1) M. Ghesmoune, M. Lebbah, and H. Azzag, "State-of-the-art on clustering data streams," *Big Data Analytics*, vol. 1, no. 1, 2016. (link)

This paper introduces several stream data clustering algorithms (GNG, GWR, IGNG, G-Stream, BIRCH, E-Stream, HUE-Stream, ClusTree, CluStream, StreamKM++, StrAP, DenStream, SOStream, SVStream, D-Stream).[1] It does a great job of describing the advantages and disadvantages of each succinctly. One thing that I personally disagree with however is that some algorithms with offline components are categorized as data stream algorithms. The entire advantage of running data stream algorithms is (1) that the algorithms are anytime – meaning they can stop at any point and there will be something to show for the effort, and (2) that they only look at one element at a time and thus are able to interact with obscenely large databases. This is one of the main reasons I gravitated towards the ClusTree algorithm presented in this paper, which had all the properties a data stream algorithm should. In summary, this is a great paper for a quick survey of the available data stream clustering algorithms, without going too in-depth for any particular one.

2) S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Networks*, vol. 15, no. 8-9, pp. 1041–1058, 2002. (link)

This paper dives deeper into the GWR algorithm described in the first source. It is compared to the GNG algorithm and another (GCS). In the third section the author describes the method of growth, mentioning details like removing edges and associated nodes as they become less relevant. It sounds like a greedy algorithm focused on distance, and I was interested to read that K-means would be used for optimization.[2] All in all the paper is well written. The author seems to know what they are talking about, and though it isn't a page turner it isn't as long-winded as it could be.

3) P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The ClusTree: indexing micro-clusters for anytime stream mining," *Knowledge and Information Systems*, vol. 29, no. 2, pp. 249–272, 2010. (link)

This paper is the reference for the ClusTree algorithm described in the survey paper above. Like the algorithm from the previous paper, this one also has a time decay function used filter out old nodes. I didn't expect to see the use of a buffer to wait and integrate multiple objects from the stream simultaneously. It certainly would be less expensive than load balancing with each and every insertion, but there is a potential delay introduced when flushing the buffer, say in the event of a prediction request. I was surprised to see a flow diagram in place of the usual textual algorithm. It is actually much easier to read, and not something I've seen often in these technical articles. The results were thorough. They went as far as tracking the number of micro clusters maintained by their algorithm, which is something that they could have just bundled under memory consumption. Unfortunately, a link to an implementation of that particular algorithm wasn't provided. This is just a nitpick, but there is some content at the end of the 4th section that is redundant with the conclusion section. This paper was actually a short read, but that's probably just because I'm interested in this particular algorithm.

4) S. Mansalis, E. Ntoutsi, N. Pelekis, and Y. Theodoridis, "An evaluation of data stream clustering algorithms," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 11, no. 4, pp. 167–187, 2018. (link)

This article was actually more helpful than the first survey paper. It described terms like micro clusters in detail (with the other paper I had to look them up separately). It actually walked through the online/

offline topic before discussing any specific algorithms. Of course, they also described several algorithms in detail (Stream, CluStream, ClusTree, StreamKM++, SWClustering, DenStream, rDenStream, FlockStream, HDDStream, PreDeConStream, D-Stream, MuDi-Stream, DENGRIS, SWEM).[4] Table 2 would be extremely helpful for someone trying to decide which algorithm to use for a particular [big data] application. Table 7 is what I was searching for the longest time for. There is no reason not to just pick an algorithm that checks all the boxes regardless of the application. All in all, this is the best-written paper of those we've critiqued thus far. I'd recommend it as a follow-up or just replacement reading for the other survey.

5) A. Amini and T. Y. Wah, "Density Micro-Clustering Algorithms on DataStreams: A Review," *Lecture Notes in Engineering and Computer Science*, vol. 1, 2011. (link)

I read this paper in attempt to understand micro clusters, after reading the first survey paper we critiqued. I wasn't aware that there were separate online and offline steps to perform with them. I just assumed their primary purpose would be minimizing performance and memory impacts online. Anyway, the paper seems a bit too brief, but perhaps that's because of the focused topic.

6) L. Liu and T. Peng, "Clustering-based topical Web crawling using CFu-tree guided by link-context," *Frontiers of Computer Science*, vol. 8, no. 4, pp. 581–595, 2014. (link)

I read this paper because I was strongly considering using Selenium WebDriver as a web crawler data stream to feed a data stream clustering algorithm. It goes into HTML specific content a bit too often for me, because personally that's the sort of thing I'd rather figure out as I develop (rather than need a guide to it in the article). The algorithm descriptions are useful though, and would be very useful to generate a dendrogram as part of the output from a web crawl. A java implementation surprised me. I'd expect c++ for something high performance, but then again it seems like a lot of the the big data tools are implemented in Java. I'm glad to see that they mentioned feature selection as a way to eliminate useless variables.

7) E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki, "Mixture of Gaussians for distance estimation with missing data," *Neurocomputing*, vol. 131, pp. 32–42, 2014. (link)

This paper was my primary focus for the deep dive. It starts by describing the advantages and disadvantages of several methods of imputation (conditional, random draw, multiple, kNN, improved case kNN, MLEM2, imputation with simultaneous classification).[7] Then other methods are described (studying input density indirectly through conditional mean distributions, feature selection, using incomplete samples as-is for NN, hidden Markov models for speech recognition).[7] Finally, guassian imputation is explained (finite mixture models, guassians for training neural networks with missing values, single multivariate guassian used in clinical trials, guassian mixture models for high density data).[7] The author states that guassians are flexible enough to handle any data distribution.[7] The EM algorithm is explained simply – that it is used to maximize the log likelihood[7], i.e. maximize the likelihood function, which measures how well the model fits the data. I disagree with the author then the statement is made that it is not limiting to require two samples to be uncorrelated (because of using squared Euclidian distance).[7] What about longitudinal studies? A few things could have been clearer in the results. It's not immediately obvious why some columns are missing for certain algorithms or what the values in parentheses are. The data set table was useful, but the proceeding results tables seemed a bit repetitive. It would have been nice to see a succinct explanation of all the results beforehand.

8) **S. van Buuren, *Flexible imputation of missing data*. Boca Raton: CRC Press, 2018. (link)**

I didn't review this whole book, just section 1.2, which I found very helpful when it came to understanding MCAR, MAR, and MNAR. The section was well written, concise, and to the point.

9) **T. P. Morris, I. R. White, and P. Royston, "Tuning multiple imputation by predictive mean matching and local residual draws," *BMC Medical Research Methodology*, vol. 14, no. 1, May 2014. (link)**

This was another article I just skimmed for an understanding of MCAR, MAR, MNAR, and multiple imputation. Mention of these terms were few and far between, so in hindsight a less verbose article may have been more suitable for this purpose.

10) **S. M. Brown, A. Duggal, P. C. Hou, M. Tidswell, A. Khan, M. Exline, P. K. Park, D. A. Schoenfeld, M. Liu, C. K. Grissom, M. Moss, T. W. Rice, C. L. Hough, E. Rivers, B. T. Thompson, and R. G. Brower, "Nonlinear Imputation of PaO2/FIO2 From SpO2/FIO2 Among Mechanically Ventilated Patients in the ICU," *Critical Care Medicine*, vol. 45, no. 8, pp. 1317–1324, 2017. (link)**

This article was skimmed for an understanding of non-linear imputation, which as it turns out isn't an easy thing to find online. Unfortunately it isn't explicitly defined in the document.