MSCAR 2020
DATA SCIENCE TUTORIAL

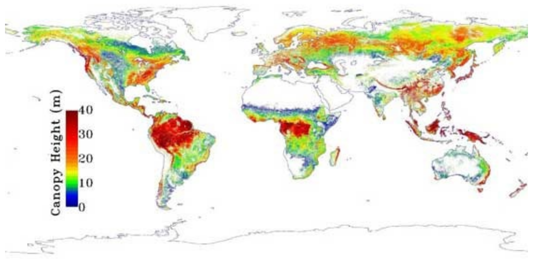
Introduction

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.



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?

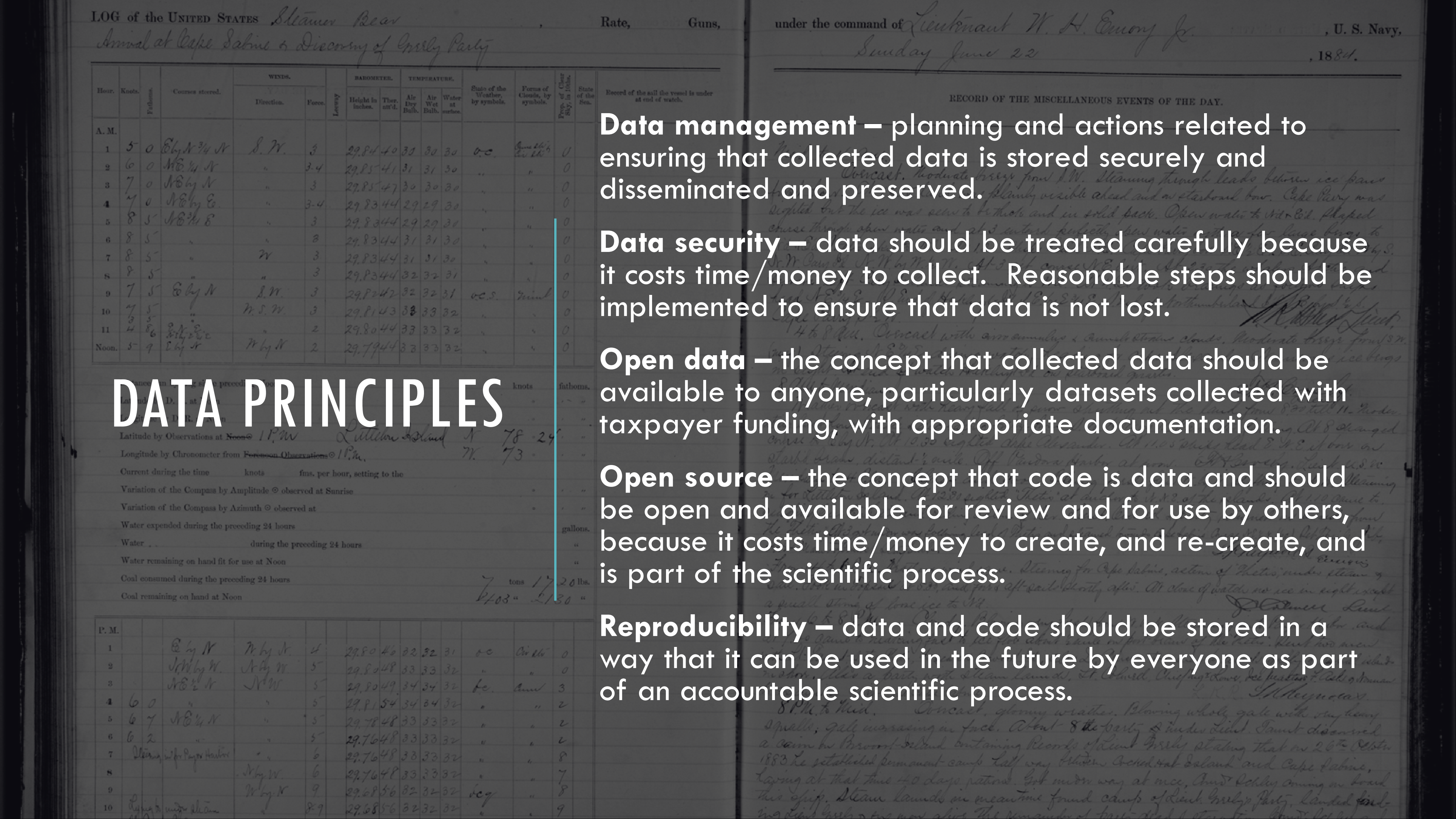
In geosciences, what are common examples of these that we have used in the past, and currently?

Data collection method	Definition	Example(s)	Advantages	Disadvantages
In situ	Measured directly by a sensor within the medium in its natural setting			
In vitro	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 COMES IN MANY FORMS

Data collection method	Definition	Example(s)	Advantages	Disadvantages
In situ	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)
In vitro	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?


LOG of the UNITED STATES Steamer Bear
Arrival at Cape Sabine & Discovery of Greely Party
under the command of Lieutenant W. H. Emory Jr. , U. S. Navy,
Sunday June 22 , 1884.
RECORD OF THE MISCELLANEOUS EVENTS OF THE DAY.

DATA PRINCIPLES

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.

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

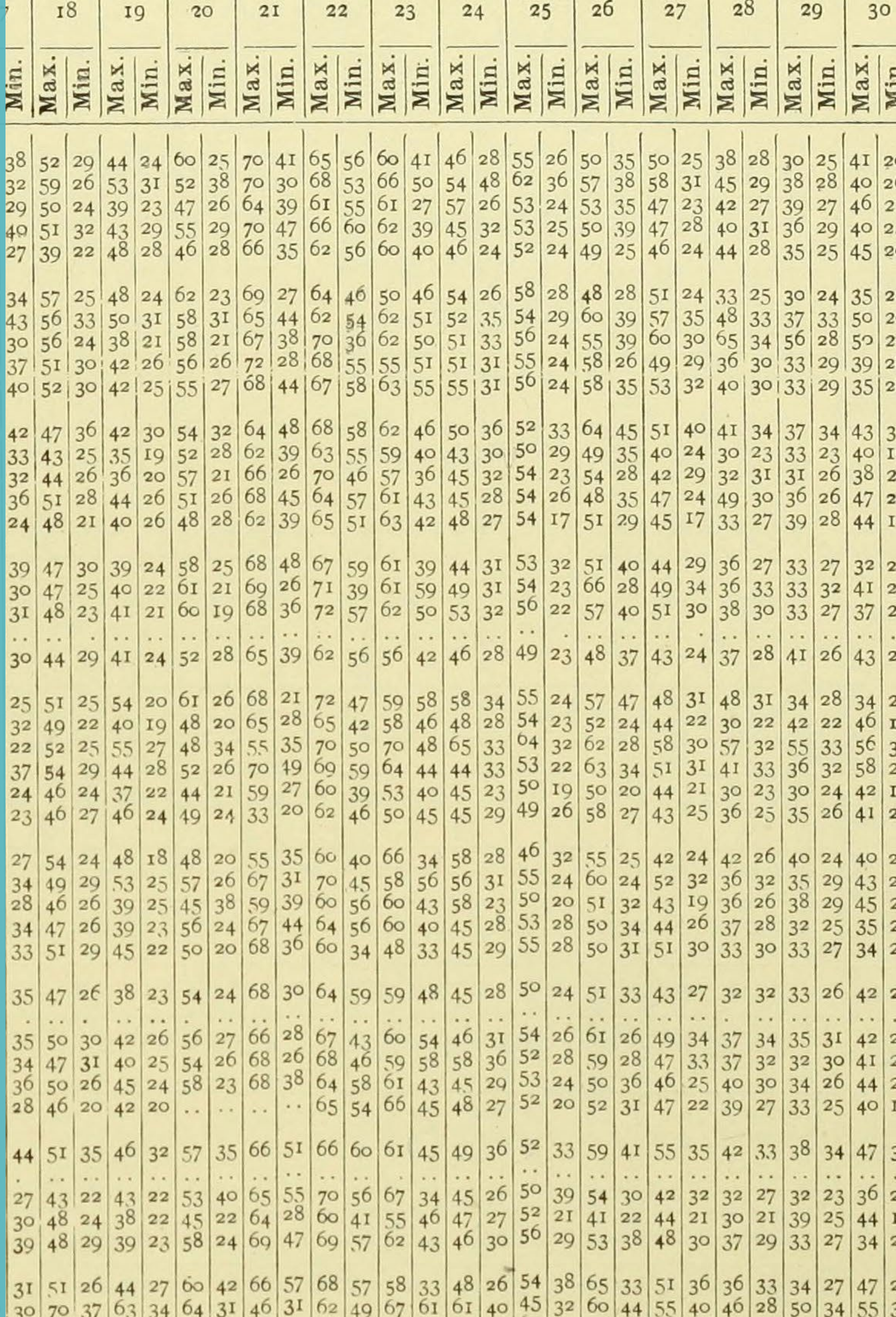
Data management: create plan for how data will be obtained and kept safe

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



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?

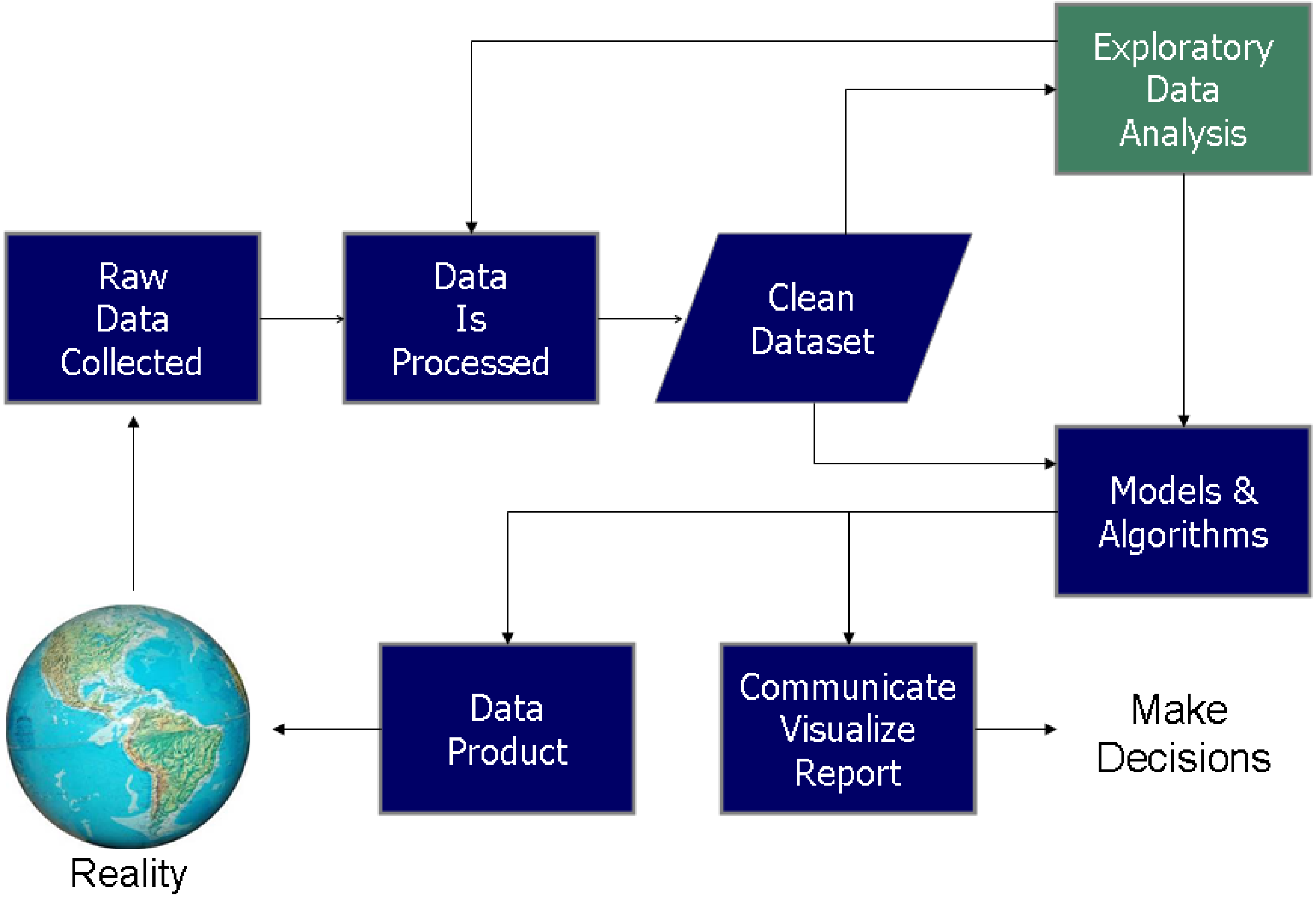
Computer hardware and software costs have decreased to the point where they are readily and easily available – and sensor advances, new technology and availability of internet + low cost devices has led to an ‘avalanche of data’

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



Source: Wikimedia.org

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!