Open Data on AWS

https://opendata.aws

Jed Sundwall, Global Open Data Lead

Agenda

Overview of Open Data on AWS

How shared data on the cloud can accelerate research

Finding data shared on AWS

Sharing data on AWS

Traditional Infrastructure

- Equipment
- Resources and Administration
- Contracts
- Cost

AWS Cloud

- No Up Front Expense Pay for what you Use
- Improve Time to Market & Agility
- Scale Up and Down
- Self-Service Infrastructure

Why does AWS care about open data?

Met Office

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

DigitalGlobe®

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

Sharing data in the cloud lets data users spend more time on data analysis rather than data acquisition.

https://opendata.aws

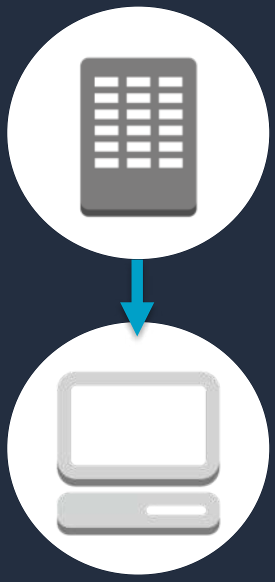
"…data must be organized, well-documented, consistently formatted, and error free. Cleaning the data is often the most taxing part of data science, and is frequently **80% of the work.**"

— Data Driven by DJ Patil and Hilary Mason

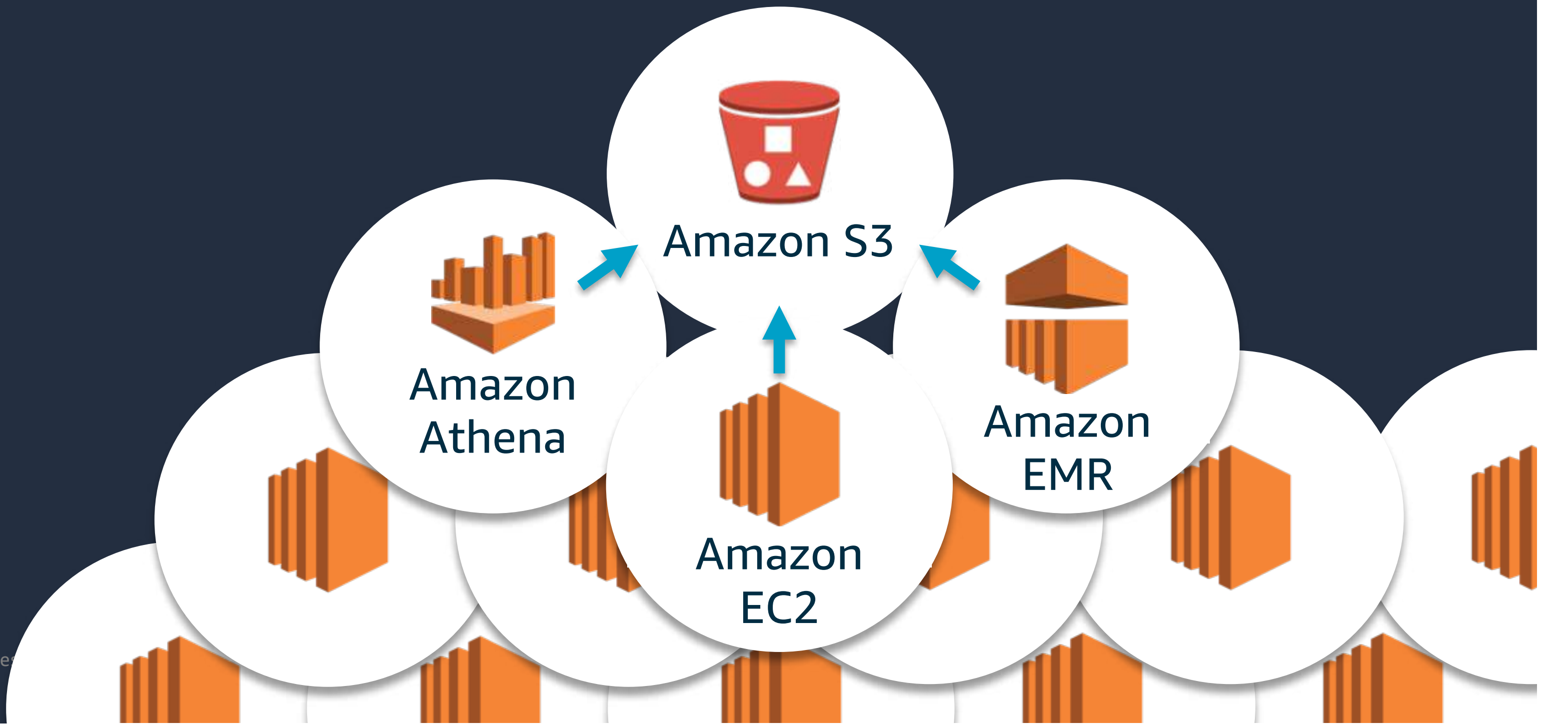
Undifferentiated heavy lifting

Flipped data flow in the cloud

Traditional approach:
Move the data to computing resources.



Cloud approach:
Move computing resources to the data.



Advantages of sharing data in the cloud

Global community of users

New services and tools

Faster pace of research

Lower cost of research

AWS Public Datasets

https://registry.opendata.aws

NASA

DWD

DEUTSCHE BÖRSE GROUP

FINNISH METEOROLOGICAL INSTITUTE

NYC Taxi & Limousine Commission

IRS

NOAA

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION U.S. DEPARTMENT OF COMMERCE

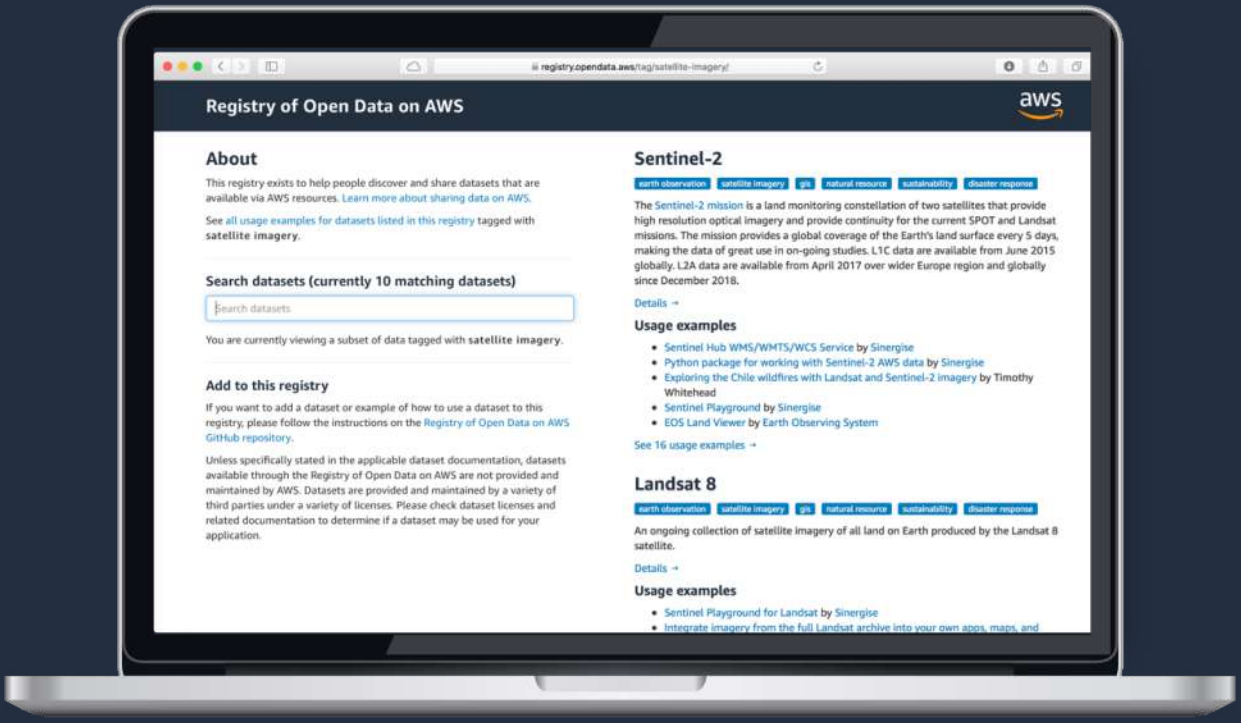
NIH National Institutes of Health

Met Office

CORNELL UNIVERSITY FOUNDED A.D. 1865

AWS Public Datasets

https://registry.opendata.aws/tag/satellite-imagery



Registry of Open Data on AWS

About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

See all usage examples for datasets listed in this registry tagged with **satellite imagery**.

Search datasets (currently 10 matching datasets)

Search datasets

You are currently viewing a subset of data tagged with **satellite imagery**.

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

Sentinel-2

The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

Details →

Usage examples

- Sentinel Hub WMS/WMTS/WCS Service by Sinergise
- Python package for working with Sentinel-2 AWS data by Sinergise
- Exploring the Chile wildfires with Landsat and Sentinel-2 imagery by Timothy Whitehead
- Sentinel Playground by Sinergise
- EOS Land Viewer by Earth Observing System

See 16 usage examples →

Landsat 8

An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.

Details →

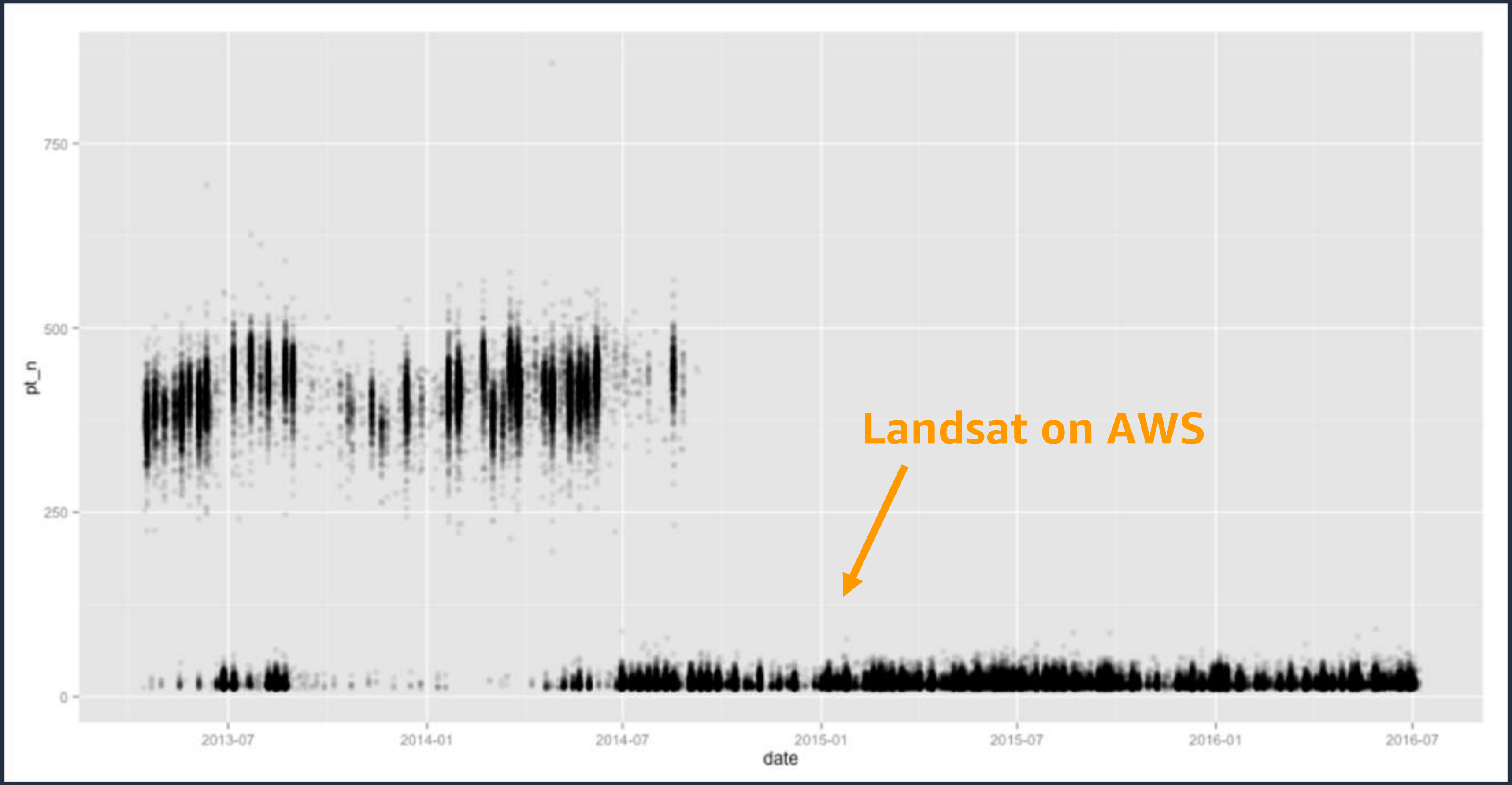
Usage examples

- Sentinel Playground for Landsat by Sinergise
- Integrate imagery from the full Landsat archive into your own apps, maps, and

"…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**."

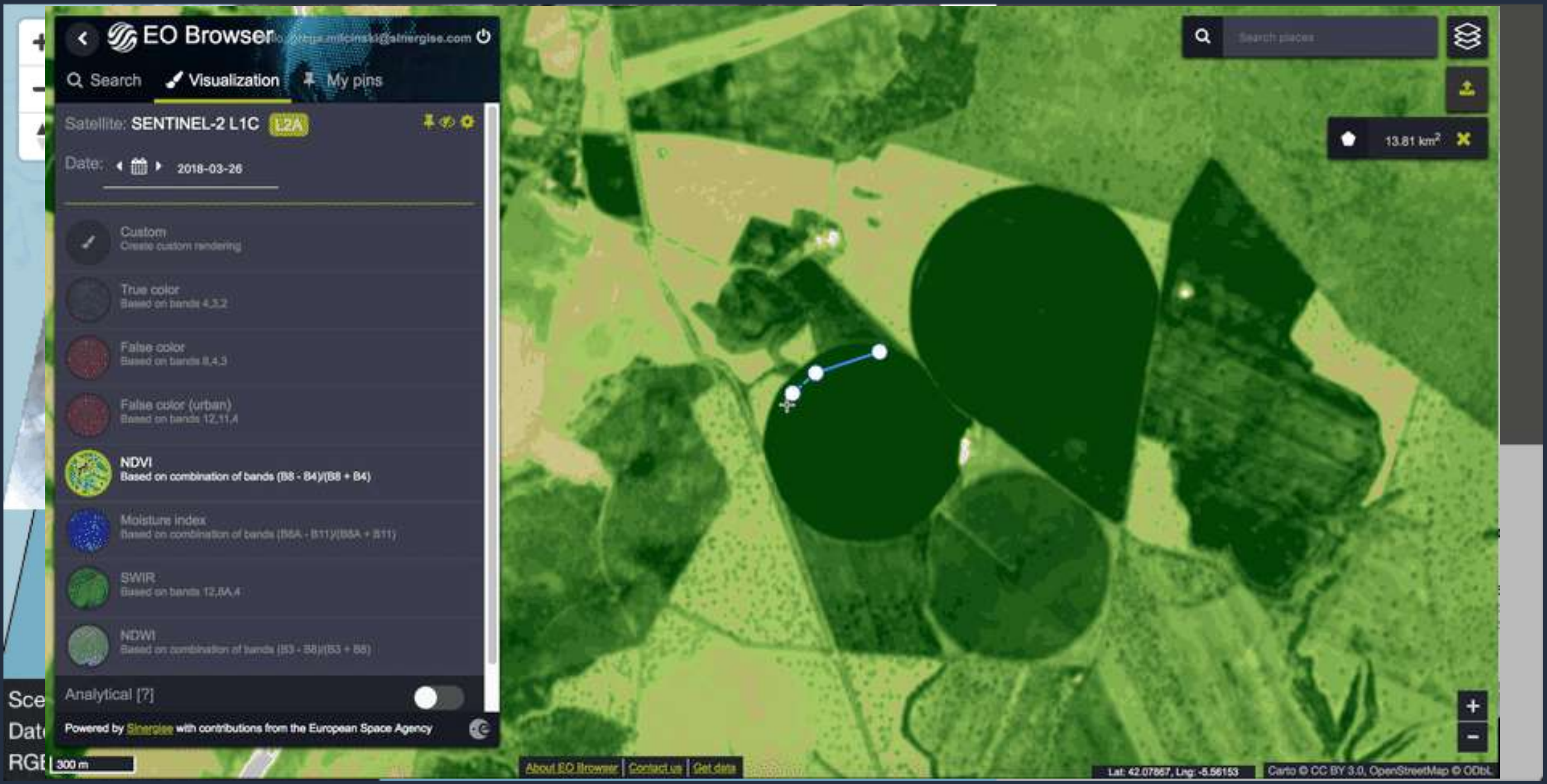
— Data Driven by DJ Patil and Hilary Mason

Undifferentiated heavy lifting



Graph by Drew Bollinger (@drewbo19) at Development Seed

Using Serverless to Visualize and Analyze Imagery



Satellite: SENTINEL-2 L1C L2A

Date: 2018-03-26

- Custom – Create custom rendering
- True color – Based on bands 4,3,2
- False color – Based on bands 8,4,3
- False color (urban) – Based on bands 12,11,4
- **NDVI – Based on combination of bands (B8 - B4)/(B8 + B4)**
- Moisture index – Based on combination of bands (B8A - B11)/(B8A + B11)
- SWIR – Based on bands 12,8A,4
- NDWI – Based on combination of bands (B3 - B8)/(B3 + B8)

Analytical [?]

Powered by Sinergise with contributions from the European Space Agency

Search places
13.81 km²
300 m
About EO Browser | Contact us | Get data
Lat: 42.07867, Lng: -8.56153
Carto © CC BY 3.0, OpenStreetMap © ODbL

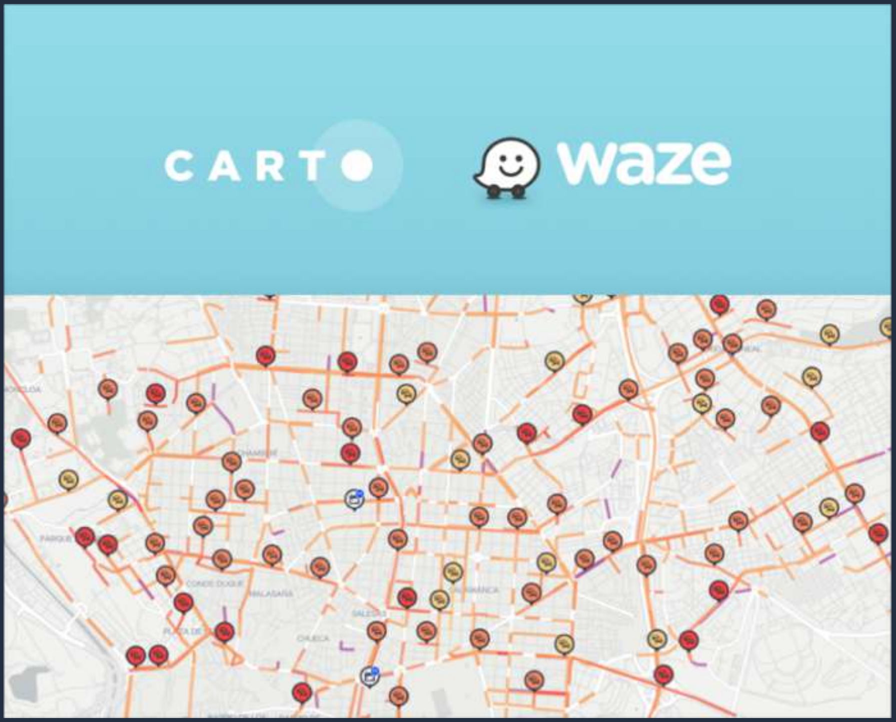
Blue Dot Observatory

https://blue-dot-observatory.com/

“By having the complete Sentinel-2 archive available on AWS and access to on-demand scalable compute resources, the workflow is simple to implement. The cost to process one month of data for about 7,000 bodies of water currently in the system is 6 EUR.”

City of Louisville WARP

“The Waze Analytics Relational-database Platform (WARP) now gives 900+ government partners access to a free platform that uses real-time and historic traffic data to improve mobility, pedestrian and bike safety, road conditions, and emergency response.”

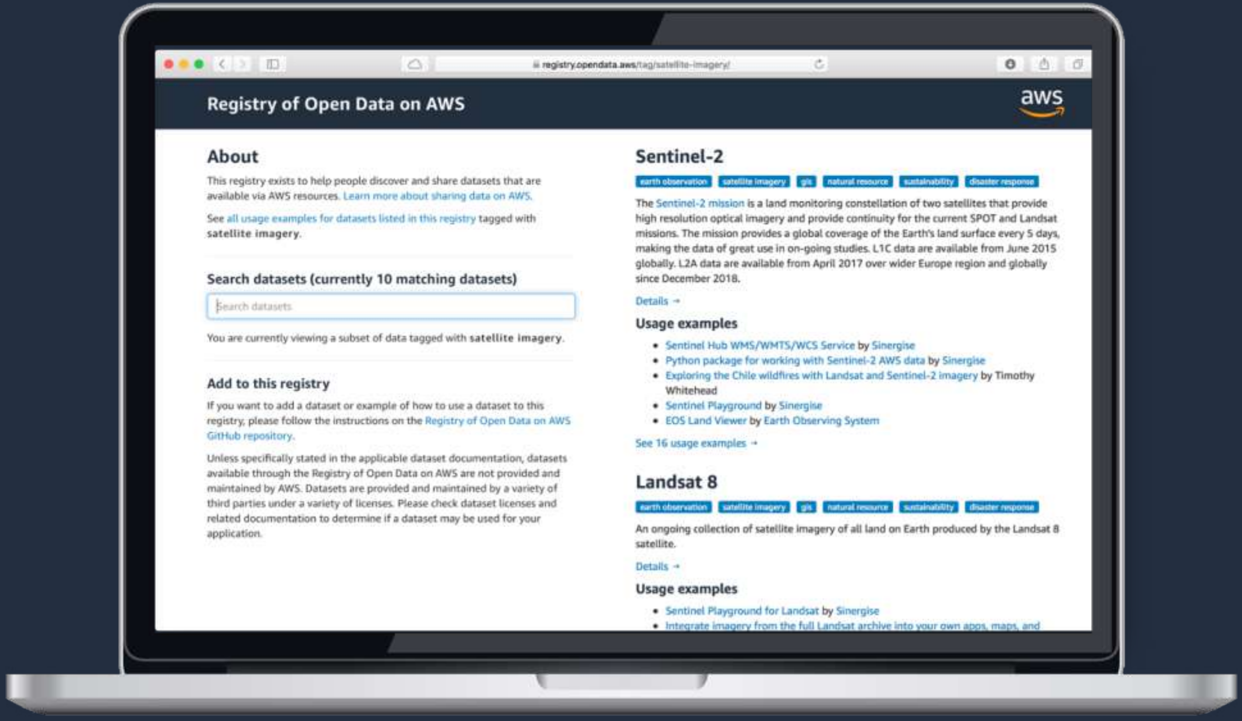


Finding data on AWS

Using the Registry of Open Data on AWS (RODA)

Registry of Open Data on AWS

https://registry.opendata.aws/



Registry of Open Data on AWS

About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

See all usage examples for datasets listed in this registry tagged with **satellite imagery**.

Search datasets (currently 10 matching datasets)

Search datasets

You are currently viewing a subset of data tagged with **satellite imagery**.

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

Sentinel-2

earth observation · satellite imagery · gis · natural resource · sustainability · disaster response

The Sentinel-2 mission is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

Details →

Usage examples

- Sentinel Hub WMS/WMTS/WCS Service by Sinergise
- Python package for working with Sentinel-2 AWS data by Sinergise
- Exploring the Chile wildfires with Landsat and Sentinel-2 imagery by Timothy Whitehead
- Sentinel Playground by Sinergise
- EOS Land Viewer by Earth Observing System

See 16 usage examples →

Landsat 8

earth observation · satellite imagery · gis · natural resource · sustainability · disaster response

An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.

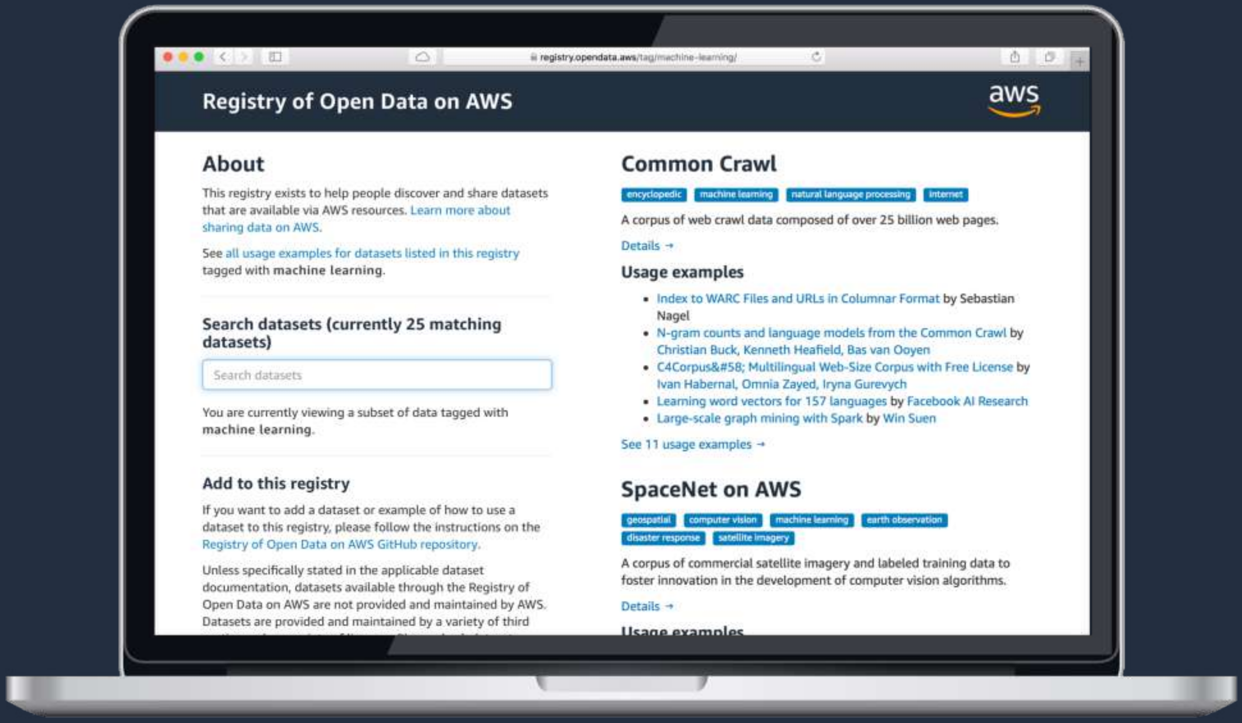
Details →

Usage examples

- Sentinel Playground for Landsat by Sinergise
- Integrate imagery from the full Landsat archive into your own apps, maps, and

Registry of Open Data on AWS – Tags

https://registry.opendata.aws/tag/machine-learning



Registry of Open Data on AWS

About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

See all usage examples for datasets listed in this registry tagged with **machine learning**.

Search datasets (currently 25 matching datasets)

You are currently viewing a subset of data tagged with **machine learning**.

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third

Common Crawl

A corpus of web crawl data composed of over 25 billion web pages.

Details →

Usage examples

- Index to WARC Files and URLs in Columnar Format by Sebastian Nagel
- N-gram counts and language models from the Common Crawl by Christian Buck, Kenneth Heafield, Bas van Ooyen
- C4Corpus&#58; Multilingual Web-Size Corpus with Free License by Ivan Habernal, Omnia Zayed, Iryna Gurevych
- Learning word vectors for 157 languages by Facebook AI Research
- Large-scale graph mining with Spark by Win Suen

See 11 usage examples →

SpaceNet on AWS

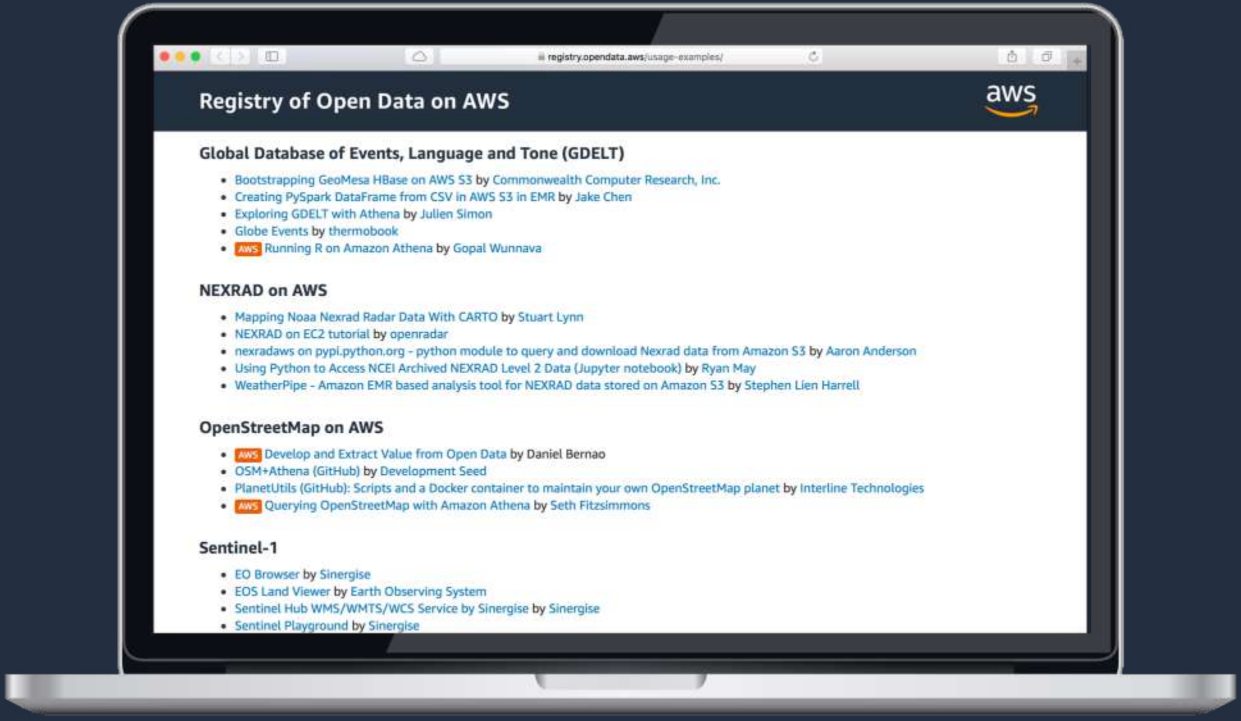
A corpus of commercial satellite imagery and labeled training data to foster innovation in the development of computer vision algorithms.

Details →

Usage examples

Registry of Open Data on AWS – Usage examples

https://registry.opendata.aws/usage-examples



registry.opendata.aws/usage-examples/

Registry of Open Data on AWS

Global Database of Events, Language and Tone (GDELT)

- Bootstrapping GeoMesa HBase on AWS S3 by Commonwealth Computer Research, Inc.
- Creating PySpark DataFrame from CSV in AWS S3 in EMR by Jake Chen
- Exploring GDELT with Athena by Julien Simon
- Globe Events by thermobook
- AWS Running R on Amazon Athena by Gopal Wunnava

NEXRAD on AWS

- Mapping Noaa Nexrad Radar Data With CARTO by Stuart Lynn
- NEXRAD on EC2 tutorial by openradar
- nexradaws on pypi.python.org - python module to query and download Nexrad data from Amazon S3 by Aaron Anderson
- Using Python to Access NCEI Archived NEXRAD Level 2 Data (Jupyter notebook) by Ryan May
- WeatherPipe - Amazon EMR based analysis tool for NEXRAD data stored on Amazon S3 by Stephen Lien Harrell

OpenStreetMap on AWS

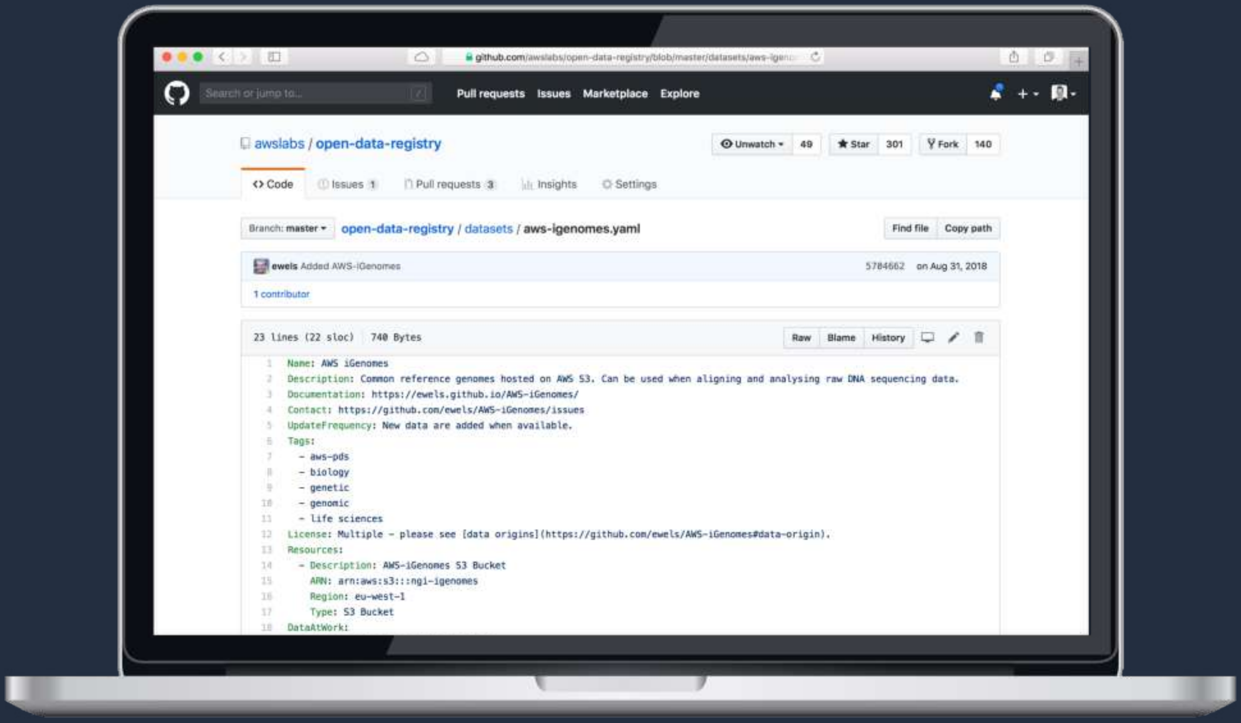
- AWS Develop and Extract Value from Open Data by Daniel Bernao
- OSM+Athena (GitHub) by Development Seed
- PlanetUtils (GitHub): Scripts and a Docker container to maintain your own OpenStreetMap planet by Interline Technologies
- AWS Querying OpenStreetMap with Amazon Athena by Seth Fitzsimmons

Sentinel-1

- EO Browser by Sinergise
- EOS Land Viewer by Earth Observing System
- Sentinel Hub WMS/WMTS/WCS Service by Sinergise by Sinergise
- Sentinel Playground by Sinergise

Registry of Open Data on AWS – How to contribute

https://github.com/awslabs/open-data-registry



open-data-registry / datasets / aws-igenomes.yaml

ewels Added AWS-iGenomes 5784662 on Aug 31, 2018

1 contributor

23 lines (22 sloc) 740 Bytes

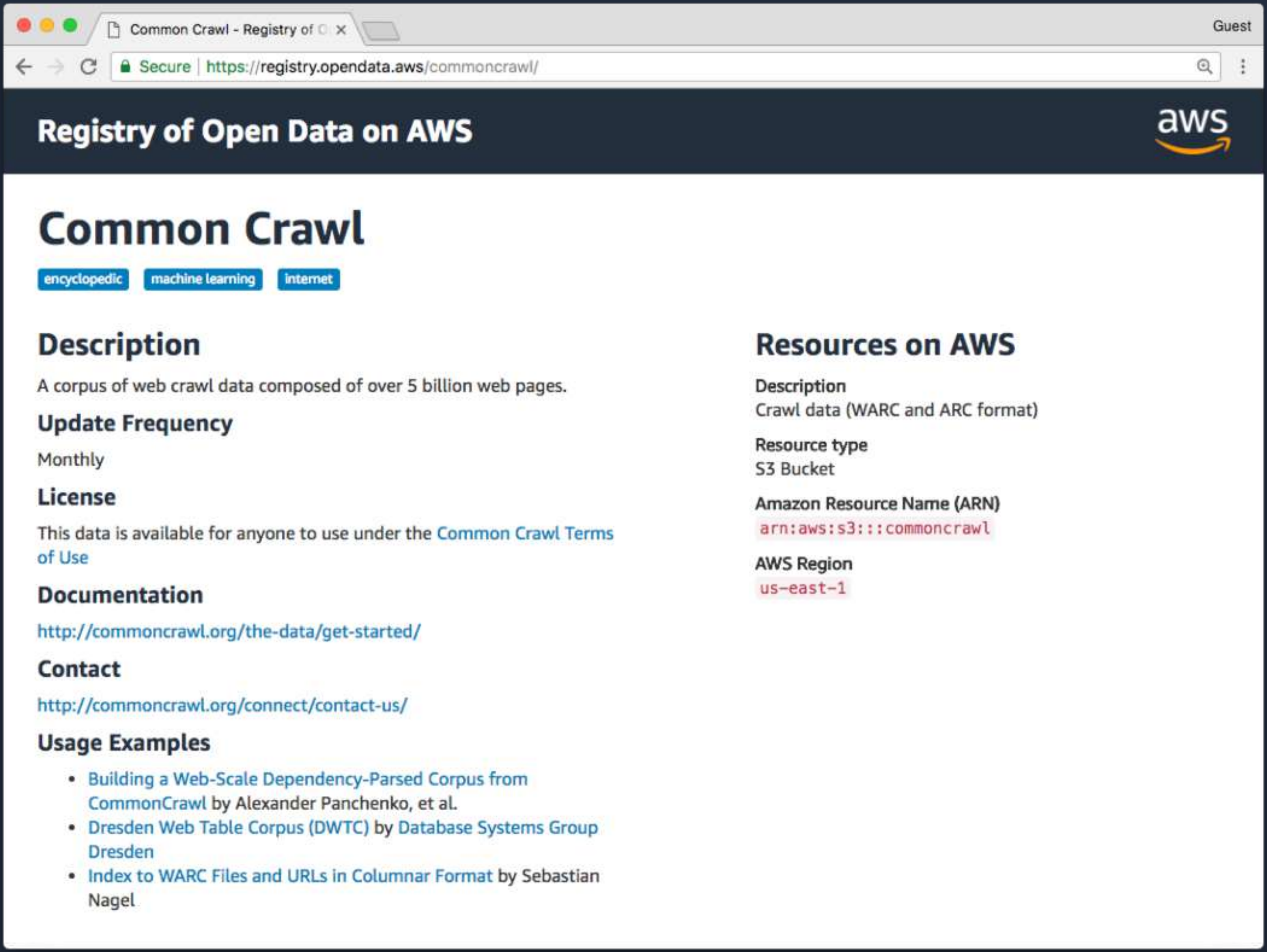
```
Name: AWS iGenomes
Description: Common reference genomes hosted on AWS S3. Can be used when aligning and analysing raw DNA sequencing data.
Documentation: https://ewels.github.io/AWS-iGenomes/
Contact: https://github.com/ewels/AWS-iGenomes/issues
UpdateFrequency: New data are added when available.
Tags:
  - aws-pds
  - biology
  - genetic
  - genomic
  - life sciences
License: Multiple - please see [data origins](https://github.com/ewels/AWS-iGenomes#data-origin).
Resources:
  - Description: AWS-iGenomes S3 Bucket
    ARN: arn:aws:s3:::ngi-igenomes
    Region: eu-west-1
    Type: S3 Bucket
DataAtWork:
```

Sharing data (on AWS)

What we’ve learned

What makes a dataset successful?

It is treated like a product.


Common Crawl - Registry of O ×
Guest
Secure | https://registry.opendata.aws/commoncrawl/

Registry of Open Data on AWS

Common Crawl

encyclopedic machine learning internet

Description

A corpus of web crawl data composed of over 5 billion web pages.

Update Frequency

Monthly

License

This data is available for anyone to use under the Common Crawl Terms of Use

Documentation

http://commoncrawl.org/the-data/get-started/

Contact

http://commoncrawl.org/connect/contact-us/

Usage Examples

- Building a Web-Scale Dependency-Parsed Corpus from CommonCrawl by Alexander Panchenko, et al.
- Dresden Web Table Corpus (DWTC) by Database Systems Group Dresden
- Index to WARC Files and URLs in Columnar Format by Sebastian Nagel

Resources on AWS

Description
Crawl data (WARC and ARC format)

Resource type
S3 Bucket

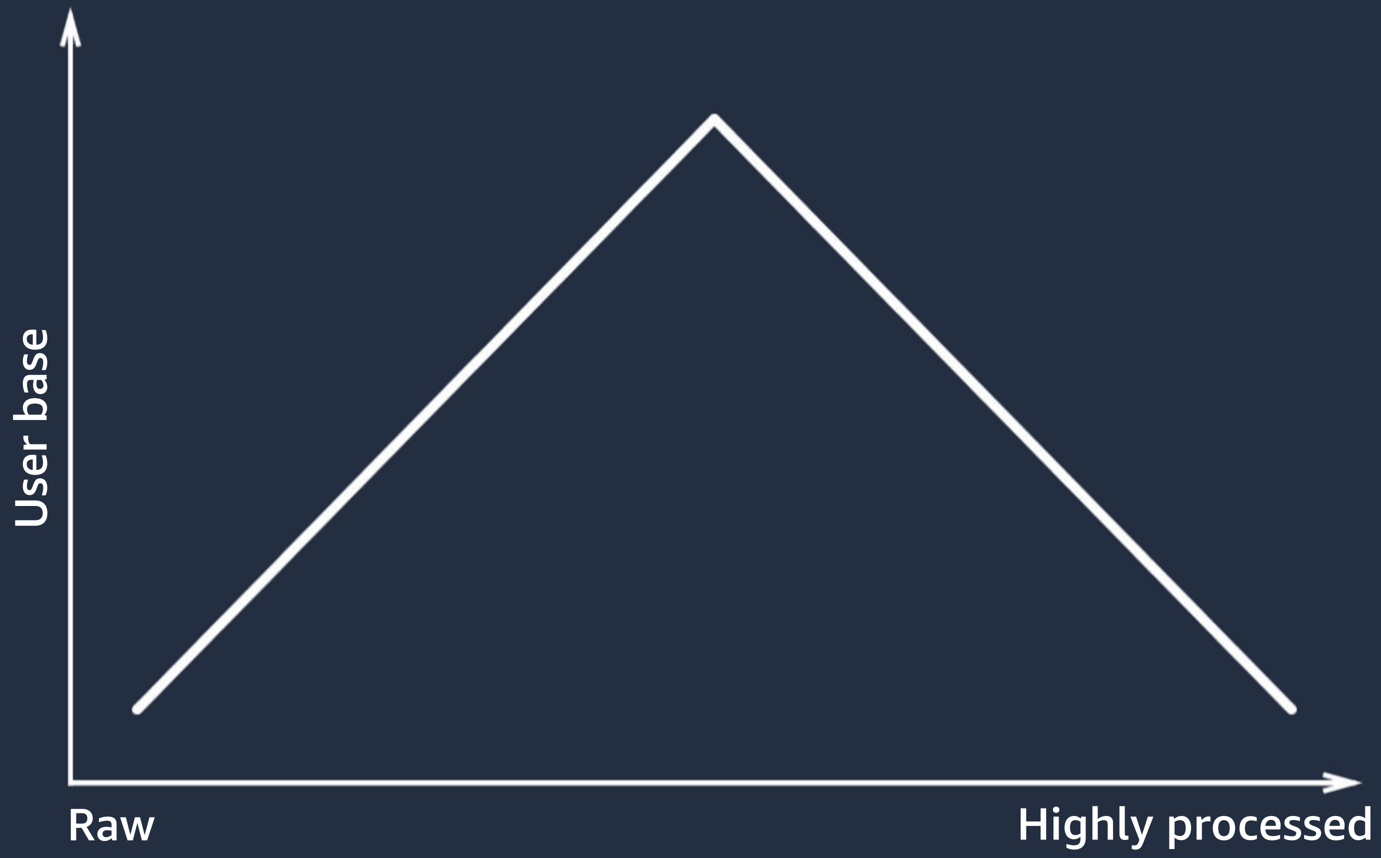
Amazon Resource Name (ARN)
`arn:aws:s3:::commoncrawl`

AWS Region
`us-east-1`

What makes a dataset successful?

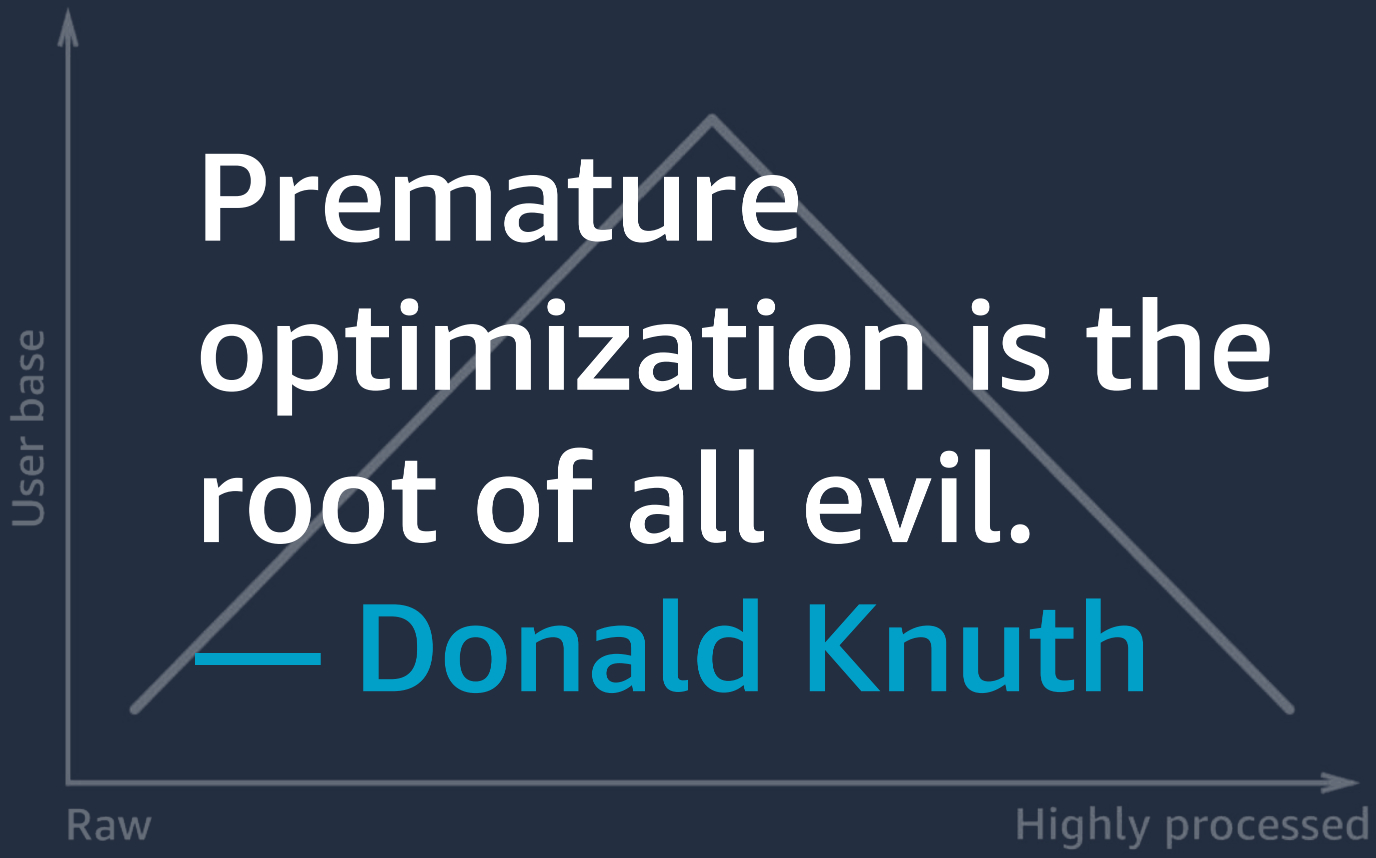
It is treated like a product.

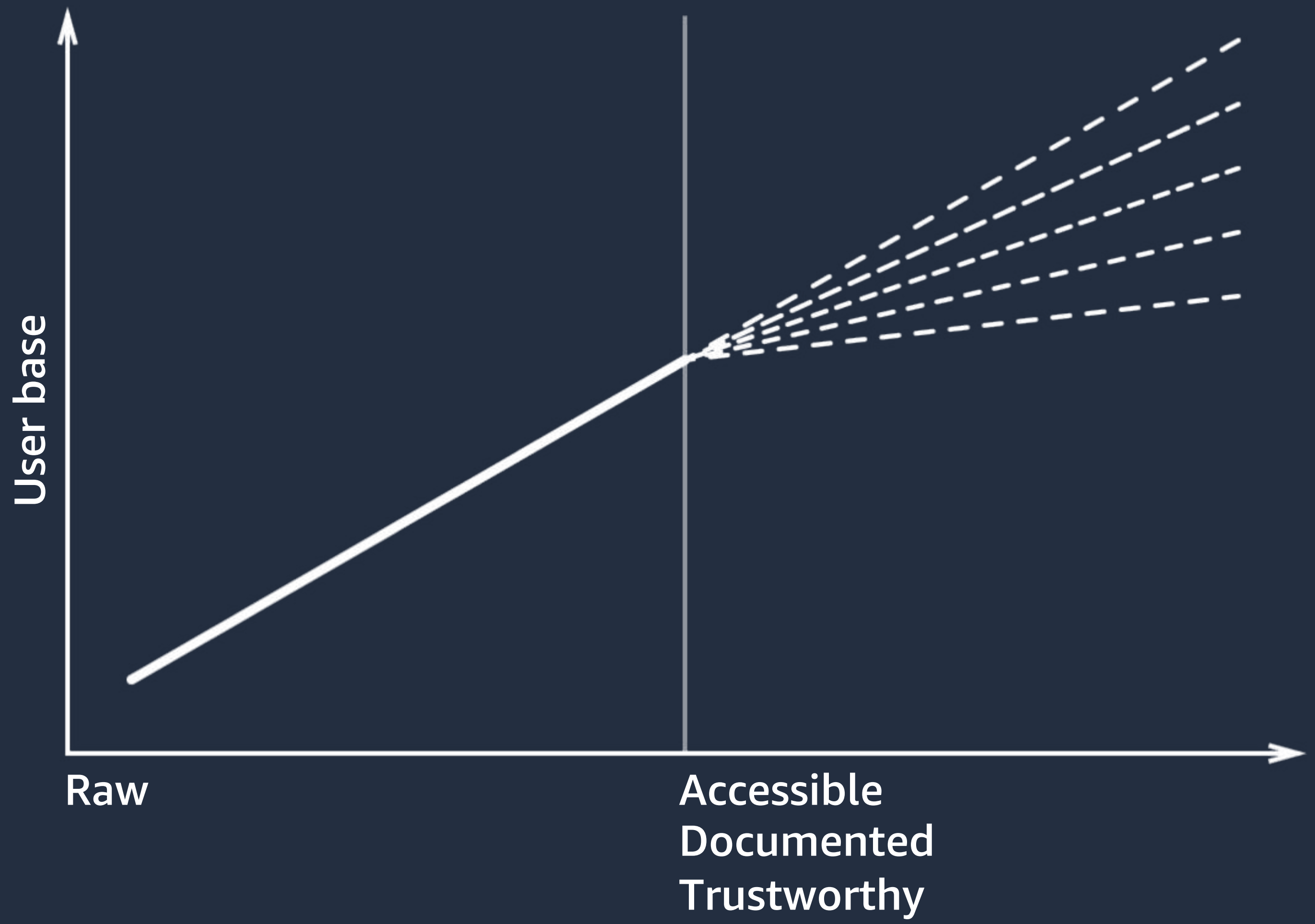
It is optimized for analysis.


User base
Raw
Highly processed

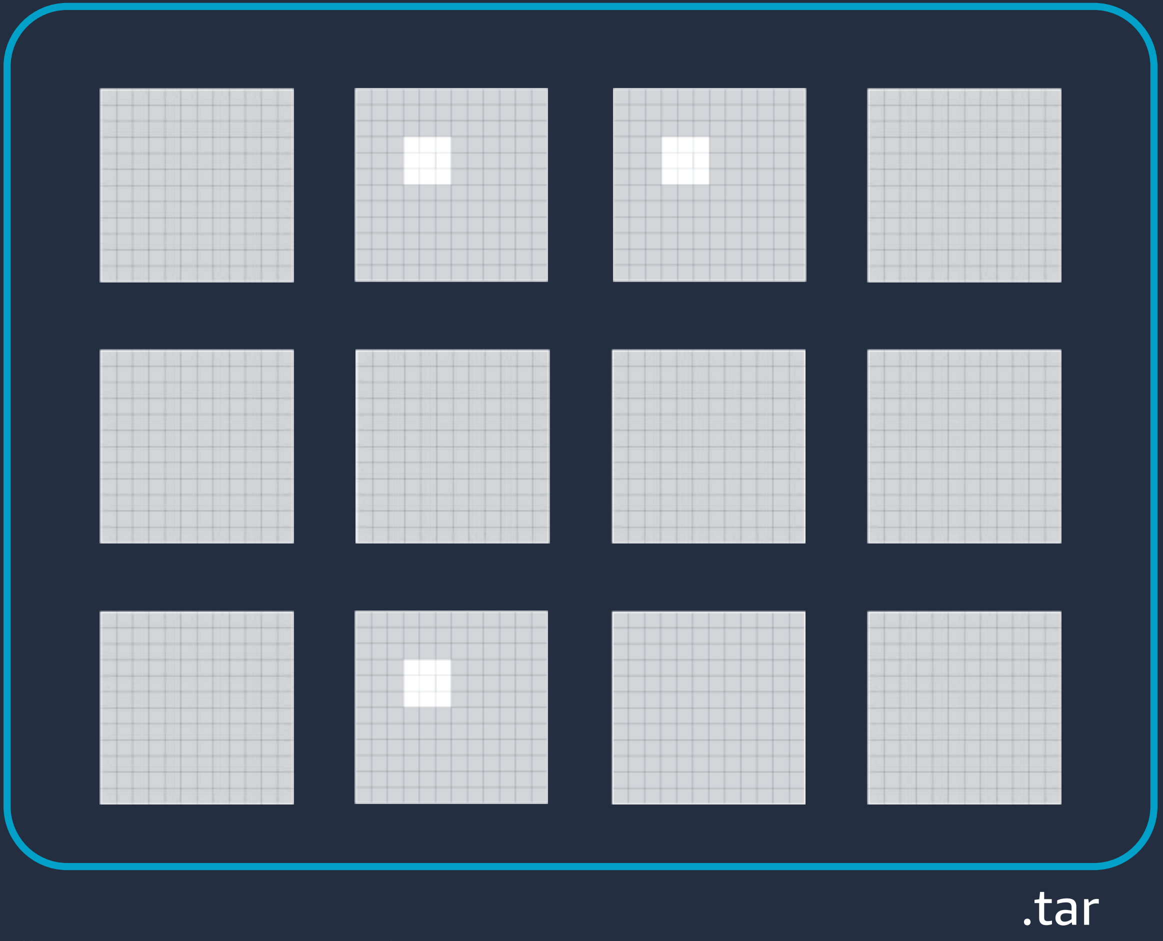
Premature optimization is the root of all evil.

— Donald Knuth

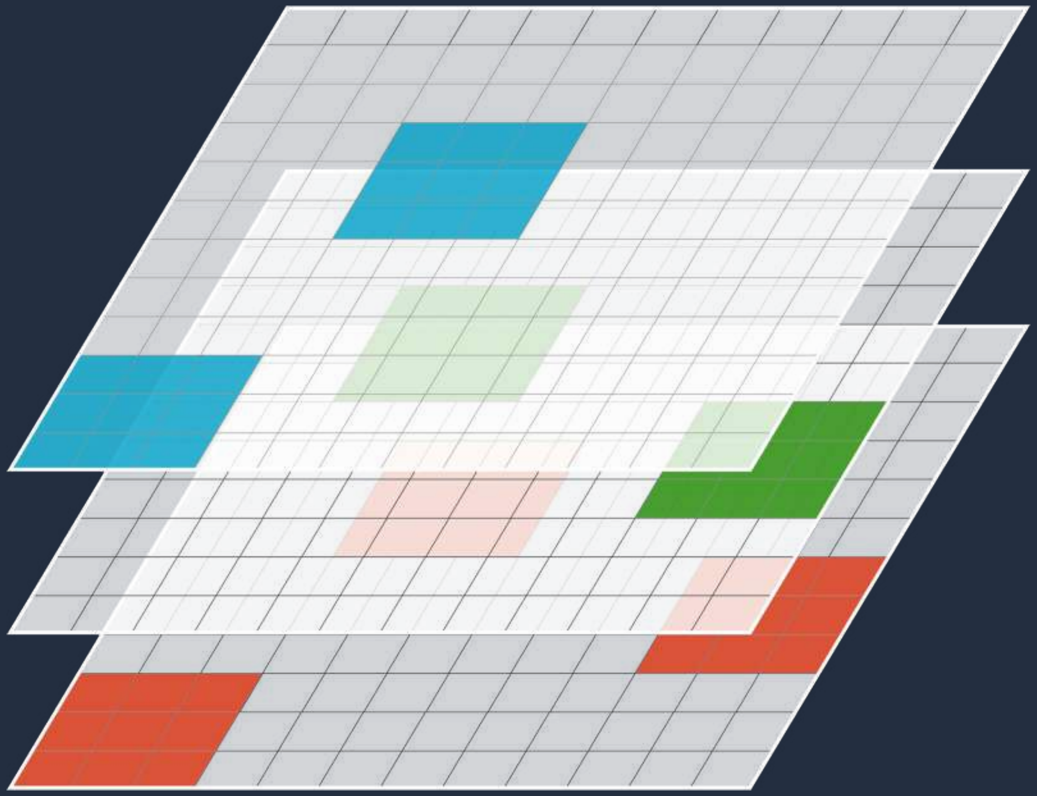



User base
Raw
Accessible
Documented
Trustworthy

The cloud-optimized GeoTIFF

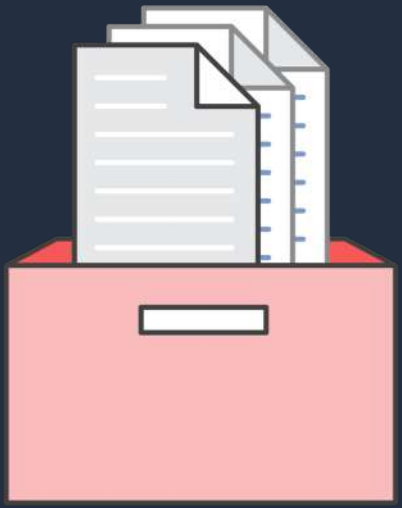


The cloud-optimized GeoTIFF



Patterns

S3 Key Index

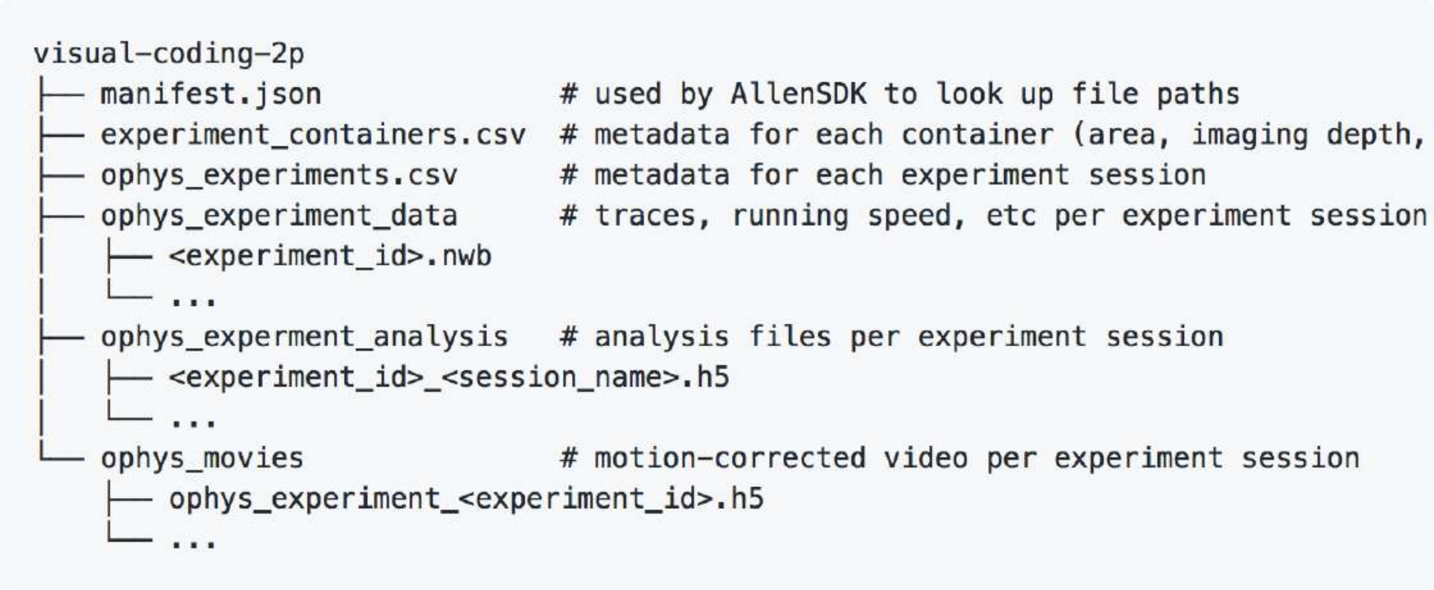


External Index

Internal Index

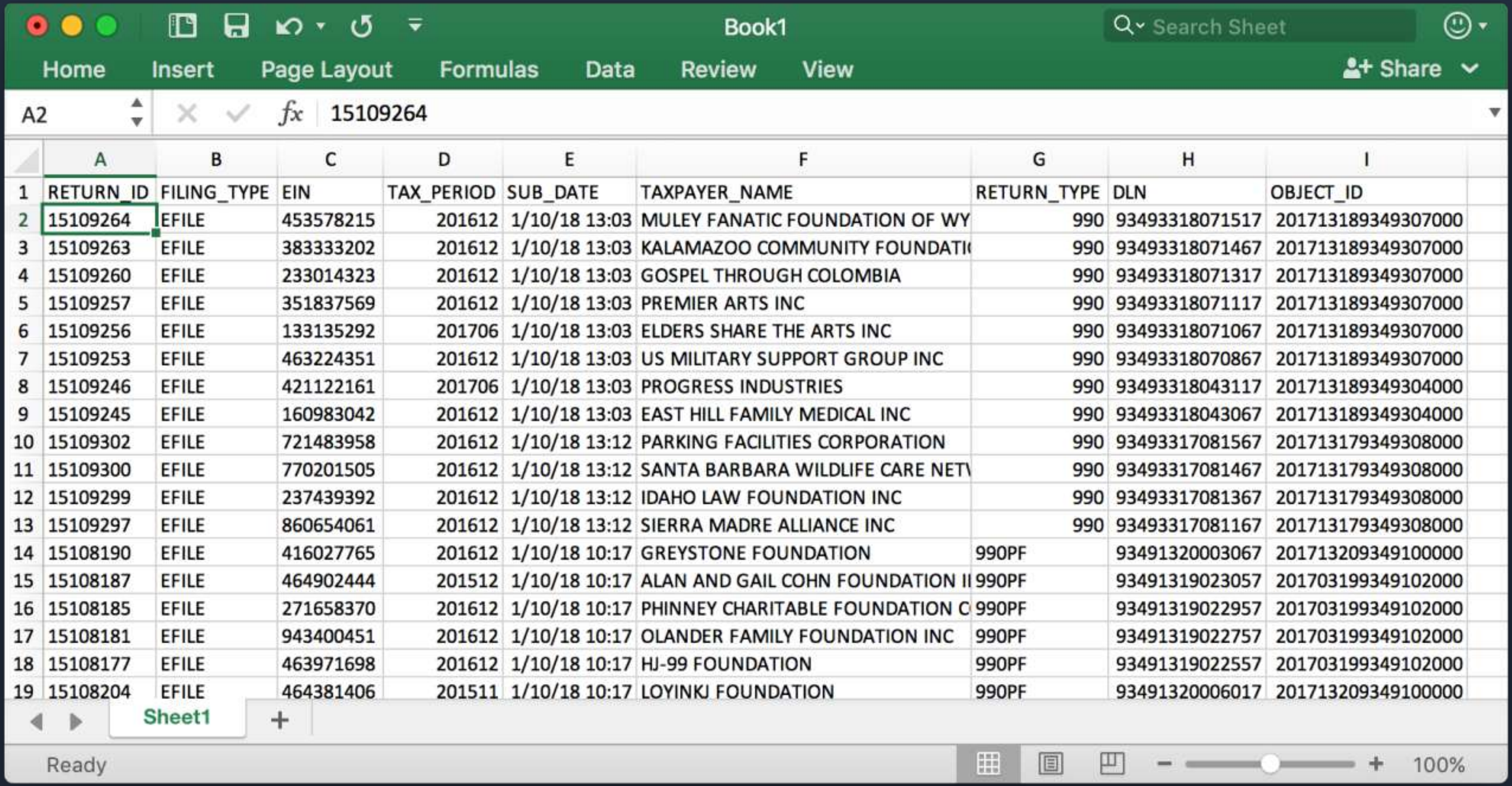
Example: Allen Brain Observatory Key Naming

https://registry.opendata.aws/allen-brain-observatory/



```
visual-coding-2p
├── manifest.json              # used by AllenSDK to look up file paths
├── experiment_containers.csv  # metadata for each container (area, imaging depth,
├── ophys_experiments.csv      # metadata for each experiment session
├── ophys_experiment_data      # traces, running speed, etc per experiment session
│   ├── <experiment_id>.nwb
│   └── ...
├── ophys_experment_analysis   # analysis files per experiment session
│   ├── <experiment_id>_<session_name>.h5
│   └── ...
└── ophys_movies               # motion-corrected video per experiment session
    ├── ophys_experiment_<experiment_id>.h5
    └── ...
```

Example: IRS 990 CSV as External Index



A2 *fx* 15109264

	A	B	C	D	E	F	G	H	I
1	RETURN_ID	FILING_TYPE	EIN	TAX_PERIOD	SUB_DATE	TAXPAYER_NAME	RETURN_TYPE	DLN	OBJECT_ID
2	15109264	EFILE	453578215	201612	1/10/18 13:03	MULEY FANATIC FOUNDATION OF WY	990	93493318071517	201713189349307000
3	15109263	EFILE	383333202	201612	1/10/18 13:03	KALAMAZOO COMMUNITY FOUNDATI(	990	93493318071467	201713189349307000
4	15109260	EFILE	233014323	201612	1/10/18 13:03	GOSPEL THROUGH COLOMBIA	990	93493318071317	201713189349307000
5	15109257	EFILE	351837569	201612	1/10/18 13:03	PREMIER ARTS INC	990	93493318071117	201713189349307000
6	15109256	EFILE	133135292	201706	1/10/18 13:03	ELDERS SHARE THE ARTS INC	990	93493318071067	201713189349307000
7	15109253	EFILE	463224351	201612	1/10/18 13:03	US MILITARY SUPPORT GROUP INC	990	93493318070867	201713189349307000
8	15109246	EFILE	421122161	201706	1/10/18 13:03	PROGRESS INDUSTRIES	990	93493318043117	201713189349304000
9	15109245	EFILE	160983042	201612	1/10/18 13:03	EAST HILL FAMILY MEDICAL INC	990	93493318043067	201713189349304000
10	15109302	EFILE	721483958	201612	1/10/18 13:12	PARKING FACILITIES CORPORATION	990	93493317081567	201713179349308000
11	15109300	EFILE	770201505	201612	1/10/18 13:12	SANTA BARBARA WILDLIFE CARE NET\	990	93493317081467	201713179349308000
12	15109299	EFILE	237439392	201612	1/10/18 13:12	IDAHO LAW FOUNDATION INC	990	93493317081367	201713179349308000
13	15109297	EFILE	860654061	201612	1/10/18 13:12	SIERRA MADRE ALLIANCE INC	990	93493317081167	201713179349308000
14	15108190	EFILE	416027765	201612	1/10/18 10:17	GREYSTONE FOUNDATION	990PF	93491320003067	201713209349100000
15	15108187	EFILE	464902444	201512	1/10/18 10:17	ALAN AND GAIL COHN FOUNDATION II	990PF	93491319023057	201703199349102000
16	15108185	EFILE	271658370	201612	1/10/18 10:17	PHINNEY CHARITABLE FOUNDATION C	990PF	93491319022957	201703199349102000
17	15108181	EFILE	943400451	201612	1/10/18 10:17	OLANDER FAMILY FOUNDATION INC	990PF	93491319022757	201703199349102000
18	15108177	EFILE	463971698	201612	1/10/18 10:17	HJ-99 FOUNDATION	990PF	93491319022557	201703199349102000
19	15108204	EFILE	464381406	201511	1/10/18 10:17	LOYINKJ FOUNDATION	990PF	93491320006017	201713209349100000

What makes a dataset successful?

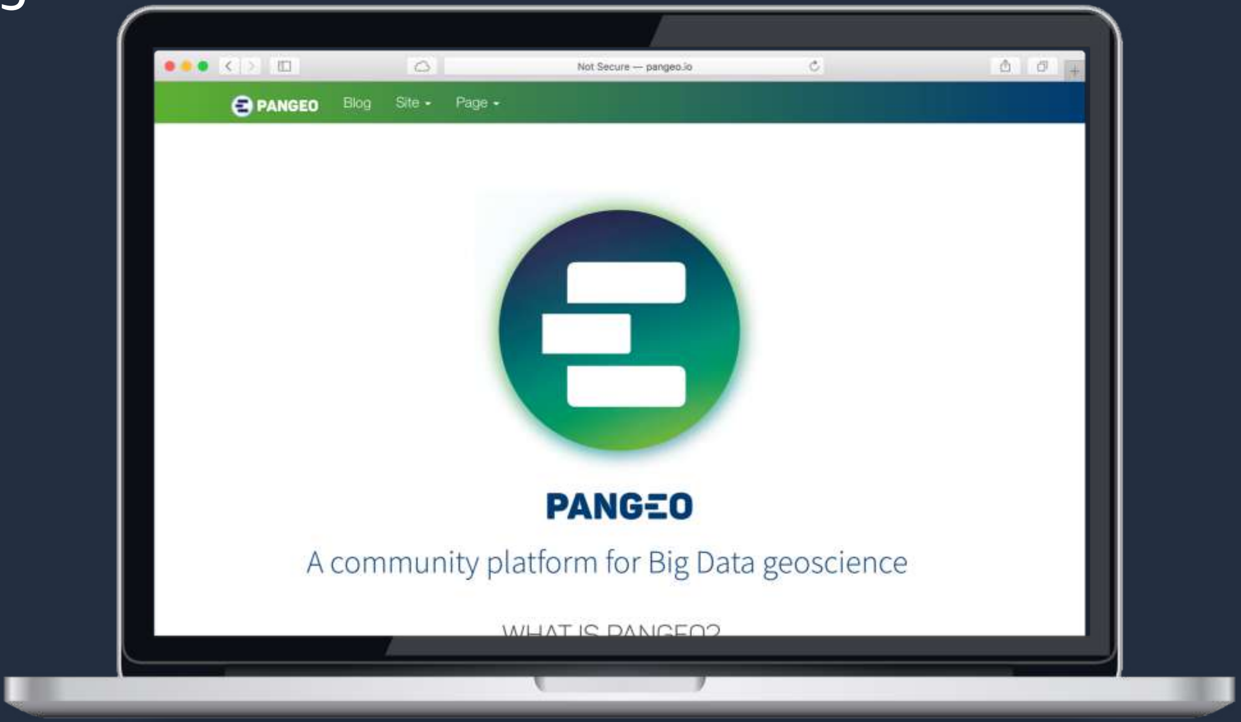
It is treated like a product.

It is optimized for analysis.

There is a community around it.

Pangeo community platform

http://pangeo.io



PANGEO

A community platform for Big Data geoscience

WHAT IS PANGEO?

Thank you!

jed@amazon.com