S-MAP Workflows on and off the Cluster

Robbie Price

Pascal Omondiagbe

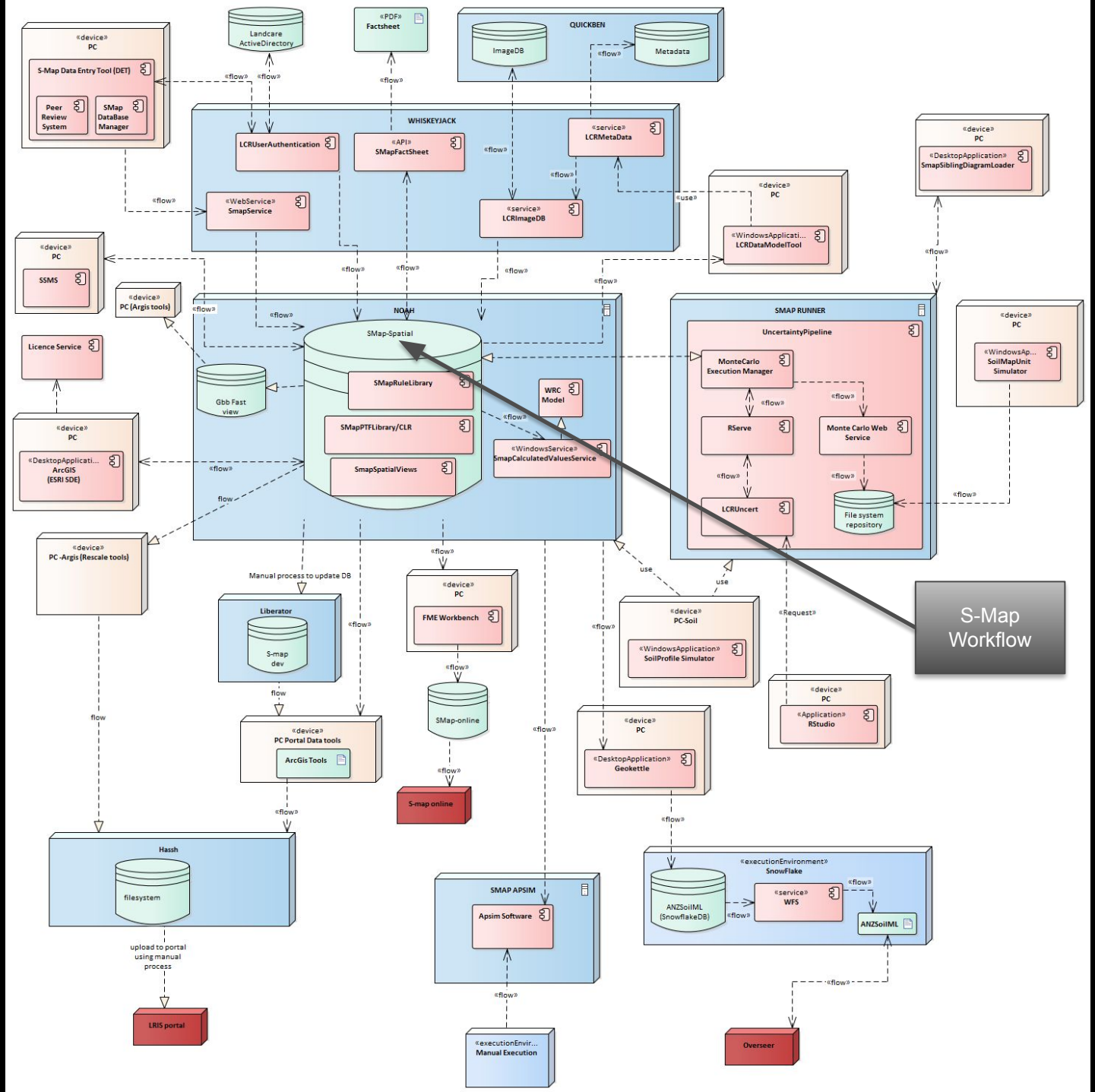
Presenting an overview of technical work to support scientists and data managers undertake repeatable spatial soils research

OutLine

- **Intro**
  - Robbie and Pascal, Staff Diagram
- **SMAP quick overview**
  - General overview
- **Digital Soil Mapping**
  - Method Overview
  - DSM Process
  - S-map Ecosystem
  - Covariate Process
  - DSM modelling Process
  - Segmentation Process
  - Post processing/Upload
- **Workflow Development**
  - Requirement
  - Retrospective
  - Workflow Principle
  - Development Process
  - DSM Workflow
  - Performance/Re-evaluate
  - S-Map Workflow Summary
- **Conclusion**

SMAP

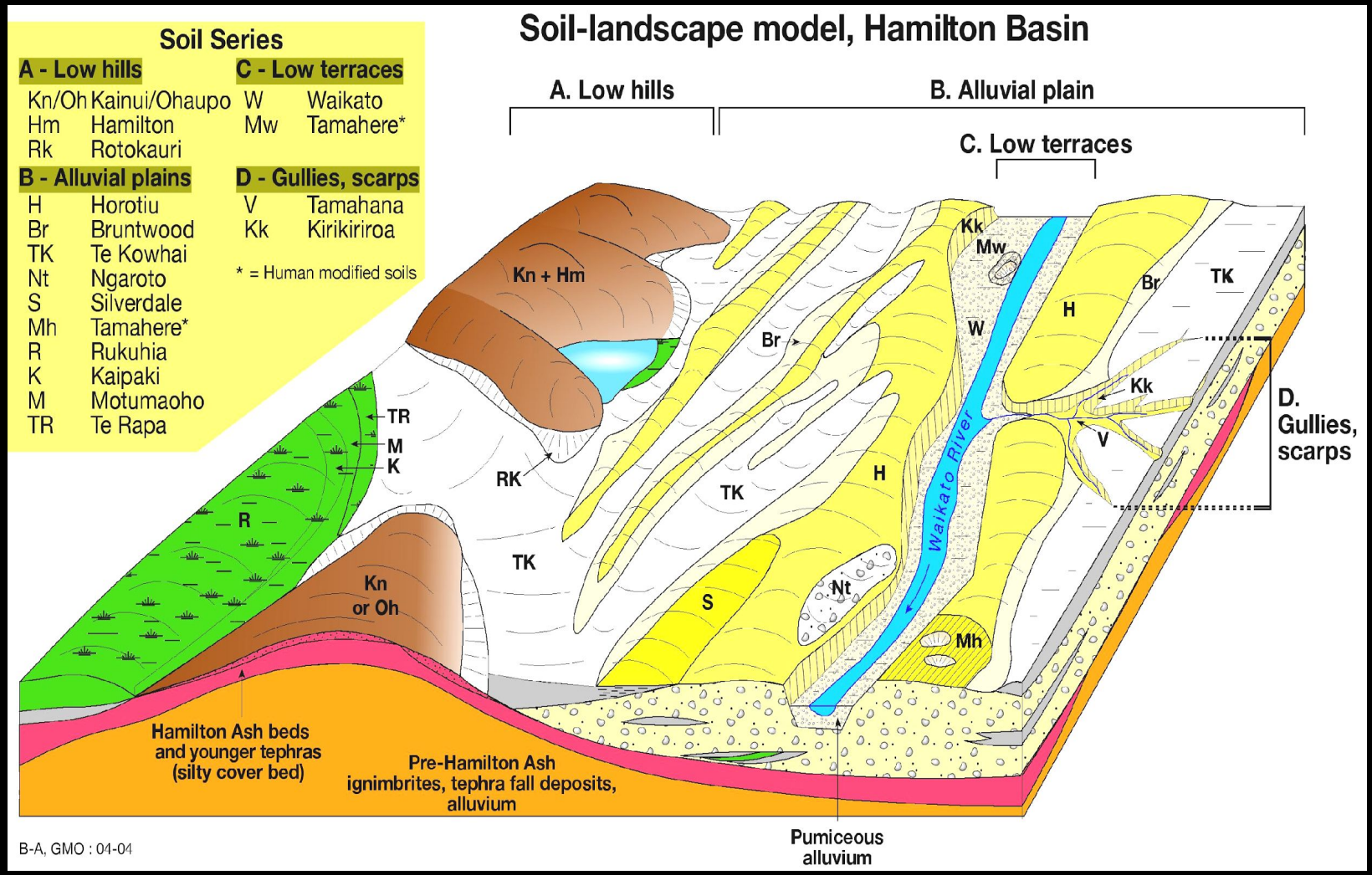
- https://smap.landcareresearch.co.nz


«device»
PC
S-Map Data Entry Tool (DET)
Peer Review System
SMap DataBase Manager
Landcare ActiveDirectory
«PDF»
Factsheet
QUICKBEN
ImageDB
Metadata
WHISKEYJACK
LCRUserAuthentication
«API»
SMapFactSheet
«service»
LCRMetaData
«WebService»
SmapService
«service»
LCRImageDB
«device»
PC
«WindowsApplicati...
LCRDataModelTool
«device»
PC
«DesktopApplication»
SmapSiblingDiagramLoader
«device»
PC
SSMS
«device»
PC (Argis tools)
NOAH
SMap-Spatial
SMapRuleLibrary
SMapPTFLibrary/CLR
SmapSpatialViews
Gbb Fast view
WRC Model
«WindowsService»
SmapCalculatedValuesService
SMAP RUNNER
UncertaintyPipeline
MonteCarlo Execution Manager
RServe
Monte Carlo Web Service
LCRUncert
File system repository
«device»
PC
«WindowsAp...
SoilMapUnit Simulator
Licence Service
«device»
PC
«DesktopApplicati...
ArcGIS
(ESRI SDE)
«device»
PC -Argis (Rescale tools)
Manual process to update DB
Liberator
S-map dev
«device»
PC
FME Workbench
SMap-online
S-map online
«device»
PC-Soil
«WindowsApplication»
SoilProfile Simulator
«device»
PC
«Application»
RStudio
«device»
PC
«DesktopApplication»
Geokettle
«device»
PC Portal Data tools
ArcGis Tools
Hassh
filesystem
upload to portal using manual process
LRIS portal
SMAP APSIM
Apsim Software
«executionEnvir...
Manual Execution
«executionEnvironment»
SnowFlake
ANZSoilML (SnowflakeDB)
«service»
WFS
ANZSoilML
Overseer
«flow»
use
«Request»
S-Map Workflow

Soils – multi scale response to geology, landform, climate and time

Pedologists

Soil-Landscape Model



Otor_28.1

Hapa_7.1

Allophanic soils from composite tephras over pumice alluvium

Otor_28 + Kinle_1.1 + Aroha_3.1

Allophanic hill complex - varying depths of Taupo Pumice and weathered tephras over weathered ignimbrite

Otor_19.1

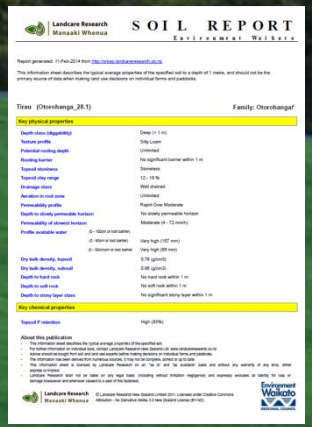
Moes_2.1 + Ngah_9.1

Pumice soils from Taupo Pumice alluvium

Taup_82.1

Allophanic soils from post 60ka composite tephras

Otor_28.1



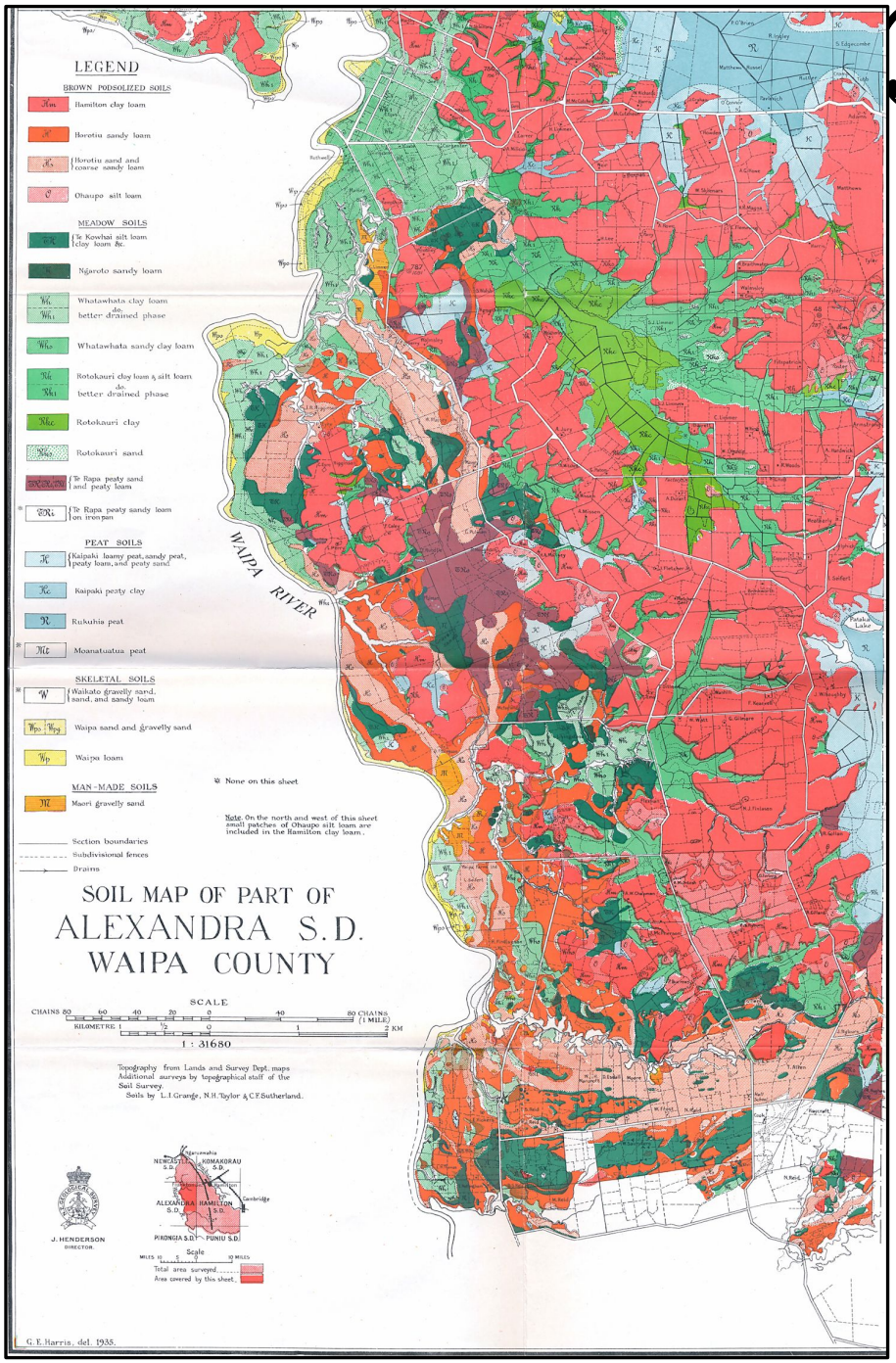
Landcare Research Manaaki Whenua

SOIL REPORT

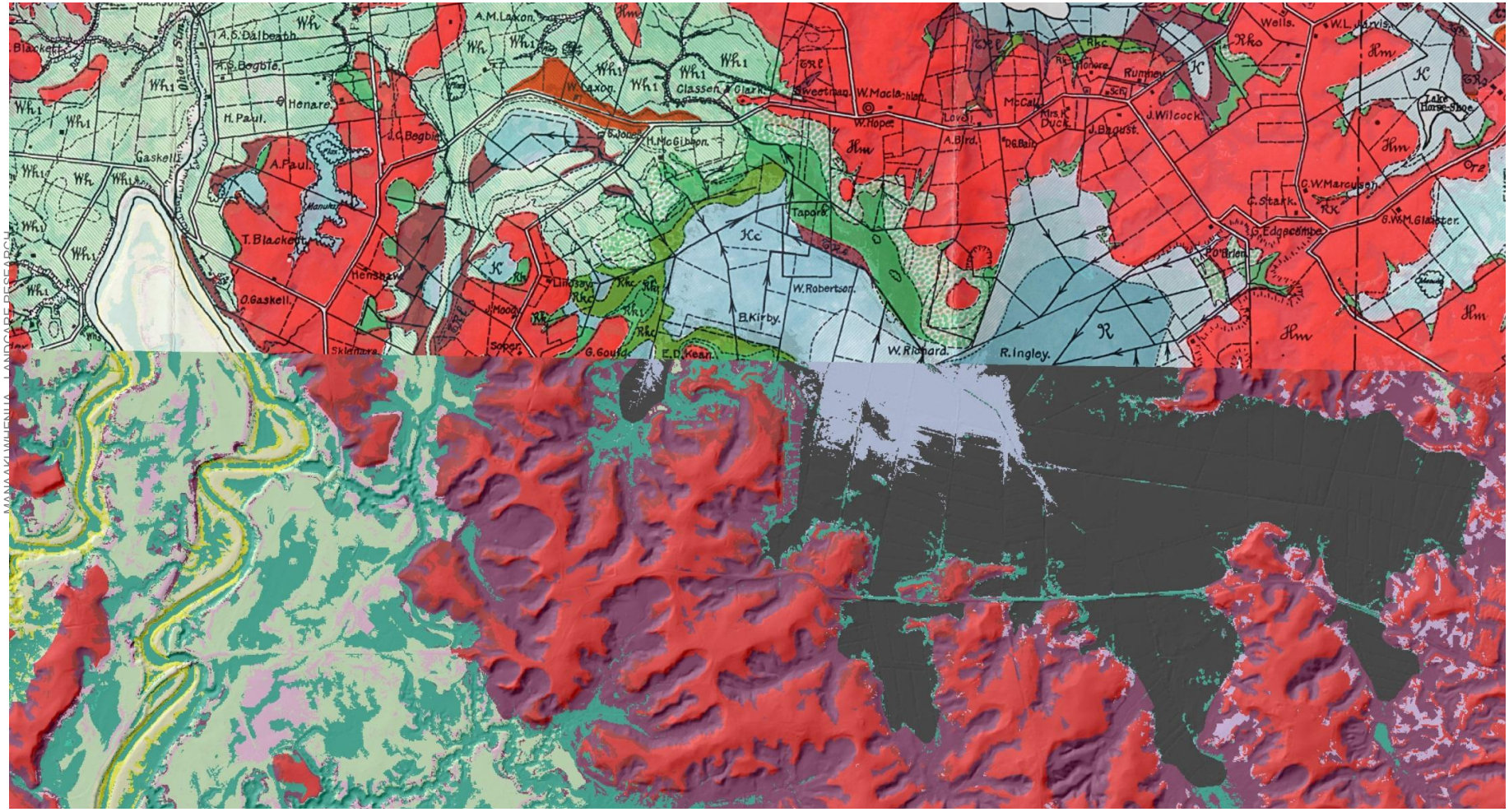
Environment Waikato

… Soil Map

- Art
- Not data driven
- Not suitable for modern demands
- Regional planning max
- Definitely not farm scale
- S-Map plenty not like

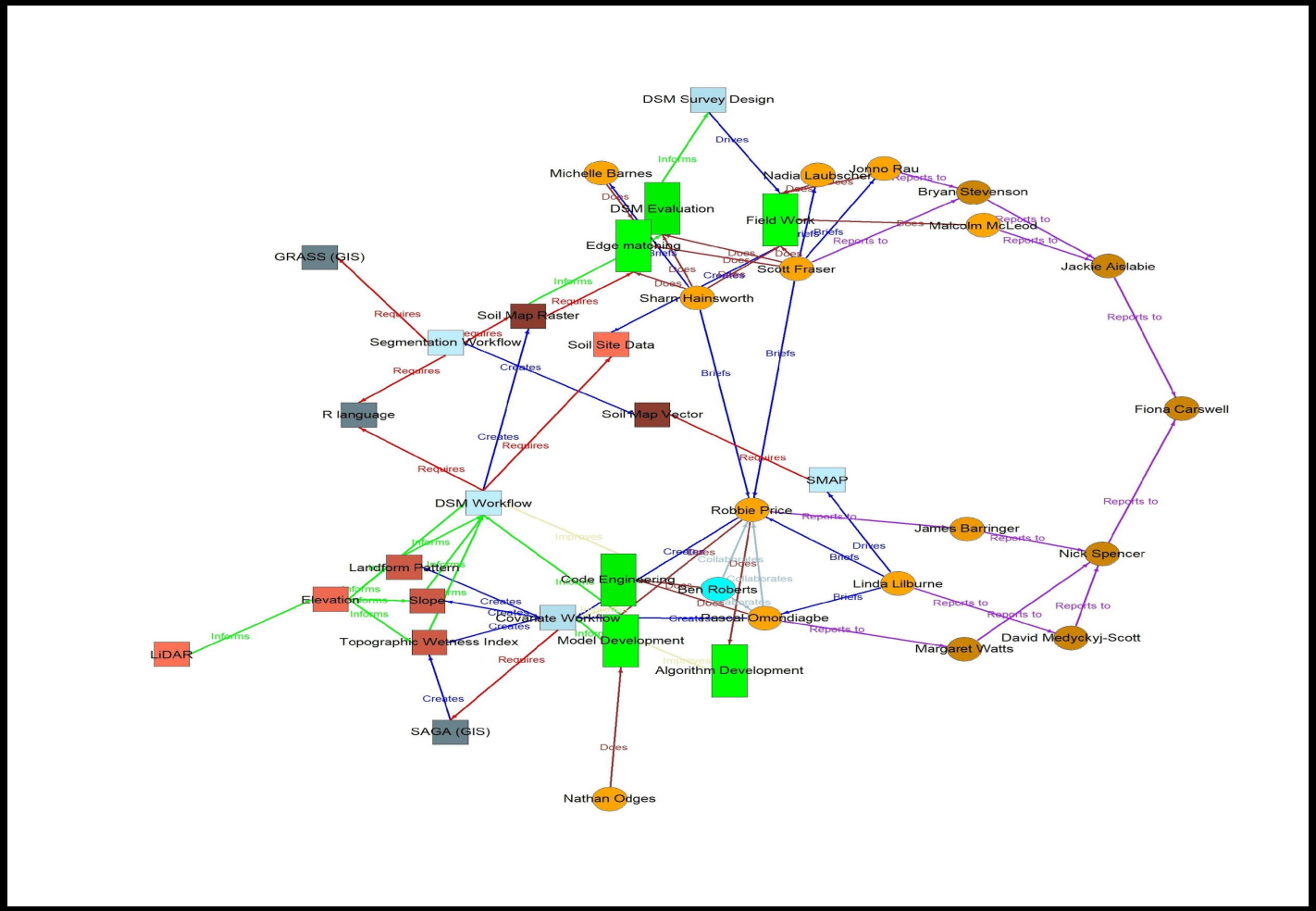


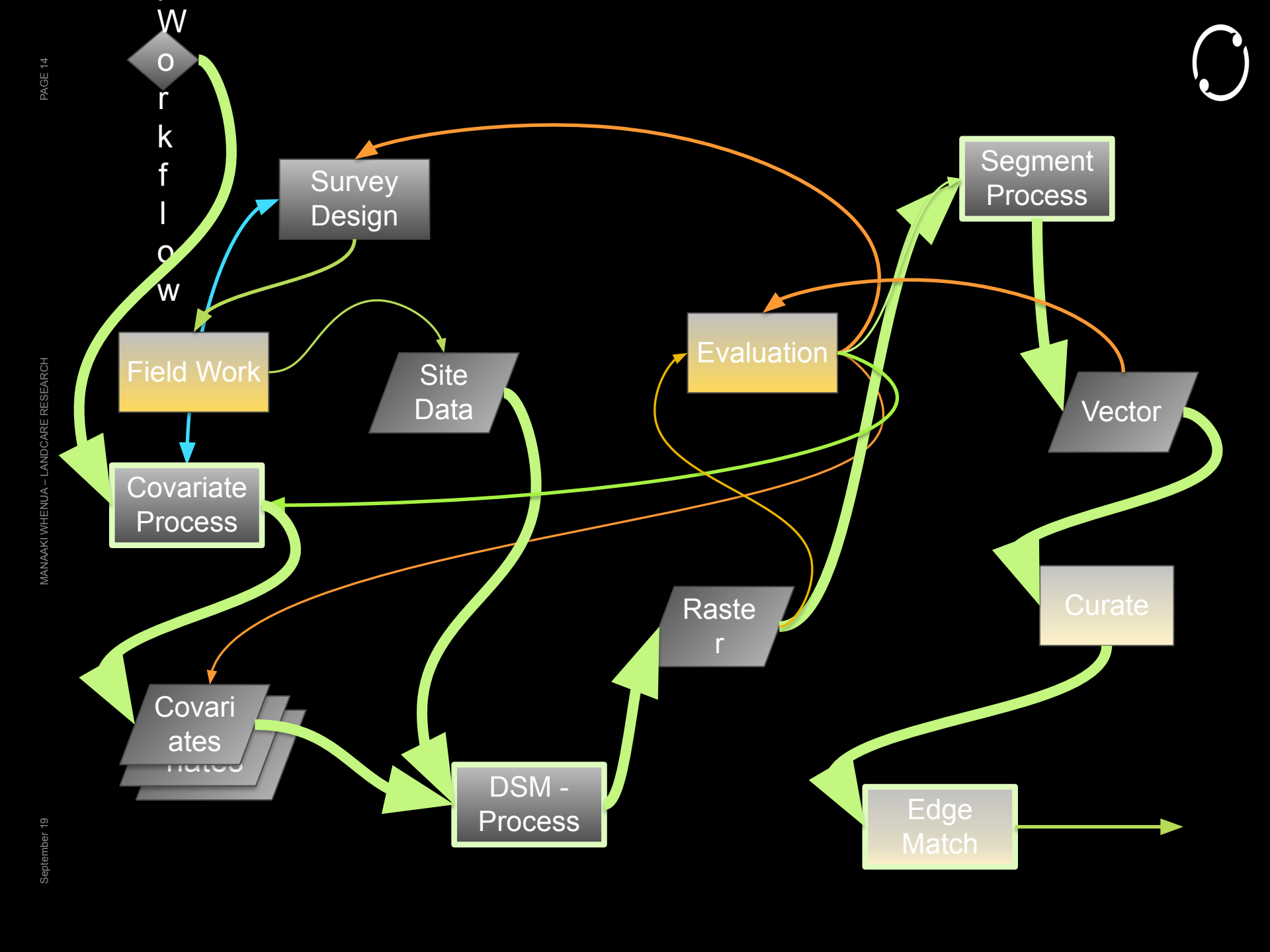
... And so digital soil mapping was born



DSM (Digital Soil Mapping) – developed by Pedologists and slotted into the normal manual way of doing things…

S-Map & Soil Mapping Ecosystem




Workflow
Survey Design
Segment Process
Field Work
Site Data
Evaluation
Vector
Covariate Process
Raster
Curate
Covariates
DSM - Process
Edge Match

Covariate Process

- LiDAR elevation model, Climate data, Geology layers
- Covariates derived using existing tools
  - ArcGIS
  - GRASS
  - SAGA
  - R
  - C/C++
  - Python
- New covariate algorithms
  - R – e.g Landscape Pattern (neighbourhood elevation range)

Workflow Progress

- Wrap existing non ESRI models in R covariate library
  - Standardised way of writing
  - Definitive Implementations as reusable functions
  - Metadata generation embedded into the workflow
- Assess alternative Implementations of ESRI models

DSM Modelling Process

- Expert feel for Area being mapped (Pedologists)
- Site Data to describe soils and locate geospatially
- Sites + Covariates + Model
  - Currently using an over-fitting model: Random Forest
- Iterative process to develop and test models and covariates
  - Choice of covariates
  - Field Sampling Strategy

Workflow Progress

- Site data “management” in Excel
- R based using Knitr to create evaluation documentation at run time
- Workflow to introduce type and play options
- Scope for parallel processing by tiling assessed but not implemented

Segmentation Process - Raster to Vector

- This area most considered an Art - tells a story
- In - Definitive Raster Soil model/s
- Out – Representative vector (linework) map
- Vector map is an inexact representation correct at the attribute level.
- Removal of small areas (minimum polygon size)
- Smoothing of boundaries

Workflow Progress

- Re-implemented ArcGIS concept in GRASS
- Workflow runs and works
- Segmentation needs complete rewrite for science-art reasons

Post Processing Issues

- Curation
  - Line work manually linked back to soils
  - Yet to gain an understanding of the process
- Edge Matching
  - Joining different datasets along boundaries without obvious artefacts
  - Traditionally done manually at end of mapping
  - Attempting to put at beginning of process
- Final Curation
  - Scripts to prepare data for upload
  - Burning in data that didn't model well
  - ArcGIS based

Retrospective

- King said I was daft to build a castle in a swamp…
  - https://www.youtube.com/watch?v=aNaXdLWt17A
- DSM is Evolving Science
  - No previous projects are significantly workflow compatible
  - Spatial processing often quite different
  - Highly dependent on the ESRI suite for GIS component
  - Metadata or documentation for intermediate steps inadequate to replicate work
    - Even where it exists
  - Some OS geospatial processing methods not particularly good
- Pedologists require flexibility into the future
- First attempt of Segmentation was done using the previous NeSI cluster (New hardware, software, rules… basically new system)

Someone wake Pascal up…

Requirement- The Brief as Given

- Turnkey end-to-end workflow for each DSM project
- Standardised Covariates with national coverage and reproducible

The Brief we wrote ourselves

- Standardise covariate algorithms
- Attempt to document prior DSM work
- Create workflow system for current and future wor0k
- Identify non workflowable aspects of SMAP work
    - Are there other approaches?
- Find holes in the existing methodology
- Identify which steps of the workflow requires HPC
- Code management integrated into the system and documentation

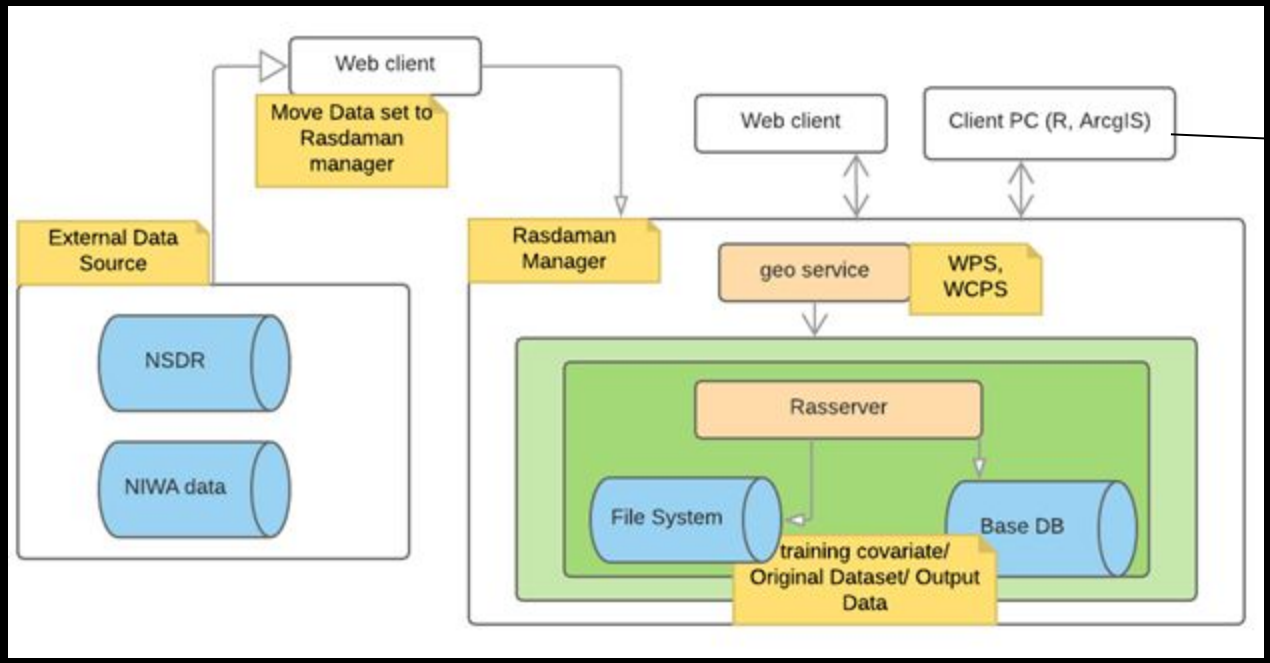
Workflow Principles

- Single language entry-point
  - R – most of our analytical code
- One code base for all platforms
  - Same code to run on local workstation as NeSI
  - Tell it where to run
- R6 Classes for wrapping models
- Generate metadata for each output
  - ISO standard XML
- Home project location for all project data (File System)

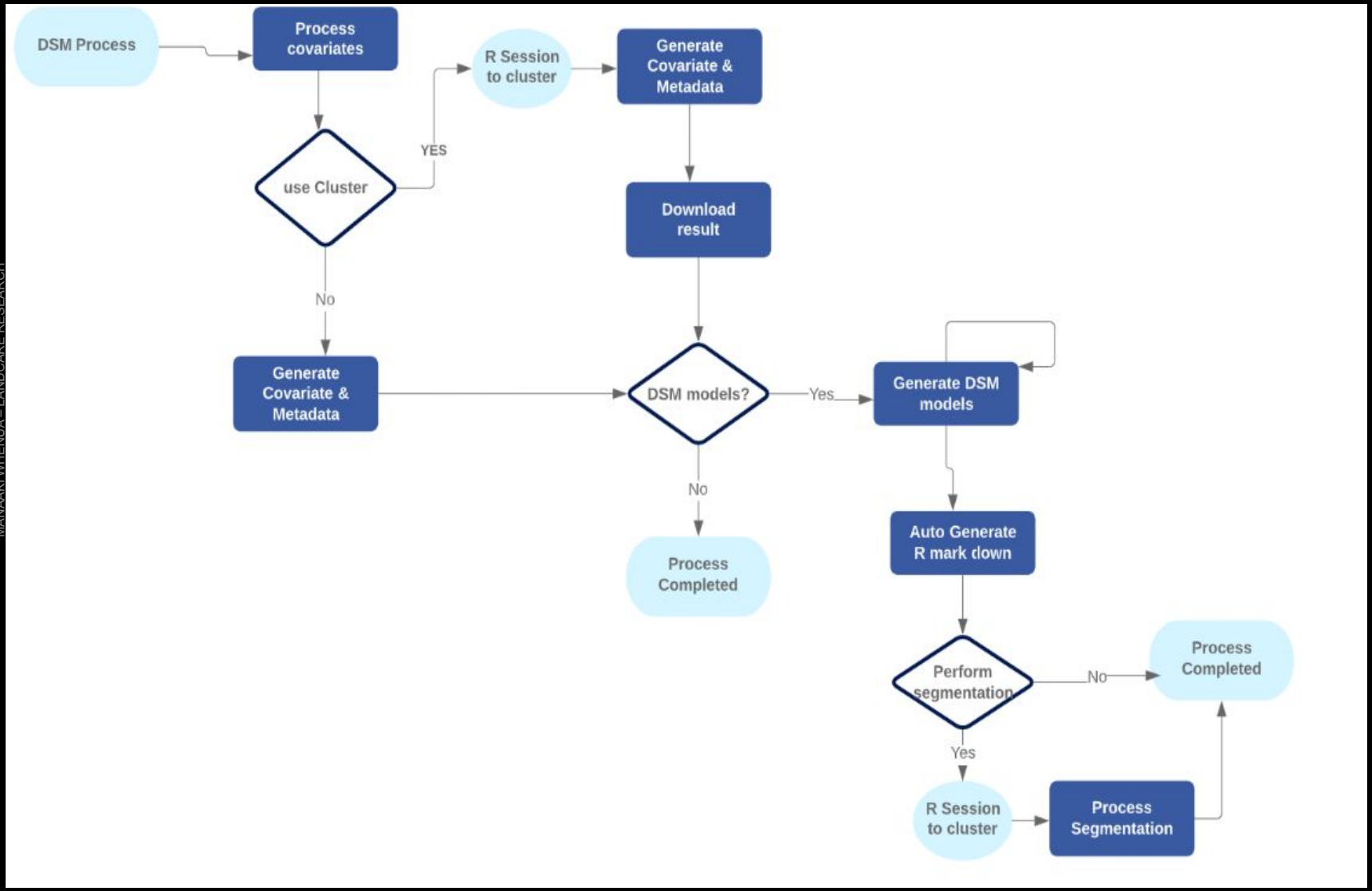
DEVELOPMENT PROCESS

- Coding Standardisation
- Metadata
- R6
- Workflow template

DSM Workflow –Initial plan



DSM Workflow –Design



- Workflow Via Drake Package

- R Template → smapDSM R Package → smapSegmentationPackage

DSM Workflow – Covariate Process

- Wrap existing non ESRI models (e.g. Saga, Grass) in R library
- Previous Grass and Saga R library exits ?
- R6 Classes for wrapping models (easier to use by The DSM guys)
- Process Locally or on the cluster
- Provide easy template to DSM guys

DSM Workflow – Covariate Process (R6 Class)

```
################################################### STEP 1  Define saga algorithm to execute (e
## Topographic Wetness Index #####
sagaAlgorithm <-function(){
  sm=smapDSM::runSaga()
  sm$path ="C:/Program Files/saga-6.3.0/saga_cmd.exe"
  sm$projectDir ="D:/Projects/smap-dsm/data/"
  sm$rasterFile="test_DEM.tif"
  sm$rasterFolder="D:/Projects/smap-dsm/data/"
  sm$resultPath="result"
  sm$outputVector="final"
  sm$sagaInit()
  sm$metaDataFirstName="pascal"
  sm$metaDataSurName="pascal"
  sm$isCluster=F
  #specify saga algorithm name
  sm$sagaAlgorithm = sm$saga$ta_hydrology$saga_wetness_index
  #get paramters
  sm$sagaParameters()
  sm$parameters$`AREA (Type: Grid (output):default value:NA)`="area"
  sm$parameters$`SLOPE (Type: Grid (output):default value:NA)`="slope"
  sm$parameters$`SUCTION (Type: Floating point:default value:10.000000)`=10.0
  sm$parameters$`AREA_TYPE (Type: Choice:default value:1)`=1
  sm$parameters$`SLOPE_TYPE (Type: Choice:default value:1)`=1
  sm$parameters$`SLOPE_OFF (Type: Floating point:default value:0.100000)`=0.1
  sm$parameters$`SLOPE_WEIGHT (Type: Floating point:default value:1.000000)`=1.0
  sm$process()
}
```

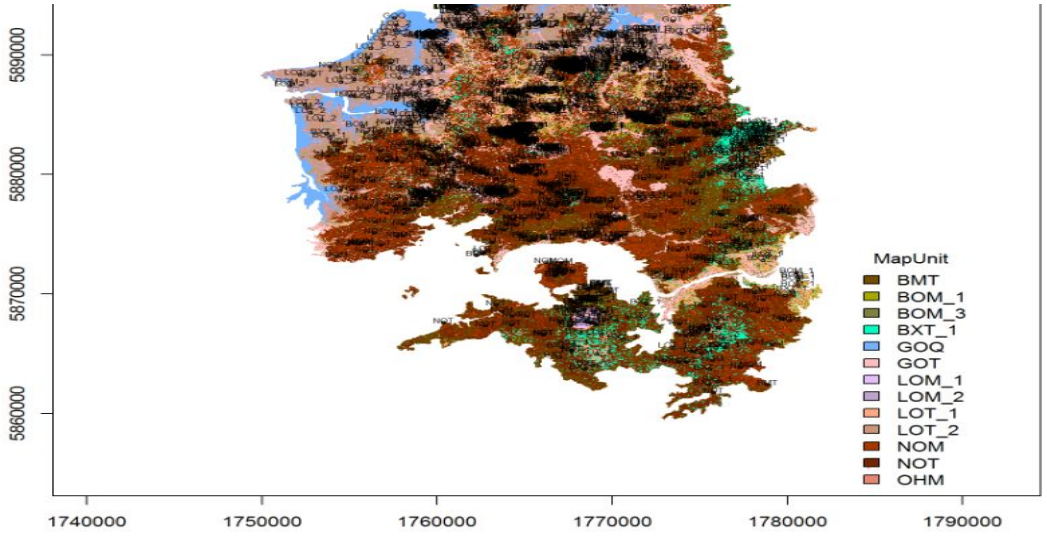
DSM Workflow – Covariates Process

```
############################ workflow Plan when processing a single raster file #
singlePlan <- drake::drake_plan(


  #wetness
  covariate=sagaAlgorithm(),

  #geomorphons
  covariate2=grassAlgorithm()

 )

clean(destroy = TRUE)
make(singlePlan, parallelism = "future", jobs = 2,force = TRUE)
config<- drake_config(singlePlan,verbose = 3)
outdated(config)
vis_drake_graph(config)
```

DSM Modelling Using Random Forest (Previous Approach)

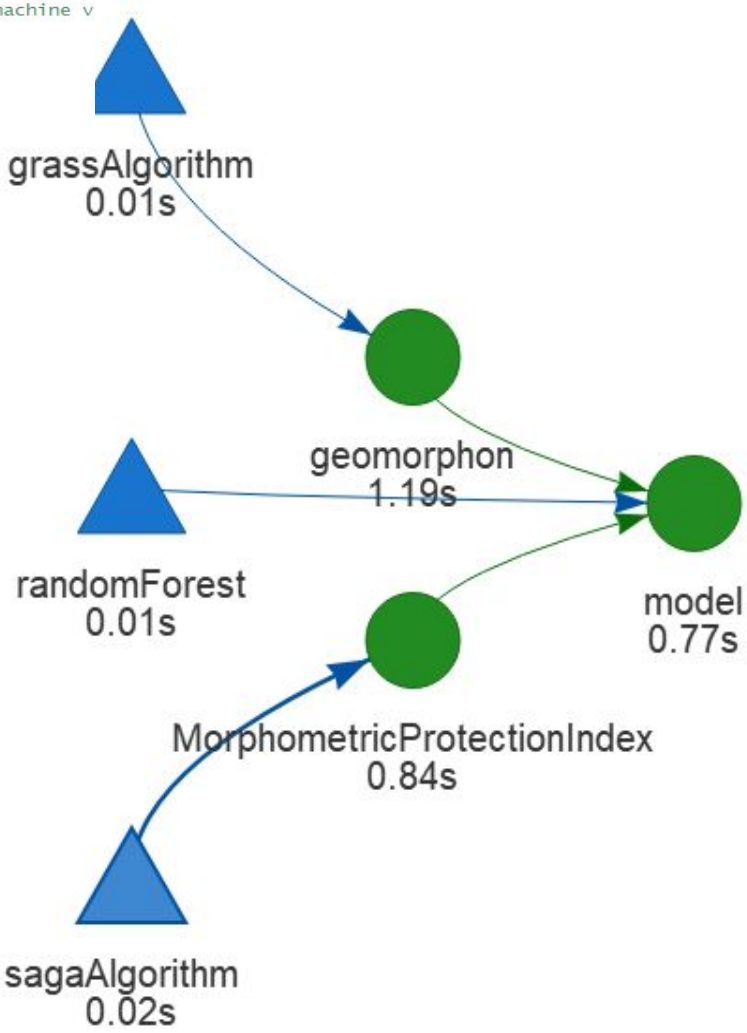

```
#-- -DEM derived
lstGrds[04] <- "P:\\Projects\\SL1705_Franklin_SMap\\Data\\GIS\\Grids\\DJP_backup30_3_2017\\F5ranklin DEm\\TerrainAtts\\mrvbf
5m.img"

lstGrds[05] <- "P:\\Projects\\SL1705_Franklin_SMap\\Data\\GIS\\Grids\\DJP_backup30_3_2017\\F5ranklin DEm\\TerrainAtts\\img
\\eudist.img"
lstGrds[06] <- "P:\\Projects\\SL1705_Franklin_SMap\\Data\\GIS\\Grids\\DJP_backup30_3_2017\\F5ranklin DEm\\TerrainAtts\\img
\\lelementx.img"
lstGrds[07] <- "P:\\Projects\\SL1705_Franklin_SMap\\Data\\GIS\\Grids\\DJP_backup30_3_2017\\F5ranklin DEm\\TerrainAtts\\img
\\lengthx.img"
lstGrds[08] <- "P:\\Projects\\SL1705_Franklin_SMap\\Data\\GIS\\Grids\\DJP_backup30_3_2017\\F5ranklin DEm\\TerrainAtts\\img
\\planc.img"
```

```
#--- run the random forest
#rfResult <- randomForest(frmFields, data = tblInput[tblTrainingData,],ntree=500,mtry= 5)
rfResult <- randomForest(frmFields, data = tblTrainingData, ntree=500,mtry= 5)

plot(rfResult)
```



DSM Workflow –Modelling Via SmapDSM Package

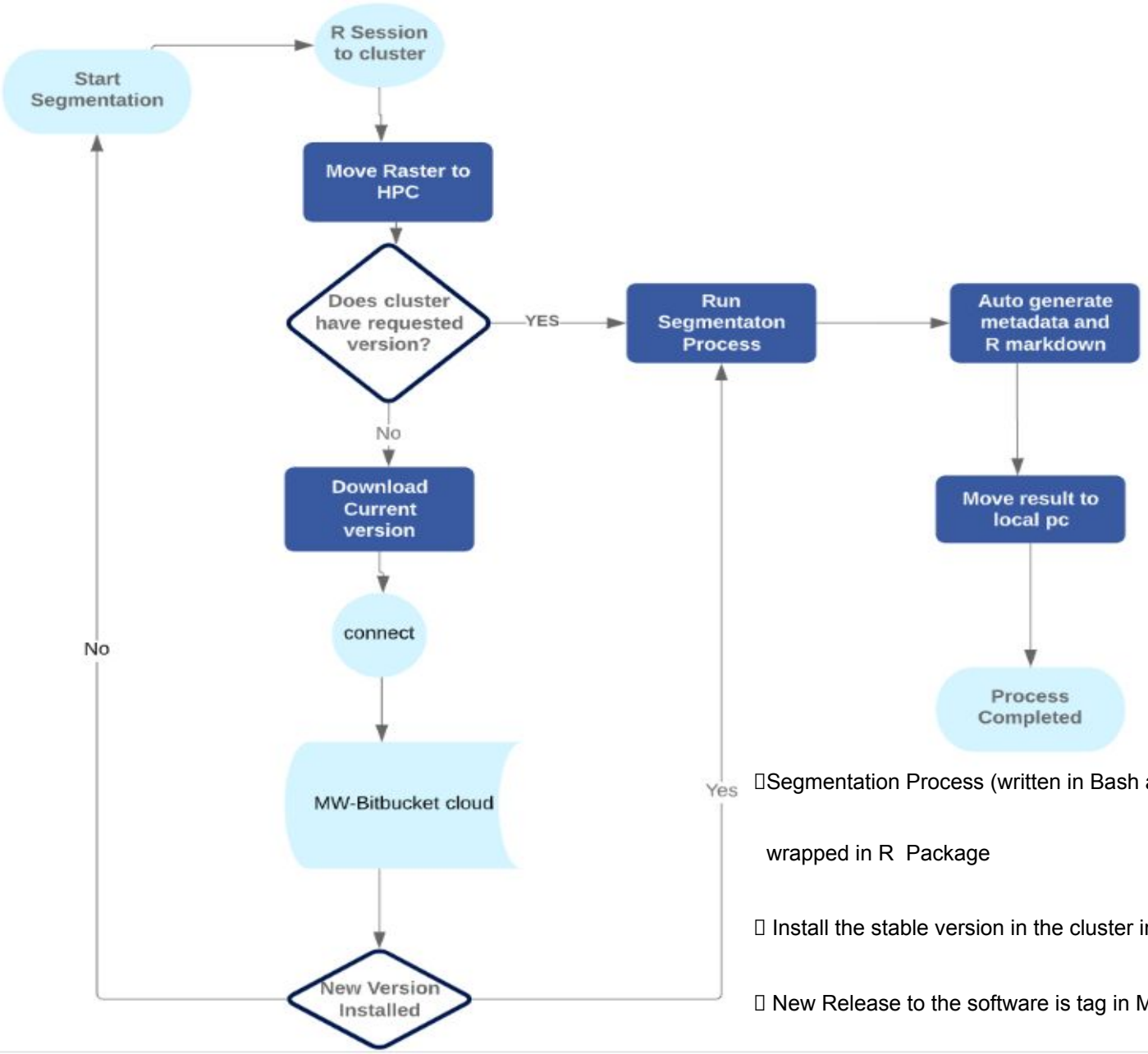
```
############################## workflow Plan when processing a single raster file
singlePlan <- drake::drake_plan(


  #Morphometric Protection Index
  MorphometricProtectionIndex=sagaAlgorithm(),

  #Calculates geomorphons (terrain forms) and associated geometry using machine v
  geomorphon=grassAlgorithm(),

  model=randomForest(MorphometricProtectionIndex,geomorphon)

 )
```



Segmentation Workflow - Initial Design

- Manually Upload segmentation script to cluster
- Manually Run segmentation process
- (you get the picture) Download Result back
- **Not Efficient ?**
- **Lots of passwords and keys**

Segmentation Workflow Re-design



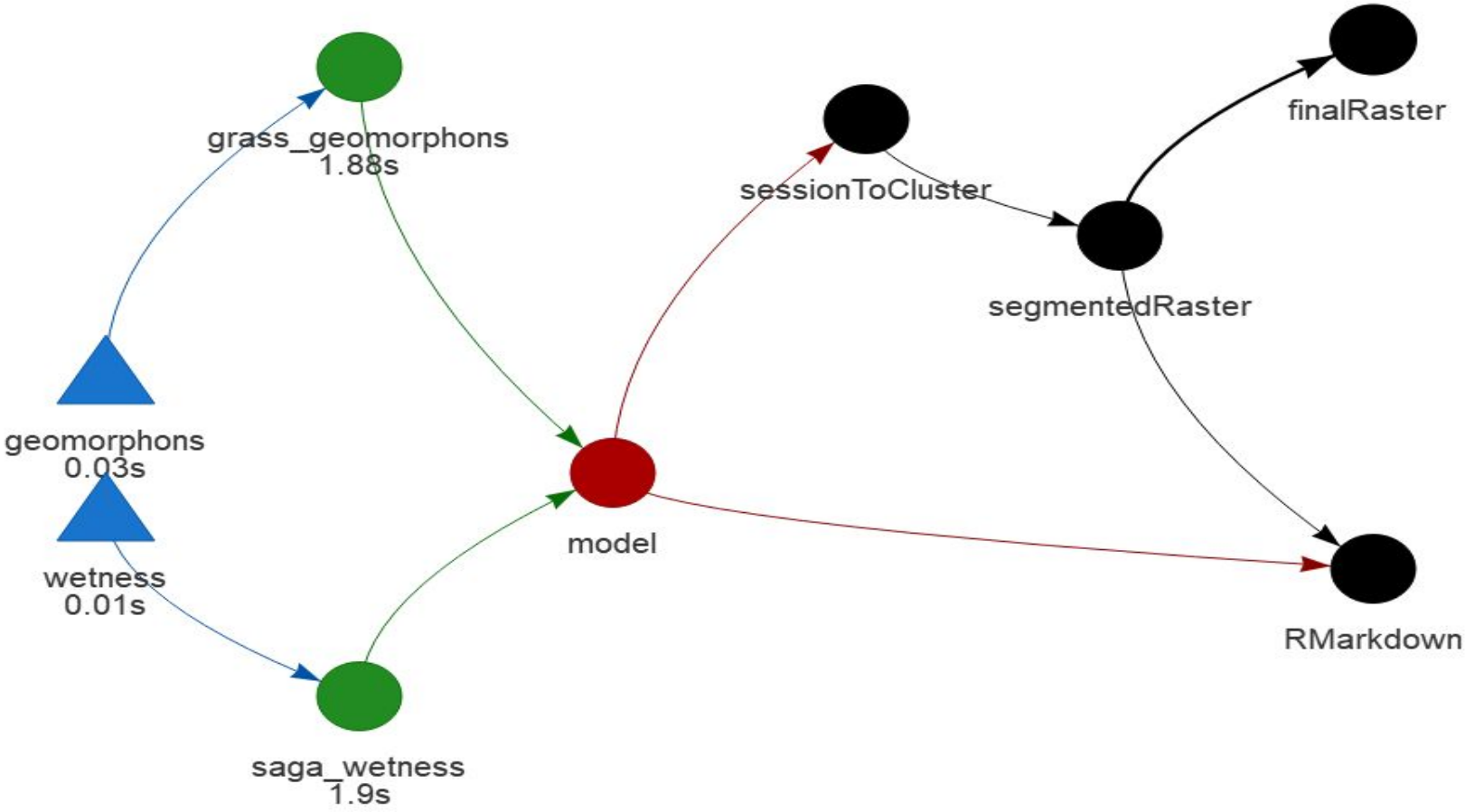
- Segmentation Process (written in Bash and uses Grass Gis)

  wrapped in R Package

- Install the stable version in the cluster inside the project directory

- New Release to the software is tag in MWLR Repository

DSM Workflow : End-to-End



Performance/Re-evaluate - Workflow

- Ben Roberts (NeSI) provides report on usage
- Evaluate the performance
- Redesign segmentation script
    - Auto-generate individual tiles
    - Auto-generate SLURM script on the fly
    - Process each tile job on different node
    - R SLURM job to stitch job together

S-Map Workflow - Summary

- Dream of an end-to-end workflow not fully realised (almost there, (not)
- Chose a single language of entry
- An over arching code protocol
- Designing a flexible coding system for evolving science
- Individual components of Workflow being built
- Code management crucial

Acknowledgements

- Linda Lilburne – funding and driving
- James Barringer – preliminary segmentation, feedback
- Scott Fraser, Sharn Hainsworth – DSM work guinea pigs (photos we stole from Scott's presentations)
- Ben Roberts (and other NeSI support)