Urban resilience, social exclusion, and poverty – (big) data challenges

Maria-Cristina Marinescu, Barcelona Supercomputing Center

Cities can`t be smart without reflecting cross-domain dependencies

NEW HOUSING IN PREVIOUSLY RAN DOWN NEIGHBOURHOOD
… housing may become unaffordable for previous neighbours
… neighbourhoods loose identity, friends/family move apart
… existing local issues spread globally (crime, STDs – Baltimore housing projects, ...)

TRAM TO AVOID TRAFFIC JAMS AND RUN DIRECTLY
… may collapse car /public transport. traffic that runs in different direction

BETTER, FASTER MAIN AVENUE
… may split neighbourhoods

WALKABLE CITIES
… commercial traffic worsens, no place to stop
… less parking space for neighbours
… possibly in places where people don`t usually walk! (e.g. steep hills)

CHANGING / DAMMING WATER COURSE FOR CITIES
… ecosystem changes
… displaced people

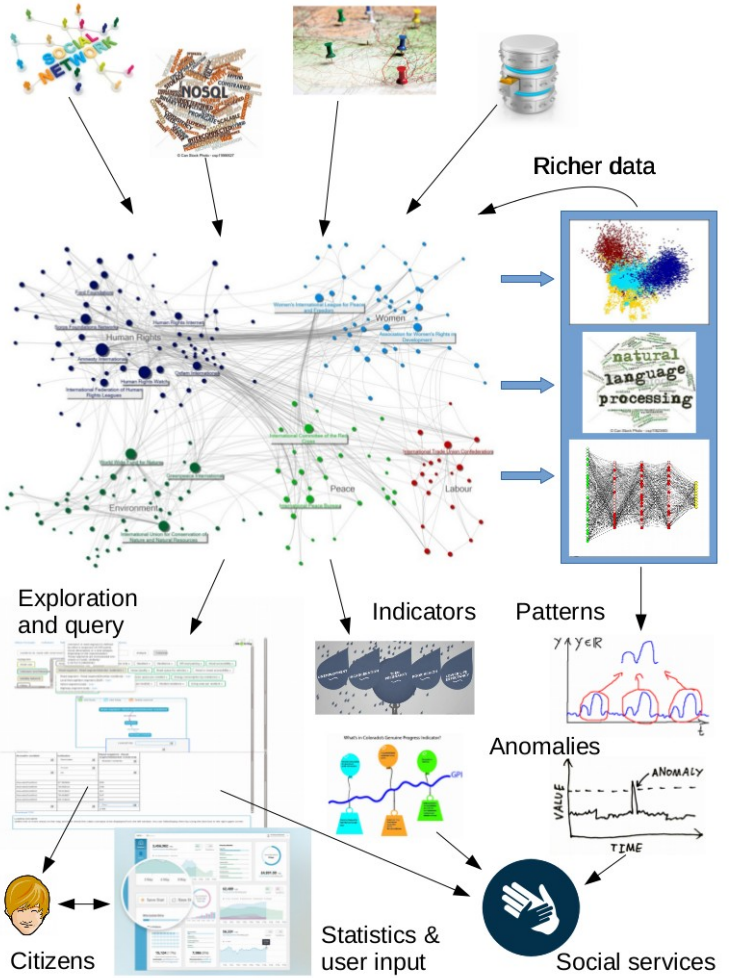
CITY FOCUSES ON LOCAL PRODUCE / RURAL REGION SPECIALIZES IN FEW HIGHLY DEMANDED PRODUCTS
… draught or plague may hit a crop → there`s no backup plan for rural region, and delayed response in city

Concrete problems in urban environments

- Need to plan, respond, optimize, predict, asses risk, etc
- E.g. *individuals at risk of*

  poverty or social exclusion
  - Water consumption data
  - Land registry
  - Income / taxes
  - Education, school attendance
  - Health services
  - Etc…

    What if no access to data?



Concrete problems in urban environments

- Need to plan, respond, optimize, predict, asses risk, etc, which requires:

Instrument and monitor the city

Integrate heterogeneous data sets

Space, time, unstructured data, etc – much more data, new types of data

Analyze and query

Understand complex patterns from big data

Compute and optimize metrics

Understand the effect of changes/decisions beforehand

Predict, including (timed) events

Simulate

Visualize

Why integration via modeling?

- ***Simplifies the development*** of applications that require integrated access to city data sources (cross-domain)
- Enables *solution reuse* as we move from one city to the next
- The world is open, changing, incomplete, and data may be faulty. Models are ***easy to evolve and maintain*** (reuse, repurpose, naturally models change) without modifying the application or the data sources.
- Standardized indicators for cities may be implemented as part of the model.
- Reasoning (inference and rules)

Interoperability challenges

Data sharing - LOD

Data cleaning

Data enrichment (metadata)

Data integration (via modeling)

Building the model: evolving ontologies, competence queries
- need complex queries!!
- need tools to make complex models user fiendly

Models as a way to integrate data:
- often lots of missing data + some data can`t be monitored
- data may not "naturally" fit the model´s relationships - semantic approximations? e.g. Legal entities / address in a block
- granularity of data issues

Format / mapping issues (data to model mapping)

Analysis

Prediction:
- garbage in / garbage out - data skew or poor quality
- correlation NOT causality; common sense is the hardest to learn by a machine (semantics can capture common sense)
- e.g. semantic helps to disambiguate (St. George on a Bike)
- understand propagating effects? e.g. earthquake

Reasoning (w. Probabilities)

Computing indicators:
- expert formulas vs via data mining
- can standardize city indicators as part of the model

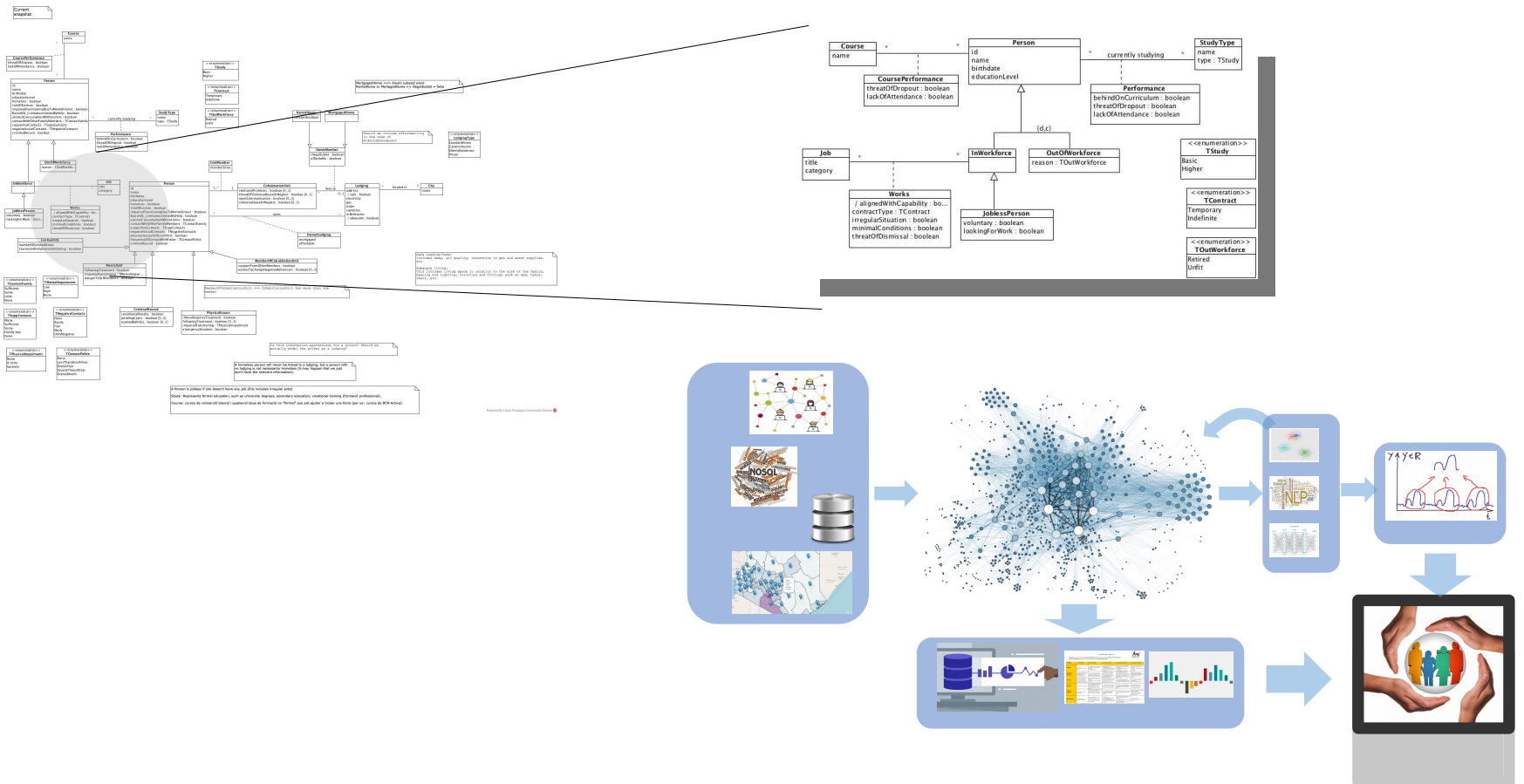
Data quality

Model constraints / programatically
Depends on the application
Some data is subjective

Data can be plain wrong – without Crossing data virtually impossible to know e.g. business doesn´t declare changes, Gentrification (AirBnB), reno w/o licence

Use case 1: Social exclusion and poverty

- LinDaFIX: Linked data for fighting inequality
  - Imprhove identification of vulnerable individuals
  - Improve / standardize the evaluation process by social workers
  - Learning patterns to try to predict hidden exclusion



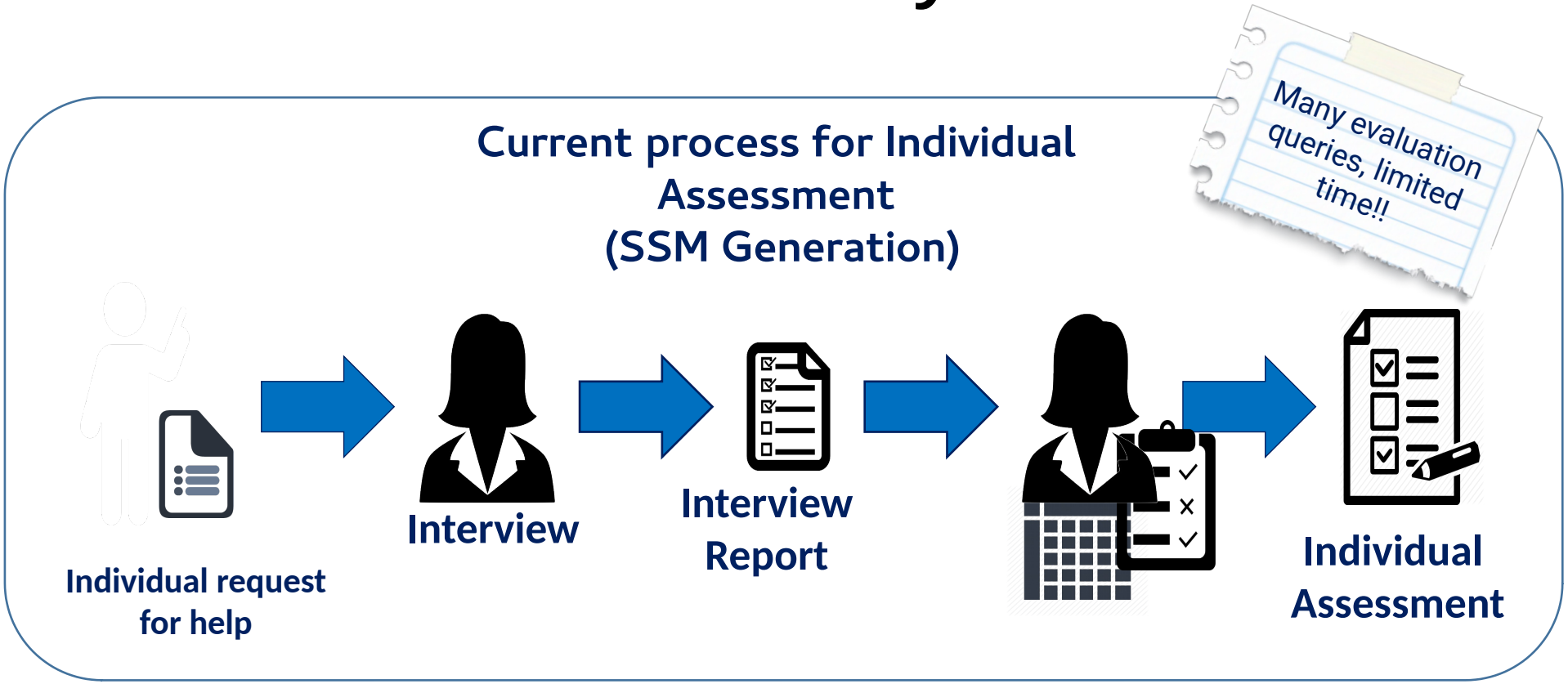
Use case 1: Social exclusion and poverty

- Need data from very many different (types of) sources: finances (income, spending, debts), work and education (qualifications, years in school, registered in programs), housing (quality of housing, illegal sublet), domestic issues, mental / physical health issues, substance use, social support, community participation, problems w the law, dependents, schooling, day care, ...
- ... but lots is missing and can disconnect the model graph
- … and some data is statistical – not enough to learn patterns, e.g. “not socially integrated may → addiction”, “drinking and living alone may → skip meds”, “drinking and small kids may → domestic violence”, “unstable job and chronic disease may → in danger of eviction” etc.
- E.g. what exists: individual and household data (address, country of origin, nationality, birthdate, gender, civil status, household_id…) - this comes from surveys; subventions approved – amount and for how many people… data from programs to connect elderly ...
- Much more data theoretically available, haven´t seen it yet...food help, homeless, health, land registry, ...

Use case 1: Social exclusion and poverty

- Rights to access the data + anonymization (laws different in different places at different times)
  - People don´t ask for help (it´s cultural; need to change the way people percieve the profile of the poor or excluded)
  - Example in Barcelona - farmacies raise alarms when people that take meds don´t show up - locally things are possible that i can`t do globally
- Data is not always clean: **incompletenes ok** - inconsistencies not ok, constraint checking as part of the model vs programatically (+ and -), Open World Assumption!!
  - same apt_id but different addresses (use additional info to understand where the error is - use ids of people living there? other data?)
  - 2 people/ 3 ids in same appt (are 2 of them the same person?) vs 4 people/ 3 ids (are there underage individuals living there?)
  - addresses with MANY inhabitants – usually social services locations!
  - Married / separated underage (or under 16)
- Data comes in CSV, Excel, text, RelDB

Self-sufficiency matrix



Why use SSM?

- Holistic view of an individual
- "Standard" way of evaluation
- Accountability of decision making
- Evaluation of policies' impact
- Proven usefulness / already in use

- Social workers could take advantage of an approximate initial snapshot to focus on real causes
- Semantic model and indicators can help!

What we are doing

- Build a top level model for social explusion from the literature – Social Services dept not very :)
- Build a more targeted model starting from SSM
- Populating data – many data missing
- Computing indicators based on SSM
  - Some data is currently not real! ...proof-of-concept :(
  - What if data is not available when intetrviewing?
- We absolutely need data to learn patterns
  - Especially for hidden exclusion / poverty!!
  - Looking for data – Mexico, online GB data (most open data is statistical…), other city halls?? (very difficult, even when they have a commitment); we could also apply NLP but documents also usually talk about statistics

Use case 2: Urban resilience

- Project:
  - Model water system (to improve resilience), contamination, transportation, energy, integrated infrastructures (to reduce CO2 footprint, energy consumption) - all in Barcelona (mostly w. City council / depts)
    - Data is pretty basic (sensors, components: taxi stands, electric vehicles, smart towers, aggregated gas/electricity consumption per building, ...) and doesn´t populate the model well enough
    - Statistical data it´s not enough, but people don`t want to publish non-aggregated data
  - Integrated with urban planning model (w. IBM and the Barcelona Urban Ecology agency):
    - We received cleaned data (problem examples: intersecting geometrical shapes, building heights in meters and floor numbers didn´t match) - they decided beforehand which source to believe!
    - Uneven data between neighbourhoods: different relationships are populated, different space granularity
    - Very little time data
    - Computing sustainability indicators

Use case 2: Urban resilience

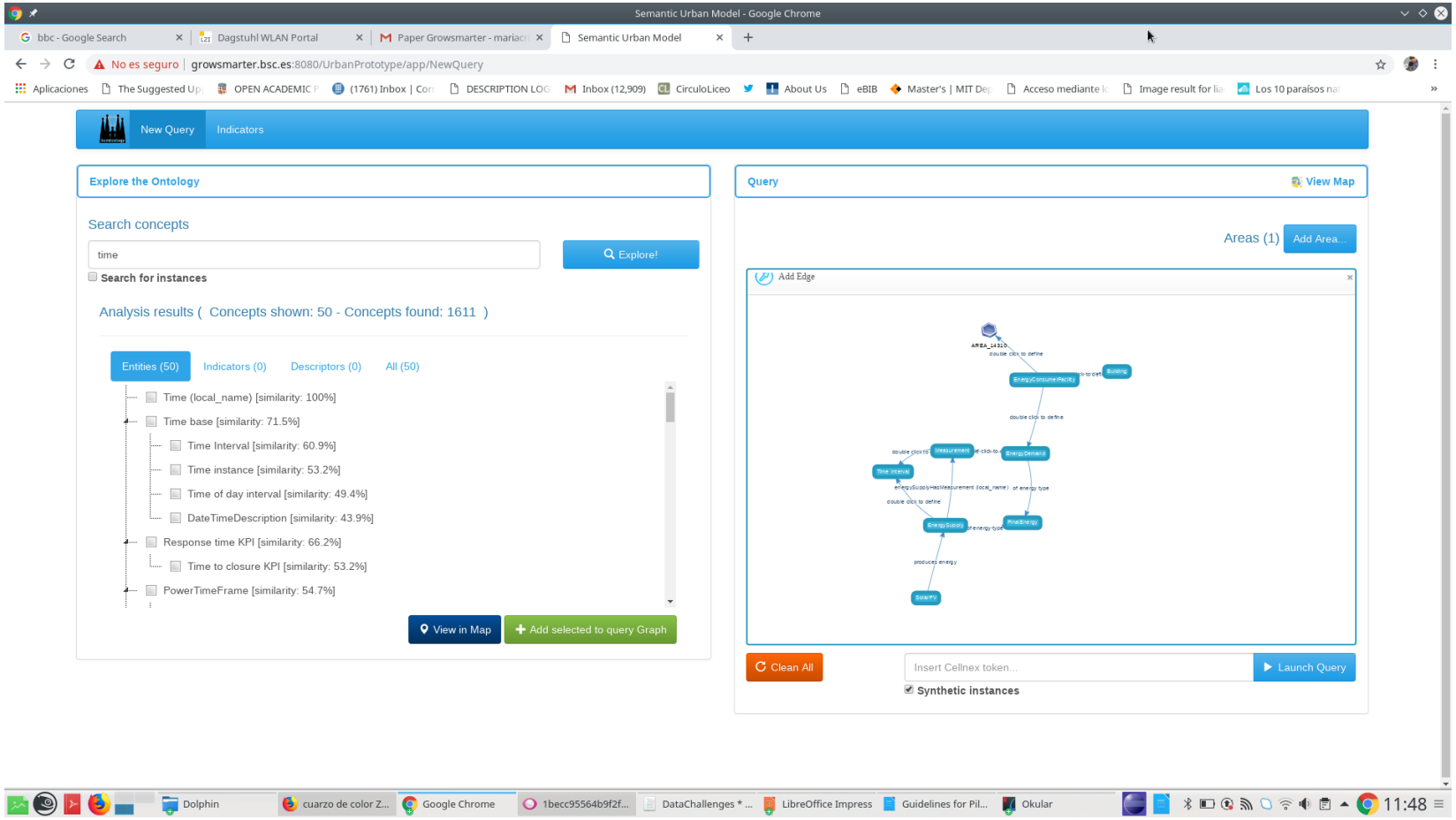
Model water system, contamination, transportation, energy, integrated infrastructures

Urban planning

- Building models is costly – semi-automation of the process? tools around the models, which are complex for domain users
- Surprise: People tend not to stress cross-domain interconnections in the model because it´s something that was not possible before – preliminary step to see the utility of semantic integration

- [UNHabitat – urban resilience model: modeling, no data, they wanted to (1) compute indicators based on survey data rather than (open) data, (2) simulate cascaded effects.]

Use case 2: Urban resilience



High-level objectives

- Data cleaning
- Approximate queries
- Cultural issues
- ML (probabilistic models) and common sense models
- Data quality
- Value of using ontologies
- Ontology discovering and mapping
- How can i make better use of data, which lots of times is mostly statistic?
- ...etc etc

Questions?

maria.marinescu@bsc.es