



AGTC
genetic tax. consultants

Orchard: Branching and State Management for Big Data Pipelines

Peter Bellagh, Christopher Blazer, Christian Buskirk, Jorden Kreps, Curtis Rose

CS Faculty Mentor : Dr. Viacheslav Fofanov, SICCS

Client and Mentor:

Dr. Viacheslav Fofanov

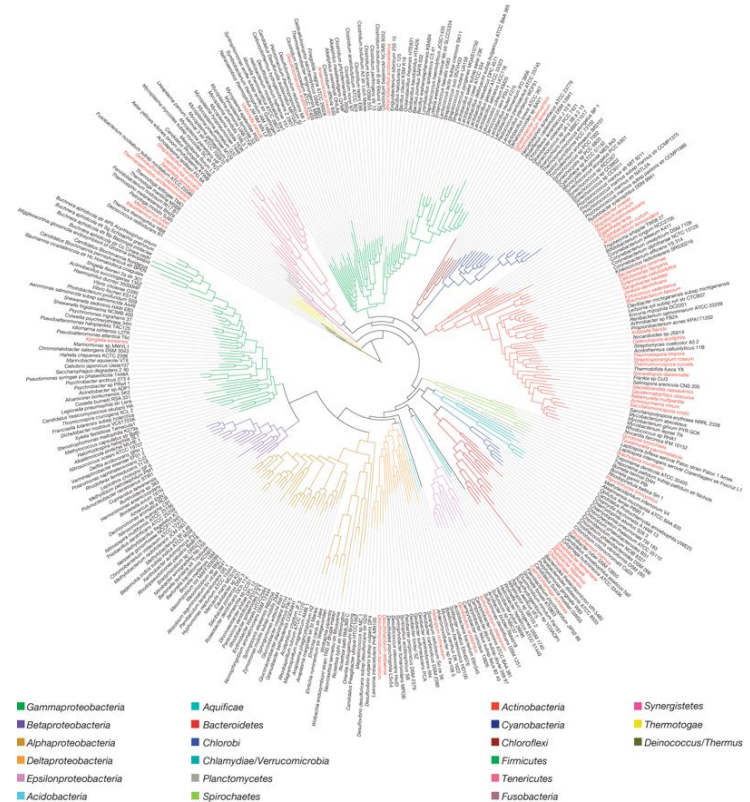
Assistant Professor at NAU School of Informatics, Computing, and Cyber Systems

Dr. Fofanov's research focuses on applications of High Throughput Sequencing data to pathogen detection in complex clinical and environmental samples.



An Introduction: Large Data Sets and Bioinformatics

Group	Total # of species	Total # of unique identifiers	Amount of sequence data
Plants and Fungi	261,135	54,019,636	119.9 GB
Vertebrates	67,311	90,338,474	542.8 GB
Eukaryotes	310,220	16,455,537	76.8 GB
Viruses	114,766	1,528,683	2.3 GB
Bacteria	281,314	3,848,470	80.0 GB
Other	47,338	5,420,682	5.7 GB



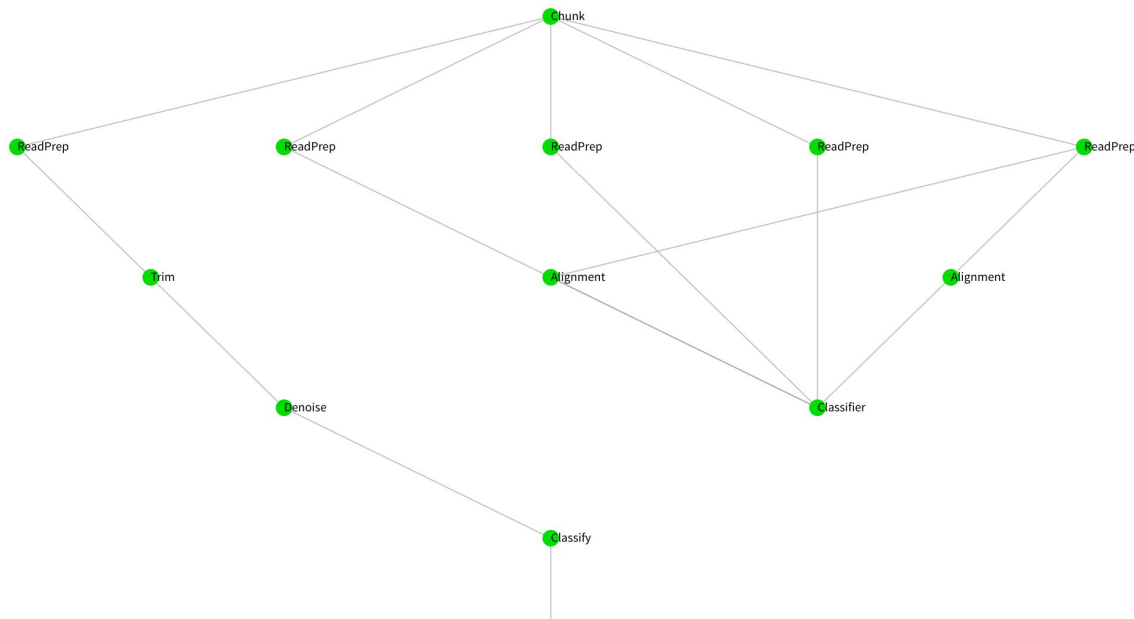
The Problem

Resource Intensive

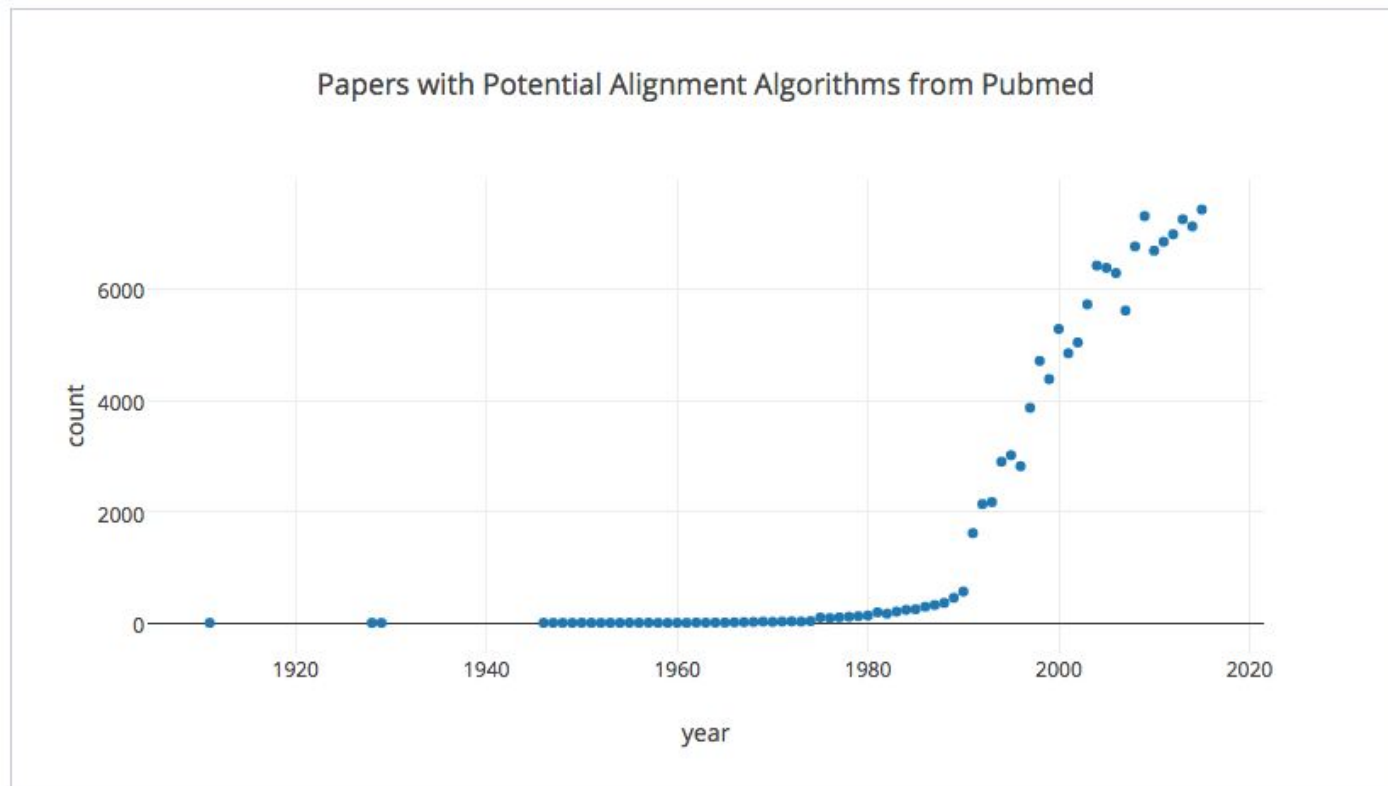
Complex Module Structure

Reuse

Branching



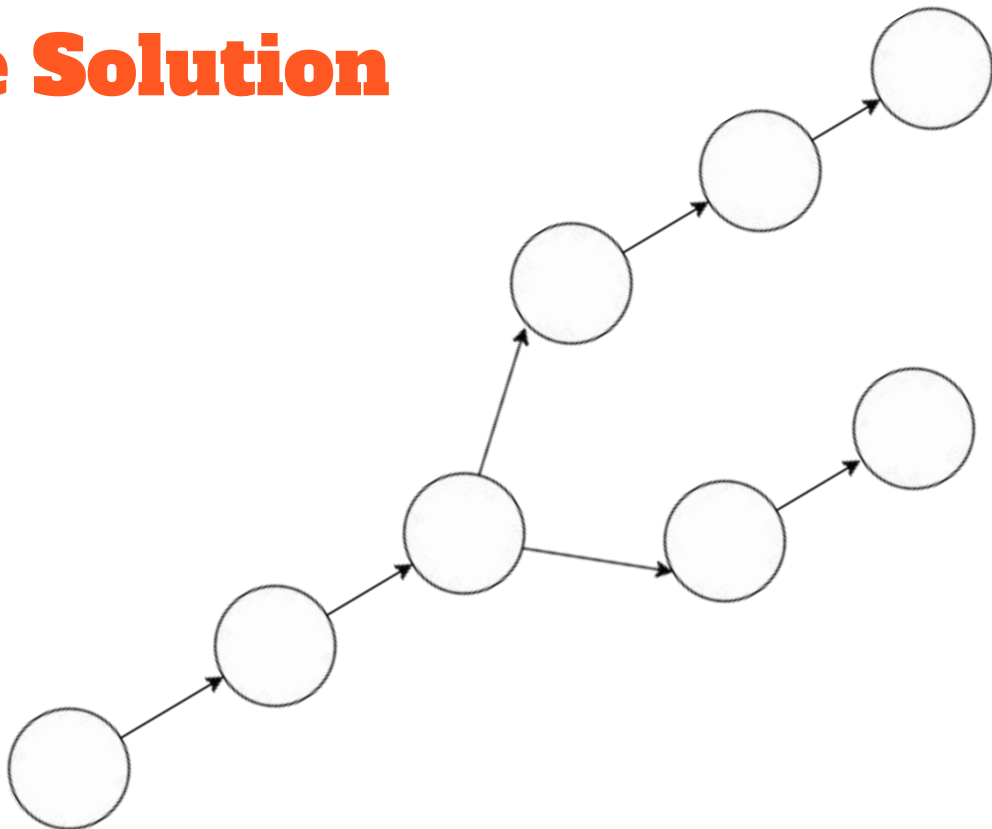
Future Maintainability



Luigi: The Base Solution

Project Requirements

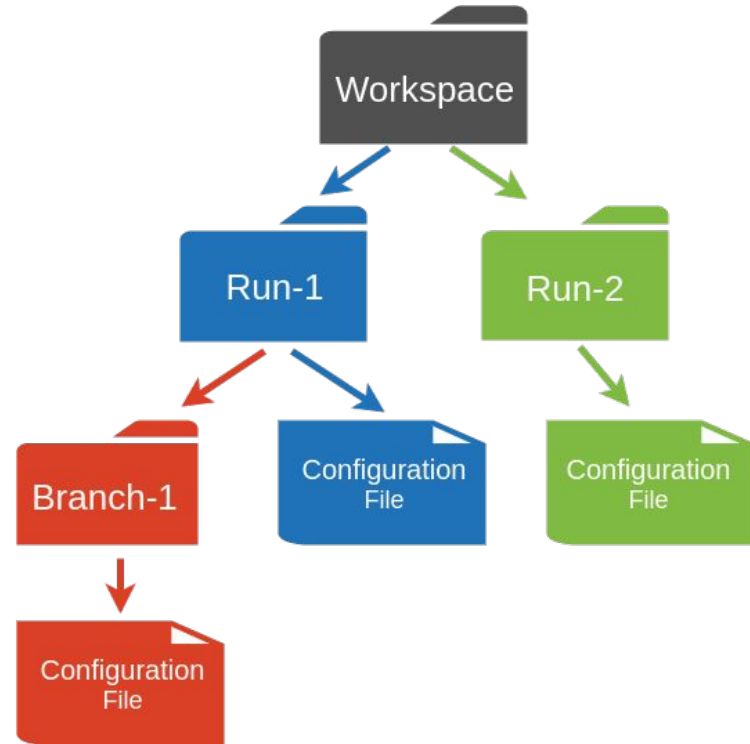
- Visualization
- Modularity
- Branching
- State Saving
- Easy Configuration
- Future Maintainability
- Rollout



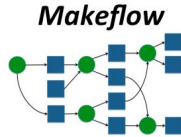
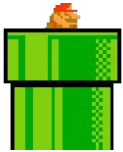
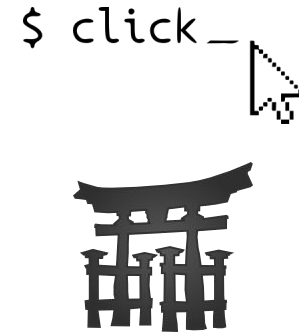
Orchard: The Expanded Solution

Project Requirements

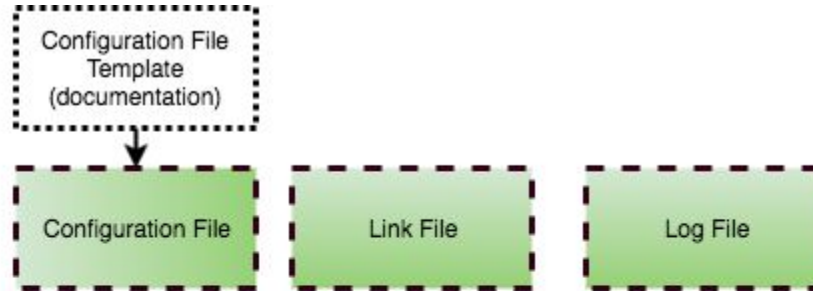
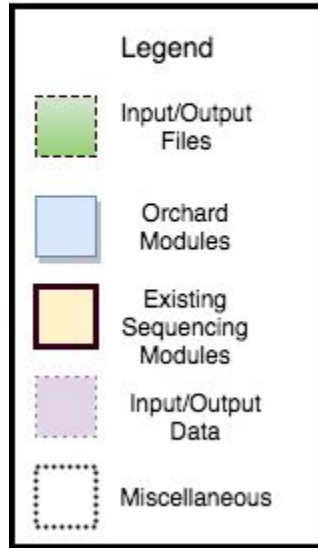
- Visualization
- Modularity
- Branching
- State Saving
- Easy Configuration
- Future Maintainability
- Rollout



Implementation Overview



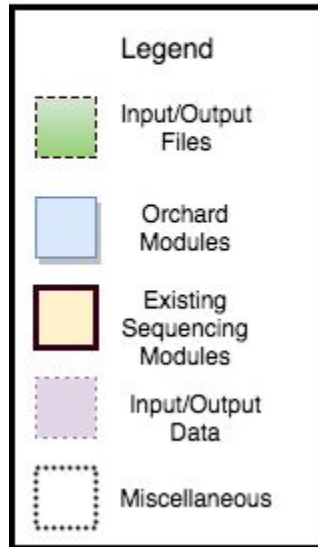
Architecture Overview



Link File

```
orchard — vim orchard/data/link.yml — 80x40
1  modules:
2  - name: ModuleOne
3    arguments:
4      - name: infile
5        command: --out
6        isBranch: false
7      - name: digit
8        command: -d
9  - name: ModuleTwo
10   dependencies:
11     - ModuleOne
12   arguments:
13     - name: infile
14     - name: outfile
15       command: --out
16       isBranch: false
17     - name: digit
18       command: -d
19   optionals:
20     - name: forward
21       command: --forward
22       isFlag: true
23     - name: reverse
24       command: --reverse
25       isFlag: true
26   exclusive:
27     - forward
28     - reverse
29 - name: ModuleThree
30   dependencies:
31     - ModuleTwo
32   arguments:
33     - name: infile
34     - name: outfile
35       command: --out
36       isBranch: false
37     - name: digit
38       command: -d
:
```

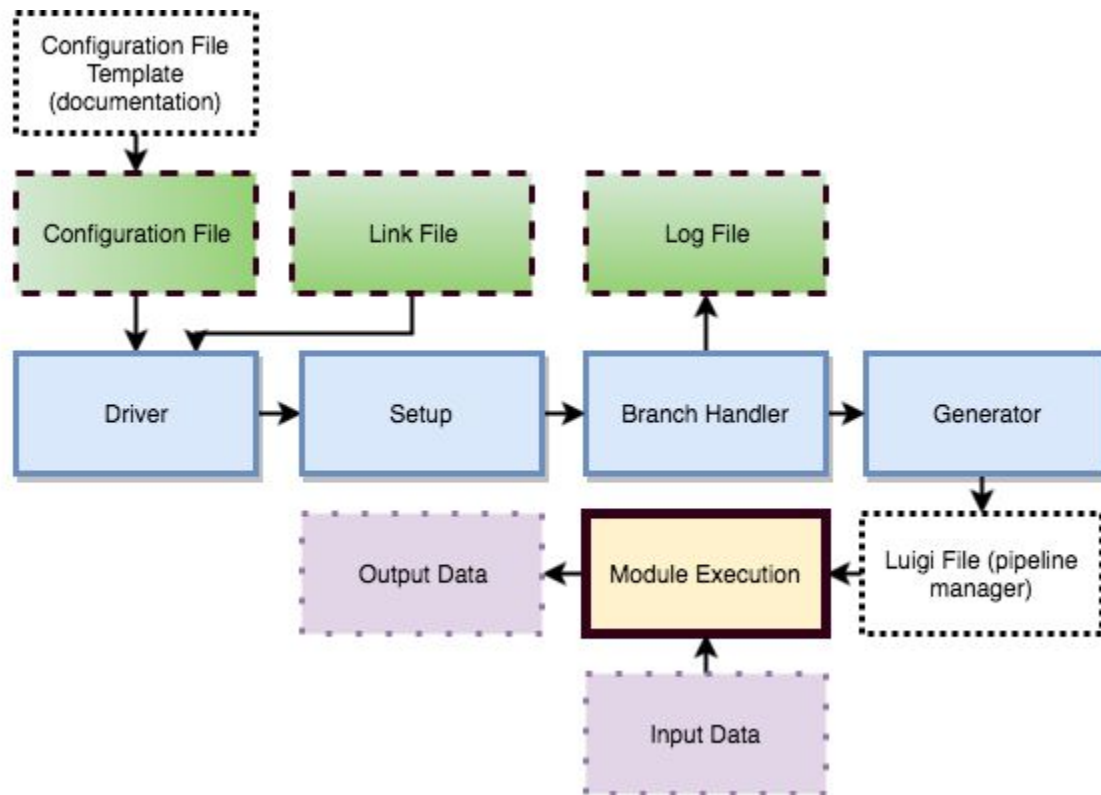
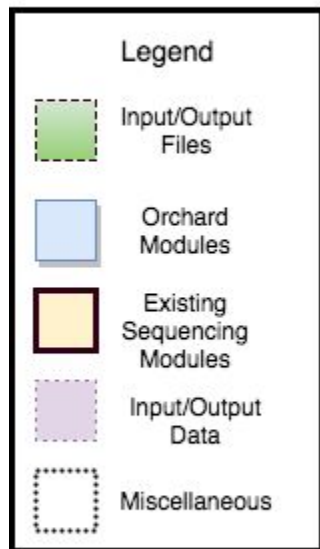
Architecture Overview



Configuration File

```
orchard — vim orchard/data/config.yml — 80x24
1  modules:
2    - name: ModuleOne
3      arguments:
4        - infile:
5          - outfile:
6            digit:
7        - name: ModuleTwo
8          arguments:
9            - infile:
10             - outfile:
11             digit:
12   optionals:
13     - forward:
14       - reverse:
15   - name: ModuleThree
16     arguments:
17       - infile:
18       - outfile:
19       - digit:
20
21
22
23
24
25
```

Architecture Overview



Working Prototype

Prototype Continued

Coding Challenges and Resolutions

Challenge 1: Abstraction

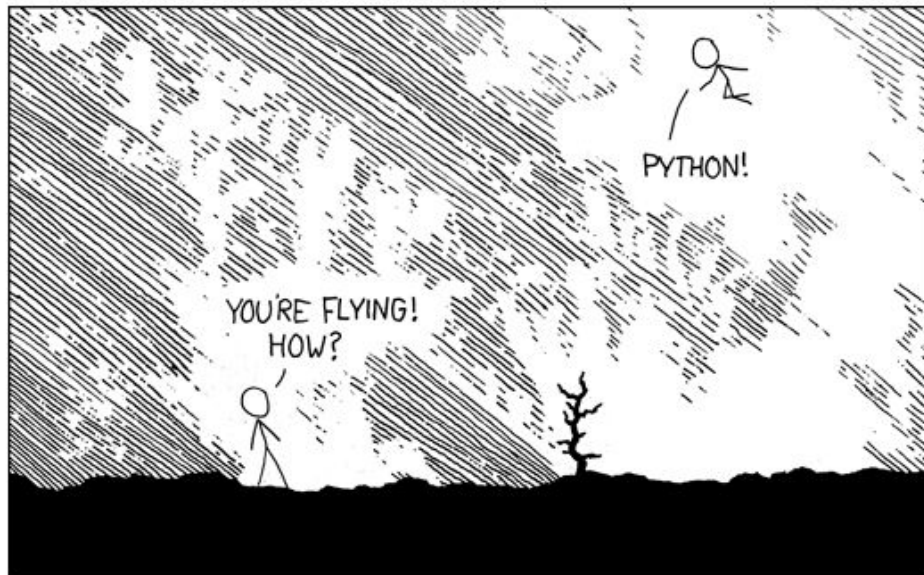
Resolution Plan:

- Configuration files
- Internal manipulation of complex data structures

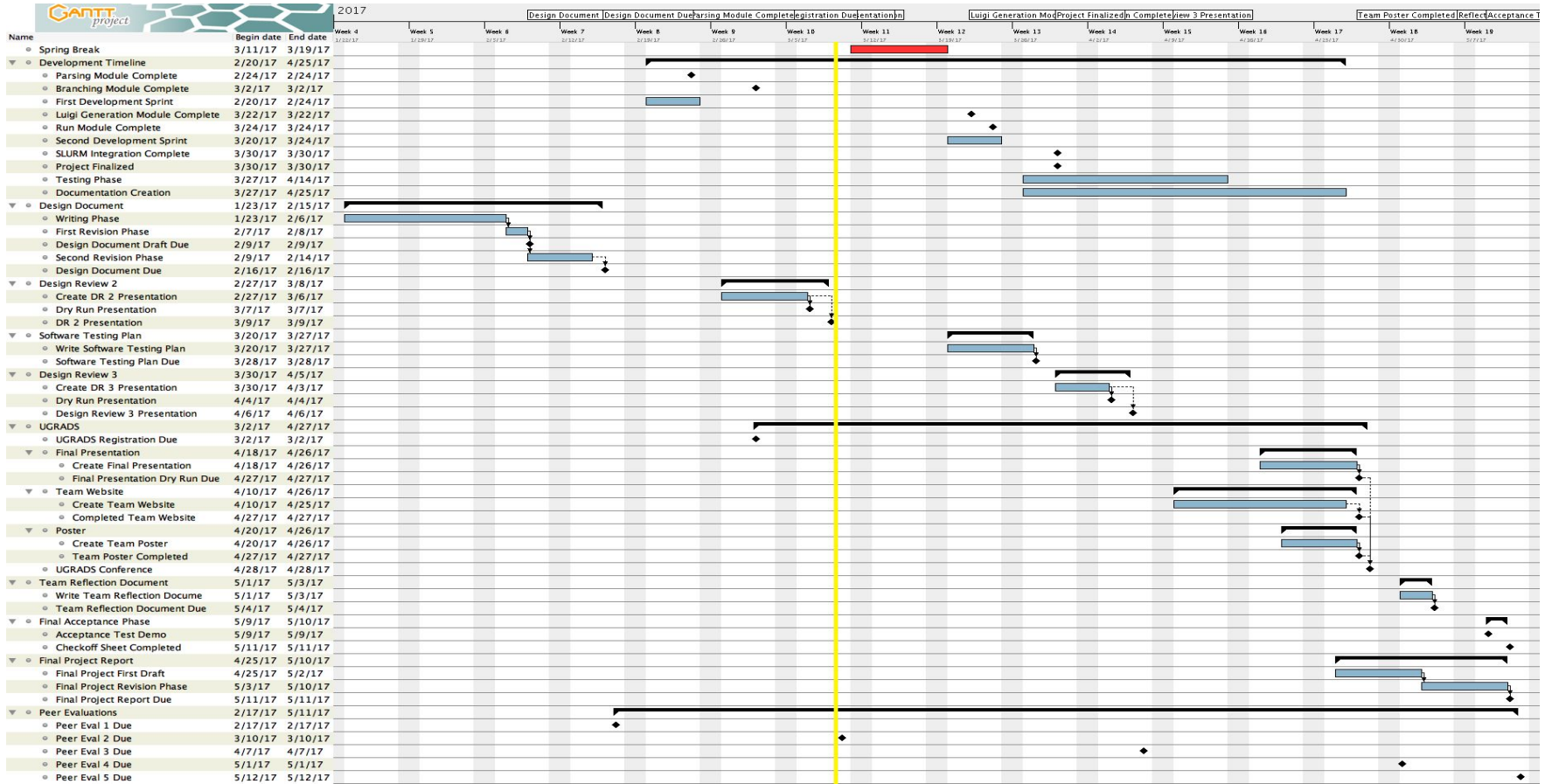
Challenge 2: Pipeline Complexity

Resolution Plan:

- Link files
- Mapping and validation of user configuration files to runnable processes



Schedule: Where are we now?

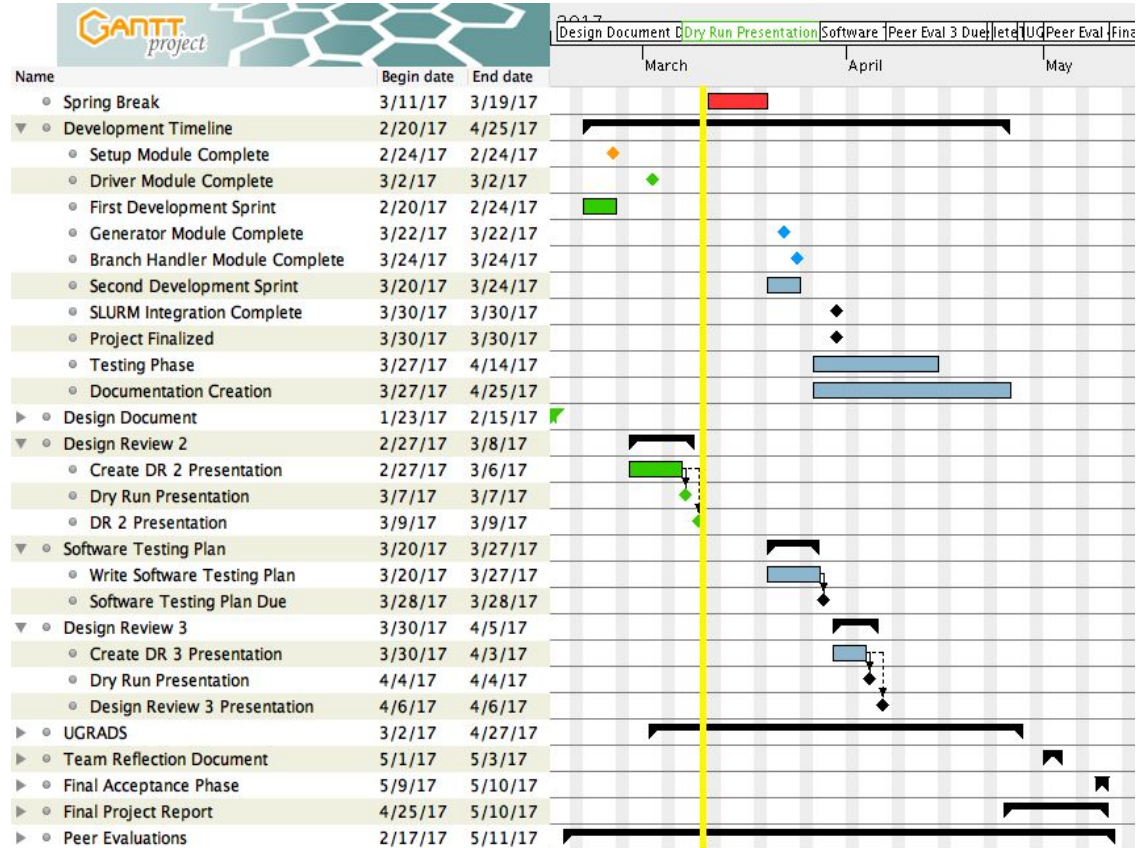


The Current Phase

Currently close to our original schedule

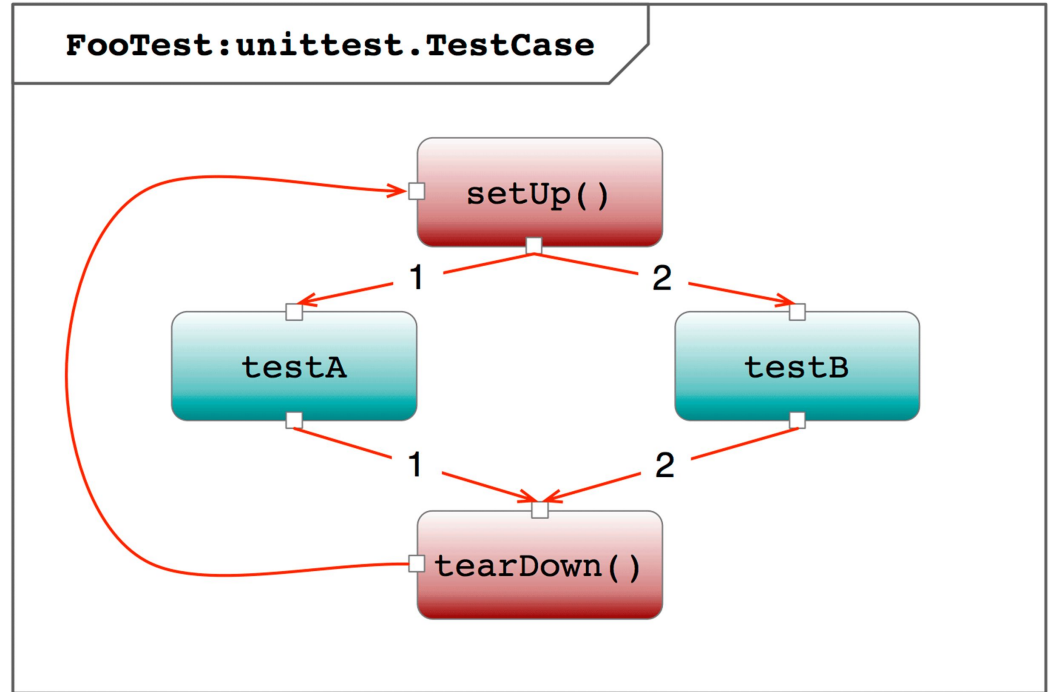
Setup module still in progress

First steps for the Generation and Branch Handler modules already in place



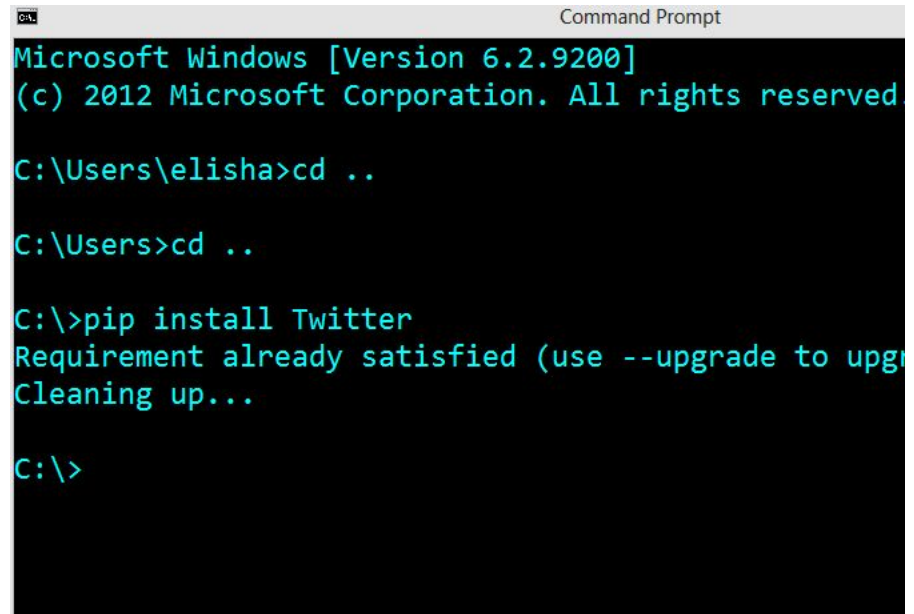
Unit Testing

- Python unittest
- Currently at 75% code coverage
- Test Driven Design



Usability Testing: Ease of Rollout

- Use of PIP to install Orchard Packages
- Internal Testing on recent Ubuntu and CentOS versions



```
Command Prompt
Microsoft Windows [Version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\elisha>cd ..

C:\Users>cd ..

C:\>pip install Twitter
Requirement already satisfied (use --upgrade to upgrade)
Cleaning up...

C:\>
```

Usability Testing: User Testing

Observations and Testing
of Three Major Groups:

- Low Technical Skill
- Medium Technical Skill
- High Technical Skill



In Conclusion: The Big Picture

Modern pipelines are responsible for handling tremendous amounts of data

This requires them to reuse processed data wherever possible

Advances in the field also require fast development cycles of internal modules

Basic solutions such as Luigi do not cover all of these aspects

With Orchard these missing features will be addressed and covered

Questions?



ORCHARD