European/International Joint PhD
in Social Representations and Communication
International Summer School 2016

European Commission REA-Research Executive Agency
FP7 - PEOPLE Initial Training Networks
So.Re.Com. Joint-IDP
(PITN-GA-2013-607279)

Funded by the European Union

From the use of big data to meta-analysis in urban and transportation studies in the society of algorithms

Sylvain Lassarre
GRETTIA/COSYS – IFSTTAR
With contributions from
Latifa Oukhellou, Gerard Scemama
GRETTIA/COSYS – IFSTTAR
Gilbert Saporta CNAM

- Tranportation systems and mobility (reminder)
- Digital revolution and mobility
  - From the users perspective
  - From the actors perspective
- Big data in the web and transportation
- Algorithms for big data
  - Predictive models
- Conclusion

Networks as macro technical systems

- Interconnection
  - physical
  - Flow : persons, goods, energy, information
- Intermediation
  - market/economy. Linking consumers and suppliers of goods and services.
- Three layers
  - low: infrastructure : lattice plus hierarchy
  - medium : infostructure : control-command devices
  - high : final services to consumers
- Three components
  - Sensors
  - Communications
  - Big data

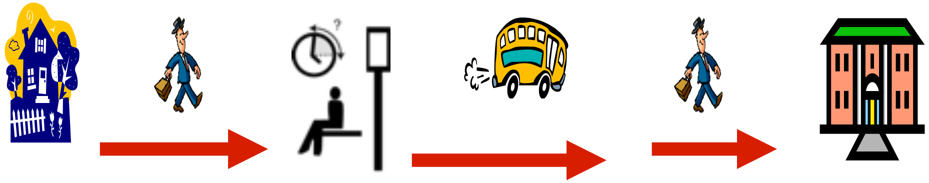
Transportation networks

- Rail and air transport yes
  - train = first physical artificial space coupled with an information system , the telegraph.
  - plane (heavier than air) under control because of radar (from the 2nd world world), wins the competition over the airship (lighter than air)

- Road, waterways and sea transport half-half
  Motorways yes. BRT too

What is spatial mobility ?

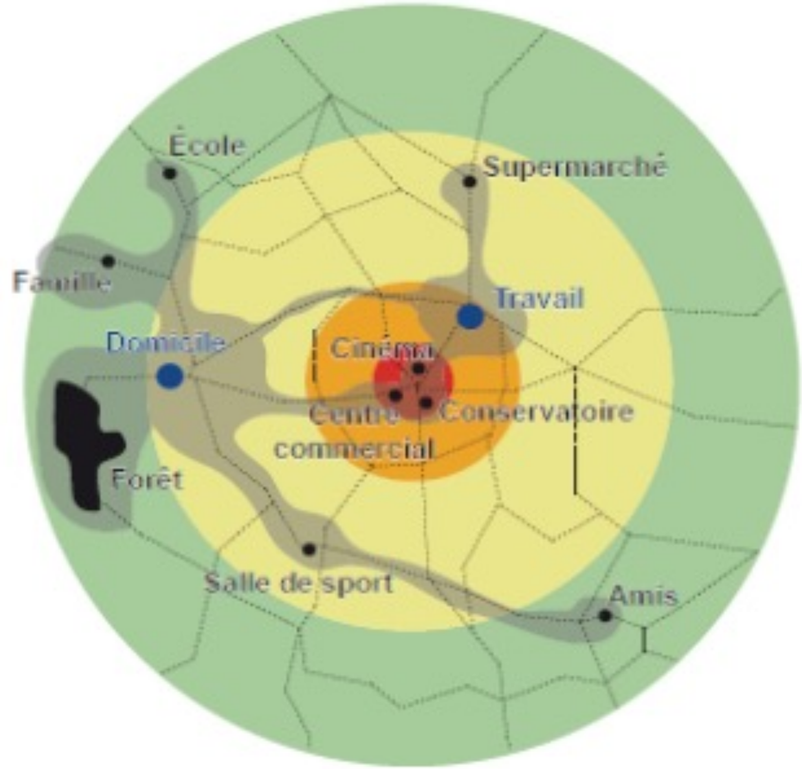
- Urban, persons/goods
- Daily , activities, trips, modes



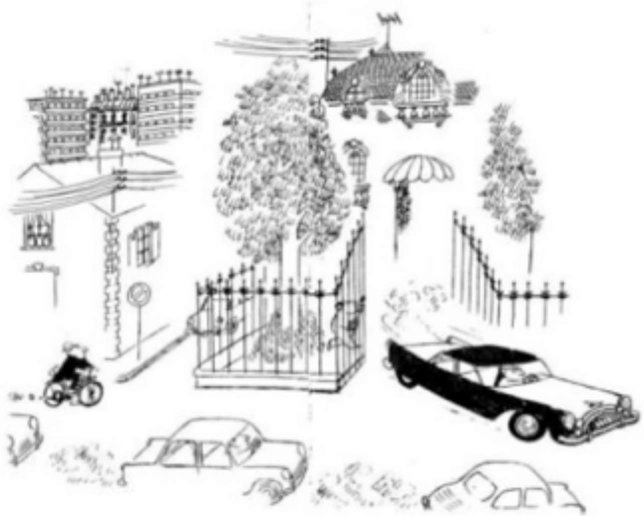
Territorial anchoring

- Travels as an expression of spatially anchored lifestyles (S. Carpentier)
- Coupling Home/transport

Les mobilités quotidiennes: représentations et pratiques. Vers l'identité de déplacement (2007)



Socio-economical anchoring



Sempé

Social anchoring

Trajectories and traffic flow theories

- Eulerian representation of the flow by function:
  - Fluid = speed V(x,t)
  - Counting vehicles and users at sites

- Lagrangian representation of the flow by individual particles
  - Particle = vehicle position (x,y,z,t) continuous/ discontinuous (sampling)
  - Tracking of vehicles/users on the network

Urban mobility patterns Universal laws

Schneider CM, Belik V, Couronne T, Smoreda Z, Gonzalez MC. 2013 Unravelling daily human mobility motifs. J R Soc Interface 10: 20130246. http://dx.doi.org/10.1098/rsif.2013.0246

- Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A Tale of Many Cities: Universal Patterns in Human Urban Mobility. PLoS ONE 7(5): e37027. doi:10.1371/journal.pone.0037027

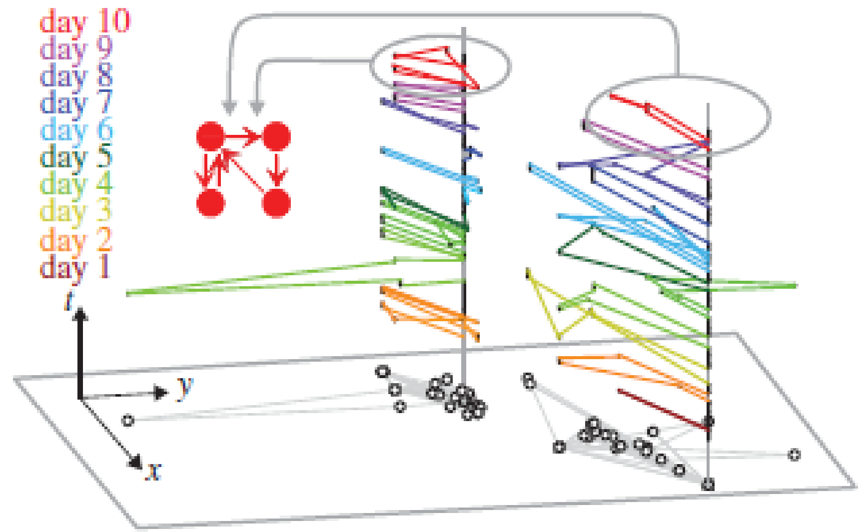
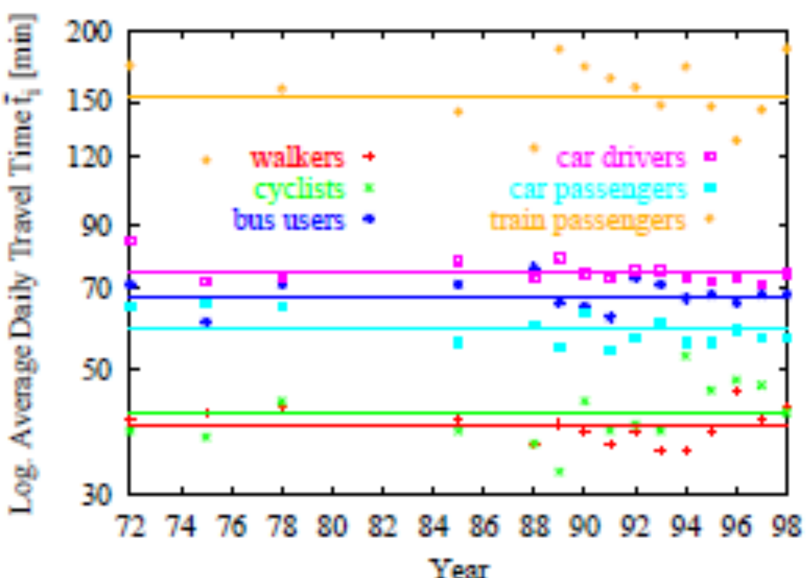


Figure 1. Decomposition of the mobility profile over 10 days into daily mobility patterns for two anonymous mobile phone users. The home location of each user is highlighted and connected over the entire observation period with a grey line. While the entire mobility profiles (black circles and grey lines in the *xy*-plane) are rather diverse, the individual daily profiles (brown to red from bottom to top for different days) share common features. The aggregated networks consist of $N = 16$ (22) nodes and $M = 37$ (43) edges with an average degree of $\langle k \rangle = 2M/N = 4.6$ (3.9). By contrast, the daily average number of nodes is $\langle N \rangle = 4.4 \pm 1.8$ (3.9 ± 1.3), and the average number of edges is $\langle M \rangle = 5.3 \pm 2.8$ (4.2 ± 2.2). The left user prefers commuting to one place and visits the other locations during a single tour, whereas the right user prefers to visit the daily locations during a single tour. On the last day, both users visit not only four locations, but also share the same daily profile consisting of two tours with one and two destinations, respectively.

- Number of places visited
- Time spent (Travel Time budget constant)
- Zahavi, Y., The TT-relationship: A Unified Approach to Transportation Planning. Traffic Engineering and Control, pp. 205-212, 1973.
- Kölbl, R. & Helbing, D., Energy laws in human travel behaviour. New Journal of Physics, 5, pp 48.1–48.12, 2003.



Unimodal

Activity	Speed (km/h)	Energy Consumption (kJ/min)
Sitting on a chair		1.5
Standing, relaxed		2.6
Standing, restless		6.7
Walking on even path	4	14.1
	5	18.0
Cycling on even path	12	14.7
Car, roads		4.2
Car, test drive		8.0 (5.9–12.6)
Car, in city, rush hour		13.4

Quantified traveller

Jariyasunant, J., Abou-Zeid, M., Carrel, A., Ekambaram, V., Gaker, D., Sen- gupta, R., and Walker, J. L. (2013). Quantified traveler: Travel feedback meets the cloud to change behavior. *Journal of Intelligent Transportation Systems*, published online 31/10/13. DOI:10.1080/15472450.2013.856714

- Distance per trip

Digital Revolution and Mobility

- Intelligent transport systems and smart mobility
- Digital and smart citizens and consumers, User centric Apps on smartphones (GPS+accelerometer)
  - Quantified self mobility
  - Crowdsensing mobility (provider)
  - Platforms : carsharing, … (co-producer)
- Digital and mobility actors :
  - equipment of transportation places and vehicles, in smart cities (Site centric)
    - Stations (ticketing) , connected vehicles, cars, ….
  - Better knowledge of behaviors than individuals
  - Better planification of mobility (less expensive, more energy efficient,reliable, shorter than « go faster »)
  - Multimodality, regulation

Quantified traveller

- Moves = activity diary

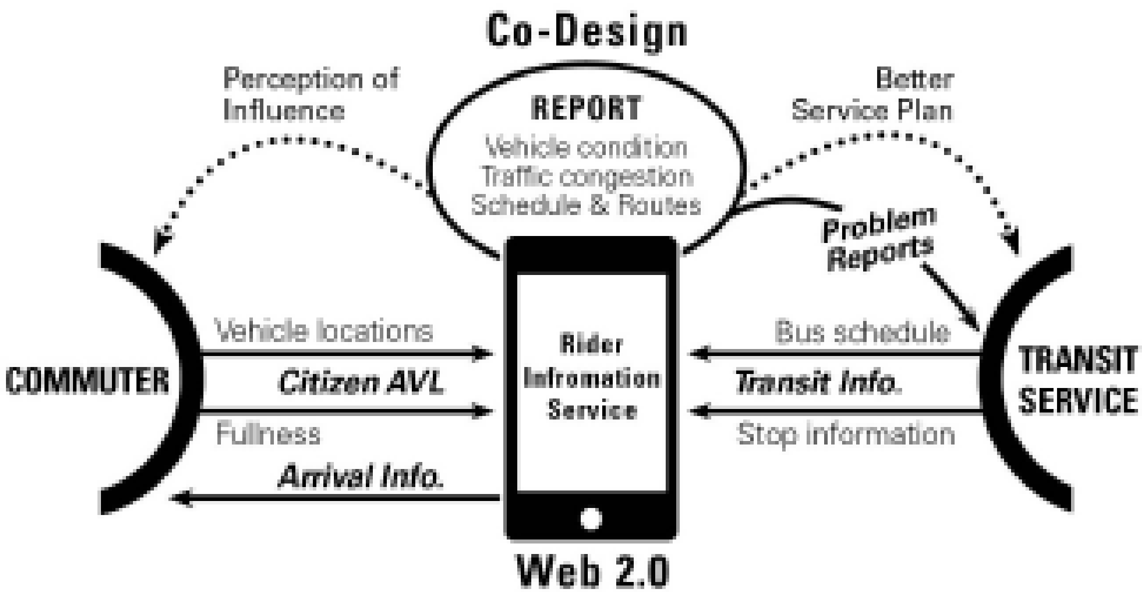
From individual to collective mobility Conditions for change

- Homo economicus/homo socialis
- Changing the frame, the representation
- Measuring collective value created
- From quantified self to quantified commons
- Small worlds or communties
- Finding the good incentives

- Alain Rallet (Université paris Sud), Jean Marc Josset (Orange labs)

Mobile Crowdsensing and transportation

- Community (Tranquilien)



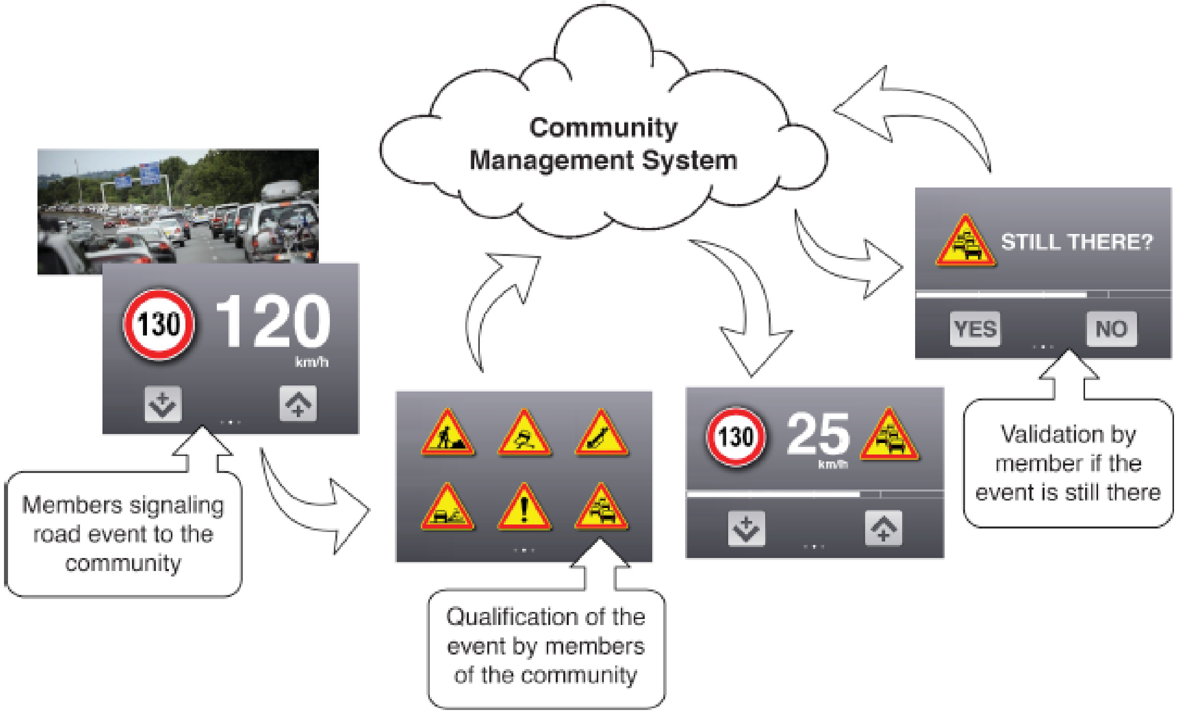
- Privacy protection and geo-localisation

- Waze

- Motivations for participation (sharing)
- Critical mass
- (semi)-trust

- **Asking**: People are more likely to contribute if they are asked, and if they are asked specifically/individually.
- **Intrinsic Motivation**: People will contribute if they perceive an intrinsic motivation, such as their own enjoyment in doing the work. In addition, people perceive value in helping others and in helping groups of people they feel an affiliation towards.
- **Rewards**: People will contribute for different kinds of rewards including praise, increased reputation, an increase in privileges, and financial compensation.

- Speed cameras Alert and more
- Coyotte and co (driving asssitant) (Pauzié)



Tweets on transportation

- Expressive data on the web
- Signals without context except time and geolocalisation; mimetism and contagion
  - Microblogging , text (ungrammatical). Content about real world events
    - Incidents (Normal, degraded, perturbed situations) in transportation system
    - Traveller's opinions
    - Information on journey needs
- Mining of tweets (Topic detection and tracking) (Gal-Tzur)
- Opinion mining and sentiment analysis

Problems

- Monotonous and repetitive quantified self
- Communication and energy consumption (battery)
- Trivial generality or oriented opinion with tweets (+ biaised)

- Who is the (co-) owner of the data footprints?

- Privacy : both desires :exposed and protected
- Illusion of trade-off between security/privacy and service effectiveness
  - Rather asymetry of information and absence of alternative
  - No possibility ex ante to control, rather ex post control of algorithms

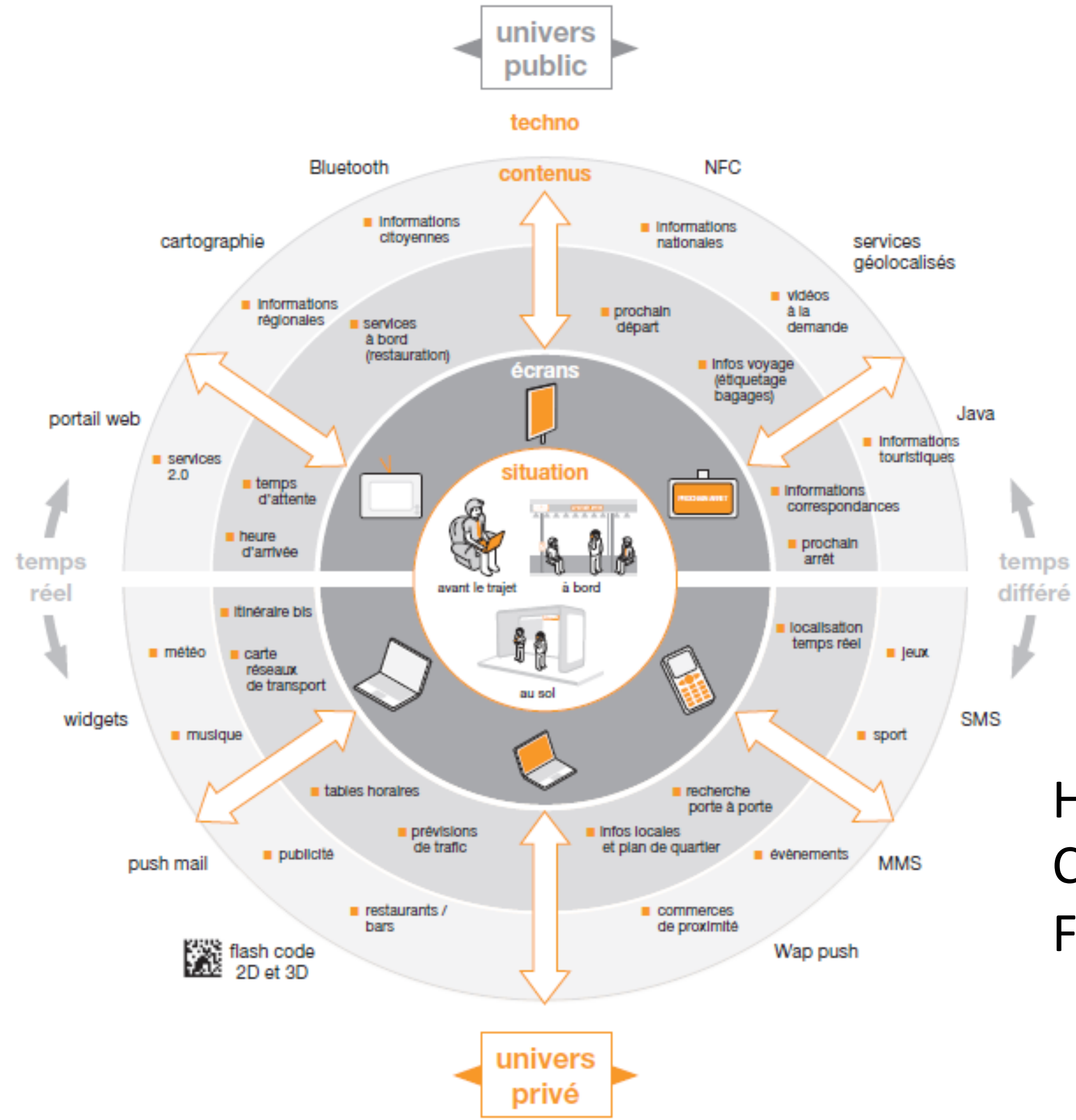
Actors of the urban transportation (eco)systems

- State and government (transportation laws)
- Local authorities , Network authorities, Transit authorities (regulator, operator), Mobility authorities
- Public and private transport operators
  - Bus, train, metro, tram + stations
  - Taxi, VTC, shuttle (van, car, two-wheeler, three-wheeler)
- Car rental companies, autoshare bicycleshare companies (services)
- Carsharing platforms
- Telephone operators, Google and co., … (Multimodal Information system)
- Households and individuals (consumer, user, citizen)
- Social networks
- Mobility generators (companies, schools, hypermarkets, festivals, …)

Information and transportation

- BtoC oriented
- Real time
- Multi sensors
- Multimedia

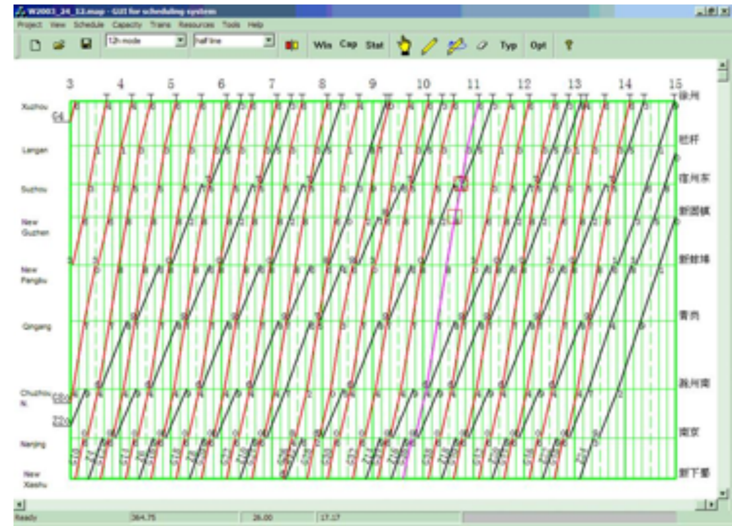
- Ticketing
- Automatic counting
  - Sensors and cameras
- Tracking
  - GPS
  - Mobile phone



(Source : Orange Business Services, octobre 2009)

Hyperconnected Consumer
From Orange Labs

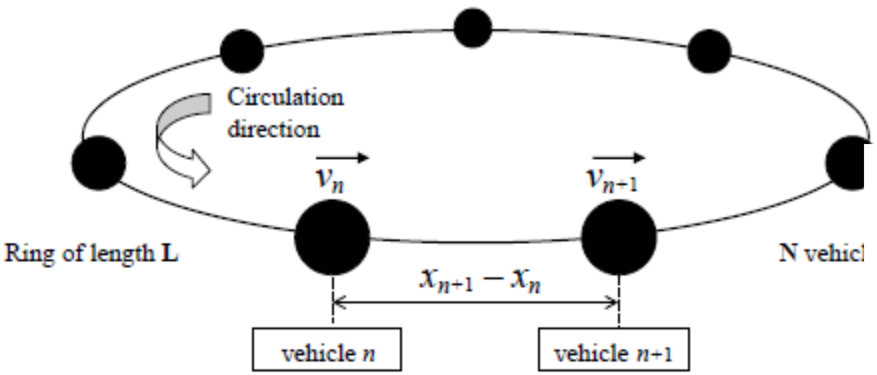
Regulation and optimisation and safety

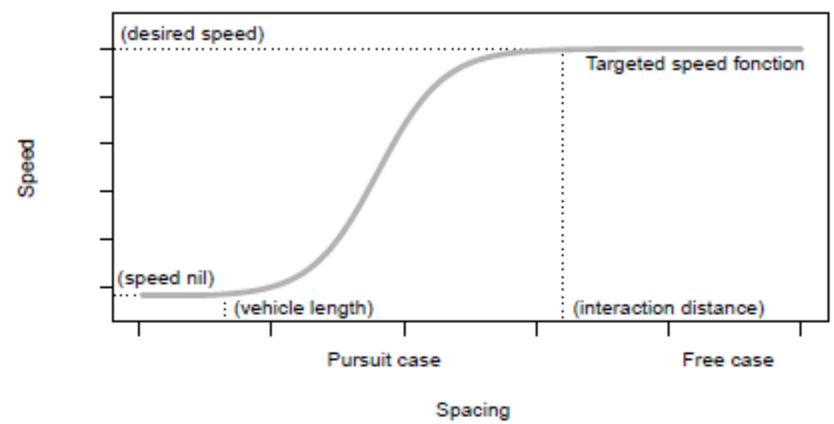


- Buses
  - Headways and bus bunching
  - Trade-off: Reliability and travel time
- Trains and metros
  - (Re)Scheduling
- Lorries and cars
  - Autonomous vehicle with sensors : lidar, radar, cameras, …
  - Naturalistic driving or drowning by numbers
    - Hundreds of signals of all nature
    - From incidents to accidents (triggering)

Stability of dynamic systems

$$\dot{x}_n(t + T^r) = \mathcal{R}\left\{x_{n+1}(t) - x_n(t), \dot{x}_{n+1}(t)\right\}$$





$$\dot{x}_n(t + T^r) = \mathcal{V}\{x_{n+1}(t) - x_n(t)\}$$

$$\ddot{x}_n(t) = \frac{1}{T^r}\left(\mathcal{V}\left\{x_{n+1}(t) - x_n(t)\right\} - \dot{x}_n(t)\right)$$

Linear stability analysis of first-order delayed car-following models on a ring
Antoine Tordeux, Michel Roussignol, and Sylvain Lassarre
Phys. Rev. E 86, 036207 – Published 12 September 2012

Problems about automation

- Algorithms for solving driving tasks ? In everyday situations
  - Telsa fatal accident

- Security (malveillance, attack)
  - Protection of communication (cryptage)
  - Control at distance by hackers

Another revolution

DEFINING THE DATA REVOLUTION

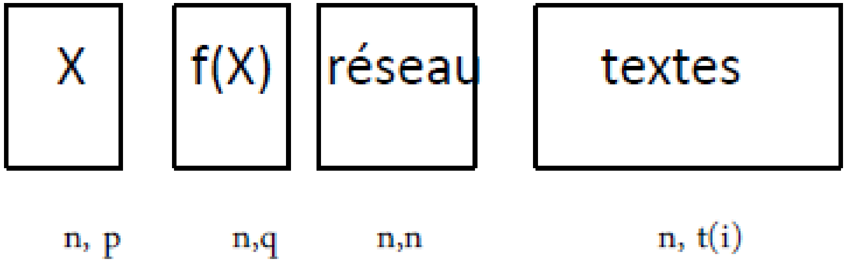
'The data revolution is: an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the 'Internet of Things,' and from other sources, such as qualitative data, citizen-generated data and perceptions data; A growing demand for data from all parts of society.'

UN Secretary-General's Independent Expert Advisory Group on a Data Revolution (A World That Counts report, page 6)

- Big Data appears for the first time 1997:
  - –Cox & Ellsworth (NASA) «Managing Big Data for Visualisation» *ACM SIGGRAPH '97*
- Data Science is much older
  - –P. Naur 1960
  - –IFCS (Kobe, 1996)"Data Science, classification, and related methods“
  - –Journal of Data Science since 2003

- Origin:
  Data from web, social networks
  Connected objects

- Volume
- Velocity (peak)

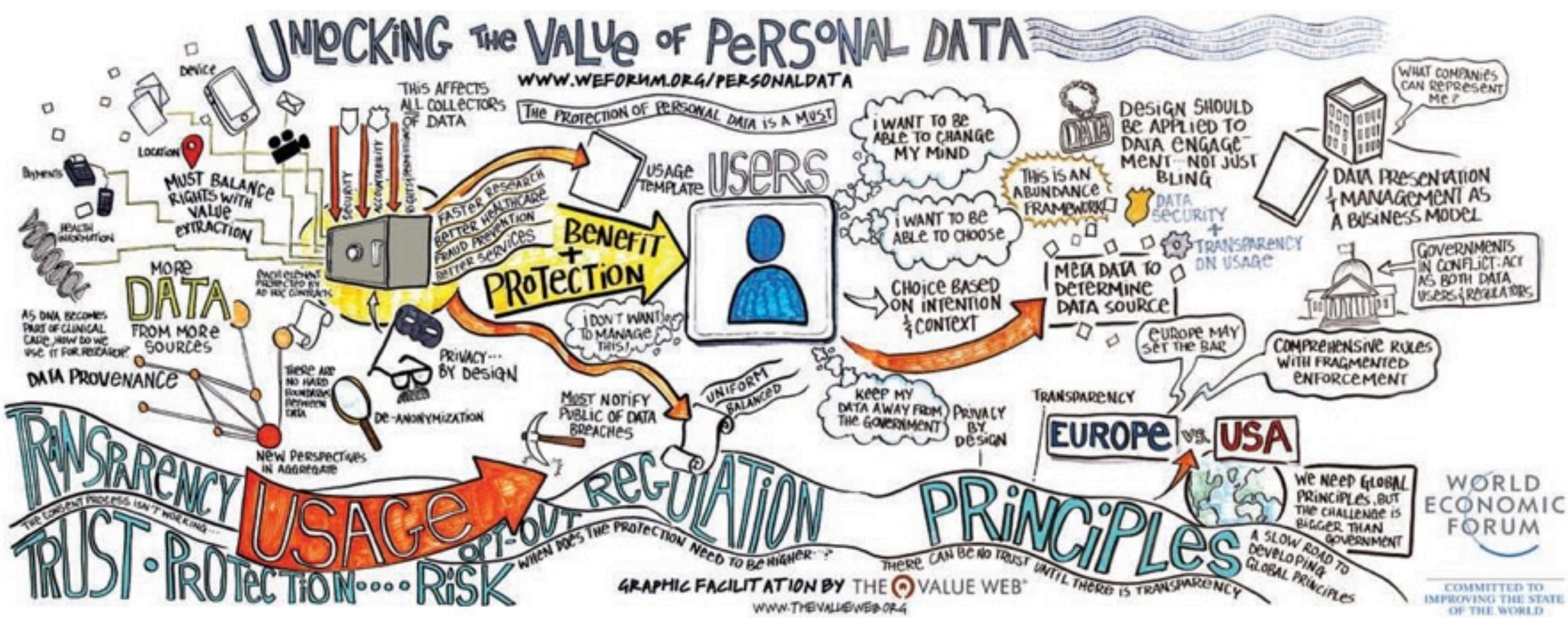


- Variety: numerical, categorical data , graphs (social networks), texts, videos, etc.

- Not structured, without context, very noisy

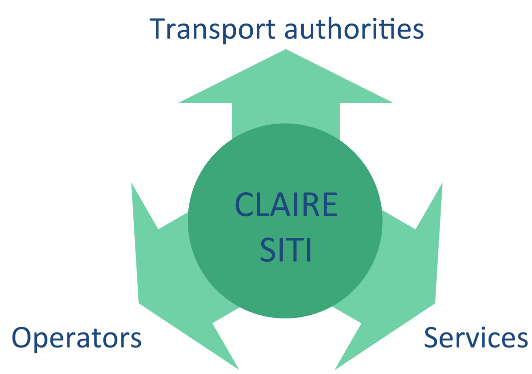
Big Data

- Supply : network, timetables (open data)
- Demand : storyboard, GPS, traces , footprints
  - vehicle (car, bus, …),
  - individual : smartphone, phone, ticketing, tweet



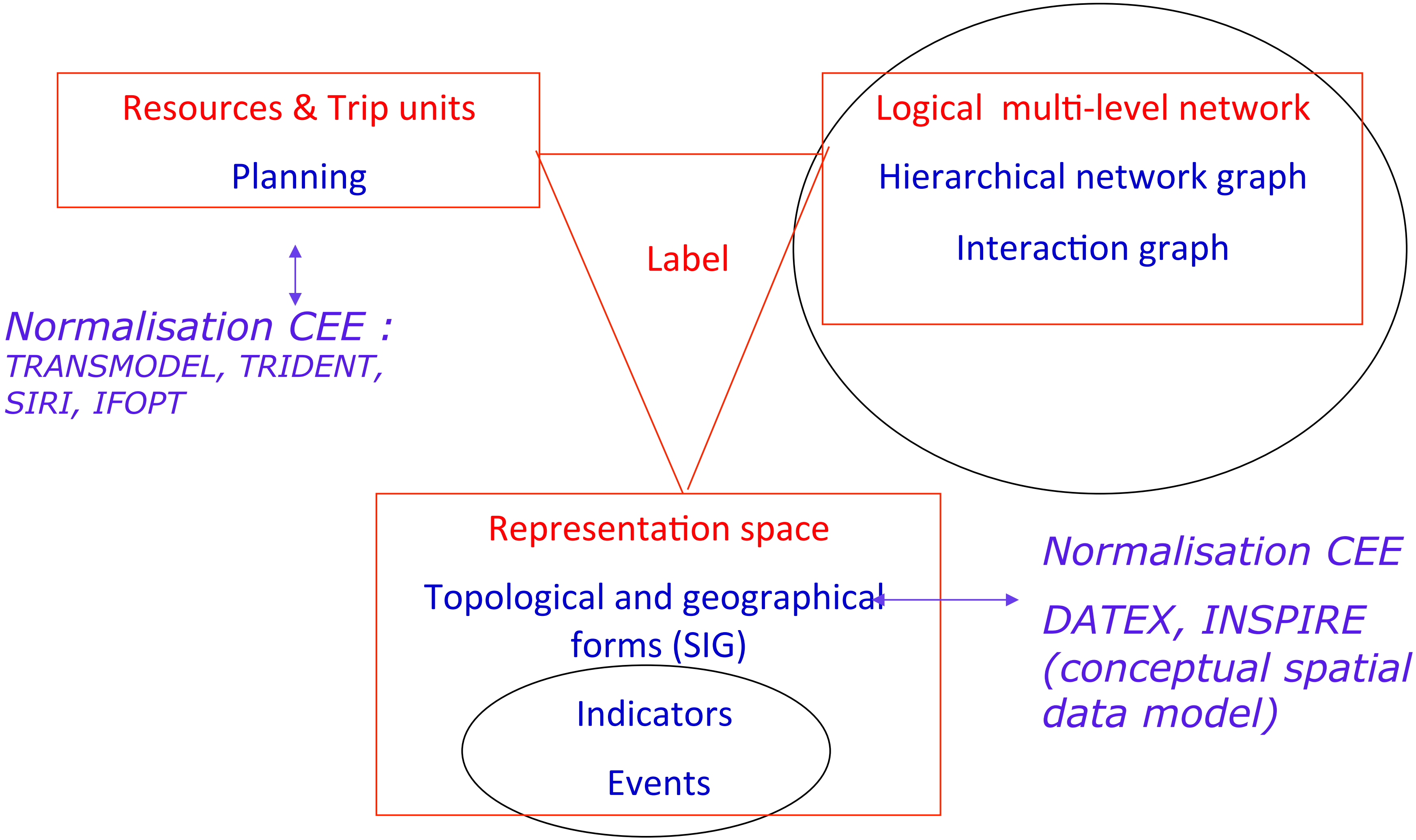
CLAIRE-SITI : A reference system for intermodality

- A GENERIC MULTIMODAL DATA MODEL
  - Any type of network (road, public transport, alternative modes)
  - Any type of indicator (congestion, time adherence, regularity, availability, reliability, sustainability)
  - Any type of event



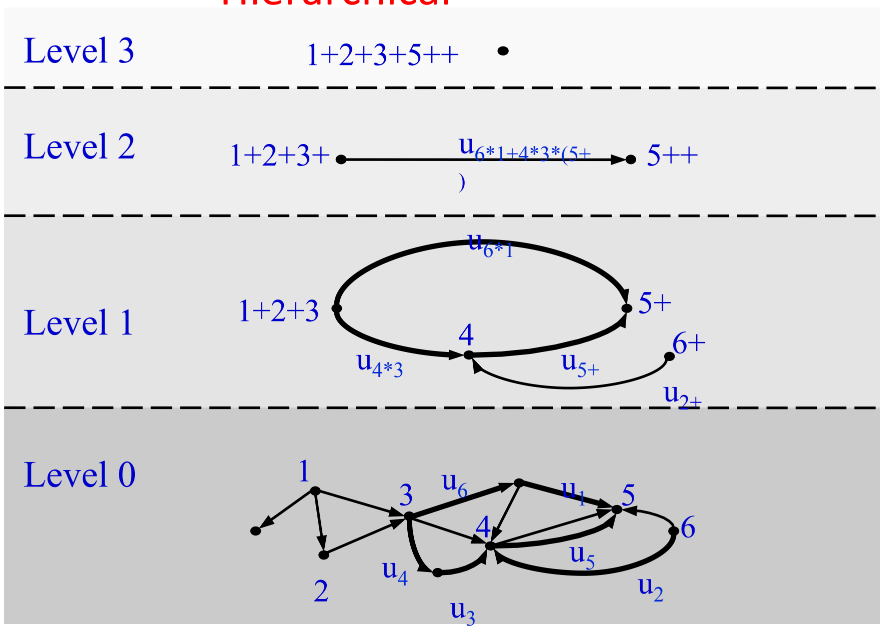
- AN ANALYSIS ENGINE WITH FUNCTIONS
  - observatory,
  - monitoring,
  - diagnosis,
  - decision/operation action
- A TOOL THAT
  - Support the development of public policies for a sustainable mobility
  - Can be integrated in service and industrial chains
  - Enhance research on Intermodality

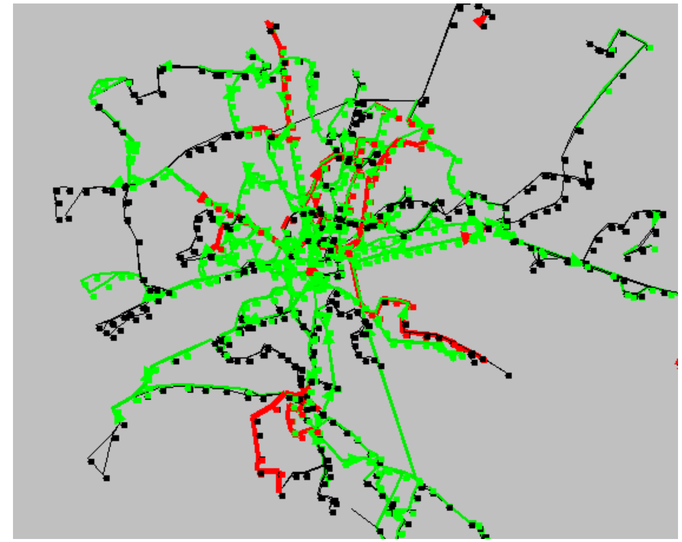
Generic Model



Structure : hierarchised multi-level & interaction graph

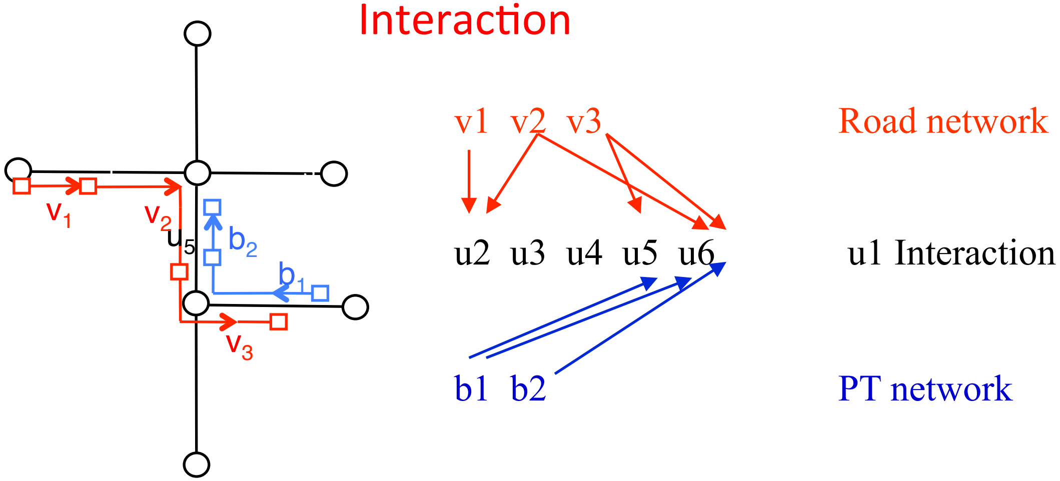
Hierarchical

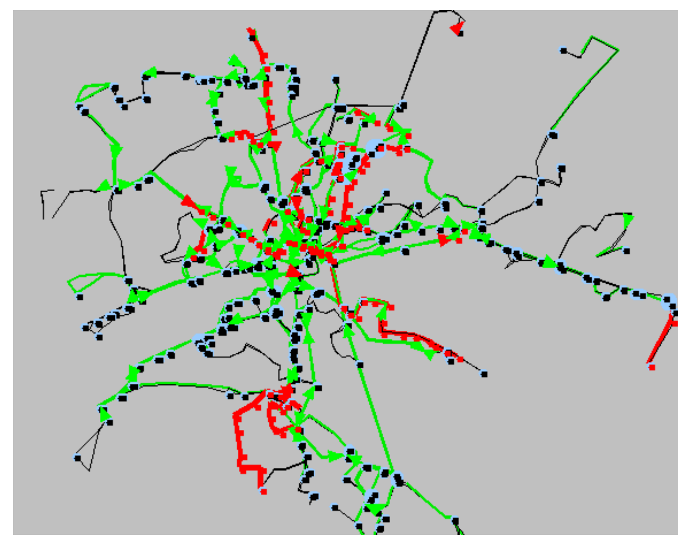




Detailed network : stops

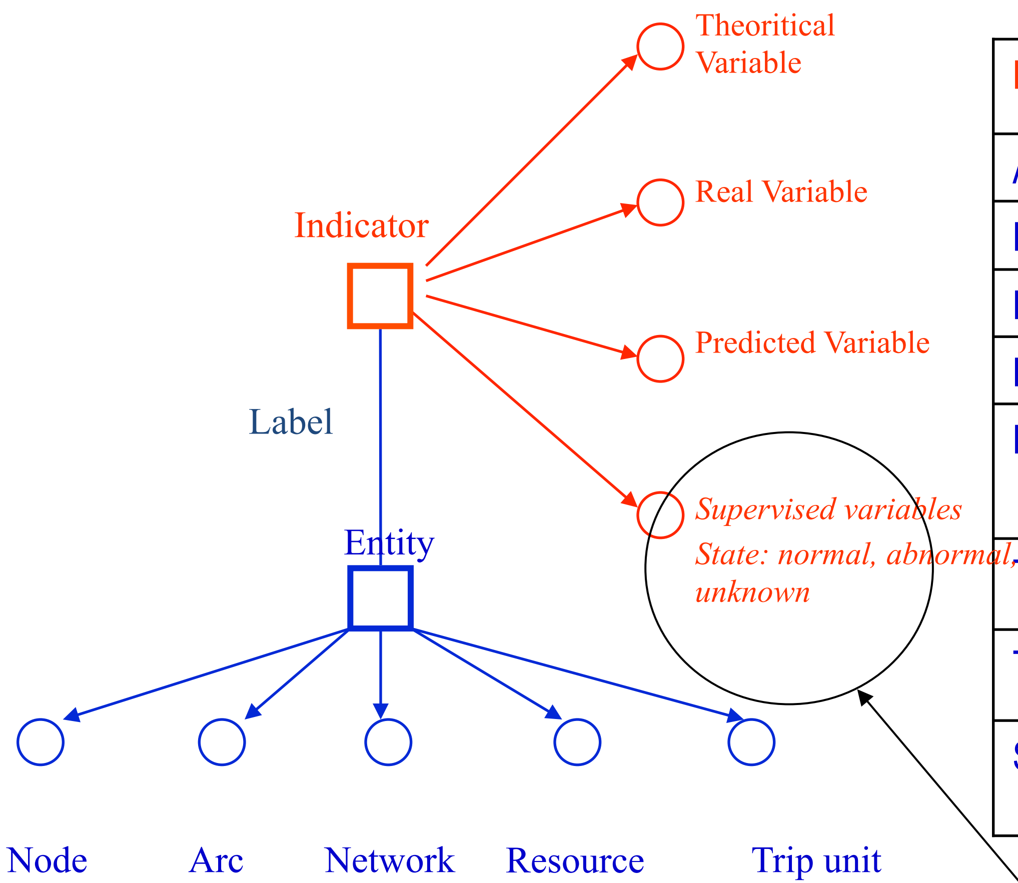
Interaction





Transfer network

Multi-criteria : Indicators & supervised variables

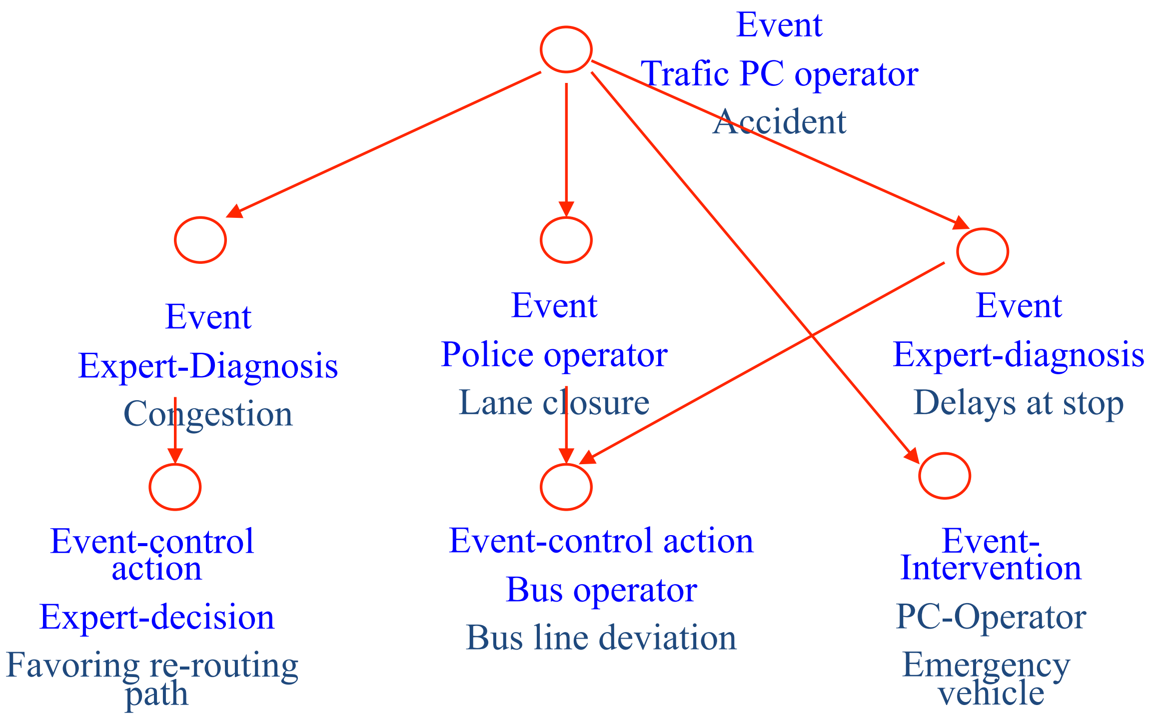


LOGIC	Variables
Adherence	Arrival time, Delay
Regularity	Waiting time, Frequency
Reliability	Commercial speed
Demand	Load
Ressources	Driver break and relief, vehicle speed,
Transfert	Transfer time
Traffic	Flow, Occupancy, Speed
Sustainability	Carbon Monoxyde, Hydorcarbon Pollutants

Level of Service (LoS)

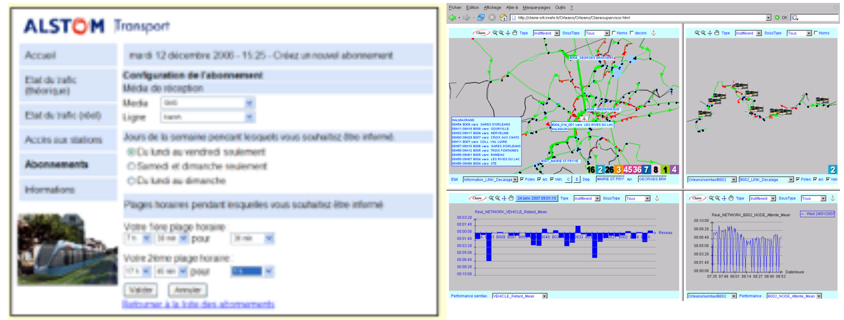
Event modeling

Event(type, sub_type, author, Causes, Effects, start-time, end-time, From, To, ….)

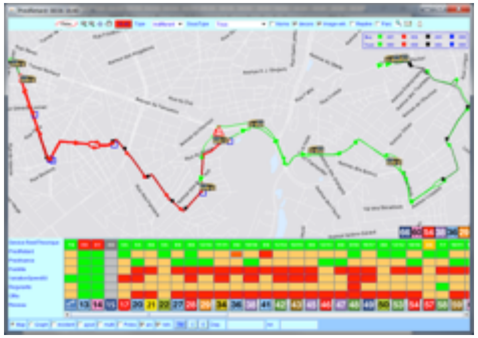


BATERI : Certification des Données des SI dans le transport

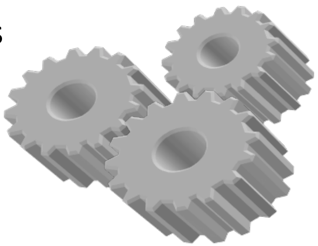
P@ss-ITS : serveur d'information multimodale en conditions perturbées



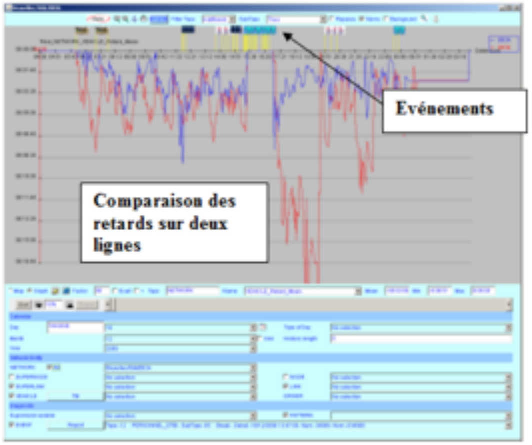
CLAIRE-SITI : supervision et décision multicritère



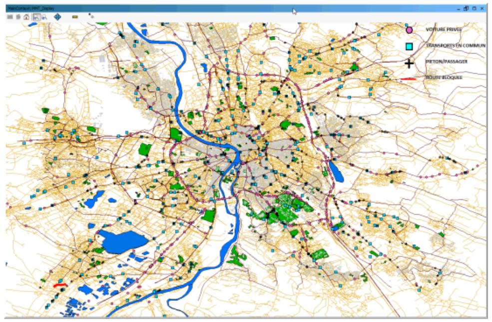
SINERGIT : nouveaux services d'information fondées sur les mobiles et les GPS



CLAIRE-SITI : Observatoire pour le suivi de la qualité de service

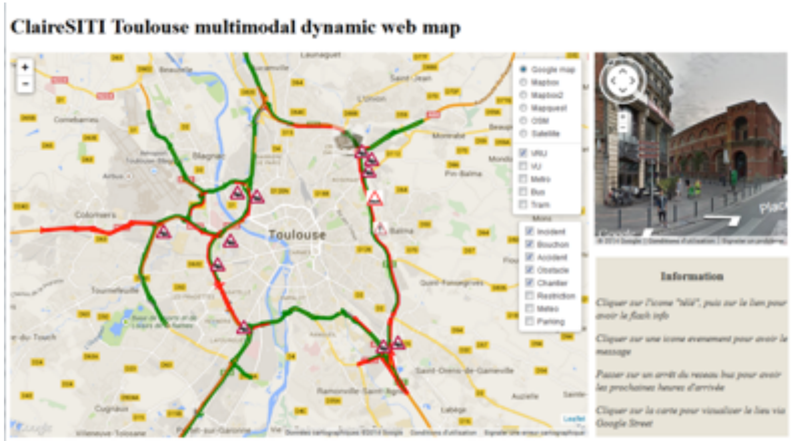


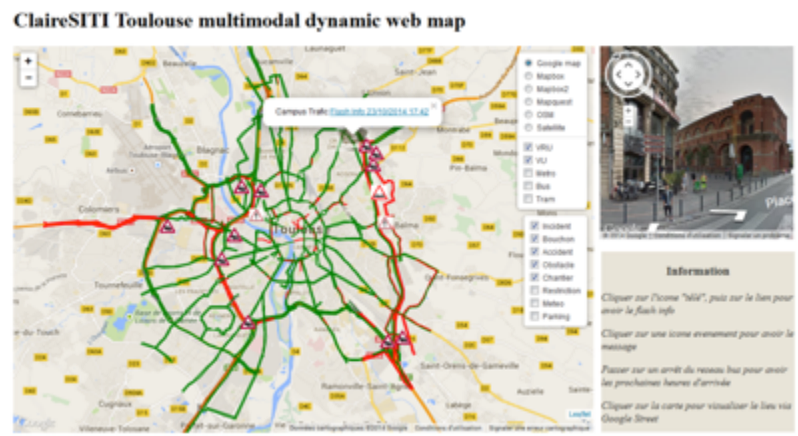
Instant mobility : Multimodal Multi-agents simulation SM4T)

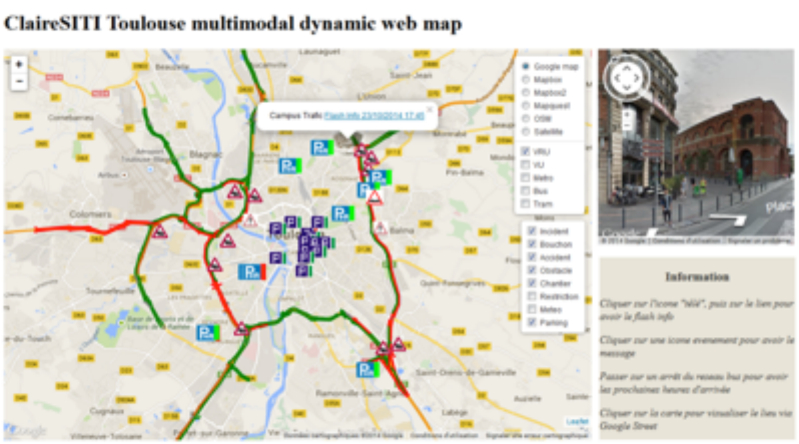


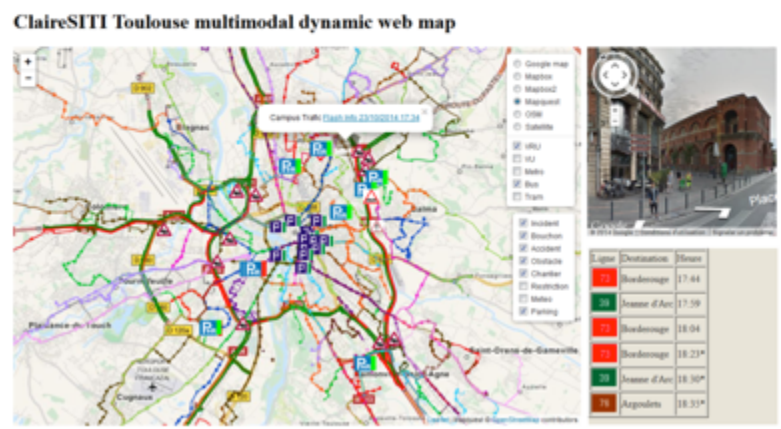
NAVITRANSPORTS : outil mobile de navigation dans le TC

Multimodal Dynamic web map









Four families of digital information and computation

Position toward the web	Aside	Above	Into	Under
Data	Views	Links (documents)	Likes	Footprints
Population	Representativ e sample	Communities, vote	Social network	Individual behaviors
Computation	Vote by clicks User centric Site centric	Meritocratic ranking	Benchmark	Machine learning
Principle for algorithm	**Popularity**	**Authority** (in the web network) Counterstrategy	**Reputation** (Knowhow)	**Prediction** **Big data**

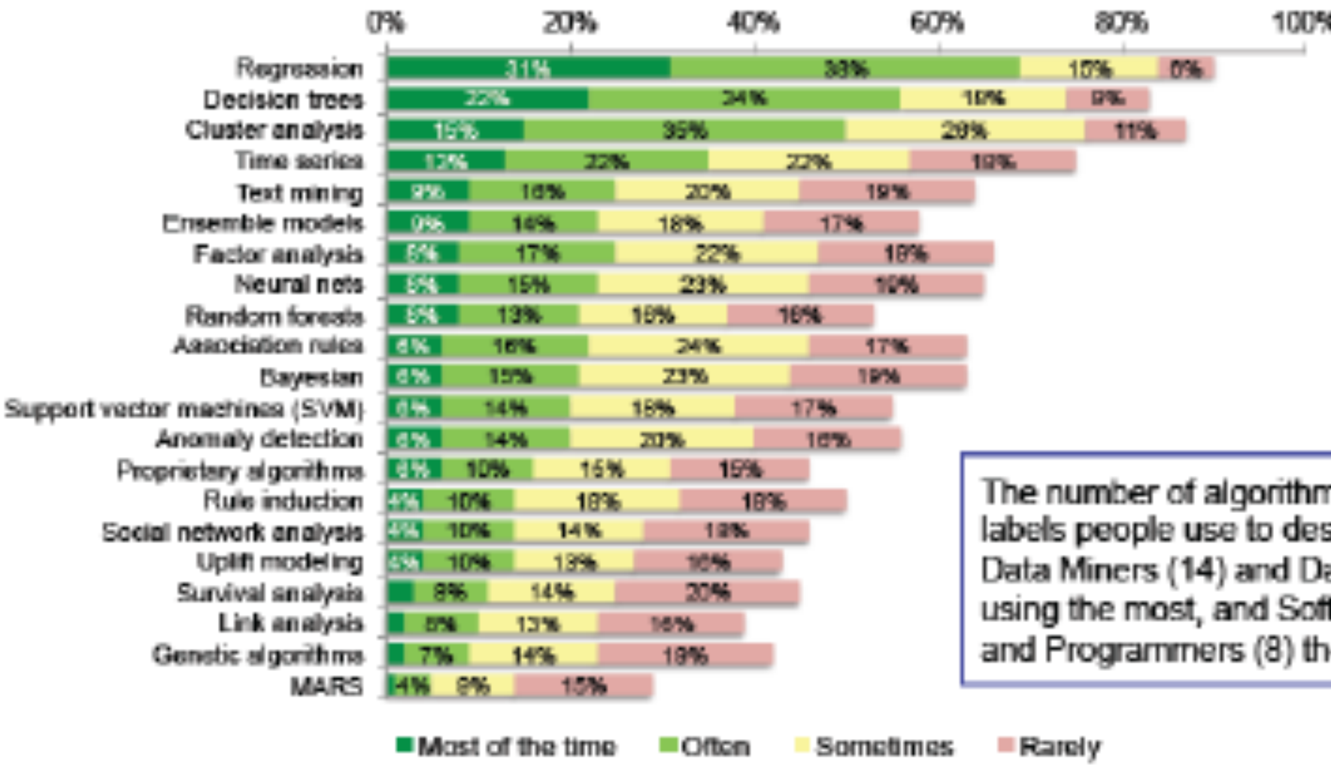
Dominique Cardon (2016) A quoi rêvent les algorithmes. Seuil.

Big Data Analytics

- Exploratory or unsupervised
  - Factorial analysis, k-means
  - Association rules
- Predictive or supervised
  - Regression models, with regularisation, trees ..
  - Black box models (neuronal network, Support Vector Machine, ..)

Algorithms

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.
- The average respondent reports typically using 12 algorithms. People with more years of experience use more algorithms, and consultants use more algorithms (13) than people working in other settings (11).



The number of algorithms used varies by the labels people use to describe themselves, with Data Miners (14) and Data Scientists (14) using the most, and Software Developers (9) and Programmers (8) the fewest.

Question: What algorithms / analytic methods do you TYPICALLY use? (Select all that apply)

A new vision of «models»

- Classical vision : models to understand
  - Provide some understanding of the data and the mechanism that generated them through a sparse representation of a random phenomenon. Usually requires the help of a statistician and a domain expert. **Generative model**
  - a model must be simple, and its parameters interpretable relative to the domain of application: elasticity, odds ratio, etc.
  - Find general patterns linked to important explanatory variables (social capital)
  - Econometric models

Prédire n'est pas expliquer
René Thom ES**H**EL (1991)

- Vision «Big Data Analytics»: **predictive model**
  - look for regularities (Habitus) with few hypothesis
  - predictive capacity on new observations :«generalisation »
  - different from goodness of fit to data (predict the past)
- A very accurate model of the data behaves unsteadily on new data: the phenomenon of overtraining or overfitting
- A very robust model (rigid) does not give a good fit to the data
  - models from data («data driven»); In Data Mining and Machine learning a model is nothing more than **an algorithm**
  - set of contingent micro-theories for probable behavior
  - support conformism (dividu Deleuze no history no representation)

The model is no more an input for the computation, but an output.

Extraction of passenger travel patterns : **passengers with similar transport habits**

- **Observed variables**

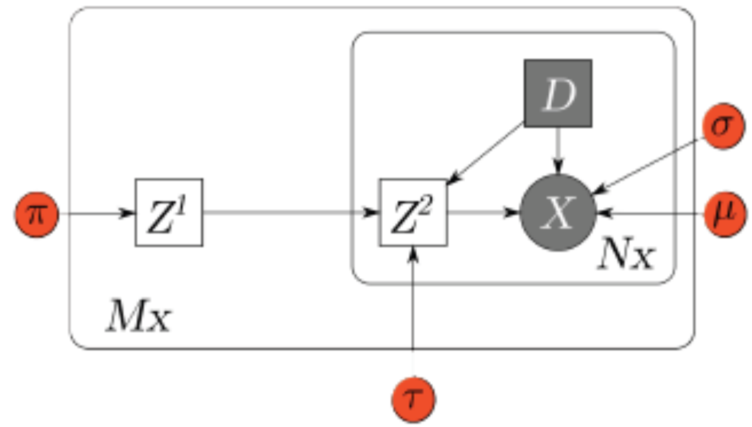
  D : day of the week the trip was made

  X : trips time generated using a normal distribution

- **Latent variables**

  Z_1 : Passenger membership to one of the K clusters

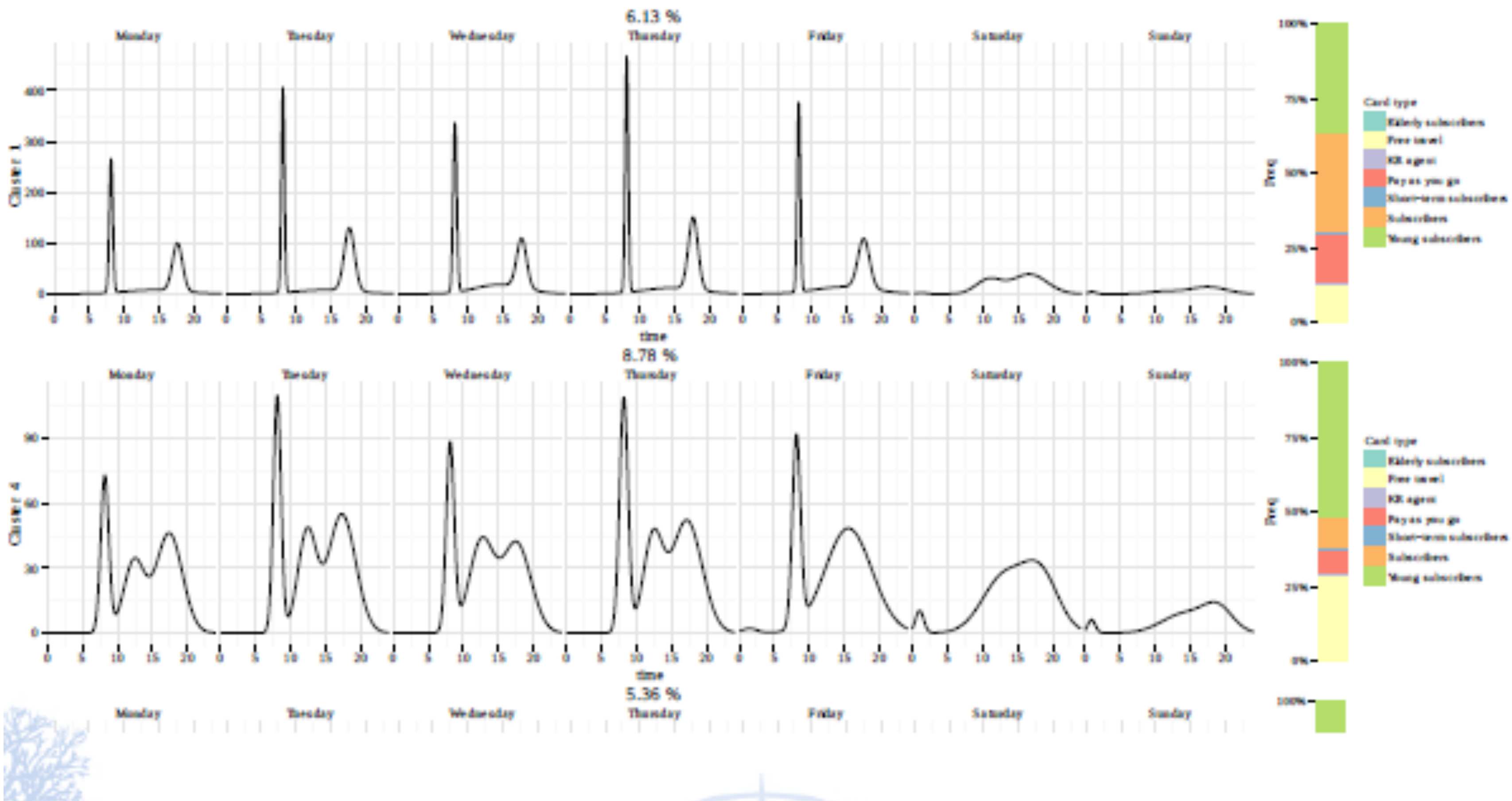
  Z_2 : Trip membership to one of the gaussians describing the temporal activity of the cluster (distribution of the trip hours made by the passengers belonging to a given cluster is modeled by a mixture of gaussians)



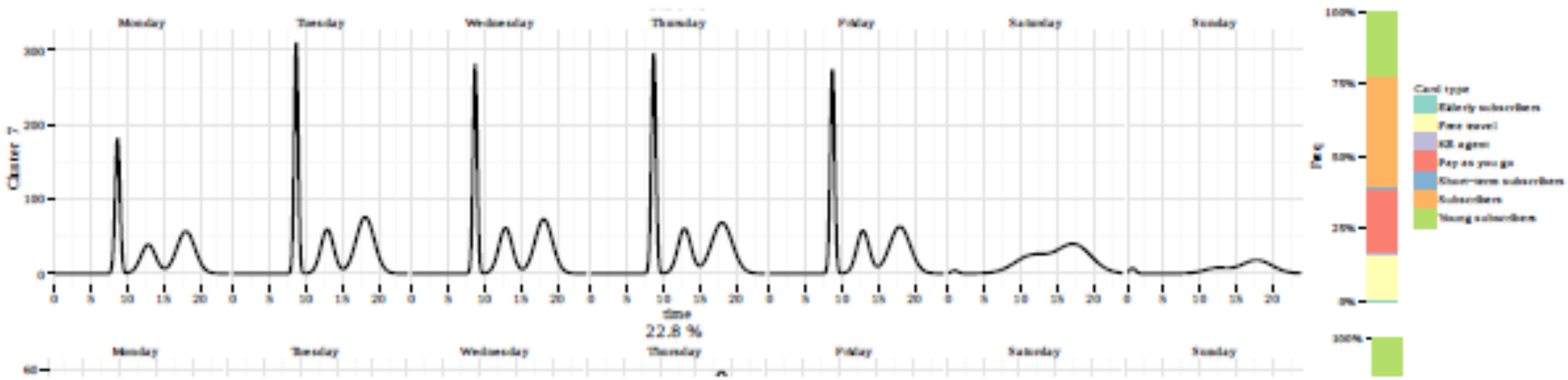
Source : Ticketing
Card number, day, time (hour)

[A-S Briand et al 2015] A-S. Briand, E. Côme, M-K. El Mahrsi, L. Oukhellou, A Mixture Model Clustering Approach for Temporal Passenger Pattern Characterization in Public Transport, IEEE DSAA (Data Science and Advanced Analytics, Paris 2015) , extension dans JDSA (Journal of Data Science and Analytics)

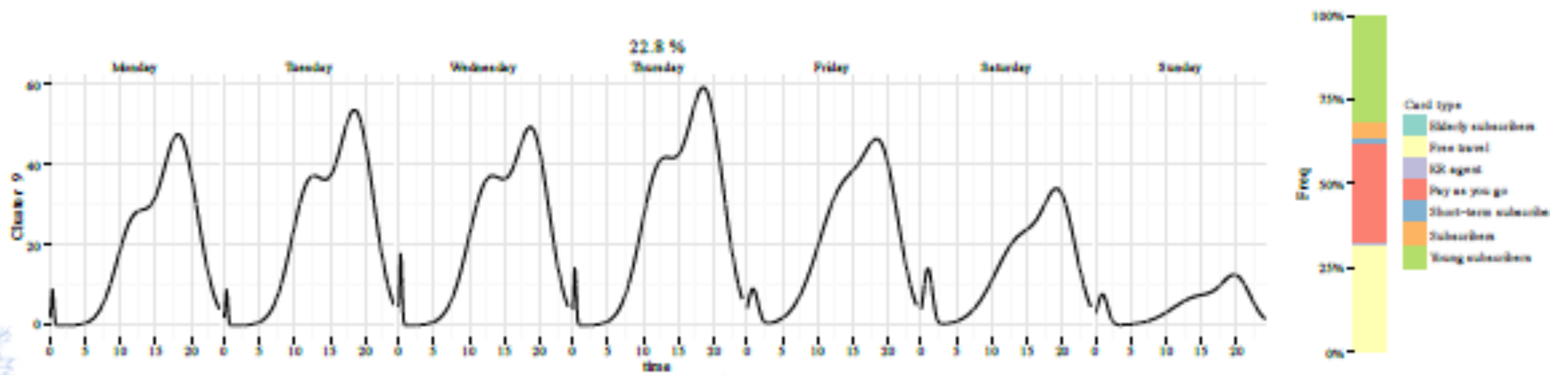
Mobility patterns



Probability density to be in the public transportation system

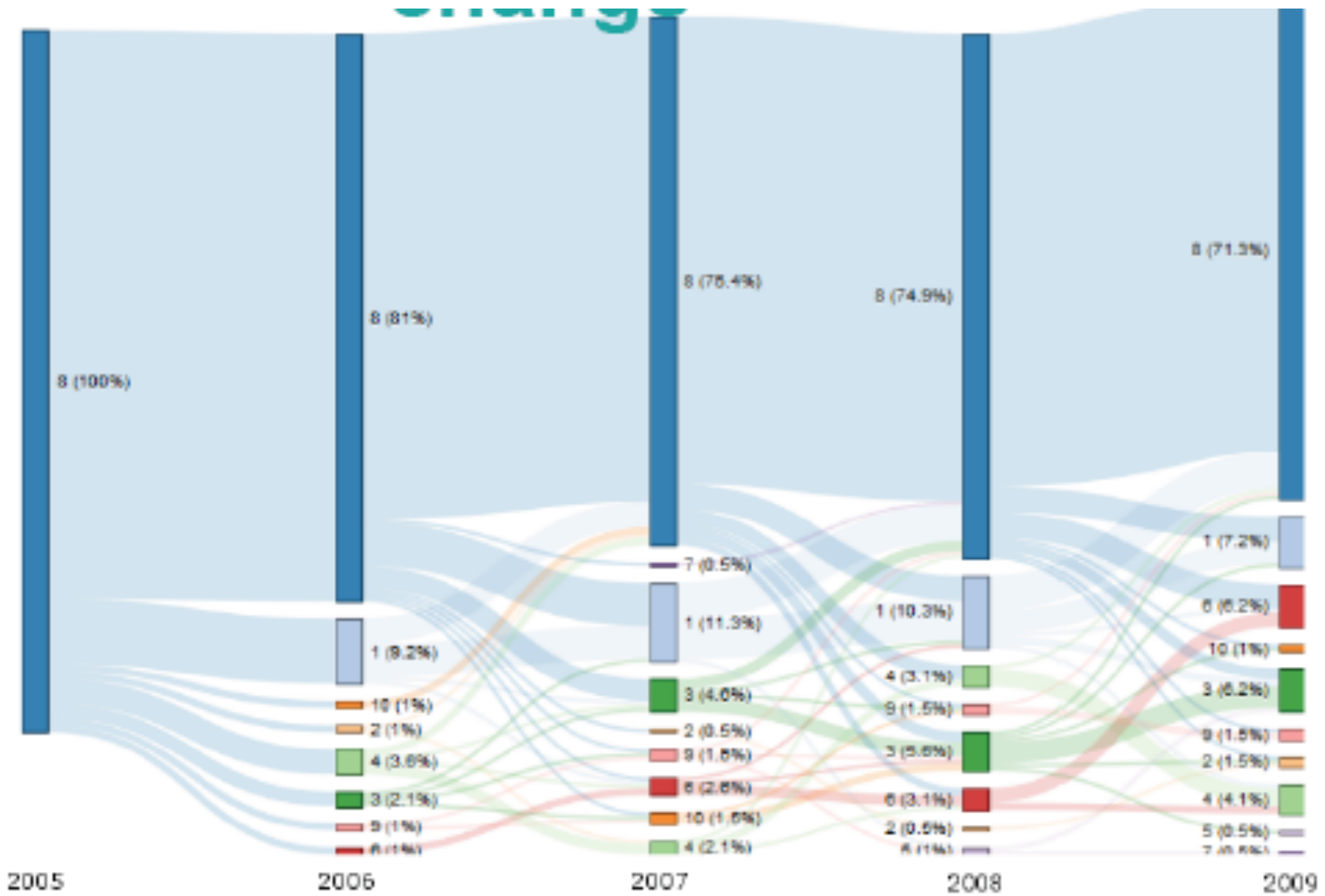


Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Cluster 7
time
Freq
Card type
Elderly subscribers
Free travel
KR agent
Pay as you go
Short-term subscribers
Subscribers
Young subscribers
22.8 %



22.8 %
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Cluster 9
time
Freq
Card type
Elderly subscribers
Free travel
KR agent
Pay as you go
Short-term subscribe
Subscribers
Young subscribers

Cluster change Quebec Public transport



- «New» models from Machine Learning
  - Neuronal networks and deep learning
  - SVM (Support Vector Machine)
  - Association rules and reputation systems (eg Amazon)
  - Random forests (decision trees combination)
  - Stacking and meta-models
- The «feature engineering»
  - A feature is a piece of information that might be useful for prediction. Any attribute could be a feature, as long as it is useful to the model.

Complexity and trade_off bias/variance

- Learning theory by Vapnik (VC dimension)
- Consistence if convergence between generalisation error and learning error.
- Beyond AIC (Akaike information criterion) and BIC (Bayesian information criterion)

Agregation of models

- Why choosing between models?
- Set methods : combine the predictions of different models
- Stacking
  - Linear combinaison of m prédictions obtained by differents models
  - First idea : linear regression
    - Foster the most complex models: overfitting

- Solution: use the predicted values without one unit i
- Améliorations:
- –Linear Combinaisons with positive coefficients (sum equal 1)
- –Régression PLS or other regularising method because the m predictions are very correlated

$$\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \dots, \hat{f}_m(\mathbf{x})$$

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j(\mathbf{x}) \right)^2$$

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j^{-i}(\mathbf{x}) \right)^2$$

- Advantages
  - –Better prediction than with the best model
  - –Possibility of mixing models of different natures: trees , ppv, neural networks etc.

The validation problem

- Need to matchMachine Learning and statistics
  - –A good model is one which predicts well
  - –Difference between goodness of fit and prediction
  - –Three samples to choose among models for learning, testing and validation

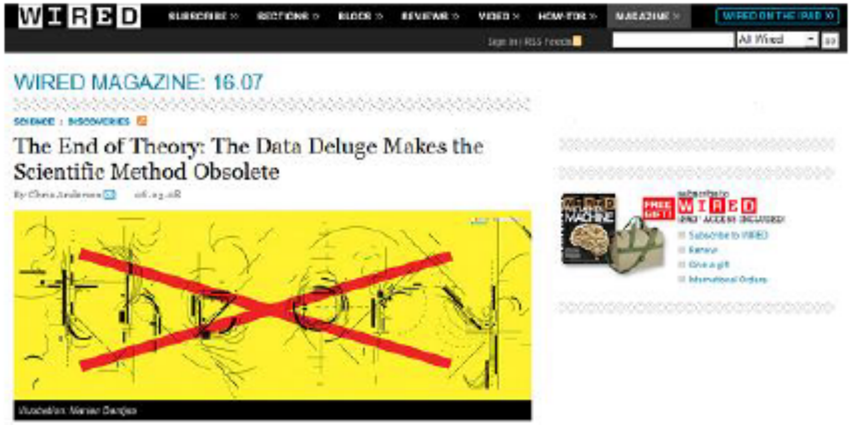
- Learning: to estimate the parameters of models
- Test : to choose the best model
  - Reestimation of the final model **: with all available data**
- Validation :to estimate the performance on future data
  - Estimate the parameters ≠ estimate the performance

But

- Correlation is not causality…
- The influence of a factor is not measured by its regression regression (P. Bühlmann)
  - «Every things equal» is difficult to sustain
  - Varying a predictor causes change in other predictors(intervention vs correlation)
  - Need for a causal diagram
- Big data require a specific appraoch
- Old methods remain effective, mainly for unsupervised methods
- Which statisticians forBig Data?

The end of science?

- Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.



Conclusion

- Mobility in an era of change
  - Decline of the conflict automobile versus Public tranport (mass transit)
  - New comers : mobility 2.0, collaborative economy, sustainability and eco-slow mobility
- Big data in tranportation
  - Already done by main actors
  - Obstacles for individual mobility data collection
  - Derived measurements through mobile phones
- Algorithms
  - From eulerian to lagrangian models for regulation in real time
  - Predictive models and The end of science? for trafic states anf their dynamics in a transportation network

Bibliography

- Saporta G.(2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings, Physica Verlag, 315-322*
- Dominique Cardon (2016) A quoi rêvent les algorithmes Nos vies à l’heure des *big data*. Seuil.