Data Science and Analytics Applied to PV Power Plants and Materials Degradation & Lifetime

Roger H. French
Laura S. Bruckman

SDLE Research Center
Case Western Reserve University
Cleveland OH 44106
http://sdle.case.edu

Abstract

As solar power becomes a larger source of electricity and power for locations, it becomes increasingly important to fully understand and predict the power output of solar modules over their entire lifetime. Traditional solar module degradation tests are done under accelerated exposure environments, where the conditions are more aggressive than an outdoor environment, with the intent of testing the lifetime performance of a module within a more reasonable time scale. While these tests are certainly important, they can be either over or inadequately aggressive; therefore it is also critical to monitor real-world, outdoor power plants degrading under actual real-world exposure conditions. A combination of the two methods provides the best rate of change (ROC) or lifetime performance prediction of PV power plants, with indoor exposures degrading modules in a shorter time span, and outdoor modeling giving insight into the actual degradation patterns of systems and providing a comparison, by cross-correlation, of accelerated and real-world degradation.

With this in mind, the SDLE Research Center is developing data-driven modeling of ROC for PV systems based on a massive collection of time series data from numerous PV systems, both research and commercially fielded, including a variety of ages, brands, module types, and climate zones. To analyze and manage data from diverse PV plants, we have developed Energy- CRADLE, an automated data acquisition, management and analytics pipeline. The Energy- CRADLE is built in a high performance computing (HPC) environment which leverages distributed computing features of HBase/Hadoop and Spark cluster for distributed storage and parallel computing. We have also developed R and Python packages for integrating with HBase tables. For cross-sectional study of running on 100s of PV systems, we use fleets of parallel jobs via the SLURM workload manager.

While commercially fielded PV power plant data sources may be of a lower quality than research focused PV sites, being able to use data from commercial plants greatly increases the length of time series datasets available for analysis, making it a unique, at-scale resource for cross-sectional studies of thousands of PV systems. This large scale data collection is used to determine what the degradation patterns of real world systems are as a function of location, climatic zone, PV module and inverter brands and what factors might affect the behavior of PV modules over time. The current scope of the data available includes thousands of PV system inverters located across hundreds of sites with power capacities from single modules to hundreds of megawatt plants, located across many different climate zones.

Given the large scale, heterogeneity and diversity of the data between the PV systems, a method had to be developed to determine the rate of change, or the rate at which the power output changes over time, for each PV system consisting of PV modules and their inverter. As this data comes from many sources, there are inconsistencies between datasets, such as different available variables, data quality, or the data capture interval, that the method had to be able to accommodate. The Month-by-Month (MbM) method was developed at the SDLE Research Center with these problems in mind, being able to handle various intervals of data, as well as different variables, the most common of which being different irradiance measurements. The MbM method consists of three models, the **β** Pseudo-month Predictive Model divides the data into 30 day long “pseudo-months” where it is assumed that negligible degradation occurs over the 30 day time period. A multiple linear regression model is built for each pseudo month based on the given environmental variables, such as irradiance, temperature, and wind speed. Once a model has been built for each month, representative weather conditions are determined for the given PV system which are the average temperature, the average wind speed, and the minimum value of all the peak irradiances for each pseudo-month. The representative weather conditions are applied to each **β** model and predicted power outputs for each month are determined. Once the predicted power for each month has been determined, the **ξ** Piecewise Regression Model uses a weighted regression to calculate the rate of change of the system (%/year) from the slope and y-intercept of the predicted power over time. The **ξ** model is weighted to the standard errors for each **β** model, improving the robustness of the method by reducing the influence or noisy or less precise pseudo-months. Once the rate of change for each system is determined, a **γ** Cross-Sectional model of the rate of change as a function of the metadata for the PV systems, such as module brand or climate zone, providing insights into the factors causing more or less severe power loss in these outdoor PV systems.

Seasonal decomposition is used to reduce the impact of seasonality on the calculated rate of change. Fluctuations in power can be seen as a yearly cycle with the seasons, potentially influencing the calculated rate of change. Performing time series seasonal decomposition to isolate the seasonal and trend components of the power time series so as to reduce the influence of seasonality on the ROC results. Clear sky identification is the latest addition to the MbM analysis pipeline. Modeled weather data, derived from satellite imagery combined with an empirical atmospheric model, is pulled from SolarGIS for each PV system as supplemental weather data. By comparing the modeled weather conditions from SolarGIS and the measured weather conditions from the system, the clear sky, or points at which there was no cloud cover, can be identified. This identification is done using the PVLib-Python open source library. Clear sky identification has many benefits. Isolating clear sky points can reduce the noise of the data and ensure that the conditions are similar between two given points. Most importantly, however, is it can be used to track sensor drifting which can be highly problematic in long term time series. The SolarGIS data can also be used as a supplement if sensor drifting is observed, as the SolarGIS data will not drift over time.

SDLE Research Center: Acknowledgements

Projects

CWRU Faculty
- Roger French, Laura Bruckman, Alexis Abramson, Jennifer Carter, Mehmet Koyüturk

Post-doctoral Research Associates
- Jennifer Braid, Nick Wheeler, Rojiar Haddadian, Wei-Heng Huang

Graduate Students
- Devin Gordon, Yu Wang, Donghui Li, Alan Curran, Addison Klinke
- Justin Fada, Arash Khalilnejad, Ahmad Karimi, Xuan Ma,
- Rhener Zhang, Menghong Wang, JiQi Liu,

Undergraduates
- Andrew Loach, Lucas Fridman, Yiyang Sheng, Jonathan Ligh, Silas Ifeanyi
- Noah Tietsort, Kevin Nash, Rachel Swanson, Abhi Devahti, Jonah Larson

High School: Sheina Cundiff, Dominique Gardner, Precious Flanders

SDLE Staff: Chris Littman, Rich Tomazin

Outline

Peta-byte & Peta-flop Computing

Supervised Classification of PV Cell Electroluminescent Images

Time-series Analysis of Real World PV Systems

Machine Learning of Real-world I-V Curves of PV Modules

Data Analytics of Complex Systems

Materials are parts of a Complex Systems:

- Coatings on Complex Substrates
- Used in Complex Environmental Exposures and Climate Zones

Materials Degradation Predictive & Mechanistic Models

- Predictive Modeling of Materials Degradation
- Mechanistic Network Models To Guide New Materials Development
- Cross-correlation of Real-world and Accelerated Studies for Service Life

Image Processing

- Develop Pipeline Methodology, Apply to Historical Datasets
- Cluster Output and Compare Cell-level Heterogeneity with I-V

Machine Learning

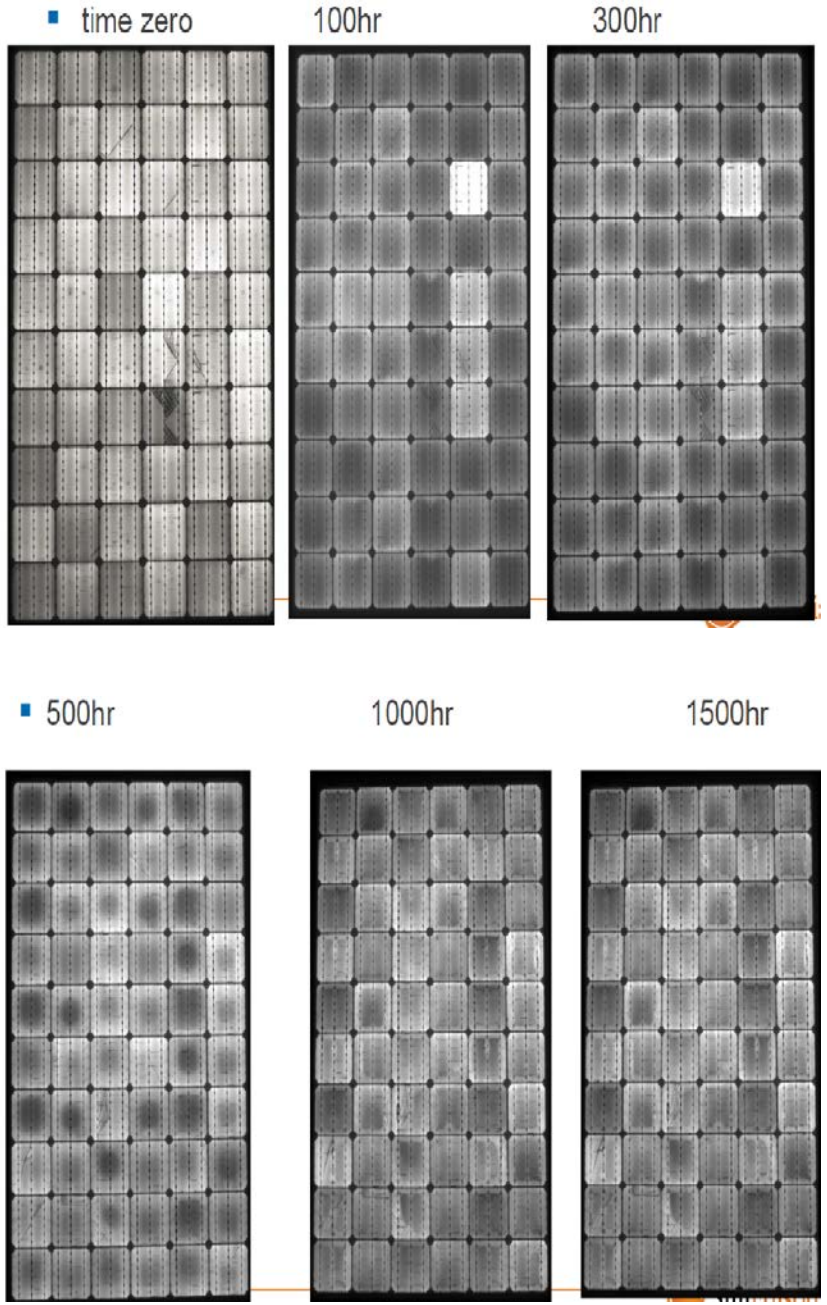
- Classifying Stages of Degradation: Identifying Feature Change Over Time
- Cluster Cell Behavior to Model of Ensemble Performance
- Determine Features Variation with Indoor Testing, Compare / Contrast

Time Series Analysis

- High Performance Computation / Data Pipelining for Rd Analysis
- Subset Datasets by Climate, Module Brand, Inverter Brand

Sample Sets of Systems such as PV power plants

- Sample Set Segmentation, Identify Performance Changes and Variations



Data Science:
Informatics and Analytics

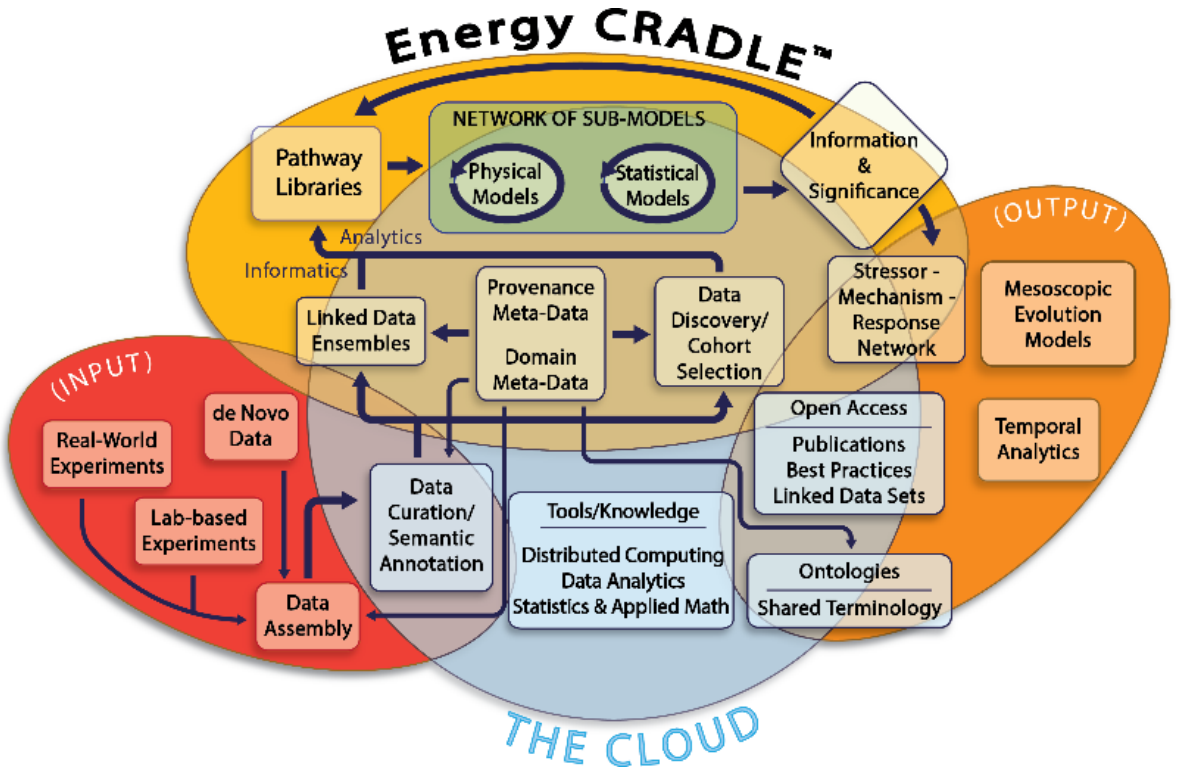
SDLE Research Center
Case Western Reserve University
Cleveland OH 44106

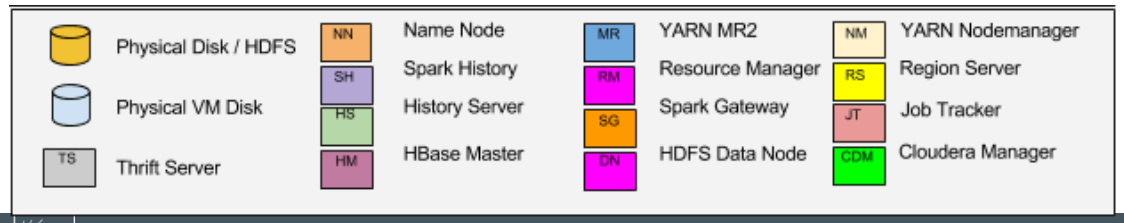
Energy-CRADLE v2.1 Architecture: Petabyte and Petaflop Computing

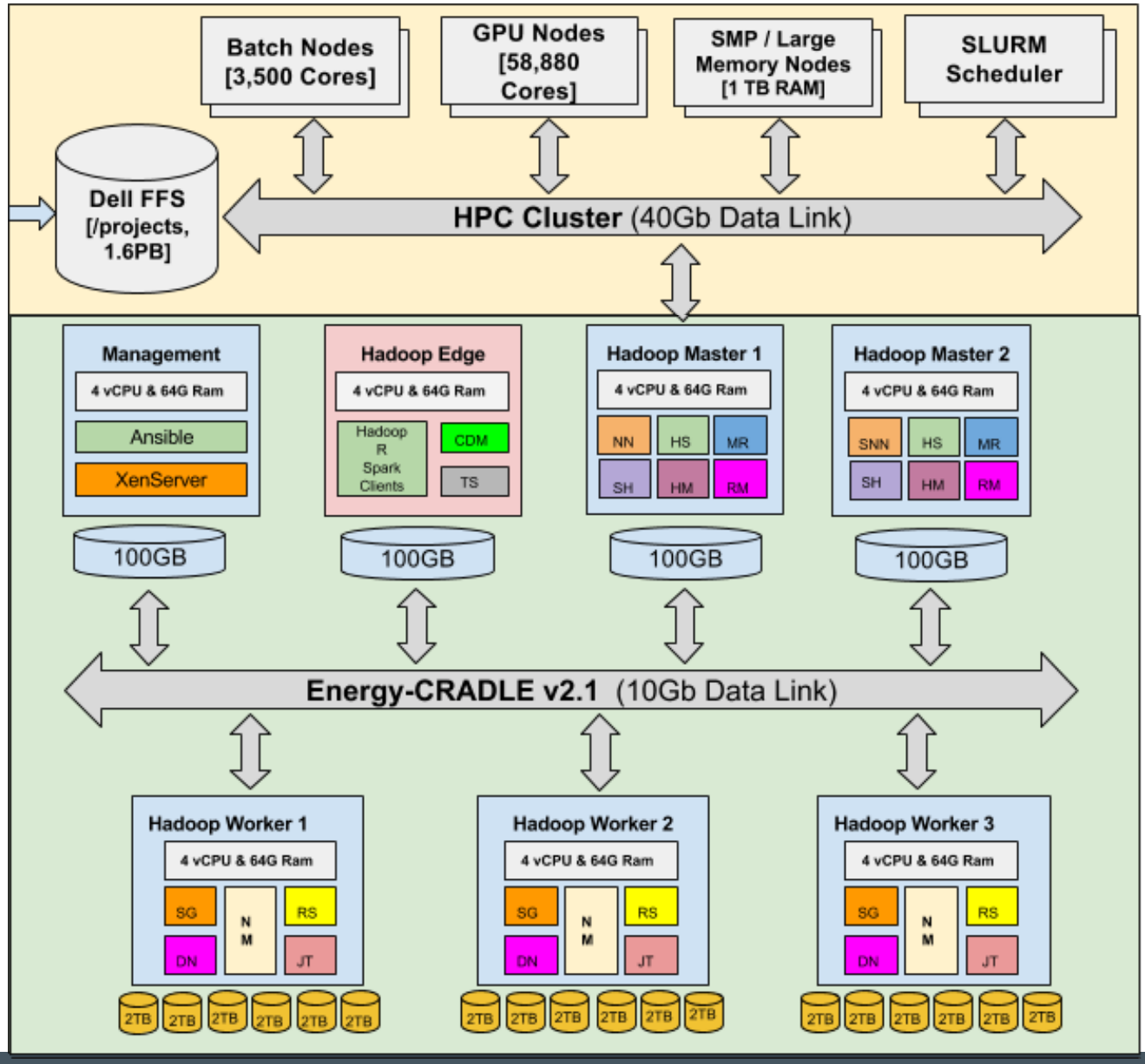
National Strategic Computing Initiative 2015

Hadoop/Hbase/MapReduce/Spark

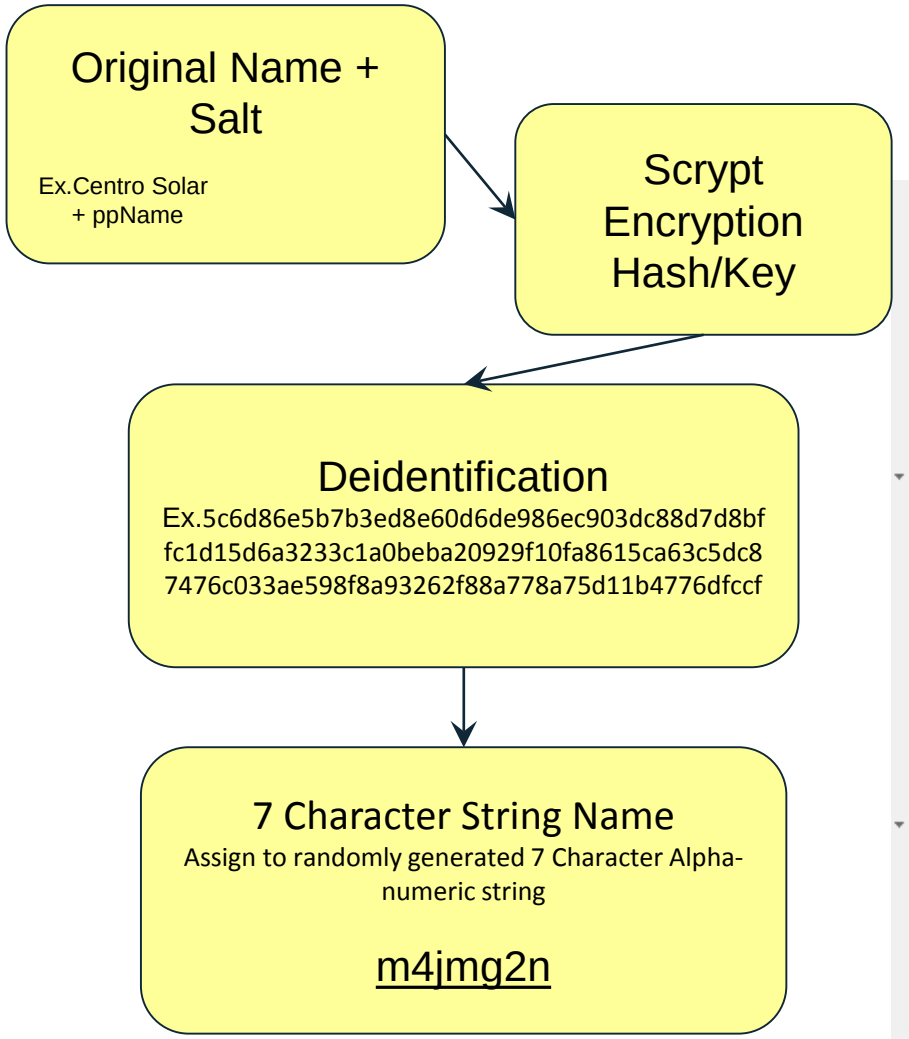
Based on Cloudera CDH5 distribution







Overview: Deidentification & Anonymization

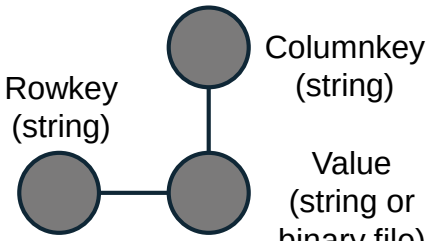


Allows us to analyze and report data

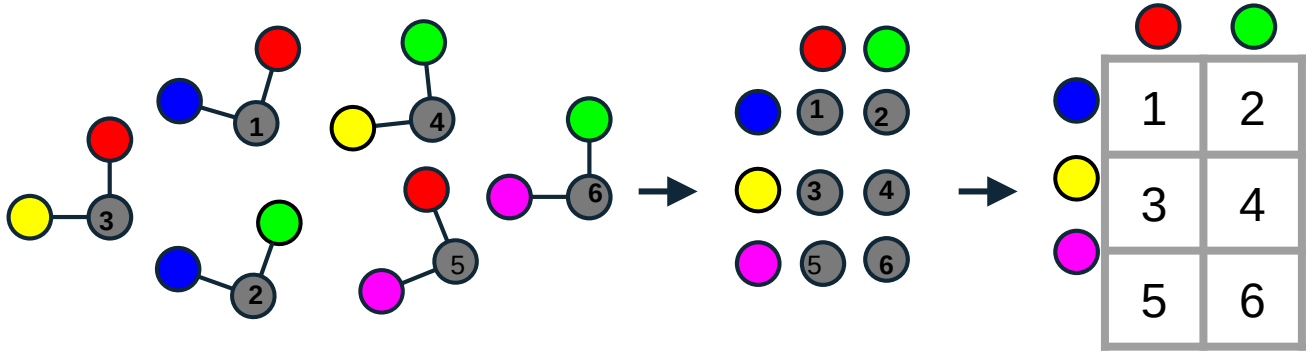
- Without being biased by brand or backsheet composition

```
## SDLE scrypt function
## take the password, creat the scrypt key derivation, concadinate the key derivation into a string.

set.seed(10);salt <- sample(1:10, 32, replace=TRUE)

#pw need to a character sting
#salt need to be 32 number

scryptkey <- function(pw , salt){
  password <- charToRaw(pw)
  keys <- scrypt::scrypt(password, salt, 65536, 8, 1)
  keys <- paste(keys, collapse = "")
  return(keys)
}

##example
tb<- read.csv("H:/Downloads/Book1.csv", stringsAsFactors = FALSE  )
## concatenate col
tb.scrypt<-matrix(NA, nrow = nrow(tb),ncol = 3)
set.seed(10);salt <- sample(1:10, 32, replace=TRUE)
for ( i in 1: nrow(tb)){
  brand_key <-scryptkey(tb[i,1],salt)
  model_key <-scryptkey(tb[i,2],salt)
  serial_number_key <-scryptkey(tb[i,3],salt)
  tb.scrypt[i,]<-c(brand_key, model_key, serial_number_key)
  print(i)
}
tb.scrypt<- data.frame(tb.scrypt)
colnames(tb.scrypt)<-c("brand_key","model_key", "serial_number_key")
write.csv(tb.scrypt,file = "scryptedcolumns.csv",row.names = FALSE)
```

Energy-CRADLE: Hadoop/Hbase Schema & NoSQL DB Abstraction



HBase 'Triple'



Combines Lab data (Spectra, Images etc.) With Time-series Data (PV Power Plant Data)

High Performance PV Data Analytics: Petabyte Data Warehouse In A Petaflop HPC Environment

- In-place Analytics: Distributed R-analytics in Hadoop/HDFS
- In-memory Data Extraction: To Separate HPC Compute Nodes

IEEE JPV

A non-relational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites

Yang Hu, *Member, IEEE,* Venkat Yashwanth Gunapati, Pei Zhao, Devin Gordon, Nicholas R. Wheeler, Mohammad A. Hossain, *Member, IEEE,* Timothy J. Peshek, *Member, IEEE,* Laura S. Bruckman, Guo-Qiang Zhang, *Member, IEEE,* and Roger H. French, *Member, IEEE*

Hu, Y., V. Y. Gunapati, P. Zhao, D. Gordon, N. R. Wheeler, M. A. Hossain, T. J. Peshek, L. S. Bruckman, G. Q. Zhang, R. H. French. "A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites." IEEE Journal of Photovoltaics 7, no. 1 (January 2017): 230–36. doi:10.1109/JPHOTOV.2016.2626919.

Hbase

Non-relational Database

- Enables new variables
- To be created at any time

Grad your data from Hbase

- Cradle_Get
- Straight into memory
- Don’t download local copies

Write back your analytic results

- Into Hbase
- You can create new variables
- No fixed RDBMS Schema

Your data Is continually enriched

- With New Results
- Available for next analyses

Semantic Web Standard, W3C

- Resource Description Format
- They are just rdf triples

  Using RDF Schema

HBase Conventions

SDLE Research Center

March 17, 2017

HBase Conventions

This document contains all commonly used variables and provides the corresponding HBase name for each along with other useful information pertaining to those variables. At the end of this document are tables of commonly used string inputs for relevant variables.

Variables

Common Variables

Deprecated Name	HBase Name	Unit/Standard	Format (if applicable)
time stamp	tmst	local	dd-mm-yyyy hh-mm-ss
time stamp	tutc	UTC	yyyy-mm-dd_hh:mm:ss

Meta Variables

Deprecated Name	HBase Name	Unit/Standard	Format (if applicable)
longitude	long	degrees	ddd-mm-ss-C
latitude	lati	degrees	ddd-mm-ss-C
zip code	zpcd	–	12345
location	loca	–	–
K-G climate zone	kgcz	–	Csa
K-G group	kggr	–	C
K-G type	kgty	–	s
K-G subtype	kgst	–	a
7-digit alpha numeric	rcid	string	abc1234
sample number	sanm	string	sa12345_00
sample location	salc	string	–
material type	maty	string	–
material brand	mabr	stirng	–
material model number	mamo	string	–
material fabrication	mafb	string	–
thickness (nominal)	thck	num	–
thickness units	thun	string	–
exposure	exps	string	–
retention step	rest	num	–
module model	modm	128 digit alpha numeric	–
inverter model	invm	128 digit alpha numeric	–
module supplier	mods	128 digit alpha numeric	–
inverter supplier	invs	128 digit alpha numeric	–
square feet	sqft	–	–

Weather Variables

Deprecated Name	HBase Name	Unit/Standard	Format (if applicable)
Ambient_temp	AirTemp_Avg & temp	Celcius	–
Wind_speed	WindSpd_Avg & wspa	m/s	–
time at max wind speed	WindSpd_TMx & wspt	m/s	–
wind direction	WindDir_Avg & wdra	–	–
wind maximum	WindSpd_Max & wspm	m/s	–
GHI (global horz. irr.)	ghir	W/m^2	–
POA (plane of array)	poay	–	–
barometric pressure	bprs	–	–
rel. humidity	relh	–	–
rain fall	rnfl	–	–
rain duration	rndr	–	–
rain intensity	rnin	–	–
source	noaa or wsta or sgis	–	–

Power Variables

Deprecated Name	HBase Name	Unit/Standard	Format (if applicable)
AC_power	iacp	kW	–
AC_energy	iace	kWh	–
DC_Voltage	dcvt	V	–
DC_Current	dccr	A	–
Module_temp	modt	C	–
power max	pwmx	W	–
fill factor	ffff	%	–
current max power	imxp	A	–
current short circuit	ishc	A	–
voltage max power	vmxp	V	–
voltage short circuit	vshc	V	–
series resistance	rssr	Ohm	–
shunt resistance	rssh	Ohm	–

Energy Variables

Deprecated Name	HBase Name	Unit/Standard	Format (if applicable)
MeterType	mtrt	–	–
CustomerID	cstm	–	–
FacilityType	fclt	–	–
ConditionedSquareFootage	cnsf	–	–
SystemID	stmi	–	–
SystemType	stmt	–	–
Hierarchy	hrrc	–	–
PointID	pnti	–	–
CommodityName	cmdn	–	–
CommodityType	cmdt	–	–
PreviousRawValue	prvr	–	–
PresentRawValue	prsr	–	–

J. Sun and Q. Jin, “Scalable RDF store based on HBase and MapReduce,” *ICACTE*, 2010, vol. 1, pp. V1-633-V1-636.

ETL and Data Ingestion to Hbase

ETL: Extract, Transform, Load

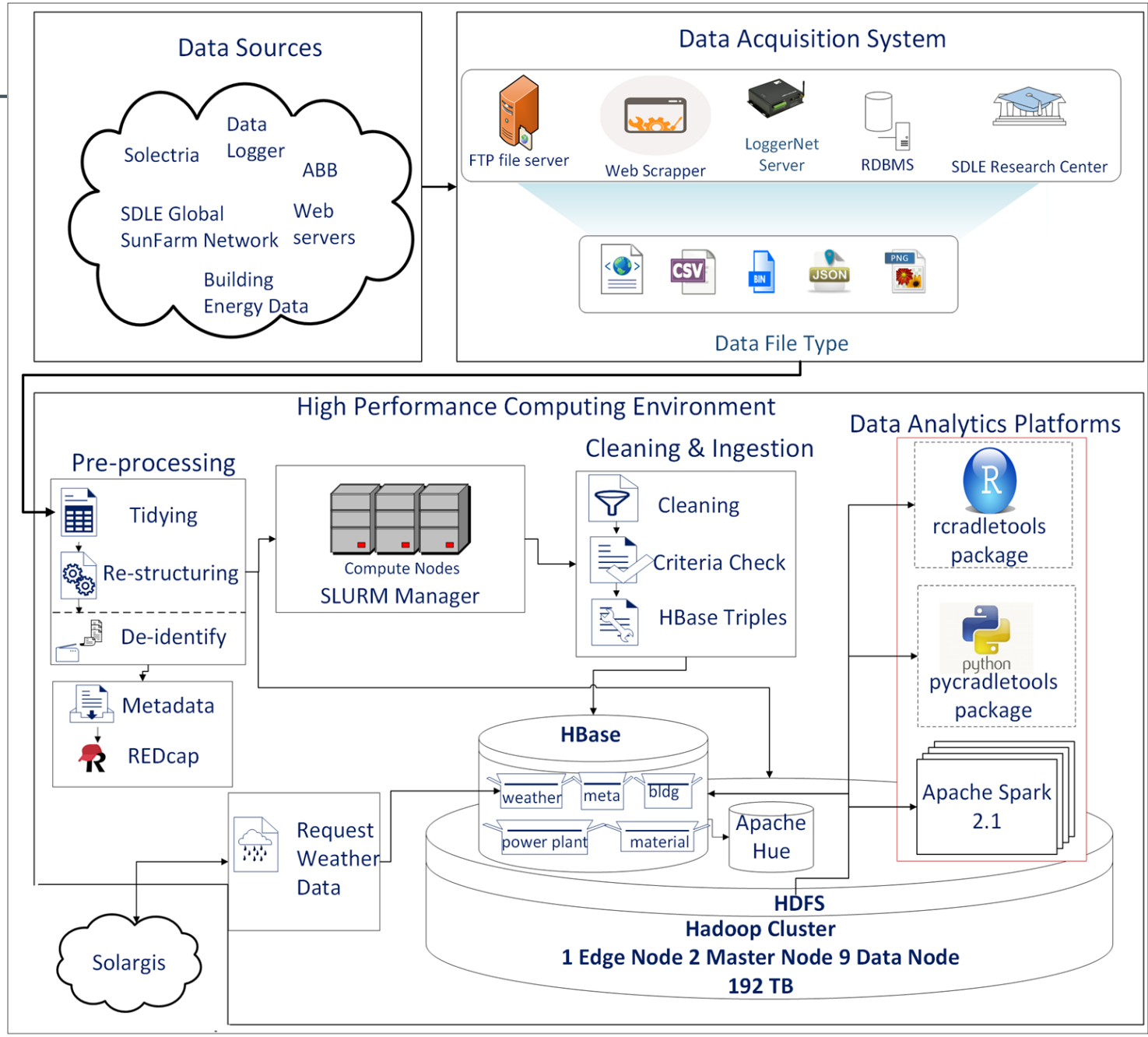
- Standard process for data acquisition
- Typically into an RDBMS system
  Relational Database Management System

Data Ingestion

- Used for NoSQL Databases like Hbase
- Preparing the data for inclusion in Triples
  Rowkey
  Columnkey
  Value

Hbase Tables

- Metadata: information about the data
- Weather: Weather & Irradiance Time-series data
- Power Plant: PV Power Plant Time-series data
- Buildings: Building Electrical Time-series data
- Materials: Spectral, Image data of Materials

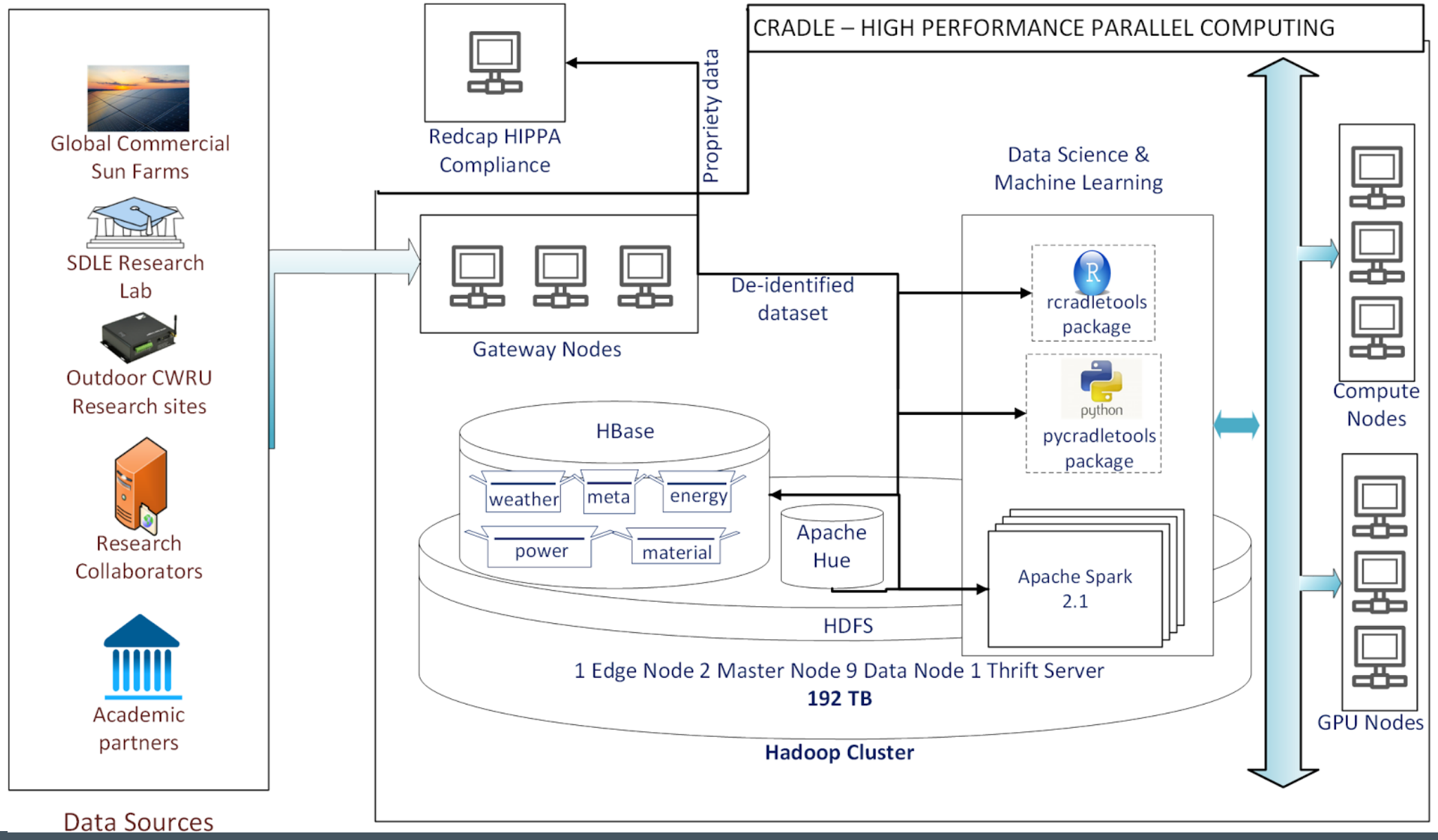


Data Analytics Pipelines

Using R & Pyton

In-place Analytics

Write-back
all Results
into Hbase



Open Data Science Tool Chain

Using Open-source tools

Reproducible Research

- Using Rmarkdown reports
- Python Jupyter Notebooks
- Add new data
- Recompile your report
- All new figures and report!
- Well Documented Code/Reports

High Level Scripting Languages: R, Python

- Similar to MatLab

Rstudio Integrated Development Environment

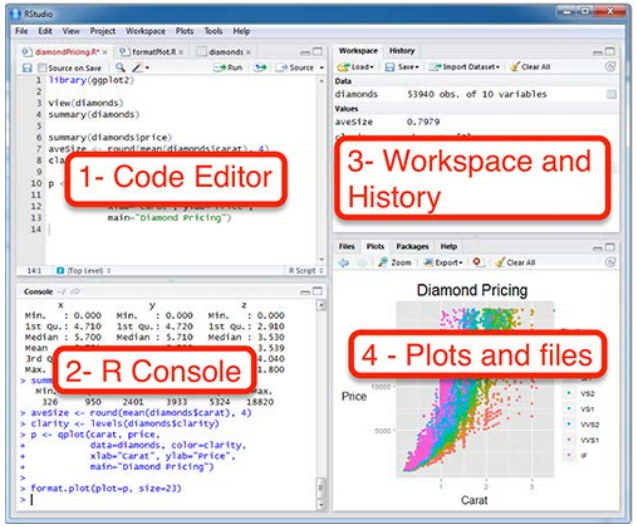
- Commercially Supported

Git Repositories for Code Version Control

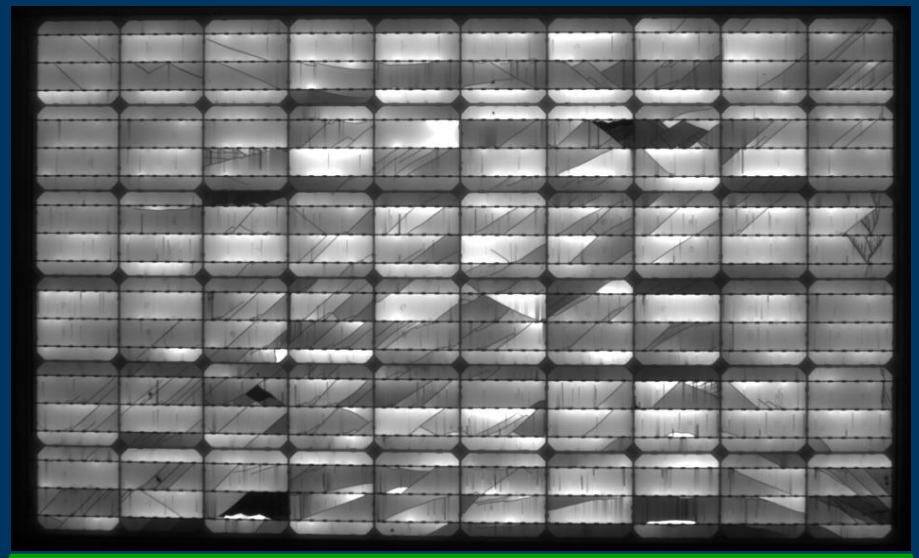
- Share code scripts with colleagues
- Share project data and reports with others

Github, BitBucket, GitLab for Collaboration

- Website hosting your Code Repositories



Electroluminescent Image Processing and Cell Degradation Type Classification via Computer Vision and Statistical Learning Methodologies



Mono-Si Al-BSF TC800

Justin S. Fada[1],

[1]SDLE Research Center, CWRU,
Cleveland, OH.

justin.fada@case.edu

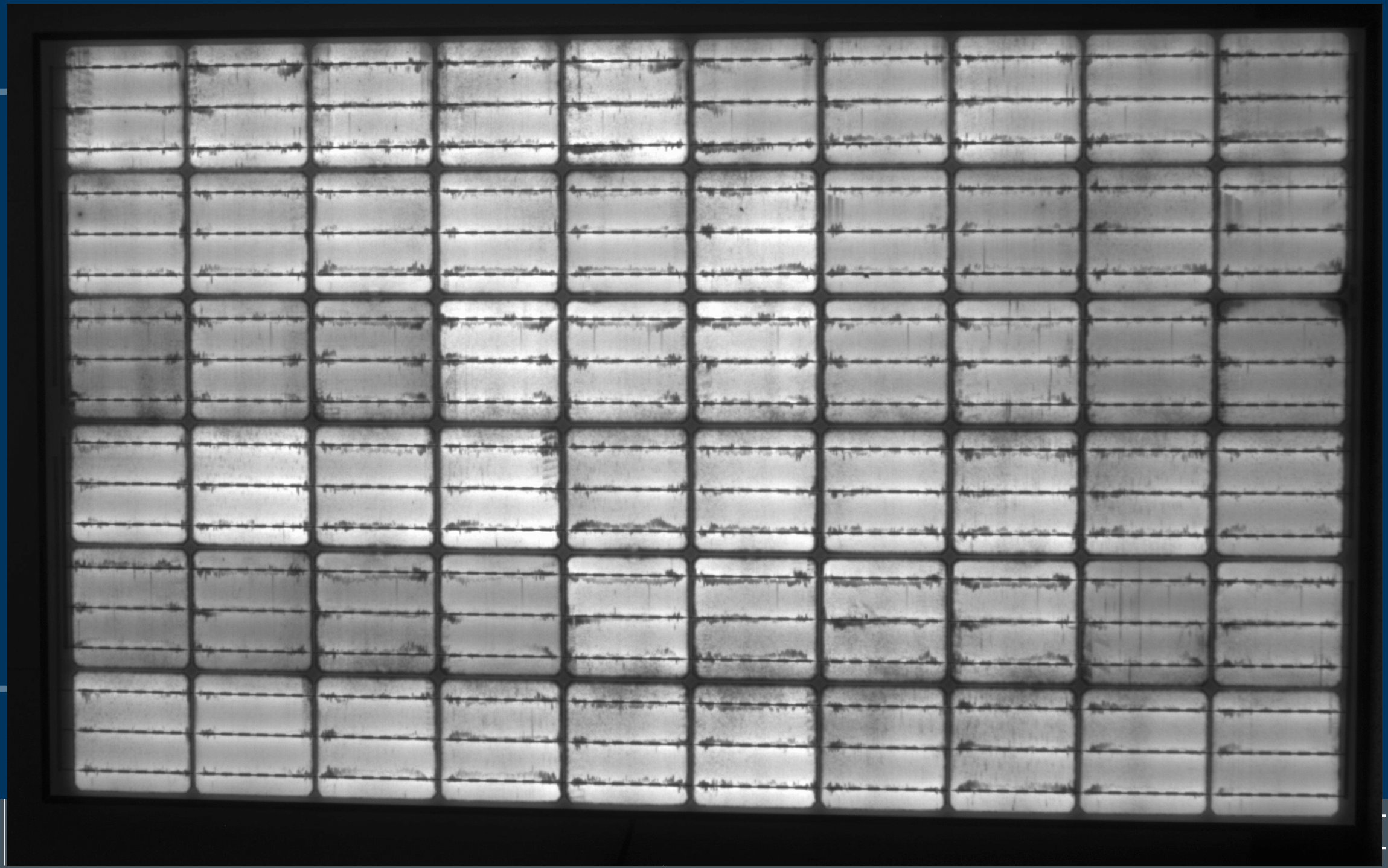
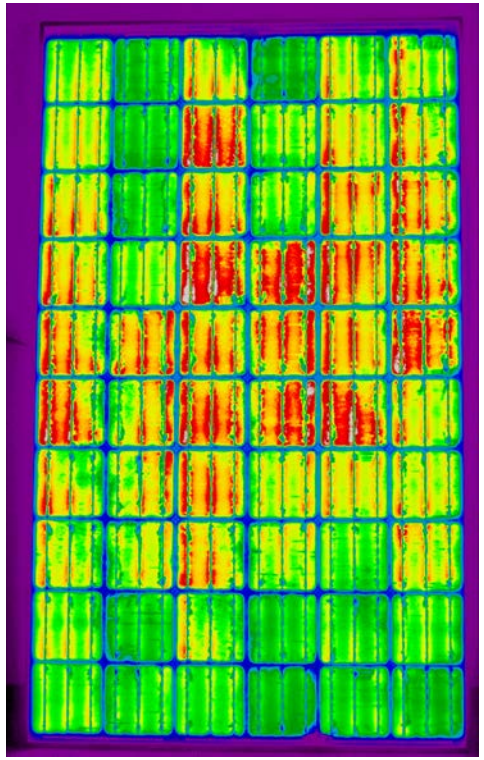
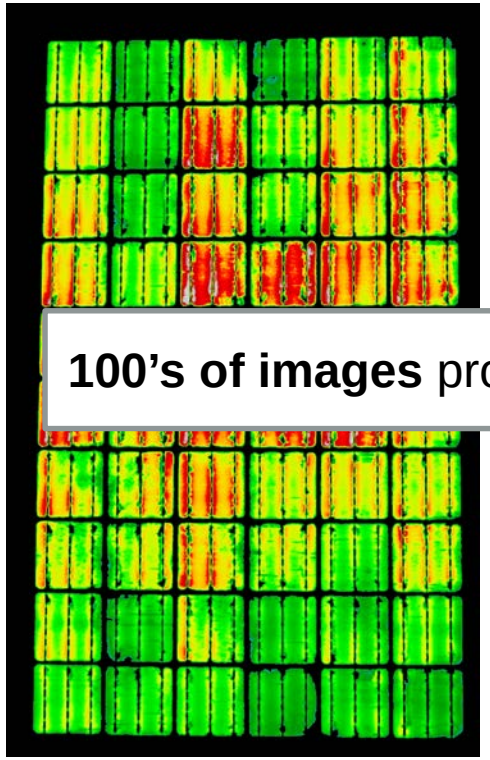
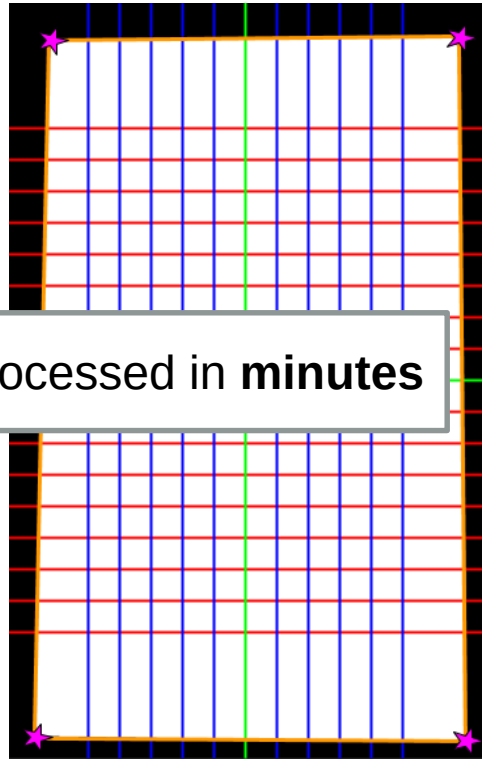
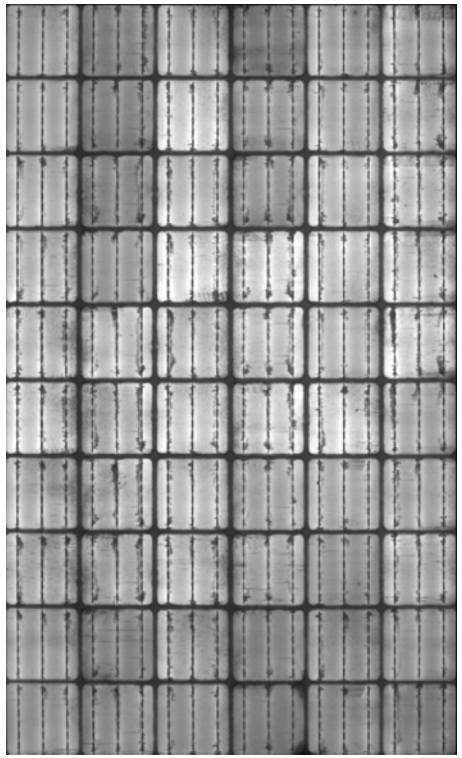


Image Processing Pipeline → Planar Indexed EL Images









100's of images processed in **minutes**

Median Filter
(Spectral Color Mapping)
- Removes high/low values
- Retains edges

Background Threshold
(Spectral Color Mapping)
- Sets non-active cell regions to 0

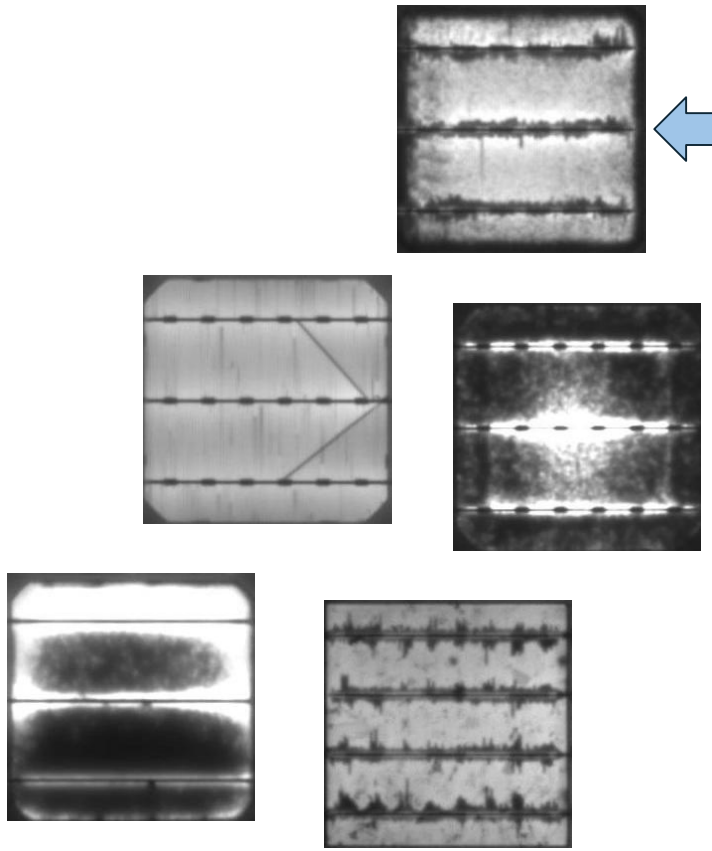
Convex Hull/
Corner Finding
- "Shrink-wraps" the active regions
- Edge finding
- Regression edge point fitting
- Corner finding

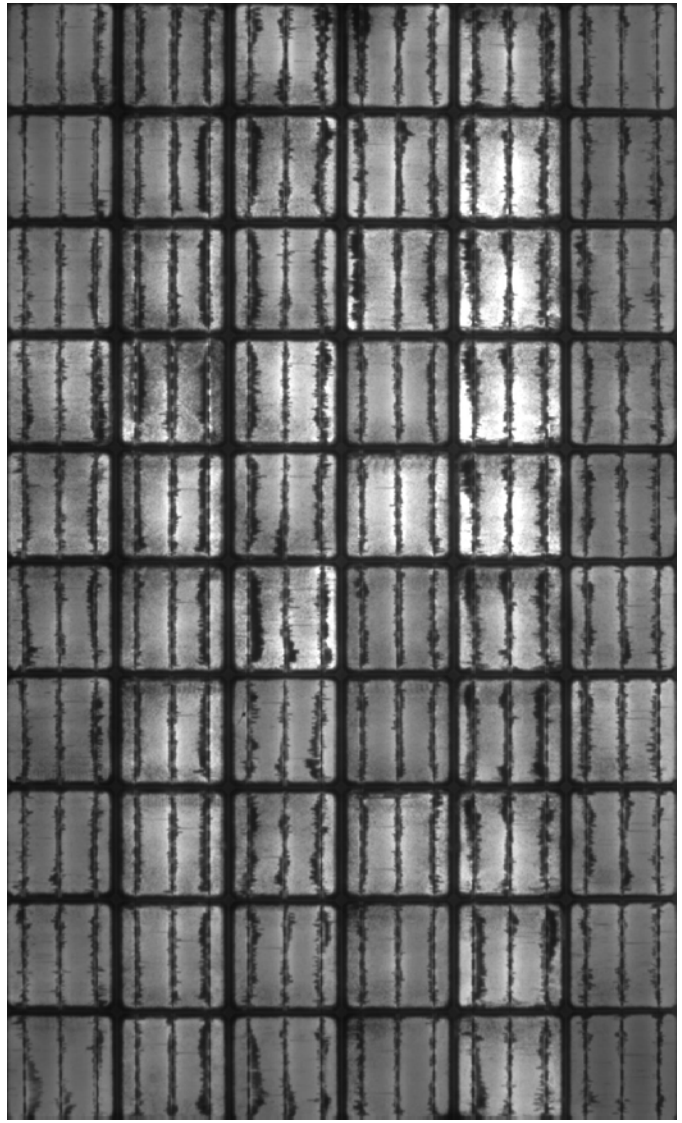
Planar Indexed EL Image
(Perspective Transformed)
- 4 point planar mapping
- Re-orients module

Planar Indexed Module → Individual Cell Images

Cell Extraction

- Starts with planar index module
- Simple matrix slicing used to extract cells
  - Further refined image processing would result in lost information
- Results in single cell images
  - Many features present
  - Resembles face recognition problem





Let’s do something with this..Machine Learning Classification

“the detection of cell cracks by EL imaging has not been automated successfully yet”

– IEA-PVPS Task 13 Review of Failures of PV Modules - 2014

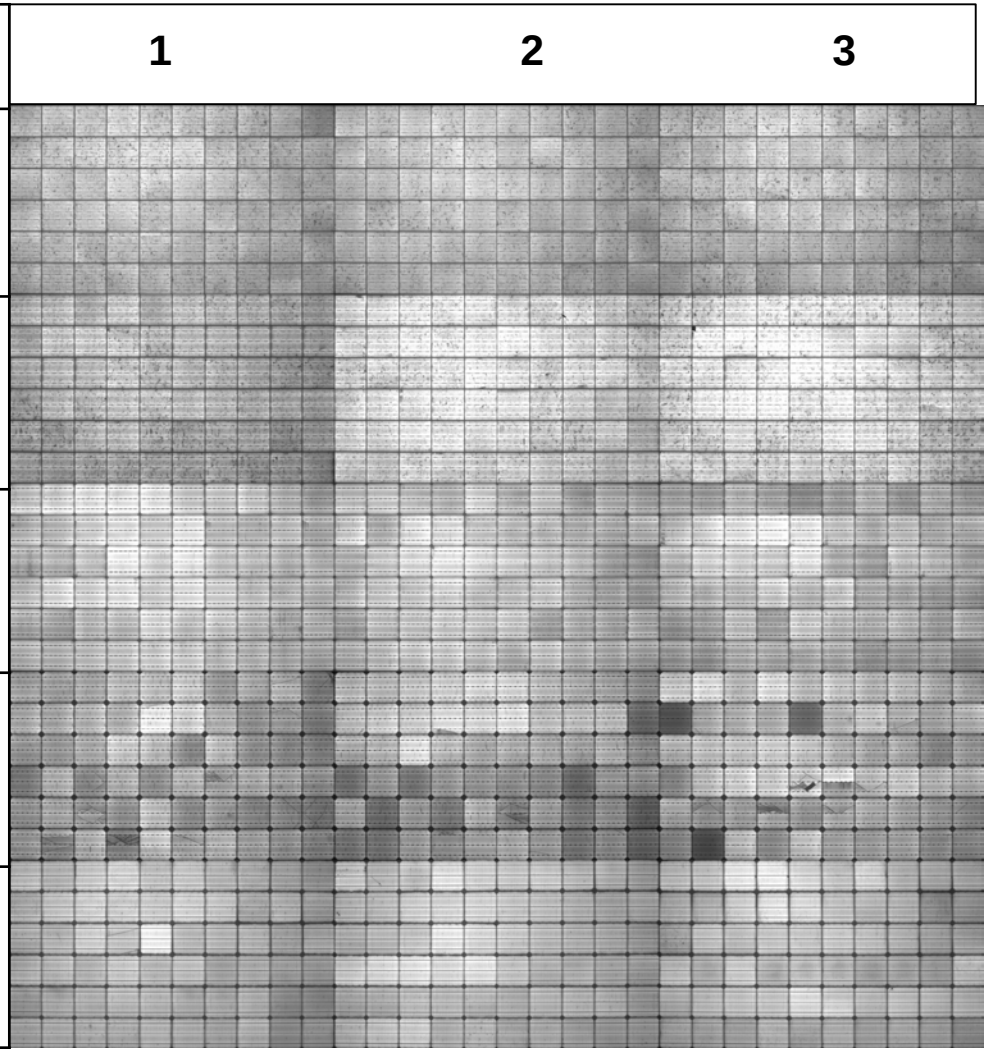
Data Set

- 15 Damp-heat modules
  - 5 Brands (A,B,C,D,E)
  - 3 Samples per brand
  - 3 Wafer types represented
    - Mono-Si
    - Mono-PERC
    - Multi-Si
- Six 500 hour steps
  - 500-3000 hours
  - IEC61215

Machine Learning Classification

- 3 Algorithms
  - Support Vector Machines
  - Random Forest
  - Artificial Neural Network

Brand	Material
A	Multi-Si Al-BSF
D	Multi-Si Al-BSF
B	Mono-Si PERC
C	Mono-Si Al-BSF
E	Mono-Si Al-BSF



M. Kontges *et al.*, “Performance and Reliability of Photovoltaic Systems: Review of Failures of PV Modules,” International Energy Agency, IEA-PVPS Task13-01:2014, Mar. 2014.

Supervised Machine Learning Classification

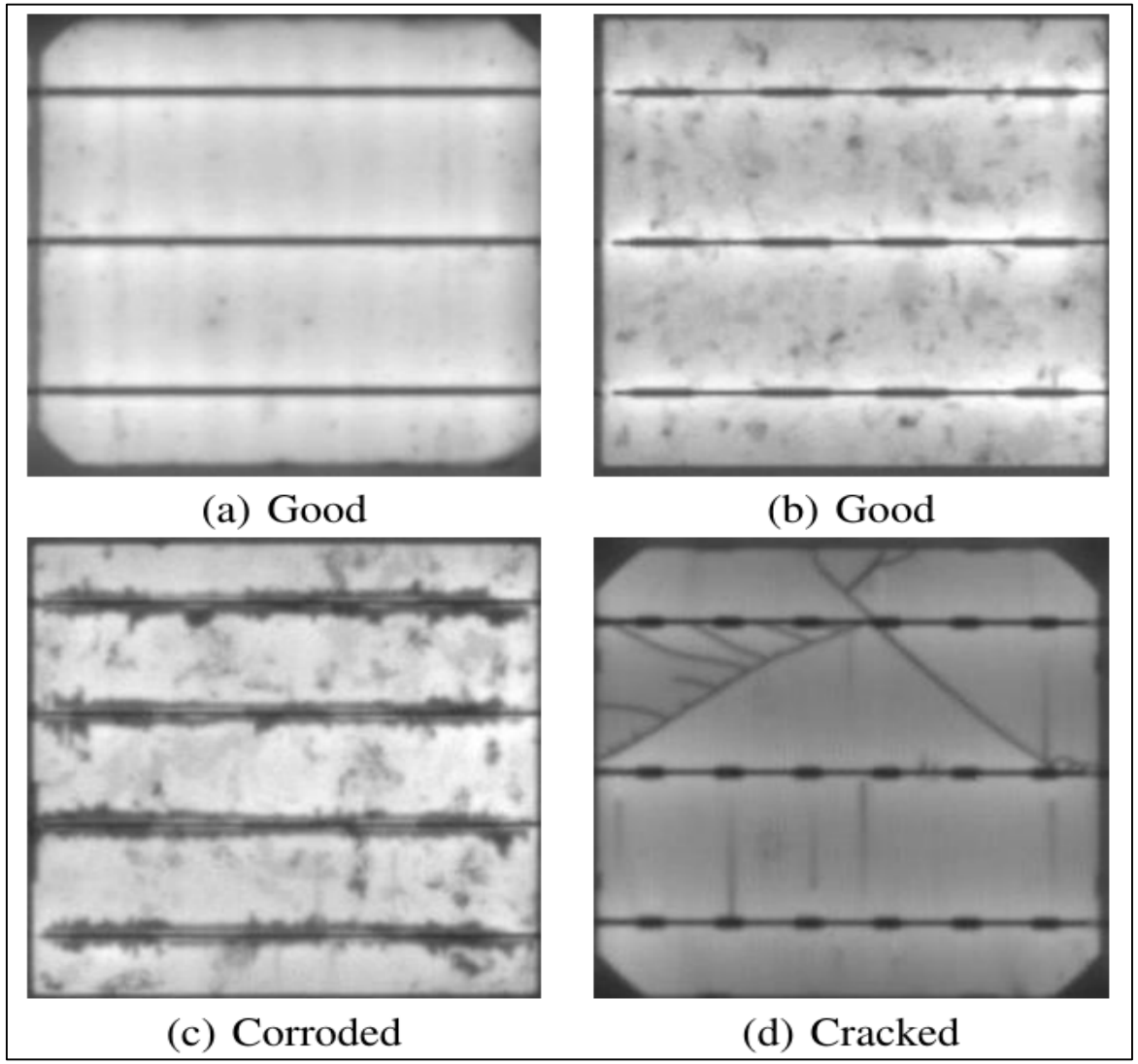
Supervised machine learning

- Uses a **training and testing framework**
- Training set contents are known
- Stratification for testing and training sets
  - Samples chosen by degradation category
  - 80:20 training to testing set ratio

3 degradation categories well represented

- Good/clean
- Busbar corroded
- Cracked

14,200 cell images of the 3 chosen modes



Cell Degradation Mode Classification Results

Support Vector Machine (SVM)

- Separating hyperplanes
- Best for overall accuracy
- Fastest training time
- Best accuracy per feature

Random Forest (RF)

- Decision tree
- High accuracy
- Slower training time
- Low crack accuracy

Artificial Neural Network (ANN)

- Interconnected nodal groups
- High accuracy
- Very slow training

Algorithm Accuracy (%)			
	SVM	RF	ANN
Accuracy (%)	**98.77**	96.90	98.13
Training Time (s)	**85.52**	90.12	2250

Feature Accuracy (%)			
	SVM	RF	ANN
Good	**100**	99	100
Corroded	**99**	98	99
Cracked	**85**	27	78

SVM Optimization

- K-fold cross-validation
- Algorithm parameter optimization (gamma and C)
- Resolution optimization

Time-series Analysis of Real World PV Systems

Yang Hu

Materials Science and Engineering
SDLE Research Center
Case Western Reserve University

Real-world Data Source: CWRU SDLE Global SunFarm Network

SDLE PV Data Covers ~3.4 GW

Encompasses 1.92% of Global PV Power Production

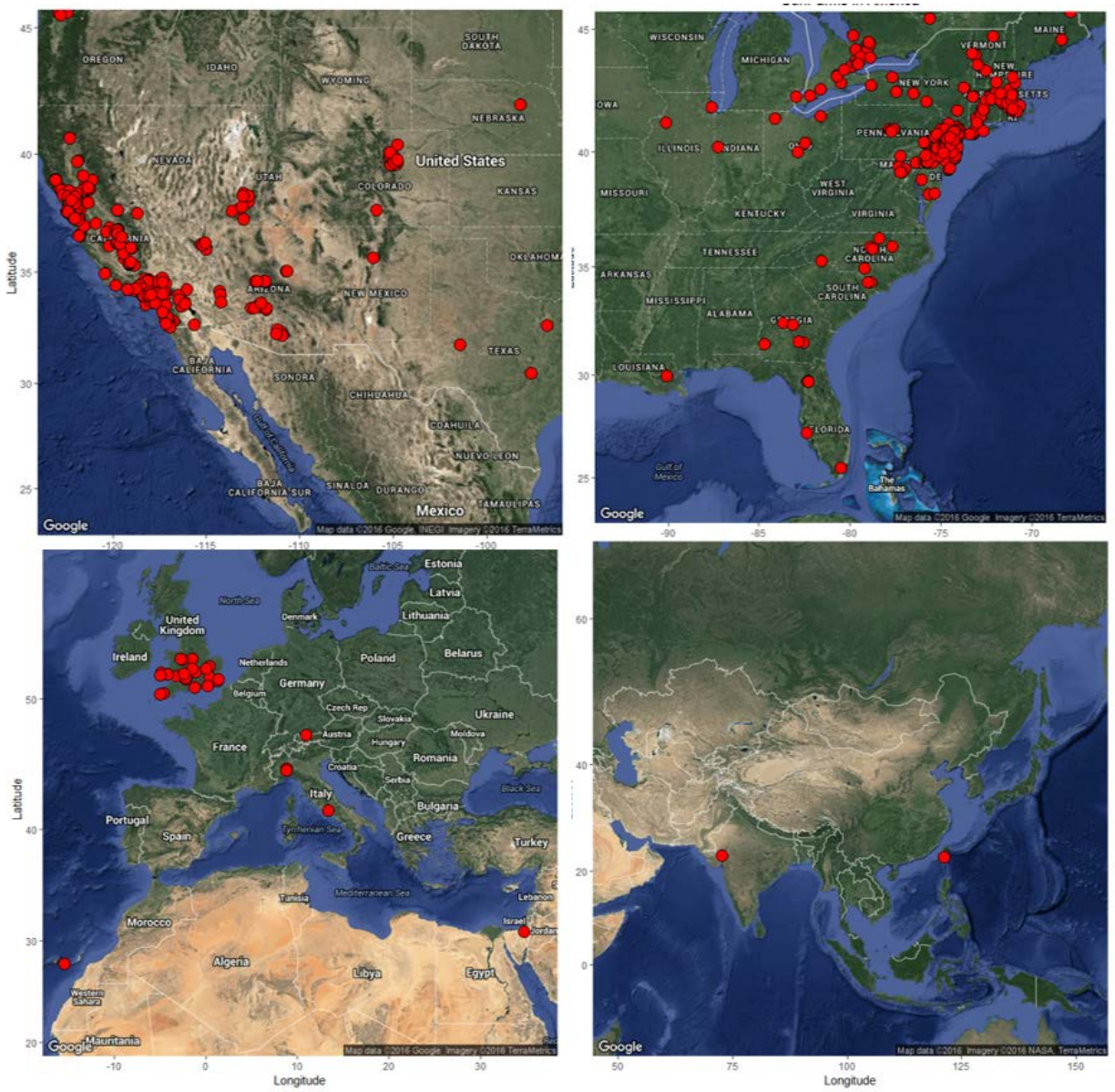
- 787 PV Project Sites
- 5638 PV Systems (Inv. & Modules)
- 60 PV Module Brands/Models
- 38 PV Inverter Brands/Models
- Across 13 Köppen-Geiger Climatic Zones
- Single Modules to 265 MW plants
- Going Back Up To 15 years

Epidemiological PV Populations

- Of Time-series data streams
- Real-world power production
- Real World Exposure Conditions
- Operating Over Real Time-scales

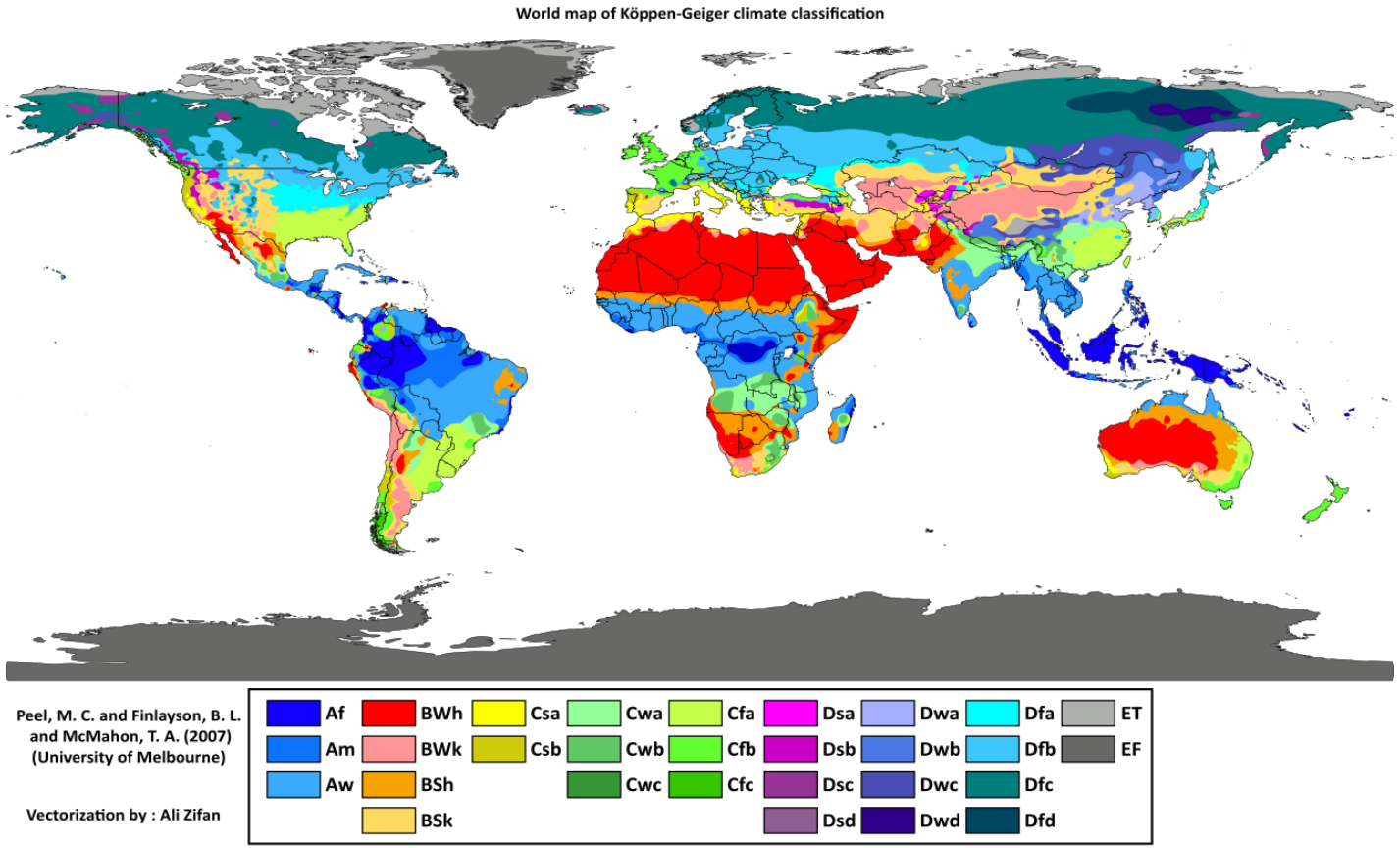
11 Different Companies Have Signed On

- To our Data Use Agreement



Köppen-Geiger Climatic Zones

Type = Humidity/Wetness
Subtype = Temperature/Temperature Range



Generated based on precipitation and temperature

- Begun in 1884, further classified 1954
- Consistent and comprehensive climatic zones

29 total K-G Climatic Zones defined

- Understand environmental stressors

Group	Type	SubType	Description	Criterion
A-Tropical $T_{min} \geq +18$℃	f		Rainforest	Pmin ≥ 60mm
	m		Monsoon	Pann≥ 25(100-Pmm) mm
	w		Savanna	Pmin < 60 mm in winter
B-Arid $P_{ann} < 10\ P_{th}$	w		Desert	Pann ≤ 5 Pth
	s		Steppe	Pann > 5 Pth
		h	Hot	$T_{ann} \geq +18$ °C
		k	Cold	$T_{ann} < +18$ °C
C-Temperate -3 °C < T_{min} < +18 °C	s		Dry Summer	
	w		Dry Winter	$P_{smax} > 10\ P_{wmin}$, $P_{wmin} < P_{smin}$
	f		Without dry season	Not Cs or Cw
		a	Hot Summer	$T_{max} \geq +22$ °C
		b	Warm Summer	$T_{max} < +22$ °C, 4 $T_{mon} \geq +10$ °C
		c	Cold Summer	$T_{max} < +22$ °C, 4 $T_{mon} < +10$ °C, $T_{min} > -38$ °C
D-Cold(Continental) $T_{min} \leq -3$ °C	s		Dry Summer	$P_{smin} < P_{wmin}$, $P_{wmax} > 3\ P_{smin}$, $P_{smin} < 40$ mm
	w		Dry Winter	$P_{smax} > 10\ P_{wmin}$, $P_{wmin} < P_{smin}$
	f		Without dry season	Not Ds or Dw
		a	Hot Summer	$T_{max} \geq +22$ °C
		b	Warm Summer	$T_{max} < +22$ °C, 4 $T_{mon} \geq +10$ °C
		c	Cold Summer	$T_{max} < +22$ °C, 4 $T_{mon} < +10$ °C, $T_{min} > -38$ °C
		d	Very cold Winter	$T_{max} < +22$ °C, 4 $T_{mon} < +10$ °C, $T_{min} \leq -38$ °C
E-Polar $T_{max} < +10$ °C	T		Tundra	$T_{max} \geq 0$ °C
	F		Frost(Ice cap)	$T_{max} < 0$ °C

Peel, M. C.; Finlayson, B. L.; McMahon, T. A. (2007). "Updated world map of the Köppen–Geiger climate classification".Hydrol. Earth Syst. Sci. 11: 1633–1644. doi:10.5194/hess-11-1633-2007. ISSN 1027-5606

KG-CZ Look Up R and Python Package:

Using KG-CZ Datasets from

vetmeduni vienna

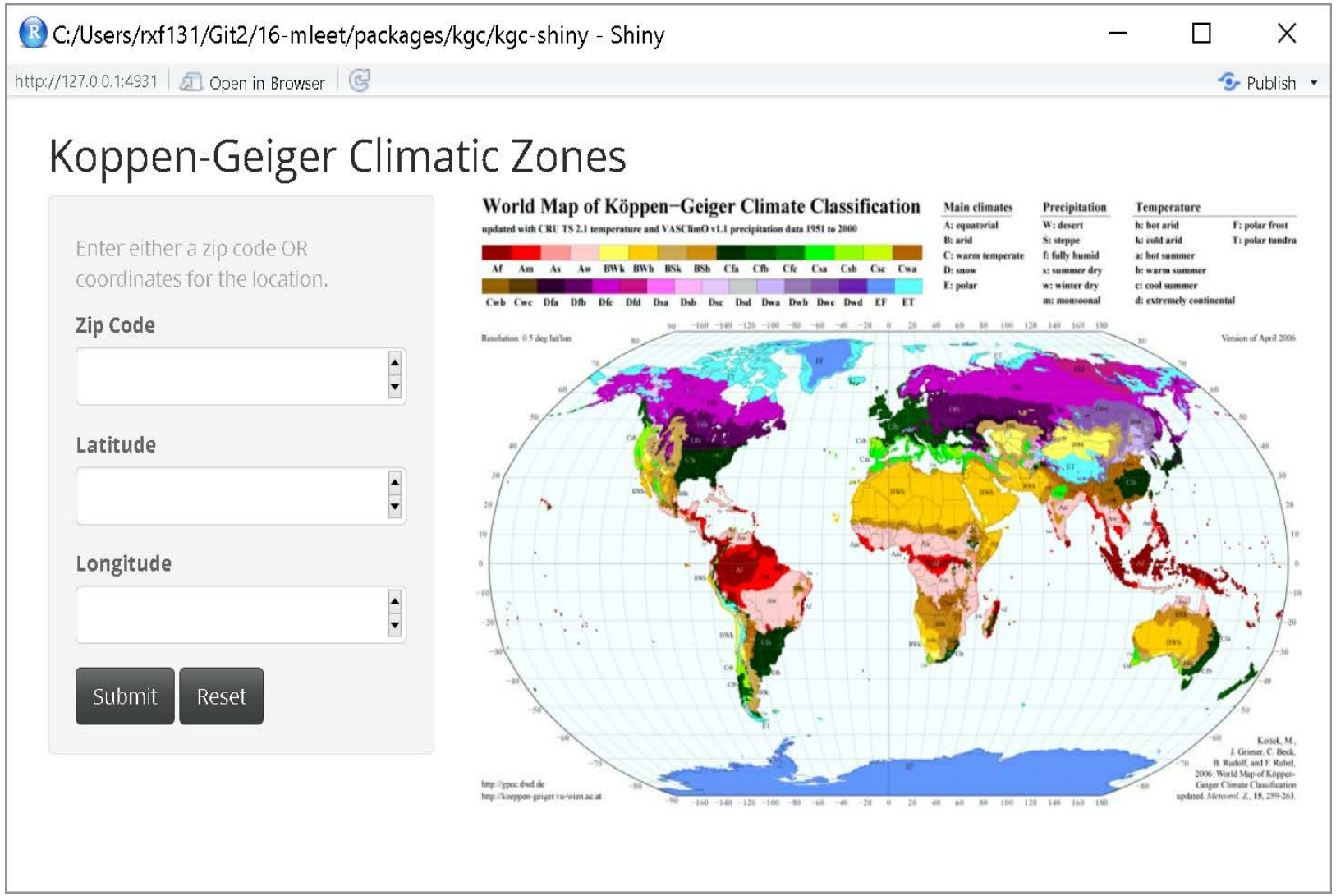
- Provide ASCII lookup table
- Google Earth kmz files
- Gis grid files

Make R package

- To publish on CRAN

With Shiny GUI for single lookups

And Functions for automated lookups



Month-by-Month Change Rate (R_c) Method

Underlying assumption:

- Train an un-biased regression model
- System performance change is a long-term phenomena
  - No obvious degradation within 30 days

Data analytic procedure

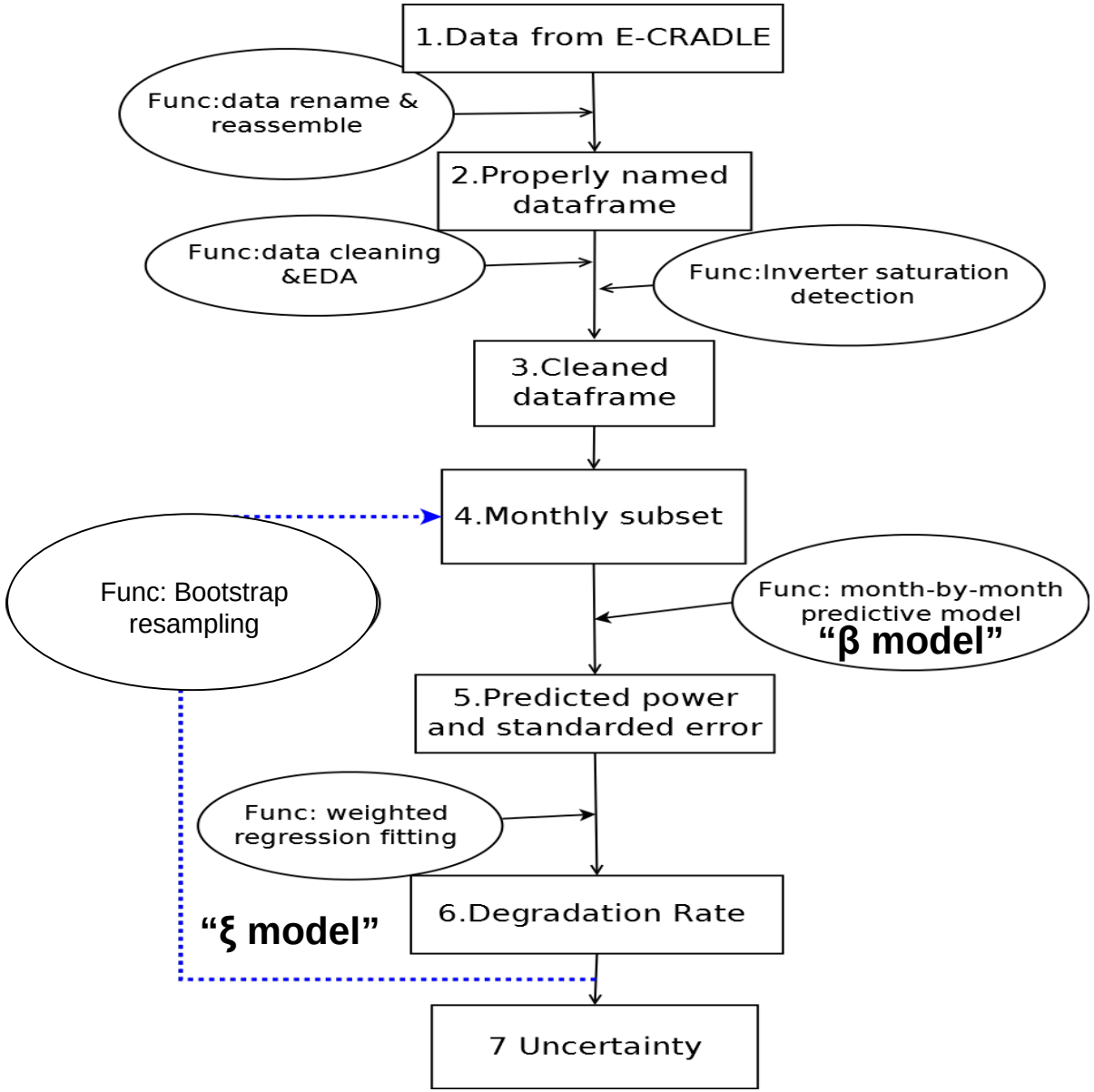
- Use all data, not excessive filtering
- Categorize data by age
  - Every 30 days considered a pseudo-month

Psuedo-Month Predictive model (β model)

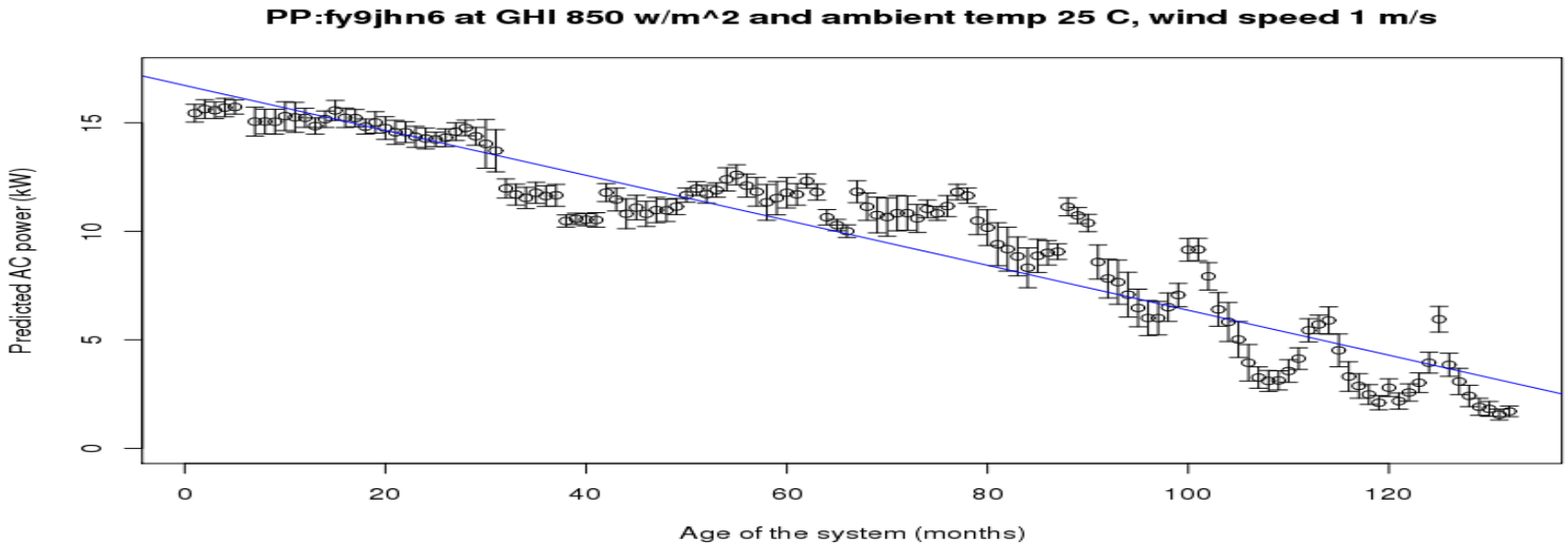
- Linear regression model
- A snapshot of the system status
- Predict system performance each Month
  - To same climate condition
- Use monthly regression models

Longitudinal Regression Model (ξ model)

- Don't assume linear degradation rate
  - Enable Piece-wise Regression Models of Change Rate
- Use bootstrap approach to estimate the uncertainty



Power Plant fy9jhn6: 15 years, BSh Arid-Steppe-Hot



Roc_1 = -2.99%

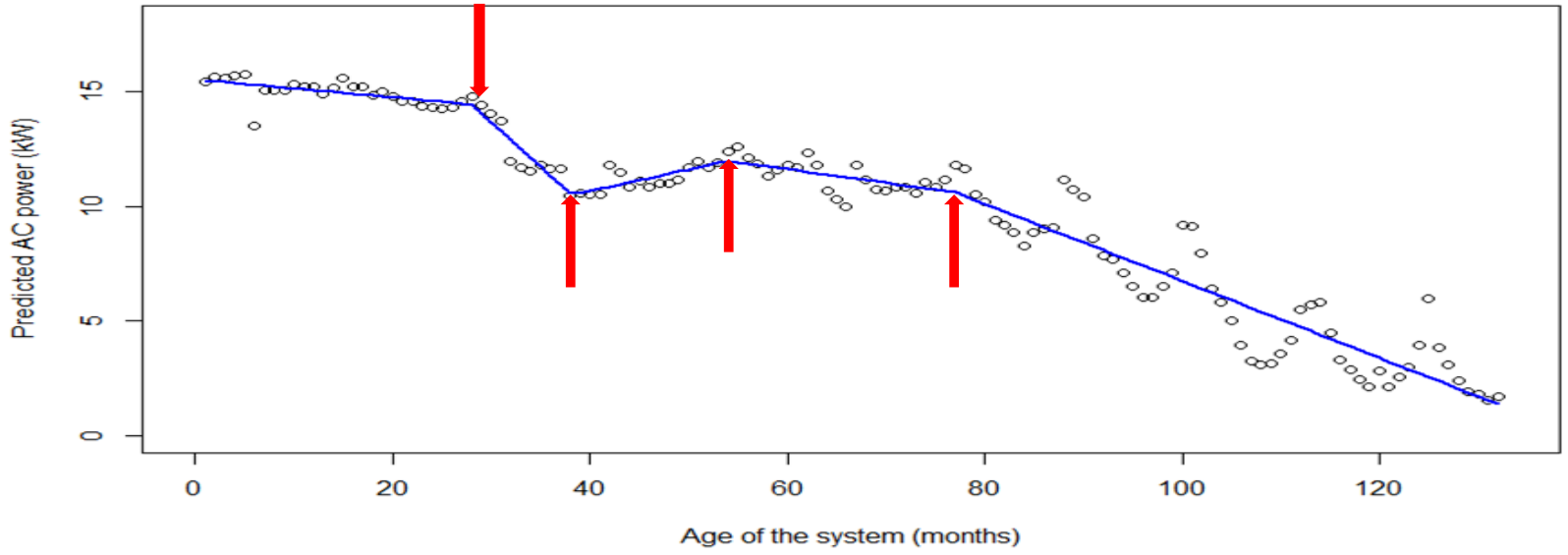
Roc_2 = -20.1%

Roc_3 = 6.31%

Roc_4 = -3.63%

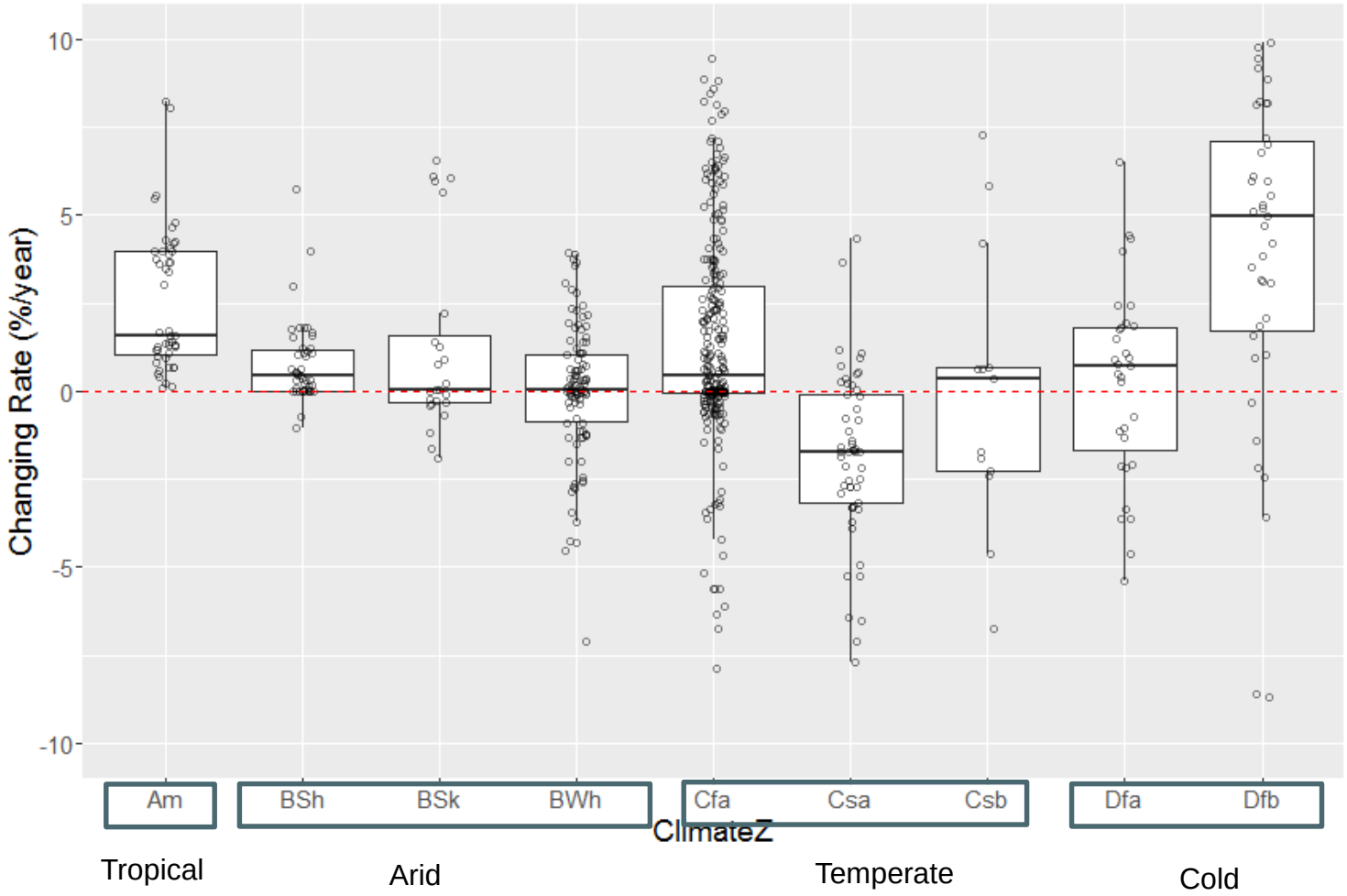
Roc_5 = -11.8%

R^2 of overall fitt 0.95



Now Study a population of PV Systems!

Analysis of variance cross 9 Climate Zones



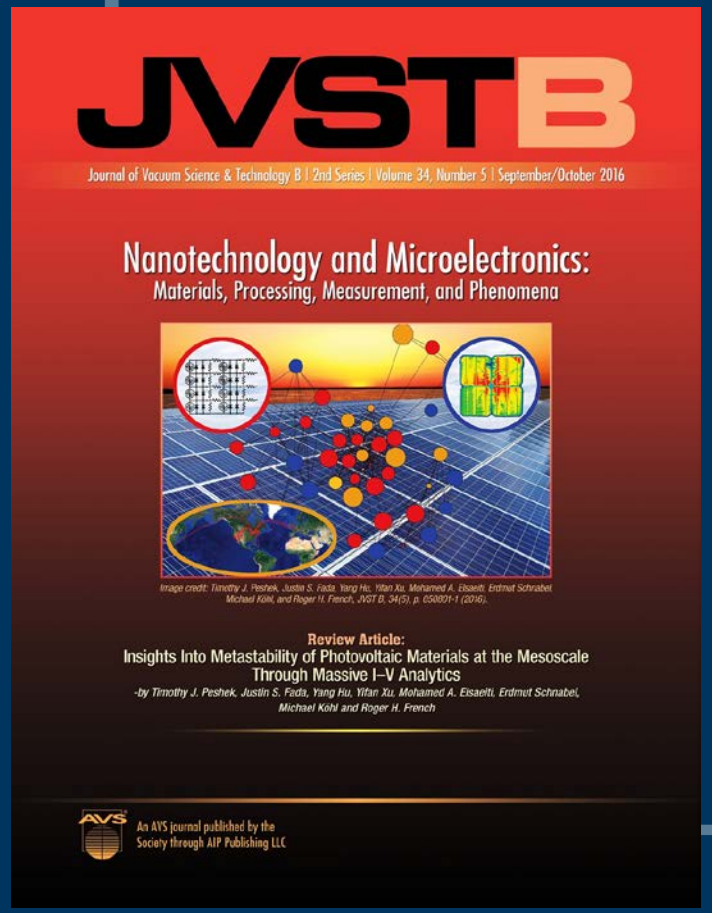
Mean of R_C in each Climate Zone show large variance

- Dfb mean is over +5%/year
- Csa mean is about -2%/year

There may be several confounding variables that influence the changing rate

Develop statistical models help solve the problem.

Machine Learning for I-V,P_{mp} Time-series Datastreams:



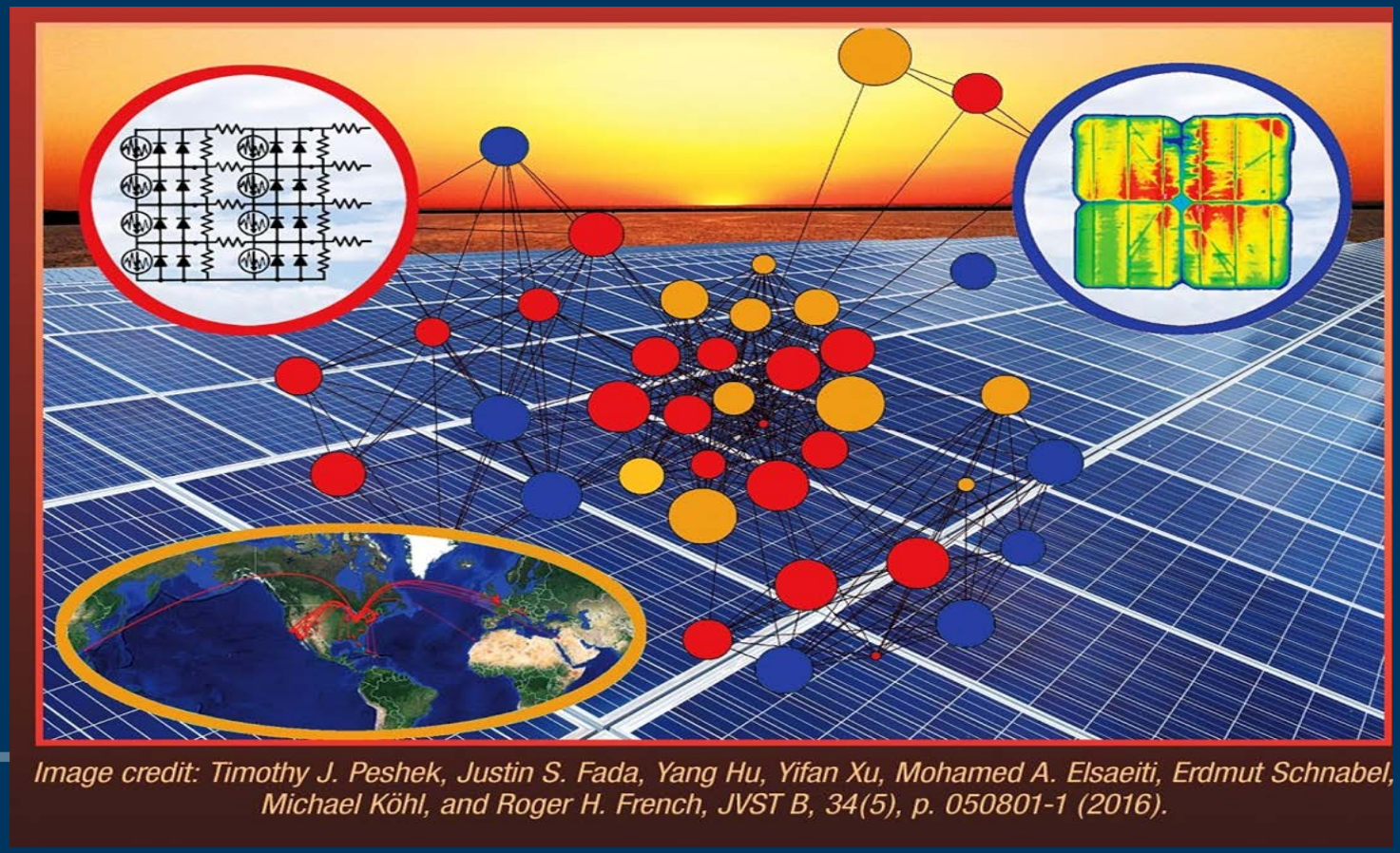


Image credit: Timothy J. Peshek, Justin S. Fada, Yang Hu, Yifan Xu, Mohamed A. Elsaeiti, Erdmut Schnabel, Michael Köhl, and Roger H. French, JVST B, 34(5), p. 050801-1 (2016).

http://dx.doi.org/10.1116/1.4960628

Peshek, T. J., Fada, J. S., Hu, Y., Xu, Y., Elsaeiti, M. A., Schnabel, E., Köhl, M. & French, R. H. Insights into metastability of photovoltaic materials at the mesoscale through massive I–V analytics. *Journal of Vacuum Science & Technology B* **34,** 50801 (2016).

Motivation

PV Modules are modeled as Diode I-V Models

“Step” I-V curves commonly seen, but neglected

- Indoor or outdoor I-V measurements
- “normal” I-V curves
- Smooth half parabolic shaped
- “step” I-V curves
- One or more string bypassed

Hard to describe with a diode model

- Usually eliminated from I-V analysis
- Usually related to partially shading or irradiance change

“Heterogeneity” within PV module

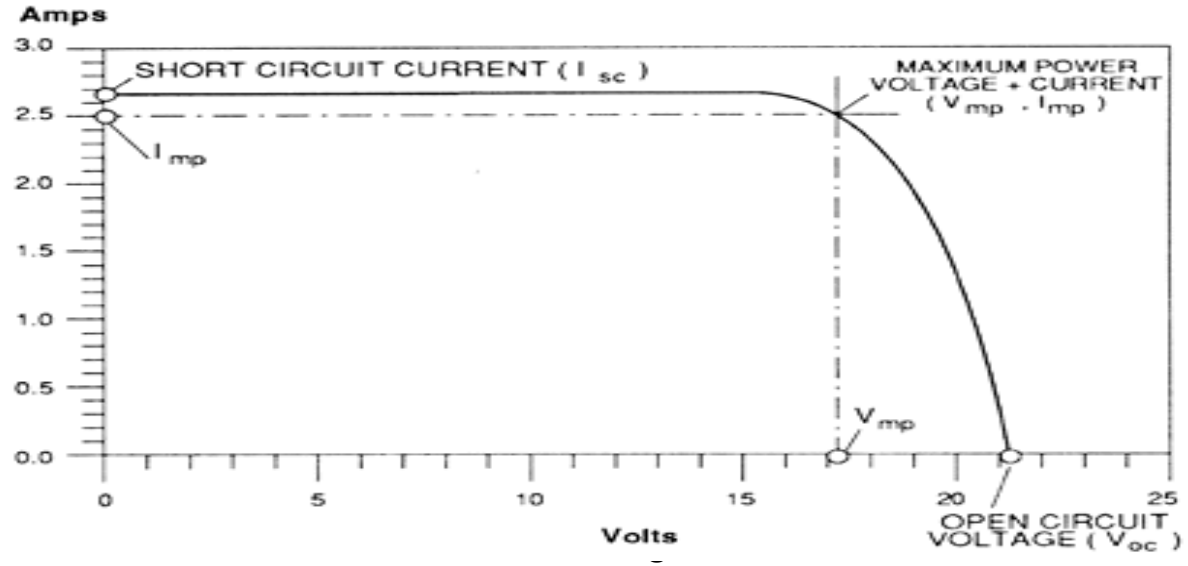
- 60 cells are not identical
- Difference introduced in manufacturing
- Non-uniform degradation

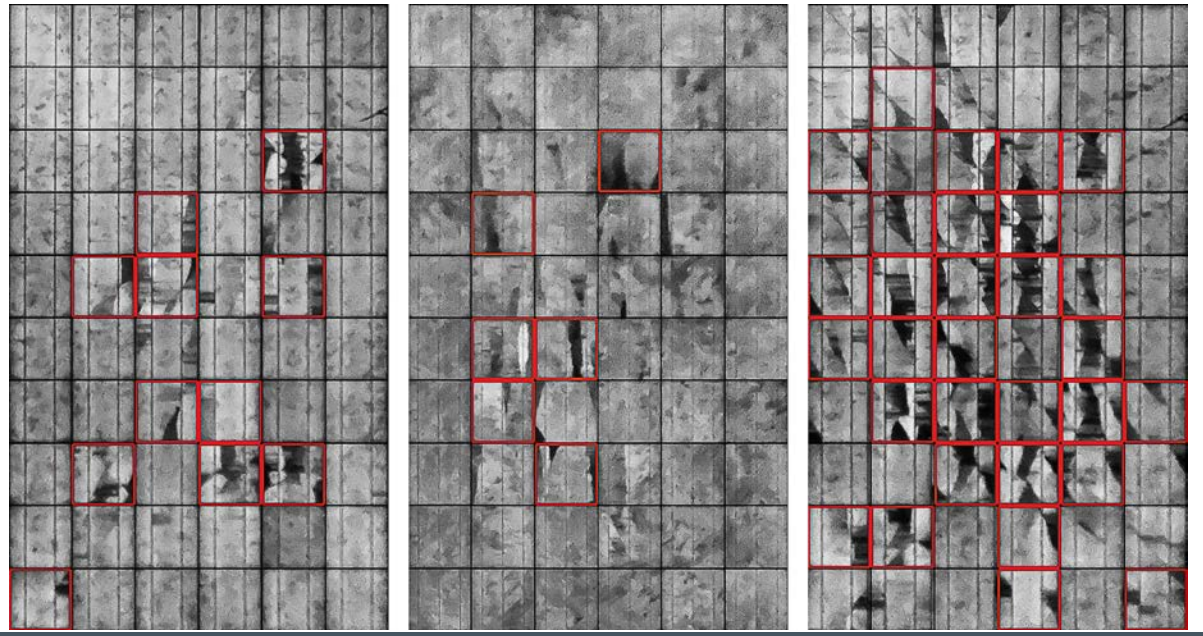
“Step” I-V curve can provide

Usefully insights about heterogeneity in PV module

Classify massive outdoor I-V curves

Non-destructive measurement to detect heterogeneity





I-V, P_{mp} Time-series data stream Sources

Fraunhofer ISE

I-V curve measurement every 5 minutes

P_{mp} measurements in between

Gran Canaria (GC)

- 2010-2015
- Arid-Steppe-Cold (BSK)
- **0.75 million I-V curves**

Mount Zugspitze (UFS)

- 2010-2015
- Polar-Tundra (ET)
- **0.85 million I-V curves**

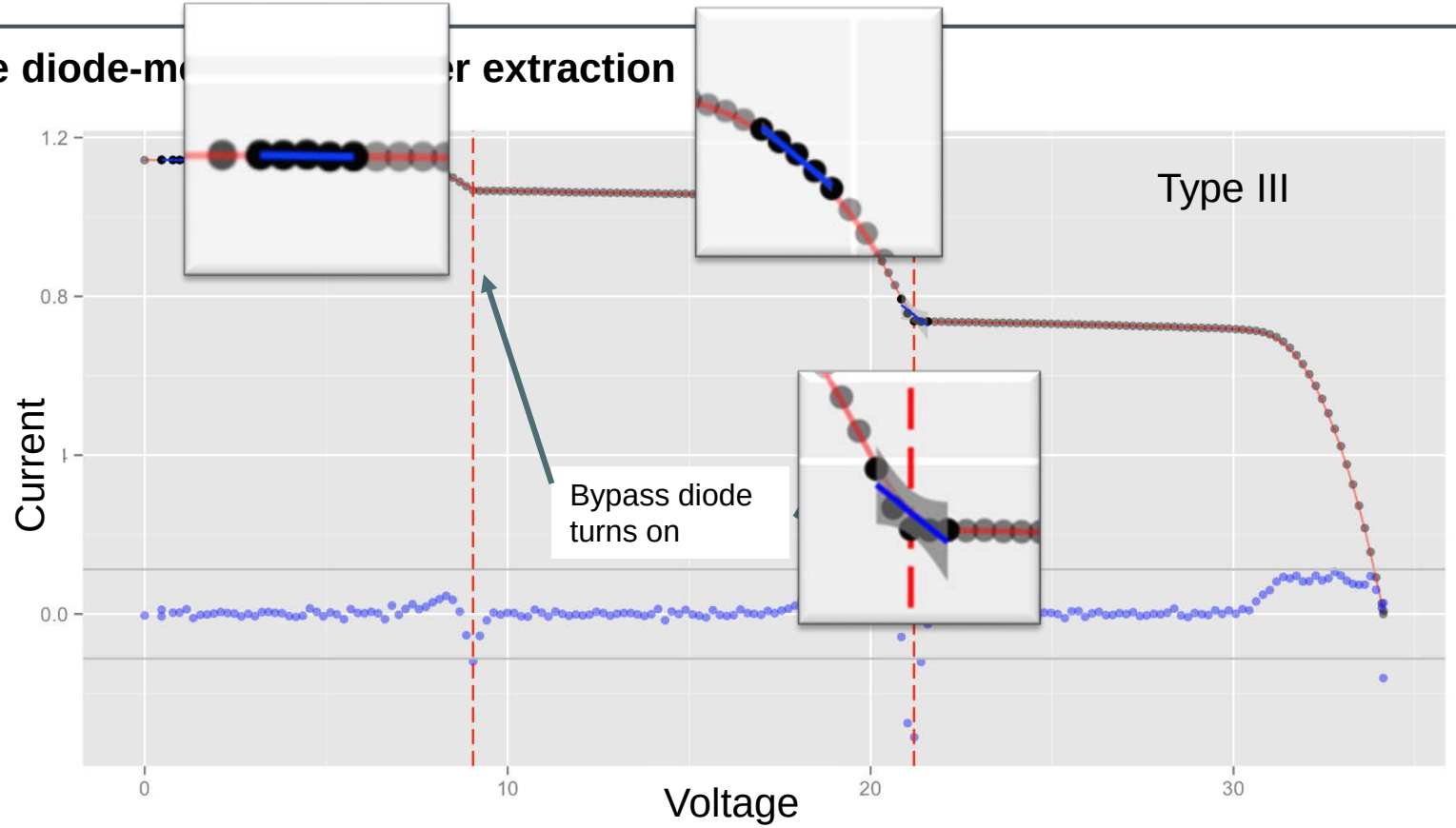
Negev Desert (NEG)

- 2012-2015
- Arid-Desert-Hot(BWh)
- **0.55 million I-V curves**

- Two module samples on each site
- Two different brands on the same site
- The same brands across three sites

Non-parametric Regression: Excess Residual Detection Method

Beyond simple diode-m r extraction



Done using R Analytics and Python for data processing

- In Hadoop-based Energy-CRADLE

Underlying Machine Learning Procedures:

Local linear regression fitting + Residual Thresholding

- (Parameters are tuned on a training set)

Classify I-V curve into five categories

- **Type I** : V_{oc} only
- **Type II** : V_{oc} + one bypass diode turns on
- **Type III** : V_{oc} + two bypass diodes turn on
- Few.points : less than 10 measured points
- Low.amps : I_{sc} lower than 1 amp

Fraunhofer ISE sites I-V curve study: Type I curves

Classification of 2.2 million I-V curves

- Accomplished using CWRU's
- High performance computing cluster
- 12 cores CPU 8G of memory
- About 4.5 hours
- Type I II and III 40-45%
  of the whole data set

Proportion of type I curves

- 6 samples
- At all three locations

In 2010 and 2011

- I-V curves are almost 100% Type I
- All four samples on site GC and UFS
- From 2012 to 2015 the proportion of Type I
- Start dropping on both GC and UFS sites
- On NEG site, for sample 311,
  - almost 100% Type I on 2012

- No artificial shading or irradiance augmentation on these three sites
- Type I curve in the first year >99.5%, except sample 328
- Only transient shading/irradiance change during I-V measurement
- Declination of the percentage of Type I curve is related to degradation

Energy & Materials Data Science: Encompassing Broader Opportunities

Where we started: Lifetime and Degradation Science

- Focusing on PV Modules Degradation Over 25 year
  - Now Shifting Focus to 50 years
  - And To High Efficiency c-Si PERC Modules

Expanding Across All Data Types

- Time Series Analysis of Power/Energy Data: Power Plants, Building Energy Efficiency
- Spectral Analysis: Materials Degradation and Mechanistic Identification
- Image Processing: Electroluminescent, Thermographic, Optical, Video Images

Expanding Beyond Time-series Analysis and Network Modeling

- Machine Learning
- Ensemble Modeling
- Deep Learning
- From NoSQL Databases, to NoSQL Document Databases

Expanding Beyond Long Term Degradation, Into Data-driven Analysis & Modeling

- Solar Irradiance Forecasting
- Research and Data Text Mining
- Information Security and CyberSecurity (VerisDB)

CASE SCHOOL
OF ENGINEERING

CASE WESTERN RESERVE
UNIVERSITY