

Existing Open Data Practices in Astro- and Particle Physics

Serguei Vorobiov, CAC-UNG

**Open Data Workshop
UNG campus, Ajdovščina, November 6th, 2024**

FAIR data vs open data

“**Open data** is data that is openly accessible, exploitable, editable and shareable by anyone for any purpose. Open data is licensed under an open license” (Wikipedia)

FAIR data (formulation in 2016):

Findable: Data should be easy to find for both humans and computers.

Accessible: Once found, data must be accessible and retrievable.

Interoperable: Data should be integrated with other datasets and systems.

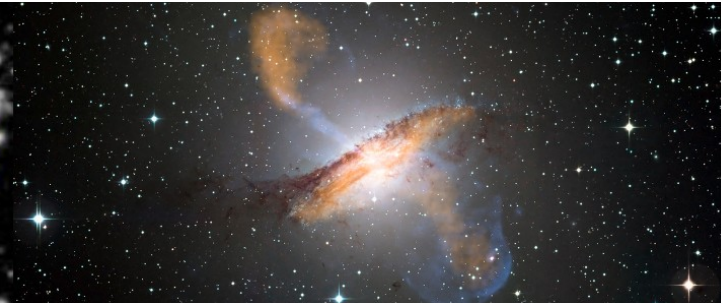
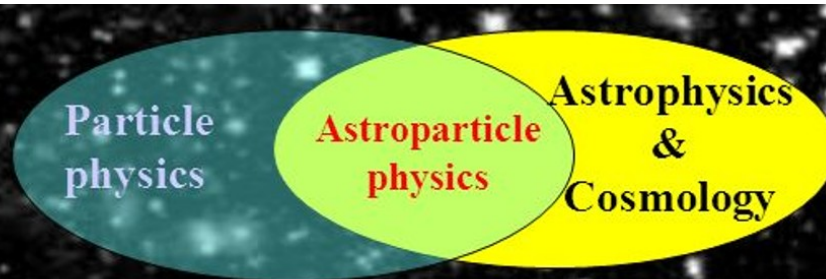
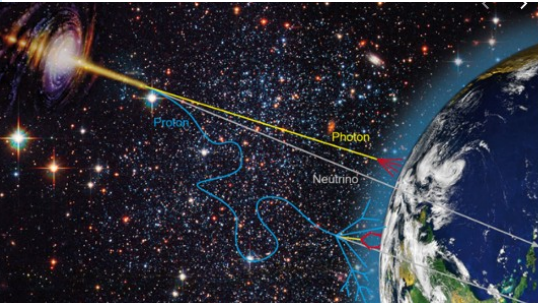
Reusable: Data must be reusable with clear licenses and provenance.

- Open data concept emphasizes **openness**, FAIR data concept – **quality and usability**
- While FAIR data may be open, it could also be restricted or controlled access

Astroparticle field: linking Universe & particles

Astroparticle physics, field at the interface between astrophysics and particle physics:

- studying radiation and particles of astronomical origin by means of experiments based on elementary particle detector techniques.
- probing astrophysics, cosmology and fundamental interactions by ground-based, underground (under water, under ice) and space-based detectors.
- detecting cosmic rays, gamma rays, neutrinos, looking for their sources, studying their production/propagation mechanisms, verifying the astrophysical models.
- **open data** is instrumental in fostering the **multimessenger astronomy** – the coordinated observation of cosmic events using multiple types of “messenger” signals: photons, neutrinos, cosmic rays, and gravitational waves – enabling faster, richer scientific discoveries.

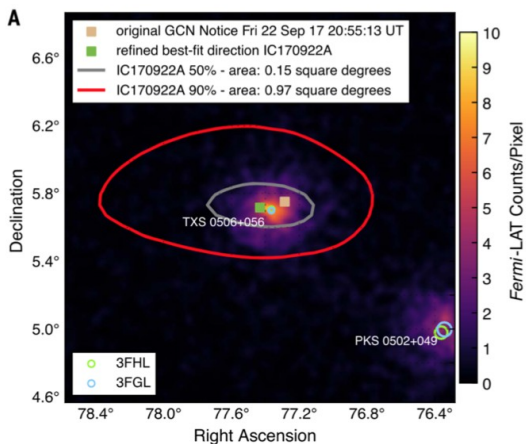
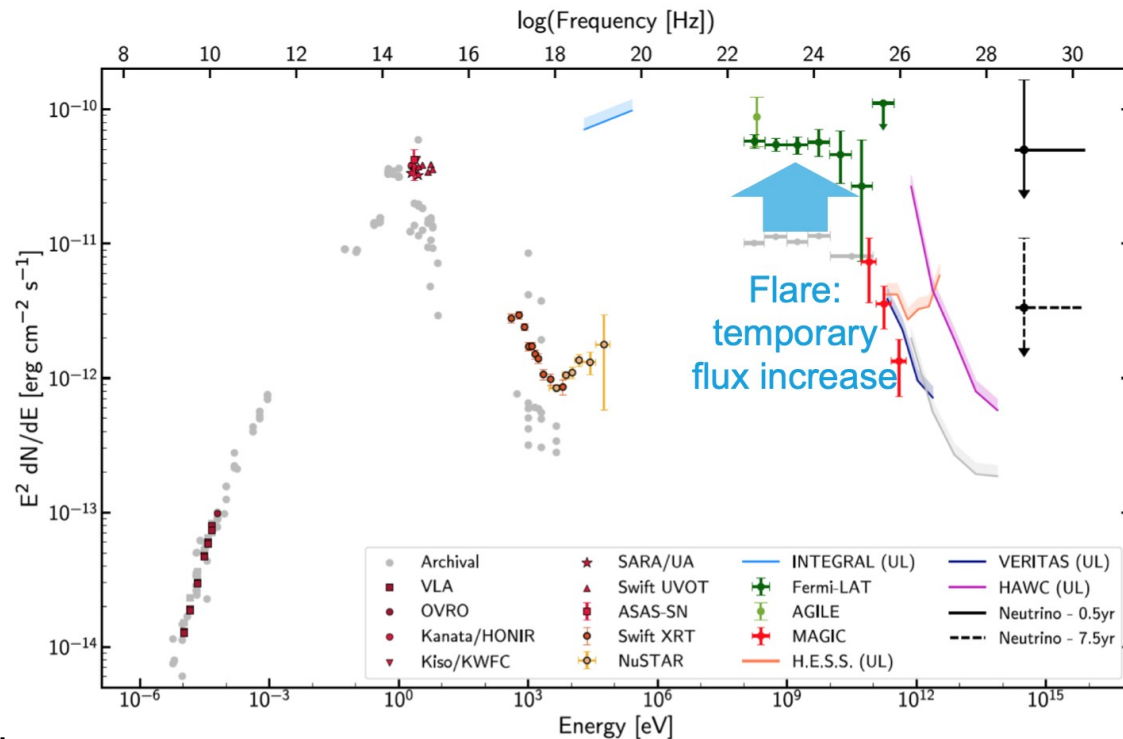
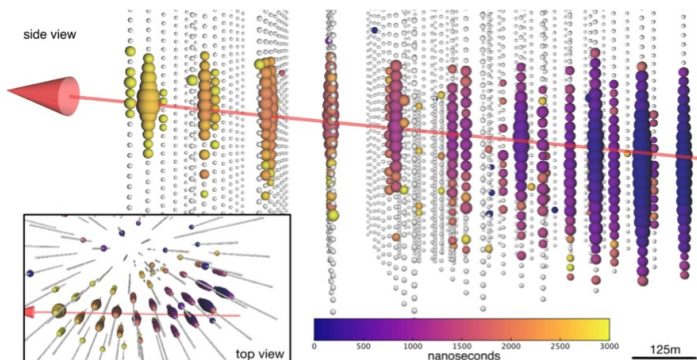


IceCube neutrino from the flaring AGN TXS 0506+056

September 22nd, 2017:

A ~290 TeV neutrino coincident with a blazar flare

An alert distributed 43 seconds after neutrino event



GCN Circular issued 4 hours later with refined direction

Observed blazar flare (Fermi-LAT, MAGIC), 3 σ correl.

$$z = 0.3365 \pm 0.0010$$

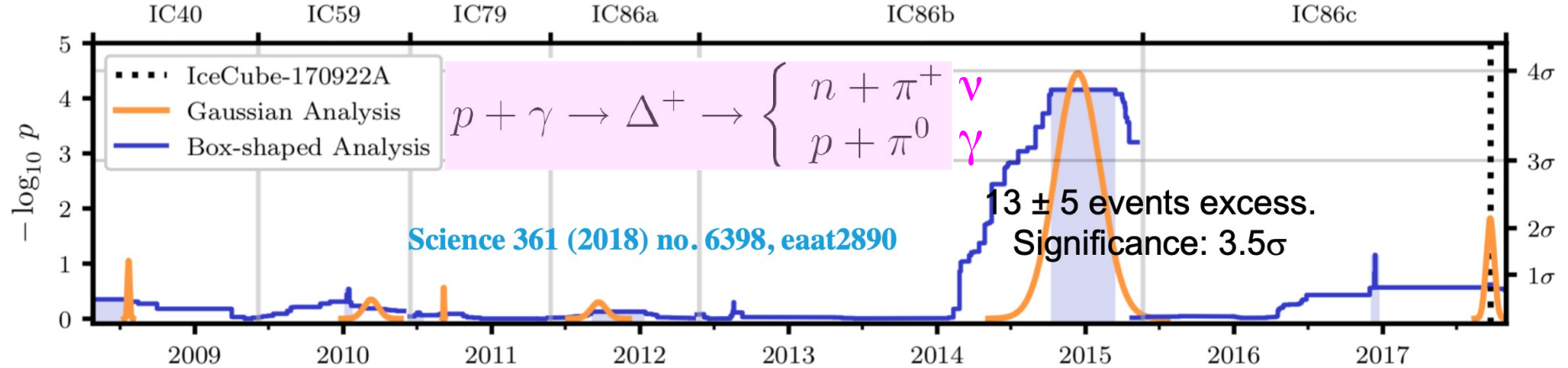
Paiano et al, 2018

Follow-up EM campaign: color, archival data: gray
First known source of VHE astrophysical neutrinos

Science 361 (2018) no. 6398, eaat1378

Analysis of IceCube archival neutrino data

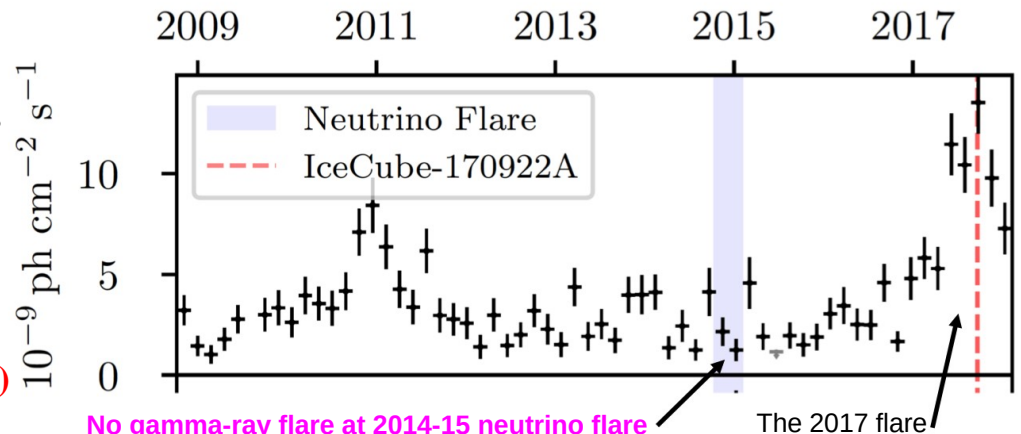
Archival neutrino data search reveals an “orphan” neutrino flare from TXS 0506+056 in 2014-15



Padovani et al, MNRAS 480 (2018) 192
 (UN Open Universe initiative/Virtual Observatory protocols)
 TXS 0506+056 is the only counterpart of all neutrino emissions

Britzen et al, A&A 630 (2019) A103
 (archival VLBA 15 GHz observations between 2009 and 2018)
 TXS 0506+056: a curved jet or two jets in collision

Lipunov et al, ApJL 896 (2020) L19
 (archival 16 yr MASTER telescope network data; Gaia catalog)
 The 2017 flare: anticorrelation of the optical and neutrino flux



Astrophysics: pioneer in open data practices

1970s: the groundwork for **shared data** laid by observatory projects and surveys such as **NRAO's VLA**. **FITS (Flexible Image Transport System) format** development since 1979, 1st standard released in 1981.

1983: **Infrared Astronomical Satellite (IRAS)**, the first mission to openly publish datasets, ultimately made freely available, standardizing the practice for space missions.

since 2000: **Sloan Digital Sky Survey (SDSS)** releasing comprehensive open datasets, including images, spectra, and photometry, accessible through a simple interface.

2002: **Virtual Observatory (VO)** initiative in order to develop standards and protocols for the exchange and re-use of digital data collections. Further boosted with **open source tools** (Astropy, PyVO etc.).

since 2009: **NASA's Kepler mission** launching and releasing time-series data on star brightness levels – data instrumental for exoplanet discovery, encouraging professional astronomers and **citizen scientists** to contribute to findings.

since 2013: **ESA's Gaia mission** adopted an open data policy, with periodic data releases providing the most precise star maps to date.

International Virtual Observatory Alliance (IVOA)

- Formed in 2002
- 23 VO members
- >50 data centers adhering to VO standards
- Working groups: Applications, Data Access Layer, Data Modelling, Grid and Web Services, Resource Registry, Semantics, VO Query Language, VOTable;
- Several Interest Groups



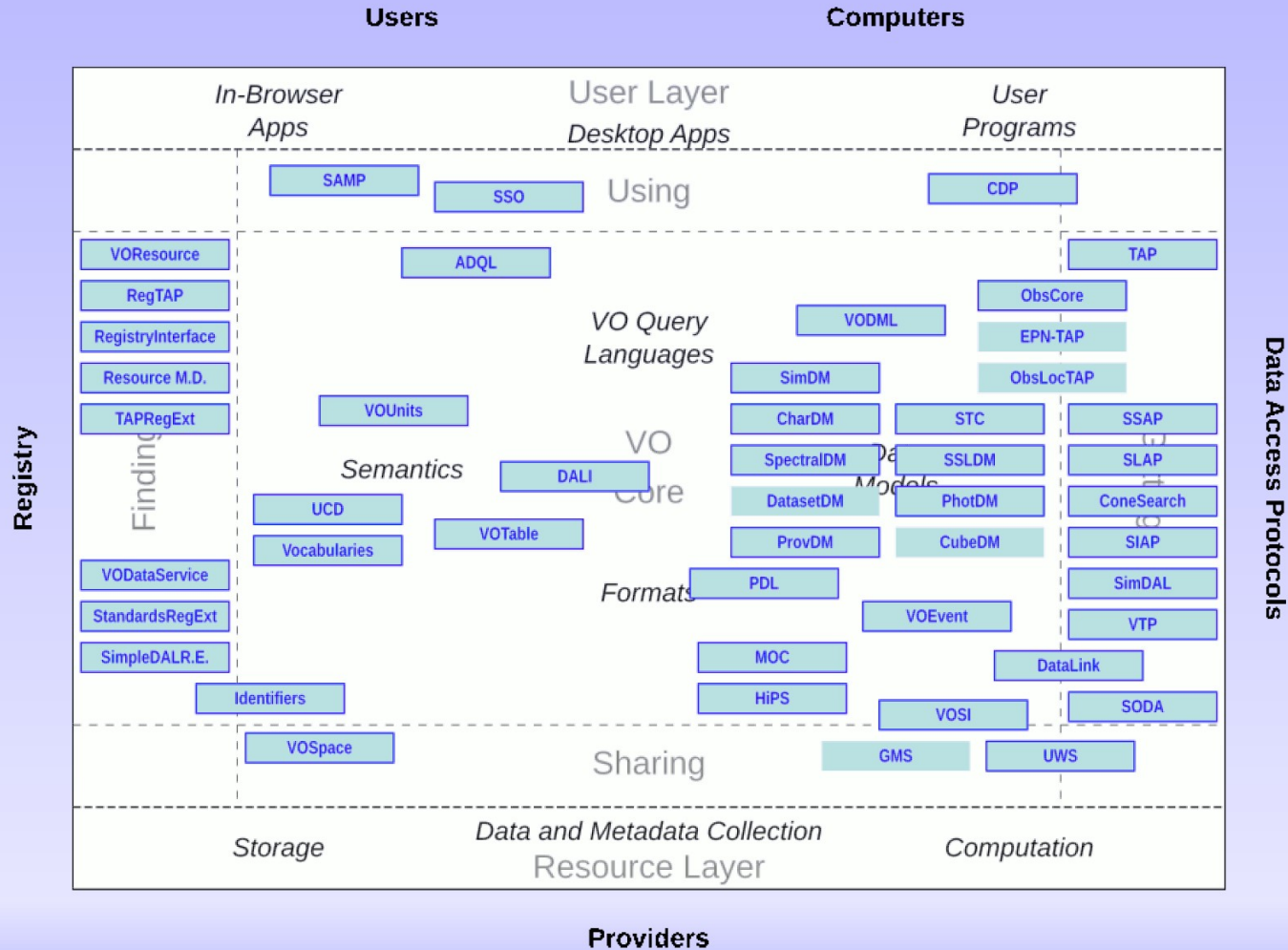
- Argentine Virtual Observatory
- Armenian Virtual Observatory
- AstroGrid, United Kingdom
- Australian All-Sky Virtual Observatory
- Brazilian Virtual Observatory
- Chinese Virtual Observatory
- Canadian Virtual Observatory
- Chilean Virtual Observatory
- European Space Agency
- European Virtual Observatory
- German Astrophysical Virtual Observatory
- Japanese Virtual Observatory
- Kazakhstan Virtual Observatory
- Netherlands Virtual Observatory
- Observatoire Virtuel France
- Russian Virtual Observatory
- Square Kilometer Array Observatory
- South African Astroinformatics Alliance
- Spanish Virtual Observatory
- Italian Virtual Observatory
- Ukrainian Virtual Observatory
- US Virtual Observatory Alliance
- Virtual Observatory India

Astronomy + space physics + geosciences

<https://ivoa.net/>

VO standards

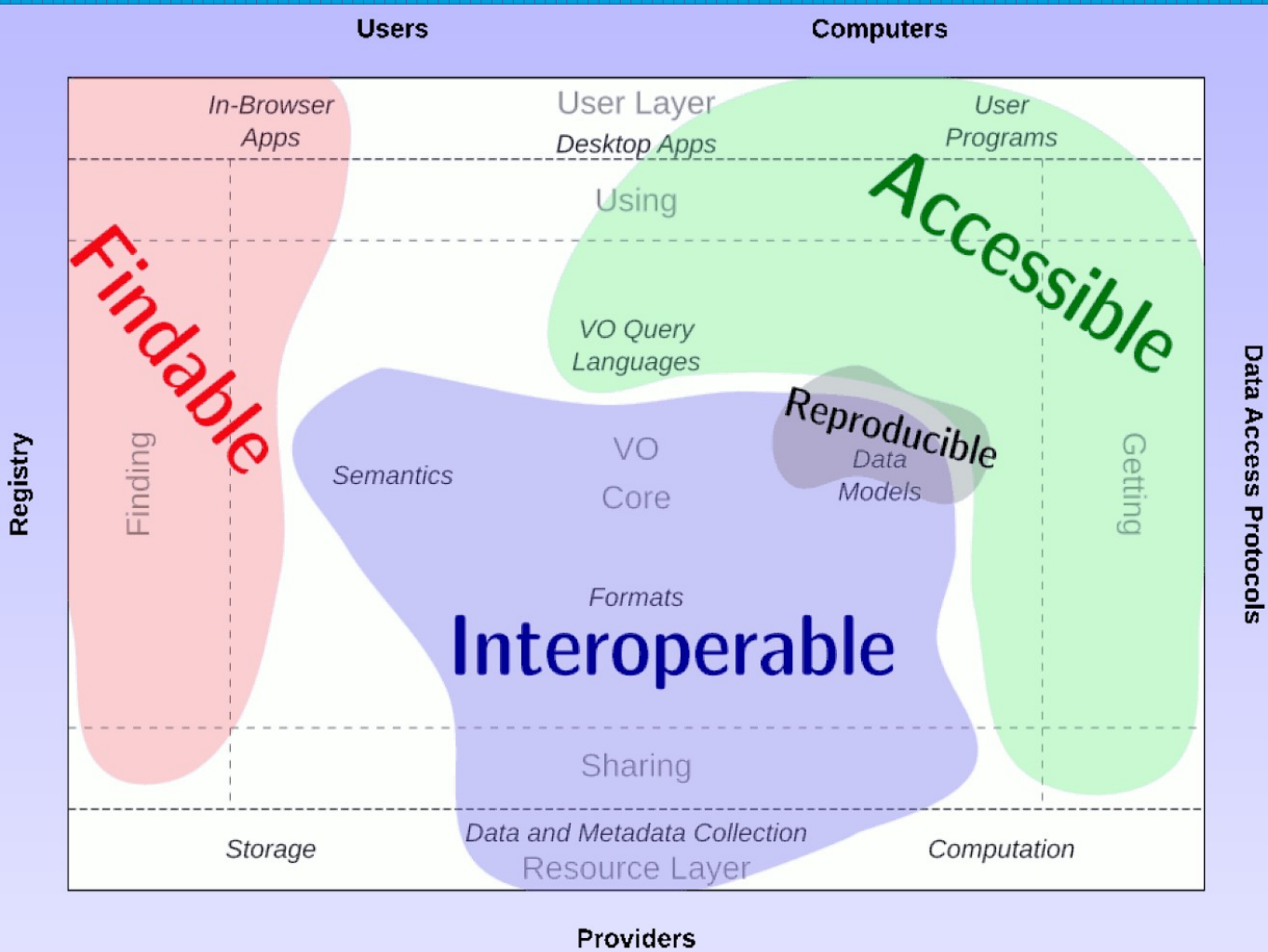
- **Findable:** VO Registry (a big collection of detailed metadata)
- **Accessible:** Query protocols tailored to specific data types (e.g. SIAP for images, SSAP for spectra, SCS for catalogue data) or of general purpose
- **Interoperable:** rich metadata in VOTable, light semantics, data models, machine-readable unit strings, agreed-upon vocabularies, Authentication & Authorisation, ...
- **Reusable (Reproducible):** Provenance data model



VO standards mapping to the FAIR principles

- **Findable:** data centers publish registries that are gathered by full registries
- **Accessible:** clients access data by querying the VO services via SQL-like queries
- **Interoperable:** a set of interoperability conventions on several levels
- **Reusable (Reproducible):** Provenance data model

Data centers offering about 30000 individual resources, publishing 100s of millions of data sets for clients (TOPCAT, Aladin), libraries (pyVO) or web pages (ESAsky, ...)

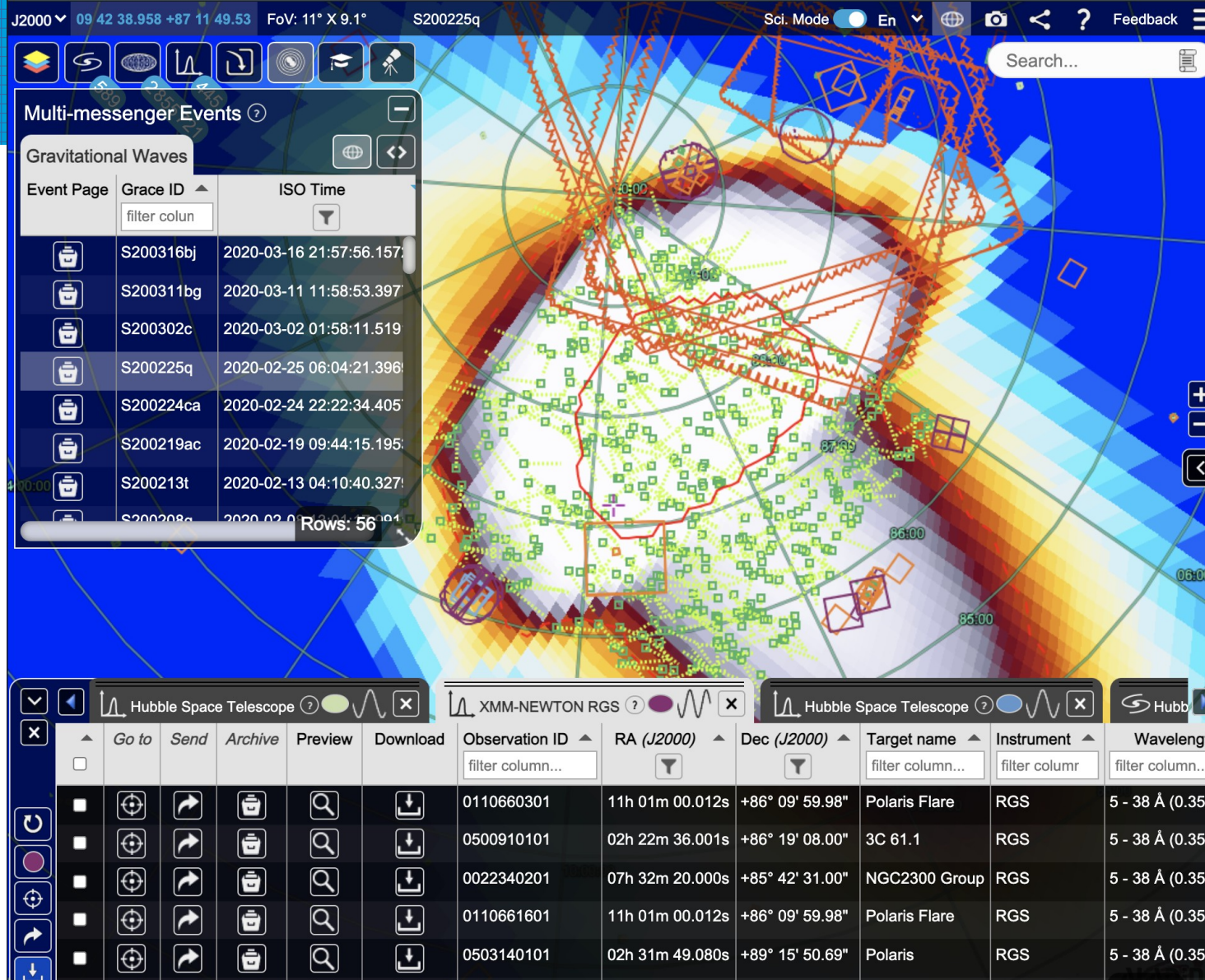


ESASky

Multi-messenger astronomy with ESASky (from IVOA Newsletter - March 2022)

ESASky: tool for exploring the multi-wavelength sky

- Included gravitational wave (GW) event probability maps on the sky, allowing users to look for electromagnetic counterparts for the GW events and using ESASky to quickly access all available archival electromagnetic data
- Plans to include the IceCube neutrino footprint and other multi-messenger data



Particle physics: more recent shift to openness

- 1990s:** At **CERN**, data sharing within collaborations established, public data access minimal due to the competitive nature and complexity of the experiments. First discussions on data-sharing principles.
- 2010:** **Large Hadron Collider (LHC) Open Data Initiative** launched by **CERN**, aiming to release public datasets from LHC experiments such as ATLAS and CMS.
- 2014:** **CERN** launched the **Open Data Platform**, making CMS data from 2010-2011 runs (2 PBytes) available, providing access to collision data and simulations as well as to original analysis software.
- 2019:** The **American Physical Society (APS)** endorsed data sharing in particle physics and related fields, encouraging researchers to adopt **FAIR** principles.
- 2020:** **CERN** adopted a formal **Open Data Policy** aiming at making all LHC data available via its Open Data platform <http://opendata.cern.ch>, with an embargo period between about 5-10 years. **Publications by non-collaboration members** currently at about five publications per year.
- 2021:** making available data from **CERN LEP** experiments (1990s), the **DESY PETRA** experiment JADE (1980s), some actions for the **DESY HERA** experiments' data (data taking ended in 2007) underway.

CERN Open Data policy for LHC experiments

<http://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments>

policy for the release of Open Data at the various data levels, **made public on Dec. 11, 2020:**

Level 1: Publications (open since a long time, e.g. arXiv, Inspire, open access journals), plus related information in machine readable form, as well as binned and unbinned likelihoods

Level 2: Education and Outreach, simplified derived data sets

Level 3: Fully calibrated reconstructed data sets as used internally by the collaborations for their analysis (mainly pioneered by CMS since 2014, see also <https://indico.cern.ch/event/882586/>), data to be partially released by the collaborations with a typical embargo time of about 5 years, with full release after 10 years or at the close of the collaborations (CMS and ALICE: since 2014, LHCb: 800 TB of Run 1 data (2011-2012) released in Dec. 2023, ATLAS: released their complete Run 2 data in 2020, with portions of their Run 1 data becoming available even earlier).

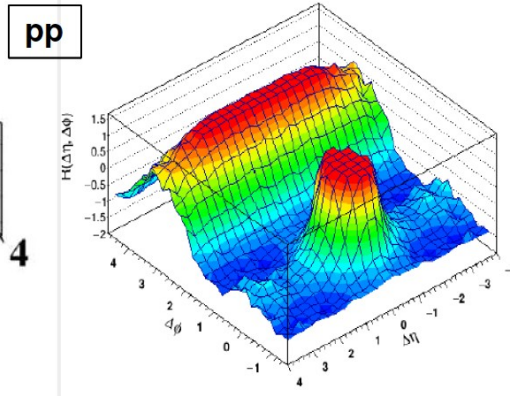
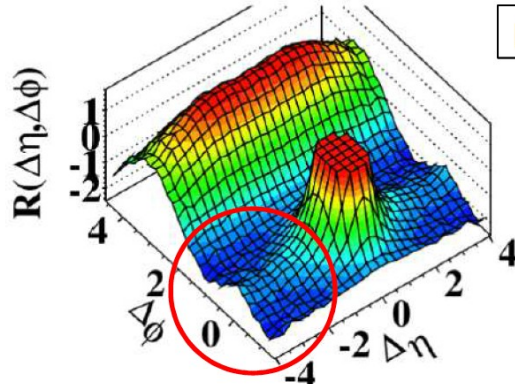
Level 4: Raw data: usually not useful to outsiders, mostly not being released

In a cross-experiment Open Data analysis, "Ridge" in long range particle correlations, most cited non-Higgs LHC result (JHEP 1009 (2010) 091) is reproducible on 2010 CMS Open Data

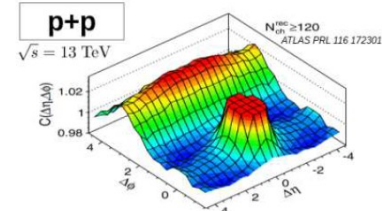
CMS Paper
JHEP 1009 (2010) 091

CMS Open Data
(summer student on office desktop)

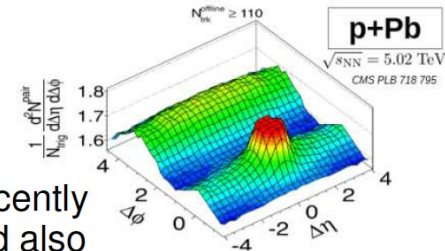
(d) CMS $N \geq 110$, $1.0 \text{ GeV}/c < p_T < 3.0 \text{ GeV}/c$



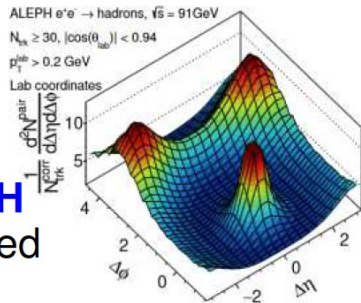
ALICE pp
Open Data
not yet
analyzed



CMS recently
released also
Heavy Ion
Open Data

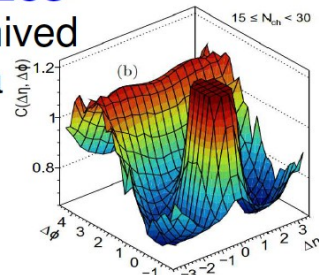


e⁺e⁻
ALEPH
archived
data



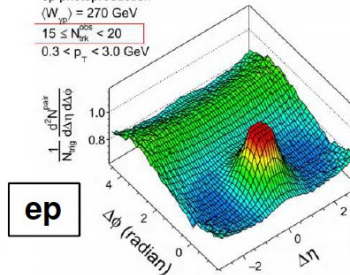
Phys.Rev.Lett.
123 (2019) 212002

ZEUS
archived
data

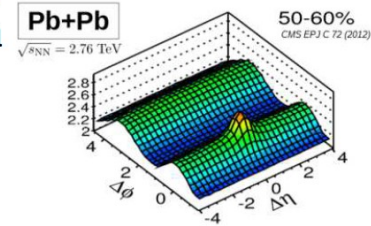


JHEP 04 (2020) 070

H1 Preliminary
ep photoproduction
(W_{γ*}) = 270 GeV
15 ≤ Nch < 20
0.3 < pT < 3.0 GeV

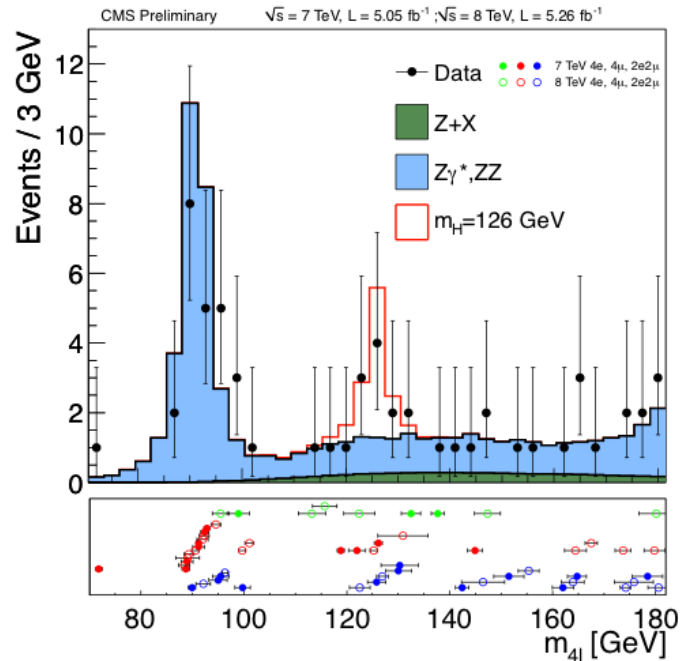


H1
archived
data





ROOT enables statistically sound scientific analyses and visualization of large amounts of data: today, more than 1 exabyte (1,000,000,000 gigabyte) are stored in ROOT files. [The Higgs was found with ROOT!](#)



As high-performance software, ROOT is written mainly in C++. You can use it on Linux, macOS, or Windows; it works out of the box. ROOT is [open source](#): use it freely, [modify it](#), [contribute to it!](#)

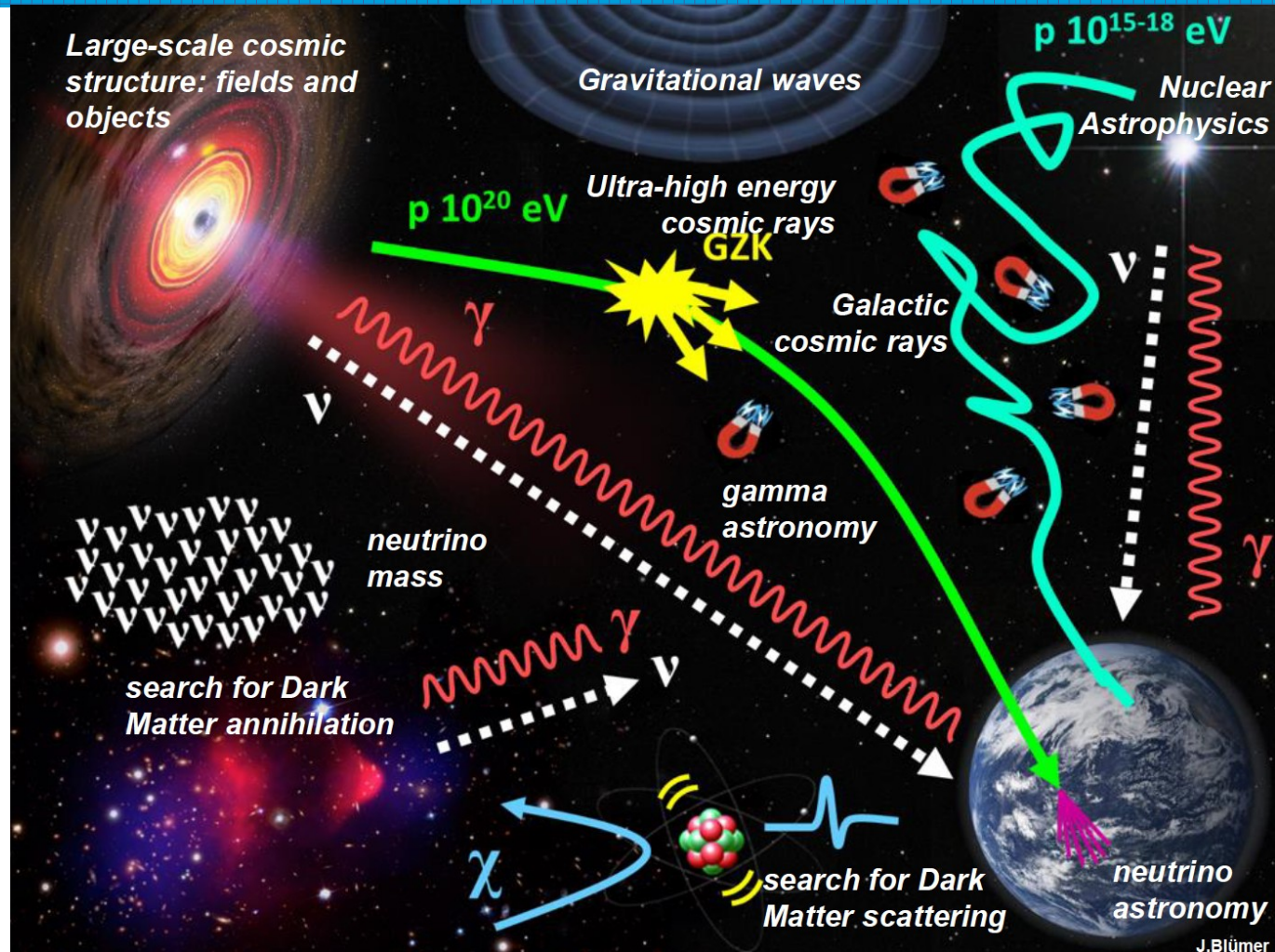
ROOT comes with an [incredible C++ interpreter](#), ideal for fast prototyping. Don't like C++? ROOT integrates super-smoothly with Python thanks to its [unique dynamic and powerful Python \$\Leftrightarrow\$ C++ binding](#). Or what about [using ROOT in a Jupyter notebook](#)?

Astroparticle physics: understanding multi-messenger and dark Universe

Data providers:

- **Observatories** operated by international organisations (CTAO, ESO, ...)
- **Experiments** operated by collaborations / institutional groups, depending on the size of an experiment

=> variety of data policies



Astroparticle physics: open data commitments

since 2008: **NASA's Fermi Gamma-ray Space Telescope** releasing gamma-ray data openly soon after launch.

since 2012: **The IceCube Neutrino Observatory** began publicly releasing neutrino detection data. IceCube continues to release datasets ~annually.

2013: **The Cherenkov Telescope Array (CTA)** project committed to open data principles even before construction. CTA Observatory plans to release its data to the public after about one year proprietary period.

2017: **The Pierre Auger Observatory** released open cosmic ray datasets, allowing external researchers to analyze ultra-high-energy cosmic rays independently => **see the Auger Open Data talk by Jon Paul**

2018: **H.E.S.S. collaboration** released a small subset of its archival data consisting in total of 27.9 hours observations of the Crab nebula, PKS 2155–304, MSH 15–52 and RX J1713.7–3946 taken with the H.E.S.S.-1 array. The release consists of event lists and instrument response functions in FITS format.

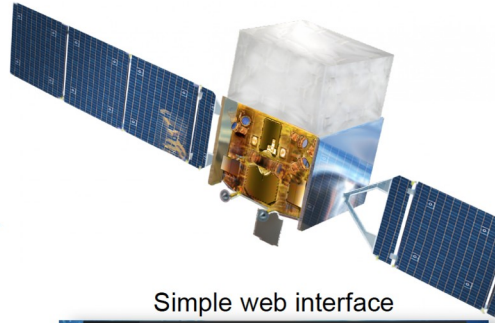
Occasionally, **H.E.S.S.**, **MAGIC** and **VERITAS** collaborations publish supplementary data in individual studies. Data sharing is primarily within collaboration members or upon request. **HAWC** and **LHAASO** collaborations have similar restricted data policy.

Fermi: a benchmark for gamma-ray data production

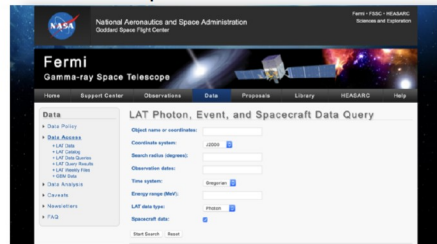
Fermi Gamma-ray Telescope

A showcase how observatories work

- Gamma-ray telescope operated by NASA
—> survey instrument
(+DOE, institutions in France, Germany, Japan, Italy and Sweden)
- All high-level data products available to the community (<1 day)
- Prompt data (e.g., GRBs) and notification (within <~15 s of detection)
- Services by NASA:
 - data (events+calibration) in common formats (FITS)
 - open source software tools
 - user support desk & documentation
 - cross-mission accessibility (e.g. through NASA's HEASARC archive)
 - catalogues



Simple web interface



Most publications not by the Fermi Team

1	2021JHEAp...39...40C 4FGLeo. Classifying Fermi-LAT uncertain gamma-ray sources by machine learning analysis Chen, Guizhao; Kowalewicz, Milos; La Mura, Giovanni	2021/03			
2	2021ApJS...252...13A First Fermi-LAT Solar Flare Catalog Ajello, M.; Baldini, L.; Bastieri, D. and 82 more	2021/02			
3	2021MNRAS...516...264T Annihilating Dark Matter Search with 12 Years of Fermi-LAT Data in Nearby Galaxy Clusters Thorpe-Morgan, Charles; Melby-Lind, Dany; Stagen, Christoph-Alexander and 2 more	2021/01			
4	2021MNRAS...516...216O Searching for signatures of chaos in γ -ray light curves of selected Fermi-LAT blazars Ostapenko, O.; Tarnopolski, M.; Zywucka, N. and 1 more	2021/01			
5	2021MNRAS...510...5297A Locating the gamma-ray emission region in the brightest Fermi-LAT γ -ray-spectrum radio quasar Achariyya, Abhishek; Chackwaj, Pritika M.; Brown, Anthony M.	2021/01			
6	2021ASA...645A...62B Ornstein-Uhlenbeck parameter extraction from light curves of Fermi-LAT observed blazars Burd, Paul R.; Kehlmann, Luca; Wagner, Sarah M. and 3 more	2021/01	cited: 1		
7	2020ApJ...895...114L Fermi-LAT Observations of V549 Vel 2017: A Subluminous Gamma-Ray Nova? Fermi-LAT Observations of V549 Vel 2017: A Subluminous Gamma-Ray Nova?	2020/12	cited: 1		

Fermi: a benchmark for gamma-ray data production

Fermi Gamma-ray Telescope

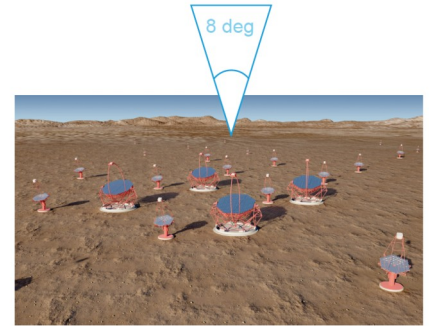
A showcase how observatories work

- Gamma-ray telescope operated by NASA
→ survey instrument
(+DOE, institutions in France, Germany, Japan, Italy and Sweden)
- All high-level data products available to the community (<1 day)
- Prompt data (e.g., GRBs) and notification (within <~15 s of detection)
- Services by NASA:
 - data (events+calibration) in common formats (FITS)
 - open source software tools
 - user support desk & documentation
 - cross-mission accessibility (e.g. through NASA's HEASARC archive)
 - catalogues

CTA Gamma-ray Telescope

Future observatory for gamma-ray astronomy.

- Gamma-ray telescope operated by CTAO
→ **pointed instrument**
+developed by a worldwide collaboration
- All high-level data products available to the community (open data after proprietary period)
- Prompt data (e.g., GRBs) and notification (within ~100s of detection)
- Services by CTAO:
 - data (events+calibration) in common data formats (FITS)
 - open source software tools
 - user support desk & documentation
 - cross-mission accessibility (possibly through ESO archive)
 - catalogues



Key difference to Fermi!
Observation proposals led by Principal Investigators

CTA Data policy not finalised yet (e.g., 1 year of proprietary period; afterwards open).

Similar to X-ray telescopes
(XMM, Chandra)
>50% of all publications based on archival data

VHE gamma-ray community effort

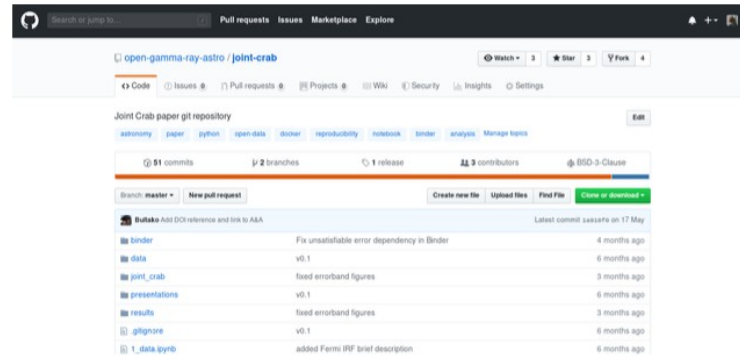
- common high-level data formats :
 - event level plus instrument response functions
 - science data

data level	description	size
DL0	raw output of DAQ	~ TB / tel. / night
DL1	calibrated quantities (charge, arrival time)	~ 10 GB / night
DL2	reconstructed shower parameters	~ 10 ² MB / run
DL3	reduced γ ray candidates + response functions	~ 10 ² kB
DL4	science data products: spectra, light curves, skymaps	~ 10 kB

- public software tools



- workflows and archiving



- elaborating scenarios for public data archives (e.g., first H.E.S.S. data release (arXiv:1810.04516, zenodo))

Altogether, preparing the upcoming CTA Observatory



IceCube Data Policy

Appendix E: Dissemination and Sharing of IceCube Research Results and Data

This defines the IceCube strategy for providing access to research results and data by the broader research community. NSF policies and guidance promote efforts by grantees to produce the timely publication of results and to make data and software available to other researchers. In addition, the Parties to the Antarctic Treaty agree that, to the greatest extent feasible and practicable, scientific observations and results from Antarctica shall be exchanged and made freely available.

IceCube is a facility-class experiment with the primary goal to identify sources of astrophysical neutrinos. NSF supports a wide range of approaches to the release of facility data, e.g., the particle physics model where data is exclusively available to members of the collaboration and the astronomy model where data are readily made public.

The Large Hadron Collider experiments follow the particle physics model; the Atacama Large Millimeter/submillimeter Array (ALMA) – the astronomy model; and, the Wilkinson Microwave Anisotropy Probe (WMAP) – an intermediate model. IceCube is similar to WMAP and large air shower experiments where data is collected, analyzed, published and released.

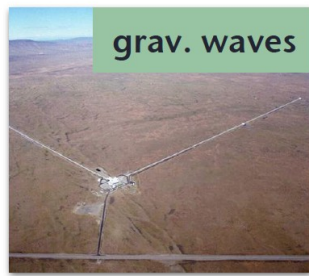
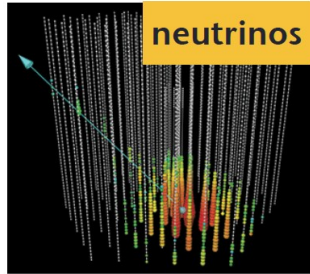
The public release of data in a scientifically meaningful way is not a trivial undertaking.

Currently there are three ways to access IceCube data:

1. IceCube Collaboration Membership
2. Associate Membership
3. Direct Access to IceCube Public Data Pages

The neutrino event IC-170922A sent a trigger to AMON ...

The Astrophysical Multi-Messenger Observatory Network



AMON enables searches for multi-messenger coincidences using particles representing the *four fundamental forces*.

Triggering observatories:

- ▶ Transmit “sub-threshold” candidate events to AMON in real-time.

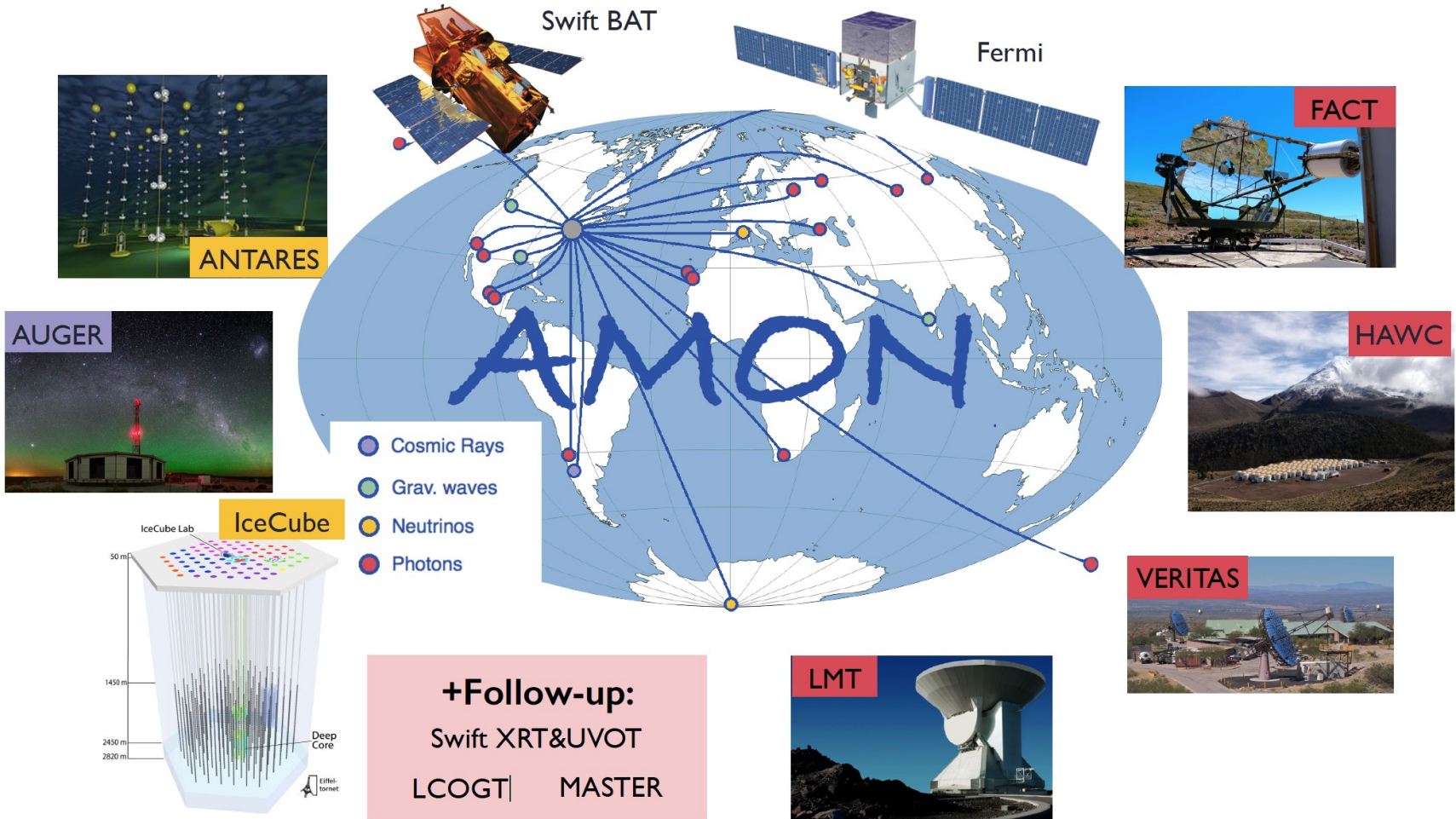
AMON:

- ▶ Provides framework for *real-time coincidence searches* of data in direction & time.
- ▶ Broadcasts *real-time alerts* (via VOEvent & GCN).
- ▶ Enables *archival analysis* of sub-threshold data.

Follow-up observatories:

- ▶ Respond to AMON alerts.
- ▶ Provide optical feedback on potential multi-messenger transients.

AMON partners



AMON case study: IceCube

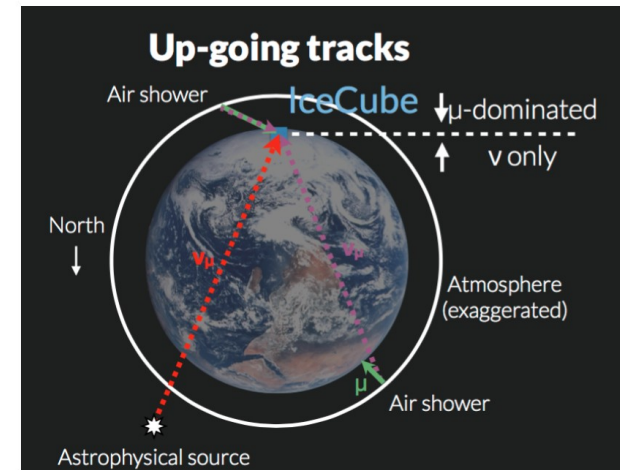
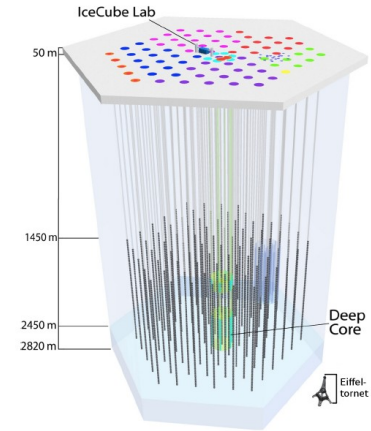
IceCube cannot detect ν sources alone:

▶ ~ 10 HESE/year

▶ $\sim 1/200$ from $z < 0.1$, with high p_{cosmic} , and $\theta \sim 1^\circ$
e.g. Ahlers&Halzen 2014, Waxman 2013

Realtime IceCube reconstructed events trigger AMON:

- ▶ **HESE** - 3 track-like highest p_{cosmic} events followed up by Swift (*NASA 1215235 award*) ~ 1 signal event
- ▶ **EHE** - 6 events /year ~ 3 signal events.
- ▶ **Up-going tracks** - Multiplets generate AMON alert
- ▶ IceCube alerts distributed by AMON/GCN



sketch from Claudio Kopfer

Challenges

Technical Challenges:

- Large volume of data (~petabytes per year) and the need for specialized infrastructure. Accessibility of this infrastructure can be a hurdle, especially for smaller institutions.
- Complexity of data formats (e.g., FITS, HDF5, ROOT) and the steep learning curve for new researchers and the general public – a limiting factor for their engagement.

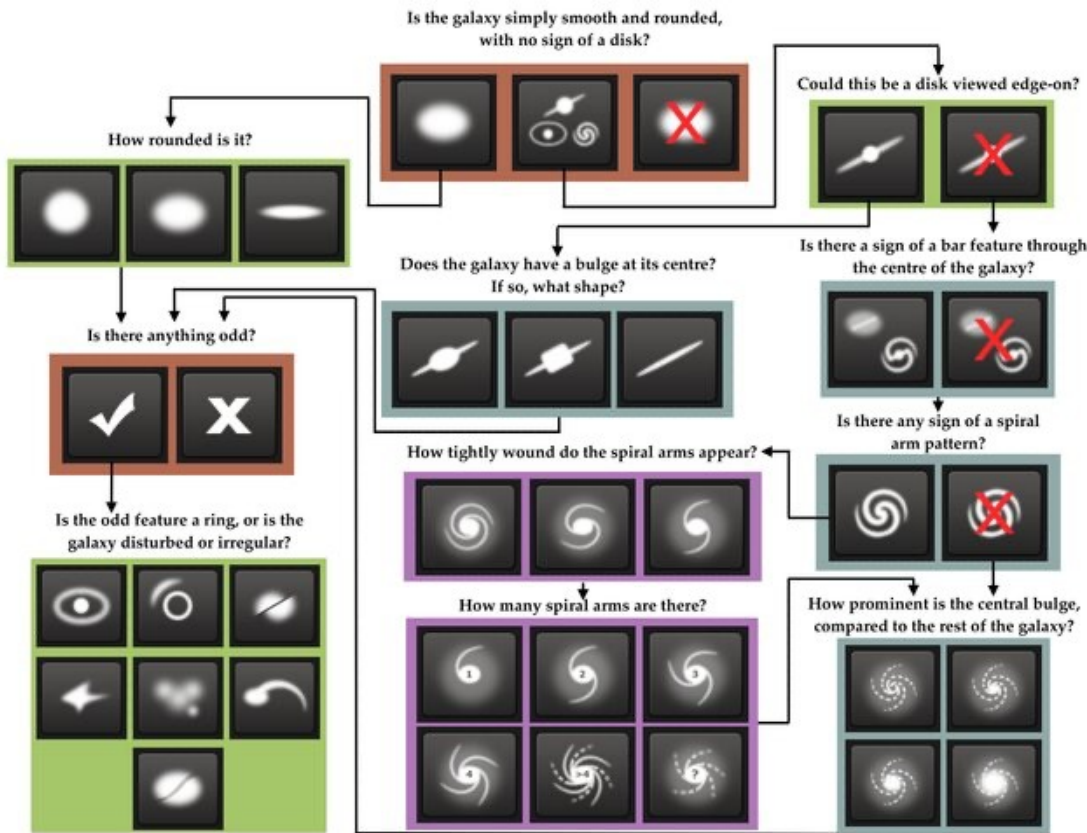
Data Standardization Issues:

- Different standards across experiments or observatories, complicating cross-disciplinary use and impeding multi-experiment data integration.
- Inconsistent or incomplete metadata – often due to varying standards between institutions – can make data challenging to discover, interpret, and reproduce accurately. Metadata standardization and thorough documentation are crucial for FAIR principles but can be time-intensive.

Multimessenger Data Coordination:

- Standardized Alert Protocols like those implemented by the GCN and AMON networks are helping, but further work is needed to ensure that all observatories follow standardized protocols for timely, public alerts, crucial for coordinated follow-up. More cross-collaborative effort is also needed to account for differences in data collection timelines, formats, and alert systems.

Applying machine learning methods to open data



Galaxy Morphology Classification

(based on Sloan Digital Sky Survey galaxy images)

Summary: Galaxy Zoo, a citizen science project, provided a large dataset of galaxy images classified by volunteers. Machine learning models were later trained on this dataset to automate galaxy classification.

Results: The ML models achieved high accuracy, allowing detailed morphological studies of galaxies and insights into galaxy evolution. This research has helped map the distribution and shapes of millions of galaxies, advancing knowledge of large-scale cosmic structures.


Publication: Dieleman, S., Willett, K. W., & Dambre, J. (2015). "Rotation-invariant convolutional neural networks for galaxy morphology prediction." *Monthly Notices of the Royal Astronomical Society*, 450(2), 1441-1459.

"The application of these algorithms to larger sets of training data will be critical for analysing results from future surveys such as the Large Synoptic Survey Telescope."

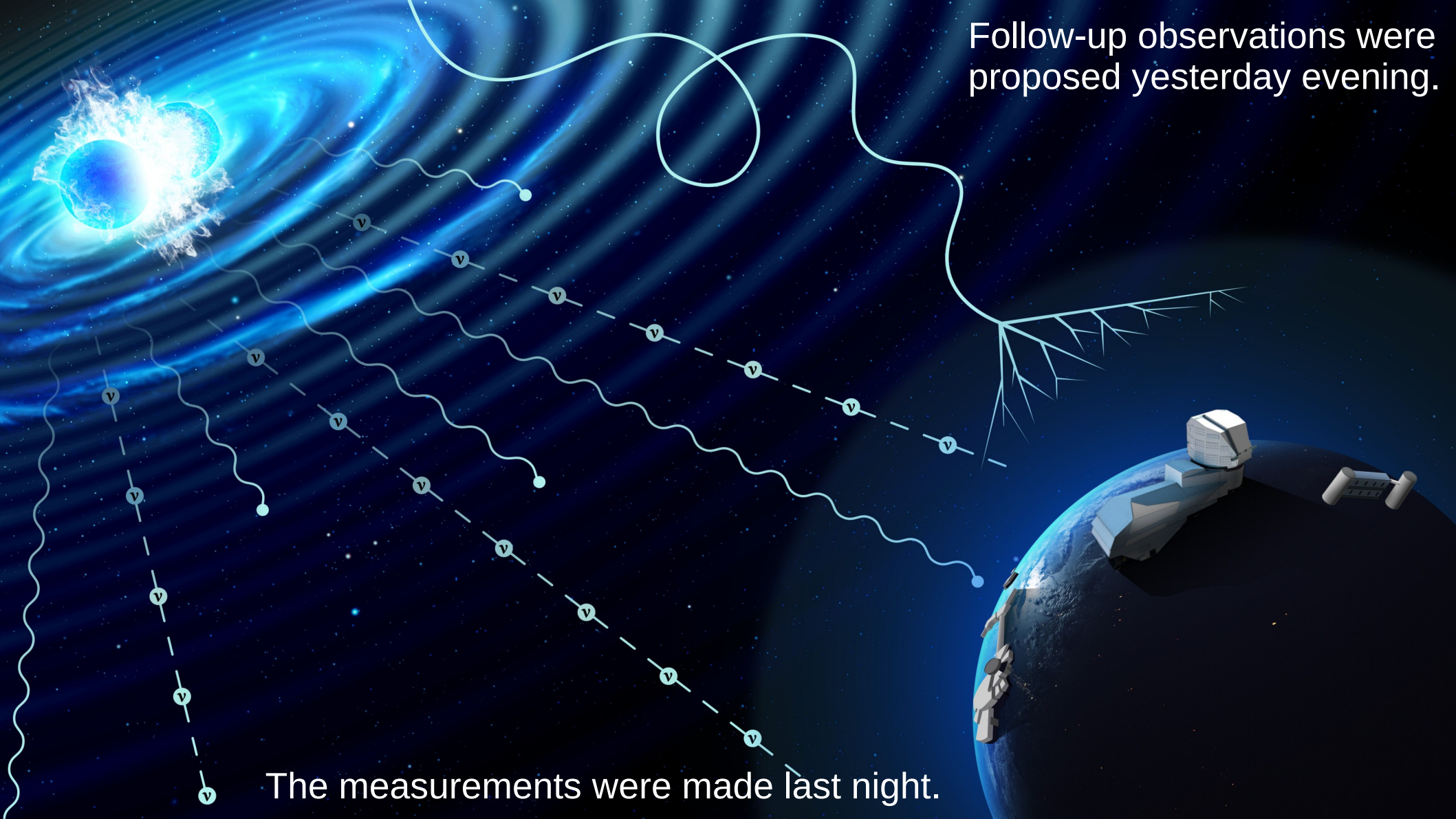
Conclusions & Outlook

- **Open data** has become fundamental in astrophysics, particle, and astroparticle physics, enhancing collaboration, reproducibility, and transparency, while accelerating innovation. A recent shift toward openness, marked by data-sharing initiatives and accessible resources, is driving breakthroughs like the multi-messenger observation of GW170817—a neutron star merger detected in both gravitational waves and gamma rays—and the identification of blazar TXS 0506+056 as a high-energy neutrino source.
- Across these fields, robust efforts are underway to develop and implement **FAIR-compliant data policies**, with a wide array of supportive tools, standards, protocols, and software already in use (Virtual Observatory in astrophysics, CERN's Open Data Portal in particle physics, ...). The challenges of astroparticle physics data, often more complex than traditional astrophysics or particle physics data, call for additional coordination and technical advancements to meet FAIR principles effectively.
- **Machine learning** also plays a transformative role in these domains, enhancing the analysis of both proprietary and open data to reveal new insights and optimize research methodologies.

Yesterday morning, a cosmic explosion
was discovered and announced.



This morning everyone can download the data
and combine them with own/other observations.



Follow-up observations were proposed yesterday evening.

The measurements were made last night.

To reach all this, we need a FAIRly open data!