



Actua's AI Activities Series

Activity 2

Decision Trees: Classifying Space Objects

Decision Trees: Classifying Space Objects

If you're accessing this activity directly, did you know there are eight other activities in this series up on our website? If you find yourself unfamiliar with any of the AI concepts and terminology introduced in these activities, please refer to our [AI Glossary](#). These activities also follow a space exploration narrative when done in order. It is recommended to complete the activities in order but they can also be done on their own.

You and your group mates are astronauts and scientists aboard the Actua Orbital Station. Unfortunately, your station just got bombarded by magnetic rays and your electronics have begun to shut down! The only one who can save you is the station's AI, DANN. DANN stands for Dedicated Actua Neural Network, and it's gone a little loopy. Brush up on your technical skills, learn about AI, and save yourself and your crewmates!

Now that we've learned the basics of AI in "[Introduction: What is AI?](#)", we can begin to fix DANN, and we can start with our scanner! The Actua Orbital Station has a large scanner to monitor and track space objects, both near and far. It looks like after the damage from the magnetic rays, the space object classifier was reset. As one of your repairs aboard the station, Mission Control has asked that you propose and evaluate a decision tree so that you can bring the space object classifier back online. Once we do that, we can move on to studying an experiment in "[Regression Analysis: Making Predictions using Data](#)".

Activity Summary

In this activity, participants will create and evaluate decision trees. Decision trees are an approach to sorting objects or data into different types by asking questions about them. Participants will develop different questions for a decision tree by looking at an example dataset. They will then test how well their decision tree works by seeing if it can correctly label a testing dataset.

Developed by Actua, 2022.



Delivery Environment	Activity Duration	Intended Audience
Classroom	75 minutes	Grades 9-12 (Ages 13-18)

Achievement Goals

Learning Goals

Learning goals are statements referring to the understanding, knowledge, skills or application participants acquire during the activity. **Following this activity, participants will:**

- **Create and evaluate** decision trees
- **Classify and test** datasets
- **Determine** how best to use and sort data for different purposes
- **Understand** the role of decision trees in computer algorithms

Logistics (Timing, Group Size, Materials)

Section Title	Time	Group Size	Materials
Opening Hook: 20 Questions	10 minutes	<i>Small groups</i>	<p>Facilitators</p> <ul style="list-style-type: none"> • List of items to use for twenty questions (see Activity Procedure step #2 for this section) <p>Group Size</p> <ul style="list-style-type: none"> • Pen/pencil • Paper
Activity 1: Making decisions with	15 minutes	<i>Entire group</i>	



Section Title	Time	Group Size	Materials
trees			
Activity 2: Classifying Space Objects	30 minutes	<i>Individually or in small groups</i>	Group Size <ul style="list-style-type: none"> • Either Computer access or Printed training dataset (<i>Appendix C</i>) • Pen/pencil • Paper
Activity 3: Evaluating a Tree	10 minutes	<i>Individually or in small groups</i>	Group Size <ul style="list-style-type: none"> • Printed testing dataset (<i>Appendix C</i>) • Pen/pencil • Calculator
Reflection & Debrief	10 minutes	<i>Entire Group</i>	



Safety Considerations

Safety considerations have been provided below to support safety during this activity, however they are not necessarily comprehensive. It is important that you review the activity and your delivery environment to determine any additional safety considerations that you should be implementing for the delivery of these activities.

Online Safety

Some components of this activity require the use of devices connected to the internet.

- Facilitators should review the provided videos and read/explore provided websites and materials to determine if they are suitable for their participants.
- Where applicable, facilitators should remind participants to stay on task and only use links provided within this activity.

Activity Procedure

Opening Hook: 20 Questions

Twenty Questions is a game where you try to guess an object or animal by asking a series of questions. Each question that you ask is used to narrow down what the object or animal might be. With a partner or in small groups:

1. Designate one group member as the “answerer”. This person is responsible for answering the questions that are asked and can only say “yes” or “no”.
2. The answerer must secretly choose an item from the list below:
 - a. Airplane
 - b. Bear
 - c. Cell phone
 - d. Desk
 - e. Eagle
 - f. The Moon
 - g. Pigeon
 - h. School bus
 - i. Space shuttle
 - j. Truck
 - k. Turtle
 - l. Rock
3. Everyone else in the group must take turns asking questions to try to figure out what the chosen item is. Only questions that can be answered with “yes” or “no” can be asked.



4. Have someone record each question that is asked. You can ask up to 20 questions.
5. If you successfully guess the item, choose a new answerer and play again (go back to step 2). Make sure that everyone has a chance to be the answerer before continuing.
6. Once you're done playing, take a look at the questions that have been asked:
 - a. Do any of the questions apply to more than one list item?
 - b. How can you use the questions you have to divide the list items into two groups? What about three groups? Four groups?

Section 1: Making Decisions with Trees

Imagine you have a pile of different objects or things that have all been mixed together. Decision trees are a way to un-mix things (e.g. animals, objects) and divide them into groups. These groups are sometimes called labels or classes. You can do this by asking questions about features, also known as the details that can be used to identify objects. The goal of a decision tree is to make the different types of objects in your pile as separate as possible.

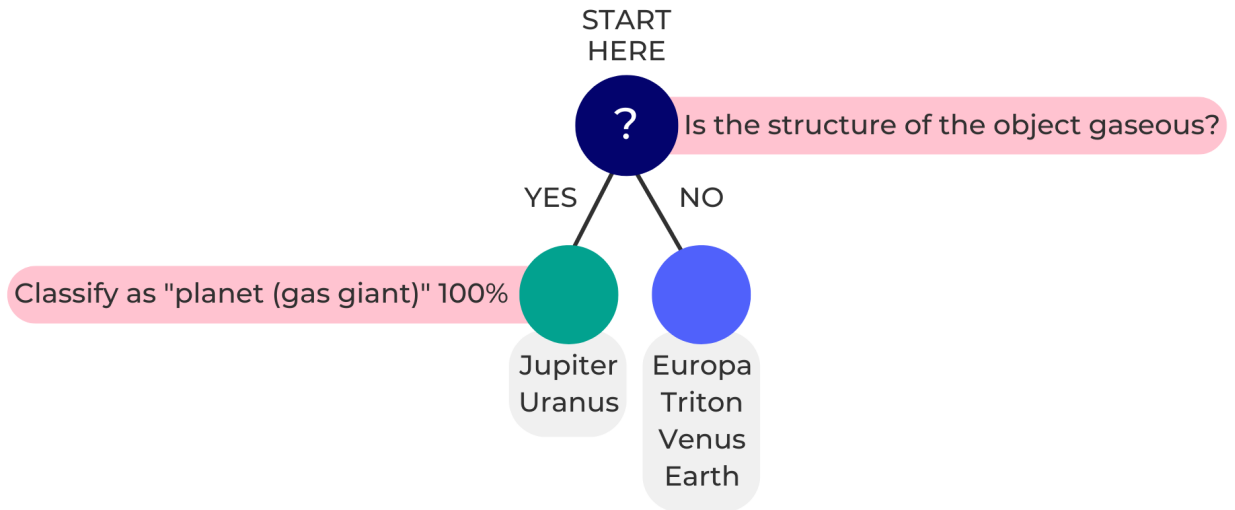
1. To learn about decision trees, let's work through an example together. We'll be trying to divide up the celestial objects below.

Features		Label	Name (extra)
Structure	Orbit		
Solid	Planet	Moon	Europa
Solid	Planet	Moon	Triton
Gaseous	Sun	Planet (gas giant)	Jupiter
Gaseous	Sun	Planet (gas giant)	Uranus
Solid	Sun	Planet (terrestrial)	Venus
Solid	Sun	Planet (terrestrial)	Earth



2. Look at the table above. How many different types of objects are there? (*Hint: look at the label column*)
 - a. There are three different types of objects: moons, and two types of planets (gas giants and solid/terrestrial).
3. We need to use the information in the table to come up with questions that identify each type of object. Look at the “structure” feature and take a moment to formulate the first question that you think we should ask to start sorting our data.
 - a. The question should be something like, “Is the structure of the object gaseous?”
4. We can also start drawing the beginning of our decision tree:
 - a. Draw the top of the tree. This is called the “root node”. Write the first question beside it.
 - b. Draw two lines coming out of the root node. These are called “branches” and there should be one for each possible answer to the node (“yes” and “no”). Label the branches.
 - c. Look at each item in the table above and sort each item using the first question (i.e. ask the question and put the item in either the “yes” group or the “no” based on the answer).
 - d. Once you’re done sorting the items based on the first question, look at the groups that have been made. Do any of the groups have only one type of object in them?

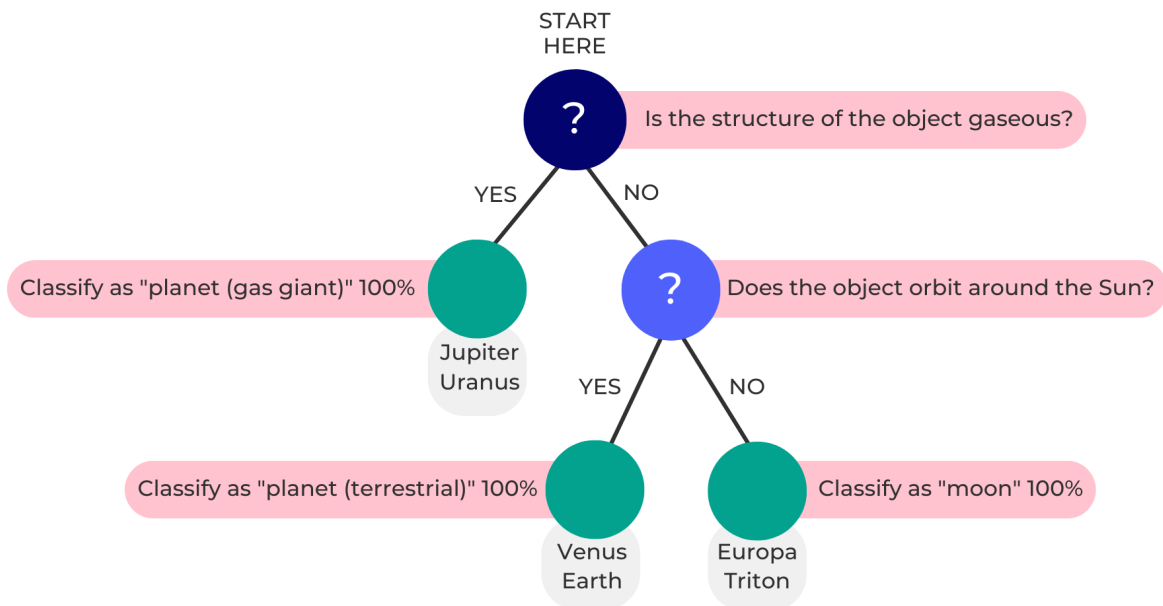




5. You should have a group that *only* contains gas giants. This is good, because it takes care of one of the types of objects that we started with.
6. You may notice that there's a bit of extra text included: "Classify as 'planet (gas giant)' 100%". What do you think this text means? Why do you think the 100% means?
 - a. This marks a decision to stop asking any more questions, since we think the items are as separate as possible. We instead choose to "classify" them, i.e. put a label on them.
 - b. This 100% represents how confident we are in our decision. Since there are only gas giants in this category, we can have full confidence in our decision.
7. The other group still has more than one type of object in it. We can try posing more questions to separate it further. *What do you think we should ask for the next question? (Hint: look at the table and think about which features we've already asked about and what remains to be asked)*
 - a. The question should be something like, "Does the object orbit around the Sun?"



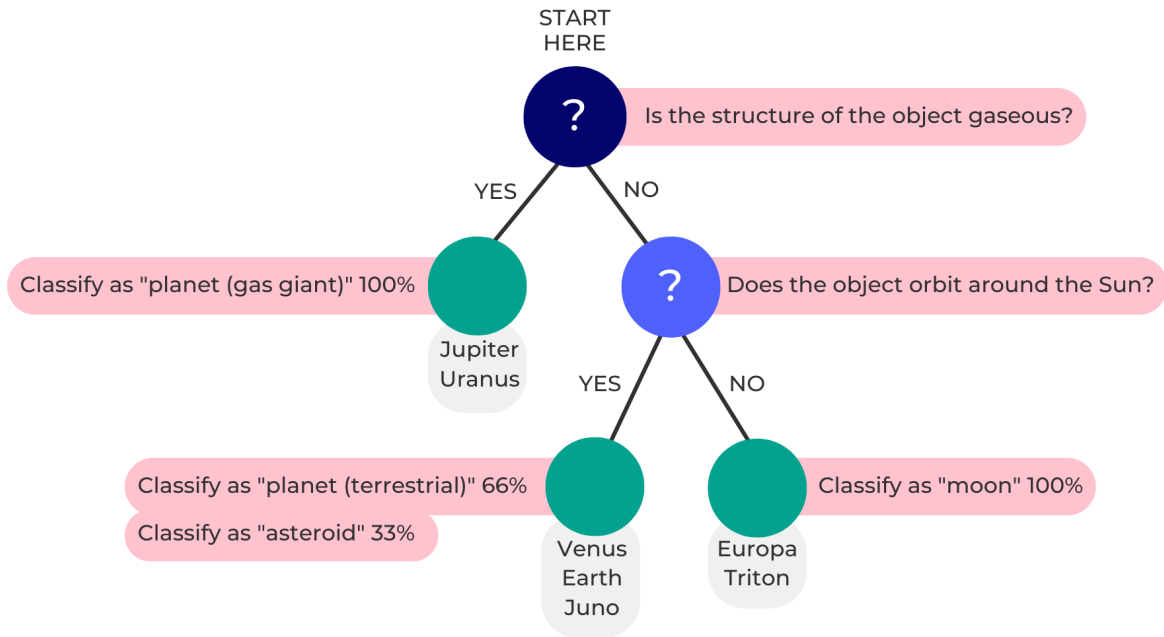
8. Follow the same steps as above to continue drawing your tree.
 - a. Look at each item that was in your second group and sort each item using your second question.
 - b. Once you're done sorting the items based on the second question, look at the groups that have been made. Do any of the groups have *only one* type of object in them? If so, add decisions to them like we did for the gas giants.



9. The example above is notable because after two questions, we have a strategy for separating our objects into the three classes that we had defined in a way that makes sure they're not mixed with any other classes (which is why we can have 100% confidence in each decision). This is not always the case. It's possible to have data that you can't fully separate.



10. What if we included a new object? Include Juno (solid, orbit: Sun), an asteroid, in our objects and then try to separate it out. How does this change our tree? Our last tree might look instead like:



11. Notice how our confidence in our classification decision has changed:
- a. Since only 2 out of 3 objects in this group are a planet, we only have a 66% chance that we're correctly classifying an object as a terrestrial planet in this category.
 - b. Since 1 out of 3 objects in this group is an asteroid, there's a 33% chance that an object in this category should be classified as an asteroid.



Section 2: Classifying Space Objects

Examining the space objects dataset

1. Now let's take a look at the space objects dataset that we will be working with. This dataset has been generated to resemble the data that the station will use. The space objects dataset contains five labelled types of celestial objects (asteroid, comet, junk, meteoroid, satellite), and has three features for each object:
 - a. material composition, either **organic** (containing water or other organic compounds) or **inorganic** (metal and rocks)
 - b. size, measured as the approximate diameter of the object (in metres)
 - c. distance, as either **local** (inside or near the Earth's orbit) or **far** (well outside Earth's orbit)
2. There are two subsets within this dataset:
 - a. a **training dataset**, which you will use to build your decision tree
 - b. a **testing dataset**, which you can use to test your decision tree
3. You will use the training dataset to create the questions that you will use to separate the data, then test to see how accurately your tree performs using your testing dataset. Both sets are included as appendices.

Creating your decision tree

The goal of a decision tree is to separate data into distinct groups (as much as possible—perfect separation may not be possible) based on known, comparable features. This separation is achieved by asking questions of the data that results in it being split. The space objects dataset has three features that we can use.

Two of these features, “material composition” and “distance”, are of a type that we dealt with in the previous example: they are *categorical* data. This means that they can only take a specific, limited range of values. The third feature, “size”, is *continuous* data, meaning that it can be any number. For each of the categories of space objects in this activity, though, size will be within a certain range. With this in mind, { individually / in small groups / as a large group}, complete the tasks below.



Task 1: Categorical features

1. Develop questions based on the **categorical** features. If you get stuck, think back to how questions were used in the example. *“Is the structure of the object solid?”* Adapt those questions to the features of the space object dataset.
2. Use the questions that you developed to draw a decision tree. This means choosing which questions to ask and in what order you want to ask them.
3. Apply your decision tree to the **training** dataset, noting at each step the composition of the groups and at each step along the way and the probabilities for the final classification decisions (as the fraction/percentage of each object type in the group).

Task 2: Categorical and continuous features

1. Now, develop questions based on the **continuous** data feature. If you get stuck, here’s a hint for how to deal with continuous data: *think about the range of sizes (minimum and maximum) for each object and how you might be able to use this to separate them. Are some objects obviously bigger or smaller than others?*
2. Create a new decision tree that combines both types of features.
3. Apply your decision tree to the training dataset, noting at each step the composition of the groups and at each step along the way and the probabilities for the final classification decisions (as the fraction/percentage of each object type in the group).

Section 3: Evaluating a tree

After you’ve developed your decision tree, you can use the **testing** dataset to see how well it works. To do this:

1. For each object in the testing dataset, apply your decision tree to see where it would be classified. Keep track of these outcomes, since you will need them in the next step.



2. Once you have sorted each row of the testing dataset into the leaf nodes of your tree, score the outcome for each leaf by multiplying the number of objects by their classification confidence (the probability of being right). For example:
 - a. If you have 5 asteroids at a leaf node that has a 100% confidence for the decision “Classify as asteroid”, $5 \times 100\% = 5 \times 1 = 5$.
 - b. If you have 3 comets and 2 meteoroids at a leaf node that has a 60% confidence for the decision “Classify as comet” and a 40% confidence for the decision “Classify as meteoroid”, $(3 \times 60\%) + (2 \times 40\%) = (3 \times 0.6) + (2 \times 0.4) = 2.6$.
3. Add up the scores across all of your leaf nodes. This number should not exceed the number of objects in the testing dataset. The highest possible value, in this case, should be 10, but that assumes that the data is perfectly sorted, which may not be the case.
4. If you have time, try this testing process on both of your trees and compare your results.

Reflection & Debrief

Having tested your decision tree, compare your scores and trees in groups or as a class. A higher score indicates a tree that should be more effective at sorting data. Consider the following questions:

1. What approaches were used to decide on which questions should be asked?
2. How did you determine the order to ask your questions in?
3. How many layers/levels did your tree have? Can you see any connection between the number of layers/levels and a tree’s testing score?
4. Did you ever repeat a question?
5. Considering both the example and the task dataset,
 - a. What impact does training data have on the structure and design of a decision tree?
 - b. What impact do the data features have on the structure and design of a decision tree?
 - c. How can you prove that your decision tree is as effective as possible?



Decision trees are just one example of a strategy that can be used to classify or sort data. Data, however, isn't always available in easily readable formats like tables and spreadsheets. In subsequent activities, you will look at other ways that classification can occur, including using machine vision and listening to recognize specific images and sounds.

Delivery Adaptations

How might you adapt the time, space, materials, group sizes, or instructions to make this activity more approachable or more challenging? **Modifications** are ways to make the activity more accessible, **extensions** are ways to make the activity last longer or more challenging.

Modifications

Simplify the dataset

- The dataset for the activity is designed so that it can't be perfectly separated ("junk" and "satellite" have overlapping size ranges and are both "local" and "inorg"). If you think that this would be confusing, you can remove one of those categories or instruct students to create a "junk or satellite" category (e.g. "human-made objects").

Extensions

Expand the example (planets and moon in the solar system) dataset

- The example was done with only a small subset of the possible objects and the categories in our solar system. Instead of using the generated dataset, have students work together to build a new set of objects.

Optimizing the decision tree

- The scoring method provided in Activity 3 works to give a general sense of how good a tree performs and it should let you compare different trees to each other. Comparing the trees made by different groups, what trees performed the best?



- Is there a relationship between the number of layers (height) of a tree and how well it performed?

Terms of Use

Prior to using this activity or parts thereof, you agree and understand that:

- It is your responsibility to review all aspects of this activity and ensure safety measures are in place for the protection of all involved parties.
- Any safety precautions contained in the “Safety Considerations” section of this write-up are not intended as a complete list or to replace your own safety review process.
- Actua shall not be responsible or liable for any damage that may occur due to your use of this content.
- This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. For more information, please see <https://creativecommons.org/licenses/by-nc-sa/4.0/>.
- You may adapt the content for your program (remix, transform, and build upon the material), providing appropriate credit to Actua and indicating if changes were made. No sharing of content with third parties without written permission from Actua.

About Actua

Actua is Canada’s leading science, technology, engineering and mathematics (STEM) youth outreach network, representing a growing network of over 40 universities and colleges across the country. Each year 350,000 young Canadians in over 500 communities nationwide are inspired through hands-on educational workshops, camps and community outreach initiatives. Actua focuses on the engagement of underrepresented youth through specialized programs for Indigenous youth, girls and young women, at-risk youth and youth living in Northern and remote communities. For more information, please visit us online at www.actua.ca and on social media: [Twitter](#), [Facebook](#), [Instagram](#) and [YouTube](#)!

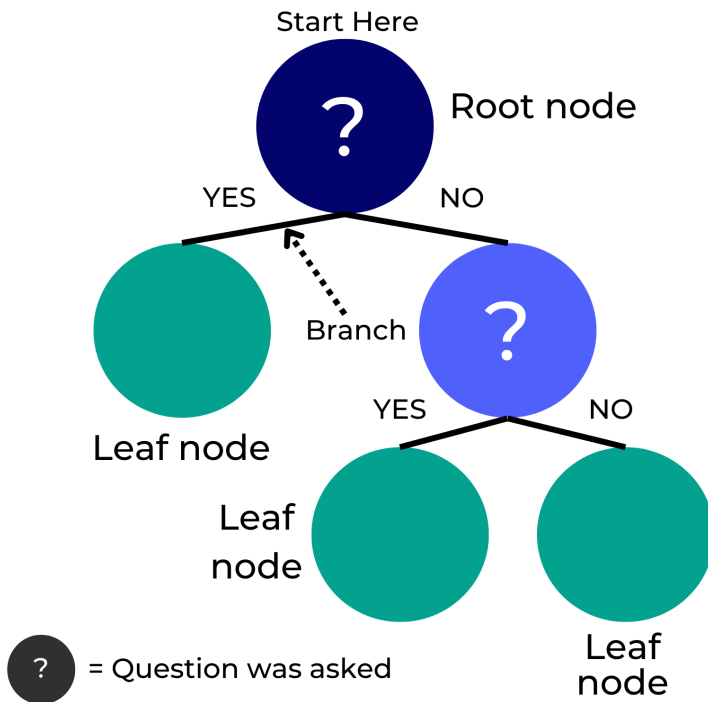


Appendices

Appendix A: Career & Mentor Connections

- Statistician
- Computer Programmer
- Data Analyst
- AI Programmer

Appendix B: Background Information



DECISION TREES

A decision tree is structured as an upside-down tree. The top of a decision tree is called the “root” or “root node”. Questions are asked at nodes, so this is where the first question is asked. From the root, there are two branches, each representing an answer to the question, and at the end of each of these branches is another node. When there are no more questions to ask at a node, no branches get added and that node is called a “leaf”. While the decision tree questions that are asked can have

multiple answers, most often they are binary, i.e. there are only two possible answers that the data can fit into.



Appendix C: Supporting Resources

DATASET SAMPLES

Space objects training dataset

Features			Label
Material composition	Size (d, metres, approx.)	Distance (local or far)	
inorg	27.32	far	asteroid
inorg	164.06	far	asteroid
inorg	169.13	far	asteroid
org	657.43	far	comet
org	879.37	far	comet
org	905.20	far	comet
org	957.84	far	comet
inorg	0.21	local	junk
inorg	2.51	local	junk
inorg	3.68	local	junk
inorg	4.40	local	junk
inorg	9.33	local	junk
inorg	9.44	local	junk
inorg	4.03	local	meteoroid
inorg	4.38	local	meteoroid
inorg	4.92	far	meteoroid
inorg	6.55	far	meteoroid
inorg	24.40	local	satellite
inorg	25.76	local	satellite
inorg	26.87	local	satellite



Space objects testing dataset

1. Use the data below to test to see how well your decision tree performs.

Features			Label
Material composition	Size (d, metres, approx.)	Distance (local or far)	
inorg	157.02	far	asteroid
inorg	181.30	far	asteroid
org	656.60	far	comet
org	700.02	far	comet
org	939.23	far	comet
inorg	1.48	local	junk
inorg	4.02	local	meteoroid
inorg	12.28	local	satellite
inorg	23.04	local	satellite
inorg	26.87	local	satellite

