

Transformative Opportunities from Data Science and Big Data Analytics: Applied to Photovoltaics

by Laura S. Bruckman

Understanding the overall lifetime performance of photovoltaic (PV) modules is essential to continue the cost reduction of solar energy, thereby increasing its contribution to the world's electricity needs and sustainability goals.^{1,2} In order to reach the 2030 SunShot goal of \$0.03/kWh,^{3,4} the power degradation rates (representing the annual reduction in power output by a PV module) must be lowered to the 0.2%/year goal, so as to increase the lifetime of PV modules installed in diverse climates zones. Predicting the PV module performance over their 20- to 30-year product warranty or lifetime is typically done using traditional reliability approaches such as pass/fail testing and materials qualification, yet this has proven insufficient. Current PV qualification tests have led to failures in real-world PV applications.⁵ Historical data from installed PV systems is the ideal source for understanding the magnitude and causes of module power loss and degradation, and for identifying how to extend the lifetime of PV modules to 50 years.

The largest set of historical data available is time series power data (typically one or five minute interval datastreams) from commercial and research PV power plants. Handling such extensive PV data sets, which can extend for 20 years, required new informatics and analytical approaches to derive scientific insights. A big data analytics approach, utilizing Hadoop⁶ and a non-relational data warehouse was developed to handle the large volume real-time data streams.⁷⁻¹⁰ Graduate and undergraduate students, across numerous academic departments, needed to learn data cleaning and assembly, statistical analysis, coding, data-driven modeling, and statistical and machine learning. This also requires the use of open and reproducible science methods, with shared code and data augmenting traditional journal publications.¹¹

In this paper, we describe some of the challenges and opportunities associated with acquiring the data, structuring the data, and performance analytics in meaningful ways while also respecting the privacy concerns of collaborators across the PV value chain. The approach we take, which we refer to as *engineering epidemiology*, draws upon medical research study designs and protocols for understanding PV modules, components, and materials under accelerated exposures and real-world, in-use, conditions. Domain knowledge of materials science, combined with network models of materials, components, and systems, allows the capture of multiple phenomena as a system of equations, for the understanding of which particular mechanisms are induced by multivariate stressors, and how those relate to meaningful overall performance metrics across dimensions and temporal scales.

We believe as data science and big data analytics grow in the solar field, the cost of PV electricity will continue to decrease by improving module lifetimes, performance, and decreasing the operational and maintenance burden on commercial PV plant owners. We also articulate some of the key emerging needs, such as greater use of image analysis of modules during manufacturing, large-scale image analysis of installed PV, greater use of current and voltage (I-V) curves, and improved solar forecasting.

Long-Term Data Acquisition

Initial Setup of Outdoor Example

Monitoring commercial field power data provides a big data opportunity to predict power output for PV fields, monitor PV module degradation or inverter failure, and understand the impact of weather on module performance and lifetime. Commercial PV fields have a large set of streaming power data collected at small time intervals (e.g., every 1, 5, or 15 min). This type of data is not difficult to generate, since module inverters (string or micro) report it automatically. However, handling and obtaining this data required developing novel methods for solar analytics.

We developed an outdoor test facility in Cleveland, OH, with 148 modules from 20 brands purchased on the open market.¹² The time series data is ingested into a non-relational database based on Hadoop, Hbase, and Spark called CRADLE (Fig. 1).⁷ This type of database allows for data to be ingested as triples with row and column keys making an agile database that doesn't require researchers to know a priori all data that is planned to be collected. This initial project led the way for us to identify the best practices of data handling, including a wholly separate metadata database, which allows for proprietary information (e.g., brand, manufacturer, cell type, specific location) to be kept separate from the power time series data. RedCAP was chosen in this case since it has high security requirements (HIPAA compliant) and prevents access to sensitive data based on user rights.¹³ A seven-digit alphanumeric code, based on salting the variable value and then hashing with scrypt, is used to de-identify and anonymize the proprietary identifiers associated with each PV power plant data source. Then multiple sources of data can be merged and analyzed together, increasing the statistical power of an analysis. All analytical results can be written back to CRADLE, enriching the dataset, and providing new results as inputs for further analysis. Code is developed using agile software development tools¹⁴ including Git code versioning,¹⁵ Jira issue tracking, and Slack team communications. This allows for everyone to share in the analytical code development, testing, packaging, and validation in a reproducible way.¹⁶

Data Handling with Corporate Partners

By developing methodology on our own test site, we were then capable of safely ingesting and maintaining data from multiple competitive sources. However, to get information beyond just the power output, we also needed weather data from sensors on the ground and from satellite sources. Sensor data at sites have a high prevalence of sensor soiling, inaccurate data capture, and outages, so satellite data can fill in missing data. Data from corporate partners needs to contain metadata about a particular site or sample. All data generated should be provided without pre-cleaning by the corporate partners, for things such as missing data or failed inverters. The best format for data is plain text, as opposed to the less robust proprietary binary file formats. Data are typically captured from accessing an API or by web scraping (e.g., Selenium). Flaws in the data are important for the researcher to clean reproducibly across all sets of data. CRADLE now includes 3.4 GW of time series power and weather

(continued on next page)

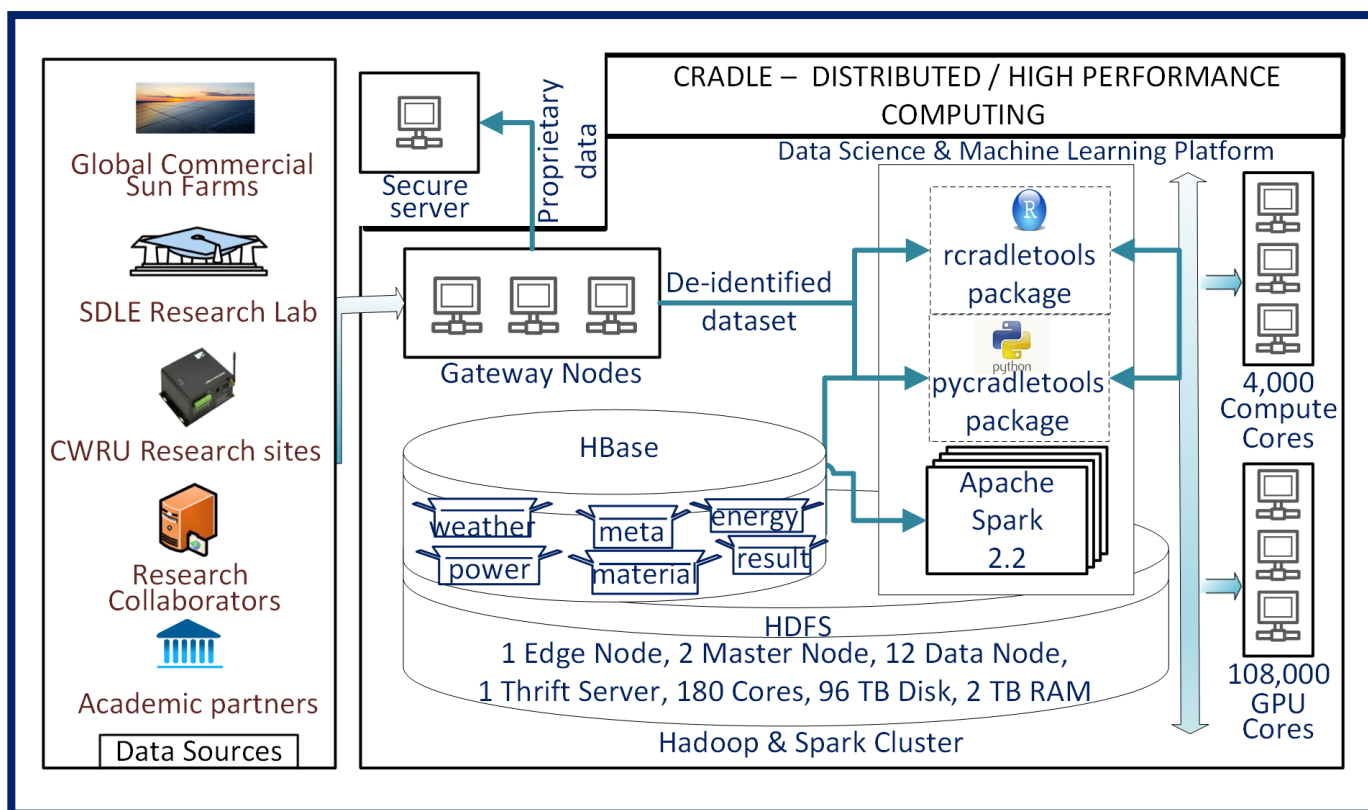


FIG. 1. The diagram of the non-relational database for CRADLE showing the ingestion of multiple data sources.

sensor data from 1486 PV sites (different c-Si cell technologies)¹⁷ and 7000 PV inverters over those 15 years.⁷ CRADLE is used for data from materials, components, and modules, not simply time series power data. The impact of performance loss rate determination strongly impacts the module manufacturers and how they develop their products.

Figure 2 shows the PV value chain where data comes from material, component, and module manufacturers to improve module reliability. This data can inform PV plant developers and owners on choices for commercial sites. Much of the module data is not open data since the manufacturer wants to protect intellectual property, but these data can be obtained from the developers and owners who have a vested interest in understanding failure and reliability. A third party incorporating data along the PV value chain will help move the entire industry forward and benefit society to increase the impact of PV renewable energy. Assembling data, developing and sharing codes and tools, and reporting research results to the whole PV value chain, as opposed to just the PV research community, is of increasing importance.

Inferring Degradation Mechanisms for Improving Materials, Components, and Systems

Data-Driven Modeling

Packaging materials (e.g., encapsulants, backsheets, framing, junction boxes) are key to extending the lifetime of PV modules. Backsheets, typically a three-layer polymer laminate on the back of modules, are important for module safety due to the high dielectric breakdown strength of poly(ethylene terephthalate) (PET). Backsheet degradation occurs as cracks,¹⁸ delamination, bubbling, burns, and discoloration.^{19,20} Data-driven models of PET show that populations of the same material can diverge in performance under accelerated exposure conditions. Multiple samples need to be exposed to capture

the distribution in the behavior of materials or systems. Evaluations during exposures (in-use or accelerated) need to be taken step-wise through time in order to develop a robust predictive model of behavior (<stressors|response>) by increasing data density.

A data-driven mixed-effect model was developed for PET degradation and showed that there is a change point shown in the degradation where damage accumulation arising from an initial mechanism reached a critical threshold and transitioned to a new mechanism. This change point behavior in PET yellowing and hazing would not be observed if samples were not characterized regularly during exposures.^{21,22} PET studied under both outdoor and accelerated lab-based exposure conditions show not just different stages of degradation, but also the synergistic effects of combined stressors since moisture as dew is an important stressor along with temperature and irradiance.²³ In PET degradation, the presence of water leads to synergistic photo-hydrolytic degradation in outdoor exposures, beyond photolytic degradation; simple exposure to humidity alone does not induce this photo-hydrolytic degradation.²¹ This is an example where combined effects in a material or complex system are not simply additive, and cannot be studied “one at a time, in isolation” as a hallmark of the scientific method (controlling all other variables when studying the one of interest).²⁴ Field studies are essential to ensure that the degradation occurring in accelerated exposures is the same mode, even if it has a different rate.²⁵ This requires accurately relating stress and degradation, while cross-correlating between accelerated and real-world conditions.²⁶

Mechanistic Understanding

Network modeling (netSEM package)²⁷⁻²⁹ provides a useful method to identify and quantify the rank-ordered degradation mechanisms and pathways activated in PV module degradation (Fig. 3). Network modeling determines the strength of relationships between stressors, mechanistic, and response variables (stressors|mechanism|response>) using step-wise regression across the network to identify parallel and sequential degradation pathways from the stressors to the system responses.³⁰ This is done using statistical learning and multiple and

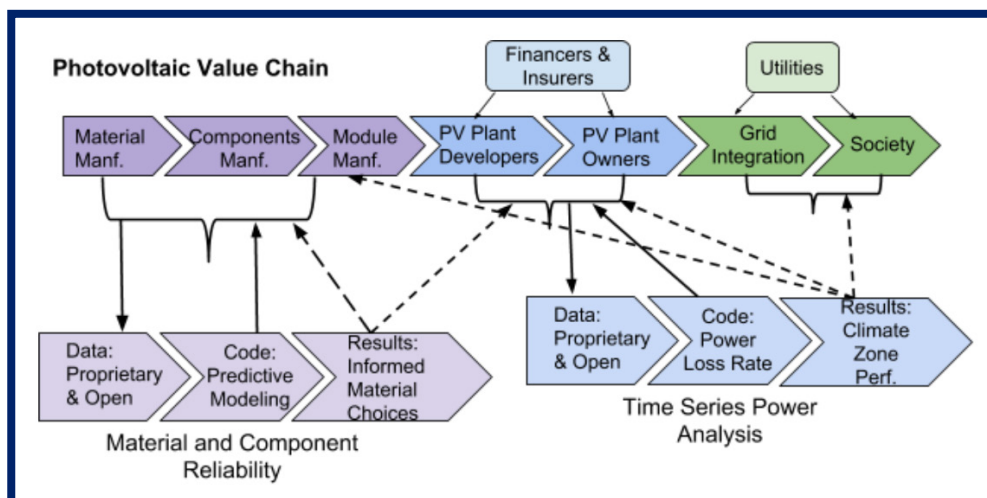


FIG. 2. The PV value chain has many sources of proprietary and open data. Data from materials, components, and modules manufacturers can be merged to improve the reliability of PV modules by informing material choices (purple). The PV developers and owners (blue) have their own data sets that can be informed from the module manufacturers. All these data work to improve grid integration and society (green) with an increase of reliable renewable PV energy.

multivariate regression with functional forms constrained by the fundamental physics and chemistry of the materials. This allows for different materials, material grades, or different systems to be compared in their degradation pathways based on various stressors. It is key to study the degradation of real-world modules in PV system field studies because degradation mechanisms occur in the field differently than under accelerated exposures.^{31,32} The lifetime and reliability of novel technologies need to be compared to established technologies; laboratory and field data from new technologies needs to be rapidly compared to the previous technologies. This requires a robust accelerated exposure methodology beyond the current pass/fail standard testing.

Opportunities for Advanced Data Analytics in Solar Energy

Image Analysis

Image-based photovoltaic module characterization comprises a broad category of techniques that provide unique, high-density, detailed information at the cell, material, and module levels. Imaging methods such as fluorescence, electroluminescence, photoluminescence, visual imagery, and thermography have each been well developed in laboratory settings, and have all recently been used outdoors,³³⁻³⁷ with many being commercialized for real-world implementation. Electroluminescence,^{38,39} fluorescence,⁴⁰ and photoluminescence imaging are typical lab-based image techniques

Recent advances in analysis of electroluminescence images of PV modules have enabled extraction of I-V characteristics of the module and individual cells.^{42,45} Automated processing of these image types is paramount for the utility of these techniques to be embraced by the PV community. Images from PV modules in accelerated and real-world conditions can be used to determine lifetime performance with open-source tools to process and evaluate these images in an automated fashion.

Integrating PV onto the Electrical Grid

Accurate solar forecasting is necessary as PV generation becomes a larger portion of the electrical grid in the U.S.⁴⁶ and elsewhere around the world.⁴⁷ It is necessary for the solar energy community to develop and validate technologies that enhance situational awareness of photovoltaic systems, from utility scale to behind-the-meter, and storage, to support reliability, resilience, and operation of power systems with high PV penetration.⁴⁸⁻⁵⁰ This involves characterizing and forecasting physical and cyber factors that impact resilience and then using graph-based modeling of thousands of PV systems to determine the normal expected performance and the power and energy lost due to these resilience factors. Spatio-temporal geographic information systems^{51,52} (e.g., IBM's PAIRS)⁵³ enable the study of historical weather events such as hurricanes, hail and high winds, along with vegetation and trees, and enables forecasting their future likelihood in different regions and areas. Historical power time series from behind the meter, distributed and utility scale PV power plants can be integrated with spatio-temporal weather data to study resiliency factors on historical power and energy production of PV systems.

Challenges

Data-Enabled Workforce

One challenge to the research community is a workforce that is educated in science and engineering and data enabled with a good understanding of statistics⁵⁴ to produce a T-shaped researcher trained in both the domain science and data science skills so that they can easily use data science tools in research and the PV industry.^{55,56} This workforce should be well versed in the benefits of reproducible research for a company so that teams of researchers can work and develop code together and data is available for future analysis.^{16,57}

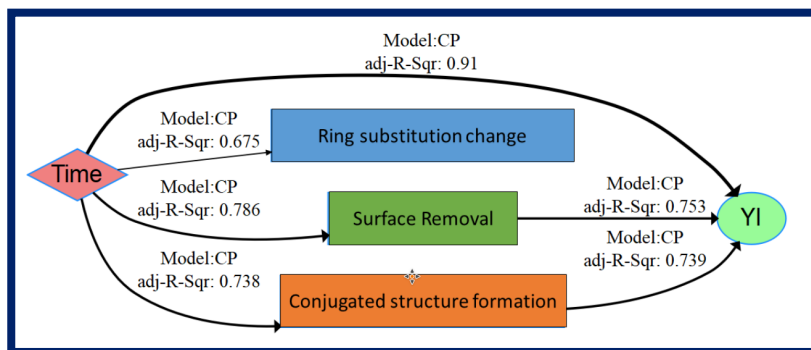


FIG. 3. A degradation pathway model from the netSEM package for a PET backsheets under irradiance and water spray accelerated conditions (all relationships are change point (CP) with adjusted R^2 noted). Mechanisms were determined by FTIR analysis and yellowness index was the performance metric. This model shows that the formation of conjugated structures is leading to yellowness index while water spray is removing the degraded material surface.

(continued on next page)

Reproducible research does not mean that the results have the correct interpretation and the ability to evaluate modeling results.⁵⁸ In the broad solar community, reproducible methods to monitor PV module performance are key to maintaining and growing PV in the grid. A solar community workforce would also be exposed to open source coding languages like Python and R. These open source tools can be used to streamline coding practices within a company with well-supported code options.^{59,60}

Open Data

Data needs to be in a format that is useable. This requires agreed-upon metadata, file formats, and a databook with information describing what headers relate to in the database. Professional societies can play a key role in reviewing and organizing open data after publication; however, this runs the risk of being difficult and time consuming for the professional society. Researchers uploading their own data into an open source repository (e.g., GitHub, Bitbucket, GitLab) can enable collaborations, research validation, and address new research problems. Software and data sharing will accelerate research discovery in the solar energy field. Sandia has shared many PV software tools in their PVLIB Git repository.⁶¹

Open software and data that is archived (with a digital object identifier (DOI)) should be considered in promotion and tenure within academia. Universities are beginning to promote open science practices, and publishing data is paramount in an open culture. Researchers who are moving in that direction should be rewarded through the promotion and tenure process especially since curated data sets are research products that must be balanced against time spent on other published works.⁶²⁻⁶⁴

Conclusions

Data science and big data analytics will drive down the cost of installed solar power by increasing the reliability of modules. While machine learning is a valuable method, for learning and prediction in the solar field, data science techniques with data-driven statistical models are necessary. Data science will inform optimal module architecture and material choices, improving reliability and extending lifetimes beyond the current 25-30 year modules. Big data analytics will begin to give accurate prediction of power output for solar fields based on climate zones, local weather conditions, and accurate lifetime prediction. This informs financial organizations, power plant owners and grid operators to reduce the variability of renewable energy.

Academia needs to support students to be data enabled in the solar field by teaching data handling, statistics, and coding along with traditional science and engineering. The solar industry will benefit from open data sets available for analysis and prediction of PV module lifetime, degradation rates, and maintenance. These open data sets will also help teach the next generation of students. ■


© The Electrochemical Society. DOI: 10.1149/2.F07191if.

About the Author



LAURA S. BRUCKMAN earned her PhD in analytical chemistry from the University of South Carolina with Prof. M. Myrick, focusing on image analysis and predictive modeling. Her research is focused on a data science approach to materials degradation. She is an expert on the application of quantitative spectroscopic techniques and image analysis to understand material degradation in relation to particular exposure stressors. A material data science approach using statistical analytics is used to develop degradation network pathway (netSEM models) diagrams for material systems. These pathway diagrams elucidate the impact of materials and

stressors and their relationship to overall degradation and lifetime performance loss. By encompassing data from both in-use and accelerated experiments of the degradation of materials, lifetime predictions can be made for material systems under a wide variety of use conditions. She may be reached at laura.bruckman@case.edu.

 <https://orcid.org/0000-0003-1271-1072>

References

1. International Energy Agency, World Energy Outlook 2018, Organization for Economic Co-operation and Development (OECD) / IEA, Paris, France, 2018.
2. A. Guterres, The Sustainable Development Goals Report 2018 | Multimedia Library - United Nations Department of Economic and Social Affairs, United Nations, Department of Economic and Social Affairs, (2018). <https://www.un.org/development/desa/publications/the-sustainable-development-goals-report-2018.html> (accessed January 1, 2019).
3. Department of Energy Sunshot 2030, <https://www.energy.gov/eere/solar/sunshot-2030> (accessed January 4, 2018)
4. R. Jones-Albertus, D. Feldman, R. Fu, K. Horowitz, and M. Woodhouse, *Prog. Photovoltaics*, **24**, 1272 (2016).
5. D. C. Jordan and S. R. Kurtz, *Prog. Photovoltaics*, **21**, 12 (2013).
6. Apache Software Foundation, Apache Hadoop, Apache Software Foundation. (2018). <http://hadoop.apache.org/>.
7. Y. Hu, V. Y. Gunapati, P. Zhao, D. Gordon, N. R. Wheeler, M. A. Hossain, T. J. Peshek, L. S. Bruckman, G. Q. Zhang, and R. H. French, *IEEE J. Photovoltaics*, **7**, 230 (2017).
8. S. Ghemawat, H. Gobioff, and S.-T. Leung, in *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, ACM (2003) <http://doi.acm.org/10.1145/945445.945450>.
9. F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, in *Proceedings of the 7th Conference on Usenix Symposium on Operating Systems Design and Implementation - Volume 7*, p. 205 (2006).
10. J. Dean and S. Ghemawat, *Commun. ACM*, **51**, 107 (2008).
11. J. S. S. Lowndes, B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang, and B. S. Halpern, *Nat. Ecol. Evol.*, **1**, 160 (2017).
12. Y. Hu, M. A. Hosain, T. Jain, Y. R. Gunapati, L. Elkin, G. Q. Zhang, and R. H. French, in *2013 IEEE Energytech*, p. 1 (2013). doi:10.1109/EnergyTech.2013.6645317.
13. P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, J. G. Conde, *J. Biomed. Inf.*, **42**, 377 (2009).
14. R. Hoda, N. Salleh, and J. Grundy, *IEEE Software*, **35**, 58 (2018). doi:10.1109/MS.2018.29011318.
15. J. D. Blischak, E. R. Davenport, and G. Wilson, *PLoS Comput. Biol.*, **12**, e1004668 (2016).
16. V. Stodden, F. Leisch, and R. D. Peng, *Implementing Reproducible Research*, CRC Press, New York (2014).
17. J. L. Braid, A. J. Curran, J. Sun, E. J. Schneller, J. S. Fada, J. Liu, M. Wang, A. J. Longacre, J. Dai, B. D. Huey, K. O. Davis, J.-N. Jaubert, L. S. Bruckman, and R. H. French, in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion*, p. 1261, IEEE (2018).
18. A. G. Klinke, A. Gok, S. I. Ifeanyi, and L. S. Bruckman, *Polym. Degrad. Stab.*, **153**, 244 (2018).
19. M. A. Quintana, D. L. King, T. J. McMahon, and C. R. Osterwald, in *Conference Record of the Twenty-Ninth IEEE Photovoltaic Specialists Conference*, 2002., p. 1436, (2002). doi:10.1109/PVSC.2002.1190879.
20. W. Gambogi, Y. Heta, K. Hashimoto, J. Kopchick, T. Felder, S. MacMaster, A. Bradley, B. Hamzavtehrany, L. Garreau-Iles, T. Aoki, K. Stika, T. J. Trout, and T. Sample, *IEEE J. Photovoltaics*, **4**, 935 (2014). doi:10.1109/JPHOTOV.2014.2305472.
21. A. Gok, D. K. Ngendahimana, C. L. Fagerholm, R. H. French, J. Sun, and L. S. Bruckman, *PLoS ONE*, **12**, e0177614 (2017). doi:10.1371/journal.pone.0177614.

22. J. C. Hemminger, Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science | U.S. DOE Office of Science (SC), U.S. Department of Energy Basic Energy Sciences Advisory Committee, (2015). <http://science.energy.gov/bes/community-resources/reports/abstracts/#CFME> (accessed January 27, 2016).
23. D. A. Gordon, W.-H. Huang, D. M. Burns, R. H. French, and L. S. Bruckman, *PLoS ONE*, **13**, e0209016 (2018). doi:10.1371/journal.pone.0209016.
24. D. A. Gordon, Z. Zhan, and L. S. Bruckman, *Polym. Degrad. Stab.*, **161**, 85 (2019). <http://www.sciencedirect.com/science/article/pii/S0141391019300126>.
25. A. Fairbrother, S. Julien, K.-T. Wan, L. Ji, K. Boyce, S. Merzlic, A. Lefebvre, G. O'Brien, Y. Wang, L. Bruckman, R. French, M. Kempe, and X. Gu, in *Reliability of Photovoltaic Cells, Modules, Components, and Systems X*, vol. 10370, p. 1037003, International Society for Optics and Photonics (2017) doi:10.1117/12.2272488.
26. C. Ferrara and D. Philipp, *Energy Procedia*, **15**, 379 (2012). doi:10.1016/j.egypro.2012.02.046.
27. W.-H. Huang, N. Wheeler, A. Klinke, Y. Xu, W. Du, A. Gok, D. Gordon, Y. Wang, J. Liu, A. Curran, J. Fada, X. Ma, J. Braid, L. S. Bruckman, and R. H. French, netSEM: Network Structural Equation Modeling, (2018).
28. L.S. Bruckman, N.R. Wheeler, J. Ma, E. Wang, C.K. Wang, I. Chou, J. Sun, and R.H. French, *IEEE Access*, **1**, 384 (2013). doi:10.1109/ACCESS.2013.2267611.
29. R. H. Hoyle, Handbook of Structural Equation Modeling, 1st ed., The Guilford Press, New York (2012).
30. R.H. French, R. Podgornik, T.J. Peshek, L.S. Bruckman, Y. Xu, N.R. Wheeler, A. Gok, Y. Hu, M.A. Hossain, D.A. Gordon, P. Zhao, J. Sun, and G.-Q. Zhang, *Curr. Opin. Solid State Mater. Sci.*, **19**, 212 (2015).
31. A. Fairbrother, M. Boyd, Y. Lyu, J. Avenet, P. Illich, Y. Wang, M. Kempe, B. Dougherty, L. Bruckman, and X. Gu, *Sol. Energy*, **163**, 62 (2018).
32. M. Köntges, G. Oreski, U. Jahn, M. Herz, P. Hacke, K. Weiss, G. Razongles, M. Paggi, D. Parlevliet, T. Tanahashi, and R. H. French, Assessment of Photovoltaic Module Failures in the Field, IEA PVPS Task 13 Report, (2017).
33. U. Jahn, M. Herz, M. Köntges, D. Parlevliet, M. Paggi, I. Tsanakas, J. S. Stein, K. A. Berger, S. Ranta, R. H. French, M. Richter, T. Tanahashi, E. Ndrio, G. Mathiak, B. Weinreich, D. Hottenrott, J. S. Fada, B. Kubicek, C. Buerhop-Lutz, and V. Medici, Review on IR and EL Imaging for PV Field Applications, IEA-PVPS Task 13, 2018.
34. R. Dubey, S. Chattopadhyay, S. Zachariah, S. Rambabu, H. K. S.A. Kottantharayil, B.M. Arora, K.L. Narasimhan, N. Shiradkar, and J. Vasi, in *2018 IEEE 7th World Conference on Photovoltaic Energy*, p. 0098, IEEE (2018). doi:10.1109/PVSC.2018.8548080.
35. B. Doll, T. Pickel, O. Schreer, C. Zetzmann, J. Teubner, C. Buerhop, J. Hauch, C. Camus, and C. J. Brabec, in *Infrared Sensors, Devices, and Applications VIII*, vol. 10766, p. 107660K, International Society for Optics and Photonics (2018). doi:10.1117/12.2320518.
36. G.A. d R. Benatto, M. Chi, O.B. Jensen, A.A.S. Lancia, N. Riedel, B. Iandolo, R.S. Davidsen, O. Hansen, S. Thorsteinsson, and P.B. Poulsen, in *2018 IEEE 7th World Conference on Photovoltaic Energy*, p. 0395, IEEE (2018). doi:10.1109/PVSC.2018.8547416.
37. R. Bhoopathy, O. Kunz, M. Juhl, T. Trupke, and Z. Hameiri, *Prog. Photovoltaics*, **26**, 69 (2018). doi:10.1002/pip.2946.
38. M. Dhimish, V. Holmes, B. Mehrdadi, and M. Dales, *J. Sci.: Adv. Mater. Dev.*, **2**, 199 (2017). doi:10.1016/j.jsamd.2017.05.005.
39. A. Pozza and T. Sample, *Prog. Photovoltaics*, **24**, 368 (2016). doi:10.1002/pip.2717.
40. J. Schlothauer, S. Jungwirth, M. Köhl, and B. Röder, *Sol. Energy Mater. Sol. Cells*, **102**, 75 (2012). doi:10.1016/j.solmat.2012.03.022.
41. M. Kontges, S. Kurtz, C. Packard, U. Jahn, K. Berger, K. Kato, T. Friesen, H. Liu, M. Van Isehegam, IEA-PVPS Task 13: Review of Failures of PV Modules, Technical Report. (2014).
42. A. M. Karimi, J. S. Fada, J. Liu, J. L. Braid, M. Koyutürk and R. H. French, in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion*, p. 1261, IEEE (2018).
43. T. Potthoff, K. Bothe, U. Eitner, D. Hinken, and M. Köntges, *Prog. Photovoltaics*, **18**, 100 (2010).
44. A.S. Rajput, J.W. Ho, Y. Zhang, S. Nalluri, and A.G. Aberle, in *Light, Energy and the Environment 2018 (E2, FTS, HISE, SOLAR, SSL) (2018), paper OM3D.5*, p. OM3D.5, Optical Society of America (2018) doi:10.1364/OSE.2018.OM3D.5.
45. M. Wang, X. Ma, W.-H. Huang, J. Liu, A. J. Curran, E. Schnabel, M. Köhl, K. O. Davis, J. Brynjarsdóttir, J. L. Braid, and R. H. French, in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion*, p. 0778, IEEE (2018).
46. R. Margolis, C. Coggeshall, and J. Zuboy, "SunShot vision study." *US Dept. of Energy*, Washington, D.C. (2012).
47. J. Zhang, B.-M. Hodges, S. L. Hendrik, F. Hamman, B. Lehman, J. Simmons, E. Campos, V. Banunaryanan, and J. Tedesco, *Sol. Energy*, **122**, 804 (2015). doi:10.1016/j.solener.2015.09.047.
48. E.C. Kara, C.M. Roberts, M. Tabone, L. Alvarez, D.S. Callaway, and E.M. Stewart, *Sust. Energy, Grids Net.*, **13**, 112 (2018).
49. M. Lave, M. J. Reno, and R. J. Broderick, *Sol. Energy*, **118**, 327 (2015).
50. M. Lave, R. J. Broderick, and M. J. Reno, *Sol. Energy*, **151**, 119 (2017).
51. N. Savage, *Nature*, **558**, S19 (2018). doi:10.1038/d41586-018-05484-4.
52. S. Lu, X. Shao, M. Freitag, L. J. Klein, J. Renwick, F. J. Marianno, C. Albrecht, and H. F. Hamann, in *2016 IEEE International Conference on Big Data (Big Data)*, p. 2672 (2016). doi:10.1109/BigData.2016.7840910.
53. L. J. Klein, F. J. Marianno, C. M. Albrecht, M. Freitag, S. Lu, N. Hinds, X. Shao, S. Bermudez Rodriguez, and H. F. Hamann, in *2015 IEEE International Conference on Big Data (Big Data)*, p. 1290 (2015). doi:10.1109/BigData.2015.7363884.
54. D. Donoho, *J. Comput. Graph. Stat.*, **26**, 745 (2017).
55. D. Hughes, and R. H. French, Crafting a Minor to Produce T-Shaped Graduates, T-SUMMIT 2016: Transformational Approaches to Creating T-shaped Professionals, National Academies, Washington DC (2016).
56. Business Higher Education Forum, Creating a Minor in Applied Data Science | BHEF, The Business Higher Education Forum, 2016. <http://www.bhef.com/publications/creating-minor-applied-data-science> (accessed August 16, 2016).
57. R. D. Peng, *Science*, **334**, 1226 (2011).
58. *Nature*, **533**, 437 (2016).
59. S. Sonnenburg, M.L. Braun, C.S. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, K.-R. Müller, F. Pereira, C.E. Rasmussen, G. Rätsch, B. Schölkopf, A. Smola, P. Vincent, J. Weston, and R. Williamson, *J. Mach. Learn. Res.*, **8**, 2443 (2007).
60. E. Maxwell, *Innovations: Technol., Gov., Global.*, **1**, 119 (2006).
61. Will Holmgren, PVLib/PVLib-Python (2019), <https://github.com/pvlib/pvlib-python> (accessed December 11, 2016).
62. D. Moher, F. Naudet, I.A. Cristea, F. Miedema, J.P.A. Ioannidis, and S.N. Goodman, *PLoS Biol.*, **16**, e2004089 (2018). doi:10.1371/journal.pbio.2004089.
63. E. C. McKiernan, *PLoS Biol.*, **15**, e1002614 (2017).
64. B.A. Nosek, G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck, C.D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D.P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E.L. Paluck, U. Simonsohn, C. Soderberg, B.A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E.J. Wagenmakers, R. Wilson, and T. Yarkoni, *Science*, **348**, 1422 (2015). doi:10.1126/science.aab2374.