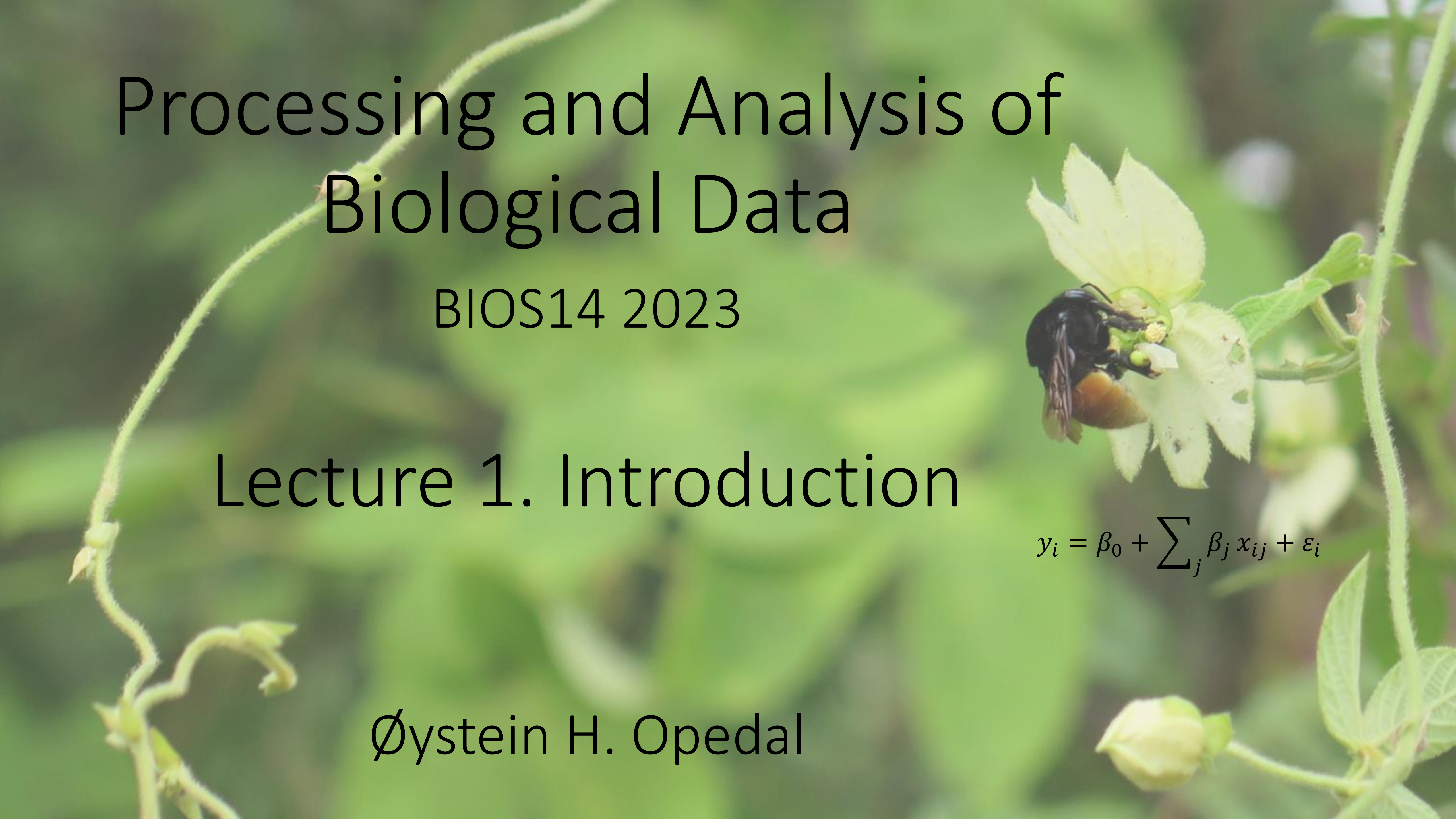


Processing and Analysis of Biological Data

BIOS14 2023

Lecture 1. Introduction

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i$$

Øystein H. Opedal

Introduction

- What is this course?
  - Analysis courses are different from other courses, and this analysis course is different from other analysis courses
- Who are we?
- Who are you? Participant backgrounds in studies, research, programming. Do you already have your own data?

Outline for first two weeks

- Today
  - General introduction to quantitative analysis
  - Summary statistics and programming (exercise)
  - Paper discussion: Houle et al. 2011
  - Fraud, Reproducibility, and GitHub
- Monday
  - The linear model: lecture and exercises
  - Paper discussion: Wasserstein & Lazar 2016 + Berner & Amrhein 2022
- Wednesday
  - Analysis of Variance (ANOVA) – First report with feedback

What is this course

- Essentials of quantitative analysis, including scientific programming, statistical modelling, quantitative interpretation, and presentation
- This course will develop as we go, depending on our progress and needs. Thus, the schedule is provisional in terms of topics.
- Which kinds of data handling/analyses do YOU need in your work?
- New from last year:
  - One extra teacher per session
  - 2 practise reports with feedback + midterm
  - Some exercises optional

Literature

- Two books – for reference
- Lecture notes
- Papers

Course representative – giving feedback during the course

- Elected by the students on the course during timetabled hours
- One to two students on the course tasked with acting as a channel of communication between students and teaching staff
- Raise problems that arise
- Check-in meetings during the course and at the end of the course
- Reminding students about course evaluations and reviewing them
- Everyone is to be informed as to who is the course representative
- Support from the Lund Science Students' Union (LUNA)

Measurement and meaning in biology

- Measurements are the link between the biology we want to study and our quantitative analyses
- Choice of measurements should, in principle, be based on the theoretical context of the work

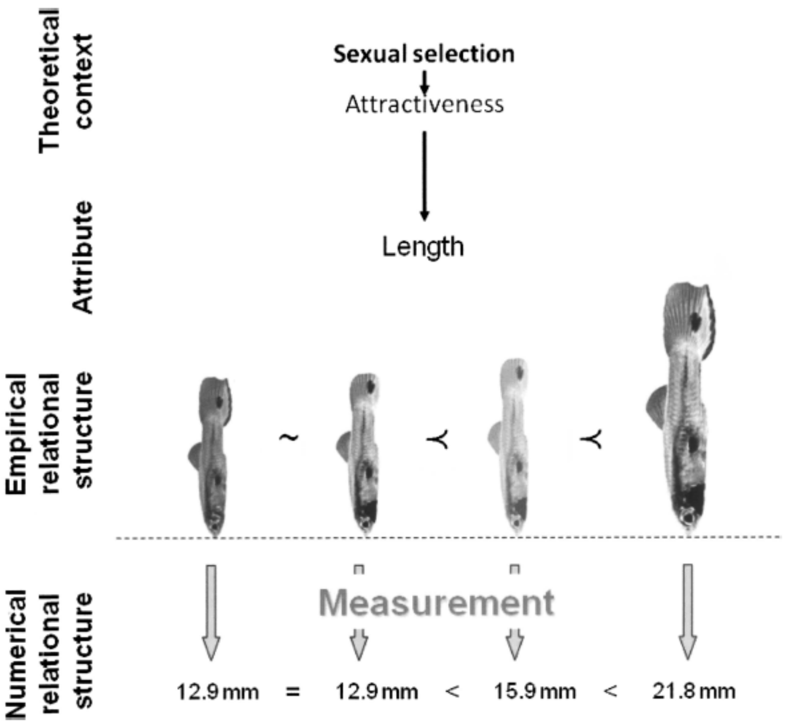

Theoretical context
Sexual selection
Attractiveness
Attribute
Length
Empirical relational structure
~
≺
≺
Numerical relational structure
Measurement
12.9 mm = 12.9 mm < 15.9 mm < 21.8 mm

Figure 1. The Measurement Process

We imagine a study where the theoretical context is sexual selection. Within this context, we focus on attractiveness of males and then on the hypothesis that size influences attractiveness in the guppy. Size can be measured in many ways, so the concept of "size" was referred to the attribute "overall length of the fish" under the hypothesis that females prefer large males and treat the tail as part of the body. The middle third of the figure represents the empirical relations that are at the root of measurement. For empirical comparison of length, fish could be aligned with their noses against a flat surface, and the identity of the fish that extends farther noted. For a pair of fish *A* and *B*, the possible results are that *A* extends farther than *B*, which we can represent as $A \succ B$; that we cannot decide whether *A* extends farther than *B*, $A \sim B$; and that *A* extends less far than *B*, $A \prec B$. Representational measurement theory proves that the conclusions that can be drawn about length on the basis of the pairwise empirical comparisons can also be drawn from a numerical system consisting of numbers (*a* and *b*) assigned to lengths of A and B plus a mapping of the empirical relations $\succ$, $\sim$, $\prec$ among fish to the relations $>$, $=$, and $<$ among the numbers.

Scale types

- How many of you have heard about scale types before?
- Can you give an example of a scale type?

Scale types

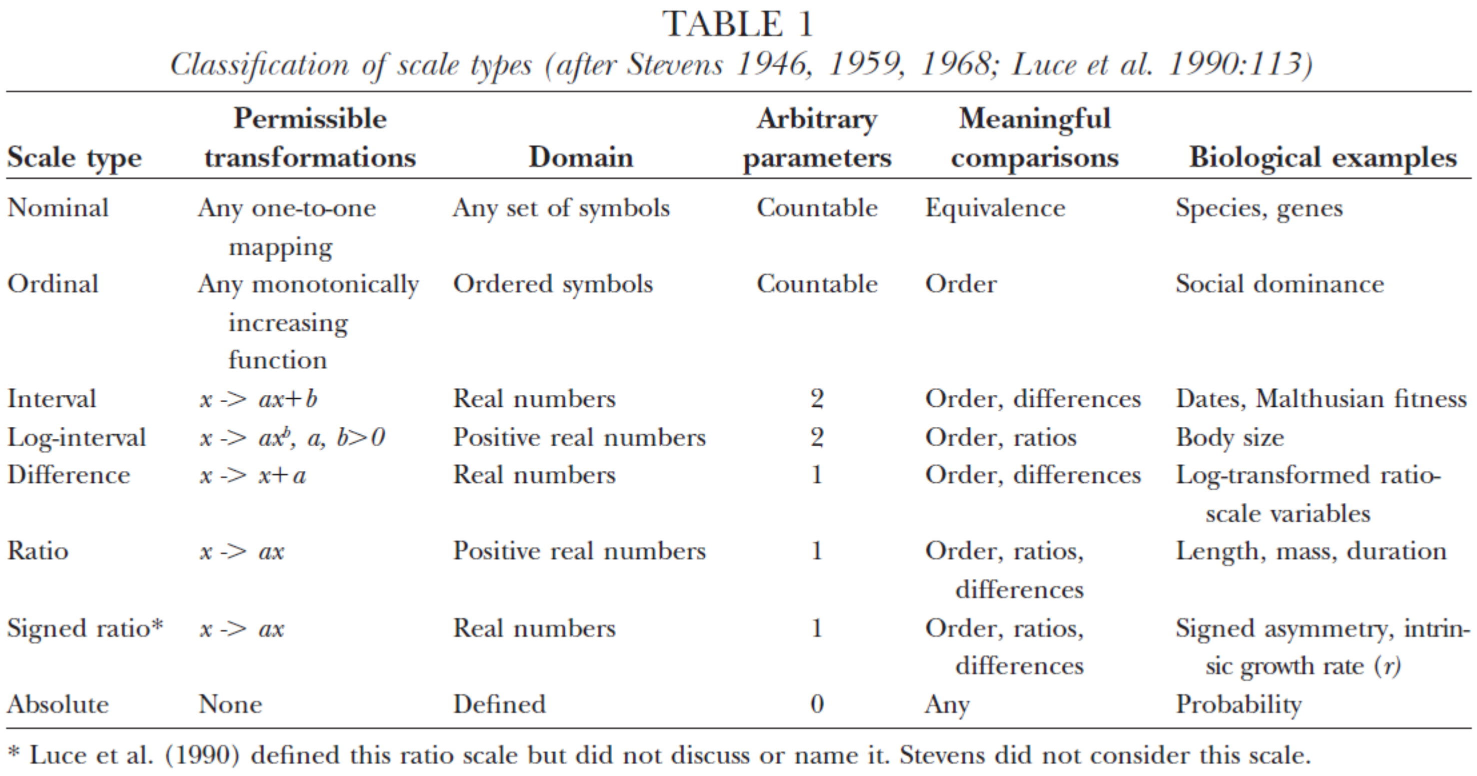


TABLE 1

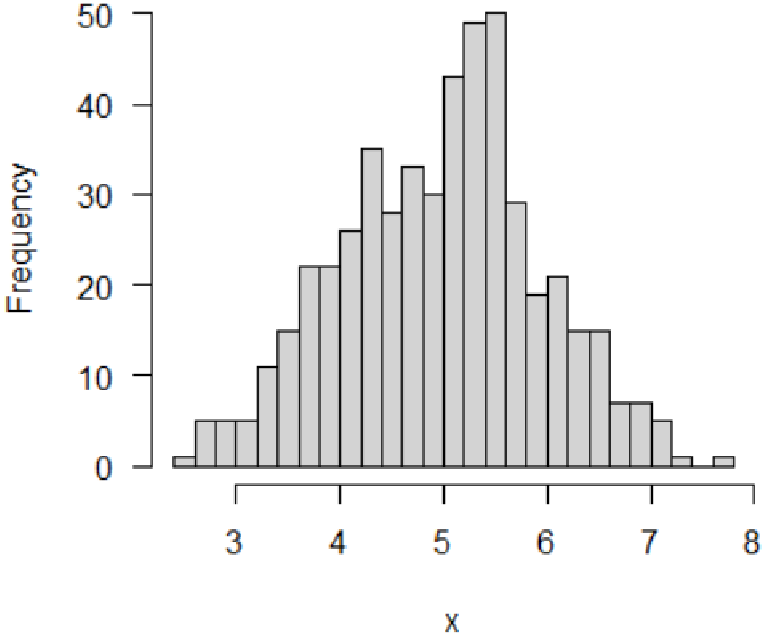
Classification of scale types (after Stevens 1946, 1959, 1968; Luce et al. 1990:113)

Scale type	Permissible transformations	Domain	Arbitrary parameters	Meaningful comparisons	Biological examples
Nominal	Any one-to-one mapping	Any set of symbols	Countable	Equivalence	Species, genes
Ordinal	Any monotonically increasing function	Ordered symbols	Countable	Order	Social dominance
Interval	$x \to ax+b$	Real numbers	2	Order, differences	Dates, Malthusian fitness
Log-interval	$x \to ax^b,\ a,\ b>0$	Positive real numbers	2	Order, ratios	Body size
Difference	$x \to x+a$	Real numbers	1	Order, differences	Log-transformed ratio-scale variables
Ratio	$x \to ax$	Positive real numbers	1	Order, ratios, differences	Length, mass, duration
Signed ratio*	$x \to ax$	Real numbers	1	Order, ratios, differences	Signed asymmetry, intrinsic growth rate (r)
Absolute	None	Defined	0	Any	Probability

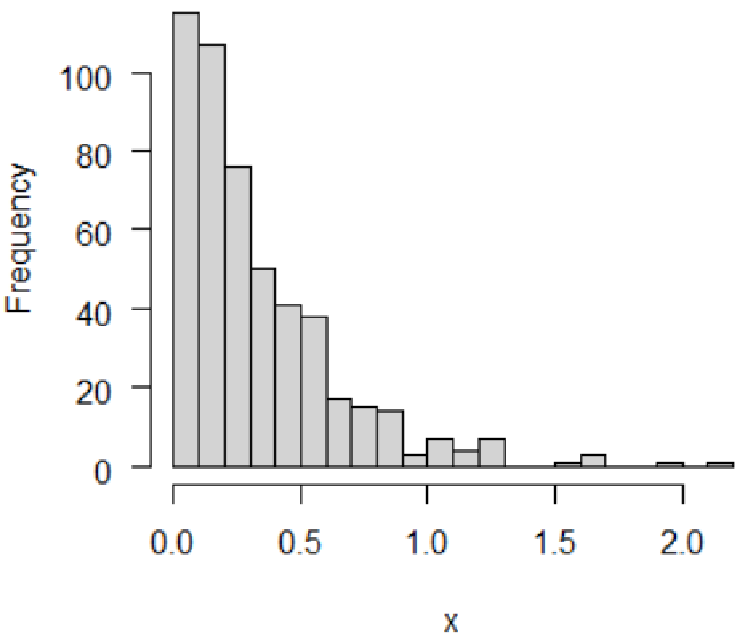
* Luce et al. (1990) defined this ratio scale but did not discuss or name it. Stevens did not consider this scale.

The importance of summary statistics

- Which is higher, the mean or the median?



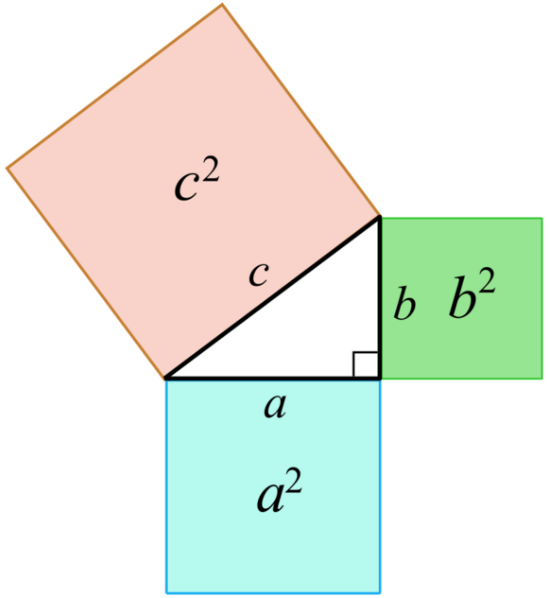
Mean ≈ Median



Mean > Median

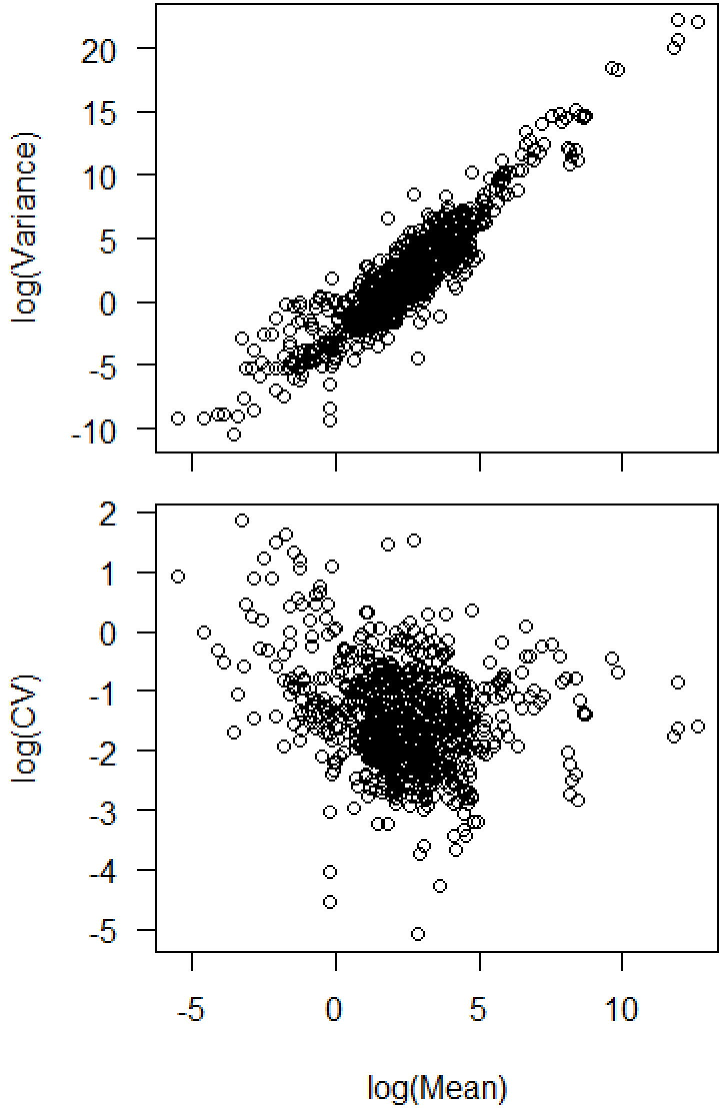
Quantifying data dispersion

- Standard deviations and variances describe the dispersion of data.
- Variances are additive, standard deviations are not (Pythagoras!)
- The variance and standard deviation are measures of dispersion in the data, **not** the certainty of an estimate. Therefore, unlike the standard error, the standard deviation should not be given with the ± sign.

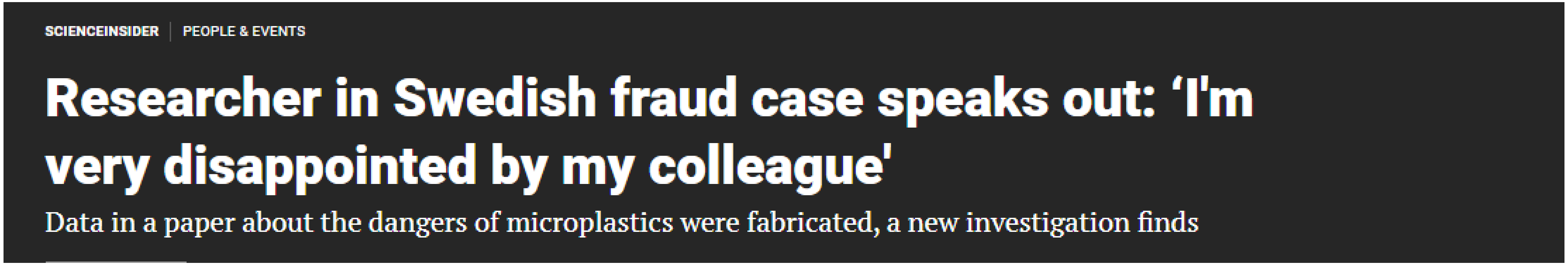


Larger structures are more variable than small structures

- Comparing small to large structures (e.g. mice to elephants) requires expressing variance on a common scale
- The coefficient of variation [SD(x)/E(x)] gives the standard deviation as a proportion of the mean, and removes the mean-variance relationship
- (Real data on ~1000 plant traits)



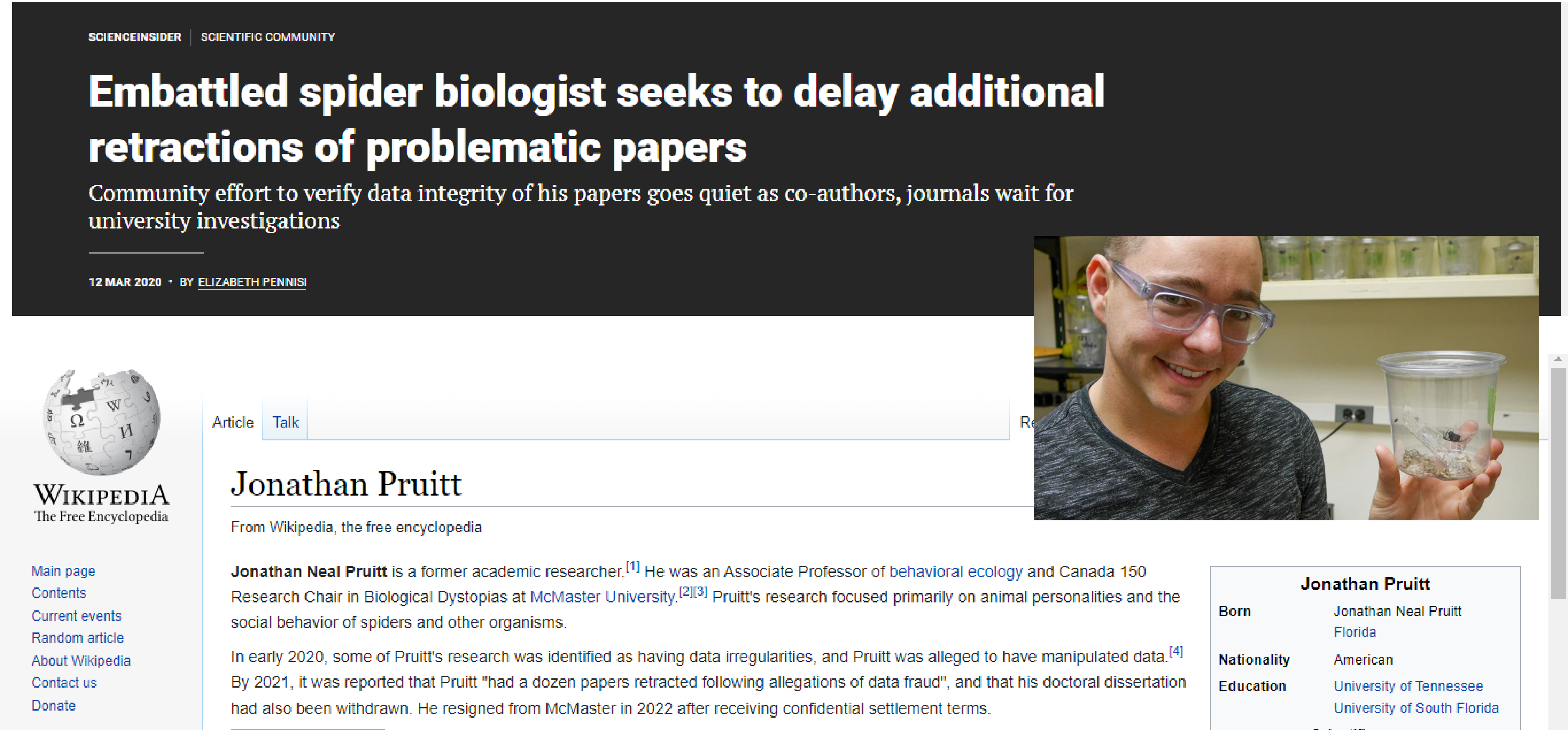
Fraud, Reproducibility, and GitHub



SCIENCEINSIDER | PEOPLE & EVENTS

Researcher in Swedish fraud case speaks out: ‘I'm very disappointed by my colleague'

Data in a paper about the dangers of microplastics were fabricated, a new investigation finds



SCIENCEINSIDER | SCIENTIFIC COMMUNITY

Embattled spider biologist seeks to delay additional retractions of problematic papers

Community effort to verify data integrity of his papers goes quiet as co-authors, journals wait for university investigations

12 MAR 2020 • BY ELIZABETH PENNISI

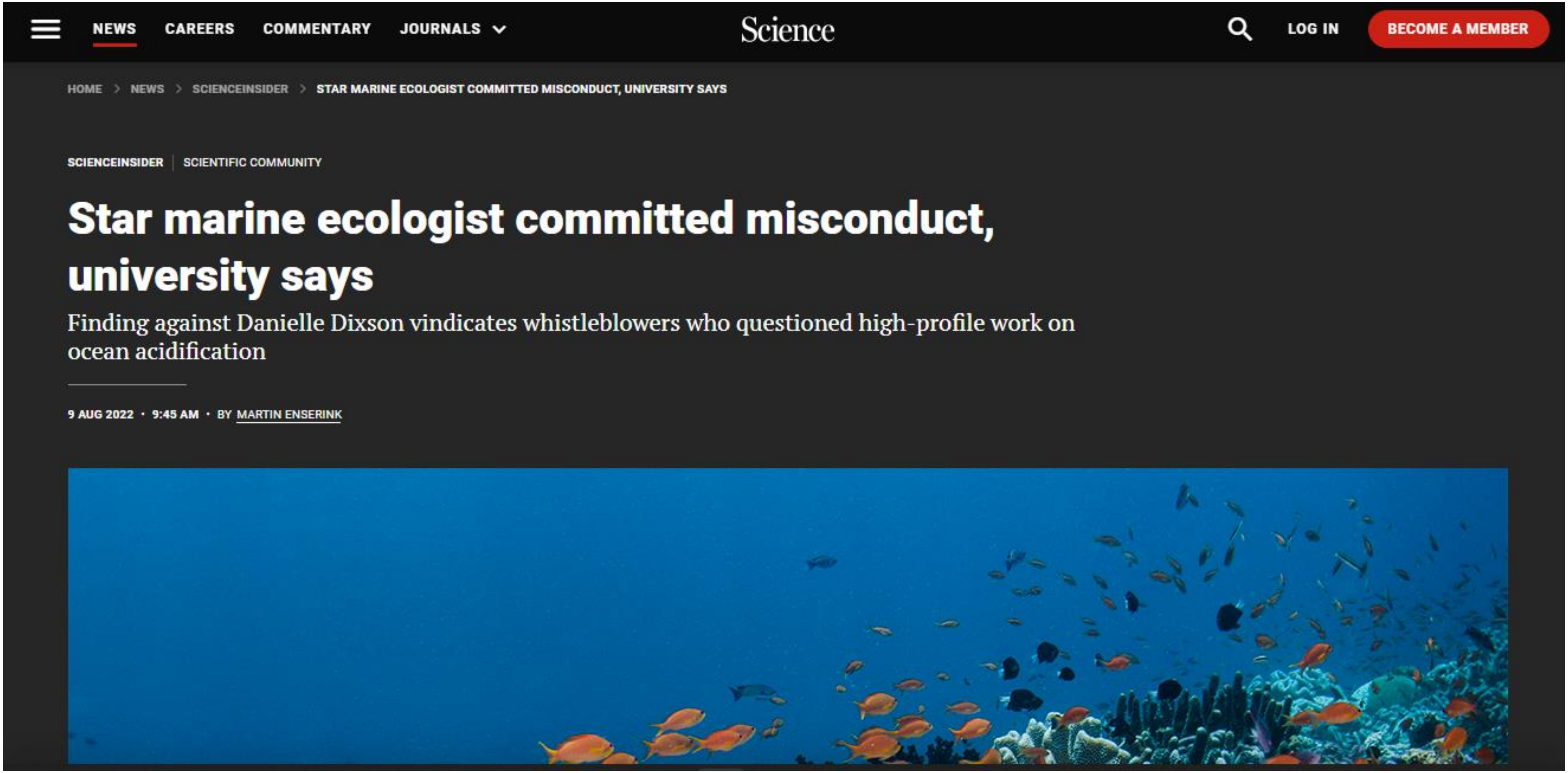
Jonathan Pruitt

From Wikipedia, the free encyclopedia

Jonathan Neal Pruitt is a former academic researcher.[1] He was an Associate Professor of behavioral ecology and Canada 150 Research Chair in Biological Dystopias at McMaster University.[2][3] Pruitt's research focused primarily on animal personalities and the social behavior of spiders and other organisms.

In early 2020, some of Pruitt's research was identified as having data irregularities, and Pruitt was alleged to have manipulated data.[4] By 2021, it was reported that Pruitt "had a dozen papers retracted following allegations of data fraud", and that his doctoral dissertation had also been withdrawn. He resigned from McMaster in 2022 after receiving confidential settlement terms.

Jonathan Pruitt	
Born	Jonathan Neal Pruitt Florida
Nationality	American
Education	University of Tennessee University of South Florida



HOME > NEWS > SCIENCEINSIDER > STAR MARINE ECOLOGIST COMMITTED MISCONDUCT, UNIVERSITY SAYS

SCIENCEINSIDER | SCIENTIFIC COMMUNITY

Star marine ecologist committed misconduct, university says

Finding against Danielle Dixson vindicates whistleblowers who questioned high-profile work on ocean acidification

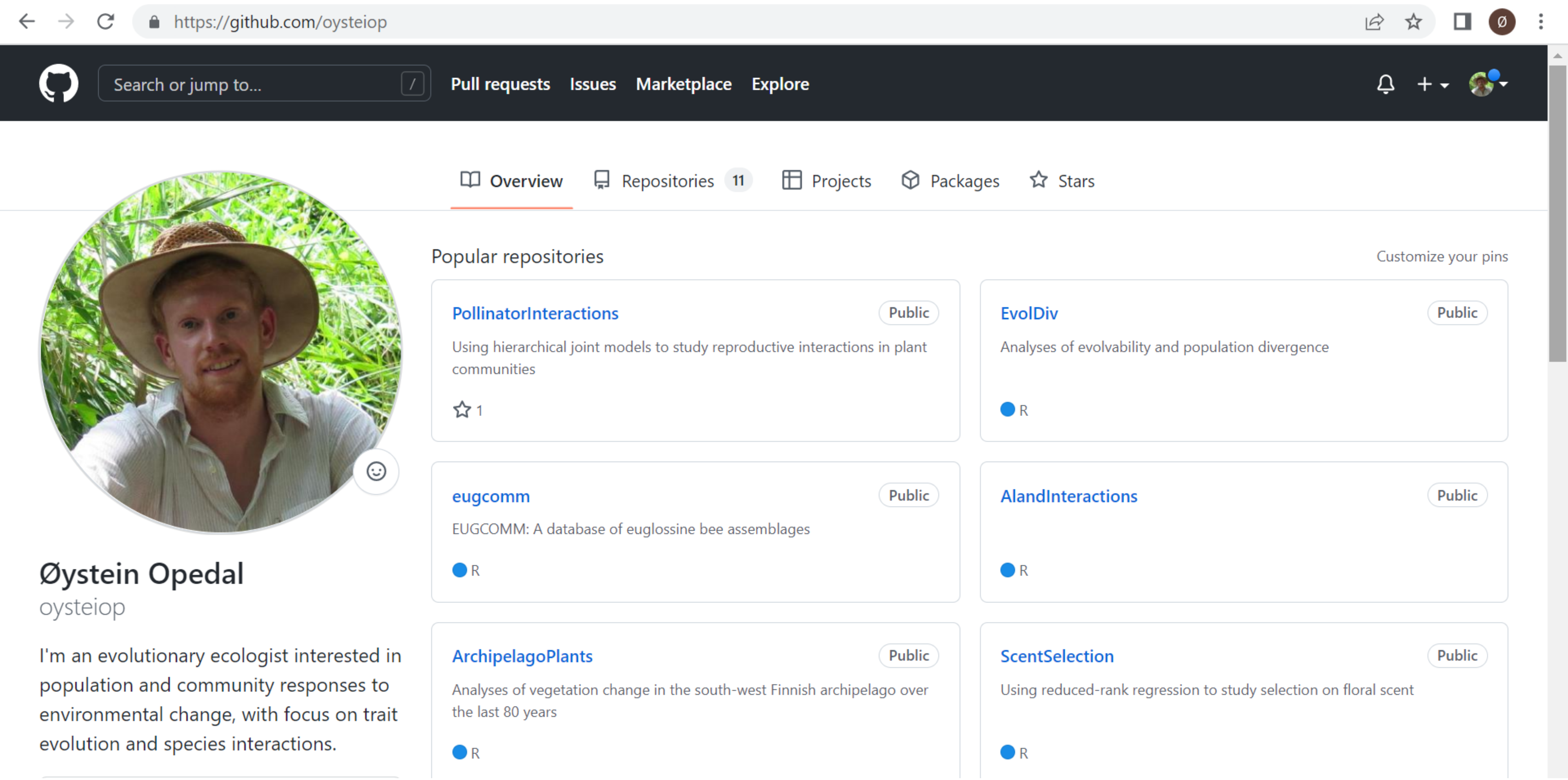
9 AUG 2022 • 9:45 AM • BY MARTIN ENSERINK

Scientific publishing in the age of fraud

- These incidents have led to new requirements for scientific publications
- Most journals now require publication of data
- More and more require publication of analysis code, some even have ‘code editors’!
- To publish your master/PhD chapters, you will have to produce reproducible code, normally in R.
- This course will teach you the necessary tools to do so

Reproducible analysis with GitHub

- GitHub (and Git as such) is a system for version control of e.g. analysis code, but also a way of backing up our analyses, and to publish the code (and data) when the paper is published
- Easily integrated with R-Studio, no need to mess around with command line (unless you want to).



Øystein Opedal

oysteiop

I'm an evolutionary ecologist interested in population and community responses to environmental change, with focus on trait evolution and species interactions.

Popular repositories

PollinatorInteractions

Using hierarchical joint models to study reproductive interactions in plant communities

☆ 1

EvolDiv

Analyses of evolvability and population divergence

● R

eugcomm

EUGCOMM: A database of euglossine bee assemblages

● R

AlandInteractions

● R

ArchipelagoPlants

Analyses of vegetation change in the south-west Finnish archipelago over the last 80 years

● R

ScentSelection

Using reduced-rank regression to study selection on floral scent

● R

Getting started with R (and GitHub)

- Install R, R Studio and Git
- Create a GitHub account
- Create a repository for this course
- Clone your repository into R
- Start exercises