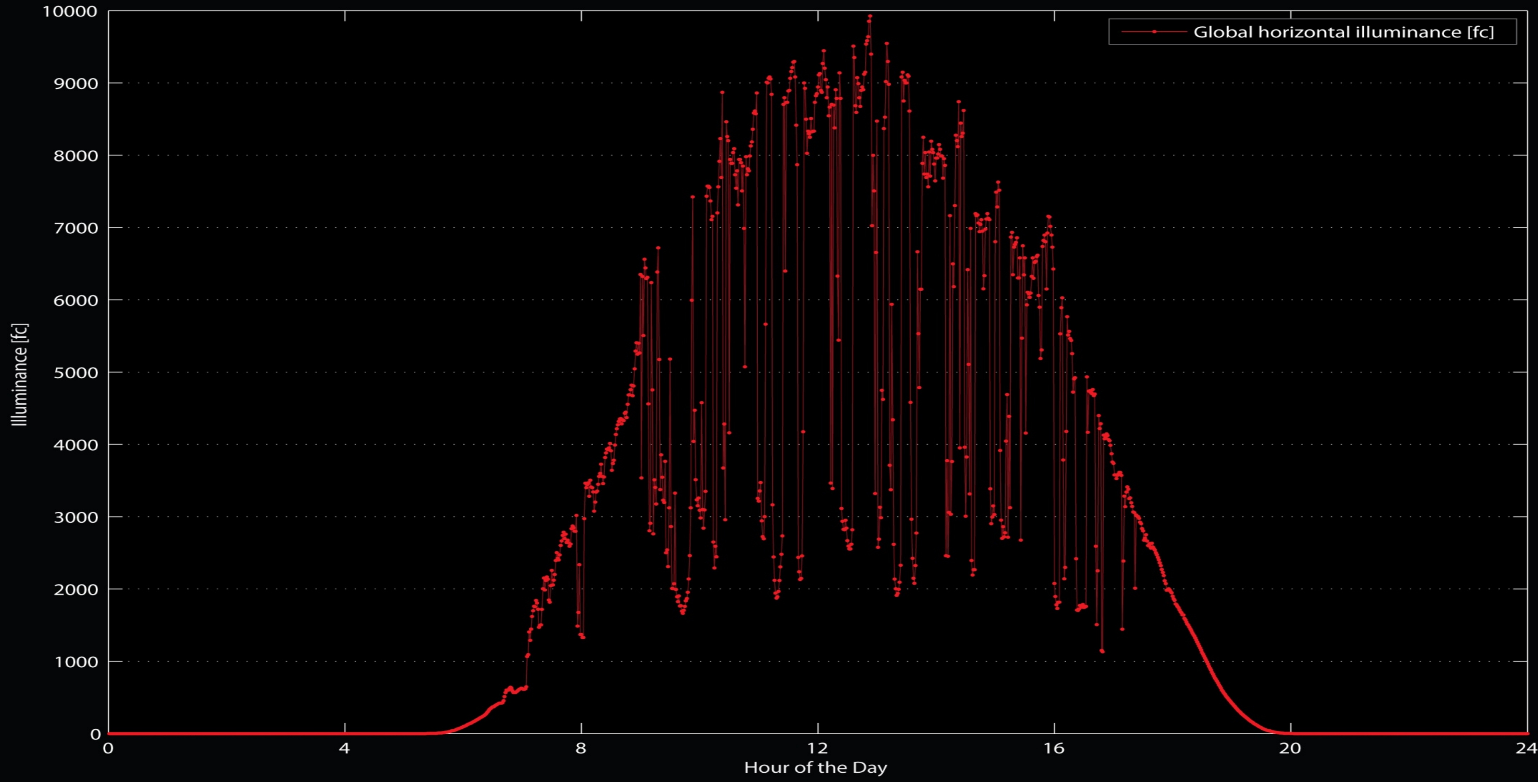
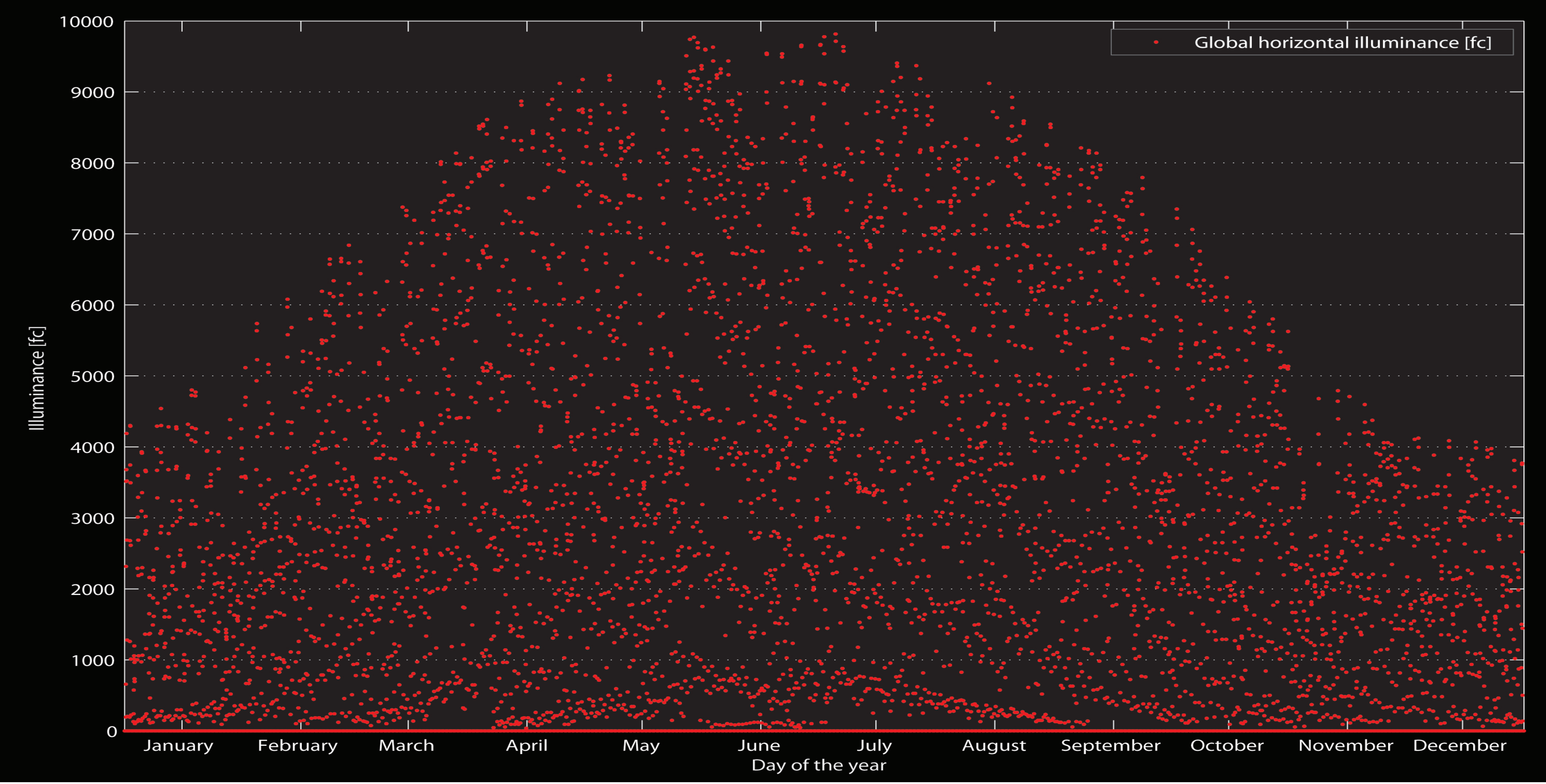
Predicting Solar Irradiance Data Using Machine Learning

Matt Franks, Associate Principal
2019 Radiance Workshop
August 22, 2019

ARUP

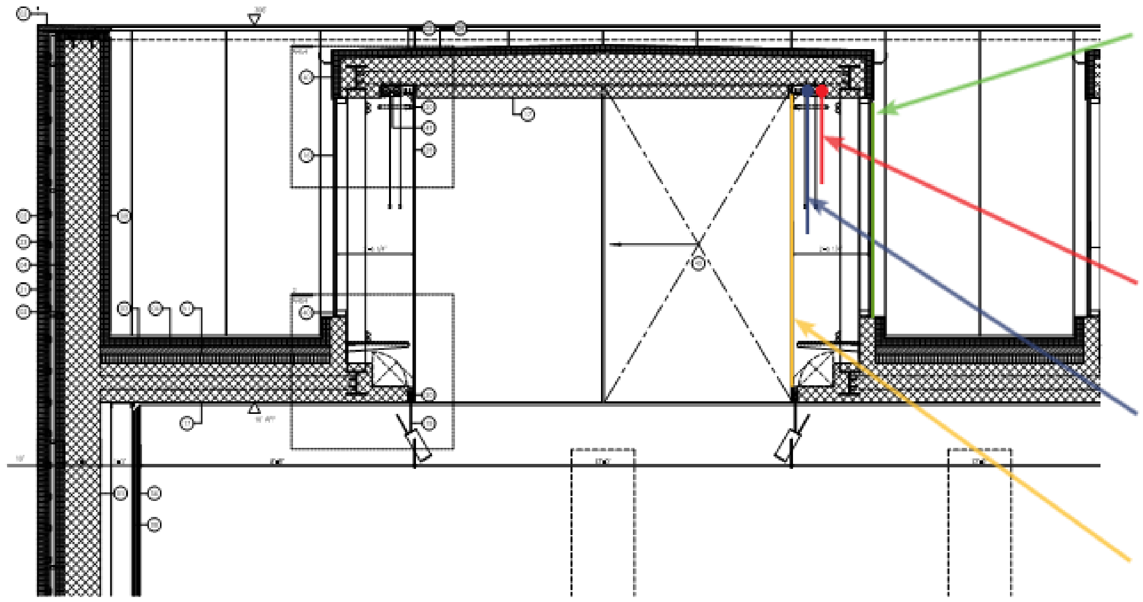

Global horizontal illuminance [fc]
Illuminance [fc]
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
4
8
12
16
20
24
Hour of the Day

One Day



Global horizontal illuminance [fc]
Illuminance [fc]
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
January
February
March
April
May
June
July
August
September
October
November
December
Day of the year

One Year



Section Detail - Gallery 2

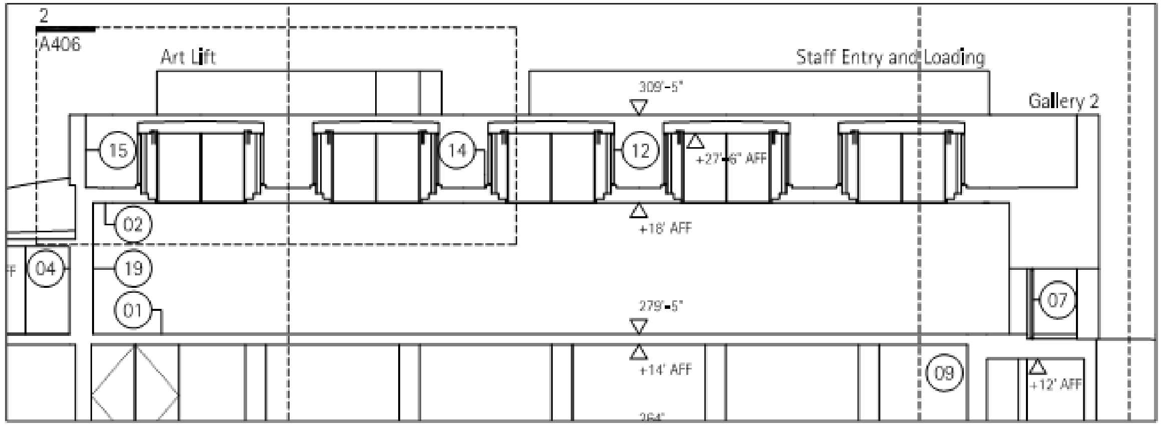
Roof Monitor Layers:

Low iron insulated glazing unit with laminated diffusing inner lite to provide ultra violet filter. Low iron glass is used to maximise colour rendering. Low-e coatings are as neutral in color terms as possible to maintain color rendering of the skylight glass unit of 97 or above. The laminated inner layer will be diffusing to mitigate direct sunlight penetration.

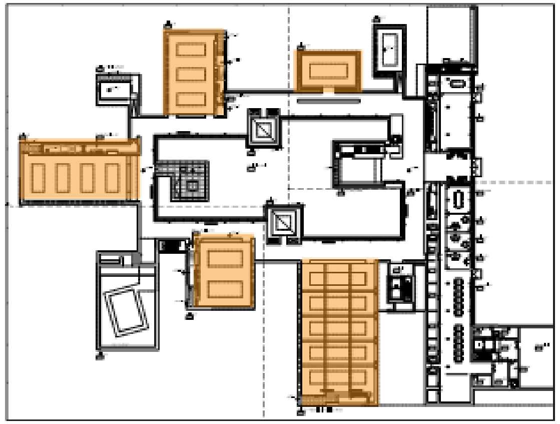
Motorised blackout roller shade to reduce daylight exposure outside museum open hours and allow for flexibility in the allowance of daylight into the gallery.The blackout shade should be provided with side-channels to eliminate light spill around the edges of the shade.

Motorised dimout roller shade to allow for reduction of light levels passing through the skylight system. The shade shall be an open-weave materials with 5% openness and a 10% to 15% visible light transmission, to be determined.

Interior diffusing glass to further diffuse directionality of light and obscure view of structure, roller shades and lighting. This shall be laminated with a diffusing interlayer, and be operable to allow easy access for maintenance. The interior glazing will have an acid-etch finish to reduce interior specular reflections.



Section - Gallery 2



Keyplan

3.6 Galleries 2, 4, 6, 8, and 10

3.6.1 Approach

Galleries 2, 4, 6, 8, and 10 will have similar daylight system designs, consisting of a roof monitor system. Refer to the architectural plans for the arrangement and dimensions of roof monitors in each gallery.

It is expected that Gallery 2 will generally be used to display parts of the permanent collection – typically a mixture of oil paintings, photographs, sculpture.

It is also recognized that Gallery 2 would at times be used for mixed media collections, which means that some works on paper may be displayed along with oil paintings and sculpture. Blackout, if required, is proposed to be provided by the deployment of roller shades in the roof monitors.

Galleries 4, 6, 8, and 10 will be used for more permanent exhibitions. It is understood that upon completion of construction Galleries 4 and 6 will exclude daylight due to the nature of their exhibits, however provisions for daylighting will be included in the design.

3.6.2 Proposed daylight system

The ceiling consists of a roof monitor system which introduces generous but controlled daylight into the gallery below. The images to the left illustrate the proposed system, which consists of a number of layers:

- Exterior vertical diffusing glass, running in the east-west direction
- Interior motorized blackout shade
- Interior motorized roller shade
- Interior diffusing Glass

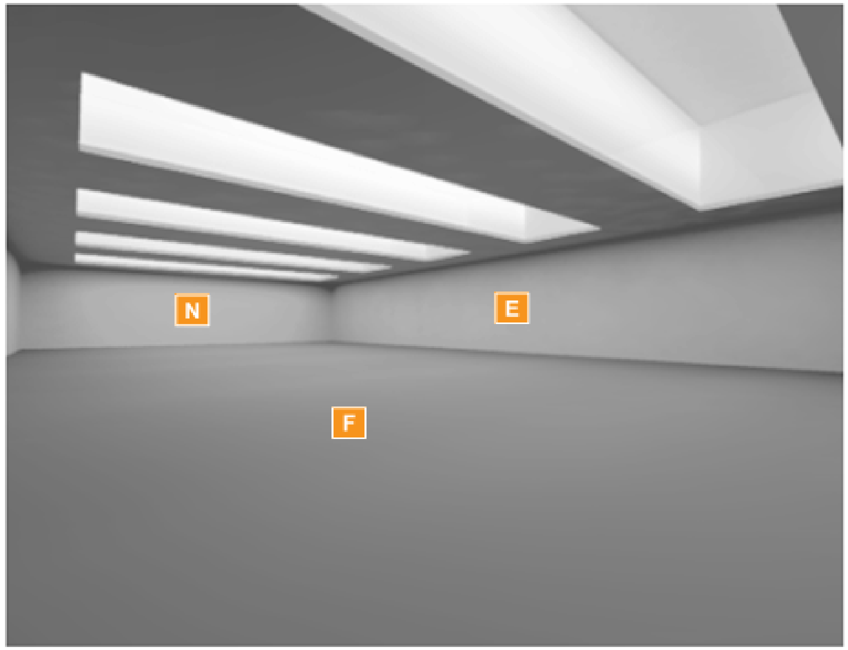
These sets of layers will occur on both the north and south sides of the roof monitor. By allowing sunlight to be diffused through the layers of the southern glazing, and northern skylight to be transmitted and diffused through the north-facing glazing, the lighting conditions in the gallery will vary through out the day as sun position and weather patterns change.

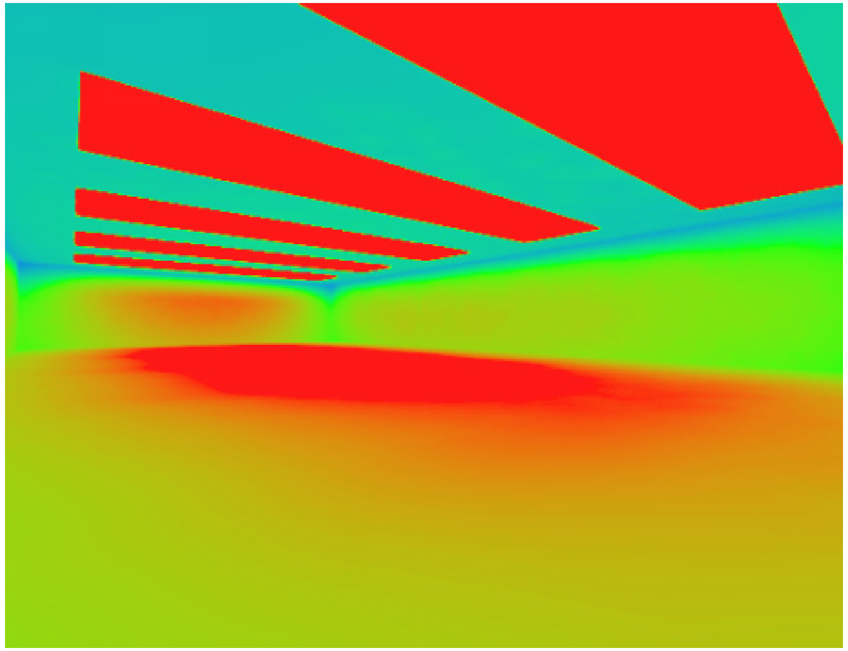
4.4 Gallery 2

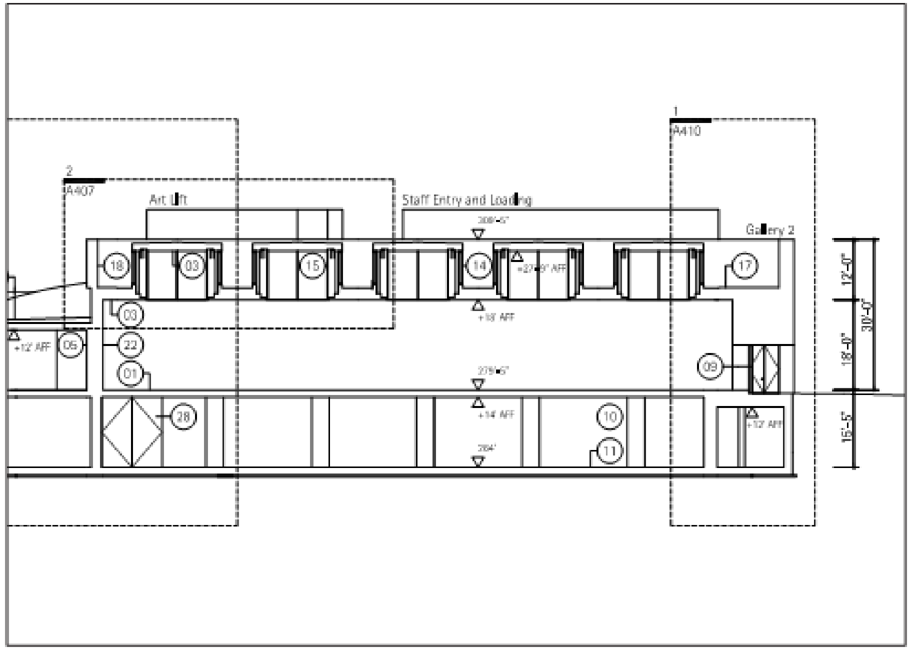
- Outer Glazing Transmittance: 53%
- Inner Glazing Transmittance: 64%
- Wall Reflectance: 75%
- Concrete Reflectance: 70%
- Floor Reflectance: 50%
- Calculation Time: 12:00 p.m. on date indicated
- Measurement points are at location indicated in images. Unless otherwise indicated images shown are for March 31, overcast conditions.

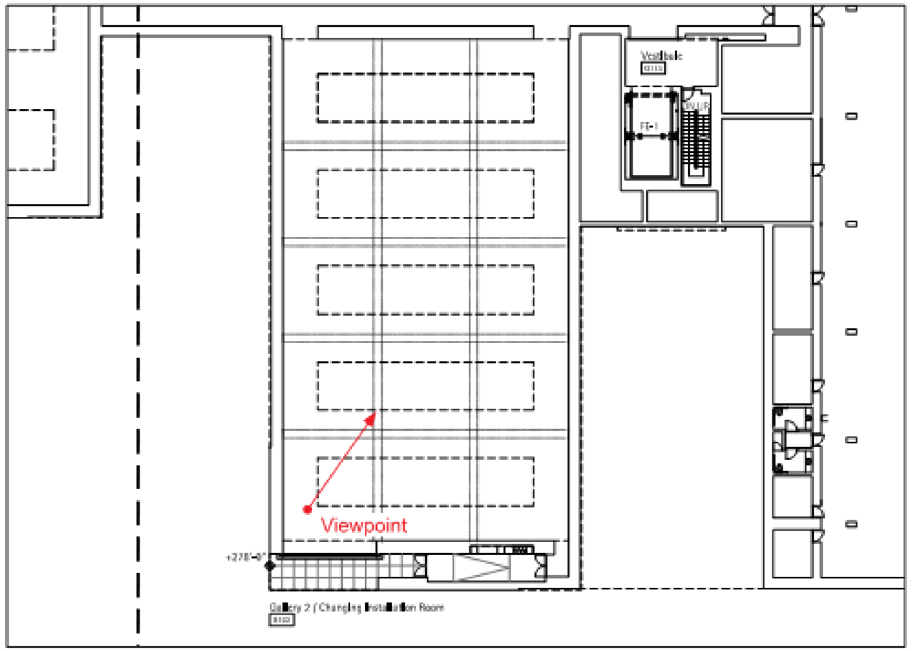
Day	Weather	N (fc)	E (fc)	S (fc)	W (fc)	F (fc)
Mar 21	Overcast	114	93	106	94	143
	Sunny	351	257	220	262	388
Jun 21	Overcast	119	100	116	101	153
	Sunny	266	200	215	203	298
Dec 21	Overcast	72	59	65	59	91
	Sunny	149	119	92	118	186

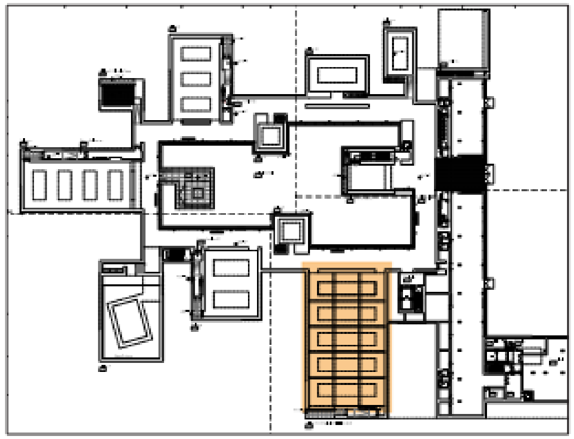
N, E, S, W, measurement points are on North, East, South, and West walls respectively; F measurement point is horizontal illuminance on floor. S and W points are not pictured.



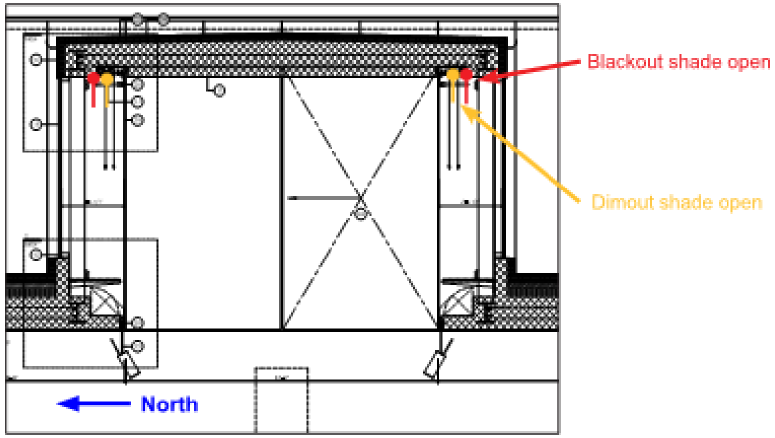








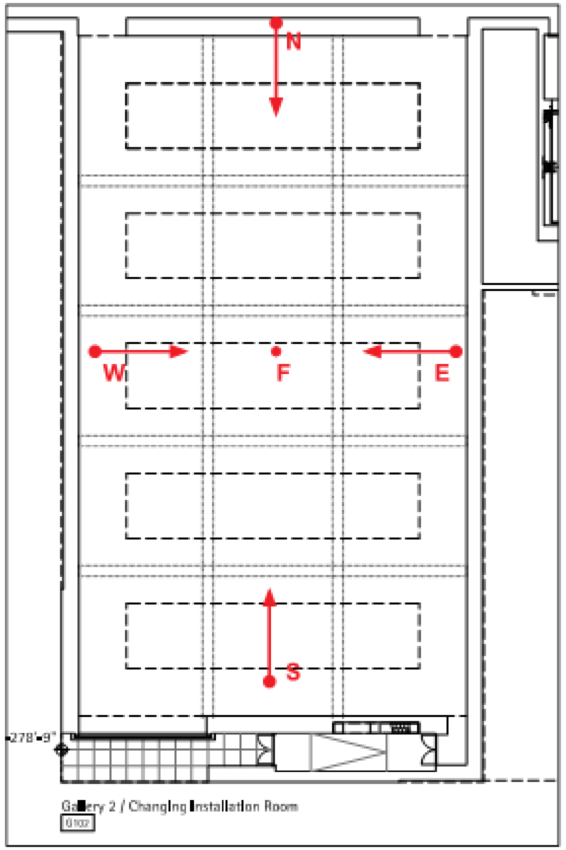
Keyplan



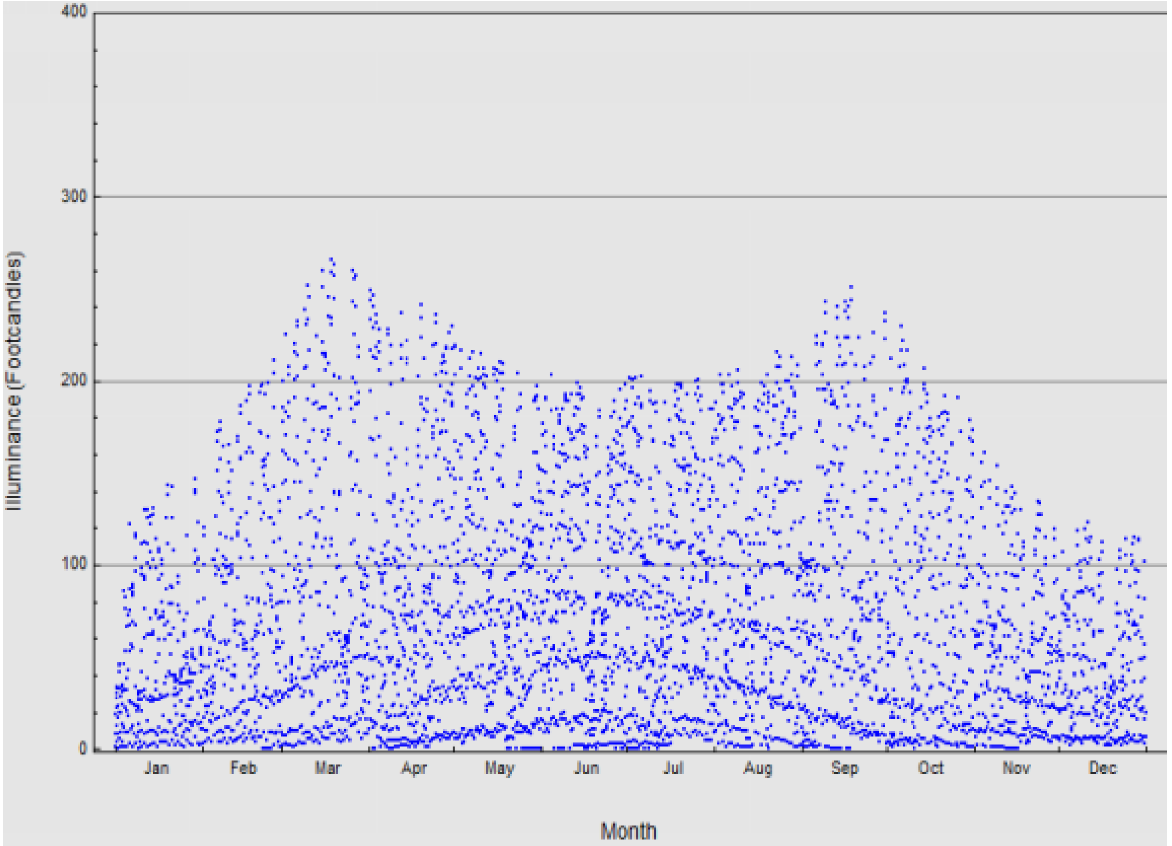
Gallery 2 - Roof Monitor Section

Day	Weather	North (fc)	East (fc)	South (fc)	West (fc)	Floor (fc)
Mar 21, 12:00 p.m.	Overcast	114	93	106	94	143
	Sunny	351	257	220	262	388

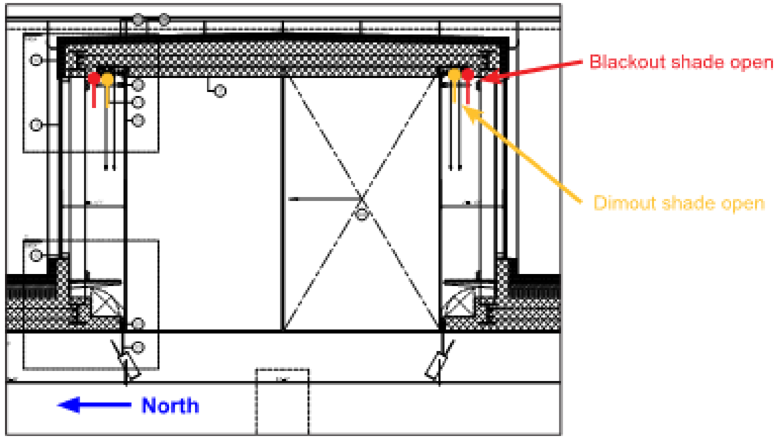
	North (k-fc-hr)	East (k-fc-hr)	South (k-fc-hr)	West (k-fc-hr)	Floor (k-fc-hr)
Annual Cumulative Exposure	457	350	362	353	524



Gallery 2 - Plan



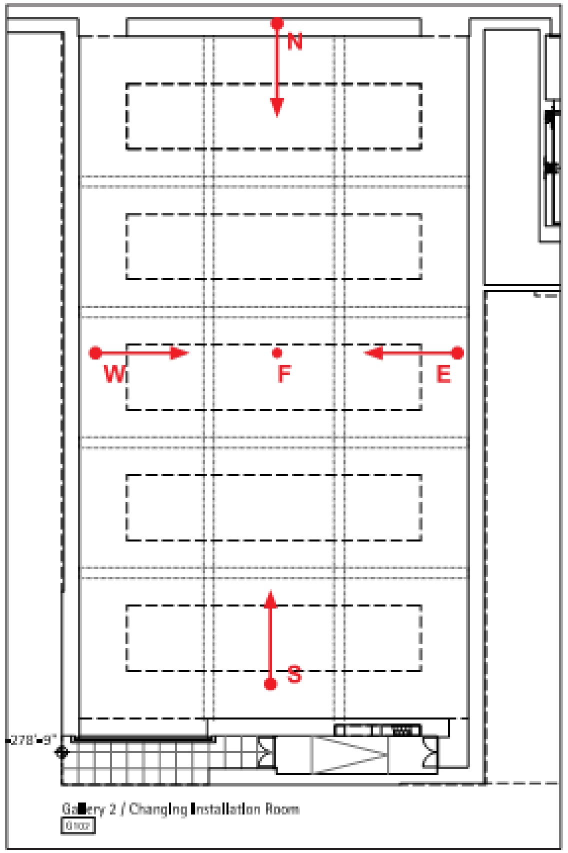
Annual Illuminance Profile on East Wall at Point E



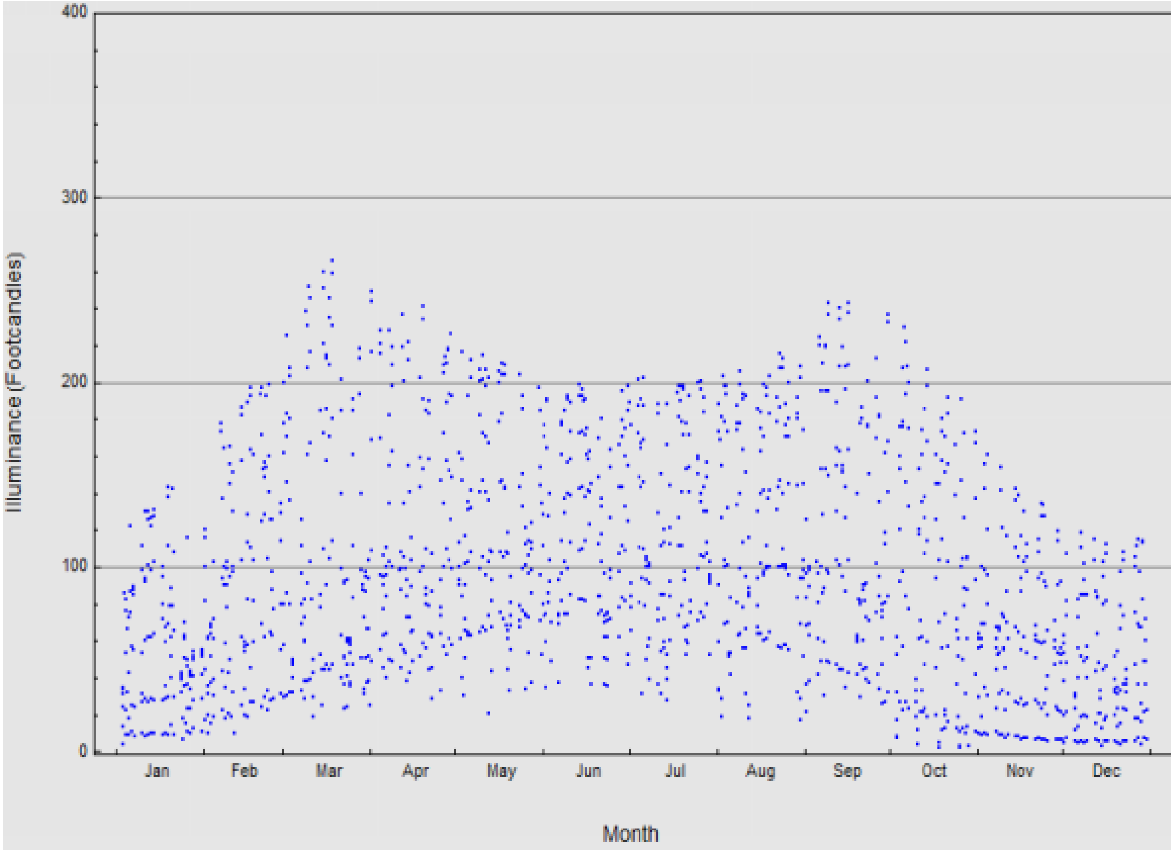
Gallery 2 - Roof Monitor Section

Day	Weather	North (fc)	East (fc)	South (fc)	West (fc)	Floor (fc)
Mar 21, 12:00 p.m.	Overcast	114	93	106	94	143
	Sunny	351	257	220	262	388

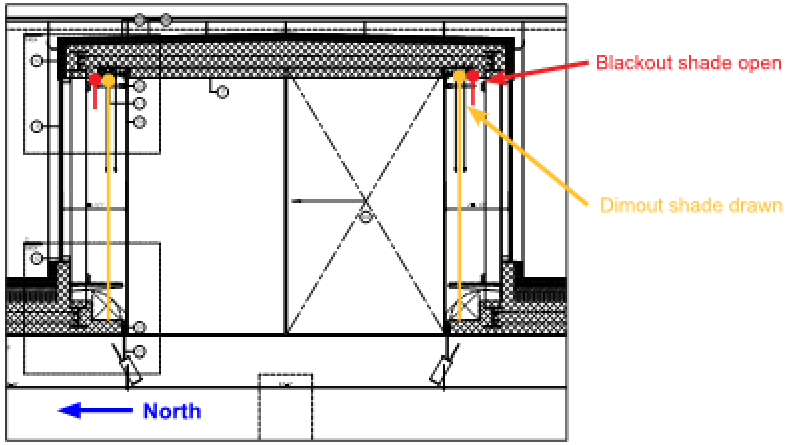
	North (k-fc-hr)	East (k-fc-hr)	South (k-fc-hr)	West (k-fc-hr)	Floor (k-fc-hr)
Annual Cumulative Exposure	199	152	156	154	227



Gallery 2 - Plan



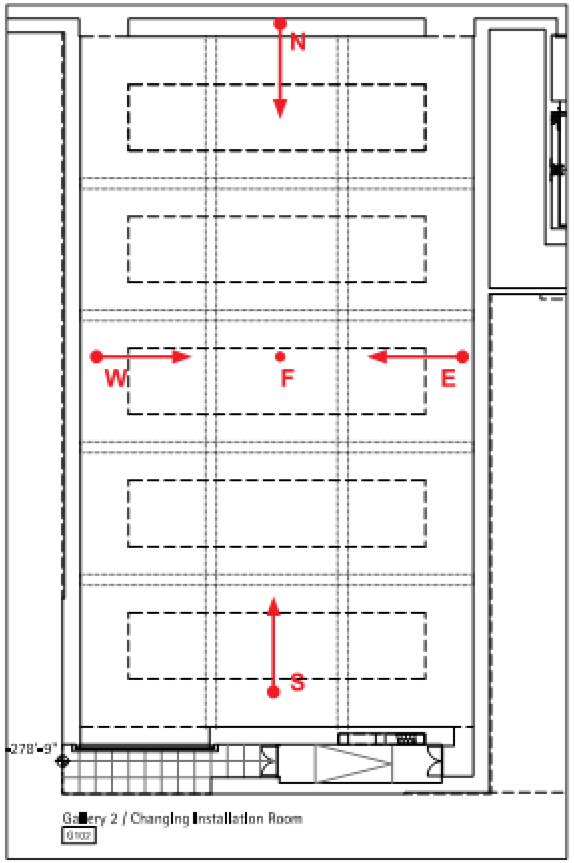
Annual Illuminance Profile on East Wall at Point E



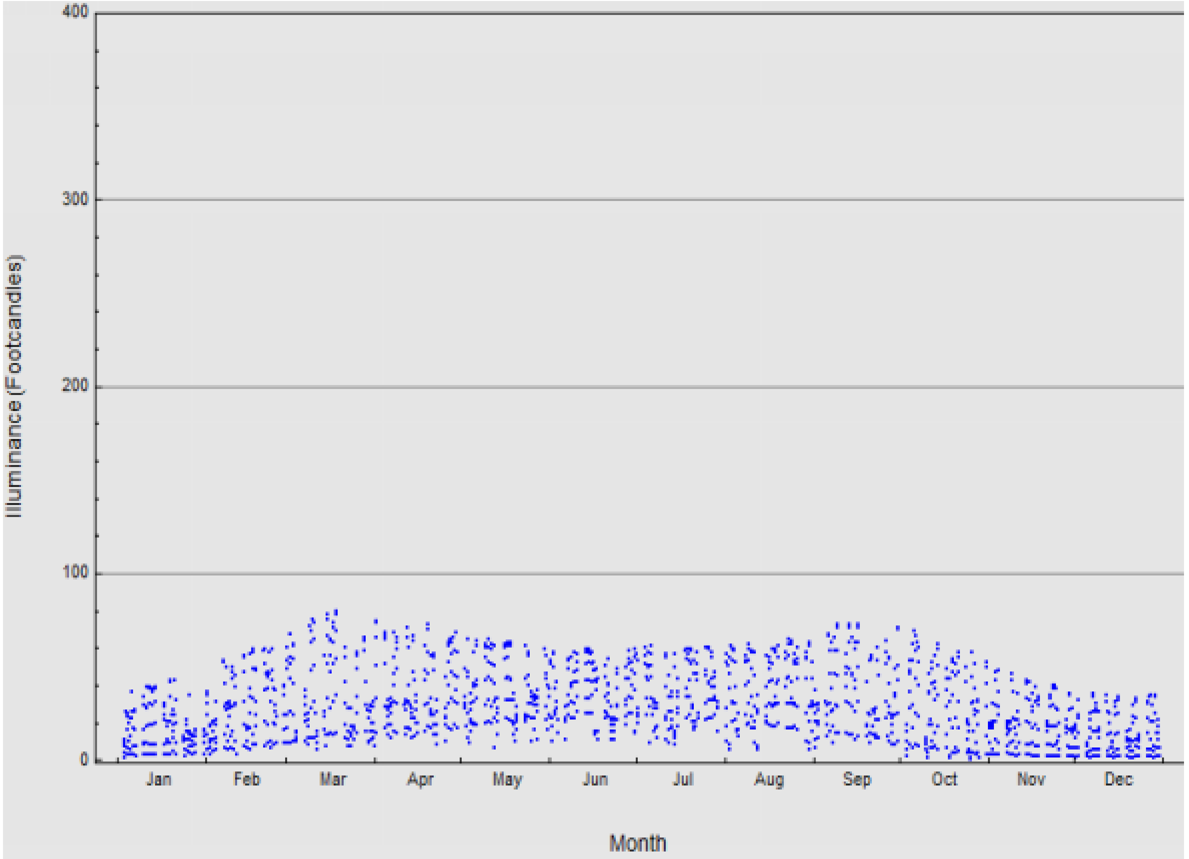
Gallery 2 - Roof Monitor Section

Day	Weather	North (fc)	East (fc)	South (fc)	West (fc)	Floor (fc)
Mar 21, 12:00 p.m.	Overcast	33	27	31	28	42
	Sunny	123	79	66	79	116

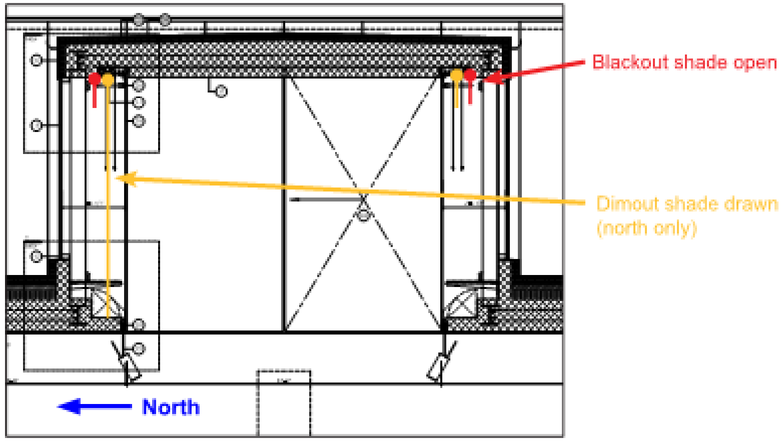
	North (k-fc-hr)	East (k-fc-hr)	South (k-fc-hr)	West (k-fc-hr)	Floor (k-fc-hr)
Annual Cumulative Exposure	59	46	47	46	68



Gallery 2 - Plan



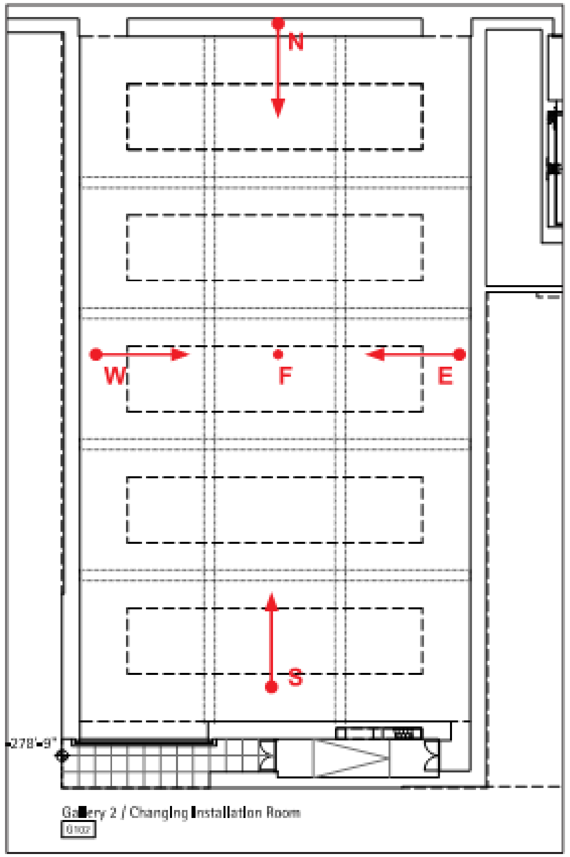
Annual Illuminance Profile on East Wall at Point E



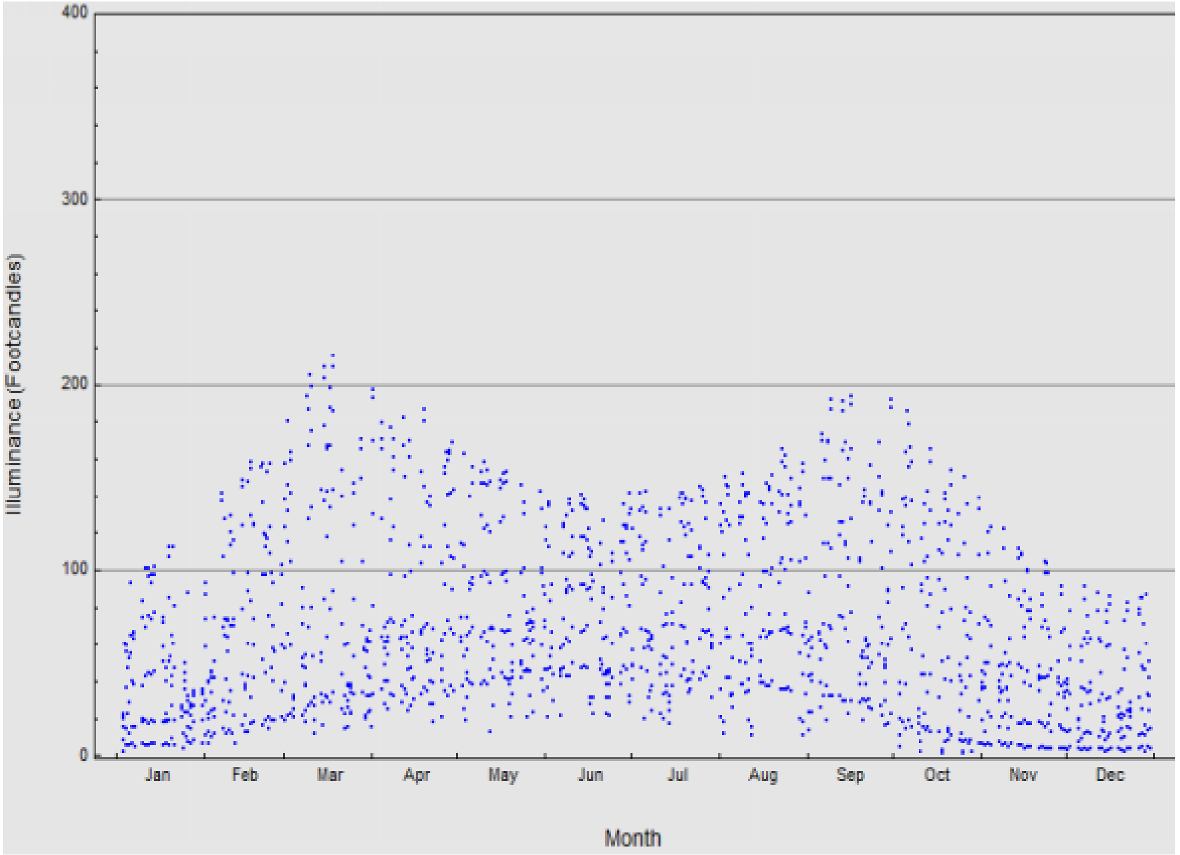
Gallery 2 - Roof Monitor Section

Day	Weather	North (fc)	East (fc)	South (fc)	West (fc)	Floor (fc)
Mar 21, 12:00 p.m.	Overcast	90	60	50	61	91
	Sunny	345	212	138	214	293

	North (k-fc-hr)	East (k-fc-hr)	South (k-fc-hr)	West (k-fc-hr)	Floor (k-fc-hr)
Annual Cumulative Exposure	169	108	81	109	163



Gallery 2 - Plan



Annual Illuminance Profile on East Wall at Point E

Why might reality be different than what was predicted?

- Real reflectances differ from those assumed
- Dirt more or less than assumed
- Constructed dimensions differ from design
- Inaccuracy of calculation methods

Why might reality be different than what was predicted?

• Real reflectances differ from those assumed	**+/- 5%**
• Dirt more or less than assumed	**+/- 5%**
• Constructed dimensions differ from design	**+/- 1%**
• Inaccuracy of calculation methods	**+/- 5%**

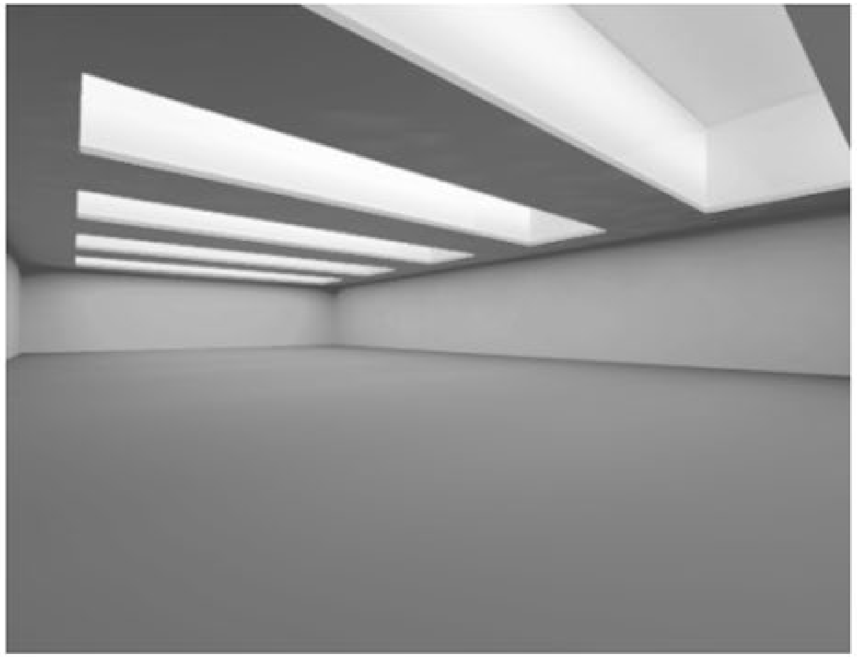
X45 1841
ESCAPE
JLG

GS-1930

2.3 Illuminance Distribution

The images on this page show comparisons between the computer model daylight distribution simulation, both in greyscale and falsecolor luminance.

The luminance distribution images show reasonable uniformity as well as agreement with the computer simulated distribution. Note that the slight dropoff in the center of wall on the right side of the image is due to the model construction, which consists of a mirror to replicate the appearance of two additional clerestories.



Computer Simulation - Gallery 2 - Greyscale

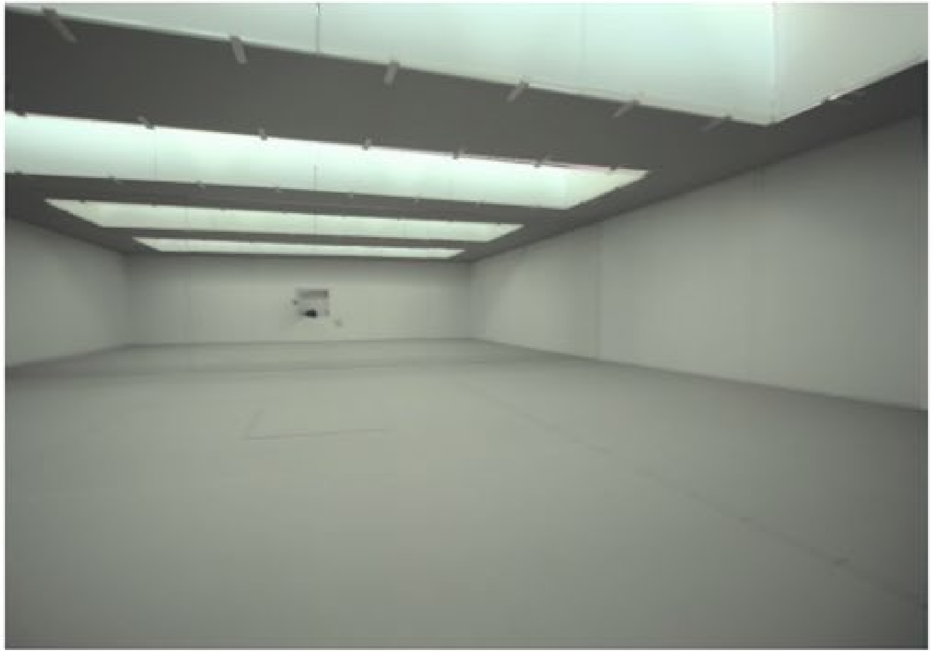
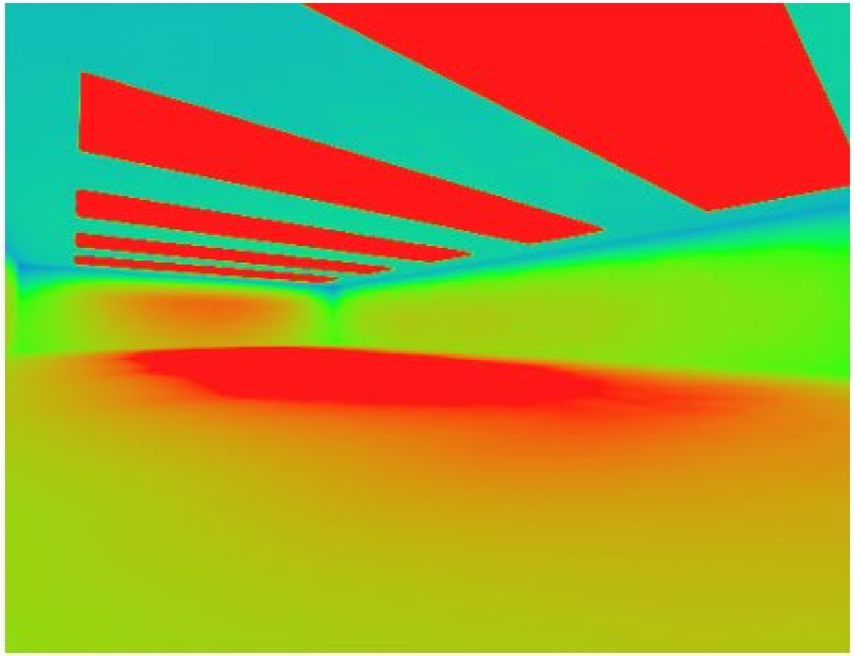


Photo - Gallery 2 Model



Computer Simulation - Gallery 2 - Falsecolor

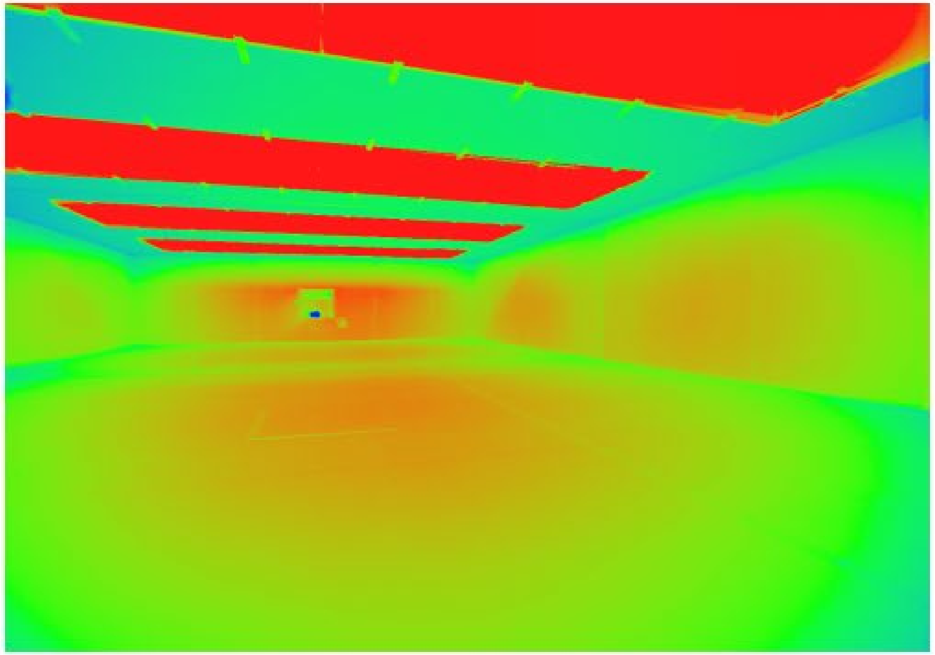
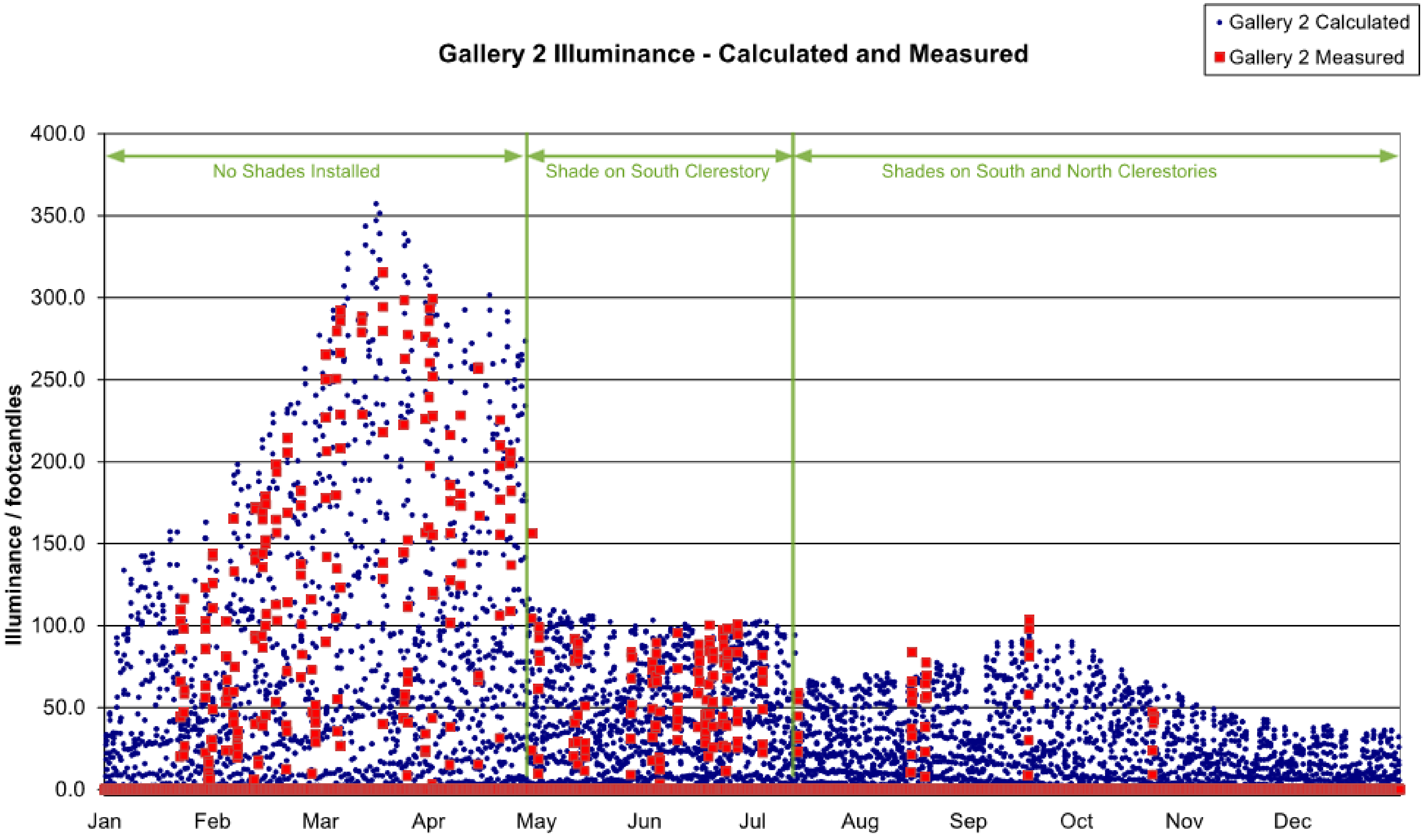


Photo - Gallery 2 Model - Falsecolor



2 Gallery 2

2.1 Illuminance Levels - Annual

The scatter plot on the left side of this pages shows a comparison between the hourly illuminance data calculated for the north wall of Gallery 2 for each hour of the year based on typical weather data from the November 15, 2012 daylighting report (blue dots), overlayed with hourly data measured from the Gallery 2 model on days that measurements was possible (red squares).

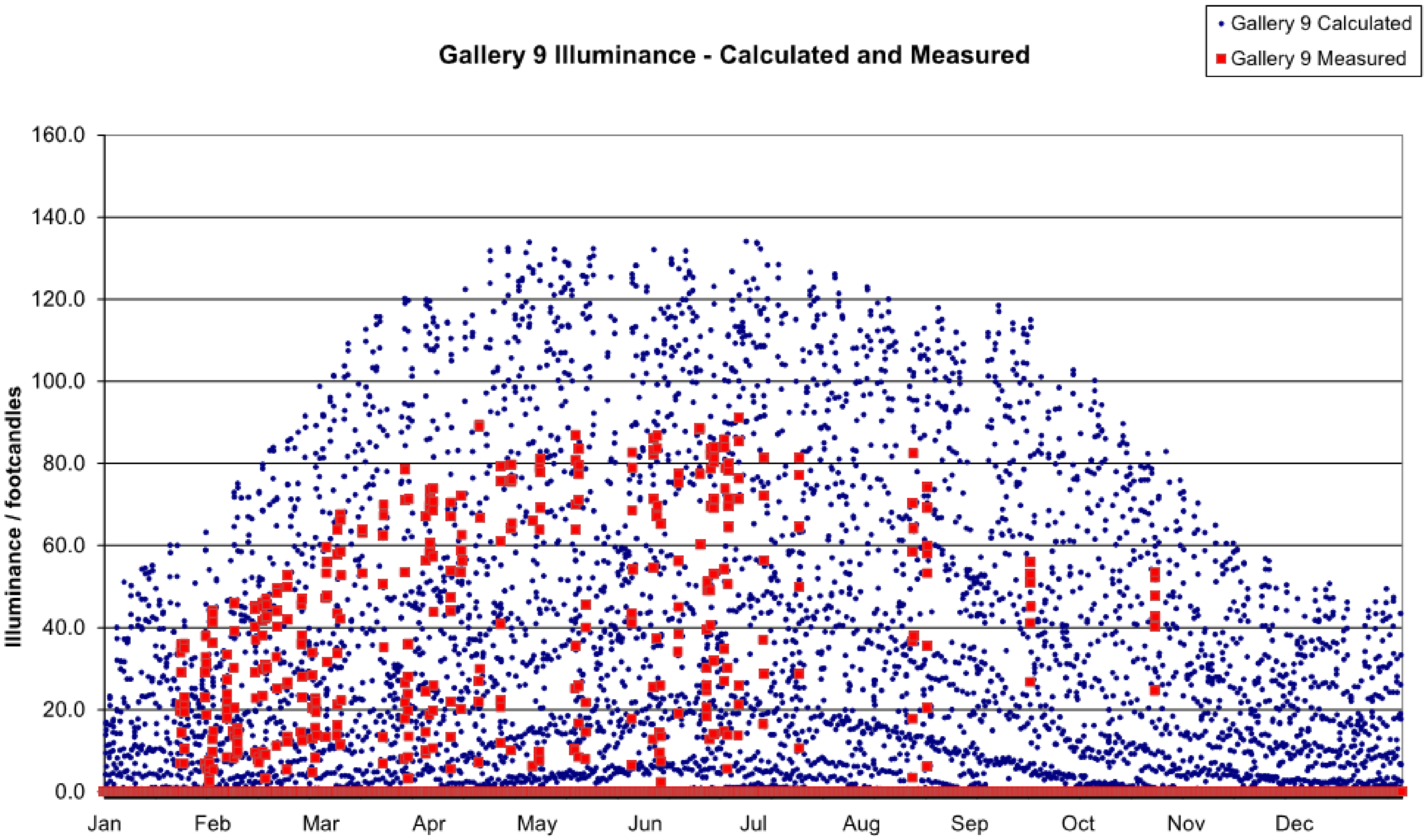
Indicated on the scatter plot is the times that the three different shade configurations were installed in the model.

- No shades from January to April 30.
- Shades on only the south clerestory from May 1 to July 14.
- Shades on both the north and south clerestories for the remainder of the year.

The shade material used in the model was the shade material currently specified, Mermet Screen Vision:

- 10% openness
- 29% visible light transmittance
- white color

The general trend of the data indicates fairly close correlation between the computer model and the measured illuminance, with illuminance peaks at similar levels.



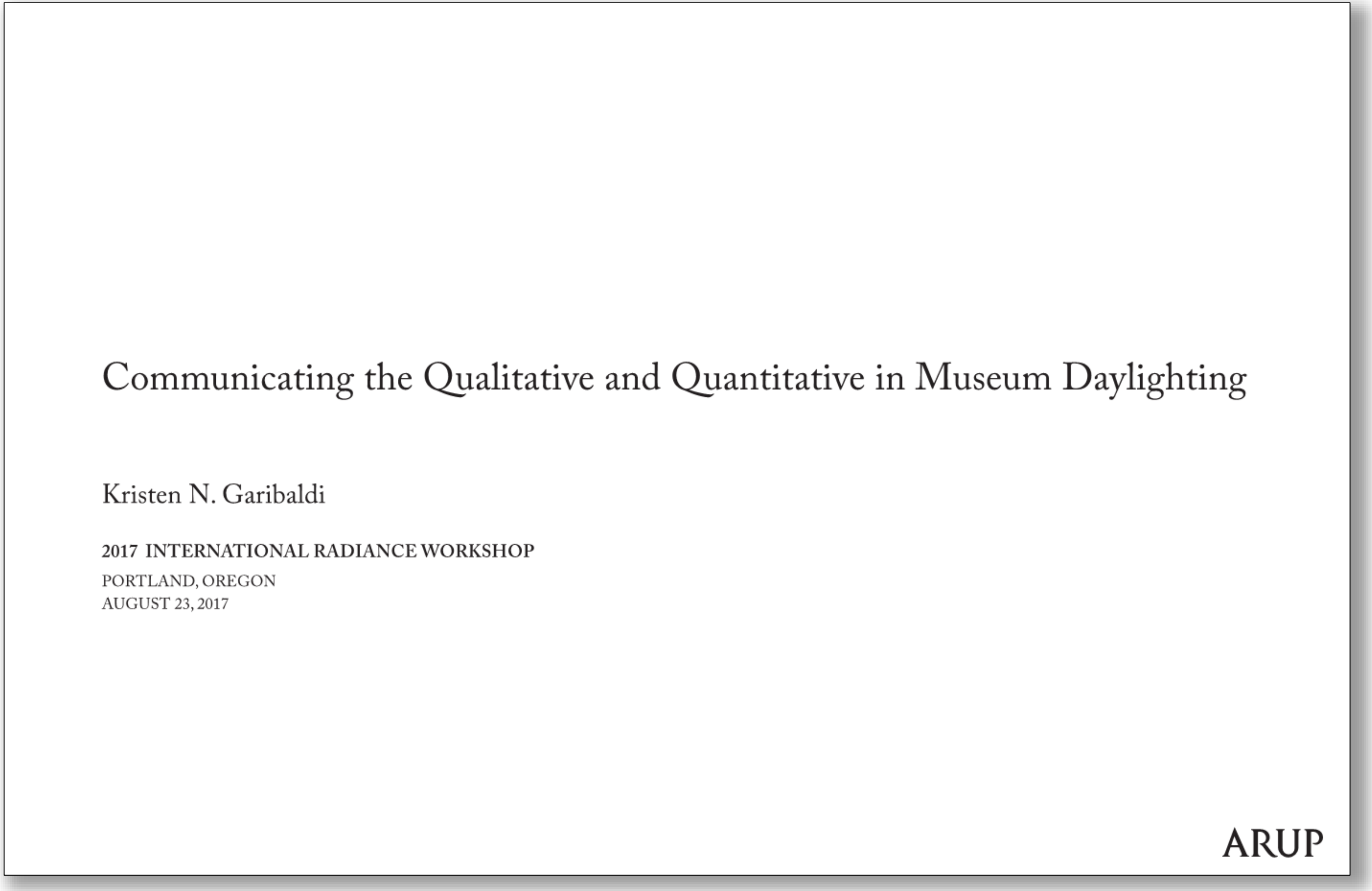
3 Gallery 9

3.1 Illuminance Levels - Annual

The scatter plot on the left side of this pages shows a comparison between the hourly illuminance data calculated for the west wall of Gallery 9 for each hour of the year based on typical weather data from the November 15, 2012 daylighting report (blue dots), overlayed with hourly data measured from the Gallery 9 model on days that measurements was possible (red squares).

The general trend of the data indicates fairly close correlation between the computer model and measurements in relative terms, however it can be seen from scatter plot that the gallery 9 measurements are in the range of 30% lower than predicted.

There are several factors that may be contributing to the difference between the calculated and measured values. These are discussed on the following pages.



Communicating the Qualitative and Quantitative in Museum Daylighting

Kristen N. Garibaldi

2017 INTERNATIONAL RADIANCE WORKSHOP

PORTLAND, OREGON
AUGUST 23, 2017

ARUP

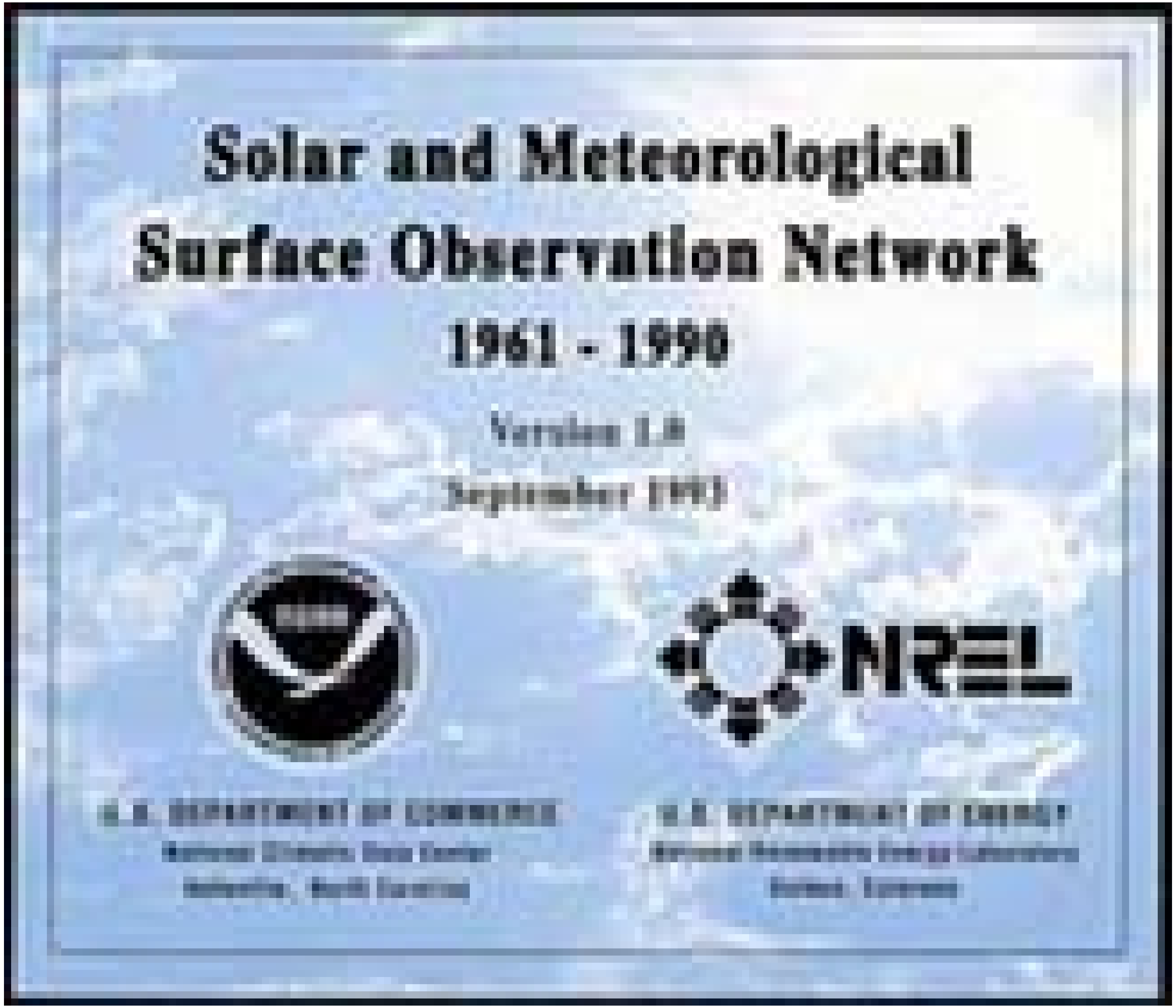
Problem (?)

- It is difficult to measure direct and diffuse illuminance (irradiance) separately.

$200
1” diameter
2 ounces

$6,000
6” diameter
2 lbs

But…



Solar and Meteorological
Surface Observation Network
1961 - 1990

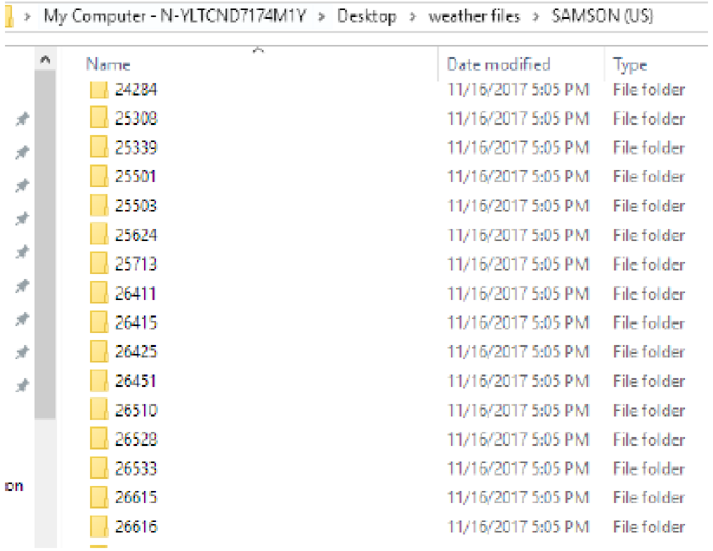
Version 1.0
September 1993

U.S. DEPARTMENT OF COMMERCE
National Climatic Data Center
Asheville, North Carolina

U.S. DEPARTMENT OF ENERGY
National Renewable Energy Laboratory
Golden, Colorado

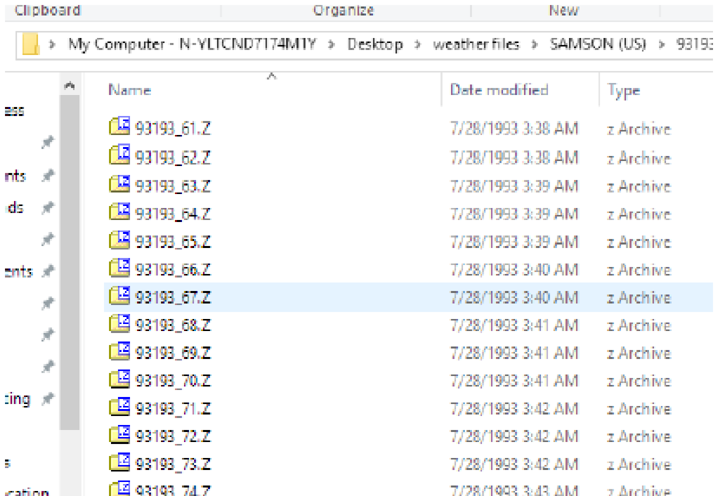
DATA SET	
WHAT	Weather Data
FORMAT	TMY
LOCATION	Local Storage

236 Folders


My Computer - N-YLTCND7174M1Y › Desktop › weather files › SAMSON (US)
on

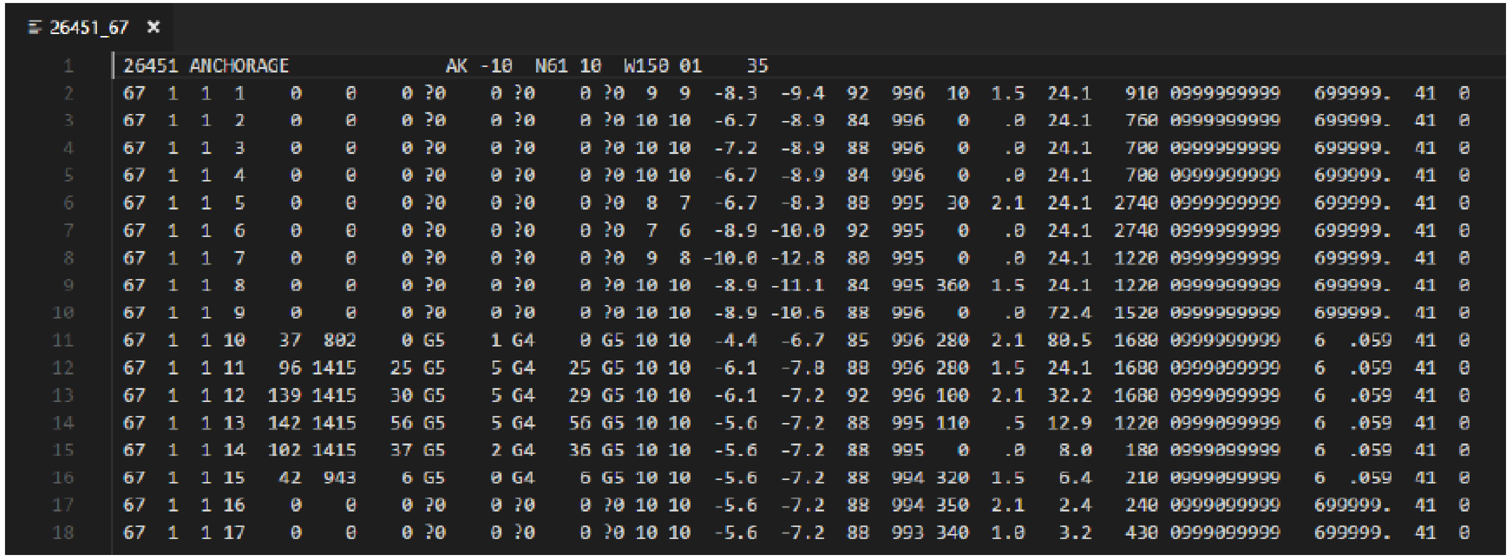
Name	Date modified	Type
24284	11/16/2017 5:05 PM	File folder
25308	11/16/2017 5:05 PM	File folder
25339	11/16/2017 5:05 PM	File folder
25501	11/16/2017 5:05 PM	File folder
25503	11/16/2017 5:05 PM	File folder
25624	11/16/2017 5:05 PM	File folder
25713	11/16/2017 5:05 PM	File folder
26411	11/16/2017 5:05 PM	File folder
26415	11/16/2017 5:05 PM	File folder
26425	11/16/2017 5:05 PM	File folder
26451	11/16/2017 5:05 PM	File folder
26510	11/16/2017 5:05 PM	File folder
26528	11/16/2017 5:05 PM	File folder
26533	11/16/2017 5:05 PM	File folder
26615	11/16/2017 5:05 PM	File folder
26616	11/16/2017 5:05 PM	File folder

30 Files per Folder


Clipboard
Organize
New
My Computer - N-YLTCND7174M1Y › Desktop › weather files › SAMSON (US) › 93193
ess
nts
ds
ents
ting
s
cation

Name	Date modified	Type
93193_61.Z	7/28/1993 3:38 AM	z Archive
93193_62.Z	7/28/1993 3:38 AM	z Archive
93193_63.Z	7/28/1993 3:39 AM	z Archive
93193_64.Z	7/28/1993 3:39 AM	z Archive
93193_65.Z	7/28/1993 3:39 AM	z Archive
93193_66.Z	7/28/1993 3:40 AM	z Archive
93193_67.Z	7/28/1993 3:40 AM	z Archive
93193_68.Z	7/28/1993 3:41 AM	z Archive
93193_69.Z	7/28/1993 3:41 AM	z Archive
93193_70.Z	7/28/1993 3:41 AM	z Archive
93193_71.Z	7/28/1993 3:42 AM	z Archive
93193_72.Z	7/28/1993 3:42 AM	z Archive
93193_73.Z	7/28/1993 3:42 AM	z Archive
93193_74.Z	7/28/1993 3:43 AM	z Archive

1 header row (same for 30 files)



```
26451 ANCHORAGE                AK -10  N61 10  W150 01     35
67  1  1  1    0    0    0 ?0    0 ?0    0 ?0  9  9  -8.3  -9.4  92  996  10  1.5  24.1   910 0999999999   699999.  41  0
67  1  1  2    0    0    0 ?0    0 ?0    0 ?0 10 10  -6.7  -8.9  84  996   0   .0  24.1   760 0999999999   699999.  41  0
67  1  1  3    0    0    0 ?0    0 ?0    0 ?0 10 10  -7.2  -8.9  88  996   0   .0  24.1   700 0999999999   699999.  41  0
67  1  1  4    0    0    0 ?0    0 ?0    0 ?0 10 10  -6.7  -8.9  84  996   0   .0  24.1   700 0999999999   699999.  41  0
67  1  1  5    0    0    0 ?0    0 ?0    0 ?0  8  7  -6.7  -8.3  88  995  30  2.1  24.1  2740 0999999999   699999.  41  0
67  1  1  6    0    0    0 ?0    0 ?0    0 ?0  7  6  -8.9 -10.0  92  995   0   .0  24.1  2740 0999999999   699999.  41  0
67  1  1  7    0    0    0 ?0    0 ?0    0 ?0  9  8 -10.0 -12.8  80  995   0   .0  24.1  1220 0999999999   699999.  41  0
67  1  1  8    0    0    0 ?0    0 ?0    0 ?0 10 10  -8.9 -11.1  84  995 360  1.5  24.1  1220 0999999999   699999.  41  0
67  1  1  9    0    0    0 ?0    0 ?0    0 ?0 10 10  -8.9 -10.6  88  996   0   .0  72.4  1520 0999999999   699999.  41  0
67  1  1 10   37  802    0 G5    1 G4    0 G5 10 10  -4.4  -6.7  85  996 280  2.1  80.5  1680 0999999999   6   .059  41  0
67  1  1 11   96 1415   25 G5    5 G4   25 G5 10 10  -6.1  -7.8  88  996 280  1.5  24.1  1680 0999099999   6   .059  41  0
67  1  1 12  139 1415   30 G5    5 G4   29 G5 10 10  -6.1  -7.2  92  996 100  2.1  32.2  1680 0999099999   6   .059  41  0
67  1  1 13  142 1415   56 G5    5 G4   56 G5 10 10  -5.6  -7.2  88  995 110   .5  12.9  1220 0999099999   6   .059  41  0
67  1  1 14  102 1415   37 G5    2 G4   36 G5 10 10  -5.6  -7.2  88  995   0   .0   8.0   180 0999099999   6   .059  41  0
67  1  1 15   42  943    6 G5    0 G4    6 G5 10 10  -5.6  -7.2  88  994 320  1.5   6.4   210 0999099999   6   .059  41  0
67  1  1 16    0    0    0 ?0    0 ?0    0 ?0 10 10  -5.6  -7.2  88  994 350  2.1   2.4   240 0999099999   699999.  41  0
67  1  1 17    0    0    0 ?0    0 ?0    0 ?0 10 10  -5.6  -7.2  88  993 340  1.0   3.2   430 0999099999   699999.  41  0
```

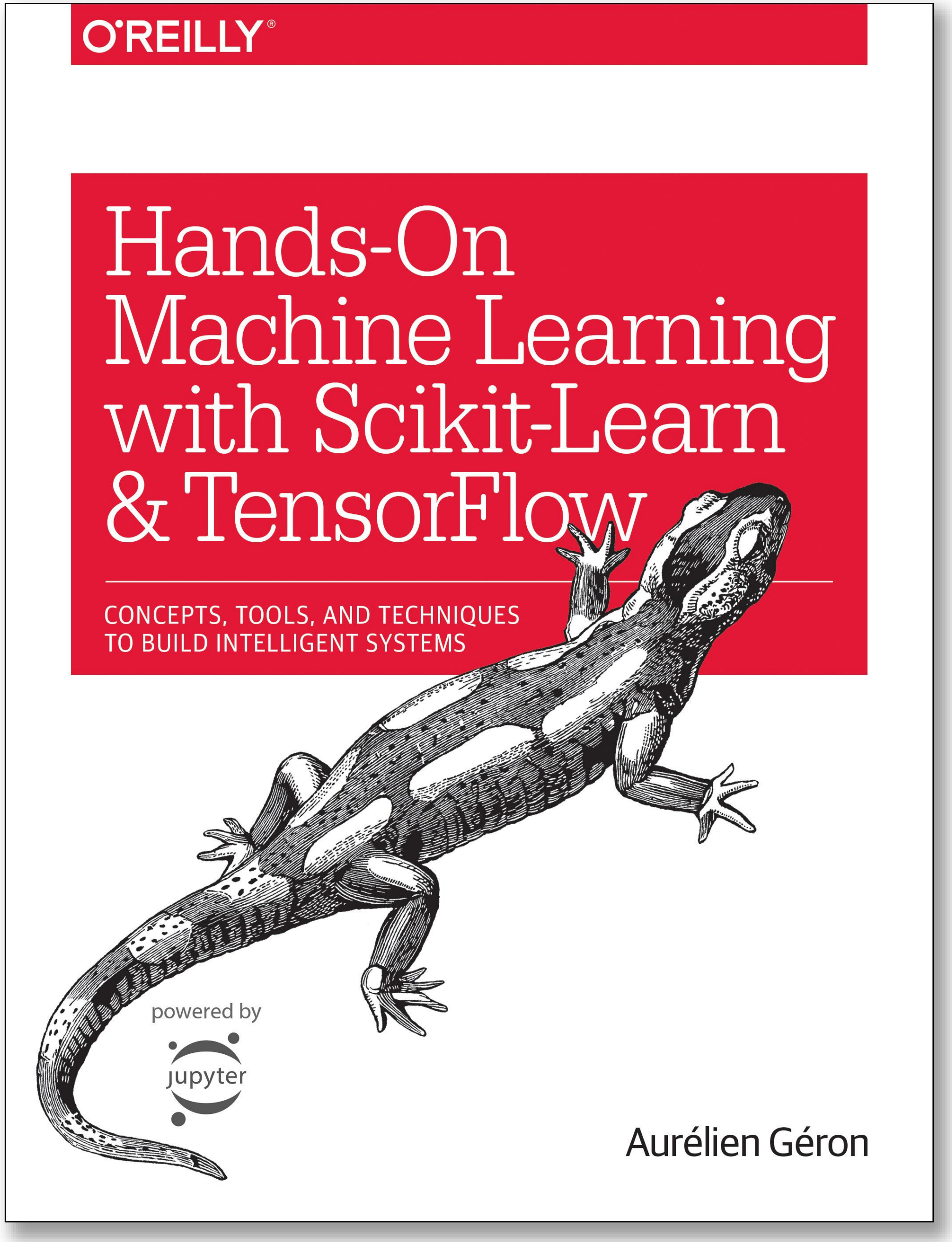
8760 rows of data, around 10 relavent columns

62 million total rows

Machine Learning!

- “the science and art of programming computers so they can learn from data.” (Geron)

- Machine learning uses data to “learn” and predict outcomes rather than using explicit algorithms or rules, and works well for problems that have no known algorithm based solution, but have lots of available data to learn from.



O'REILLY®

Hands-On Machine Learning with Scikit-Learn & TensorFlow

CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS

powered by jupyter

Aurélien Géron

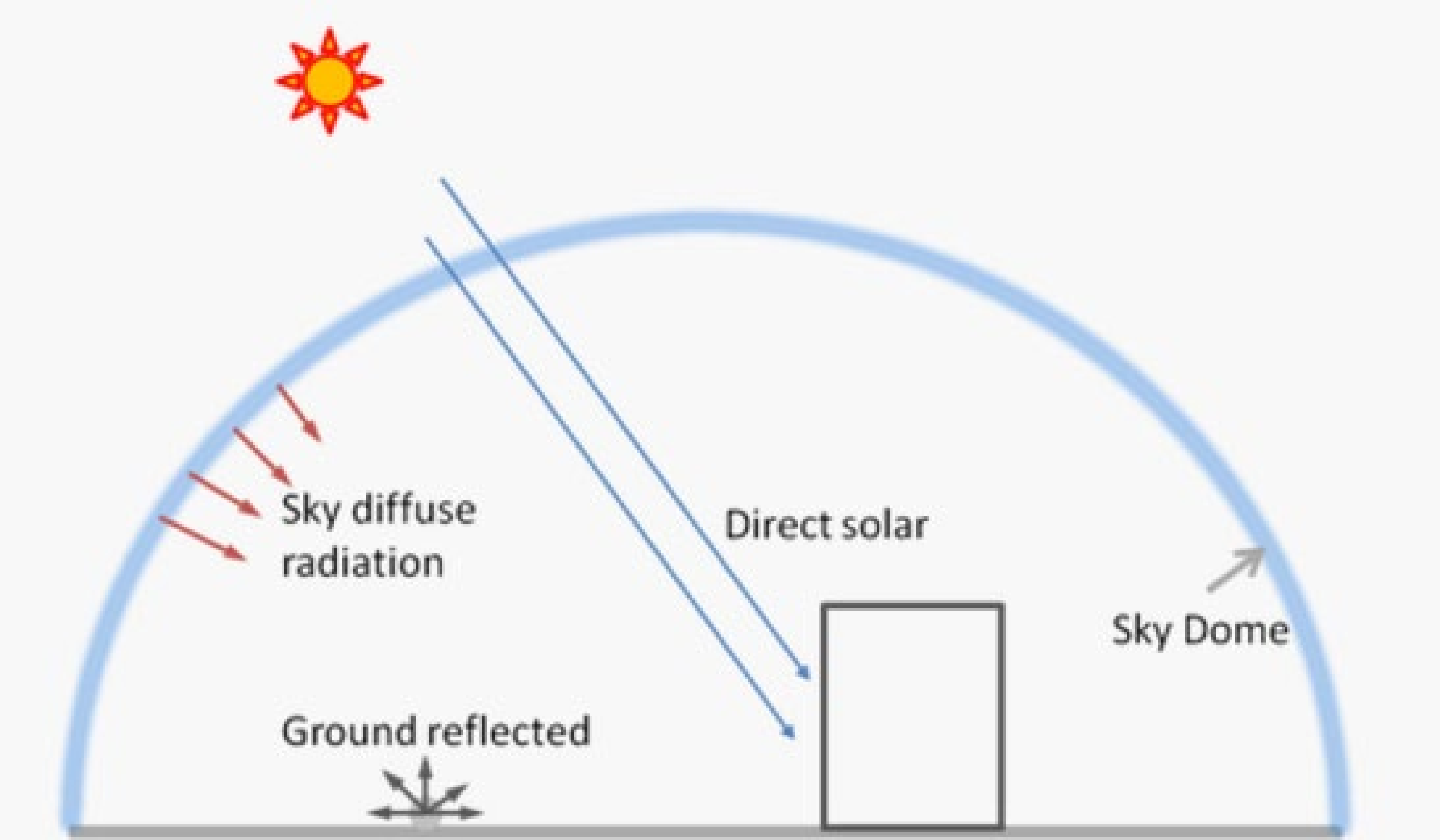
When to use machine learning:

1. Tasks involve a function that maps well-defined inputs to well-defined outputs
2. Large (digital) datasets exist or can be created containing input-output pairs
3. Tasks provide clear feedback with clearly definable goals and metrics
4. No long chains of logic or reasoning that depend on diverse background knowledge or common sense
5. Tasks do not require detailed explanations for how the decision was made
6. Tasks have a tolerance for error and no need for provably correct or optimal solutions
7. The phenomenon or function being learned should not change rapidly over time
8. No specialized dexterity, physical skills, or mobility is required

From “What Can Machine Learning Do? Workforce Implications” Erik Brynjolfsson and Tom Mitchell, Science Magazine, Dec 22, 2017

Hypothesis

- Data normally used:
  - Month
  - Day
  - Hour
  - Latitude
  - Longitude
  - Direct Illuminance (DIR)
  - Diffuse Illuminance (DIF)
- Data we also have:
  - Global Illuminance (GLOB)



Direct, sky diffuse and ground reflected radiation

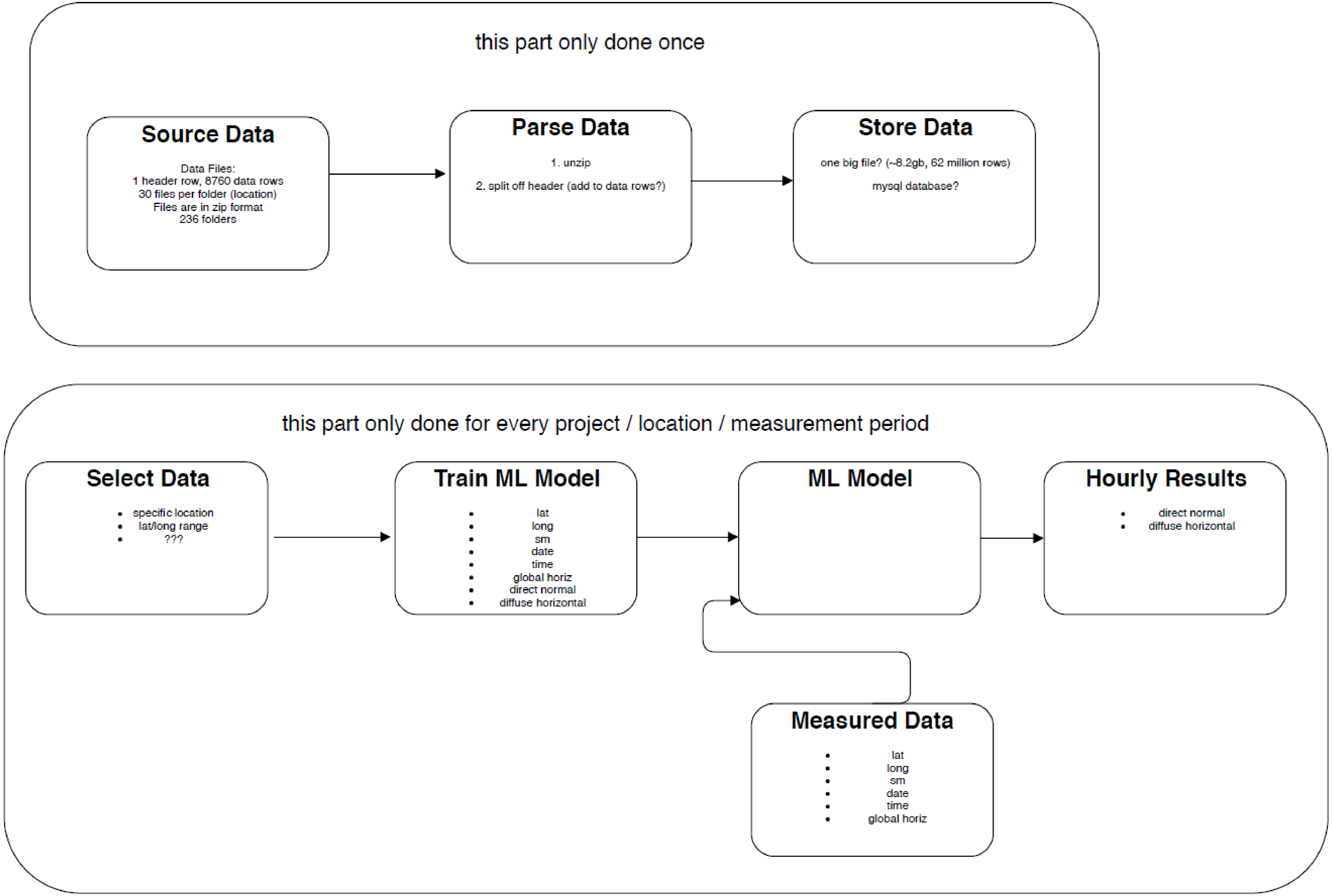
$$\mathrm{DIR} + \mathrm{DIF} = \mathrm{GLOB}$$

Hypothesis

- Data normally used:
  - Month
  - Day
  - Hour
  - Latitude
  - Longitude
  - Direct Illuminance (DIR)
  - Diffuse Illuminance (DIF)
- Data we also have:
  - Global Illuminance (GLOB)

- If we have:
  - Month
  - Day
  - Hour
  - Latitude
  - Longitude
  - Global Illuminance (GLOB)
- Can we predict:
  - Direct Illuminance (DIR)
  - Diffuse Illuminance (DIF)

Process



this part only done once

Source Data

Data Files:
1 header row, 8760 data rows
30 files per folder (location)
Files are in zip format
236 folders

Parse Data

1. unzip
2. split off header (add to data rows?)

Store Data

one big file? (~8.2gb, 62 million rows)

mysql database?

this part only done for every project / location / measurement period

Select Data

- specific location
- lat/long range
- ???

Train ML Model

- lat
- long
- sm
- date
- time
- global horiz
- direct normal
- diffuse horizontal

ML Model

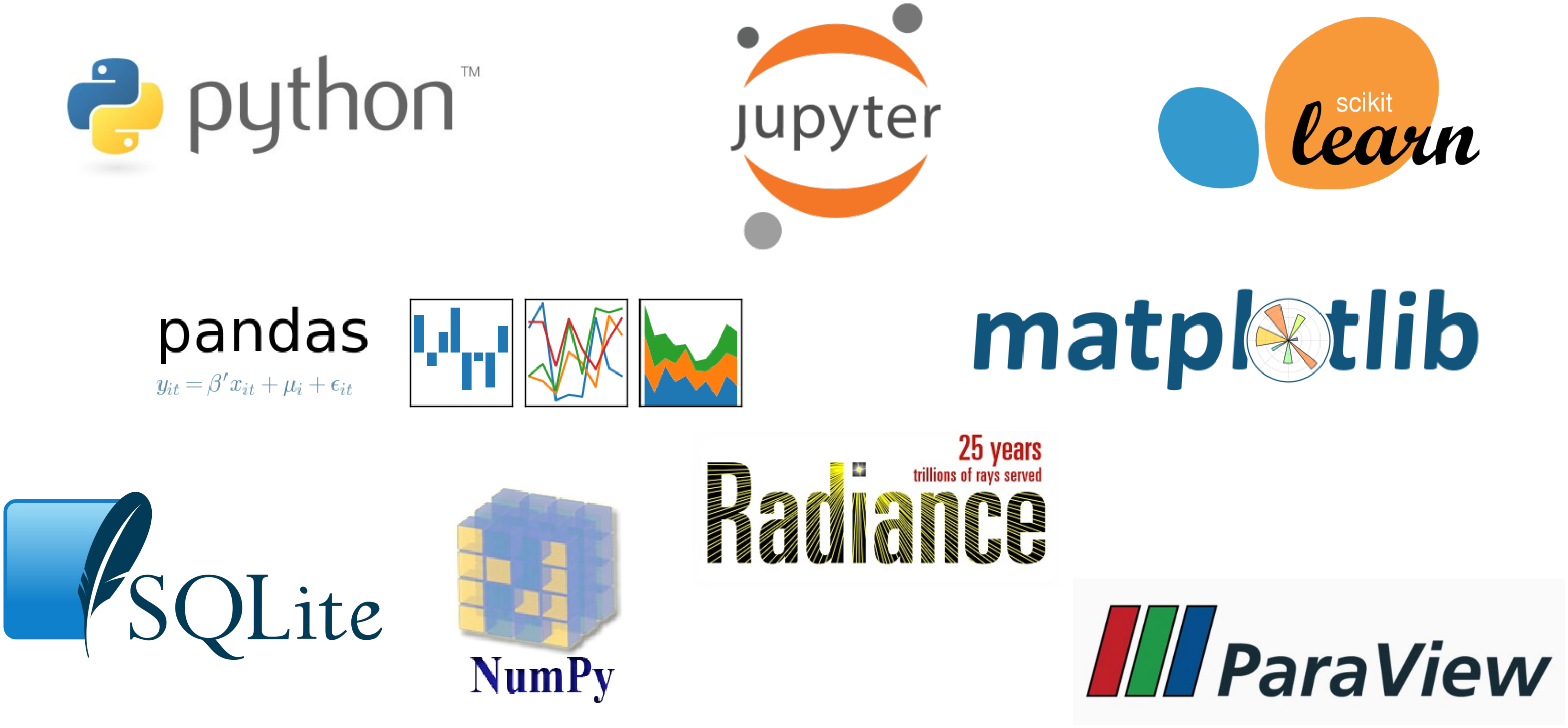
Hourly Results

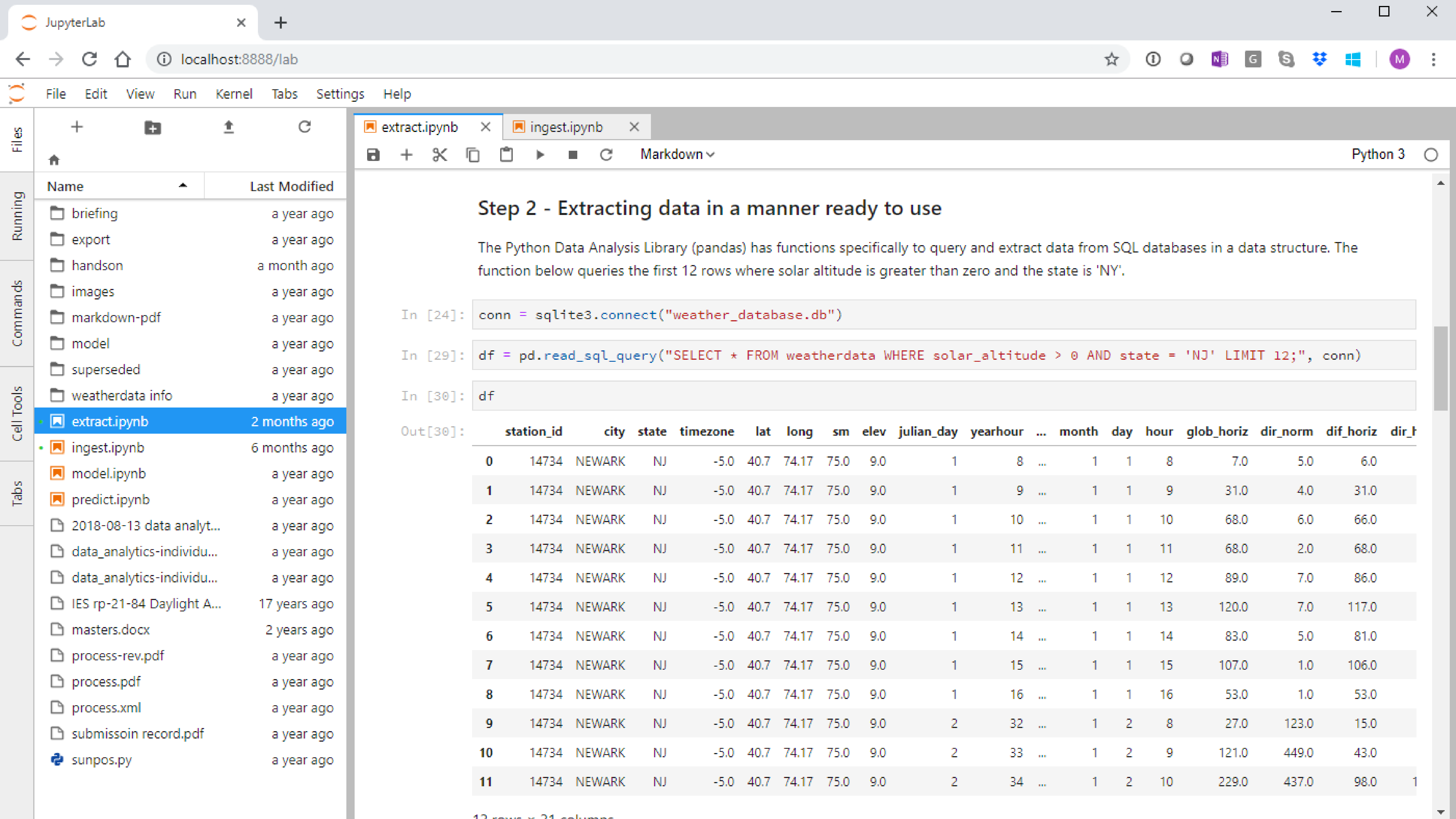
- direct normal
- diffuse horizontal

Measured Data

- lat
- long
- sm
- date
- time
- global horiz

Tools


python™
jupyter
scikit learn
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
matplotlib
25 years
trillions of rays served
Radiance
SQLite
NumPy
ParaView



Name	Last Modified
briefing	a year ago
export	a year ago
handson	a month ago
images	a year ago
markdown-pdf	a year ago
model	a year ago
superseded	a year ago
weatherdata info	a year ago
extract.ipynb	2 months ago
ingest.ipynb	6 months ago
model.ipynb	a year ago
predict.ipynb	a year ago
2018-08-13 data analyt...	a year ago
data_analytics-individu...	a year ago
data_analytics-individu...	a year ago
IES rp-21-84 Daylight A...	17 years ago
masters.docx	2 years ago
process-rev.pdf	a year ago
process.pdf	a year ago
process.xml	a year ago
submissoin record.pdf	a year ago
sunpos.py	a year ago

Step 2 - Extracting data in a manner ready to use

The Python Data Analysis Library (pandas) has functions specifically to query and extract data from SQL databases in a data structure. The function below queries the first 12 rows where solar altitude is greater than zero and the state is 'NY'.

```
In [24]: conn = sqlite3.connect("weather_database.db")

In [29]: df = pd.read_sql_query("SELECT * FROM weatherdata WHERE solar_altitude > 0 AND state = 'NJ' LIMIT 12;", conn)

In [30]: df
```

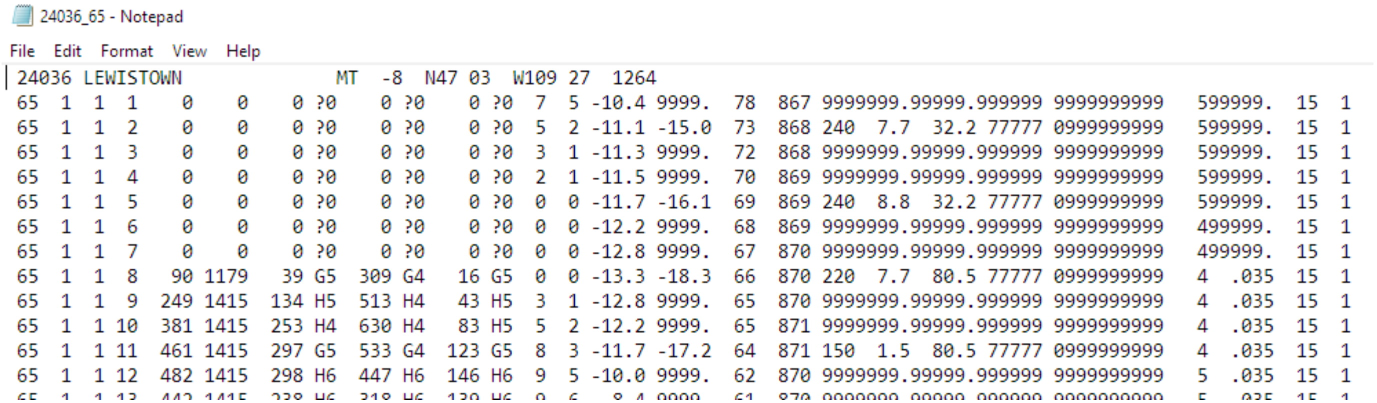
Out[30]:

	station_id	city	state	timezone	lat	long	sm	elev	julian_day	yearhour	...	month	day	hour	glob_horiz	dir_norm	dif_horiz	dir_h
0	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	8	...	1	1	8	7.0	5.0	6.0	
1	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	9	...	1	1	9	31.0	4.0	31.0	
2	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	10	...	1	1	10	68.0	6.0	66.0	
3	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	11	...	1	1	11	68.0	2.0	68.0	
4	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	12	...	1	1	12	89.0	7.0	86.0	
5	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	13	...	1	1	13	120.0	7.0	117.0	
6	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	14	...	1	1	14	83.0	5.0	81.0	
7	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	15	...	1	1	15	107.0	1.0	106.0	
8	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	16	...	1	1	16	53.0	1.0	53.0	
9	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	2	32	...	1	2	8	27.0	123.0	15.0	
10	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	2	33	...	1	2	9	121.0	449.0	43.0	
11	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	2	34	...	1	2	10	229.0	437.0	98.0	1

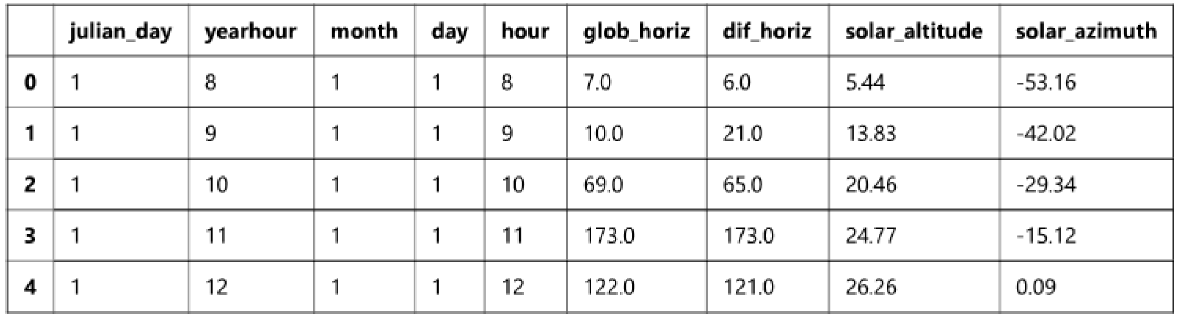
12 rows × 21 columns

Part 1 – Get data into useable format

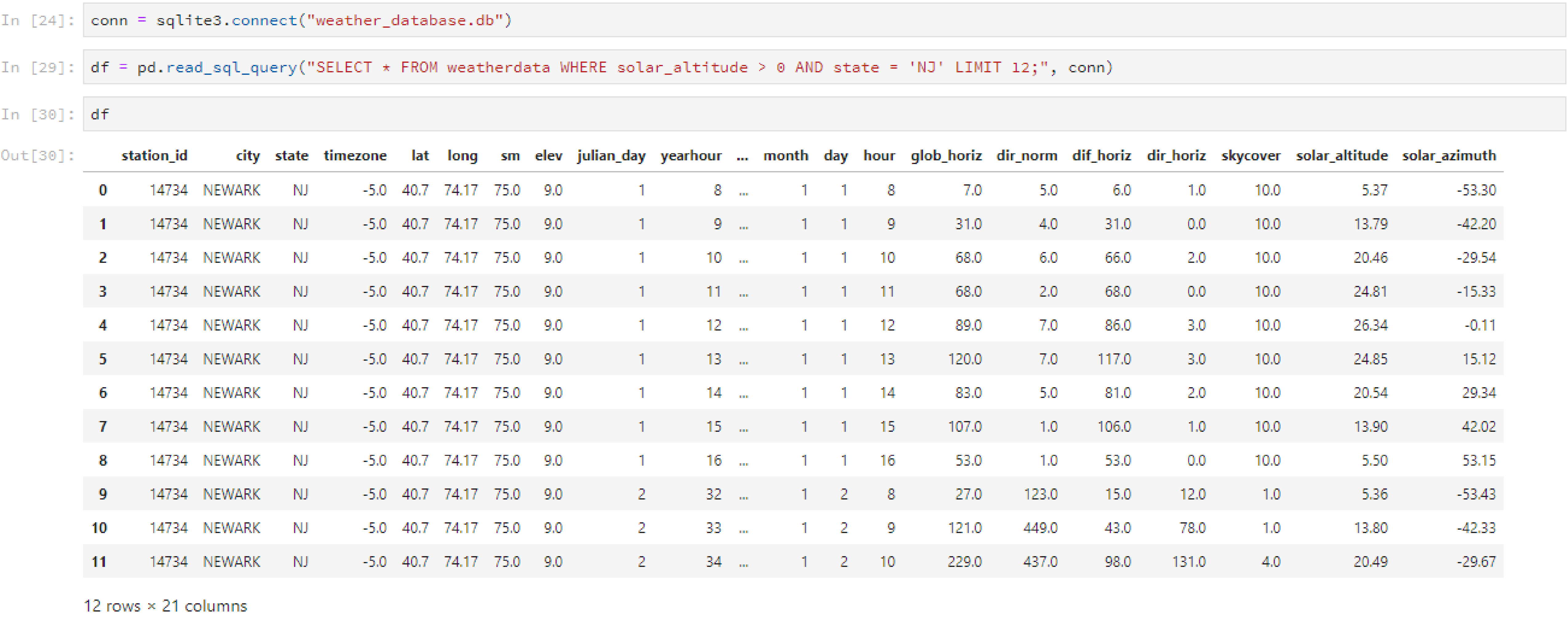
- Unzip data files
  - Each group of 30 files within its own folder
- Add in header data to each line
  - Add lat/long info based on day/time/location
- Read into database



```
24036 LEWISTOWN               MT   -8  N47 03  W109 27  1264
65  1  1  1    0    0    0 ?0    0 ?0    0 ?0  7  5 -10.4 9999.  78  867 9999999.99999.999999 9999999999   599999.  15  1
65  1  1  2    0    0    0 ?0    0 ?0    0 ?0  5  2 -11.1 -15.0  73  868 240  7.7  32.2 77777 0999999999   599999.  15  1
65  1  1  3    0    0    0 ?0    0 ?0    0 ?0  3  1 -11.3 9999.  72  868 9999999.99999.999999 9999999999   599999.  15  1
65  1  1  4    0    0    0 ?0    0 ?0    0 ?0  2  1 -11.5 9999.  70  869 9999999.99999.999999 9999999999   599999.  15  1
65  1  1  5    0    0    0 ?0    0 ?0    0 ?0  0  0 -11.7 -16.1  69  869 240  8.8  32.2 77777 0999999999   599999.  15  1
65  1  1  6    0    0    0 ?0    0 ?0    0 ?0  0  0 -12.2 9999.  68  869 9999999.99999.999999 9999999999   499999.  15  1
65  1  1  7    0    0    0 ?0    0 ?0    0 ?0  0  0 -12.8 9999.  67  870 9999999.99999.999999 9999999999   499999.  15  1
65  1  1  8   90 1179   39 G5  309 G4   16 G5  0  0 -13.3 -18.3  66  870 220  7.7  80.5 77777 0999999999   4  .035  15  1
65  1  1  9  249 1415  134 H5  513 H4   43 H5  3  1 -12.8 9999.  65  870 9999999.99999.999999 9999999999   4  .035  15  1
65  1  1 10  381 1415  253 H4  630 H4   83 H5  5  2 -12.2 9999.  65  871 9999999.99999.999999 9999999999   4  .035  15  1
65  1  1 11  461 1415  297 G5  533 G4  123 G5  8  3 -11.7 -17.2  64  871 150  1.5  80.5 77777 0999999999   4  .035  15  1
65  1  1 12  482 1415  298 H6  447 H6  146 H6  9  5 -10.0 9999.  62  870 9999999.99999.999999 9999999999   5  .035  15  1
```



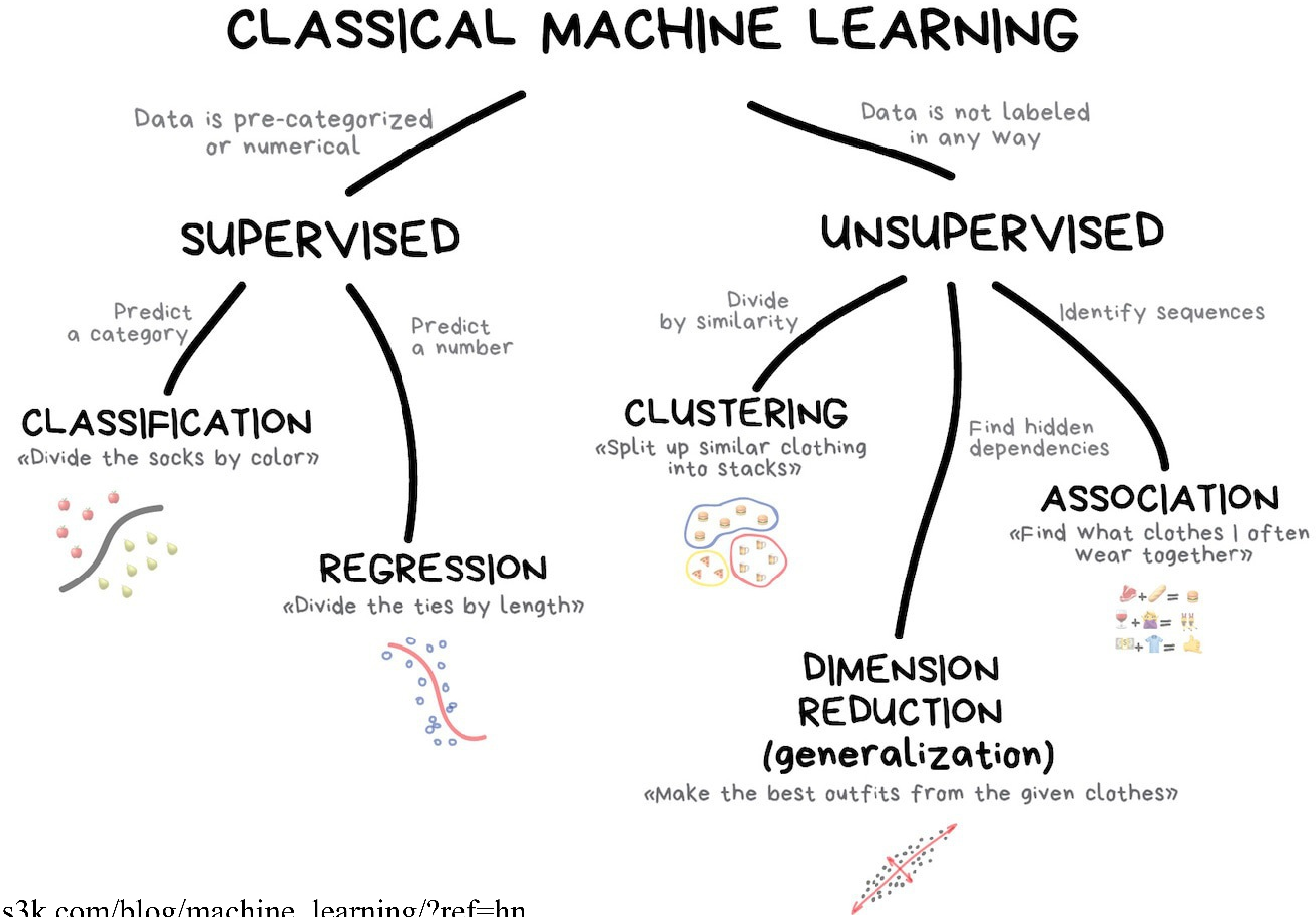
	julian_day	yearhour	month	day	hour	glob_horiz	dif_horiz	solar_altitude	solar_azimuth
0	1	8	1	1	8	7.0	6.0	5.44	-53.16
1	1	9	1	1	9	10.0	21.0	13.83	-42.02
2	1	10	1	1	10	69.0	65.0	20.46	-29.34
3	1	11	1	1	11	173.0	173.0	24.77	-15.12
4	1	12	1	1	12	122.0	121.0	26.26	0.09



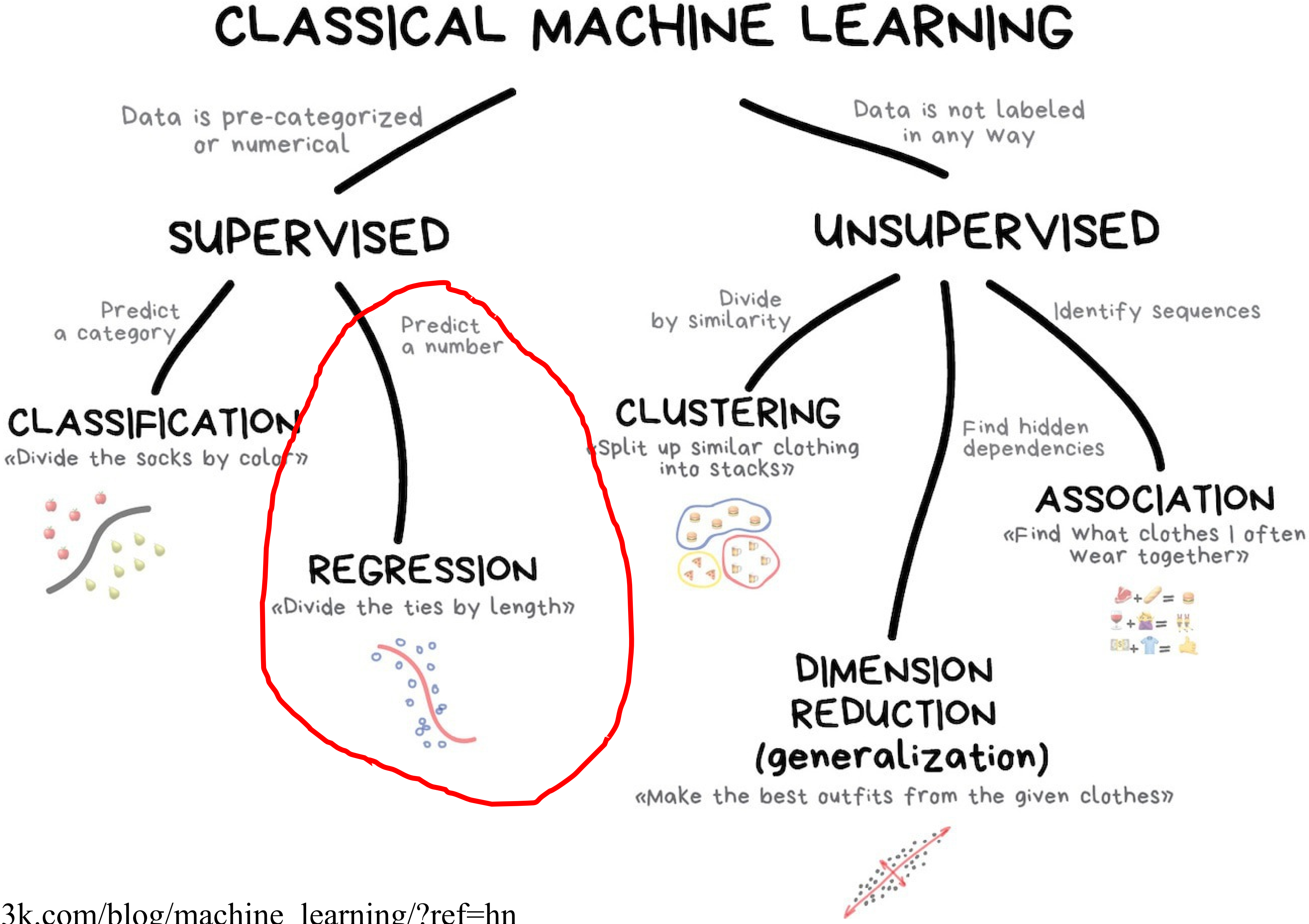
```
In [24]: conn = sqlite3.connect("weather_database.db")

In [29]: df = pd.read_sql_query("SELECT * FROM weatherdata WHERE solar_altitude > 0 AND state = 'NJ' LIMIT 12;", conn)

In [30]: df

Out[30]:
```

	station_id	city	state	timezone	lat	long	sm	elev	julian_day	yearhour	...	month	day	hour	glob_horiz	dir_norm	dif_horiz	dir_horiz	skycover	solar_altitude	solar_azimuth
0	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	8	...	1	1	8	7.0	5.0	6.0	1.0	10.0	5.37	-53.30
1	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	9	...	1	1	9	31.0	4.0	31.0	0.0	10.0	13.79	-42.20
2	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	10	...	1	1	10	68.0	6.0	66.0	2.0	10.0	20.46	-29.54
3	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	11	...	1	1	11	68.0	2.0	68.0	0.0	10.0	24.81	-15.33
4	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	12	...	1	1	12	89.0	7.0	86.0	3.0	10.0	26.34	-0.11
5	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	13	...	1	1	13	120.0	7.0	117.0	3.0	10.0	24.85	15.12
6	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	14	...	1	1	14	83.0	5.0	81.0	2.0	10.0	20.54	29.34
7	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	15	...	1	1	15	107.0	1.0	106.0	1.0	10.0	13.90	42.02
8	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	1	16	...	1	1	16	53.0	1.0	53.0	0.0	10.0	5.50	53.15
9	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	2	32	...	1	2	8	27.0	123.0	15.0	12.0	1.0	5.36	-53.43
10	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	2	33	...	1	2	9	121.0	449.0	43.0	78.0	1.0	13.80	-42.33
11	14734	NEWARK	NJ	-5.0	40.7	74.17	75.0	9.0	2	34	...	1	2	10	229.0	437.0	98.0	131.0	4.0	20.49	-29.67

12 rows × 21 columns

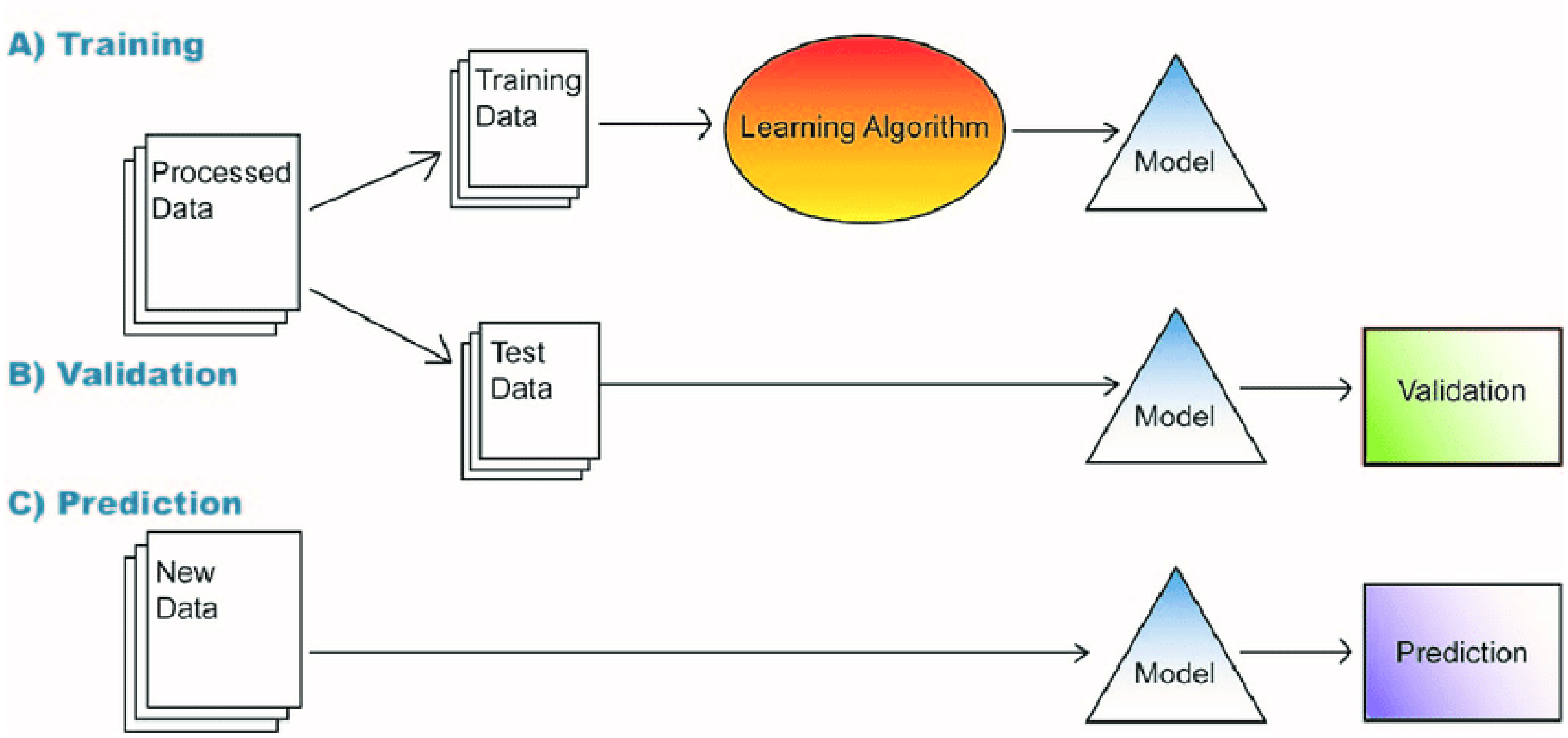

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
Find hidden dependencies
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

From https://vas3k.com/blog/machine_learning/?ref=hn

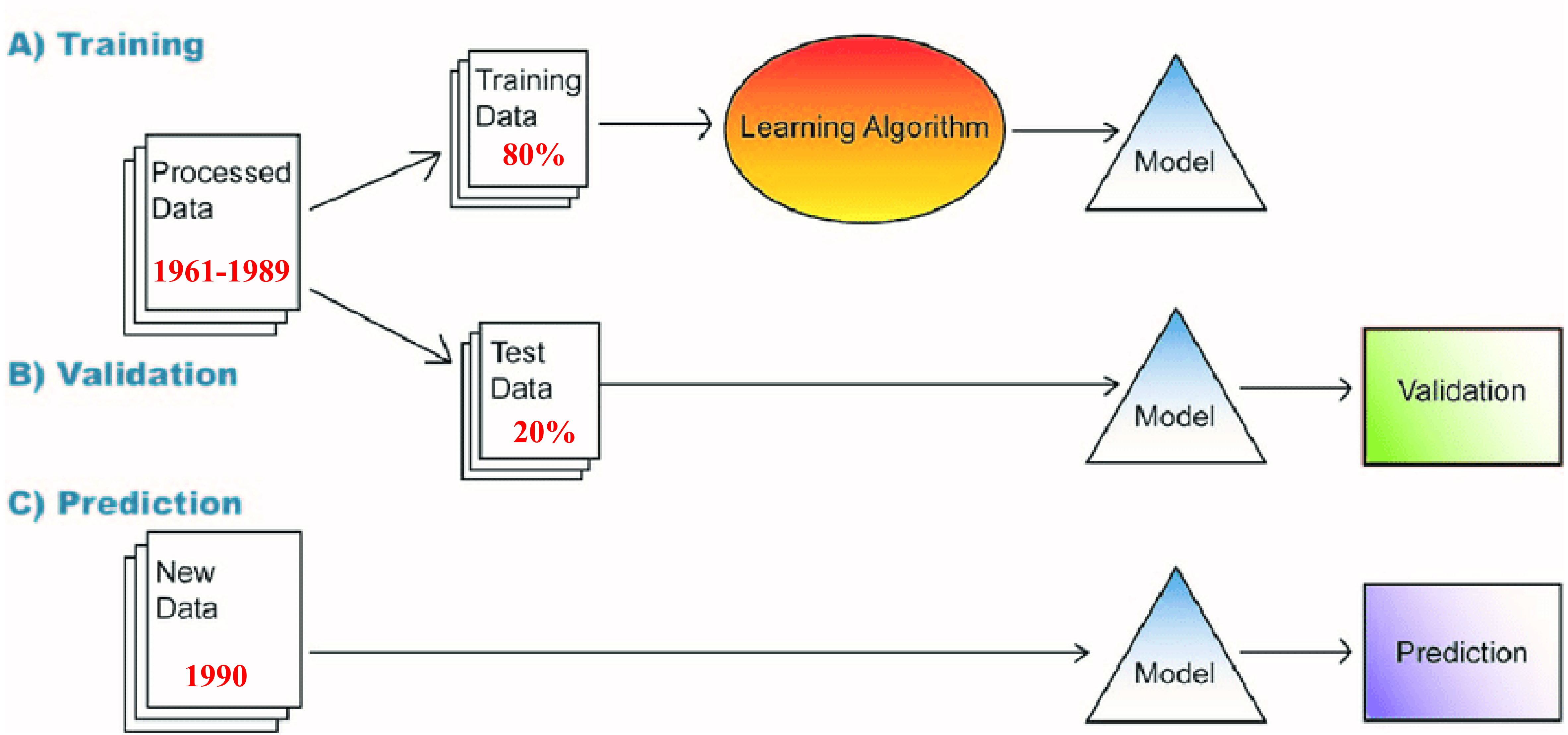

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
Find hidden dependencies
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

From https://vas3k.com/blog/machine_learning/?ref=hn

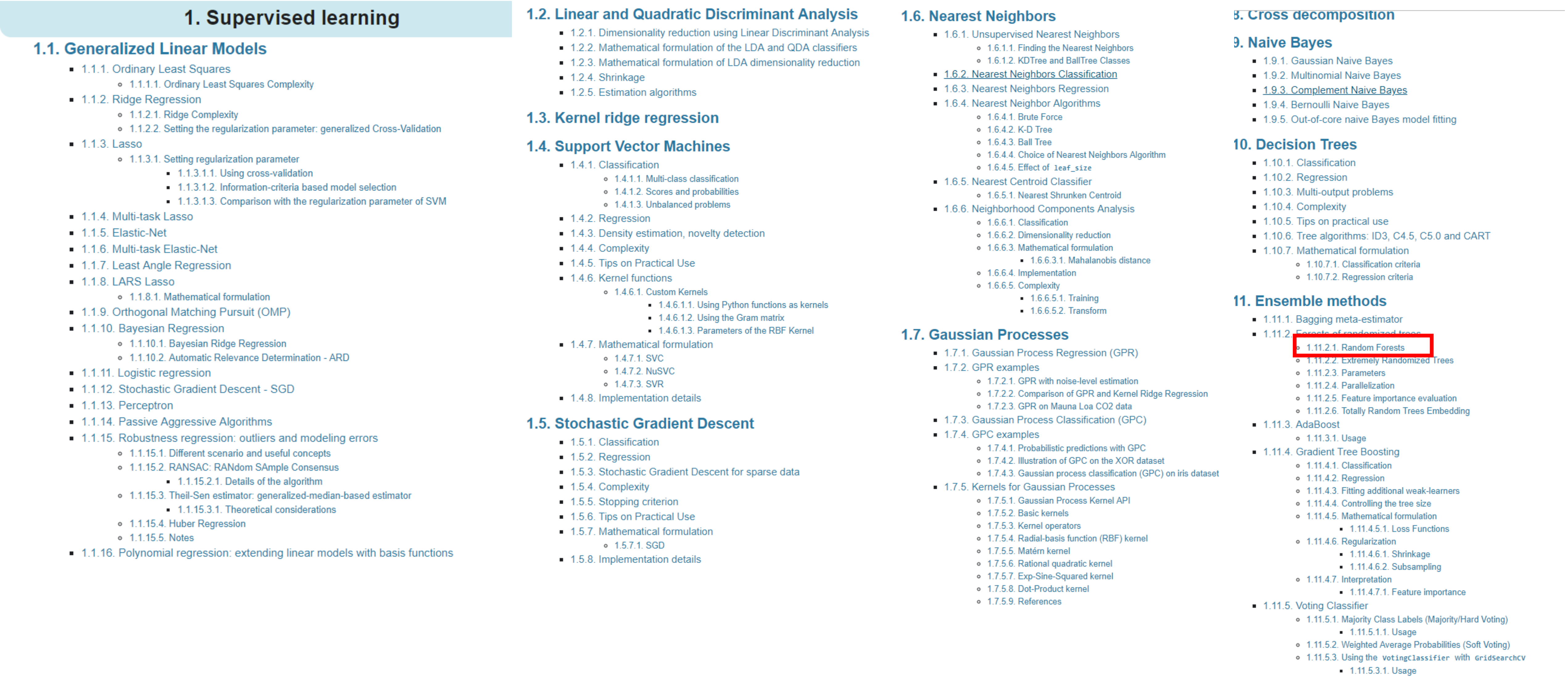
Part 2 – Create machine learning model



Part 2 – Create machine learning model



Part 2 – Select train, validate model



1. Supervised learning

1.1. Generalized Linear Models

- 1.1.1. Ordinary Least Squares
  - 1.1.1.1. Ordinary Least Squares Complexity
- 1.1.2. Ridge Regression
  - 1.1.2.1. Ridge Complexity
  - 1.1.2.2. Setting the regularization parameter: generalized Cross-Validation
- 1.1.3. Lasso
  - 1.1.3.1. Setting regularization parameter
    - 1.1.3.1.1. Using cross-validation
    - 1.1.3.1.2. Information-criteria based model selection
    - 1.1.3.1.3. Comparison with the regularization parameter of SVM
- 1.1.4. Multi-task Lasso
- 1.1.5. Elastic-Net
- 1.1.6. Multi-task Elastic-Net
- 1.1.7. Least Angle Regression
- 1.1.8. LARS Lasso
  - 1.1.8.1. Mathematical formulation
- 1.1.9. Orthogonal Matching Pursuit (OMP)
- 1.1.10. Bayesian Regression
  - 1.1.10.1. Bayesian Ridge Regression
  - 1.1.10.2. Automatic Relevance Determination - ARD
- 1.1.11. Logistic regression
- 1.1.12. Stochastic Gradient Descent - SGD
- 1.1.13. Perceptron
- 1.1.14. Passive Aggressive Algorithms
- 1.1.15. Robustness regression: outliers and modeling errors
  - 1.1.15.1. Different scenario and useful concepts
  - 1.1.15.2. RANSAC: RANdom SAmple Consensus
    - 1.1.15.2.1. Details of the algorithm
  - 1.1.15.3. Theil-Sen estimator: generalized-median-based estimator
    - 1.1.15.3.1. Theoretical considerations
  - 1.1.15.4. Huber Regression
  - 1.1.15.5. Notes
- 1.1.16. Polynomial regression: extending linear models with basis functions

1.2. Linear and Quadratic Discriminant Analysis

- 1.2.1. Dimensionality reduction using Linear Discriminant Analysis
- 1.2.2. Mathematical formulation of the LDA and QDA classifiers
- 1.2.3. Mathematical formulation of LDA dimensionality reduction
- 1.2.4. Shrinkage
- 1.2.5. Estimation algorithms

1.3. Kernel ridge regression

1.4. Support Vector Machines

- 1.4.1. Classification
  - 1.4.1.1. Multi-class classification
  - 1.4.1.2. Scores and probabilities
  - 1.4.1.3. Unbalanced problems
- 1.4.2. Regression
- 1.4.3. Density estimation, novelty detection
- 1.4.4. Complexity
- 1.4.5. Tips on Practical Use
- 1.4.6. Kernel functions
  - 1.4.6.1. Custom Kernels
    - 1.4.6.1.1. Using Python functions as kernels
    - 1.4.6.1.2. Using the Gram matrix
    - 1.4.6.1.3. Parameters of the RBF Kernel
- 1.4.7. Mathematical formulation
  - 1.4.7.1. SVC
  - 1.4.7.2. NuSVC
  - 1.4.7.3. SVR
- 1.4.8. Implementation details

1.5. Stochastic Gradient Descent

- 1.5.1. Classification
- 1.5.2. Regression
- 1.5.3. Stochastic Gradient Descent for sparse data
- 1.5.4. Complexity
- 1.5.5. Stopping criterion
- 1.5.6. Tips on Practical Use
- 1.5.7. Mathematical formulation
  - 1.5.7.1. SGD
- 1.5.8. Implementation details

1.6. Nearest Neighbors

- 1.6.1. Unsupervised Nearest Neighbors
  - 1.6.1.1. Finding the Nearest Neighbors
  - 1.6.1.2. KDTree and BallTree Classes
- 1.6.2. Nearest Neighbors Classification
- 1.6.3. Nearest Neighbors Regression
- 1.6.4. Nearest Neighbor Algorithms
  - 1.6.4.1. Brute Force
  - 1.6.4.2. K-D Tree
  - 1.6.4.3. Ball Tree
  - 1.6.4.4. Choice of Nearest Neighbors Algorithm
  - 1.6.4.5. Effect of `leaf_size`
- 1.6.5. Nearest Centroid Classifier
  - 1.6.5.1. Nearest Shrunken Centroid
- 1.6.6. Neighborhood Components Analysis
  - 1.6.6.1. Classification
  - 1.6.6.2. Dimensionality reduction
  - 1.6.6.3. Mathematical formulation
    - 1.6.6.3.1. Mahalanobis distance
  - 1.6.6.4. Implementation
  - 1.6.6.5. Complexity
    - 1.6.6.5.1. Training
    - 1.6.6.5.2. Transform

1.7. Gaussian Processes

- 1.7.1. Gaussian Process Regression (GPR)
- 1.7.2. GPR examples
  - 1.7.2.1. GPR with noise-level estimation
  - 1.7.2.2. Comparison of GPR and Kernel Ridge Regression
  - 1.7.2.3. GPR on Mauna Loa CO2 data
- 1.7.3. Gaussian Process Classification (GPC)
- 1.7.4. GPC examples
  - 1.7.4.1. Probabilistic predictions with GPC
  - 1.7.4.2. Illustration of GPC on the XOR dataset
  - 1.7.4.3. Gaussian process classification (GPC) on iris dataset
- 1.7.5. Kernels for Gaussian Processes
  - 1.7.5.1. Gaussian Process Kernel API
  - 1.7.5.2. Basic kernels
  - 1.7.5.3. Kernel operators
  - 1.7.5.4. Radial-basis function (RBF) kernel
  - 1.7.5.5. Matérn kernel
  - 1.7.5.6. Rational quadratic kernel
  - 1.7.5.7. Exp-Sine-Squared kernel
  - 1.7.5.8. Dot-Product kernel
  - 1.7.5.9. References

8. Cross decomposition

9. Naive Bayes

- 1.9.1. Gaussian Naive Bayes
- 1.9.2. Multinomial Naive Bayes
- 1.9.3. Complement Naive Bayes
- 1.9.4. Bernoulli Naive Bayes
- 1.9.5. Out-of-core naive Bayes model fitting

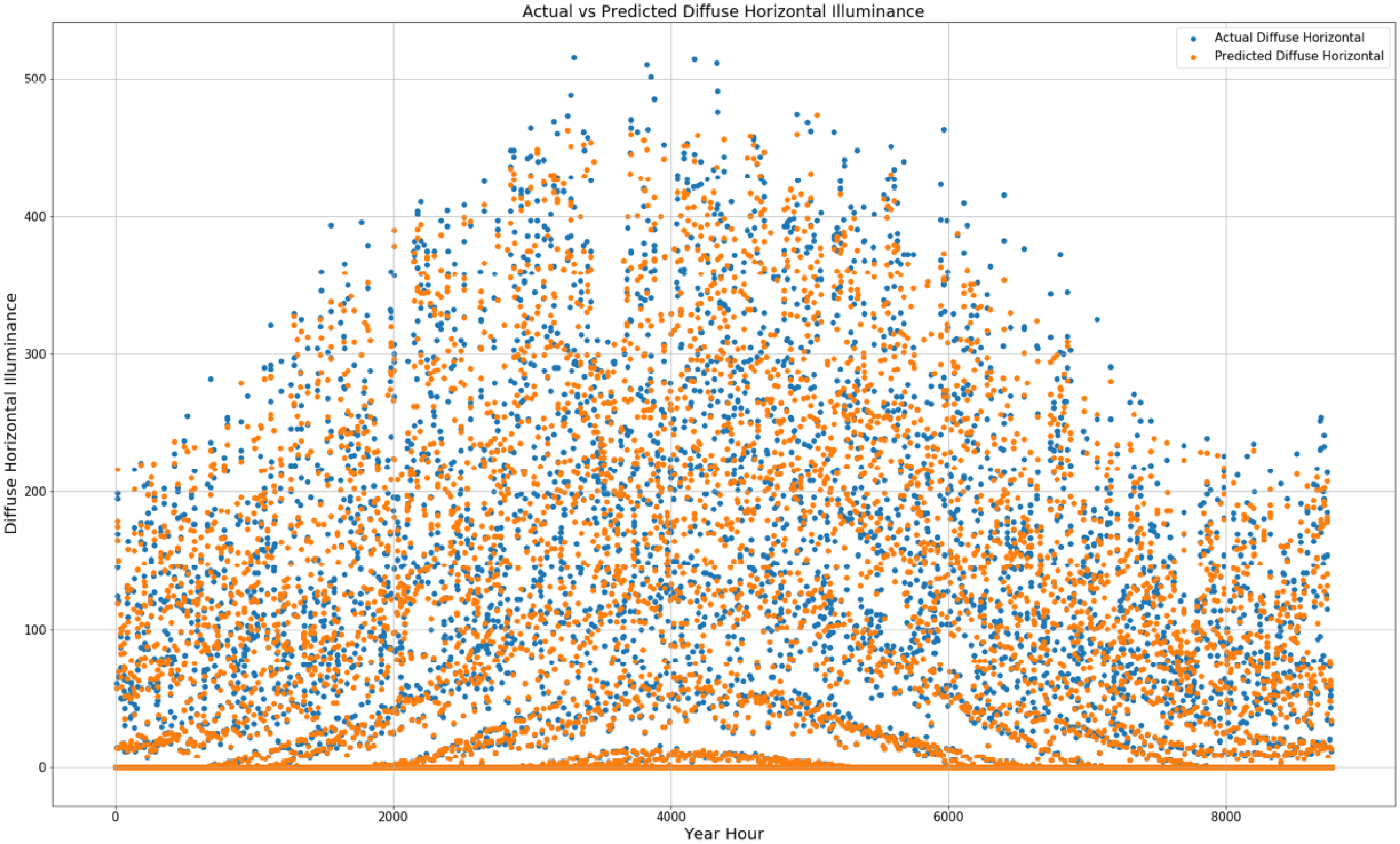
10. Decision Trees

- 1.10.1. Classification
- 1.10.2. Regression
- 1.10.3. Multi-output problems
- 1.10.4. Complexity
- 1.10.5. Tips on practical use
- 1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART
- 1.10.7. Mathematical formulation
  - 1.10.7.1. Classification criteria
  - 1.10.7.2. Regression criteria

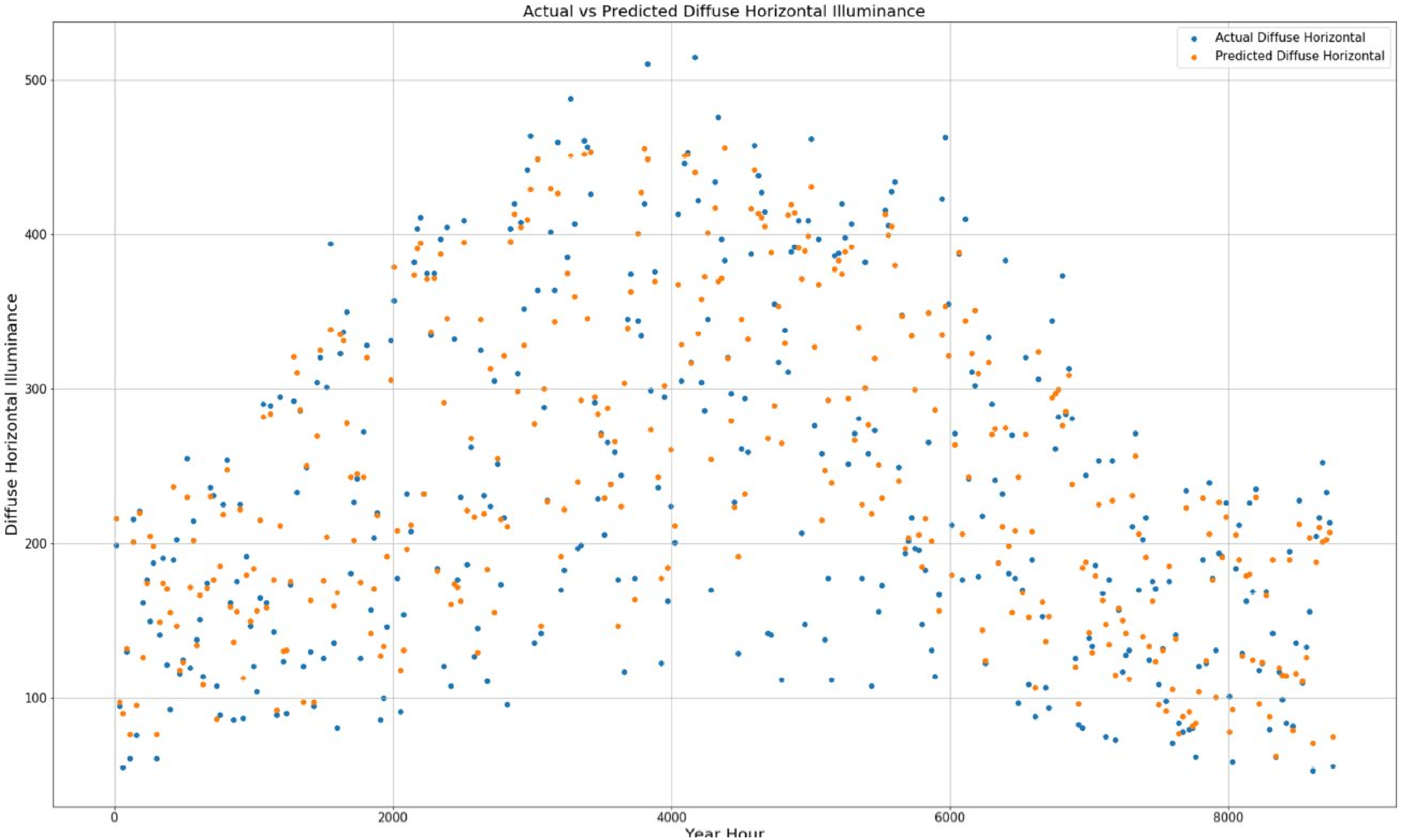
11. Ensemble methods

- 1.11.1. Bagging meta-estimator
- 1.11.2. Forests of randomized trees
  - 1.11.2.1. Random Forests
  - 1.11.2.2. Extremely Randomized Trees
  - 1.11.2.3. Parameters
  - 1.11.2.4. Parallelization
  - 1.11.2.5. Feature importance evaluation
  - 1.11.2.6. Totally Random Trees Embedding
- 1.11.3. AdaBoost
  - 1.11.3.1. Usage
- 1.11.4. Gradient Tree Boosting
  - 1.11.4.1. Classification
  - 1.11.4.2. Regression
  - 1.11.4.3. Fitting additional weak-learners
  - 1.11.4.4. Controlling the tree size
  - 1.11.4.5. Mathematical formulation
    - 1.11.4.5.1. Loss Functions
  - 1.11.4.6. Regularization
    - 1.11.4.6.1. Shrinkage
    - 1.11.4.6.2. Subsampling
  - 1.11.4.7. Interpretation
    - 1.11.4.7.1. Feature importance
- 1.11.5. Voting Classifier
  - 1.11.5.1. Majority Class Labels (Majority/Hard Voting)
    - 1.11.5.1.1. Usage
  - 1.11.5.2. Weighted Average Probabilities (Soft Voting)
  - 1.11.5.3. Using the `VotingClassifier` with `GridSearchCV`
    - 1.11.5.3.1. Usage

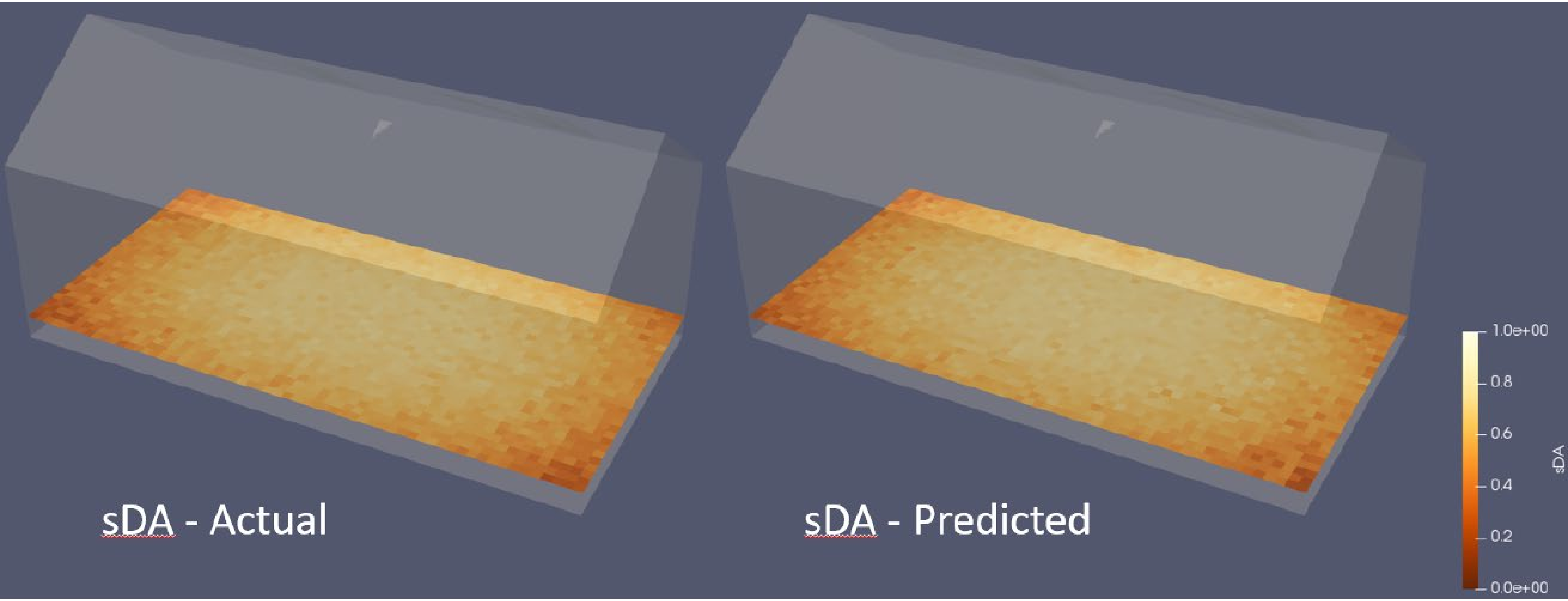
Results – Annual Hourly



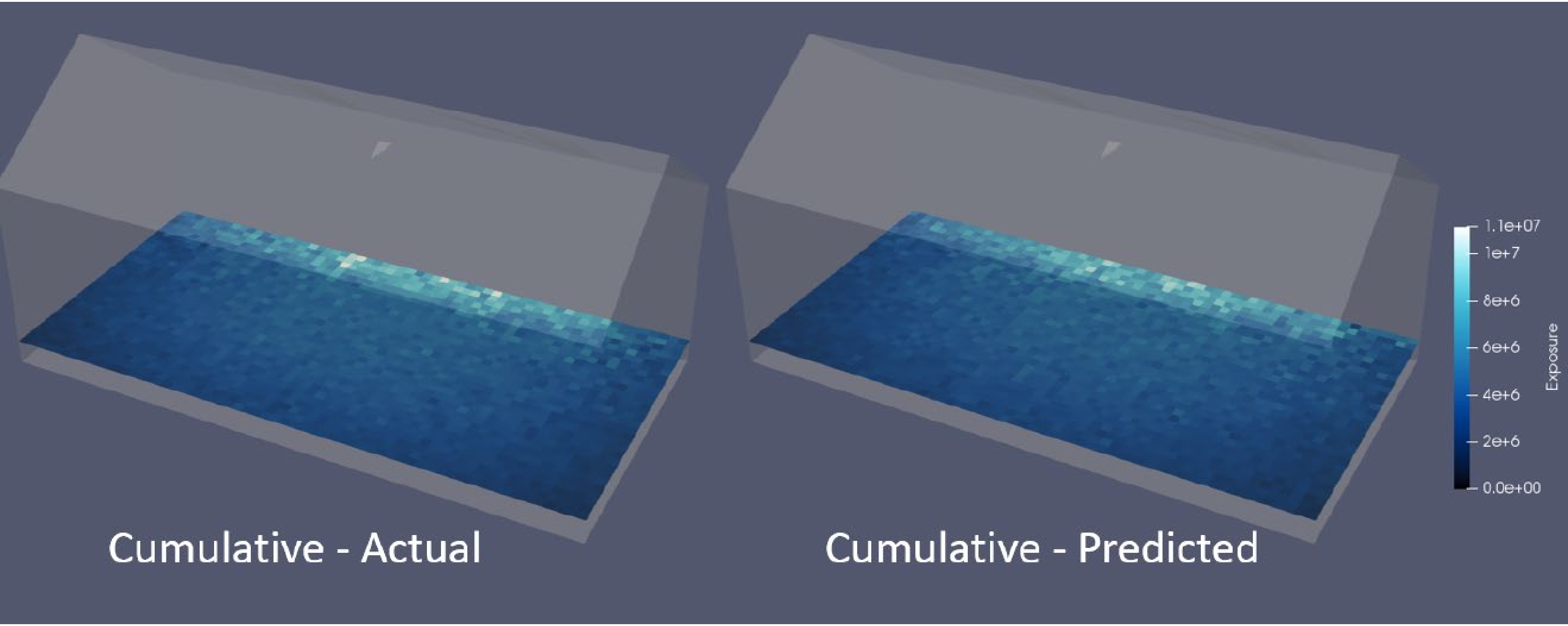
Results – Annual Hourly – 12p.m. only



Analysis with Predicted Data - sDA



Analysis with Predicted Data - Cumulative



But wait!

Aren’t there already ways to do this?

Existing Models

- Erbs et al., 1982 (ER)
- Orgill and Hollands, 1977 (OH)
- Reindl et al., 1990 (RE)
- Lam and Li, 1996 (LL)
- Skarteveit and Olseth, 1987 (SO)
- Louche et al., 1991 (LO)
- Maxwell, 1987 (MA)
- Vignola and McDaniels, 1984 (VM)

Sokol Dervishi and Ardeshir Mahdavi. Computing diffuse fraction of global horizontal solar radiation: A model comparison. Solar Energy, 2012

Error Metrics

- Mean Bias Deviation (MBD)
- Relative Error (RE)
- Root Mean Squared Deviation (RMSD, RMSE)

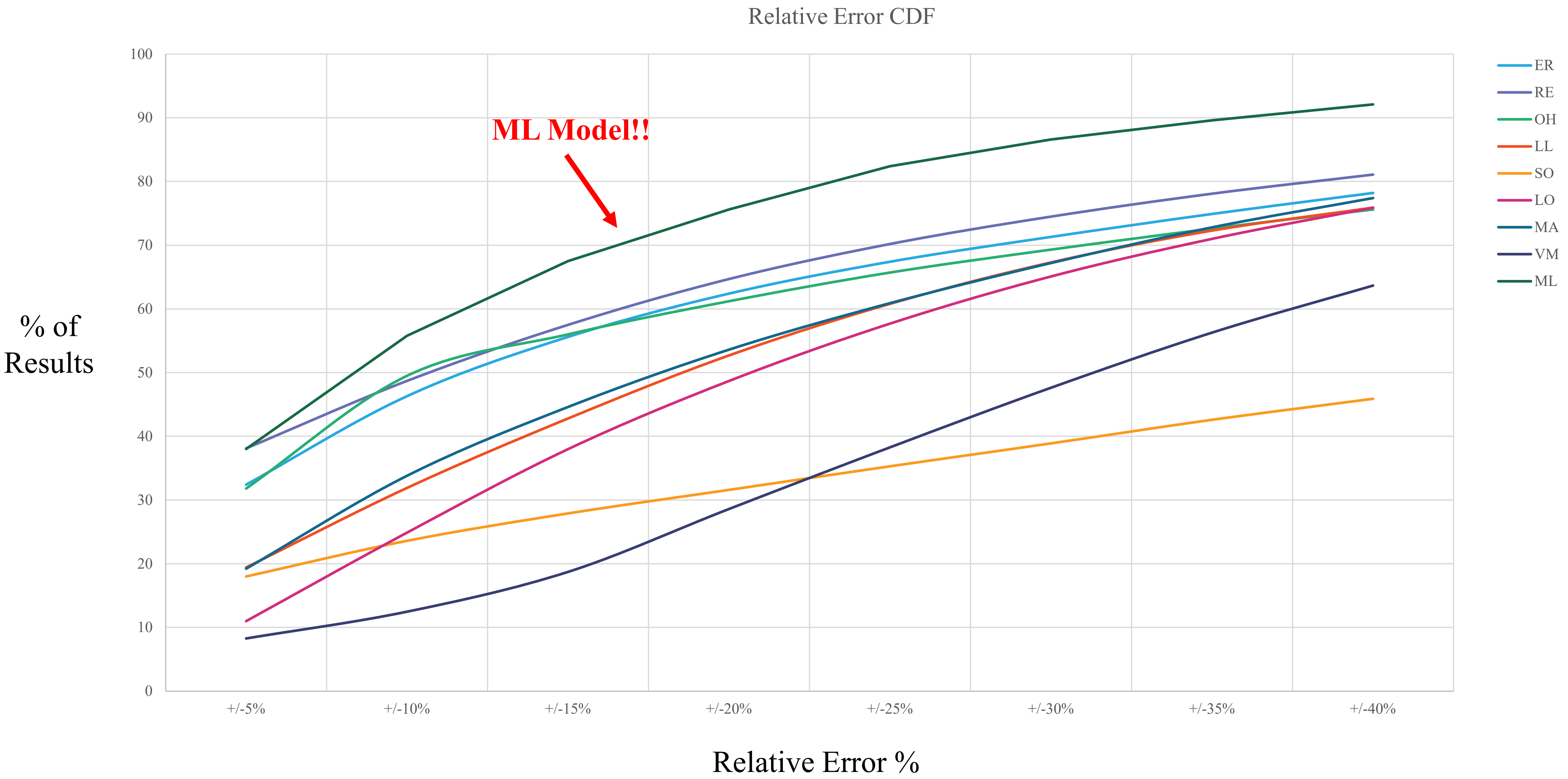
Error Metrics – MBD and RMSD

Model	**MBD (%)**	**RMSD (W/m^{-2})**
ER	-9.2	37.4
RE	-10.5	41.6
OH	-13.3	43.1
LL	11.9	45.7
SO	-98.3	199.9
LO	19.5	29.6
MA	21.1	33.2
VM	-60.38	50.4

Error Metrics – MBD and RMSD

Model	MBD (%)	RMSD (W/m^{-2})
ER	-9.2	37.4
RE	-10.5	41.6
OH	-13.3	43.1
LL	11.9	45.7
SO	-98.3	199.9
LO	19.5	29.6
MA	21.1	33.2
VM	-60.38	50.4
ML	**-5.87**	**35.32**

Error Metrics – Relative Error CDF



Relative Error CDF

ML Model!!

Takeaways

- Solution looks promising – needs more development
  - ML trained on only one NY weather station, ~131,000 measurements
  - 62M measurements in data set
  - Does using more than one weather station improve results, ie within a radius of target location where similar climate conditions are expected?
  - Does using all 236 weather stations improve results?
- Databases are very useful!
- Python/Jupyter environment worked well for this type of development.

Thank you!

77 Water Street
New York, NY 10005

matt.franks@arup.com

+1 212 896 3000

ARUP