Eurographics Conference on Visualization (EuroVis) 2017
B. Kozlíková, T. Schreck, and T. Wischgoll (Editors)

An Argument Structure for Data Stories

Robert Kosara

Tableau Research

Abstract

Many data stories in journalism do not have a story arc, but rather present facts without much structure. This mirrors the popular inverted pyramid style of writing that presents the most important information up front, to be followed by evidence.
We have found a subset of stories that follow a more structured approach, however. These stories begin with a claim or question, but do not immediately present that as the conclusion. Instead, they then present pieces of evidence that are only tied together, and back to the initial claim, at the end.
In this paper, we formalize and discuss this structure, and present a few examples. We believe that this is a viable and practical model for data stories more generally, and certainly a stronger arc than most existing stories today.

1. Introduction

A common way of writing in journalism is called the *inverted pyramid*, sometimes also the *inverted triangle*: state the most important piece of information in the headline, then follow that with the next-most important information in the opening (*sub-hed* or *lede* in journalism terminology), and then continue adding information.

Many data stories in news media follow this structure, which means that they have clearly-defined and structured beginnings, but they simply end because there is no more information to provide. Instead of having a defined ending, they peter out.

The inverted pyramid structure has the advantage that it allows readers to get the gist of the story even if they only read the title or a few sentences. But it also lacks a clear ending, and thus only really shapes one end of the story, leaving the other one open. Is there a better way?

More classical stories like they are told in movies, novels, etc., don't tend to reveal all the information up front, but only provide some expectations early on, and then often contain twists or reveals towards the end. Crucially, the ending ties together the events of the story and provides a form of closure.

In following news graphics over the last few years, we have observed a pattern that bridges the gap between the classic story structure and the inverted pyramid: it starts out by asking a question or making a claim, then provides evidence, and finally closes by tying the evidence back to the initial claim or question.

In this paper, we present this structure, discuss a number of examples from news media that employ it, and argue that it is a viable and generalizable pattern that provides a clearer structure than the inverted pyramid (and is more suited to data than the classic story structures).

2. Related Work

Efforts have been made to formalize the structure of data visualization, such as Card and Mackinlay's overview of the design space [CM97], Wilkinson's *Grammar of Graphics* [Wil05], or languages like VizQL [STH02].

Much of the research in stories is not very formalized or easily applicable to data. One exception is work on comics, in particular Cohn's model of the structure of the story. He defines five types of frame in a comic: Establisher (E), Initial (I), Prolongation (L), Peak (P), and Release (R) [Coh12]. In common four-pane comics, he finds the EIPR structure (Establisher followed by an Initial, then Peak, then Release) the most common. Cohn also shows that there are elements of a visual grammar present in the way we read sequential images, which he verifies by measuring the brain's confusion response to elements that are missing or out of order [CPJ*12].

Amini et al. [ARL*15] have applied Cohn's classification scheme to data videos. They found all the components in many of the videos, but the relative weighting is quite different from that in comics. Expressed in a regular expression-like format, the dominant structures they found were: E+I+PR+ and E+I+P. The latter structure is unusual in comics, but it might be a good structure for news stories that don't have a punchline that requires an additional frame for maximum effect (which is what the Release does). The former structure is partly due to videos showing a variety of views for each point, rather than focusing on a single one like a comic.

We do not suggest that Cohn's EIPR structure is necessarily the best fit for data stories, but his structural classification of the frames provides a good framework to analyze news pieces. While the frames in a comic are generally presented at once, they are read sequentially. The examples discussed below all present the frames

DOI: 10.2312/eurovisshort.20171129

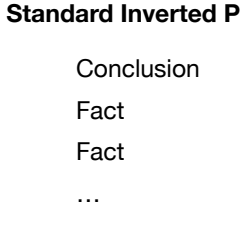

Figure 1: *The classic inverted pyramid structure starts with the conclusion and then presents a list of facts without a clear ending.*

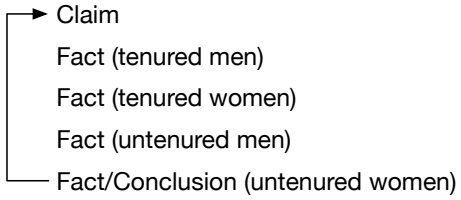

Figure 2: *The structure of Tenure Pipeline at Harvard Business School [Fai14] consists of a claim, three facts, and one final combined fact/conclusion.*

sequentially. We do not believe that this has a significant impact on the applicability of the model, however. Cohn also uses sequential presentation in some of his studies [CPJ*12].

Existing research has looked at the sequencing [HDHR*13] and framing [HD11] of data presentation steps, but not the structure of the actual argument being made.

Sequences like the inverted pyramid are often cited in writing about journalism [Har11]. They have inspired the idea of *leading with the ending* for presentations [NK15]. A model that is closer to the one presented below is called the *hourglass structure* [Sch17] for presentations: give a preview, present the argument, then end with conclusions. This structure has not been studied in the context of pure data storytelling or news graphics, however.

We note the difference between the model presented below, as well as the hourglass model, and the Martini glass structure described by Segel and Heer [SH10]. The latter describes ways of interaction with a story, where the reader is led through a narrative before being given the ability to interact with the data at the end. These two structures are entirely orthogonal, however. Some of the case studies presented below follow the Martini glass (we point them out), but not all of them do.

3. The CFO Pattern

We propose a model that mirrors Cohn's EIP structure and follows the E+I+P pattern found by Amini et al. Unlike a comic, it consists of the following elements (we picked some non-obvious letters to avoid confusion with Cohn's model).

Claim (C) or Question. A claim is made or a question asked. Crucially, it is not assumed to be self-evident or easily answered.

Facts (F), Evidence. Evidence is presented. This can be elaborate and the facts do not necessarily tie back to the claim or question in an obvious way.

Explanation (X). Optionally, facts can be interspersed with explanations of mechanisms, how to read the visualization, etc.

Conclusion (O). The last step ties the evidence presented back to the claim or question. In some cases, the final fact also acts as the conclusion. We denote this as F_O.

Similar to Cohn, we use the letters to denote the sequence in a way similar to simple regular expressions (+ means repetition, parentheses group). In the simplest case, it will be CF+O (a claim followed by one or more facts and a conclusion). More complex structures are possible, however, that return to the initial question multiple times in a structure that looks like C(F+O)+ or C(F+O)+O. In the former structure, the conclusion of the final block also serves as the overall conclusion, while the latter would have a separate overall conclusions step (Section 4.3).

It is also conceivable to have a structure with an overall claim that is followed by smaller claims, in a pattern like C(CF+O)+O. We have not encountered such a structure yet, however.

In our notation, the simple inverted pyramid would be denoted OF+ (Figure 1). The CFO pattern differs from it not only in its formal structure, but also in that it provides a clearly-defined end to the story. Readers need to read or watch the entire story to get the argument laid out for them. This might compel them to consume more of the story than otherwise. The pattern also has a defined arc, making it more comparable to classical stories.

4. Case Studies

We have observed this pattern in a number of examples and believe that it is of value. The examples listed below are exemplars, however, rather than an exhaustive set based on a systematic study. We group them by the way the claim or question is framed.

Provide Proof One use is to provide evidence: the title makes a claim or statement that has the reader think, *Oh yeah? Prove it!* The story then provides the additional information in a way that is not just a laundry list, but builds from the weakest to the strongest evidence (see Sections 4.1 and 4.2 for examples).

Question A variation of this is the question, which usually has a foregone conclusion. The title asks a question and then answers it in much the same way as the previous point (Sections 4.3 and 4.4).

Long Explanation Some stories need to explain not just the evidence itself, but also the way it is being presented (Sections 4.4 and 4.5).

All but the *Tenure Pipeline* (Section 4.1) example allow some interaction after the conclusion, thus technically fitting the Martini glass structure. The only piece that lets the reader explore data that was not covered in the initial narrative is *For the Elderly, Diseases that Overlap* (Section 4.5), however.

4.1. Tenure Pipeline at Harvard Business School

This piece by Hannah Fairfield from 2014 [Fai14] consists of five steps (Figure 2). The first shows the total number of male and female faculty members at Harvard Business School (HBS) broken down by gender. This is accompanied by the claim that HBS "says it wants to improve the gender balance among faculty members, but it is far from that goal without extensive hiring." Step two lays out the tenured male faculty members on a timeline by the number of years they have been at HBS. Step three adds the female tenured professors, clearly a much smaller number. Step four then adds the male tenure pipeline, i.e., the male untenured professors, some fraction of whom will receive tenure at HBS. Their number is almost twice that of the tenured male professors (120 untenured vs. 76 tenured). The final step adds the female untenured professors – clearly a much smaller number than the men (and almost the exact same relation between untenured and tenured: 34 vs. 19). It does not appear likely that the gender balance is going to change anytime soon.

This final step is thus both a piece of evidence and the conclusion of the piece. Without the final step, the story is incomplete.

Claim: *Harvard Business School says it wants to improve the gender balance among faculty members, but it is far from that goal without extensive hiring.*

Structure: CFFFF_{O}: Claim, Fact (tenured men), Fact (tenured women), Fact (untenured men), Fact/Conclusion (untenured women)

4.2. 2016 Was the Hottest Year on Record

Tom Randall and Blacki Migliozzi make a clear claim the title of this piece: *Not a Hoax: 2016 Was the Hottest Year on Record* [RM17]. The evidence presented are the temperatures of all years from 1880 to 2016, which are shown in one continuous animation, though drawn year by year. Each year is drawn as a line chart with the months along the x axis, thus building up a chart that slowly creeps upwards, with new highest yearly averages highlighted and tracked by a reference line. At the end, the final line very clearly sits above all the others, presenting both the final piece of evidence and the conclusion.

Even though the structure is very simple and linear, there is a strong arc: make a claim at the top, then present the evidence. The piece almost entirely relies on the data to tell its own story. But because of the way the temperatures build up, the peak occurs at the very end. There is a certain lack of subtlety that makes this story quite powerful and memorable.

After the animation has run, the reader can mouse over the lines to see the years and temperatures labeled.

Claim: *No Hoax: 2016 Was the Hottest Year on Record*

Structure: CF+F_{O}: Claim, Facts (continuous years 1880-2015), Fact/Conclusion (2016)

4.3. What's Really Warming the World?

At the top of this piece that is related to the previous one, Eric Roston and Blacki Migliozzi ask a simple question: "Skeptics of manmade climate change offer various natural causes to explain why the Earth has warmed 1.4 degrees Fahrenheit since 1880. But can these account for the planet's rising temperature?" [RM15] The viewer is then taken through a number of possible scenarios by scrolling down (Figure 3): the Earth's orbit, the sun, volcanos, all three combined, deforestation, ozone pollution, aerosol pollution, and finally greenhouse gases. In a final step, the viewer is invited to try it him- or herself by being able to turn different potential causes on and off and comparing their effects to the measured temperature increase.

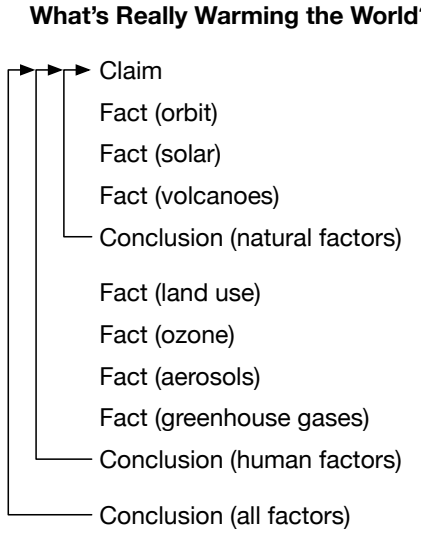

Figure 3: *The structure of What's Really Warming the World? [RM15] consists of a claim, facts, two sections with their own conclusions, and one overall conclusion.*

The structure of this piece is more complex than most we have observed. There are two separate sections, one for natural factors, the other for human causes. Each presents a number of individual facts and ends in its own partial conclusion. The piece ends in an overall conclusion that summarizes all of them and compares them to the observed temperature anomaly.

After the conclusion, readers are able to pick different potential causes and combine them to see how they compare to the observed temperature anomaly. Even though no additional data can be explored this way, being able to create new combinations makes this story fit the Martini glass structure.

Claim: *Skeptics of manmade climate change offer various natural causes to explain why the Earth has warmed 1.4 degrees Fahrenheit since 1880. But can these account for the planet's rising temperature? Scroll down to see how much different factors, both natural and industrial, contribute to global warming, based on findings from NASA's Goddard Institute for Space Studies.*

Structure: CFFFOFFFFOO: Claim, Fact (orbit), Fact (solar), Fact (volcanoes), Conclusion (natural factors); Fact (land use), Fact (ozone), Fact (aerosols), Fact (greenhouse gases), Conclusion (human factors); Conclusion (all factors)

4.4. Turning a Corner?

Amanda Cox's *Turning a Corner?* [Cox09] turns a chart of a single time series, industrial production, into a two-dimensional connected scatterplot [HKF15]. The title asks a question, and the lede claims that "the economy is poised to turn around, but that the climb out of the current downturn will be a long one."

This piece has the most explicit explanation steps. In the first step following the initial one, the reader is first introduced to the idea of a business cycle and how it is shown in four quadrants of a two-dimensional graph. The following step shows the way the data is transformed (change in last six months on horizontal axis, amount compared with trend on vertical). The next four steps show recent recessions with animations illustrating developments over time, ending with the then-current one (the piece was published in 2009). After that, another explanation step introduces the idea of a leading indicator by showing how industrial production has closely followed the indicator with a roughly six-month lag before. In the final step, while the main timeline is still squarely in the downturn quadrant, the leading indicator is pointing towards recovery – thus presenting the conclusion that ties back to the initial claim.

After the end of the narrative, the reader is able to browse through the data to see parts of the timeline that have been skipped.

Claim: *A chart of industrial production [...] suggests that the economy is poised to turn around, but that the climb out of the current downturn will be a long one.*

Structure: CXXFFFFXO: Claim, Explanation (four parts of the business cycle), Explanation (how to read the chart), Fact (1970s), Fact(1990s), Fact (2007), Fact (2009), Explanation (what is a lading indicator), Conclusion (leading indicator points to recovery)

4.5. For the Elderly, Diseases That Overlap

Matthew Bloch and Hannah Fairfield investigate the diseases elderly people living in assisted-living centers suffer from [BF13]. They show that most people have multiple chronic conditions.

To walk readers through the argument, they create a Venn diagram from individual human shapes (each representing 600 residents of assisted-living centers). The initial claim step shows all of them in a circle. Two fact steps show the prevalence of Alzheimer's and high blood pressure individually. The fourth step is a mix of explanation and fact, showing the overlap between Alzheimer's and high blood pressure. Step five presents the amount of heart disease in that population by itself. The conclusion in step six finally shows all three conditions presented before in all their combinations, with single conditions only making up 40% of the population, and the remaining 60% having two or all three.

One further step allows readers to combine conditions from a larger list than was covered in the story, making this a true Martini glass design as well.

Claim: *More than 700,000 people live in assisted-living centers [...]. Most of the residents have multiple chronic health conditions.*

Structure: CFFFF_{O}: Claim, Fact (Alzheimer's), Fact (high blood pressure), Fact (Alzheimer's and high blood pressure), Fact (heart disease), Fact/Conclusion (Alzheimer's, high blood pressure, heart disease in all their possible combinations)

5. Discussion

The CFO pattern is a simple structure, but provides cohesion between the steps as well as a bracket around the entire piece, that the inverted pyramid lacks.

In contrast to the inverted pyramid, the pattern we present here not just defines the beginning of the story, but structures the entire piece by tying the end back to the beginning. It also includes elements of classical story without trying to emulate it too closely: it mirrors the Aristotelian story arc (beginning, middle, end) that has held audiences' attention for millennia [NK15]. The presence of a recognizable beginning and an attempt at resolution have been recognized as the universal building blocks of true narratives across all cultures [Aus11]. The initial claim also sets an expectation, which the conclusion satisfies. We tend to want to see a story that has been teased to its completion [Zei27].

While there are similarities between the CFO and EIP patterns, they are quite different. The Claim or Question (C) seems similar to the Establisher (E) in Cohn's model, but differs in that the expectation in a comic is set by the Initial (I) instead of the Establisher. The Facts (F) and Explanations (X) act mostly as what Cohn considers Prolongations (L) that continue the action set in motion by the Initial. Finally, the Conclusion (O) sits at the end of an argument structure, whereas the Peak (P) is typically followed by one or more frames (Release (R) in Cohn's model of comics, but also in classical story). In an argument, the conclusion is the end; that is not the case for the climax in a classical story.

There is certainly also overlap between this model and Cohn's, but the descriptors serve different purposes and are much more specific to data stories or arguments.

Going further, there are many possible questions we might ask about this structure. Should the motivation be a claim or a question? Which of the many possible structures outlined above work better than others? Is a separate conclusion better than a combined fact and conclusion? What number of steps works well to keep people's attention until the end? Is substructure like the small conclusions in *What's Really Warming the World?* effective?

6. Conclusions and Future Work

Structure in data stories is a topic that is difficult to formalize and classify. Most existing data stories seem to follow the inverted pyramid format, which provides very little actual structure.

In collecting the examples above and extracting their underlying pattern, we hope to have contributed a piece of structure that is both analytically useful and practically applicable.

The examples also show that narrative structure can not only coexist with the need for quick information in journalism, but be quite effective in combination.

This is only a small first step. We believe that there are other patterns that are being used by the authors of data stories, but that have yet to be discovered and formalized.

References

[ARL*15] AMINI F., RICHE N. H., LEE B., HURTER C., IRANI P.: Understanding Data Videos: Looking at Narrative Visualization through the Cinematography Lens. In *Proceedings CHI* (2015), pp. 1459–1468.

[Aus11] AUSTIN M.: *Useful Fictions: Evolution, Anxiety, and the Origins of Literature*. University of Nebraska Press, 2011.

[BF13] BLOCH M., FAIRFIELD H.: For the elderly, diseases that overlap. http://nyti.ms/16YUqAo, 2013.

[CM97] CARD S. K., MACKINLAY J. D.: The Structure of the Information Visualization Design Space. In *Proceedings Information Visualization* (1997), IEEE CS Press, pp. 92–99.

[Coh12] COHN N.: Visual Narrative Structure. *Cognitive Science 37*, 3 (Nov. 2012), 413–452.

[Cox09] COX A.: Turning a corner? http://www.nytimes.com/interactive/2009/07/02/business/economy/20090705-cycles-graphic.html, 2009.

[CPJ*12] COHN N., PACZYNSKI M., JACKENDOFF R., HOLCOMB P. J., KUPERBERG G. R.: (Pea)Nuts and Bolts of Visual Narrative: Structure and Meaning in Sequential Image Comprehension. *Cognitive Psychology 65*, 1 (Aug. 2012), 1–38.

[Fai14] FAIRFIELD H.: Tenure pipeline at harvard business school. http://nyti.ms/1cqTsOM, 2014.

[Har11] HART J.: *Storycraft*. The University of Chicago Press, 2011.

[HD11] HULLMAN J., DIAKOPOULOS N.: Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 17*, 12 (2011), 2231–2240.

[HDHR*13] HULLMAN J., DRUCKER S., HENRY RICHE N., LEE B., FISHER D., ADAR E.: A Deeper Understanding of Sequence in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 19*, 12 (2013), 2406–2415.

[HKF15] HAROZ S., KOSARA R., FRANCONERI S.: The Connected Scatterplot for Presenting Paired Time Series. *IEEE Transactions on Visualization and Computer Graphics 22*, 9 (2015), 2174–2186.

[NK15] NUSSBAUMER KNAFLIC C.: *Storytelling with Data*. Wiley, 2015.

[RM15] ROSTON E., MIGLIOZZI B.: What's really warming the world? http://www.bloomberg.com/graphics/2015-whats-warming-the-world/, 2015.

[RM17] RANDALL T., MIGLIOZZI B.: No hoax: 2016 was the hottest year on record. http://www.bloomberg.com/graphics/hottest-year-on-record/, 2017.

[Sch17] SCHWABISH J.: *Better Presentations*. Columbia University Press, 2017.

[SH10] SEGEL E., HEER J.: Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1139–1148.

[STH02] STOLTE C., TANG D., HANRAHAN P.: Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 8*, 1 (2002), 52–65.

[Wil05] WILKINSON L.: *The Grammar of Graphics*. Springer, 2005.

[Zei27] ZEIGARNIK B.: Das Behalten erledigter und unerledigter Handlungen. *Psychologische Forschung 9*, 1 (1927), 1–85.