# Big data aware applications

**Genoveva Vargas Solar**
French Council of scientific research, LIG-LAFMIA, France
Genoveva.Vargas@imag.fr
http://www.vargas-solar.com/bigdata-fest

# Open data

Prof David Hakken http://www.soic.indiana.edu/people/profiles/hakken-david.shtml

# Definitions



- Description of facts
- Reproducible without ambiguities
- Parts of larger information or knowledge
- Can be expressed and stored in digital formats

single pieces
of information
of every nature

Data → Information → Knowledge → Wisdom

http://digitalcuration.blogspot.com/2009/05/what-are-data.html

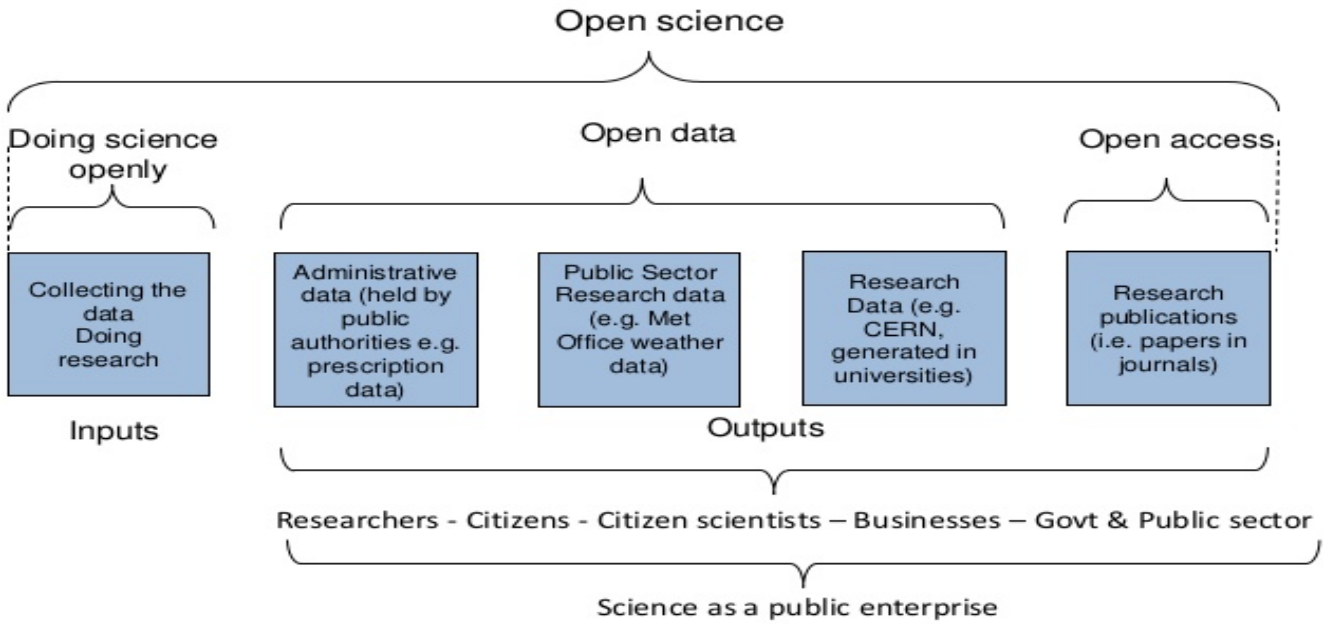
# “open”



- A piece of content or data is open if anyone is free to use, anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and share-alike.”

http://www.opendefinition.org/okd/

# Open Science

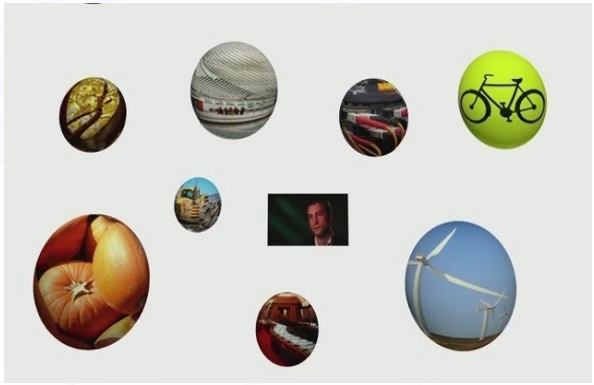



Source: A Revolution in Open Science: Open Data and the Role of Libraries (Professor Geoffrey Boulton at LIBER 2013)

# Now …


http://www.trentinoopendata.eu

## Open […] Data



http://opengovernmentdata.org/film/

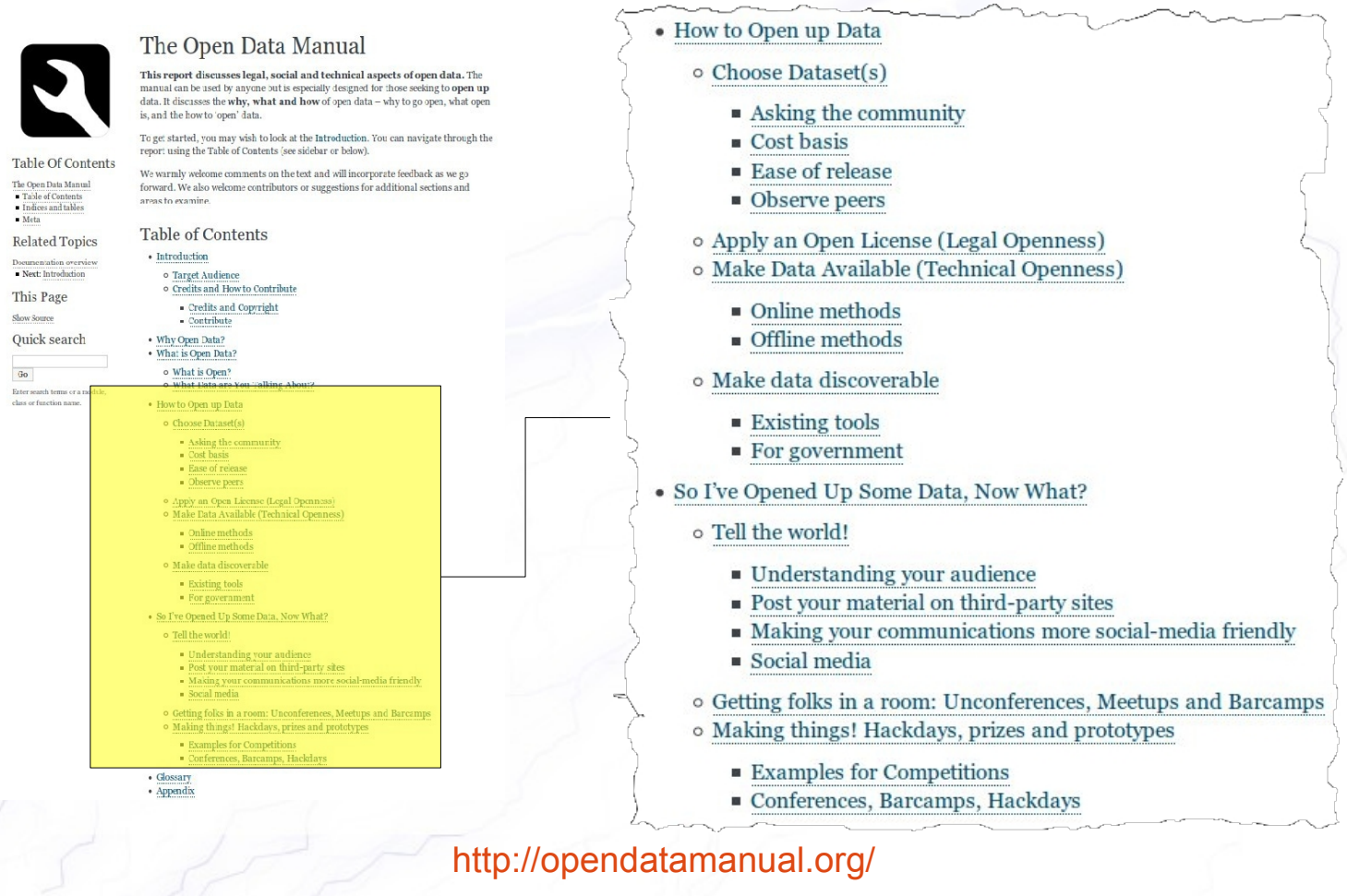
## Open Government Data

## Open Linked Data


## Open Community Data

http://www.ted.com/talks/lang/eng/tim_berners_lee_the_year_open_data_went_worldwide.html

# The correct way




http://opendatamanual.org/

# Linked open data

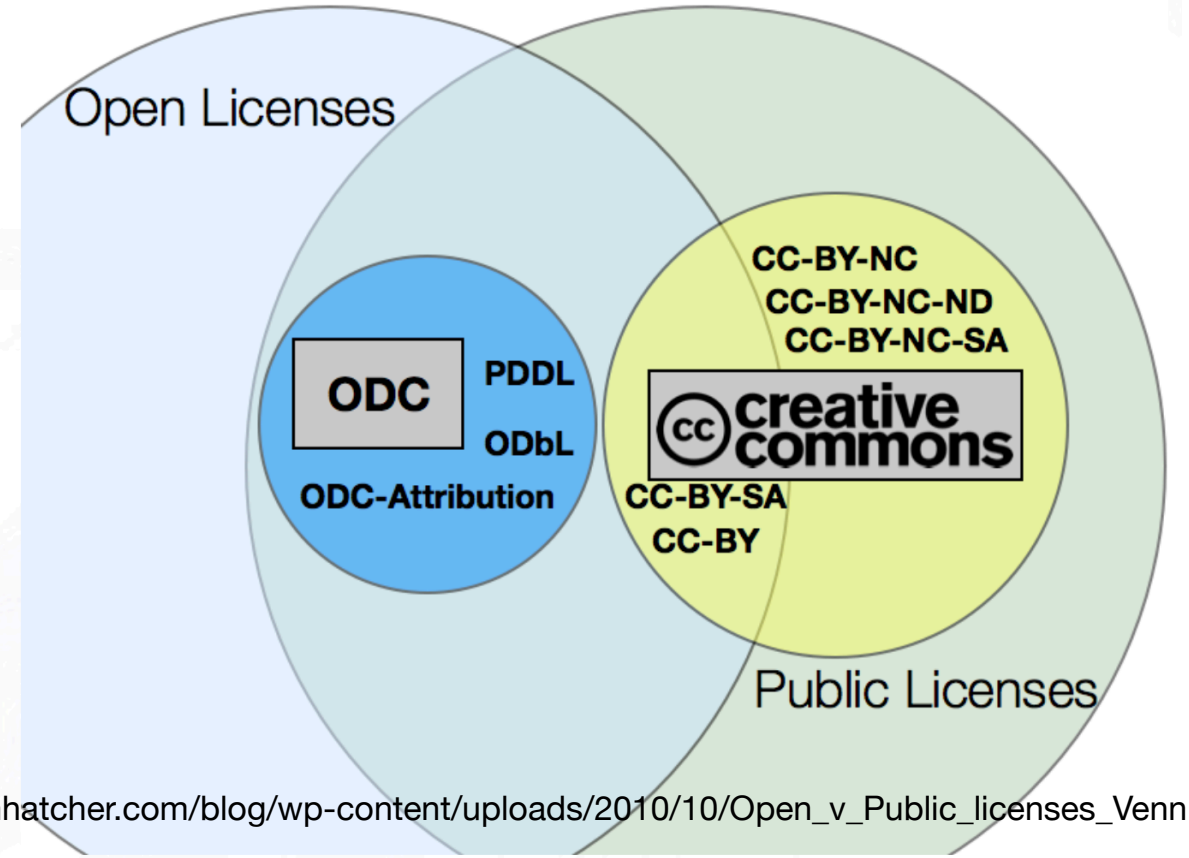


- Make your stuff available on the Web
  - Whatever format under an open licence
- Make it available as structured data
  - E.g., Excel instead of an image scan of a table
- Use non proprietary formats
  - CSV instead of excel
- Use URIs to identify things so that people can point at your stuff
- Link your data to other data to provide context

http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

# Licensing






http://www.jordanhatcher.com/blog/wp-content/uploads/2010/10/Open_v_Public_licenses_Venn.002-001.png

# Open Data Licenses

| LICENSE | BY | SA | Comment |
|---|---|---|---|
| Open Data Commons Public Domain Dedication and Licence (ODC PDDL) | N | N | Dedicate to the Public Domain (all rights waived) |
| Open Data Commons Attribution License | Y | N | Attribution for data(bases) |
| Open Data Commons Open Database License (ODbL) | Y | Y | Attribution-ShareAlike for data(bases) |
| Creative Commons CCZero | N | N | Dedicate to the Public Domain (all rights waived) |

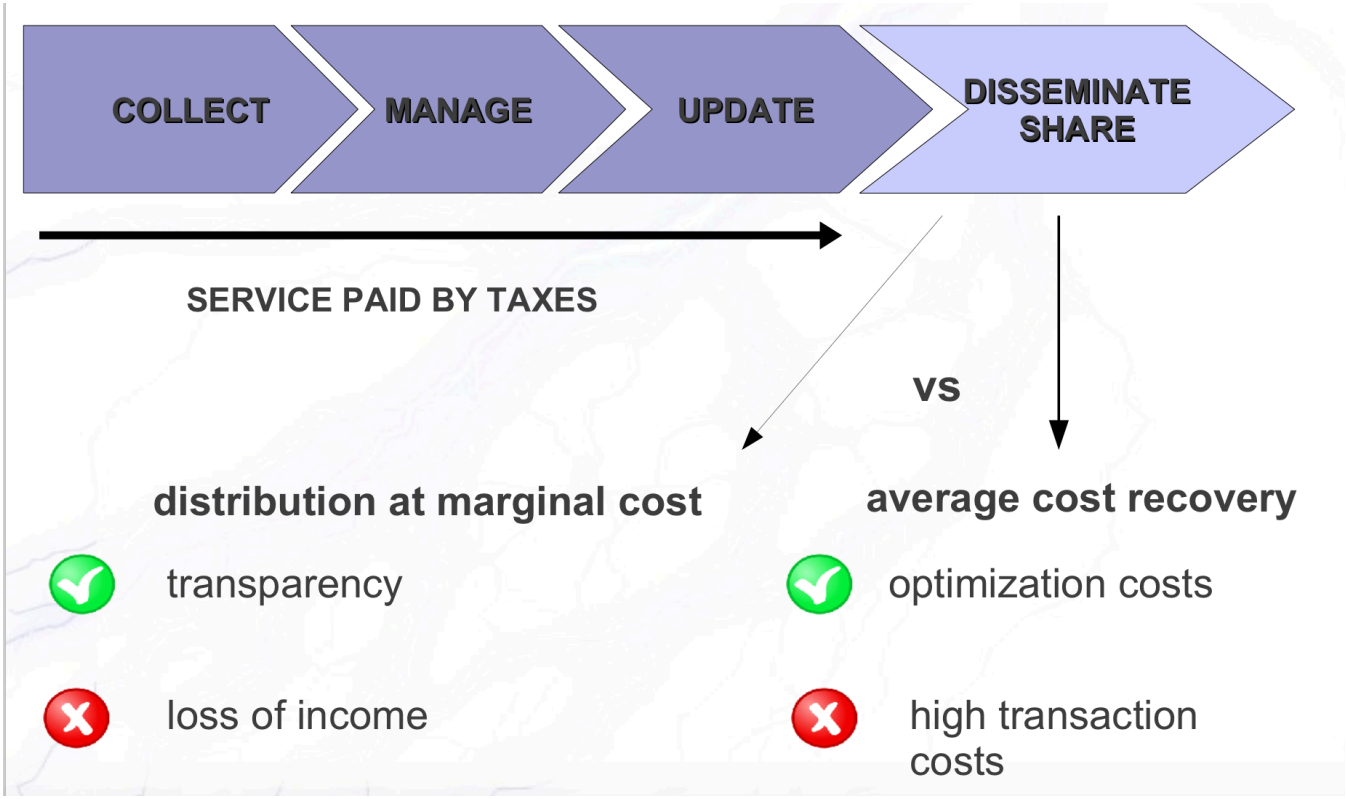
http://www.opendefinition.org/licenses/

http://www.opendefinition.org/guide/data/

**What Legal (IP) Rights Are There in Data(bases)**

# Public affairs and data





# How to make this possible



- Leadership: also/primarily political
- Crisis: budget cuts
- Heroes: people that believe is possible but also ready to lose
- Pressure from citizens
- Innovative companies (ex. Data publica)

# Data “tools”



- Compehensive Knowledge Archive Network
- About the project http://ckan.org
- The project http://thedatahub.org
- The italian hub http://it.ckan.net
- Europe support http://publicdata.eu/
- World http://opendatasearch.org/
- Get the data http://getthedata.org/
- DATAPKG
  - http://packages.python.org/datapkg/

# Open knowledge initiatives

http://www.okfn.org/

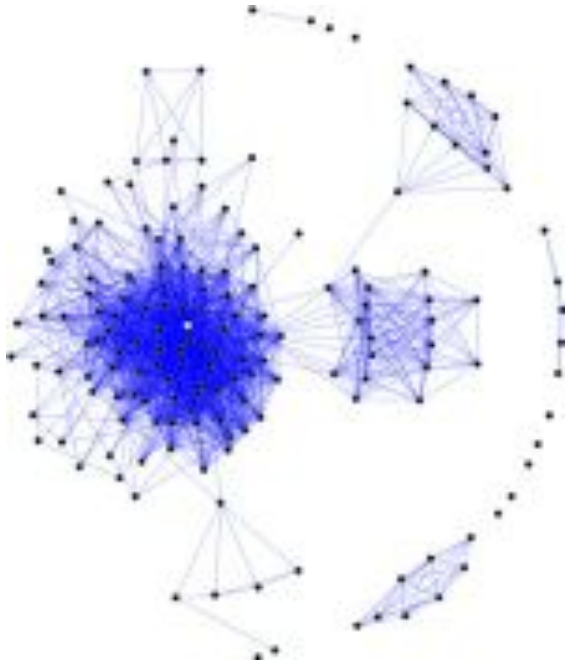


- http://opengovernmentdata.org/
- http://opendatachallenge.org/
- http://wheredoesmymoneygo.org/
- http://openspending.org/
- http://energy.publicdata.eu/ee/
- http://publicdomainworks.net/
- http://bibliographica.org/

# Viral marketing

Laks V. S. Lakshmanan (University of British Columbia)

# Online Social Networking Sites

# Social Networks & Media

**SarcasticRover**

@SarcasticRover

*Not the real @marscuriosity...*
*like I care.*

4th Rock From the Sun · http://mars.jpl.nasa.gov/msl/

Follow

**291** TWEETS

**257** FOLLOWING

**63,197** FOLLOWERS

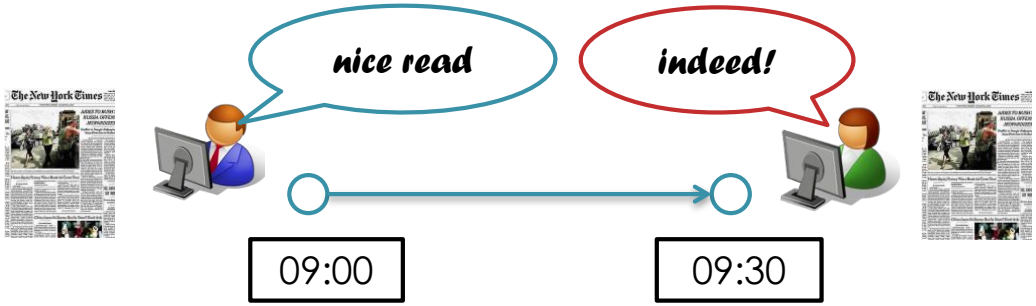
Oh sure, I can't think of anything I'd rather be doing than driving around a wasteland looking at dirt for the rest of my life.

**SarcasticRover** 3 days ago

I'm really glad all you fricken hipsters took a vague interest in science for 8 hours. Thanks for that.

**SarcasticRover** 3 days ago

# Information Propagation




People are **connected** and perform **actions**

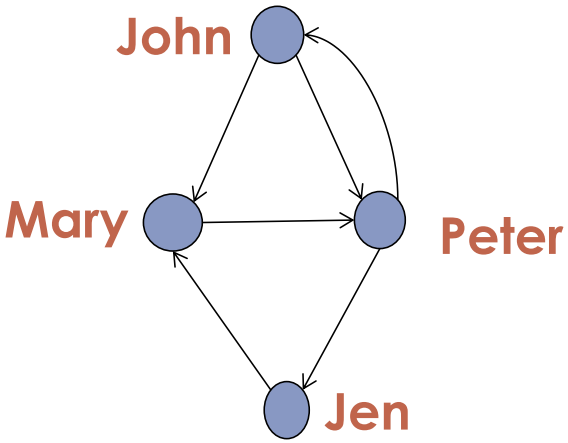
friends, fans, followers, etc.

comment, link, rate, like, retweet, post a message, photo, or video, etc.

# Basic Data Model

**Graph**: users, links/ties

$G = (V, E)$



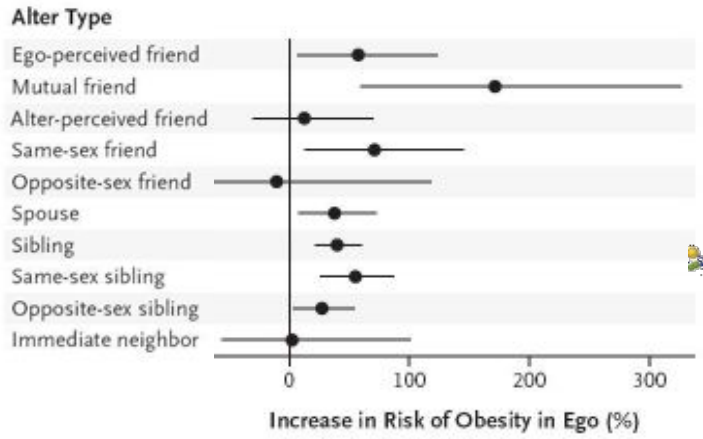

**Log**: user, action, time

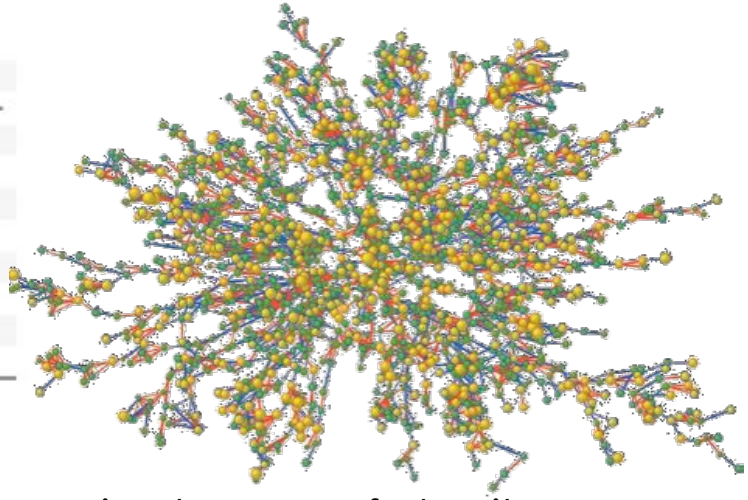
$A = \{\langle u_1, a_1, t_1 \rangle, \ldots\}$

| User | Action | Time |
|---|---|---|
| John | Rates with 5 stars "*The Artist*" | June 3rd |
| Peter | Watches "*The Artist*" | June 5th |
| Jen | … | … |

# Social Influence: Real-world Story I

12K people, 50K links, medical records from 1971 to 2003



Obese Friend → 57% increase in chances of obesity

Obese Sibling → 40% increase in chances of obesity

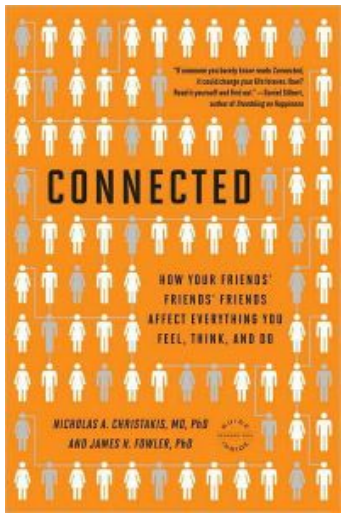
Obese Spouse → 37% increase in chances of obesity

***[Christakis and Fowler, New England Journal of Medicine, 2007]***

# Social Influence: Real-world Story II

Key to understanding people is understanding ties between them.

Your friend's friends' actions and feelings affect your thoughts, feelings and actions!

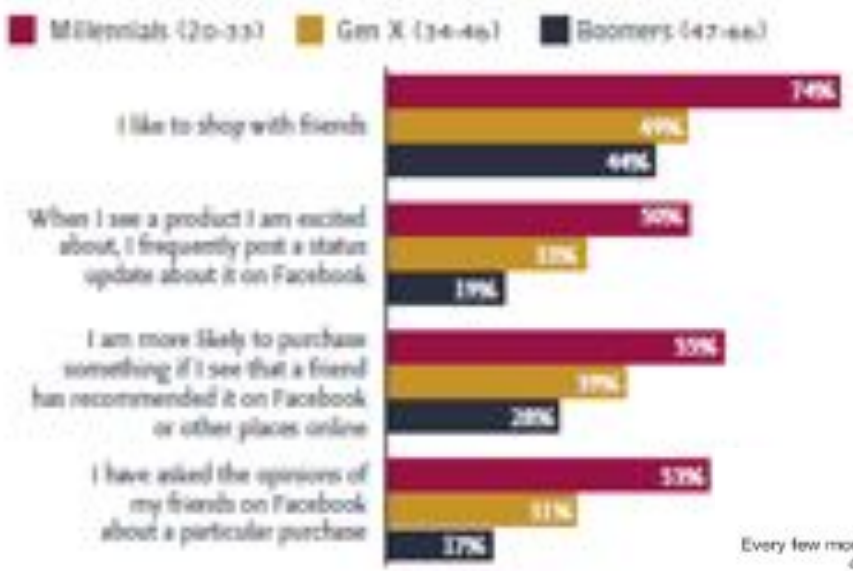



- **Back pain:** spread from West to East in Germany after fall of Berlin Wall
- **Suicide:** well known to spread throughout communities on occasion
- **Sex practices:** e.g., growing prevalence of oral sex among teenagers
- **Politics:** the denser your connections, the more intense your convictions

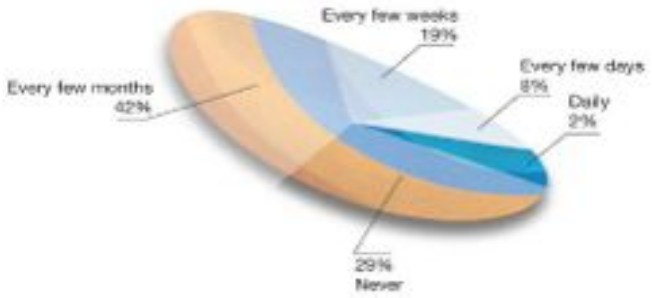
*[Christakis and Fowler 2011]*

# Application: viral marketing

**Purchase decisions are increasingly influenced**
by opinions of friends in Social Media




How frequently do you share recommendations online?



# Viral/Word-of-Mouth Marketing

- Idea: **exploit social influence for marketing**
- Basic assumption: word-of-mouth effect
  - Actions, opinions, buying behaviors, innovations, etc. propagate in a social network
- Target users who are likely to produce word-of-mouth diffusion
  - Additional reach, clicks, conversions, brand awareness
  - Target the influencers

# Transitivity of trust

- **Trust** is associated with the belief of an agent in the assertions by other agents; it is neither necessary nor sufficient for influence
- The **Web of Trust** from the early 1990s
  - Public Key Certification
  - *Advogato*: propagate trust through links
- **Transitive social importance** from the late 1940s
  - Seeley 1949, Wei 1952, Katz 1953: transitive importance computation
  - Reinvented as **PageRank** [Page et al. TR 1998]
  - TrustRank [Gyongyi et al. VLDB 2004], EigenTrust, Trust/distrust propagation

# Identifying influencers

- Influencers increase brand awareness. product conversions through WoMM
  - Influencers advocate a brand
  - Influencers influence purchasing actions

# Identifying influencers: start-ups

- **Klout**
  - Measure of overall influence online (initially Twitter, now FB and LinkedIn)
  - Score = function of true reach, amplification probability and network influence
  - Claims score to be highly correlated to clicks, comments and retweets

- **Peer Index**
  - Identifies/Scores authorities on the social web by topic

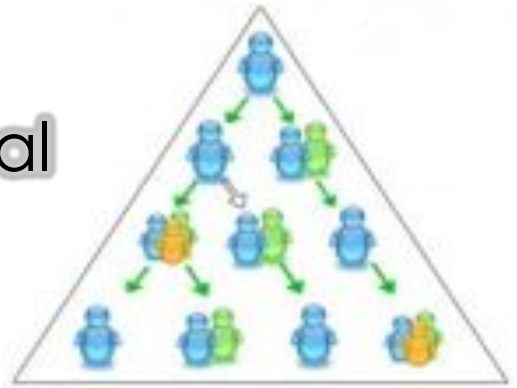
- **SocialMatica**
  - Ranks 32M people by vertical/topic, claims to take into account quality of authored content

- **Influencer50**
  - Clients: IBM, Microsoft, SAP, Oracle and a long list of tech companies

+ Svnetwork, Bluecalypso, CrowdBooster, Sproutsocial, TwentyFeet, EmpireAvenue, Twitaholic, and many others …

# Viral marketing & The Influence Maximization Problem

- Problem statement:
  - **find a seed-set of influential people such that by targeting them we maximize the spread of viral propagations**
- Focus of **Part II** of this tutorial



# The first definition of IM problem: A Markov random fields formulation

- Each node $i$ has random variable $X_i$, indicating bought the product or not, $\boldsymbol{X} = \{X_1, \ldots, X_n\}$
- Markov random field formation: $X_i$ depends on its neighbors' actions $\boldsymbol{N_i} \subseteq \boldsymbol{X}$
- Marketing action $\boldsymbol{M} = \{M_1, \ldots, M_n\}$
- **Problem**: find a choice of $\boldsymbol{M}$ that maximizes the revenue obtained from the result of $\boldsymbol{X}$

*[Domingos & Richardson KDD 2001, 2002]*

# Major stochastic diffusion models

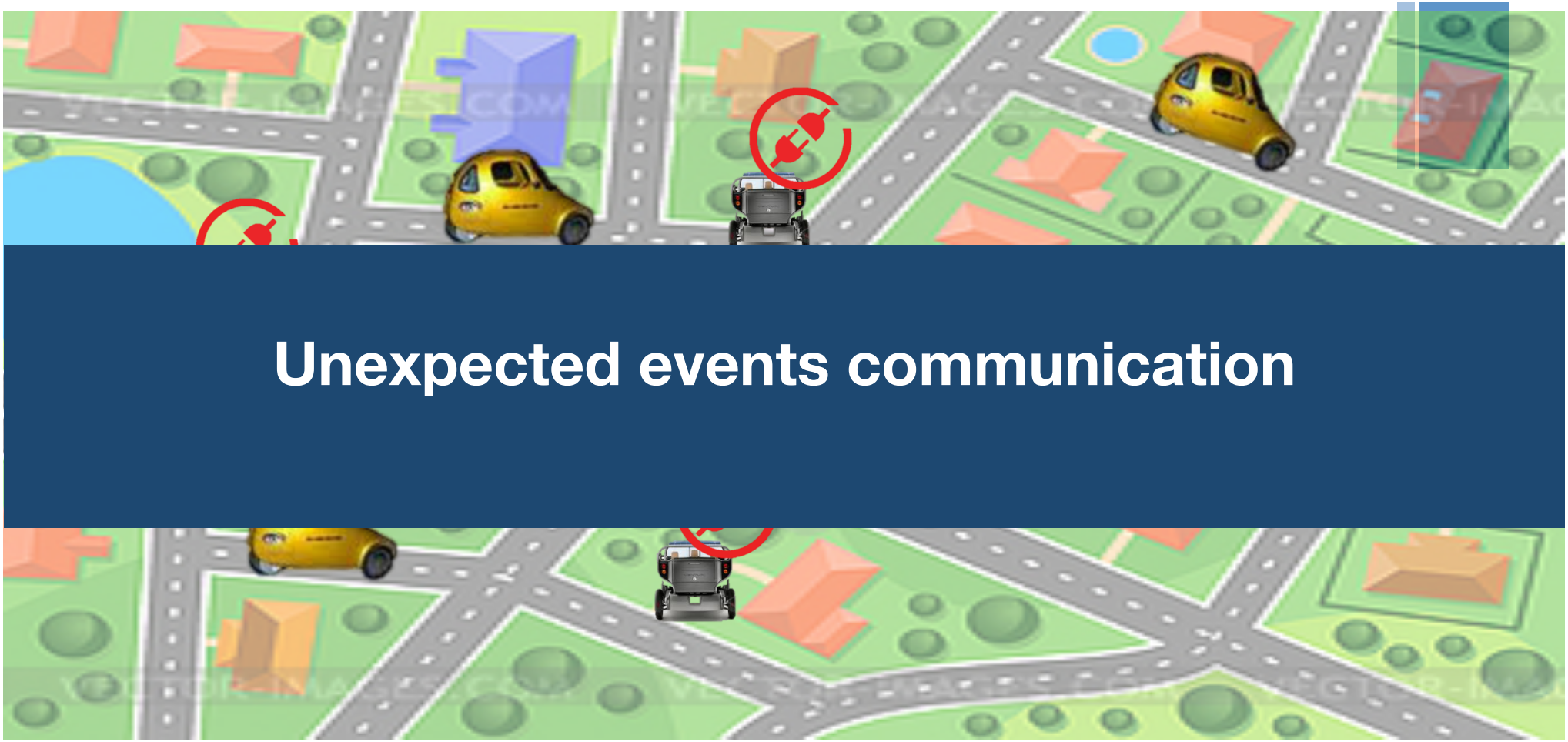
- **Independent cascade** (IC) model
- **Linear threshold** (LT) model
- General threshold model
- Others
  - Voter model
  - Heat diffusion model

# Smart transport

# Vehicles position
# Energy levels

novemb

ecautoxis



# Unexpected events communication

novembre 27, 2014

# Queue length at the recharging stations location

novembre 27, 2014

# **Decision making** for the autonomous vehicles to help **piloting the vehicles** to their **destination**

ecautoxis

novembre 27, 2014

# Ensuring **vehicles availability, service continuity** avoiding **accidents**

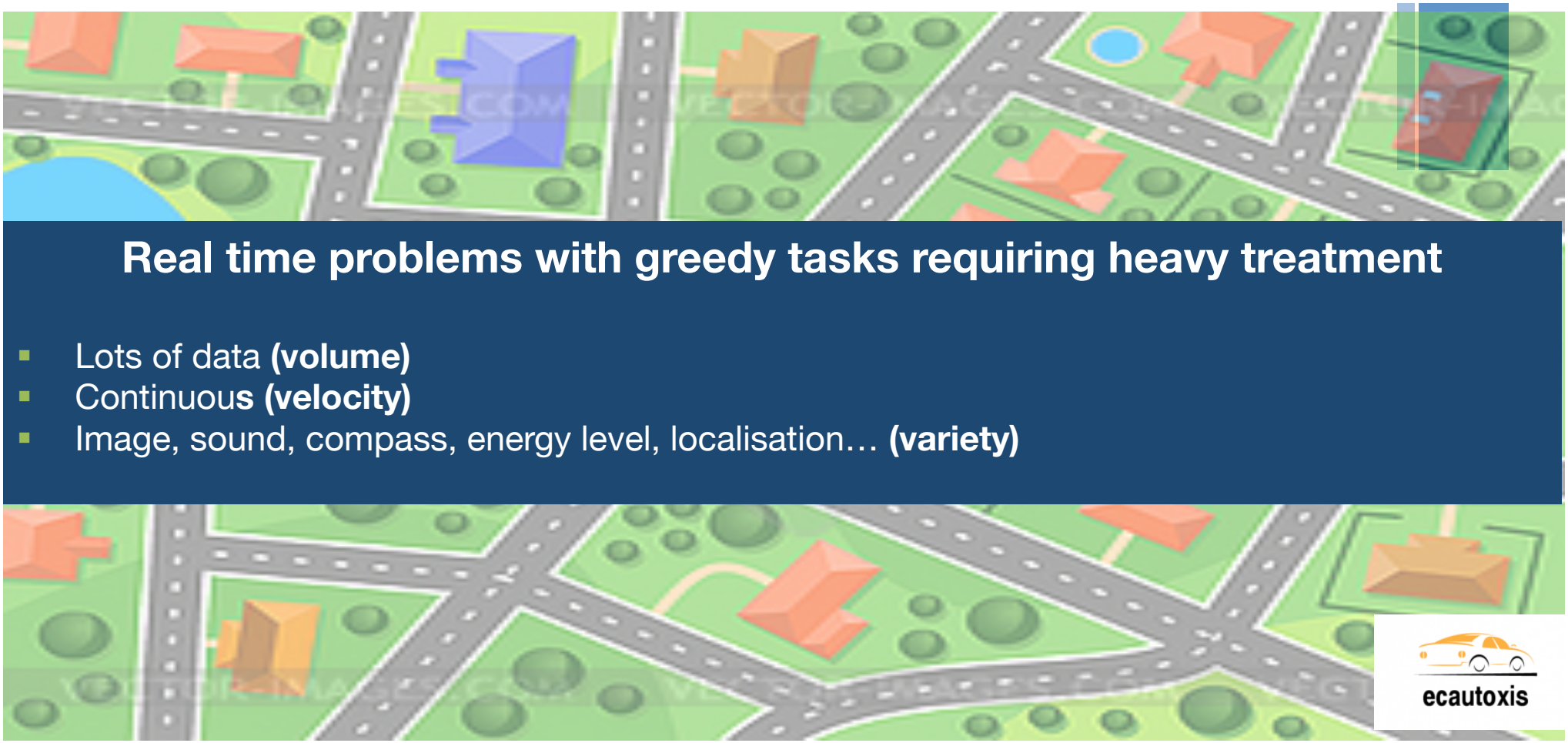
ecautoxis

novembre 27, 2014

# Ensuring **optimal recharging, through mobile recharging units**

ecautoxis

novembre 27, 2014



# Real time problems with greedy tasks requiring heavy treatment

- Lots of data **(volume)**
- Continuous **(velocity)**
- Image, sound, compass, energy level, localisation… **(variety)**

ecautoxis

# Big data and intelligent transport

- Transdec: big data for transportation
  - http://imsc.usc.edu/intelligent-transportation.html
- How big data drives intelligent transportation, Rocky Moutain Institute
  - http://www.greenbiz.com/blog/2012/08/15/how-big-data-drives-intelligent-transportation
- Real-Time Data Capture and Management
  - http://www.its.dot.gov/data_capture/data_capture.htm
- Traffic analytics

# Current transport projects and apps

## BIG DATA + CLOUD

novembre 27, 2014

# Cloud: services for greedy big data processing

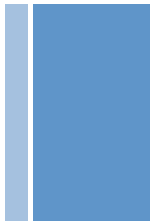
Service Oriented architecture: *everything as a service*


- Access via a network, on demand, self-service, computer resource
- Elasticity, flexibility, and unlimited computing, storage and memory resources for executing greedy operations on Big data

# Problem statement

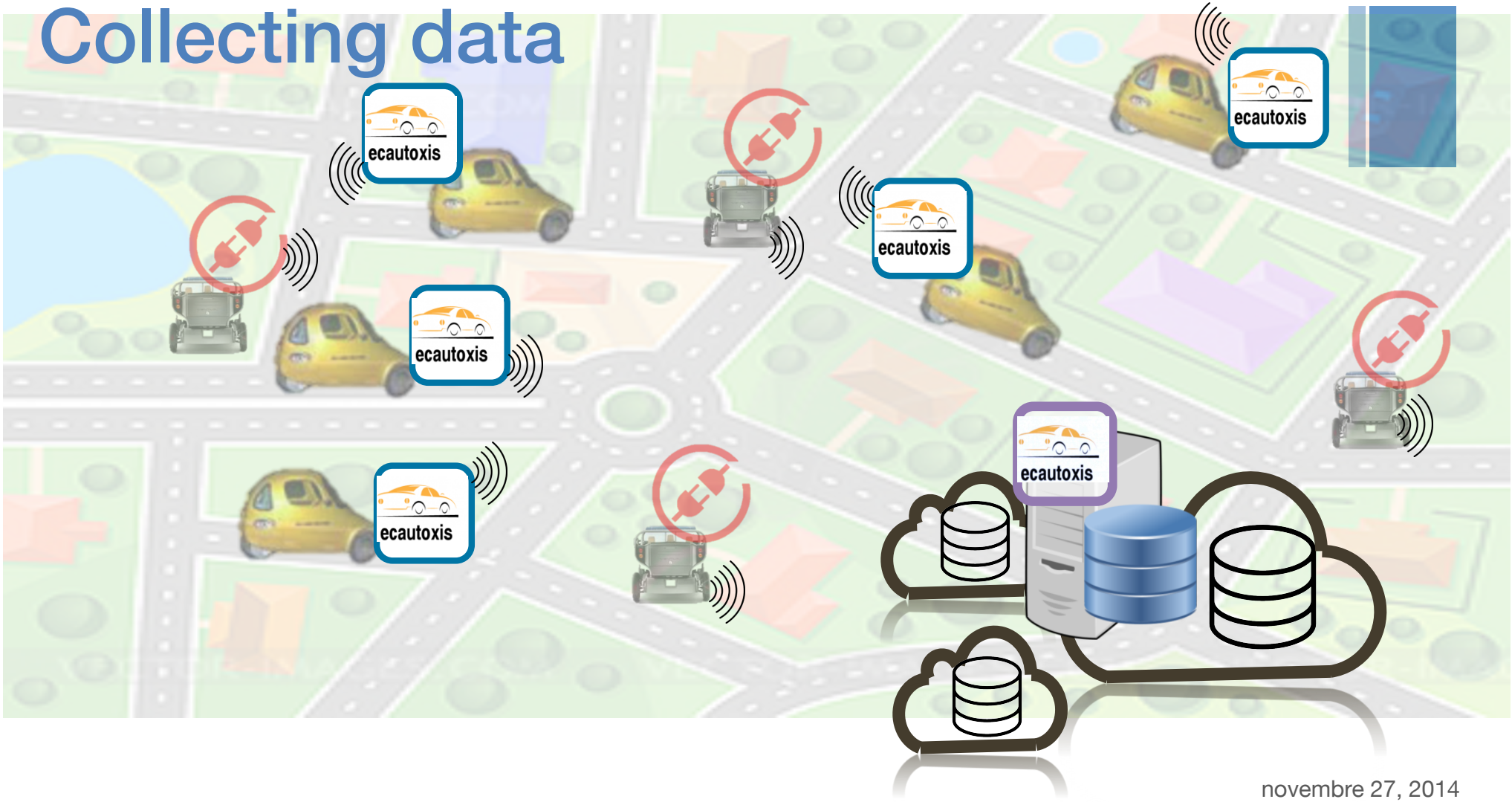
- Data collection (what sources ?: compass, video stream, LADAR…)
- Data storage (keep or not and how long : missed parked car or someone crossing the road)
- Data communication strategy optimise network (rate of communication, who's initiative)
- Scalability ( if we need extra vehicles: make it work with a 100 and with a 1000)
- polyglot programming (different programming for different needs)
- Data → information (image video → information de localisation)

# Objectives

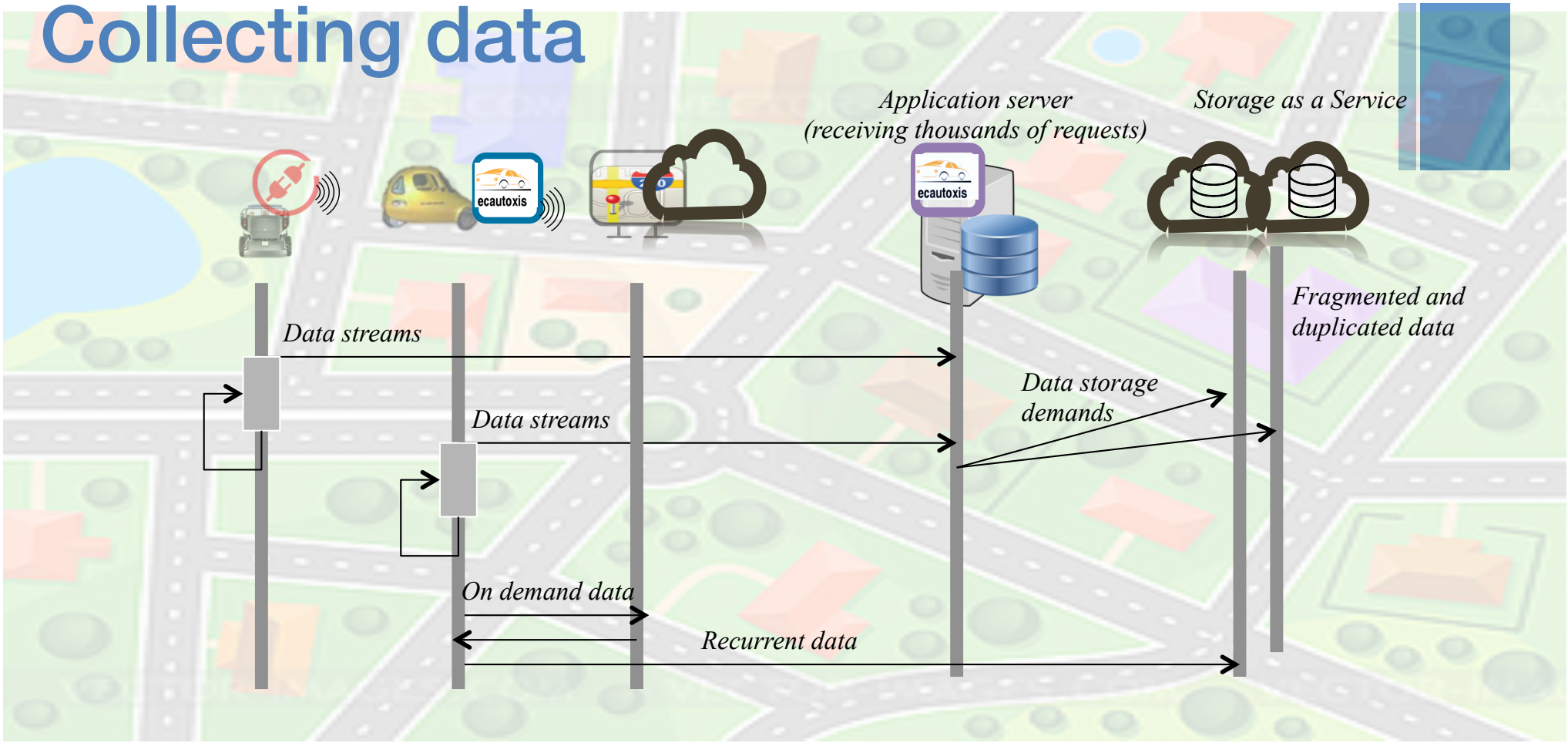


- Develop service using big data for decision making
- Using Cloud and Streaming as tools
- Insuring that big data, cloud and streaming work well together

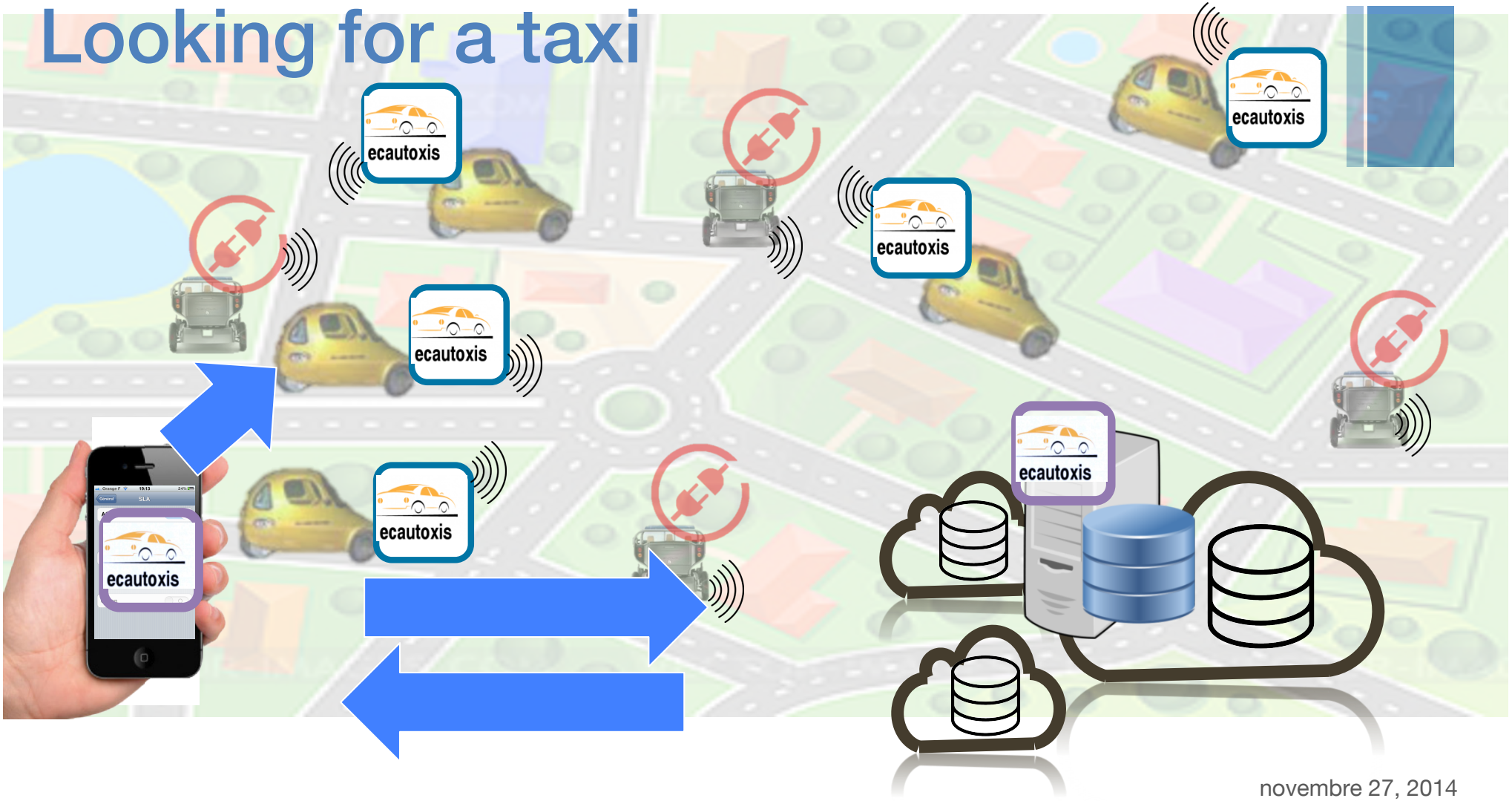
# Collecting data



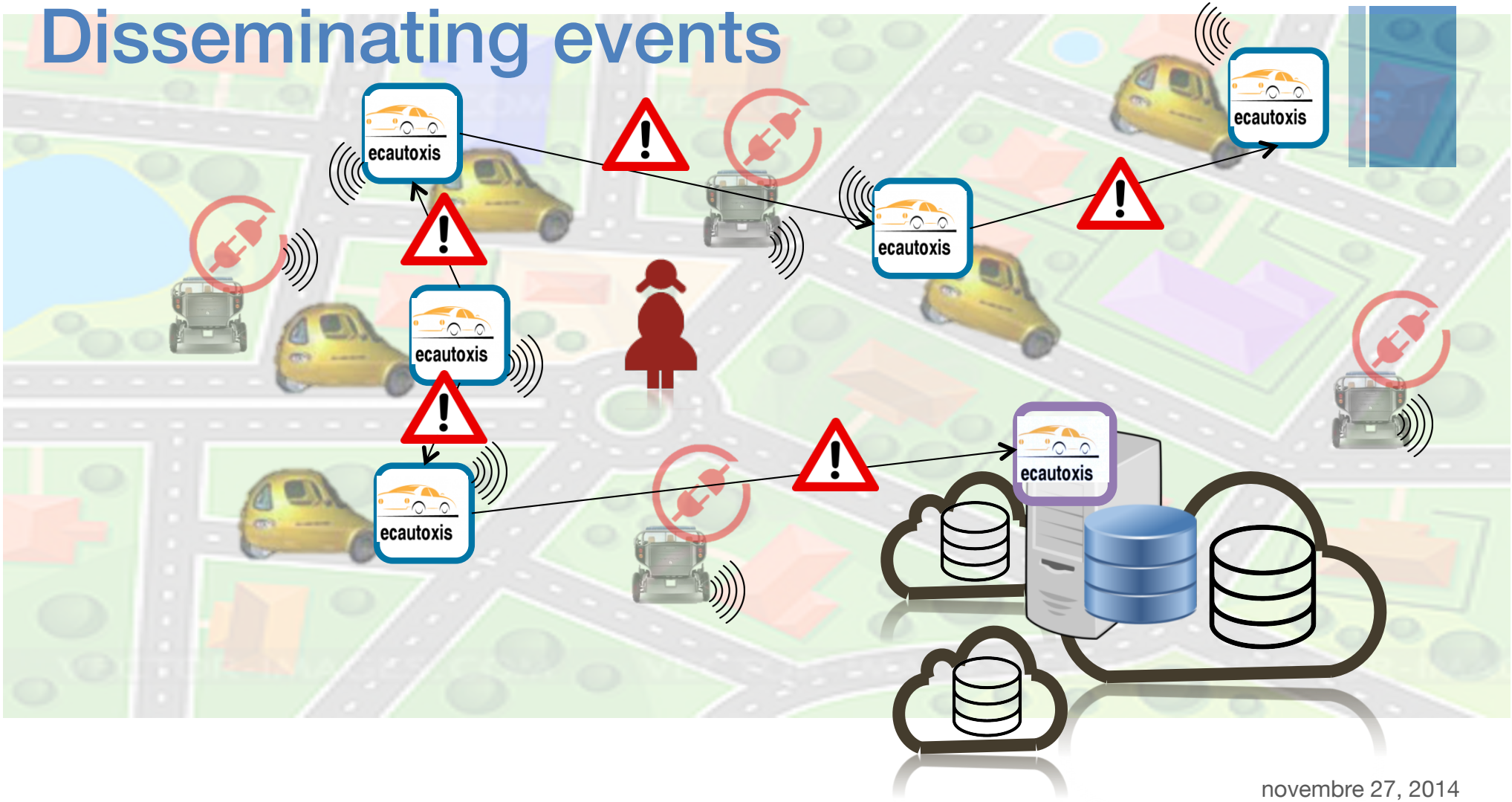
# Collecting data



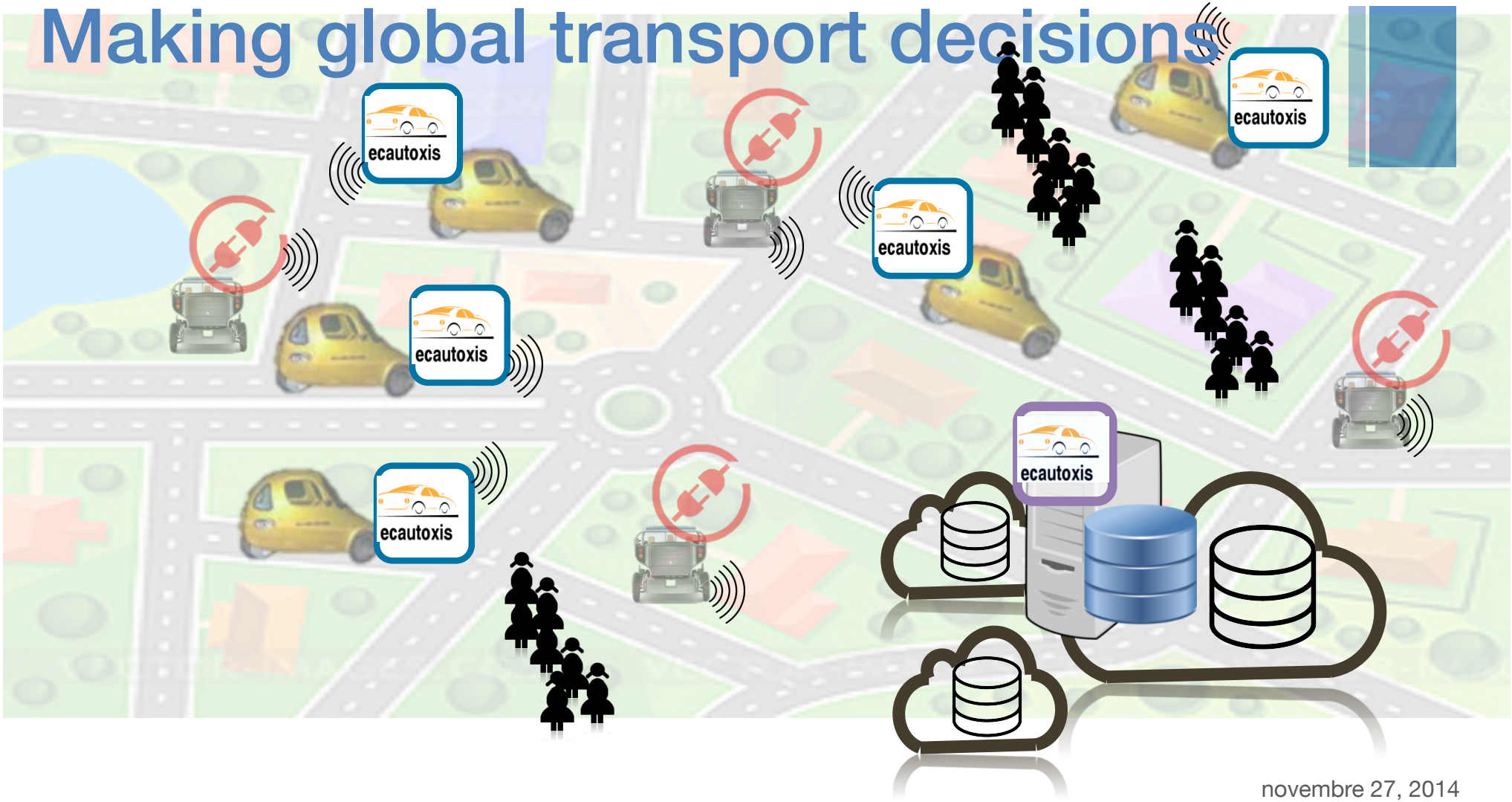
# Looking for a taxi



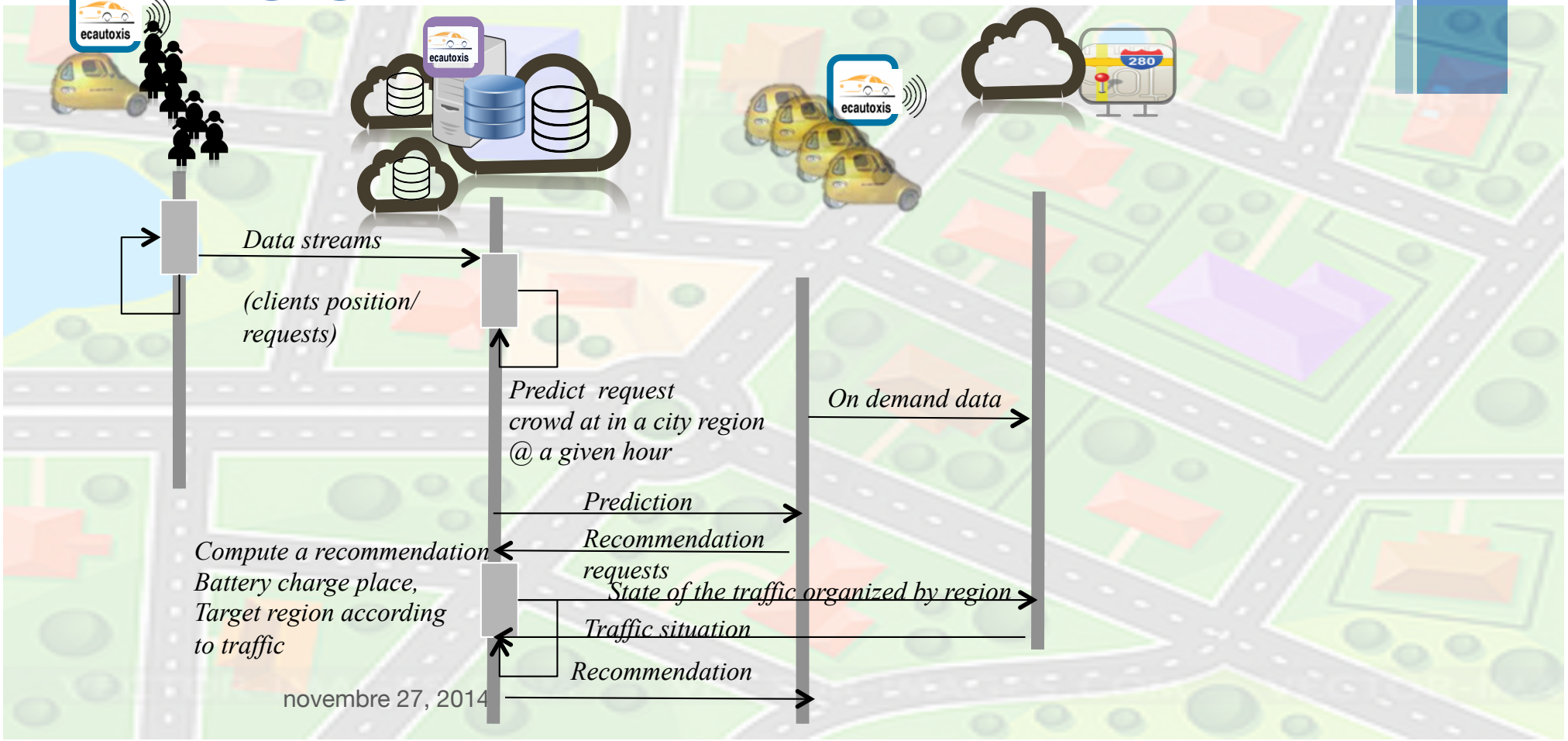
# Disseminating events



# Making global transport decisions



# Making global transport decisions



# Thanks Merci Gracias

*Contact*: Genoveva Vargas-Solar, CNRS, LIG-LAFMIA
Genoveva.Vargas@imag.fr
http://www.vargas-solar.com/teaching