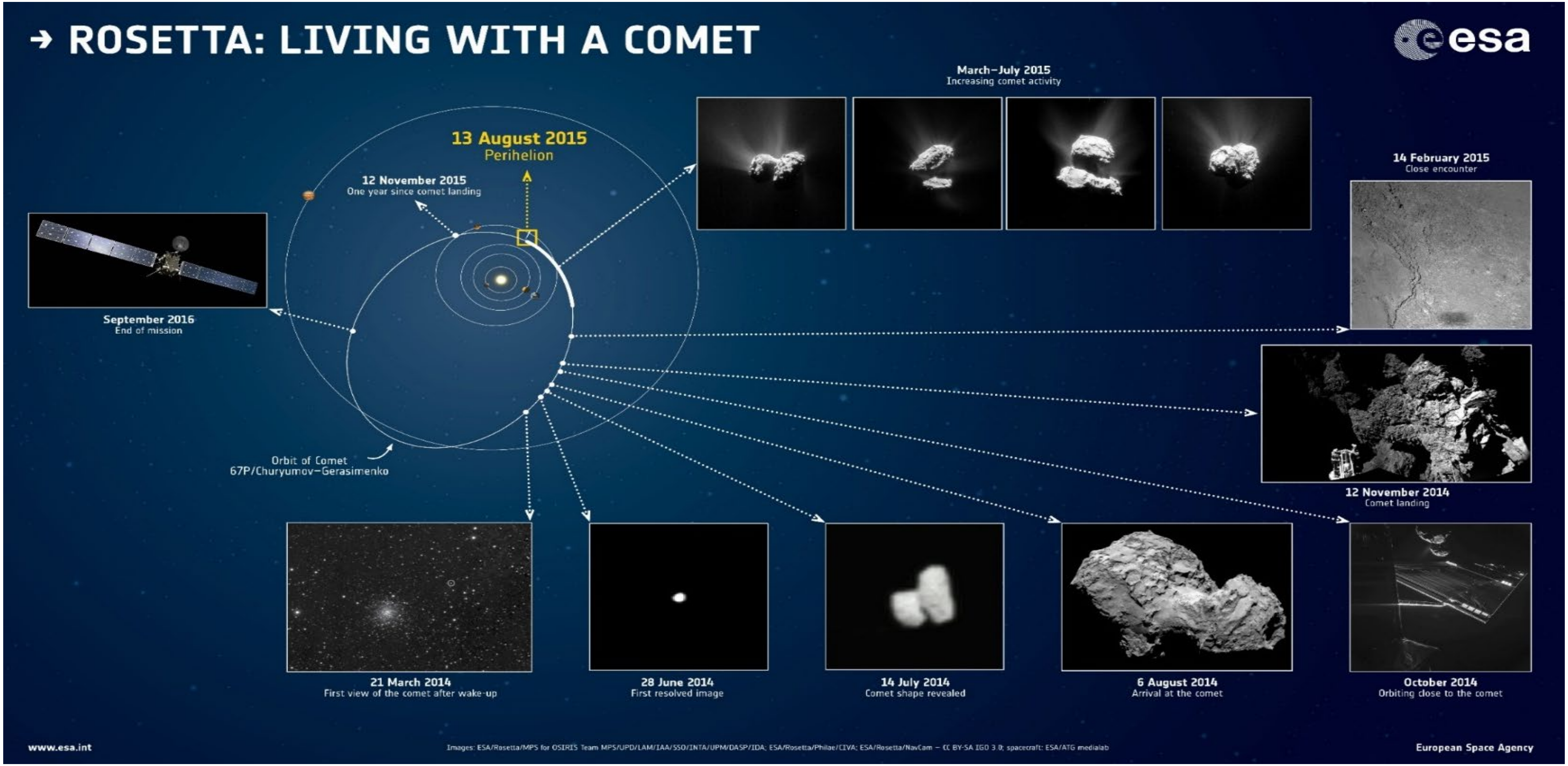
CAS-ADS: Module 2

Comparative statistical analysis of cometary outgassing

(based on data from ESA's Rosetta mission)

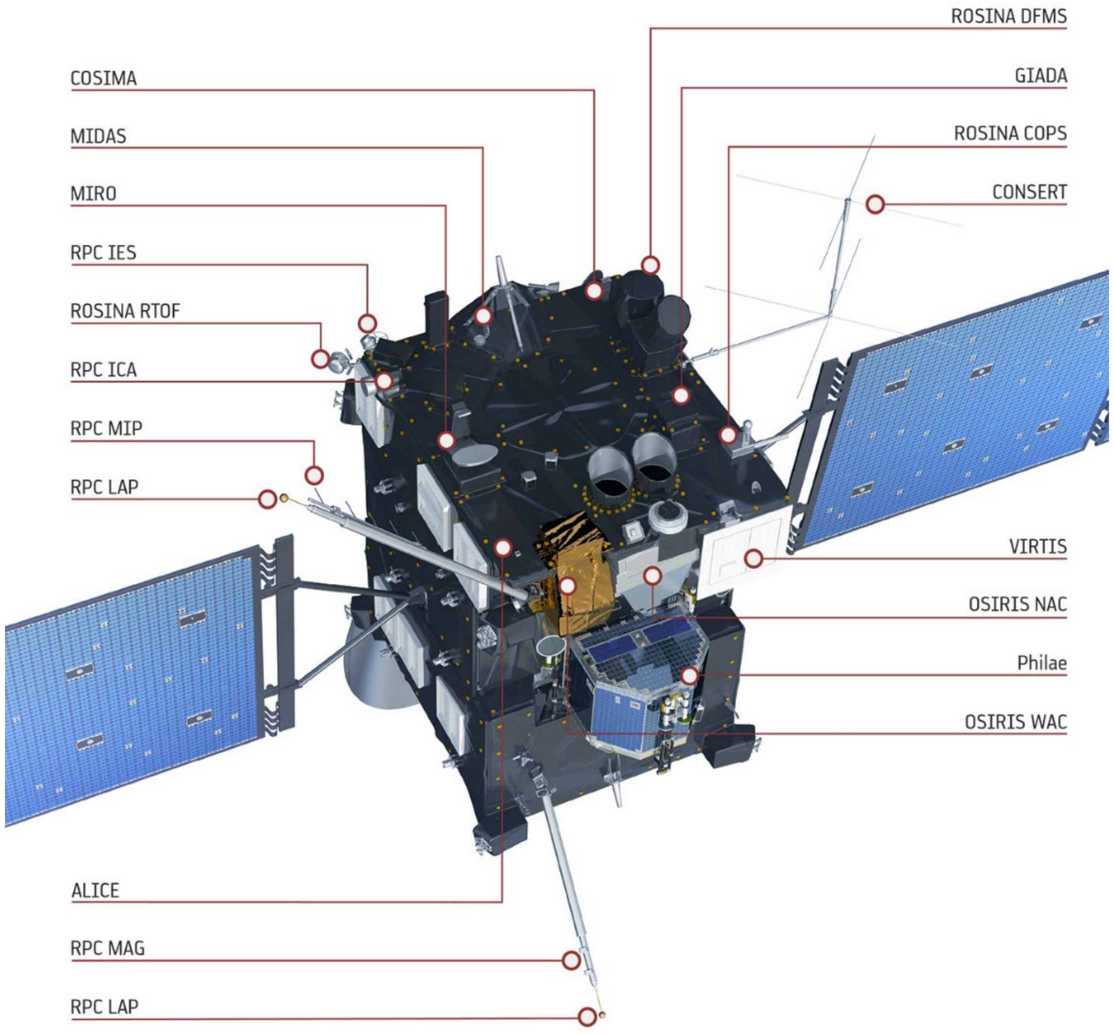
Janine Kocher & Nora Hänni

Rosetta: mission timeline and orbit



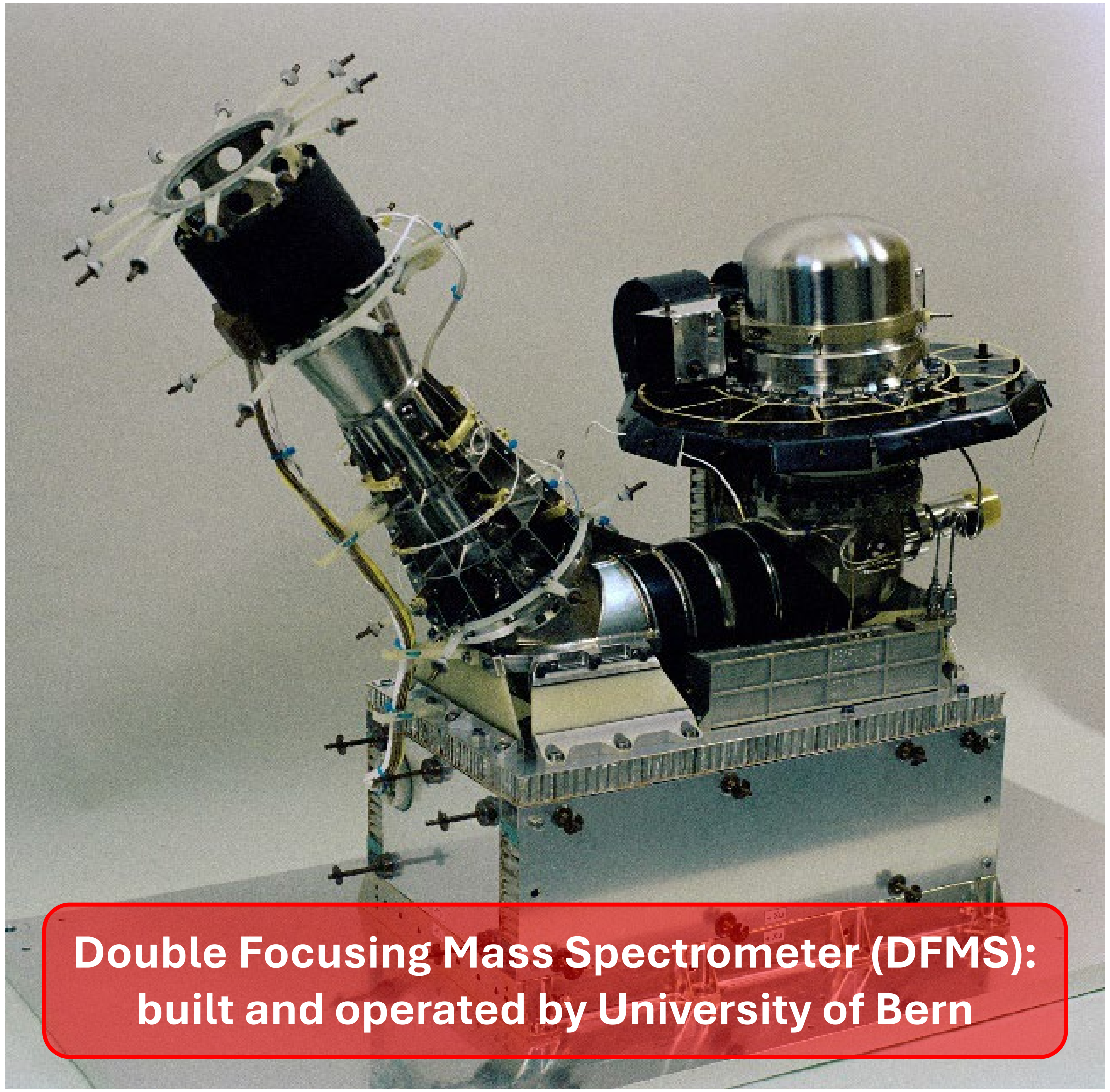
Rosetta: instrumentation

• OSIRIS	Camera	28 kg
• **ROSINA/DFMS**	**Gas mass spectrometer**	**35 kg**
• COSIMA	Dust mass spectrometer	20 kg
• GIADA	Dust flux analyzer	4.5 kg
• MIDAS	Dust microscope	5.5 kg
• VIRTIS	Infrared spectrometer	23 kg
• MIRO	Microwave experiment	16.2 kg
• ALICE	Ultraviolet spectrometer	2.2 kg
• RPC	Plasma instruments	5.7 kg
• RSI	Radio experiment	0.0 kg
• CONSERT	Comet nucleus sounder	2.0 kg
• Lander	10 Experiments	26.7 kg

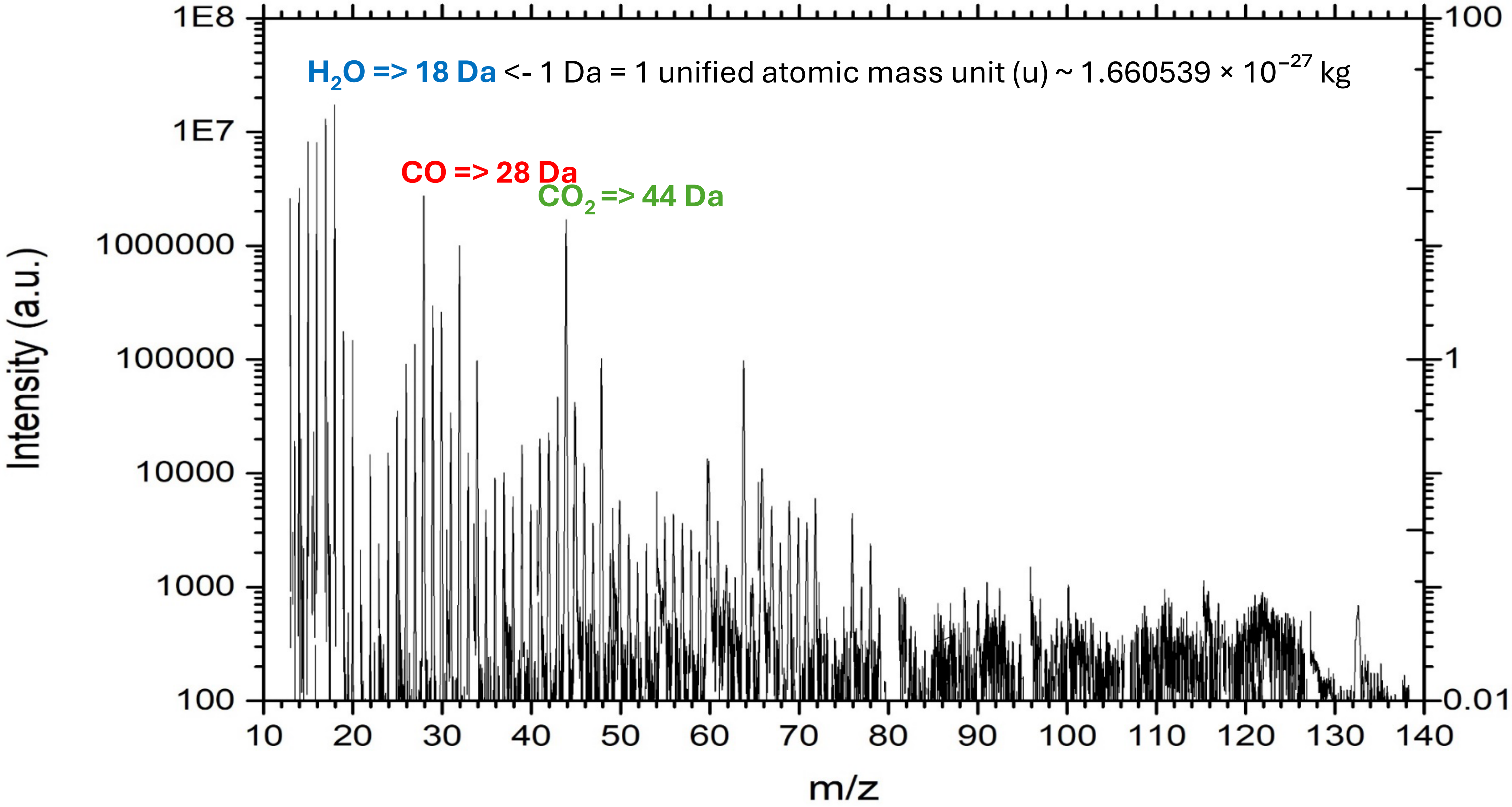


Rosetta: instrumentation

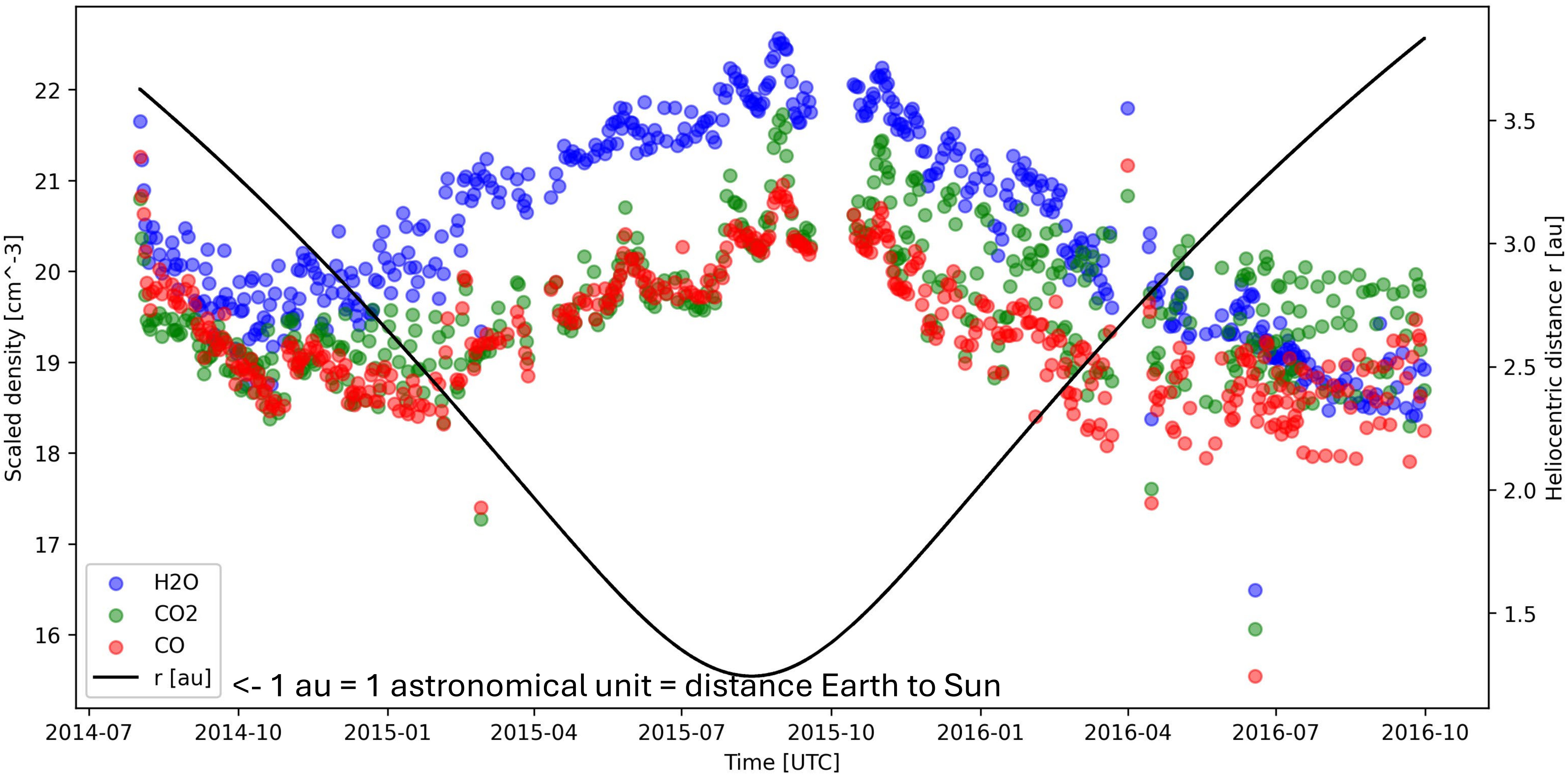
• OSIRIS	Camera	28 kg
• **ROSINA/DFMS**	**Gas mass spectrometer**	**35 kg**
• COSIMA	Dust mass spectrometer	20 kg
• GIADA	Dust flux analyzer	4.5 kg
• MIDAS	Dust microscope	5.5 kg
• VIRTIS	Infrared spectrometer	23 kg
• MIRO	Microwave experiment	16.2 kg
• ALICE	Ultraviolet spectrometer	2.2 kg
• RPC	Plasma instruments	5.7 kg
• RSI	Radio experiment	0.0 kg
• CONSERT	Comet nucleus sounder	2.0 kg
• Lander	10 Experiments	26.7 kg



DFMS raw data: counts as a function of mass-per-charge (m/z)



DFMS derived data: local densities (interpol. time series)



Investigated data set: time series of geometries and densities

Data file: CombineROSINA_rowAB_new.log (tab-separated -> can be directly handled as pandas dataframe)

```
AcquisitionTime AcquisitionTimeEt [s]   DistCGSC [km]   DistCGSUN [au]   NadirAngle [deg]    SunAngle [deg]   Lon [deg]    Lat [deg]    LocalTime [h]    PhaseAngle [deg]
2014-08-01T10:05:59 460159626.18    845.65  3.627   1.04    172.40  -78.65  37.65   12.42   8.65    44.86   52454   0.555   6.276136e+05    2.562130e+05    8.818728e
2014-08-01T10:06:59 460159686.18    845.44  3.627   1.03    172.38  -79.13  37.65   12.42   8.65    44.86   52454   0.555   7.811392e+05    3.188873e+05    1.097595e
2014-08-01T10:07:59 460159746.18    845.23  3.627   1.02    172.37  -79.61  37.64   12.42   8.65    44.86   52454   0.555   7.806682e+05    3.186951e+05    1.096933e
2014-08-01T10:08:59 460159806.18    845.02  3.627   1.01    172.35  -80.09  37.64   12.42   8.66    44.86   52454   0.556   7.815821e+05    3.190682e+05    1.098217e
2014-08-01T10:09:59 460159866.18    844.81  3.627   0.99    172.33  -80.57  37.64   12.42   8.66    44.86   52454   0.556   7.765560e+05    3.215280e+05    1.106684e
2014-08-01T10:10:59 460159926.18    844.60  3.627   0.98    172.31  -81.05  37.63   12.42   8.66    44.86   52456   0.556   7.674921e+05    3.277723e+05    1.128176e
2014-08-01T10:11:59 460159986.18    844.39  3.627   0.96    172.29  -81.54  37.63   12.42   8.67    44.86   52456   0.556   7.567504e+05    3.336825e+05    1.148519e
2014-08-01T10:12:59 460160046.18    844.18  3.627   0.94    172.27  -82.02  37.63   12.42   8.67    44.86   52456   0.556   7.468902e+05    3.390013e+05    1.166826e
2014-08-01T10:13:59 460160106.18    843.97  3.627   0.93    172.25  -82.50  37.62   12.42   8.68    44.86   52456   0.556   7.460450e+05    3.395683e+05    1.168777e
2014-08-01T10:14:59 460160166.18    843.76  3.627   0.91    172.23  -82.98  37.62   12.42   8.68    44.86   52456   0.556   7.460865e+05    3.405431e+05    1.172133e
2014-08-01T10:15:59 460160226.18    843.55  3.627   0.89    172.20  -83.46  37.62   12.43   8.68    44.86   52456   0.556   7.461097e+05    3.415151e+05    1.175478e
2014-08-01T10:16:59 460160286.18    843.34  3.627   0.86    172.18  -83.94  37.61   12.43   8.69    44.86   52453   0.556   6.362924e+05    2.920732e+05    1.005302e
2014-08-01T10:17:59 460160346.18    843.13  3.627   0.84    172.15  -84.43  37.61   12.43   8.69    44.86   52453   0.556   7.448587e+05    3.428784e+05    1.180171e
2014-08-01T10:18:59 460160406.18    842.92  3.627   0.82    172.12  -84.91  37.61   12.43   8.69    44.86   52453   0.556   7.435744e+05    3.432617e+05    1.181490e
```

print(df.describe())

	AcquisitionTimeEt [s]	DistCGSC [km]	DistCGSUN [au]	...	nHC3N [cm^-3]	nNG_N2 [cm^-3]	nRG_N2 [cm^-3]
count	4.540320e+05	454032.000000	454032.000000	...	454032.000000	4.540320e+05	4.540320e+05
mean	4.935478e+08	102.025748	2.471913	...	1281.183874	3.543576e+07	8.036930e+07
std	2.017161e+07	116.135904	0.795915	...	12738.336695	6.425173e+07	1.799453e+08
min	4.601596e+08	3.900000	1.243000	...	0.000000	4.067578e+01	0.000000e+00
25%	4.747050e+08	22.860000	1.686000	...	9.215120	1.016184e+07	0.000000e+00
50%	4.937398e+08	48.540000	2.515000	...	30.734445	1.847062e+07	0.000000e+00
75%	5.103436e+08	143.150000	3.212000	...	91.770980	3.479048e+07	9.579008e+07
max	5.284972e+08	1265.680000	3.833000	...	646759.800000	1.463048e+09	1.645793e+10

Unique key | Geometry stuff | Densities (26 species) | Total density for normalization

Data pre-treatment

	AcquisitionTimeEt [s]	DistCGSC [km]	DistCGSUN [au]	...	nHC3N [cm^-3]	nNG_N2 [cm^-3]	nRG_N2 [cm^-3]
count	4.540320e+05	454032.000000	454032.000000	...	454032.000000	4.540320e+05	4.540320e+05
mean	4.935478e+08	102.025748	2.471913	...	1281.183874	3.543576e+07	8.036930e+07
std	2.017161e+07	116.135904	0.795915	...	12738.336695	6.425173e+07	1.799453e+08
min	4.601596e+08	3.900000	1.243000	...	0.000000	4.067578e+01	0.000000e+00
25%	4.747050e+08	22.860000	1.686000	...	9.215120	1.016184e+07	0.000000e+00
50%	4.937398e+08	48.540000	2.515000	...	30.734445	1.847062e+07	0.000000e+00
75%	5.103436e+08	143.150000	3.212000	...	91.770980	3.479048e+07	9.579008e+07
max	5.284972e+08	1265.680000	3.833000	...	646759.800000	1.463048e+09	1.645793e+10

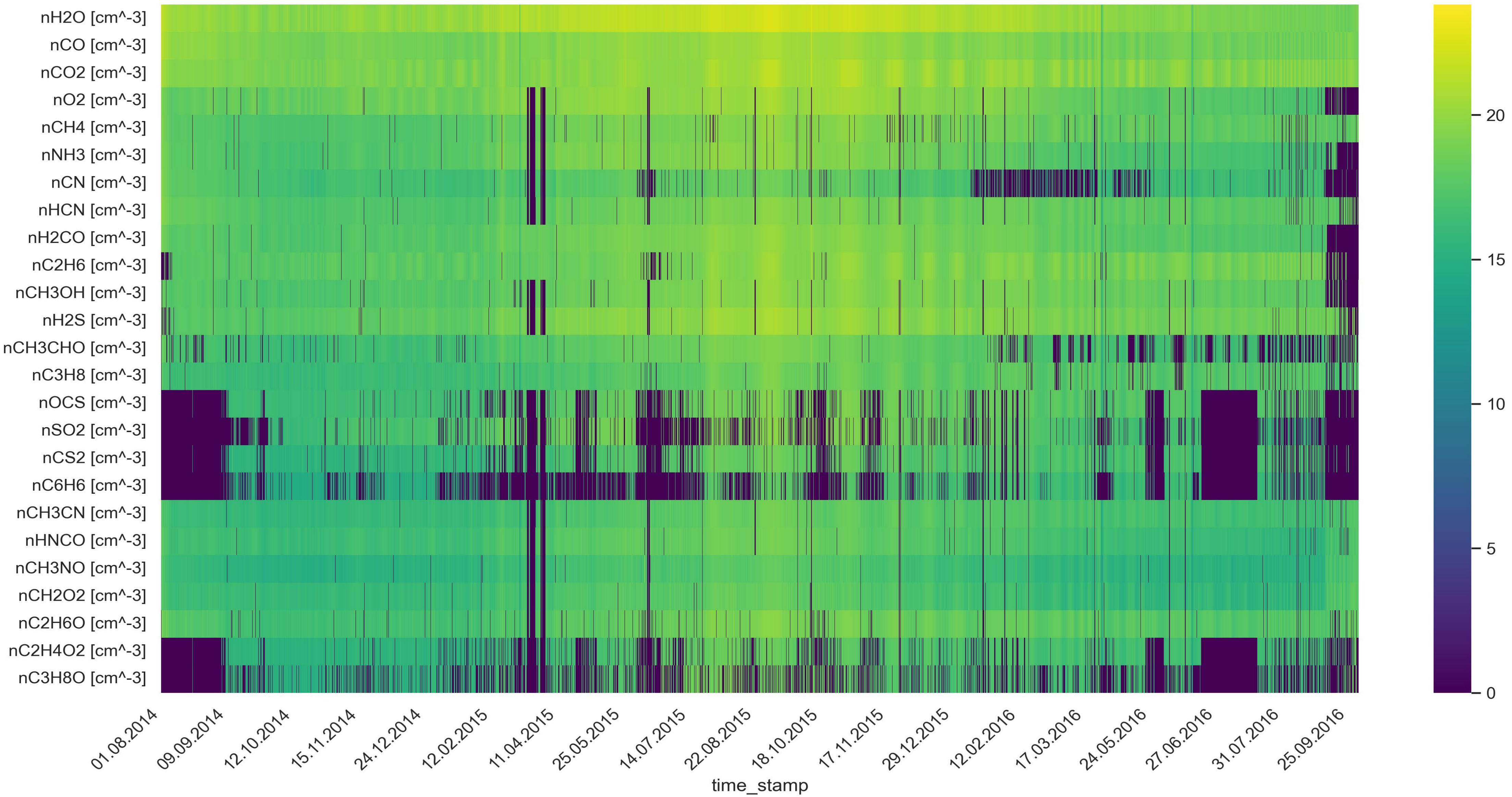
Dealing with zero densities: Zero density means the species was not fittable (either signal too small or spectrum faulty):

1) Replace zeroes with 1 so log(1)=0 again yields original value **(we chose this option for now)**
2) Exclude these values **(probably better option for correlation investigation)**

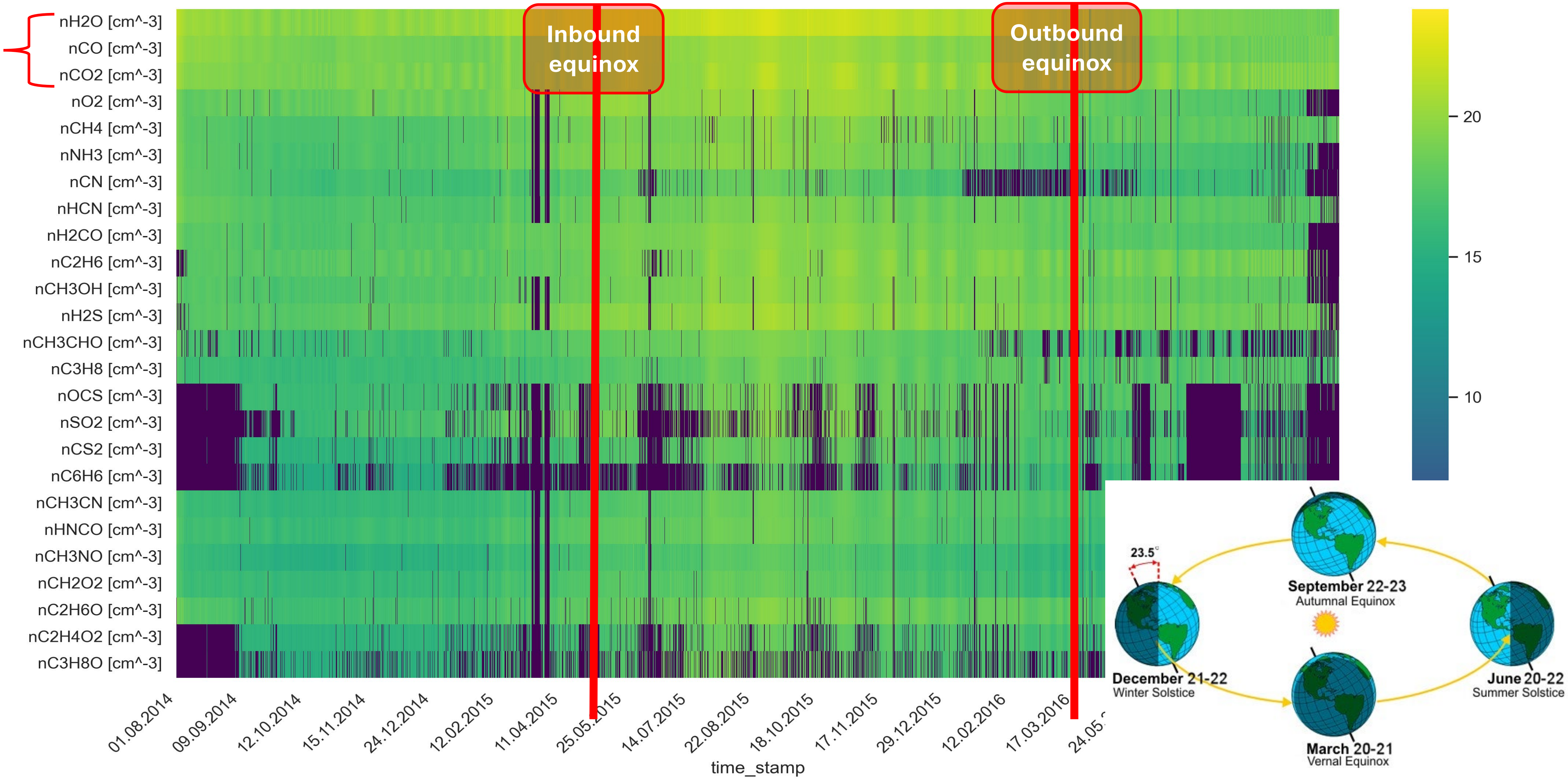
Log-transformation: Gives more weight to the less abundant species. If no log-transformation, all correlations would be dominated by the major species (water, carbon dioxide, carbon monoxide).

Density correction: As the spacecraft changes distance to the comet, the local gas density is affected. This physical effect can be corrected for by multiplying the densities with (DistCGSC)^2.

Scaled densities visualized as heatmap



Data subsets for comparison: Δt=2m around in-/outbound equinox



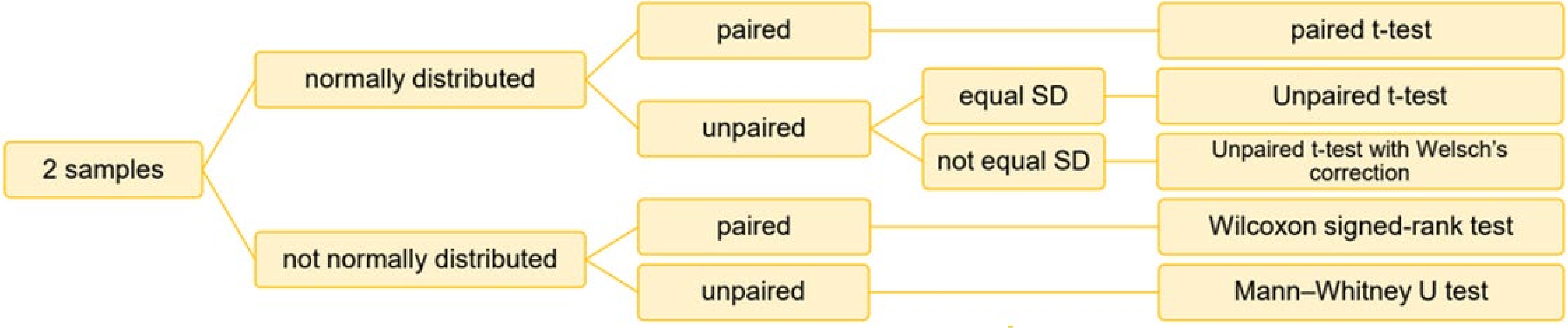
[Source: https://www.weather.gov/cle/seasons]

Hypotheses test

<u>Question:</u>
Does the data from the main species from the inbound equinox behave like the data from the outbound equinox or are those significantly different?

We have 2 samples we want to compare. To find the matching statistical hypothesis test one must answer the following questions:

- Are the samples normally distributed?
- Are our samples paired or unpaired?



Which assumptions does our data fullfill?

Is our data normally distributed?

Use Q-Q-plot to check if our data is normally distributed:

- The Q-Q-plot compares the quantiles of the dataset against the quantiles of the chosen distribution (-> nomal distribution)

- If the points in the plot form a straight line the dataset can be considered normally distributed
  ➡ not fulfilled for the selected data/samples

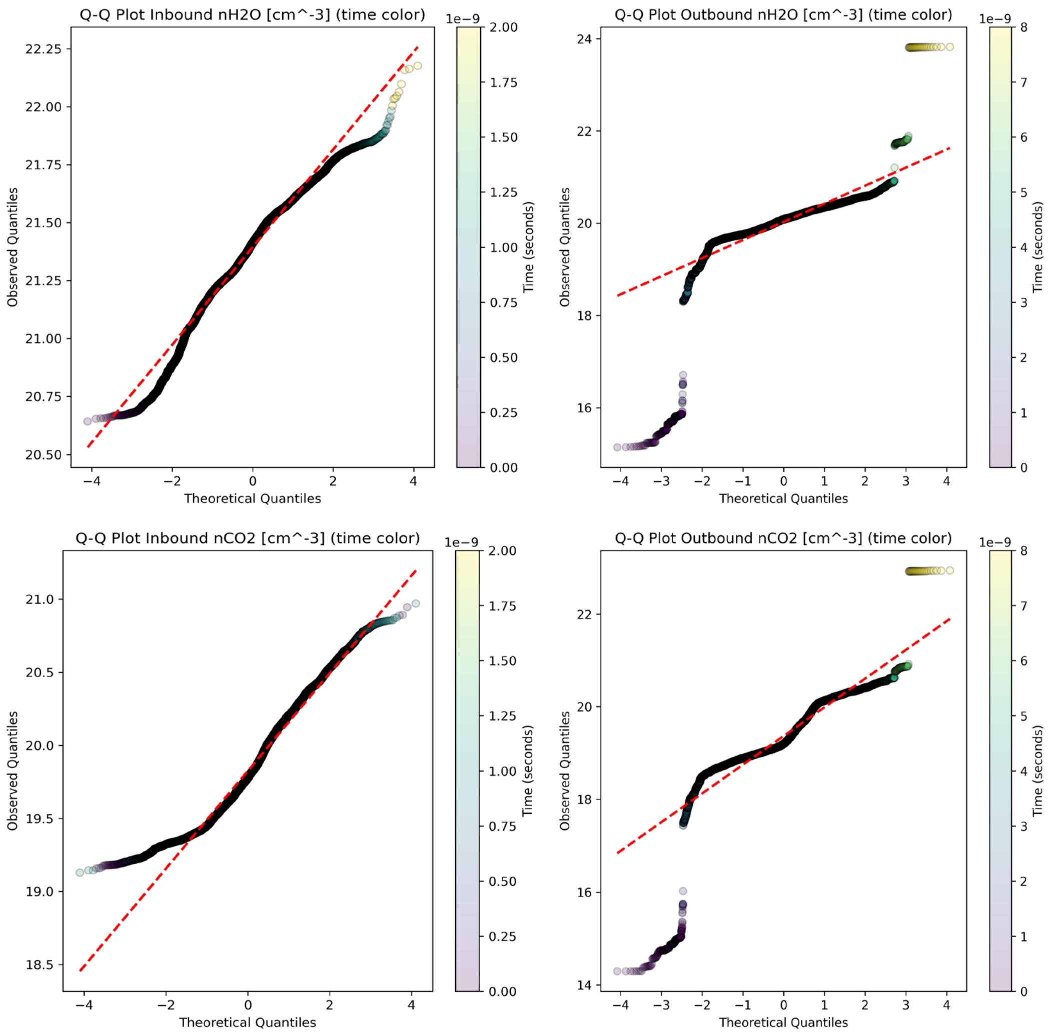
Is our data paired or unpaired?

Our data is **unpaired** since the measurements are independent - we do not have the same sample size in both intervals (same time interval but not same sample size)

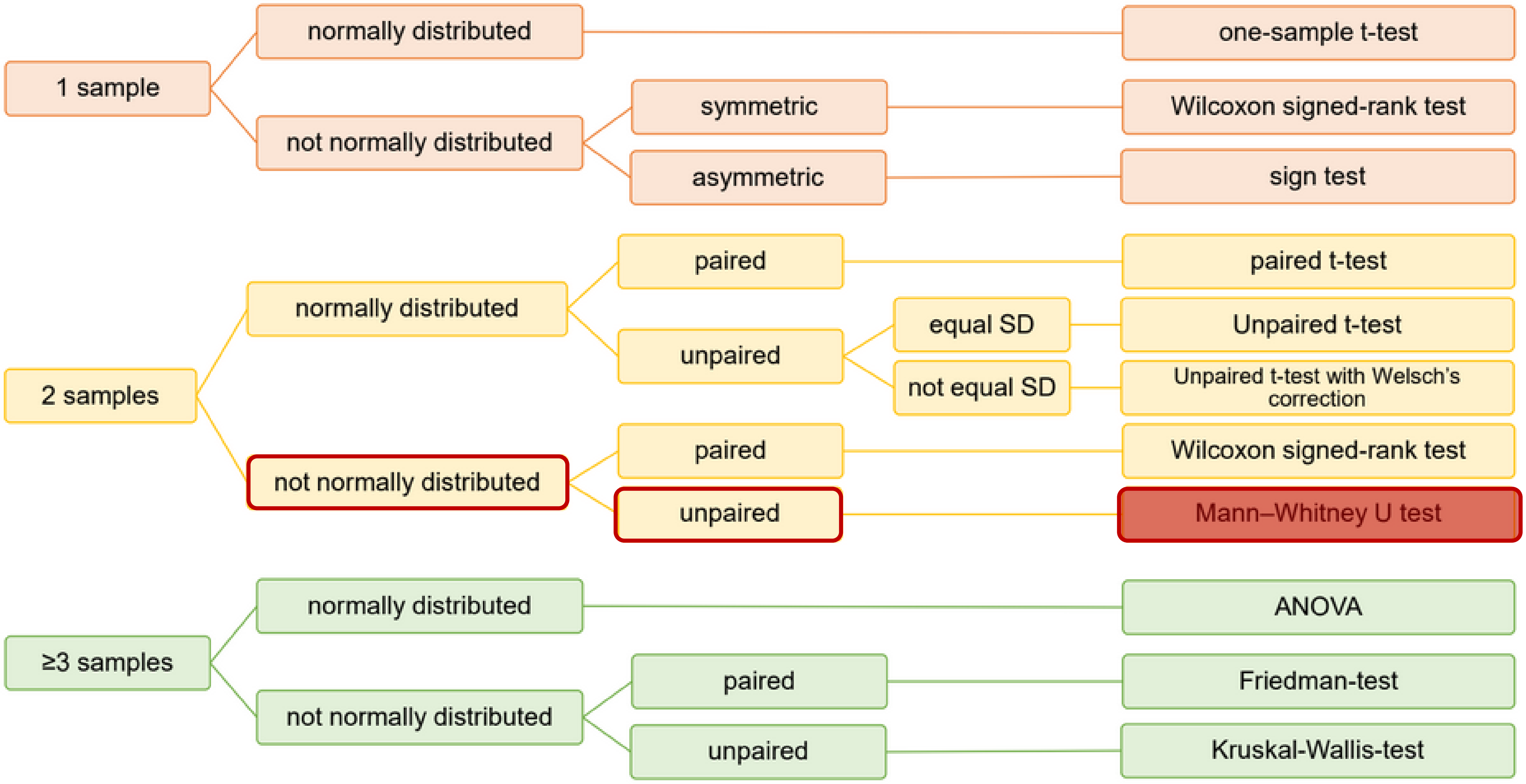
Inbound Outbound

H2O

CO2



Tests – which test is appropriate?



Hypotheses and result

We consider a Mann-Whitney U test for our hypothesis:

Samples:

- Densities from the inbound equinox (time interval: 2015-05-10 +/- 30 days)
- Densities from the outbound equinox (time interval: 2016-03-21 +/- 30 days)

Goal: Test whether the densities of the 3 main species differ significantly?
With significance level: 5%

For every main species we define the following H0 and H1 hypothesis:

H0:
The two samples inbound and outbound densities for the 3 main species come from the same distribution i.e. are not significantly different.

H1:
There is a significant difference between the two samples.

Hypotheses test result

How to interpret the results:

- The U-value is the main measure in the Mann-Whitney U Test. It compares the ranking of the values from both groups compared to each other. It is the base value to calculate the corresponding p-value.
- If the resulting p-value is smaller than the set significance level (here 0.05) one can reject the H0 hypotheses

```
Mann-Whitney U test (Inbound vs Outbound):
nH2O [cm^-3]: U-statistic=1035089854.00, p-value=0.000e+00
nCO [cm^-3]: U-statistic=1021465898.00, p-value=0.000e+00
nCO2 [cm^-3]: U-statistic=766707469.00, p-value=0.000e+00
```

Result:
We obtain very small p-values and very large test statistics (ranked sum).
Therefore H0 can be rejected.

Meaning for the comet setting:
The densities of that species inbound and outbound are statistically distinguishable, which could reflect seasonal, heliocentric, or cometary activity changes. The comet has changed during the perihelion passage.

Outlook

Having a look at the Pearson correlation matrix one can see that certain species seem to correlate a lot and others are nearly correlated.

So open questions are:

- Does the Pearson matrix change with time and if yes over what time periods?
- Can we explain clusters in the Pearson matrix as a linear combination of the others?
- Should we apply filters to the data, e.g., filter out angles/instrument off-pointing etc.?
- How should we treat zero values (these are either from not fittable or missing spectra or true zeros)?