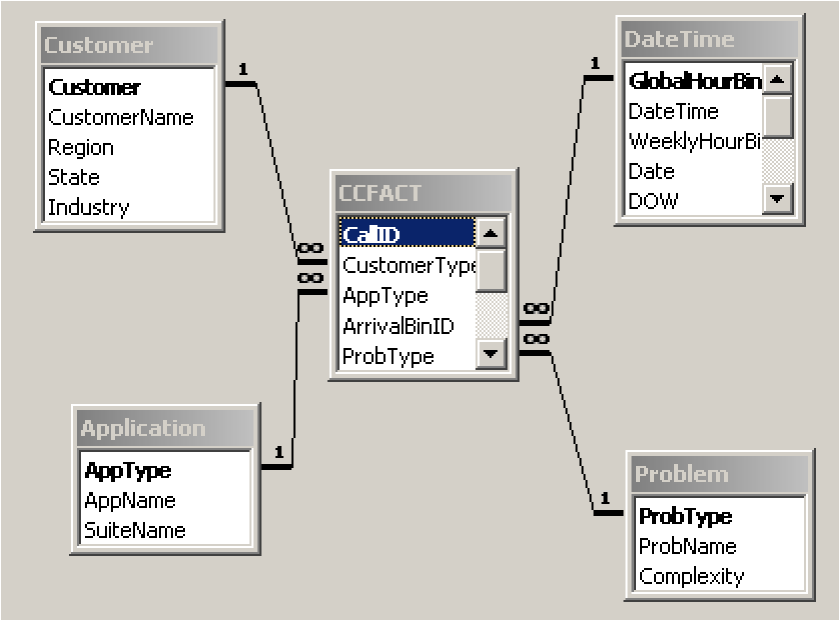
Supplements for Modeling

Example: Logical Data Model of Call Center Data warehouse.



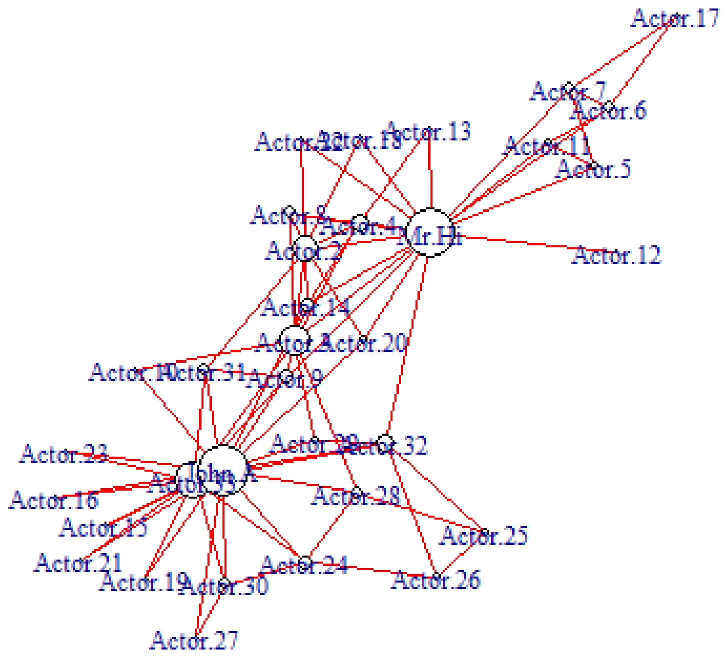
Characteristics of Data Warehouse Logical Data Model

1. It has all the entities which will be used in data warehouse
2. It shows all possible attributes of all entities
3. It depicts the relationships between all entities

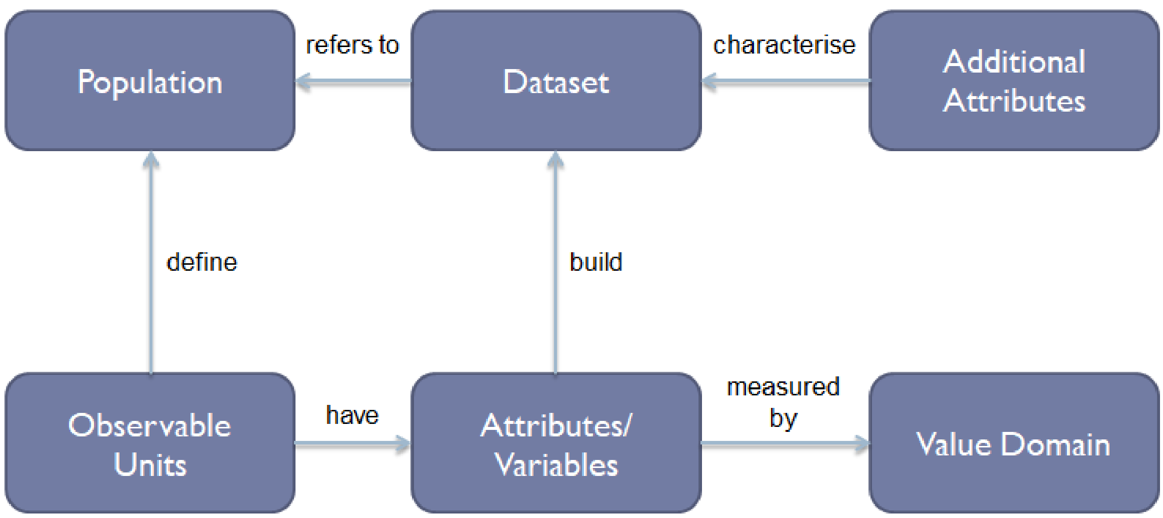
Source: http://www.bidw.org/datawarehousing/logical-data-model/ (Blogger)

Example Network graph of a Karate Club

(A standard example used for demonstration of social networks)



Example: Statistical language (top level)



Example of an a terminology system ConceptNet

Precursor of an ontology

Distance measures in networks:

There are many different centrality measures for undirected networks. The most important are the degree of vertices, the betweenness of a vertex and the closeness of a vertex.

Degree is the number of edges passing this vertex. A vertex with high degree has more connections to other vertices.

Betweenness of a vertex v is defined by the fraction of shortest paths (connections) between vertices which pass the defined vertex v compared to all shortest paths between the vertices. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

Closeness is a measure of the distance of a vertex to all other vertices. A small value of closeness shows that the vertex has small distance to other vertices.

The density compares the number of edges in the network to all possible edges between the vertices. It is always between 0 and 1. A high density indicates that the network has many connections.

Examples of graph structures

1 BPMN

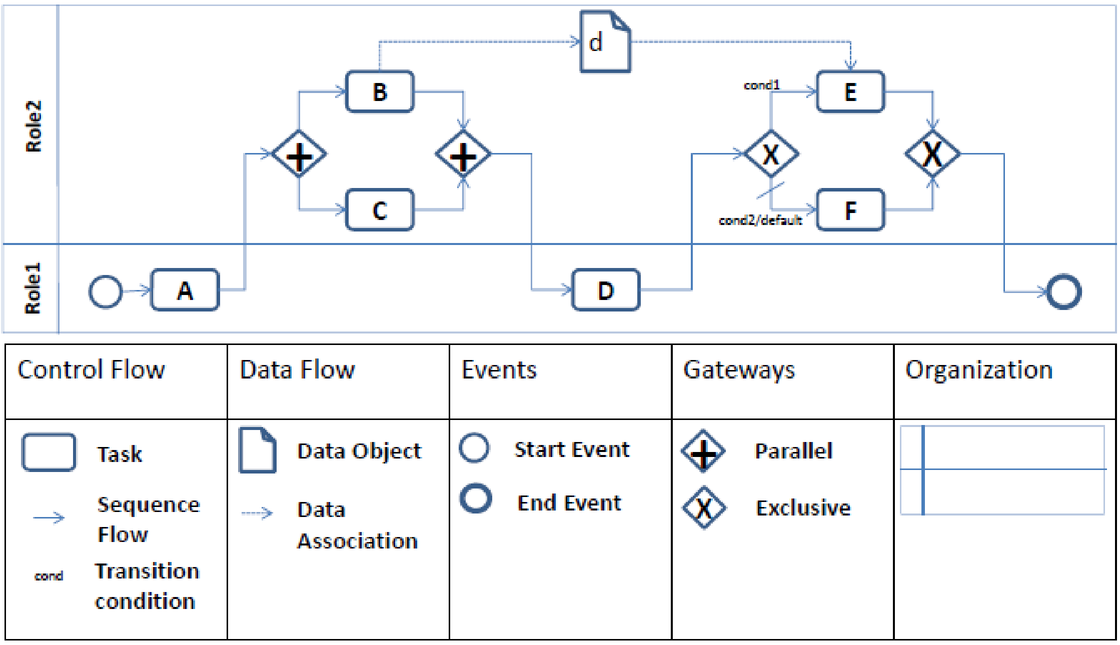


Fig. 2.3 Example Process Model in BPMN Notation; Selected Elements in BPMN 2.0

2. Petri nets

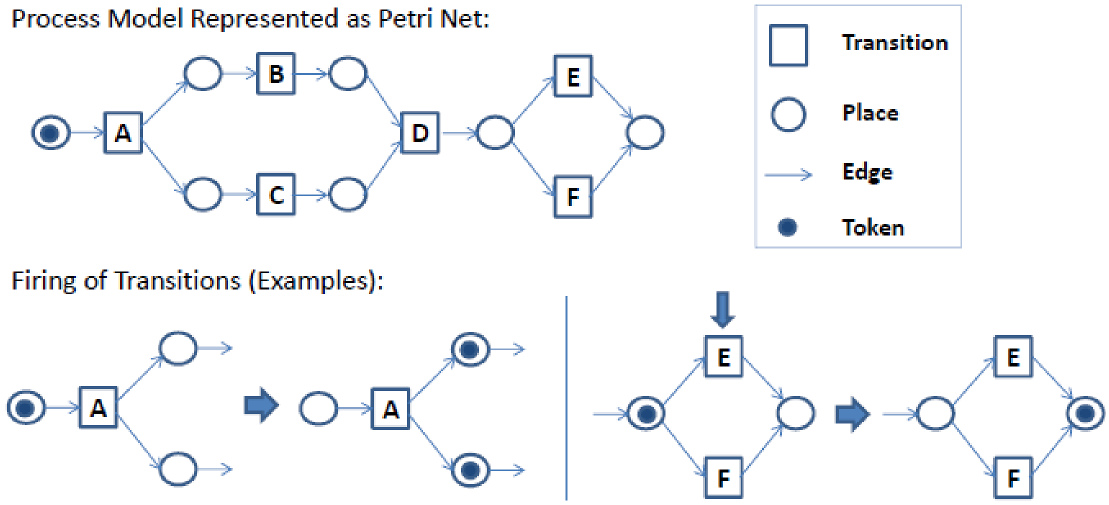


Fig. 2.4 Example Process Model in Petri Net Notation

Joint probability – Bayes Theorem – Independence

Joint Probabilities

Usage Pattern	Age Group		marginal
	young	old	
high	0.2	0.1	0.3
moderate	0.3	0.2	0.5
inactive	0.1	0.1	0.2
marginal	0.6	0.4	1.0

Conditional Probabilities given Usage

Usage Pattern	Age Group		marginal
	young	old	
high	0.67	0.33	1
moderate	0.6	0.4	1
inactive	0.5	0.5	1
marginal	0.6	0.4	1.0

Conditional Probabilities given Age

Usage Pattern	Age Group		marginal
	young	old	
high	0.33	0.25	0.3
moderate	0.50	0.50	0.5
inactive	0.17	0.25	0.2
marginal	0.6	0.4	1.0

Fig. 2.5 Joint probabilities and conditional probabilities

Generalization of this example leads to the relation between conditional distribution, joint distribution and marginal distribution formulated in terms of likelihoods.

$$p(y|x) = \frac{p(x,y)}{p(x)} \qquad p(x|y) = \frac{p(x,y)}{p(y)} \tag{2.11}$$

An application of Bayes Theorem

Links to problems with Web Data:

Confirmation Bias

Network neutrality

Ricardo Baeza-Yates: Bias in the Web (Lecture TU 2017)

Selected References

Models

Frigg, Roman and Stephan Hartmann. Models in Science.
https://stanford.library.sydney.edu.au/archives/spr2009/entries/models-science/

Modeling

Karagiannis, Dimitris and Harald K¨uhn. Metamodelling Platforms. Int'l Conference ECWeb Dexa, p. 182 (2002).

Hand, David, Mannila, Heikki, and Smyth, Padhraic. Principles of Data Mining. The MIT Press, Cambridge, MA and London (2001).

van der Aalst, Wil M.P. et al.: Process Mining Manifesto, Business Process Management Workshops, Springer, pp. 169–194 (2012).

Business Process Model and Notation (BPMN), Version 2.0, OMG Document Number: formal/2011-01-03, Standard document URL: http://www.omg.org/spec/BPMN/2.0, (2011).

Knowledge Representation

Gruber, Thomas R. A translation approach to portable ontology specifications. Knowledge Acquisition 5(2):199–220 (1993).

Guarino, Nicola. Formal Ontology in Information Systems. Int'l Conference Frontiers in Artificial Intelligence and Applications, IOS Press, pp. 3–15 (1998).

Gruber, Thomas R. Ontologies. In: Encyclopedia of Database systems, Springer (2008).

Hepp, Martin. Ontologies: State of the Art, Business Potential, and Grand Challenges. Ontology Management, Computing for Human Experience 7, pp. 3–22 (2008).

Spies, M. Einführung in die Logik . Spektrum Verlag, Heidelberg (2004).

Graph theory and algorithms

10. Dieter Jungnickel. Graphen, Netzwerke und Algorithmen (3rd edition). BIWissenschaftsverlag (1994) (in German).

11. Sedgewick, Robert, Wayne Kevin. Algorithms (4th edition). Addison Wesley (Pearson Education) (2011).

Statistics

Hodges, J.L. Jr., Lehmann, Erich L. Basic Concepts of Probability and Statistics (Classics in Applied Mathematics 48). SIAM (2004).

Crawley, Michael, J. Statistics: An Introduction Using R. J. Wiley (2005).

Hand, David, J.: Information Generation. Oneworld Publication, Oxford (UK) (2007).

Temporal Modeling

Galton, Antony: "Temporal Logic". The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/ fall2008/entries/logic-temporal/.

Combi, Carlo, Keravnou-Papailiou, Elpida, Shahar, Yival: Temporal Information Systems in Medicine. Springer (2010).

Shahar, Y and Musen, M.A.: RESUME: A temporal-abstraction system for patient monitoring.
Computers and Biomedical Research 26(3) (1993).

Quality and quality management

ISO8402: Quality Management and Quality Assurance. https://www.iso.org/standard/20115.html

Batini, Carlo and Scannapieco, Monica: Data Quality – Concepts, Methodologies and Techniques. Springer DCSA (2006).

Apel, Detlef, Behme, Wolfgang, Eberlein, R¨udiger, Merighi, Christian. Datenqualität erfolgreich steuern. TDWI Europe, Hanser, Munich (2010)(in German).