Research in Computer Science and IT: How and Where Are We (Going)?

A Chilean Perspective

Leopoldo Bertossi[⋆]
Carleton University
Ottawa, Canada

[⋆]: Faculty Fellow IBM CAS

Contents:

A. The research community: here and there

B. At the light of my own research

C. Getting there

A. The Research Community

The Kind of Research

- A personal assessment of international and national research community and activity in CS and IT

- Based on a bit more than 20 years of research in CS (roughly 10 in Chile, 10 in Canada)

- With a previous incarnation as a mathematician, for several years, including my PhD studies

- For more information and details:

  Bertossi, L. "En 20 Años de Computación: Una Visión muy Personal".
  Revista Bits de Ciencia, N. 5, 2011. Universidad de Chile.

  `http://people.scs.carleton.ca/~bertossi/papers/BITSREV5LeoBertossi.pdf`

- Perspectives from both the classical science and computer science/technology communities

Practice and dissemination of research in CS/IT:

- Researchers tend to collaborate, few with single authors
- Papers have several authors, but not necessarily with members of a same local team
- International collaboration is common and appreciated
- Conferences are crucial for exchange of ideas and fast communication of results
- More published in conference proceedings than journals
- Papers accepted at top conferences can be more prestigious than journal papers
- Each area has a few of those high-impact conferences

Conference vs. Journal:

- Much controversy in these days (cf. *Comm. ACM*)
- Too many conferences, WSs; many narrow and low quality
- Good researchers are overloaded with paper reviewing
- Little time to do a proper review
- Technical/mathematical details seldom checked
- Experiments suggest "applicability", not support claims
- Experiments are not replicable
- Sloppy writing, no clear problems/contributions
- Research becomes conference/deadline driven

Other “problems” in our area:

- Most problems around conference publications persist beyond them
- (Multidisciplinary) research may be just applied CS/IT
- Little awareness of other authors’ results
- …

General view and conclusions:

- Very active and effervescent community
- Possibility of high impact
- Relatively short transition from research to products, applications, teaching
- Potential for multidisciplinary research and applications
- Many opportunities for collaboration
- A lot of fun!
- Research and publication standards leave to be desired
- Very easy to become and stay as a mediocre “researcher”, still “publishing”, but without any impact

Should we care?

Dangerous for countries like Chile!

- The basis of the area is still underdeveloped, shaky and fragile
- Relatively young discipline (first academic departments of CS/IT created in early 80s)
- The creation of critical mass has progressed very slowly
- Very easy to imitate and adopt the "bad" practices
- The overall environment, as a country, increases exposure those risks and vulnerability

Exaggeration? In Chile:

- Classic scientific disciplines have experienced much stronger growth in the same period (math, physics, ...)
- CS/IT is not even close to the level achieved by them
- In the late 80s and early 90s many fresh PhDs in CS returned, from the best universities

  A tiny fraction of them are still active and internationally recognized researchers
- Many faculty members publish only in local/regional venues
- Faculty assessment criteria usually less strict than in classic scientific disciplines
- Very little R&D in industry

- There has been no state policy, program, or funding agency specifically supporting the area

  Peculiarities of the area have not been considered or understood

- Difficult to attract good students to academy and research
  - Academic salaries in the area are rather low, in absolute and relative terms
  - Study of CS/IT only through an Engineering degree (no Bachelor of CS)

Let's face these (and other) problems as challenges!

More on the bright side:

- There is a good basis for growth
- There are cases encouraging optimism

Some advantages and strengths:

- Very good high schools and universities
- Very good students: solid background, hard-working, bright, ...
- Country well-positioned in Latin America (LA)

  Attractive for good regional graduate students (potentially)
- State scholarship program
- National research funding agency with continuity and several strengths
- Some (but still few) universities offer competitive salaries
- Chileans tend to return to Chile after the PhD

- Main universities appreciate research activity on site

  Impact on resources, teaching load, hiring, promotion, etc. (with some caveats though, cf. later)

- In CS/IT research Chile is top in LA

  In terms on impact, recognition, quantity/quality of publications and researchers

- World-renowned researchers can be found in some areas: data management, web research, algorithms, ...

  In data management and semantic web research Chile plays in the world champions league

Fortunately, I contributed to this success ...

How has been success in data management and semantic web research possible in Chile?

B. A Trajectory and a Certain Kind of Research

In 10 years since the early 90's:

- Started with mathematical training and background in mathematical logic

Strengths that could be appropriately exploited, make a difference and a contribution

- Relevant/crucial for several areas of CS, particularly
  - Data Management
  - Knowledge Representation in AI

Some steps:

- Right after PhD in math, I went for a postdoc in CS

- Contacted top researchers in both DM and KR, seeking collaboration

I could help them with my background, and I did

- Since early in the CS career: tried publishing at the highest level, with top researchers

And I could become involved hands-on in my new areas ...

- Taught a second/third-year course on “logic for computer science”, taken mandatorily rather early in the studies

Taught in an WS style, I could identified the best and more theoretically-oriented students

- Invited them to research group, as peers, all learning from each other

- Exposed students to research, new material, discussions, local conferences, and invited speakers

- We all targeted relevant publication venues and conferences

- Stayed close to local industry, not much technical joint activity, but contacts were valuable in many regards, until now ...

- Started formal international collaboration with relevant researchers abroad

Funding from Fondecyt, and the Conicyt's Binational Programs (USA, Germany, France, ...)

- Visitors:
  - Did research in Chile
  - Gave talks
  - Interacted technically/socially/individually with the students
  - Participated in thesis committees, even at masters' level (forcing students to write and present in English)

- Spent my summer vacations doing research in Northern hemisphere

- Took sabbaticals abroad (not very common ...)

- Students were invited to do the masters with the group

- Encouraged to submit masters' research to international conferences, and apply for PhD studies to excellent places abroad

- Most of my students came back, to academia

Applying the same (but improved) recipe with their students, with a multiplicative effect

- Now I have academic grandchildren who are internationally recognized researchers, and they are or coming back to Chile

- We concentrated on fundamental research in data management and knowledge representation

- Not much technological infrastructure needed

Just scientific background, brain power, and a scientific attitude

A certain kind of research:

I. Consistent query answering

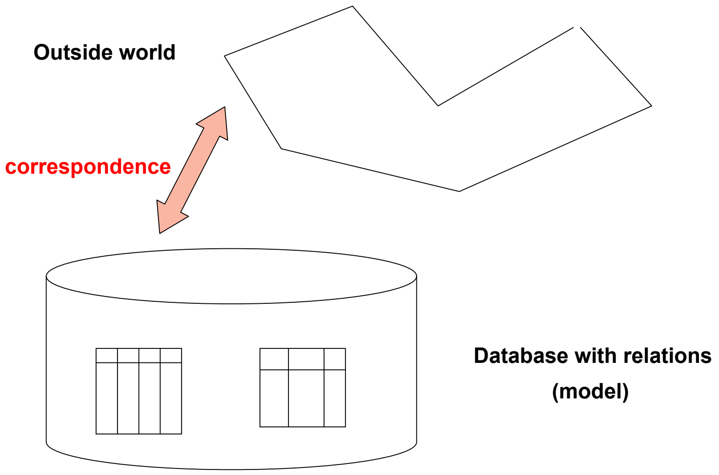
II. Contexts and applications

Of a fundamental nature, but with potential practical impact

It has had impact on the DB research community

I has been used and adopted by more applied researchers

I. Consistent Query Answering



A database instance D is a model of an outside reality

An integrity constraint on D is a condition that D is expected to satisfy in order to capture the semantics of the application domain

A set IC of integrity constraints (ICs) helps maintain the correspondence between D and that reality

A database may not satisfy a given set of integrity constraints

What is the consistent data in an inconsistent database?

What are the consistent answers to a query posed to an inconsistent database?

We may want to retrieve consistent data? But what is that?

How good will be our algorithms?

A precise characterization (mathematical definition) was needed

We provided one (Arenas,Bertossi,Chomicki; PODS99)

Intuitively, the consistent data in an inconsistent database D is invariant under all minimal ways of restoring D's consistency

That is, consistent data persists across all the minimally repaired versions of the original instance: the repairs of D

Example: Instance D violates FD: $Name \rightarrow Salary$

$Employee$	$Name$	$Salary$
	$page$	5K
	$page$	8K
	$smith$	3K
	$stowe$	7K

Two minimal repairs if only deletions/insertions of whole tuples are allowed: D_1, resp. D_2

$Employee$	$Name$	$Salary$
	$page$	5K
	$smith$	3K
	$stowe$	7K

$Employee$	$Name$	$Salary$
	$page$	8K
	$smith$	3K
	$stowe$	7K

$(stowe, \text{7K})$ persists in all repairs: it is consistent information

$(page, \text{8K})$ does not; actually it participates in violation of FD

A consistent answer to a query $\mathcal{Q}$ from a database D is an answer that can be obtained as a usual answer to $\mathcal{Q}$ from every possible repair of D wrt IC (a given set of ICs)

- $\mathcal{Q}_1 : Employee(x, y)?$

  Consistent answers: $(smith, \text{3K}), (stowe, \text{7K})$

- $\mathcal{Q}_2 : \exists y Employee(x, y)?$

  Consistent answers: $(page), (smith), (stowe)$

CQA may be different from classical data cleaning!

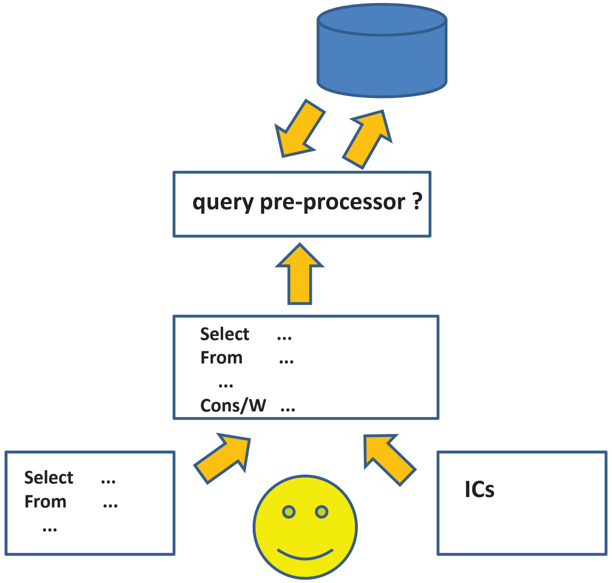
However, CQA is relevant for data quality; an increasing need in business intelligence

It also provides concepts and techniques for data cleaning

A whole area of research was opened!

Next DBMSs should provide more flexible, powerful, and user friendlier mechanisms for dealing with semantic constraints

In particular, allowing and answering queries requesting consistent data



Why not an enhanced SQL?

```
SELECT      Name, Salary
FROM        Employee
CONS/W      FD: Name -> Salary;
```

(FD not maintained by the DBMS)

Paradigm shift: ICs are constraints on query answers, not on database states!

II. Contexts and Data Quality

A hospital table with data about the temperatures of patients

TempNoon

	Patient	Value	Time	Date
1	Tom Waits	38.5	11:45	Sep/5
2	Tom Waits	38.2	12:10	Sep/5
3	Tom Waits	38.1	11:50	Sep/6
4	Tom Waits	38.0	12:15	Sep/6
5	Tom Waits	37.9	12:15	Sep/7

Is this quality data?

If not, anything to clean? What?

We don't know It depends ...

Table is supposed to contain *temperature measurements for Tom taken at noon by a certified nurse with an oral thermometer*

Is this quality data? We still don't know ...

Maybe we can say something about the time

Maybe good enough for time to be “around noon” (meaning?)

Questions about the quality of this data make sense in a broader setting

The quality of the data depends on “the context”

A context that allows us to:

- provide meaning and semantics (disambiguation)
- make sense of the data
- assess data quality
- support data cleaning

Etc.

Contexts So Far: We find “contexts” in several places in CS: databases, semantic web, KR, mobile applications, ...

Usually used for “*context aware* ... search, databases, applications, devices, ...”

Most of the time no explicit notion of context, but mechanisms that take into account/computation some contextual notions

Usually, time and geographic location, i.e. particular *dimensions*

In our opinion, there is a lack of fundamental research in the area, specially for data management

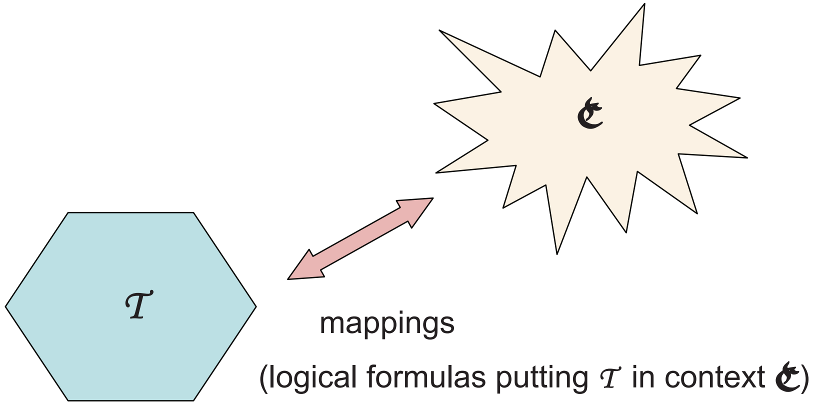
Precise and formalized notions of context are rather absent

Contexts that can be implemented/used in a principled manner in data management

A Vision for Contexts:

A general notion and theory of context have still to be developed

We envision it as follows:


$\mathcal{T}$
$\mathfrak{C}$
mappings
(logical formulas putting $\mathcal{T}$ in context $\mathfrak{C}$)

- A logical theory $\mathcal{T}$ is the one that has to be "put in context"

  For example, a relational database can be seen as a theory

- The context is another logical theory, $\mathcal{C}$

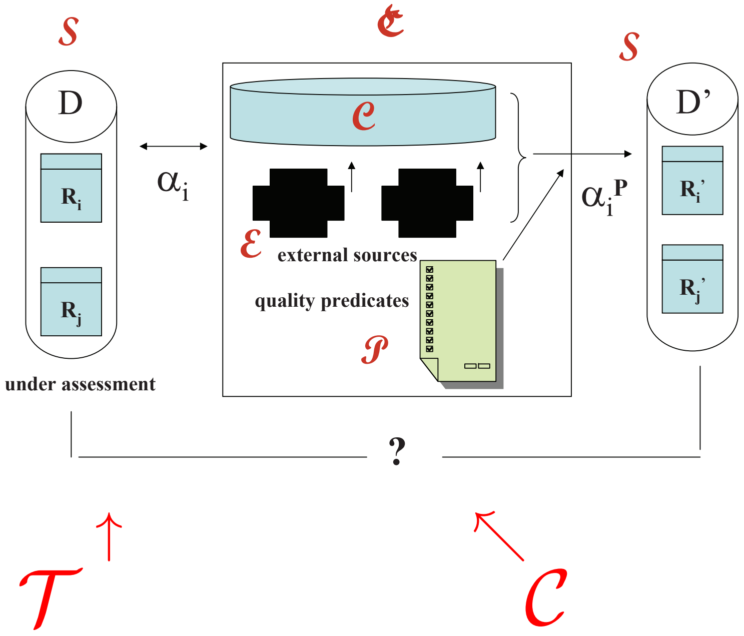
  $\mathcal{T}$ and $\mathcal{C}$ may share some predicate symbols

- Actually, the connection between $\mathcal{T}$ and $\mathcal{C}$ is established through: connection predicates and mappings

Long-term research, with many open problems!

In particular, for applications in data management (WS VLDB'10)

In our data quality scenario, the database under assessment is mapped into the context, for further data quality analysis, imposition of quality requirements, and cleaning



The context can, e.g. capture and narrow down semantics

- Defining in $\mathcal{C}$ predicates that are used in $\mathcal{T}$ (e.g. "time close to noon")

- Contributing in $\mathcal{C}$ with additional constraints for predicates used in $\mathcal{T}$, e.g. integrity constraints for table TempNoon

- Dimensions and points of view for analysis and understanding of $\mathcal{T}$'s knowledge
- Specifying and using notions of relevance
- Explanation, diagnosis, causality
- Capturing commonsense assumptions and practices

C. What Can We Do?

Natural questions:

- Any lessons that we can learn and conclusions we can draw?
- Are there ways in which we can improve?
- What is crucial for change in the right direction?
- Lessons from success in data management and semantic web research in Chile?

Have contributed to the latter and been a close witness of it

Have seen a lot, including many mistakes and missed opportunities

Possible to make some remarks and offer some advice (thinking of Chilean reality)

As a PhD applicant:

- Choose carefully your:
  - Field
  - University: Language should not be a consideration when choosing a place
  - Supervisor: crucial in many aspects:

    knowledge, source of research problems, mentoring, experience, recognition, visibility, contacts, recommendations, ...

As a PhD student:

- Read a lot; broaden basis and perspective

  Stay informed about publications in main conferences and journals

- Be independent researcher, with *your own* research problem (as opposed to a group's problem)

- See always the broad picture in your area

- Try to submit to the most prestigious and high-impact conferences and publication venues

- Stay in direct research relationship with your supervisor

- Avoid:
  - becoming a lower layer in a pyramid
  - being supervised by more advanced grad students
  - becoming a small part of a huge research project (machine)
  - being just an implementer

As a young, post PhD, researcher:

- Go for a postdoc!
- Start with a tenure-track mentality and attitude (even if there is not tenure system)
- Write as many papers as possible based on PhD thesis work, including journal papers
- Do not stick exclusively and forever to your PhD theme or supervisor, diversify

  Explore also other subjects and collaborators
- Time right after the PhD may be the most difficult

  People expect a lot and you feel the same as before
- Your priority is to reestablish your research program (or design one) in new location

- Avoid heavy and complex administrative work

- Not many resources (equipment) are necessary for doing successful research

  Fundamental research is fine, necessary, visible, has impact

  In Chile we have the right material ...

  And in CS/IT applications are never too far

- Participate in competitions for grants and win them

  Not only for funding, but also recognition by peers and prestige

- Seek mentorship and stay connected to international community

  Apply to international collaboration programs, bring *relevant researchers* to Chile, visit them, publish with them

- Target high-impact conferences and journals
- In Chile ISI publications are crucial
  - There are low-quality/irrelevant journals that are ISI
  - There are high-impact publication venues (e.g. proceedings of top conferences) that are not ISI

  Thus, target both relevant ISI journals and high-visibility conferences (not incompatible goals)

Conclusions

- A successful research program in CS/IT is possible in Chile

With international impact, recognition and projection

- It is possible to identify factors of success

And identify and avoid sources of mistakes

- It is easy to get the wrong messages from the international community

Identify the relevant players and research

As a researcher stay focused, on track, and relevant

- There is human capital in Chile of the best quality

Attract and nurture the best students