nature research

Peer Review Information

Journal: Nature Computational Science
Manuscript Title: Using Sequences of Life-events to Predict Human Lives
Corresponding author name(s): Professor Sune Lehmann

Editorial Notes:

Transferred manuscripts This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Computational Science

Reviewer Comments & Decisions:

Decision Letter, initial version:

Date: 21st August 23 16:25:44
Last Sent: 21st August 23 16:25:44
Triggered By: Fernando Chirigati
From: fernando.chirigati@us.nature.com
To: sljo@dtu.dk
BCC: fernando.chirigati@us.nature.com
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-23-0677-T
Message: ** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Professor Lehmann,

Your manuscript "Using Sequences of Life-events to Predict Human Lives" has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- All of the referees mention issues related to presentation. We would recommend merging subsections 2.1 and 2.2 into one subsection, which summarizes model and data development in enough detail that makes it easy to understand the results; more details can come in the Methods section. Also, make sure to highlight the main, most important and impactful results to make the research more clear to our broader audience. Overall, please make sure to follow our referees' suggestions on this matter.
- Please add a sensitivity/robustness analysis of the results with respect to the data selection criteria, as suggested by Referee #2.
- Referee #3 suggests the addition of other baselines experiments – comparing baseline models with other datasets for each of the prediction tasks – to highlight the contributions coming from the data itself. We believe that such comparisons are important and would strengthen the work.
- Referees mention several missing discussions; in particular, on the model's portability / broad applicability to other tasks and countries/regions; on the potential limitations of the data and its sampling period; on the heterogeneity across gender and age groups; on how the model will further computational social science as a field; and on the potential applications and implications of the study in general.

In addition to these points, it would also be beneficial to address the following concerns:

- Related to the issue of broad applicability, would it be possible and feasible to add experiments with similar or related data from other countries/regions to better show how well the proposed model (even without such a detailed data as the one used in the paper) works? This would likely boost the impact of the paper.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System

(MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

Reviewers comments:

Reviewer #1 (Remarks to the Author):

I thank the authors for their hard work on writing this manuscript and conducting the studies, and the editor for inviting me to review it.

This is a set of well-executed, clearly described, and interesting exploratory studies employing state-of-the art methods and a unique dataset. I learnt a lot from reading it and applaud the authors for their efforts. I really like the life2vec model and think that this approach would likely be a source of many useful insights and accurate predictions.

It is quite rare for me, but I have little in terms of critical feedback to offer to the authors. Sorry! I had some questions and concerns (e.g., initially I was not sure whether the authors conducted cross-validation/out-of-sample prediction), but I was able to address those concerns by reading the supplement.

I do not understand why the authors decided to predict a single personality item (there is more random error in a single item than in the entire scale). I would strongly recommend predicting all Big Five domain scores. The current setup of the prediction task may rise the eyebrows of social scientists (and some may even suspect that the authors opportunistically picked this single item as to maximize the prediction accuracy).

The following is a matter of taste and not criticism: It seems that this manuscript was initially written with CS audience in mind - perhaps it would be worth resuffling the contents a little bit to make it more easily palatable for the general audience. In particular, the goal of some of the analyses (e.g., the visual exploration of the mortality prediction space) is unclear to me. I know that in CS conference

proceedings scholars often list many "curious" results and interesting plots illustrating what's happening in the data - I have published a few manuscripts like that myself. Yet, this journal is read by broader readership that may expect more clarity in terms of what we are trying to prove here, how different analyses relate to each other, etc. I would even risk a statement that eliminating (i.e., moving to supplement) some of the side plots/analyses to focus on the core findings (predictability / diagnostic utility of the life2vec approach) may boost the impact of the paper.

Reviewer #2 (Remarks to the Author):

Key results:

The paper presents a transormer-based deep neural network model for embedding life-related tokens and representing life-sequences as vectors. These vectors can be used in downstream machine learning tasks to predict e.g., early mortality, personality types, professional occupation, etc. In addition, by using state-of-the-art model inspection tools, the model features can be clustered to gain scientific insights in e.g., features related to increased early mortality, or effects of the features on model's predictions. Using this model the authors present results of such inspection for early mortality and personality type prediction tasks.

Validity:

Validation is very solid. The auhtors include all the necessary statistical significance tests, compare their model with a range of the state-of-the-art and baseline models and show an increased performace of their model. The loss functions and the model's architecture are presented with great detail. I have only missed a more detailed robustness analysis w.r.t. to data filtering and preprocessing. For example, the authors presents four data selection criteria on pages 14-15 (section 4.2) but do not make additional experiment nor do they discuss the sensitivity/robustness of the results w.r.t. those filters.

Significance:

The paper, the results and in the particular the model have a potentially significant impact. With more and more of our lives and our activities being subjects to prediction from machine learning models, the work presented in the paper is a very important step in a systematic and responsible way to approach this development.

Data:

The data is impressive containg all the health/occupation records of the entire population of a European country. It is understandable that this data can not be made accesible to the general public. The authors show in an impressive way what is possible with data of such quality, granularity, and scale.

Methodology:

The methods used are state-of-the-art and clearly demonstrate how deep learning

methods used in another domain (natural language processing) can be adapted to additional domains and tasks. The machine learning aspects of the paper are very well described, systematically applied and analyzed (e.g., trhough hyperparameter optimization) and compared with additional ML as well as traditional models.

Analytical approach:

Statistical treatment of the results is excellent.

Suggested improvements:

My suggestions are mainly concerned with presentation issues. In particular:
- Section 2 (Results) presents a mixture of methods (model presentation and validation) and the results. I would consided sections 2.1 and 2.2. to be model presentation and not results. I understand that the model needs to be described first before the results can be presented but the current presentation is both too long for such early sections in the paper and still leaves some important details of the model not clearly described. For example, the idea of segments is not clear from the sections 2.1 and 2.2. Therefore, I suggest compressing the sections 2.1 and 2.2 to a single (one page) section but trying to describe all the elements (and only those) that are needed to understand the sections 2.4 and 2.5. Also, section 2.3 is the validation of the model and can be potentially moved to section 4 in order to get to the results (2.4 and 2.5) earlier.
- Description of artifical language and one example (similar to E.7 can be also a part of section 4)
- I am still not completly clear what are the A, B, C segments? Are these parts of a day? How they are labeled? What happens if two events fall into the same segment?
- Some minor issues. Rewrite the first sentence at the top of the second page ('While it is known...'). Currently, it is difficult to understand. Also, page 42, point 4, you are missing closing brackets.

In addition, I would also like to see some additional discussion on the model's portability. How would we use the model for e.g., prediction professional occupation, academic success, etc.? Would be possible to use the token embeddings and embed people from another country, another region, etc?

Clarity and context:

Clarity, see above. The related work is very well cited.

My expertise:

I am knowledgeble in all ML/statistical methods the authors use in the paper.

Reviewer #2 (Remarks on code availability):

The code is well documented and professionaly written. Apart from the data not being available the experiments seem to be fully reproducible.

Reviewer #3 (Remarks to the Author):

Review report for "Using Sequences of Life-events to Predict Human Lives" by Savcisens et al.

I love this paper. I believe this paper has the potential to become a landmark paper in my field. I can see it being commonly referred to as the "life2vec" paper, in the years to come. Over the past decade, transformer-based models have transformed a wide range of fields and domains. Here the authors apply the model to the Danish registry data, which captures life trajectories at a remarkable level of scale and detail. They show that life2vec offers superior predictive power for mortality risks and personality nuances, and the associated embedding space further offers insights into individual life trajectories and developments. But I feel the key to thinking about the contribution of this paper is through the many new directions that this paper will open. Indeed, I can't help but wonder how the authors could use life2vec to further their established lines of research on social networks and human mobility. For my own field of the science of science, I also see myself wondering if one could combine this approach with publication data or inventor information – to unpack the secrets of invention, creativity, collaborations, and more. Of course, some of these will necessarily be limited by privacy and other data restrictions. But my point is, when combined with the many advances in computational social science, the possibilities are seemly endless. And life2vec may as well open a new chapter in our quantitative understanding of human behavior. Overall, I recommend the paper for publication. My comments below should only be considered as further suggestions to amplify the impact and contribution of the paper.

While the data used in this paper is impressive, it's important to outline its potential limitations. For example, this paper uses Danish national registers from 1st January 2008 until 31st December 2015. This is an important fact that should be highlighted in the main text and deserves further discussions (currently it's mentioned in Figure caption and the method section 4.2). For example, what's the implication of this sampling period? For people who were born before 2008, what information then do you have on them? How would that affect the results of the paper? For people of different ages, you would have different lengths of observations, either in the absolute sense or relative sense. How would that affect the prediction tasks and their evaluations? Also for the mortality prediction task, it seems you're predicting whether a person will be alive in 2020. Is it true? Then where does the ground truth come from in this case, if your data stops at 2015? I feel there should be a paragraph early in the main text to highlight the various potential issues and concerns as well as in the discussion section to talk about potential limitations.

I urge the authors to think more and more carefully about the overall setup of the prediction tasks. Namely, what should you compare against? Conceptually, I see two advances in this paper: data & model. Right now, the paper compares itself with other ML models trained on their data, which is good. But in doing so, it also narrows its contribution to just the model, missing the opportunity to highlight the importance of their data and its potential to reveal fundamental insights into human behavior. At some level, one may think the predictive power lies in the data rather than the method, where life2vec is just a powerful way of aggregating the remarkably diverse range of information that you have about an individual's life trajectory. Following this thought, then one might ask how well your method will perform based on models trained on just the health data. This also gets to the question of what's the baseline

here. The current paper suggests that the baseline is other ML models trained on their data. But if one agrees that one major advance of this paper is its data, which integrates numerous dimensions of life – age, income, occupation, health, etc – then one could compare the predictions with baselines that just look at the health information to predict mortality. In other words, maybe the predictive power comes not from one single factor but a combination of many facets in life. Maybe you've tried to do this already with "life table", but that variable seems a bit too crude. This would further highlight the value of your data and how this paper represents a fundamentally new approach to understanding human behavior and life outcomes.

Figure 2D shows large heterogeneity across gender and age groups. You wrote that "the model performs better on a younger cohort of the population and on a cohort of females." Can you say more about this? For example, from the figure it seems that the differences between female and male are not always significant. Also the variances seem to be larger for young cohorts. Could you explain more about why this is the case? Also for Figure 3, it seems that for Q1 and Q4, the life2vec model is not significantly better than RNN. Do you have any thoughts on why that is? Of course here it also relates to my preceding comment. If you can compare against models trained on "competing data", it would also help you bypass this issue.

I also feel section 2.3 should be more tightly integrated with sections 2.4&2.5. Right now they feel separated. After reading 2.3, I find myself dying to learn about what explains these superior predictive powers of life2vec, and the embedding space analyses can be quite informative in this regard.

The methods section feels too scattered. There are various subsections and short paragraphs. But for many of them, it's hard to find an internal logic that leads from one to another. Instead, I find myself jumping from one section to another, quite disoriented. Overall, I feel the methods section can be much better organized. And again, that would be important given the potential contribution of the paper – if the paper is as successful as I expect it to be, many of your readers will read through the methods section in detail.

Can you elaborate more in the discussion section on how life2vec will further computational social science as a field?

Hope these comments are helpful. I want to congratulate the authors for breaking new ground with this paper.

Author Rebuttal to Initial comments

Editor

E1 While we ask you to address all of the points raised, the following points need to be substantially worked on:

Thank you for clarifying the key points to focus on. We have indeed answered all of the reviewer comments in detail, see below. We have paid extra attention to the points you raise below. As always, we summarize the changes we have implemented in the revised MS in the point-by-point responses.

E2 All of the referees mention issues related to presentation. We would recommend merging subsections 2.1 and 2.2 into one subsection, which summarizes model and data development in enough detail that makes it easy to understand the results; more details can come in the Methods section. Also, make sure to highlight the main, most important and impactful results to make the research more clear to our broader audience. Overall, please make sure to follow our referees' suggestions on this matter.

This is an important point! Looking back at our initial MS, we see that there was ample room for improvement.

We have indeed merged and focused sections 2.1 and 2.2. Further we have updated the overall flow of the MS to make the research easier to understand for a broad audience, and moved many technical details to the Methods section, as suggested. In general, we have followed the guidance provided by the reviewers for the rewrite & restructuring.

E3 Please add a sensitivity/robustness analysis of the results with respect to the data selection criteria, as suggested by Referee #2.

We agree with the importance of including more robustness tests and have added a number of new results on these topics.

Specifically, as suggested by Reviewer 2, we did a complete robustness evaluation of the concept space (see Tab. A5, also included below). We did this by pretraining life2vec on different subsets of data, varying both the age of individuals in the training data and the duration of the life-sequences. Then, we confirmed that the concept space converges to a similar structure in all cases. (For this, we use the same procedure we previously applied to determine that the concept space was robust with respect to the overall population used to train the model, see Methods: Evaluation of the Concept Space).

Table A5: Similarity of the concept space. Comparison of the original concept space (generated from pretraining on the full dataset described in the Methods Chapter) to the concept space generated with different dataset versions (after 30 epochs of pretraining). We follow the evaluation procedure described in the Methods Chapter to calculate the significance values (p-values are adjusted with the Benjamini-Hochberg procedure).

Data Composition	Spearman's ρ	Corrected p-value
Age: 25-35 Period: 2008-2015	.672	<.005
Age: 45-55 Period: 2008-2015	.691	<.005
Age: 55-70 Period: 2008-2015	.651	<.005
Age: 25-70 Period: 2008-2011	.687	<.005

E4 Referee #3 suggests the addition of other baselines experiments – comparing baseline models with other datasets for each of the prediction tasks – to highlight the contributions coming from the data itself. We believe that such comparisons are important and would strengthen the work.

This is another excellent suggestion that we have worked hard to make sure we fully make use of.

To investigate the role of the specific parts of the dataset, we evaluated *life2vec* on different sub-versions of our dataset formed by excluding tokens from the vocabulary. Beyond the full dataset, we look at **only** labor data (salary, workplace, industry, position, etc.), partial labor data that does not use information about the workplace, industry, position, and labor force, and finally, a version that uses only partial labor data and health data.

In Table A3 (reproduced below), we show the performance of the model on these datasets. The predictive performance degrades as we remove more tokens, and highlighting that *life2vec* does indeed use information from the labor events (as opposed to only health-data) to predict the mortality.

Table A3: Corrected Matthew's Correlation Coefficient (MCC) and Area under the Lift (AUL) on the Mortality Prediction Task. 95%-Confidence intervals for the MCC based on the stratified bootstrapping. Evaluation of the `life2vec` model on various data. Models are pretrained from scratch using these datasets and then finetuned accordingly. **Full Labour & Health** corresponds to the `life2vec` model described in the main manuscript. **Only Full labor** corresponds to the dataset without any health-related events. **Partial Labor & Health** contains health-related events and partial-labor events (without the specification of Work Industry, Sector, Position and Labour Force). **Partial Labor** consists only of partial-labor events (without the specification of Work Industry, Sector, Position and Labour Force).

Data	MCC, 95%-CI	AUL	Accuracy, 95%-CI	F1-Score, 95%-CI	Vocab Size
Full Labor & Health	**0.413 [0.410, 0.422]**	**0.845**	**0.788 [0.782, 0.794]**	**0.443 [0.435, 0.450]**	**2043**
Partial Labor & Health	0.375 [0.367, 0.384]	0.837	0.784 [0.777, 0.790]	0.399 [0.393, 0.406]	1034
Only Full Labor	0.319 [0.312, 0.327]	0.809	0.764 [0.758, 0.770]	0.340 [0.335, 0.345]	1290
Only Partial Labor	0.278 [0.271, 0.285]	0.782	0.737 [0.731, 0.743]	0.305 [0.300, 0.310]	281

E5 Referees mention several missing discussions; in particular, on the model's portability / broad applicability to other tasks and countries/regions; on the potential limitations of the data and its sampling period; on the heterogeneity across gender and age groups; on how the model will further computational social science as a field; and on the potential applications and implications of the study in general.

Yes. This is a good point. In the restructured MS, we have added a subsection (*life2vec as a foundation model*) that focuses on many of these: portability to other regions and the role of the training data discussed above. In the revised *Discussion* section, we mention further limitations and have expanded on how the model will further computational social science as a field.

E6 In addition to these points, it would also be beneficial to address the following concerns:

- Related to the issue of broad applicability, would it be possible and feasible to add experiments with similar or related data from other countries/regions to better show how well the proposed model (even without such a detailed data as the one used in the paper) works? This would likely boost the impact of the paper.

We also discuss this issue in the new *life2vec as a foundation model* section. The summary is that while there are no similar or related data we can explore, there are indications in the literature that the embedding spaces could be portable, and the model could be useful in other contexts in that sense.

Reviewer 1

R1.1 I thank the authors for their hard work on writing this manuscript and conducting the studies, and the editor for inviting me to review it.

This is a set of well-executed, clearly described, and interesting exploratory studies employing state-of-the art methods and a unique dataset. I learnt a lot from reading it and applaud the authors for their efforts. I really like the life2vec model and think that this approach would likely be a source of many useful insights and accurate predictions.

It is quite rare for me, but I have little in terms of critical feedback to offer to the authors. Sorry! I had some questions and concerns (e.g., initially I was not sure whether the authors conducted cross-validation/out-of-sample prediction), but I was able to address those concerns by reading the supplement.

Thank you for the kind words!

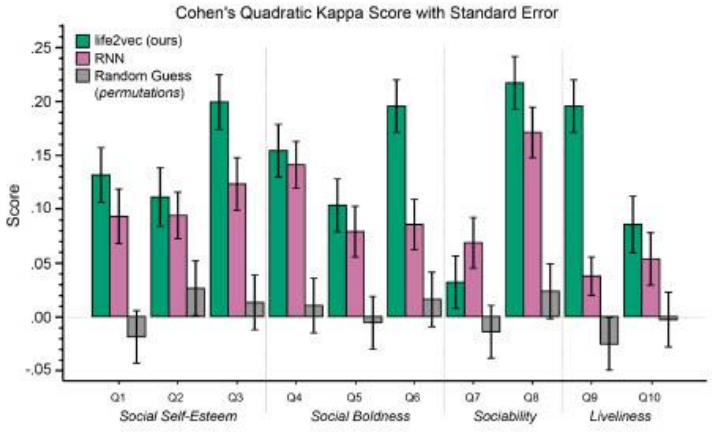
R1.2 I do not understand why the authors decided to predict a single personality item (there is more random error in a single item than in the entire scale). I would strongly recommend predicting all Big Five domain scores. The current setup of the prediction task may rise the eyebrows of social scientists (and some may even suspect that the authors opportunistically picked this single item as to maximize the prediction accuracy).

This is a good point! Actually, we initially focused on predicting the aggregated personality scores (extraversion, neuroticism, etc.), but after consulting with personality psychologists, we changed this strategy.

It turns out that when we consider at life-outcomes, individual questionnaire items are more appropriate. The most recent psychology research recommends focusing on outcomes rather than aggregated measures when analyzing life-outcomes [1-3], because individual items tend to be more correlated with life-outcomes than aggregate measures.

For that reason, we hope the Reviewer understands that we have chosen to continue to focus on individual items in the revised manuscript.

That said, we do realize that it could appear as “cherry-picking” that we selected a random item per extraversion-facet. To avoid this perception as well as to ensure that we are being as general as possible, we now predict all items from the extraversion personality dimension. The results are shown below.



Cohen's Quadratic Kapa Score with Standard Error (per Personality Item): Overview of performance on the items belonging to the Extraversion Facets. Models predict all ten items simultaneously (i.e., it is a multiclass and multitarget prediction task).

In almost all cases, our performance is better than RNN, which is the best of the competing methods, providing a more general foundation for our claims on personality prediction.

Note that In Q7, RNNs fare marginally better than *life2vec*, we address this in the revised main text as well as discuss the overall differences between RNNs and *life2vec* in the new section "*life2vec* as a foundation model".

Below, for the Reviewer's convenience, we include all of the actual questions related to extraversion. Those are answered on a scale from 1 to 5, where 1 represents "strongly disagree," and 5 is "strongly agree."

Overview of questions:

1. I feel that I am an unpopular person
2. I feel reasonably satisfied with myself overall
3. I sometimes feel that I am a worthless person
4. When I'm in a group of people, I'm often the one who speaks on behalf of the group

5. In social situations, I'm usually the one who makes the first move
6. I rarely express my opinions in group meetings
7. The first thing that I always do in a new place is to make friends
8. I prefer jobs that involve active social interaction to those that involve working alone
9. Most people are more upbeat and dynamic than I generally am
10. On most days, I feel cheerful and optimistic

Overview of the Extraversion facets [4] :

1. Social Self-Esteem: tendency to have positive self-regard
2. Social Boldness: comfort within a variety of social situations
3. Sociability: tendency to enjoy social interactions
4. Liveliness: one's typical enthusiasm and energy

References:

1. René Mõttus, Timothy C Bates, David M Condon, Daniel K Mroczek, and William R Revelle. Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. 2022
2. Anne Seeboth and René Mõttus. Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. European Journal of Personality, 32(3):186–201, 2018
3. Ross David Stewart, René Mõttus, Anne Seeboth, Christopher John Soto, and Wendy Johnson. The finer details? the predictability of life outcomes from big five domains, facets, and nuances. Journal of personality, 90(2):167–182, 2022
4. Lee, Kibeom, and Michael C. Ashton. "Psychometric properties of the HEXACO personality inventory." Multivariate behavioral research 39.2 (2004): 329-358.

R1.3 The following is a matter of taste and not criticism: It seems that this manuscript was initially written with CS audience in mind - perhaps it would be worth resuffling the contents a little bit to make it more easily palatable for the general audience. In particular, the goal of some of the analyses (e.g., the visual exploration of the mortality prediction space) is unclear to me. I know that in CS conference proceedings scholars often list many "curious" results and interesting plots illustrating what's happening in the data - I have published a few manuscripts like that myself. Yet, this journal is read by broader readership that may expect more clarity in terms of what we are trying to prove here, how different analyses relate to each other, etc. I would even risk a statement that eliminating (i.e., moving to supplement) some of the side plots/analyses to focus on the core findings (predictability / diagnostic utility of the life2vec approach) may boost the impact of the paper.

This is a really nice and constructive comment. We have substantially rewritten and restructured the revised main text of the revised version of the manuscript, taking into account these comments. We hope the reviewer likes the updated version!

Reviewer 2

R2.1 Key results:

The paper presents a transormer-based deep neural network model for embedding life-related tokens and representing life-sequences as vectors. These vectors can be used in downstream machine learning tasks to predict e.g., early mortality, personality types, professional occupation, etc. In addition, by using state-of-the-art model inspection tools, the model features can be clustered to gain scientific insights in e.g., features related to increased early mortality, or effects of the features on model's predictions. Using this model the authors present results of such inspection for early mortality and personality type prediction tasks.

Thank you for the nice summary.

R2.2 Validity:

Validation is very solid. The auhtors include all the necessary statistical significance tests, compare their model with a range of the state-of-the-art and baseline models and show an increased performace of their model. The loss functions and the model's architecture are presented with great detail. I have only missed a more detailed robustness analysis w.r.t. to data filtering and preprocessing. For example, the authors presents four data selection criteria on pages 14-15 (section 4.2) but do not make additional experiment nor do they discuss the sensitivity/robustness of the results w.r.t. those filters.

This is a very good point and connects to comments from other reviewers. To address this comment, we have run a number of new experiments related to the underlying data and filtering.

The four criteria are

1. participants are alive and lived in Denmark on the 31st of December 2015,
2. have at least 12 records in the labor data during the year 2015,
3. have consistent sex and birthday attributes over the whole residency period,
4. are between 25 and 70 years old on the 31st of December 2015,
5. Data runs from 2008-2015.

Here, we have added a "fifth" criterion about the date-span of the data set. We have now varied these and examined the resulting embedding spaces.

Criterion 2 and 3 are difficult to vary, as these concern 'unusual' individuals, e.g., people who have moved in and out of Denmark, who have very little or inconsistent data (only around 3% of all individuals have fewer than 12 labor events over 2015). Meanwhile, around 50% of people have less than 108 labor events, and around 99% have less than 500 labor events from 2008-2015.

Specifically, Table A5 in the appendix (also included below) provides an overview. We pretrained the model on different variations of a dataset by varying the inclusion requirements:

a. Individuals between 25-35 years old and a timespan of 2008-2015 (criterion 1,4),
b. Individuals between 45-55 years old and a timespan of 2008-2015 (criterion 1,4),
c. Individuals between 55-70 years old and a timespan: 2008-2015 (criterion 1,4),
d. Individuals between 25-70 years old and a timespan of 2008-2011 (criterion 5)

We find that the space of concepts is indeed robust and – for every sub-dataset – we find a high correlation with the original concept space (using the procedure outlined in Methods: Robustness of the Concept Space).

Table A5: Similarity of the concept space. Comparison of the original concept space (generated from pretraining on the full dataset described in the Methods Chapter) to the concept space generated with different dataset versions (after 30 epochs of pretraining). We follow the evaluation procedure described in the Methods Chapter to calculate the significance values (p-values are adjusted with the Benjamini-Hochberg procedure).

Data Composition	Spearman's ρ	Corrected p-value
Age: 25-35 Period: 2008-2015	.672	<.005
Age: 45-55 Period: 2008-2015	.691	<.005
Age: 55-70 Period: 2008-2015	.651	<.005
Age: 25-70 Period: 2008-2011	.687	<.005

Further, we have also looked at the role of the dataset in terms of prediction, so we have pre-trained and fine-tuned the model on the Early Mortality Prediction task, studying how different versions of the vocabulary, ranging from *only* a limited set of labor information to the full dataset impact the predictive ability. These experiments are in Appendix Table A3 (reproduced below).

Table A3: Corrected Matthew's Correlation Coefficient (MCC) and Area under the Lift (AUL) on the Mortality Prediction Task. 95%-Confidence intervals for the MCC based on the stratified bootstrapping. Evaluation of the `life2vec` model on various data. Models are pretrained from scratch using these datasets and then finetuned accordingly. **Full Labour & Health** corresponds to the `life2vec` model described in the main manuscript. **Only Full labor** corresponds to the dataset without any health-related events. **Partial Labor & Health** contains health-related events and partial-labor events (without the specification of Work Industry, Sector, Position and Labour Force). **Partial Labor** consists only of partial-labor events (without the specification of Work Industry, Sector, Position and Labour Force).

Data	MCC, 95%-CI	AUL	Accuracy, 95%-CI	F1-Score, 95%-CI	Vocab Size
Full Labor & Health	**0.413 [0.410, 0.422]**	**0.845**	**0.788 [0.782, 0.794]**	**0.443 [0.435, 0.450]**	**2043**
Partial Labor & Health	0.375 [0.367, 0.384]	0.837	0.784 [0.777, 0.790]	0.399 [0.393, 0.406]	1034
Only Full Labor	0.319 [0.312, 0.327]	0.809	0.764 [0.758, 0.770]	0.340 [0.335, 0.345]	1290
Only Partial Labor	0.278 [0.271, 0.285]	0.782	0.737 [0.731, 0.743]	0.305 [0.300, 0.310]	281

We find that the performance of the model degrades as expected, with the lowest performance on the partial labor dataset. We also see that including Health information significantly boosts the performance of the model.

Including this robustness analysis has further strengthened our confidence in the model, and we hope the reviewer finds them satisfactory.

R2.3 Significance:

The paper, the results and in the particular the model have a potentially significant impact. With more and more of our lives and our activities being subjects to prediction from machine learning models, the work presented in the paper is a very important step in a systematic and responsible way to approach this development.

Thank you for the kind words!

R2.4 Data:

The data is impressive containg all the health/occupation records of the entire population of a European country. It is understandable that this data can not be made

accesible to the general public. The authors show in an impressive way what is possible with data of such quality, granularity, and scale.

We agree and are thankful for this comment. Simultaneously, we underscore that data are, in fact, available to other researchers via Statistics Denmark (an affiliation to a Danish University is required). We are currently working on perhaps releasing the embedding spaces within this same context.

R2.5 Methodology:

The methods used are state-of-the-art and clearly demonstrate how deep learning methods used in another domain (natural language processing) can be adapted to additional domains and tasks. The machine learning aspects of the paper are very well described, systematically applied and analyzed (e.g., through hyperparameter optimization) and compared with additional ML as well as traditional models.

Thank you so much for this kind comment.

R2.6 Analytical approach:

Statistical treatment of the results is excellent.

Fantastic!

R2.7 Suggested improvements:

My suggestions are mainly concerned with presentation issues. In particular:
- Section 2 (Results) presents a mixture of methods (model presentation and validation) and the results. I would consided sections 2.1 and 2.2. to be model presentation and not results. I understand that the model needs to be described first before the results can be presented but the current presentation is both too long for such early sections in the paper and still leaves some important details of the model not clearly described. For example, the idea of segments is not clear from the sections 2.1 and 2.2. Therefore, I suggest compressing the sections 2.1 and 2.2 to a single (one page) section but trying to describe all the elements (and only those) that are needed to understand the sections 2.4 and 2.5. Also, section 2.3 is the validation of the model and can be potentially moved to section 4 in order to get to the results (2.4 and 2.5) earlier.

Thank you for your careful thoughts on how we can restructure the paper.

This advice aligns with comments from the other Reviewers. We have followed your advice and completely restructured the paper. We hope the reviewer finds that the revised MS is improved.

R2.8 Description of artifical language and one example (similar to E.7 can be also a part of section 4)

Great idea. We have rewritten the MS, also incorporating this idea.

R2.9 I am still not completly clear what are the A, B, C segments? Are these parts of a day? How they are labeled? What happens if two events fall into the same segment?

Thank you for pointing out the lack of clarity in the MS.

The segments are inspired by the Next Sentence Prediction Task. In earlier BERT versions, two sentences were passed, and the model had to predict whether these two sentences followed each other or not. Every token of the 1st sentence would be part of the 1st segment, etc.

In *life2vec*, segments mark separate events (to help distinguish between events on separate days). We want the assignment of segments to be *independent* of *any covariates* (since they do not have any particular meaning). We therefore chose to include 3 *segment types* since it is rare to have more than three events on the same day (but we could chosen a different number of segments).

The segment assignment starts with A (each token of the first event is marked as a segment A), the next event is marked B (even if this event happens on the next day), the next event is marked C, the next is marked A and so on.

We thus ensure that (1) in case three events happen on the same day, each event has a different segment, (2) any single event is equally likely to be marked as either A, B, or C independent of the number of events on a particular day.

All that being said, this is probably too much detail for this point in the MS, and we have removed a discussion from the main text and added a more detailed description of segments to Methods (also see SI: E7).

R2.10 Some minor issues. Rewrite the first sentence at the top of the second page ('While it is known...'). Currently, it is difficult to understand. Also, page 42, point 4, you are missing closing brackets.

Thank you for the attention to detail. We have fixed all these!

R2.11 In addition, I would also like to see some additional discussion on the model's portability. How would we use the model for e.g., prediction professional occupation, academic success, etc.? Would be possible to use the token embeddings and embed people from another country, another region, etc?

Good point! We have added an entire new section to the revised main text which discusses aspects of generalizability. There we discuss the portability of the embeddigns in details

R2.12 Clarity and context:

Clarity, see above. The related work is very well cited.

My expertise:

I am knowledgeble in all ML/statistical methods the authors use in the paper.

Reviewer #2 (Remarks on code availability):

The code is well documented and professionaly written. Apart from the data not being available the experiments seem to be fully reproducible.

Thank you again for all the helpful comments!

Reviewer 3

R3.1 I love this paper. I believe this paper has the potential to become a landmark paper in my field. I can see it being commonly referred to as the "life2vec" paper, in the years to come. Over the past decade, transformer-based models have transformed a wide range of fields and domains. Here the authors apply the model to the Danish registry data, which captures life trajectories at a remarkable level of scale and detail. They show that life2vec offers superior predictive power for mortality risks and personality nuances, and the associated embedding space further offers insights into individual life trajectories and developments. But I feel the key to thinking about the contribution of this paper is through the many new directions that this paper will open. Indeed, I can't help but wonder how the authors could use life2vec to further their established lines of research on social networks and human mobility. For my own field of the science of science, I also see myself wondering if one could combine this approach with publication data or inventor information – to unpack the secrets of invention, creativity, collaborations, and more. Of course, some of these will necessarily be limited by privacy and other data restrictions. But my point is, when combined with the many advances in computational social science, the possibilities are seemly endless. And life2vec may as well open a new chapter in our quantitative understanding of human behavior. Overall, I recommend the paper for publication. My comments below should only be considered as further suggestions to amplify the impact and contribution of the paper.

Thank you so much for these encouraging comments. Having worked hard on this project for years, it is genuinely fantastic to see a review that appreciates the work.

R3.2 While the data used in this paper is impressive, it's important to outline its potential limitations. For example, this paper uses Danish national registers from 1st January 2008 until 31st December 2015. This is an important fact that should be highlighted in the main text and deserves further discussions (currently it's mentioned in Figure caption and the method section 4.2).

This is an excellent point, and we have emphasized this in the paper, both in the introduction and in the discussion.

R3.3 For example, what's the implication of this sampling period? For people who were born before 2008, what information then do you have on them? How would that affect the results of the paper? For people of different ages, you would have different lengths of observations, either in the absolute sense or relative sense. How would that affect the prediction tasks and their evaluations? Also for the mortality prediction task, it seems

you're predicting whether a person will be alive in 2020. Is it true? Then where does the ground truth come from in this case, if your data stops at 2015? I feel there should be a paragraph early in the main text to highlight the various potential issues and concerns as well as in the discussion section to talk about potential limitations.

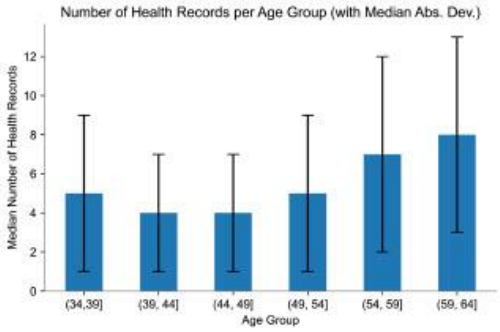
These are all excellent questions. Below, we answer each one in turn.

R3.3a: For people who were born before 2008, what information then do you have on them? How would that affect the results of the paper?

Answer: While DST has information about the labor information about residents before 2008, we decided on a cutoff of 1st January 2008. Data before 2008 is available, but it is aggregated on a yearly basis (and thus not available on the same level of granularity). Meanwhile, detailed health information is available before 2008. However, we set a cutoff on the 1st of January 2008 to align it with the labor data. It is entirely possible that including earlier data could boost the performance of the model

R3.3b: For people of different ages, you would have different lengths of observations, either in the absolute sense or relative sense.

Answer: Since we limit the timespan from 2008 to 2015 and set the age requirement - most people in our dataset have more than 80 records (at a minimum, 12 salaries per year over the 8-year period. When it comes to health-events, they are indeed a bit more frequent in sequences of older people, but the difference is relatively low (an average of ~5 in the younger cohort to an average of ~7 in the older cohort), see plot below. So, actually, the number of observations is quite balanced across age groups, both in a relative and absolute sense.



Note that the larger number of health-observations for the youngest group could potentially indicate that this group is more active in terms of contacting health professionals when they have symptoms of disease.

R3.3c: How would that affect the prediction tasks and their evaluations?

Answer: While age is not strongly correlated with the number of observations, there is a significant heterogeneity in sequence lengths (notice the large error bars in the plot above). For the mortality prediction task, we do consider model performance in terms of ***age*** in Figure 2D, as a function of ***sequence length*** in Figure 2B, and as a function of ***number of health events*** in Figure 2C.

We find that more health events tend to improve predictive performance and that performance depends slightly on age. Based on the results of the auditing of the early-mortality predictions, however, sequence length does not have a significant effect on the performance of the model.

R3.3d: Also for the mortality prediction task, it seems you're predicting whether a person will be alive in 2020. Is it true? Then where does the ground truth come from in this case, if your data stops at 2015?

Answer: We have access to data up to the current time, so we have ground truth labels (i.e., mortality and emigration data) for the period 2015-2020. The reason why we cut sequences at the end of 2015 is to avoid any kind of information leakage (when predicting mortality). We have updated the main text to make this aspect clear.

Once again, we fully agree that all of the questions above are important considerations. We have added a paragraph early in the main text to address limitations. Further, we have updated the substantially revised main text to make the answers to these questions more clear to the reader.

R3.4 I urge the authors to think more and more carefully about the overall setup of the prediction tasks. Namely, what should you compare against? Conceptually, I see two advances in this paper: data & model. Right now, the paper compares itself with other ML models trained on their data, which is good. But in doing so, it also narrows its contribution to just the model, missing the opportunity to highlight the importance of their data and its potential to reveal fundamental insights into human behavior. At some level, one may think the predictive power lies in the data rather than the method, where life2vec is just a powerful way of aggregating the remarkably diverse range of information that you have about an individual's life trajectory. Following this thought,

then one might ask how well your method will perform based on models trained on just the health data. This also gets to the question of what's the baseline here. The current paper suggests that the baseline is other ML models trained on their data. But if one agrees that one major advance of this paper is its data, which integrates numerous dimensions of life – age, income, occupation, health, etc – then one could compare the predictions with baselines that just look at the health information to predict mortality. In other words, maybe the predictive power comes not from one single factor but a combination of many facets in life. Maybe you've tried to do this already with "life table", but that variable seems a bit too crude. This would further highlight the value of your data and how this paper represents a fundamentally new approach to understanding human behavior and life outcomes.

This is a highly insightful comment - and accordingly, our answer has a number of parts.

R3.4a Firstly, we address the aspect of "maybe the predictive power comes not from one single factor but a combination of many facets in life. Maybe you've tried to do this already with "life table", but that variable seems a bit too crude".

As we see it, our logistic regression baseline behaves something along the line of life-tables but with many more variables. The regression uses around 2000 variables, much more than the 2 variables we put in the life tables. As shown in Table A2 (included) below, the performance of logistic regression is substantially better than for the life-tables. We see from the model-performance listed in this table, however, that the neural networks can extract interesting and non-linear effects from the data that are not available to the regression models.

life2vec, which is a "foundation"-type model and able to make general predictions around - not just the early mortality but any topic in the data set - fares better than the specialized neural networks (feedforward and RNN, which can only perform this specific task). The performance gap between specialized neural nets and *life2vec* is not something that is trivially expected (in fact, the naive expectation is that the models should perform similarly). This implies that the more general model forms a superior representation of the complex and high-dimension data.

In the revised MS, we have added a discussion of this point.

Table A2: Corrected Matthew's Correlation Coefficient (MCC) and Area under the Lift (AUL) on the Mortality Prediction Task. 95%-Confidence intervals for the MCC based on the stratified bootstrapping. In both cases, the higher value is preferred. Model Size specifies the number of trainable parameters, we performed a hyperparameter tuning on RNN, FFNN, and Logistic Regression.

Model	MCC, 95%-CI	AUL	Accuracy, 95%-CI	F1-Score, 95%-CI	Model Size
L2V	**0.413 [0.410, 0.422]**	**0.845**	0.788 [0.782, 0.794]	0.443 [0.435, 0.451]	8.4m
RNN-GRU	0.369 [0.361, 0.378]	0.834	0.778 [0.771, 0.783]	0.395 [0.389, 0.402]	1.5m
FFNN	0.340 [0.332, 0.348]	0.822	0.768 [0.762, 0.774]	0.345 [0.339, 0.350]	8.4m
Logistic Reg	0.149 [0.142, 0.155]	0.735	0.639 [0.633, 0.645]	0.201 [0.198, 0.204]	2.0k
Life Tables	0.059 [0.051, 0.066]	0.650	0.555 [0.548, 0.562]	0.161 [0.158, 0.164]	3
Random	-.005 [-0.011, 0.002]	0.497	0.496 [0.489, 0.503]	0.132 [0.128, 0.135]	-
Majority Class	0.0	0.497	0.5	-	-

R3.4b Next, we find the reviewer's considerations around decoupling the different datatypes to understand if something special happens in the combination of many data sources ("Following this thought, then one might ask how well your method will perform based on models trained on just the health data".)

To answer this part of the question, we now evaluate the performance of *life2vec* on a number of data variations to determine the contribution of the variables (see Table A3). Specifically, we consider

- **Full labor** (all of the labor data),
- **Partial labor** (a subset of labor data containing only the following variables: Sex, Country of Origin, Birthday, Income Level, Labor-Force Interval, Municipality of Residency, and a Tax Bracket). We chose these variables as they remove information that is related to the employer (Industry, Sector, Occupation, and Labor Force Status),
- **Partial labor** and **health** (including all the health data),
- **Full labor** and **health.**

Note that we cannot run the analysis "only health" as many individuals only have a few health events. In the analysis above, we keep the cohort constant to understand the effect of changing the underlying data.

In the revised SI B, we show all of this in Table A3 (also reproduced below).

Table A3: Corrected Matthew's Correlation Coefficient (MCC) and Area under the Lift (AUL) on the Mortality Prediction Task. 95%-Confidence intervals for the MCC based on the stratified bootstrapping. Evaluation of the `life2vec` model on various data. Models are pretrained from scratch using these datasets and then finetuned accordingly. **Full Labour & Health** corresponds to the `life2vec` model described in the main manuscript. **Only Full labor** corresponds to the dataset without any health-related events. **Partial Labor & Health** contains health-related events and partial-labor events (without the specification of Work Industry, Sector, Position and Labour Force). **Partial Labor** consists only of partial-labor events (without the specification of Work Industry, Sector, Position and Labour Force).

Data	MCC, 95%-CI	AUL	Accuracy, 95%-CI	F1-Score, 95%-CI	Vocab Size
Full Labor & Health	**0.413 [0.410, 0.422]**	**0.845**	**0.788 [0.782, 0.794]**	**0.443 [0.435, 0.450]**	**2043**
Partial Labor & Health	0.375 [0.367, 0.384]	0.837	0.784 [0.777, 0.790]	0.399 [0.393, 0.406]	1034
Only Full Labor	0.319 [0.312, 0.327]	0.809	0.764 [0.758, 0.770]	0.340 [0.335, 0.345]	1290
Only Partial Labor	0.278 [0.271, 0.285]	0.782	0.737 [0.731, 0.743]	0.305 [0.300, 0.310]	281

Here, we see that the performance really does depend on having all of the data – that the predictive power comes – not from one single factor – but a combination of all of the facets we include. For example, it is interesting to see that (beyond health) the full labor data makes a large difference, both with and without the health data.

We are thankful to the reviewer for these questions. These consideration around the role of data strike right at the heart of what we are trying to understand with the paper. The revised main text now features a subsection focusing on this topic.

R3.5 Figure 2D shows large heterogeneity across gender and age groups. You wrote that "the model performs better on a younger cohort of the population and on a cohort of females." Can you say more about this? For example, from the figure it seems that the differences between female and male are not always significant. Also the variances seem to be larger for young cohorts. Could you explain more about why this is the case?

Good point! The very short answer is that young people generally do not die, and when they do, it's clear from the health records. That will result in higher model performance.

To unpack this - and explain the change in variance, we created Fig. A.12 (included below). Here, we can see that the younger cohort of people indeed has fewer deaths (A.12.B) and that there is a large uncertainty around this cohort. This is clear from the fact that there is a larger fraction of "unlabeled samples" (individuals for whom we do not know their mortality-statue, e.g., because they are not currently residing in Denmark), see Fig A.12.C. The fraction of unlabelled samples will also contribute to a higher uncertainty.

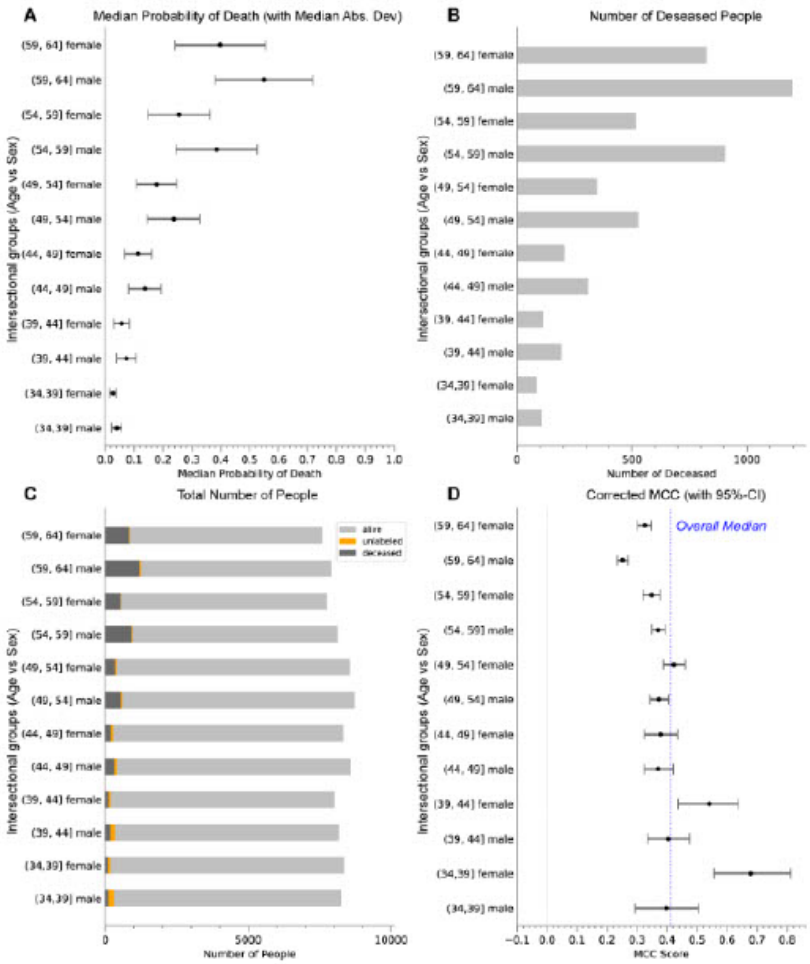


Figure A.12: Detailed evaluation of the life2vec model based on the intersection of **sex** and the **age groups**.

R3.6 Also for Figure 3, it seems that for Q1 and Q4, the life2vec model is not significantly better than RNN. Do you have any thoughts on why that is? Of course here it also relates to my preceding comment. If you can compare against models trained on "competing data", it would also help you bypass this issue.

This is an excellent comment and relates to an important distinction that we did not address clearly in the previous version of the main text.

Our model is what you might term a *Foundation*-type model that trains a highly general embedding (the concept space), which is similar to the word embeddings of Large Language models. Just as word embeddings have many distinct use-cases, our model can use this space to create a wide range of predictions pertaining to more-or-less anything related to the concepts in that space (in the MS, we look at mortality, personality, and emigration … but we could have chosen many other prediction tasks). Thus, the generality of *life2vec* is what makes it interesting to interpret its embedding spaces.

In contrast, the RNN is highly specialized and optimized for the personality prediction task, and its embedding is only meaningful for that specific problem.

We do not necessarily expect that *life2vec* is better on some tasks than an RNN trained for that task on the same data. In fact, it is reasonable to expect the performance to be roughly the same. The fact that *life2vec* does a bit better is likely because the self-attention mechanism allows tokens to interact across the entire sequence, capturing nuanced long-term effects.

We have updated the revised main text to make this aspect more clear.

R3.7 I also feel section 2.3 should be more tightly integrated with sections 2.4&2.5. Right now they feel separated. After reading 2.3, I find myself dying to learn about what explains these superior predictive powers of life2vec, and the embedding space analyses can be quite informative in this regard.

Good point. Inspired by this comment (and by comments from other reviewers), we have significantly restructured the paper.

R3.8 The methods section feels too scattered. There are various subsections and short paragraphs. But for many of them, it's hard to find an internal logic that leads from one to another. Instead, I find myself jumping from one section to another, quite disoriented. Overall, I feel the methods section can be much better organized. And again, that would be important given the potential contribution of the paper – if the paper is as successful as I expect it to be, many of your readers will read through the methods section in detail.

We have restructured the methods to be more coherent.

R3.9 Can you elaborate more in the discussion section on how life2vec will further computational social science as a field?

Yes. As part of our substantial rewrite, and we have also strengthened the discussion (including a further discussion of *life2vec* will further computational social science as a field).

R3.10 Hope these comments are helpful. I want to congratulate the authors for breaking new ground with this paper.

YES! These comments have been incredibly valuable and helped us improve the paper in many dimensions. Thank you for taking the time to read the paper carefully and the insightful & thoughtful feedback.

Decision Letter, first revision:

Date: 3rd November 23 11:22:35
Last Sent: 3rd November 23 11:22:35
Triggered By: Fernando Chirigati
From: fernando.chirigati@us.nature.com
To: sljo@dtu.dk
CC: computationalscience@nature.com
BCC: fernando.chirigati@us.nature.com
Subject: AIP Decision on Manuscript NATCOMPUTSCI-23-0677A
Message: Our ref: NATCOMPUTSCI-23-0677A

3rd November 2023

Dear Dr. Lehmann,

Thank you for submitting your revised manuscript "Using Sequences of Life-events to Predict Human Lives" (NATCOMPUTSCI-23-0677A). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please note that, once we send our requirements, we will likely request for a quick turnaround (approximately 5 days) for the revision to be submitted to us, as we would like to publish your paper before the end of the year. Please let us know if you have any questions or if you think this won't be possible.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please remember to choose, using the manuscript system, whether or not you want to participate in transparent peer review.**
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed.

Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science. Please do not hesitate to contact me if you have any questions.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

I thank the authors for their hard work on revising the manuscript. I feel that all my remarks were successfully addressed.

Reviewer #2 (Remarks to the Author):

I thank the authors for taking care of all of the issues and questions that I raised in the previous round of reviews. The paper is now ready for publication.

Reviewer #2 (Remarks on code availability):

The code looks professionaly developed and is well documented. It should be possible to reproduce the experiments from the paper.

Reviewer #3 (Remarks to the Author):

Fantastic job! Well done! Congratulations!

Final Decision Letter:

nature research

Date: 15th November 23 11:08:39
Last Sent: 15th November 23 11:08:39
Triggered By: Fernando Chirigati
From: fernando.chirigati@us.nature.com
To: sljo@dtu.dk
CC: computationalscience@nature.com
BCC: fernando.chirigati@us.nature.com
Subject: Decision on Nature Computational Science submission NATCOMPUTSCI-23-0677B
Message: 15th November 2023

Dear Dr. Lehmann,

I am delighted to tell you that your manuscript NATCOMPUTSCI-23-0677B has been accepted for publication in Nature Computational Science.

As discussed, due to the exceptional nature of your work, we will publish your paper on an accelerated schedule. **Please carefully review the details below and contact us immediately at computationalscience@nature.com if you have any travel plans or other conflicts that may make you unable to respond to us for the next 5-7 days.**

In approximately 2 business days you will receive a link to choose the appropriate publishing options for your paper and complete the appropriate grant of rights necessary to publish your work. As it is vital that this process not be delayed, we strongly encourage you to whitelist the email address do-not-reply@springernature.com to ensure that this message is received.

You will receive a link to your electronic proof via email with a request to make any necessary corrections as soon as possible. You will find that we have made minor changes to enhance the clarity of the text and to ensure that your paper conforms to the journal's style so we ask that you review these proofs carefully to ensure that we have not inadvertently introduced errors or altered the sense of your text in any way.

Please return your proof within 24 hours of receiving it. If you have any questions about your proofs or anticipate any delays please contact rjsproduction@springernature.com immediately.

Once a publication date is set for your paper, the Springer Nature press office will be in touch with the full embargo details. We request that you do not send out your own publicity or contact any journalists until you hear from us that the paper has a confirmed publication date.

If you would like to inform your Public Relations or Press Office about your paper, we suggest that you do so immediately to allow them as much time as possible to prepare an appropriate press release and organize publicity if they choose to do so. Please include your manuscript tracking number NATCOMPUTSCI-23-0677B and the name of the journal, which they will need if they contact our press office.

Please note that Nature Computational Science is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to the Protocol Exchange. Protocol Exchange is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available, assigned DOIs for ease of citation and fully searchable through nature.com. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated page to collect all your lab Protocols. By uploading your Protocols to Protocol Exchange, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at www.nature.com/protocolexchange/. Further information can be found at www.nature.com/protocolexchange/about .

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

Best,
Fernando

--
Fernando Chirigati, PhD

Chief Editor, Nature Computational Science
Nature Portfolio

P.S. Click here if you would like to recommend Nature Computational Science to your librarian - this will link directly to the Recommend page.

http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**