Unveiling the Morphological Characteristics of Seeds: A Comparative Data-Driven Approach to Understanding Domestication and Wild Traits

CAS in Applied Data Science
Module 2

[redacted]

9 Oct. 2023

Painting from the tomb of Menna, agricultural scribe of Pharaoh

Jared Diamond´s questions:

“why were so few species domesticated, in so few locations ?”

Why were some species domesticated and others not is especially hard to answer because only a handful of species were domesticated and these are usually unrelated taxonomically and grow in different locations and climates.

→ Domestication events are rare but foundational events of civilisations.

→ The two groups of plants that were most often domesticated: the grasses (sugar source), and the legumes (protein source).

→ One plant group of 300 species, the pulses (Viciae), was domesticated exceptiona[redacted] exceptional location: the fertile crescent.

Painting from the tomb of Menna, agricultural scribe of Pharaoh

Study System

Vicia faba
(Faba bean)

Lens culinaris ssp. Culinaris
(Lentils)

Lathyrus sativus
(Sweet pea)

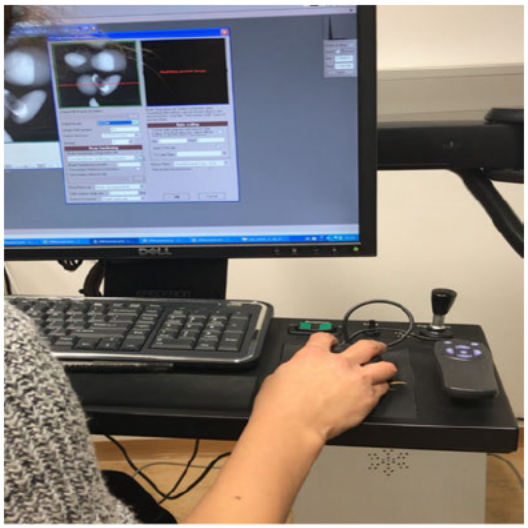
Pisum sativum ssp. Sativum
(Peas)

Data collection

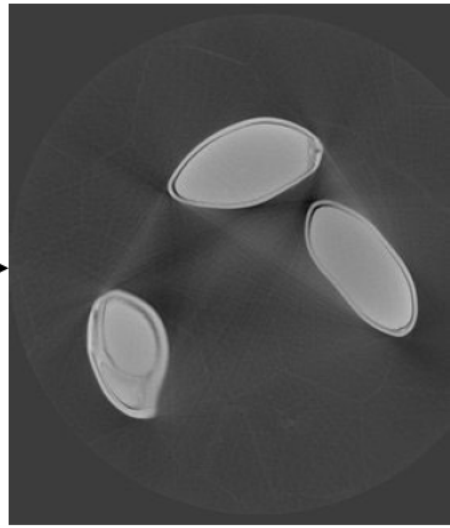
Seeds (mostly from herbarium loan)

Mounted in Styrofoam
(X-Ray transparent)

Transfer to CT scanner



Scan programming

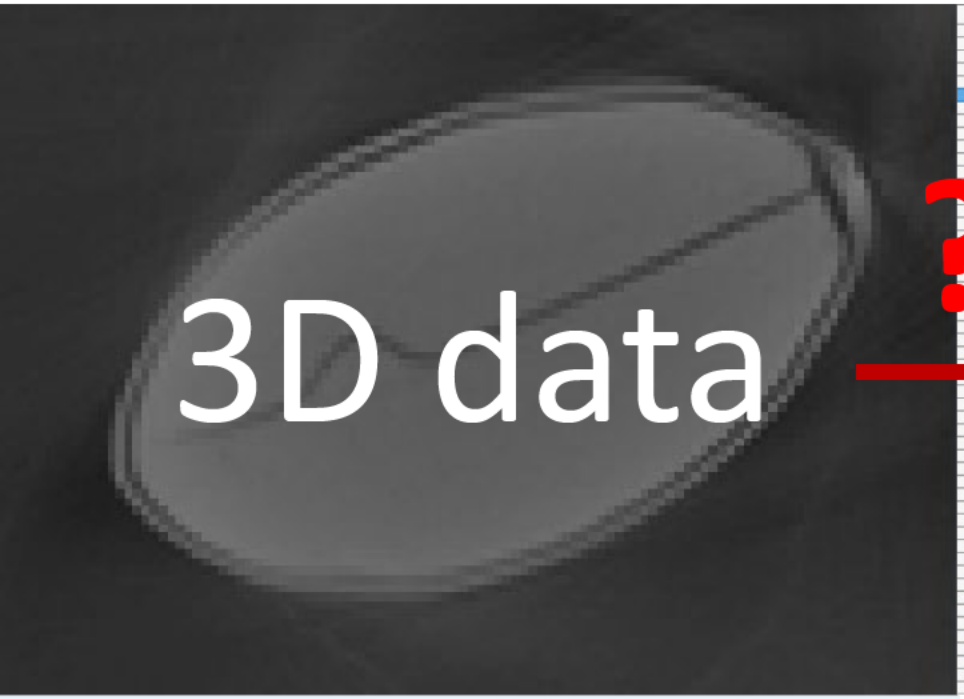


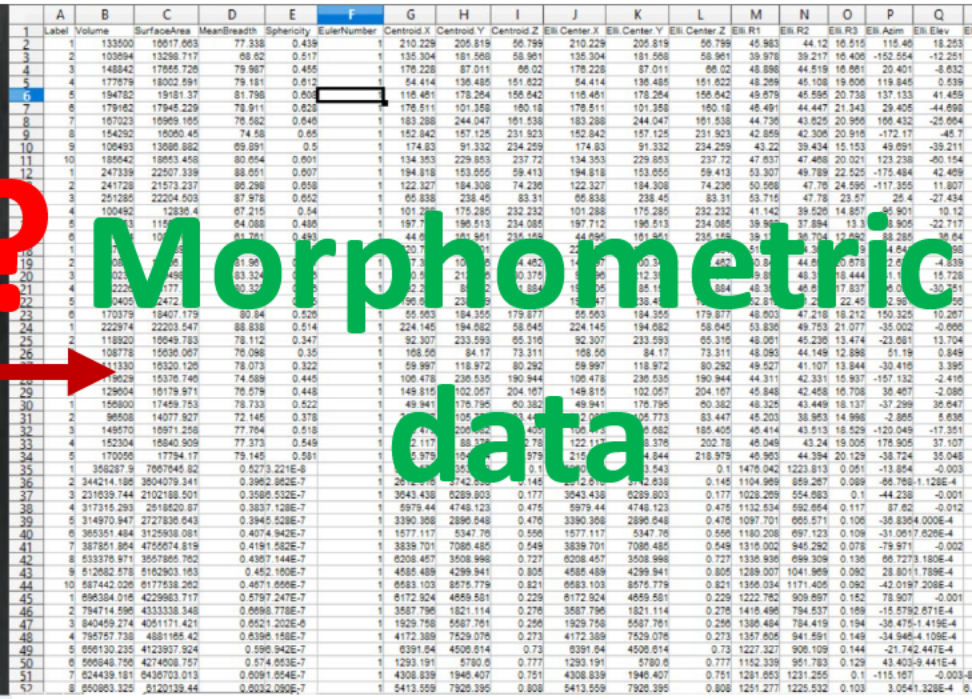
Scan data: stack of grayscale pictures

Image and data processing pipeline

- How do we get from

This to that





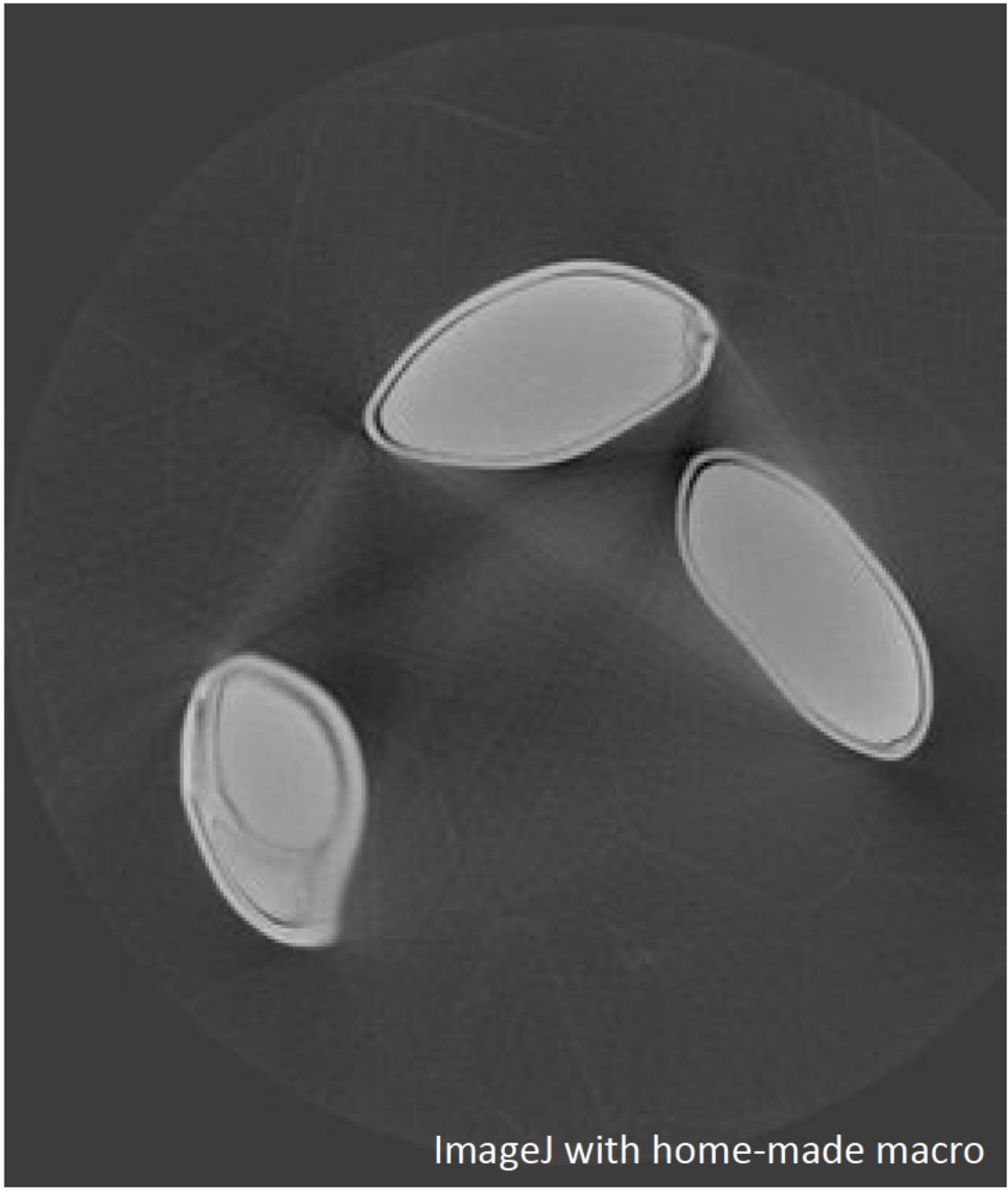
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

*Processing_macro.ijm

```
1  run("8-bit");
2
3  run("Next Slice [>]");run("Next Slice [>]");
4  run("Next Slice [>]");run("Next Slice [>]");
5  run("Next Slice [>]");run("Next Slice [>]");
6  run("Next Slice [>]");run("Next Slice [>]");
7  run("Next Slice [>]");run("Next Slice [>]");
8  run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast...");
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity e
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch | Kill | persistent | Show Errors | Clear



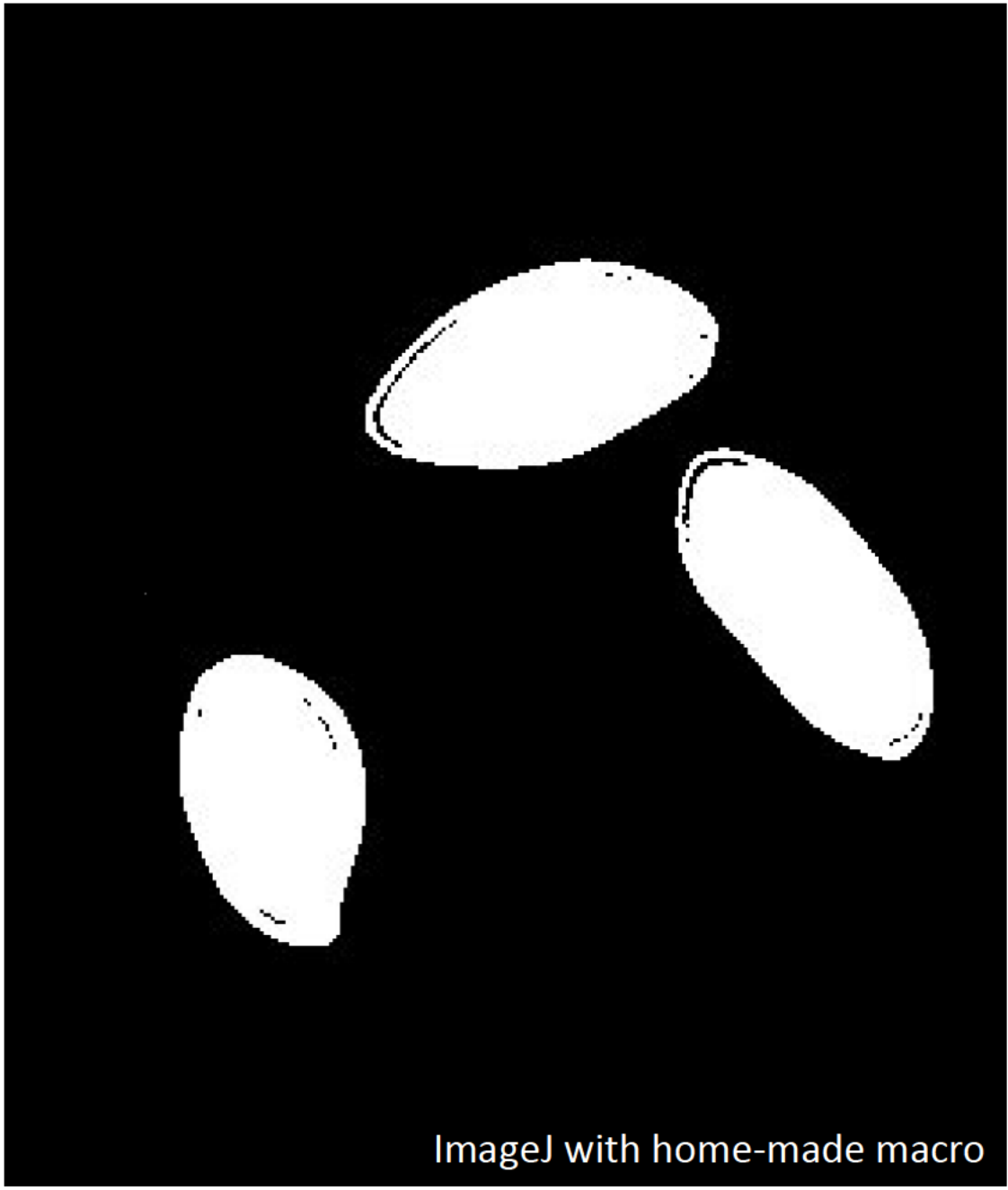
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

*Processing_macro.ijm

```
1  run("8-bit");
2
3  run("Next Slice [>]");run("Next Slice [>]");
4  run("Next Slice [>]");run("Next Slice [>]");
5  run("Next Slice [>]");run("Next Slice [>]");
6  run("Next Slice [>]");run("Next Slice [>]");
7  run("Next Slice [>]");run("Next Slice [>]");
8  run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast...");
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity e
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch Kill persistent Show Errors Clear



ImageJ with home-made macro

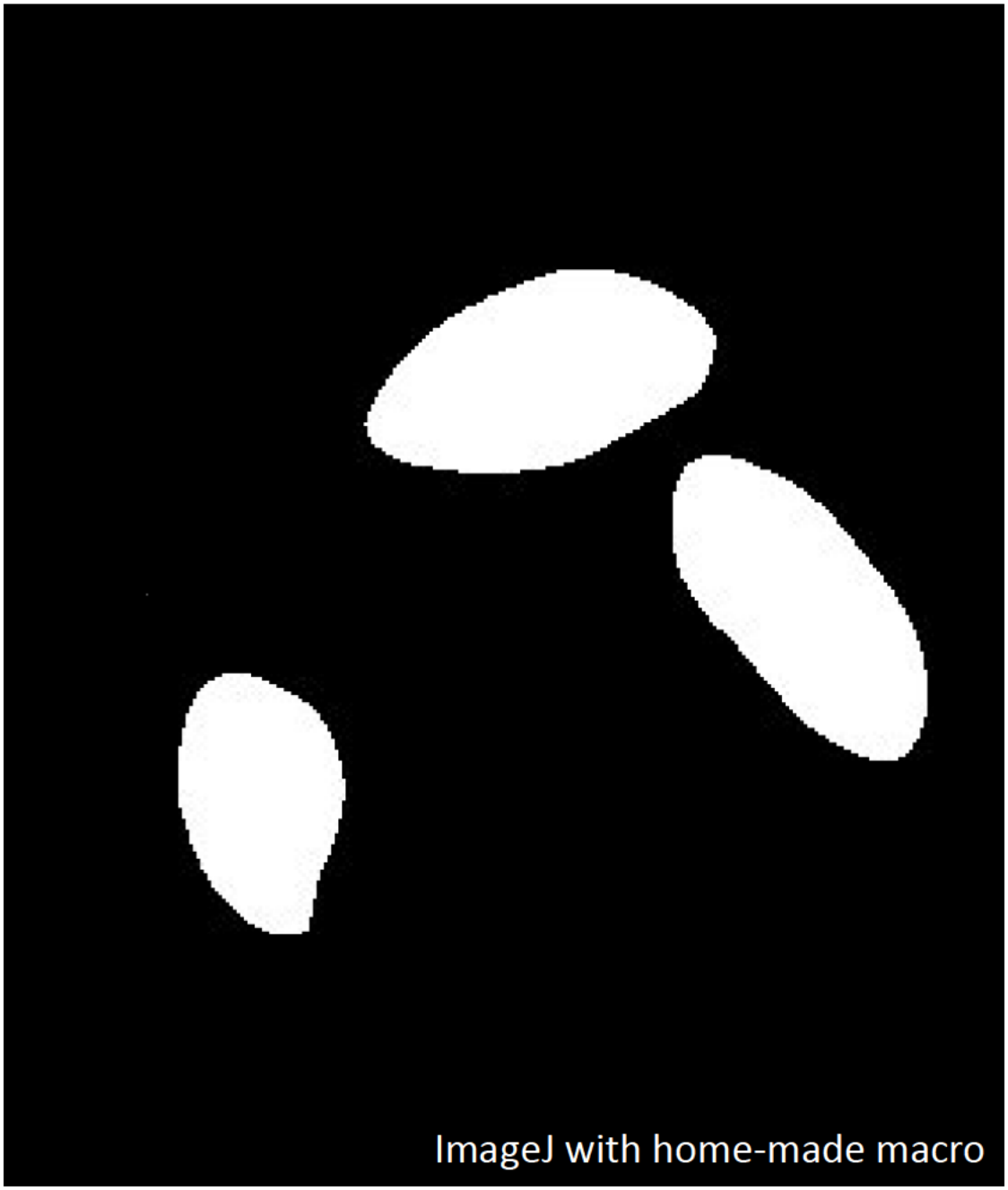
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

*Processing_macro.ijm

```
1  run("8-bit");
2
3  run("Next Slice [>]");run("Next Slice [>]");
4  run("Next Slice [>]");run("Next Slice [>]");
5  run("Next Slice [>]");run("Next Slice [>]");
6  run("Next Slice [>]");run("Next Slice [>]");
7  run("Next Slice [>]");run("Next Slice [>]");
8  run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast...");
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity e
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch Kill persistent Show Errors Clear



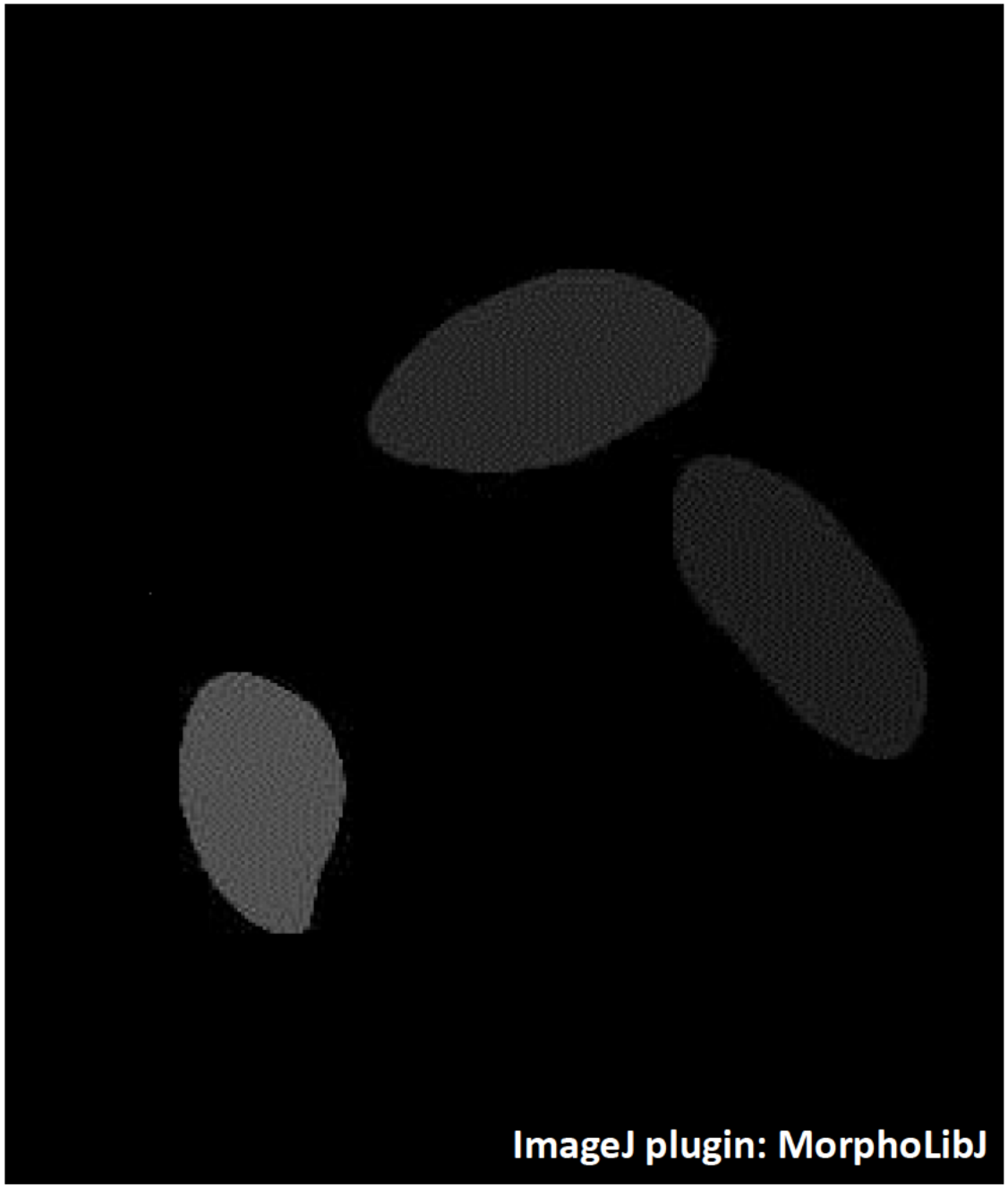
*Processing_macro.ijm

File Edit Language Templates Run Tools Tabs Options

*Processing_macro.ijm

```
1  run("8-bit");
2
3  run("Next Slice [>]");run("Next Slice [>]");
4  run("Next Slice [>]");run("Next Slice [>]");
5  run("Next Slice [>]");run("Next Slice [>]");
6  run("Next Slice [>]");run("Next Slice [>]");
7  run("Next Slice [>]");run("Next Slice [>]");
8  run("Next Slice [>]");run("Next Slice [>]");
9
10 //run("Brightness/Contrast...");
11 setMinAndMax(0, 64000);
12
13
14 run("Convert to Mask",
15 "method=Li background=Dark");
16
17 run("Invert", "stack");
18 run("Erode", "stack");
19 run("Dilate", "stack");
20 run("Dilate", "stack");
21 run("Fill Holes", "stack");
22 run("Erode", "stack");
23 run("Invert", "stack");
24
25
26 run("Connected Components Labeling",
27 "connectivity=26 type=[8 bits]");
28
29 run("Analyze Regions 3D",
30 "volume surface_area mean_breadth sphericity
31
32
33 saveAs("Results");
34 saveAs("Tiff");
35 run("Close All");
```

Batch Kill persistent Show Errors Clear

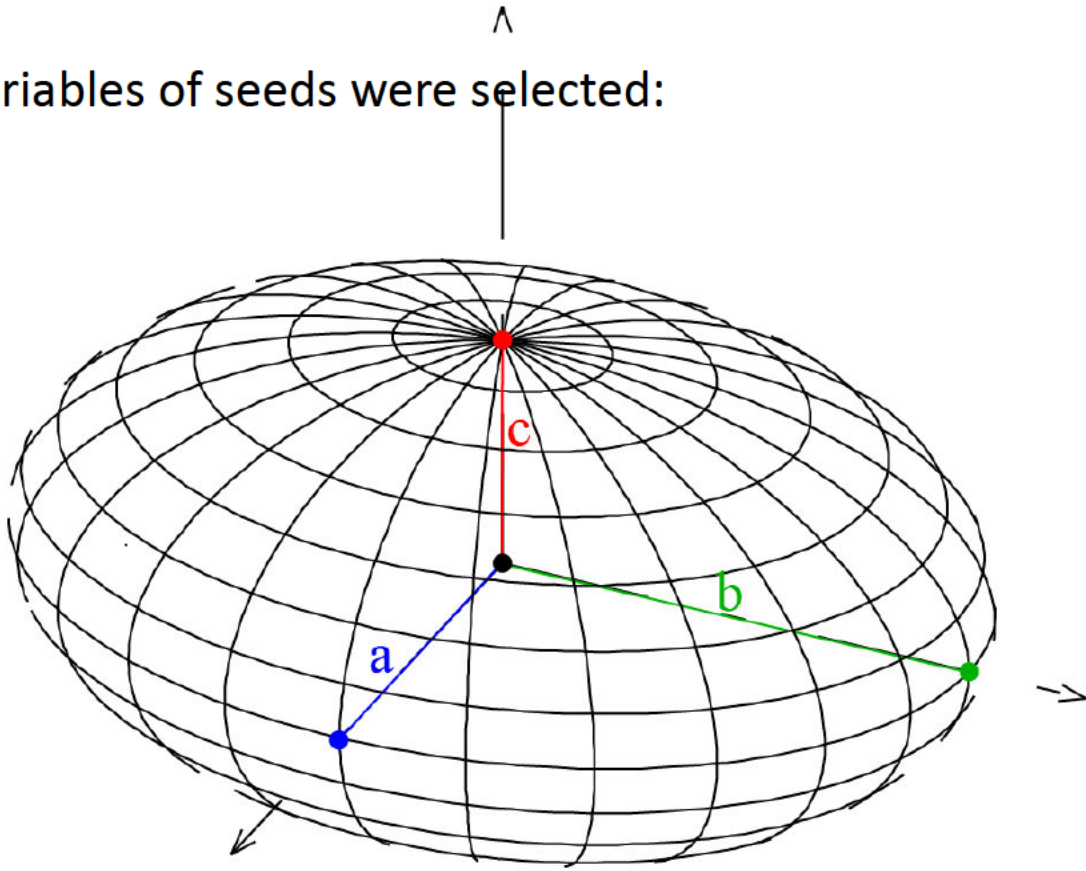


Variables & Sampling

For this project, five morphological variables of seeds were selected:

1) Seed volume
2) Seed surface
3) Fitted ellipsoid radius 1 (**a**)
4) Fitted ellipsoid radius 2 (**b**)
5) Fitted ellipsoid radius 3 (**c**)

a > **b** > **c**



Genus	domesticate	wild	wild_relative
Lathyrus	22	349	42
Lens	19	71	97
Pisum	29	0	0
Vicia	35	442	15

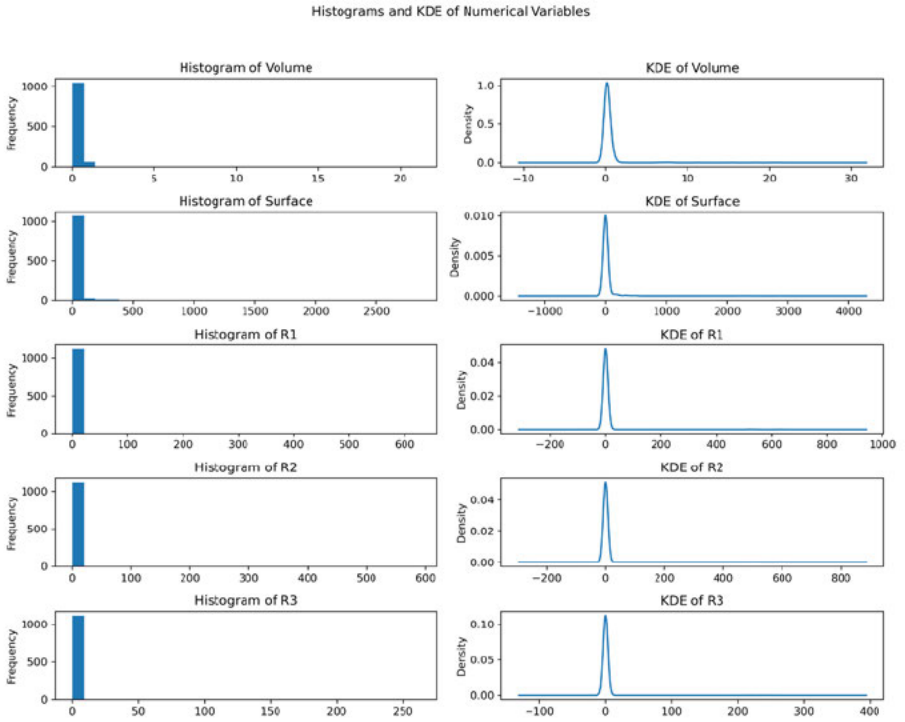
MorphoLibJ can, among other functions model the best fitting ellipsoid on the seeds

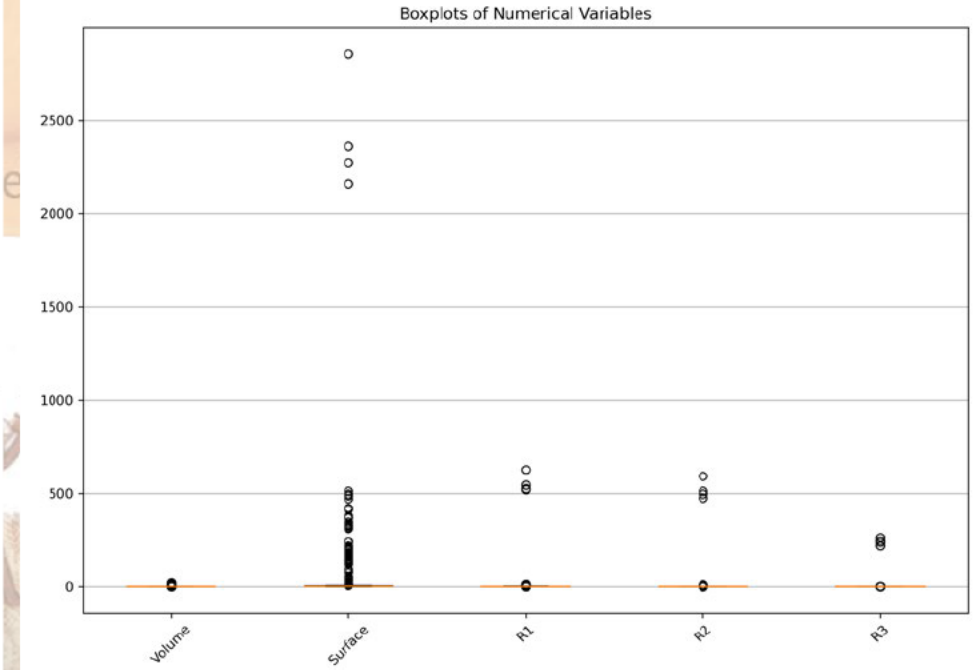
ImageJ plugin: MorphoLibJ

Exploratory & Descriptive Statistics

summary_statistics

	count	mean	std	min	25 %	50 %	75 %	max	skewness	kurtosis
Volume	1121.0	0.36717648145115966	1.305946527761704	1.82e-06	0.081070254	0.167251036	0.357289921	21.18948939	11.885008134613416	159.4513013098811
Surface	1121.0	21.043653425181983	153.5104821830469	0.000722492	1.111470709	1.709605634	3.07659192	2854.976568	14.254893636870085	225.28185626744795
R1	1121.0	2.8404765330918824	33.159058808582174	0.01071862	0.36737438	0.451246266	0.599969944	627.677985	16.761042843758823	282.21865153978246
R2	1121.0	2.5261981063630685	31.047535151458945	0.01071862	0.327794306	0.391379842	0.51857484	590.701925	16.813085565099282	283.8840101069201
R3	1121.0	1.1040347338929528	14.378174152964226	0.0	0.1474451	0.241722524	0.326241456	262.917655	16.786339761365596	281.586054749796

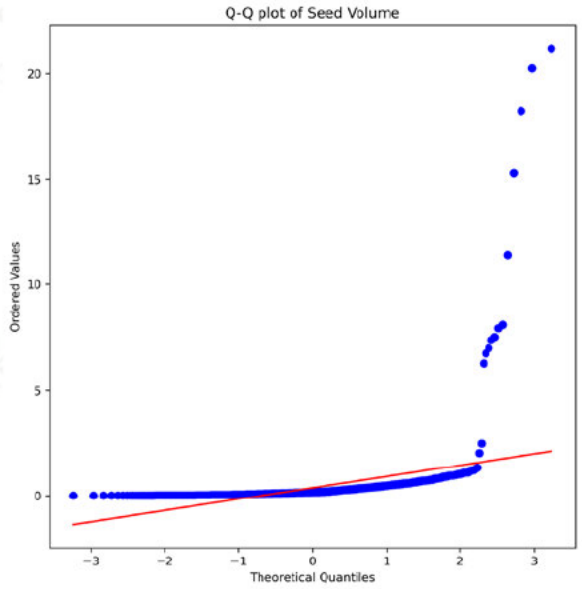


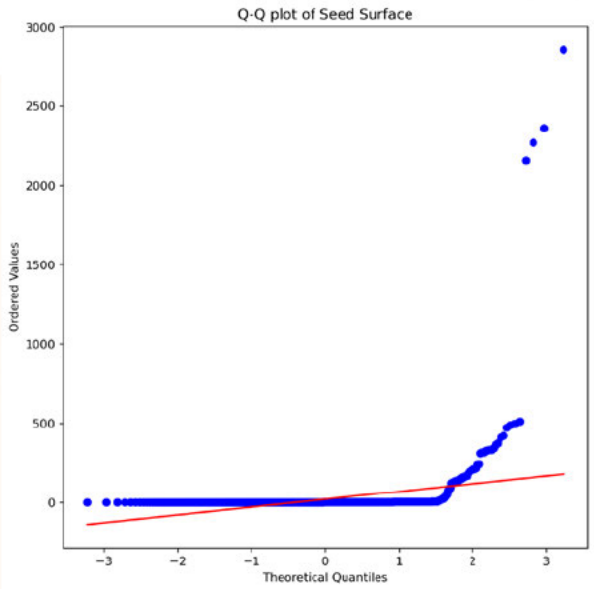


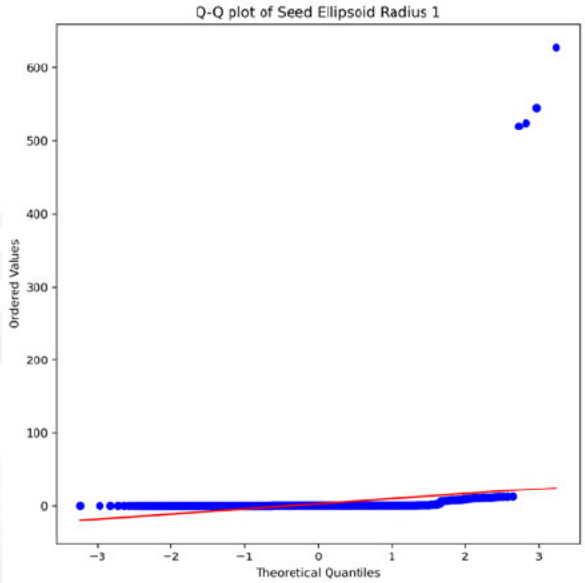
Painting from the tomb of Menna, agricultural scribe of Pharaoh

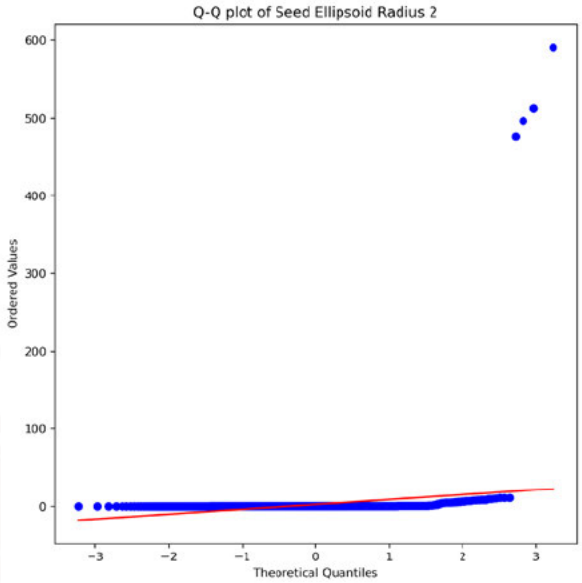
Exploratory & Descriptive Statistics

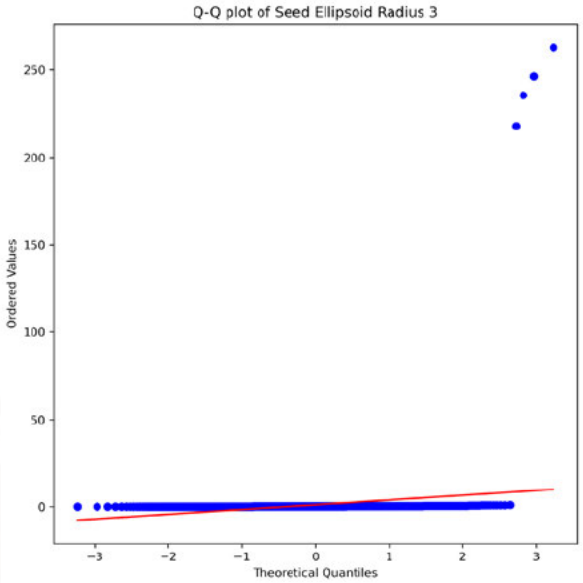
Q-Q Plot



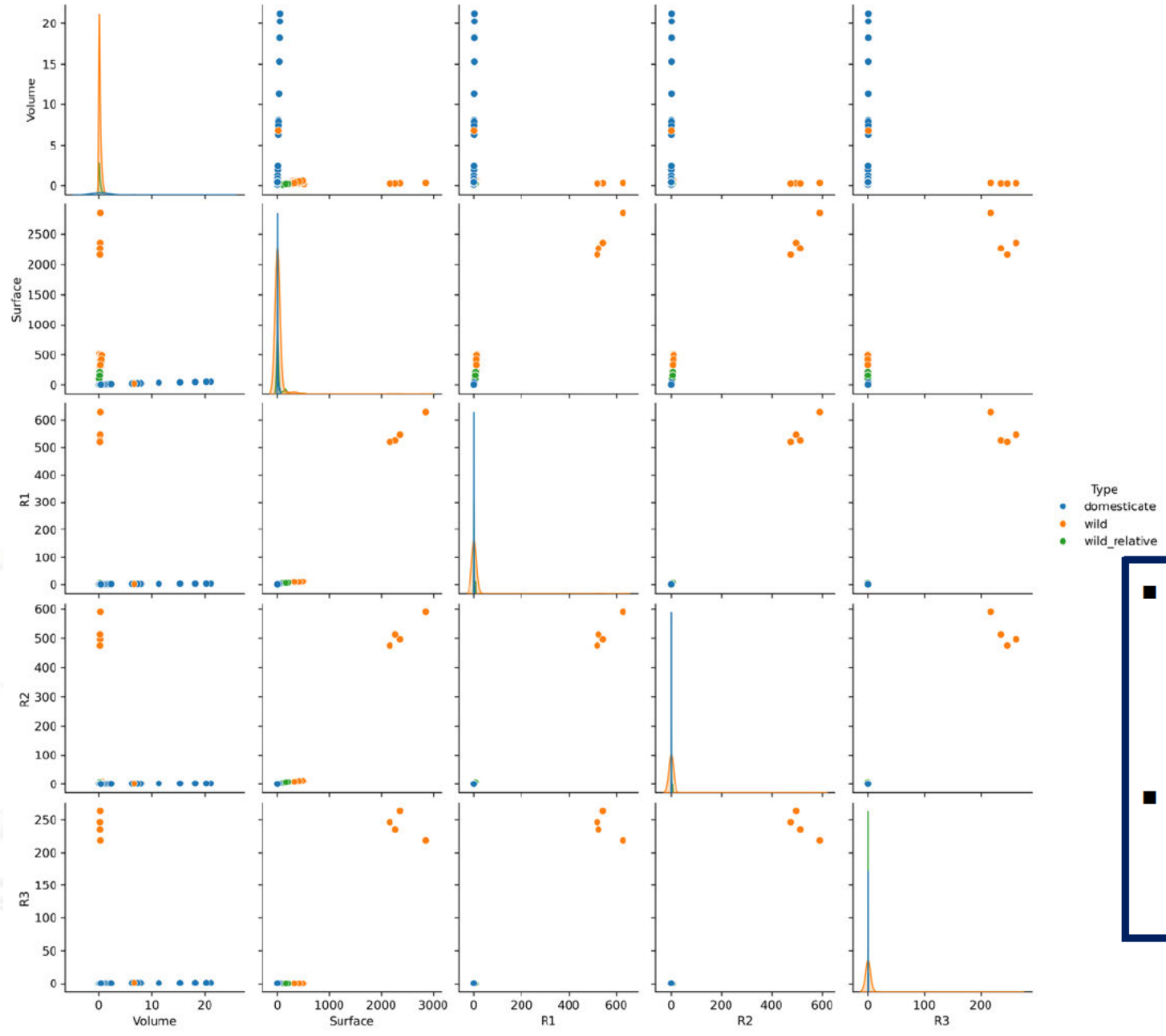








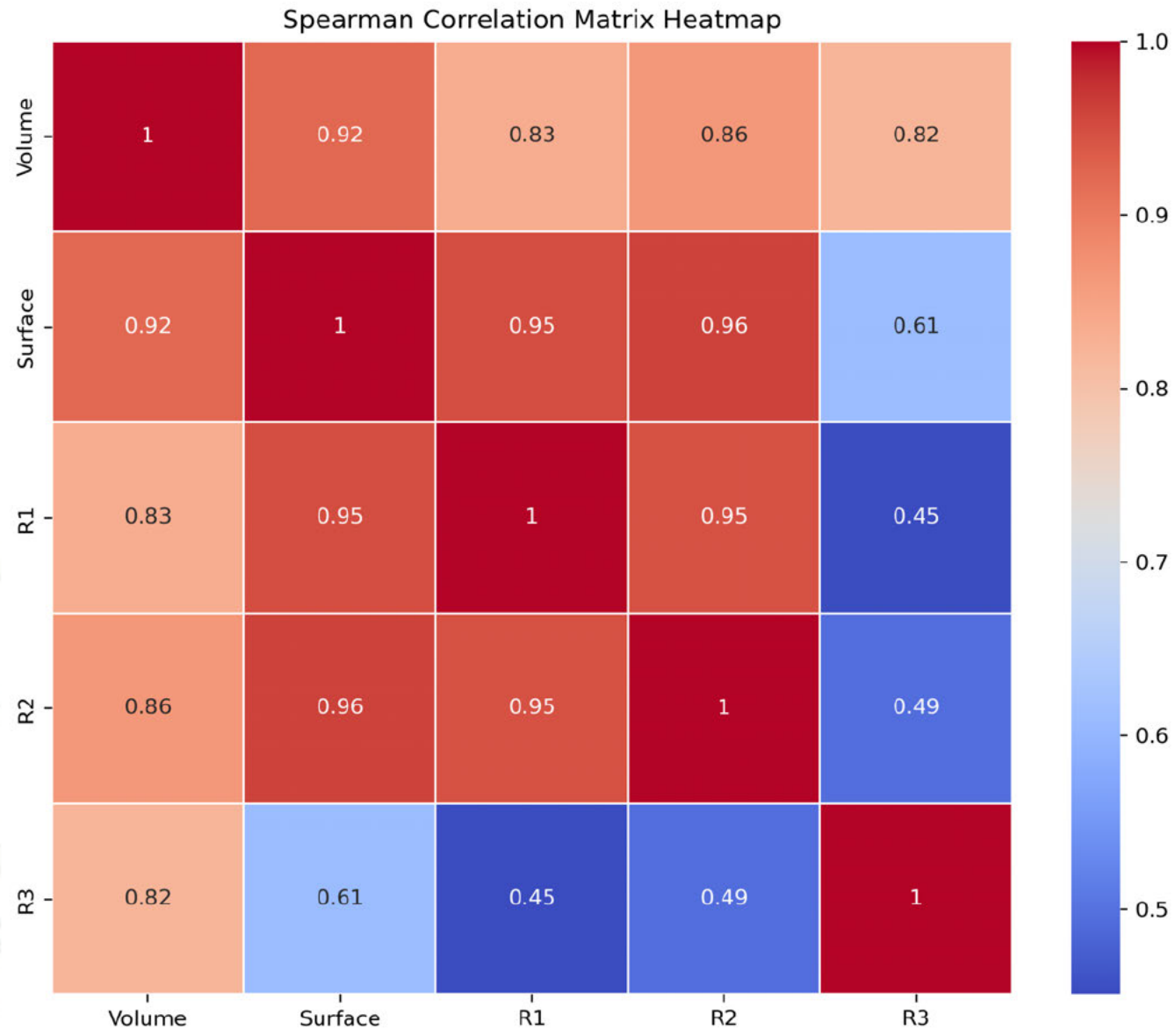
Exploratory & Descriptive Statistics



Pair Plot

- The data contains outliers among wild species
- Volume seems to contain most of the variation

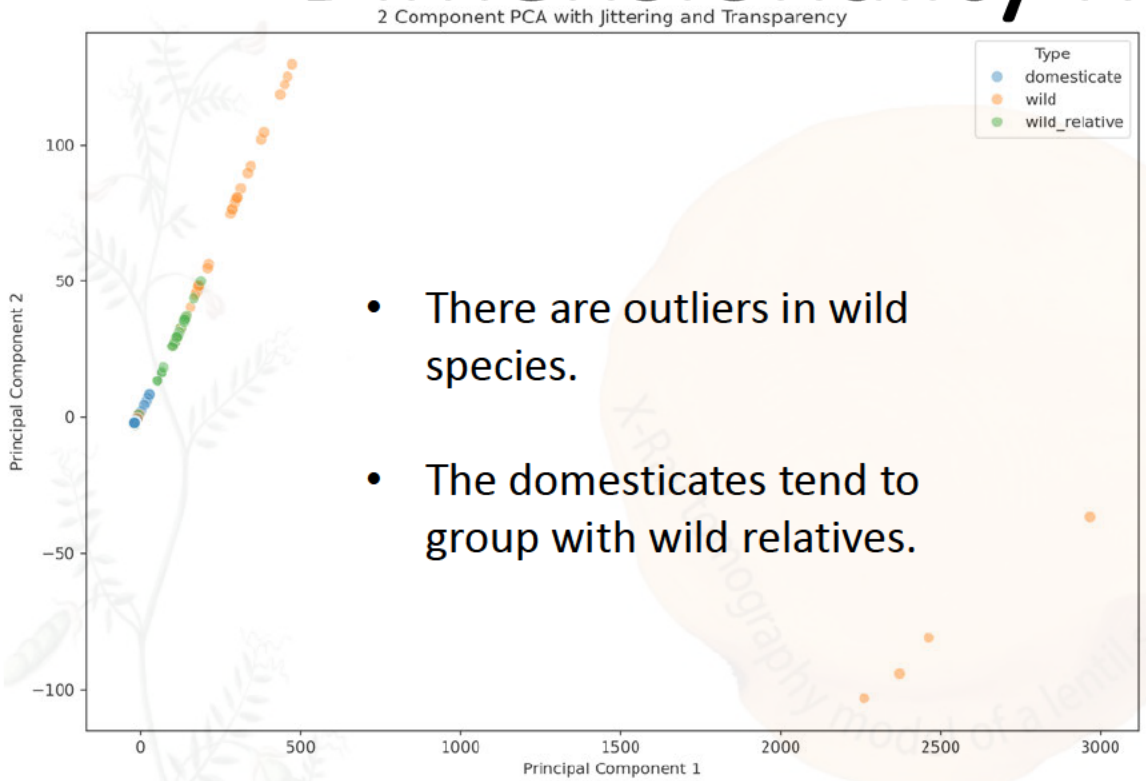
Correlations of Variables



Most variables are strongly correlated With each other, except R3 which quantifies the thickness of the seed.

Painting from the tomb of Menna, agricultural scribe of Pharaoh

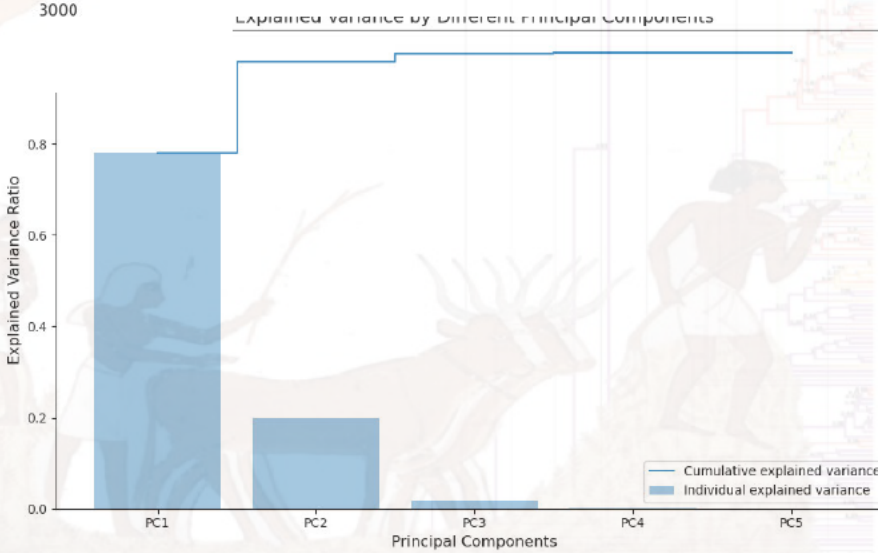
Dimensionality Reduction



- There are outliers in wild species.
- The domesticates tend to group with wild relatives.

	Volume	Surface	R1	R2	R3
PC1	0.003078	0.490235	0.505095	0.504593	0.499925
PC2	0.999901	0.010402	-0.004430	-0.006337	-0.005485
PC3	-0.013313	0.845547	-0.164571	-0.190315	-0.470712
PC4	0.003269	-0.208335	0.395399	0.524535	-0.724644
PC5	0.000812	0.034755	-0.749290	0.658779	0.058022

- The loading on PC1 are the surface, R1, R2, and R3 (ca. 80% variation).
- The loading on PC2 is the volume (ca. 20% of variation).



Painting from the tomb of Menna, agricultural scribe of Pharaoh

Hypotheses Testing

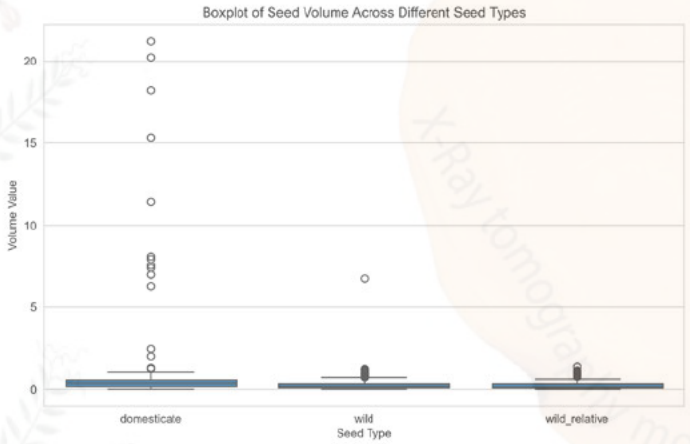
To investigate the hypotheses considering the non-normal distribution of the data, non-parametric statistical tests were carried out:

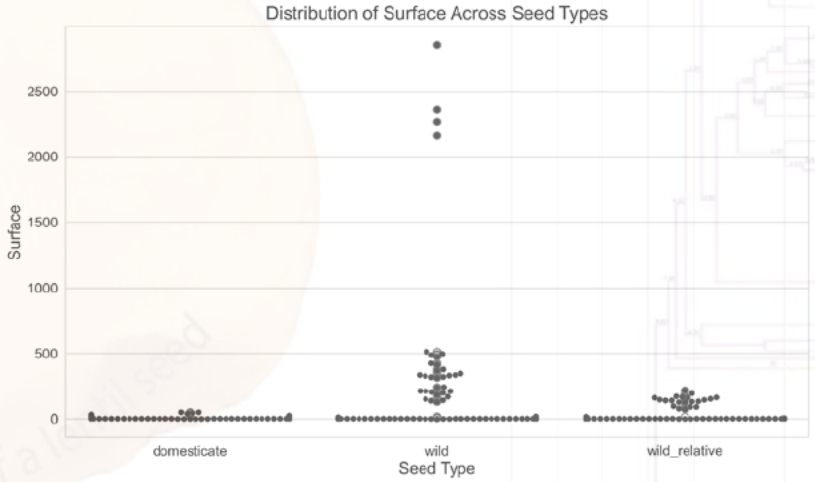
1) Kruskal-Wallis H-test which is used to compare three or more independent samples.

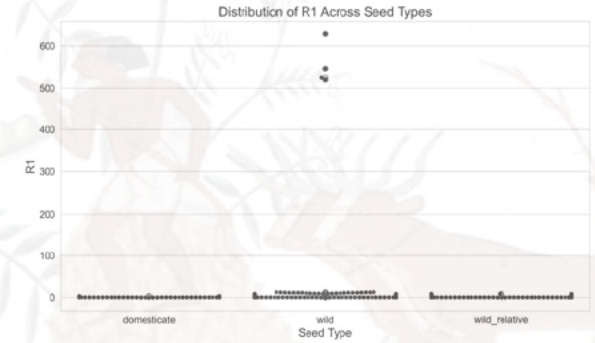
2) Mann-Whitney U test which is used for determining whether two independent samples were drawn from a population with the same distribution.

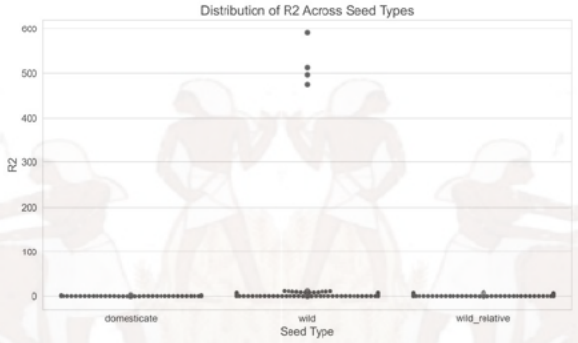
Hypotheses Testing

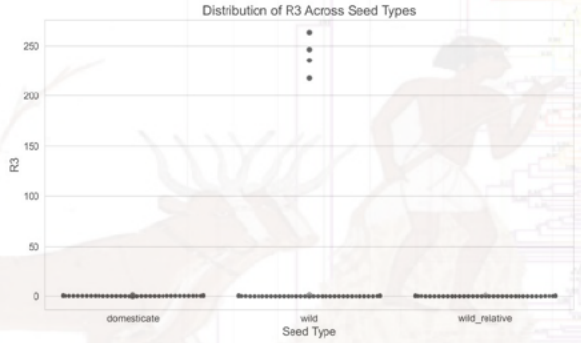
Q1: Do the morphological variables differ significantly among the three seed types (domesticate, wild relative, and wild species)?











Painting from the tomb of Menna, agricultural scribe of Pharaoh

Hypotheses Testing

Q1: Do the morphological variables differ significantly among the three seed types (domesticate, wild relative, and wild species)?

- *H*0: There is no difference in [morphological variable] across different seed types.
- *Ha*: There is a difference in [morphological variable] across different seed types.

kruskal_results

	Statistic	p-value
Volume	35.609249371729085	1.851610976155955e-08
Surface	25.628994166410322	2.721038134967955e-06
R1	23.63835721240094	7.362002332333 1e-06
R2	37.37842816608115	7.644999346542194e-09
R3	44.723763318846565	1.9424873101891612e-10

Our analyses reject the null hypothesis and accept that there are differences across groups for all variables.

Given that $p < 0.05$, I proceed to post-hoc analysis to know which groups significantly differ from each other.

Painting from the tomb of Menna, agricultural scribe of P

Hypotheses Testing

Q1: Do the morphological variables differ significantly among the three seed types (domesticate, wild relative, and wild species)?

posthoc_results_Volume

	domesticate	wild	wild_relative
domesticate	1.0	1.5972274441724896e-07	5.048144349941972e-08
wild	1.5972274441724896e-07	1.0	0.24864524937954008
wild_relative	5.048144349941972e-08	0.24864524937954008	1.0

posthoc_results_Surface

	domesticate	wild	wild_relative
domesticate	1.0	2.320328254633047e-06	0.013322566436874968
wild	2.320328254633047e-06	1.0	0.2545792590074899
wild_relative	0.013322566436874968	0.2545792590074899	1.0

posthoc_results_R1

	domesticate	wild	wild_relative
domesticate	1.0	1.144130111716078e-05	0.05960613620866917
wild	1.144130111716078e-05	1.0	0.1095706071396374
wild_relative	0.05960613620866917	0.1095706071396374	1.0

posthoc_results_R2

	domesticate	wild	wild_relative
domesticate	1.0	2.6504097181519644e-07	0.18661154147056103
wild	2.6504097181519644e-07	1.0	0.0008760299449781818
wild_relative	0.18661154147056103	0.0008760299449781818	1.0

posthoc_results_R3

	domesticate	wild	wild_relative
domesticate	1.0	0.00010592241814652835	1.1233064148379653e-10
wild	0.00010592241814652835	1.0	8.565281770459626e-06
wild_relative	1.1233064148379653e-10	8.565281770459626e-06	1.0

Summary:

- The "domesticate" seed type is significantly different from both the "wild" and "wild relative" seed types, except for R1 &R2 of wild relatives.

- No significant difference is observed between the "wild" and "wild relative" seed types except for R2 & R3.

Hypotheses Testing

Q2-1: Are the distribution of seed variables the same between domesticate and wild relative Types?

Q2-2: Are the distribution of seed variables the same between wild relative and wild Types?

Hypotheses Testing

Q2-1: Are the distribution of seed variables the same between domesticate and wild relative Types?

mannwhitney_results_domesticate_vs_wi

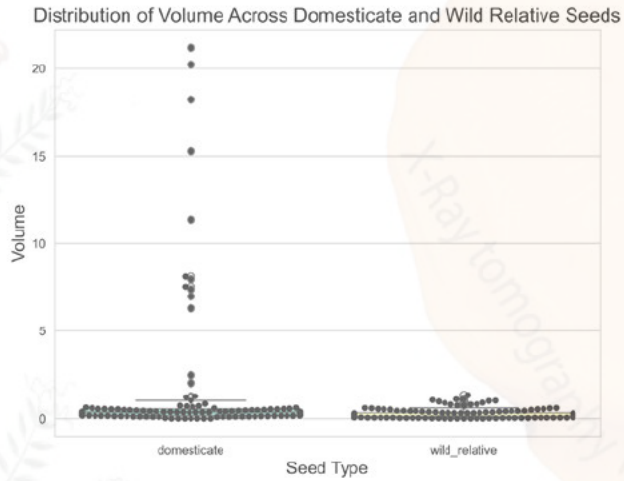
Variable	Statistic	p_value
Volume	11343.0	3.7247067746726724e-08
Surface	9509.0	0.016174553896296563
R1	9236.0	0.0519280192956 4013
R2	9062.0	0.0989913589559 0888
R3	11689.0	1.143251527473281e-09

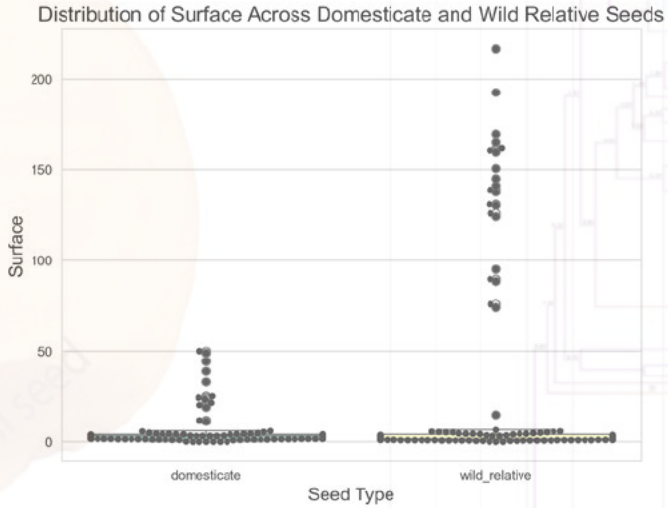
Summary:

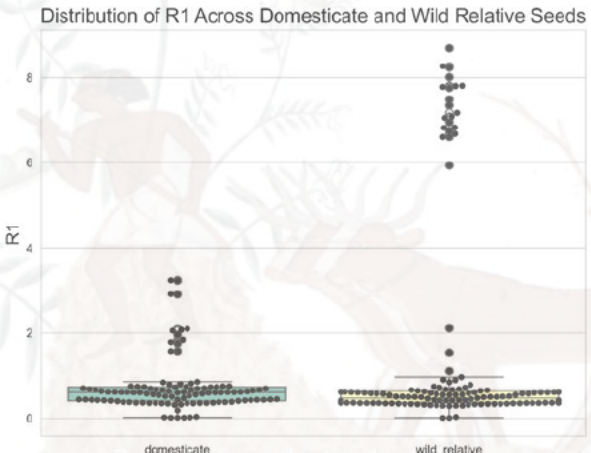
- **Volume** and **R3** have very small p-values, suggesting strong evidence of differing distributions between the two seed types for these variables.
- **Surface** also shows a significant difference but with a larger (yet still below 0.05) p-value, suggesting moderate evidence of differing distributions.
- **R1** and **R2** do not show statistically significant differences in the distributions between the two seed types.

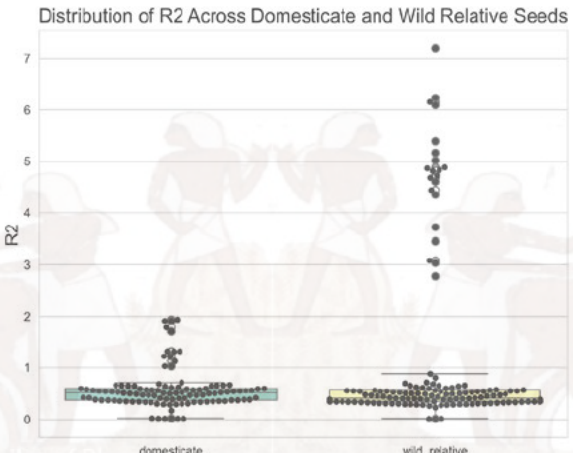
Hypotheses Testing

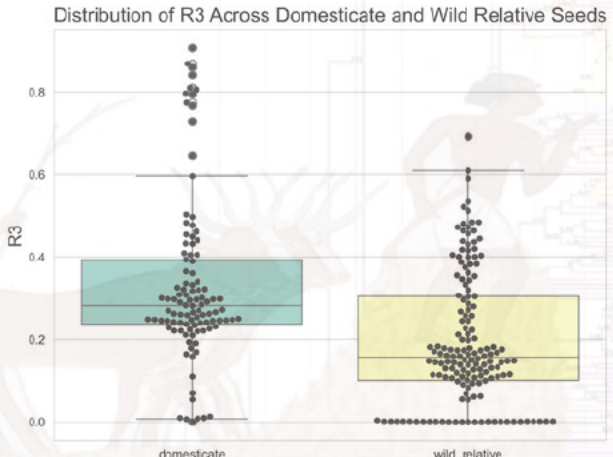
Q2-1: Are the distribution of seed variables the same between domesticate and wild relative Types?











Hypotheses Testing

Q2-2: Are the distribution of seed variables the same between wild relative and wild Types?

mannwhitney_results_wild_vs_wild_relati

Variable	Statistic	p_value
Volume	72267.0	0.0789564081806023
Surface	60852.0	0.0997298965376448
R1	59586.0	0.0430106886578626 4
R2	54350.0	0.00033752726385039055
R3	82379.0	1.8282534430370777e-06

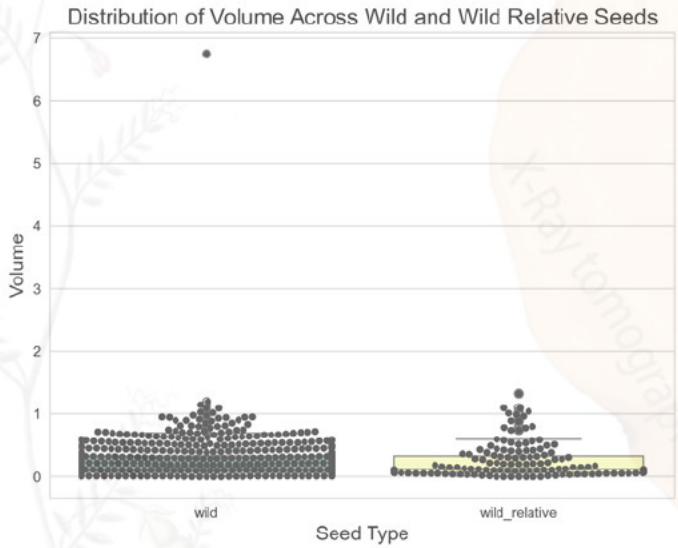
Summary:

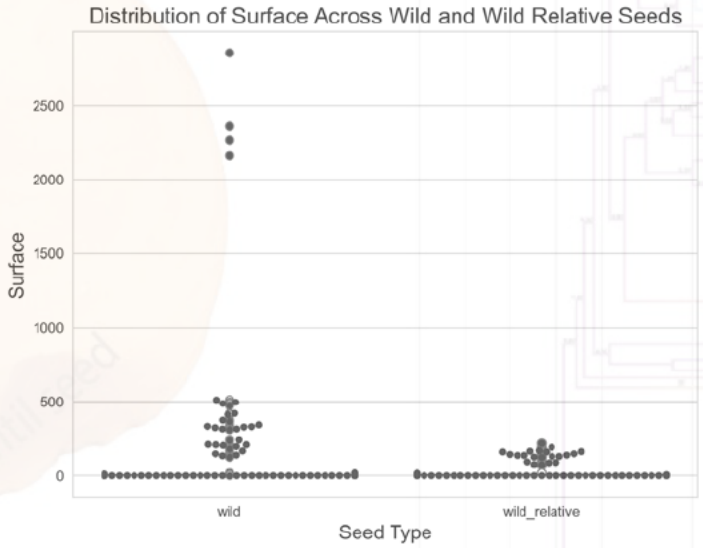
• Variables **R1**, **R2**, and **R3** demonstrate statistically significant differences between the two groups being compared.

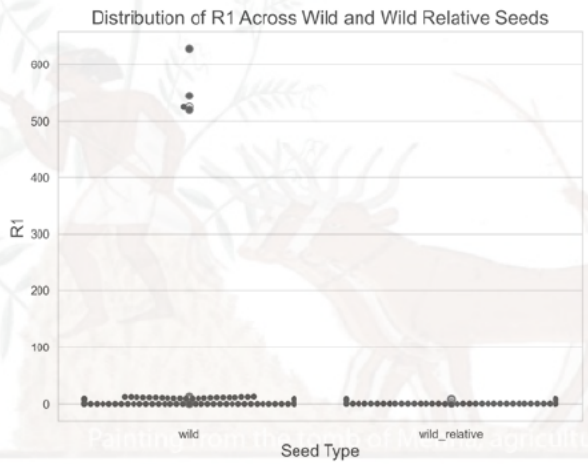
• Variables **Volume** and **Surface** do not show statistically significant differences based on the conventional 0.05 threshold.

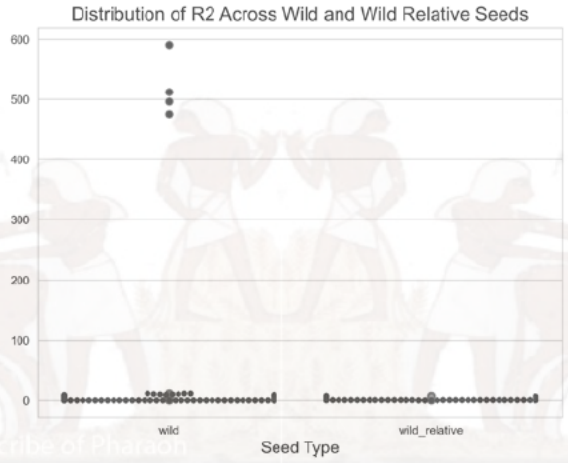
Hypotheses Testing

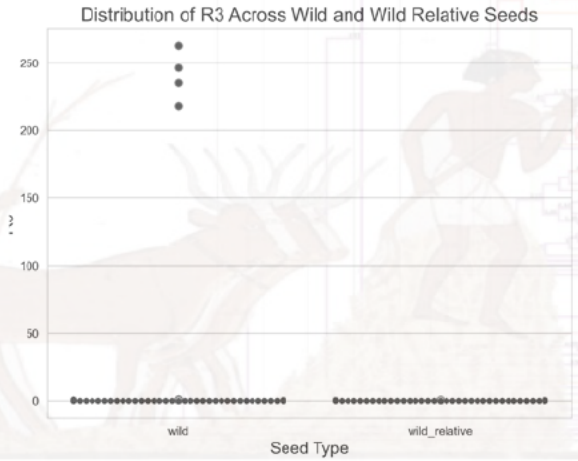
Q2-2: Are the distribution of seed variables the same between wild relative and wild Types?











Hypotheses Testing

a **logistic regression model** was carried out to understand if some traits made some species suitable to be the target of domestication (domesticate Type were removed from the data).

Can Morphological Variables Predict domestication?

Logit Regression Results

Dep. Variable:	IsWildRelative	No. Observations:	1016
Model:	Logit	Df Residuals:	1010
Method:	MLE	Df Model:	5
Date:	Sun, 08 Oct 2023	Pseudo R-squ.:	0.09378
Time:	14:32:48	Log-Likelihood:	-391.70
converged:	True	LL-Null:	-432.24
Covariance Type:	nonrobust	LLR p-value:	5.017e-16

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5713	0.126	-12.495	0.000	-1.818	-1.325
Volume	1.3184	0.465	2.837	0.005	0.407	2.229
Surface	-0.0834	0.016	-5.363	0.000	-0.114	-0.053
R1	0.2983	0.257	1.159	0.246	-0.206	0.803
R2	2.4684	0.547	4.513	0.000	1.396	3.540
R3	-6.4733	1.062	-6.094	0.000	-8.555	-4.391

Summary:

- **Model Significance**: The LLR p-value is quite low, suggesting the model as a whole is significant.
- **Effect of Variables**:

✓ Volume, Surface, R2, and R3 are statistically significant predictors of IsWildRelative.

- This results are preliminary and model needs more validation etc.

General Conclusion

- The difference in volume between domesticates and wild species and wild relatives is the expected outcome of selective breeding for yield.

- The fact that the significant differences between wild relatives and other wild species involve ellipsoid radii suggests a role in seed shape, especially tickness (R3).

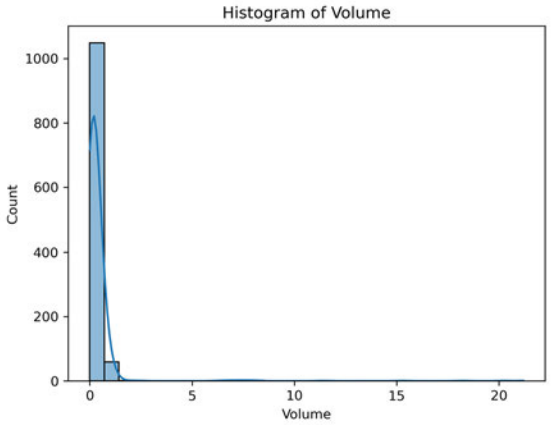
- Taken together, our results suggest that stone age humans had a preference for some types of seeds, but that these preferences were not simply limited to volume, but also included other factors such as seed shape, hinting at the role of other factors such as seed handling and storage.

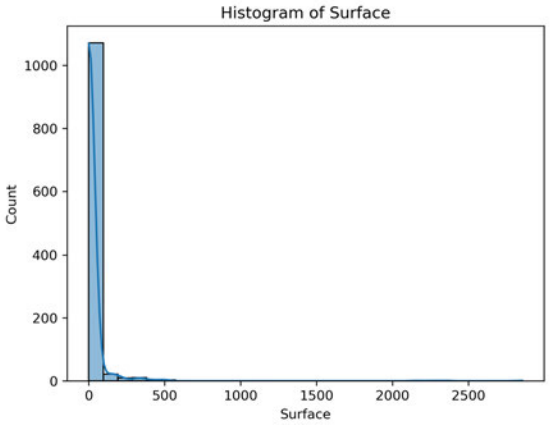
Painting from the tomb of Menna, agricultural scribe of Pharaoh

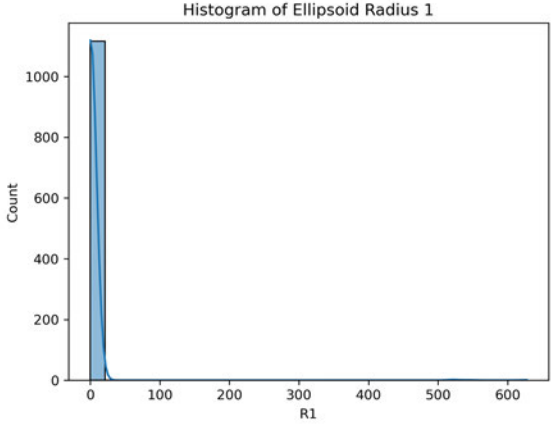
Thank you very much for your attention!

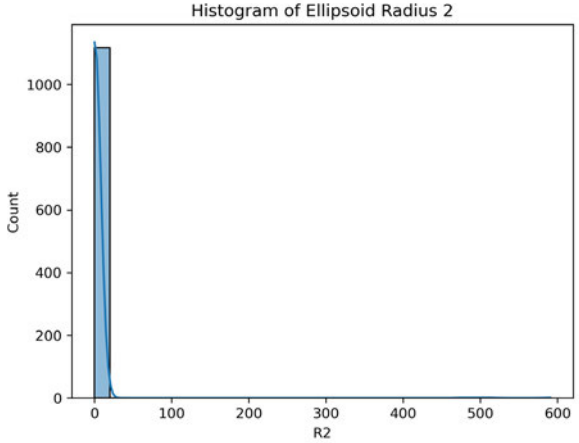
Various libraries were used for this project:

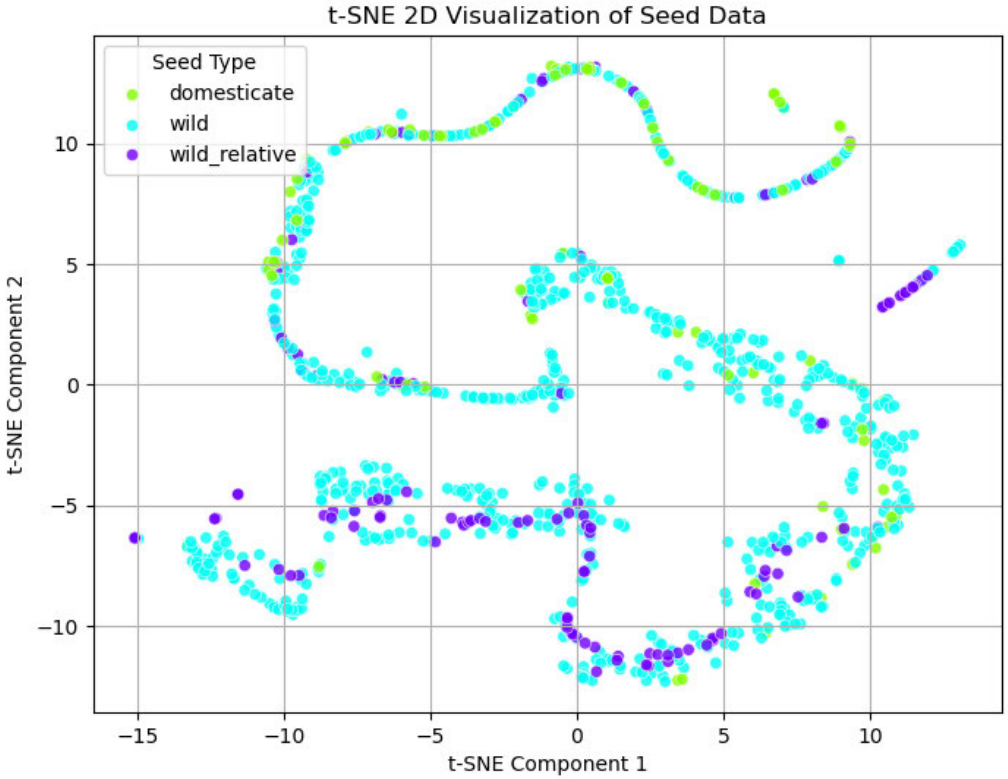
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.preprocessing import StandardScaler
from scipy.stats import Kruskal
from scikit_posthocs import posthoc_dunn
import scipy.stats as stats
import scikit_posthocs as sp
import statsmodels.api as sm
```

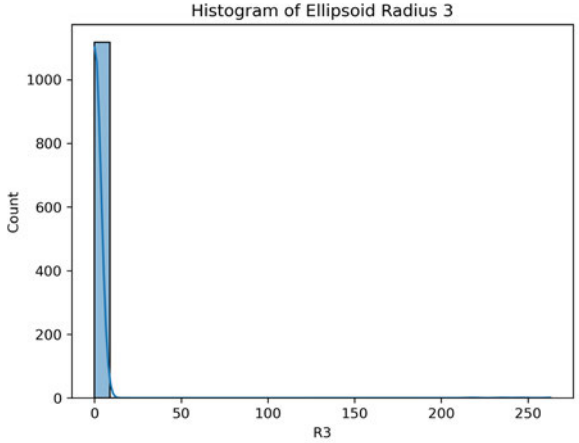

Histogram of Volume
Count
Volume
0
200
400
600
800
1000
0
5
10
15
20


Histogram of Surface
Count
Surface
0
200
400
600
800
1000
0
500
1000
1500
2000
2500


Histogram of Ellipsoid Radius 1
Count
R1
0
200
400
600
800
1000
0
100
200
300
400
500
600


Histogram of Ellipsoid Radius 2
Count
R2
0
200
400
600
800
1000
0
100
200
300
400
500
600


t-SNE 2D Visualization of Seed Data
Seed Type
domesticate
wild
wild_relative
t-SNE Component 2
t-SNE Component 1
10
5
0
−5
−10
−15
−10
−5
0
5
10


Histogram of Ellipsoid Radius 3
Count
R3
0
200
400
600
800
1000
0
50
100
150
200
250

Hypotheses Testing

To conduct a hypothesis test with the goal of determining which variables predict domestication in our seed dataset, a **logistic regression model** was used. This approach is widely used for binary outcome variables.

Can Morphological Variables Predict domestication?

Logit Regression Results

Dep. Variable:	IsDomesticate	No. Observations:	1121
Model:	Logit	Df Residuals:	1115
Method:	MLE	Df Model:	5
Date:	Sun, 08 Oct 2023	Pseudo R-squ.:	0.1028
Time:	14:36:20	Log-Likelihood:	-312.72
converged:	True	LL-Null:	-348.56
Covariance Type:	nonrobust	LLR p-value:	4.548e-14

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5001	0.169	-14.788	0.000	-2.831	-2.169
Volume	1.2369	0.413	2.998	0.003	0.428	2.045
Surface	-0.0502	0.042	-1.187	0.235	-0.133	0.033
R1	-0.1162	1.145	-0.101	0.919	-2.361	2.128
R2	0.9125	1.363	0.669	0.503	-1.759	3.584
R3	-1.4980	1.300	-1.152	0.249	-4.046	1.050

Summary:

- **Model Significance**: The LLR p-value is quite low, suggesting the model as a whole is significant.

- **Effect of Variables**:
  - ✓ **Volume** is the only predictor that provides a significant contribution to explaining the variance in IsDomesticate.
  - ✓ **Surface**, **R1**, **R2**, and **R3** do not appear to significantly contribute to the model.