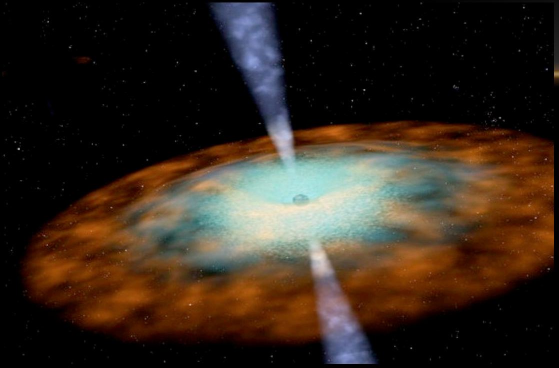
A PRINCIPAL COMPONENT ANALYSIS OF THE PROPERTIES OF GALAXIES WITH MASER ACTIVITY



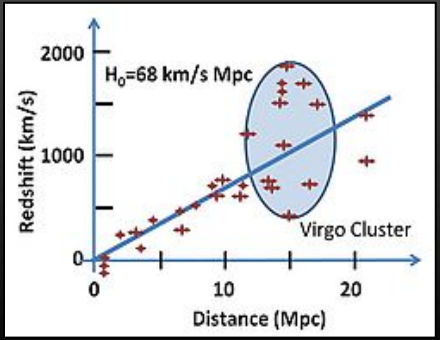
Nathan DiDomenico

Anca Constantin

With thanks to:

Jimmy Corcoran, Emil Christensen, and Jamil Guevara

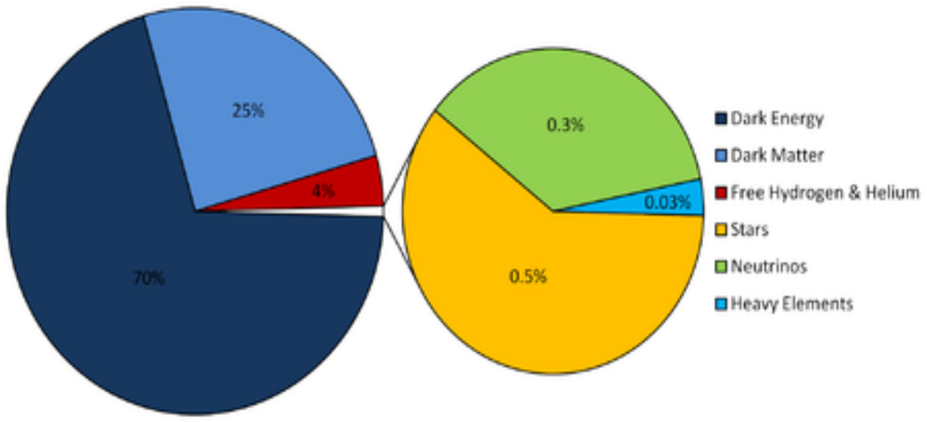
Hubble's Law and Supermassive Black Holes



The rate of a object's recession from the Earth is proportional to their distance from the Earth:

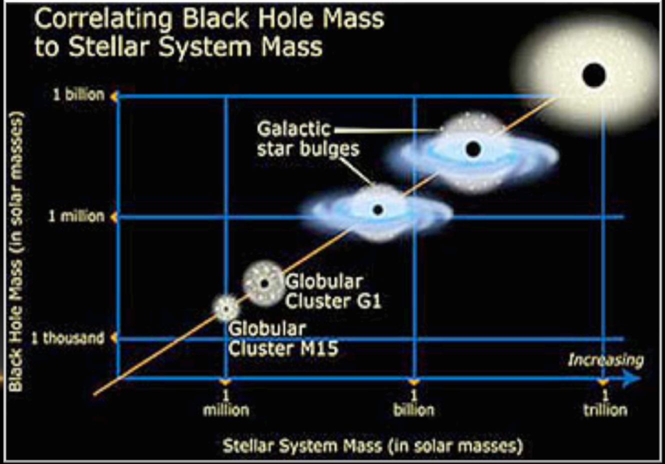
$$v=H_0D$$

- A Very accurate value for H_0 is necessary for constraining cosmological models
- In order to measure H_0 we need accurate values for a object's distance and velocity.

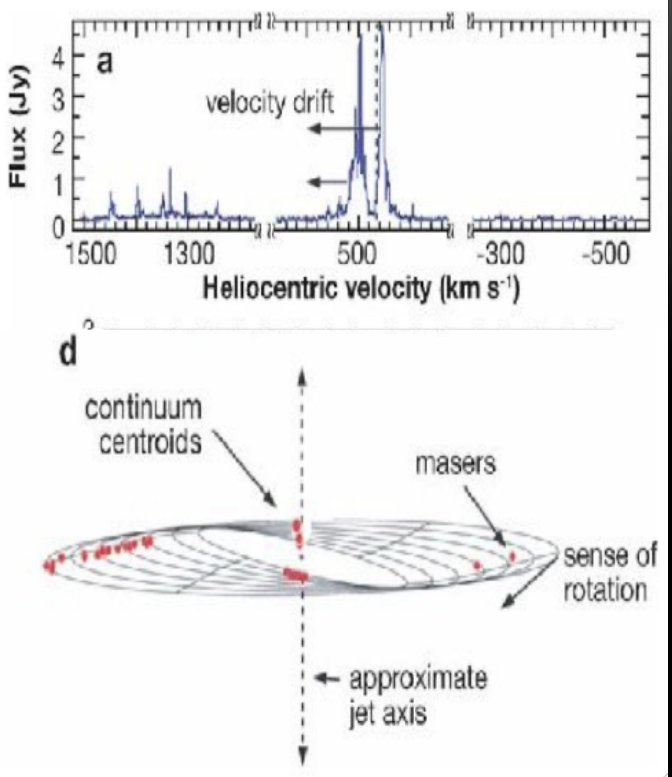


A black hole that is on the order of 10^6 to 10^9 M_{sun}. It is thought that a supermassive black hole exists in most, and possibly every galaxy's center.

How did they form? How do they affect galaxy formation?



Mega-masers are A-mase-ing!



Using Megamasers Disks to Find Distances and BHMs

NGC 4258: the prototype for this method.

1. Measure v_r from Doppler shifts
2. Multiple observations to measure a_c (v^2/r)
3. $dv/d\theta = D(v/r)$ Keplerian thin disk model Solve for D!
4. Solve for D to constrain cosmological models (H_0)! or find black hole masses.

One Such Calculation is not Enough

3% accuracy for H_0 requires:

100 galaxies like NGC 4258

Or 10 more distant (> 50 Mpc), more luminous systems.

The detection rate of maser systems remains low

- Surveys for maser systems have had a <3% success rate for Galaxies with any maser emission, even fewer maser disks.

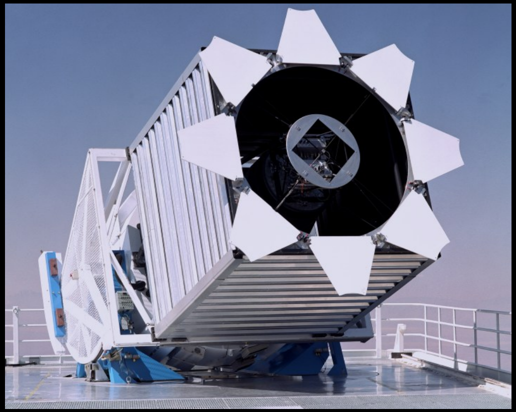
OUR PROJECT: PROPERTIES OF GALAXIES WITH H_2O MASERS

What produces megamaser emission in some galaxies and not in others?

First we need data, and for the first time ever there is plenty of it!

Megamaser Cosmology Project (MCP):
- Uses the GBT, VLBA, VLA, and others to find H2O megamasers
- ~3000 galaxies surveyed, 146 detected in H2O maser emission

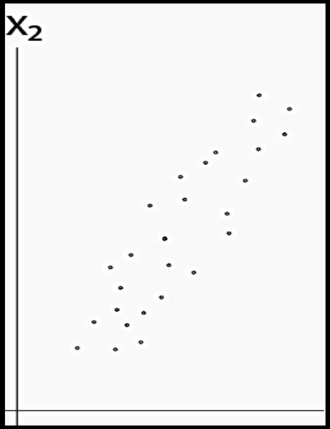
Sloan Digital Sky Survey (SDSS):
- A survey of a quarter of the sky
- Final dataset includes 230 million celestial objects, including spectra of about 10^6 galaxies

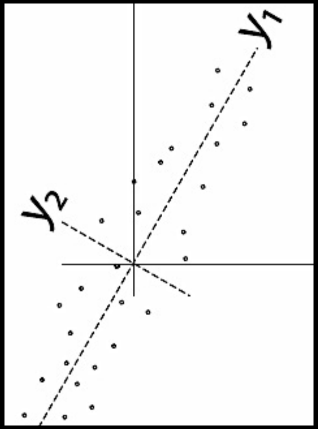


Our data:
- A cross match between these two data sets revealed ~1200 galaxies including 44 maser detections.
- We employ redshifts, line fluxes, luminosities, stellar population properties, proxies for black hole masses, etc.

Principle Component Analysis (PCA):

What is the data telling us?





PCA determines which dynamics are important, which are redundant and which are noise by finding the Principle Components, the directions of the most variance in the data.

Let **X** be the original data set, where each *column* is a single object of our data set

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_3 & z_3 \end{pmatrix}$$

1) Subtract the mean

2) Find **C**, the covariance matrix

- covariance is a measure of how much two random variables change together
- variance is the variation of the values of a variable
- We may express the these as a dot products
  - $\sigma^2_{ab} \equiv (1/n)\,\mathbf{ab}^T$
- the covariance matrix $C_X \equiv (1/n)XX^T$.

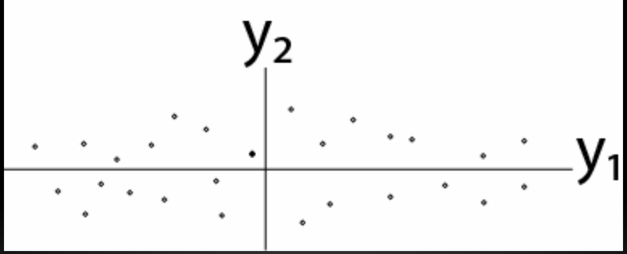
- the diagonal terms (C_{ii}) show the *variance* of the i^{th} data type.
- The off-diagonal terms (C_{ij}) are the *covariance* between the data types.

In English:

- In the diagonal terms, large values correspond to interesting structure.
- In the off-diagonal terms large magnitudes correspond to high redundancy.

So the goal becomes: Find some orthonormal matrix P as in Y = PX such that C is a diagonal matrix (i.e. find a matrix that will change the basis to one where the covalence matrix is diagonalied)

C is a square symmetric *m* × *m* matrix and can therefore be diagonalied by a matrix of its eigenvectors.



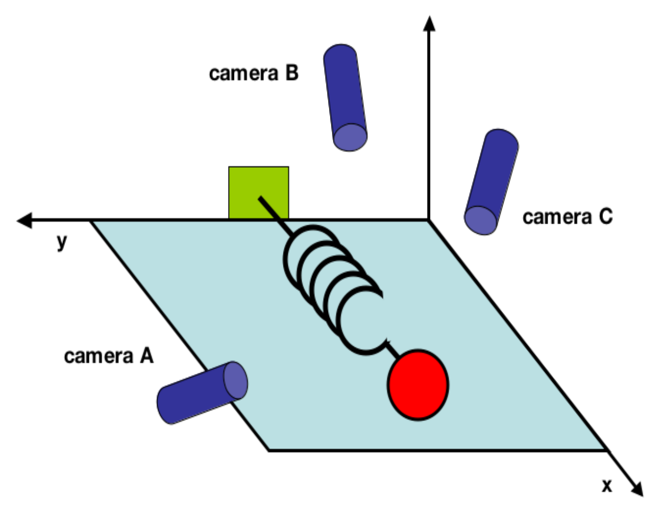
A toy example:

Pretend we are studying the motion of the physicist's ideal spring. Of course the underlying dynamics can be expressed as a function of a single variable x. (It moves in a line, it is one dimensional).

But we don't know:

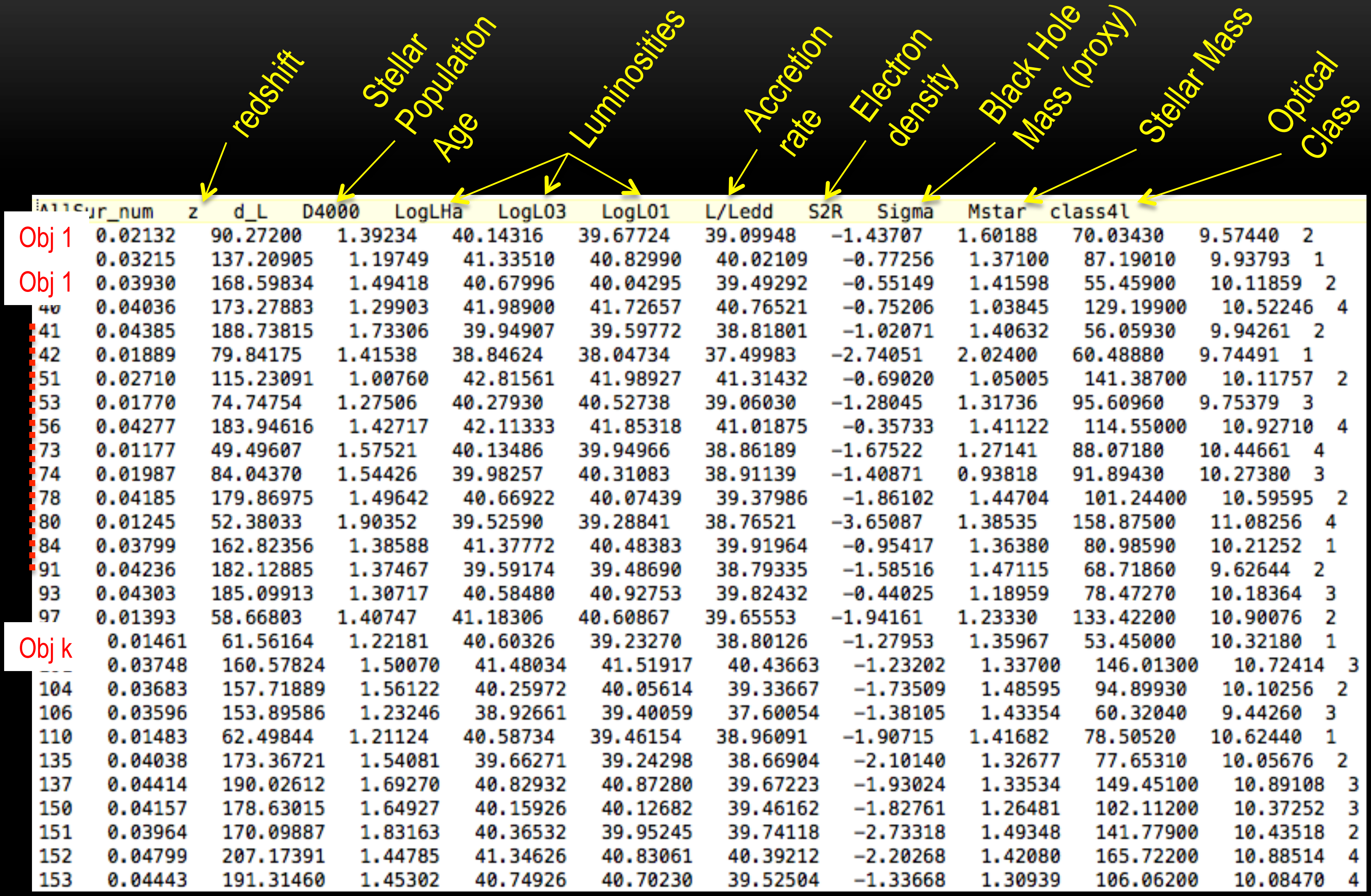
- which or how many, axes and dimensions are important to measure. We measure the ball's position in a three-dimensional space.
- We do not even know what are the real "**x** ", "**y** " and "**z** "axes, so we choose some arbitrary angles with respect to the system.

How do we get from this data set to a simple equation of **x**



PCA analysis would tell us that the process is purely linear and (in terms of where the cameras are currently located) what axis that line is.

A Mere Sample of our Glorious Data



AllSur_num	z	d_L	D4000	LogLHa	LogLO3	LogLO1	L/Ledd	S2R	Sigma	Mstar	class4l
	0.02132	90.27200	1.39234	40.14316	39.67724	39.09948	-1.43707	1.60188	70.03430	9.57440	2
	0.03215	137.20905	1.19749	41.33510	40.82990	40.02109	-0.77256	1.37100	87.19010	9.93793	1
	0.03930	168.59834	1.49418	40.67996	40.04295	39.49292	-0.55149	1.41598	55.45900	10.11859	2
40	0.04036	173.27883	1.29903	41.98900	41.72657	40.76521	-0.75206	1.03845	129.19900	10.52246	4
41	0.04385	188.73815	1.73306	39.94907	39.59772	38.81801	-1.02071	1.40632	56.05930	9.94261	2
42	0.01889	79.84175	1.41538	38.84624	38.04734	37.49983	-2.74051	2.02400	60.48880	9.74491	1
51	0.02710	115.23091	1.00760	42.81561	41.98927	41.31432	-0.69020	1.05005	141.38700	10.11757	2
53	0.01770	74.74754	1.27506	40.27930	40.52738	39.06030	-1.28045	1.31736	95.60960	9.75379	3
56	0.04277	183.94616	1.42717	42.11333	41.85318	41.01875	-0.35733	1.41122	114.55000	10.92710	4
73	0.01177	49.49607	1.57521	40.13486	39.94966	38.86189	-1.67522	1.27141	88.07180	10.44661	4
74	0.01987	84.04370	1.54426	39.98257	40.31083	38.91139	-1.40871	0.93818	91.89430	10.27380	3
78	0.04185	179.86975	1.49642	40.66922	40.07439	39.37986	-1.86102	1.44704	101.24400	10.59595	2
80	0.01245	52.38033	1.90352	39.52590	39.28841	38.76521	-3.65087	1.38535	158.87500	11.08256	4
84	0.03799	162.82356	1.38588	41.37772	40.48383	39.91964	-0.95417	1.36380	80.98590	10.21252	1
91	0.04236	182.12885	1.37467	39.59174	39.48690	38.79335	-1.58516	1.47115	68.71860	9.62644	2
93	0.04303	185.09913	1.30717	40.58480	40.92753	39.82432	-0.44025	1.18959	78.47270	10.18364	3
97	0.01393	58.66803	1.40747	41.18306	40.60867	39.65553	-1.94161	1.23330	133.42200	10.90076	2
	0.01461	61.56164	1.22181	40.60326	39.23270	38.80126	-1.27953	1.35967	53.45000	10.32180	1
	0.03748	160.57824	1.50070	41.48034	41.51917	40.43663	-1.23202	1.33700	146.01300	10.72414	3
104	0.03683	157.71889	1.56122	40.25972	40.05614	39.33667	-1.73509	1.48595	94.89930	10.10256	2
106	0.03596	153.89586	1.23246	38.92661	39.40059	37.60054	-1.38105	1.43354	60.32040	9.44260	3
110	0.01483	62.49844	1.21124	40.58734	39.46154	38.96091	-1.90715	1.41682	78.50520	10.62440	1
135	0.04038	173.36721	1.54081	39.66271	39.24298	38.66904	-2.10140	1.32677	77.65310	10.05676	2
137	0.04414	190.02612	1.69270	40.82932	40.87280	39.67223	-1.93024	1.33534	149.45100	10.89108	3
150	0.04157	178.63015	1.64927	40.15926	40.12682	39.46162	-1.82761	1.26481	102.11200	10.37252	3
151	0.03964	170.09887	1.83163	40.36532	39.95245	39.74118	-2.73318	1.49348	141.77900	10.43518	2
152	0.04799	207.17391	1.44785	41.34626	40.83061	40.39212	-2.20268	1.42080	165.72200	10.88514	4
153	0.04443	191.31460	1.45302	40.74926	40.70230	39.52504	-1.33668	1.30939	106.06200	10.08470	4

The Results of the PCA analysis

Sample of Maser Hosts

data	70%	89%	96%	100%
z	0.4499	0.066	0.1012	0.0562
D4000	0.4404	0.1525	0.1031	0.1752
LogLO3	-0.4403	-0.022	0.0962	-0.4482
L/Ledd	0.4087	-0.0571	0.1827	-0.8422
S2R	0.4487	0.0648	0.0831	0.1362
Sigma	-0.2029	0.6141	0.7426	0.0679
Mstar	0.016	0.7664	-0.6149	-0.1813

Amount of variance accounted for

PC 1

PC 2

…

Sample of Galaxies Without Masers

data	43%	68%	83%	92%
z	-0.5326	-0.1015	-0.1402	0.2337
D4000	-0.4804	0.2625	-0.1399	0.2166
LogLO3	0.49	-0.0008	-0.2943	0.4598
L/Ledd	-0.0531	-0.6279	-0.4251	0.4217
S2R	-0.4919	-0.04	0.0181	-0.0387
Sigma	0.024	0.6794	-0.0512	0.5198
Mstar	0.0121	0.2517	-0.831	-0.4874

PC 1

PC 2

…

THE HOMEWORK

- We have the eigenvectors now we need to find what they mean physically.
- How do correlations suggested by the PCA results differ across the samples? What correlations are important and what can we learn from them?
- Run the PCA for different subsets of measurements
- We will combine the two samples but include a “strength of H2O emission” (L_{H20})
- Add measurements in other wavelengths e.g. WISE (Wide Field Infrared Survey Explorer) data (from Jimmy Corcoran)
    - Crucial for mapping the link between dust properties and the masing activity.