

Provenance Scenarios

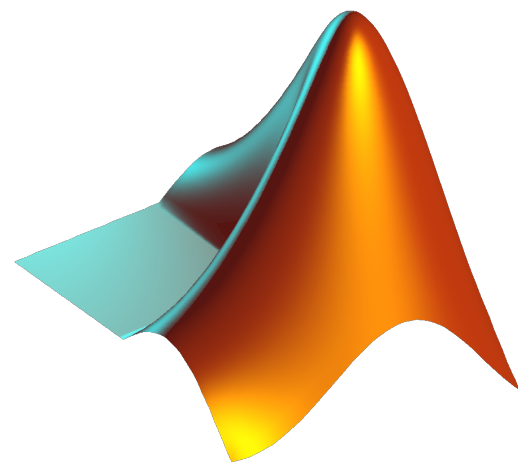
As a *<role>*, I want to *<goal>* so I can *<reason>*.

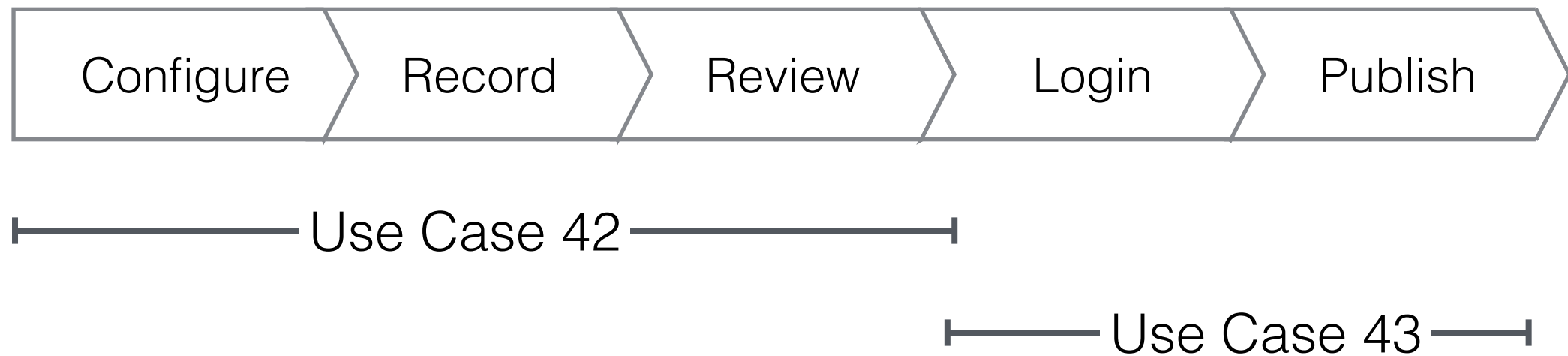
Use Case 41

As a data analyst using R or Matlab, I want to easily document my script with standardized comments so I can understand the workflow of the script.

Use Cases 42, 43

As a data analyst using R or Matlab, I want to keep track of my data input files, data output files and scripts so I can review my runs and potentially choose those to share with colleagues through an established DataONE repository.





Will be using Matlab code examples

Configure

Record

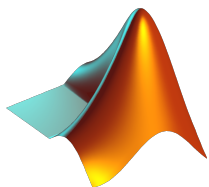
Review

Login

Publish

my_script.m

```
1  import org.dataone.client.configure.Configuration;
2
3  config = Configuration;
4
5  set(config, ...
6      'baseDirectory', ...
7      '/Users/cjones/matlab/runs' );
8
9  set(config, ...
10     'sourceRepositoryBaseURL', ...
11     'https://mercury-ops2.ornl.gov/MSTMIP/mn' );
```



Configure

Record

Review

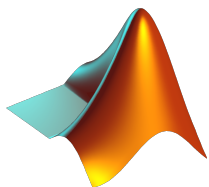
Login

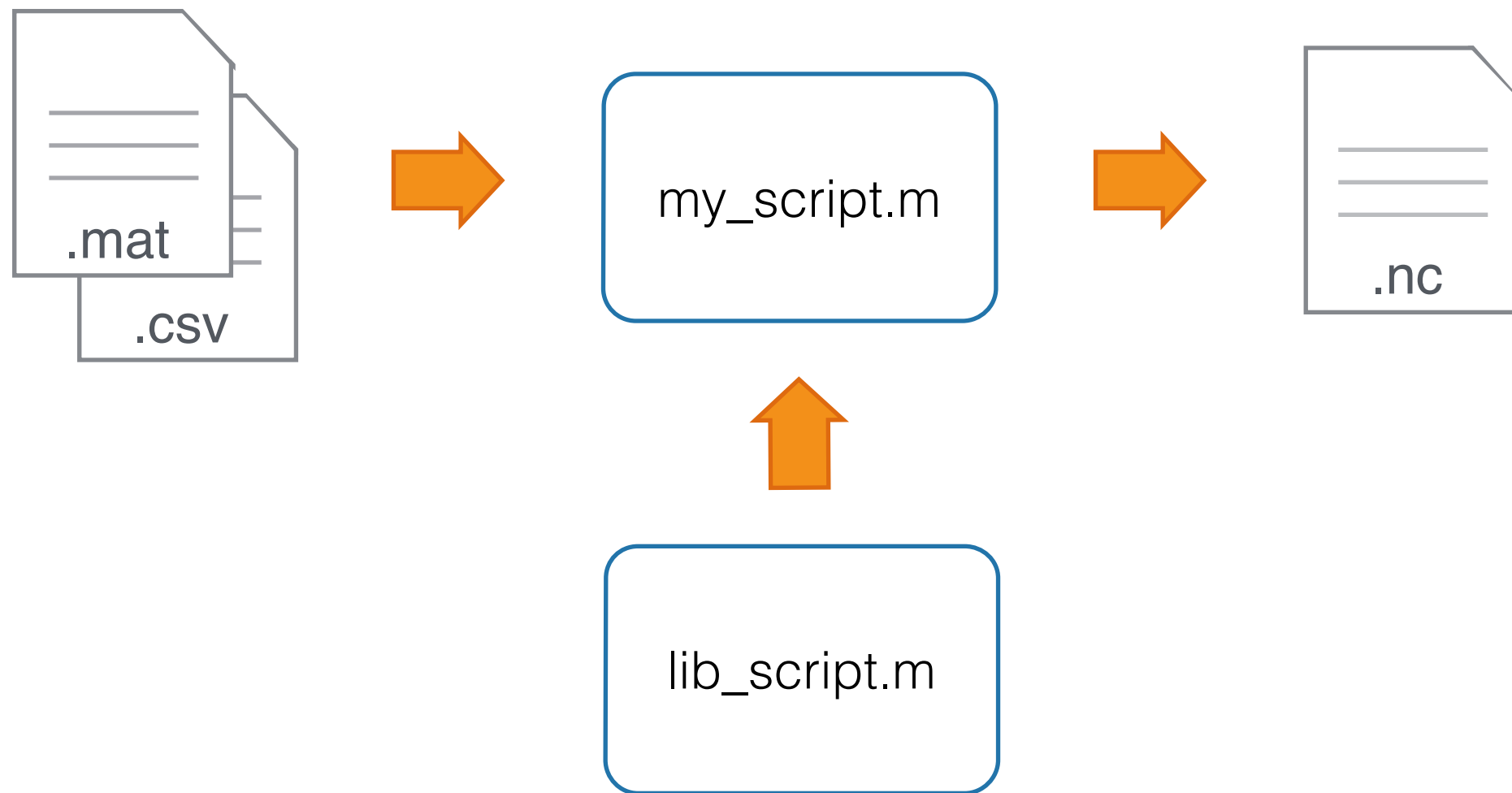
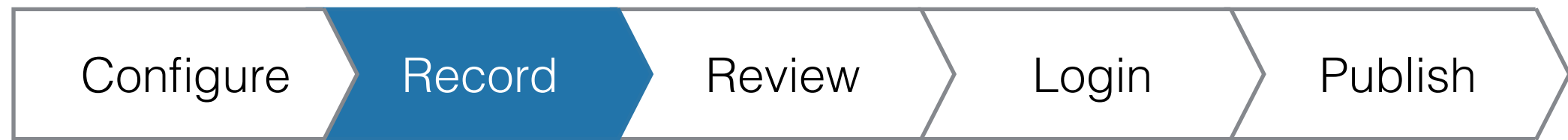
Publish

my_script.m

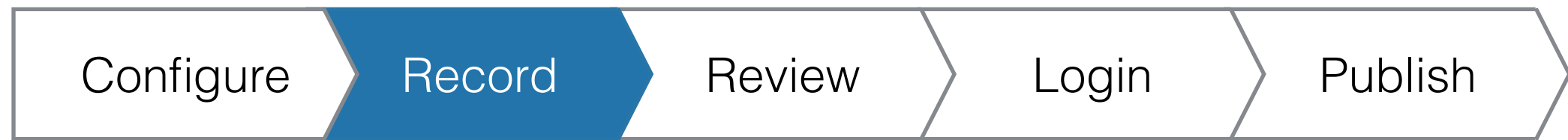
```
1 import org.dataone.client.configure.Configuration;
2 import org.dataone.client.configure.LabelParser;
3
4 parser = LabelParser;
5 parser.parse; % looks for config in comments
6
7 % dlprov:ingestStep
8 % prov:used, '/Users/cjones/data.nc'
9 some_array = ...
10     get_input_data('/Users/cjones/data.nc');
```

Alternative: develop a cross-language
markdown-like annotation syntax and parsers



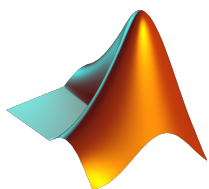


`record()` tracks input files, output files, referenced scripts, and the main script



my_script.m

```
1  import org.dataone.client.run.RunManager;  
2  
3  runManager = RunManager;  
4  
5  runId = runManager.record('/Users/cjones/my_script.m');  
6  
7  % run your model analysis here  
8  
9  
10
```

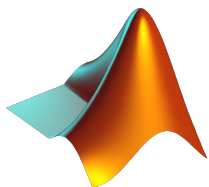


record() hides details of using
insertRelationship() under the hood



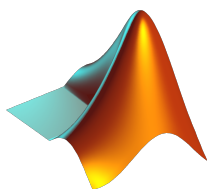
my_script.m

```
1  runManager.list(); % Prints out summary about all runs
2
3  runManager.view(runId); % More details about a run
4
5
6
7
8
9
10
```





```
1  runManager.list(); % Prints out summary about all runs
2
3  runManager.view(runId); % More details about a run
4
5
6
7
8
9
10
```



The record/review cycle is iterative



my_script.m

```
1 runManager.view(execution.1.1) ;
```

```
Run: execution.1.1
```

```
-----
```

```
run start time: Mon Sep 8 13:01
```

```
run end time: Mon Sep 8 13:02
```

```
Matlab version: 2014a
```

```
Operating system: Mac OS X 10.9.5
```

```
Host name: laurenshome
```

```
Data Package: datapackage.1.1
```

```
-----
```

```
figure.1.1 was generated by execution.1.1
```

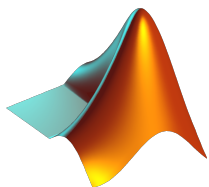
```
execution.1.1 used data.1.1
```

```
execution.1.1 used script.1.1
```

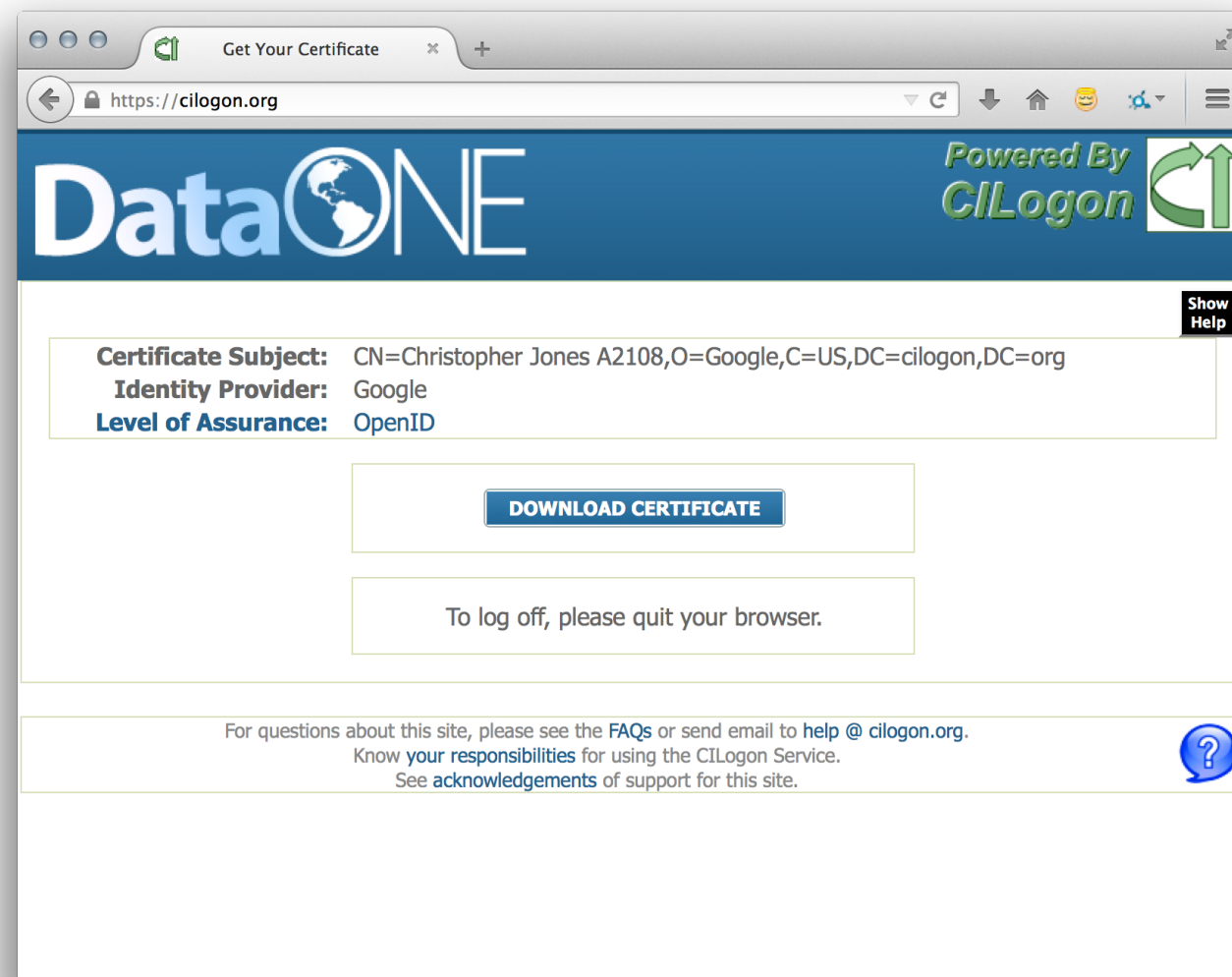
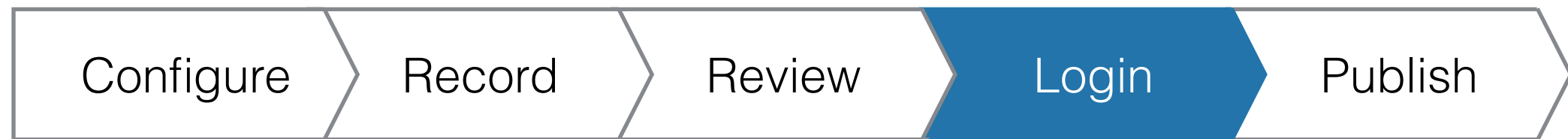
```
data.1.1 is documented by metadata.1.1
```

```
figure.1.1 is documented by metadata.1.1
```

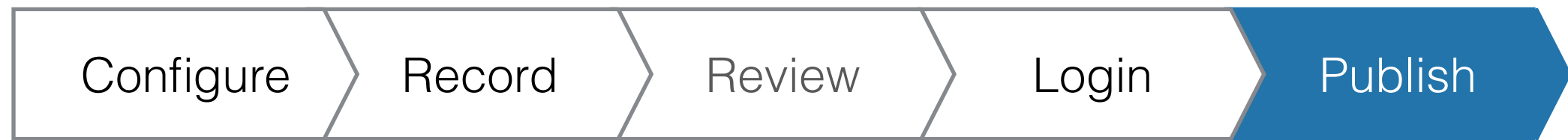
```
script.1.1 is documented by metadata.1.1
```



We need to discuss what should be shown in summaries and details



Login to DataONE and download your certificate

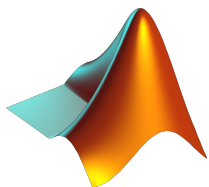


my_script.m

```
1 % Publish the data package to the repository
2 runManager.publish('datapackage.1.1');
3
```

RunManager will:

- create identifiers for each Data Package member
- upload each to the target DataONE Member Node
- upload the Data Package with all PROV relationships



UC 42/43 Todos

- Discuss the envisioned process (12-month goals)
 - Will it work? Are there gaps?
- Configuration Step:
 - Decide on programmatic vs inline-comment
 - Determine what configuration options are required
- Review Step
 - Discuss what is shown in summary and detail views

Use Case 43

As a data analyst using R or Matlab, I want to publish my data and their history so I can share them with colleagues through an established DataONE repository.

Use Case 44

As a scientist, I want to be able to examine the original datasets used in a derived dataset I've found through DataONE so I can understand the history and composition of the derived dataset.

Use Case 45

As a scientist, I want to be able to find all derived datasets in DataONE that use my dataset so I can understand how my data are being used and by which colleagues.

Use Case 46

As a scientist reviewing derived tables or figures, I want to be able to examine the original datasets and the original script used to generate them so I can understand their history and composition.

Search

View

Download

The screenshot shows a web browser window with the URL <https://cn.dataone.org/onemercury/>. The page features the DataONE logo and navigation links: About, Participate, Resources, Education, and Data. The main section is titled "ONE Mercury" and "A DataONE Search Tool for Scientific Data". It includes a "Search For:" input field with a hint: "Hint: boolean operators and phrases are allowed. ex: precipitation or (rain and 'moisture content')". To the right of the search field is a "Results/Page" dropdown set to "10" and a "SEARCH" button. Below the search field are three buttons: "Show/Hide Advanced Options", "Clear All", and "Help". The interface is divided into three main search sections: "Fielded Search" with three "Full Text" input fields and "AND" operators; "Date Search" with radio buttons for "Collection Date", "Publication Date", and "Either", a "during" dropdown, and date input fields; and "Geographic Search" which includes a map of the world with labels for various countries like Germany, France, Italy, Spain, Turkey, China, and others. The map also shows "United States" and "North".



Doe, John. 2014. Seabirds of the Gulf of Alaska and North Pacific.
([seabirds.2.1](#))

Species_List **Download**

60.607666	-145.87834	8:14
60.607666	-145.87984	8:14
60.607838	-145.88133	8:15
60.607838	-145.88266	8:15
60.607998	-145.88417	8:15
60.607998	-145.8855	8:15
60.607998	-145.88699	8:16
60.60817	-145.88834	8:16
60.608334	-145.89	8:16
60.608334	-145.89116	8:16
60.608498	-145.89265	8:17
60.608498	-145.89418	8:17
60.608665	-145.89568	8:17

Description

Object Name

Size

Authentication

Format Name

Online Distribution Info


Species_List

189493

514259fe89f514259fe89fng5142vrf...

csv

seabirds.3.1

 This table was derived from [knb.485.1](#), [jstocking.4.9](#), and [jstocking.3.4](#) using the program [seabirds.6.1](#)

View
derivation
history






Doe, John. 2014. Seabirds of the Gulf of Alaska and North Pacific.
([seabirds.2.1](#))

Seabird_Survey_Script [Download](#)

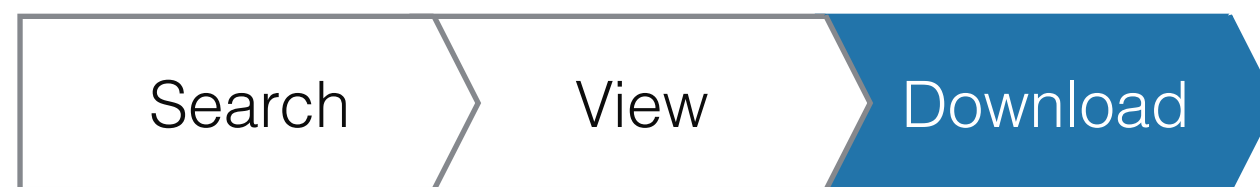
```
> samp.size = function(x)
+ {
+   n = length(x) - su
+   nas = sum(is.na(x)
+   out = c(n, nas)
+   names(out) = c("",
+   out
+ }
> ls()
[1] "nums"      "samp.siz
> samp.size(nums)
      NAs
24      1
```

Description	
Object Name	Seabird_Survey_Script
Size	189493
Authentication	514259fe89f514259fe89fng5142vrf...
Format Name	application/octet-stream
Online Distribution Info	seabirds.4.1

 This program generated [seabirds.5.1](#) using [seabirds.3.1](#)

View
generation
history





Doe, John. 2014. Seabirds of the Gulf of Alaska and North Pacific.
([seabirds.2.1](#))

Species_List **Download**

60.607666	-145.87834	8:14
60.607666	-145.87984	8:14
60.607838	-145.88133	8:15
60.607838	-145.88266	8:15
60.607998	-145.88417	8:15
60.607998	-145.8855	8:15
60.607998	-145.88699	8:16
60.60817	-145.88834	8:16
60.608334	-145.89	8:16
60.608334	-145.89116	8:16
60.608498	-145.89265	8:17
60.608498	-145.89418	8:17
60.608665	-145.89568	8:17

Description

Object Name

Size

Authentication

Format Name

Online Distribution Info


Species_List

189493

514259fe89f514259fe89fng5142vrf...

csv

seabirds.3.1

 This table was derived from [knb.485.1](#), [jstocking.4.9](#), and [jstocking.3.4](#) using the program [seabirds.6.1](#)

Follow links
to download



UC 44/45/46 Todos

- Provenance UI Design
- Review and Modify (tomorrow morning session)