

# Whole Tale: The *Experience* of Research

**Whole-Tale:** *Reproducible, computational narratives*

**YesWorkflow:** *Revealing workflow, provenance from scripts*

**Kurator:** *Automating data cleaning workflows*

**EulerX:** *Agreeing to disagree about variant taxonomies*

**Bertram Ludäscher**

**`ludaesch@illinois.edu`**

Director, Center for Informatics Research in Science & Scholarship (CIRSS)

School of Information Sciences (iSchool@Illinois)

& National Center for Supercomputing Applications (NCSA)

& Department of Computer Science (CS@Illinois)



School of  
**Information Sciences**  
The iSchool at Illinois

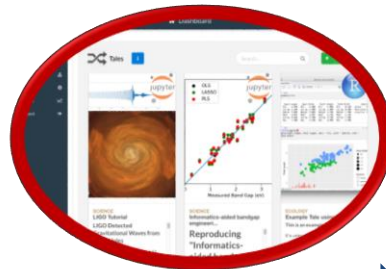
*BCoN Workshop*  
*2018-02-13..14 U Kansas*

# Whole Tale: The next step in the evolution of the scholarly article: The “Living” Paper

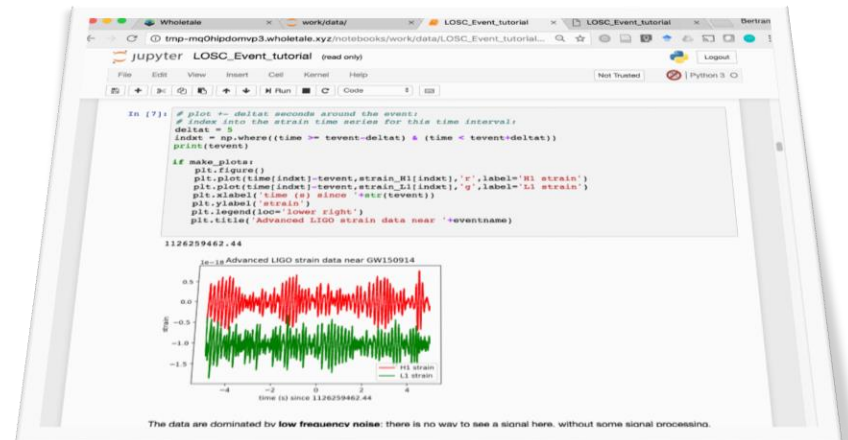
- 1<sup>st</sup> Generation:
  - **narrative** (prose)
- 2<sup>nd</sup> Generation: **plus** ...
  - name .. identify .. include (access to) **data**
- 3<sup>rd</sup> Generation: **plus** ...
  - name .. reference .. include **code** (software) ..
  - and **provenance** ... and **exec environment** (containers)



Whole Tale



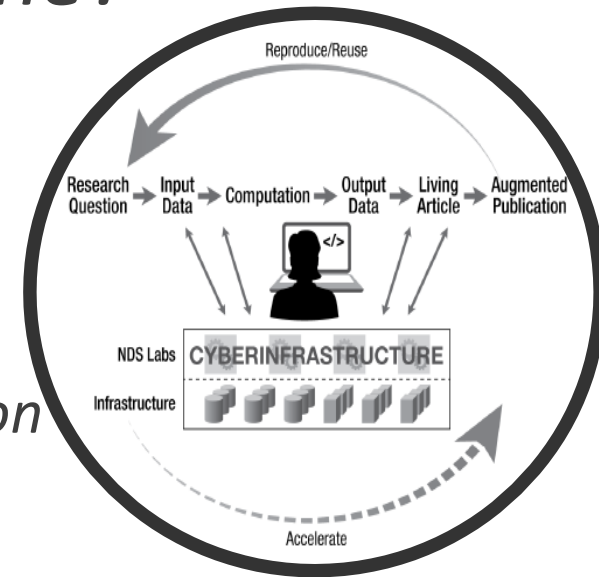
Whole Tale **Dashboard**



# Whole Tale: *What's in a name?*

## (1) Whole **Tale** $\Leftrightarrow$ Whole **Story**:

- *Support (computational / data) scientists*
- *... along the complete research lifecycle*
- *... from experiment to (new kind of) publication*
- *... and back!*



## (2) Whole **Tale** $\Leftrightarrow$ for the **Long Tail of Science**

- *Easy sharing of your computational narratives, data, and exec-env since 2017!*
- *Power applications for everyone!*



# The **Whole Tale**: *Merging Science and Cyberinfrastructure Pathways*

## NSF-DIBBS award (5 years, 5 institutions)

- **Illinois (NCSA & iSchool)**
  - Bertram Ludäscher (PI), MT Campbell (PM) [Kandace Turner], Victoria Stodden (coPI), Matt Turk (coPI), Kacper Kowalik (sw-architect), Craig Willis (dev)
- **U of Chicago**
  - Kyle Chard (coPI), Mihael Hategan (dev)
- **UT Austin/TACC**
  - Niall Gaffney (coPI), Siva Kulasekaran (dev)
- **U Notre Dame**
  - Jarek Nabrzyski (coPI), Ian Taylor (sw-dev), Adam Brinckman (dev)
- **UCSB/NCEAS**
  - Matt Jones (coPI), Bryce Mecum (dev)





# Whole Tale **Motivation**

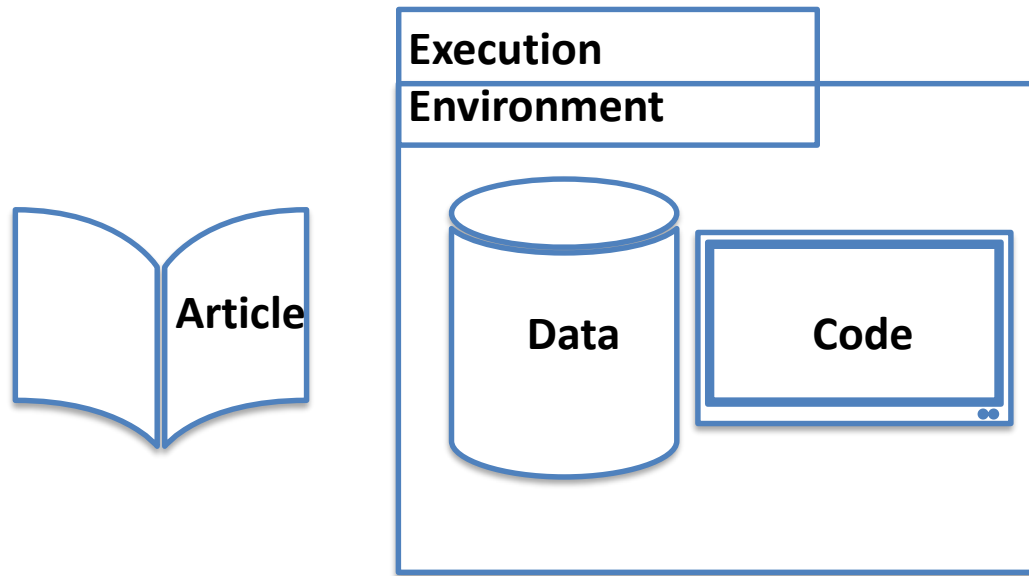
- Can't reproduce result because:
  - Don't know how to run analysis
  - Can't get the software running
  - Can't pay for the computer or compute power the result was computed on

*Source: Bryce Mecum, WT team @NCEAS*



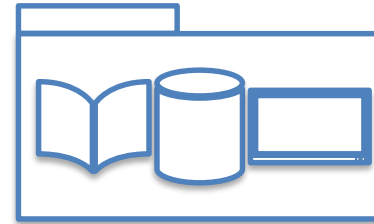
# Whole Tale Vision

## Addressing reproducibility

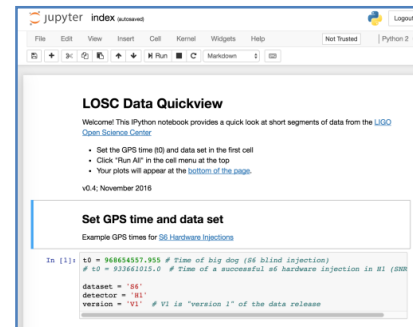
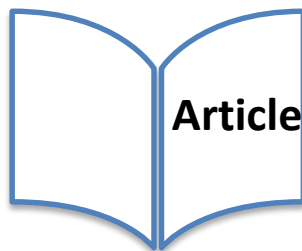


100

- Living publication  
(data + code + environment)

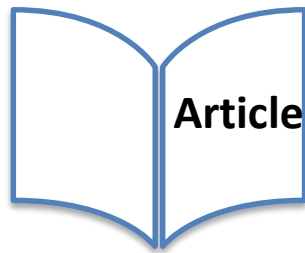


- Facilitate reproducibility
- Encourage investigation of results making it easy to recreate the environment the result was created in



# Whole Tale Vision

## Addressing reproducibility

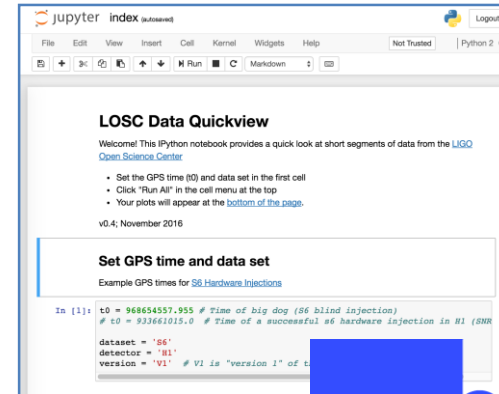
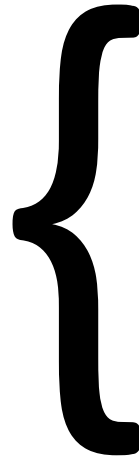


+

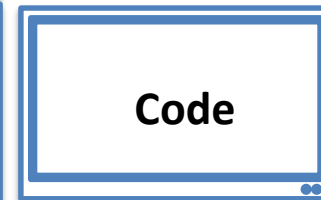
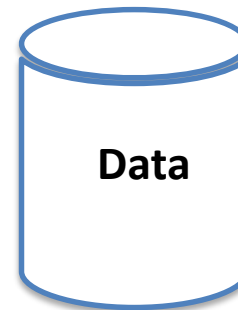


**Tale**

# Whole Tale Vision

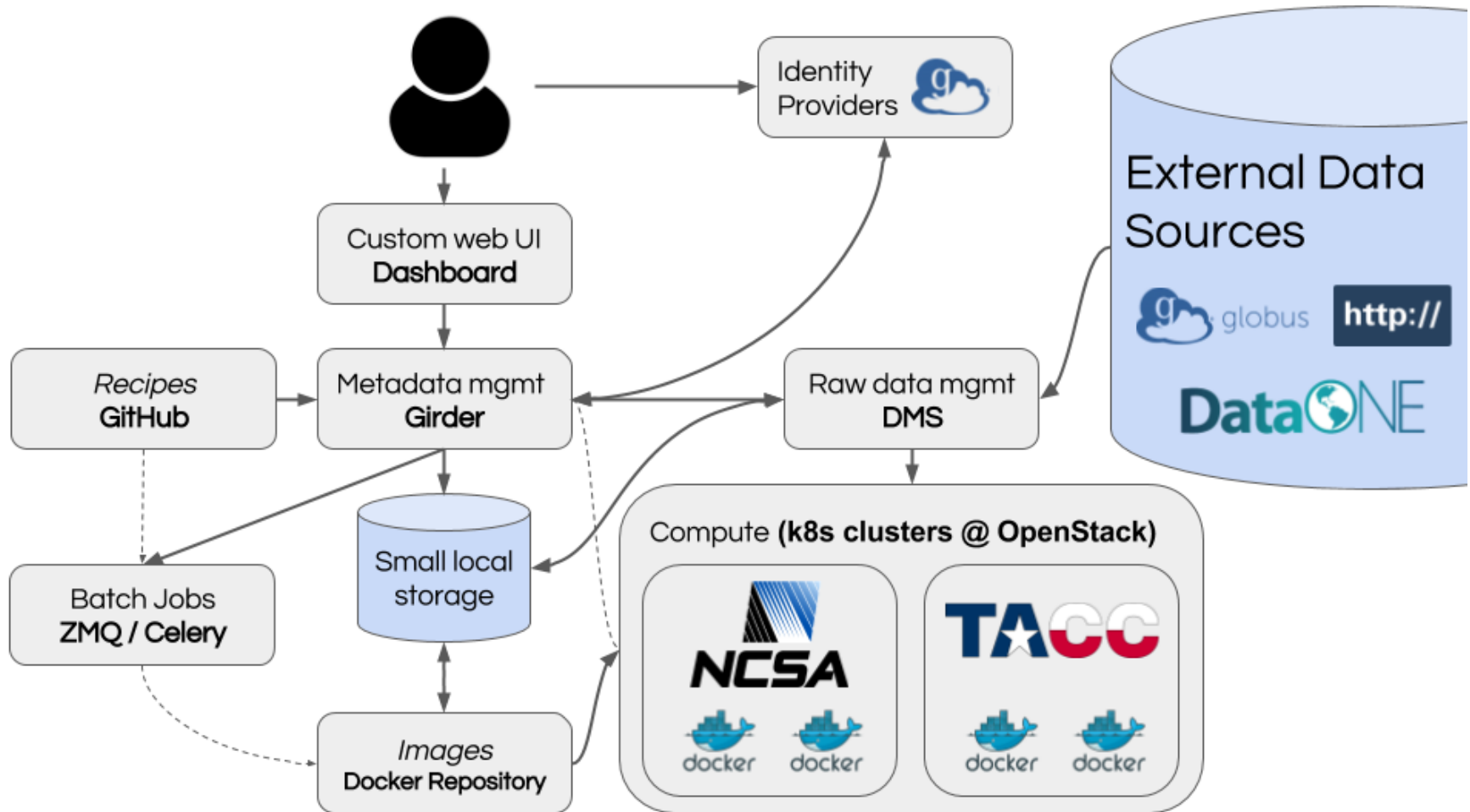


D1PROV





# WT Architecture



# DEMO: (re-)use existing tale or ...

WHOLE TALE

Dashboard

Bertram Ludaescher

Home

Tales

Catalog

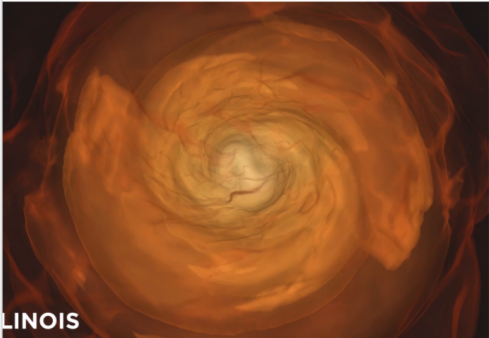
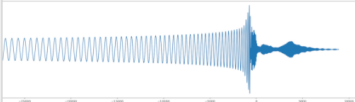
My Data

Compose

Status

Logout

Tales



SCIENCE

LIGO Tutorial

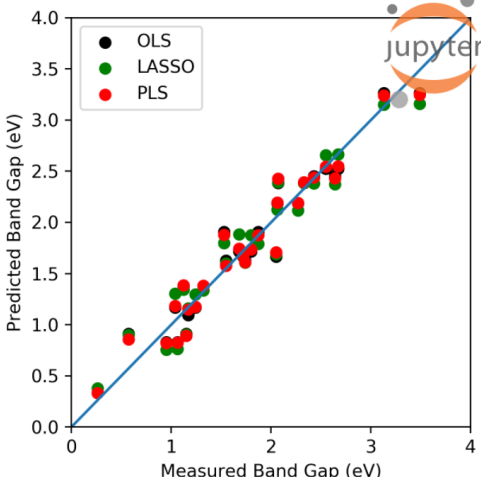
LIGO Detected Gravitational Waves from Black Holes

On September 14, 2015 at 5:51 a.m. Eastern Daylight Time (09:51 UTC), the twin Laser Interferometer Gravitational-Wave Observatory (LIGO) detected gravitational waves from the merger of two black holes.

Kacper Kowalik

Launch

Tales



SCIENCE

Informatics-aided bandgap engineering...

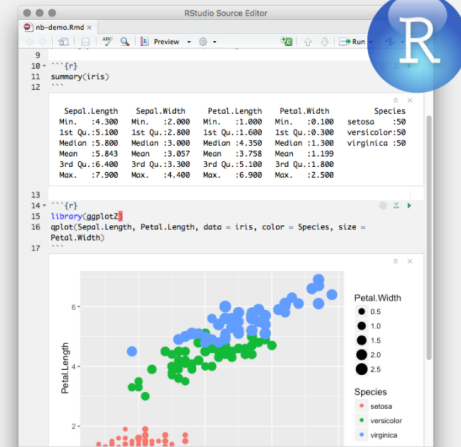
Reproducing "Informatics-aided bandgap engineering for solar materials"

This notebook shows how to replicate the main findings of a 2014 paper by Dey *et al* that used machine learning to predict the bandgap of solar materials.

Logan Ward

Launch

Tales



ECOLOGY

Example Tale using R

This is an example Tale

It's using R Studio.

Kacper Kowalik

Launch

1

2

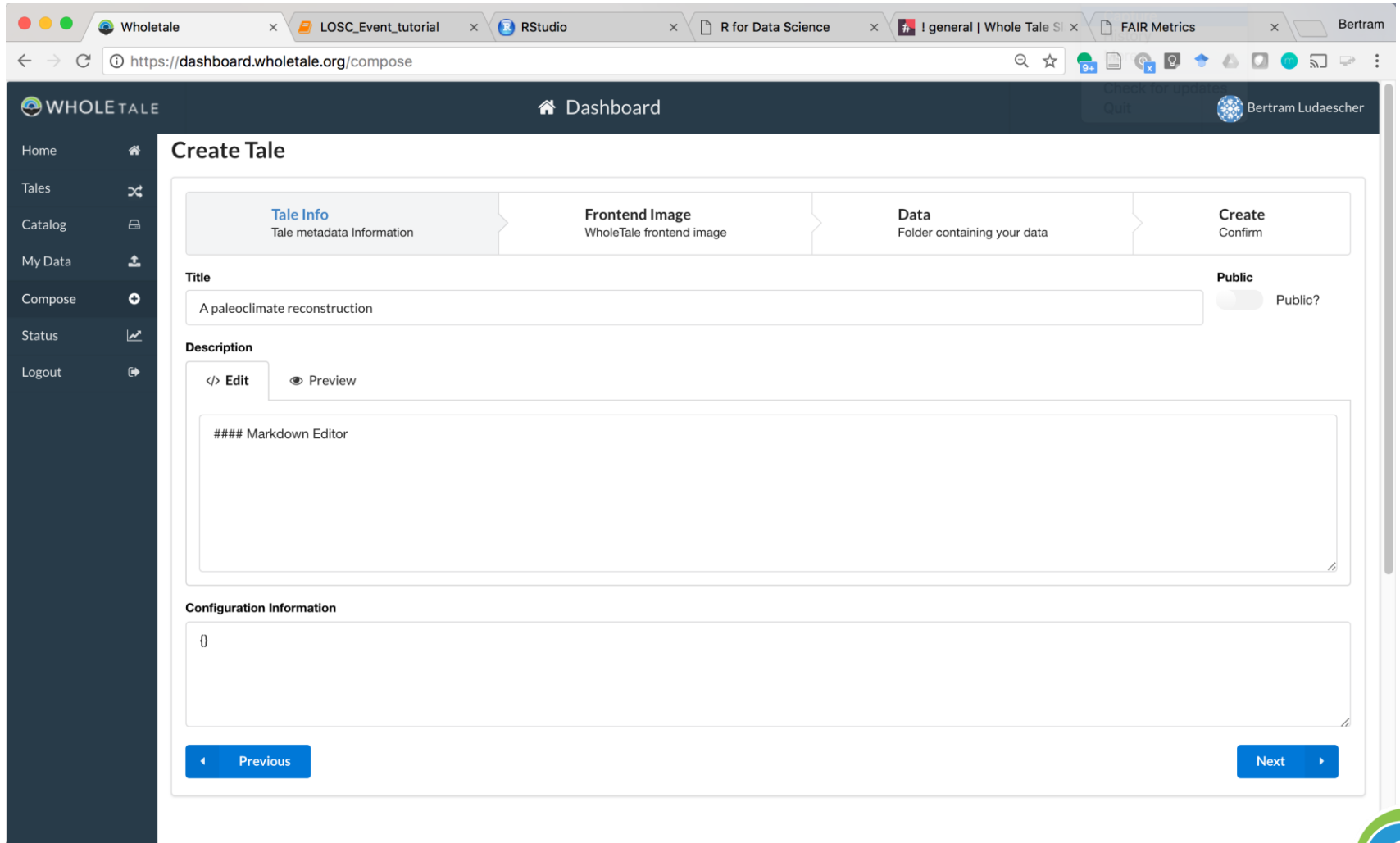
3

4

Ludäscher: Whole-Tale++

11

# ... create a *new Tale*!



The screenshot shows the 'Create Tale' interface in the WholeTale dashboard. The browser tabs at the top include 'Wholetale', 'LOSC\_Event\_tutorial', 'RStudio', 'R for Data Science', 'general | Whole Tale S...', 'FAIR Metrics', and 'Bertram'. The URL bar shows 'https://dashboard.wholetale.org/compose'. The dashboard header includes the 'WHOLETALE' logo, a 'Dashboard' link, and a user profile for 'Bertram Ludaescher'. A left sidebar contains navigation links: Home, Tales, Catalog, My Data, Compose (active), Status, and Logout. The main content area is titled 'Create Tale' and features a four-step progress bar: 'Tale Info' (active), 'Frontend Image', 'Data', and 'Create'. The 'Tale Info' step contains a 'Title' field with the text 'A paleoclimate reconstruction', a 'Public' toggle switch, and a 'Description' field with a 'Markdown Editor' placeholder. At the bottom, there are 'Previous' and 'Next' navigation buttons.

WHOLETALE Dashboard

Check for updates  
Quit

Bertram Ludaescher

### Create Tale

Tale Info  
Tale metadata Information

Frontend Image  
WholeTale frontend image

Data  
Folder containing your data

Create  
Confirm

**Title**

A paleoclimate reconstruction

**Public** ☐ Public?

**Description**

[Edit](#) [Preview](#)

#### Markdown Editor

**Configuration Information**

{ }

[Previous](#) [Next](#)



WholeTale x LOSC\_Event\_tutorial x RStudio x R for Data Science x ! general | Whole Tale S x FAIR Metrics x Bertram

https://dashboard.wholetale.org/compose

# WHOLETALE

## Dashboard

Bertram Ludaescher


### Create Tale


**Tale Info**  
Tale metadata Information


**Frontend Image**  
WholeTale frontend image


**Data**  
Folder containing your data

**Create**  
Confirm

  
Rstudio

  
Jupyter Notebook

  
Jupyter with Spark

  
Refine

Previous

Let's use this Frontend ...

coming soon:  
add your own Frontend Images!

Next

WholeTale Dashboard - Create Tale

Navigation: Home, Tales, Catalog, My Data, Compose, Status, Logout

Dashboard Tabs: Tale Info (Tale metadata Information), Frontend Image (WholeTale frontend image), **Data** (Folder containing your data), Create (Confirm)

Name	Created	Last Modified	Size
Data	6 months ago	6 months ago	7.91 kB
Home	6 months ago	6 months ago	34.61 kB
Workspace	6 months ago	6 months ago	0

OR

[Select From Registered Datasets](#)

[Previous](#) [Next](#)

**Data Binding:**  
Bring in your own (local) data,  
or link to existing data  
(DataONE search, Globus, ..)





## Create Tale

<b>Tale Info</b> Tale metadata Information	<b>Frontend Image</b> WholeTale frontend image	<b>Data</b> Folder containing your data	<b>Create</b> Confirm
---	---	--	--------------------------

### Confirm Details

#### Frontend

Rstudio

#### Folder

Home

#### Title

A paleoclimate reconstruction

#### Public

☐ Public?

#### Description

#### Markdown Editor

#### Configuration Information

```
{}
```

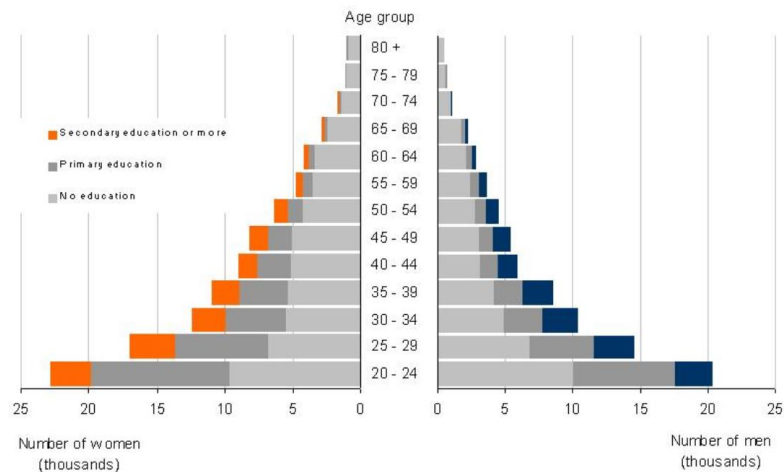
Previous

Submit



# A paleoclimate reconstruction

By **Bertram Ludaescher**



Category: **SCIENCE**

Description

Markdown Editor

## Tale Tools

Instantiate Tale

Launch Tale

Instantiation Name

Tale Name (optional)



Wholetale x RStudio x RStudio x LOSC\_Event\_tutorie x R for Data Science x \* general | Whole T x FAIR Metrics x Bertram

tmp-m9ikiluctaq.prod.wholetale.org

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console ~/

```

*** building package inaiices
** installing vignettes
** testing if installed package can be loaded
* DONE (tidyverse)

The downloaded source packages are in
  '/tmp/Rtmpo441Mz/downloaded_packages'
> library(tidyverse)
— Attaching packages — tidyverse 1.2.1 —
✓ ggplot2 2.2.1 ✓ purrr 0.2.4
✓ tibble 1.3.4 ✓ dplyr 0.7.4
✓ tidyr 0.8.0 ✓ stringr 1.2.0
✓ readr 1.1.1 ✓ forcats 0.2.0
— Conflicts — tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag() masks stats::lag()
> mpg
# A tibble: 234 x 11
  manufacturer model displ year cyl trans drv cty hwy
    <chr>      <chr> <dbl> <int> <int> <chr> <chr> <int> <int>
1 audi      a4      1.8  1999  4 auto(l5) f 18 29
2 audi      a4      1.8  1999  4 manual(m5) f 21 29
3 audi      a4      2.0  2008  4 manual(m6) f 20 31
4 audi      a4      2.0  2008  4 auto(av) f 21 30
5 audi      a4      2.8  1999  6 auto(l5) f 16 26
6 audi      a4      2.8  1999  6 manual(m5) f 18 26
7 audi      a4      3.1  2008  6 auto(av) f 18 27
8 audi a4 quattro 1.8  1999  4 manual(m5) 4 18 26
9 audi a4 quattro 1.8  1999  4 auto(l5) 4 16 25
10 audi a4 quattro 2.0  2008  4 manual(m6) 4 20 28
# ... with 224 more rows, and 2 more variables: fl <chr>, class <chr>
> ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
+   geom_boxplot()
> ggplot(data = mpg) +
+   geom_boxplot(mapping = aes(x = reorder(class, hwy, FUN = median), y = hwy)) +
+   coord_flip()
> |

```

Environment History

To Console To Source

```

install.packages("tidyverse")
install.packages("tidyverse")
library(tidyverse)
mpg
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
  geom_boxplot()
ggplot(data = mpg) +
  geom_boxplot(mapping = aes(x = reorder(class, hwy, FUN = median), y = hwy)) +
  coord_flip()

```

Files Plots Packages Help Viewer

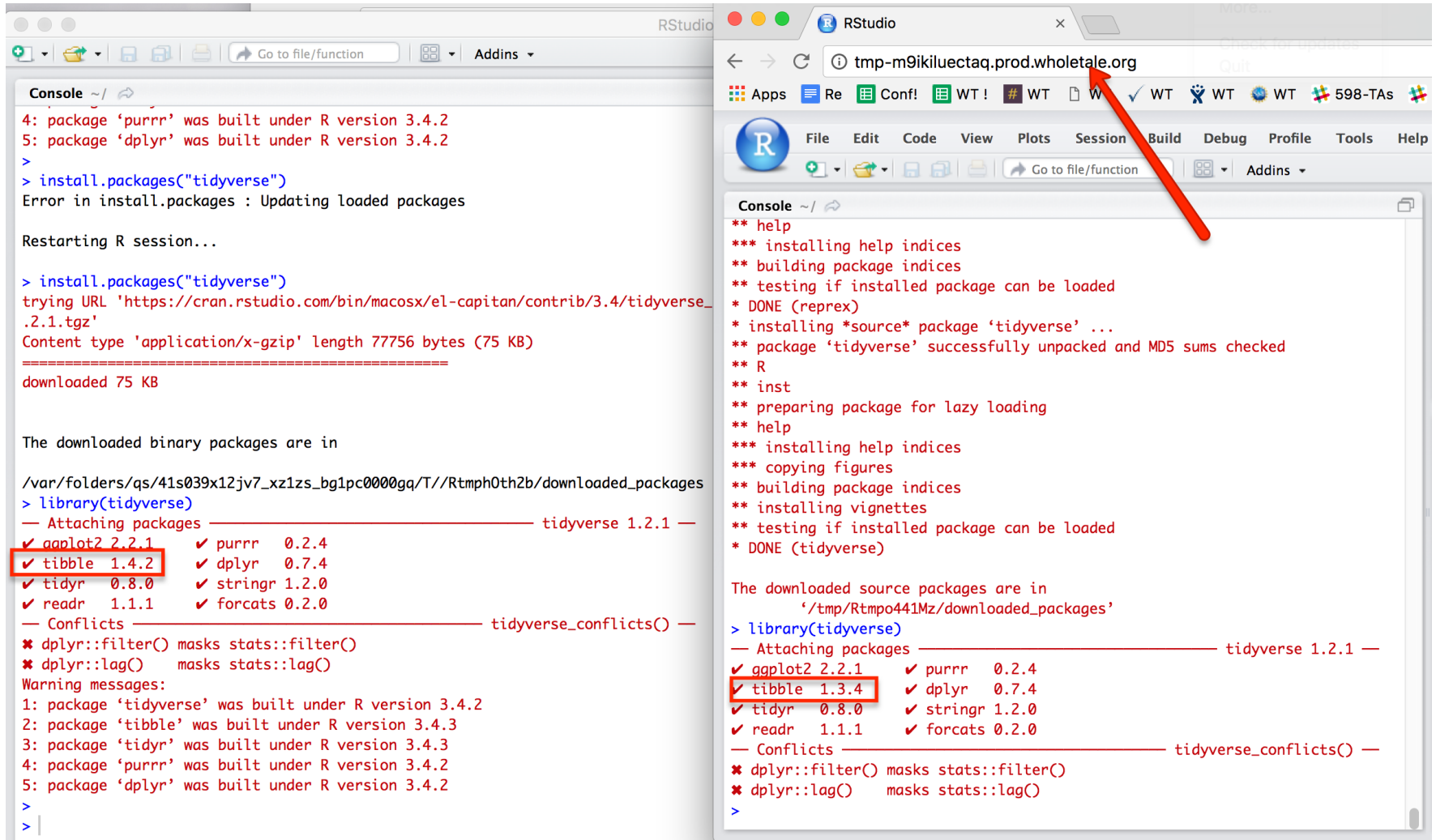
Zoom Export Publish

reorder(class, hwy, FUN = median)

hwy

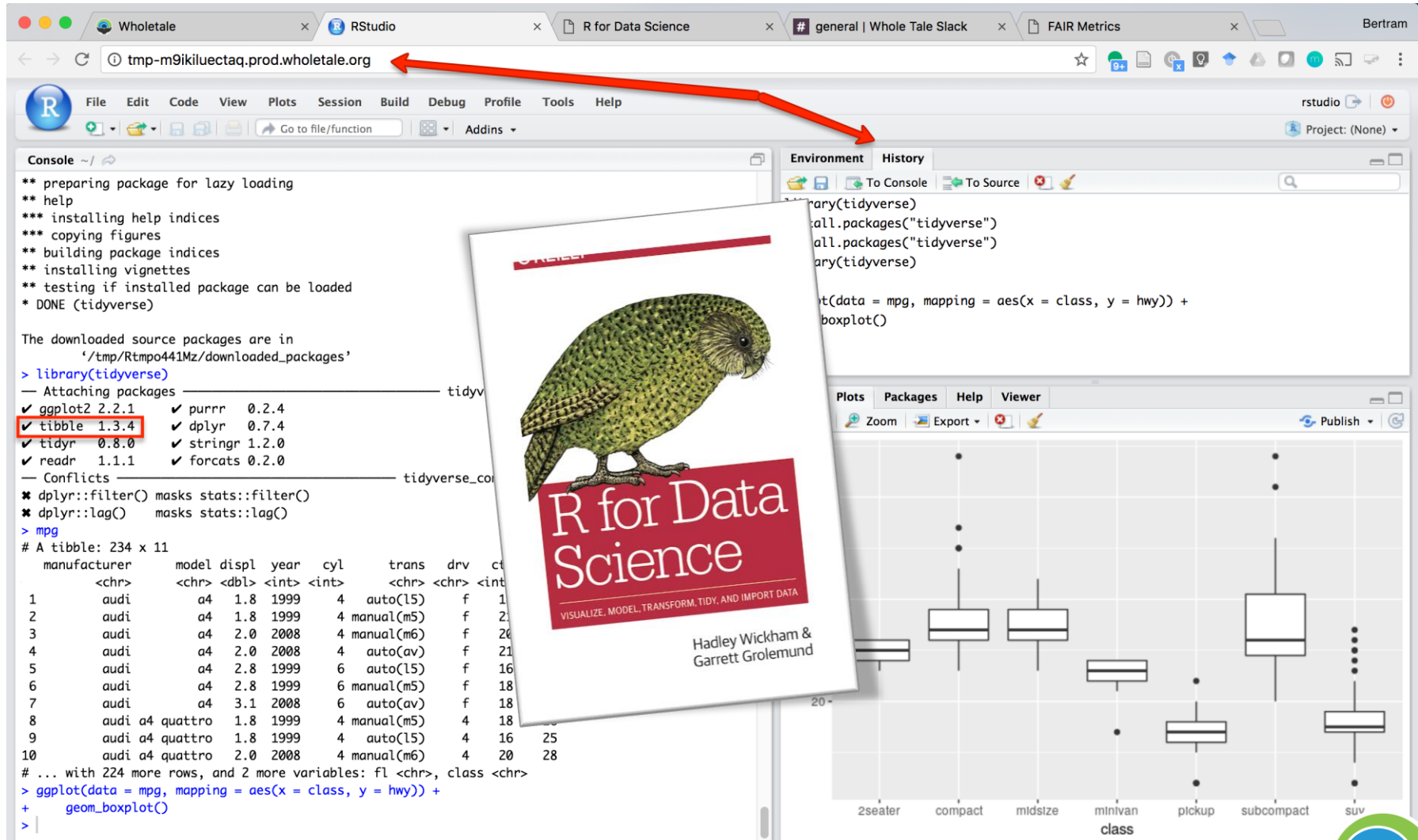


# Running with RStudio: Locally or on WT...



*You're up and running quickly on Whole Tale !!*

# Maybe you just want to use WT to learn *R for Data Science* ...



The screenshot shows the RStudio interface with a browser window at the top displaying 'tmp-m9ikiluctaq.prod.wholetale.org'. A red arrow points from the browser address bar to the RStudio console. The console shows the installation of the tidyverse package and the loading of the mpg dataset. The Environment pane on the right shows the loaded packages and the mpg dataset. The Plots pane at the bottom shows a boxplot of highway mileage (hwy) by car class (class).

**Console Output:**

```
** preparing package for lazy loading
** help
*** installing help indices
*** copying figures
** building package indices
** installing vignettes
** testing if installed package can be loaded
* DONE (tidyverse)

The downloaded source packages are in
  '/tmp/Rtmpo441Mz/downloaded_packages'
> library(tidyverse)
— Attaching packages — tidyv
✓ ggplot2 2.2.1      ✓ purrr  0.2.4
✓ tibble 1.3.4       ✓ dplyr  0.7.4
✓ tidyr  0.8.0       ✓ stringr 1.2.0
✓ readr  1.1.1       ✓ forcats 0.2.0
— Conflicts — tidyverse_co
* dplyr::filter() masks stats::filter()
* dplyr::lag() masks stats::lag()
> mpg
# A tibble: 234 x 11
  manufacturer    model displ  year  cyl  trans      drv  ct
  <chr>          <chr>  <dbl>  <int> <int> <chr>  <chr> <int>
1 audi          a4      1.8    1999   4    auto(l5) f     1
2 audi          a4      1.8    1999   4    manual(m5) f     2
3 audi          a4      2.0    2008   4    manual(m6) f    26
4 audi          a4      2.0    2008   4    auto(av) f    21
5 audi          a4      2.8    1999   6    auto(l5) f    16
6 audi          a4      2.8    1999   6    manual(m5) f    18
7 audi          a4      3.1    2008   6    auto(av) f    18
8 audi a4 quattro 1.8    1999   4    manual(m5) f    18
9 audi a4 quattro 1.8    1999   4    auto(l5) f    16
10 audi a4 quattro 2.0    2008   4    manual(m6) f    20
# ... with 224 more rows, and 2 more variables: fl <chr>, class <chr>
> ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
+   geom_boxplot()
> |
```

**Environment Pane:**

```
library(tidyverse)
all.packages("tidyverse")
all.packages("tidyverse")
library(tidyverse)

mpg
  t(data = mpg, mapping = aes(x = class, y = hwy)) +
  boxplot()
```

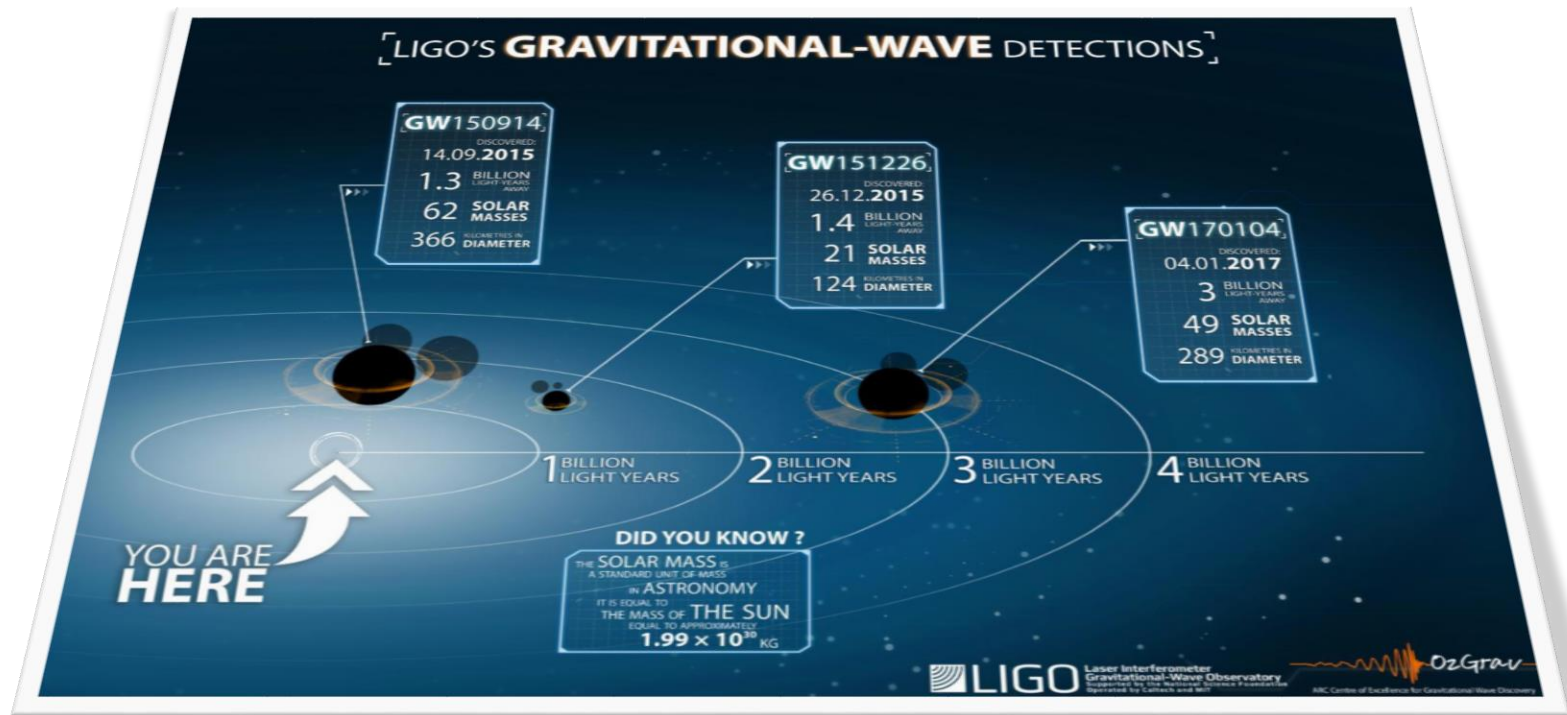
**Plots Pane:**

Boxplot showing highway mileage (hwy) by car class (class). The x-axis categories are 2seater, compact, midsize, minivan, pickup, subcompact, and suv. The y-axis represents highway mileage (hwy).





# An example Tale: LIGO gravitational wave detection (tutorial notebook)



LOSCEvent\_tutorial

tmp-yb6ue5tscpep.prod.whole tale.org/notebooks/work/data/LOSCEvent\_tutorial.ipynb

jupyter LOSCEvent\_tutorial (read only)

File Edit View Insert Cell Kernel Help

Run Code

Running the LIGO tutorial inside of Whole Tale ...

## BINARY BLACK HOLE SIGNALS IN LIGO OPEN DATA

Version 1.63, 2017 Sept 11

Welcome! This IPython notebook (or associated python script LOSCEvent\_tutorial.py ) will go through some typical signal processing tasks on strain time-series data associated with the LIGO Event data releases from the LIGO Open Science Center (LOSC):

- View the tutorial as a [web page, for GW150914](#).
- After setting the desired "eventname" below, you can just run the full notebook.

Questions, comments, suggestions, corrections, etc: email [losc@ligo.caltech.edu](mailto:losc@ligo.caltech.edu)

**This tutorial is intended for educational purposes. The code shown here is not used to produce results papers published by the LIGO Scientific Collaboration, which instead rely on special purpose analysis software packages.**

For publicly available, gravitational-wave software analysis packages that are used to produce LSC and Virgo Collaboration results papers, see <https://losc.ligo.org/software/>.

For technical notes on this tutorial, see [https://losc.ligo.org/bbh\\_tutorial\\_notes/](https://losc.ligo.org/bbh_tutorial_notes/).

### Table of Contents

- [Intro to signal processing](#)
- [Download the data](#)
- [Set the event name to choose event and the plot type](#)
- [Read in the data](#)
- [Plot the ASD](#)
- [Binary Neutron Star detection range](#)

LOSC\_Event\_tutorial x

tmp-yb6ue5tscepeprod.wholetale.org/notebooks/work/data/LOSC\_Event\_tutorial.ipynb

jupyter LOSC\_Event\_tutorial (read only) Logout

File Edit View Insert Cell Kernel Help

Not Trusted Python 3

Run Code

## Read in the data

We will make use of the data, and waveform template, defined above.

```
In [5]: #-----
# Load LIGO data from a single file.
# FIRST, define the filenames fn_H1 and fn_L1, above.
#-----
try:
    # read in data from H1 and L1, if available:
    strain_H1, time_H1, chan_dict_H1 = r1.loadaddata(fn_H1, 'H1')
    strain_L1, time_L1, chan_dict_L1 = r1.loadaddata(fn_L1, 'L1')
except:
    print("Cannot find data files!")
    print("You can download them from https://losc.ligo.org/s/events/"+eventname)
    print("Quitting.")
    quit()
```

## Data Gaps

**NOTE** that in general, LIGO strain time series data has gaps (filled with NaNs) when the detectors are not taking valid ("science quality") data. Analyzing these data requires the user to [loop over "segments"](#) of valid data stretches.

In this tutorial, for simplicity, we assume there are no data gaps - this will not work for all times! See the [notes on segments](#) for details.

## First look at the data from H1 and L1

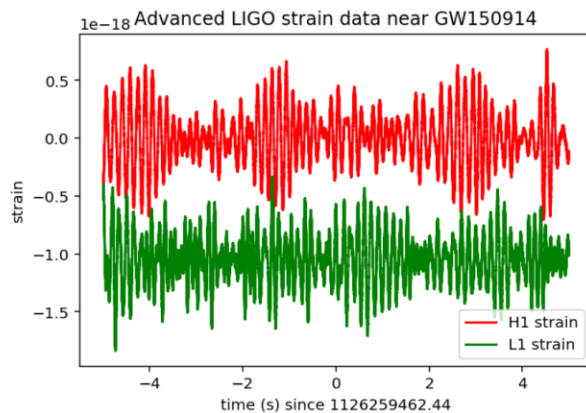
```
In [6]: # both H1 and L1 will have the same time vector, so:
time = time_H1
# the time sample interval (uniformly sampled!)
dt = time[1] - time[0]
```

```
In [7]: # plot +/- deltata seconds around the event:
# index into the strain time series for this time interval:
deltat = 5
indx = np.where((time >= tevent-deltat) & (time < tevent+deltat))
print(tevent)

if make_plots:
    plt.figure()
    plt.plot(time[indx]-tevent, strain_H1[indx], 'r', label='H1 strain')
    plt.plot(time[indx]-tevent, strain_L1[indx], 'g', label='L1 strain')
    plt.xlabel('time (s) since '+str(tevent))
    plt.ylabel('strain')
    plt.legend(loc='lower right')
    plt.title('Advanced LIGO strain data near '+eventname)
```

Executed  
Step

1126259462.44



Code and interleaved output ..

```

# loop over the detectors
dets = ['H1', 'L1']
for det in dets:
    if det is 'L1': sspec = Pxx_L1.copy()
    else:           sspec = Pxx_H1.copy()
    sspecfr = sspec[fr]
    # compute "inspiral horizon distance" for optimally oriented binary; FINDCHIRP Eqn D2:
    D_BNS = np.sqrt(4.*np.sum(htilda2/sspecfr)*df)/SNRdet
    # and the "inspiral range", averaged over source direction and orientation:
    R_BNS = D_BNS/Favg
    print(det+' BNS inspiral horizon = {0:.1f} Mpc, BNS inspiral range    = {1:.1f} Mpc'.format(D_BNS,R_BNS))

```

```

H1 BNS inspiral horizon = 169.4 Mpc, BNS inspiral range    = 74.8 Mpc
L1 BNS inspiral horizon = 147.1 Mpc, BNS inspiral range    = 64.9 Mpc

```

Code and output ...

## BBH range is >> BNS range!

NOTE that, since mass is the source of gravity and thus also of gravitational waves, systems with higher masses (such as the binary black hole merger GW150914) are much "louder" and can be detected to much higher distances than the BNS range. We'll compute the BBH range, using a template with specific masses, below.

... explanations!

## Whitening

From the ASD above, we can see that the data are very strongly "colored" - noise fluctuations are much larger at low and high frequencies and near spectral lines, reaching a roughly flat ("white") minimum in the band around 80 to 300 Hz.

We can "whiten" the data (dividing it by the noise amplitude spectrum, in the fourier domain), suppressing the extra noise at low frequencies and at the spectral lines, to better see the weak signals in the most sensitive band.

Whitening is always one of the first steps in astrophysical data analysis (searches, parameter estimation). Whitening requires no prior knowledge of spectral lines, etc; only the data are needed.

To get rid of remaining high frequency noise, we will also bandpass the data.

The resulting time series is no longer in units of strain; now in units of "sigmas" away from the mean.

We will plot the whitened strain data, along with the signal template, after the matched filtering section, below.

Ludäscher: Whole-Tale++



```
plt.grid()
plt.xlabel('time (s)')
plt.ylabel('v/c')
#plt.title(eventname+' template v/c')
```

/opt/conda/lib/python3.6/site-packages/ipykernel\_launcher.py:5: DeprecationWarning: object of type <class 'float'> cannot be safely interpreted as an integer.  
"""

Properties of waveform template in GW150914\_4\_template.hdf5

Waveform family = b'lalsim.SEOBNRv2'

Masses = 41.74, 29.24 Msun

Mtot = 70.98 Msun, mfinal = 67.43 Msun

Spins = 0.35, -0.77

Freq at inband, peak = 43.05, 169.84 Hz

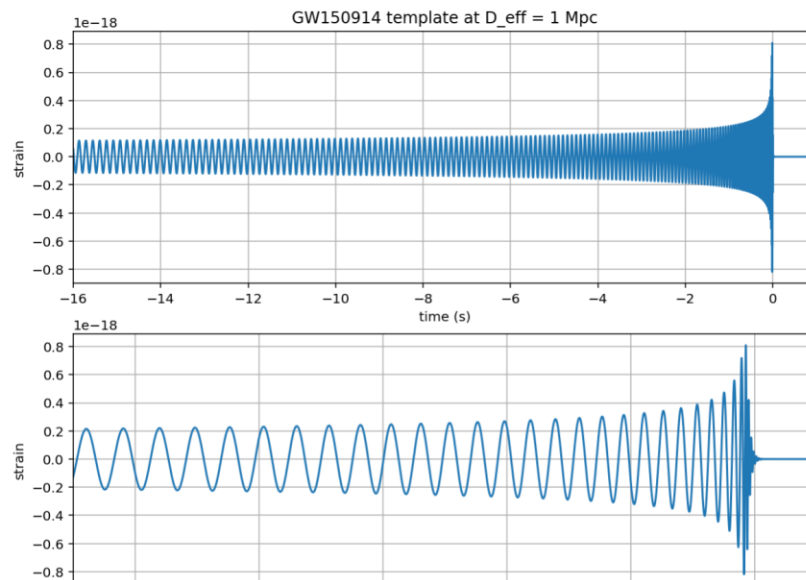
Time at inband, peak = -0.08, -0.02 s

Duration (s) inband-peak = 0.06 s

N\_cycles inband-peak = 4

v/c at peak = 0.57

Radius of final BH = 199 km



noticeably more high-pitched, and the signal will be easier to hear.

```
In [19]: # function that shifts frequency of a band-passed signal
def reqshift(data,fshift=100,sample_rate=4096):
    """Frequency shift the signal by constant
    """
    x = np.fft.rfft(data)
    T = len(data)/float(sample_rate)
    df = 1.0/T
    nbins = int(fshift/df)
    # print T,df,nbins,x.real.shape
    y = np.roll(x.real,nbins) + 1j*np.roll(x.imag,nbins)
    y[0:nbins]=0.
    z = np.fft.irfft(y)
    return z

# parameters for frequency shift
fs = 4096
fshift = 400.
speedup = 1.
fss = int(float(fs)*float(speedup))

# shift frequency of the data
strain_H1_shifted = reqshift(strain_H1_whitenbp,fshift=fshift,sample_rate=fs)
strain_L1_shifted = reqshift(strain_L1_whitenbp,fshift=fshift,sample_rate=fs)

# write the files:
write_wavfile("../"+eventname+"_H1_shifted.wav",int(fs), strain_H1_shifted[indxd])
write_wavfile("../"+eventname+"_L1_shifted.wav",int(fs), strain_L1_shifted[indxd])

# and the template:
template_p_shifted = reqshift(template_p_smooth,fshift=fshift,sample_rate=fs)
write_wavfile("../"+eventname+"_template_shifted.wav",int(fs), template_p_shifted[indxt])
```

### Listen to the frequency-shifted template and data

```
In [20]: fna = "../"+eventname+"_template_shifted.wav"
print(fna)
Audio(fna)
```

../GW150914\_template\_shifted.wav

Out[20]:



From data to figures  
... to sound!

```
GPS start, GPS stop and length of all data in this file = 1126259446.0 1126259477.9997559 131072
Number of segments with DQflag CBC_CAT3 = 1
GPS start, GPS stop and length of segment 0 in this file = 1126259446.0 1126259477.9997559 131072
Number of segments with DQflag NO_CBC_HW_INJ = 1
GPS start, GPS stop and length of segment 0 in this file = 1126259446.0 1126259477.9997559 131072
```

## Comments on sampling rate

LIGO data are acquired at 16384 Hz ( $2^{14}$  Hz). Here, we have been working with data downsampled to 4096 Hz, to save on download time, disk space, and memory requirements.

This is entirely sufficient for signals with no frequency content above  $f_{\text{Nyquist}} = f_s/2 = 2048$  Hz, such as signals from higher-mass binary black hole systems; the frequency at which the merger begins (at the innermost stable circular orbit) for equal-mass, spinless black holes is roughly  $1557 \text{ Hz} \cdot (2.8/M_{\text{tot}})$ , where 2.8 solar masses is the total mass of a canonical binary neutron star system.

If, however, you are interested in signals with frequency content above 2048 Hz, you need the data sampled at the full rate of 16384 Hz.

## Construct a csv file containing the whitened data and template

```
In [23]: # time vector around event
times = time-tevent
# zoom in on [-0.2,0.05] seconds around event
irange = np.nonzero((times >= -0.2) & (times < 0.05))
# construct a data structure for a csv file:
dat = [times[irange], strain_H1_whitenbp[irange], strain_L1_whitenbp[irange],
        template_H1[irange], template_L1[irange] ]
datcsv = np.array(dat).transpose()
# make a csv filename, header, and format
fncsv = "../"+eventname+'_data.csv'
headcsv = eventname+' time-'+str(tevent)+' \
            (s),H1_data_whitened,L1_data_whitened,H1_template_whitened,L1_template_whitened'
fmtcsv = ",".join(["%10.6f"] * 5)
np.savetxt(fncsv, datcsv, fmt=fmtcsv, header=headcsv)

print("Wrote whitened data to file {}".format(fncsv))

Wrote whitened data to file ../GW150914_data.csv
```

Save / export data ..

# New & Upcoming Features in WT ...

- Add your **own Frontends** (e.g. OpenRefine, ..)
- Persistent, shared or personal files: /work/{ home, data }
- WT **“Derived Tales”**:
  - take a tale; modify it to your liking; and publish as a derived work
- WT **“Take-Out”**:
  - Want to run your tales elsewhere?
  - *Take-out* your tale and run on your on (or cloud) platform
- WT **“Scale-Out”**:
  - If the WT-dashboard isn't enough → run your own WT system!
- WT **Provenance support**:
  - ... via DataONE provenance tools, ProvONE model (W3C PROV extension)
  - ... via **YesWorkflow**
- **Interest in joining a WT Biodiversity Informatics Working Group!?**
  - We already have: archaeology & ecology, astronomy, materials science
  - Your input wanted! (is WT developing something useful for you?)
  - Try out WT, create some examples (in R, Python, ...) and provide feedback!
  - => fund **a summer intern!**

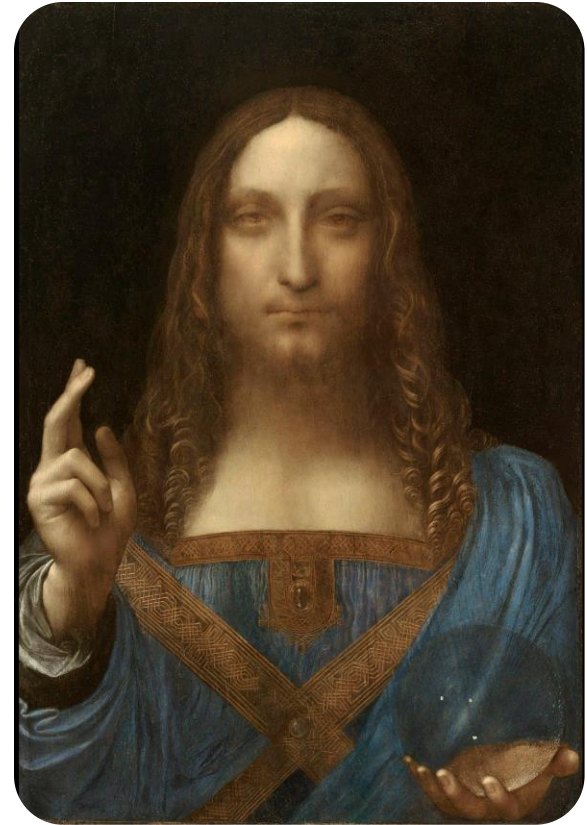
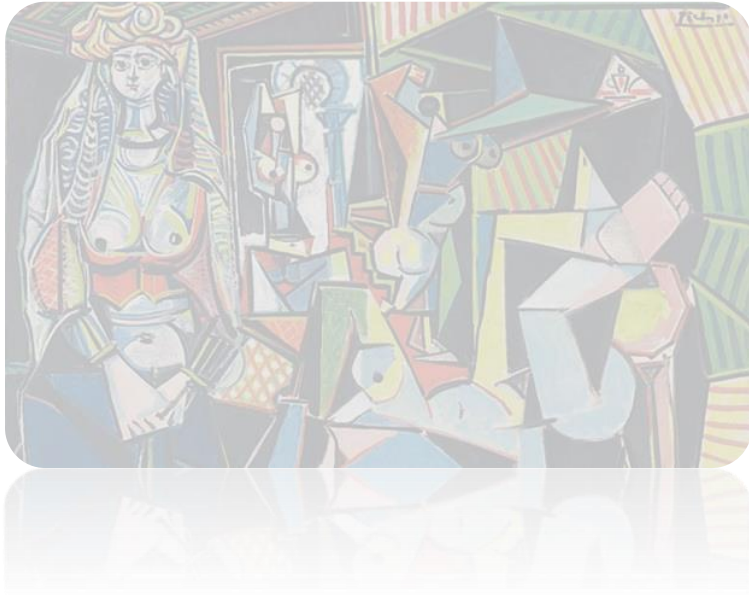
# Provenance (*Lineage*) matters ...



- One of these sold for \$180M, the other one for \$22K (*but could be worth more ... definitely maybe ...*)
- Which one would you like to own?



# Provenance (*Lineage*) matters ...



- One of these sold for \$180M, the other one for ...
- ... **\$450M !!!**

# Provenance is: *keeping records ...*

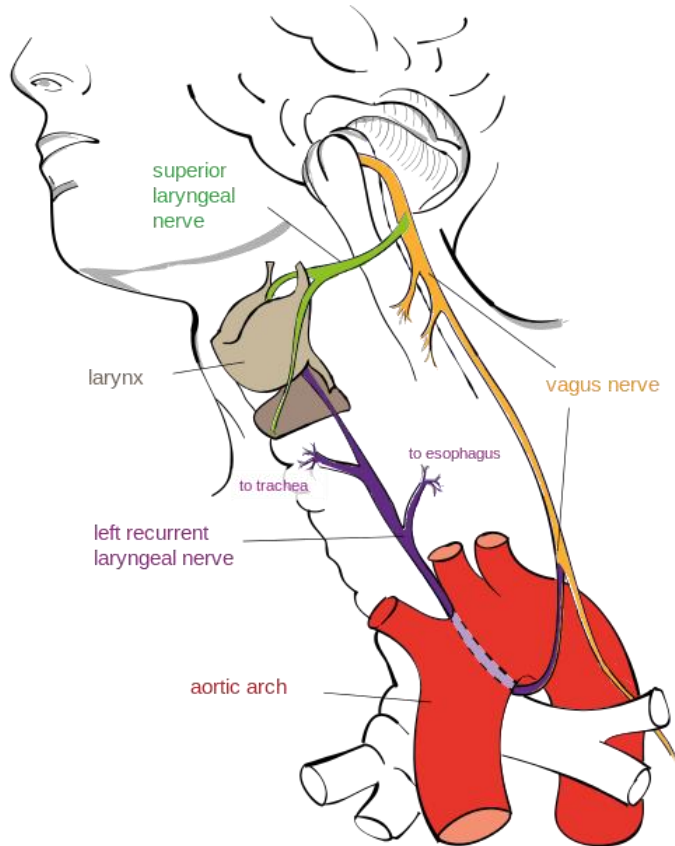


- Grand Canyon's rock layers are a record of the early **geologic history** of North America. The ancestral puebloan granaries at Nankoweap Creek tell archaeologists about more recent **human history**. (By Drenaline, licensed under CC BY-SA 3.0)
- Not shown: **computational archaeologists** reconstructing past climate from multiple tree-ring databases → computational provenance is key for **transparency** & **reproducibility**



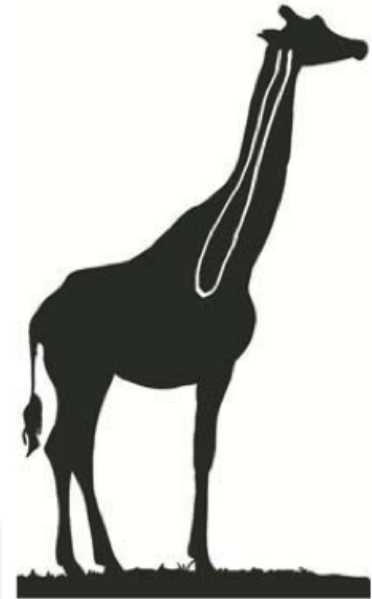
# ... and provenance is:

## *Understanding* what happened!



Author: Jkwchui (Based on drawing by Truth-seeker2004)

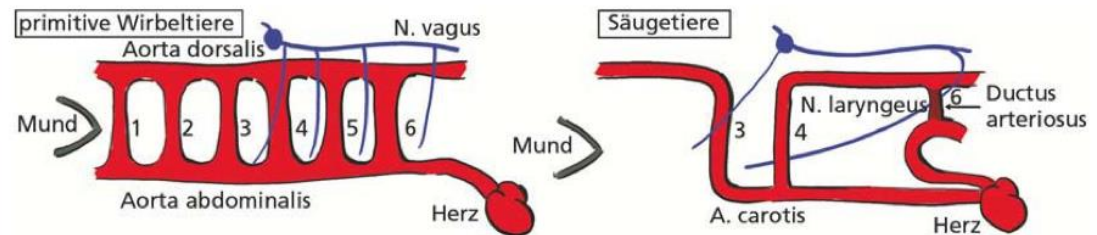
Zrzavý, Jan, David Storch, and Stanislav Mihulka. *Evolution: Ein Lese-Lehrbuch*. Springer-Verlag, 2009.



5.17 Suboptimale evolutionäre Konstruktionslösung:

334

5 Adaptation

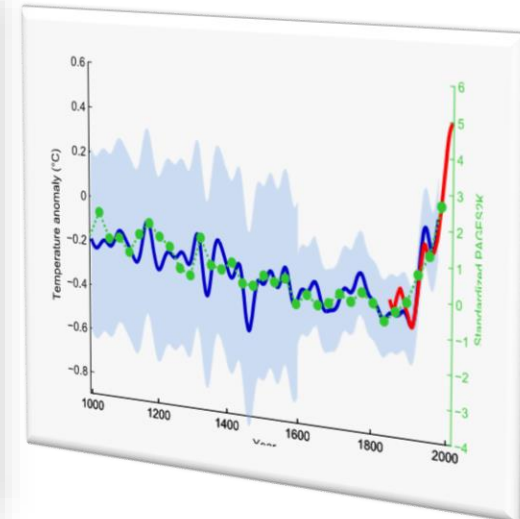
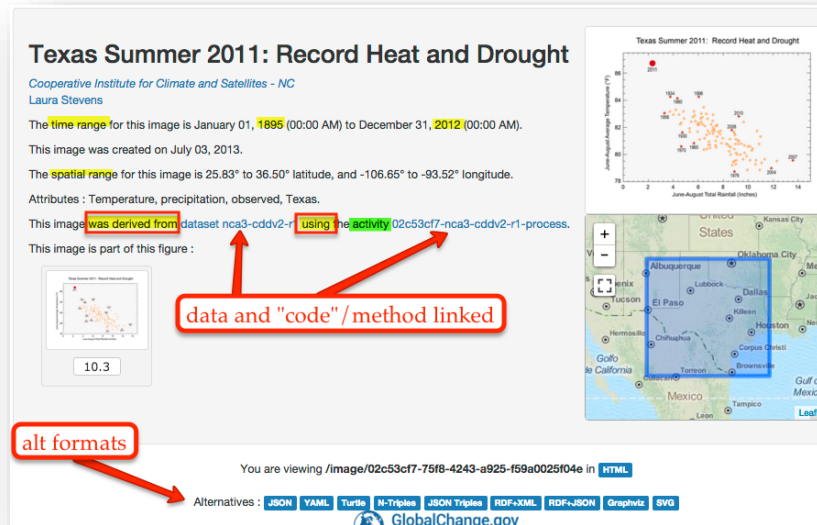
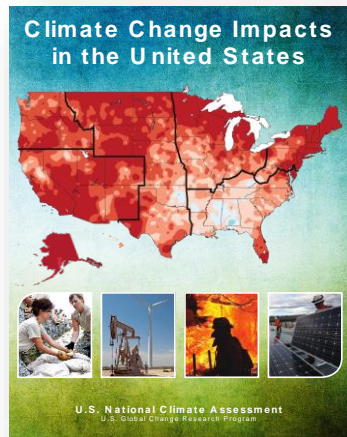


5.16 Evolution der Schleife des rückläufigen Kehlkopfners (Nervus laryngeus recurrens) der Wirbeltiere. Dieser Nerv stellt den vierten Ast des Nervus vagus dar. Bei ursprünglichen Wassertieren sandte der Vagusnerv seine Äste zu den Kiemenarterien, die die Bauch- und die Rückenaorta verbanden. Während der Phylogenese der Wirbeltiere haben sich allerdings die Kiemenbögen und mit ihnen auch die Kiemenarterien verändert und das Herz wurde nach kaudal verschoben. Aus der sechsten Arterie wurde bei den Säugetieren der Ductus arteriosus; der vierte Ast des Vagus, der heute den Kehlkopf (Larynx) innerviert, liegt stets *hinter* der ehemaligen sechsten Arterie, also hinter dem Ductus arteriosus. Daher führt dieser Nerv vom Gehirn aus nach hinten, windet sich unter dem Ductus hindurch und kehrt nach vorne zurück, um den Larynx zu innervieren.



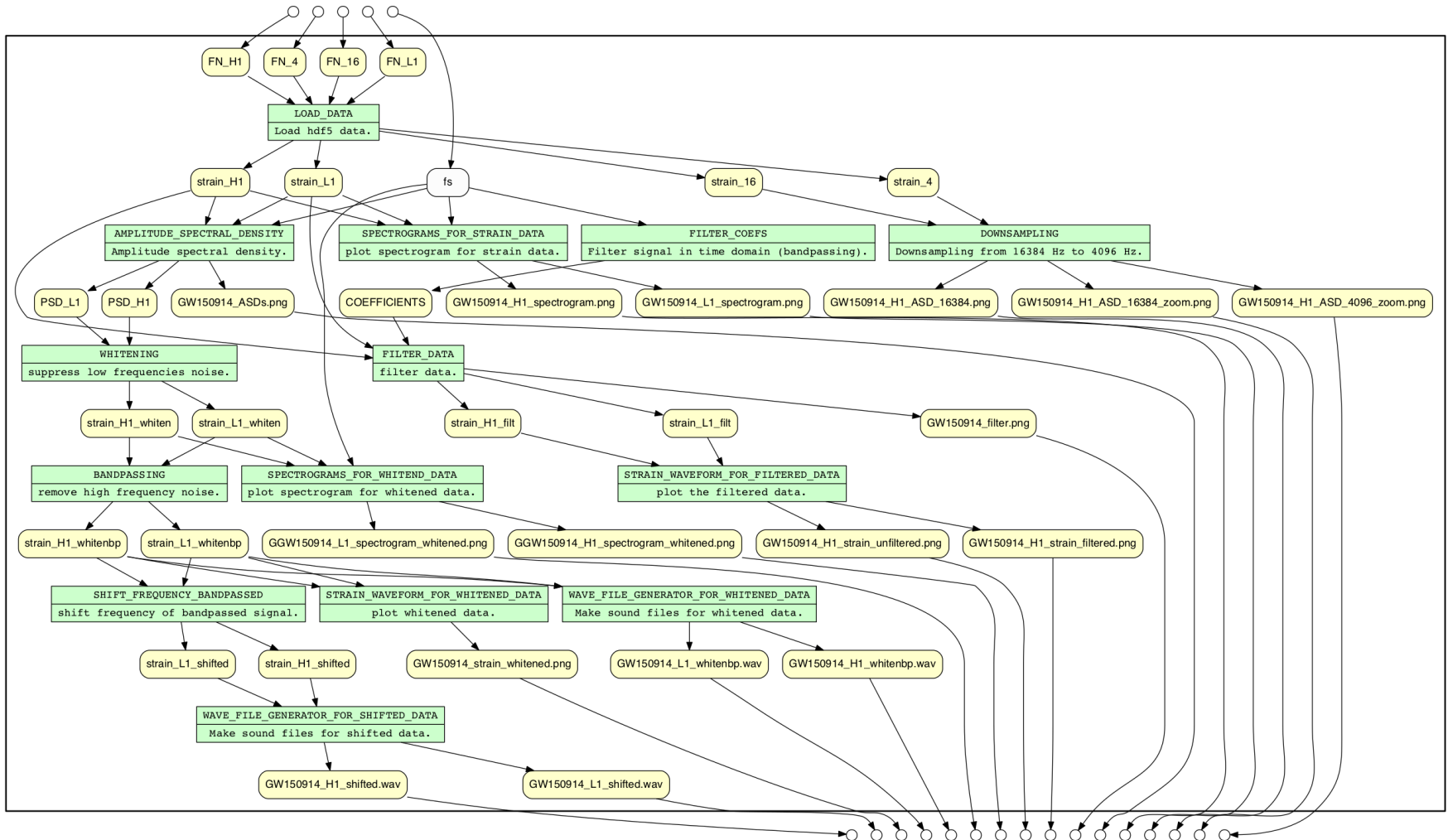
# Computational Provenance ...

- Origin, **processing history** of artifacts
  - **data products**, figures, ...
  - also: underlying **workflow**
- ➔ understand **methods**, **dataflow**, and **dependencies**



# YesWorkflow: How does the LIGO script produce its results??

GRAVITATIONAL\_WAVE\_DETECTION



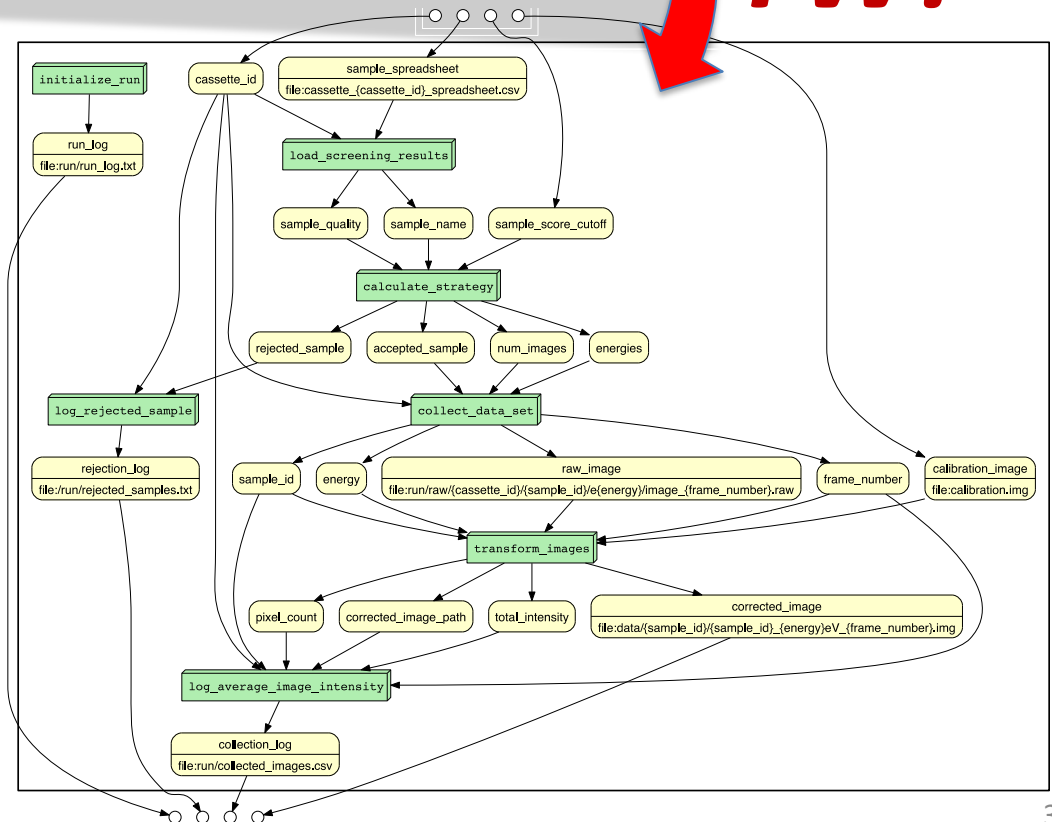
# YesWorkflow: Prospective & Retrospective Provenance ... (almost) for free!

```
# @BEGIN collect_data_set
# @PARAM cassette_id @PARAM accepted_sample @PARAM num_images @PARAM energies
# @OUT sample_id @OUT energy @OUT frame_number
# @OUT raw_image_path @AS raw_image
# @URI file:run/raw/{cassette_id}/{sample_id}/e{energy}/image-{frame_number}.raw
run_log.write("Collecting data set for sample {0}".format(accepted_sample))
sample_id = accepted_sample
for energy, frame_number, intensity, raw_image_path in collect_next_image(
    cassette_id, sample_id, num_images, energies,
    "run/raw/{cassette_id}/{sample_id}/e{energy}/image-{frame_number}.raw"):
    run_log.write("Collecting image {0}".format(raw_image_path))
# @END collect_data_set
```

**@BEGIN .. @END ..**  
**@IN .. @OUT ..**  
**@URI .. @LOG ..**

**YW!**

- **YW annotations** in a (Python, R, ...) script **recreate a workflow view** from the script ...

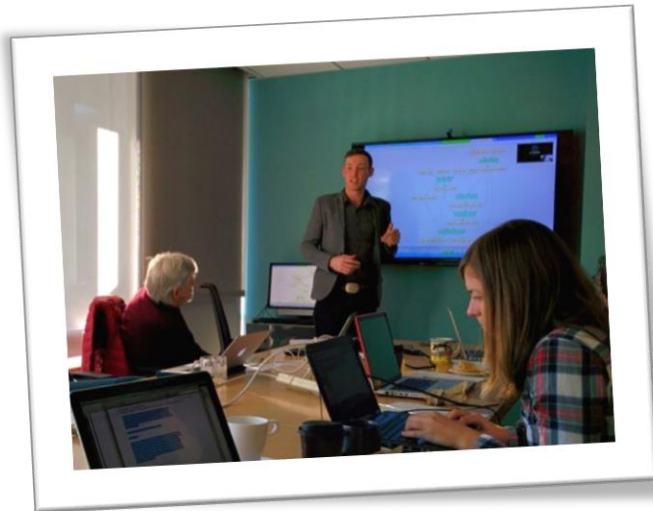


# Paleoclimate Reconstruction (openSKOPE.org)

- ... explained using YesWorkflow!

Kyle B., (computational) archaeologist:

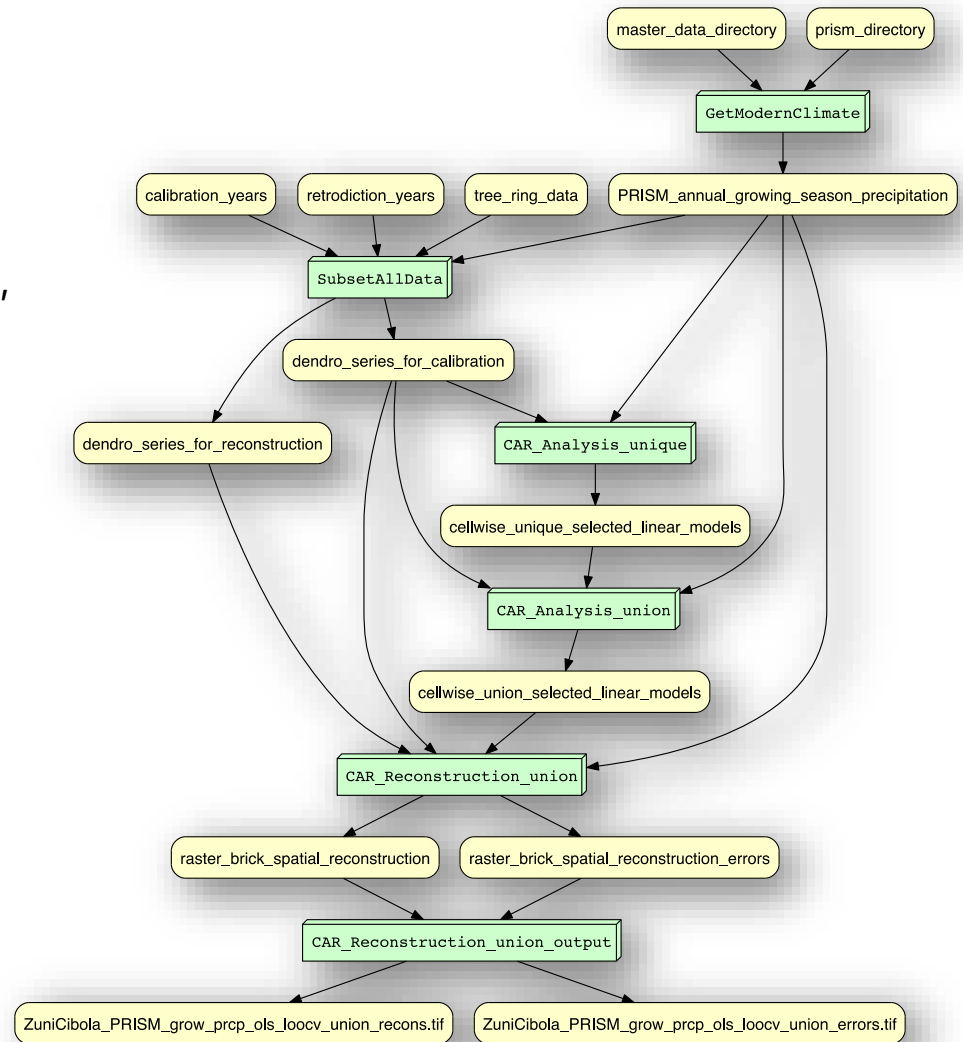
*"It took me about 20 minutes to comment. Less than an hour to learn and YW-annotate, all-told."*



SKOPE + Kurator  
+ DataONE  
Data Observation Network for Earth

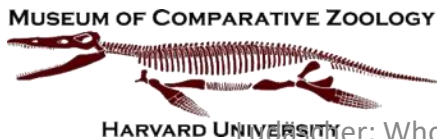
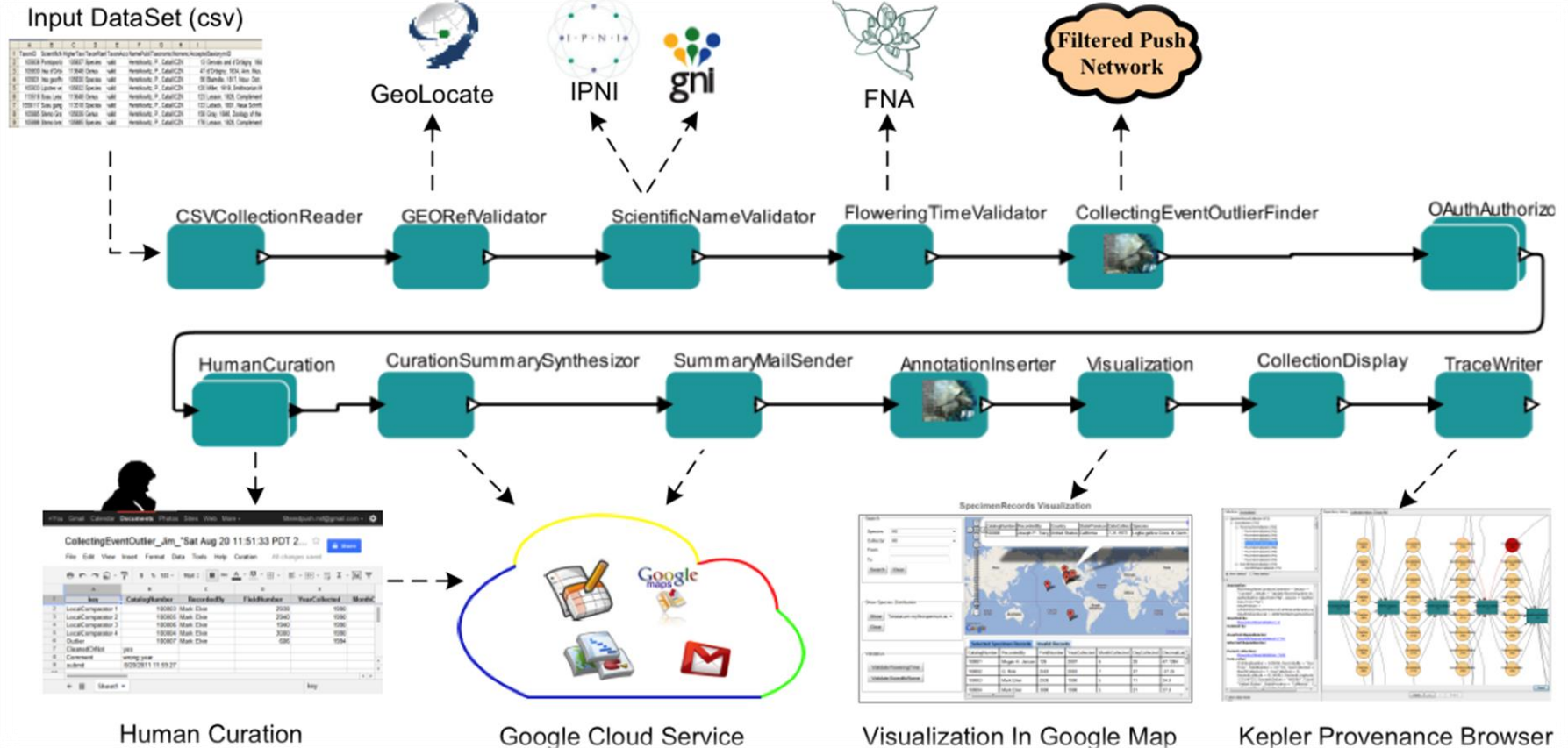
=> YesWorkflow.org

Ludäscher: Whole-Tale++



# Data Curation Workflows

(Filtered-Push ... Kepler ... **Kurator** projects)





<http://kurator.acis.ufl.edu/kurator-web/>

Kurator provides scientific workflow tools for data quality improvement of natural history collections and other biodiversity data. Kurator Web is a set of a user friendly web interface to configure and launch curation workflows while maintaining provenance. Kurator-Akka and the Kurator YesWorkflow data curation software and code are available on [GitHub](#). For more information about Kurator, please visit our [wiki](#).

**File Aggregator:**

Aggregates two files into one file.

**Date Validator:**

Validates event date fields and fills in missing dates from atomic event date fields.

### Georeference Validator:

Performs validation of the georeference fields and fills in or transposes missing or inconsistent coordinates.

### Vocabulary Maker:

Creates a vocabulary file with fields for the original values, the standard values, and vetted values.

**Controlled Field Assessor:**

Creates a report of counts of distinct geographic values and provides recommended values.

### Field Value Counter:

Creates a report of counts of distinct values and recommended values for values that are not standard.

### Property Parser:

Parses "dynamicProperties" field in a DarwinCore-Archive spreadsheet and creates separate fields for each value.

**Darwinizer:**

Creates a new file with as many field names standardized to Darwin Core as possible.

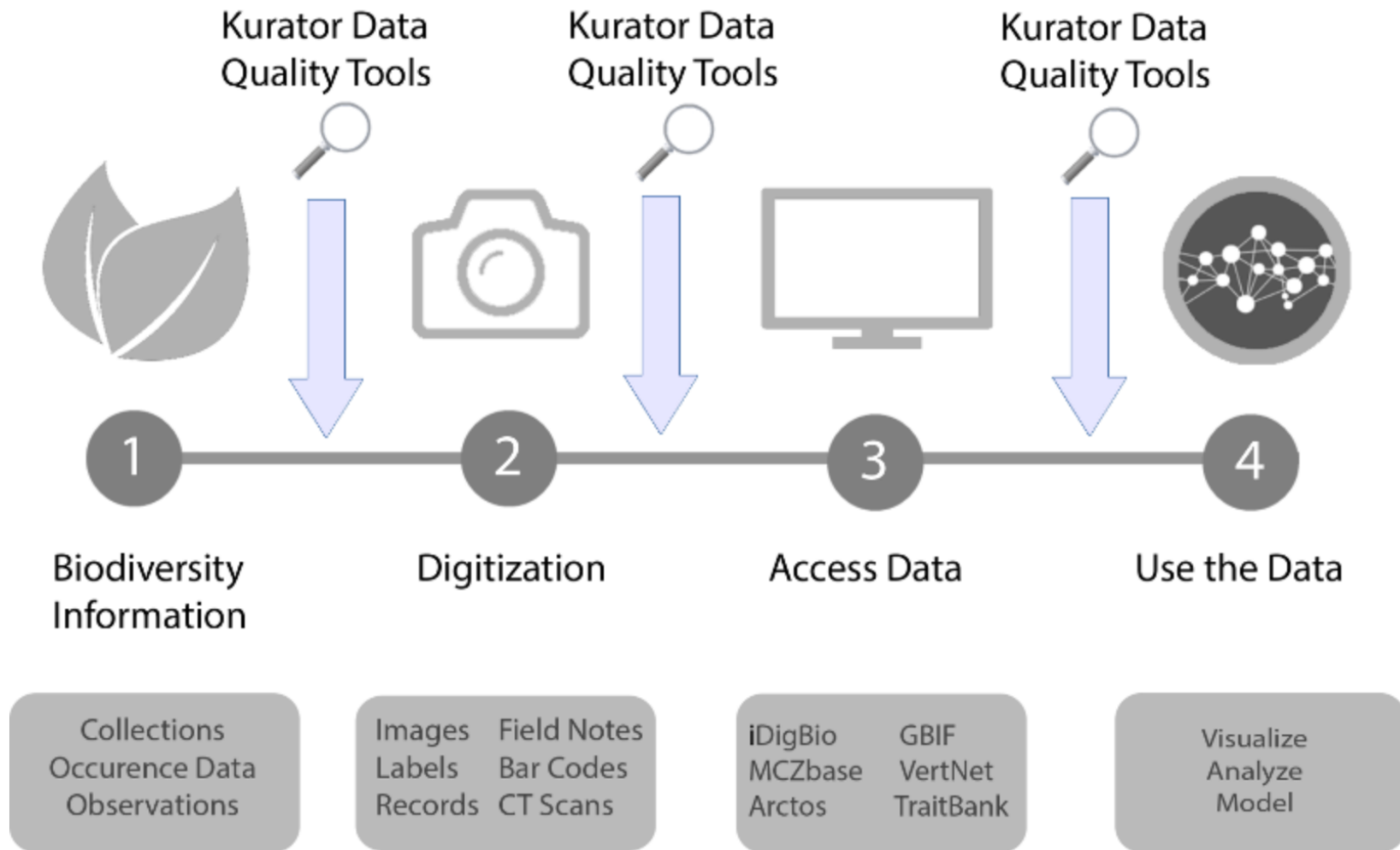
### Geography Cleaner:

Creates a new occurrences file with standardized geography and original geography saved in new fields.

**Geography Assessor:**

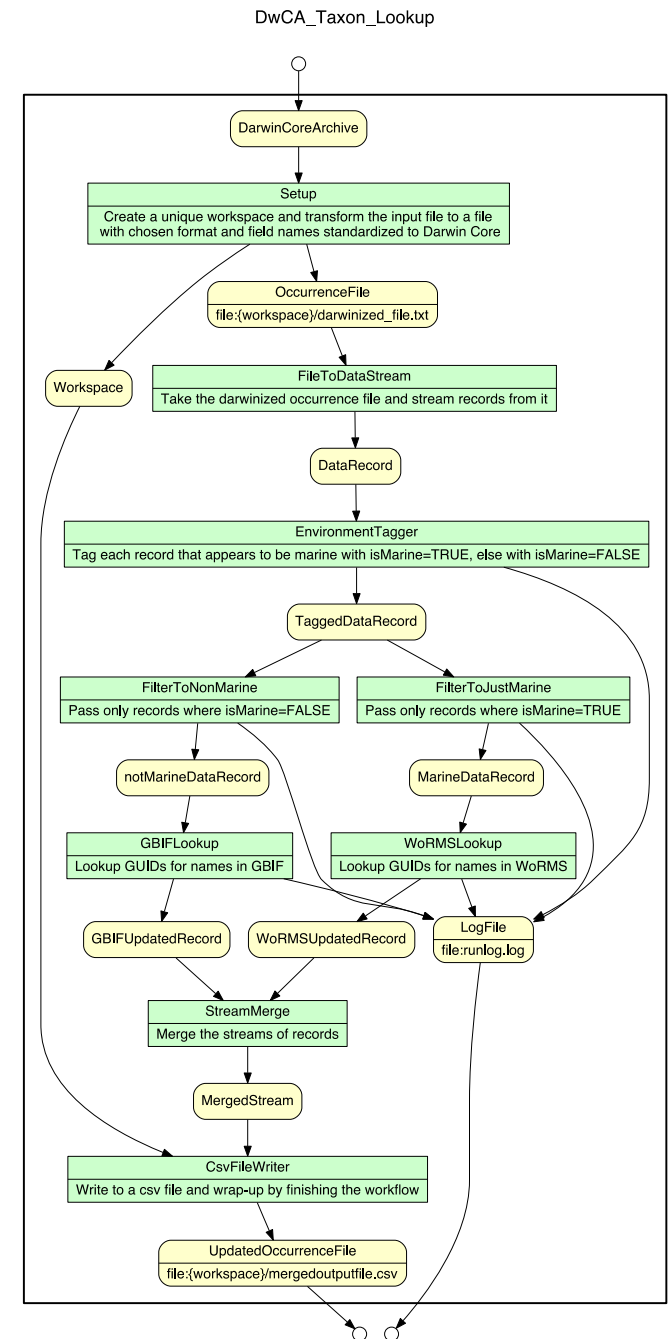
Creates file containing the recommendations to standardize distinct combinations of higher geography.

# How Kurator fits in to the biodiversity data workflow



# DwCA Taxon Lookup Workflow

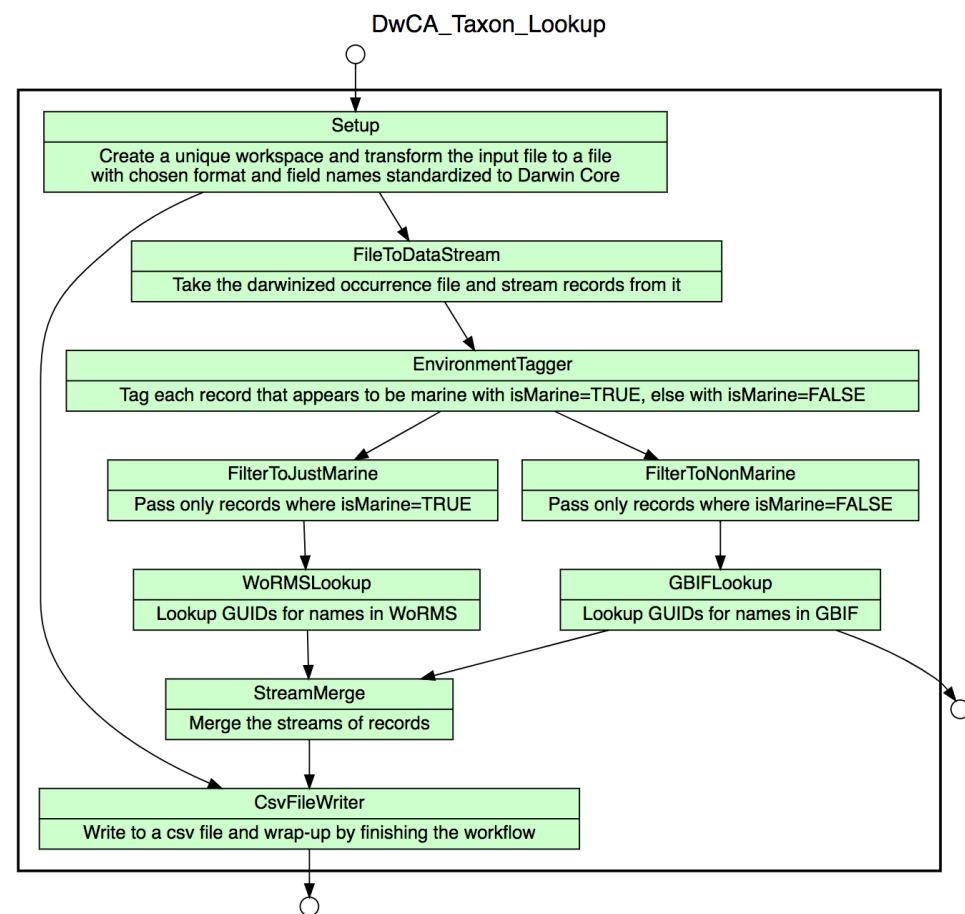
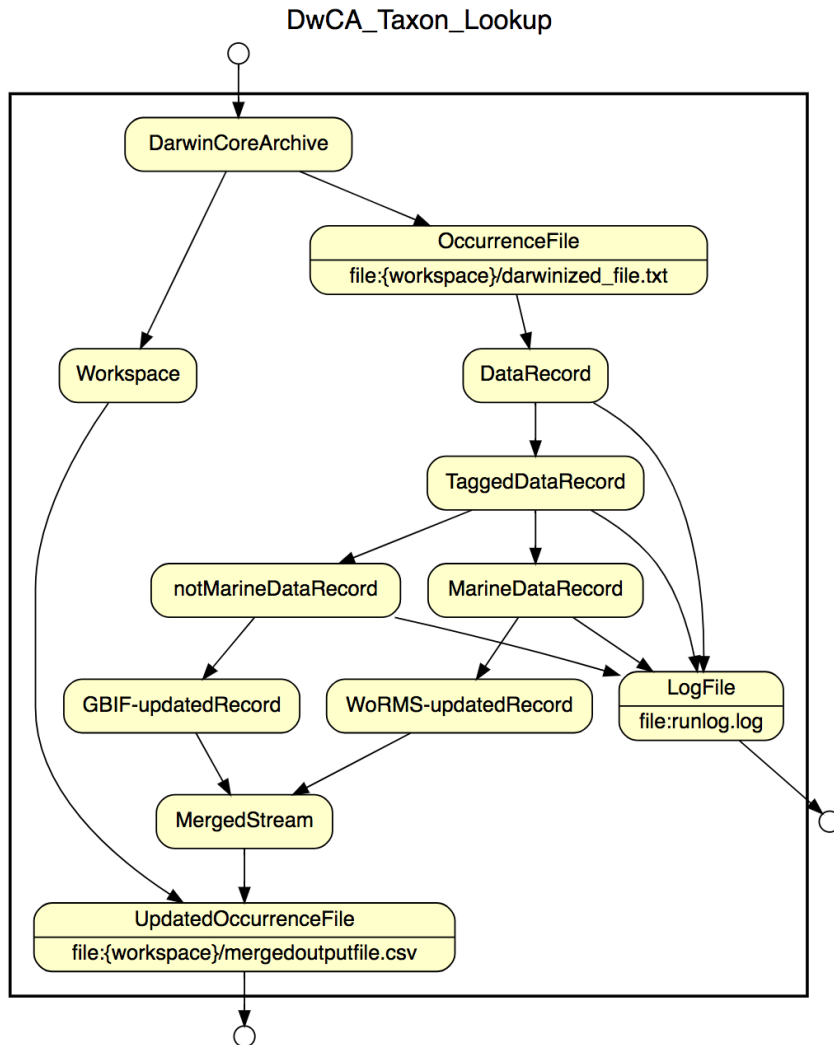
- Declare *inputs*, *outputs*, and *steps* of a script (or wf) with *YW annotations* to ...
  - communicate **provenance graphically** (via graphviz)
  - **combine** different forms of provenance
  - **query** provenance
- Simple **YW annotations** in comments:
  - **@BEGIN Step, @END Step**
  - **@IN Data, @OUT Data**
  - **@URI Template, @LOG Pattern**





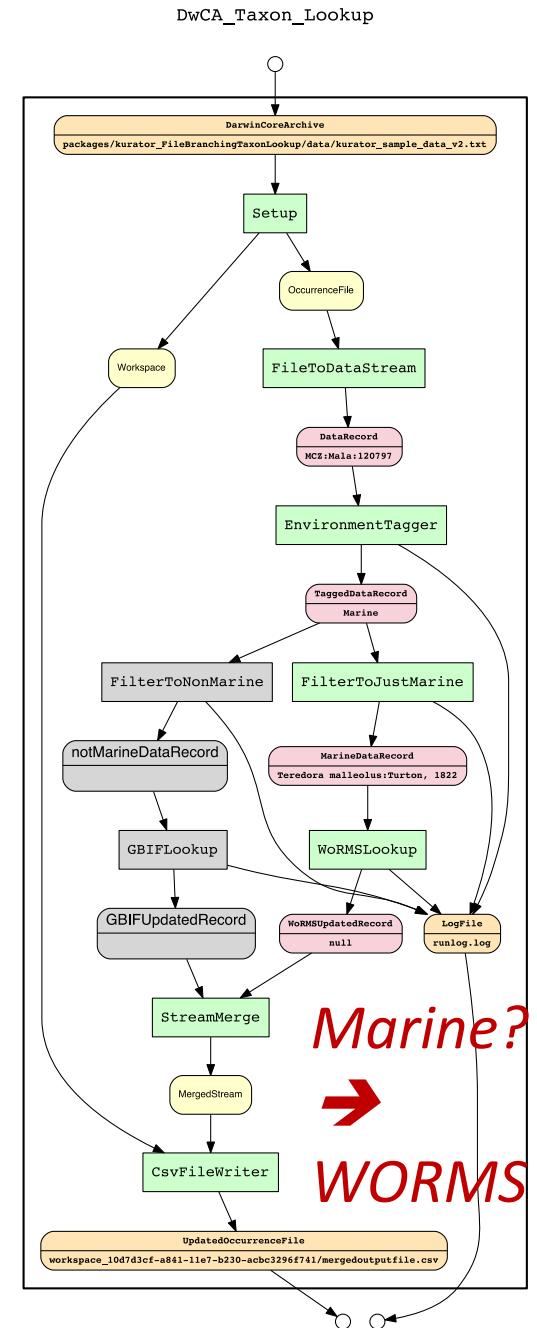
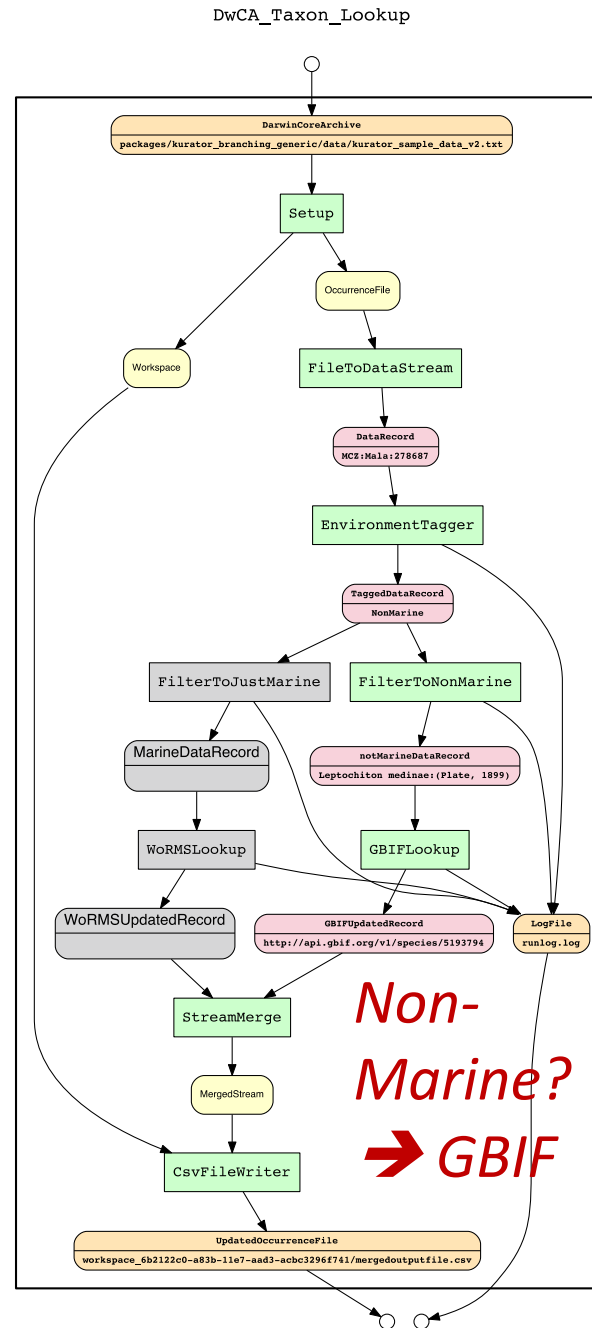
# Taxon Lookup Workflow:

## *Data View and Process View*



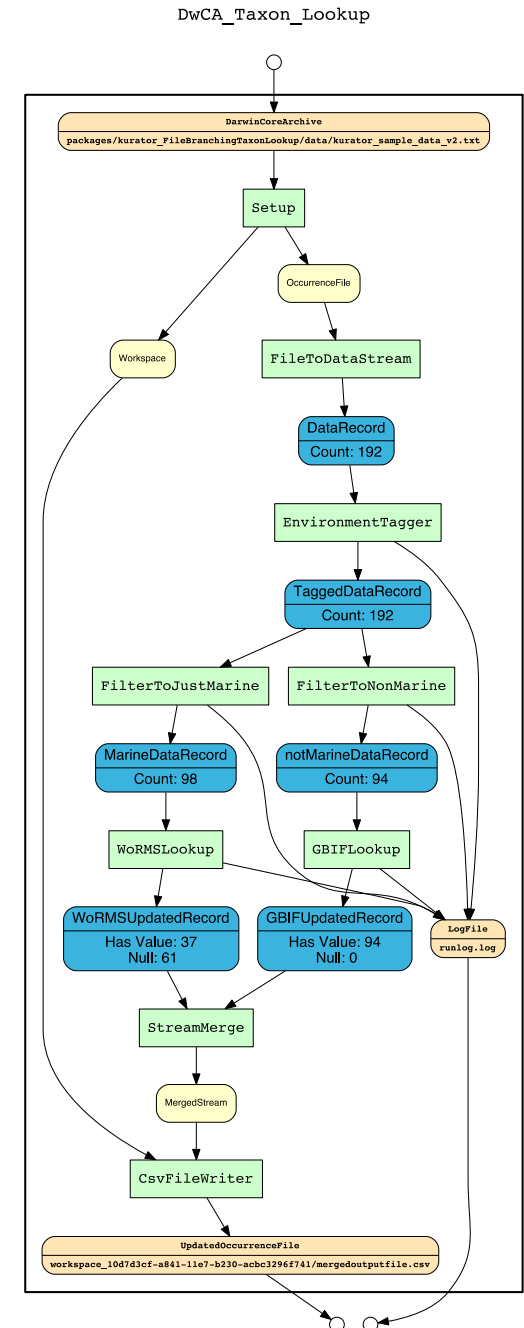
# The story of two individual records

- One took the **GBIF** route, while ...
- ... the other went all **WORMS**!



# The aggregate story ..

- *How many records were observed as inputs or outputs of workflow steps?*
- *Were there any NULL values? How many?*



# YesWorkflow Summary

- Lightweight **YW annotations** can be added easily to your scripts to reap **workflow benefits**

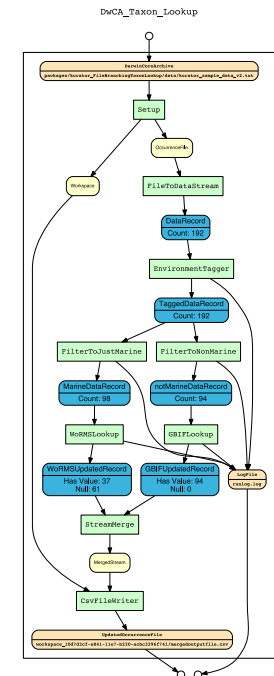
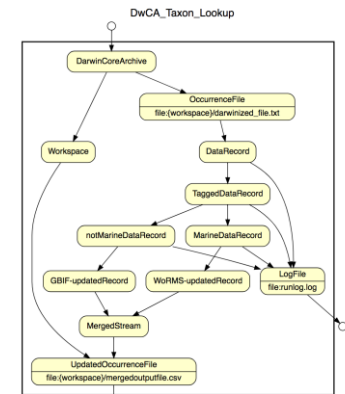
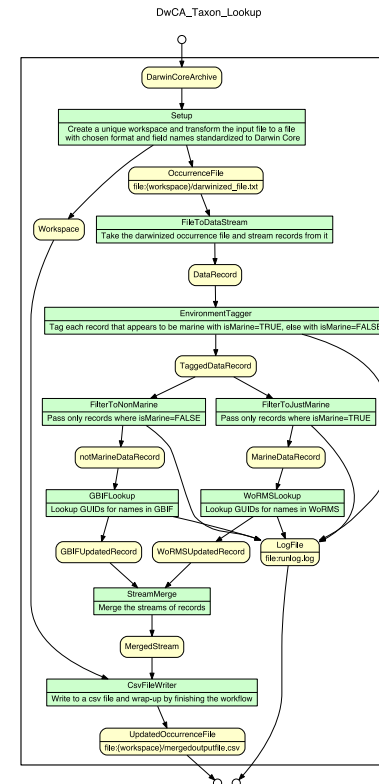
- **Documentation** of what's important
- Visualization of **dependencies**
- **Querying provenance** (*prospective, retrospective, and hybrid*)

➔ make **provenance actionable**

➔ **provenance for self!**

=> [github.com/yesworkflow-org/yw](https://github.com/yesworkflow-org/yw)

=> [try.yesworkflow.org](https://try.yesworkflow.org)



(Disclaimer) <https://github.com/idaks/dataone-ahm-2016-poster>

# Demo Time

<https://github.com/idaks/wt-prov-summer-2017>

<https://github.com/yesworkflow-org/yw-idcc-17>

The screenshot displays a presentation slide titled "wf\_recon\_complete\_graph\_all\_observables.gv". The slide is divided into two main sections: a terminal window on the left and a workflow diagram on the right.

**Terminal Window:** The terminal shows the execution of a script named "C3\_C4\_map\_present\_NA\_with\_comments.m". The script loads input data (SYNMAP land cover classification map) and processes it to generate output files. The terminal output includes the following commands and results:

```
nodatavalue = -999.0;

%% Load input: SYNMAP land cover classification map; also read coordinate variables to re-use them later
% @BEGIN fetch_SYNMAP_land_cover_map_variable
% @in SYNMAP_land_cover_map_data @URI inputs/land_cover/SYNMAP_NA_QD.nc
% @out lon @AS lon_variable
% @out lat @AS lat_variable
% @out lon_bnds @AS lon_bnds_variable
% @out lat_bnds @AS lat_bnds_variable

grass_type=[19,20,21,22,23,24,25,26,27,38,41,42,43];
sncid=netcdf.open('inputs/land_cover/SYNMAP_NA_QD.nc','NC_NOWRITE');
fvid=netcdf.inqVarID(sncid,'biome_frac');
frac=netcdf.getVar(sncid,fvid);
tvid=netcdf.inqVarID(sncid,'biome_type');
type=netcdf.getVar(sncid,tvid);

lon_vid=netcdf.inqVarID(sncid,'lon');
lon=netcdf.getVar(sncid,lon_vid);
lat_vid=netcdf.inqVarID(sncid,'lat');
lat=netcdf.getVar(sncid,lat_vid);

--(DOS)--- C3_C4_map_present_NA_with_comments.m 5% (17,0) Git-master (MATLAB Fill)

*shell* 1 C3C4
/Users/ludaesch-admin/git/dataone-ahm-2016-poster/examples/C3C4:
total used in directory 56 available 87640315
drwxr-xr-x 11 ludaesch-admin staff 374 Oct 5 19:52 .
drwxr-xr-x 40 ludaesch-admin staff 1360 Oct 5 19:52 results
drwxr-xr-x 8 ludaesch-admin staff 272 Oct 5 19:52 facts
drwxr-xr-x 4 ludaesch-admin staff 136 Oct 5 19:52 views
drwxr-xr-x 8 ludaesch-admin staff 272 Oct 5 19:51 ..
-rw-r--r-- 1 ludaesch-admin staff 12292 Oct 5 19:36 .DS_Store
-rwxr-xr-x 1 ludaesch-admin staff 125 Oct 5 19:36 clean.sh

-:%%- C3C4 Top (1,0) (Dired by date)
no changes added to commit (use "git add" and/or "git commit -a")
bash-3.2$ pwd
/Users/ludaesch-admin/git/dataone-ahm-2016-poster/examples/LIGO
bash-3.2$ cd ..
bash-3.2$ cd C3C4/
bash-3.2$ ./clean.sh
bash-3.2$ ./make.sh
bash-3.2$

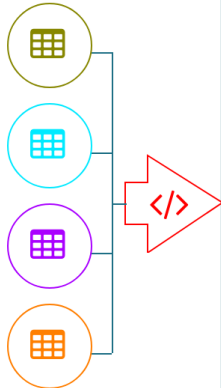
-:*** *shell* Bot (140,10) (Shell:run)
```

**Workflow Diagram:** The diagram illustrates the data processing workflow. It starts with input data (SYNMAP land cover classification map) and coordinate variables (lon, lat, lon\_bnds, lat\_bnds). These inputs are processed through several steps: "fetch\_SYNMAP\_land\_cover\_map\_variable", "fetch\_monthly\_mean\_air\_temperature\_data", "precipitation\_data", "Rain\_Matrix", "Tair\_Matrix", "line\_pixels\_for\_grass", "C3\_Data", "C4\_Data", "lon\_variable", "lat\_variable", "lon\_bnds\_variable", "lat\_bnds\_variable", "or\_C3\_fraction", "generate\_netcdf\_file\_for\_C4\_fraction", and "generate\_netcdf\_file\_for\_C3\_fraction". The final output is a netCDF file named "outputs/SYNMAP\_PRESENTVEG\_C4Grass\_RelafRac\_NA\_v2.0.nc".

# DataONE: Search and Provenance Display

## Data Table, Image, and Other Data Details

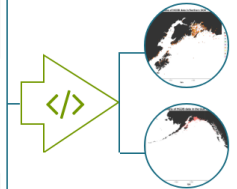
4 sources



### Data Table

Entity Name	Total_Aromatic_Alkanes_PWS.csv										
	<a href="#">Download</a>										
Description	Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK										
Object Name	Total_Aromatic_Alkanes_PWS.csv										
Online Distribution Info	<a href="https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9">https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9</a>										
Size	2801033 byte										
Text Format	<table><tr><td>Number of Header Lines</td><td>1</td></tr><tr><td>Record Delimiter</td><td>#x0A</td></tr><tr><td>Attribute Orientation</td><td>column</td></tr><tr><td colspan="2"><b>Simple Text</b></td></tr><tr><td>Field Delimiter</td><td>,</td></tr></table>	Number of Header Lines	1	Record Delimiter	#x0A	Attribute Orientation	column	<b>Simple Text</b>		Field Delimiter	,
Number of Header Lines	1										
Record Delimiter	#x0A										
Attribute Orientation	column										
<b>Simple Text</b>											
Field Delimiter	,										
Number Of Records	12142										

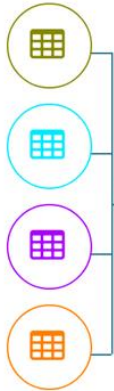
2 derivations



# DataONE: Search and Provenance Display

## Data Table, Image, and Other Data Details

4 sources



**Source Program**

**Total\_PAH\_and\_Alkanes\_GoA\_Hydrocarbons\_Clean.R**

Citation

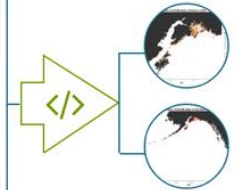
[View »](#)

This program generated the data you are currently viewing, **Total\_Aromatic\_Alkanes\_PWS.csv**.

This program used **PAH.csv**, **Sample.csv**, **Non-EVOS\_SINs.csv** and (and 1 more ).

Alkanes_PWS.csv
from PAH, Alkane and Sample tables documenting samples collected after the oil spill in Prince William Sound, AK
Alkanes_PWS.csv
<a href="https://dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9">https://dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9</a>

2 derivations



Text Format

Number of Header Lines	1
Record Delimiter	#x0A
Attribute Orientation	column
<b>Simple Text</b>	
Field Delimiter	,

Number Of Records

12142



# Adding YesWorkflow to DataONE

25 inputs

**Other Entity**

Entity Name: C3\_C4\_map\_present\_NA\_with\_comments.m

Download

Data Object Type: text/plain

Physical Structure Description:

Object Name: C3\_C4\_map\_present\_NA\_with\_comments.m

Size: 13962

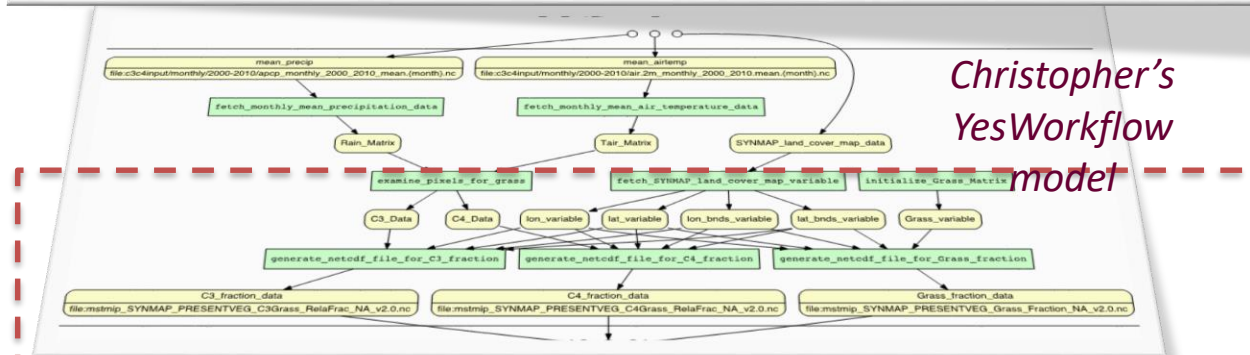
Externally Defined Format: text/plain

Online Distribution Info: [https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program\\_014c5a89-011b-4125-bdb5-af0475020e1a](https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_014c5a89-011b-4125-bdb5-af0475020e1a)

6 outputs

iew more

*Yaxing's script with inputs & output products*



*Christopher's results can be traced back all the way to Yaxing's input*

8 inputs

**Other Entity**

Entity Name: GrasslandWUE.m

Download

Data Object Type: text/plain

Physical Structure Description:

Object Name: GrasslandWUE.m

Size: 4443

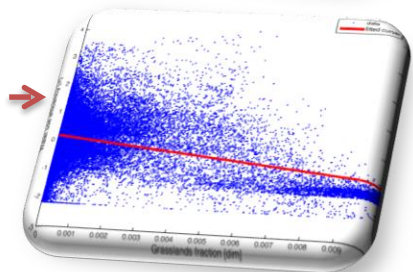
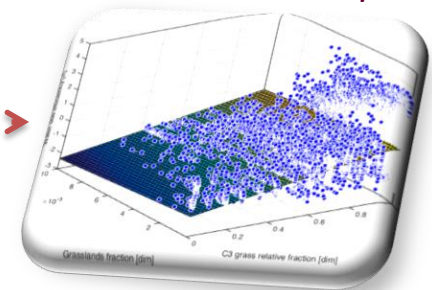
Externally Defined Format: text/plain

Online Distribution Info: [https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program\\_92511b50-c6b2-4949-9ee2-b176a46bd913](https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_92511b50-c6b2-4949-9ee2-b176a46bd913)

2 outputs

iew more

*Christopher using Yaxing's outputs as inputs for his script*





# Agreeing to Disagree: Reconciling Conflicting Taxonomic Views using a Logic-based Approach

Yi-Yun Cheng<sup>1</sup>, Nico Franz<sup>2</sup>, Jodi Schneider<sup>1</sup>, Shizhuo Yu<sup>3</sup>, Thomas Rodenhauen<sup>4</sup>, Bertram Ludäscher<sup>1</sup>

<sup>1</sup>School of Information Sciences, University of Illinois at Urbana-Champaign; <sup>2</sup>School of Life Sciences, Arizona State University;

<sup>3</sup>Department of Computer Science, University of California at Davis; <sup>4</sup>School of Information, University of Arizona

## INTRODUCTION

**Tina:** Hey Amy, can you recommend a signature dish from where you live?

**Amy:** Oh, definitely the half-smokes from the Northeast! They are these tasty half-pork and half-beef sausages.

**Tina:** What a coincidence! We have half-smokes in the South, too! Where do you live in the Northeast? New York? Boston?

**Amy:** Wrong guesses! Where do you live in the South?

**Tina and Amy together:** Washington, D.C.

[The two of them look at each other, confused.]



Figure 1. National Diversity Council map (NDC) vs. Census Bureau map (CEN)

“In the face of incompatible information or data structures among users or among those specifying the system, attempts to create unified knowledge categories are futile. Rather, parallel or multiple representational forms are required...” (Bowker & Star, 2000).

## RELATED WORK

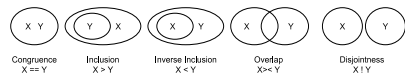
### Taxonomy Alignment Problems (TAP)

Taxonomies  $T_1$ ,  $T_2$  are inter-linked via a set of input articulations  $A$ , defined as RCC-5 relations, to yield a “merged” taxonomy  $T_3$ .

### Euler/X

**Articulations** – a constraint or rule that defines a relationship (a set constraint) between two concepts from different taxonomies.

### Region Connection Calculus (RCC-5)



**Possible Worlds** – When encoding and solving TAPs via ASP, the different answer sets represent alternative taxonomy merge solutions or possible worlds (PWs).

• Github link:  
<https://github.com/EulerProject/ASIST17>  
• Email: [bertram.lud@illinois.edu](mailto:bertram.lud@illinois.edu)



## CASE 1 RESULTS: CEN vs. NDC

- State-level alignments are all congruent (Bottom-up)
- Inferred new articulations for regional-level alignments

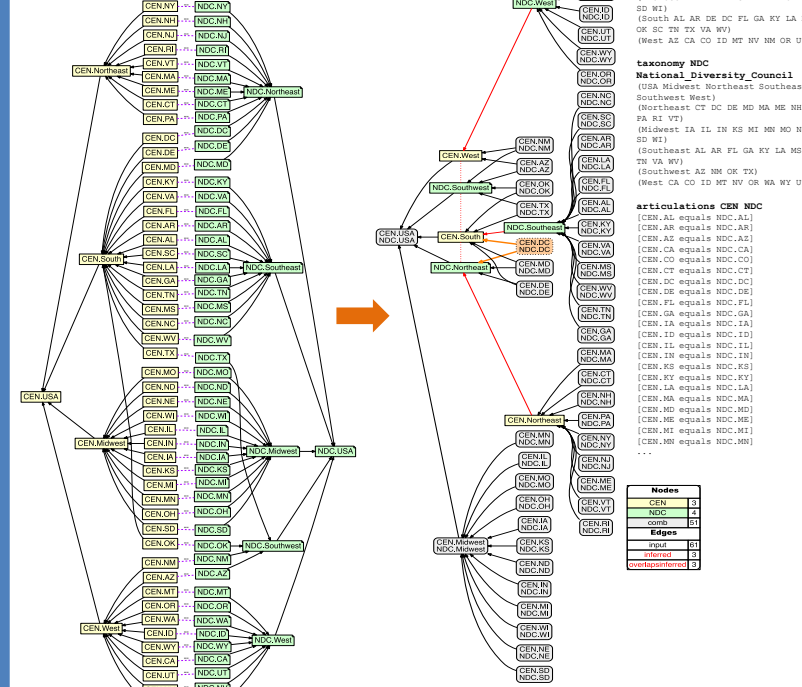


Figure 3. (Left) CEN-NDC taxonomy alignment problem with 49 input articulations between  $T_{CEN}$  and  $T_{NDC}$ . Figure 4. (Right) The unique possible world (PW)  $T_3$  reconciling  $T_{CEN}$  and  $T_{NDC}$  via inferred relationships

## CASE 2 RESULTS: CEN vs. TZ

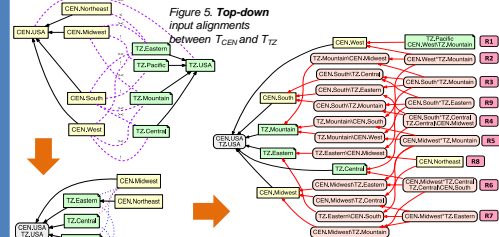


Figure 6. The unique PW for the  $T_{CEN}$  with  $T_{TZ}$  alignment

Figure 10. Combined concepts solution for  $T_{CEN}$  and  $T_{TZ}$

## RESEARCH DESIGN

- Step 1. Supply input taxonomies  $T_1$  and  $T_2$
- Step 2. Formulate RCC-5 articulations between  $T_1$  and  $T_2$
- Step 3. Iteratively edit articulations in Euler/X

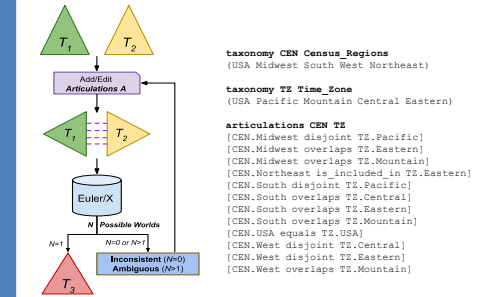


Figure 2. The process of aligning taxonomies  $T_1$  and  $T_2$  with Euler/X

## CONCLUSION

- Our logic-based taxonomy alignment approach can be used to solve crosswalking issues. We will be able to mitigate the membership condition problems that occur in equivalent crosswalking.
- RCC-5 approach preserves the original taxonomies while providing an alignment view. We can solve data integration problems that happen in the more coarse-grained relative crosswalking, which otherwise is subjected to information loss.
- Our study also underscores the benefits of designing different alignment workflows (Bottom up vs. Top-down) to match the needs of specific taxonomy alignment problems. Bottom-up approach: seems to work well whenever we have non-overlapping relationships at the leaf-level (lowest-level) articulations, and we are not sure how the higher-level concepts should be aligned.

Top-down approach: seems favorable when there is an expectation of certain higher-level articulations in conjunction with under-specified, complex, and often overlapping leaf-level relations.

## Acknowledgments

Support of the authors' research through the National Science Foundation is kindly acknowledged (DEB-1155984, DBI-1342595, and DBI-1643002). The authors thank Professor Kathryn La Barre for her comments and suggestions. We would also like to thank Dr. Laetitia Navarro and Jeff Terstriep for help with creating map overlays in QGIS.

*Tracing taxonomic names (concepts!) over time ...*

***For another time?***

# Non-unitary syntheses of systematic knowledge

**Nico Franz**

School of Life Sciences, Arizona State University

***CIRSS Seminar – Center for Informatics Research in Science and Scholarship***

February 17, 2017 – iSchool, University of Illinois Urbana-Champaign

# Taxonomic concept alignment, *Andropogon glomeratus-virginicus* complex, spanning across 11 classifications authored 1889-2015

- **36** unique taxonomic names
- **88** taxonomic concept labels  
☐ name sec. author strings
- **Alignment** by A.S. Weakley  
☐ row position = congruence
- **1/36 names** with unique 1 : 1 name : meaning cardinality across all classifications
- *Andropogon virginicus*
- **Source:** Franz *et al.* 2016<sup>1</sup>

1	13	17	24	31	33
sec. Hackel (1889)	sec. Small (1933)	sec. Blomquist (1948)	sec. Hitchcock & C. (1950)	sec. RAD (1968)	sec. Godfrey & W. (1979)
<i>A. virginicus</i> var. <i>glauca</i> subvar. <i>glauca</i>	<i>A. capillipes</i>	<i>A. capillipes</i>	<i>A. capillipes</i>	<i>A. virginicus</i>	<i>A. capillipes</i>
4	14	18	25	32	34
<i>A. virginicus</i> var. <i>glauca</i> subvar. <i>dealbatus</i>	<i>A. capillipes</i>	<i>A. capillipes</i>	<i>A. capillipes</i>	<i>A. virginicus</i>	<i>A. capillipes</i>
5					
<i>A. virginicus</i> var. <i>viridis</i> subvar. <i>genuinus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>
7	15	20	27		36
<i>A. virginicus</i> var. <i>viridis</i> subvar. <i>genuinus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>
<i>A. virginicus</i> var. <i>viridis</i> subvar. <i>genuinus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>
<i>A. macrourus</i> var. <i>glaucoptis</i>	<i>A. glomeratus</i>	<i>A. virginicus</i> var. <i>glaucoptis</i>	<i>A. virginicus</i> var. <i>glaucoptis</i>	<i>A. virginicus</i>	<i>A. glaucoptis</i>
9	16	21	28		38
<i>A. macrourus</i> var. <i>hirsutior</i>	<i>A. glomeratus</i>	<i>A. glomeratus</i> (?)	<i>A. virginicus</i> var. <i>hirsutior</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>abbreviatus</i>
10		23	29		37
<i>A. macrourus</i> var. <i>abbreviatus</i>	<i>A. glomeratus</i>	<i>A. glomeratus</i>	<i>A. glomeratus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>abbreviatus</i>
11			30		
<i>A. macrourus</i> var. <i>genuinus</i>	<i>A. glomeratus</i>	<i>A. virginicus</i> var. <i>tenuispathus</i>	<i>A. glomeratus</i>	<i>A. virginicus</i>	<i>A. virginicus</i> var. <i>abbreviatus</i>
12		22			

39	53	67	79	89
sec. Campbell (1983)	sec. Campbell (2003)	sec. Weakley (2006)	sec. BONAP (2014)	sec. Weakley (2015)
<i>A. virginicus</i> var. <i>glauca</i> "drylands variant"	<i>A. virginicus</i> var. <i>glauca</i> "drylands variant"	<i>A. capillipes</i> "drylands variant"	<i>A. capillipes</i>	<i>A. capillipes</i>
42	56	69	80	90
<i>A. virginicus</i> var. <i>glauca</i> "wetlands variant"	<i>A. virginicus</i> var. <i>glauca</i> "wetlands variant"	<i>A. capillipes</i> "wetlands variant"	<i>A. capillipes</i>	<i>A. dealbatus</i>
43	57	70		91
<i>A. virginicus</i> var. <i>virginicus</i> "old-field variant"	<i>A. virginicus</i> var. <i>virginicus</i> "old-field variant"	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> "old-field variant"
45	59	72	82	93
<i>A. virginicus</i> var. <i>virginicus</i> "smooth variant"	<i>A. virginicus</i> var. <i>virginicus</i> "smooth variant"	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> var. <i>virginicus</i>	<i>A. virginicus</i> "smooth variant"
46	60			94
<i>A. virginicus</i> var. <i>virginicus</i> "deceptive variant"	<i>A. virginicus</i> var. <i>decipiens</i>	<i>A. virginicus</i> var. <i>decipiens</i>	<i>A. virginicus</i> var. <i>decipiens</i>	<i>A. virginicus</i> var. <i>decipiens</i>
47	61	73	83	95
<i>A. glomeratus</i> var. <i>glaucoptis</i>	<i>A. glomeratus</i> var. <i>glaucoptis</i>	<i>A. glaucoptis</i>	<i>A. glaucoptis</i>	<i>A. glaucoptis</i>
49	63	74	84	96
<i>A. glomeratus</i> var. <i>hirsutior</i>	<i>A. glomeratus</i> var. <i>hirsutior</i>	<i>A. glomeratus</i> var. <i>hirsutior</i>	<i>A. hirsutior</i>	<i>A. hirsutior</i>
50	64	76	85	97
<i>A. glomeratus</i> var. <i>glomeratus</i>	<i>A. glomeratus</i> var. <i>glomeratus</i>	<i>A. glomeratus</i> var. <i>glomeratus</i>	<i>A. glomeratus</i> var. <i>glomeratus</i>	<i>A. glomeratus</i> var. <i>glomeratus</i>
51	65	77	87	99
<i>A. glomeratus</i> var. <i>pumilus</i>	<i>A. glomeratus</i> var. <i>pumilus</i>	<i>A. tenuispathus</i>	<i>A. glomeratus</i> var. <i>pumilus</i>	<i>A. tenuispathus</i>
52	66	78	88	100

<sup>1</sup> Franz *et al.* 2016. Names are not good enough: reasoning over taxonomic change in the *Andropogon* complex. Semantic Web Journal (IOS). doi:10.3233/SW-160220

# High-elevation fir trees of western North America

AZ NM

CO

WY

MT

AB

eBC

wBC

WA

OR

Distribution

*Abies lasiocarpa*  
*var. arizonica*

*Abies lasiocarpa* var. *lasiocarpa*

USDA - ITIS

*Abies bifolia*

*Abies lasiocarpa*

Flora North America

A

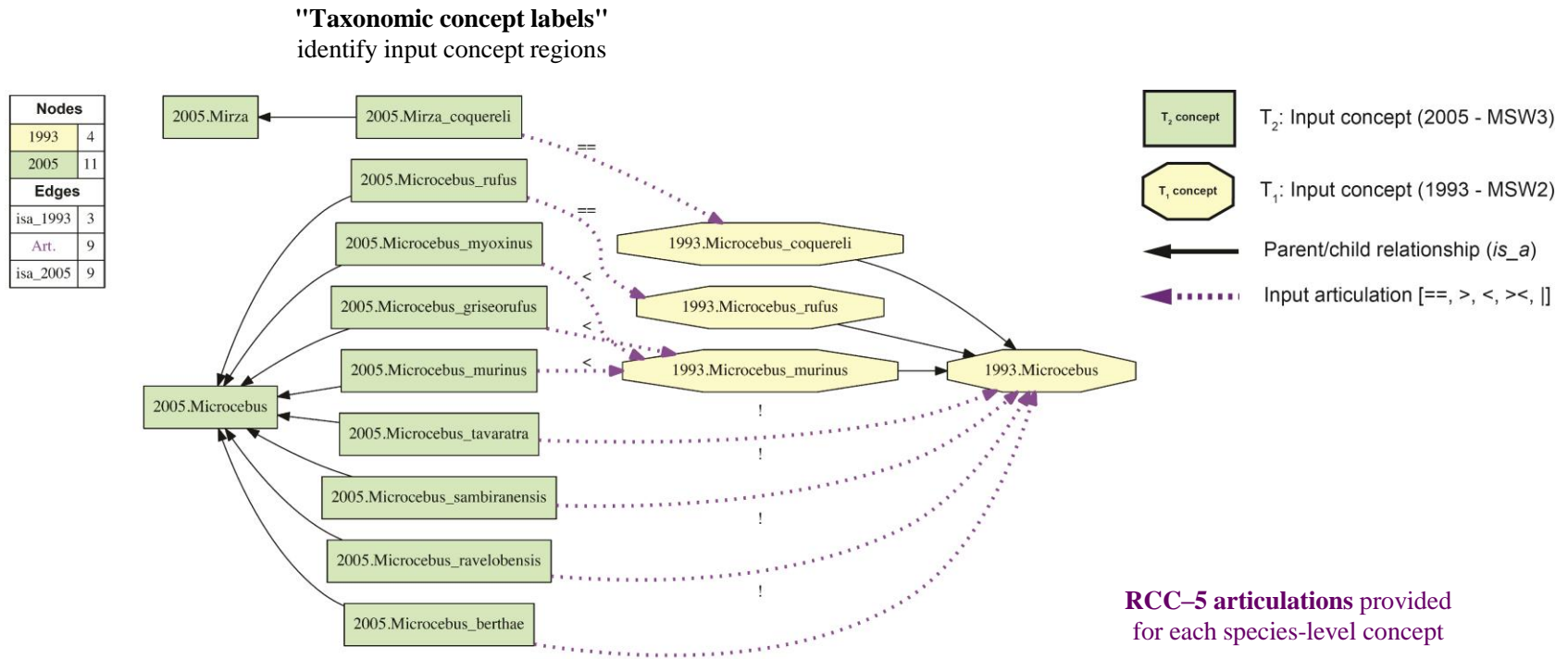
B

C

Minimal concepts

## Use case 1.a. Aligning *Microcebus* + *Mirza* sec. MSW3 (2005)

- Input visualization: MSW3 (2005) versus MSW2 (1993)



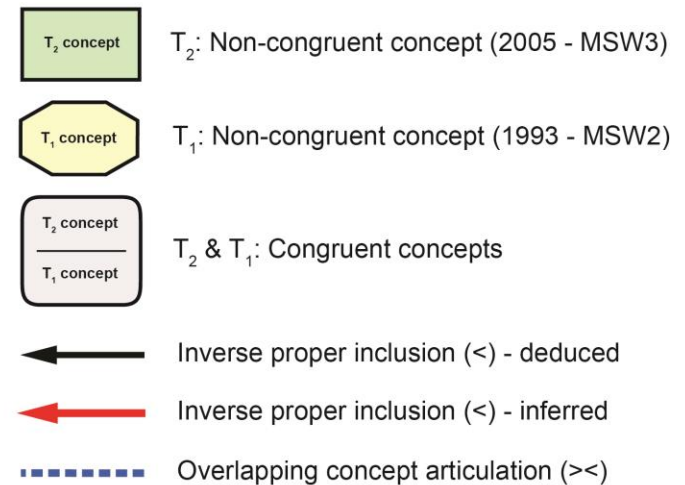
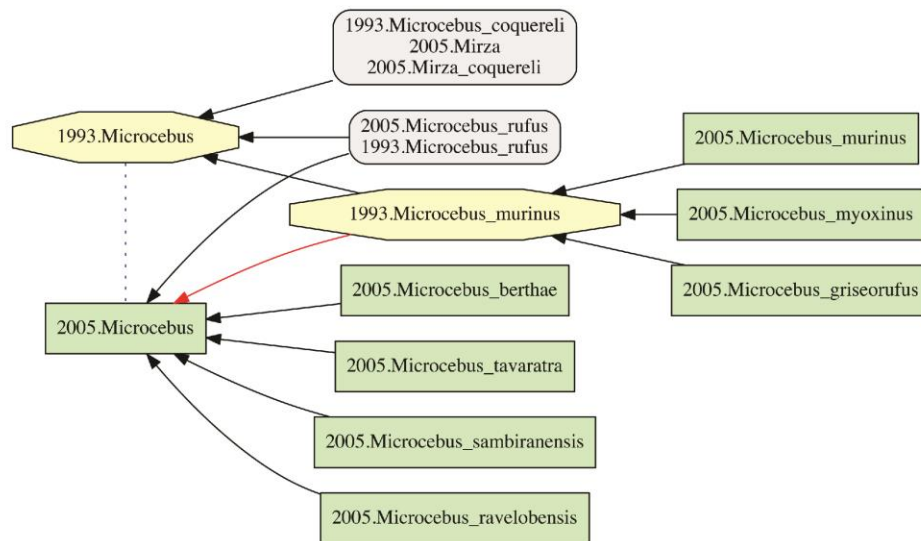
Source: Franz et al. 2016. Two influential primate classifications logical aligned. doi:10.1093/sysbio/syw023



## Use case 1.a. Aligning *Microcebus* + *Mirza* sec. MSW3 (2005)

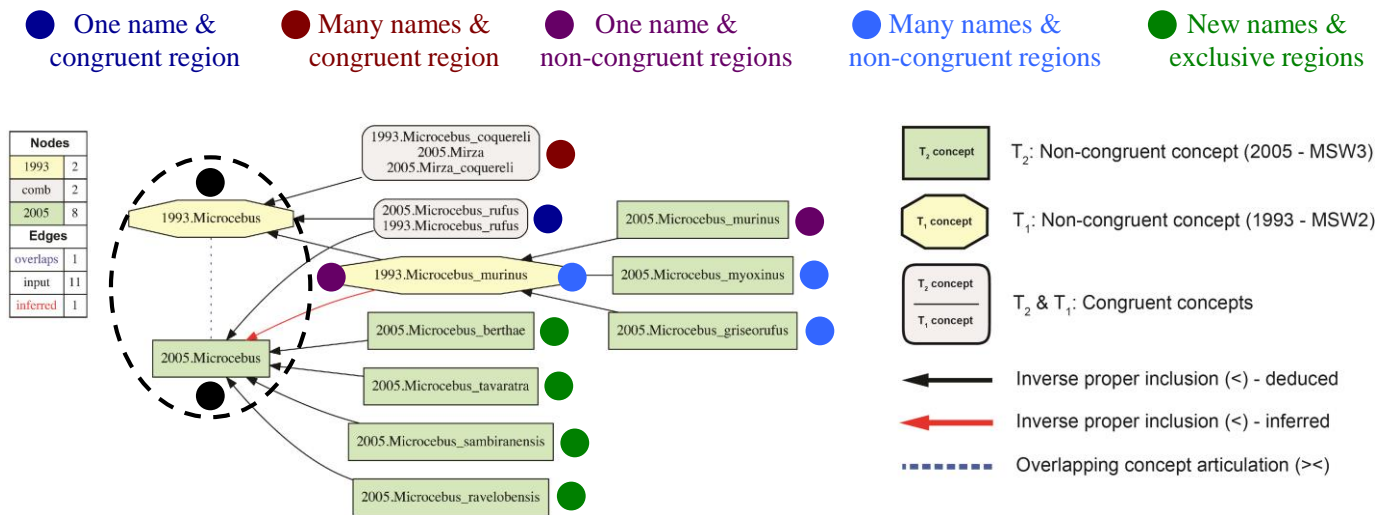
- Alignment visualization: "grey means taxonomically congruent"

Nodes	
1993	2
comb	2
2005	8
Edges	
overlaps	1
input	11
inferred	1



## Use case 1.a. Aligning *Microcebus* + *Mirza* sec. MSW3 (2005)

- Alignment visualization: "grey means taxonomically congruent"



- Application of **coverage constraint**: parent-to-parent articulations ( $><$ ) are fully defined by alignment signal propagated from their respective children.  
 ➔ Sensible when **complete sampling of children is intended**.

## 1 in 3 names is *unreliable* across MSW2/MSW3 classifications

TABLE 4. Analysis of taxonomic name:meaning relations for the entire Prim-UC alignment (800 input concepts), grounded in the MIRs (Table 3)

Rank	sec. Groves (2005)	sec. Groves (1993)	== : =	== : ≠	> : =	< : =	>< : =	Totals
Species	376	233	151	17	1	55	0	224
Genus	69	60	44	0	7	6	2	59
Subfamily	9	10	3	0	3	1	0	7
Family	15	13	5	2	1	0	1	9
Order	1	1	0	0	1	0	0	1
Totals	470	317	203	19	13	62	3	300

Notes: Relations are categorized by taxonomic rank (for shared MSW2/MSW3 ranks only), and emphasize concept pairs with the same name (=) and/or congruent meanings. Legend: == : => taxonomic congruence, same name(s); == : ≠> taxonomic congruence, different names; > : => taxonomic proper inclusion, same name(s); < : => taxonomic inverse proper inclusion, same name(s); >< : => taxonomic overlap, same names(s).

TABLE 5. Analysis of taxonomic congruence and name reliability for six Prim-UC partitions (Table 2)

Partition	sec. Groves (2005)	T <sub>1</sub> concepts	Actual == articulations	Relative congruence (%)	Reliable names	Unreliable names	Reliability ratio
1	Primates	317	283	89.3	203	97	2.1 : 1
2	Primates-HLO*	24	13	54.2	8	12	1 : 1.5
3	Strepsirrhini	77	74	96.1	45	49	1 : 1.1
4	Haplorrhini**	114	98	86.0	79	45	1.8 : 1
5	Catarrhini	125	111	88.8	79	63	1.3 : 1
6	Hominoidea	23	24	100	14	14	1 : 1

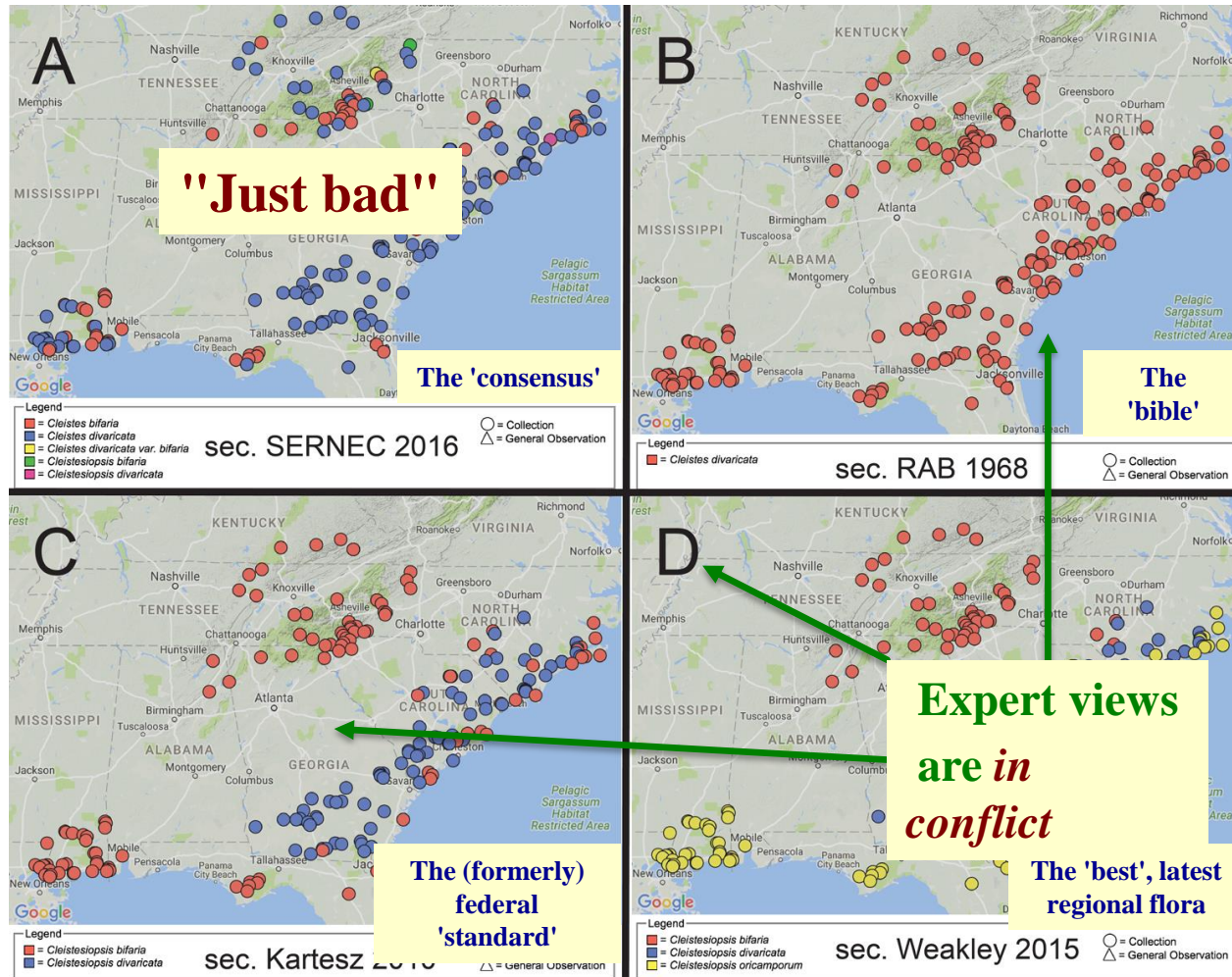
Notes: Relative congruence is understood as the quotient of the number of congruent concepts and number of concepts in the concept-poorer taxonomy (T<sub>1</sub>; sec. Groves 1993). The quotient may be greater than 100% if the concept-richer taxonomy has “redundant” concepts (i.e., multiple concepts with superseding ranks that are taxonomically congruent; see Gregg 1954). Reliable names are of the == : = type in Table 4. Unreliable names are of the [== : ≠, > : =, < : =, >< : =] types in Table 4. The reliable : unreliable ratio is adjusted to 1 for the smaller value. \*HLO = Higher Levels Only. The range of taxonomic ranks is limited to ordinal to subfamilial level.

\*\*Excluding Catarrhini sec. Groves (2005).

Source: Franz et al. 2016. Two influential primate classifications logical aligned. doi:10.1093/sysbio/syw023

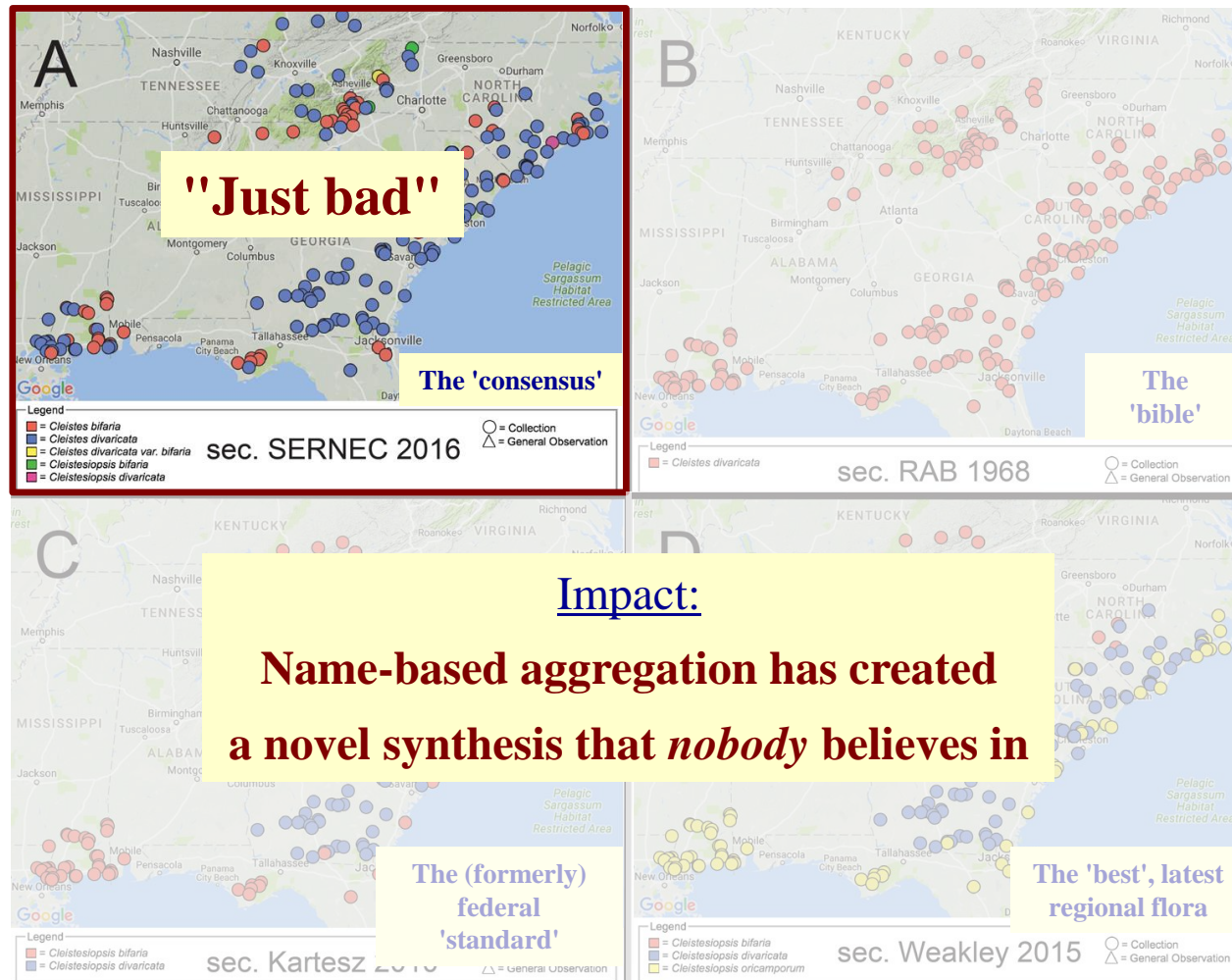


"Controlling the taxonomic variable"



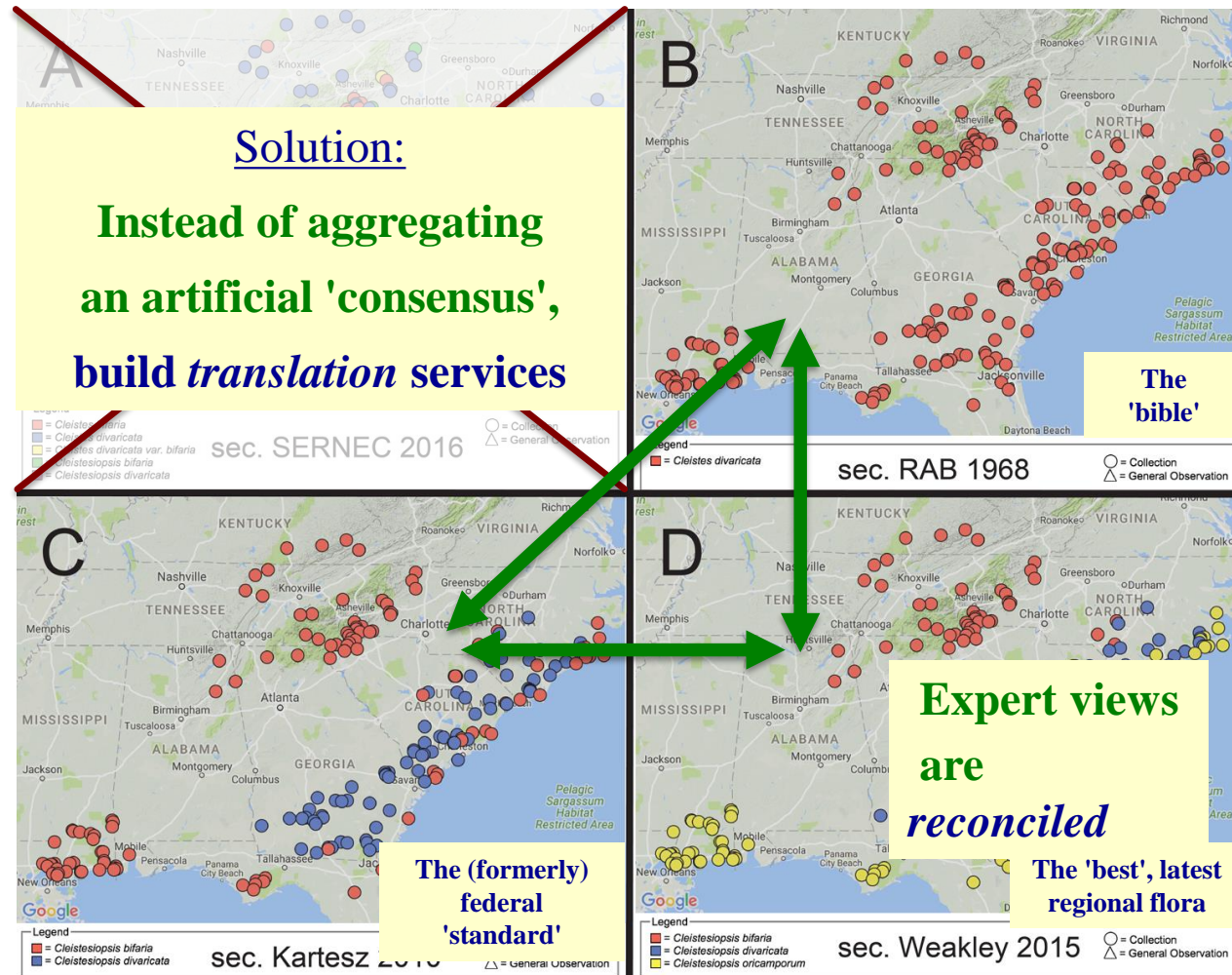
Source: Franz et al. 2016. Controlling the taxonomic variable: [...]. RIO Journal. doi:10.3897/rio.2.e10610

"Controlling the taxonomic variable"



Source: Franz et al. 2016. Controlling the taxonomic variable: [...]. RIO Journal. doi:10.3897/rio.2.e10610

"Controlling the taxonomic variable"



Source: Franz et al. 2016. Controlling the taxonomic variable: [...]. RIO Journal. doi:10.3897/rio.2.e10610

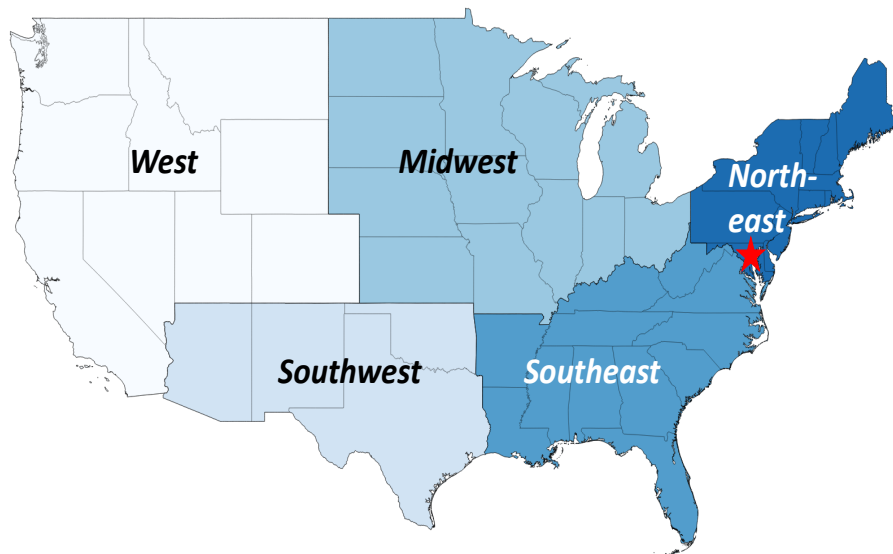
# Leaving taxon and species headaches ...

- To illustrate Euler think of a simpler use case:
- Agreeing to disagree!
- ... when there are **multiple**, legitimate perspectives
- Sorting things out!
  - Euler as a taxon concept (& name) “microscope” ...
  - .. or “time machine” ?

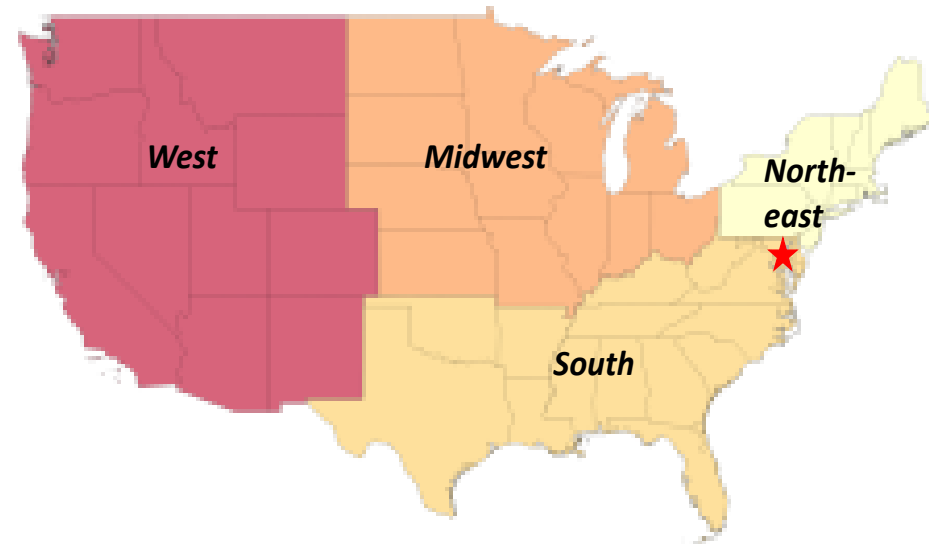


# Two Taxonomies: **NDC** vs **CEN**

*“...in the face of incompatible information or data structures among users or among those specifying the system, attempts to create unitary knowledge categories are futile. Rather, parallel or multiple representational forms are required”* [Bowker & Star, 2000, p.159]



**National Diversity Council map (NDC)**

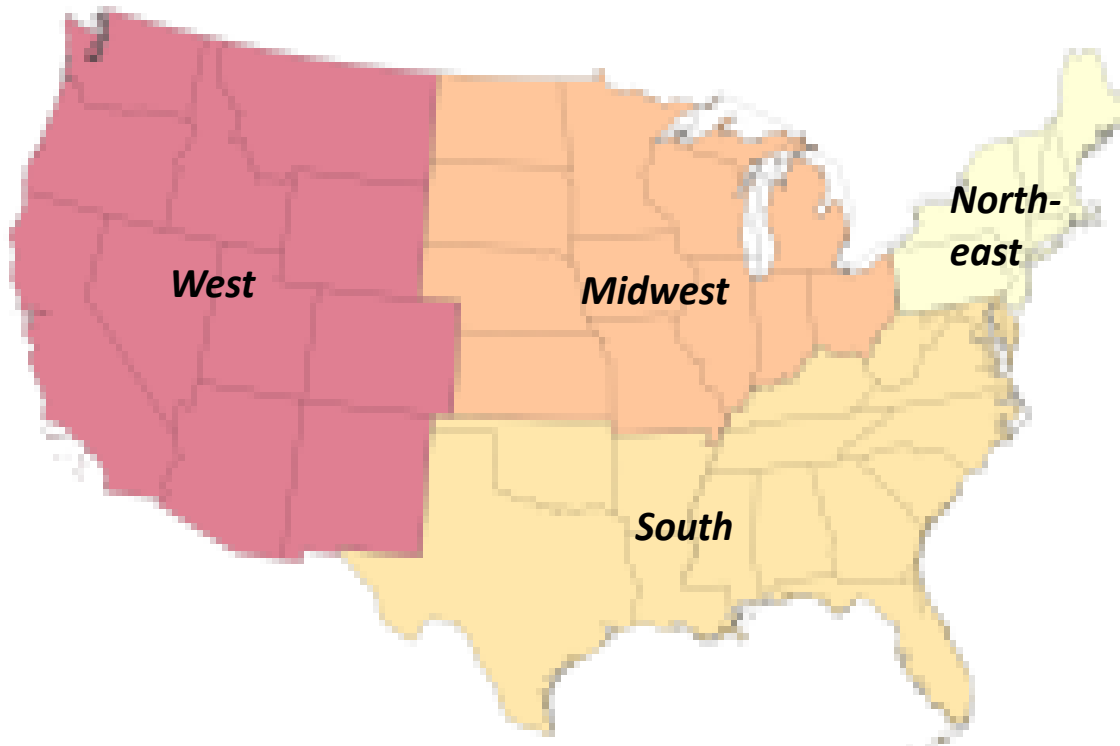


**US Census Buero map (CEN)**

Source: **Yi-Yun (Jessica) Cheng** (PhD student, iSchool @ Illinois)

# The taxonomies

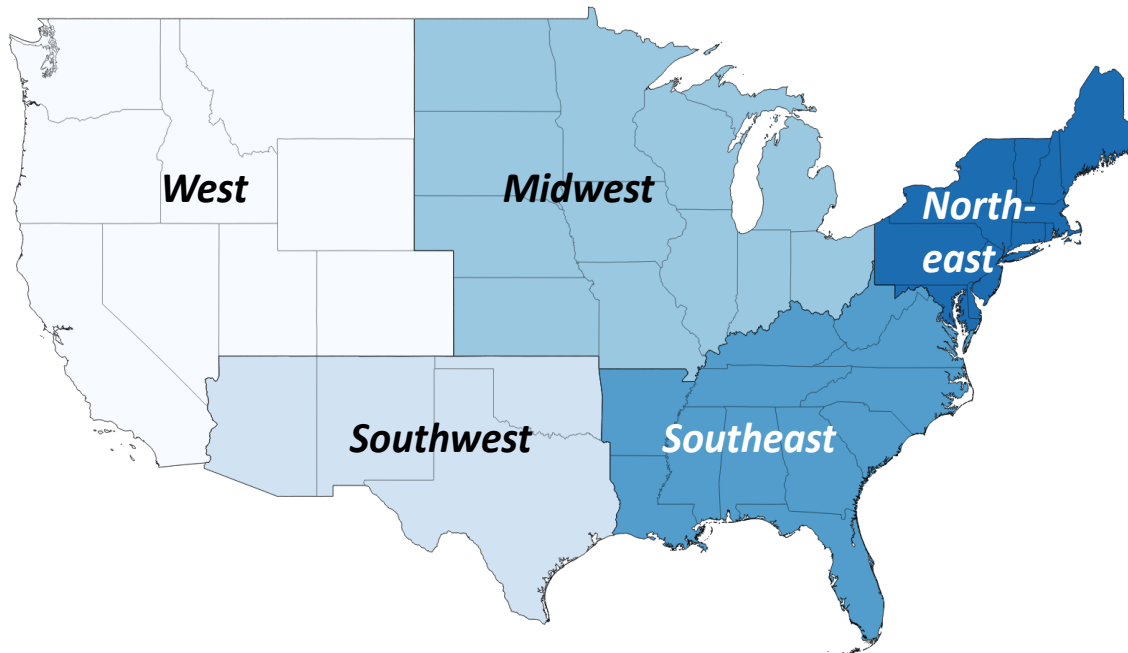
- The Census Regions Map (CEN), consists of **four** regions: West, Midwest, Northeast, and South, i.e., the contiguous 48 states and Washington D.C.





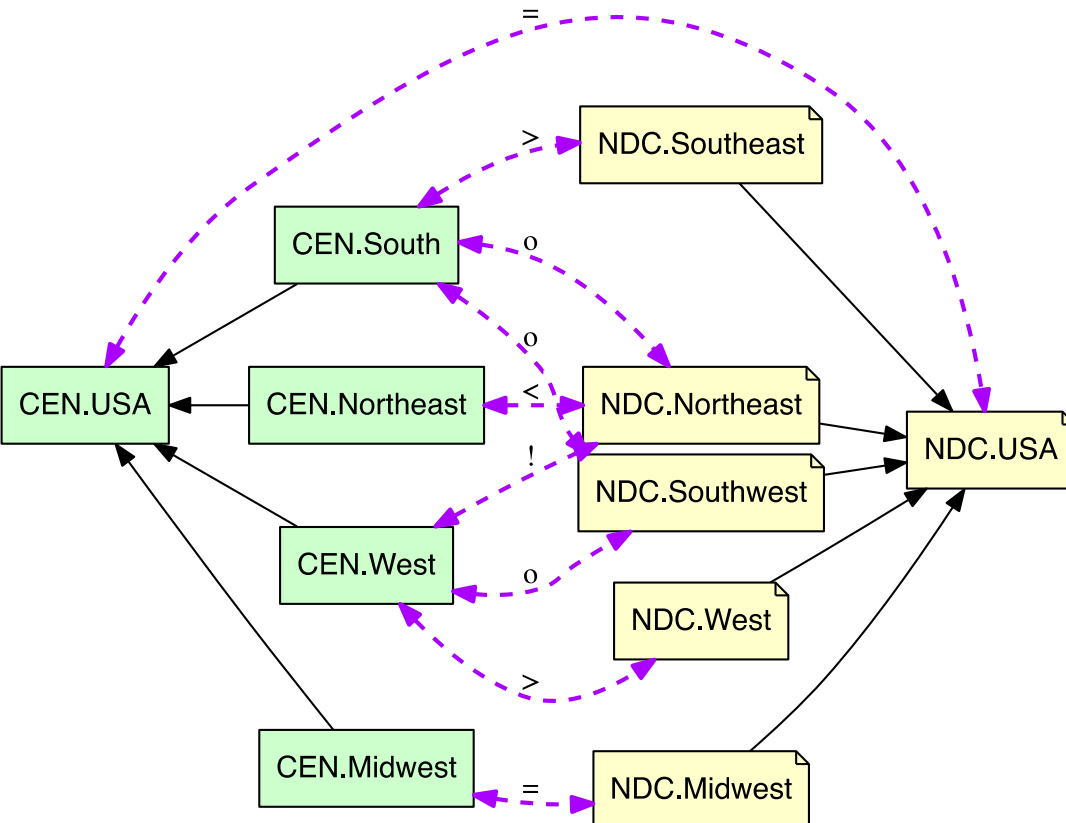
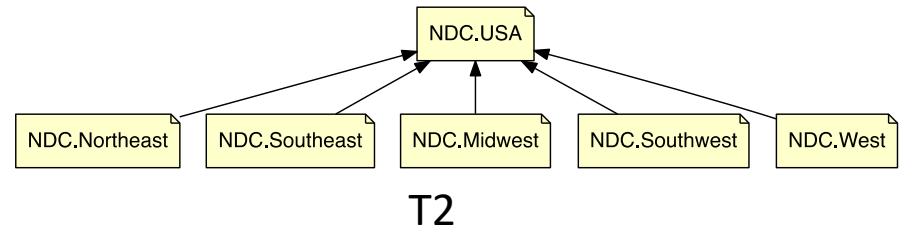
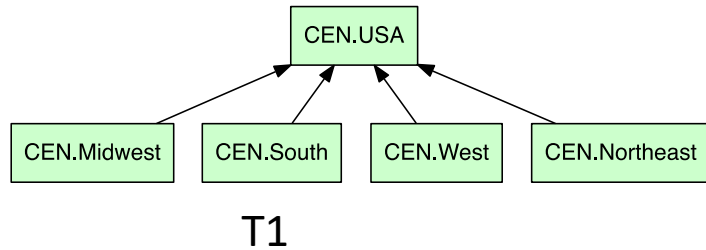
# The taxonomies

- The National Diversity Council Map (NDC), consists of **five** regions: West, Southwest, Midwest, Northeast, Southeast, the 48 states and Washington D.C.



- NDC splits South into SW and SE
- Do NDC and CEN agree on “West”? “Midwest”? ...
- How can we sort this out?

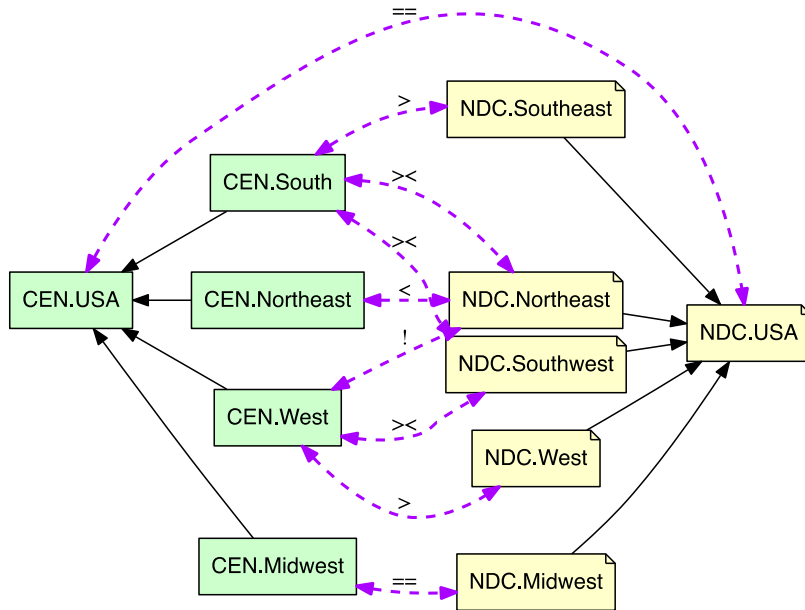
# Sorting things out ...



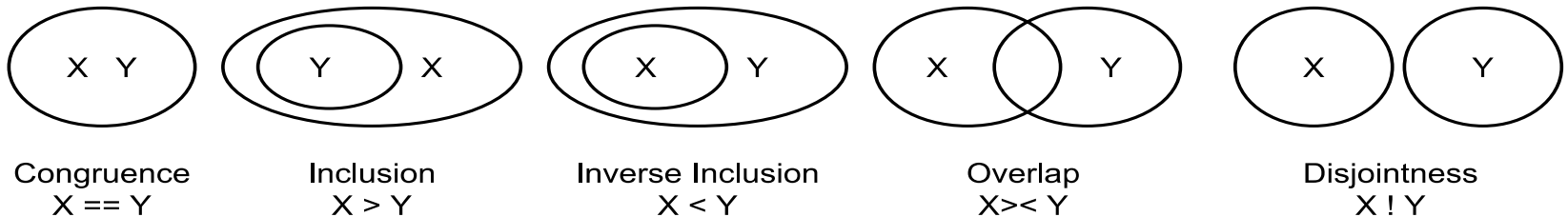
Ludascher: Whole-Tale++

- **Given:**
  - taxonomies T1, T2
  - and relations T1 ~ T2  
(*articulations, alignment*)
- **Find:**
  - merged taxonomy T3
- **Such that:**
  - T1, T2 are **preserved**
  - all pairwise relations are **explicit**

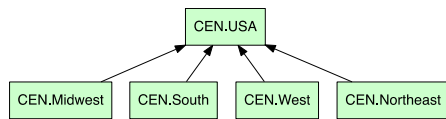
# 5 ways to relate concepts (regions)



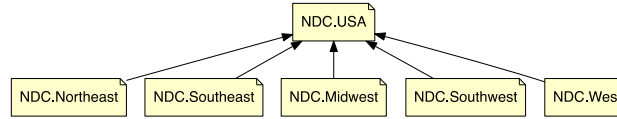
- **Idea:** relate concepts X and Y with *articulations*
- Articulation Language: **Region Connection Calculus (RCC5)**: congruence, inclusion, inverse inclusion, overlap, disjointness



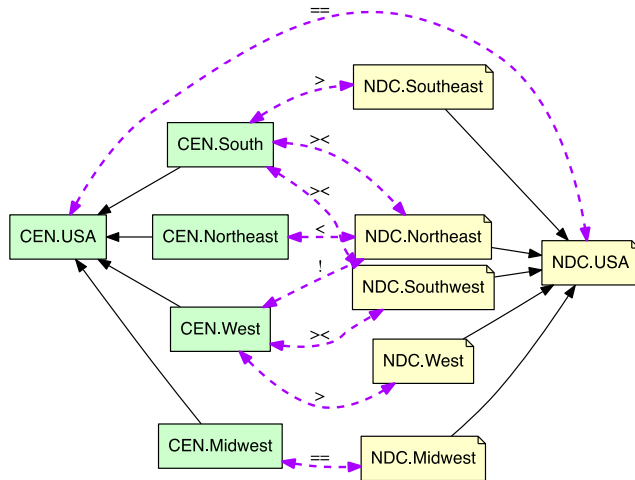
# Merged taxonomy T3



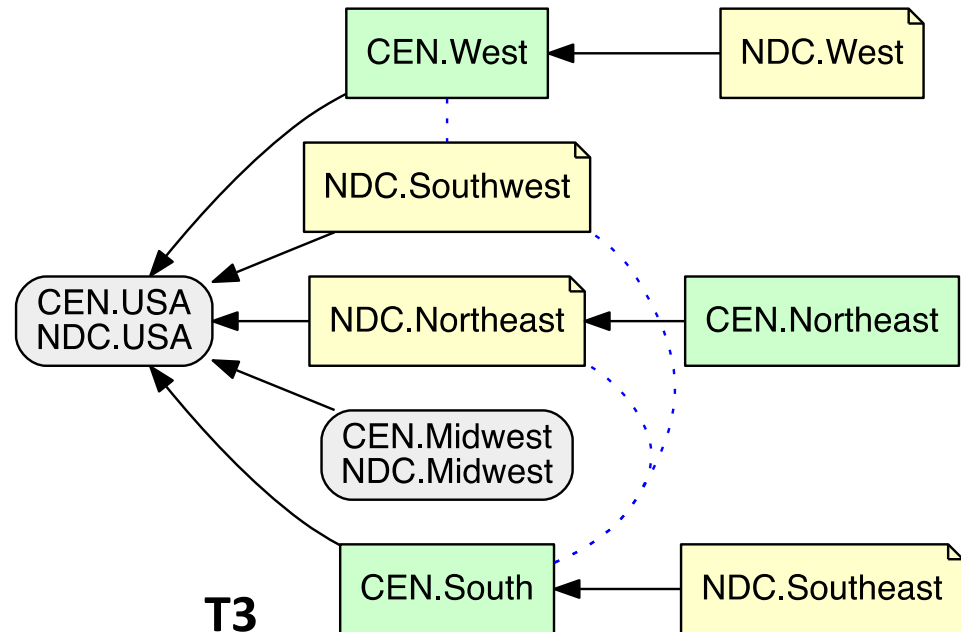
T1



T2



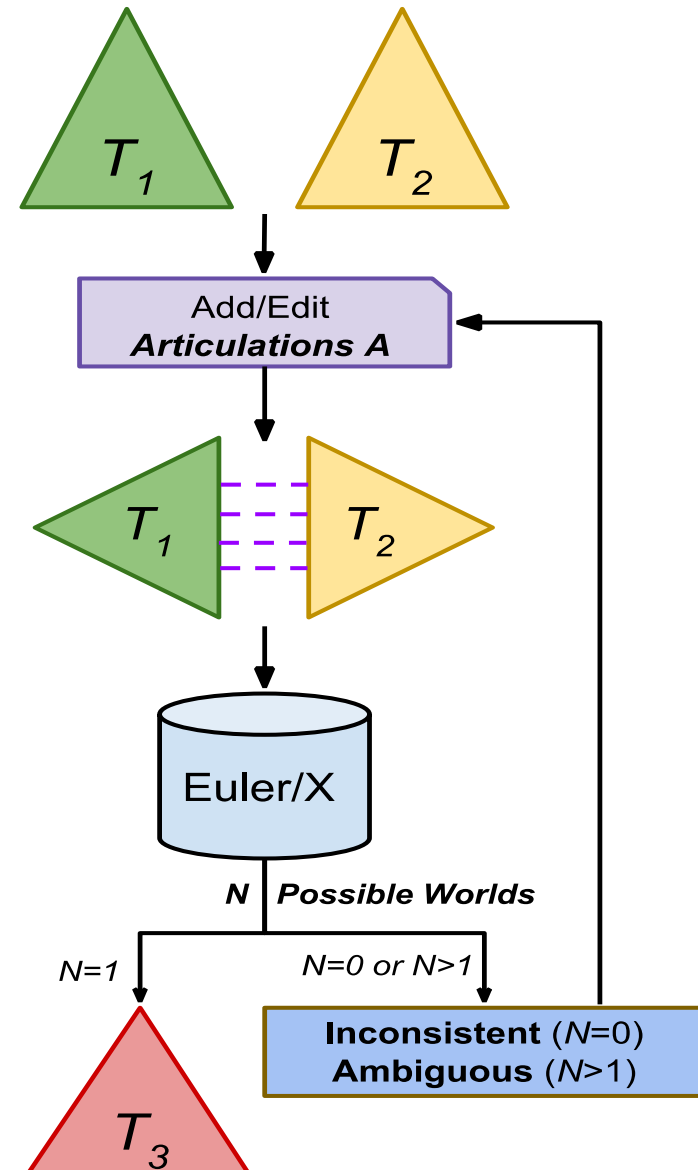
T1 ~ T2



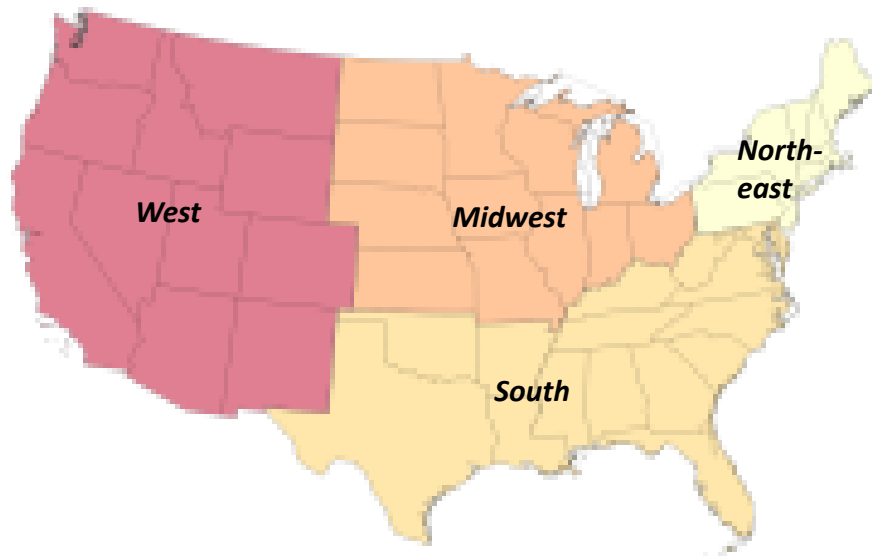
T3

# How we align two taxonomies $T_1$ and $T_2$

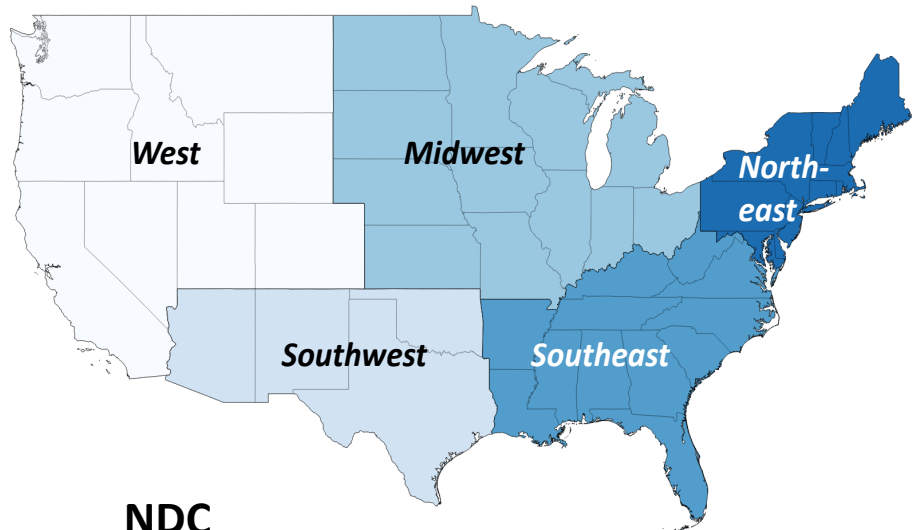
- **Step 1.** Supply input taxonomies  $T_1$  and  $T_2$
- **Step 2.** Describe the relationships between  $T_1$  and  $T_2$
- **Step 3.** Iteratively edit articulations in Euler/X
- ... but where do the *articulations* come from??
  - expert opinion
  - automatically derived from data



# Case 1: Census Region vs. National Diversity Council



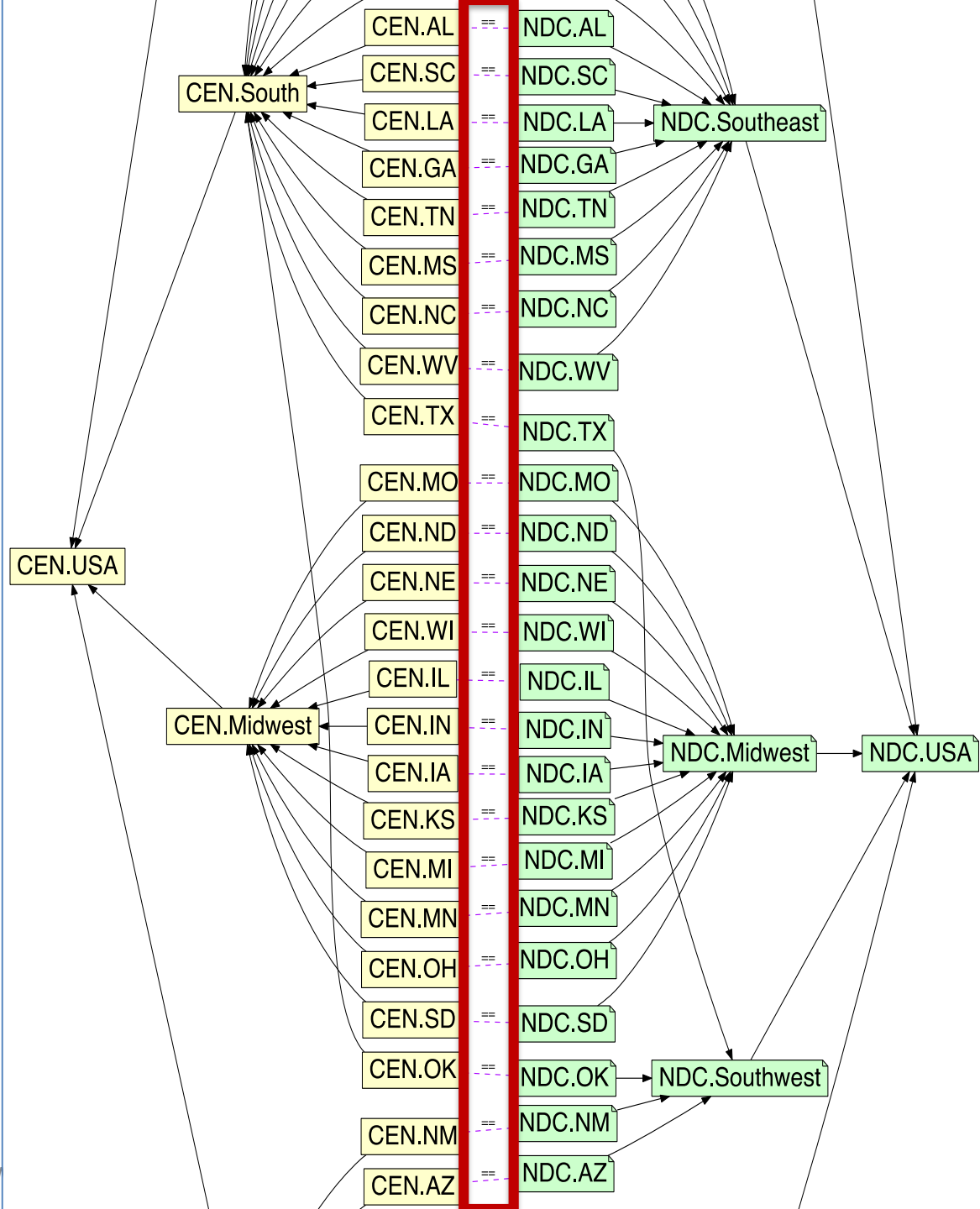
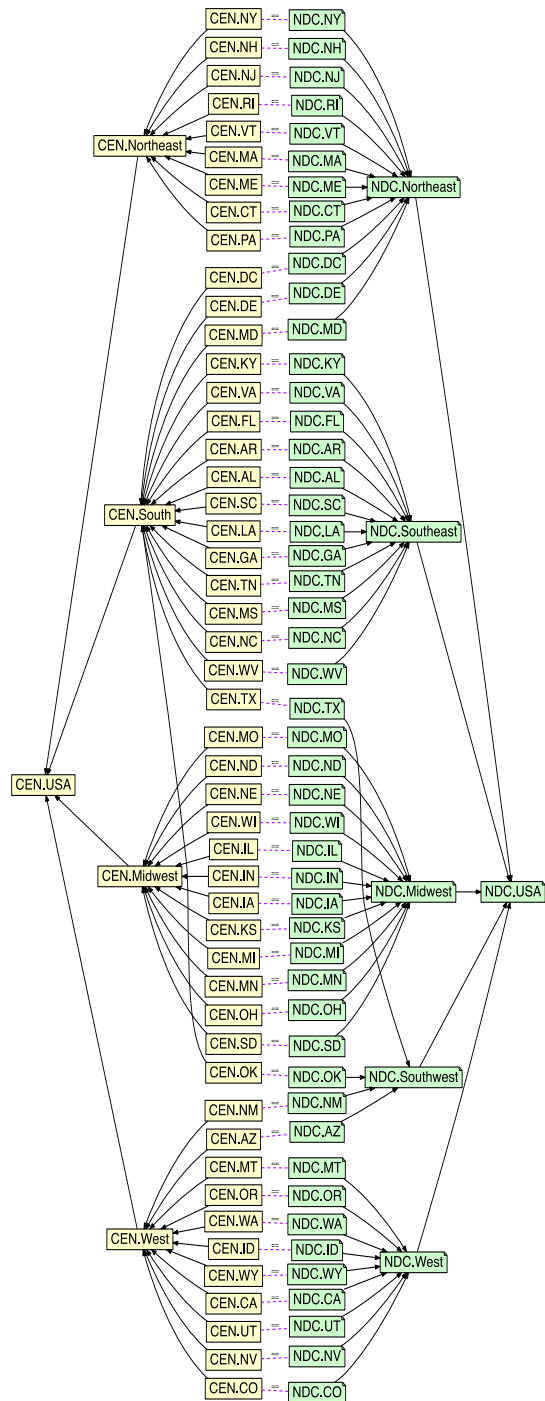
CEN



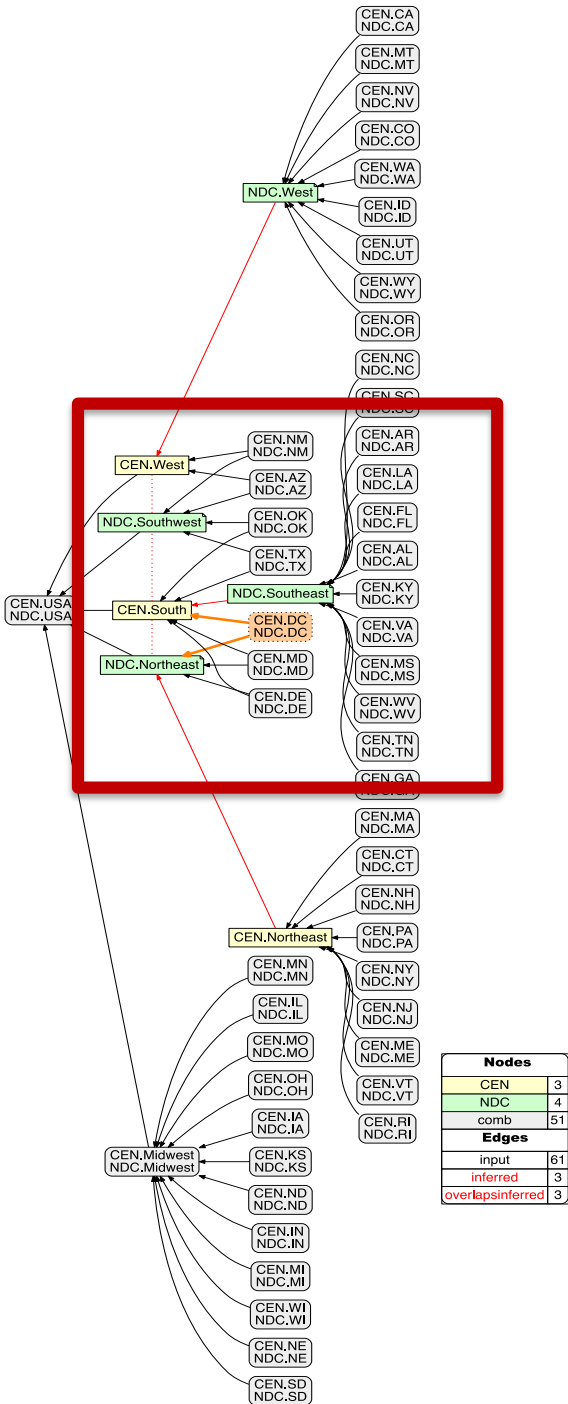
NDC

- ... but where do the *articulations* come from??
  - automatically derived from data
  - expert input

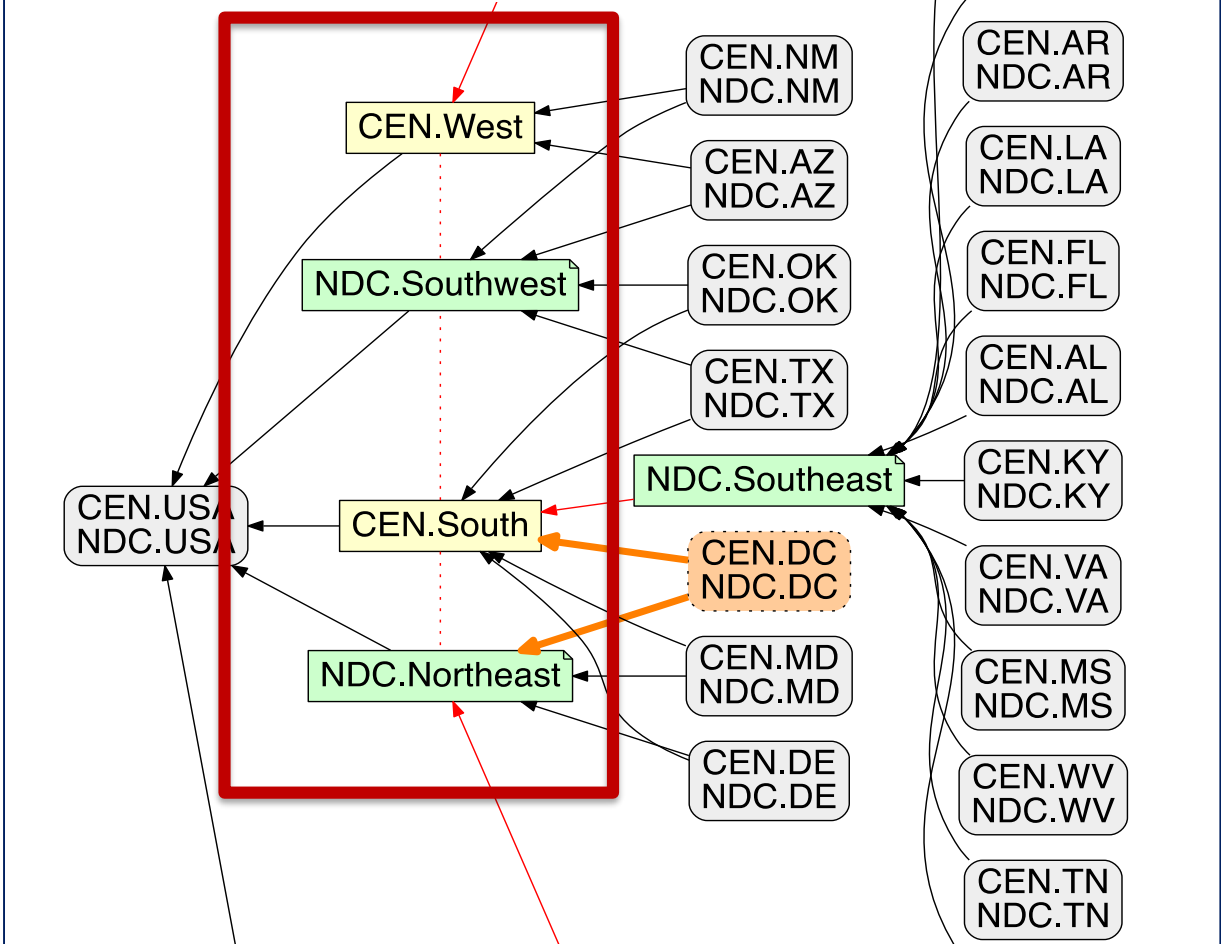




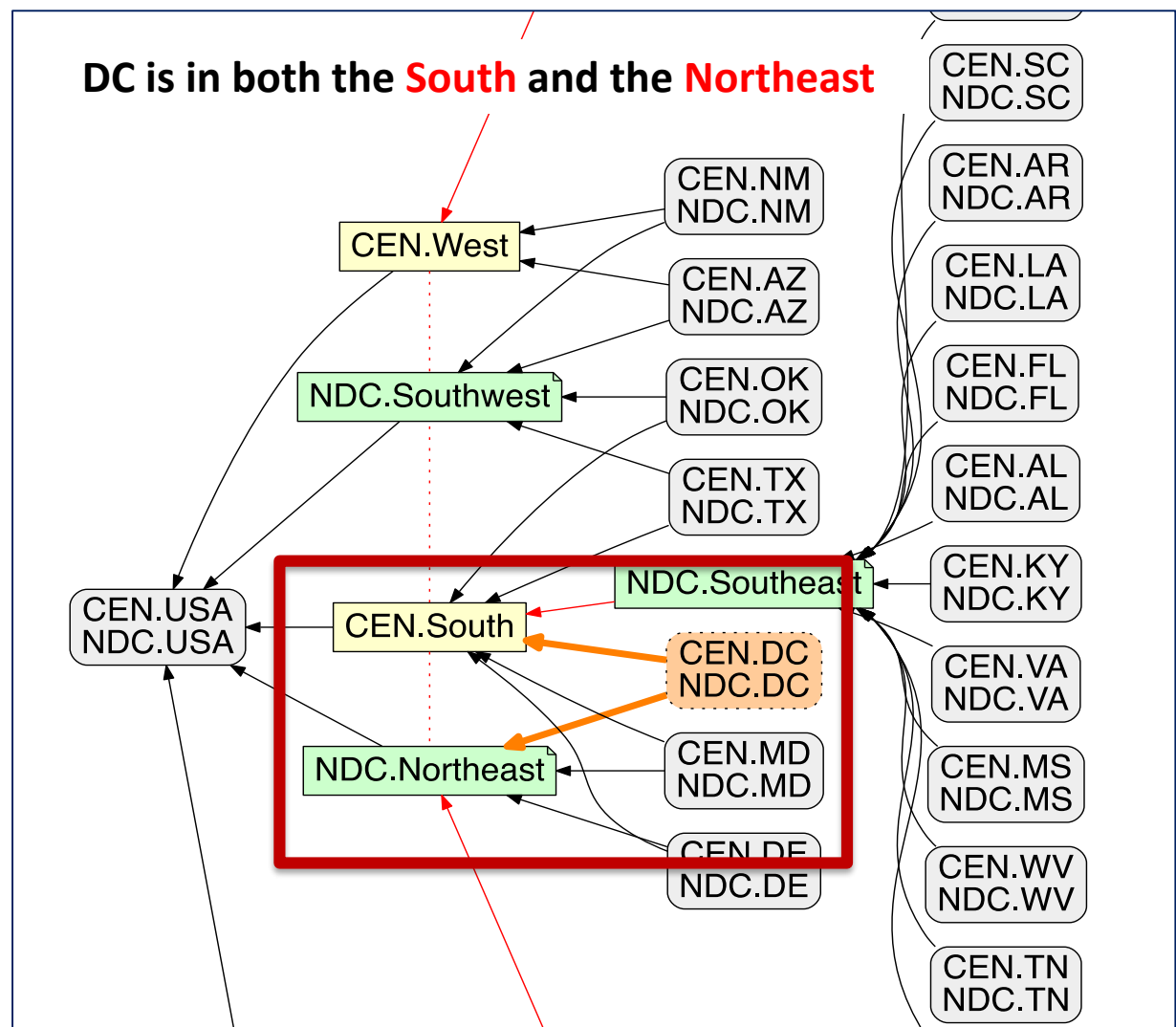
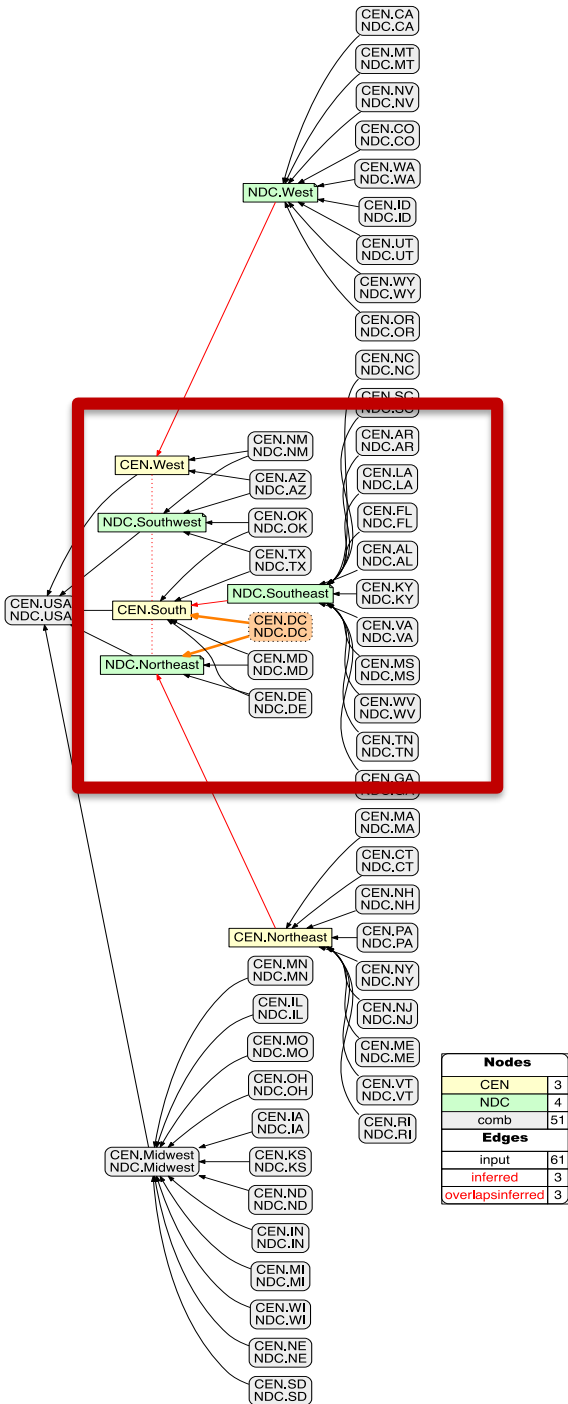




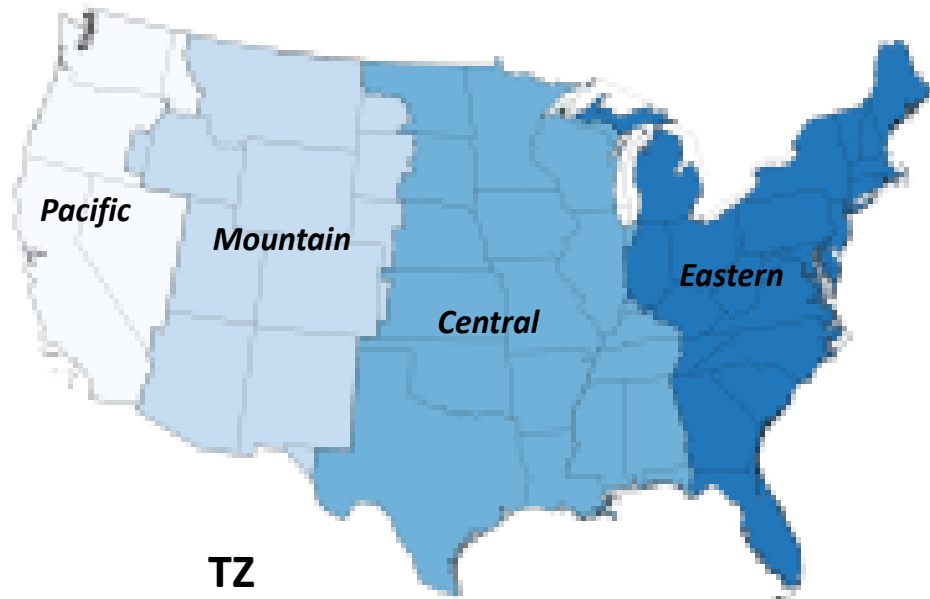
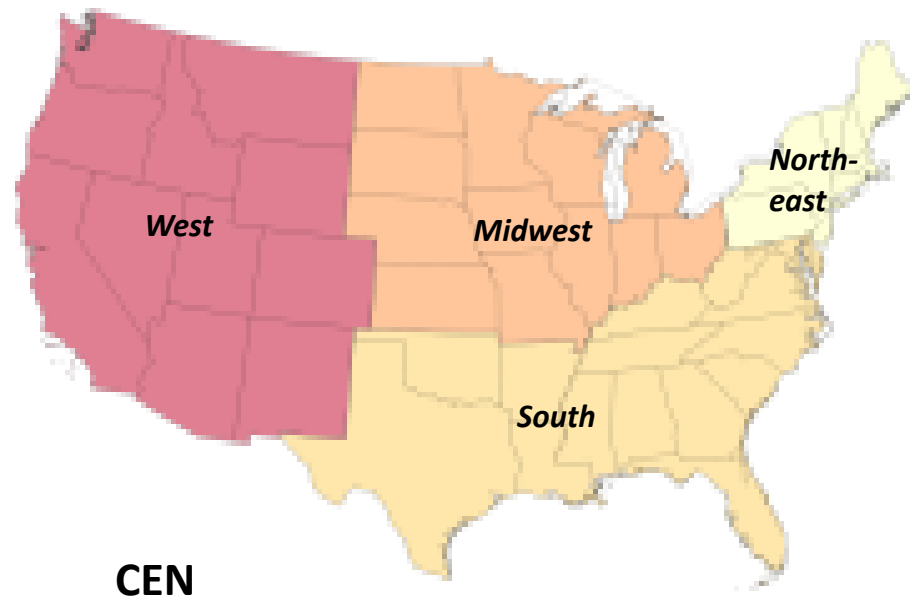
The overlapping relations are  
**automatically derived from data**



Nodes	
CEN	3
NDC	4
comb	51
Edges	
input	61
inferred	3
overlapsinferred	3

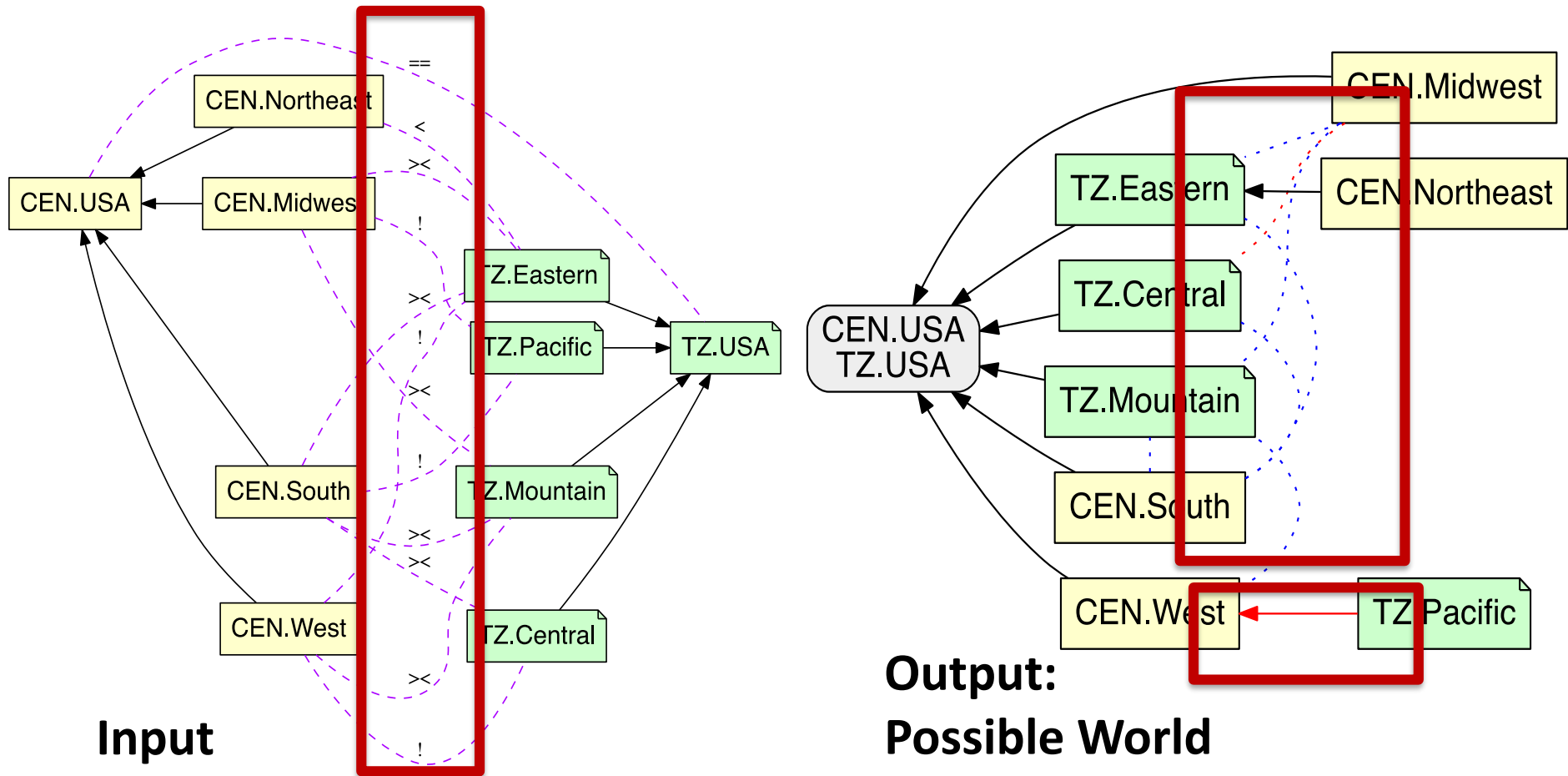


# Case 2: Census Region vs Time Zone



- ... but where do the *articulations* come from??
  - automatically derived from data
  - **expert input**

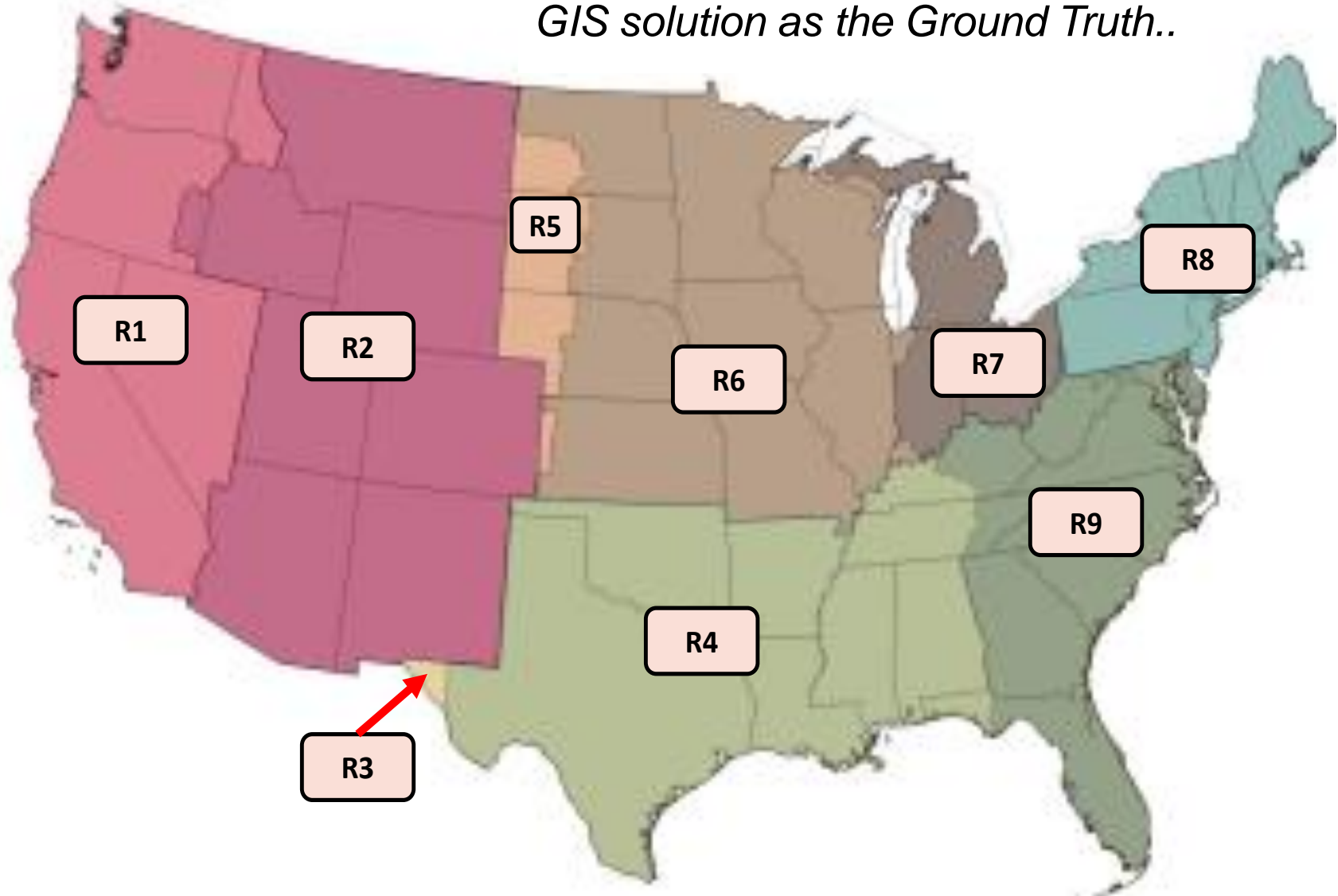
# Top-down regional alignment



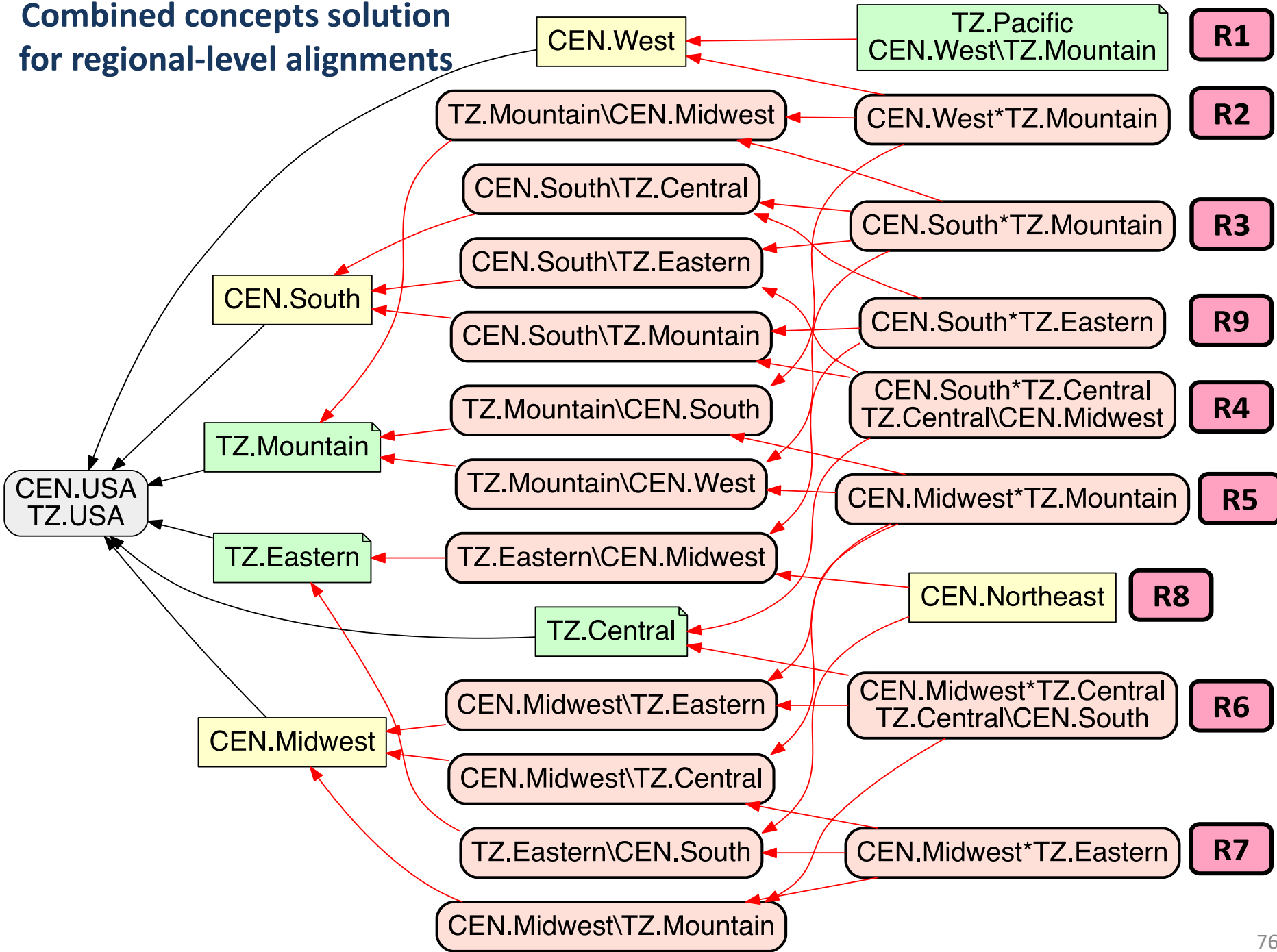


# How do we know if our 'expert articulations' are correct?

*GIS solution as the Ground Truth..*



Combined concepts solution  
for regional-level alignments



# Do the taxonomies have to be spatial in order to use RCC-5?

- **No!** The more typical cases for taxonomy alignment are usually between *non-spatial* taxonomies
  - for which no “GIS route” or direct visual cues about regional extensions are available
  - the use of RCC-5 as an alignment vocabulary is a suitable approach to perform a wide range of multi-hierarchy reconciliations

# Conclusion & Discussion

- Underscores the benefits of designing different alignment workflows (Bottom-up vs. Top-Down)
  - Bottom-up: non-overlapping relationships at the lowest-level articulations, not sure how to align the higher-level concepts
  - Top-Down: when there is often overlapping leaf-level relations.. Expert input will frequently be needed to establish such expectations under the top-down approach



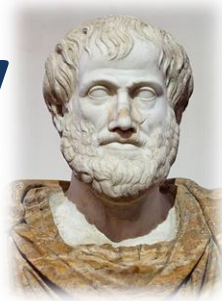
# Implications

- Logic-based taxonomy alignment approach
  - Disambiguate name-based taxonomy alignment over time
    - 40% of the concepts in biology taxonomies undergoes name change over time (Franz et al., 2016)
  - May mitigate problems in equivalent crosswalking
    - Membership condition problem that was often criticized in crosswalking
  - Preserves the original taxonomies while providing an alignment view
    - Solve data integration problems that happen in the more coarse-grained relative crosswalking



# Some History

- ... Aristotle ...
- ... Euler ...
- ...
- ... Greg Whitbread ...



- [BPB93] J. H. Beach, S. Pramanik, and J. H. Beaman. Hierarchic taxonomic databases., Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision, 1993
- [Ber95] Walter G. Berendsohn. The concept of “potential taxa” in databases. Taxon, 44:207–212, 1995.
- [Ber03] Walter G. Berendsohn. MoReTax – Handling Factual Information Linked to Taxonomic Concepts in Biology. No. 39 in Schriftenreihe für Vegetationskunde. Bundesamt für Naturschutz, 2003.
- [GG03] M. Geoffroy and A. Güntsch. Assembling and navigating the potential taxon graph. In [Ber03], pages 71–82, 2003.
- [TL07] Thau, D., & Ludäscher, B. (2007). Reasoning about taxonomies in first-order logic. Ecological Informatics, 2(3), 195-209.
- [FP09] Franz, N. M., & Peet, R. K. (2009). Perspectives: towards a language for mapping relationships among taxonomic concepts. Systematics and Biodiversity, 7(1), 5-20.

