



Check for updates

## RESEARCH NOTE




**REVISED** The rise and fall of machine learning methods in biomedical research [version 2; referees: 2 approved]Hashem Koohy  1,2<sup>1</sup>MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK<sup>2</sup>Honorary Research Fellow in Computational Biology, Zeeman Institute, University of Warwick, Coventry, UK**v2** First published: 14 Nov 2017, 6:2012 (doi: [10.12688/f1000research.13016.1](https://doi.org/10.12688/f1000research.13016.1))  
Latest published: 02 Jan 2018, 6:2012 (doi: [10.12688/f1000research.13016.2](https://doi.org/10.12688/f1000research.13016.2))**Abstract**



In the era of explosion in biological data, machine learning techniques are becoming more popular in life sciences, including biology and medicine. This research note examines the rise and fall of the most commonly used machine learning techniques in life sciences over the past three decades.



This article is included in the **Machine learning: life sciences** collection.

**Open Peer Review****Referee Status:**  

Invited Referees	
1	2
<hr/>	
<b>REVISED</b>	
<b>version 2</b>	report
published 02 Jan 2018	
<b>version 1</b>	
published 14 Nov 2017	report

- 1 **Alex Bateman** , European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK
- 2 **Konrad Förstner** , University of Würzburg, Germany

**Discuss this article**

Comments (0)

**Corresponding author:** Hashem Koohy ([hashem.koohy@rdm.ox.ac.uk](mailto:hashem.koohy@rdm.ox.ac.uk))

**Author roles: Koohy H:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Koohy H. **The rise and fall of machine learning methods in biomedical research [version 2; referees: 2 approved]** *F1000Research* 2018, **6**:2012 (doi: [10.12688/f1000research.13016.2](https://doi.org/10.12688/f1000research.13016.2))

**Copyright:** © 2018 Koohy H. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

**Grant information:** This work was supported by the Human Immunology Unit MRC Core grant (MC\_UU\_12010).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 14 Nov 2017, **6**:2012 (doi: [10.12688/f1000research.13016.1](https://doi.org/10.12688/f1000research.13016.1))

**REVISED Amendments from Version 1**

In this new version, I have tried to address all the comments and suggestions raised by both of my referees. I have therefore changed the manuscript's contents, structure and figures accordingly. Therefore, the new version has now three main figures whereas the previous version had only one figure. I have additionally made some changes to the R code and updated the online git project.

To address R1's key comments, earlier in the manuscript, I have acknowledged that this work has been motivated by a previous similar publication. I have made the distinction between ANNs and DNNs clearer throughout the manuscript. All minor points by R1 has been also addressed.

To address R2's key concerns:

I have added proper referencing to the R code to prevent from the error.

Both 'naïve bayes classifier' and 'logistic regression' models have been re-added to the manuscript and to the figures.

The issue of counting hits for DNN has been resolved and manuscript has been updated accordingly.

Figures have been re-structured and now I have three figures instead of only one. The PR has been illustrated in one separate figure.

I have also tried to address all the possible minor points.

I have added a full detailed response to each of the referees in online 'Respond or Comment' section of the manuscript.

**See referee reports**

## Introduction

Over the past three decades, biological data have grown dramatically in both size and complexity. The major contributors to the growth in size of computation biology data include, but not are not limited to, the ability of biologists to sequence complex genomes such as the human genome (1990–2003) (Lander *et al.*, 2001), the advent of new high throughput sequencing techniques (around 2008) (Marx, 2013), and most recently the very rapid advancements in single cell technologies, introduced in 2009 (Wang & Navin, 2015).

The complexity of biological data has been growing even faster, and doesn't seem to be linearly dependent on the size of data. Examples of complexity in the field of computational genomics include multiple diverse sources of technical noise, low signal to noise ratio, low numbers of biological replicates in comparative approaches, rare and usually hardly detectable mutations in non-coding regions and rare and barely identifiable cell types in complex heterogeneous systems such as the immune system and/or the brain.

At the intersection of mathematics, statistics and computer science is machine learning (ML), the *de facto* tool box in data

science for deciphering the relationship between the input and output as well as detecting significant patterns within large, complex data sets. These quantitative approaches have been shown to be effective and are becoming increasingly popular in addressing challenges such as those outlined above. Highlights of their successful applications in functional genomics include, but are not limited to, learning and characterizing chromatin states by employing unsupervised approaches such as chromHMM (Ernst & Kellis, 2012), predicting sequence specificities of DNA- and RNA-binding proteins using convolutional neural networks such as DeepBind (Alipanahi *et al.*, 2015), and employing a combination of supervised and unsupervised approach to determine the genetic and epigenetic contributors of antibody repertoire diversity (Bolland *et al.*, 2016). Nowadays it is almost impossible to publish a study on single cell assays without using dimensionality reduction methods such as Principal Component Analysis or t-SNE.

One indirect measure of the success of these techniques in extracting scientific insights from biological data is to measure the popularity and usage of machine learning algorithms in life sciences research over time (Jensen & Bateman, 2011). Motivated by Jensen *et al.*, I therefore set out to update machine learning usage in life sciences. For this I quantified what fraction of published papers in the PubMed database mention a particular technique and how these number of citations are changed each year (see methods).

## Methods

For this analysis, I used the R RISmed package (Kovalchik, 2015) to parse the publication data from NCBI. I examined publications in PubMed from 1990 to 2017 using a metric that measures the proportion of publications per year that mention the technique in the full text (Hits Per Year per Million articles published, or HPYM). The Popularity Rate (PR) of a technique was then defined as the difference between HPYMs in any two consecutive years. A positive PR shows an increase in popularity, whereas a negative PR reflects a decrease in popularity. I limited this note to 12 models listed in Table 1 which have been the most common or which showed a sharp change in popularity rate at a particular time. However, the R code is available with which any particular model during a specific period of time can be easily measured.

## Results

This analysis demonstrates that the overall popularity of machine learning methods in biomedical research has linearly increased since 1990 to 2017, but with two different slopes. From 1990 to 2000 the slope is 0.02, meaning that popularity increased only 2% per year. In 2001 (when sequencing big genomes became possible) the slope increased to 0.06, and since then it has remained constant. A maximum of 1.2% of all papers published in PubMed in any calendar year have mentioned one of the machine learning methods investigated in this study

**Table 1. Common Machine Learning Techniques in Life**

**Sciences.** This table shows 12 machine learning techniques whose popularity in life sciences have been investigated in this study. Technical note: Supervised means that the model requires training data to learn its parameters. A supervised model is used to predict the future instances. An unsupervised model doesn't require any training data and is used to detect patterns within a dataset. Dimensionality reduction models are used to project high-dimensional datasets into lower dimension space where new variables are more interpretable.

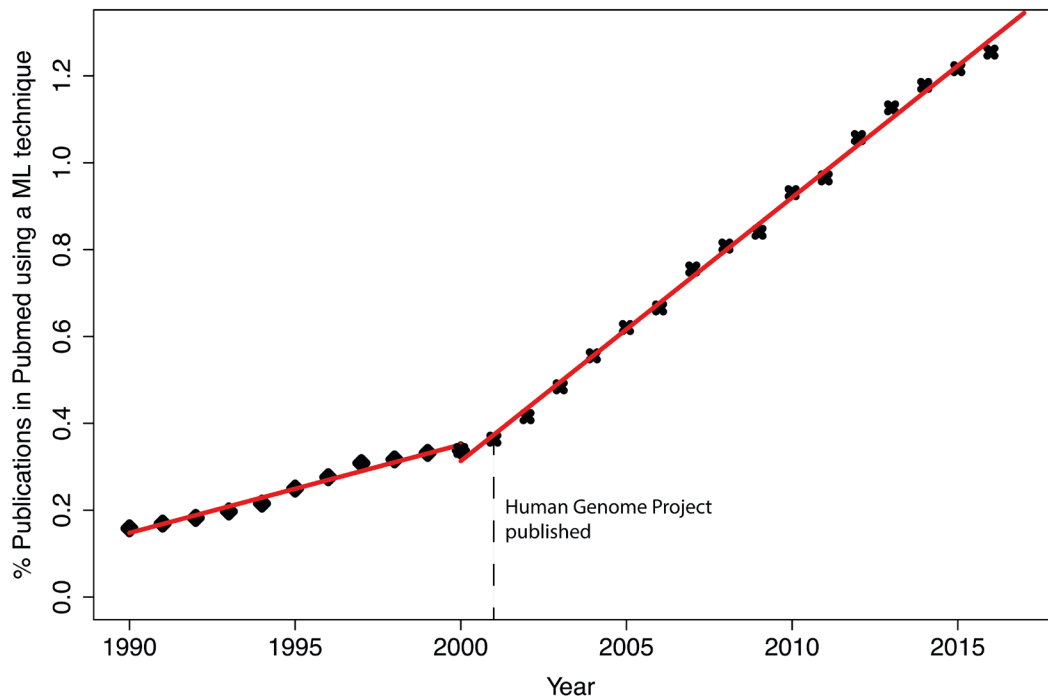
Technique	Abbreviation	Category
Random Forest	RF	Supervised
Support Vector Machine	SVM	Supervised
Artificial Neural Network	ANN	Supervised
Deep Neural Network	DNN	Supervised & Unsupervised
Principal Component Analysis	PCA	Dimensionality Reduction
Linear Regression	LR	Supervised
Markov Model	MM	Unsupervised
Decision Tree	DT	Supervised
Hierarchical Clustering	HC	Unsupervised
t-Distributed Stochastic Neighbour Embedding	t-SNE	Dimensionality Reduction
Logistic Regression Model	LogReg	Supervised
Naïve Bayes Classifier	NBC	Supervised

(Figure 1). I was expecting to see a higher usage of ML in life sciences, but without a gold standard set to compare with, I would not be able to judge if this is too high or low or just about right.

The Linear Regression (LR) models have been the most dominant machine learning techniques in the life sciences over the past three decades (Figure 2A). It is interesting to see that LR models are still highly in used despite recent appearance of sophisticated ML techniques such as ensemble-based approaches and/or Support Vector Machines and even with very recent and state of the art deep learning techniques. Although, its popularity rate has been plateaued over the past few years (Figure 3) meaning that its usage is increased linearly with a constant slope. With a constant increase of 300 HPYM, and considering its higher intercept at 1990, the linear regression models is predicted to be one of the most popular techniques over the next few years.

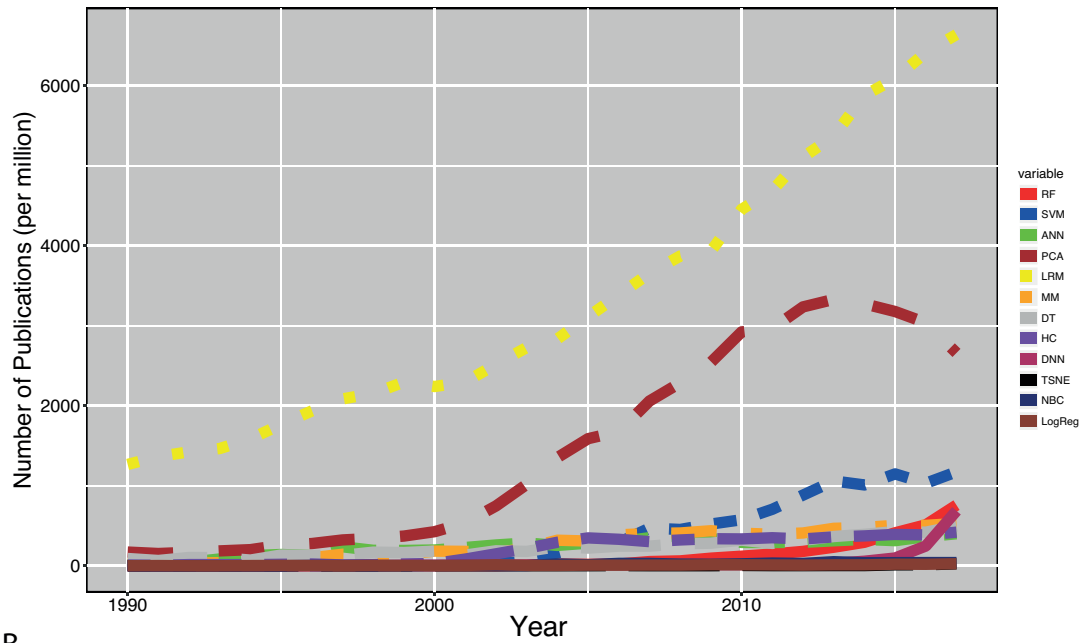
Perhaps a very surprising observation of this study is the rise and fall of Principle Component Analysis (PCA). PCA became very fashionable during 2000 to 2013. In fact, 3329 per million papers published in 2013 mentioned PCA which was the highest number of PCA usage. Since then it has been used less, although it still is the second most popular tool (Figure 2A).

In early 2000s, unsupervised Hierarchical Clustering alongside newly introduced supervised techniques Support Vector Machines (SVMs) and Random Forests (RFs), showed a sharp rise in

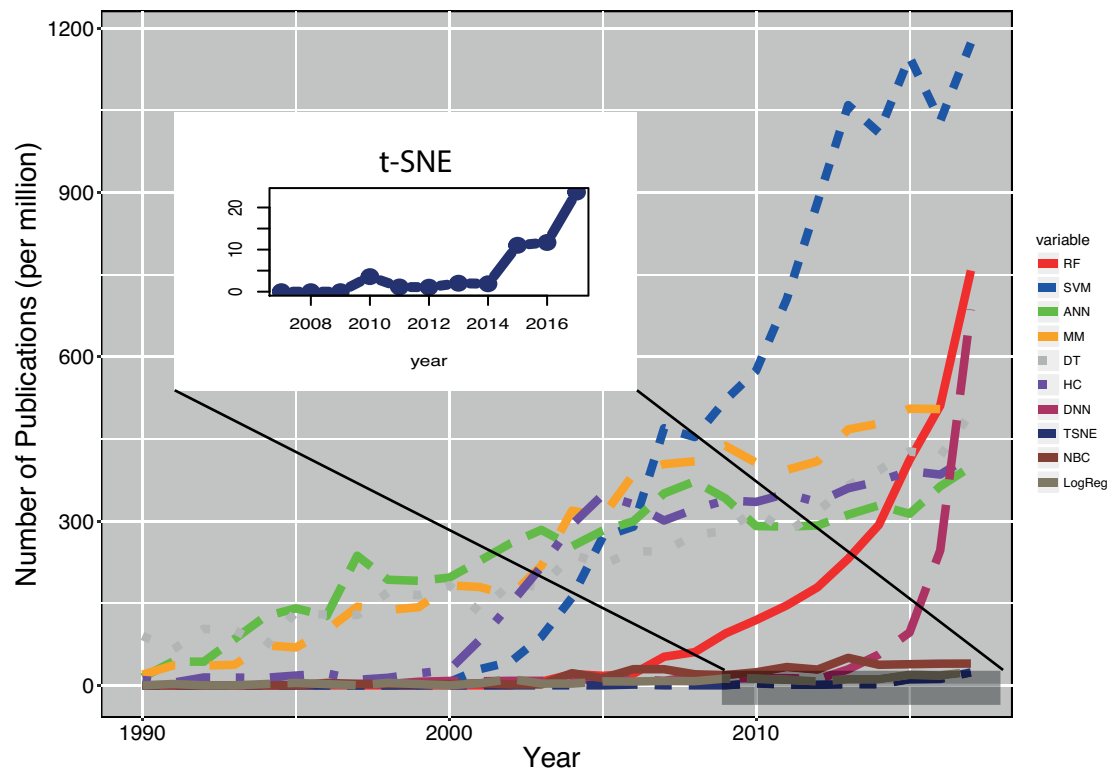


**Figure 1. Cumulative usage of all 12 machine-learning techniques used in this manuscript.** Two different linear regression models have been fitted to this data. The first one covers years from 1990 to 2000. The second one that shows a triple increase in its slope covers from 2001 till 2017. Y-axis shows the number of hits per 100 publications.

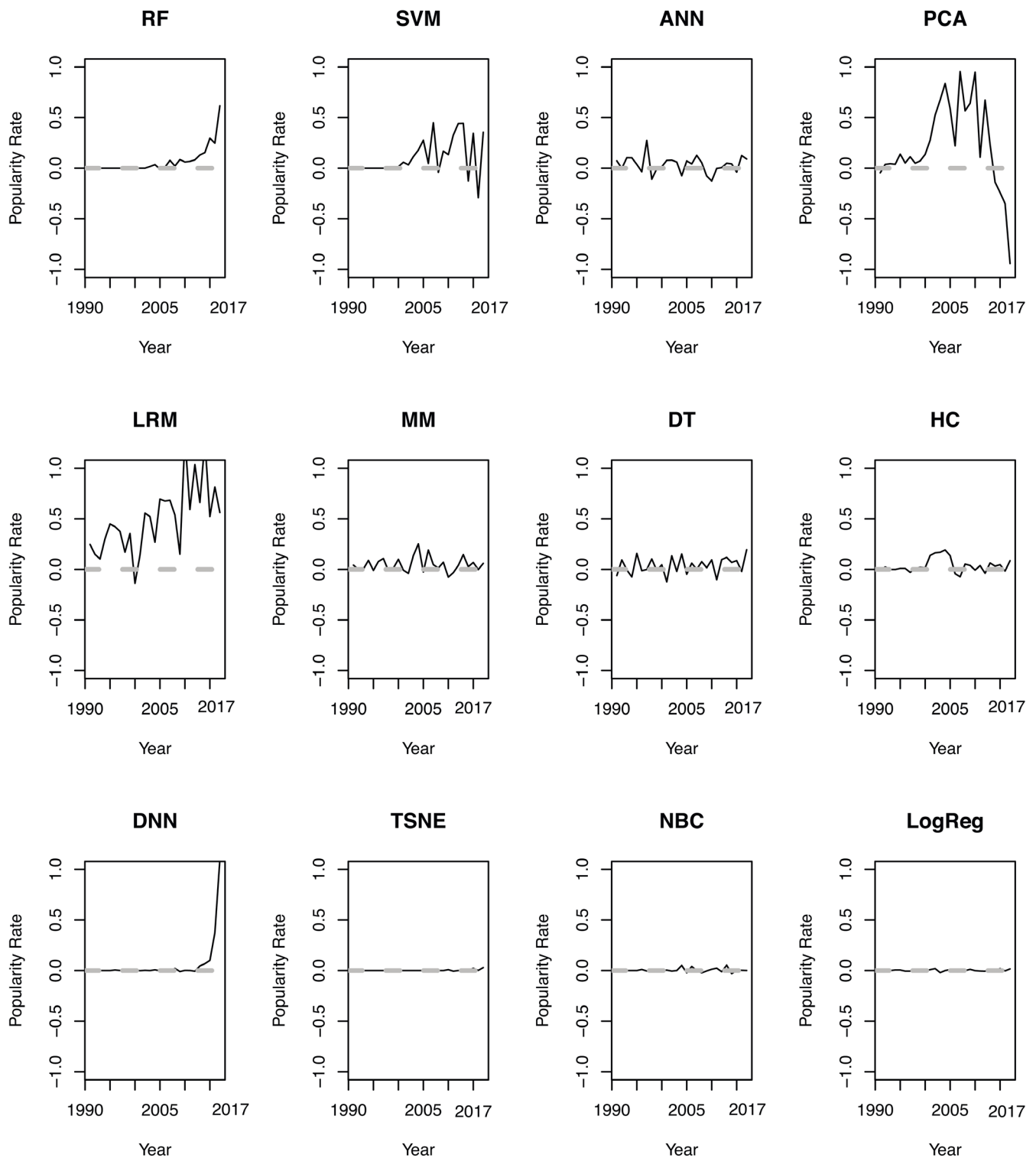
A



B



**Figure 2.** **A:** Trends of individual machine-learning techniques defined as per million hits in y-axis. **B:** Similar to **A** but without the two very highly used techniques Linear Regression and Principal Components Analysis in order to enhance clarity in usage of other not-very-commonly used techniques that were overshadowed by LR and PCAs.



**Figure 3.** An illustration of popularity rate of all 12 techniques used in this manuscript. The PR has been defined as differences of HPYMs in each two-consecutive year for each model. This number have been further re-scaled to vary only between -1 and 1.

usage, which was mainly associated to microarray data analysis. Usage of hierarchical clustering plateaued shortly after its sharp popularity rise in 2000. SVMs kept their popularity longer, for almost a decade in fact, but subsequently dropped to an almost negligible popularity rate (Figure 3). RFs on the other hand, showed less popularity at the beginning of their arrival, but later on (after 2013) they were ranked the second highest in popularity after Deep Neural Networks (DNN) (Figure 2A, 2B and Figure 3).

During the period of 1990–2017, Artificial Neural Networks (ANNs) have demonstrated considerable fluctuations in popularity (Figure 2B and Figure 3). ANNs in the early 1990's after Linear Regression and PCA, were the most commonly used techniques until early 2000, when they lost their popularity to MM, HCs and SVMs and even later to RFs. However, since 2013, a sub-family of ANN known as Deep Neural Networks (DNNs) made their way into the life sciences, and their usage since then has increased remarkably, so that DNNs currently have the highest popularity rate (Figure 3).

The dimensionality reduction technique t-distributed Stochastic Neighbour Embedding (t-SNE) published in 2008, has become quickly tailored to all sorts of single cell techniques. It is therefore not surprising to see that t-SNE usage has also been very rapidly growing over the past few years (Figure 2B).

**Dataset 1. The text file contains the raw data underlying the results presented in this study, i.e. the number of publications in PubMed mentioning each machine learning technique from 1990–2017. These data is further normalized per million for downstream analysis**

<http://dx.doi.org/10.5256/f1000research.13016.d184022>

## Discussion

I have illustrated the rise and fall of ML techniques in life sciences from 1990 to the present day. I chose this period because I believe this is the transition period for life scientists to join the big-data club. With the same R code used in this study to parse the publication data from NCBI, it would be possible to look at any period of time.

It was not very surprising to see LR models as the most commonly used model in the field, since:

a) LR models are one of the oldest ML methods that have been in use in almost any field,

b) Parameters in LR models can be learned by using a training data with just a few data samples.

c) A lot of other models can be placed under this umbrella, for instance by first applying a transformation function.

It was, however, surprising to see the sharp rise and fall of PCA. Perhaps a contributing factor to PCA being the most dominant dimensionality reduction method available in this period was its easy-to-use implementation in R. The question still remains as to why its popularity decreased from 2008 onwards. Perhaps the arrival of more versatile models such as RFs and SVMs which are very capable of handling high dimensionality and dealing with co-linearity in biological data eased the need to use PCA. Additionally, t-SNE as a tremendously growing dimensional reduction model in the field, is establishing itself as a strong competitor for the PCA.

ANNs have been fairly popular since the 1990s until around 2004. Around that time more readily useable and less complex techniques became available, such as SVMs, RFs and MM. However, with the huge investments of giant information companies such as Google leading to very impressive applications of DNNs and other sub-families of ANNs, in various disciplines, DNNs, currently has the sharpest popularity rate (Figure 3).

I appreciate that there are limitations to this study. For instance, for the majority of comparative analyses of gene expression, researchers use a differential expression software and/or package, but cite only the package name and not the underlying statistical or ML technique used in the package. These cases have not been covered in this study. However, this study can be considered as an approximation of the extent to which machine learning techniques are used in life sciences.

This note can be considered as an update of a similar study by Jensen *et al.* (Jensen & Bateman, 2011), in which the authors investigated the rise and fall of a number supervised machine learning techniques in life sciences. Here, I have gone beyond the abstracts and searched the full text of each paper, for the usage of both supervised and unsupervised ML technique.

## Data and software availability

**Dataset 1:** The text file contains the raw data underlying the results presented in this study, i.e. the number of publications in PubMed mentioning each machine learning technique from 1990–2017. These data is further normalized per million for downstream analysis. DOI, [10.5256/f1000research.13016.d184022](https://doi.org/10.5256/f1000research.13016.d184022) (Koohy, 2017).

R code used to parse the publication data from NCBI is available at: [https://github.com/hkoohy/Machine\\_Learning\\_in\\_Life\\_Sciences](https://github.com/hkoohy/Machine_Learning_in_Life_Sciences)

Archived source code as at the time of publication: <http://doi.org/10.5281/zenodo.1039642> (hkoohy, 2017).

License: GNU GENERAL PUBLIC LICENSE

### Competing interests

No competing interests were disclosed.

### Grant information

This work was supported by the Human Immunology Unit MRC Core grant (MC\_UU\_12010).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

### Acknowledgments

I am very grateful to David Sims, Edward Morrissey and Supat Thongjuea for critical reading of the manuscript and for their invaluable comments. I acknowledge both referees of this manuscript for their very fair and un-biased comments that have immensely improved the clarity and quality of this note.

## References

Alipanahi B, Delong A, Weirauch MT, *et al.*: **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.** *Nat Biotechnol.* 2015; **33**(8): 831–838.

[PubMed Abstract](#) | [Publisher Full Text](#)

Bolland DJ, Koohy H, Wood AL, *et al.*: **Two mutually exclusive local chromatin states drive efficient V(D)J recombination.** *Cell Rep.* 2016; **15**(11): 2475–2487.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nat Methods.* 2012; **9**(3): 215–216.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

hkoohy: **hkoohy/Machine\_Learning\_in\_Life\_Sciences: First release.** (Version V1.0.0). *Zenodo.* 2017.

[Data Source](#)

Jensen LJ, Bateman A: **The rise and fall of supervised machine learning**

**techniques.** *Bioinformatics.* 2011; **27**(24): 3331–3332.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Koohy H: **Dataset 1 in: The rise and fall of machine learning methods in biomedical research.** *F1000Research.* 2017.

[Data Source](#)

Kovalchik S: **RISmed: download content from NCBI databases.** *R package version.* 2015.

[Reference Source](#)

Lander ES, Linton LM, Birren B, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature.* 2001; **409**(6822): 860–921.

[PubMed Abstract](#) | [Publisher Full Text](#)

Marx V: **Biology: The big challenges of big data.** *Nature.* 2013; **498**(7453): 255–260.

[PubMed Abstract](#) | [Publisher Full Text](#)

Wang Y, Navin NE: **Advances and applications of single-cell sequencing technologies.** *Mol Cell.* 2015; **58**(4): 598–609.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)



# Open Peer Review

Current Referee Status:



## Version 2

Referee Report 08 January 2018

doi:10.5256/f1000research.14767.r29390



**Konrad Förstner** 

Core Unit Systems Medicine, Institute for Molecular Infection Biology, University of Würzburg, Würzburg, Germany

I thank Dr. Koohy for addressing my raised issues adequately.

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

## Version 1

Referee Report 06 December 2017

doi:10.5256/f1000research.14114.r28432



**Konrad Förstner** 

Core Unit Systems Medicine, Institute for Molecular Infection Biology, University of Würzburg, Würzburg, Germany

In the manuscript "The rise and fall of machine learning methods in biomedical research" the author has generated a quantitative perspective on the usage of machine learning methods in the life sciences. For some of the methods a hypothesis about the underlying reason for an increased or decrease popularity are discussed. The code for performing the analysis is available on GitHub and - like the retrieved PubMed data - has been deposited at Zenodo.

I have several major objections / question / suggestion for the author:

- I tried to reproduce the analysis using RStudio 1.1.383 with the deposited RStudio project but got the following error when executing the R chunks in the file *Machine\_Learning\_Trends.Rmd*: "Error in library(informationRetrieval) : there is no package called 'informationRetrieval'" The file *informationRetrieval.R* is located in another subfolder and I guess this just needs proper referencing inside of the project.
- The author states that he has selected widely used machine learning methods used in life sciences. I would have expected Naive Bayes classifiers in the list of most popular methods. A

simple PubMed search for "naive bayes classifier" OR "naive bayesian classifier" return twice as many hits as for "deep neural networks" (but over a longer time span):

- <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22naive+bayes+classifier%22+OR+%22naive+bayesian+classifier%22>
- <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22deep+neural+networks%22>
- Similar issue for logistic regression: The analysis in the provided file *Machine\_Learning\_Trends.Rmd* actually contains the counting of publications containing logistic regression that shows a large (206,619 at the time of writing) and growing number of this but this method has not been discussed in the manuscript and is not displayed in the plots.
  - <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22logistic%20regression%22>
- The counting of hits for deep neural networks (DNN) is not done properly. Looking at the code to count the number of hits of different search terms shows that the author use "artificial neural networks" and "deep neural networks" and "deep learning" as search term for DNN (see code selection at the bottom of this section). I think using the search term "artificial neural network" for both ANN and DNN is not sound and changes the story of DNN (a special form ANN) significantly. Either DNN is treated as subset of ANN and only ANN are plotted or DNN and ANN are treated separately and the search term "artificial neural network" is not used for DNN. Furthermore the search term "deep learning" results in numerous unrelated hits before 2010 (e.g. PMID: 8936230, 9165817, 9487168, 10463930).
 <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22deep+learning%22> (then click on the "Result by year" histogram).
- The authors tries to explain the underlying reasons for the gain or loss of certain ML methods. In Figure 1 one of the publications of the human genome is placed in the year 2000 without any citation. The human draft genome was published in 2001 (International Human Genome Sequencing Consortium, Nature 409, 860–921, 2001, <https://doi.org/10.1038/35057062>) and it would be interesting to see what the author is referring to.
- The Popularity Rate (PR) introduced here is not plotted anywhere directly but is the slope of edges between the data points of two consecutive years. The author should consider visualizing this measurement of change as well.
- The curve plotted in Fig 1 A is nearly reassembled by the LRM curve in Fig 1 B. Is the observation in Fig 1 A maybe only an observation of the dominating LRM method? I do not understand why Fig 1 A can actually look nearly exactly like the LRM curve considering the other methods e.g. the PCA curve.

Code selection regarding ANN and DNN

```
'''
```

```
ANN_hits <- get_normliazed_number_of_hits(years = YEARS, query="artificial neural network[tw]",
db="pubmed", normalization_value=1000000)
```

```
NN_term <- "(artificial neural networks[tw] OR deep neural networks[tw] or deep learning[tw])"
```

```
DNN_hits <- get_normliazed_number_of_hits(years=YEARS, query=NN_term, db="pubmed",
normalization_value=1000000)
```

```
'''
```

Minor issues:

- Figure 1

- Style: The different lines are hard to distinguish by color only - maybe consider an additional discriminator (e.g. dashed lines for a subset); Next to Fig 1 C is a lot of white space. Placing the t-SNE subplot to a different location (e.g. the middle of Fig 1 C) would make it possible to use this space more efficiently.
- Maybe think rearranging the whole figure. Figure 1 C is a subplot of figure 1 B like the t-SNE plot is a subplot of Figure 1 C
- "de facto" should be written in italic font
- The link to RISmed should use the link indicated at the page itself that says "Please use the canonical form <https://CRAN.R-project.org/package=RISmed> to link to this page."
- For Linear Regression Model sometimes "LRM" and sometime "LR" is used in the manuscript
- In order to understand which biological approaches / questions that are influencing the usage of different ML method the association of those methods with certain MeSH term would be interesting. Either as part of this manuscript or a future one.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 22 Dec 2017

**Hashem Koohy**, Oxford University

I thank Dr Konrad Forstner for his time in evaluating the manuscript and for his very detailed comments/suggestions that I believe will immensely enhance the quality of the manuscript.

In the following I address the issues raised by Konrad.

In the manuscript "The rise and fall of machine learning methods in biomedical research" the author has generated a quantitative perspective on the usage of machine learning methods in the

life sciences. For some of the methods a hypothesis about the underlying reason for an increased or decrease popularity are discussed. The code for performing the analysis is available on GitHub and - like the retrieved PubMed data - has been deposited at Zenodo.

I have several major objections / question / suggestion for the author:

- I tried to reproduce the analysis using RStudio 1.1.383 with the deposited RStudio project but got the following error when executing the R chunks in the file *Machine\_Learning\_Trends.Rmd*: "Error in library(informationRetrieval) : there is no package called 'informationRetrieval'" The file *informationRetrieval.R* is located in another subfolder and I guess this just needs proper referencing inside of the project.

I have changed the package structure and added a cell on top of the *rdm* file so that it should easily find the \*.r codes and source them.

- The author states that he has selected widely used machine learning methods used in life sciences. I would have expected Naive Bayes classifiers in the list of most popular methods. A simple PubMed search for "'naive bayes classifier" OR "naive bayesian classifier" return twice as many hits as for "deep neural networks" (but over a longer time span):
  - <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22naive+bayes+classifier%22+OR+%22>
  - <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22deep+neural+networks%22>

In my very initial analysis I had naïve bayes classifier, but to my surprise after normalization it was overshadowed by other highly used techniques. I therefore took it out. Now, for the completion, I have added it again.

- Similar issue for logistic regression: The analysis in the provided file *Machine\_Learning\_Trends.Rmd* actually contains the counting of publications containing logistic regression that shows a large (206,619 at the time of writing) and growing number of this but this method has not been discussed in the manuscript and is not displayed in the plots.

Similarly, I had logistic regression models in my initial analysis. And for the same reason I took it out from the final submission and left to the reader to test if they wish. It has been added again.

- <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22logistic%20regression%22>
- The counting of hits for deep neural networks (DNN) is not done properly. Looking at the code to count the number of hits of different search terms shows that the author use "artificial neural networks" and "deep neural networks" and "deep learning" as search term for DNN (see code selection at the bottom of this section). I think using the search term "artificial neural network" for both ANN and DNN is not sound and changes the story of DNN (a special form ANN) significantly. Either DNN is treated as subset of ANN and only ANN are plotted or DNN and ANN are treated separately and the search term "artificial neural network" is not used for DNN. Furthermore the search term "deep learning" results in numerous unrelated hits before 2010 (e.g. PMID: 8936230, 9165817, 9487168, 10463930). <https://www.ncbi.nlm.nih.gov/pubmed/?term=%22deep+learning%22> (then click on the "Result by year" histogram).

I really apologize for this. I have corrected in the code, and changed the manuscript accordingly.

- The authors tries to explain the underlying reasons for the gain or loss of certain ML methods. In Figure 1 one of the publications of the human genome is placed in the year 2000 without any citation. The human draft genome was published in 2001 (International Human Genome Sequencing Consortium, Nature 409, 860–921, 2001, <https://doi.org/10.1038/35057062>) and it would be interesting to see what the author is referring to.

I apologize again for the confusion. I was in fact referring 2001 IHGSC paper, as I had cited in the manuscript. I have changed the figure to make it clear.

- The Popularity Rate (PR) introduced here is not plotted anywhere directly but is the slope of edges between the data points of two consecutive years. The author should consider visualizing this measurement of change as well.

Very valid point. I have restructured the manuscript and the figures. Now, I have a separate figure for this.

- The curve plotted in Fig 1 A is nearly reassembled by the LRM curve in Fig 1 B. Is the observation in Fig 1 A maybe only an observation of the dominating LRM method? I do not understand why Fig 1 A can actually look nearly exactly like the LRM curve considering the other methods e.g. the PCA curve.
- Great observation. Both figures are very similar, though with different slopes and intercepts. In order to check if the cumulative figure is dominant by LRM, in a separate task, I filtered out LRM and made the cumulative figure. Although in both full-model and filtered-model we can see two different slopes (from 1990 to 2000, from 2001 to 2017), not surprisingly, the full model fits better.

I think what happens is that around the time that PCA starts declining, we have an almost exponential increase from other models such as RF, SVM and later on from DNN. These collectively delute effect of PCA decline.

Code selection regarding ANN and DNN

...

```
ANN_hits <- get_normliazed_number_of_hits(years = YEARS, query="artificial neural
network[tw]", db="pubmed", normalization_value=1000000)
```

```
NN_term <- "(artificial neural networks[tw] OR deep neural networks[tw] or deep learning[tw])"
DNN_hits <- get_normliazed_number_of_hits(years=YEARS, query=NN_term, db="pubmed",
normalization_value=1000000)
```

...

Minor issues:

- Figure 1
  - Style: The different lines are hard to distinguish by color only - maybe consider an additional discriminator (e.g. dashed lines for a subset); Next to Fig 1 C is a lot of white space. Placing the t-SNE subplot to a different location (e.g. the middle of Fig 1 C) would make it possible to use this space more efficiently.
  - Maybe think rearranging the whole figure. Figure 1 C is a subplot of figure 1 B like the t-SNE plot is a subplot of Figure 1 C
  - As suggested, I have restructure the manuscript and the figures. The manuscript now has three main figures which are hopefully more clearer than the previous version.
- "de facto" should be written in in italic font
- Corrected.
- The link to RISmed should use the link indicated at the page itself that says "Please use the canonical form <https://CRAN.R-project.org/package=RISmed> to link to this page."
- For Linear Regression Model sometimes "LRM" and sometime "LR" is used in the manuscript
- I have corrected for this.

- In order to understand which biological approaches / questions that are influencing the usage of different ML method the association of those methods with certain MeSH term would be interesting. Either as part of this manuscript or a future one.
- This is a very interesting point. Though as suggested is beyond this manuscript.

**Competing Interests:** I have no competing interest with Dr Konrad Forstner.

Referee Report 27 November 2017

doi:10.5256/f1000research.14114.r28048



**Alex Bateman** 

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

I should firstly point out that I was co-author on the 2011 editorial published in Bioinformatics titled, "The rise and fall of supervised machine learning techniques"<sup>1</sup>. Therefore I was momentarily surprised to be invited to review a paper with such a similar title. That editorial was only a page and a half long and only really scratched the surface of the interesting topic of the prevalence of use of machine learning in the biosciences. The author cites our 2011 paper and mentions that the current article can be considered an update of it. However, that is only mentioned in the very final paragraph of the paper. It would seem reasonable to me to make that one of the first things mentioned in the paper. Of course I am far from a neutral observer on this point.

Overall I think that the article presents sound and interesting science and should be published within F1000Research. I think it provides a timely update to the 2011 editorial and expands it with some nice extra details. The article increases the number of ML methods investigated from 5 to 10. Most notably linear regression models are included which top the league table.

I noticed an inconsistency in the data presented for ANNs in this new paper compared to the 2011 paper. Why is that? The numbers for ANN are considerably lower in this article. Is that because DNNs are split out from ANNs? Throughout the paper it says that ANNs have become known as DNNs. That is not correct. DNNs are a subtype of ANNs. So all DNNs are ANNs, but not all ANNs are DNNs. That needs correction throughout.

The following statement does not read well:

"The sharp increase usage in popularity rate of DNNs over the past few years (Figure 1C) suggests that DNNs will take the PR lead again in the coming years."

After multiple readings I would presume that PR lead means it has the highest popularity rate. DNNs would have more than 300 more mentions per million papers per year. Firstly that sentence is very confusing to understand for a reader. For the first two readings I thought you were saying that DNNs would take the lead from LRMs, which would seem unlikely. On third reading I thought you meant that the slope of DNNs would overtake LRMs, but clearly it has already done that. I think you should rethink that sentence or take it out.

#### Minor points:

Page 2. At he intersection -> At the intersection

Page 2. You mention that a surprising maximum of 1.2% of all paper mention one of the 10 ML techniques. Why is that surprising? Is it too low, too high? Please explain.

Page 2. NCBI database is mentioned. NCBI has a lot of databases, please specify which one.

Page 3. less used less -> used less

## References

1. Jensen LJ, Bateman A: The rise and fall of supervised machine learning techniques. *Bioinformatics*. 2011; **27** (24): 3331-2 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** Author of editorial upon which this article has built.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 22 Dec 2017

**Hashem Koochy**, Oxford University

I thank Dr. Alex Bateman for his time in evaluating the manuscript as well as for his very valuable comments.

In fact, I was inspired by Alex's commentary. I therefore apologize for not appropriately mentioning this earlier in the manuscript. I have made this change to the manuscript and I hope that is clear enough now.

As Alex has suggested, I have made the distinction between ANNs and DNNs clear in corresponding paragraph and change the manuscript accordingly.

I have also addressed the minor points accordingly.

I hope the current version of manuscript meets Alex's standards and consequently is clearer for the readers now.

**Competing Interests:** I have no competing interest.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

F1000Research