



# Variety in Big Data: A Cities Perspective

---

**Mark S. Fox**

Professor of Industrial Engineering & Computer Science

Senior Fellow, Global Cities Institute

University of Toronto

[msf@eil.utoronto.ca](mailto:msf@eil.utoronto.ca), [www.eil.utoronto.ca](http://www.eil.utoronto.ca)

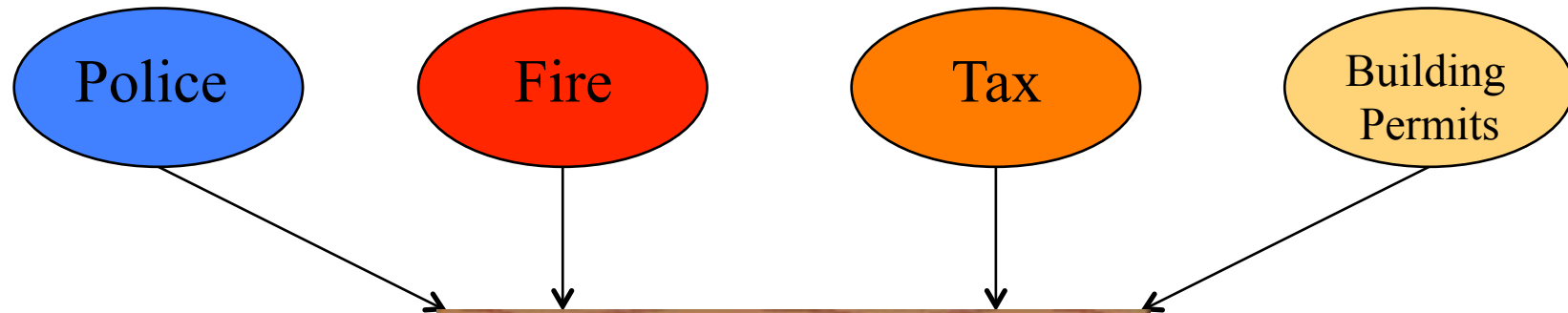
27 March 2014

# Big Data in the City

- NYC building owners were illegally converting them into rooming houses that contained 10 times the number people they were designed for.
  - Posed a number of problems, including fire hazards, drugs, crime, disease and pest infestations.
  - There are over 900,000 properties in New York City and only 200 inspectors who received over 25,000 illegal conversion complaints per year.
- How to distinguish nuisance complaints from those worth investigating?
  - Current methods resulted in **13%** of inspections issuing vacate orders.



# Data From Many Sources



- New York's Analytics team created a dataset combining data from 19 agencies:
  - Buildings, preservation, police, fire, tax, and building permits.

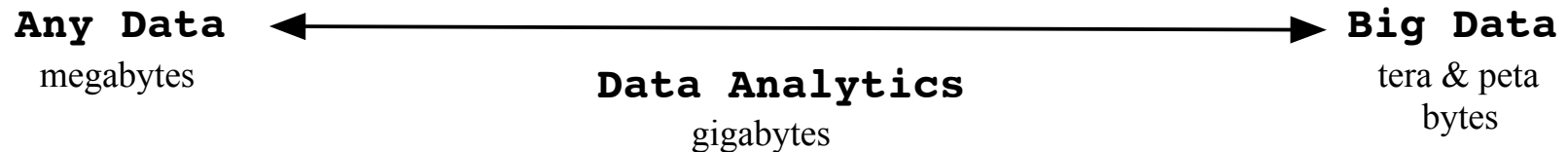
- By combining data analysis with expertise gleaned from inspectors, able to rate complaints such that in **70%** of inspector visits led to vacate orders.



# How big is **BIG**

---

- Most organizations are making the conversion from gut-based to fact-based decision making, and *any* facts will do!



- And the data? It’s “not so big”.

**Big Data = Data Analytics**





# The Real Bottlenecks

---

- Assuming you know what data you need  
(Big IF: often we do not know without doing some research, nor can you just dump all the data into a data mining tool and let the algorithm figure it out):
  - Where can I find it?
  - What are the attributes? What do they mean?
  - Are they equivalent to the attributes from other data sets?
  - Is the data correct? Complete? Can I trust it?



# Key Distinction

---

- Over the last 3 months I taught a course titled “Big Data and Global Cities” where each student did a project using data from Global Cities.
- **Data:** A set of values that are created by a repeatable, standardized, calibrated process.
  - Sensors.
- **Information:** A set of values that are created by a process that is inherently uncertain.
  - Determining the number of homeless people,
  - Municipal financial data reporting.



# Global City Indicators

---

A city can be defined as ‘smart’ when investments in human and social capital and traditional (transport) and modern (ICT) communication infrastructure fuel sustainable economic development and a high quality of life, with a wise management of natural resources, through participatory action and engagement. (Caragliu et al. 2009)



# Measurement Gap

---

- World Bank funded a study, by the University of Toronto, of city performance indicators at 9 cities.
- Belo Horizonte, Brazil
- Bogota, Colombia
- Cali, Colombia
- King County, Washington State, USA
- Montreal, Canada
- Toronto, Canada
- Vancouver, Canada
- Porto Alegre, Brazil
- Sao Paulo, Brazil



# Measurement Gap

---

- World Bank funded a study, by the University of Toronto, of city performance indicators at 9 cities.
- Belo Horizonte, Brazil
- Bogota, Colombia
- Cali, Colombia
- King County, Washington State, USA
- Montreal, Canada
- Toronto, Canada
- Vancouver, Canada
- Porto Alegre, Brazil
- Sao Paulo, Brazil

**Total of 1100 indicators across 9 pilot cities – only 2 comparable.**





# World Bank Vision (Hoornweg et al., 2006)

- **Objective:** clear, well defined, precise and unambiguous, simple to understand.
- **Relevant:** directly related to the objectives.
- **Measurable and replicable:** easily quantifiable, systematically observable.
- **Auditable:** valid, subject to third-party verification, quality controlled data (legitimacy across users).
- **Statistically representative** at the city level.
- **Comparable/ Standardized** longitudinally (over time) and transversally (across cities).
- **Flexible:** can accommodate continuous improvements to what is measured and how. Have a formal mechanism for all cities and interested parties to comment on.
- **Potentially Predictive:** extrapolation over time and to other cities that share common environments.
- **Effective:** tool in decision making as well as in the planning for and management of the local system.
- **Economical:** easy to obtain/inexpensive to collect. Use of existing data.
- **Interrelated:** indicators should be constructed in an interconnected fashion (social, environmental and economics).
- **Consistent and sustainable over time:** frequently presented and independent of external capacity and funding support.

## Over 100 indicators defined.

### City Services

- Education
- Finance
- Governance
- Recreation
- Transportation
- Wastewater
- Energy
- Fire and Emergency Services
- Health
- Safety
- Solid waste
- Urban Planning
- Water

### Quality of Life

- Civic Engagement
- Economy
- Shelter
- Culture
- Environment
- Social Equity
- Technology and Innovation

**Over 250 cities involved.**



# Student/Teacher Ratio

---

**2008:**      **Student/teacher ratio**

*World Bank, (2008), "Global City Indicators Program Report: Preliminary Final Report", April 2008.*

**2012:**      **Student/teacher ratio**

**Numerator: Number of Students**

**Denominator: Number of Teachers**

*Global City Indicators Facility: Website User Guide. October 2012.*



# 2014: Student/Teacher Ratio (STR)

---

- "The student/teacher ratio shall be expressed as the number of enrolled primary school students (numerator) divided by the number of full-time equivalent primary school classroom teachers (denominator).
- The result shall be expressed as the number of students per teacher.
- Private educational facilities shall not be included in the student/teacher ratio.
- One part-time student enrolment shall be counted as one full-time enrolment; in other words a student who attends school for half a day should be counted as a full-time enrolment.
- If a city reports full-time equivalent (FTE) enrolment (where two half day students equal one full student enrolment), this shall be noted.
- The number of classroom teachers and other instructional staff (e.g. teachers' aides, guidance counselors), shall not include administrators or other non-teaching staff.
- Kindergarten or preschool teachers and staff shall not be included.
- The number of teachers shall be counted in fifth time increments, for example, a teacher working one day per week should be counted as 0.2 teachers, and a teacher working three days per week should be counted as 0.6 teachers."

## Over 100 indicators defined and submitted to ISO

### City Services

- Education
- Finance
- Governance
- Recreation
- Transportation
- Wastewater
- Energy
- Fire and Emergency Services
- Health
- Safety
- Solid waste
- Urban Planning
- Water

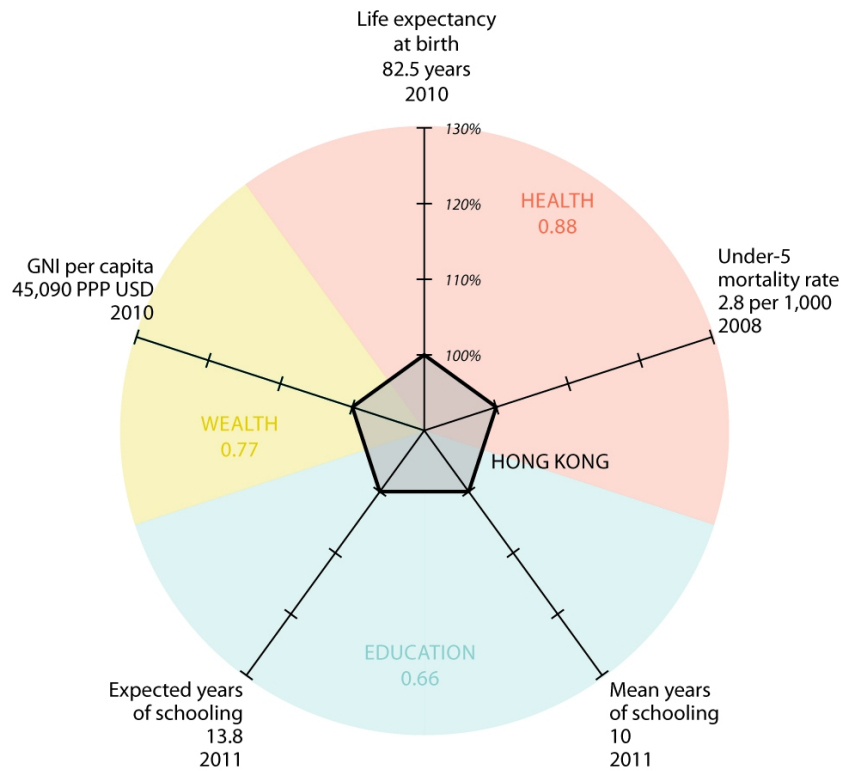
### Quality of Life

- Civic Engagement
- Economy
- Shelter
- Culture
- Environment
- Social Equity
- Technology and Innovation

***ISO 37120 – Sustainable  
Development and Resilience of  
Communities – Indicators for City  
Services and Quality of Life (under  
TC268***

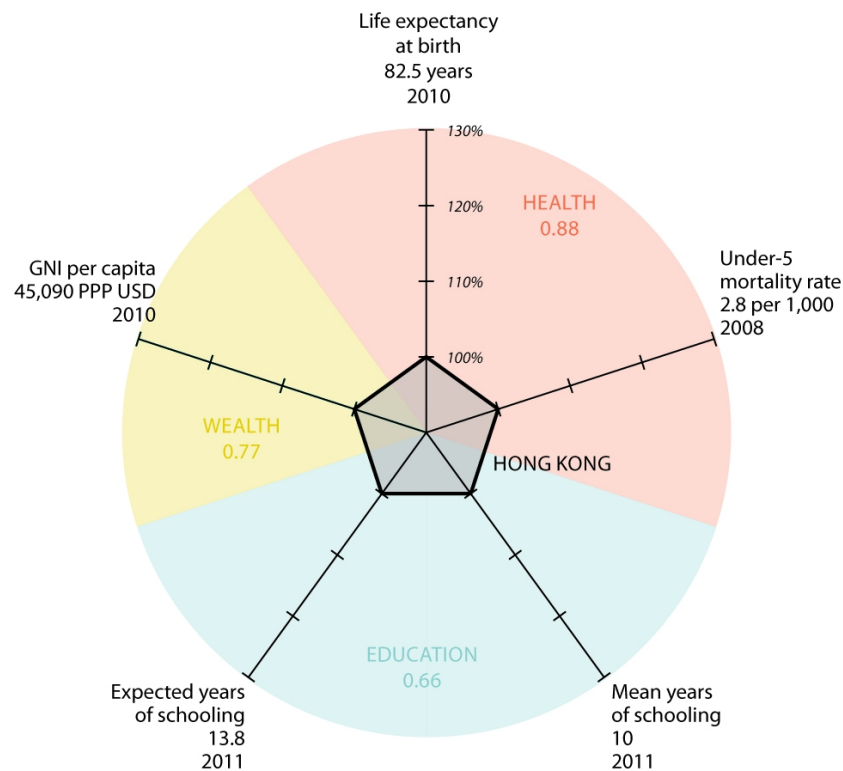


# Computer Science Vision (slightly provocative)



# Ontology Engineering Vision

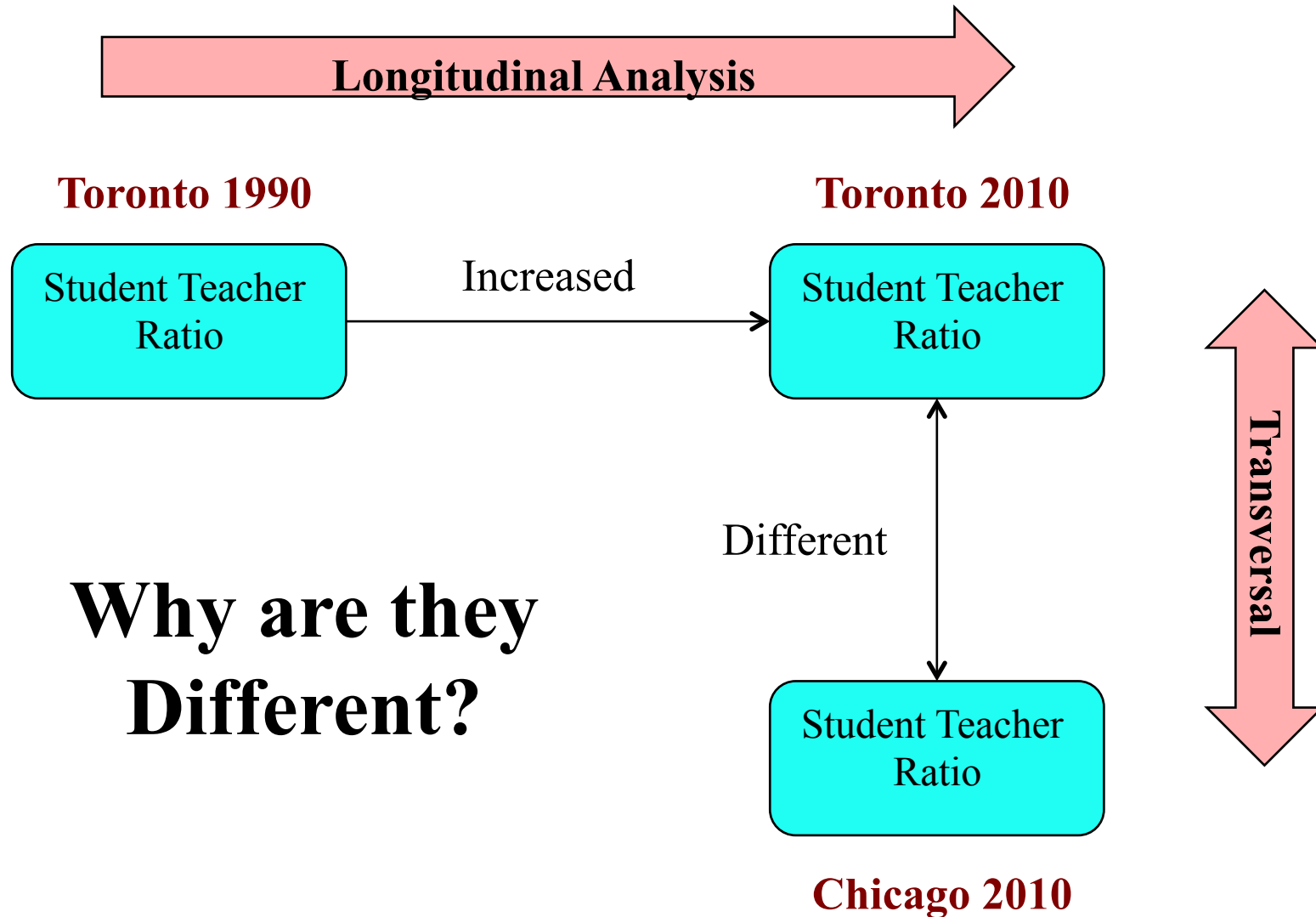
## Transition from Visualization to Analysis



- Automate the analysis of city performance
  - Perform **longitudinal** and **transversal** analyses, and
  - Determine the **root causes** of differences, using data from across the semantic web.



# Automated Analysis





# Step 1: Providing the Ontologies to Represent City Data

---

Fox, M.S., (2013), “A Foundation Ontology for Global City Indicators”, Global City Institute Working Paper #3,



# Modeling Gap

---

- What type of number is a STR?
  - Unit of measurement?  
Meters? Grams?
  - Ratio, Ordinal,  
Nominal?
  - Scale? Kilos?

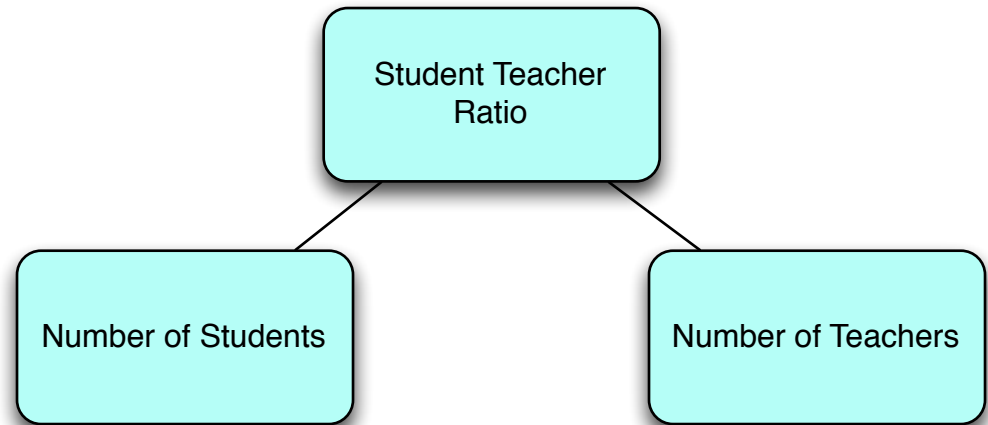
Student Teacher  
Ratio





# Modeling Gap

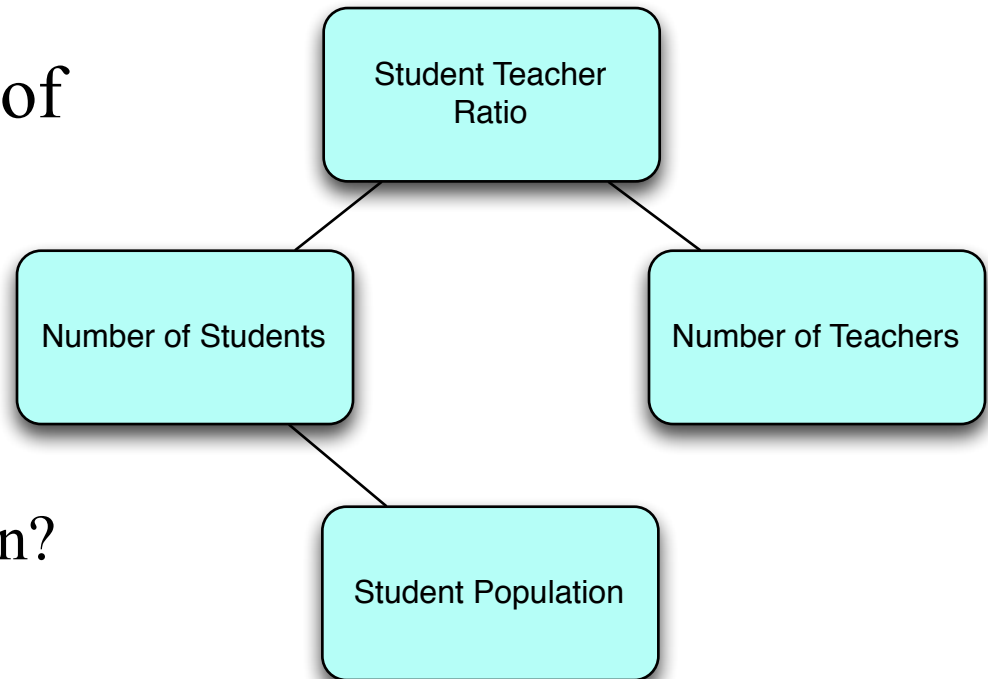
- What is an STR composed of?
  - If a “division”, what is the numerator?  
denominator?
- What kind of numbers are these?
  - Ordinal?
  - Units? Kilo?





# Modeling Gap

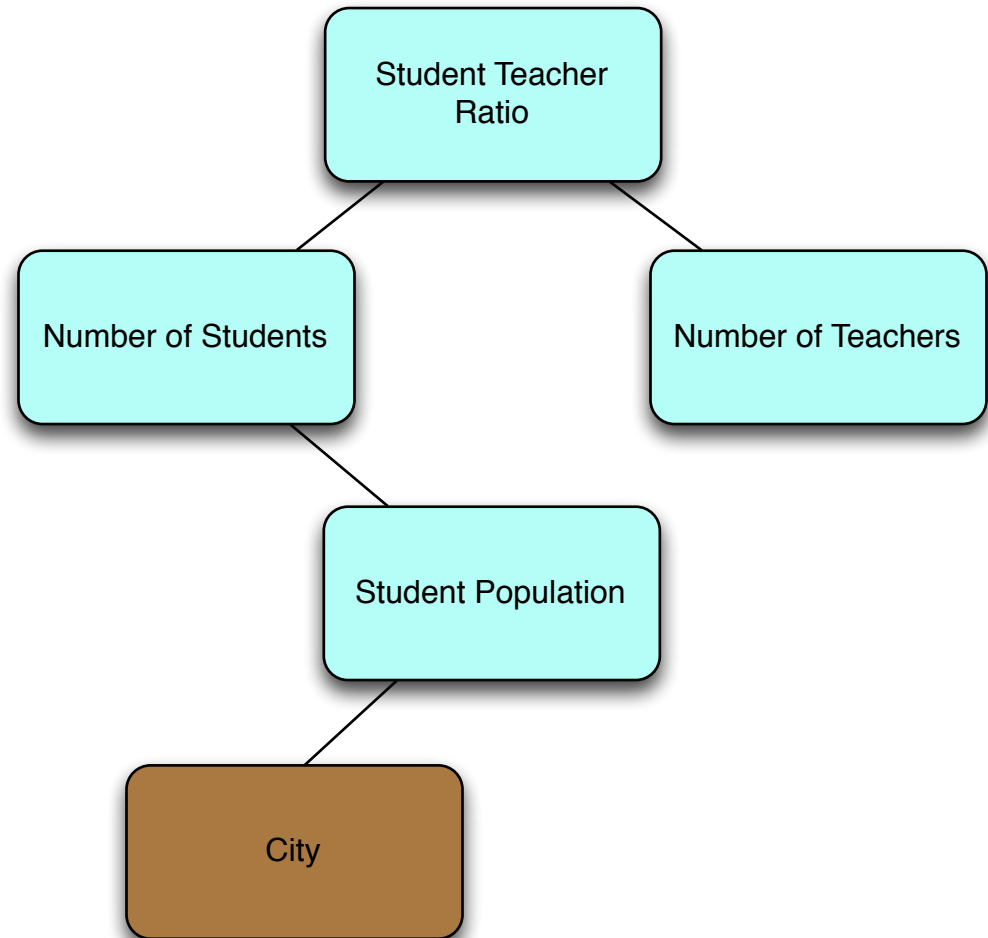
- What is the “number of students” representative of?
  - Is it a statistic? Or a property of a set?
  - What is the Population?





# Modeling Gap

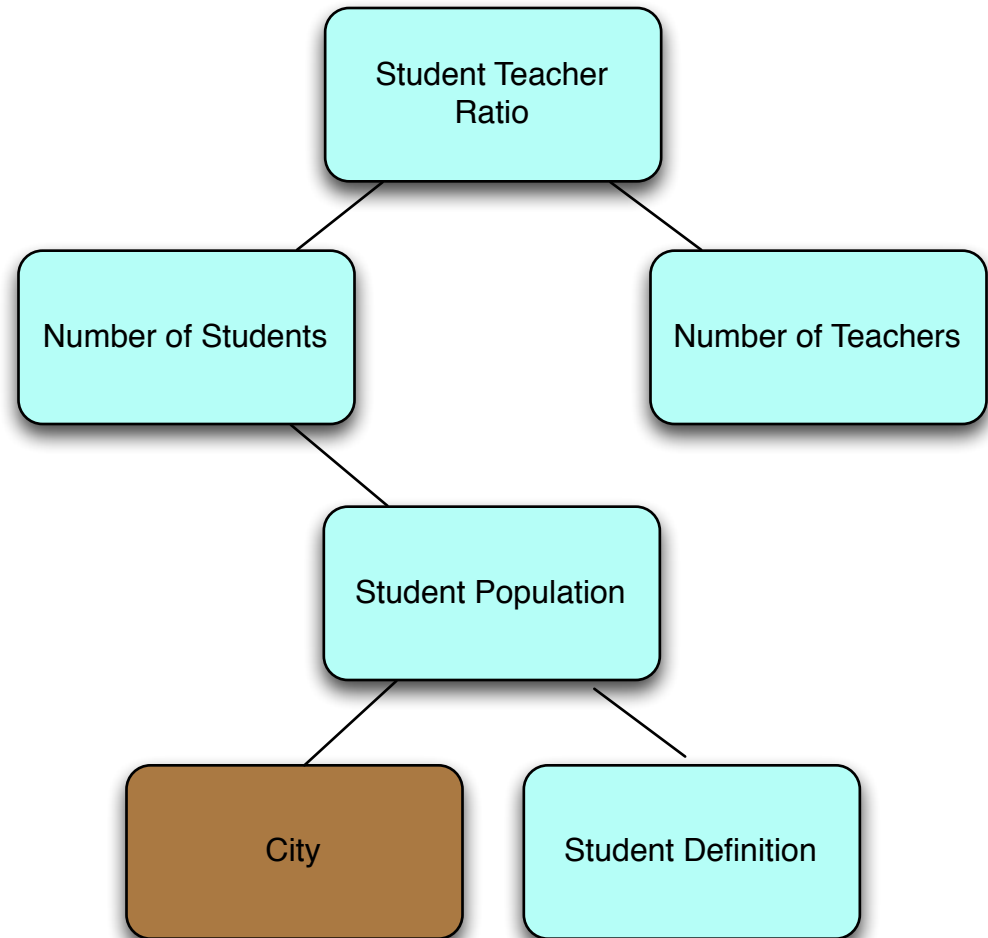
- How are members of the Population determined?
  - Where is the population drawn from?
  - Toronto ON?  
Toronto OH?





# Modeling Gap

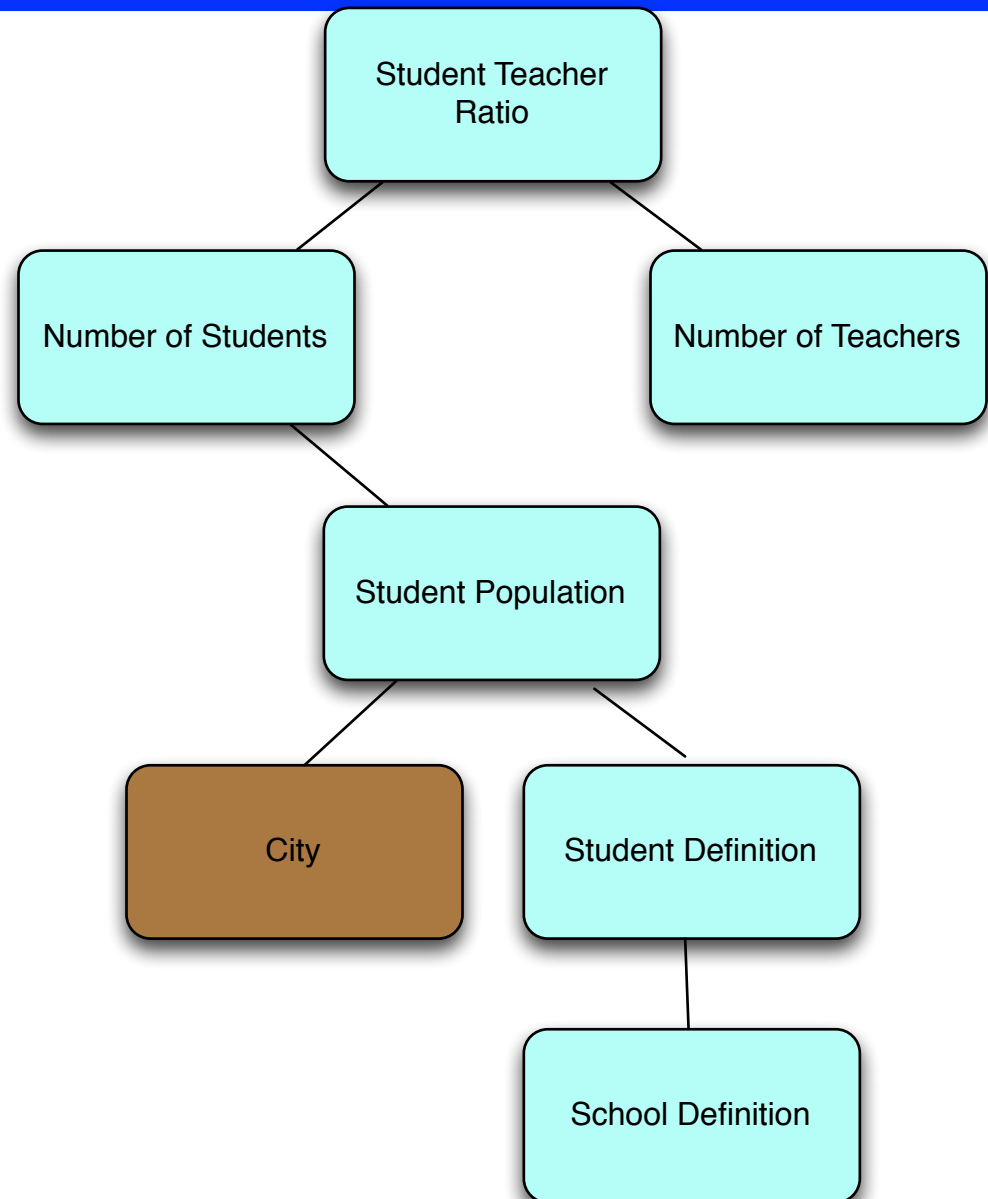
- How are members of the Population determined?
  - What is a student?
  - Full or part time?
  - Regular or special?
  - Primary or secondary grades?





# Modeling Gap

- What schools are included in defining students?
  - Public, Private?
  - What are the primary grades?



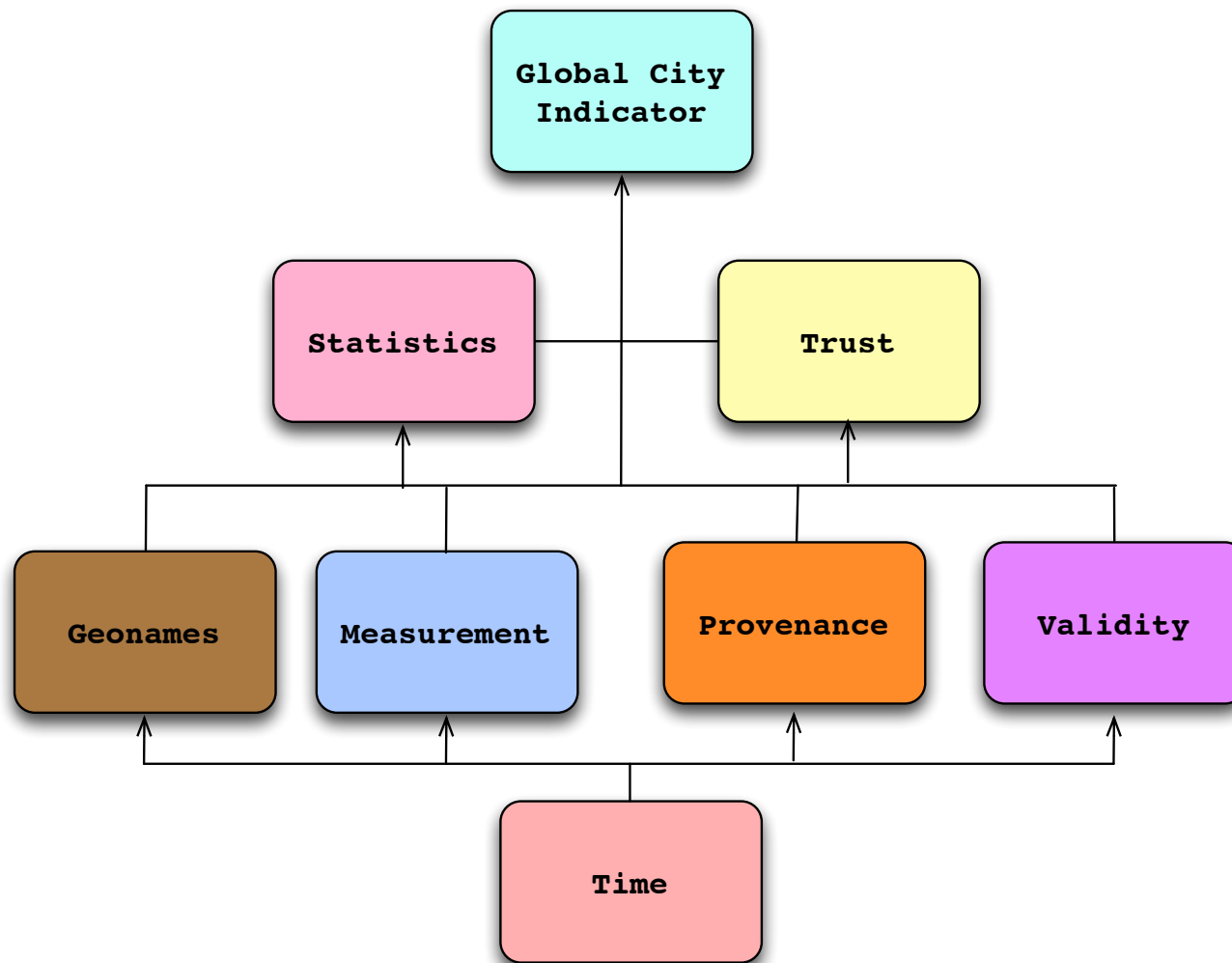


# Modelling Gap

---

- **Provenance**
  - Who, when, how
- **Validity & Belief**
  - Effective time period
  - Degree of belief
- **Trust**
  - In the city, organization, individual
  - In beliefs, performance

# Consequence



- A single indicator requires the integration of many types of ontologies.

# And Includes More Specific Ontologies



| Category   | Indicator  | 3. Placename       | 4. Measurement | 5. Statistics | 6. Provenance | 7. Time | 8. Validity | 9. Trust | 10. Govt Finance | 11. Govt Organization | 12. Economics | 13. Census | 14. Environment | 15. City Infrastructure | 16. Edu, Tech, Innov | 17. Health, Safety, Emerg. |
|------------|--|--------------------|----------------|---------------|---------------|---------|-------------|----------|------------------|-----------------------|---------------|------------|-----------------|-------------------------|----------------------|----------------------------|
|            |  | Profile Indicators |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
| Government | Type of government (e.g. Local, Regional, County)  |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Gross Operating Budget (US\$)                      |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Gross Operating Budget per capita (US\$)           |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Gross Capital Budget (US\$)                        |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Gross Capital Budget per capita (US\$)             |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
| Economy    | Average household income [US\$]                    |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Annual inflation rate (avg. of last 5 years) [%]   |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Cost of living [US\$]                              |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Income distribution [GINI Coefficient]             |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Country's GDP [US\$]                               |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Country's GDP per capita [US\$]                    |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | City Product per capita [US\$]                     |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | City product as a % of country's GDP               |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Total employment                                   |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Employment % change based on the last 5 years      |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Number of businesses per 1000 population           |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Annual avg. unemployment rate                      |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Commercial/Ind. assessment as % of total assess't. |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
| People     | Total population                                   |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | Population density (per sq. kilometer)             |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | % of country's population                          |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |
|            | % of population that are children (0-14)           |                    |                |               |               |         |             |          |                  |                       |               |            |                 |                         |                      |                            |





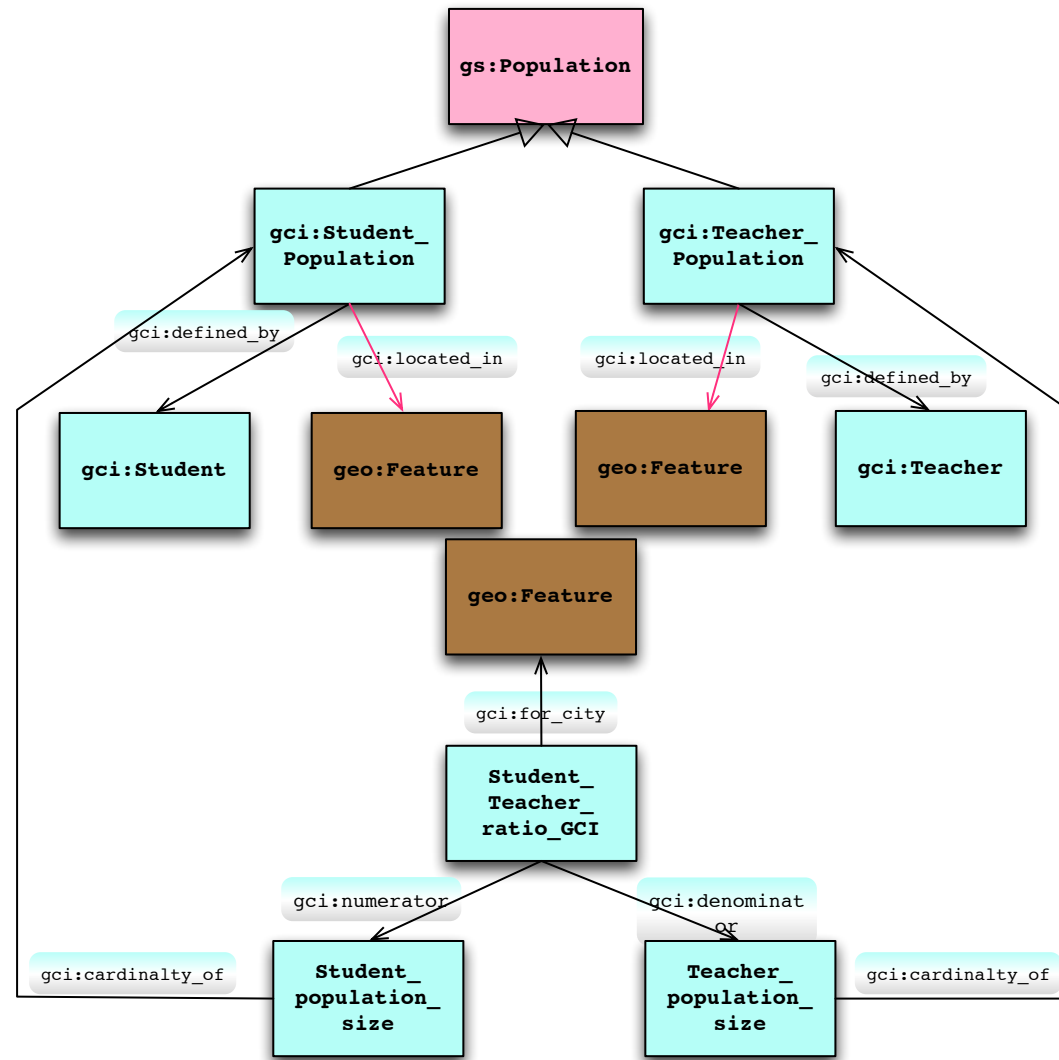
## **Step 2: Determine Consistency of Merged Data Using Bridge Axioms**

---

With the integration of information from multiple sources, we need to guarantee that instances are consistent with their definitions and with other instances.

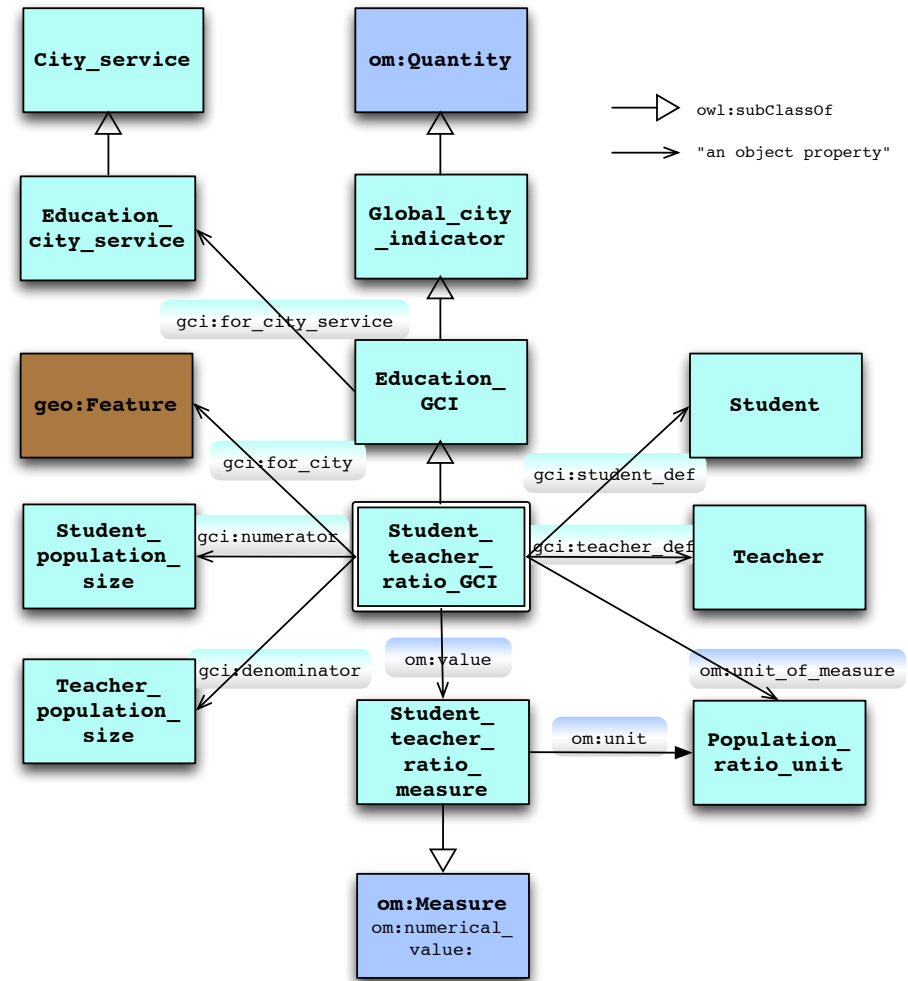
# Placename Rules

- **Rule G1:** The city for the STR being measured is the same as the cities where its numerator and denominator are measured.



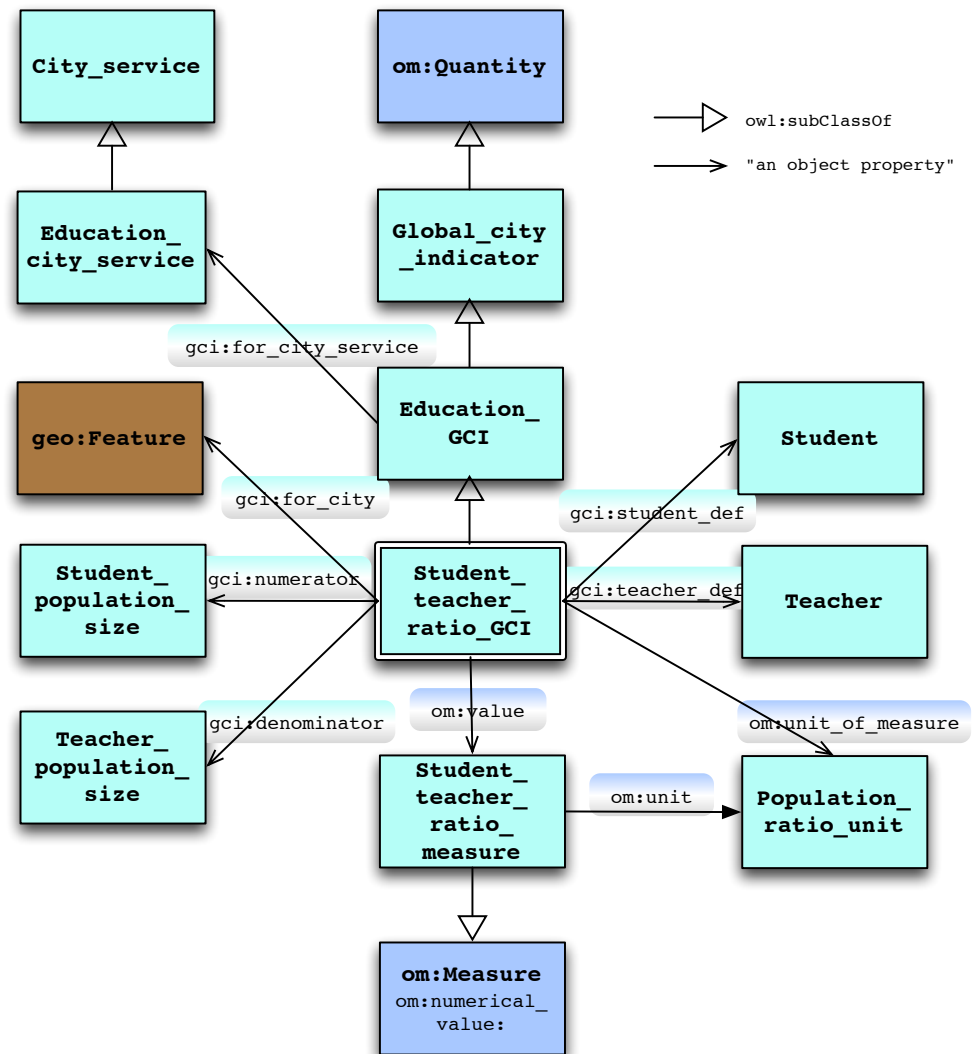
# Measurement Rules

- **Rule M1:** The numerator and denominator of a `gci:Student_teacher_ratio_GCI` are the correct type (as specified by the GCI).
- **Rule M2:** The numerator and denominator of the `gci:Student_teacher_ratio_GCI` are consistent with the numerator and denominator of its unit of measure.
- **Rule M3:** If the numerator and denominator of a `gci:Student_teacher_ratio_GCI` are the same type, then they should have the same units (scale).
- **Rule M4:** The units of the actual measurement are the same as defined by GCI it is a measure of.
- **Rule M5:** The value of the `gci:Student_teacher_ratio_measure` is equal to the value of the `gci:Student_teacher_ratio_GCI` numerator divided by the denominator.



# Population Rules

- Rule S1:** The definitions of student and teacher for the `gci:Student_teacher_ratio_GCI` are the same as used by its numerator and denominator.





# Meta Information Rules

---

- **Validity**
  - **Rule V1:** The effective time period for which an indicator is valid is contained within the effective time periods of its numerator and denominator.
  - **Rule V2:** The effective period for an indicator is after the time the indicator was generated.
- **Provenance**
  - **Rule P1:** If two versions of the same indicator exist, then they are inconsistent with each other if different methods were used to generate them.
- **Trust**
  - **Rule T1:** The trustee in a trust relationship is the same as the `pr:wasAttributedTo` Agent for an indicator.
  - **Rule T2:** The trusted certainty degree of an indicator is less than or equal to the indicator's certainty assigned by its creator.



# Step 3: Analysing the Data

---



# Conclusion

---

- The automated analysis of city indicators requires a high degree of fidelity.
  - *Fidelity refers to the degree to which a model reproduces the state of a real world object, feature or condition. Fidelity is therefore a measure of the realism of a model.*
- Fidelity requires a semantically rich core of both foundational and applied ontologies.
- Sadly, most ontologies are simply vocabularies with limited definitions, hence limiting their value.



# References

---

- Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., and Zednik, S., (2012), “PROV Model Primer”, <http://www.w3.org/TR/prov-primer>.
- Caragliu, A; Del Bo, C. & Nijkamp, P (2009). "[Smart cities in Europe](#)". *Serie Research Memoranda 0048* (VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics).
- Fox, M.S., (2013), “A Foundation Ontology for Global City Indicators”, Global City Institute Working Paper #3, [http://media.wix.com/ugd/672989\\_bfb12b835c845d2b3773f49d2a8aa308.pdf](http://media.wix.com/ugd/672989_bfb12b835c845d2b3773f49d2a8aa308.pdf)
- Fox, M.S., and Huang, J., (2005), “Knowledge Provenance in Enterprise Information”, *International Journal of Production Research*, Vol. 43, No. 20., pp. 4471-4492.
- Hoornweg, D., et al., (2006), “City Indicators: Now to Nanjing”, Third World Urban Forum, Vancouver.
- Huang, J., and Fox, M.S, (2006), "An Ontology of Trust – Formal Semantics and Transitivity," *Proceedings of the International Conference on Electronic Commerce*, pp. 259-270.
- Hobbs, J.R., and Pan, F., (2006), “Time Ontology in OWL”, <http://www.w3.org/TR/owl-time>.
- Pattuelli, M.C., (2003), “The GovStat Ontology: Technical Report”. The GovStat Project, Integration Design Laboratory, School of Information and Library Science, University of North Carolina at Chapel Hill, <http://ils.unc.edu/govstat/papers/govstatontology.doc>.
- Rijgersberg, H., Wigham, M., and Top, J.L., (2011), “How Semantics can Improve Engineering Processes: A Case of Units of Measure and Quantities”, *Advanced Engineering Informatics*, Vol. 25, pp. 276-287.