# Scoop Tutorial

## Introduction

Hi, my name is [name], and I’ll be walking you through our study. If your cell phone is on, please set it to silent mode.

[Wait.]

I’ll be reading from this script to be consistent in the information I provide you and the other people taking part in this study. Please don’t discuss this study with anyone, as we don’t want other participants to receive any advance information.

You’ll be using software that automatically labels messages by topic. Our research is about testing software like this. “Testing” means that you will be deciding when the computer’s outputs are right or wrong. If the computer makes mistakes, “testing” will help you find them.

Let’s get started. Let me know if you have any questions as we go along.

[Wait for questions.]

Press the “OK” button to start the application. If your screen doesn’t look like the projector, please raise your hand.

[Driver clicks OK. Wait until everyone has it open.]

The computer has tried to label these messages by topic. There are a lot of messages -- too many to test all of them individually. You’ll try to decide how good the computer is, or whether it makes too many mistakes to be useful.

At the top of the window are messages with their subjects and dates listed.

[Wait. Driver circles the message subjects and dates.]

You can sort messages in different ways. Let's sort by topic by clicking on the “Topic” column, like this:

[Driver sorts by Topic. Wait.]

You can group the messages in different ways by sorting on each column. For now, let’s go back to sorting them by date.

[Driver sorts by Date. Wait.]

Most computer programs don’t have a “Correctness” column like this one, so it has a tooltip to tell you what it means.

[Driver hovers over Correctness column until tooltip appears.]

### CONTROL:

Can someone please read this tooltip?

### NON-CONTROL

The [ confidence / unusualness / relevant words ] column also has a tooltip. Can someone please read these tooltips?

[Wait until someone answers]

Right, the Correctness column basically tells you which messages have been tested, and whether they were right or wrong. The question marks you see mean that none of these messages have been tested yet.

### NON-CONTROL

The [ confidence / unusualness / relevant words ] column tells you how much the computer thinks you should test each message, and why.

[\*\*Insert widget definition here.]

Now let’s scroll back to the top and sort the list by date again.

[Driver sorts by date.]

Click on the third subject line labeled, “rm.split.” You may need to read the whole message to decide whether the computer’s label is right.

[Driver clicks on the message and hovers over the message body. Wait.]

The subject of this message isn’t very clear, but the text clearly discusses motorcycles. I think the computer’s label is right. Let’s tell the computer it’s right, so that we can keep track of which messages we’ve tested. Click on the question mark in the correctness column.

[Driver clicks on widget. Wait to make sure everyone else does too.]

A slider with X’es and check marks should appear.

[Driver circles around the slider.]

If you hover over the slider, a tooltip will appear explaining what the different marks mean. Could somebody read them out loud now?

[Wait for someone to read them.]

Right, the X’es mean the computer is wrong, and the checks mean it is right. Because we’re really certain that this message is right, let’s use the big checkmark. Drag the slider over to the big checkmark.

[Driver drags the slider to the big checkmark. Wait to make sure everyone follows.]

Did anyone see anything change on screen?

[Wait for responses.]

Right! There’s a checkmark in the Correctness column, which reminds us that this message has been tested. The computer can also learn from its mistakes, so sometimes it may change these topic predictions.

[Driver hovers over Topics column. Wait.]

### NON-CONTROL

The [ confidence / unusualness / relevant words ] icon goes away after you test the message. The History column shows you what the [ confidence / unusualness / relevant words ] and Correctness icons used to be, to help you keep track of things that changed.

If the computer thinks a message is very similar to the one you just tested, it will automatically get tested as well. Let’s sort on the “History” column to see if anything else was tested by our check mark.

[Driver sorts on History column. Wait for everyone to catch up.]

When the computer tests a message, it uses a gray icon instead of a black one. Can someone tell me one of the messages that was tested by the computer?

[Wait for response.]

Good. Now let’s sort by date again.

[Driver sorts by Date.]

Look at the progress bar.

[Driver circles the progress bar.]

The progress bar tells us how many messages have been tested, and also tells us how many were right and how many were wrong. If you hover over the bar, a tooltip will tell you exactly how many messages have been tested.

[Driver hovers over bar until tooltip appears.]

Testing high priority messages may help you find the computer’s mistakes quickly. Testing low-priority messages may help you eliminate groups of messages that aren’t likely to have mistakes.

Let's move on to another message. Please click on the 5th message from the top, “SAAB Mailing List”.

[Driver selects message. Wait.]

After reading through the message, I think it may be wrong, but I’m not sure because it talks about “mailing lists” but doesn’t specify cars or computers. Let’s click on the question mark and look at our options.

[Driver clicks question mark. Wait.]

Which option should I use to tell the computer that it may be wrong?

[Driver hovers over slider until tooltip appears to remind participants about it. Wait until someone responds.]

A small X is good for situations like this. Let’s drag the slider over to the small X.

[Driver drags slider to small X. Wait.]

### NOT CONTROL

Like before, several things changed.

The [ confidence / unusualness / relevant words ] icon disappeared and the Correctness column shows our small X.

[Driver circles around these areas.]

Also, look at the progress bar. It increased a bit, and we can see how there’s one color for check marks, and another color for X marks.

[Driver circles around these sections of the progress bar.]

As you can see, both checks and X’es increase the progress bar.

The History column changed, too. Look at the changes from when we checked off the third message – the icons slid to the right and got smaller. Each time you test a message, the History icons slide over and shrink down, until they disappear off the edge of the screen. This is because the History column only shows you what recently changed.

Let’s say I changed my mind about this message. I can undo a test by clicking on the small X and moving the slider back to the middle, like this:

[Driver drags slider back to the question mark. Wait.]

We’ve used the big checkmark and the small X; there is also a small checkmark and a big X. For instance, I might see a message labeled “Cars” that talks about engines, but I don’t know a lot about car engines, so I might not be confident enough to say the computer is definitely right. The small checkmark is for messages that might be right, but I’m not 100% certain.

The big X is for messages that are definitely wrong, like a message labeled Motorcycles but only contains the sentence “How late is the library open?”

If you know a topic is wrong, and also know what it SHOULD be, you can fix the topic instead of using the X mark. Let’s tell the computer that this message should be labeled “Cars” instead of “Computers” by clicking on the Topic menu and selecting “Cars.”

[Driver clicks the topic menu to display it, and slowly selects “Cars” from the list.]

Notice how a big checkmark has appeared? This is because you told the computer the right topic. Since you chose the topic, the computer assumes it’s right.

## Subsequent Versions

Thanks for your help so far. Let’s go through the [next / final] version. Press the “OK” button to start a new variant of the application.

### NOT CONTROL AND ALREADY DESCRIBED NON-CONTROL VERSION

You’ve already seen a version similar to this.

[\*\*insert the widget definition here\*\*]

### CONTROL

A few things have changed. The [ confidence / unusualness / relevant words ] and History columns are gone, as is the progress bar. You can still place checkmarks and X’es to test like before, but the computer won’t also test similar messages.

### FIRST NON-CONTROL VERSION

A number of things have changed. The [ confidence / unusualness / relevant words ] column is new, and has a tooltip. Can someone please read this tooltip?

[Wait until someone answers]

Right. The [ confidence / unusualness / relevant words ] column tells you how much the computer thinks you should test each message, and why.

[\*\*Insert widget definition here.]

Now let’s scroll back to the top and sort the list by date again.

[Driver sorts by date.]

Let’s look at the third message, labeled “rm.split”, and change its Correctness to a big check.

[Driver slowly checks off the message. Wait.]

The [ confidence / unusualness / relevant words ] icon, which tells you how much the computer thinks you should test each message, goes away after you test the message.

The History column shows you what the [ confidence / unusualness / relevant words ] and Correctness icons used to be, to help you keep track of things that changed.

[Driver circles History column header]

If the computer thinks a message is very similar to the one you just tested, it will automatically get tested as well. Let’s sort on the “History” column to see if anything else was tested by our check mark.

[Driver sorts on History column. Wait for everyone to catch up.]

When the computer tests a message, it uses a gray icon instead of a black one. Can someone tell me one of the messages that was tested by the computer?

[Wait for response.]

Good. Now let’s sort by date again.

[Driver sorts by Date.]

Look at the progress bar.

[Driver circles the progress bar.]

The progress bar tells us how many messages have been tested, and also tells us how many were right and how many were wrong. If you hover over the bar, a tooltip will tell you exactly how many messages have been tested.

[Driver hovers over bar until tooltip appears.]

Testing high priority messages may help you find the computer’s mistakes quickly. Testing low-priority messages may help you eliminate groups of messages that aren’t likely to have any mistakes.

Let's change the topic of the 5th message, “SAAB Mailing List”, to “Cars,” like we did earlier.

[Driver changes topic of message to “Cars”.]

Look at the progress bar. It increased a bit, and we can see how there’s one color for check marks, and another color for X marks.

[Driver circles around these sections of the progress bar.]

As you can see, both checks and X’es increase the progress bar.

The History column changed, too. Look at the changes from when we checked off the third message – the icons slid to the right and got smaller. Each time you test a message, the History icons slide over and shrink down, until they disappear off the edge of the screen. This is because the History column only shows you what recently changed.

Everything else behaves just like the last version.

## CONFIDENCE Definition

There’s a multi-colored pie graph inside a green square.

[Driver circles the pie widget. Wait.]

The brighter the green square is, the higher the priority for testing. The pie explains the priority; each slice show how strongly the computer believes the message goes with the topic of that color. So, if the pie’s slices are all about the same size, the computer is really having trouble choosing among them. To see an example of this, sort the messages based on “Confidence”, like on the screen.

[Driver sorts by confidence.]

(\*\*Only bring up if someone says something.)

If your screen does not look exactly like ours, don’t worry about it. Each computer may be a little different.

To see exactly how confident the computer is, hover over a pie to bring up a tooltip.

[Driver hovers until tooltip appears. Wait.]

The dark green square is there because the computer isn’t very confident about this message, so it thinks its more important to test this message than a message with a light green square, like at the bottom of the screen.

[Driver scrolls to the bottom of the list, pauses, then scrolls back up.]

## UNUSUALNESS Definition (aka CosSimilarity)

There’s a partially filled circle inside a green square. The darker the green the square is, the higher the priority for testing. The circle explains the priority; a mostly white circle means the computer has seen many other messages with words like this one, so it thinks this message is normal. A mostly blue circle means the computer thinks this message is unusual, so it may not be very good at choosing its topic. To see the messages that are the most unusual, please sort by “Unusualness”.

[Driver sorts by Unusualness.]

(\*\*Only bring up if someone says something.)

If your screen does not look exactly like ours, don’t worry about it. Each computer may be a little different.

To see exactly how unusual the message is, hover over a circle to bring up a tooltip.

[Driver hovers until tooltip appears. Wait.]

The dark green square is there because the computer thinks this message is very unusual, so it thinks its more important to test this message than a message with a light green square, like at the bottom of the screen.

[Driver scrolls to the bottom of the list, pauses, then scrolls back up.]

## RELEVANCE Definition

There’s a large number inside a green square. The darker the green the square is, the higher the priority for testing. The number explains the priority; it represents how many words in the message are also on the computer’s list of the 20 most important words. By having more of these words in a message, the computer is more likely to know the topic of the message. To see the messages that have the fewest relevant words, sort by Relevance now.

[Driver sorts by Relevance. Wait.]

The dark green square is there because without many relevant words, the computer isn’t certain how to predict a topic, so it thinks its more important to test this message than a message with a light green square, like at the bottom of the screen.

[Driver scrolls to the bottom of the list, pauses, then scrolls back up.]

## End of Tutorial

I’m going to give you a couple of minutes to practice finding the computer’s mistakes. If you forget what some of the icons mean, the tooltips may help remind you. You can also ask us if you have questions. You’ll have two minutes to practice finding mistakes now.

[Wait for two minutes.]

Alright, let’s move on.

### IF FIRST RUN THROUGH

When you think you’ve found all of the computer’s mistakes, click the “I’m done!” button. Let’s click it now.

[Driver slowly clicks “I’m Done!” button. Waits.]

Once you confirm that you are done, you cannot go back and test more messages. We’re done testing these messages, so we’ll click “Yes, I’m done testing” now.

[Driver clicks “Yes, I’m done testing”. Waits.]

A new application window should automatically appear. Your task is to find all of the computer’s mistakes. Afterwards, we’ll ask you how accurate you think the computer is. You won’t have enough time to look at every message, so pick tests that help you find mistakes fast. You’ll have ten minutes to work. If you finish early, click “I’m Done!” and wait until everyone has finished. Try to find all of the computer’s mistakes now.

### SUCCESSIVE RUN THROUGHS

Click the “I’m done!” button to close the application. A new application will appear momentarily. Like before, you have 10 minutes to find the computer’s mistakes. If you finish early, click “I’m Done!” and wait until everyone has finished. Try to find all of the computer’s mistakes now.

## After Main Task

[Driver should block projector so nothing is on-screen. Wait 10 minutes, or until everyone is finished.]

Alright, we’re going to move on. Please click the “I’m Done!” button if you haven’t already done so. A series of questions are being handed out. Please answer them as best you can, there are no wrong answers. When you finish, let one of the helpers know and we’ll collect your answers.

[Wait until everything has been collected. Driver un-blocks projector.]

## End of Session

Alright, that completes the experiment. Thank you so much for all of your help. Please stay seated and we will be coming round to give you your incentives.