# Constraint representation towards precise data-driven storytelling

Yu-Zhe Shi*
The Hong Kong University of Science and Technology

Haotian Li†
The Hong Kong University of Science and Technology

Lecheng Ruan‡
Peking University

Huamin Qu§
The Hong Kong University of Science and Technology

**ABSTRACT**

Data-driven storytelling serves as a crucial bridge for communicating ideas in a persuasive way. However, the manual creation of data stories is a multifaceted, labor-intensive, and case-specific effort, limiting their broader application. As a result, automating the creation of data stories has emerged as a significant research thrust. Despite advances in Artificial Intelligence, the systematic generation of data stories remains challenging due to their hybrid nature: they must frame a perspective based on a seed idea in a top-down manner, similar to traditional storytelling, while coherently grounding insights of given evidence in a bottom-up fashion, akin to data analysis. These dual requirements necessitate precise constraints on the permissible space of a data story. In this viewpoint, we propose integrating constraints into the data story generation process. Defined upon the hierarchies of interpretation and articulation, constraints shape both narrations and illustrations to align with seed ideas and contextualized evidence. We identify the taxonomy and required functionalities of these constraints. Although constraints can be heterogeneous and latent, we explore the potential to represent them in a computation-friendly fashion via Domain-Specific Languages. We believe that leveraging constraints will facilitate both artistic and scientific aspects of data story generation.

**Index Terms:** Data-driven storytelling, structural representation, domain-specific language, constraint programming.

## 1 INTRODUCTION

Data-driven storytelling is a powerful vehicle for conveying ideas persuasively to target audiences. It has been widely applied in various scenarios [3, 20, 35, 7], including science education, clinical diagnosis with therapy interpretation, product popularization, and public policy advocacy, among others. Creating data stories is a comprehensive and costly effort that integrates multiple processes: mining relevant data, interpreting data insights, organizing textual narratives, and rendering visual illustrations [39, 24]. Hence, the automatic generation of data stories is crucial for broader applications, despite its inherent complexity [51, 25, 26].

The major challenge of data-driven storytelling arises from the dual requirement of *communicating subjective knowledge and insights* while *supporting them with objective evidence*. When talking about a data story, we expect the insights behind the proposed therapy to align with the patterns entailed in clinical data regarding a healthcare propaganda; we expect the elements of the story to fairly reflect events happening in the real world regarding a theme-based news report; we expect to see photorealistic illustration animations with physically-real rendered ocean currents in an advocacy about marine pollution; and we expect to derive actionable messages that meet our realistic requirements when inquiring about government's record of export trade data. These *twisted* expectations reveal the hybrid nature of data stories — a blend of the imaginative aspects of storytelling and the grounded basis of evidence.

Data stories are not pure stories. A conventional story is constructed solely based on a core seed idea and is derived from the seed idea in a top-down manner. Stories make rationalization of scenarios, characters, and plots for a self-consistent virtual world with its unique dynamics [12]. In contrast, data stories are required to be persuasive. Every piece of material used to organize the story should be grounded in evidence from the real world, thus limiting the extent of rationalization compared to purely fictional narratives.

At the same time, data stories are not mere summarizations or reports of data. Those evidence-based approaches aim to reflect every aspect of data in detail, reconstructing the objective material without information loss in a bottom-up fashion. In contrast, data stories are required to convey specific, and sometimes opinionated, ideas rather than neutral ones [6]. All the evidence should compactly support the target idea, and the irrelevant information should be discarded, thereby refining the interpretation from a uniform one.

Putting together, data-driven storytelling *intertwines* the methodologies of top-down storytelling and bottom-up evidence-based analysis. If we consider a spectrum indicating the proportion of evidence and rationalization, data stories would lie in the middle of the two endpoints, mediating the properties of traditional stories and data reports. While a vast space of reasonable stories can be created through the lens of highly diversified perspectives, subjected to the authors' personal and societal contexts [17, 5, 6, 23, 22, 31], there should exist a *boundary* that restricts the space to a compact and permissible one, where the dual requirements of data stories, conveying ideas and grounding evidence, can be both satisfied.

This boundary can take various forms, such as domain-specific knowledge to ensure the integrity of data analysis, external memory of key variables to maintain logical coherence in the textual narrative, or models for model-based generative rendering to create physically-real visual illustrations. Regardless of its type, this boundary excludes the possibility of generating ambiguous or problematic data stories, thereby enhancing the precision of data story generation. The bounded permissible space, shaped by both seed ideas and the boundary, maximizes the flexibility of storytelling while ensuring that the story remains firmly grounded in evidence.

We refer to such boundaries as the **constraints** of data stories. Constraints, in contrast to production sets, determine whether a final product is *feasible* under given requirements rather than defining the final product itself, which has been introduced to facilitate effective visualization creation, such as recommendation and verification [34, 8, 41, 52, 40]. In the context of data-driven storytelling, the production set is the seed idea of the story, which shapes the perspective to frame, the position to hold, and the context to include. Accordingly, we have various types of constraints at different hierarchies to ensure that data stories become what they should be.

Given the varied requirements of constraints, we hold the position that constraints are significant for generating data stories that are precisely coherent with both the author's intention and the grounded evidence. Unfortunately, although various generative tools have been developed, the efficient representation of constraints remains under-researched. This is not a trivial problem — constraints are heterogeneous and often latent. As aforementioned, constraints can be first-order rules, higher-order rules, structural

*e-mail: syz@autodsl.org
†e-mail: haotian.li@connect.ust.hk
‡e-mail: ruanlecheng@ucla.edu
§e-mail: huamin@cse.ust.hk

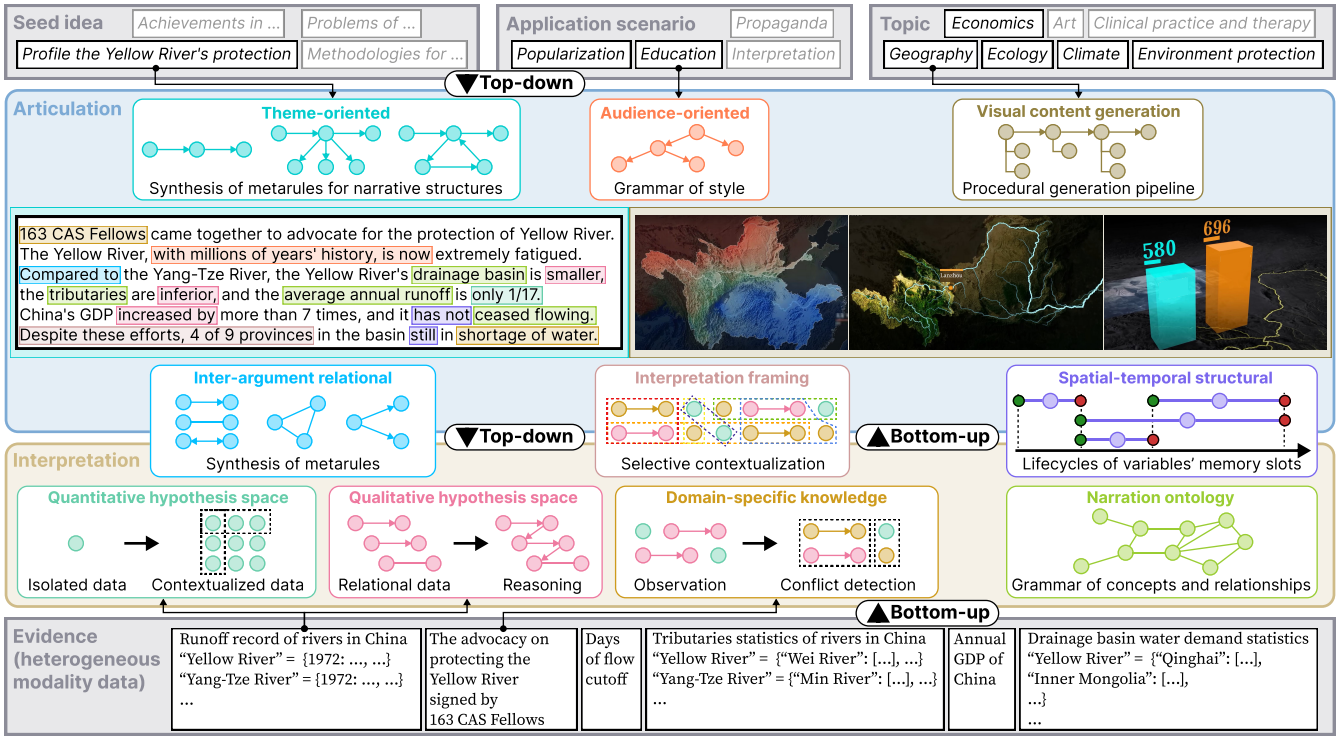


Figure 1: **The architecture of data-driven storytelling with hierarchical constraints.** We present intuitive illustrations of the representations with blocks (see Sec. 3.3). The colors highlighting textual narratives and visual illustrations are encoded according to their respective constraints.

data abstraction, or even more specialized types according to domain knowledge. Additionally, constraints such as domain-specific knowledge and grammar for audience-oriented adaptation can be tacit knowledge held by domain experts [2], making them difficult to be represented for computation [44]. With these challenges unaddressed, we suggest that the integration of constraints for data-driven storytelling still requires interdisciplinary research efforts.

In this perspective article, we first introduce the architecture of our cognitively inspired framework of constraint representation for data-driven storytelling (Sec. 2). Within this framework, we systematically study the requirements of constraints in different hierarchies, demonstrating with running examples (Fig. 1 and Sec. 3). We also explore the potential of representing constraints with Domain-Specific Languages (DSLs) (Tab. 1), along with the challenge and potential solutions for generalization across different domains and scenarios. Concluding with general discussions (Sec. 4), we hope this paper provides the data-driven storytelling community with a fresh perspective on enhancing both *the creativity of art* and *the preciseness of science* in data story generation.

## 2 Framework Architecture

Our theoretical framework is grounded in the cognitive foundations of human perspective framing [43], aligning with the underlying logic of data-driven storytelling — selectively integrating grounded supporting information to align with the central objective. Specifically, the constraints are defined across two hierarchies: **interpretation** and **articulation**. These hierarchies can be incorporated into two bidirectional pathways that characterize human information processing. The top-down view begins with a specific seed idea, moves to the connections between arguments and visualizations, and finally **interprets** individual pieces of evidence to support the arguments. Conversely, the bottom-up view starts with a piece of quantitative data insight, progresses to the qualitative logical relationships upon the insights, and ultimately returns to the narrations and illustrations that **articulate** these insights.

**Top-down** There are several constraints to incorporate in the generation process from the seed idea to the arguments. We need a constraint with second-order rules to structure the narrative and illustrations, such as parallel arguments or progressive arguments [48]. We also require a constraint with first-order rules to identify the relationships between arguments, such as contrast, enumeration, or analogy [11]. Additionally, a structural variable management mechanism is necessary to maintain logical coherence, such as managing a temporally varied key quantity in the data. Also, a grammar-based constraint is needed to tailor the narrative organization specifically for the target audience, as the same story should be told differently in contexts of scientific popularization versus therapy interpretation. Furthermore, the generation from arguments to evidence should also be constrained. While evidence itself cannot be "generated", the context of evidence can be carefully selected to support the arguments.

**Bottom-up** There are several constraints to incorporate in the extraction process from quantitative data insights to qualitative data insights. First, we need a constraint with flexible hypothesis spaces, equipped with composable functions to test arbitrary statistical relationships between arbitrary pairs or groups of variables, such as condition, correlation, and causality [15]. Upon this layer, we need a constraint with combinatorial logic functions that abstract the quantities to relationships, such as opposites, comparisons, and superlatives. Additionally, we need a constraint that represents domain-specific knowledge to contextualize the data appropriately and detect potential insights. For example, in many cases, anomaly data can be insightful but can only be detected with an understanding of what constitutes normal data. Domain-specific knowledge also maintains the correctness of both quantitative and qualitative analyses, minimizing the possibility of misleading or deceptive in-

terpretations. Furthermore, as the narration must be coherently aligned with the data of interest, the intersection between domain-specific knowledge and the scope of the selected data should also serve as a constraint on the grammar of the narration.

## 3 CONSTRAINT REPRESENTATION

In this section, we specify the taxonomies and functions of the constraints targeting precise data-driven storytelling. To ensure the descriptions are understandable, we first introduce a running example that goes throughout the remainder of this paper (Sec. 3.1). Afterwards, we discuss the hierarchies to define the constraints (Sec. 3.2), and explore their representation approaches (Sec. 3.3).

### 3.1 An exemplar data story

We carefully select a relatively ideal data story, *The Protection of the Yellow River*[1], as our running example. The rationale for selecting this piece of work comes in three-fold: (i) all data insights are properly interpreted in the context of corresponding background knowledge; (ii) the narrative structure and style are adaptively optimized for both the topic itself and the target audience's expectations regarding a data story for education and advocacy; (iii) the visual illustrations upon the narrations are coherent and the photorealistic render visualizations are impressive. The story comes as follows.

> In 1997/98, 163 Chinese Academy of Sciences Fellows came together to advocate for the protection of Yellow River, when it failed to flow into the sea for nearly two-thirds of the year. At its worst, 90% of the lower reaches of the river dried up completely. The Yellow River, with millions of years' history, is now extremely fatigued.
>
> Compared to the Yang-Tze River, the Yellow River's situation is even more dire. The drainage basin is significantly small, the tributaries are significantly inferior, and the average annual runoff is a mere 58B$m^3$, compared to the Yang-Tze River's 975.5B$m^3$, a ratio of just 1/17.
>
> A significant portion of the Yellow River's water, about 60%, comes from just 28% of the drainage basin above Lanzhou. The large water demand of 69.6B$m^3$ for industry further strain the river, against its 58B$m^3$ runoff.
>
> Consequently, the first flow cutoff occurred in 1972 and became frequent over the next two decades. By 1997, the flow cutoff problem had become increasingly severe, culminating in an unprecedented 226 days without flow.
>
> [Measures have been implemented ...] From 1999 to 2022, China's GDP increased by more than 7 times, and the Yellow River has not ceased flowing.
>
> However, the serious water and soil erosion of the Loess Plateau has made the Yellow River the river with the highest sediment content in the world.
>
> [Measures have been implemented ...] By 2022, compared to 1949, the vegetation coverage of the Loess Plateau had increased by 59% to 65%, and the amount of sediment entering the Yellow River had decreased by nearly 88%, from 1.6B tons to 193M tons.
>
> Despite these efforts, four of the nine provinces in the Yellow River drainage basin still suffer from extreme or severe water shortages, suggesting that the protection of the Yellow River still demands continuous effort.

The data story is compressed due to space limits. We keep all narrations based on data insights and discard pure descriptive text for succinctness, *e.g.*, the descriptions of measures and actions.

[1]Please refer to the original video posted at `https://youtu.be/jq_74a58wtA?si=9LXODKpOipbmpWsS` (with English and Chinese captions)

### 3.2 The hierarchies of constraints

We define the constraints across the hierarchies of data-driven storytelling as the primary taxonomy. The hierarchical structure is intrinsically in line with the current consensus of the community [9].

#### 3.2.1 Seed idea to articulation

The seed idea is the starting point and the overall guidance of a data story. The seed idea reflects the internal factors, such as societal background, motivation, position, and perspective of the author regarding the topic [29]. For a high-quality data story, the narratives should be coherently subjected to the seed idea. Additionally, the narratives should also be adapted to external factors, such as the requirements of the presenter, the requirements of the target audience, the context of the presentation, and the objective of the presentation.

In our example, the seed idea is likely to be *"To profile the past, the present, and the future of the Yellow River's protection"*, thus the narrative structure is organized as a sequence of *problem-measure-result* triplets along the temporal dimension, within linear logic. If the seed idea changes to *"The current problems of the Yellow River's protection"*, *"The achievements in the Yellow River's protection"*, or *"Methodologies for the Yellow River's protection"*, the narrative structure would vary accordingly. We refer to this constraint as the **theme-oriented constraint**.

Our exemplar data story is a video for popularization education, thus the narrative comes in a teaching style. If the presentation scenario changes to official propaganda or professional interpretation, the narrative style will vary accordingly. We refer to this constraint as the **audience-oriented constraint**.

As the topic covers disciplines such as geography, natural environment protection, climate, ecology, and economy, the visual illustrations are expected to be photorealistic. Specifically, data stories regarding geography and environment are also expected to visualize the data insights on 2D maps, 3D terrain models, or 3D globe models, and the visualization elements should come in a scientific style, *i.e.*, the animations are physically real. For videos illustrating data stories of other topics, such as operational process instructions, historical event interpretations, or public policy advocacy videos, the languages of visual expression vary accordingly. We refer to this constraint as the **visual content generation constraint**.

#### 3.2.2 Articulation to interpretation

Derived from and constrained by the seed idea, arguments are organized following a narrative structure. The arguments are not isolated at all — instead, they are supported by the interpretations of evidence, and also, are then articulated in the global structure to fit with the context of each other [43]. Similar to the former hierarchy, this hierarchy also only cares about the constraints derived from the high-level internal and external factors, rather than the data.

The connections between arguments specify the narrative structure. Different narrative structures determine distinct styles of articulation between arguments. In our example, the relationship *"contrasting"* is a representative use case. To emphasize the severe condition the Yellow River was facing around 1997, by contrasting with the condition of the Yang-Tze River; and also, to highlight the achievements made by the protection actions, by contrasting key indicators before and after the treatments. We refer to this constraint as the **inter-argument relational constraint**.

Each argument must be supported by a piece of evidence. Argument and evidence are bridged by interpretations [21], which selectively construct different contexts as a premise, thus drawing distinct conclusions given the same set of observations. Most trickily, all aspects of the conclusions can be logically self-consistent. A typical case comes in the last sentence of our exemplar story — given the same observation, which can be objectively described as *"among the 9 provinces in the Yellow River drainage basin, 5*

*provinces are free from water shortage while 4 provinces are suffering extreme or severe water shortages"*. Interestingly, in the story, the context *"despite these efforts, 4 of 9 provinces are suffering..."* modifies the audience's uniform prior on water shortage and constructs a context to increase their expectation of water shortage, thus enhancing the significance to the continuously protect the Yellow River. Through this frame of perspective, the interpretation of a piece of evidence becomes subjective, and it is logically correct. Let us consider another context, where we say *"5 of 9 provinces are free from water shortages"*, the interpretation can be totally different — we may be relieved that more than half of the provinces in the Yellow River drainage basin do not have the problem of water shortage. Thus, such interpretation is subjected to the arguments' roles, which is further determined by the author's perspective. We refer to this constraint as the **interpretation framing constraint**.

Along the development horizon of the story, some key concepts, patterns, and illustrations can be referenced by different arguments. They are thought to be invariant throughout the development process of the story, or they should be manipulated in a closure way, where each modification is coherent to its interaction with other concepts or patterns in the story. In our example, two distinct arguments call the concept *"annual runoff of the Yellow River"*, which are required to keep in mind that these two concepts are identical without any variants, sharing the same quantity. We refer to this constraint as the **spatial-temporal structural constraint**.

#### 3.2.3 Evidence to interpretation

Interpretation from evidence is the foundation of a data story. In contrast with the former two hierarchies, which are defined from the top-down view, this hierarchy is defined from the bottom-up view, where we are detecting insights from the data, and only from the data, regardless of the high-level factors.

Elementary data insights are entailed in quantities. However, the absolute quantities of individual variables may not be insightful. Insights come from the combinations of and the associations between different variables. In our example, the fact that *"the average annual runoff of the Yellow River is 58Bm$^3$"* is not insightful at all, since it is sufficiently a large number for ordinary people. However, contextualizing this piece of data together with *"the average annual runoff of the Yang-Tze River is 975.5Bm$^3$"* makes the analysis insightful — although people cannot make precise perception of the absolute quantities of 58B and 975.5B, *i.e.*, "how large they are", the distinction between the two distributions is trivial — significantly, the former is much smaller than the latter, conveying the idea of the Yellow River's inferior condition to the audience. Such insights come from the appropriate construction of hypothesis spaces over statistics variables, which are subjected to the feasible combinatorial spaces of individual variables. We refer to this constraint as the **quantitative hypothesis space constraint**.

Logical insights are built upon the data insights and come from reasoning over logical relationships between variables for reasoning, which take a further step from evidence to interpretation. In our example, *"the annual water demand for industry is 69.6Bm$^3$, larger than the annual runoff of the Yellow River"* entails a logic proposition *"Industry is extremely water demanding"*. Given another proposition *"Beijing-Tianjin-Hebei metropolitan area is a major industry area of China and relies on the Yellow River for water"* and the general commonsense *"Industry is the major driving force of GDP"*, we can easily find the outcome *"China's GDP increased by more than 7 times while the Yellow River has not ceased flowing"* quite interesting and persuasive, demonstrating the effect of practical measurements such as optimizing water allocation policy. These insights come from multiple-step logical deduction, induction, and abduction, which are subjected to feasible hypothesis spaces spanned by logic operators. We refer to this constraint as the **qualitative hypothesis space constraint**.

In addition to those insights that can be detected by contextualization, another family of insights seems not insightful even in the context of associated variables or logic chains. They usually come from *anomaly*. Although seeming normal, those data or logic may become abnormal when contextualized in domain-specific knowledge, which is sometimes not as trivial as general commonsense. In our example, *"60% water of the Yellow River comes from just 28% of the drainage basin above Lanzhou"* is insightful only given the background knowledge of the positive correlation between water runoff and the area of drainage basin. Similarly, in *"163 Chinese Academy of Sciences (CAS) Fellows came together to advocate for the protection"*, the quantity 163 will seem insightful only with the background knowledge that there were about 300 CAS Fellows in total by 1998. Such insights, coming without intrinsic statistical or logical patterns of interest themselves, can only be detected under the constraint of background knowledge. We refer to this constraint as the **domain-specific knowledge constraint**.

These constraints for interpretation we have analyzed will be integrated into the narrations ultimately. Thus, along with the high-level factors, the ontology extracted from the data also constrains the generation of narrations. Specifically, ontology defines the terminologies and the relationships between them, according to the domain reflecting the data. Our example comes with the domain-specific terminologies, such as *"drainage basin area"*, *"the number and length of tributaries"*, *"average annual runoff"*, and *"days of flow cutoff"*, and also the relationships between them, such as *"the significant positive correlation between the scale of tributaries and average annual runoff"*, *"the weak negative correlation between average annual runoff and days of flow cutoff"*, and *"a tributary is counted in the drainage basin of its mainstream"*. This ontology is then integrated into the grammar of narratives to ensure that there are neither missing nor redundant terminologies and relationships regarding the source domain. Namely, the grammar thereby *compactly* tailors the knowledge and insights behind the evidence. We refer to this constraint as the **narration ontology constraint**.

### 3.3 DSL as constraint representation

Given the taxonomy of the constraints w.r.t. their requirements, we explore how to *represent* these constraints. Summarizing the properties and requirements of the ten constraints, we suggest that structural representations tailored for the constraints' definitions, namely DSLs, may become the appropriate approaches. We also showcase the utilities of DSL-based constraints in our exemplar (Tab. 1).

#### 3.3.1 The rationale of DSL-based constraints

The constraints we have discussed indeed share some commonalities: they are required to be complete when verifying the generated content, avoiding open-ended cases; they must be consistent as the story progresses, avoiding "magical modifications"; they are also expected to precisely encode knowledge of various granularity. These properties — completeness, consistency, and multiple-granularity — naturally fit the advantages of symbolic representation, in particular through programming languages, the symbolic representation with the highest expressive capacity [47, 10, 18, 38].

Unfortunately, a major part of programming languages, the General-Purpose Languages (GPLs) such as C/C++, Python, and Java, may not be the best candidates to represent the constraints — programs written in those languages can become extremely complicated, thus hindering both machine program generation and human understanding [33, 14]. As GPLs maintain a general set of syntactic and semantic features to cover all aspects of usage, GPL programs for a narrower set of usage are also built from those features from a relatively low level of abstraction. In contrast, DSLs that consider a specific set of usage only introduce features tailored for the target domain, thus enjoying a higher level of abstraction. This results in simple programs only with features echoing the domain-specific

Table 1: **Demonstrations of constraint representations**

| Constraint | Showcase in our exemplar | DSL-based constraint representation |
|---|---|---|
| Theme-oriented | *The motivational seed idea of the data story is to profile the past, the present, and the future of the Yellow River's protection.* | `time_linear(P, Q, R) :- past(P), present(Q), future(R).`<br>`protection(X, Y, Z) :-`<br>`problem(X), measure(Y),`<br>`result(Z).`<br>`nested(time_linear, protection).` |
| Audience-oriented | *The data story comes for popularization education, thus the narrative should come in a teaching style, targeting for specialized audience group.* | `teaching_narrative ::= engage \|`<br>`explore \| explain \| ...`<br>`engage ::= "Let us" \| ...`<br>`...` |
| Visual content generation | *The topic is about geography, natural environment protection, climate, ecology, and economy, the visual illustration is expected to be photorealistic.* | `3D_terrain = SurfaceModeling(`<br>`size = [L: 100, W: 100, H: 50],`<br>`camera = [X: 20, Y: 30, Z: 45],`<br>`base_model = "YellowRiver.asm",`<br>`...)` |
| Inter-argument relational | *Compared to the Yang-Tze River, the Yellow River's drainage basin is significantly smaller, the tributaries are significantly inferior, and the average annual runoff is significantly lower.* | `YR_vs_YT(X, Y) :-`<br>`duplicate(contrast(X, Y)).`<br>`YR_vs_YT(YR, YT) :-`<br>`smaller_basin(YR, YT),`<br>`inferior_tributary(YR, YT),`<br>`lower_runoff(YR, YT).` |
| Interpretation framing | *Despite these efforts, four of the nine provinces in the Yellow River drainage basin still suffer from extreme or severe water shortages.* | `water_supplying ::= sufficient \|`<br>`shortage \| severe_shortage`<br>`interpret(water_supplying ->`<br>`!sufficient).` |
| Spatial-temporal structural | *The average annual runoff is a mere 58Bm$^3$... The large water demand for industry further strain the river, against its 58Bm$^3$ runoff.* | `YR_runoff = new memory slot X.`<br>`...`<br>`YR_industry = new memory slot Y.`<br>`cmp(YR_runoff, YR_industry) :-`<br>`cmp(call(X), call(Y)).` |
| Quantitative hypothesis space | *The average annual runoff is a mere 58Bm$^3$, compared to the Yang-Tze River's 975.5Bm$^3$, a ratio of just 1/17.* | `dyadic_relation(X, Y) ->`<br>`stat_prop(X, Y).`<br>`YR_runoff = [], YT_runoff = [].`<br>`stat_prop(YR_runoff, YT_runoff).` |
| Qualitative hypothesis space | *From 1999 to 2022, China's GDP increased by more than 7 times, and the Yellow River has not ceased flowing.* | `corr(GDP, YR_industry).`<br>`corr(YR_industry, YR_cutoff).`<br>`map([GDP.pre, GDP.post] ->`<br>`[YR_cutoff.pre, YR_cutoff.post]).` |
| Domain-specific knowledge | *A significant portion of the Yellow River's water, about 60%, comes from just 28% of the drainage basin above Lanzhou.*<br>*In 1997/98, 163 Chinese Academy of Sciences Fellows came together to advocate for the protection of Yellow River.* | `corr(water_pc, basin_pc).`<br>`water_pc = 60, basin_pc = 28.`<br>`abnormal(corr(water_pc,`<br>`basin_pc)).`<br>`num_Fellow = 300.`<br>`abnormal(!significance(num_Fellow,`<br>`163)).` |
| Narration ontology | *The terminologies and the relationships between them, according to the domain reflecting the data, should be integrated into the narratives ultimately.* | `terminology ::= drainage_basin \|`<br>`basin_area \| annual_runoff \| ...`<br>`relationship ::=`<br>`corr(basin_area, annual_runoff)`<br>`\| in_drainage_basin_of(tributary,`<br>`mainstream) \| ...` |

requirements, such as domain knowledge, which are easy to synthesize by machines, and are also easy to learn, understand, and use by domain experts without programming experience.

The constraints for data-driven storytelling are heterogeneous, such as representing structures, knowledge, models, and calculations, respectively. In addition, some of them are tacit knowledge of domain experts, which requires fine-grained domain-specific knowledge injection. Consequently, they are appropriate to be represented with DSLs — one DSL for one specific type and one specific domain. Such considerations are generally acknowledged. For example, there is a variety of DSLs developed for creating diverse visualizations targeting specific domains efficiently [32].

In the following paragraphs, we discuss the utilities of DSL-based constraints, according to the abstraction levels they are work-

ing on [1]: (i) **syntax-level constraints** care about the structures of structural representations, such as trees, graphs, and cycles, and also the mechanisms of symbolic calculation, such as unary, dyadic, and multiple operators; (ii) **semantics-level constraints** consider the exact meanings of variables, operators, and functions, echoing the ontology of the reference model from the source domain; (iii) **execution-level constraints** synthesize and interpret programs dynamically, namely linking and contextualizing unit components in the programs and verifying their global consistency.

#### 3.3.2 Syntax-level constraint

Among the ten constraints, theme-oriented constraint and inter-argument relational constraint are working on the syntax level. Their major utilities are generating meta-level templates, *a.k.a*. metarules [13], for defining a feasible space and permissible operations upon the space. Afterwards, the narrations are generated by grounding the space without conflicts with the constraint.

Theme-oriented constraint shapes the narration structures, which are usually tree-based or graph-based. For different seed ideas, we may exploit different narrative structures to maximize their communication bandwidths. According to the theories of arguments [48], we may use linear structure for temporal-related contents, multi-headed structure for spatial-related contents, and non-monotonic logical structure for contents with subjective judgments. We can also locally nest different types of logic for mixed purposes. Similarly, inter-argument relational constraint implements the narration structures. There are sequences for progressive arguments, branches for alternative arguments, recursions for repeatedly updating arguments, and parallels for contrasting arguments.

#### 3.3.3 Semantics-level constraint

Among the ten constraints, audience-oriented constraint, visual content generation constraint, interpretation framing constraint, and narration ontology constraint are working on the semantics level. Their major utilities are ensuring the specific meanings of the generated content to be consistent with general commonsense and domain-specific knowledge, and also to be complete for use.

Audience-oriented constraint and narration ontology constraint both shape the generation space of textual narrations. The former comes from a higher level, *i.e*., external factors of the data story, while the latter comes from a lower level, *i.e*., the given evidence. These two constraints are usually represented as deterministic or probabilistic Context-Free Grammars (CFGs), controlling the style and scope of the narratives [18]. To constrain the style, we leverage the combination rules of specialized keywords, sentence structures, and transitions between sentences. For example, we use engaging transitions like *"... Now it is the turn to do it together ..."* in data stories for education; we exploit keywords with sense-of-belonging, such as *"our community"* in data stories for advocacy; and we employ sentences with superlative statements, such as *"... is the highest/ best of ..."* in data stories for propaganda. To constrain the scope, we map the ontology from the corresponding domain of the evidence to the abstract grammar of narratives, both completing the concepts inside the scope and removing those out of the scope.

Interpretation framing constraint can be viewed as a probabilistic CFG, which is a tree with intermediate nodes spanning the *world*, *i.e*., all possible candidate meanings, of specific concepts or events [37]. The world is the context for interpretation, mostly coming from the general commonsense. The process of framing is reweighing the candidates belonging to the same world.

Visual content generation constraint is the *model* for model-based generation. For data stories on topics related to natural sciences, clinical practices, and engineering, visual illustrations are often required to be physically real. Despite the current advancements of Artificial Intelligence Generated Content (AIGC) techniques, generating photorealistic videos that are physically real is still challenging because elementary physical properties, such as the spatial-temporal dynamics, are latent and long-tail distributed in datasets, implying that they may not be correctly extracted during training. Instead, they may be induced as shortcuts. This is the drawback of model-free generation by nature. Consequently, we may leverage the physical constraints provided by DSLs for 3D modeling, such as Blender[2] — we can synthesize Blender code for programs rendering a 3D model, precisely edit the model by modifying the program, and explicitly constraint the model with physical properties from *the first principle*. Furthermore, for data stories on topics related to history, public policy, and business, visual illustrations are usually expected to be animated drawings rather than photorealistic videos. However, those animations can be a sophisticated combination of components, such as spatially articulated objects and temporally varied scenes, where the similar-sample-based end-to-end generative models may be struggling [45]. The straightforward solution also leverages a rule-based model generating local states, layout and rendering configurations, topological relationships, and temporal state transitions of the components.

#### 3.3.4 Execution-level constraint

Among the ten constraints, spatial-temporal structural constraint, quantitative hypothesis space constraint, qualitative hypothesis space constraint, and domain-specific knowledge constraint are working on the execution level. Their major utilities are generating hypothesis spaces dynamically and verifying them in real-time.

Quantitative and qualitative hypothesis space constraints are dynamically generating hypothesis spaces to detect any possible data insights, *i.e*., evidence of interest. A piece of interesting evidence with insight comes from the shift from one way of explanation to another, akin to the moment of representation shift in problem-solving [4, 19, 36]. Analogous to the representation of a problem, which determines *selecting what information of the problem into solving it*, our hypothesis spaces consider *putting which pieces of evidence together to explain them*. For example, the data on *"the Yellow River's annual runoff in 1998"* possesses multiple contexts, such as the 1998 annual runoff of other rivers in China, the Yellow River's annual runoff in other years, the industry water demand of the Yellow River drainage basin in 1998, and the annual runoff of the upper Lanzhou part of the Yellow River. The hypothesis space indicates the structure of observable variables, *e.g*., dyadic or triadic, and the type of verification, *e.g*., statistical testing functions or logical reasoning functions. Thereby insights can be detected at the shifting from isolated data to contextualized data. Domain-specific knowledge constraint works with quantitative and qualitative hypothesis space constraints. While the latter two put evidence together in different frames, the former puts grounded background knowledge, either procedural or declarative knowledge, together with evidence, to create insightful context shifting.

Spatial-temporal structural constraint is a robust infrastructure for demonstrating the detected insights in the narrations. The variables are called in distinct parts and by various means, necessitating the maintenance of numerical integrity and logical consistency. On the temporal dimension, the lifecycles of the invariant are tracked to avoid inconsistency between different calls. Also, variables being modified with specific calculations are constrained with preconditions and postconditions for state transition tracing. On the spatial dimension, the relative changes of variables are tracked, such as duplication of variables, chaining among triple variables, inversion between dual variables, and recursion on multiple variables.

## 4 General discussions

In this perspective, we propose integrating constraints into the automatic generation of data stories to facilitate the creativity in storytelling alongside the preciseness in data analysis. We investigate

[2] https://www.blender.org/

the requirements of these constraints and explore the possibility of representing them through DSLs based on a realistic example. It is important to note that the proposed taxonomy of constraints may not be entirely mutually exclusive and collectively exhaustive — there may be other specific constraints that are significant in different data stories and cannot be perfectly categorized within our definition of constraints. Our primary aim is to provide a structured framework for the data-driven storytelling community, which may, in turn, inspire the development of a fine-grained taxonomy of constraints and their corresponding implementation techniques.

### 4.1 Integrating constraints into the current workflow

The data-driven storytelling community has made significant efforts in developing powerful tools for creating data stories. Currently, there are two schools of thought on the generative models — multi-stage pipelines and end-to-end approaches. We propose that constraints should be incorporated into both approaches, through implicit and explicit methods, respectively. In multi-stage pipelines, different categories of constraints can be mapped to corresponding stages [26] — such as analysis [50], planning [54], implementation [49], and communication [16] — and the modules within these pipelines can be modified according to these constraints. Additionally, the generated content of different modules can be verified against the relevant constraints, thereby implicitly integrating constraints into the workflow. For end-to-end approaches, which feature a higher degree of integration across the entire workflow [30], constraints can be explicitly added by appending a constraint layer to the final output of the tools and verifying the generated content against these constraints. In this way, we outline a framework for integrating constraints into the generation workflow, which may inspire further research on refining the individual tools within the workflow with constraints at a more granular level.

### 4.2 Automating the entire workflow with constraints

Our ultimate goal is to automate the entire workflow of generating data stories, necessitating the automatic synthesis of constraints rather than manual specification. Code generation techniques facilitate the automatic synthesis of constraint programs [53, 27], given simple instructions, requirements, or textual narrations to be verified. Subsequently, the satisfaction of these constraints is verified through language features, such as answer set planning over logic programs [28], which has been applied to constrain visualizations with design theories [34]. Although this is a rudimentary approach to utilizing constraints, it treats constraint program verification as *"first-class citizens"*, ensuring determinism. All uncertainties, ambiguities, and factual errors introduced by generative models with non-deterministic nature, such as the hallucination of Large Language Models (LLMs), are subject to the top-level constraints.

Consequently, this framework preserves inherent freedom of the random creativity characteristic of AIGC, while simultaneously ensuring that this creativity operates within a secure environment. We hope this straightforward yet self-consistent framework can serve as an accessible starting point for exploring the synthesis of constraints through active interaction with generative models. Indeed, the framework includes objective, subjective, and context-dependent constraints. Quantitative and qualitative hypothesis space constraints, spatial-temporal structural constraint, and visual content generation constraint are exactly objetive constraints, reflecting data and the physical world. In contrast, theme-oriented constraint, audience-oriented constraint, and interpretation framing constraint are relatively subjective, influenced by human preference. Additionally, domain-specific knowledge constraint and narration ontology constraint are context-dependent, sensitive to the exact scope of the evidence and the target story. We would like to clarify that the implementation of those heterogeneous constraints according to different requirements, such as transforming domain-specific knowledge into corresponding constraints, is beyond the scope of this perspective article. Nonetheless, explicitly disentangling objective, subjective, and context-dependent constraints within our proposed framework and exploring their respective implementations represents a significant direction for future research.

### 4.3 On the generality of constraints

While it is theoretically possible to automate the entire workflow for generating data stories, a crucial challenge remains: full automation is contingent upon the availability of predefined constraint sets, *i.e.*, the DSLs for constraint representation. However, the origin of these DSLs poses a problem, as they are not readily available like off-the-shelf programming languages. In current practices, most DSLs are manually designed through the collaborative efforts of computer scientists and domain experts, a process that is both time-consuming and costly. This may be acceptable for specific applications requiring only a single DSL library, as DSL design is a *once-and-for-all* endeavor there. Unfortunately, DSLs for representing constraints in data stories span multiple categories, diverse requirements, and an ever-expanding range of domains. For instance, within the ten constraint categories, there exist a vast array of potential DSL instances. For the audience-oriented constraint, the DSL syntax must be tailored to one specific audience group; for the domain-specific knowledge constraint, the DSL semantics must encode the background knowledge of a particular domain of expertise; and let alone the narration constraint, the grammar of the DSL must be designed on-the-fly based on the available evidence. Although it is conceivable that we derive a comprehensive set of constraints covering all potential domains, namely the so-called "one-size-fits-all general constraint", such an endeavor would result in a constraint system of prohibitively complexity, rendering it intractable for both machine and human end-users.

The highly varied and frequently evolving demands for DSLs are difficult to meet through human effort alone. Even if we manage to manually craft these DSLs, the progress in automated data story generation would be undermined — we would merely be shifting human labor from one part of the workflow to another, even potentially increasing the overall labor required. Consequently, we find ourselves in a dilemma: GPLs, which easily accessible, are unsuitable for representing constraints due to their overwhelming complexity, whereas DSLs, which simplify specialized language features, inherently lack generalizability across different domains. To address this dilemma, rather than waiting for a universally applicable constraint to emerge, a more practical solution might involve automating the design of DSL-based constraints.

This solution is both feasible and evaluable. By adopting the AutoDSL approach [42], which combines bottom-up data-driven approaches and top-down principle-derived methods, we can automatically create DSL-based constraints for data-driven storytelling based on relevant materials and design principles [46]. The resulting DSLs can be evaluated both quantitatively and qualitatively [14]. Quantitative evaluation checks the mapping from ontology elements in the reference model, *i.e.*, concepts and relations in the domain corpus, to DSL constructs of constraints; while qualitative evaluation takes the design guidelines of DSL as questions for assessing the DSL-based constraints, from a user-centric perspective. This joint pipeline of design and evaluation leads to a promising future where DSL-based constraints are designed automatically, AIGC tools produce the necessary content and assets for data stories, and the constraints are synthesized and verified automatically. The integration of these approaches will enable content creators to script and implement their data stories more seamlessly.

### 4.4 Valuing humans in data-driven storytelling

Concerns may arise regarding the fully automation of data story generation and its potential severe impact on the content creation

ecosystem. It appears that integrating constraints with generators could bridge the gap between creativity and preciseness, potentially marginalizing content creators. However, humans remain indispensable even in a future where constraints are fully realized. Firstly, constraints are not generators. While constraints define a feasible space for generation, generators determine which specific points within the space are sampled as the generated content. This indicates that the output space of AIGC tools remains significantly larger than the ideal output space that aligns with content creators on latent dimensions, such as aesthetic and ideological considerations. This disparity underscores the necessity for human-machine collaboration tools [26]. Moreover, constraints can be latent. Even with automated DSL design tools, not all constraints can be specified purely based on domain corpora. Some constraints require tacit knowledge from domain experts, necessitating a human-in-the-loop approach. Lastly, neither generators nor constraints can substitute for higher-level cognitive processes involving human factors, such as comprehensing, interpreting, and evolving the *meaning* of data stories for humanity, and consequently, the metaphysical planning of seed ideas. Indeed, AIGC tools with constraints may merely alleviate human content creators from elementary technical tasks, thereby allowing them to concentrate on intention alignment, knowledge externalization, and metaphysical thinking.

## Acknowledgments

This work has been partially supported by RGC GRF Grant 16210722. The authors wish to thank Leixian Shen, Liwenhan Xie, Xian Xu, Hanlu Ma, Yanna Lin, Zhonghua Sheng, Shuchang Xu, Yuying Tang, Lin Gao, and Leni Yang for helpful discussions.

## References

[1] H. Abelson and G. J. Sussman. *Structure and interpretation of computer programs*. The MIT Press, 1996.

[2] G. Abend. The meaning of 'theory'. *Sociological Theory*, 26(2):173–199, 2008.

[3] R. Alharbi, O. Strnad, L. R. Luidolt, M. Waldner, D. Kouřil, C. Bohak, T. Klein, E. Gröller, and I. Viola. Nanotilus: Generator of immersive guided-tours in crowded 3d environments. *IEEE Transactions on Visualization and Computer Graphics*, 29(3):1860–1875, 2021.

[4] P. M. Auble, J. J. Franks, and S. A. Soraci. Effort toward comprehension: Elaboration or "aha"? *Memory & Cognition*, 7(6):426–434, 1979.

[5] R. W. Bybee. Scientific inquiry and science teaching. *Scientific inquiry and nature of science: Implications for teaching, learning, and teacher education*, pp. 1–14, 2006.

[6] K. K. Cetina. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press, 1999.

[7] Q. Chen, Z. Li, T.-C. Pong, and H. Qu. Designing narrative slideshows for learning analytics. In *2019 IEEE Pacific Visualization Symposium (PacificVis)*, 2019.

[8] Q. Chen, F. Sun, X. Xu, Z. Chen, J. Wang, and N. Cao. Vizlinter: A linter and fixer framework for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):206–216, 2021.

[9] F. Chevalier, M. Tory, B. Lee, J. van Wijk, G. Santucci, M. Dörk, and J. Hullman. From analysis to communication: Supporting the lifecycle of a story. In *Data-Driven Storytelling*, pp. 151–183. AK Peters/CRC Press, 2018.

[10] N. Chomsky. *Syntactic Structures*. Mouton de Gruyter, 1957.

[11] J. G. Crammond. The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication*, 15(2):230–268, 1998.

[12] E. Dubourg and N. Baumard. Why imaginary worlds? the psychological foundations and cultural evolution of fictions with imaginary worlds. *Behavioral and Brain Sciences*, 45:e276, 2022.

[13] W. Emde and C.-R. Rollinger. The discovery of the equator or concept driven learning. In *International Joint Conference on Artificial Intelligence*, 1983.

[14] M. Fowler. *Domain-specific languages*. Pearson Education, 2010.

[15] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[16] B. D. Hall, L. Bartram, and M. Brehmer. Augmented chironomia for presenting data to remote audiences. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 2022.

[17] W. Heisenberg. Physics and philosophy: The revolution in modern science. *Physics Today*, 11(9):36, 1958.

[18] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Longman Publishing Co., Inc., 1996.

[19] J. Kounios and M. Beeman. The aha! moment: The cognitive neuroscience of insight. *Current Directions in Psychological Science*, 18(4):210–216, 2009.

[20] D. Kouřil, O. Strnad, P. Mindek, S. Halladjian, T. Isenberg, M. E. Gröller, and I. Viola. Molecumentary: Adaptable narrated documentaries using molecular visualization. *IEEE Transactions on Visualization and Computer Graphics*, 29(3):1733–1747, 2021.

[21] T. S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press: Chicago, 1970.

[22] B. Latour. *Science in action: How to follow scientists and engineers through society*. Harvard University Press, 1987.

[23] B. Latour and S. Woolgar. *Laboratory life: The construction of scientific facts*. Princeton University Press, 1986.

[24] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale. More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5):84–90, 2015.

[25] H. Li, Y. Wang, Q. V. Liao, and H. Qu. Why is ai not a panacea for data workers? an interview study on human-ai collaboration in data storytelling. *arXiv preprint arXiv:2304.08366*, 2023.

[26] H. Li, Y. Wang, and H. Qu. Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

[27] J. T. Liang, C. Yang, and B. A. Myers. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 2024.

[28] V. Lifschitz. Answer set planning. In *Logic Programming and Nonmonotonic Reasoning: 5th International Conference, LPNMR'99*, 1999.

[29] H. E. Longino. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press, 1990.

[30] J. Lu, W. Chen, H. Ye, J. Wang, H. Mei, Y. Gu, Y. Wu, X. L. Zhang, and K.-L. Ma. Automatic generation of unit visualization-based scrollytelling for impromptu data facts delivery. In *Pacific Visualization Symposium (PacificVis)*, 2021.

[31] M. Lynch. *Scientific practice and ordinary action: Ethnomethodology and social studies of science*. Cambridge University Press, 1993.

[32] A. M. McNutt. No grammar to rule them all: A survey of json-style dsls for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):160–170, 2022.

[33] M. Mernik, J. Heering, and A. M. Sloane. When and how to develop domain-specific languages. *ACM Computing Surveys (CSUR)*, 37(4):316–344, 2005.

[34] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, 2018.

[35] E. Mörth, S. Bruckner, and N. N. Smit. Scrollyvis: Interactive visual authoring of guided dynamic narratives for scientific scrollytelling. *IEEE Transactions on Visualization and Computer Graphics*, 29(12):5165–5177, 2022.

[36] S. Ohlsson. Restructuring revisited: I. summary and critique of the gestalt theory of problem solving. *Scandinavian Journal of Psychology*, 25(1):65–78, 1984.

[37] R. Reiter. On closed world data bases. In *Readings in Artificial Intelligence*, pp. 119–140. Elsevier, 1981.

[38] S. J. Russell and P. Norvig. *Artificial intelligence a modern approach*. Prentice Hall Press, 2010.

[39] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, 2010.

[40] S. Shankar, H. Li, P. Asawa, M. Hulsebos, Y. Lin, J. Zamfirescu-Pereira, H. Chase, W. Fu-Hinthorn, A. G. Parameswaran, and E. Wu. Spade: Synthesizing assertions for large language model pipelines. *arXiv preprint arXiv:2401.03038*, 2024.

[41] L. Shen, E. Shen, Z. Tai, Y. Song, and J. Wang. Taskvis: Task-oriented visualization recommendation. In *EuroVis (Short Papers)*, 2021.

[42] Y.-Z. Shi, H. Hou, Z. Bi, F. Meng, X. Wei, L. Ruan, and Q. Wang. AutoDSL: Automated domain-specific language design for structural representation of procedures with constraints. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[43] Y.-Z. Shi, S. Li, X. Niu, Q. Xu, J. Liu, Y. Xu, S. Gu, B. He, X. Li, X. Zhao, et al. PersLEARN: Research training through the lens of perspective cultivation. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[44] Y.-Z. Shi, M. Xu, W. Han, and Y. Zhu. To think inside the box, or to think out of the box? Scientific discovery via the reciprocation of insights and concepts. *arXiv preprint arXiv:2212.00258*, 2022.

[45] Y.-Z. Shi, M. Xu, J. E. Hopcroft, K. He, J. B. Tenenbaum, S.-C. Zhu, Y. N. Wu, W. Han, and Y. Zhu. On the complexity of Bayesian generalization. In *International Conference on Machine Learning*, 2023.

[46] Y.-Z. Shi, Q. Xu, F. Meng, L. Ruan, and Q. Wang. Abstract Hardware Grounding towards the Automated Design of Automation Systems. In *International Conference on Intelligent Robotics and Applications*, 2024.

[47] A. Tarski. *Introduction to Logic and to the Methodology of Deductive Sciences*. Dover Publications, 1946.

[48] S. E. Toulmin. *The uses of argument*. Cambridge University Press, 1958.

[49] A. Tyagi, J. Zhao, P. Patel, S. Khurana, and K. Mueller. Infographics wizard: Flexible infographics authoring and design exploration. In *Computer Graphics Forum*, 2022.

[50] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):895–905, 2019.

[51] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, and H. Qu. Ai4vis: Survey on artificial intelligence approaches for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5049–5070, 2021.

[52] J. Yang, P. F. Gyarmati, Z. Zeng, and D. Moritz. Draco 2: An extensible platform to model visualization design. In *IEEE Visualization and Visual Analytics (VIS)*, 2023.

[53] D. Zan, B. Chen, F. Zhang, D. Lu, B. Wu, B. Guan, W. Yongji, and L. Jian-Guang. Large language models meet nl2code: A survey. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[54] J. Zhao, S. Xu, S. Chandrasegaran, C. Bryan, F. Du, A. Mishra, X. Qian, Y. Li, and K.-L. Ma. Chartstory: Automated partitioning, layout, and captioning of charts into comic-style narratives. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1384–1399, 2021.