Harnessing spatial representations of image collections in the discovery of cause-and-effect relationships in data

Aljaz Kotnik and Graham Pullan

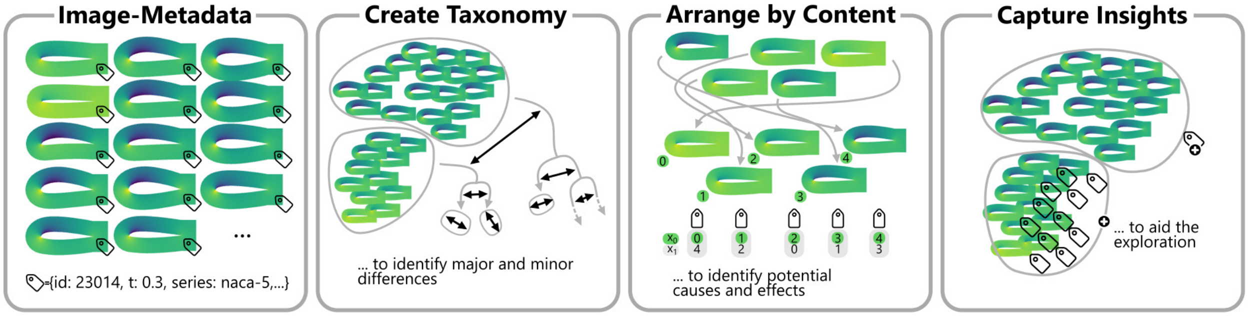
Image-Metadata
...
={id: 23014, t: 0.3, series: naca-5,...}
Create Taxonomy
... to identify major and minor differences
Arrange by Content
0
1
2
3
4
x_0
x_1
0 4
1 2
2 0
3 1
4 3
... to identify potential causes and effects
Capture Insights
... to aid the exploration

Fig. 1. An example image-metadata dataset of air flow around airfoil designs, represented by small multiples. Interactive taxonomy creation, on-screen position harnessing, and insight capture support the discovery and interpretation of cause-and-effect relationships to generate knowledge about the underlying physical mechanisms.

Abstract—Discovering cause-and-effect relationships, the links between input parameters to an experiment or simulation and the observed outcome, is fundamental to the scientific process. An understanding of the underlying mechanisms responsible for these connections is the key transferable learning to be derived. In many cases, data collected to study these mechanisms is best represented by images. The scientist's domain knowledge is crucial for image interpretation and in cases where results contain large numbers of images, the necessary human-data interaction becomes the bottleneck of the analysis and discovery process. To accelerate this step, we propose a conceptual framework to reduce the number of detailed image analyses and comparisons required to gain a full understanding of cause-and-effect relationships. Our approach allows scientists to visually encode their interpretation of an image set by interactively manipulating the on-screen location of the images. To identify cause-and-effect relationships, this spatial arrangement is then correlated with the input and outcome metadata associated with each image. A variety of tools to meaningfully group images, capture knowledge represented by the groups, and establish the representative behavior of a group and its variability, are incorporated to provide guidance to the practitioner and accelerate the process. We illustrate the utility of the framework with examples including classification of images using machine learning, analysis of metal micro-structures in material science and engineering simulations of the flow around airfoils.

Index Terms—Process/Workflow Design ; Physical & Environmental Sciences, Engineering, Mathematics ; Data Analysis, Reasoning, Problem Solving, and Decision Making ; Mixed Initiative Human-Machine Analysis ; Data Clustering and Aggregation

1 Introduction

In many fields of science and engineering, the analysis of image-based data to uncover or explain behavior is a key part of daily workflow. For example, in aerospace engineering the air flow around proposed designs must be simulated and controlled; in medicine, 2D and 3D imaging techniques are a routine part of diagnosis; in meteorology, oceanography, and geography, satellite images quantify changes in the environment. Additional data associated with the image, called 'metadata', can be valuable during this analysis. Metadata might include information about the candidate design or computer simulation configuration, patient history, or temperature measurements on the ground. These image-based datasets are becoming increasingly large and require dedicated tools for efficient analysis.

Fig. 2 shows an abstract view of datasets containing metadata and images, depicting process inputs and outputs. Analysis of cause-and-effect relationships is possible based on pairs of input and output data. Potential causality relationships need to be supported by a valid explanation of the behavior of the underlying mechanisms. In many cases, these are captured in images, either because of convenience, or because the data can best be summarised and presented in a visual way, such as results of engineering simulations. The input and output data can contain several images or metadata variables.

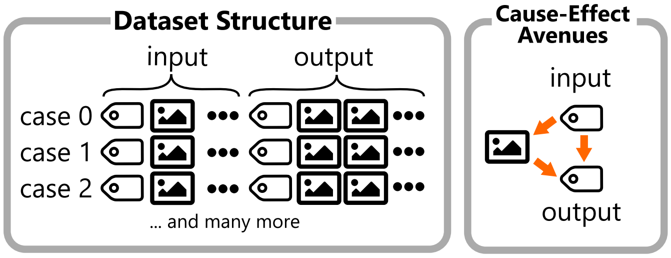

Fig. 2. The structure of the data and the available cause-and-effect relationships. Each case can associate input & output metadata with several images. The output images capture the domain-specific mechanism behavior. The dataset need not have all input/output and metadata/image combinations. The most valuable relationships to discover are between metadata and output images.

In datasets with many metadata variables, there may be several un-

- *Aljaz Kotnik and Graham Pullan are with the Whittle Laboratory, University of Cambridge, UK. Emails: {ak2164, gp10006}@cam.ac.uk*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

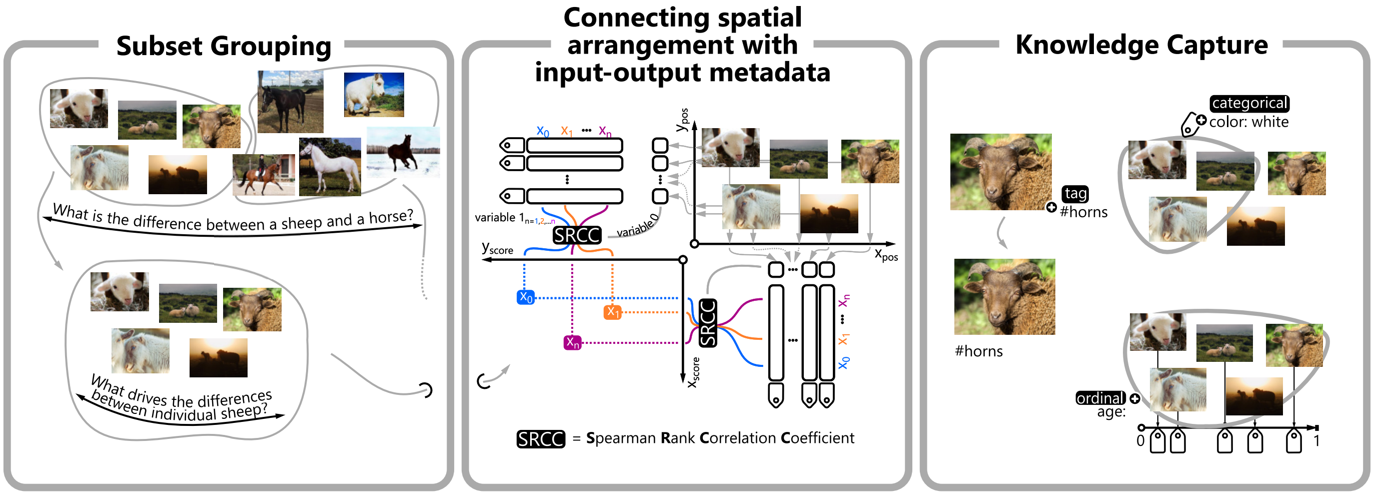
Subset Grouping
What is the difference between a sheep and a horse?
What drives the differences between individual sheep?
Connecting spatial arrangement with input-output metadata
x0 x1 ··· xn
variable 1n=1,2,...n
SRCC
variable0
yscore
ypos
xpos
xscore
x0
x1
xn
SRCC = Spearman Rank Correlation Coefficient
Knowledge Capture
categorical
color: white
tag
#horns
#horns
ordinal
age:
0
1

Fig. 3. The three parts of the analysis process: Subset grouping is used to create taxonomies. The link between spatial arrangement and metadata is used to identify and study possible cause-effect relationships. Knowledge capture allows the user to create additional metadata on-the-fly, thus enabling further analysis.

derlying cause-and-effect relationships to discover, and systematically verifying them can be time consuming. Interactive small multiple visualization allows the user to first establish trends in individual features of the small multiples by rearranging them on-screen. However, any correlation between the spatial arrangement of the small multiples and the associated metadata remains hidden. While metadata variables can be visually encoded to the small multiples by changing their properties (changing the image size, position, border color,...), only a small number variables can be encoded at the same time.

To effectively analyze and verify cause-and-effect relationships, an intuitive, interactive user interface, allowing the user to match the changes of features shown by small multiples with the corresponding metadata trends, is required.

We propose a generic three-part approach for exploring cause-and-effect relationships of small multiples through interactive on-screen spatial arrangement:

Subset grouping. Image-based datasets typically include subset groups whose members are related. Identifying these groups is a key step in the discovery of cause-effect relationships in large image-based datasets. Based on the interactivity features demonstrated by Lekschas *et al.* [7], our approach enables the user to analyse the entire dataset in an efficient manner.

Connecting spatial arrangement to input-output metadata. Inspired by investigators finding patterns in data by arranging physical pictures on tabletops, interactive on-screen spatial arrangement of small multiples is used to encode the user's domain knowledge. The resulting on-screen arrangement can be used to find correlated metadata variables and recommend them to the user for consideration.

Knowledge capture. Lastly, we introduce interactive tagging as a way to capture knowledge by allowing the user to generate additional metadata on-the-fly. These tags, based on observations of the existing data, are then available for use in further exploration of the data.

We illustrate the utility of this approach with examples from analysis and classification of photographs, study of metal micro-structures in material science, and engineering simulations of flows around airfoils.

After a discussion of related work in Section 2, the paper proceeds with a description of the three elements of the framework in Section 3, followed by methods used to support and accelerate the framework in Section 4. Finally, Section 5 demonstrates our implementation of the proposed framework using a walk-through of cause-effect relationship discovery in airfoil design.

In Section 5 an airfoil dataset is introduced and analysed. A demo featuring this data is available at: `https://aljazkotnik.github.io/harnessingspace/`.

2 Related Work

Small multiples. The importance of small multiples as a visual representation for data analysis has been succinctly described by Tufte [10]:

> "At the heart of quantitative reasoning is a single question: Compared to what? Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, small multiples are the best design solution."

Small multiples are a series of similar images intended for comparison that are visualised simultaneously [10]. To facilitate comparison they usually have the same size, and the same range of data where possible and applicable. In science and engineering, small multiples are appropriate for a wide range of data from material science (e.g. steel micrography images [3]), genomics (genome interaction matrices [6]), machine learning image datasets [7], and from computational simulations of candidate designs in engineering [9].

Small multiples interactions. Small multiples form the basis of interactive user interfaces that enable efficient analysis of image-based datasets. In the simple example of the use of small multiples as icon representations of files in a folder, some familar interactions are available. The file metadata (filename, date modified, extension type, etc) can be used to arrange the order of the files, but the user may not be free to position the icons manually. More advanced examples typically involve data-driven on-screen arrangement of small multiples [3, 6, 7]. Some implementations allow the user to manually arrange the small multiples for use with other functionality [6, 7].

When faced with a collection of small multiples, the user may naturally wish to organize collections of images into groups. File icons can be grouped by placing them in a common folder. Lekschas *et al.* have created a general framework, PILING.JS, for the interactive collection of related small-multiples into 'piles' [7]. The aggregation of small multiples into piles can be done via categorical metadata variables, and by on-screen proximity. Metadata variables can be used to control the spatial arrangement of the small multiples and, if the metadata includes an embedding based on image similarity, piles of similar small multiples can be created quickly. Piling is a versatile approach for the user to navigate and organise a dataset.

Tools for small multiples exploration. The image data of each small multiple is high dimensional and, in general, the associated metadata is also high dimensional. Techniques to provide a meaningful two-dimensional embedding for the on-screen arrangement of small multiples have been demonstrated. The UHCS micrography dataset was visualised by embedding the images into 2D using t-SNE [11], Lekschas *et al.* [6] used t-SNE to arrange parts of a genome interaction matrix in 2D, and Lekschas *et al.* [7] used UMAP to embed sketches of necklaces from the Google Quickdraw Sketches dataset.

Summary. The on-screen arrangement of small multiples, and the ability to group them, are key components of interactive interfaces for the manipulation of small multiples. In the present work, we contribute to both of these elements and also introduce methods for the correlation of small multiple spatial location (both as individuals and as groups) with input-output metadata to allow the discovery, and then capture, of cause-and-effect relationships hidden in large image-based datasets.

3 Components of the Framework

The interactive discovery of cause-effect relationships from datasets of images and metadata can be dividing into three tasks: subset grouping; connecting spatial arrangement to input-output metadata; and, knowledge capture. This process is shown schematically in Fig. 3 and we now discuss each step in turn.

3.1 Subset grouping

The objective of the first step is to accelerate the analysis by following a ‘divide and explore’ approach that separates images into meaningful groups. This aids the cause-effect discovery process in two ways: first, the user can correlate a representative of each subset group with the input-output metadata, rather than having to analyse each image separately; second, the smaller variations within a subset group can also be correlated with the metadata to understand the mechanisms driving the detailed variations of the images. The focus is on allowing the practitioner to use their domain knowledge to aid the process as much as possible.

To illustrate the benefit of obtaining a taxonomy of groups, we use the example of a researcher studying the differences between the animals of the Animals-10 image dataset [2]. This dataset consists of images of 10 different animal types, and thus contains differences between species, as well as within a species population. To gain a full understanding of the dataset, the researcher must understand both types of differences, as well as the cause of these differences.

One approach is to compare individual dataset members, and progressively improve the researcher’s understanding of the dataset. When following this approach, all the images must be analyzed in detail to get a complete understanding of the data. Alternatively, the researcher could first establish an appropriate taxonomy of groups of all the dataset members, thus allowing separate analysis of differences between the species, and differences between species members. For example, instead of analyzing a set of 500 horses and sheep, the user can compare a representative horse and a representative sheep, and then analyze 200 horses and 300 sheep separately, within the context of the individual species. Fig. 3 includes examples of questions that may be posed between and within individual taxonomy groups. The separate analysis of smaller groups already requires fewer pairwise comparisons of members to be made, thus accelerating the analysis. Furthermore, the similarity of taxonomy group members allows subtle differences between members to be observed faster.

Using a taxonomy to first classify dataset members is intuitive in the biological setting, but it can be applied more generally to datasets with many comparable members. For example, engineering simulations of air flow around an object can contain effects such as wakes and shock waves, which could be used to form a basis for case classification. Metal alloy micrography images can be grouped based on the observed crystalline structures, as shown in Fig. 4. The defining property of such a taxonomy is that each group must represent a meaningful self-contained set. Many clustering algorithms exist however there is no common definition of a cluster [5]. While clustering algorithms can be used to form groups, the analyst must use their domain knowledge to ensure the resulting clusters are meaningful. Thus, the user’s interactive input is crucial when creating a dataset taxonomy.

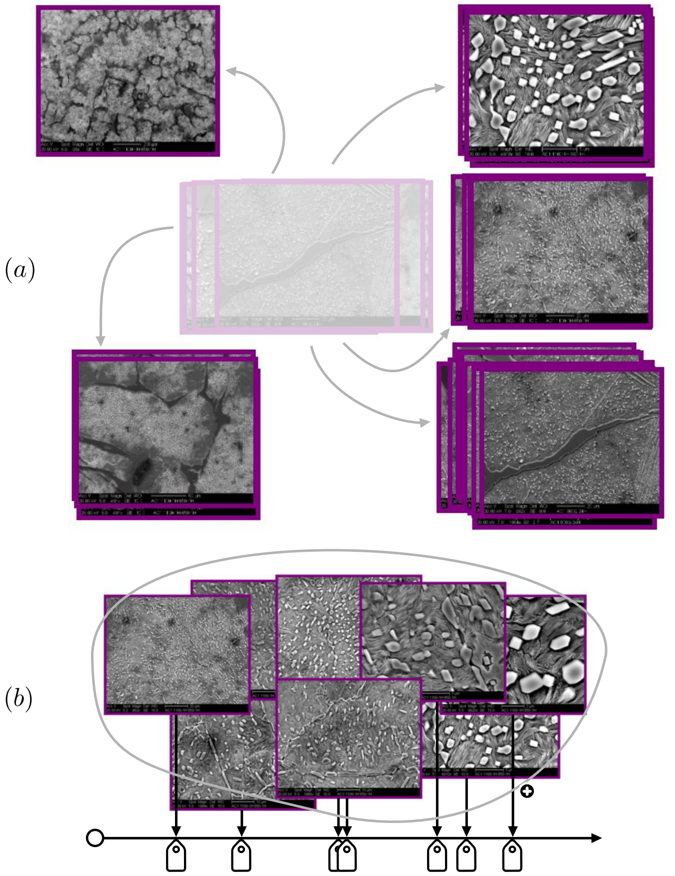

Fig. 4. (a) Creating a taxonomy of the UHCS (UltraHigh Carbon Steel) images [3] based on micro-constituent shape and size allows analysis of micro-structure dependence on annealing process. (b) The images were taken at different magnifications, which needs to be taken into account when comparing images. On-screen arrangement can be used to capture the progression of micro-constituent sizes.

The goal of comparing individual group members is to gain an understanding of the domain specific behavior, captured by the images, responsible for subtle input-output linkages. By arranging the images spatially such that features of interest gradually change, the user can encode a progression that can be correlated to the metadata. Machine learning tools can find embeddings that arrange images by their similarity, but the user is always required to verify the results. As the user is the final arbiter, they should have ultimate freedom to arrange the images meaningfully.

We note that creating a taxonomy is not always a linear process. The user may observe that a different grouping captures the fundamental differences better. This does not mean that the original observations were invalid - they may be of interest in future analyses. Tools to capture the user’s observations for the duration of the exploration are required (see Sect. 3.3).

3.2 Connecting spatial arrangement to metadata

Trends can be found by separately analyzing either the metadata or the small multiples in isolation, but working with both simultaneously allows the cause-and-effect relationships to be identified. The user can ask question of the data in two ways: by mapping from the metadata to the small multiples, or *vice versa*.

The user may want to know how a metadata parameter impacts the features shown by small multiples (question mapping from metadata to the small multiples). HiPiler [6] and Piling.js [7] allow the user to rearrange the multiples on-screen using the associated metadata. The user can then directly compare the small multiples to identify any trends

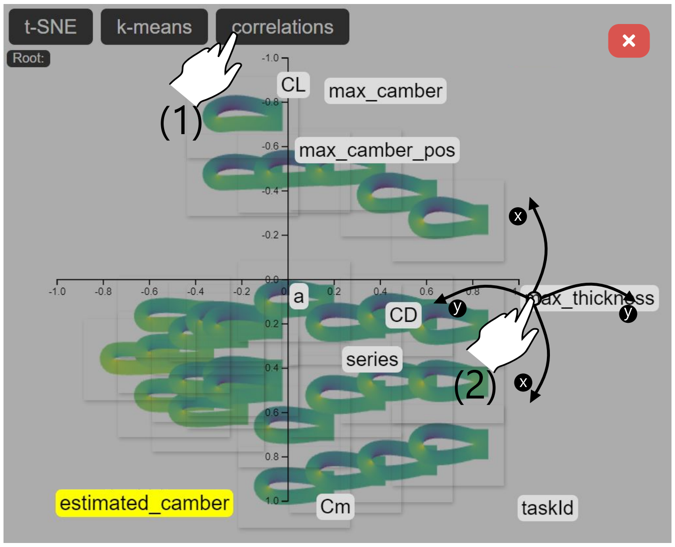

Fig. 5. The correlation menu showing the connection of spatial arrangement to metadata. Each metadata variable is visualised as a label. Correlations with user added tags are drawn in yellow. Note that categorical variables also have SRCC calculated. The top-left corner coordinates of the labels are the SRCC values between the metadata variable and the on-screen position of small multiples. By dragging the label perpendicularly to an axis, the user can arrange the small multiples according to the metadata variable along that axis.

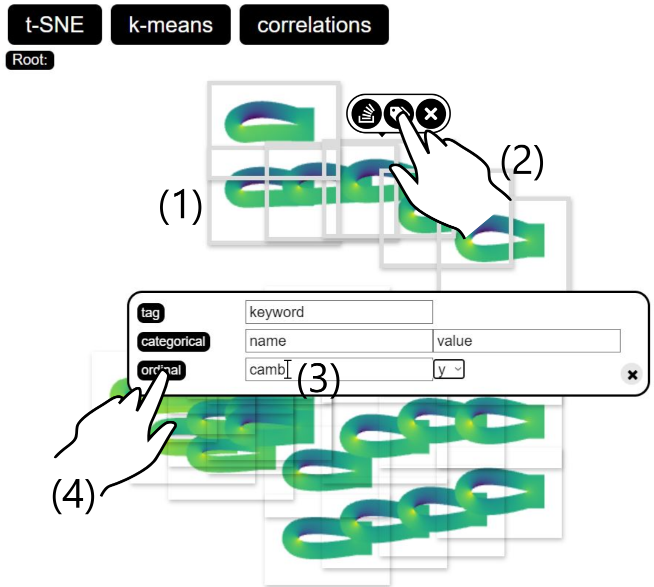

Fig. 6. An example showing the interactive tagging. (1) The lasso tool is used to select small multiples to tag, (2) tagging is selected, (3) the name of the tag is typed in, and the axis to take the values from selected, (4) the tag is added by pressing the ‘ordinal’ button.

associated with the metadata variable used for the arrangement. For datasets with hundreds of metadata variables, the process of sequentially arranging the small multiples by one metadata parameter, then the next, could be time-consuming.

The process can be accelerated by first identifying the changes that are visible in the small multiples, and then finding the associated causes and effect in the metadata (question mapping from the small multiples to the metadata). This approach relies on the analyst interacting with the data to encode their domain knowledge, such that the correlations in the metadata can be found. We suggest that the encoding is done via on-screen arrangement of the small multiples.

The discovery of metadata related to the on-screen arrangement pattern consists of two steps. The first step is to interactively arrange the small multiples on-screen into a meaningful pattern, and the second is finding variables that correlate well with the arrangement in the x and y coordinate directions. The first step allows domain knowledge to be used to interpret the small multiples data and encode the domain knowledge.

In the second step, shown in Fig. 3, the correlation between the on-screen arrangement and the metadata variables is calculated using the SRCC (Spearman’s Rank Correlation Coefficient), which assesses how well the relationship between two variables can be described using a monotonic function. The use of SRCC requires the user to arrange the small multiples such that the changes either decrease or increase monotonically in the vertical and horizontal directions. One input variable of the SRCC calculation is a metadata variable, and the other is the vertical or horizontal position of the small multiple on-screen. Therefore, the SRCC is calculated twice for every metadata variable. The SRCC can also be used to calculate correlations for categorical character label variables if the labels are first encoded with an integer value in order of appearance along the vertical and horizontal dimension on the screen. The results of the calculations are visualised in a correlation menu shown in Fig. 5. By clicking and dragging the metadata variable labels, as shown in Fig. 5 (2), the user can rearrange the small multiples using that metadata variable.

To calculate the SRCC between the on-screen arrangement and categorical metadata variables, the categorical variables must first be mapped to numbers. The midpoints of all small multiples corresponding to a particular label are calculated. The calculated midpoints are used to determine the order in which the labels broadly appear on-screen, and are therefore used to create the mapping. As the mapping may be different along the other axis, a new mapping is calculated and used when calculating the SRCC for that axis.

The vertical and horizontal positions used are distances in pixels from a reference point. This allows the user to also encode the degree of difference between individual small multiples into the on-screen arrangement by positioning small multiples the user judges to have smaller relative differences closer together, while positioning those with larger differences farther apart. The distances between individual small multiples can therefore help to better identify metadata variables of interest.

Using on-screen arrangement to encode the domain knowledge restricts the user to search for potential cause-and-effect relationships of a maximum of two features simultaneously (horizontally and vertically). Additional dimension encodings could be implemented, such as size and border color of the small multiples.

As well as finding potential cause-and-effect relationships between the metadata and individual group members, the spatial arrangement and correlation approach can also be used to study the connections between groups (piles) of images and the metadata. After the piles are created, the task for the analyst is to explain the cause of inherent differences between the piles. The approach discussed for analyzing individual small multiples can also be applied to piles. The analyst can spatially arrange the piles to form a meaningful trend or pattern, metadata variables correlated with the spatial arrangement are automatically detected, and the analyst can use those as clues to generate a domain-specific explanation for the observed pattern.

The cases corresponding to the small multiples can have several images associated with them, each showing a different aspect of the domain specific behavior. By creating piles based on one set of images, and then switching the view to another, the user can see how insights created when analyzing the first set correlate with the second. In this way, on-screen arrangement can be seen as a lens through which to look for cause-and-effect relationships between image-based data.

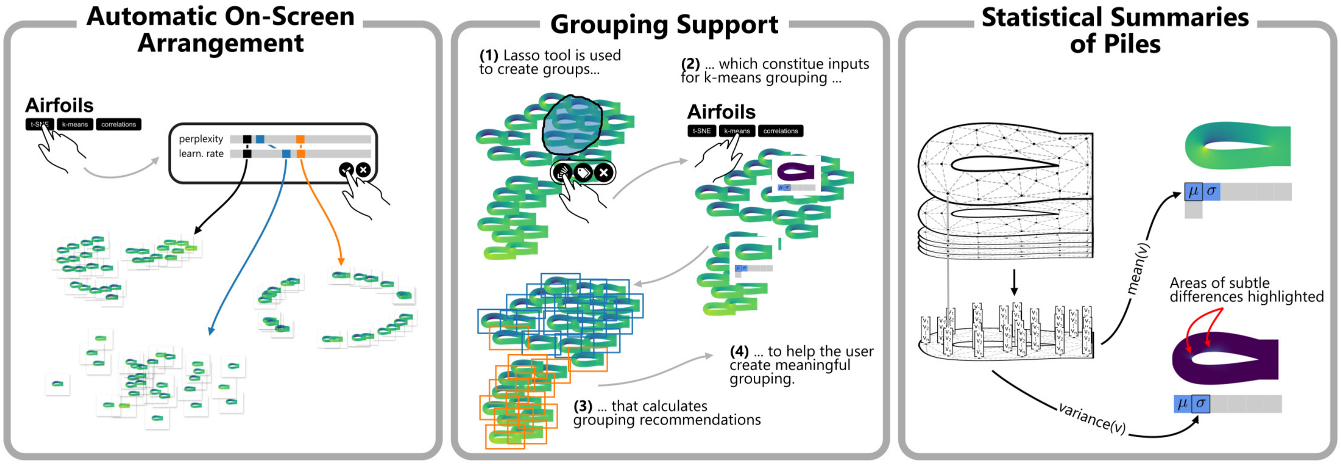
Automatic On-Screen Arrangement
Airfoils
t-SNE
k-means
correlations
perplexity
learn. rate
Grouping Support
(1) Lasso tool is used to create groups...
(2) ... which constitue inputs for k-means grouping ...
Airfoils
t-SNE
k-means
correlations
(4) ... to help the user create meaningful grouping.
(3) ... that calculates grouping recommendations
Statistical Summaries of Piles
mean(v)
μ σ
Areas of subtle differences highlighted
variance(v)
μ σ

Fig. 7. The t-SNE controls can be adjusted to find the most convenient grouping. The k-means interface allows quick grouping of all small multiples between the piles specified by the user. If the images have comparable underlying data the group members can be summarized using statistical covers.

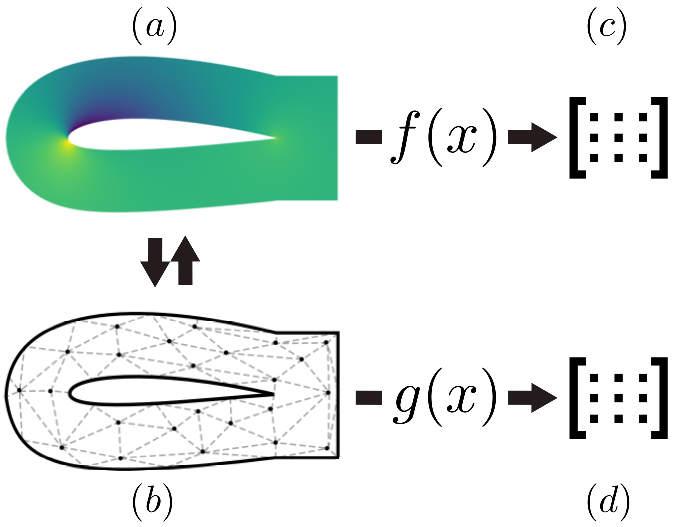

Fig. 8. The data used when comparing the images can be the image itself (a), the underlying data (b), where applicable, or an appropriate transformation of either (c),(d).

3.3 Interactive tagging

When the user creates meaningful piles of small multiples based on domain specific insights, they effectively temporarily encode their insights to the exploration session. However, the encoding is lost when the user chooses to disband the pile to try a different approach. To retain the insights for the long-term, they can be captured by adding the appropriate descriptive tags to the metadata, much like keywords are attached to scientific publications. These tags can later be used to search for cases, filter them, or aggregate several cases into a pile. The piles are the ideal basis to use for the tagging. Examples of tagging are presented conceptually in Fig. 3, and a screenshot of a demo is shown in Fig. 6.

Two distinct types of tags can be added to the metadata: loose and name-value pair tags. Loose tags allow the user to capture specific insights about a particular group of small multiples, such as identifying the animals shown in the image, a particular flow structure, or crystalline structure. Name-value pair tags allow the comparison of all small multiples that have the appropriate name-value tag: the name provides a dimension along which the multiples can be compared; the 'value' allows direct comparison of individual multiples. Furthermore, the name-value tags can be used for on-screen arrangement of small multiples, finding correlations with other metadata variables, and to study interactions between features the user identifies during exploration.

After interpreting a series of small multiples, the user may want to capture the progression of a specific feature and use it for further analysis. On-screen spatial arrangement allows ordinal name-value tags to be captured. The user can arrange the small multiples in a meaningful order, and the non-dimensional on-screen position of the image can be captured as the value. The approach is flexible and can be applied to images with any content.

4 Acceleration Methods to Support the Framework

We implement a range of methods to accelerate the cause-and-effect discovery framework outlined above. The methods are illustrated in Fig. 7 and are divided into three categories: first, creation of meaningful piles by arranging similar small multiples together on-screen; second, automatic grouping of small multiples among established piles; third summarizing the behavior of the piles.

Before discussing these support methods, we note that each of them relies on the availability of data allowing individual dataset members to be compared to one another. This data may be available as metadata or may have to be derived from the small multiples themselves, Fig. 8. Rendered images can be converted into underlying data by sampling the image at several locations; for example, grayscale images can be treated as a set of 2D arrays whose values can be compared element-wise. Alternatively, images can undergo further processing and the derived metrics used for comparison; for example the activation map from a layer of a CNN could be used. Whether the comparative data is obtained from metadata, or from the small multiples, the following methods are applied in the same way.

We now discuss each of three categories of support method in turn.

4.1 Grouping support

Piling.js [7] gives an overview of interactive techniques to form piles of small multiples. The two main tasks of pile forming are: selection of the appropriate number of piles; and, allocation of the small multiples to their corresponding group. Piling.js allows the user to form piles manually (by dragging and dropping small multiples, or by selecting several multiples using a lasso tool) and automatically (using a grid applied to the viewport such that all small multiples that fall into the same grid cell form a pile). The manual approach relies on the user to distribute the small multiples, which can be time intensive if the images

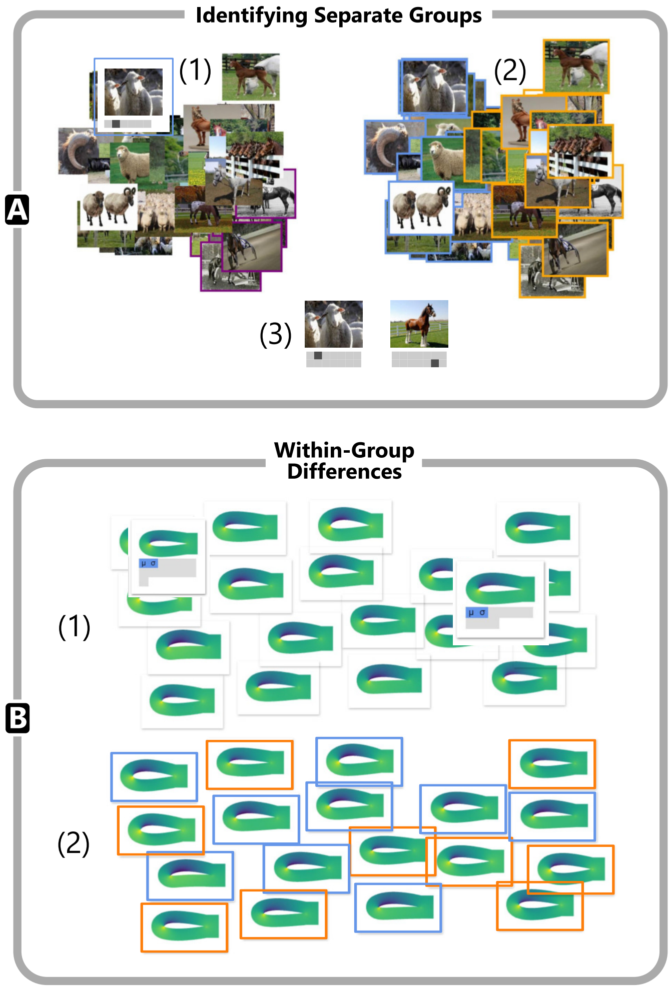

Fig. 9. A-(1) The user is grouping some horses (purple border), after having created a pile of sheep. The piles form a basis for k-means clustering A-(2) with the clustering results encoded through border color. A-(3) The clustering results can be used to arrive at complete sheep and horses piles. B-(1) The user creates piles representing the extremes of observed behavior. B-(2) The k-means segments the small multiples and supports the user in roughly arranging them based on the observed behavior.

are inconveniently arranged on-screen. The benefit of manual piling is that it allows the user to aggregate small multiples based on *their assessment* of the represented data, as opposed to using the associated metadata or the on-screen position. The automatic approach relies on the user to implicitly select the number of piles by selecting dimensions of the grid that will be used to aggregate the multiples. This approach creates piles quickly, however the piles obtained may not be appropriate for further knowledge extraction and cause-effect discovery.

We propose a third, augmented, approach to pile creation, which combines the advantages of both the manual and automatic approaches. The advantage of manually specifying piles is that the analyst uses their domain knowledge to form a meaningful pile. In doing so, the user specifies the number of clusters (number of piles on-screen), and captures some insight, as the members form a meaningful group that is in some way different to the rest of the multiples. A clustering algorithm, that works on the small multiple data, can then complete the assignment of all multiples into the established piles.

The augmented approach can be used to identify separate groups, or to segment a single group. Fig. 9 A-(1) shows the user creating initial piles, that are used to create the assignment for all the small multiples shown in Fig. 9 A-(2). The user can create initial piles based on fundamental differences between the small multiples and use the results to complete the piling. Alternatively, augmented grouping can be used to segment members of a particular group. In that case the user can identify the extreme members by creating piles based on them, Fig. 9 B-(1), and use the clustering results, encoded by the border color, Fig. 9 B-(2), to rearrange the small multiples.

4.2 Automatic on-screen arrangement

On-screen arrangements can be based on the metadata variables, the images themselves, or image encodings (using CNNs, for example). Each of these is likely to have a much higher dimensionality than the two-dimensional, on-screen coordinate system. Dimensionality reduction techniques such as t-SNE [11] and UMAP [8] have been introduced to address the challenge of embedding high-dimensional data into a 2-D visualization space and these have been applied to the spatial arrangement of small multiples [6, 7].

The most general approach to use dimensionality reduction for on-screen arrangement is to add embeddings to the metadata, and then use the metadata to position the small multiples. This allows the arrangement to be pre-calculated by any dimensionality reduction technique available to the user. This is the approach taken in an example on the PILING.JS website, where the UMAP arrangement is accomplished using pre-computed UMAP embeddings.

As an alternative to the pre-computed strategy, we integrate the dimensionality reduction process into the application itself to allow interactive control of the low dimensional embedding. A benefit of this approach is that the user may manually create a pile, for example, and then request t-SNE to provide a two-dimensional embedding of the remaining small multiples (i.e. not the full dataset). In addition, the dimensionality reduction technique may have parameters that control the resultant segmentation and clustering. For t-SNE, the three settings are learning rate, perplexity, and the dimensionality of the resulting embedding. In an interactive approach, perplexity and learning rate can be adjusted and the t-SNE re-run, as shown in Fig. 7. Both the recalculation of t-SNE after the manual creation of piles, and after the changing of t-SNE parameters, require the user's involvement to determine if the spatial arrangements are meaningful; we view an interactive approach as key to maintaining the user's effective engagement with the process.

4.3 Statistical summaries of piles

When analyzing a pile of small multiples, the two main tasks are: establishing an image that is representative of the members of the pile; finding the variation in the members of the same pile and even determining which areas of the images contain the greatest differences between the pile members (this is particularly helpful when these differences are small, or located on part of the image only, as is often the case in science and engineering visualization).

HIPILER [6] provides an 'average' and 'variance' cover for matrix visualizations. These covers are created by calculating the mean and variance for each matrix cell using the data from all pile members. This approach can be extended to greyscale images by interpreting the pixel values as the matrix values. When interpreting images as matrices, it is assumed that individual pixels are directly comparable.

We extend the matrix visualization approach to the general case of data stored in two-dimensional structured arrays. Engineering simulation datasets often contain scalar field data stored on a structured (i, j) grid; in this case, although the size of the grid is the same for each member of the dataset, the coordinates of each point (x_{ij}, y_{ij}) and its scalar value v_{ij} are different in each case. A small multiple that is representative of the pile can be generated by obtaining the mean coordinate and value, at each i, j pair, from the set of data associated with the pile. The calculation is illustrated in Fig. 7. In the same way, taking the variance of the value each point, and using the mean coordinates, allows the user to discover which areas of the image are associated with the greatest differences between the pile members. This approach can also be applied to sets of structured grid data where the size of the arrays is not the same for each member of the dataset, or

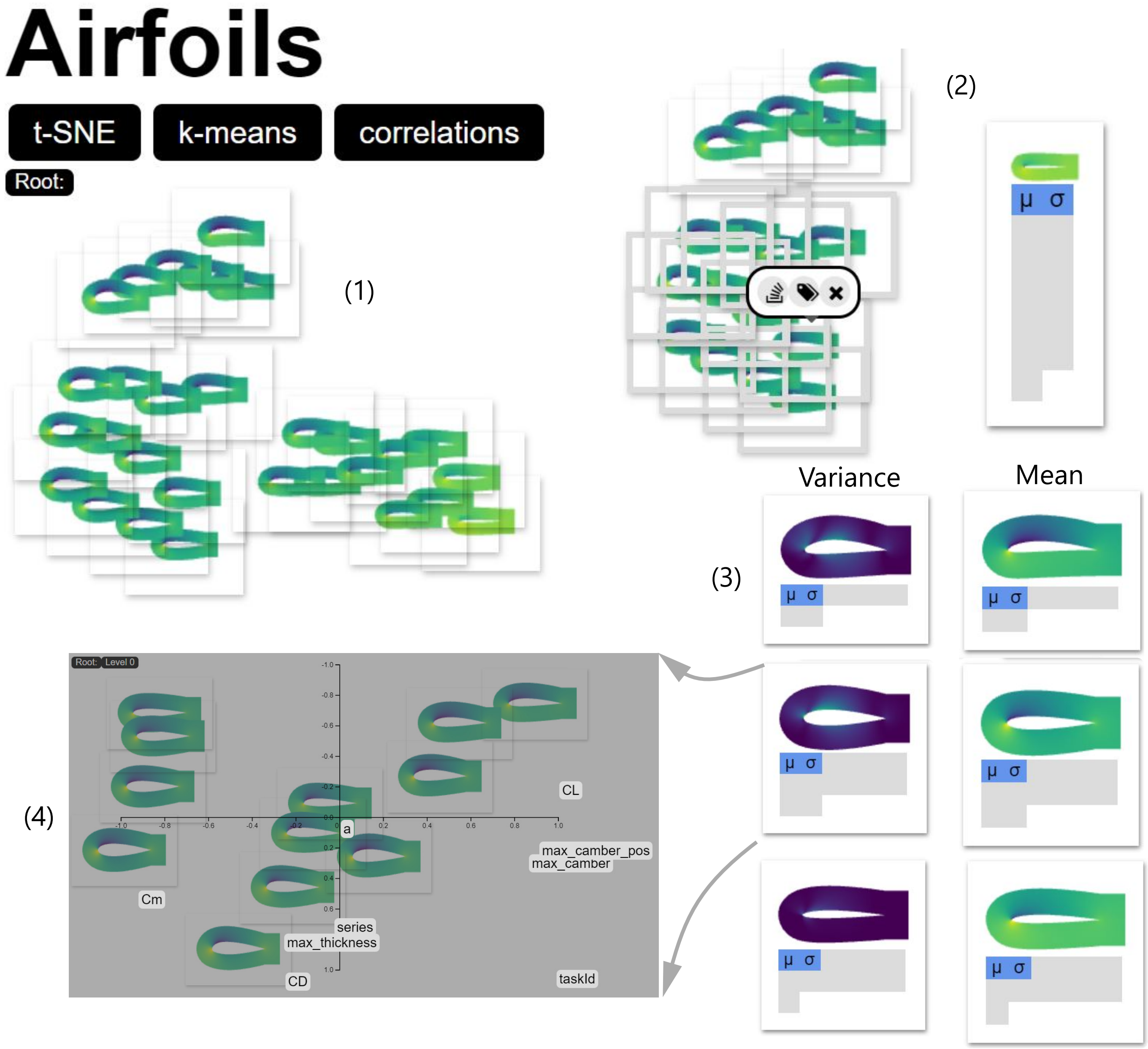

Fig. 10. (1) The initial arrangement of small multiples using t-SNE suggests three clusters. (2) Piles are created based on the observed behavior. (3) The statistical covers show that individual groups contain airfoils of different general shapes. The variance cover shows that most differences appear on the airfoil top side. (4) Arranging the multiples within a group allows the user to see how the images change with respect to a particular metadata parameter or vice versa.

if unstructured (triangulated) data is used, provided the data can be sampled, in a meaningful full way, on to a structured grid of fixed size.

In situations where a comparable structured array for each small multiple is not available (a dataset of photographs, for example) a ‘representative pile member’ can be obtained, instead of representing the average of the members of the pile. This can be achieved using the multi-dimensional centroid of the group members’ metadata, and then defining the member closest to the centroid as the most representative.

5 Demonstration case study

As an illustrative case study, we now show how the process and tools described enable the discovery of cause-and-effect relationships in an example from engineering design. A dataset of simulated flow around different airfoil designs was created by evaluating 29 airfoils (NACA 4 and 5-series airfoils [1]) using the XFOIL flow simulation code [4]. A walk-through analysis example is shown in Fig. 10, and a working demo is available at: `https://aljazkotnik.github.io/harnessingspace/`.

The result of the evaluation of a single airfoil are 2D scalar fields consisting of arrays of size 181 × 26 on which the flow variables, such as horizontal and vertical flow velocity components, and pressure, are stored. For our illustration, the image data, and input-output metadata are:

image:	airfoil shape and pressure scalar field
input metadata:	airfoil max thickness
	airfoil max non-symmetry
	angle of oncoming air to airfoil
	speed of oncoming air
output metadata:	lift coefficient
	drag coefficient
	moment coefficient

The input metadata describes both the shape of the airfoil and the conditions of the oncoming air flow (left-to-right in the images). The output metadata provides a high-level description of the airfoil performance.

The scalar fields are rendered using WebGL, and the images allow the engineer to understand the mechanisms that underpin the airfoil performance. The t-SNE arrangement of small multiples is based on the underlying 2D array of scalar field data itself, and, by changing the t-SNE parameters, the user can try to find a better arrangement. The arrangement in Fig. 10(1) shows 3 major clusters, which are piled in Fig. 10(2).

The mean covers in Fig. 10(3) show that the bottom group contains thin airfoils, the top group members have flat lower airfoil surfaces, and the middle ones have blunt leading edges. The variance cover shows areas of high differences between the group members in yellow. In this case most of the differences appear on the top airfoil surface for the top and middle groups. This location on the variance covers corresponds to the blue (low pressure) areas on the mean covers, indicating that most of the differences are due to the differences of the blue (low pressure) areas on the top surfaces of the airfoils.

Entering the group and arranging the multiples by the lift coefficient, CL, in the horizontal direction and the drag coefficient, CD, in the vertical direction allows the identification of other correlated variables, as well as the changes in the small multiples, Fig. 10(4). Maximum thickness appears to have an impact on drag, but not on lift. Camber correlates with lift, but not with drag. The flows in the top left corner have relatively large blue areas that represent low air pressure on the top side of the airfoils. Once the user is happy with the analysis of this pile they can exit it and focus on other piles.

In this example, the dataset is comprised of the results from 29 simulations. With current compute capability, datasets of hundreds of designs are routinely created. The analysis process described – interactive arrangement and grouping of small multiples, correlation with input-output metadata, and knowledge capture – injects valuable acceleration into the discovery of key cause-and-effect relationships.

6 Conclusions

We have demonstrated a ‘divide and explore’ approach for the discovery of cause-and-effect relationships in datasets containing images and associated input-output metadata. The approach is based on the interactive spatial arrangement of small multiples, and is comprised of three parts:

1. **Subset grouping.** The interactive piling approach presented by Lekschas *et al.* [7] was used as the inspiration for the user interface. Piles are the natural way to create taxonomies of the data. Furthermore, they provide a host for statistical summaries of the pile members, whether that information is encoded in the pile properties, or in a statistical cover image. Where possible, we extend the variance cover concept to the underlying data from which the images are rendered.

2. **Connecting spatial arrangement to input-output metadata.** The spatial arrangement of individual small multiples, or of subset groups, is a powerful technique for capturing the user’s domain knowledge in the analysis of the dataset. We provide statistical tools to guide the identification of cause-and-effect relationships by correlating the two-dimensional spatial location with the input-output metadata.

3. **Knowledge capture.** On-the-fly tagging is implemented to allow the user to capture insights as they occur. Adding keywords to the data allows the user to save the taxonomy of groups. Name-value pair tags allow the user to compare the images through custom-created values. The capture of ordinal name-value tags is found to be a useful mechanism to encode quantitative information for images whose underlying data is unstructured (triangulated).

To accelerate the process, interactive interfaces have been added for the arrangement of the small multiples using t-SNE, and for grouping using k-means. For t-SNE, one-dimensional parameters can be specified using sliders, and for k-means the piles are used to simultaneously specify the number of clusters to find, as well as the starting points.

In both the three-step process, and the tools used to accelerate it, we emphasise the importance of interactivity as a route to fully utilizing the user’s domain knowledge.

The approach has been described using example datasets from image processing, material science and aerospace engineering, but has broad utility to other datasets comprised of images and input-output metadata.

7 Future work

The extension of the present work to time-dependent data would enable additional classes of measurement and simulation data to be analysed. This can be achieved either through static representations (directly via a time series, or via Fourier transforms, wavelets or modal decomposition) or by using a dynamic video small multiples arranged by appropriate stationary metrics (e.g. time-averaged values). Dynamic small multiples, whose representation responds to a time slider for example, would open up many new opportunities for the domain practitioner to analyse datasets of multiple time-varying cases.

Acknowledgments

The authors wish to thank Dr Marcus Meyer for his support and advice during this work.

References

[1] I. H. Abbott, A. E. Von Doenhoff, and L. S. Stivers Jr. Summary of airfoil data. Technical report, 1945.

[2] A. Corrado. Animals-10. `https://www.kaggle.com/alessiocorrado99/animals10`, 2019.

[3] B. L. DeCost, M. D. Hecht, T. Francis, B. A. Webler, Y. N. Picard, and E. A. Holm. Uhcsdb: ultrahigh carbon steel micrograph database. *Integrating Materials and Manufacturing Innovation*, 6(2):197–205, 2017. doi: 10.1007/s40192-017-0097-0

[4] M. Drela. Xfoil: An analysis and design system for low reynolds number airfoils. In *Low Reynolds number aerodynamics*, pp. 1–12. Springer, 1989.

[5] V. Estivill-Castro. Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, 4(1):65–75, {jun} 2002. doi: 10.1145/568574.568575

[6] F. Lekschas, B. Bach, P. Kerpedjiev, N. Gehlenborg, and H. Pfister. Hipiler: Visual exploration of large genome interaction matrices with interactive small multiples. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):522–531, 2018. doi: 10.1109/TVCG.2017.2745978

[7] F. Lekschas, X. Zhou, W. Chen, N. Gehlenborg, B. Bach, and H. Pfister. A generic framework and library for exploration of small multiples through interactive piling. *IEEE Transactions on Visualization and Computer Graphics*, 10 2020. doi: 10.1109/TVCG.2020.3028948

[8] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018.

[9] G. Pullan, T. Chuan, D. Wong, and F. Jasik. Enhancing web-based cfd post-processing using machine learning and augmented reality. In *AIAA Scitech 2019 Forum*, p. 2223. AIAA, 2019. doi: 10.2514/6.2019-2223

[10] E. R. Tufte, N. H. Goeler, and R. Benson. *Envisioning information*, vol. 2. Graphics press Cheshire, CT, 1990.

[11] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86), {nov} 2008.