THE UNIVERSITY *of* EDINBURGH
informatics

Applied Machine Learning (AML)

Class Starting at 4:10pm

Oisin Mac Aodha • Siddharth N.

Applied Machine Learning

Week 8: Clustering and Non-Linear Dimensionality Reduction

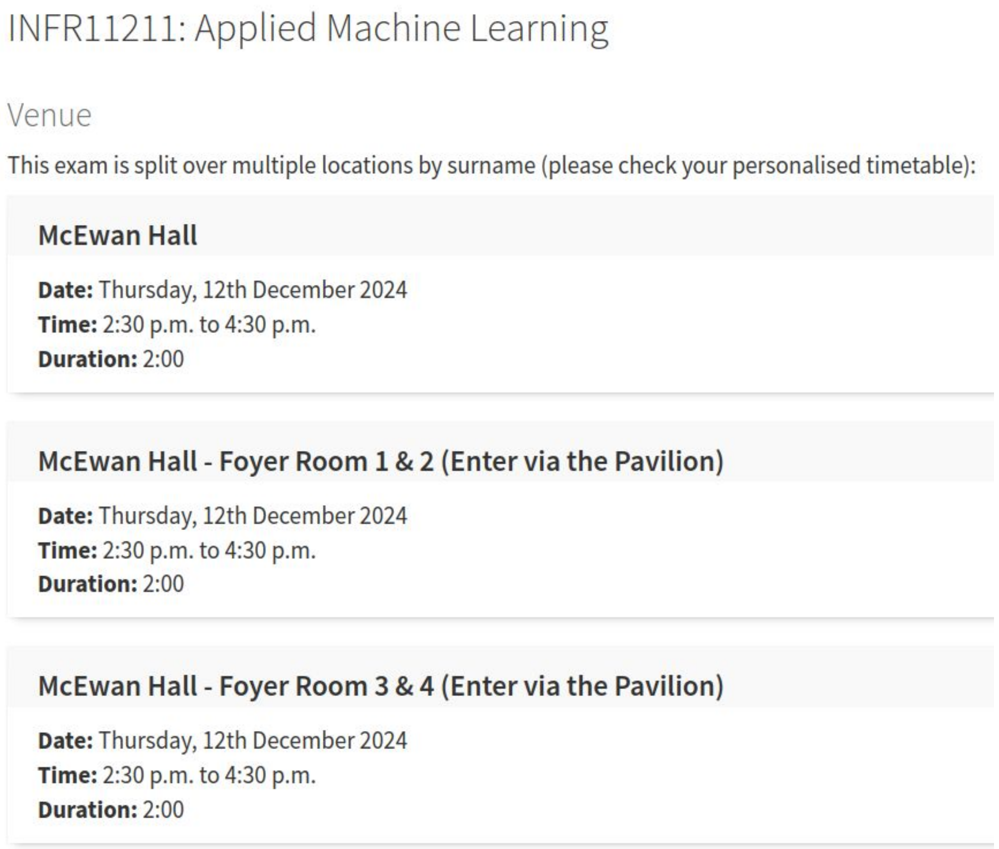
This slides will be made available on the project website after the class.
This session will be recorded.

Overview

1) Outline your tasks this for week
2) Discussion of Week 7’s topics

Exam

- ***Thu, 12th Dec 2024 [14:30-16:30]***
- ***McEwan Hall [Foyer Rooms]***
  - ***split by surname; check personalised timetable***
- On campus and will be **closed book**



INFR11211: Applied Machine Learning

Venue

This exam is split over multiple locations by surname (please check your personalised timetable):

McEwan Hall

Date: Thursday, 12th December 2024
Time: 2:30 p.m. to 4:30 p.m.
Duration: 2:00

McEwan Hall - Foyer Room 1 & 2 (Enter via the Pavilion)

Date: Thursday, 12th December 2024
Time: 2:30 p.m. to 4:30 p.m.
Duration: 2:00

McEwan Hall - Foyer Room 3 & 4 (Enter via the Pavilion)

Date: Thursday, 12th December 2024
Time: 2:30 p.m. to 4:30 p.m.
Duration: 2:00

- Format: 2/3 questions as in IAML (INFR10069) before 2020
- Exams in 2020 and 2121 were “open book” - less relevant
- Past exam papers are available here: https://exampapers.ed.ac.uk
- We do not provide past exam solutions

Coursework Submission

- ***Thu, 21st Nov 2024 - 12:00***
- Instructions for submission - by early next week [week 9]
  - Only report due on 21st
  - Supplementary materials [report LaTeX + code + Readme] (submit at later date - details TBA)

- **NOTE**: Lateness & Extension Policy
  - *No deadline extensions allowed on any account* ***[Rule 2]***
  - See course information page for further details

Week 8: Your tasks for this week

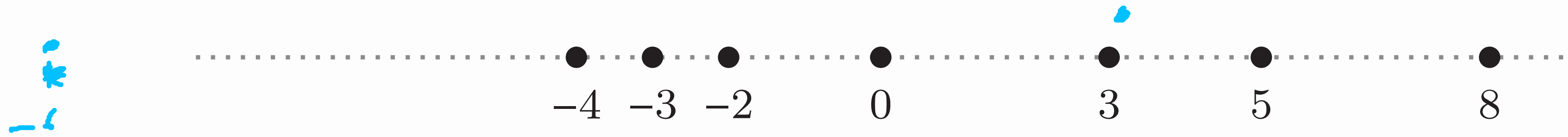
1) Complete Tutorial 3
2) Watch videos for week 8
    - **Recommender Systems** and **Neural Networks**
3) Ask questions on Piazza if stuck
4) Continue working on the coursework
5) Start **Lab 4** which takes places next week - link in week 9

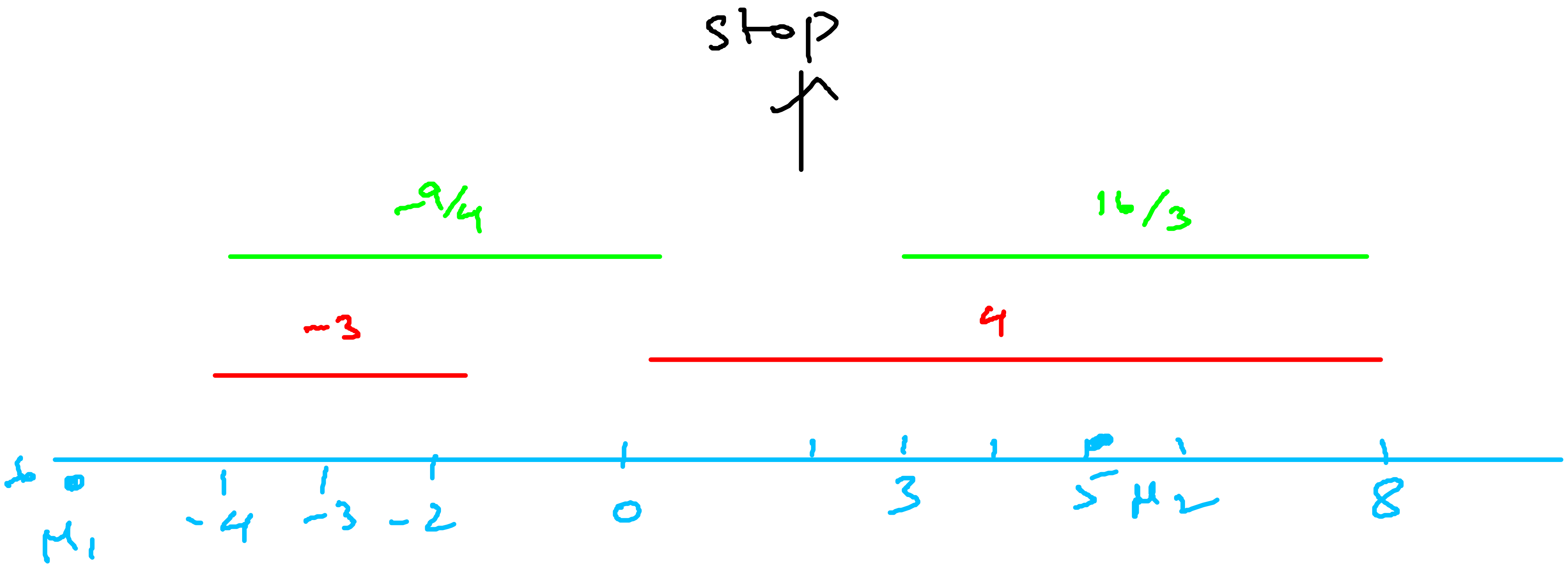
K-Means Example

Consider the following dataset where every instance is represented by a single numeric attribute: $\{-4, -3 - 2, 0, 3, 5, 8\}$. Make a sketch plot of the data.

Run the K-Means clustering algorithm on the data above. Assume $K = 2$ and that the starting means are set as $\mu_1 = -6$, and $\mu_2 = 5$. List the instances in each cluster after the first and second iteration. After how many iterations would you stop the algorithm?

Plotted on a line, the data points would look like this:




stop
-9/4
16/3
-3
4
-4
-3
-2
0
3
5
8
μ1
μ2

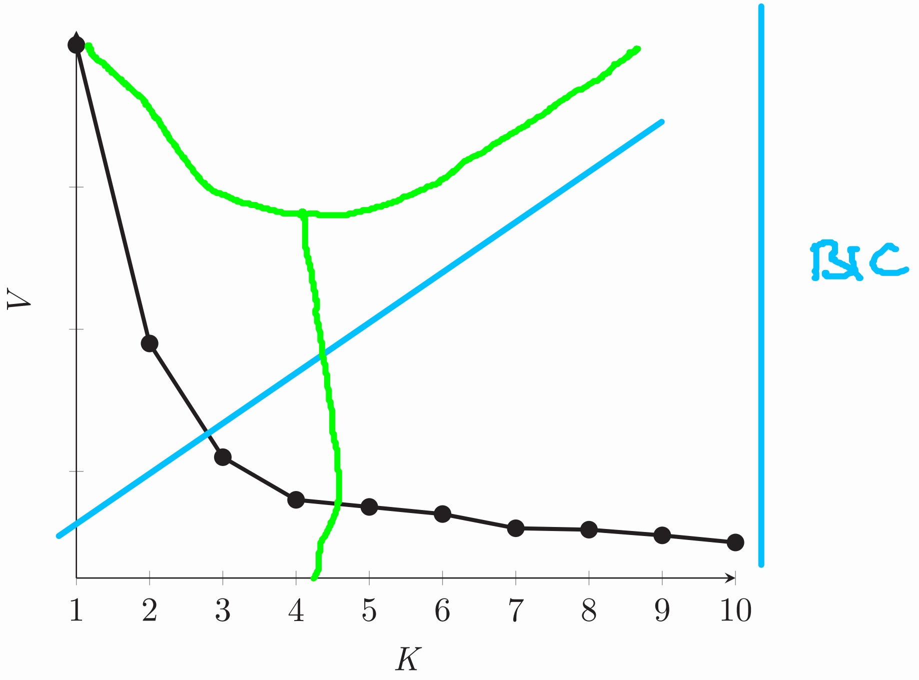
Estimating Number of Clusters

$BIC = K \log N$

How many clusters does your data have?

- Get (K) from class labels (e.g. digits 0…9)

- Find an "appropriate" K: optimise for V
  - Run K-Means for $K = 1, 2, \ldots$; choose K with smallest V
  - **Issue:** What is V when $K = N$?
    – choose best K on *validation* data
  - Choose visually from a ***elbow*** plot
    – point that maximises the 2nd derivative of V



Intrinsic Evaluation: Supervised

Key Idea: Evaluate relationship between *pairs* of data points $\boldsymbol{x}_l, \boldsymbol{x}_m$

Rand Index (RI)

- $+$: $\boldsymbol{x}_l, \boldsymbol{x}_m$ are in the same cluster
- $-$: $\boldsymbol{x}_l, \boldsymbol{x}_m$ are in different clusters

		Predicted (C)	
		+	−
True ($\mathcal{R}$)	+	TP	FN
	−	FP	TN

$$\mathrm{RI} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \text{Accuracy!}$$

Intrinsic Evaluation: Supervised

Issue: Expected value of RI of two *random* partitions $\neq 0$ (or any constant)

Adjusted Rand Index (ARI)

	$\boldsymbol{c}_1$	$\boldsymbol{c}_2$		$\boldsymbol{c}_U$	sum
$\boldsymbol{r}_1$	N_{11}	N_{12}	$\cdots$	N_{1U}	a_1
$\boldsymbol{r}_2$	N_{21}	N_{22}	$\cdots$	N_{2U}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\boldsymbol{r}_V$	N_{V1}	N_{V2}	$\cdots$	N_{VU}	a_V
sum	b_1	b_2	$\cdots$	b_U	N

$N_{ij} = |\boldsymbol{r}_i \cap \boldsymbol{c}_j| \qquad \binom{N}{2} = \frac{N(N-1)}{2}$

$$\text{RI} = \sum_{ij} \binom{N_{ij}}{2}$$

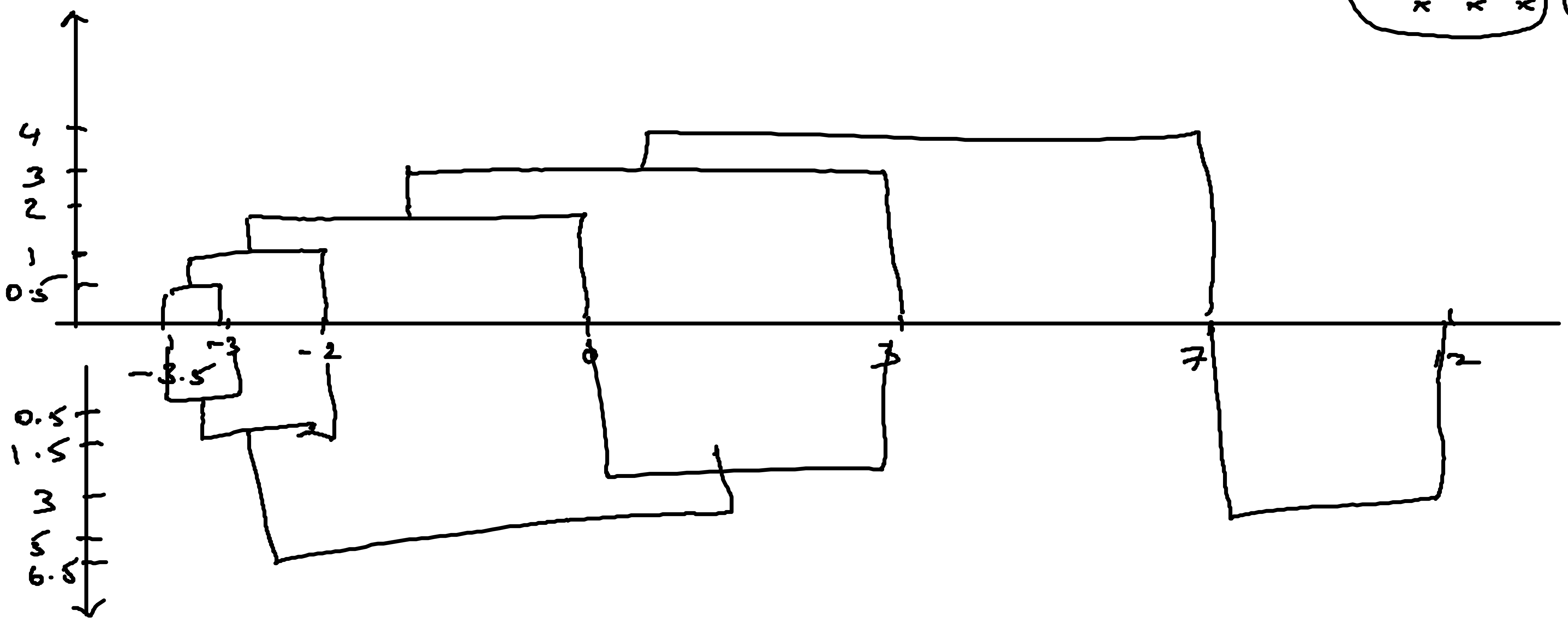
$$\text{Expected RI} = \frac{1}{\binom{N}{2}} \left[\sum_v \binom{a_v}{2} \cdot \sum_u \binom{b_u}{2} \right]$$

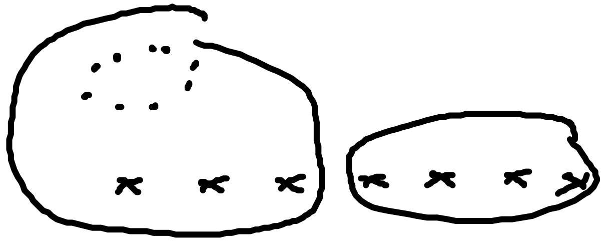
$$\text{Max RI} = \frac{1}{2} \left[\sum_v \binom{a_v}{2} + \sum_u \binom{b_u}{2} \right]$$

$$\text{ARI} = \frac{\text{RI} - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}}$$

Consider the following dataset, where every instance is represented by a single numeric attribute: $\{-3.5, -3, -2, 0, 3, 7, 12\}$

- Run the single-link clustering algorithm on the dataset above until two clusters remain. List the instances in each of the two clusters.
- Run the complete-link clustering algorithm on the dataset above until two clusters remain. List the instances in each of the two clusters.
- Provide a qualitative description of the difference between the two clusterings.


4
3
2
1
0.5
-3.5
-3
-2
0
3
7
12
0.5
1.5
3
5
6.5



Wed demo

- https://pair-code.github.io/understanding-umap/
- https://jlmelville.github.io/uwot/umap-examples.html
- https://distill.pub/2016/misread-tsne/

Visualisation with t-SNE and UMAP

- Hyperparameters *really* matter
  - t-SNE: perplexity
  - UMAP: # neighbours, minimum distance
- Cluster sizes *do not* mean anything
- Cluster distances *may not* mean anything
- Seeing patterns in randomness!

Can be like tasseography—reading tea leaves!