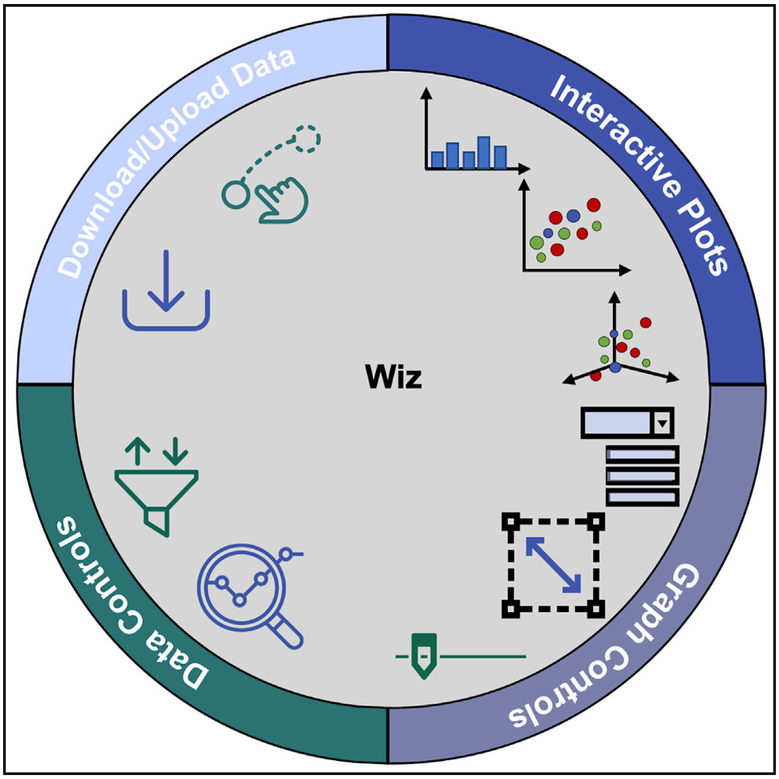
Descriptor

Patterns

Wiz: A Web-Based Tool for Interactive Visualization of Big Data

Graphical Abstract



Highlights

- Freely accessible web app for interactive big data visualization
- Built modularly for easily incorporation of new data analytics/ visualization features
- Lowers the barrier to entry for visualizing large datasets

Authors

Christopher Balzer, Rama Oktavian,
Mohammad Zandi,
David Fairen-Jimenez,
Peyman Z. Moghadam

Correspondence

df334@cam.ac.uk (D.F.-J.),
p.moghadam@sheffield.ac.uk (P.Z.M.)

In Brief

Data have become bigger than ever thanks to strides in computing and high-throughput experimental methods. Visualizing these large and complex datasets effectively requires the ability to quickly explore and interpret subsets of the data. Most tools for effective visualization require user programming or expensive commercial software. Wiz is a freely accessible web app that anyone with a browser can use to interactively visualize their data in multi-dimensions without any knowledge of a programming language.

Balzer et al., 2020, Patterns *1*, 100107
November 13, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.patter.2020.100107

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

Wiz: A Web-Based Tool for Interactive Visualization of Big Data

Christopher Balzer,[1,2] Rama Oktavian,[1] Mohammad Zandi,[1] David Fairen-Jimenez,[3,*] and Peyman Z. Moghadam[1,4,*]
[1]Department of Chemical and Biological Engineering, University of Sheffield, Sheffield S1 3JD, UK
[2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[3]Adsorption & Advanced Materials Laboratory (A^{2}ML), Department of Chemical Engineering & Biotechnology, University of Cambridge, Cambridge CB3 0AS, UK
[4]Lead Contact
*Correspondence: df334@cam.ac.uk (D.F.-J.), p.moghadam@sheffield.ac.uk (P.Z.M.)

https://doi.org/10.1016/j.patter.2020.100107

THE BIGGER PICTURE As data become bigger and more complex, ready-to-use visualization techniques are crucial in discovering patterns in unstructured data. Most methods to visualize big data effectively require access to commercial software or expertise in a programming language. Here, we introduce Wiz, a freely accessible web app that anyone with a browser can use to interactively visualize their data in multi-dimensions. Wiz serves as a framework for interactive dashboards that can be utilized by researchers to make their publication data available or commercially for Industry 4.0 applications via serving real-time data from process equipment. Wiz will continue to grow in functionality to be a one-stop-shop for both interactive data visualization and data analysis in the coming years.

1 2 **3** 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

In an age of information, visualizing and discerning meaning from data is as important as its collection. Interactive data visualization addresses both fronts by allowing researchers to explore data beyond what static images can offer. Here, we present Wiz, a web-based application for handling and visualizing large amounts of data. Wiz does not require programming or downloadable software for its use and allows scientists and non-scientists to unravel the complexity of data by splitting their relationships through 5D visual analytics, performing multivariate data analysis, such as principal component and linear discriminant analyses, all in vivid, publication-ready figures. With the explosion of high-throughput practices for materials discovery, information streaming capabilities, and the emphasis on industrial digitalization and artificial intelligence, we expect Wiz to serve as an invaluable tool to have a broad impact in our world of big data.

INTRODUCTION

Scientific data become bigger and more complex each year. The development of new simulation techniques for materials characterization and generation of databases for various functional materials has led to the emerging area of large-scale computational screening where properties of thousands, or millions, of materials can be assessed using fast computing clusters.[1–3] In parallel, development of new, automated instruments allows researchers to perform high-throughput experiments to design and discover new materials.[4,5] While this idea has long been essential for industrial-driven fields, such as drug discovery, the last decade has seen an uptick in high-throughput screenings in various fields within material science, chemistry, and biotechnology.[6–10] Such experimental/computational efforts generate high-dimensional datasets with complex relationships between the variables. The ability to derive meaningful relationships from such large datasets depends on our access to analysis tools. In particular, data visualization tools play an essential role in understanding and communicating results from large datasets.

Standard data visualization is quite routine in developed software, such as Python, MATLAB, R, and Excel; however, static, 2D images naturally fail to capture the full story of a complex

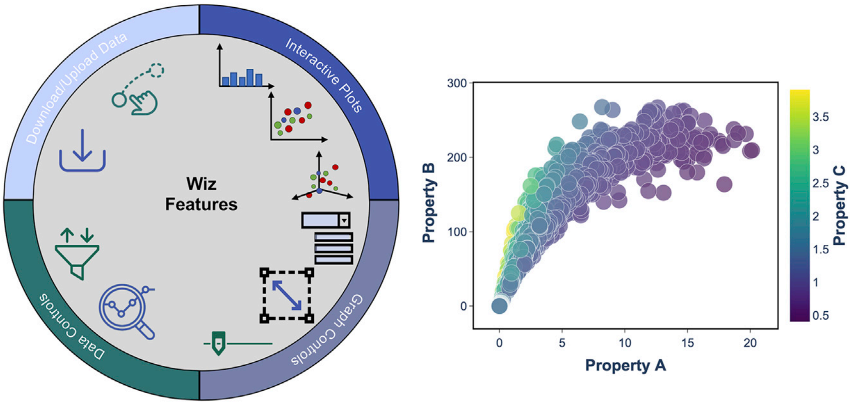


Figure 1. Graphical Overview of Wiz

Diagram of main features of Wiz (left). *Interactive plots*: Wiz provides a range of graph types with hover- and click-driven insights. *Graph controls*: dropdowns, data sliders, and in-graph zooming/panning allow users to quickly switch between datasets and focus on the most important parts of their data. *Data controls*: built-in analysis and data filtering allows users to manipulate their data without affecting the underlying data. *Download/upload data*: users can drag-and-drop datasets into Wiz and can download their plot data and export their figures with publication quality. Prototypical plot produced in the Wiz app (right).

dataset. As outlined in the classical work by Shneiderman,[11] effective visualization of large datasets requires some degree of interactivity, or a user-controlled experience. In academia, many online journals have started fighting against static figures in favor of interactive, or "living" figures.[12] Interactive figures give readers the ability to zoom in-on, click on, and pan across the underlying data in a figure. Not only does this give readers increased ability to understand authors' conclusions but living figures enhance access to the underlying dataset, facilitate reproducibility, and encourage potentially new conclusions to be made about a dataset. However, without universal adoption of interactive figures over static figures, the responsibility to create interactive figures falls on individual researchers.

Interactive data visualization and exploration can be a daunting task for many scientists. Data visualization tools and packages exist for a variety of programming languages. Some Python examples include Plotly,[13] Altair,[14] pygal,[15] and to a lesser extent Bokeh,[16] Gleam,[17] and Matplotlib.[18] However, most visualization tools are declarative, which means a user must identify the columns or series to plot. Declaration requires (1) user knowledge of the programming language and (2) previous knowledge of the dataset. Many scientists find programming cumbersome to learn, use, or transfer to others. In addition, some users want to quickly plot many relationships from their datasets without having to declare each plot. Commercial software, such as Tableau,[19] Sisense,[20] JMP,[21] or Biovia's Pipeline Pilot[22] provide tools for data visualization but also require licensed software and/or do not have fully interactive plots. Thus, there is a need for a widely accessible, easy-to-use, and easy-to-build platform to create interactive visualizations.

In this work, we address these challenges by creating a web-based data visualization tool built with Dash by Plotly.[23] Named Wiz, the intention of the web app is to explore the relationships across large and complex datasets easily, quickly, and interactively. *Web-based* indicates that the app is accessible online, anytime at https://wiz.shef.ac.uk. Importantly, users require no programming skills to visualize desired datasets and can simply navigate to the above URL and begin using the app. The idea of web-based apps has been seen before in domain-specific applications,[24] but we envision Wiz to be a general tool across disciplines. While not limited to the following fields, we imagine that Wiz is best suited for datasets arising from screenings (computational and experimental) in the fields of materials science, chemistry, biological systems, and in numeric machine learning applications, enabling acquisition of *new* knowledge by exploring existing pieces of knowledge. The remainder of this work will outline the functionality of Wiz through examples in various applications.

RESULTS AND DISCUSSION

Overview of the App Build and Features

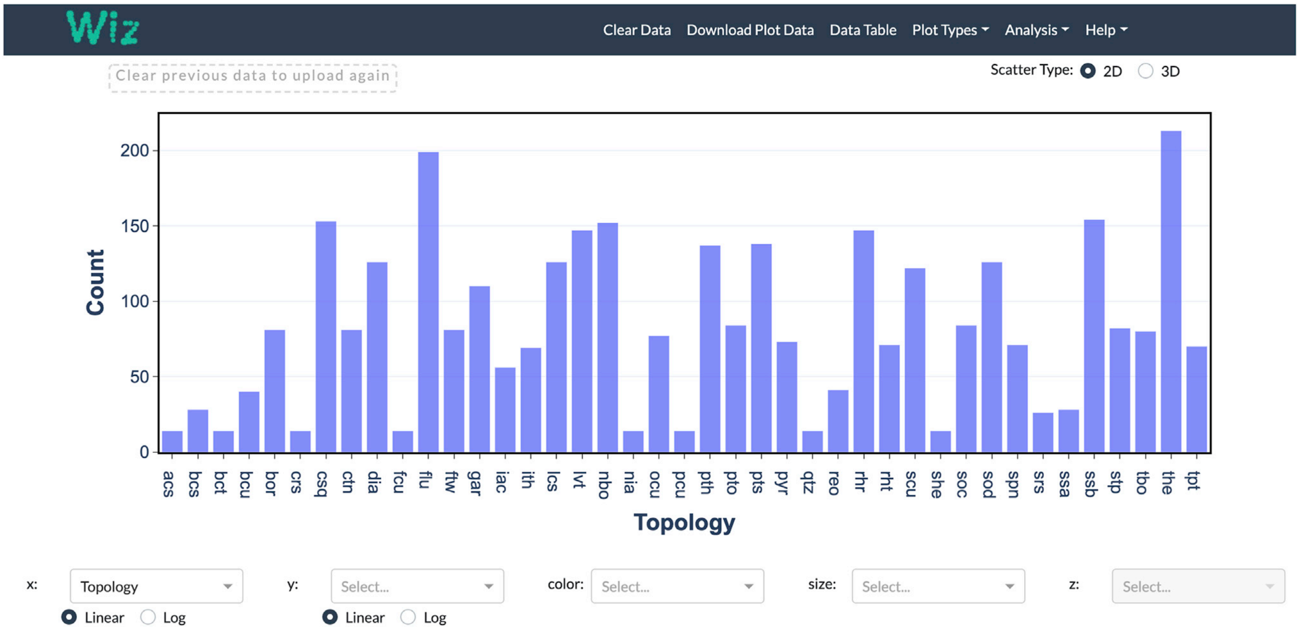
Wiz is built with Dash by Plotly,[23] a Python framework for building analytical web applications. A similar framework exists for R called Shiny.[25] The great benefit of Dash for scientists is that no Java or HTML is required. Dash itself is declarative and reactive, making the creation of basic applications easy for those familiar with Python. Most importantly, Dash already has the framework for interactive visualization with Plotly.

Wiz builds on the idea of Dash by making a visualization tool that requires no programming ability. To that end, Wiz removes the need to program routines for data upload, data filtering/processing, and the plotting commands. Anyone with a compatible dataset can create several types of stunning, interactive graphs by simply going to https://wiz.shef.ac.uk and uploading their dataset. Wiz has four main features that make interactive plotting easier than ever. These features are outlined graphically in Figure 1. We have also provided a public version of the Wiz code that can be used by other researchers to further develop or use in their own applications, https://github.com/peymanzmoghadam/Wiz.

On the backend, Wiz is highly modular, such that new features and graph types can be readily implemented. Each page within the app contains different essential components that makeup the layout (i.e., links, dropdowns, upload buttons, datatables) that are implemented using Dash. User interaction with an app page fires callbacks at the heart of the interactive experience. While the plotting routines and backend implementation of Wiz are well established, to the best of our knowledge we are the first to put the pieces together in such an easy-to-use, relevant app for data visualization. Combined with robust hosting through the University of Sheffield, Wiz is a one-of-a-kind multi-user platform.

Using Wiz across Science Domains

A number of example datasets are included in the Supplemental Information (Tables S1, S2, S3, S4, S5, and S6). A step-by-step



Scatter Type: ● 2D ○ 3D

x: Topology ● Linear ○ Log

y: Select... ● Linear ○ Log

color: Select...

size: Select...

z: Select...

Figure 2. User Interface of Wiz

Histogram of categorical data of over 3,000 metal-organic frameworks displayed in multiple graph types in Wiz. The user interface of Wiz emphasizes the graph with easy-access controls for the various graph components. Here, the ''Topology'' variable is selected in the axis selector. When only one axis is selected, a histogram is displayed in the graph window. Data from Moghadam et al.[30] (Table S1).

guide of the following examples shows the layout of the app and explains key functions of Wiz. While the applications in computational screenings and machine learning applications are discussed here, any compatible dataset can be visualized in Wiz without user programming or downloading software. Wiz is transferable between anyone with a browser and easy to use for scientists and non-scientists alike. Furthermore, the Supplemental Information includes a Video S1 to demonstrate the features in real time. For the most thorough guide, users should visit the Help documentation at https://wiz.shef.ac.uk/help.

High-Throughput Screenings for Materials Design and Discovery: The Value in Visualizing Structure-Property Relationships

Development of new materials would be greatly accelerated if we had a better understanding of the key properties that need to be optimized. Identification of such properties, and therefore top-performing materials, often require complex and time-consuming calculations and/or experiments. In such cases, development of data-driven insights and structure-property relationships are essential to reduce the search space and guide efforts toward selection of promising materials. High-throughput computational and experimental screenings allow us to study hundreds, thousands, or millions of materials to develop a mechanistic understanding of their performance. Such strategies have streamlined design in biomaterials, polymers, ionic liquids, nanomaterials, energy materials, and many more fields.[6,26–29] The examples below showcases a number of high-throughput screening studies in porous crystals called metal-organic frameworks (MOFs). In these examples, we highlight the immense value in structure-property relationships, which connect physical, geometric, or chemical properties to performance parameters from thousands of simulations. These relationships guide understanding of critical components of performance and guide experimental efforts to create better materials.

Figure 2 shows the user interface of Wiz and an example of creating histograms for a categorical dataset from Moghadam and colleagues.[30] The dataset contains physical and mechanical properties for over 3,000 MOFs with 68 attributes (Table S1). Once the data are uploaded into Wiz, one can simply use the dropdown menus to start analyzing the data (see Figure 2). For example, from the dropdown lists, if the ''Topology'' attribute is selected, a histogram is plotted. From Figure 2, one can see how the minimalist design of graph/data controls emphasizes the graph itself.

Often in MOFs, the properties are not static over an entire temperature or pressure space. In another example, nearly 3,000 MOFs were assessed for their oxygen storage through Monte Carlo simulations.[7] A key performance indicator for oxygen storage is the uptake of oxygen, which varies with pressure. Figure 3 shows two 4D plots comparing the gravimetric and volumetric uptake for 3,000 structures at two different pressures, 20 and 100 bar. For each plot, physical properties, i.e., density and the cavity diameter are shown using rainbow color scale and size, respectively, making the plot 4D. Once in ''3D'' mode, Wiz users can pick a fifth dimension (z axis) to create interactive 5D plots (see demo Video S1 in the Supplemental Information). Importantly, the data points are clickable and can display the values for the plotted dimensions (Figure 3A). Wiz makes it easy to upload and switch between multiple datasets with dropdowns for each of the axes. The slider at the bottom of the user interface displays the filenames, or sheet names, of each dataset. The utility is not only in the visualization itself, but the ease

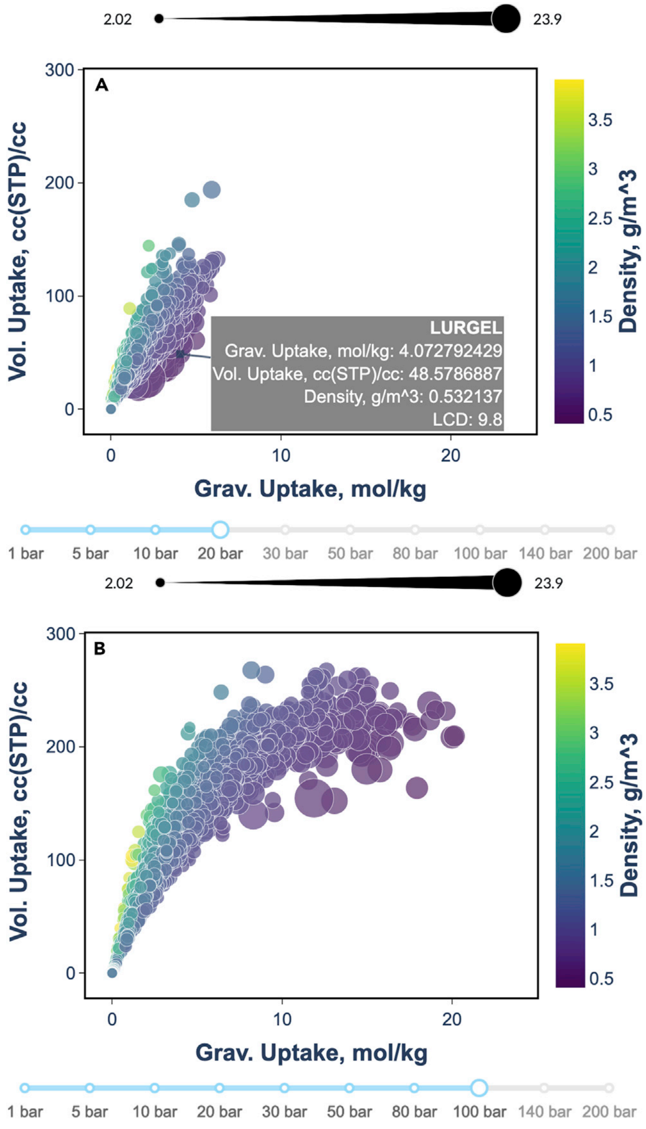


Figure 3. Multi-Dimensional Interactive Wiz Example Plots

Oxygen storage capacity for 3,000 MOFs at different pressures plotted in Wiz. The gravimetric and volumetric oxygen uptake, MOF density (color), and cavity diameter (size, Å) at 20 bar (A) and 100 bar (B). Inset in (A) shows clickable points effect where the full data are displayed for that point. Each data point represents a different structure. Wiz makes it easy to switch between pressures using the slider at the bottom of the user interface. Data are collected from Fairen-Jimenez et al.[7] (Table S2).

to upload potentially complex datasets and switch between them. For example, the data in Figure 3 come from a single Microsoft Excel file, where the data at each pressure are in the different sheets. However, different sheets do not have to have the same feature variables. Each web page in the app has more detail on what file types are accepted and formatting the input datasets.

Machine Learning and Large Datasets

Machine learning algorithms use statistical methods to learn, or make predictions, based on underlying relationships in a dataset. These datasets can exceed thousands of instances or hundreds of features. Previous examples showed how powerful Wiz for visualizing relationships between feature-like variables. The following example illustrates how Wiz can be used at different stages of the machine learning pipeline. Figures 4A–4C show data from the Movie Lens Dataset collected by Grouplens, a research group at the University of Minnesota.[31] These data consist of 100,000 movie ratings of 1,700 movies by 1,000 users. Wiz can be used for both initial visualization and analysis. Figure 4A shows histograms of the movie ratings versus the genre of the movies. Wiz automatically switches between plot types depending on whether the data are categorical. Here, the categorical abscissas automatically generates box-and-whisker plots—a convenient way of displaying data in terms of their quartiles while identifying outliers—describing how the data are distributed between genres. In the context of machine learning, visualization of the raw dataset with Wiz aides in identification of outliers and can help generate ideas for feature engineering. With Wiz, one can also perform basic principal-component analysis (PCA) or linear discriminant analysis (LDA) on a dataset. PCA gives a way of visualizing the relatedness (correlation) between descriptors in a dataset as understanding the dimensionality of a dataset. LDA reduces the dimensionality of a dataset to best separate classes of data (e.g., movie genres, top rated movies, most popular movies). As a demonstration, we show Wiz's usefulness in visualizing latent matrix factorization for recommender systems using the MovieLens dataset. In such a sparse dataset (many user/rating combinations missing), we can apply an SVD-like learning model to learn a "recommender" matrix from the product of two matrices—one for the user latent factors and one for the movie latent factors.[32] After training the recommender matrix, observing the learned movie latent factor matrix gives insight into what the recommender system learned. Figure 4B shows the Wiz plot generated from LDA on the movie latent factor matrix. Details of the training can be found in the Supplemental Information. The movies were divided into classes based on their average rating, signified by different colors in the plot. In Figure 4B, the top 5% rated movies are distinct from the lowest 5% rated movies, showing that the recommender system learned some structure related to the average user rating. With a dropdown selector, conducting PCA and LDA is automatic and easy to visualize. Figure 4C shows an example scree plot generated by Wiz from the latent matrix factorization, where the interactive data hovering makes analyzing variance contributors fast and easy. Other examples can be found at https://wiz.shef.ac.uk/examples. Note that only the 2D projection is produced from both PCA and LDA in Wiz.

Wiz can handle datasets exceeding 50,000 instances by utilizing WebGL plot elements for large datasets, as opposed to SVG elements for smaller datasets. Figure 4D shows an example of visualizing a large dataset, possessing 100,000 instances, as well as filtering that dataset in Wiz. For such a large dataset, often only a small portion is of interest. One of the most useful features of Wiz is the ability to filter graph data in-place. For example, the data in Figure 4D are filtered such that data points above a threshold value of 15 (x-data) is not plotted. The filtering process is as easy as typing inequalities or search terms in the data table tab ("<15," here). The ability to filter data easily without editing the underlying dataset is powerful for handling a dataset. Each of the datasets are provided in the Supplementary Information

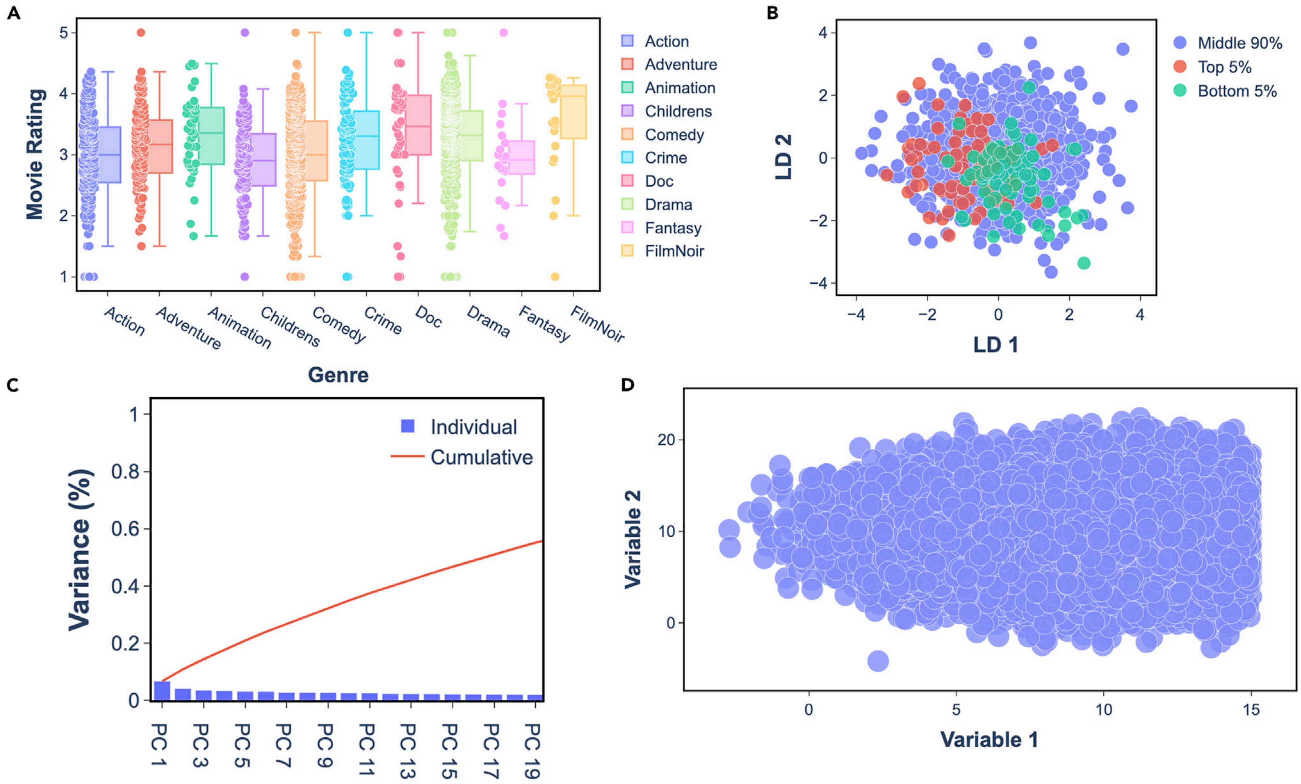


Figure 4. Exploration of Wiz Data-Handling Features

(A) Average movie rating versus movie genre for MovieLens dataset[31] as a box-and-whisker plot. Understanding the distribution and outlying data can aid in data processing before training.

(B) Example LDA projection using the top and bottom 5% rated movies as classes. Wiz's analysis makes seeing the structure in the learned matrix fast and easy.

(C) Scree plot output from PCA on MovieLens dataset (first 20 of 50 PCs shown).

(D) Data filter functionality shown on a manually generated dataset with 100,000 instances (Table S3).[33] The data are filtered in-place to show data points less than 15 (x axis). (A–D) are exported directly from Wiz.

for the user to investigate these features on their own (Tables S1, S2, S3, S4, S5, and S6).

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Peyman Z. Moghadam is the lead contact of this study and can be reached by email: p.moghadam@sheffield.ac.uk.

Materials Availability

This study did not generate new materials.

Data and Code Availability

The Wiz website is hosted by the University of Sheffield and can be freely accessed at https://wiz.shef.ac.uk/. All data uploaded to Wiz are only stored during the user session via cache and removed after the session is ended. The public version of Wiz is available in a Github repository https://github.com/peymanzmoghadam/Wiz.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100107.

ACKNOWLEDGMENTS

P.Z.M. thanks the Corporate Information and Computing Services (CiCS) and Partnerships and Regional Engagement at the University of Sheffield for providing partial funds for the project. D.F.-J. thanks the Royal Society for funding through University Research Fellowships. P.Z.M. also thanks John Dale from the University of Sheffield, and Yosof Badr and David Moss from Siemens for useful discussions. M.Z. acknowledges the Knowledge Exchange funding (X/013145) from the University of Sheffield. R.O. acknowledges Indonesia Endowment Fund for Education (LPDP) for funding his doctoral study and also acknowledges Muhammad Rifaldi from Brawijaya University for assisting in the design of front cover for this paper. The authors also thank the University of Sheffield for providing infrastructure to host Wiz.

AUTHOR CONTRIBUTIONS

D.F.-J. and P.Z.M. conceptualized the study. C.B. created the Wiz app and all documentation for Wiz under supervision of D.F.-J. and P.Z.M. P.Z.M. wrote all scripts to host Wiz and authored the initial draft of the manuscript. R.O. and M.Z. contributed to design and testing of Wiz. All authors contributed to manuscript review.

DECLARATION OF INTERESTS

P.Z.M. has financial interest through Monoclinic Ltd. D.F.-J. has financial interest through Immaterial Ltd. The other authors declare no competing interests.

Received: May 16, 2020
Revised: July 14, 2020
Accepted: August 25, 2020
Published: September 23, 2020

REFERENCES

1. Groom, C.R., Bruno, I.J., Lightfoot, M.P., and Ward, S.C. (2016). The Cambridge Structural Database. Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater. *72*, 171–179.

2. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al. (2016). PubChem substance and compound databases. Nucleic Acids Res. *44*, D1202–D1213.

3. Swain, M.C., and Cole, J.M. (2016). ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. J. Chem. Inf. Model. *56*, 1894–1904.

4. Entzeroth, M., Flotow, H., and Condron, P. (2009). Overview of high-throughput screening. Curr. Protoc. Pharmacol. *44*, 9.4.1–9.4.27.

5. Szymański, P., Markowicz, M., and Mikiciuk-Olasik, E. (2012). Adaptation of high-throughput screening in drug discovery-toxicological screening tests. Int. J. Mol. Sci. *13*, 427–452.

6. Cooper, C.B., Beard, E.J., Vázquez-Mayagoitia, Á., Stan, L., Stenning, G.B.G., Nye, D.W., Vigil, J.A., Tomar, T., Jia, J., Bodedla, G.B., et al. (2019). Design-to-device approach affords panchromatic co-sensitized solar cells. Adv. Energy Mater. *9*, 1802820.

7. Fairen-Jimenez, D., Fantham, M., Farha, O.K., Islamoglu, T., Goswami, S., Moghadam, P.Z., Exley, J., Kaminski, C.F., and Snurr, R.Q. (2018). Computer-aided discovery of a metal–organic framework with superior oxygen uptake. Nat. Commun. *9*, 1378.

8. Korth, M. (2014). Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods. Phys. Chem. Chem. Phys. *16*, 7919–7926.

9. Shevlin, M. (2017). Practical high-throughput experimentation for chemists. ACS Med. Chem. Lett. *8*, 601–607.

10. Moghadam, P.Z., Li, A., Wiggin, S.B., Tao, A., Maloney, A.G.P., Wood, P.A., Ward, S.C., and Fairen-Jimenez, D. (2017). Development of a Cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future. Chem. Mater. *29*, 2618–2625.

11. Shneiderman, B., Eyes have it: a task by data type taxonomy for information visualizations. In IEEE Symposium on Visual Languages, Proceedings; 1996; pp 336–343.

12. Perkel, J.M. (2018). Data visualization tools drive interactivity and reproducibility in online publishing. Nature *554*, 133–134.

13. Plotly Technologies Inc. (2020). Collaborative data science. https://plot.ly/.

14. VanderPlas, J., Granger, B.E., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., and Sievert, S. (2018). Altair: Interactive Statistical Visualizations in Python. Journal of Open Source Software *3*, 1057.

15. Kozea (2016). Pygal: dynamic SVG charting library. http://pygal.org/.

16. Bokeh Development Team (2018). Bokeh: Python library for interactive visualization. https://bokeh.pydata.org/.

17. Robinson, D. (2014). Gleam. https://github.com/dgrtwo/gleam.

18. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering *9*, 90–95.

19. Tableau Software, LLC (2020). Tableau. https://tableau.com/.

20. Sisense, Inc (2020). Sisense. https://sisense.com/.

21. SAS (1989-2020). JMP (Cary, NC, SAS Institute).

22. Dassault Systèmes, BIOVIA. (2020). Pipeline Pilot (San Diego: Dassault Systèmes).

23. Plotly Technologies, Inc. (2020). Dash by Plotly. https://dash.plotly.com/.

24. Holzinger, A., and Zupan, M. (2013). KNODWAT: a scientific framework application for testing knowledge discovery methods for the biomedical domain. BMC Bioinformatics *14*, 191.

25. RStudio, Inc. (2017). Shiny: Web application framework in R. https://shiny.rstudio.com/.

26. Le, T., Epa, V.C., Burden, F.R., and Winkler, D.A. (2012). Quantitative structure-property relationship modeling of diverse materials properties. Chem. Rev. *112*, 2889–2919.

27. Goldsmith, B.R., Boley, M., Vreeken, J., Scheffler, M., and Ghiringhelli, L.M. (2017). Uncovering structure-property relationships of materials by subgroup discovery. New J. Phys. *19*, https://doi.org/10.1088/1367-2630/aa57c2.

28. Wang, Y., Zhang, W., Chen, L., Shi, S., and Liu, J. (2017). Quantitative description on structure–property relationships of Li-ion battery materials for high-throughput computations. Sci. Technol. Adv. Mater. *18*, 134–146.

29. Ostuni, E., Chapman, R.G., Holmlin, R.E., Takayama, S., and Whitesides, G.M. (2001). A survey of structure-property relationships of surfaces that resist the adsorption of protein. Langmuir *17*, 5605–5620.

30. Moghadam, P.Z., Rogge, S.M.J., Li, A., Chow, C.-M., Wieme, J., Moharrami, N., Aragones-Anglada, M., Conduit, G., Gomez-Gualdron, D.A., Van Speybroeck, V., and Fairen-Jimenez, D. (2019). Structure-mechanical stability relations of metal-organic frameworks via machine learning. Matter *1*, 219–234.

31. GroupLens (1998). MovieLens 100K dataset. https://grouplens.org/datasets/movielens/100k/.

32. Funk, S. (2006). Netflix Update: Try this at Home. http://sifter.org/∼simon/journal/20061211.html.

33. Bache, K., and Lichman, M. (2020). UCI machine learning repository. http://archive.ics.uci.edu/ml/.