Developing Science Gateways with Clowder for the Permafrost Discover Gateway

Luigi Marini,

Todd Nicholson, Kastan Dey, Chandi Witharana, Ingmar Nitze, Gala Wind, Lauren Walker, Robyn Thiessen-Bock, Chris Jones, Matt Jones, Kenton McHenry, Rajitha Udawalpola, Ehsan Bhuiyan, Jason Cervenec, Bidhya Yadav, Amber Budden, Michael Brubaker, Guido Grosse, Ben Jones, Aiman Soliman, Anna Liljedahl

Sept 22nd, 2022

Woodwell Climate Research Center



National Center for Supercomputing Applications

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIG

## **Science Gateway**

- **Goal:** Make models and tools developed by the PDG accessible to a wider set of user, beyond the original creators
- A science gateway is a resource to simplify access to community specific tools, applications and data collections for researchers, educators and students using user friendly, online interfaces

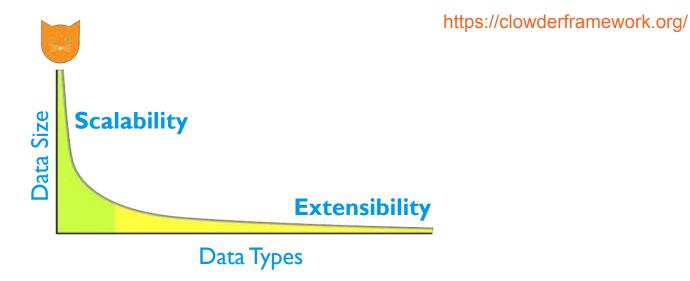
# **Data Pipelines**

- 1. Mapping Application for Arctic Permafrost Land Environment (MAPLE): detects ice wedge polygons from high resolution optical imagery archived at Polar Geospatial Center. Uses Deep Learning Convolutional Neural Networks and requires GPU (Python, Tensorflow) (*Chandi Witharana and Team*)
- 2. Permafrost Region Disturbances (PRDs): Lake area change, fire scars and retrogressive thaw slumps from Landsat images pre-processed using Google Earth Engine, plus additional machine learning and geospatial analysis (Python, Javascript, scikit-learn) (*Ingmar Nitze*)
- 3. Arctic Satellite Joint Product (ASJP): distills the high volume of archived NASA public satellite data to the essential community variables of interest that can be made available quickly and efficiently (Fortran) (*Gala Wind*)



# **Clowder and Long Tail Data**

Clowder is a <u>customizable</u> and <u>scalable</u> data management framework to support any <u>data format</u> and multiple <u>research domains</u>



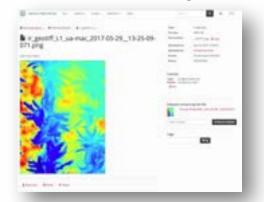
Heidorn, P. Bryan. "Shedding Light on the Dark Data in the Long Tail of Science." Library Trends, vol. 57 no. 2, 2008, p. 280-299. Project MUSE, doi:10.1353/lib.0.0036.



#### **CLOWDER OVERVIEW**

#### UPLOAD, SHARE, INDEX HETEROGENOUS DATA AND **METADATA** File visualizations, License, tags

- **Upload** files, tag, license and organize them in datasets and collections
- Selectively share datasets with collaborators and • publish them to the Internet
- Add well structured **metadata** to files and datasets using JSON Linked Data (JSON-LD)
- Automatically trigger execution of custom code on files and datasets using a Cloud based extraction system
- Visualize data and metadata in the browser using • custom previewers



#### Structured metadata

Metallata	60
-	100
) (see the	
- Address by Long Boundation Sec (1, 1973)	-
Lonather Rame (Husto A. (K). Lonather (K) (L) Longther (K) (L)	
- salar is up that a fact ( 20)	
- Milel & Lag Rent of Sec 7, 215	
Address for the tensor file (in the tensor	

Remarkan Alexander	
Manage Users of Space Plue Tower Data	
the second	
linear line largert	
and the second second second	
and the second se	
Description and the second sec	

#### Controlled vocabularies

wantania		19
Anto Instantiana		
	**	( )
and a	Test to	a state
Teriste Rates	(±.	-5.
No.	where a product of the	
	adata to a trip a gaget	
- March 1	ada an	and the second se
	and the second s	
Advention of	anti-method anguget	
Lattude	appropriate and	
Langthoise	Ante-Teach	
	All a family and a second	

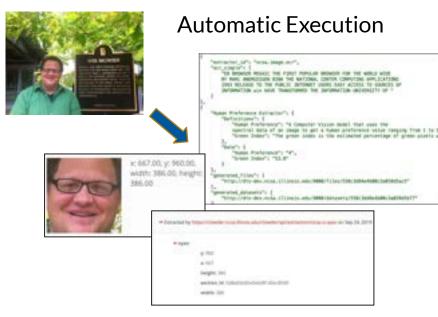
#### Advanced search

earch			inert letter
· · · · · · · · · · · · · · · · · · ·	The second secon	the series     the of the sector and and the sector and and     the sector and the sector and and     the sector and the	



#### **CLOWDER OVERVIEW**

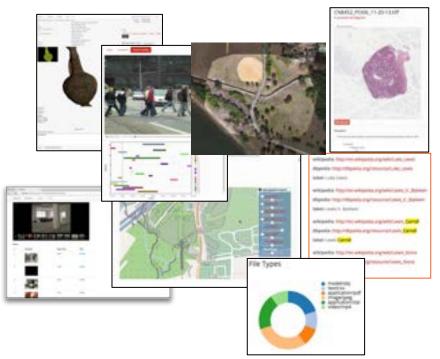
#### MANUAL AND AUTOMATIC DATA ANALYSIS AND VISUALIZATION



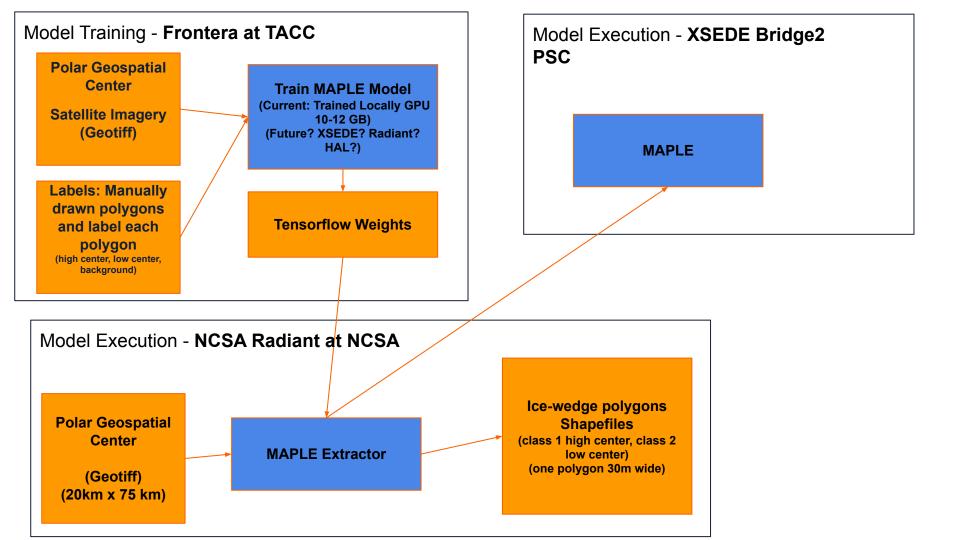
Develop extractors in Python, R and any other language



#### Visualizations







## **MAPLE on XSEDE Bridges2**

- Requires GPU resources not currently available on Radiant.
- Solution use ssh to submit job to XSEDE Bridges2
- **File Extractor**: individual files are transferred to XSEDE. A slurm job is created, run when resources are available, and then uploaded.
- **Dataset Extractor**: Start with empty dataset and path to input files on bridges. Determine which files have not been run, create slurm jobs, then upload. Checks to see which files do not need to be uploaded or run.



## **MAPLE Clowder Extractor**

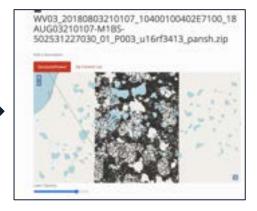
#### Input - GeoTiff



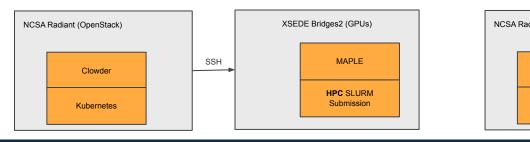
Submit



#### **Output - Shapefile**



#### Current







#### IBM-ILLINOIS DISCOVERY ACCELERATOR INSTITUTE



The Grainger College of Engineering

## **Portable Parallel Pipelines**

- Prototyped a Clowder extractor to Launch Parsl jobs on local Kubernetes cluster
- Prototyped Ray extractor to launch Ray jobs on CodeFlare provisioned clusters



https://parsl-project.org/

•••• RAY

https://www.ray.io/



https://codeflare.dev/



## **MAPLE Visualization Pipeline in Ray On Delta**

		-	S. 3					
2	0 0							
	Name of Street	A. De Anton						
								_
				222				
			1000				-	
			-	222				
			-					
			 River for all line	1111				
			-	222	-			
			*****	222				
			 -	122	-			
			-		No.			
			ALC: N	122	1100	10.00		
			1000					
			-	-		-		
			-					
			1000	100				
			-					
			E		1127.			

https://www.ncsa.illinois.edu/research/projecthighlights/delta/

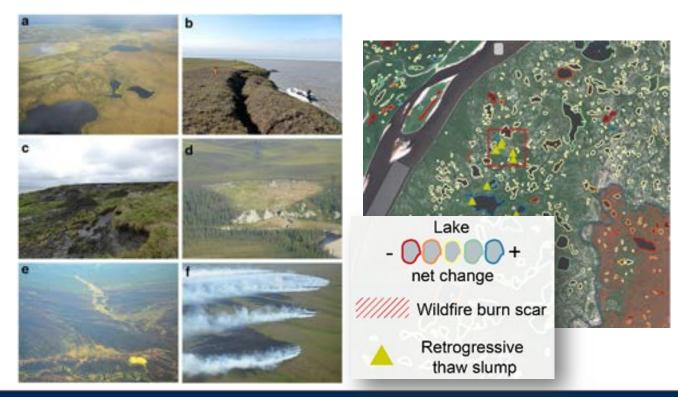


MAPLE file count • 10s of thousands. • Each image requires a dedicated GPU (or 8GB yram) for ML infrence								
-	min, No, star	machicster	total, No. court					
high_ice	8.536743e-05	1641.900	22370					
low, lot	\$536743e-05	1253,986	5456					
medium, his	8.536743#-05	1254.654	11209					
water_clipped	9.535743e-05	1071.421	8807					

Kastan Dey, Robyn Thiessen-Bock, Lauren Walker, Matt Jones

#### LandsatTrends - Permafrost Region Disturbances (PRD)

- Input Landsat
   & Sentinel
- Identify
  - Thaw slumps
  - Fire scars
  - Lake changes

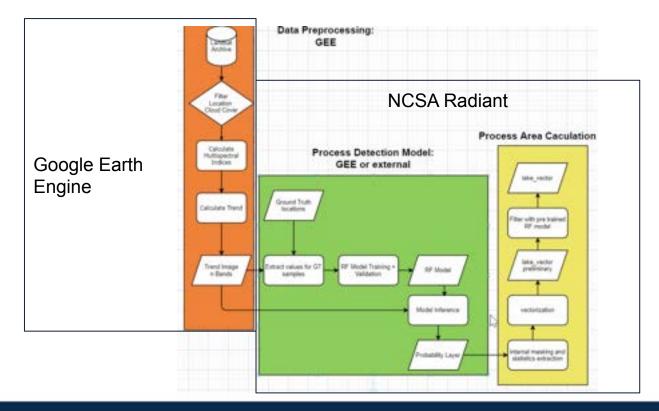






Ingmar Nitze et al. 2018

## LandsatTrends Pipeline





## **LandsatTrends Extractors**

- 1. Preprocessing Extractor
  - a. submits preprocessing steps to Google Earth Engine using Python API
  - b. Data from Google Earth Engine is transferred to Google Cloud using GC API
  - c. Data in Google Cloud is uploaded to Clowder
- 2. Inference Extractor
  - a. Run model on local files



# **Google Earth Engine API**

• Instead of using the google earth console or Google drive, images can be downloaded directly to a location on disk.

GEE Javascript Console



Local Python Code







#### LandsatTrends Detection Model and Area Calculation

#### Inputs - GeoTiffs



Submit



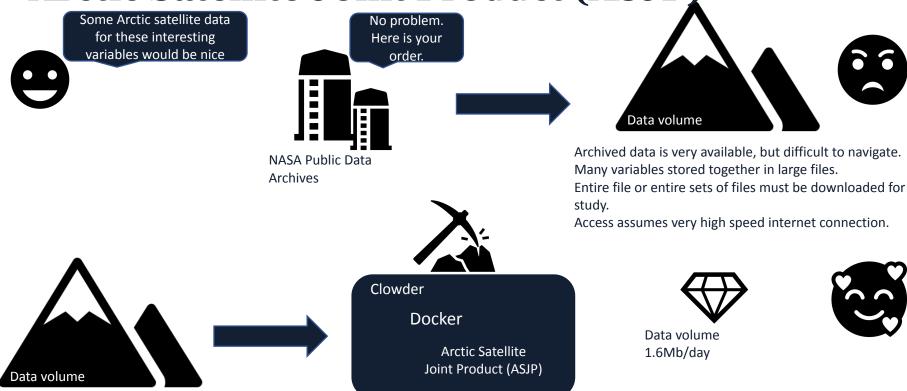
#### Outputs - GeoTiffs and Shapefiles

lake_change_grid_3000_netchange.tif maptif pr 16, 200 4148 #100101	A freedaat # hites # Mine	
lake_change_grid_3000_grossgain.tif mapent pris.coc 4518 #16001 #	A Second and Million Million	
lake_change_grid_3000_grossloss.tif	A Desetlari Winters	





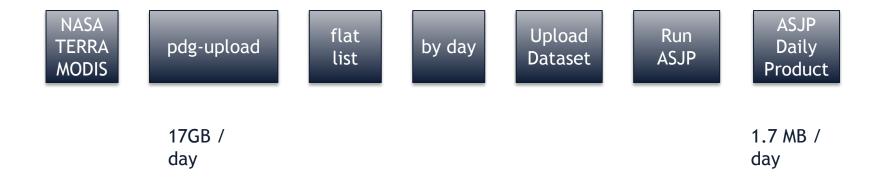
### **Arctic Satellite Joint Product (ASJP)**



Arctic Satellite Joint Product with help of Docker and Clowder distills the high volume of archived satellite data to the essential

community variables of interest that can be made available quickly and efficiently.

#### Gala Wind @ NASA





### Flat list of files

MOD06 L2.A2022156.1145.061.2022158133116.hdf MOD06\_L2.A2022156.1150.061.2022158132342.hdf M0D06 L2.A2022156.1155.061.2022158132334.hdf MOD06 L2.A2022156.1200.061.2022158132329.hdf MOD06 L2, A2022156, 1205, 061, 2022158132440, hdf MOD06\_L2\_A2022156.1210\_061.2022158132337.hdf M0D06\_L2.A2022156.1215.061.2022158132505.hdf MOD06\_L2.A2022156.1220.061.2022158132446.hdf MOD06 L2.A2022156.1225.061.2022158132455.hdf MOD06 L2.A2022156.1230.061.2022158132504.hdf MOD06 L2, A2022156, 1235, 061, 2022158132440, hdf M0D06\_L2\_A2022156\_1240\_061\_2022158133421\_hdf MOD06 L2.A2022156.1245.061.2022158133822.hdf MOD06 L2. A2022156. 1250. 061. 2022158133724. hdf MOD06 L2.A2022156.1255.061.2022158134130.hdf MOD06 L2, A2022156, 1300, 061, 2022158133235, hdf M0D06\_L2.A2022156.1305.061.2022158134313.hdf M0D06\_L2.A2022156.1310.061.2022158133842.hdf MOD06 L2.A2022156.1315.061.2022158133914.hdf MOD06 1.2 A2022156 1320 061 2022158134158 hdf

M0014.A2022150.0600.061.2022151092320.hdf M0D14.A2022150.0605.061.2022151090022.hdf M0014.A2022150.0610.061.2022151090442.hdf M0014\_A2022150\_0615\_061\_2022151090650\_hdf M0014, A2022150, 0620, 061, 2022151090110, hdf M0014.A2022150.0625.061.2022151090543.hdf M0014.A2022150.0630.061.2022151085940.hdf M0014.A2022150.0635.061.2022151090739.hdf M0014 A2022150 0640 061 2022151090812 hdf M0014, A2022150, 0645, 061, 2022151090248, hdf MOD14, A2022150, 0650, 061, 2022151090121, hdf M0D14.A2022150.0655.061.2022151091117.hdf M0014\_A2022150\_0700\_061\_2022151090007\_hdf M0014.A2022150.0705.061.2022151090633.hdf M0D14 A2022150 0710 061 2022151090757 hdf M0014.A2022150.0715.061.2022151090405.hdf M0014 A2022150 0720 061 2022151085812 hdf M0014.A2022150.0725.061.2022151085744.hdf MOD14.A2022150.0730.061.2022151085752.hdf M0014\_A2022150\_0735\_061\_2022151090432\_hdf

MOD29E1D.A2022148.061.2022151212855.hdf MOD29E1D.A2022149.061.2022151221800.hdf MOD29E1D\_A2022150\_061\_2022151223616\_hdf MOD29E1D.A2022151.061.2022152234746.hdf MOD29E1D\_A2022152\_061\_2022153140412\_hdf M0029E1D.A2022153.061.2022154094511.hdf M0D29E1D.A2022153.061.2022158013908.hdf MOD29E1D.A2022153.061.2022159224825.hdf M0029E1D, A2022154, 061, 2022158014936, hdf M0029E1D\_A2022155\_061\_2022156085054\_hdf M0D29E1D, A2022156, 061, 2022158014700, hdf M0D29E1D.A2022157.061.2022158090314.hdf M0029E1D\_A2022158\_061\_2022159080709\_hdf MOD29E1D.A2022159.061.2022160075838.hdf M0029E1D.A2022160.061.2022161080643.hdf MOD29E1D.A2022161.061.2022162092018.hdf M0029E10\_A2022162\_061\_2022163084023\_hdf M0029E1D.A2022163.061.2022164080310.hdf M0029E1D\_A2022164\_061\_2022165101941\_hdf MOD29E1D.A2022165.061.2022166075322.hdf

## Folders by day

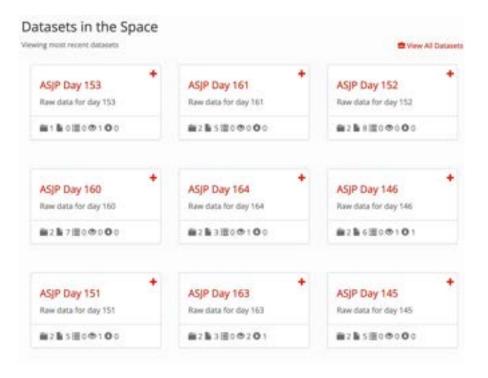
lmarini@pdg-upload:/data/modaps\_by\_day\$ ls
142 143 144 145 146 147 148 149 151 152 153 154 155 156 157 158 159 160 161
162 163 164

lmarini@pdg-upload:/data/modaps\_by\_day/163\$ ls MOD06 MOD14 MOD09CMG.A2022161.061.2022163043428.hdf MOD29E1D.A2022162.061.2022163084023.hdf MOD10C1.A2022161.061.2022163054356.hdf

#### MOD09CMG.A2022161.061.2022163043428.hdf



### **Raw Data per Day**



NCSA | NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS

### **Private Dataset per Day**

ies Metad	fata Extractions Visualizations Co	mmenta (5)			
ked Films (Ø) +					
	MOD06	& Download			
î	MOD14	≜ Cownicad			
•	MOD10C1.A2022161.061	1.2022163054356.hdf	•		
1.00	application/s-hdf Jun 13, 2022	& Downkast	man befreite	Aryge Spaceforms joined Product	
	5.0 MB 10 2 10 0 2 📋	* Follow	prig landsattend	landtattenni natatat	
		Move	pdg tandtument takaanalysis	sale analysis landsat extractor	
			prig-maple 2.0	MAPLE 3.5 adapted as estimated	



#### **Public Data**

Dwner: Luigi Ma	Outputs	All Rights Reserv	ed Luigi Marini
reated on Mar dd-creatar(s)	25, 2022		
dd a descriptio			
Add Files     ✓ Create Fold	▲ Download All Files ▲ Archive All er ▲ Submit for extraction	2 Unarchive All	Deleta * follow
Files Met	adata Extractions Visualizations	Comments (0)	
performance.			
Marked Files (B)			
Marked Files (B)	ASJP Day 164.nc4		
	ASJP Day 164.nc4		± Download



## **Recovering from Failure**

- Keep track of uploaded data uploaded (checkpoint)
  - which files are uploaded successfully, which files are not
- Avoid creating duplicate datasets, uploading duplicate files
- Do not upload partial files being written to disk
- Always confirm integrity of uploaded files before deleting original files on disk

Number of the second se	Ba.H	Attest 14	NUMBER OF TAXABLE PARTY.
Intelections in my little work with A work of AND THE ARE AND ADDRESS AND	RT1RelationsT1994818	APTIMUM ADMITTHE	877 March 101 March 101
Internetage, by stay 1010 140 Million AMIN' TO ADD THE SECOND SECTION AND ADDRESS OF	ALT I MADE DOWNED DOWNED	APT I Note Apres 2 (Section	All there is not the little
Testamology, by page (0.021102) 4021010 40400144.0000140.000140.000140.000140.0001	ATT I Make NUMBER OF STREET	ALL PROPERTY AND ADDRESS OF TAXABLE	\$1100-00-0010-00-00
International State of the Stat	ATT CRIME-SCHOLD DEADLE	OF CONTRACTORS OF A	81108-0-00109-00
International for the PERFECTATION AND INCOME. ADDITIONAL PROPERTY AND ADDITIONAL AD	ACT COMPARISON (CARDING	IN TRADUCTION OF TRADE	demonstration and
International Society (\$1507102100014-000014-000021-01.0000.001.00021-000000014-0	BTT Disartition (17 (house)	OF COMPACT OR ADD	ST-MANHOUSED MADE
Hardwinnedages, bug, dags \$5000 100 WHICH A MILLION ADDRESS AD 1800, DR1, MILLION ADDRESS ADD	ACCOMPANY AND ADDRESS OF	MITTRADUCTION AND	10110-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0
HARAFESTING AND AND A CONTRACT AND A CONTRACT OF A CONTRACT AND A	REPORT OF A PARTY OF A PARTY	STORESSING CONTRACTOR	distance in the second second
Managements in the Part of Million Amode is American and the part of the second state	RTTNAMENDUST (SALDIN	In the second second	-
Instalmentages, her, shay 10000 million and the addition of states and a state water states and	AT CONSISTENCE Distance	ET OblahimET (SAED)	61106-00-00-01-01-01
Administration, in: dog/0002140/00014/00014.4002140.0012.001.0002140004012.007	ADT COMMITTERNET (PARTY)	ET 1944-BENET/SAED	ET-TRAINING TOTAL
International Statements of the statement of the statemen	AT CONSISTENCE SAME	BIT I MARKAGENET CRAMEN	47-19-0-10-14/1-10-4
Intervention, in: Page 2010 Told MICELANCO CLASSED FOR THE DRIVEN AND DESCRIPTION	ALT I DOWNLOW COMPANY	ALC: Determined Constant	All these spectra made
Value includes, by Jug 1986 148 MORT AMOUNT AM	\$171 Manufacture (1) (1)	ALL DAMAGEMENT CONTINUES.	STORAGE MADE
Idela Transform, No. and William & William & WORD 14, ADD 14,	ACCORDING TO AND ADDRESS OF TAXABLE	AT DESCRIPTION CONTRACTOR	distant for the second second
International Society and Party and	ACT OR ADDRESS OF THE OWNER	BITTING BUILDINGS TRANSPORT	01108-00-00-00-00-00-00-00-00-00-00-00-00-0
Intelligence by they been well a should be added the second and added and they been and	BT TReach Invit Unrolated	OT DRAWNING COMMON	ST-BANGOLOUP MAD
International Statement in the West of Manual Advances and the statement of the	ATTING-ISTATISTICS	8711040-00-001010-0010	BTT Data Advances Made
Value and a state of the second	ALL CONTRACTOR OF CONTRACTOR	ATT TRANSPORT	11-IN-MARKING-THE
Manhatimentagine, Say, Many WARD THE MATCH A MARCH TALADARD THE 2008 AND ADDRESS AND ADDRESS AND	RT19AL-BOART18ARA	OT Distance in the state	STORAGE COMPANY STATE
Industry and the second s	ADT 1 Marca Robustle (Marchill	Er 1964 Automativitestille	611566-60407-3140
Valuetreningen, Sen skey Malaring Malaring Malaring Malaring Sciences and Sciences and	ALL DRIVENING TRAFTIC	ET Underfallent Prinkter	All Designations
Industry and an any first water and the second statement of the second statement and	MITTRACE BUILDINGS CHAPTER	APT 1 Mindre Automatic Victoria and Victoria	APP Designation of the local
Value Transmission, Say, Mary 1950, "Add MARCO 4, AMOD 14, AMOD 141, 2020, AM 2020 142000000, AM	ALL Discription Designation	and close-surveil close and	distance and the set
Chalacteristics, Inc. Mag. 2002 Tol: MICH 14 MICH 14 ADDITION OF A DISTANCE OF A DISTA	ATT I MADE AND TAKEN	APT 1 Market and 1 To Ave. (1)	-
International Inc. And Statement Without A Without A construct A state and party addressed and	All International Contents	All Condensations and statements	An Inc. of Concession, Name

# **Clowder v2 - Join the Development!**

https://github.com/clowder-framework/clowder2

	Pas Measure					1988	
here a second se	here		-	194		Anna	- Call
ater	Contraction and the	to Chail Mang				Nere Salarin Index 107	1
Transit (	a non exercision (a)	$A_{13} \approx 2000, 2000, Physics Charleson$	-	-		Decer Thereforeg President Ang. 77, 2022, 21 (8-76)	
State State State	🔒 incertante, pre jung, and 👾	Aug 12 2020 2120 PM for Chair Mang.	-	-	A Dorte	All and a second s	
To angle in some	S intermetter (S)	Aug 12 2022 21:20 PM to Chair Mang.		-	# Dobe	Last result don 17, 2011 (1-4040)	
-		Aug 17, 2010, 21, 22 PM by Charl Marg.	100 1000	-		Developin 2 Last territopini Norm	
And card Inspect	a set franker (a)	And 12 YO R DOLLAR PROVIDE HAVE		-	-		





CLOWDER v2

### **Thank you!**

https://permafrost.arcticdata.io

https://clowderframework.org/



Navigating the New Arctic Awards # 1927872, 1927723, 1927729, 1927720, 1927920, & 2052107



