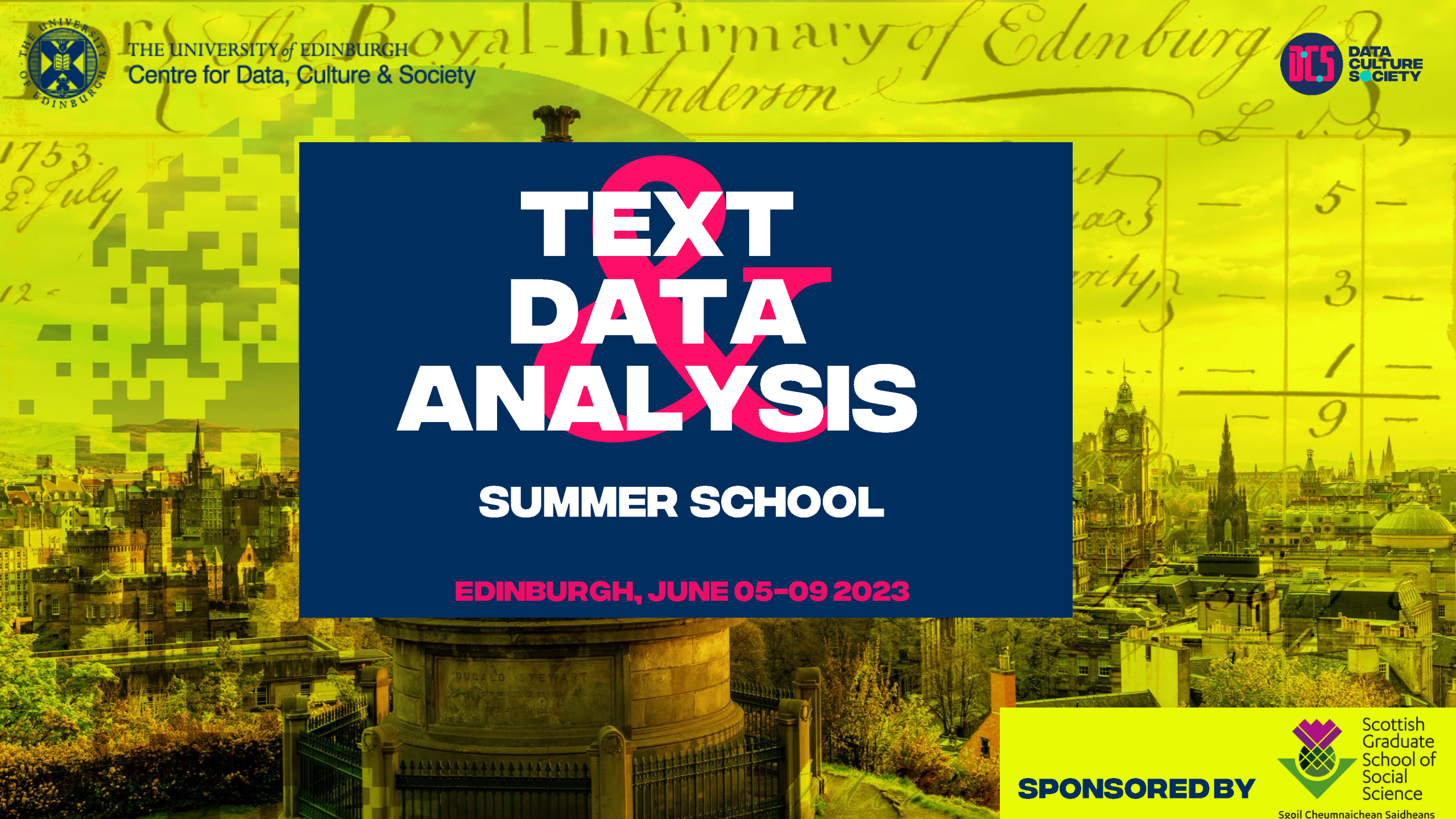
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

TEXT & DATA ANALYSIS

SUMMER SCHOOL

EDINBURGH, JUNE 05–09 2023



SPONSORED BY

Scottish Graduate School of Social Science

Sgoil Cheumnaichean Saidheans

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

TODAY'S SCHEDULE

Seminar: Exploring "Fishy" Data - an Adventure in Multivariate Statistics and the Cult of J.W. Tukey

Hands-on session 1: Statistics 1

Hands-on session 2: Statistics 2

Keynote Lecture: The Boundaries of Digitised Content: Designing Research Projects within Collection Constraints

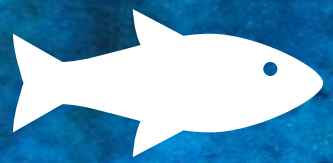
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

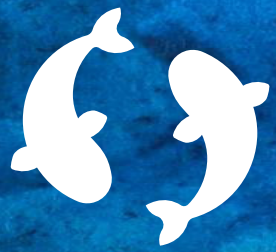
EXPLORING "FISHY" DATA - AN ADVENTURE IN MULTIVARIATE STATISTICS AND THE CULT OF J.W. TUKEY

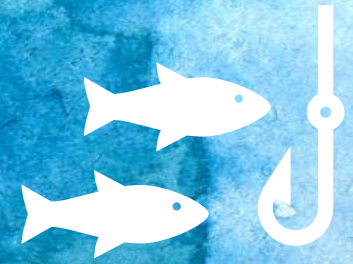
Dr Sam Leggett,

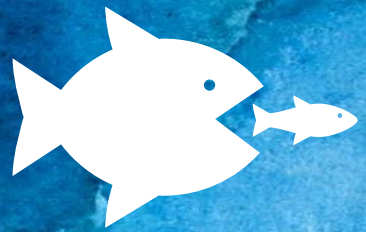
Lecturer in Archaeology at the University of Edinburgh

Exploring "Fishy" Data - an Adventure in Multivariate Statistics and the Cult of J.W. Tukey









Dr Sam Leggett

@samleggs22 sam.leggett@ed.ac.uk

School of History, Classics and Archaeology

University of Edinburgh

THE UNIVERSITY of EDINBURGH

LEVERHULME TRUST

TW: Some slides in the presentation contain images of human remains

My Stats Rant

- There's a lot of bad practice in terms of stats in archaeology
- Reliance on outliers, but usually only visually, without doing any formalized outlier analysis
- Lots of uncertainty and error propagation in isotopic analyses (my sub-discipline)
- Decided to completely avoid p-values
- Big fan of the "Johns" - Tukey and Kruschke - so used a Tukey-eque Exploratory Data Analysis and "New Statistics" approach

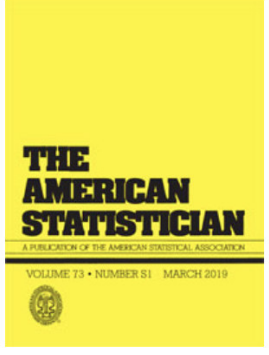
'..."statistically significant"—don't say it and don't use it.'

Wasserstein, Schirm, and Lazar (2019)
"Moving to a World Beyond 'p < 0.05.'"

The American Statistician 73:sup1

The "New" Statistics and ASA's ban on statistical significance

'…"statistically significant"–don't say it and don't use it.'



The American Statistician, Volume 73, Issue sup1 (2019)

Volume 73, 2019 | Vol 72, 2018 | Vol 71, 2017 | Vol 70, 2 | See all volumes and issues

sue 4 | Issue 3 | Issue 2 | Supplement 1 | Issue 1

Browse this journal

- Latest articles
- Current issue
- List of issues
- Special issues
- Open access articles
- Most read articles
- Most cited articles

Download citations | Download PDFs | Browse by section (All) | Display order (Default)

Statistical Inference in the 21st Century: A World Beyond p < 0.05

Editorial

Editorial

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

Pages: 1-19

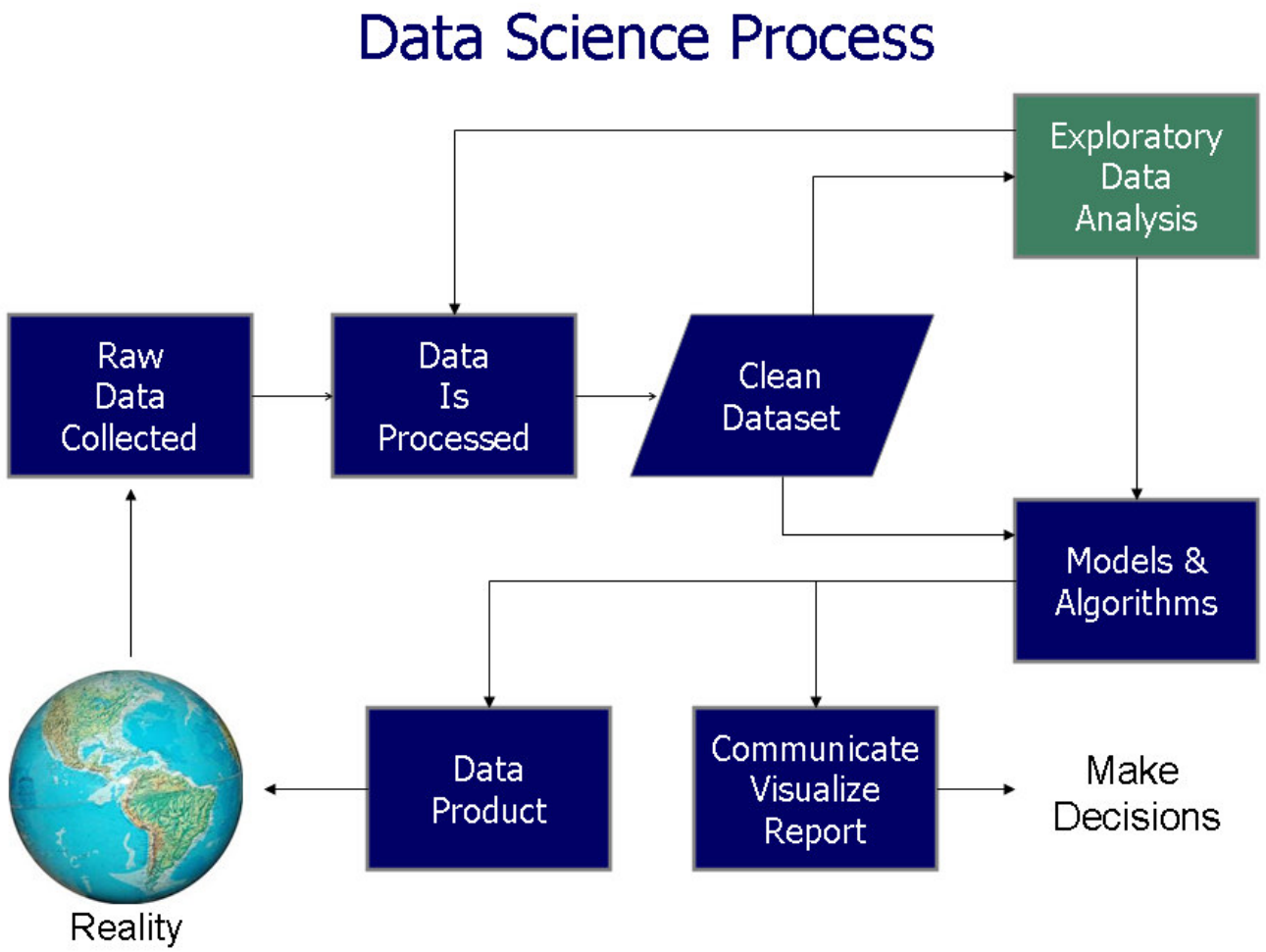
Published online: 20 Mar 2019

336534 Views

1435 CrossRef citations

1,419 Altmetric

John W. Tukey and Exploratory Data Analysis (EDA)



"actively incisive rather than passively descriptive, with real emphasis on the discovery of the unexpected" (Tukey, 1977)

Arose in reaction to an over-emphasis/reliance on NHST (aka confirmatory data analysis) – you need to understand and see your data to be able to suggest what hypotheses to test (many statisticians now abhor HST full stop)

This doesn't mean we don't start with theories/ideas/hypotheses to test – but that shouldn't be our statistical starting point

Same Stats, Different Graphs:

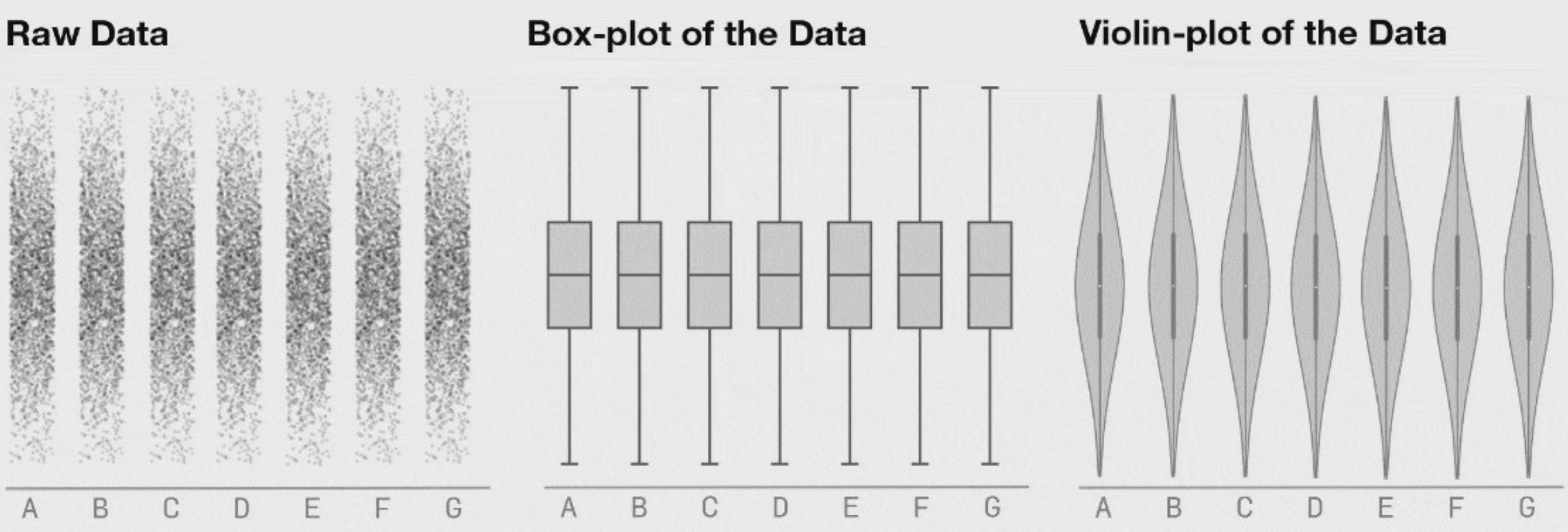
Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka
George Fitzmaurice

AUTODESK. RESEARCH

Why the mean (and other summary stats) aren't enough

http://dx.doi.org/10.1145/3025453.3025912



Case Study: The Fish Event Horizon & Meta-analysis

'Dark Age Economics' revisited: the English fish bone evidence AD 600-1600

James H. Barrett,[1] Alison M. Locker[2] & Callum M. Roberts[3]

When did the market economy come to Europe? Fish might seem an unlikely commodity to throw light on the matter, but the authors use fish bones from English sites to offer a vivid account of the rise and rise of the market as a factor in European development from the late tenth century.

Keywords: England, Europe, medieval, Dark Age, fish, trade, market

Introduction

Twenty-two years ago – when Richard Hodges (1982) published his influential monograph *Dark Age Economics* – two observations regarding early medieval economy seemed clear. Firstly, the transition from exchange of high-value prestige goods to low-value staples (and thus, in his view, from gift-exchange to market trade, from proto-urban settlements to true towns and from substantivist to formalist economics) was central to an understanding of European socio-economic change. Secondly, although complex and uneven in detail, this transition could be dated to the tenth and eleventh centuries. Hodges was, of course, not alone in these observations. The growth of trade and urbanism had long played an important role in defining the Viking Age (e.g. Arbmann 1939; Jankuhn 1956; Blindheim 1975; Bencard 1981). Moreover, *Dark Age Economics* was one contribution to a movement within medieval archaeology that was heavily influenced by economic and neo-evolutionary anthropology (e.g. Grierson 1959; Callmer 1977; Randsborg 1980; Jankuhn 1982). It thus found an audience primed for either reception or resistance (Astill 1985; Sawyer 1989).

Since then, however, archaeology has confirmed the existence of early (particularly eighth century) antecedents to many of North-western Europe's first towns, and of other early markets without urban populations (Cowie & Whytehead 1988; Hill *et al.* 1990; Ulriksen 1994; Kemp 1996; Feveile & Jensen 2000; Gardiner *et al.* 2001; see contributions in Hansen & Wickham 2000; Hill & Cowie 2001; Prestell & Ulmschneider 2003). Concurrently, accessible surveys of the relevant historical evidence have emphasised the existence and scale of commercial transactions – including the exchange of basic staple goods – in Carolingian times (e.g. Verhulst 1995; 2002). Wider paradigm shifts within archaeology have also peripheralised the neo-evolutionary basis of Hodges' original argument (Gosden 1999:88-105; Gerrard 2003:172, 217-231). It is thus not surprising to find that interpretations have

1 *Department of Archaeology, The King's Manor, University of York, York, YO1 7EP, England. (Email: jhb5@york.ac.uk)*
2 *L'Ensoleillee', 20 bld de Garavan, 06500 Menton, France. (Email: glocker@monaco.mc)*
3 *Environment Department, University of York, York, YO10 5DD, England. (Email: cr10@york.ac.uk)*

Received: 9 February 2004; Accepted: 29 March 2004

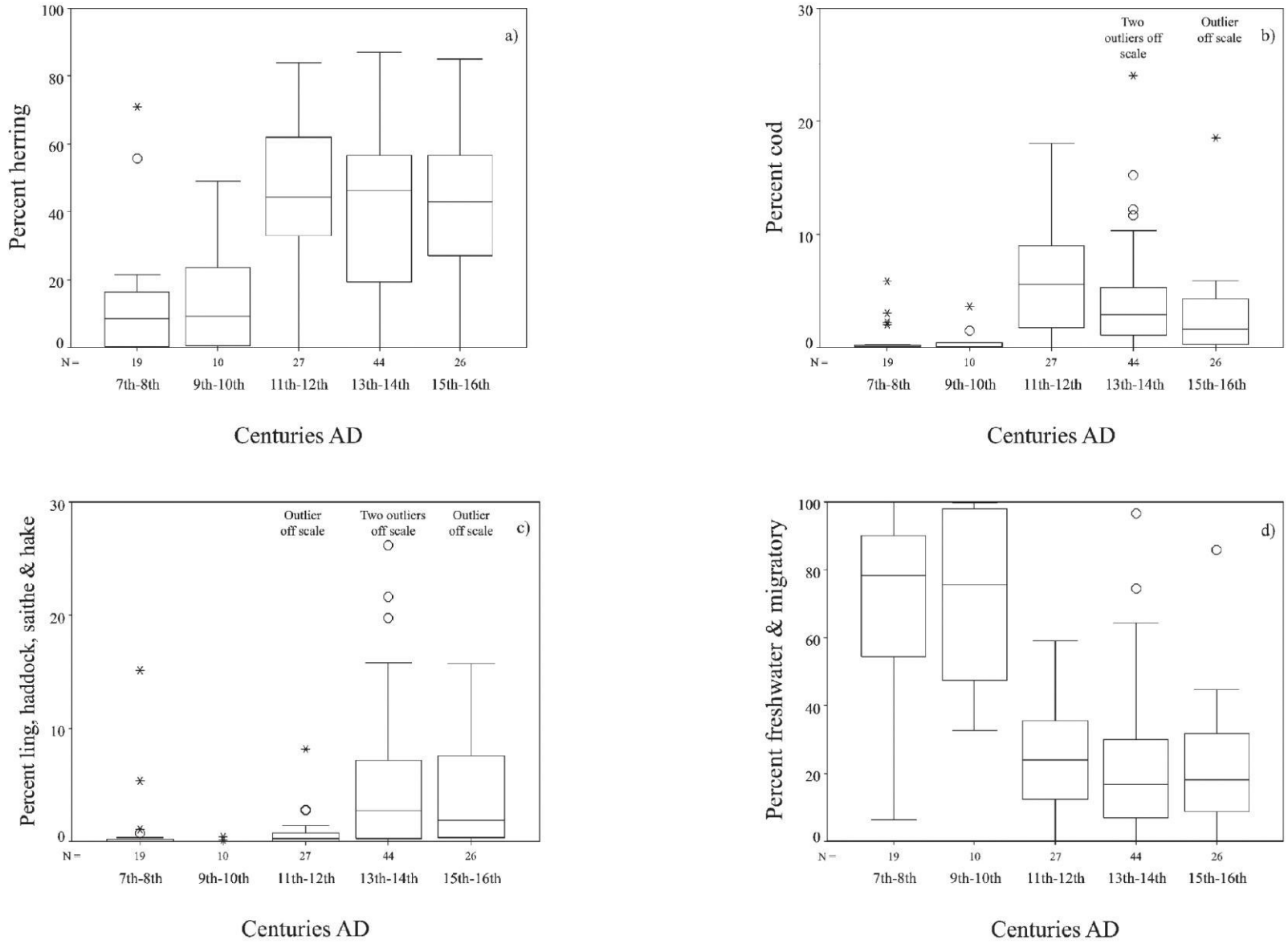


Figure 3. (a through c) Boxplots showing the percentages of common marine species in English fish bone assemblages from AD 600 to 1600 (based on the number of identified specimens). (d) For comparison, the percentage of freshwater and migratory taxa is also shown – based on cyprinids, pike, perch, eel, smelt, salmonids and flatfish (many of which are probably flounder, which enters fresh water).

'Dark Age Economics' revisited: the English fish bone evidence AD 600-1600

James H. Barrett,[1] Alison M. Locker[2] & Callum M. Roberts[3]

When did the market economy come to Europe? Fish might seem an unlikely commodity to throw light on the matter, but the authors use fish bones from English sites to offer a vivid account of the rise and rise of the market as a factor in European development from the late tenth century.

Keywords: England, Europe, medieval, Dark Age, fish, trade, market

Introduction

Twenty-two years ago – when Richard Hodges (1982) published his influential monograph *Dark Age Economics* – two observations regarding early medieval economy seemed clear. Firstly, the transition from exchange of high-value prestige goods to low-value staples (and thus, in his view, from gift-exchange to market trade, from proto-urban settlements to true towns and from substantivist to formalist economics) was central to an understanding of European socio-economic change. Secondly, although complex and uneven in detail, this transition could be dated to the tenth and eleventh centuries. Hodges was, of course, not alone in these observations. The growth of trade and urbanism had long played an important role in defining the Viking Age (e.g. Arbmann 1939; Jankuhn 1956; Blindheim 1975; Bencard 1981). Moreover, *Dark Age Economics* was one contribution to a movement within medieval archaeology that was heavily influenced by economic and neo-evolutionary anthropology (e.g. Grierson 1959; Callmer 1977; Randsborg 1980; Jankuhn 1982). It thus found an audience primed for either reception or resistance (Astill 1985; Sawyer 1989).

Since then, however, archaeology has confirmed the existence of early (particularly eighth century) antecedents to many of North-western Europe's first towns, and of other early markets without urban populations (Cowie & Whytehead 1988; Hill *et al.* 1990; Ulriksen 1994; Kemp 1996; Feveile & Jensen 2000; Gardiner *et al.* 2001; see contributions in Hansen & Wickham 2000; Hill & Cowie 2001; Prestell & Ulmschneider 2003). Concurrently, accessible surveys of the relevant historical evidence have emphasised the existence and scale of commercial transactions – including the exchange of basic staple goods – in Carolingian times (e.g. Verhulst 1995; 2002). Wider paradigm shifts within archaeology have also peripheralised the neo-evolutionary basis of Hodges' original argument (Gosden 1999:88-105; Gerrard 2003:172, 217-231). It is thus not surprising to find that interpretations have

1 *Department of Archaeology, The King's Manor, University of York, York, YO1 7EP, England. (Email: jhb5@york.ac.uk)*
2 *L'Ensoleillee', 20 bld de Garavan, 06500 Menton, France. (Email: glocker@monaco.mc)*
3 *Environment Department, University of York, York, YO10 5DD, England. (Email: cr10@york.ac.uk)*

Received: 9 February 2004; Accepted: 29 March 2004

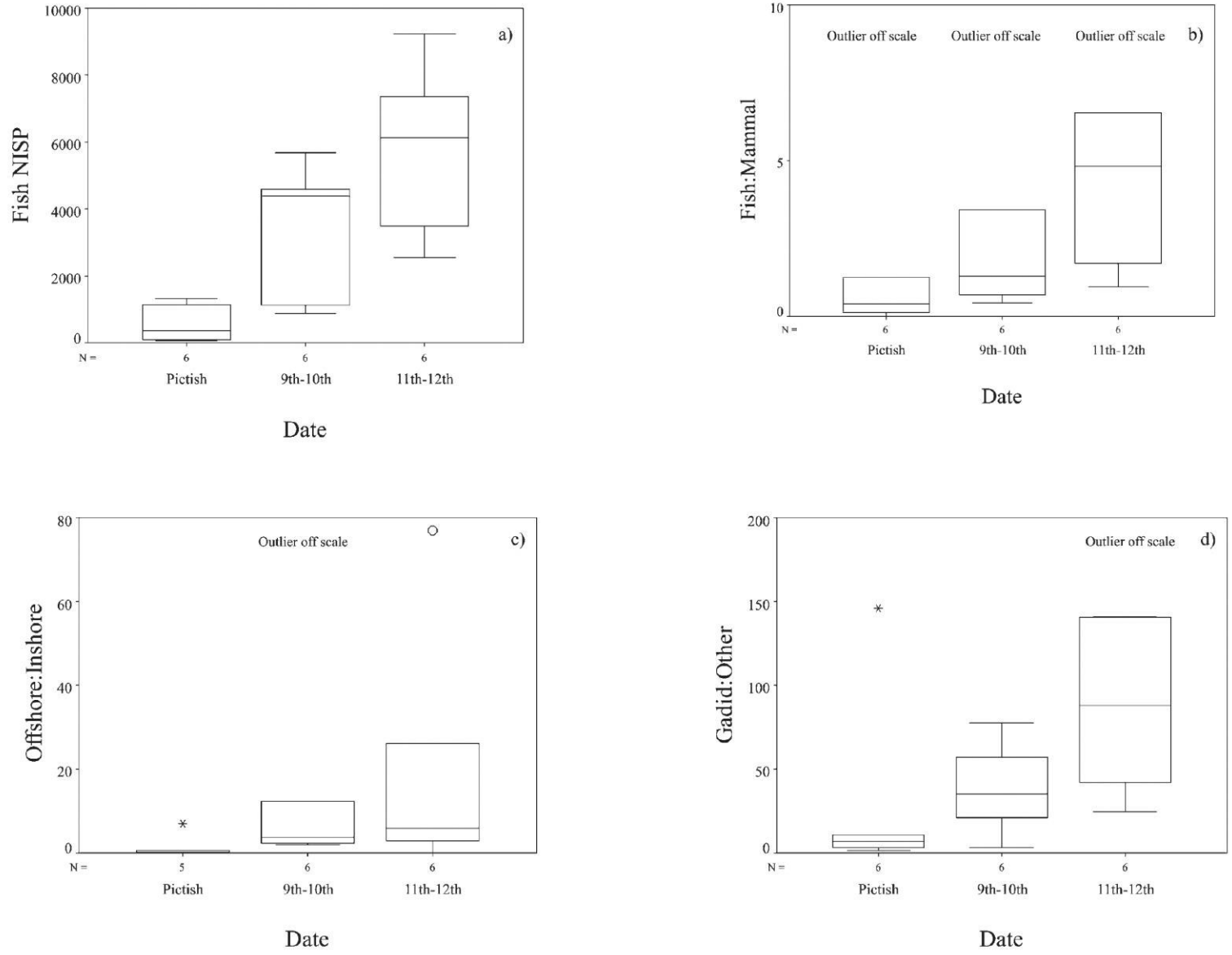
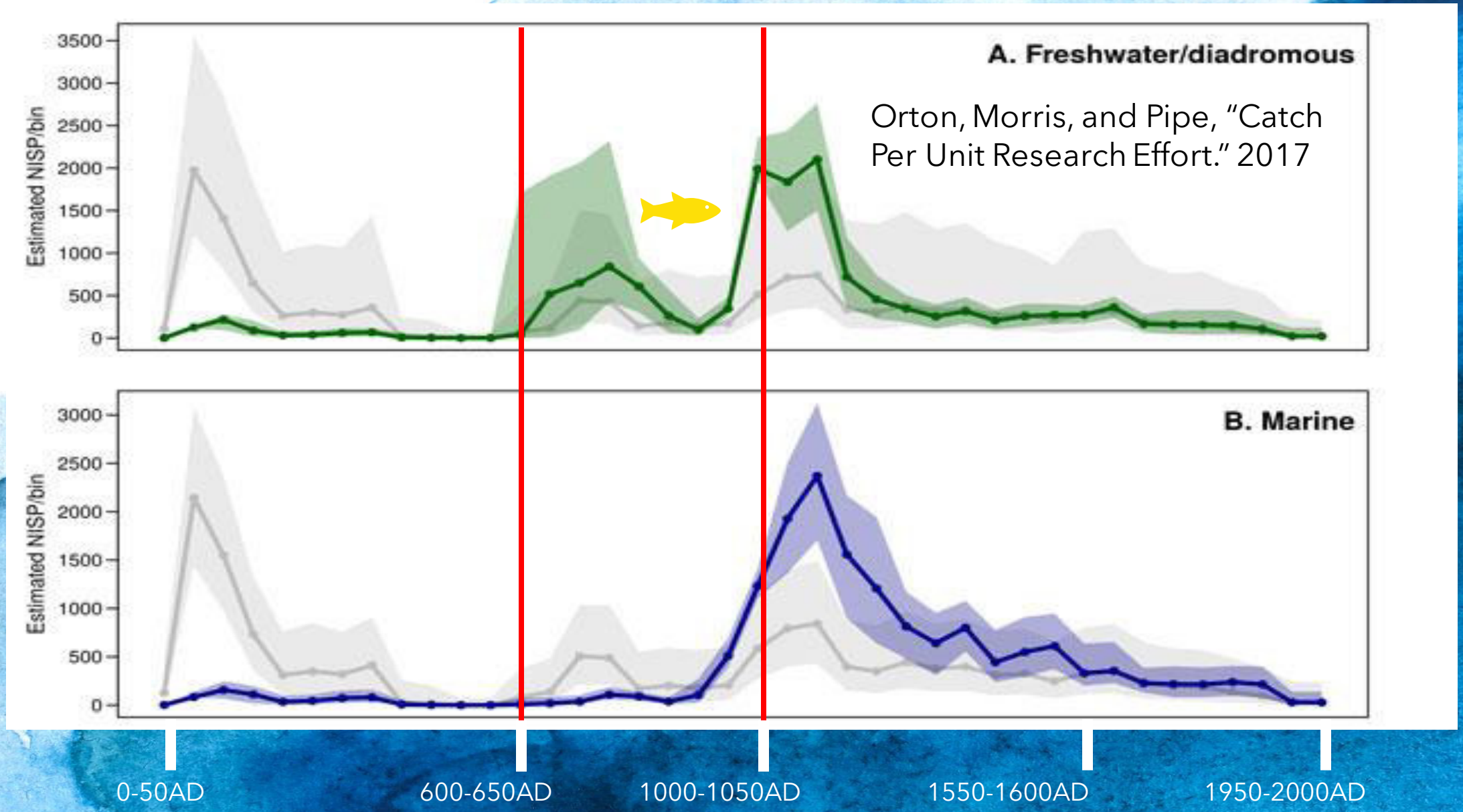


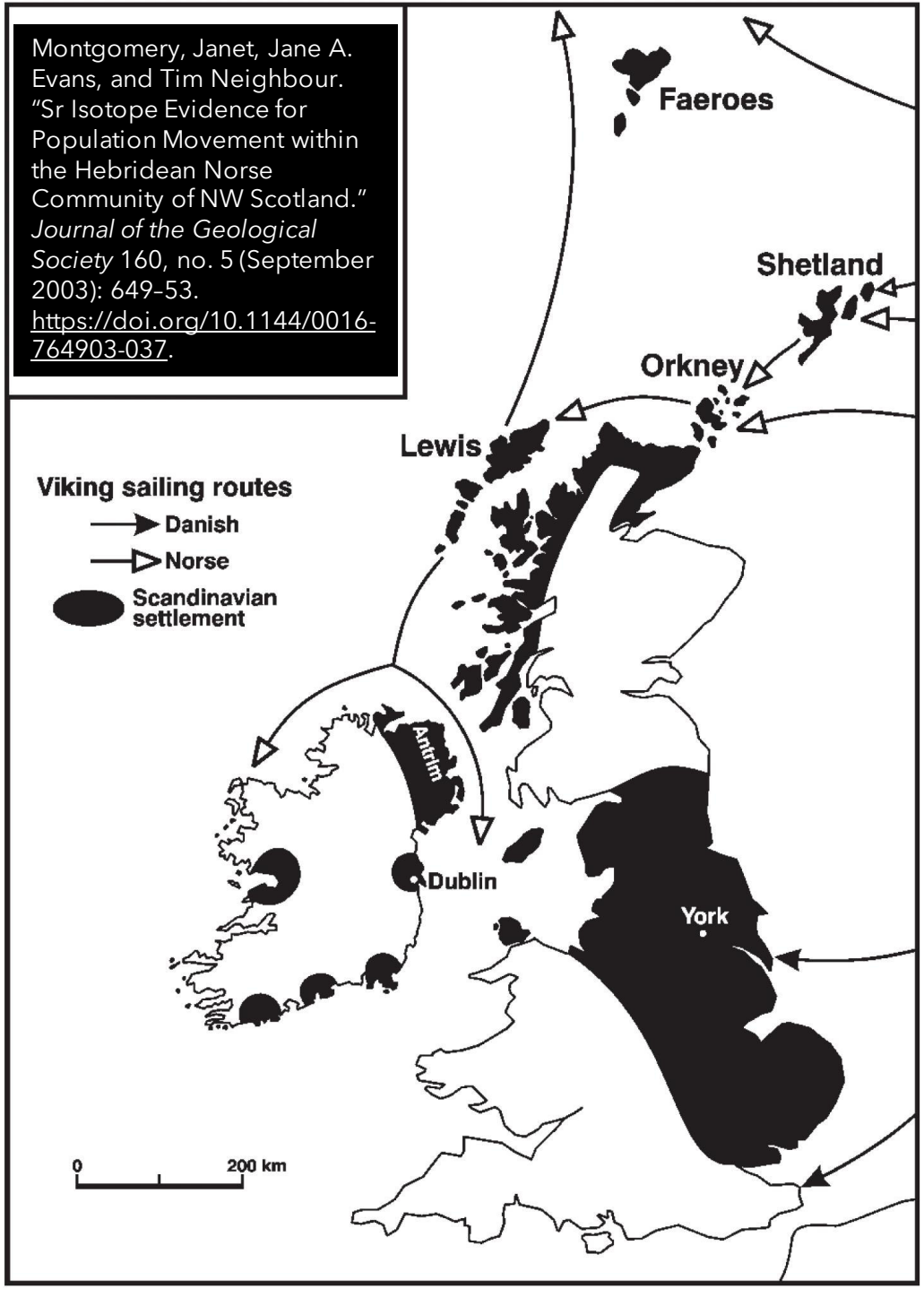
Figure 4. Boxplots showing increases in the intensity of fishing, and the importance of cod and related species, in northern Scotland during both the ninth/tenth and eleventh/twelfth centuries AD. The preceding 'Pictish' period covers approximately the fourth to eighth centuries. (a) The number of fish bones recovered. (b) The ratio of fish bone to mammal bone. (c) The ratio of inshore to offshore taxa – based on a comparison of ling (Molva molva) and Torsk (Brosme brosme) to rocklings (Ciliata or Gaidropsarus species), wrasse (Labridae) and cottids (Cottidae). (d) The ratio of cod family to all other fish. The data are based on NISP figures and have been taken from Barrett and Oltmann (1998); Barrett et al. (1999; 2001) and references therein.

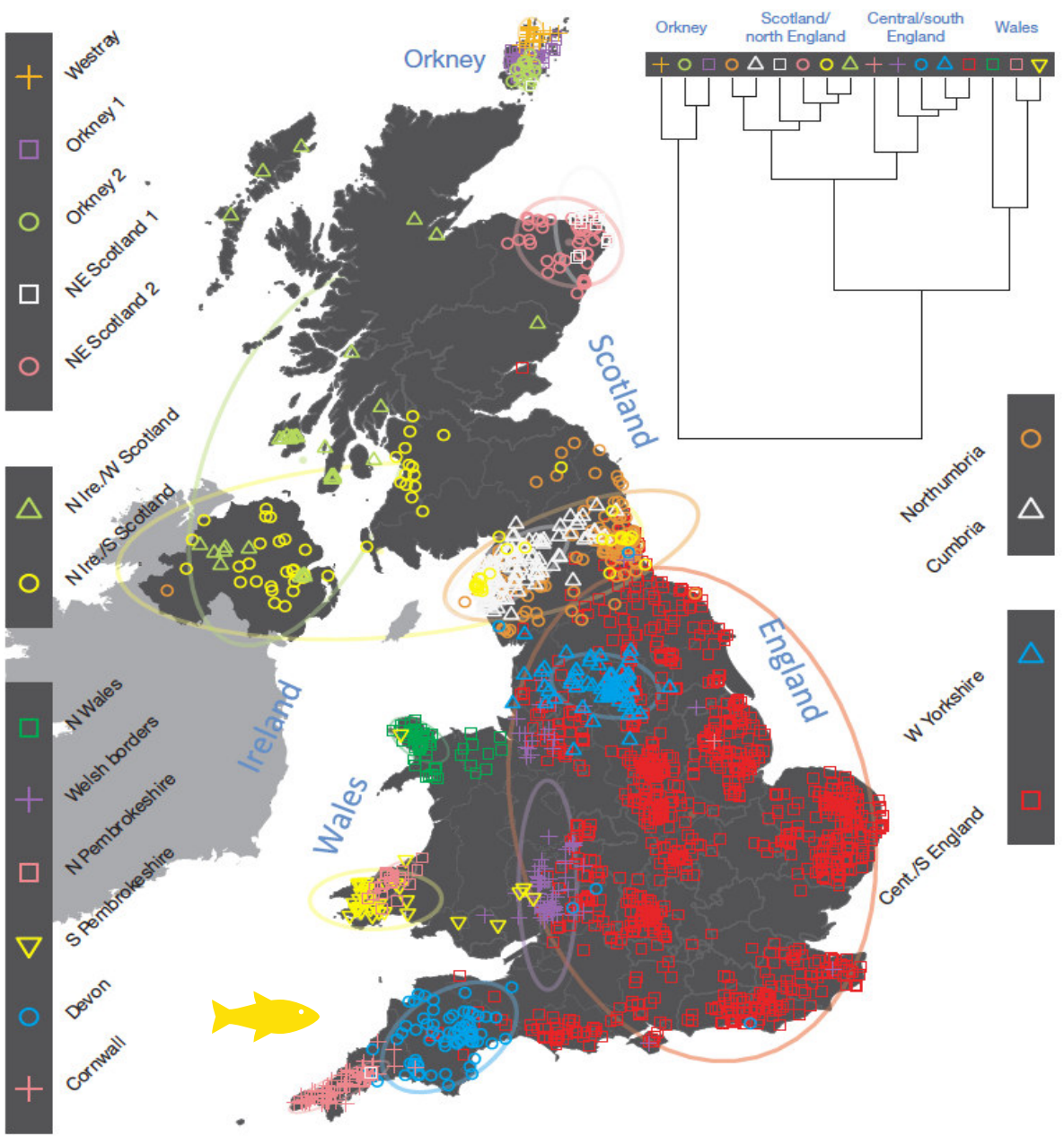


Orton, Morris, and Pipe, "Catch Per Unit Research Effort." 2017

Fish Event Horizon(s)

Montgomery, Janet, Jane A. Evans, and Tim Neighbour. "Sr Isotope Evidence for Population Movement within the Hebridean Norse Community of NW Scotland." *Journal of the Geological Society* 160, no. 5 (September 2003): 649–53. https://doi.org/10.1144/0016-764903-037.



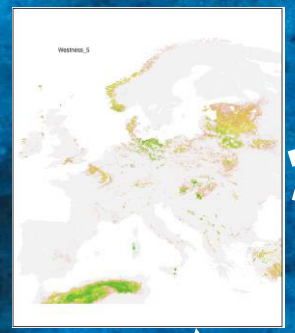


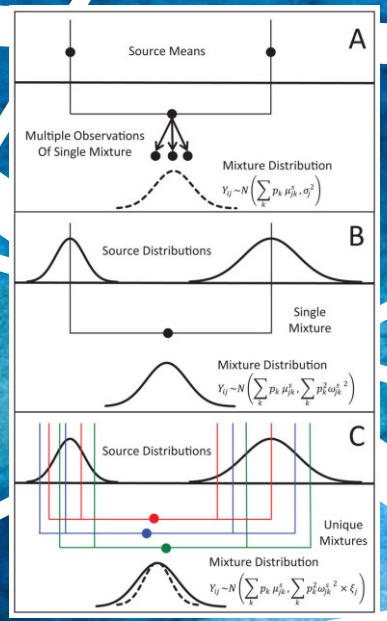
Leslie et al. 2015 "The Fine-Scale Genetic Structure of the British Population."

The ArchaeoFINS project

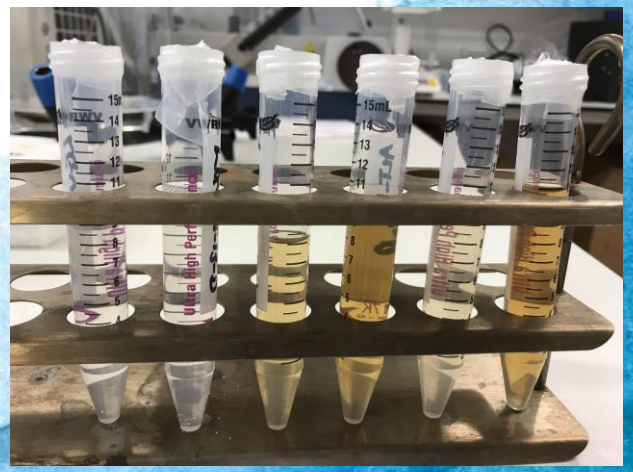
- When do fish disappear from and re-enter diets in Scotland at a scale detectable by biomolecular techniques?
- Does marine consumption uptake correlate with major migration events or socio-cultural shifts (e.g. Vikings)?
- How do the diet and mobility signatures in Scotland and Ireland fit within the broader context of Western Europe?

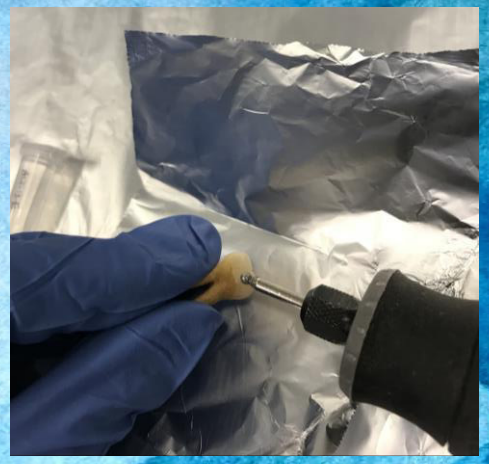
The ArchaeoFINS project

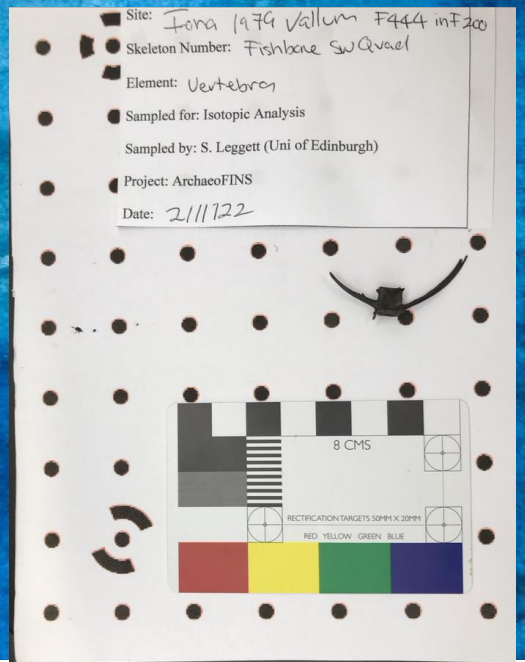




ArchaeoFINS – expanding the dataset – current work

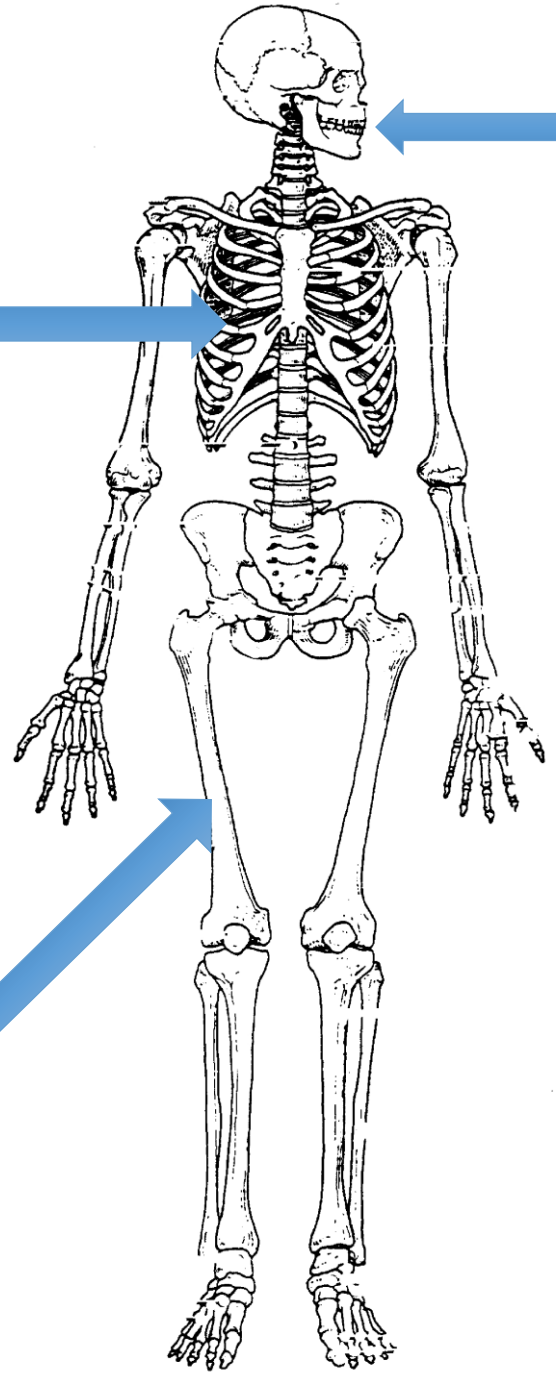






Tissue Differences

Ribs (<10yrs before death)



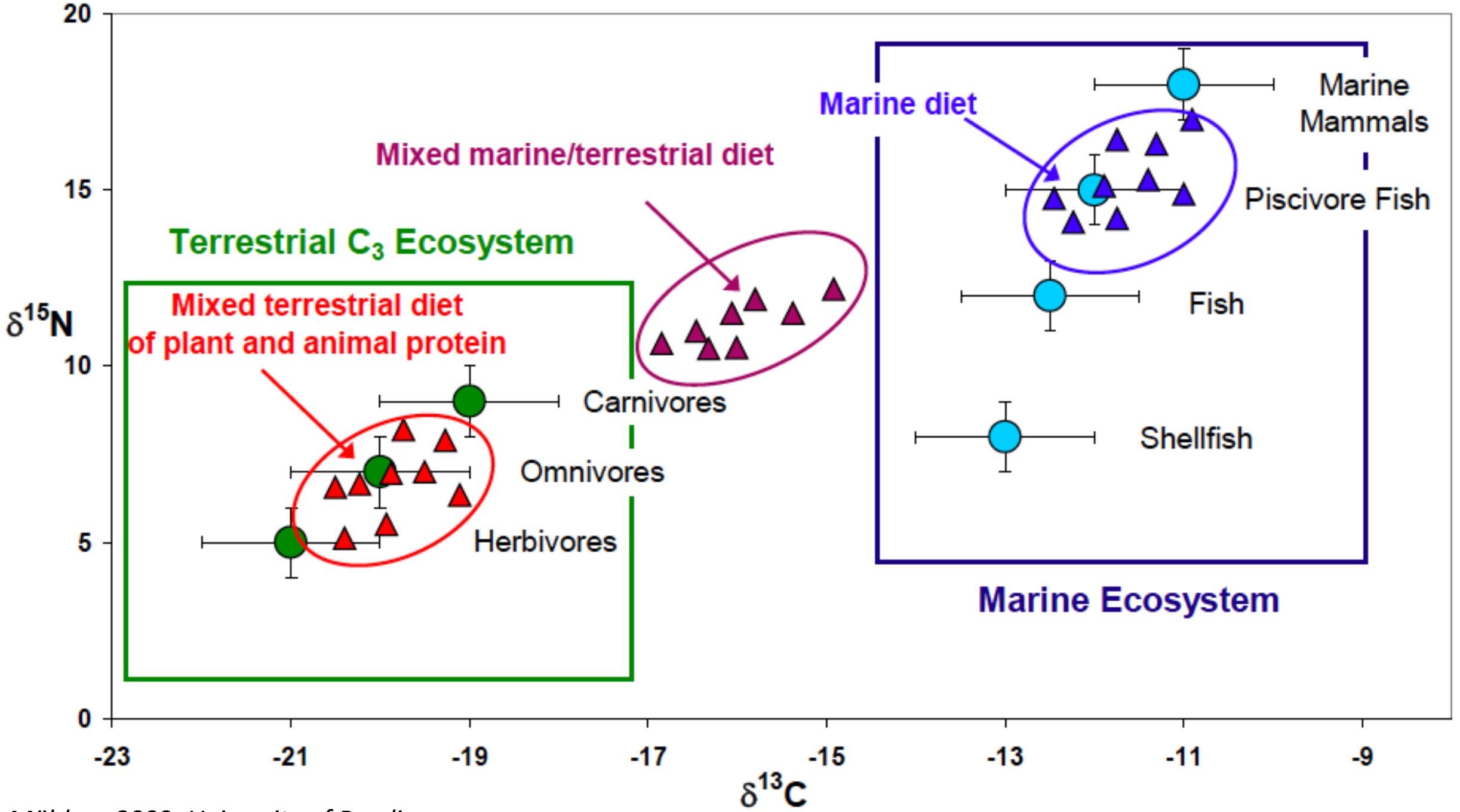
Teeth (childhood)

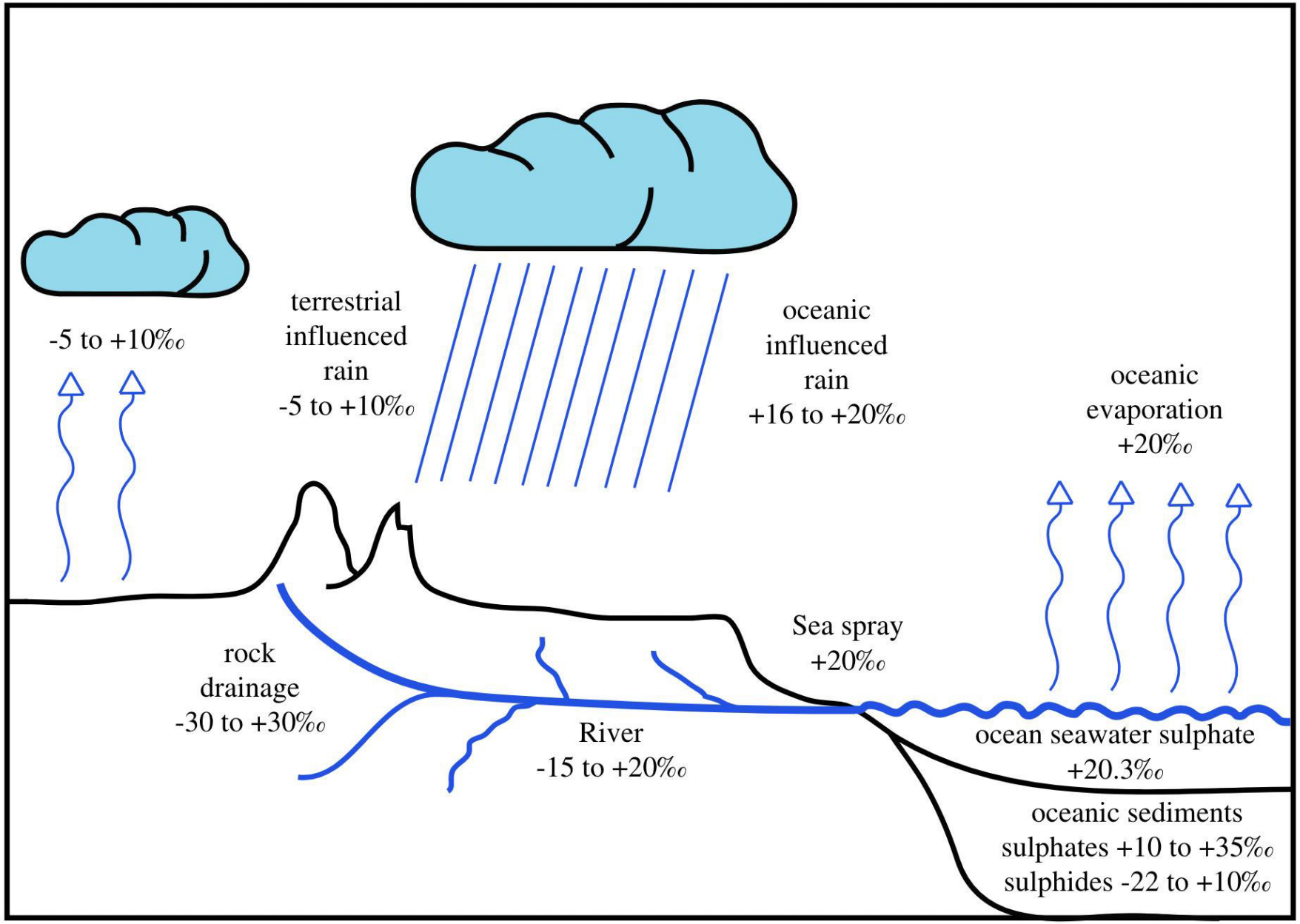
2nd Pre-molars and 2nd molars

Enamel: 6-7/7-8 years old

Dentine: 12-14/14-16 years old

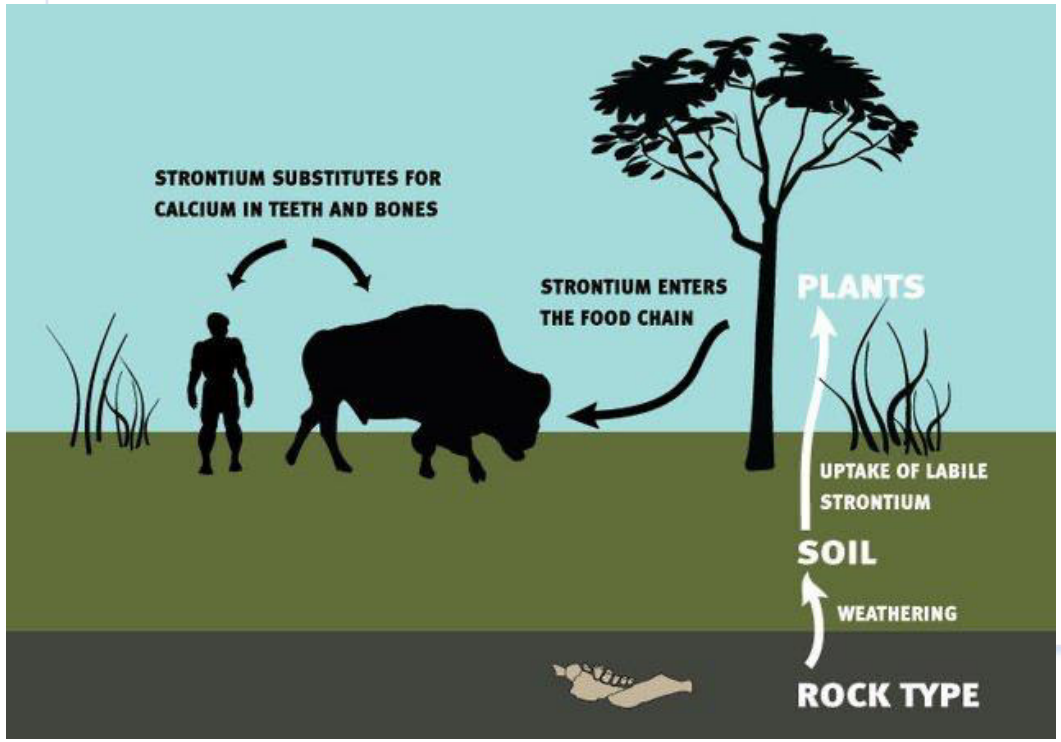
Cortical Bone (20 years+)

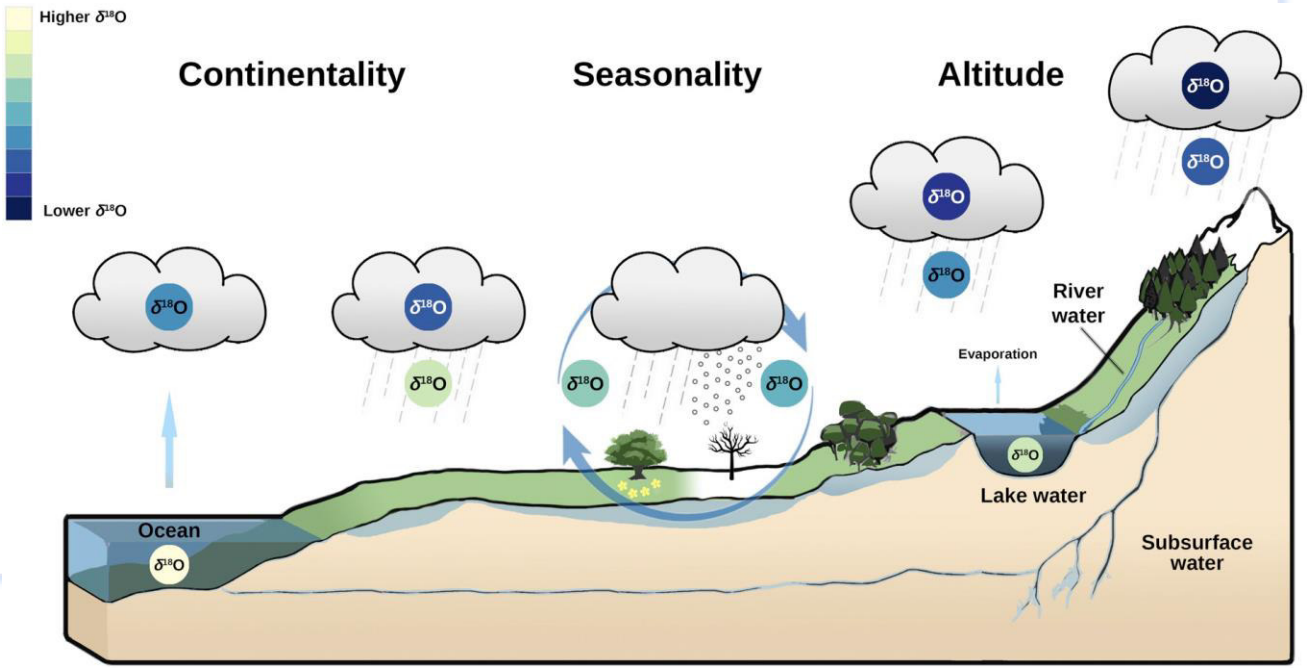




Nehlich, Olaf. 'The Application of Sulphur Isotope Analyses in Archaeological Research: A Review'. *Earth-Science Reviews* 142 (March 2015): 1–17. https://doi.org/10.1016/j.earscirev.2014.12.002.

Isotopic Analysis – you are What you Eat/Drink and Where It Came From

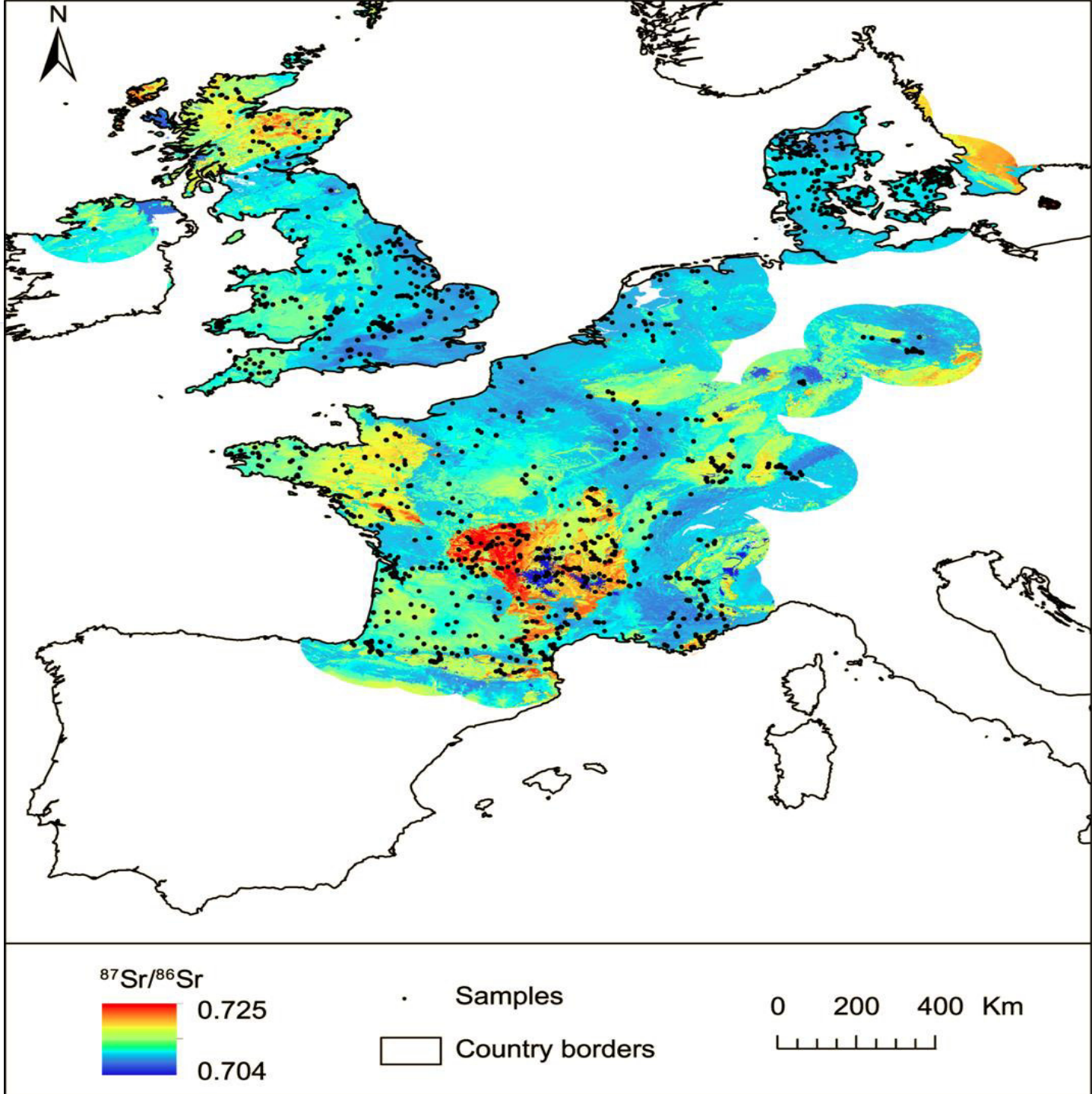




http://timeteam.lunchbox.pbs.org/time-team/experience-archaeology/isotope-analysis/

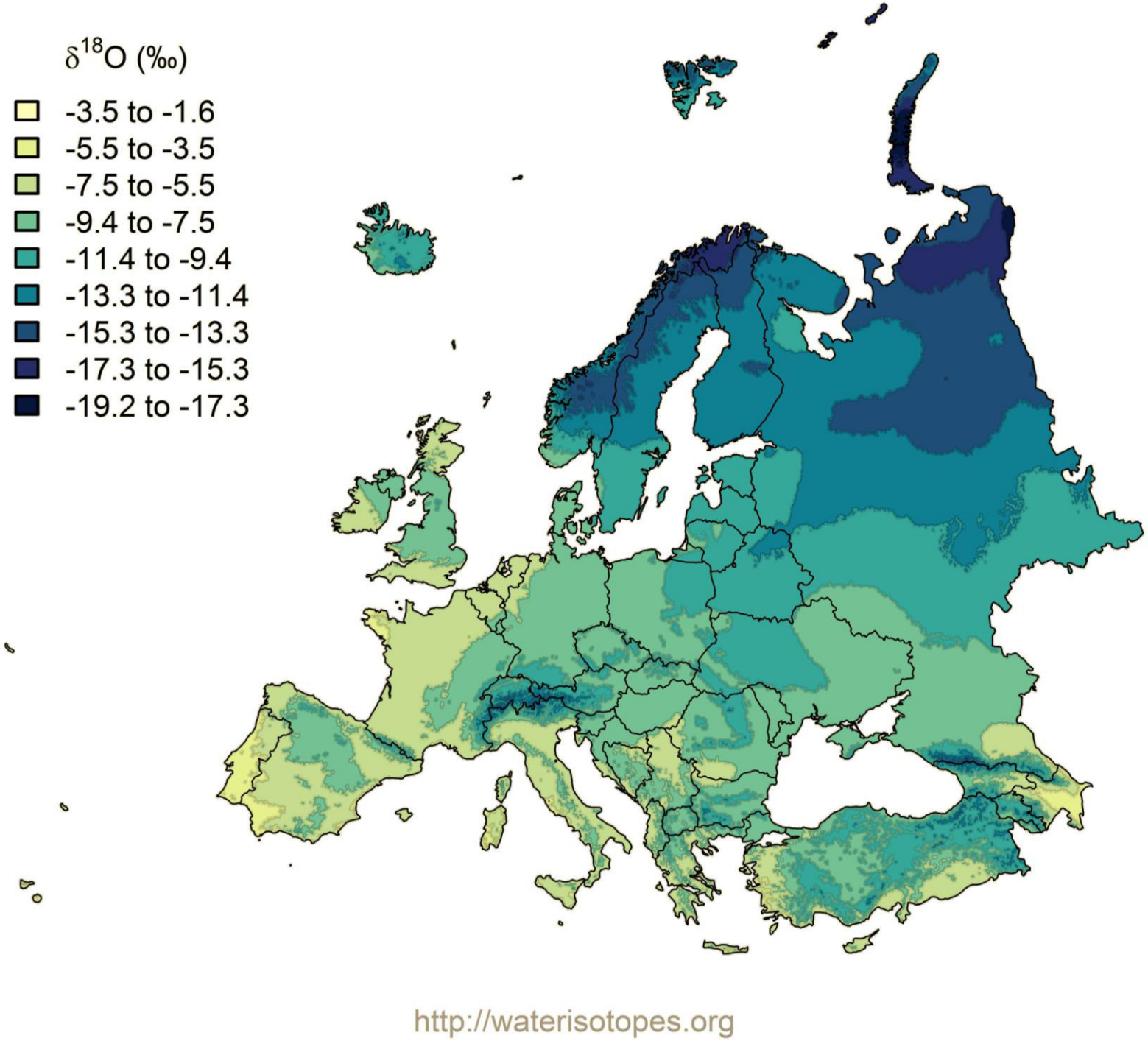
Pederzani and Britton, (2019) "Oxygen Isotopes in Bioarchaeology."

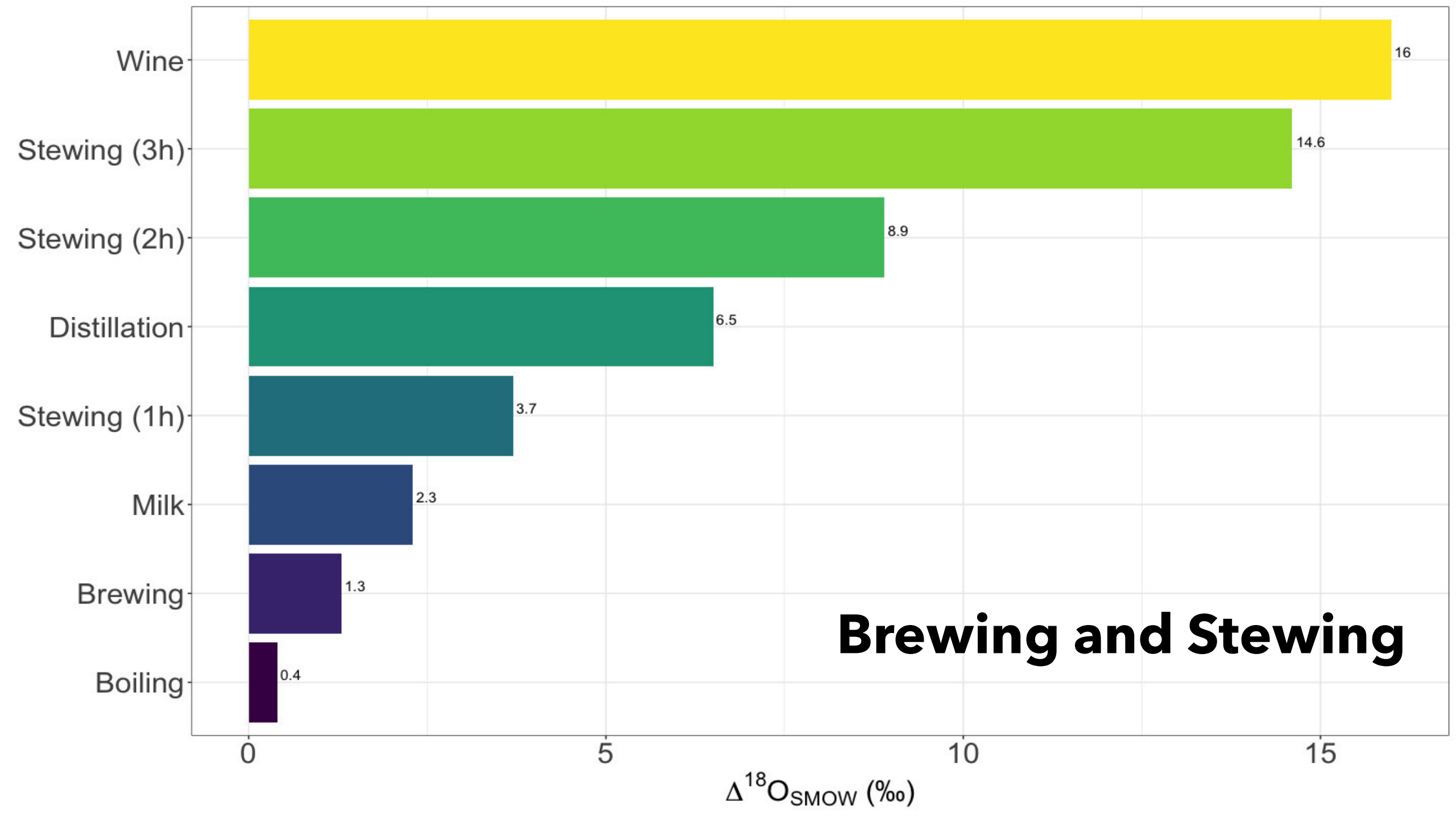
ISOTOPIC PROVENANCING
GEOLOGY

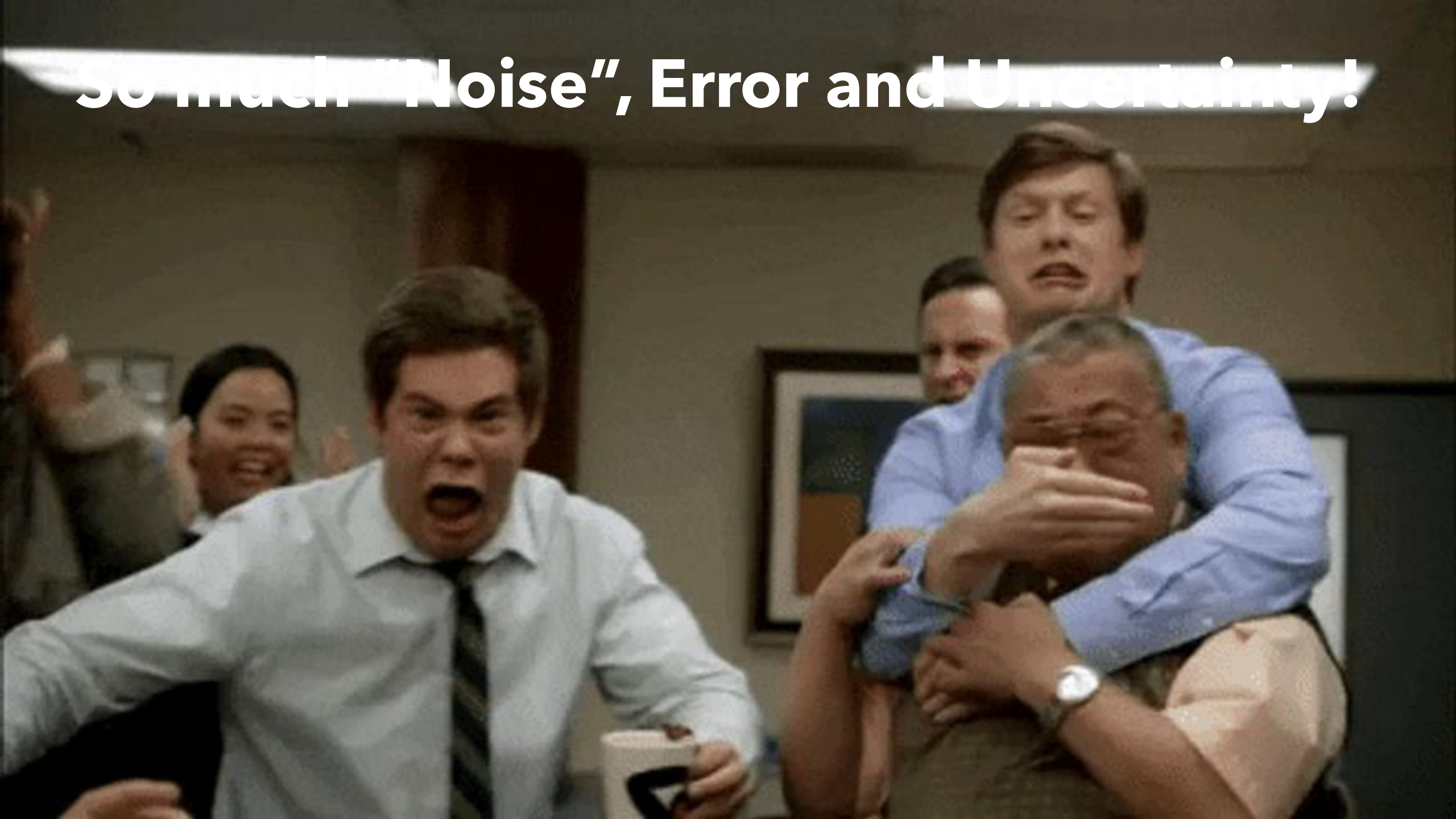


Bataille et al. 2018 “A Bioavailable Strontium Isoscape for Western Europe.”

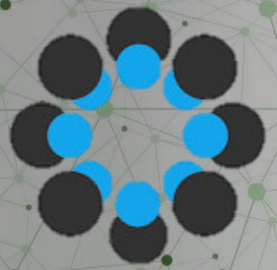
ISOTOPIC PROVENANCING
CLIMATE




Wine
Stewing (3h)
Stewing (2h)
Distillation
Stewing (1h)
Milk
Brewing
Boiling
16
14.6
8.9
6.5
3.7
2.3
1.3
0.4
0
5
10
15
Δ18OSMOW (‰)
Brewing and Stewing


So much "Noise", Error and Uncertainty!

New Statistics and EDA to the Rescue!



The following graphs/stats were produced in R/Rstudio – largely in the tidyverse. All data & code from published papers/pre-prints available through my Github and/or OSF accounts:

https://github.com/samleggs22

https://osf.io/jmuxw/

Early medieval diet, mobility and the FEH

Fishhooks from Wood Quay Dublin

North Atlantic

Leggett, 'A Hierarchical Meta-Analytical Approach to Western European Dietary Transitions in the First Millennium AD'. https://doi.org/10.1017/eaa.2022.23

Leggett, S, A Rose, E Praet, and P Le Roux. *Ecology* 102, no. 6 (2021): e03349. https://doi.org/10.1002/ecy.3349.

- Human Tooth Enamel
- Faunal Bone Collagen
- Human Dentine Collagen
- Human Bone Collagen

0 750 1500 km

Atlantic & Arctic Norway

Scotland & Scottish Isles

Skagerrak-Kattegat-Jutland Basin

Baltic

Irish Sea

England

Frisia & Saxony

Normandy/ Neustria

Austrasia & Burgundy

Austro-Hungary & Bavaria

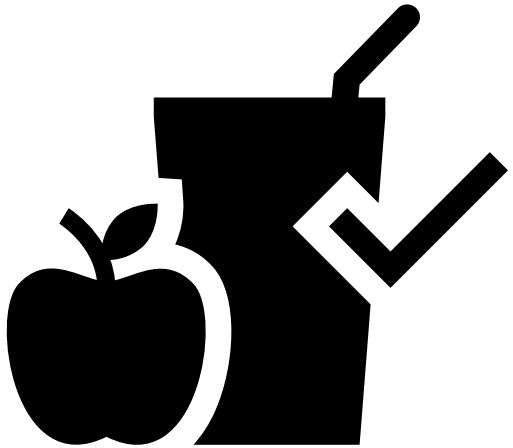
Po Valley

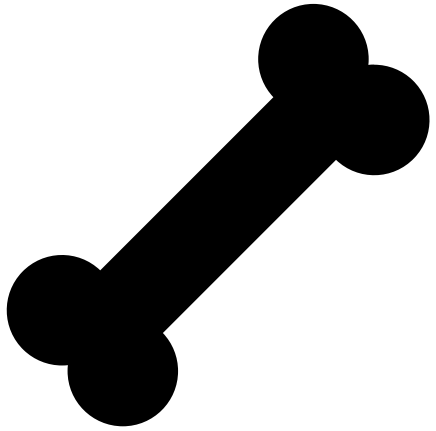
Croatia

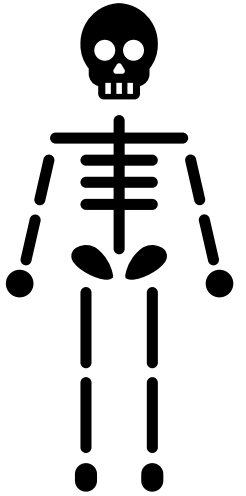
Inland and Western Iberia

Balearic Islands & Tyrrhenian Sea

Greece

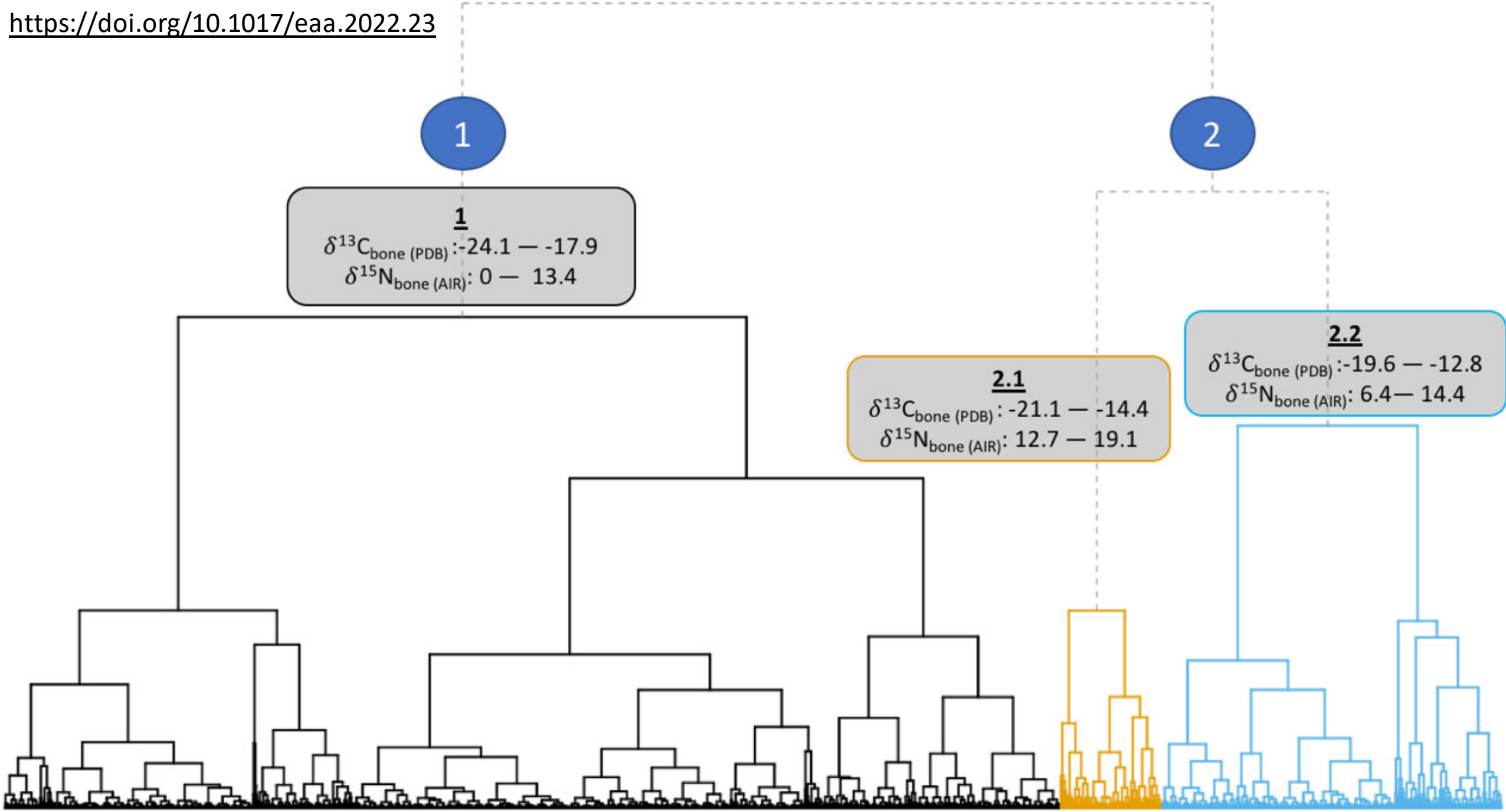


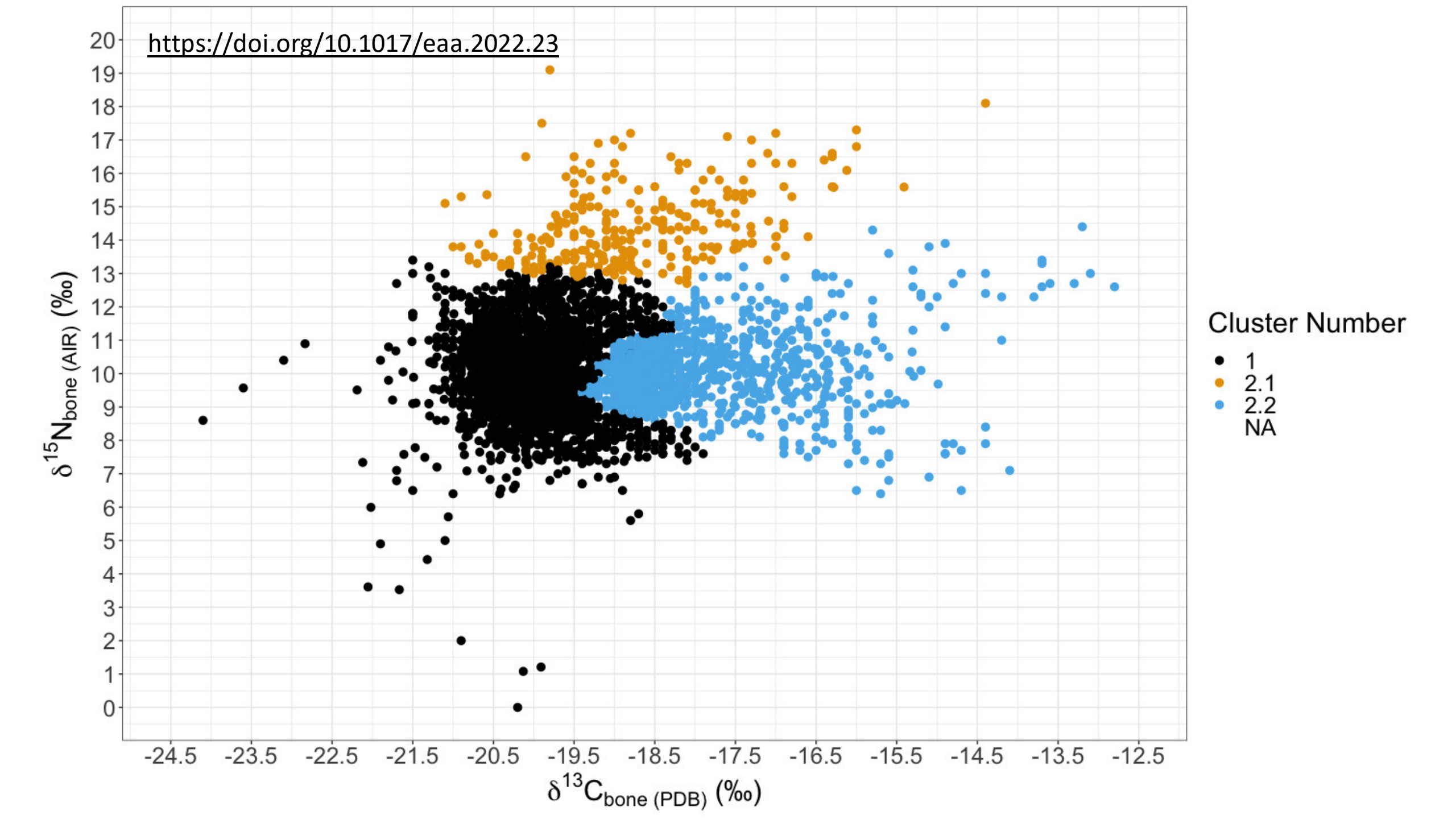


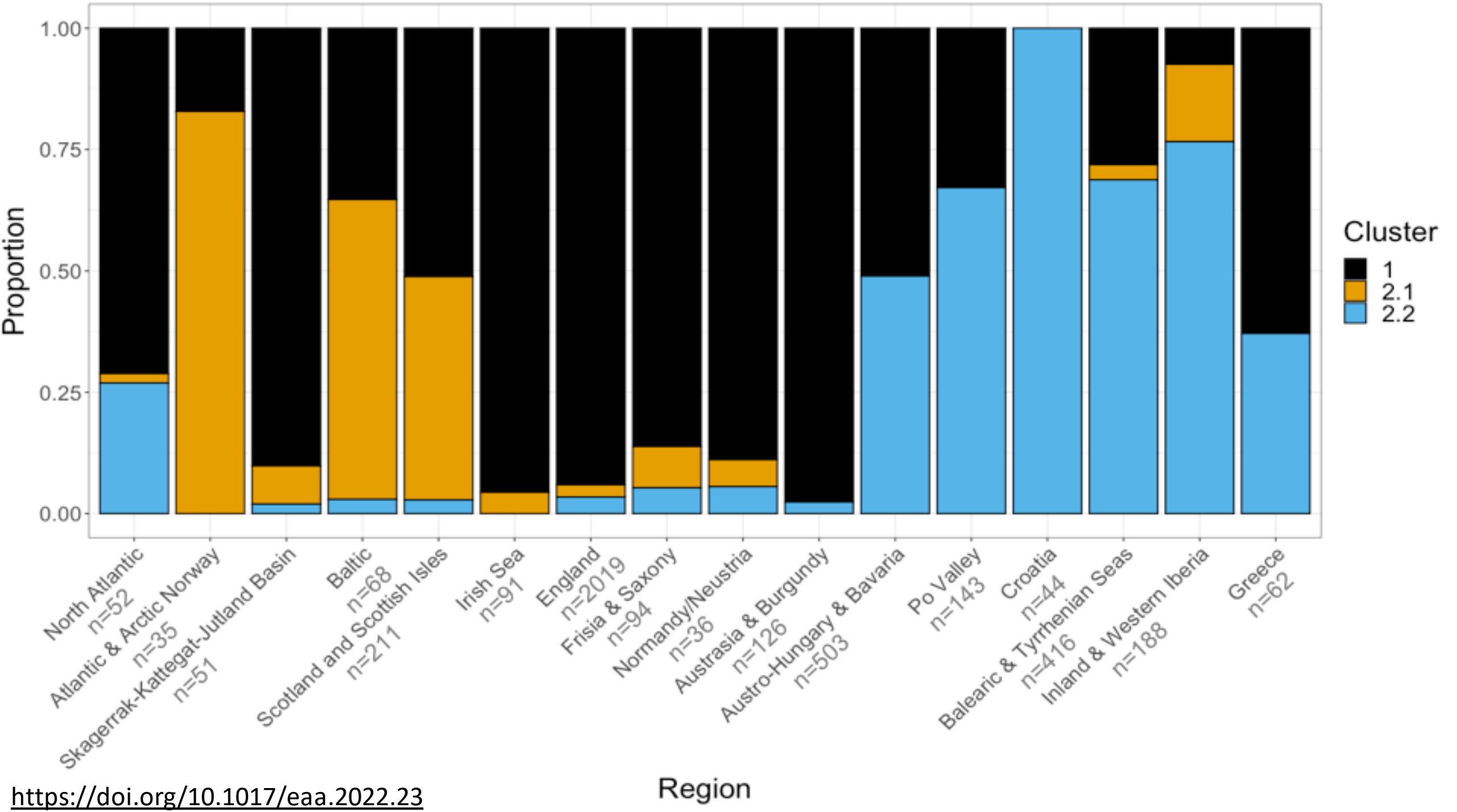


Later Life Diet – Postcranial Bones

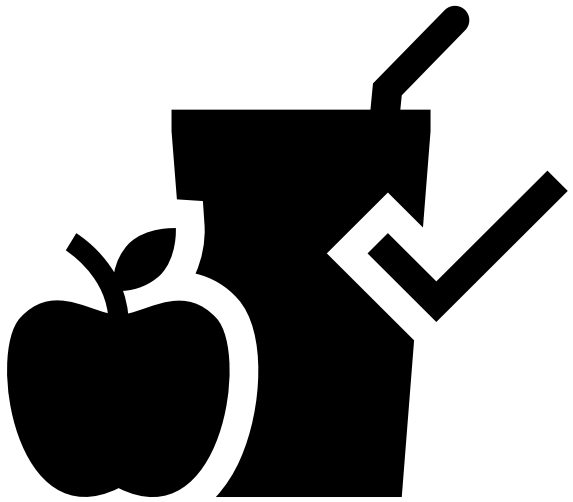
https://doi.org/10.1017/eaa.2022.23

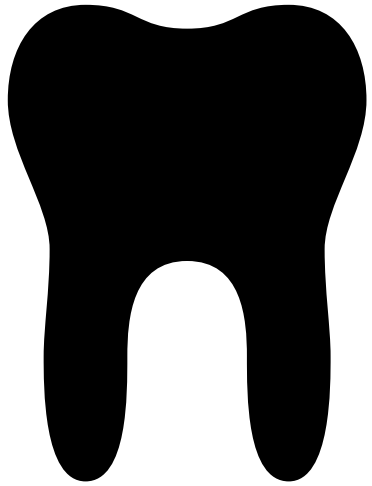



https://doi.org/10.1017/eaa.2022.23
δ15Nbone (AIR) (‰)
δ13Cbone (PDB) (‰)
Cluster Number
1
2.1
2.2
NA

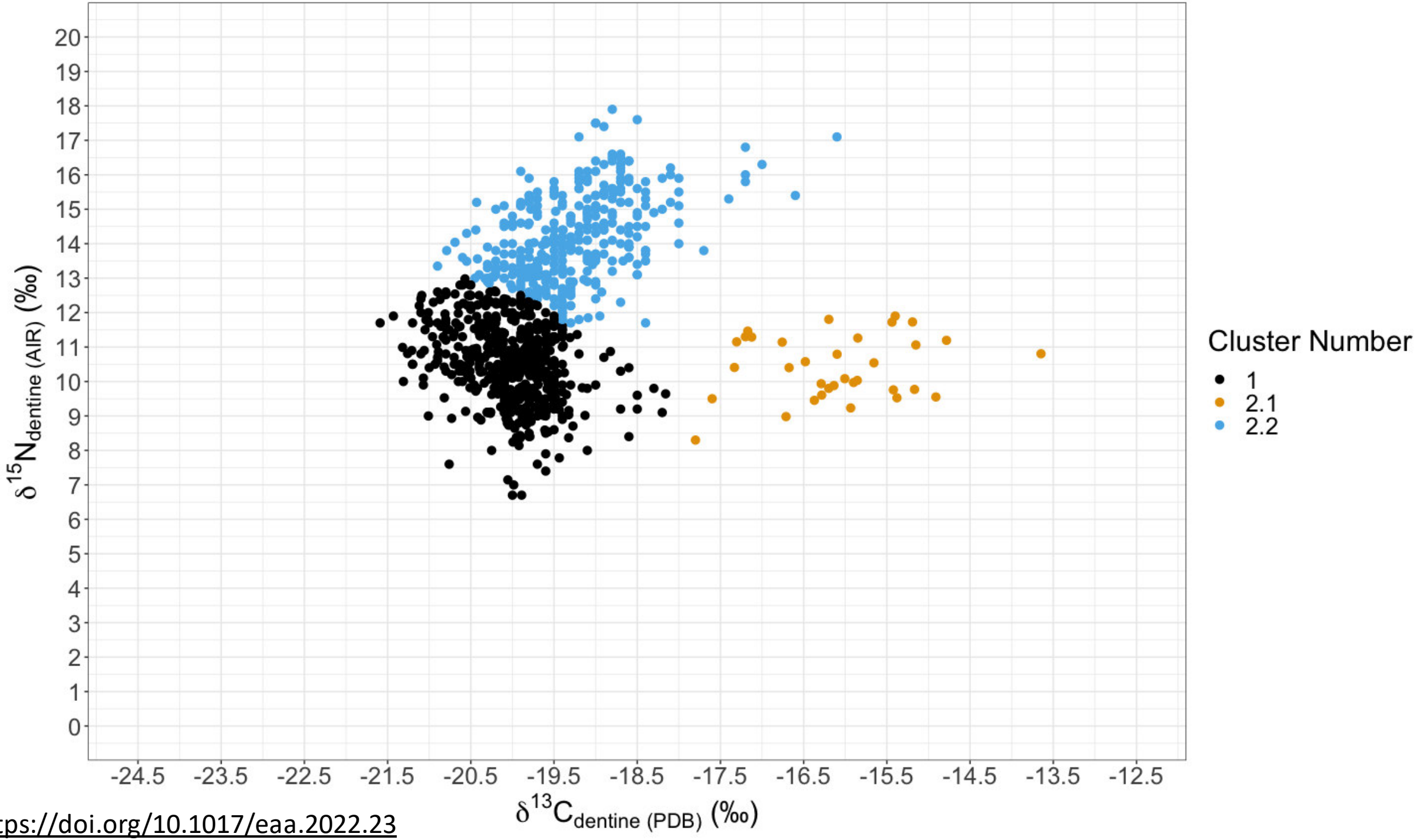


https://doi.org/10.1017/eaa.2022.23

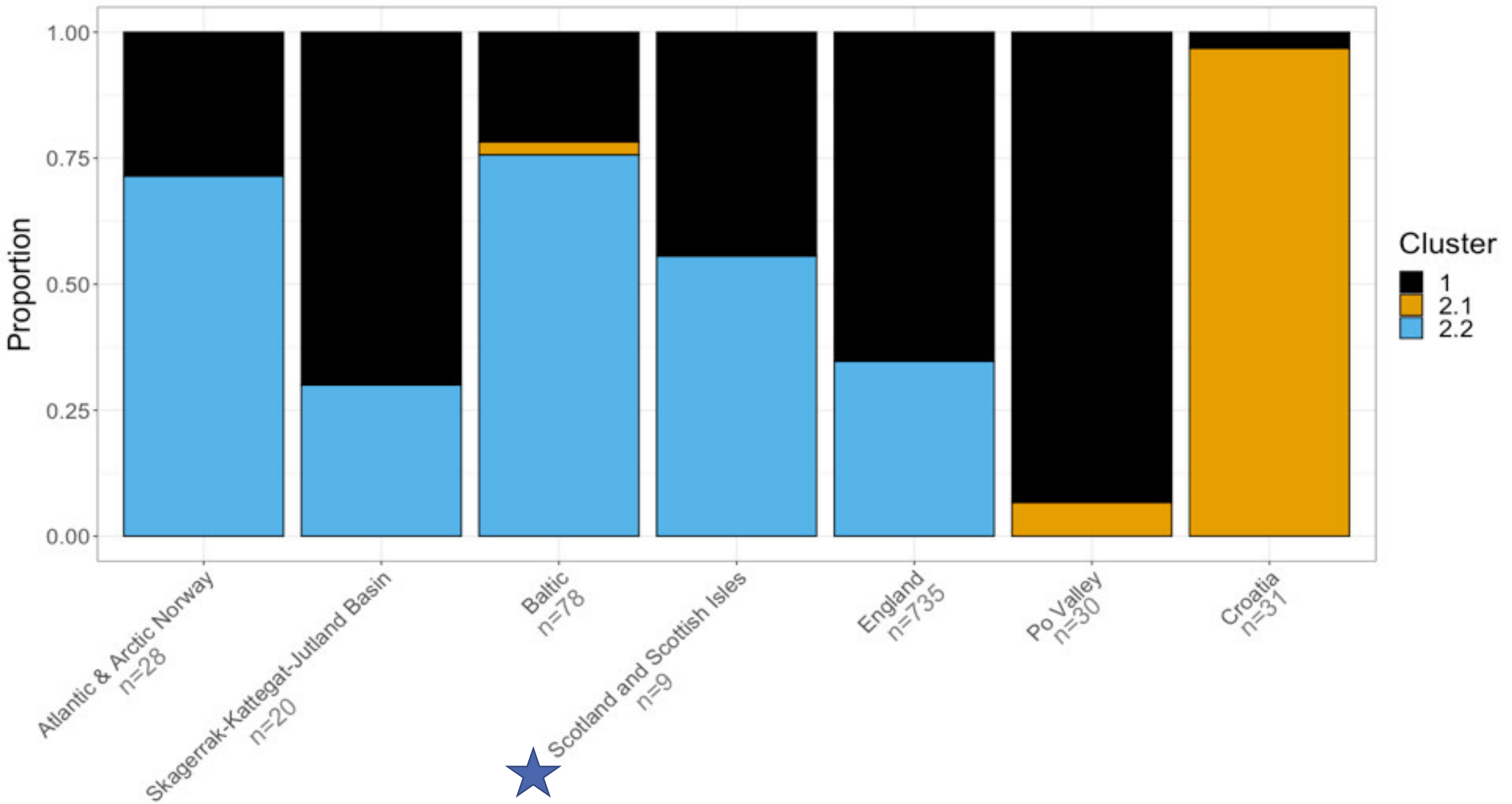




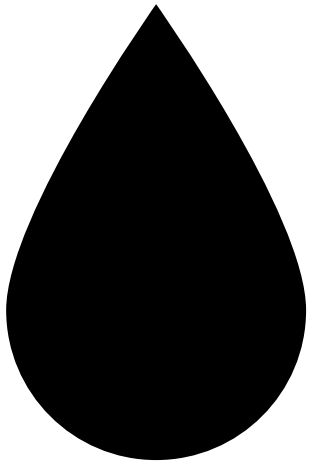
Childhood Diet - Dentine

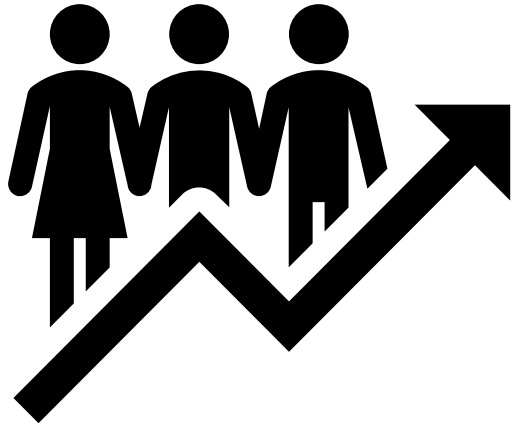


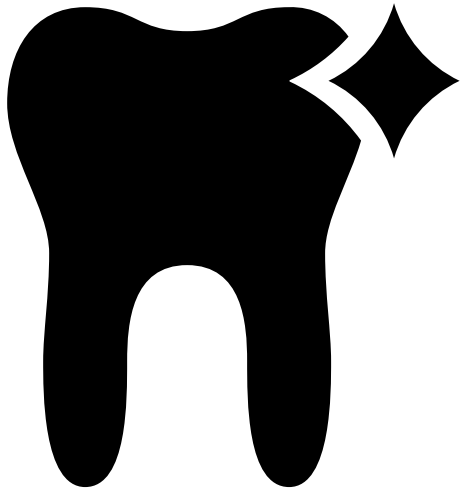
https://doi.org/10.1017/eaa.2022.23



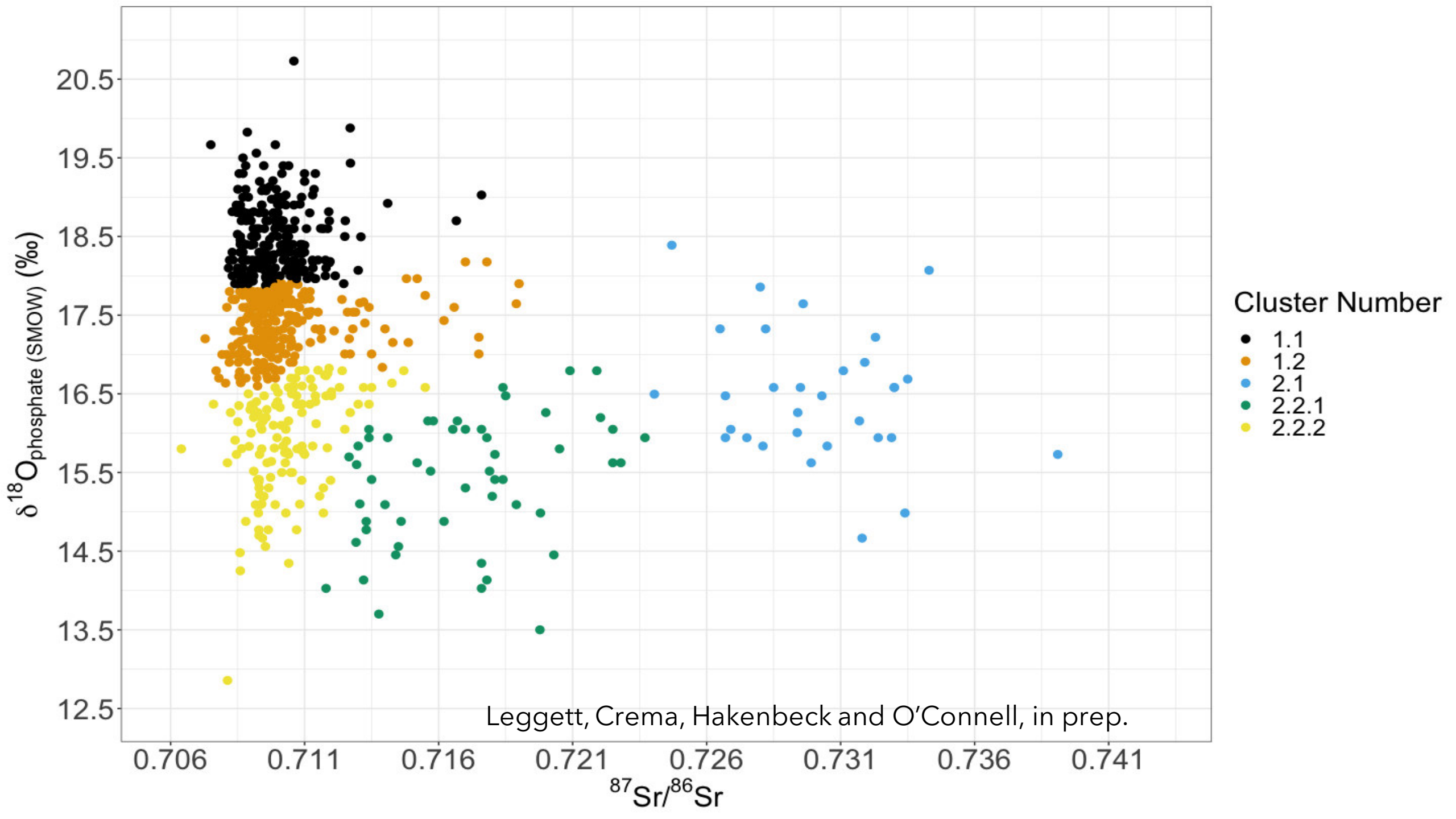
https://doi.org/10.1017/eaa.2022.23

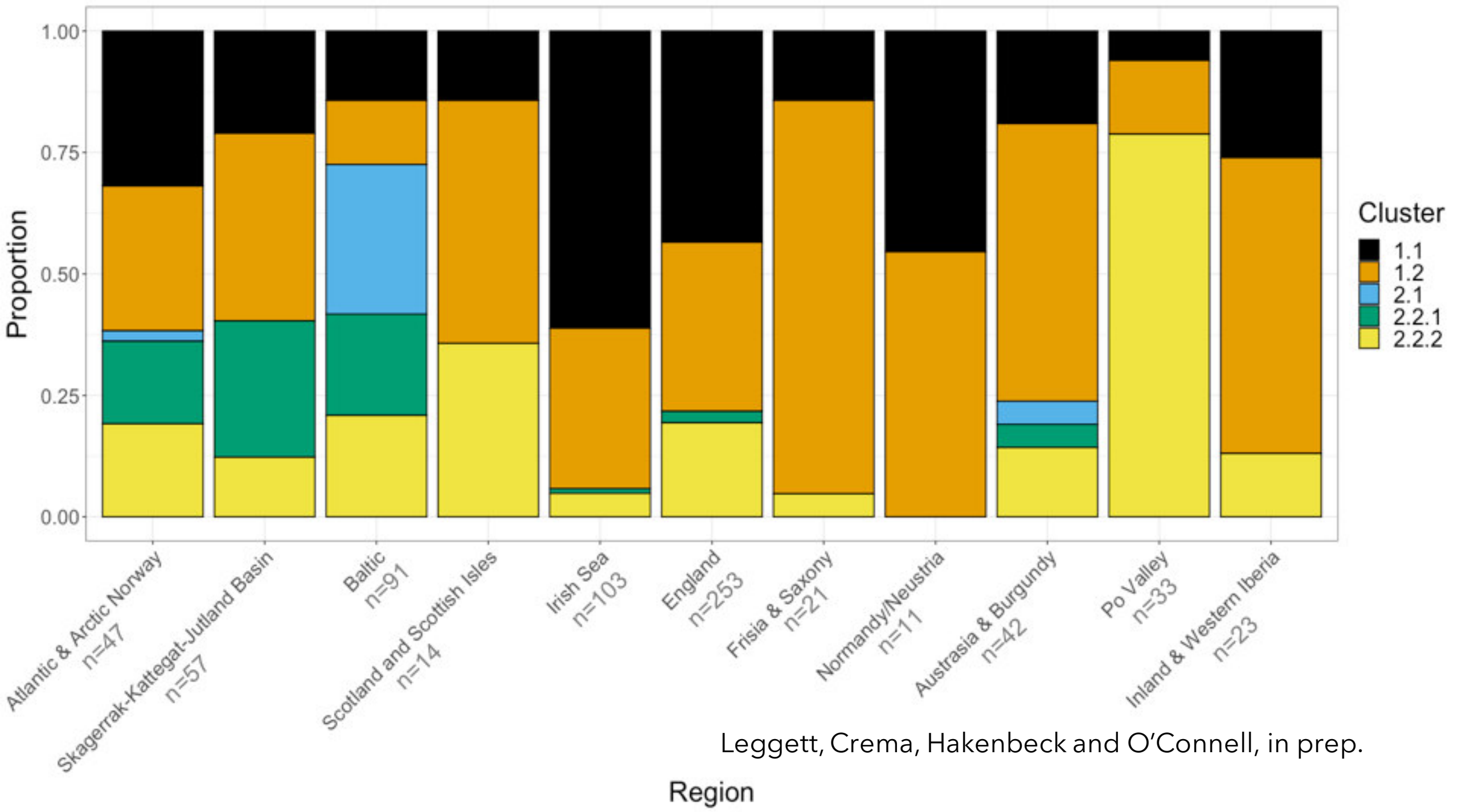




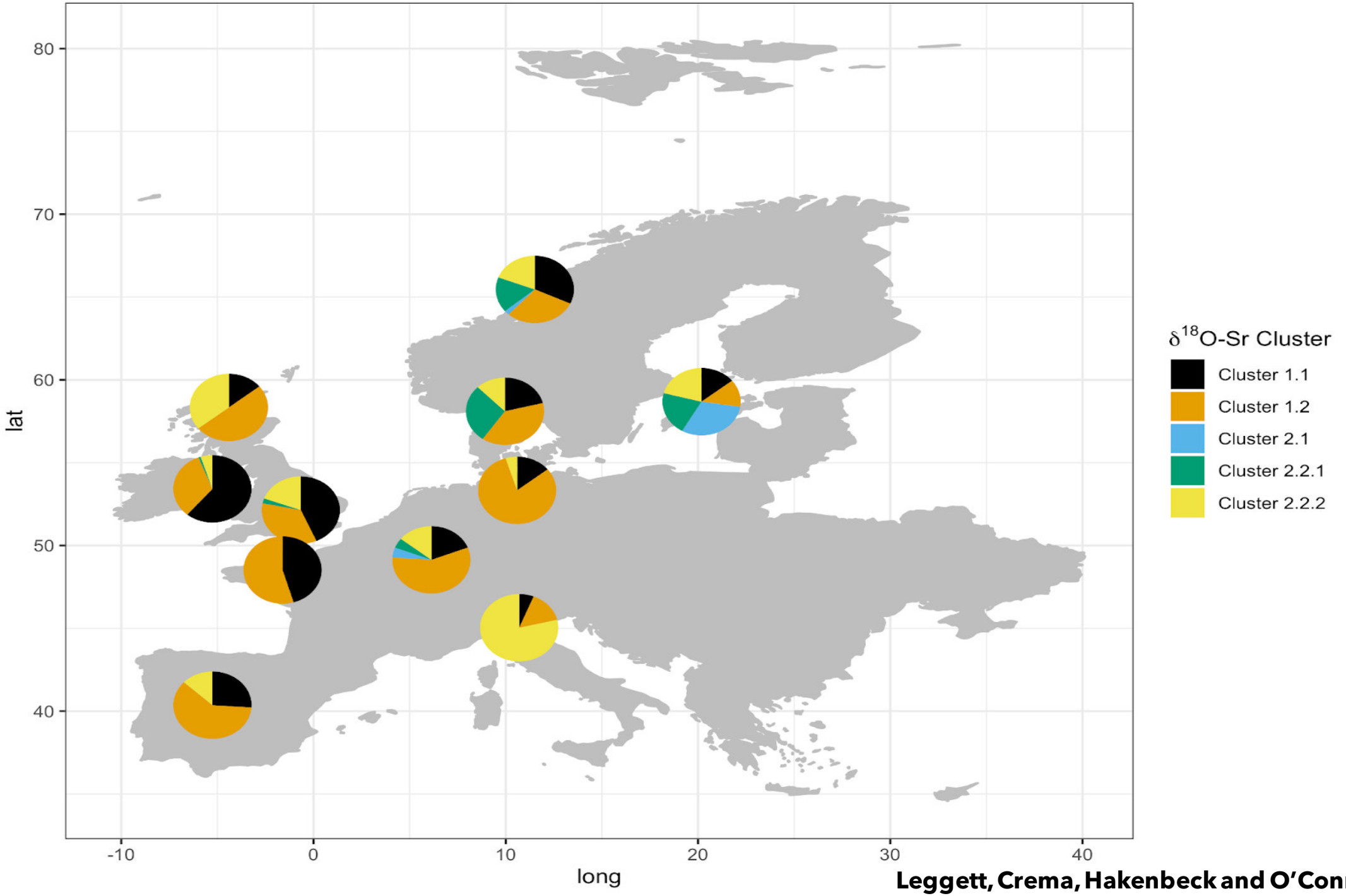


Mobility – Tooth Enamel

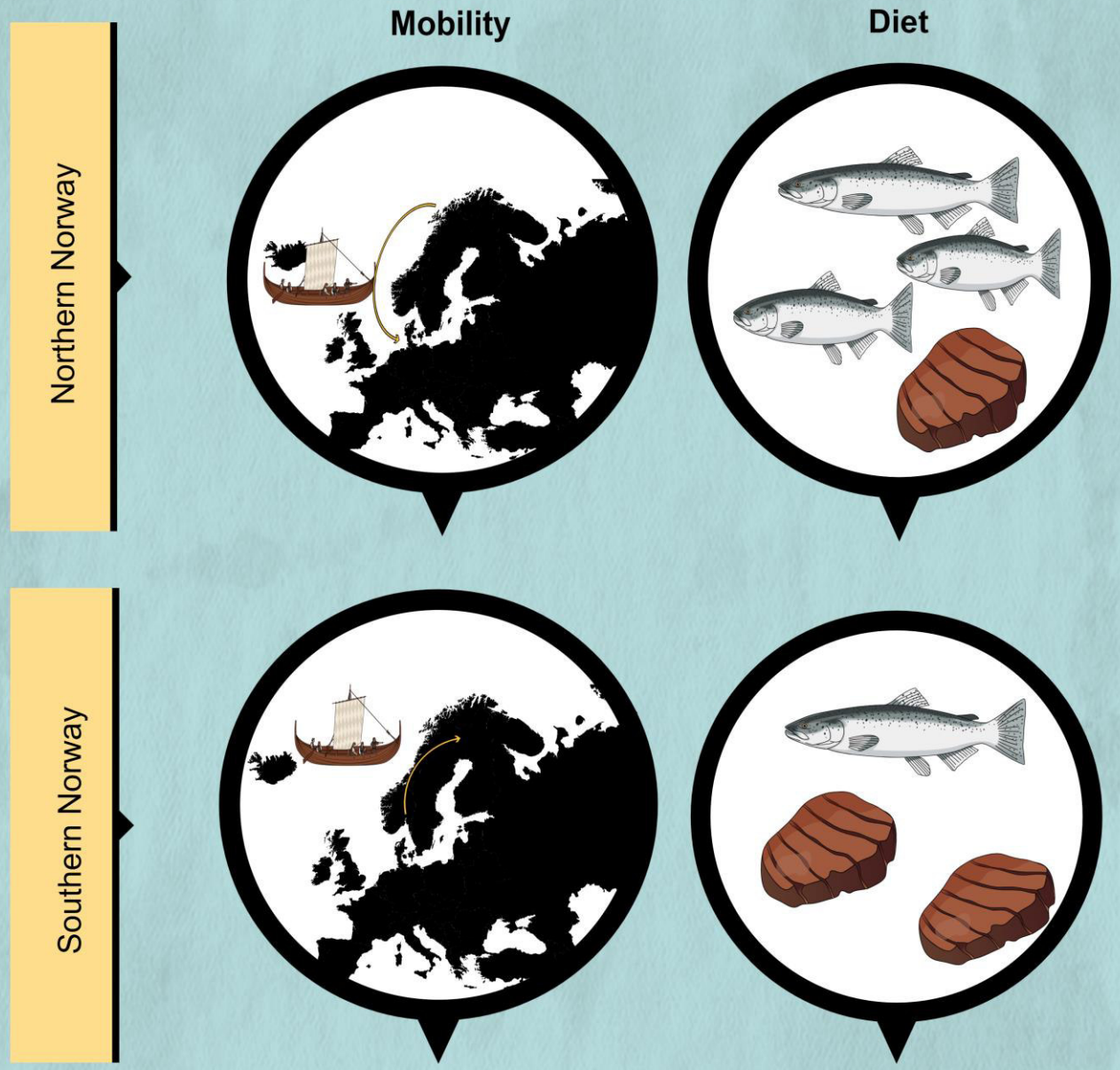

20.5
19.5
18.5
17.5
16.5
15.5
14.5
13.5
12.5
δ18Ophosphate (SMOW) (‰)
0.706
0.711
0.716
0.721
0.726
0.731
0.736
0.741
87Sr/86Sr
Cluster Number
1.1
1.2
2.1
2.2.1
2.2.2
Leggett, Crema, Hakenbeck and O'Connell, in prep.

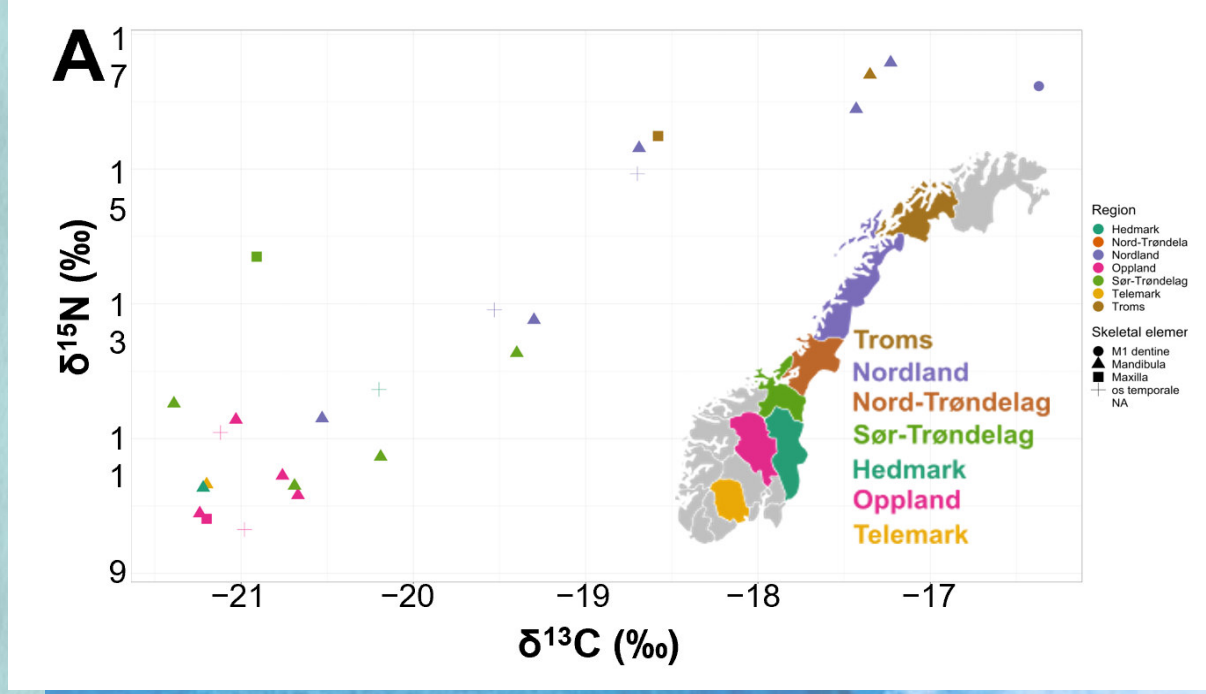


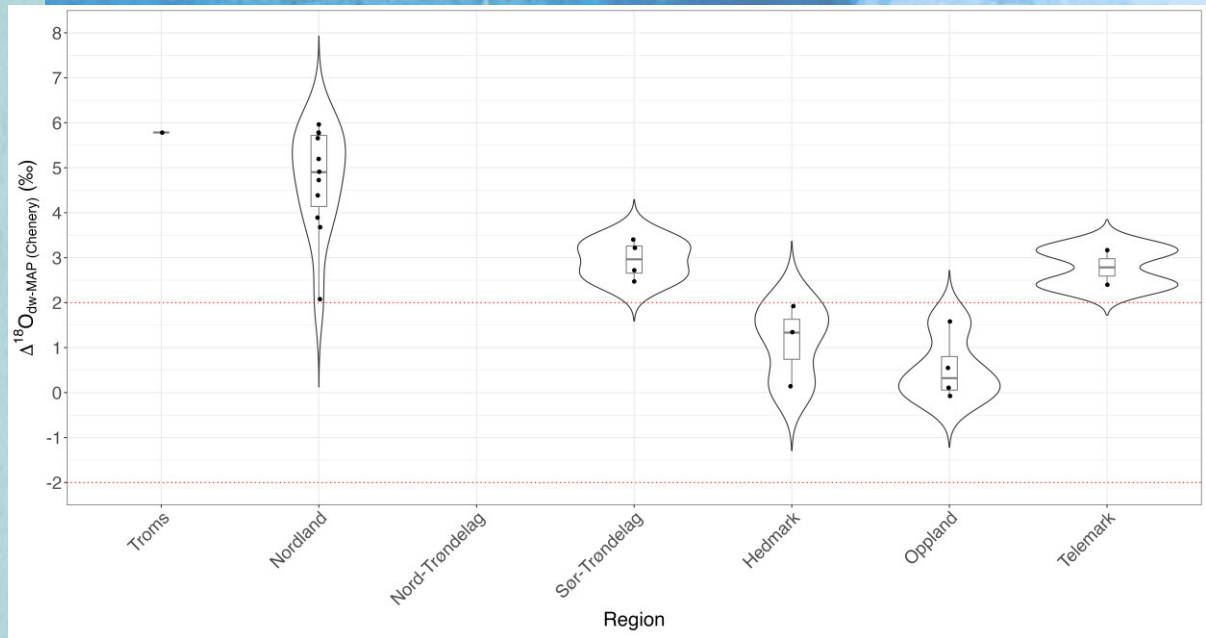
Leggett, Crema, Hakenbeck and O'Connell, in prep.



Leggett, Crema, Hakenbeck and O'Connell, in prep.

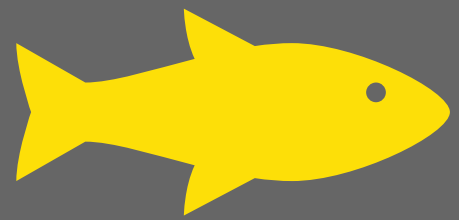






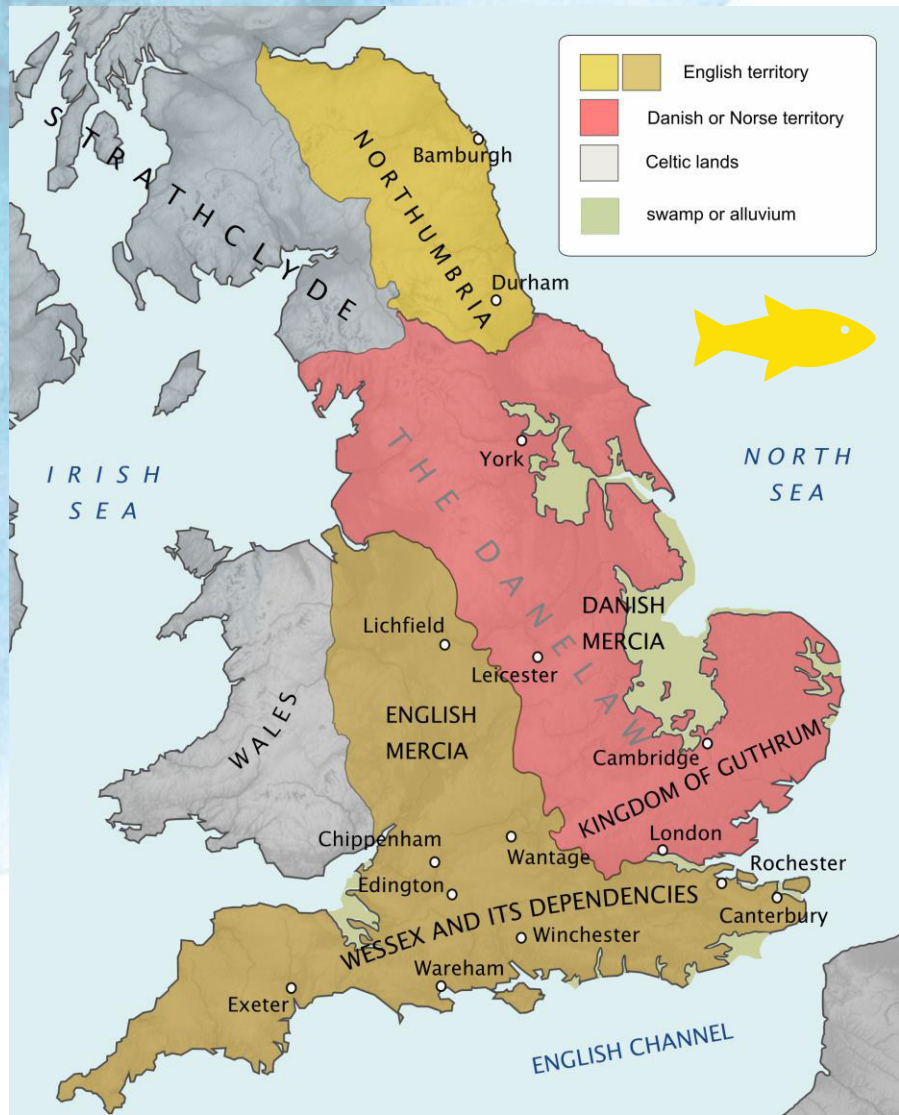
Strand, Lisa Mariann, Sam Leggett, and Birgitte Skar. 'Multi-Isotope Variation Reveals Social Complexity in Viking Age Norway'. *iScience*, September 2022, 105225. https://doi.org/10.1016/j.isci.2022.105225.

Fishy eaters in early medieval Britain – are they all (northern) Scandinavians?

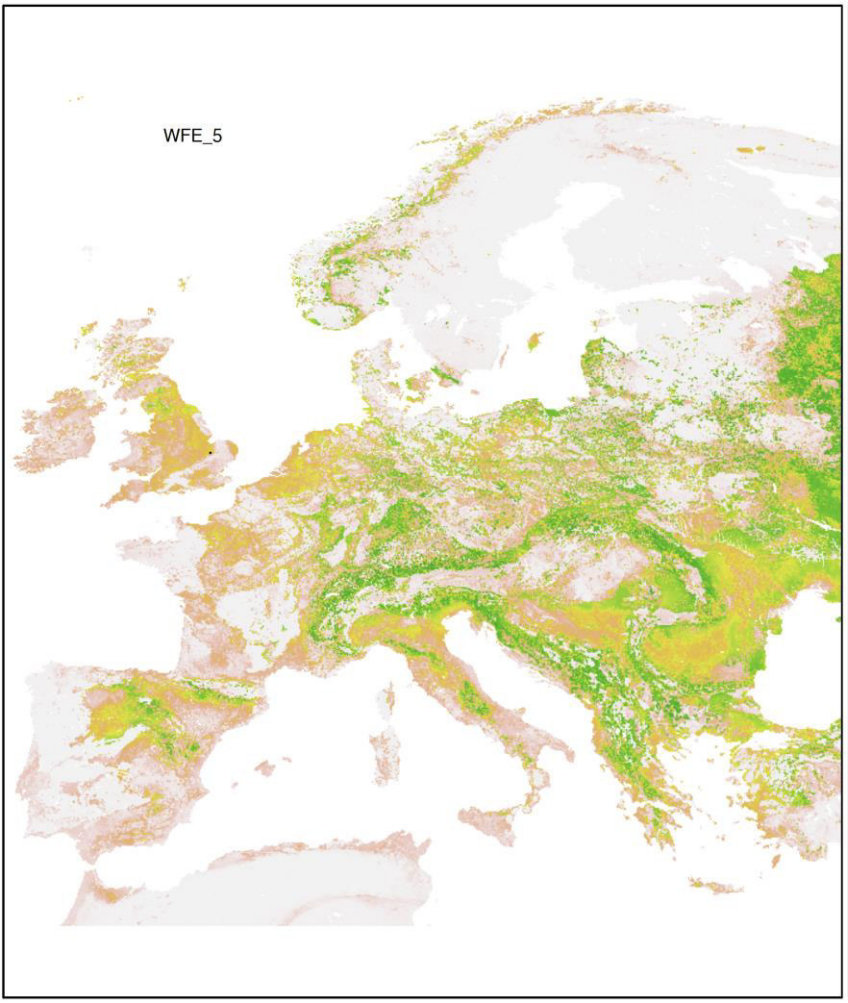


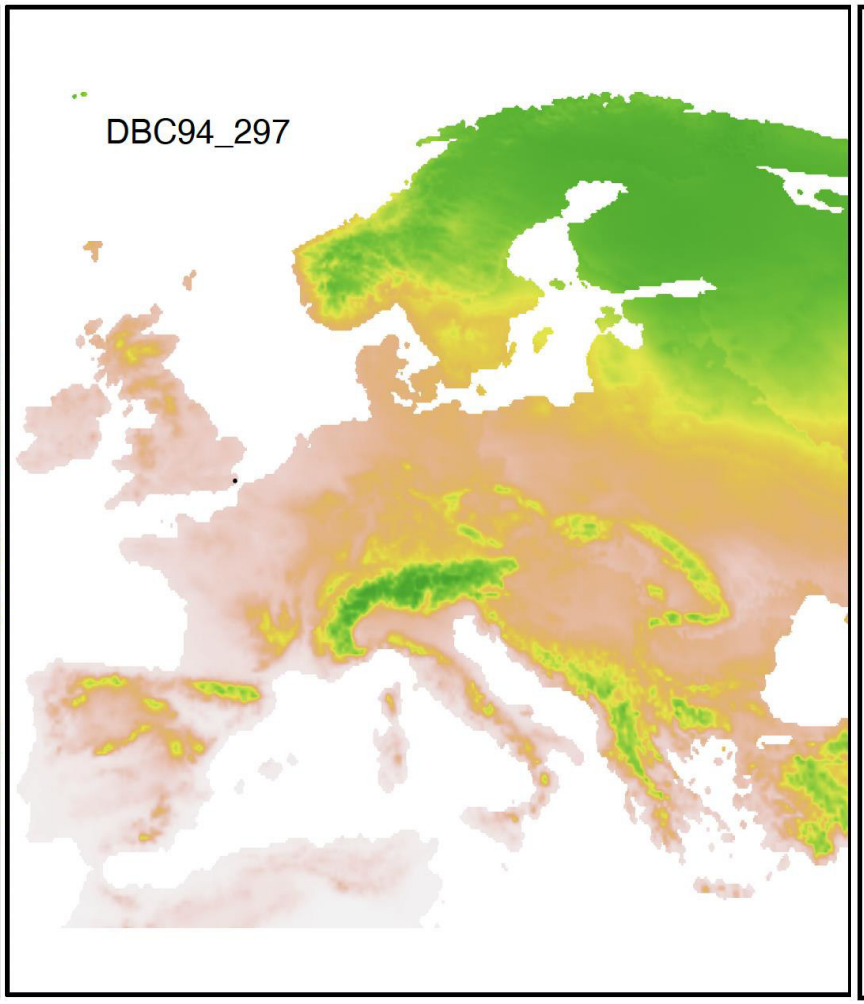
Possible Adulthood "Fish" eaters in England

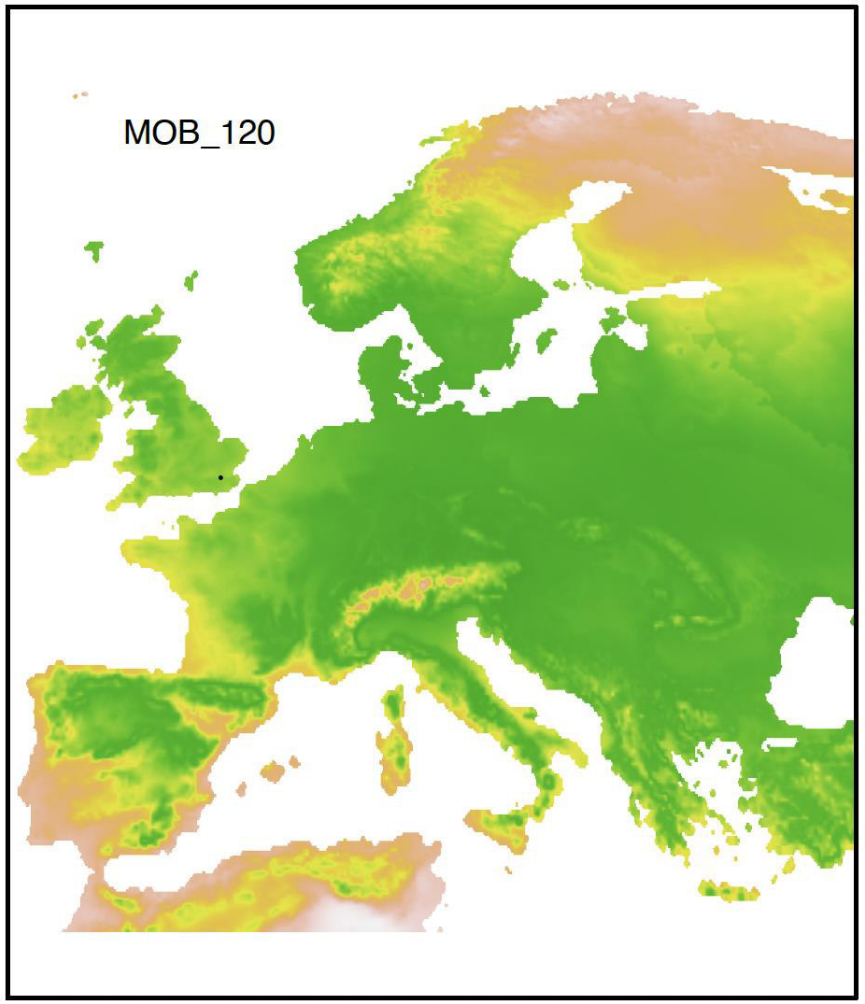
- **Barton upon Humber – 1 outlier**
- **York sites – Belle Vue, Fishergate**
- Berinsfield – 1 outlier
- Bishopstone – 1 outlier
- **Black Gate Newcastle**
- Buckland Dover – 1 outlier
- Burgh Castle
- Caister-by-Yarmouth/Caister-on-Sea
- **Norwich Castle cemeteries**
- Grange Road Sewerage Extension Cambridge
- Holborough – 1 outlier
- **Stoke Quay Ipswich**
- **Ketton Quarry**
- **Kirkdale**
- **Masham**
- Melbourn Water Lane – 1 outlier
- Mill Hill, Deal – 2 outliers
- Portway Andover – 1 outlier
- **Priory Orchard Godalming**
- **Raunds Furnells**
- **Repton**
- South Acre
- **St John's College Oxford**
- Wasperton – 1 outlier
- Westgarth Gardens
- Westfield Farm Ely
- **Yarnton – 2 outliers**



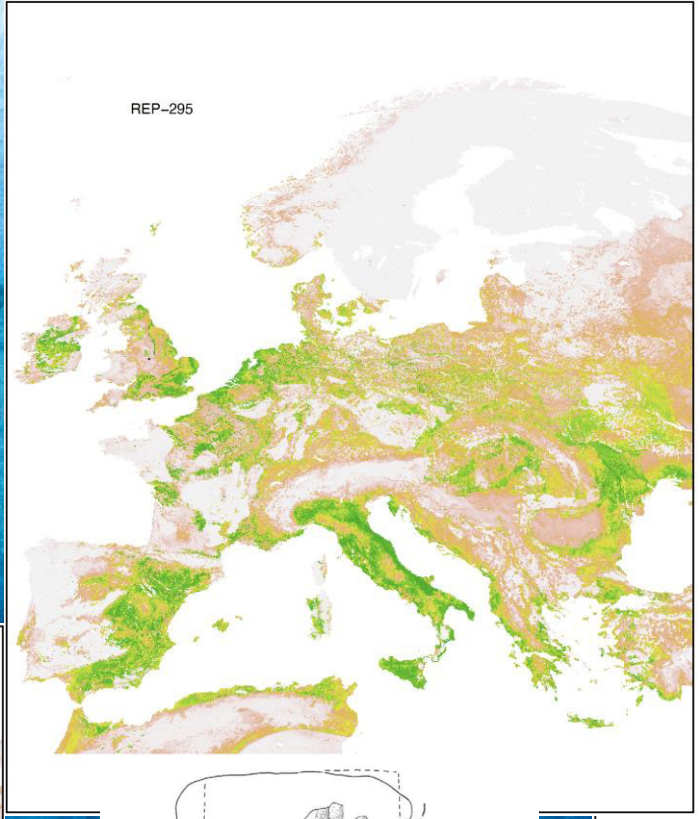
For more detail see: Leggett, Sam, and Tom Lambert. "Food and Power in Early Medieval England: A Lack of (Isotopic) Enrichment." *Anglo-Saxon England*, April 20, 2022, 1–33. https://doi.org/10.1017/S0263675122000072.

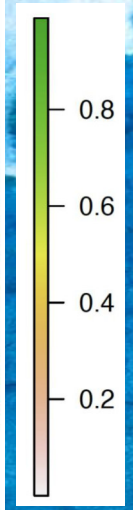

WFE_5


DBC94_297

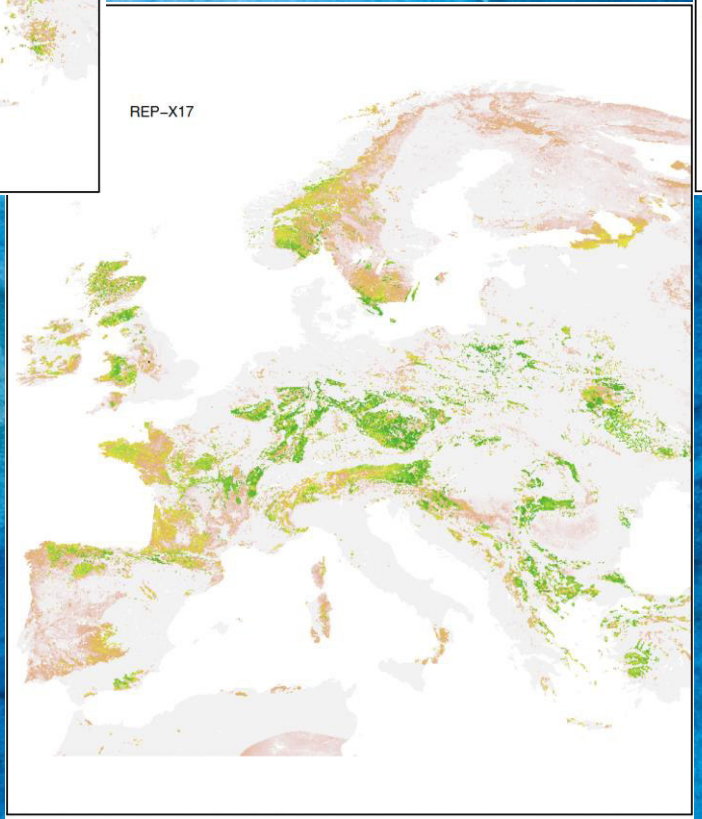

MOB_120

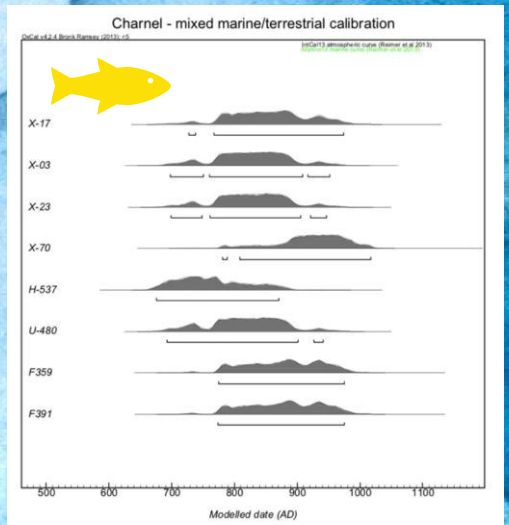

0.8
0.6
0.4
0.2



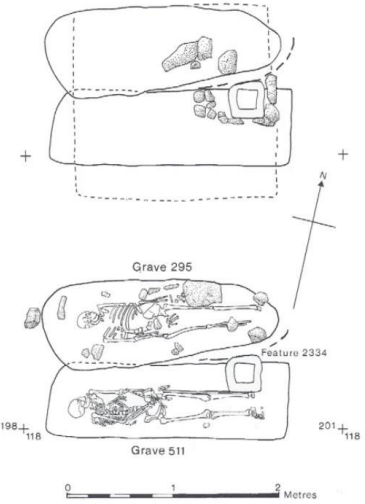


https://archaeology.co.uk/articles/features/resolving-repton.htm

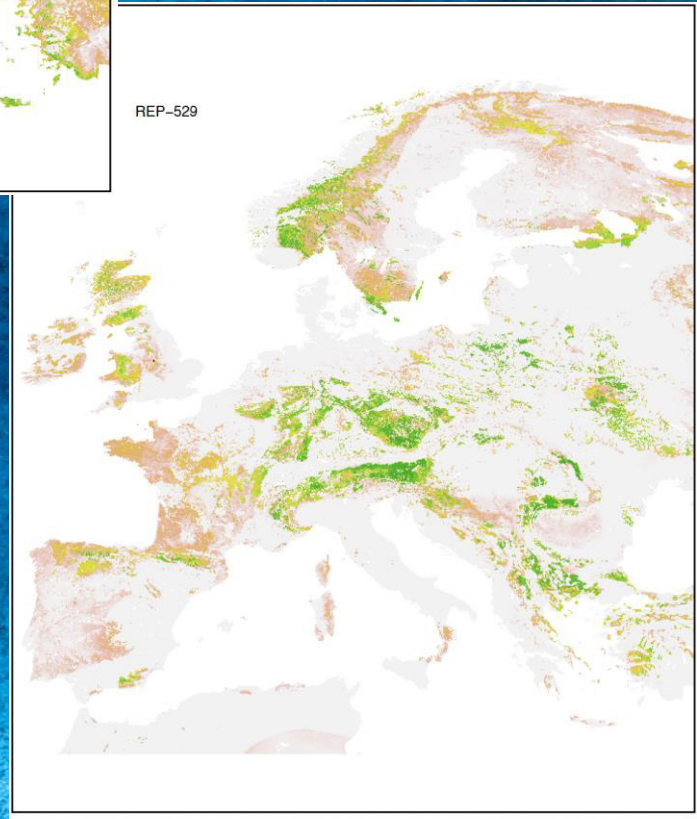




New radiocarbon dates from Repton have confirmed that the charnel dead are consistent with a 9th-century date, meaning that the human remains could have come from the Viking Great Army. [Image: Cat Jarman]



Recent aDNA analysis of G5ll and G295 has suggested that the two men may be father and son. [Image: © Martin Biddle/ Judith Dobie]



The Fish Event Horizon in Early Medieval Britain

- Clearly there in the zooarchaeological record, especially associated with urban and high-status sites
- The earliest pre-Viking "fishy" outliers are Scandinavian migrants
- In the lead up to the 11th century few coastal or riverine communities are consuming fish in isotopically visible amounts
- Seems to be a foodway predominantly associated with Scandinavian communities in the Danelaw in England
- To a lesser extent also clear in Christian communities – tricky to disentangle with Danelaw and local baseline impact in many cases (e.g. East Anglia)
- Scotland has a more striking shift – especially on Orkney (Barrett et al. 2001; Barrett 2003; Barrett and Richards 2004)
- Well established (isotopically) in human diets by the later medieval period (Müldner 2016; Rose 2021)
- Evidence for community assimilation (change from 'fishy' childhood to more terrestrial adulthood)

Stats/Data viz take aways

- Think about the questions you're asking or the story you want to tell and analyze/visualize accordingly
- So much more you can say without p-values using EDA
- Consider the shape/distribution of the data and not just dive into look for outliers or compare means

Some key readings

- Calin-Jageman, Robert J., and Geoff Cumming. 'The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known'. *The American Statistician* 73, no. sup1 (29 March 2019): 271–80. https://doi.org/10.1080/00031305.2018.1518266.
- Kassambara, Alboukadel. *Machine Learning Essentials: Practical Guide in R*. Frankreich: CreateSpace Independent Publishing Platform, 2018.
- Kaufman, Leonard, and Peter Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Förlag: Wiley, 2009.
- Kruschke, John. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. Boston: Academic Press, 2014.
- Kruschke, John K., and Torrin M. Liddell. 'The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective'. *Psychonomic Bulletin & Review* 25, no. 1 (1 February 2018): 178–206. https://doi.org/10.3758/s13423-016-1221-4.
- Lightfoot, Emma, and Tamsin C. O'Connell. "On the Use of Biomineral Oxygen Isotope Data to Identify Human Migrants in the Archaeological Record: Intra-Sample Variation, Statistical Methods and Geographical Considerations." *PLOS ONE* 11, no. 4 (April 28, 2016): e0153850. https://doi.org/10.1371/journal.pone.0153850.
- Tukey, John W. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioural Science : Quantitative Methods. Reading, Mass.: Addison-Wesley, 1977.
- Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA Statement on P-Values: Context, Process, and Purpose." *The American Statistician* 70, no. 2 (April 2, 2016): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

National Museums Scotland

University of BRISTOL

HISTORIC ENVIRONMENT SCOTLAND | ÀRAINNEACHD EACHDRAIDHEIL ALBA

AOC Archaeology Group

LEVERHULME TRUST

THE UNIVERSITY of EDINBURGH

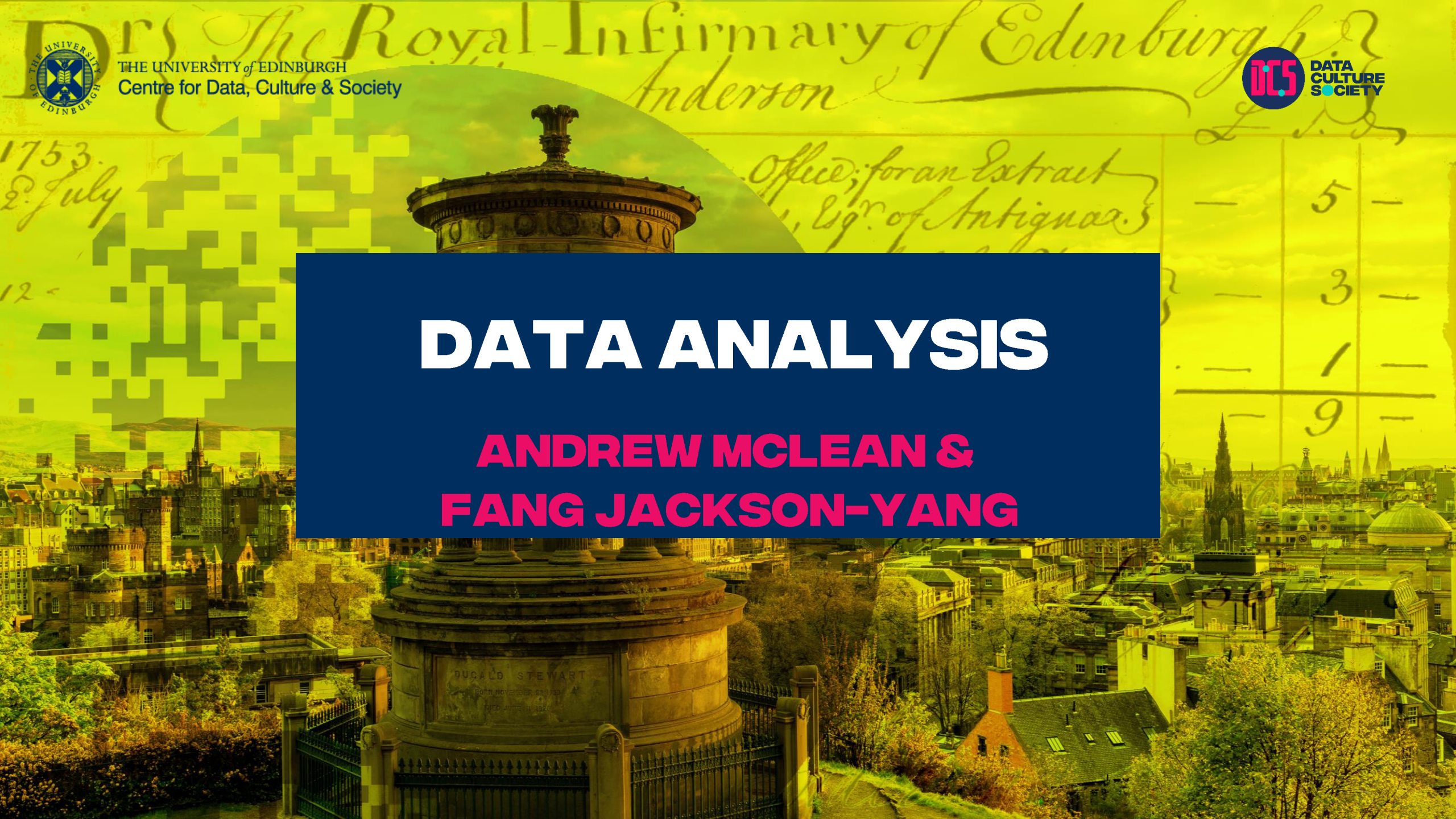
SUERC
Scottish Universities Environmental Research Centre

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

COFFEE BEAK

WE ARE GOING TO RESTART AT 11:00



DATA ANALYSIS

ANDREW MCLEAN &
FANG JACKSON-YANG

OUTLINE

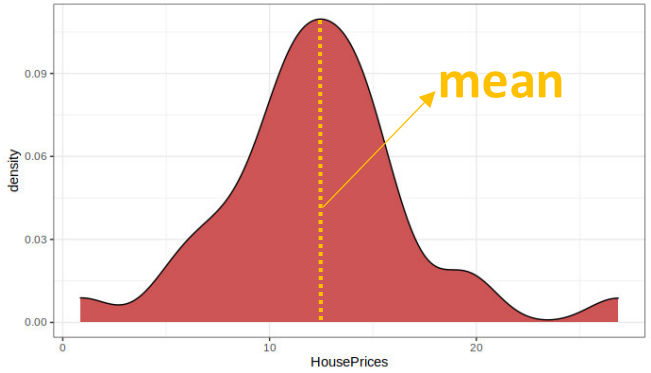
- Descriptive Statistics
- Inferential Statistics: Linear Regression
- Inferential Statistics: Null Hypothesis Testing

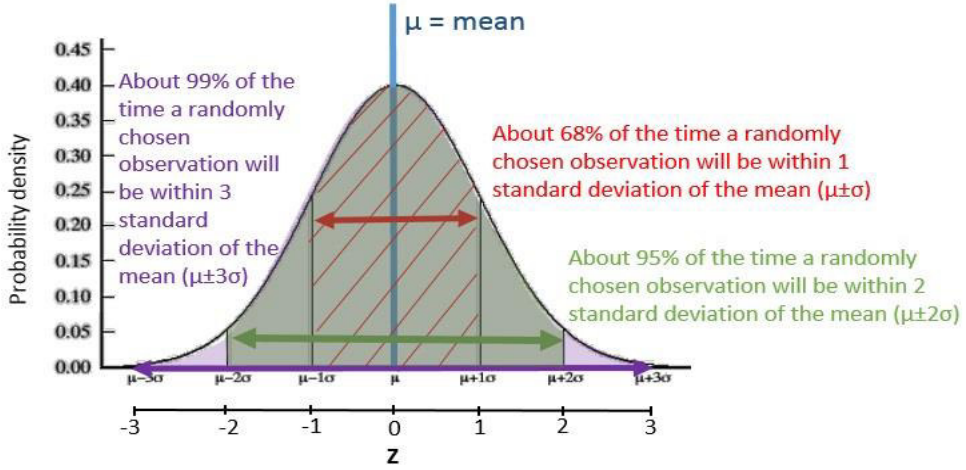
STATISTICS

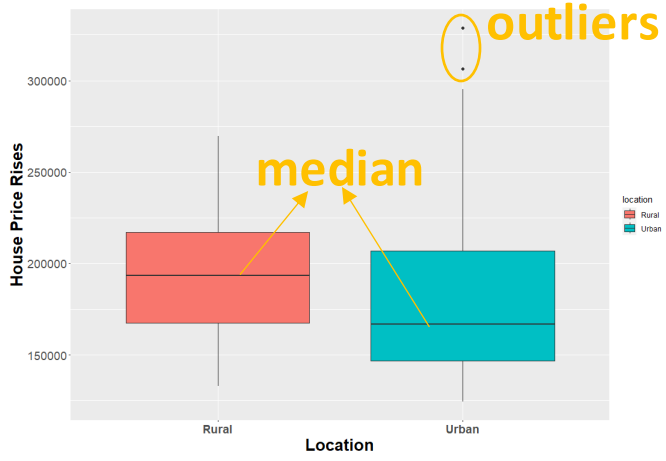
- Two main types: Descriptive & Inferential.
- Descriptive statistics are used to summarise data.
- Inferential statistics are used to make inferences (and predictions) about the world based on the observed data.

DESCRIPTIVE STATISTICS

- Distribution & central tendency
- Mean, median, standard deviation (sd)





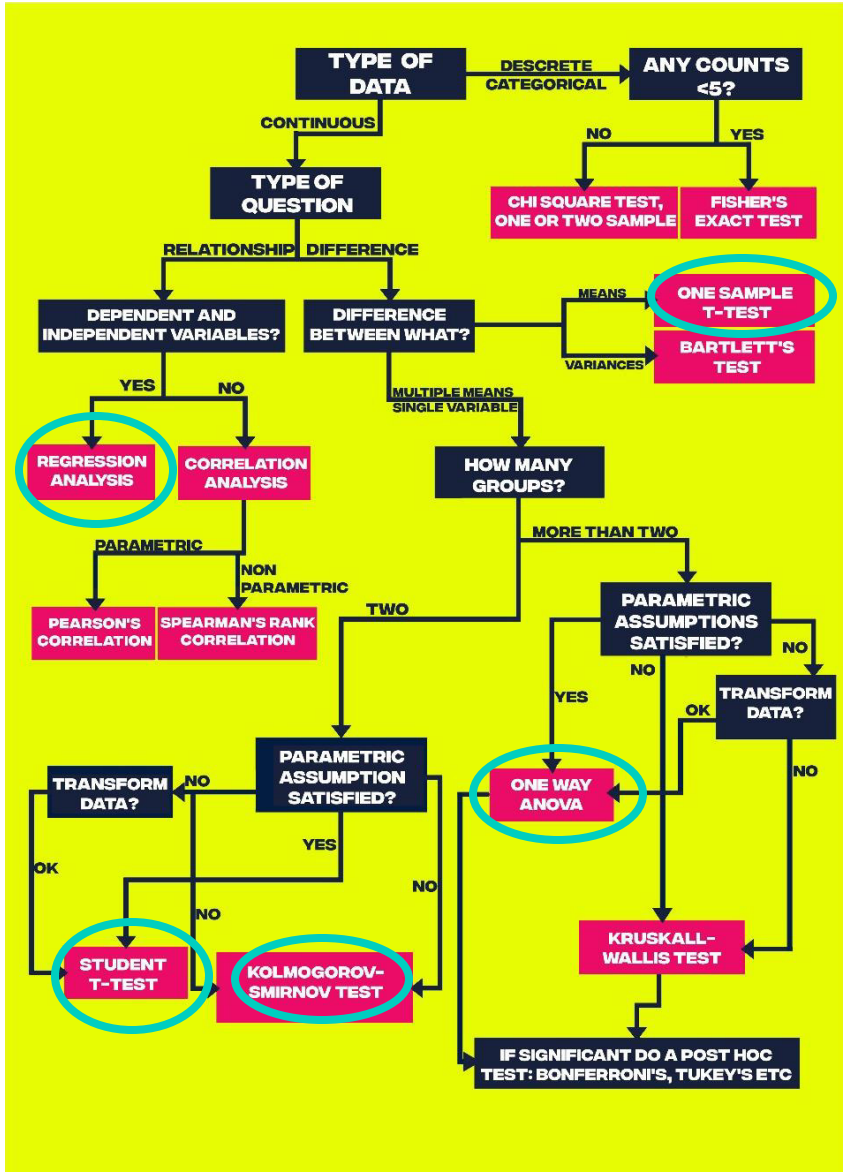


Resource: http://dev1.ed-projects.nyu.edu/statistics/normal-distribution/

INFERENTIAL STATISTICS

LOTS OF CHOICES TO BE MADE

- Nature of the research question
- Nature of the data
- Objectives



LINEAR REGRESSION

- **Linear Model:** $Y \sim \beta_0 + \beta_1 X + \varepsilon$ (in R: **lm(Y ~ X, data)**)

- **Model Interpretation**

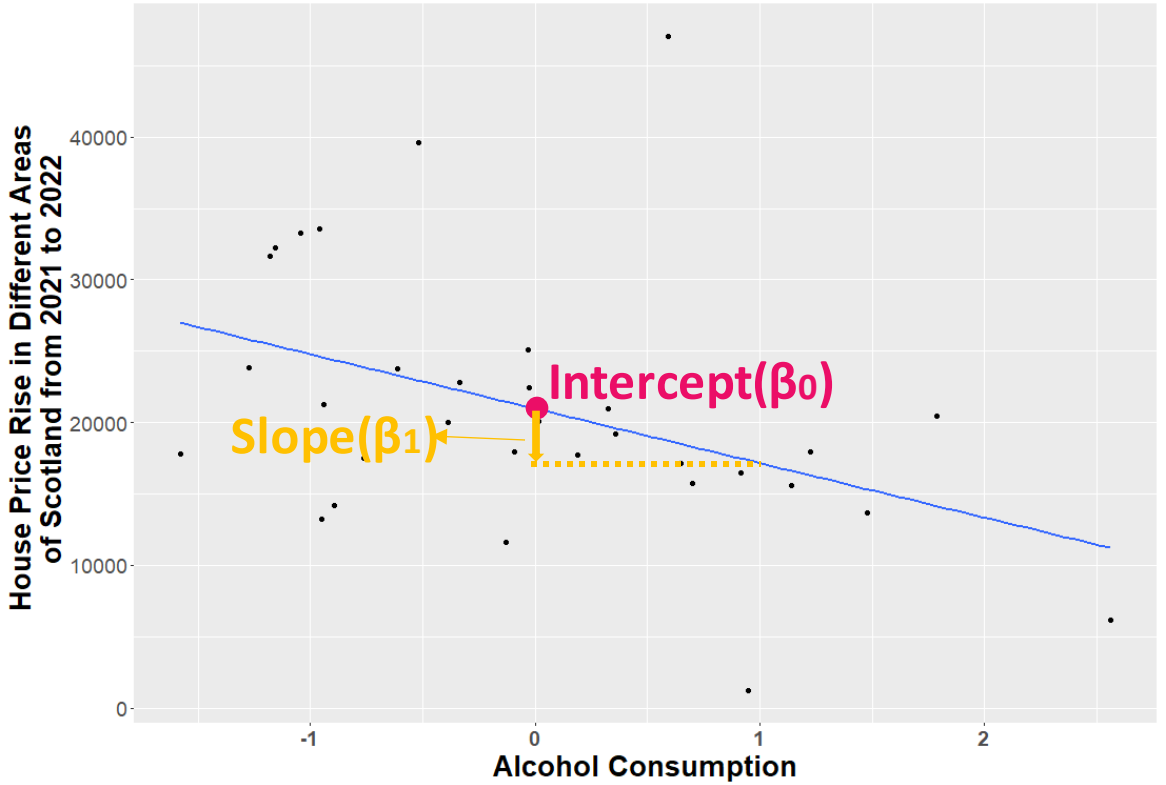
- **Intercept**: the expected value of y when x is 0.
- **Slope**: the rate of change (number of units by which y increases as x increases by 1 unit).

- **Model evaluation**

- How much variance does the model account for?
- How much does the model fit improve over chance?

- **Model comparison**

- Does model fit improve if we add an additional independent variable? (in R: **anova(**m1, m2**)**)



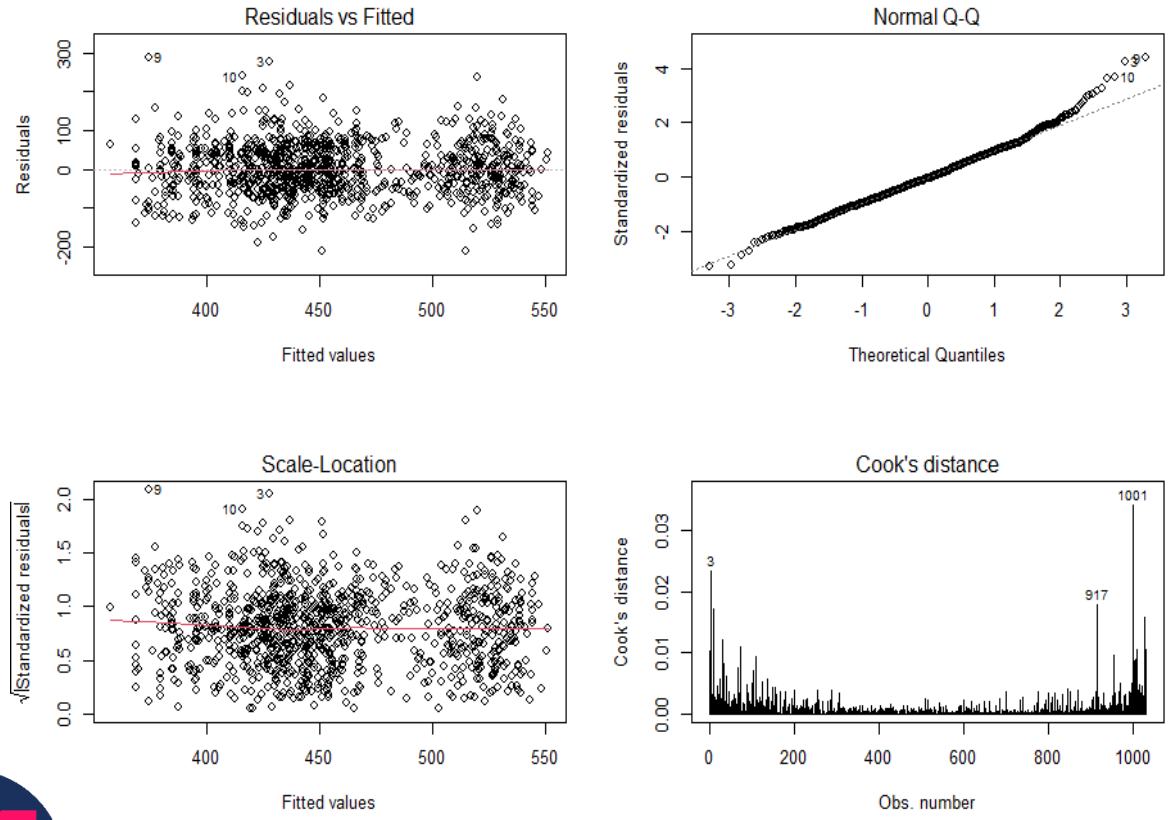
ASSUMPTION CHECK

FOR LINEAR REGRESSION MODELS:

- Linearity of relationships
- Independence of residuals
- Normality of residuals
- Equal variances for residuals
- Ideally no collinearity & no outliers

FOR T-TESTs:

- Normally distributed data

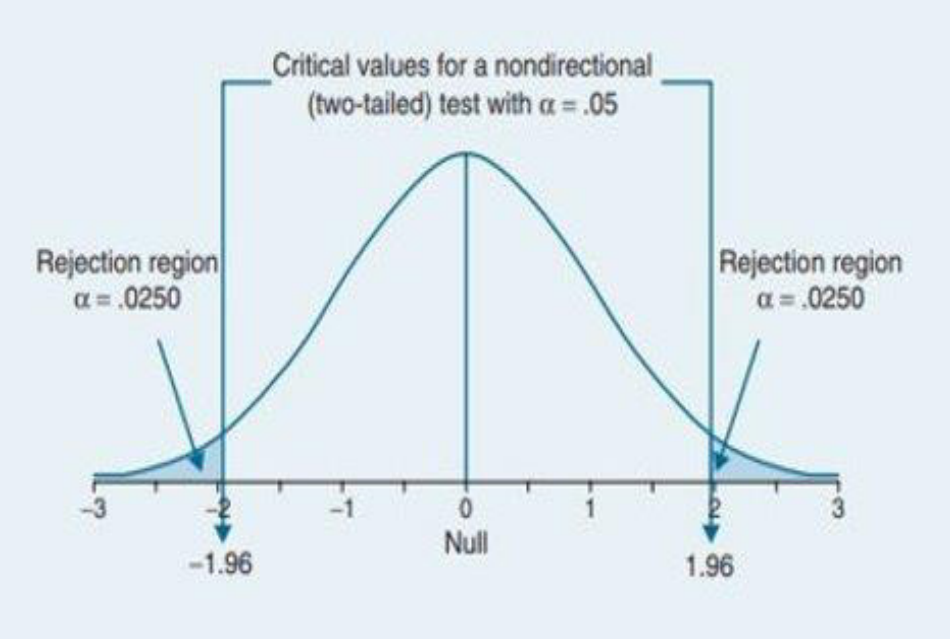


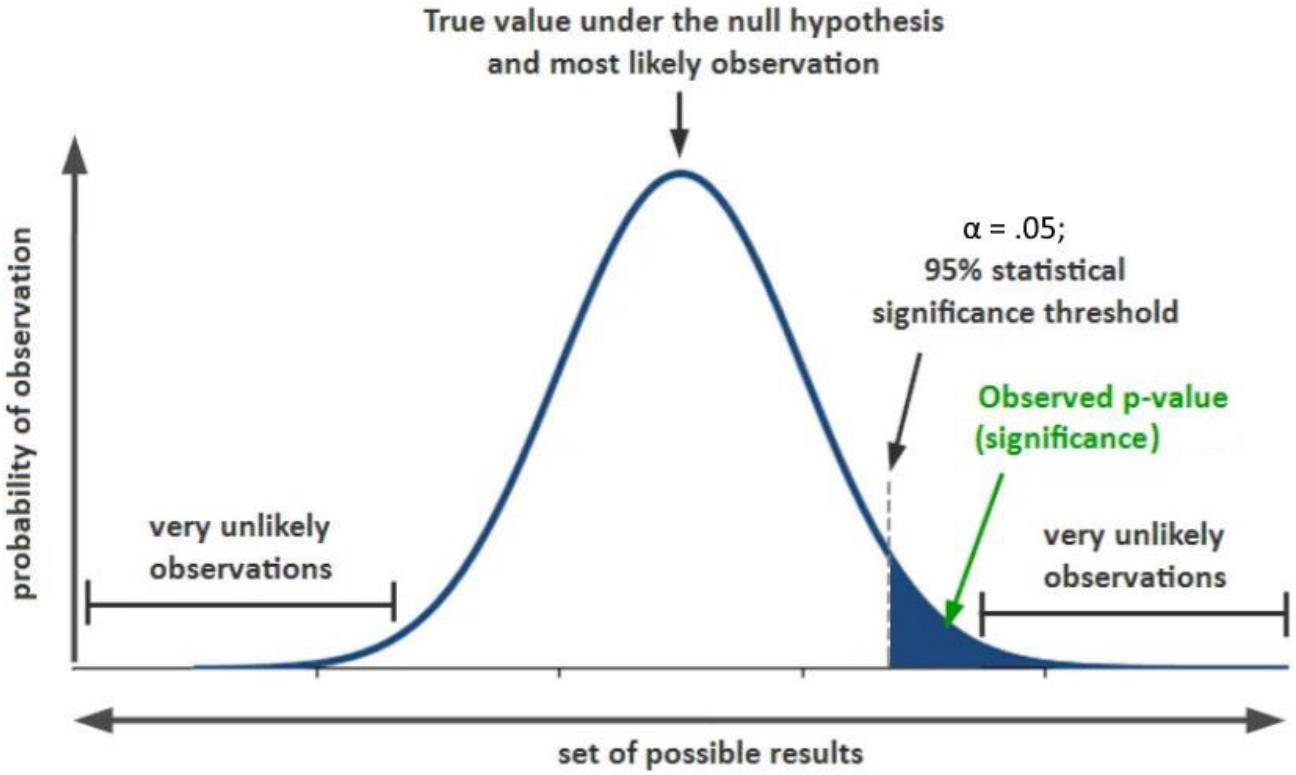
NULL HYPOTHESIS TESTING (NHT)

- Hypotheses are claims researchers make about the world based on some data.
- NHT is a formal method for validating hypotheses, by comparing a theoretically non-null hypothesis (H_1) with a null hypothesis (H_0).
  - H_1: House price rises associated with alcohol consumption.
  - H_0: House price rises not associated with alcohol consumption.
- The goal is to accept or reject **H0** (not H_1) and then make inferences about H1.
  - H_1: Pandas are black and white.
  - H_0: Some pandas are not.

NULL HYPOTHESIS TESTING (NHT)

- Goal: Reject or accept H_0





Resources: https://medium.com/datadriveninvestor/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e; https://www.sagepub.com/sites/default/files/upm-binaries/40007_Chapter8.pdf; https://learningstatisticswithr.com/lsr-0.6.pdf

NULL HYPOTHESIS TESTING (NHT)

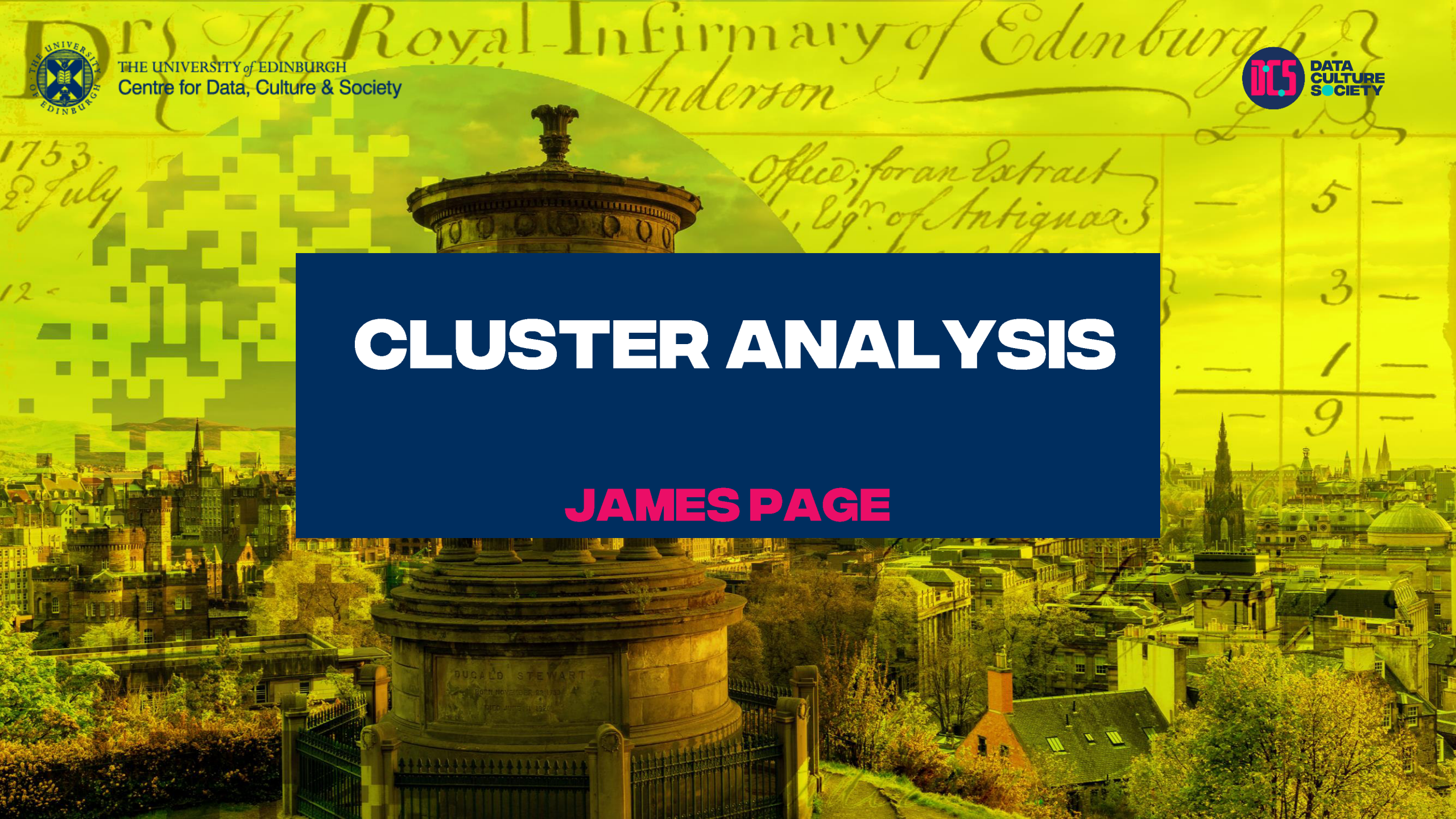
- Two types of errors in decision making

		Decision	
		Retain H_0	Reject H_0
H_0 Truth	True	1 - α = probability of correct retention	α (type I error rate)
	False	β (type II error rate)	1 - β = power of the test

Resources: https://medium.com/datadriveninvestor/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e; https://www.sagepub.com/sites/default/files/upm-binaries/40007_Chapter8.pdf; https://learningstatisticswithr.com/lsr-0.6.pdf


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TIME FOR R


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
LUNCH BEAK
WE ARE GOING TO RESTART AT 13:30


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
CLUSTER ANALYSIS
JAMES PAGE

WHAT IS CLUSTER ANALYSIS

- Cluster analysis is a way of dividing objects in a dataset into groups made up of similar characteristics.
- Many different types of cluster analysis. We will only cover a few.
- Have urban and rural local authorities been similarly impacted by the cost of living crisis?

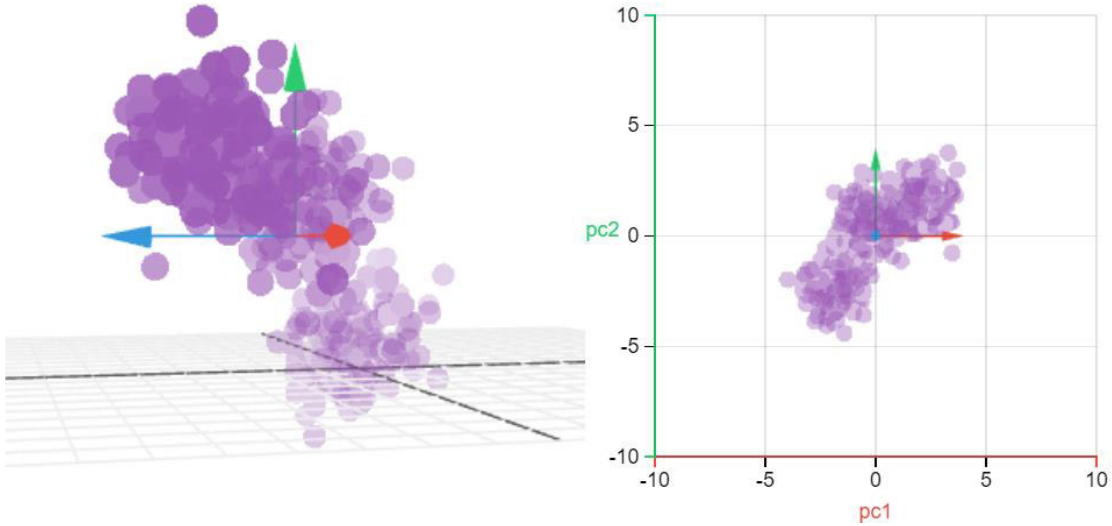
PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is a technique used to emphasize variation and bring outline patterns in a dataset

When to Apply?

- 3 or more continuous variables
- 1 or more categorical variable (not mandatory)

When there are more than two variables, it can be difficult to isolate underlying patterns in a dataset because it's hard to see through a cloud of data.



HOW IT WORKS

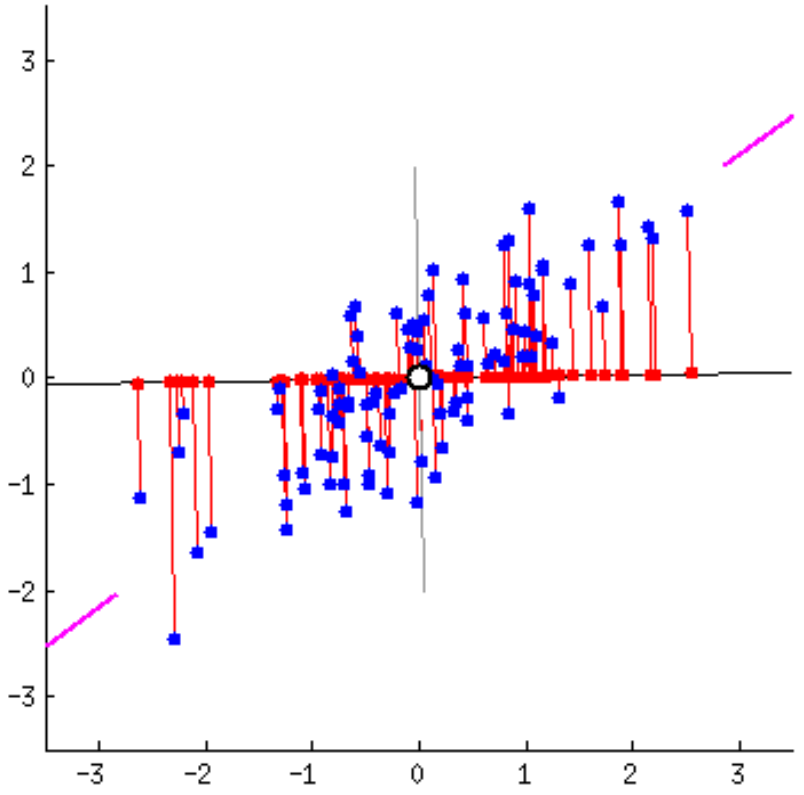
PCA can be used to find out which of our variables cause the most variation in our dataset

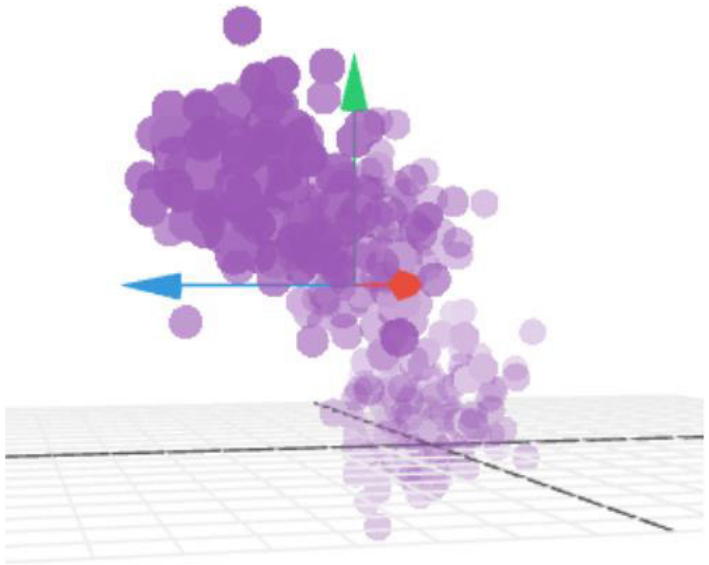
- Theoretically, variables with high variation will have a higher chance of creating patterns when plotted.
- These Principal Components (PCs) are ranked.
  - PC1 contains the highest variation.
  - PC2 the second, PC3 the third etc.

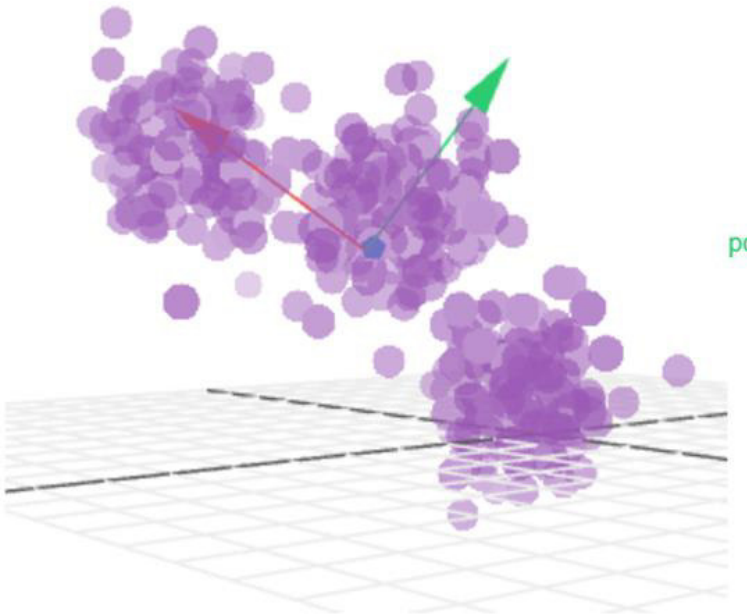
HOW IT WORKS

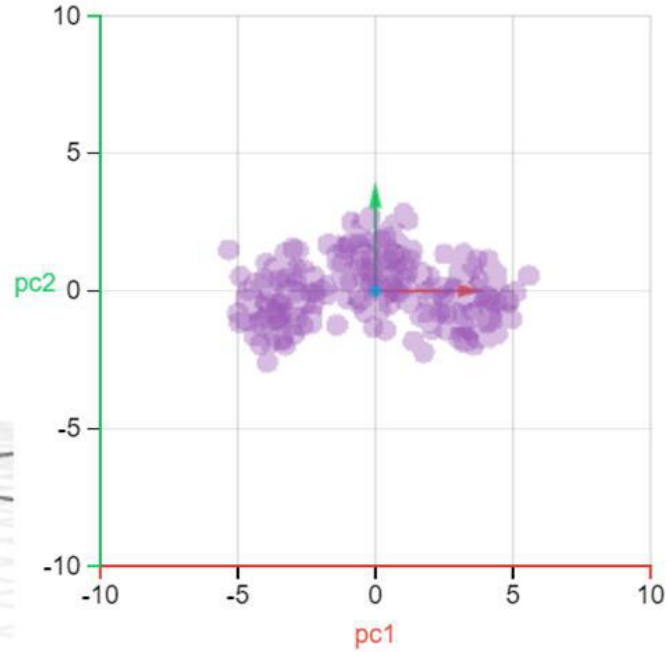
As most variation are captured by PC1 and PC2 so we can use them to replot our sample in two dimensions, discarding the other PCs.

- To do this, PCA generates a new coordinate system in which every point has a new (x,y) value in relation to their distance from the lines for PC1 and PC2.
- The axes don't actually mean anything physical; they're combinations of the original variables.







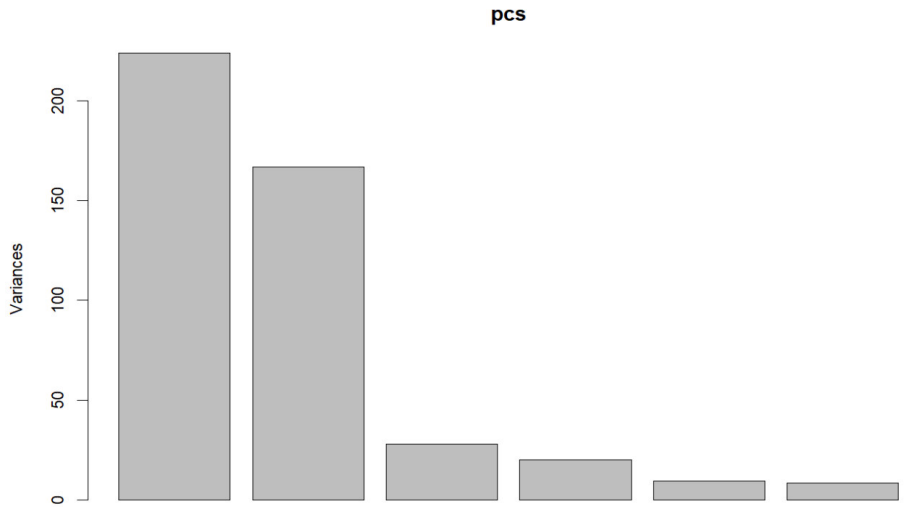

10
5
pc2
0
-5
-10
-10
-5
0
5
10
pc1

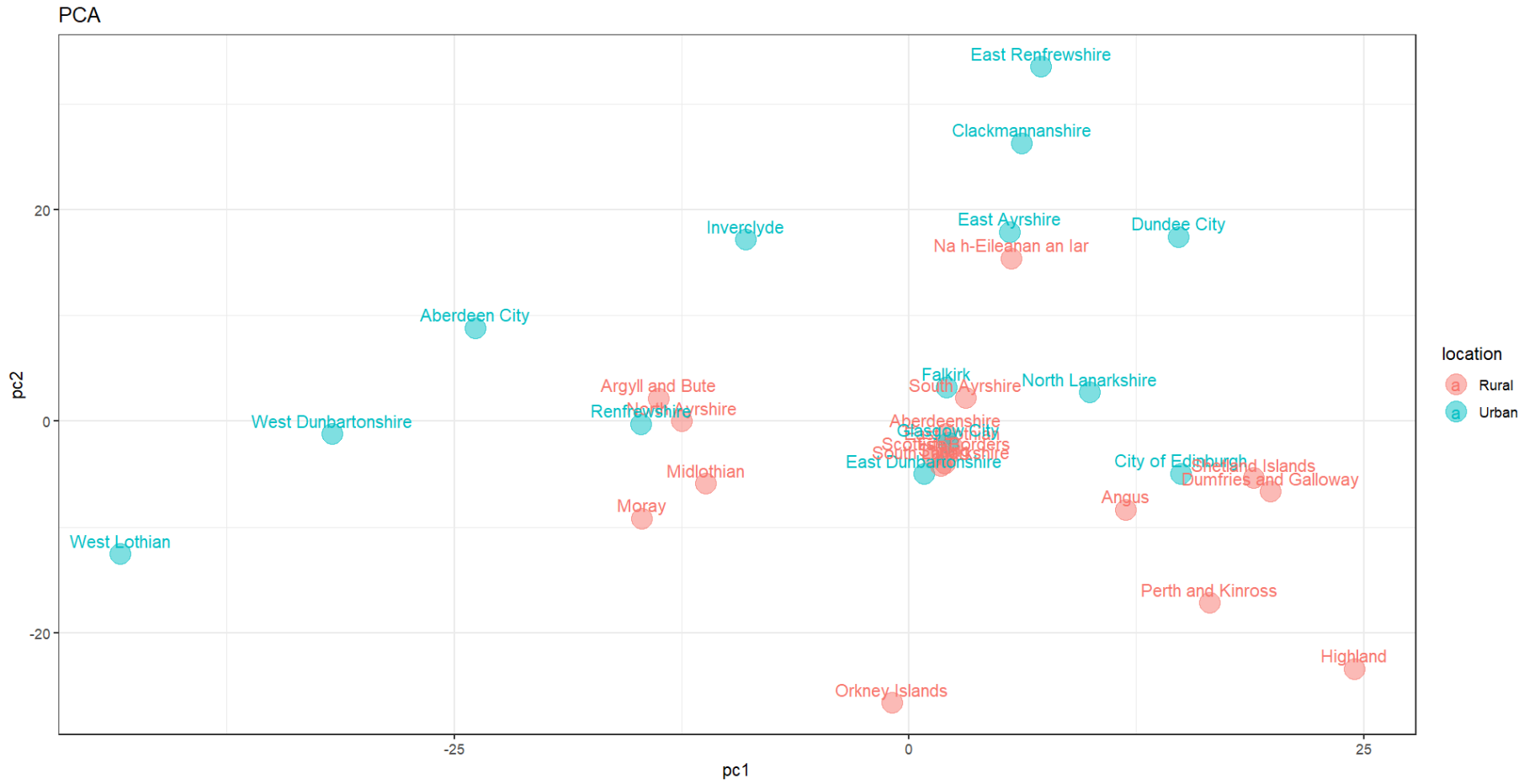

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TIME FOR R

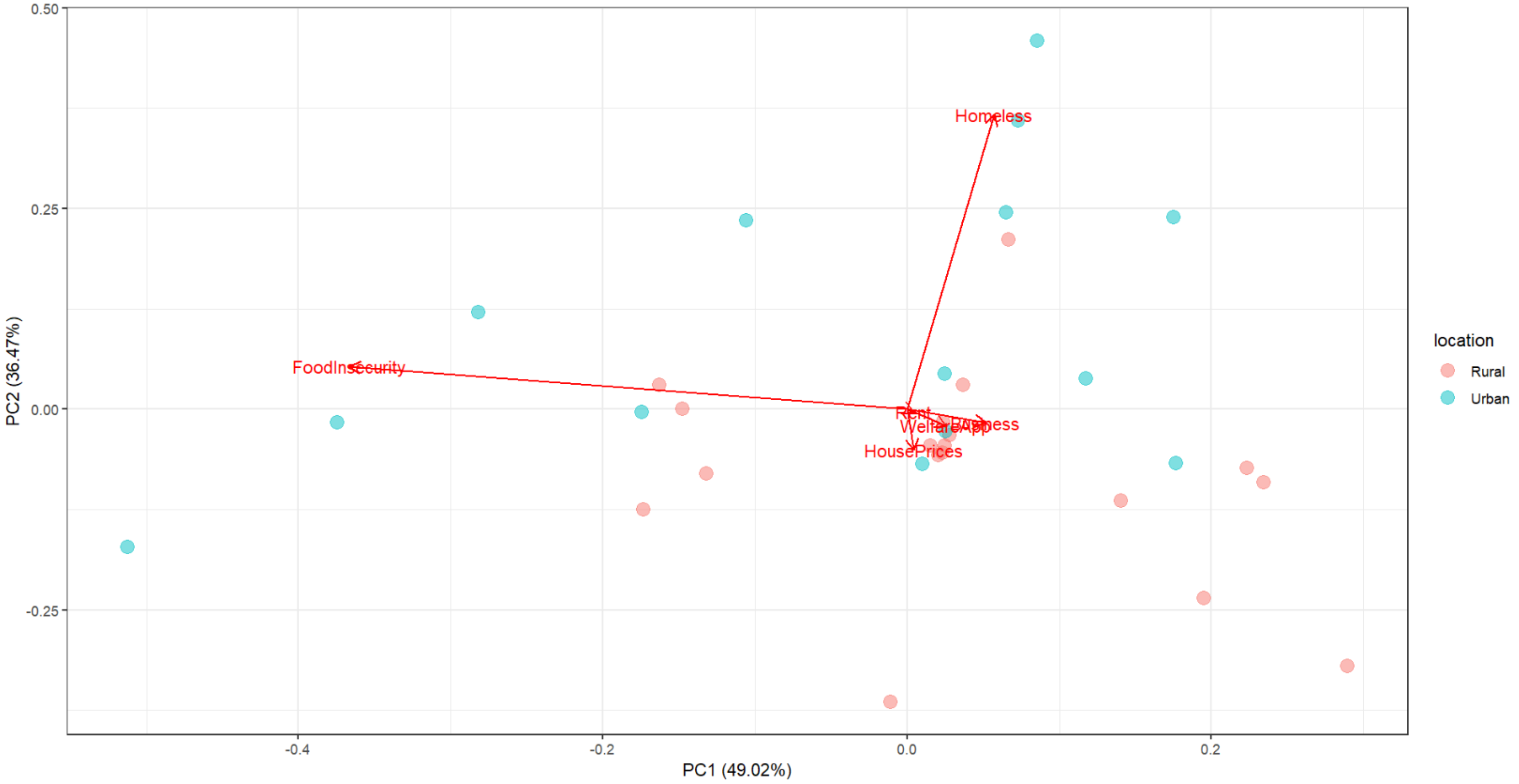
PC1 and PC2 contain the most variation, making *FoodInsecurity* and *Homeless* are the most relevant variables to explain sample variation.
As PC3 – PC6 contain little variation, we'll drop them from the visualisation.

	PC1	PC2	PC3	PC4	PC5
FoodInsecurity	-0.97612329	0.141440711	-0.09352987	0.003465612	-0.09400554
HousePrices	0.01189536	-0.134901830	-0.04923299	0.953871554	-0.26264953
WelfareApp	0.06787445	-0.053996720	-0.97628024	-0.004114028	0.19795540
Homeless	0.15291102	0.978165860	-0.05193203	0.123025965	-0.04033791
Business	0.13757367	-0.044611235	-0.13032841	-0.207233200	-0.74394656
Rent	0.01125115	-0.009042371	0.12658096	0.178938917	0.57263271

	PC6
FoodInsecurity	0.09788239
HousePrices	-0.01950371
WelfareApp	0.01228988
Homeless	0.01871627
Business	0.60472783
Rent	0.78983641

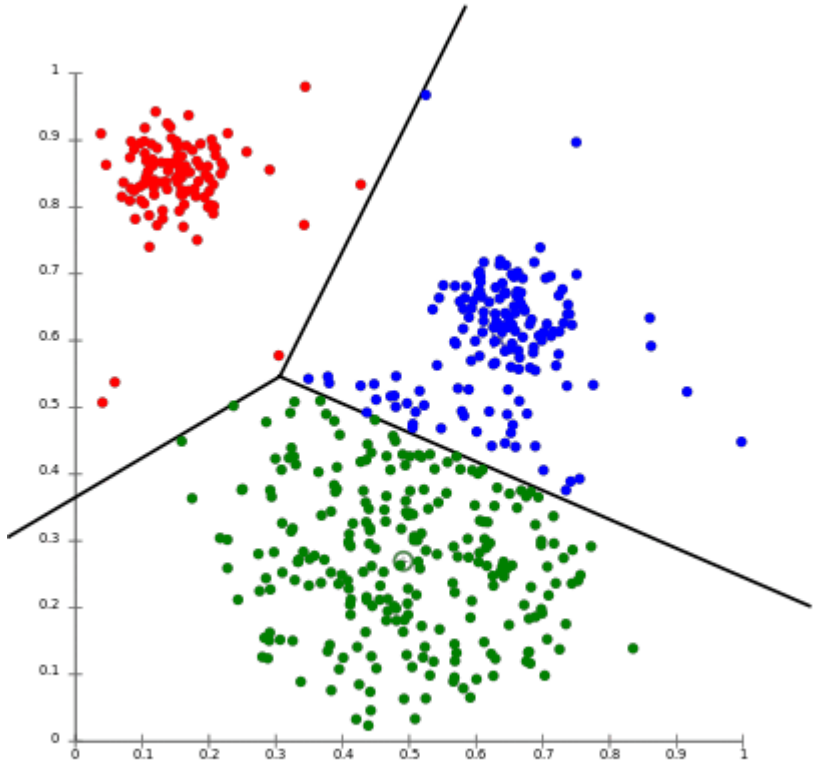



PCA
East Renfrewshire
Clackmannanshire
Inverclyde
East Ayrshire
Dundee City
Na h-Eileanan an Iar
Aberdeen City
Falkirk
South Ayrshire
North Lanarkshire
Argyll and Bute
Renfrewshire
West Dunbartonshire
Aberdeenshire
East Dunbartonshire
City of Edinburgh
Shetland Islands
Dumfries and Galloway
Midlothian
Angus
Moray
West Lothian
Perth and Kinross
Highland
Orkney Islands
pc2
pc1
20
0
-20
-25
0
25
location
Rural
Urban


0.50
0.25
0.00
-0.25
PC2 (36.47%)
-0.4
-0.2
0.0
0.2
PC1 (49.02%)
Homeless
FoodInsecurity
Rent
WelfareApp
HousePrices
location
Rural
Urban

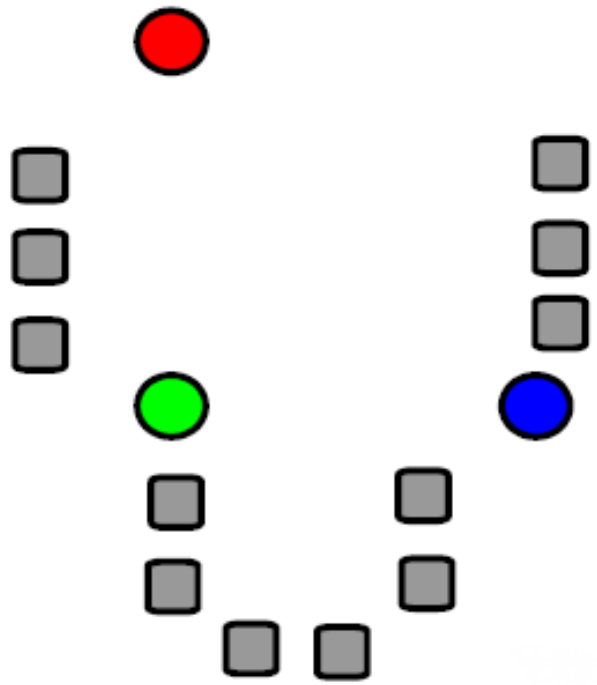
K-MEANS CLUSTERING

- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- The objective of K-means is simple: group similar data points together and discover underlying patterns.
- To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.



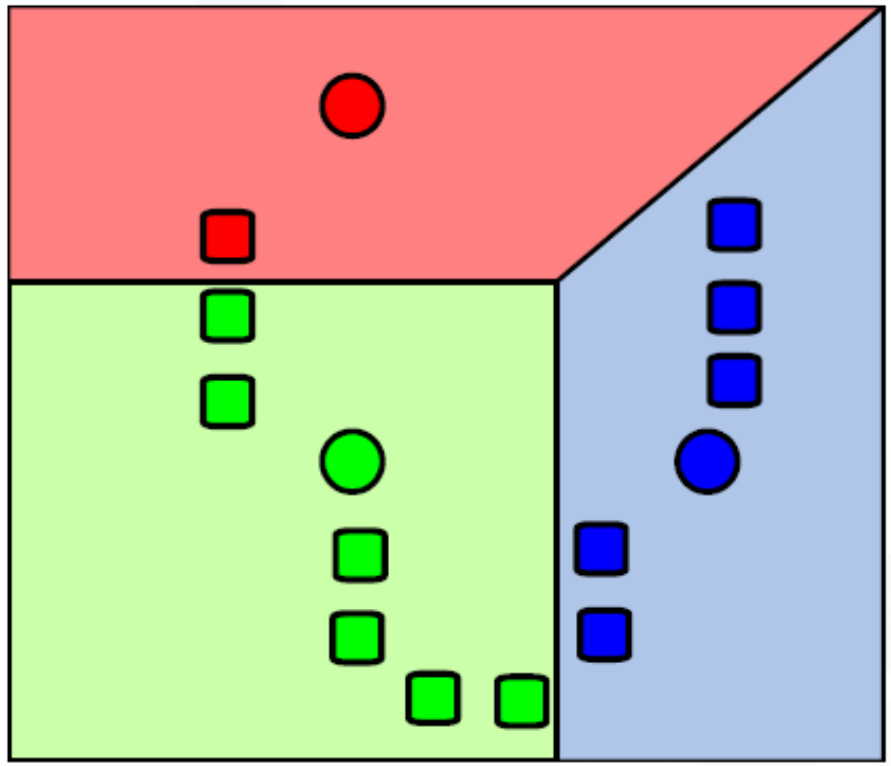
K-MEANS CLUSTERING

- To begin, we define a target number k, which refers to the number of centroids we need in the dataset.
- A centroid is the imaginary or real location representing the centre of the cluster.
- Initially, a point is randomly selected as the centre of each group.



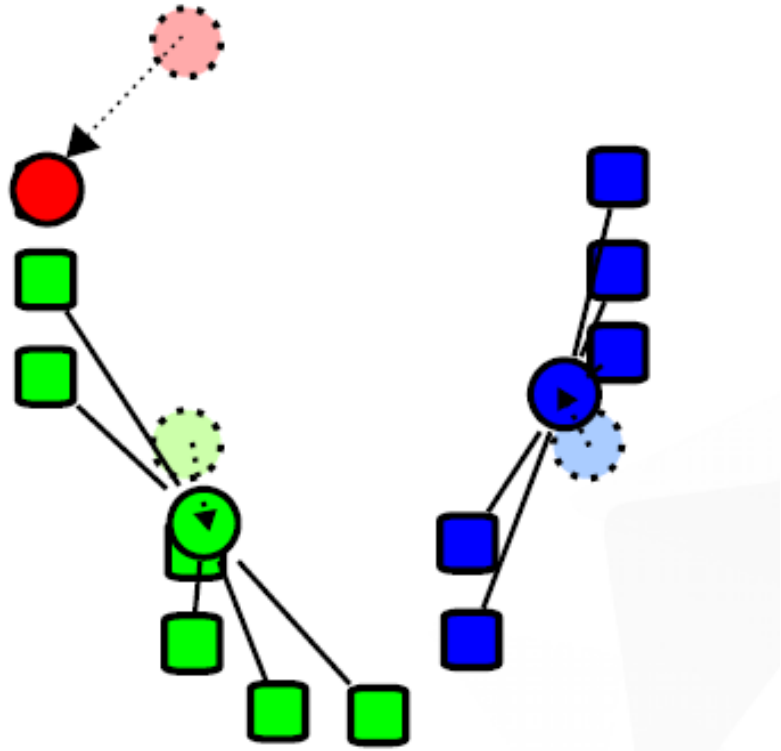
K-MEANS CLUSTERING

- Samples are then split based on Euclidean distance.



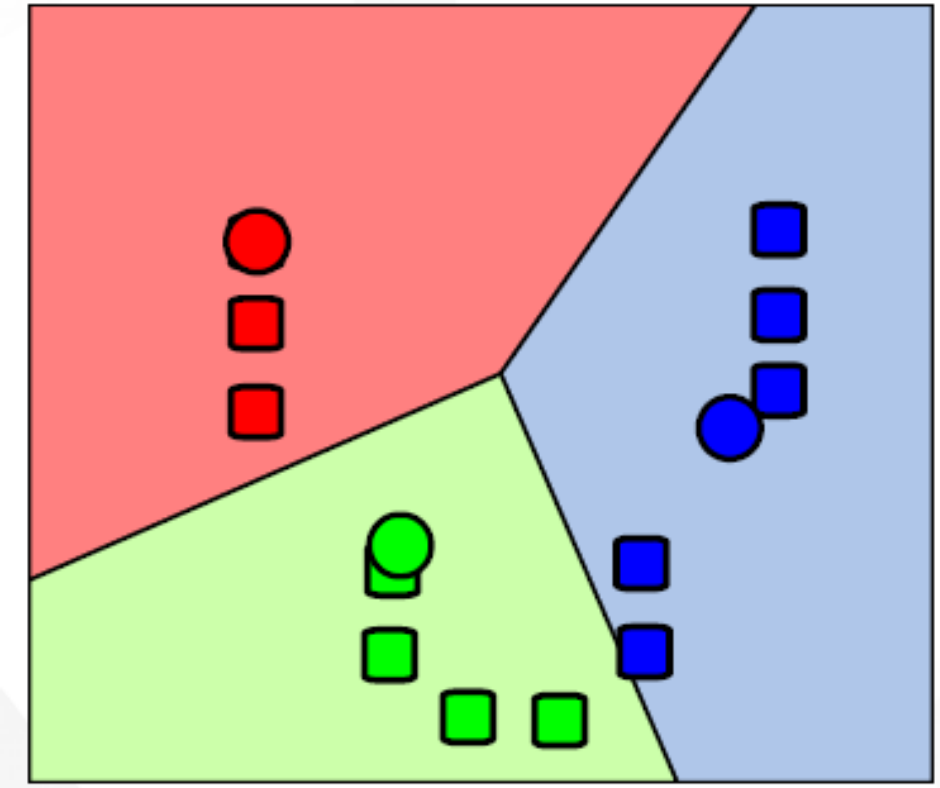
K-MEANS CLUSTERING

- To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster.
- It then performs iterative (repetitive) calculations to optimize the positions of the centroids.



K-MEANS CLUSTERING

- The sample is then reclassified based on the new centres.
- The algorithm halts creating and optimizing clusters when either:
  - The centroids have stabilized — there is no change in their values because the clustering has been successful.
  - The defined number of iterations has been achieved.



K-MEANS CLUSTERING

- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- However, it does have some drawbacks.
- Its performance is usually not as competitive as those of the other sophisticated clustering techniques because slight variations in the data could lead to high variance.
- Clusters are assumed to be spherical and evenly sized, something which may reduce the accuracy of the results.


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TIME FOR R

LET'S TRY FOR OURSELVES...

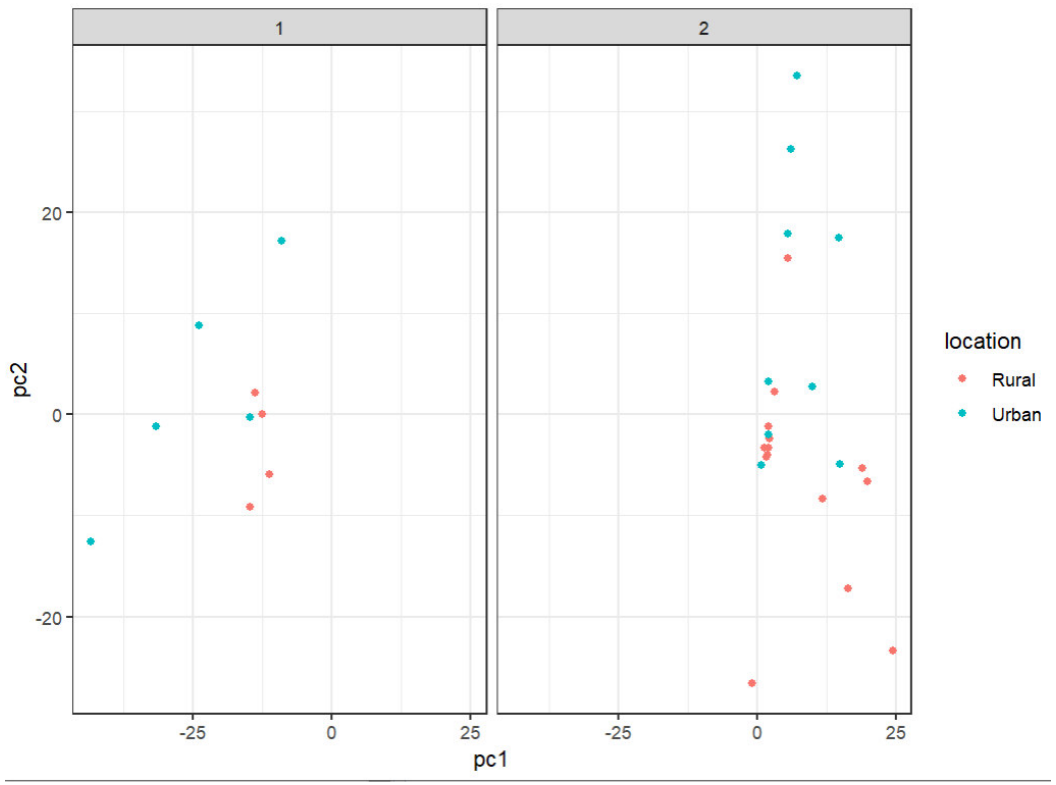
- The algorithm gives you quite a lot of information.
- These include the average values of the different variables of the cluster where each local authority was assigned.

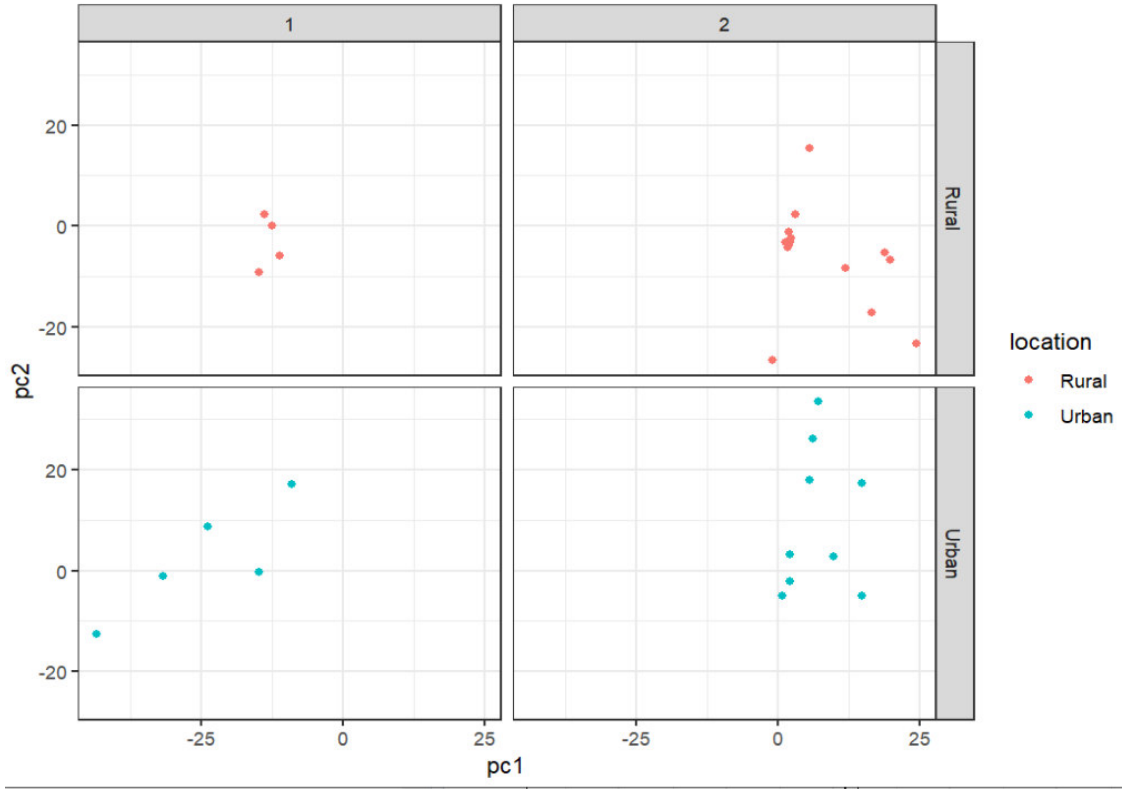
```
K-means clustering with 2 clusters of sizes 9, 23

Cluster means:
  FoodInsecurity HousePrices WelfareApp Homeless  Business     Rent
1       21.165556    10.27556 -0.5711111 3.592222 -6.123333 4.666667
2       -5.136957    13.11609  1.5360870 8.196087 -2.870435 5.582609

Clustering vector:
 [1] 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2 1 1 1 2 1 2 2 2 1 2 2 2 2 2 1 1

Within cluster sum of squares by cluster:
[1] 2487.647 6909.752
 (between_SS / total_SS =  33.7 %)
```


1
2
20
0
-20
pc2
-25
0
25
pc1
location
Rural
Urban


1
2
Rural
Urban
20
0
-20
pc2
-25
0
25
pc1
location
Rural
Urban


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
COFFEE BEAK
WE ARE GOING TO RESTART AT 15:30

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

KEYNOTE LECTURE

PROF. MELISSA TERRAS

THE UNIVERSITY of EDINBURGH

Searching for AI's killer app for Humanities research

Professor Melissa Terras

Professor of Digital Cultural Heritage
Director, Edinburgh Centre for Data, Culture & Society
Director of Research, Edinburgh Futures Institute

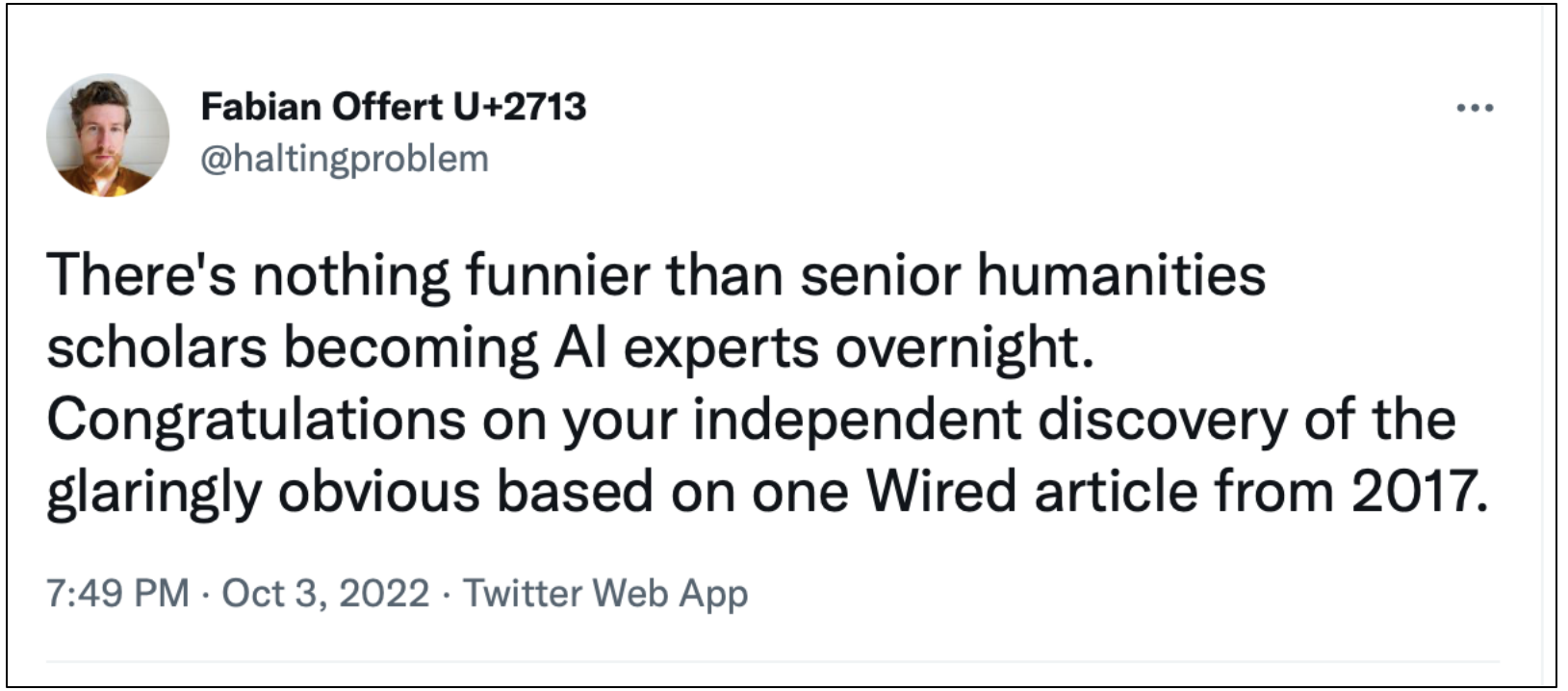
m.terras@ed.ac.uk
@melissaterras

Overview

- (What gives me the right to have an opinion?)
- The long history of Artificial Intelligence (AI) in the Humanities
- But the difficulties for AI for the Humanities
- Speculative Design Process – what would a killer AI App in the Humanities look like?
  - Three AI tools I wished existed to help me with my Humanities research!
- How can we build these type of AI applications in the humanities?
- And why the AI table needs us.

Artificial Intelligence / Humanities

- AI
  - The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. (OED)
- Humanities
  - The branch of learning concerned with human culture; the academic subjects collectively comprising this branch of learning, as history, literature, ancient and modern languages, law, philosophy, art, and music. (OED)



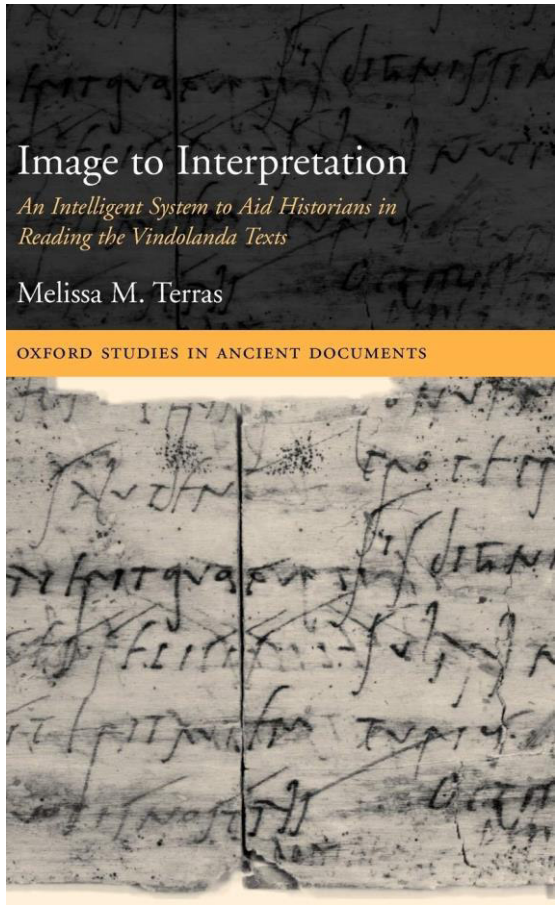
Fabian Offert U+2713
@haltingproblem

There's nothing funnier than senior humanities scholars becoming AI experts overnight. Congratulations on your independent discovery of the glaringly obvious based on one Wired article from 2017.

7:49 PM · Oct 3, 2022 · Twitter Web App

https://twitter.com/haltingproblem/status/1577007965264957440?s=20&t=Zd9tR1wYCm6OxM3LscGxHg

PhD in Expert Systems



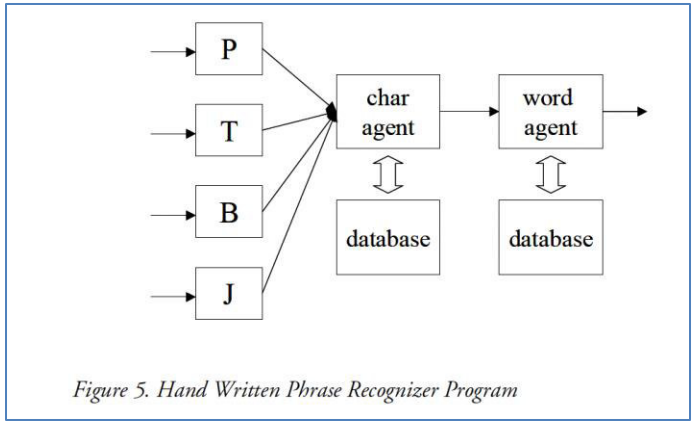
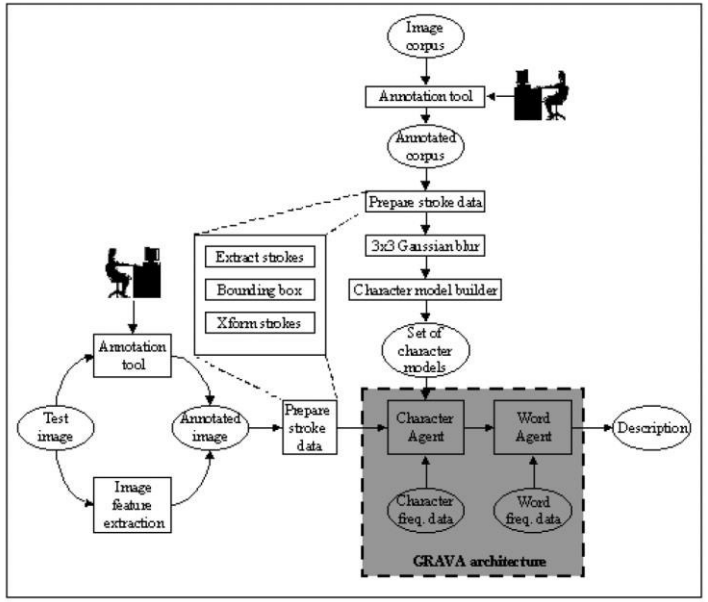
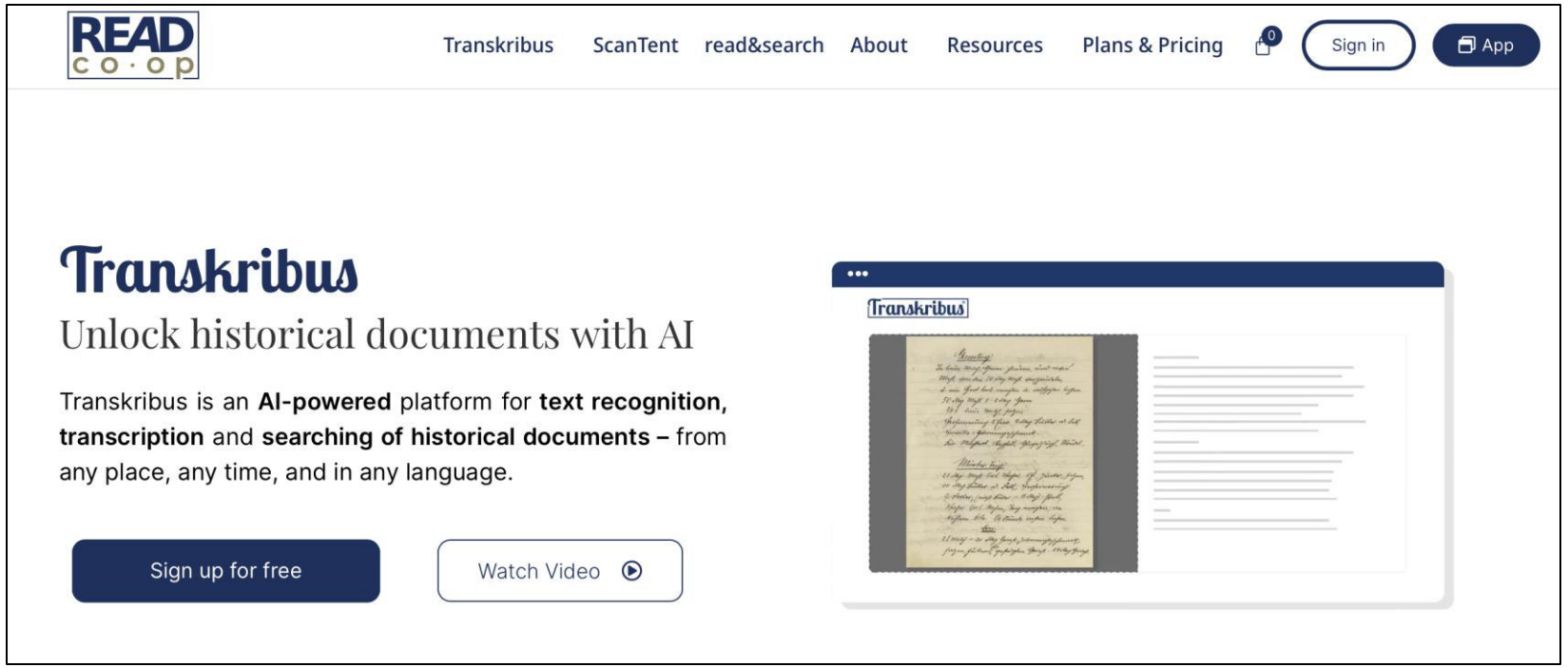
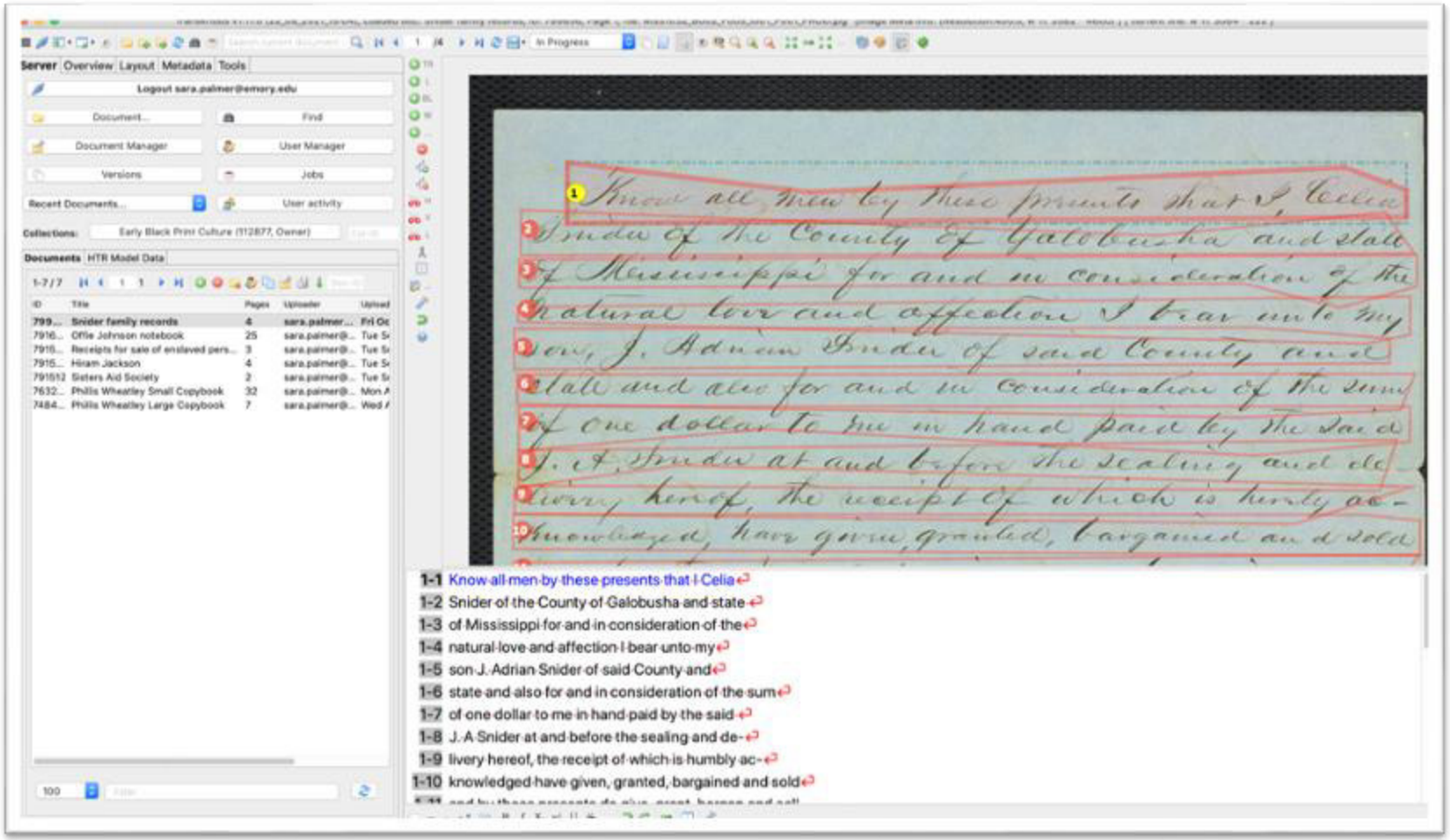


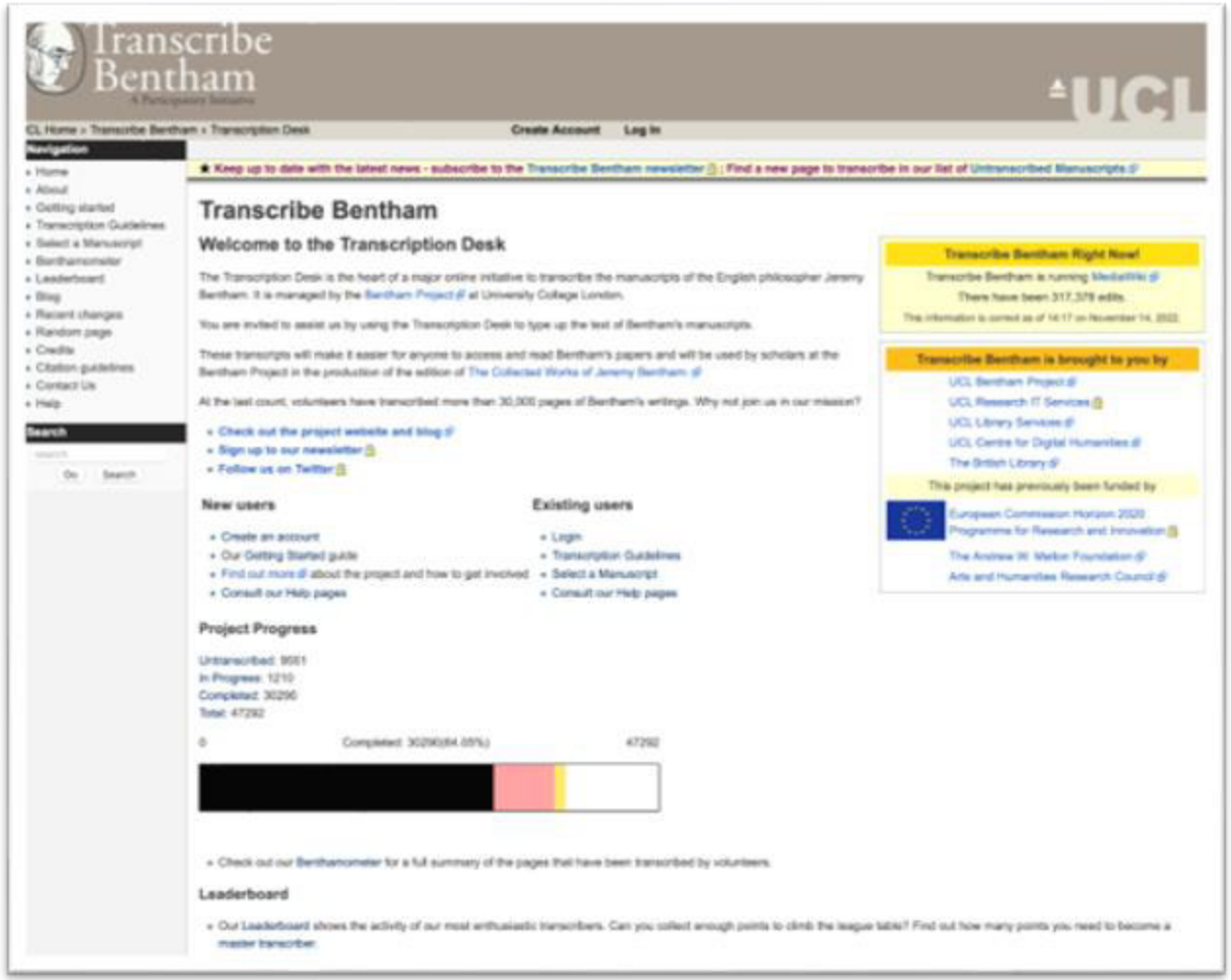
Figure 5. Hand Written Phrase Recognizer Program





https://readcoop.eu/transkribus/?sc=Transkribus


Server Overview Layout Metadata Tools
Logout sara.palmer@emory.edu
Document...
Find
Document Manager
User Manager
Versions
Jobs
Recent Documents...
User activity
Collections:
Early Black Print Culture (112877, Owner)
Documents HTR Model Data
Know all men by these presents that I Celia
Snider of the County of Galobusha and state
of Mississippi for and in consideration of the
natural love and affection I bear unto my
son, J. Adrian Snider of said County and
state and also for and in consideration of the sum
of one dollar to me in hand paid by the said
J. A. Snider at and before the sealing and de-
livery hereof, the receipt of which is humbly ac-
knowledged, have given, granted, bargained and sold
1-1 Know all men by these presents that I Celia
1-2 Snider of the County of Galobusha and state
1-3 of Mississippi for and in consideration of the
1-4 natural love and affection I bear unto my
1-5 son J. Adrian Snider of said County and
1-6 state and also for and in consideration of the sum
1-7 of one dollar to me in hand paid by the said
1-8 J. A Snider at and before the sealing and de-
1-9 livery hereof, the receipt of which is humbly ac-
1-10 knowledged have given, granted, bargained and sold



http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham

Members of READ-COOP SCE

Many hands make light work.

More than 100 institutions and private persons have already joined the coop. Every single member is an important pillar for the success of Transkribus and the further development of Transkribus. Become a member now >

135
Members

30
Countries

https://readcoop.eu/members/

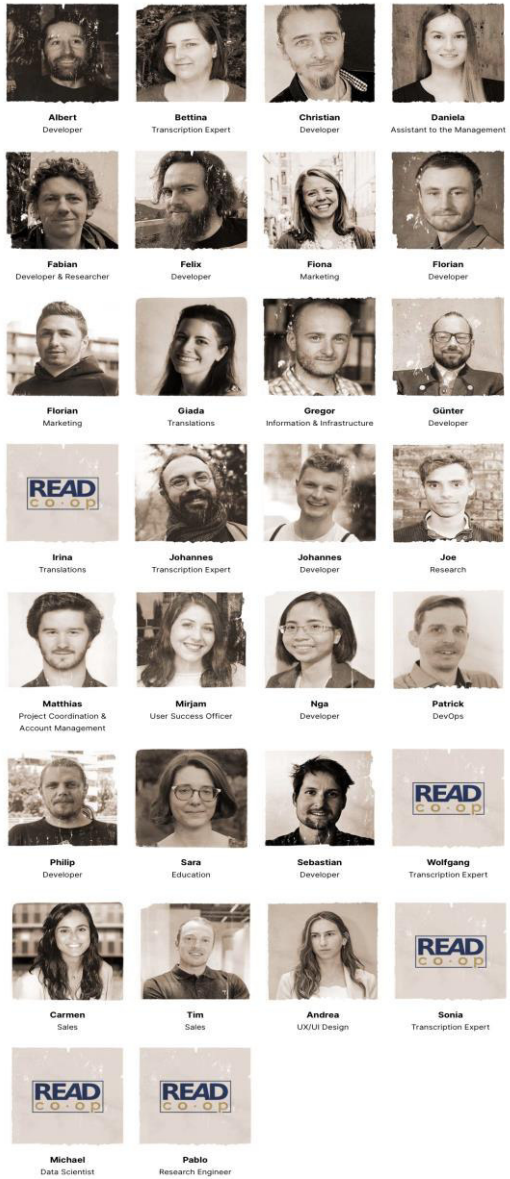
As of 2nd June 2023…

Totals

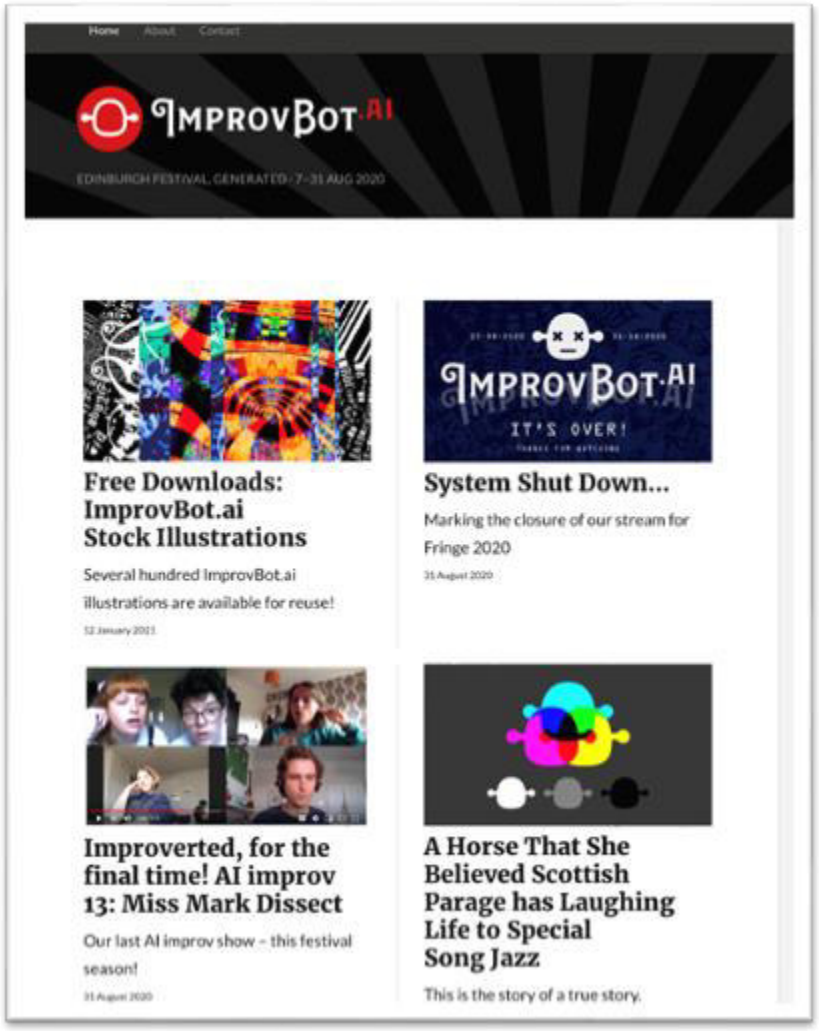
Total Active Users	Total Images	Total HTR Models
130,866	47.2M	21,339

Board

Günter Mühlberger, Chair
Melissa Terras, Scholarly Director
Andy Stauder, Managing Director



Albert, Developer
Bettina, Transcription Expert
Christian, Developer
Daniela, Assistant to the Management
Fabian, Developer & Researcher
Felix, Developer
Fiona, Marketing
Florian, Developer
Florian, Marketing
Giada, Translations
Gregor, Information & Infrastructure
Günter, Developer
Irina, Translations
Johannes, Transcription Expert
Johannes, Developer
Joe, Research
Matthias, Project Coordination & Account Management
Mirjam, User Success Officer
Nga, Developer
Patrick, DevOps
Philip, Developer
Sara, Education
Sebastian, Developer
Wolfgang, Transcription Expert
Carmen, Sales
Tim, Sales
Andrea, UX/UI Design
Sonia, Transcription Expert
Michael, Data Scientist
Pablo, Research Engineer



A Horse That She Believed Scottish Parage has Laughing Life to Special Song Jazz

Prepare to bring you the stirring story of The Concert of Giddi Shane and internationally acclaimed traditional music and returns to the Edinburgh Fringe with a celebration of successful designers and the female burst in the ruin and the silly mother. This is the story of a true story.

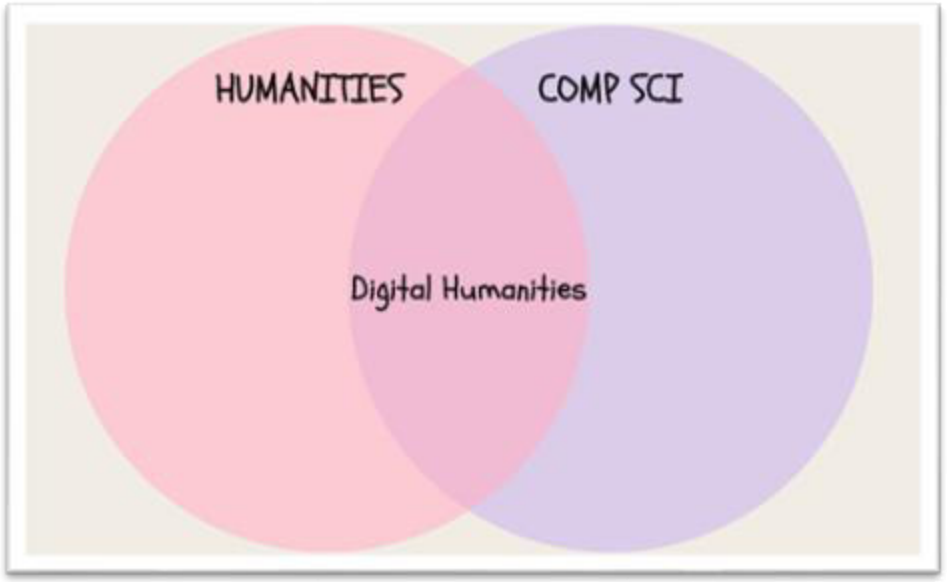
Join the music to Granchon's traditional comedy by Decasent Features, the Show Out of the Boy of the Book of on a deceptive new comedy about the works of modern Comedy Central and ghosts with theatre, supporting charity women and desperate top strange cardboard arguments.

The above show blurb was generated by The Bot, using a recurrent neural network on 2,098,140 words of Fringe show listings (2011-2019).

https://improvbot.ai

My AI Experience

- Interdisciplinarity
- Teamwork
- Common Task
- Different Perspectives
- .. I'm firmly in the Digital Humanities
- Jack of all trades etc etc



History of AI and the Humanities?

Kathleen Booth in 1964, with a newly installed IBM 1620 data-processing system at the University of Saskatchewan, where she undertook work on machine translation and neural networks. Photograph: University of Saskatchewan

https://www.theguardian.com/technology/2022/nov/10/kathleen-booth-obituary

AI's Long History in the Humanities

AI in Planning an Archaeological Excavation

Work type: paper

Mythili Rao, Ashok Marathe, Milind Vaishampayan

Languages: English

Presented at ACH/ALLC / ACH/ICCH / ALLC/EADH - 1989 - Toronto - University of Toronto

THE INDEX OF DIGITAL HUMANITIES CONFERENCES

WIZDOM - a flexible multi-purpose tutorial system based on AI-TECHNIQUES

Work type: paper

Juergen Handke

Languages: English

Presented at ACH/ALLC / ACH/ICCH / ALLC/EADH - 1990 - University of Siegen - Universität Siegen

https://dh-abstracts.library.virginia.edu/works?ordering=year&text=%22AI%22&conference=&work_type=

JOURNAL ARTICLE

An Editor's Expert System

Get access >

GORDON DIXON

Literary and Linguistic Computing, Volume 1, Issue 3, 1986, Pages 136–142,
https://doi.org/10.1093/llc/1.3.136
Published: 01 January 1986

Cite Permissions Share ▾

Abstract

The work described in this paper is concerned with the development of microcomputer software to aid the process of academic editing. Editorial procedures are represented as a network of operations that are designed to provide both advice and assistance. To achieve these aims knowledge has been encapsulated in the network so that an editor can interact with the system which is able to act as an aide memoire, an advisor, a production facility and an extending memory. This expert system has been formed from the experience gained in editing the Association for Literary and Linguistic Computing (ALLC) Bulletin.

Issue Section: Articles

https://academic.oup.com/dsh/article-abstract/1/3/136/1034530

Visual Geometry Group, University of Oxford

Bodleian Ballads Search

In cooperation with Bodleian Libraries UNIVERSITY OF OXFORD

File: Choose File no file selected or URL: | Upload and Search Whole sheet

Representative examples of images in the database

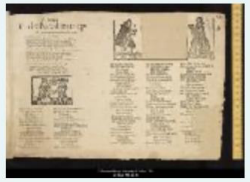
List of images in the database

Images 1 to 20 of 916. Next

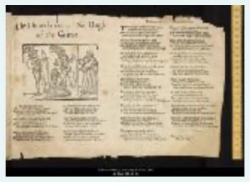
4o Rawl. 566(1)

4o Rawl. 566(2)

4o Rawl. 566(8)



4o Rawl. 566(9)



http://imagematch.bodleian.ox.ac.uk:8000/page0

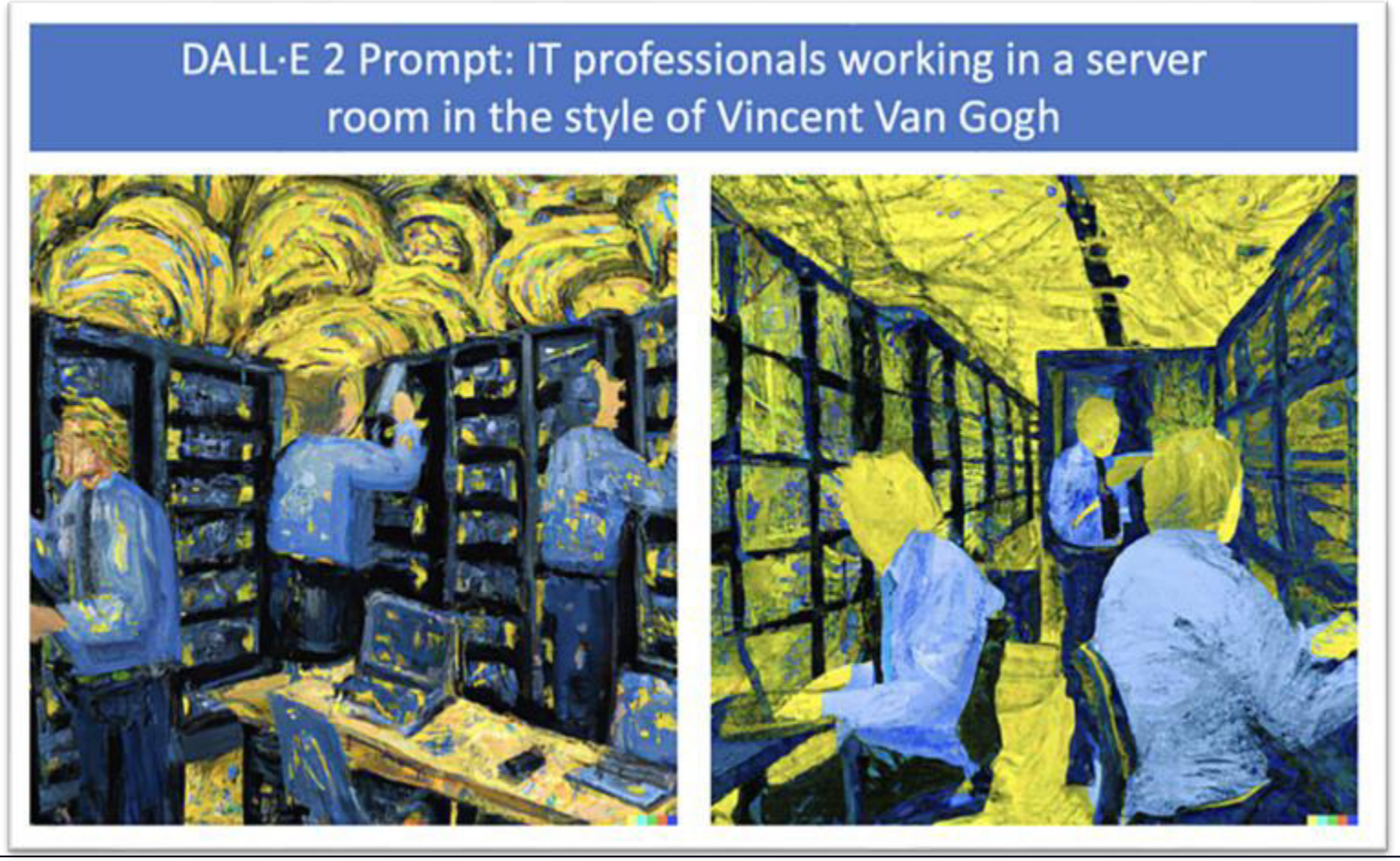
Why didn't these projects scale…

- Niche
- Specific domain
- Need training
- Need updating/ sustaining
- Effort worth the investment?
- Lack of funding/ opportunities
- Market changes/ approach changes
- "AI Winter"

Nowadays – AI Spring – AI Everywhere

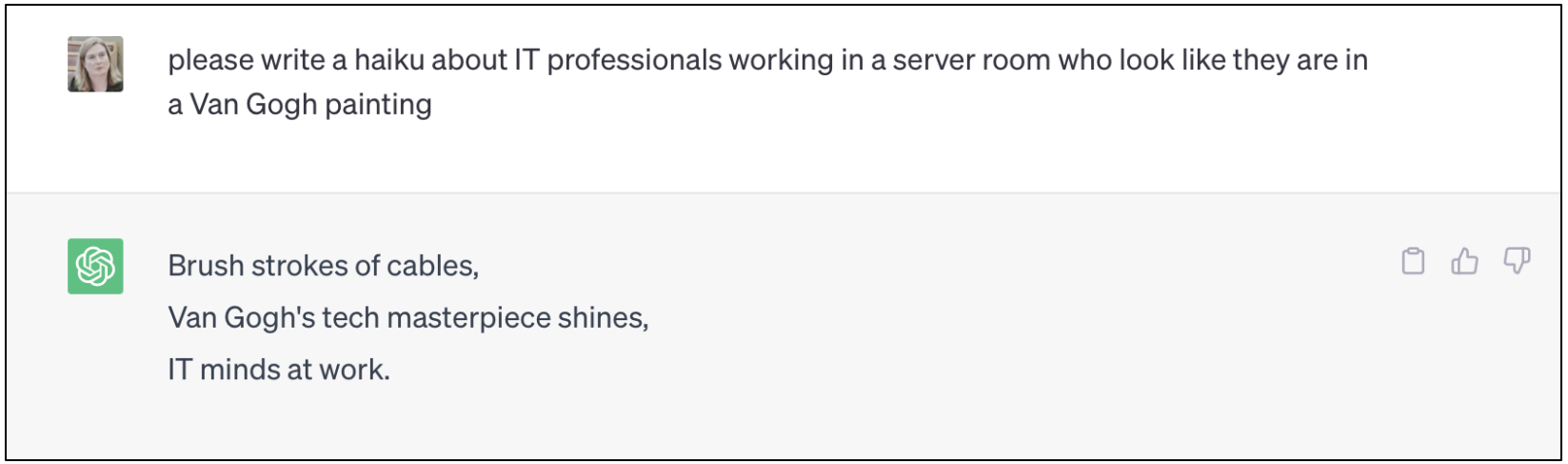
- In our search environments
- In our shopping environments
- In our writing environments
- In our artistic environments
- … Yet where is it designed for purely humanistic research?

This season's latest…

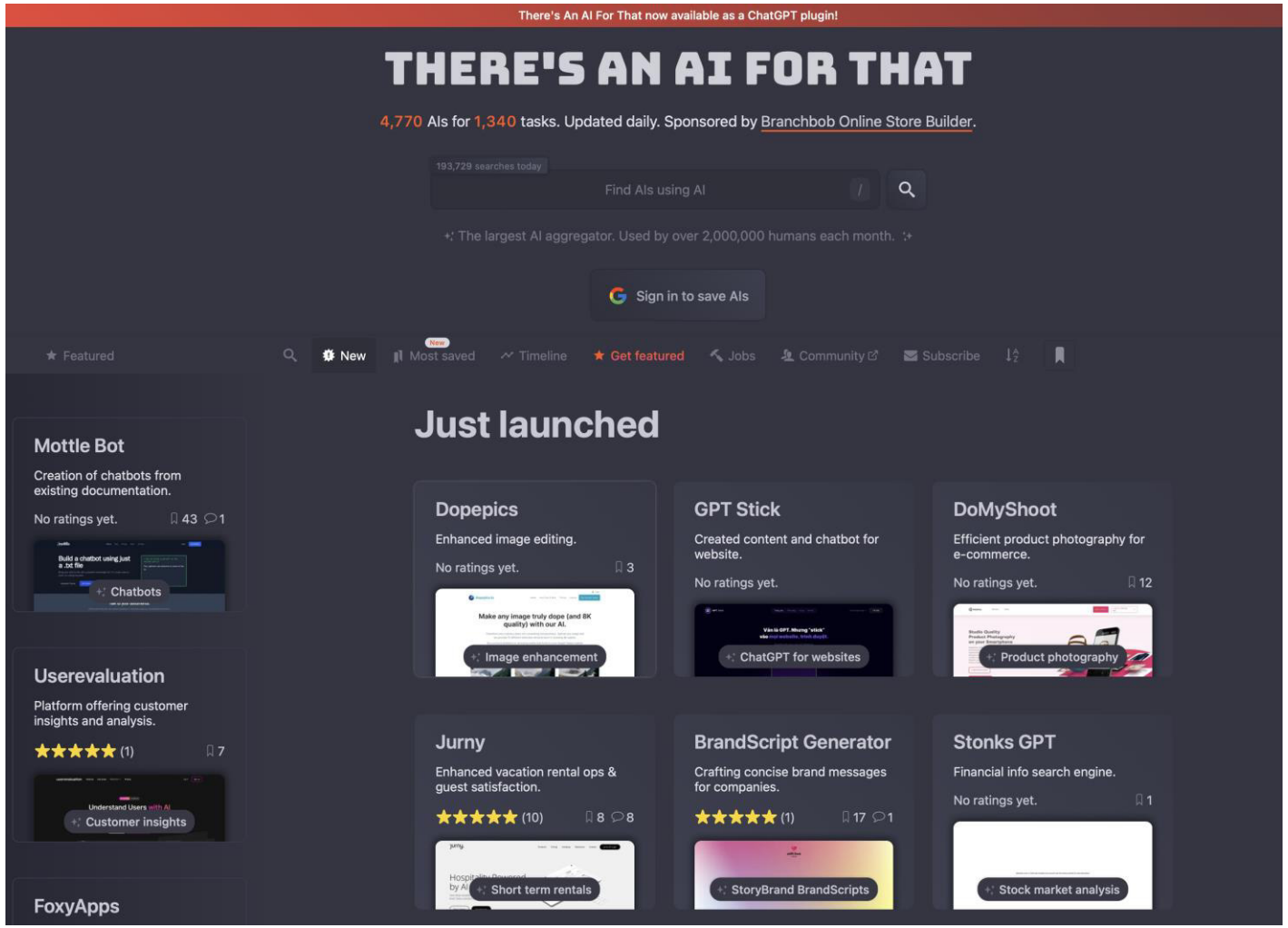


https://community.spiceworks.com/topic/2462271-art-imitates-life-dall-e-2-ai-generates-art-featuring-it-professionals

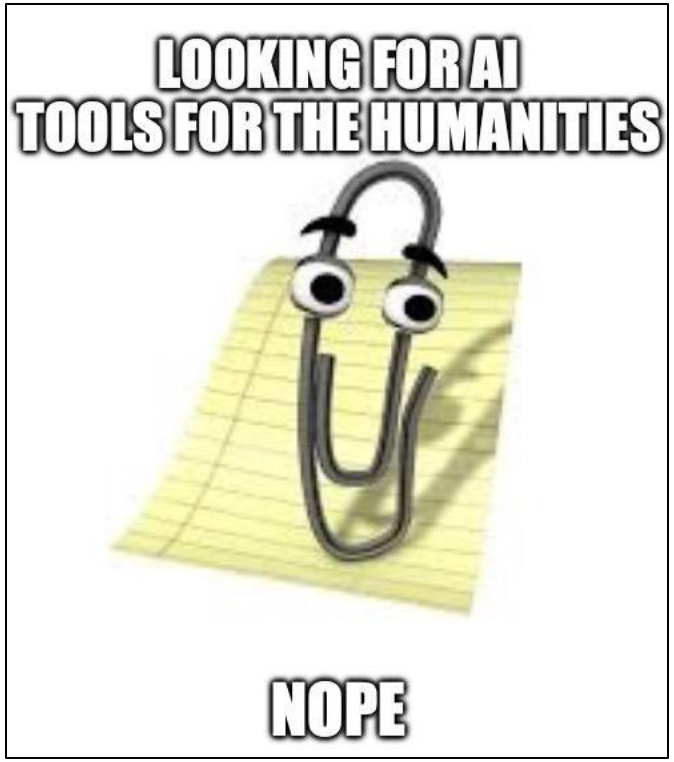
This season's latest…



https://community.spiceworks.com/topic/2462271-art-imitates-life-dall-e-2-ai-generates-art-featuring-it-professionals



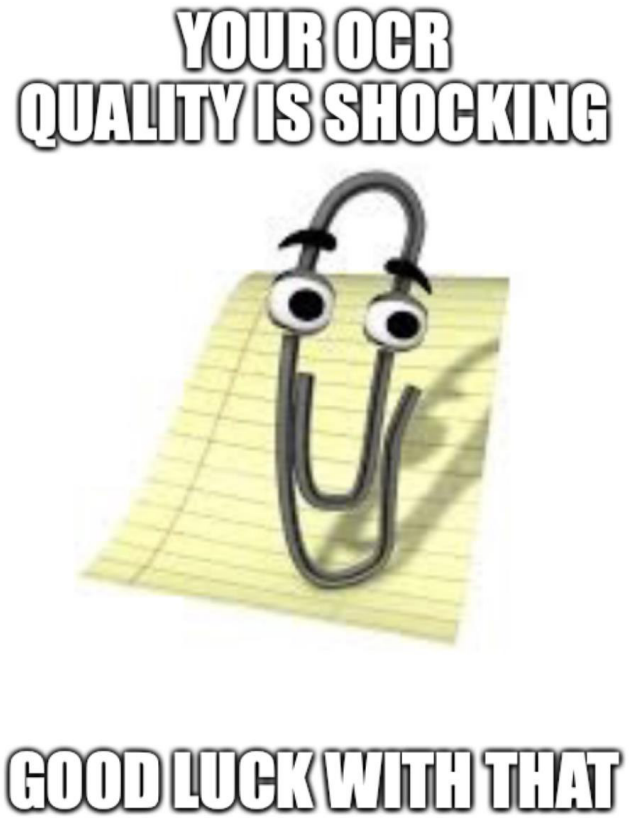
https://theresanaiforthat.com


LOOKING FOR AI
TOOLS FOR THE HUMANITIES
NOPE

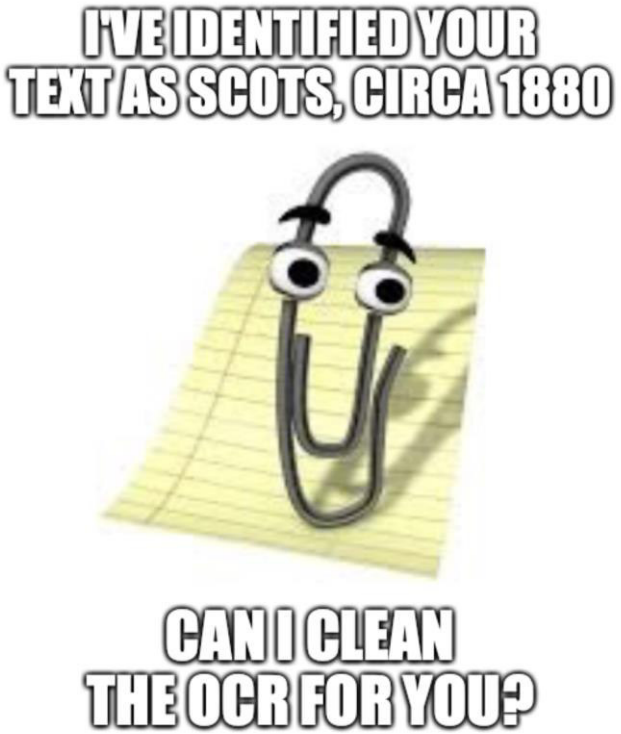
Speculative Design

- “Design futuring approaches, such as speculative design, design fiction and others, seek to (re)envision futures and explore alternatives.”
- Articulate possibilities
- Generate new directions for work
- Analyze own and others work
- Kozubaev, Sandjar, Chris Elsden, Noura Howell, Marie Louise Juul Søndergaard, Nick Merrill, Britta Schulte, and Richmond Y. Wong. "Expanding modes of reflection in design futuring." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-15. 2020.
- Three AI tools I wished existed for my Humanities/ DH Research… and what that tells us.

1: OCR Correction

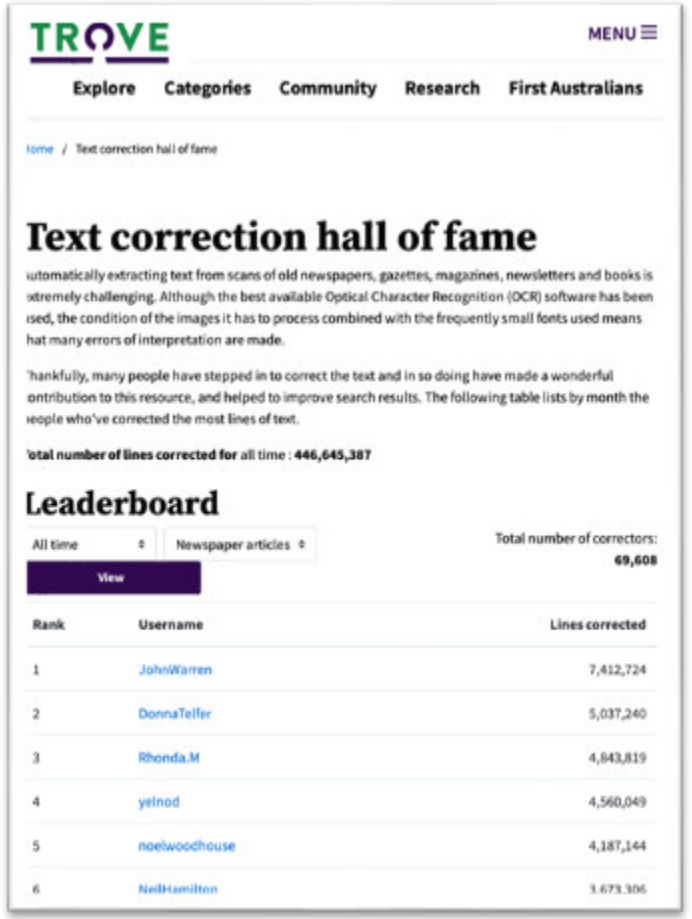


1. An AI system to provide domain specific OCR cleaning



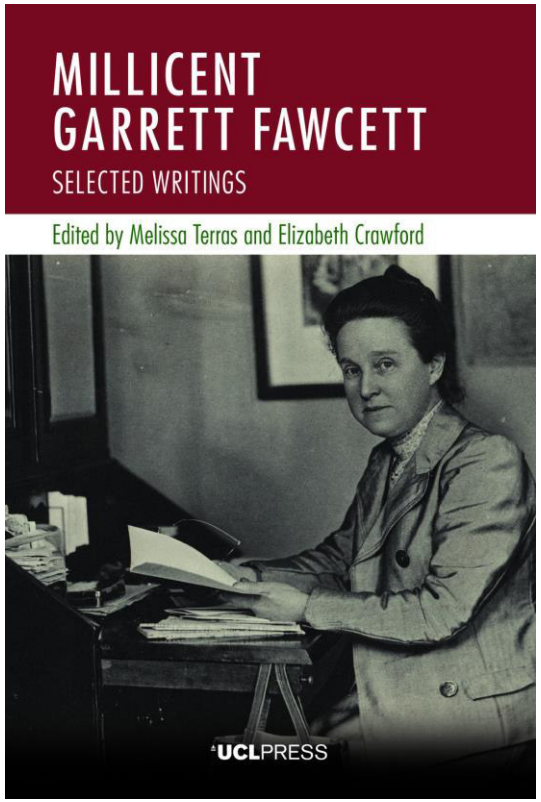
This would require:

- A way to differentiate between checked/cleaned mass OCR and dirty OCR for training
- Getting access to the cleaned OCR/ open licensing
- Building subject level AND local granularity
- Preparing our digitized collections for the robots!
- Access to ML infrastructure
  - Specific LLMs



https://trove.nla.gov.au/landing/community/hallOfFame

2. An AI system to find source quotations across mass digitized systems



"Things won are done; joy's soul lies in the doing"

Troilus and Cressida (1.2.173)

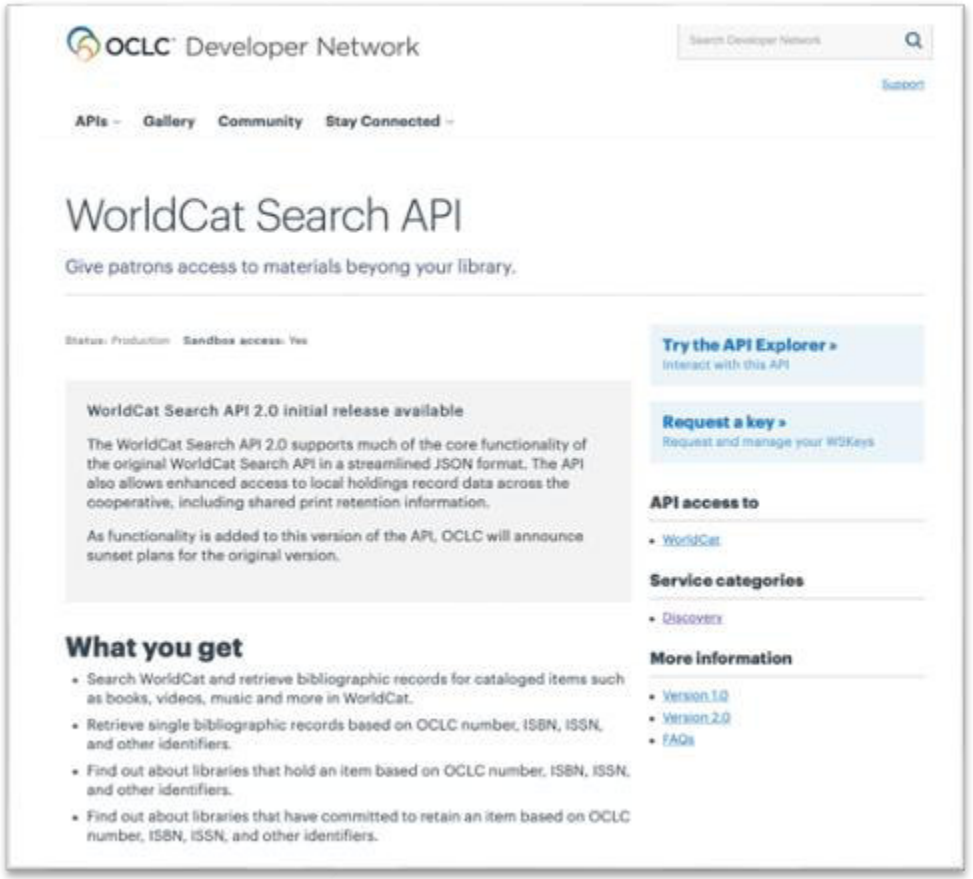
https://discovery.ucl.ac.uk/id/eprint/10149793/1/Millicent-Garret-Fawcett.pdf

This would require

- Cleaning up of OCR! (hello!)
- Federation of mass digitized content
- Agreement on search APIs between major commercial suppliers
- Ability to weight algorithms as you learn a subject's habit (try *The Times*, then *The Evening* News, then *The Scotsman*…)
- Rapid cycling, reviewing, learning, agility, improvement…

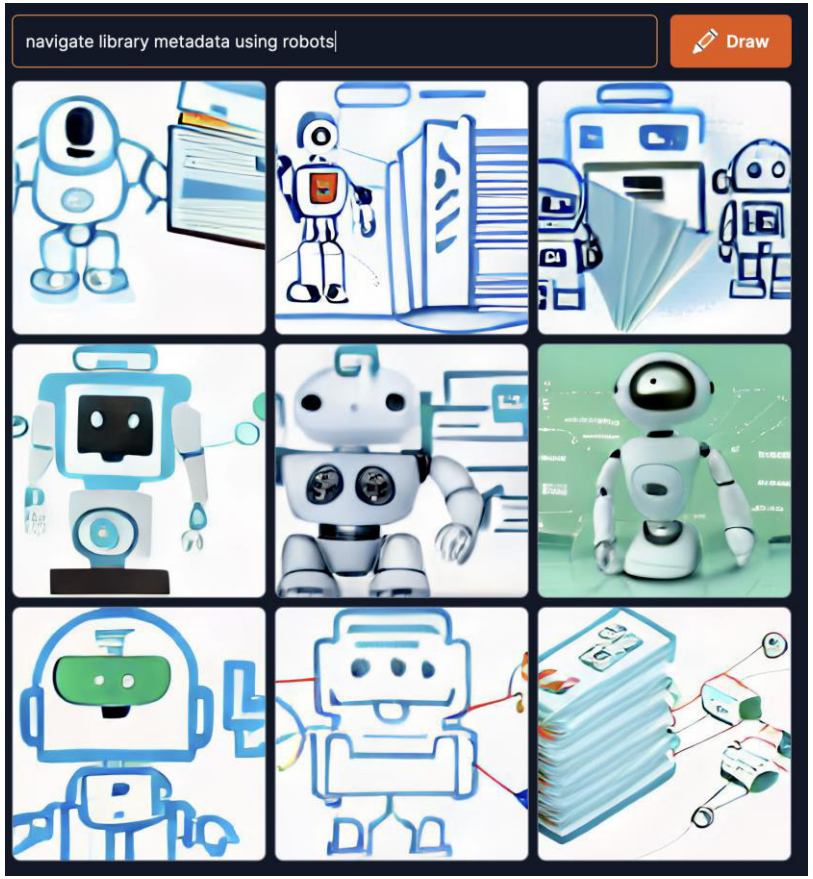
3. An AI system to find me new stuff in library databases!

- Often things you are looking for have same size/ publisher/ keywords
- But different cataloguers
- “Find me anything that may be a suffrage pamphlet”



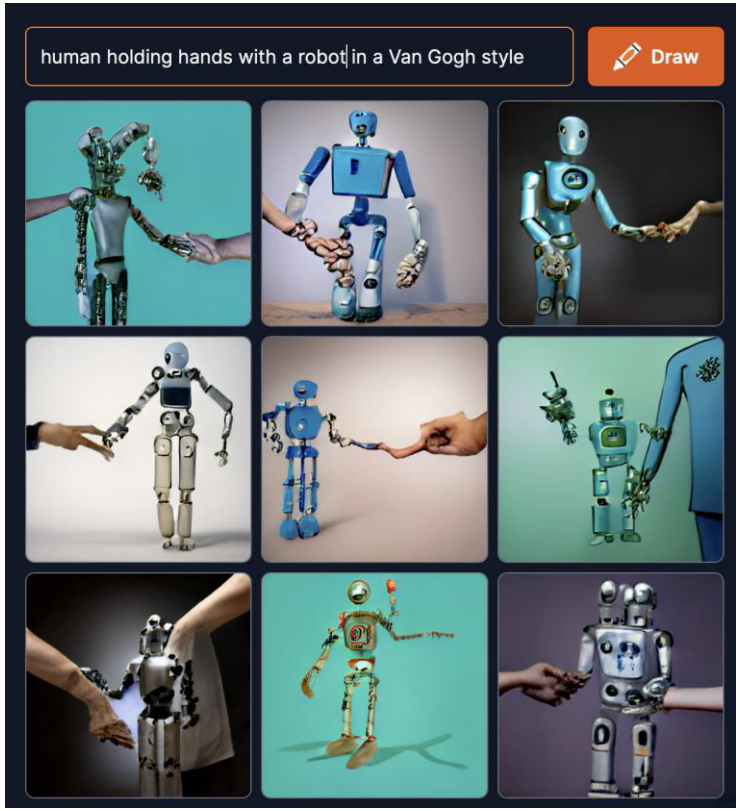
This would require

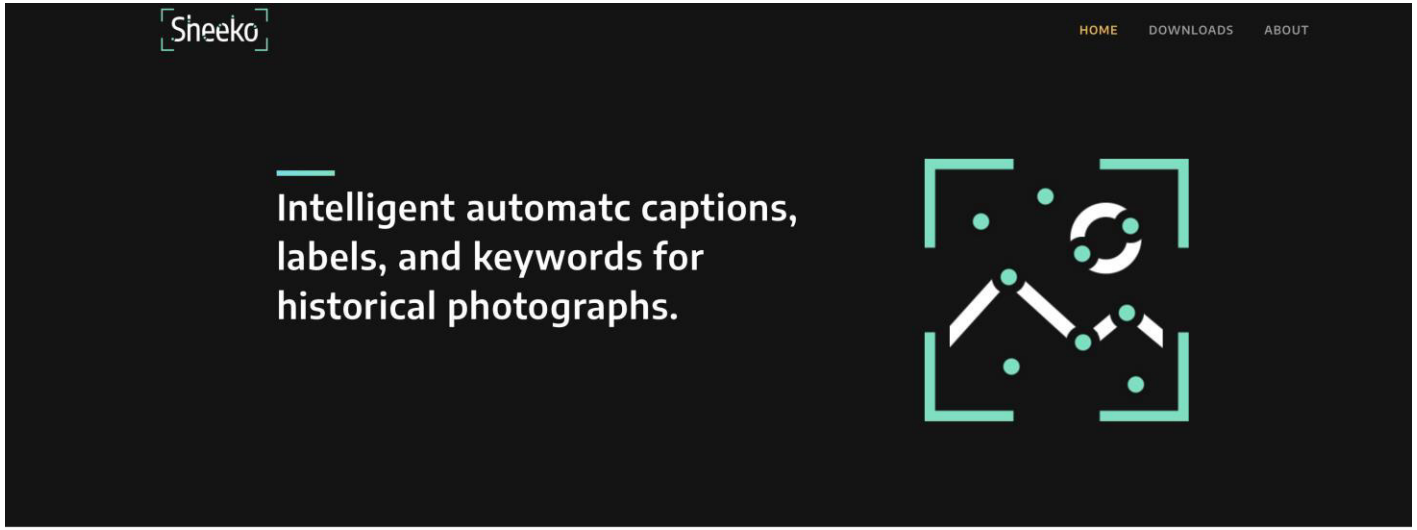
- Cleaning catalogue data
- Access to OCLC API
- Access to Library Metadata
- (which needs to be prepared for & ingested by the robots)
- Clear understanding of search structure, terms, importance
- Tweaking weighting – symbiosis with system
- Way to narrow down searches rapidly – UX challenges



Get AI to do the boring stuff

- (I want to decide what is boring, for my own research question. Let's acknowledge boring is relative)
- What is rote? What can be automated? Is that harder to define in the Humanities?
- Administrative processes?
- Where can we tweak, and control?
- Let computers do things at scale, and *bring things to me* to use my capacity to sort/analyse/synthesise
- We've built it for HTR. Where else?





Automatic Metadata Generation

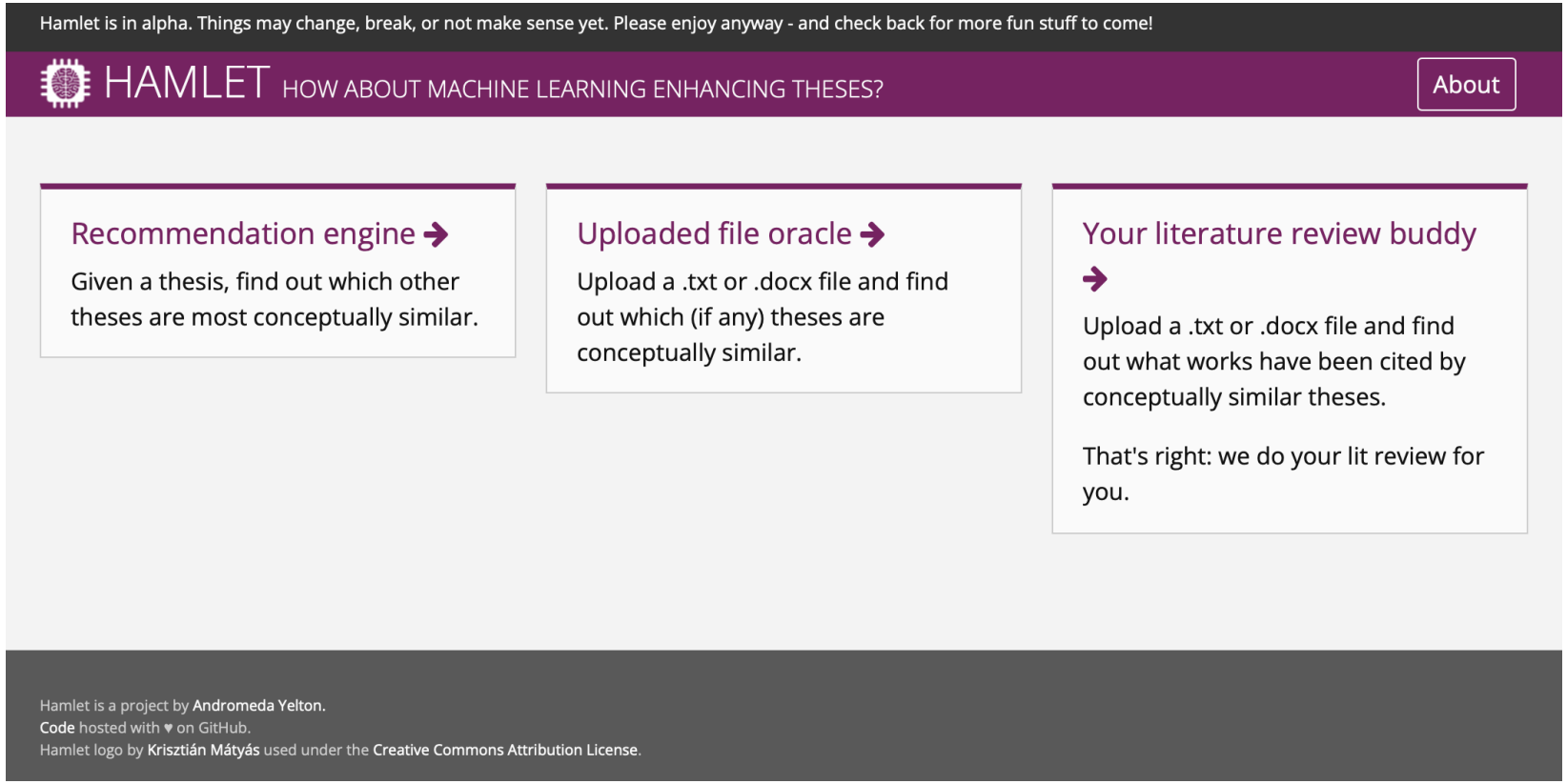
High-quality metadata is the bedrock of Digital Library systems, as it helps in users discovering the unique content in various collections housed in digital libraries.

Creating metadata is a time-intensive manual process and is done by experts who are trained in metadata schemas, and taxonomies; but the process is a bottleneck in adding content to digital libraries.

Project Sheeko examines the use of machine learning techniques to extract useful information from images that will assist metadata experts.

Learn More

https://sheeko.org



https://hamlet.andromedayelton.com – MIT Graduate Theses

melissa terras
@melissaterras

This is my biggest fear about where we are with language based generative AI - an absolute crumbling of digital infrastructure that allows you to find anything real - or know what is legit. We'll need some fair trade certification for hand-hewn words, soon...

Anosognosiogenesis @pookleblinky · May 30

Wait til amazon scammers start selling [redacted] AI generated fake books. You look up Stephen King's the Stand and the real one is hidden in a crowd of [redacted]awful 5,000 word regurgitations, etc.

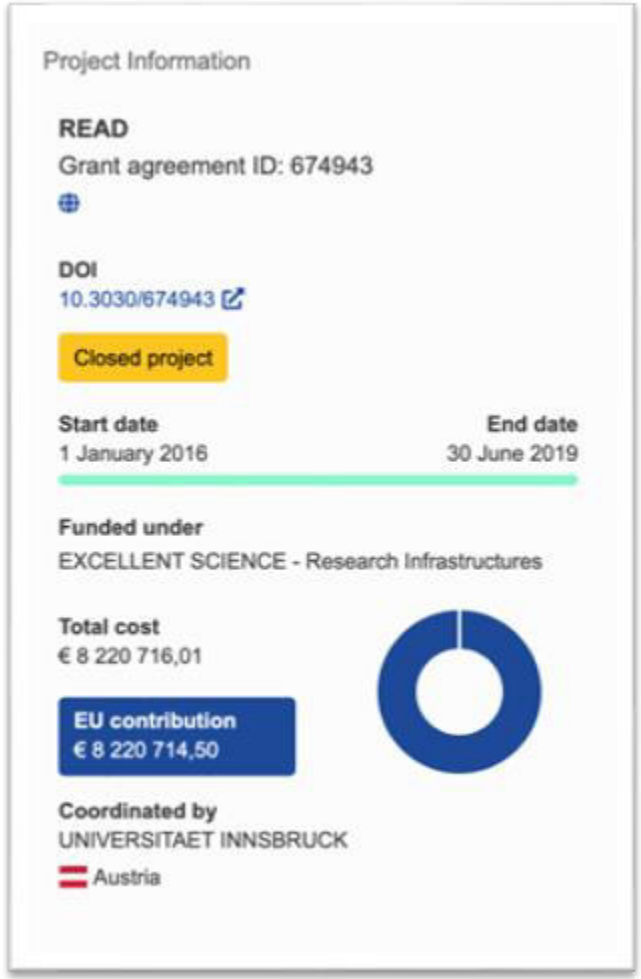
Show this thread

10:42 AM · May 30, 2023 · **5,593** Views

https://twitter.com/melissaterras/status/1663480975139938305

How do we build them?

- Digital Infrastructure isn't cheap.
- No Business Model. Not Revenue generating.
  - Transkribus had over 11m EUR funding before it stood on its own two feet…
- There's only so much voluntary "open science" we can do.
- Lack of joined up thinking re mass-digitized collections: silos
- Need for digitization of *more* content
- Lack of funding for the Arts and Humanities in comparison to science

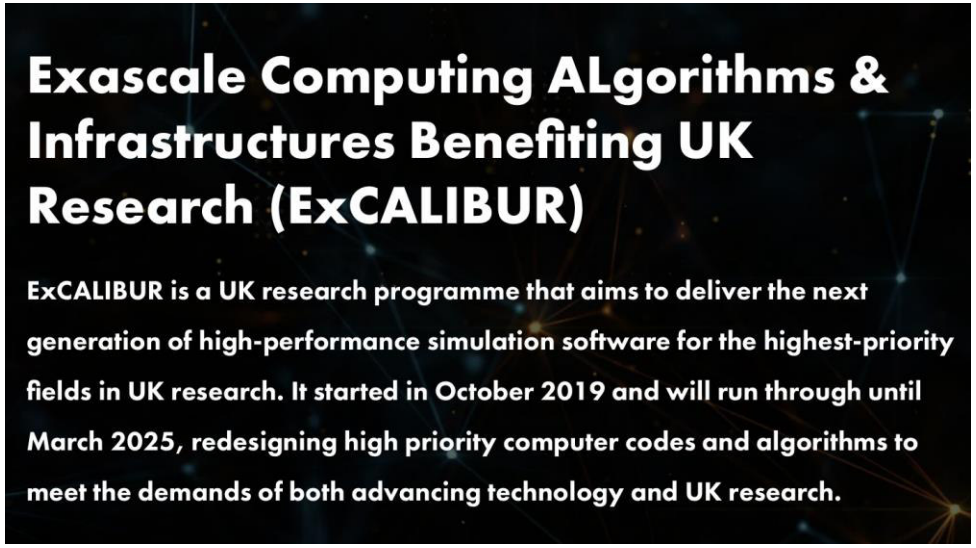


	2022-23 (£m)	2023-24 (£m)	2024-25 (£m)
Core R&I Budgets[2], of which:	**4,881**	**5,553**	**5,999**
AHRC	71	65	70
BBSRC	300	318	326
EPSRC	621	647	661
ESRC	121	119	122
MRC	548	587	615
NERC[3]	288	311	325
STFC[4]	531	544	575
Research England[5]	1,730	2,163	2,333
Innovate UK[6]	669	799	970

https://www.ukri.org/wp-content/uploads/2022/08/UKRI-190822-CorporatePlan2022to2025.pdf

Lack of competitiveness in market conditions

Lack of skills set. Or framing of problem? Or size and structure of data? Or Access to Infrastructure?



https://www.archer2.ac.uk, https://excalibur.ac.uk

New business models to sustain infrastructure

European Cooperative Society (SCE)

A European Cooperative Society (SCE) is an optional legal form of a cooperative. It aims to facilitate cooperatives' cross-border and trans-national activities. The members of an SCE cannot all be based in one country. The SCE is required to unite residents from more than one EU country.

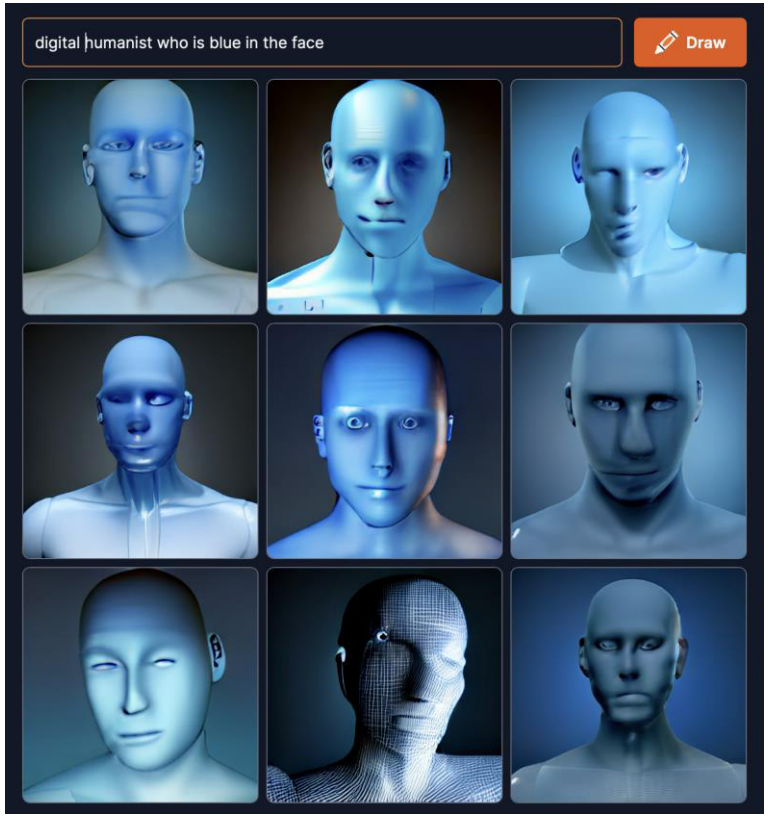
What the SCE is

- a legal entity that allows its members to carry out common activities, while preserving their independence
- its principal object is to satisfy its members' needs and not the return of capital investment
- members benefit proportionally to their profit and not to their capital contribution.

https://single-market-economy.ec.europa.eu/sectors/proximity-and-social-economy/social-economy-eu/cooperatives/european-cooperative-society-sce_en

Do Humanists just need more digital training? Sigh and repeat.

- Upskilling
- Upskilling of management/boards
- Different needs for onboarding to infrastructures
- Prioritisation of interdisciplinary work
- Means to work in teams
- Lobby for the importance of our worldview(s)
- Lobby for what we can *contribute*



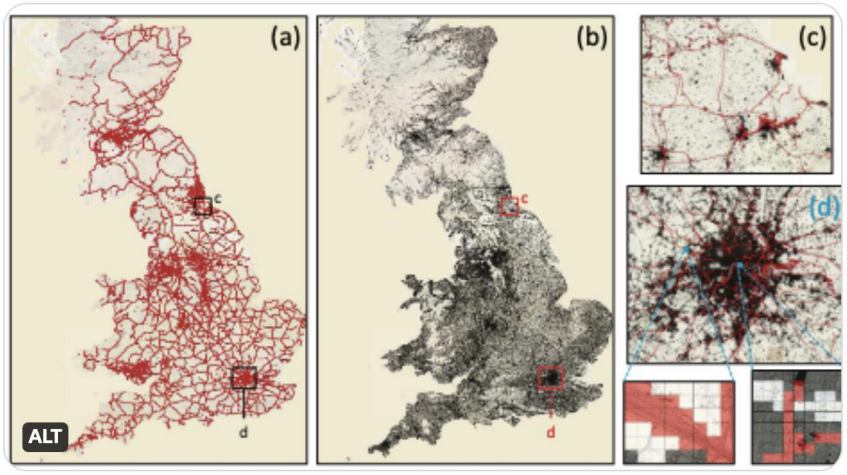
Because (Digital) Humanists can contribute

- Ethical approaches
- Understanding Biases
- Power Dynamics
- History of Cultures, Society
- Languages, Literatures
- Psychology
- Communications
- Legal frameworks
- UX/ Transparency
- We need ambition…

Kalle Westerling, PhD Retweeted

Eileen Clancy
@clancynewyork

Wowza. Python library to analyze map collections. Allows users with little computer vision expertise to 1) retrieve maps from web 2) preprocess, divide into patches 3) annotate 4) train, fine-tune, evaluate deep neural network models 5) create structured data. @LivingwMachines



Ruth Ahnert @RuthAhnert · 4h
Very proud of my colleagues for publishing on MapReader, a major outcome from @LivingwMachines by @khetiwe24 @kasra_hosseini @technocene and Kaspar Beelen. Bravo! dl.acm.org/doi/10.1145/35...

https://living-with-machines.github.io/MapReader/
https://livingwithmachines.ac.uk

THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute

MENU

Creating the world's first Scottish Gaelic Speech Recognition system

EFI Home / Creating the world's first Scottish Gaelic Speech Recognition system

Researchers are strengthening Scottish Gaelic resources – using automatic speech and handwriting recognition to advance Gaelic language technology

EFI supports interdisciplinary and data-driven research which focusses on navigating an increasingly complex future. Lying at the intersection between ethnography, linguistics and data-driven innovation, Dr Will Lamb and his team are creating and refining the world's first Scottish Gaelic Speech Recognition System.

Automatic Speech Recognition

Automatic Speech Recognition (ASR) technology can be used to translate spoken language into written text. It is used for many purposes in the lives of majority language speakers, for example in subtitling, voice assistant software and dictation services. For minority languages however, these services are often unavailable or inaccurate.

ASR and minority languages

ASR technology needs to be 'trained' with real language input to become more accurate. ASR systems analyse spoken language data, for example, audio recordings from native speakers, and learn about its patterns and structures. The more natural language data there is available, the more accurate the ASR technology can become.

For majority languages such as English there is a wealth of 'real world' language data available, for example, from literature, television, news and radio. This data can be used for training an increasingly accurate ASR system. But for minority languages such as Scottish Gaelic there is usually less of this data available. This means that ASR systems for minority languages struggle to reach the high levels of accuracy that are possible for majority languages.

Cambridge Digital Humanities

Search

About People Research Learning MPhil What's On Opportunities Media

Home > Research > Projects

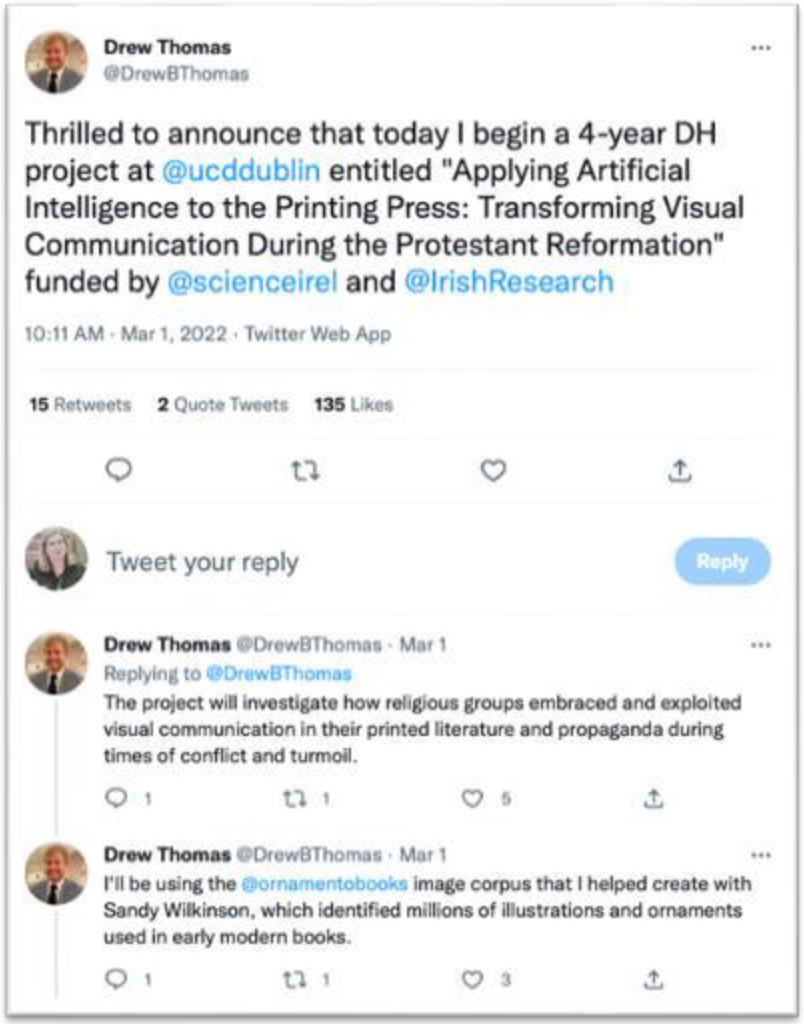
Digital Approaches to the Capture and Analysis of Watermarks

Project title: Digital approaches to the capture and analysis of watermarks using the manuscripts of Isaac Newton as a test case
Funder: AHRC
Scheme: NEH/AHRC New Directions for Digital Scholarship in Cultural Institutions
UK PI: Scott Mandelbrote, University of Cambridge
Collaborating institutions: Cambridge Digital Humanities
Dates: February 2021–August 2023

This project uses innovative imaging techniques and computer vision to provide an idea of the organisation and chronology of a body of manuscript material (taking the papers of Isaac Newton (1642-1727) as its sample). The dispersal of Newton material through the saleroom in the twentieth century has made accurate comparison of the physical evidence extremely difficult, generating a puzzle in terms of the historical order of the collection. The project is a collaboration between the University of Cambridge, the National Archives, and King's College, Cambridge in the UK and Indiana University, the Huntington Library, and the Science History Institute in the USA, and is funded jointly by the AHRC and the NEH under their call for 'New Directions for Digital Scholarship in Cultural Institutions'. We are building on expertise created at the École des Chartes and by the École des Ponts Paris Tech.

The background to the project is provided by the history of work on Isaac Newton's papers. The Newton Project began an electronic edition of the manuscripts of Isaac Newton in 1998. The Chymistry of Isaac

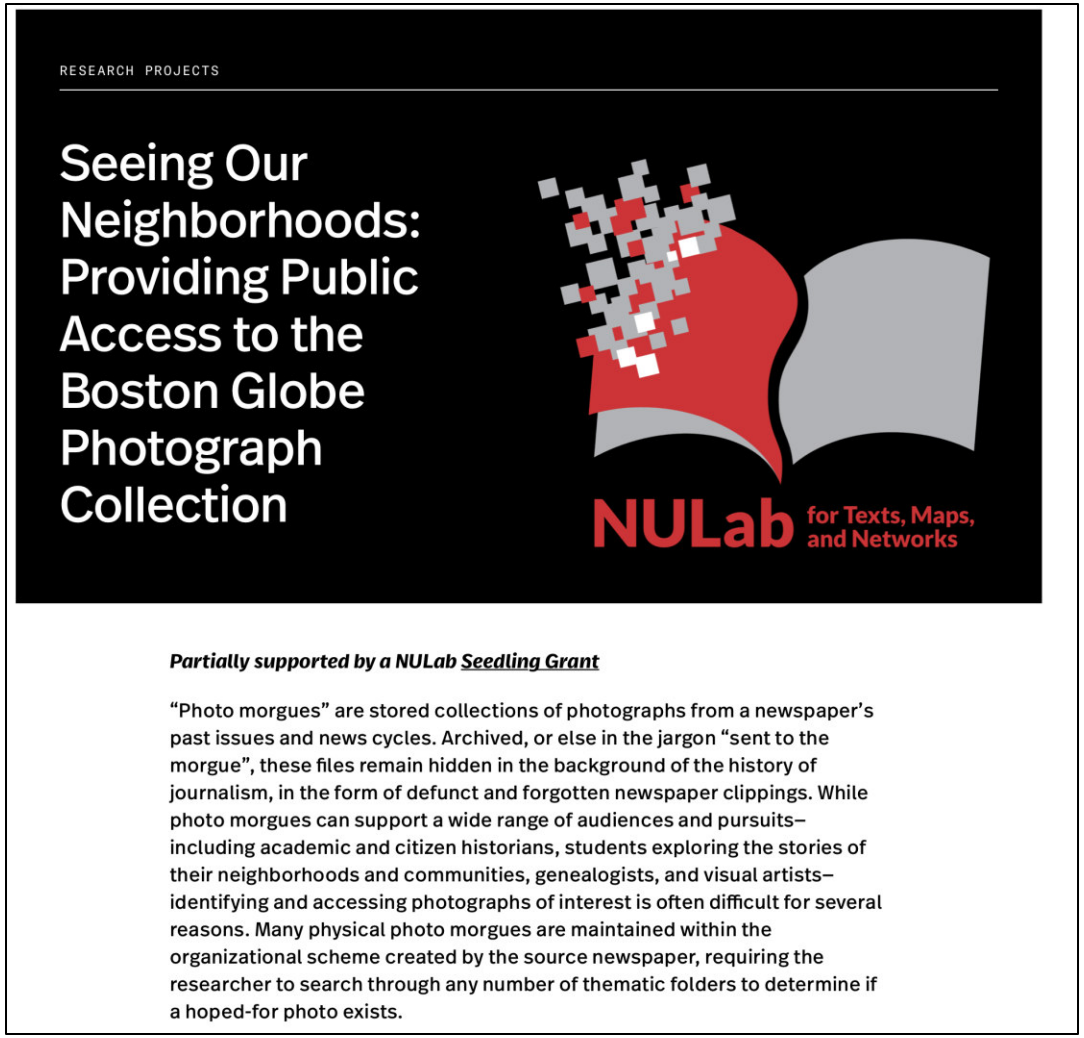
https://www.cdh.cam.ac.uk/research/projects/digital-approaches-to-the-capture-and-analysis-of-watermarks/



https://twitter.com/DrewBThomas/status/1498601847413694467?s=20&t=1iGu2o9Ld_xAxzvDuhNC7A

https://www.lancaster.ac.uk/digging-ecm/

https://unlockingarchives.com



https://cssh.northeastern.edu/nulab/seeing-our-neighborhoods/

Archaeology and Ancient History

UNIVERSITY OF LEICESTER

Study Research Partnerships and Enterprise Alumni Giving About

< Archaeology and Ancient History

< Research

< New Approaches to the Material World

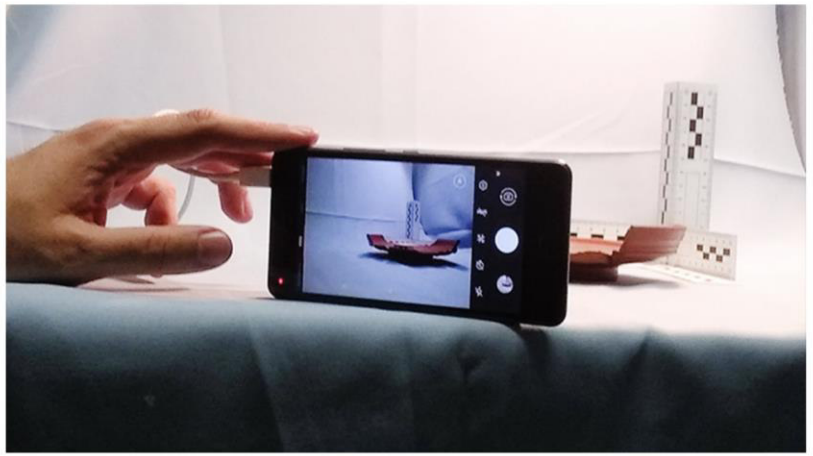
v Arch-I-Scan

About Arch-I-Scan

Meet the team

How to Get Involved

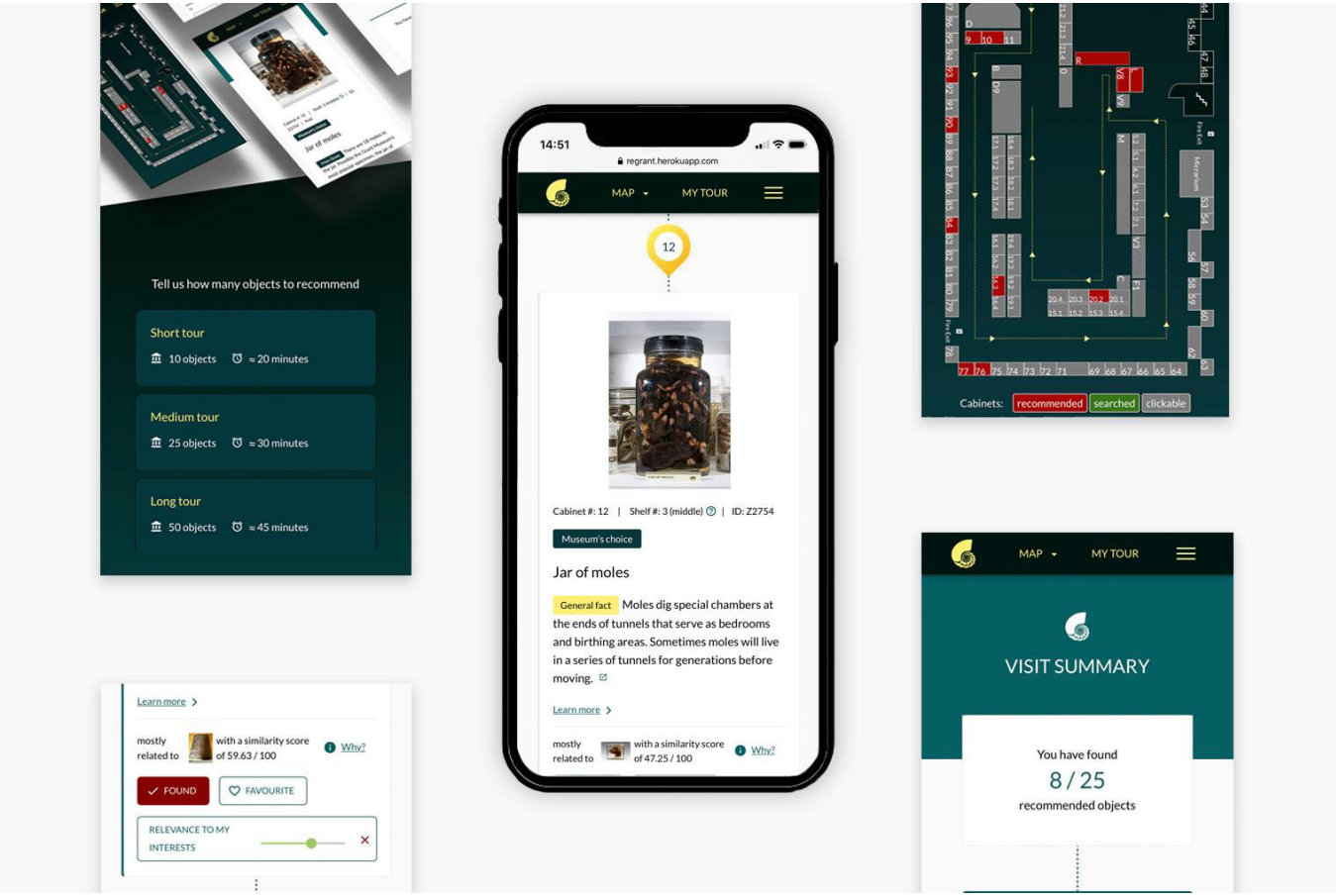
Arch-I-Scan



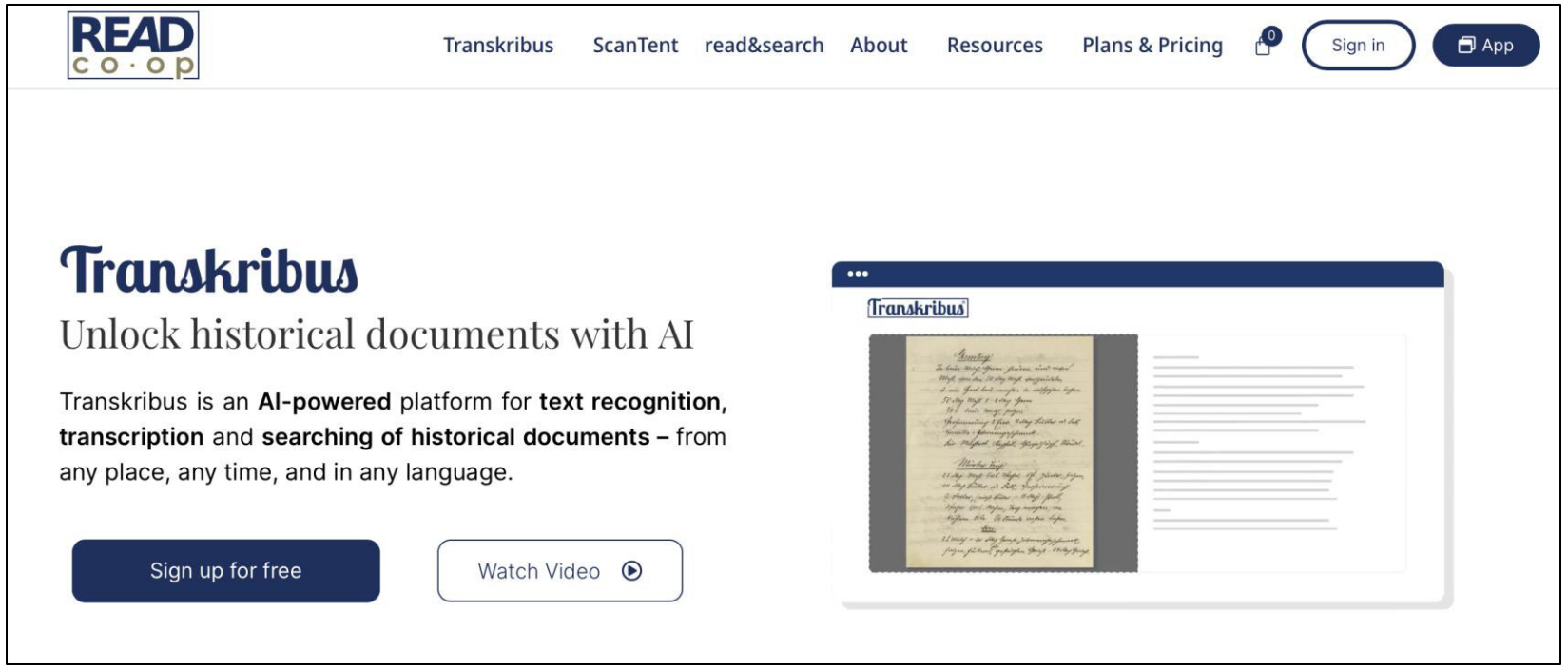
The Arch-I-Scan project is developing a state-of-the-art image-recognition and machine-learning service to automatically identify and record Roman pottery vessels and sherds. Arch-I-Scan is an interdisciplinary project, funded by the Arts and Humanities Research Council, and led by Professor Penelope Allison (**School of Archaeology and Ancient History**) and Professor Ivan Tyukin (**School of Mathematics**).

https://le.ac.uk/archaeology/research/new-approaches-to-the-material-world/arch-i-scan

https://www.nb.no/maken/



https://regrant.herokuapp.com/#/home



https://readcoop.eu/transkribus/?sc=Transkribus

THE UNIVERSITY of EDINBURGH

Searching for AI's killer app in the Humanities

Professor Melissa Terras

Professor of Digital Cultural Heritage
Director, Edinburgh Centre for Data, Culture & Society
Director of Research, Edinburgh Futures Institute

m.terras@ed.ac.uk
@melissaterras

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

RECEPTION

THE COUNTING HOUSE

34 West Nicholson Street

Edinburgh EH8 9DD

www.cdcs.ed.ac.uk