Presented by:

ODSC

ACCELERATE AI

ACCELERATE AI EAST

BOSTON | April 30-May 4

2019

THE LEADING DATA SCIENCE CONFERENCE

Big Data and Mobility Analytics

What can we learn from the way things (and humans!) move?

ODSC Boston
2019.05.01
Arturo Opsetmoen Amador

Outline

Cities as economy drivers
- Mobility and economy in a global world
- Mobility as a fundamental part of everyday life

From static to dynamic: location data (MA)
- IoT
- Big Data
- Telco networks as IoT devices
- Mobility Analytics in Norway

The Privacy Challenge
- Re-identification mechanisms
- GDPR
- Privacy enhancing technologies
- Is it really anonymous?

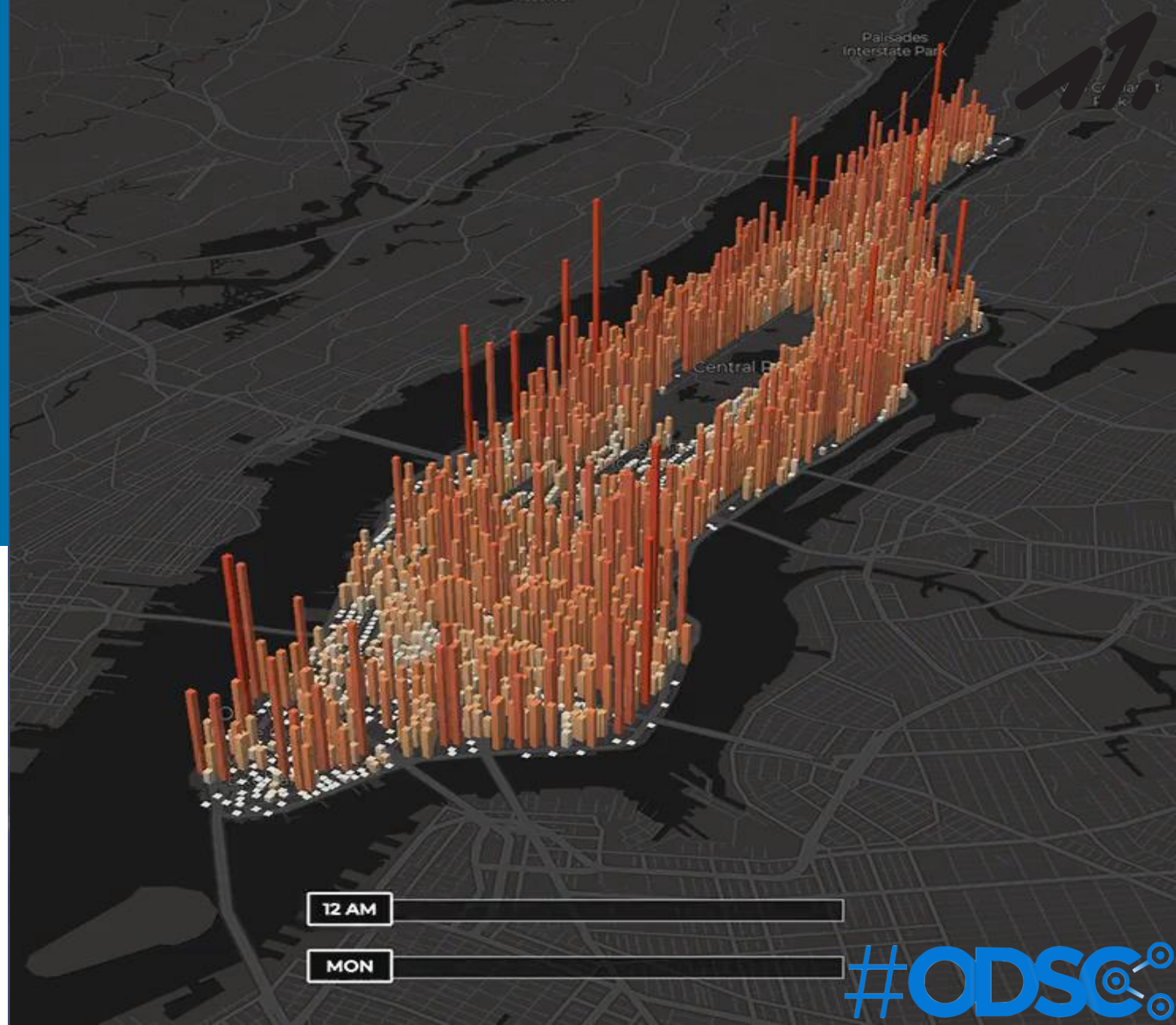
Location data enrichment: merging data sources
- Wastewater epidemiology
- Drinking water, electricity, ammonia
- AirBnB

Big Data technologies and intelligent maritime navigation
- The business case for mar. nav.
- Norwegian navigation data
- Analytics for intelligent maritime navigation

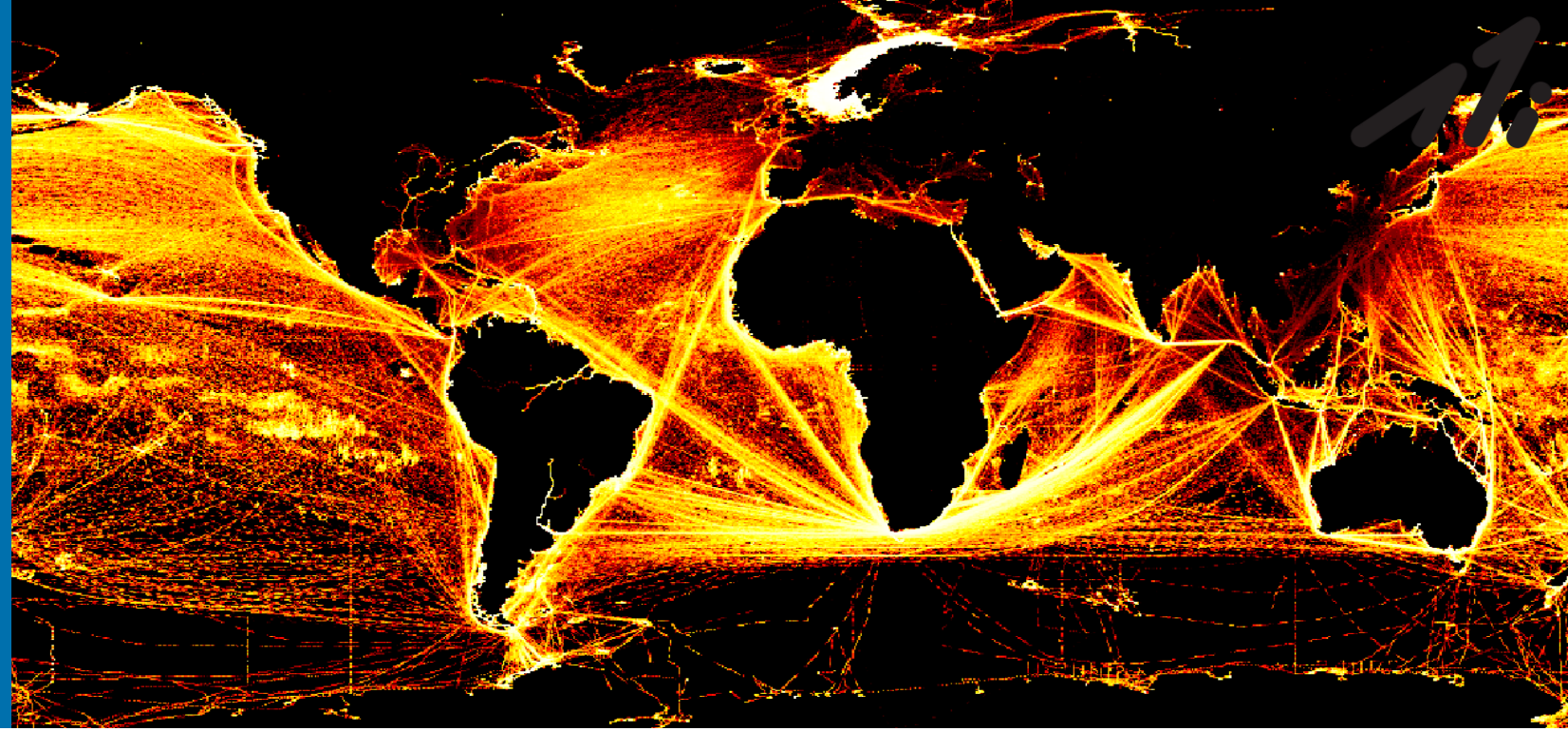
#ODSC

Cities as drivers of economy

Today cities have become the world's dominant demographic and economic clusters



Mobility in a global economy



Top industries in Norway

- Petroleum
- Natural gas
- Shipping
- Fishing

… All of them highly dependent on transport of goods!

Mobility:

Fundamental part of life

60.2 %

"Market share" of busses as public transport option

183 million

Trips taken in public transport

15.8%

Passenger increase since March 2017

3,462 MNoK

Ticketing revenue public transport last year

Internet of Things

12 %

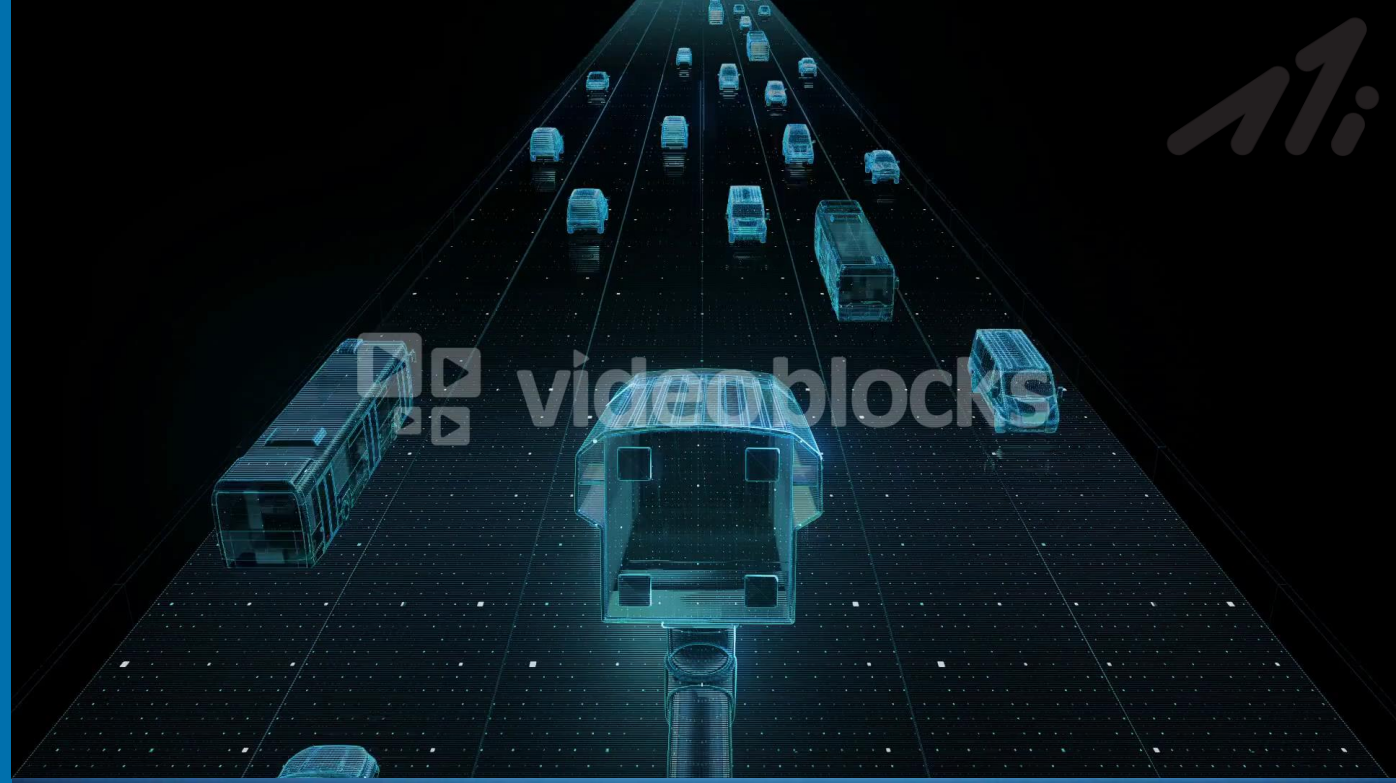
Average yearly increase in the number of connected IoT devices

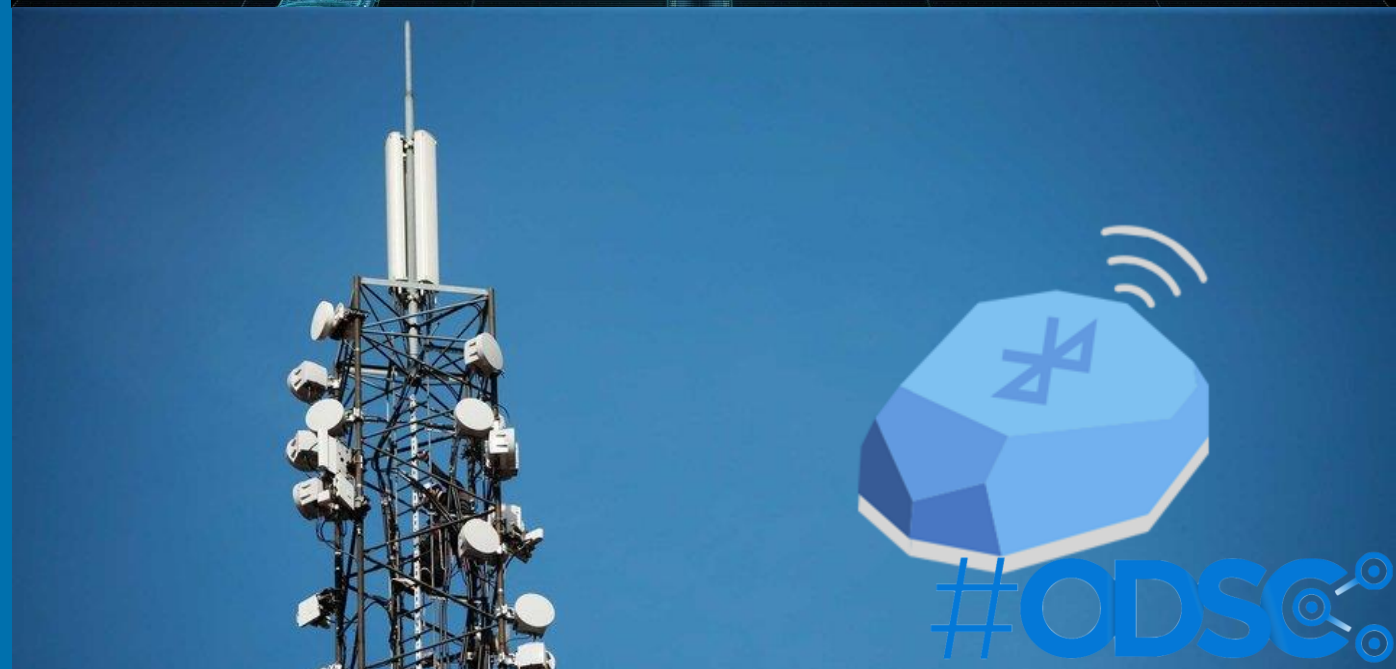
17 billion

Number of connected IoT devices in 2017

125 billion

Number of connected IoT devices expected in 2030





The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

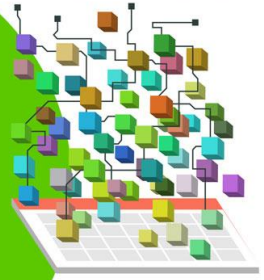
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

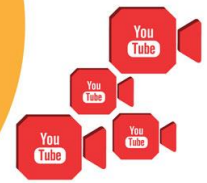
Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

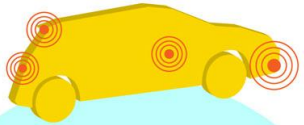
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

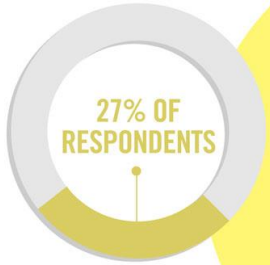


Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



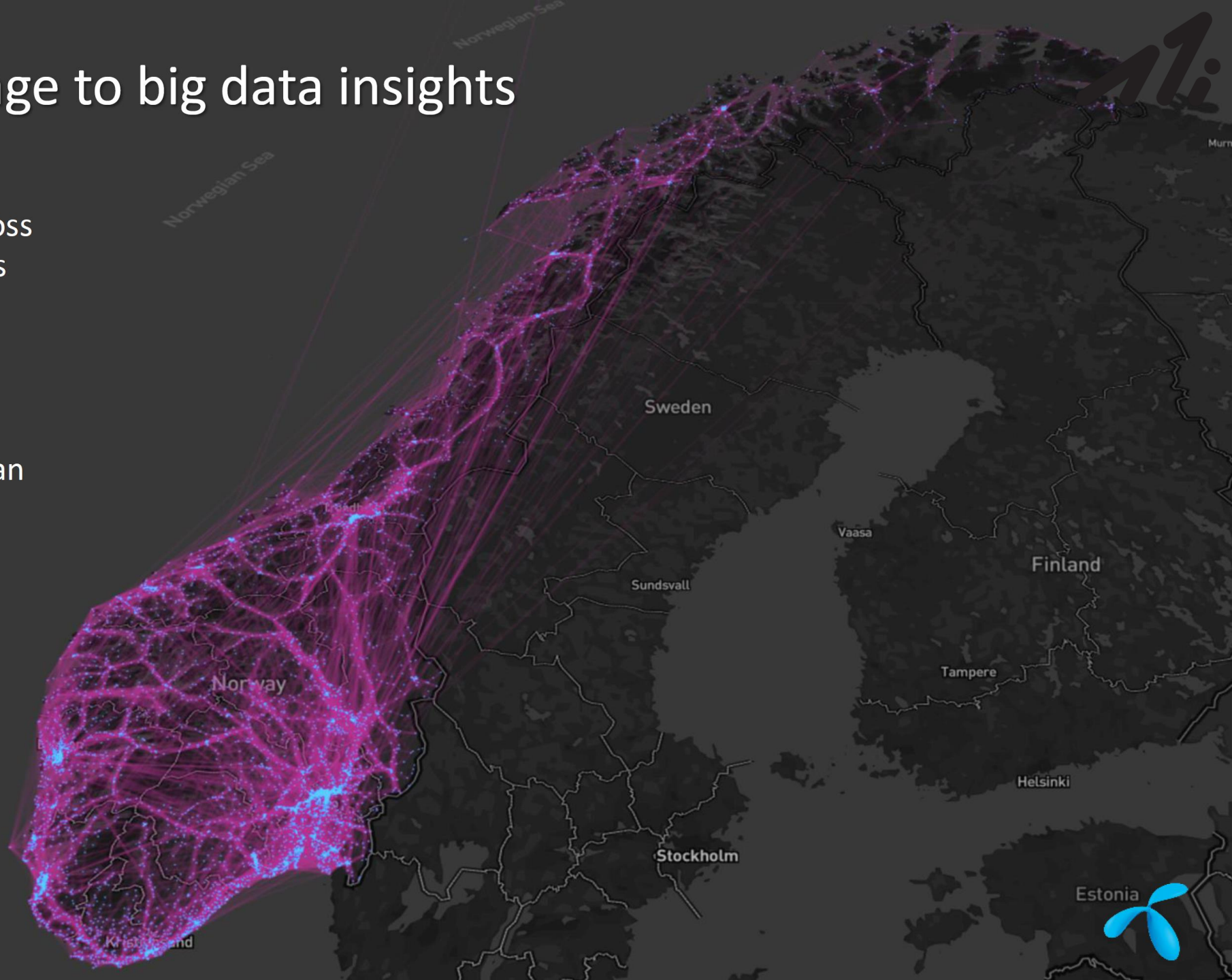
in one survey were unsure of how much of their data was inaccurate

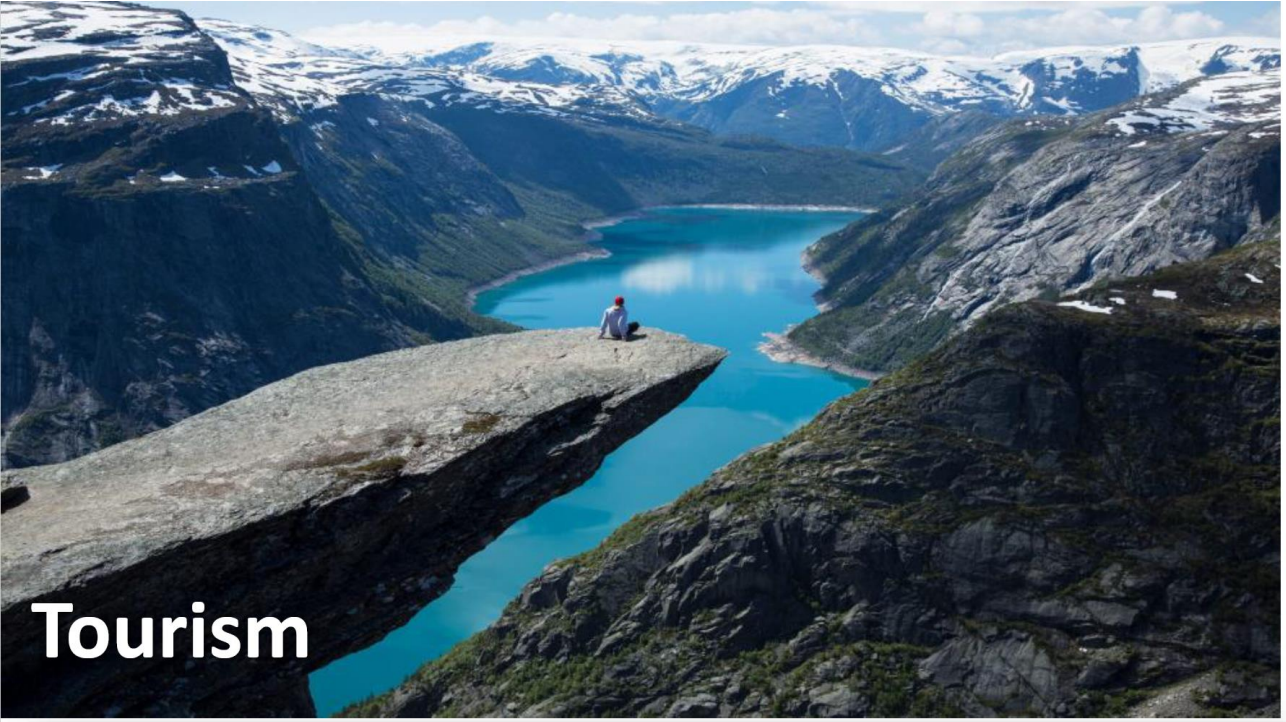
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

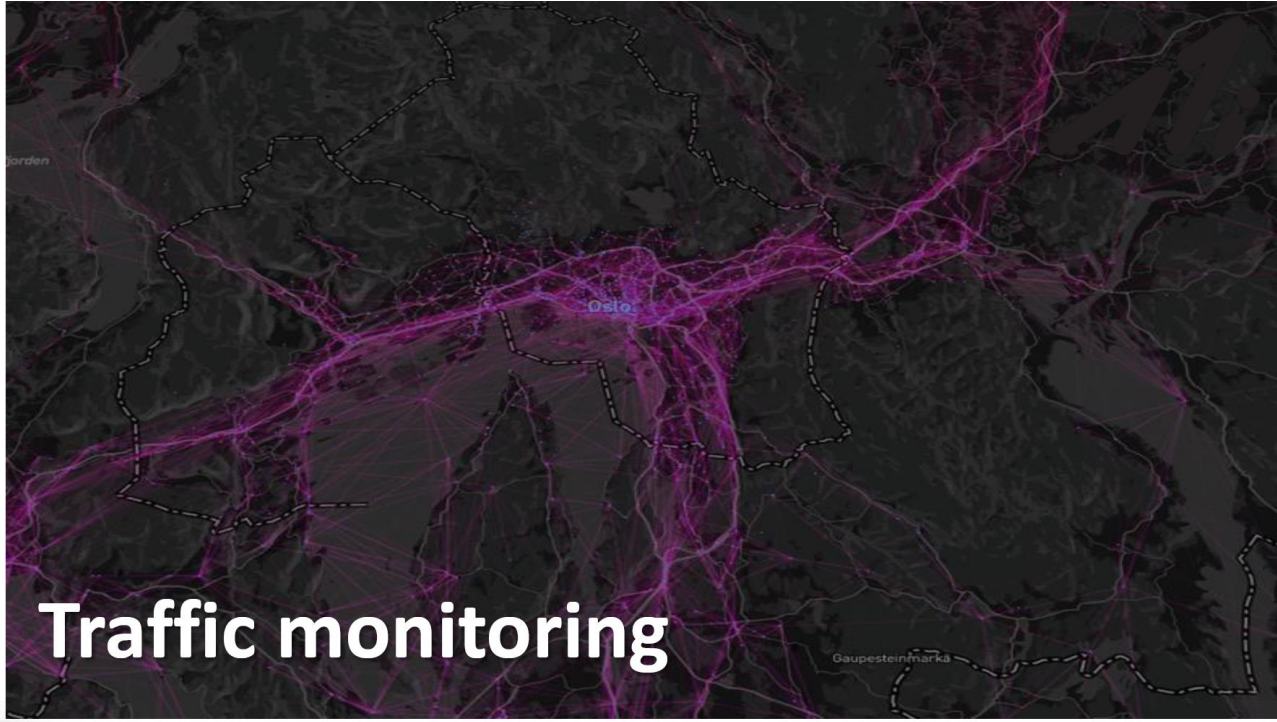
#ODSC

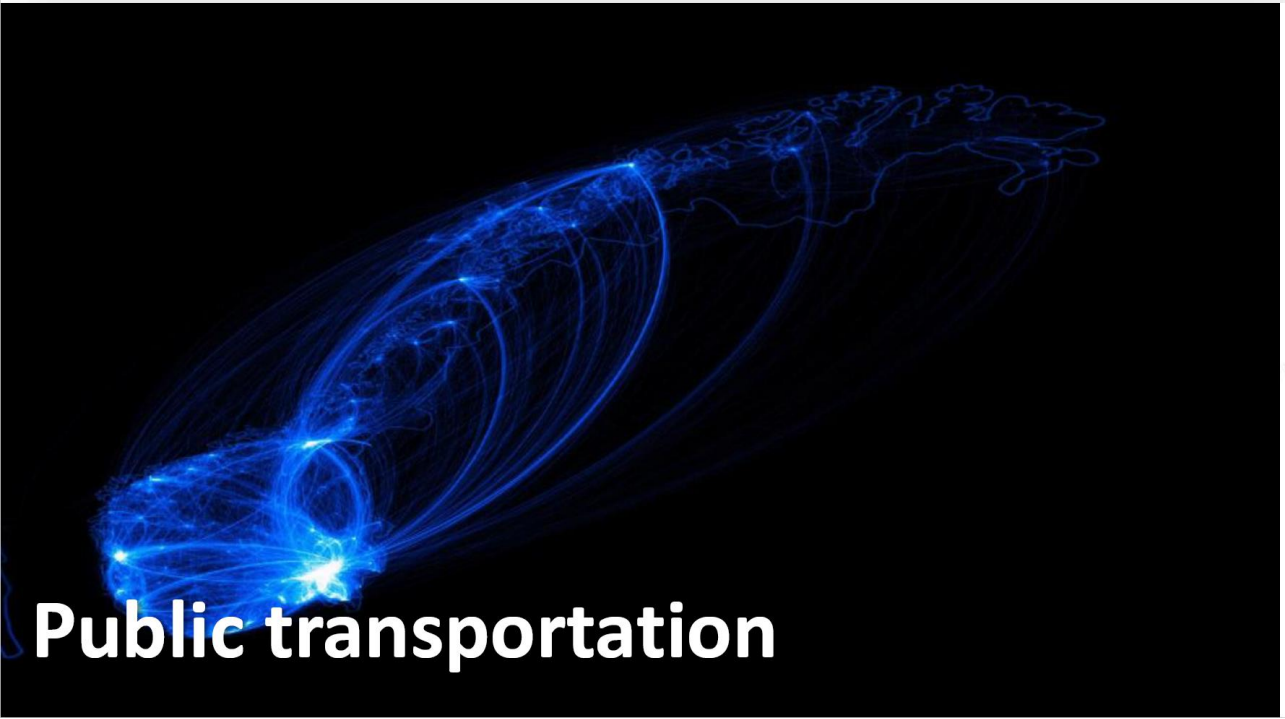
From great coverage to big data insights

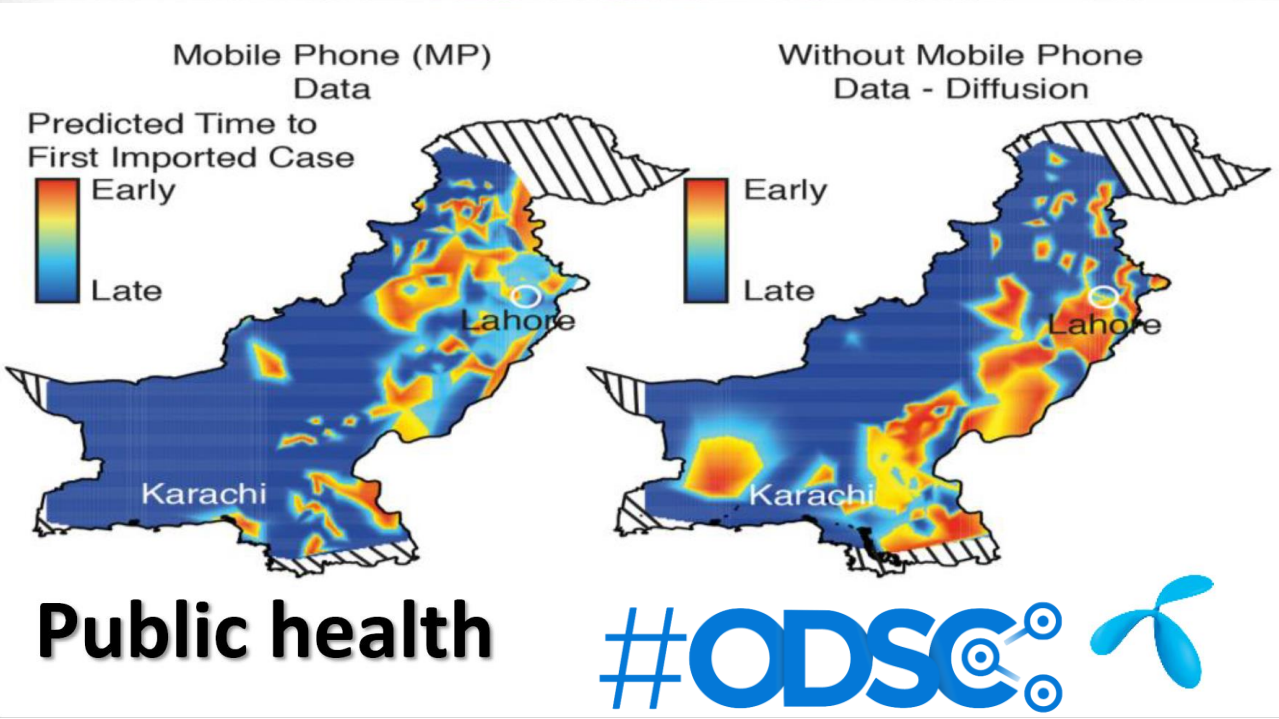
- 2.3 million people moving across Norway in a period of 12 hours
- 1.5 TB of signaling data in a snapshot
- Telenor empowering Norwegian society through big data innovation




Tourism


Traffic monitoring


Public transportation


Mobile Phone (MP) Data
Predicted Time to First Imported Case
Early
Late
Lahore
Karachi
Without Mobile Phone Data - Diffusion
Early
Late
Lahore
Karachi
Public health
#ODSC

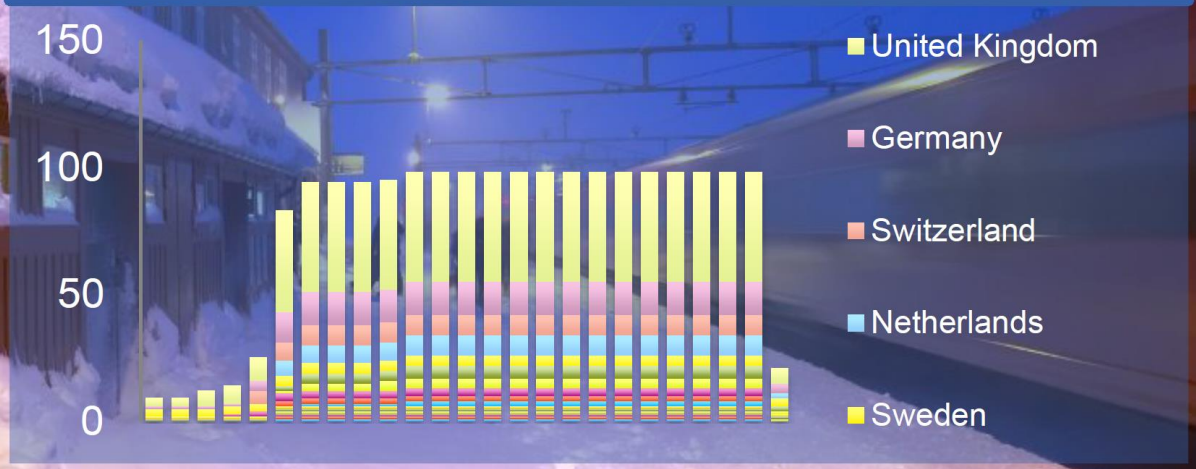
Telenor discovering foreign tourists at Finse

33,850 foreign travelers from 53 countries during February'17

Discovered the top nationalities of foreign travelers

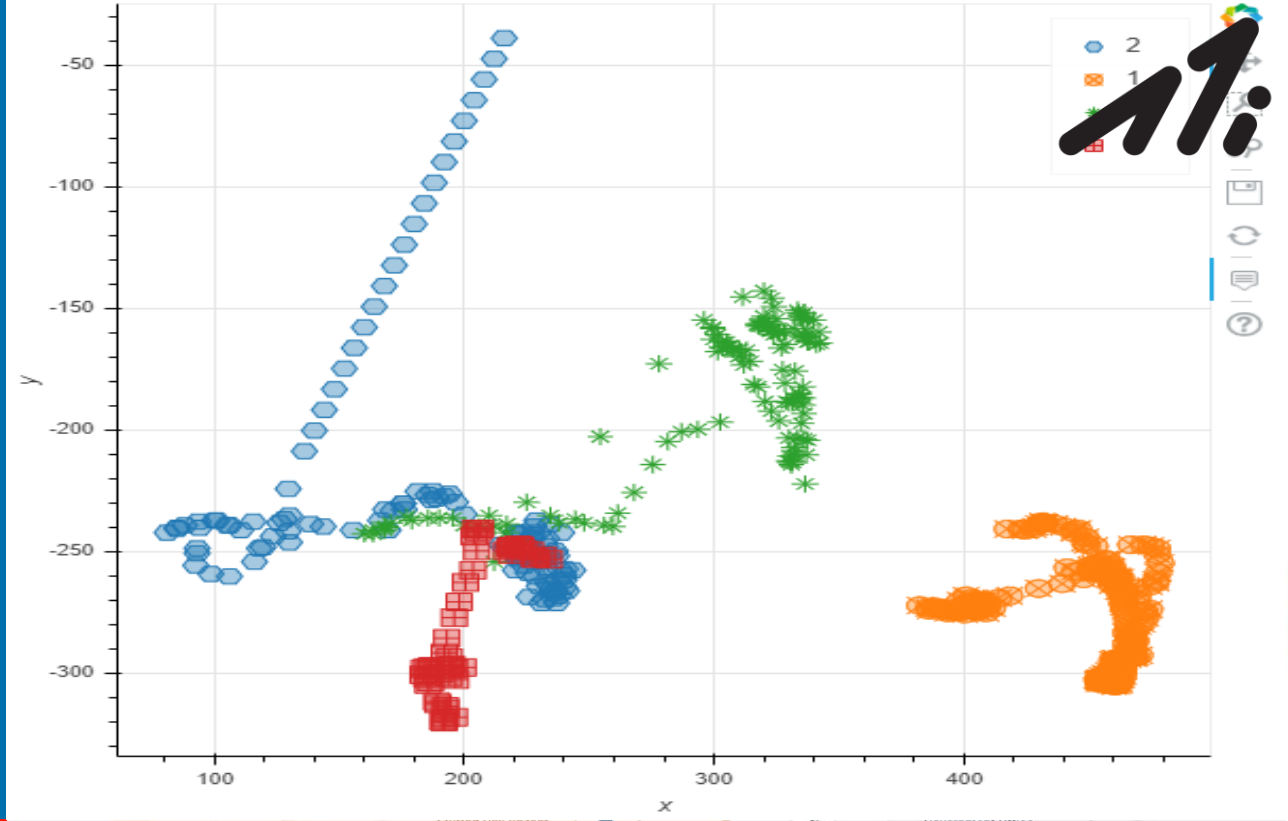
Discovered preferences on time of travel of foreign passengers

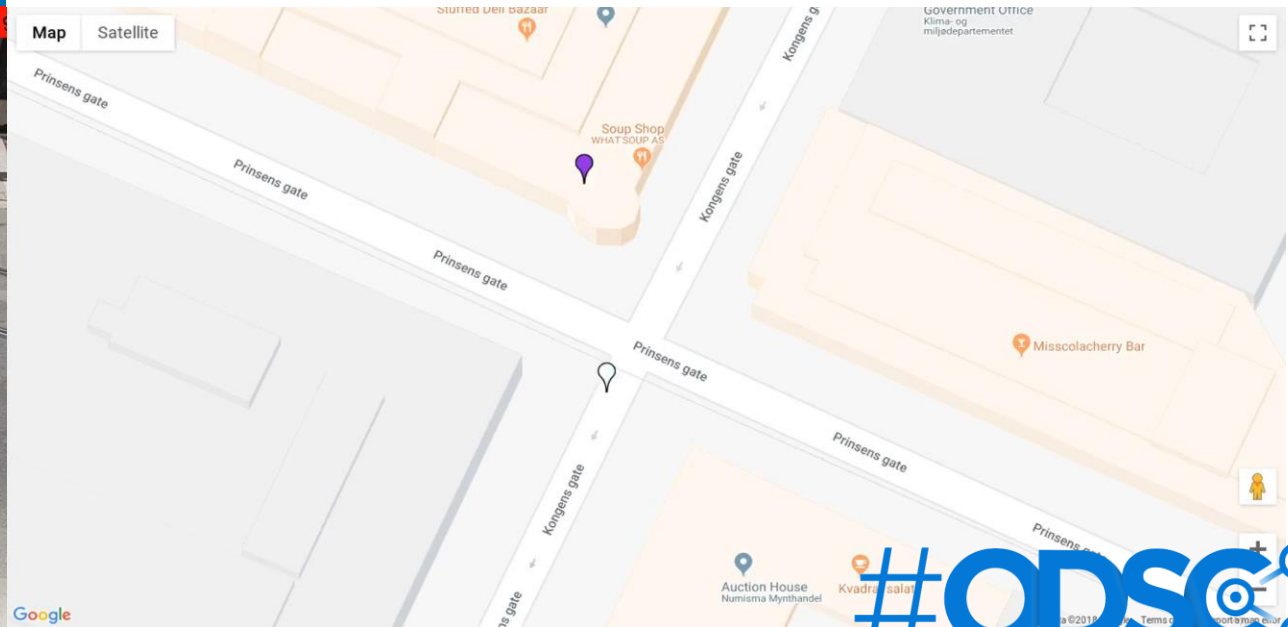
From signaling data: measure a train arriving to Finse



#ODSC

Mobility Analytics and trends in Artificial Intelligence





Mobility Analytics in Norway

Telenor's MA

- Telco location data
- Don't look for active user consent
- Can analyze data only at aggregated level
- Subject to GDPR regulations

Telia's Crowd Insights

- Telco location data
- Don't look for active user consent
- Can analyze data only at aggregated level
- Subject to GDPR regulations

Unacast's Real World Graph

- GPS location data from Mobile phones
- Not subject to GDPR
- Operating in USA
- Can analyze data at individual level

Consulting

- Helping location data generators develop their Mobility Analytics service
- Helping gov orgs develop MA solutions
- Smart city platform
- Self-driving buss

The privacy challenge

- Lack of anonymization
- Privacy breaches
- Unethical actions
- Improper data masking

GDPR regulations

- Adopted in 2016
- Came into force in July 2018
- Fines up to 20,000,000 EU or 4% of global turnover

Art. 6 GDPR
Lawfulness of processing

1. Processing shall be lawful only if and to the extent that at least one of the following applies:

(a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;

(b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;

(c) processing is necessary for compliance with a legal obligation to which the controller is subject;

(d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;

(e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

(f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Point (f) of the first subparagraph shall not apply to processing carried out by public authorities in the performance of their tasks.

#ODSC

Privacy Enhancing Technologies

Encryption

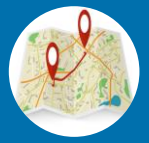
For security reasons data should be encrypted. Both data on transit and at rest. This imposes severe penalties in performance

Masking

Masking techniques, such as hashing can be used instead of encryption, they might be a good compromise between security and performance

Extrapolation

In some cases, it is possible to extrapolate from a customer base to a population estimate. This introduces uncertainties but increases protection of privacy



Path obfuscation

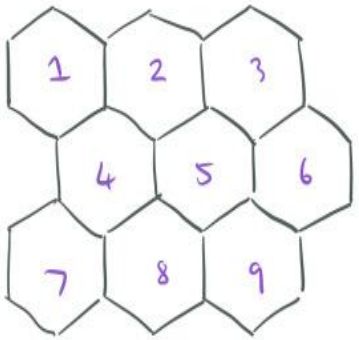
By introducing pseudo-random noise, we can further protect privacy from the risk of re-identification by inference. This will decrease data quality

Aggregation algorithms

Aggregation techniques such as k-anonymity can strengthen privacy frameworkds by avoiding exposure of individuals. See l-variety, t-closeness...

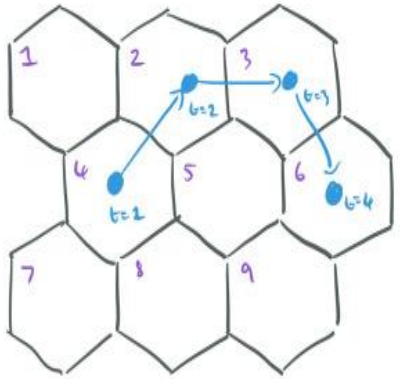
Recovering trajectories from Ash

Spatial aggregation - Tesselation



Base stations (locations) M=9

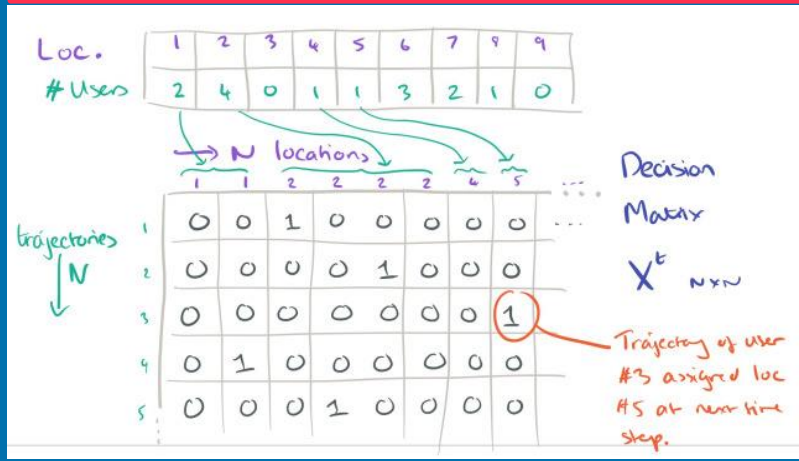
Source – Individual trajectories



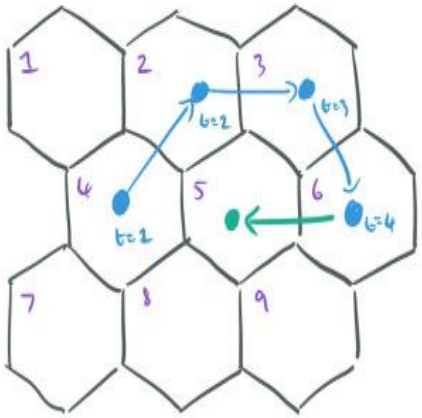
Example user trajectory at t=4 for user i

$S_i^t = [4, 2, 3, 6]$

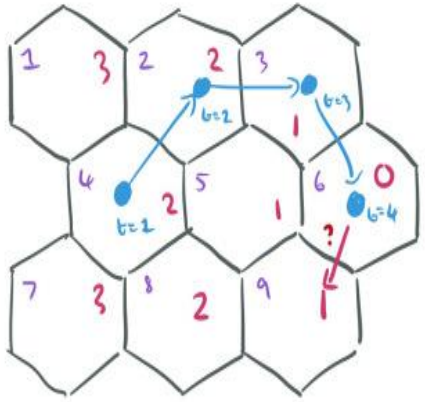
Occupancy matrix



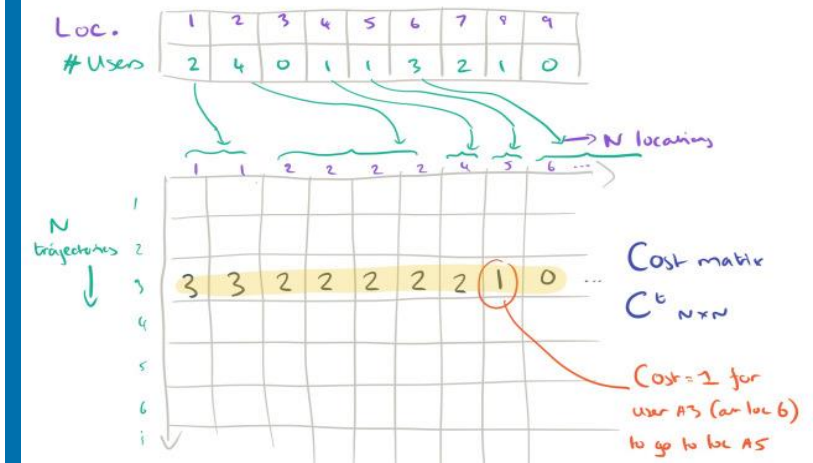
What happens after *i* time steps?



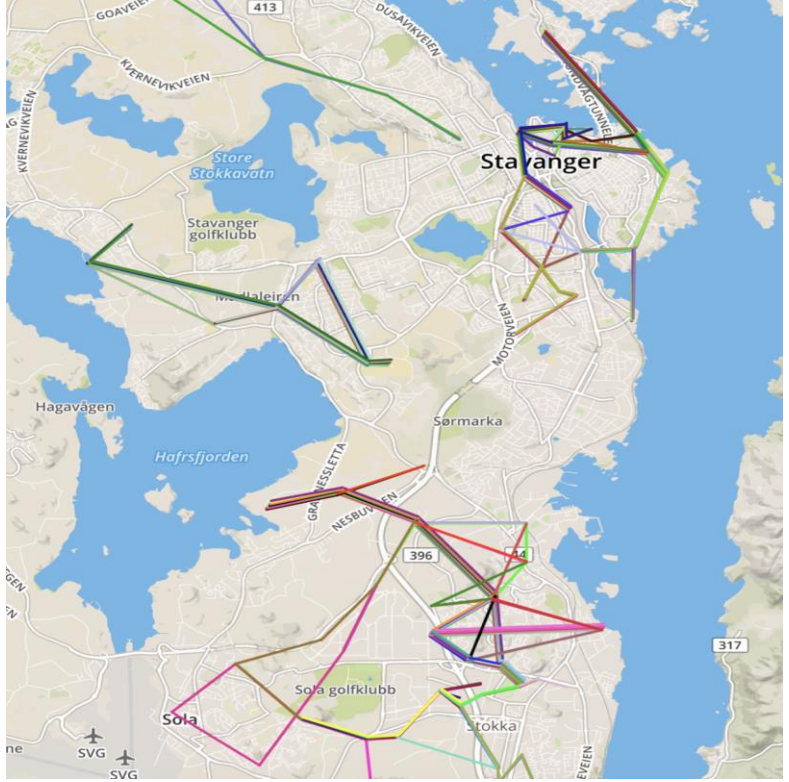
The cost of moving



Cost matrix



Testing of a privacy framework



Extremelly dificult to recover trajectories!

- Took measurements of mobility in Stavanger
- Around 100,000 individual trajectories
- Used a very good privacy framework
- Just a few trajectories were recovered

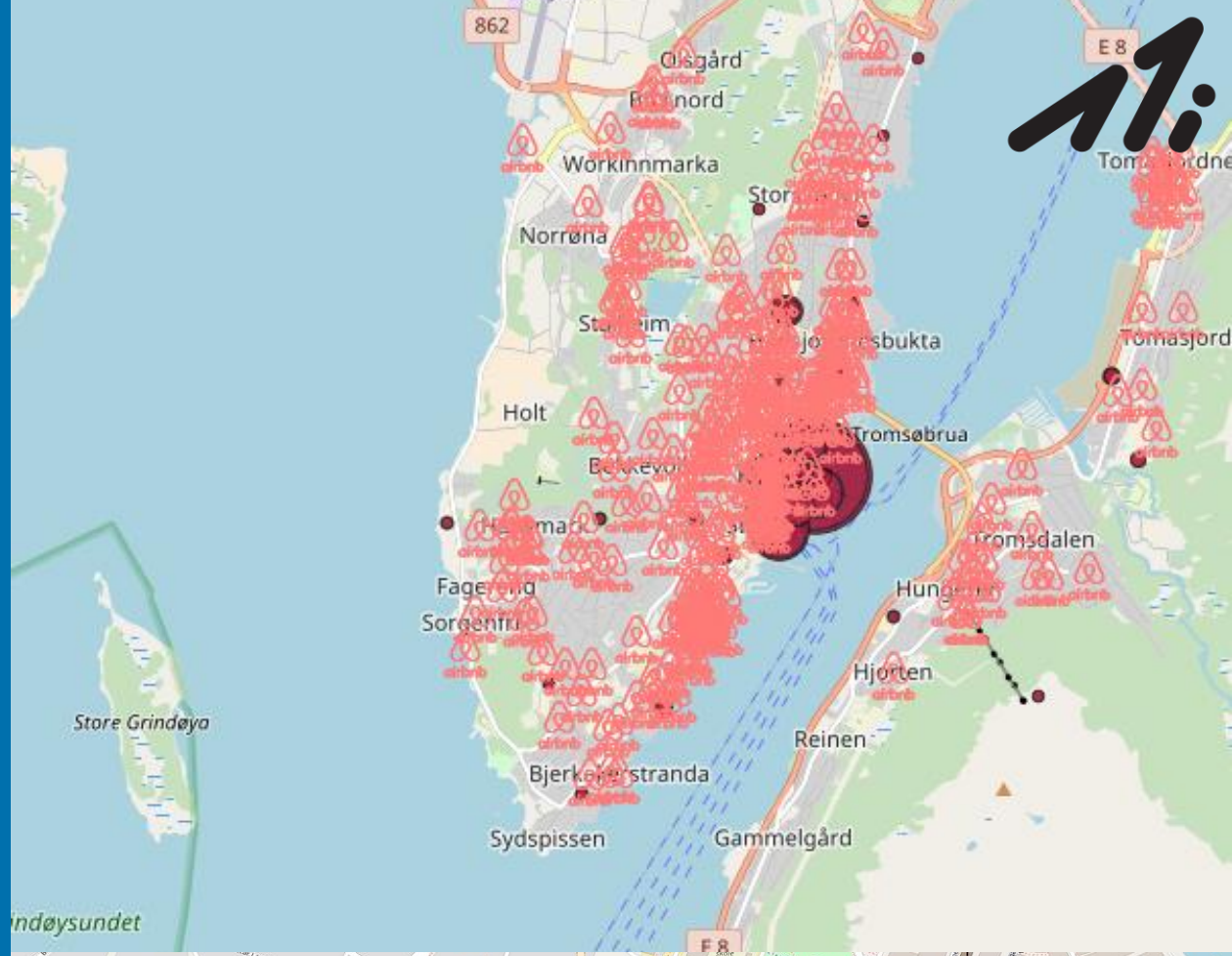
#ODSC

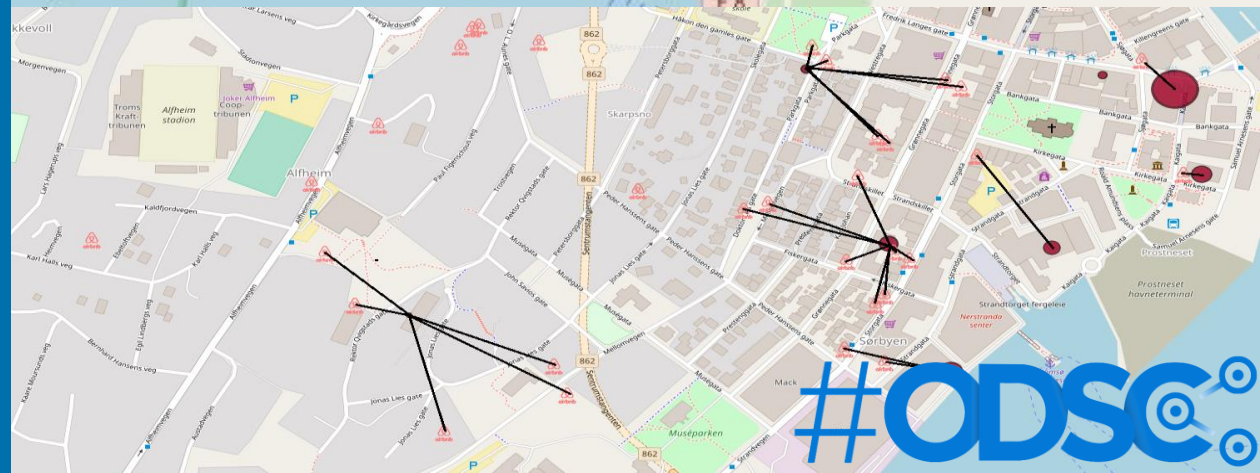
Data enrichment: AirBnB in Tromsø

- 300 housing offers listed
- Capacity for 884 persons (Largest ‘house’ for 16 people!)
- Potential for revenue per night: 225.655 NOK
- Average guest satisfaction rating: 96,47/100
- 5838 reviews available to perform sentiment analysis

Tourist preferences AirBnB Vs Hotels

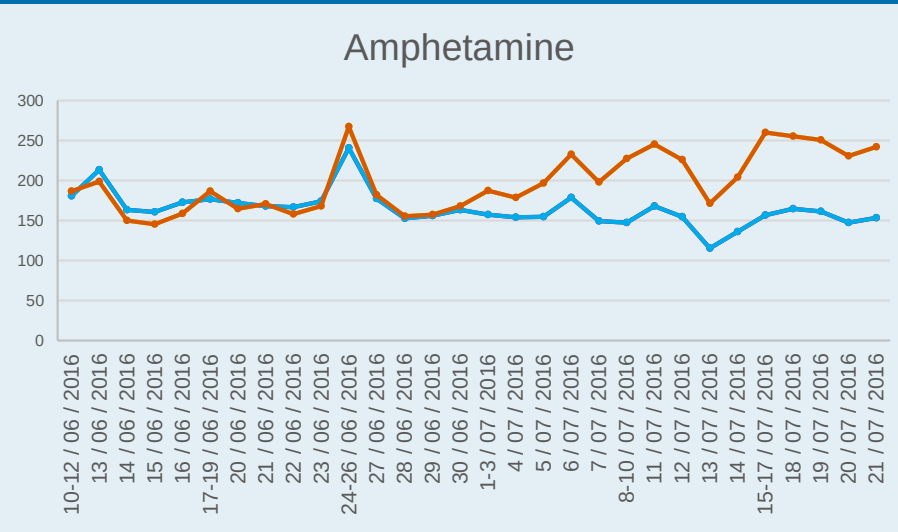
- We took the 300 AirBnB listings and calculated the closest base station (out of 96 base stations in Tromsø) providing cellular coverage for AirBnB
- We took the presence of tourists in these basestations between 03:00 and 04:00 (overnight)





Data enrichment: Mobility meets wastewater...

- Measured the mobility behaviour of people in Oslo
- Correlated population dynamics with drug consumption measurements
- Discovered the best time of the year to run anti-drug campaigns



Environmental Science & Technology

Article

Use of Mobile Device Data To Better Estimate Dynamic Population Size for Wastewater-Based Epidemiology

Kevin V. Thomas,*[,†,‡] Arturo Amador,[§] Jose Antonio Baz-Lomba,[†] and Malcolm Reid[†]

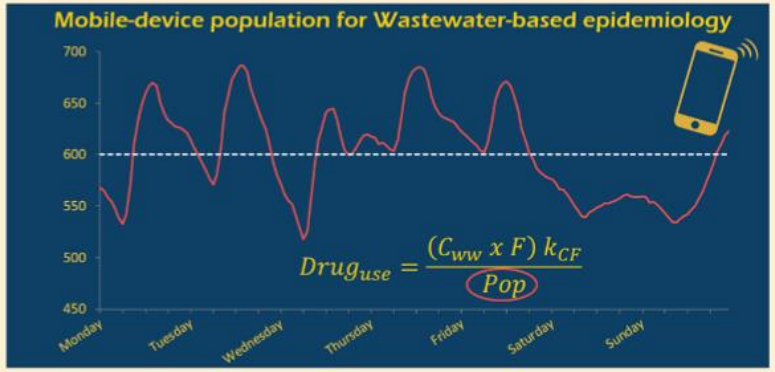
[†]Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, NO-0349 Oslo, Norway

[‡]Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 39 Kessels Road, Coopers Plains, Queensland 4108, Australia

[§]Telenor ASA, Snarøyveien 30, NO-1360 Fornebu, Norway

Supporting Information

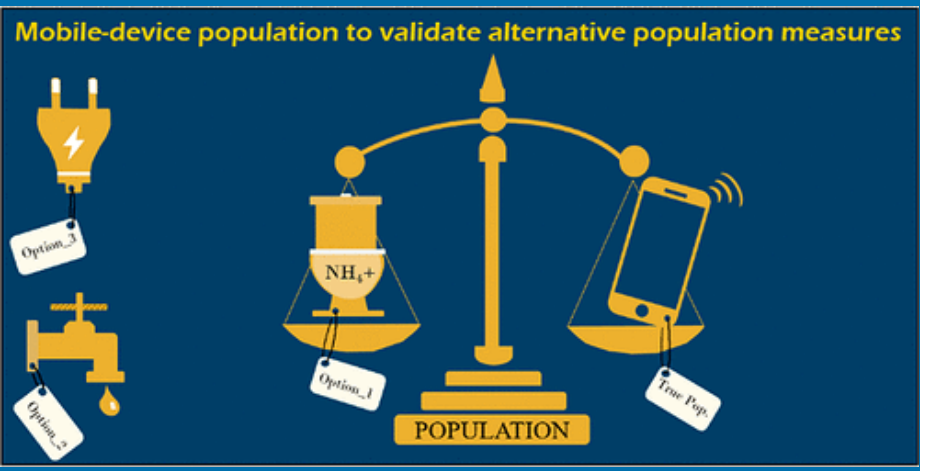
ABSTRACT: Wastewater-based epidemiology is an established approach for quantifying community drug use and has recently been applied to estimate population exposure to contaminants such as pesticides and phthalate plasticizers. A major source of uncertainty in the population weighted biomarker loads generated is related to estimating the number of people present in a sewer catchment at the time of sample collection. Here, the population quantified from mobile device-based population activity patterns was used to provide dynamic population normalized loads of illicit drugs and pharmaceuticals during a known period of high net fluctuation in the catchment population. Mobile device-based population activity patterns have for the first time quantified the high degree of intraday, week, and month variability within a specific sewer catchment. Dynamic population normalization showed that per capita pharmaceutical use remained unchanged during the period when static normalization would have indicated an average reduction of up to 31%. Per capita illicit drug use increased significantly during the monitoring period, an observation that was only possible to measure using dynamic population normalization. The study quantitatively confirms previous assessments that population estimates can account for uncertainties of up to 55% in static normalized data. Mobile device-based population activity patterns allow for dynamic [illegible] much improved temporal and spatial trend analysis.

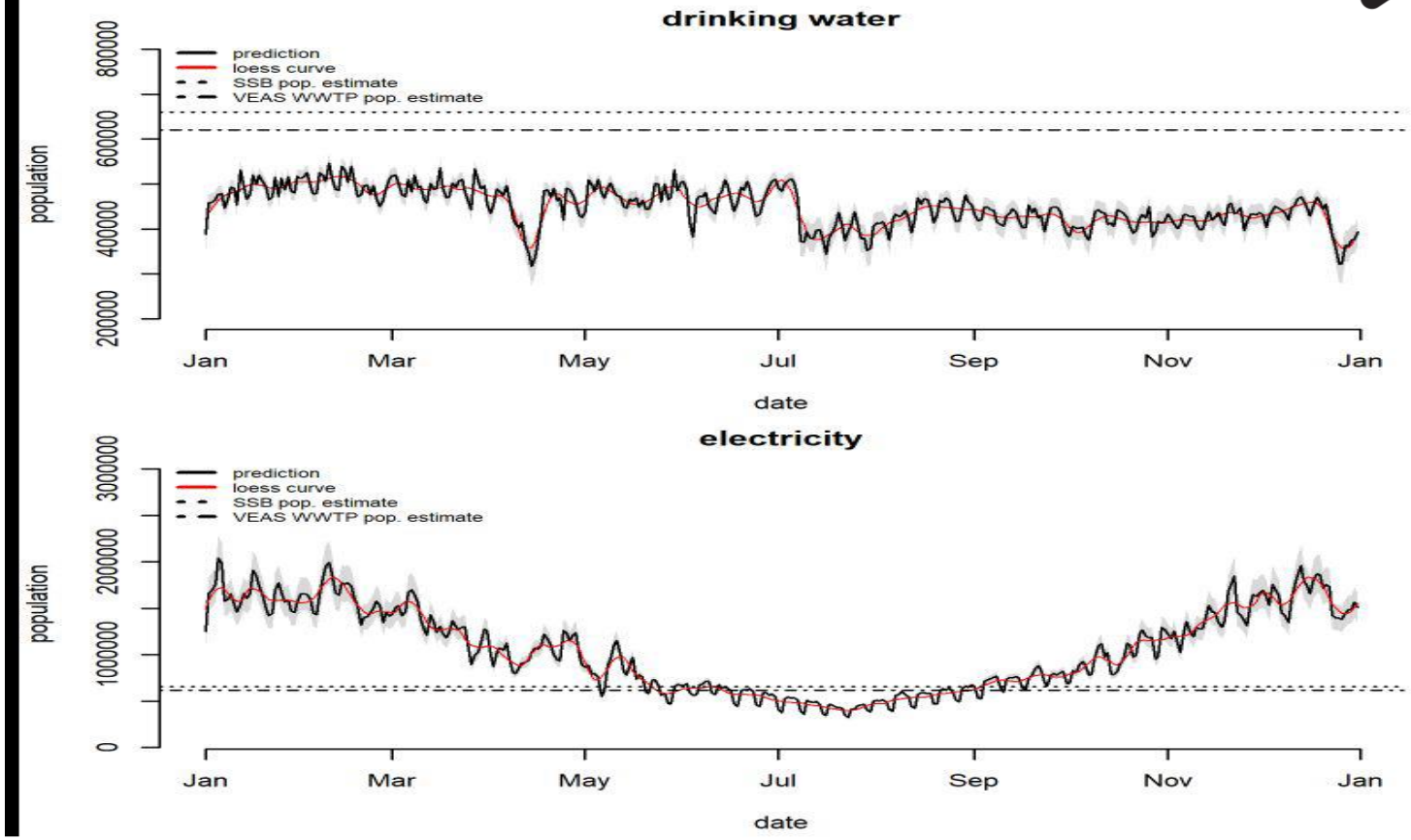


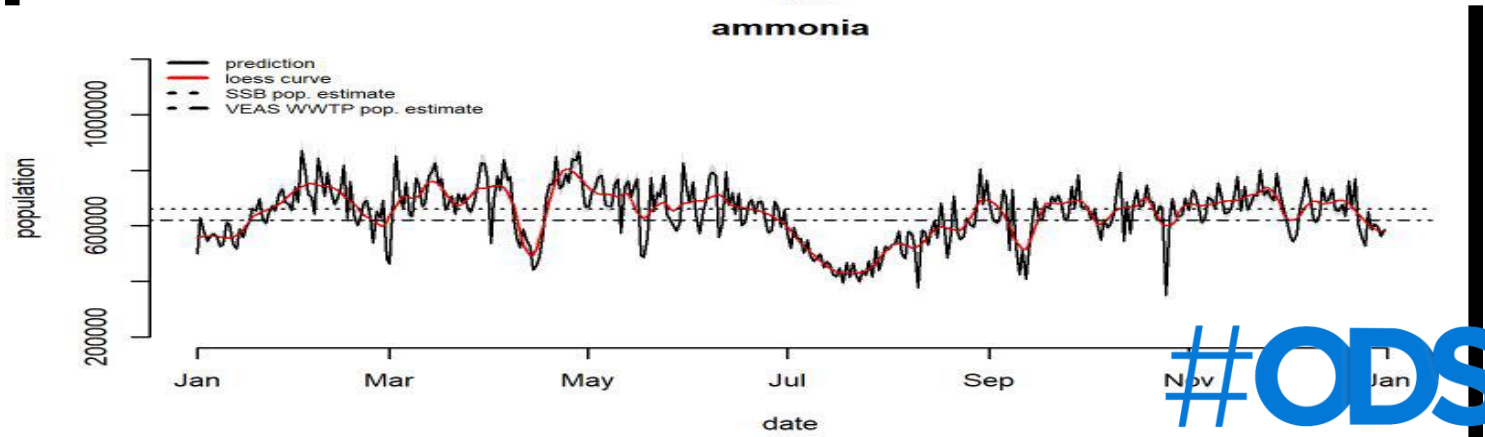
#ODSC

Assessing Alternative Population Size Proxies in a Wastewater Catchment Area Using Mobile Device Data

- Measured the mobility behaviour of people in Oslo
- Correlated population dynamics with drug consumption measurements
- Discovered the best time of the year to run anti-drug campaigns

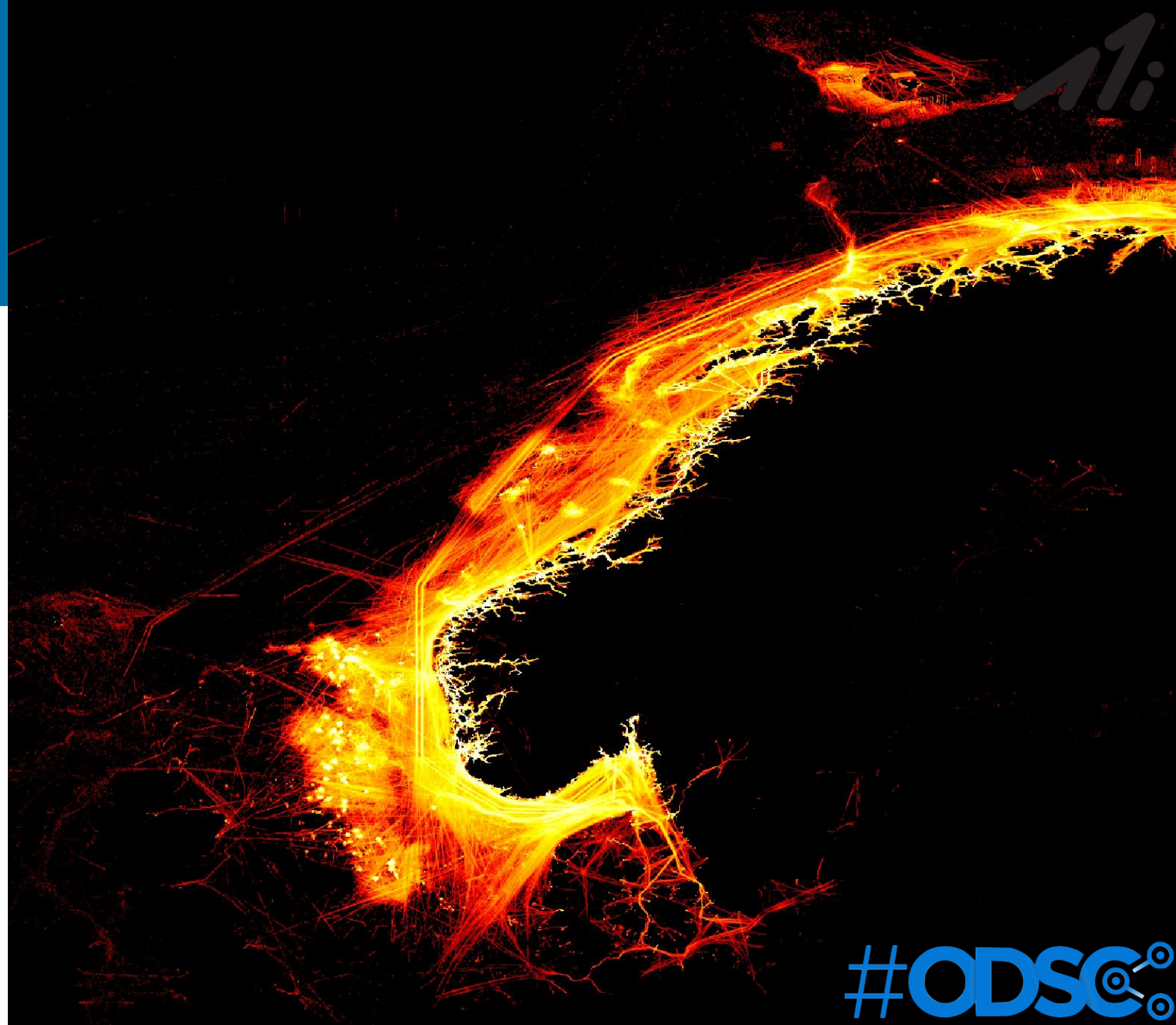




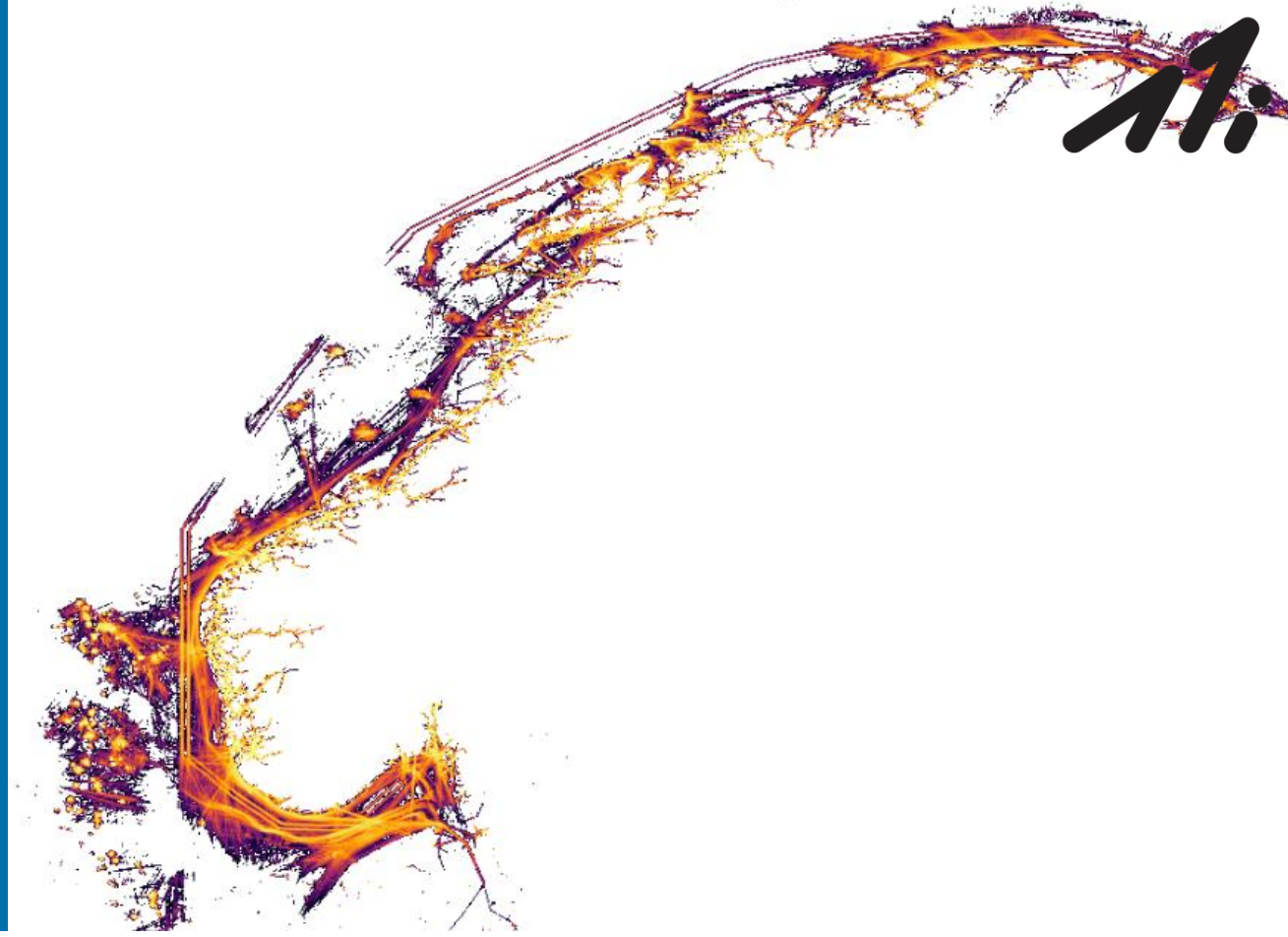


Big Data technologies for intelligent maritime navigation

- Maritime transportation provides the best method to transport goods over large distances
- The overall volume of trade by this means is steadily growing every year
- Concerns in maritime safety and security are growing together with the industry!



Analytics for intelligent maritime navigation



Port demand prediction

- Use RNN to predict port demands/saturation
- Given enough historical data, we can utilize deep learning for route prediction

Travel time and distance

- Spatio Temporal-Neural Networks offer a framework to predict travel distances between ports
- Travel time can be predicted (Time of day, atmospheric conditions, etc.)

Route optimization

- Deep Reinforcement Learning. By observing reward signals and following feasibility ru

#ODSC