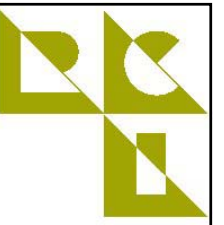
Resources Computing International Ltd

Incomplete and missing data in geoscience databases

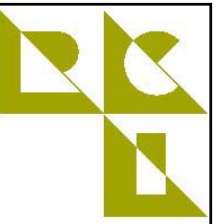
Towards the OWA relational model ?

Stephen Henley



Not just geoscience

- The title says this is about geoscience – but the conclusions are much more widely applicable



Geoscience data

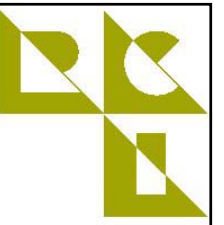
- Very commonly may be
  - Imprecise
  - Incomplete
  - Missing

Typical imprecise data

Sample	SiO2 %	Cu ppm
#101	53.5	128
#102	49.2	185
#103	66.3	163

Typical imprecise data

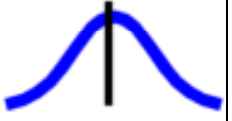
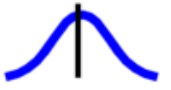
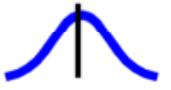
Sample	SiO2 %	Cu ppm
#101	53.5	128
#102	**49.2**	185
#103	66.3	163



What is “49.2% SiO2” ?

- A recorded value from a laboratory
- Imprecise: the true value could be 49.2%, 49.21% or 48.55%?
- because of **instrumental errors**
- and because of **sampling errors**
- The full data item should include “49.2” AND data about the error distribution

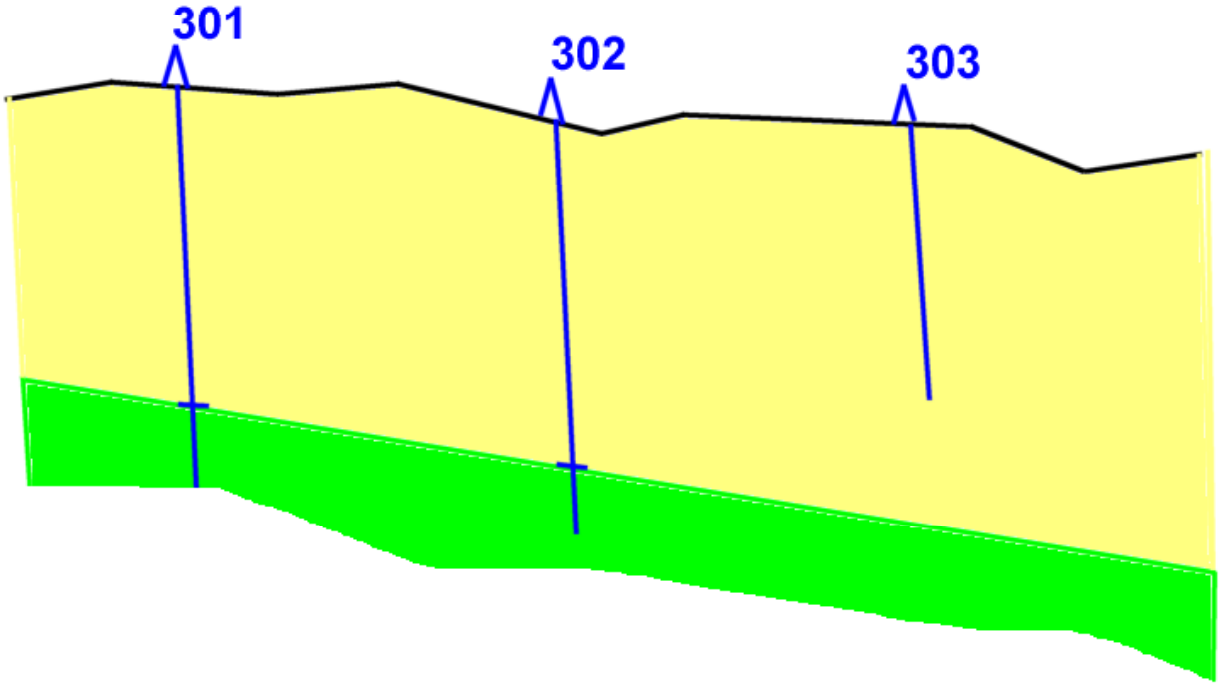
Each value has its own error distribution

Sample	SiO2 %	Cu ppm
#101	~53.5 	~128 
#102	**~49.2** 	~185 
#103	~66.3 	~163 

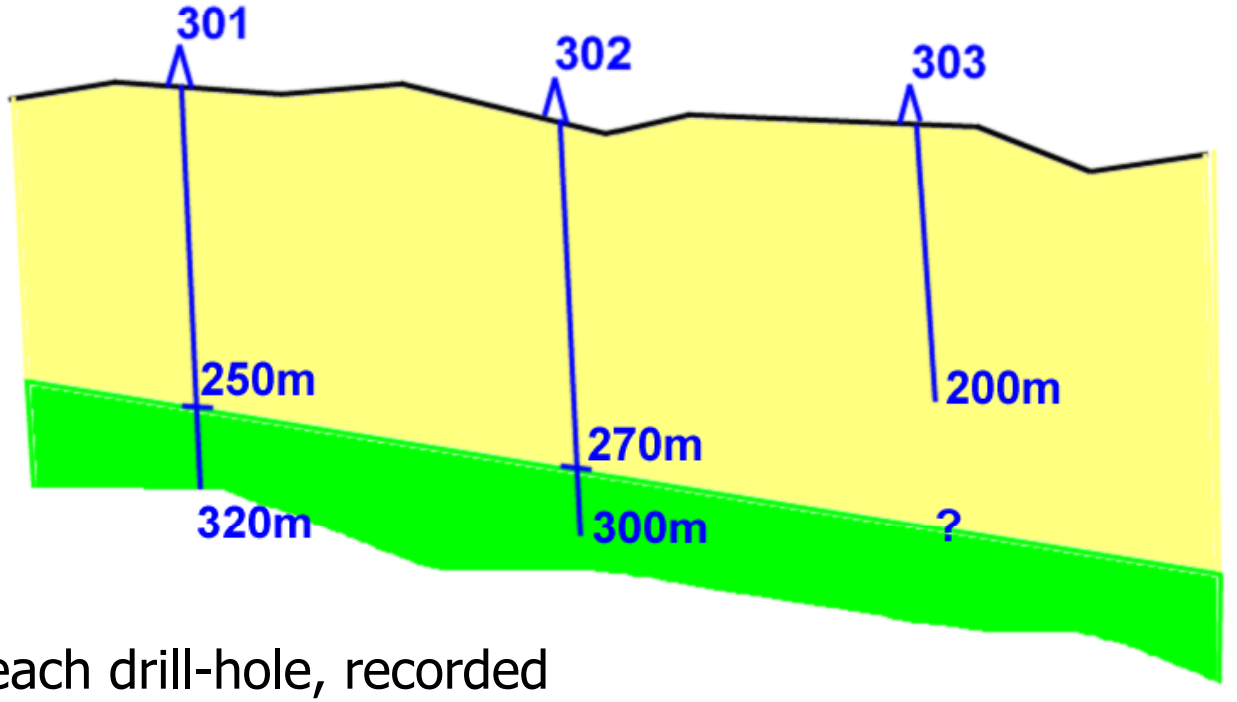
What about queries ?

- Given the SiO2 value of "~ 49.2"
  - Query **"WHERE SiO2 >50 …"**
  - Not **TRUE** or **FALSE** but **P = 0.317** (for example)
- So the simple **2VL does not apply** – instead a continuous scale of probability estimates from P=0 (FALSE) to P=1 (TRUE)
- Related to 'fuzzy logic' – but let's not go there today !

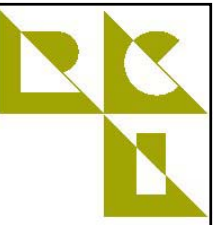
Incomplete data



Incomplete data

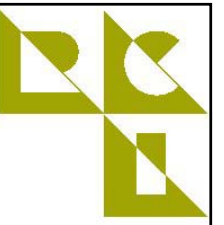


For each drill-hole, recorded
Total depth and depth to 'green' rock unit



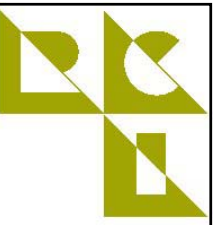
Incomplete data

Hole_ID	Total_Depth	D_green
#301	320.0	250.0
#302	300.0	270.0
#303	200.0	**Unknown ?**



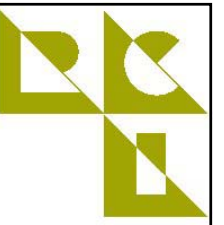
Incomplete data

Hole_ID	Total_Depth	D_green
#301	320.0	250.0
#302	300.0	270.0
#303	200.0	**> 200.0**



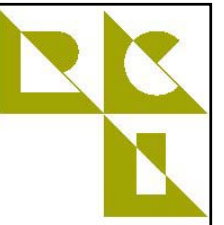
Incomplete data

- This value ">200.0" is semi-quantitative
- (another similar example – "below detection limit" in chemical analysis data – e.g. "< 5 ppm")
- It is not a NULL so Chris Date ought to be quite happy about it
- BUT queries will not always give unambiguous TRUE or FALSE



Querying a tuple with D_green value ">200"

- WHERE D_green > 150 … **TRUE**
- WHERE D_green < 100 … **FALSE**
- WHERE D_green > 250 … **UNKNOWN**
- WHERE D_Green <250 … **UNKNOWN**
- **So 2VL (true/false) is inadequate here also**



Missing data

- Very often there are genuine gaps in data sets, for many possible reasons
  - Samples not collected
  - Observations not taken
  - Instrumental malfunction
  - . . . 1001 other possible reasons
- These gaps may be **single data items** or **whole rows (tuples)**

Missing data item

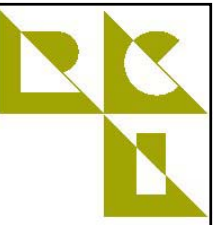
Sample	SiO2 %	Cu ppm
#101	53.5	128
#102	-	185
#103	66.3	163

Missing data item

- We know sample #102 must have a SiO2 value – we just don't know what it is
- So this value is just "missing". It's not "inapplicable" (which might justify re-designing the database)
- If we use Chris Date's 'CWA relational' model then we are not allowed 'NULL' so how do we represent this ?

CWA: Avoiding NULL

- Several proposed methods to get around the prohibition of NULL, including –
  - Default-value solutions (Chris Date)
  - Other suggestions (Hugh Darwen and Fabian Pascal)



The default-value 'solution' as proposed by Date

- Instead of a global 'null'
- A default value defined separately for each domain
- If a legitimate value for the domain, how are missing values distinguished from actual values ?
- If not a legitimate value for the domain, it's just another sort of 'null' – no better, but more complicated, so in fact worse

Proposals by Darwen and Pascal

- Different in detail, but both involve decomposition to ‘hide’ the missingness of data values

Decompose into 'null-free' relations

Sample	SiO2%	Cu ppm
#101	53.5	128
#102	-	185
#103	66.3	163

Sample	SiO2 %
#101	53.5
#103	66.3

Sample	Cu ppm
#101	128
#102	185
#103	163

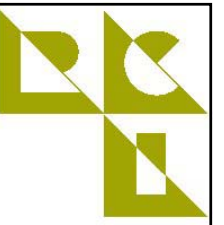
In this way …

- We certainly get rid of the 'NULL'
- Any missing data item is expressed instead as a **missing tuple** in a binary relation

The CWA states that …

- where **r** is any relation and **t** is **any possible** tuple that conforms to the heading of **r** :-
- If **t** appears in the body of **r**, then it is a true instantiation of the predicate (i.e. the corresponding proposition is considered to be true);
- conversely, if **t** does not appear in the body of **r**, then it is a false instantiation (i.e. the corresponding proposition is considered to be false)

– **Date & Darwen 1998, 2000, …**



…. SO

- Under the CWA, any tuple that is legitimate but is missing is assumed to represent a FALSE proposition.
- So what about our decomposed 'null-free' relations ? …

No tuple for sample #102 in the SiO2 relation

Sample	SiO2 %
#101	53.5
#103	66.3

Sample	Cu ppm
#101	128
#102	185
#103	163

Under the CWA …

- There are infinitely many possible legitimate tuples for sample #102: for example

Sample	SiO2%
#102	51.2

or

Sample	SiO2%
#102	45.5

- **But NONE of them is included**
- So **ALL** are interpreted as **FALSE** propositions
- This means that **under the CWA interpretation,** for sample #102 **there is NO acceptable value** of SiO2 **–** it does not mean that the value is merely unknown.

This implies that …

- If we have any missing data, then the CWA is not appropriate.
- Does this mean we can't use the relational model for geoscience data ?
- Of course not. Just that the narrow 'CWA' version of relational, defined by Date, Darwen, & Pascal, is inadequate
- - but is that really the only game in town ?

Codd **was** right

- We need to revert to the "true" relational model as defined by Codd – which ALLOWS for the reality, that there will always be missing and incomplete data
- Codd's 1979 RM/T paper and his 1990 book leave many unanswered questions – but they do allow us to use the open world assumption
- This does not restrict us to 2VL but uses a 3VL - including truth value UNKNOWN.

So let's take a look at the truth tables

- 2VL – two valued logic for CWA
- Then extended to allow for probabilities
- Then 3VL as needed by OWA

CWA - 2VL

NOT	
T	F
F	T

AND	T	F
T	T	F
F	F	F

OR	T	F
T	T	T
F	T	F

T represents TRUE

F represents FALSE

2VL with probabilities

NOT	
T	F
p	1-p
F	T

AND	T	p(A)	F
T	T	p(A)	F
p(B)	p(B)	p(A∩B)	F
F	F	F	F

OR	T	p(A)	F
T	T	T	T
p(B)	T	p(A∪B)	p(B)
F	T	p(A)	F

T represents **p=1;** **F** represents **p=0**

p(A∩B), **p(A∪B)** in general need statistical computation

OWA - 3VL

NOT	
T	F
U	U
F	T

AND	T	U	F
T	T	U	F
U	U	U	F
F	F	F	F

OR	T	U	F
T	T	T	T
U	T	U	U
F	T	U	F

T represents TRUE

F represents FALSE

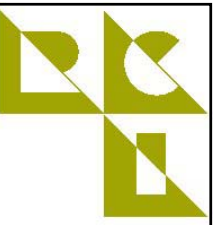
U represents UNKNOWN

Conclusions

- If any data are imprecise, incomplete, or missing, then CWA and 2VL are inadequate
- Imprecise data need a probabilistic approach – is this an extension of CWA / 2VL ?
- If we have any incomplete (e.g. truncated) or missing data we need OWA / 3VL

Conclusions

- A database is not about what IS, but about what IS KNOWN.
- Perfectly reasonable to use the CWA about what IS –
- – but not about what IS KNOWN – precisely because 'I don't know' has to be a valid answer: hence truth value UNKNOWN must be legal



Conclusions

- 3VL need not be scary. It isn't actually much more complex than 2VL
- **Relational databases can use the richness of the Open World Assumption. We just need to do it right.**

Some final words from E.F.Codd (1990)

- *In developing the relational model, I have tried to follow Einstein's advice,* ***"Make it as simple as possible, but no simpler".*** *I believe that in the last clause he was discouraging the pursuit of simplicity to the extent of distorting reality.*
- Is insistence on CWA and 2VL perhaps distorting reality ? A little TOO simple ?