Datatrack:
An R package for managing data in an experimental workflow

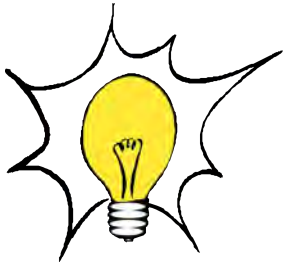
Data versioning and provenance considerations
In interactive scripting

Philip Eichinski, Paul Roe
Queensland University of Technology,
Brisbane, Australia

Overview

- Datatrack R package allows easy record-keeping of provenance metadata within the R scripting environment during small-scale exploratory development.
- Simple integration requires minimal learning or modifications of coding style
- Allows visual exploration of provenance metadata within R studio to assist choosing input during interactive scripting

scientific question



idea

coding
testing
small data

Automation
Distribution
etc

Pegasus

Kepler

VisTrails

coding
testing
small data

- Loss of REPL interactivity
- Learning new software
- Learning new language (workflow language)
- Many unneeded features
- Switching between environments

SWfMS

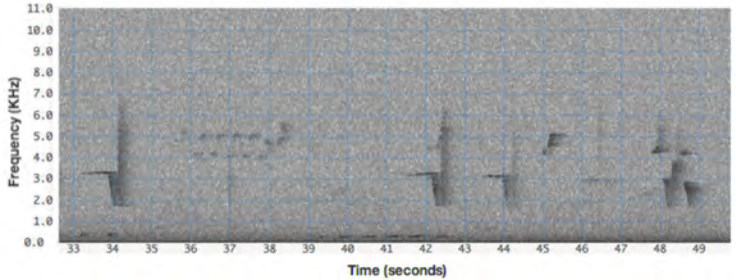
Pegasus

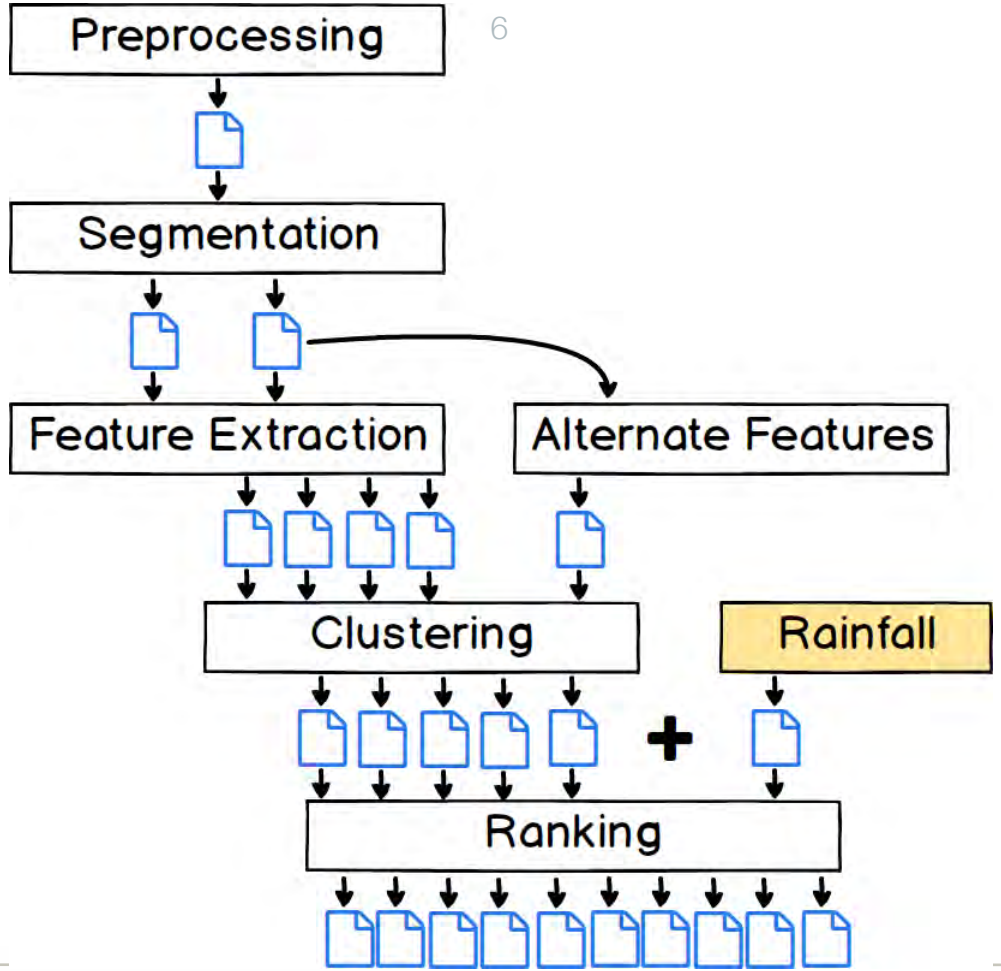
Kepler

VisTrails


Boo!


database


Frequency (KHz)
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
Time (seconds)


Preprocessing
Segmentation
Feature Extraction
Alternate Features
Clustering
Rainfall
Ranking

Data Provenance

- Information about data required to reproduce it

- Necessary for selecting the desired inputs to a step of a workflow when run in isolation.

Data Provenance
for decision-making in interactive scripting

- Which parameters were used to produce the data?
- Which other data was used as input to produce the data (and their parameters): ***data dependencies*?**

Data Provenance
for decision-making in interactive scripting

Recorded by Datatrack via
wrappers for read and write functions.

Writing Data

- Ability to write data along with provenance metadata

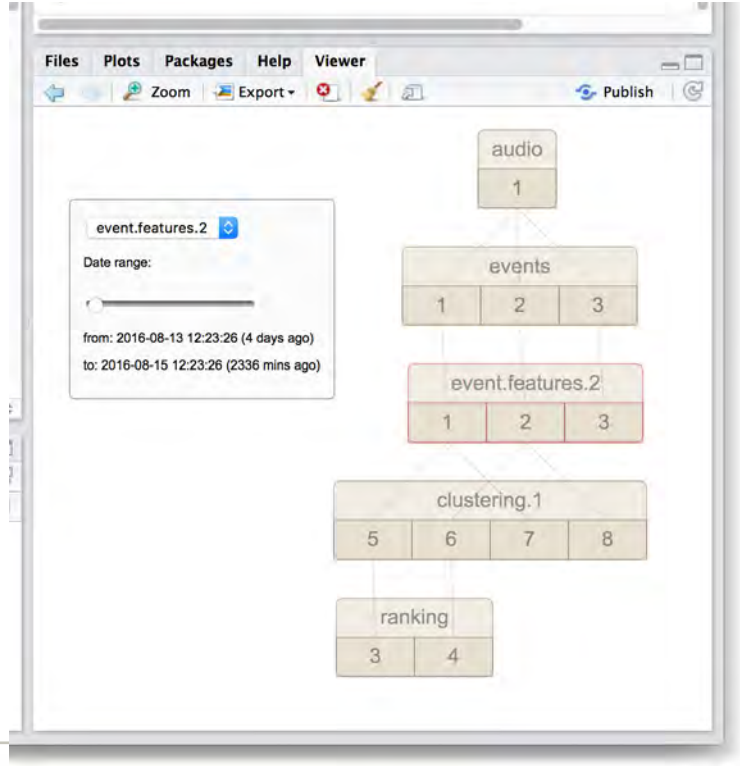
```
writeDataobject(mydata,
                name = ‘my.data.output’,
                ... additional metadata as parameters ...
```

- Which parameters were used when generating the data
- Which other data objects that were used when generating the data

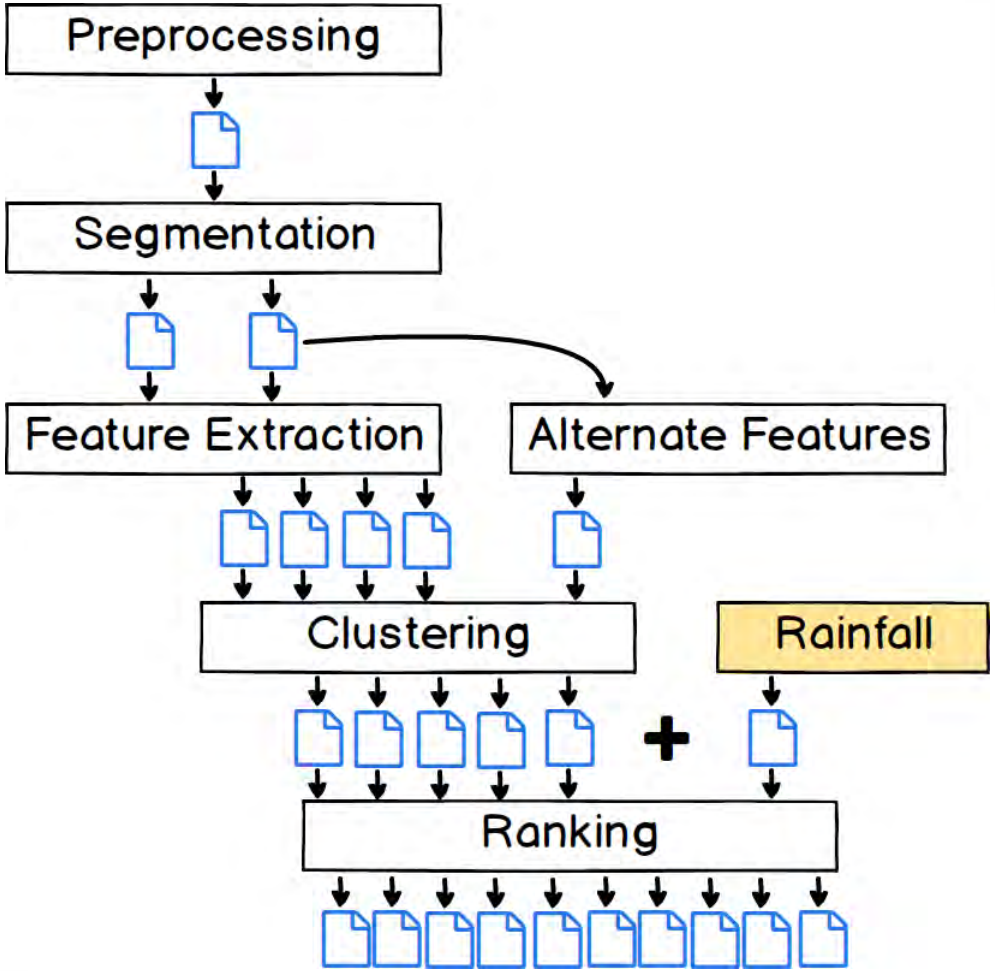
Reading Data

- Ability to view the dependency graph of existing data to assist selection when reading data

```
readDataobject(
        'event.features.2')
```



Demo


Preprocessing
Segmentation
Feature Extraction
Alternate Features
Clustering
Rainfall
Ranking

Considerations

- Tracking of users: the “who” of provenance
- Tracking of code versions and environment information
- Generating versions and overwriting data
- Cyclic data dependencies

Summary

- Datatrack R package allows easy record-keeping of provenance metadata within the R scripting environment during small-scale exploratory development.
- Simple integration requires minimal learning or modifications of coding style
- Allows visual exploration of provenance metadata within R studio to assist choosing input during interactive scripting

Thank You

philip.eichinski@qut.edu.au

https://github.com/peichins/datatrack

Implementation

- Metadata stored as a single csv
- Dependency graph visualization written in javascript using D3.js
- Inserted into R Studio viewer using Html Widgets package.