

Why the Title of This Workshop Should Be “Learning Relational Statistical Models from Data”

Stuart Russell

Computer Science Division
University of California
Berkeley, CA 94720-1776
russell@cs.berkeley.edu

The assumption underlying the title “Learning Statistical Models from Relational Data” is that relational data exist.

Data are actual observations. The observation of some random variable X gives us a value x ; procedurally speaking, we obtain x and we assign it (with certainty) as the value of X . We can have certainty in this sense because X may be defined as the random variable whose value will be obtained by the observation procedure (e.g., “the outcome of the next coin flip”). It can, of course, be the case that X corresponds to a noisy measurement of some other quantity W (e.g., X is the “measured battery charge remaining” and W is the “true battery charge remaining”). Again, we typically know the connection between W and X with certainty. Since many observations are noisy, the ability to distinguish true and measured values is important.

When we observe a tuple “ $r(a, b)$ ” (say, in a database), of what random variable is this an observation? Here are some possible answers:

- It is an observation of the Boolean random variable $R(A, B)$. That is, we are uncertain as to whether object A is related by R to object B , and the observation settles the issue with certainty. This is the sense in which relational data exist.
- “ a ” was mistyped, and the observation actually concerns the Boolean random variable $R(C, B)$.
- “ b ” was mistyped, and the observation actually concerns the Boolean random variable $R(A, D)$.
- The data entry clerk was confused about the argument order and the observation actually concerns the Boolean random variable $R(B, A)$.
- “ r ” was mistyped, and the observation actually concerns the Boolean random variable $S(A, B)$.
- Both “ a ” and “ b ” were mistyped, etc., etc.

Hence, unless the model-builder is absolutely certain of the correctness and uniqueness of the identifiers used in the atom—i.e., there is no noise in the observation—it is better to view the observation not as a relational datum, but as an observation of three string-valued random variables. In theory, it could amount to an observation of any of the Boolean random variables corresponding to relations between pairs of objects.

The appropriate level of scepticism depends on the application. Data entry software may be such that confusing the *Supervisor* relation with the *OfficeNumber* relation is impossible. The first argument “ a ” may be a long identifier, such that the probability of misidentification is negligible provided the software does not allow assertions about new identifiers. (Note, however, that my social security number is sometimes used by a small printing company in Indiana, thanks to a transposition of two digits.) Even if the identifiers in one database are never confused, however, we may need to merge (or learn models from) two different databases that use different identifiers. In that case, there is often uncertainty as to which identifiers are equivalent—e.g., is the owner of account “9999-999999”, whose name is recorded as “Stewart Russell” at “263 Hilcrest”, the same person as the owner of account “1234-567890”, whose name is recorded as “Stuart J. Russell” at “263 Hillcrest Rd.”?

The conflation of tuple observations with relational data has been noticed in the context of Web data. Some projects have viewed the existence of a link between two URLs (with appropriate anchor text) as an observation of a particular relation between objects somehow connected with those URLs. Commentators have observed that this could cause difficulties. The home page of a student may contain “student at Palo Alto Junior College” but of course the preceding words might be “My chihuahua Tiggy is a” or “There is no truth in the rumor that I was once a”.

The discussion above assumes that data are already available in tuple form, whether noisy or not. In many actual applications, data are gathered not from relational databases but from text, Web pages, speech, cameras, instruments, etc. Unique identifiers (other than perhaps for the observation objects themselves, e.g., the URL as an identifier of a Web page (if we neglect time)) are generally unavailable in such data, as are relation names. For example, papers have reference lists that “refer” to other papers, but it may require a good deal of sophisticated probabilistic reasoning to work out which papers those might be. Finally, there is the question of the origin of the “refers” relation. We cannot address this question at all if data are already relational.