Harnessing Big Data

Workshop 5 | THREDBO 2015 | Santiago de Chile

Chair: Marcela Munizaga Rapporteur: Gabriel Sánchez-Martínez

Canada Chile Japan Singapore UK Uruguay USA

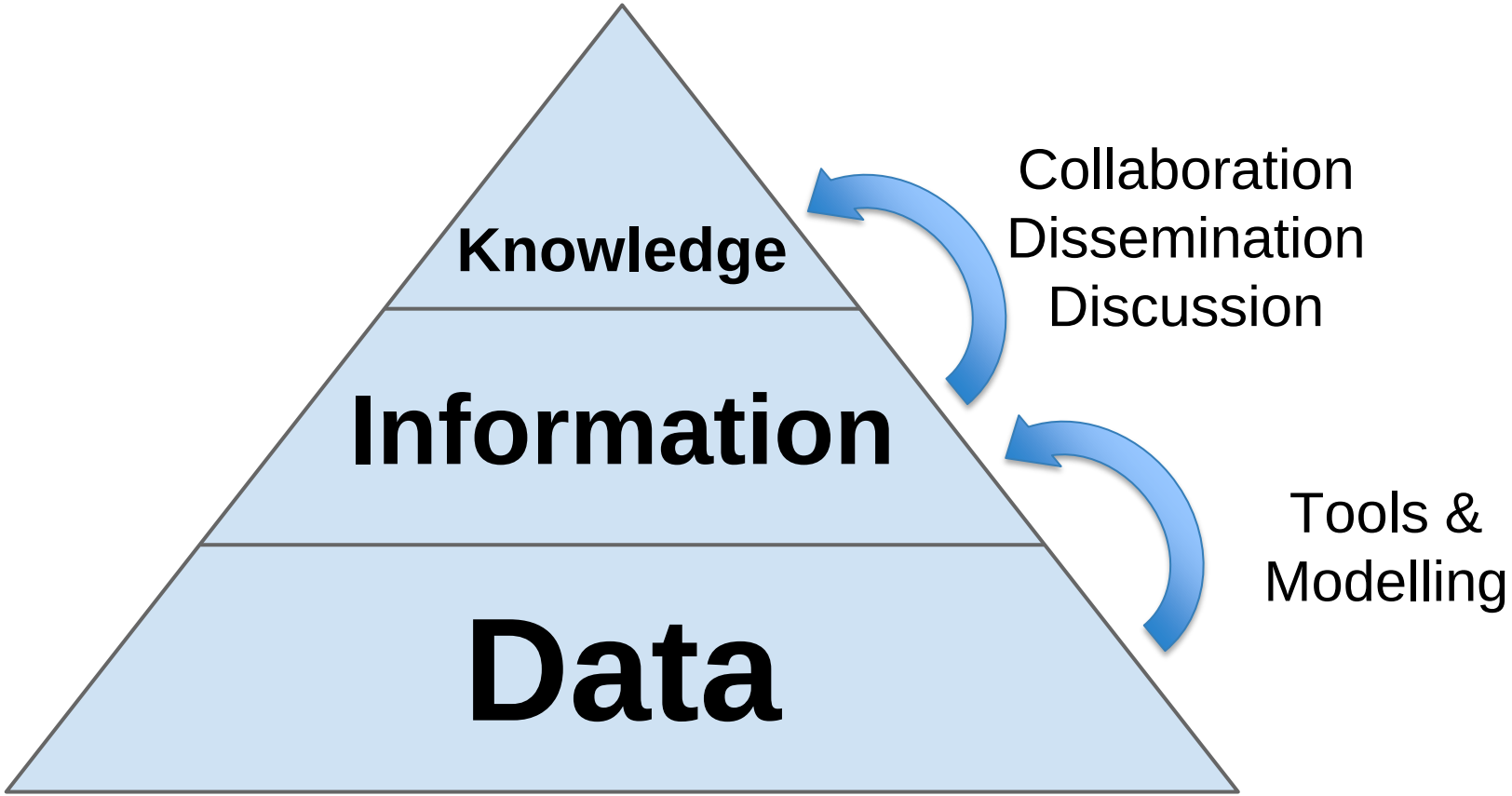
THE UNIVERSITY OF SYDNEY

TransitUC

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

The Open University

Workshop 5 discussion


Knowledge
Information
Data
Collaboration
Dissemination
Discussion
Tools &
Modelling

Identify how **Big Data** is and can be *leveraged* to improve **urban transportation** and **quality of life**

What has been done so far?

What are the outstanding challenges?

How do we bridge the gap?

Data Sources

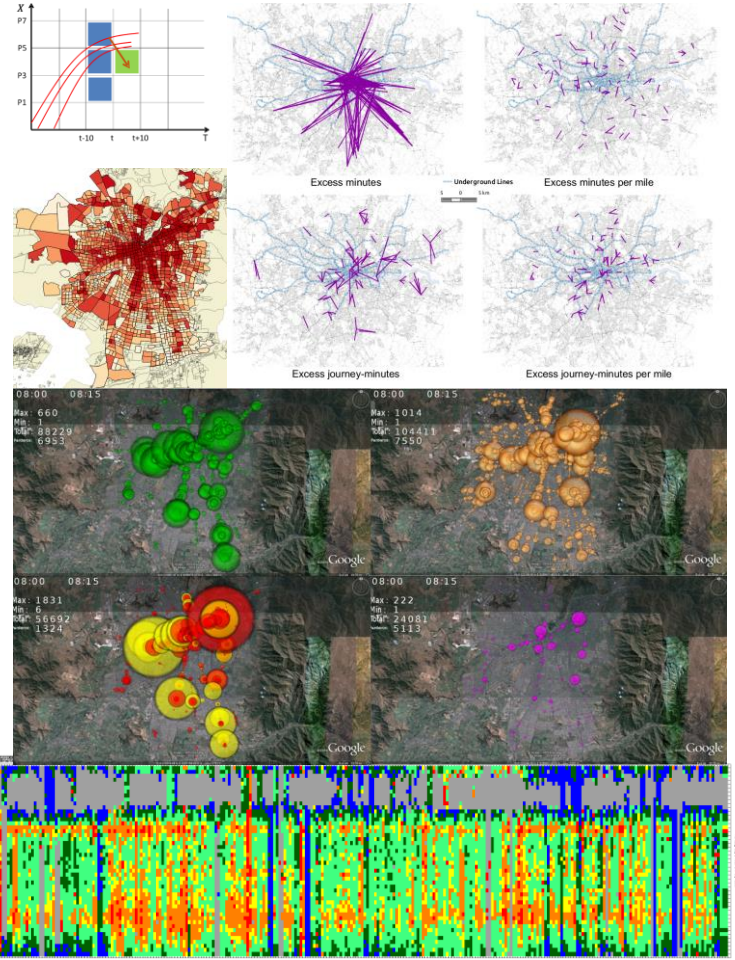
Tools

Statistical
Programming
Optimization
Simulation
Visualization
Business intelligence
GIS/Mapping
Databases
Data Warehousing

State of the Art

Inference of travel patterns
Operations monitoring
Performance evaluation
Real-time predictions
Human behavior analysis
Data mining

→ Focused on public transport

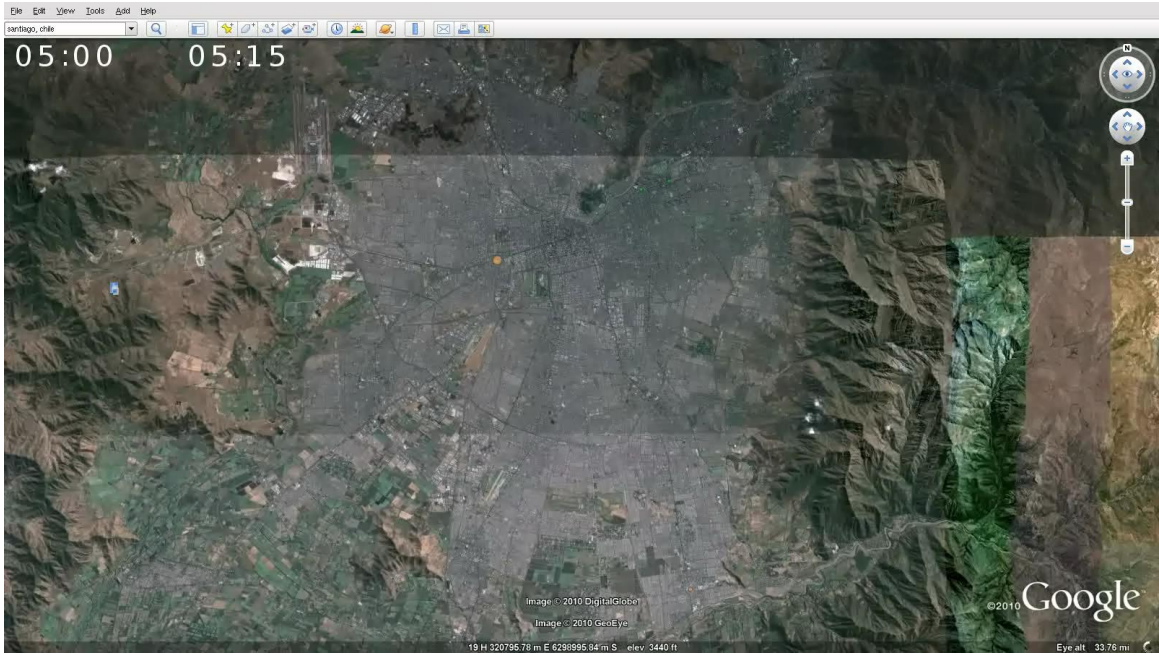


Findings

Wide diversity in utilization
Developing research area
Low cost, high-definition, but partial view
Significant progress ⟺ much more to be done!
Huge potential

Research Needs

- New data sources
- New tools
  - cleaning
  - fusion
  - analysis
  - visualization
- Predictive models
- Validation
- Indicators



Policy Recommendations

Standardization of data
Share successes and failures
Share methodologies, approaches, and practices
Disseminate!
Open data
Make cities smarter
Safeguard user privacy **and** access to data
Data-driven contract evaluation

Synthesis

More thinking, workshops, including practitioners and operators
Use big data not just for planning and control → many other uses, across stakeholders
More research, more applications, more sharing → more fun!

Jacqueline Arriagada, Owen Bull, Charles Fleurent, Ricardo Giesen, Brendon Hemily, Felipe Hernández, Rodrigo Hurtado, Guillermo Maureira, Antonio Mauttone, Marcela Munizaga, Toshiyuki Nakamura, Carolina Palma, Cristóbal Pineda, Gabriel Sánchez-Martínez, Hiroshi Shimamoto, George Sun, Sebastián Tamblay, Alan Valdez, Cecilia Viggiano

What is Big Data?

- Large amounts of data that did not exist until now, as a consequence of increased automation, improvements in sensing, storage and communication technologies
  - overwhelming, often exceeding our capabilities to understand it and make use of it.
- New tools and skills are needed to process and analyze
  - e.g. collaboration with computer scientists
  - Other fields (e.g. astronomy) have bigger datasets, perhaps we could benefit from reaching out to specialists outside of the transport field.
- Characteristics of big data in computer science
  - volume (massive)
  - velocity (needs to be processed in real-time)
  - variety (i.e. different kinds of data)

Specific Goals

Enhance service and operations planning
- Measure reliability as experienced by passengers
- Finding the factors leading to unreliability
- Real-time control and incident management

Understanding passengers
- Develop passenger profiles, understand travel patterns and behavior
- Encourage and predict behavioural change
- Provide customer information and interfaces

Planning
- Evaluate urban planning and policies (economic growth, quality of life, environmental impacts, equity)
- Improve network planning

Competition and Innovation
- Create new business models
- Provide demand-responsive services
- Innovative products and services (autonomous vehicles)
- Performance monitoring (for contract tendering and incentives)
- Make better use of existing resources

Contribute to other areas/fields
- Town planning
- Location
- Project evaluation

Data Sources

Automated vehicle location (GPS)
Automated fare collection
Automated passenger counting
Network and scheduling data (GTFS)
Crowdsourcing (Waze, Moovit, Twitter)
Urban sensors (CCTV, Wi-Fi, Bluetooth)
Geospatial data (zoning, land use)

Data Sources

Different data sources have different ownership and access issues. Negotiating data access in contracts is very important. Academics must build trust to gain access to data.

Public Transportation-Related (directly)

- Automated vehicle location (GPS) (if negotiated in contract)
- Automated fare collection (if negotiated in contract)
- Automated passenger counting (if negotiated in contract)
- GTFS
- Network and scheduling data
- Vehicle Health Monitoring
- Travel Surveys

Others

- Crowdsourcing (e.g. Waze)
- CCTV, traffic coils
- Wi-Fi and Bluetooth detectors, Internet of Things, ubiquitous computing
- Surveys
- Geospatial data, e.g. zoning, household and job densities
- Weather
- Vehicle sensors

Typology of Uses of Data

- Corporate / Executive Monitoring and Reporting
  - Dashboards (External consumption)
  - Executive Management (KPIs / alarms)
  - Monitoring of Operator Service Contracts
- Service / Operations Planning
  - Service and Performance Monitoring (e.g. On-Time Performance, Schedule Adherence, Loads vs capacity (crowding), Ridership profiles, OD Matrices, Reliability
  - Strategic Analyses (e.g. Corridor planning, Alternatives analysis)
  - Tactical Analyses
  - Policy Analyses (e.g. Benchmark against peers, accessibility/equity)
- Scheduling (e.g. Actual running times)
- Line Management (e.g. Operator Performance)
- Real-Time Pro-active Operational Control (e.g. prediction of gaps / bunching, placement/insertion of standby buses)
- Exploratory / Data Mining (e.g. Safety Analyses, Accessibility and Equity)
- Users / Customers
  - Real-time information about services
  - Customer experience

Are we missing data sources?

- Sociodemographic
- Land use data, GIS
- Detailed and unified description of urban infrastructure related to public transport, e.g. number of lanes, segregated busway, tramways, metro lines, traffic signals
- Sensors (weight for load profiles, transit priority system)
- Traffic data
- Data about events
  - streets blocked
  - protests
- Weather

Missing Data Sources - mobile phone data

Feedback from users e.g. collaborative and opportunistic tools (moovit, tiramisu).



Through their smartphones, users become part of the sensing system and contribute data, while simultaneously gaining access to it.

Essential when transit user information required for multi-leg and multimodal journeys (routes, timetables) is missing.

Contributes to bridging the gap between operators, regulators and users. This increases communication and perceived legitimacy.

Are we taking enough advantage of available data sources?

some are, others are not - who are we? (academics vs operators vs planners)
data fusion
future travel demand
to promote public transport use
for applications related to cycling and pedestrians
sometimes an agency needs to be encouraged
    contract performance-based penalties and incentives
It can be challenging to know which agencies are taking advantage of big data.
    importance of leaders who promotes usage of big data
    hiring of graduates from university partnership programs
CCTV
100% fleet equipped with APC and modern hardware
need to have teams that have both transportation and computation expertise
researchers should show immediate practical value to agencies (through case studies - a variety to meet the needs of different agencies), because it encourages further research and partnership with researchers

High Level Challenges for Transit Authorities / Operators

- Lack of Understanding by Senior Management / Policy Board
- Corporate Data Management Challenges
- Challenges Related to Ensuring Data Quality
- Challenges in Using ITS Data Once Cleaned

Constraints to access data

Political
Ownership (e.g. vendors may own the data). Sometimes data is for sale.
Trust
Skill (being able to do something with it, even if you have it)
Worries and legal constraints about confidentiality, privacy
Commercial sensitivity (e.g. equal accessibility by all bidders, Uber)
Data not being stored, proprietary sensors and systems
Poor data quality (sensor failures, data cleaning)
Poor documentation and understanding of data structures
Lack of standards, open, data structures
    help agencies give out data without having dedicated staff

Issues with AVL/GPS data

- Cleaning and filtering is necessary
- Sometimes the polling interval is too long
  - Chile every 30 s, sometimes every 90 s
- Imprecise geocoding of bus stops
- Multiple bus stop inventories and bus stop IDs
- Data matching problems between Scheduling / CAD/AVL systems
- Issues with first/last stop and detecting the beginning and ending of a trip (e.g. Layover locations, Loops, Branches, Negative loads, etc.)
- Recognizing / addressing Corrupt or ”Bad Day” Data in automatic reports of KPIs (e.g. snowstorms, protests, sports events)
- Factoring up samples / biases from data sources
- Adjusting to ever-changing schedules and stop locations

Categories of predictive data

real-time control, e.g. bunching
demand prediction
operations planning applications, e.g. identifying factors leading to the characteristics of service
connected vehicles for safety and mobility applications, V2X (vehicles to anything), Internet of Things, M2M
electric, driverless vehicles

Modelling & Methods

Destination and transfer inference
Linear and logistic regression
Simulation
Clustering
Classification

- Making big data smaller so that it can be an agent of change
- Advantages of open data

Who are we missing?

Bus operators (private)
Passengers
City planners
Politicians
Private transportation providers (e.g. Uber)
Computer and data scientists and engineers
IT people within agencies (more/others)
Traffic controllers
Big data and business intelligence/analytics vendors

What are we missing?

Fare evasion
Sociodemographic data
    link with surveys
Social media (e.g. people “checking in”)
Fare elasticities
Long-term behavior change (because of card churning)
Travel purpose
    infer work, home, other activities
Correlating customer experience and quality of service perception
Politics and political pressures
Social networks (e.g. households plan trips together)
Social equity

Challenges - Morning 09-01

- Issues with data
- Endogeneity
- Sometimes big data does not provide the information of interest
- There may be behavior biases with people having smart cards vs. those paying cash

measuring equity of service and fares
endogeneity issue: more accessible real estate tends to be more expensive
governments seldom have access to sociodemographic data tied to AFC
sociodemographics may be obtained voluntarily if an incentive is provided

Privacy

Big Data projects can capture a large amount of personal or commercially sensitive data. Perceived legitimacy of institutions collecting and analyzing personal data is important to secure citizen’s participation in schemes.Care must be taken to safeguard this information and to make it available only in anonymized or aggregated forms top prevent its misuse (e.g, by terrorists or criminals).
Additionally, it is important that users feel in control of their personal information (e.g, through opt-in schemes). And that they perceive that their personal information contributes to the creation and fair distribution of the value created through big data.

Are we obtaining enough value from the data?

We are just starting. Many challenges remain.

Thousands of transit systems, but very few getting the benefit.

We should be more user-oriented.

The data is relevant for non-transport issues, such as equity. Other perspectives can be added.

Transit signal priority.

How can we contribute to diminish the gap?

Create a network of researchers, industry, and transit agencies working with transit and big data.
    Transformative Data sub-committee of TRB
publications of methodologies may not reach operators
make the business case, help senior management make their case, short case studies
encourage open data and standards
    give more data feeds to encourage app developers to take advantage of big data and provide free solutions that people can use.
take advantage of trends, e.g. popularity of smart cities
    the focus of the debate should be a successful city, not transit in itself. Transit follows.
encourage it through contract incentives/penalties that can lead to financial gain from data-driven approaches
provide a big benefit from contributing data in a standard format, e.g. Google Transit and GTFS
    e.g. load profiles, crowding metrics, running time analyses
bring the dialogue to associations such as SIBRT and UITP
    create a big data category of award. good publicity.
agencies hiring students after research in transit, hence funding for graduate studies in transit by agencies
getting governments to disseminate knowledge, e.g. through professional capacity development programs

Are the estimations we obtain reliable?

More validation is required.
In general we do think it is reliable.
It provides a high-definition picture of part (sometimes most) of the elements, but we completely miss other elements
e.g. fare evasion, sociodemographics, detailed bus stop visit data.
More trust on some data types than others.
data errors vs. model errors

Who is a more effective agent to make big data useful? Are we its creators or opportunists?

- Researchers are creating methods to use the data
- Some are installing sensors out in the city and want to make a business out of it
- There is a trend to publishing data, e.g. timetables, real-time vehicle locations, but not real-time loads
  - encourages citizens to create apps, free for the transit agency

Smart Cities

Smart cities create opportunities for change, bringing stakeholders together (even stakeholders who usually would not think about transport) and creating momentum for big data applications.

The smart city vision provides opportunities to build leadership, trust, and partnerships

There are different models of smart cities: more efficient, competitive, sustainable. The use of technology to monitor urban flows (water,energy,transport) is emerging as a dominant vision.

Cities that use the data available in the benefit of the citizens, with different agents working coordinately for a common goal

Cities in which different agencies share data to solve problems that may not be transit, e.g. healthcare.

- e.g. bicycle system collaborating with transit
- e.g. targeting where security should be to make people safer on transit
- e.g. super transit app that aggregates all transport modes, including taxi, uber, transit, bicycle
- e.g. Mexico City making driver's license a smart card that can also be used to pay for transit and bicycle share

What are the perils of big data?

Different branches of government have different data and they do not share data or expertise.

Failures may lead to key people losing trust on big data and not supporting it in the future.

Opening too much data may lead to someone making bad use of data, e.g. someone without an understanding of transportation making incorrect policy decisions, e.g. terrorists using open data to plan attacks. This exacerbates #2. Technical expertise is required.

Human Behavior

- For what purpose could we like to use big data?
  - Strategic-level (e.g. infrastructure planning)
  - Tactical-level (e.g. improving bus service)
  - Operational-level (e.g. scheduling, maintenance)
- Extracted data level
  - Aggregated (by location, time, social demographic)
  - Individual
- For what mode?
  - Public Transport(metro,rail,bus),Private car,Taxi,Commercial vehicle,bicycle, on foot
- Which time lenge?
  - Daily,Weekly,Monthly,Yearly(short-term/long term)

Human Behavior (cont.)

- How to analyze big data to understand travelers' behavior better?
  - Data fusion (e.g. combination of big data and survey)
  - Visualize travelers' behavior (e.g. by location, by time, by day, by social demographic)
  - Categorizing travelers' behavior
  - Longitudinal analysis (e.g. panel analysis)
  - Behavior experiment with big data

Linking to Customer Experience

We need to link the use of data for monitoring performance, etc. to the experience of the customer, as recommended by the European Union Standard on Quality of Public Transport (EN 13816 / EN 15140). More research is needed on how to measure / link to customer experience

Can we use this type of data to develop predictive models?

Most of the data has been to analyze the past. Can we use data to develop predictive models? What are the challenges?

One of the challenges is online cleaning and validation of data feeds.

Some statistical learning tools and others borrowed from computer science may become increasingly useful.

Hi-tech agencies have dashboards and real-time KPIs

USA TODSS, etc.

It's not only a matter of developing tools. The learning process of practitioners is just as important. Business strategies should be developed and implemented. Education of transit system staff.

Sometimes interfaces require human interaction through non-standard graphical UIs, instead of standard machine-readable feeds.

Prediction of hardware failure. Vehicle and infrastructure health monitoring. Optimization of maintenance plan.

Running a vehicle more intensely than others to accelerate the appearance of problems and learn what the future maintenance challenges will be. "Hare"

real-time mining social media to identify events and problems as or before they occur

transfer coordination in low-frequency service when users alert the agency they wish to transfer

What tools do we use?

Statistical
Programming
Optimization
Visualization
Business intelligence
GIS/Mapping
Databases and Data Warehousing

Statistical Tools

Matlab
R
Stata
SPSS
Excel/ACCESS

Programming

- Python
  - sklearn - machine learning package
  - SciPy/NumPy/matplotlib
- C++
- C#
- Java
- Shell scripts

Optimization Packages

CPLEX
Gurobi
Minos
GLPK

Visualization and Business Intelligence

QlikView
iGraph
Tableau
D3.js
CartoDB

Business Intelligence
- Rapid Miner
- SAP/BI
- IBM/Cognos
- Oracle/BI
- MicroStrategy
- HASTUS-Analytics

GIS and Mapping

ArcGIS
QGIS
TransCAD
Google Earth
OpenStreetMap

Databases and Data Warehousing

Relational Databases
- PostreSQL / PostGIS
- Oracle
- SQL Server

Data Warehouse
- TeraData
- Oracle

What types of tools are we missing?

Do we need tools that are specific to transport?

What can be done to increase tangible realized gains from big data?

Make the business case, show how to save money
Make sure the public owns the data
Quality assurance of data and sensors, identifying standards of accuracy
Producing user-friendly interfaces
the will of managers and senior staff can make a difference
Improve communication of benefits to authorities and senior management
Technology transfer through universities, consultants
Provide standard tools and data formats
Emphasis on visualization and presentation to communicate information from big data
Provision of customized information to users
Service planning based on detailed demand information
Project evaluation methodologies should require best practices when it comes to using this data
Measuring, documenting, and sharing before-after studies about projects using big data to show its value
Avoiding the success bias. Report failures too.

Can new data improve contracting?

- Can improve performance measuring, which can be used in contracts to give incentives/penalties.
- Can new data be used to build better (e.g. more objective) cases for investments in public transport?
    - Yes, but standardize the methods, KPIs, and the specific mapping of data to KPIs
- Who is responsible for establishing and measuring KPI?
- KPIs should be used not only for incentives and penalties but also for resource allocation and operations planning. The authority should be involved.
- Use KPI’s to increase visibility of performance and encourage competition among transportation providers.
- The level of sophistication of the provider and the authority are positively correlated. Contract structure and level of competition also influence sophistication and use of KPIs.
- Can we use big data to provide operators with tools to serve customers better? - the contract could require use of these tools - using the tools (for example, real time control strategy) may be more effective than describing requirements with rules...

Potential to improve planning

service & infrastructure planning
operations planning & management

Outstanding Challenges

- Sharing new tools and knowledge so that best practices are adopted across the industry.
- Balancing utility and user privacy
- Getting a complete picture
  - when the data does not cover all users (e.g. biases)
  - when the data does not cover all the key elements (e.g. sociodemographics)
- Lack of standard formats makes it difficult to generalize analysis tools
- Lack of understanding and support from senior management
- Lack of technical expertise

Conclusions

We should make efforts to bridge the gap between academics/researchers and operators.
Predictive indicators.
We need standardization of data
Even though we have done a lot, we need to spread the best practices to many operators, not just the most hi-tech ones.
Use big data not just for planning and control, but for many other uses and across stakeholders.
Recognize that we are having a high-definition view of only part of the system. Keep the limitations in mind.
Share methodologies, approaches and practices. Disseminate.
More thinking, workshops, including perhaps more practitioners, operators.
So far we have focused on the most obvious applications. Much remains to be done.
We can incorporate many other data sources, and use data-fusion.
We must safeguard both user privacy and the lack of legal and regulatory restrictions to access the data, and take advantage of this momentum.
Data can be used to improve service contracts through better monitoring.

Outstanding Challenges

- Setting minimum standards
- Standardization
- Spreading
- Obtaining more value