Introduction to Big Data Sources

- from the perspective of geospatial & social data

Dr. Yuanzheng Shao

Geo-computation Center for Social Sciences, Wuhan University

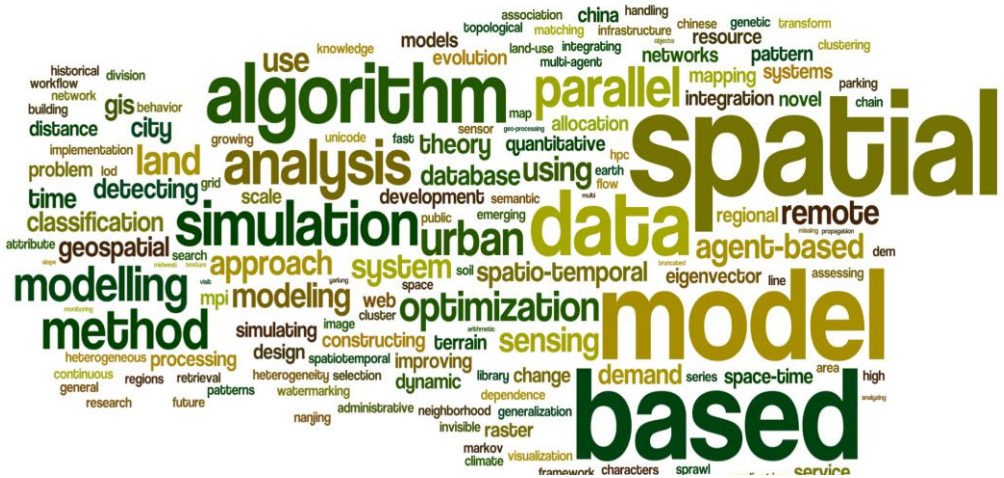
Outline

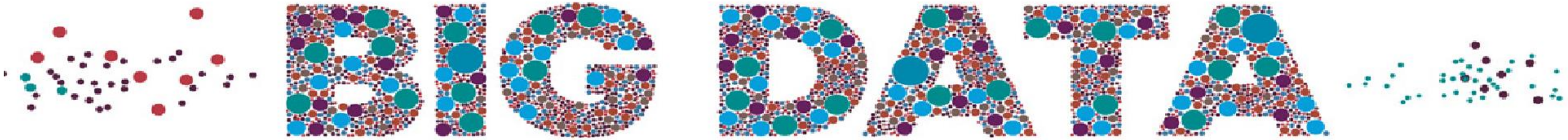
1. Background
2. Spatial Big Data
   - Remote Sensing Data
   - Major Earth Observation Satellites & Data Products
   - Chinese Satellites - Luojia-1
3. Social Big Data
   - Mobility Data
   - Airline Flight Data
   - High-speed Train Data
   - Social Media Data
4. CEOS WGISS CWIC

1. Background - Social Computing & Geo-computation

Social computing

Geo-Computation





A Brief Introduction About GCSS

Geocomputation Center for Social Sciences (GCSS):

- Established on Jan. 2018
- Hosted by the State Key Lab of Survey and Remote Sensing Information Engineering at Wuhan University.
- Co-sponsored by Wuhan University, the University of Chicago and China Data Institute
- Chair: Jianya Gong
- Co-Chairs: Luc Anselin and Shuming Bao

社会地理计算联合研究中心
Geocomputation Center for Social Science

THE UNIVERSITY OF CHICAGO

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

LIESMARS

CDL

The Objectives of GCSS

- Promote research on new theory and methodology
- Develop core technology for spatial data analysis and intelligent Information service
- Provide an open platform for international collaboration

Resources for COVID-19 & Global Research

Resources for COVID-19 | China

projects.iq.harvard.edu/chinadatalab/resources-covid-19

HARVARD UNIVERSITY HARVARD.EDU

China Data Lab

HOME People Resources Partners Events About

RESOURCES

- Enter the Lab
- Data Case Studies
  - Resources for COVID-19
- Publications
- Presentations

Resources for COVID-19

https://dataverse.harvard.edu/dataverse/cdl_dataverse

This project aims to provide an information infrastructure for the spatial study of the new novel coronavirus (COVID-19), which was first detected in Wuhan City, Hubei Province, China and spread to other parts of China and dozens of countries and regions in the world. The objectives of this project are:

- To provide data support for the spatial study of COVID-19 at local, regional and global levels with information collected and integrated from different sources.
- To facilitate quantitative research on spatial spreading and impacts of COVID-19 with advanced methodology and technology.
- To promote collaborative research on the spatial study of COVID-19 on the China Data Lab, Dataverse and WorldMap platforms.
- To build research capacity for future collaborative projects.

This project is a joint effort by scholars and professionals from the Center for Geographical Analysis at Harvard University, the Geo-Computation Center for Social Sciences at Wuhan University, the China Data Institute, and the RMDS Lab. The final data will be deployed to dataverse.harvard.edu for research sharing.

Contact: spatialdatalab@list.fas.harvard.edu

Resources for COVID-19

dataverse.harvard.edu/dataverse/2019ncov

HARVARD Dataverse

Add Data Search About User Guide Support Sign Up Log In

Resources for COVID-19 (China Data Lab)

Harvard Dataverse > China Data Lab Dataverse > **Resources for COVID-19**

Contact Share

Data Development Code News Report Research Papers

Search this dataverse... Find Advanced Search

Dataverses (6)
Datasets (0)
Files (0)

Dataverse Category
Research Group (6)

Publication Year
2020 (6)

1 to 6 of 6 Results Sort

Data (China Data Lab) 2020-2-11

Research Papers (China Data Lab) 2020-2-11

Workflows (China Data Lab) 2020-2-11

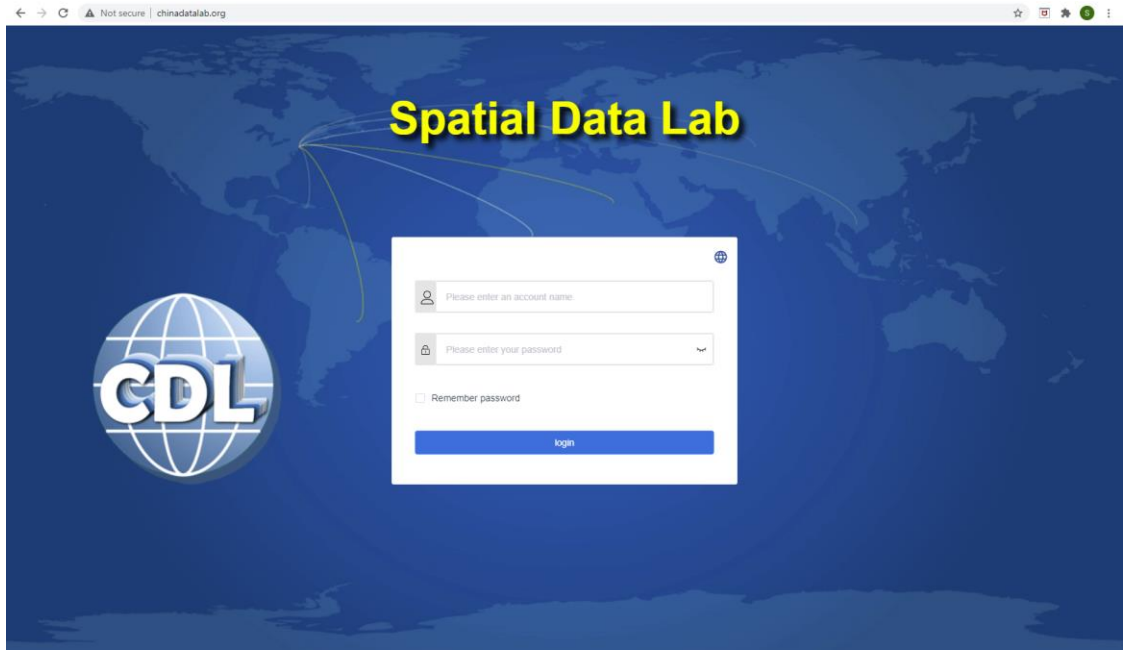
Web Sites (China Data Lab) 2020-2-11

News Report (China Data Lab) 2020-2-11

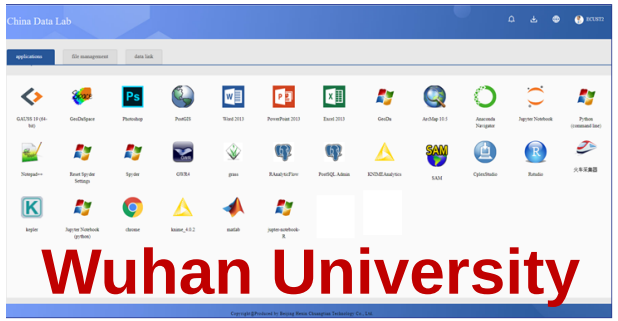
Feedback

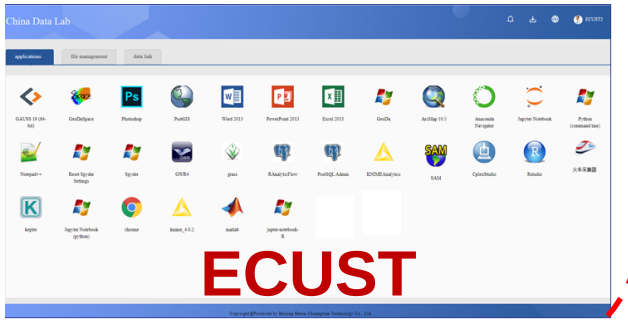
Spatial Data Lab on the Cloud

SDL at Harvard



SDL in China





http://harvard.chinadatalab.org

http://chinadatalab.cn

2 Spatial Big Data

© WUHAN UNIVERSITY.

2. Introduction - Spatial big data

Spatial Big Data—be this natively geocoded content, geographical metadata, or data that itself refers to spaces and places—has become a pervasive presence in the spaces and practices of everyday life. Beyond preoccupations with "the geotag" and with mapping geocoded social media content, this special theme explores what it means to encounter and experience spatial Big Data as a quotidian phenomenon that is both spatial, characterized by and enacting of material spatialities, and spatializing, configuring relations between subjects, objects, and spaces in new and unprecedented ways.

Source: Agnieszka Leszczynski, Jeremy Crampton, "Introduction: Spatial Big Data and everyday life"

Remote Sensing Data

Remote sensing has become one of the most important methods used to quickly and directly acquire information on the Earths surface. In recent years, with development of environmental information science, remote sensing data have played an important role in many research fields, such as atmospheric science, ecology, soil contamination, water pollution, environmental geology, environmental soil science, volcanic phenomena and evolution of the Earths crust.

Remote sensing data has several concrete and special characteristics: multi-source, multi-scale, high-dimensional, dynamic-state, isomer, and non-linear characteristics.

Source: Peng Liu, "A survey of Remote Sensing Data"

According to the Index of Objects Launched into Outer Space maintained by United Nations Office for Outer Space Affairs (UNOOSA), there are currently **5774** satellites currently orbiting the planet.

As of May 2020, there are currently **2666** operational satellites.

Remote Sensing Data Applications

Application - Forestry

1) reconnaissance mapping:
- forest cover type discrimination
- agroforestry mapping

2) Commercial forestry:
- clear cut mapping / regeneration assessment
- burn delineation
- infrastructure mapping / operations support
- forest inventory
- biomass estimation

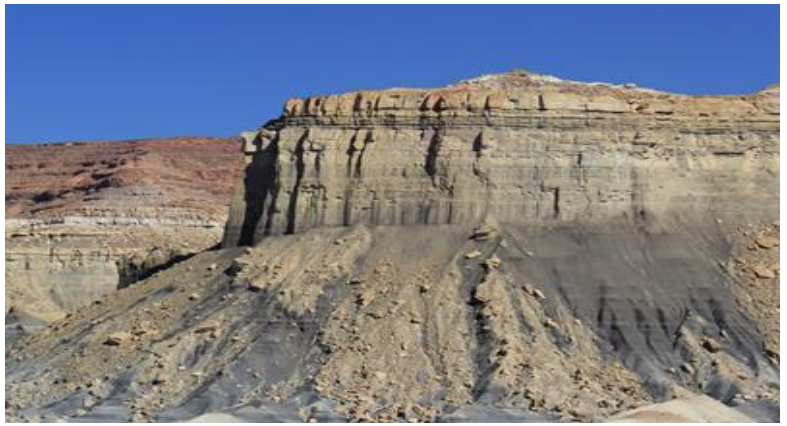
3) Environmental monitoring

Conservation authorities are concerned with monitoring the quantity, health, and diversity of the Earth's forests.
- deforestation (rainforest, mangrove colonies)
- species inventory
- watershed protection (riparian strips)
- coastal protection (mangrove forests)
- forest health and vigour

Remote Sensing Data Applications

Application - Geology

- surficial deposit / bedrock mapping
- lithological mapping
- structural mapping
- sand and gravel (aggregate) exploration/ exploitation
- mineral exploration
- hydrocarbon exploration
- environmental geology
- geobotany
- baseline infrastructure
- sedimentation mapping and monitoring
- event mapping and monitoring



2. Introduction - Remote Sensing Data

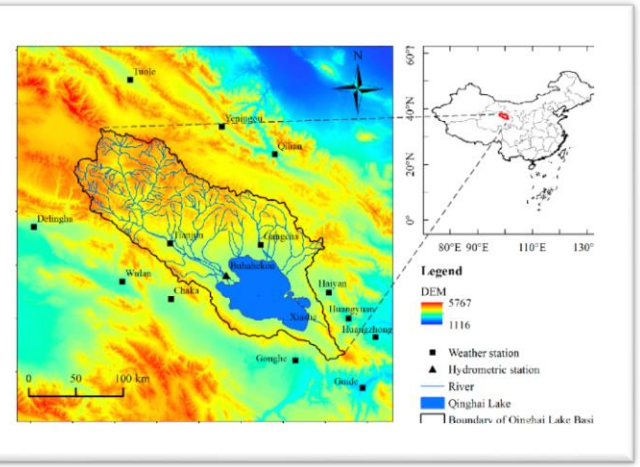
Application - Geology

- surficial deposit / bedrock mapping
- lithological mapping
- structural mapping
- sand and gravel (aggregate) exploration/ exploitation
- mineral exploration
- hydrocarbon exploration
- environmental geology
- geobotany
- baseline infrastructure
- sedimentation mapping and monitoring
- event mapping and monitoring

Remote Sensing Data Applications

Application - Hydrology

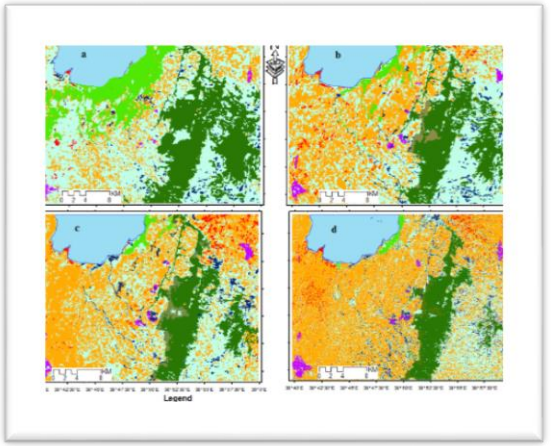
- wetlands mapping and monitoring,
- soil moisture estimation,
- snow pack monitoring / delineation of extent,
- measuring snow thickness,
- determining snow-water equivalent,
- river and lake ice monitoring,
- flood mapping and monitoring,
- glacier dynamics monitoring (surges, ablation)
- river /delta change detection
- drainage basin mapping and watershed modelling
- irrigation canal leakage detection
- irrigation scheduling



Remote Sensing Data Applications

Application - Land cover & Land use

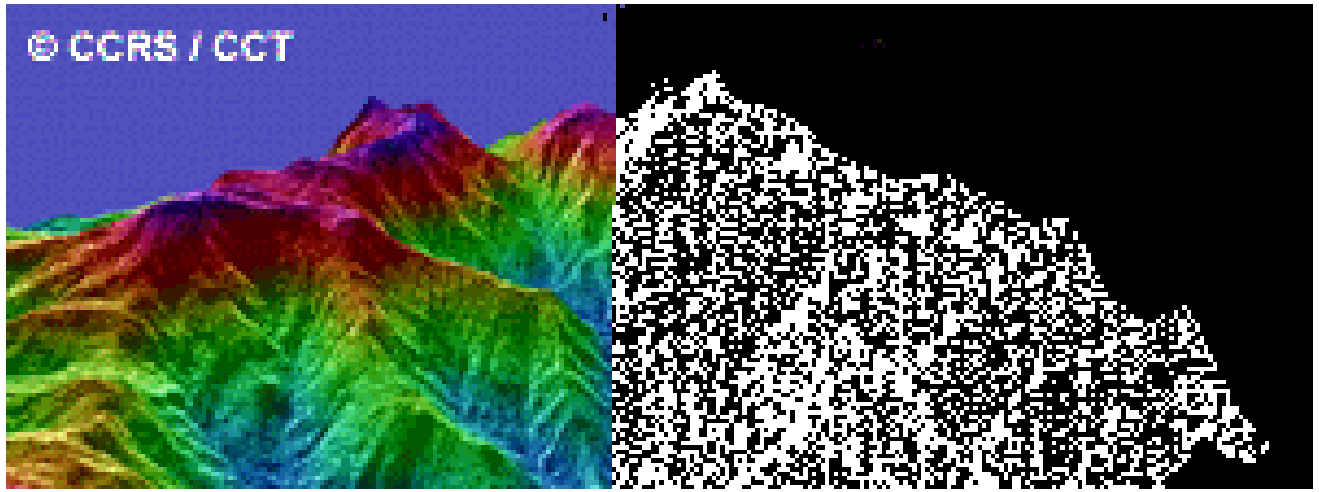
- natural resource management
- wildlife habitat protection
- baseline mapping for GIS input
- urban expansion / encroachment
- routing and logistics planning for seismic / exploration / resource extraction activities
- damage delineation (tornadoes, flooding, volcanic, seismic, fire)
- legal boundaries for tax and property evaluation
- target detection - identification of landing strips, roads, clearings, bridges, land/water
- interface



Remote Sensing Data Applications

Application - Mapping

- planimetry
- digital elevation models (DEM's)
- baseline thematic mapping / topographic mapping



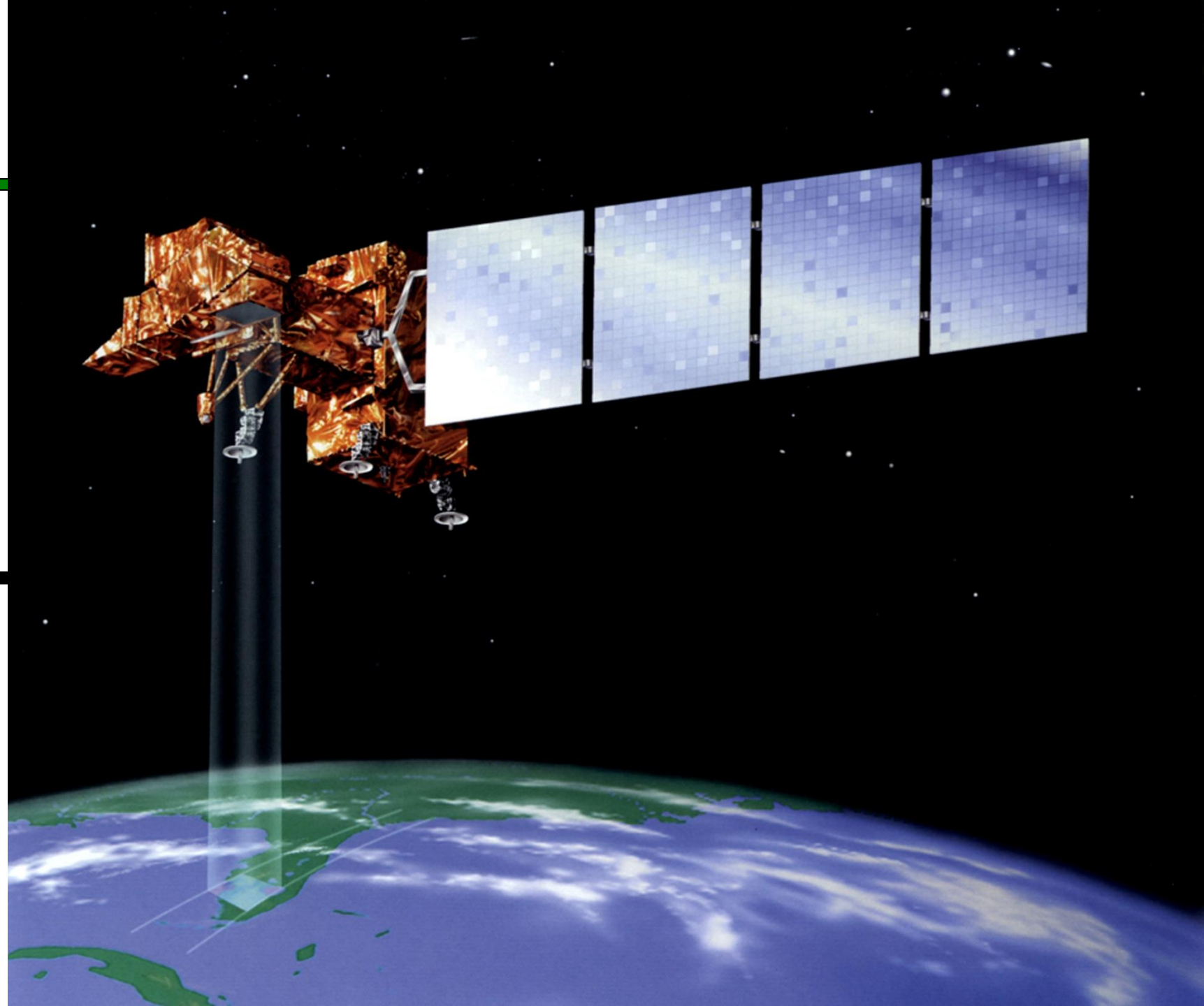
Major Earth Observation satellites

- Landsat
- SPOT
- Ikonos
- AVHRR
- Seawifs
- GOES
- Radarsat
- Meteosat
- ESA Satellites (ENVISAT)
- Japanese Satellites
- Terra A-Train

Major Application Areas

- Environmental Concerns
- Engineering
- Agriculture
- Forestry And Natural Resources
- Geology
- Oceanography
- Land use, Geography, Urban Planning
- Atmospheric Sciences
- Hydrology

Landsat Data



Landsat Data

The longest-running enterprise for acquisition of satellite imagery of Earth.

Swath Width: 185 km

Orbit Altitude: 705 km

Spectral bands : 8

Spatial resolutions : 15 ~ 60 meters

The temporal resolution : 16 days

Equatorial Crossing : at around 10 a.m. local solar time

One Landsat scene : about 115 miles long and 115 miles wide.

Divided into scenes for easy downloading.

Rich band

High resolution

Short repeat cycle

Easy to download & classify

Large map area

USGS science for a changing world

EarthExplorer

https://earthexplorer.usgs.gov/

Landsat Data Applications

- Agriculture
- Cartography
- Geology
- Forestry
- Regional planning
- Surveillance
- Education

……

Agriculture, Forestry and Range Resources	Land Use and Mapping	Geology	Hydrology	Coastal Resources	Environmental Monitoring
Discriminating vegetative, crop and timber types	Classifying land uses	Mapping major geologic features	Determining water boundaries and surface water areas	Determining patterns and extent of turbidity	Monitoring deforestation
Measuring crop and timber acreage	Cartographic mapping and map updating	Revising geologic maps	Mapping floods and flood plain characteristics	Mapping shoreline changes	Monitoring volcanic flow activity
Precision farming land management	Categorizing land capabilities	Recognizing and classifying certain rock types	Determining area extent of snow and ice coverage	Mapping shoals, reefs and shallow areas	Mapping and monitoring water pollution
Monitoring crop and forest harvests	Monitoring urban growth	Delineating unconsolidated rocks and soils	Measuring changes and extent of glacial features	Mapping and monitoring sea ice in shipping lanes	Determining effects of natural disasters
Determining range readiness, biomass and health	Aiding regional planning	Mapping volcanic surface deposits	Measuring turbidity and sediment patterns	Tracking beach erosion and flooding	Assessing drought impact

SPOT Data

SPOT Data

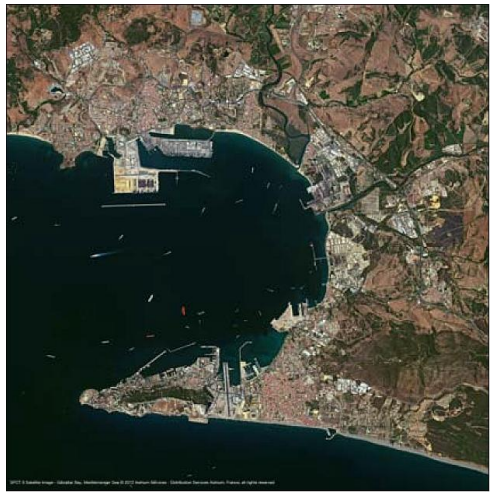
SPOT

- A commercial high-resolution optical imaging Earth observation satellite system.
- Run by Spot Image, based in Toulouse, France.
- Initiated by the CNES in the 1970s
- Developed in association with the SSTC and the Swedish National Space Board.

https://www.usgs.gov/centers/eros/science/usgs-eros-archive-commercial-satellites-spot-north-american-data-buy?qt-science_center_objects=0#qt-science_center_objects

SPOT Data Applications

- Agriculture
- Planning, land use and landcover
- Cadastral mapping
- Cartography and topography
- Urban planning
- Forestry
- Natural reserve management and planning
- Natural hazard and pollution monitoring
- Geology, mineral and oil exploration
- Water resources
- Coastal and ocean studies
- Monitoring and surveillance



Ikonos Data

Ikonos Data

IKONOS was a commercial Earth observation satellite, and was the first to collect publicly available high-resolution imagery at 1- and 4-meter resolution. It collected multispectral and panchromatic imagery.

The capability to observe Earth via space-based telescope has been called "one of the most significant developments in the history of the space age", and IKONOS brought imagery rivaling that of military spy satellites to the commercial market.

IKONOS imagery began being sold on 1 January 2000, and the spacecraft was retired in 2015.

https://www.usgs.gov/centers/eros/science/usgs-eros-archive-commercial-satellites-cdp-imagery-ikonos?qt-science_center_objects=0#qt-science_center_objects

Ikonos Data Applications

Revolutions Around the Earth	14.7, every 24 hours
Altitude	681 kilometers
Resolution at Nadir	0.82 meters panchromatic; 3.28 meters multispectral
Resolution 26° Off-Nadir	1.0 meter panchromatic; 4.0 meters multispectral
Image Swath	11.3 kilometers at nadir; 13.8 kilometers at 26° off-nadir
Dynamic Range	11-bits per pixel
Image Bands	Panchromatic, blue, green, red, near IR

Ikonos satellite provides panchromatic, natural-color, color-infrared and stereo images. In particular, this is useful for cartographic, photogrammetric and various remote sensing applications.

Its applications include both urban and rural mapping of natural resources and of natural disasters, tax mapping, agriculture and forestry analysis, mining, engineering, construction, and change detection.

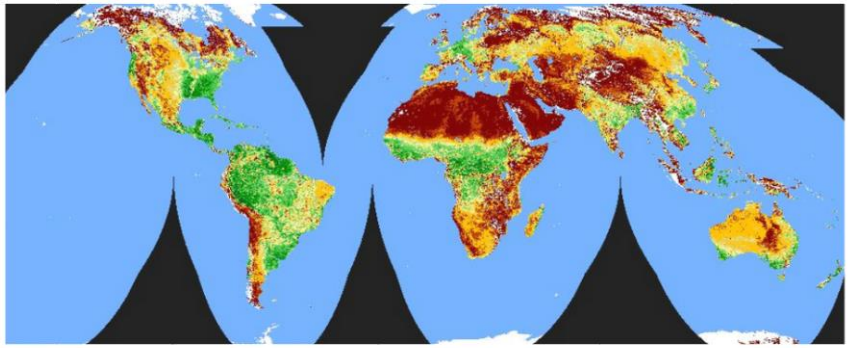
AVHRR Data

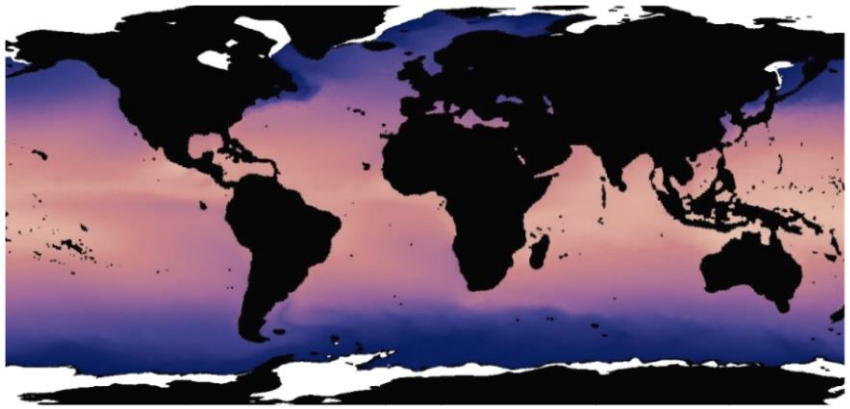
AVHRR Data

The Advanced Very-High-Resolution Radiometer

A space-borne sensor that measure the reflectance of the Earth in five spectral bands (relatively wide by today's standards.)

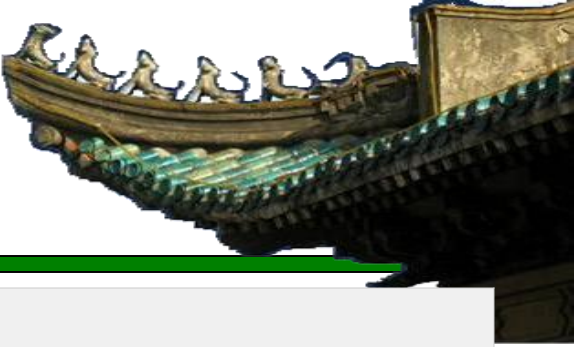
A multispectral sensor with six spectral bands.
Includes red, thermal, mid and near-infrared bands.
But over time, their spectral ranges have varied.





https://www.usgs.gov/centers/eros/science/usgs-eros-archive-advanced-very-high-resolution-radiometer-avhrr?qt-science_center_objects=0#qt-science_center_objects

AVHRR – spectral bands and specifications



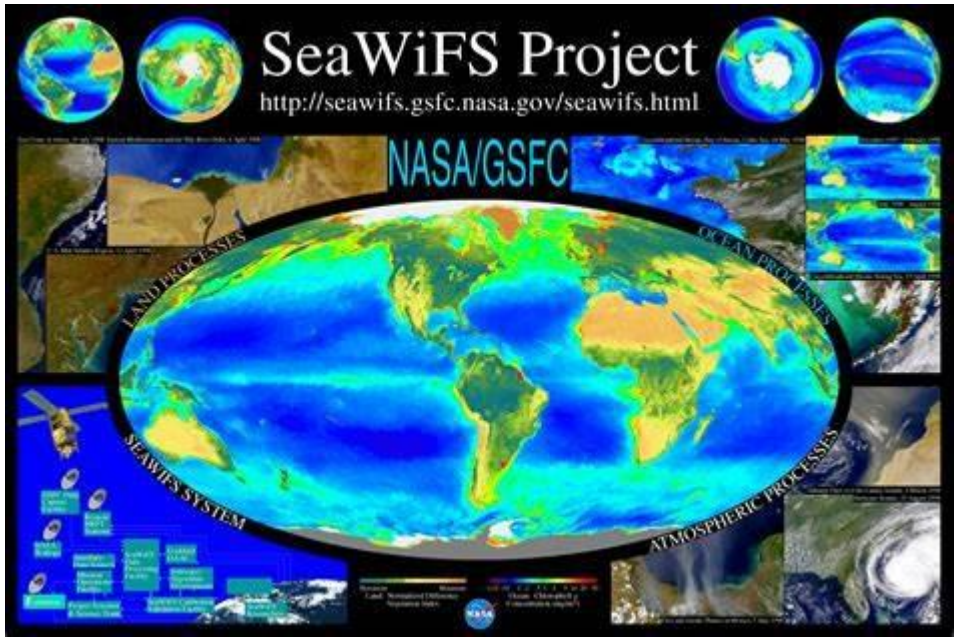
Band	Name	Spectral Range	Applications
Band 1	Red	0.58-0.68	Urban, vegetation, snow/ice, daytime clouds
Band 2	Near IR	0.725-1.00	Vegetation, land/water boundaries, snow/ice, flooding
Band 3A	Mid IR	1.58-1.64	Vegetation, snow/ice detection, dust monitoring
Band 3B	Thermal	3.55-3.93	Surface temperature, wildfire detection, nighttime clouds, volcanic eruptions
Band 4	Thermal	10.30-11.30	Surface temperature, wildfire detection, nighttime clouds, volcanic eruptions
Band 5	Thermal	11.5-12.50	Sea surface temperature, water vapor path radiance

SeaWIFS Data

SeaWIFS Data

SeaWIFS was a satellite-borne sensor designed to collect global ocean biological data. Active from September 1997 to December 2010, its primary mission was to quantify chlorophyll produced by marine phytoplankton.

Swath width : ~2875 km at the equator
Ground resolution : ~1 km / pixel at nadir



The near-IR bands (765 and 865 nm) are used to derive the atmospheric path radiance, which is then subtracted from the total signal to obtain water leaving radiance (Lw) in the visible.

https://oceancolor.gsfc.nasa.gov/forum/oceancolor/topic_show.pl?tid=85

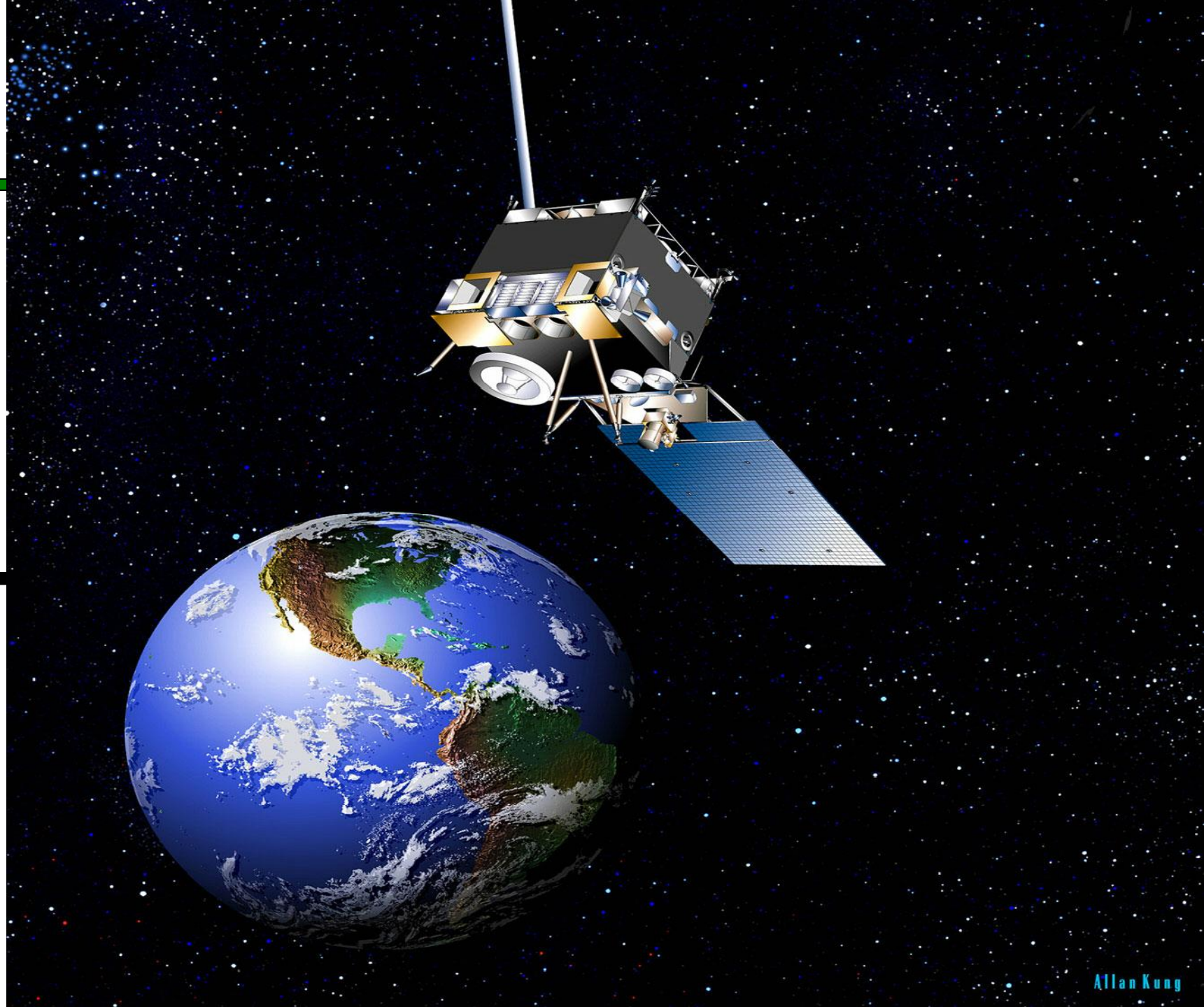
SeaWIFS Data Applications

SeaWiFS' primary application is to sensing of color variations in the Earth's oceans. Specifically, the color variations are indicative of phytoplankton content, suspended sediments, and the presence of dissolved organic material (DOM or yellow matter).

The purpose of the SeaWiFS Project is to provide quantitative data on global ocean bio-optical properties to the Earth science community. Subtle changes in ocean color signify various types and quantities of marine phytoplankton (microscopic marine plants), the knowledge of which has both scientific and practical applications.

https://oceancolor.gsfc.nasa.gov/SeaWiFS/

GOES Data

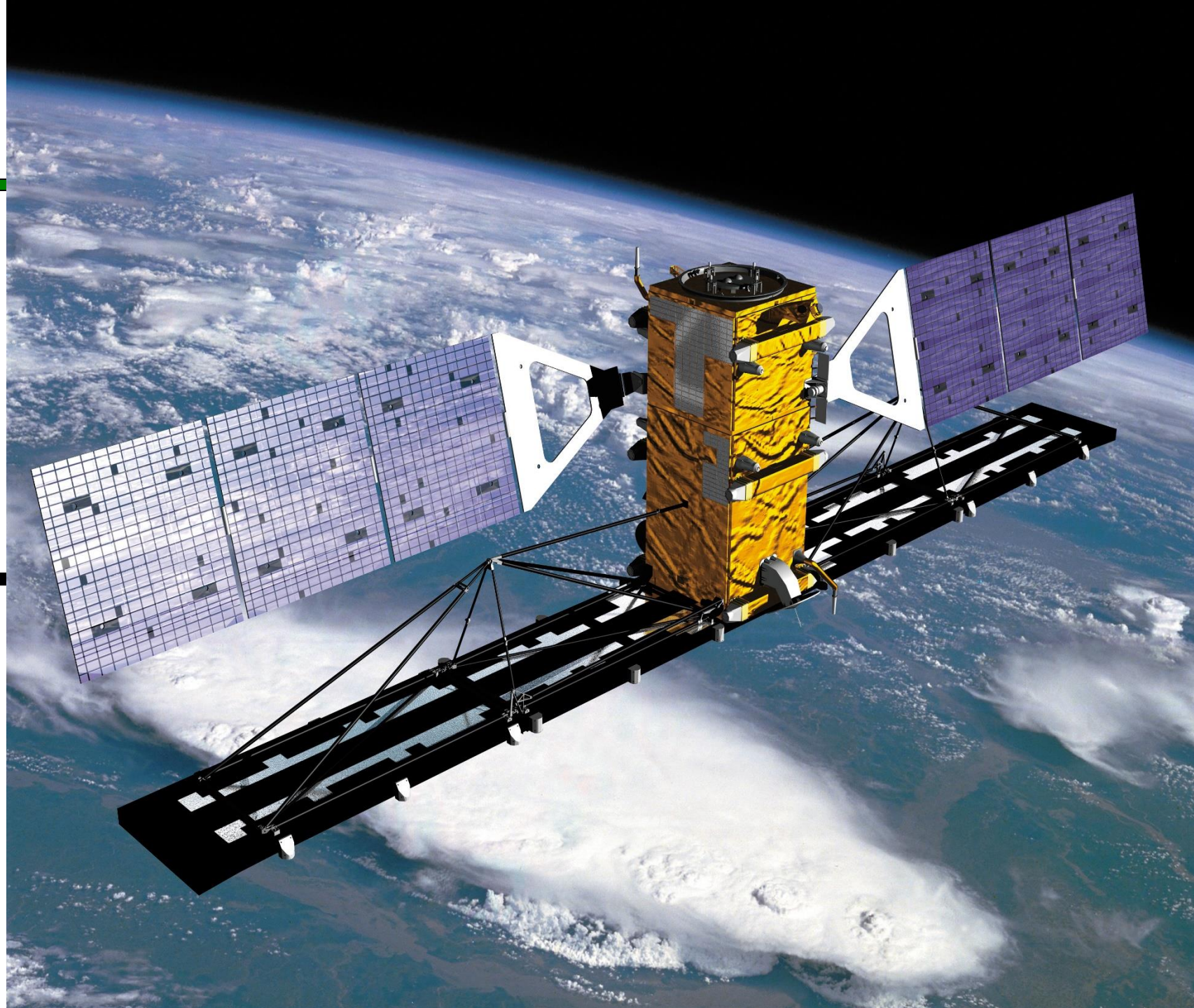


GOES Data

The Geostationary Operational Environmental Satellite system (GOES), operated by the United States' National Environmental Satellite, Data, and Information Service (NESDIS), supports weather forecasting, severe storm tracking, and meteorology research. Spacecraft and ground-based elements of the system work together to provide a continuous stream of environmental data

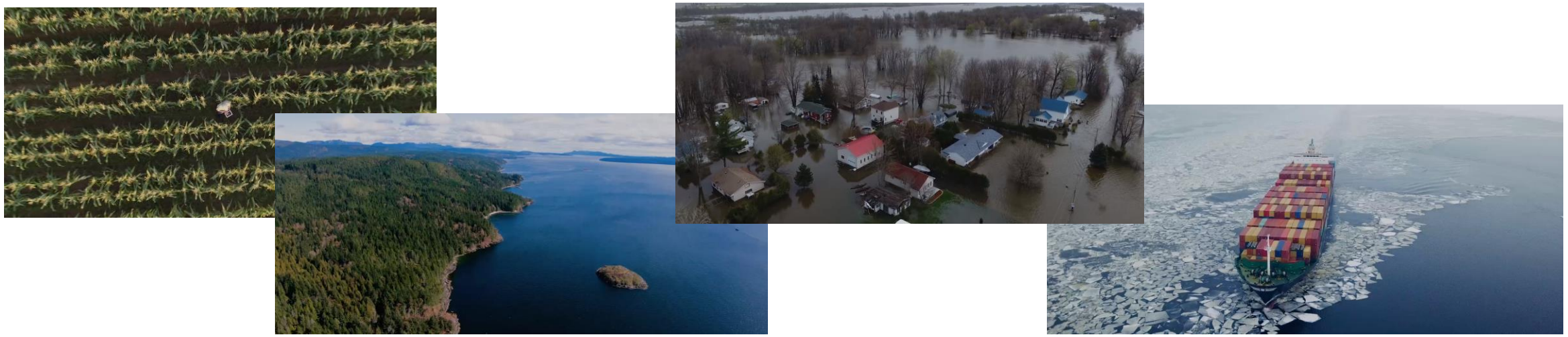
https://www.star.nesdis.noaa.gov/goes/

Radarsat Data



Radarsat Data Application

RADARSAT was designed to provide detailed information on sea ice and terrestrial ice sheets for climate research, to produce radar imagery for geographical applications in oceanography, agriculture, forestry, hydrology, and geology, and to provide real-time products for arctic ocean navigation including ice surveillance.



https://www.asc-csa.gc.ca/eng/satellites/radarsat/default.asp

RADARSAT

ESA Data

adm aeolus
cryosat 2
swarm
earthcare
goce
smos
LIVING PLANET
EARTH EXPLORERS

ESA Data - ENVISAT

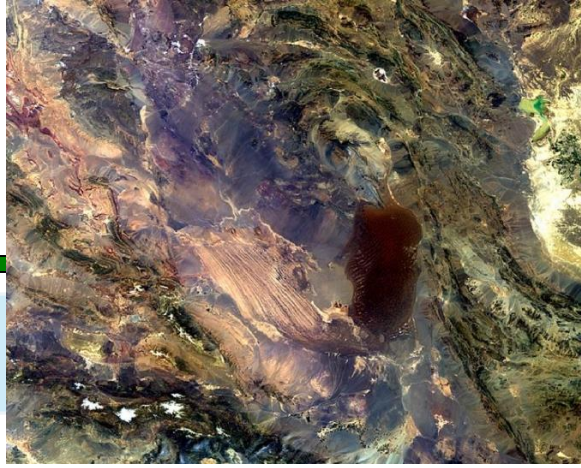
The European Space Agency (ESA) is Europe’s gateway to space. Its mission is to shape the development of Europe’s space capability and ensure that investment in space continues to deliver benefits to the citizens of Europe and the world.

Envisat was ESA's successor to ERS.
The largest civilian Earth observation mission.(8 tons)

Envisat had 10 instruments to provide continuous observation and monitoring of Earth's land, atmosphere, oceans and ice caps

https://www.esa.int/

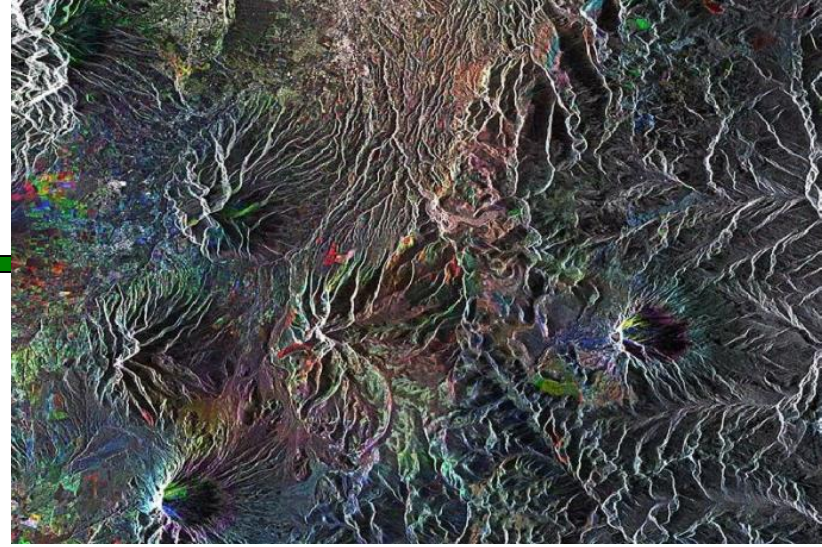
ENVISAT Data Applications: Atmosphere



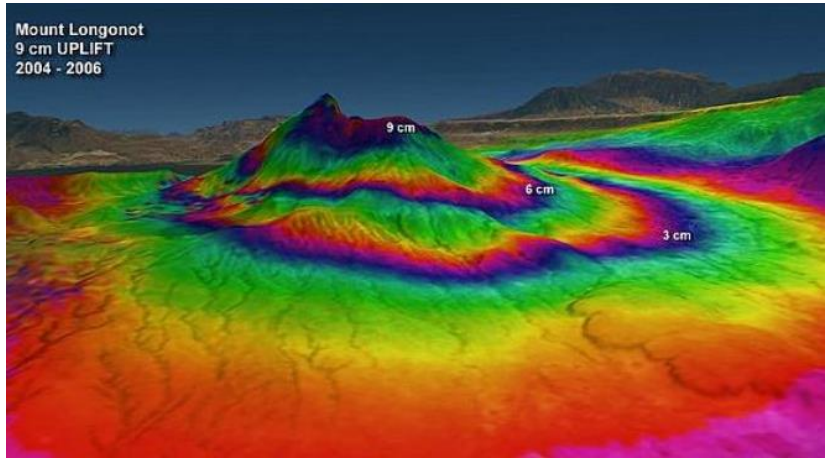
Atmosphere

	ASAR	GOMOS	RA-2	MERIS	MIPAS	MWR	LR	SCIA	AATSR	DORIS
Clouds				x	x			x	x	
Water Vapour		x		x	x	x			x	
Radiation Budget		(x)		x	(x)	x		(x)	x	
Temperature / Pressure		x		x	x			x		
Trace Gases		x			x			x		
Aerosols		x		x	x			x	x	
Turbulence		x								

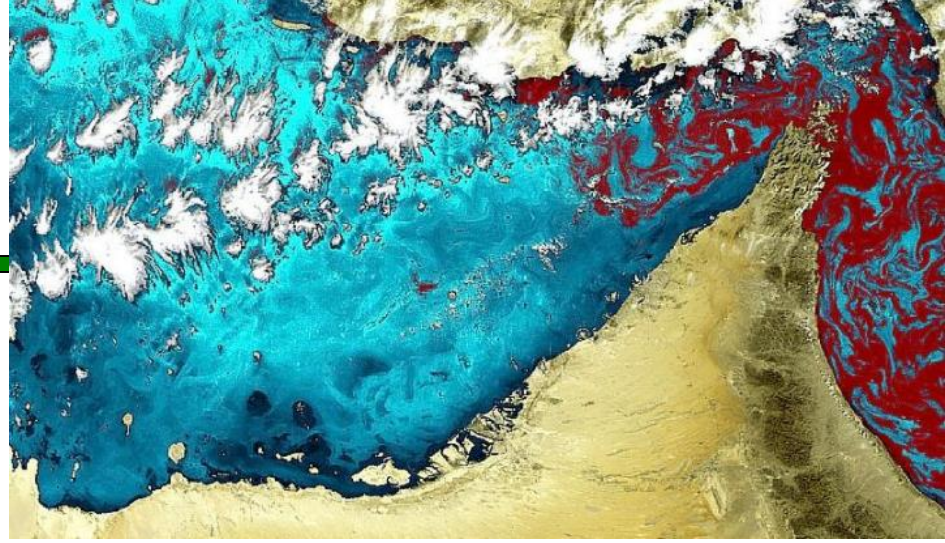
ENVISAT Data Applications: Land



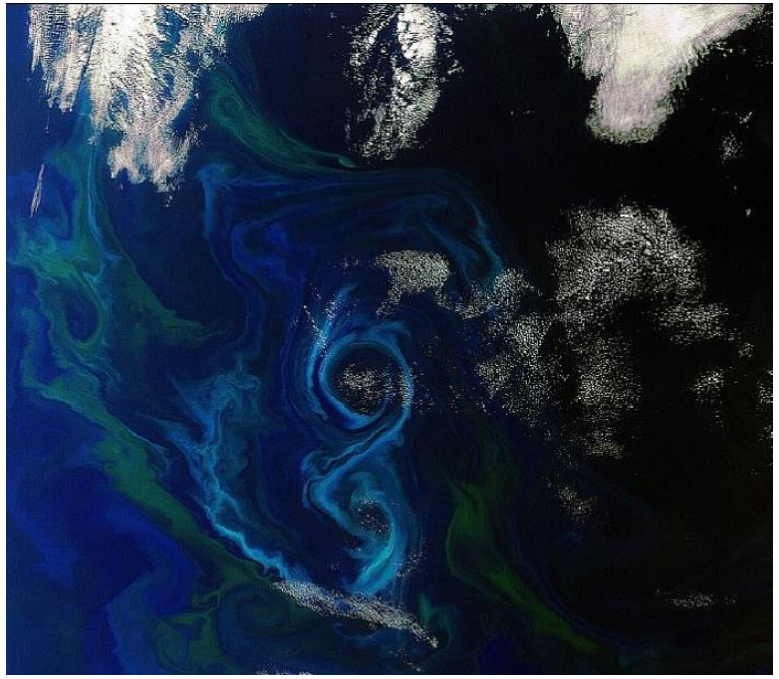
Land	ASAR	GOMOS	RA-2	MERIS	MIPAS	MWR	LR	SCIA	AATSR	DORIS
Surface Temperature						(x)			(x)	
Vegetation Characteristics	x			x					x	
Agriculture and Forestry	x			(x)					(x)	
Surface Elevation	x		x				x		x	x
Geology and Topography	x		(x)						x	
Hydrology Parameters	x		(x)	x		(x)			x	
Flooding	x									
Fire									x	



ENVISAT Data Applications: Ocean



Ocean	ASAR	GOMOS	RA-2	MERIS	MIPAS	MWR	LR	SCIA	AATSR	DORIS
Ocean Colour				x						
Sea Surface Temperature									x	
Surface Topography			x				x			x
Turbidity				x						
Wave Characteristics	x		x							
Near Surface Wind	x		x							
Current Features	x		x							
Marine Geoid			x							
Global Circulation			x						x	
Ocean Fronts	x			x					(x)	



ALOS Data

ALOS Data

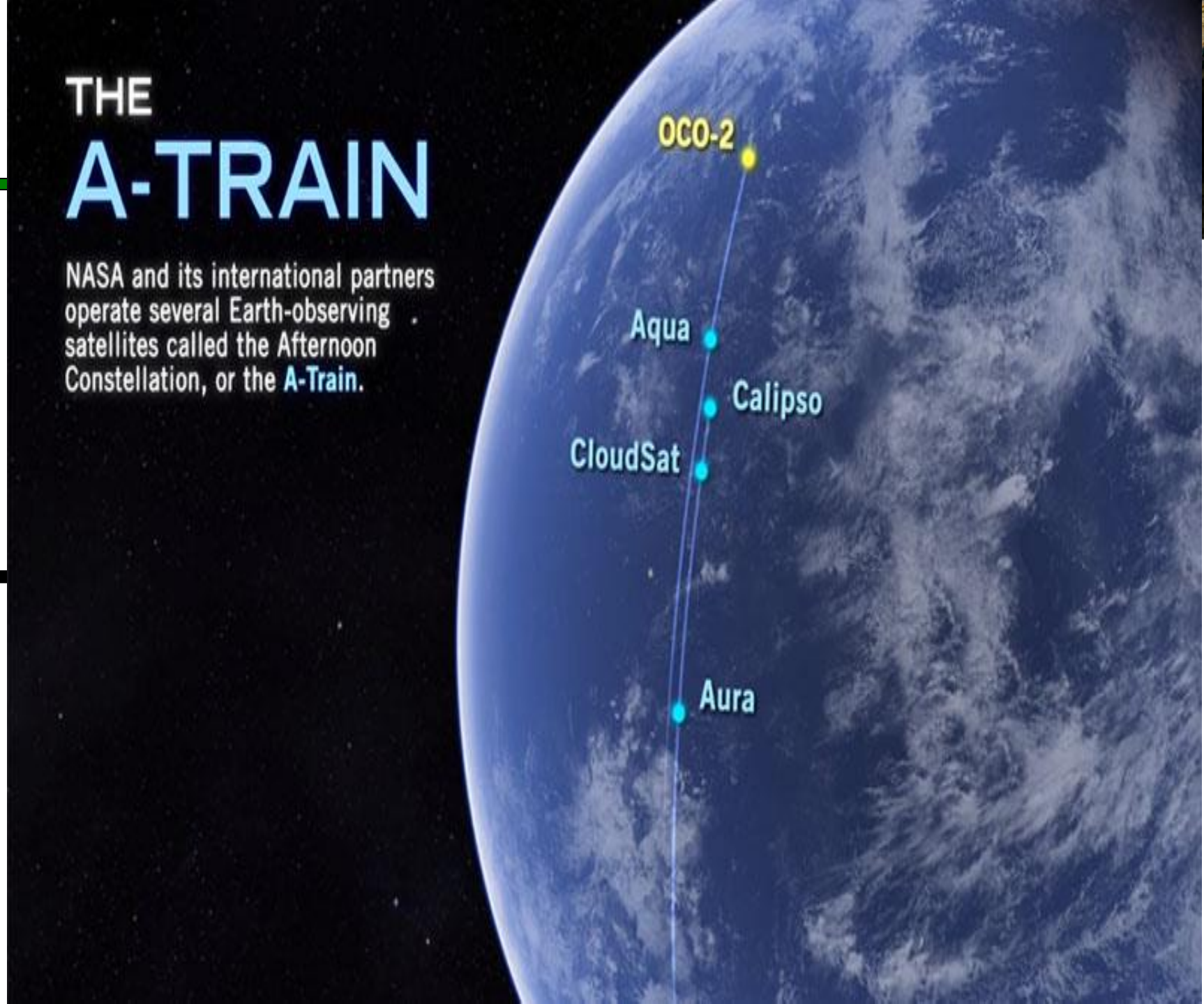
ALOS is one of the largest Earth observing satellites. Its objectives are:

1. to provide maps for Japan and other countries including those in the Asian-Pacific region (Cartography)
2. to perform regional observation for "sustainable development", harmonization between Earth environment and development (Regional Observation),
3. to conduct disaster monitoring around the world (Disaster Monitoring),
4. to survey natural resources (Resources Surveying),
5. to develop technology necessary for future Earth observing satellite (Technology Development)

Observation parameter	ALOS (launch 2006)	ALOS-2 (launch 2014)
Spatial resolution	- Strip map: 10 m - ScanSAR: 100 m	- Strip map: 3 m /6 m /10 m - ScanSAR: 100 m - Spotlight: 1 m x 3 m

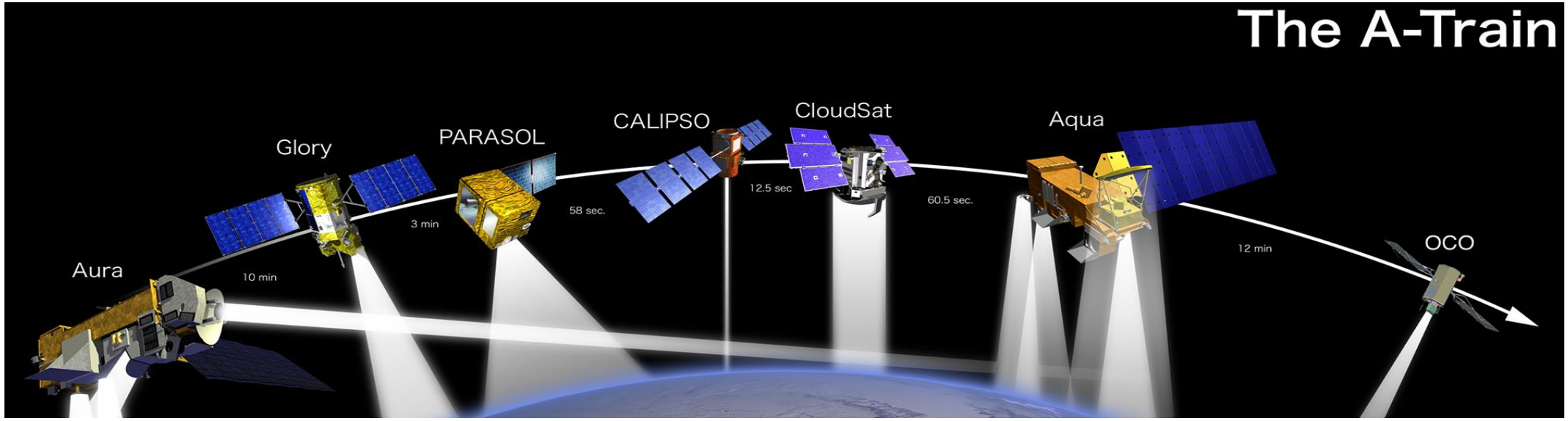
https://www.eorc.jaxa.jp/ALOS/en/dataset/dataset_index.htm

NASA A-Train


THE
A-TRAIN
NASA and its international partners operate several Earth-observing satellites called the Afternoon Constellation, or the A-Train.
OCO-2
Aqua
Calipso
CloudSat
Aura

NASA A-Train Data

The A-train is a satellite constellation of five Earth observation satellites of varied nationality in sun-synchronous orbit at an altitude of 705 km above the Earth. The orbit, at an inclination of 98.14°, crosses the equator each day at around 1:30 pm solar time, giving the constellation its name and crosses the equator again on the night side of the Earth, at around 1:30 am.

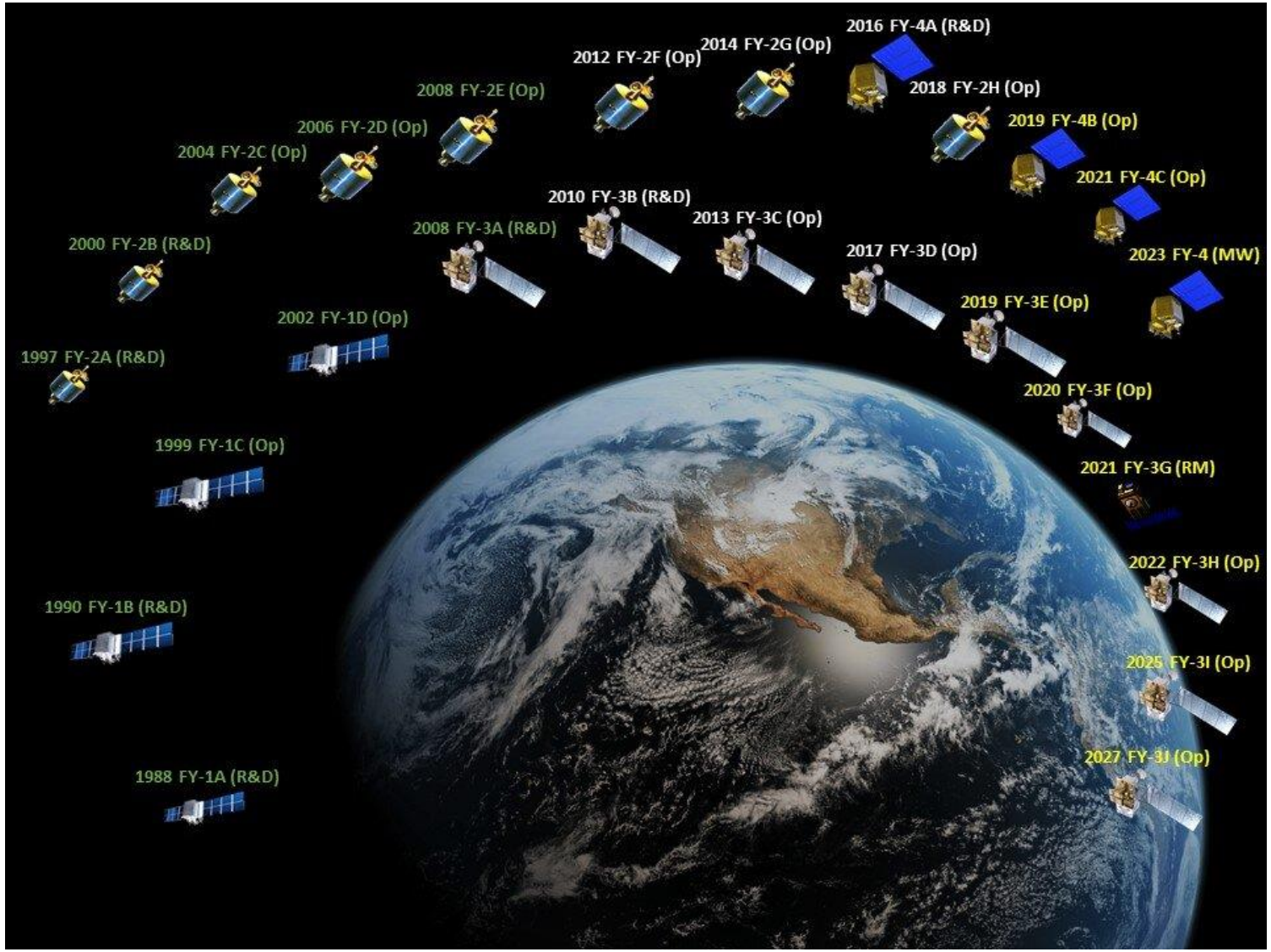


https://atrain.nasa.gov/

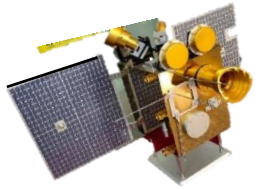
NASA A-Train Data Applications

Satellite	Related Applications
Aqua	Agricultural Efficiency, Air Quality, Carbon Management, Coastal Management Disaster Management, Ecological Forecasting, Water Management
Aura	Agricultural Efficiency, Air Quality, Public Health
PARASOL	Agricultural Efficiency, Air Quality, Carbon Management, Ecological Forecasting, Energy Management
CALIPSO	Air Quality, Aviation, Homeland Security, Public Health
CloudSat	Aviation, Weather Prediction
GCOM-W1	Agricultural Efficiency, Coastal Management, Disaster Management, Homeland Security, Water Management, Weather Prediction
OCO-2	Air Quality, Carbon Management, Public Health

Chinese Satellite Data



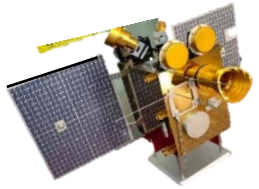
Luojia 1-01

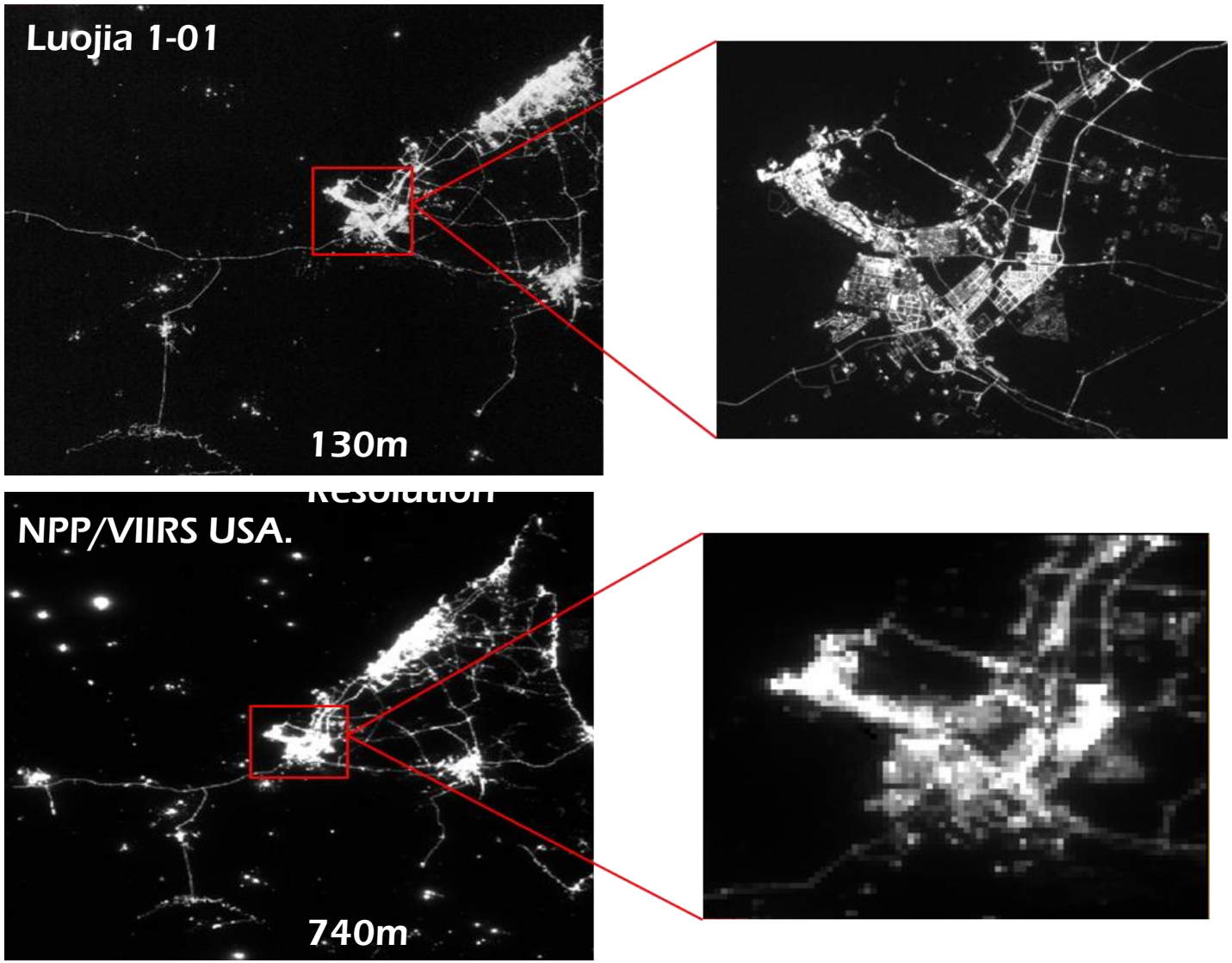


- Launched on June 2, 2018.
- A CubeSat (6U) sized satellite built by the Wuhan University.
- The satellite acts as a prototype for a future 60-80 earth observation satellite constellation.
- Luojia 1-01 features an imager with 100m ground resolution.
- 130 meters surface resolution of night light remote sensing data
- 5900 scenes of effective data have been obtained
- 230000 scenes of data have been distributed to more than 4000 users for free from more than 20 countries.
- Used by Beidou and GPS navigation.

http://59.175.109.173:8888/index.html

Luojia 1-01

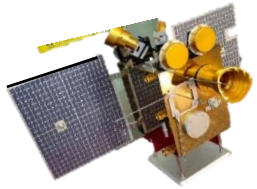


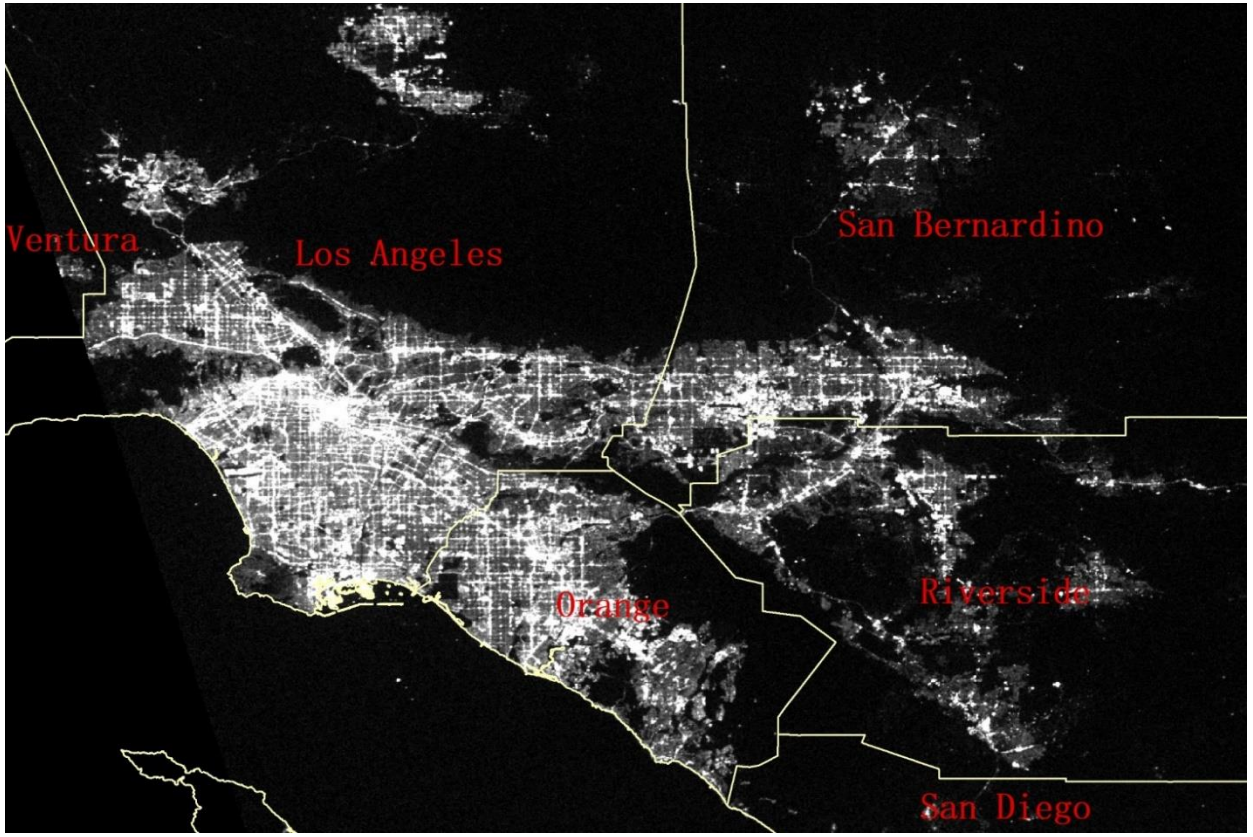


High sensitivity luminous camera

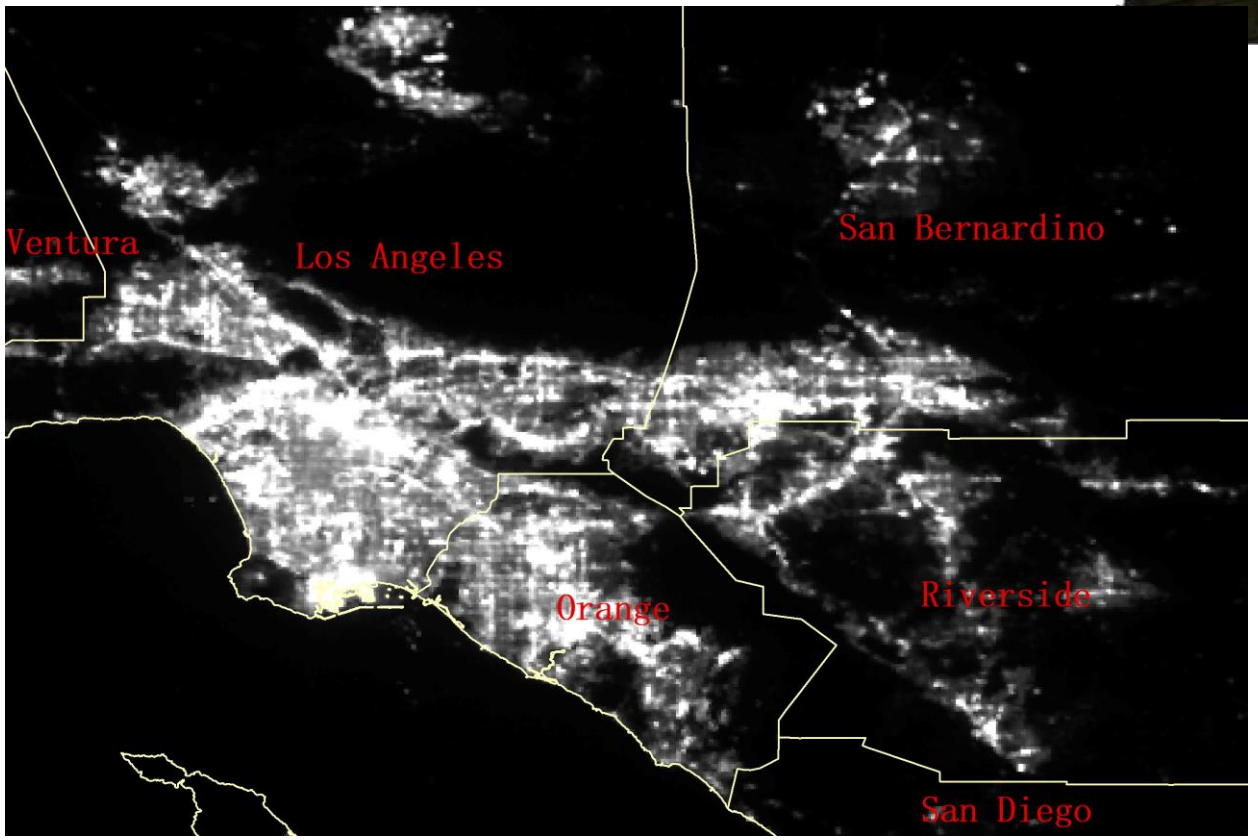
High resolution image data

Luojia 1-01



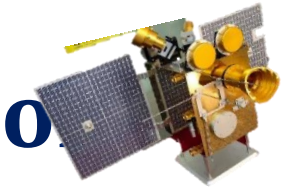


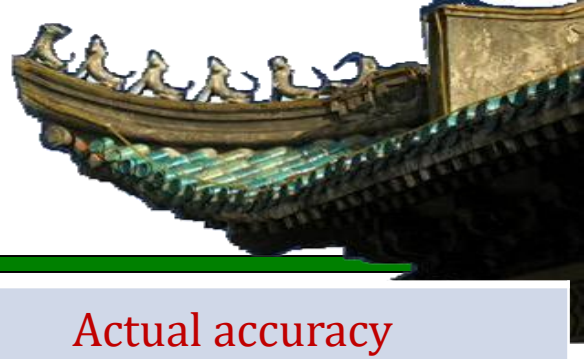
Luojia 1-01 night-time remote sensing image of Los Angeles, USA (June 6, 2018)



S-NPP /VIIRS night-time remote sensing image of Los Angeles, USA

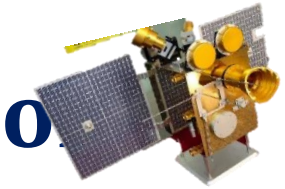
Luojia 1-01 Data Applicatio





Type	Example	Data	Actual accuracy
GDP Analysis	Correlation analysis between China's nighttime lighting and GDP in 2012	NPP_VIIRS	Cor.Coeff（R^2） 0.8702
Population analysis	Spatial analysis of coastal population based on census and multi-source night light data	NPP_VIIRS DMSP_OLS	Cor.Coeff（R^2）0.785\0.798
Analysis of power consumption	Correlation analysis between night lighting and power consumption in China in 2012	NPP_VIIRS	Cor.Coeff（R^2）0.8961
Carbon Emission	Correlation analysis between global light distribution and carbon dioxide emissions	DMSP_OLS	Cor.Coeff（R^2）0.73
Urban Expansion	Urbanization pattern and process in Yangtze River Delta Based on night light data	DMSP_OLS	The extraction error of each city is within 7%, the average error is .52
Urbanization	Analysis of urban change in China based on night light data	DMSP_OLS	Most of the area extraction coincidence > 75%
Poverty	World poverty map based on noctilucent remote sensing data	DMSP_OLS	Cor.Coeff（R^2）0.7217

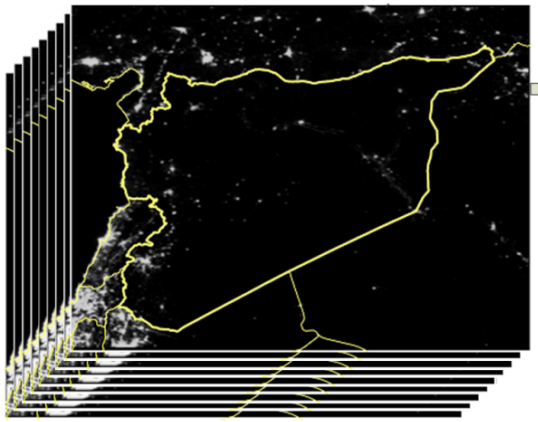
Luojia 1-01 Data Applicatio



Syria war impact assessment

The temporal and spatial data clustering technology is used to mine the temporal and spatial change patterns of Syrian luminous.

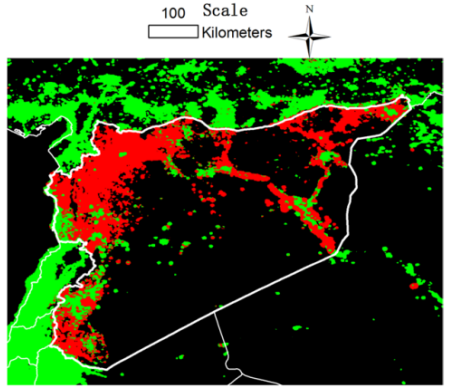
The clustering results of the two categories show that the luminous patterns are divided by the Syrian border line, and the clustering results of the three categories also reflect similar patterns.



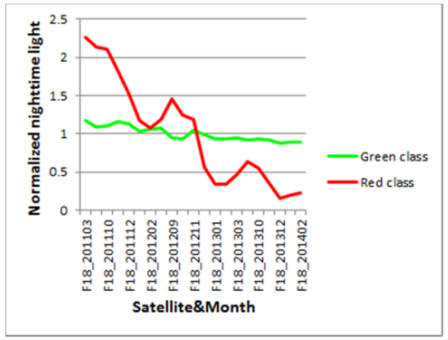
Time series night-time light images

Normalization

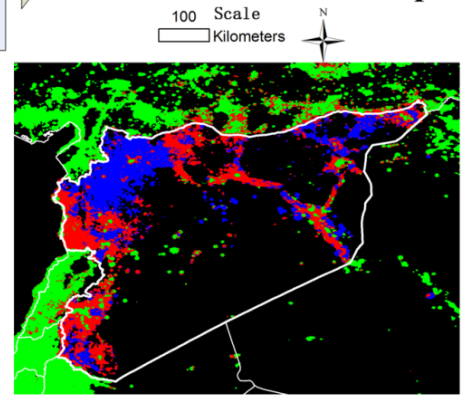
Clustering analysis



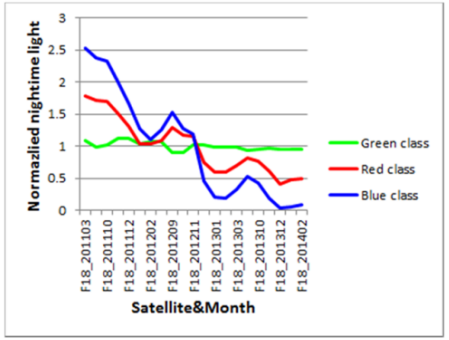
Two-class based class map



Two-class based class centers



Three-class based class map



Three-class based class centers

Luojia 1-01 Data Applicatio

Assessment of unbalanced regional development in China

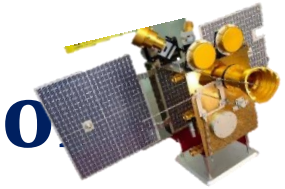
- Regional unbalanced development has always been the important problems in China's regional economy

A primary school in Beijing , China

A village primary school in Guizhou, China

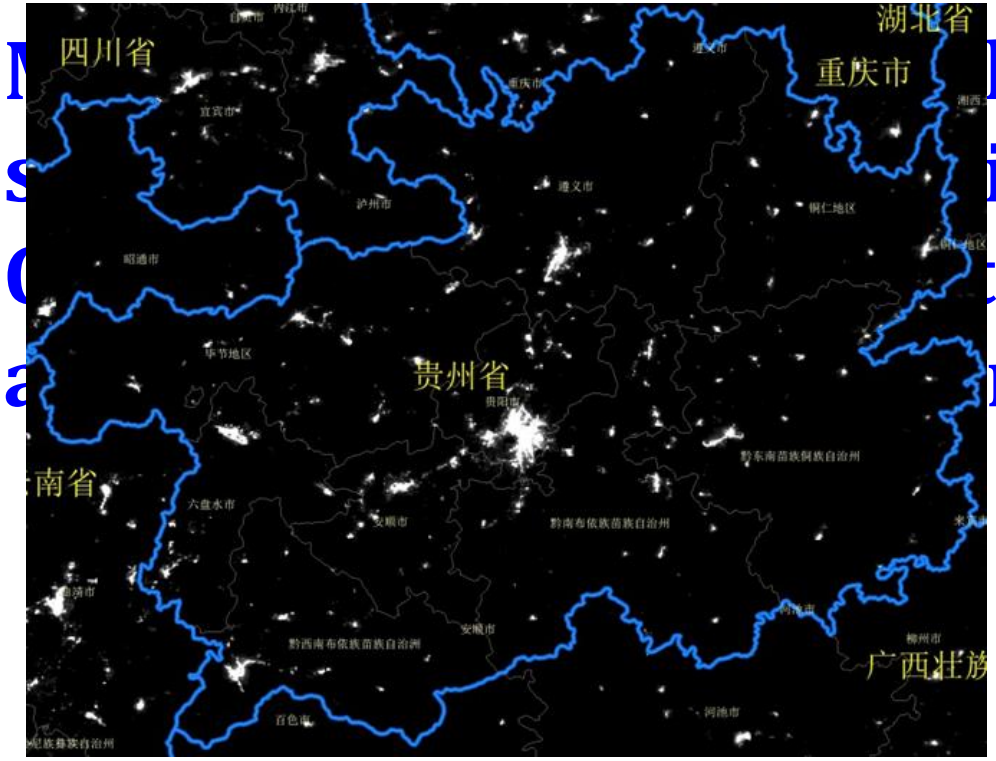
- The central government of China has strengthened the policy of regional coordinated development

Luojia 1-01 Data Applicatio

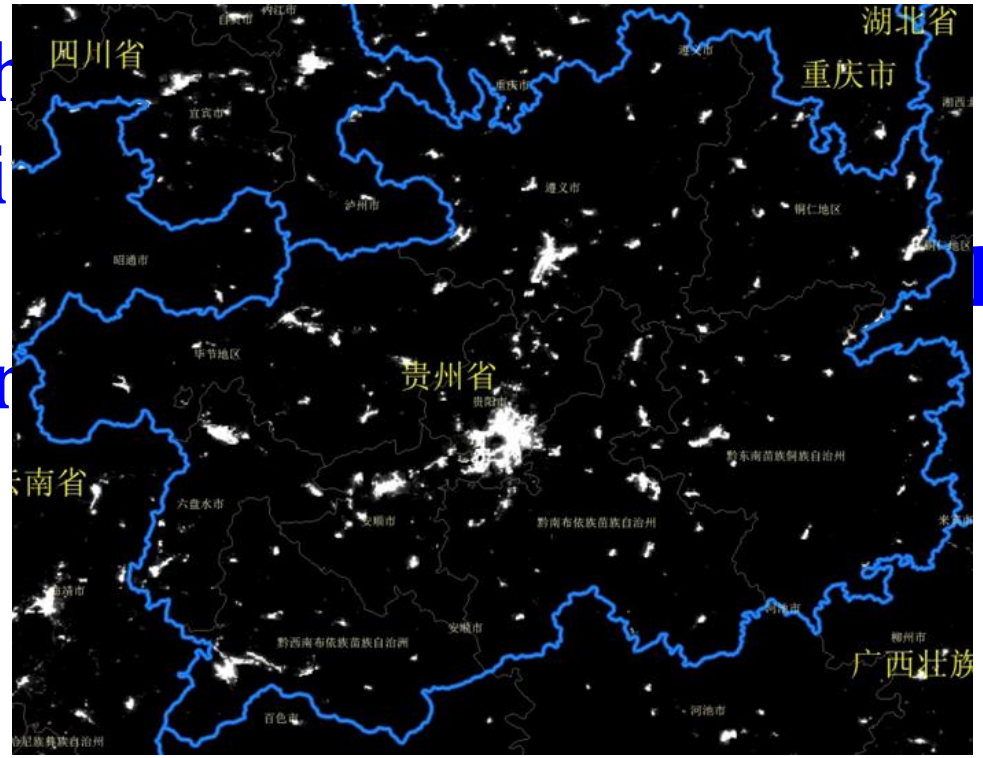


Assessment of unbalanced regional development in China

- M… h… s… i… C… h… a… r…

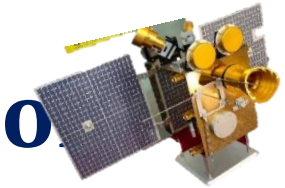


Guizhou province – 2012



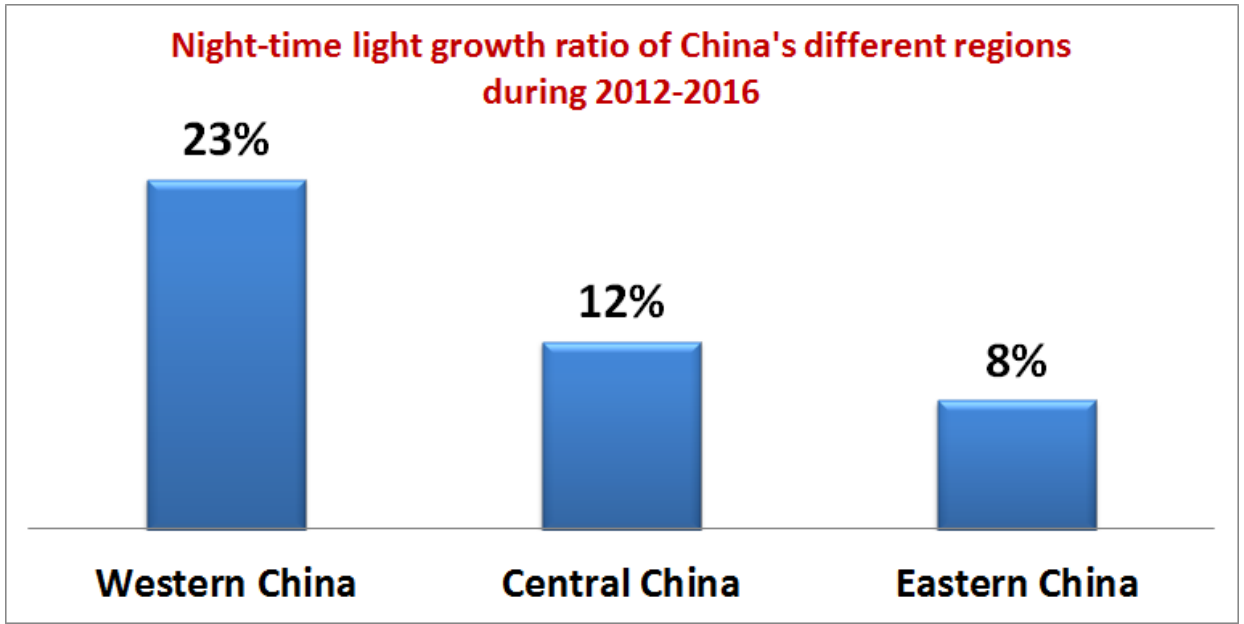
Guizhou province – 2016

Luojia 1-01 Data Applicatio



Assessment of unbalanced regional development in China

- According to our analysis, Guizhou, Chongqing, Xinjiang, Hunan, Qinghai, Sichuan, Anhui, Jiangxi and Hubei are the 10 provinces with the fastest growth in lighting. All of these provinces are located in the West or central part of China, and none of them is located in the east of China.
- The above analysis shows that the urban construction and development speed in the West and central China is faster than that in the East, which proves that the regional coordinated development policy has played a role.



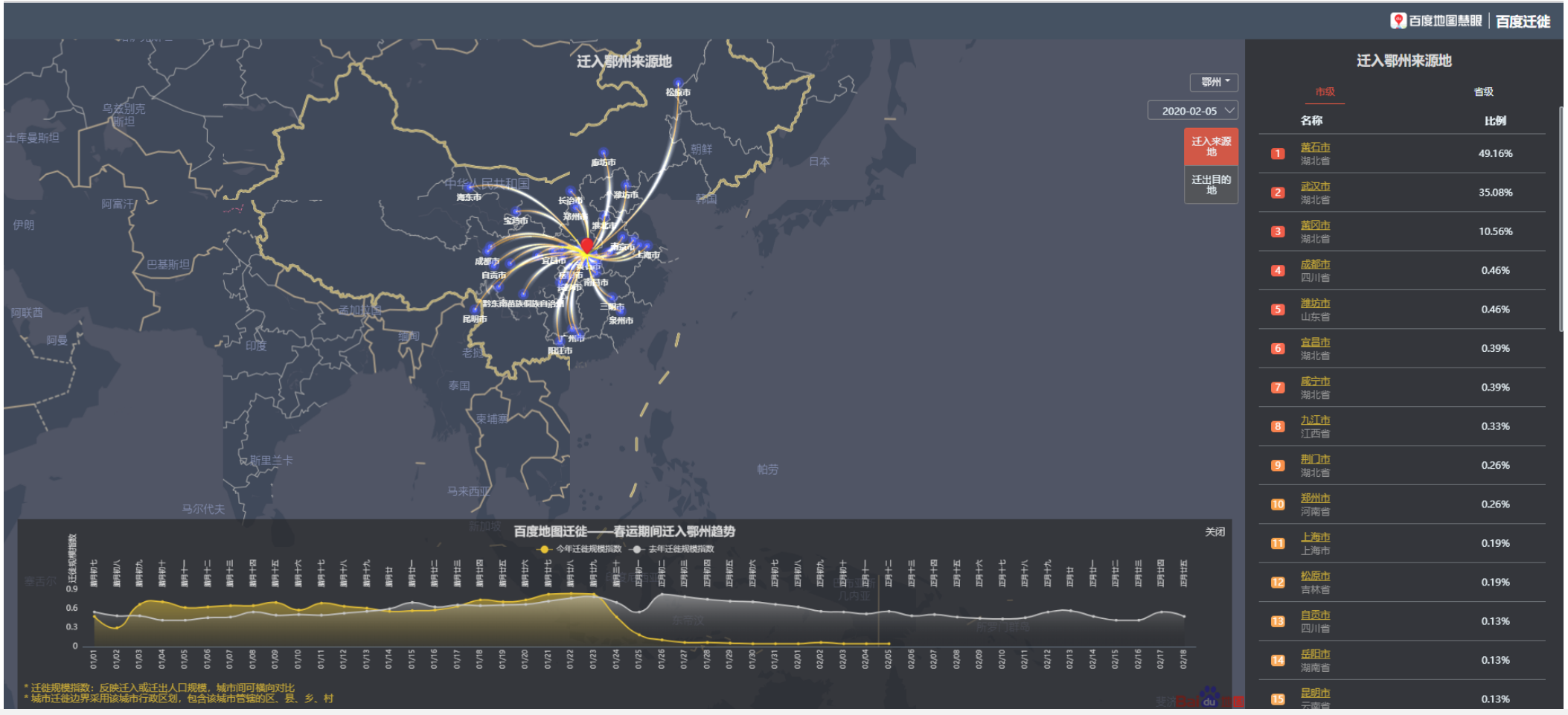
LIESMARS

3 Social Big Data

© WUHAN UNIVERSITY.

3. Social Big Data: Mobility Data

Mobility data



GPS

IP

Base station

WiFi

Source: https://qianxi.baidu.com/?from=shoubai#city=0

Social Big Data: Mobility Data

Mobility data

The proportion of people moving to the source area: the proportion of people who move to a city in a day in other cities

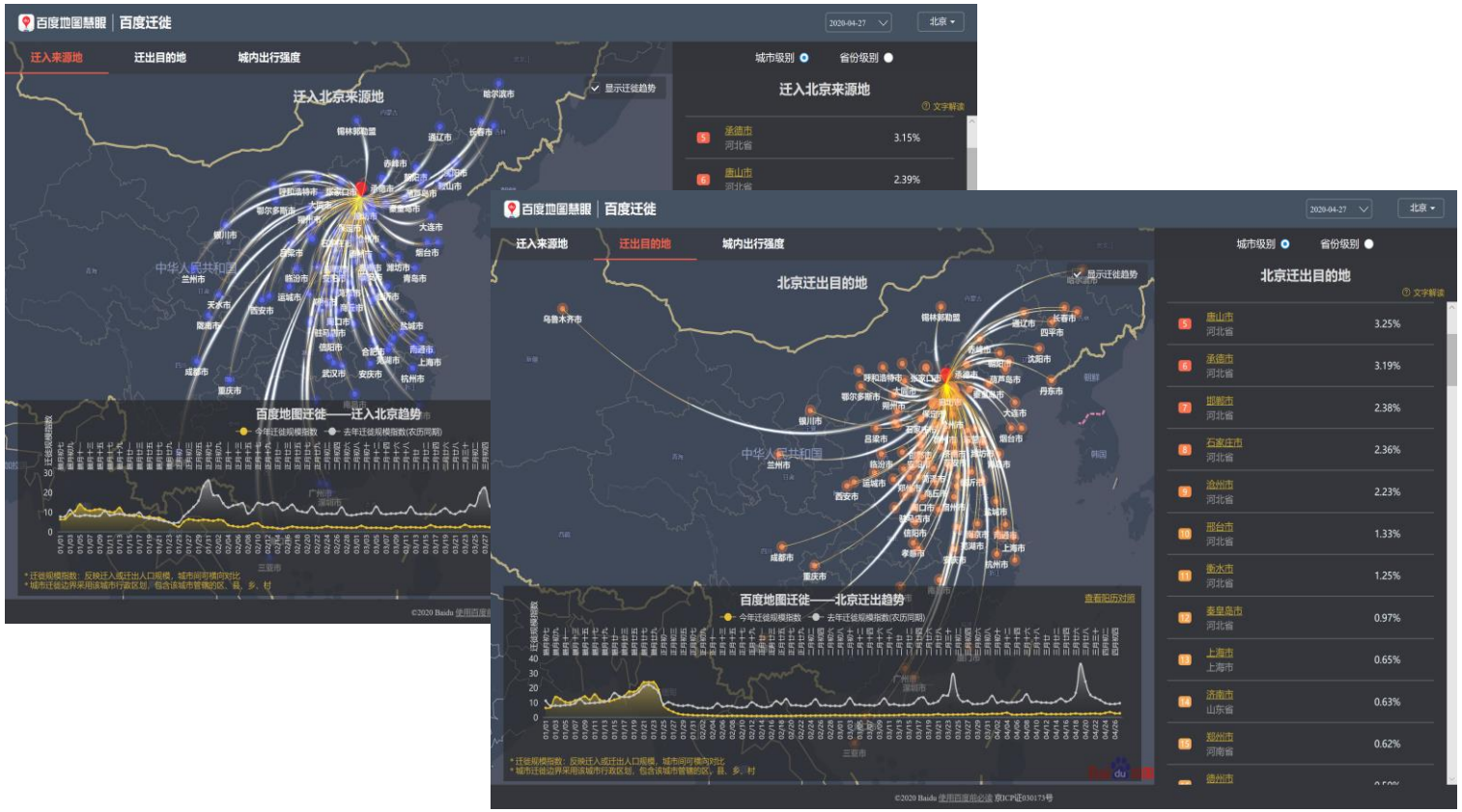
The proportion of people moving out of a city to other cities (Top100) in a day

City-level:

Field	Type
Place of immigration	Str
Province of the immigration	Str
Place to move out	Str
Province of the Emigration	Str
Time	yyyy-MM-dd
Proportion of population	Float

Province-level:

Field	Type
Province of the immigration	Str
Province of the Emigration	Str
Time	yyyy-MM-dd
Proportion of population	Float

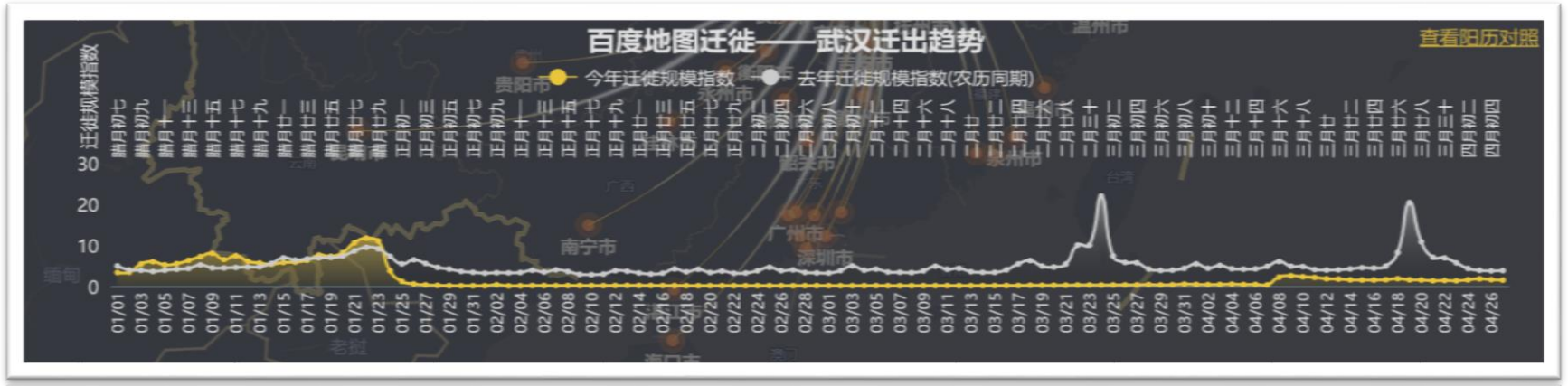


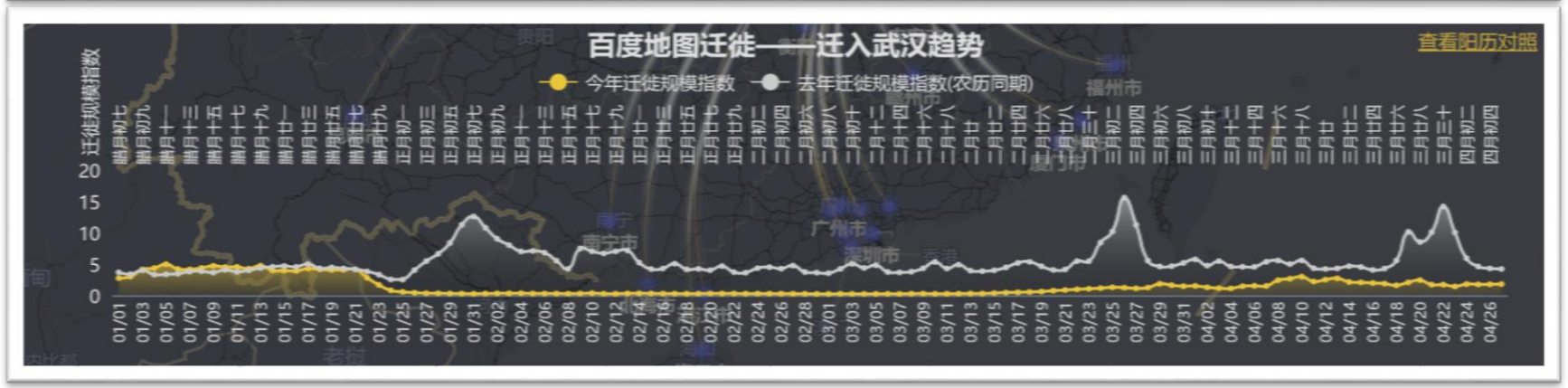
Social Big Data: Mobility Data

Mobility data

Migration trend: the change of the number (scale) of immigrants in a city over time

Emigration trend: the change of the number (scale) of emigration in a city over time

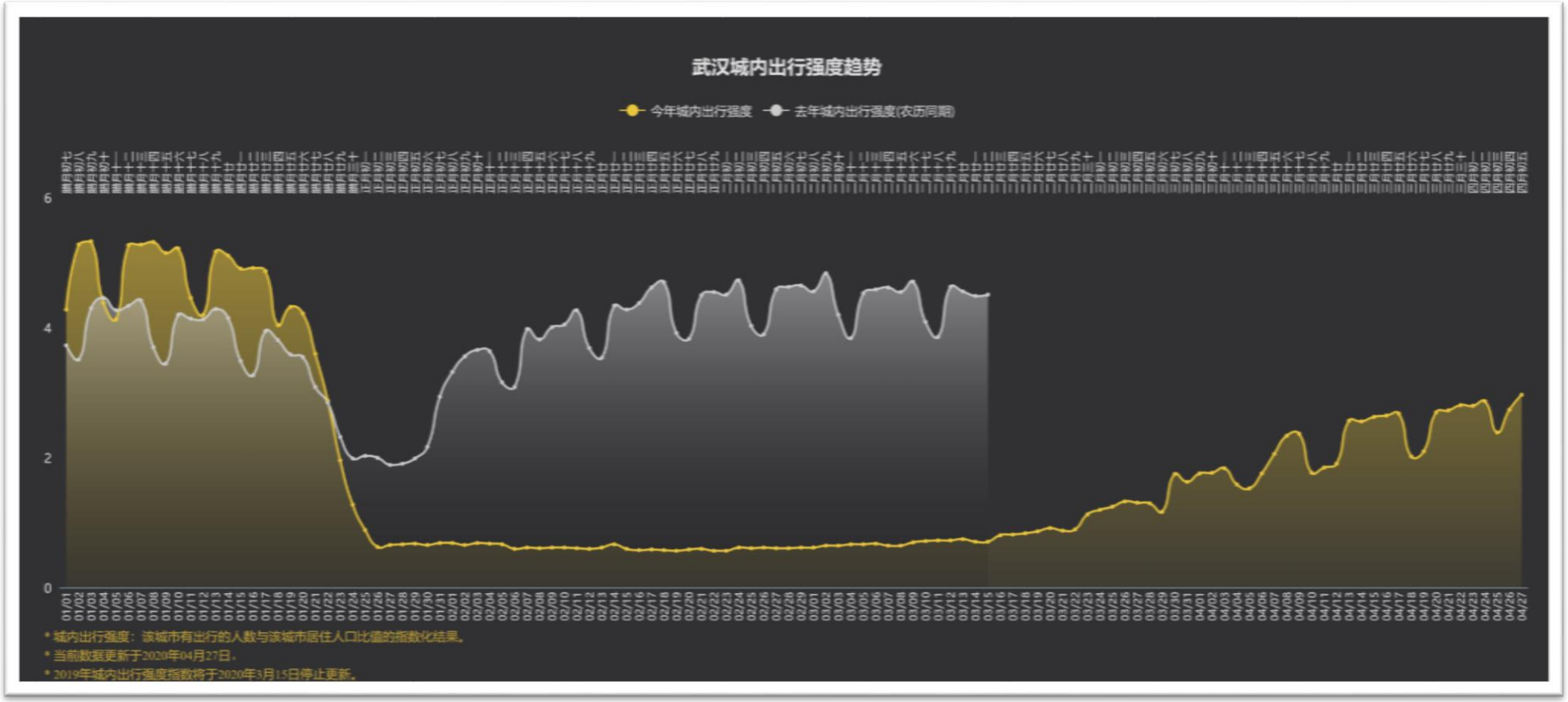




Social Big Data: Mobility Data

Mobility data

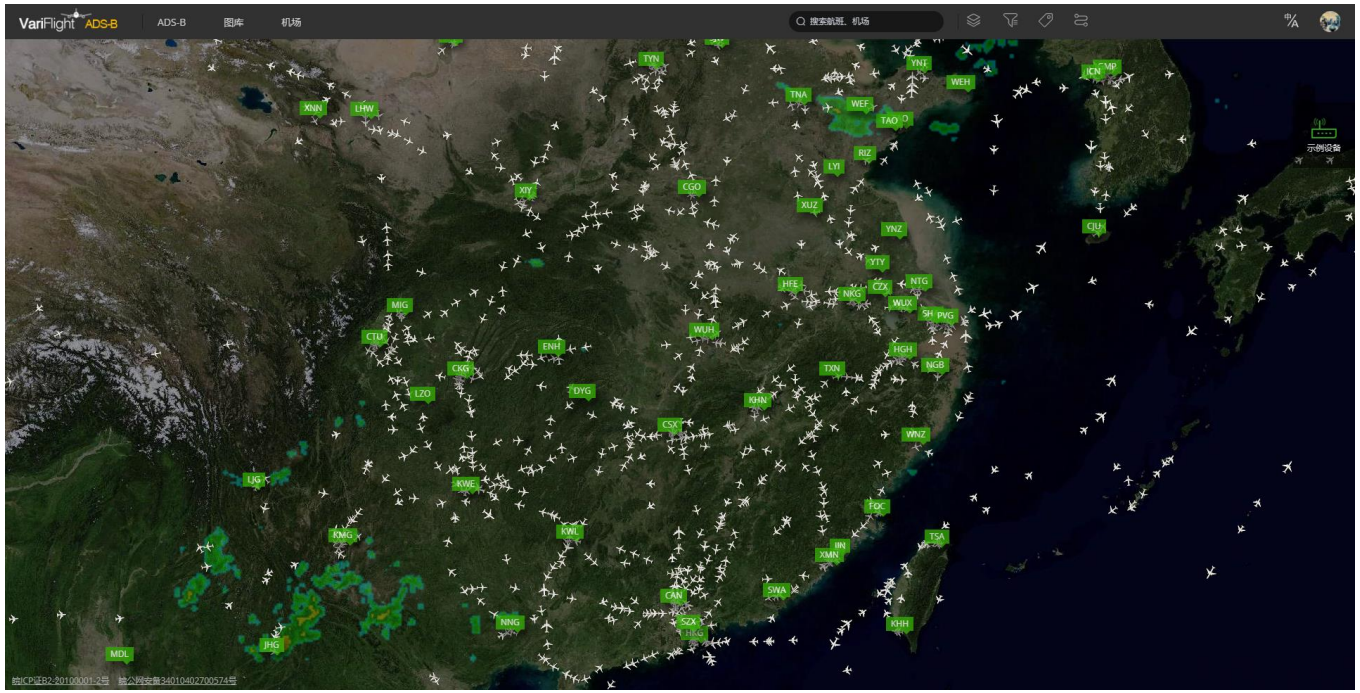
Travel intensity: the index result of the ratio of the number of people who travel in the city to the resident population in the city

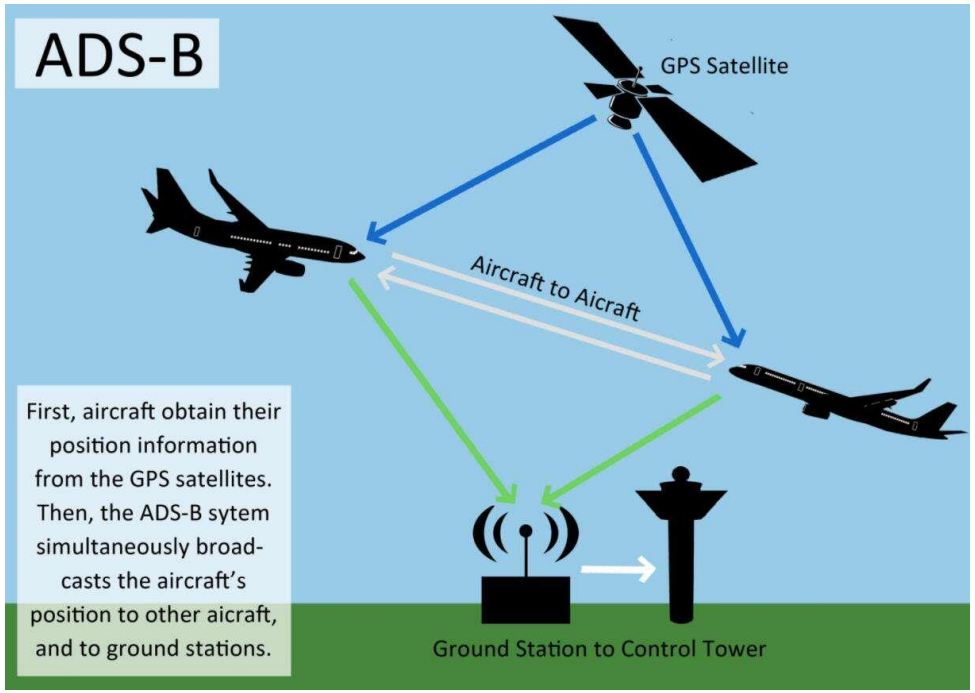


Social Big Data: Flight Data

Airline flight data

- The historical flight path data from VariFlight ADS-B (Automatic Dependent Surveillance - Broadcast) (https://flightadsb.variflight.com/)
- The airport data from OpenFlights (https://openflights.org/data.html)

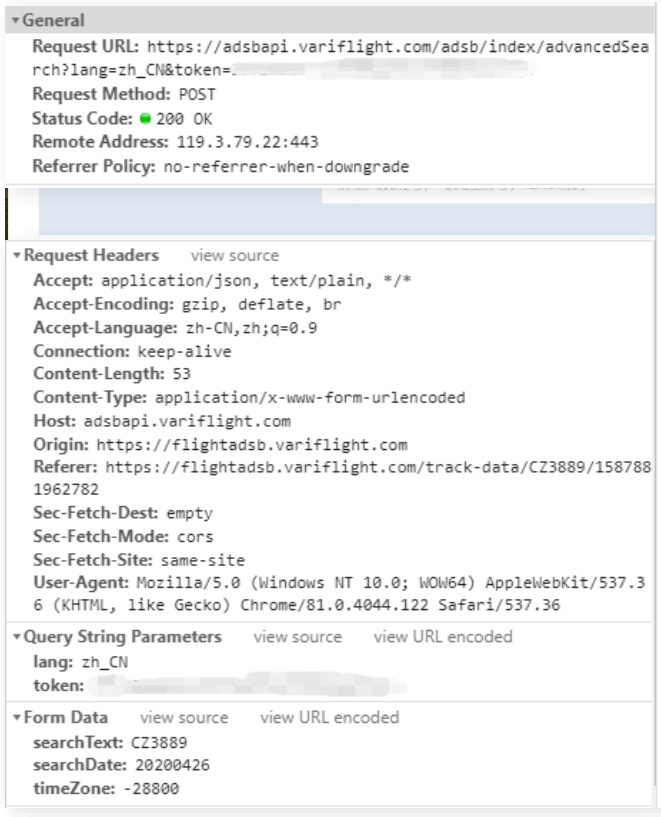




Social Big Data: Flight Data

Airline flight data

Observe the rules of web page interaction	Switch flight number; switch date.
Find the correct response data	Use the Google Chrome Development Tool.
Determine the request parameters	Observe the request header to determine crawling method.
Use python program to crawl data	Except crawling data, program also transform the json to the excel
Request result output	Excel file output



Source: https://flightadsb.variflight.com/track-data

Social Big Data: Train Data

High-speed train data

站序	车站	车次	出发时间 到达时间	历时	商务座 特等座	一等座	二等座	高级软卧 （下）	软卧 （下）	动卧	硬卧 （中）	软座	硬座	无座	其他
1	上海虹桥	G10	10:00 ----	----	--	--	--	--	--	--	--	--	--	--	--
2	南京南	G10	11:02 11:00	01:00 当日到达	--	--	--	--	--	--	--	--	--	--	--
3	济南西	G10	13:01 12:59	02:59 当日到达	--	--	--	--	--	--	--	--	--	--	--
4	北京南	G10	---- 14:28	04:28 当日到达	--	--	--	--	--	--	--	--	--	--	--

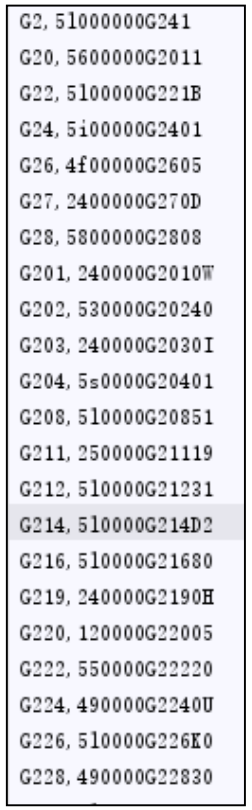
Website: http://www.12306.cn

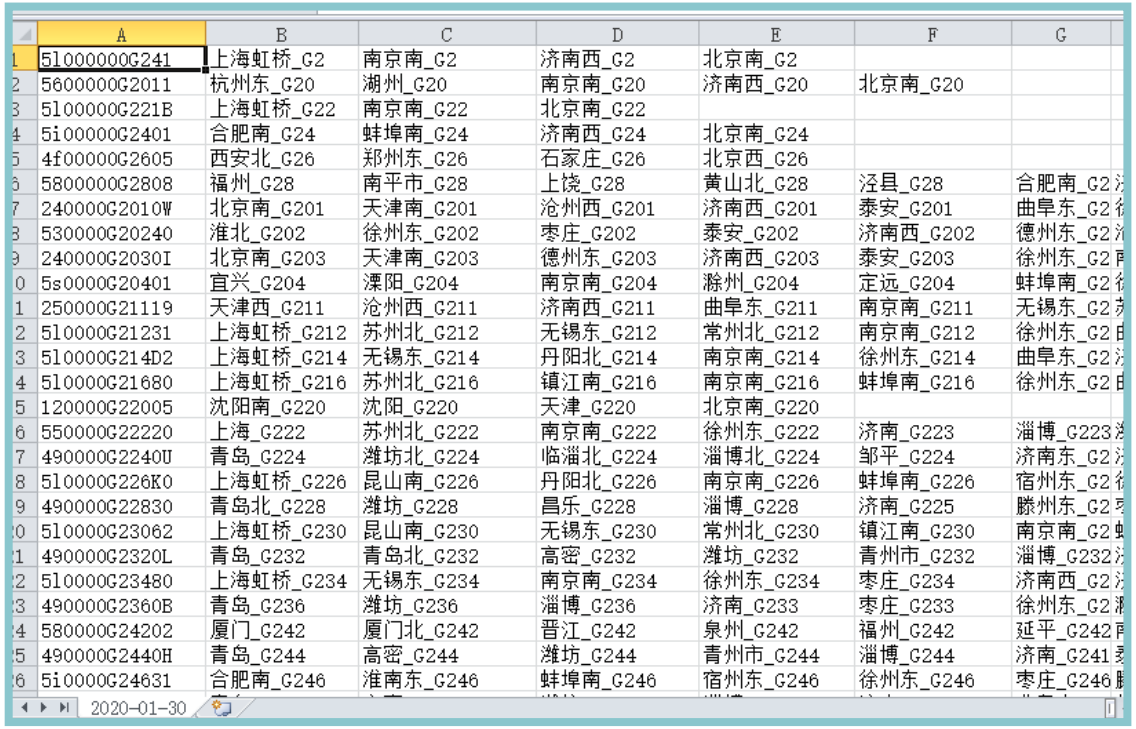
Social Big Data

High-speed train data

Search train No. by keywords and date and get as much Train No. as possible.

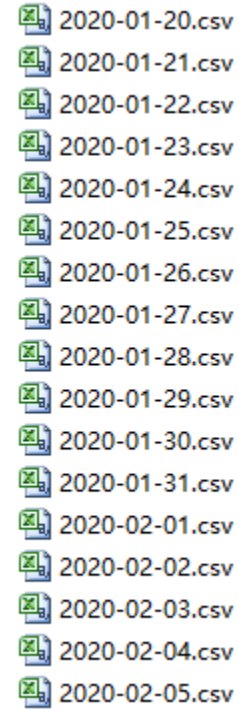
2200+ train No. have been collected.





	A	B	C	D	E	F	G
1	51000000G241	上海虹桥_G2	南京南_G2	济南西_G2	北京南_G2		
2	5600000G2011	杭州东_G20	湖州_G20	南京南_G20	济南西_G20	北京南_G20	
3	5100000G221B	上海虹桥_G22	南京南_G22	北京南_G22			
4	5i00000G2401	合肥南_G24	蚌埠南_G24	济南西_G24	北京南_G24		
5	4f00000G2605	西安北_G26	郑州东_G26	石家庄_G26	北京西_G26		
6	5800000G2808	福州_G28	南平市_G28	上饶_G28	黄山北_G28	泾县_G28	合肥南_G2
7	240000G2010W	北京南_G201	天津南_G201	沧州西_G201	济南西_G201	泰安_G201	曲阜东_G2
8	530000G20240	淮北_G202	徐州东_G202	枣庄_G202	泰安_G202	济南西_G202	德州东_G2
9	240000G2030I	北京南_G203	天津南_G203	德州东_G203	济南西_G203	泰安_G203	徐州东_G2
10	5s0000G20401	宜兴_G204	溧阳_G204	南京南_G204	滁州_G204	定远_G204	蚌埠南_G2
11	250000G21119	天津西_G211	沧州西_G211	济南西_G211	曲阜东_G211	南京南_G211	无锡东_G2
12	510000G21231	上海虹桥_G212	苏州北_G212	无锡东_G212	常州北_G212	南京南_G212	徐州东_G2
13	510000G214D2	上海虹桥_G214	无锡东_G214	丹阳北_G214	南京南_G214	徐州东_G214	曲阜东_G2
14	510000G21680	上海虹桥_G216	苏州北_G216	镇江南_G216	南京南_G216	蚌埠南_G216	徐州东_G2
15	120000G22005	沈阳南_G220	沈阳_G220	天津_G220	北京南_G220		
16	550000G22220	上海_G222	苏州北_G222	南京南_G222	徐州东_G222	济南_G223	淄博_G223
17	490000G2240U	青岛_G224	潍坊北_G224	临淄北_G224	淄博北_G224	邹平_G224	济南东_G2
18	510000G226K0	上海虹桥_G226	昆山南_G226	丹阳北_G226	南京南_G226	蚌埠南_G226	宿州东_G2
19	490000G22830	青岛北_G228	潍坊_G228	昌乐_G228	淄博_G228	济南_G225	滕州东_G2
20	510000G23062	上海虹桥_G230	昆山南_G230	无锡东_G230	常州北_G230	镇江南_G230	南京南_G2
21	490000G2320L	青岛_G232	青岛北_G232	高密_G232	潍坊_G232	青州市_G232	淄博_G232
22	510000G23480	上海虹桥_G234	无锡东_G234	南京南_G234	徐州东_G234	枣庄_G234	济南西_G2
23	490000G2360B	青岛_G236	潍坊_G236	淄博_G236	济南_G233	枣庄_G233	徐州东_G2
24	580000G24202	厦门_G242	厦门北_G242	晋江_G242	泉州_G242	福州_G242	延平_G242
25	490000G2440H	青岛_G244	高密_G244	潍坊_G244	青州市_G244	淄博_G244	济南_G241
26	5i0000G24631	合肥南_G246	淮南东_G246	蚌埠南_G246	宿州东_G246	徐州东_G246	枣庄_G246

2020-01-30

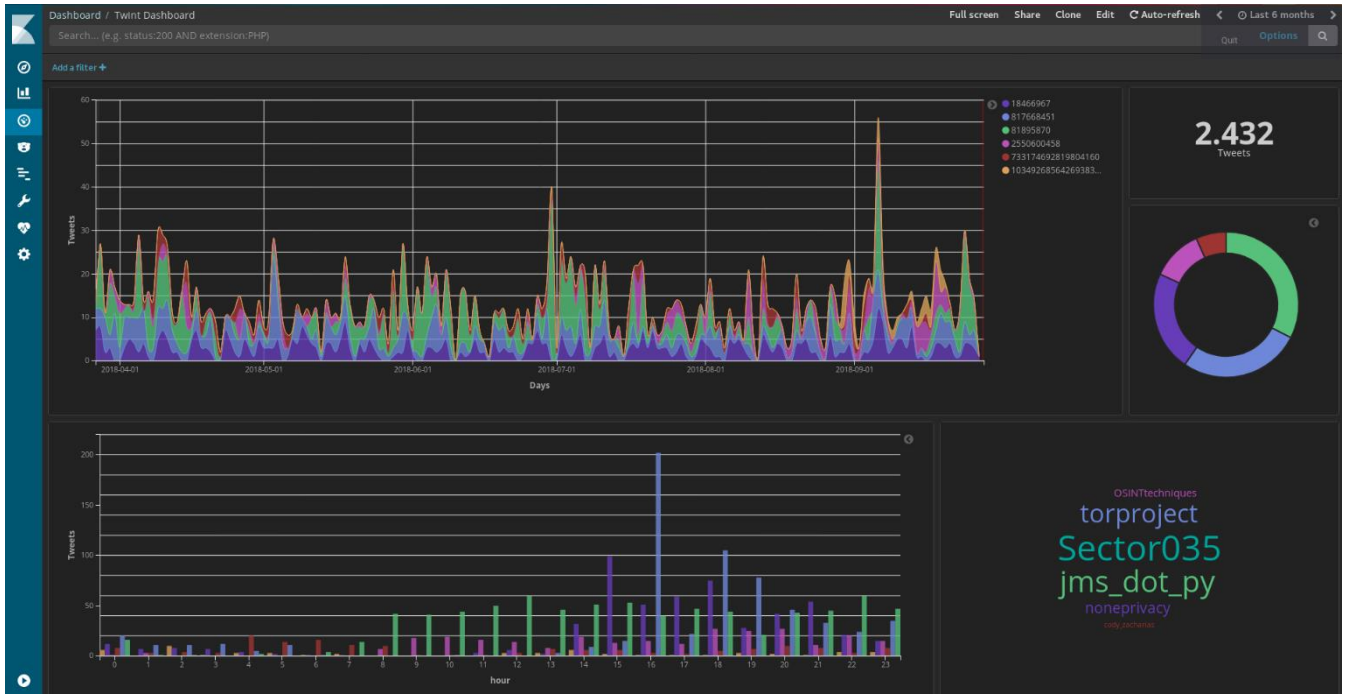


Train Data: 2020-01-20 to 2020-03-31

Social Big Data: Social Media Data

Social media data

- An advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles **without** using Twitter's API.
- Utilize Twitter's search operators to let scrape Tweets from specific users, scrape certain topics, hashtags & trends, or sort out *sensitive* information from Tweets like e-mail and phone numbers.



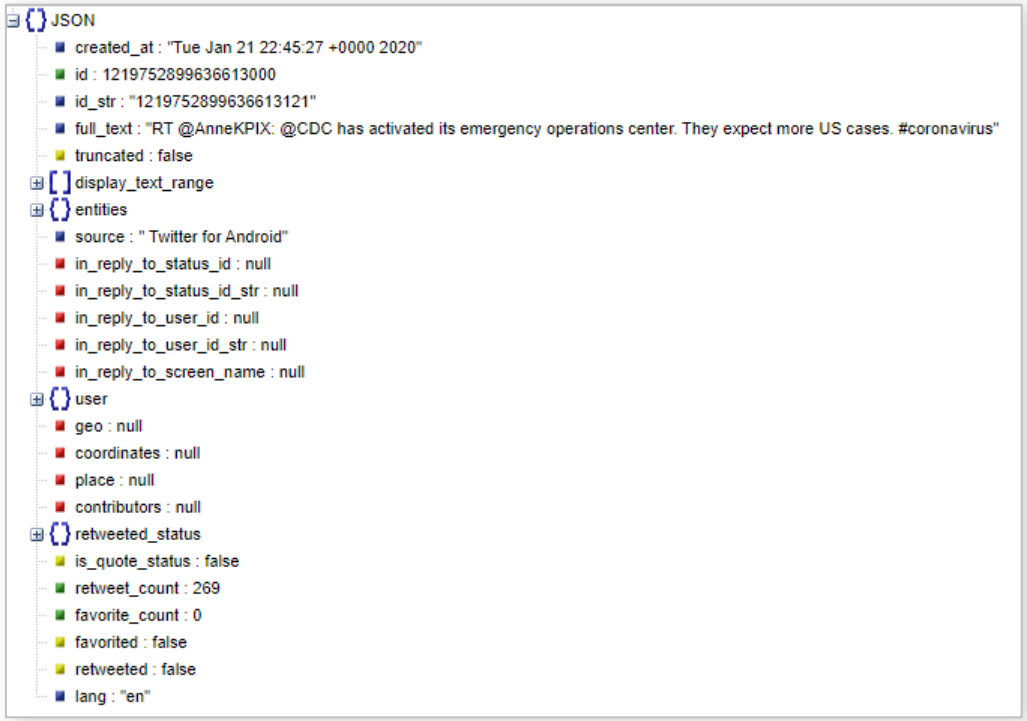
Twint: https://github.com/twintproject/twint

Social Big Data: Social Media Data

Social media data

Public COVID-19-TweetIDs: https://github.com/echen102/COVID-19-TweetIDs

Convert JSON to CSV



```
'tweet_id',
'time',
'content',
'reply_to_userid',
'reply_to_username',
'source',
'geo',
'retweet_count',
'user_id',
'user_name',
'user_location',
'user_description',
'user_followers',
'user_friends',
'user_created_time',
'user_favourites',
'user_geo_enabled',
'user_lang',
'user_statuses',
'user_verified',
```

```
'retweet_tweet_id',
'retweet_created_time',
'retweet_content',
'retweet_source',
'retweet_reply_to_userid',
'retweet_reply_to_username',
'retweet_retweets',
'retweet_favorites',
'retweet_lang',
'retweeted_user_info',
'retweet_user_id',
'retweet_user_name',
'retweet_user_location',
'retweet_user_description',
'retweet_user_followers',
'retweet_user_friends',
'retweet_user_create_time',
'retweet_user_geo_enabled',
'retweet_user_favourites',
'retweet_user_statuses',
'retweet_user_lang',
'retweet_user_verified'
```

Limitations: With no geographic information; Delayed dataset updates

Social Big Data: Social Media Data

Social media data

Geo-Tweets from the Center for Geographic Analysis at Harvard

Descriptions: https://doi.org/10.7910/DVN/3NCMB6

Geo-tweet Archive Geo-located tweets or "geo-tweets" since late 2012.

Approximately 1% of tweet records contain geographic coordinates.

Geo-tweet Records:

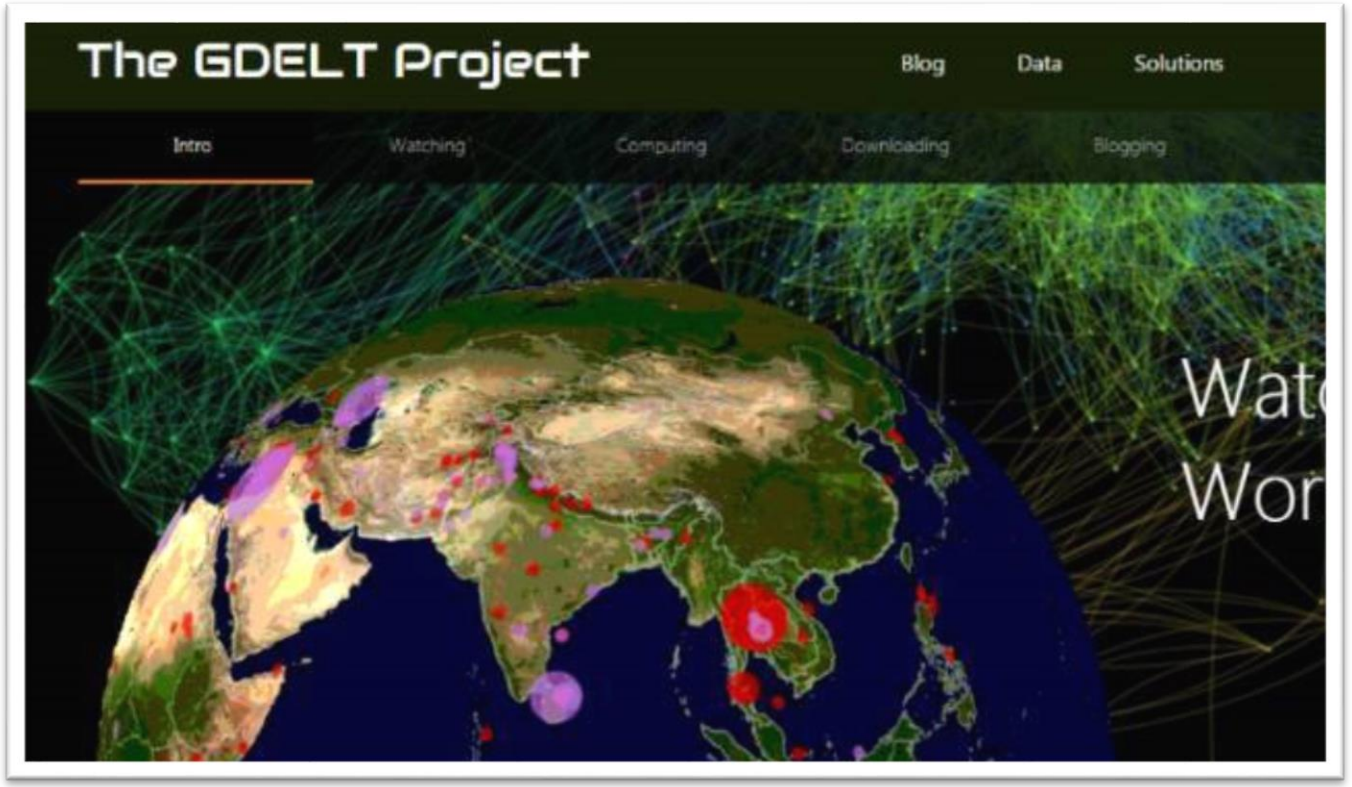
- tweet_id (bigint)
- time (timestamp without time zone)
- lat (double precision)
- lon (double precision)
- goog_x (double precision)
- goog_y (double precision)
- sender_id (integer)
- sender_name (character varying(50))
- source (character varying(200))
- reply_to_user_id (integer)
- reply_to_tweet_id (bigint)
- place_id (character varying(40))
- tweet_text (character varying(500))

Social Big Data: Global News Data

Global News data

The GDELT Project monitors the world's broadcast, print, and web news from nearly every country in over 100 languages and identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images, and events.

It offers a free open platform for public.



Source: https://www.gdeltproject.org

4. CEOS WGISS CWIC

- Heterogeneity among existing data providers
  - Different web portal
  - Different query language and queryables
  - Different metadata model
  - Different data access method

ACADEMY OF OPTO-ELECTRONICS
CHINESE ACADEMY OF SCIENCES

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

NASA

LTSRF
GHRSST Long Term Stewardship and Reanalysis Facility
NOAA NODC

USGS
science for a changing world

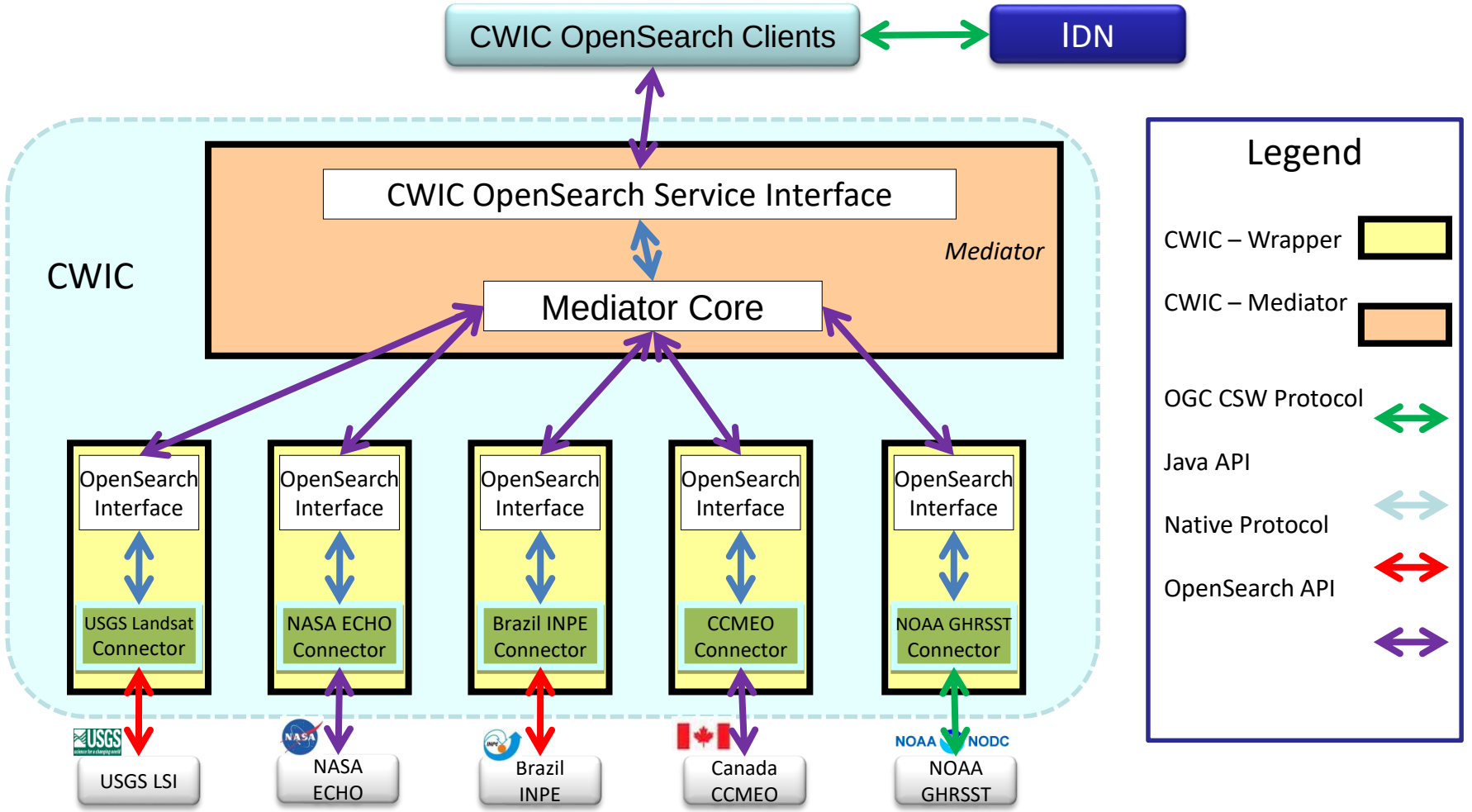
INPE

Government of Canada
Gouvernement du Canada
Canada

इसरो isro

Soure: http://cwic.wgiss.ceos.org

4. CEOS WGISS CWIC



Soure: http://cwic.wgiss.ceos.org

Web Sites

Public Data: http://chinadatalab.net

Cloud Platform: http://chinadatalab.org