To data warehouse or not to data warehouse…

Marlies Van der Wee
Samaneh Bahreini

Mobility Data Forum, May 2016
Rise in mobility data, from different sources, setting different requirements…
Main research questions

- How can the large amount of mobility data be handled and structured?

- What is the best political, economic, social and technical strategy for Flemish mobility data?

- What is the value of this data for businesses, governments and research institutions?
How to handle mobility data?
Categorization of mobility data

- Levels: primary, secondary and tertiary
- Types: structured, unstructured and semi-structured
- Sources: internal, external
- Born digital or born analogue
- Features of the data source
  - Frequency of collection
  - Spatial coverage
  - Equipment requirements
- Usage goals
Categorization of mobility data example

<table>
<thead>
<tr>
<th>Mobility data</th>
<th>Features of data sources</th>
<th>Usage goal</th>
<th>Type</th>
<th>Source</th>
<th>Level</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic flow data</td>
<td>In-road sensors</td>
<td>Incident detection</td>
<td>Structured</td>
<td>Internal</td>
<td>Primary</td>
<td>Speed, Travel time, Volume, Vehicle classification, Occupancy</td>
</tr>
<tr>
<td></td>
<td>Inductive loops</td>
<td>Congestion or queue identification</td>
<td>Semi-structured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Magnetic induction coils</td>
<td>Vehicle detection</td>
<td>Analogue Digital</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Magnetometers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pneumatic tubes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Piezoelectrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weigh-in-motion equipment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Truck weighing technologies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Typical collection Equipment</td>
<td>Spatial coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equipment</td>
<td>Temporal coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usually spaced at &lt;= 1.6 km; by lane</td>
<td>Report at 20-to 60-second intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Different requirements for data

- Raw or pre-processed?
- Individual or aggregated?
- Real-time or historical?
- Continuous or sample interval?

Pyramid diagram with levels:
- Data
- Information
- Knowledge
- Wisdom
Single point of contact

One responsible within the company that manages the available data

- Integration from multiple sources
- Analysis and reporting (though manual)
- Historical data
- No massive data sets
- No automated analysis and reporting (hence slow)
- No real-time data
Data portal

Integrated, centralized website with search engine

- Multiple sources (though not integrated)
- Searchable
- Both real-time and historical data

- No actual data storage
- No automated analysis and reporting
Data lake

Large storage repository, no requirements with regards to type or standard
"Give me whatever you have"

- ✔ Massive data sets, multiple sources
- ✔ Reporting
- ✔ Both real-time and historical data
- ✗ No integration of sources
- ✗ No standardization
- ✗ Limited reporting, no analysis
Database

Organized, structured collection of standardized data

- Standardized and structured
- Searchable
- Both real-time and historical data
- Reporting, no automated analysis
- No data cleaning
- Limited verification on data quality
Data warehouse

Intelligent database

- Searchable, reporting, analysis
- Data processing, cleaning, checking
- Both real-time and historical data
- Expensive in installation and maintenance
Different requirements for data need different handling systems

One responsible within the company

Centralized online overview of available data

Quasi unlimited storage, "give me what you have"

Storage and reporting on standardized data

Storage, reporting and analysis on standardized data
Cost estimation of the data handling systems

- Based on
  - Literature review
  - Experts interview
  - BEMES model as a cost calculation tool for technical equipment
Cost estimation of the data handling systems

Comparing the total cost per TB of data per year
PEST for mobility data in Flanders
PEST analysis

Political

Economic

Socio-cultural

Technical
PEST: mobility data in general

- Political
  - Privacy
  - Security
  - Ownership

- Economic
  - Business model potential
  - Data collection cost
  - Data storage cost

- Socio-cultural
  - Value for individuals
  - Value for society
  - Value for the environment

- Technical
  - Quality
  - Accuracy
  - Reliability
PEST: data handling systems

- Organizational politics
- Responsibilities

- Energy footprint of data centers

- Cost-benefit trade-off depending on characteristics and needs

- Infrastructure and components
  Amount of data
Value of mobility data
Big data opportunities

- New business models
- Intelligence on own business or competitors
- Enriching existing products/services
- Increase user experience
- Enter new markets
- Increase efficiency/productivity
- App developers
- End users
- Government and local authorities
- Research community
Data users

Data requirements

Data handling systems
Data requirements

- Standardized
- Searchable
  - Categorized
  - Structured
- Minimal delay (latency)
- Pre-processed
  - Cleaned
  - Missing data points filled
- Processed
  - Analysed
From data users to requirements

- App developers
- Researchers
- Citizens
- Logistic companies
- Real estate companies
- Authorities/policy makers

Data
- Standardized
- Minimal delay

Information
- Searchable
- Pre-processed

Knowledge
- Processed

Wisdom
From handling systems to requirements

<table>
<thead>
<tr>
<th>Following standards</th>
<th>Searchable</th>
<th>Min delay</th>
<th>Pre-processed</th>
<th>Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image36.png" alt="Image" /></td>
<td><img src="image37.png" alt="Image" /></td>
<td><img src="image38.png" alt="Image" /></td>
<td><img src="image39.png" alt="Image" /></td>
<td><img src="image40.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Quantitative analysis

- Find the nearest and cheapest place to park

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Types of benefits</th>
</tr>
</thead>
</table>
| Economic competitiveness  | ▪ Travel time saving  
                          ▪ Vehicle operation saving  
                          ▪ Improve local and regional transit  
                          ▪ Increased city parking occupancy                                             |
| Quality of life           | ▪ Reduction in the commute time, thus more satisfaction  
                          ▪ Decrease stress level through providing more convenient way of finding parking in the city |
| Environmental impact      | ▪ Reduction in travel time and thus reduction of emissions                          |
| Safety                    | ▪ Reduction of vehicles looking for a parking spot within the network as result of improved guidance on parking availability in the area |
Quantitative analysis

- Comparison of costs and benefits for Ghent

5 year comparison

Open data portal: €500,000
Societal benefits: €180 million

COSTS
REVENUES
Quantitative analysis: Proximus analytics

Proximus analytics

Improved market insights
Get to know your customer better through Proximus analytics
Quantitative analysis: Proximus analytics

- Benefits of data analytics
  - Direct dialogue with your customers
  - Re-develop your products
  - Perform risk analysis
  - Smarter marketing

5 year comparison

Data warehouse: €80 million

Benefits for SMEs: €140 million
To data warehouse or not to data warehouse?

depends strongly on the specific case or application…
Thank you for your attention!

Any questions?

saminbahrini@intec.ugent.be
marlies.vanderwee@intec.ugent.be