Probing Heterogeneous Pretraining Datasets with Small Curated Datasets

Gregory Yauney [1 2] **Emily Reif** [3] **David Mimno** [1]

Abstract

Language models rely on increasingly large web-scraped datasets for pretraining. The size of these datasets prevents manual curation, and existing automated quality filters are heuristic and limited. Characterizing these datasets is an open problem. We present preliminary work on documenting and visualizing pretraining datasets by mapping their similarity to downstream benchmark datasets, which are often hand-curated and more focused in style and content. We show this method finely characterizes popular pretraining datasets, supplementing existing characterizations that can be used for quality filtering.

1. Introduction

Large language models have achieved success by pretraining on web-scale datasets. Larger pretraining datasets empirically improve language modeling performance (Kaplan et al., 2020), with recent models like PALM being trained on 780 billion tokens (Chowdhery et al., 2022). However, extremely large datasets like the Pile and C4 have proven difficult to characterize beyond website source and word frequencies (Gao et al., 2020; Dodge et al., 2021; Schaul et al., 2023). How can we characterize web-scraped datasets, especially with respect to their downstream uses?

Pretraining datasets are too large for manual curation, so dataset constructors curate based on "quality" proxies such as domain (Nguyen et al., 2022; Xie et al., 2023), heuristics (Raffel et al., 2020), or classifiers (Chowdhery et al., 2022; Du et al., 2022). A common approach narrowly defines high-quality documents as those that a classifier predicts are more similar to books and Wikipedia (Chowdhery et al., 2022), though this has limitations (Longpre et al., 2023). Are there more data-centric ways to evaluate the quality of pretraining data?

[1]Cornell University [2]Work done as a student researcher at Google Research. [3]Google Research. Correspondence to: Gregory Yauney <gyauney@cs.cornell.edu>.

Data-Centric Machine Learning Research Workshop at ICML, Honolulu, Hawaii, USA. 2023.

We present preliminary work on using small datasets to describe large pretraining datasets. While pretraining datasets are largely undocumented, downstream benchmark datasets are often meticulously constructed and validated (Paullada et al., 2020). We contribute a clustering-based "dataset probing" method to find the documents in a pretraining dataset that are similar to known small dataset "probes." We find different probe datasets cover different clusters of popular language model pretraining datasets. Origin-based domains are often used as proxies for pretraining document content (Nguyen et al., 2022), and our method provides a supplemental data-centric characterization that does not simply reproduce domain source. Our method is a step towards more multipolar notions of quality, and we hope it can provide more granularity for heterogeneity-based pretraining dataset curation.

Related work. Data documentation is becoming more prevalent (Bender & Friedman, 2018; Gebru et al., 2021) but is hard to perform at scale (Bandy & Vincent, 2021; Dodge et al., 2021). Unsupervised methods like topic modeling (Blei, 2012; Li & McCallum, 2006), or embedding the dataset (McInnes et al., 2018; Van der Maaten & Hinton, 2008) and visualizing it (Smilkov et al., 2016; Wexler et al., 2019; Bolukbasi et al., 2021) are all used to characterize large datasets. These can uncover patterns in the corpus but do not leverage better-understood datasets. Visualization is compatible with existing dataset documentation frameworks (Crisan et al., 2022; Pushkarna et al., 2022). Pillutla et al. (2021) and Assogba et al. (2023) both compare clustered pairs of datasets, but for the purpose of evaluating generated data rather than describing a large heterogeneous dataset.

2. Dataset probing

Given a pretraining dataset and a set of small probe datasets, our goal is to find which pretraining documents are similar to probe documents. We represent documents using Sentence-T5 (Ni et al., 2022; Wolf et al., 2020) embeddings, though others can be used. We perform k-means clustering on the pretraining dataset, query the closest cluster center to all the documents in the probe datasets, and calculate the percentage of each probe dataset in each cluster. This method is simple and fast; once a pretraining dataset has been clustered, many small probe datasets can be used.

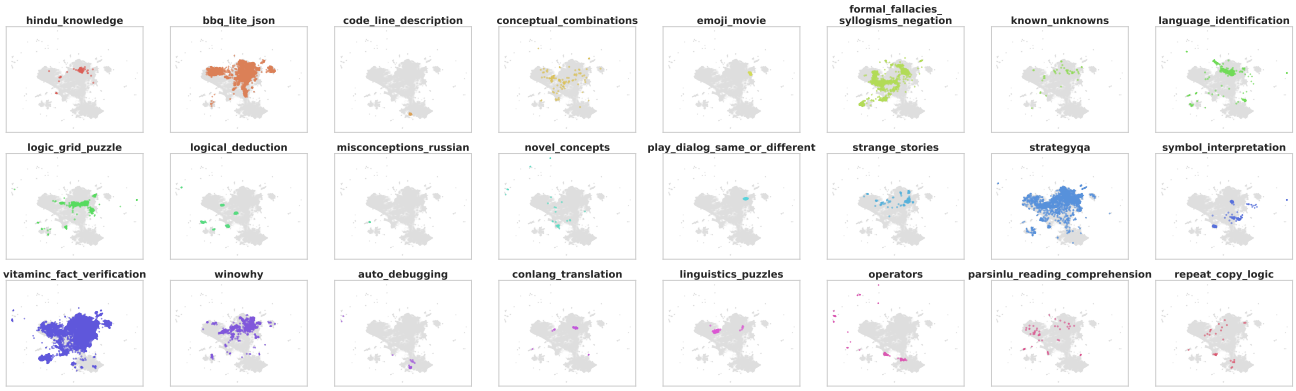

Figure 1. Probe datasets from the BIG-bench Lite suite map to different parts of the Pile. Each dataset is drawn with a different color.

Datasets. We use the Pile (Gao et al., 2020) as the pretraining dataset since it is publicly available and has been studied extensively (Xie et al., 2023). Appendix C performs the analysis on C4 (Raffel et al., 2020). Due to the datasets' size, we perform experiments on a 100,000-document sample of each. We probe with BIG-bench Lite, a set of 24 diverse downstream datasets that range in size from 16 to 43,735 documents (Srivastava et al., 2022).

3. Results

Figure 1 uses UMAP (McInnes et al., 2018) to visualize the overlap when probing the Pile with BIG-bench Lite. We UMAP the pretraining dataset and project the probe datasets into the resulting low-dimensional space. Different probe datasets cover different portions of the Pile.

Figure 2 shows the percentage of each of the 24 probe datasets that is in each of $k = 60$ clusters (Appendix A shows robustness to number of clusters). The probes categorize parts of the dataset. Clusters 0, 16, and 25 are covered by documents from many of the probe datasets, so these may be similar to the question-answering style. Others, like 20, 34, and 44, contain documents primarily from only a few probe datasets. For example, cluster 34 is similar to documents in language_identification and parsinlu (a Persian NLU dataset), two datasets that have a significant amount of non-Latin text, as documents in cluster 34 do upon inspection (see Appendix D for more examples). Clusters 0, 8, 34, and 40 account for only 8.5% of the pretraining data but 43.8% of all of the probe documents, indicating that additional probes could further characterize the Pile.

Each document in the Pile is annotated with one of 22 domains from which it was collected. Appendix B shows that dataset probes do not simply recover domains and that many domains are distributed across clusters. For example, cluster 44 contains several different domains, but it contains all the documents of symbol_interpretation.

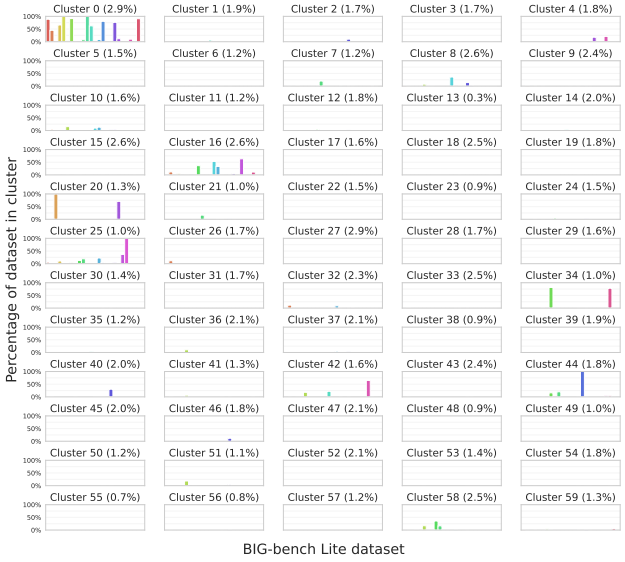

Figure 2. Documents from BIG-bench Lite datasets cover different clusters of the Pile. Percentages refer to the amount of the Pile in each cluster. Dataset colors correspond to Figure 1.

4. Conclusion and future work

We have demonstrated a data-centric method for characterizing large web-scraped pretraining datasets that leverages small curated datasets. We have focused on comparing pretraining datasets with better curated benchmark datasets, but the approach can be extended to other kinds of large and small datasets in modalities beyond text. Xie et al. (2023) and Longpre et al. (2023) have recently shown that heterogeneity at the domain level improves downstream performance, and our work may provide the basis for more targeted quality interventions. Future work could also build on training data attribution methods (Pruthi et al., 2020; Akyürek et al., 2022) rather than embeddings for a label-mediated notion of document similarity.

Acknowledgements

We would like to thank Daphne Ippolito, Katherine Lee, Daniel Smilkov, Nikhil Thorat, Ann Yuan, and Lucas Dixon.

References

Akyürek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., and Guu, K. Tracing knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*, 2022.

Assogba, Y., Pearce, A., and Elliott, M. Large scale qualitative evaluation of generative image model outputs. *arXiv preprint arXiv:2301.04518*, 2023.

Bandy, J. and Vincent, N. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *TACL*, 6:587–604, 2018.

Blei, D. M. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Viégas, F., and Wattenberg, M. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Crisan, A., Drouhard, M., Vig, J., and Rajani, N. Interactive model cards: A human-centered approach to model documentation. In *FAccT*, 2022.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *EMNLP*, 2021.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, 2022.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Li, W. and McCallum, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, 2006.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., and Ippolito, D. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. Quality not quantity: On the interaction between dataset design and robustness of CLIP. *arXiv preprint arXiv:2208.05516*, 2022.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *ACL Findings*, 2022.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *NeurIPS Workshop on Machine Learning Retrospectives, Surveys, and Meta-analyses*, 2020.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. *NeurIPS*, 2021.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. *NeurIPS*, 2020.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *FAccT*, 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

Schaul, K., Chen, S. Y., and Tiku, N. Inside the secret list of websites that make AI like ChatGPT sound smart. *Washington Post*, 2023. URL https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning.

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F. B., and Wattenberg, M. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, 2020.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., and Yu, A. W. DoReMi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023.

A. Robustness to number of clusters

Figure 3 repeats the analysis with 10 clusters, and Figure 4 does the same with 100 clusters. In both cases, we can see that there are a few clusters where most of the probe datasets fall. Using more clusters provides a finer-grained view of which pretraining documents are similar to which probe datasets.

B. Dataset probes are different from domain origin

Figures 5 and 7 (next page) repeat the analysis on the Pile but visualize the domain of documents in the Pile rather than probe with smaller downstream datasets (essentially probing with existing subsets of the Pile). Comparing with Figures 1 and 2 shows that probe datasets do not simply uncover the domain structure of the Pile.

C. Probing C4

Figures 6 and 8 (next page) reproduce the probing analysis with C4 as the pretraining dataset. We see qualitatively similar results as with the Pile. Clusters 9, 41, and 45 account for only 4.1% of the pretraining data but 48.3% of all of the probe documents. Clusters 41, 49, and 2 are covered by documents from many of the probe datasets, but clusters like 42 and 45 contain documents primarily from only 1 probe dataset. Most clusters (42 out of 60) contain fewer than 1% of the probe documents.

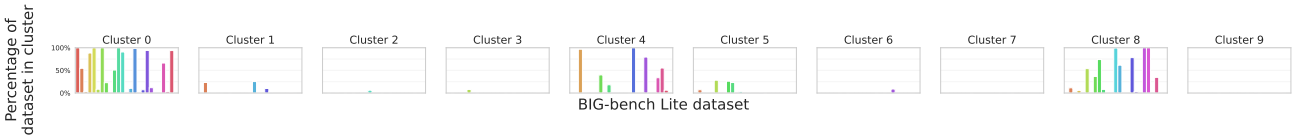

Figure 3. The distribution of BIG-bench Lite documents over $k = 10$ clusters. Compare to Figure 2. Dataset colors are the same as in Figures 1 and 2.

Figure 4. The distribution of BIG-bench Lite documents over $k = 100$ clusters. Compare to Figure 2. Dataset colors are the same as in Figures 1 and 2.

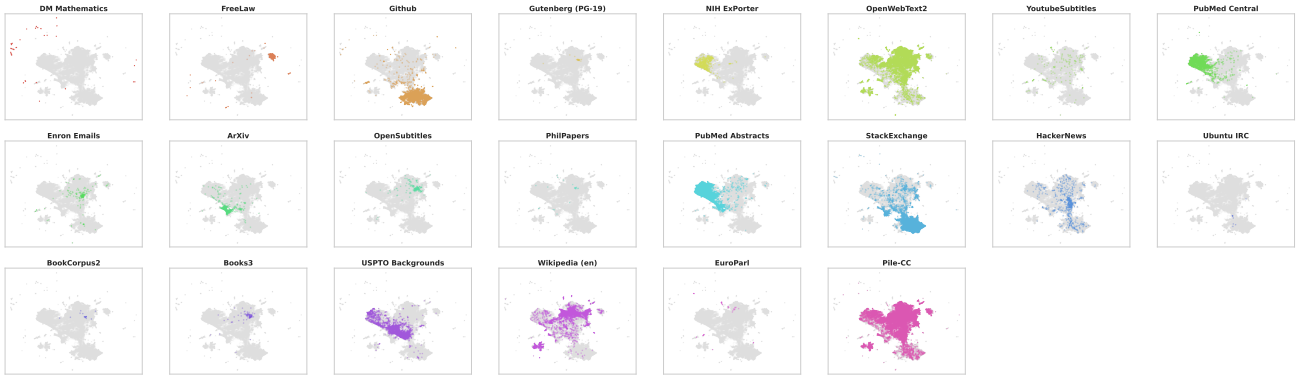

Figure 5. Domain source for documents in the Pile cluster in different parts of the embedding space. Comparing this figure to how Figure 1 probes the Pile with BIG-bench Lite datasets, we can see that BIG-bench Lite datasets do not only reproduce the domain structure.

Figure 6. Probe datasets from the BIG-bench Lite benchmark suite map to different parts of C4.

Figure 7. Clusters are typically composed of documents from just a few Pile domains. Percentages here are normalized **per cluster**, unlike in Figures 2 and 8. Domain colors correspond to Figure 5.

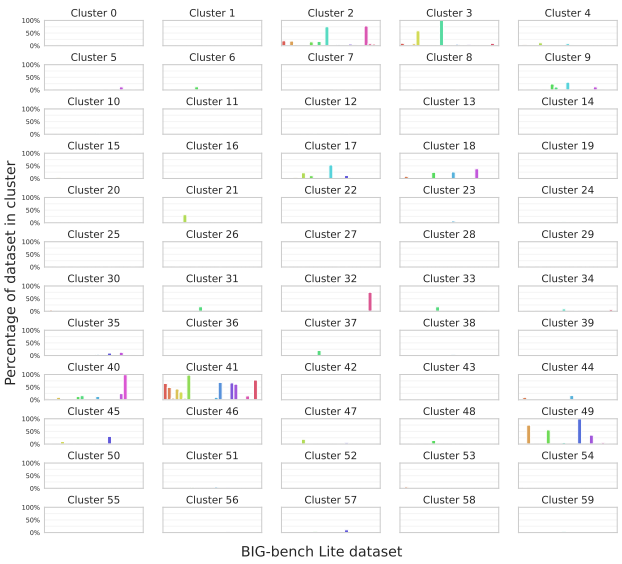

Figure 8. Most clusters of C4 are not covered by any BIG-bench Lite documents. The majority of BIG-bench Lite datasets are in just a few clusters. Dataset colors correspond to Figure 6.

D. Randomly sampled documents from example clusters

Here are the first 200 characters from each of four random examples from the clusters mentioned in Section 3.

Cluster 0:

- The three-piece girl group may not be back together in a reunion sense, but Beyoncé and Kelly Rowland contribute guest verses and make appearances in “Say Yes,” the new music video from Michelle Willi
- Some of the designs included in this collection are Pictured. You must have the necessary hardware and software to transfer these designs to your embroidery machine. Shipping for this Item is only $1.
- David Avrom Bell is a Princeton historian who discusses his recent scathing critique of Steven Pinker’s latest book, Enlightenment Now.

Cluster 16:

- Is it grammatically correct to say, She went missing"? What is the rule?" If the phrase troubles your ear, you’re not alone! You hear or see the expression a lot these days. When a person disappears,
- - The rules are set to keep order and Team SFI is here to enforce it. SFI strives to be a friendly and active community for international fans of SHINee. This requires effort from the members and the
- Tag Archives: Spider E-book & Grey Squirrel E-book February has been an amazing month. I love writing in the start of a fresh year, seems that all possibilities are achievable and the cold air is eno

Cluster 20:

- Q: pytest AttributeError: Metafunc instance has no attribute ’parameterize’ I am trying to parameterize my tests using pytest. I modified my conftest.py file to parameterize the fixture via pytest_g
- Class Title: Functions, generators and ducks Facilitator: TM Date: 060815 These notes are by: TM duck typing: note this may be much better left for OOP sessions explanation of duck typing failed, a m
- Q: How can I calculate variance of a very large random variable? I’m implementing an algorithm which recieves as input samples from a random variable with an unknown distribution. The random variabl

Cluster 25:

- Found cat didn’t know what to call it. Found here unisex cat names http://allcatsnames.com/unisex-kitten-names full list of names for cats. 2017-1-24 NO.11 Hi. Is it possible to order this head on t
- This Cat is a Superhero After Potentially Saving a Little Boys Life This cat is a Superhero. When a little boy was playing at home, a stray dog runs out of nowhere and viciously attacks him. Luckily
- If you live near rattlesnake habitat, there’s always the chance of you or your pets bumping into one. All encounters differ in many ways, but generally humans are more aware of what the creature is an

Cluster 34:

- Για την ανάγκη να υπάρξουν χαμηλότεροι φορολογικοί συντελεστές, αλλά και βελτίωση της φορολογικής συμμόρφωσης, έκανε λόγο ο Κυριάκος Μητσοτάκης, επισημαίνοντας πως «η Νέα Δημοκρατία είναι το κόμμα το
- Zobacz wideo Sprawą nagrania we wtorek nie zajmowała się prokuratura, ale, jak ustaliliśmy w Ministerstwie Spraw Wewnętrznych i Administracji, bada ją policja. REKLAMA Funkcjonariusze wystąpili do
- Como se había estimado, durante el mes pasado la fábrica de aerogeneradores de media y baja potencia ubicada en la localidad neuquina de Cutral Co ya está dando sus primeros pasos. El emprendimiento

Cluster 44:

- // // MHCommonLabelItemViewModel.m // WeChat // // Created by senba on 2017/9/21. // Copyright © 2017 CoderMikeHe. All rights reserved. // #import "MHCommonLabelItemViewModel.h" @implementation
- Q: Displaying attribute table on top I would like to pin (fix) the opened attribute table to the QGIS window or display the table on top. The attribute table is opening as another window (and it is
- Introduction Regular expressions are a well recognized way for describing string patterns. The following regular expression defines a floating point number with a (possibly empty) integer part, a non