# Data
# Observatory

## Llamado a Propuestas de Valor

*An opportunity to invest in the leadership of Chile on the data-centric era*

Ministerio de Economía, Fomento y Turismo

Gobierno de Chile

DATAOBSERVATORY

# Astronomy in the 21st Century

In order to explain our universe nature, its origins, and destiny, Astronomy, as a Natural Science, depends on the careful observation of the sky, the extraction of datasets from those observations, and the elaboration of theories from these datasets. A XVI-XVII century example is the emergence of *Philosophiæ Naturalis Principia Mathematica* by Isaac Newton (1643-1727). Systematic data extraction from sky observations done by Tycho Brahe (1546-1601), enabled Johannes Kepler (1531-1630) to deduct the laws of planetary motion, a result on which Newton build to elaborate his general laws of motion in 1687, still confirmed nowadays in activities that range from real estate construction to space exploration.

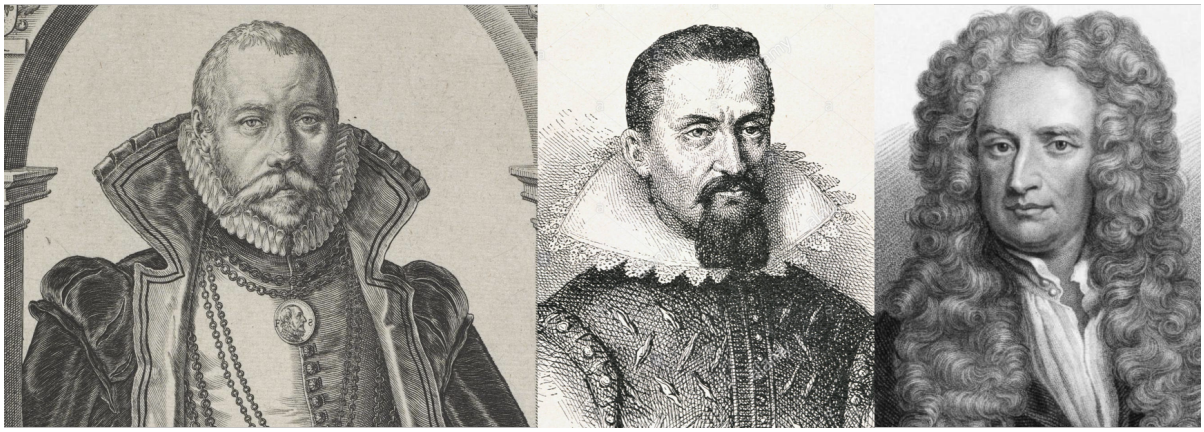

*Figure 1. Brahe, Kepler & Newton*

Over the last century the technology of astronomical observatories improved dramatically, the understanding of the origin and destiny of our universe has evolved accordingly, but there is still more to discover. The unknowns of the accelerated expansion of the universe is an example of the latter. The Nobel of Physics was awarded in 2011 to Riess, Schmidt, and Perlmutter for demonstrating that indeed our universe grows at an increasing rate. We call Dark Energy to the cause of this observed expansion, if our current understanding of nature is correct, it amounts to 68% of our universe, and we don't know what it is.

This progress in technology ignited a transformation in the way astronomy works: the knowledge that emerged from individual's minds now flows from multi-disciplinary teams using data-centric tools. On one hand, data blooms from observatories; on the other, data bursts from cosmological simulations on computing clusters. Telescopes will produce zetta-scale datasets over the next decade (Quinn et al., 2015 & Szalay 2001), and theoretical astrophysics will generate similar data volumes and challenges.

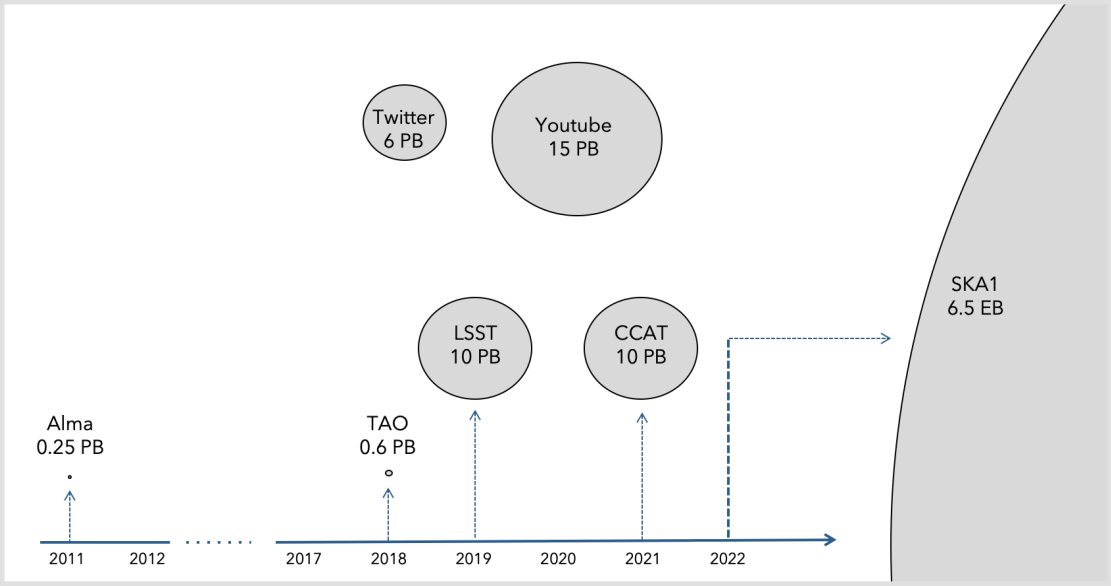


*Figure 2. Total data volumes in astronomy compared with web 2.0 (Ministerio de Economía, 2018 based in SKAO, 2013). For the reader reference, the time required to download a petabyte of data with the average cell phone internet connection (4G) is approx. 25 years.*

# Chile, a capital of ground Astronomy

There are many reasons why the Chilean Atacama Desert is wonderful for astronomy. Its remote mountaintop locations and its shielding by mountain ranges prevent potential urban light pollution; the photons acquired on its summits and plateaus in high-altitude go through a thin atmosphere, reducing its distortions further; its dryness prevents absorption of light by water vapor; the cold Humboldt Current of the Pacific Ocean and the Pacific Anticyclone reduce the formation of high clouds, allowing a larger number of clear nights for observation than other locations; and the laminar wind flow from the Pacific Ocean prevents the formation of atmospheric turbulence which leads to lower scintillation and sharper images.

Since the 60s, the collaboration between the Chilean government and international observatories has brought 40% of Earth's telescopes to our territory. That share will grow to around 60% in 2021 (Catanzaro, 2014 & Unda-Sanzana, 2018). The inauguration of instruments in the next decade will further enshrine the Atacama Desert as a capital of ground astronomy. The volume of astronomical data acquired in the Atacama desert will go from around 1 PB/year today, to 16.5 PB/year in 2021 (EY, 2017).

# Government Policy

Chile international relationship with astronomy began in 1849 when the US Navy Astronomy expedition led by James T. Gillis brought state of the art telescopes. These instruments were bought by the government to establish the National Astronomical Observatory (OAN) in 1852, on the mountaintop of the Santa Lucía hill in Santiago.

The Chilean Government most known scientific policy defines that 10% of the observation time is reserved for local astronomers. This rule started in 1961 and has had a positive impact. Undergraduate programs have raised exponentially in the last 20 years leading to an explosive increase of astronomers, and Astronomy is the most productive scientific field of the country (figure 3).

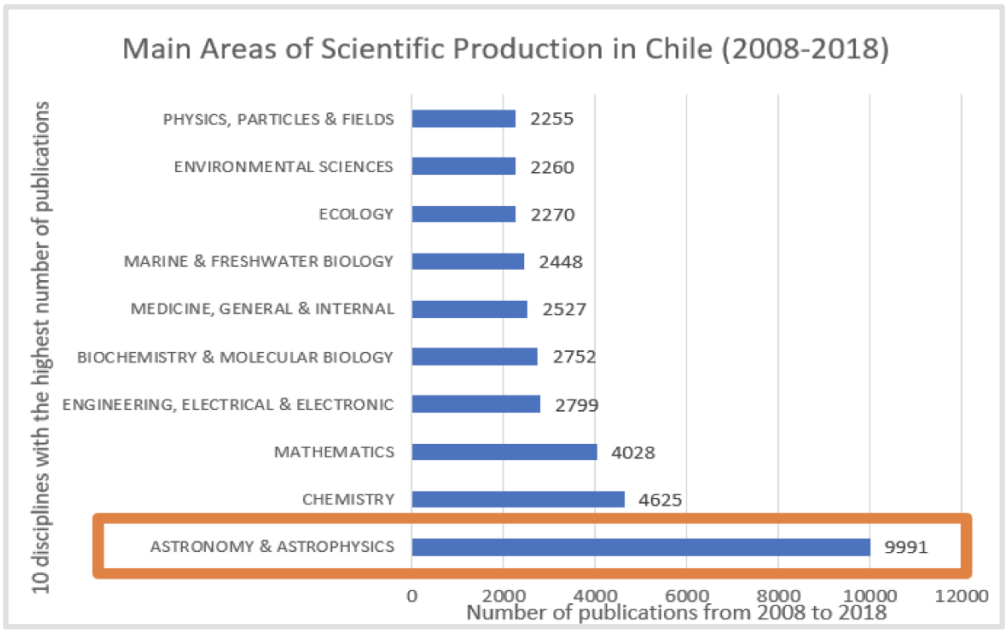


*Figure 3: Publications from 2008 to 2018. Source: Guridi et al., 2018 with data from CONICYT.*

## Chile´s unique position in Astronomy

Build on the natural advantages and the long-lasting positive relationship with international observatories, the Government of Chile has a unique position as the only public or private organization in the world to have already in place agreements with observatories from the United States, Canada, Europe, Korea, and Japan. Each time one of these observatories is installed in the country, a treaty is signed. Until 2018, the Government had signed 22 treaties and is in the process of negotiating 5 more for next decade. The Government has identified Astronomy as a priority field, and is leveraging its position in order to generate the right conditions for a public-private partnership, and make the best of the two sectors working together for a common goal: the development of the region digital economy; the government is offering all the coordination capacity of the public sector, calling for the private sector to contribute with all the entrepreneurship, talent and market development capacity.

## Astronomy as a capacity building engine

In order to solve its daunting data-centric challenges, astronomy has developed talent, technology and infrastructure for the following groups of data-centric tasks:

- **Data Acquisition and Generation tasks:** astronomy projects demand data that comes from observations or simulations, the data is acquired and generated using advanced sensors, has to be quality-assured, representing true physical phenomena, and has to be in the appropriate units to work with it.
- **Data Access & Governance tasks:** to use and reuse the data, has to be standardized, stored and indexed, enabling it for search and discovery by the initial project, or new ones.
- **Data Analysis tasks:** data obtained from archives or directly from simulations or observations is analyzed (either real-time or in ways not time-dominated) to obtain value in a broad sense, new knowledge or insights towards it.
- **Data Exploration and Visualization tasks**: In doing each of these tasks, persons explore and visualize data, which means bringing it back to human-scale, for people to choose what to do next in this non-linear data-centric process.

Even though Astronomy is different to other activities in many aspects, it is similar in these groups of data-centric tasks, hence the developed talent, technology and infrastructure is useful to other activities. If the right coordination between astronomy and productive sector exists, the capacities developed for Astronomy can be transferred to other activities.

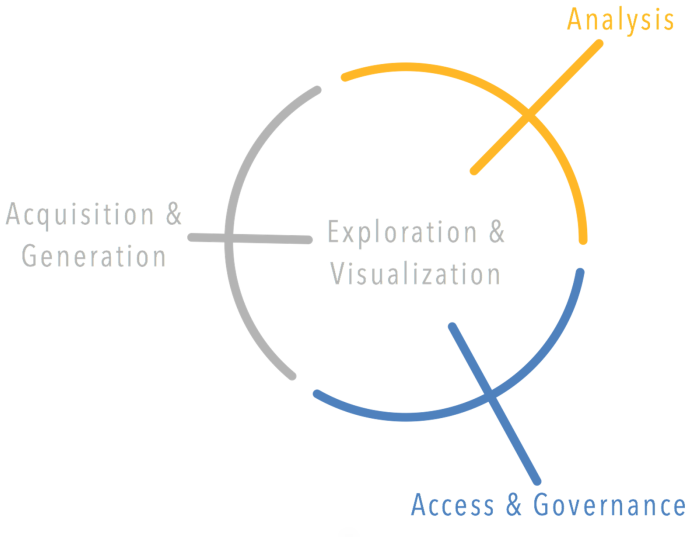


*Figure 4. Astronomy Data-Centric Tasks.*

# Astronomy as market maker

From a technical perspective, the astronomical data acquired in Chile is not only large but also complex and in many cases requires real-time analysis. Regarding its volume, there is no field in the Latin American region that compares with the 16.5 PB/year prospected for 2021 (EY, 2017). Regarding its complexity and variability, includes data from diverse instruments observing our universe at different wavelengths that go from gamma rays to the millimeter and sub-millimeter end of the spectrum, including visible and infrared bandwidths. Finally, regarding the analysis velocity, the signal captured by observatories is digitized and has to be real-time processed to remove instrument and observing systems artifacts and to calibrate the instruments for diverse observing modes, aiming for data to be in the appropriate physical units for scientific analysis. Also, in the era of multi-messenger astronomy, the astronomical data produced in Chile requires state-of-the-art machine learning techniques to enable coordinated follow-up observations of relevant phenomena in Chile and elsewhere.

The forefront capacities related with acquiring, transporting, storing, analyzing, visualizing and offering access to this exa-scale data critical for the progress in the fields of space exploration and astronomy across-continents, constitutes an exciting and new market for IT industry interested in investing in the Latin American region, and an enabler for other activities in this region to progress in the same data-driven direction, and hence for other new and exciting markets to emerge.

# The Data Observatory, a way to transform the opportunities in reality

Based on Chile's Government unique position in global Astronomy, the potential of the field as a capacity generator and market maker, and the recommendations from international experts[1], the strategy of the Chilean Government, coordinated by the Ministry of Economy, Foreign Affairs Ministry and The National Commission for Science and Technology Research (CONICYT), together with International Observatories consists in the creation of a neutral-broker organization to enable bi-directional transference between forefront data-centric fields (beginning with Astronomy) and IT industry, to foster the digital economy of the Latin American region. The Data Observatory, as a non profit organization, will be the body in charge of building from the Government existing coordination capacity to generate a data science pole of global impact. As such, the Data Observatory will not only provide and train the most capable human capital in data science, but it will also contribute to generate the markets of the future.

# The Data Observatory vision and mission

The DO will be a non-profit-organization at the heart of industry, government and astronomy collaboration.

The DO **vision** is to be at the vanguard of data-centric innovation, leading in the production of data-centric solutions, talent, and social capital for the Latin-American region. This translates to the **mission** of hosting datasets of global value acquired and generated in the Latin American region, and enabling their maximal exploitation by the global scientific community, the industry, and the public, facilitating data access, analysis, exploration, visualization, and governance.

In accomplishing this mission the DO activities will be compliant with the following principles:

- shall prioritize involvement in fields at the vanguard of data-centric requirements,

[1] **Dr. Massimo Tarenghi** (Italy-Chile, who directed the construction of the most advanced telescopes of the 20th century including the NTT in La Silla, the VLT in Paranal, and ALMA in Chajnantor), **Dr. Robert Williams** (USA, former Director of the Sace Telescope Science Institute and President of the International Astronomical Union), **Dr. Chris Smith** (USA, National Science Foundation Chief of Large Facilities, former AURA Director), **Dr. Peter Quinn** (Australia, Executive Director of the International Center for Radio Astronomy Research, former ESO Director for Data Operations), **Dr. Mario Hamuy** (Chile, member of the Chilean Academy of Sciences, former CONICYT president), **Dr. María Teresa Ruiz** (Chile, president of the Chilean Academy of Sciences), **Dr. Andrew Conolly** (USA, Director of the Data Intensive Research in Astrophysics and Cosmology at the University of Washington, current leader of the Large Synoptic Survey Telescope (LSST) Data Management Team) and **Dr. Alex Szalay** (USA, Director of the Institute for Data Intensive Engineering and Science at the John Hopkins University and former leader of the Sloan Digital Sky Survey (SDSS) Data Processing Team), among others.

- shall foster multi-directional transfer between selected fields and DO-members,
- shall work in coordination with DO-members avoiding competing with them,
- shall aim at complementing DO-members capacities,
- shall promote the use of open and public standards,
- shall promote open access to knowledge,
- shall promote access focused in intended audiences when deploying data,
- and shall be financially sustainable over time.

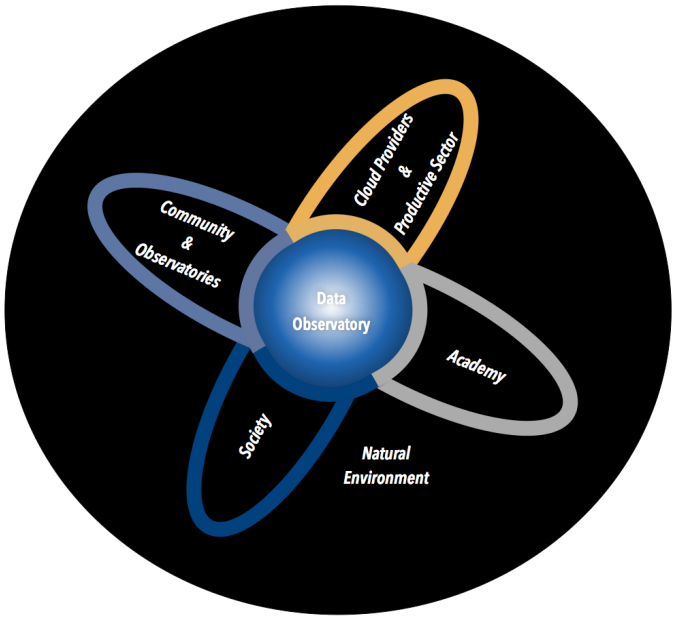


*Figure 5. The Data Observatory*

## Data Observatory Lines of Work Value Proposition

The DO will operate in the 4 following lines of work:


*Figure 6. The Data Observatory lines of work*

The **dataset and dataset products** line of work will acquire, and offer analysis, exploration & visualization, access & governance tools to maximize the exploitation in a broad sense of the most valuable datasets, mixing public and other modes of access to maximize the value of the data.

The **challenge management** line of work will adopt or generate challenges related to the most valuable datasets, and develop specific data-centric solutions to solve these challenges, and to enable exploitation of the data beyond its originating field. Results will be solutions at TRL6 to TRL9.

The **talent formation** line of work will create learning materials and provide mentors to foster the formation of experts for the future needs of digital economy.

And the **outreach** line of work will create a space for multi-sector collaboration, focused at building trust between a diverse group of stakeholders, and growing strong community ties with Chile and the Latin American audiences interested in Astronomy and other data-centric domains.

These line of works will create a spectrum of value for a diverse set of stakeholders:

For **Infrastructure as a Service Providers (Cloud and others)**, it will provide an attractive pool of data to be hosted at their infrastructure, this data will increase their infrastructure associated tools usage and brand value, and through the DO challenges with diverse stakeholders, it will increase the spectrum of potential customers on one hand and talent for the provider on the other. Finally, it will broaden the provider impact, specifically to the area of talent, technology, and infrastructure capacity development.

For **Data Producers**, it will give untethered access to their public data, and their preferred specific-audiences access to other datasets they provide to the DO. It will enable collaboration with a large and active data science community, boosting the productivity of their data in a broad sense that includes research publications and can lead to talent, technology and infrastructure development for the Latin American region. Finally, through the DO challenges line it will provide new tools to exploit their data with DO solutions at TRL6 to TRL9, including potential royalties if these tools were generated using their data.

For **DO Trainees**, it will provide untethered access to all public data in the DO, including access to specially curated datasets for education and support materials and specific research needs, and it will provide preferred Data Producer access to other datasets. It will also provide opportunities to develop relevant Data Science skills and to participate as a protagonist in challenges that are defining the future of the data science field and the data-centric activities in a broad spectrum of domains. Finally, it will provide increased domain exposure, to the domains of the DO data producers.

For **Small, Medium-sized Enterprises (SMEs) and Large Enterprises**, if they have data-centric capacities, it will provide opportunities to scale up these capacities in volume,

variability (complexity) or velocity to serve new and more profitable markets, and no matter if they already have capacities or not, it will provide data-driven capacities (talent formation environment, solution development environment), skilled and experienced Data Professionals to hire for projects or indefinitely, will provide ways to evaluate the commercial viability for developed solutions and royalties if the solutions are successful. Finally, it will increase their brand value.

For **Research Organisations**, it will provide untethered access to public data, and access to curated datasets for education and support materials or solving specific research challenges, it will also provide skilled and experienced Data Professionals, interaction with domain experts, and commercial opportunities for researched solutions, including potential royalties. For these organization, it will also provide opportunities to increase brand value and to broaden the organization impact.

For **Researchers**, it will provide untethered access to public data, and access to specially curated datasets for research challenges, a pool of skilled and experienced data professionals, and interaction with domain experts, it will also provide commercial opportunities for researched solutions and potential royalties.

For **Higher Education Entities**, it will provide untethered access to public data, and access to specially curated datasets for education and support materials, training materials and tutors.

For **Schools**, it will provide untethered access to public data and access to specially curated datasets for education and support materials, training materials and tutors, a place to obtain hands-on peta-scale data science and engineering training.

For **Data-gazers**, it will provide untethered access to public data, a place to obtain hands-on peta-scale data science and engineering training and philanthropy opportunities.

# How to participate?

In the concept development of the DO, there are 3 ways envisioned for stakeholders to participate.

## Founding members

The DO will have 4 founding members, one of them being the Chilean Government. Founding members are those that contribute to the DO in an amount equal to or greater

than the Chilean government initial investment, whose membership length will be agreed on a case by case basis, using the Chilean government case as a baseline. DO founding members will be part of the DO governance with the Chilean government according to the diagram below and will have, in addition, the following rights:

- designate ⅗ of the board and ⅔ of the supervisory and auditing committee,
- access to the DO curated data sets and DO-generated challenges,
- access to all DO outreach activities,
- and when Challenge management activities result in a product that will be commercialized by the DO Stakeholder, participation in the agreements that will be generated to correctly assign IP results.

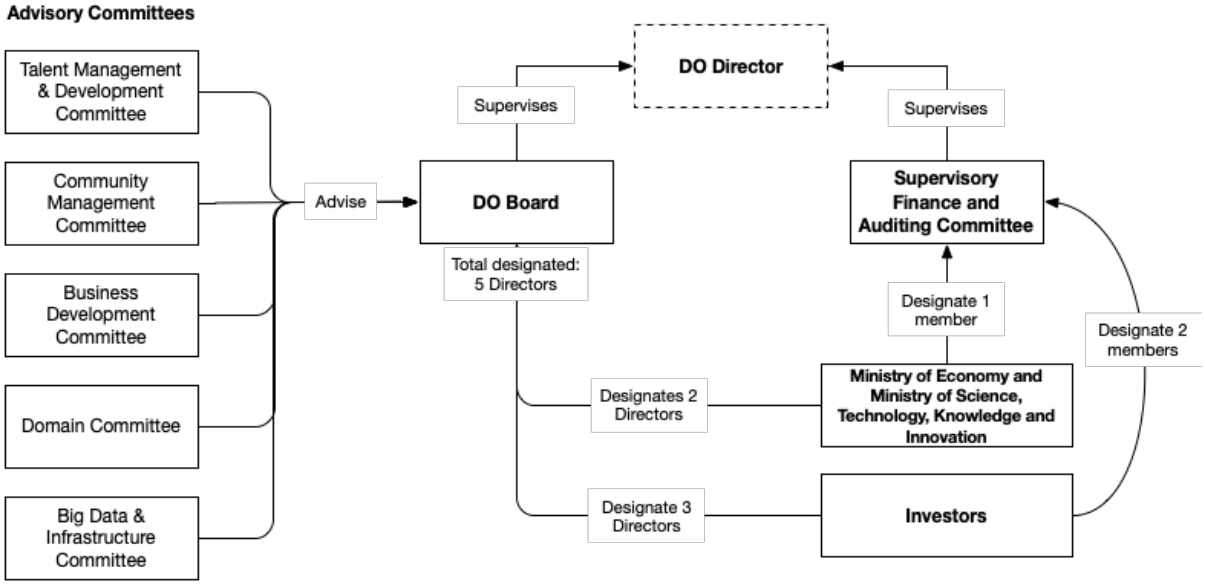


*Figure 6. Data Observatory Board Composition. The DO Board will be composed of 5 members designated by the Government (2) and the Founding Members (3), the committees will provide expert advice to the board on the mission of the DO. Also, the founding members will designate ⅔ of the supervisory finance and auditing committee and the government the remaining third.*

## Memberships

There are a number of ways to become a DO member:

1. being a DO stakeholder that pays an annual membership-fee, varying based on the size of the entities and the entity mission;
2. being a DO stakeholder that provides a dataset deemed valuable by the DO in accordance with its mission. Data is considered valuable when associated challenges that enable the DO to be at the vanguard of data-centric activities can be identified. The length of the membership, in this case, will be the period over what data is contributed and remains valuable.
3. being a DO stakeholder that provides cloud capacities deemed valuable by the DO in accordance with its mission. Cloud capacity is considered valuable when enables addressing DO challenges.
4. being a DO stakeholder that provides talent deemed valuable by the DO in accordance with its mission.
5. the DO board can define new ways to become DO member.

DO Members will have the following rights:

- designate members of the advisory committees,
- access to the DO curated data sets and DO-generated challenges,
- access to all DO outreach activities,
- and when Challenge management activities result in a product that will be commercialized by the DO Stakeholder, if the member is included in the Challenge, participation in the agreements that will be generated to correctly assign IP results.

## Other ways to participate

1. Create a specific challenge hosted by the DO
2. Require access to a specific dataset curated by the DO
3. Require access to a specific talent formation material by the DO
4. Access to DO Open content