Lab 3: Parasite Specificity + Host Range

Today's Lab

1. Go over basics of data manipulation in R
2. Code along example working with GMPD
3. Set you free to answer your own question

What is R and why should should I care about knowing how to use it?

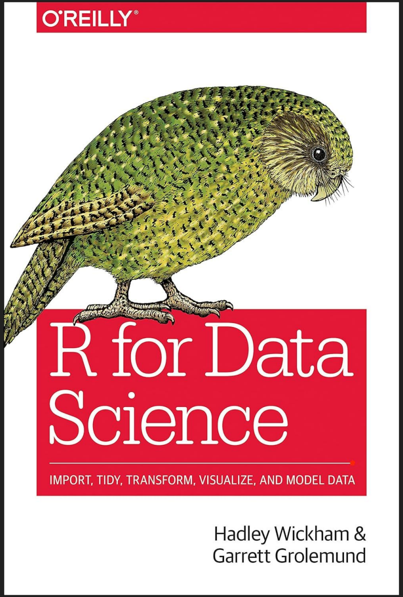
Statistical programming language

Based on precursor language called S

Good at working with data

- Manipulation
- Analysis
- Visualization

Free, Open-Sourced, Reproducible!



O'REILLY®

R for Data Science

IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA

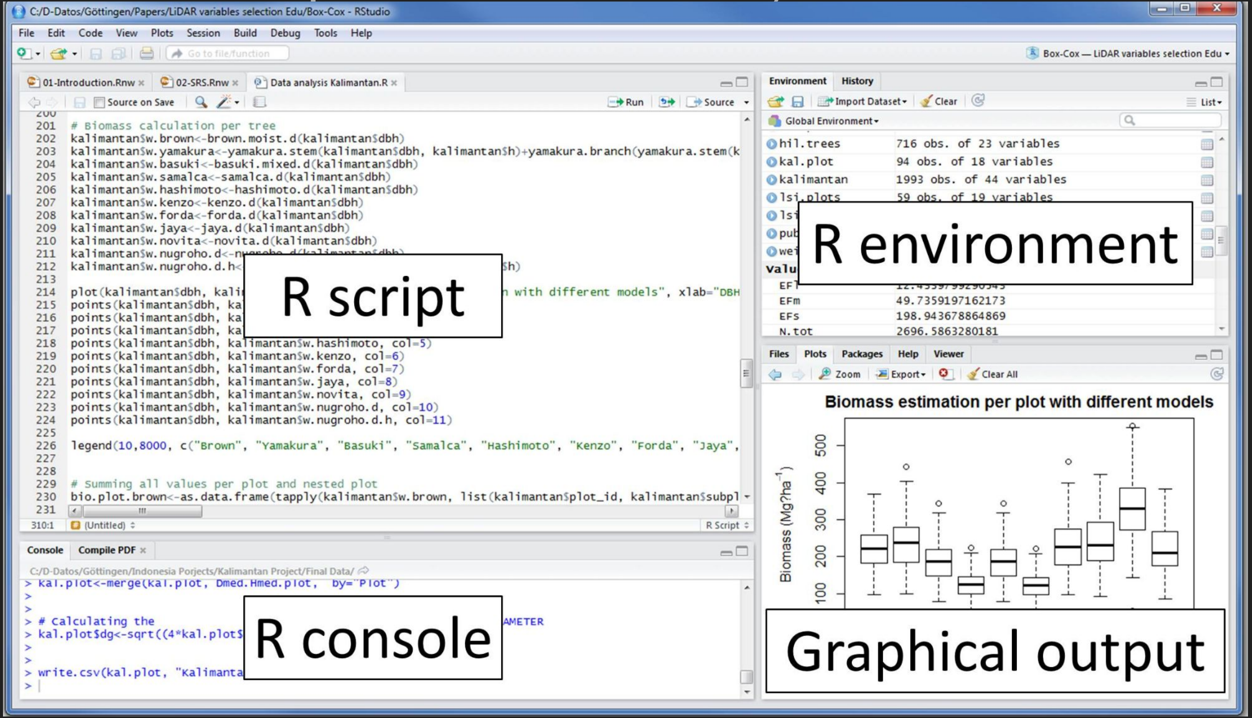
Hadley Wickham &
Garrett Grolemund

CRAN R

The Comprehensive R
Archive Network

Basic Terminology

R is the language, we'll be interacting with it through **RStudio** (an IDE: Integrated Development Environment)



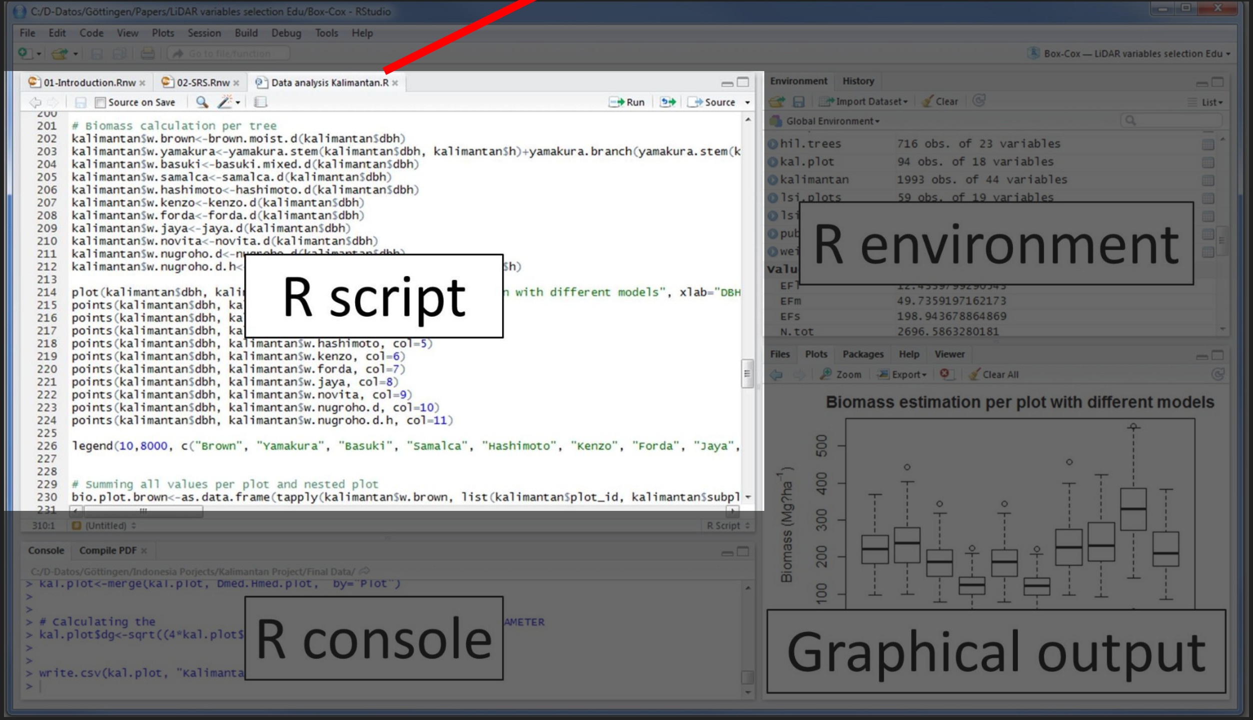
```
# Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yamakura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
kalimantan$w.nugroho.d<-n
kalimantan$w.nugroho.d.h<                              $h)

plot(kalimantan$dbh, kali                              n with different models", xlab="DBH
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)

legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",


# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
```

```
hil.trees     716 obs. of 23 variables
kal.plot      94 obs. of 18 variables
kalimantan    1993 obs. of 44 variables
lsi.plots     59 obs. of 19 variables
lsi
pub
wei
Valu
 EF
 EFm          49.7359197162173
 EFs          198.943678864869
 N.tot        2696.5863280181
```

```
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
> kal.plot<-merge(kal.plot, Dmed.Hmed.plot,  by="Plot")
>
>
> # Calculating the                              AMETER
> kal.plot$dg<-sqrt((4*kal.plot$
>
>
> write.csv(kal.plot, "Kalimanta
>
```

Basic Terminology

R Scripts: The file where your code lives! Creating scripts allows you to know exactly what you did, and reproduce it every time.



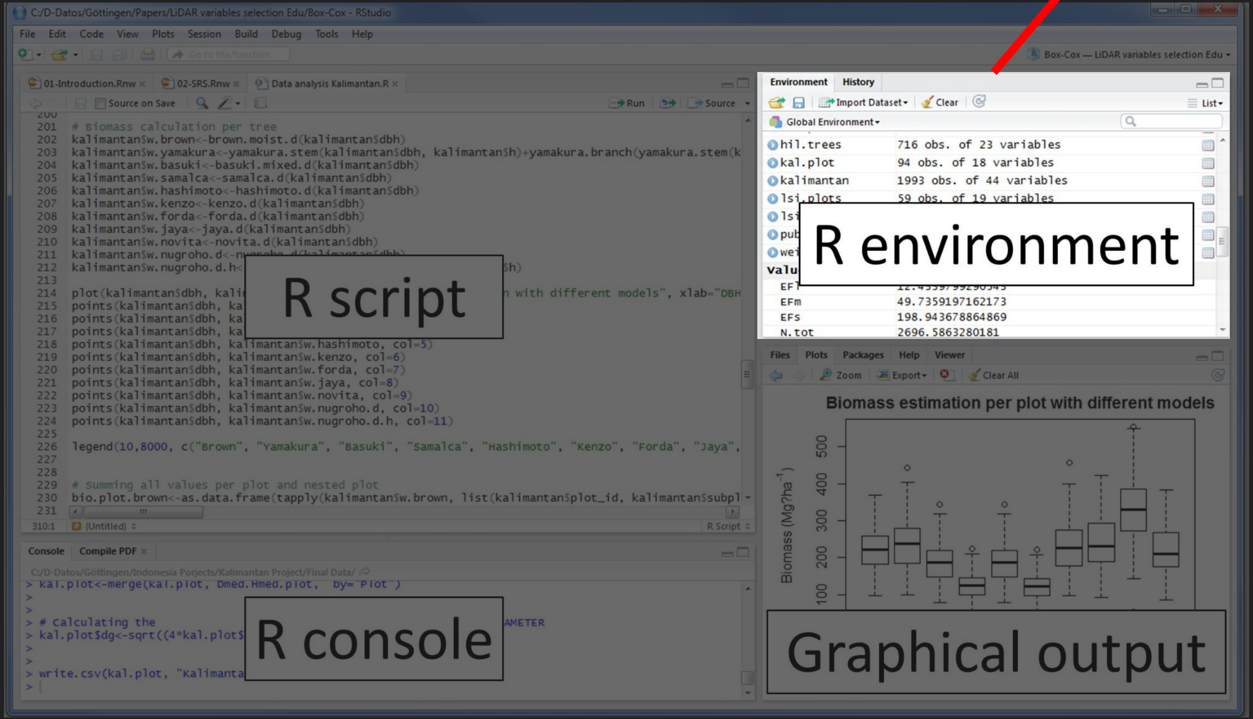
```
# Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yamakura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
kalimantan$w.nugroho.d<-n
kalimantan$w.nugroho.d.h<                                $h)

plot(kalimantan$dbh, kali                                n with different models", xlab="DBH
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)

legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",


# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
```

```
hil.trees     716 obs. of 23 variables
kal.plot      94 obs. of 18 variables
kalimantan    1993 obs. of 44 variables
lsi.plots     59 obs. of 19 variables
lsi
pub
wei
Valu
EFm           49.7359197162173
EFs           198.943678864869
N.tot         2696.5863280181
```

```
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
> kal.plot<-merge(kal.plot, Dmed.Hmed.plot,  by="Plot")
>
>
> # Calculating the                                AMETER
> kal.plot$dg<-sqrt((4*kal.plot$
>
>
> write.csv(kal.plot, "Kalimanta
>
```

Basic Terminology

R Environment: List of all the objects you're currently working with, loaded into your memory. Data objects, custom functions, etc.



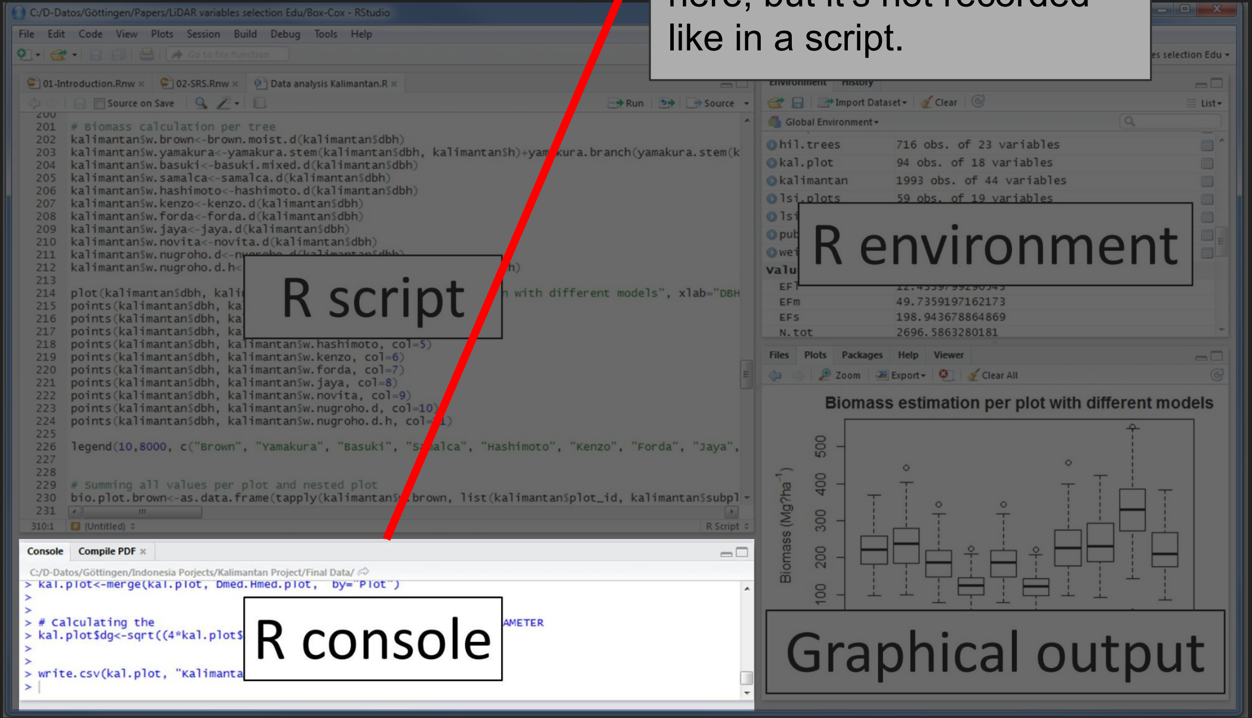
```
# Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yamakura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
kalimantan$w.nugroho.d<-nugroho.d(kalimantan$dbh)
kalimantan$w.nugroho.d.h<                        $h)

plot(kalimantan$dbh, kali                        n with different models", xlab="DBH
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)

legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",


# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
```

```
hil.trees      716 obs. of 23 variables
kal.plot       94 obs. of 18 variables
kalimantan     1993 obs. of 44 variables
lsi.plots      59 obs. of 19 variables
lsi
pub
wei
Valu
EF
EFm            49.7359197162173
EFs            198.943678864869
N.tot          2696.5863280181
```

```
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
> kal.plot<-merge(kal.plot, Dmed.Hmed.plot,  by="Plot")
>
>
> # Calculating the                          AMETER
> kal.plot$dg<-sqrt((4*kal.plot$
>
>
> write.csv(kal.plot, "Kalimanta
>
```

Basic Terminology

R Console: Shows the most recent code and commands you've run. You can also type a run code here, but it's not recorded like in a script.



R script

```
# Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yam kura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
kalimantan$w.nugroho.d<-nugroho.d(kalimantan$dbh)
kalimantan$w.nugroho.d.h<                                        h)

plot(kalimantan$dbh, kali                                  h with different models", xlab="DBH
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)

legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",


# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
```

R console

```
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
> kal.plot<-merge(kal.plot, Dmed.Hmed.plot,  by="Plot")
>
>
> # Calculating the                                AMETER
> kal.plot$dg<-sqrt((4*kal.plot$
>
>
> write.csv(kal.plot, "Kalimanta
> 
```

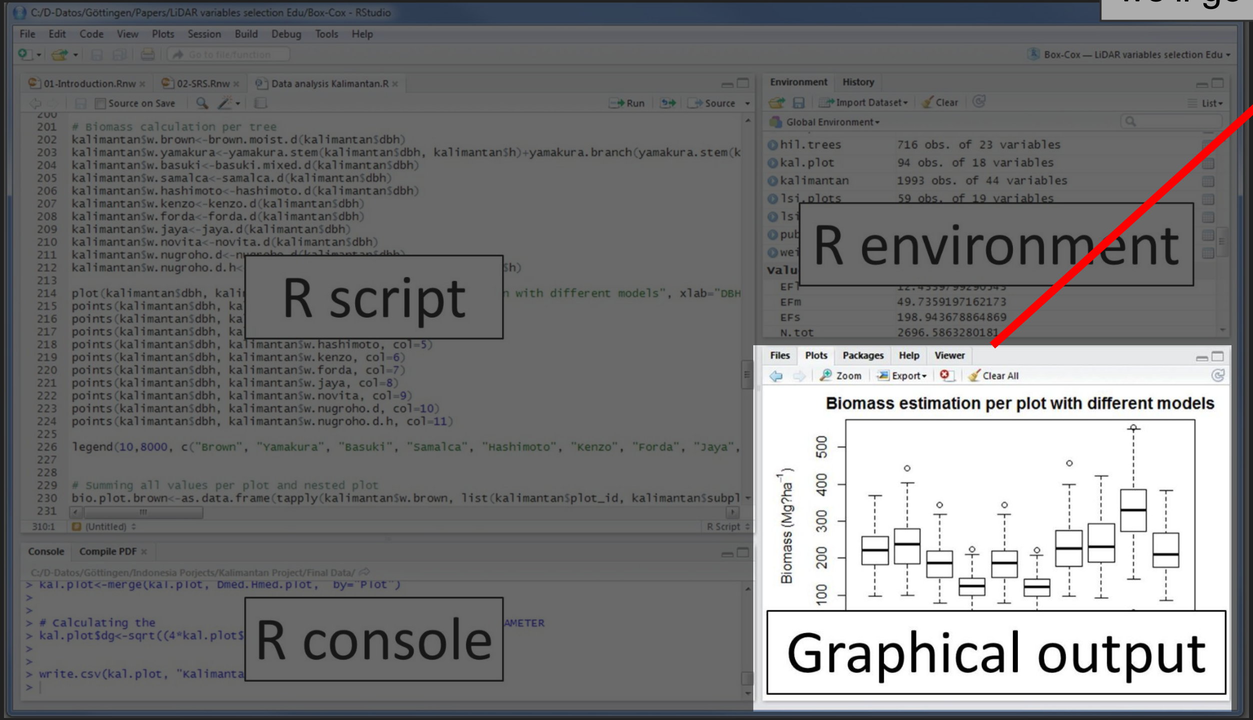
R environment

```
hil.trees      716 obs. of 23 variables
kal.plot       94 obs. of 18 variables
kalimantan     1993 obs. of 44 variables
lsi.plots      59 obs. of 19 variables
lsi
pub
wei
Valu
EFl            12.4339799290545
EFm            49.7359197162173
EFs            198.943678864869
N.tot          2696.5863280181
```

Graphical output

Basic Terminology

<u>Other Stuff:</u> There are a few tabs here, all of them useful. Most of the names are self-explanatory, and we'll go over them in class



R script

```
# Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yamakura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
kalimantan$w.nugroho.d<-nugroho.d(kalimantan$dbh)
kalimantan$w.nugroho.d.h<                                $h)

plot(kalimantan$dbh, kali                                n with different models", xlab="DBH
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)

legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",


# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
```

R console

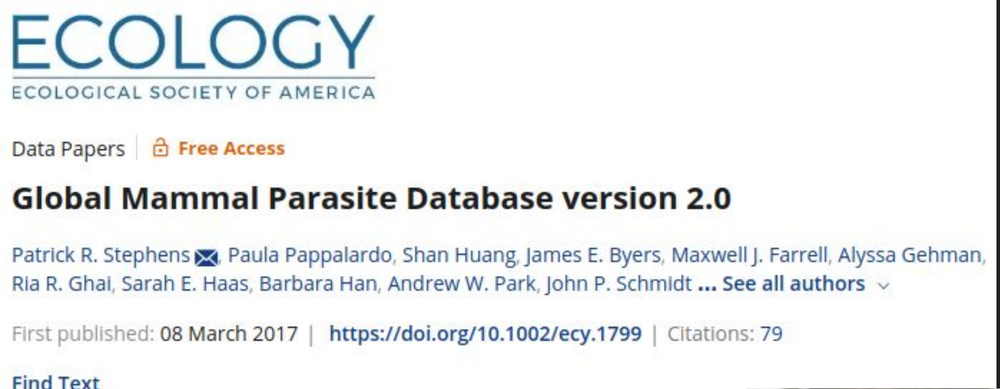
```
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
> kal.plot<-merge(kal.plot, Dmed.Hmed.plot,  by="Plot")
>
>
> # Calculating the                          AMETER
> kal.plot$dg<-sqrt((4*kal.plot$
>
>
> write.csv(kal.plot, "Kalimanta
> 
```

R environment

```
hil.trees        716 obs. of 23 variables
kal.plot         94 obs. of 18 variables
kalimantan       1993 obs. of 44 variables
lsi.plots        59 obs. of 19 variables
lsi
pub
wei
Valu
  EF
  EFm            49.7359197162173
  EFs            198.943678864869
  N.tot          2696.5863280181
```

Graphical output

Our Dataset



ECOLOGY

ECOLOGICAL SOCIETY OF AMERICA

Data Papers | Free Access

Global Mammal Parasite Database version 2.0

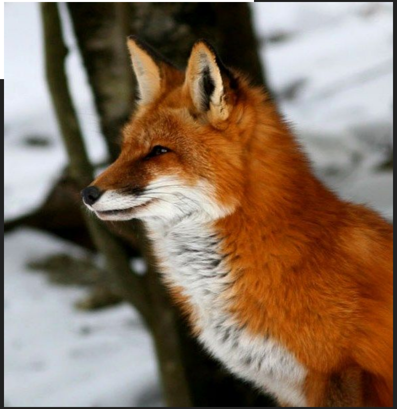
Patrick R. Stephens, Paula Pappalardo, Shan Huang, James E. Byers, Maxwell J. Farrell, Alyssa Gehman, Ria R. Ghai, Sarah E. Haas, Barbara Han, Andrew W. Park, John P. Schmidt **... See all authors**

First published: 08 March 2017 | https://doi.org/10.1002/ecy.1799 | Citations: 79

Find Text

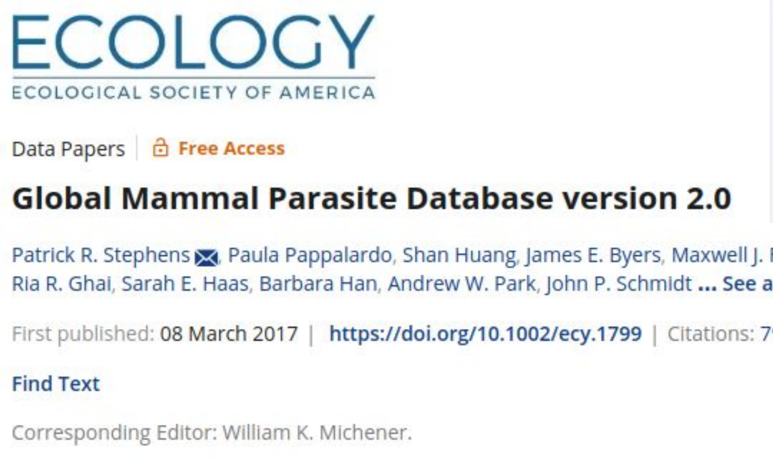
Corresponding Editor: William K. Michener.

https://parasites.nunn-lab.org/



Hosts: Ungulates, Carnivores, and Primates

Our Dataset



ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Data Papers | Free Access

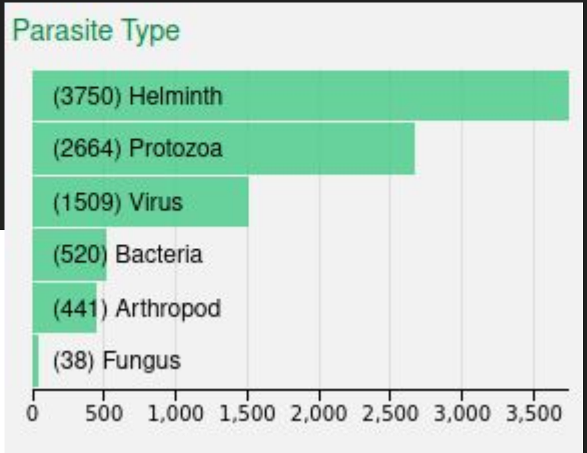
Global Mammal Parasite Database version 2.0

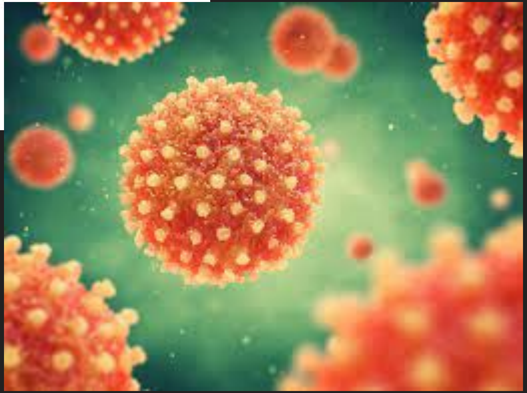
Patrick R. Stephens, Paula Pappalardo, Shan Huang, James E. Byers, Maxwell J. Farrell, Alyssa Gehman, Ria R. Ghai, Sarah E. Haas, Barbara Han, Andrew W. Park, John P. Schmidt ... See all authors

First published: 08 March 2017 | https://doi.org/10.1002/ecy.1799 | Citations: 79

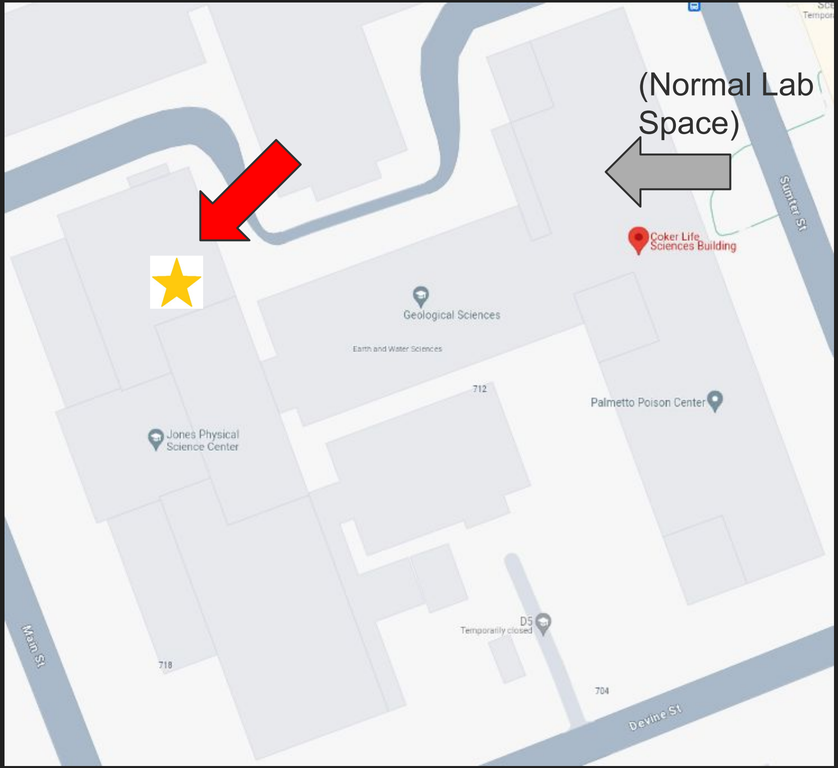
Find Text

Corresponding Editor: William K. Michener.





Where is lab today?

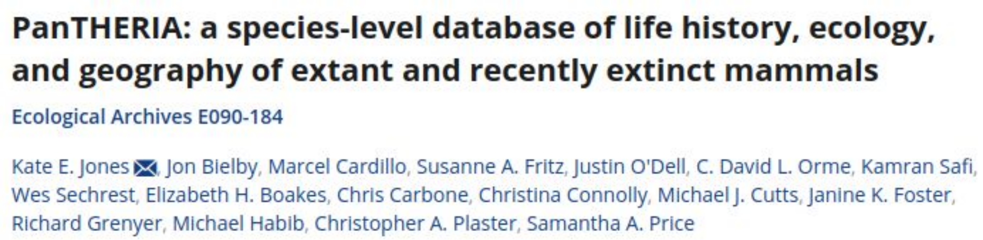


(Normal Lab Space)

The Question:

What are predictors of parasite richness across mammalian hosts?

Data I’ve given you



PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals

Ecological Archives E090-184

Kate E. Jones ✉, Jon Bielby, Marcel Cardillo, Susanne A. Fritz, Justin O'Dell, C. David L. Orme, Kamran Safi, Wes Sechrest, Elizabeth H. Boakes, Chris Carbone, Christina Connolly, Michael J. Cutts, Janine K. Foster, Richard Grenyer, Michael Habib, Christopher A. Plaster, Samantha A. Price

Host traits: extracted from PanTHERIA

Morphometric data: Adult body mass, neonatal body mass, adult forearm length, basal metabolic rate…

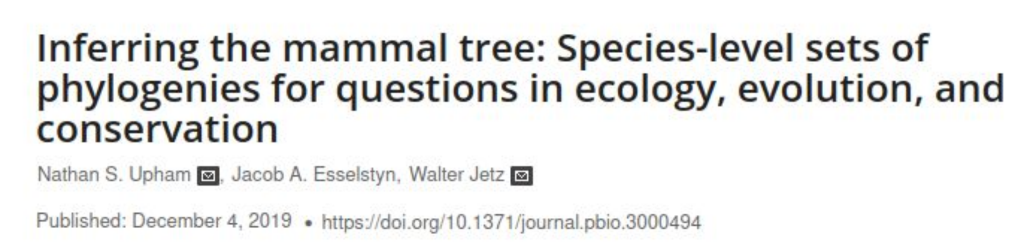
Pace of Life data: Gestation length, litter size, age at eyes opening, age at first reproduction…

Life-History/ecological data: Habitat breadth, population density, trophic level

Geographic Range: Total range size, mean latitude, max/min latitude, temperature or PET across range…

Read the MetaData!!!

Data I’ve given you



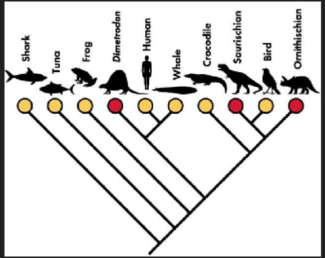
Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation

Nathan S. Upham, Jacob A. Esselstyn, Walter Jetz

Published: December 4, 2019 • https://doi.org/10.1371/journal.pbio.3000494

Host phylogenetic information

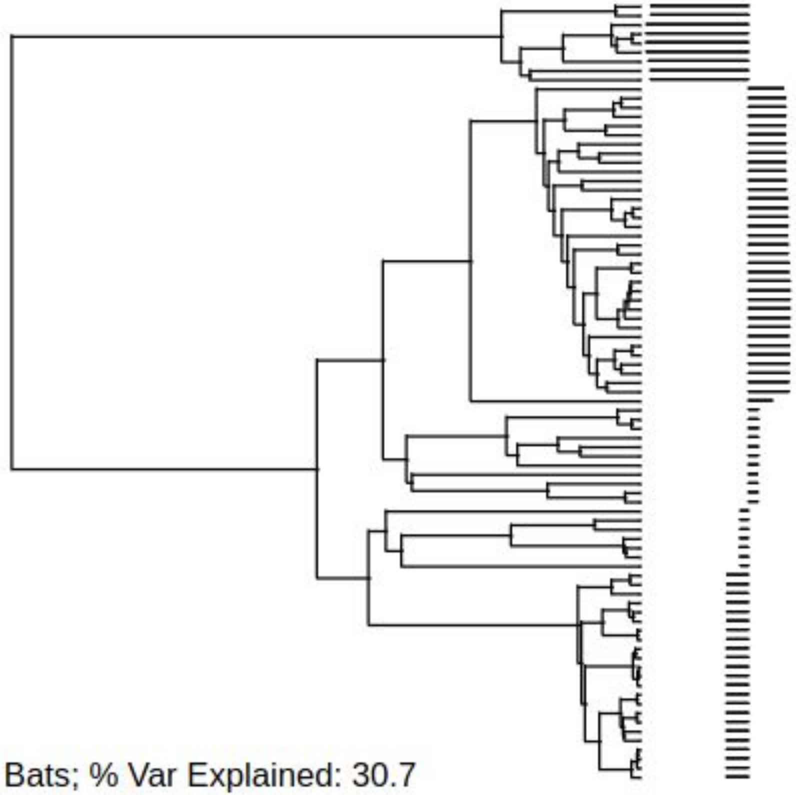
Based on Phylogenetic Tree From Upham, Esselstyn, and Jetz (2019).

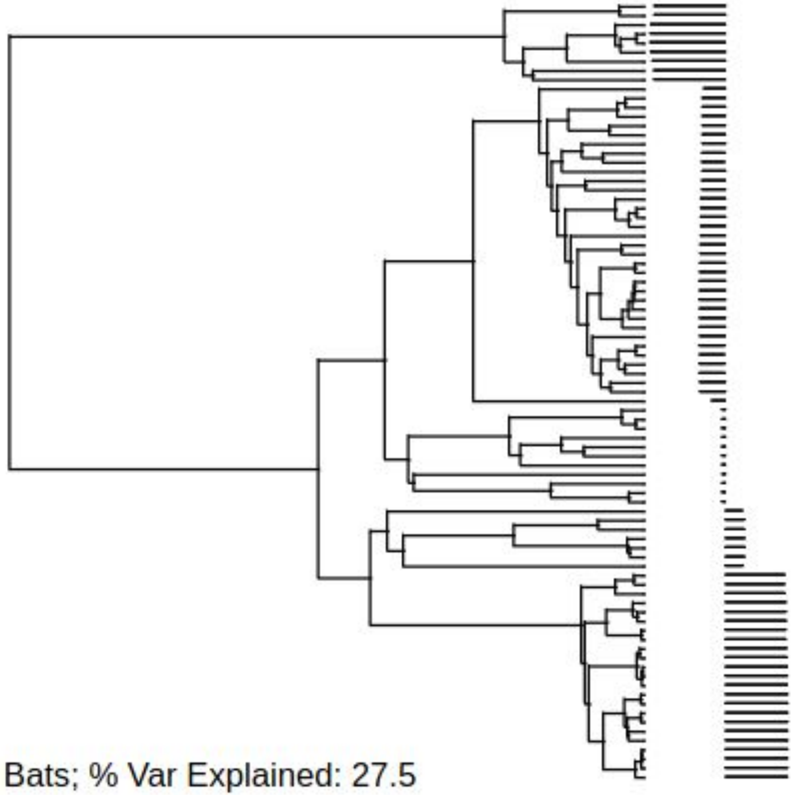


How do we convert this to something we can use for prediction?

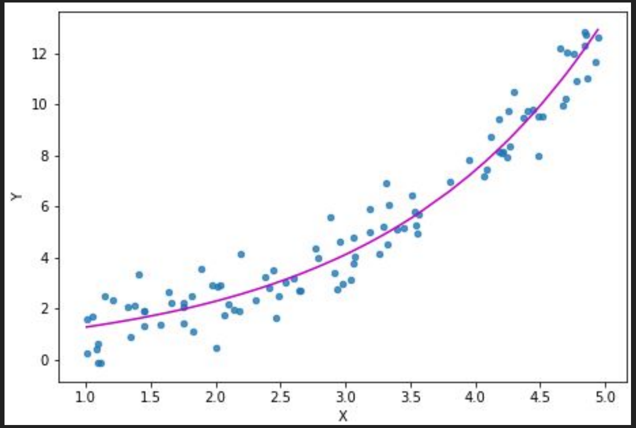
Different approaches, but the one I chose was eigenvalue decomposition

Long story short, a way of condensing a complicated phylogeny into a series of continuous axis. You lose information doing this, but what’s left can still be quite useful!


Bats; % Var Explained: 30.7


Bats; % Var Explained: 27.5

Generalized Linear Models



Really a general term, describing a really simple framework

Data = Model + Error

$$\hat{Y}=\beta_0 + \beta_1 X$$

If this looks like linear regression, that’s because it is!

The “Generalized” form however just means that we’re specific about our assumptions about error distributions in a way that means we don’t just have to use continuous numerical predictors.

R takes care of the majority of this for us, but we need to know how to interpret them

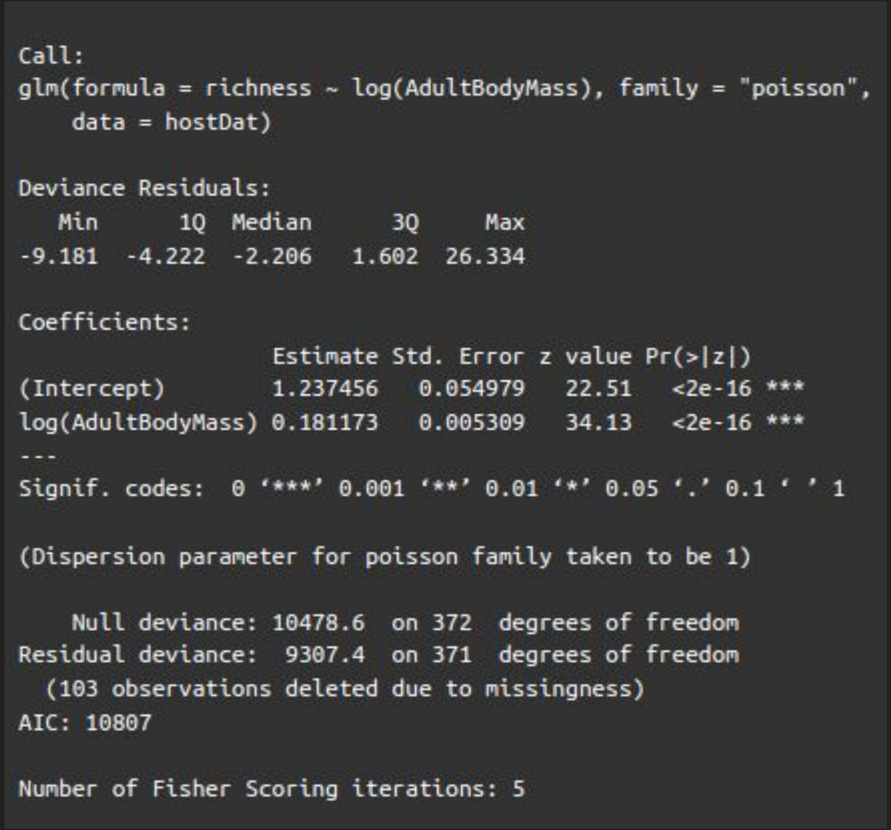
Generalized Linear Models in R

In R, we can make a glm with the `glm()` function (go figure!)

?glm() will show you that there’s many arguments, but the most important ones for our purposes are the “formula” and “data”

When you use `summary()` to look at a glm model, it’ll look something like this…

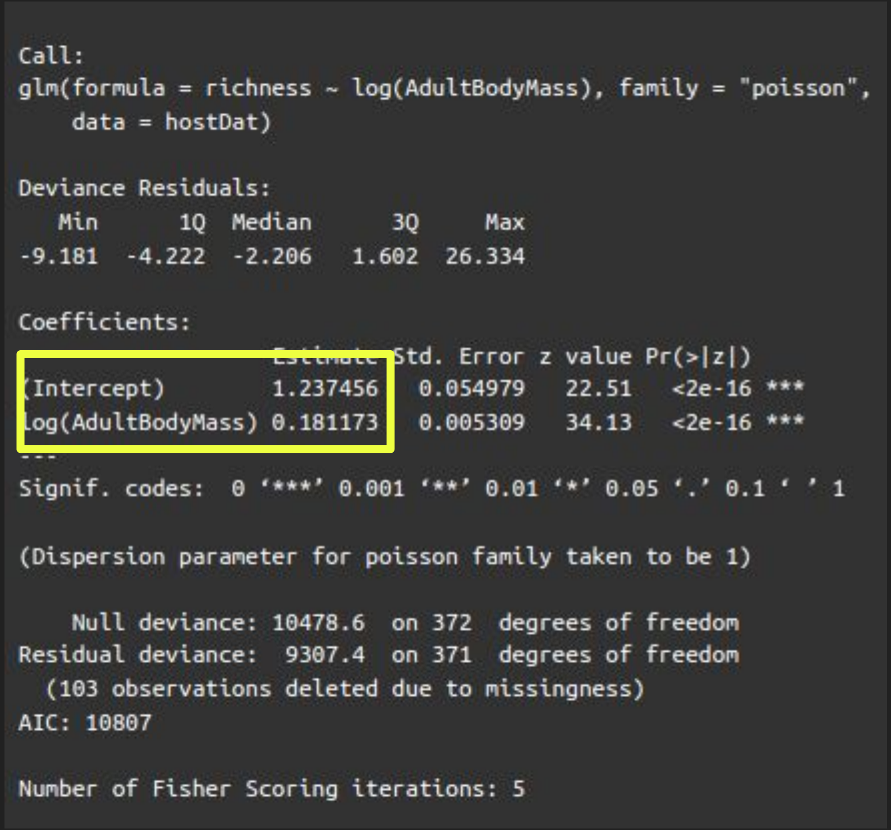
This has a lot of information, but the ones I want you to look at for interpretation are…



```
Call:
glm(formula = richness ~ log(AdultBodyMass), family = "poisson",
    data = hostDat)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-9.181  -4.222  -2.206   1.602  26.334

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.237456   0.054979   22.51   <2e-16 ***
log(AdultBodyMass) 0.181173   0.005309   34.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10478.6  on 372  degrees of freedom
Residual deviance:  9307.4  on 371  degrees of freedom
  (103 observations deleted due to missingness)
AIC: 10807

Number of Fisher Scoring iterations: 5
```

Generalized Linear Models in R

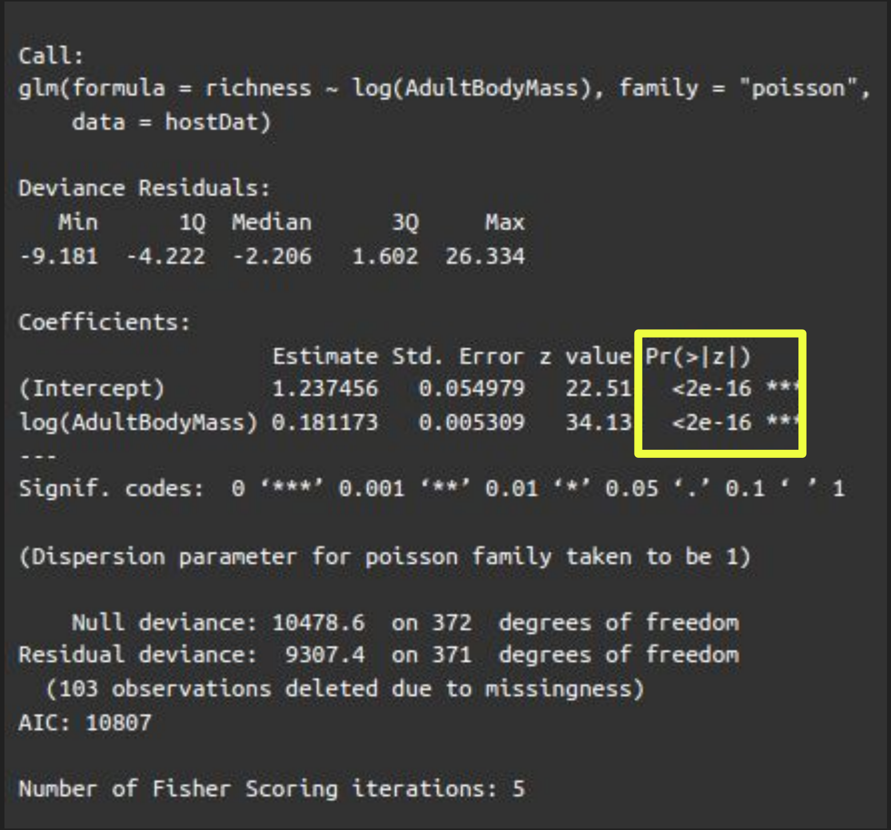
1. The coefficient estimates
   a. These describe the relationships between your predictors and the response variable
   b. Can use these to understand whether the relationship is positive or negative (for our case-a poisson link function, mean y changes by exp(β1) per unit change of x)



```
Call:
glm(formula = richness ~ log(AdultBodyMass), family = "poisson",
    data = hostDat)

Deviance Residuals:
   Min       1Q   Median       3Q      Max  
-9.181   -4.222   -2.206    1.602   26.334  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)        1.237456   0.054979   22.51   <2e-16 ***
log(AdultBodyMass) 0.181173   0.005309   34.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10478.6  on 372  degrees of freedom
Residual deviance:  9307.4  on 371  degrees of freedom
  (103 observations deleted due to missingness)
AIC: 10807

Number of Fisher Scoring iterations: 5
```

Generalized Linear Models in R

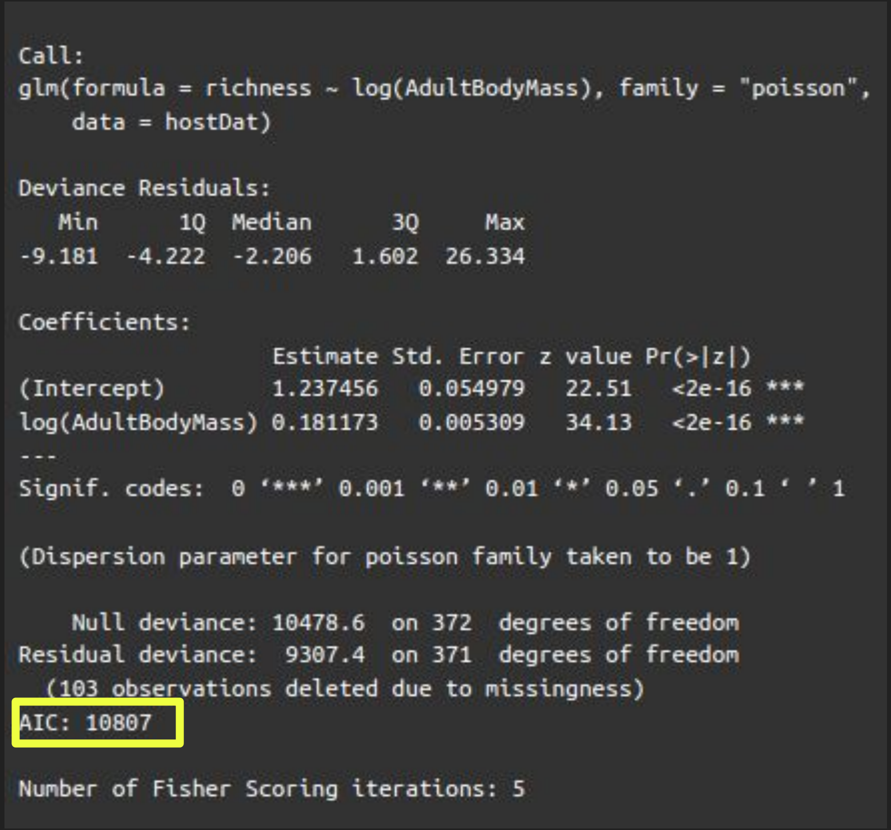
1. The coefficient estimates
   a. These describe the relationships between your predictors and the response variable
   b. Can use these to understand whether the relationship is positive or negative (for our case-a poisson link function, mean y changes by exp(β1) per unit change of x)
2. P-values; give us an idea of significance
   a. P > 0.05, then the term is significant (and helps explain variation in Y).



```
Call:
glm(formula = richness ~ log(AdultBodyMass), family = "poisson",
    data = hostDat)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-9.181  -4.222  -2.206   1.602   26.334

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.237456   0.054979   22.51   <2e-16 ***
log(AdultBodyMass) 0.181173   0.005309   34.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10478.6  on 372  degrees of freedom
Residual deviance:  9307.4  on 371  degrees of freedom
  (103 observations deleted due to missingness)
AIC: 10807

Number of Fisher Scoring iterations: 5
```

Generalized Linear Models in R

1. The coefficient estimates
    a. These describe the relationships between your predictors and the response variable
    b. Can use these to understand whether the relationship is positive or negative (for our case-a poisson link function, mean y changes by exp($\beta 1$) per unit change of x)
2. P-values; give us an idea of significance
    a. $P > 0.05$, then the term is significant (and helps explain variation in Y).
3. Akaike Information Criterion (AIC)



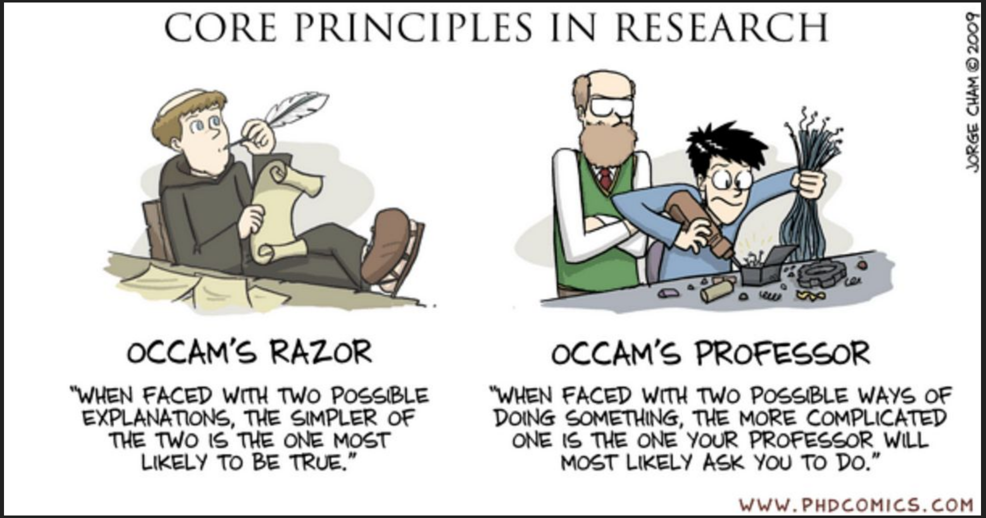
```
Call:
glm(formula = richness ~ log(AdultBodyMass), family = "poisson",
    data = hostDat)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-9.181   -4.222   -2.206    1.602   26.334

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.237456   0.054979   22.51   <2e-16 ***
log(AdultBodyMass) 0.181173   0.005309   34.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10478.6  on 372  degrees of freedom
Residual deviance:  9307.4  on 371  degrees of freedom
  (103 observations deleted due to missingness)
AIC: 10807

Number of Fisher Scoring iterations: 5
```

Model Comparison through AIC

When we're comparing models, there's often a tradeoff between goodness of fit and **parsimony**.

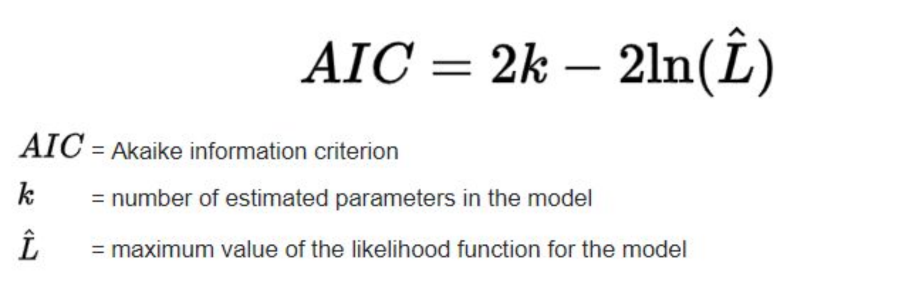
The idea of **parsimony** stem's from Occam's razor-when comparing among equally supported explanations, the one that requires the fewest assumptions is usually correct.

There are many ways you can evaluate model performance in light of these two ideas, but one of the most common is **AIC**


CORE PRINCIPLES IN RESEARCH
JORGE CHAM © 2009
OCCAM'S RAZOR
"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."
OCCAM'S PROFESSOR
"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."
WWW.PHDCOMICS.COM

Model Comparison through AIC

AIC: Akaike Information Criterion

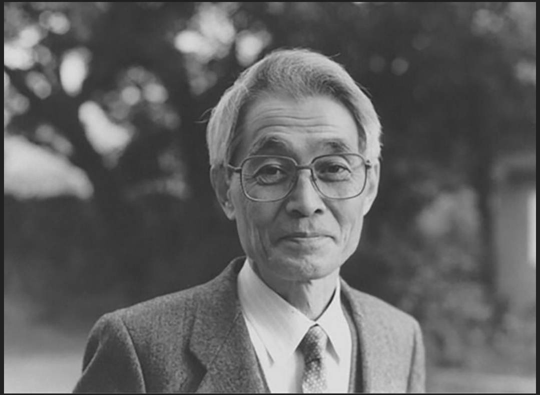


$$AIC = 2k - 2\ln(\hat{L})$$

AIC = Akaike information criterion

k = number of estimated parameters in the model

$\hat{L}$ = maximum value of the likelihood function for the model



The lower the AIC score, the better.

Models are penalized by the number of parameters (k), but rewarded by goodness of fit (L)

Model with the lowest AIC score generally has most support given the data, though if the difference between models is <2 you can’t really distinguish between them

Note: You can only compare models trained on the same data! (Otherwise the comparison is meaningless!)

Mystery Host(s)

How many parasites do you think they have?

We can make a guess using the `predict()` function!