

# <sup>1</sup> daltoolbox: Leveraging Experiment Lines for Modular and Reproducible Data Analytics

<sup>3</sup> **Eduardo Ogasawara**  <sup>1</sup>, **Ana Carolina Sá**  <sup>1</sup>, **Antonio Castro**  <sup>1</sup>, **Caio Santos**  <sup>1</sup>, **Diego Carvalho**  <sup>1</sup>, **Diego Salles**  <sup>1</sup>, **Eduardo Bezerra**  <sup>1</sup>, **Esther Pacitti**  <sup>3</sup>, **Fabio Porto**  <sup>2</sup>, **Janio Lima**  <sup>1</sup>, **Lucas Tavares**  <sup>1</sup>, **Rafaelli Coutinho**  <sup>1</sup>, **Rebecca Salles**  <sup>4</sup>, and **Vinicius Saidy**  <sup>1</sup>

<sup>7</sup> 1 Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ), Brazil 2 National  
<sup>8</sup> Laboratory of Scientific Computing (LNCC), Brazil 3 University of Montpellier, LIRMM, France 4  
<sup>9</sup> National Institute for Research in Digital Science and Technology (INRIA), France 5 Petróleo Brasileiro  
<sup>10</sup> S.A. (Petrobras), Brazil

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 06 November 2025

Published: unpublished

## License

Authors of papers retain copyright and release the work under a

Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/))

## <sup>11</sup> Summary

<sup>12</sup> The **daltoolbox** package provides an open-source framework for constructing modular and  
<sup>13</sup> reproducible data analytics workflows in R. Built upon the concept of *Experiment Lines* (*EL*)  
<sup>14</sup> ([Marinho et al., 2017](#)), **daltoolbox** enables the definition of flexible experiment families through  
<sup>15</sup> the composition of alternative preprocessing, modeling, and evaluation steps. This design  
<sup>16</sup> allows researchers and practitioners to create, compare, and evolve analytical workflows with  
<sup>17</sup> minimal code modification. The package integrates with external R and Python libraries,  
<sup>18</sup> fostering interoperability and transparency in experimental data analysis.

## <sup>19</sup> Background

<sup>20</sup> The rapid expansion of data-driven research across domains such as finance, healthcare, and  
<sup>21</sup> environmental sciences has increased the need for tools that support reproducibility, modularity,  
<sup>22</sup> and flexibility in data analytics. Researchers often need to construct and compare multiple  
<sup>23</sup> workflows, each differing in transformation methods, learning algorithms, or evaluation criteria.  
<sup>24</sup> However, managing this variability is time-consuming and error-prone when using traditional  
<sup>25</sup> scripting or static pipeline tools. Scientific workflow systems have advanced reproducibility but  
<sup>26</sup> often lack the flexibility required for experimentation.

<sup>27</sup> The concept of *Experiment Lines* (*EL*) ([Marinho et al., 2017](#)), derived from software product line  
<sup>28</sup> engineering, extends workflow design by introducing **variability** (alternative components) and  
<sup>29</sup> **optionality** (configurable presence or absence of steps). **daltoolbox** operationalizes *EL* principles  
<sup>30</sup> for data analytics, providing a practical, code-based framework for managing experimental  
<sup>31</sup> diversity.

## <sup>32</sup> Statement of Need

<sup>33</sup> Data analytics workflows frequently require the exploration of multiple preprocessing, modeling,  
<sup>34</sup> and evaluation alternatives. Managing these alternatives often leads to repetitive code,  
<sup>35</sup> fragmented design, and limited traceability, which hinder reproducibility across experiments.  
<sup>36</sup> **daltoolbox** was developed to address this challenge by enabling modular and flexible experiment  
<sup>37</sup> definition through a unified interface.

<sup>38</sup> The **target audience** includes researchers, educators, and data practitioners who require

39 transparent, reproducible workflows for experimentation in classification, regression, clustering,  
40 and time series prediction. The package is particularly valuable in academic and applied  
41 research contexts, where multiple analytical alternatives must be compared under controlled  
42 conditions.

43 daltoolbox provides a consistent syntax and modular architecture that facilitate systematic  
44 experimentation. Users can easily modify, replace, or omit workflow components, allowing  
45 efficient exploration of design alternatives while preserving reproducibility and transparency.

## 46 State of the Field

47 Several tools exist for designing machine learning workflows. Visual environments such as  
48 **WEKA** (Witten et al., 2016), **Orange** (Demsar et al., 2013), and **KNIME** (Berthold et al.,  
49 2009) are widely used for education and prototyping but offer limited flexibility for dynamic  
50 reconfiguration. Frameworks such as **Scikit-learn** (Pedregosa et al., 2011) and **MLlib** (Meng et  
51 al., 2016) provide robust APIs but focus on static pipelines rather than structured workflow  
52 variability. AutoML systems like **Auto-WEKA** (Kotthoff et al., 2017) and **Auto-sklearn** (Feurer  
53 et al., 2015) automate model selection but reduce user control and transparency.

54 daltoolbox differentiates itself by offering explicit modeling of variability and optionality,  
55 allowing controlled exploration of alternatives. This focus on transparency and user-driven  
56 design complements rather than replaces existing ML frameworks, positioning daltoolbox as an  
57 intermediary layer for reproducible experimentation.

## 58 Main Features

- 59   ■ Unified API for **transformation**, **classification**, **regression**, and **clustering**.
- 60   ■ Explicit modeling of *optional* and *variable* workflow components.
- 61   ■ Modular operators for scaling, normalization, and dimensionality reduction.
- 62   ■ Easy substitution of models and preprocessing steps without code refactoring.
- 63   ■ Visualization utilities for model comparison and interpretation.
- 64   ■ Interoperability with external R and Python libraries.
- 65   ■ Comprehensive documentation and testing, distributed under the MIT license.

## 66 Example Usage

```
# Define a tiny workflow runner once
DemoWorkflow <- function(model, prep, train, test) {
  prep <- fit(prep, train)
  train <- transform(prep, train)
  model <- fit(model, train)
  predict(model, test)
}

# Scenario A: skip transformation (no-op) + KNN
prep_a <- dal_transform() # no-op transformer
model_a <- cla_knn("rain", levels = c("yes", "no"), k = 3)
preds_a <- DemoWorkflow(model_a, prep_a, train, test)

# Scenario B: min-max normalization + Random Forest
prep_b <- minmax()
model_b <- cla_rf("rain", levels = c("yes", "no"))
preds_b <- DemoWorkflow(model_b, prep_b, train, test)
```

<sup>67</sup> This pattern shows how a single workflow function enables testing alternative pipelines by  
<sup>68</sup> switching only the prep or model component, without refactoring code.

## <sup>69</sup> Acknowledgements

<sup>70</sup> This work was partially supported by **CNPq**, **CAPES**, and **FAPERJ**. The authors acknowledge  
<sup>71</sup> the DAL community and institutional partners for their support.

## <sup>72</sup> References

- <sup>73</sup> See paper.bib for the complete list of references.
- <sup>74</sup> Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K.,  
<sup>75</sup> & Wiswedel, B. (2009). KNIME - the konstanz information miner: Version 2.0 and beyond.  
<sup>76</sup> *Journal of Machine Learning Research*, 11, 3191–3195.
- <sup>77</sup> Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M.,  
<sup>78</sup> Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J.,  
<sup>79</sup> Zitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of*  
<sup>80</sup> *Machine Learning Research*, 14(71), 2349–2353.
- <sup>81</sup> Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Ef-  
<sup>82</sup> ficient and robust automated machine learning. *Advances in Neural Information Processing*  
<sup>83</sup> *Systems*, 2962–2970.
- <sup>84</sup> Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2017). Auto-WEKA  
<sup>85</sup> 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of*  
<sup>86</sup> *Machine Learning Research*, 18, 1–5.
- <sup>87</sup> Marinho, A., Oliveira, D. de, Ogasawara, E., Silva, V., Ocana, K., Murta, L., Braganholo,  
<sup>88</sup> V., & Mattoso, M. (2017). Deriving scientific workflows from algebraic experiment  
<sup>89</sup> lines: A practical approach. *Future Generation Computer Systems*, 68, 111–127. <https://doi.org/10.1016/j.future.2016.08.016>
- <sup>91</sup> Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D.  
<sup>92</sup> B., Amde, M., Owen, S., Xin, D., Franklin, M. J., Zadeh, R., Zaharia, M., & Talwalkar, A.  
<sup>93</sup> (2016). MLLib: Machine learning in apache spark. *Journal of Machine Learning Research*,  
<sup>94</sup> 17, 1–7.
- <sup>95</sup> Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,  
<sup>96</sup> Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,  
<sup>97</sup> Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python.  
<sup>98</sup> *Journal of Machine Learning Research*, 12, 2825–2830.
- <sup>99</sup> Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine*  
<sup>100</sup> *learning tools and techniques* (4th ed.). Morgan Kaufmann. ISBN: 978-0-12-804357-8