

# Generative AI Models

## ECE 598 LV – Lecture 24

Lav R. Varshney

19 April 2022

- Interpretability

- Why?

- What?

- How?



**EXPLAINABILITY IN DATA SCIENCE: INTERPRETABILITY,  
REPRODUCIBILITY, AND REPLICABILITY**

Payel Das and Lav R. Varshney

# **Explaining Artificial Intelligence Generation and Creativity**

*Human interpretability for novel ideas and artifacts*

(to appear in *IEEE Signal Processing Magazine*)



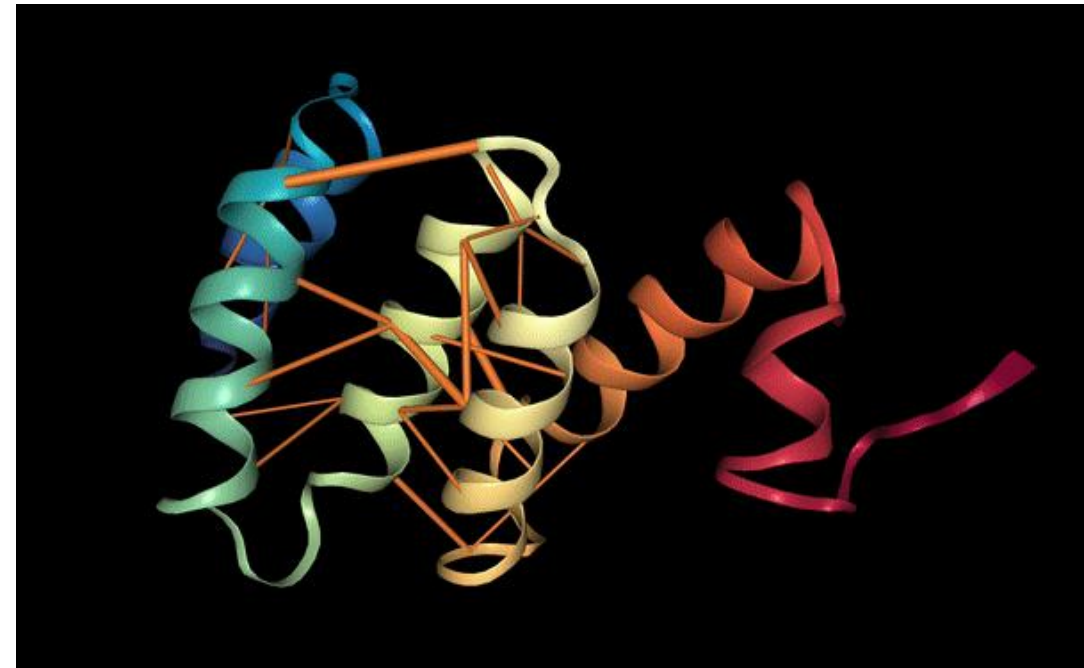
## Dimensions of interpretability [Selbst and Barocas, 2018]

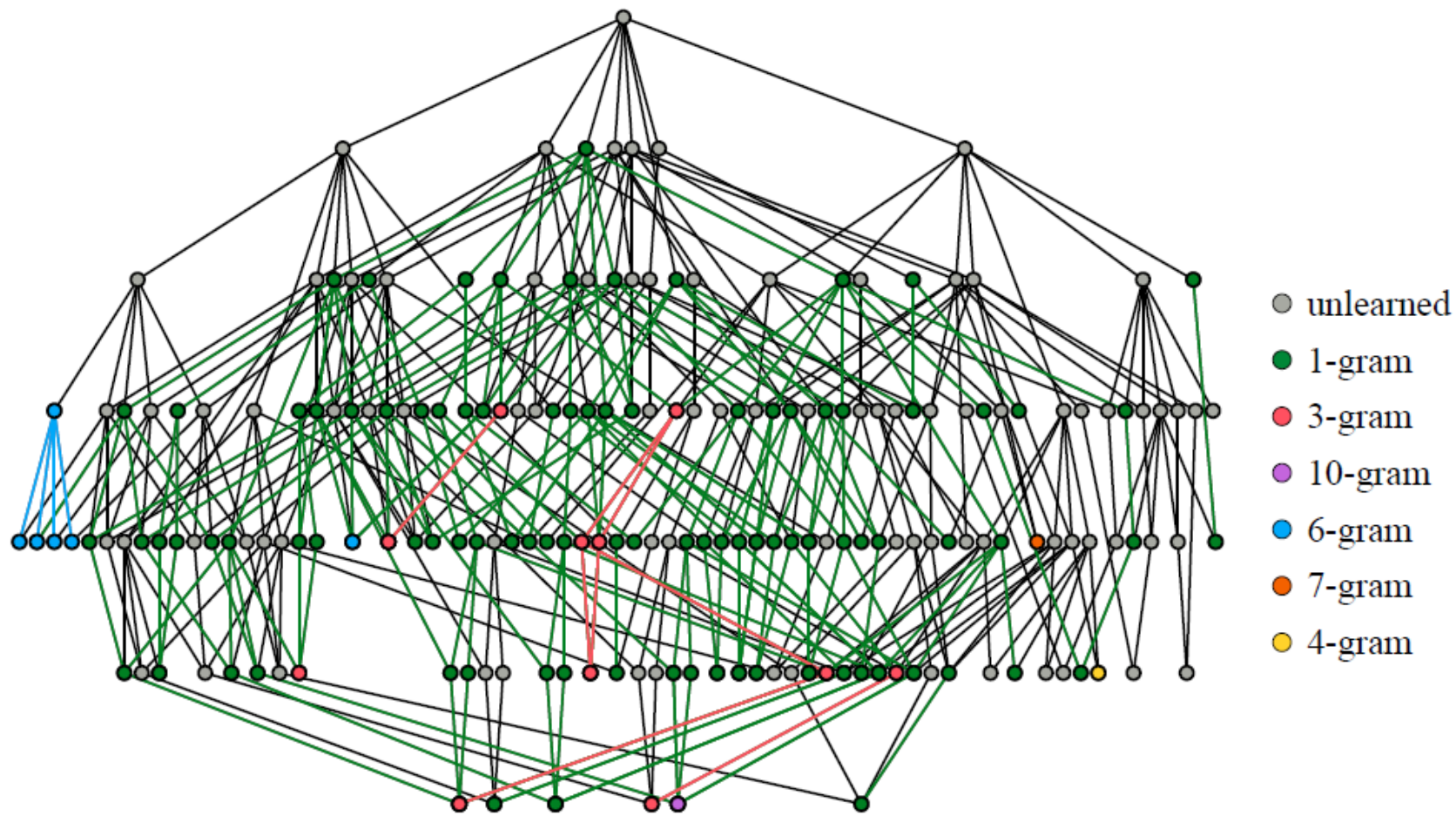
- What sets machine learning models apart from other algorithms are their *inscrutability* and *nonintuitiveness*
  - Inscrutability suggests that models available for direct inspection may defy understanding,
  - Nonintuitiveness suggests that even where models are understandable, they may rest on apparent statistical relationships that defy intuition
  - Most extant work on interpretable ML/AI only addresses inscrutability, but not nonintuitiveness
- Dealing with inscrutability requires providing a sensible description of rules; addressing nonintuitiveness requires providing satisfying explanation for why the rules are what they are

For numerous settings, may need technical solutions to both inscrutability and nonintuitiveness

# BERTology

- <https://www.youtube.com/watch?v=1aaWuAum5HY>
- <https://www.youtube.com/watch?v=pFf4PltQ9LY>



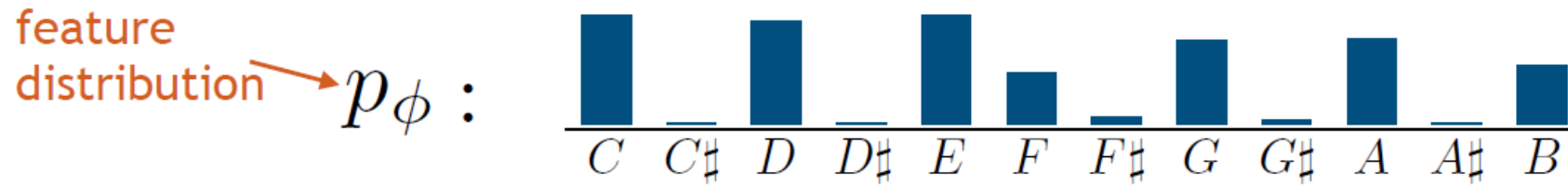


Visualization of Bach's music mind for writing chorales. The underlying directed acyclic graph signifies an upside-down information lattice.

# Simple human-interpretable rules

Compositional Rule Examples:

feature  $\phi$  : pitch class in the soprano voice



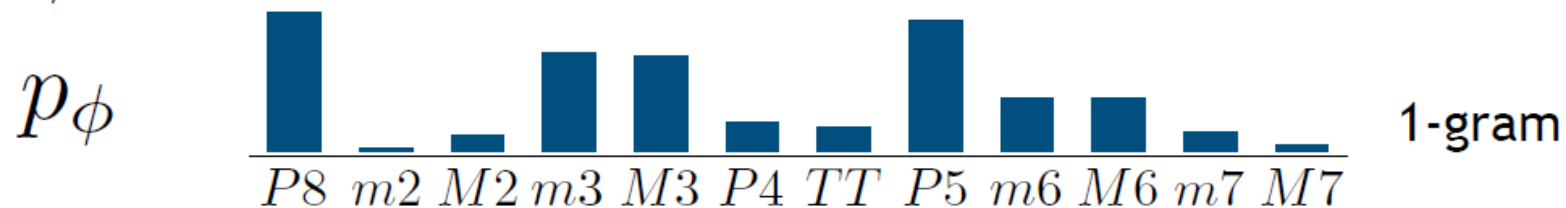
This rule can be interpreted or translated to:

“The soprano voice is built on a diatonic scale.”

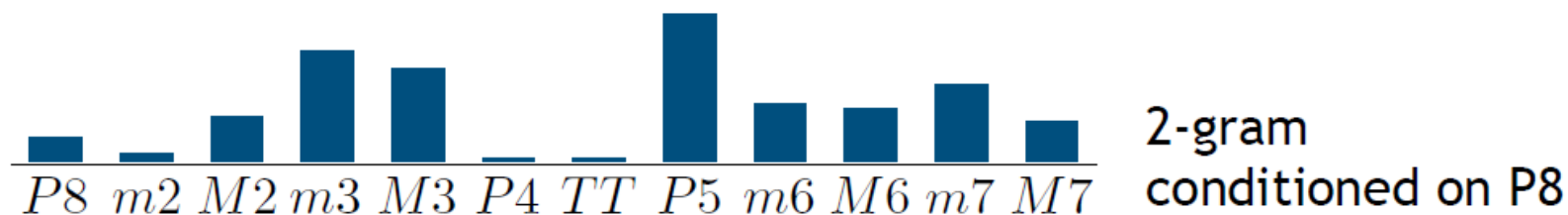
# Hierarchical concept learning

Compositional Rule Examples:

$\phi$  : interval class between soprano and bass



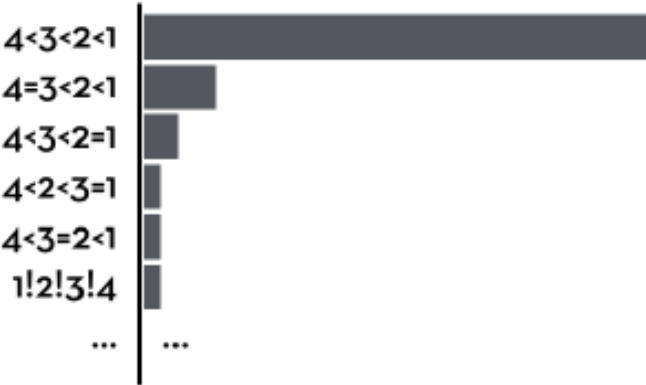
“Individual perfect octaves (P8s) are favored as most consonant.”



“Parallel perfect octaves (P8s) are uncommon.”

This form of compositional rules are in fact human-interpretable

window: (1,2,3,4)  
basis feature: order  
n-gram: 1



Score Range	# of Students
50	3
[40,50)	7
[30,40)	2
[20,30)	4
[10,20)	1
[0,10)	1
0	5

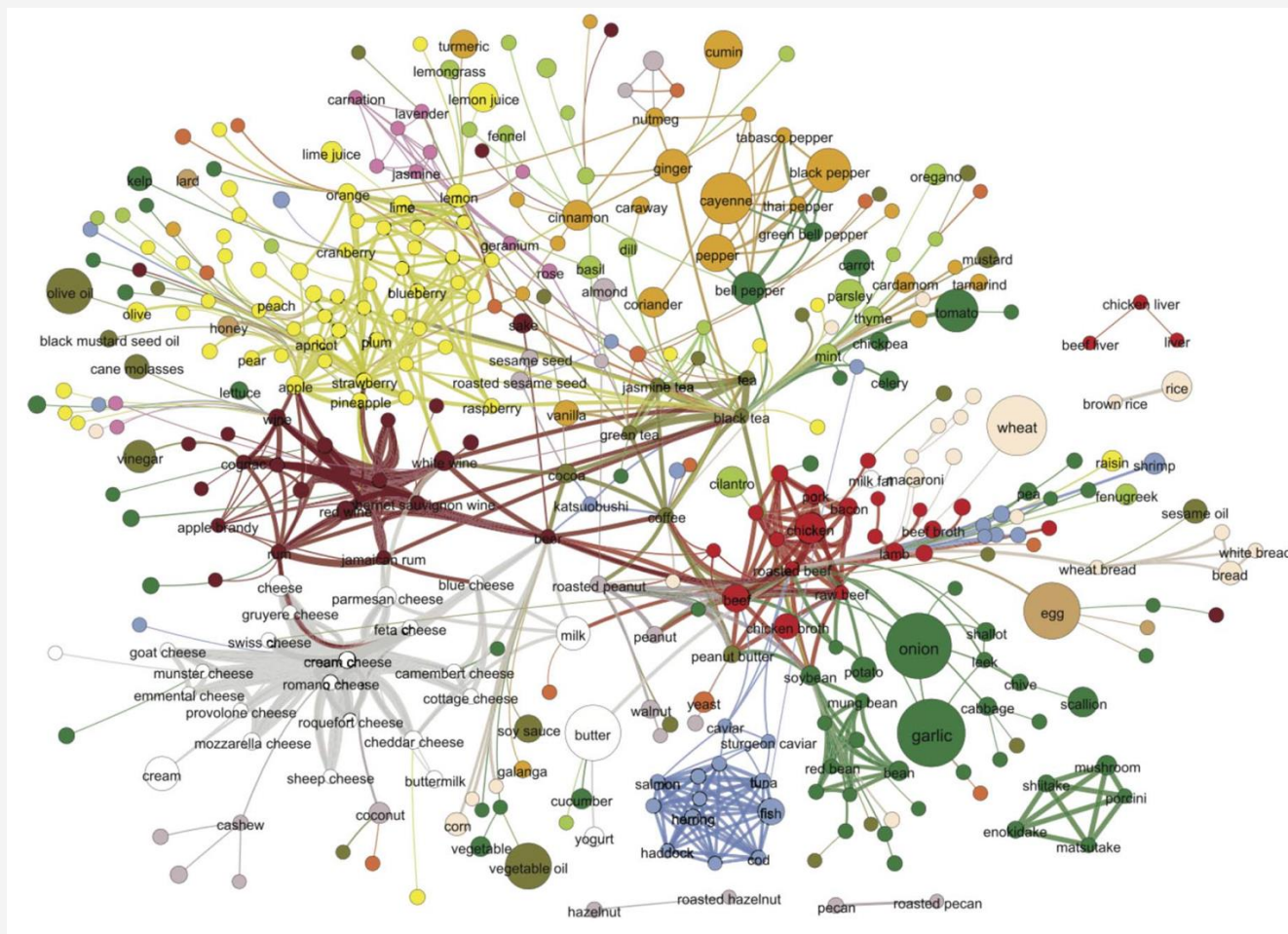
Table 1: Students' final scores.

# Explaining Creative Artifacts

- Due to widespread deployment of AI in settings with people, interest in providing explanations for their results—whether to enable action, provide basis for evaluation, or for intrinsic reasons
  - Past work in XAI focused on decisions and predictions: what about creative ideas or artifacts?
  - **Inverse problem formulation and algorithms to provide *post hoc* human-understandable rationales for creative artifacts** (not models a la BERTology)
    - **Insight from behavioral sciences:** the human process of creativity is largely through associative chains, so also provide explanations via associative chains
    - **Algorithm idea / solution concept:** Traveling salesman path through a knowledge graph of associative elements (nearby associations are easier to understand)
1. Culinary recipe of a new spice mixture that can be used for pastries: thyme, clove, tangerine peel oil, french lavender, and lavender flower.
  2. An English sentence “After hearing the music, I woke up in the morning and opened my eyes, after which I had breakfast at the kitchen table”.
  3. A Hindi sentence “पुराना भारतीय सरकार अन्य अंतर्राष्ट्रीय सरकार के उल्लंघन का कारण है” (the old Indian government was the cause of other international governments unravelling).

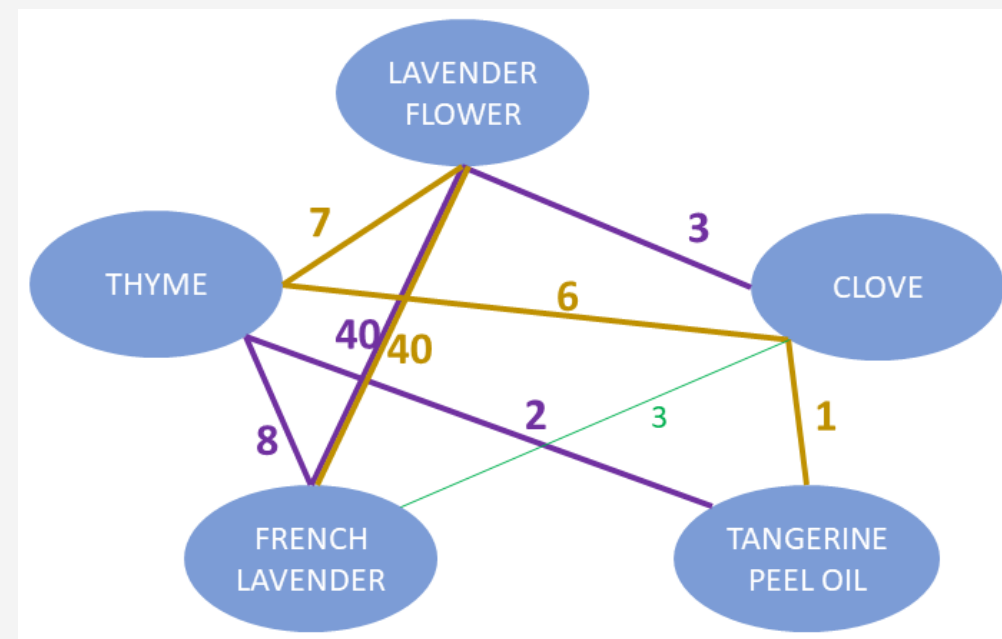


# Examples with explicit concepts and their relationships



Shared flavor compound network as knowledge network yields explanations via flavor pairing hypothesis in culinary science [Ahn, et al., 2011]

lavender flower and clove are used together since they share methyl benzoate, carvone, and linalyl acetate as flavor compounds



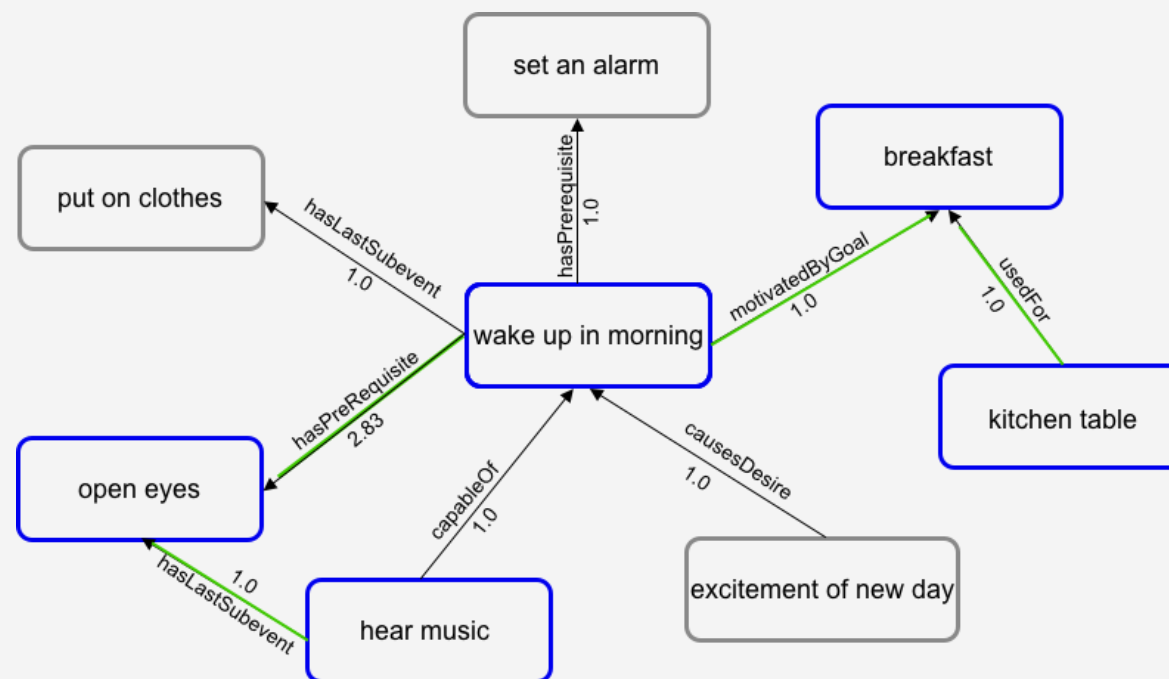
2-best TSP paths in the subgraph, where edge weights are number of shared flavor compounds (treated as strength of association)

# Examples with explicit concepts and their relationships

Use concept extraction and then map to the ConceptNet knowledge graph, which represents natural language concepts with relations represented using commonsense reasoning

Notice that the TSP tour is not in the same order as the sentence syntax, but is governed by the strongest semantic relationships

The edge labels provide human-interpretable explanations, such as “waking up in the morning is motivated by the goal of breakfast” even though this isn’t really in the sentence



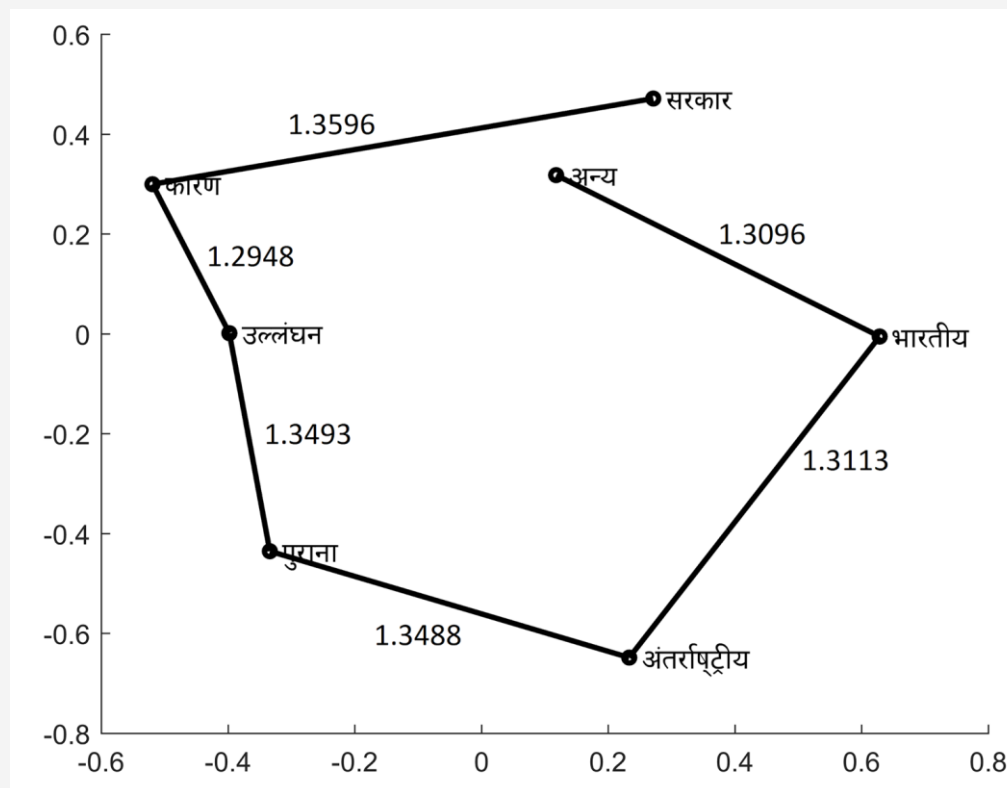
2. An English sentence “After hearing the music, I woke up in the morning and opened my eyes, after which I had breakfast at the kitchen table”.

## Example with implicit concept relationships

There is no common “ConceptNet” for Hindi or other low-resource languages

Instead use implicit relationships from distances in a word embedding space

3. A Hindi sentence “पुराना भारतीय सरकार अन्य अंतर्राष्ट्रीय सरकार के उल्लंघन का कारण है” (the old Indian government was the cause of other international governments unravelling).



# Define TSP path length as a measure of novelty

Creative artifacts that connect semantically distant concepts are very novel

Use computational geometry to characterize stochastic sampling-based creativity algorithms

**Theorem 1** (Beardwood et al. (1959)). *Let  $\{X_1, \dots, X_n\}$  be a set of i.i.d. random variables in  $\mathbb{R}^d$  with bounded support. Then the length  $L_n$  of the shortest TSP tour through the points  $\{X_i\}$  satisfies*

$$\frac{L_n}{n^{(d-1)/d}} \rightarrow \beta_d \int_{\mathbb{R}^d} f(x)^{(d-1)/d} dx$$

*with probability 1 as  $n \rightarrow \infty$ , where  $f(x)$  is the absolutely continuous part of the distribution of the  $\{X_i\}$  and  $\beta_d$  is a constant that depends on  $d$  but not on  $f(x)$ .*

Taking  $f(x)$  as the creativity algorithm, one can directly characterize level of novelty it produces

**Theorem 2** (Hero et al. (2002)). *Let  $\{X_1, \dots, X_n\}$  be a set of i.i.d. random variables in  $\mathbb{R}^d$  with bounded support. Let  $L_n$  be the length of the shortest TSP tour through the points  $\{X_i\}$ . Then the following estimator for Renyi entropy*

$$\hat{H}_\gamma(X) = \frac{1}{1-\gamma} (\ln L_n / n^\gamma - \ln \beta),$$

*where  $\gamma = (d-1)/d$  and  $\beta$  is a fixed constant independent of  $f$ , is an asymptotically unbiased and almost surely consistent estimator of the Renyi entropy of  $f$ .*

TSP path length is consistent estimator for Shannon entropy (in large dimensions), so semantic novelty intertwined with statistical novelty

