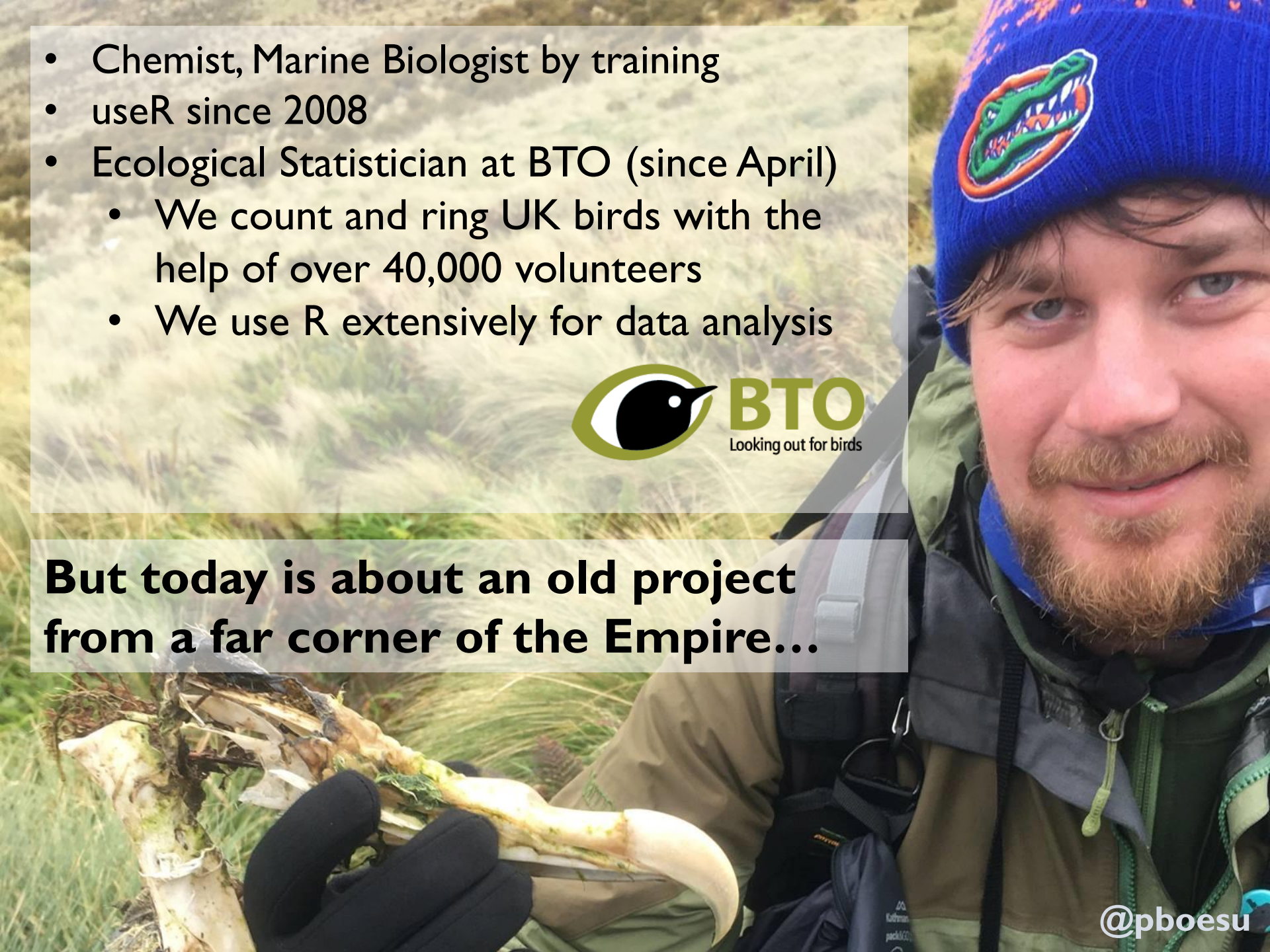
Counting and weighing penguins really fast with Rcpp

Philipp Boersch-Supan
British Trust for Ornithology
@pboesu

- Chemist, Marine Biologist by training
- useR since 2008
- Ecological Statistician at BTO (since April)
  - We count and ring UK birds with the help of over 40,000 volunteers
  - We use R extensively for data analysis

BTO
Looking out for birds

But today is about an old project from a far corner of the Empire…



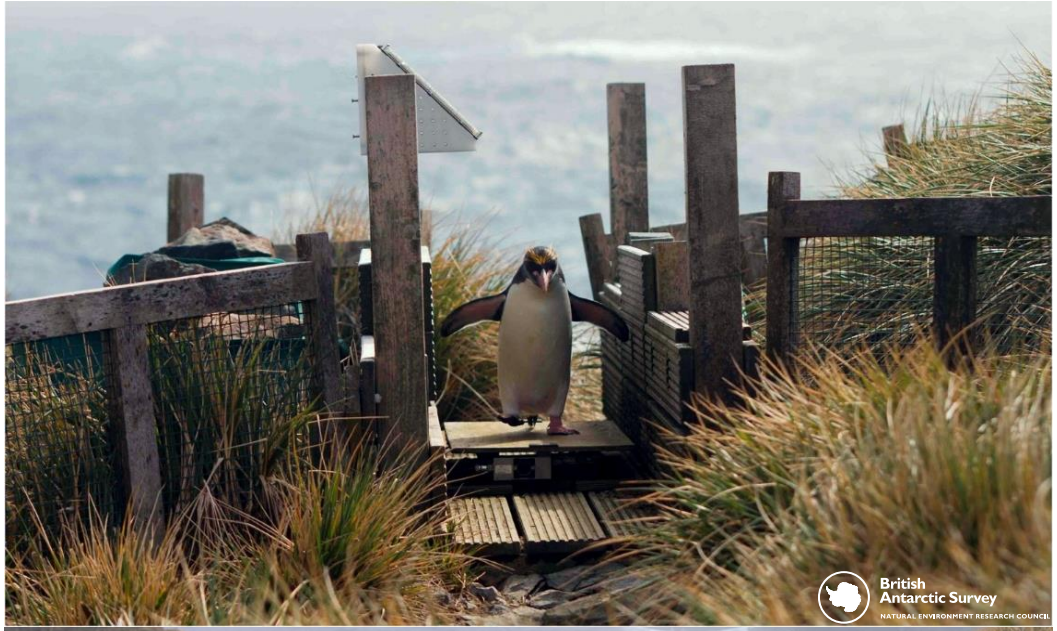
@pboesu

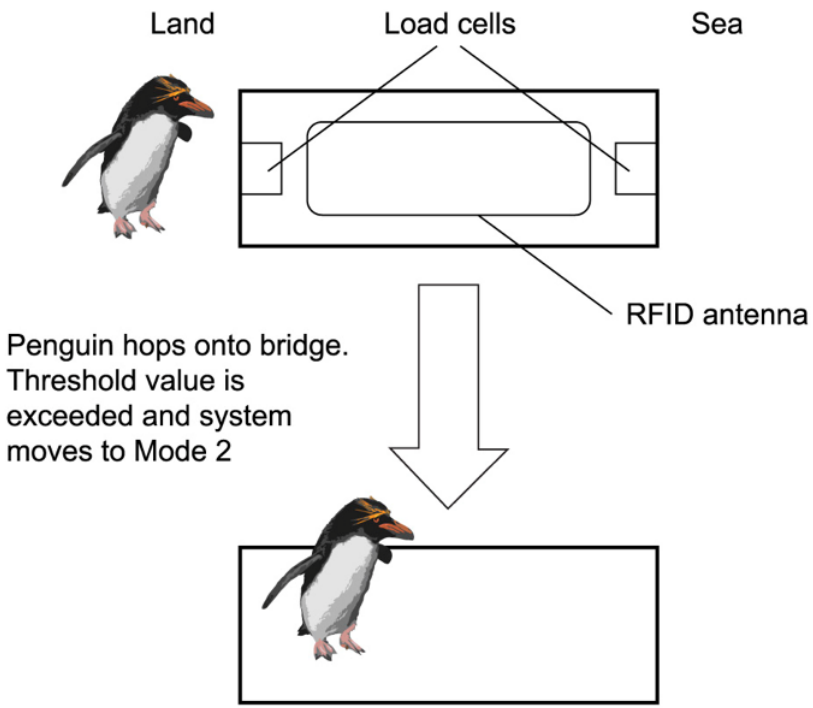
The **macaroni penguin** *Eudyptes chrysolophus*

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

@pboesu

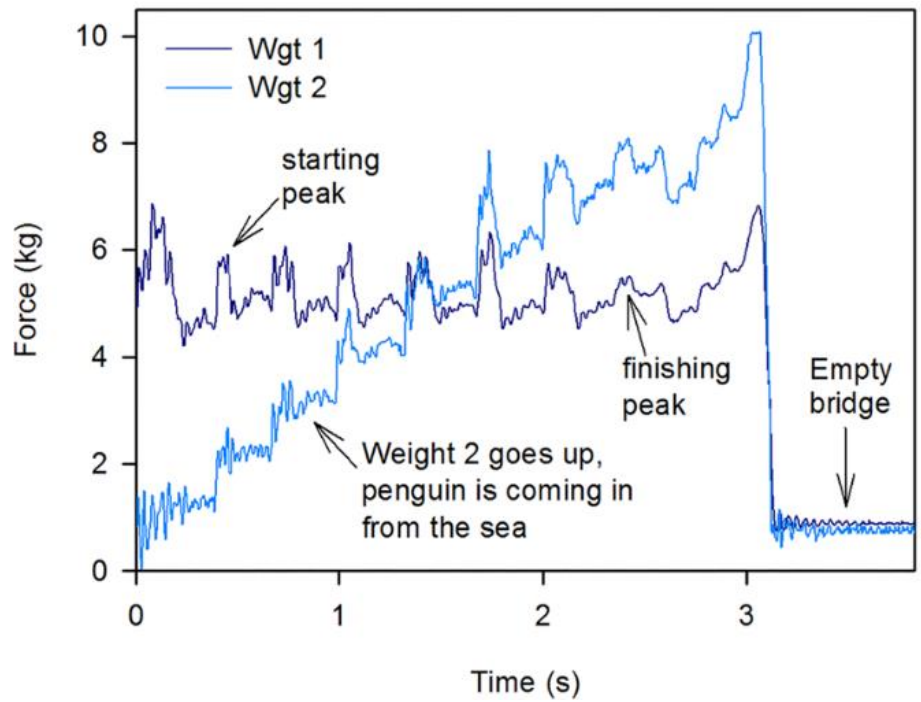
The penguin weighbridge of the British Antarctic Survey





@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

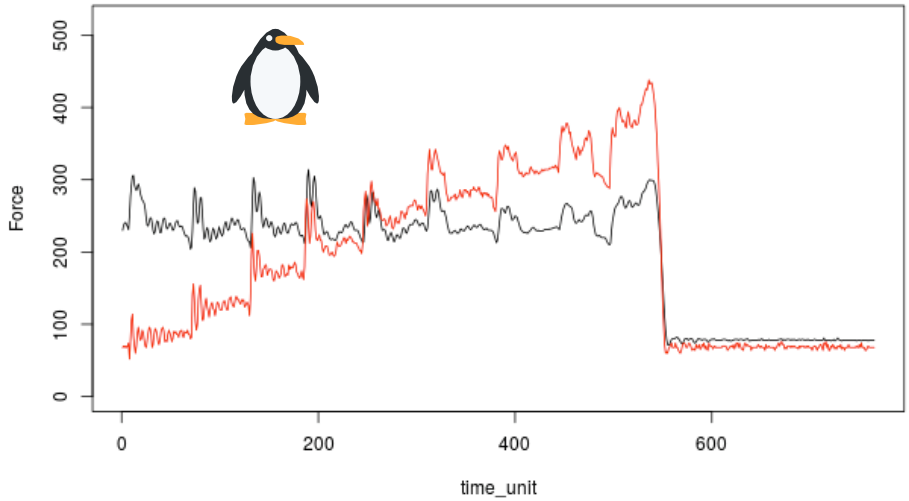
Challenges: accurate mass requires high-frequency sampling

50,000-80,000 crossings/season = **40 – 60 million raw data points/season**

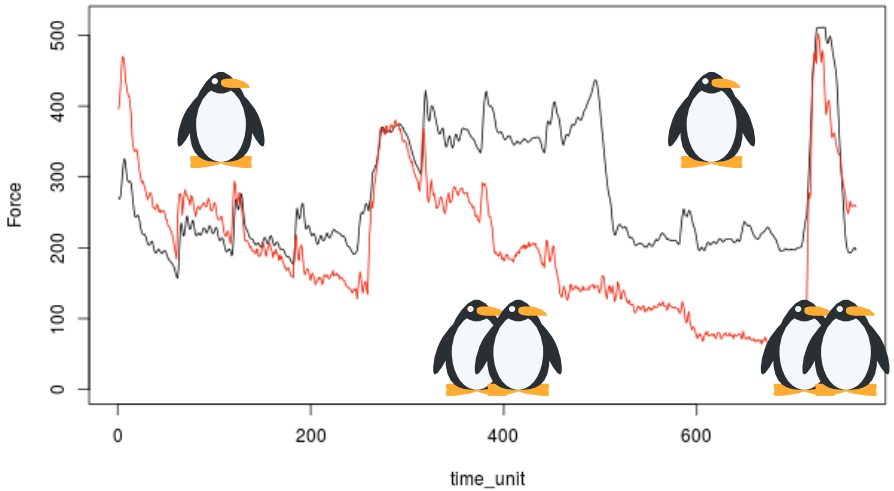
Afanasyev et al. 2015 PLoS ONE

and, penguins don't play by the rules…

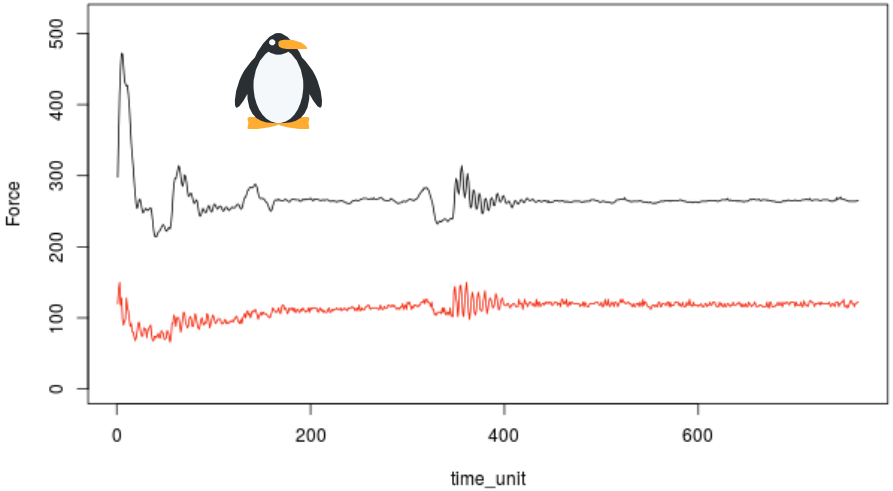
design case: 1 walking bird



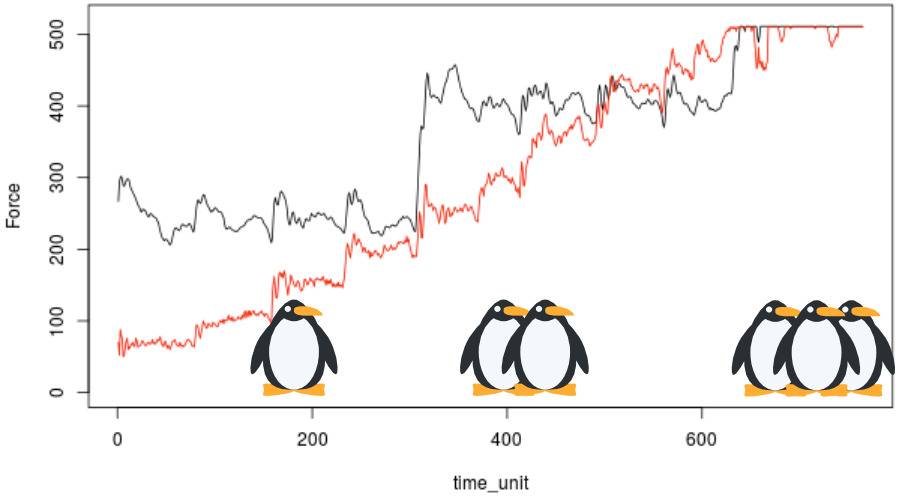
too busy: 3 birds in succession



not busy enough: 1 standing bird



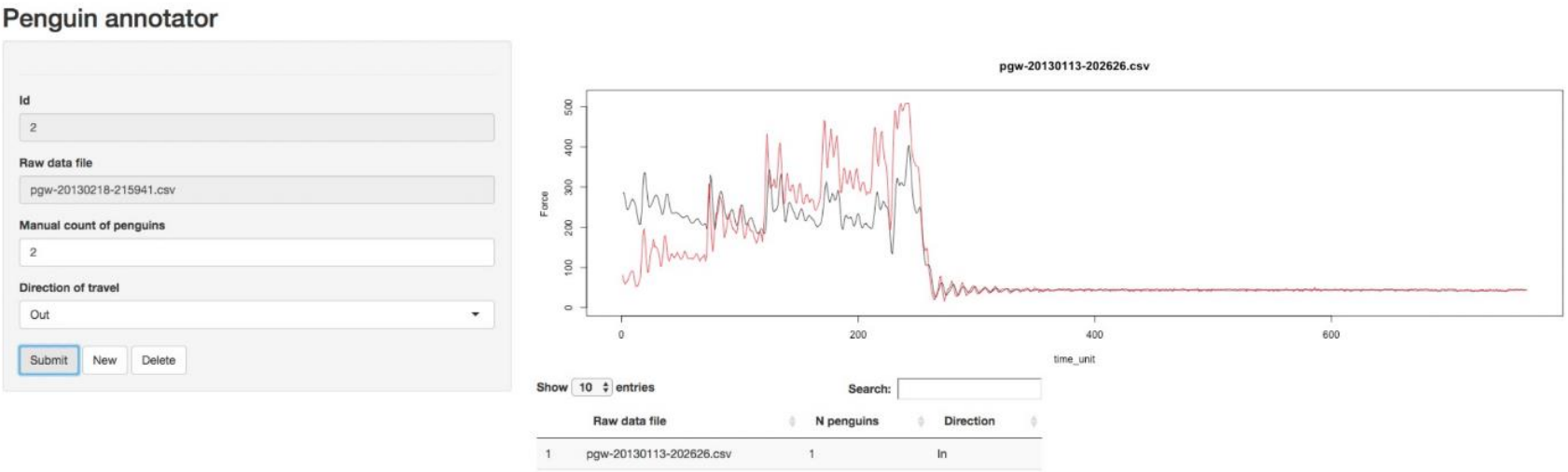
way too busy: 3 birds in succession



Easy-ish to separate for the human eye, less easy to automate

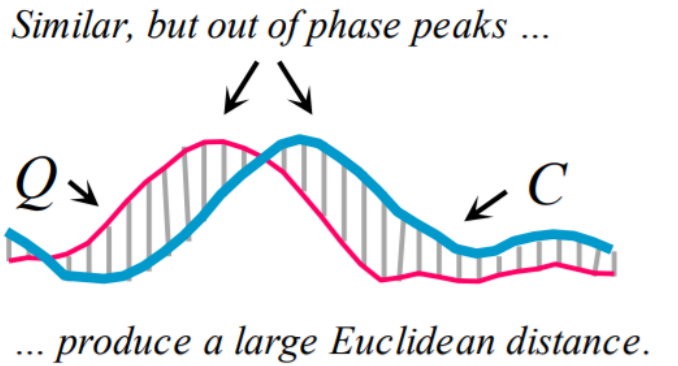
@pboesu

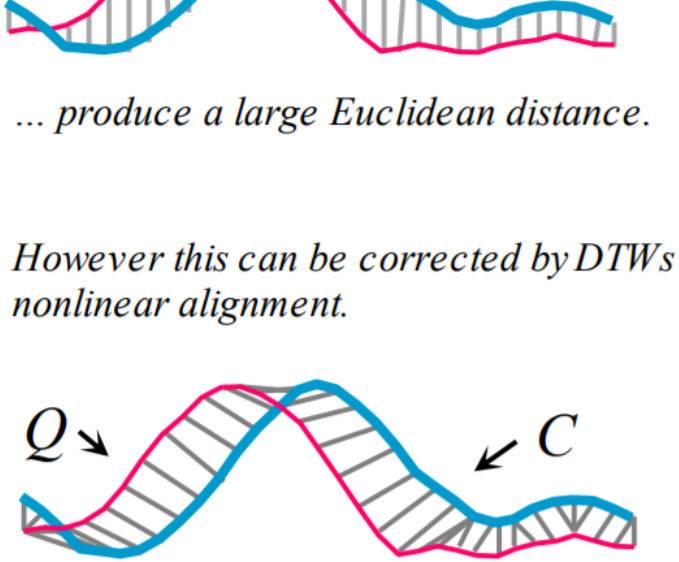
Step 1: The penguin annotator (shiny + RPostgres)

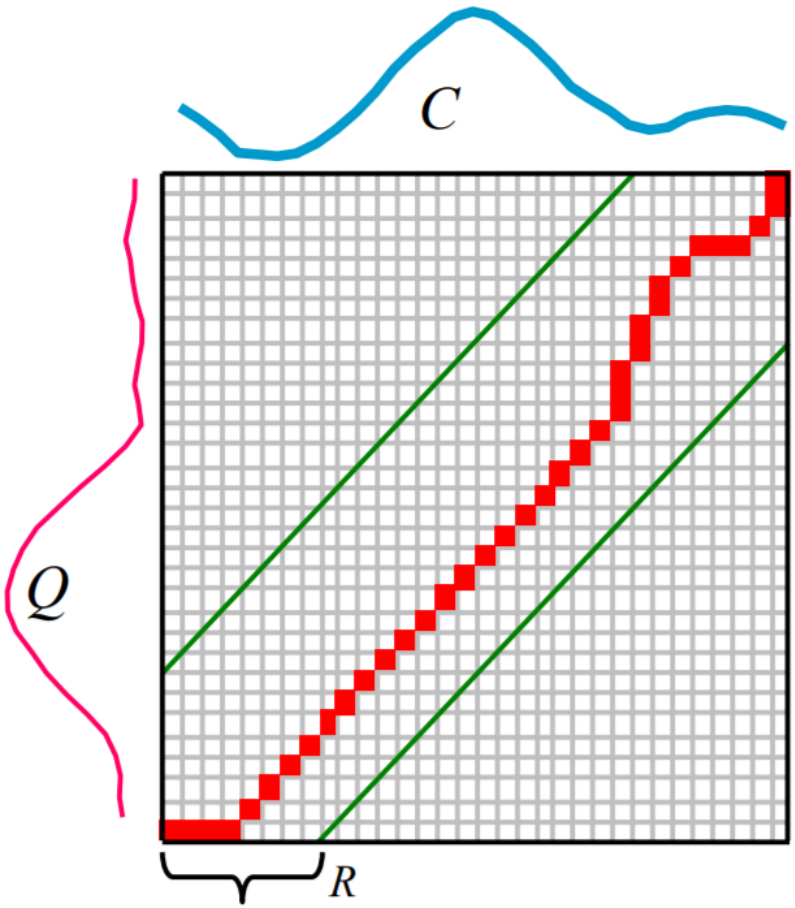


- Easy manual classification (but 50k+ files/yr prohibitively time intensive)
- **Need a classifier that has 'time-series shape recognition'**

Step 2: Dynamic Time Warping





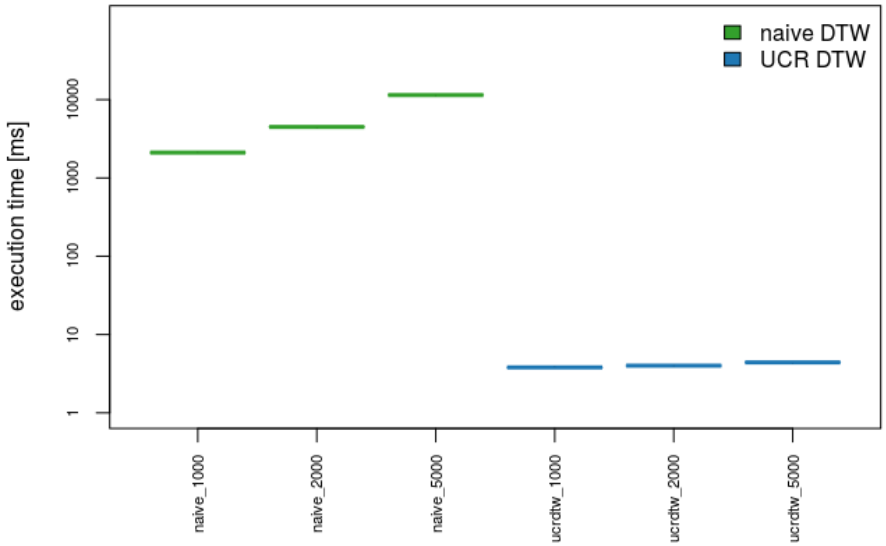


- implemented in R package dtw (using fast C routines), very good accuracy
- **but too costly (~ 1 minute) to compute full warping path for each crossing**

Rakthanmanon et al. 2012, Proc SIGKDD

Ultrafast Dynamic Time Warping to the rescue!

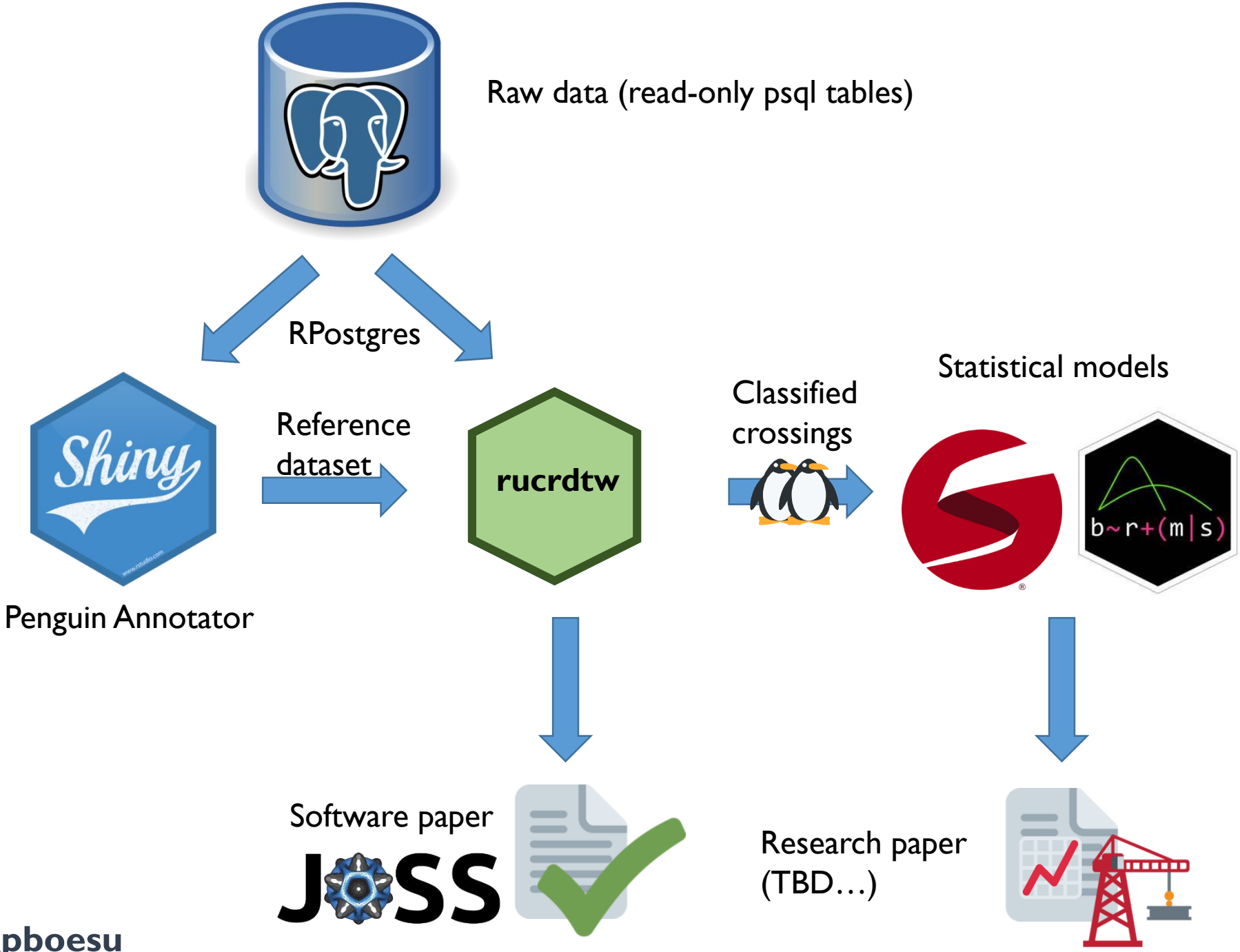
- 'best match' search, exploits early abandoning
- 2-3 orders of magnitude faster than naïve DTW
- Open C++ source available



- Thanks to Rcpp, draft R pkg **rucrdtw** working in a few hours
  - (Maëlle-approved version took a couple weeks more)

- Freely available on CRAN: `install.packages("rucrdtw")`
- **A single season of penguin crossings can now be classified in <1hr**

Rakthanmanon et al. 2012, Proc SIGKDD


Raw data (read-only psql tables)
RPostgres
Shiny
Reference dataset
rucrdtw
Classified crossings
Statistical models
b~r+(m|s)
Penguin Annotator
Software paper
JOSS
Research paper (TBD…)

@pboesu

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

Thank you for listening!

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

Thank you to

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

BAS