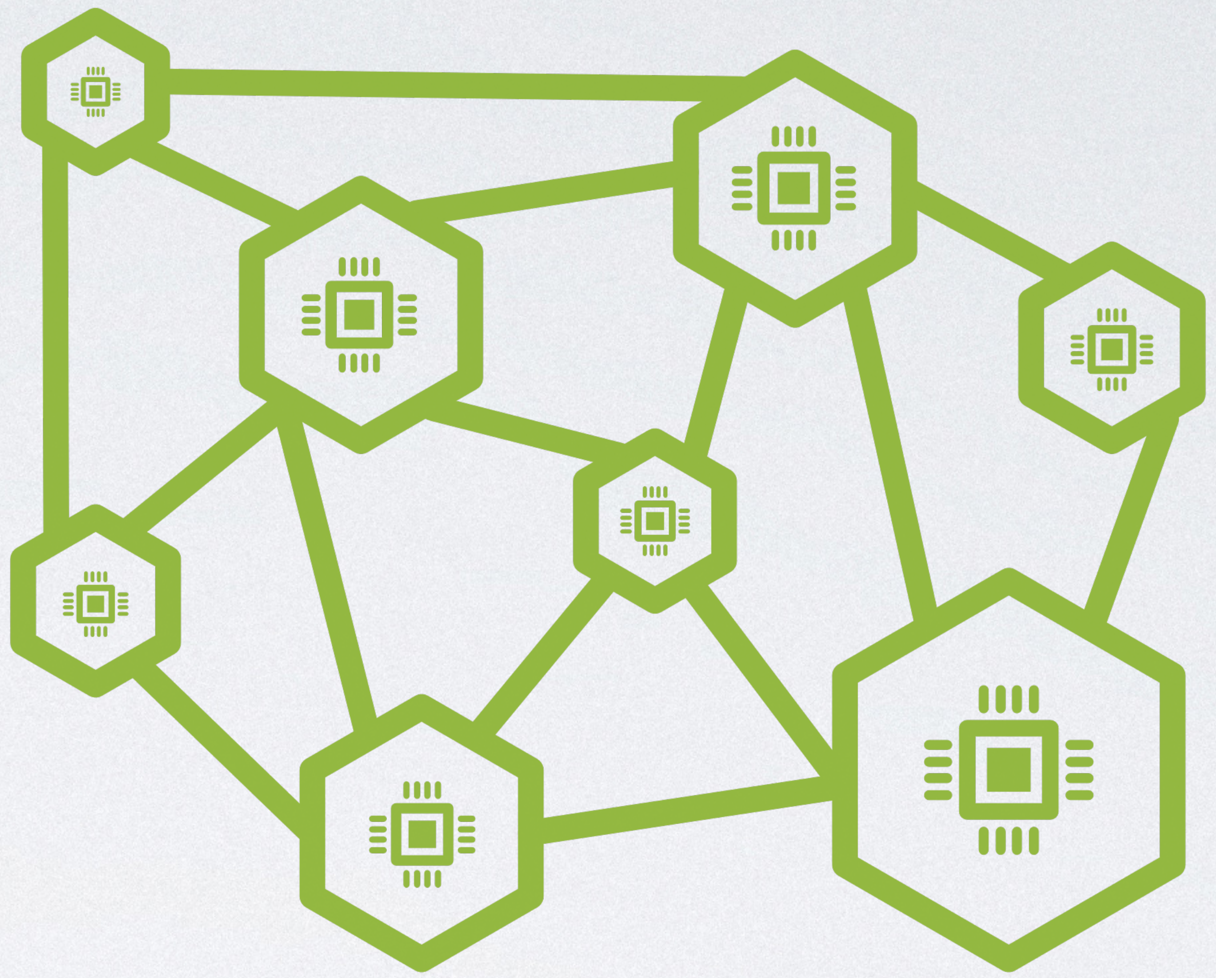
DIFFERENTIALLY PRIVATE LEARNING FROM LABEL PROPORTIONS (DP-LLP)

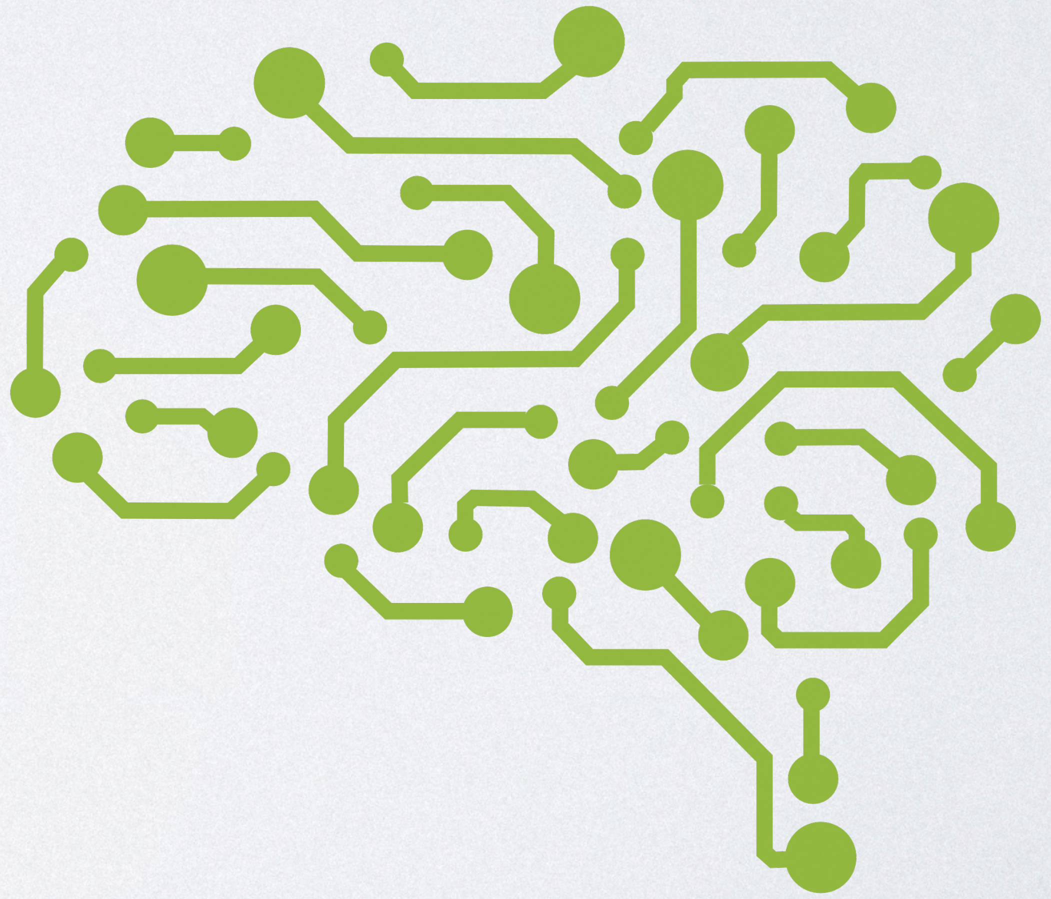
For Privacy Preserving Route Planning

Timon Sachweh

AGENDA

- Motivation
- State of Research
- DP-LLP
- Evaluation Results / Comparison
- Conclusion

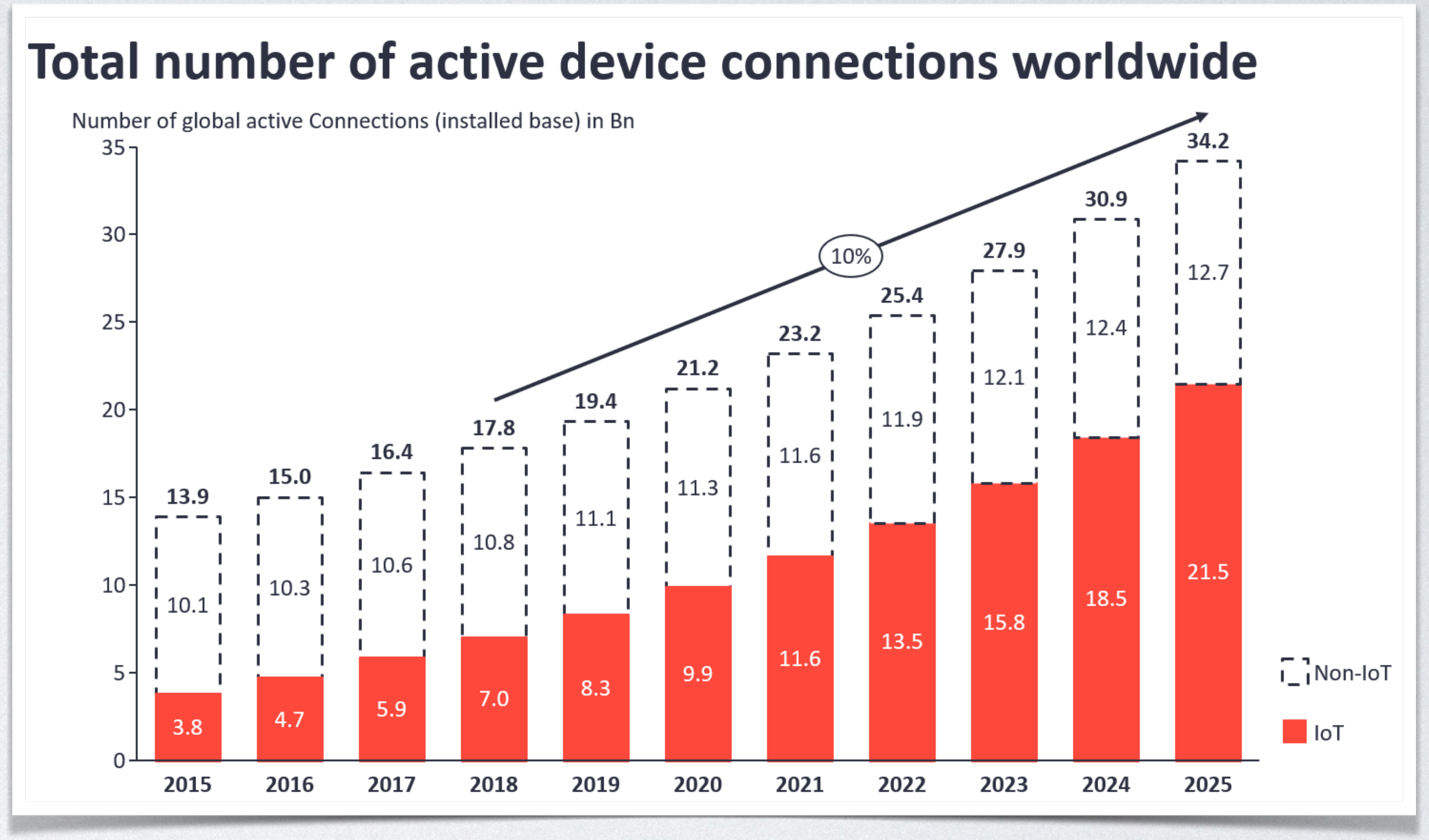




MOTIVATION

MOTIVATION

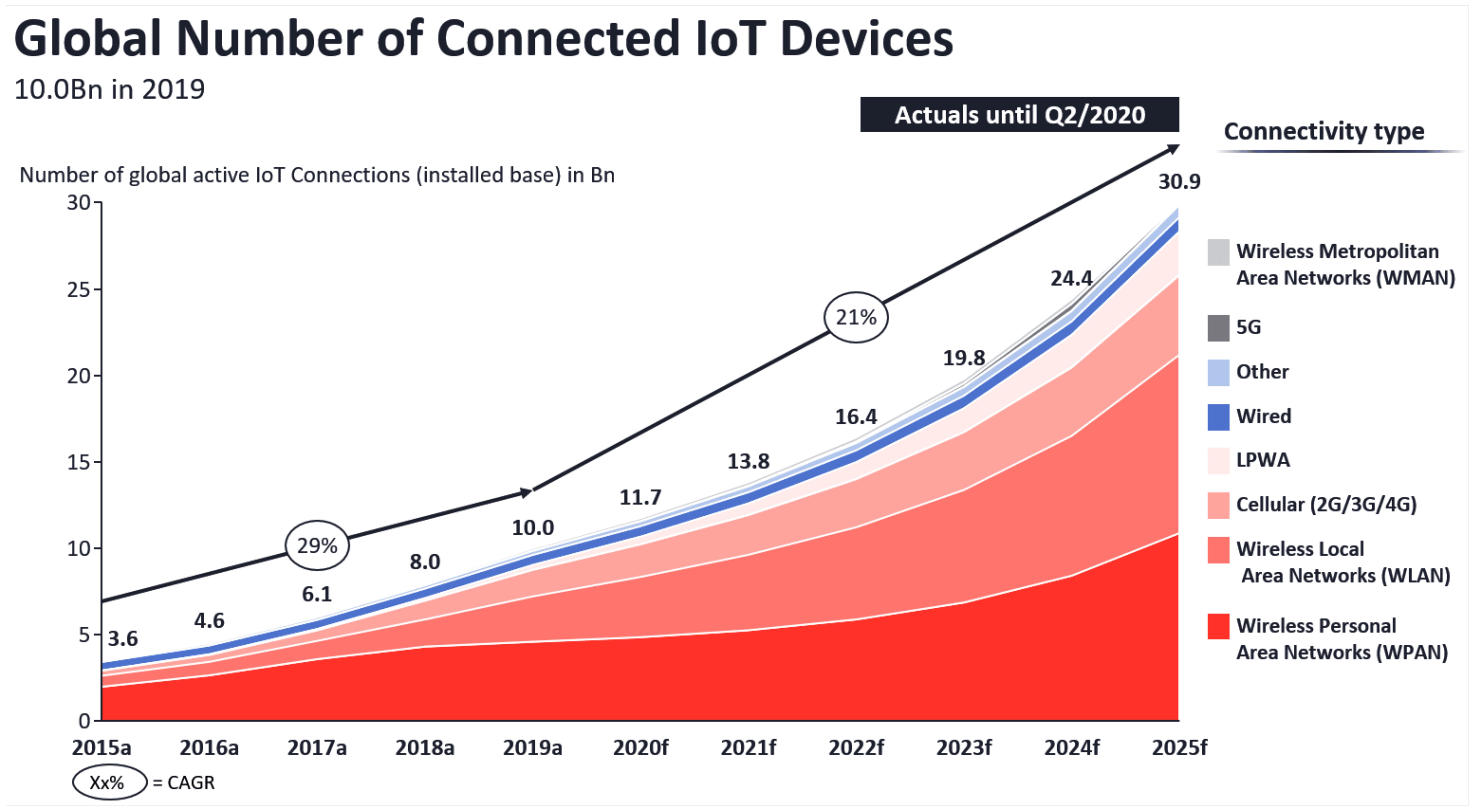
- cross-linking and internet of things ➝ more and more data volume and flow
- decentralised collection of data
- high requirements on data privacy ➝ GDPR
  - data exploitation for business models requires resolving the conflict between data use and data protection
- **goal:** enable data use while ensuring data protection



Source(s): IoT Analytics - Cellular IoT&LPWA Connectivity Market Tracker 2010-25

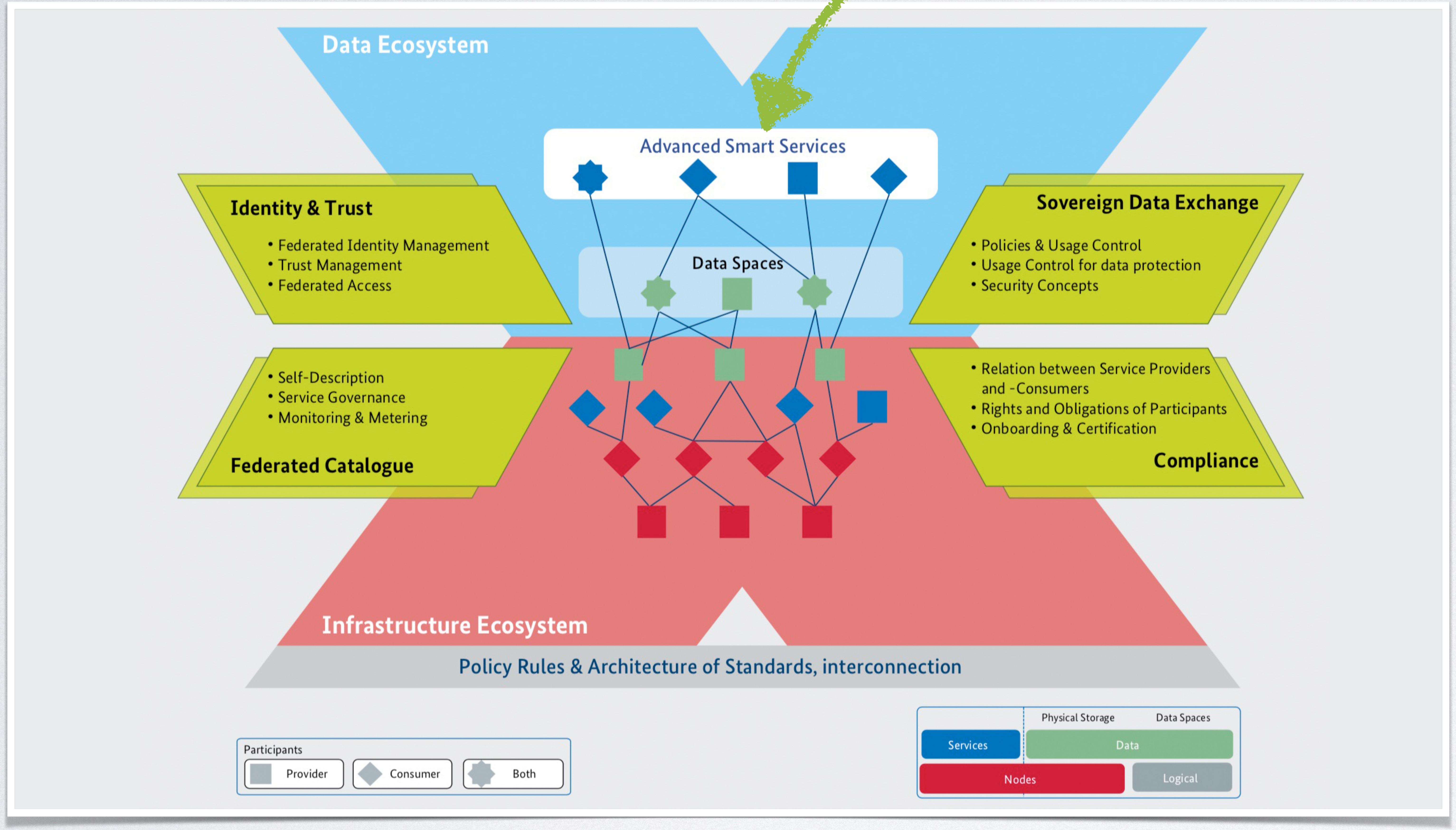
Source(s): https://www.jet-software.com/datenmaskierung/gdpr/

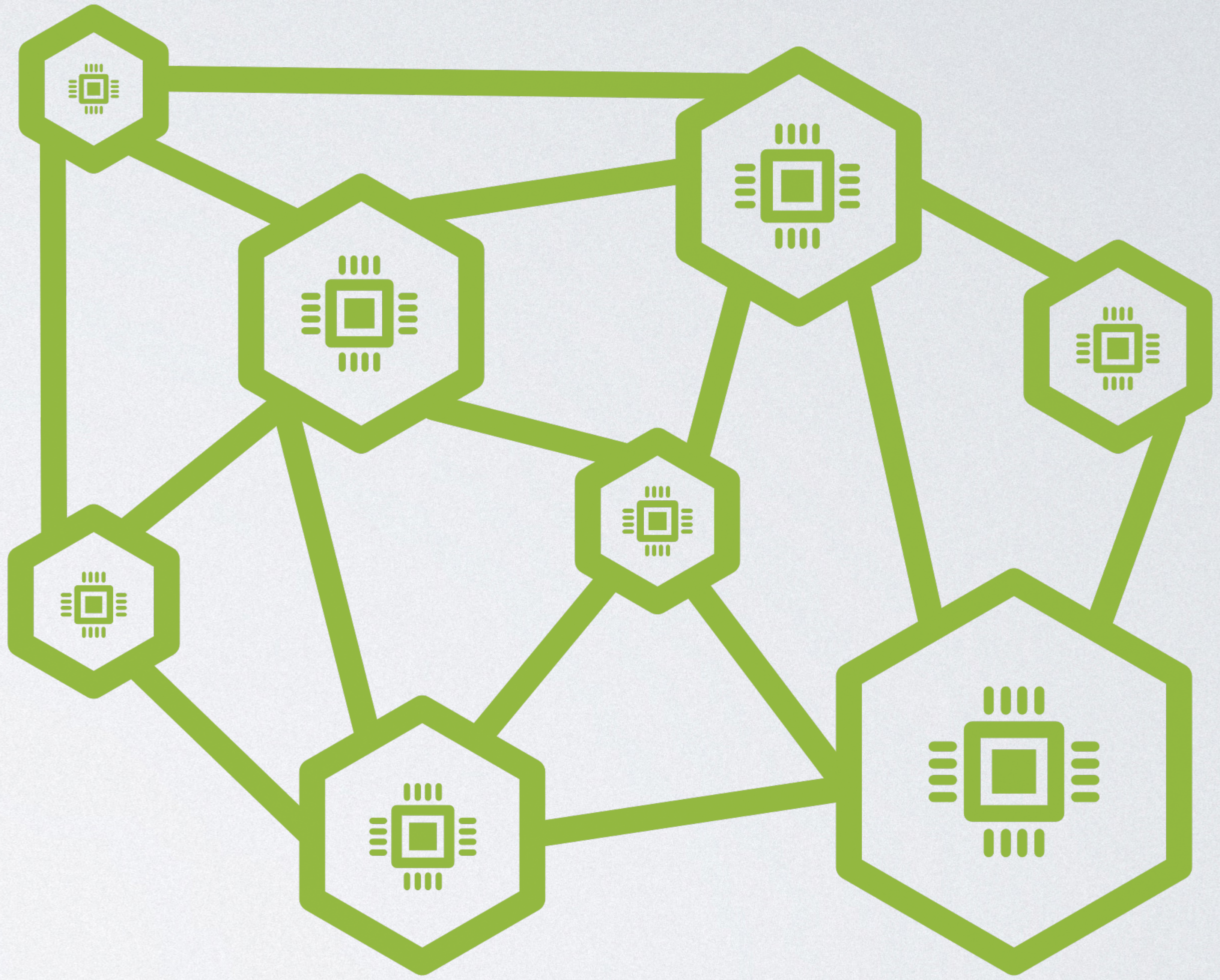
MOTIVATION



Source(s): IoT Analytics - Cellular IoT&LPWA Connectivity Market Tracker 2010-25

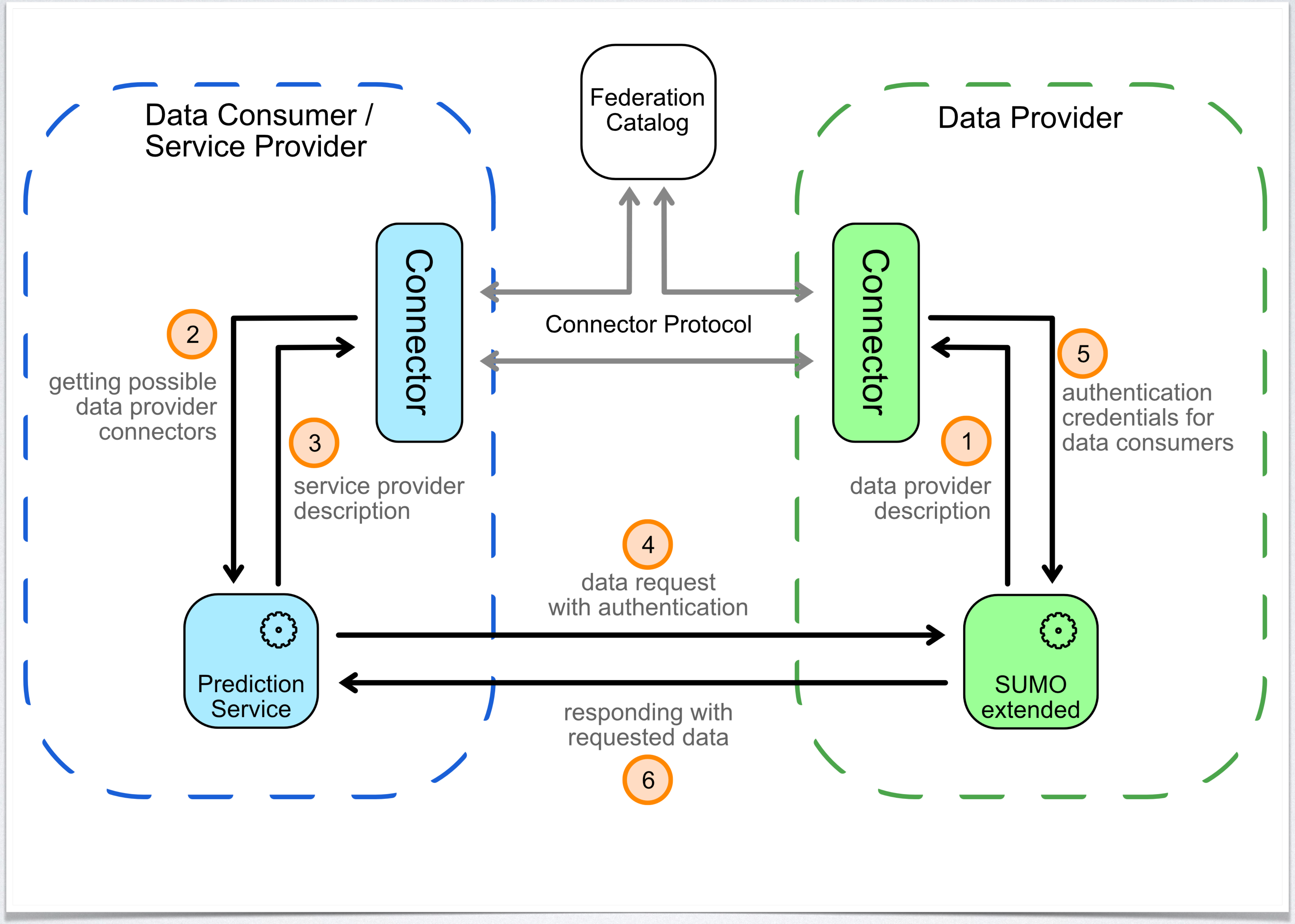
MOTIVATION





Source: https://www.heise.de/news/Bundeswirtschaftsminister-Gaia-X-als-weltweiter-Goldstandard-fuer-Cloud-Dienste-4774826.html

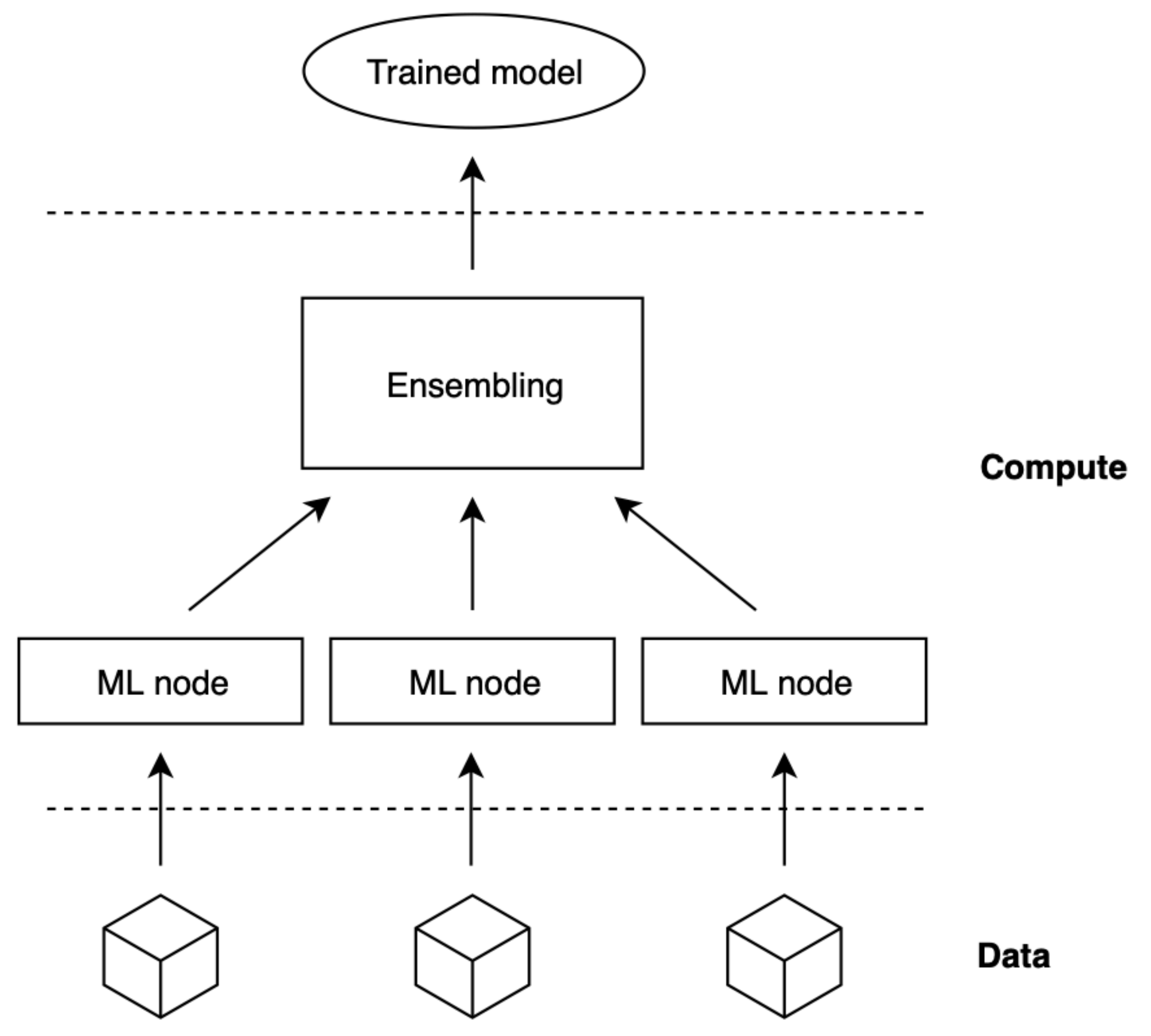
MOTIVATION



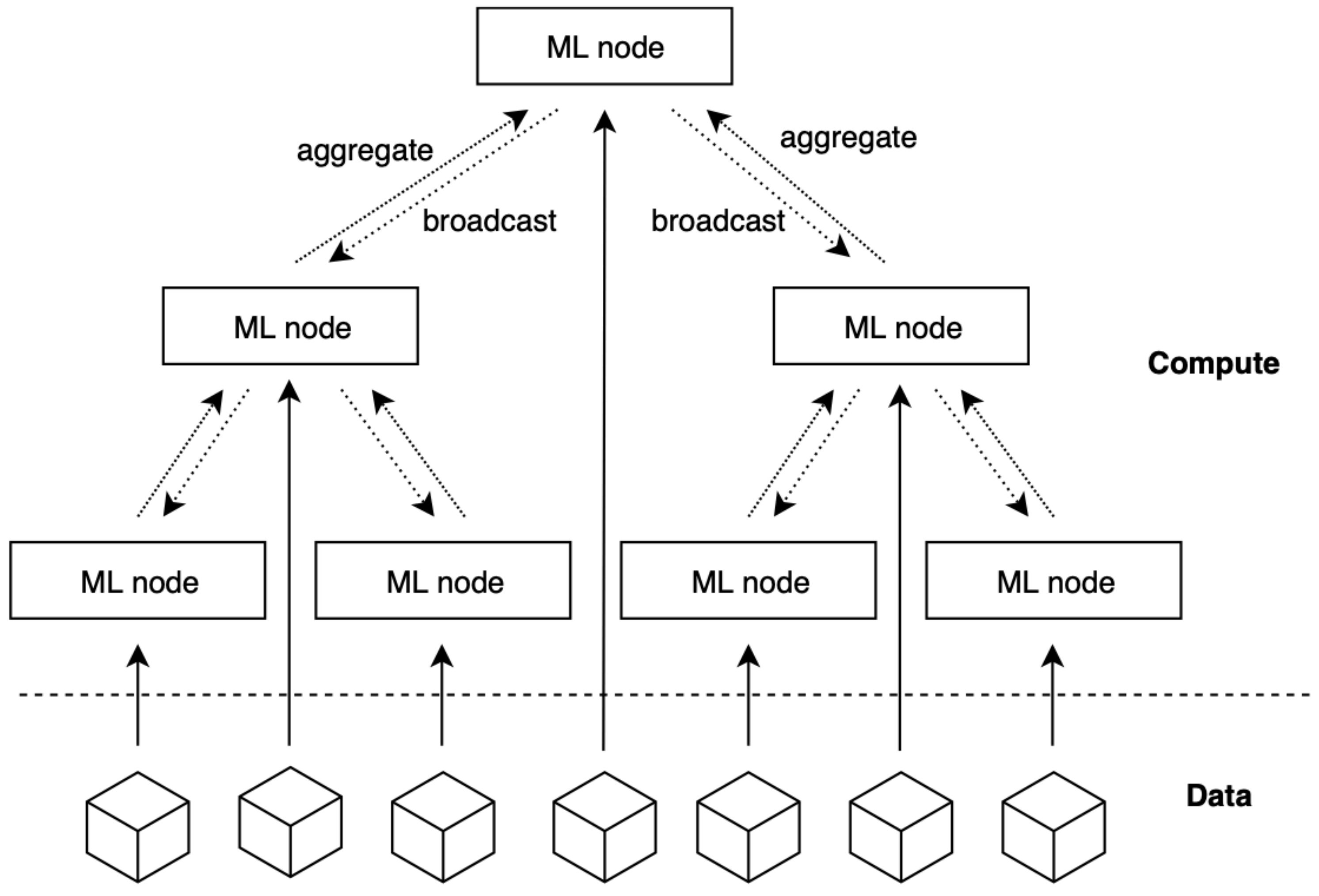
Federation Catalog
Data Consumer / Service Provider
Data Provider
Connector
Connector
Connector Protocol
2
getting possible data provider connectors
3
service provider description
5
authentication credentials for data consumers
1
data provider description
4
data request with authentication
Prediction Service
SUMO extended
responding with requested data
6

STATE OF RESEARCH

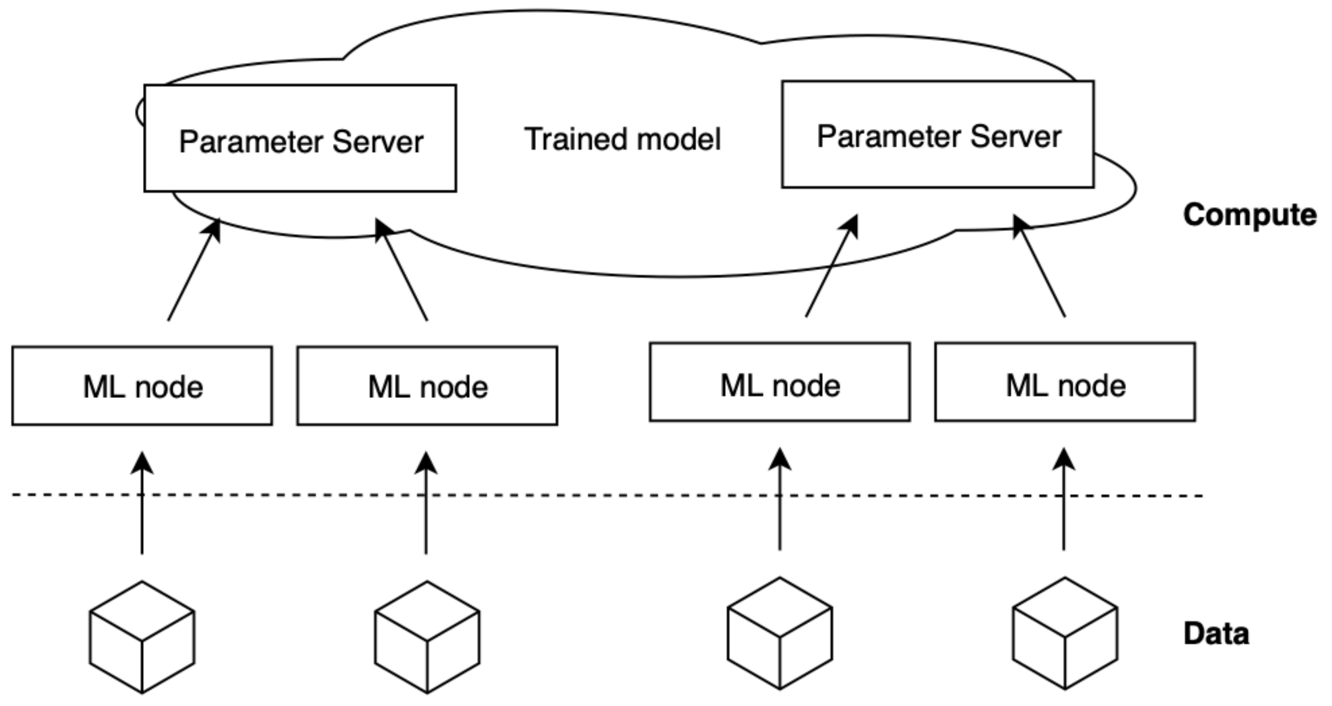
STATE OF RESEARCH



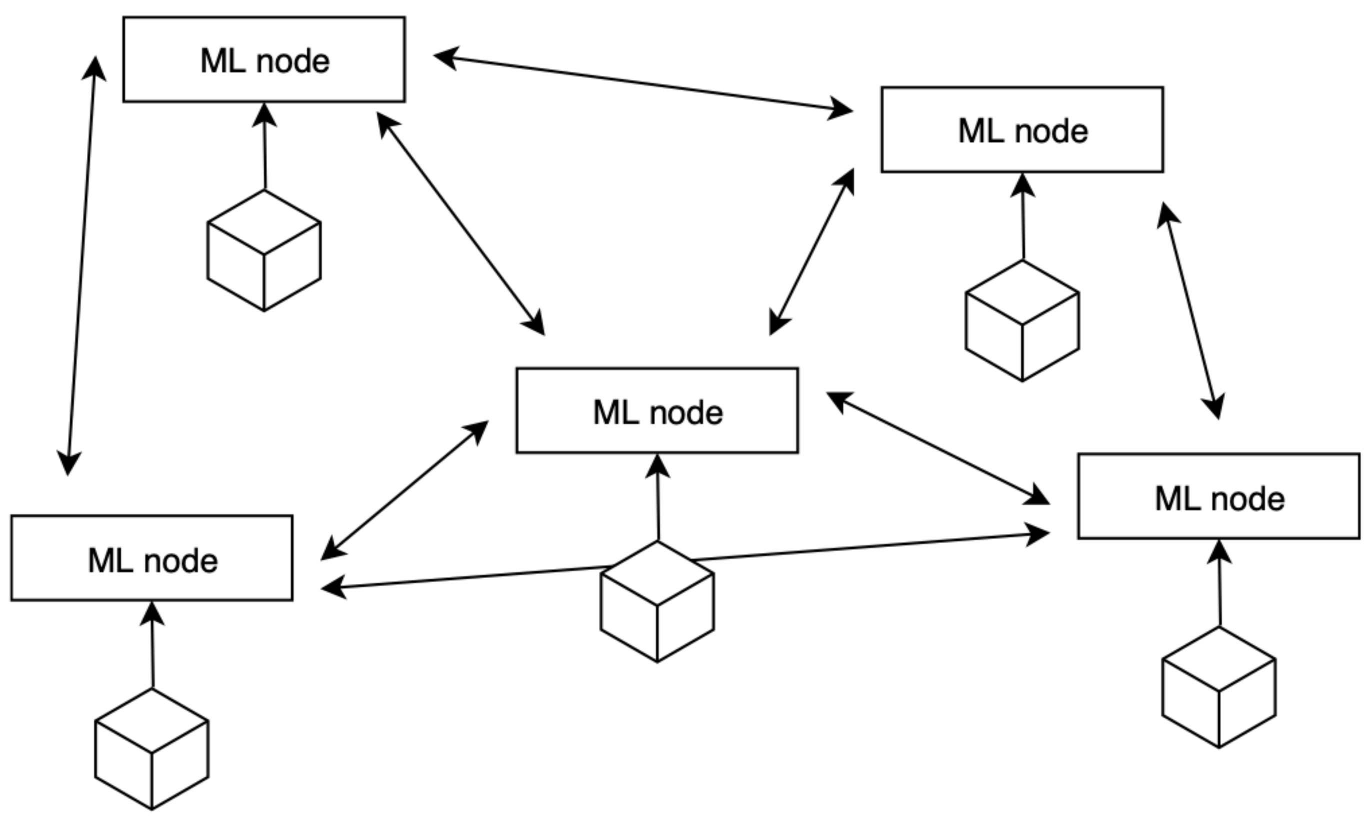
(a) Centralized (Ensembling)



(b) Decentralized (Tree)



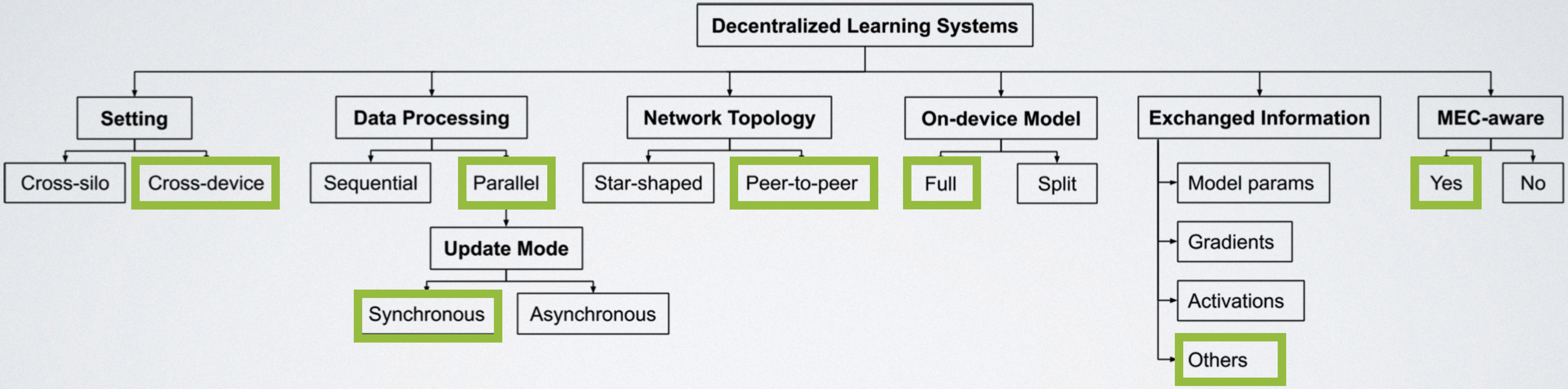
(c) Decentralized (Parameter Server)



(d) Fully Distributed (Peer to Peer)

[Verbraeken et. al., 2020]

STATE OF RESEARCH



Source: Bellavista, P., Foschini, L., & Mora, A. (2021). Decentralised learning in federated deployment environments: A system-level survey. *ACM Computing Surveys (CSUR)*, *54*(1), 1-38.

STATE OF RESEARCH

Aggregation | Masking | Encryption

Basic Idea

hiding individual data in those of many by aggregating data that is distributed in time or space

+ requesting node receives only aggregated data
  → individual data points no longer identifiable or traceable

+ aggregation can also be seen as transformation in a certain way
  - models can be learned/trained locally
  - learned weights of the model (here aggregation) are used to train the global model

STATE OF RESEARCH

Aggregation | Masking | Encryption

Basic Idea

hiding individual data in those of many by aggregating data that is distributed in time or space

- depending on the implementation, decentralised nodes in the network still receive **original data** (cf. SMART **[He et al., 2007]**)
  - possible attacker could impersonate neighbouring nodes in the network
- depending on the algorithm, only certain aggregation functions are possible (Min, Max, Mean, Sum …)
  - **[Zhang et al., 2019]** offers a wide range of applicable functions (increases possibilities)

STATE OF RESEARCH

Aggregation **Masking** Encryption

Basic Idea

original data is enriched so that the exact distribution as well as the absolute data values do not match the original

- by inserting random values (**camouflage**-values) no clear statement can be made, which data was really collected/measured

+ relatively **easy** and computational efficient

− a consideration must be made as to how far the data set can be manipulated

How much loss of information can be tolerated?

STATE OF RESEARCH

Aggregation | Masking | **Encryption**

Basic Idea

original data is first encrypted and aggregated with other encrypted data (→Term: **Secure Aggregation**)

[Zhang, 2011]

- there are procedures that decrypt the data again at the destination node, or not
- \+ in 2nd approach, homomorphic encryption is used to learn on the encrypted data
  - the values are encoded, but relations between different values are still identifiable
- − high data transfer is needed
- − high computation capacity required at the edge nodes for encryption/ decryption

DIFFERENTIALLY PRIVATE LEARNING FROM LABEL PROPORTIONS (DP-LLP)

LLP - OVERVIEW

Conditions

- limited memory (IoT)
- less data overhead for algorithm
- less computation capacity at the decentralised nodes

Approach

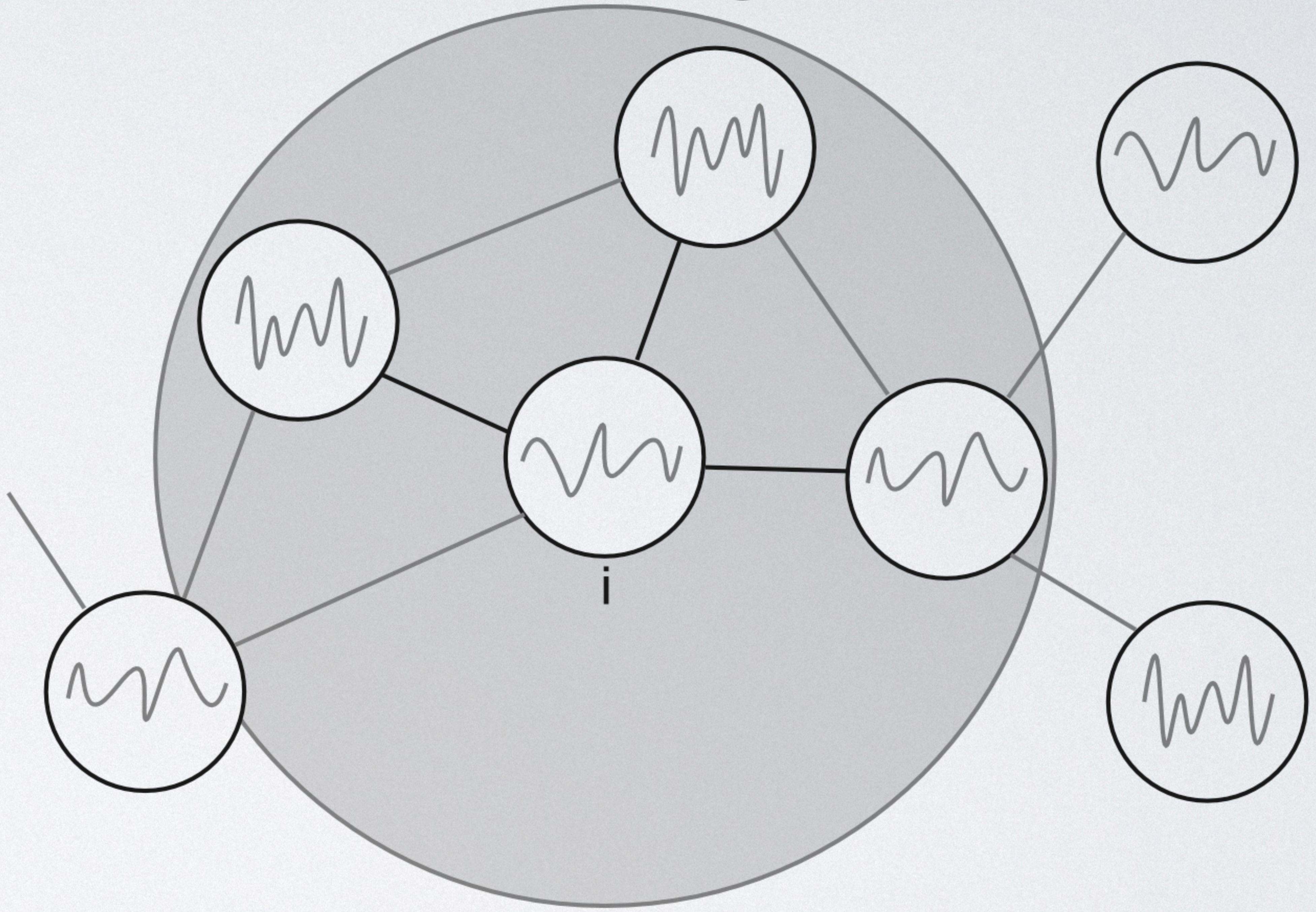
local data is not aggregated over multiple nodes of a graph, but is time-referenced for each node

→each node can only identify its own original data, if only modified data is passed to external properties

[Stolpe et. al., 2015]

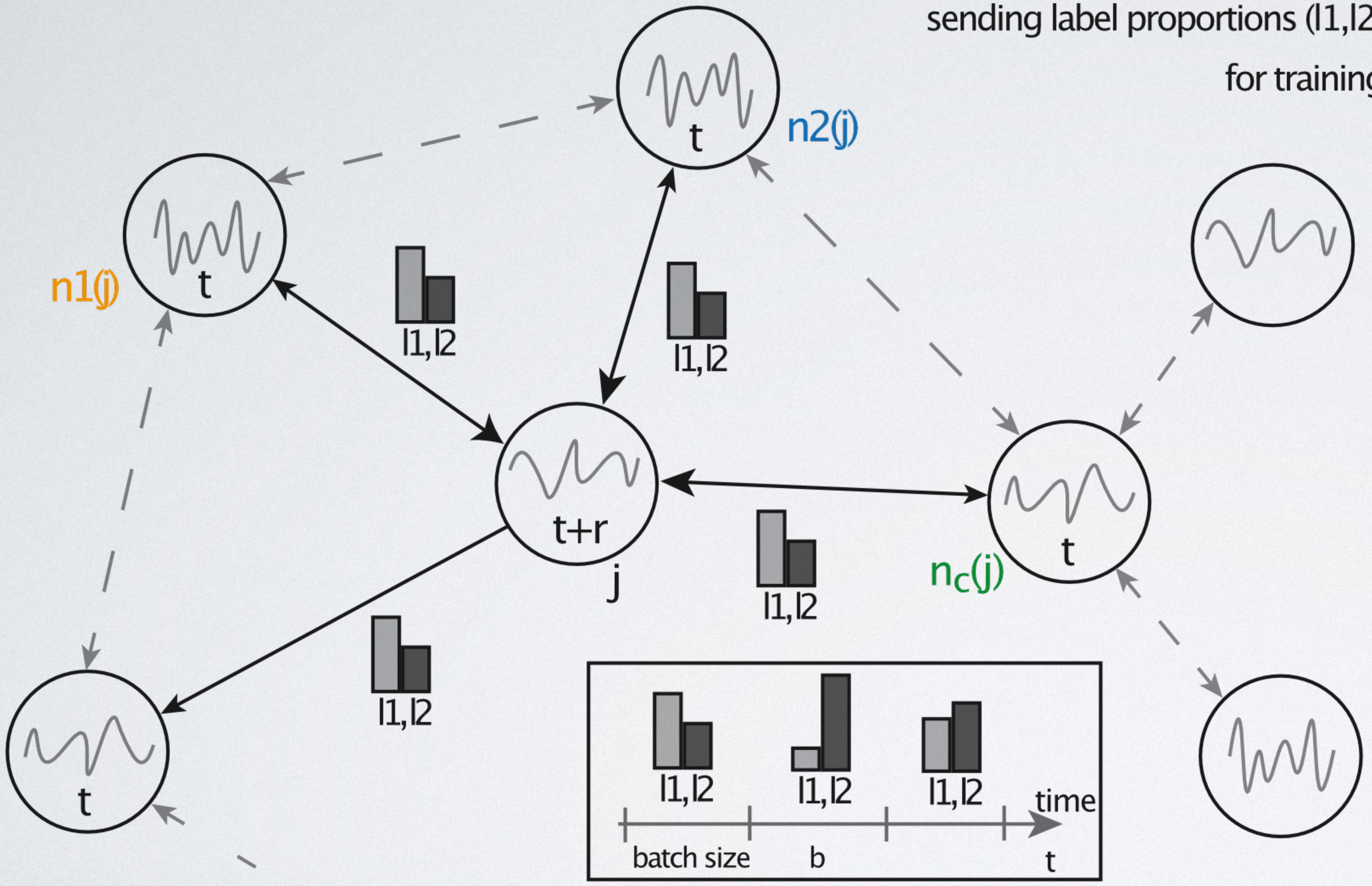
LLP - OVERVIEW

- data cannot be send between all nodes inside the network (would be to much data traffic)
- algorithm assumes, **close** locations have similar behaviour (also in prediction)
- therefore we need to have knowledge about distances



[Stolpe et. al., 2015]

LLP - BASIC ALGORITHM



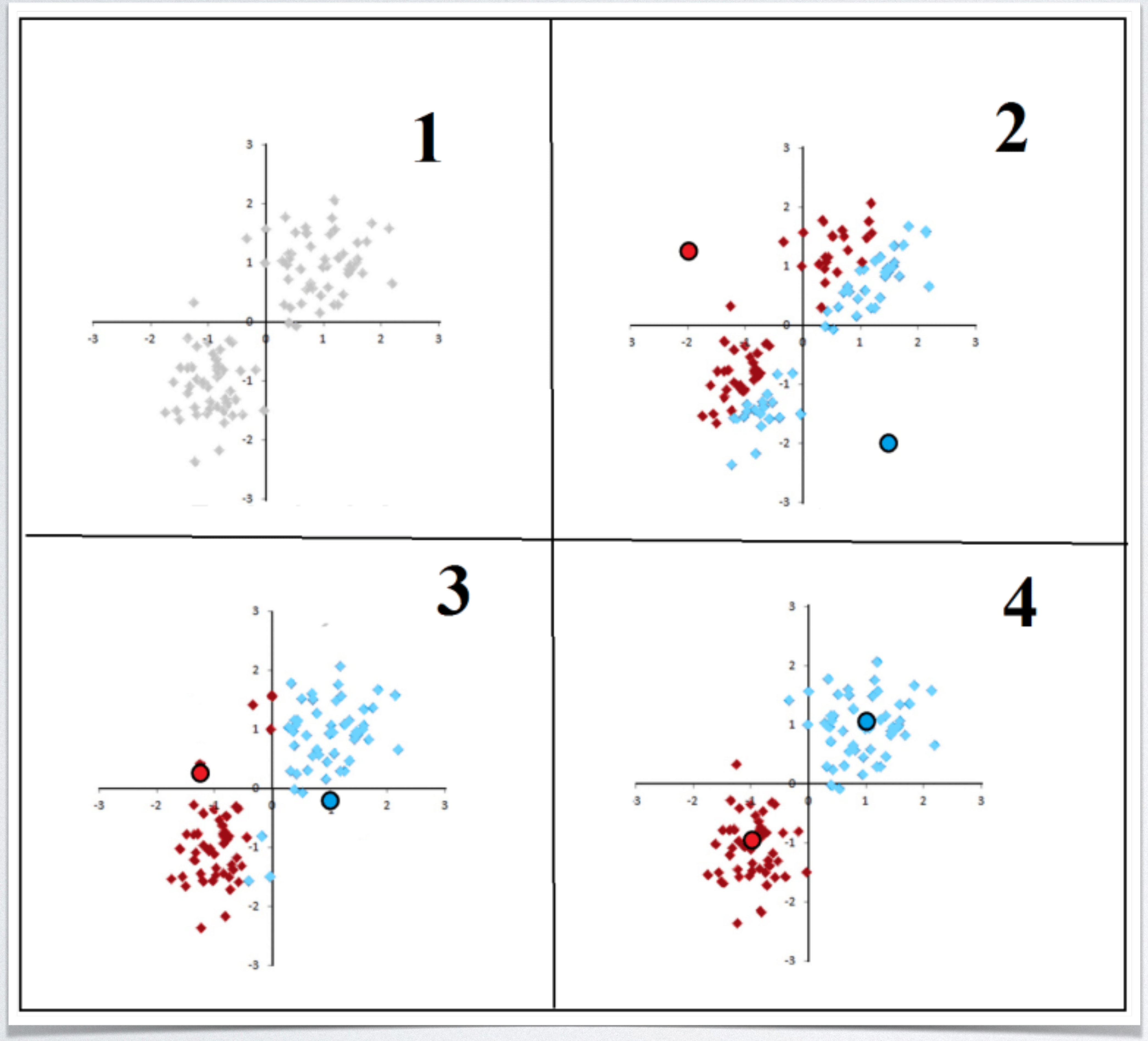
batch size: b
window-size: w
window: x_i

	1	...	w–2	w–1	w	y(j)
x1(j)	0.24	...	0.31	0.76	0.81	l1
x2(j)	0.41	...	0.76	0.81	0.61	l1
x3(j)	0.72	...	0.81	0.61	0.11	l2
x4(j)	0.21	...	0.61	0.11	0.91	l1
...	...	...	...	...	...	...

} b

[Stolpe et. al., 2015]

LLP - BASIC ALGORITHM



1. Assignment Step

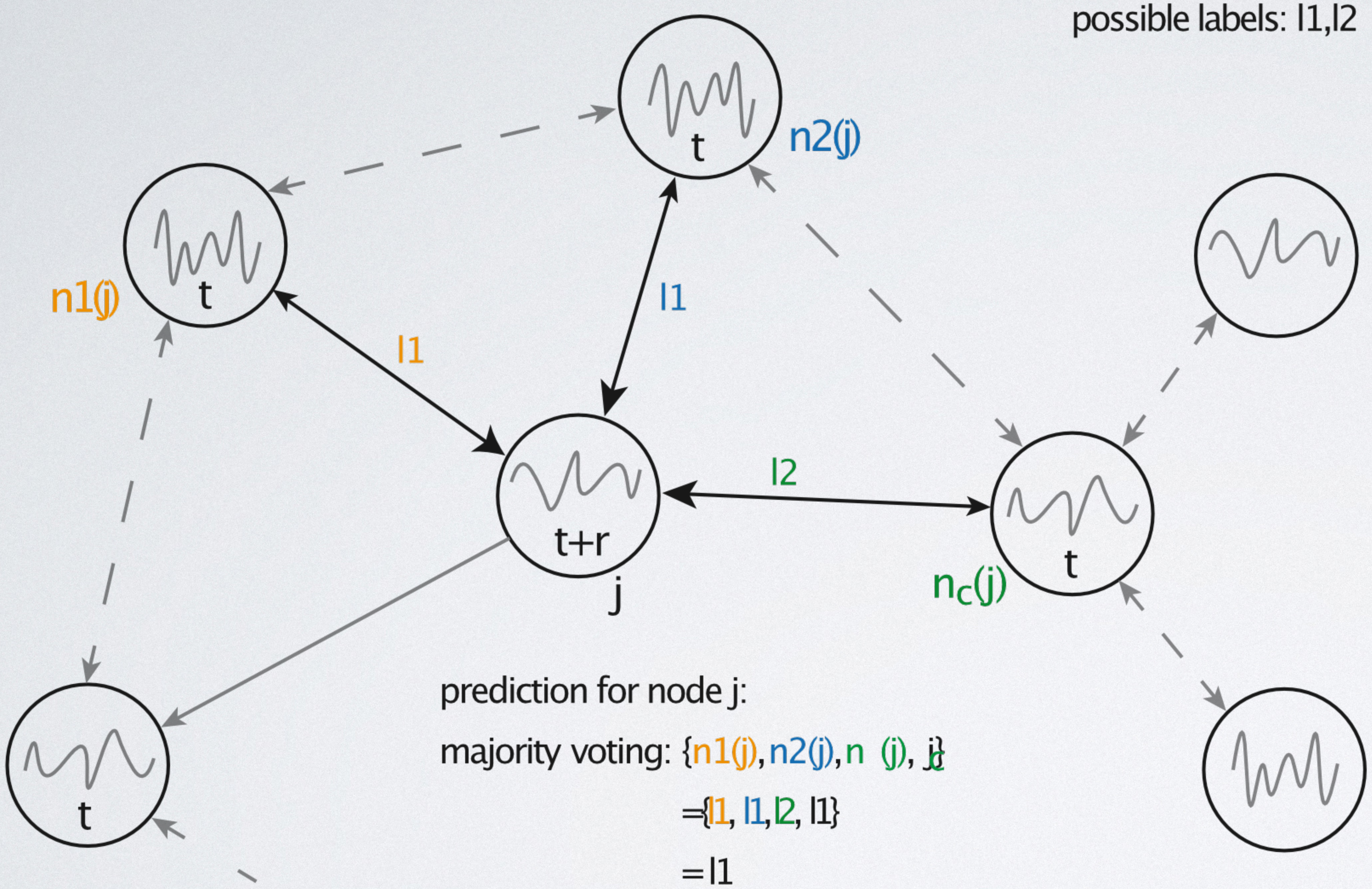
$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \le j \le k \right\}$$

2. Update Step

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

LLP - BASIC ALGORITHM

possible labels: l1,l2



prediction for node j:

majority voting: $\{n1(j), n2(j), n_c(j), j\}$

$= \{l1, l1, l2, l1\}$

$= l1$

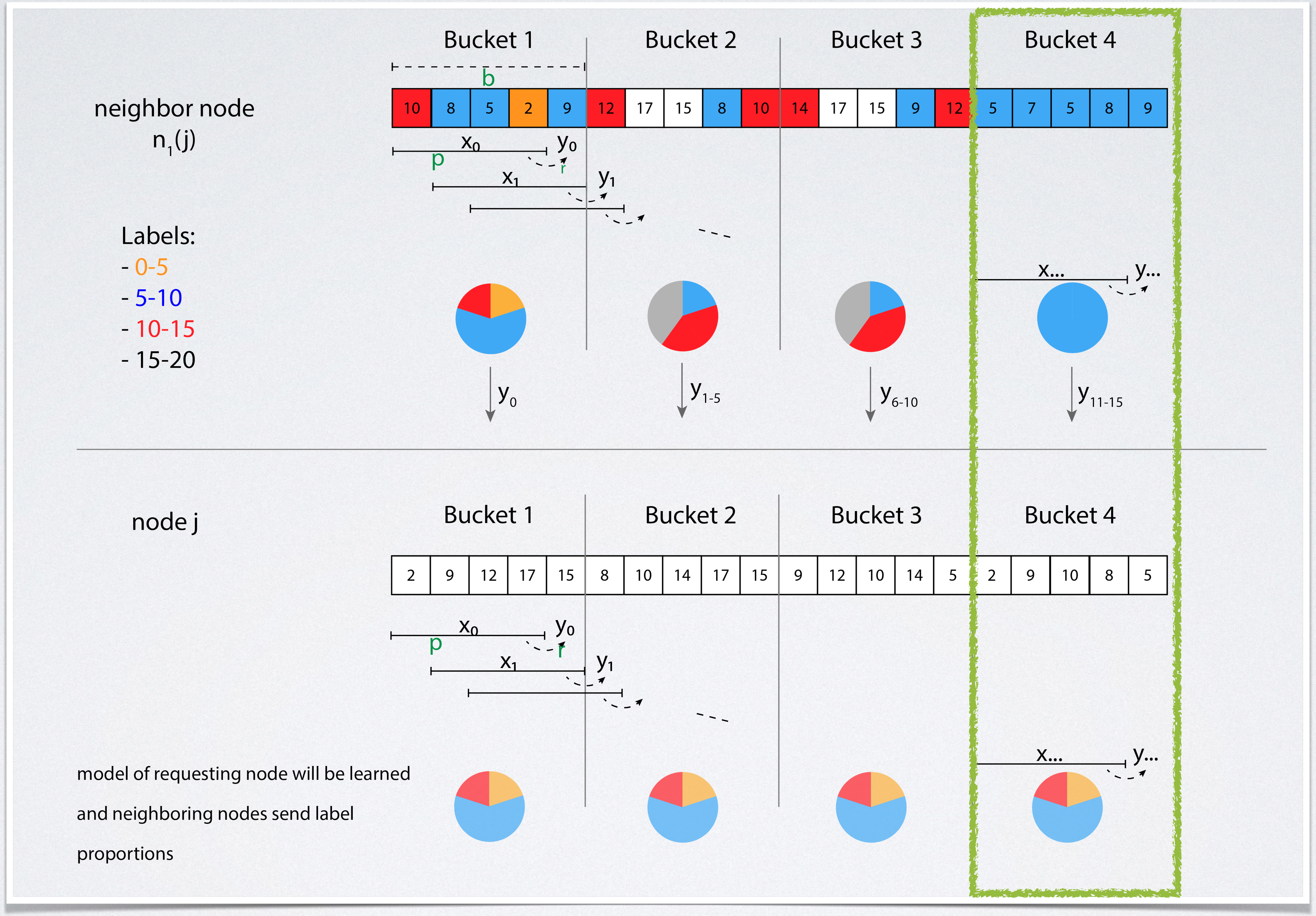
batch size: b
window-size: w
window: x_i

	1	...	w–2	w–1	w	y(j)
x1(j)	0.24	...	0.31	0.76	0.81	l1
x2(j)	0.41	...	0.76	0.81	0.61	l1
x3(j)	0.72	...	0.81	0.61	0.11	l2
x4(j)	0.21	...	0.61	0.11	0.91	l1
...	...	...	...	...	...	...

}b

[Stolpe et. al., 2015]

LLP - DATA LEAKAGE



DIFFERENTIALLY **P**RIVATE **L**EARNING FROM **L**ABEL **P**ROPORTIONS (DP-LLP)

DP-LLP - DIFFERENTIAL PRIVACY

Definition: Differential Privacy

A randomised algorithm $M : D \rightarrow R$ with domain D and range R is ϵ-differentially private if for all $S \subseteq R$ and for any $D', D'' \in D$ such that $||D' - D''||_1 \leq 1$:

$$\Pr\left[M(D') \in S\right] \leq e^{\epsilon} \Pr\left[M(D'') \in S\right]$$

[Dwork et al., 2014]

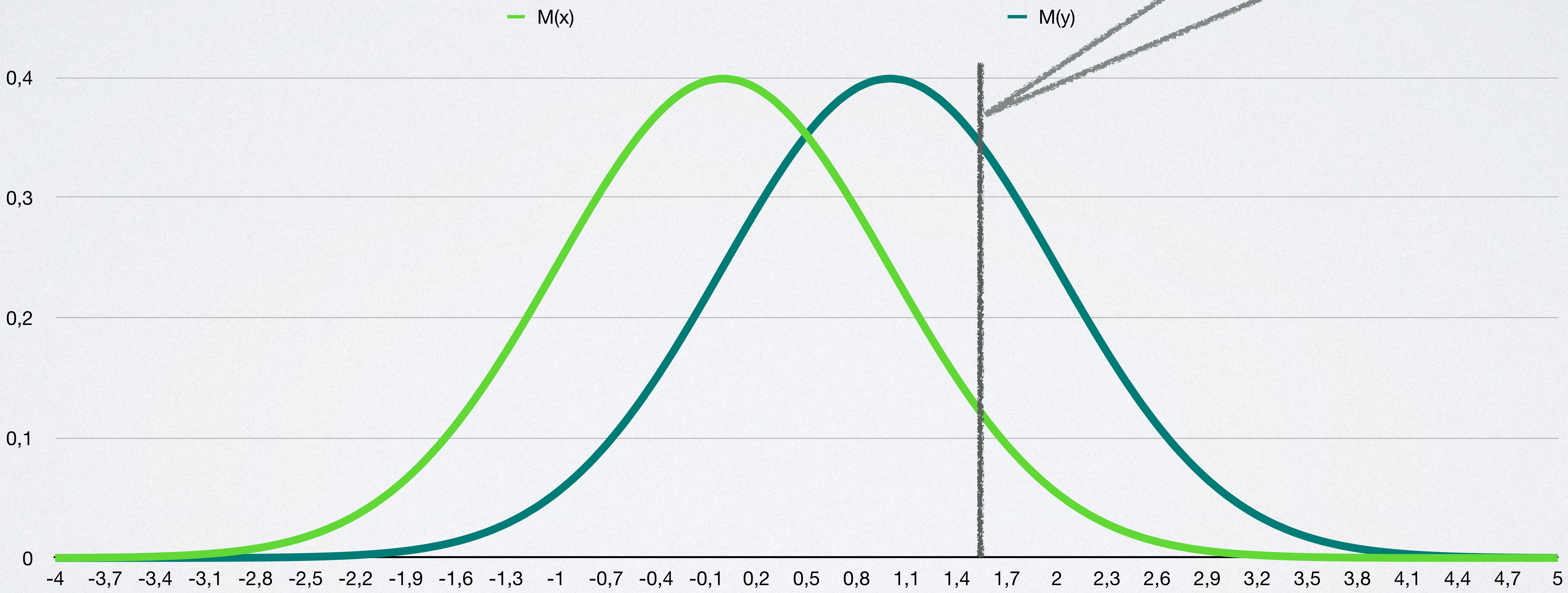
Explanation

- with ϵ a degree of deviation between the datasets D' and D'' can be specified, wich is allowed
- $\epsilon = 0$: totally privacy compliant, no difference between the two datasets
- the higher ϵ is chosen, the more noticeable the missing date is in the dataset

DP-LLP - DIFFERENTIAL PRIVACY

	M(x)	M(y)
Mean	0	1
Stdv.	1	1

max vertical distance limited by e^{ϵ}



DP-LLP - DIFFERENTIAL PRIVACY

Definition: Sensitivity

The l_1-sensitivity of a function $f : D \rightarrow R$ is:

$$\triangle f = \max_{D', D'' \in D,\ ||D' - D''||_1 = 1} \quad ||f(D') - f(D'')||_1$$

[Dwork et al., 2014]

Explanation

- sensitivity indicates the factor/value by which a single datum can influence the dataset in the worst-case scenario
- using sensitivity to regulate how much **noise** must be calculated on the data to be **differentially private**

DP-LLP - DIFFERENTIAL PRIVACY

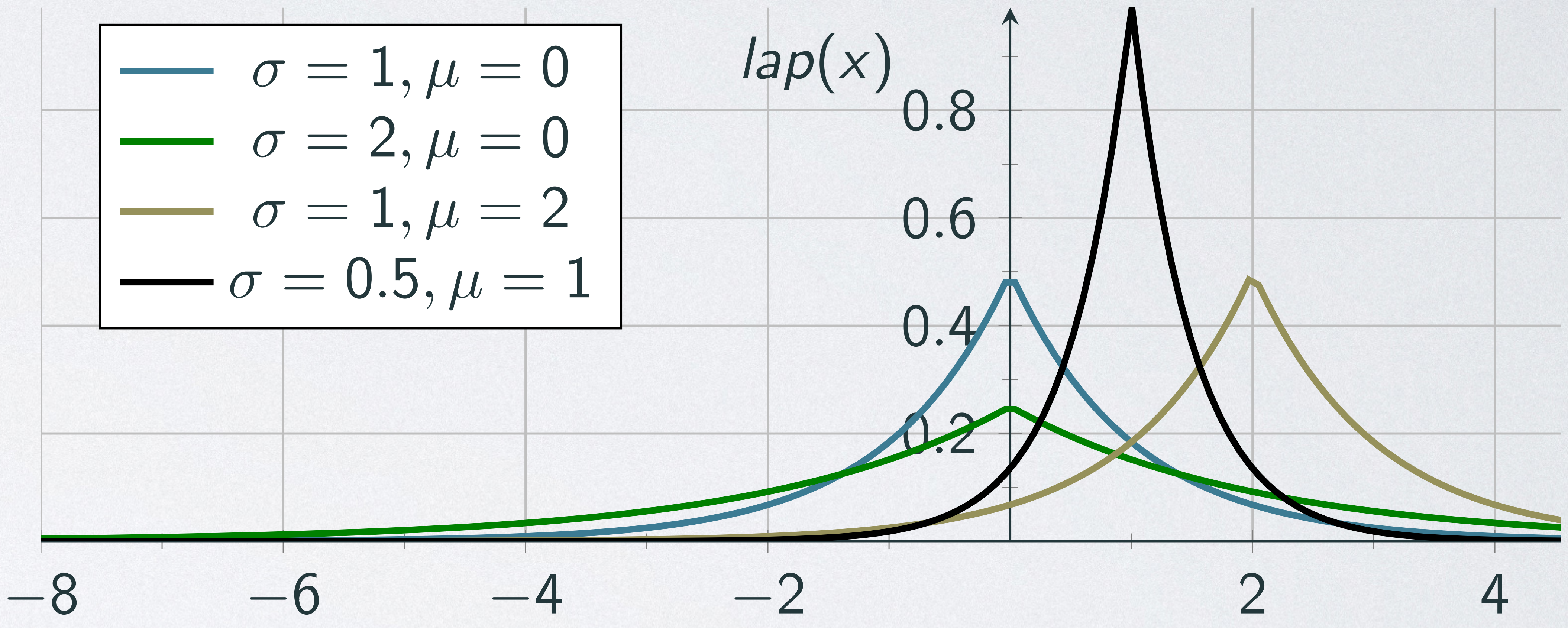
Definition: Laplace Distribution

With ϵ, $\triangle f$ as l_1-sensitivity function given and D as real data points given:

$$Pr(R = x \,|\, D = \text{trueworld}) = \frac{1}{2\sigma} e^{\frac{-|x-\mu|}{\sigma}}$$
$$= \frac{\epsilon}{2 \triangle f} e^{-\frac{|x-f(D)|\epsilon}{\triangle f}}$$

[Dwork et al., 2014]

Example



DP-LLP - DIFFERENTIAL PRIVACY

Proof

$$\frac{Pr[M(D') \in S]}{Pr[M(D'') \in S]} \leq e^{\epsilon}$$

$$\frac{\frac{\epsilon}{2 \triangle f} e^{-\frac{|x-f(D')|\epsilon}{\triangle f}}}{\frac{\epsilon}{2 \triangle f} e^{-\frac{|x-f(D'')|\epsilon}{\triangle f}}} \leq e^{\epsilon}$$

$$\frac{e^{-\frac{|x-f(D')|\epsilon}{\triangle f}}}{e^{-\frac{|x-f(D'')|\epsilon}{\triangle f}}} \leq e^{\epsilon}$$

triangle equation: $|a| - |b| \leq |a - b|$

$$e^{-\frac{|x-f(D')|\epsilon}{\triangle f} + \frac{|x-f(D'')|\epsilon}{\triangle f}} \leq e^{\epsilon}$$

$$e^{\frac{\epsilon}{\triangle f}|f(D')-f(D'')|} \leq e^{\epsilon}$$

distance between $f(D')$ and $f(D'')$ cannot be higher than $\triangle f$

$$\rightarrow \frac{|f(D') - f(D'')|}{\triangle f} \leq 1$$

DP-LLP - DIFFERENTIAL PRIVACY

Now, we know, **how** we can add noise to prevent leakage of private data! And we also know, that we can **prevent** exploitation of private data to a **specific degree**!

But where can this noise be applied?

DP-LLP - ALGORITHM

gaia-x Hub Germany

technische universität dortmund

Require: $B_1, \ldots, B_h, Y$
Ensure: $Q(j)$

```
for i in 1..h do
    for j in 1..|Y| do
        Q(j)_{i,j} ← sum(B_i == Y_i)
    end for
    m ← sum(Q(j)_i)
    for j in 1..|Y| do
        Q(j)_{i,j} ← Q(j)_{i,j} + lap(e = 0, s = 1/ϵ)
        clip Q(j)_{i,j} to bounds [0.001, m]
    end for
    normalize Q(j)_i
end for
```

calculating upper clipping value

calculating noise for each aggregated value + adding to the private data

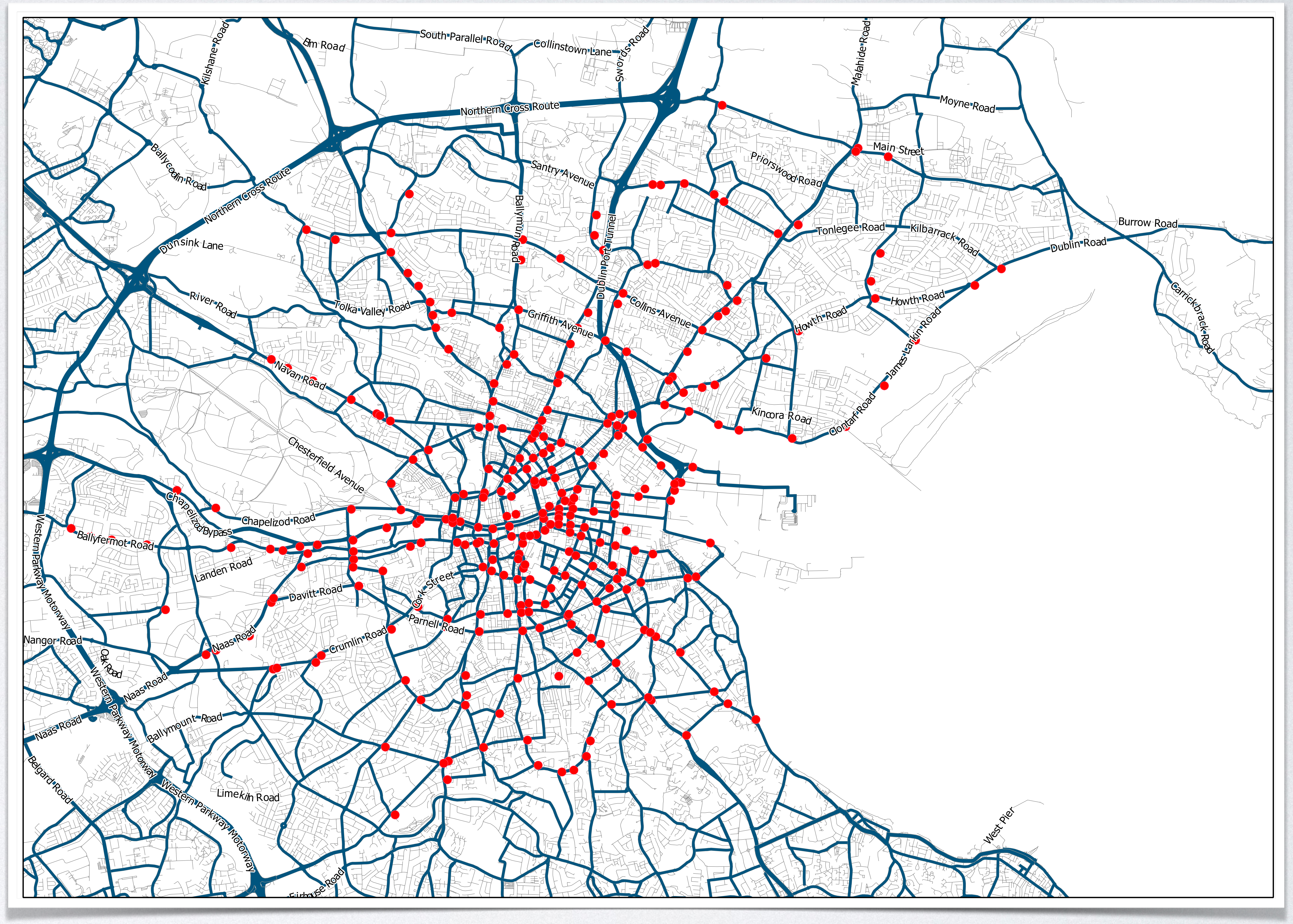
clipping to prevent negative values

EVALUATION

PERFORMANCE STATISTICS

EVALUATION: DATASET

- traffic volume data from **S**ydney **C**oordinated **A**daptive **T**raffic **S**ystem
- sensors at each traffic signal
- contains 5 minute averaged values of traffic flow
- data from *January 2013*
- continuous values, that are mapped to *5 class labels*: $[0, \frac{1}{52}, \frac{6}{52}, \frac{16}{52}, \frac{30}{52}, 1]$



Overview of traffic flow sensors [McCann, 2014] gaia-x Hub Germany

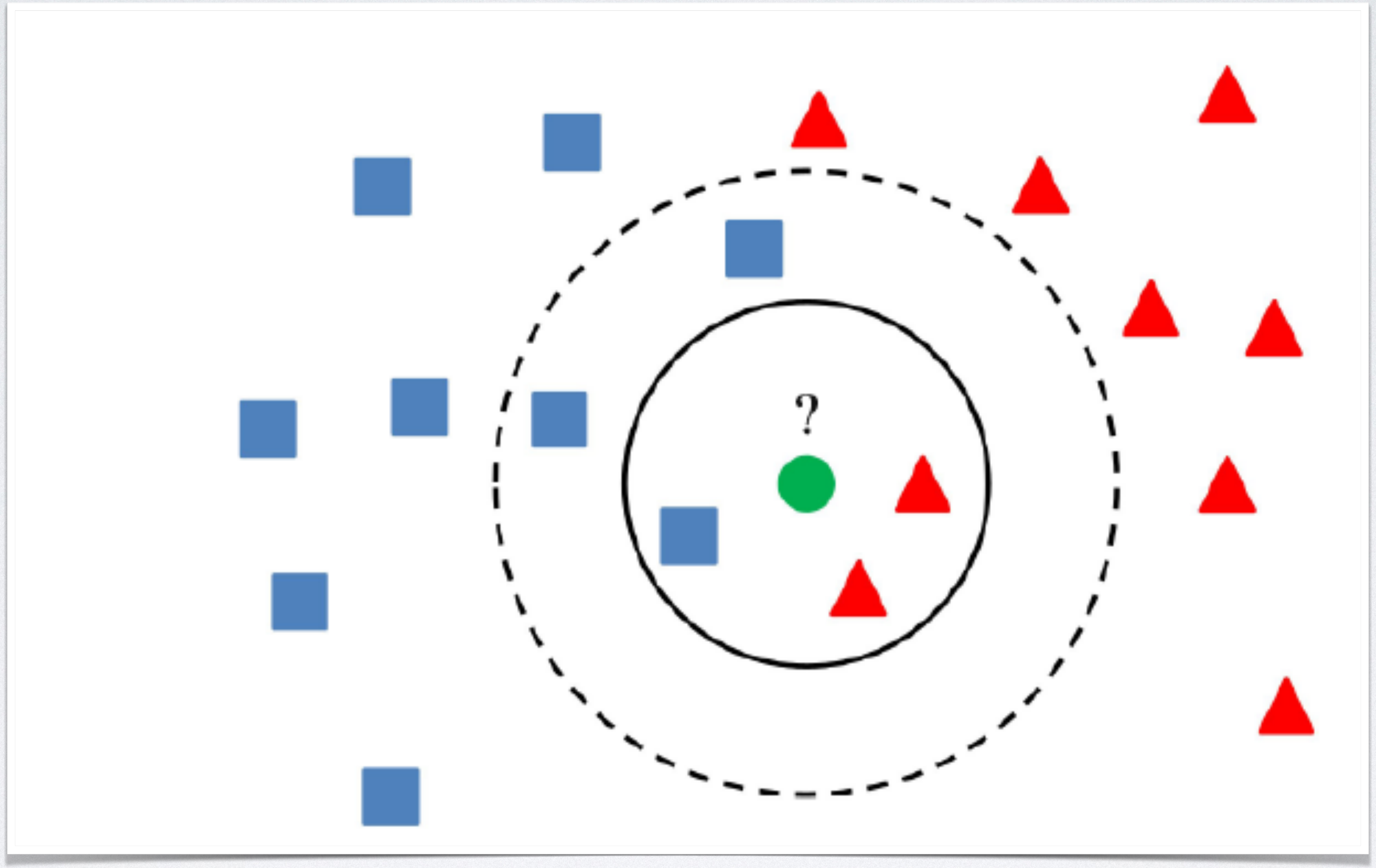
EVALUATION: DATASET

technische universität dortmund

- 57 parking sensors, which are located in Hamilton, Australia
- two labels (used=red, free=green)
- only value changes are stored
  → preprocessing in timeframes is necessary
- 391,444 entries between 2019 and 2021

Overview of parking sensors
[Southern Grampians Shire Council, 2021]

LLP - OVERVIEW



[Alaliyat, 2022]

EVALUATION: SETUP

Steps

1. modification batch size b
2. modification cluster size c
3. modification ϵ
4. comparison results between datasets

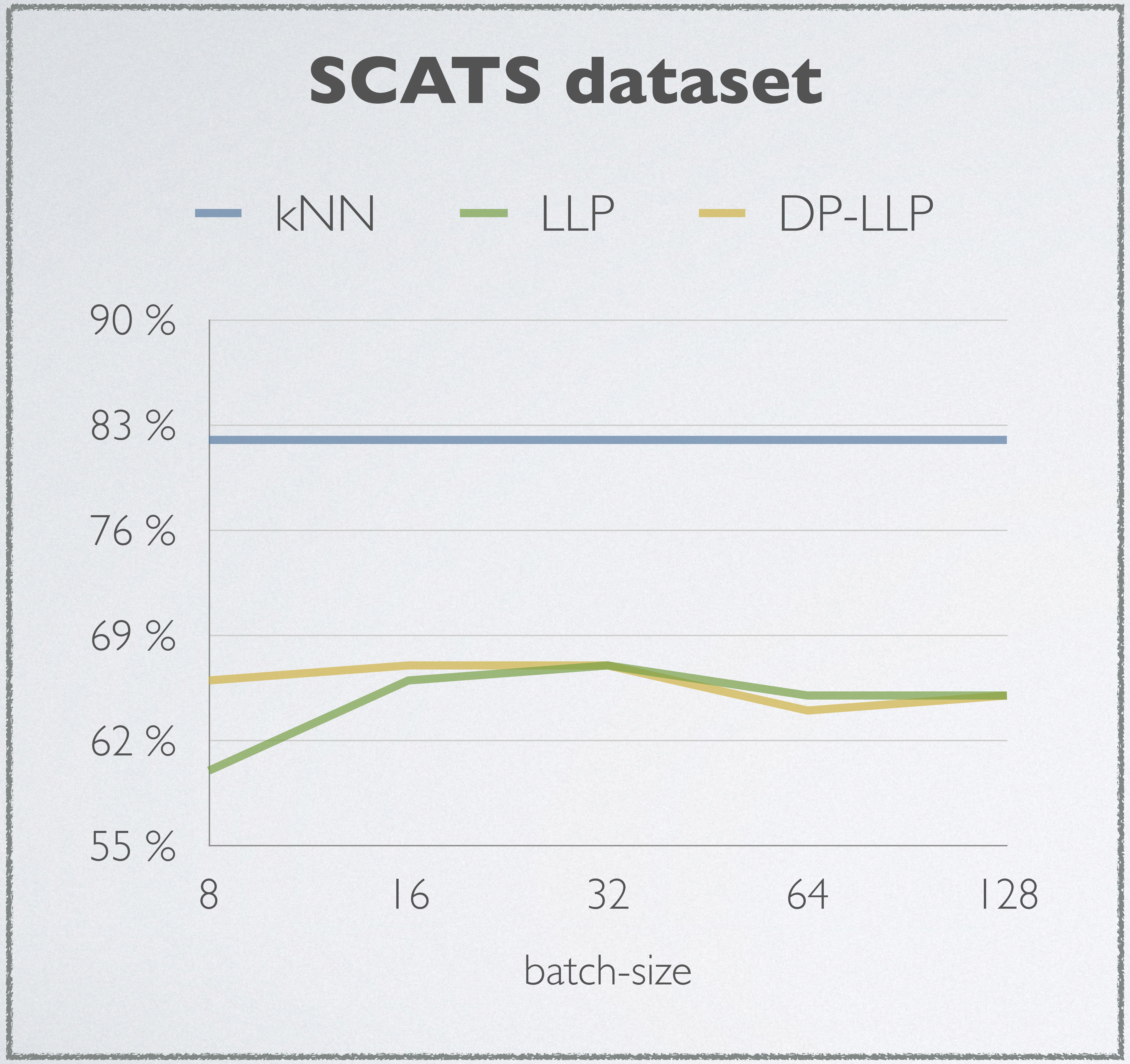
General Information

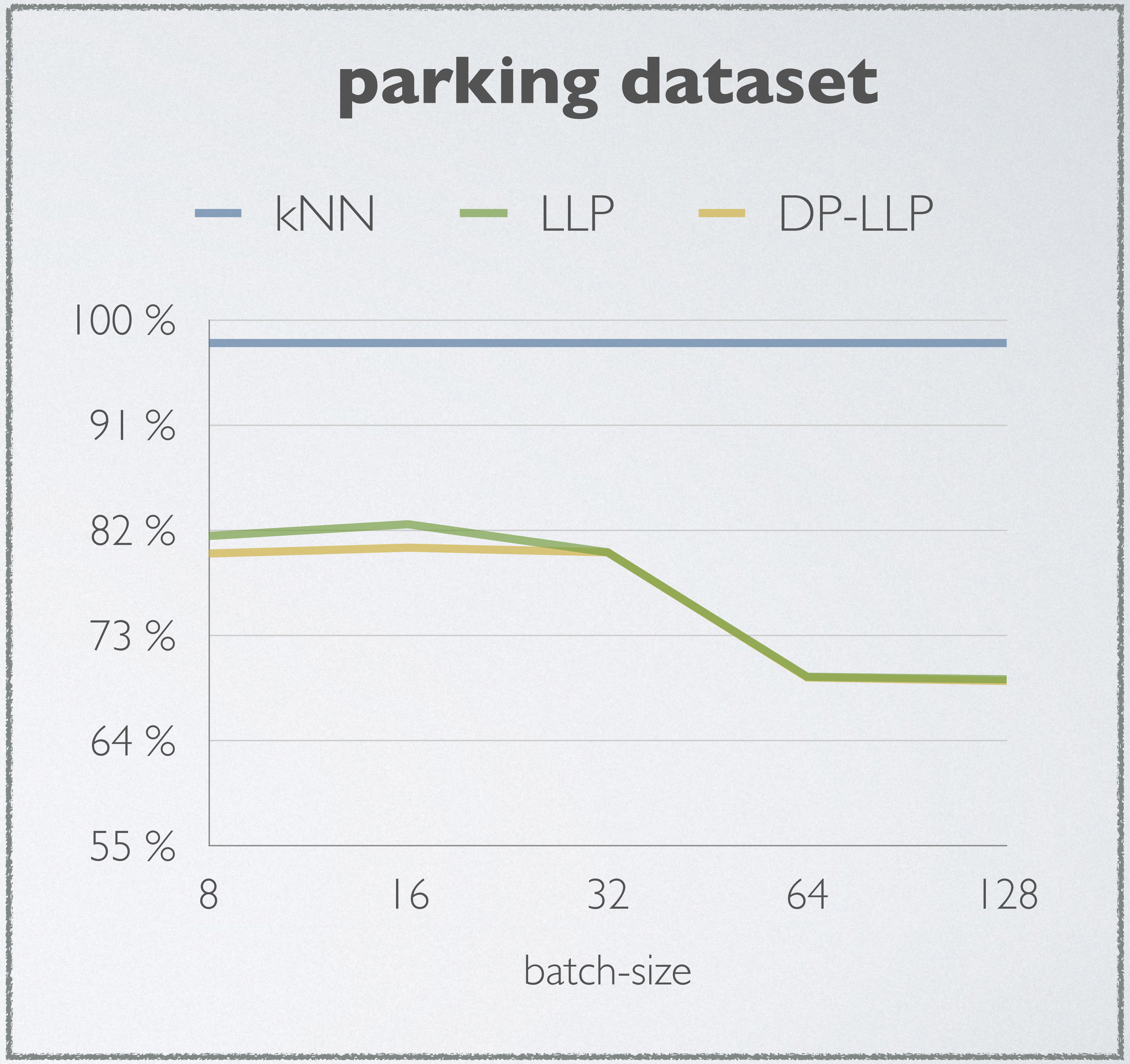
- cross validation with 10 folds
- Evaluation on the two types of datasets
  - vehicle flow
  - parking sensors
- comparison with k-nearest neighbor

$$acc = \frac{tp + tn}{tp + fp + tn + fn}$$

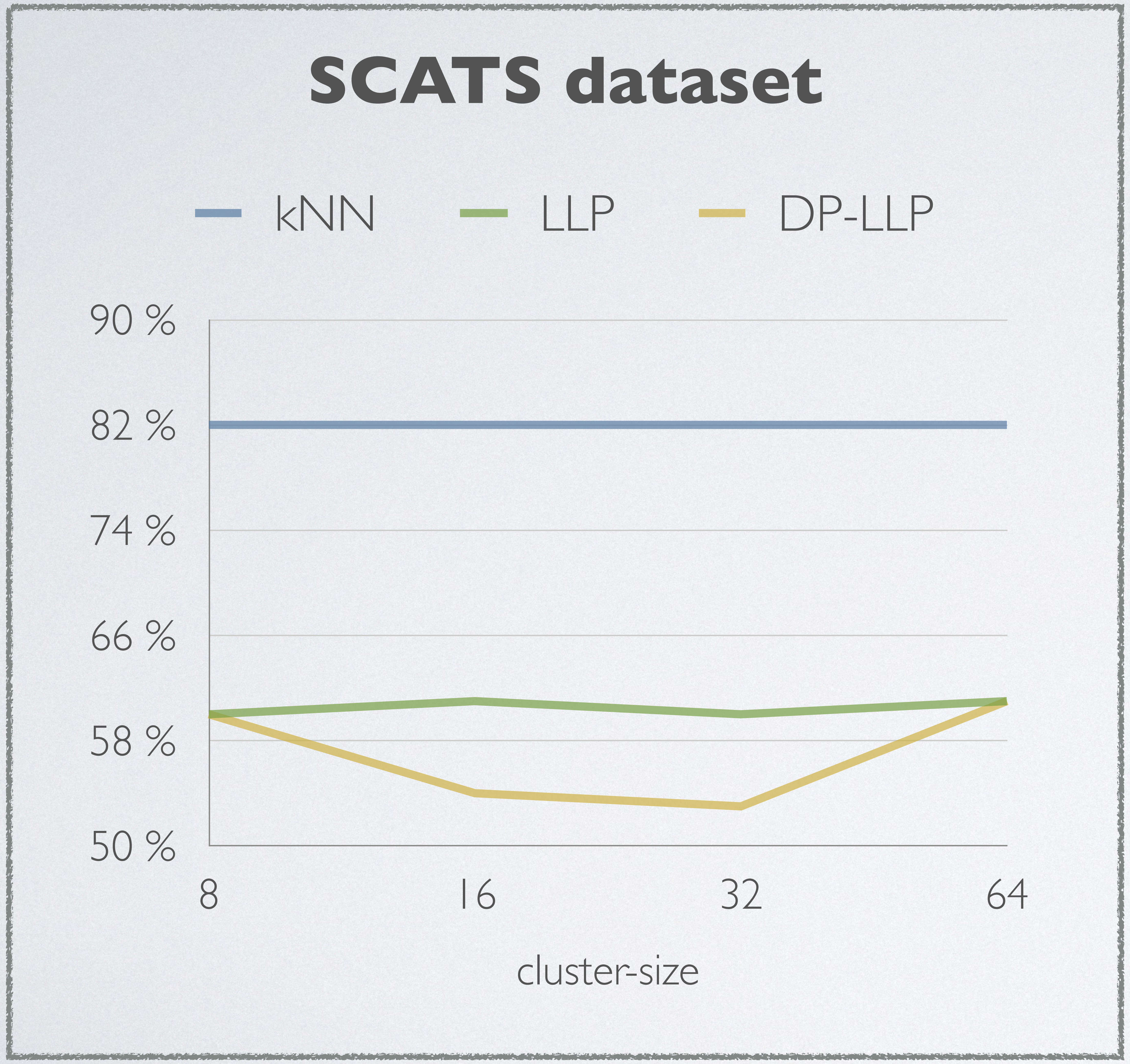
predicted \ actual	positiv	negativ
positiv	TP	FN
negativ	FP	TN

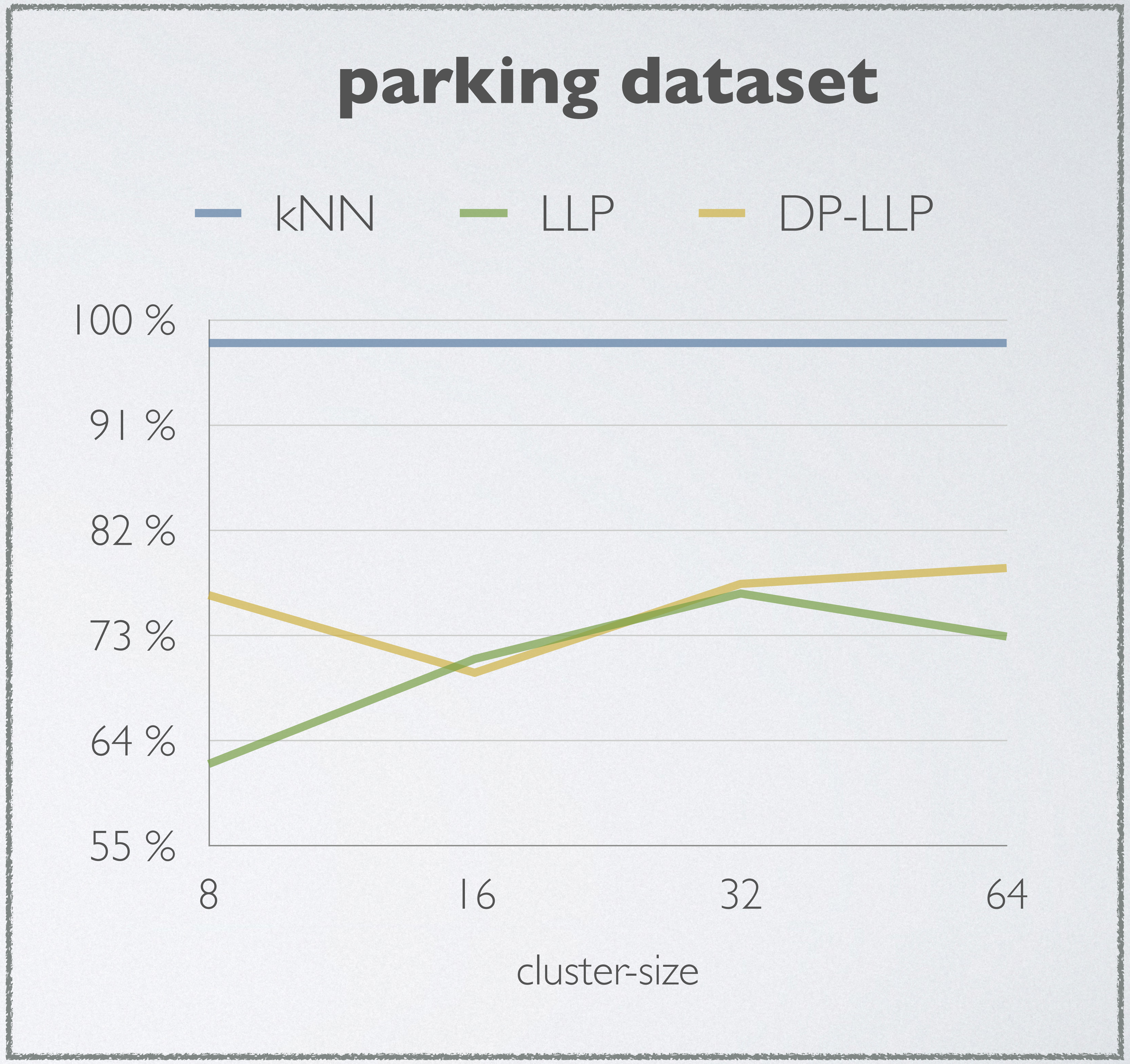
EVALUATION - BATCH-SIZE





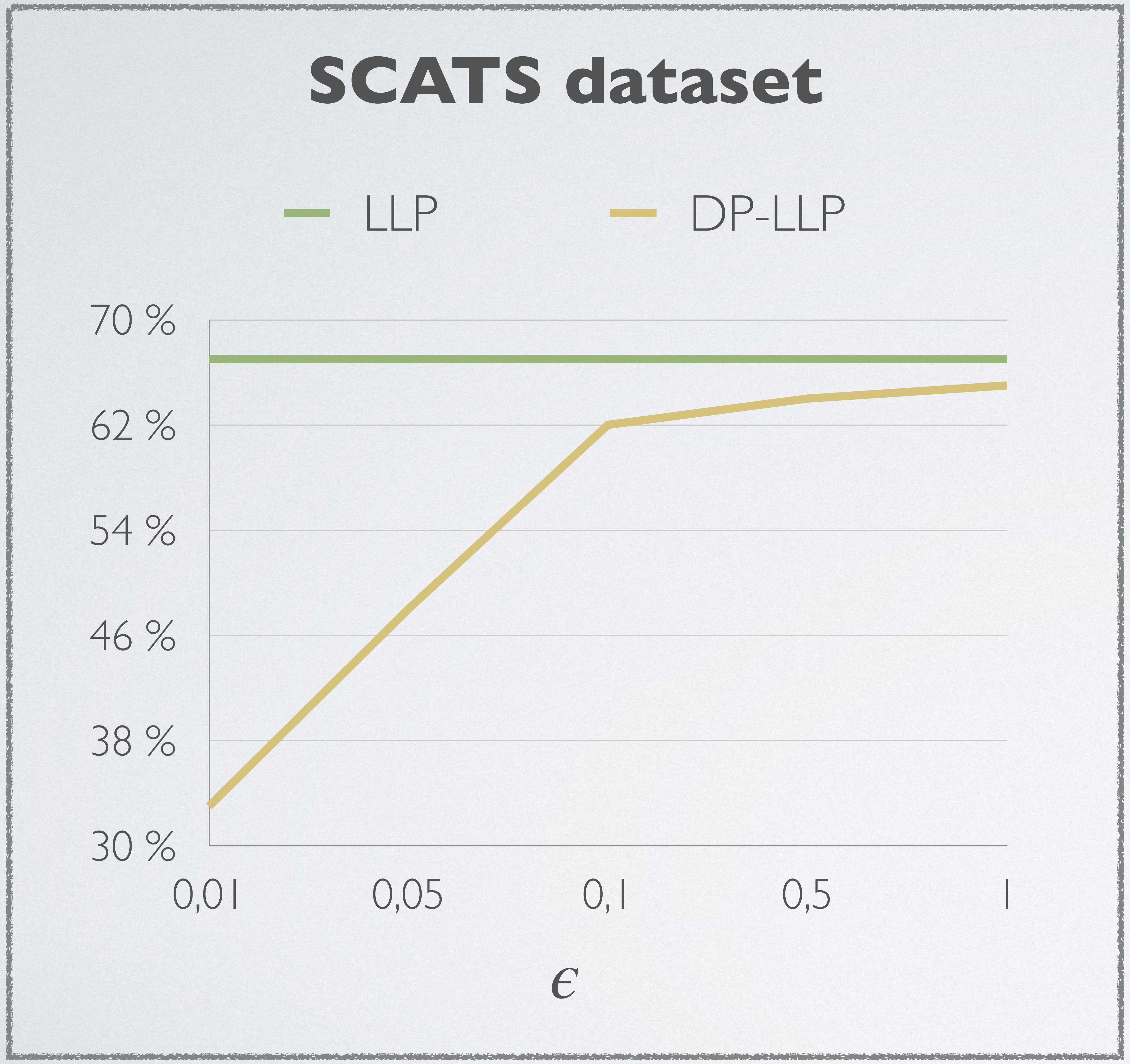
EVALUATION - CLUSTER-SIZE

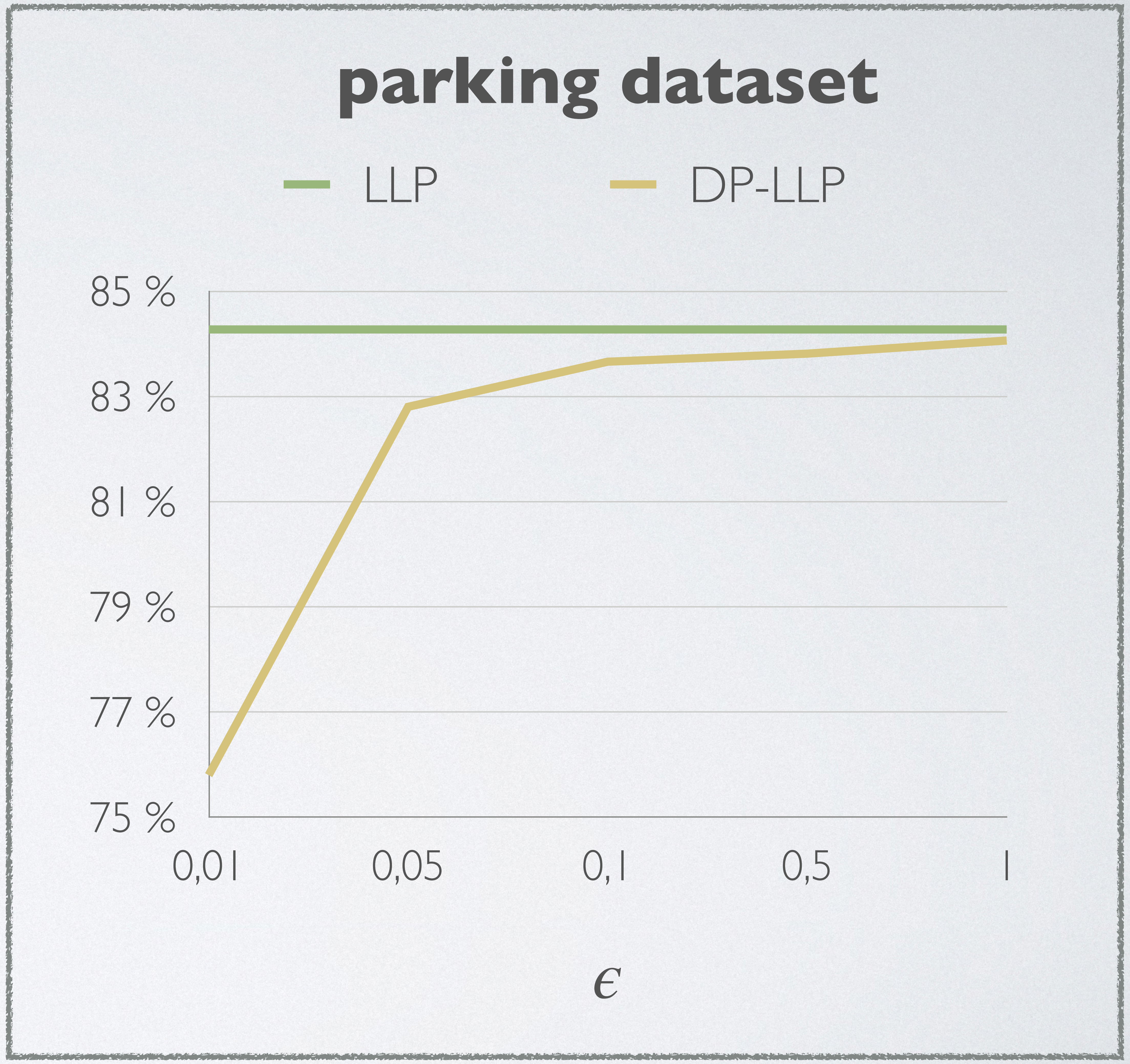




EVALUATION - CLUSTER-SIZE

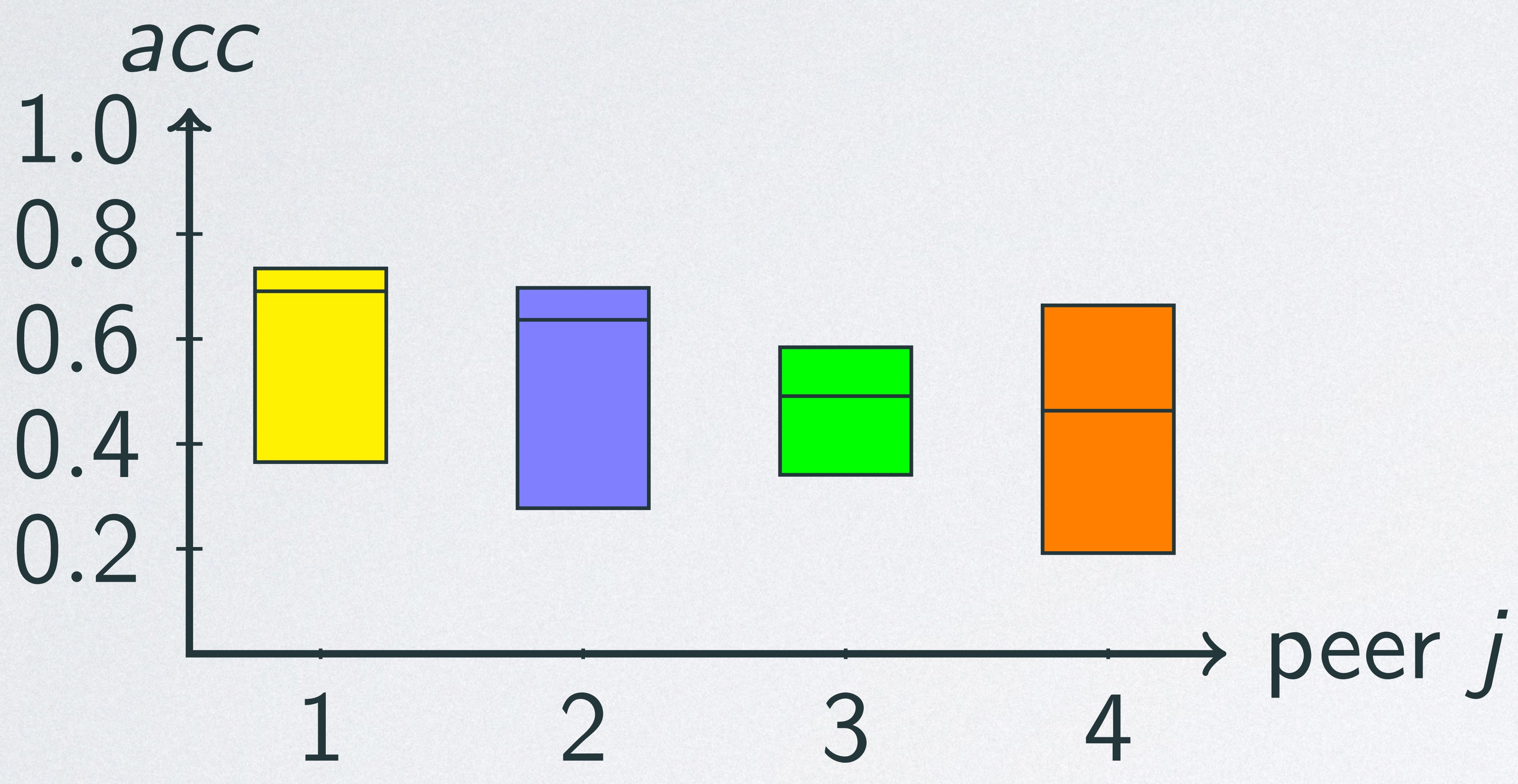
EVALUATION - ϵ



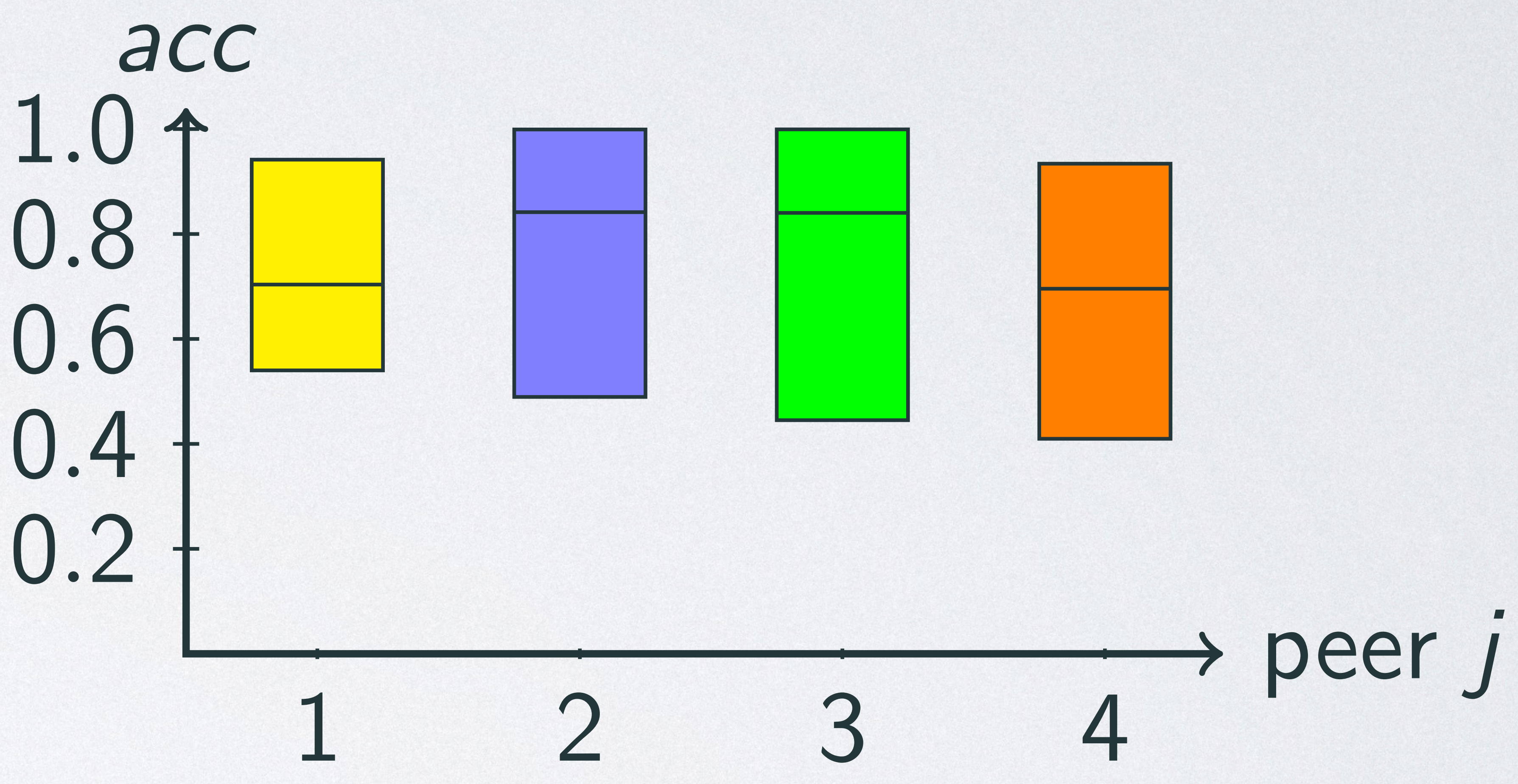


EVALUATION - DATASETS

SCATS dataset



parking dataset



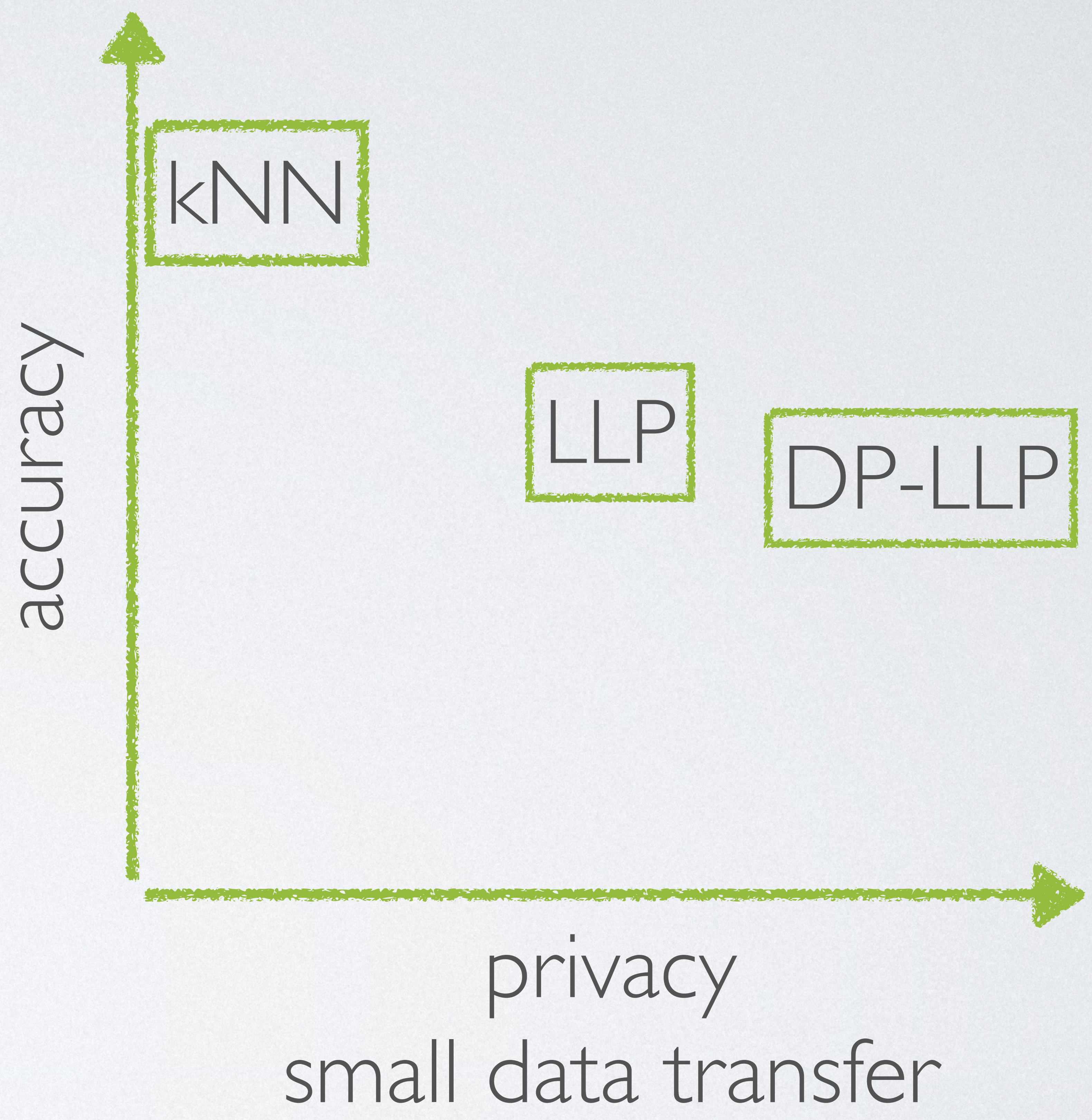
CONCLUSION

gaia-x Hub Germany

technische universität dortmund

CONCLUSION

- the **DP-LLP** algorithm reaches nearly the same performance as the LLP algorithm
- by choosing ϵ, companies can set the tradeoff between **data privacy** and **accuracy**
- most influence on the **accuracy** is achieved by varying the **batch-size** and ϵ
- performance not as good as centralised approaches (but **significantly less data** is used)



FUTURE POSSIBILITIES

implementation of **Online Learning**

merging typical **Federated Learning** (exchange of weights) with **DP-LLP**

more complex **ML**-model inside the **DP-LLP**

DISCUSSION

LITERATURE

Dwork, C., Roth, A., et al. (2014).
The algorithmic foundations of differential privacy.
Foundations and Trends in Theoretical Computer Science, 9(3-4):211-407.

McCann, B. (2014).
A review of scats operation and deployment in dublin.
In Proceedings of the 19th JCT traffic signal symposium & exhibition.

Southern Grampians Shire Council (2021).
Smart community & open data portal: Parking sensors.

Stolpe, M., Liebig, T. and Morik, K. (2015).
Communication-efficient learning of traffic flow in a network of wireless presence sensors.
In Proc. of the Workshop on Parallel and Distributed Computing for Knowledge Discovery in Data Bases (PDCKDD), CEUR Workshop Proceedings, CEUR-WS.

Hub Germany

LITERATURE

technische universität dortmund

Zhang, W. (2011).
Secure Data Aggregation, pages 1104-1105.
Springer US, Boston, MA.

Alaliyat, S. (2022).
Video-based Fall Detection in Elderly's Houses.