Diagram Schemas: What, Why, How + www.Diagrams.org

Robert P. Futrelle

Biological Knowledge Laboratory, College of Computer & Information Science
Northeastern University, Boston, MA 02115
http://www.ccs.neu.edu/home/futrelle/

Abstract

Diagrams play a critical role in recording and communicating information that is difficult to express or comprehend using only text. In contemporary biology research papers, figures account for approximately 50% of the content, amazingly enough. (This counts the space occupied by the figures per se as well as the text of the captions and the discussion of the figures in the body of the paper.) Though there are many systems for indexing and searching text content, there is virtually nothing available today for figures.
This poster looks at several aspects of diagrams, particularly the categorization problem. Categorization must be understood in order to properly understand and design useful diagram-related systems. This poster revises some of the material in the corresponding paper in the Diagrams 2004 proceedings.

www.Diagrams.org - The World of Diagrams

Diagrams.org is intended to be a comprehensive and scholarly site about a broad range of topics related to diagrams. It launches 19 March 2004. Its primary content is a gallery of diagrams, typically drawn from journals, with accompanying information such as the original caption, body text of the paper that discussed the diagram, citation, e-copy of the full paper, and notes by Futrelle. The other material is a collection of essays by Futrelle (typically illustrated) on a range of diagram-related topics such as strata and sections, the arrow, sketching, external cognition, layout and semantics for indexing. Future plans for the site include a large bibliography, news, a list of major researchers in the field (past and present), and an archived mailing list. I am looking forward to receiving contributions to the site from other scholars and practitioners in the future. Diagrams.org complements my already existing BioNLP.org site, to cover both graphics and text.

What?

Diagram schema - A conceptualization, including categorization as well as the conception, design, production and use of diagrams. Schemas and most of the concepts below are highly context-dependent.

Categorization - The placement of a diagram in our conception of artifacts in the world [7].

Ontology - For specification purposes: A description of the concepts and relationships in a domain which is then used to enable communication. Often hierarchical.

Instance - Any specific, "concrete" diagram.

Exemplar - A remembered or stored diagram example (instance).

Prototype - An abstract summary representation of a class of diagrams.

Family resemblance - Useful within complex and difficult to define classes such as "game", "furniture", "diagram".

Central - A common reference class, e.g., toward the middle of the sequence: diagram, quantitative plot, Cartesian plot, line data, line plot with points and error bars.

Intensional definition - The necessary and sufficient conditions that an item of a class must obey.

Extensional definition - A collection of instances which serve to define a class.

Why?

Diagram schema aid in the design of comprehensive diagram-related systems, from drawing tools to publications, indexes and online document reading systems.

Categorization is needed to specify diagrams during production, for indexing and retrieval and for users in specifying the type of diagrams they want to see. Ontologies can help us organize and relate a variety of diagram classes.

The concepts: instance, exemplar and prototypes show the variety of ways we can view categories.

The notions of family resemblance and centrality are important in building systems for people to use, because they are very much in tune with the way people deal with diagrams and what the expect diagram systems to offer.

Intensional and extensional definitions of diagram classes are critical in the design of computer systems that deal with diagrams. Intensional definitions can be elaborated to describe not just diagram classes but to describe their entire internal structure. Extensional definitions are well-suited for use in interactive systems where examples are presented to users.

How?

All systems that offer users a choice of diagram classes, e.g., statistical packages and spreadsheets, employ exemplars and have organized the choice of classes according to some implicit or labeled ontology.

Discussions of diagrams typically begin with central examples and work outward from there.

When writers or lecturers design diagrams to communicate results, they must choose the class of diagrams right down to specific design parameters. They then must generate instances that hold the information and data they wish to communicate. They have to choose diagram designs that are familiar to their audience.

To go more deeply into diagram structure and content it is necessary to build computer-based systems that use some mix of the following two approaches:

Machine learning, which can deal with categorization by using large collections of examples, either in supervised or unsupervised learning paradigms.

Grammar-based diagram parsing, which requires detailed intensional descriptions [4]. The author has successfully implemented such a diagram parsing system [5].

Unfortunately, virtually all electronically published diagrams today are in raster format, so we have embarked on a major project to build a system that can "vectorize" diagrams, converting them from pixels to objects such as parametrically described lines, polygons and curves.

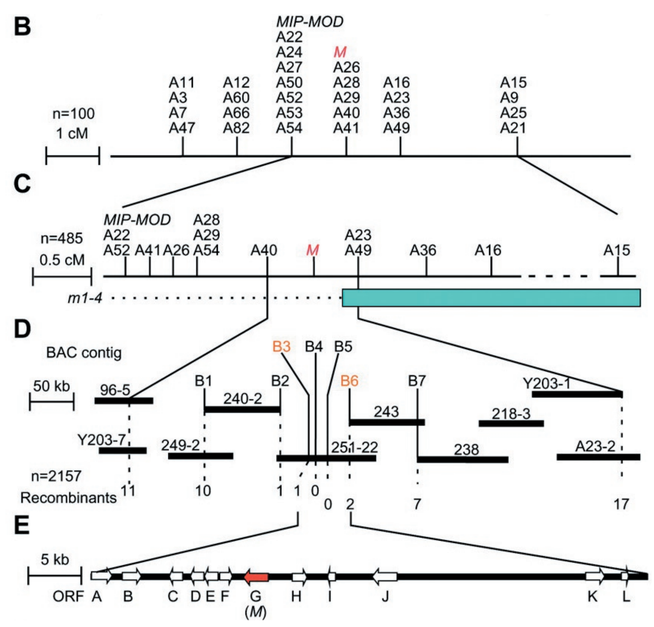

Figure 1. Molecular biology is a field in rapid evolution. With new and more complex phenomena and structures being discovered every day, many adaptations and extensions of "standard" diagrams are constantly appearing. This is a complex gene diagram. It includes views at various levels of resolution via the "expander lines". It also includes extensive text labeling.
(Source: Science **303**, pg1516, 2004)

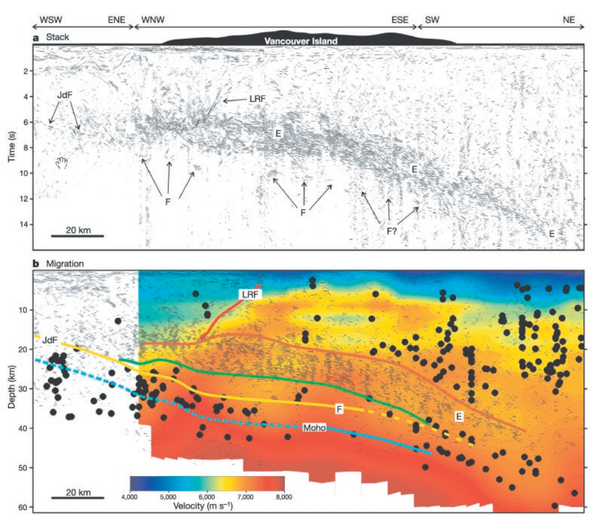

Figure 2. This is not a photograph of a natural object in any conventional sense. Rather it is a visualization of seismic records in the vicinity of Vancouver Island, Canada. The raw data is overlaid with drawn lines and text labels. Such annotations began with the invention of photography. Even before photography, labels and callouts in etchings were common. A color key is used -- people are attuned to color keys in nature (blue versus dirty water, live green versus dead brown vegetation). Schemas for the production of such pseudocolor plots are now common. A central example of such coding would be the daily weather map in a newspaper, coding temperatures from blue (cold) to red (hot), which are nature-based metaphors.
(Source: Nature **428**, pg163, 2004)

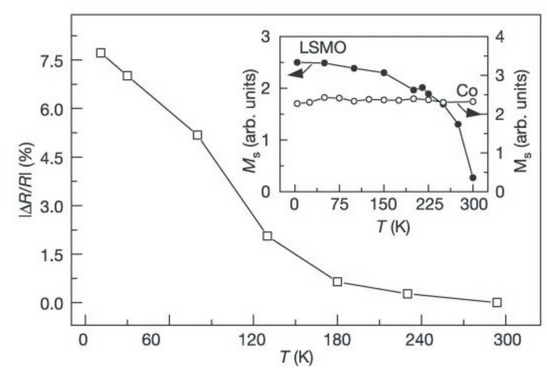

Figure 3. This is a classic Cartesian (x,y) data graph, in this case with an additional inset graph. The associated schema is also quite standard: the marshaling of the dependent data values and the choice of axis scales and their description. All technical people know how to read such diagrams. Only a specialist will understand the units involved, much less the semantics in the specialized paper This is often the case for diagrams in highly technical papers. The closest we can come to an exemplar (remembered instance) of diagrams of this class would be an individual's memory of a specific diagram of some importance or one used as a starting point such as the "standard" chart types available in a spreadsheet application or statistical package. This diagram is a prototypical example because it exhibits features that are central to a commonly understood class.
(Source: Nature **427**, pg821, 2004)

Annotated Bibliography

1. Diagrams.org is a brand-new site (beta in March 2004). The site will attempt to go well beyond the well-known work by Tufte [2], because it will include extensive material on diagram research, cached diagram source papers, diagram research papers and more:
Futrelle, R. P. (2004). Diagrams.org. The World of Diagrams (web site). *http://www.diagrams.org*.

2. This is a famous book, both a scholarly work and a popular "coffee table" book:
Tufte, E. R. (1997). *Visual Explanations*. Cheshire, CT: Graphics Press.

3. This is a recent comprehensive collection of papers on diagrams, with some omissions, e.g., diagram/visual-language parsing:
Anderson, M., Meyer, B., & Olivier, P. (Eds.). (2002). *Diagrammatic Representation and Reasoning*. London: Springer-Verlag.

4. This collection focuses on visual languages, including parsing:
Marriott, K., & Meyer, B. E. (Eds.). (1998). *Visual Language Theory*: Springer Verlag.

5. The author has built a diagram parsing system which is described here, including screen shots:
Futrelle, R. P. (1998). The Diagram Understanding System Demonstration Site. *http://www.ccs.neu.edu/home/futrelle/diagrams/demo-10-98/*.

6. Reasoning with and about diagrams is well covered in this collection:
Glasgow, J., Narayanan, N. H., & Chandrasekaran, B. (Eds.). (1995). *Diagrammatic Reasoning*. Menlo Park, CA/ Cambridge, MA: AAAI Press/ MIT Press.

7. Categorization is treated in this authoritative and delightful book:
Lakoff, G. (1987). *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. Chicago: U. of Chicago Press.