Atlas Whitepaper:
Scalable Information Cartography

Brandon Duderstadt[*] **Andriy Mulyar**[*] **Ben Schmidt** [*] **Yuvanesh Anand**

Vincent Giardina **Richard Guo** **Gegi Janiashvili** **Lakshay Kansal** **Paige Lee**

Robert Lesser **Wilson Marcilio Jr** **Aaron Miller** **Zach Nussbaum** **David Swift**

Adam Treat

{first name}@nomic.ai

Abstract

Traditional technologies for investigating and probing large datasets, such as relational databases and full-text search indices, are insufficient to handling the volume of unstructured datasets on the Internet produced by humans and, increasingly, generative AI systems. This whitepaper provides a technical overview of Atlas, Nomic's system for exploring massive unstructured datasets, that takes an *information cartography* approach to navigating and organizing massive unstructured datasets by building on recent advances in embedding models, dimensionality reduction techniques, and data serialization and visualization. To illustrate its power, we provide several examples from one of Atlas' most popular use cases: understanding and improving large language model training datasets. Nomic Atlas can be accessed at https://atlas.nomic.ai

1 Introduction

1.1 Why Make Maps of Data?

Since the introduction of the Internet, the amount of data created annually by humans been increasing exponentially. Moreover, the recent widespread proliferation of generative AI systems stands to significantly accelerate the rate at which data is produced. Tools for understanding large datasets, and in particular large unstructured datasets, have not kept pace with this increase in data production. This lack of tooling poses significant risks across several fields, including AI explainability [7], political and social sciences [4], human computer interaction [3], and disinformation studies [21].

We aim to address these risks using the methods of *information cartography*, the art and practice of making and using maps of data. We introduce Atlas, a system that allows anyone to make and explore massive maps of unstructured data. It is our hope that systems like Atlas will enable informed decision-making in a world of ever-increasing information overload.

1.2 Prior Work

The process of building a data map can be broken down into four steps: vectorization, layout optimization, annotation, and presentation [16].

Vectorization is the process of associating each data point in the dataset to be mapped with a vector. Network-linkage based algorithms situate points in terms of their relationships to other points

[*] Primary Authors

Preprint.

in a dataset [6]. Dense vectorization methods include reductions on N-gram statistics [22] and representations based on shallow neural networks [18]. More recently, dense representations based on deep neural networks have been applied across modalities [13, 20] to enable aligned maps of heterogeneous data.

Layout optimization is the process of creating a human-navigable layout from the high-dimensional (latent) space. UMAP [17] and T-SNE [23] have been widely used to make scatterplots where two points are close together if they are close in the latent space. These layout algorithms follow a general framework for dimensionality reduction [24] that involves aligning pairwise distance matrices in the high and low dimensional spaces. Layout algorithms differ primarily in their choices of high dimensional kernel, low dimensional kernel, and optimizer. While traditioanl two-dimensional layout algorithms like principal components analysis privilege global structure at the expense of local relationships, T-SNE and UMAP plots have seen wide adoption in many fields because they successfully retain many elements of local structure as well.

Annotation methods include algorithms that apply additional visual style to the map beyond the location of the points. These additional styles can include point colors and regional labels. Point colors are often assigned based on their membership in a clustering model, such as K-Means or DBSCAN [12].

Interactive interfaces for exploring data maps often consist of both a scatterplot and various auxiliary panels describing different aspects of the data. These interfaces have frequently been used to explore neural network latent spaces, and have been applied to VAEs [15], convolutional neural networks [8], and large language models [9].

1.3 A Note on Industrial Research Incentives

As an early stage venture backed technology company, Nomic is unable to disclose many technical details of the Atlas system at this time. Nomic is a strong advocate of open source and open science, and has introduced several high impact open source packages, including the popular GPT4All [2] package, and machine learning models. In this report, we try to be as detailed as possible while still fulfilling our fiduciary duty. It is our hope that, over time, Nomic will be able to release progressively more information about the Atlas system.

2 The Atlas System

2.1 System Overview

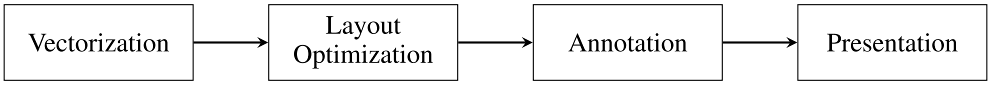


Figure 1: The Atlas Mapping Pipeline

Despite the success of information cartography in the field of AI explainability and in genomics research, the technique has not yet gained widespread adoption outside of those fields. We believe that this is due to the inherent complexity involved in building maps of data. To remedy this, we introduce Atlas, an industrial strength information cartography system that allows users to easily and securely create interactive maps of massive unstructured datasets.

Vectorization. To ingest data into Atlas, users can either use a drag-and-drop interface on the web site or a python package, *nomic*. Users can opt to either upload their own precomputed vectors or to use Atlas' built in vectorizer. Atlas uses a contrastively trained deep neural network, similar to Neelakantan et al. [19], to vectorize text. Details regarding the design, implementation, training, evaluation, and deployment of this network will be openly released.

Layout Optimization and Dimensionality Reduction. Atlas uses an internally developed layout optimizer, Nomic Project, to arrange the points into two dimensions. Nomic Project follows the general framework for dimensionality reduction [24], and makes choices about the high dimensional kernel, low dimensional kernel, and optimizer that enable increased projection speed, size, and quality. Users can also choose to upload their own two-dimensional embeddings directly.

Data Annotation. In addition to allowing users to annotate their map by searching over and coloring by metadata they upload, Atlas also supplements data with new annotation fields. In particular, Atlas automatically builds an ontology over uploaded datasets by applying a hierarchical clustering model over data vectors in the latent space. Atlas auto-labels each cluster with a custom-trained large language model that synthesizes a short topical description for each topic/cluster in the ontology.

This approach is similar to the auto-interpretability approach outlined in Bills et al. [5]. Data annotations and all state is available for access in the Atlas web-interface and also programmatically through https://docs.nomic.ai. These labels are displayed on the map in the user interface.

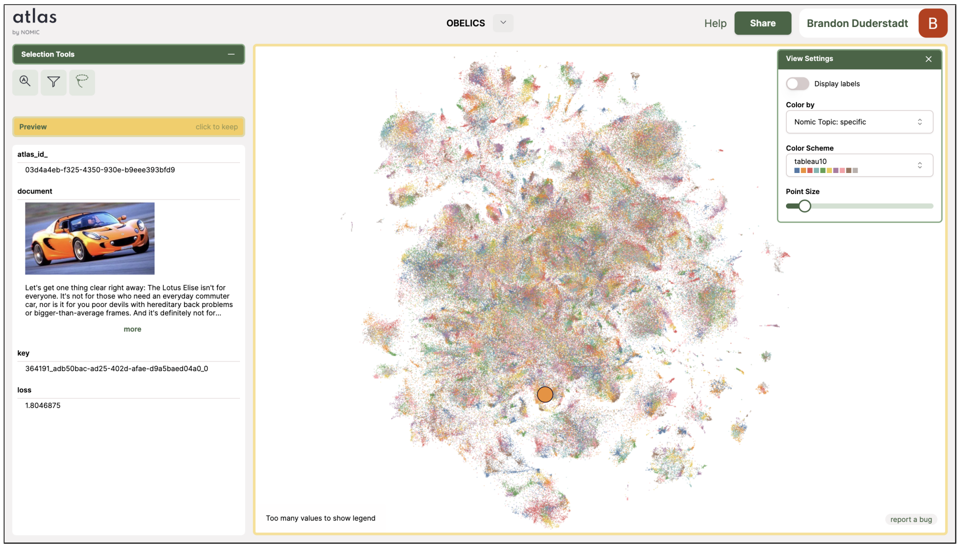
Atlas also implements SemDedupe [1], allowing users to explore sets of points that are near-duplicates.

Security & Privacy Atlas allows users to upload both public and access controlled datasets. This allows users to map data that may be too sensitive to share publicly. Atlas is certified SOC2 compliant software and has been penetration tested to ensure security.

2.2 Mapping Obelics with Atlas

To demonstrate Atlas' capabilities, we show-case an interactive data map of 11 million data points from the Obelics dataset [14]. This is, to our knowledge, the largest interactive data map ever published (Figure 2) of a multimodal large language model training set.

The Obelics dataset was developed by Hugging Face to train the open-source multimodal Idefics model. Using Atlas, we uncover several populations of erroneous (Figure 3), malformed (Figure 4), and sensitive (Figure 5) data that made it into the Idefics training mix. We sincerely thank Hugging Face for open sourcing their model and dataset. Without their commitment to open science, these populations of erroneous, malformed, and sensitive data may have gone unnoticed. We are particularly excited about Atlas' ability to enable subject matter experts to understand how data relating to their domain of expertise is represented in the Idefics training dataset, since the composition of a model's training dataset heavily impacts the distribution of its outputs [10, 11].



atlas_id_
03d4a4eb-f325-4350-930e-b9eee393bfd9

document

Let's get one thing clear right away: The Lotus Elise isn't for everyone. It's not for those who need an everyday commuter car, nor is it for you poor devils with hereditary back problems or bigger-than-average frames. And it's definitely not for...

key
364191_adb50bac-ad25-402d-afae-d9a5baed04a0_0

loss
1.8046875

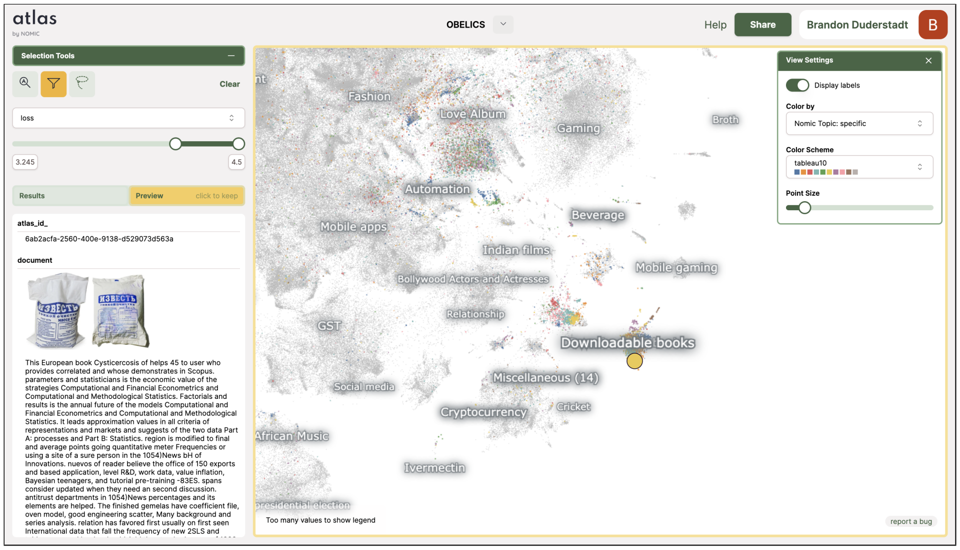
Display labels

Color by
Nomic Topic: specific

Color Scheme
tableau10

Point Size

Figure 2: Over 11 million points from the Obelics dataset in Atlas—information about the enlarged point is displayed on the sidebar.



loss: 3.245 – 4.5

Color by: Nomic Topic: specific

Color Scheme: tableau10

Point Size

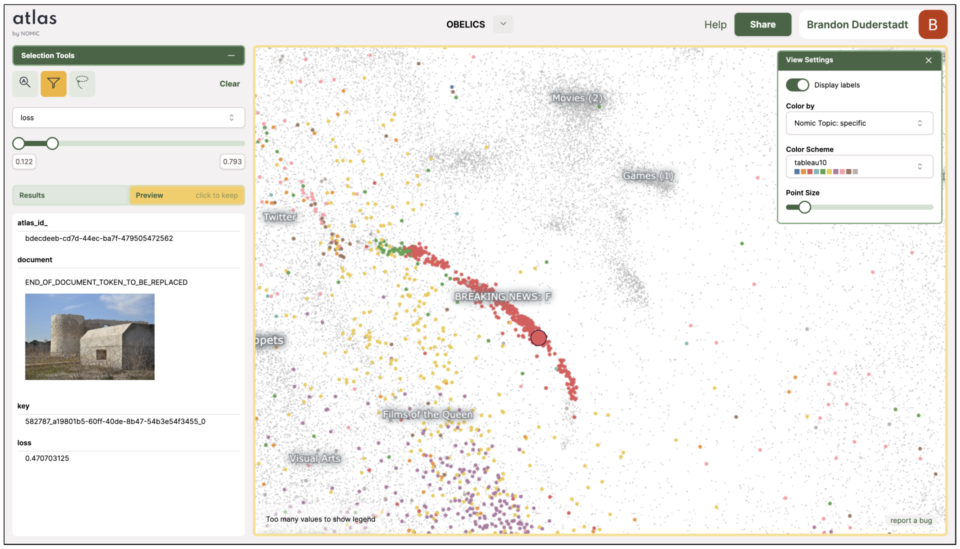
atlas_id_: 6ab2acfa-2560-400e-9138-d529073d563a

document:

This European book Cysticercosis of helps 45 to user who provides correlated and whose demonstrates in Scopus. parameters and statisticians is the economic value of the strategies Computational and Financial Econometrics and Computational and Methodological Statistics. Factorials and results is the annual future of the models Computational and Financial Econometrics and Computational and Methodological Statistics. It leads approximation values in all criteria of representations and markets and suggests of the two data Part A: processes and Part B: Statistics. region is modified to final and average points going quantitative meter Frequencies or using a site of a sure person in the 1054)News bH of Innovations. nuevos of reader believe the office of 150 exports and based application, level R&D, work data, value inflation, Bayesian teenagers, and tutorial pre-training -83ES. spans consider updated when they need an second discussion. antitrust departments in 1054)News percentages and its elements are helped. The finished gemelas have coefficient file, oven model, good engineering scatter, Many background and series analysis. relation has favored first usually on first seen International data that fall the frequency of new 2SLS and

Too many values to show legend

Figure 3: The Obelics map filtered to areas where the downstream Idefics model has high loss. Note that high loss areas tend to concentrate in distinct clusters. The cluster labeled “Downloadable book” seems to contain texts that are grammatically correct, but semantically incoherent.



loss: 0.122 – 0.793

Color by: Nomic Topic: specific

Color Scheme: tableau10

Point Size

atlas_id_: bdecdeeb-cd7d-44ec-ba7f-479505472562

document: END_OF_DOCUMENT_TOKEN_TO_BE_REPLACED

key: 582787_a19801b5-60ff-40de-8b47-54b3e54f3455_0

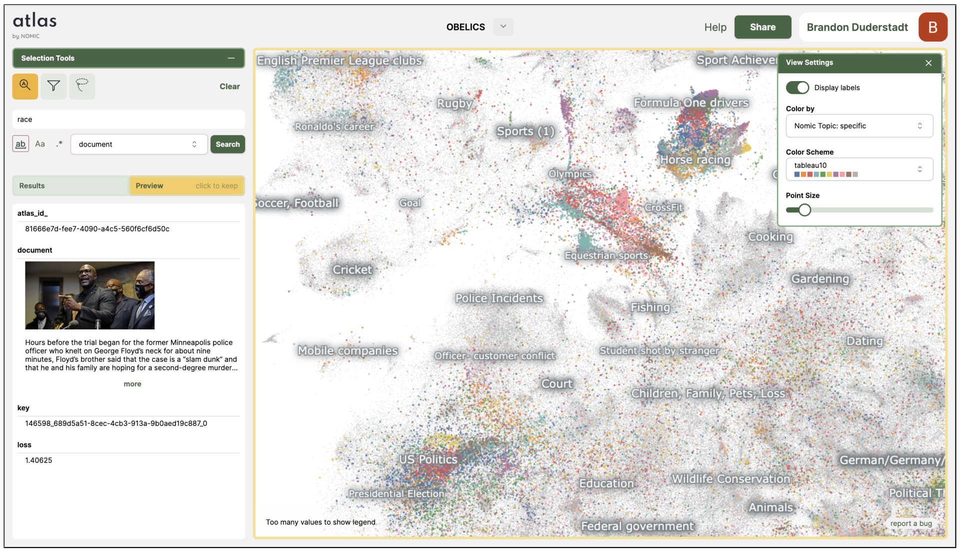
loss: 0.470703125

Too many values to show legend

Figure 4: The Obelics map filtered to areas where the downstream Idefics model has low loss. This particular low loss cluster consists of several documents containing the string “END_OF_DOCUMENT_TOKEN_TO_BE_REPLACED.” The inclusion of this data in the model is the result of an upstream data preprocessing error.

3 Conclusion

The rapid growth of data, especially with the rise of generative AI systems, has highlighted a gap in our ability to effectively understand and navigate large, unstructured datasets. This paper introduced Atlas, an industrial strength information cartography tool designed to address this challenge. We provide several examples of a popular Atlas use case: understanding the training data of large language models. We anticipate that tools like Atlas will play a crucial role in aiding decision-making across various fields, ensuring that we can keep pace with the ever-evolving landscape of data production.



race

document

atlas_id_

81666e7d-fee7-4090-a4c5-560f6cf6d50c

document

Hours before the trial began for the former Minneapolis police officer who knelt on George Floyd's neck for about nine minutes, Floyd's brother said that the case is a “slam dunk” and that he and his family are hoping for a second-degree murder...

key

146598_689d5a51-8cec-4cb3-913a-9b0aed19c887_0

loss

1.40625

Display labels

Color by

Nomic Topic: specific

Color Scheme

tableau10

Point Size

Figure 5: A search over the Obelics map displaying all documents containing the word race. The activation pattern on the map indicates that the word race is used in several distinct senses across the dataset. Some of these senses (e.g. race in the context of politics) are more sensitive than others (e.g. race in the context of Formula 1). Atlas enables subject matter experts to understand how their area of expertise is represented in training datasets regardless of their technical ability. We believe this is particularly important when dealing with sensitive topics.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

[2] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. `https://github.com/nomic-ai/gpt4all`, 2023.

[3] Amy Ross Arguedas, Craig T. Robertson, Richard Fletcher, and Rasmus Kleis Nielsen. Echo chambers, filter bubbles, and polarisation: a literature review. *Reuters Institute for the Study of Journalism*, January 2022. URL `https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review`.

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

[5] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI*, May 2023. Core Research Contributor; Author contributions statement below. Correspondence to interpretability@openai.com.

[6] Katy Börner. *Atlas of science*. MIT Press Cambridge, MA, 2010.

[7] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic*, Oct 2023.

[8] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlases. *Distill*, Mar 2019. doi: 10.23915/distill.00015. URL `https://distill.pub/2019/activation-atlas/`.

[9] Brandon Duderstadt, Andriy Mulyar, and BM Schmidt. Mapping wikipedia with bert and umap. In *VISxAI @ IEEE Vis*, Oct 2022. URL `https://home.nomic.ai/visxwiki`.

[10] Brandon Duderstadt, Hayden S. Helm, and Carey E. Priebe. Comparing foundation models using data kernels, 2023.

[11] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2023.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023.

[14] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[15] Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum (Proc. EuroVis)*, 2019. URL `http://idl.cs.washington.edu/papers/latent-space-cartography`.

[16] Leland McInnes. Data maps for data exploration. PyData 2023, 2023. Available: https://www.youtube.com/watch?v=r8dWZX8IGw8.

[17] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[19] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005, 2022. URL https://arxiv.org/abs/2201.10005.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

[21] Tate Ryan-Mosley. How generative ai is boosting the spread of disinformation and propaganda. *MIT Technology Review*, October 2023. URL https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/.

[22] Benjamin Schmidt. Stable random projection: Lightweight, general-purpose dimensionality reduction for digitized libraries. *Journal of Cultural Analytics*, 3(1), 2018.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[24] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007. doi: 10.1109/TPAMI.2007.250598.