



Self-regulatory Information Sharing in Participatory Social Sensing

Evangelos Pournaras, Jovan Nikolic, Pablo Velasquez, Marcello Trovati, Nik Bessis, Dirk Helbing

Opportunities

**MIT
Technology
Review**

Big Data and Analytics: Here, There, and Everywhere

In this content collection, MIT Technology Review looks at the data explosion from multiple angles. Editors and contributors examine how big data is revolutionizing shale-oil production—and how the big-data boom may be leaving poorer nations behind. They also look at the growing role of data analytics in everything from increasing crop production to gauging driving efficiency.

01 This Car Knows Your Next Misstep Before You Make It

Researchers trained a computer to recognize the behavior that precedes a particular maneuver.

October 1, 2015



Advertisement

02 Trick That Doubles Wireless Data Capacity Stands Up in Cell Network Tests

Major wireless carriers have begun testing a technology that can double the capacity of any wireless data connection.

September 30, 2015



WIRED

BUSINESS

CULTURE

DESIGN

GEAR

SCIENCE

SECURITY

Log in / Register

Search

SHARE

Subscribe

Data Is the New Oil of the Digital Economy

DATA IS THE NEW OIL OF THE DIGITAL ECONOMY

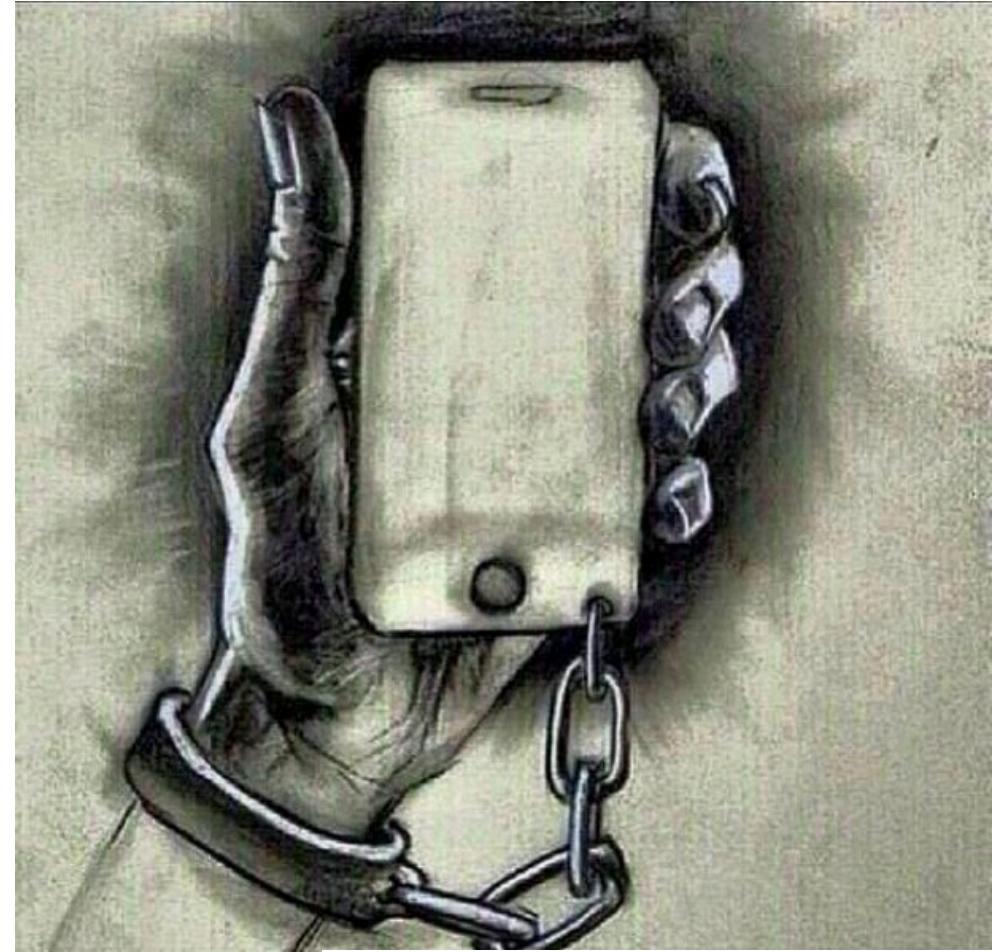
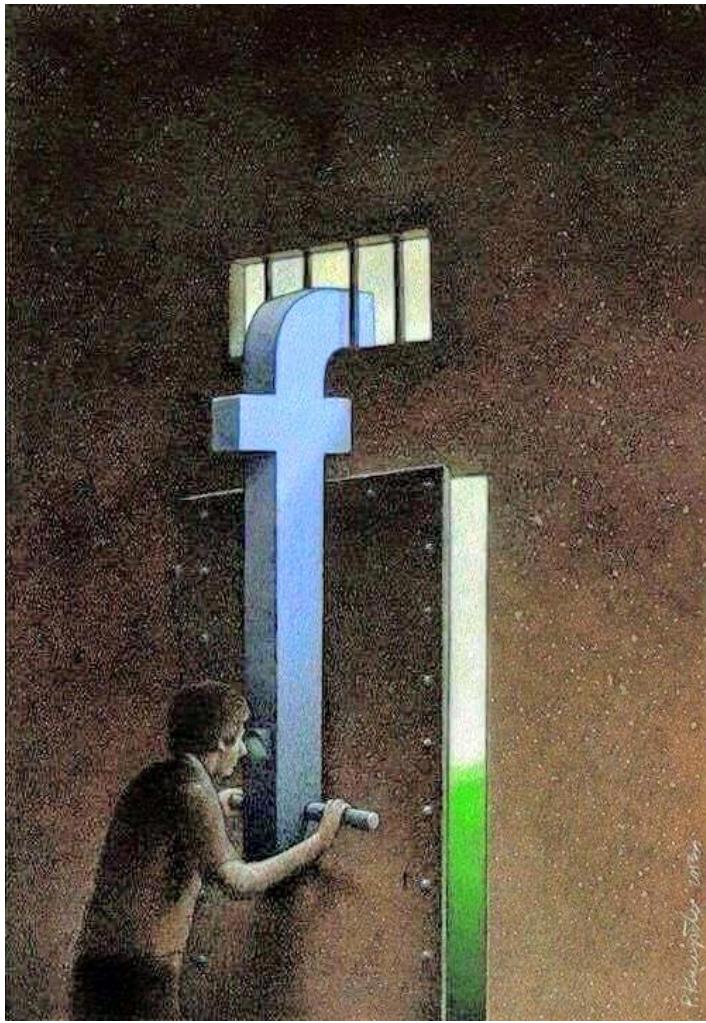


Image: verifex/Flickr

DATA IN THE 21st Century is like Oil in the 18th Century: an immensely, untapped valuable asset. Like oil, for those who see Data's fundamental value and learn to extract and use it there will be huge rewards.

We're in a digital economy where data is more valuable than ever. It's the key to the smooth functionality of everything from the government to local companies. Without it, progress would halt.

Challenges



Challenges

Harvard Business Review

THE MAGAZINE BLOGS VIDEO BOOKS CASES WEBINARS COURSES

Guest Subscribe today and get access to all current articles and HBR online archive.

HBR Blog Network



Big Data's Dangerous New Era of Discrimination

by Michael Schrage | 8:00 AM January 29, 2014

Comments (30) [Email](#) [Twitter](#) [LinkedIn](#) [Facebook](#) [Google+](#) [Print](#)

Congratulations. You bought into Big Data and it's paying off Big Time. You slice, dice, parse and process every screen-stroke, clickstream, Like, tweet and touch point that matters to your enterprise. You now know exactly who your best — and worst — customers, clients, employees and partners are. Knowledge is power. But what kind of power does all that knowledge buy?

Big Data creates Big Dilemmas. Greater knowledge of customers creates new potential and power to discriminate. Big Data — and its associated analytics — dramatically increase both the dimensionality and degrees of freedom for **detailed discrimination**. So where, in your corporate culture and strategy, does value-added personalization and segmentation end and harmful

ZDNet

White Papers Hot Topics Downloads Reviews Newsletters

US Edition Internet of Things Mobility Research Windows Enterprise Software

ARE YOUR HR & FINANCE SYSTEMS BASED ON TECH OLDER THAN 2000?

MUST READ [Hackers jump on the Shellshock Bash bandwagon](#)

Topic: **Big Data** Follow via: [RSS](#) [Email](#)

Why big data evangelists should be sent to re-education camps

Summary: *Big data is a dangerous, faith-based ideology. It's fuelled by hubris, it's ignorant of history, and it's trashing decades of progress in social justice.*

By Stilgherrian for The Full Tilt | September 19, 2014 -- 07:13 GMT (00:13 PDT)

[Follow @stilgherrian](#) [Get the ZDNet Big Data newsletter now](#)

Comments 27 [Votes](#) 7 [Share](#) 459 [Twitter](#) 692 [LinkedIn](#) Share more +

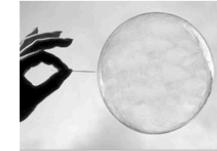
The last time I wrote about big data, in July, I called it a **big, distracting bubble**. But it's worse than that. Big data is an ideology. A religion. One of its most important gospels is, of course, at Wired.

In 2008, Chris Anderson talked up a thing called The Petabyte Age in **The End of Theory: The Data Deluge Makes the Scientific Method Obsolete**.

"The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all," he wrote.

Declaring the **scientific method** dead after 2,700 years is quite a claim. Hubris, even. But, Anderson wrote, "There's no reason to cling to our old ways." Oh, OK then.

Now, this isn't the first set of claims that correlation would supersede causation, and that the next iteration of computi



Challenges

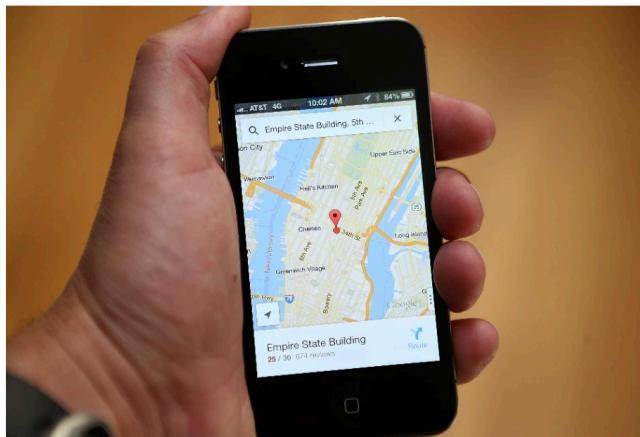


Existing **social mining** practices
threaten **social cohesion**



*“surveillance has become
increasingly privatized, commercialized
and participatory”, Julie E. Cohen*

Recent Views



We may think Google Maps is free, but we actually pay by giving it access to valuable data—our geo locations. (Photo by Justin Sullivan/Getty Images)

Forbes / Opinion

APR 1, 2016 @ 03:48 PM

2,051 VIEWS

Privacy Is The New Money, Thanks To Big Data



Omri Ben-Shahar
CONTRIBUTOR

I write about law, economics, and consumer markets

FOLLOW ON FORBES (4)



FULL BIO >

Opinions expressed by Forbes Contributors are their own.

The Apple/FBI showdown was the recent installment in an unfolding legal battle over privacy protection. Beginning with the Snowden revelations, it is widely thought that the major threat to our privacy in the digital era comes from the power of Big Government to access personal information stored in devices and websites. As this debate rages, we are losing sight of the other enterprise of personal data collection—known as “Big Data”—which is subject to less popular interest, but is far grander in scope, involves higher stakes and numerous ongoing legal battles.

The FBI or NSA data collection is Small Data. It focuses on meta-data or on few targeted individuals under investigation. In contrast, Big Data business is *really* big. I am talking about the collection of personal data by websites, mobile apps, retailers, insurance companies – any commercial entity that receives information from people. In the old brick-and-mortar world, firms had Pendaflex files about their customers, neatly tucked away in file cabinets. If you walked into a supermarket or bookstore and browsed the shelves, there would be no record of this activity. In the digital world, people leave their prints everywhere. The sum of our activities – where we browse, shop, or drive; what we read, eat, or own; who we chat with, like or love – is collected, neatly organized by algorithms, smartly analyzed by sophisticated software, and used or sold primarily for marketing purposes. It does not decay or gather dust, and it is never forgotten.

How The Citizen Data Scientist Will Democratize Big Data



Bernard Marr
CONTRIBUTOR

I write about big data, analytics and enterprise performance

FOLLOW ON FORBES (264)



FULL BIO >

Opinions expressed by Forbes Contributors are their own.

The rise of the citizen data scientist is a subject which is creating a lot of excitement at the moment. Put simply (and a bit bluntly) businesses, particularly larger ones with more mature **Big Data** analytical operations, are finding that it is too important to be left solely in the hands of the data scientists.

For a start – one reason is that there simply aren’t enough of them. That isn’t to say that data scientists – by which I mean staff with a formal education in business intelligence, statistics and roles purely involving data analytics – are no longer needed. They are, and I believe people with these backgrounds will continue to play a crucial role. But there is an ever growing plethora of tools and services designed to facilitate Big Data analytics outside of the IT lab and across the organization as a whole.

This is enabling the rise of what has been termed the “Citizen Data Scientist”. In fact, last year analysts at **Gartner** predicted that the demand for these people will increase five times more quickly than the demand for “traditional”, highly skilled data scientists.

Retailer Sears, for example, recently empowered 400 staff from its business intelligence (BI) operations to carry out advanced, Big Data driven customer segmentation – work which would previously have been carried out by specialist Big Data analysts, probably with PhDs. The move is said to have created hundreds of thousands of dollars’ worth of efficiencies in data preparation costs alone. Exploratory analysis, visualization and putting insights into action is also taken care of by this new class of Citizen Data Scientist.

Sears used tools provided by Platfora to allow its BI staff to effectively retrain and repurpose themselves as Big Data analysts. Platfora VP of products Peter Schlamp told me “customer segmentation is a very complex problem. It is not something your average Excel user can do.



Citizen Data Scientists (Source: Shutterstock)

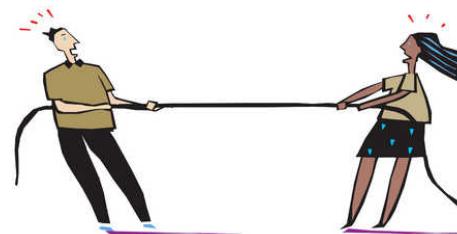
Opposing Views in Information Sharing

Fragmentation, polarization,
deceived citizens, social injustice,



Big Data Analytics vs. Privacy-preservation

More data,
more information, more
knowledge, more security,
more business opportunities,
more prosperity

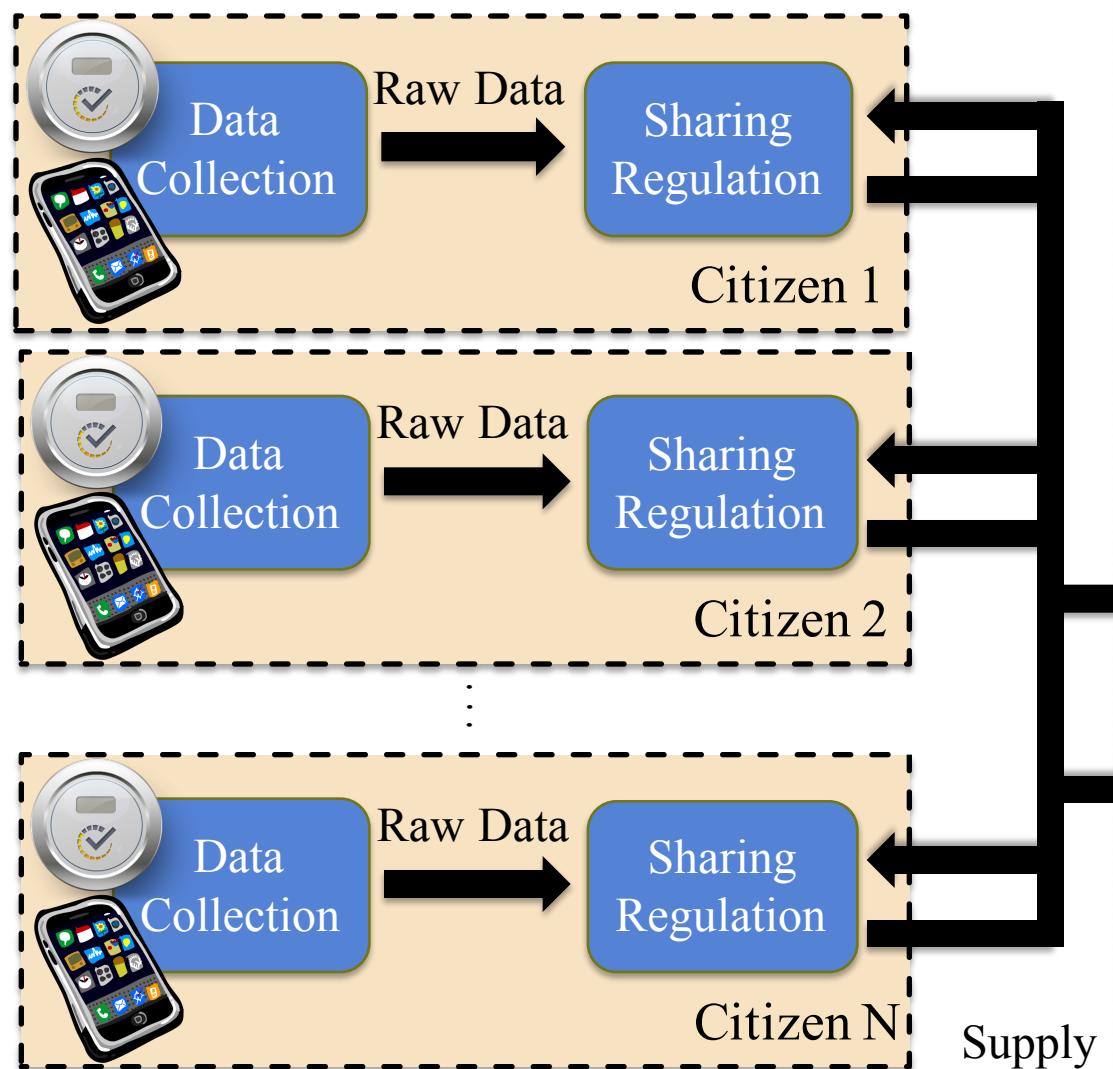


Less data,
less information, less surveillance, less
discrimination, more freedom/justice,
more social cohesion,
more prosperity

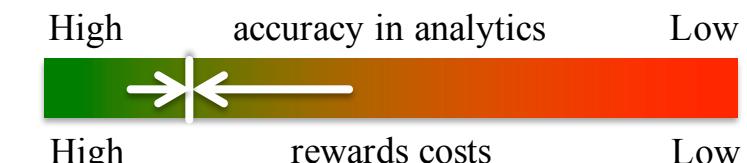
Research communities cancel out their contributions

How to bridge this gap?

Information Sharing



Citizen Selections



Aggregator Selections

Summarized Data

Rewards

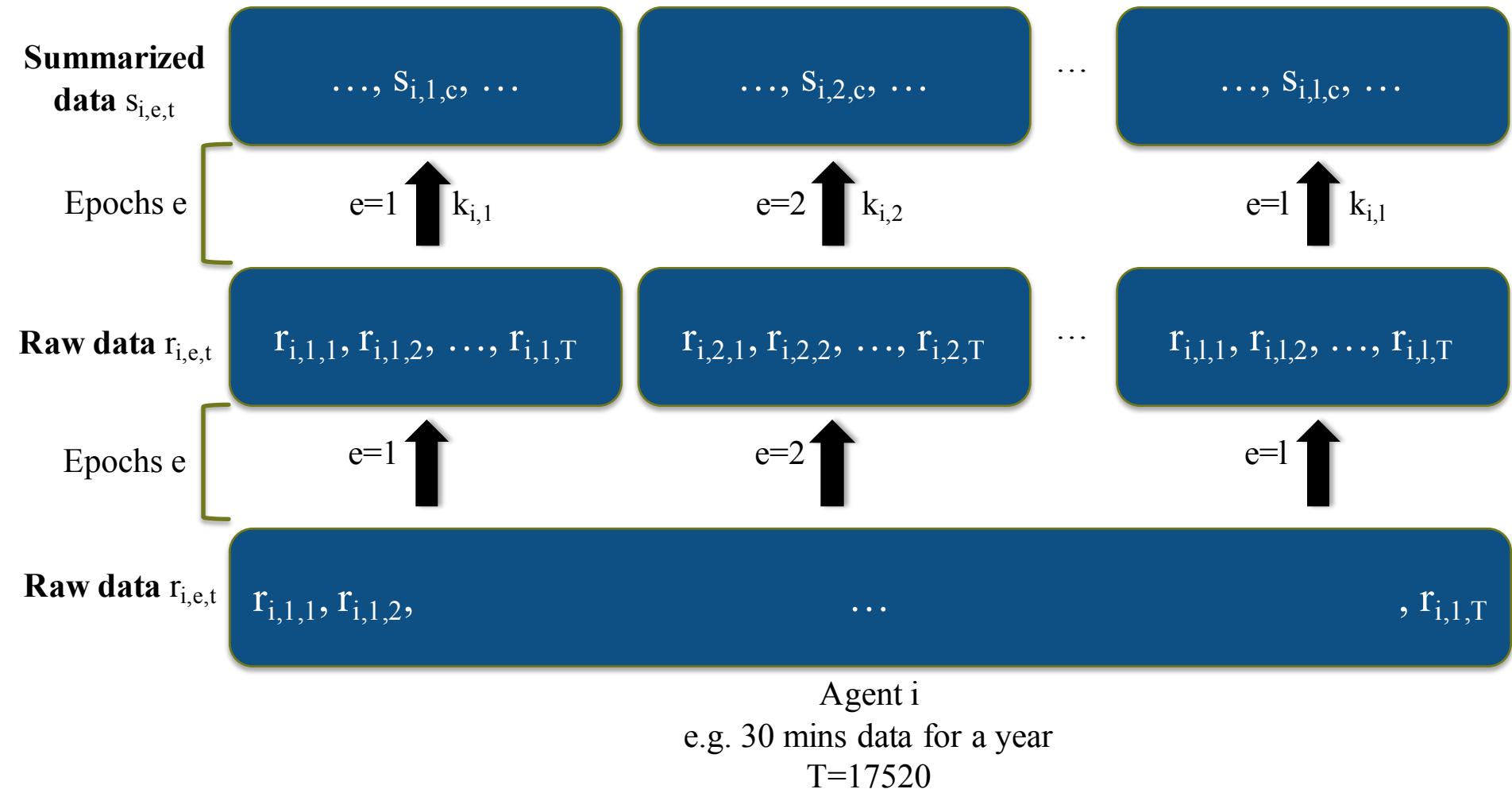
Analytics

Incentivization

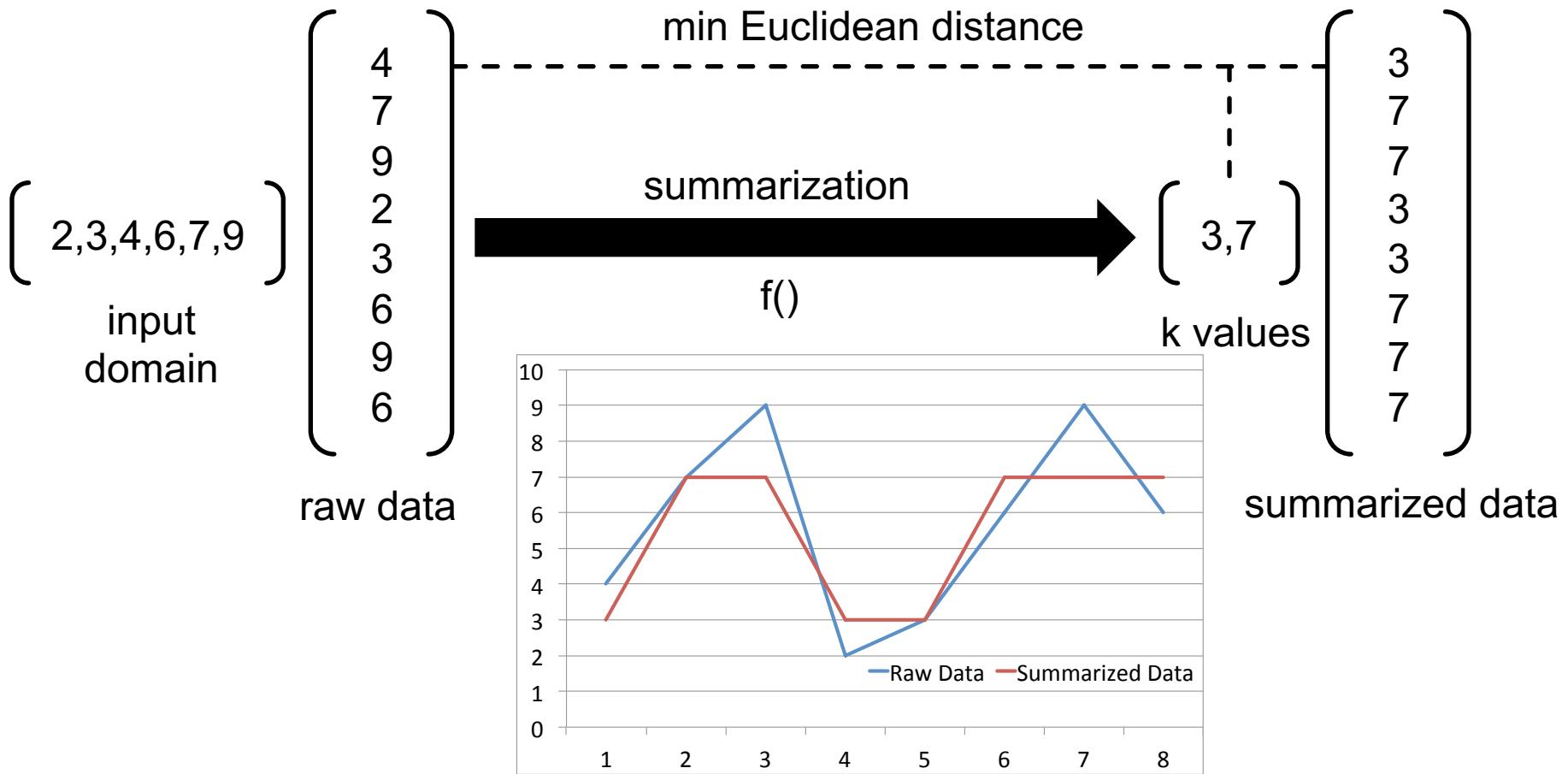
Demand

Supply

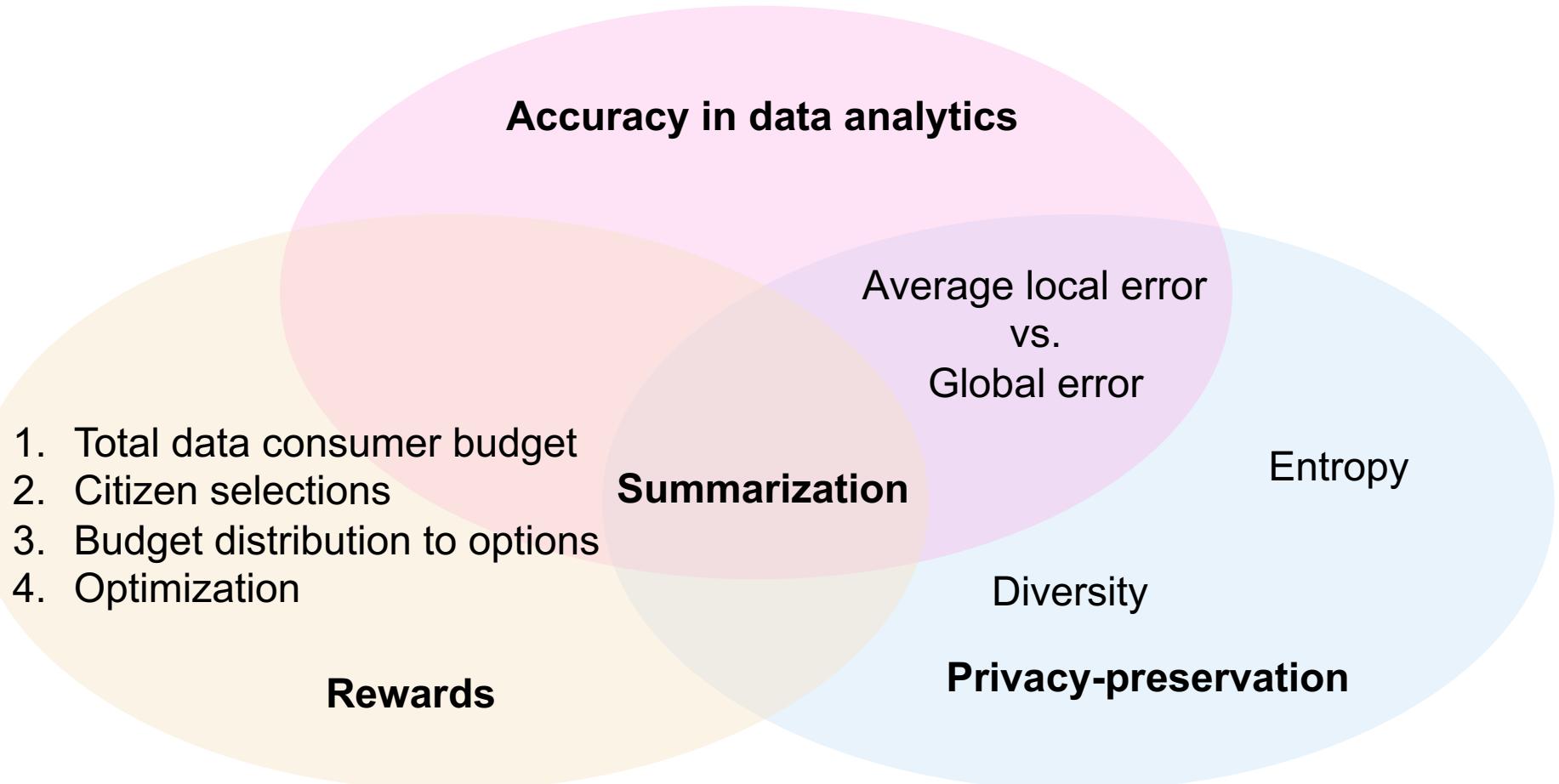
Privacy-preserving Information Sharing



Summarization



The Trade-offs of Information Sharing



The Trade-offs of Information Sharing

Symbol	Interpretation
i	An agent index
e	An epoch index
t	A time index within an epoch
T	Epoch duration
$R_{i,e}$	Sequence of raw data
$r_{i,e,t}$	A record of raw data
$S_{i,e}$	Sequence of summarized data
$s_{i,e,t}$	A record of summarized data
$f_s()$	Summarization function
j	An index for a possible summarization value
$c_{i,e,j}$	A possible summarization value
$k_{i,e}$	The number of possible summarization values
$/$	Number of epochs
$\alpha_{i,e}$	Summarization metric
$D_{i,e}$	Sequence of raw or summarization data
$H(D_{i,e})$	Entropy
$p_{i,e,j}$	Probability of a possible value occurring in an epoch
n_t	Occurrence or not of possible value at time t
$\beta_{i,e}$	Diversity
m_t	Change or not between two consecutive time periods t and $t + 1$
$\epsilon_{i,e,t}$	Local error
$\varepsilon_{i,e,t}$	Global error
n	Number of participating citizens
$\epsilon_{e,t}$	Average local error among citizens
γ_e	Total rewards that data aggregators are willing to provide
$P_r()$	Probability density function for rewards
z	Number of discrete participation levels
$P_s()$	Probability density function for summarization
$\gamma_{i,e}$	Rewards provided to agent i

Average local error

$$\epsilon_{e,t} = \frac{1}{n} \sum_{i=1}^n \epsilon_{i,e,t}$$

$$\epsilon_{i,e,t} = \frac{|r_{i,e,t} - s_{i,e,t}|}{|r_{i,e,t}|}$$

Entropy

$$H(D_{i,e}) = - \sum_{j=1}^{k_{i,e}} p_{i,e,j} \log_2 p_{i,e,j},$$

$$p_{i,e,j} = \frac{1}{T} \sum_{t=1}^T n_t, \quad n_t = \begin{cases} 1 & \text{if } c_{i,e,j} = d_{i,e,t}, \\ 0 & \text{if } c_{i,e,j} \neq d_{i,e,t}, \end{cases}$$

Diversity

$$\beta_{i,e} = \frac{1}{T-1} \sum_{t=1}^{T-1} m_t, \quad m_t = \begin{cases} 1 & \text{if } d_{i,e,t} = d_{i,e,t+1}, \\ 0 & \text{if } d_{i,e,t} \neq d_{i,e,t+1}, \end{cases}$$

Global error

$$\varepsilon_{e,t} = \frac{\left| \sum_{i=1}^n r_{i,e,t} - \sum_{i=1}^n s_{i,e,t} \right|}{\left| \sum_{i=1}^n r_{i,e,t} \right|},$$

Rewards

$$\gamma_{i,e} = \frac{\gamma_e * P_r(\alpha_{i,e})}{n * P_s(\alpha_{i,e})}.$$

Unsupervised learning

Implementation

Survey questions

Privacy preferences

Survey answers → summarization range

My household may decide to be more aware of the amount of electricity used by appliances we own or buy.

ECBT - Smart Grid
6435 participants

1 sensor
1 year

Datasets

Nervousnet
154 participants
several sensors
4 days

Several implementation algorithms

Summarization - Clustering

Fixed: Manual selection

Empirical: Citizens' preferences, semi-automated

Customizable – number of clusters

Algorithmic: Fully-automated, data-driven



Measurements & variables	ECBT	Nervousnet
Privacy	✓	✓
Accuracy	✓	✓
Costs & Rewards	✓	X
Epoch length	daily & weekly	daily
Summarization level	fixed, empirical & algorithmic	fixed & algorithmic
Number of citizens	✓	✓
Several sensor types	X	✓
Analytics	summation	average

Evaluation & Research Questions

Does summarization improve privacy?

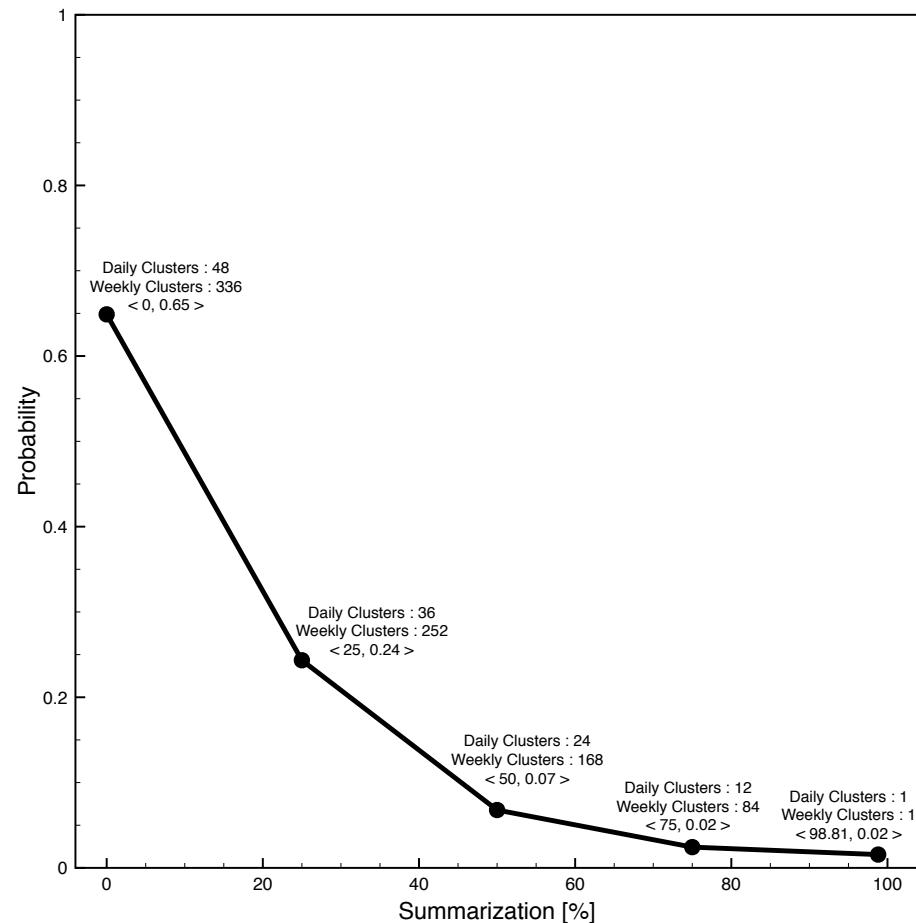
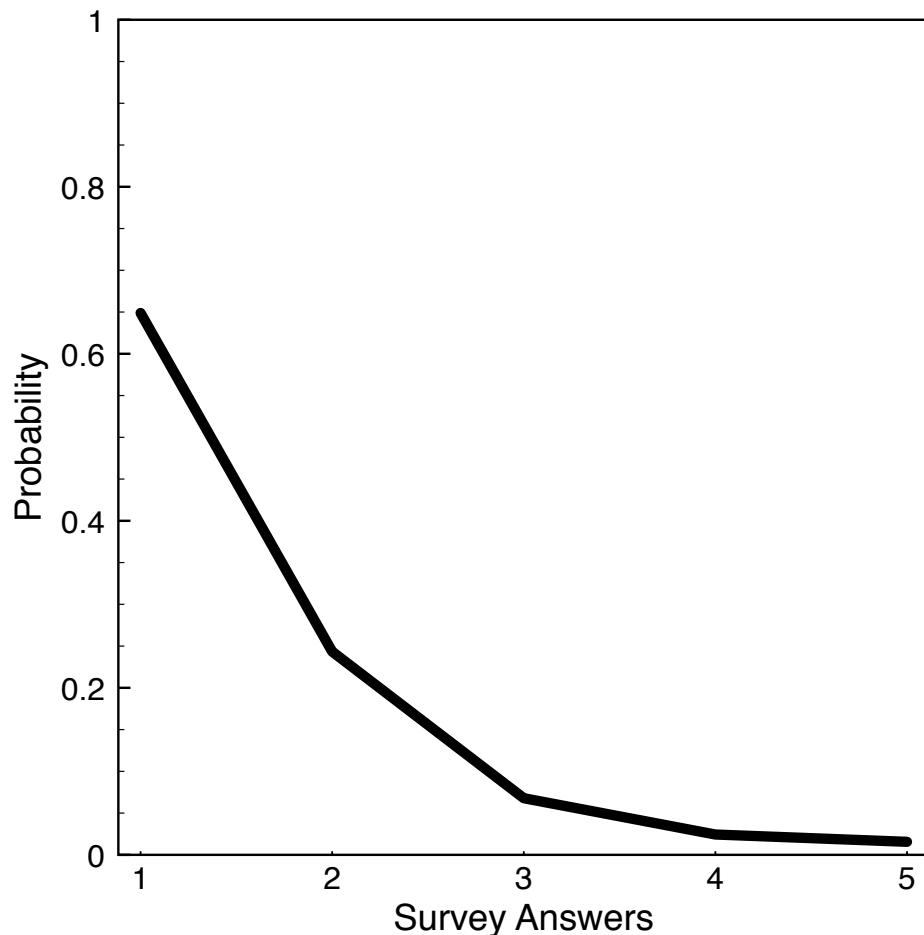
How does participation level influence privacy?

Which are the trade-offs between privacy & accuracy in analytics?

Does sensor/information type influence these trade-offs?

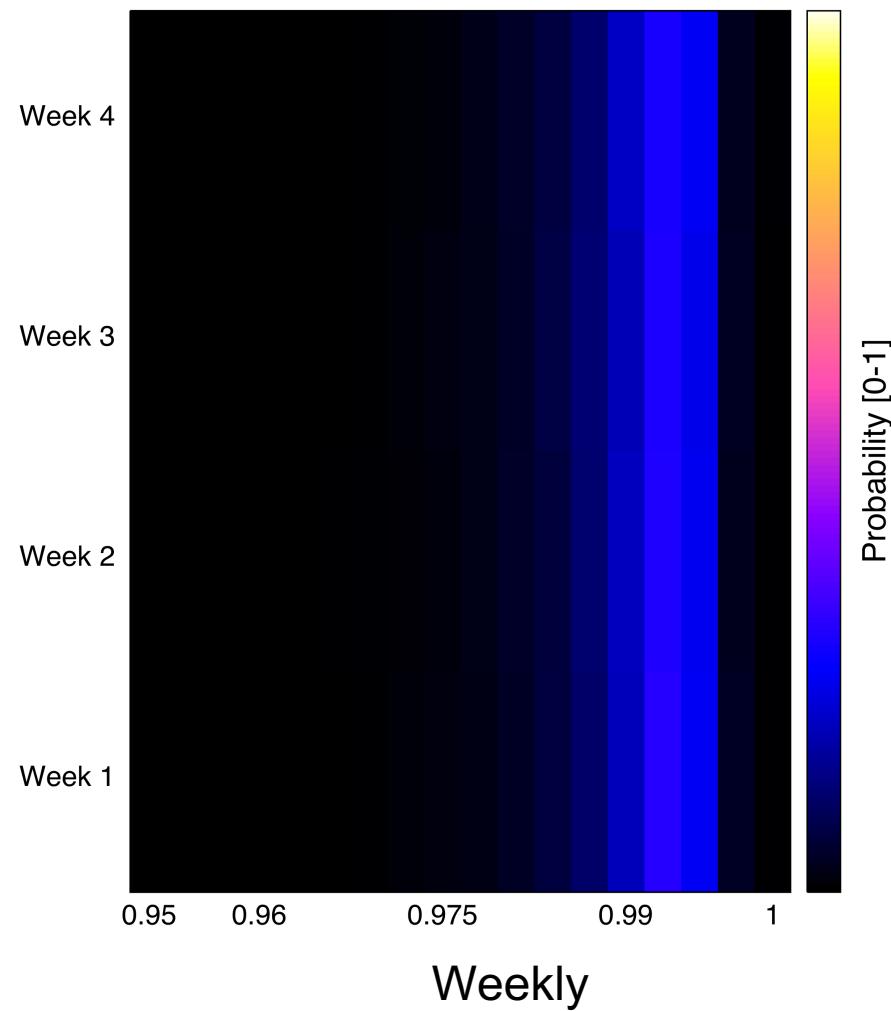
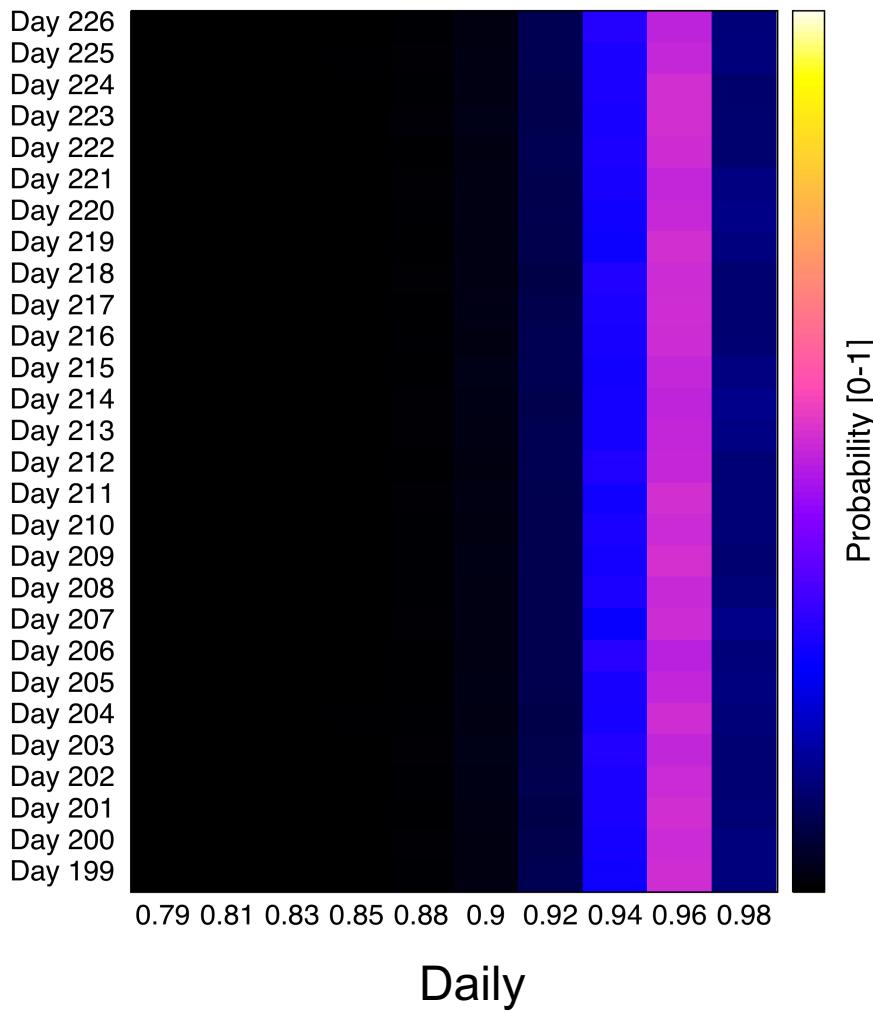
How rewards can be fairly distributed given citizens' selections?

Empirical Summarization Values – Smart Grid

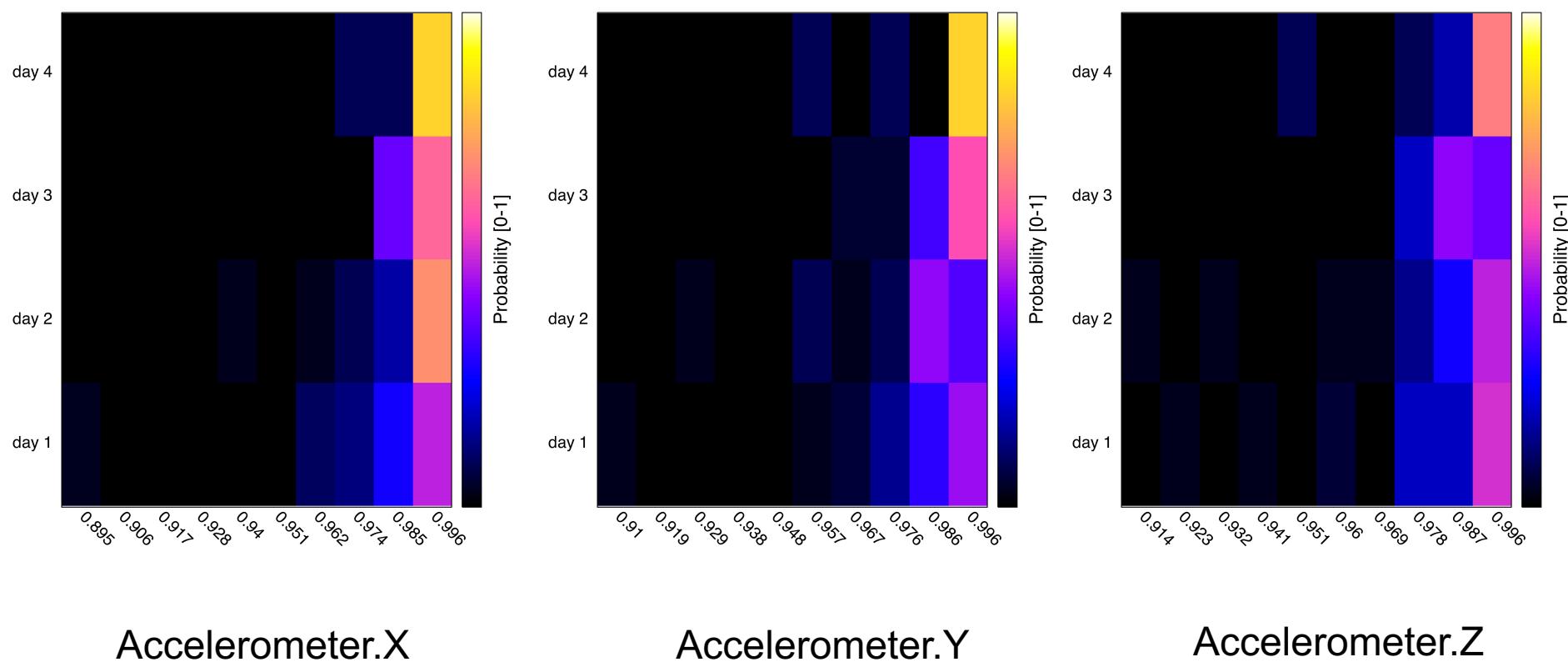


Daily summarization

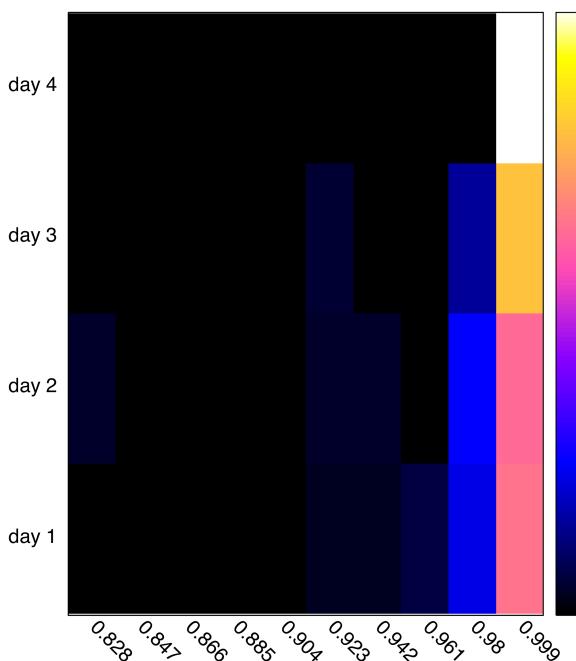
Algorithmic Summarization Values – Smart Grid



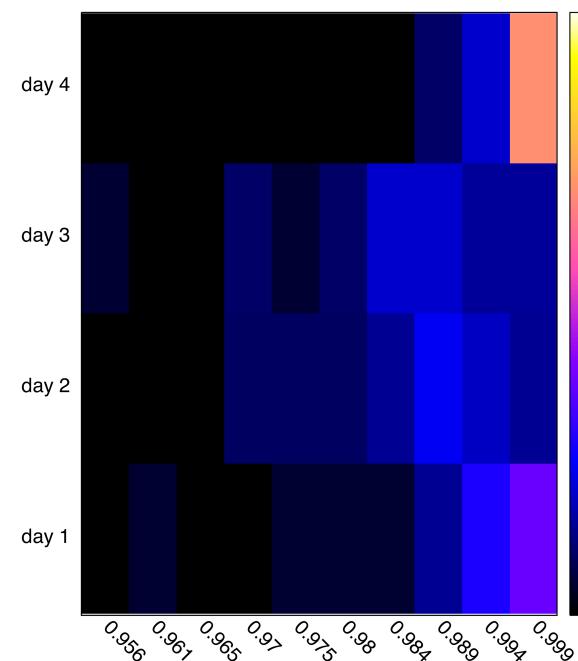
Algorithmic Summarization Values - Nervousnet



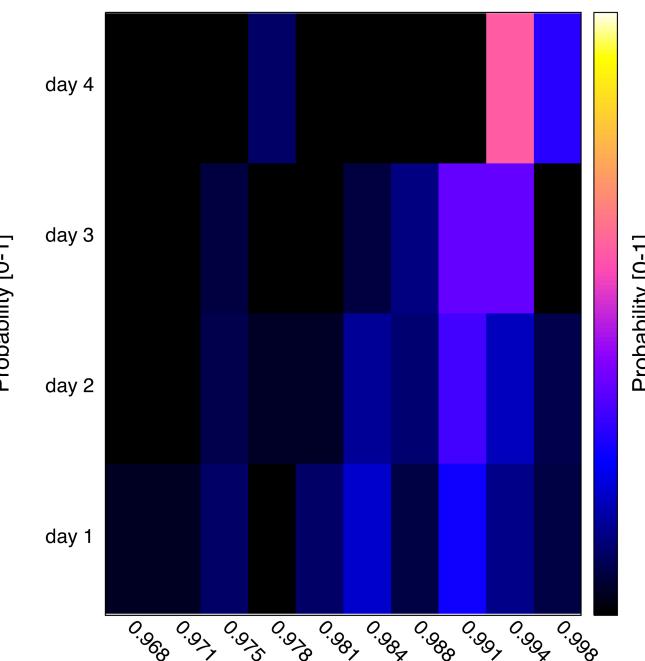
Algorithmic Summarization Values - Nervousnet



Battery



Light

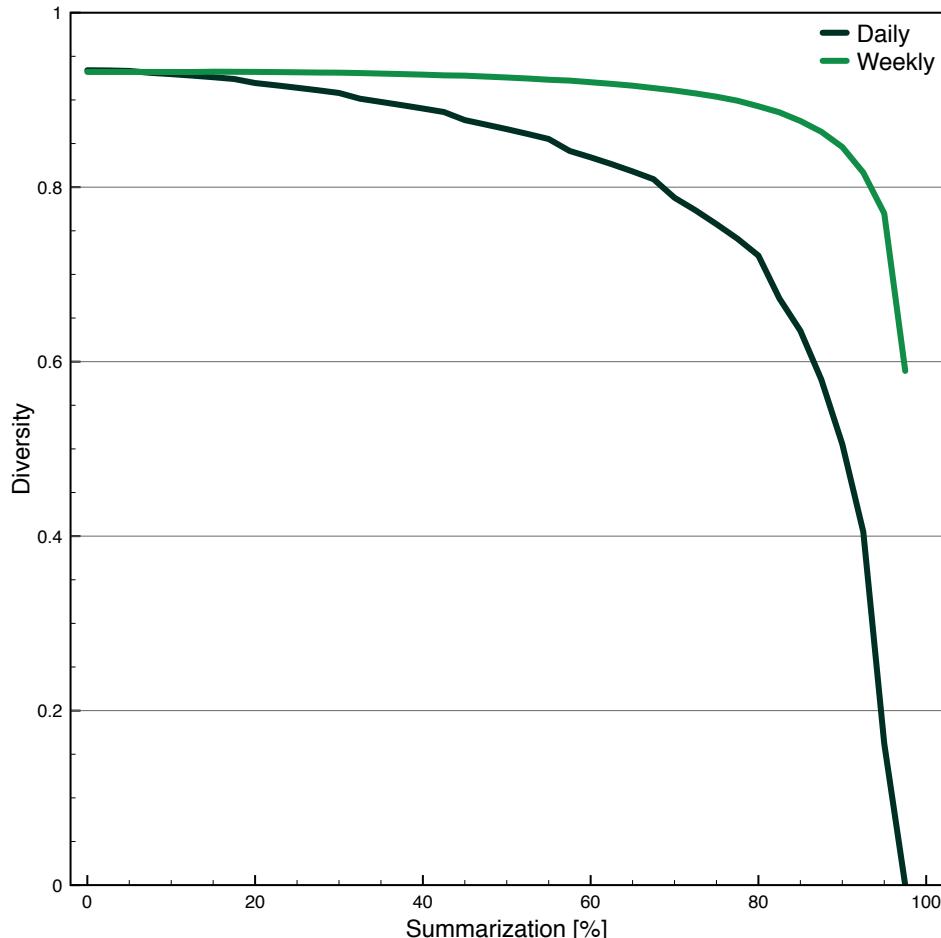
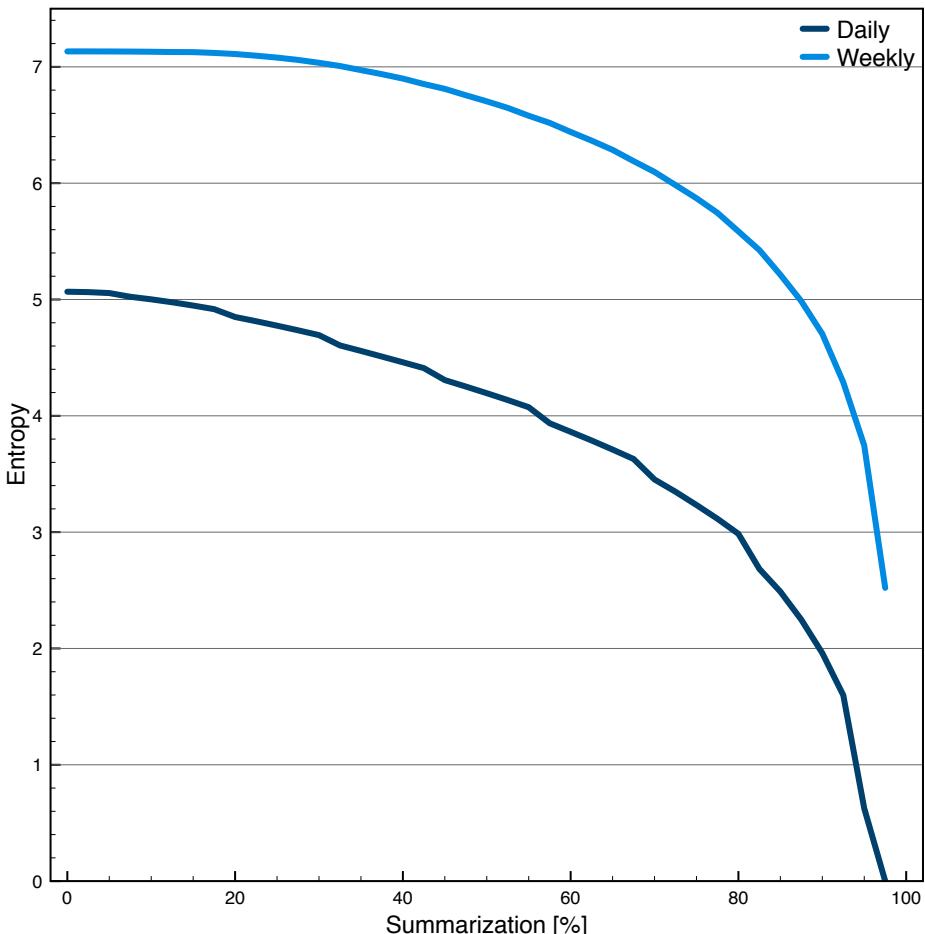


Noise

Does summarization improve privacy?

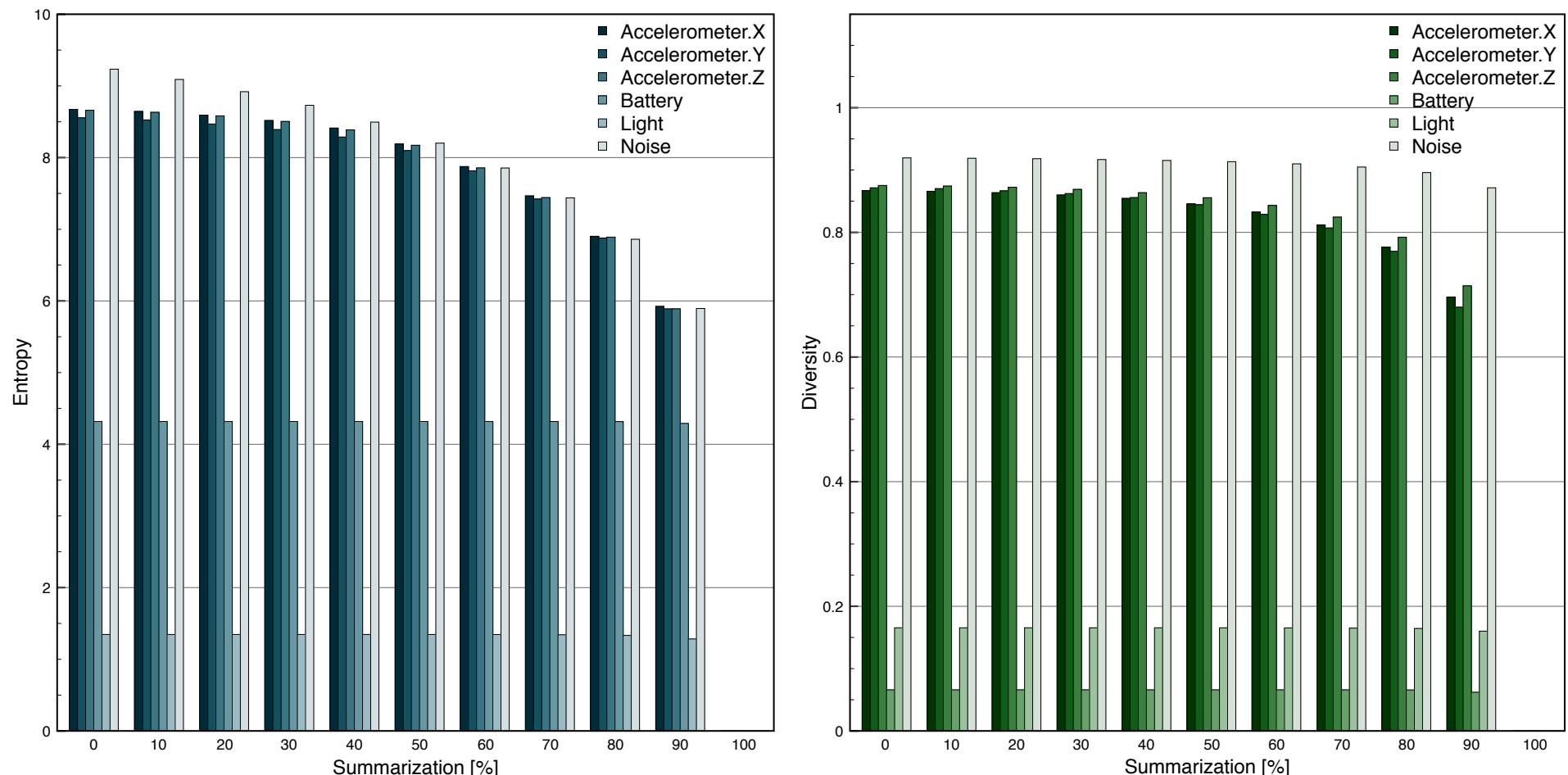
Does sensor/information type
influence these trade-offs?

Privacy-preservation – Smart Grid



Fixed summarization levels

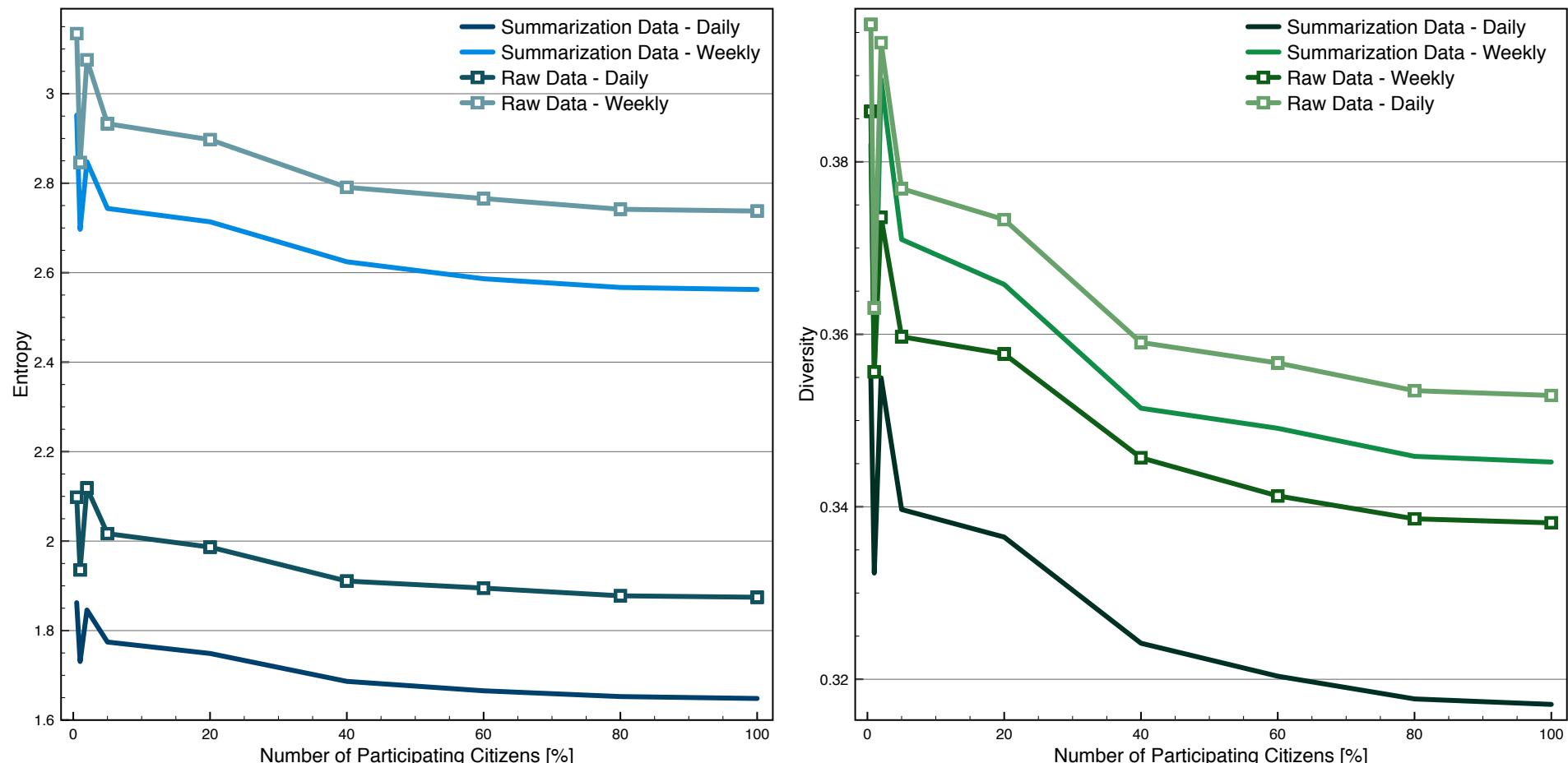
Privacy-preservation – Nervousnet



Fixed summarization levels

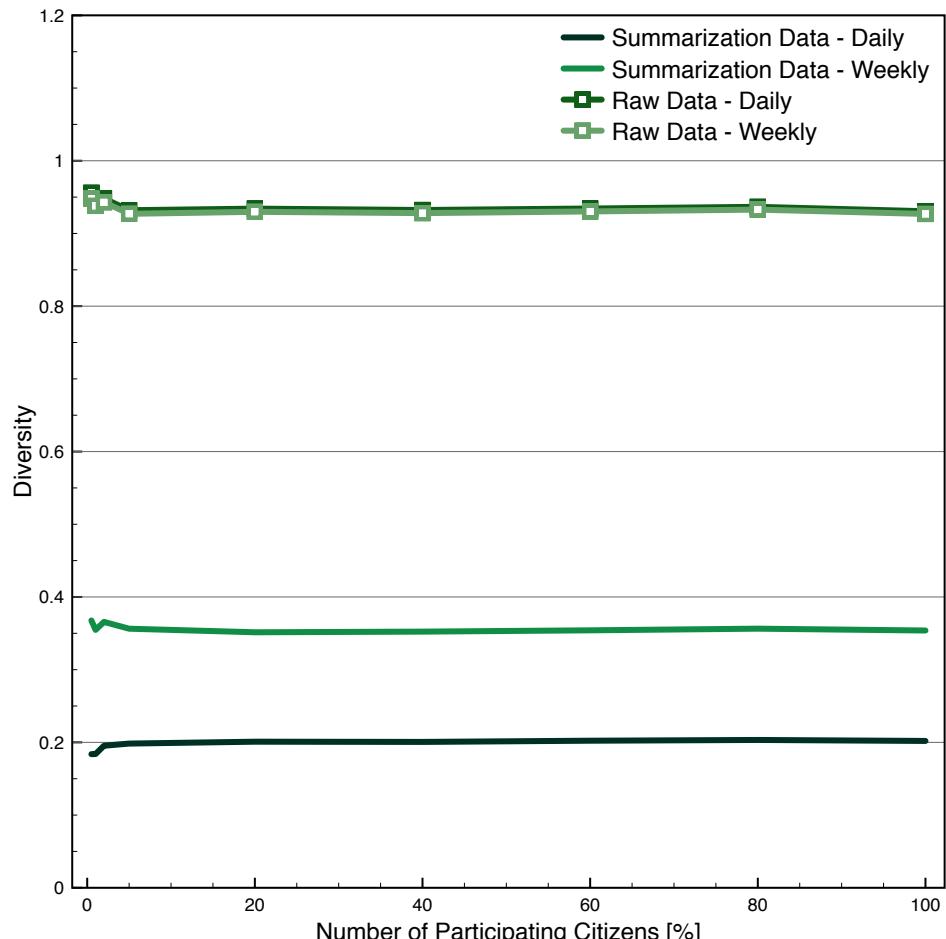
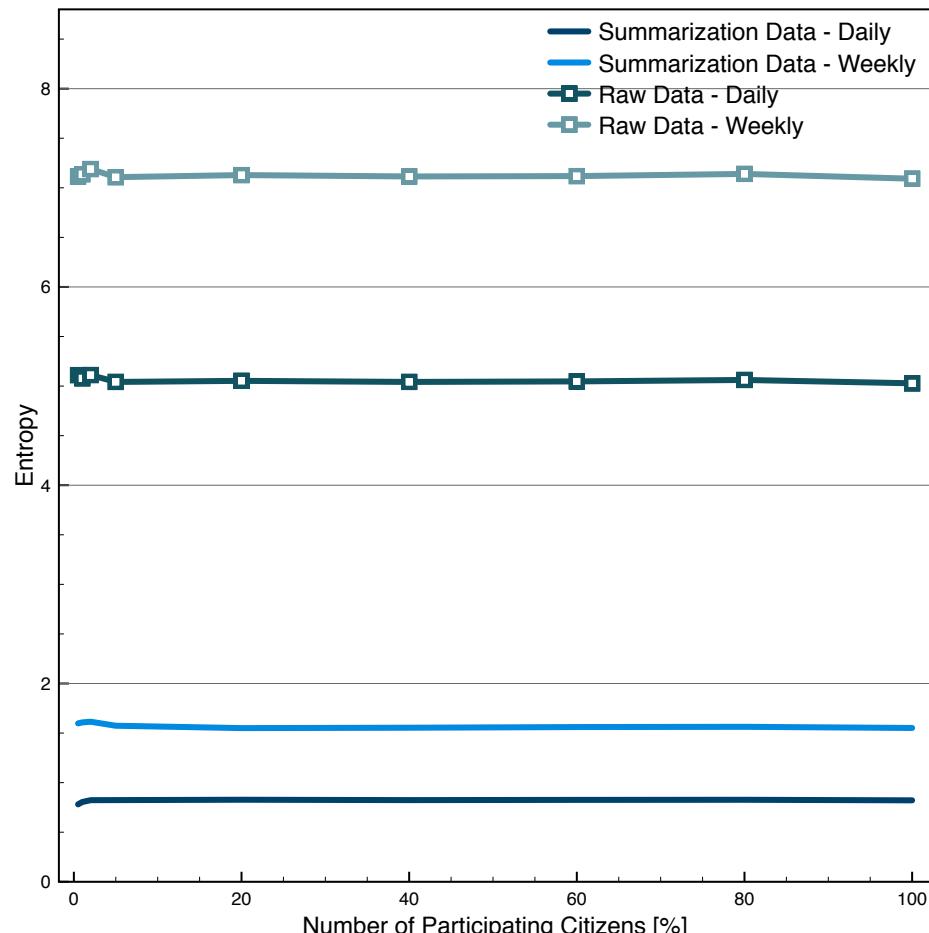
How does participation level influence privacy?

Privacy-preservation – Smart Grid



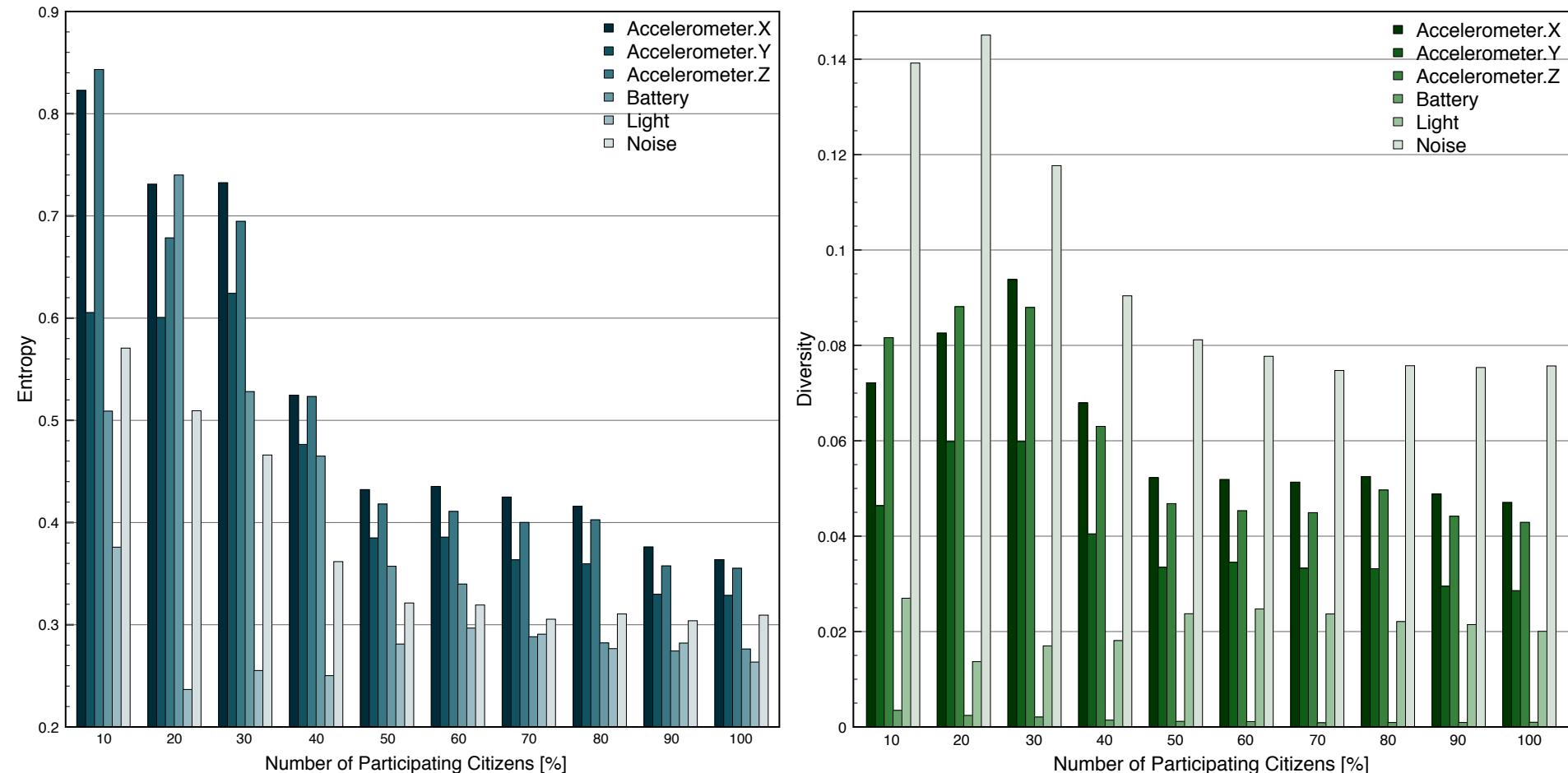
Empirical summarization levels

Privacy-preservation – Smart Grid



Algorithmic summarization levels

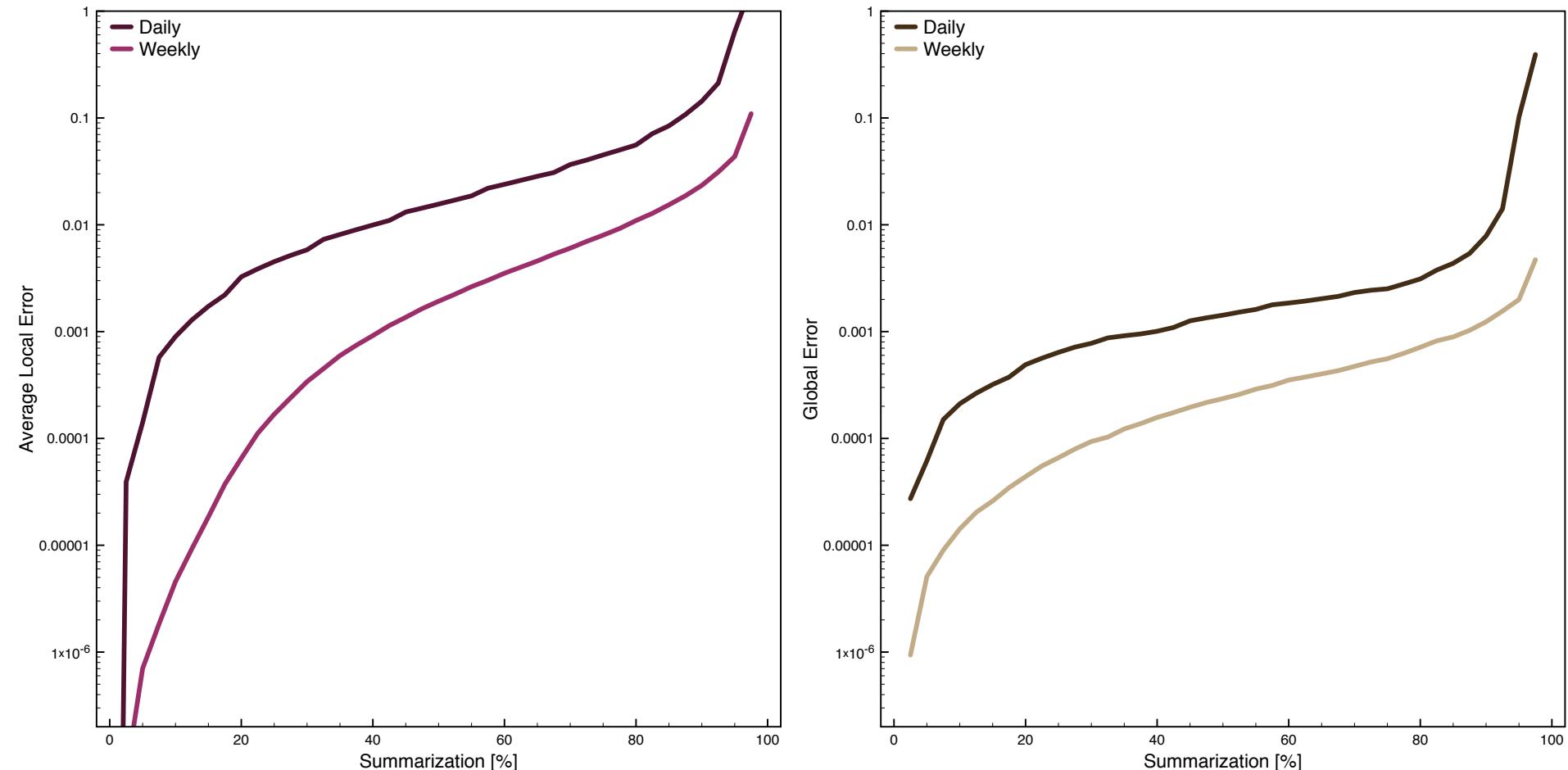
Privacy-preservation – Nervousnet



Algorithmic summarization levels

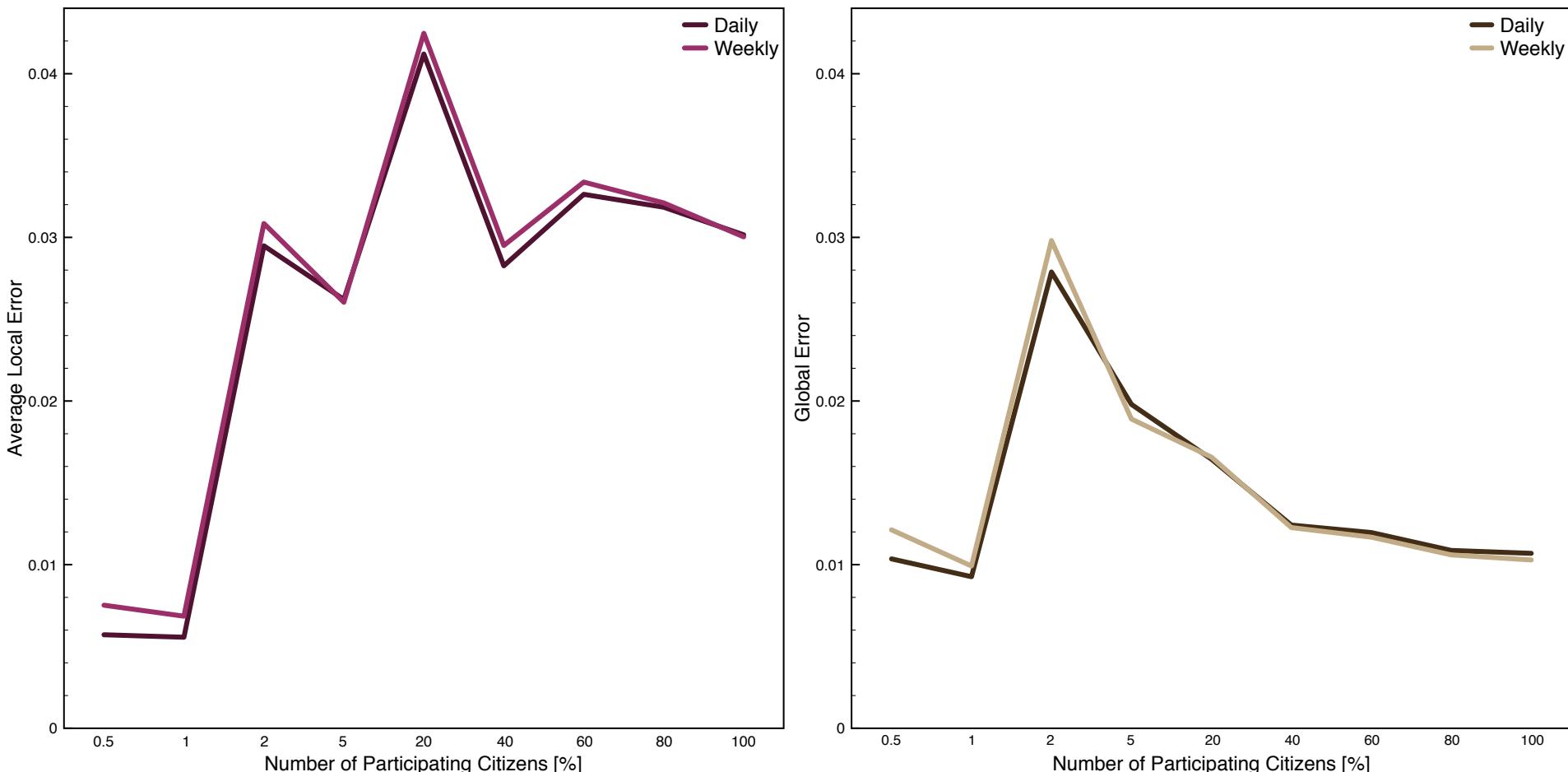
Which are the trade-offs between
privacy & accuracy in analytics?

Privacy vs. Accuracy – Smart Grid



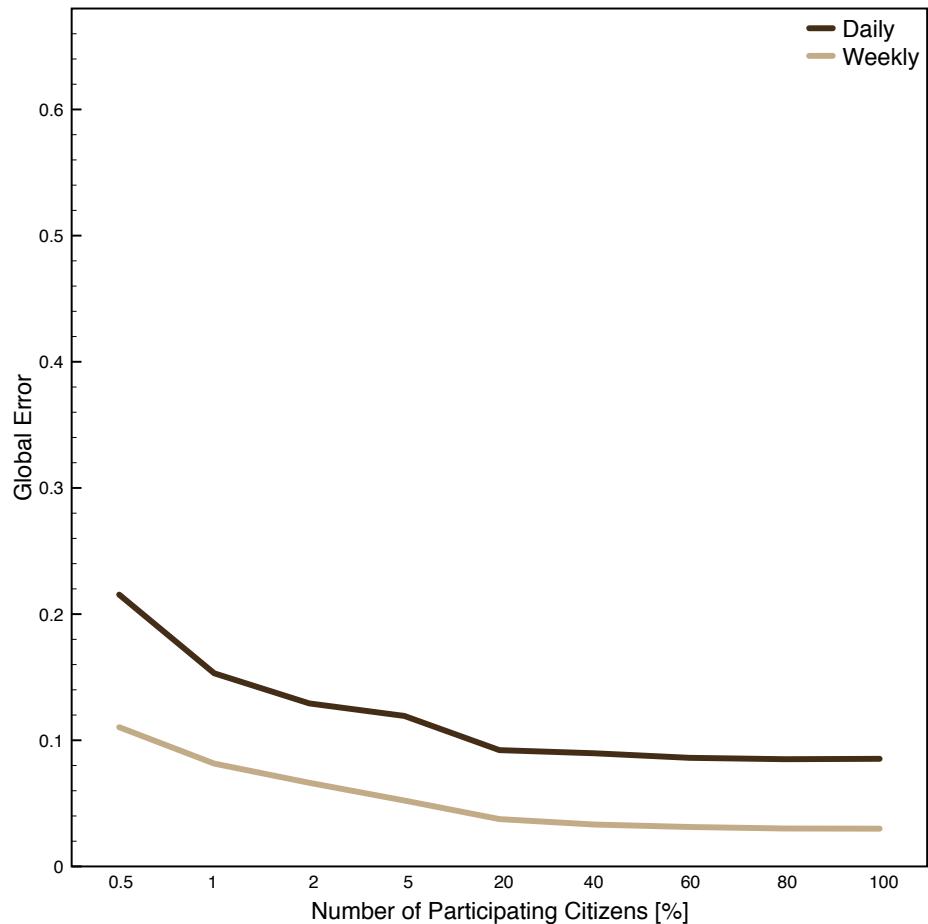
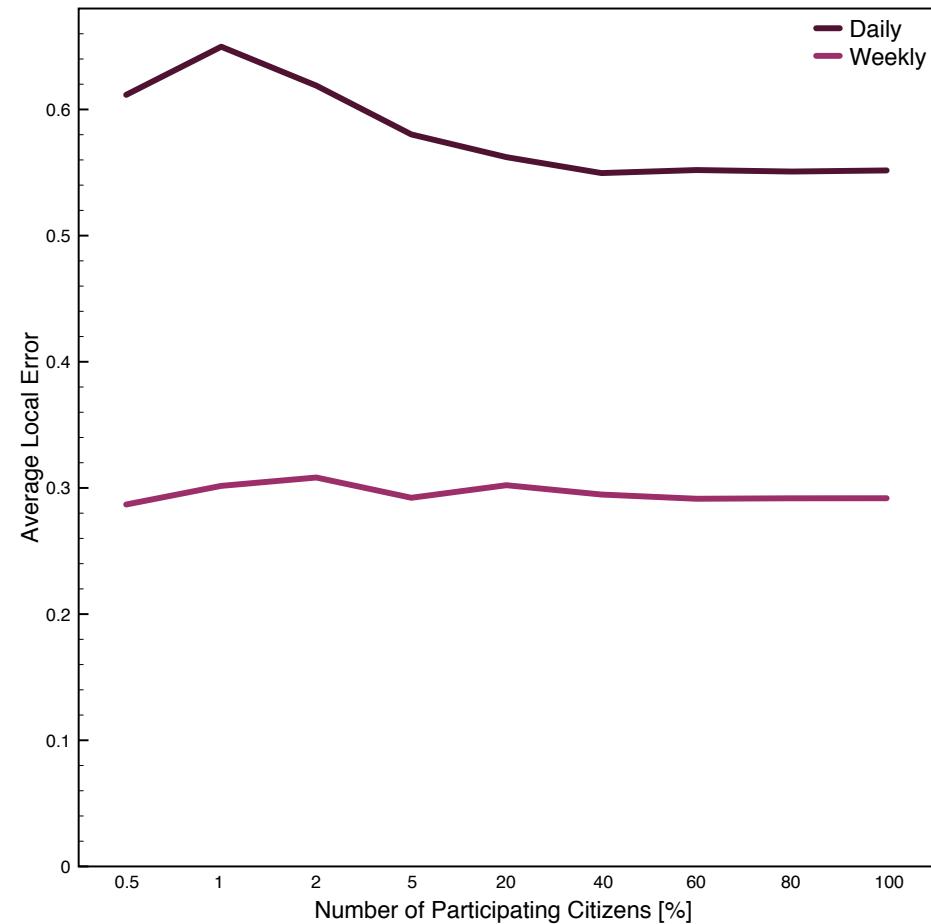
Fixed summarization levels

Privacy vs. Accuracy – Smart Grid



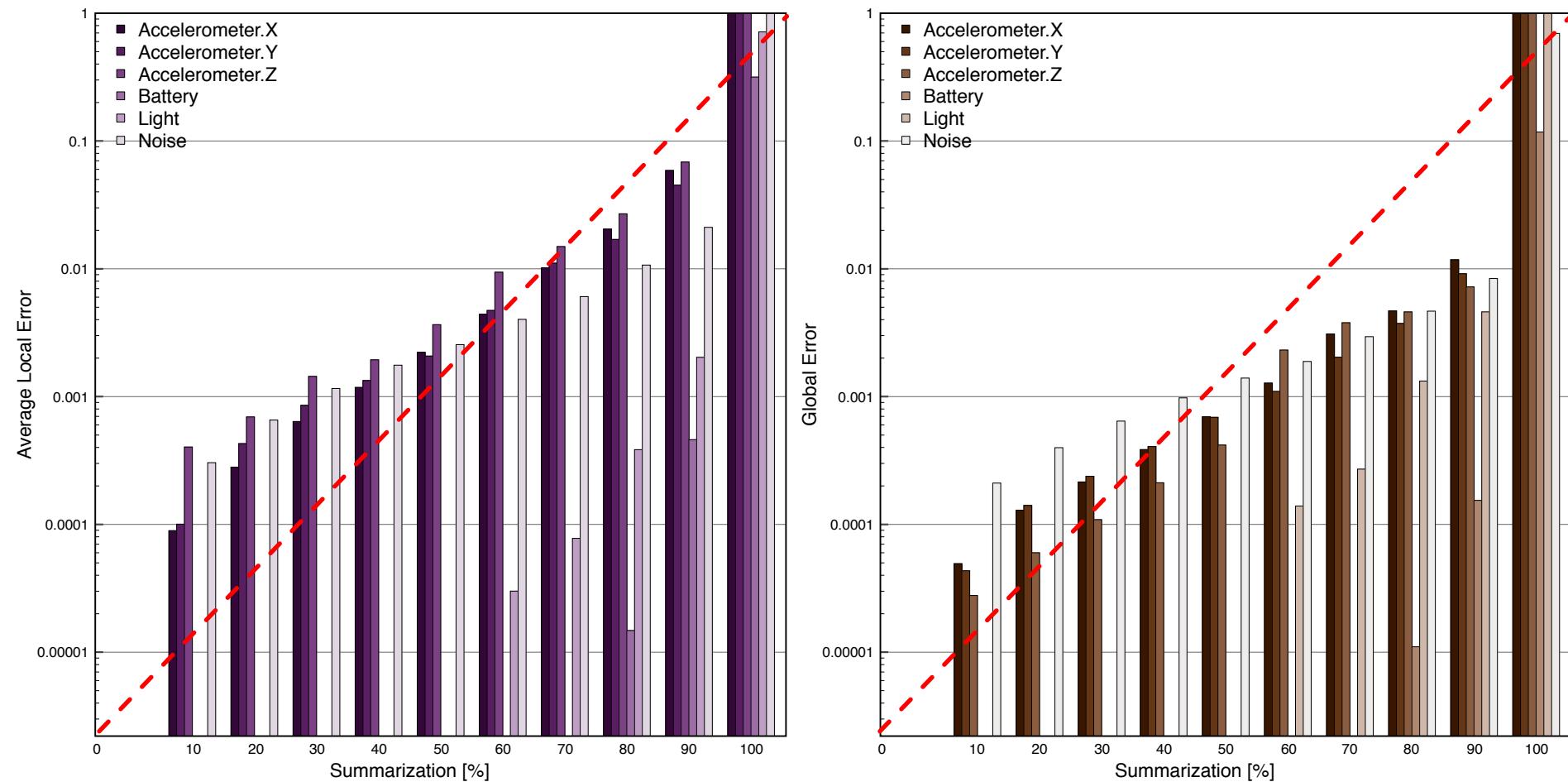
Empirical summarization levels

Privacy vs. Accuracy – Smart Grid



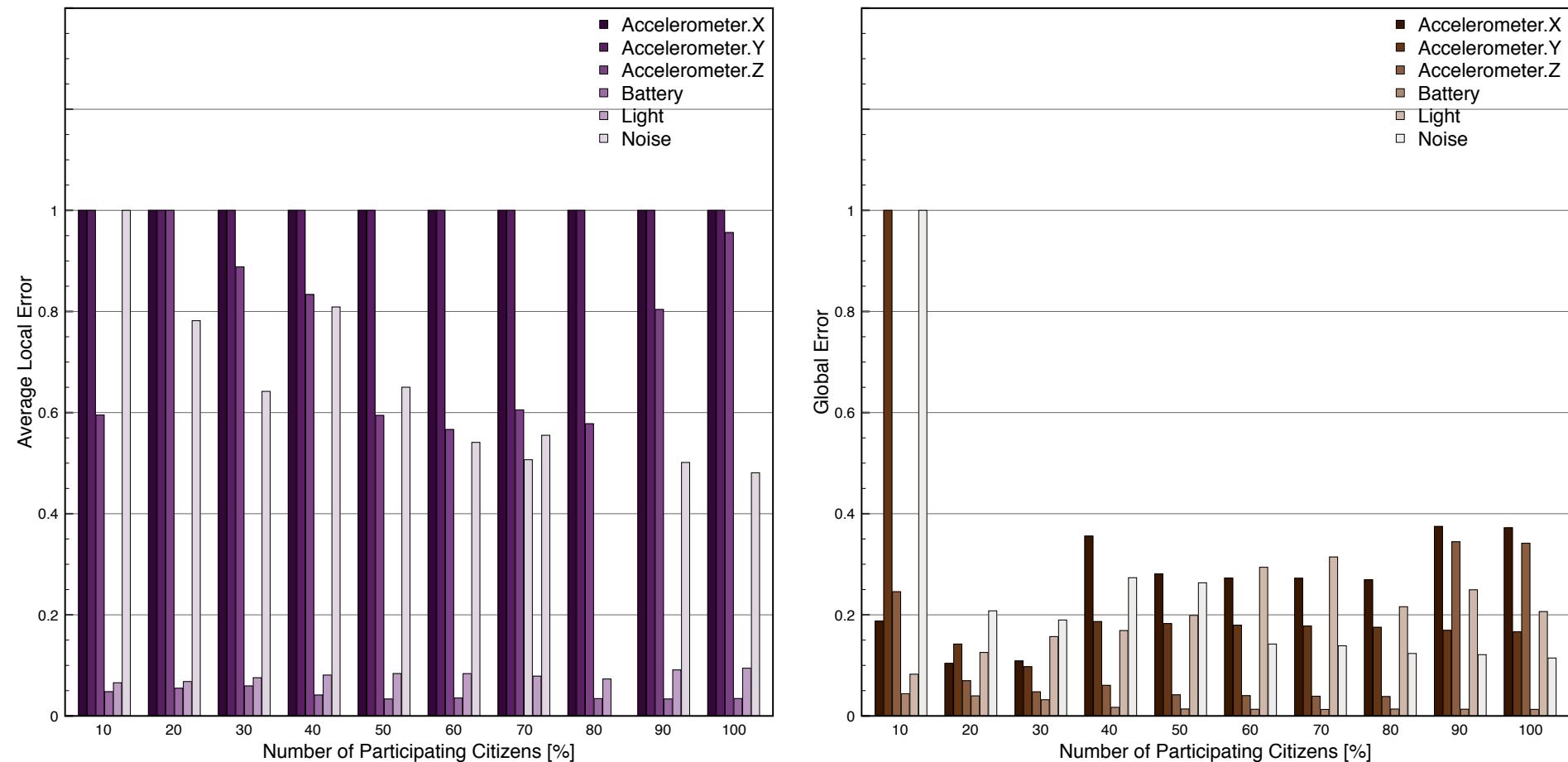
Algorithmic summarization levels

Privacy vs. Accuracy – Nervousnet



Fixed summarization levels

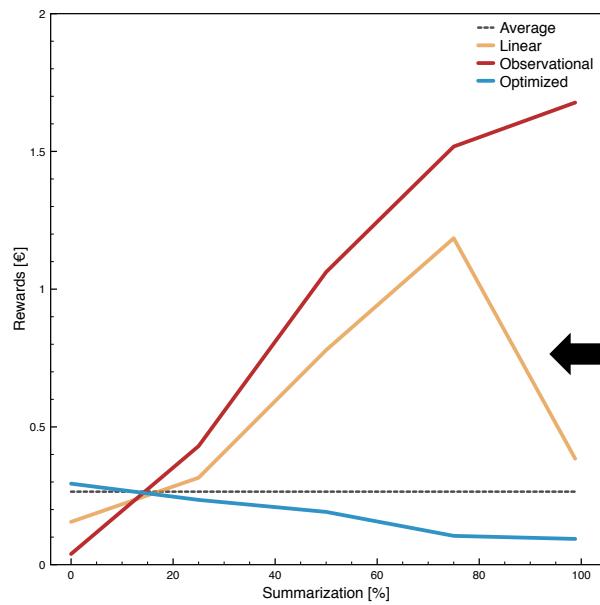
Privacy vs. Accuracy – Nervousnet



Algorithmic summarization levels

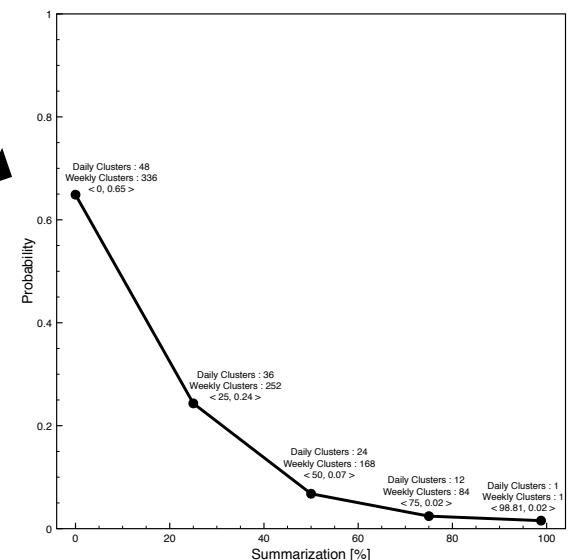
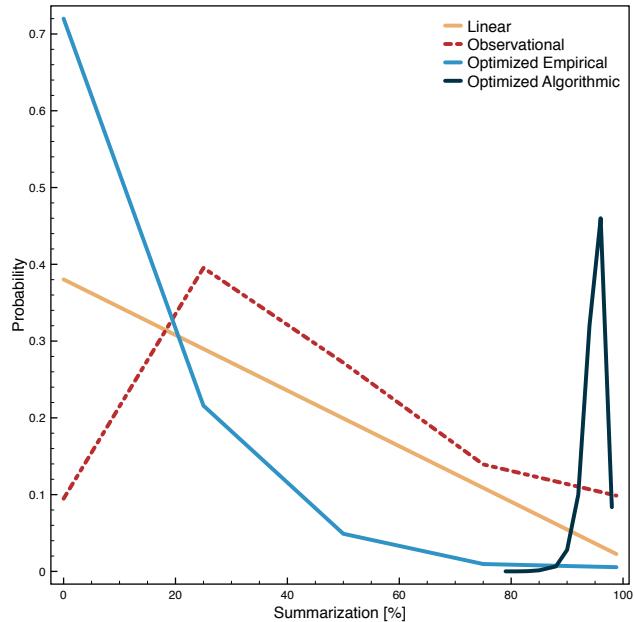
How rewards can be fairly distributed
given citizens' selections?

Rewards – Smart Grid

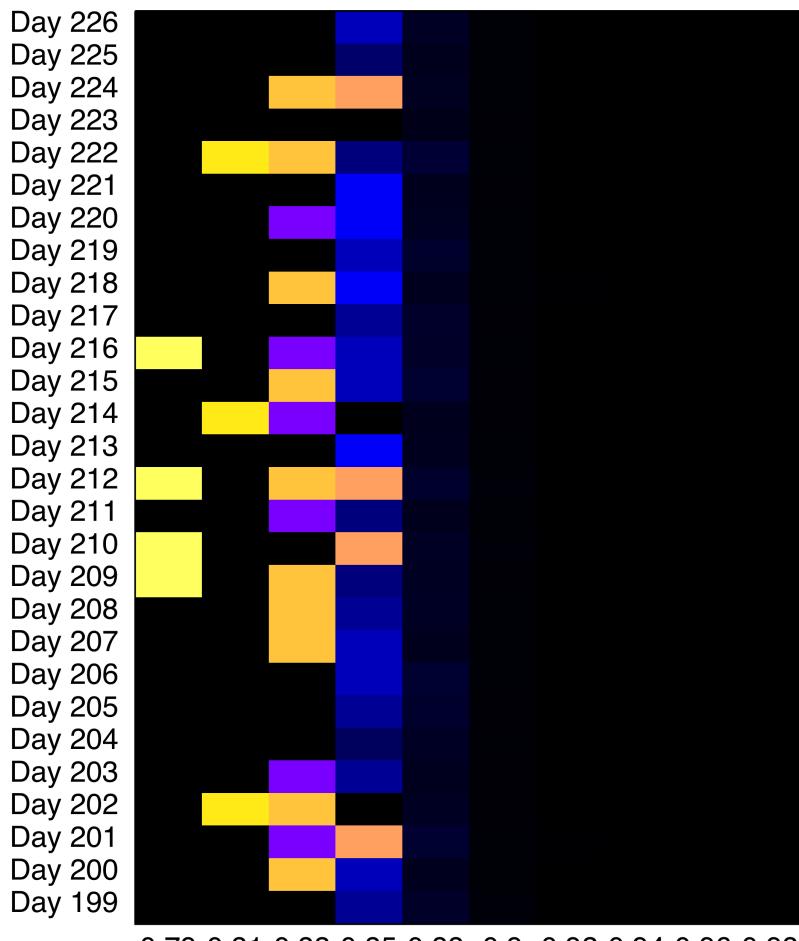


$$\gamma_{i,e} = \frac{\gamma_e * P_r(\alpha_{i,e})}{n * P_s(\alpha_{i,e})}$$

Total budget Number of citizens

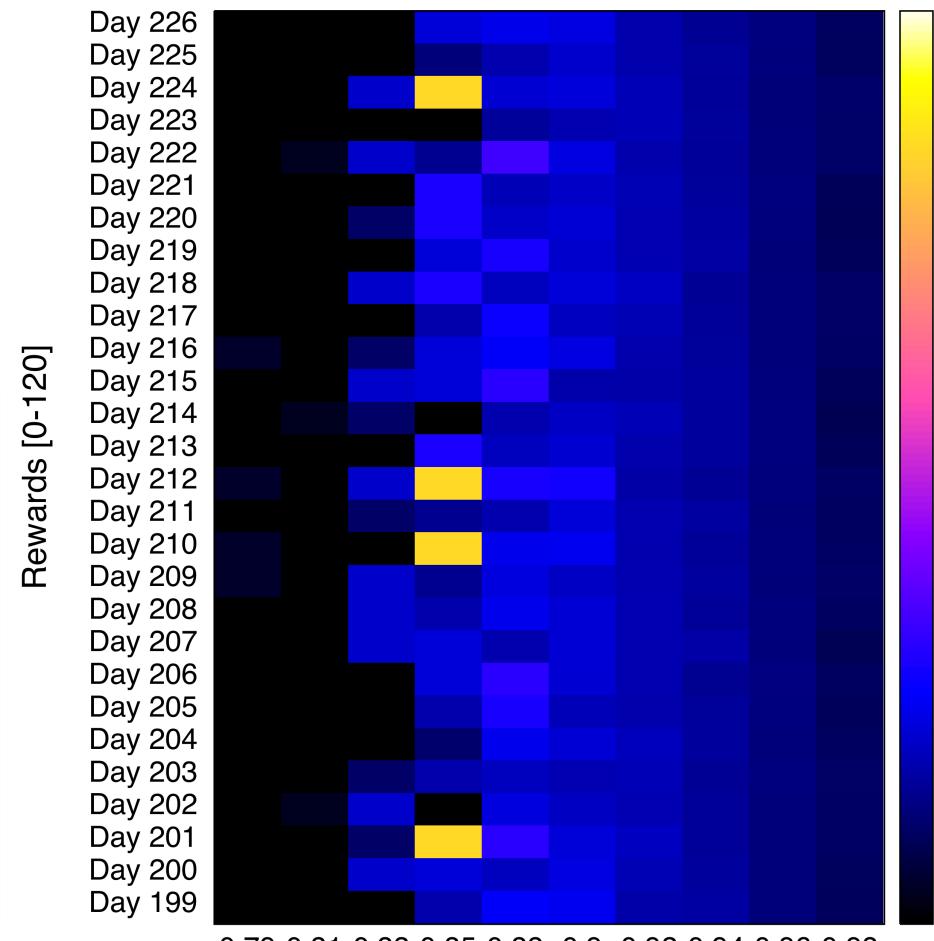


Rewards – Smart Grid



Linear

Summarization



Optimized

Rewards [0-2]

Rewards [0-120]

Conclusions

Higher summarization, higher privacy-preservation

More participants, higher privacy-preservation

Sensor types influence privacy-preservation & accuracy

Local errors cancel out resulting in low global errors

Incentivization can be optimized to be fair

Looking ahead

Improve performance further

Adapt to different types of data

Robust to inference

An ecosystem of summarization algorithms!

<http://www.nervousnet.ethz.ch/hackathon/>

HACKATHON

Making it easy!

Tutorials, code utilities and more info

<https://github.com/epournaras/Nervousnet-Hackathon-Challenge>

Submission

Select your .zip file:



output.zip

 output.zip  Remove  Browse ...

Submit

NervousnetTeam

Submissions

Your team has submitted 2 times.

Submission	Submitted	State	Entropy	Diversity	Avg Local Error	Global Error
1	2016-04-01T10:18:41.406Z	Analysed	1.2269450302542217	0.2746648936170221	0.3617792470281442	0.035184004592273194
2	2016-04-01T10:19:55.337Z	Analysed	2.9050783804260867	0.6848636018237065	0.06689308456640566	0.0054395033494120245

Questions?

Evangelos Pournaras, Jovan Nikolic, Pablo Velasquez, Marcello Trovati, Nik Bessis and Dirk Helbing, *Self-regulatory Information Sharing in Participatory Social Sensing*, The European Physical Journal Data Science, 5:14, 2016

www.evangelospournaras.com

epournaras@ethz.ch