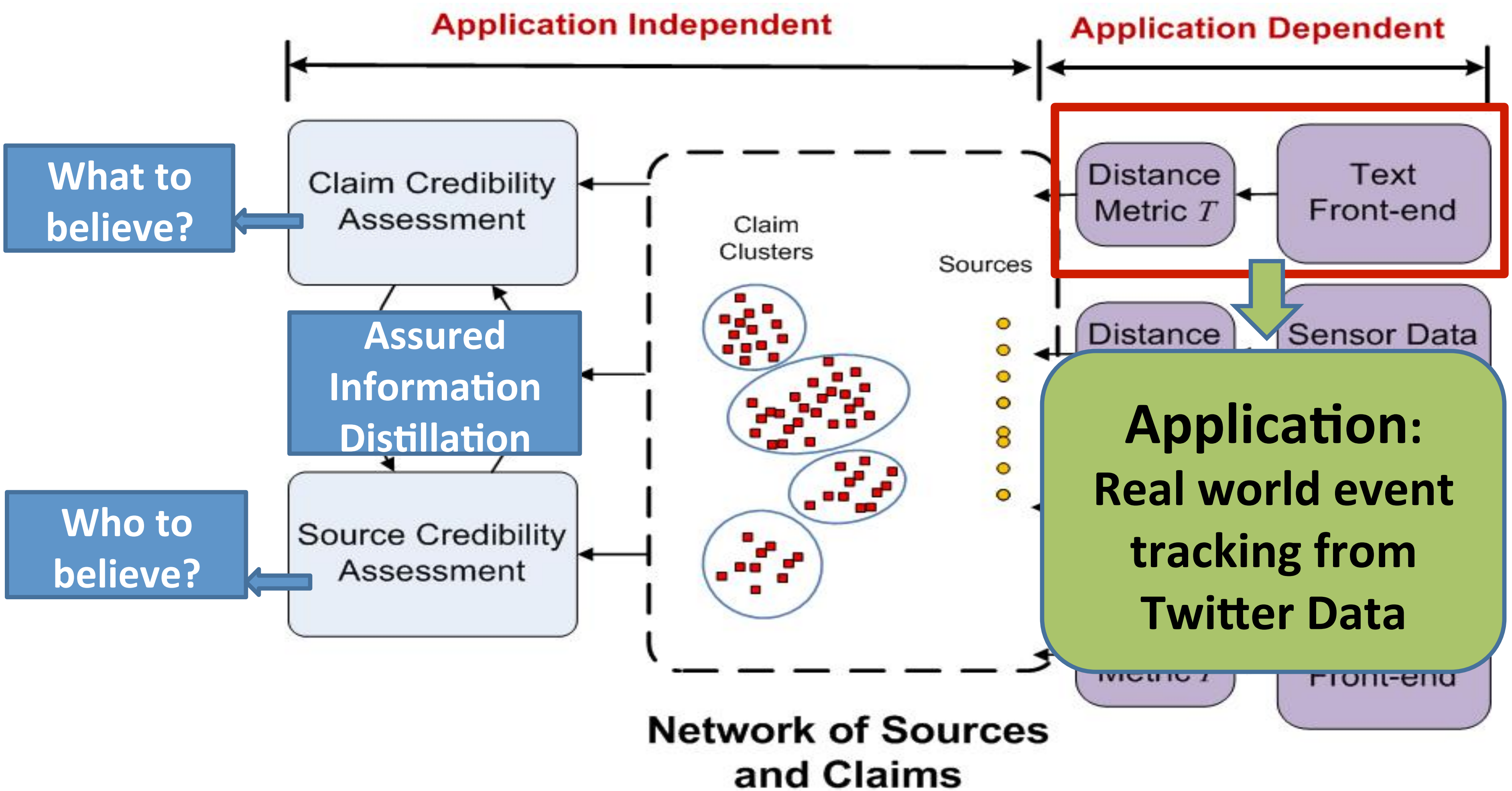
Github: Apollo Source Code

https://github.com/ApolloFactFinder/Apollo



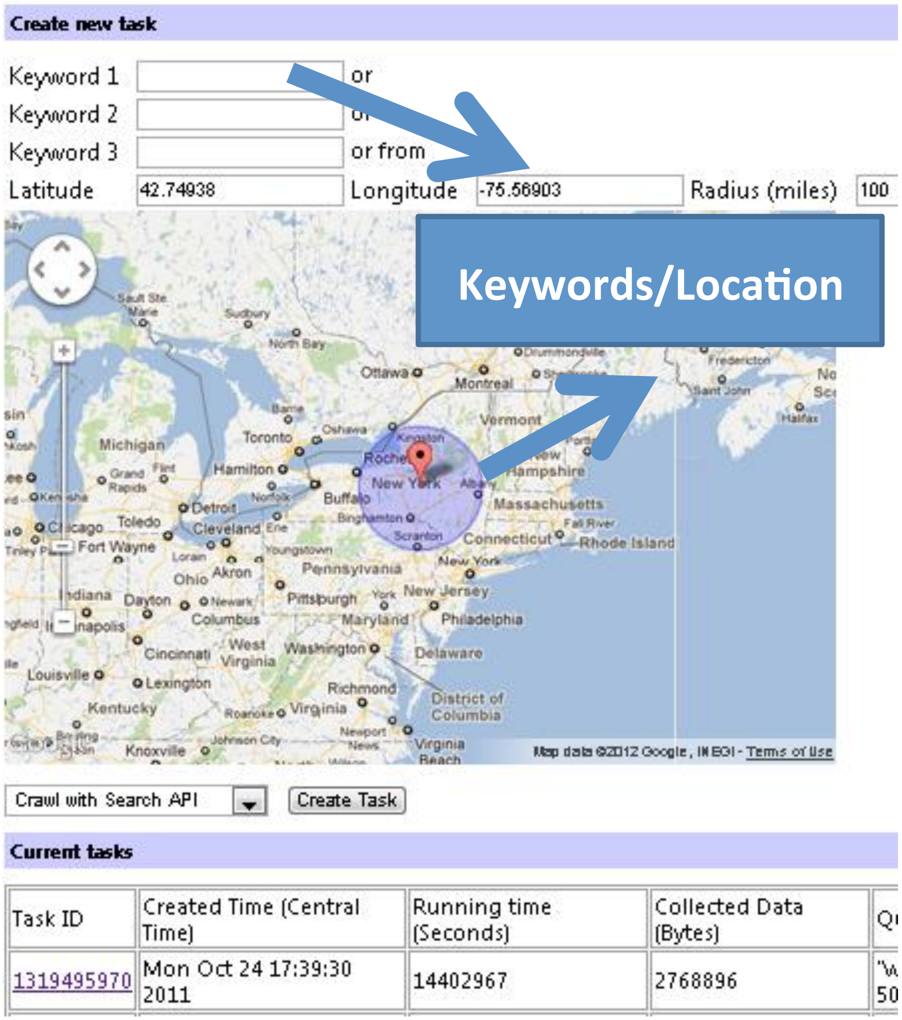
Evaluation of EM using Twitter Data

The Apollo Fact-finder

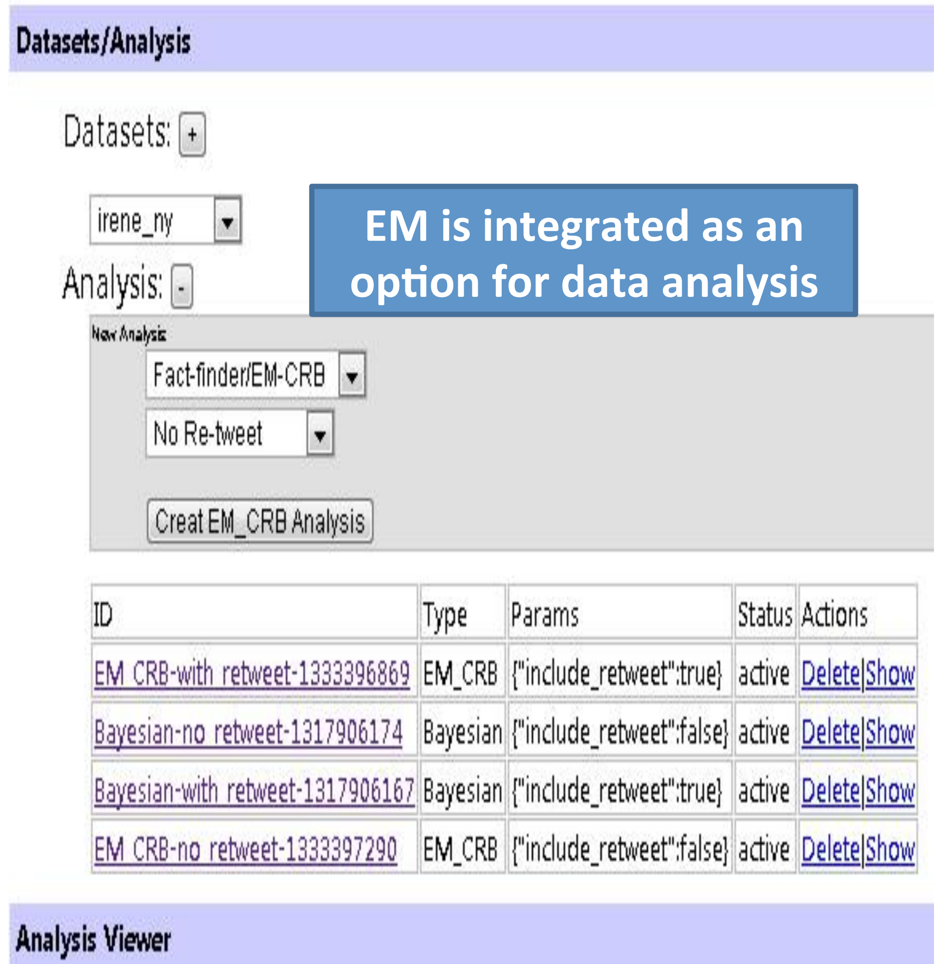


EM is Integrated with Apollo

A Real World Application



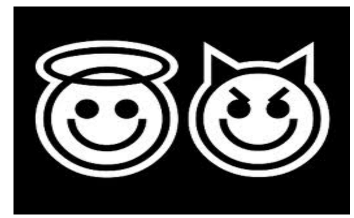
Data Collection Frontend



Information Analysis Frontend

Real World Event Tracking from Twitter Data

Difficult to verify truthfulness of tweets



Difficult to find out trustworthiness of sources

Datasets collected:

Egypt Unrest

Hurricane Sandy

Japan Tsunami and Nuclear Event

Boston Bombing

Egypt Unrest
(1.5M Tweets, Feb. 2011)

Media Report	Tweet found by EM
The leadership of Egypt's **ruling National Democratic Party resign**, including Gamal Mubarack, the son of Hosni Mubarak. Hossam ... party ... y-general.	Leadership of **Egypt's ruling party resigns** http://bit.ly/hebSGm.

2011-2-3 1:02 (CNN)

2 hours before

2011-2-2 22:03 (Tweets)

Top correct tweets found by EM matches well with Media Reports

Hurricane Sandy
(2 M Tweets, Nov. 2012)

#	Media	Tweet found by EM
1	More than 60 million Americans are braced for the impact of Hurricane Sandy after forecasters said it could be the biggest storm ever to hit the US mainland.	US East Coast braced for storm: People in the US East Coast are urged to prepare for the arrival of Hurricane Sandy... http://t.co/j7AYo6I9
2	New York Governor Andrew Cuomo says New York City subway, bu[illegible]d train service will be suspended sta[illegible]g at 7pm ET ahead of Hurricane Sa[illegible]y.	Governor Cuomo Announces Public Transportation Shutdown At 7 PM Tonight: ATTN: New Yorkers [illegible]
5	Hurricane Sandy, one of the largest storms to ever hit the East Coast, pummeled Massachusetts on Monday with punishing winds and dangerously high seas, flooding some coastal areas, and cutting power to some 385,000 homes.	[illegible] East Coast, pummeled Massachusetts: Hurricane Sandy, one of the largest storms to ever h... http://t.co/hKqFeLf6
6	Hurricane Sandy caused destruction up the Eastern Seaboard from North Carolina to Maine, with much of the damage centered in New York and New Jersey.	The aftermath from every angle: Hurricane Sandy leaves a trail of damage across the East Coast. http://t.co/Ah0EPafh
7	NEW YORK (AP) Superstorm Sandy knocked out a quarter of the cell towers in an area spreading across 10 states, [illegible]	Hurricane Sandy has knocked out 25 percent of all cell towers, cable services in 10 [illegible]owski...
10	On the East Coast, Election Day was hectic. Sixty-six polling places in New York had to be moved. The governor also signed an executive order, so New Yorkers could vote anywhere.	Sandy disrupts voting; U.S. Northeast braces for new storm http://t.co/SJZupk8G

Media Report	Tweet found by EM
New Jersey and New York, struggling to recover from the wreckage of Sandy, **were staggered today by a gas shortage**	Live gas chat: Hurricane Sandy causes **massive lines at N.J. gas stations** http://t.co/9IAbN0l2

Top correct tweets found by EM matches well with Media Reports

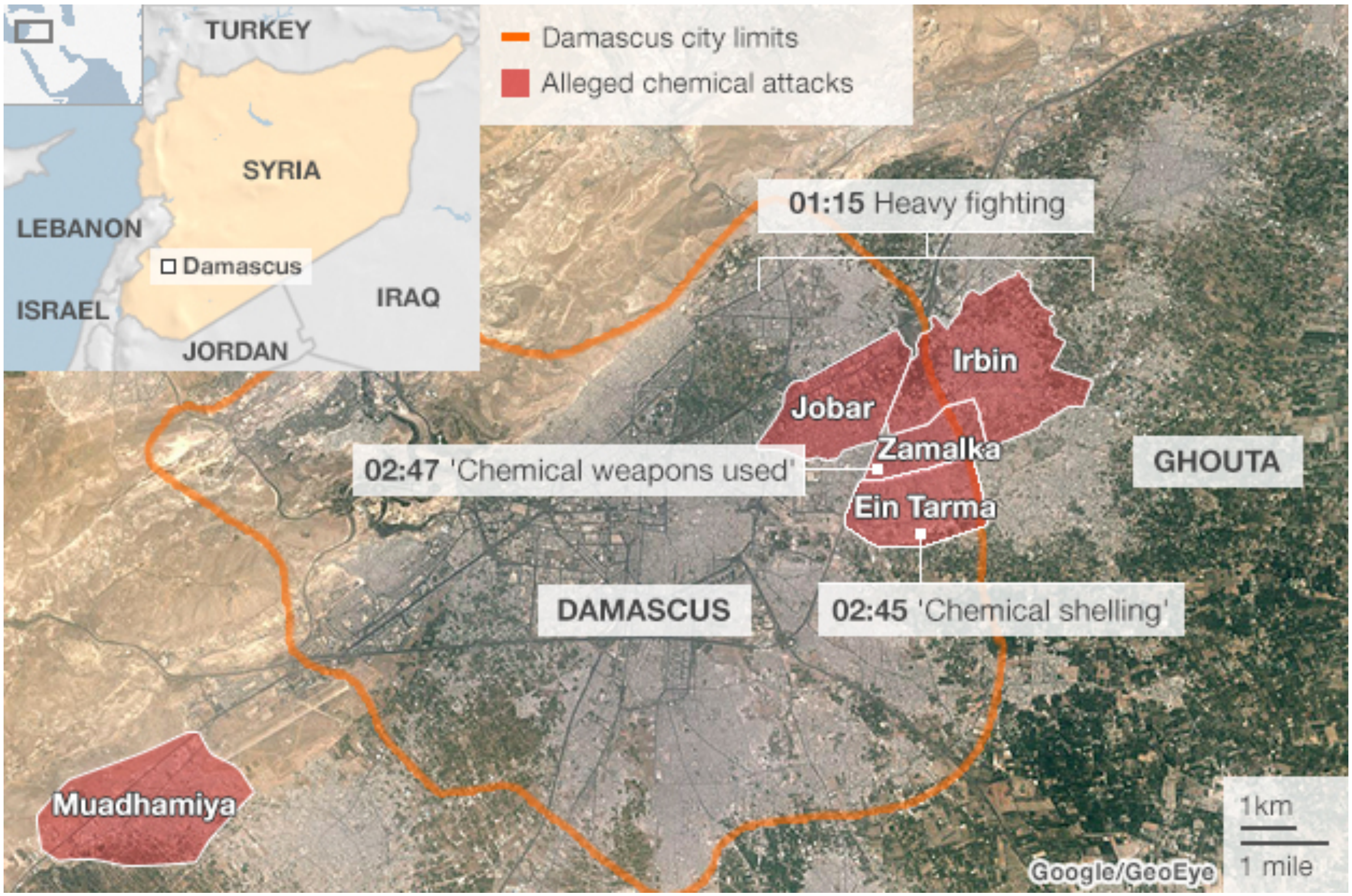
Japan Nuclear Disaster (1.0M Tweets, March 2011)

Media Articles	Top Tweets Returned by Triage Tool
The first earthquake hits Japan	RT @Hu.PostWorld: BREAKING: Massive 7.9 earthquake reportedly rocks Japan, 19-foot tsunamis feared http://hu..to/ezzmQb
The government warns the possibility of radiation leak	RT @Reuters: Japan warns of radiation leak from quake-hit plants http://t.co/iAFcDZg
Large number of dead and missing were reported	RT @BreakingNews: Latest Japan quake toll: 398 dead, 805 missing - Kyodo
Prediction of high probability of nuclear meltdown at Fukushima	RT @Reuters: FLASH: #Japan nuclear authorities say high possibility of meltdown at Fukushima Daiichi No. 1 reactor - Jiji
Cooling system at Fukushima nuclear plant failed	RT @komonews: RT @Reuters: Japan's nuclear safety agency says Fukushima Daiichi Nuclear Plant No. 3 reactor's emergency cooling system not functioning

Japan Nuclear Disaster (1.0M Tweets, March 2011)

Media Articles	Top Tweets Returned by Triage Tool
The first earthquake hits Japan	RT @Hu.PostWorld: BREAKING: Massive 7.9 earthquake reportedly rocks Japan, 19-foot tsunamis feared http://hu..to/ezzmQb
The government warns the possibility of radiation leak	RT @Reuters: Japan warns of radiation leak from quake-hit plants http://t.co/iAFcDZg
Large number of dead and missing were reported	RT @BreakingNews: Latest Japan quake toll: 398 dead, 805 missing - Kyodo
Prediction of high probability of nuclear meltdown at Fukushima	RT @Reuters: FLASH: #Japan nuclear authorities say high possibility of meltdown at Fukushima Daiichi No. 1 reactor - Jiji
Cooling system at Fukushima nuclear plant failed	RT @komonews: RT @Reuters: Japan's nuclear safety agency says Fukushima Daiichi Nuclear Plant No. 3 reactor's emergency cooling system not functioning

Example: Early Warning
Syrian WMD Attack, August 21st, 2013



Example:
Syrian WMD Attack

Triage Result: Recommended for Viewing
Medecins Sans Frontieres says it treated about 3,600 patients with 'neurotoxic symptoms' in Syria, of whom 355 died http://t.co/eHWY77jdS0
Weapons expert says #Syria footage of alleged chemical attack "difficult to fake" http://t.co/zfDMujaCTV
U.N. experts in Syria to visit site of poison gas attack http://t.co/jol8OlFxnf via @reuters #PJNET
Syria Gas Attack: 'My Eyes Were On Fire' http://t.co/z76MiHj0Em
Long-term nerve damage feared after Syria chemical attack http://t.co/8vw7BiOxQR
Syrian official blames rebels for deadly attack http://t.co/76ncmy4eqb
Assad regime responsible for Syrian chemical attack, says UK government http://t.co/pMZ5z7CsNZ
US forces move closer to Syria as options weighed: WASHINGTON (AP) — U.S. naval forces are moving closer to Sy... http://t.co/F6UAAXLa2M
400 tonnes of arms sent into #Syria through Turkey to boost Syria rebels after CW attack in Damascus --> http://t.co/KLwESYChCc
UN Syria team departs hotel as Assad denies attack http://t.co/O3SqPoiq0x
Vehicle of @UN #Syria #ChemicalWeapons team hit by sniper fire. Team replacing vehicle & then returning to area.
International weapons experts leave Syria, U.S. prepares attack. More @ http://t.co/4Z62RhQKOE
Military strike on Syria would cause retaliatory attack on Israel, Iran declares http://t.co/M950o5VcgW
Asia markets fall on Syria concerns: Asian stocks fall, extending a global market sell-off sparked by growing ... http://t.co/06A9h2xCnJ

Example:
Syrian WMD Attack

Triage Result: Recommended for Viewing
Medecins Sans Frontieres says it treated about 3,600 patients with 'neurotoxic symptoms' in Syria, of whom 355 died http://t.co/eHWY77jdS0
Weapons expert says #Syria footage of alleged chemical attack "difficult to fake" http://t.co/zfDMujaCTV
U.N. experts in Syria to visit site of poison gas attack http://t.co/jol8OlFxnf via @reuters #PJNET
Syria Gas Attack: 'My Eyes Were On Fire' http://t.co/z76MiHj0Em
Long-term nerve damage feared after Syria chemical attack http://t.co/8vw7BiOxQR
Syrian official blames rebels for deadly attack http://t.co/76ncmy4eqb
Assad regime responsible for Syrian chemical attack, says UK government http://t.co/pMZ5z7CsNZ
US forces move closer to Syria as options weighed: WASHINGTON (AP) — U.S. naval forces are moving closer to Sy... http://t.co/F6UAAXLa2M
400 tonnes of arms sent into #Syria through Turkey to boost Syria rebels after CW attack in Damascus --> http://t.co/KLwESYChCc
UN Syria team departs hotel as Assad denies attack http://t.co/O3SqPoiq0x
Vehicle of @UN #Syria #ChemicalWeapons team hit by sniper fire. Team replacing vehicle & then returning to area.
International weapons experts leave Syria, U.S. prepares attack. More @ http://t.co/4Z62RhQKOE
Military strike on Syria would cause retaliatory attack on Israel, Iran declares http://t.co/M950o5VcgW
Asia markets fall on Syria concerns: Asian stocks fall, extending a global market sell-off sparked by growing ... http://t.co/06A9h2xCnJ

An Example Timeline (Abbreviated)

Charlie Hebdo Shooting at Paris (Jan. 07-11, 2015)

- Distilling hundreds of thousands of raw tweets into a summary of the event (told in tweets selected by Apollo)

Example: The Paris Shooting in Tweets (Apollo Summary)

- Jan 7: Massacre at French magazine office: **Gunmen have attacked** the Paris office of the **French satirical magazine**... http://t.co/6DXhg5taXI
- Jan 7: French officials **identify 3 suspects** in Paris terror attack that **left 12 dead:** http://t.co/gpkbQNoACy #Fox #News #AN247

Example: The Paris Shooting in Tweets (Apollo Summary)

- Jan 8: Paris terror attack **suspect surrenders to police**: One of three suspects in the Charlie Hedbo attack in Paris W... http://t.co/jHtEJR5HLL
- Jan 8: Tensions running high as **two people injured in separate Paris shooting** http://t.co/UzkQCtIR9J

Example: The Paris Shooting in Tweets (Apollo Summary)

- Jan 8: Paris terror suspects reportedly **spotted in northern France**, police flood scene | Fox News http://t.co/cDpNEtrsnm
- Jan 8: French police **swarm forest 'larger than Paris'** in hunt for Charlie Hebdo jihadist assassins http://t.co/5mxovSzp54

Example: The Paris Shooting in Tweets (Apollo Summary)

- Jan 9: **Police surround Paris attack suspects**, at least one hostage reported taken | http://t.co/fs1opaqDF8
- Jan 9: **French police kill Paris massacre suspects**, hostage-taking ally in separate raids | http://t.co/gI5EjMpYO9

Example: The Paris Shooting in Tweets (Apollo Summary)

- Jan 10: THE AL QAEDA CONNECTION **Yemen terror group claims it directed Paris massacre** as cops kill http://t.co/trPDUXLFIK
- Jan 11: #BREAKING Prosecutor: **Man who killed kosher market hostages**, policewoman, now **linked to third shooting in #Paris**. http://t.co/Hhi2WtX02u

Demo

- http://apollo.cse.nd.edu/now/
- http://apollo.cse.nd.edu/analysis/browse3.html