Open Data on AWS

https://opendata.aws

Why does AWS care about open data?

Sharing data on AWS makes it accessible to a large and growing community of researchers, entrepreneurs, and enterprises who use the AWS Cloud.

Many AWS customers supply data to the public to accelerate research and product development.

Many AWS customers use data shared on AWS to create new products and services.

aws

The AWS Open Data program expands access to data by staging it for analysis in the cloud.

https://opendata.aws

“No matter who you are, most of the smartest people work for someone else.”

— Bill Joy, co-founder of Sun Microsystems

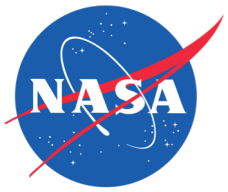
aws

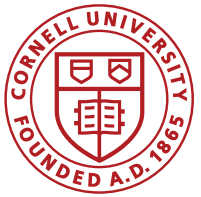
How can we maximize access to data?

aws

AWS Public Datasets

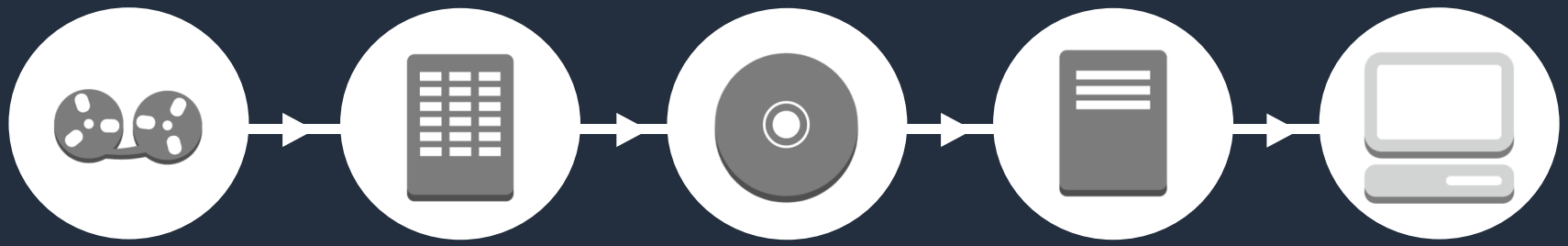
https://registry.opendata.aws





aws

Traditional data acquisition



"…data must be organized, well-documented, consistently formatted, and error free. Cleaning the data is often the most taxing part of data science, and is frequently **80% of the work**."

— Data Driven by DJ Patil and Hilary Mason

aws

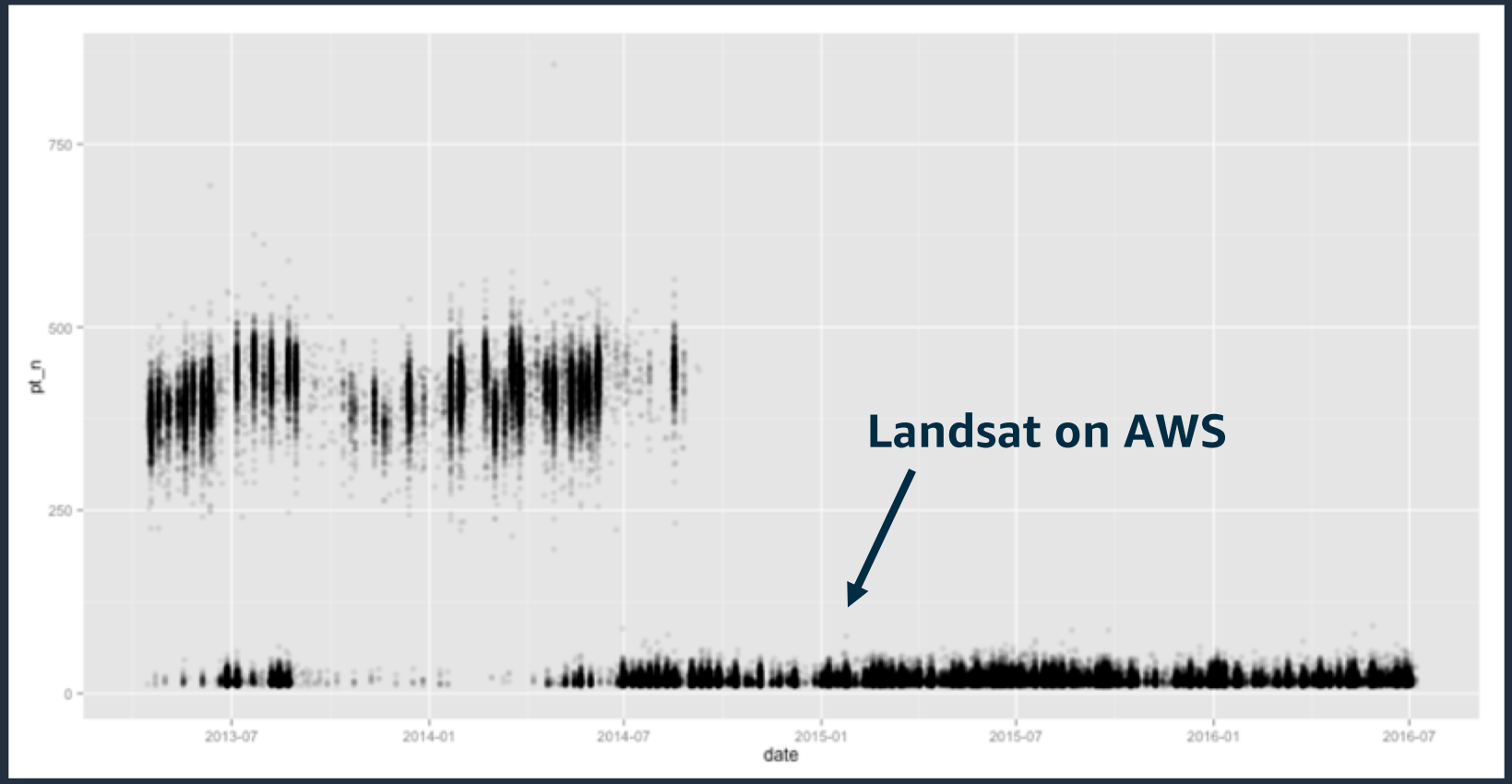
Undifferentiated heavy lifting

aws

It’s easier to bring algorithms to data than to bring data to algorithms.

It is more efficient to bring computing resources to data than to download and copy data to local computing resources.

aws



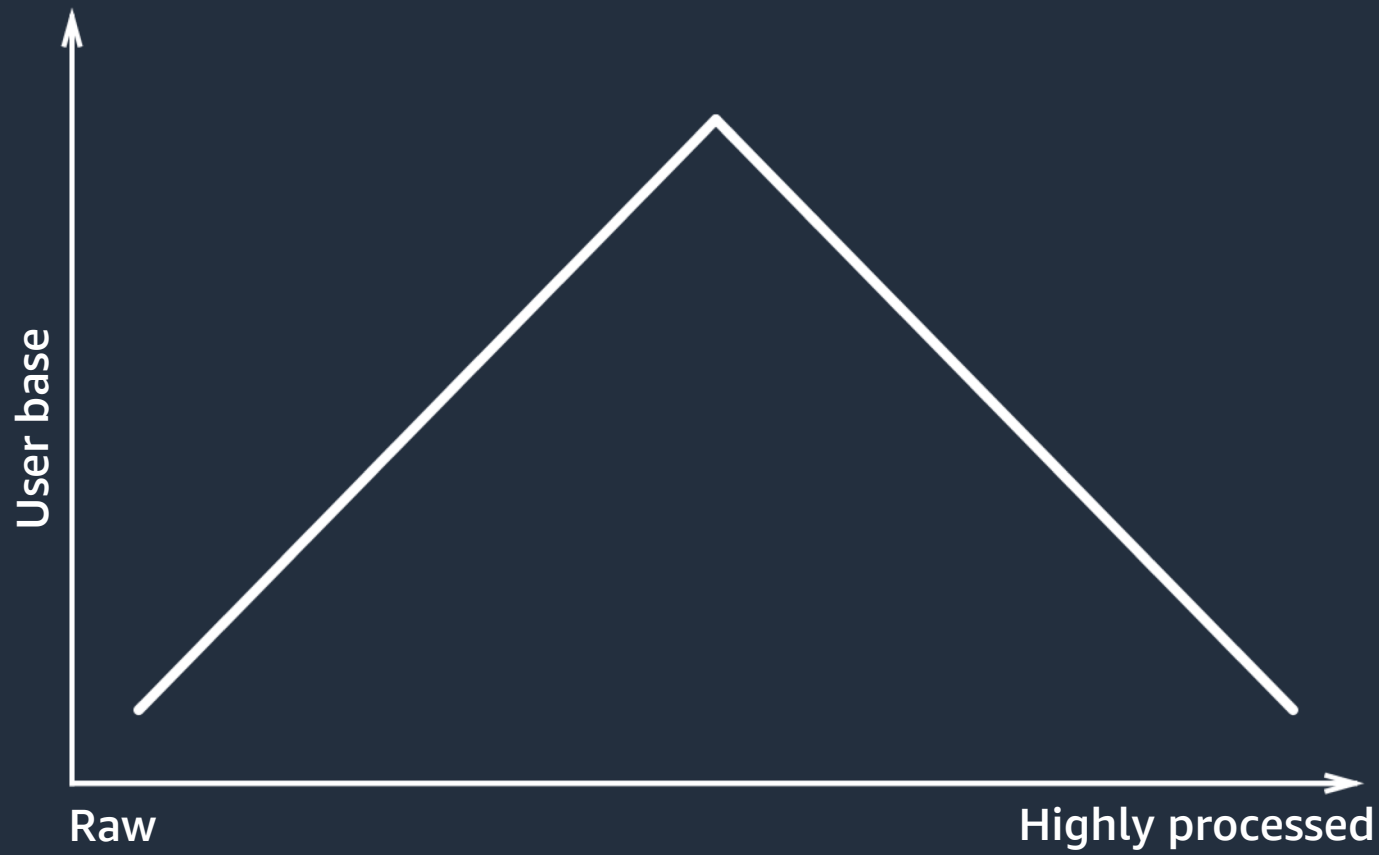
Graph by Drew Bollinger (@drewbo19) at Development Seed

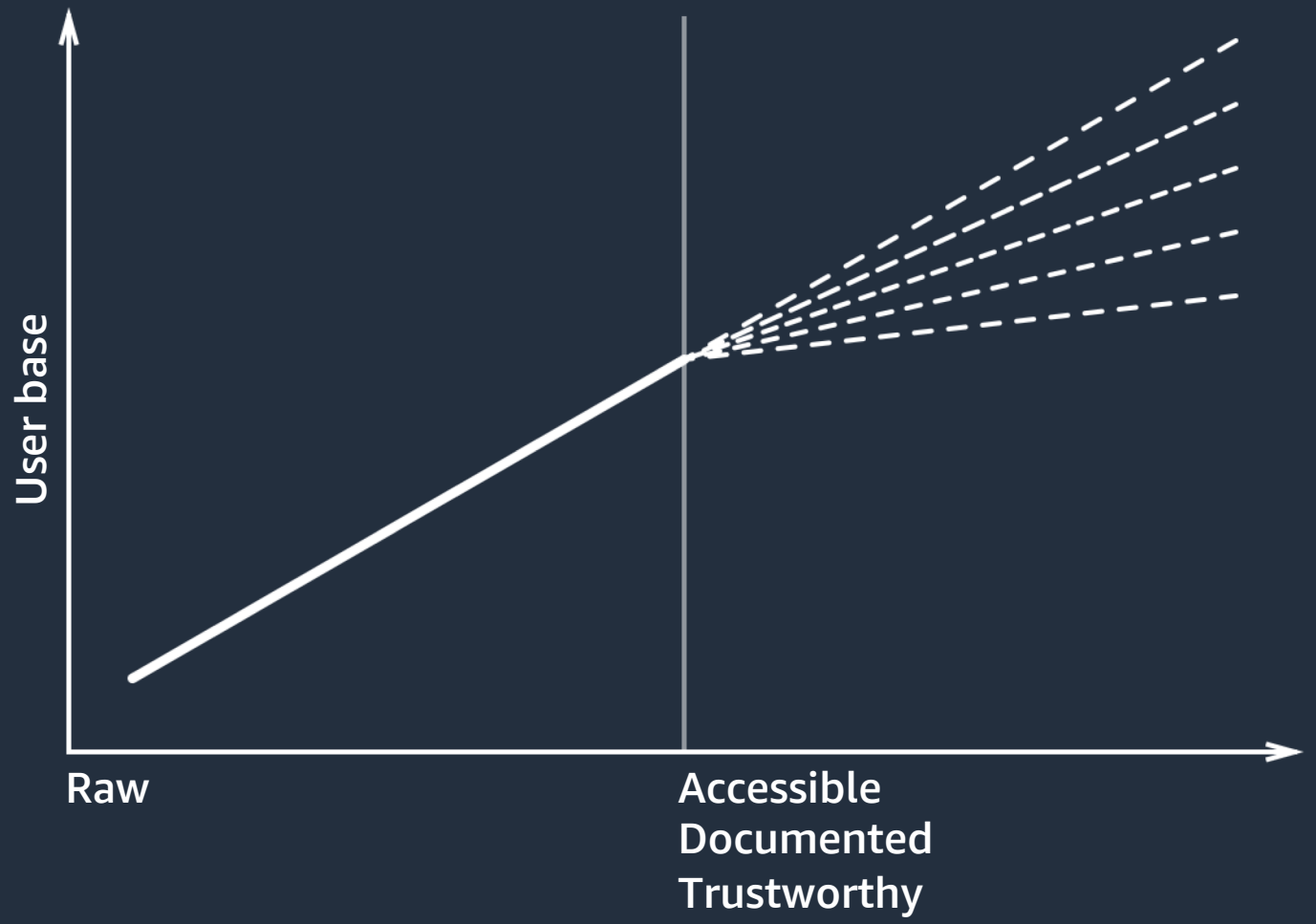
How can we maximize access to data?

https://registry.opendata.aws

https://aws.amazon.com/earth/research-credits

aws





ESIP Summer 2018 Meeting

Earth Science Information Partners

aws

Many users = many use cases and tools

ESIP Summer 2018 Meeting

The universe of meteorological data users is expanding to include:

- Economists
- Software developers (web and app developers)
- Young students
- Amateurs

These users have different:

- Skills
- Tools
- Needs

aws

Big is different

ESIP Summer 2018 Meeting

- Object storage is different than file storage
- Toolmakers must keep up with emerging formats
- A number of cloud-friendly formats are emerging
    - Cloud-optimized GeoTIFF (cogeo.org)
    - Zarr
    - NetCDF to Parquet/ORC

aws

Archive of Convenience

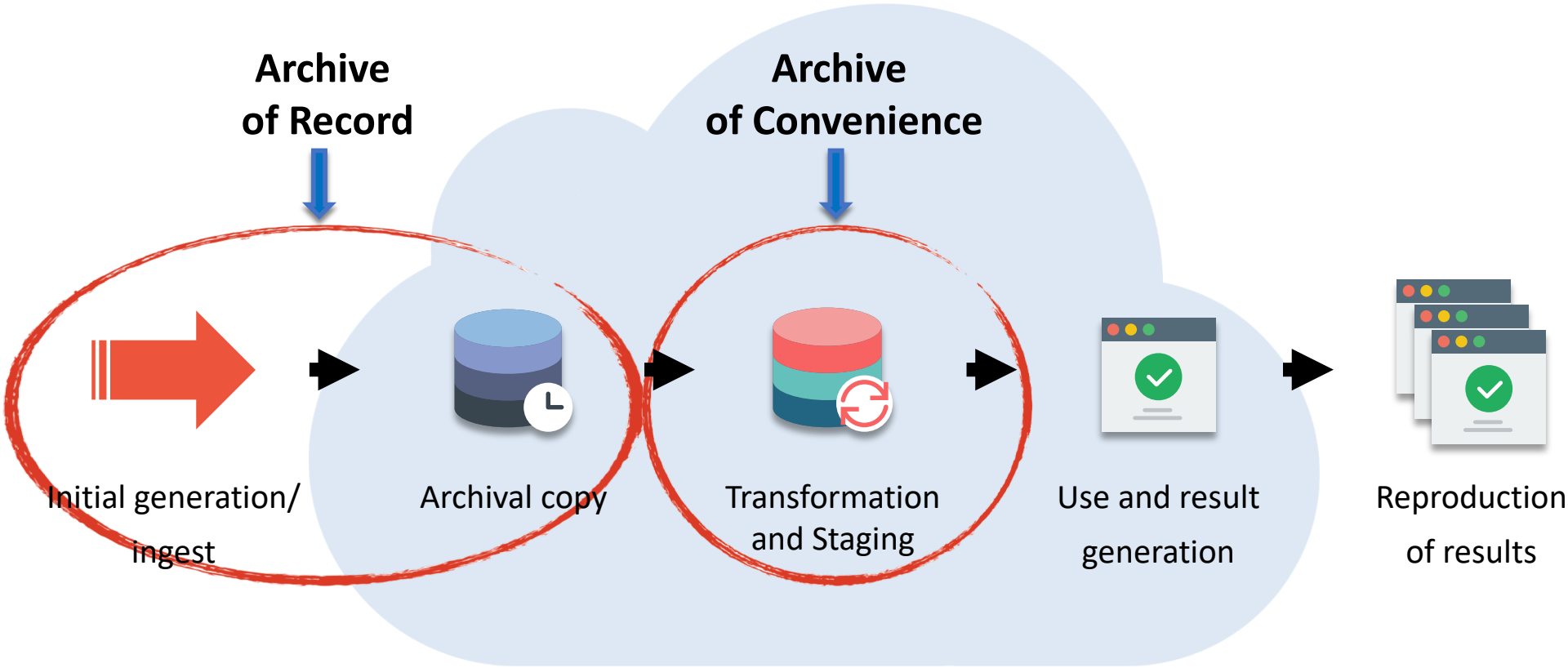
ESIP Summer 2018 Meeting

Element 84 presented an approach to managing complex datasets on the cloud called the Archive of Convenience.

- Remove the **undifferentiated heavy lifting** of data wrangling
- Exploit **ephemeral** nature of the cloud for both storage and services
- Created with **automated, repeatable, trustworthy** processes
- Strive for **compatibility** between datasets, formats, and tooling
- **Use case** Oriented

aws

Lifecycle of data in the cloud



Pangeo

ESIP Summer 2018 Meeting

The PanGeo community (pangeo.io) is experimenting with ways to share large volumes of NetCDF/HDF in the cloud.

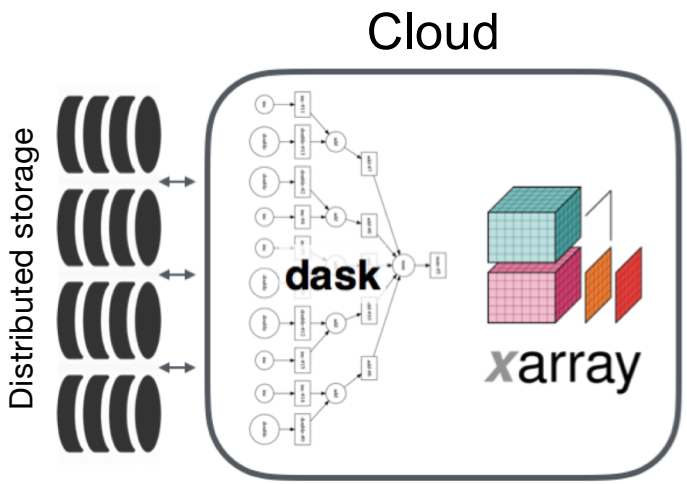
- https://bit.ly/big-data-non-portal
- Zarr
  - Simple format, clear specification
  - Data is chunked and can be accessed in parallel
  - lightweight global and variable metadata stored as JSON
  - Free, open-source software
  - Read/write using Xarray

aws

PANGEO

Pangeo Architecture

"Analysis Ready Data" stored on globally-available distributed storage.



Parallel computing system allows users deploy clusters of compute nodes for data processing.

Dask tells the nodes what to do.

Jupyter for interactive access on remote systems

end user

Xarray provides data structures and intuitive interface for interacting with datasets

Jed Sundwall
jed@amazon.com