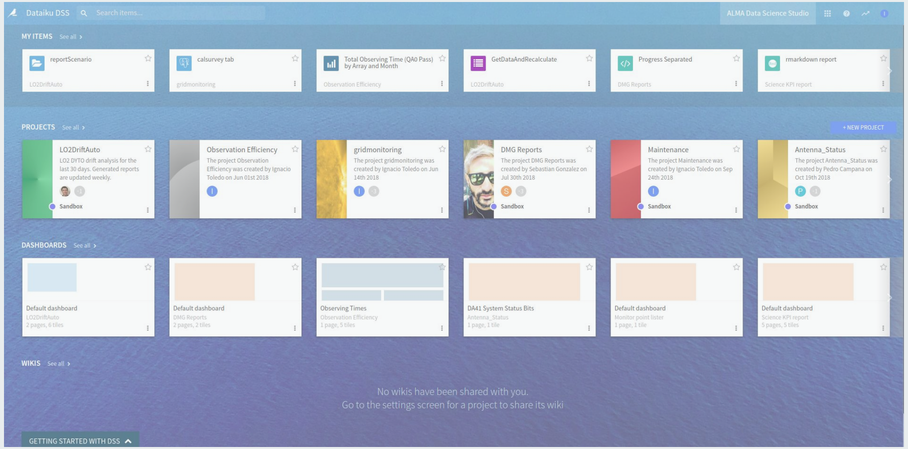
Data Science, not Software Engineering.

Exploring a workflow for ALMA operations.



reportScenario
LO2DriftAuto

calsurvey tab
gridmonitoring

Total Observing Time (QA0 Pass) by Array and Month
Observation Efficiency

GetDataAndRecalculate
LO2DriftAuto

Progress Separated
DMG Reports

rmarkdown report
Science KPI report

LO2DriftAuto
LO2 DYTO drift analysis for the last 30 days. Generated reports are updated weekly.
Sandbox

Observation Efficiency
The project Observation Efficiency was created by Ignacio Toledo on Jun 01st 2018

gridmonitoring
The project gridmonitoring was created by Ignacio Toledo on Jun 14th 2018

DMG Reports
The project DMG Reports was created by Sebastian Gonzalez on Jul 30th 2018
Sandbox

Maintenance
The project Maintenance was created by Ignacio Toledo on Sep 24th 2018
Sandbox

Antenna_Status
The project Antenna_Status was created by Pedro Campana on Oct 19th 2018
Sandbox

Default dashboard
LO2DriftAuto
2 pages, 6 tiles

Default dashboard
DMG Reports
2 pages, 2 tiles

Observing Times
Observation Efficiency
1 page, 5 tiles

DA41 System Status Bits
Antenna_Status
1 page, 1 tile

Default dashboard
Monitor point lister
1 page, 1 tile

Default dashboard
Science KPI report
5 pages, 5 tiles

No wikis have been shared with you.
Go to the settings screen for a project to share its wiki

Ignacio Toledo, JAO.

ADASS xxvii

Abstract

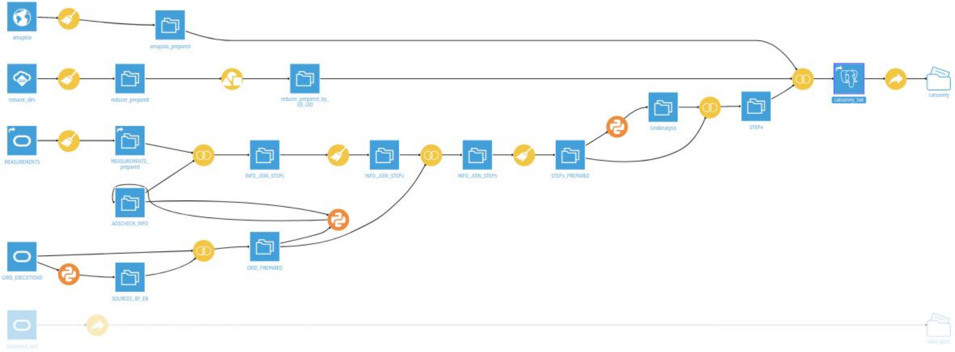
In the last few years Data science has emerged as a discipline of its own to address problems where data is usually heterogeneous, complex and abundant. In a nutshell, data science allows to provide answers to situations where a hypothesis can be formulated and later can be either confirmed or rejected following standard scientific methodology using data as raw material.

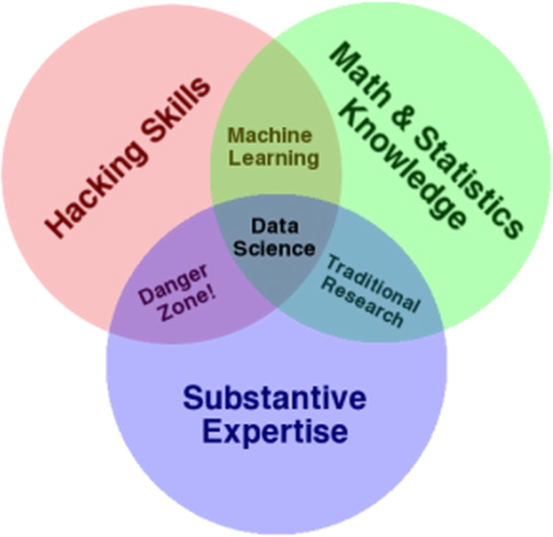
Data science has been called differently depending of the domain (business intelligence, operational management, astroinformatics) and it has been recently in the center of a hype related to artificial intelligence and machine learning. It has been quickly adopted by the digital industry as the tool to distill information of massive operational data sets. Among the many tools data science requires (mathematics, statistics, domain knowledge of the data sets, …), IT infrastructure and software is by far the most visible and there is at present a whole ecosystem available as open source projects. The downside of this is data science is commonly confused with IT and software development, which creates conflicts between engineering- and scientific- mindsets, and leads to wrongly applying software development methodologies to it neglecting the experimental nature of the problem. In summary, creating the data lab becomes more important than answering questions with it. In the domain of ALMA operations, there are many instances that can be identified and described as data science cases or projects ranging from monitoring array elements to understand performances and predict faults for engineering operations to routine monitoring of calibrators for science operations purposes.

We have identified already around 30 different initial questions (or data science cases) and found that several of them have been addressed through individual efforts. In parallel, several enabling platforms or frameworks have appear in the ecosystem that provides data scientists with both the “laboratory equipment” to conduct their “experiments” as well as enabling tools for collaboration, versioning control, and deploying results in production with a quick turnaround. This talk aims to summarize the results of our exploration to apply data science workflows to resolve ALMA operations issues, identify suitable platforms that are already in use by the industry, share our experience in addressing specific ALMA operations data cases, and discuss the technical and sociological challenges we encountered along the way.

Overview

- What is Data Science (at least for the next 10 minutes)
- Why focusing in the differences with Software Engineering?
- ALMA - Data iku Experience:
    - Don’t reinvent the wheel
    - Take advantage of your strengths
    - Socialize, communicate, collaborate
    - Our challenge: Data Engineering!

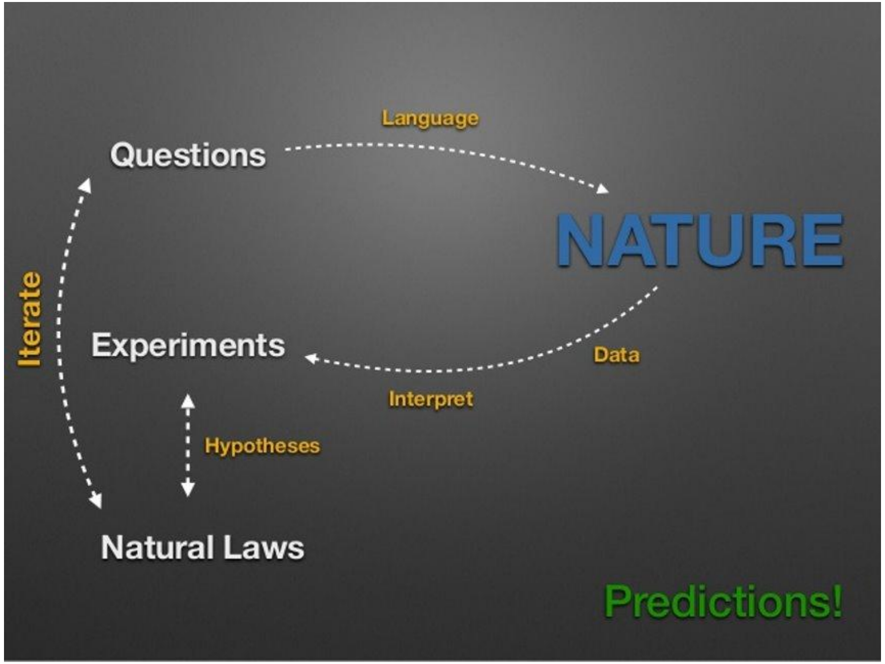


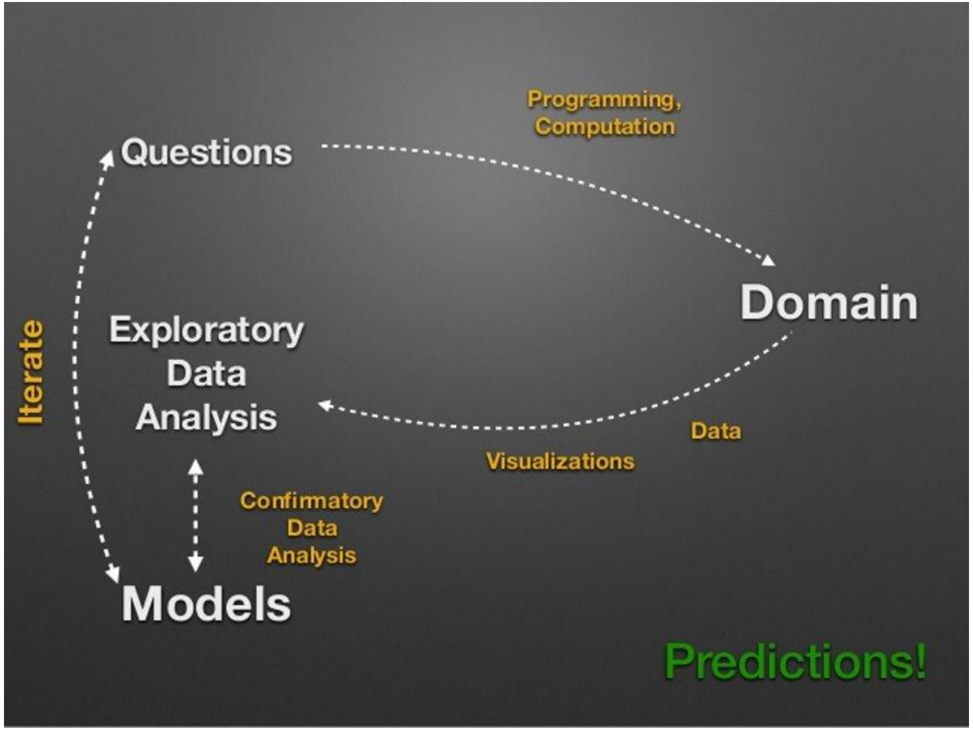


“Data science is the discipline of making data useful.”

“Data science is science… it just starts with raw data exploration instead of physical observation”

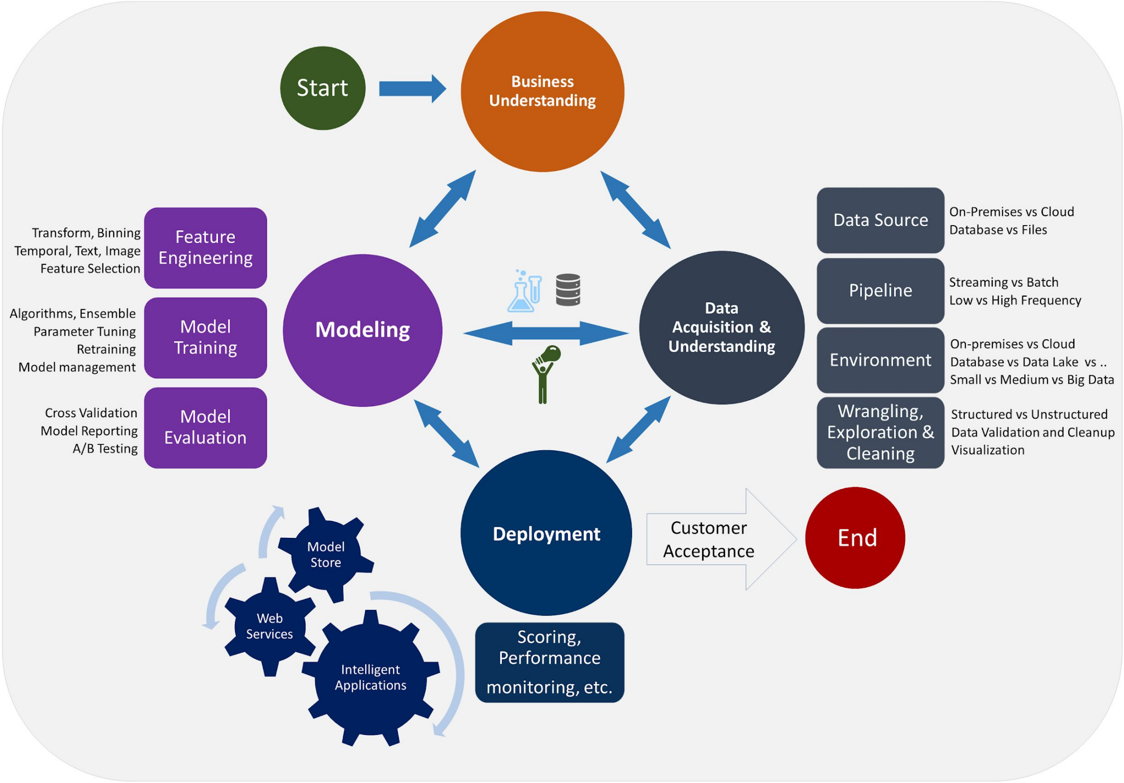
What is Data Science?

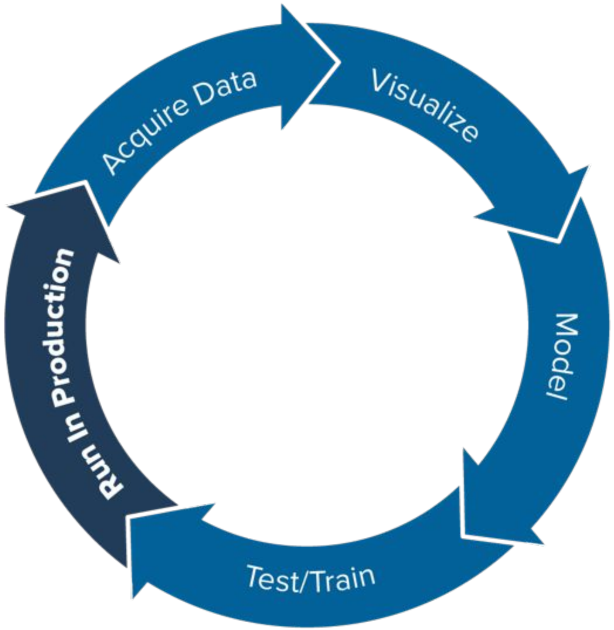




A new methodology for an old process

Data Science Lifecycle





A different methodology

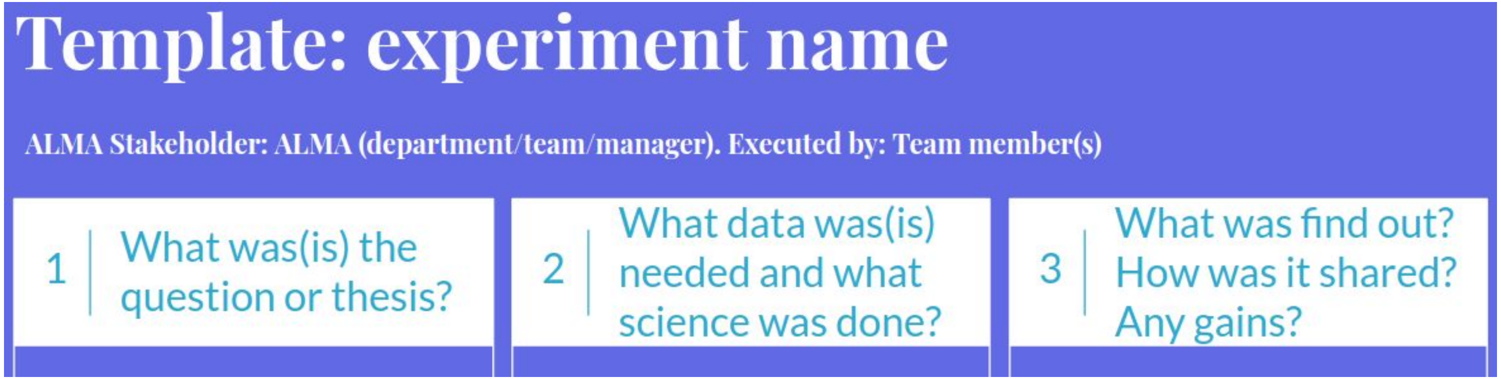
How is the observatory time distributed among the different executives?

Can we automate early detection or prediction of hardware failures leading to observation downtime or degradation? Can we automate diagnosis?

There is a need for a general set of tools to explore relationships between disparate arbitrary data streams as a support for problem investigation.

Hardware malfunction fault detection and diagnosis (FDD)

Given the current configuration and the forthcoming, how much pressure do we have of Science Projects, by band and LST?



Template: experiment name

ALMA Stakeholder: ALMA (department/team/manager). Executed by: Team member(s)

1 | What was(is) the question or thesis?

2 | What data was(is) needed and what science was done?

3 | What was find out? How was it shared? Any gains?

Detect and diagnose IFPs amplifiers degradations on any antenna on 2 polarizations each, 4 basebands each polarization. Degradation leads to TP detectors readings calibration drift.

Quantify the dependency of photonic reference failure rate as a function of various factors to identify strategies for improvement

Track Needed Grid Executions, reduction & Ingestions: Observing Log of Grid data. Which sources need to be observed? Which sources/datasets need to be reduced/ingested ? Who is reducing the data ?

Are these software engineering or data science problems? Hint: check the outcome.

What are we doing at ALMA?

The Data Science initiative

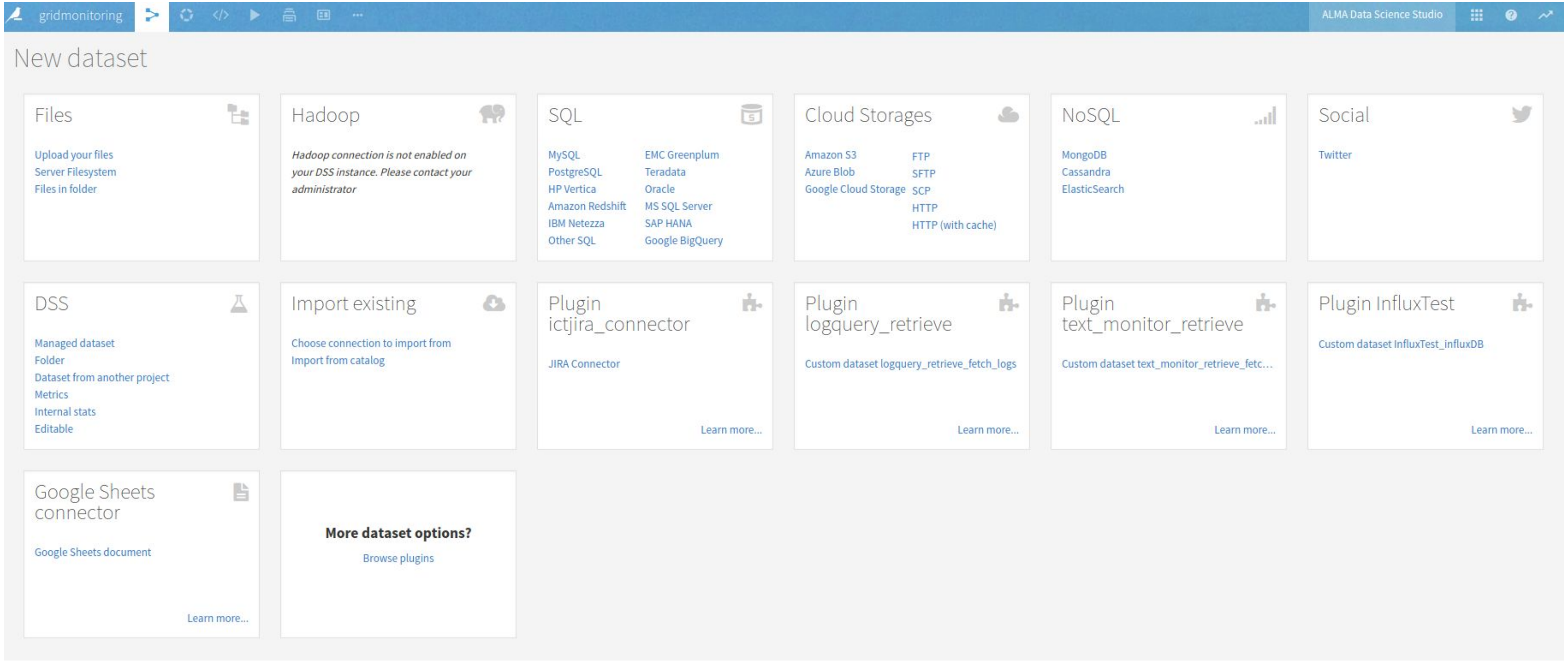
- Not much resources: do not reinvent the wheel!
- Astronomers and engineers with statistical, mathematical, domain knowledge and coding skills. Work with them!
- Collaborate, socialize, communicate!
- Show results.

A data science platform collaboration

ALMA

data
iku

Make data accessible



gridmonitoring

ALMA Data Science Studio

New dataset

Files

Upload your files
Server Filesystem
Files in folder

Hadoop

Hadoop connection is not enabled on your DSS instance. Please contact your administrator

SQL

MySQL
PostgreSQL
HP Vertica
Amazon Redshift
IBM Netezza
Other SQL
EMC Greenplum
Teradata
Oracle
MS SQL Server
SAP HANA
Google BigQuery

Cloud Storages

Amazon S3
Azure Blob
Google Cloud Storage
FTP
SFTP
SCP
HTTP
HTTP (with cache)

NoSQL

MongoDB
Cassandra
ElasticSearch

Social

Twitter

DSS

Managed dataset
Folder
Dataset from another project
Metrics
Internal stats
Editable

Import existing

Choose connection to import from
Import from catalog

Plugin ictjira_connector

JIRA Connector

Learn more...

Plugin logquery_retrieve

Custom dataset logquery_retrieve_fetch_logs

Learn more...

Plugin text_monitor_retrieve

Custom dataset text_monitor_retrieve_fetc...

Learn more...

Plugin InfluxTest

Custom dataset InfluxTest_influxDB

Learn more...

Google Sheets connector

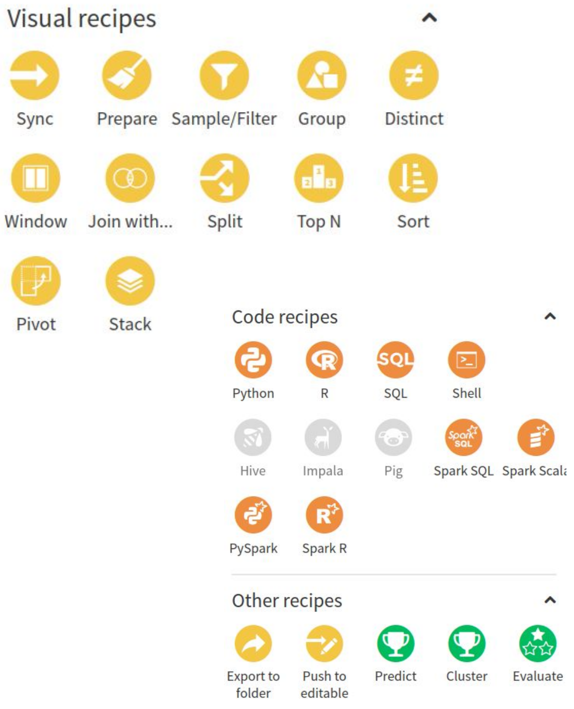
Google Sheets document

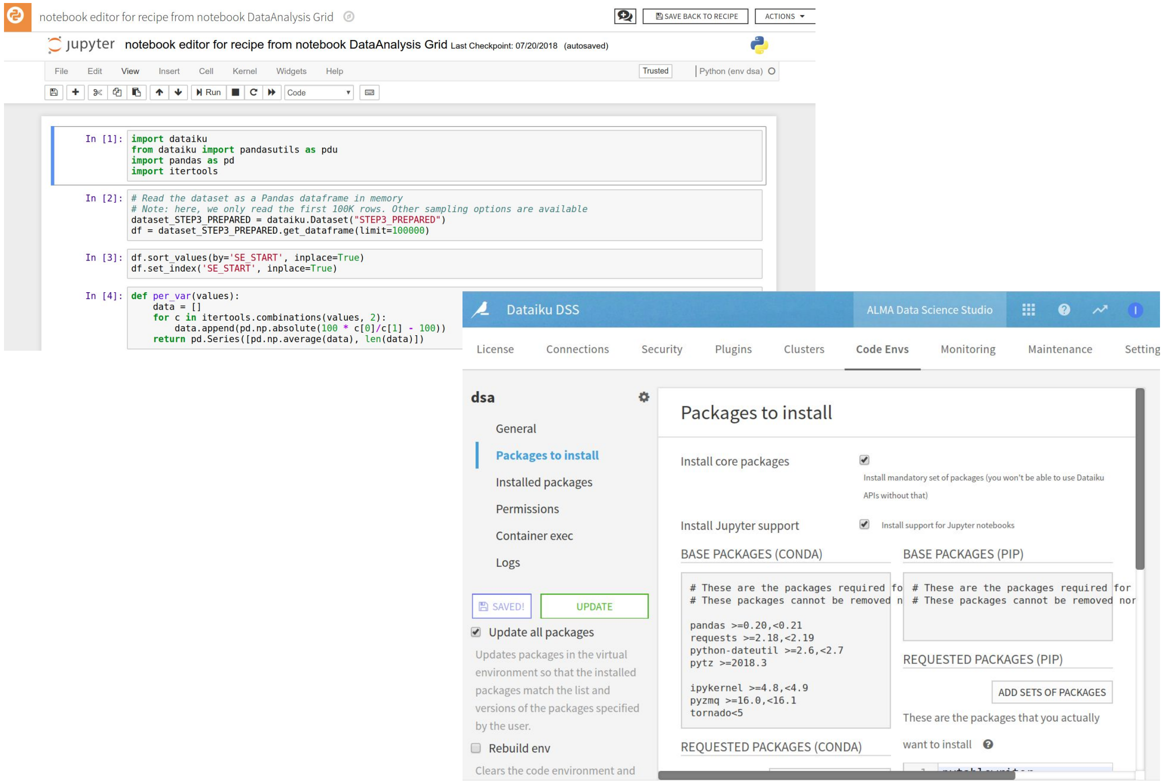
Learn more...

More dataset options?

Browse plugins

Create a laboratory with the tools needed





notebook editor for recipe from notebook DataAnalysis Grid Last Checkpoint: 07/20/2018 (autosaved)

In [1]:
```
import dataiku
from dataiku import pandasutils as pdu
import pandas as pd
import itertools
```

In [2]:
```
# Read the dataset as a Pandas dataframe in memory
# Note: here, we only read the first 100K rows. Other sampling options are available
dataset_STEP3_PREPARED = dataiku.Dataset("STEP3_PREPARED")
df = dataset_STEP3_PREPARED.get_dataframe(limit=100000)
```

In [3]:
```
df.sort_values(by='SE_START', inplace=True)
df.set_index('SE_START', inplace=True)
```

In [4]:
```
def per_var(values):
    data = []
    for c in itertools.combinations(values, 2):
        data.append(pd.np.absolute(100 * c[0]/c[1] - 100))
    return pd.Series([pd.np.average(data), len(data)])
```

Packages to install

Install core packages ☑ Install mandatory set of packages (you won't be able to use Dataiku APIs without that)

Install Jupyter support ☑ Install support for Jupyter notebooks

BASE PACKAGES (CONDA)

```
# These are the packages required fo
# These packages cannot be removed n

pandas >=0.20,<0.21
requests >=2.18,<2.19
python-dateutil >=2.6,<2.7
pytz >=2018.3

ipykernel >=4.8,<4.9
pyzmq >=16.0,<16.1
tornado<5
```

BASE PACKAGES (PIP)

```
# These are the packages required for
# These packages cannot be removed nor
```

REQUESTED PACKAGES (PIP)

These are the packages that you actually want to install

☑ Update all packages

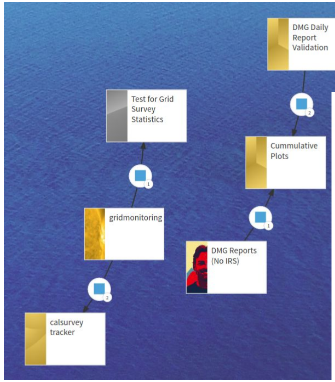
Updates packages in the virtual environment so that the installed packages match the list and versions of the packages specified by the user.

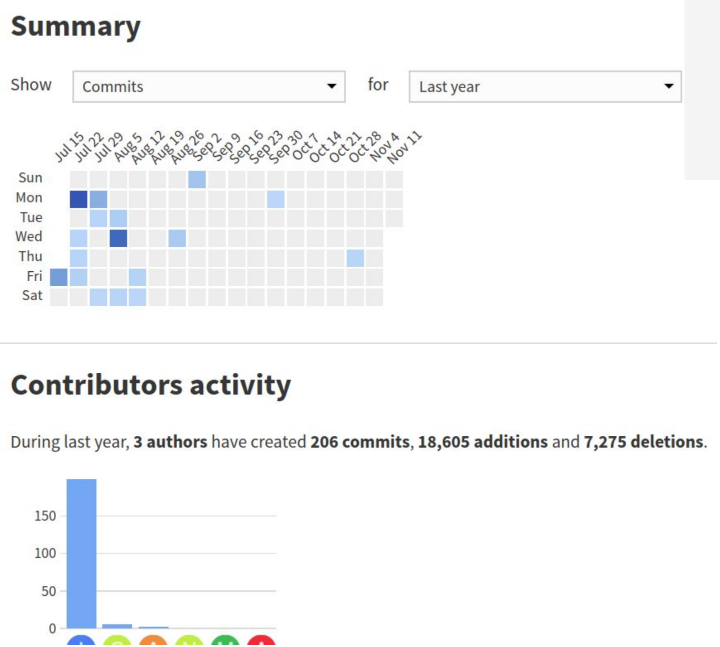
□ Rebuild env

Clears the code environment and

REQUESTED PACKAGES (CONDA)

Collaborate and socialize



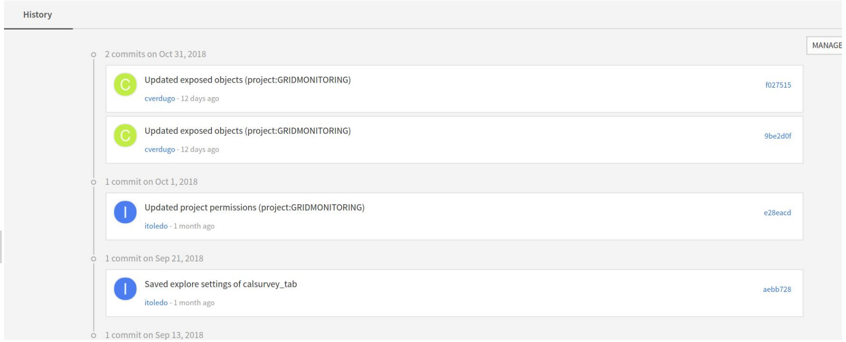


Summary

Show: Commits for: Last year

Contributors activity

During last year, **3 authors** have created **206 commits**, **18,605 additions** and **7,275 deletions**.



History

2 commits on Oct 31, 2018

- Updated exposed objects (project:GRIDMONITORING) — cverdugo - 12 days ago — f027515
- Updated exposed objects (project:GRIDMONITORING) — cverdugo - 12 days ago — 9be2d0f

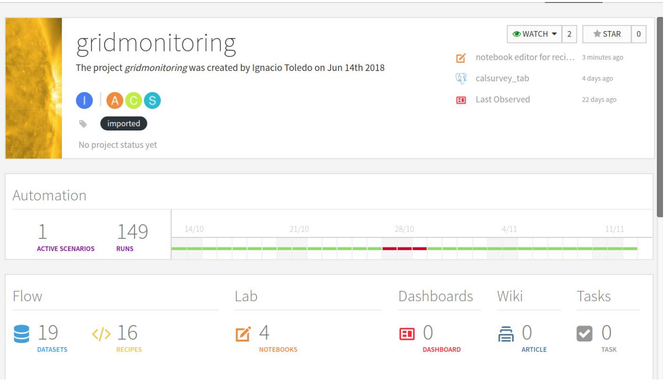
1 commit on Oct 1, 2018

- Updated project permissions (project:GRIDMONITORING) — itoledo - 1 month ago — e28eacd

1 commit on Sep 21, 2018

- Saved explore settings of calsurvey_tab — itoledo - 1 month ago — aebb728

1 commit on Sep 13, 2018



gridmonitoring

The project *gridmonitoring* was created by Ignacio Toledo on Jun 14th 2018

No project status yet

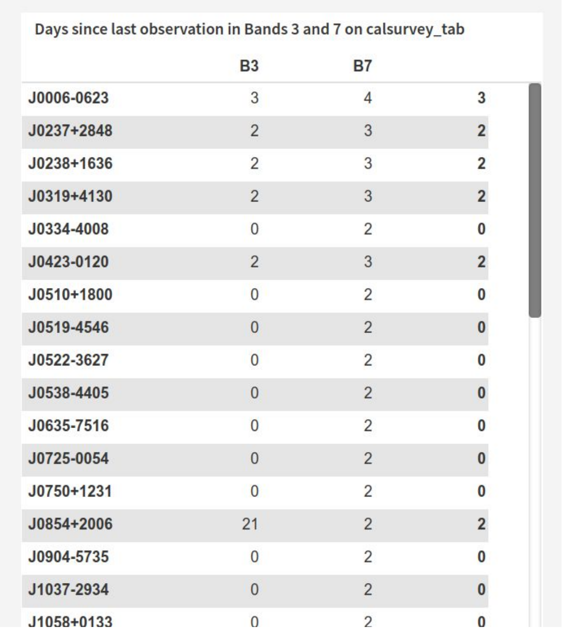
notebook editor for reci…	3 minutes ago
calsurvey_tab	4 days ago
Last Observed	22 days ago

Automation

1 ACTIVE SCENARIOS — 149 RUNS

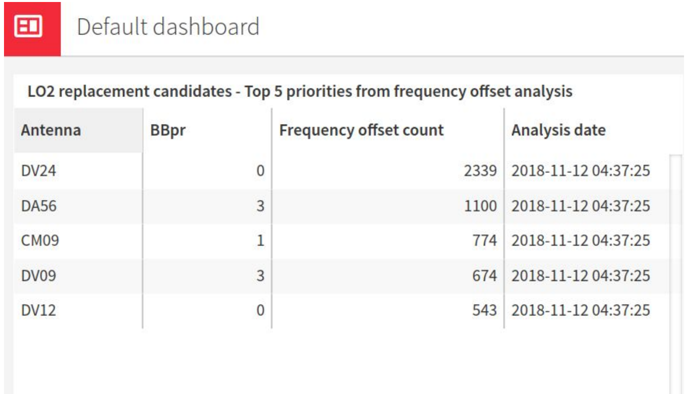
Flow		Lab	Dashboards	Wiki	Tasks
19 DATASETS	16 RECIPES	4 NOTEBOOKS	0 DASHBOARD	0 ARTICLE	0 TASK

Show the results!



Days since last observation in Bands 3 and 7 on calsurvey_tab

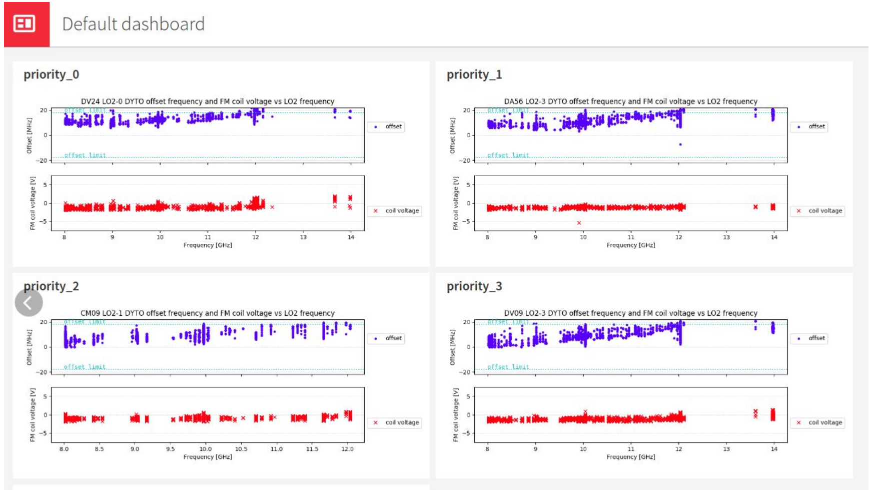
	B3	B7	
J0006-0623	3	4	**3**
J0237+2848	2	3	**2**
J0238+1636	2	3	**2**
J0319+4130	2	3	**2**
J0334-4008	0	2	**0**
J0423-0120	2	3	**2**
J0510+1800	0	2	**0**
J0519-4546	0	2	**0**
J0522-3627	0	2	**0**
J0538-4405	0	2	**0**
J0635-7516	0	2	**0**
J0725-0054	0	2	**0**
J0750+1231	0	2	**0**
J0854+2006	21	2	**2**
J0904-5735	0	2	**0**
J1037-2934	0	2	**0**
J1058+0133	0	2	**0**



Default dashboard

LO2 replacement candidates - Top 5 priorities from frequency offset analysis

Antenna	BBpr	Frequency offset count	Analysis date
DV24	0	2339	2018-11-12 04:37:25
DA56	3	1100	2018-11-12 04:37:25
CM09	1	774	2018-11-12 04:37:25
DV09	3	674	2018-11-12 04:37:25
DV12	0	543	2018-11-12 04:37:25



Default dashboard

priority_0

priority_1

priority_2

priority_3

DV24 LO2-0 DYTO offset frequency and FM coil voltage vs LO2 frequency
DA56 LO2-3 DYTO offset frequency and FM coil voltage vs LO2 frequency
CM09 LO2-1 DYTO offset frequency and FM coil voltage vs LO2 frequency
DV09 LO2-3 DYTO offset frequency and FM coil voltage vs LO2 frequency
Offset [MHz]
20
0
−20
offset limit
offset
FM coil voltage [V]
5
0
−5
coil voltage
Frequency [GHz]
8
9
10
11
12
13
14
8.0
8.5
9.0
9.5
10.0
10.5
11.0
11.5
12.0

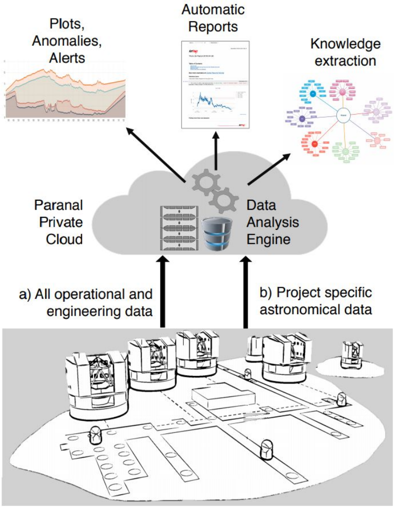
What is Next? A new challenge.

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

Framework to use modern Big Data Software Tools to improve operations at the Paranal Observatory

Eduardo Pena*, Ricardo Schmutzer, Christian Stephan, Claudio Reinero, Julien Milli, Juan C. Guerra, Juan Osorio European Southern Observatory, Alonso de Córdova 3107, Vitacura, casilla 19001, Santiago, CHILE



Questions?