EluciDATA
Innovation Lab

Data-driven innovation beyond the hype

Guest lecture Odisee | May 16th, 2018

Dr. Mathias Verbeke

sirris
driving industry by technology

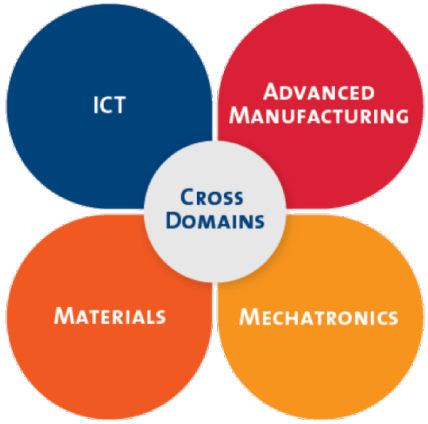
- **Collective research centre** of the Belgian technology industry
  - Non-profit organisation
  - Industry owned, +2500 member companies

MISSION

"To increase the competitiveness of companies through technological innovations"

SOFTWARE ENGINEERING
CLOUD COMPUTING
DATA INNOVATION

METALS & COMPOSITES
PLASTICS & HYBRIDS
COATINGS
NANOMATERIALS

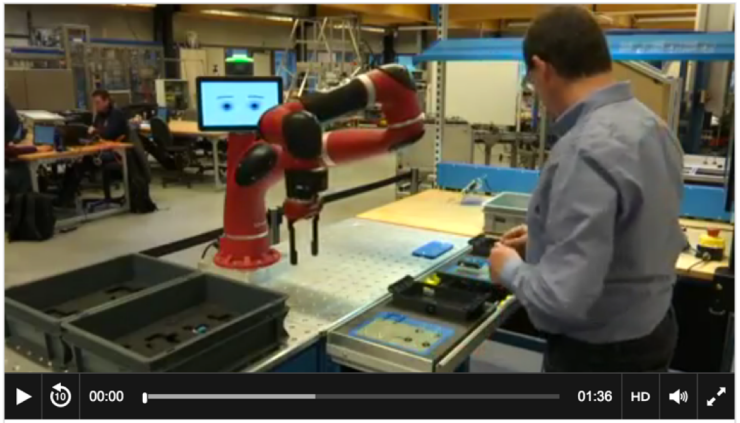


FACTORIES OF THE FUTURE
WORLD CLASS TECHNOLOGIES
ADDITIVE MANUFACTURING

ECO-MECHATRONICS
SENSORIZED FUTURE
MODEL BASED DESIGN

You might know Sirris from…

First Sawyer Cobot in the Benelux just arrived in our Smart & Digital Factory application lab in Kortrijk [video]



Een primeur in de Benelux: werknemer krijgt een "derde arm"

Primeur voor de Benelux in Kortrijk: daar staat de eerste Sawyer Cobot, een éénarmige robot die er niet is om werknemers te vervangen, maar om ze erg laagdrempelig te helpen bij hun werk. Een derde arm zeg maar. Die Sawyer Cobot staat bij Sirris, een centrum dat opgericht is door de Belgische hoogtechnologische industrie. Dit project loopt samen met de Provinciale Ontwikkelingsmaatschappij West-Vlaanderen en bedrijven kunnen er komen kijken en testen of dergelijke technologie iets voor hen is.

WETENSCHAP di 25/04/2017 - 18:05

Europe's largest climate chamber

EluciDATA

- **Collective research centre** of the Belgian technology industry
  - Non-profit organisation
  - Industry owned, +2500 member companies

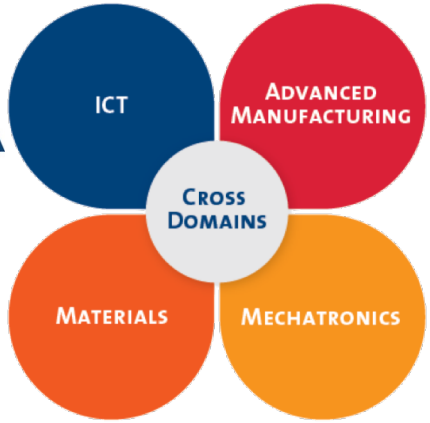
MISSION

"To increase the competitiveness of companies through technological innovations"

SOFTWARE ENGINEERING
CLOUD COMPUTING
DATA INNOVATION

EluciDATA Innovation Lab

METALS & COMPOSITES
PLASTICS & HYBRIDS
COATINGS
NANOMATERIALS



FACTORIES OF THE FUTURE
WORLD CLASS TECHNOLOGIES
ADDITIVE MANUFACTURING

ECO-MECHATRONICS
SENSORIZED FUTURE
MODEL BASED DESIGN

EluciDATA Innovation Lab Profile

MISSION

to support industry in realizing product & service innovation by facilitating the exploitation of real-world data by means of data science

KEY CHARACTERISTICS

- Focus on complex and rich data → few 'big data' applications
- Domain-agnostic → energy, manufacturing, transport, ...
- Problem-oriented → no technology push

APPROACH

- Industrial services
- R&D projects
- Individual and collective knowledge transfer

The Data Innovation Team

Elena Tsiporkova *Team Leader Data Innovation*

Tom Tourwé *Technical Lead Data Innovation*

Mathias Verbeke *Data Innovation Scientist*

Caroline Mair *Project Manager Data Innovation*

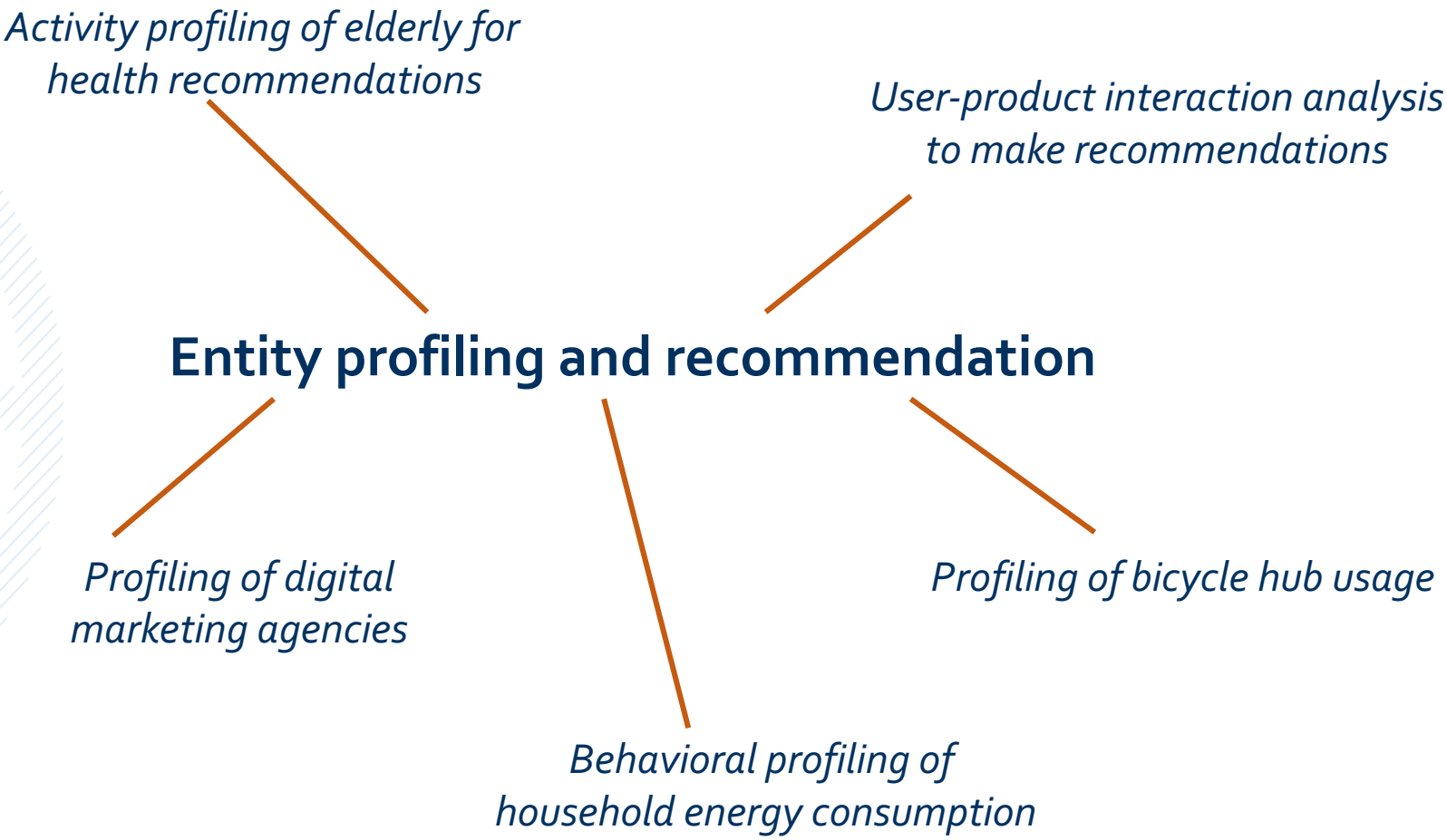
Pierre Dagnely *Data Innovation Scientist*

Nicolás González-Deleito *Project Leader Data Innovation*

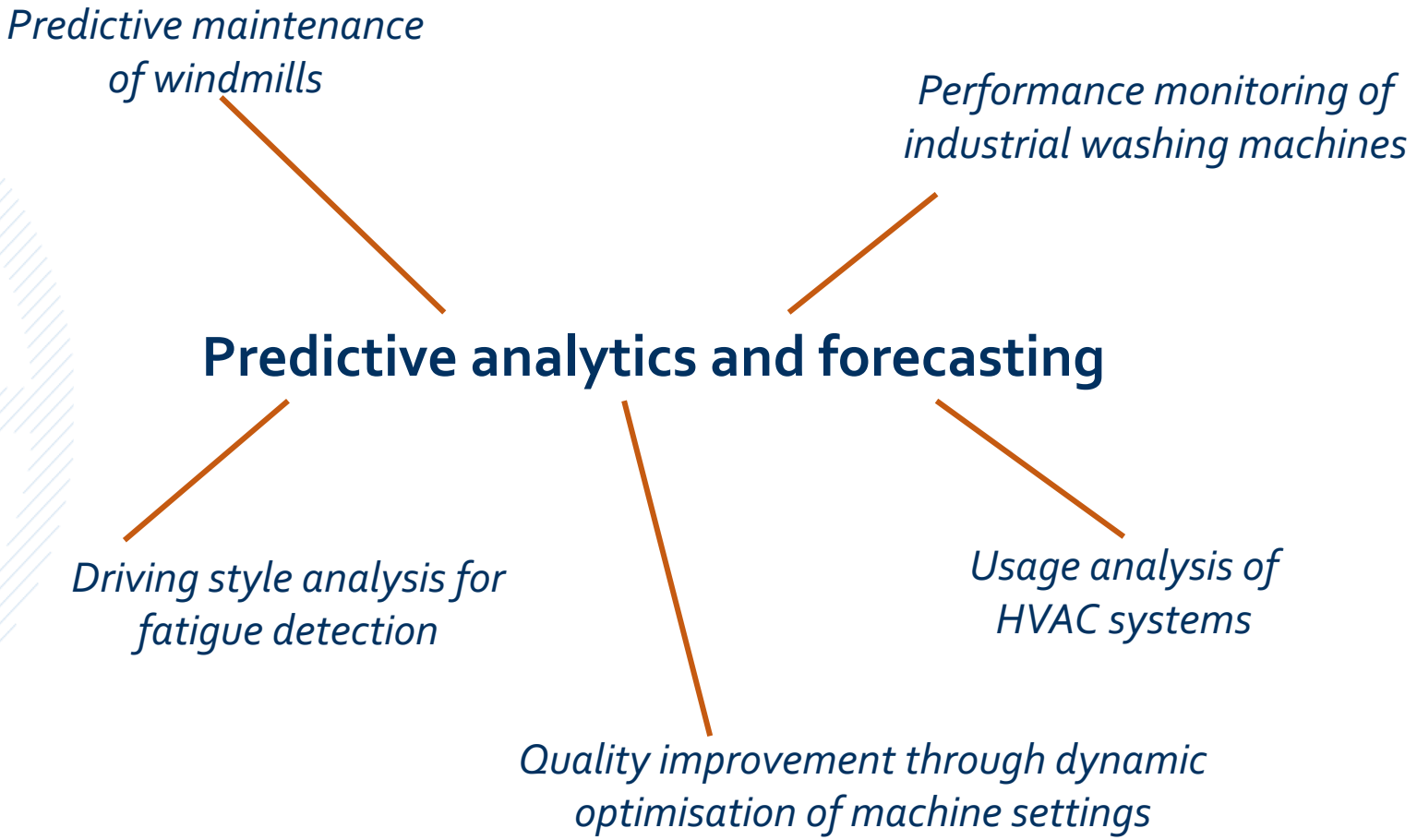
Andriy Zubaliy *Data Innovation Scientist*

Alessandro Murgia *Data Innovation Scientist*

Entity profiling and recommendation pilot projects



Predictive analytics and forecasting pilot projects

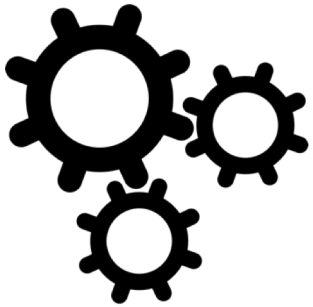


while HW/SW design remain important,

it's all about data

storage

algorithms

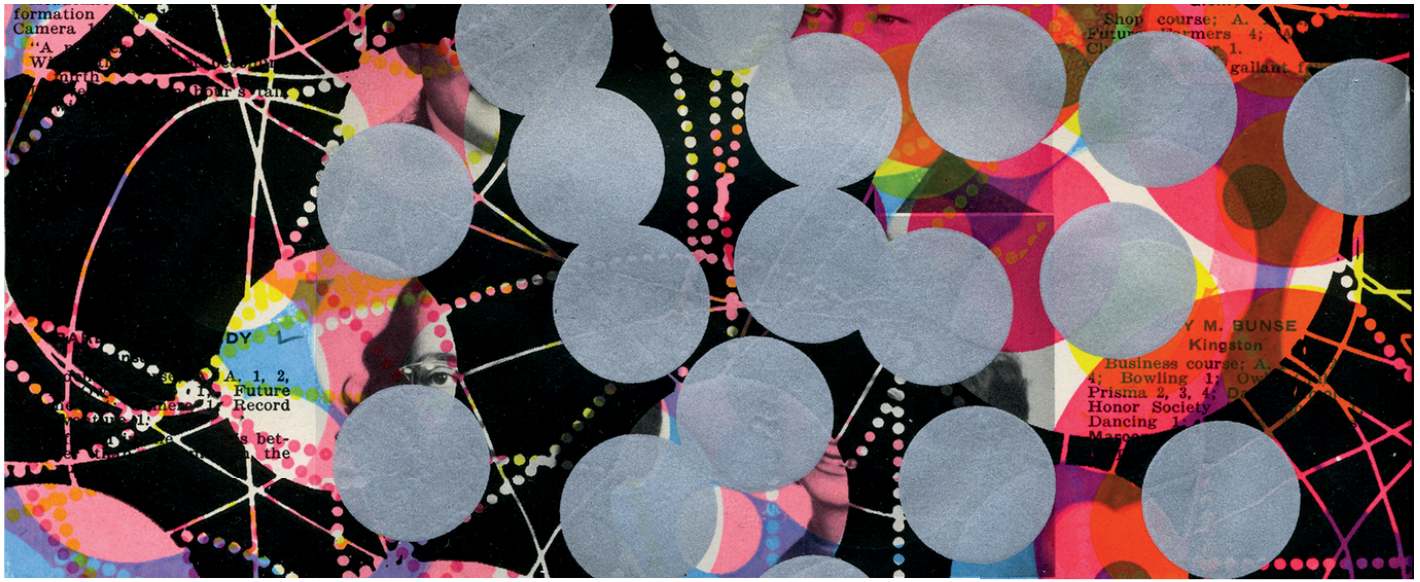


trends & patterns

privacy & security

uncovering insights

Harvard Business Review



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

The rise of data *innovation*

= The exploitation of data in industry by means of data science to realise product/service innovation (for competitive advantage)

- Previously, datasets could be analysed manually by teams of statisticians, modellers and analysts.

  Currently, this has been made impossible due to the complexity and the volume of the data that is being generated.

- Why now?
  - Computers have become far more powerful
  - Networking has become ubiquitous
  - Algorithms have been developed for broader and deeper analyses than previously possible

Applications in different domains

Healthcare and life sciences
- Lifestyle analysis (wearables)
- Bioinformatics (DNA sequencing,etc.)

Finance
- High-frequency trading

Retail
- Demand/supply forecasting
- Customer profiling & behavioral analysis
- Decision support

Environmental & Mobility
- Impact analysis
- Traffic management

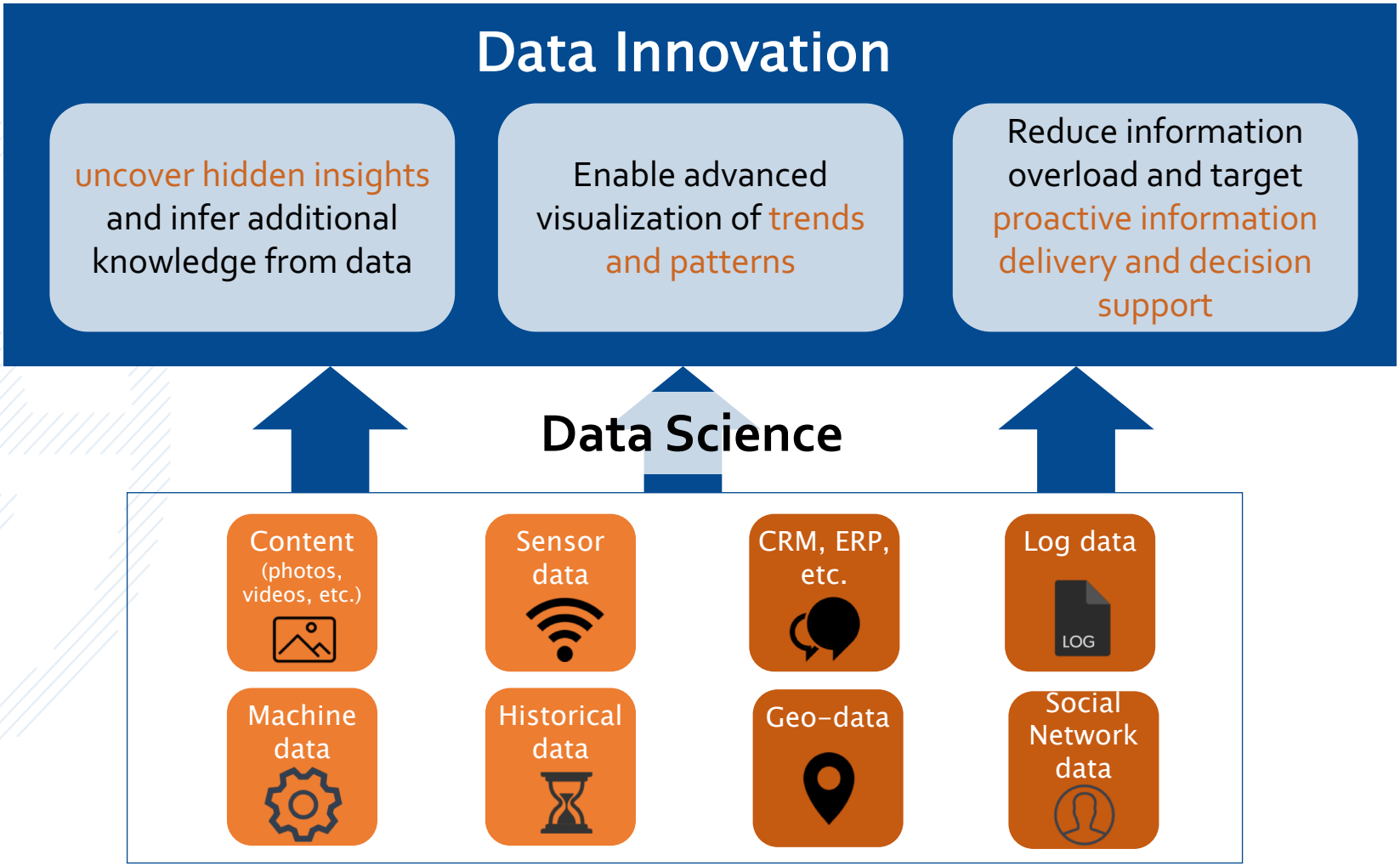
Media
- Scene analysis
- Image understanding

Manufacturing
- Predictive maintenance
- Automation (Factory of the Future)

Infrastructure
- Load balancing
- Anomaly detection

Social Networking
- Sentiment analysis
- Influence detection

New opportunities
for value creation and competitive advantage



Data intelligence confirms left isn't right

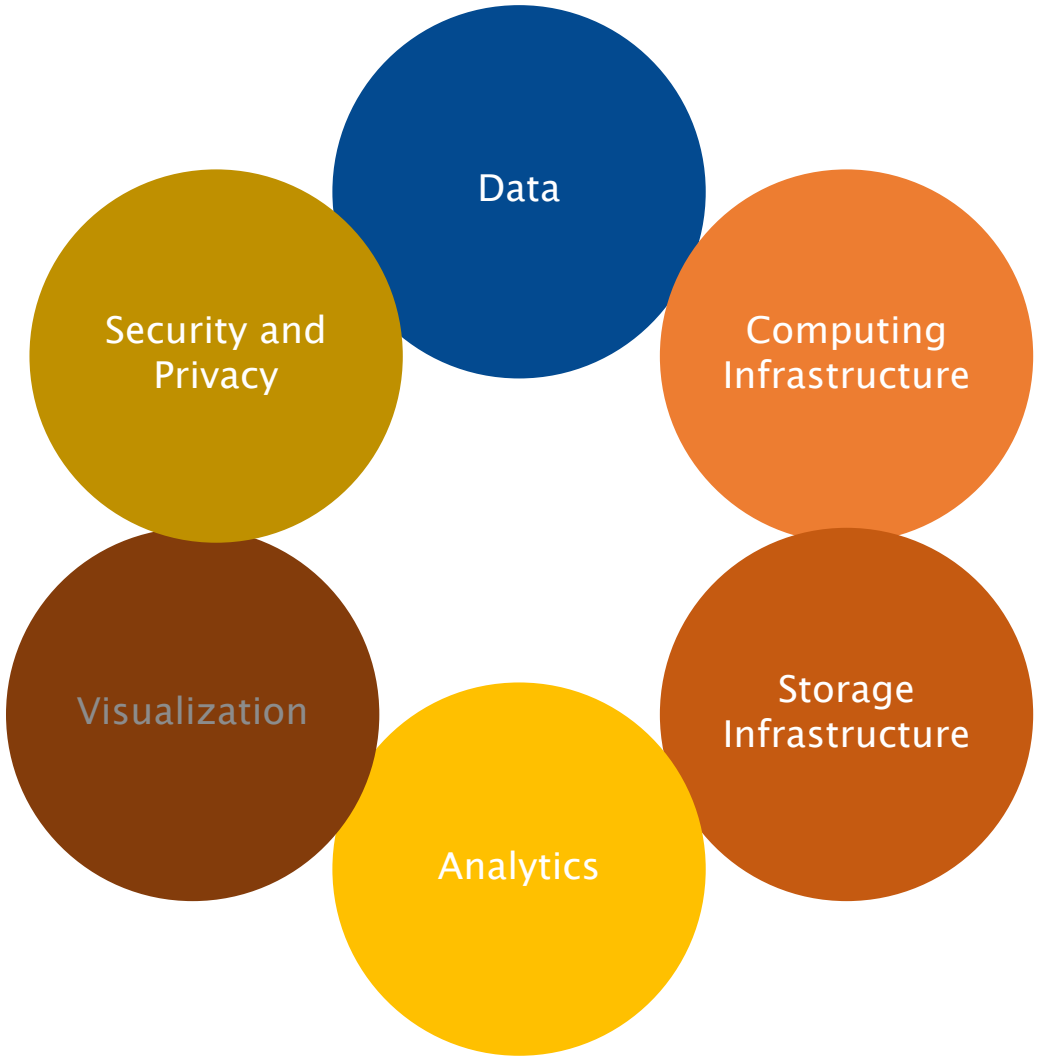
Most UPS managers have been UPS drivers and they knew that left turns mean

- idling (longer route times)
- going against traffic (less safe)

UPS optimized algorithmically delivery routes by aiming at avoiding left turns

Resulted in saving 38 million liter fuel per year, equivalent to reduction of 20.000 ton CO_2

… but also new challenges



The Art of Data Science

Statistics

collecting and analysing numerical data

Databases

Efficiently organise, store and retrieve data

Distributed systems

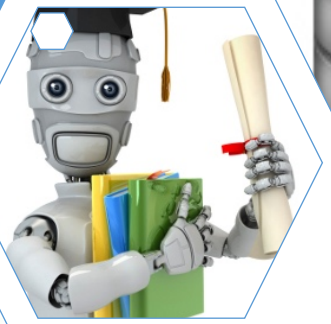
Index data from multiple machines

Machine Learning

Algorithms to learn from data

Data Mining

Search patterns in large amounts of data



Data innovation is *not* …

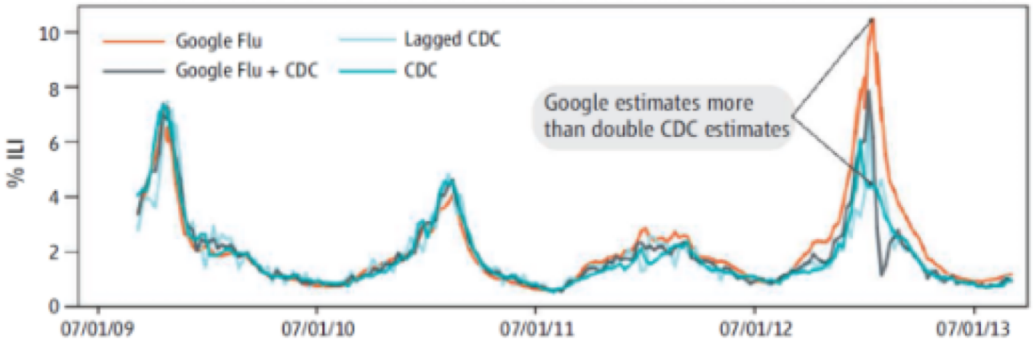
- Just data crunching → business understanding
  - Even the most intelligent algorithm will be useless without some *prior knowledge of the application domain*
- About volume → data understanding
  - Even relatively small datasets can be very complex to understand and handle
  - *Having access to the right data with the right quality is essential*
- About installing & configuring platforms, libraries or tools → data modelling & analysis
  - It's about statistics, machine learning, experimentation, trial & error, …
- Easily scalable in terms of effort and human resources → human capital
  - Few data problems can be solved with *off-the-shelf* products and IT generalists
  - Most data problems require a handcrafted approach and data scientists with deep expertise in statistics, machine learning, …

Google Flu Trends (GFT)

problematic marriage of big and small data

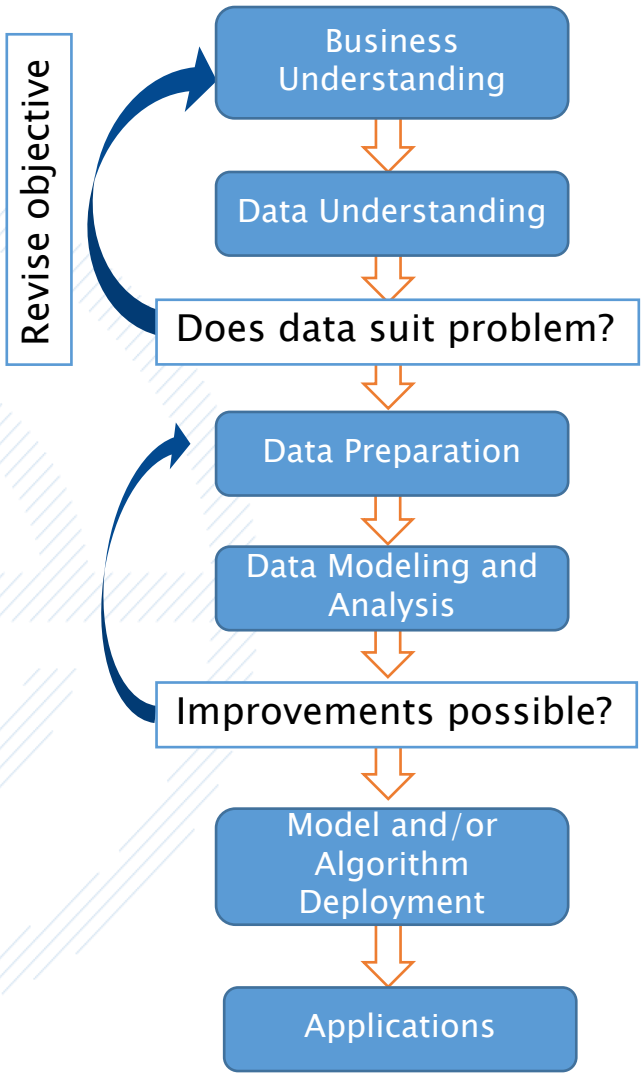
- In February 2013, Nature reported that GFT was predicting more than double the proportion of doctor visits for flu-like illness than the Centers for Disease Control and Prevention (CDC)
  - finds the best matches among 50 million search terms, which fit 1152 data points
  - high probability of finding search terms that point to flu susceptibility
  - but are structurally unrelated, and so do not predict the future

- Big Data "(Arrog/Ignor)ance"
  - Quantity of data does not mean that one can ignore foundational issues of measurement and valid data construction



Reference: http://www.sciencemag.org/content/343/6176/1203.full.pdf

The Data Science Workflow



- Analysis: what kind of task is this? Is it similar to some "standard" task

  => Business and data understanding

- Approach: How will I solve this task?

  => Data preparation and modeling

- Deployment: How to put my solution in production?

  => Deployment and applications

Unique problem vs. common solutions

- Each data science problem is unique

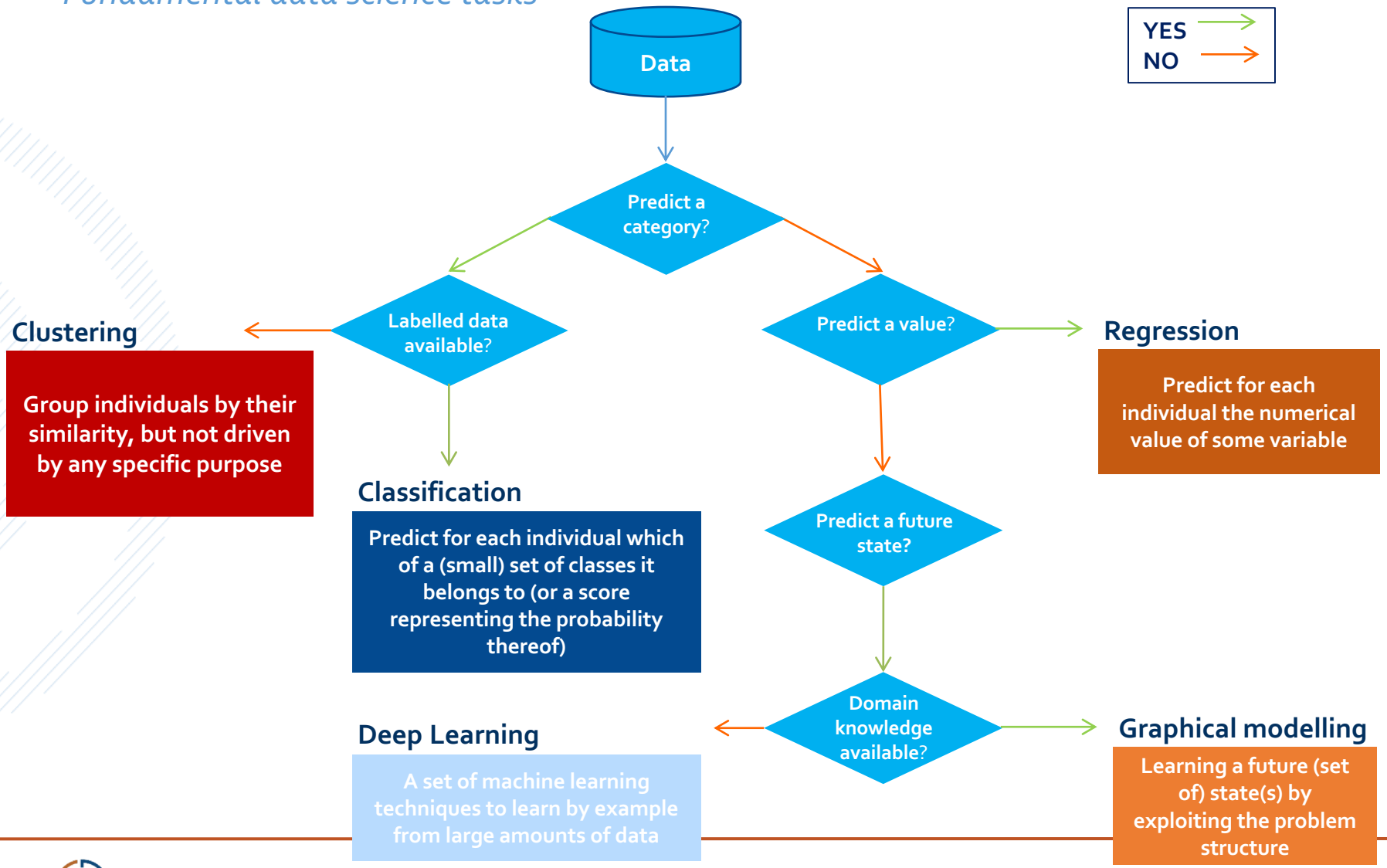
- However, there are only a handful of fundamental tasks underlying these problems

- Goal
  - Decompose the problem into subtasks
  - Identify unique subtasks and common DS tasks
  - Recompose solutions to the subtasks to obtain the overall solution

→ Decomposing and identifying these tasks is an essential data science skill

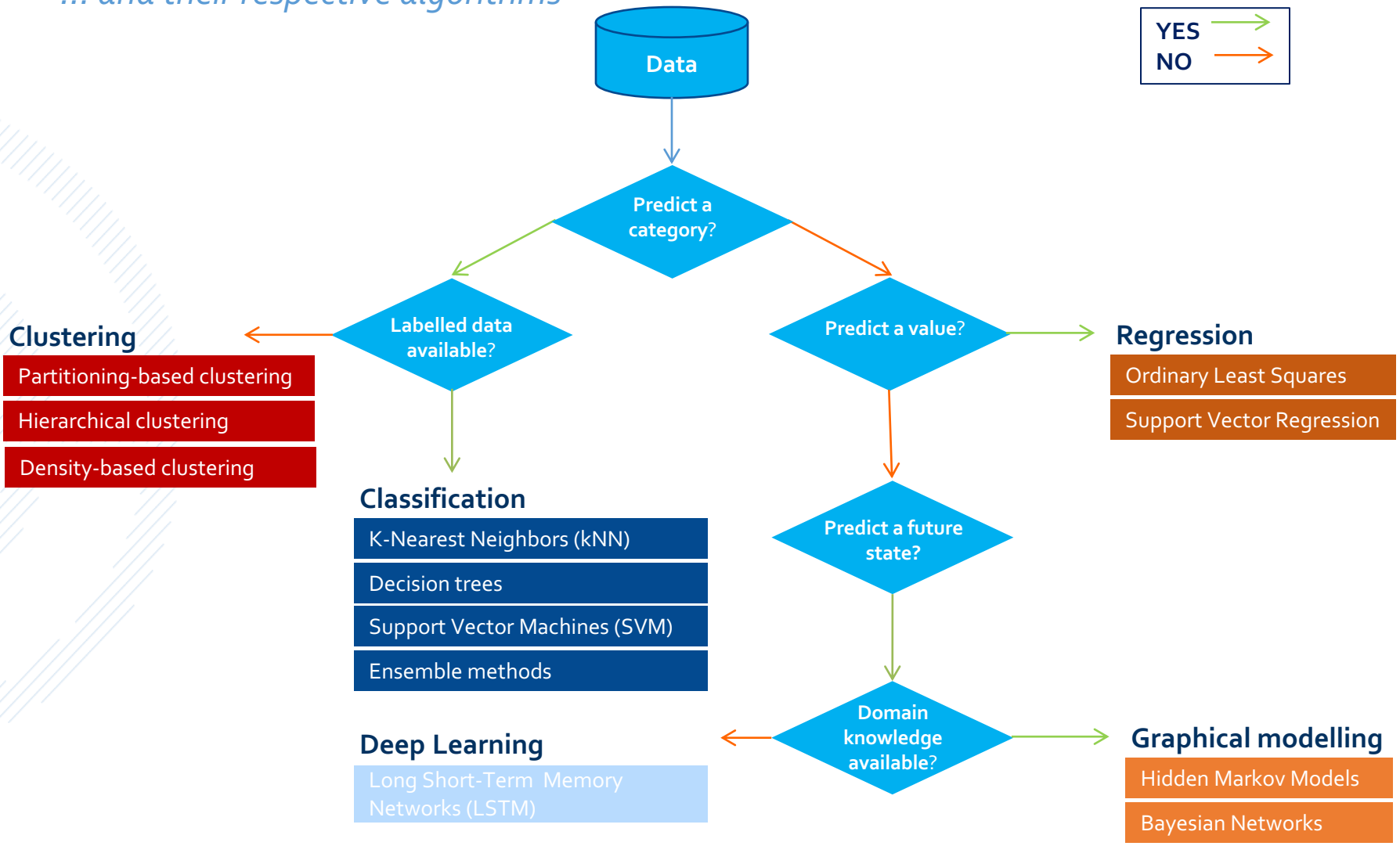
Selecting the right algorithm for the task

Fundamental data science tasks



Selecting the right algorithm for the task

... and their respective algorithms



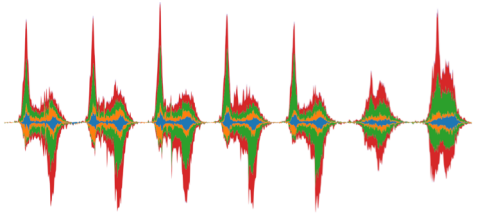
Data innovation is …

- a cross-disciplinary expertise
  - Analytical skills & system architecture expertise & domain knowledge & non-technical competences (privacy & legislation) & …
- more of an art than a science
  - Asking the right questions, selecting the right data, choosing the right algorithm for the right task, …
- a trial-and-error process for solving the unique data science problem at hand

4 example cases

1. Lifestyle monitoring & profiling

2. Traffic flow visualization



3. Spatio-temporal taxi usage

4. Product usage monitoring for smart, connected products

1st example use case

Lifestyle monitoring & profiling

- Business objective

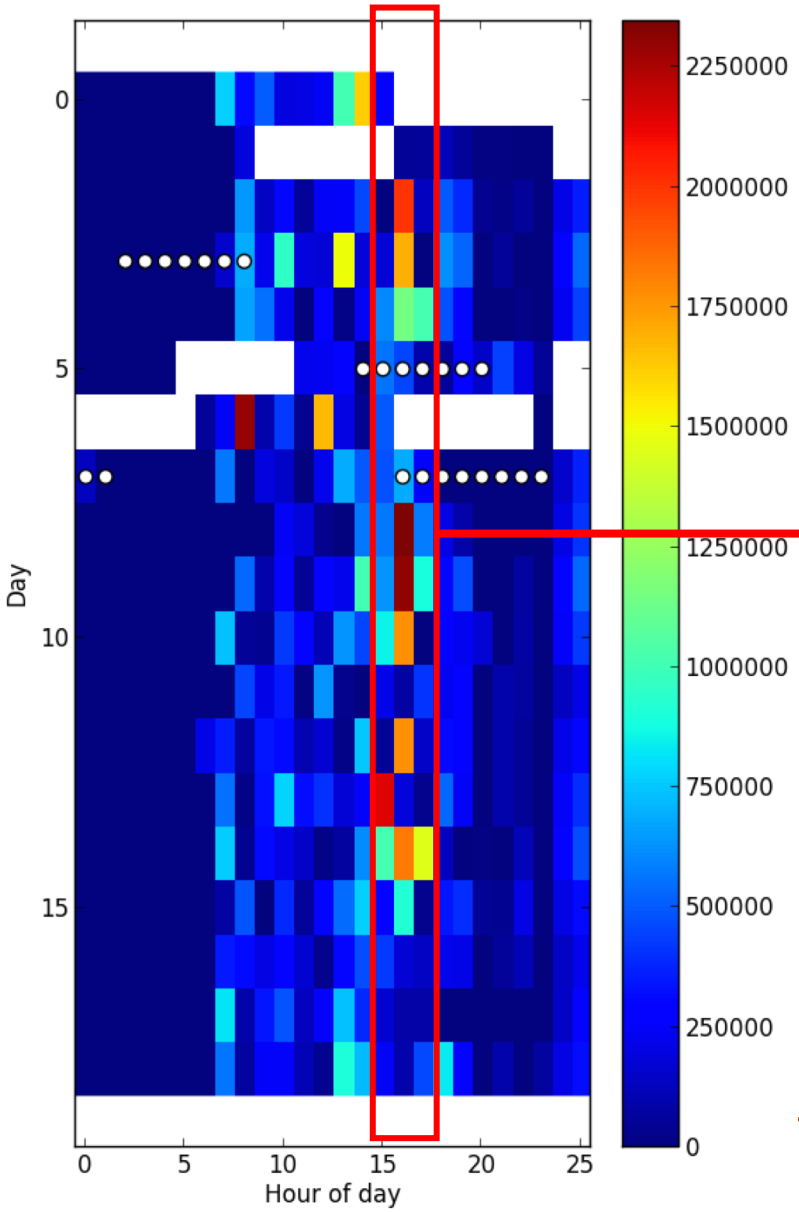
  Use mobile phone sensors to monitor the daily activities of elderly, detect short-term anomalies and long-term evolutions

- Available data

  GPS and accelerometer data obtained from 104 mobile phones during a period of ~1.5 years

- Domain knowledge
  - People are creatures of habit
  - Guidelines for healthy behavior (e.g. amount of active time, …)

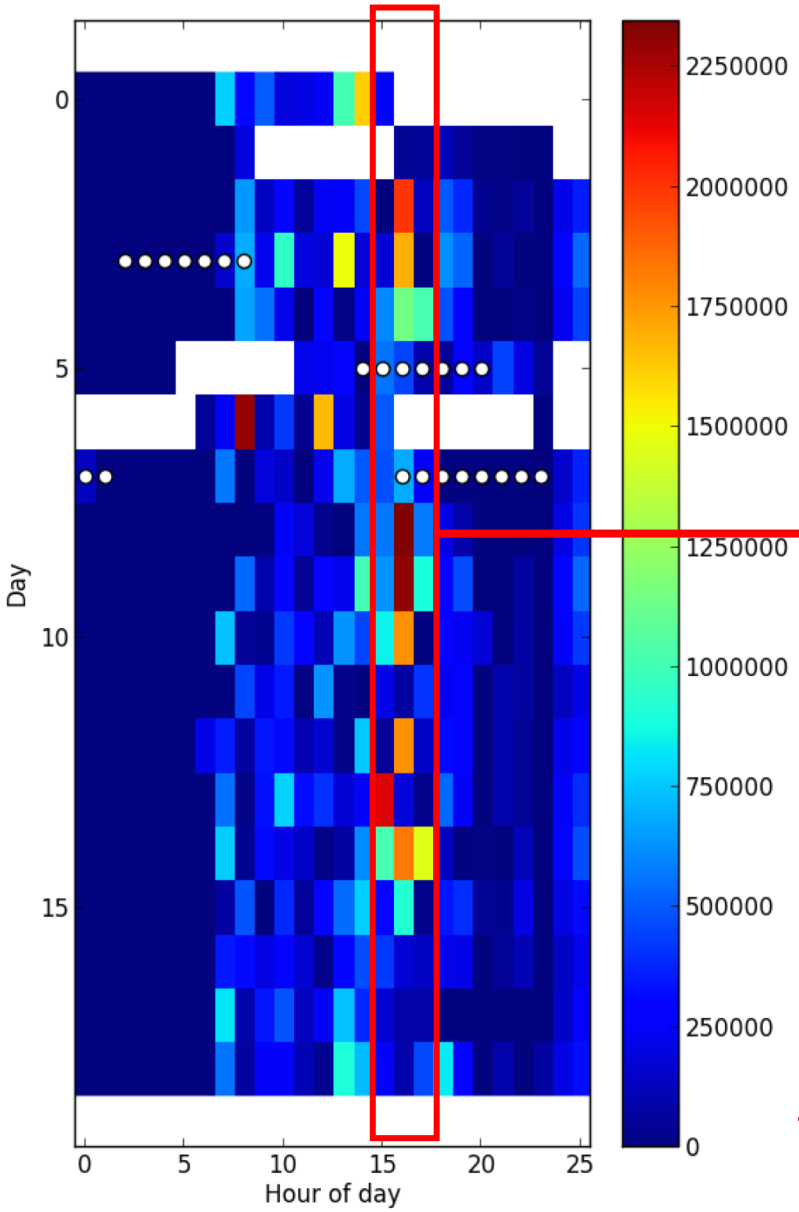
1st step: getting insights into the data



Activity intensity (in milliseconds) for a particular user on Mondays

Intense activity on Monday afternoons between 15:00 and 18:00

1st step: getting insights into the data



Intense activity on Monday afternoons between 15:00 and 18:00

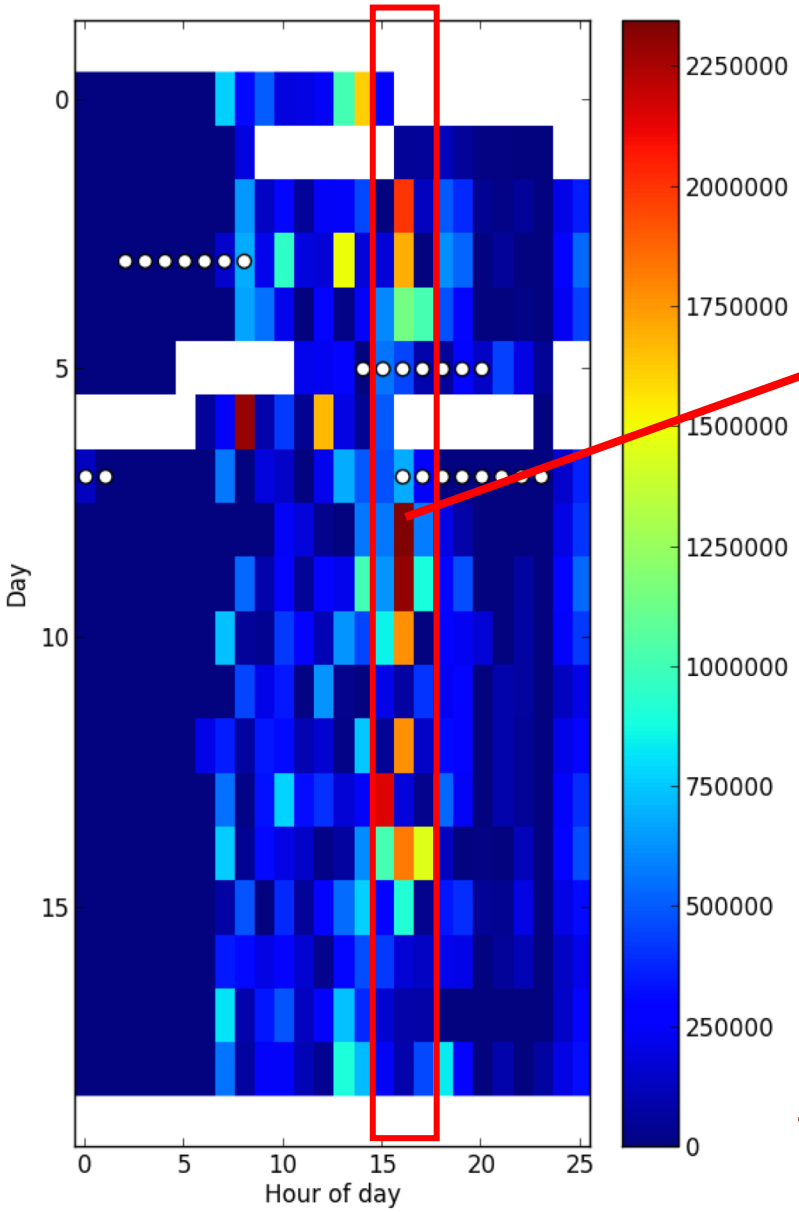
↓

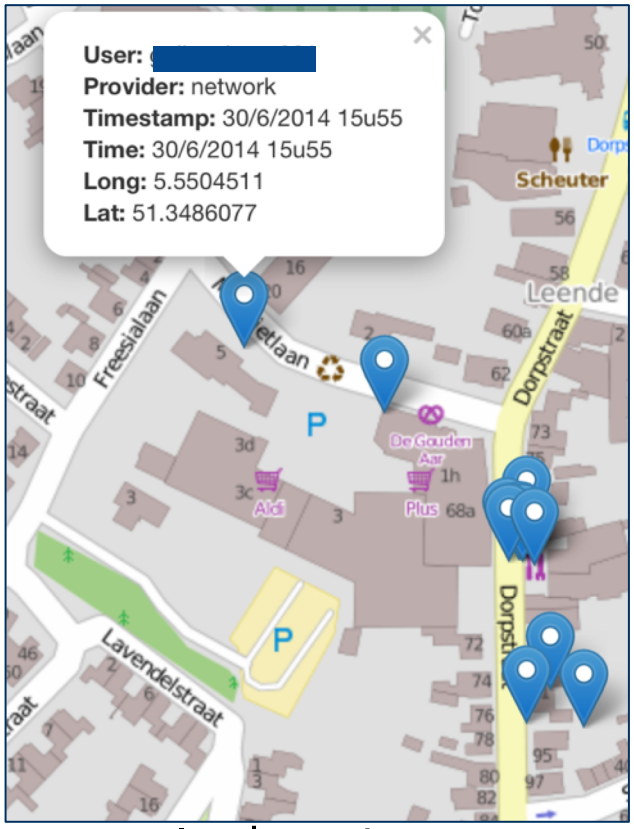
Pattern in user behaviour?

↓

Combine with location data!

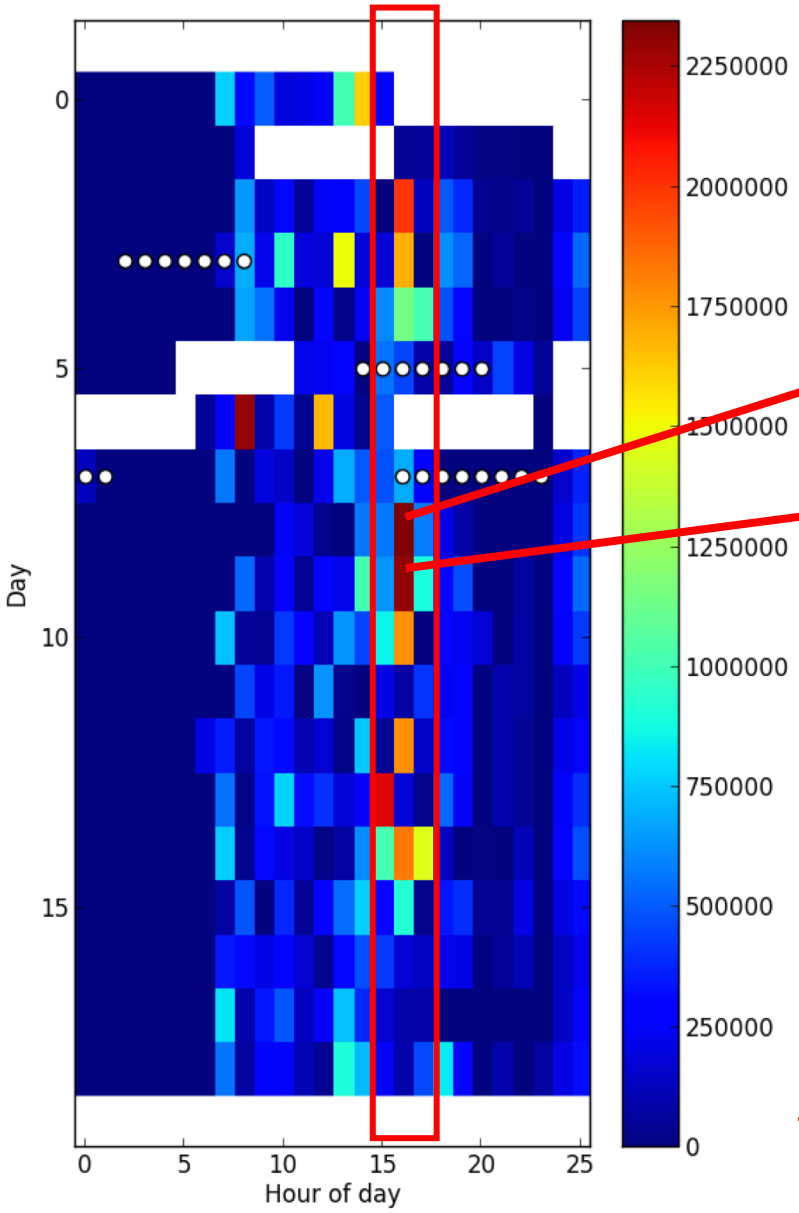
1st step: getting insights into the data

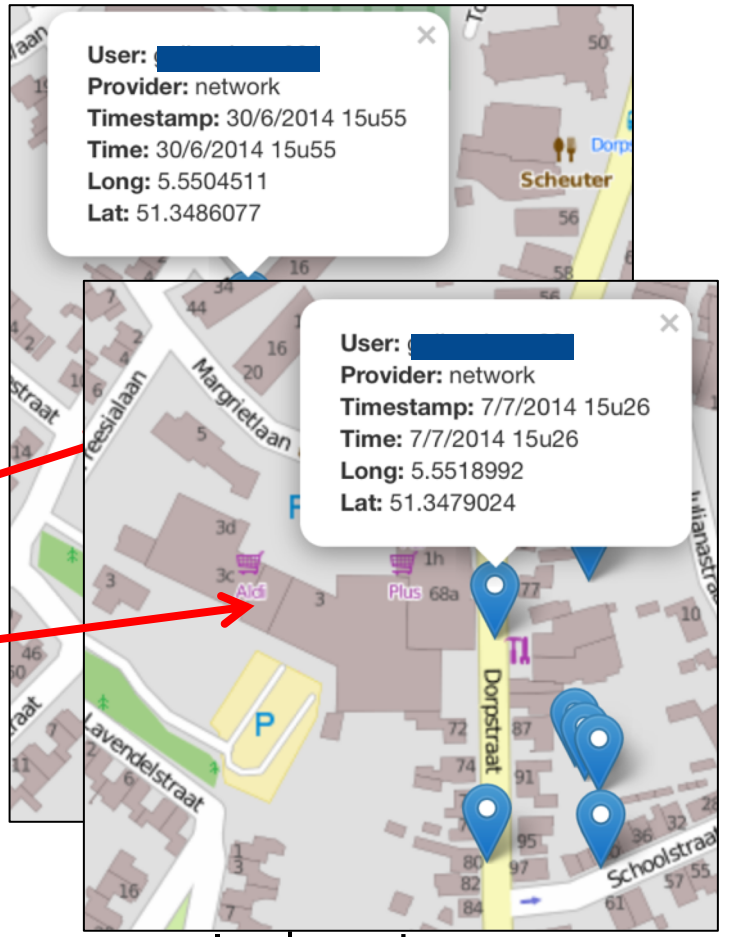




In shopping area

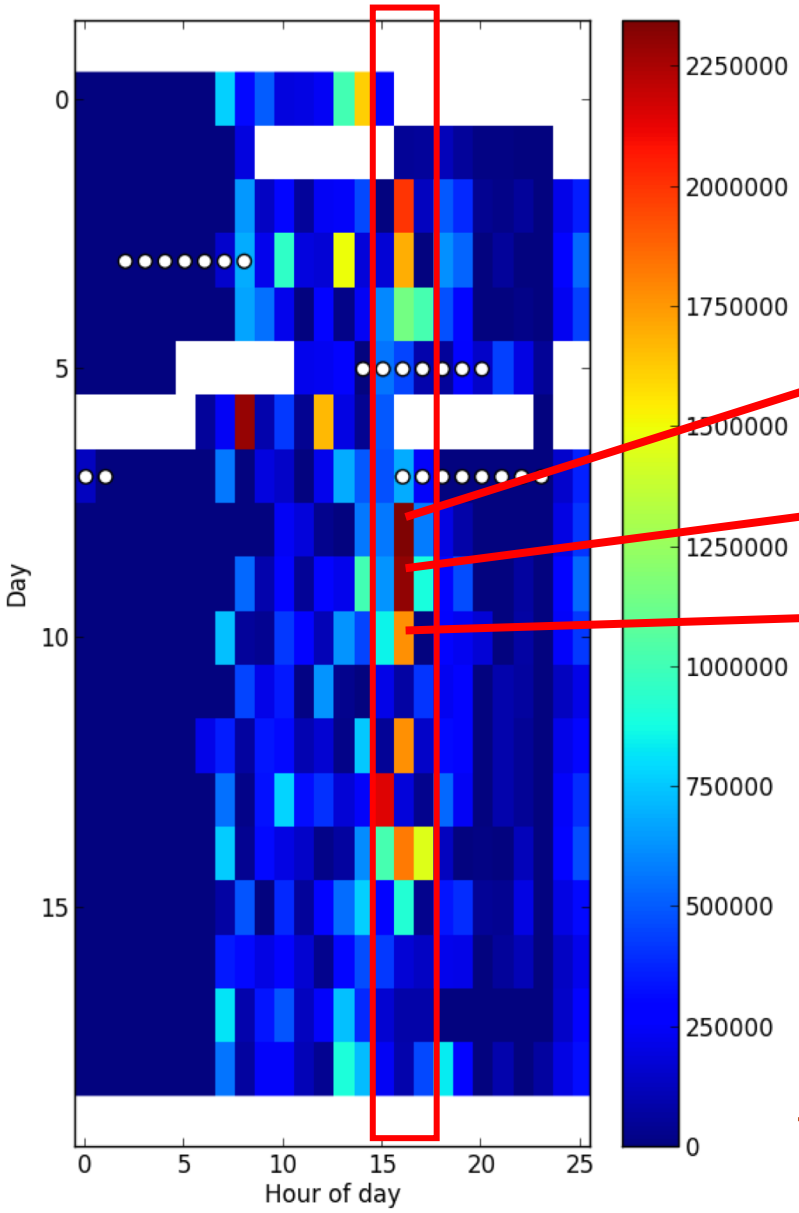
1st step: getting insights into the data

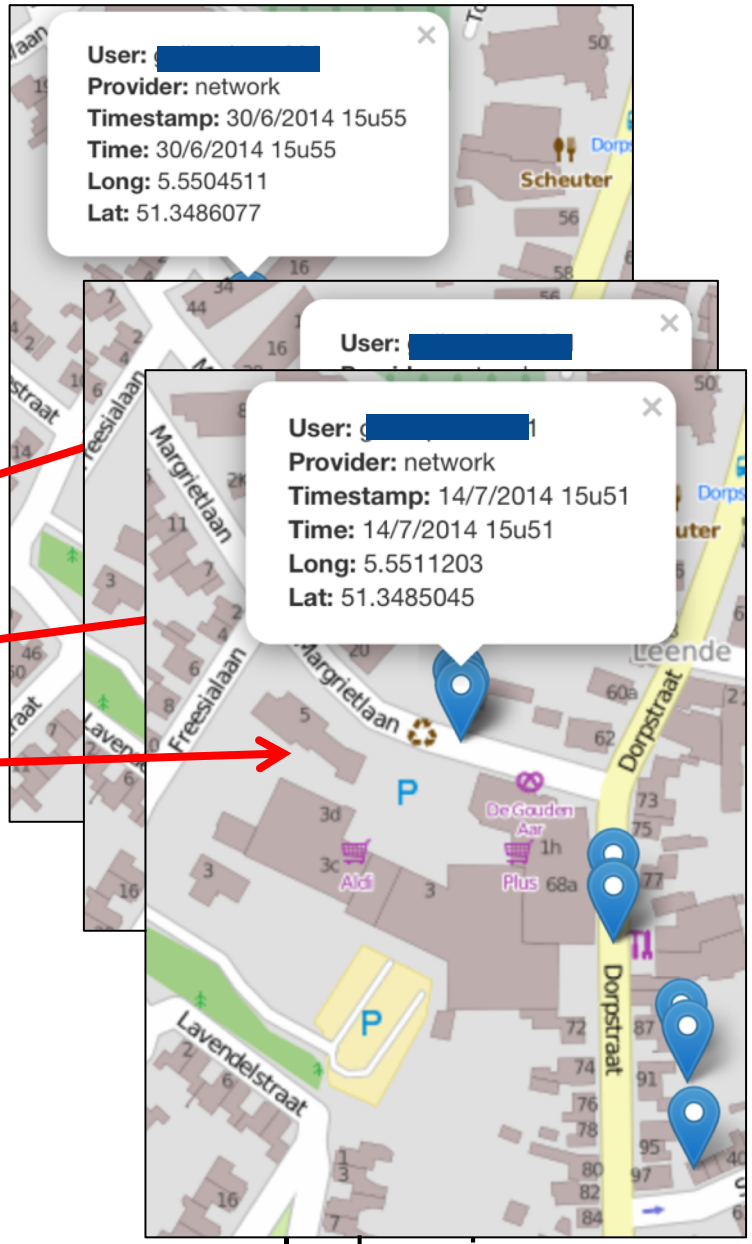




In shopping area

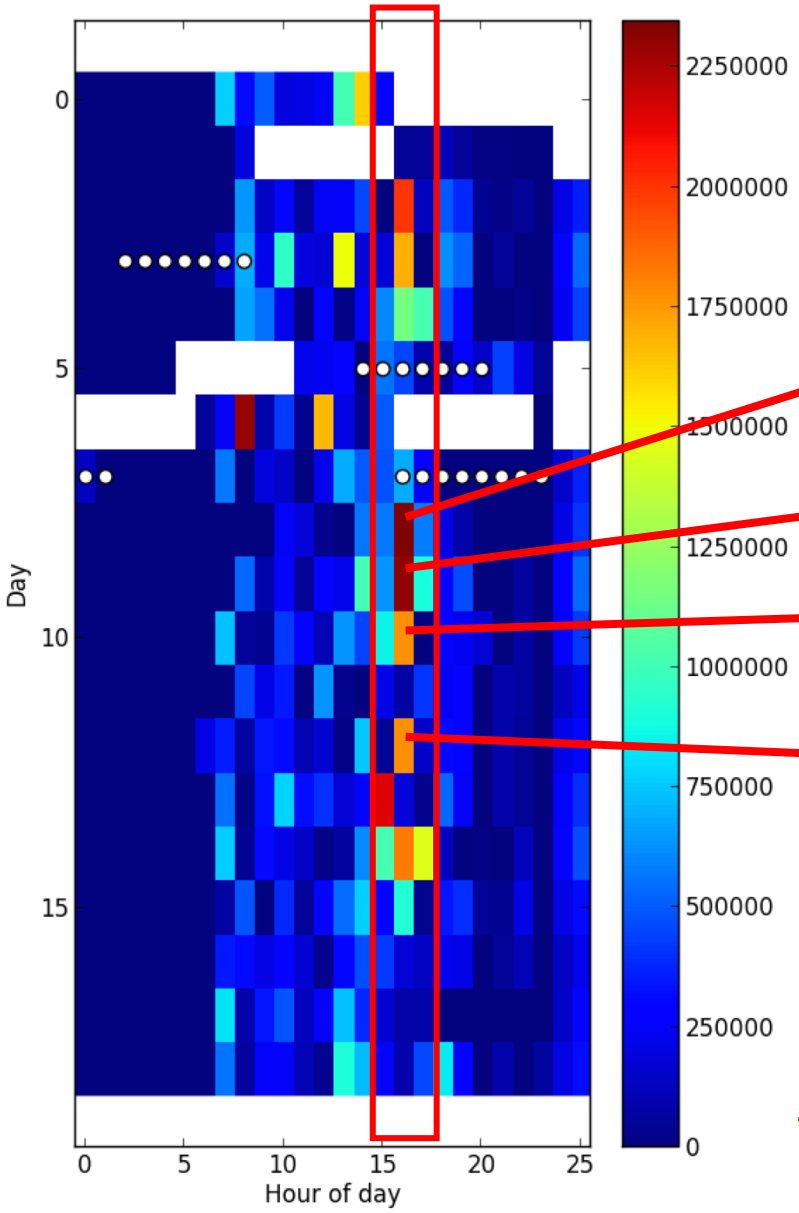
1st step: getting insights into the data

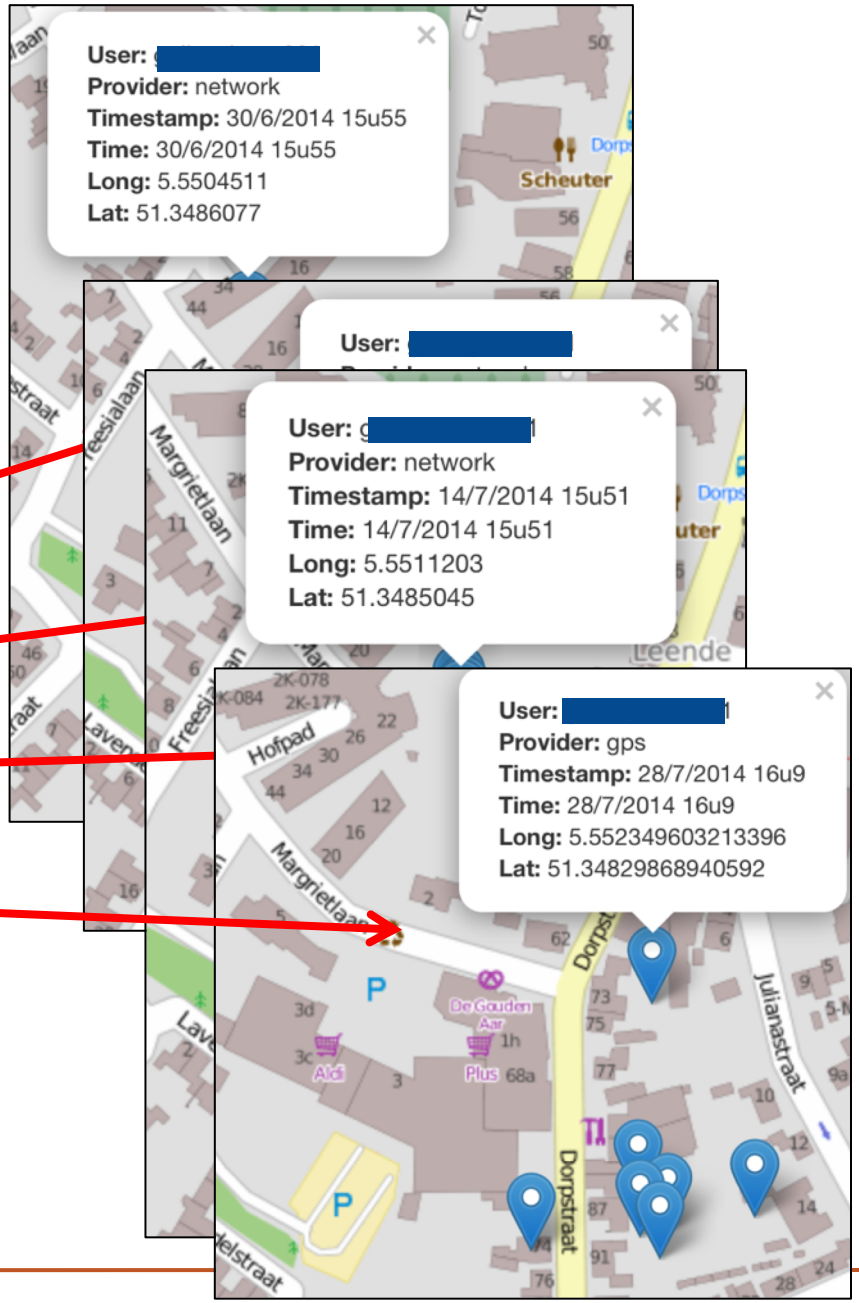




In shopping area

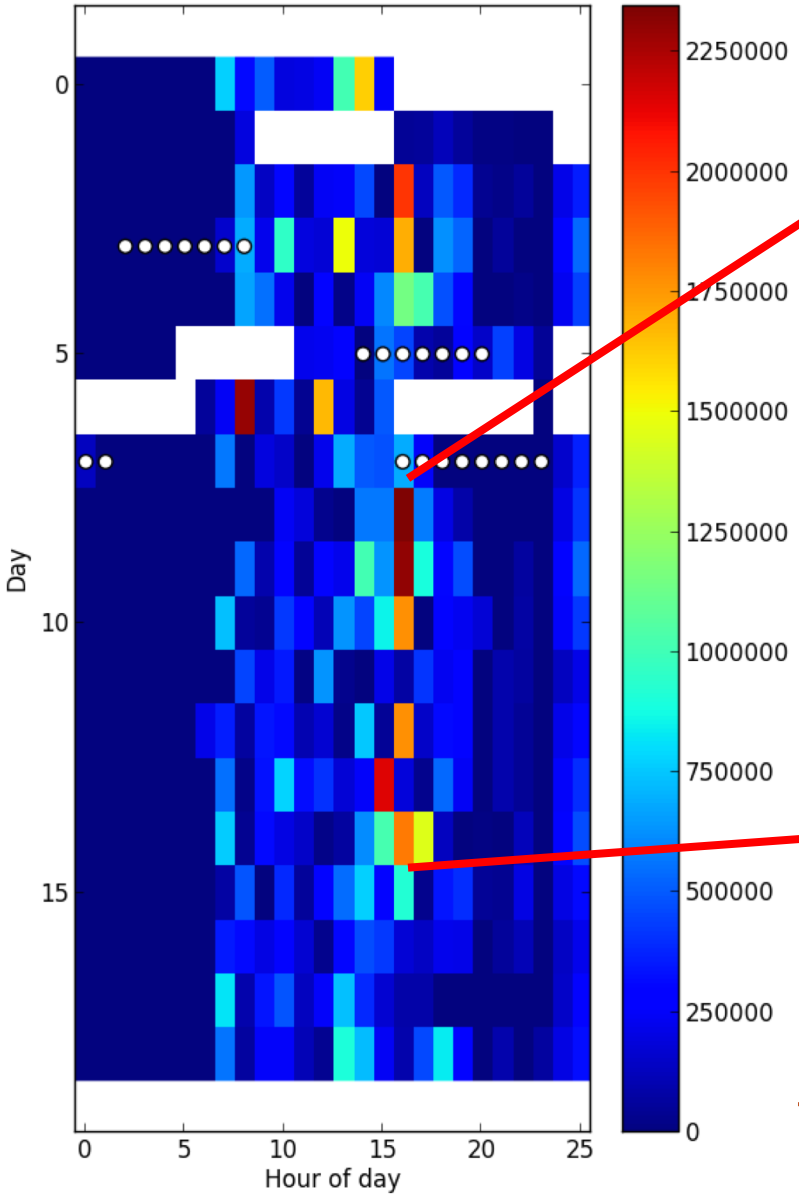
1st step: getting insights into the data





In shopping area

1st step: getting insights into the data

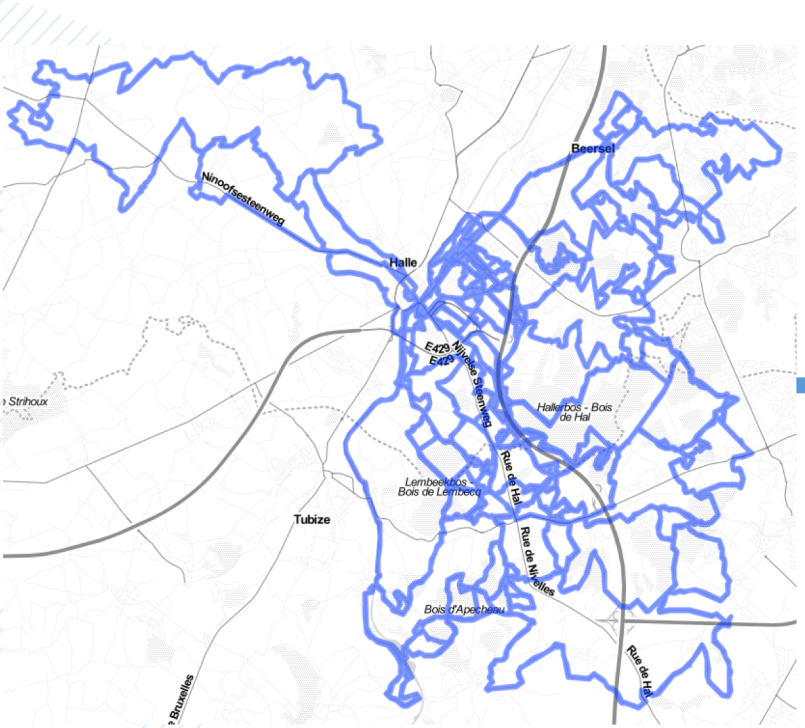


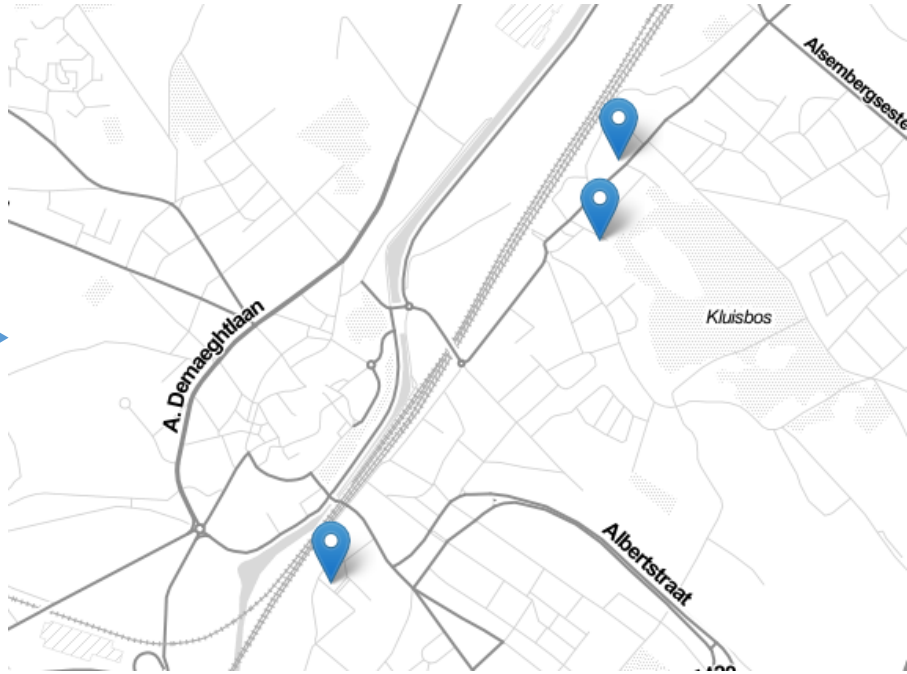
Intense activity on Monday afternoons between 15:00 and 17:00

User in shopping area

2nd step: extracting information from the data

Point-of-interest detection

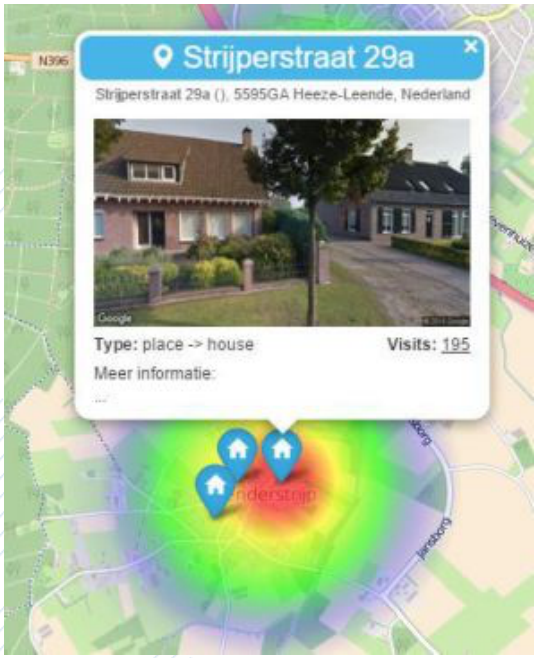




A point of interest is a location that

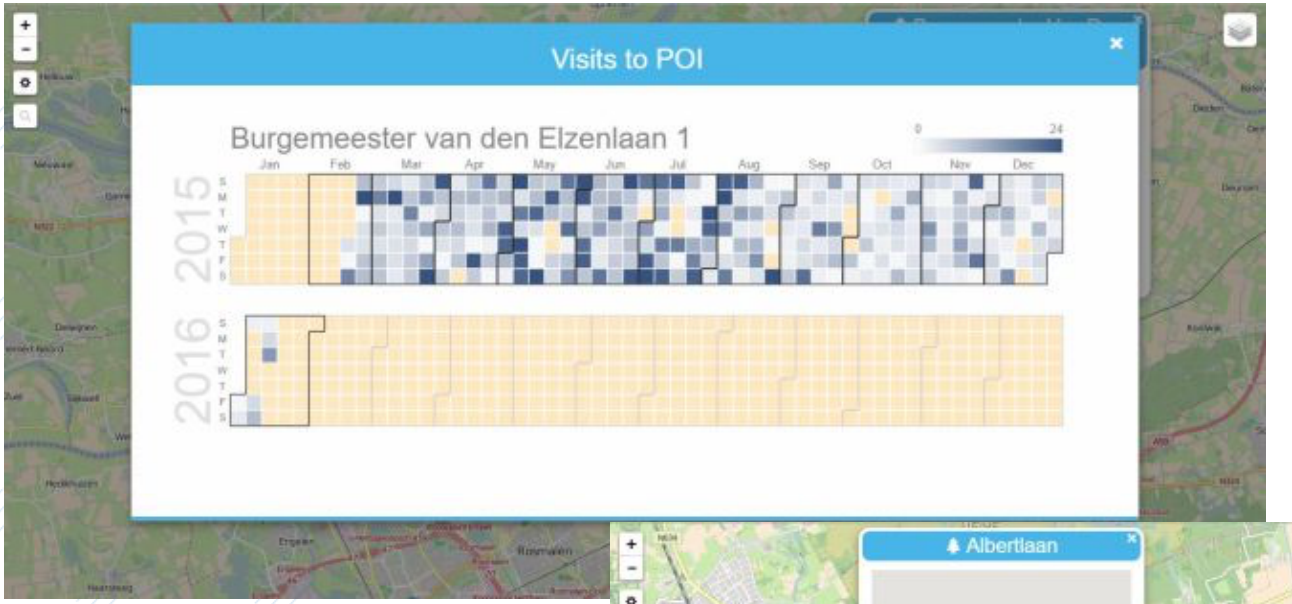
- is '*often*' visited by a user
- for a '*significant*' amount of time

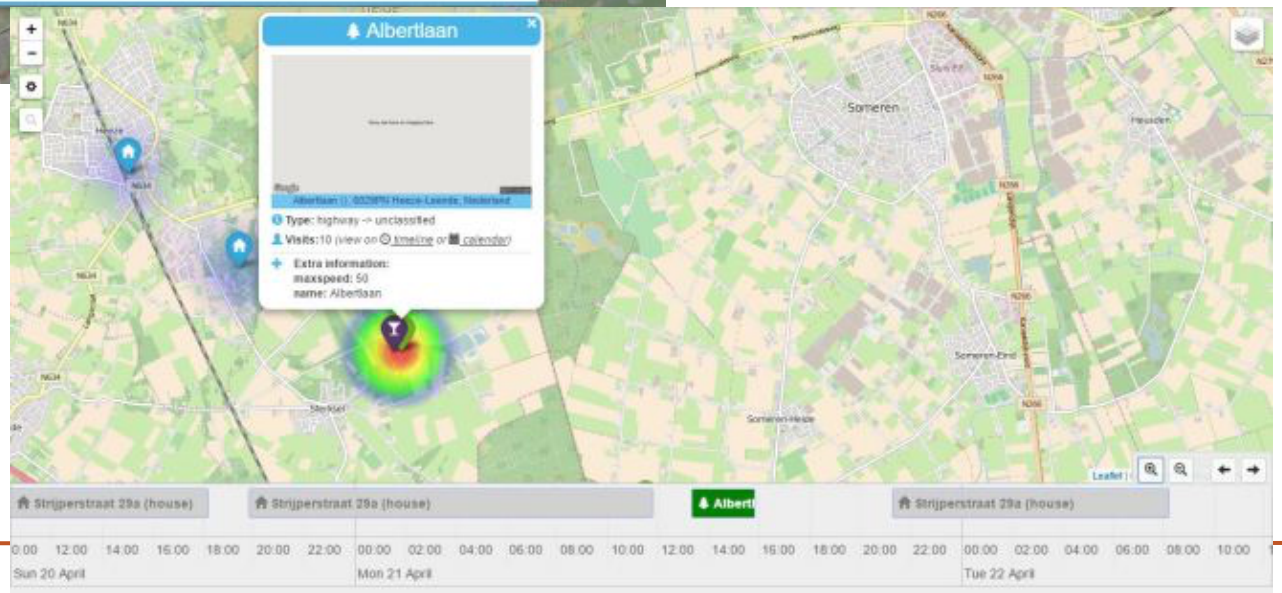
3rd step: enriching the information



- Get extra information from external sources such as Google and OpenStreetMap
  - Picture from Street View
  - Type of place
  - Opening hours
  - Contact details
  - …

4rd step: visualizing the information



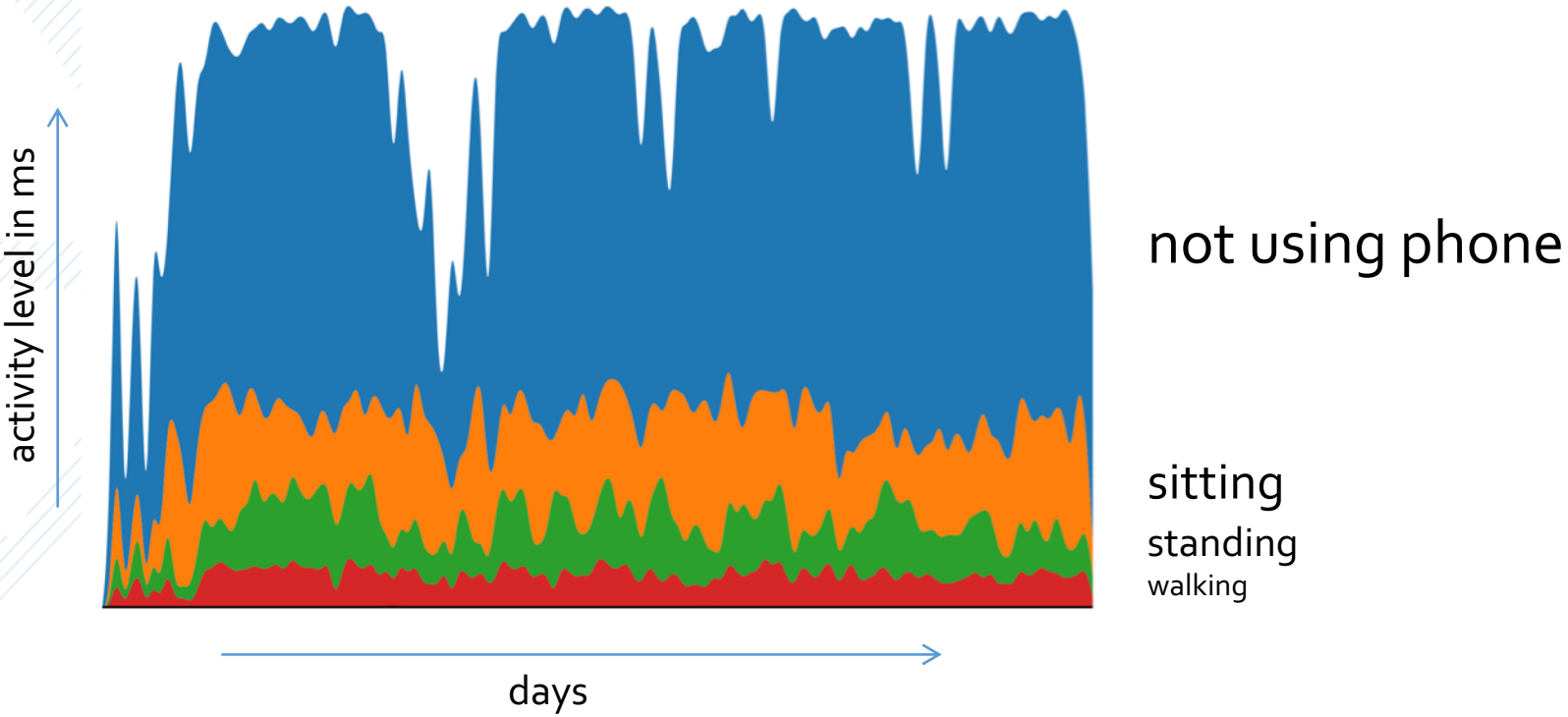


EluciDATA

BUT …

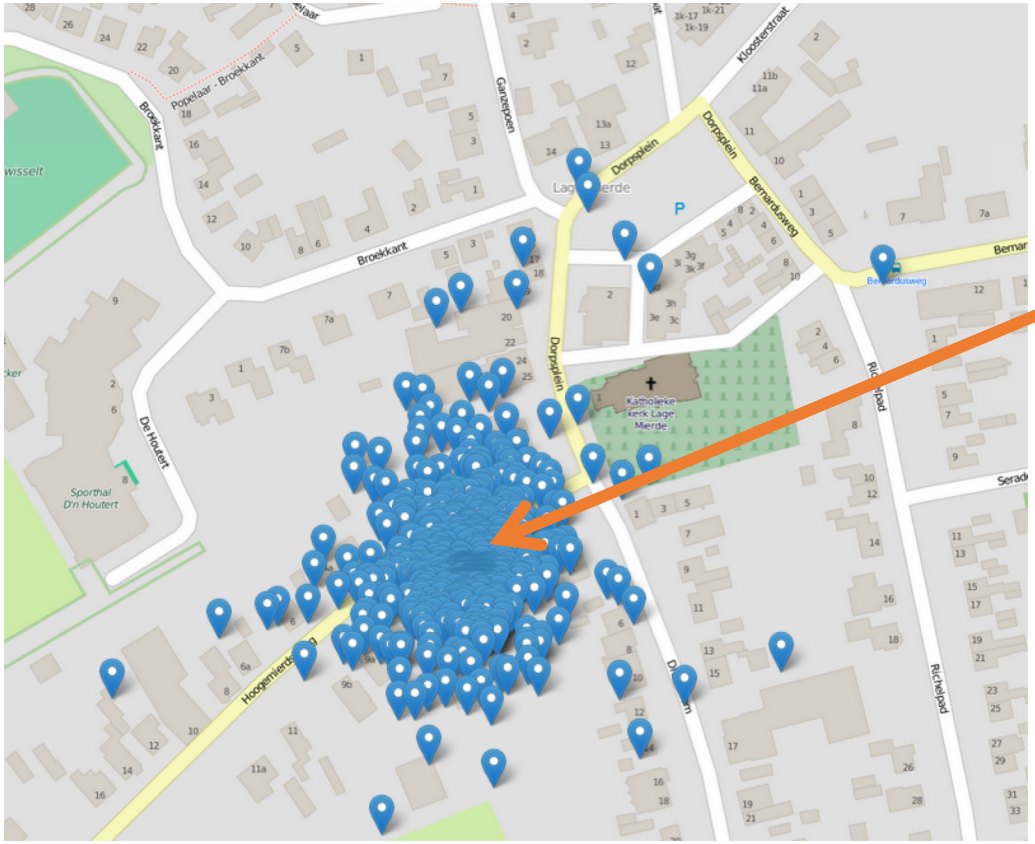
(Accelerometer) data from the most used wearable

- 2/3rd of the 15.000 hourly records do not contain useful data



BUT …

(GPS) data from the most used wearable

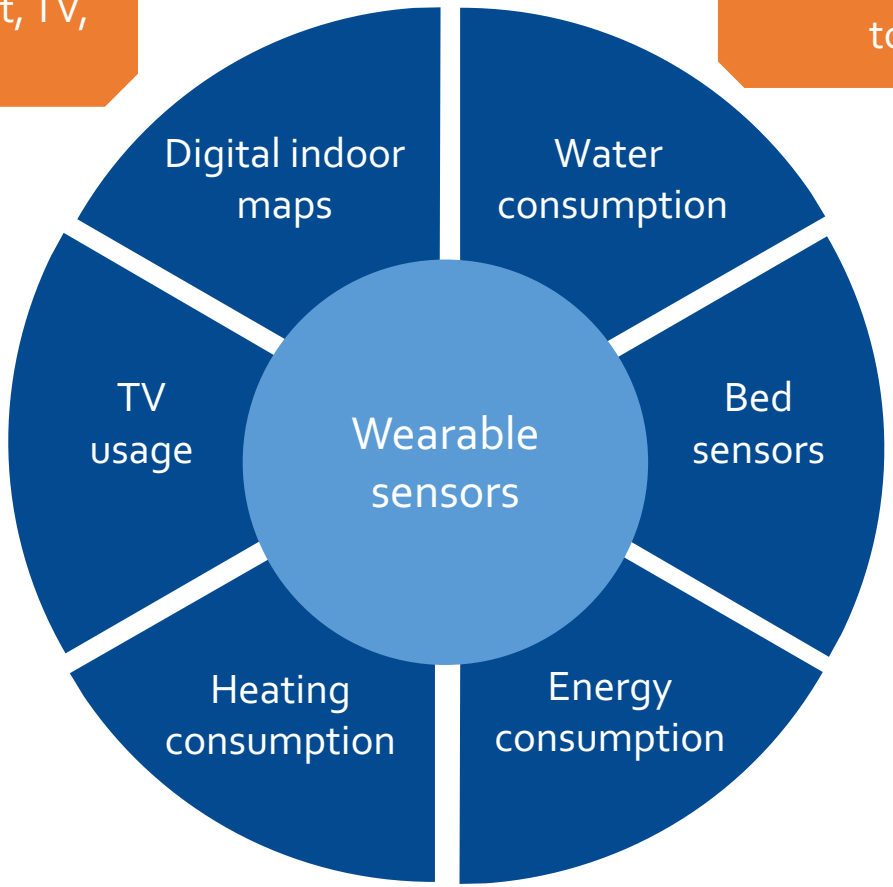


"home"

Next step: complement wearable data with context data

know the location of bathroom, kitchen, toilet, TV, bed, …

know usage of shower, toilet, taps, ….



know leisure times and duration

know sleeping quality, times and duration

derive occupancy patterns, inactivity, cooking, ….

Rationale

Connected sensors

"showering" & "sleeping" vs. "not using phone"

"bathroom" & "kitchen" vs. "home"

→ fine(r)-grained coverage and increased accuracy of locations & activities

typical hours of sleep, showering time & duration, watching TV, …

→improved baseline of 'typical' behavior

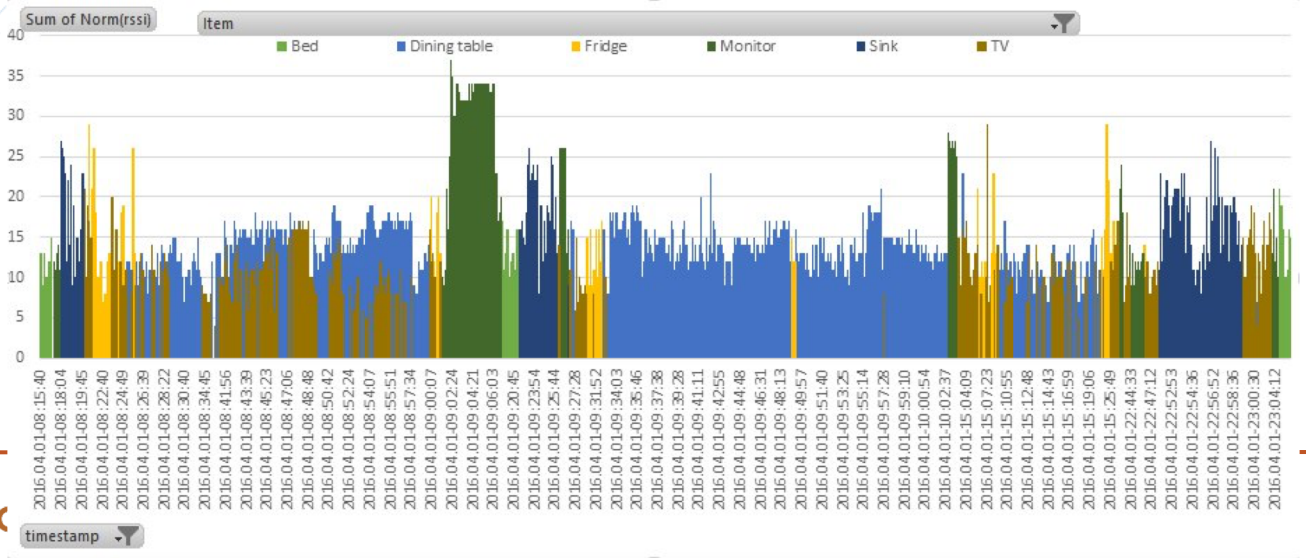
→prerequisite for anomaly detection, prediction & prevention

disturbed sleeping pattern, more TV watching, less physical activity, …

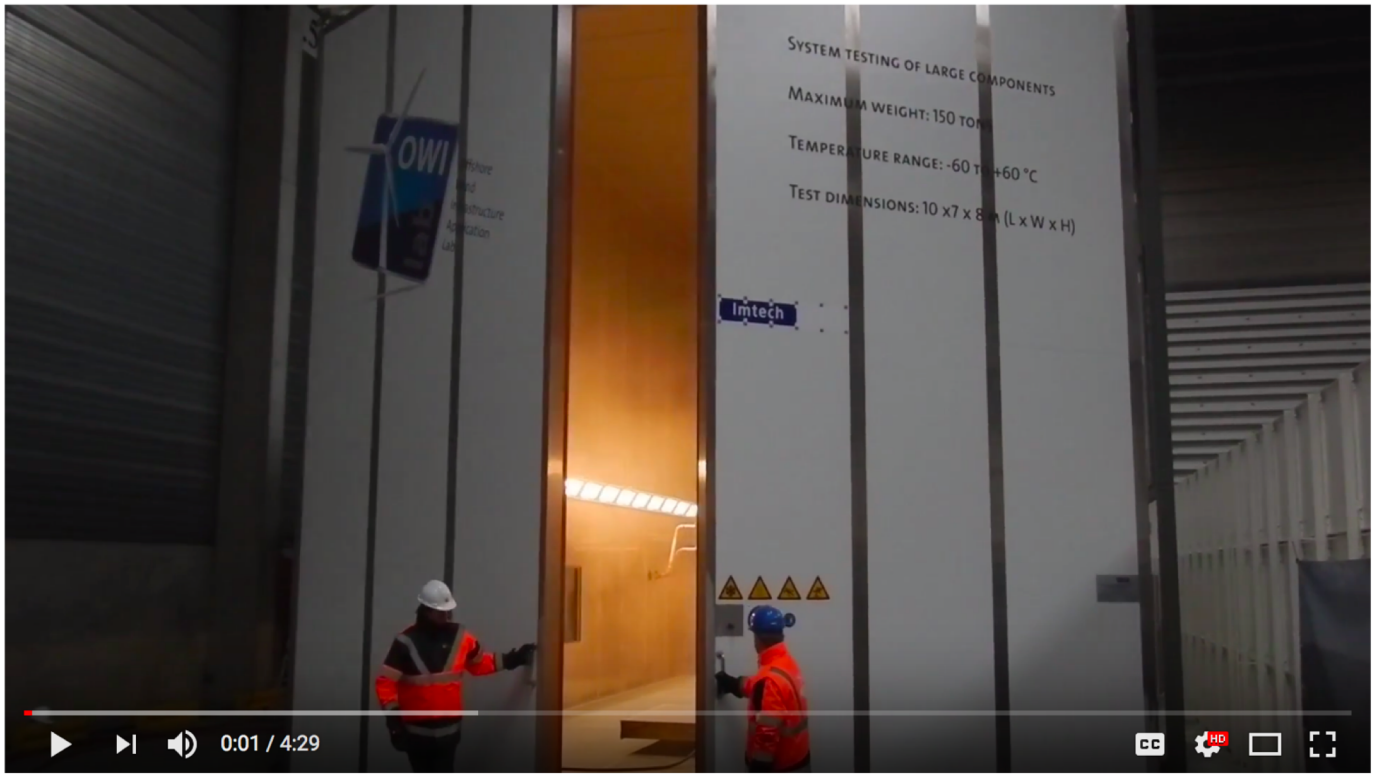
Experimental setup

+

=



Similar setup/technology can be used in different contexts



SIRRIS SMART PRO safer industrial workspace

https://www.youtube.com/watch?v=XgF29RrtJEw

2nd example use case

Traffic flow visualization

- Business goal

  Estimate the commercial value of bike-friendly infrastructure

- Available data
  - Data describing the traffic flow for 242 (biking) lanes in Flemish Brabant
  - Measured at different, only partially overlapping time intervals
  - Describes passing bikes of a particular type passing a particular GPS point at a particular time, given direction and speed

- Domain knowledge
  - Functional vs. recreational traffic

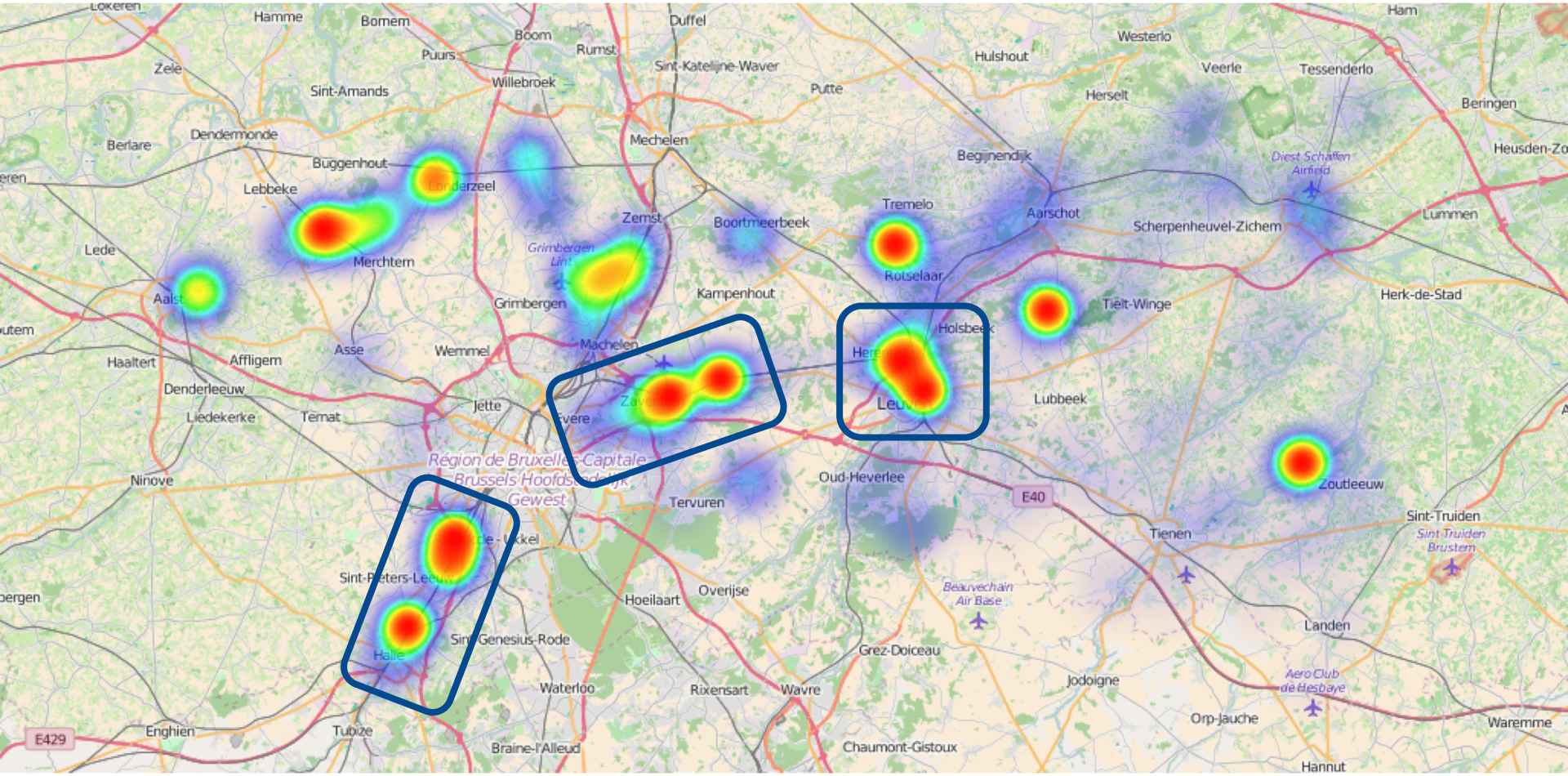
Advanced visualisation

- Advanced visualisation techniques exploit humans' extraordinary visual pattern recognition abilities

- A clever visualisation of data can already reveal interesting patterns & insights, even before any complex algorithm is applied

- The tricks of the trade lie in knowing what visualisations exist, and how to use them in an intelligent way

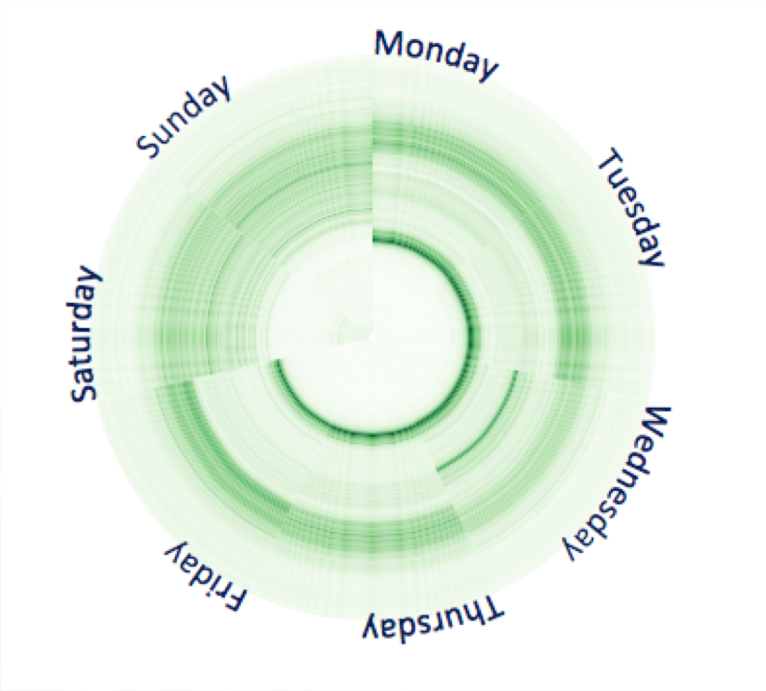
Geographical heatmap – data selection



Weekly patterns in traffic flow data

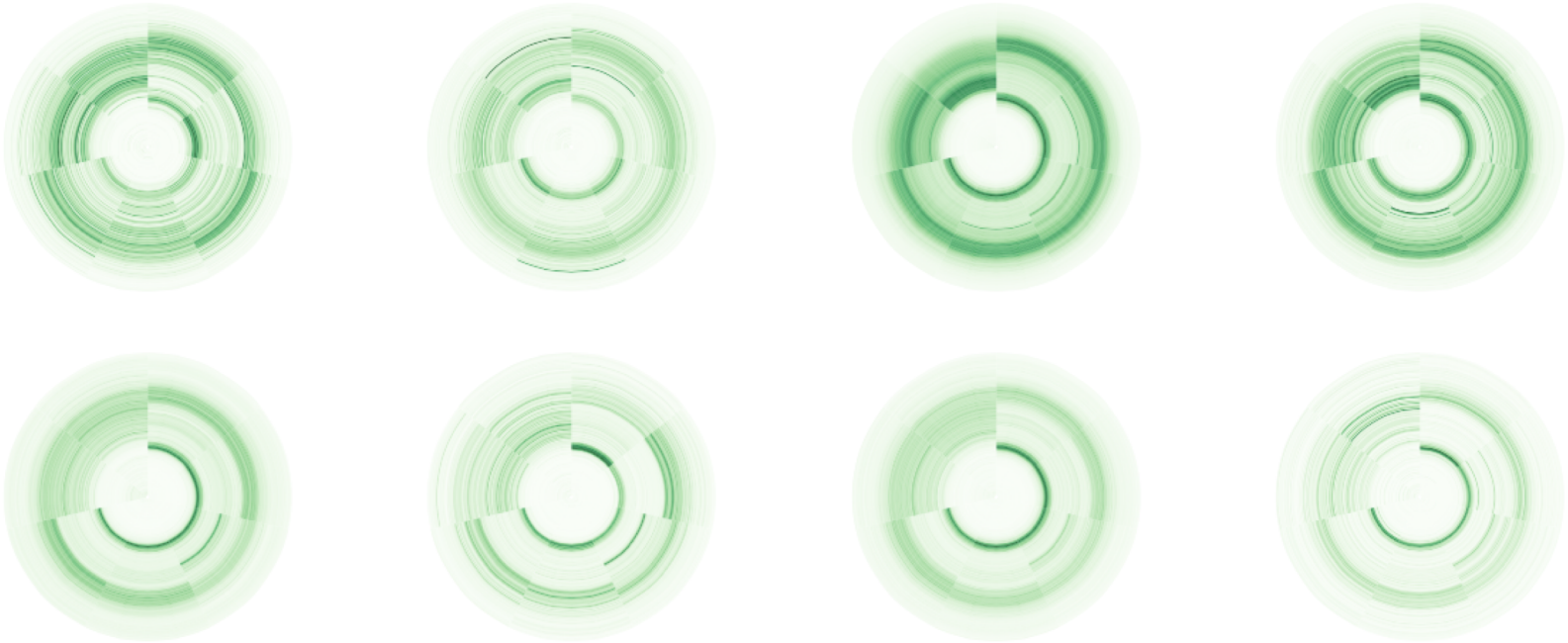
- 1 single node in the cycling network
- 10-min based measurements *further aggregated and normalized* per weekday

- Patterns
  - Morning and evening traffic on weekdays
  - Evening traffic more spread out over time
  - Wednesday has slightly less evening traffic, and slightly more afternoon traffic
  - Traffic in weekend days starts later and ends earlier
  - This conforms to the assumed "functional" traffic



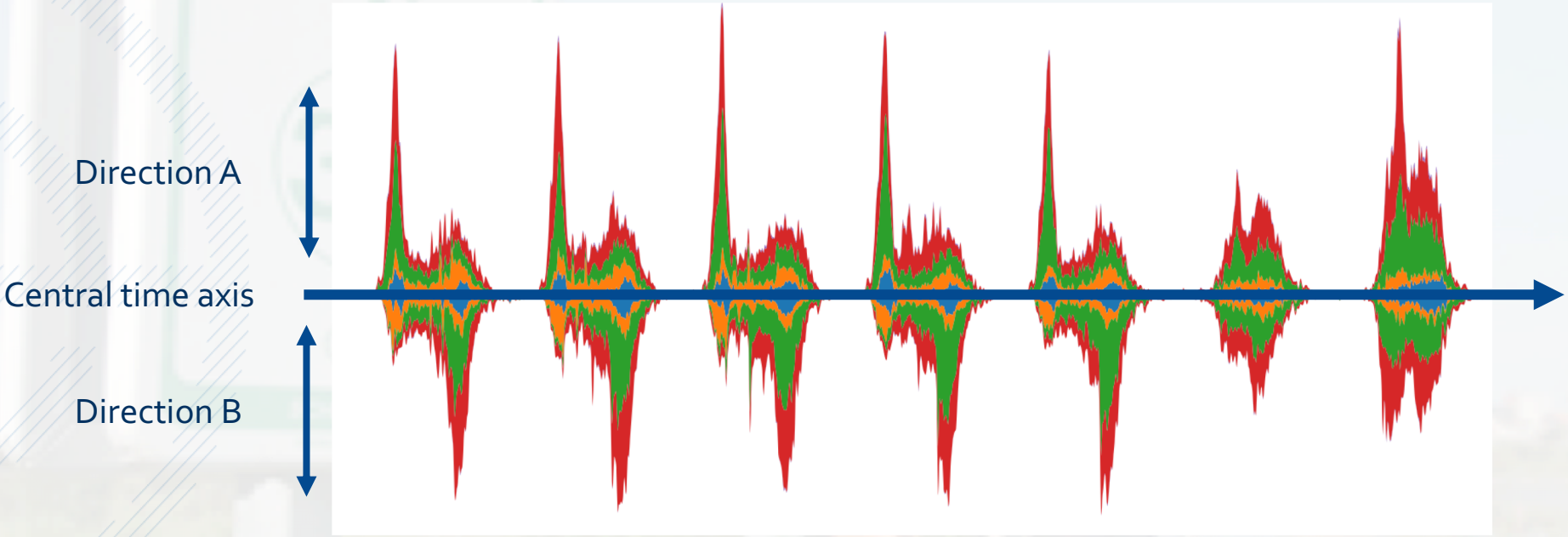
Small multiples

- Chart = circular heatmap
- Partition = different nodes in the cycling network



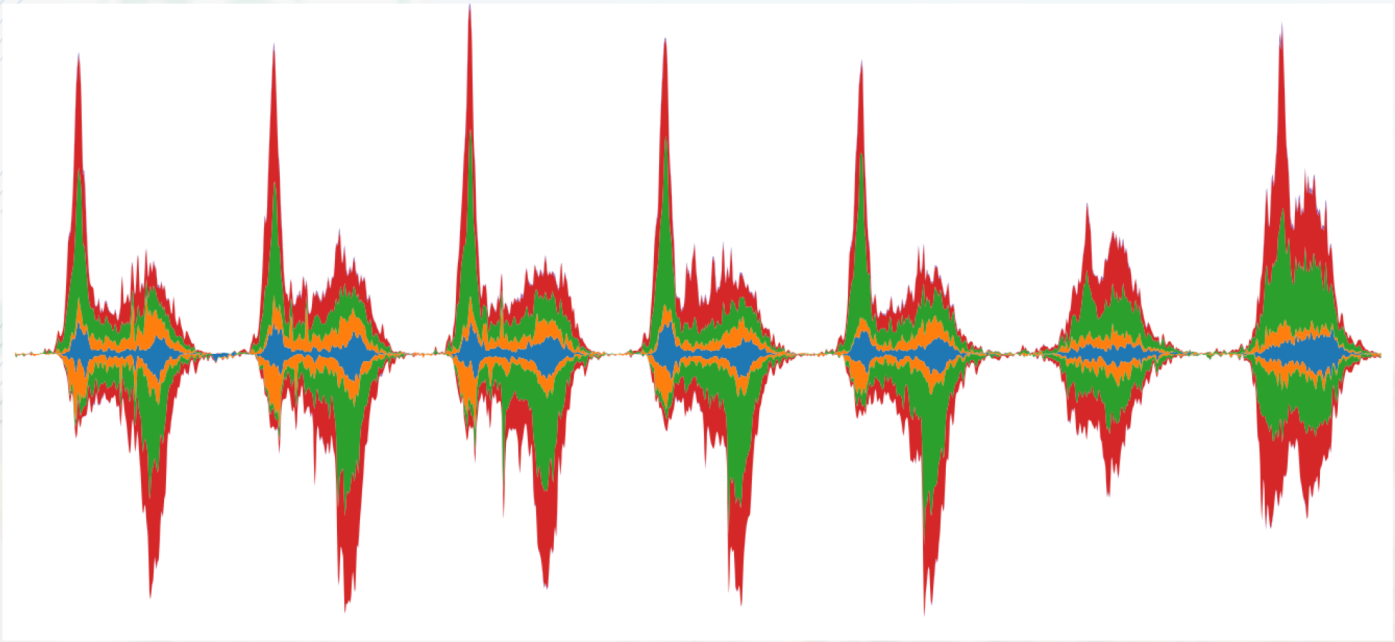
- Weekday vs. weekend patterns consistent across nodes
- Morning and evening traffic pattern consistent across nodes
- Wednesday pattern visible across certain nodes

Stream graphs



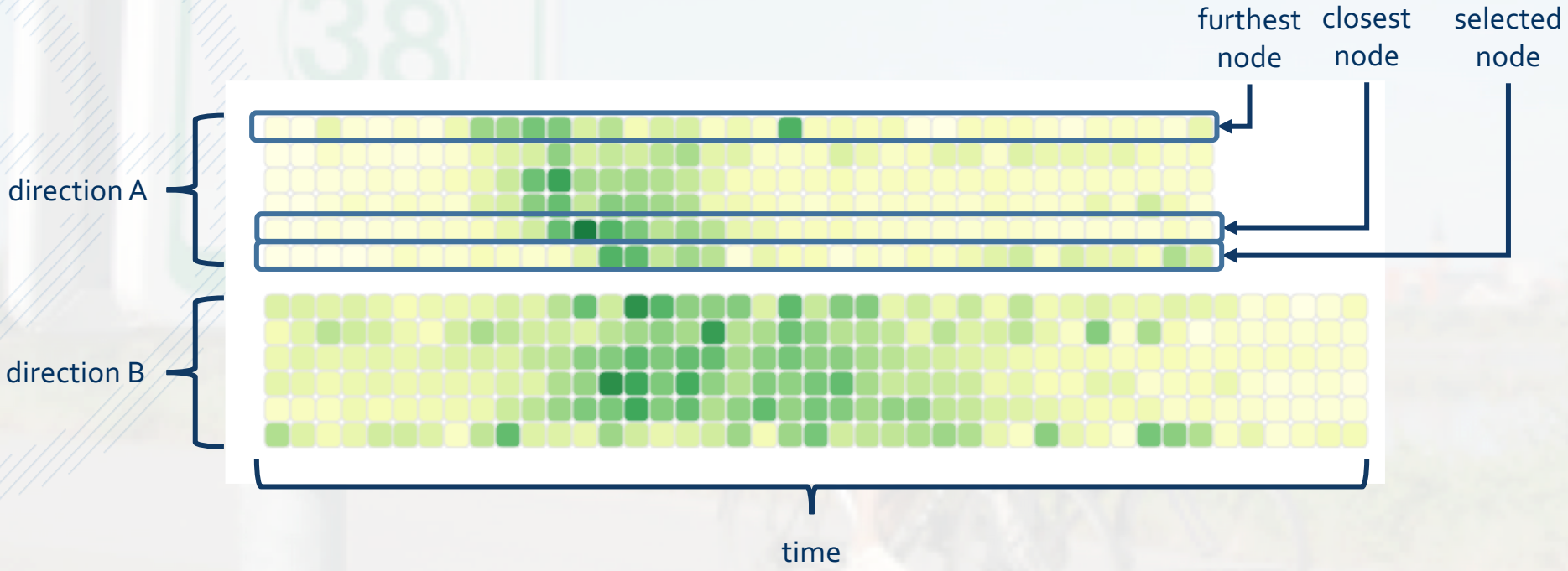
Stream graphs reveal directions

- Weekday vs. weekend patterns consistent across nodes
- Consistent morning and evening traffic on weekdays *in opposite directions*

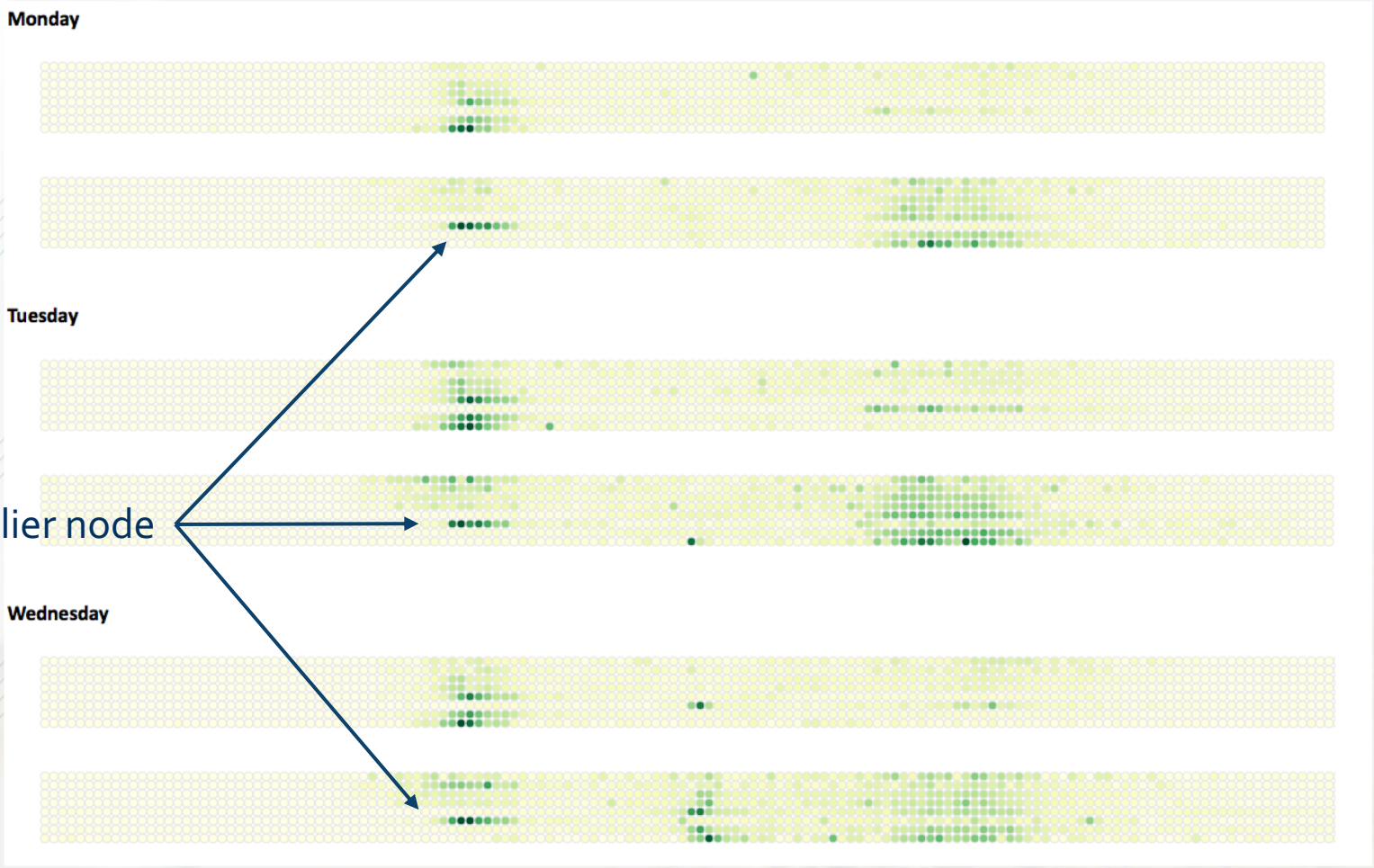


Standard heat map

- Matrix = different nodes in the network, ordered according to their geographical distances to a selected node (in the last row)
- Pairs of matrices represent traffic in the two opposing directions

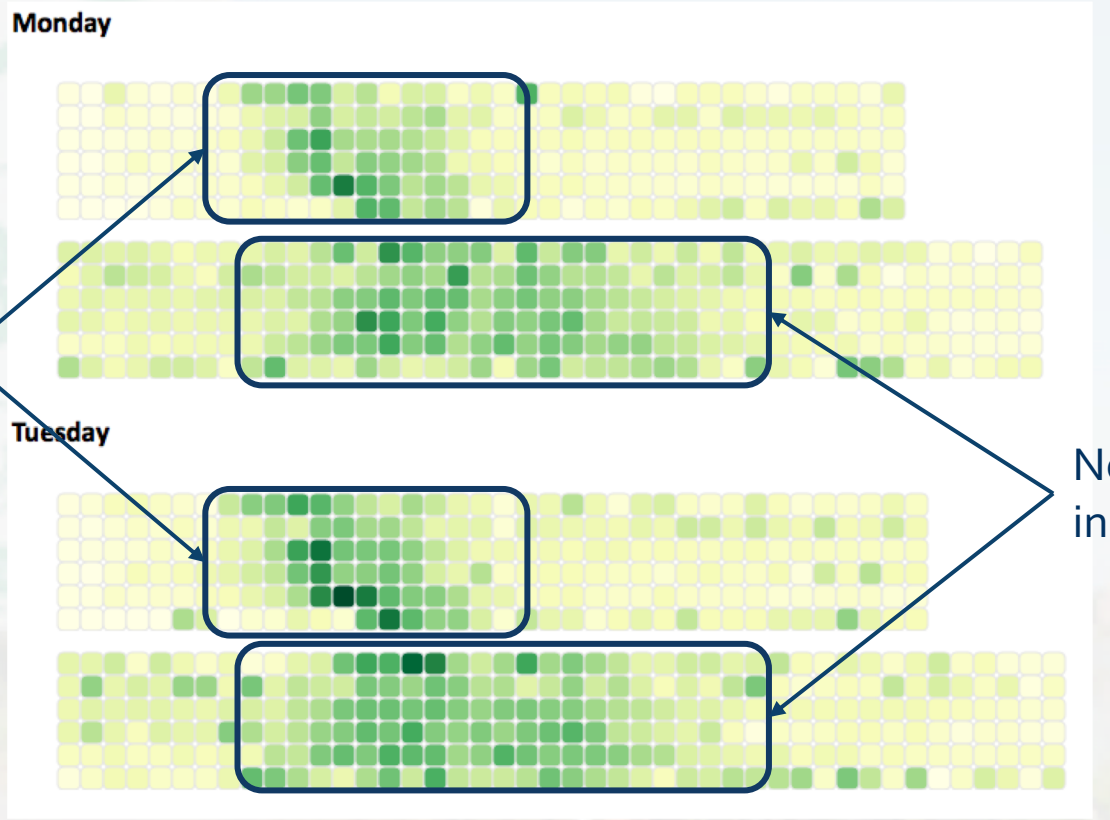


Standard heat map to detect outliers



Standard heat map to detect flow

Morning flow towards selected node



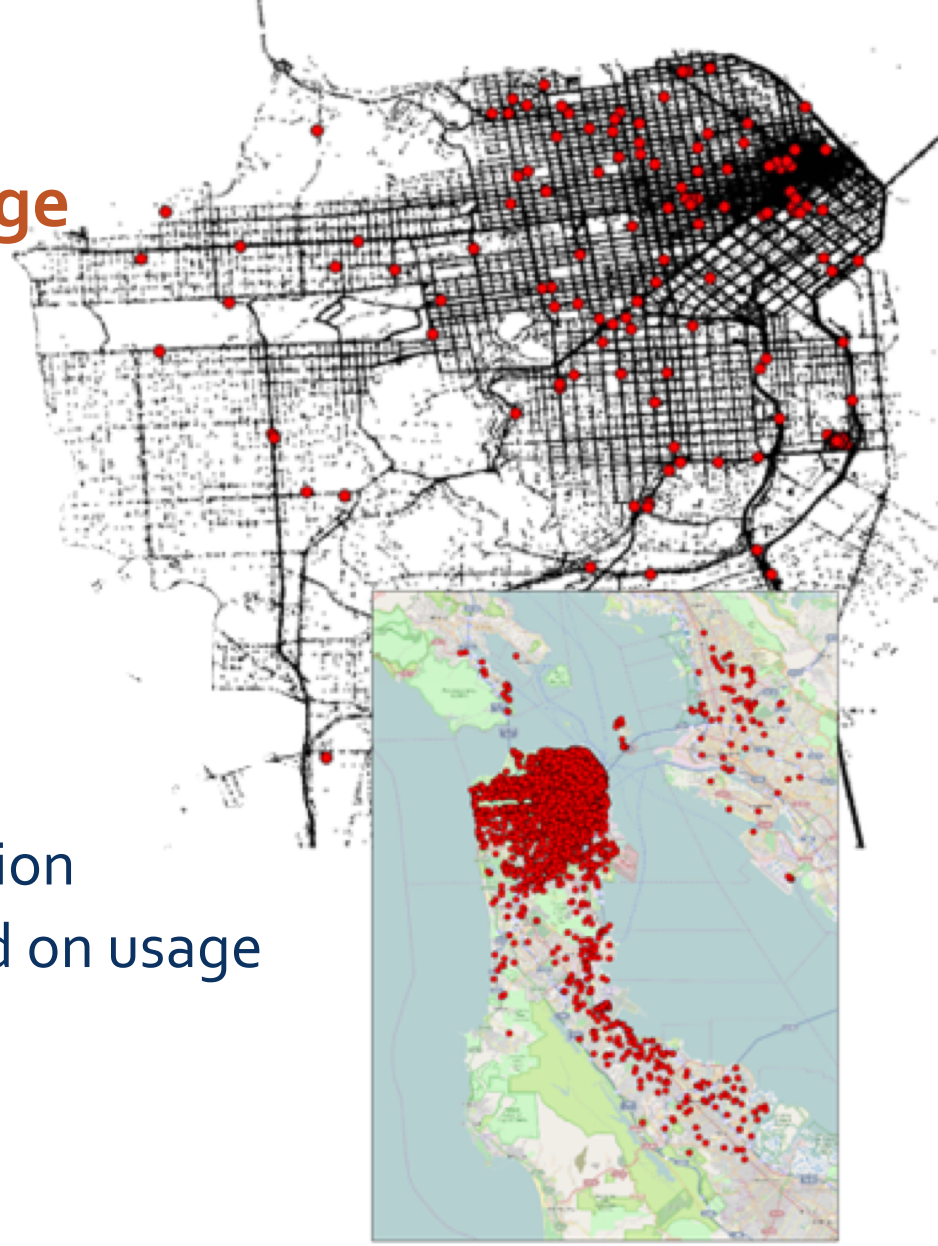
No clear pattern in afternoon flow

3rd example use case

Spatio-temporal taxi usage

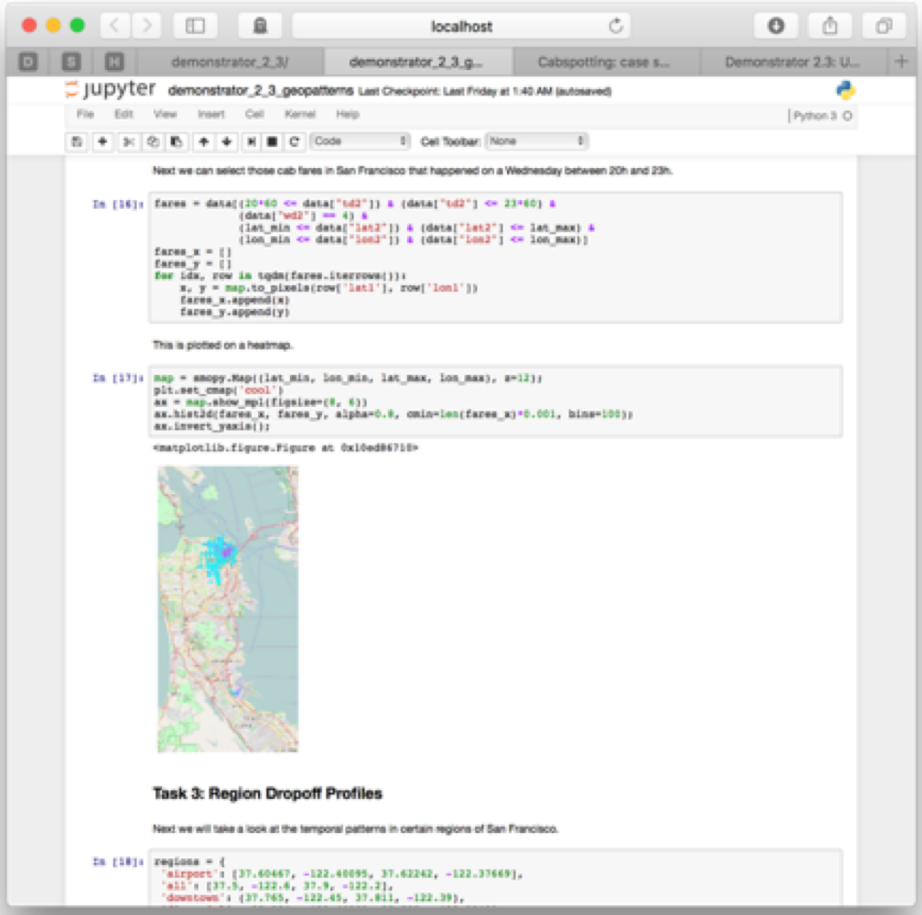
- Given:
  - Taxi fares (route + time)

- Tasks:
  - Typical duration
  - Frequent locations
  - Weekly profile for a location
  - Location clustering based on usage
  - Taxi distribution

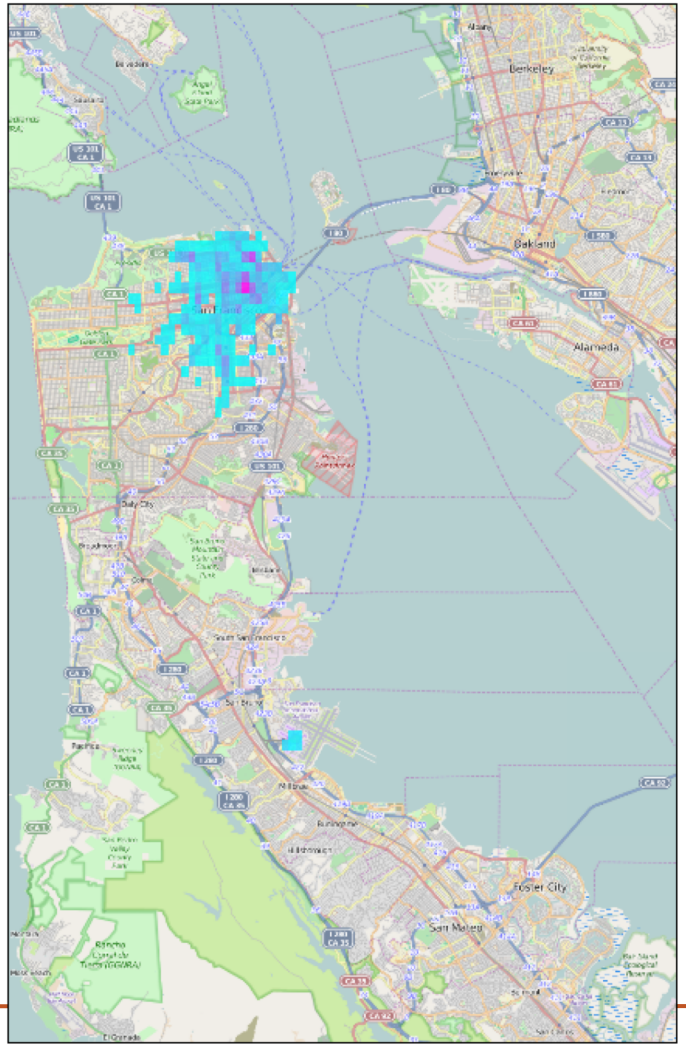


Setup

- Data: Cabspotting.org
- Analytics tools
  - IPython Notebook
  - Pandas
  - Scikit-learn
  - Folium/Smopy
    OpenStreetMap.org

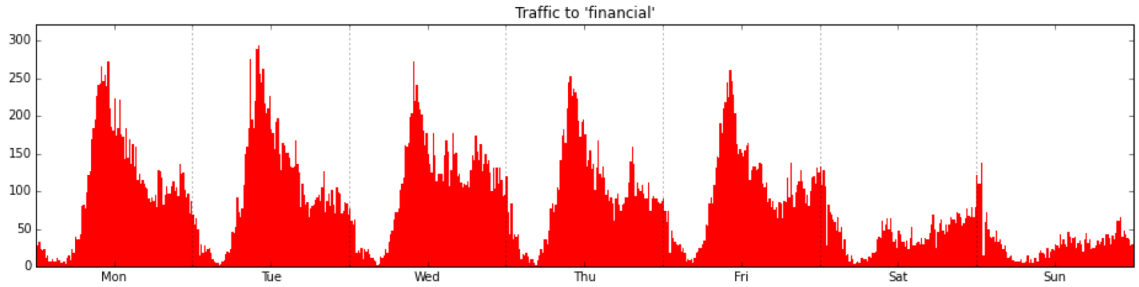


Frequent Locations

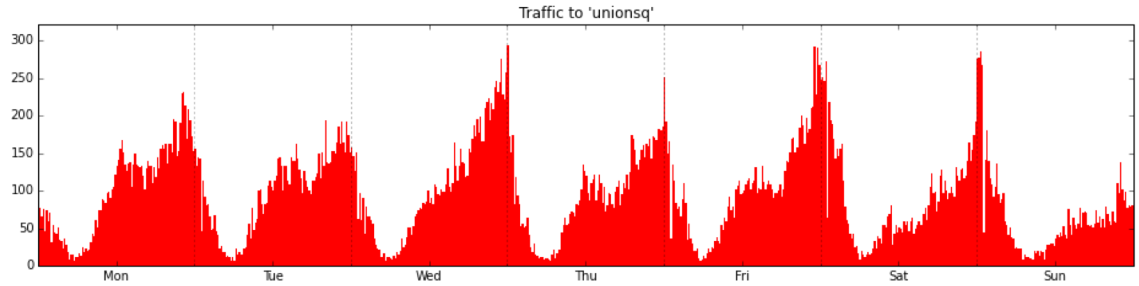


Usage Profiles

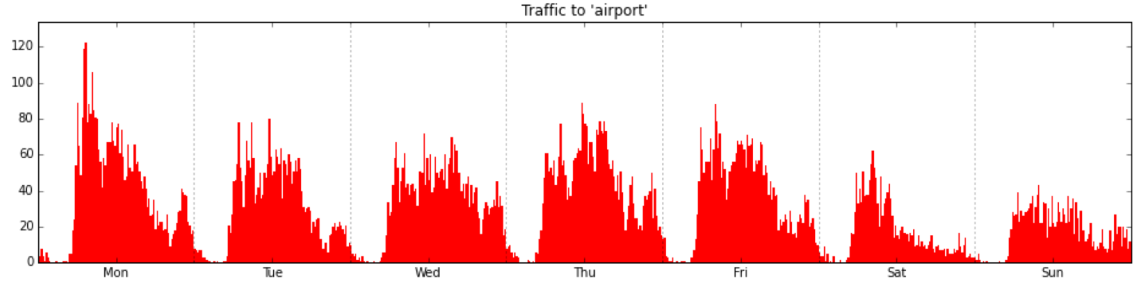
#dropoffs
Financial District



#dropoffs
Union Square



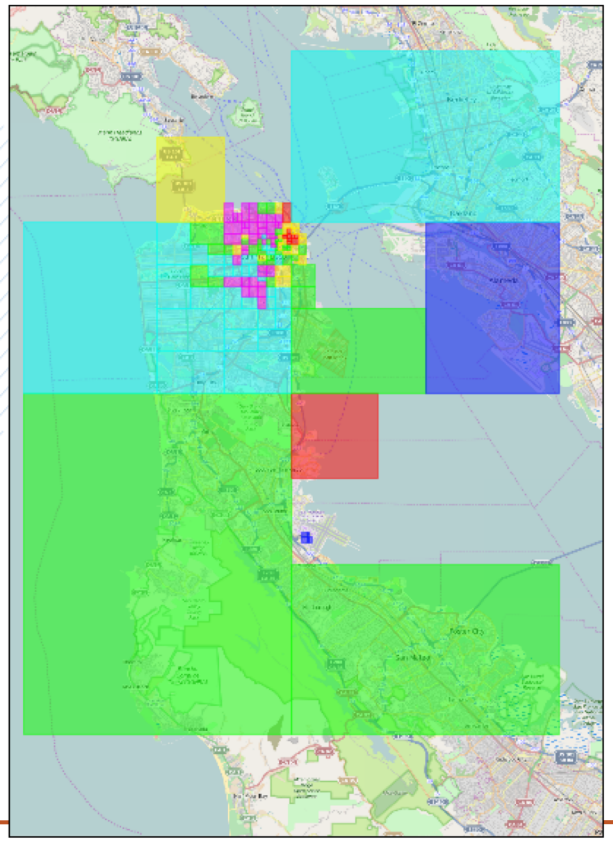
#dropoffs
Airport

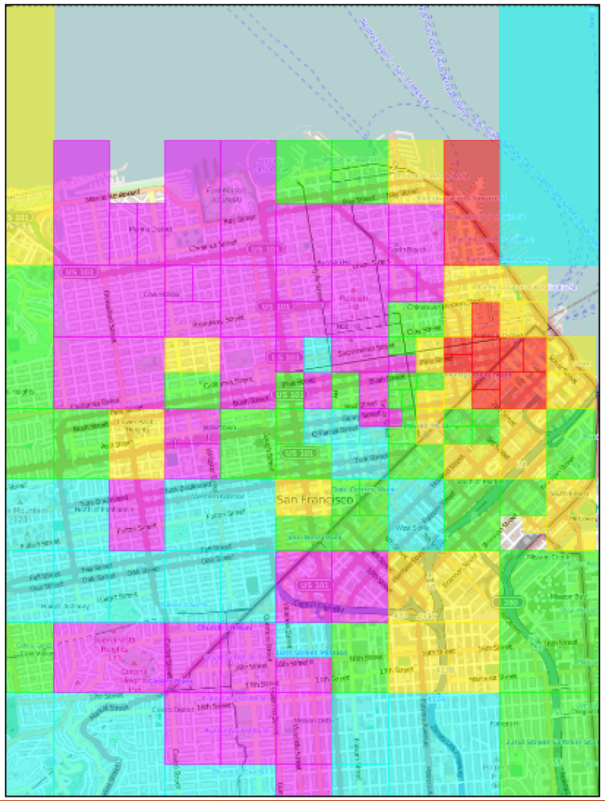


Week

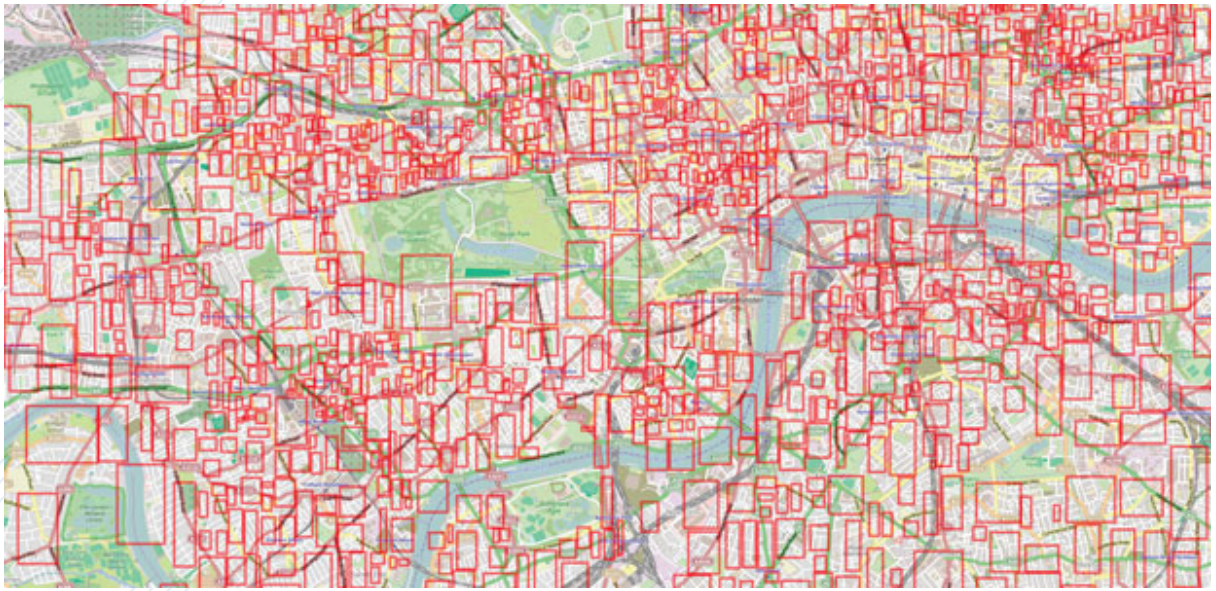
Location clustering based on usage

- Data-based segmentation (KD-tree)
- K-means clustering





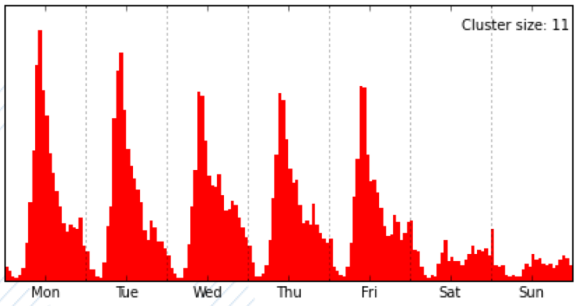
Intermezzo: Airbnb

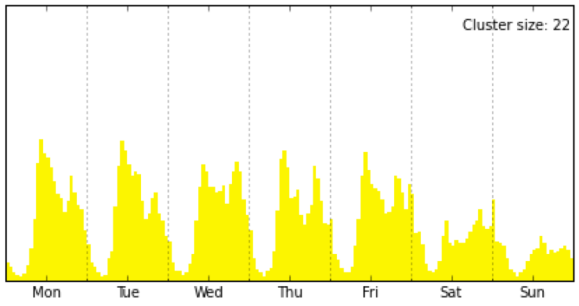


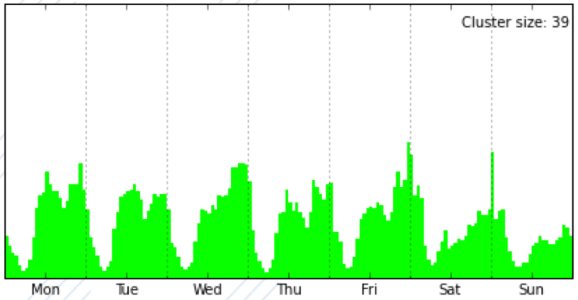
Won't You Be My Neighbor? Algorithms use historical pricing data to group properties into detailed microneighborhoods, as shown on this map of London.

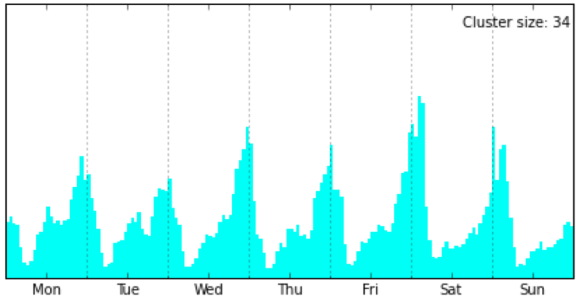
[Dan Hill (product lead at Airbnb), IEEE Spectrum, 2015]

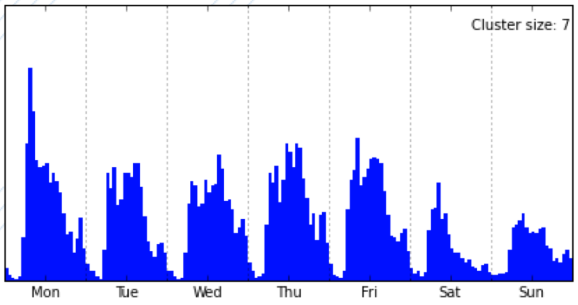
Location clustering based on usage

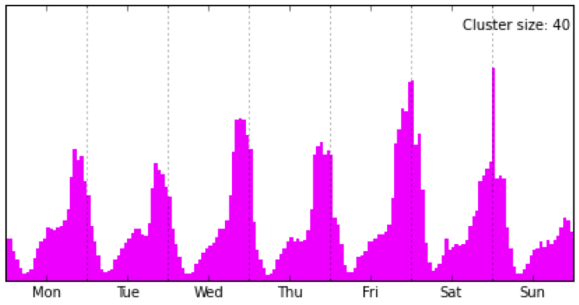


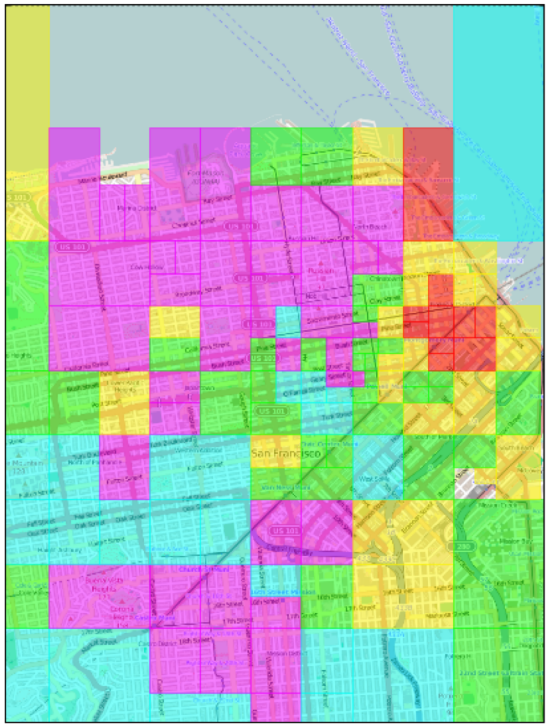




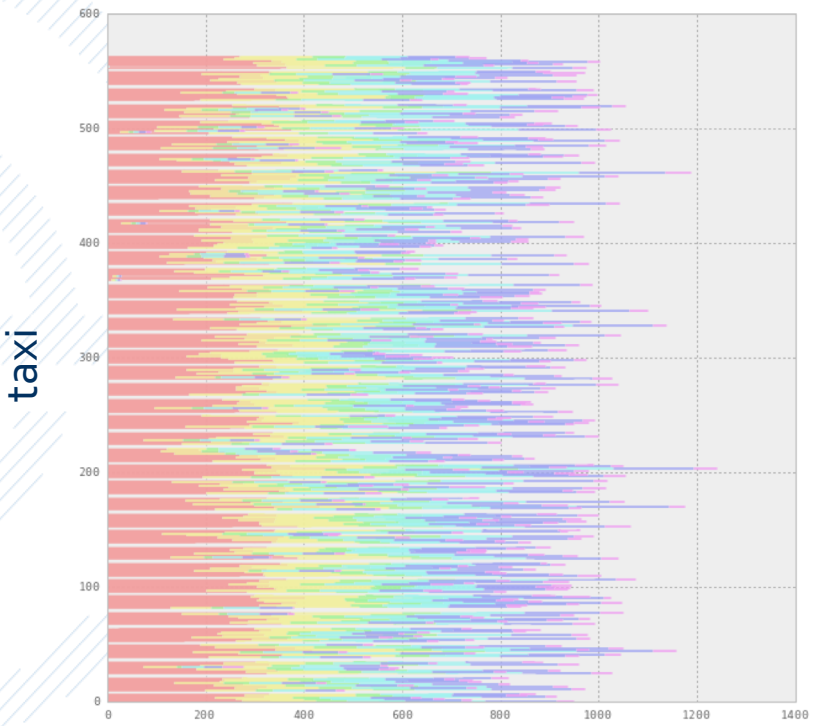




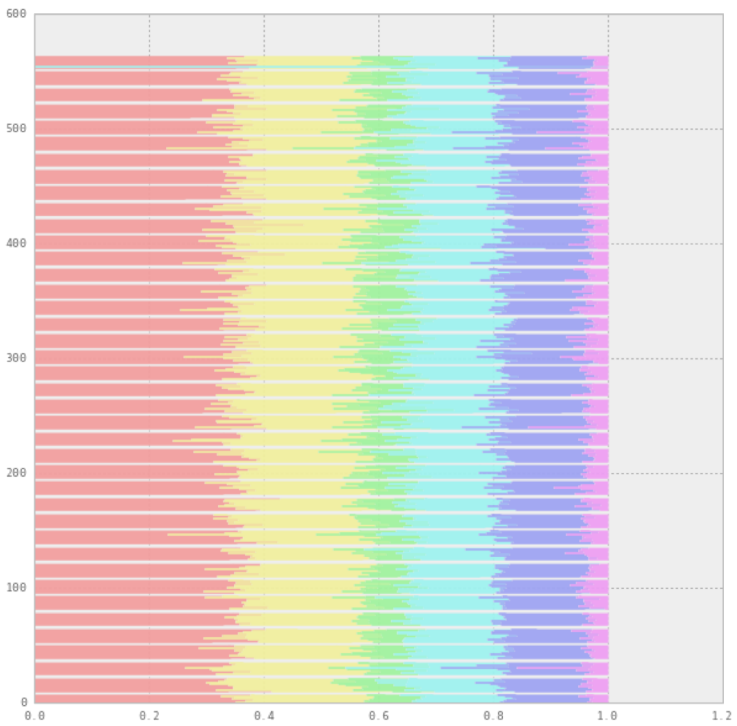




Cab distribution over the clusters



#pickups



%pickups

4th example use case
Using product usage data to make a product smarter

Wie gebruikt het product?

+

Hoe wordt het product gebruikt?

+

In welke omstandigheden wordt het product gebruikt?

+

Hoe is het product geconfigureerd?

Waarom gebruiksdata verzamelen?

Monitoring

Sensoren en externe databronnen laten toe om

- **de staat van het product**
- **de externe omgeving**
- **de werking en het gebruik van het product**

te monitoren, en indien nodig alarmen of notificaties uit te sturen

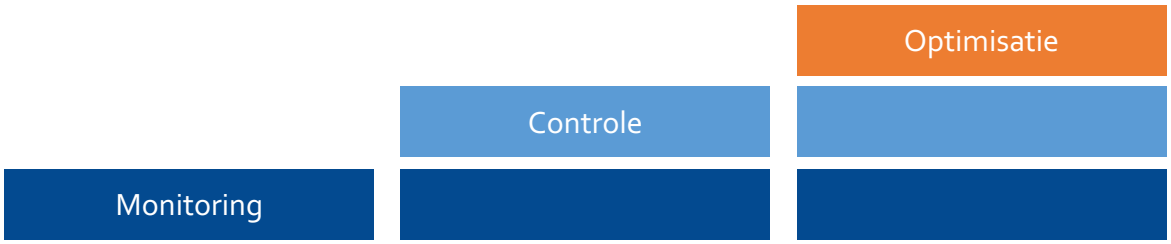
Bron: https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Waarom gebruiksdata verzamelen?

Monitoring	Controle
Sensoren en externe databronnen laten toe om ▪ de staat van het product ▪ de externe omgeving ▪ de werking en het gebruik van het product te monitoren, en indien nodig alarmen of notificaties uit te sturen	**Software in het product of in de cloud laat toe om** ▪ **de productfuncties te controleren** ▪ **de gebruikerservaring te personaliseren**

Bron: https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Waarom gebruiksdata verzamelen?



Sensoren en externe databronnen laten toe om

- de staat van het product
- de externe omgeving
- de werking en het gebruik van het product

te monitoren, en indien nodig alarmen of notificaties uit te sturen

Software in het product of in de cloud laat toe om

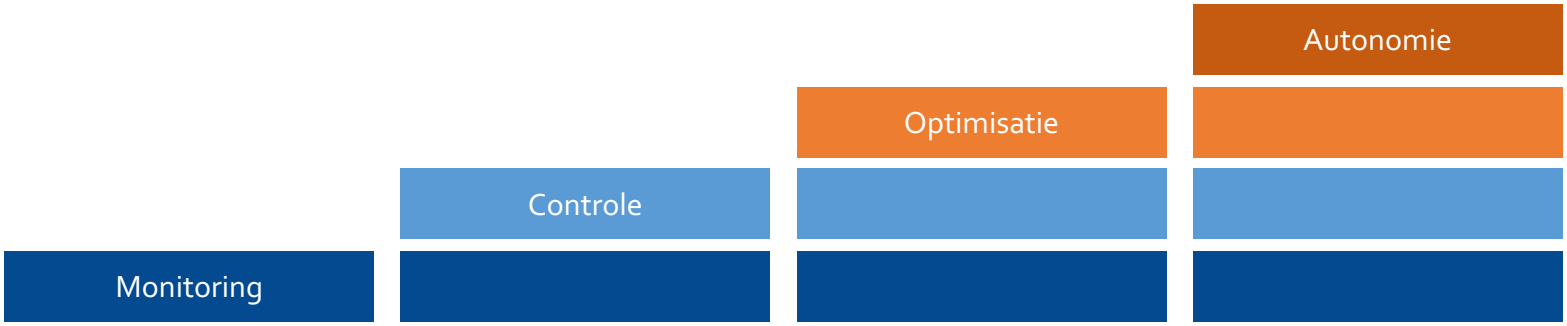
- de productfuncties te controleren
- de gebruikerservaring te personaliseren

De mogelijkheden om te monitoren en te controleren laten algoritmes toe om de werking en het gebruik van het product te optimaliseren om

- **de productperformantie te verbeteren**
- **predictieve diagnoses, service en herstellingen toe te laten**

Bron: https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Waarom gebruiksdata verzamelen?



Sensoren en externe databronnen laten toe om

- de staat van het product
- de externe omgeving
- de werking en het gebruik van het product

te monitoren, en indien nodig alarmen of notificaties uit te sturen

Software in het product of in de cloud laat toe om

- de productfuncties te controleren
- de gebruikerservaring te personaliseren

De mogelijkheden om te monitoren en te controleren laten algoritmes toe om de werking en het gebruik van het product te optimaliseren om

- de productperformantie te verbeteren
- predictieve diagnoses, service en herstellingen toe te laten

Het combineren van monitoren, controle en optimalisatie laat toe om

- **de autonome werking van het product**
- **zelf-coördinatie van de werking i.s.m. andere producten en diensten**
- **autonome productverbetering en personalisatie**
- **zelf-diagnose en service**

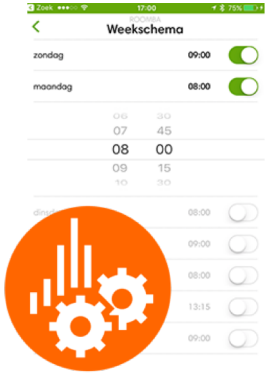
mogelijk te maken

Bron: https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

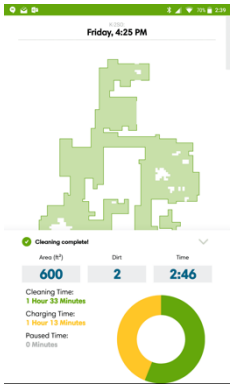
Wat is gebruiksdata?

Voorbeeld 1: Roomba robotstofzuiger

- Gebruikshygiëne (bv. filter en sensoren regelmatig schoonmaken, etc.)
- Ruimte vrij van losse obstakels
- Correcte reiniging



- Aantal reinigingstaken
- Totale tijd voor taken
- Ingesteld weekschema



- Aantal keer 'Dirt Detect'
- Hoeveelheid obstakels
- Oppervlakte woning
- Aantal kamers

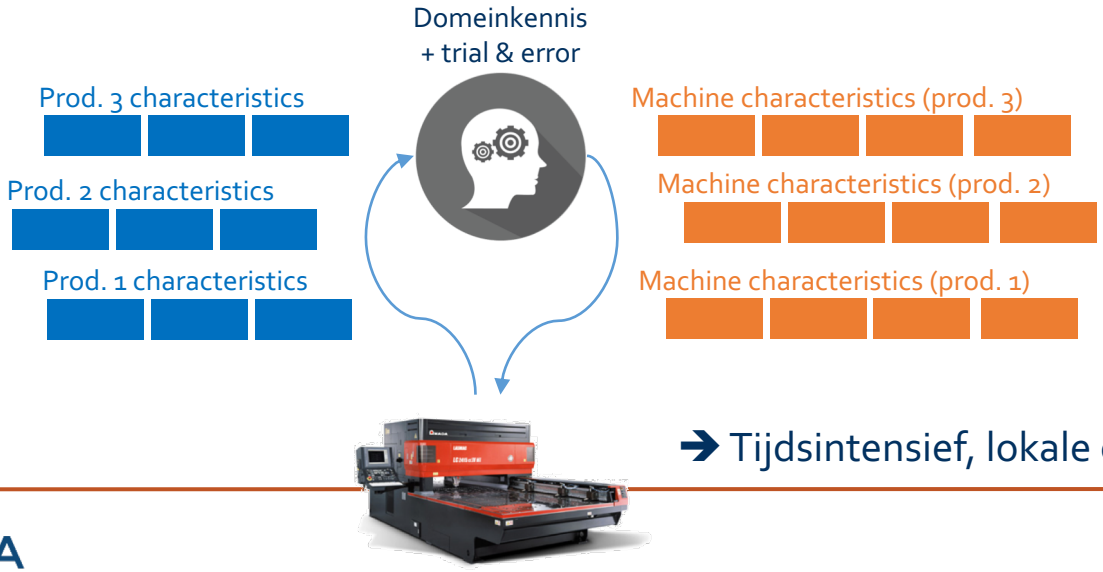
- Type batterij
- Type borstels
- Firmwareversie

Voorbeeld 2:
Optimalisatie van het product-configuratie proces

Context: industriële productiemachines (bv. laser cutters) dienen geconfigureerd te worden afhankelijk van het ruwe materiaal dat verwerkt wordt, de verwerkingssnelheid, de productvorm, etc.

Business question: Hoe kunnen de optimale machine settings bepaald worden voor een gegeven productconfiguratie?

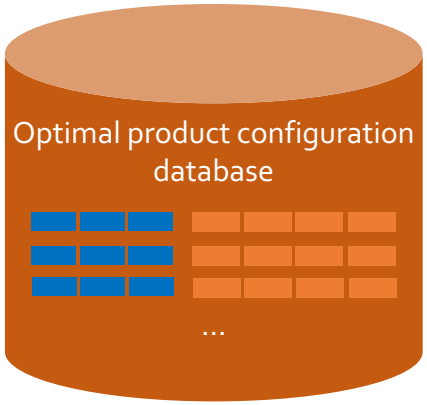
Huidige aanpak: manuele bepaling van machine settings op basis van domeinkennis en trial & error (door operator, machinebouwer, etc.)



➔ Tijdsintensief, lokale optima, etc.

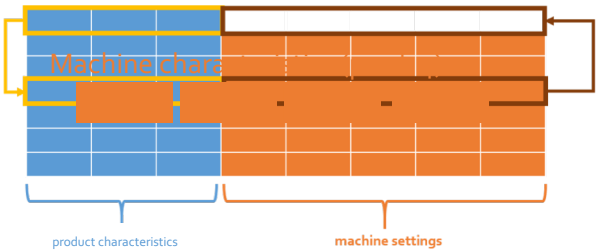
Voorbeeldcase:
Optimalisatie van het product-configuratie proces

Data-gedreven product-configuratie:



Data analytics

Result: data-gedreven methodologie voor het afleiden van een afstandsmaat tussen elk paar productkarakteristieken, om op basis hiervan machine settings af te leiden en aan te bevelen om het configuratieproces te initiëren.



From smart, connected products…

Prod. 3 characteristics

Prod. 2 characteristics

Prod. 1 characteristics

Prod. 4 characteristics

Domeinkennis + trial & error

Machine characteristics (prod. 3)

Machine characteristics (prod. 2)

Machine characteristics (prod. 1)

… to a smart, connected factory with the help of data analytics.

➔ Voordelen: tijdsefficiënter, objectiever, globale optima, etc.

Technology stack we typically use

- Data storage
  - SQL, Document-based (MongoDB, CouchDB), Graph-based (Neo4j), …

- Popular machine learning & data mining libraries and toolkits
  - R, Scikit-learn, statsmodels, Pandas, Weka, SciPy, NumPy, …

- Visualization
  - node.js, D3.js, vis.js, Matplotlib, ggplot2, …

- Deployment
  - Docker, Ansible, Chef, Jupyter notebooks, …

Most of this work happened in **collaboration with students** during their internship, MSc thesis, or summer job ...

UNIVERSITEIT GENT

Joren Van Severen
Mathias Putman
Simon Buelens

Milena Angelova

Technical University of Sofia

Jonas Maeyens
Lorenz Pensaert
Sander Sienaert
Michiel Van Lancker
Michiel Dhont
Rik Sergoynne
Glenn Coppens
Ruben Van Wanzeele
Kenneth Sterckx
Sara Mikolajczak

KATHOLIEKE UNIVERSITEIT LEUVEN

DTAI

Elena Nikolskaya
Christaan Leysen
Oscar De Somer
Thomas Kutz
Wojtek Kuberski
Laurens Teirlynck
Matthieu Vendeville
Kristijan Shirgoski

HOGESCHOOL

Jan De Geest
Mathias De Roover
Jasper Van Audenaerde
Matthias Story
Timothy Buyle
Jonas Geerts
Toon De Pauw
Robin Claus

HoGent

Kilian Hendrickx

Andriy Zubaliy

Universiteit Antwerpen

Thomas Anciaux
Bob Van Looveren

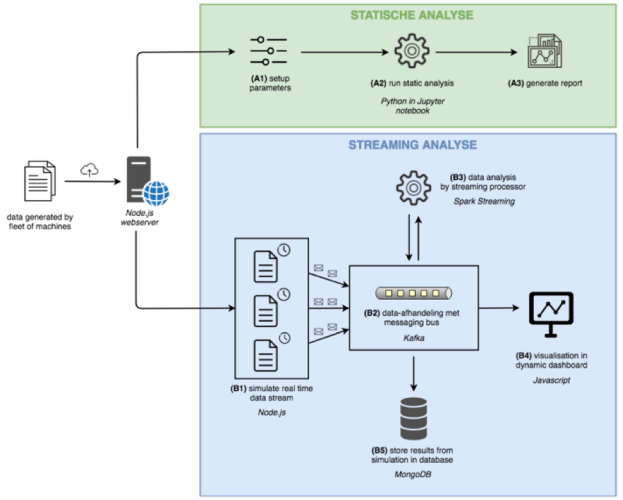
Pierre Dagnely
Steven Beyen
Florian Delporte

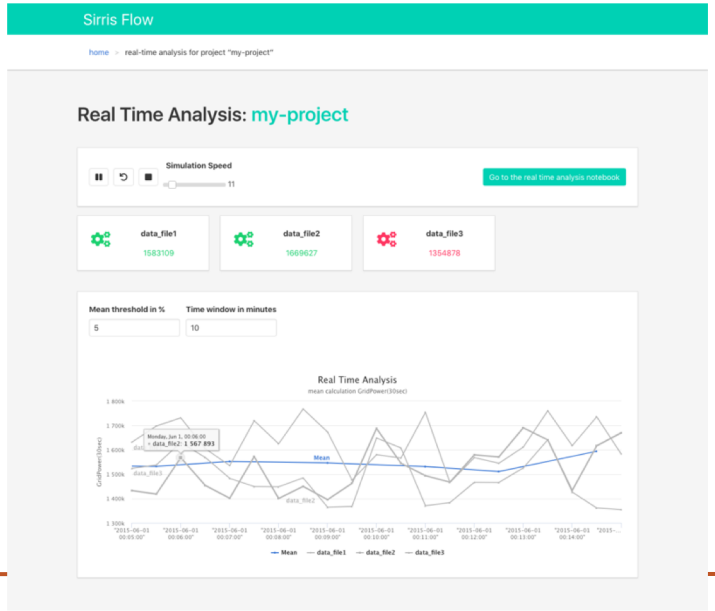
ULB

Looking for innovative internship?

- work in an international environment with strong industrial research focus
- get hands-on industrial experience, with state-of-the-art technology
- expand your knowledge and skills
- influence the subject and work on something you really like

...in an informal working environment, with flexible office hours & work location





EluciDATA

EluciDATA
Innovation Lab

Questions?

Mathias Verbeke
Mathias.Verbeke@sirris.be
+32 494 03 27 47