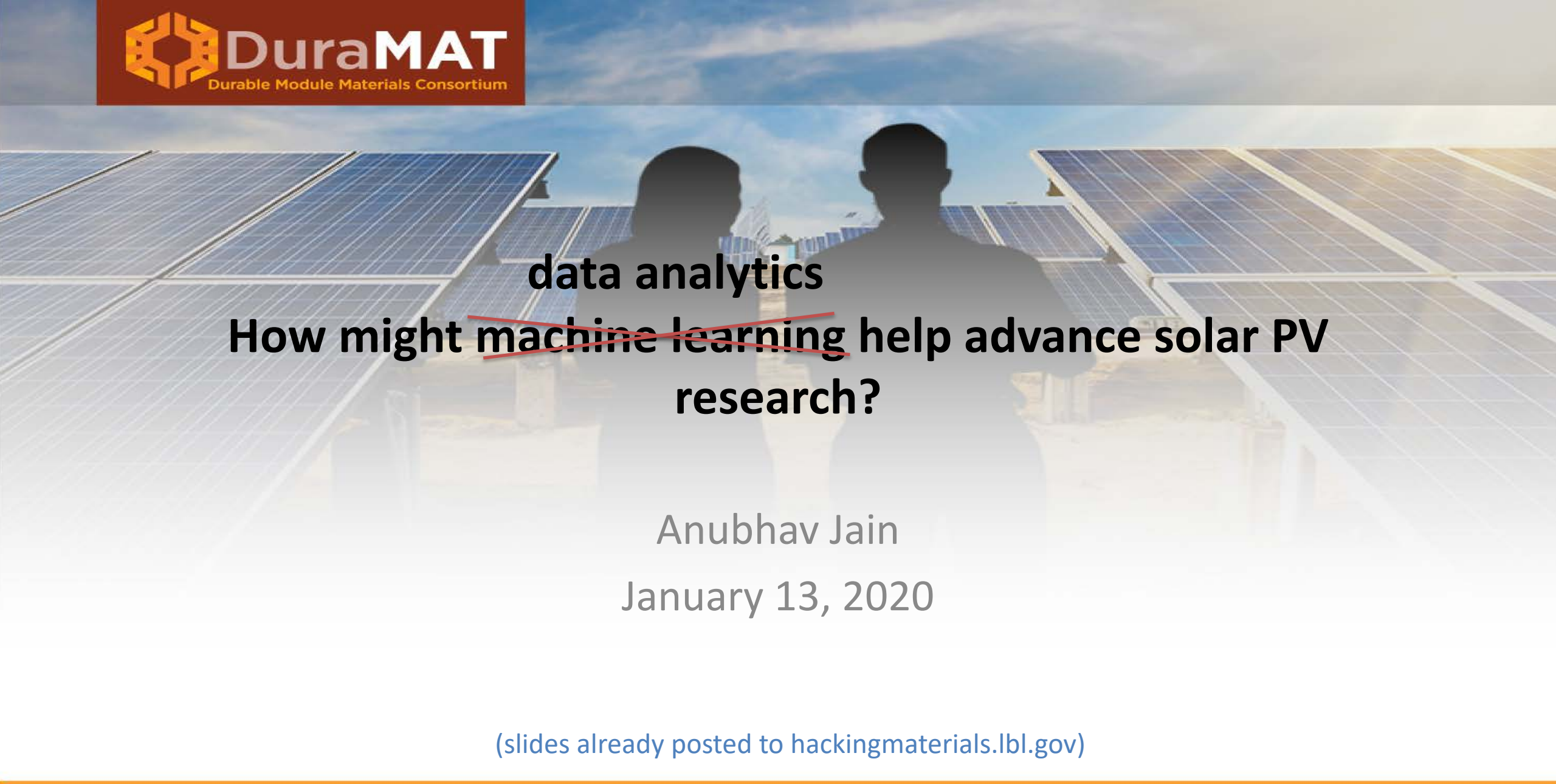


DuraMAT
Durable Module Materials Consortium

data analytics
How might ~~machine learning~~ help advance solar PV research?

Anubhav Jain

January 13, 2020

(slides already posted to hackingmaterials.lbl.gov)

Outline

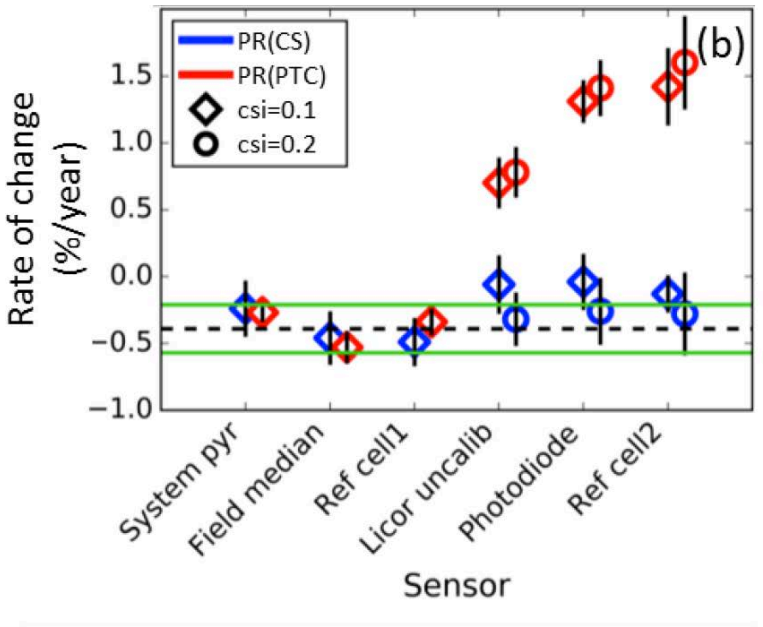
- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress)*
- Time series data
  - Extracting module parameters from production power data (*in progress*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques

Outline

- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress)*
- Time series data
  - Extracting module parameters from production power data (*in progress*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques

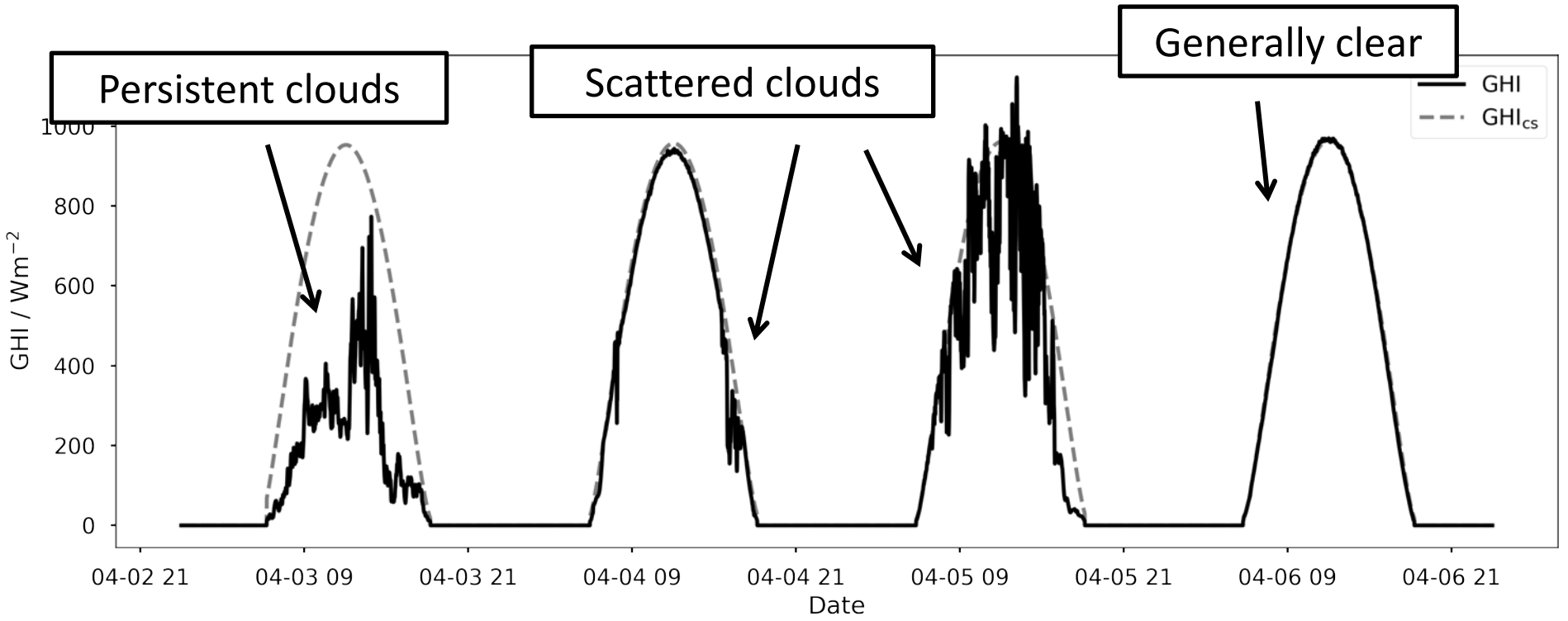
Motivation – data filtering affects analyses!

- Prior work has demonstrated that degradation rate calculations can be sensitive to the type of data filtering performed
- Restricting the data set to periods of clear sky results in more consistent and reliable fits of degradation rate
- But, how do we define a “clear sky” period?



Jordan, D. C., Deline, C., Kurtz, S. R., Kimball, G. M. & Anderson, M. Robust PV Degradation Methodology and Application. IEEE J. Photovoltaics 8, 525–531 (2018).

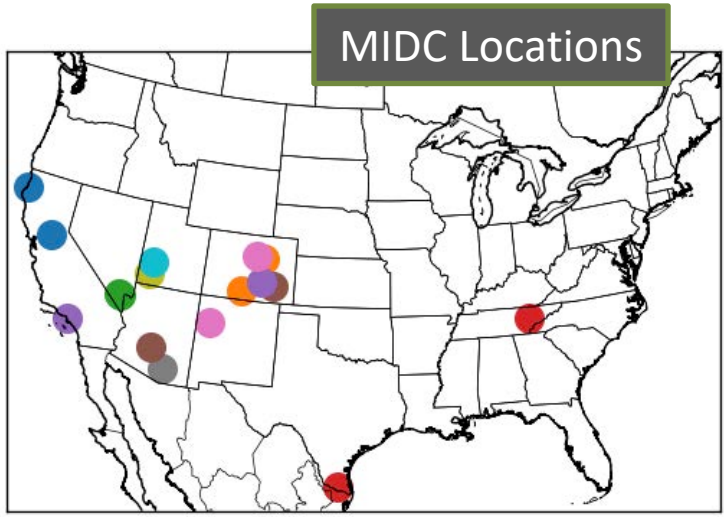
Identifying clear sky periods by examining irradiance plots



Identifying clear sky periods is relatively easy to do "by eye", at least approximately.
How about doing this automatically?

Data sets for fitting a clear sky model

- We need both input data (GHI measurements) and known output data (is the sky clear)
- We get GHI measurements from ground-based detectors in the MIDC network
- We get known clear sky labels from satellite measurements via the NSRDB database
- Some subtleties
  - difference in temporal (1 min vs 30 min) and spatial (on-site vs 4 km^2) resolutions of MIDC vs. NSRDB
  - big assumption: both data sets give a consistent picture of irradiance and sky clarity
    - Generally OK, but some data cleaning performed to remove clear violations of this assumption

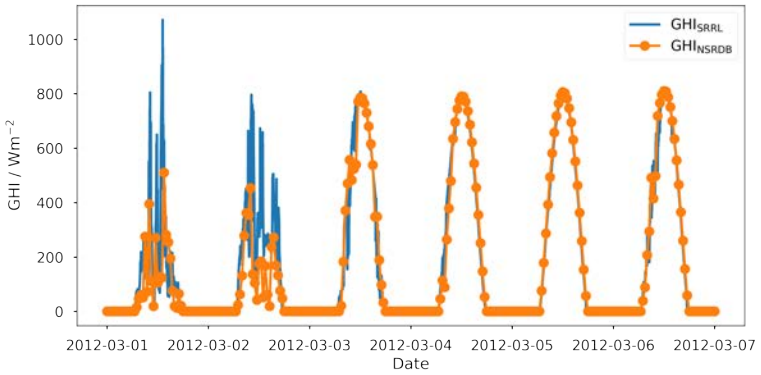


OAK RIDGE National Laboratory

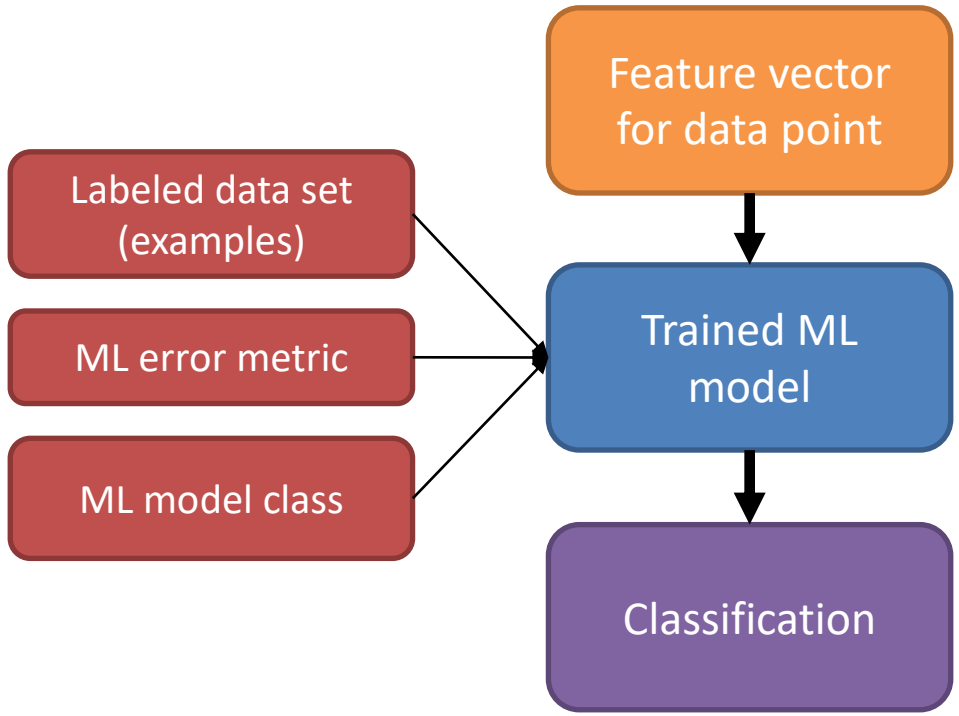
NREL NATIONAL RENEWABLE ENERGY LABORATORY

UTPA THE UNIVERSITY OF TEXAS-PAN AMERICAN

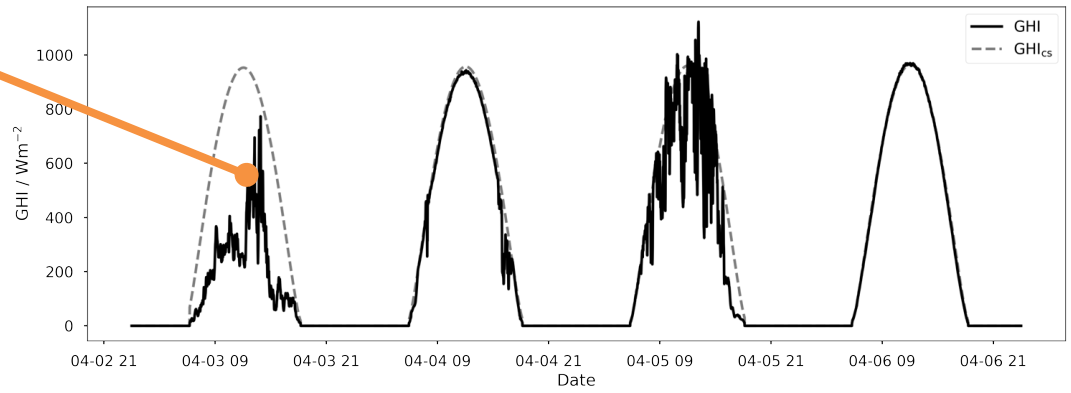
HUMBOLDT STATE UNIVERSITY



General procedure for using a labeled data set to develop a clear sky algorithm



Transform every data point into a vector of many "features" or "descriptors"

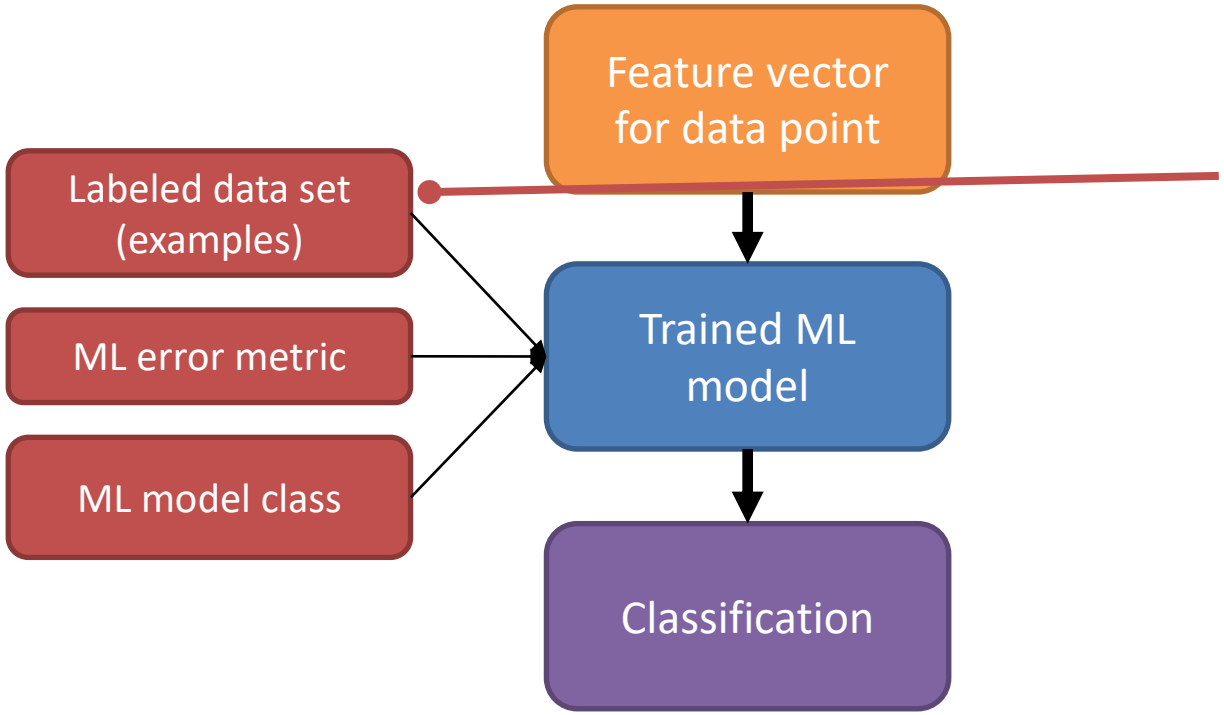


Within a certain "time window" around the target data point:

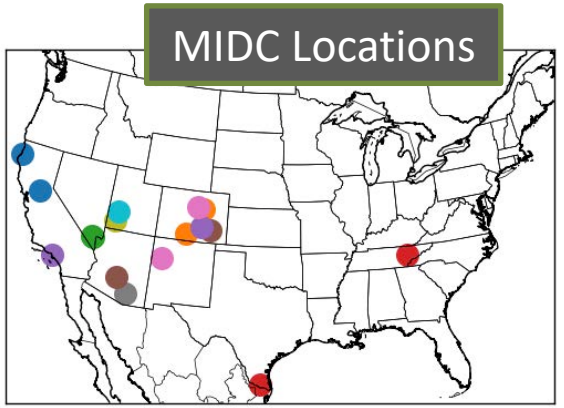
1. Difference of mean GHI and GHI_{CS}
2. Difference of maximum GHI and GHI_{CS}
3. Difference of line length of GHI vs. time curve and GHI_{CS}
4. Difference of standard deviation of slopes in GHI and GHI_{CS}
5. Maximum difference in slopes between GHI and GHI_{CS}

M. J. Reno and C. W. Hansen, "Identification of periods of clear sky irradiance in time series of GHI measurements," Renew. Energy, vol. 90, pp. 520–531, 2016.

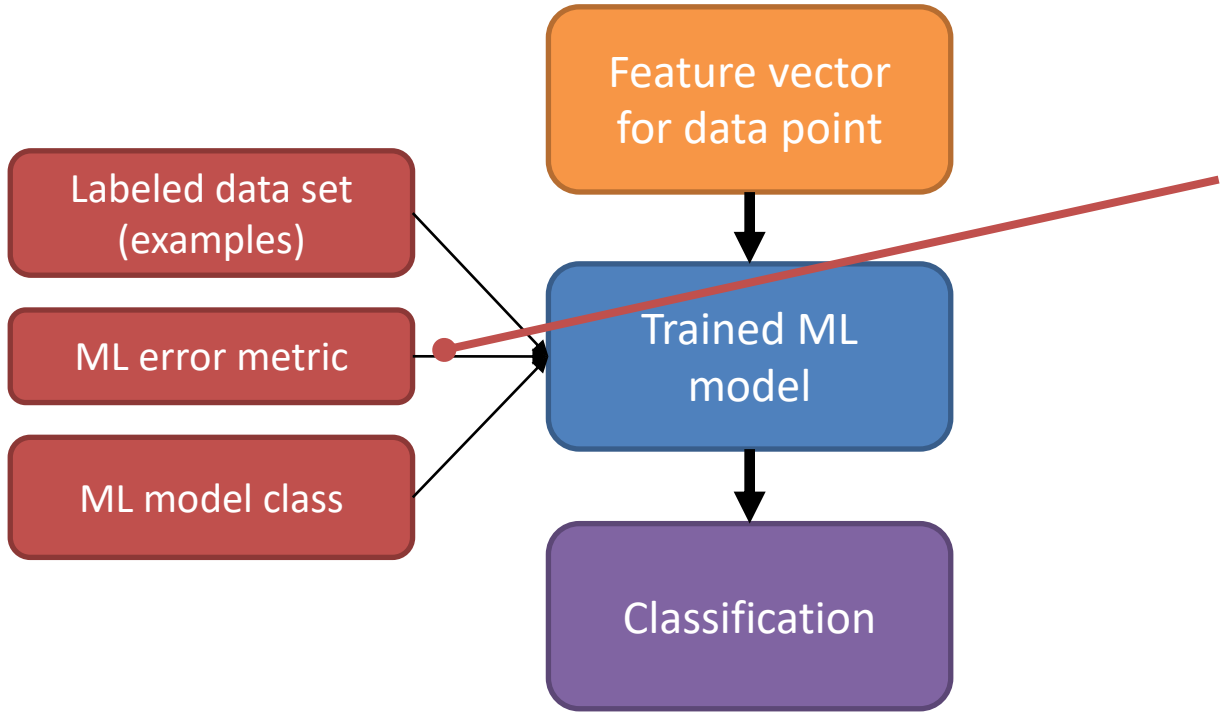
General procedure for using a labeled data set to develop a clear sky algorithm



We already described the labeled data set from NSRDB and MIDC



General procedure for using a labeled data set to develop a clear sky algorithm

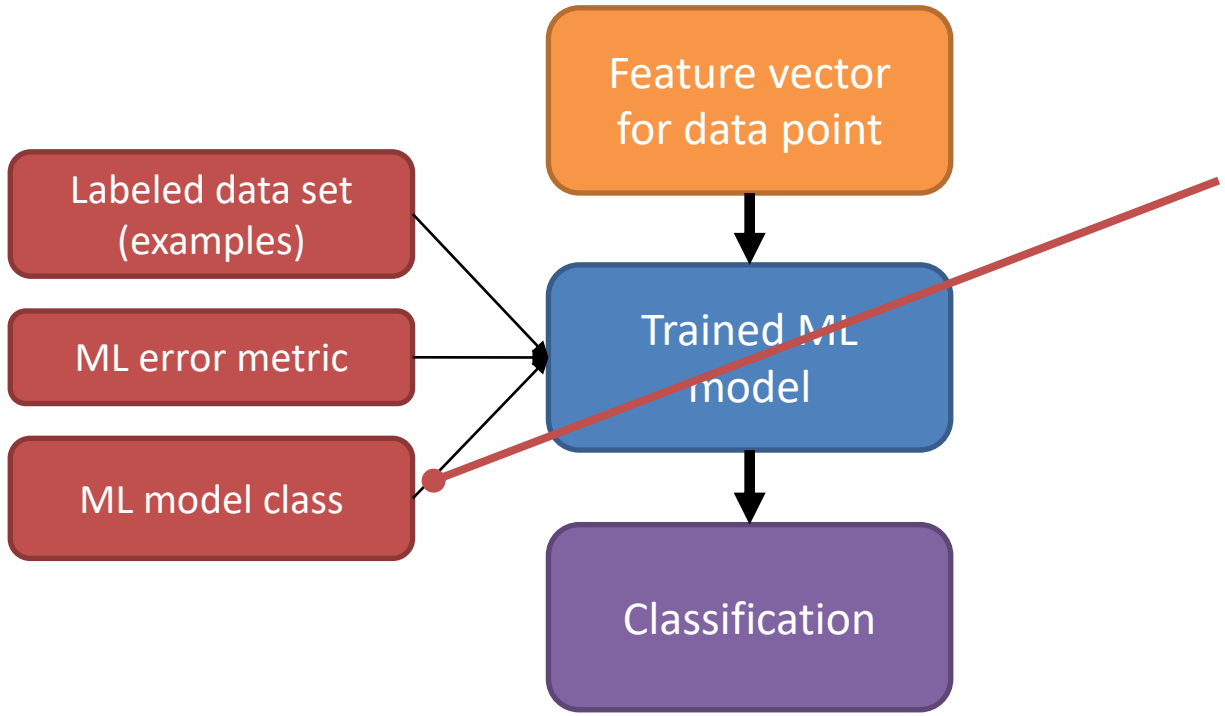


Score using the $F_{0.5}$ function

Leans towards high precision (filter out as many "unclear" points as possible at the expense of also filtering out some more "clear" data points)

$$F_\beta = (1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+\text{recall}}$$

General procedure for using a labeled data set to develop a clear sky algorithm



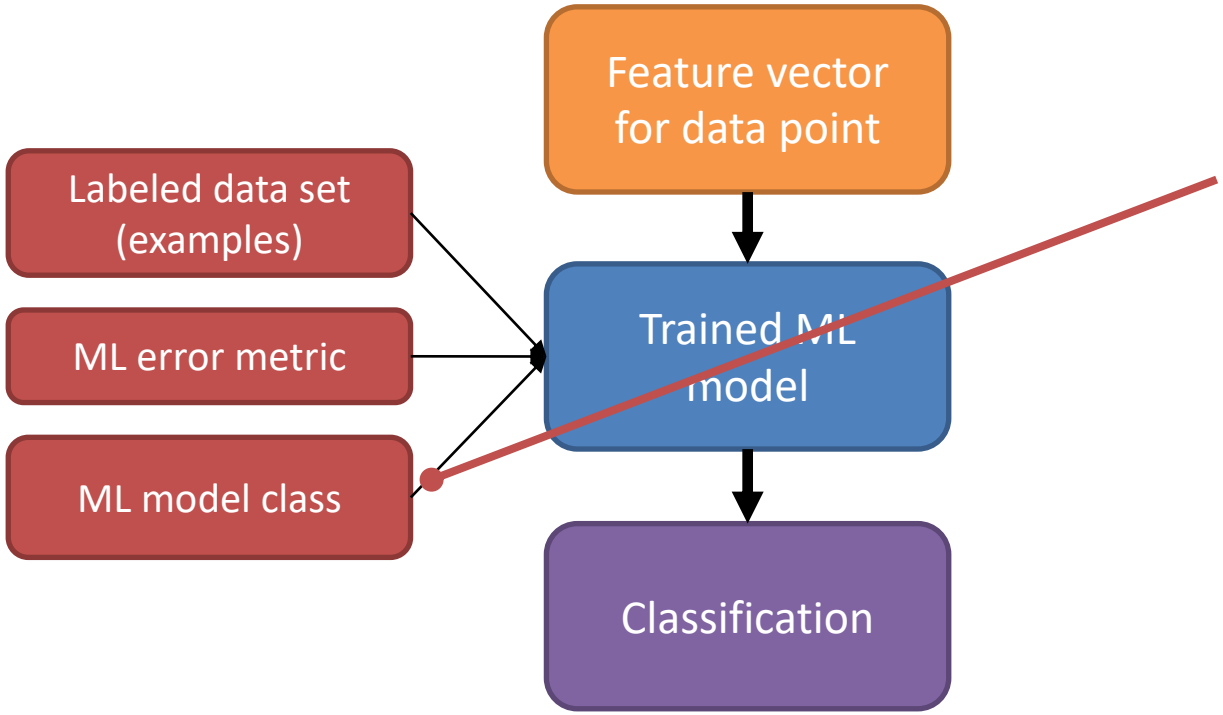
For the ML model class we found something interesting.

We first attempted “conventional” ML like random forest and regularized regression models.

These models gave good scores (both training and generalization), but visual inspection demonstrated some strange outliers from time to time.

We believe (but did not rigorously prove) that the ML models are too flexible, and started to fit on incorrect labels in the training data.

General procedure for using a labeled data set to develop a clear sky algorithm

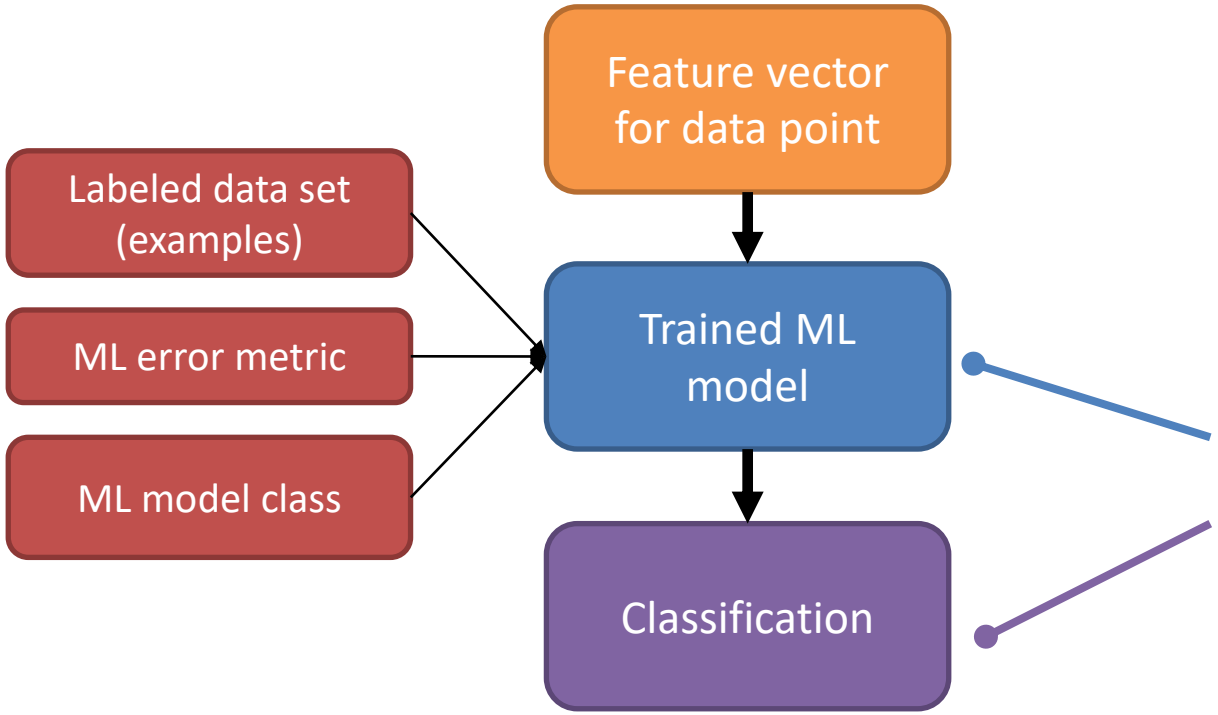


So we ended up ditching the ML models and just doing a traditional optimization to define static thresholds for the 5 features mentioned earlier.

This is just like the prior work of Reno and Hansen, except the thresholds were determined with a data set and not ”by eye”. So we are using the data as a substitute for manual tuning of thresholds.

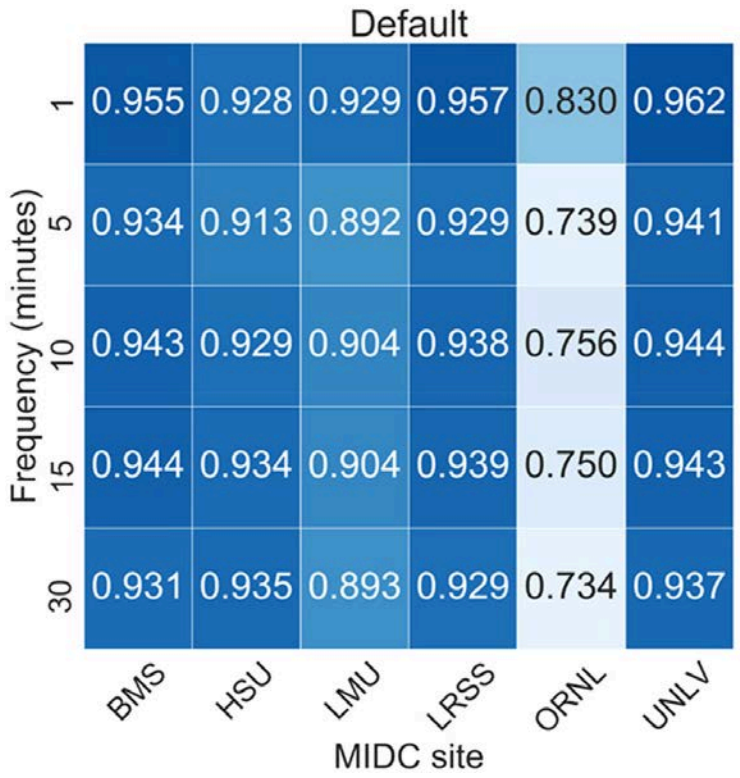
5-dimensional parameter optimization was done using the help of ML (Gaussian Process with Gradient-boosted decision trees were used to help determine which points in the space to test based on past results until hitting convergence).

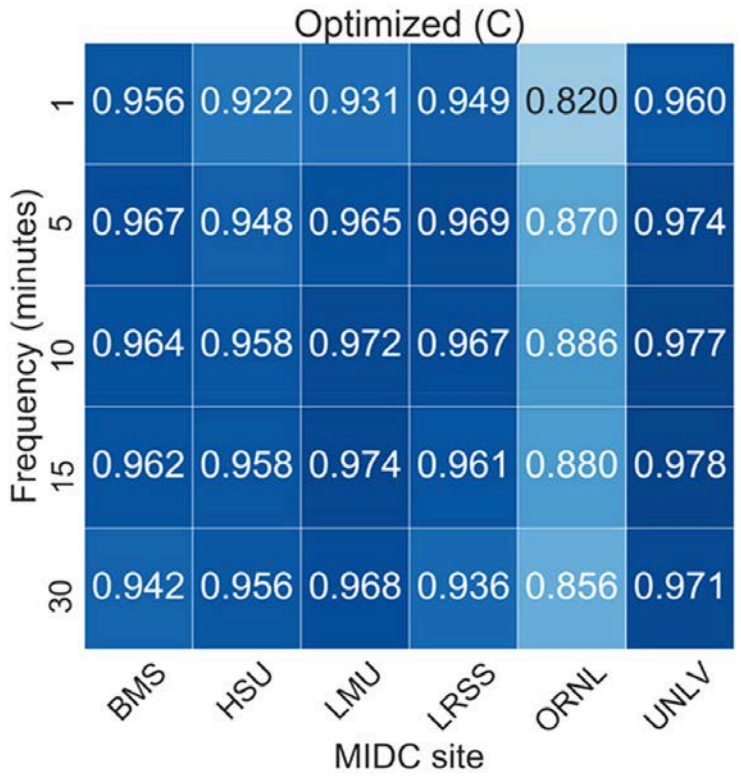
General procedure for using a labeled data set to develop a clear sky algorithm



With the trained model, we could now classify all our data points – checking against our data set as well as visual inspection of results

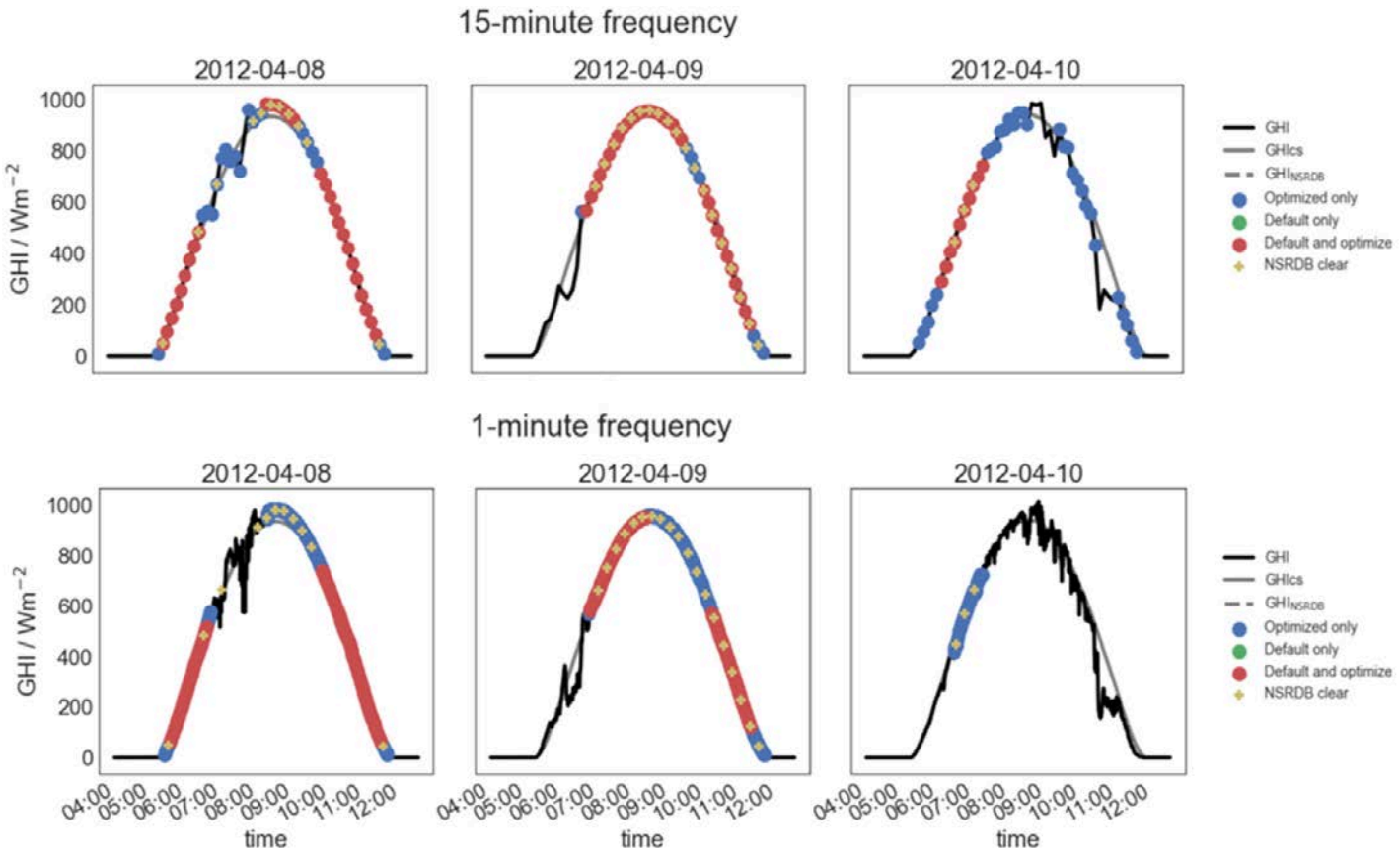
Results – optimized parameters give better scores on cross-validation





	Default	Optimized
Mean difference	75	79.144
Max. difference	75	59.152
Lower line length	-5	-41.416
Upper line length	10	77.789
Std. dev. of slopes	0.005	0.00745
Max diff. of slopes	8	68.579

Results – optimized parameters show visual differences in data quality



Blue points are all points that would ***not*** have been considered clear, for one reason or the other, using the default PVLIB implementation of the Reno & Hansen method!

Conclusions - clear sky classification

- The use of machine learning and data-driven methods helped us generate a better clear sky model
  - But when there are problems in the data set, there are also problems in the ML models – so you need to be careful and also check results visually
  - The model also assumes that irradiance will match clear sky models; other approaches are possible (e.g., see B. Meyers “statistical clear sky”)

- The updated thresholds have been contributed to PVLib-Python

- A manuscript with full details is published in JPV

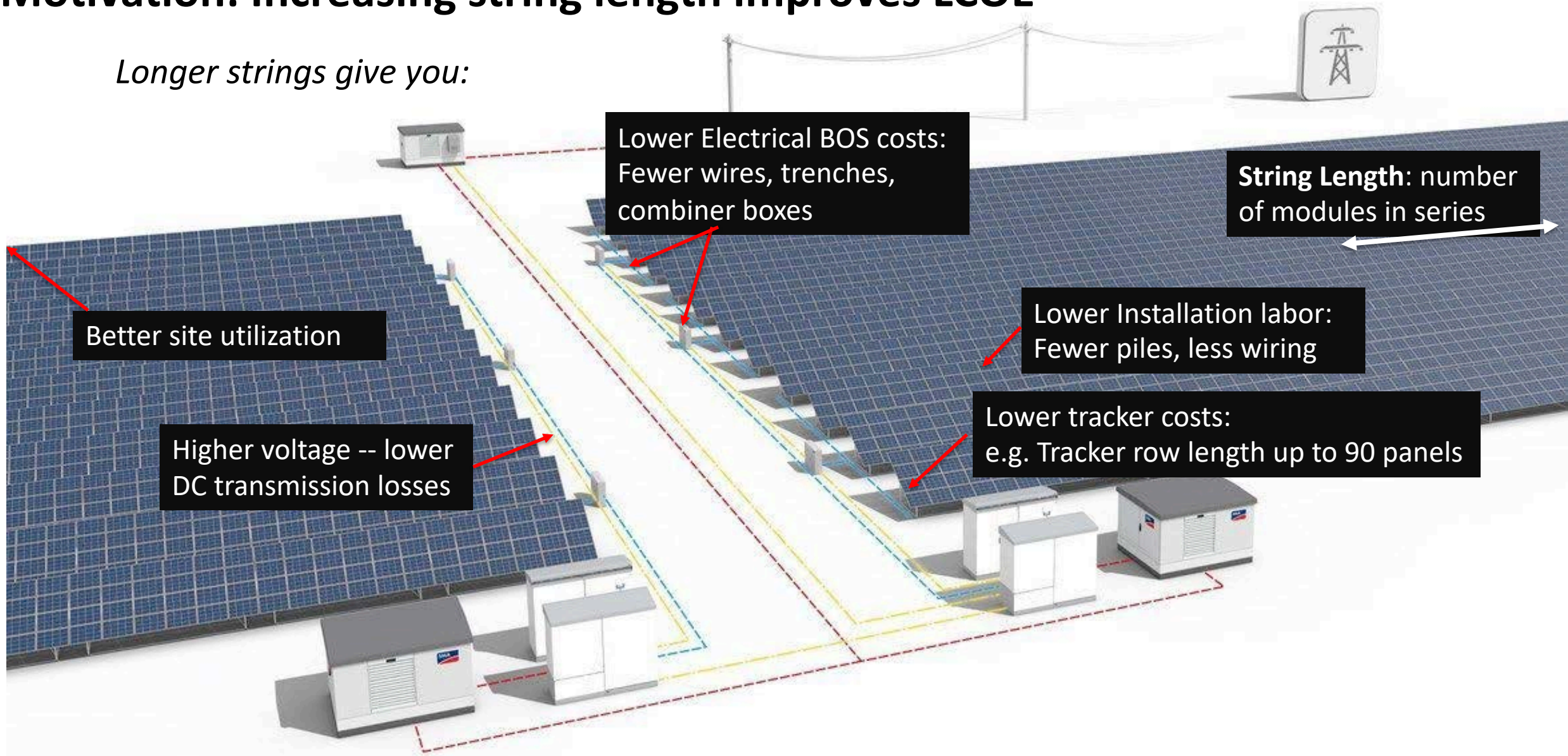
Ellis, B. H., Deceglie, M. & Jain, A. Automatic Detection of Clear-Sky Periods From Irradiance Data. IEEE Journal of Photovoltaics 998–1005 (2019). doi:10.1109/JPHOTOV.2019.2914444

Outline

- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress…)*
- Time series data
  - Extracting module parameters from production power data (*in progress…*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress…*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques (*ideas only!*)

Motivation: Increasing string length improves LCOE

Longer strings give you:

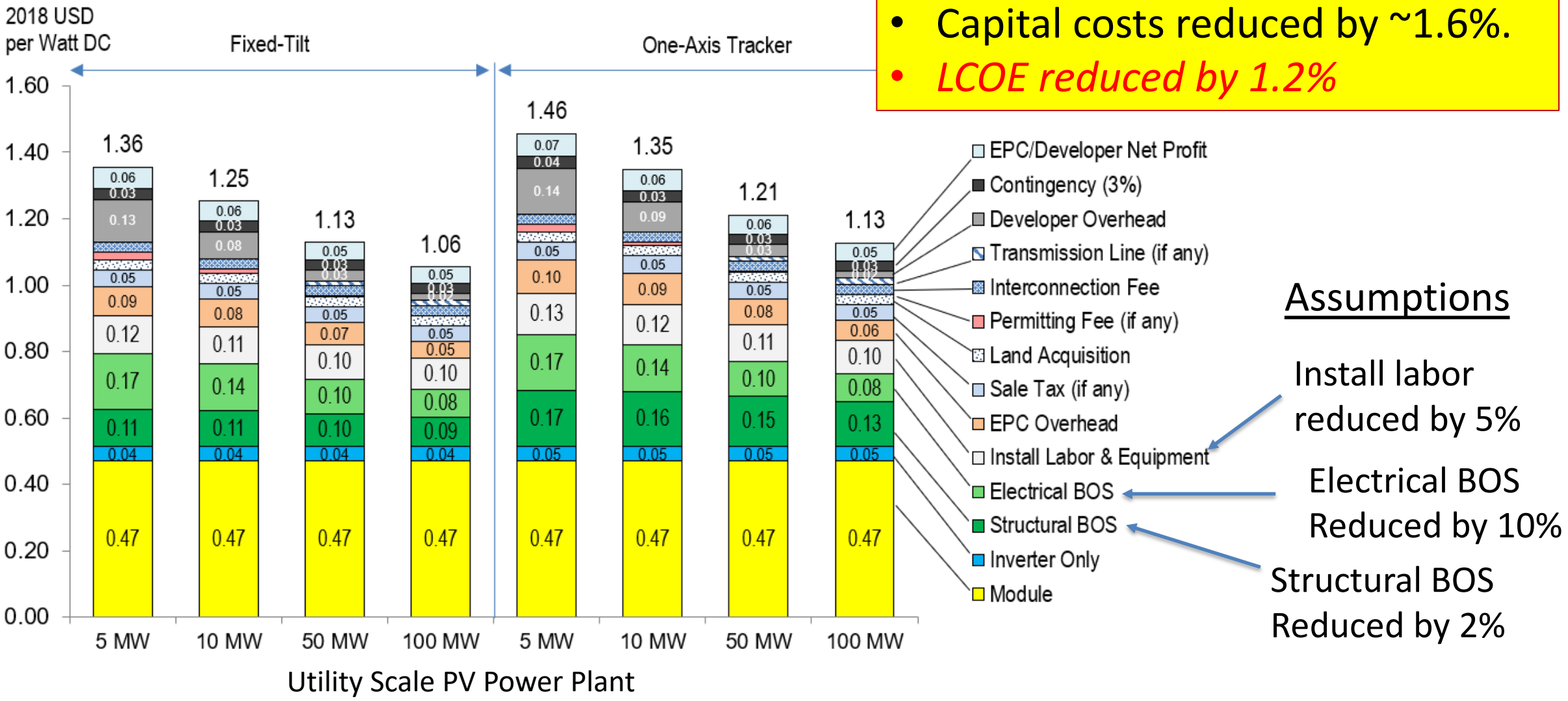


https://www.sma-america.com/industrial-systems/pv-power-plants.html
https://www.nextracker.com/2016/11/nextracker-achieves-1-solar-tracker-global-market-share-according-to-gtm-research/

LCOE Impact Estimate

Increase string length by 10%:

- Capital costs reduced by ~1.6%.
- *LCOE reduced by 1.2%*



Assumptions

Install labor reduced by 5%

Electrical BOS Reduced by 10%

Structural BOS Reduced by 2%

Fu, Ran, David Feldman, and Robert Margolis. 2018. ***U.S. Solar Photovoltaic System Cost Benchmark: Q1 2018***. Golden, CO: National Renewable Energy Laboratory. NREL/TP-6A20-72399. https://www.nrel.gov/docs/fy19osti/72399.pdf

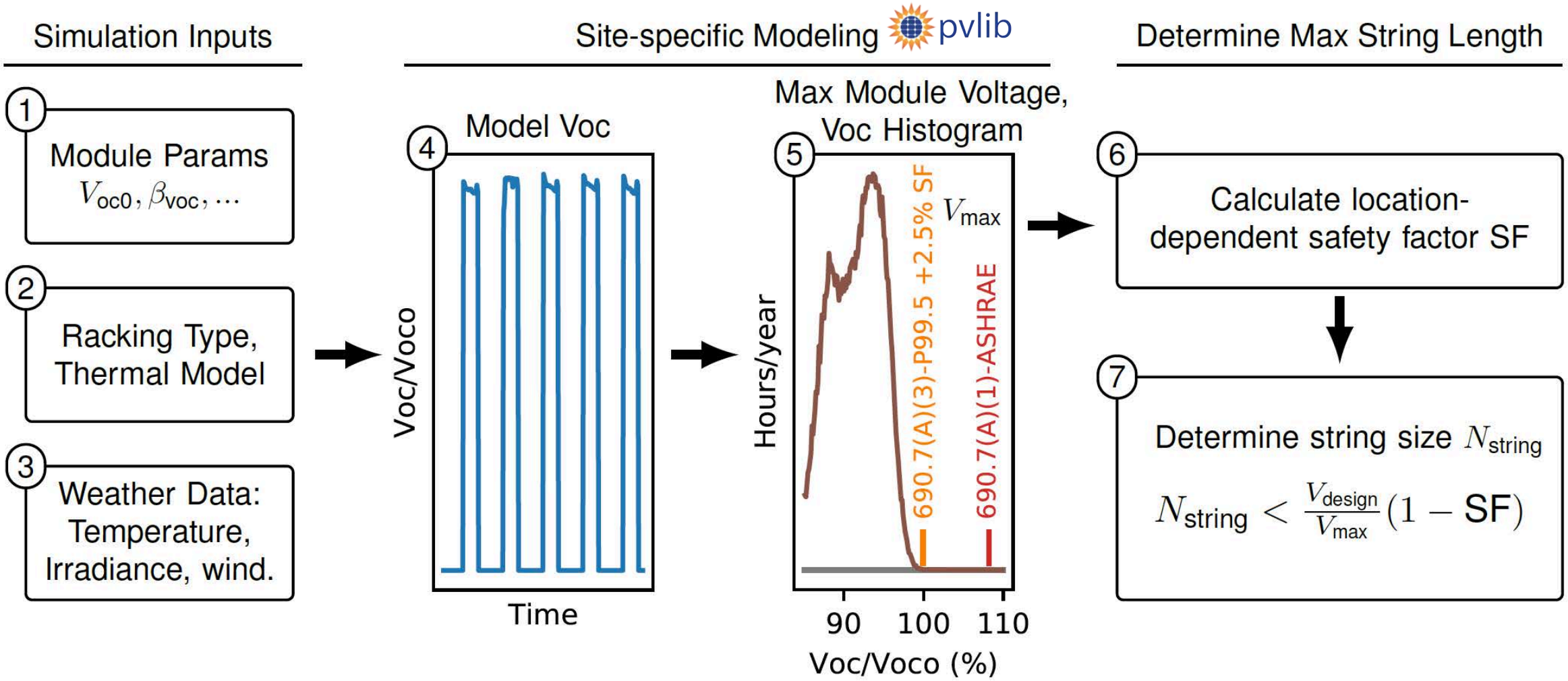
NEC 2017: Multiple valid methods for determining the string length

- 690.7(A)(1) Instruction in listing or labeling of module: The sum of the PV module-rated open-circuit voltage of the series-connected modules corrected for the lowest expected ambient temperature using the open-circuit voltage temperature coefficients in accordance with the instructions included in the listing or labeling of the module.
- 690.7(A)(3) PV systems of 100 kW or larger: For PV systems with a generating capacity of 100 kW or greater, a documented and stamped PV system design, using an industry standard method and provided by a licensed professional electrical engineer, shall be permitted.
  - *Informational Note:* One industry standard method for calculating voltage of a PV system is published by Sandia National Laboratories, reference SAND 2004-3535, Photovoltaic Array Performance Model

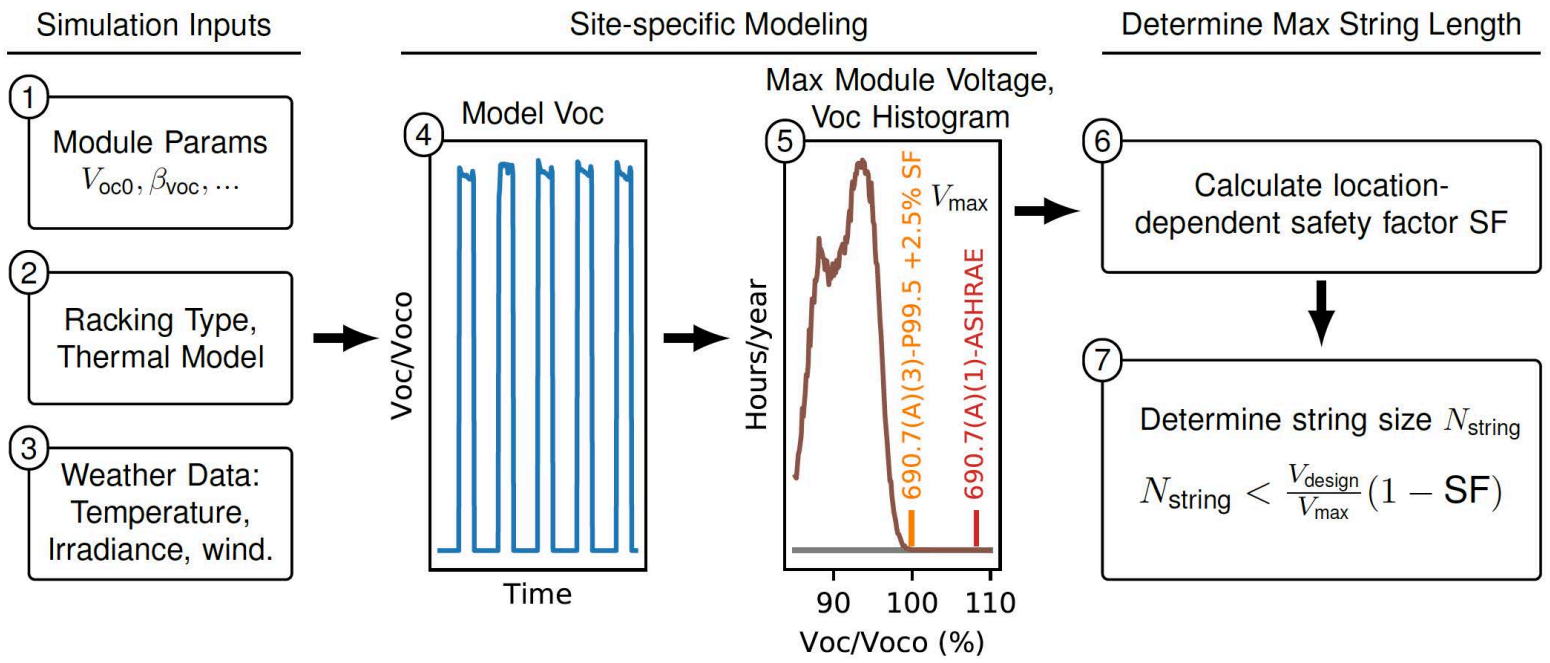
Method 1 (Traditional) uses lowest expected temperature and 1000 W/m^2.

Method 2 (Site-specific modeling) models system Voc over time

Site-specific String Length Determination. NEC 2017 690.7(A)(3) Compliant



Some AHJs accept this, some don't. *We want to make this the standard!*



Does site-specific modeling work?

Field data validates modeling method

- Used data from mobile performance and energy rating testbed (mPERT). 2014.
- 33 modules across 3 locations with integrated I-V tracers.
- 1-2 years Voc data at 5-10 minute intervals.

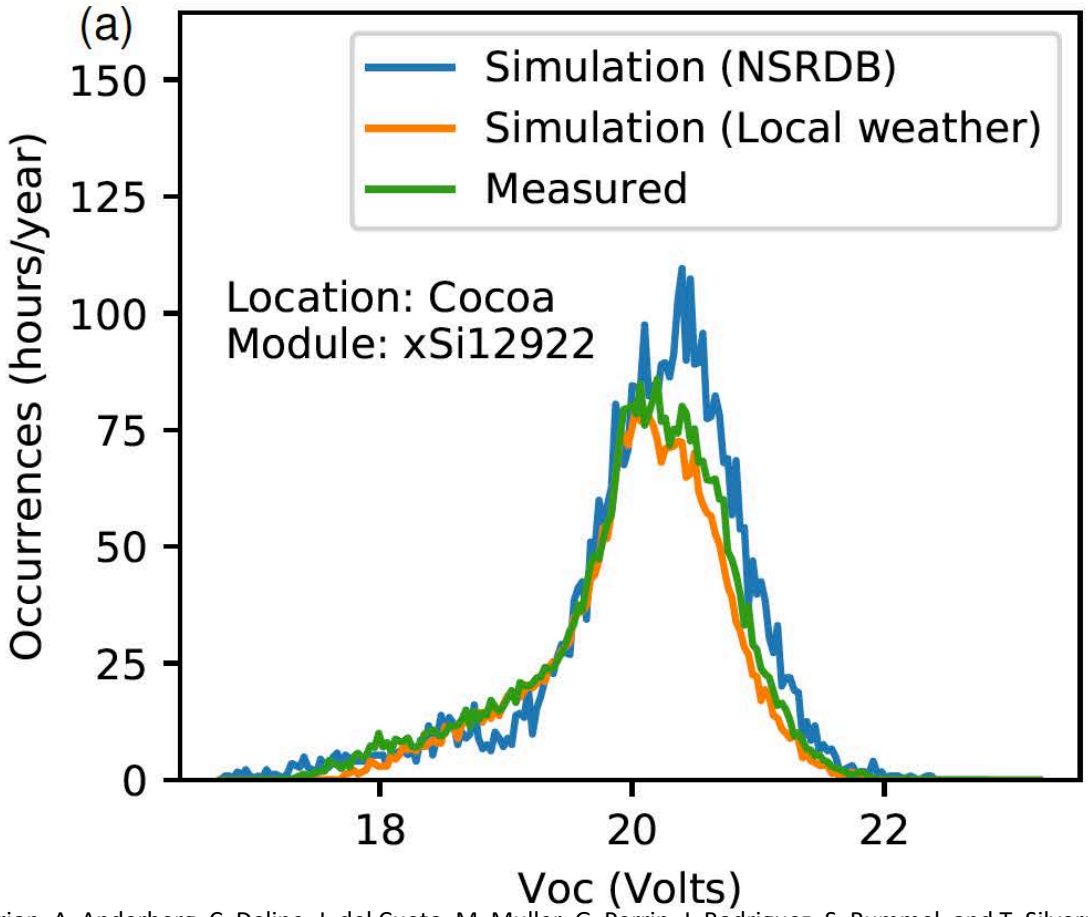
NREL PV Module Identifier	Technology	Manufacturer/ Model	File Names
xSi12922	Single-crystalline silicon	Manufacturer 1 Model A	Cocoa_xSi12922.csv Eugene_xSi12922.csv
xSi11246	Single-crystalline silicon	Manufacturer 1 Model A	Golden_xSi11246.csv
mSi460A8	Multi-crystalline silicon	Manufacturer 1 Model B	Cocoa_mSi460A8.csv Eugene_mSi460A8.csv
mSi460BB	Multi-crystalline silicon	Manufacturer 1 Model B	Golden_mSi460BB.csv
mSi0166	Multi-crystalline silicon	Manufacturer 2 Model C	Cocoa_mSi0166.csv Eugene_mSi0166.csv
mSi0188	Multi-crystalline silicon	Manufacturer 2 Model C	Cocoa_mSi0188.csv Eugene_mSi0188.csv
mSi0247	Multi-crystalline silicon	Manufacturer 2 Model C	Golden_mSi0247.csv
mSi0251	Multi-crystalline silicon	Manufacturer 2 Model C	Golden_mSi0251.csv
CdTe75638	Cadmium telluride	Manufacturer 3 Model D	Cocoa_CdTe75638.csv Eugene_CdTe75638.csv
CdTe75669	Cadmium telluride	Manufacturer 3 Model D	Golden_CdTe75669.csv
CIGS39017	Copper indium gallium selenide	Manufacturer 4 Model E	Cocoa_CIGS39017.csv Eugene_CIGS39017.csv
CIGS39013	Copper indium gallium selenide	Manufacturer 4 Model E	Golden_CIGS39013.csv
CIGS8-001	Copper indium gallium selenide	Manufacturer 5 Model F	Cocoa_CIGS8-001.csv Eugene_CIGS8-001.csv

W. Marion, A. Anderberg, C. Deline, J. del Cueto, M. Muller, G. Perrin, J. Rodriguez, S. Rummel, and T. Silverman, "User's manual for data for validating models for pv module performance," National Renewable Energy Laboratory, Golden, CO, Tech. Rep., 2014

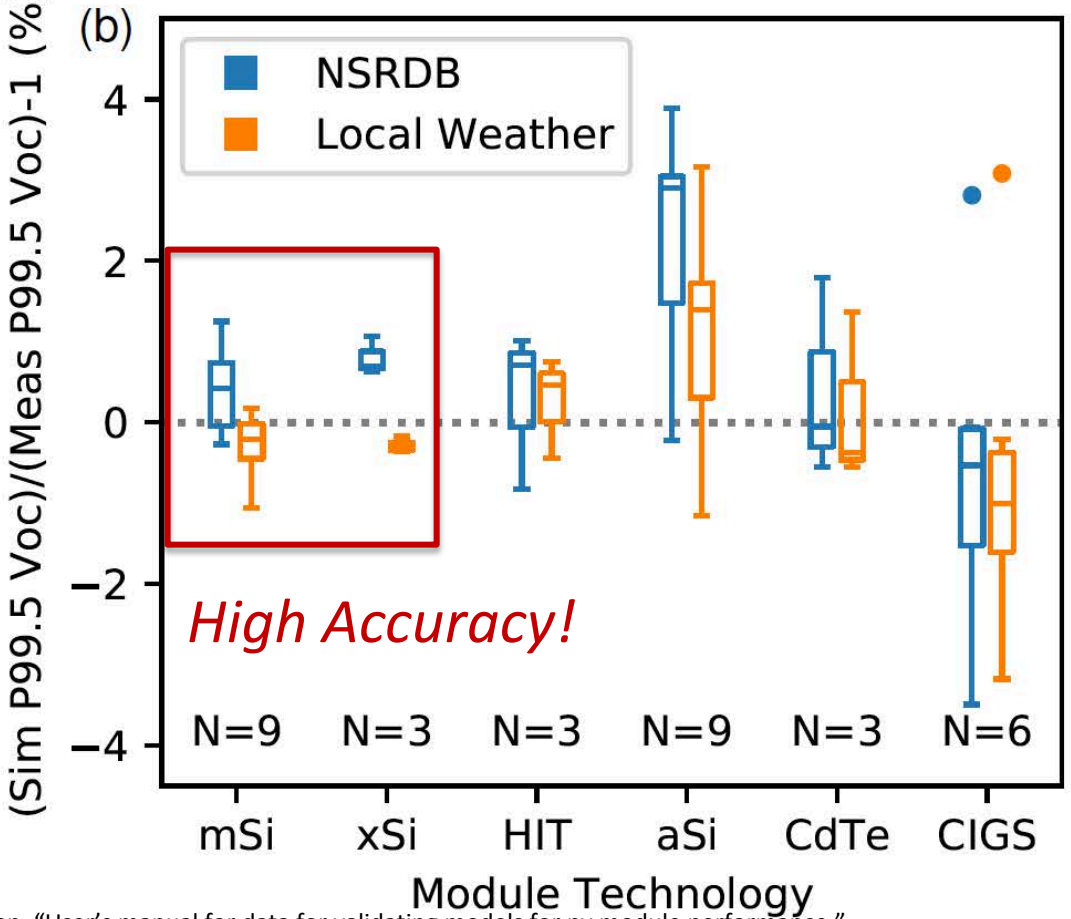
Simulated data agrees with field measurements to 1.5%

Mobile performance and energy rating testbed (mPERT),

Histogram of measured vs. modeled Voc



P99.5 Voc

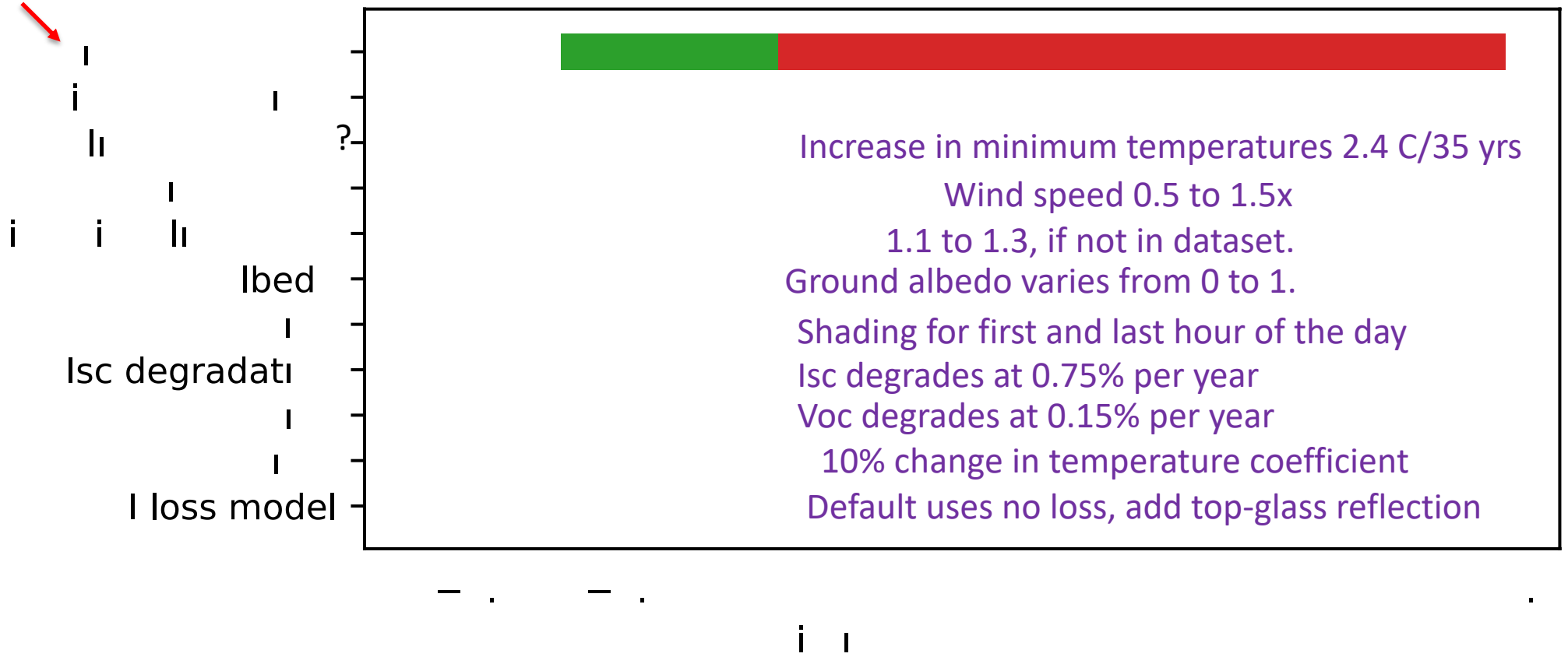


W. Marion, A. Anderberg, C. Deline, J. del Cueto, M. Muller, G. Perrin, J. Rodriguez, S. Rummel, and T. Silverman, "User's manual for data for validating models for pv module performance," National Renewable Energy Laboratory, Golden, CO, Tech. Rep., 2014

Adding in safety factors using an uncertainty analysis

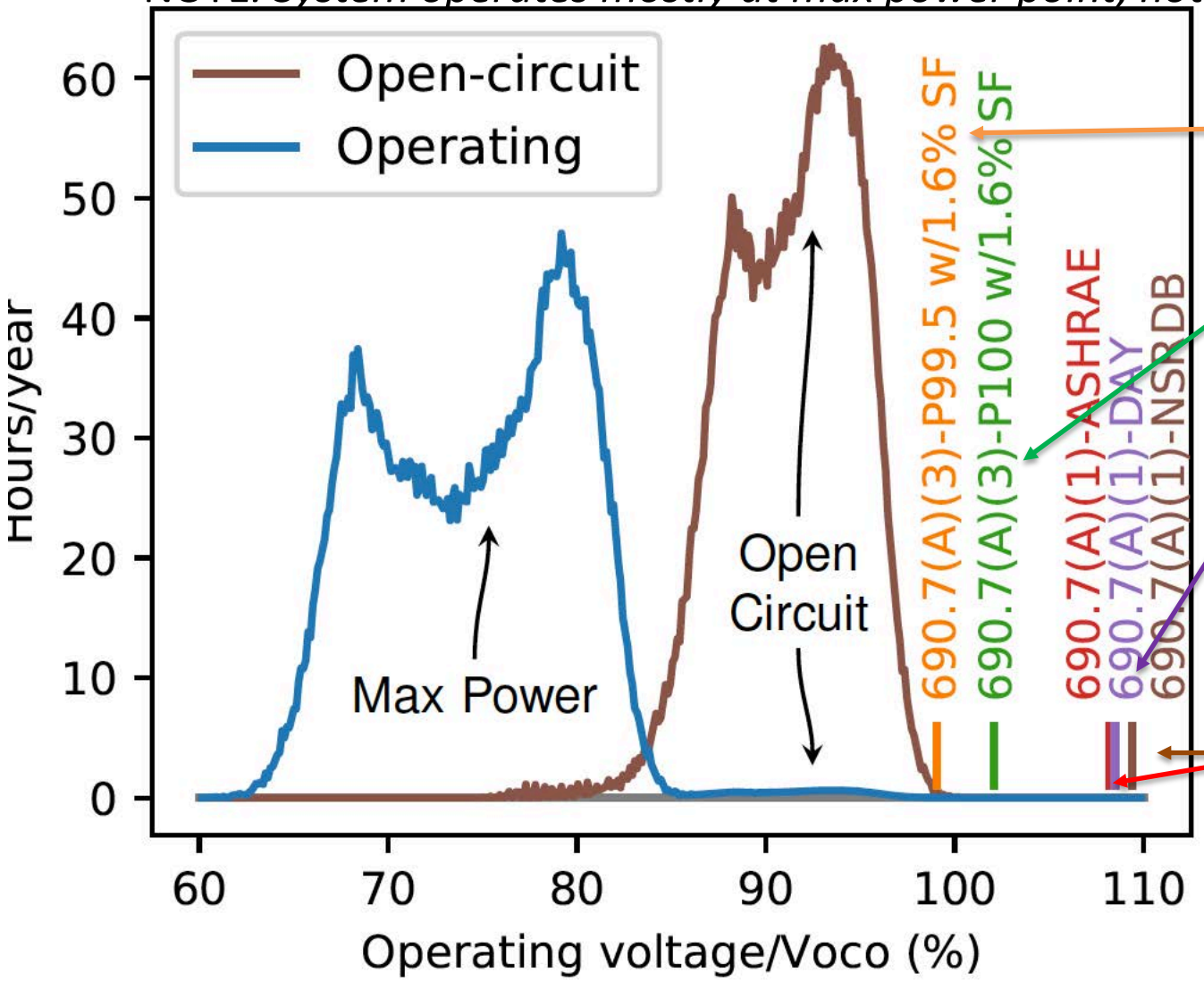
Air temperature uncertainty due to use of NSRDB - location dependent

Variation in Voc due to manufacturing inhomogeneity



And remember, you will be at V_{MPP} most of the time, not V_{oc} …

NOTE: *System operates mostly at max power point, not open-circuit!*



P99.5: Use site-specific modeling of Voc over 18 years, find 99.5 percentile Voc

Hist: Use site-specific modeling of Voc over 18 years, find highest ever predicted Voc

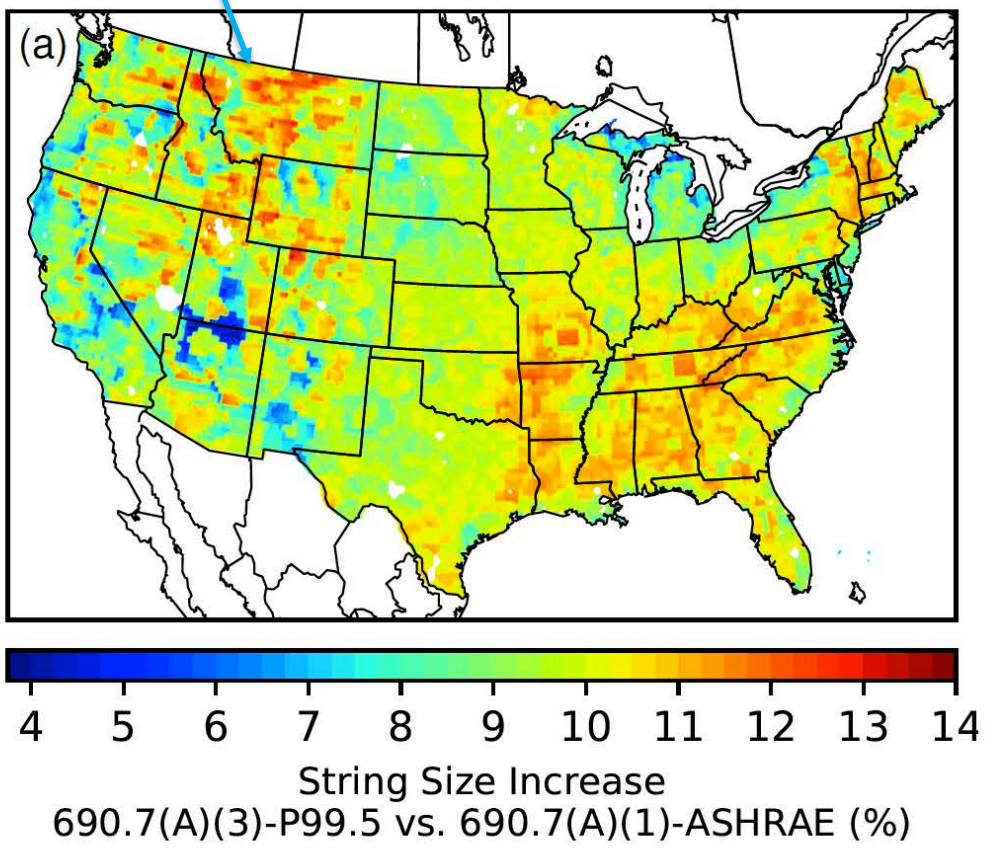
Day: uses 1000 W/m^2 and extreme annual mean minimum design dry bulb temperature *during daytime* (GHI>150 W/m^2)

Trad: uses 1000 W/m^2 and extreme annual mean minimum design dry bulb temperature. *Unnecessarily conservative.*

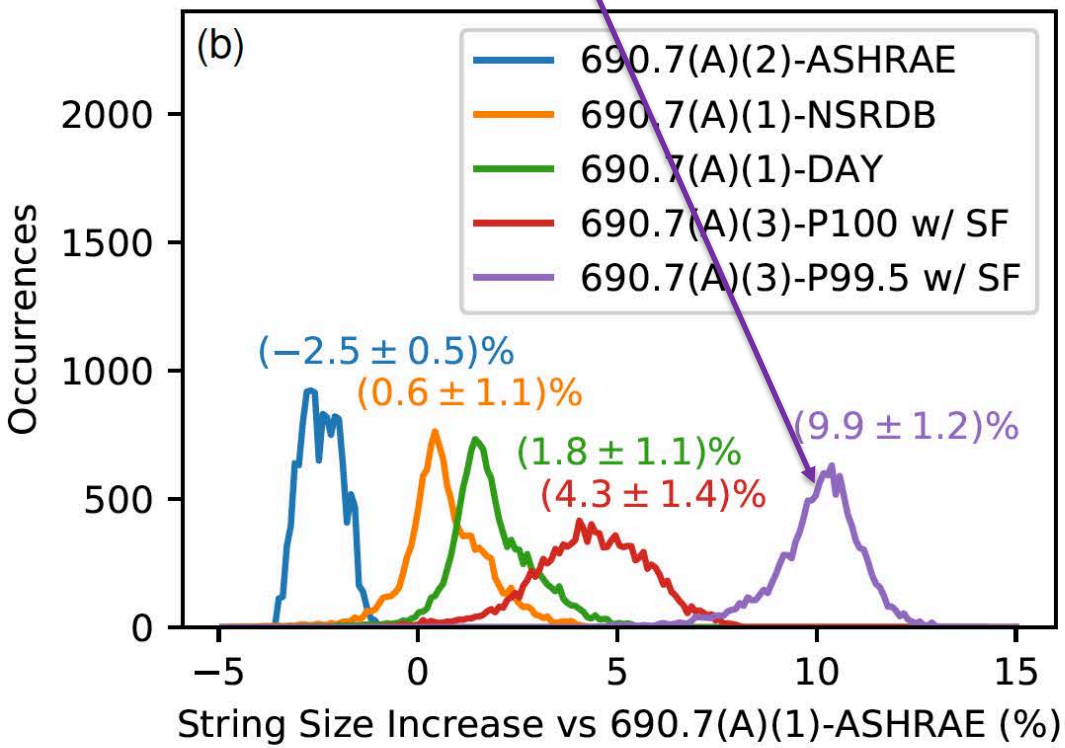
Add safety factor based on sensitivity analysis.

How much improvement from using site-specific modeling?

Higher improvement in mountainous regions: extreme cold during nighttime



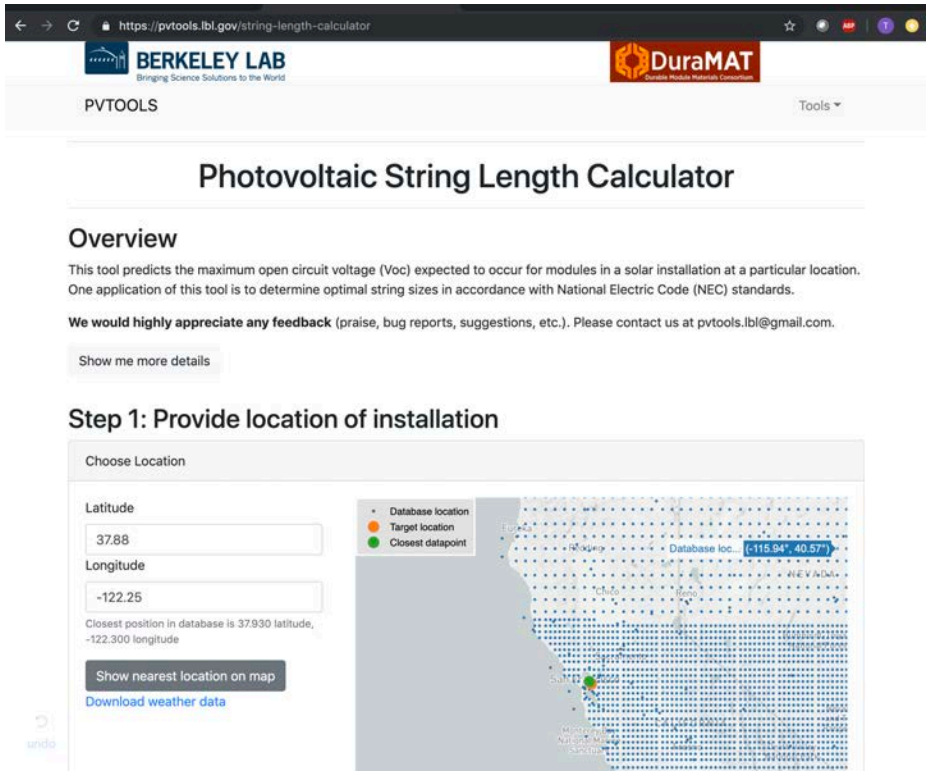
~10% Longer strings are acceptable using site-specific modeling compared to traditional!



Simulation: Fixed-tilt, south at latitude tilt, Voc temperature coefficient -0.35%/C, n_diode = 1.2

Conclusion

- Using site-specific Voc modeling, string lengths can be increased by ~10% compared to traditional method.
- Potentially reduces LCOE by ~1.2% *just by reorganizing strings.*
- NEC-2017 compliant method (trying to get explicit footnote into NEC guidelines)
- Site-specific string length design is now easy for anyone to perform using a simple web tool.
- Method available as open-source python module.
- Bifacial modeling also possible (simple method on web, more accurate modeling in Python code)



https://github.com/toddkarin/vocmax

https://pvtools.lbl.gov

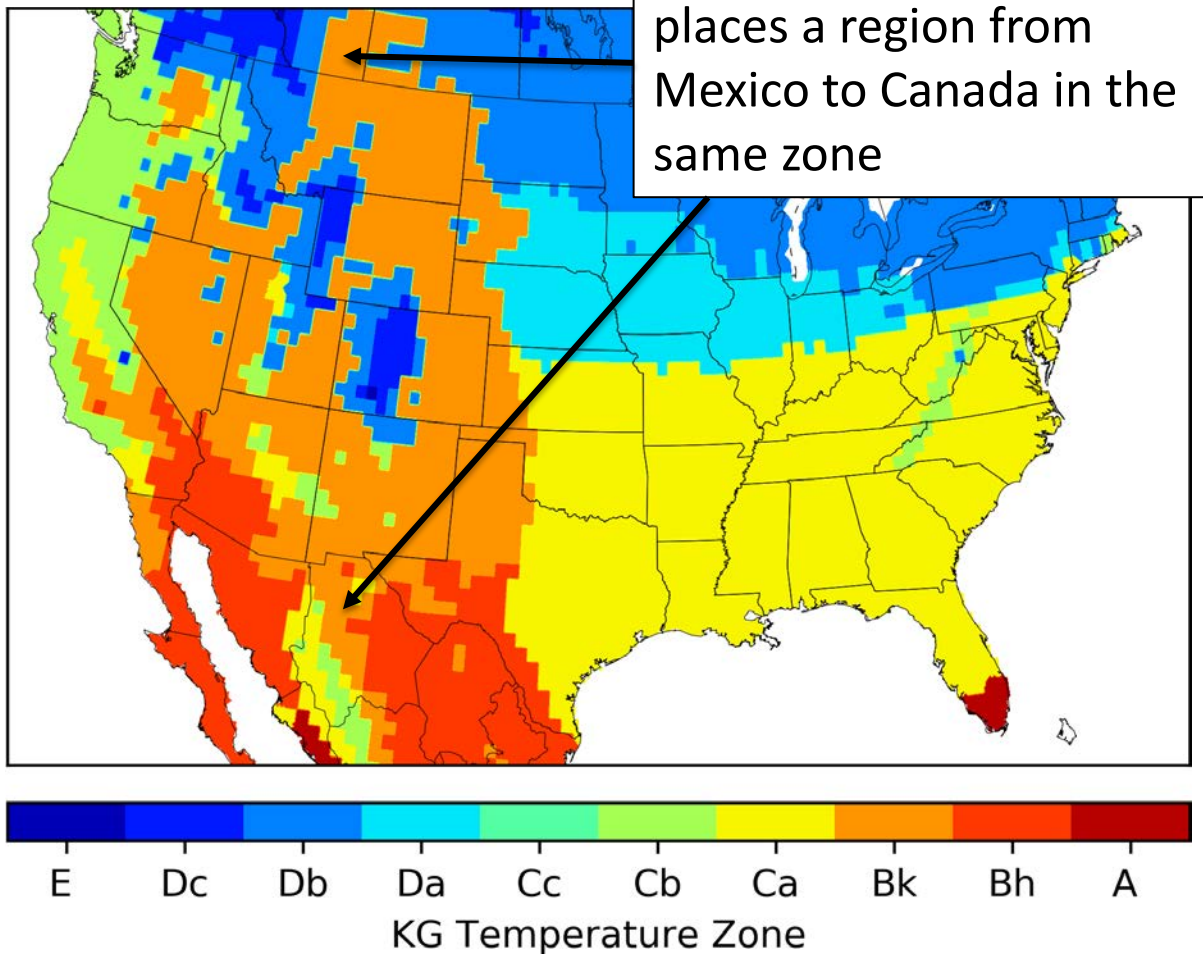
Karin & Jain, "Photovoltaic String Sizing using Site-Specific Modeling", accepted for publication, IEEE Journal of Photovoltaics

Outline

- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress)*
- Time series data
  - Extracting module parameters from production power data (*in progress*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques

Goal

- We want a scheme to determine which geographical locations are likely to see similar types and magnitudes of PV degradation
- This will help move us away from a uniform degradation rate estimation
- Can also help design climate—specific protocols for testing PV modules



Methods

- **Dataset:** PV stressors calculated using NASA global land data assimilation system (GLDAS), incorporating ground and surface measurements. 01/2010 – 01/2019.
- Module temperature calculated using PVLIB, open-rack polymer-back and roof-mount glass-back, GHI = POA.

Stressors:

- Arrhenius weighted Equivalent module temperature (T_{eq})

Activation energy 1.1 eV

$$\exp\left(-\frac{E_a}{k_B T_{\mathrm{eq}}}\right) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \exp\left(-\frac{E_a}{k_B T_m(t)}\right) dt$$

Module Temp

- Temperature cycling

$$C = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{dT_m}{dt}\, dt$$

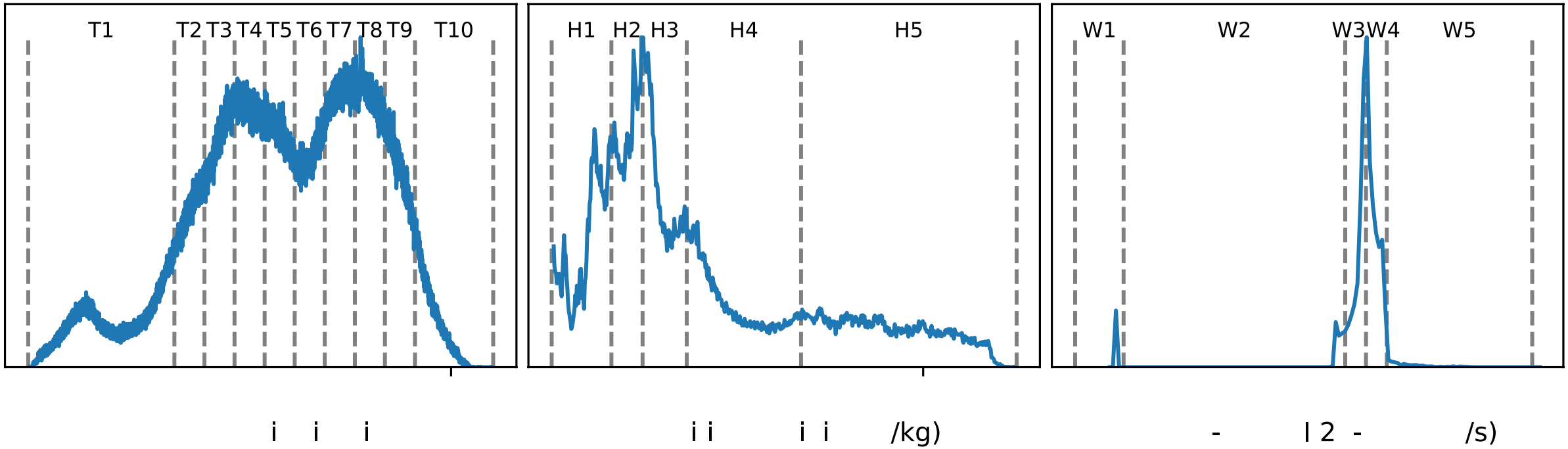
- Mean specific humidity

$$H = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \mathrm{SH}(t) dt$$

- UV stress

$$\mathrm{UV} = 0.05 \times \mathrm{GHI}$$

Define zones using thresholds on temperature, humidity and wind stressors



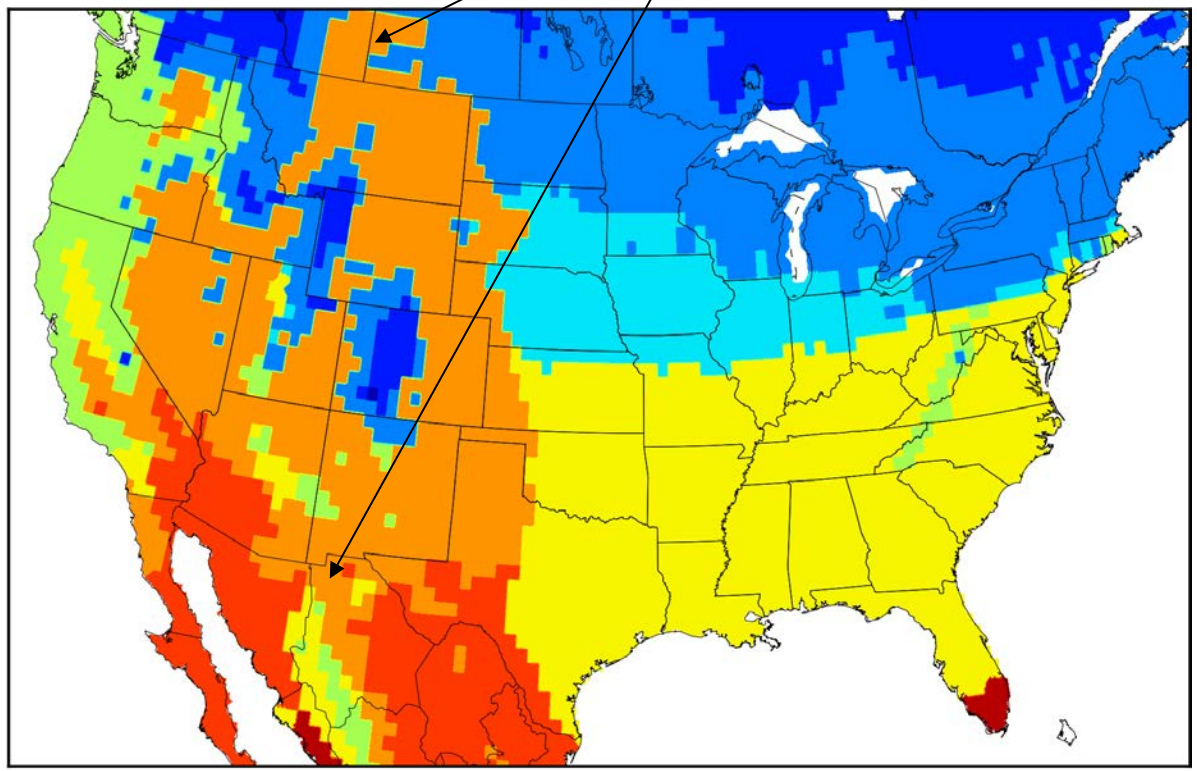
Define **9 temperature** and **5 humidity** zones based on “equal area” approach.

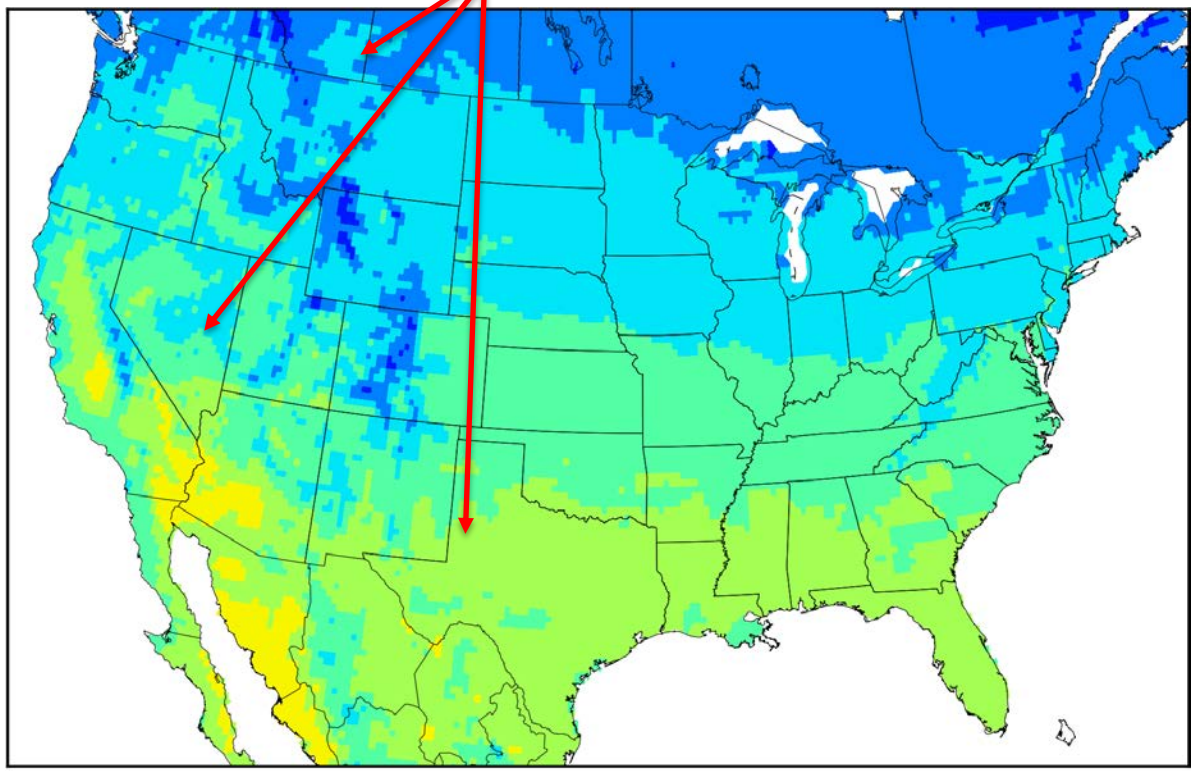
Define **5 wind zones** to to split difference between hurricane, tornado regions and less extreme wind zones.

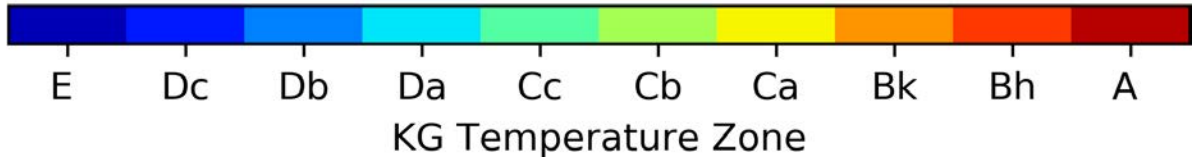
PVCZ vs. KG

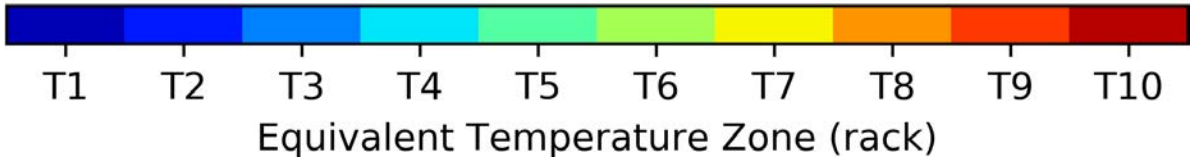
KG zone places regions from Mexico to Canada into a single zone (Bk).

PVCZ puts these areas into 3 different zones.



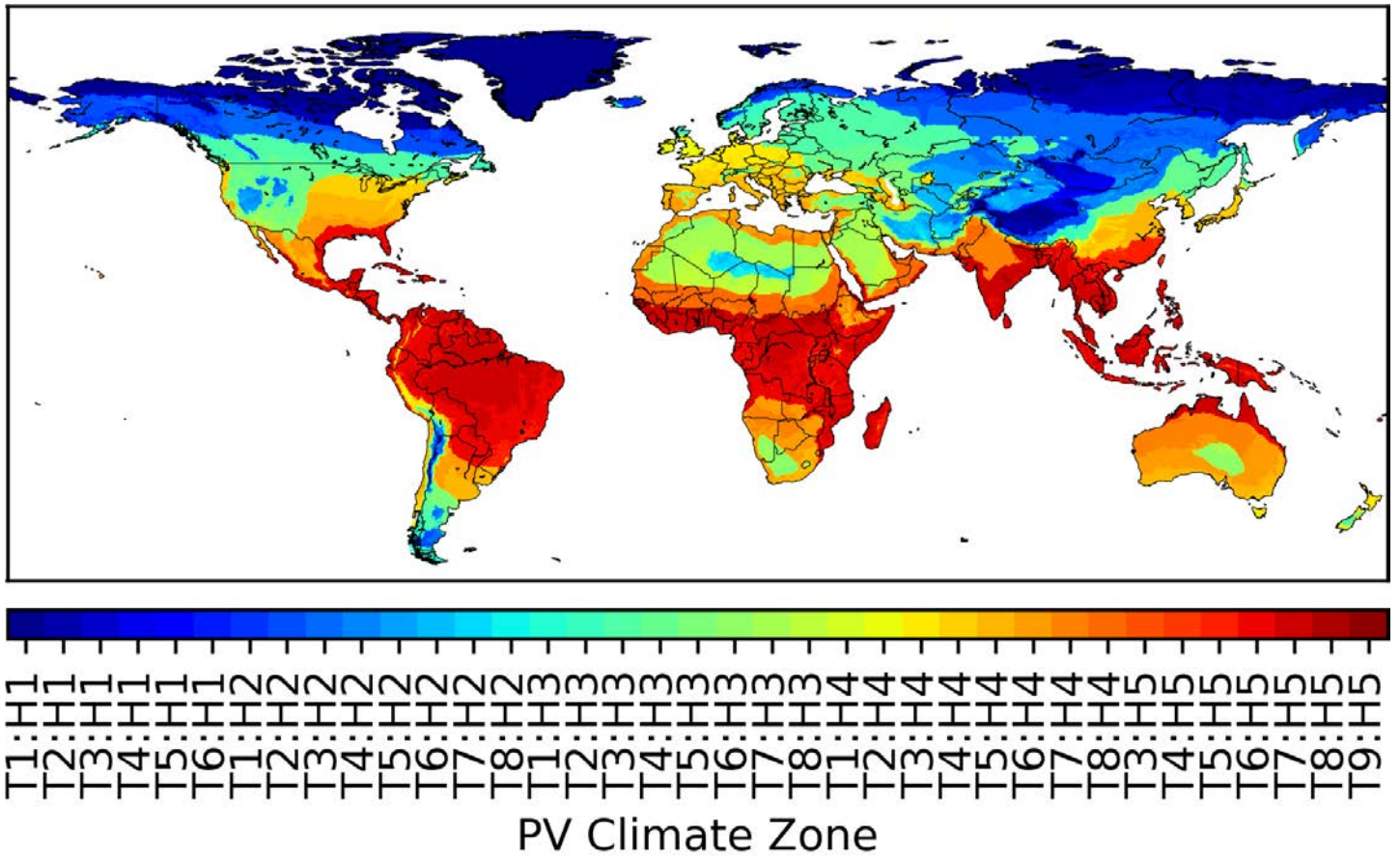






Conclusions and Outlook

- **We are developing a climate zone scheme specific to PV degradation**.
- Data freely available on datahub, open-source python package and web tool.
- Difficult to know how well it "works" – ideas welcome
- Future work is to analyze how to best combine stressors into zones that reflect real PV degradation



https://github.com/toddkarin/pvcz

https://pvtools.lbl.gov/pv-climate-stressors

Outline

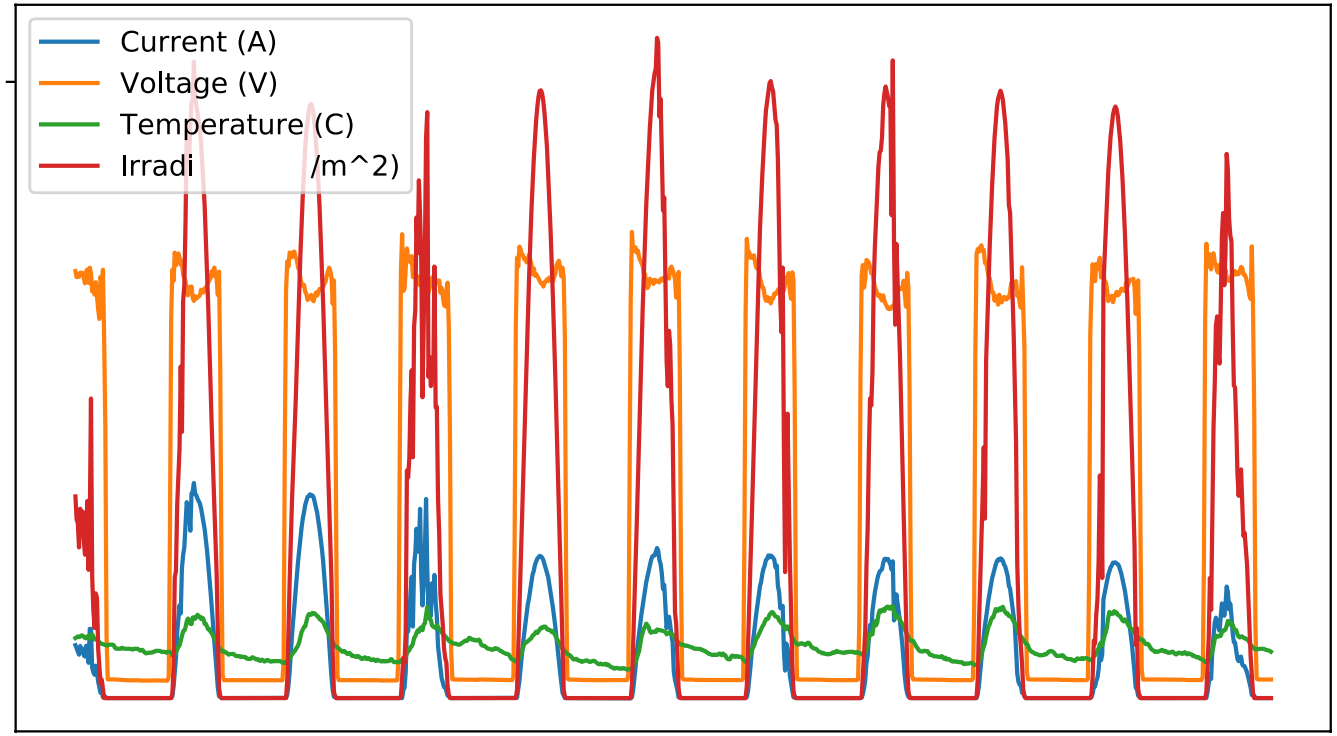
- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress)*
- Time series data
  - Extracting module parameters from production power data (*in progress*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques

Can we learn about module health from operating data?

Lots of potential problems

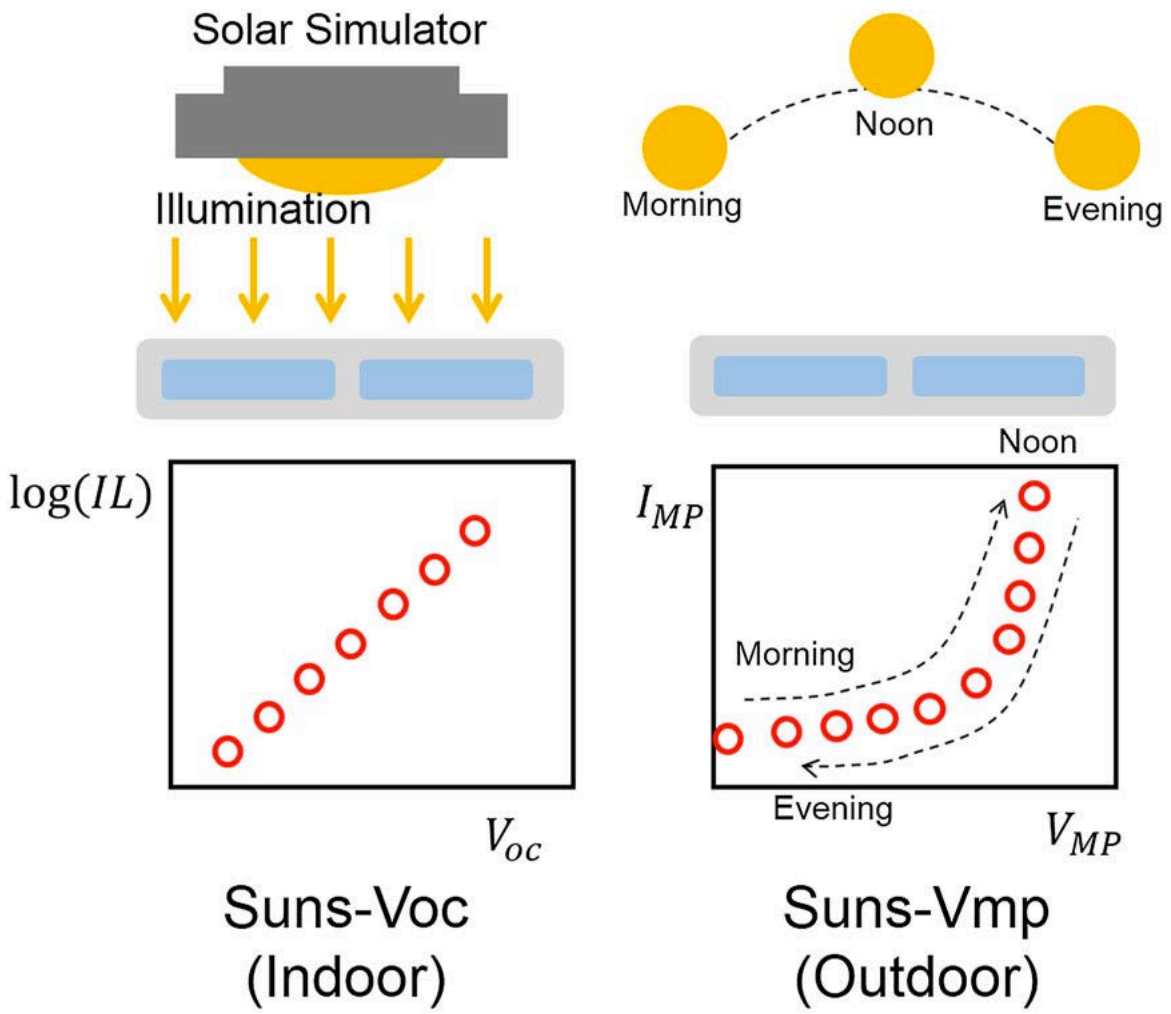
Big PV power plant

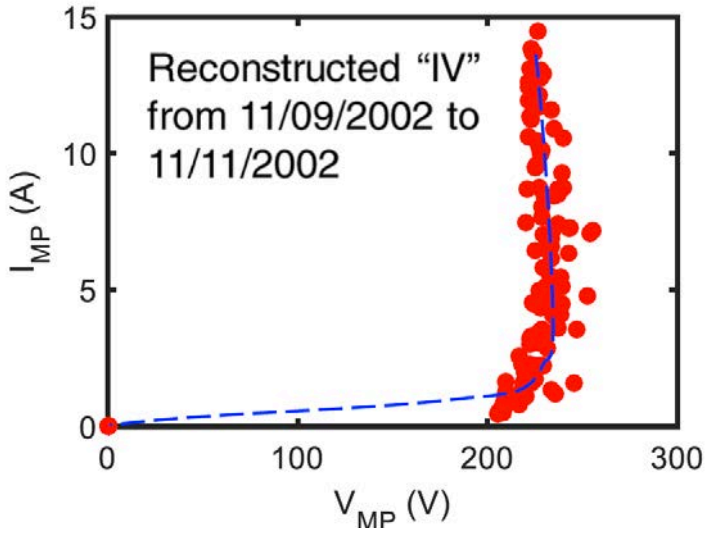
Operating data

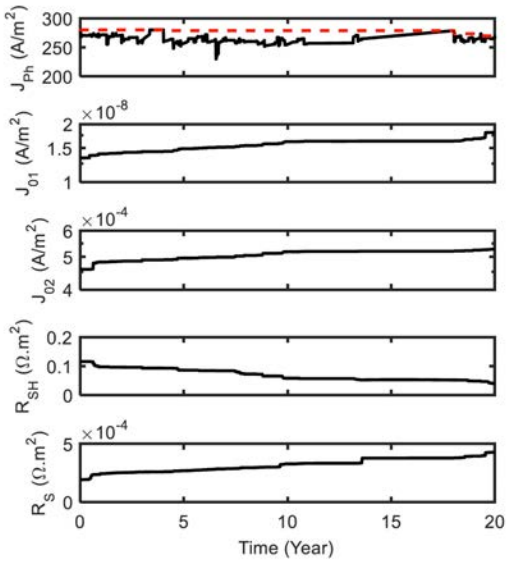


Goal: *Use operating data to extract module health without costly onsite surveys*

The recently reported “Suns-VMP” method provides a potential method

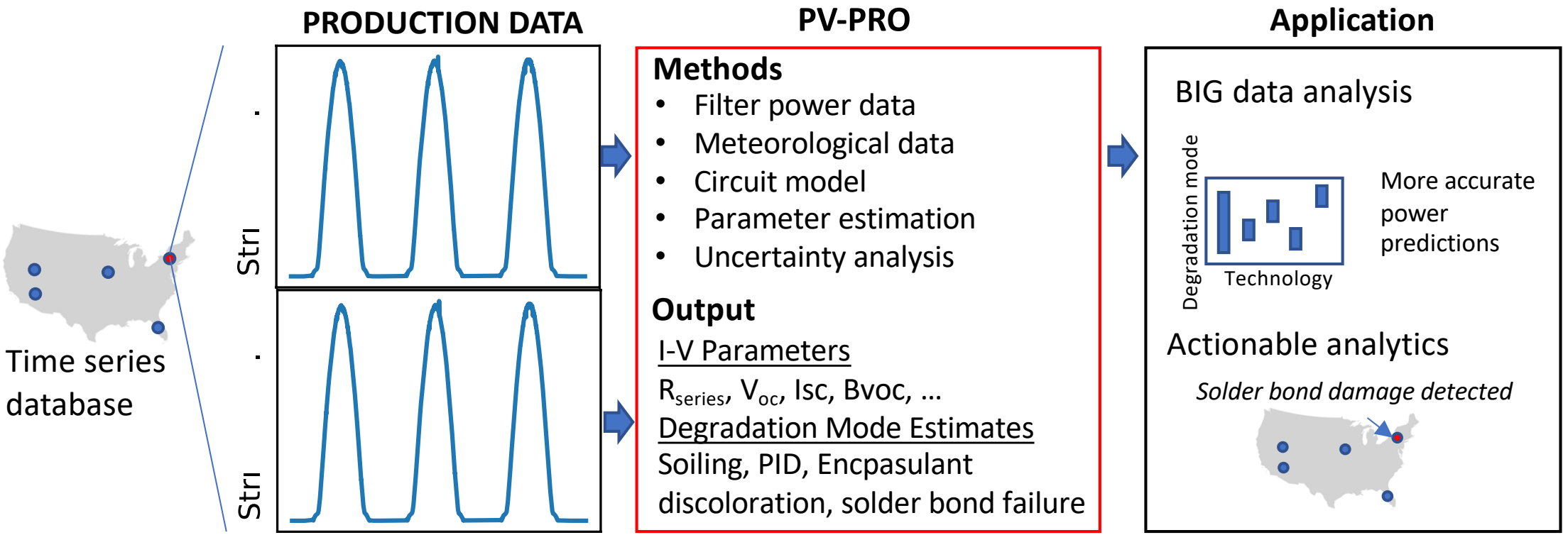






1. Sun, X., Vamsi, R., Chavali, K. & Alam, M. A. Real- time monitoring and diagnosis of photovoltaic system degradation only using maximum power point — the Suns - Vmp method. *Progress in Ph* 1–12 (2018). doi:10.1002/pip.3043

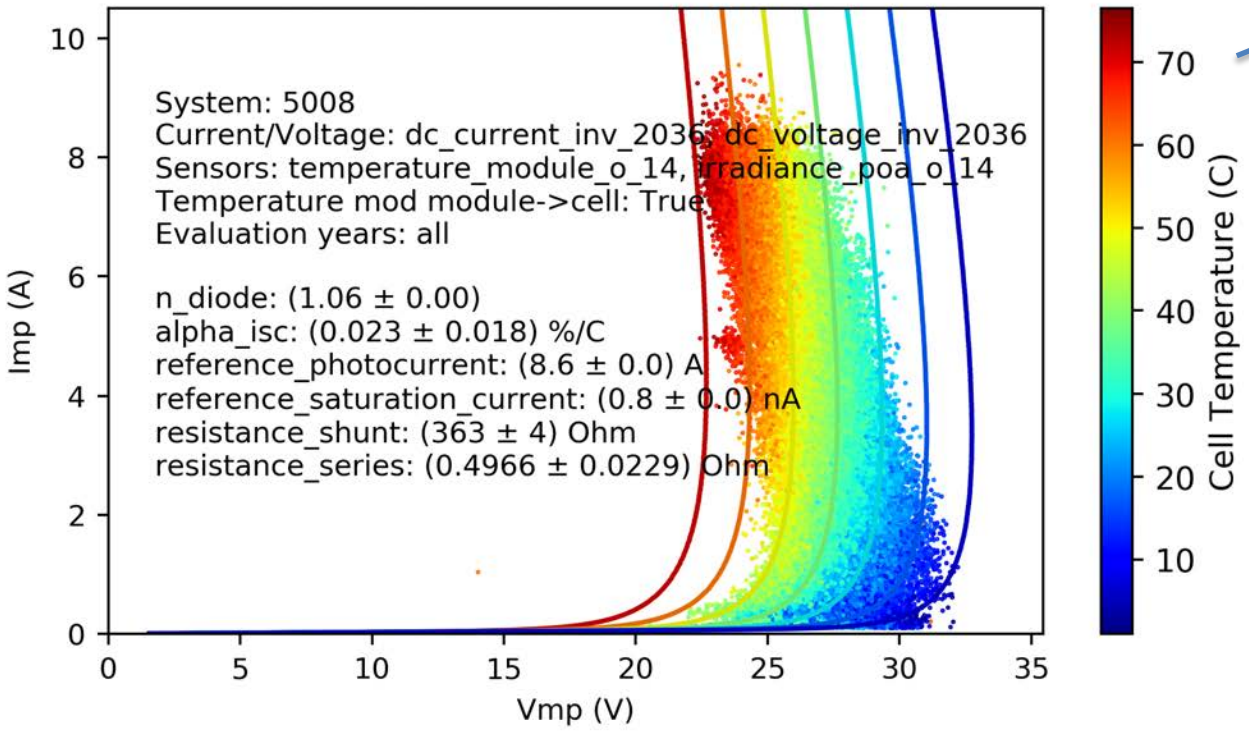
PV-Pro Project: Try to generalize Suns-Vmp over large production data in the DuraMat data hub



Maximum power point curve fit

Use equations for 6 parameter single diode model.

$$I = I_L - I_O\left[\exp\left(\frac{V + IR_S}{nV_{th}}\right) - 1\right] - \frac{V + IR_S}{R_{SH}}$$



Parameter	Fit	Datasheet
Reference photocurrent	8.6 A	8.5 A
Reference saturation current	0.8 nA	0.7 nA
alpha_isc	+0.023%/C	+0.07 %/C
ndiode	1.06	CEC: 1.03
Shunt Resistance	363 Ohm	CEC: 181 Ohm
Series Resistance	0.50 Ohm	CEC: 0.23 Ohm

Conclusions and outlook

- The Suns-Vmp method is a strategy to use *maximum power point data* to extract I-V curve information

- Our goal is to:
  - Extend the Suns-Vmp method (e.g., get Voc values before the inverter kicks on)
  - Improve the stability, reliability, and consistency of the Suns-Vmp method
  - Apply Suns-Vmp over a large data set (e.g., PVFleets) to determine degradation of circuit parameters over time

- Preliminary results are promising, but many open questions remain …

Outline

- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress)*
- Time series data
  - Extracting module parameters from production power data (*in progress*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques

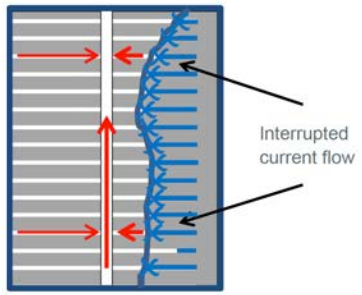
Goal – automatically analyze electroluminescence images for cracks

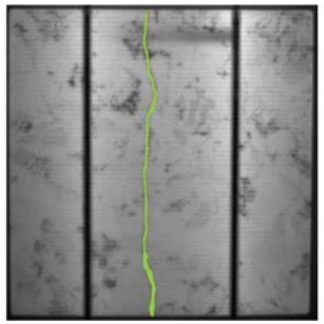
MBJ
services to know what's up

MBJ Solar Module Judgment Criteria

Analysis criteria for solar module testing in the Mobile Lab / Mobile PV-Testcenter

Date: 26.08.2019 – Revision 3.4

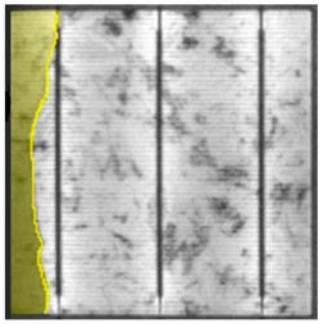




Judgment

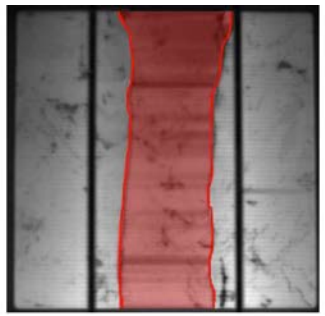
A further expansion of the cell break is not expected.

Possible cell area disconnection 0%.



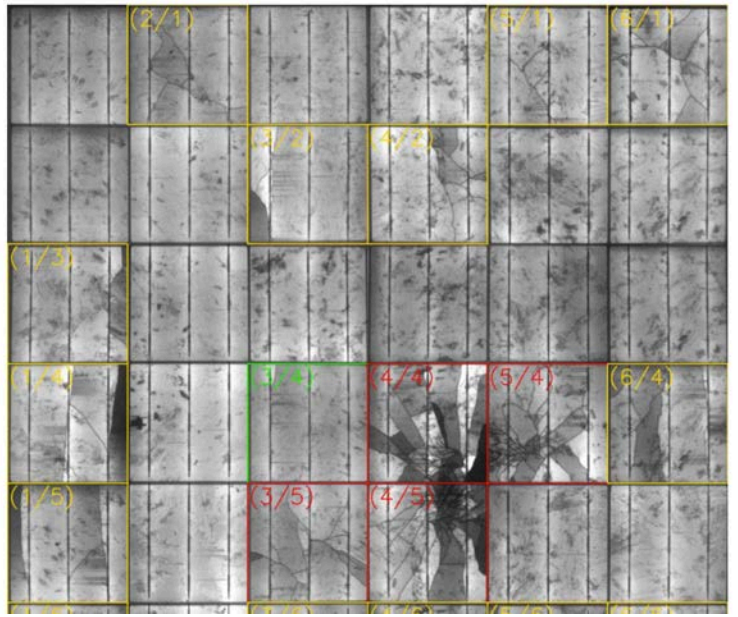
Judgment

Disconnected cell area approx. 10%.

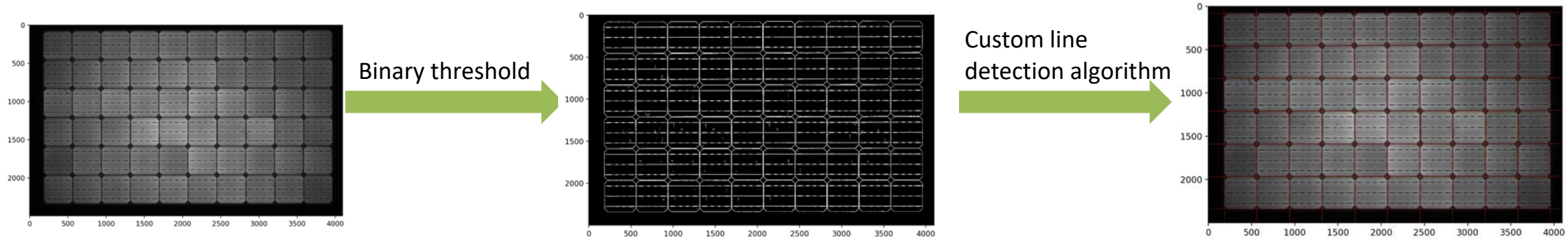


Judgment

Possible cell area more than 20%.



Current status – can automatically segment images (separate module and cells)



- 100% success rate on 47 indoor module images
- Robust to tilting, objects on side of module
- ~0.4 sec per image

Next steps

- Obtain a more diverse image library
  - PVEL may be supplying a large library of images
  - Peter Hacke has sent some images
- Hand-label cell images by crack category
- Train a neural network model to automatically classify the cracks
  - Many methodological details to work out and test
- Note that “cracked” versus “uncracked” cell classification via neural network was reported previously by Case Western

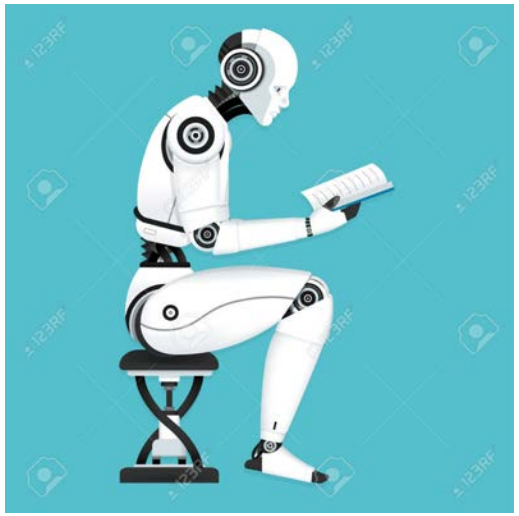
Karimi, A. M., Fada, J. S., Hossain, M. A., Yang, S., Peshek, T. J., Braid, J. L. & French, R. H. Automated Pipeline for Photovoltaic Module Electroluminescence Image Processing and Degradation Feature Classification. *IEEE Journal of Photovoltaics* 1–12 (2019). doi:10.1109/JPHOTOV.2019.2920732

Outline

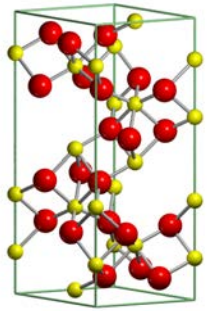
- Geospatial and climate data
  - Clear sky detection
  - Planning optimal string sizing for PV plants
  - Redefining climate zones for PV degradation analysis *(in progress)*
- Time series data
  - Extracting module parameters from production power data (*in progress*)
- Image data
  - Classifying and detecting cracks in electroluminescence images (*in progress*)
- Natural language (text) data
  - Potential opportunities for applying ML techniques

There is a lot of information in text sources, but unfortunately one cannot read through it all. Could ML do this instead?

papers to read “someday”



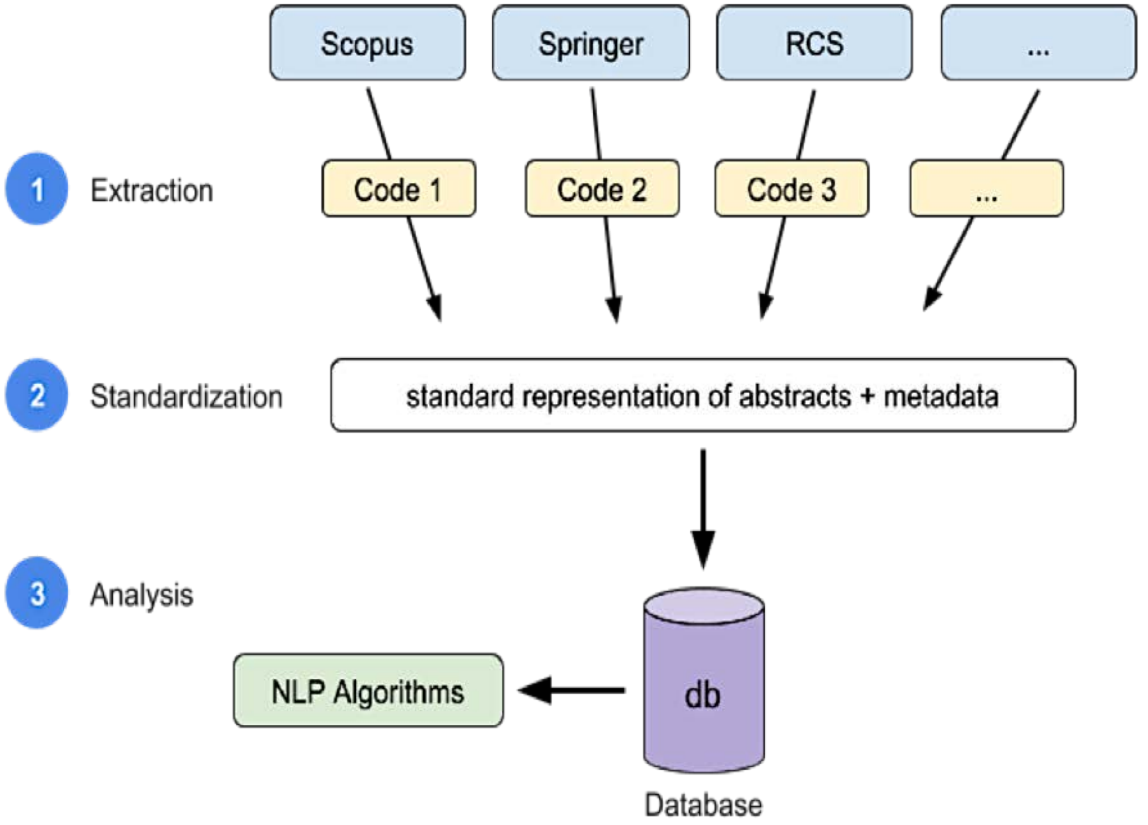
NLP algorithms



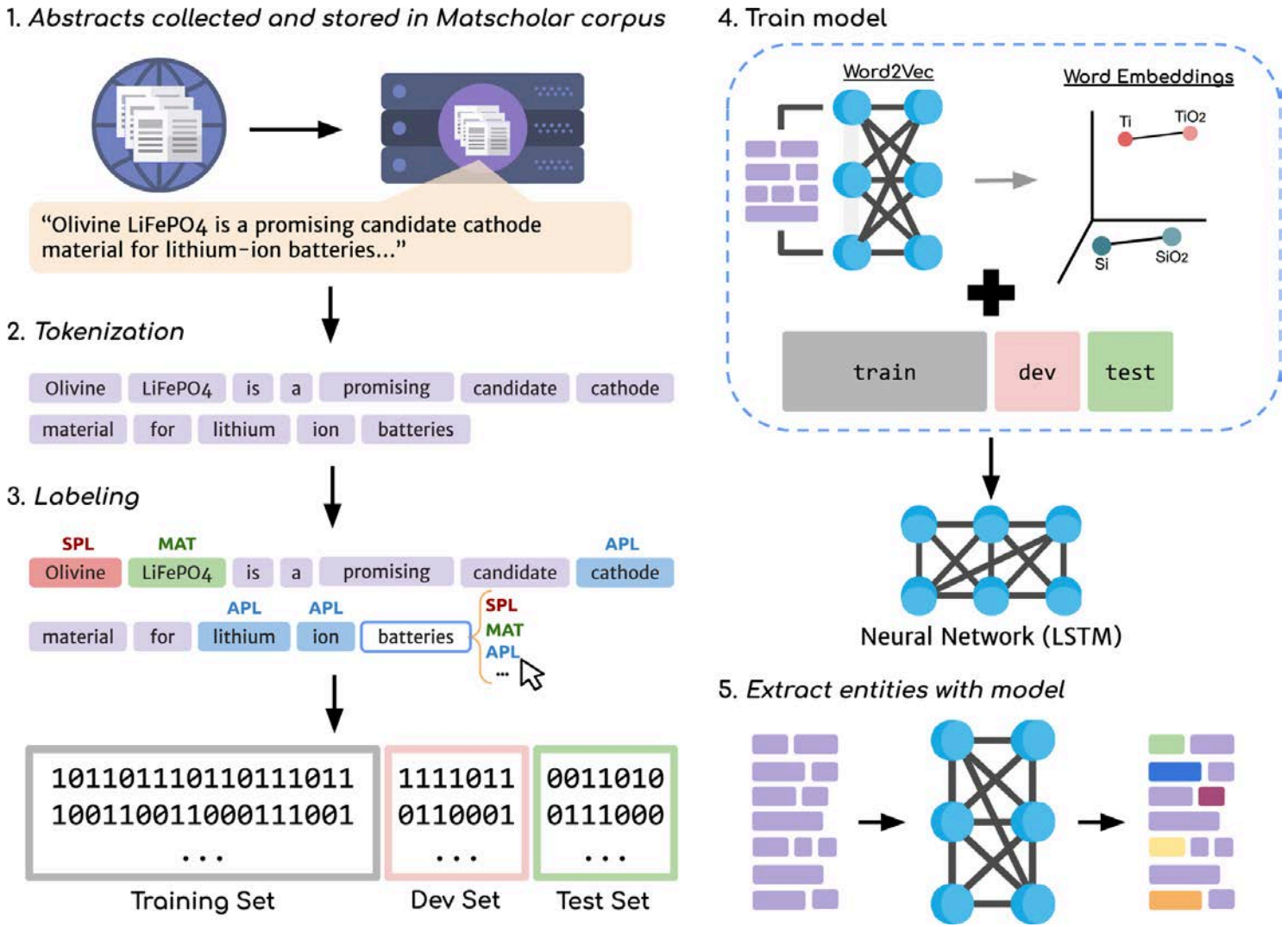
Goal: collect and organize knowledge embedded in the materials science literature

We have extracted ~2 million abstracts of relevant scientific articles

We use natural language processing algorithms to extract knowledge from all this data

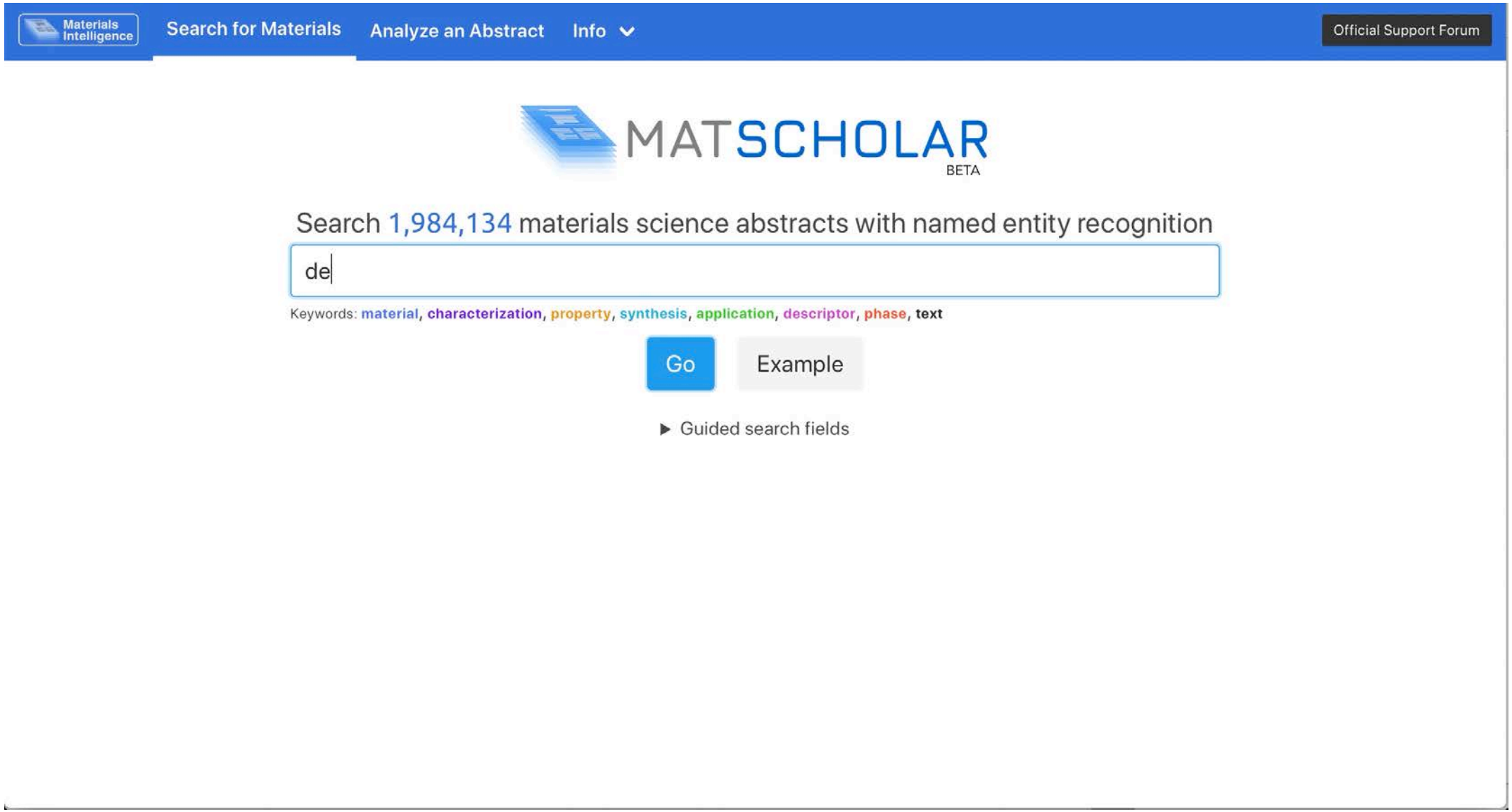


Developed algorithms to automatically tag keywords in the abstracts based on word2vec and LSTM networks

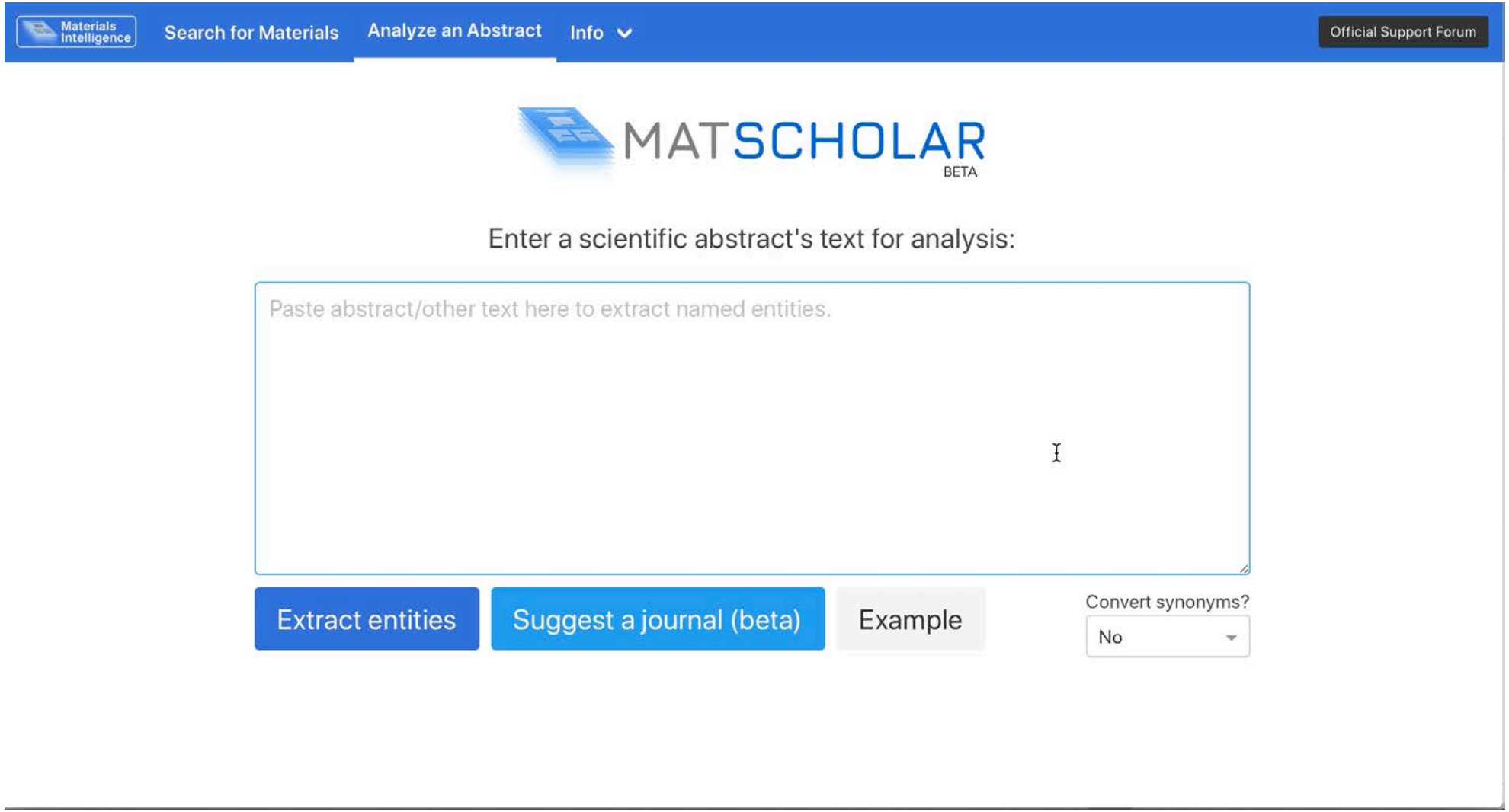


Weston, L. et al Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *J. Chem. Inf. Model.* (2019).

Now we can search!
Live on www.matscholar.com



You can also analyze abstracts on matscholar.com



What's different versus, say, Google Scholar?

- Domain-specific entity normalization
  - Recognizes that "CdTe" and "TeCd" are the same thing
  - Recognizes that "XRD" and "x-ray diffraction" are the same thing

- We can train vector representations of words that have interesting and surprising properties
  - The word vectors can make useful descriptors for machine learning algorithms
  - The word vectors can be used to predict "gaps" in the research literature, such as what materials should likely be studied for an application but haven't thus far

- Looking for applications in solar!

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. & Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98 (2019).

Overall conclusions

- There are a spectrum of opportunities for applying data analytics and machine learning to solar PV research

Do what we've done before, but improve things with data

Examples:

- refine Reno and Hansen thresholds for clear sky detection using data analysis
- Improve and apply the Suns-Vmp method to large data sets

Do things automatically that typically would be done manually

Examples:

- Formalize and automate the procedure for calculating string lengths
- Classify images into cracked / uncracked

Do things that we couldn't hope to do manually

Examples:

- Analyze large amounts of text data for advancing solar research?

Acknowledgements

Todd Karin

- String sizing
- Climate zones
- PV-Pro

Ben Ellis

- Clear sky detection

Xin Chen

- Image analysis

Also collaborations with **Mike Deceglie** (NREL), **Birk Jones** (Sandia), **Jenya Meydbray** (PVEL), **Robert White** (NREL), the PV-Pro and EL Crack detection teams, and the many people who gave us advice along the way!

This work was supported by the Durable Modules Consortium (DuraMat), an Energy Materials Network Consortium funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Solar Energy Technologies Office
(Work on natural language processing was funded by Toyota Research Institutes)

(slides already posted to hackingmaterials.lbl.gov)