

A satellite view of the Earth, showing the Americas and surrounding oceans. The image is darkened to serve as a background for text.

MSCAR 2020 DATA SCIENCE TUTORIAL

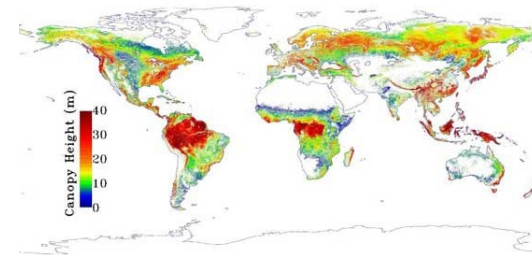
Introduction

WHAT IS GEOSCIENCES DATA?

Geosciences data is any data that is a record of a physical or descriptive characteristic of the earth system, or its inhabitants that has a geographic location associated with it. Often times, geosciences data will also have a time associated with it.

Examples of geosciences data include data from weather stations, satellite imagery of earth, locations and times of earthquakes, surveys of tree populations in a forest, economic impact data from weather disasters, output from a weather forecast or climate model, photos of glaciers, surveys of human illness related to water quality, and many others.

As you can see, the geosciences data extends well beyond what we typically think of as “weather” or “climate” data.



The background of the slide is a dark, moody photograph of a library. Tall wooden bookshelves filled with books line the walls. In the foreground, a large computer monitor is visible, displaying a complex data visualization or software interface. The overall lighting is low, creating a sense of quiet study and research.

WHAT IS GEOSCIENCES DATA?

- Geosciences data can take many forms
 - Digital: Data that is stored by a computer in binary format.
 - Analog: Data that is stored as a physical record.

What are examples of each of these?

In geosciences, what are common examples of these that we have used in the past, and currently?

Data collection method	Definition	Example(s)	Advantages	Disadvantages
<i>In situ</i>	Measured directly by a sensor within the medium in its natural setting			
<i>In vitro</i>	Measured directly by a sensor within the medium outside its natural setting			
Computationally	Using a model to simulate and understand natural systems			
By proxy	Using one dataset to infer about another that cannot be directly measured (frequently used to reconstruct data that can not be observed)			
Remote sensing	Using electromagnetic radiation to retrieve information			

DATA COMES IN MANY FORMS

Data collection method	Definition	Example(s)	Advantages	Disadvantages
<i>In situ</i>	Measured directly by a sensor within the medium in its natural setting	Thermometer	Direct measurement Can characterize instrument	Only available where you have sensor(s)
<i>In vitro</i>	Measured directly by a sensor within the medium outside its natural setting	Aerosol filter brought back to lab	Can collect and reanalyze	Cumbersome?
Computationally	Using a model to simulate and understand natural systems	Climate model	Computational power is cheap compared to obs	"All models are wrong, but some are useful"
By proxy	Using one dataset to infer about another that cannot be directly measured (frequently used to reconstruct data that can not be observed)	Measure tree rings to infer precipitation	Can allow synthetic observations past measurements	Data not collected in a controlled setting/ relationship between proxy and variable?
Remote sensing	Using electromagnetic radiation to retrieve information	Using infrared radiation to measure temperature	Can often be applied over a large area	Relationship between EM energy and variable?

DATA COMES IN MANY FORMS

LOG of the UNITED STATES Steamer Bear
Arrival at Cape Sabine & Discovery of Greely Party

Rate, Guns,

under the command of Lieutenant W. H. Emory Jr., U. S. Navy,
Sunday June 22, 1884.

Hour.	Knots.	Fathoms.	Course steered.	WINDS.		Barometer.	TEMPERATURE.					State of the Weather, by symbols.	Form of Clouds, by symbols.	Clear Sky in 10ths.	State of the Sea.	Bearing of the sail the vessel is under at end of watch.
				Direction.	Force.		Height in inches.	Ther. air.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.					
A. M.																
1	5	0	E by N 3/4 N	S. W.	3	29.84	40	32	32	30	0.0	bc	sc	0	1	
2	6	0	N by E 1/4 N	"	3-4	29.85	41	31	31	30	"	"	"	0	0	
3	7	0	N by E	"	3	29.85	47	30	30	30	"	"	"	0	0	
4	7	0	N by E	"	3-4	29.83	44	29	29	30	"	"	"	0	0	
5	8	5	N by E	"	3	29.83	44	29	29	31	"	"	"	0	0	
6	8	5	"	"	3	29.83	44	31	31	30	"	"	"	0	0	
7	8	5	"	W	3	29.83	44	31	31	30	"	"	"	0	0	
8	8	5	"	"	3	29.83	44	32	32	31	"	"	"	0	0	
9	7	5	E by N	S. W.	3	29.82	42	32	32	31	ccs	mi	"	0	0	
10	7	5	"	N. S. W.	3	29.81	43	33	33	32	"	"	"	0	0	
11	7	5	"	"	2	29.80	44	33	33	32	"	"	"	0	0	
Noon.	5	9	E by S	N by N	2	29.79	44	33	33	32	"	"	"	0	0	

RECORD OF THE MISCELLANEOUS EVENTS OF THE DAY.

Data management – planning and actions related to ensuring that collected data is stored securely and disseminated and preserved.

Data security – data should be treated carefully because it costs time/money to collect. Reasonable steps should be implemented to ensure that data is not lost.

Open data – the concept that collected data should be available to anyone, particularly datasets collected with taxpayer funding, with appropriate documentation.

Open source – the concept that code is data and should be open and available for review and for use by others, because it costs time/money to create, and re-create, and is part of the scientific process.

Reproducibility – data and code should be stored in a way that it can be used in the future by everyone as part of an accountable scientific process.

DATA PRINCIPLES

Latitude by Observations at Noon @ 1 P.M. Littleton Island N 78° 24'

Longitude by Chronometer from Forenoon Observations @ 1 P.M. W. 73°

Current during the time knots fms. per hour, setting to the

Variation of the Compass by Amplitude @ observed at Sunrise

Variation of the Compass by Azimuth @ observed at

Water expended during the preceding 24 hours gallons.

Water ... during the preceding 24 hours

Water remaining on hand fit for use at Noon

Coal consumed during the preceding 24 hours tons 17 28 lbs.

Coal remaining on hand at Noon 7403 " 2150 "

P. M.																
1			E by N	N by N	4	29.80	46	32	32	31	cc	bc	sc	0	0	
2			N by W	N by W	5	29.80	48	33	33	32	"	"	"	0	0	
3			N by E	N by W	5	29.80	49	34	34	32	bc	sc	sc	3	0	
4	6	0	"	"	5	29.81	54	34	34	32	"	"	"	2	0	
5	6	7	N by E	"	5	29.78	48	33	33	32	"	"	"	2	0	
6	6	2	"	"	5	29.76	48	33	33	32	"	"	"	2	0	
7			W by S by E	"	6	29.76	48	33	33	32	"	"	"	8	0	
8			"	N by W	6	29.76	48	33	33	32	"	"	"	7	0	
9			"	N by N	9	29.68	56	32	32	32	ccg	"	"	8	0	
10			"	"	8-9	29.68	56	32	32	32	"	"	"	9	0	

IN THIS COURSE, WE WILL USE TOOLS THAT FOLLOW DATA PRINCIPLES

Data management: create plan for how data will be obtained and kept safe

Data security: cloud computing, backups, other technologies

Open data: use publicly-available data; share online, with documentation and appropriate referencing

Open source: python + cloud/github version control

Reproducibility: Source all datasets, keep all code, and publish online

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
Aiken	38	52	29	44	24	60	25	70	41	65	56	60	41	65	56	60	41
Allendale	32	59	26	53	31	52	38	70	30	68	53	61	50	54	48	62	36
Anderson	29	50	24	39	23	47	26	64	39	61	55	60	27	57	26	53	24
Augusta, Ga.	49	51	32	43	29	55	29	70	47	66	60	62	39	45	32	53	25
Barksdale	27	39	22	48	28	40	28	66	35	62	56	60	40	46	24	52	24
Batesburg	34	57	25	48	24	62	23	69	27	64	46	50	46	54	26	58	28
Beaufort	43	56	33	50	31	58	31	65	44	62	54	62	51	52	35	54	29
Bennettsville	60	24	62	39	55	26	45	13	64	39	59	11	54	24	45	28	60
Blackville	67	35	66	41	38	29	31	26	38	18	50	17	42	20	50	34	64
Bowman	65	38	67	37	59	30	32	25	36	18	44	15	40	26	55	30	63
Charleston	66	45	65	47	58	50	50	26	37	23	46	24	48	36	57	41	64
Charlotte, N. C.	50	57	28	41	21	38	21	62	39	63	55	59	40	43	30	50	29
Cheraw (O.)	32	44	26	36	20	57	21	66	26	70	46	57	36	45	32	54	23
Clark's Hill	36	51	28	44	26	51	26	68	45	64	57	61	43	45	28	54	26
Clemson College	24	48	21	40	26	48	28	62	39	65	51	63	42	48	27	54	17
Columbia	39	47	30	39	24	58	25	68	48	67	59	61	39	44	31	53	32
Conway	30	47	25	40	22	61	21	69	26	71	39	61	59	49	31	54	23
Darlington	31	48	23	41	21	60	19	68	36	72	57	62	50	53	32	56	22
Dillon
Due West	30	44	29	41	24	52	28	65	39	62	56	56	42	46	28	49	23
Florence	25	51	25	54	20	61	26	68	21	72	47	59	58	58	34	55	24
Gaffney	32	49	22	40	19	48	20	65	28	65	42	58	46	48	28	54	23
Georgetown	22	52	25	55	27	48	34	55	35	70	50	70	48	65	33	64	32
Gillisonville	37	54	29	44	28	52	26	70	49	69	59	64	44	44	33	53	22
Greenville	24	46	24	37	22	44	21	59	27	60	39	53	40	45	23	50	19
Greenwood	23	46	27	46	24	49	24	33	20	62	46	50	45	45	29	49	26
Heath's Springs	27	54	24	48	18	48	20	55	35	60	40	66	34	58	28	46	32
Kingstree (O.)	34	49	29	53	25	57	26	67	31	70	45	58	56	56	31	55	24
Liberty	28	46	26	39	25	45	38	59	39	60	56	60	43	58	23	50	20
Little Mountain	34	47	26	39	23	56	24	67	44	64	56	60	40	45	28	53	28
Lugoff	33	51	29	45	22	50	20	68	36	60	34	48	33	45	29	55	28
Newberry	35	47	26	38	23	54	24	68	30	64	59	59	48	45	28	50	24
Pinopolis
St. Georges	35	50	30	42	26	56	27	66	28	67	43	60	54	46	31	54	26
St. Matthews	34	47	31	40	25	54	26	68	26	68	46	59	58	58	36	52	28
Saluda	36	50	26	45	24	58	23	68	38	64	58	61	43	45	29	53	24
Santuc	28	46	20	42	20	65	54	66	45	48	27	52	20
Savannah, Ga.	44	51	35	46	32	57	35	66	51	66	60	61	45	49	36	52	33
Seivern
Society Hill	27	43	22	43	22	53	40	65	55	70	56	67	34	45	26	50	39
Spartanburg	30	48	24	38	22	45	22	64	28	60	41	55	46	47	27	52	21
Stateburg	39	48	29	39	23	58	24	69	47	69	57	62	43	46	30	56	29
Summerville	31	51	26	44	27	60	42	66	57	68	57	58	33	48	26	54	38
Spartanburg	30	70	37	63	34	64	31	46	31	62	49	67	61	61	40	45	32

WHAT IS DATA SCIENCE?

“**Data science** is an interdisciplinary field that uses **scientific** methods, processes, algorithms and systems to extract knowledge and insights from **data** in various forms, both structured and unstructured...”

“Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.”

- Wikipedia, retrieved 1/10/19

WHY DATA SCIENCE?

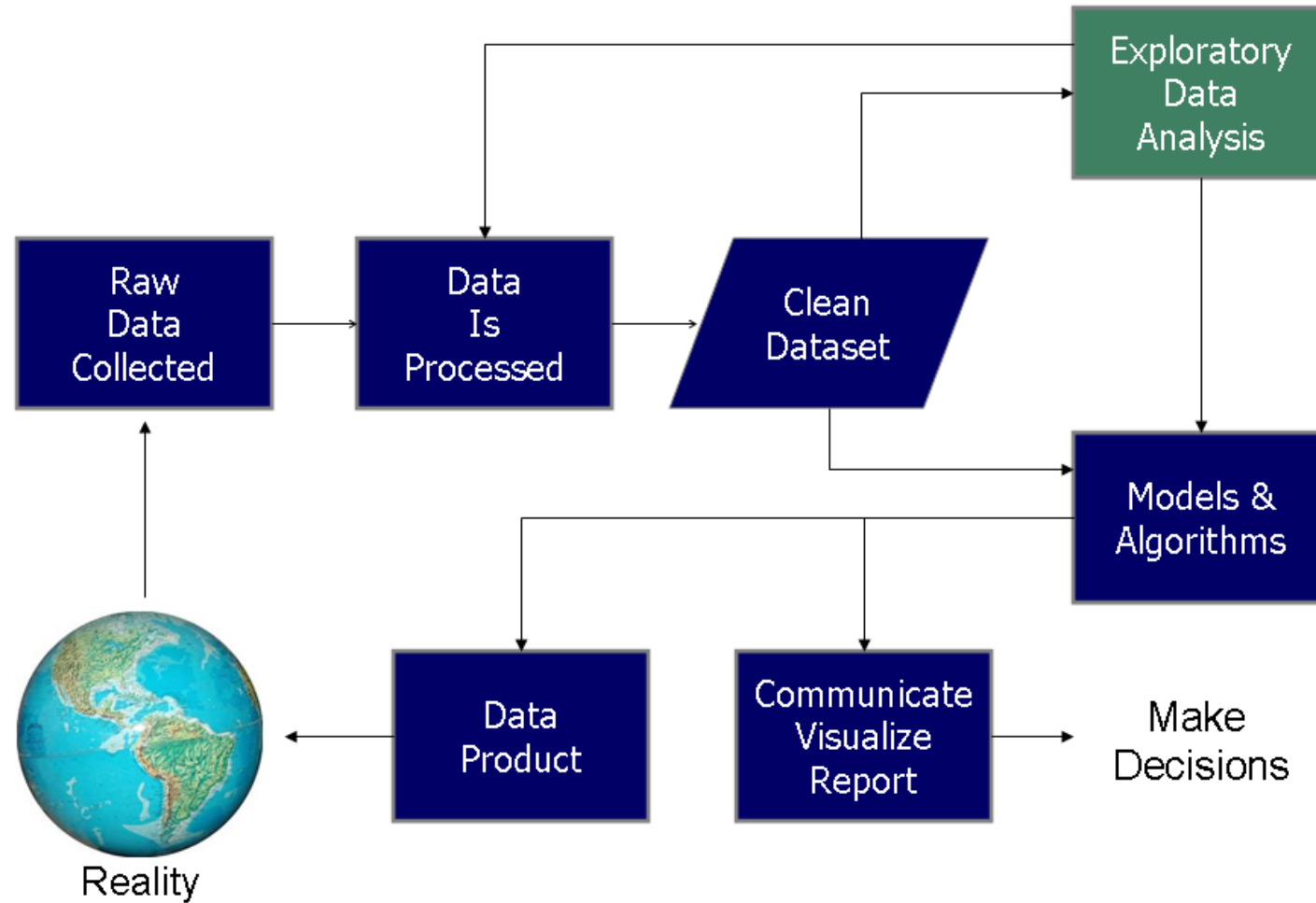
Computer hardware and software costs have decreased to the point where they are readily and easily available – and sensor advances, new technology and availability of internet + low cost devices has led to an ‘avalanche of data’

We need tools and experts to deal with the ‘data avalanche’.

Data and its analysis has value and can be used to make scientific advances, help stakeholders make decisions, which translates into economic value and can save lives.

This has led to a rapidly growing demand for expertise in dealing with data.

Data Science Process



TODAY'S AGENDA

This course will introduce you to the tools of data science and give example applications for using data science to analyze data for geophysical applications.

We will use python + jupyter notebooks + cloud computing + github.

These are not the only tools that exist, for example, the R programming language is also a popular data science platform.

The general concepts of analysis will apply to any analysis tools, so once you learn one tool, it will be easier to learn another!