

The Fellowship of the Data



Source:myblogs.pw

Sev Leonard
sev@thedatacout.com
gizm0_0

Ways to play

- View the presentation at <http://tinyurl.com/j6xpkvu>
- Follow along with the Jupyter notebooks on Github
 - <https://github.com/gizm00/pycon2016>
 - Notebooks directory will have the labs
- Run tutorial locally in Jupyter – see Github for required libs
 - Need to use CsvStorage object (more on this later)
 - Form scraping will not run
- Run Docker tutorial container - same caveats as local run
- Run Docker LAMP stack and tutorial container

Set the timer! 45min

Greetings! About me

- My pronoun.is/he
- 11 years in PDX
- 1st PyCon!
- Software and “data science” consultant
- Likes mangoes, outside, cats



The Data Scout

Your guide in the information wilderness



Class info

- Tutorial content & setup instructions:
<https://github.com/gizm00/pycon2016>
- HTTP requests, MYSQL storage are handled by the pycon2016_lamp Docker container
 - You will be able to run most of the tutorial without this if it is giving you grief
- Tutorial is based on data source mocks

About the tutorial

- Based on data gathering and organization techniques used for
www.campnear.me



Source:artistthinks.wordpress.com

Tutorial topics

- Mostly lab (Jupyter notebook) based
- Data gathering topics:
 - Working with APIs
 - Scraping webpages
 - Navigating forms
- OOD topics:
 - Creating uniform interfaces through abstraction
 - Extending classes
 - Encapsulating methods, data

Agenda

1. Overview of data sources, best practices for web data, and design techniques
2. Labs
 1. Getting data from an API
 2. Creating objects for storing and extracting data
 3. Scraping web pages
 4. Creating objects for scraping data
 5. Navigating forms
 6. Using the scraper object for reservation info
 7. Creating objects to merge data
 8. The One Pipeline

There and back again – a journey



Image source: knightleyemma.com

Recreation Information Database - RIDB

Home Data Download Data Sharing Mobile App Sign In Register

Recreation Information Database

Search RIDB catalog by state, organization, and activity

Recreation Areas Facilities

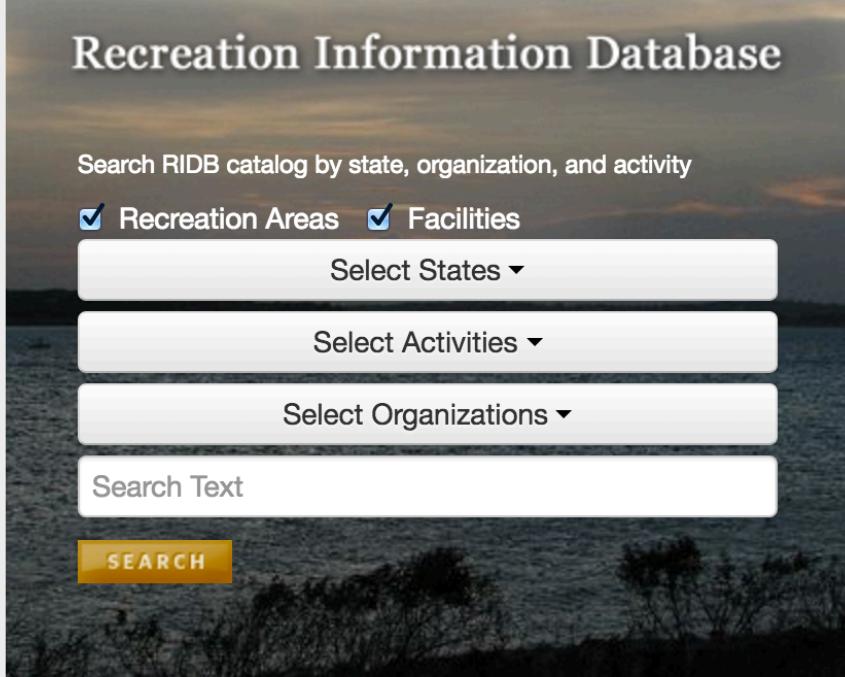
Select States ▾

Select Activities ▾

Select Organizations ▾

Search Text

SEARCH



Data Partners
U.S. Forest Service



ridb.recreation.gov

US Forest Service Websites

Search **Go**

[Site Map](#)

Gifford Pinchot National Forest

- [Home](#)
- [Special Places](#)
- Recreation**
 - [Bicycling](#)
 - [Camping & Cabins](#)
 - [Climbing](#)
 - [Fishing](#)
 - [Hiking](#)
 - [Horse Riding & Camping](#)
 - [Nature Viewing](#)
 - [OHV Riding & Camping](#)
 - [Outdoor Learning](#)
 - [Picnicking](#)
 - [Water Activities](#)
 - [Winter Sports](#)
 - [Other Activities](#)
- [Alerts & Notices](#)
- [Passes & Permits](#)
- [Maps & Publications](#)
- [Land & Resources Management](#)
- [Learning Center](#)
- [Working Together](#)
- [About the Forest](#)

Campground: Cultus Creek

Area Status: Closed X

Cultus Creek campground is located in a pristine wooded setting on the boundary of **Indian Heaven Wilderness**. The trailhead for a **Indian Heaven Trail #33**, a popular and challenging trail into the Wilderness, is located in the campground. The campground is also near traditional huckleberry picking fields which makes it quite popular in huckleberry season.

At a Glance

Current Conditions:	4/28/2016: Not accessible yet for season.
Operational Hours:	Open after snow melt, historically by mid-June. Call Mt. Adams Ranger Station at 509-395-3402 for opening date.
Fees	<ul style="list-style-type: none">▶ Camping: \$10/night/single unit; \$20/night/double unit; \$5/extravehicle fee▶ Day use: \$5/vehicle/day
Permit Info:	Wilderness permits are required when entering Wilderness. Permits are free and self-issued at trailheads.
Usage:	Light
Restrictions:	Maximum vehicle length is 32 feet.
Closest Towns:	Trout Lake, WA
Water:	No
Restroom:	Vault toilet
Passes:	<ul style="list-style-type: none">▶ Camping: 50% discount for single site camping with any of these passes: Interagency Senior, Interagency Access, Golden Age, or Golden Access.▶ Day use: A valid Recreation Pass is required at Cultus Creek/Indian Heaven trailheads at Cultus Creek Campground.

FIRE DANGER LOW TODAY!

Fire Danger Level and Information

Related Information

Campground: Cultus Creek

[News & Events](#)

Alerts & Warnings

⚠ Cowlitz Valley Flood Damage and Road Repair (last updated on 4/14/2016)

- [Help Keep Caves Open](#)
- [Spring Hours](#)
- [Climbers Bivouac Closed](#)

[View All Forest Alerts](#)

Quick Links

[Online Washington State Accessible Outdoor Recreation Guide](#)

Areas & Activities

[Find An Area](#)

Location

Latitude : 46.0476
Longitude : -121.7546
Elevation : 3996

Recreation.gov

Find Sites [x]

Availability

Arrival Date

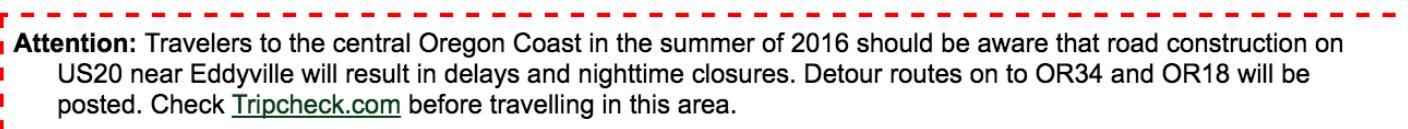
Departure Date

Not Flexible ▾

Loop

Any Loop ▾

Site #

**Attention:** Travelers to the central Oregon Coast in the summer of 2016 should be aware that road construction on US20 near Eddyville will result in delays and nighttime closures. Detour routes on to OR34 and OR18 will be posted. Check [Tripcheck.com](#) before travelling in this area.

Find Other Facilities

CAPE PERPETUA, OR

part of Siuslaw National Forest, US Forest Service



 19

Status: Open through Sun Sep 04 2016 [Season Dates](#)

Tidy data principles

1. Each variable is a column
2. Each observation is a row
3. Each type of observational unit is a table

Source: Wickham 2014 <http://www.jstatsoft.org/v59/i10/paper>

Observation: campground

Variables: lat/long, name, website, etc

Observational unit – data source (RIDB, USFS, recreation.gov)

Web data best practices



Source:kaiju.wikidot.com

1. Web data is ephemeral



Source:lotr.wikia.com

Format, availability, access requirements can change at any time

2. Just because you can get the data...

You may use the APIs if your Application is designed to help LinkedIn members be more productive and successful.

You may not use the APIs if your Application: exceeds a reasonable amount of API calls; relies fundamentally on the APIs; stores more than a LinkedIn member's profile data; or is used for hiring, marketing or selling.

1.4 Eligibility Criteria

Here is the eligibility criteria to determine if you may use the APIs pursuant to these Terms:

You may use our APIs if:

1. you are developing an Application designed to help LinkedIn members be more productive and successful across the web; for example, by augmenting their profile and professional brand in an Application for publishing, discussing and sharing content with like-minded professionals.
2. your Application is NOT expected to: have more than 250,000 lifetime members; make more than 500,000 daily calls to an API; make more than 500,000 lifetime people search calls to an API; or serve greater than 1 million daily plugin impressions.
3. your Application DOES NOT rely on access to the APIs as a fundamental aspect of your business.
4. your Application WILL NOT store or export any data from LinkedIn other than the LinkedIn Profile Data for the LinkedIn member that requested the data. **“Profile Data”** means the name, photo, headline, contact information, experience, education, summary, and location of a LinkedIn member. Profile Data excludes connections, network updates, job listings, groups, companies, and any other content.



Source:realitysandwich.com

<https://developer.linkedin.com/legal/api-terms-of-use>

... doesn't mean you can use it



A group of researchers has released a data set on nearly 70,000 users of the online dating site OkCupid. The data dump breaks the cardinal rule of social science research ethics: It took identifiable personal data without permission.

<http://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release>



Source:doblu.com

3. Terms of Use - attribute properly

The screenshot shows the top navigation bar of the Oregon State Parks website. It includes links for "Special Notices", a search bar, and "Share". Below the navigation is a menu bar with icons and text: "visit" (with a tent icon), "things to do" (with a person icon), "get involved" (with a group icon), "about us" (with a shield icon), and "shop" (with a shopping bag icon). On the left, there's the Oregon State Parks logo and the "Nature HISTORY Discovery" wordmark. The main content area has a brown header titled "Data Terms of Use". Below it, a text box contains the following text:

Artwork, photos, images and text (“data”) stored on **oregonstateparks.org** may be subject to copyright. All use of data from this website must be credited “courtesy Oregon Parks and Recreation Department,” and must be presented as originating from OPRD.

Permission to use the OPRD shield and wordmark (see image) must be obtained separately. E-mail **Jean Thompson**.

By clicking here, **I Agree**

In the bottom right corner of the content area, there is a smaller version of the Oregon State Parks logo and the "Nature HISTORY Discovery" wordmark.

Source:oregonstateparks.org

4. Be aware of rate limits, server loading

Find Sites [x]

Availability

Arrival Date

Departure Date

Not Flexible

Loop

Any Loop

Site #

Attention: Travelers to the central Oregon Coast in the summer of 2016 should be aware that road construction on US20 near Eddyville will result in delays and nighttime closures. Detour routes on to OR34 and OR18 will be posted. Check [Tripcheck.com](#) before travelling in this area.

[Find Other Facilities](#)

CAPE PERPETUA, OR

part of Siuslaw National Forest, US Forest Service



[Like](#) 19

Status: Open through Sun Sep 04 2016 [Season Dates](#)

Source:lotr.wikia.com



The data in Recreation.gov is provided for free - there is no cost to use it, and no need to contact us before incorporating Recreation.gov data into your system. (In exchange, we encourage you to provide a link to Recreation.gov and acknowledge credit, such as *Data Source: Recreation.gov*)

Web data summary

- Highly dynamic!
 - Structure of webpages
 - Access / permissions
 - API versions, output formats, parameters, rate limits
- May not be there tomorrow
- Potential to be blocked for excessive scraping or API abuse

Object Oriented Programming

- Class-based design
- SOLID principles
 - Classes should do one job
 - Instead of modifying existing classes, extend through sub classes
 - Class dependencies should be abstract, not concrete*

```
# Base Data class for extracting information from a datasource
import pandas as pd

class Data():

    def __init__(self, name):
        self.name = name
        self.df = pd.DataFrame()

    def extract(self):
```

* Abstract base classes are not Pythonic, but are used in this tutorial for teaching purposes

Read more about SOLID: <https://scotch.io/bar-talk/s-o-l-i-d-the-first-five-principles-of-object-oriented-design>

How can object oriented design help?

- Encapsulate properties that are likely to change
 - API endpoints, parameters, etc
- Build on abstractions to allow flexibility
 - Consistent interfaces for Data objects, regardless of source
- Extend existing classes to accommodate changes
 - Create WebScraper objects that can be extended
- Create testable, maintainable structures

RIDB API

Home Data Download Data Sharing Mobile App Sign In Register

Recreation Information Database

Search RIDB catalog by state, organization, and activity

Recreation Areas Facilities

Select States ▾

Select Activities ▾

Select Organizations ▾

Search Text

SEARCH

DATA PARTNERS
U.S. Forest Service



The logo of the U.S. Forest Service, featuring a green shield with a yellow border. Inside the shield, the words "FOREST SERVICE" are written along the top curve, "U.S." in large letters in the center, and "DEPARTMENT OF AGRICULTURE" along the bottom curve. A yellow pine tree is positioned in the center of the "U".

ridb.recreation.gov

RIDB API Documentation

<http://usda.github.io/RIDB>

RIDB.RECREATION.GOV

Introduction
About the RIDB
Intended Audience
API Endpoints
Authentication
Export Formats
Get All Rec Area Data
Organizations
Recreation Areas
Facilities
Get All Facilities
Get a Specific Facility
Facility Addresses
Facility Media
Facility Links
Facility Events
Facility Activities
Campsites
Permit Entrances
Tours
Activities
Events
Media
Links

Get All Facilities

This endpoint retrieves all facilities.

HTTP REQUEST

```
GET https://ridb.recreation.gov/api/v1/facilities
```

URL PARAMETERS

Parameter	Type	Required	Description
query	string	optional	query filter criteria. Searches on facility name, description, keywords, and stay limit
limit	numeric	optional	number of records to return (max 50)
offset	numeric	optional	start record of overall result set
full	string	optional	return the full record details or compact (abbreviated) details
latitude	numeric	optional	Latitude of the point in decimal degrees
longitude	numeric	optional	Longitude of the point in decimal degrees
radius	numeric	optional	Distance (in miles) by which to include search results
state	string	optional	comma delimited list of 2 character state codes
activity	string	optional	comma delimited list of activity IDs (see activities)
lastupdated	date	optional	return all records modified since this date

Example Responses

Example facility JSON response

```
{  
    "FacilityEmail": "Library.Bush@nara.gov",  
    "FacilityLongitude": -96.331389,  
    "LastUpdatedBy": 1008,  
    "FacilityDescription": "Like the other Presidential Libraries, the George Bush  
Presidential Library and Museum is also a research institution, totally integrated into  
the academic environment of Texas A&M University.<p>\r\nThe Bush Library's collections  
include 38,000,000 pages of official and personal papers, 1,000,000 photographs, 2,500  
hours of videotape, and 70,000 museum objects. These rich primary sources document George  
Bush's distinguished public career as congressman, Ambassador to the United Nations,  
Chief of the U.S. Liaison Office in China, Chairman of the Republican National Committee,  
Director of the Central Intelligence Agency, Vice-President, and President. Included in  
the Museum's exhibits are items ranging from a 1925 film of George Bush's first steps in  
Kennebunkport, Maine, to records and memorabilia from his tenure as the 41st President of  
the United States. <p>\r\nThe Museum also contains a special section is dedicated to  
former First Lady Barbara Bush and a classroom designed specifically for students from  
kindergarten through high school. <p>\r\nOperated by the National Archives and Records  
Administration (NARA), the George Bush Presidential Library and Museum is the tenth  
Presidential Library in the United States. The library is located on the Texas A&M  
University campus in College Station, Texas. \r\n",  
    "FacilityLatitude": 30.612222,  
    "FacilityTypeDescription": "Library",  
    "FacilityPhone": "979-691-4000",  
    "FacilityMapURL": "http://bushlibrary.tamu.edu/map.html",  
    "FacilityReservationURL": "",  
    "FacilityDirections": "See the map at <a href=\"http://bushlibrary.tamu.edu/map.html\">bushlibrary.tamu.edu/map.html</a>\r\n

\r\nFrom Houston: Take I-45 north to Conroe. In Conroe, take Hwy 105 west to


```

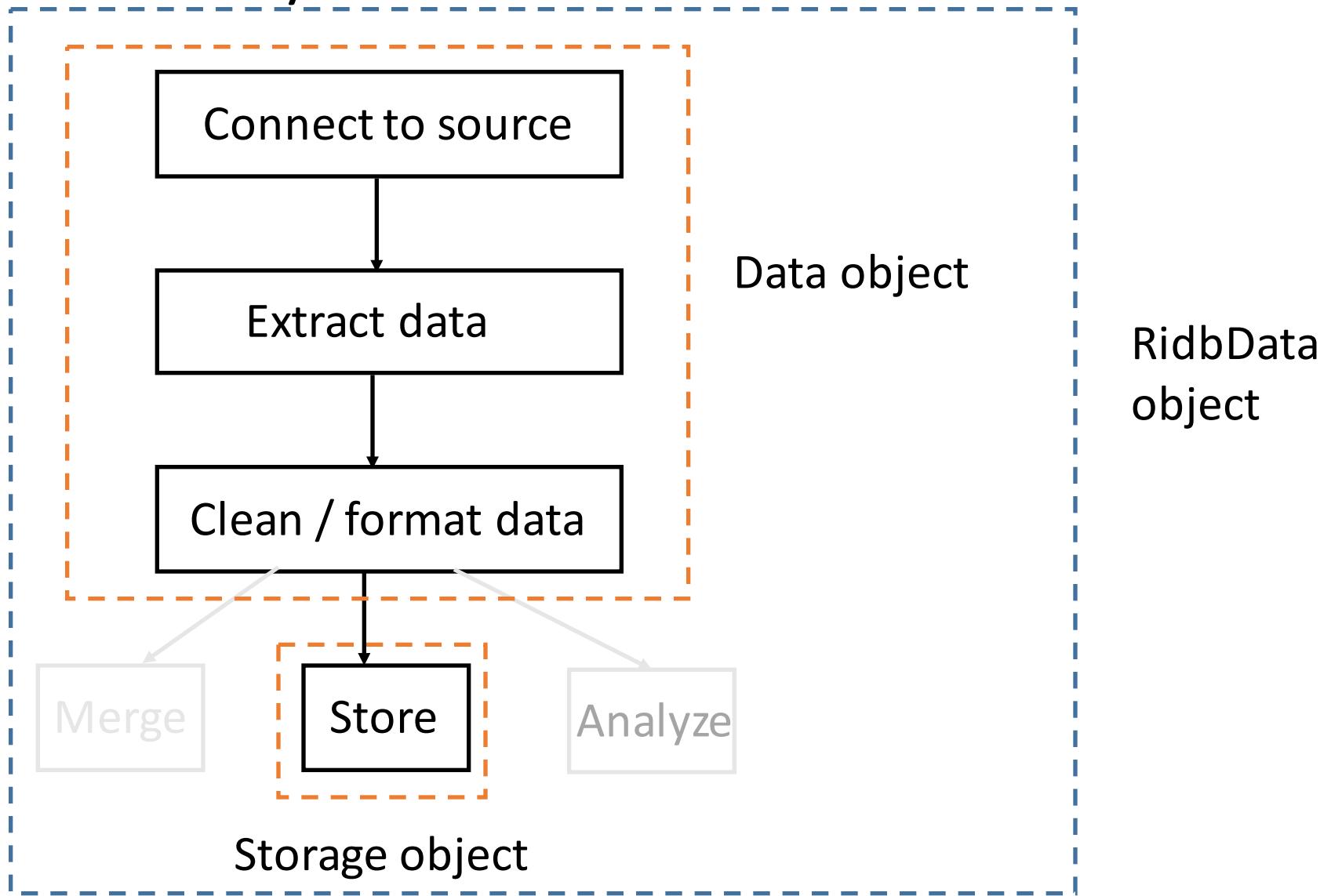
RIDB data notes

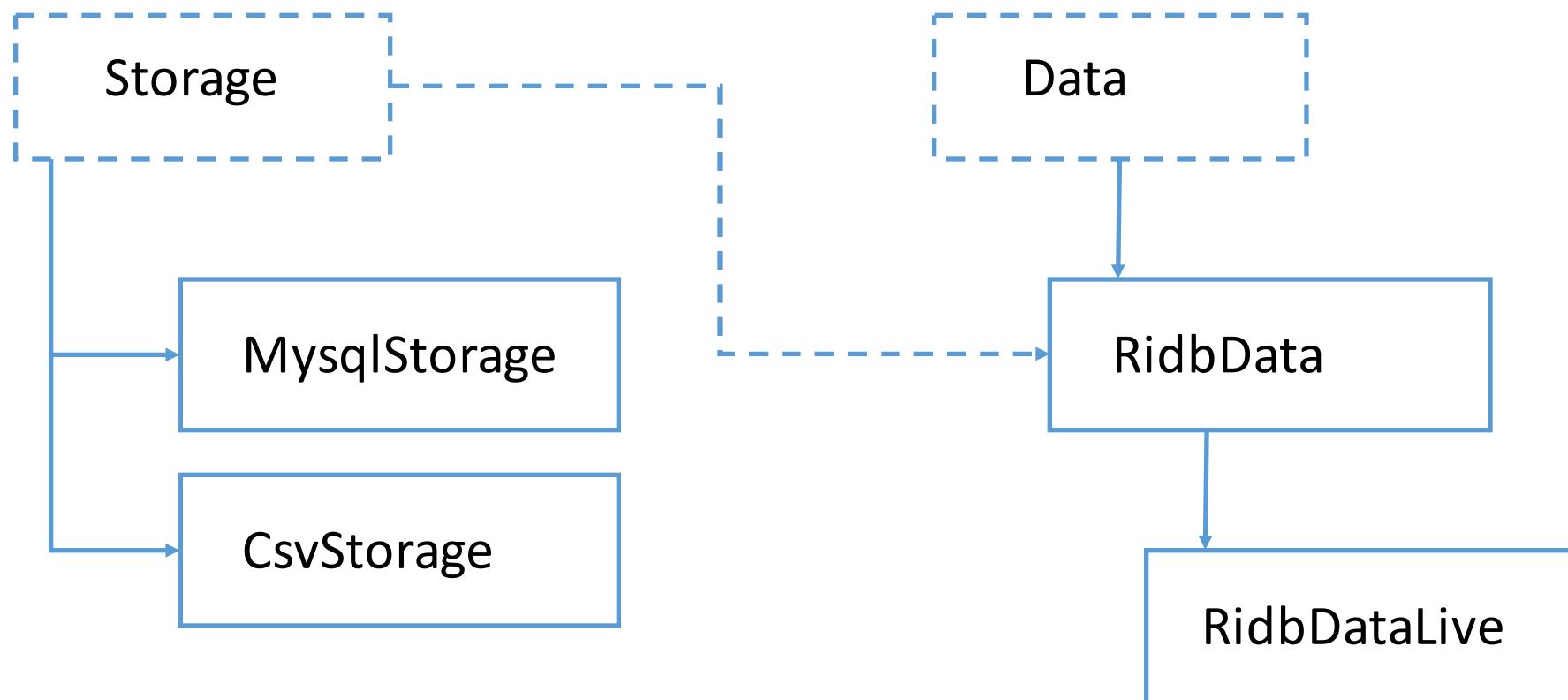


- Only has campgrounds that can be reserved – no first come, first serve
- Missing reservation URLs
- Lat / Longs not guaranteed to be accurate

Lab 1 & 2

Lab 1 & 2 summary





USFS Website scraping



USFS Data

- No API
- Links are discovered through scraping (already done)
- Pages are fairly uniform!

[Site Map](#)**Gifford Pinchot National Forest**[Home](#)[Special Places](#)**Recreation**

- [Bicycling](#)
- [Camping & Cabins](#)
- [Climbing](#)
- [Fishing](#)
- [Hiking](#)
- [Horse Riding & Camping](#)
- [Nature Viewing](#)
- [OHV Riding & Camping](#)
- [Outdoor Learning](#)
- [Picnicking](#)
- [Water Activities](#)
- [Winter Sports](#)
- [Other Activities](#)

- [Alerts & Notices](#)
- [Passes & Permits](#)
- [Maps & Publications](#)

Land & Resources Management**Learning Center****Working Together****About the Forest**

Campground: Cultus Creek

Area Status: Closed

Fire Danger Level and Information

Related Information[Campground: Cultus Creek](#)[News & Events](#)**Alerts & Warnings**

Cowlitz Valley Flood Damage and Road Repair (last updated on 4/14/2016)

- [Help Keep Caves Open](#)
- [Spring Hours](#)
- [Climbers Bivouac Closed](#)

[View All Forest Alerts](#)**Quick Links**

- [Online Washington State Accessible Outdoor Recreation Guide](#)

Areas & Activities[Find An Area](#)**Location**

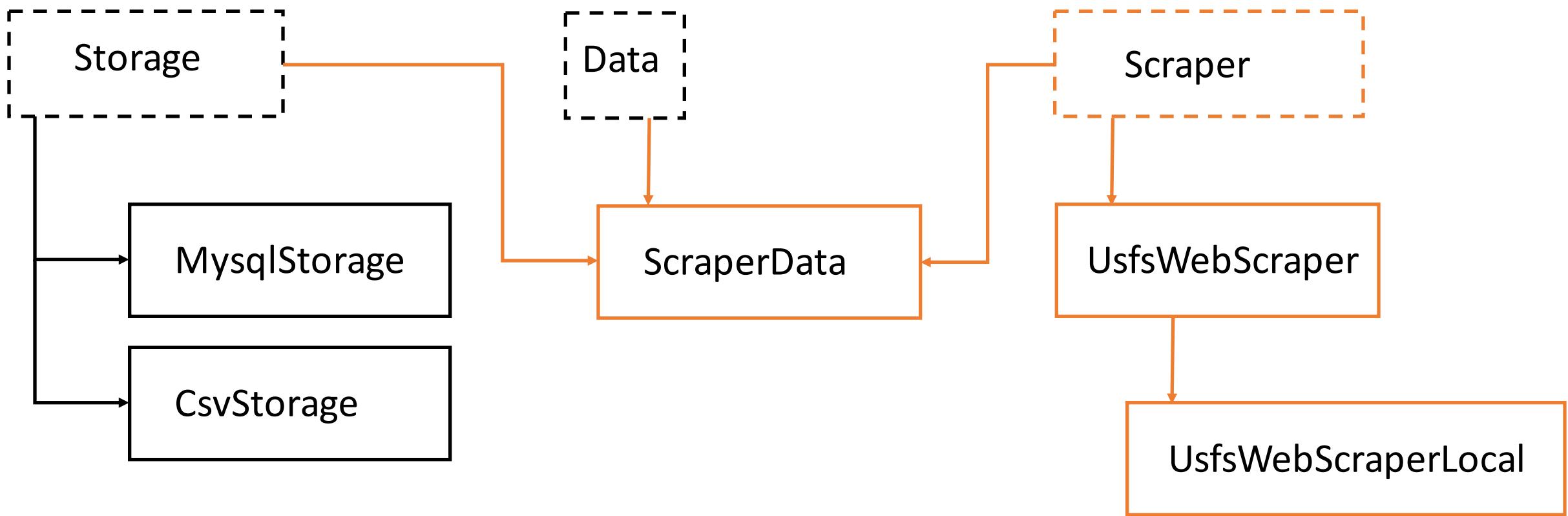
Latitude : 46.0476
Longitude : -121.7546
Elevation : 3996

At a Glance

Current Conditions:	4/28/2016: Not accessible yet for season.
Operational Hours:	Open after snow melt, historically by mid-June. Call Mt. Adams Ranger Station at 509-395-3402 for opening date.
Fees	<ul style="list-style-type: none"> ▶ Camping: \$10/night/single unit; \$20/night/double unit; \$5/extravehicle fee ▶ Day use: \$5/vehicle/day
Permit Info:	Wilderness permits are required when entering Wilderness. Permits are free and self-issued at trailheads.
Usage:	Light
Restrictions:	Maximum vehicle length is 32 feet.
Closest Towns:	Trout Lake, WA
Water:	No
Restroom:	Vault toilet
Passes:	<ul style="list-style-type: none"> ▶ Camping: 50% discount for single site camping with any of these passes: Interagency Senior, Interagency Access, Golden Age, or Golden Access. ▶ Day use: A valid Recreation Pass is required at Cultus Creek/Indian Heaven trailheads at Cultus Creek Campground.

Lab 3 & 4

Saving tutorials with OOD!



Recreation.gov reservation scraping

Find Sites [x]

Availability

Arrival Date

Departure Date

Loop

Site #

Attention: Travelers to the central Oregon Coast in the summer of 2016 should be aware that road construction on US20 near Eddyville will result in delays and nighttime closures. Detour routes on to OR34 and OR18 will be posted. Check [Tripcheck.com](#) before travelling in this area.

[Find Other Facilities](#)

CAPE PERPETUA, OR

part of Siuslaw National Forest, *US Forest Service*

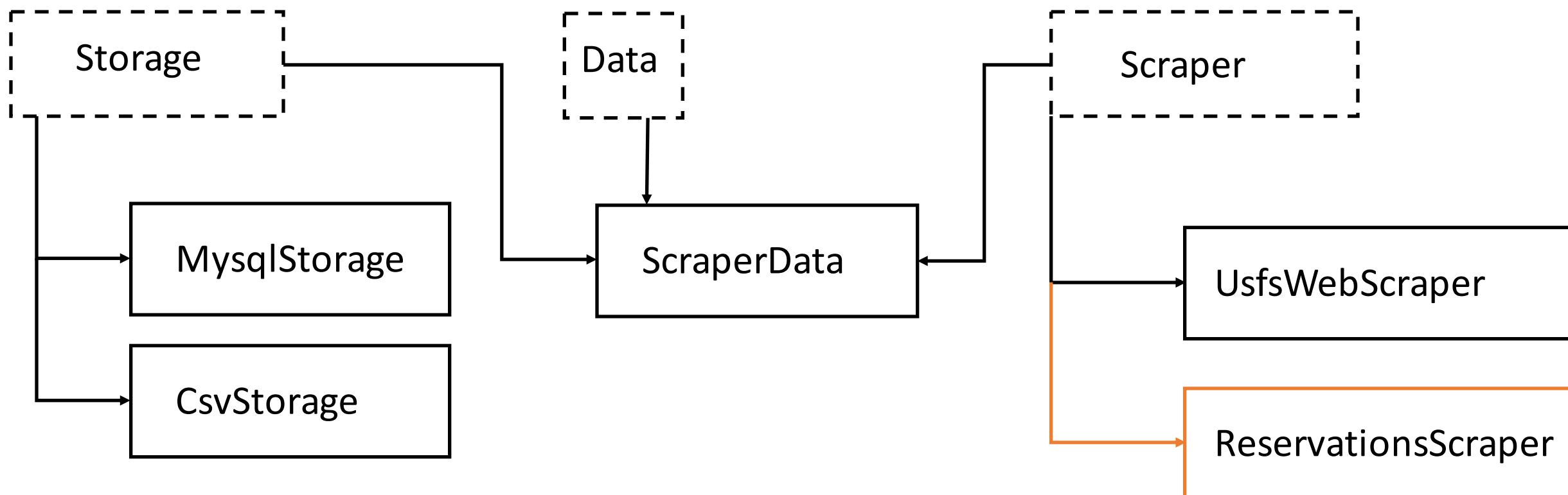


[!\[\]\(fac3df7ac48ec2707f62dd77e89888f8_img.jpg\) Like 19](#)

Status: Open through Sun Sep 04 2016 [Season Dates](#)

Lab 5 & 6

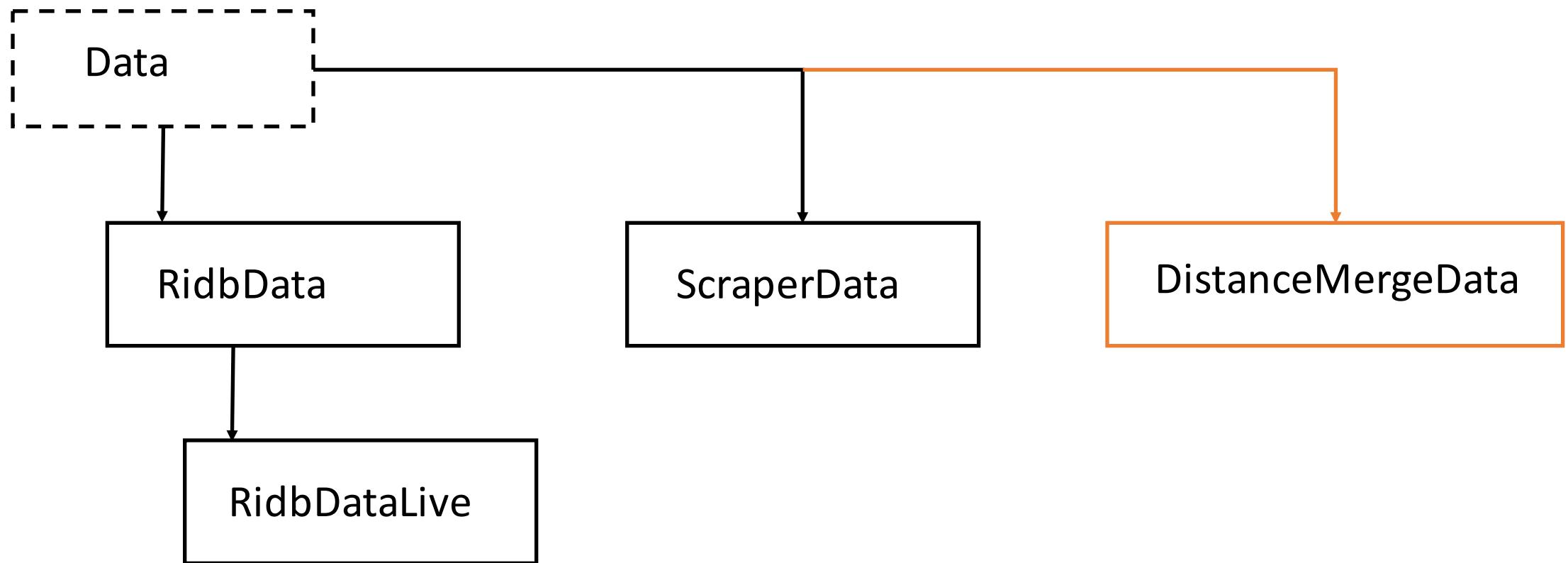
Lab 5 & 6 objects



Merging all the data – Lab 7

- Data merge is based on Lat/Long
 - But remember, lat/longs are frequently not correct!
 - Some error
 - Merge / join will be based on the distance between campgrounds being below a certain threshold

Lab 7 objects



Lab 8 - The One Pipeline

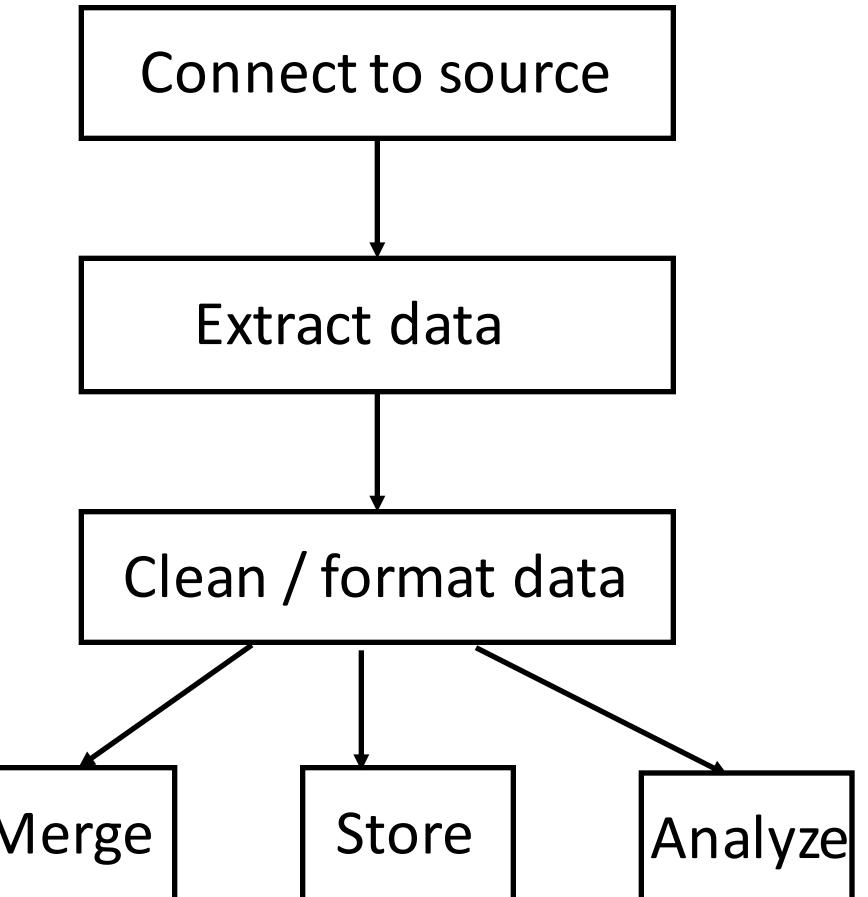
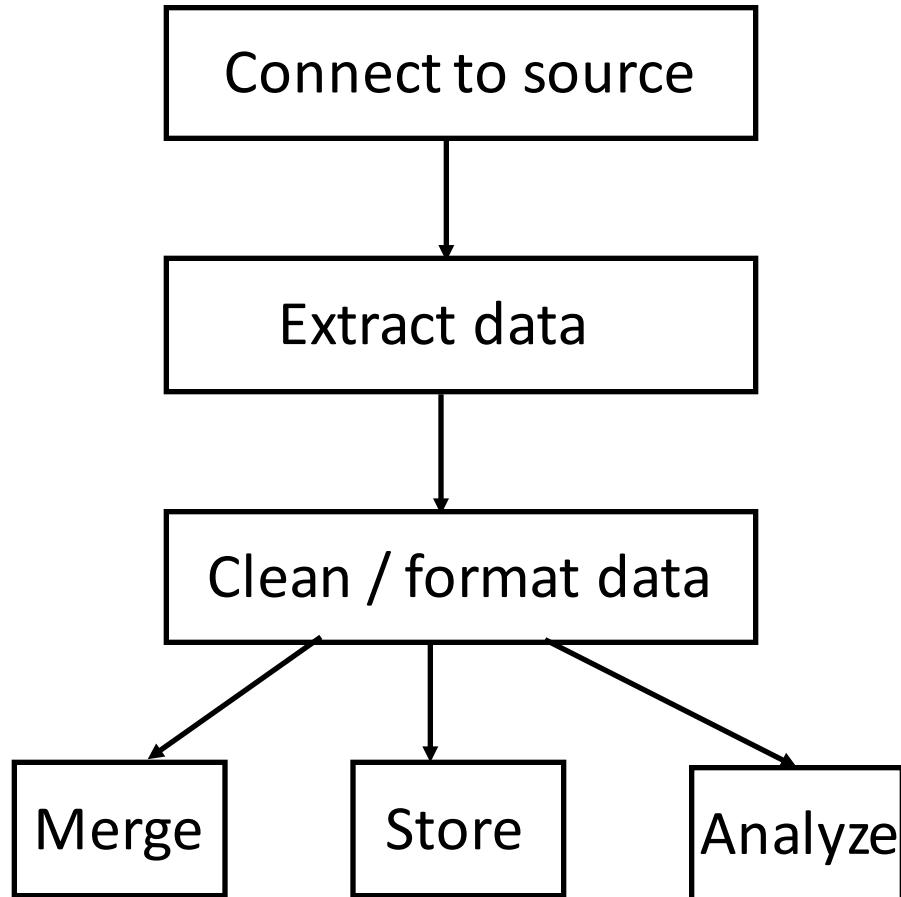
Thanks!



Image source: lotr.wikia.com

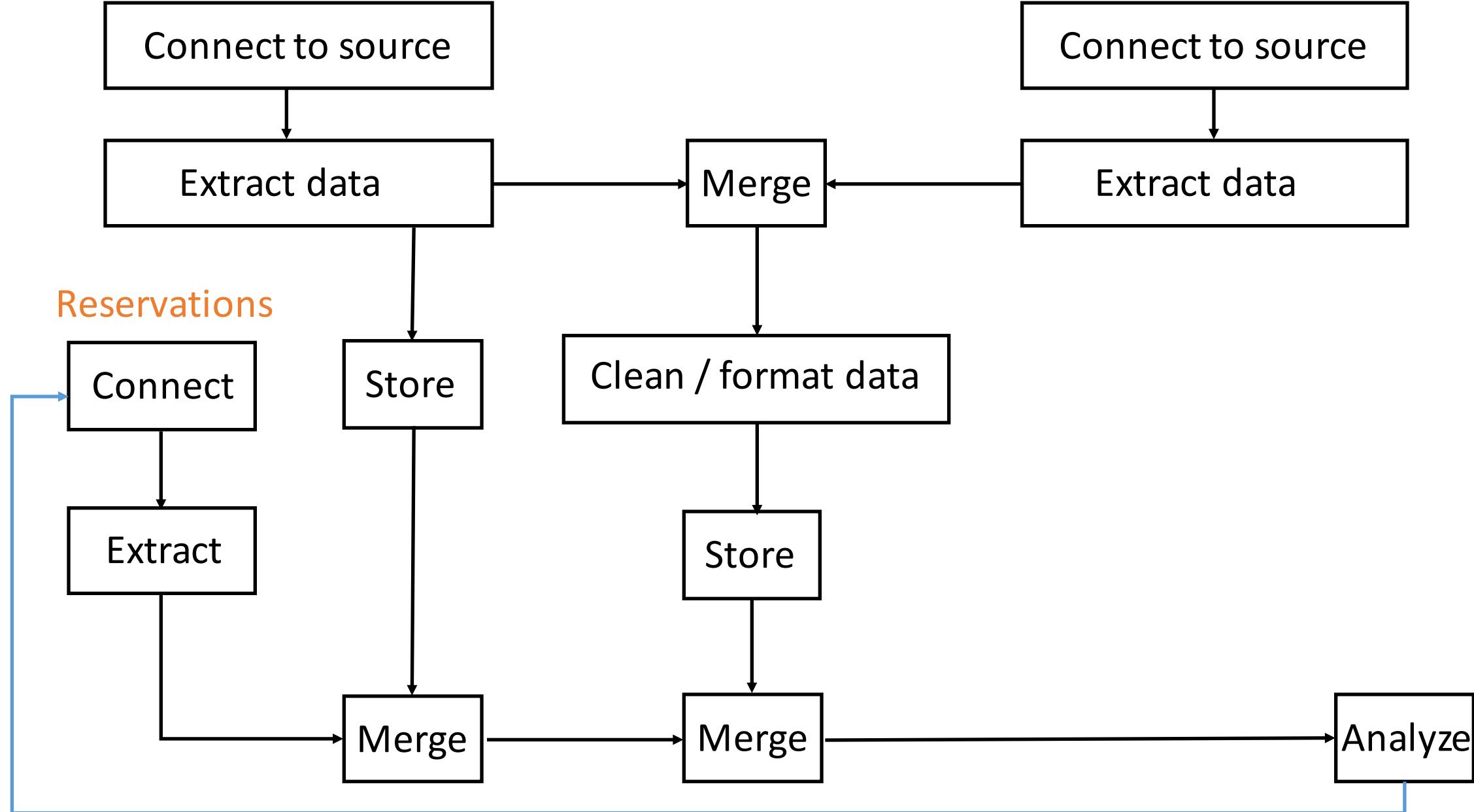
Backup

The ideal data pipeline



RIDB API

NF websites



Look across

