Determining penguin phenology
from a very large force sensor data set

Philipp Boersch-Supan British Trust for Ornithology @pboesu
Helen Peat & Phil Trathan British Antarctic Survey

Study system: **Macaroni penguin** *Eudyptes chrysolophus*

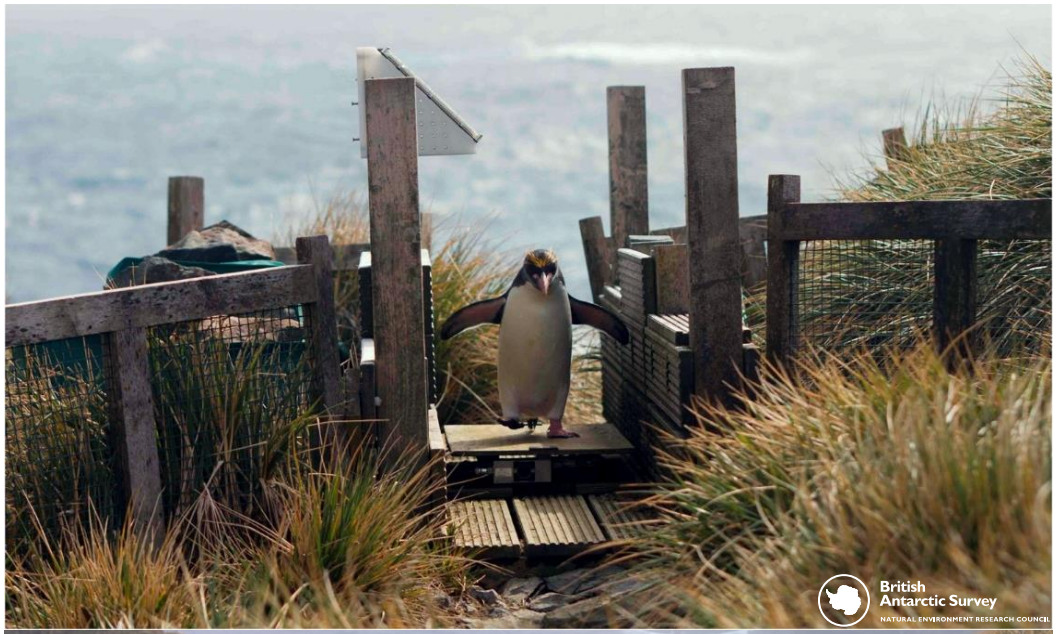
- 6 million breeding pairs globally
- single largest seabird consumer of prey biomass (krill, lanternfish)
- populations strongly declining

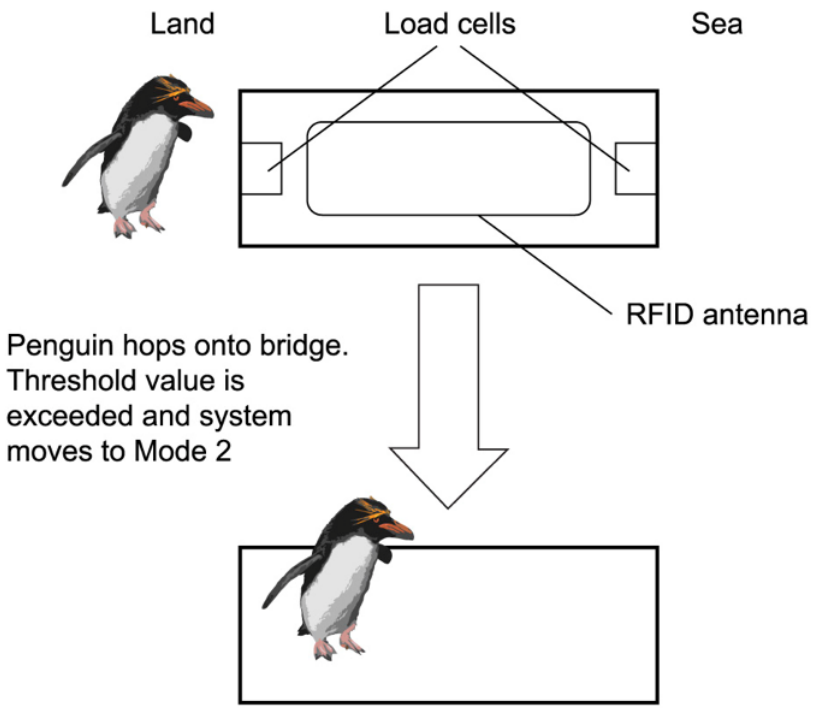
Original motivation for data collection:
Study body condition, foraging success, provisioning rates to better understand demographic rates

Image: BAS

@pboesu

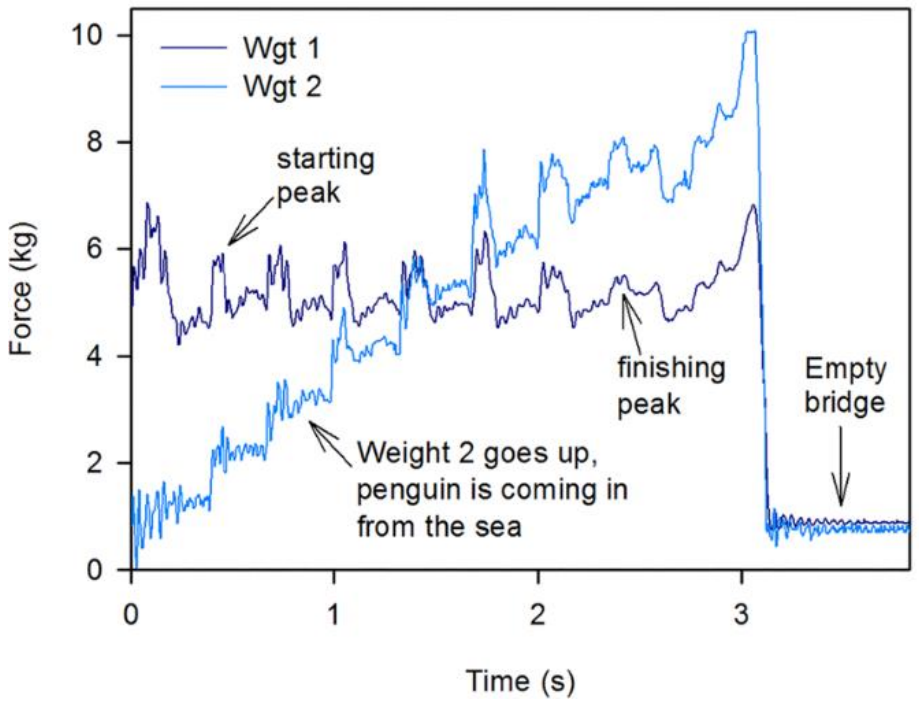
The penguin weighbridge of the British Antarctic Survey





@pboesu

Measurement principle



(1) RFID read identifies individual
(2) Integrate force over time to get penguin mass
(3) Difference between outbound and inbound mass = meal mass

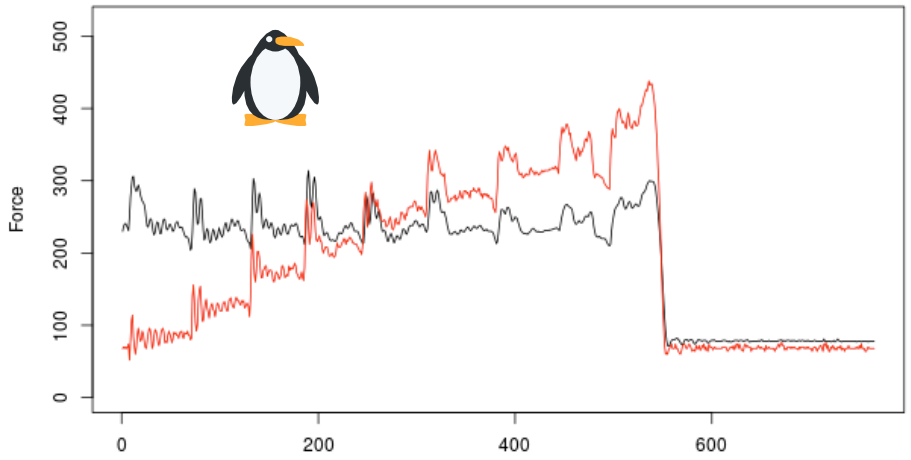
accurate mass requires high-frequency sampling:
50,000-80,000 crossings/season = **40 – 60 million raw data points/season**

Big data challenge at processing stage

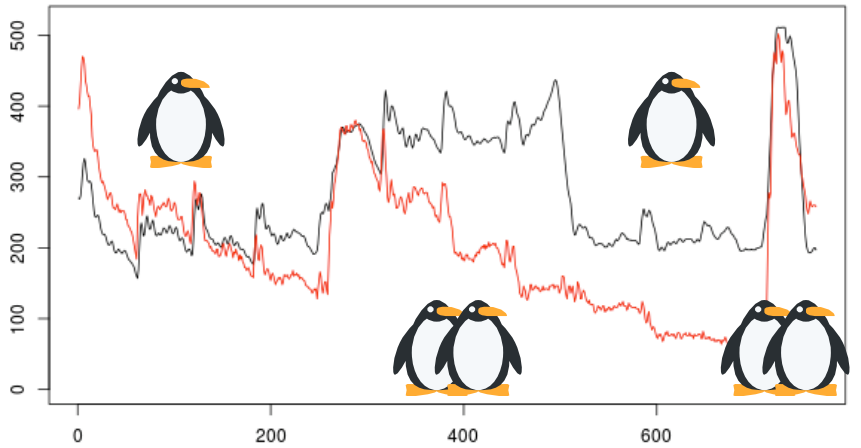
Afanasyev et al. 2015 PLoS ONE

and, penguins don't play by the rules…

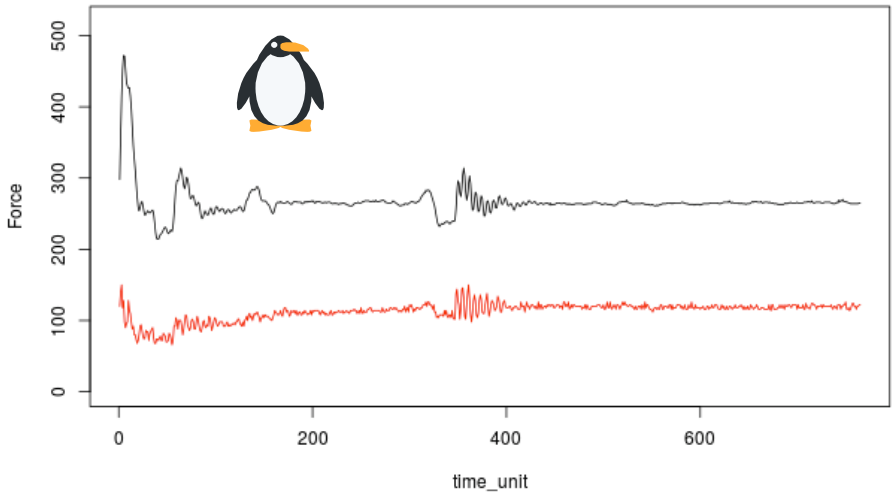
design case: 1 walking bird



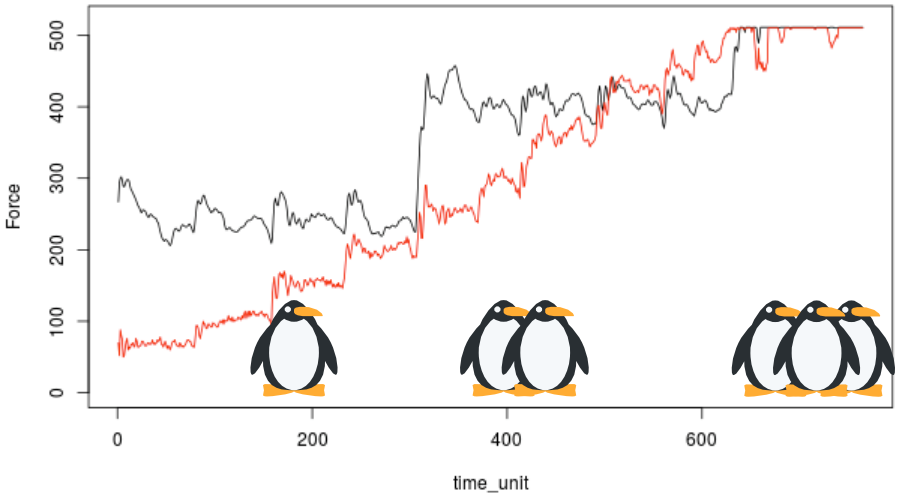
too busy: 3 birds in succession



not busy enough: 1 standing bird



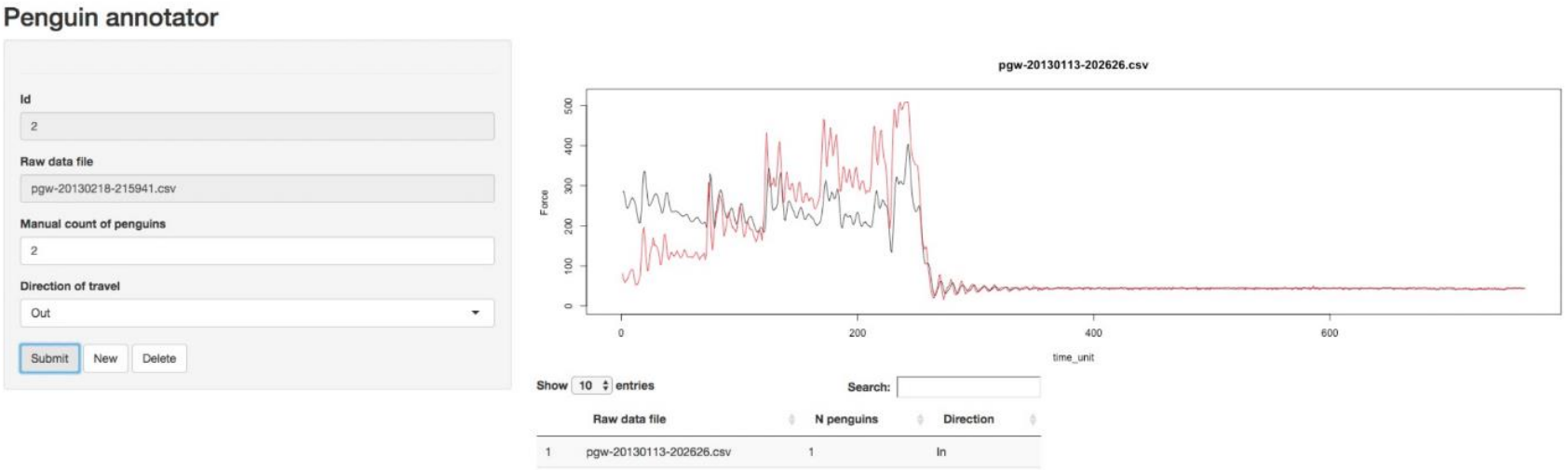
way too busy: 3 birds in succession



- Multiple crossings account for c. 30% of crossing birds
- Doesn't interfere with tag detection *per se*, but limits detection of directionality, calculation of weights, assignment of weights to individuals

@pboesu

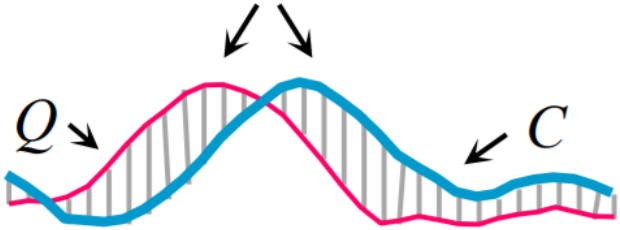
Step 1: The penguin annotator (shiny + RPostgres)



- Easy manual classification and enumeration of penguins
- but 50k+ files/yr prohibitively time intensive
- Simple decision trees failed (drastic changes in body mass during season)
- **Need a classifier that has 'time-series shape recognition'**

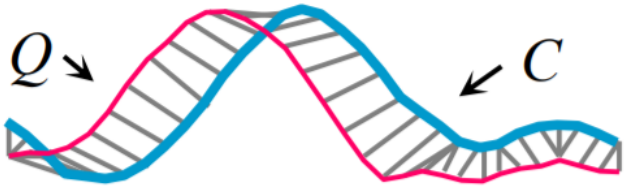
Step 2: Dynamic Time Warping

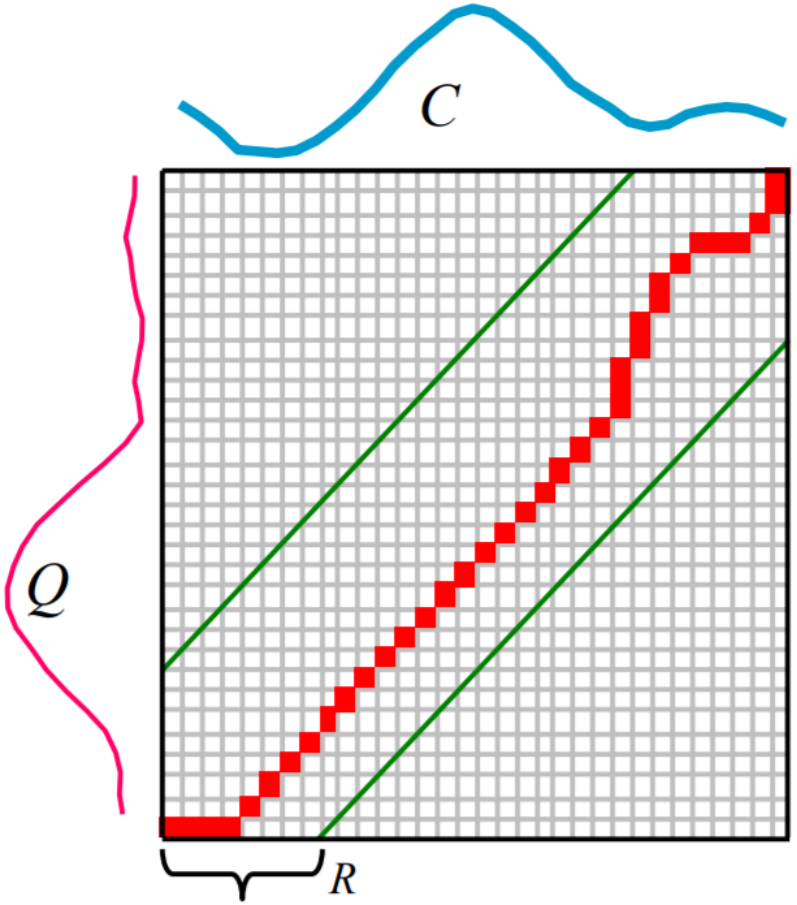
Similar, but out of phase peaks ...



... produce a large Euclidean distance.

However this can be corrected by DTWs nonlinear alignment.



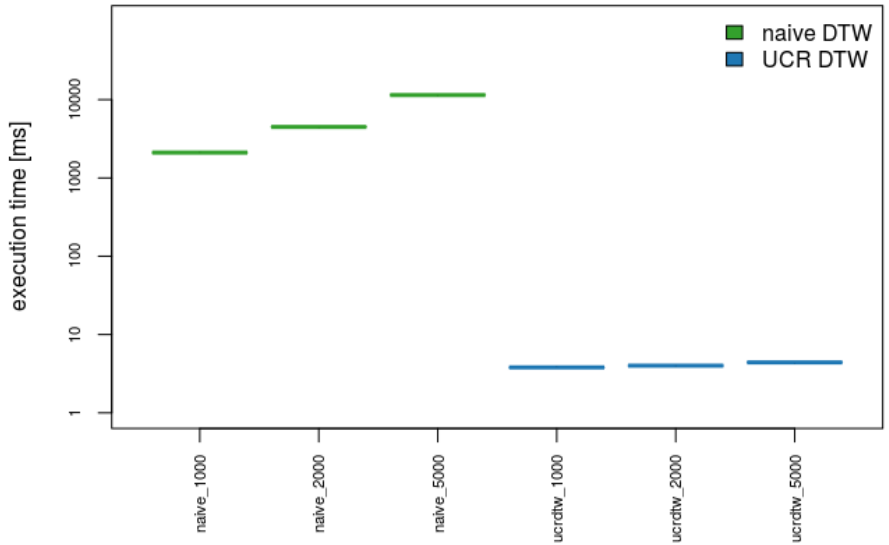


Rakthanmanon et al. 2012, Proc SIGKDD

- very good accuracy >90%
- **too costly (~ 1 min) to compute & compare full warping path for each crossing**

Ultrafast Dynamic Time Warping to the rescue!

- 'best match' search, exploits early abandoning on multiple levels
- 2-3 orders of magnitude faster than naïve DTW comparisons
- Open C++ source available from Rakthanmanon et al. 2012

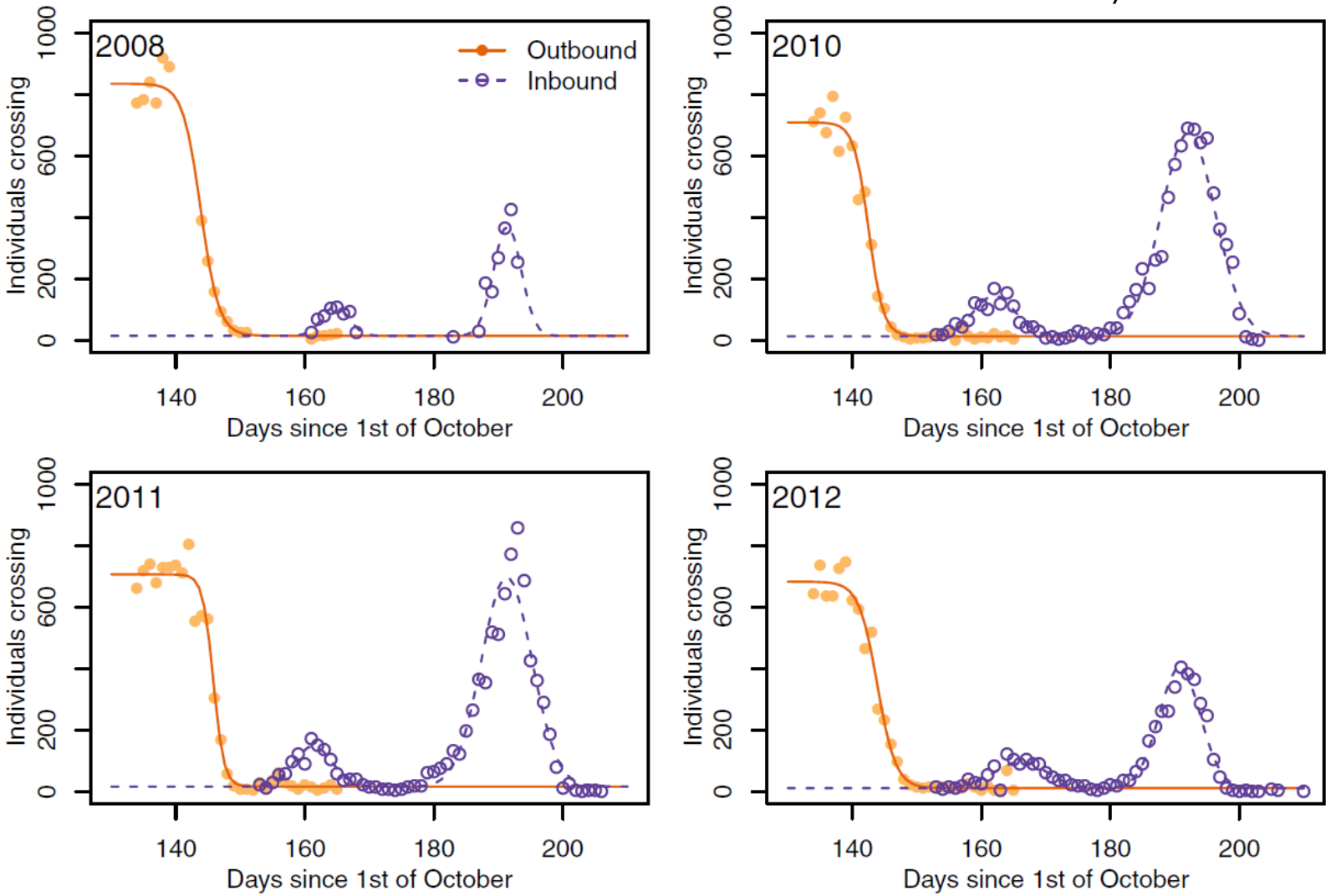


- R bindings now implemented in `install.packages("rucrdtw")`
- Could be of interest to accelerometer/acoustic data?!
- **A single season of penguin crossings can now be classified in <1hr**
  - **i.e. counts, directions for statistical modelling**

Rakthanmanon et al. 2012, Proc SIGKDD
Boersch-Supan 2016, J Open Source Software

Quantifying event timing from daily counts

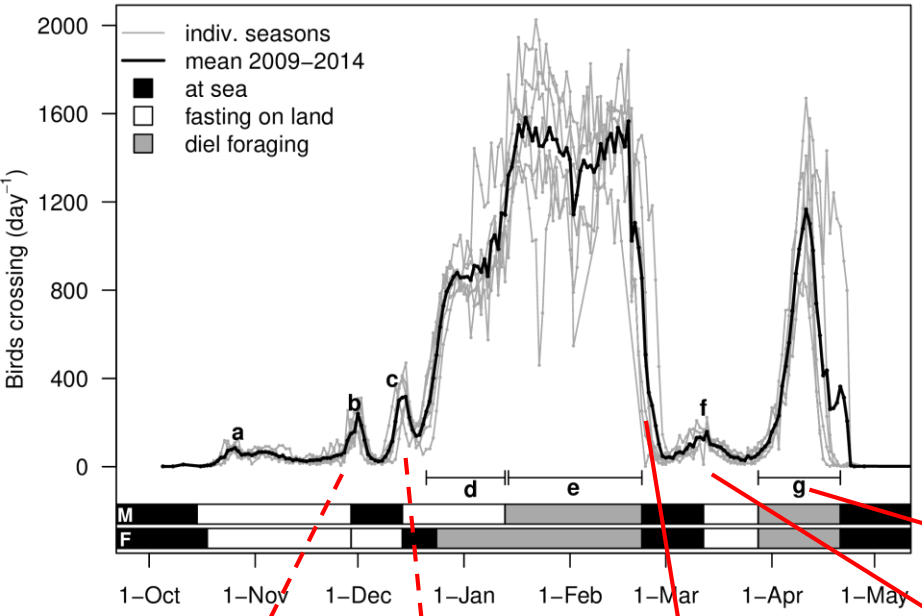
Inference approach:
Gaussian mixture model expressed as hierarchical non-linear regression
Bayesian estimation w/ informative priors

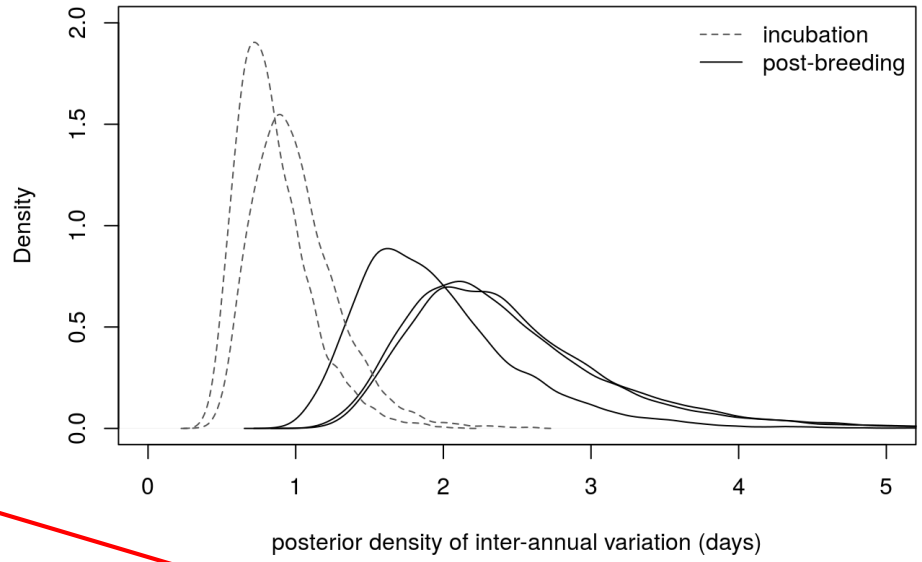


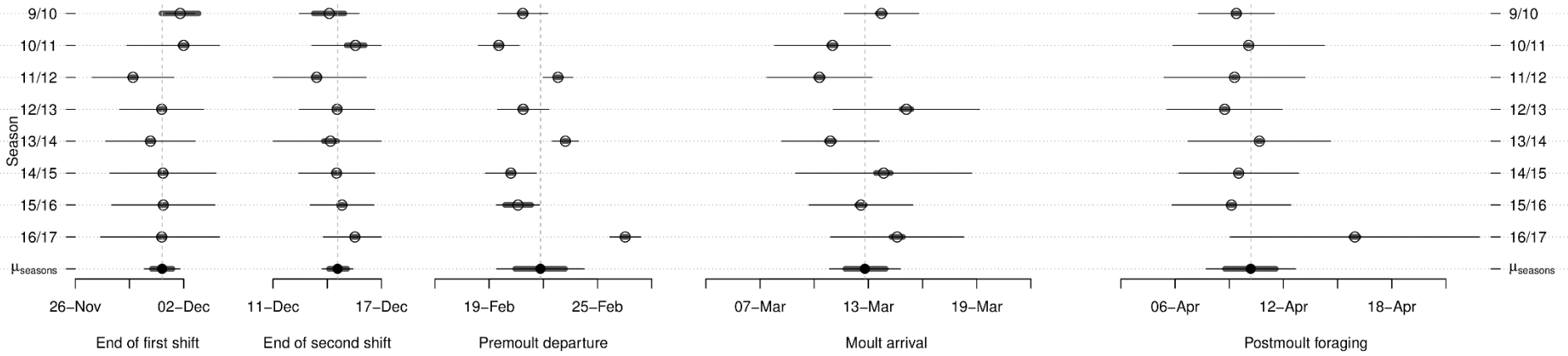
Phenological curves allow for missing data, overlap of events

Linden et al. 2017, J Avian Biol
Boersch-Supan et al. in prep

Inter-annual variation in breeding/moult timing



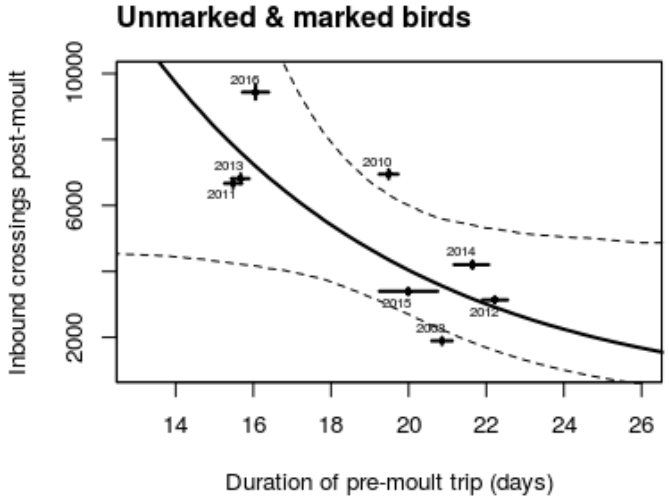


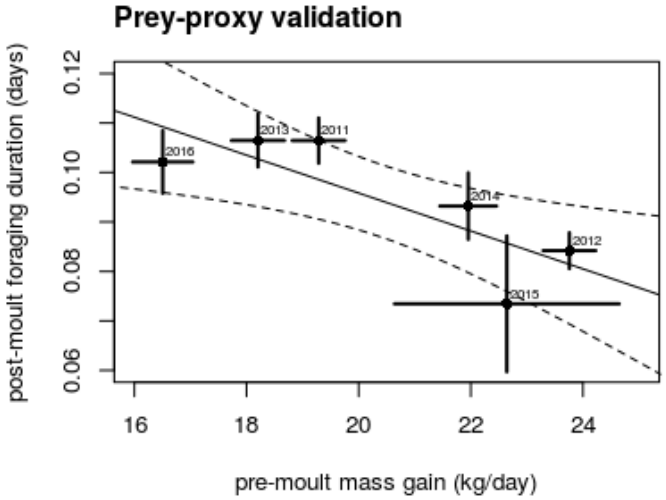


Timing of breeding events less variable among years than timing of moult/post-moult events

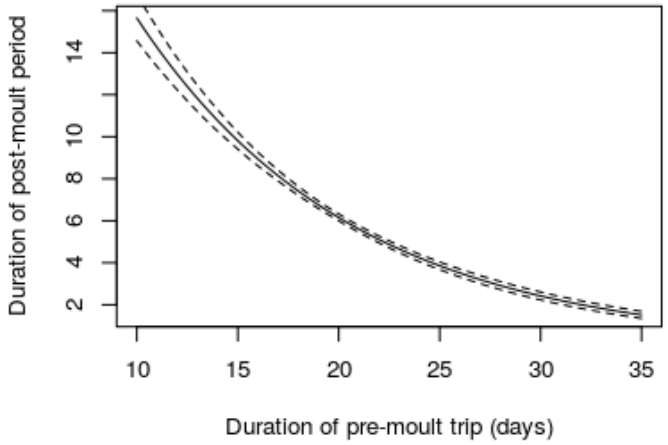
Boersch-Supan et al. in prep

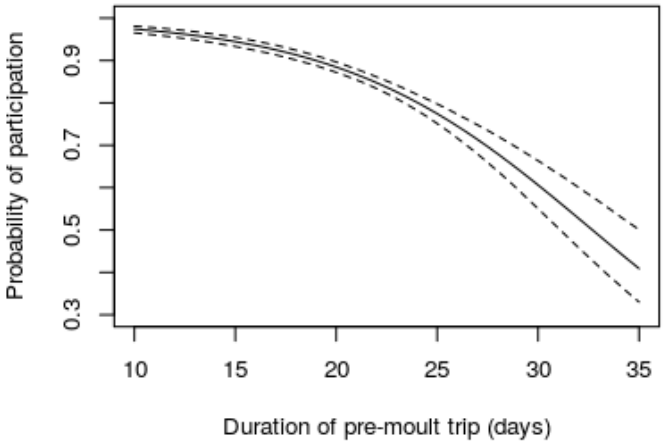
Linking post-moult movement to prey abundance





Inference approach:
errors-in-variables GLMs
Bayesian estimation
LOO-IC model selection

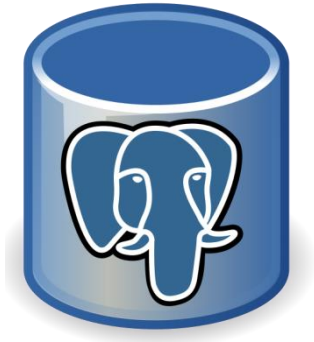




Neg. binomial hurdle model
Subject-level random effect
Bayesian estimation
LOO-IC model selection

- When pre-moult feeding is good, more birds delay migration for longer
- Krill abundance during brood-guard was not an informative predictor of post-moult activity
- Central place foraging may have energetic (thermoregulation) and/or social benefits

Workflow overview



Raw data (read-only psql tables)

"big data" but "small regression"
N = 6-8 years!

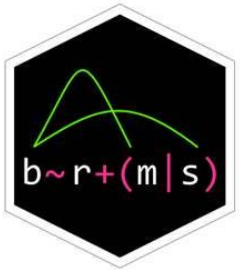
RPostgres

Statistical models

Classified crossings

Penguin Annotator

Reference dataset



Thank you for listening!

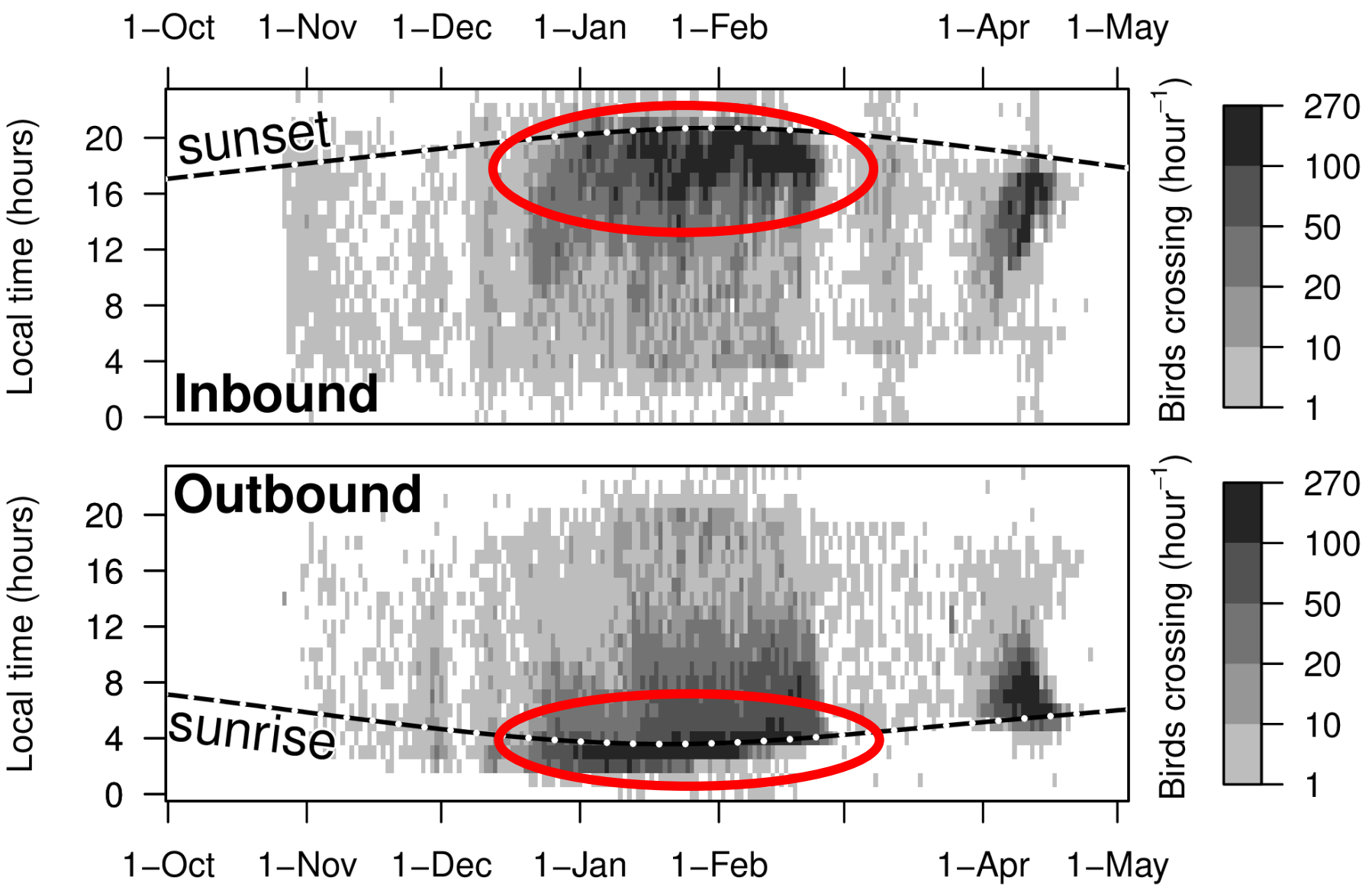
http://pboesu.github.io
pboesu@gmail.com
@pboesu

Thank you to

BAS engineers, data team & field technicians
rucrdtw reviewers & users: Maëlle Salmon, Noam Ross, Florian Pfisterer

BAS

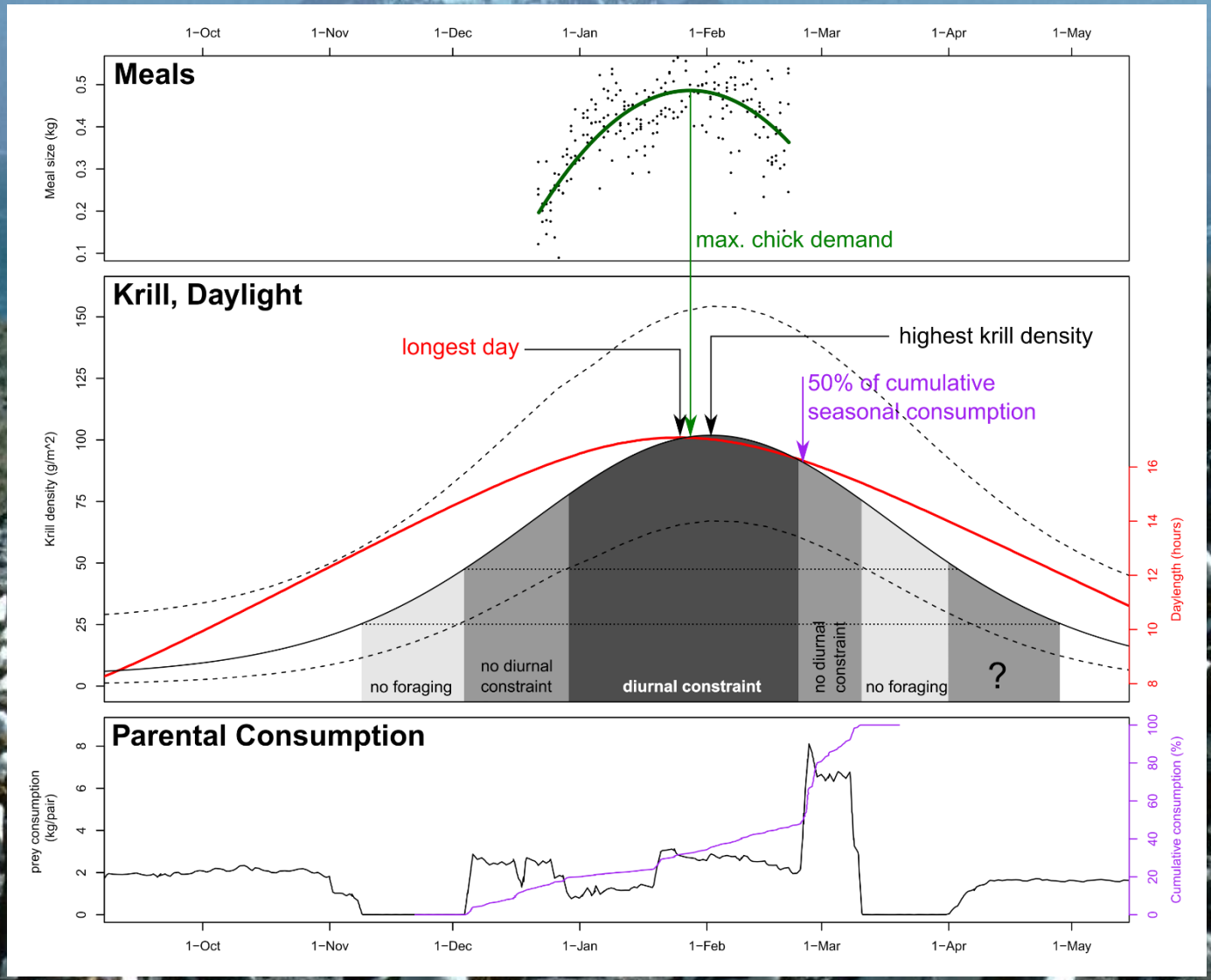
One of several results: rush hour!



Boersch-Supan et al. in prep

Penguins modulate post-breeding phenology, but not breeding phenology in response to local prey availability.

Bad news for breeding success if summer krill peak shifts permanently?



BAS

Software paper in JOSS

@JOSS_TheOJ

The Journal of Open Source Software

Submit Papers About Sign in

rucrdtw: Fast time series subsequence search in R

Philipp H Boersch-Supan

Article details

- View review »
- Download paper »
- Software repository »
- Software archive »

Submitted: 23 October 2016
Accepted: 07 November 2016

Cite as:

Boersch-Supan, (2016), rucrdtw: Fast time series subsequence search in R, Journal of Open Source Software, 1(7), 100, doi:10.21105/joss.00100

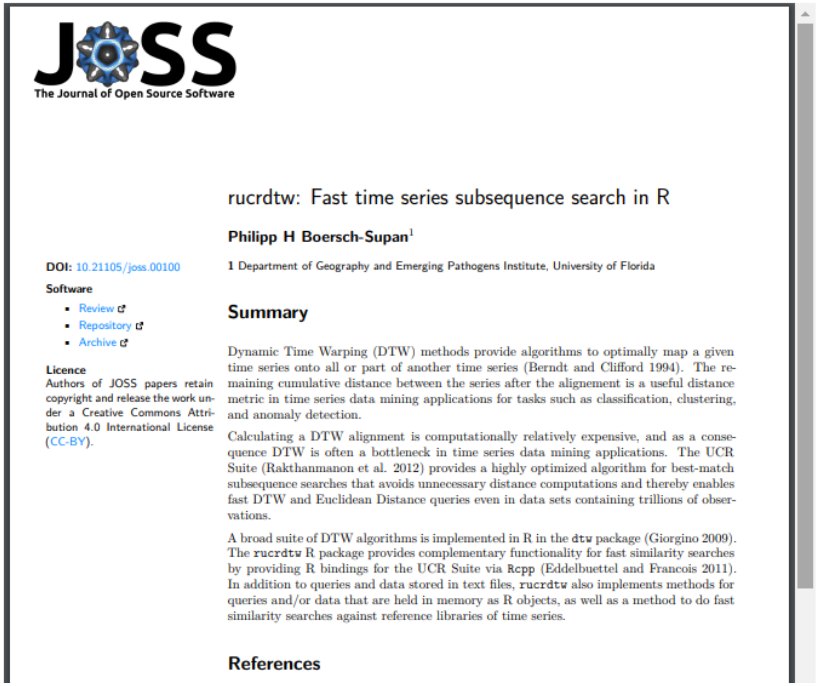
Status badge

JOSS 10.21105/joss.00100

License

Authors of JOSS papers retain copyright.

This work is licensed under a Creative Commons Attribution 4.0 International License.



JOSS
The Journal of Open Source Software

rucrdtw: Fast time series subsequence search in R

Philipp H Boersch-Supan[1]

1 Department of Geography and Emerging Pathogens Institute, University of Florida

DOI: 10.21105/joss.00100

Software
- Review
- Repository
- Archive

Licence
Authors of JOSS papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

Dynamic Time Warping (DTW) methods provide algorithms to optimally map a given time series onto all or part of another time series (Berndt and Clifford 1994). The remaining cumulative distance between the series after the alignement is a useful distance metric in time series data mining applications for tasks such as classification, clustering, and anomaly detection.

Calculating a DTW alignment is computationally relatively expensive, and as a consequence DTW is often a bottleneck in time series data mining applications. The UCR Suite (Rakthanmanon et al. 2012) provides a highly optimized algorithm for best-match subsequence searches that avoids unnecessary distance computations and thereby enables fast DTW and Euclidean Distance queries even in data sets containing trillions of observations.

A broad suite of DTW algorithms is implemented in R in the `dtw` package (Giorgino 2009). The `rucrdtw` R package provides complementary functionality for fast similarity searches by providing R bindings for the UCR Suite via `Rcpp` (Eddelbuettel and Francois 2011). In addition to queries and data stored in text files, `rucrdtw` also implements methods for queries and/or data that are held in memory as R objects, as well as a method to do fast similarity searches against reference libraries of time series.

References

- Great experience: Fast, constructive, friendly, transparent
- **Would recommend!**

@pboesu