Applied Data Science:

Statistical & Machine Learning, Deep Learning and Artificial Intelligence

Roger H. French

SDLE Research Center
Materials Science & Engineering Department
Case Western Reserve University, Cleveland OH 44106 USA

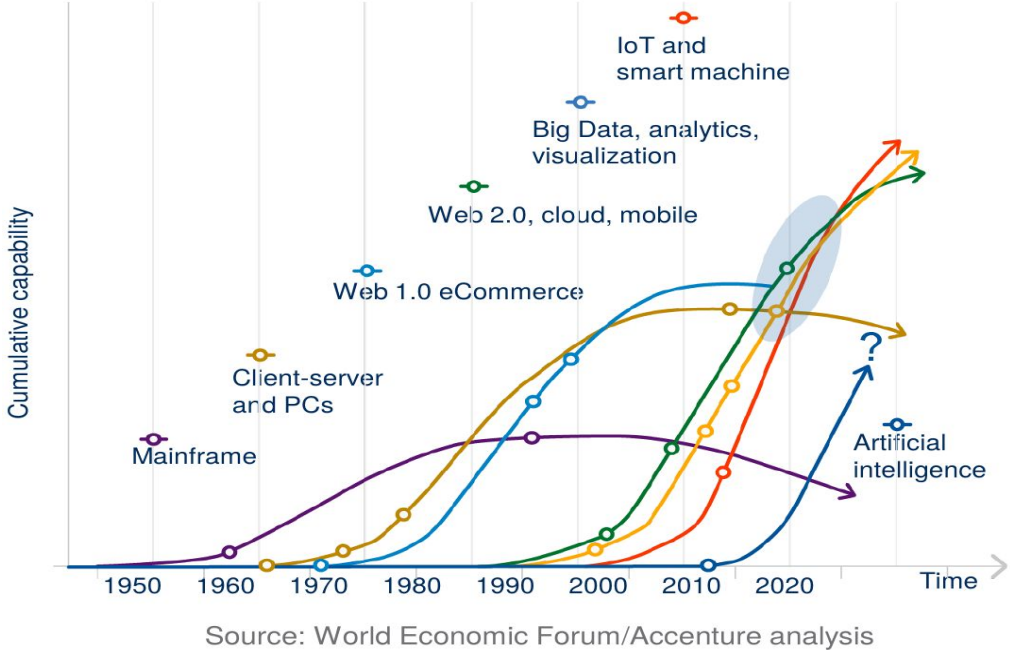
http://sdle.case.edu

Digital Transformation: Combinatorial Effects of Tech. are Accelerating Change

The falling cost of advanced technologies

- A defining characteristic of digital revolution
  - Computing
  - Internet Communications
  - Data Storage

Major role in driving digital transformation



Source: World Economic Forum/Accenture analysis

Examples of the falling cost

- of key technologies

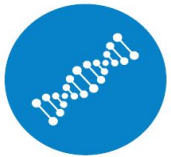
Drones

Cost per unit

2007 **$100,000**

2013 **$700**

DNA Sequencing



Cost per unit

2000 **$2.7 billion**

2007 **$10 million**

2014 **$1,000**

Solar

Cost per kWh*

1984 **$30**

2014 **$0.16**

Industry 4.0: The 4th industrial Revolution

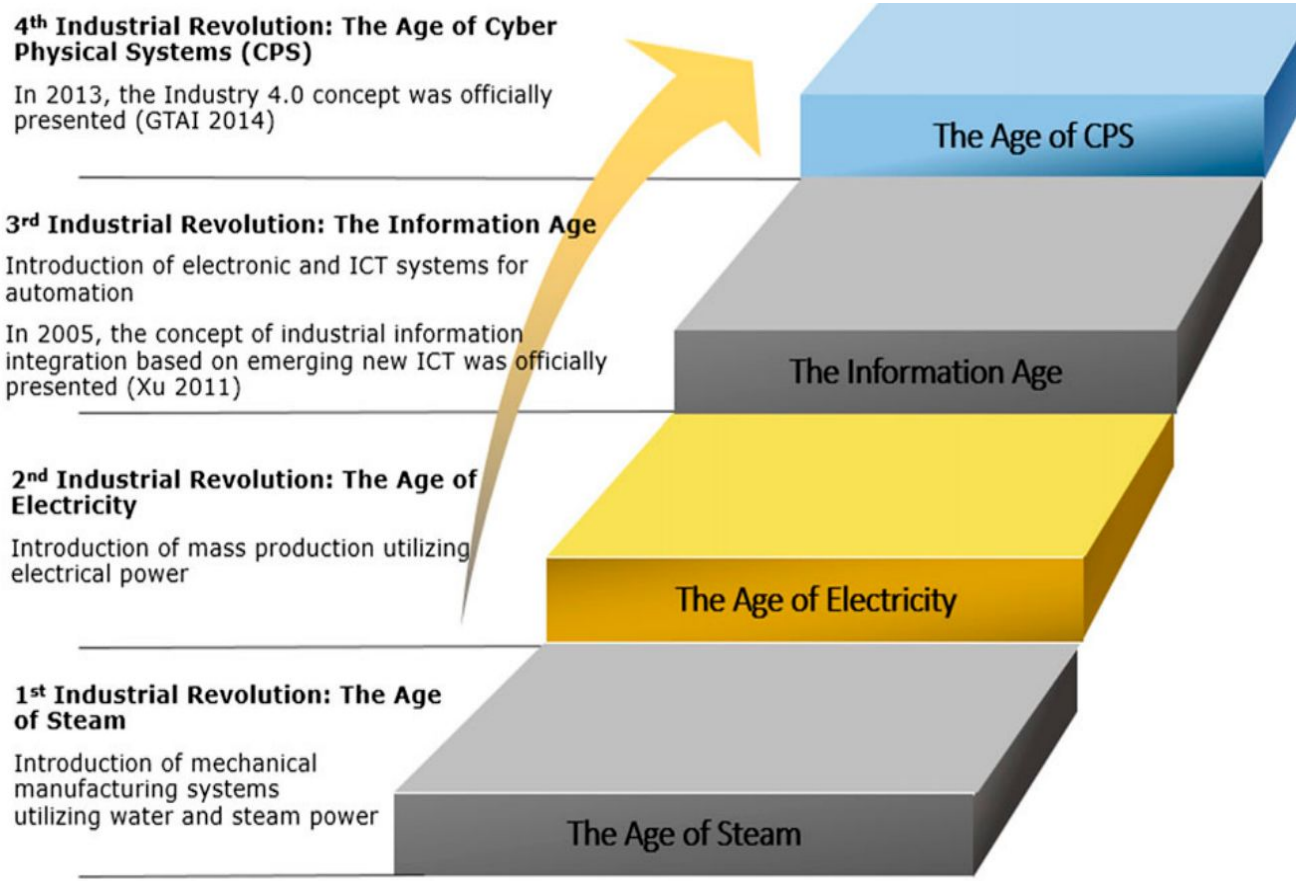
Digital Transformation

- Is transforming Industry
- Into Industry 4.0

Digital Technologies

- Will provide new flexibility
- And Industrial efficiency

Opening new opportunities



Y. Lu, Industry 4.0: A survey on technologies, applications and open research issues, Journal of Industrial Information Integration. 6 (2017) 1–10. https://doi.org/10.1016/j.jii.2017.04.005 .

CWRU Applied Data Science UG/Grad Program

THROUGH THE COLLABORATION of its business and higher education members, the Business-Higher Education Forum (BHEF) launched the National Higher Education and Workforce Initiative (HEWI) to create new undergraduate pathways in high-skill, high-demand fields such as data science and analytics. Data science and analytics must be integrated with T-shaped skills, such as critical thinking, collaboration, and effective communication, which are critical for all graduates entering the 21st century workforce. Knowledge of data science and analytics in recent years has become as fundamental as any other skill for graduates' career readiness. BHEF's Strategic Business Engagement Model with higher education addresses this demand by moving the two sectors from transactional relationships to strategic partnerships through five strategies:

1. **ENGAGE** corporate leadership;
2. **FOCUS** corporate philanthropy on undergraduate education;
3. **IDENTIFY** and tap core competencies and expertise;
4. **FACILITATE** and encourage employee, faculty, and staff engagement;
5. **EXPAND** the focus of funded research to include undergraduate education.

This case study examines how BHEF member Case Western Reserve University (Case Western Reserve) is integrating T-shaped skills into a minor in applied data science.

PROGRAM OVERVIEW

THE APPLIED DATA SCIENCE (ADS) MINOR AT CASE WESTERN RESERVE serves as a national model for undergraduate education in data science. Available to every undergraduate student across all schools at the university, this program of study requires experiential learning opportunities, embeds T-shaped skills, and allows students to master fundamental ADS concepts in their chosen domain area. From strong leadership engagement to funded undergraduate research opportunities, Case Western Reserve applied BHEF's Strategic Business Engagement Model to create a minor that responds to the fundamental need for data science in today's global business community.

Medical Mutual of Ohio
Medtronic
Philips Healthcare
Sherwin-Williams Company
Siemens
Teradata Corporation
Timken Company
University Hospitals

http://www.bhef.com/publications/creating-minor-applied-data-science

CWRU Applied Data Science UG/Grad Program

A BHEF CASE STUDY

CREATING A MINOR IN

APPLIED DATA SCIENCE

Case Western Reserve University Engages B
Leaders to Produce T-Shaped Professionals

The New York Times | https://nyti.ms/2sSkAVI

With Innovation, Colleges Fill the Skills Gap

By JOHN HANC JUNE 7, 2017

How large is the so-called skills gap?

The Manpower Group, a human resources consulting firm, says the gap, which is often defined as the difference in job skills required and the actual skills possessed by employees, is a chasm. Of the more than 42,000 employers the firm surveyed last year, 40 percent said they were having difficulties filling roles, the highest level since 2007.

Case Western Reserve University

Creating 15- or 18-credit minors may be one of the more effective strategies for preparing students to enter high-demand fields. Because a minor requires fewer credits than a major and few, if any, prerequisites, these allow colleges to be more flexible and responsive to changing industries and emerging technologies.

Case Western's minor in applied data science, for example, funnels students into this hot field from other disciplines. The students learn skills like data management, distributed computing, informatics and statistical analytics.

OVERVIEW

) MINOR AT CASE
ional model for
ience. Available to
all schools at the
quires experiential
haped skills, and
ntal ADS concepts
strong leadership
ate research
ve applied BHEF's
del to create a
ental need for data
ommunity.

l Mutual of Ohio
nic
Healthcare
n-Williams
ny
ıs
a Corporation
Company
ity Hospitals

CWRU's Applied Data Science Program: Undergraduate Minor, Graduate Certificate

Applied Data Science program

- Undergraduate Minor
- Graduate Data Science Certificate

Developed in 2014, Courses Started in 2015

- In collaboration with Business Higher Ed. Forum
- UG & Grad Course Section
  - DSCI351/451
- Now have Materials Data Science M sections
  - DSCI351M

Applied/Materials Data Science

- DSCI351M: Exploratory Data Analysis
- DSCI353M: Modeling, Prediction, Machine Learning
- DSCI352M: Materials Data Science Res. Proj., POSEV
- DSCI354: Data Visualization and Analytics

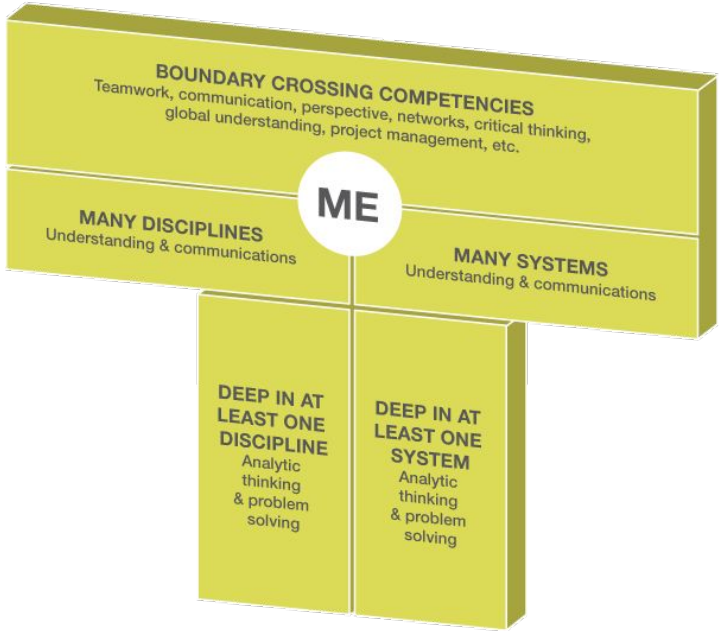
Specialization Courses

- DSCI354M: Data Visualization & Analytics
- DSCI430, Cognition and Computation
- DSCI432, Spatial Statistics for Subsurface Modeling

Now 70 to 80 students per semester

Developing T-shaped <u>Undergraduates</u>

- Deep domain knowledge (in Materials Science)
- Broad knowledge in Applied Data Science



D. Hughes, R. H. French. Crafting a Minor to Produce T-Shaped Graduates. T-Summit 2016, Washington DC, March 21, (2016). at http://tsummit.org/

Components of Applied Data Science Curriculum

Applied/Materials Data Science Core Courses

- DSCI351M: Exploratory Data Analysis
- DSCI353M: Modeling, Prediction, Machine Learning
- DSCI352M: Materials Data Science Res. Project
  - For their GitHub "Portfolio"
- DSCI 354: Data Visualization & Analytics

POSEV Concepts

- Privacy, Openness, Security, Ethics, Value

Taught from "Structure of a Data Analysis" Perspective

Agile Software Development Tools & Approach

Knuth's Literate Programming Perspective

- Integrate Code and Report Writing
- Rmarkdown, Jupyter Notebooks

Textbooks (Open Access)

- Open Intro Statistics
- Introduction to Statistical Learning with R, 2nd Edition

Taught using a Practicum Approach

Each class has two parts

- **Foundation:**
  - Statistics, Regression, ML, Time series ...
- **Practicum**
  - Code Style and commenting
  - Pipelines and Pipe operators
  - Data structures and data frames

Coding/Programming Language

- R with Rstudio IDE
- Python with Spyder (or Jupyter notebooks)

Open Data Science Toolchain

- (cross platform: Linux, Mac, Win)
- R, Python
- Rstudio, Spyder
- Markdown, Rmarkdown
  - Jupyter Notebooks for Python, R
- LaTeX engine, TexStudio
- Chrome, Firefox, html

“Structure of a Data Analysis” Perspective, For SemProj’s & Class Practicum

Part a) Define Question

- Background on the research area & critical issues
- Define the question
- Define the ideal data set
- Determine what data you can access
- Define critical capabilities, identify packages you will draw upon
- Obtain the data, define you target data structure
- Clean and tidy the data

Part b) Cleaning and Exploratory Data Analysis (EDA)

- Write your databook, defining variables, units and data structures
- Data visualization and exploratory data analysis
- Observations of trends and functional forms
- Power transformations
- Validate with reference to domain knowledge
- Evaluate the types of Modeling Approaches to take

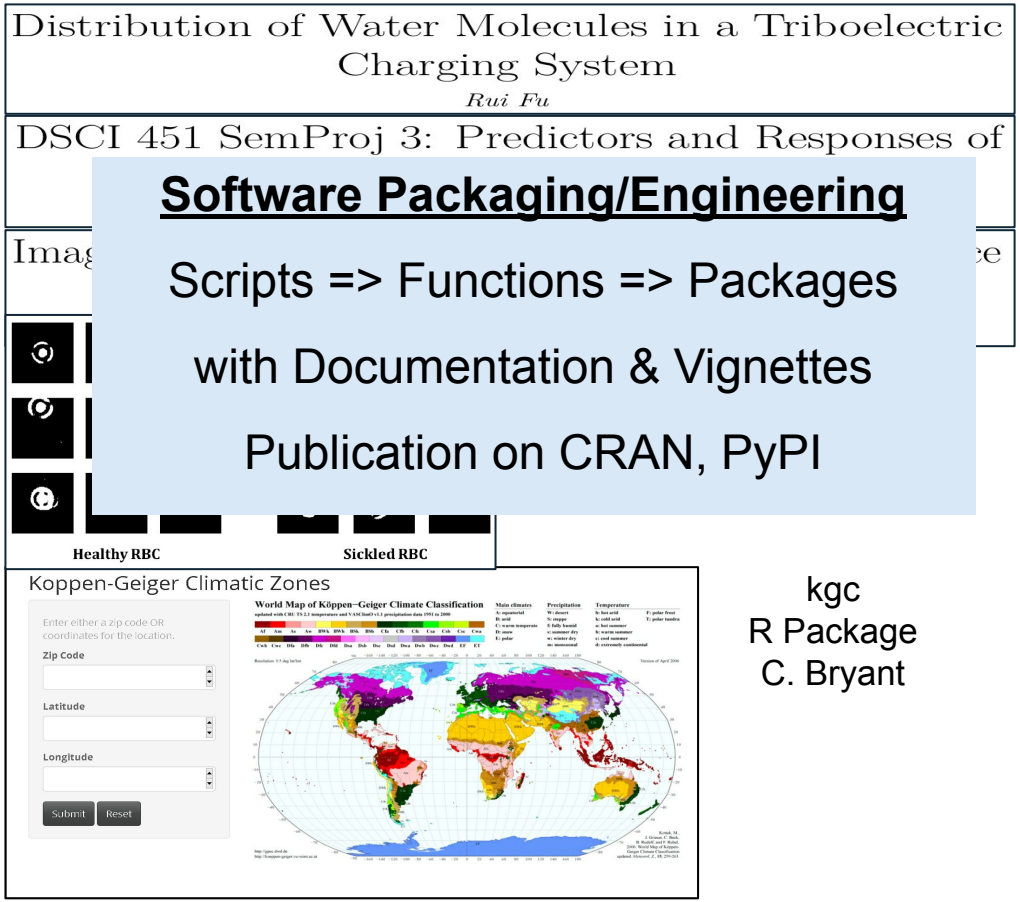
Part c) Modeling, Prediction, Machine Learning

- Types of modeling to try
- Statistical prediction/modeling
- Model selection
- Cross-validation, Predictive R2
- Interpret results
- Challenge results

Part d) Present Your Final Models and Learnings

- Present your results
- Present reproducible code
- Comparison to literature modeling approaches

Jeff Leek, JHU, Data Analytic Style



kgc
R Package
C. Bryant

Open Data Science Tool Chain

Using Open Source, Agile Tools

- Manifesto for Agile Software Development

Reproducible Research

- Using Rmarkdown reports
- Python/R Jupyter Notebooks
- When data updates
- Recompile your report
- All new figures and report!
- Well Documented Codes & Reports

High Level Scripting Languages: R, Python

- Use Machine Learning Frameworks
- Such as Keras/TensorFlow for Deep Neural Networks

Rstudio Integrated Development Environment

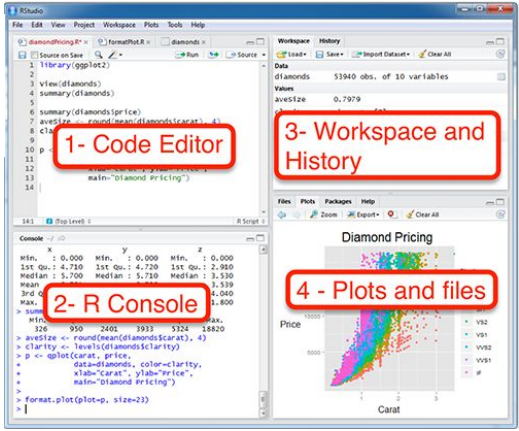
- Spyder IDE for Python

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



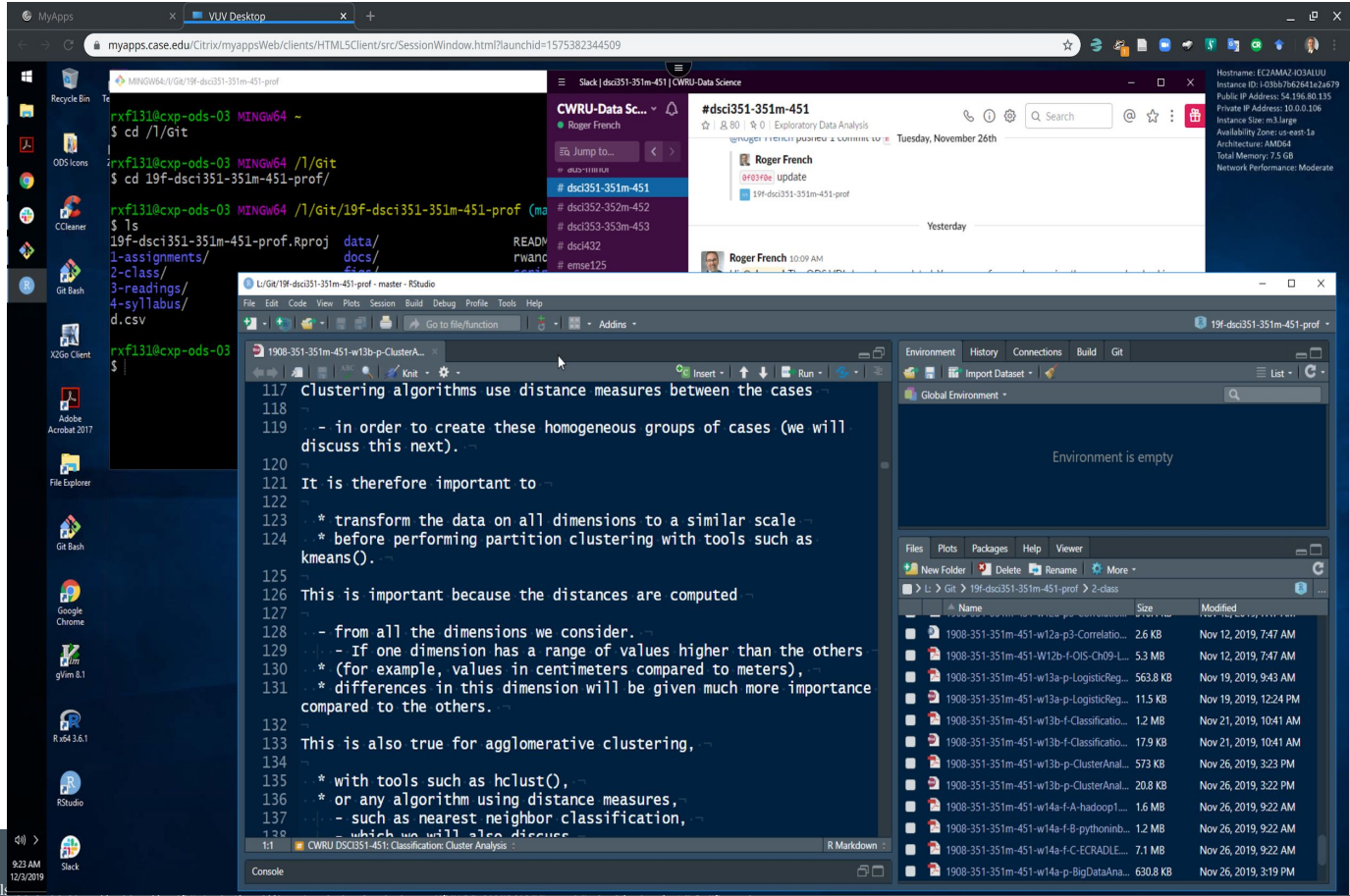
Compute Infrastructure for the ADS program

Provide students Open Data Science Computers

- Win10 Cloud Computers (Citrix)
- Hosted by CWRU
  - Scalable,
  - Good Performance
- With R, with Rstudio IDE
- Python3, with PyCharm IDE
  - “standard” R packages
  - “standard” Python packages
- Git, (git bash)
- Pandoc, LaTeX, html
- Slack
- StackExchange

Standard ODS Env.

- No time lost fixing computers
- Full install instruc. provided



CWRU Markov Data Science Cluster: Hosted by [U]Tech Res. Computing

Markov Total = 1120 CPU cores, 174k GPU cores

- 28 nodes: 2 Xeon CPUs (20 cores/CPU).
- 20 nodes: 2 Xeon CPUs with 2 Nvidia RTX2080Ti
- 1120 CPU cores and 174K GPU cores.

Enables Batch & Interactive GUI sessions

- Using either compute or GPU nodes.

Running R/Rstudio and Python3/PyCharm

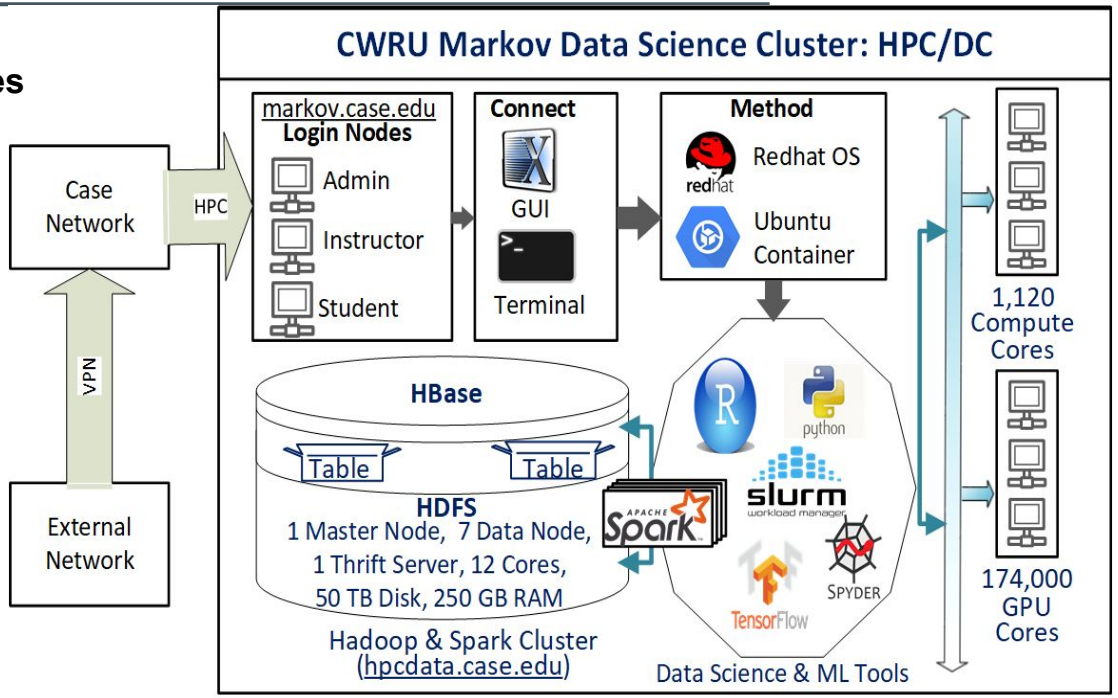
- Along with Keras/TensorFlow2/Cuda/CDNN

Implemented using the Open Data Science (ODS) Ubuntu 20.04 Container

- Based on Singularity.

Markov's Hadoop cluster: hpcdata.case.edu

- Loaded with publically available datasets



Teaching with Git & The Tools of Agile Software Development

Coursework distributed using Git Repository

- Fork the “Prof” repo
- Students tend their personal repo

Coherent Repo Structure

- Codes using relative pathing
- So codes work cross-platform

Git Sync and Pull

- For each class

Git Add, Commit, Push

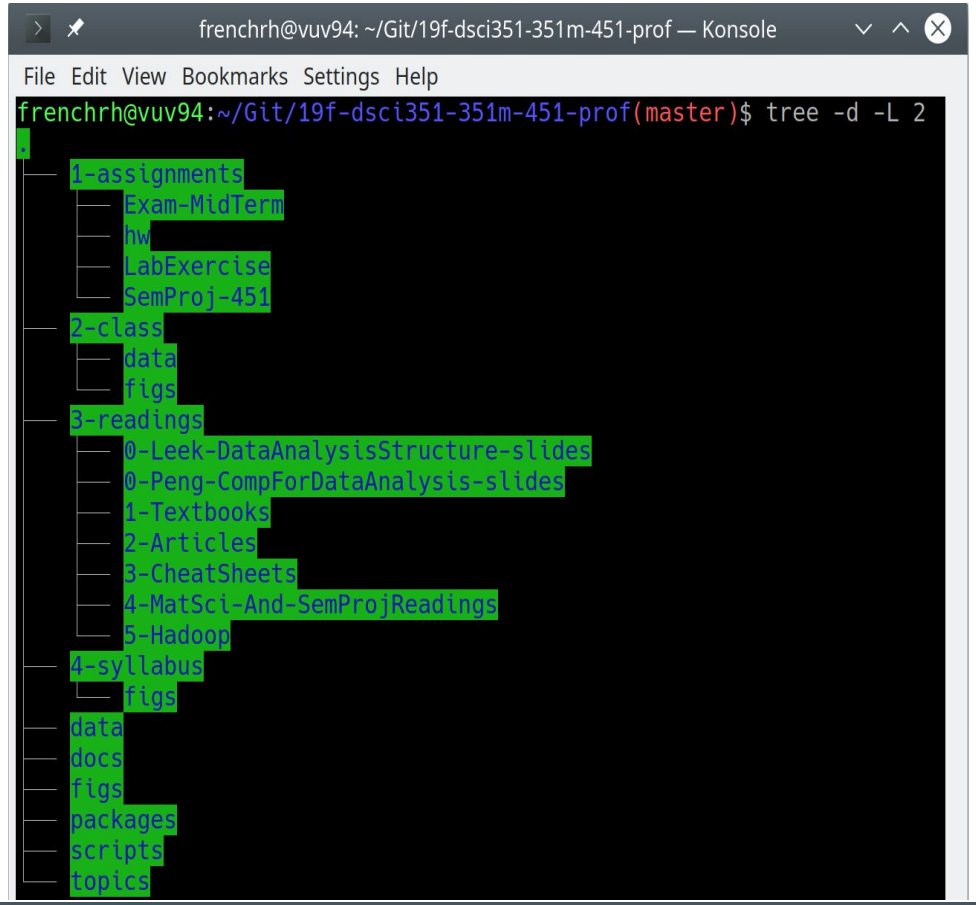
- Students own work

Class Notes in Rmarkdown

- Compiled to pdf
- With Pandoc & LaTeX

Assignments

- Traditional homeworks
- Lab Exercises: Two week assignments

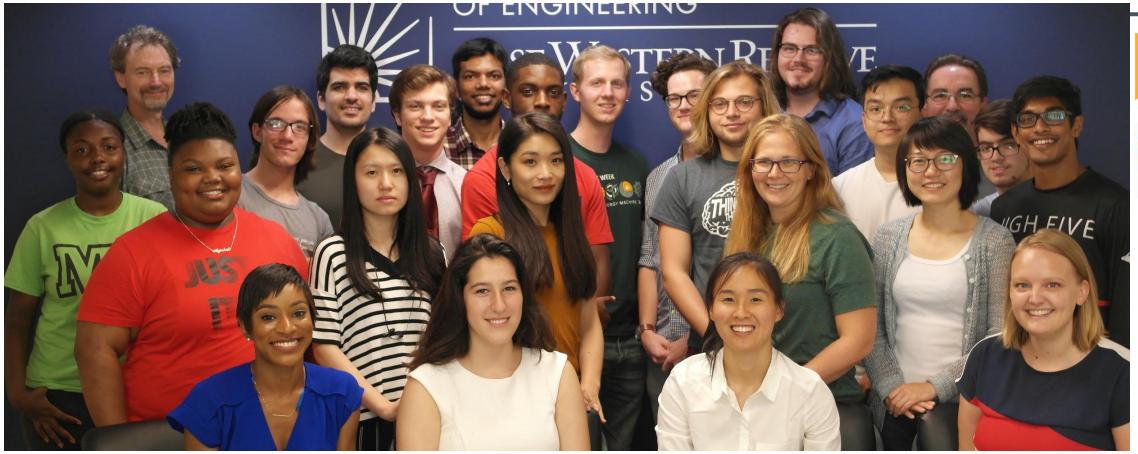


Machine Learning and Image Processing Techniques for Materials Evaluation

Roger H. French

SDLE Research Center
Case Western Reserve University

SDLE Research Center: Acknowledgements



CWRU Faculty
- Roger French, Laura Bruckman, Jeffrey Yarus, Yinghui Wu, Alp Sehirlioglu, Matt Willard, Vipin Chaudhary

Post-doctoral Research Associates
- Jay Jimenez, Pawan Tripathi, 1 Open Position

Graduate Students
- JiQi Liu, Sameera Nalin Venkat, Arafath Nihar, Raymond Wieser, Tian Wang, 3 Open PhD Positions
- Kristen Hernandez, Nat Tomczak, Will Oltjen, Liangyi Huang, Weiqi Yue
- Alex West, Steven Timothy, Deepa Bhuvanagiri, Tom Ciardi, Hein Aung, David Meshnick,

Undergraduates
- Tyler Burleyson, Carolina Whitaker, Minh Luu, Asher Baer, Daniel Arnholt, Medha Nayak
- Cora Lutes, Beck Pierce, Shreeyah Chugh, Sakin Kirti, Kehley Coleman, Shuyue Bian
- Vibha Mandayam, Guo Chen, Nathan Romig, NIkhila Balasubramaniam, Leean Jo, Max Atkinson
- Andre Yost, Asya Orhan, Lena Plover, Glenn Boyette, Nitin Chockalingam, Jakob Wegmueller
- Abhinav Khanna, Carolina Whittaker, Mirra Rassmussen, Ray Le

High School: Jack Gordon

SDLE Staff: Jonathan Steirer, Rich Tomazin

Outline

CRADLE Compute Cluster

- Distributed & High Performance Computing
- FAIRification of Datasets

Image Processing

- Nucleation & Growth of AlN Crystals from Al:Ni Melt
- Automated Analytics Pipelines

Supervised Machine Learning:

- Degradation of PV Cells Using Electroluminescent Images
- Classification with Convolutional Neural Networks
- EL + I-V Data: Data Integration with Non-Image Datasets

Unsupervised Machine Learning of EL Images

- Bag of Words and Feature Vectors

Spatiotemporal Graph Neural Networks

- Power Forecasting for PV Power Plant Fleets

Common Research Analytics & Data Lifecycle Environment (CRADLE) Compute Cluster

- Distributed & High Performance Computing[1]
- FAIRification of Datasets

1. A. Khalilnejad, et al., Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data, PLOS ONE. 15 (2020) e0240461. https://doi.org/10.1371/journal.pone.0240461.

Data Processing Framework: CRADLE

Data acquisition

- Diverse sources
- Anonymization
- Pre-processing
- Metadata

NoSQL Database system

- HBase

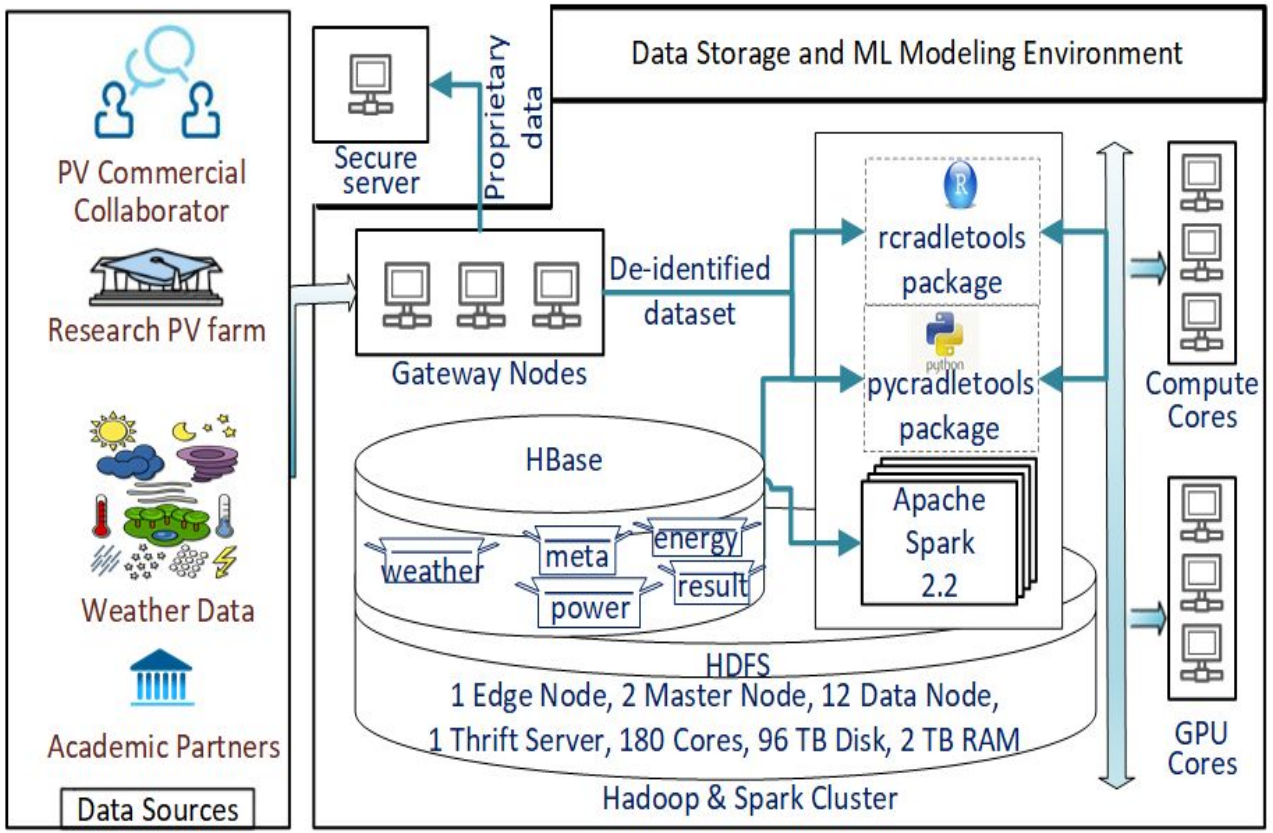
Computation

- Analysis
- Impute missing values
- R & Python3
- Tensorflow2
- Torch & Pytorch

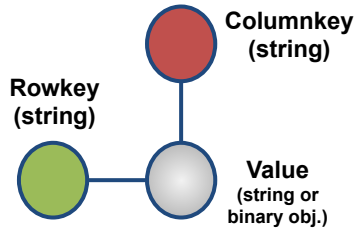
CRADLE tools

- R & Python package

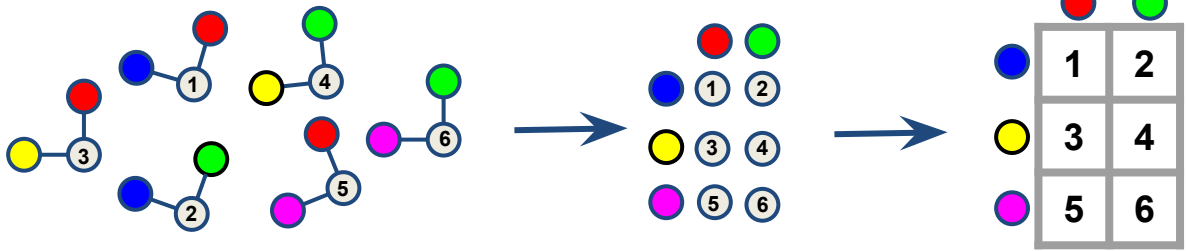
Common Research Analytics & Data Processing Environment



The “NoSQL” Database Abstraction of Hadoop/Hbase: RDF Triples



HBase RDF ‘Triple’



Combines Lab data (Spectra, Images, Videos etc.)

With Geospatiotemporal Data (PV Power Plant Data)

Distributed & High Performance Computing:

Petabyte Data Lake In A Petaflop HPC Environment

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

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*

Automated pipeline framework for processing of large-scale building energy time series data

Arash Khalilnejad[1,5], Ahmad M. Karimi[2,5], Shreyas Kamath[1,5], Rojiar Haddadian[2,5], Roger H. French[2,4,5]*, Alexis R. Abramson[3,6¤]

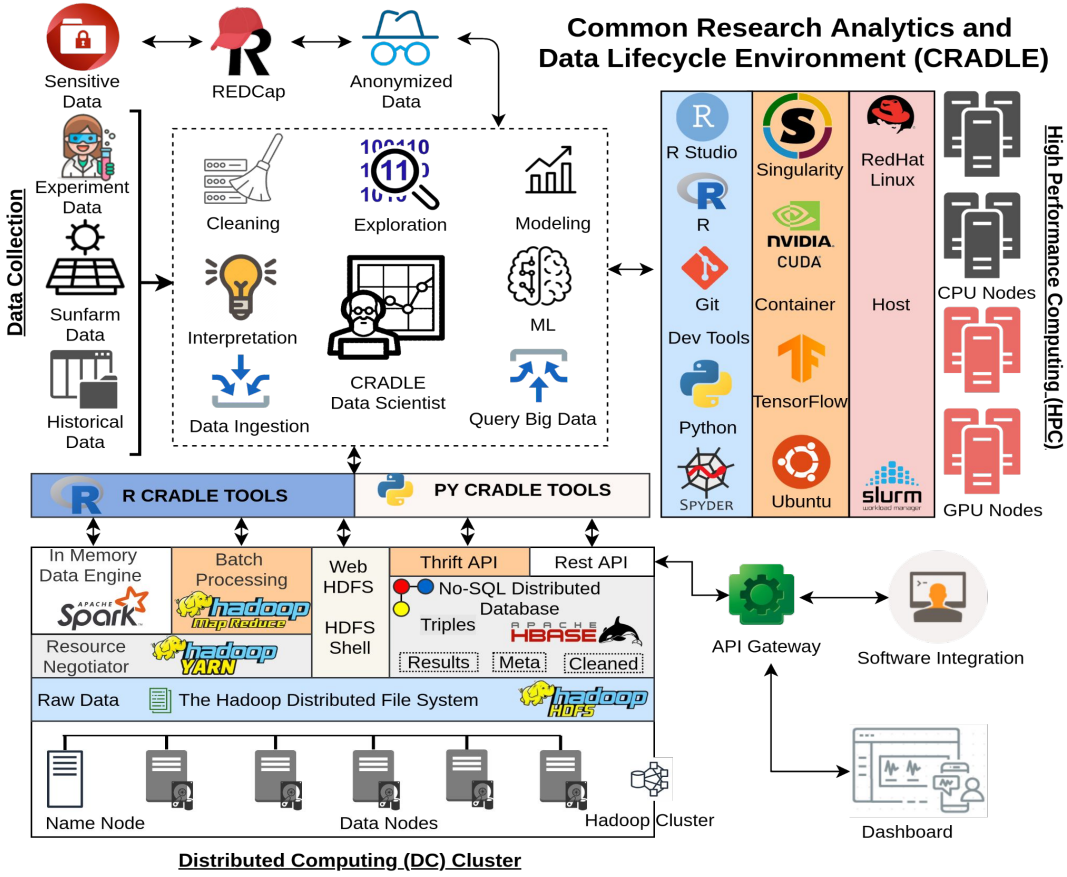
CRADLE Analytics Environment

CRADLE 2.2

- 192 Cores
- 1 Tb RAM
- 75 Tb Storage
- CDH 5.13.0

CRADLE 2.3

- 128 Cores
- 0.5 Tb RAM
- 100 Tb Storage
- CDH 5.16.2
- NSF SP-800-171
- For DOD CUI



FAIRification of Datasets and Models, Enables AI learning

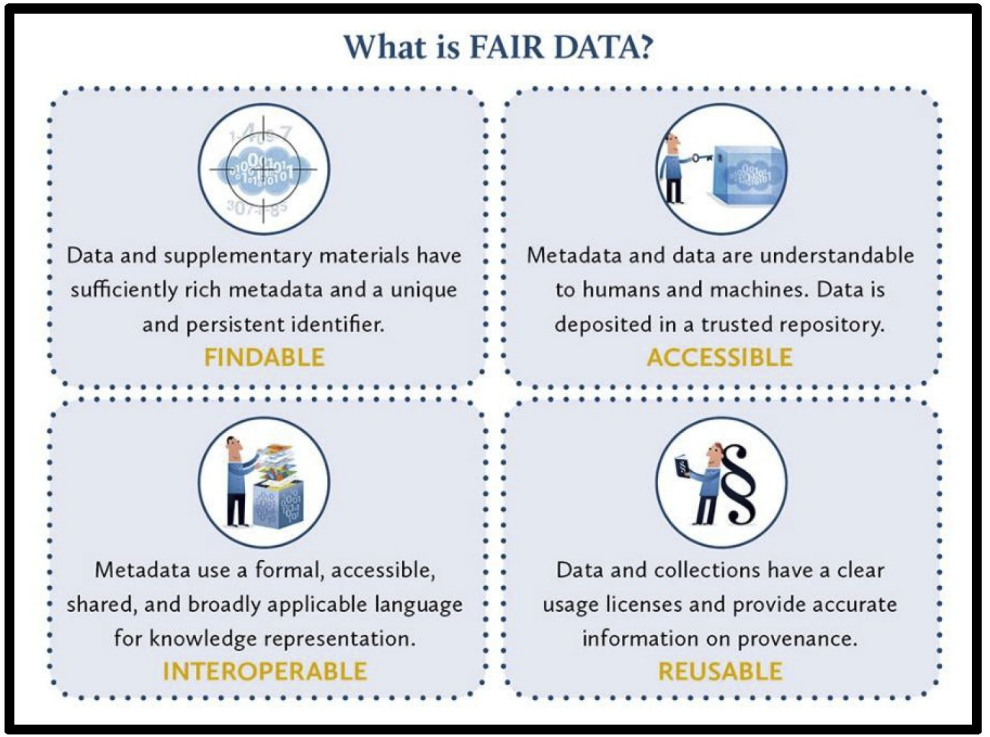
Making Datasets & Models FAIR
- By “FAIRification”

Enables Models to find Data
- And Data to find Models

So that they can advance
- Without human intervention

This is an aspect of the Semantic Web
- And Resource Description Framework
- Hbase triples are an example of RDF

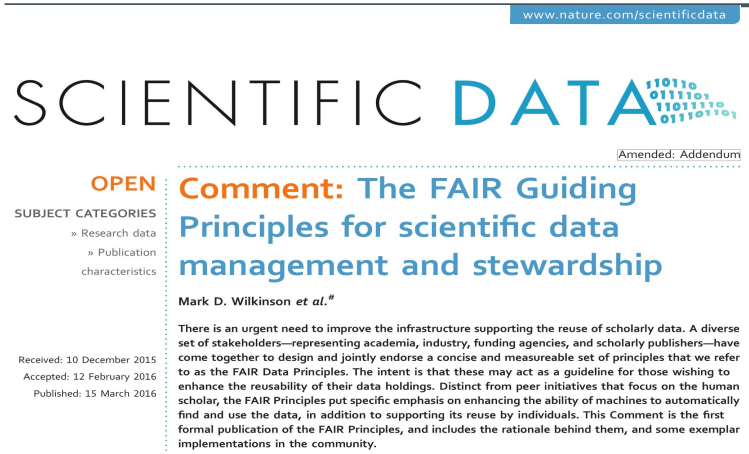
We just received a DOE SETO AI award
- For st-GNN, that includes FAIRification



M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci. Data, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.18.
https://en.wikipedia.org/wiki/Resource_Description_Framework

SDLE Research Center, Roger H. French © 2021 http://sdle.case.edu

FAIRification:


www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

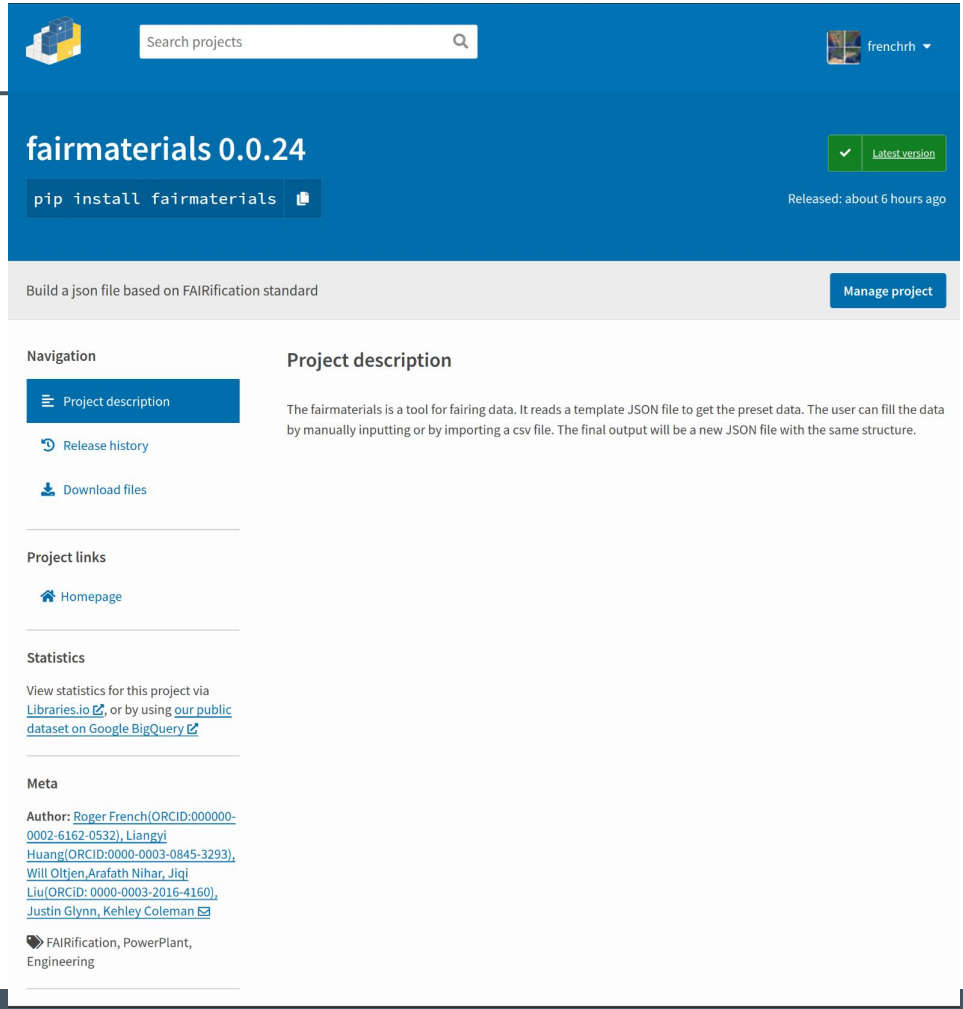
SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.



FAIRmaterials Code Package

- **Just published the Python3 package!**
  - To PyPI.org
  - Great work of Liangyi Huang, Xusheng Xiao
- **R package submitted!**
  - To CRAN
  - With Will Oltjen (of DOE-SETO PV proj.)
  - JiQi Liu and Arafath Nihar
  - And SDLE UGs

Image ML Workflow Diagram

Data Collection

Data Preprocessing

Object & Feature Extraction

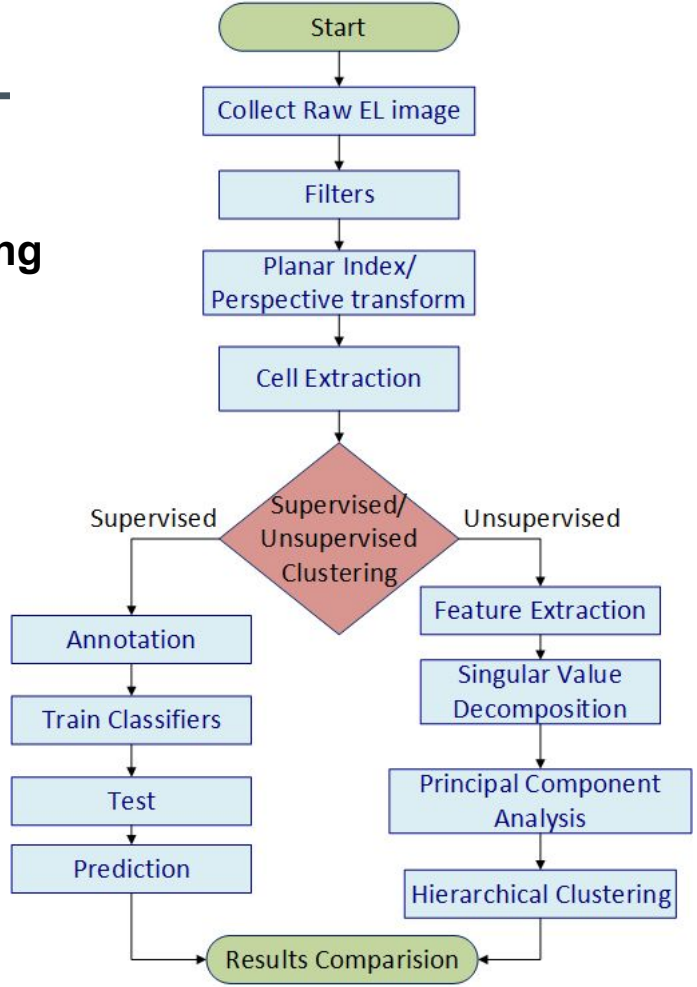
Machine Learning Application

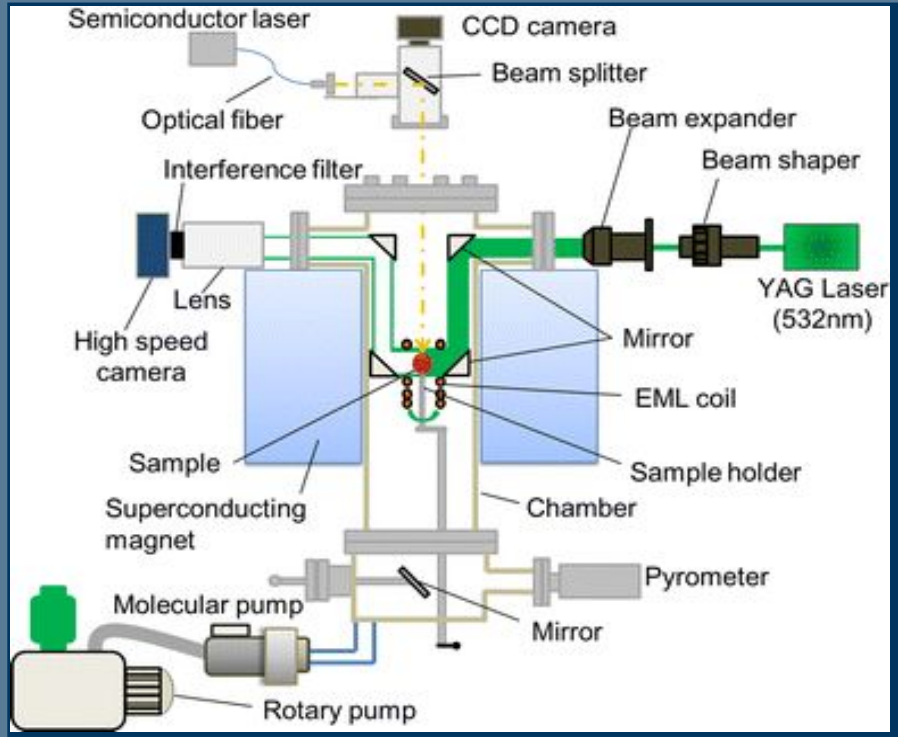
- Classification
- Clustering
- Regression

Results

Image Processing

Machine Learning





Nucleation & Growth Kinetics of AlN as quantified from Magnetic Levitation Studies

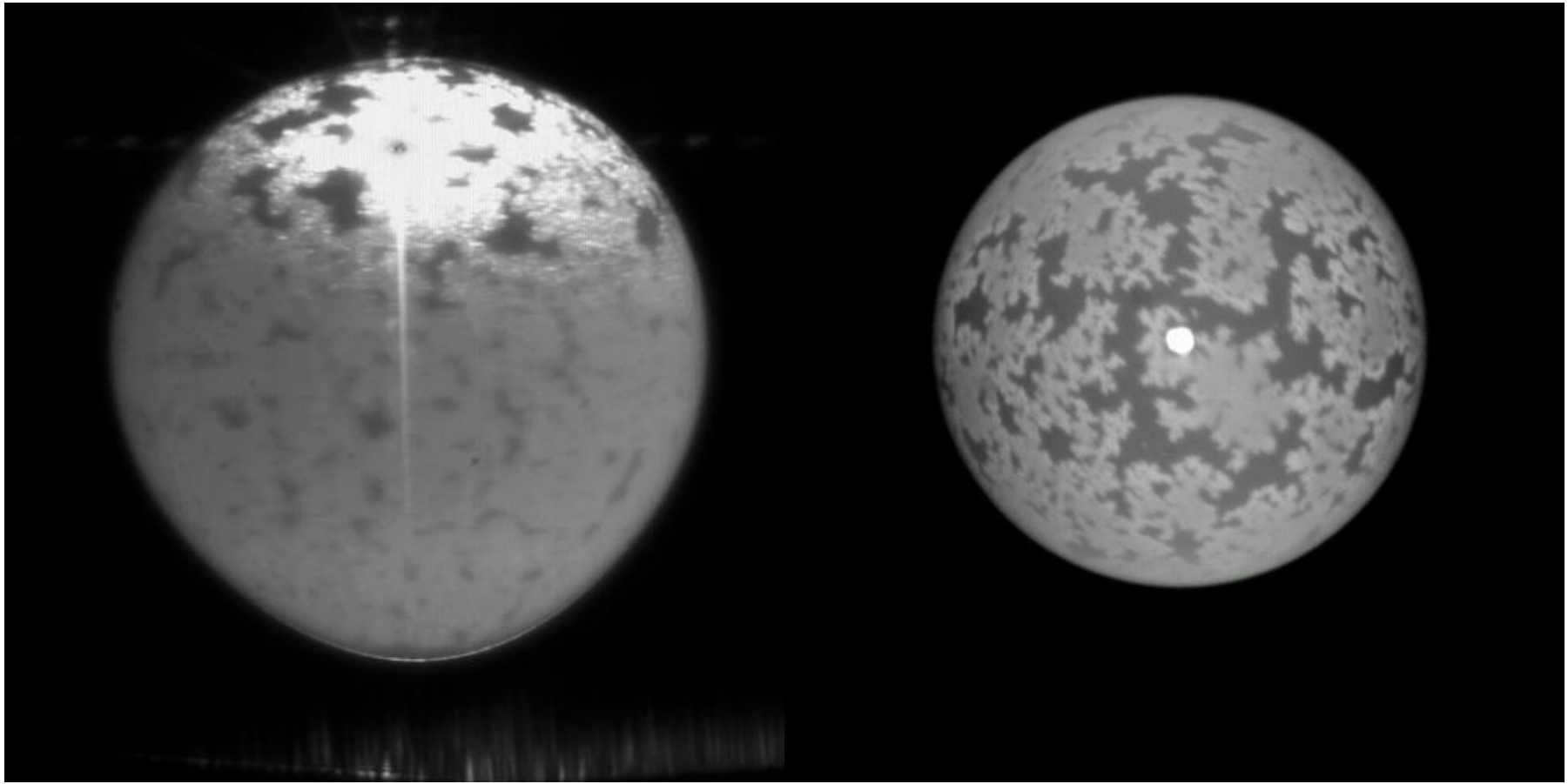
Tohoku: Hiroyuki Fukuyama (Prof), Masayoski Adachi (Assistant Prof.),

Sonoko Hamaya (M.S. student/graduated), Yuji Yamagata (M.S. student/current)

CWRU: Jennifer Carter (Assistant Prof), Roger French (Prof)

Andrew Loach (Undergrad), Justin Fada (M.S. student), Laura Wilson (Ph.D student),

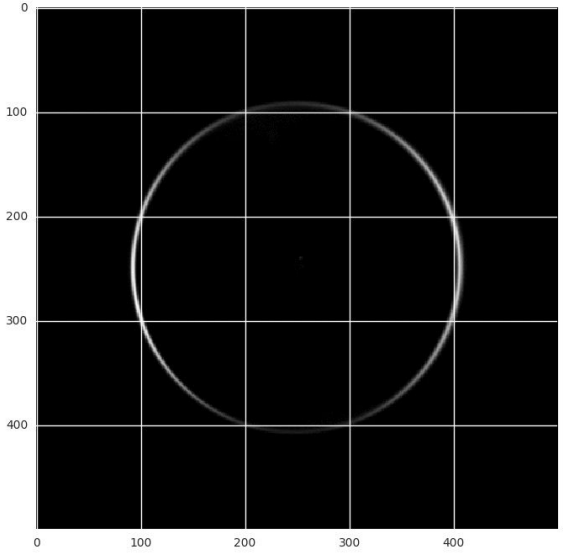
High Speed Video: Ni-50mol%Al In N_2 Gas: Top and Side View

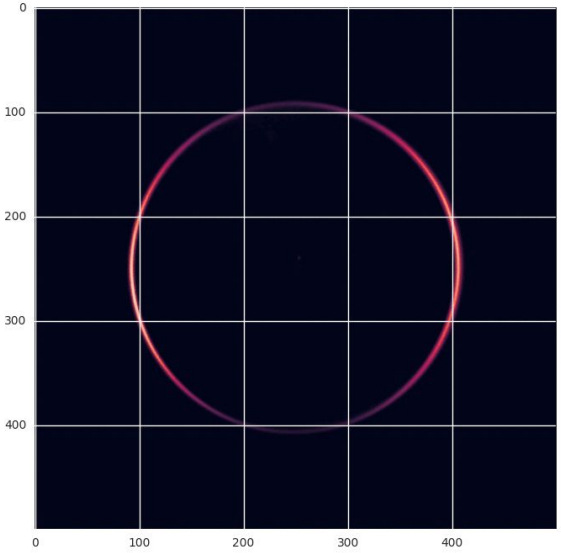


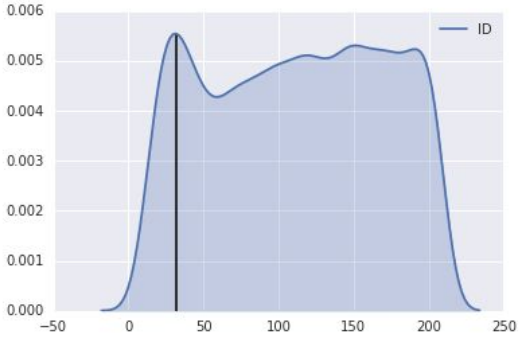
Time Dependent Crystal Nucleation

Density plots for Ni-50mol%Al top

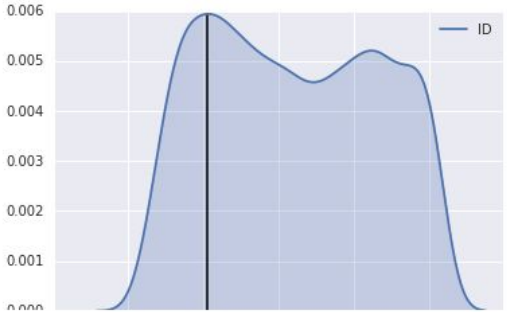
- Show that the small crystals
- Rapidly forming
- And growing as a function of time







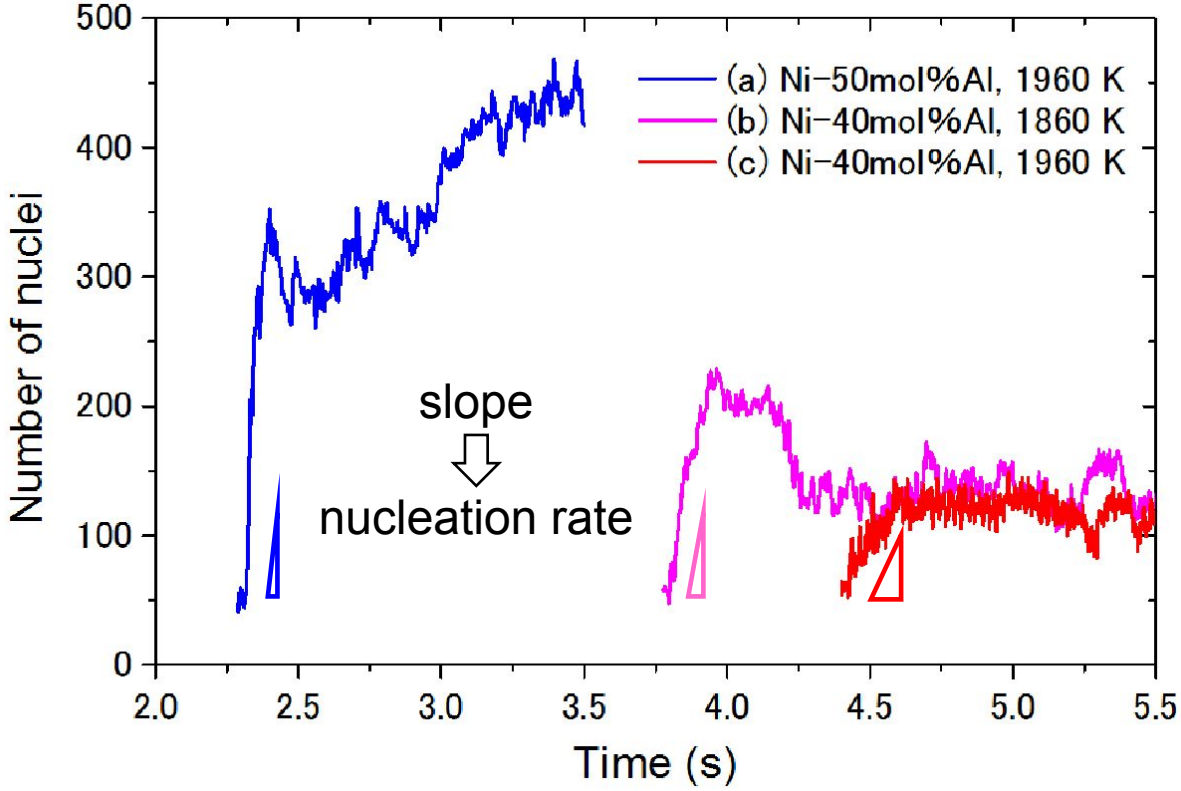

```
In [173]: p=sns.kdeplot(ff1_df['ID'], shade=True)
     ...:
     ...: x,y = p.get_lines()[0].get_data()
     ...:
     ...: max_y = y[np.argmax(y)]
     ...: max_x = x[np.argmax(y)]
     ...:
     ...: plt.vlines(max_x, 0, max_y)
     ...: plt.show()
```



Nucleation rate

Evaluated quantitatively

- by image processing.



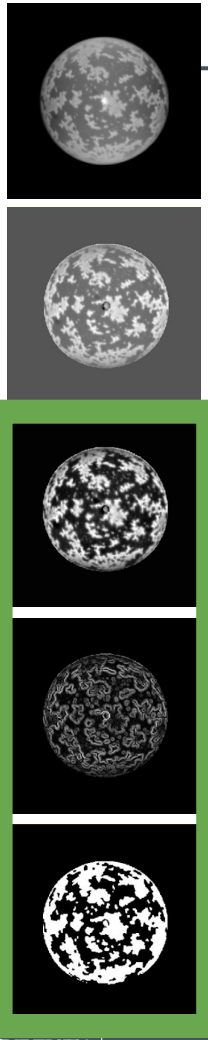
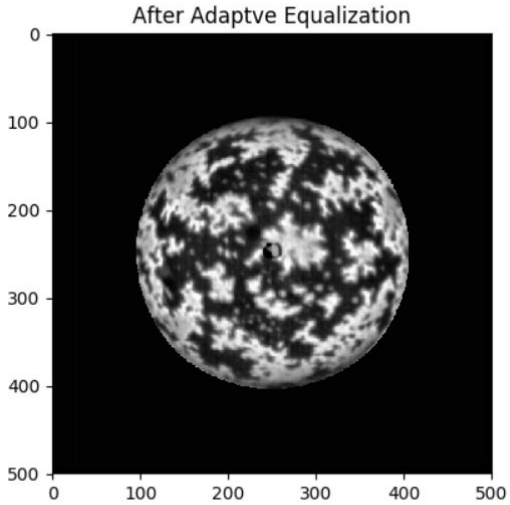
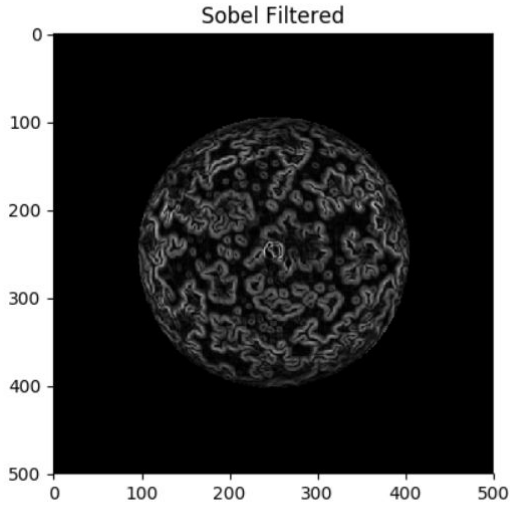
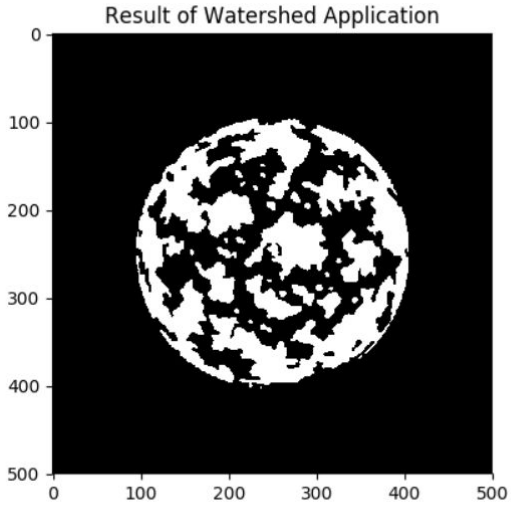


Image Procesing: Edge Finding & Watershed Segmentation

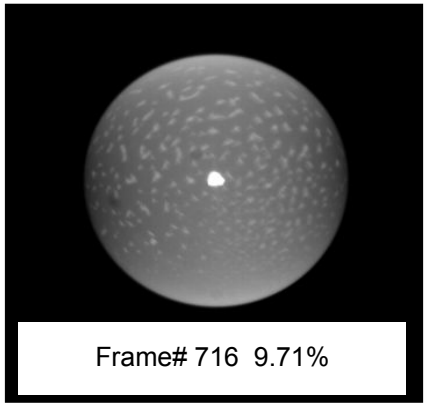
- Locally increase contrast while maintaining rank order
- Sobel filter finds intensity gradient
  - Gradient used to represent edges of crystals
- Watershed application fills in the regions
  - Uses markers to define foreground and background

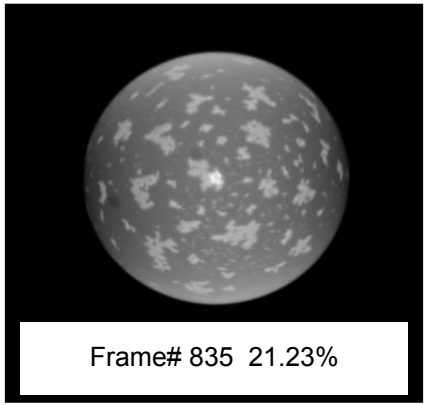


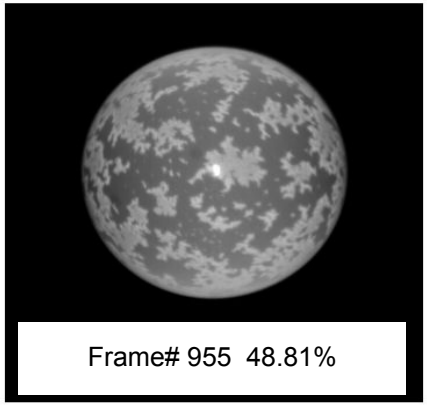


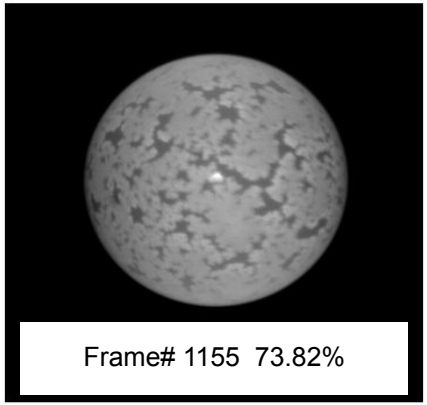


Time Dependent Surface Crystal Coverage Ni-50mol%Al









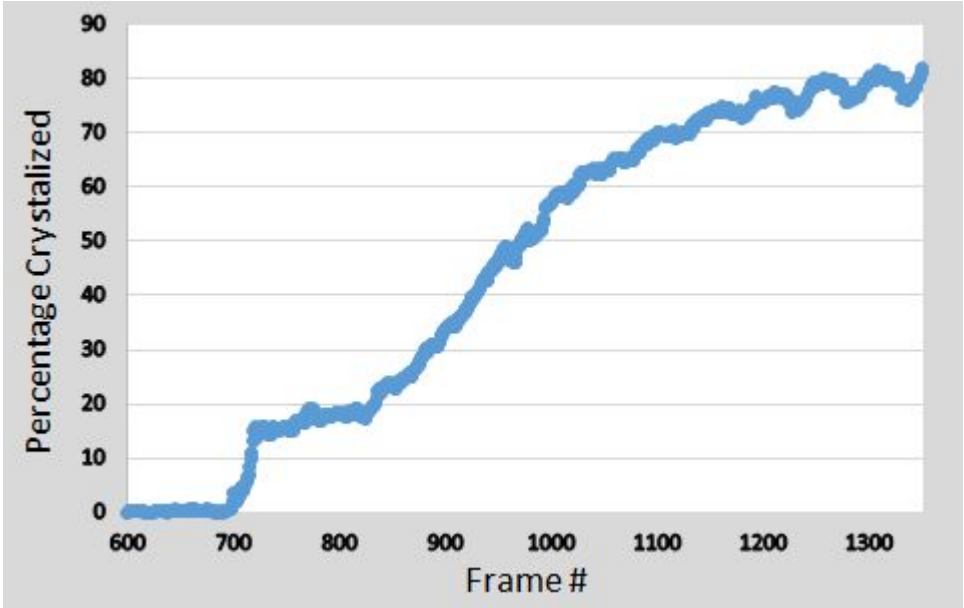


Image Process 400,000 Images With Automated Code Pipeline

Both Nucleation & Growth

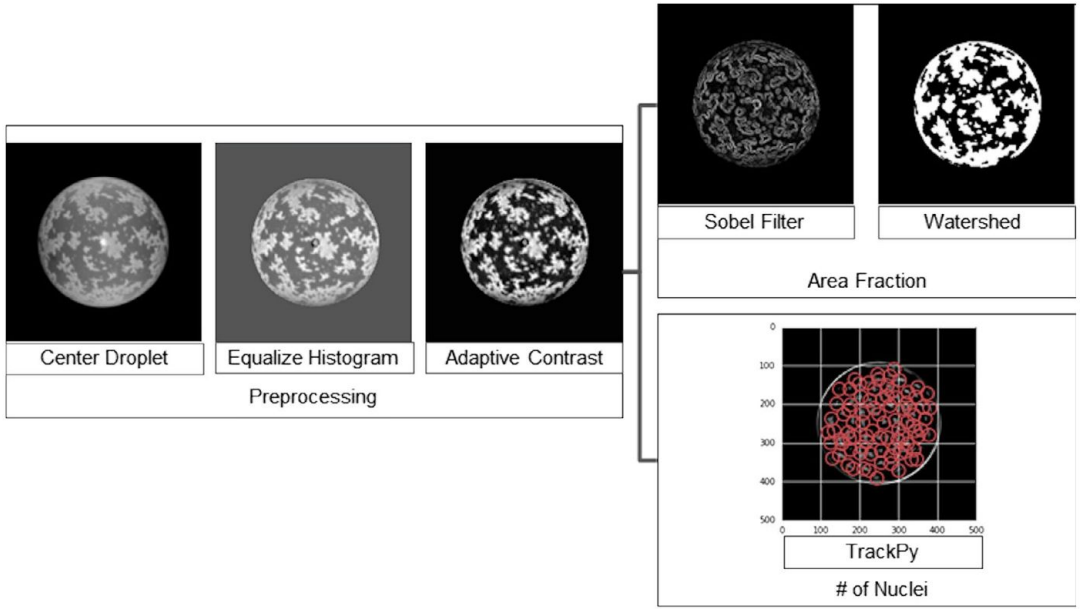


FIGURE 4 Schematic of the image processing procedures to quantify kinetic parameters of the formation of AlN on the levitated Ni-Al droplet

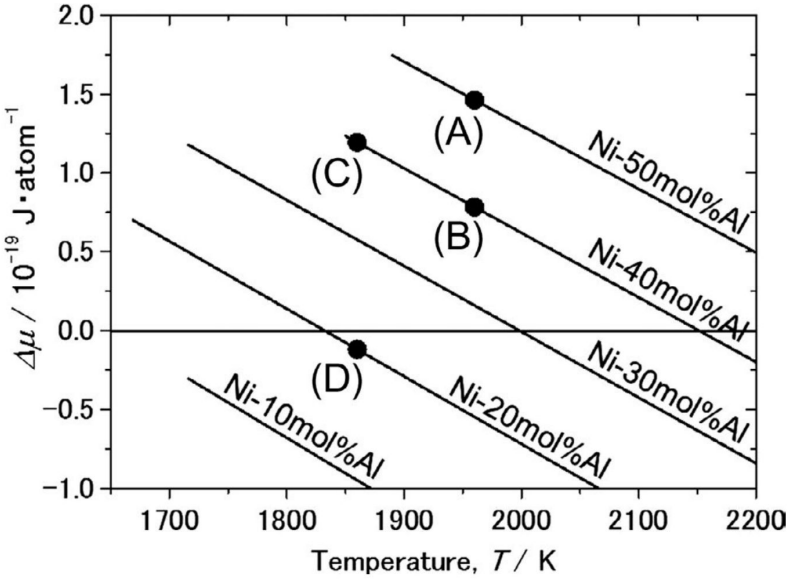
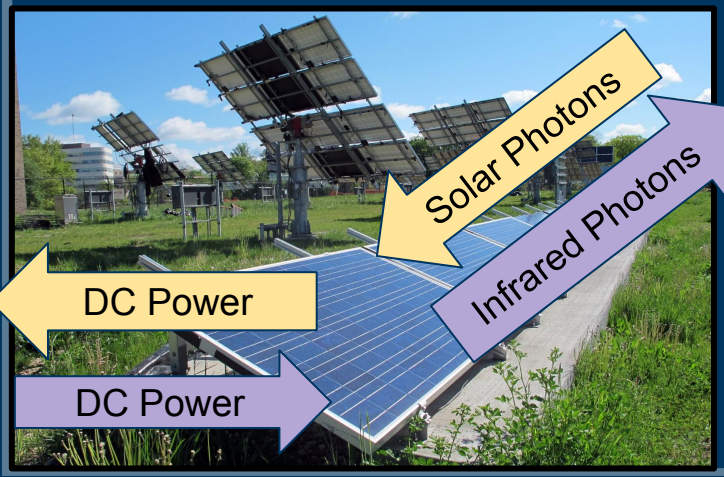
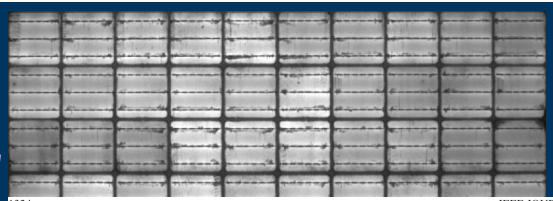


FIGURE 2 Temperature-dependent driving force of AlN formation on Ni-Al droplets for various compositions under 1 bar nitrogen partial pressure. The left-end of each line is the liquidus temperature for that alloy composition. The points (A-D) indicate that temperatures and compositions explored in this paper

Supervised Machine Learning

- Degradation of PV Cells Using Electroluminescent Images
- EL Images + I-V Data: Predict Power Loss From Images





Automated Pipeline for Photovoltaic Module Electroluminescence Image Processing and Degradation Feature Classification

Ahmad Maroof Karimi, *Graduate Student, Member, IEEE*, Justin S. Fada, Mohammad Akram Hossain, Shuying Yang, Timothy J. Peshek, Jennifer L. Braid, *Member, IEEE*, and Roger H. French, *Member, IEEE*

Generalized and Mechanistic PV Module Performance Prediction from Computer Vision and Machine Learning on Electroluminescence Images

Ahmad Maroof Karimi *†, Justin S. Fada *, Nicholas A. Parrilla *, Benjamin G. Pierce *†, Mehmet Koyutürk †, Roger H. French *‡, *Member, IEEE*, Jennifer L. Braid *‡, *Member, IEEE*
* SDLE Research Center, Case Western Reserve University, 10900 Euclid Ave., Cleveland, Ohio 44106, USA
† Department of Computer and Data Sciences, Case Western Reserve University
‡ Deapartment of Materials Science and Engineering, Case Western Reserve University

1. A.M. Karimi, et al., Automated Pipeline for Photovoltaic Module Electroluminescence Image Processing and Degradation Feature Classification, IEEE Journal of Photovoltaics. 9 (2019) 1324–1335. https://doi.org/10.1109/JPHOTOV.2019.2920732.
2. A.M. Karimi, et al., Generalized and Mechanistic PV Module Performance Prediction From Computer Vision and Machine Learning on Electroluminescence Images, IEEE J. Photovoltaics. 10 (2020) 878–887. https://doi.org/10.1109/JPHOTOV.2020.2973448.

Dataset : 10,000 PV Cell Images From a Test-to-Failure Study

15 Damp-heat modules

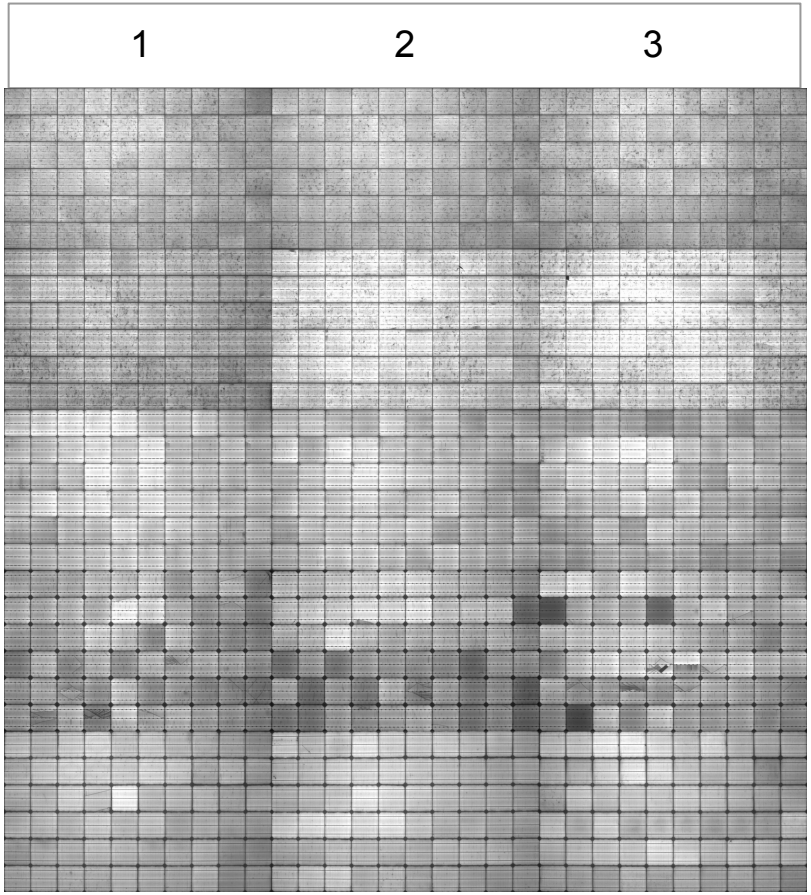
- 5 Brands (A,B,C,D,E)
- 3 Samples per brand
- 3 Wafer types
  - Mono-Si
  - Mono-PERC
  - Multi-Si

Six 500 hour steps

- 500-3000 hours
- IEC 61215
- EL images and IV curves

Additional data of same 5 brands:

- Thermal cycling
- UV irradiance
- PID +1000V, PID -1000V
- Dynamic mechanical loading

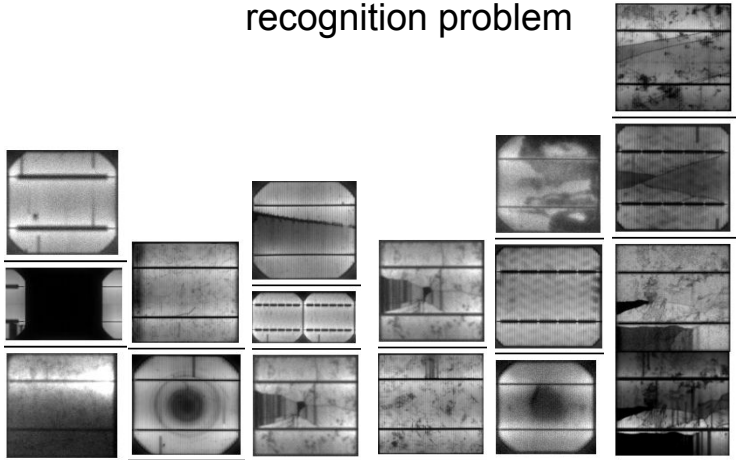


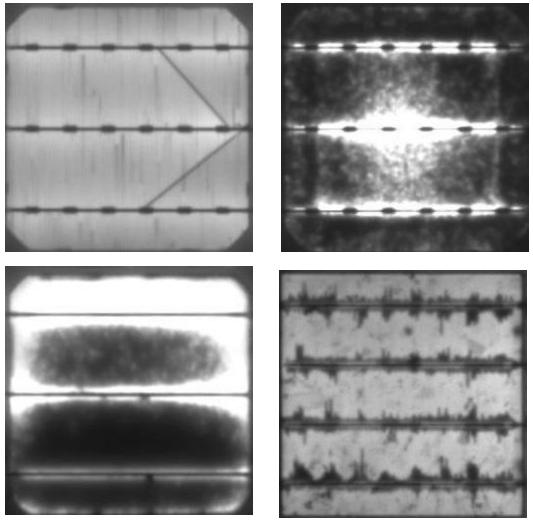
International Electrotechnical Commission, "IEC 61215 Terrestrial photovoltaic (PV) modules - Design qualification and type approval," International Electrotechnical Commission, International Standard, 2016.

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
  - Resembles face recognition problem





Supervised PV Cell Classification by Degree of Corrosion

Cell-level images receive corrosion score

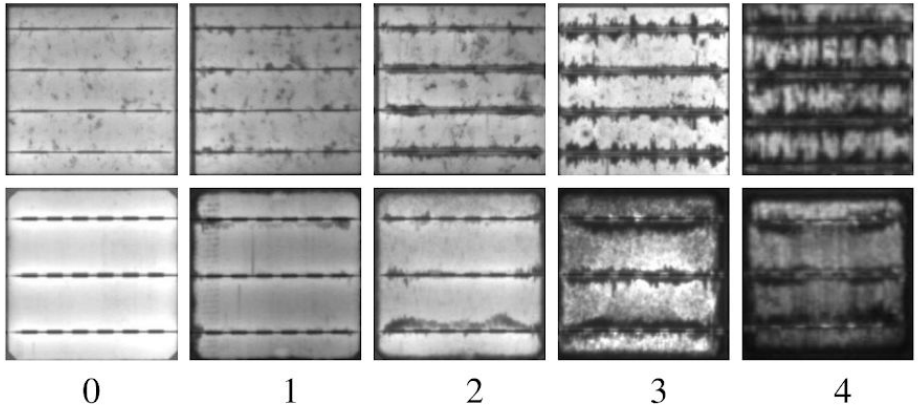
- Busbar corrosion = 0 to 4

Convolutional Neural Network

- Trained to assign corrosion score
- Based on manual classification

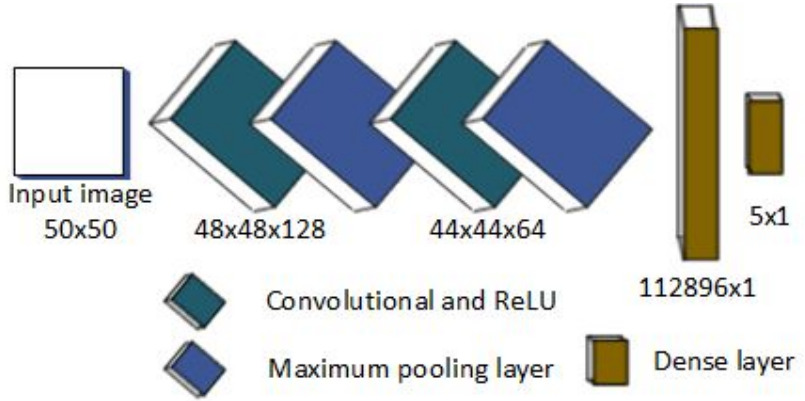
Module-level images scored as

- Average of cell-level corrosion

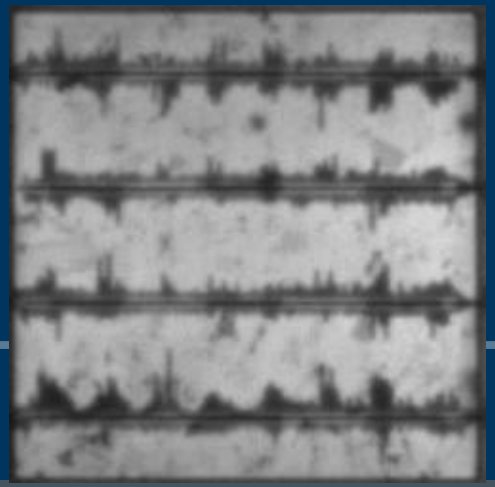


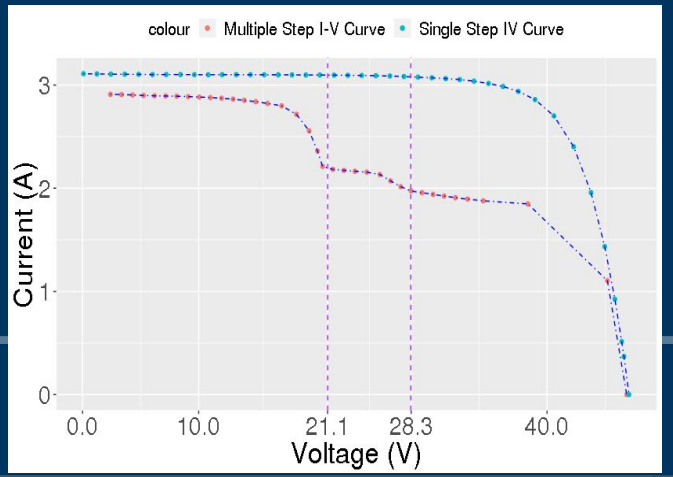
		Predicted class					Recall
		0	1	2	3	4	
Actual class	0	527	8	0	0	0	0.98
	1	20	112	5	0	0	0.81
	2	0	4	82	3	0	0.92
	3	0	0	2	62	2	0.93
	4	0	0	0	2	60	0.96
Precision		0.96	0.90	0.92	0.92	0.96	

Confusion matrix for cell classification into 5 corrosion levels in order of increasing severity from 0-4



Data Assembly & Integration for Predictive and Inferential Modeling of Photovoltaic Lifetime Performance





EL – IV Feature Correlation

Generalized EL features:

- EL median intensity (F_{med})
- Fraction dark ("inactive") pixels (F_{FDP})

Degradation-specific EL features:

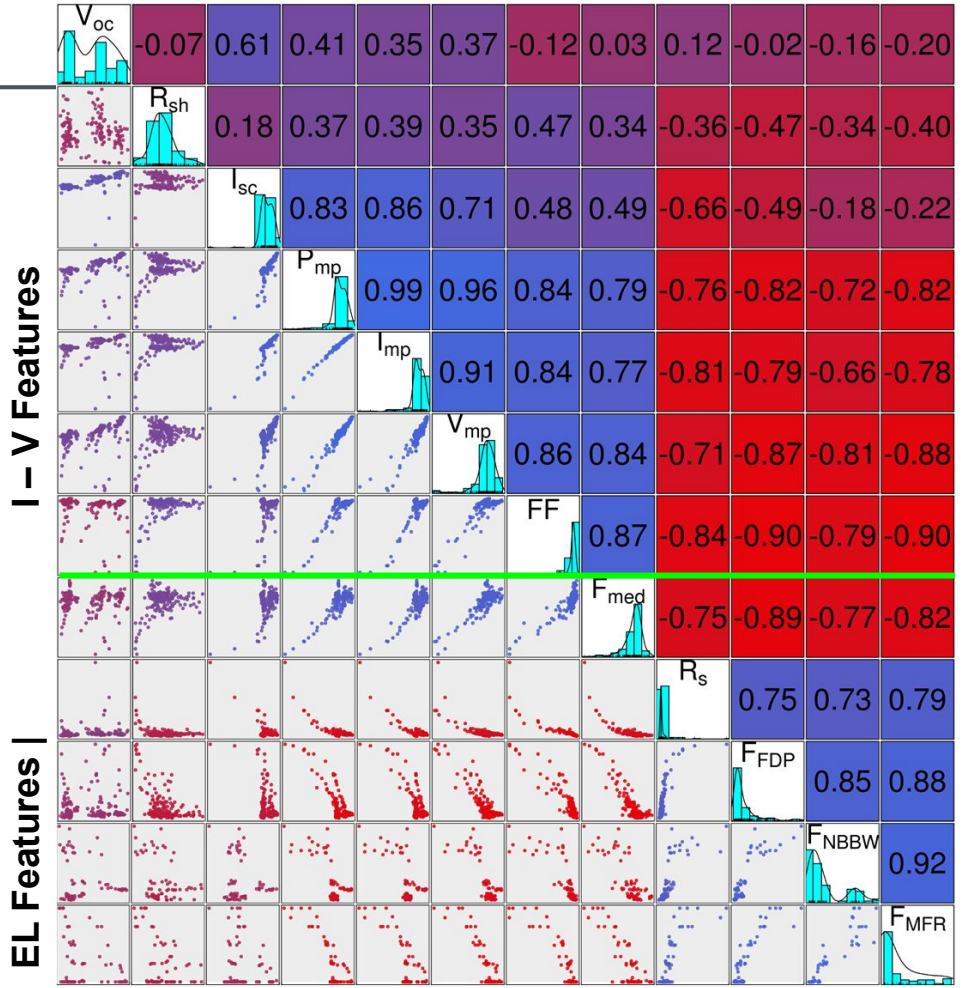
- Normalized busbar width (F_{NBBW})
- Module feature ratio (F_{MFR})

Correlation map of I-V and EL features

- Identify related features, such as
- EL median and series resistance (R_s)
- Series resistance and power (P_{mp})

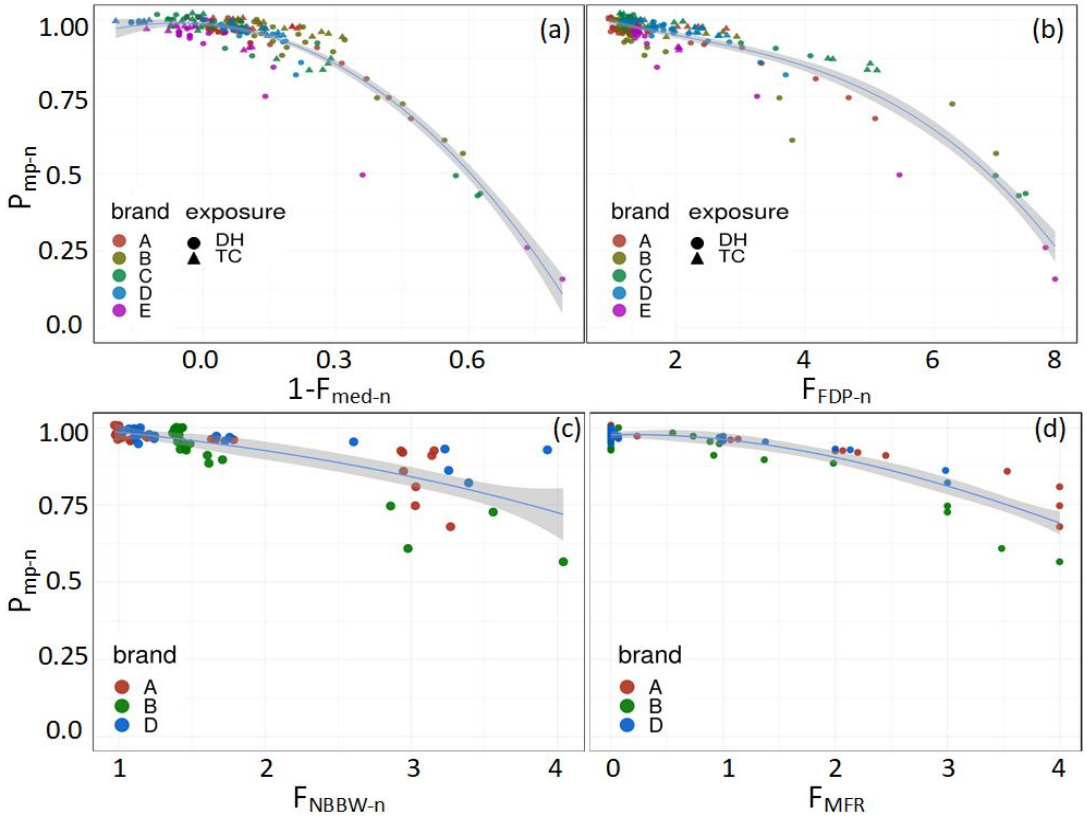
Develop predictive models for

- Overall module performance (P_{mp})
- Mechanistic degradation (R_s etc.)



PV Module Power Prediction from EL Image Features

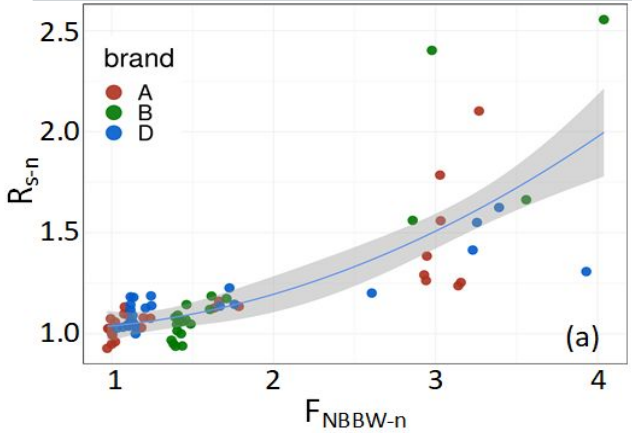
Generalized Power Prediction: Across Brands and Exposure Types

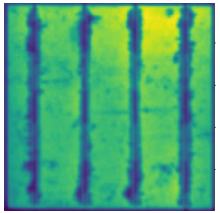


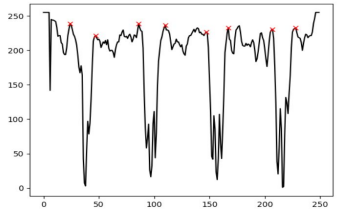
Power prediction models					
$P_{mp-n} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$					
Model	**X**	β_0	β_1	β_2	β_3
Median	$1 - F_{med-n}$	0.936	-1.396	-0.756	-0.039
Inactive Region	F_{FDP-n}	0.936	-1.502	-0.478	-0.128
NBBW	F_{NBBW-n}	0.935	-0.576	-0.064	-0.005
MFR	F_{MFR}	0.932	-0.650	-0.204	0.014

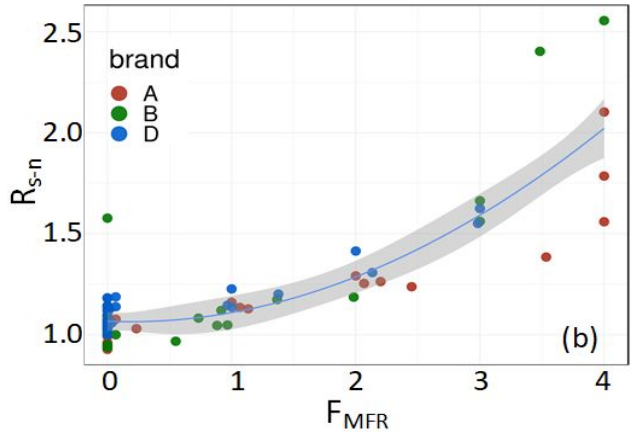
Metrics to measure the performance of the models			
	Model Type	**Mean**	**Standard dev.**
Adjusted R^2	**Median**	0.88	0.025
	Inactive Region	0.87	0.016
	NBBW	0.70	0.03
	MFR	0.81	0.017
RMSE	**Median**	11.87	5.57
	Inactive Region	12.44	4.96
	NBBW	13.35	10.53
	MFR	9.53	4.75
MAPE	**Median**	3.46	1.58
	Inactive Region	3.94	2.08
	NBBW	4.34	3.08
	MFR	2.83	1.45

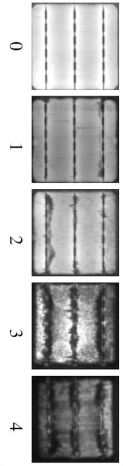
Mechanistic Degradation Prediction from EL Image Features











Predicting Series Resistance from Busbar Corrosion

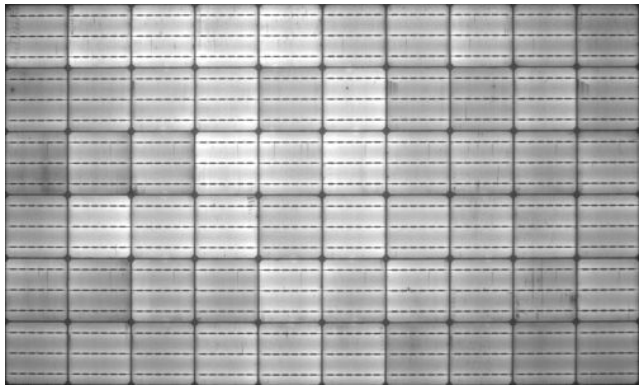
Series resistance prediction models				
$R_{s-n} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$				
Model	**X**	β_0	β_1	β_2
NBBW	F_{NBBW-n}	1.192	1.992	0.360
MFR	F_{MFR}	1.196	2.119	0.725

Metrics to measure the performance of the models			
	Model Type	**Mean**	**Standard dev.**
Adjusted R^2	**NBBW**	0.61	0.057
	MFR	0.73	0.025
RMSE	**NBBW**	0.049	0.034
	MFR	0.065	0.035
MAPE	**NBBW**	6.88	4.50
	MFR	6.82	1.75

Unsupervised Machine Learning of EL Images

- **Bag of Words and Feature Vectors**

Image Processing Pipeline



Simulated Exposure

Field Exposure

EL Imaging

CRADLE: Hadoop, Hbase, Spark

CWRU HPC Cluster

Supervised Learning w/ CNN

Unsupervised Learning

Bag of Visual Words Model

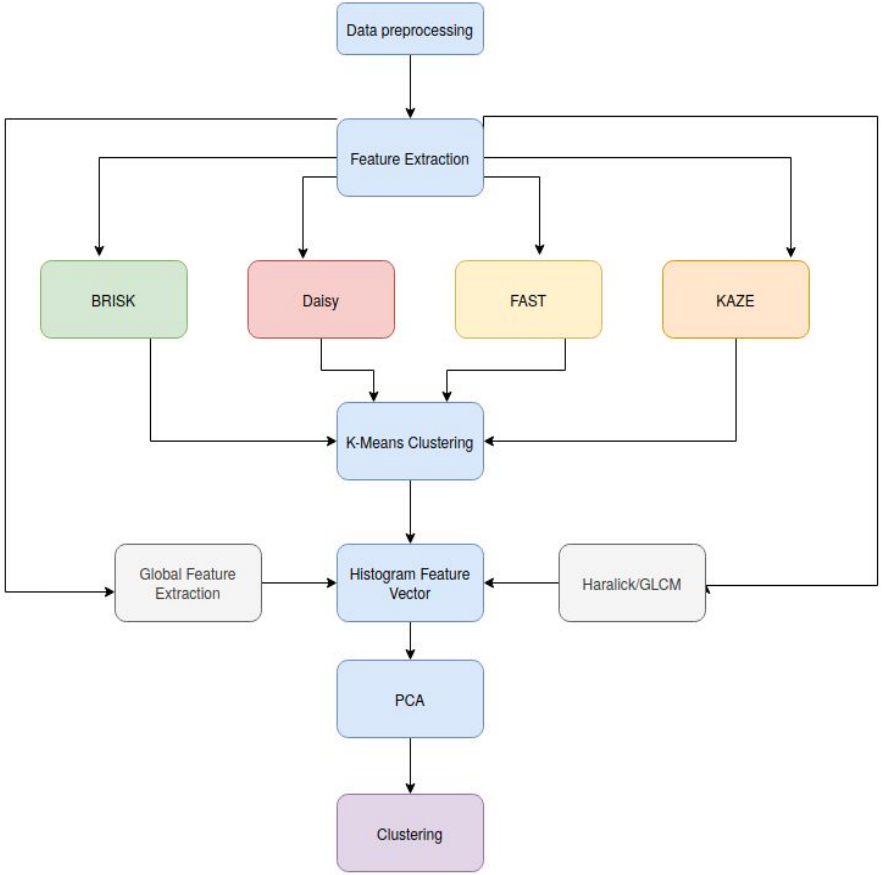
This type of model originated in NLP

- Natural Language Processing
- Finding the meaning of text

  Using word frequencies

Although each step is

- Not overly complex in itself
- The process is rather involved



Local Feature Extraction

First step: Extract important key points

- Key points represent different things
- Most commonly a shift in intensity

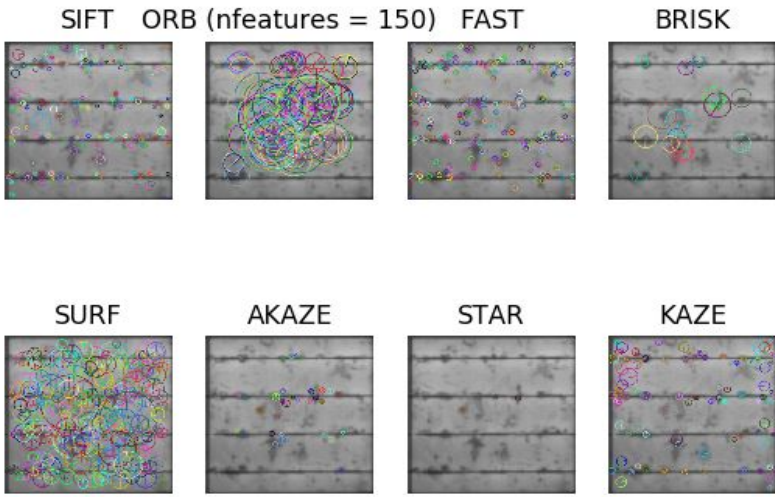
Then, each keypoint is described

- By some quantitative method
- Encoding relevant detail

We have used many algorithms

- ORB, SIFT, FAST, BRISK
- SURF, AKAZE, STAR, KAZE[1]

This descriptor is a 1D Feature Vector



$$\tau(\mathbf{p};\mathbf{x},\mathbf{y}) := \begin{cases} 1 & : \mathbf{p}(\mathbf{x}) < \mathbf{p}(\mathbf{y}) \\ 0 & : \mathbf{p}(\mathbf{x}) \geq \mathbf{p}(\mathbf{y}) \end{cases},$$

$$f_n(\mathbf{p}) := \sum_{1 \leq i \leq n} 2^{i-1} \tau(\mathbf{p};\mathbf{x}_i,\mathbf{y}_i)$$

[1] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 Intl. Conf. Comp. Vision*, Spain, Nov. 2011, pp. 2564–2571, doi: 10.1109/ICCV.2011.6126544.

Feature Vector Composition

Can compose a final feature vector using different techniques

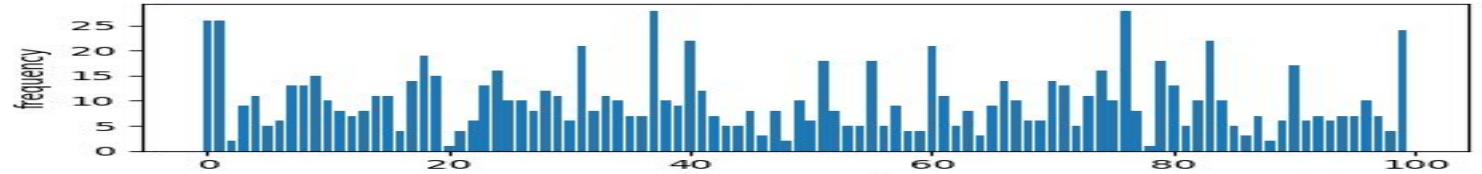
- Bag of Words
- Global feature descriptors
- Haralick (texture) features
  - Not discussed here

Example feature vector:

- BoW + Global feature descriptors
- Length = k + $pixels_{horizontal\ direction}$

Results in high dimensional Feature Vector

BoW_1	BoW_2	...	BoW_{k-1}	BoW_k	GFD_1	GFD_2	...	GFD_{x-1}	GFD_x



Clustering

Now that we have feature vectors

- We can cluster them

This can be done multiple ways

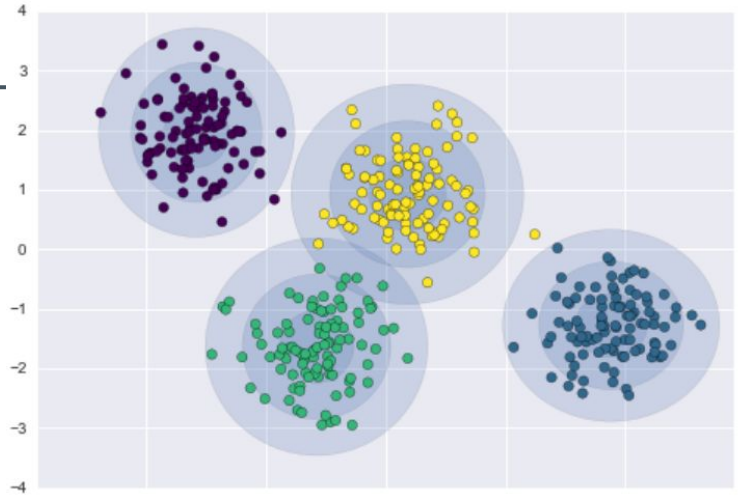
- With different clustering algorithms

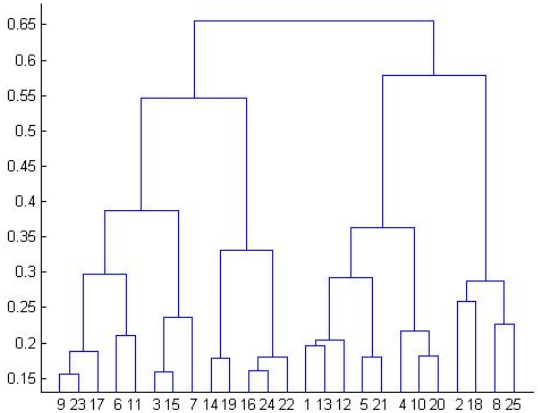
Currently, the best models appear to be

- Hierarchical clustering
- Gaussian mixture models

Models may or may not predict on new data

- May require additional learning





Classification

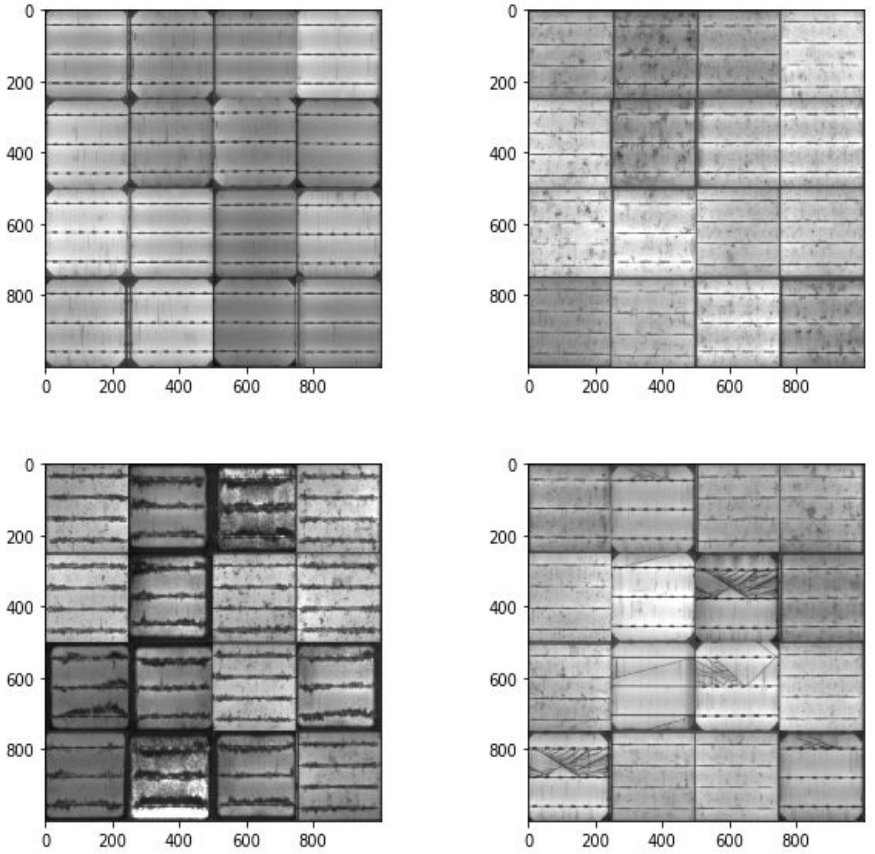
Clustering != Classification

- Although the two concepts are similar
- Clustering does not mean classification
- It only labels clusters

A human interprets clusters manually!

Clusters inherently represent

- The distribution of the data
- Rather then our expectations



Spatiotemporal Graph Neural Networks

- **Power Forecasting for PV Power Plant Fleets**

1. A.M. Karimi, et al., Spatiotemporal Graph Neural Network for Performance Prediction of Photovoltaic Power Systems, in: Proc. AAAI Conf., Virtual, 2021: p. 8. https://ojs.aaai.org/index.php/AAAI/article/view/17799.

Spatiotemporal Graph Neural Network (st-GNN)

Interest

- Leverage Information from Neighboring Nodes
- Undergoing Similar Exposure Conditions

Sequence of

- Graph Convolution Layer
- Temporal Convolutional Layer
  - 1-D Convolution

Coherence

- Spatial Dependencies
- Temporal Dependencies

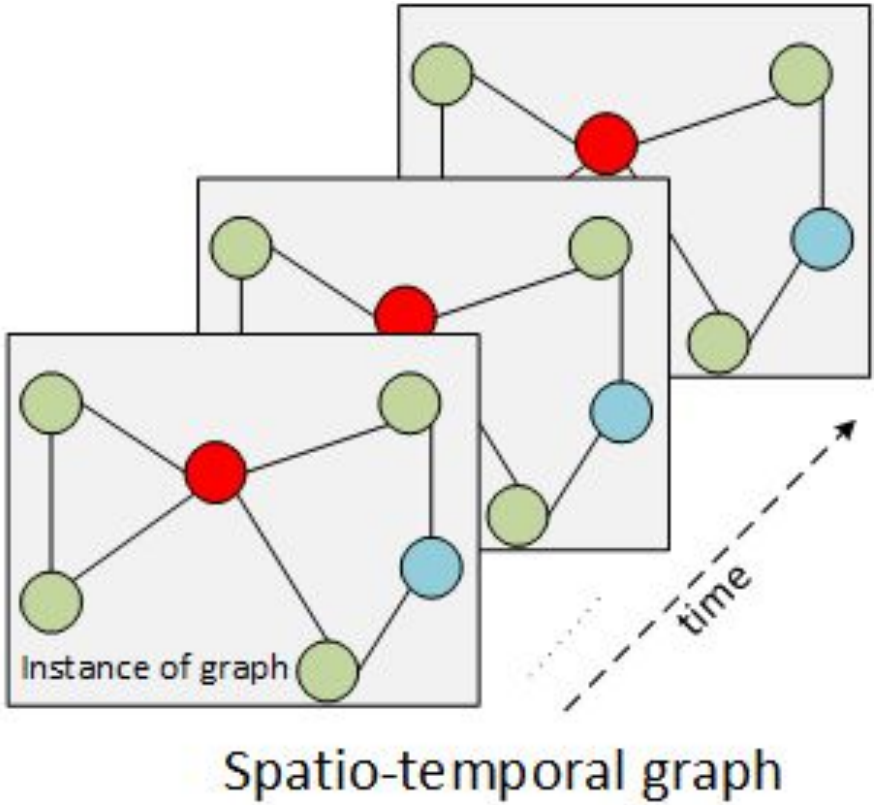
PV Power Plant Systems Dataset

- SS1 + SS2 dataset: 316 power plants
- 2 years of data (730 days)

Power forecasting models

- Power (P_{mp})

Performance Loss Rate (*PLR*)



PV System st-Graph Construction

Calculate Distance Between Two Nodes

- $d_{lon} = lon_2 - lon_1, d_{lat} = lat_2 - lat_1$

  $a = (sin(d_{lat}/2))^2 + cos(lat_1) * cos(lat_2) * (sin(d_{lon}/2))^2$

  $c = 2 * a * tan2(\sqrt{(a)}, \sqrt{(1-a)})$

  $d = R * c$

  where, R is radius of the earth

Convert Elements of Distance Matrix to Weight Matrix

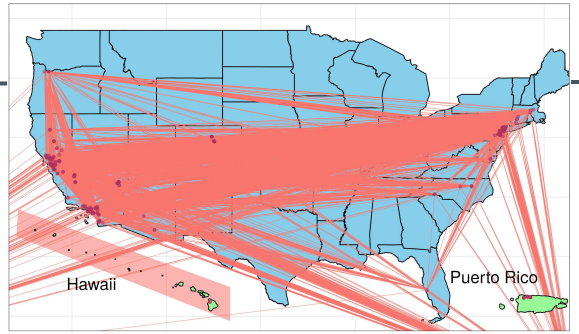
- $\varepsilon_c = 0.5$

$$w_{ij} = \begin{cases} \exp(-\frac{d_{ij}^2}{\sigma^2}), & i \neq j \text{ and } \exp(-\frac{d_{ij}^2}{\sigma^2}) \geq \epsilon \\ 0 & , \text{ otherwise.} \end{cases}$$

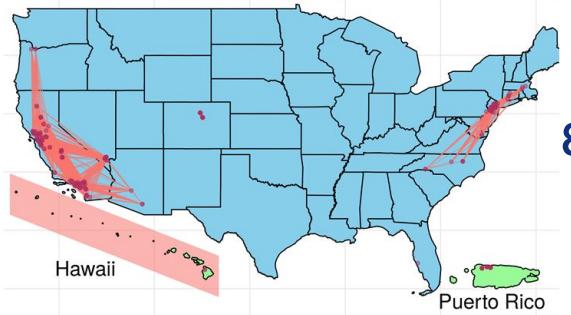
d_{ij} = distance between node i and node j

σ is normalizing constant

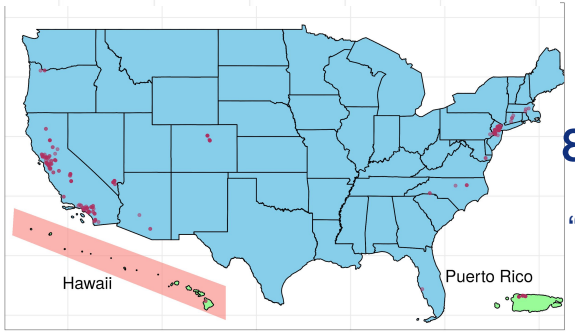
ϵ is constant which control graph sparsity



$\varepsilon_c = 0$



$\varepsilon_c = 0.5$



$\varepsilon_c = 1.0$

"One-system-at-a-time"

H. B. Goodwin, The haversine in nautical astronomy, Naval Institute Proceedings, vol. 36, no. 3 (1910)

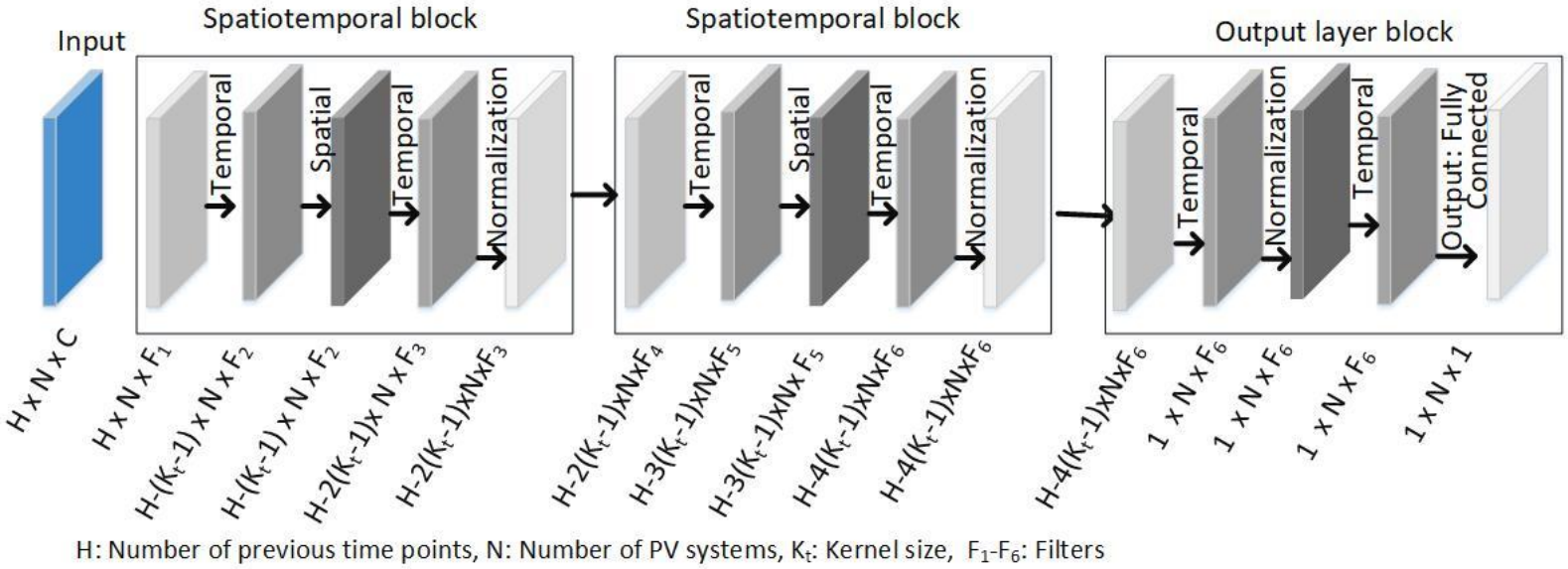
Spatiotemporal Graph Neural Network (st-GNN) Representation

Two Spatiotemporal Blocks

- Two Temporal Convolution Layers
- One Spatial Convolution Layer

Output Layer Block

- Two Temporal Convolution Layers
- Fully Connected Layer



H: Number of previous time points, N: Number of PV systems, K_t: Kernel size, F_1-F_6: Filters

H = 24 number of time lag points

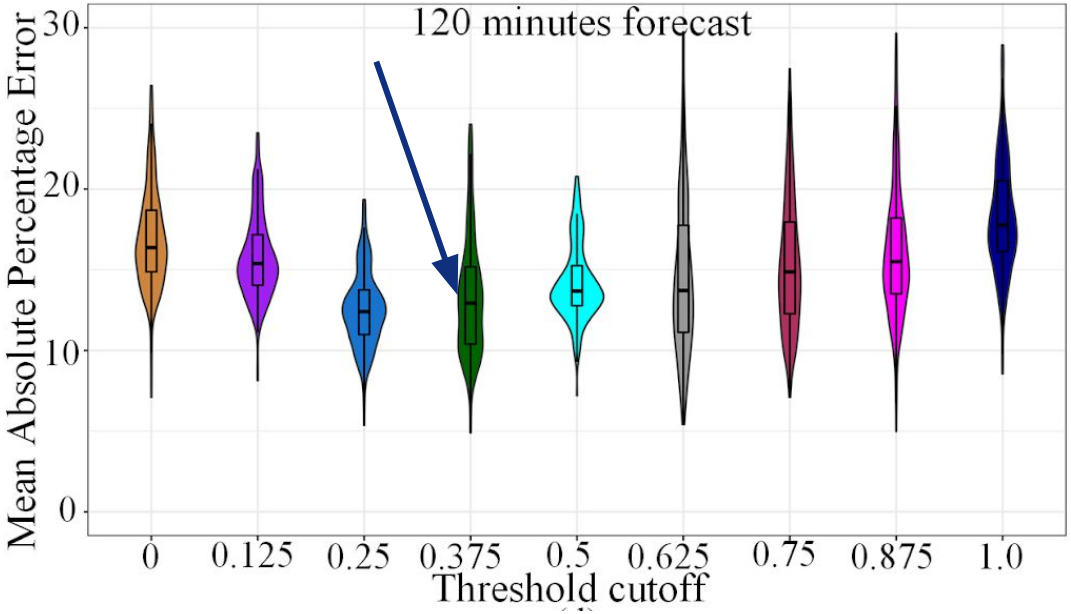
N = 316 PV Power Plant Systems

Trainable parameters:

1 Channel Network: 775,468

Spatiotemporal Graph Conv. Neural Net Model Accuracy

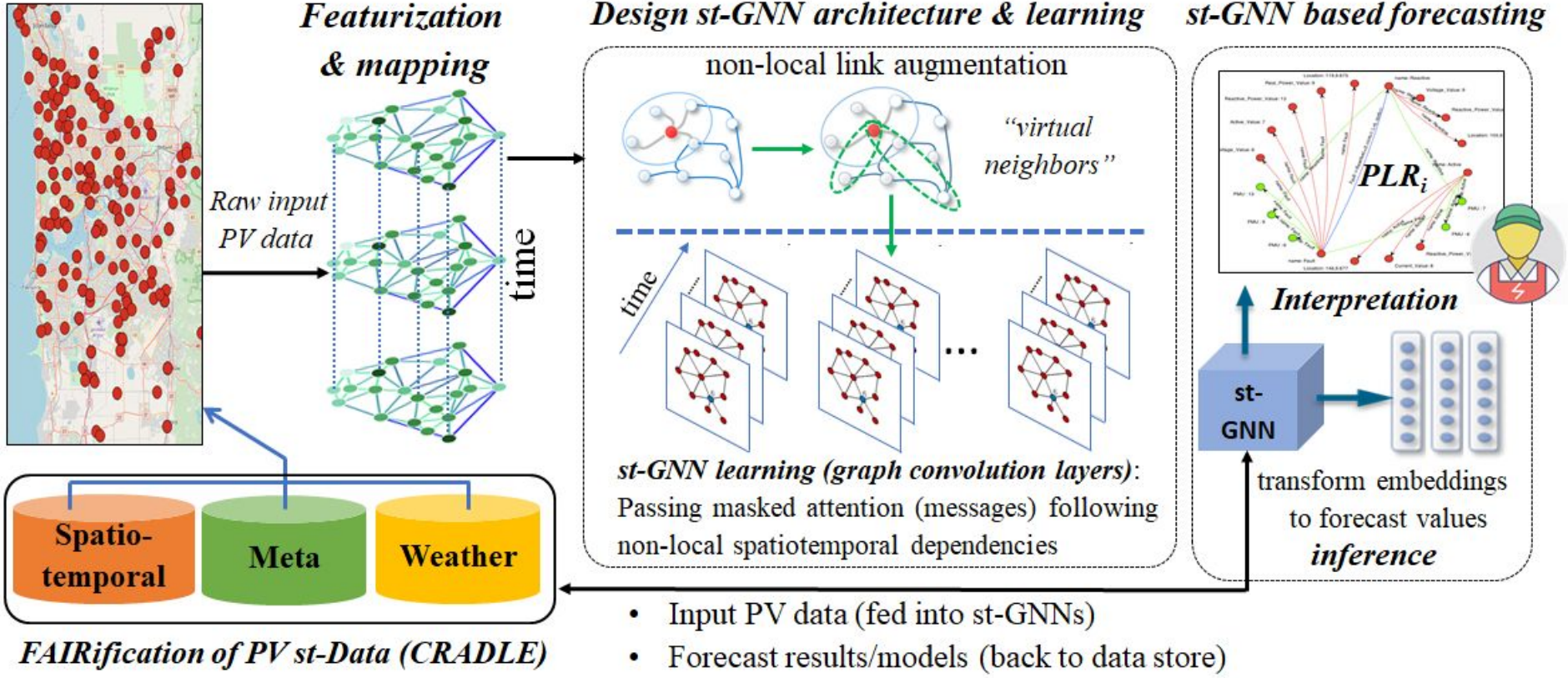
Spatiotemporal GCN & temporal convolution



Forecast (minute)	MAPE for 316 systems			
	s-t convolution		temporal convolution	
	ϵ_c=0.375		ϵ_c=1.0	
	mean	sd	mean	sd
120	11.01	5.04	18.98	5.15
105	9.31	4.36	15.63	4.57
90	8.39	3.87	13.62	4.07
75	7.67	3.36	11.78	3.54
60	7.24	2.87	10.12	2.96
45	6.28	2.61	8.42	2.45
30	4.68	2.48	6.65	2.13
15	2.75	2.37	3.92	2.01

Table 1: Mean and standard deviation of MAPE values for temporal convolution (standalone) vs spatiotemporal convolution for PV systems with optimum ϵ_c for st-GNN network.

New DOE-SETO "AI for PV" project: PV-stGNN for PLR Det.




CASE SCHOOL
OF ENGINEERING
CASE WESTERN RESERVE
UNIVERSITY