The ALMA Data Science Initiative

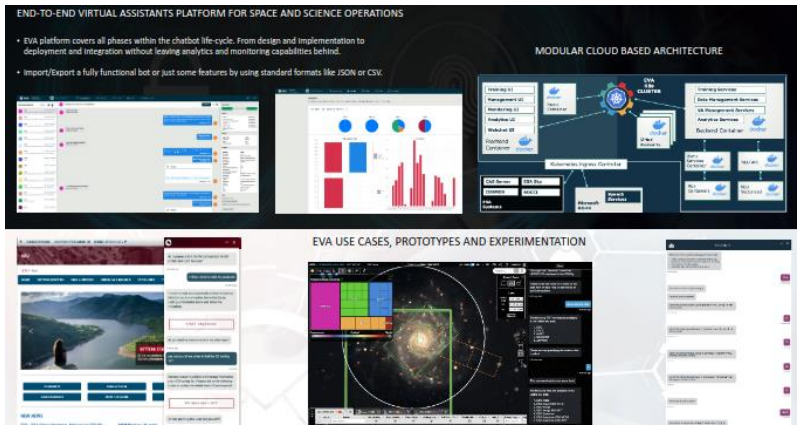
Building a Data-driven Organization to Improve Operations

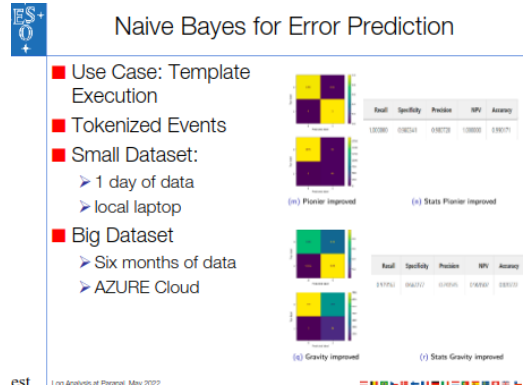
SciOps 2022: Artificial Intelligence for Science and Operations in Astronomy

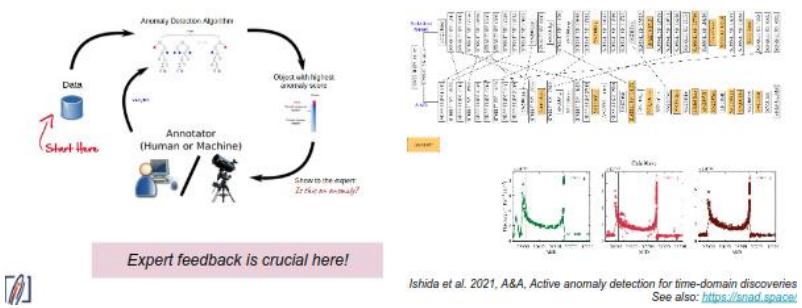
Ignacio Toledo, Joint ALMA Observatory

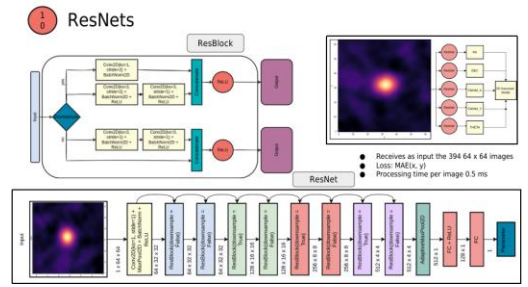
Great Conference!

- AI/ML is being developed and used, and the approach to data is slowly changing (Big change between 2019 and today)









Am I at the right place?

While I'm not an expert at all in AI/ML, I have been seeing a possible connection with your work and the challenges in operations.

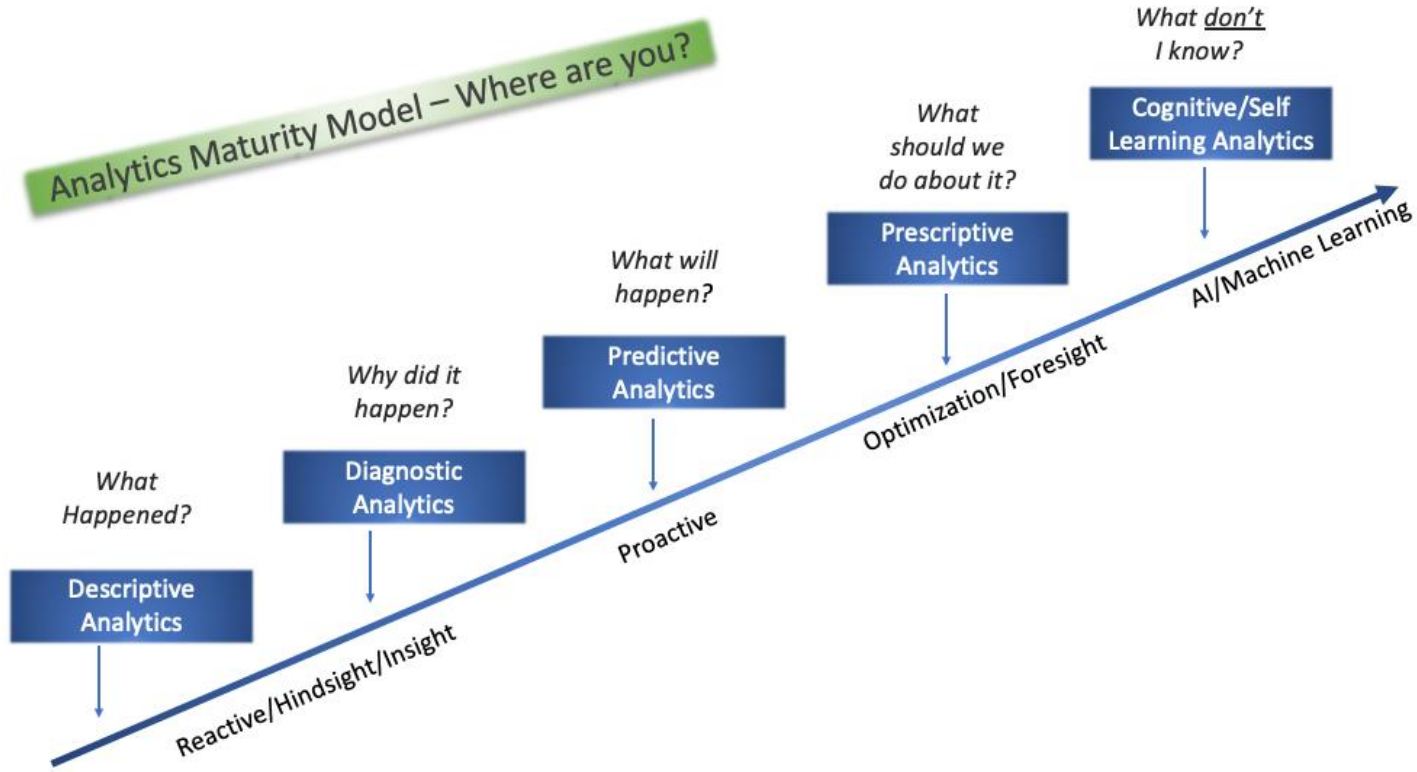
Hypothesis: we are meeting in the middle coming from different directions. If we meet there, most probably everyone will win.

AI/ML: Research vs Operations (broad generalization)

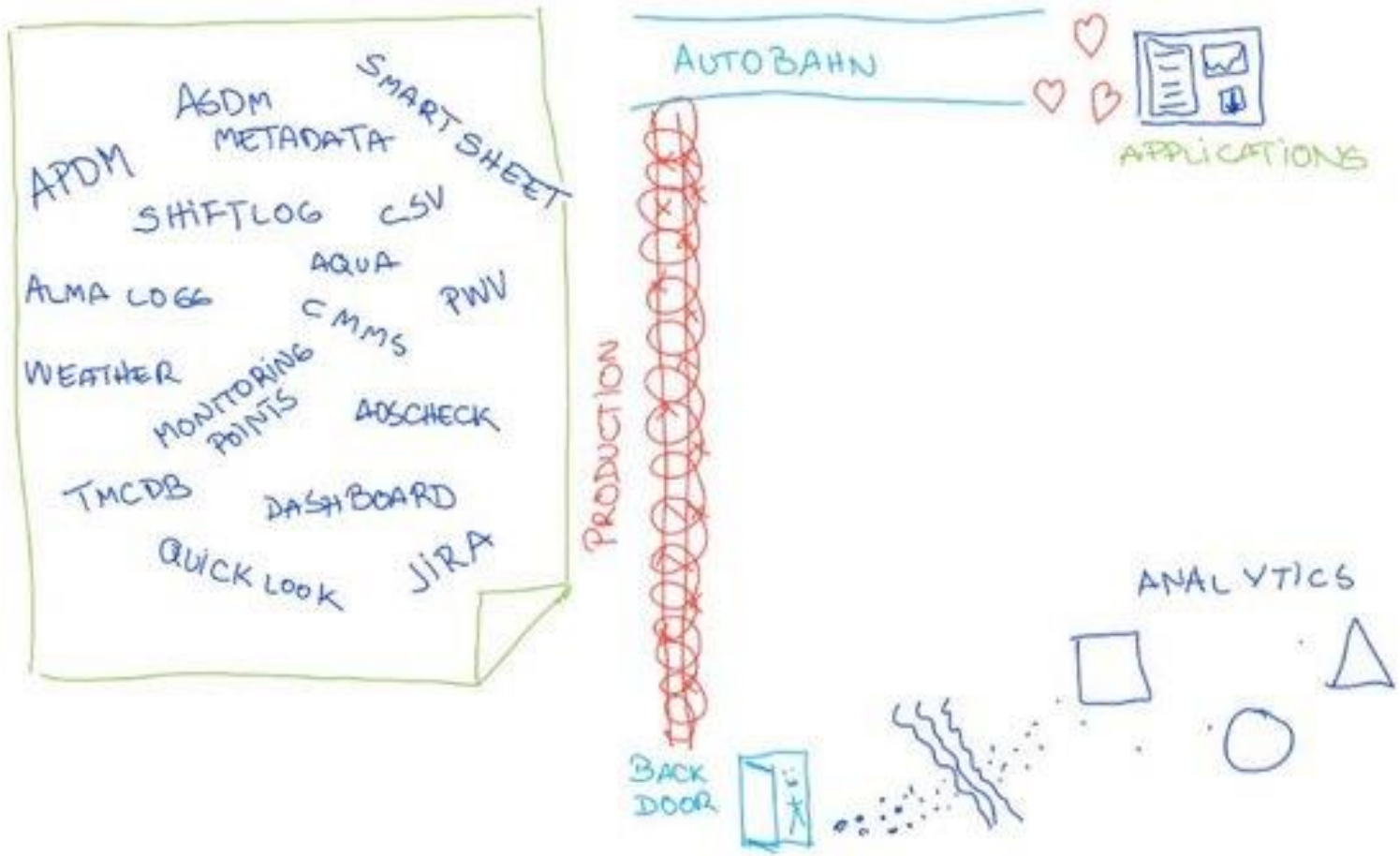
- Research looks for the best solution – Operations (companies) want to add value from the application
- Researchers are encouraged to increase the accuracy, even if takes a long time – Operations want to have solutions working quickly and that are maintainable
- Researches usually work with fixed, clean data, and in many cases with simulated data – Operations have to deal with dynamic data, that keeps changing with time
- Accordingly, researchers focus in the training and validation, and publishing a paper as soon as the results look great. Operations spend most of the time dealing with data and monitoring and maintaining data.

Twitter thread from @svpino

Analytical maturity for Organizations

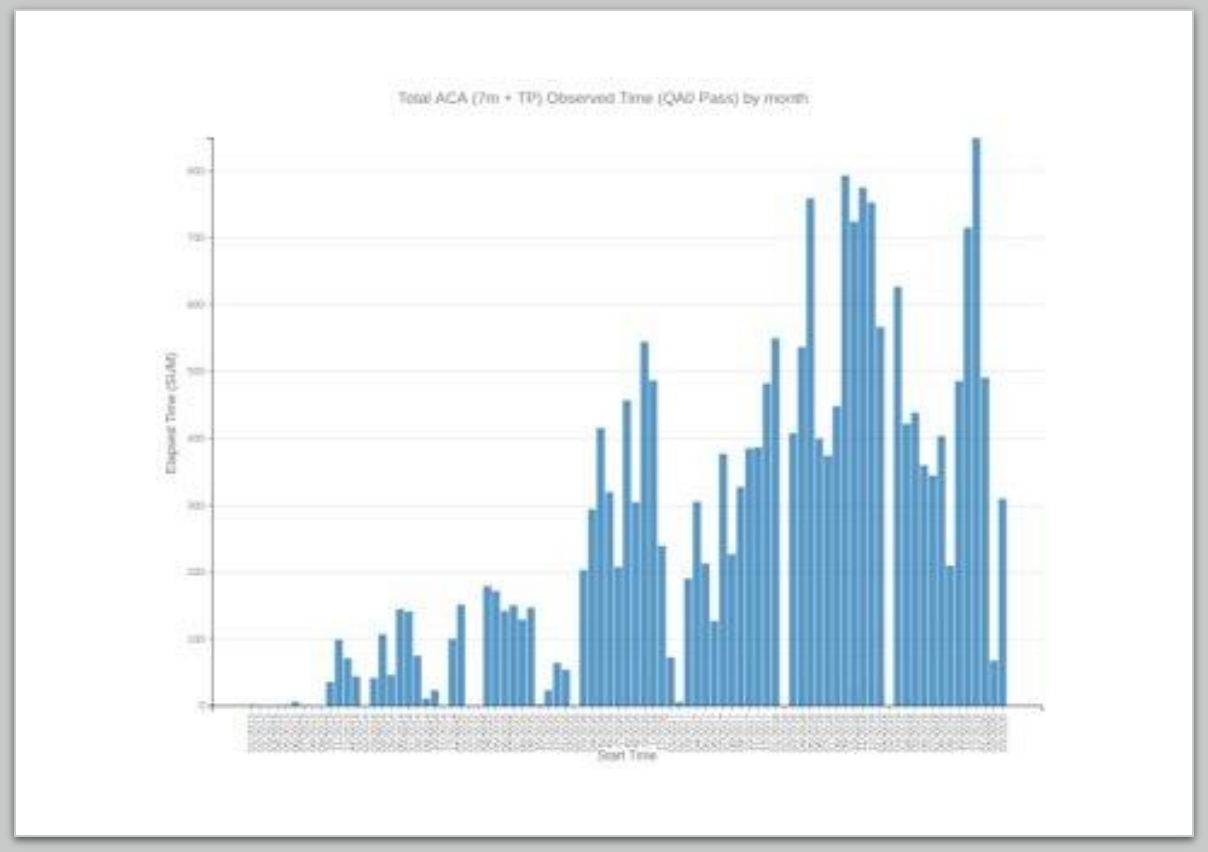


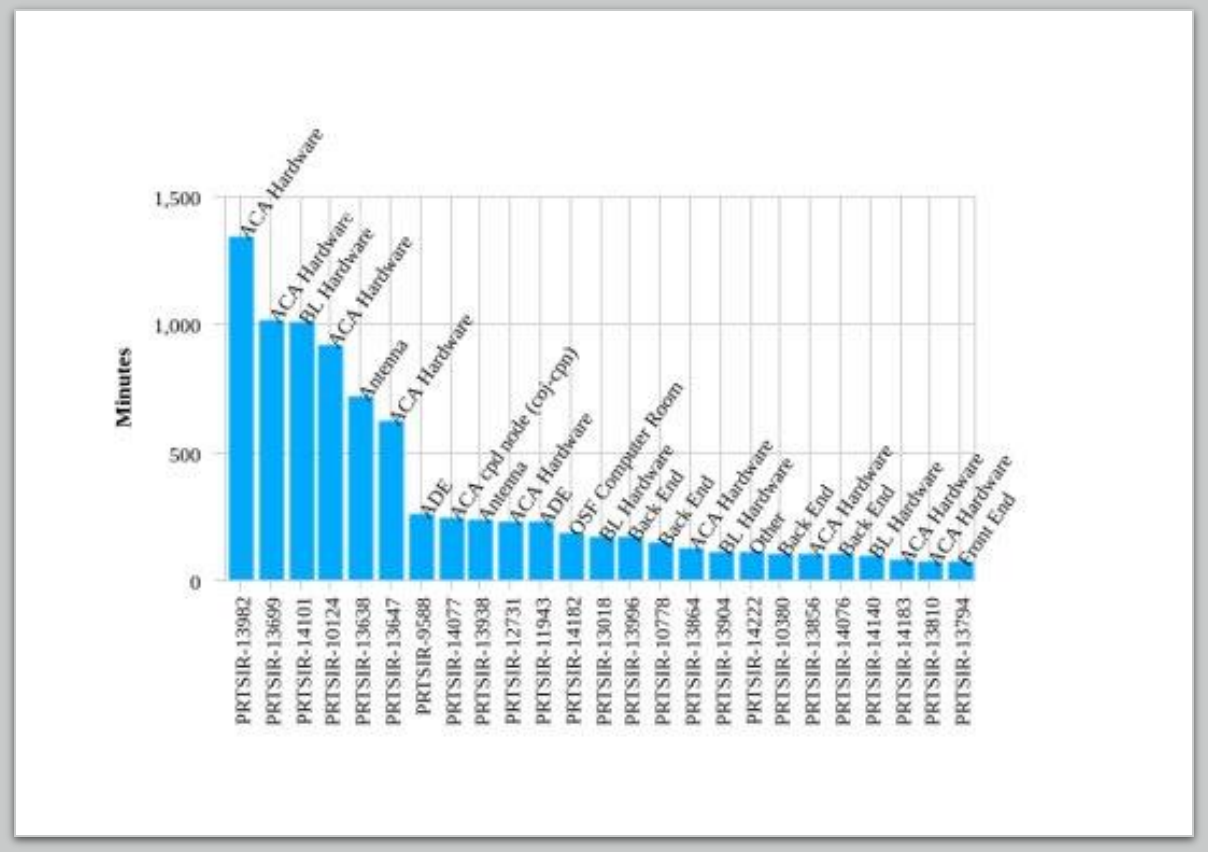
Observatory: Data for applications, not for analytics



Problems

- Duplicated work
- Decisions not consistent between departments or groups
- Not possible to automatize reports, many works lost in personal computers





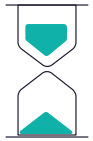
What could be done?

WE HAD A BUNCH OF PEOPLE INTERESTED IN FINDING A SOLUTION. BUT NOBODY HAD TIME TO BUILD A SOLUTION.

OPERATIONAL BUDGET IS SCARCE, NOT EASY TO INVEST IN A NEW SOLUTION WITHOUT CLEAR INDICATIONS THAT THERE WILL BE A NICE ROI.

Ikig.AI - Dataiku for good (2018)

- Free Dataiku license to fully functional version
- Free data science services (training, project support)
- Free tech support from onboarding to project delivery
- >1000 days of Ikig.AI time



DSS: An end-to-end platform

Connect data | Prep & enrich data | Create Models | Operationalize | Visualize

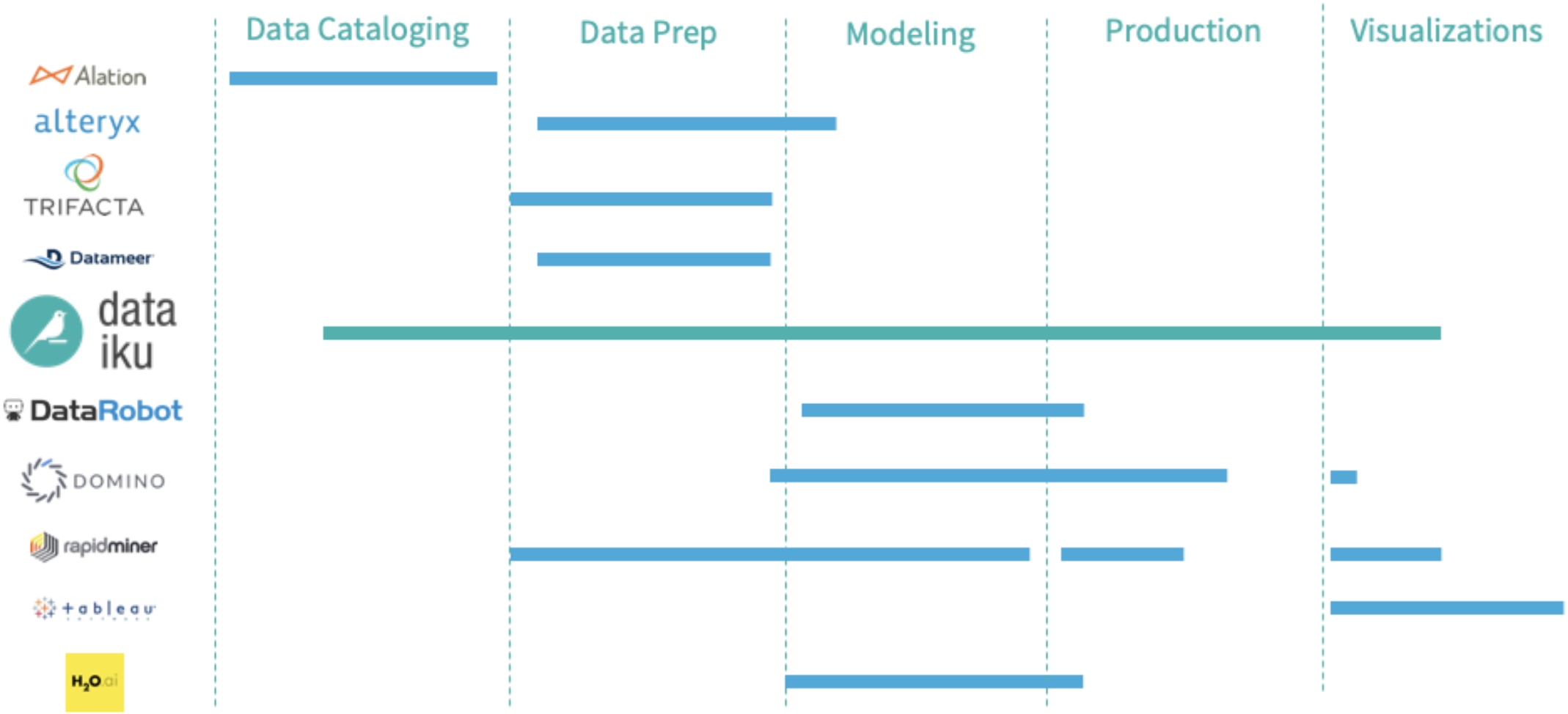
Data Engineers	Research Analyst (domain expert)	Data Scientist	IT / Data Ops	Data Analyst Or Consumer





Governance

DSS: The software landscape



DSS: An orchestration platform

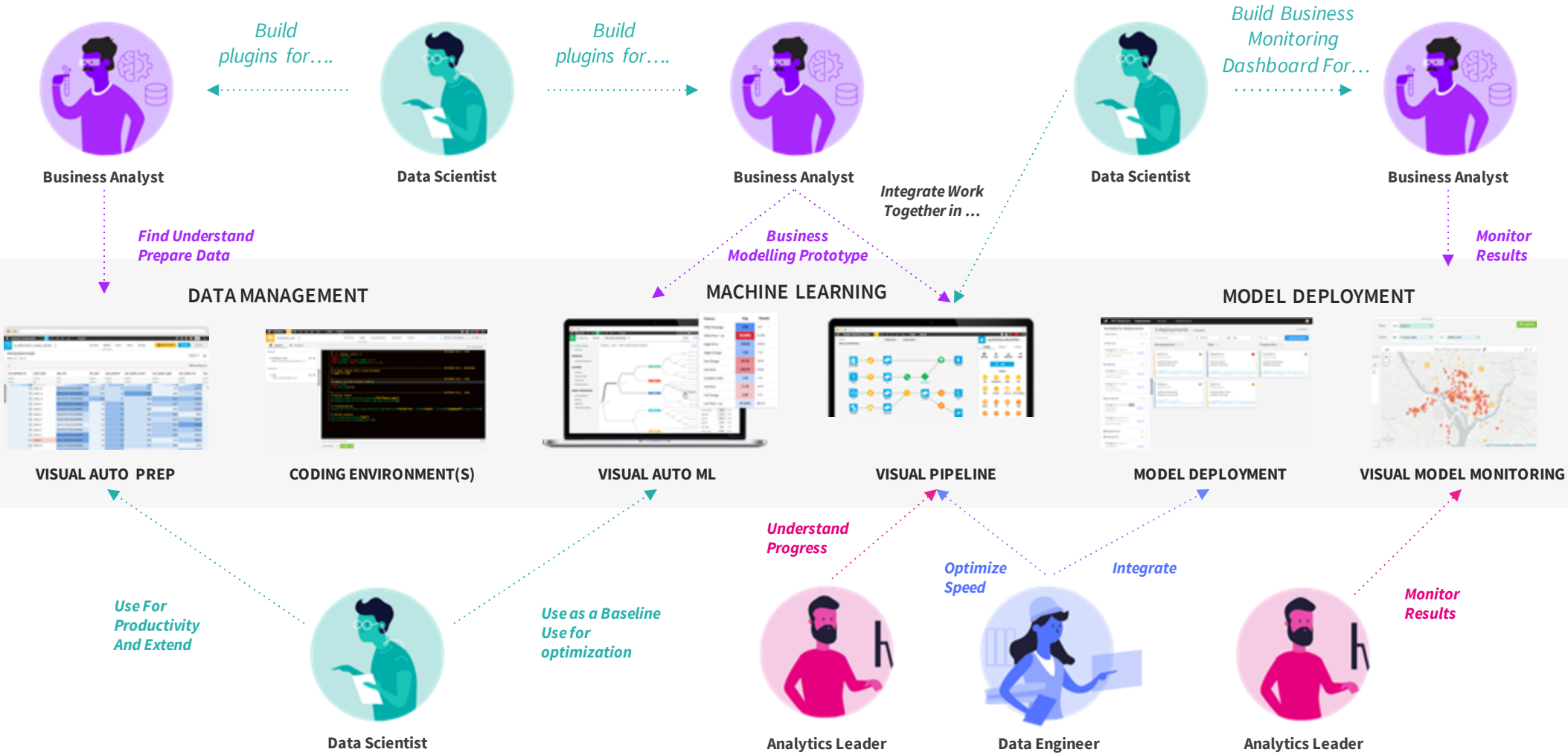
Leverage existing skills, **coding languages, and open source tools**

Use your current **infrastructure**

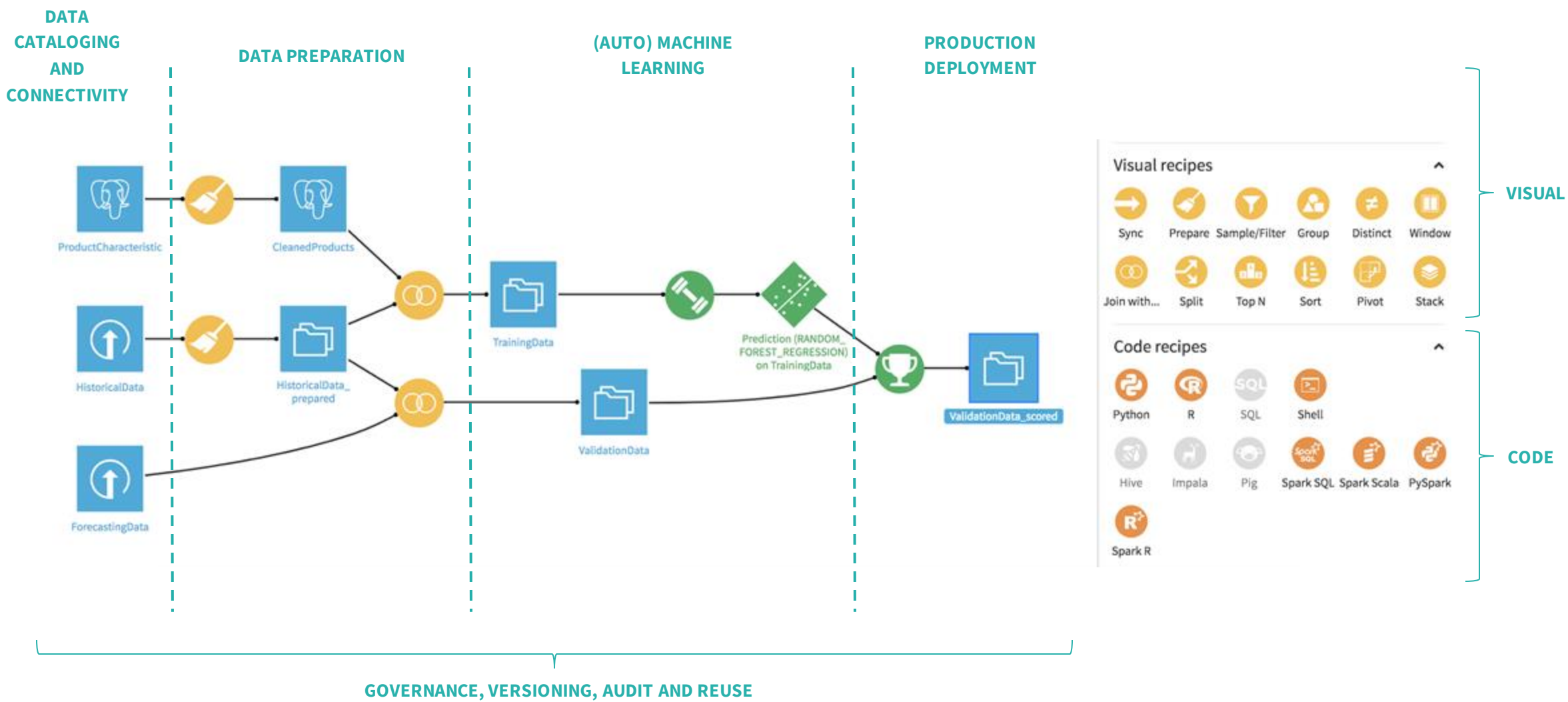
Maximise your usage of **cutting edge technologies**

Visualize and extend with different tools

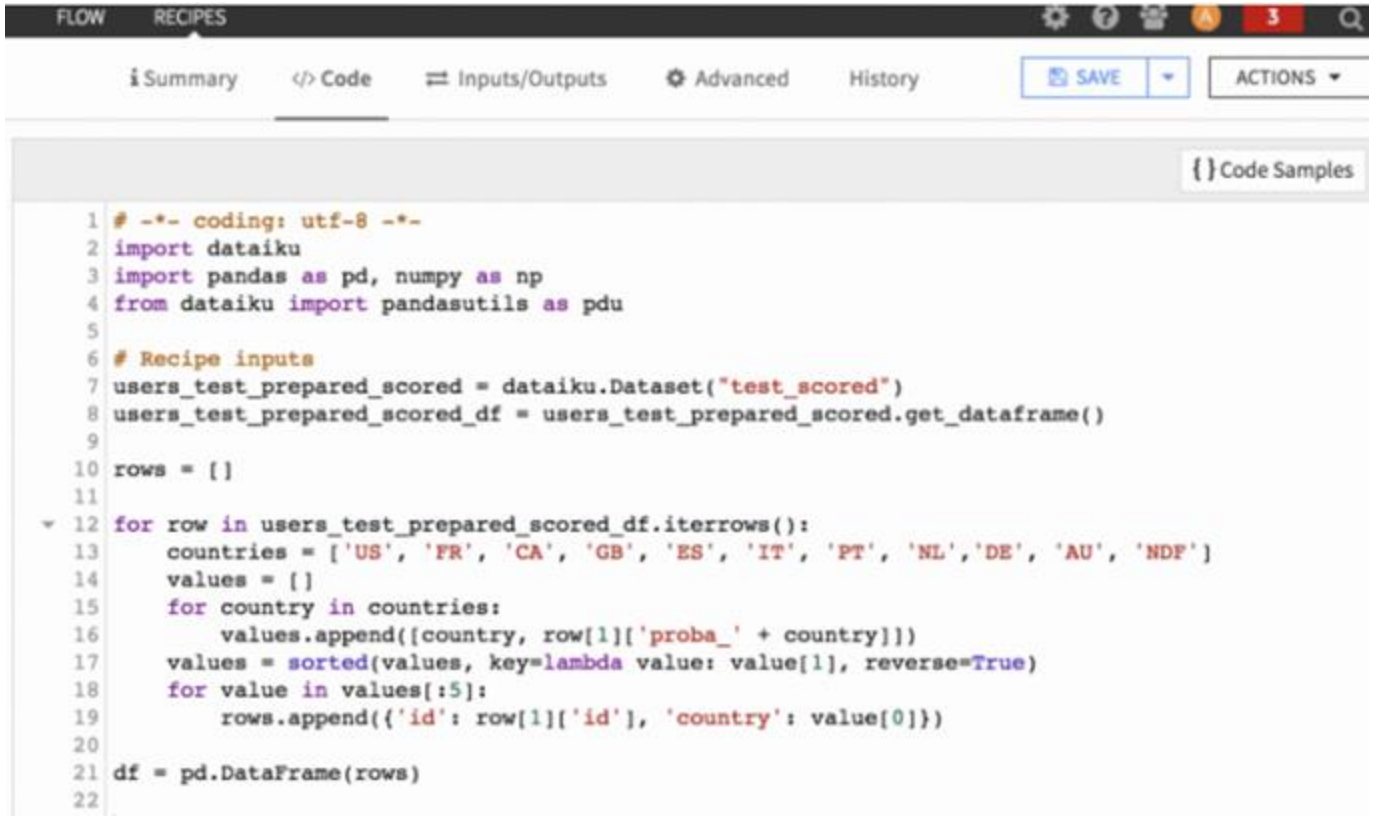
DSS: A collaboration platform



DSS for visual users



DSS for coders (R, Python...)



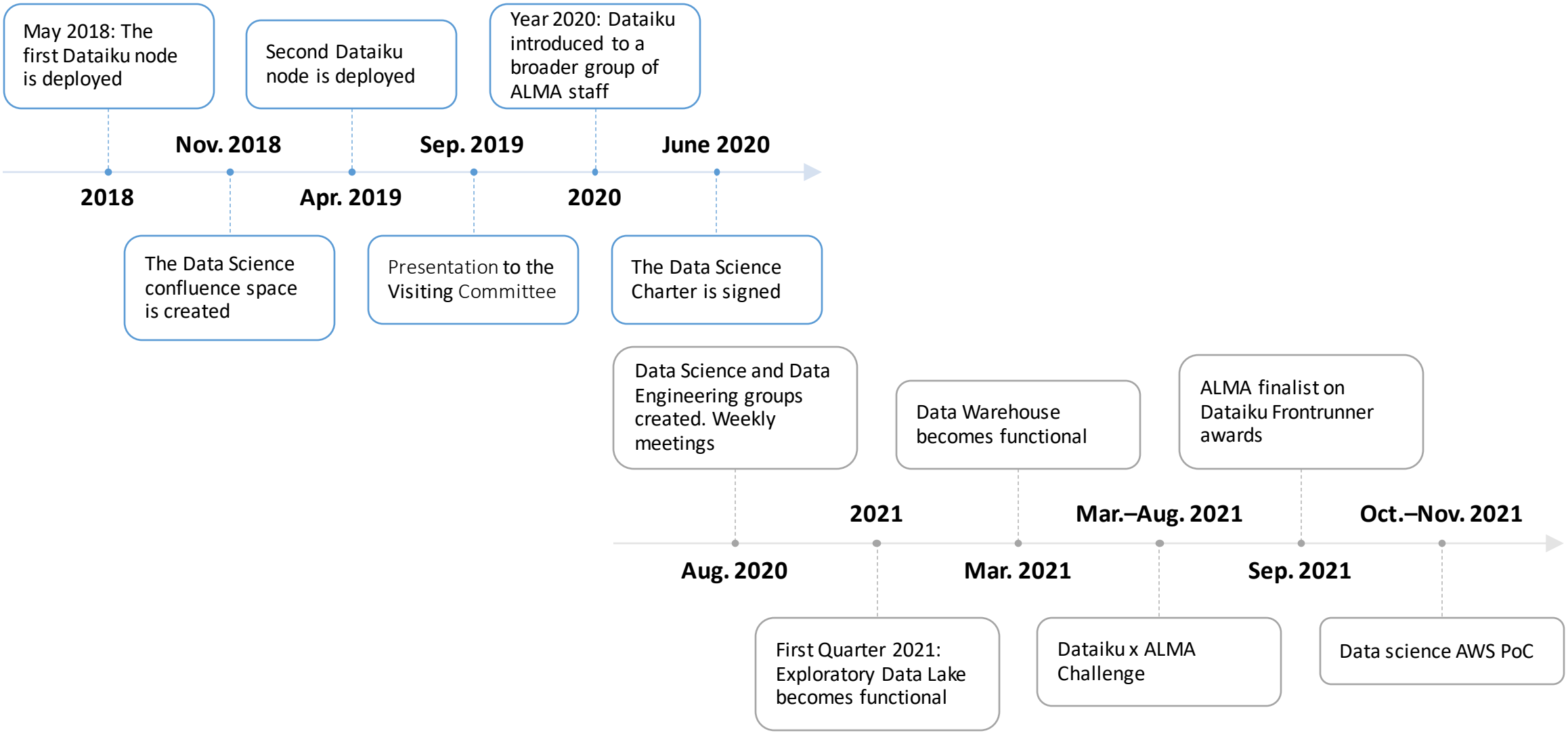
Interactive Python, R, SQL notebooks **and more**

Code and share your own recipes

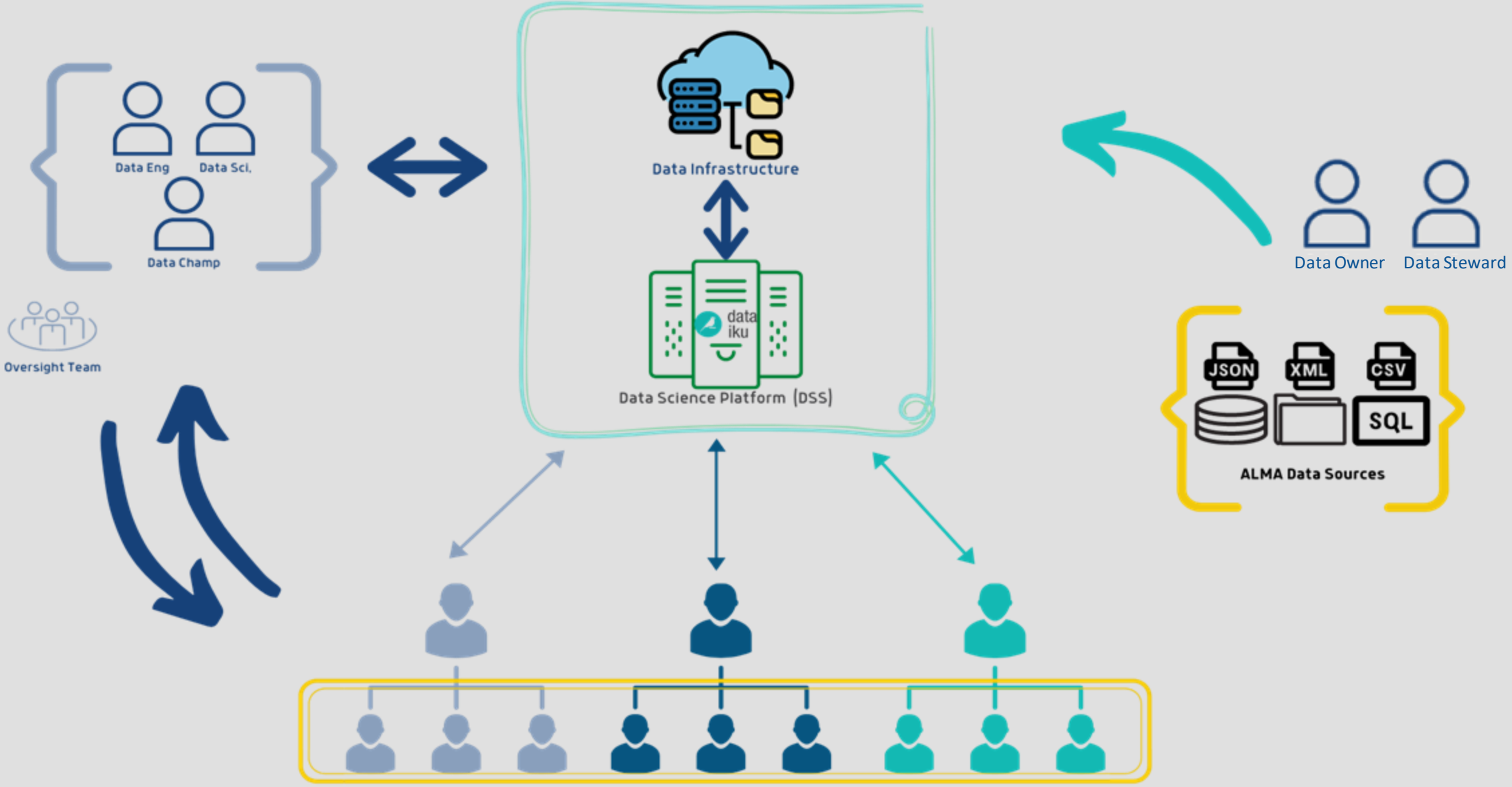
Code your own visualizations

Create reusable components and environments

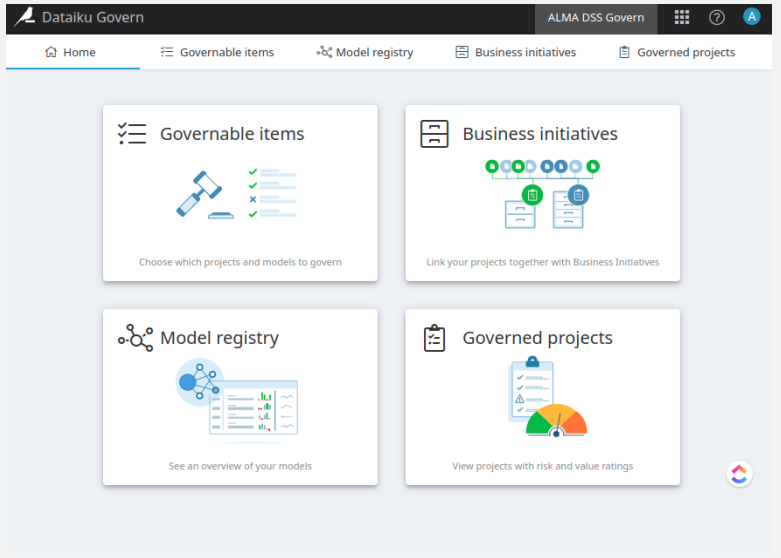
What did we learn and achieve?

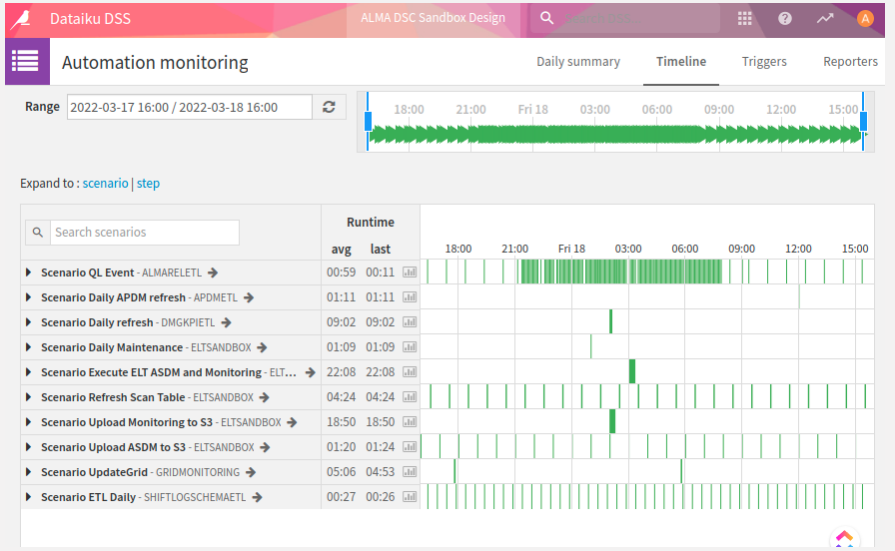


ALMA DS implementation experiment

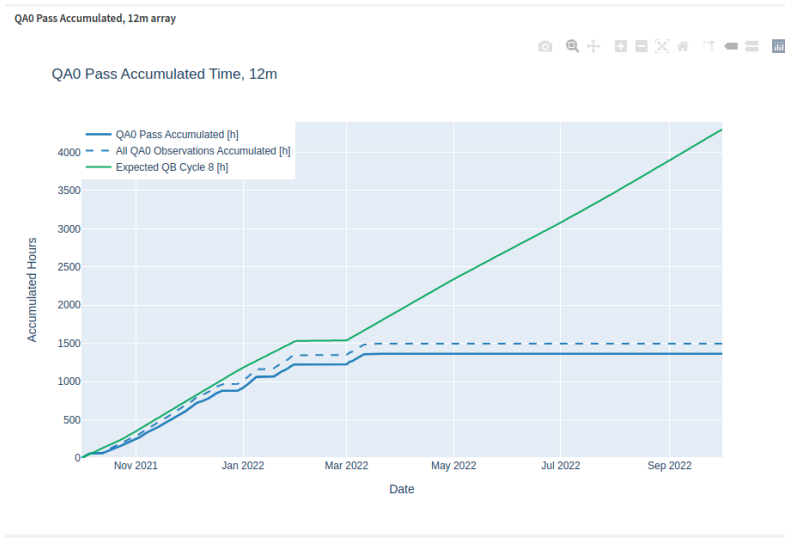


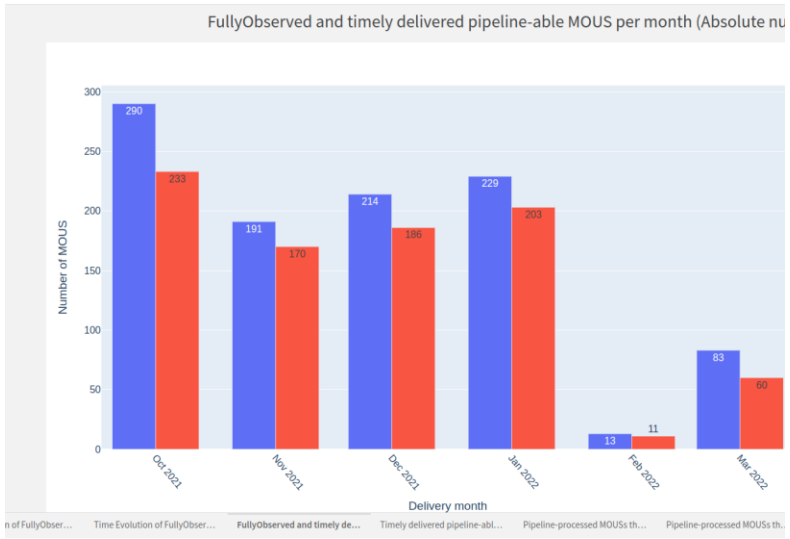
ALMA use cases, today

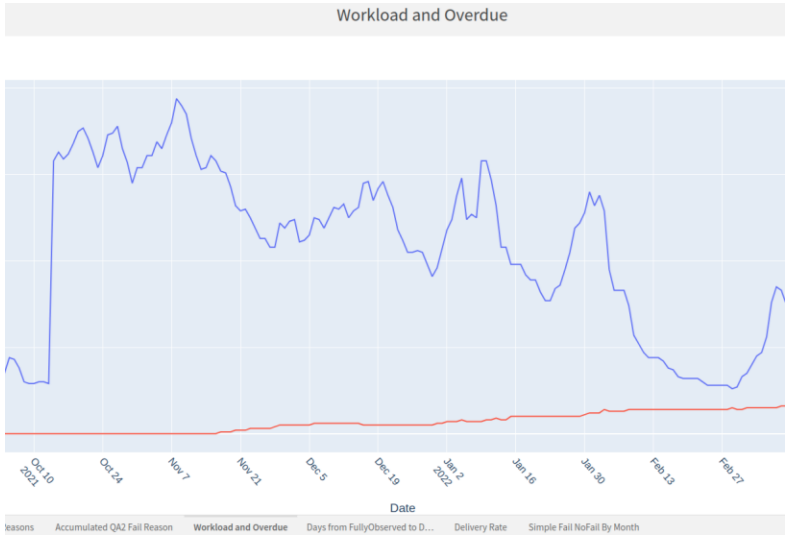


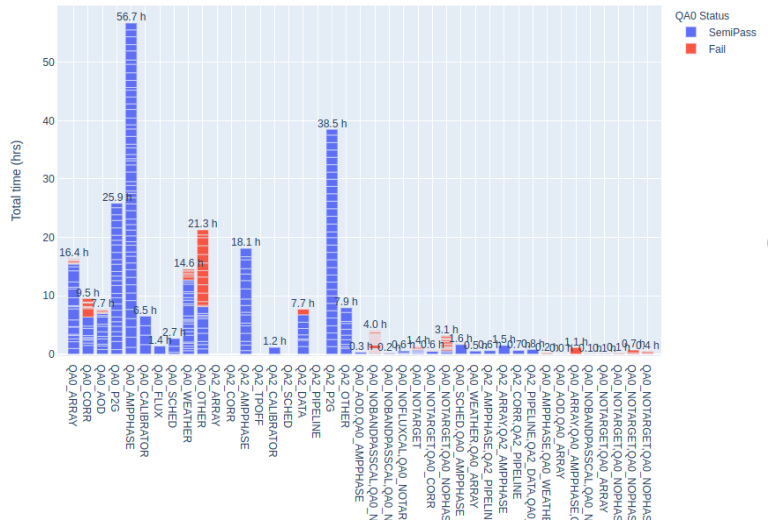


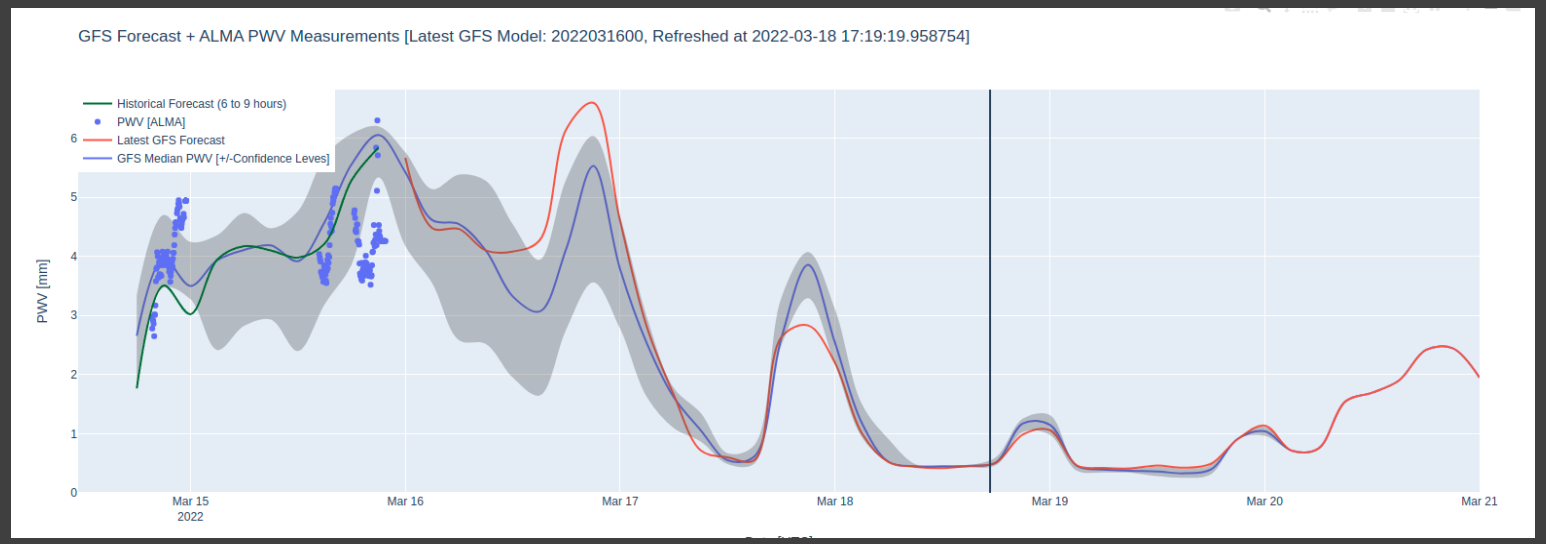
Monitoring data acquisition and processing, daily











Support for decision-making of observational short-term plans

Last update on 2022-03-18T14:53:53, and last shift end time: 2022-03-18T11:59:59

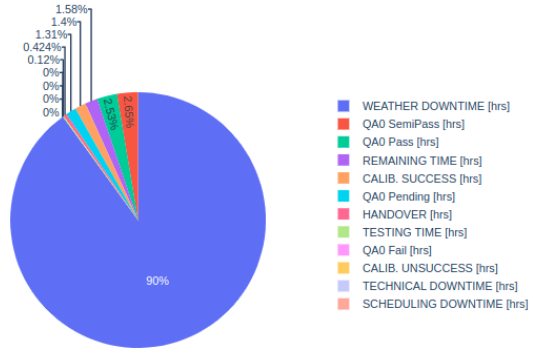
LAST SHIFT

LAST WEEK SHIFTS

SCIENCE EFFICIENCIES FOR THE LAST WEEK SHIFTS:

12M ARRAY | 7M ARRAY | TP ARRAY

Metric	Value
SHIFT DURATION [hrs]	143.621
HANDOVER [hrs]	0.609
QA0 Pass [hrs]	3.631
QA0 Pending [hrs]	1.886
QA0 SemiPass [hrs]	3.811
CALIB. SUCCESS [hrs]	2.009
WEATHER DOWNTIME [hrs]	129.226
TECHNICAL DOWNTIME [hrs]	0
SCHEDULING DOWNTIME [hrs]	0
REMAINING TIME [hrs]	2.264
OPERATION EFFICIENCY WITH CALIB. [%]	41.42
OBSERVATION EFFICIENCY WITH CALIB. [%]	41.42
TESTING TIME RATE [%]	1.249
Time lost rate due to Fails [%]	27.995
CALIBRATION RATE [%]	35.619
FAILURE RATE [%]	46.154
TIME_COMPLETENESS	1
BACKLOG RATE [%]	42.857



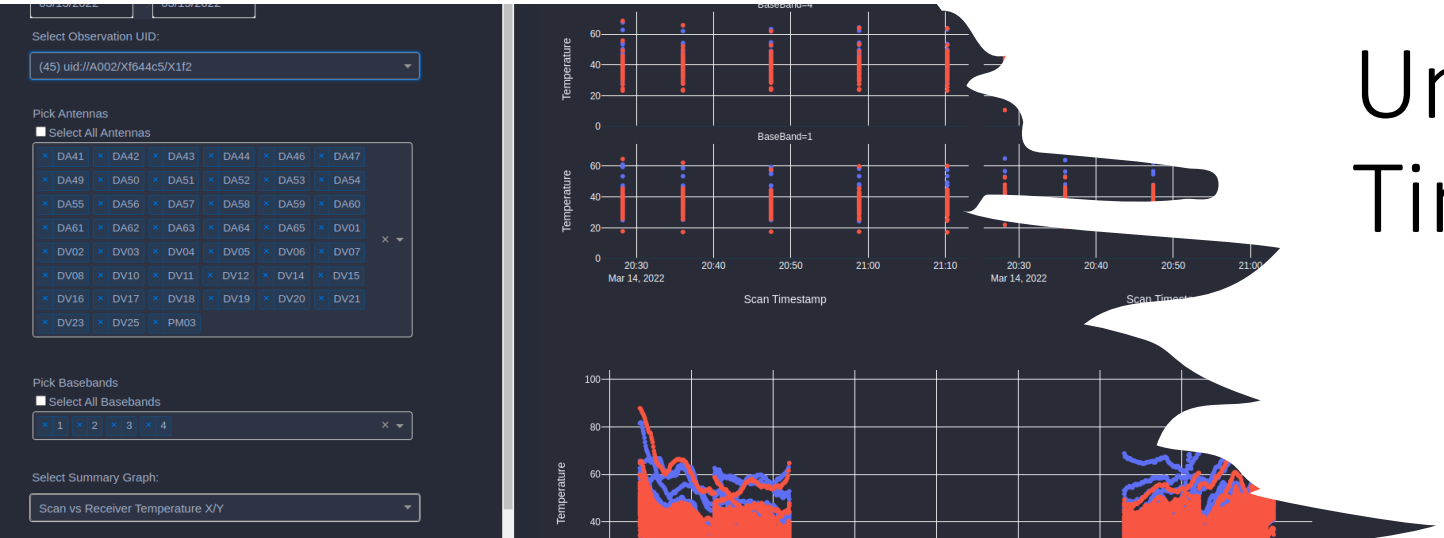
WEBAPP:

CLASS:
- GAP
- USER ERROR

PRIORITY:
- Blocker
- High
- None
- Medium
- Low

TYPE:
- Non activities shift
- Handover
- Array recreation
- Failed execution gap
- Other
- Technical downtime
- End shift
- Weather downtime
- Allocated time not well defined
- EOC shift overlapping a SCIOPS shift
- Overlapping shift
- PI executions in downtime
- Scheduling downtime
- Downtime not completed
- Aborted execution gap
- Testing time gap
- Uncertain shift start time
- Array with wrong activity
- Running execution gap
- PI execution with wrong calibration/test flag

END_TIME	CLASS	TYPE	TAGS	TIME_DURATION [hrs]	Comments	Priority	Array Fa
2022-03-12 22:47:49	GAP	Handover	NaN	1.7967	NaN	Low	Total Po
2022-03-12 21:20:13	GAP	Handover	Band_3_6_7	0.3369	NaN	Low	12 [m]
2022-03-12 21:36:56	GAP	Handover	Band_3_6_7	0.6153	NaN	Low	7 [m]
2022-03-12 21:42:20	GAP	Array recreation	PI science	0.1544	NaN	None	12 [m]
2022-03-12 21:36:55	USER ERROR	Allocated time not well defir	Timestamp non defined	0	Check the start/end time.	Medium	7 [m]
2022-03-12 22:09:09	GAP	Other	NaN	0.2361	NaN	None	7 [m]
2022-03-12 22:19:15	GAP	Other	NaN	0.0906	NaN	None	12 [m]
2022-03-12 22:46:12	GAP	Aborted execution gap	NaN	0.1006	NaN	None	7 [m]
2022-03-12 23:07:45	GAP	Failed execution gap	NaN	0.1419	Missing or not well defin…	Medium	Total Po
2022-03-13 16:12:01	GAP	Array recreation	PI science	0.4967	NaN	None	12 [m]
2022-03-13 16:23:41	GAP	Other	NaN	0.1225	NaN	None	7 [m]
2022-03-13 20:51:49	GAP	Other	NaN	0.0914	NaN	None	12 [m]
2022-03-14 20:29:37	GAP	Handover	NaN	0.0289	NaN	Low	
2022-03-14 19:15:55	GAP	Handover	IF Delay measurement (def	0.265	NaN	Low	
2022-03-14 19:48:48	GAP	Array recreation	PI science	0.2728	NaN		



Understanding Observing Time Efficiency

Current ALMA DSS Platform Activity Summary

200 PROJECTS, OF WHICH 40 OR SO ARE ACTIVELY USED BY THE SCIENCE DEPARTMENT, 5 BY THE ENGINEERING DEPARTMENT, AND A SIMILAR NUMBER BY COMPUTING.

60 SCENARIOS RUNNING (AUTOMATION).

AN AVERAGE OF 15 "ANALYSTS" USERS WORKING DAILY.

320 DASHBOARDS

3474 INSIGHTS

20 API SERVICES

1107 NOTEBOOKS (PYTHON, R)

2 ARCS EXPLORING

The meeting point: working together, collaboratively

The inspiration comes from the Dataiku x ALMA Challenge done in 2021.

Working with Quality Assurance Data to detect anomalies

How things worked out

Data:

- Started with 18 tables containing around 1 million records
- End up focusing on 2 bigger tables: 17.5 million records & 715,000 x 2 spectrums

Collaboration:

- Weekly meetings of one hour from April 15th until early September
- Usually 5 participants of the community + 2 or 3 “Dataikers” + 2 or 3 ALMAers

Process:

- From understanding the data, preparing it, modelling it thanks to the subject-matter experts help, doing exploratory work, and finally some insights and ideas put to the test
- From a really ambitious QA general process to focusing on 3 areas that could help to improve the QA and diagnostics in the future:
    - Outliers detection using non-parametric methods
    - Trend analysis (seeking changes in the behavior of a measure or metric)
    - Leveraging the use of Webapps (with dash and plotly) to produce interfaces for final consumers

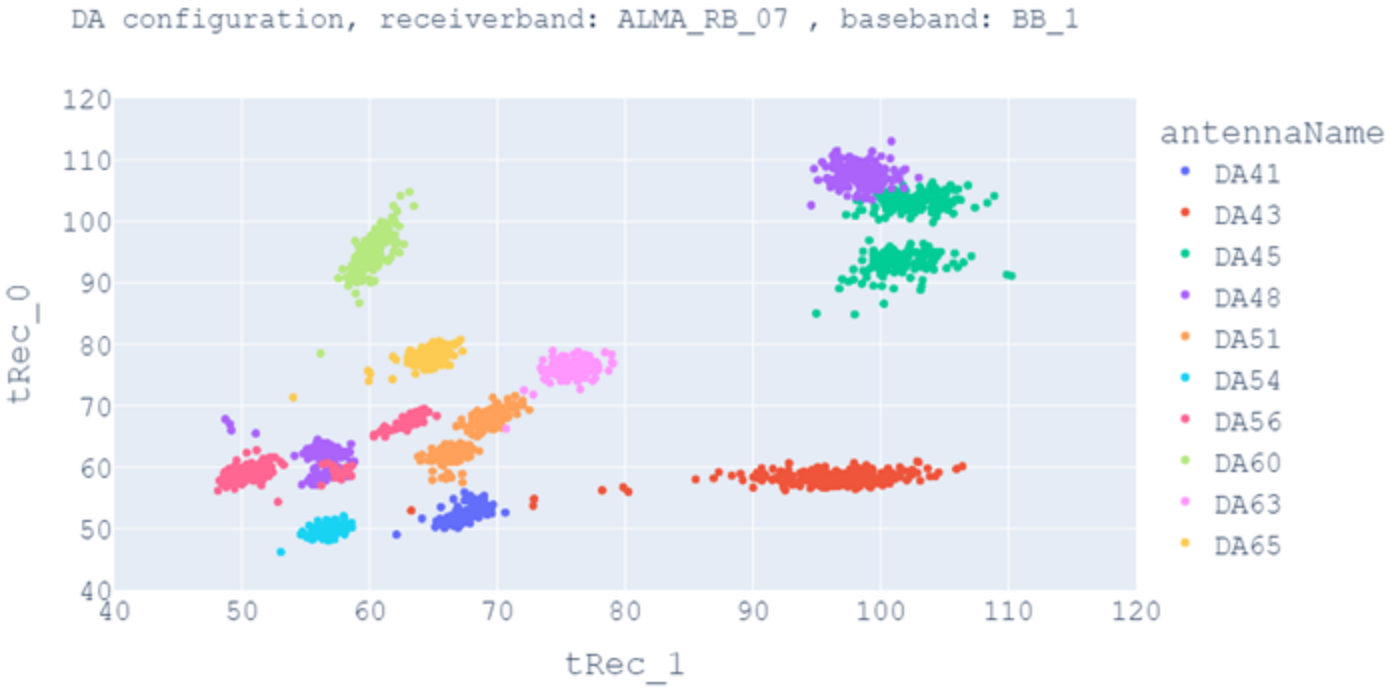
Trend Analysis and Outlier detection

Insights Antenna Receiver Temperature in History

Goal: Identify outliers in antennas or trend analysis that can support the QA process by finding 'badly' operating antennas

⇾ Focus on data of the receiver temperature of the antenna during calibration measurements.

Can we find outliers per execution block?



Exploratory visualization shows that antenna's occupy distinct space in receiver temperature, even though theoretically they are build the same.

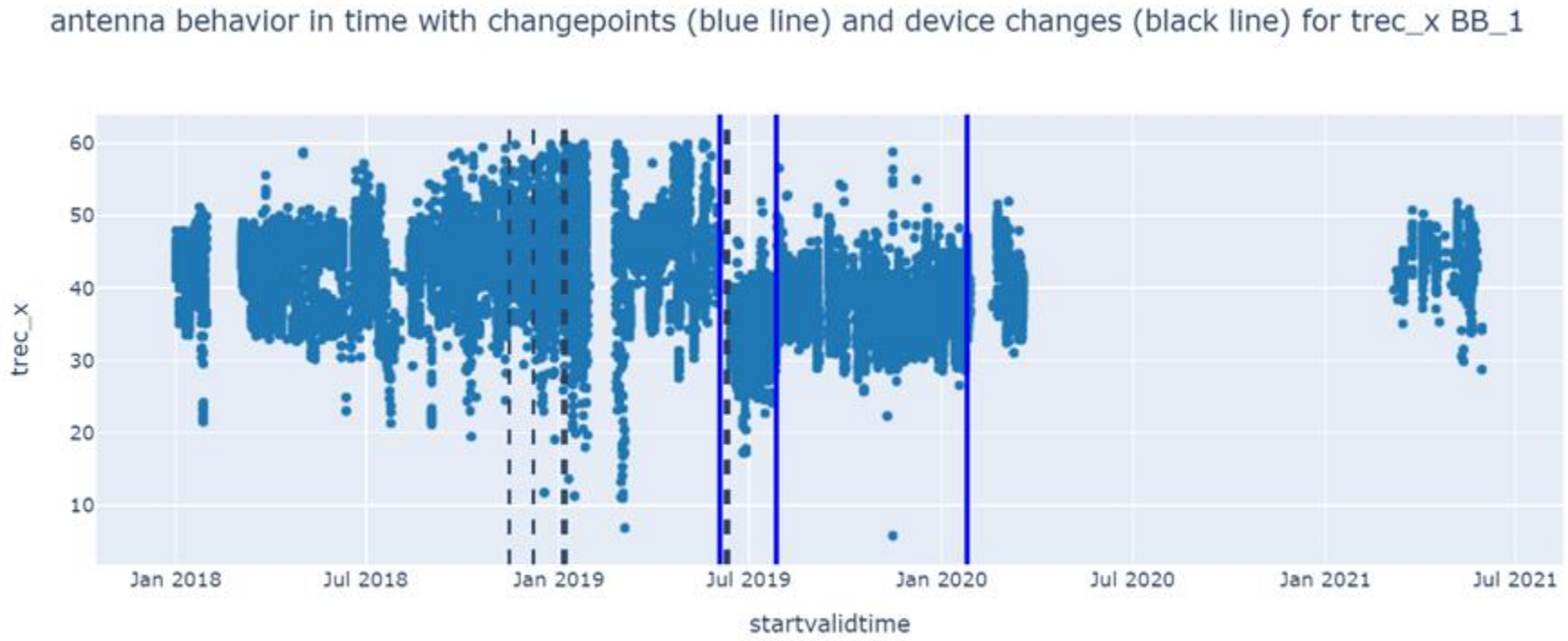
⇾ Find outliers based on the history

Analysis Antenna Temperature

Change Point Analysis

Change point detection tries to identify times when the probability distribution of a stochastic process or time series change. It can identify changes in a distribution in the mean and variance (depending on the specific algorithm and cost function chosen).

Idea: A changepoint can indicate an early warning signal for an engineer that the performance of the device is getting ‘worse’ and maintenance should be planned. Implementing changepoint analysis online since last change of device history can aid an astronomer on duty.



Summary

- AI/ML has been proved to be an essential tool to understand and improve complex tasks (either in science research or operations)
- To be applied in operations, we need to have the technologies and frameworks that solve the particular challenges of this area
- Our organizations should be building the platforms, teams and skills that will enable its use in a daily basis
  - Data Engineering teams are required.
  - Data scientists and analyst should work hand to hand with the domain experts.
  - A modern data framework is fundamental for future success.
- The risk of not building these foundations is to have a bunch of interesting and expensive projects that do not add value to the observatory.

Collaborators and users

- Tomas Staig
- Sergio Pavez
- Rosita Hormann
- Jorge Garcia
- Jose Luis Ortiz
- Mark Gallilee
- Maxs Simmonds
- Jose Lobos
- Nicolas Ovando
- Gaston Velez
- Juan Uribe
- Jorge Avarias
- Cristobal Achermann
- Barbara Sepulveda
- Cesar Zapata
- Marcos Ortega
- Takeshi Okuda
- Stephan Gairing
- Giorgio Siringo

- Rodrigo Cabezas
- Drew Brisbin
- Alejandro Barrientos
- Bernhard Lopez
- Celia Verdugo
- Kurt Plarre
- Gabriel Marinello
- Juan Cortes
- José Fernandez
- Laura Gomez
- Andres Guzman
- Mario Garces
- Matias Radiszcz
- Andres Perez-Sanches
- Carmen Toribio
- Sergio Martin
- Juan Millar
- Ruediger Kneissl

Dataiku and Friends:

- Lisa Bardet
- Leo Dreyfus-Schmidt
- Aimee Coelho
- Darien Mitchell-Tontar
- Giuseppe Naldi
- Pauline van Nies
- Niklas Muennighoff
- Matthieu Scordia
- Tom Brown
- Marc Robert
- Jack Craft
- Jordan Blair
- Bruno Carvalho
- Akshay Katre
- Augustin Ador
- AWS: Agustin Grangetto, Jorge Sierra et al
- TARS: Javier Mancilla, Ignacio Duque et al

Ikig.AI - Win-Win

- For Dataikers 
  - Outside-the-box projects for positive impact beyond business

- For your Nonprofit
  - Visibility on blog, annual conference, client portfolio
  - Free optimisation of means to advance mission

Ikigai@dataiku.com

Ikig.AI: The ALMA Observatory

- Atacama Large Millimeter/submillimeter Array
- Whole data pipeline connection
- Stakeholders collaboration - data scientists, engineers, astronomers

More info:

- "The State of Data in Astronomy"
- "ALMA Observatory: Building a Revolutionary Data Science Culture"
- "The Potential for Using Deep Learning to Improve Local Weather Forecasts"
- "Data for Good: Insights From the ALMA Volunteer Challenge"
- Webinare : Building a Data-Centric Culture at the ALMA Observatory