INSTITUT TUTTE INSTITUTE

Topology of Language

Topological Techniques for Natural Language Processing

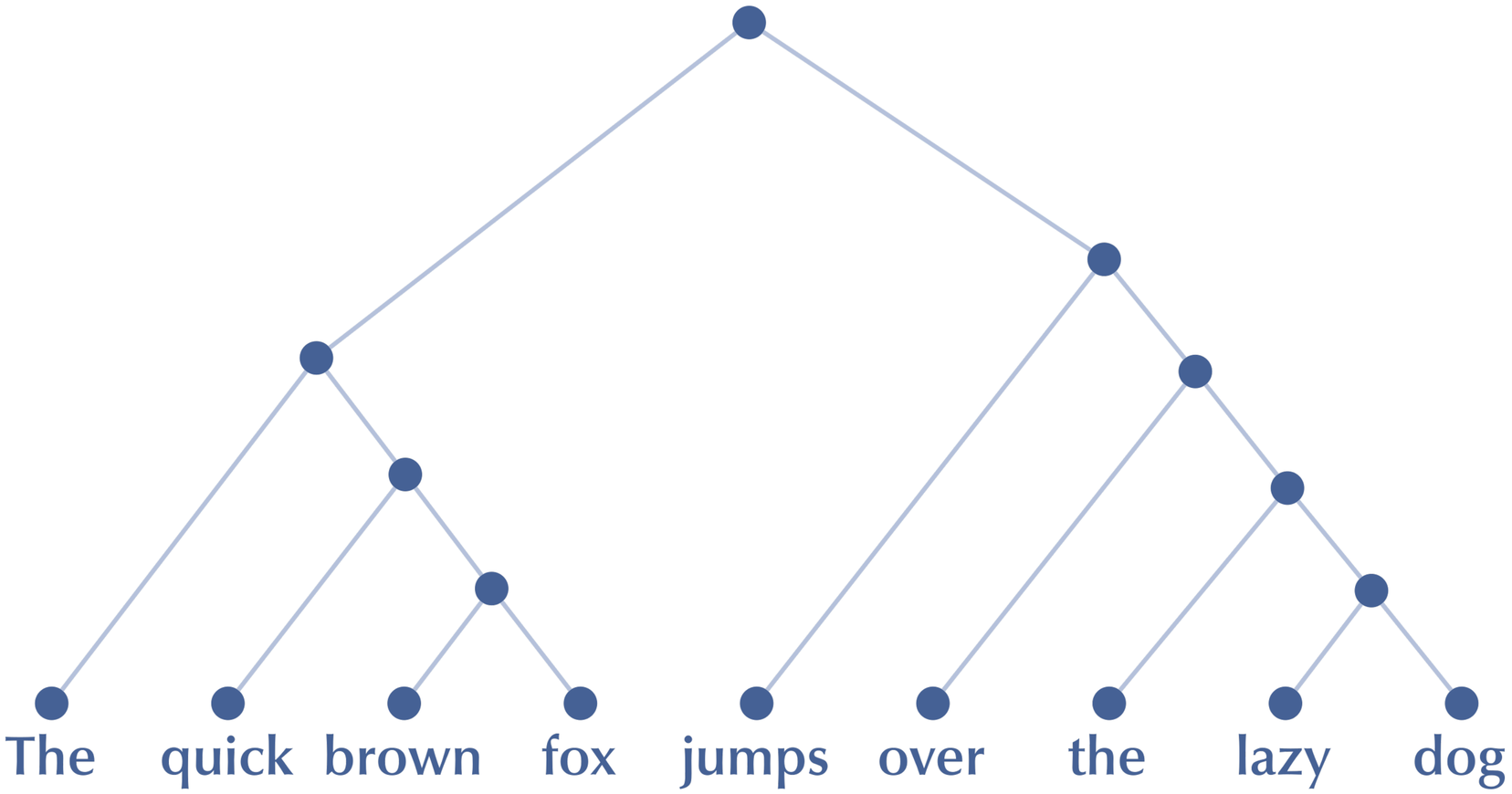
INSTITUT TUTTE INSTITUTE

Vision

INSTITUT TUTTE INSTITUTE

Language has structure

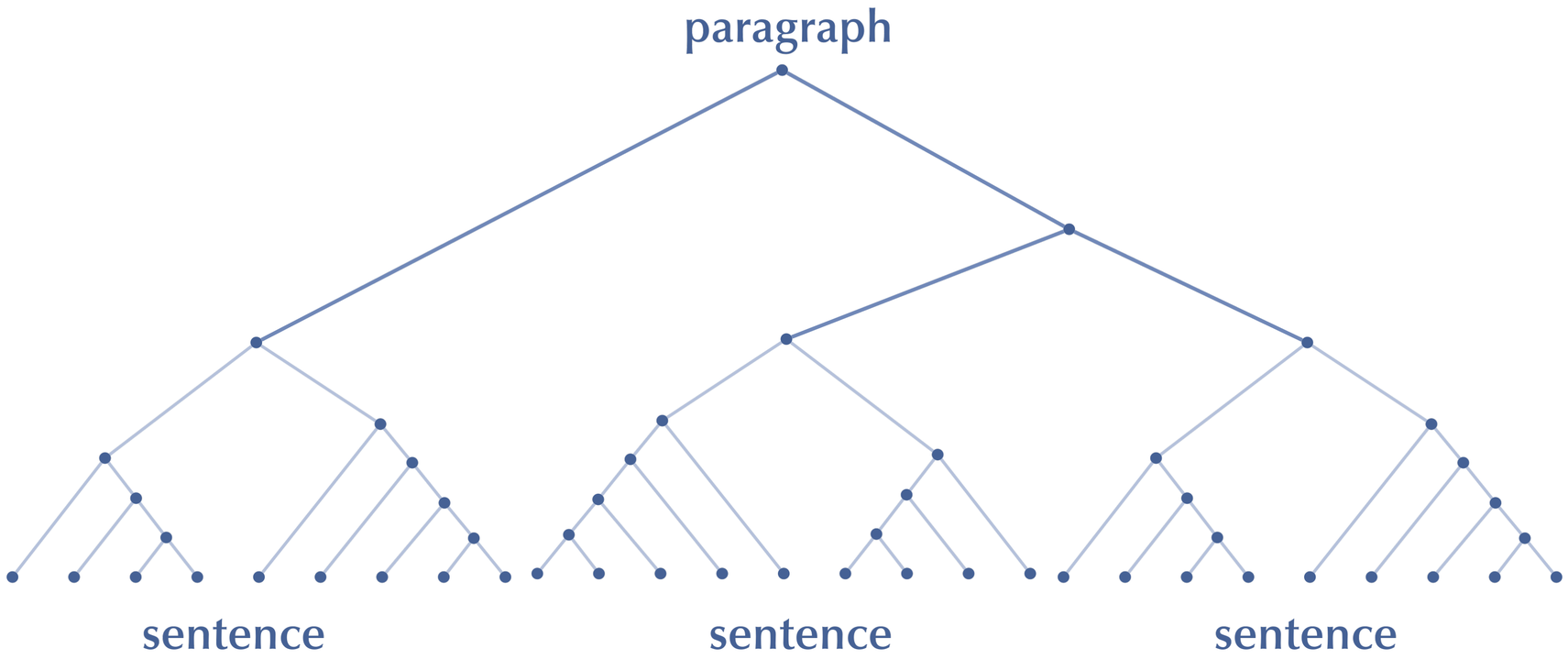
INSTITUT TUTTE INSTITUTE



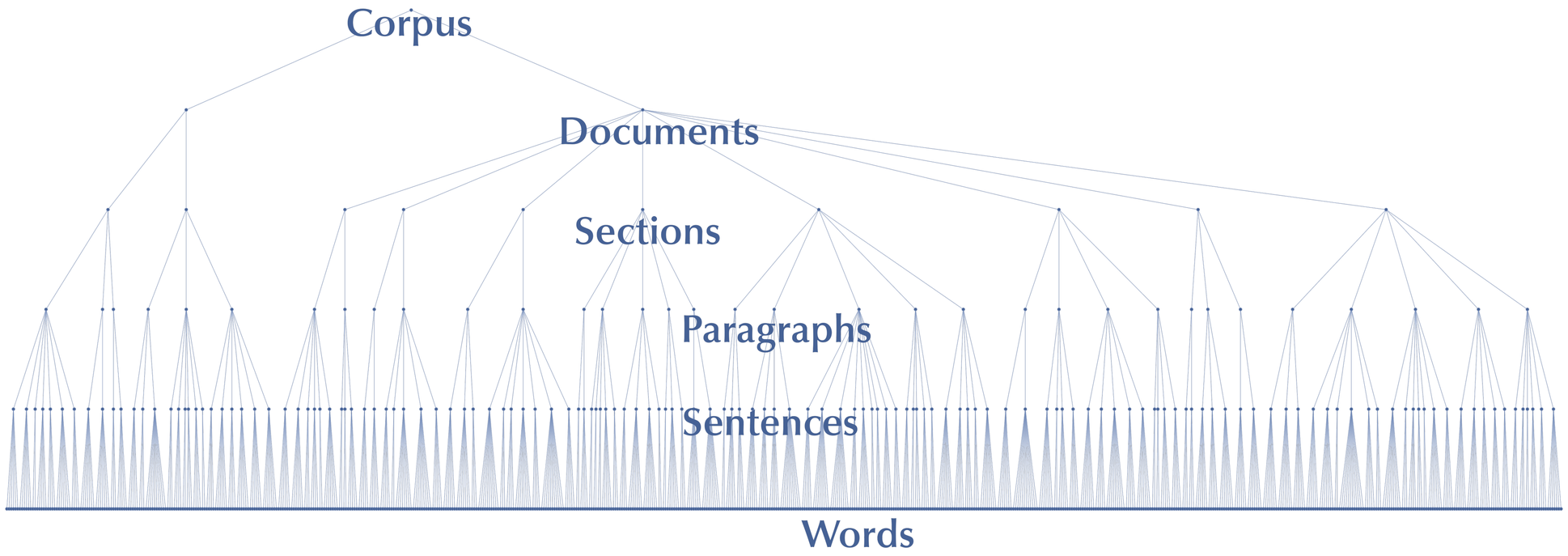
INSTITUT TUTTE INSTITUTE

Many layers of structure

INSTITUT TUTTE INSTITUTE



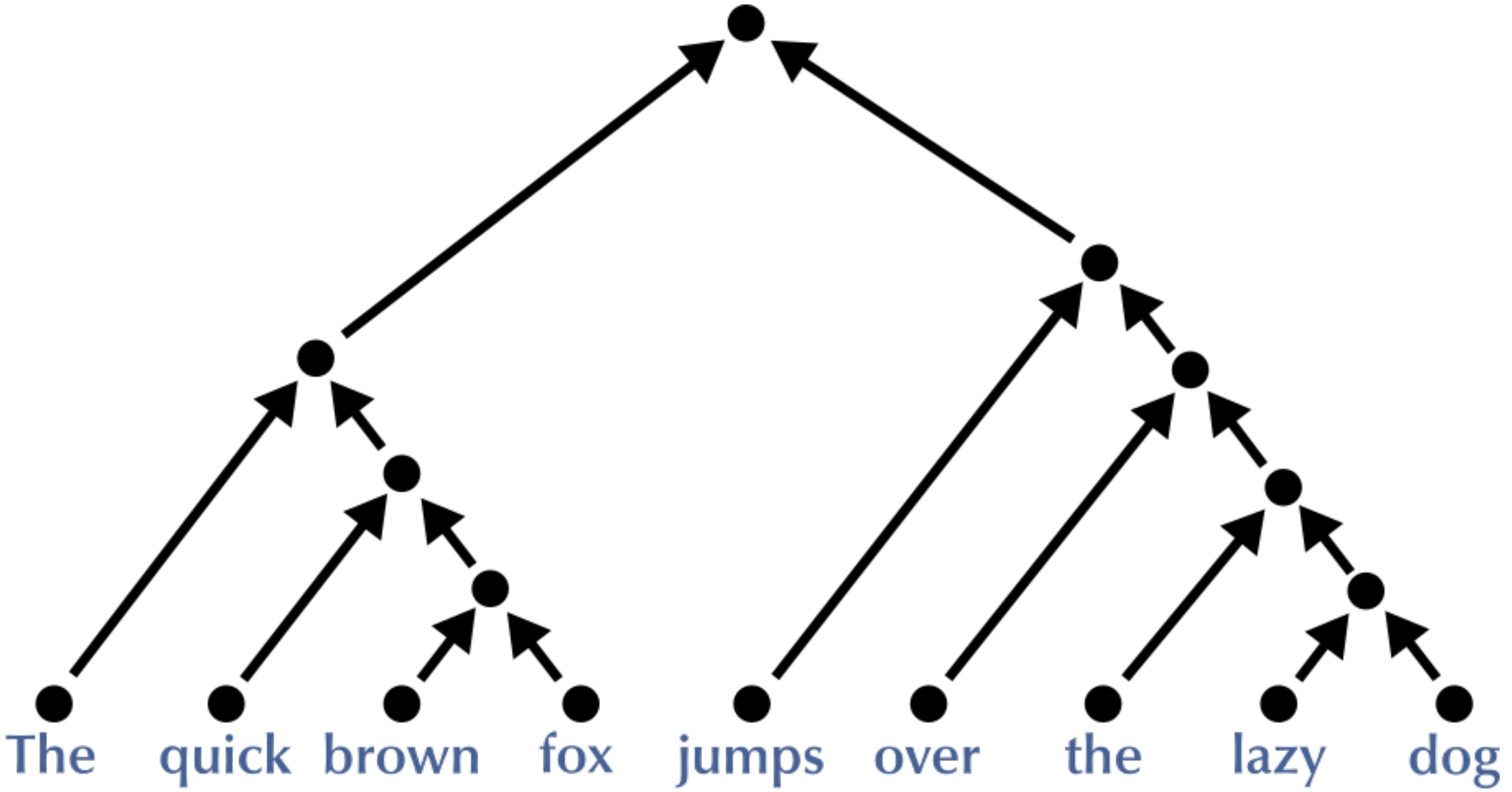
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

We could build a category from this

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

Yoneda embedding of a document?

All the constituent parts of the document

INSTITUT TUTTE INSTITUTE

Simplify: all the word usages in a document

This is the classical "bag-of-words" model

INSTITUT TUTTE INSTITUTE

What about words?
In the opposite category…

INSTITUT TUTTE INSTITUTE

A word is the set of all its usages in a corpus

INSTITUT TUTTE INSTITUTE

Simplify: all the sentences that use the word

INSTITUT TUTTE INSTITUTE

You shall know a word by the company it keeps

— John Rupert Firth

INSTITUT TUTTE INSTITUTE

Words as Distributions

INSTITUT TUTTE INSTITUTE

Overview

INSTITUT TUTTE INSTITUTE

Represent a word as a document of all the sentences that use the word

INSTITUT TUTTE INSTITUTE

Represent such a document using a bag-of-words model

INSTITUT TUTTE INSTITUTE

Represent a word as the multinomial distribution of words that occur nearby

INSTITUT TUTTE INSTITUTE

Information geometry tells us that multinomial distributions live on a manifold

INSTITUT TUTTE INSTITUTE

$$d(p,q) = \arccos\left(\sum_i \sqrt{p_i}\sqrt{q_i}\right)$$

INSTITUT TUTTE INSTITUTE

We can explore the geometry of words under this distance metric

INSTITUT TUTTE INSTITUTE

Messy Details

INSTITUT TUTTE INSTITUTE

Some problems:

- Word order matters
- Words carry different amounts of information
- Language use is noisy

INSTITUT TUTTE INSTITUTE

Words order matters:

Words used *before* a given word
Words used *after* a given word

INSTITUT TUTTE INSTITUTE

Information in words:

Approximate the information of a word in a given context and weight words accordingly

INSTITUT TUTTE INSTITUTE

Language is noisy:

Approximate a background noise model of word use and decompose the multinomial into a mixture of signal and noise distributions

INSTITUT TUTTE INSTITUTE

Where do we get models of entropy and noise?

INSTITUT TUTTE INSTITUTE

$$X \approx UV$$

Where

X is an $N \times D$ matrix

U is an $N \times d$ matrix

V is an $d \times D$ matrix

INSTITUT TUTTE INSTITUTE

Minimize

$$\sum_{i=1}^{N} \sum_{j=1}^{D} \textbf{Loss}\left(X_{ij},\ (UV)_{ij}\right)$$

Subject to constraints…

INSTITUT TUTTE INSTITUTE

Suppose

$$X \sim \mathbf{Pr}(\,\cdot \mid \Theta)$$

where

$$\Theta = UV$$

INSTITUT TUTTE INSTITUTE

Let the loss be the
negative log likelihood
of observing X given Θ

INSTITUT TUTTE INSTITUTE

Multinomial Matrix Factorization

Minimize

$$\sum_{i=1}^{N}\sum_{j=1}^{D} -(UV)_{ij} \cdot \log\left(X_{ij}\right)$$

Subject to

$$(UV)\mathbf{1} = \mathbf{1} \text{ and } (UV)_{ij} \geq 0$$

INSTITUT TUTTE INSTITUTE

Probabilistic Latent Semantic Analysis

Minimize

$$\sum_{i=1}^{N}\sum_{j=1}^{D} -(UV)_{ij} \cdot \log\left(X_{ij}\right)$$

Subject to

$$U\mathbf{1} = \mathbf{1}, V\mathbf{1} = \mathbf{1} \textbf{ and } U_{ij} \geq 0, V_{ij} \geq 0$$

INSTITUT TUTTE INSTITUTE

Θ provides a low rank multinomial distribution model

INSTITUT TUTTE INSTITUTE

Θ_{ij} is the "background" probability of word j occurring in the context of word i

$-\log(\Theta_{ij})$ is an approximation of the information carried by word j in the context of word i

INSTITUT TUTTE INSTITUTE

We can weight each entry by the (approximate) information it carries

INSTITUT TUTTE INSTITUTE

A rank 1 model gives classical TF-IDF from NLP

INSTITUT TUTTE INSTITUTE

Θ_i provides a model of the "background" multinomial of word occurrence in the context of word i

INSTITUT TUTTE INSTITUTE

We can use an EM algorithm to decompose a given multinomial into noise and signal

INSTITUT TUTTE INSTITUTE

We are back to where our overview left off…

INSTITUT TUTTE INSTITUTE

Embedding Words

INSTITUT TUTTE INSTITUTE

We have high quality and high dimensional representations of words with a natural ambient geometry …

INSTITUT TUTTE INSTITUTE

We need a low dimensional representation that preserves topology

INSTITUT TUTTE INSTITUTE

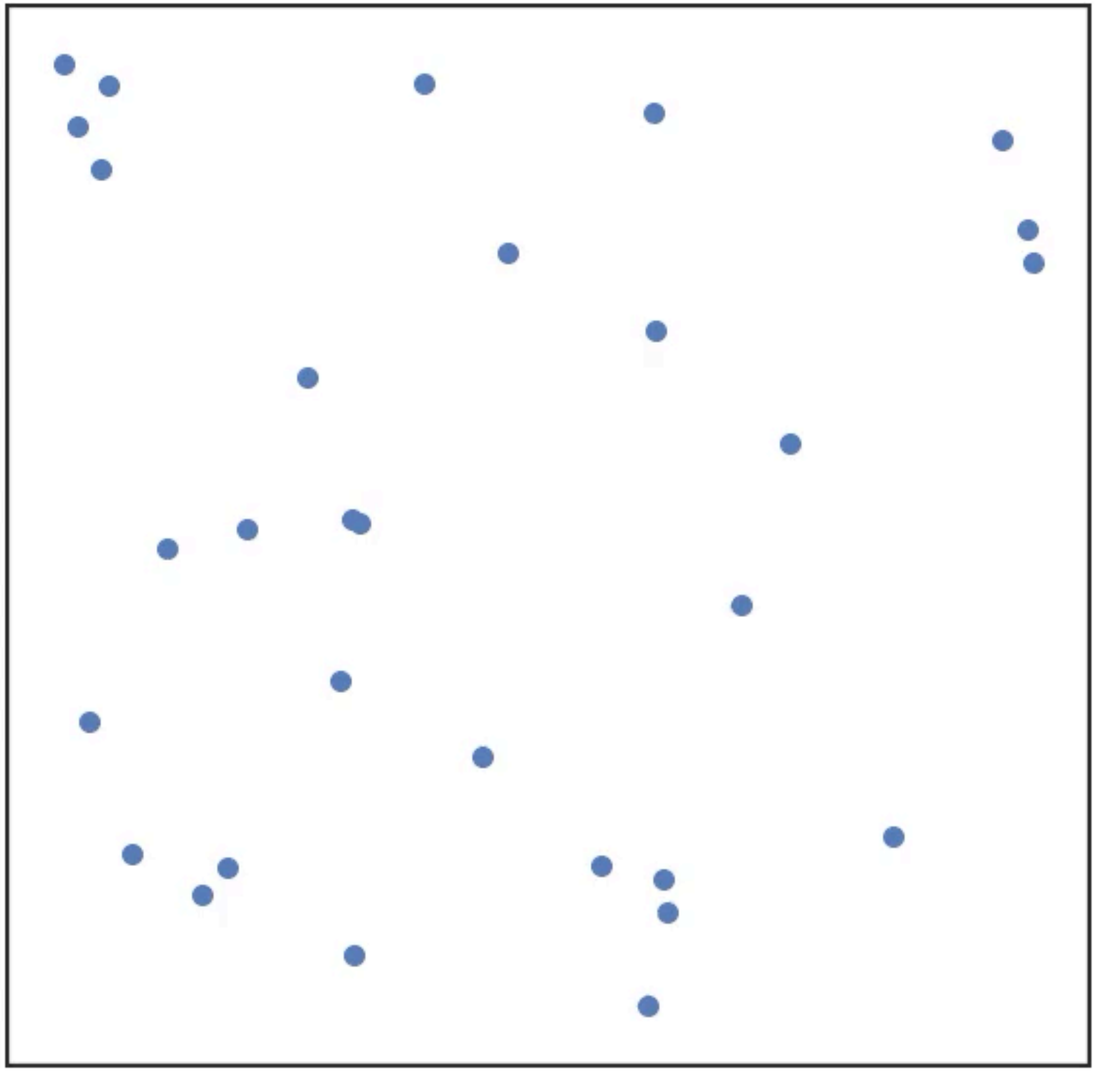
UMAP is a topology based dimension reduction algorithm

INSTITUT TUTTE INSTITUTE

Assumptions:

Uniform distribution

Locally connected



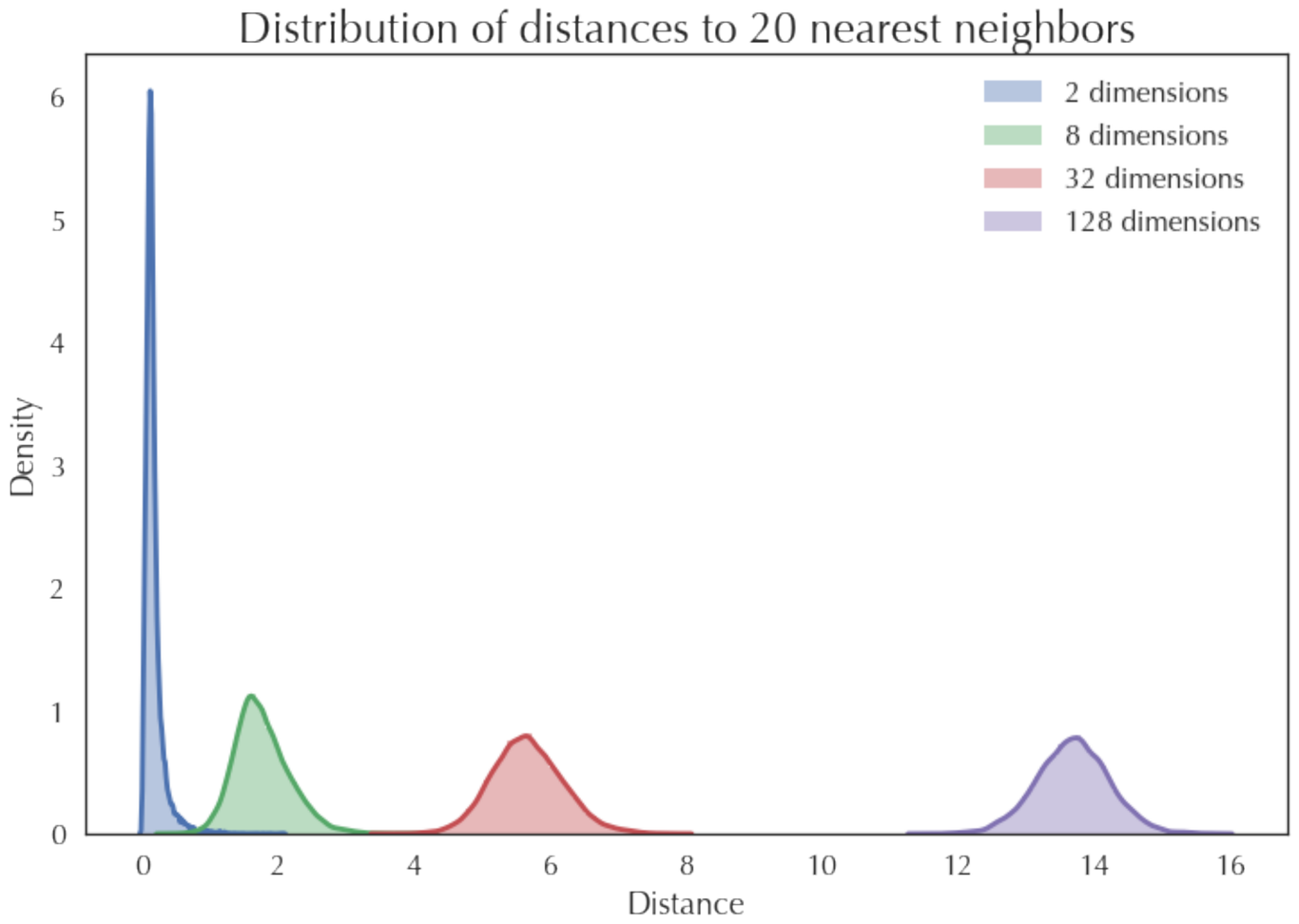
INSTITUT TUTTE INSTITUTE

Assumptions:

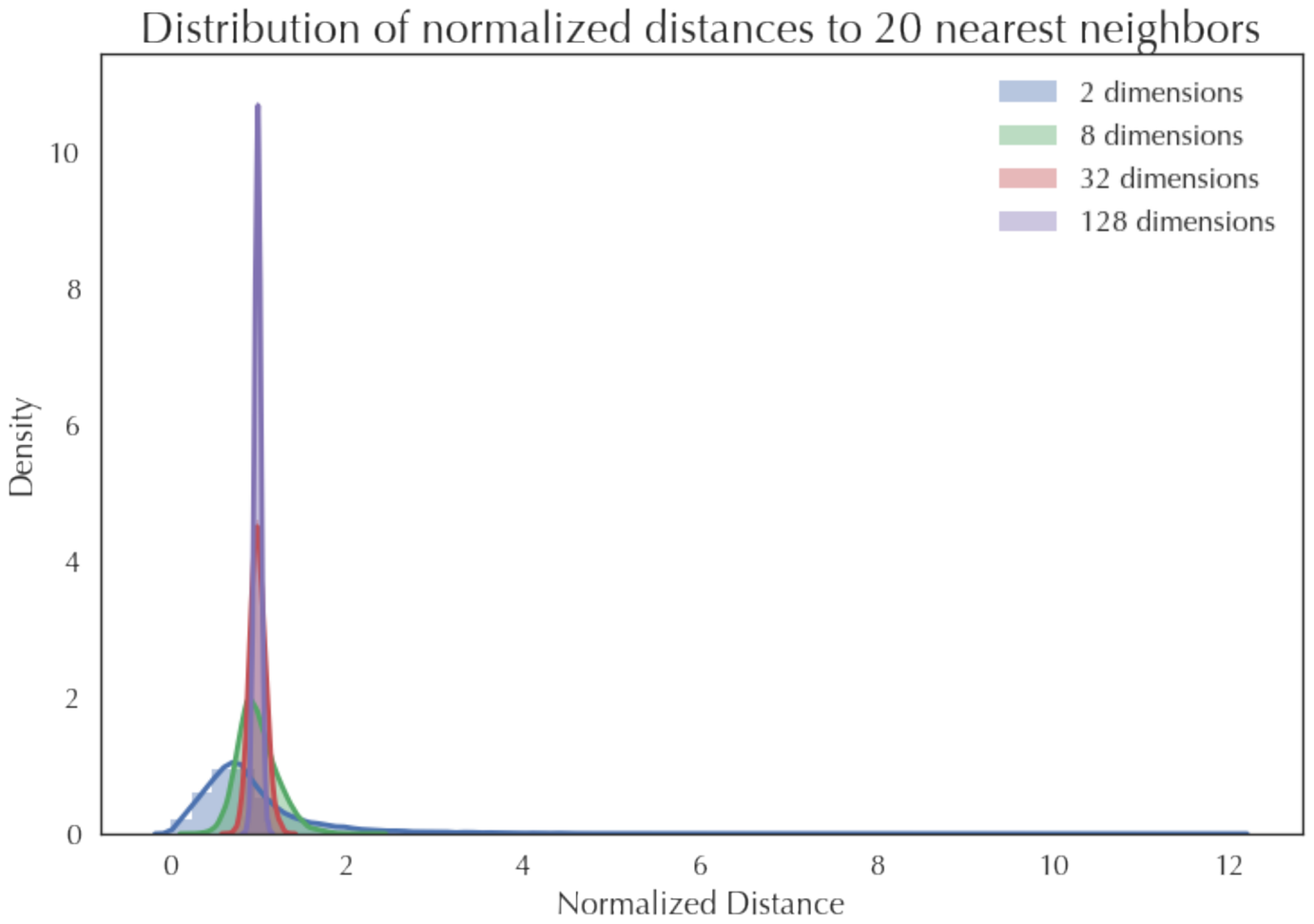
Uniform distribution

Locally connected

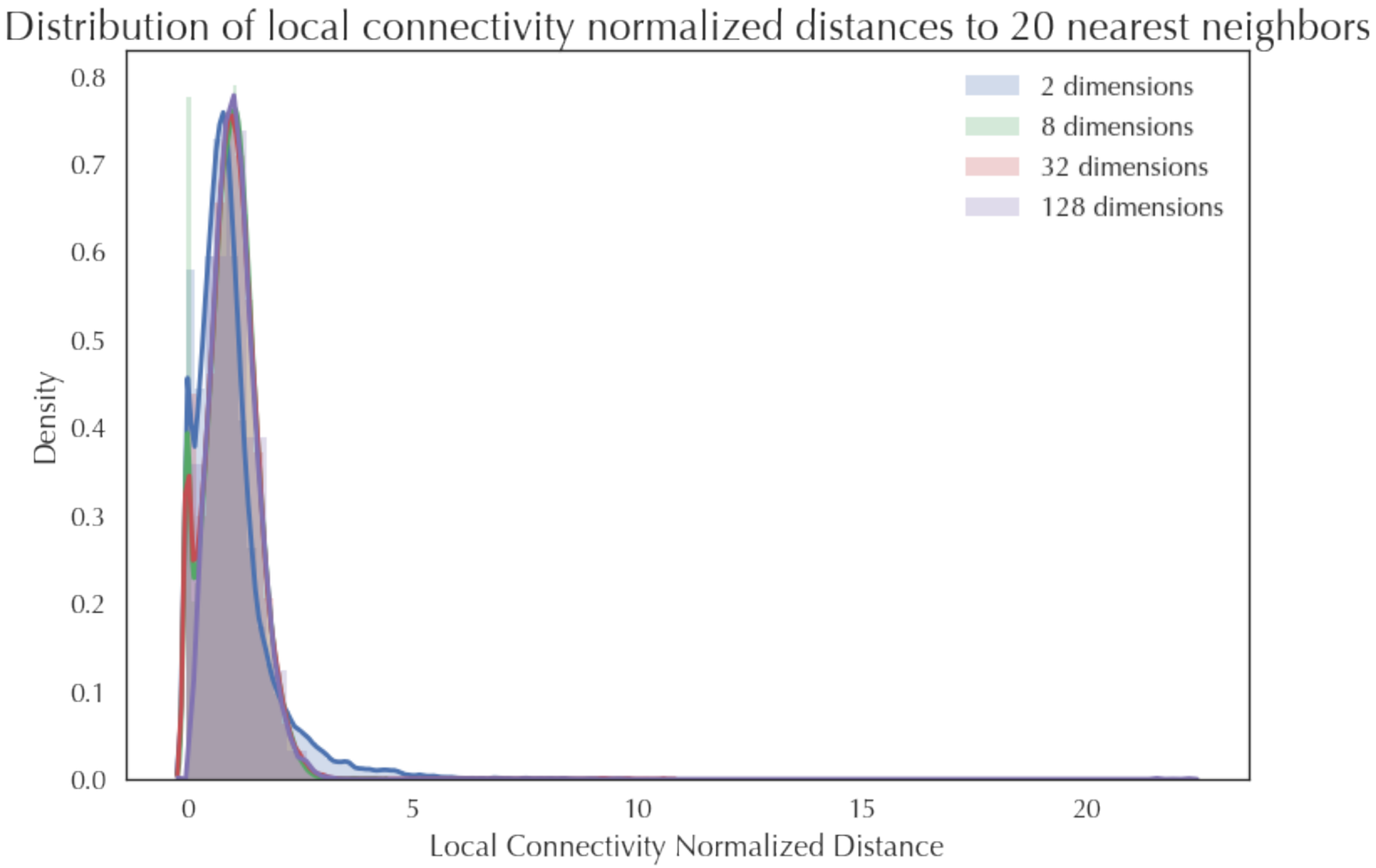
INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

We now have a filtered simplicial complex

INSTITUT TUTTE INSTITUTE

Optimize a low dimensional representation via cross-entropy

INSTITUT TUTTE INSTITUTE

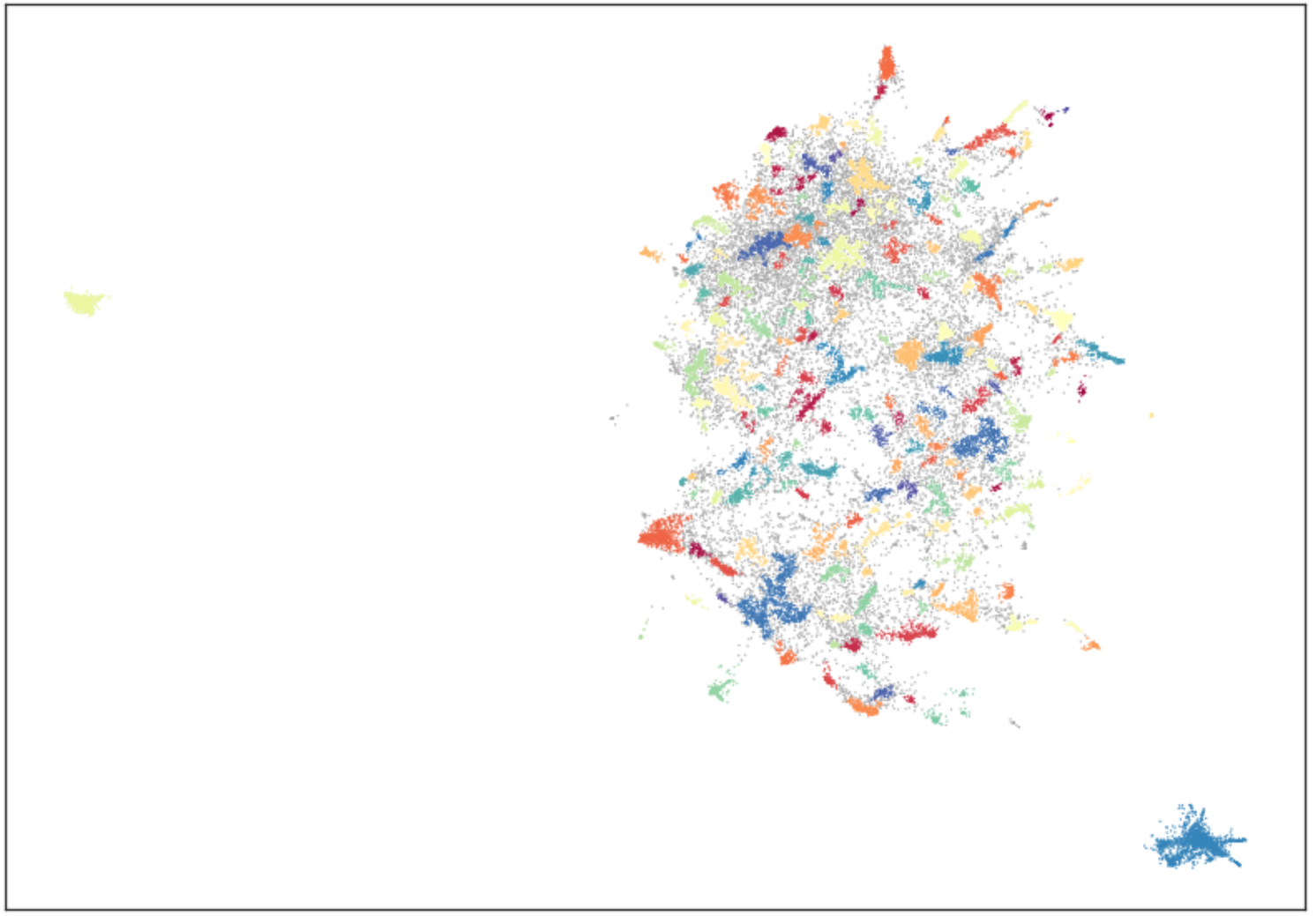
$$\sum_{a} \mu(a) \log\left(\frac{\mu(a)}{\nu(a)}\right) + (1-\mu(a)) \log\left(\frac{1-\mu(a)}{1-\nu(a)}\right)$$

INSTITUT TUTTE INSTITUTE

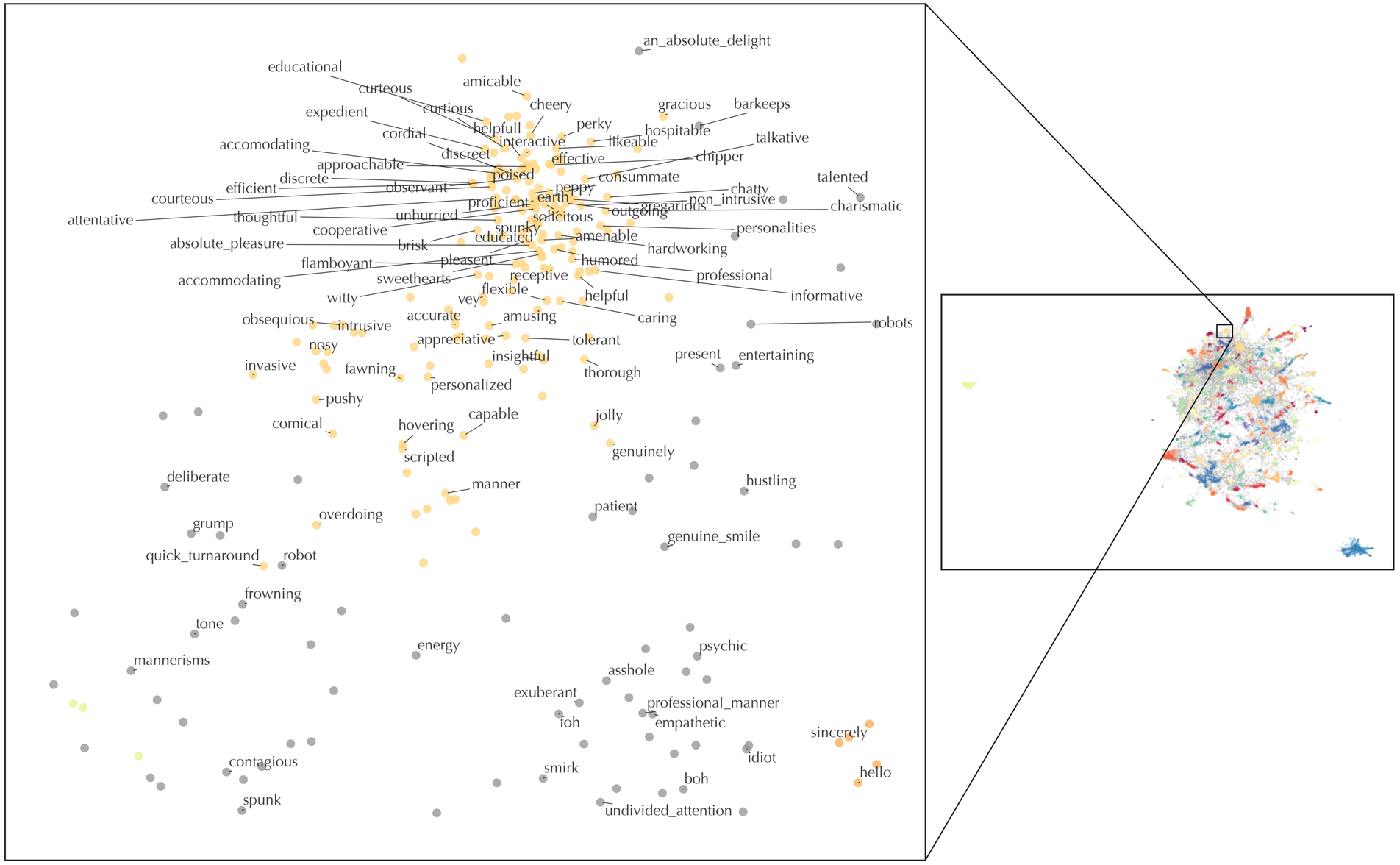
$$d(p,q) = \arccos\left(\sum_i \sqrt{p_i}\sqrt{q_i}\right)$$

INSTITUT TUTTE INSTITUTE

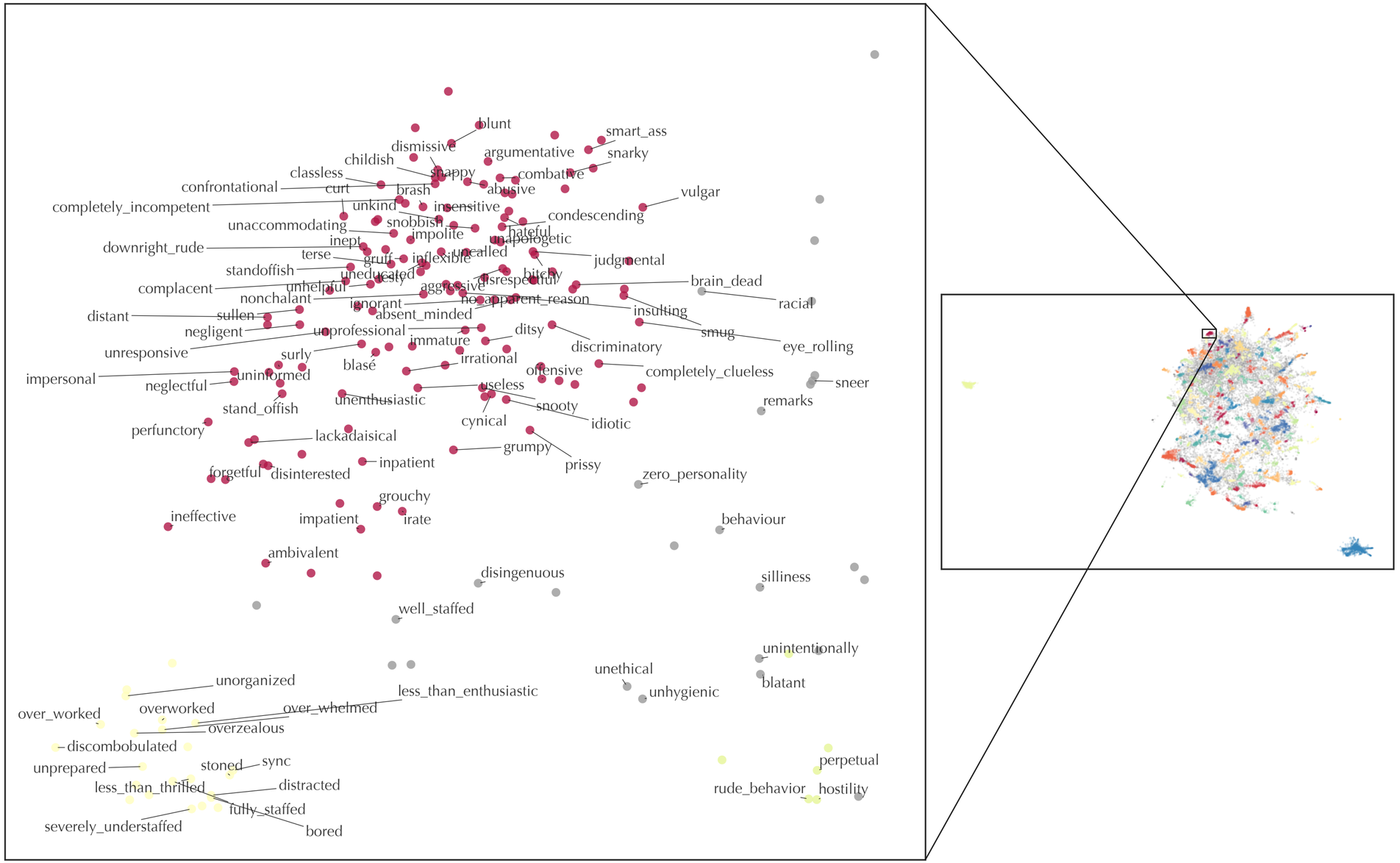
Embedding Yelp Reviews



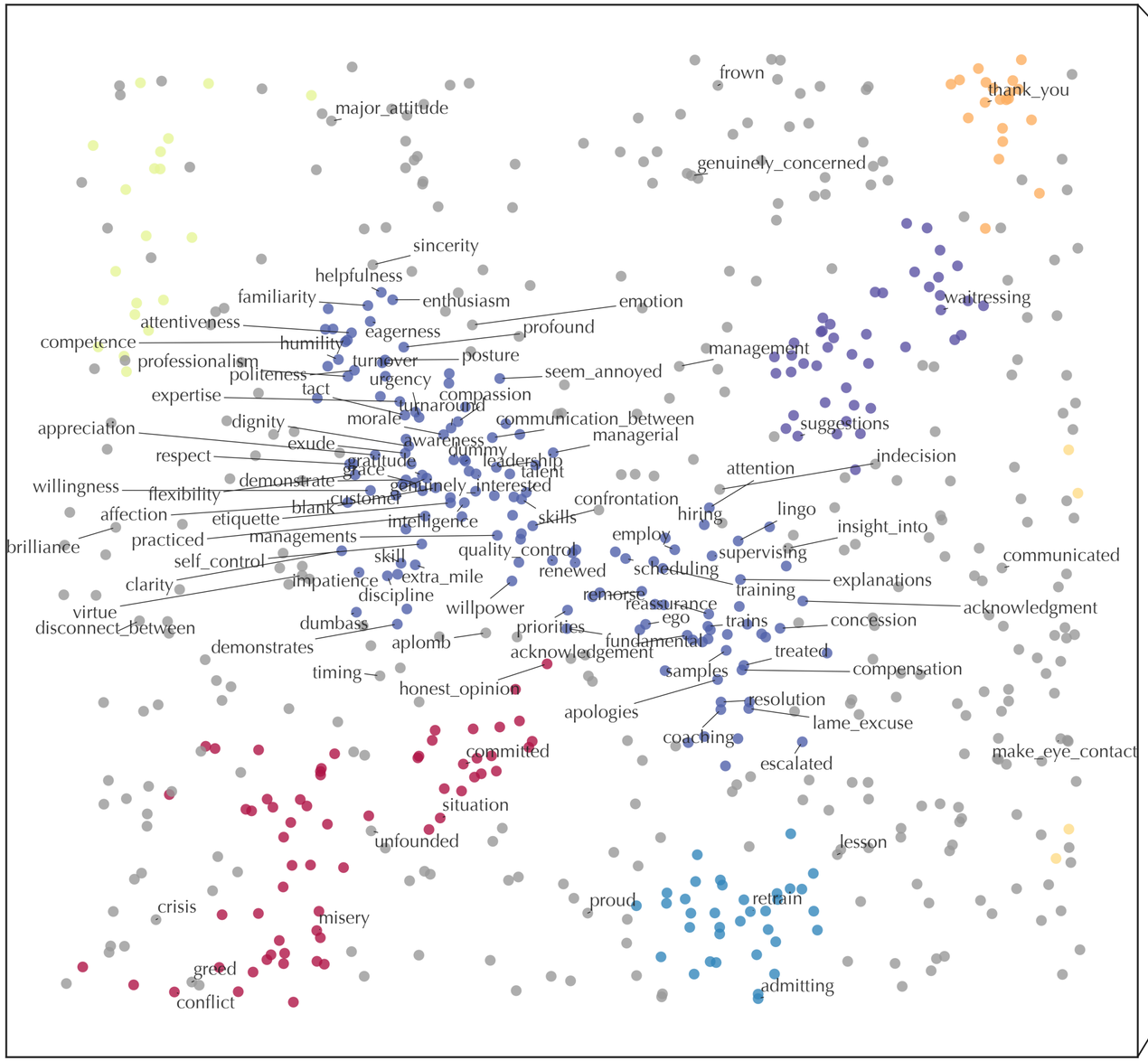
INSTITUT TUTTE INSTITUTE



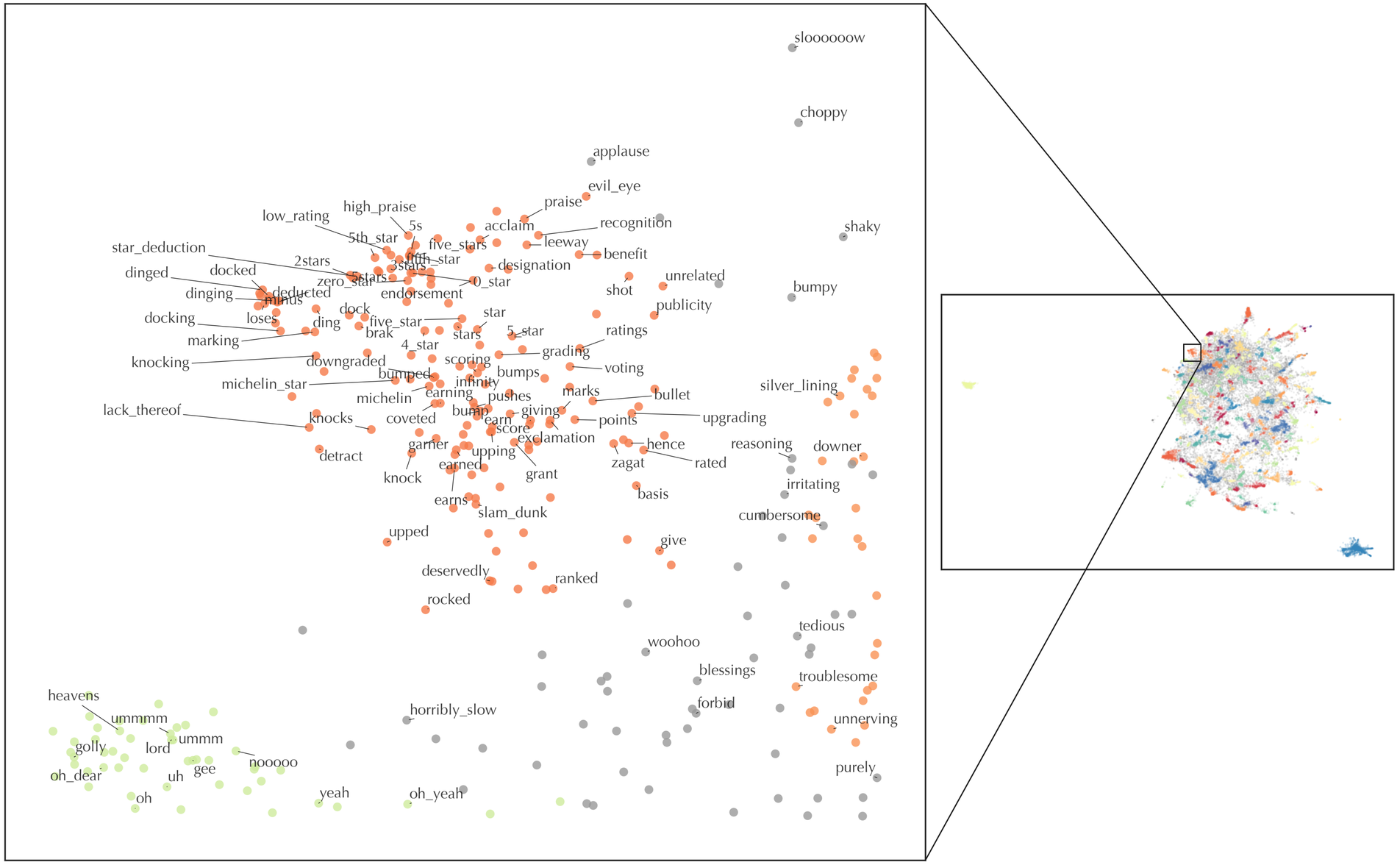
INSTITUT TUTTE INSTITUTE



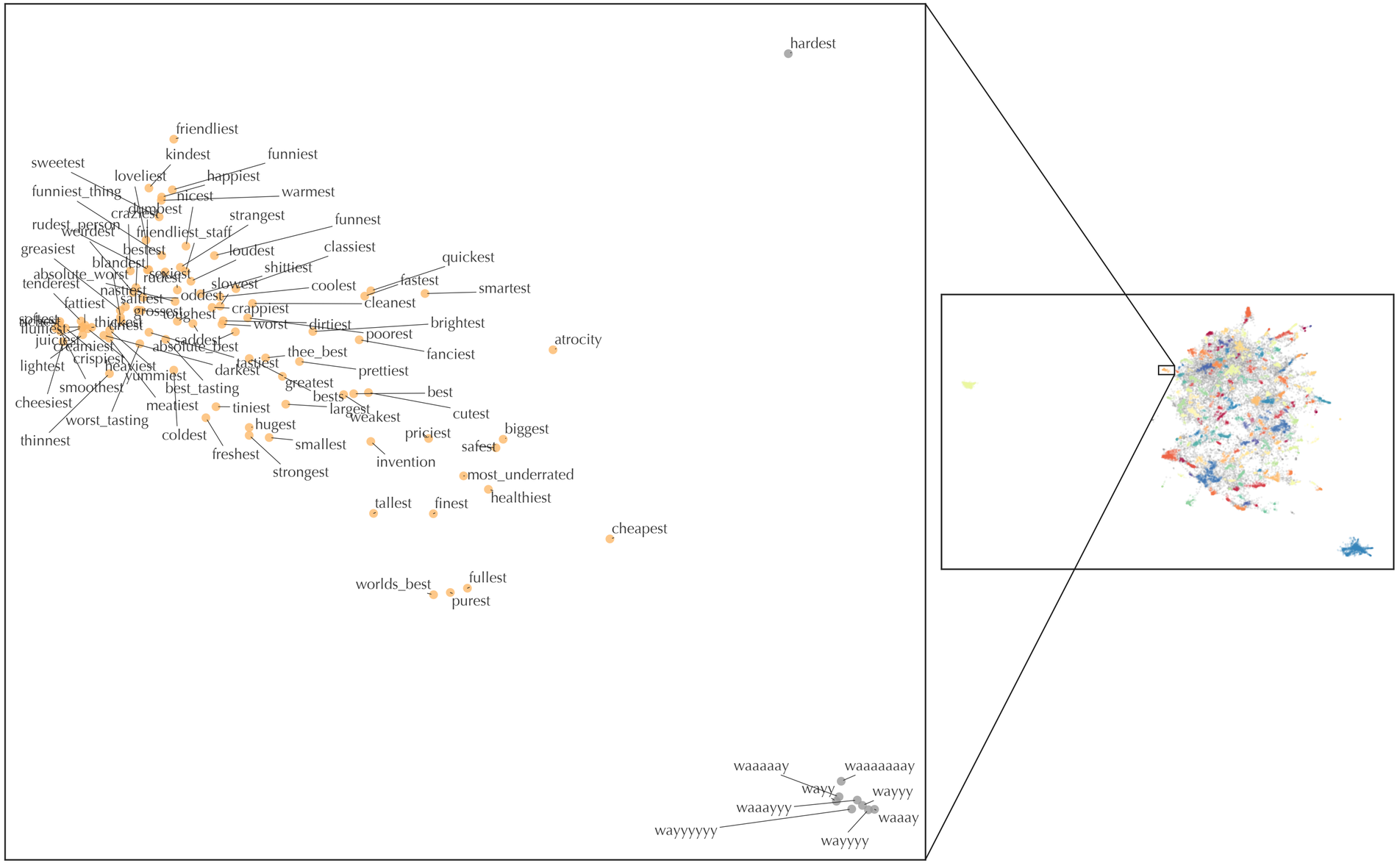
INSTITUT TUTTE INSTITUTE

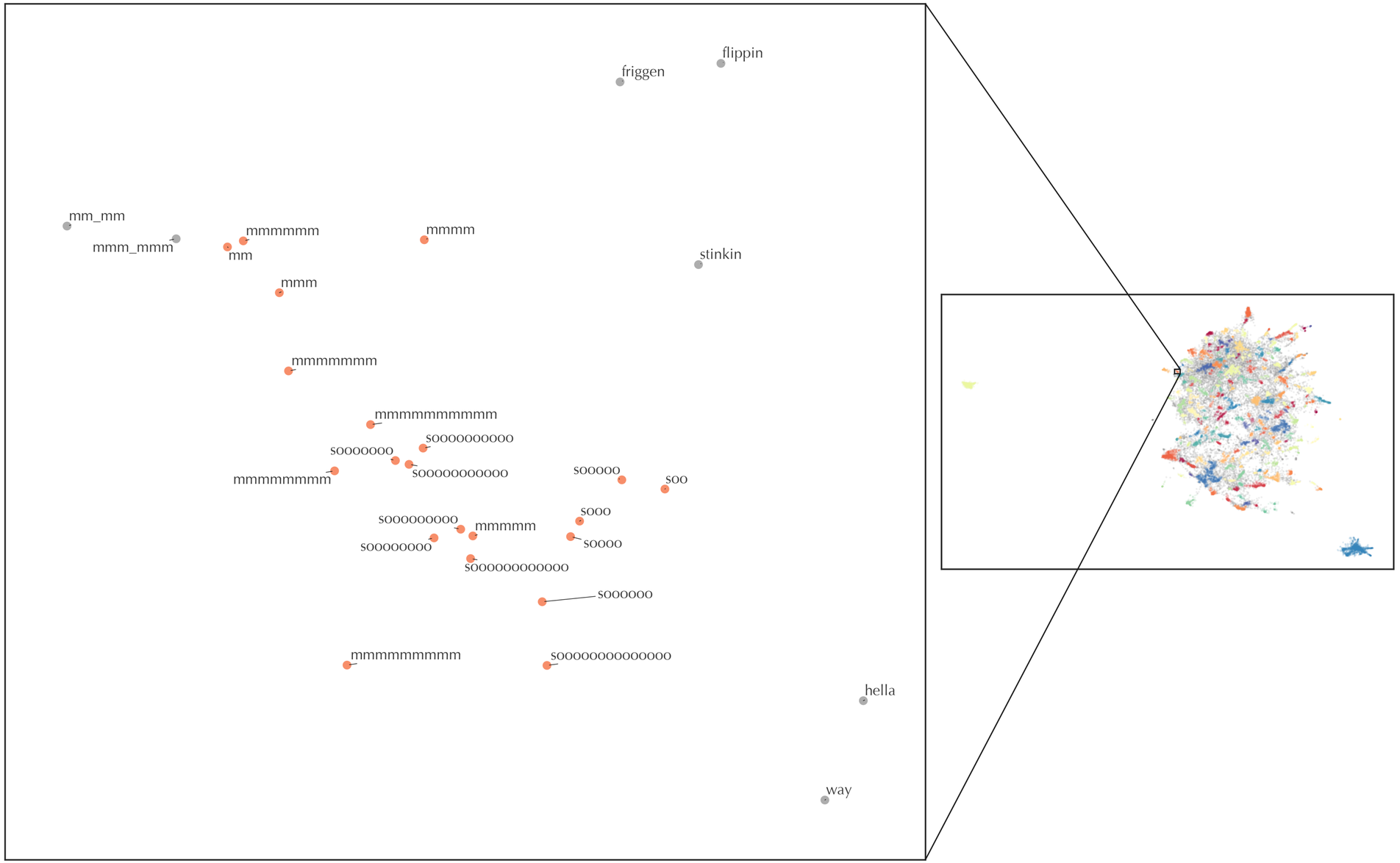


INSTITUT TUTTE INSTITUTE

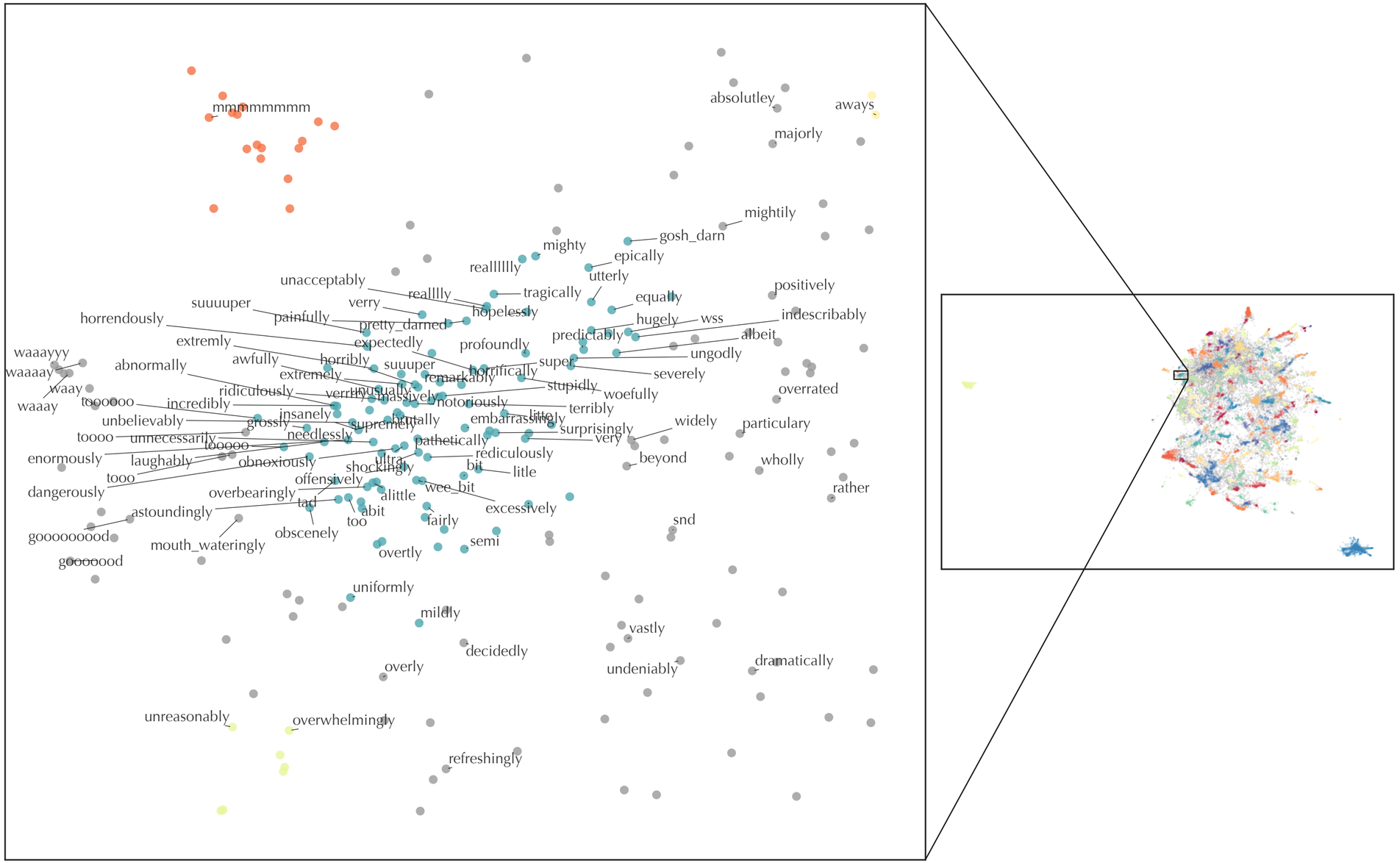


INSTITUT TUTTE INSTITUTE

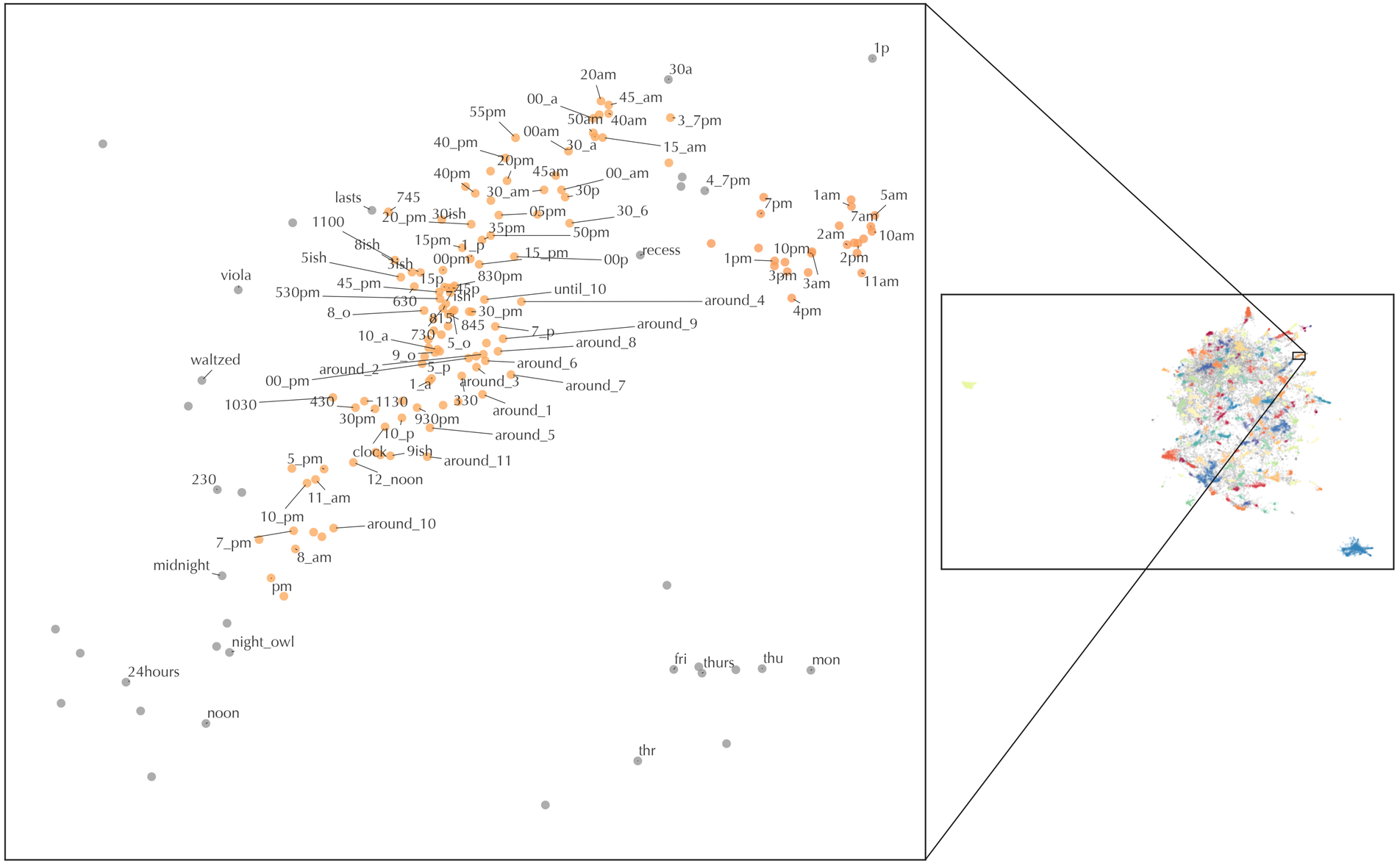



flippin
friggen
mm_mm
mmmmmm
mmmm
mmm_mmm
mm
stinkin
mmm
mmmmmmmm
mmmmmmmmmmm
soooooooooo
sooooooo
mmmmmmmmm
sooooooooooo
soooooo
soo
soooooooooo
mmmmm
sooo
sooooooooo
soooo
soooooooooooo
soooooo
mmmmmmmmmm
sooooooooooooo
hella
way

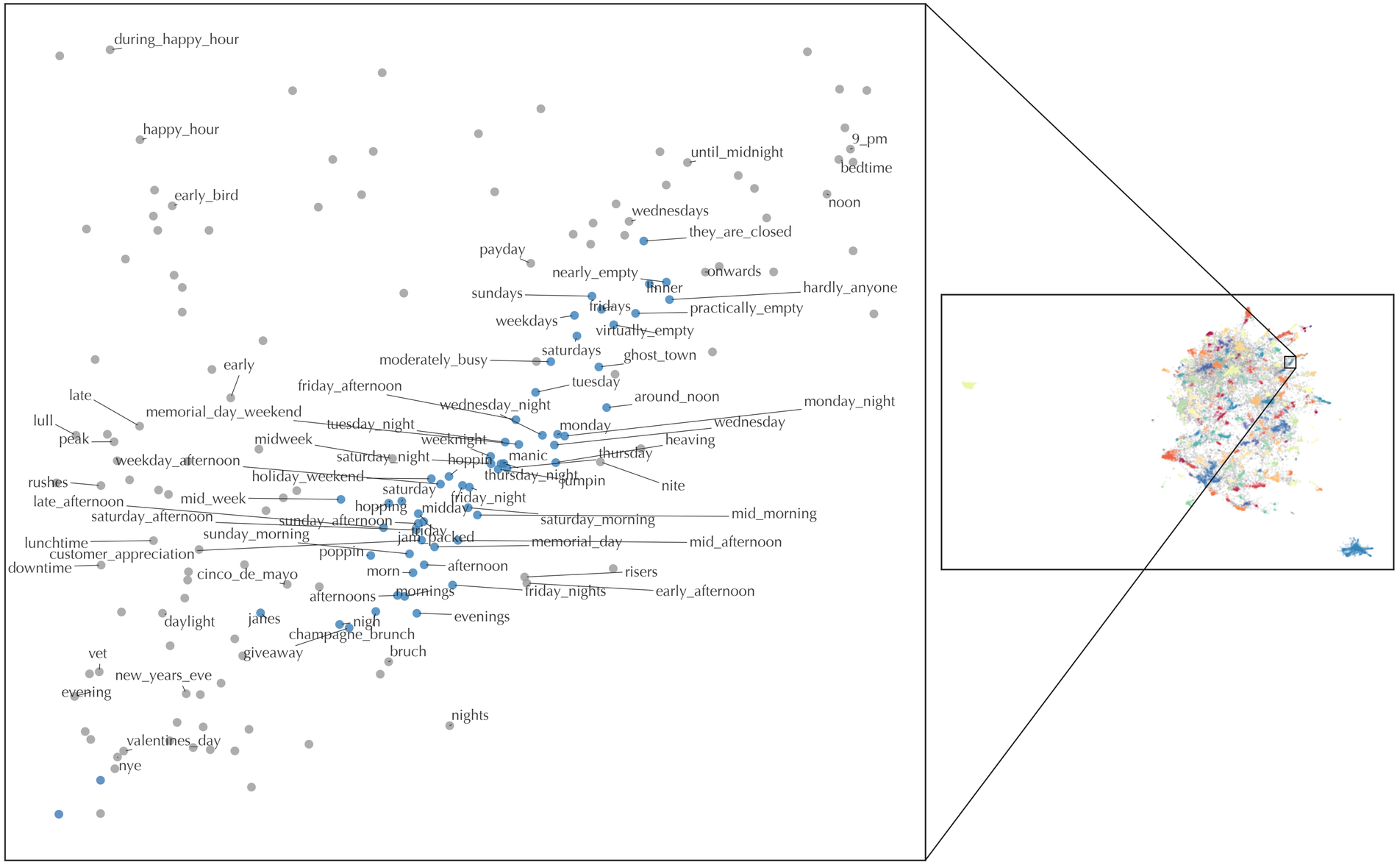
INSTITUT TUTTE INSTITUTE



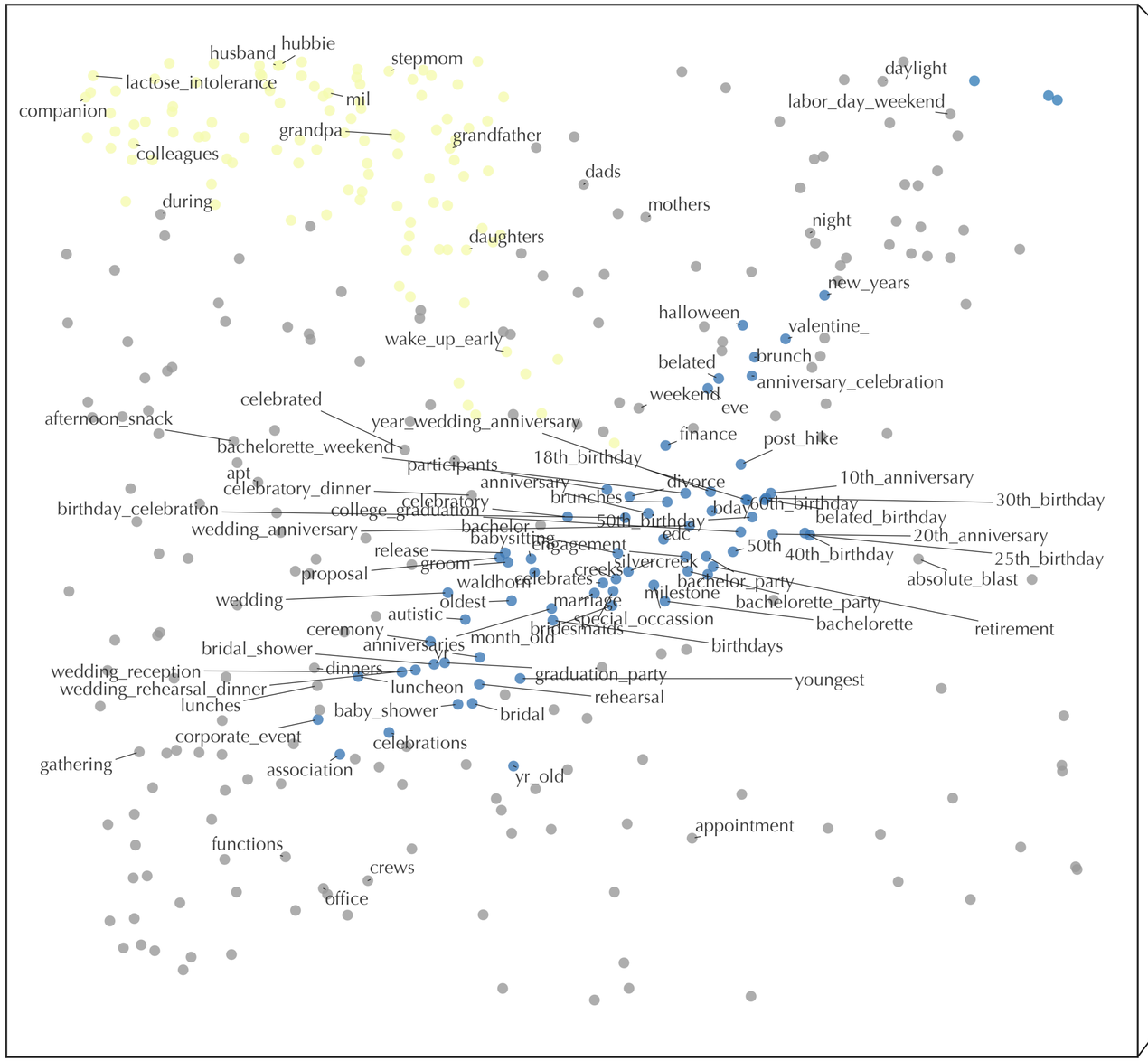
INSTITUT TUTTE INSTITUTE

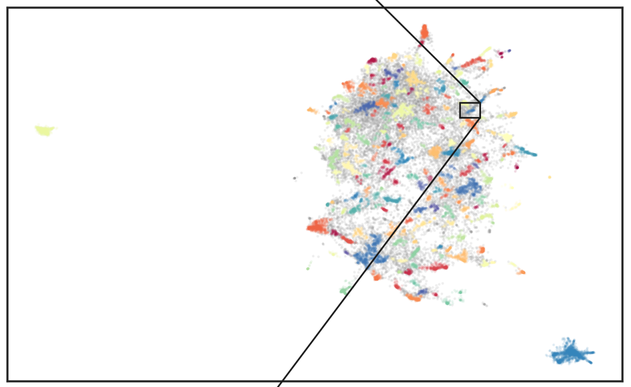


INSTITUT TUTTE INSTITUTE

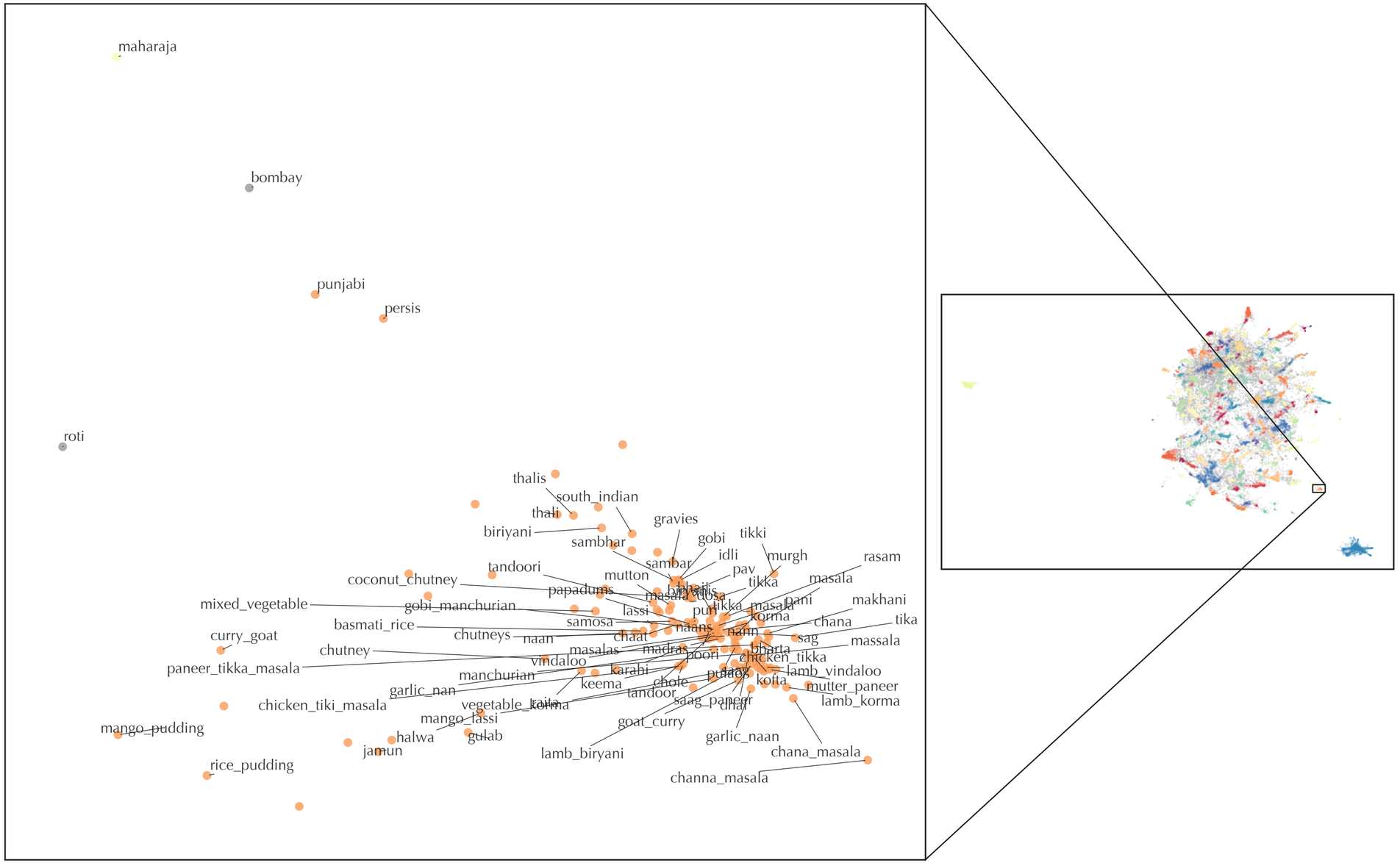


INSTITUT TUTTE INSTITUTE

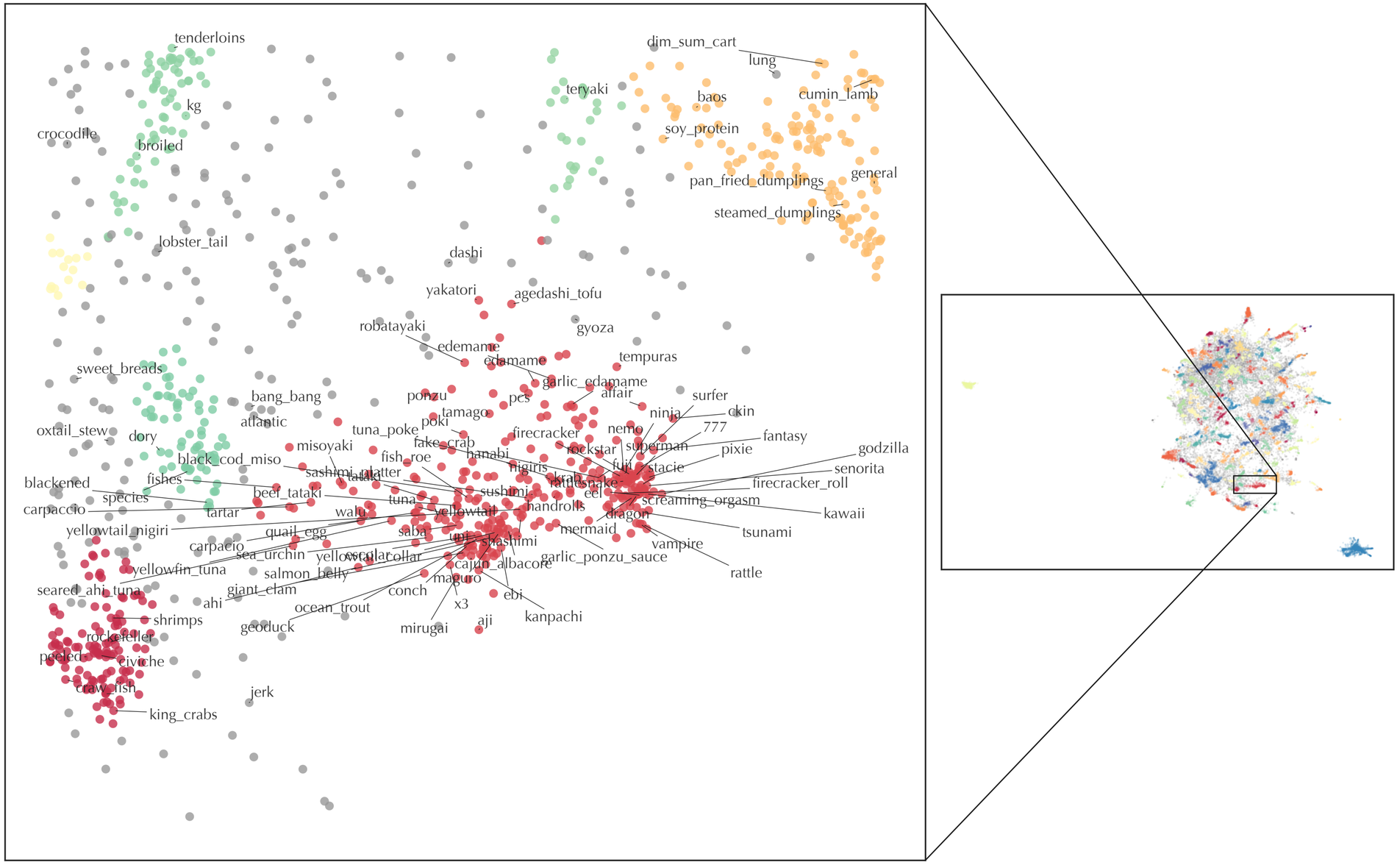




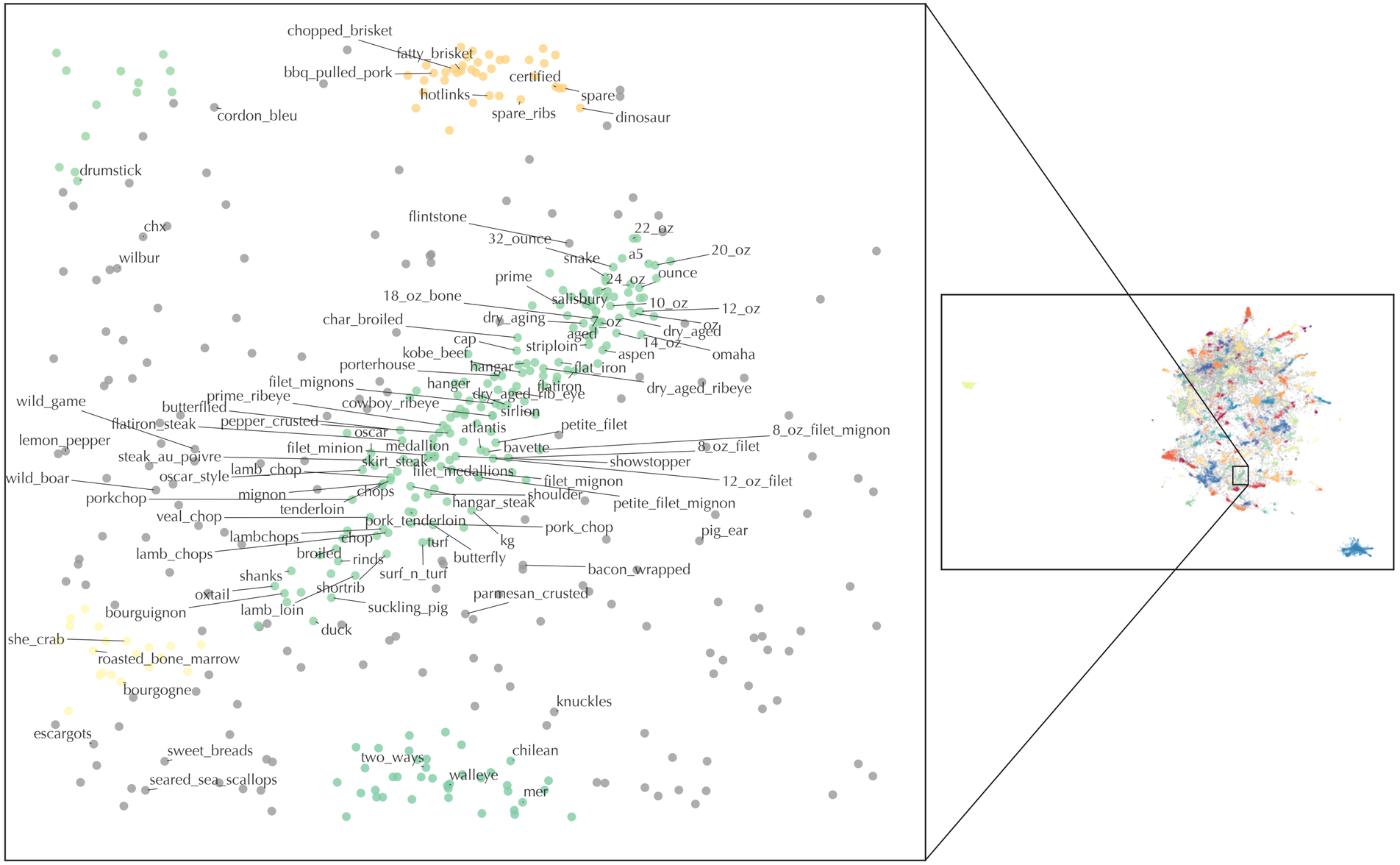
INSTITUT TUTTE INSTITUTE



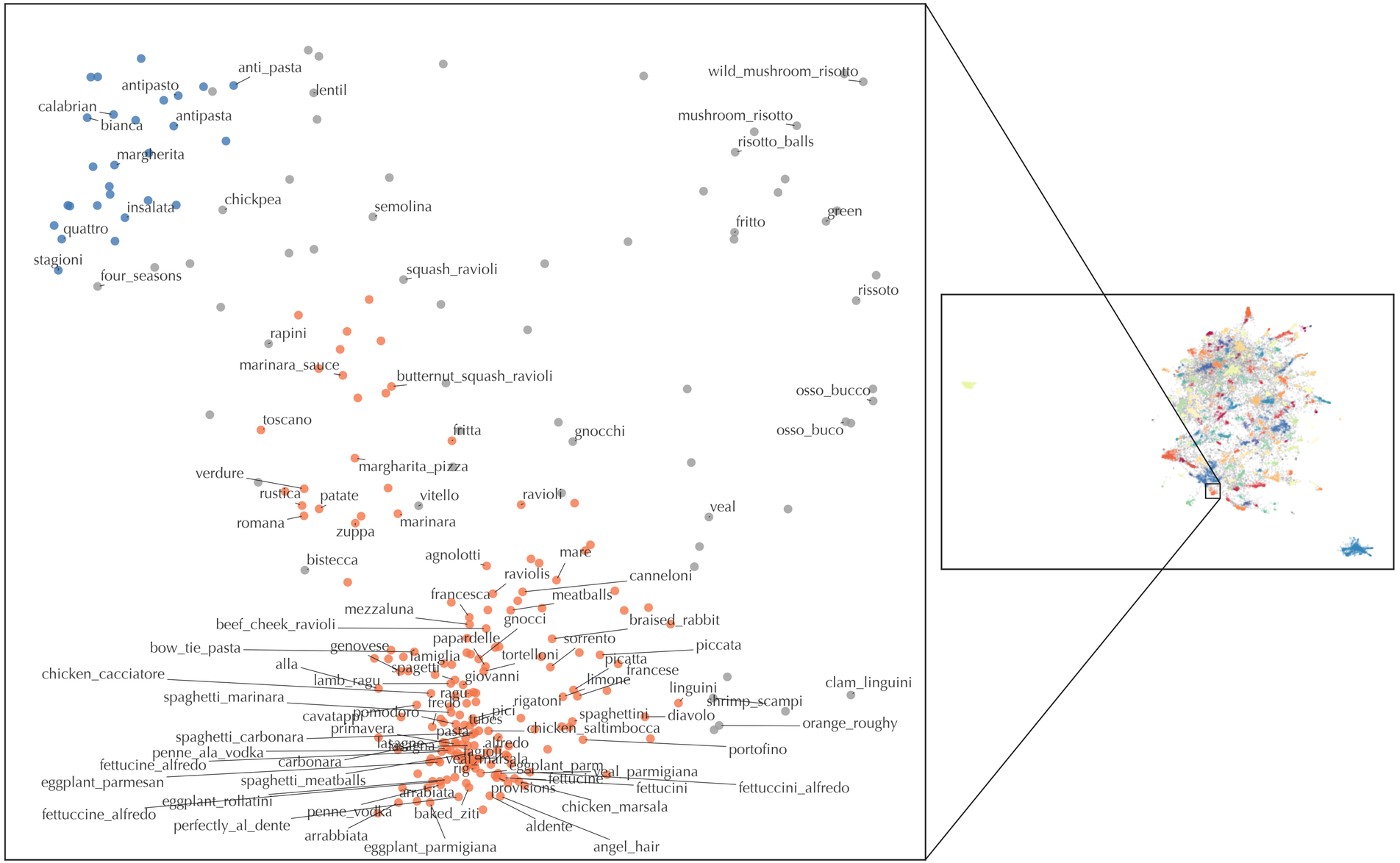
INSTITUT TUTTE INSTITUTE



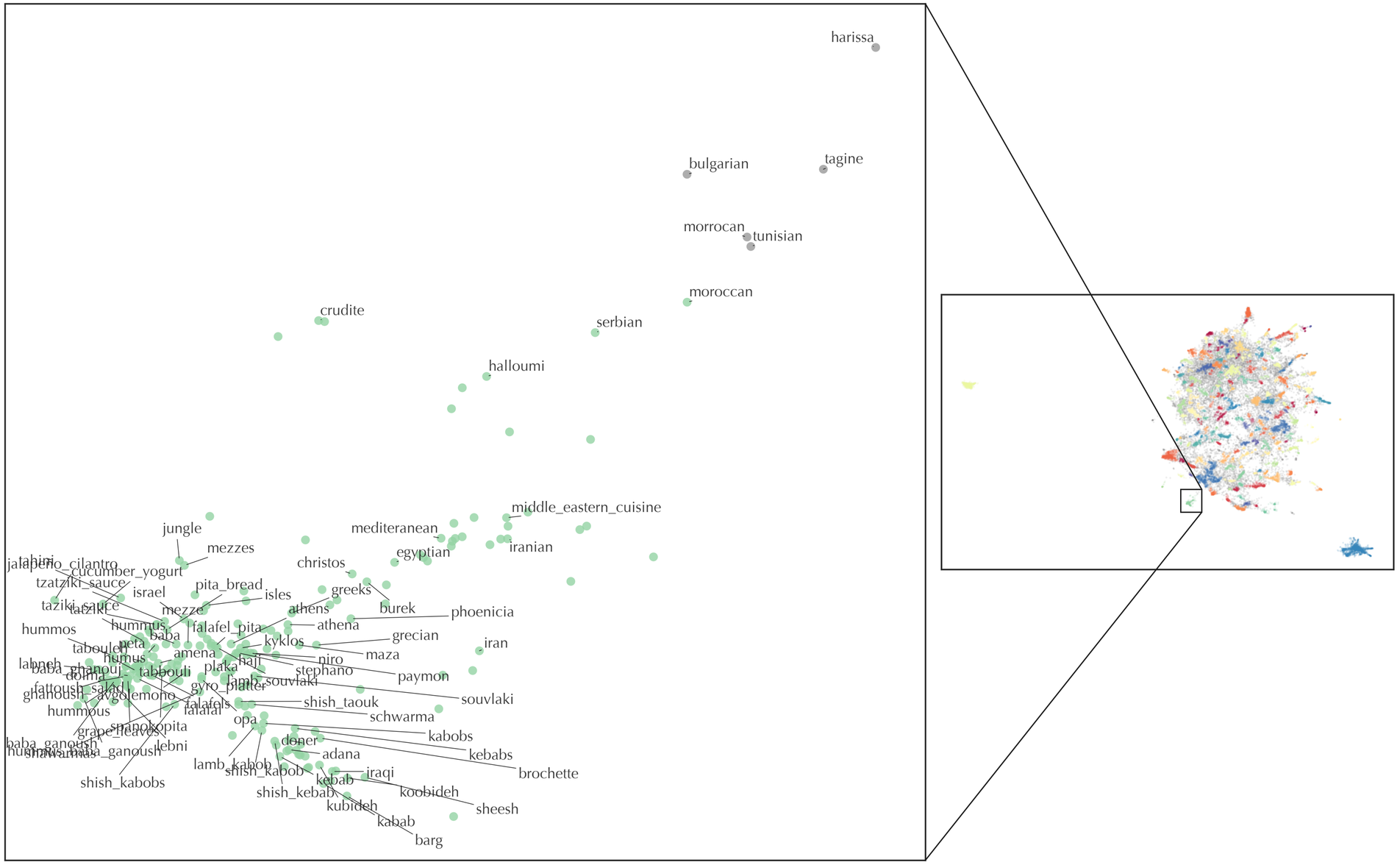
INSTITUT TUTTE INSTITUTE



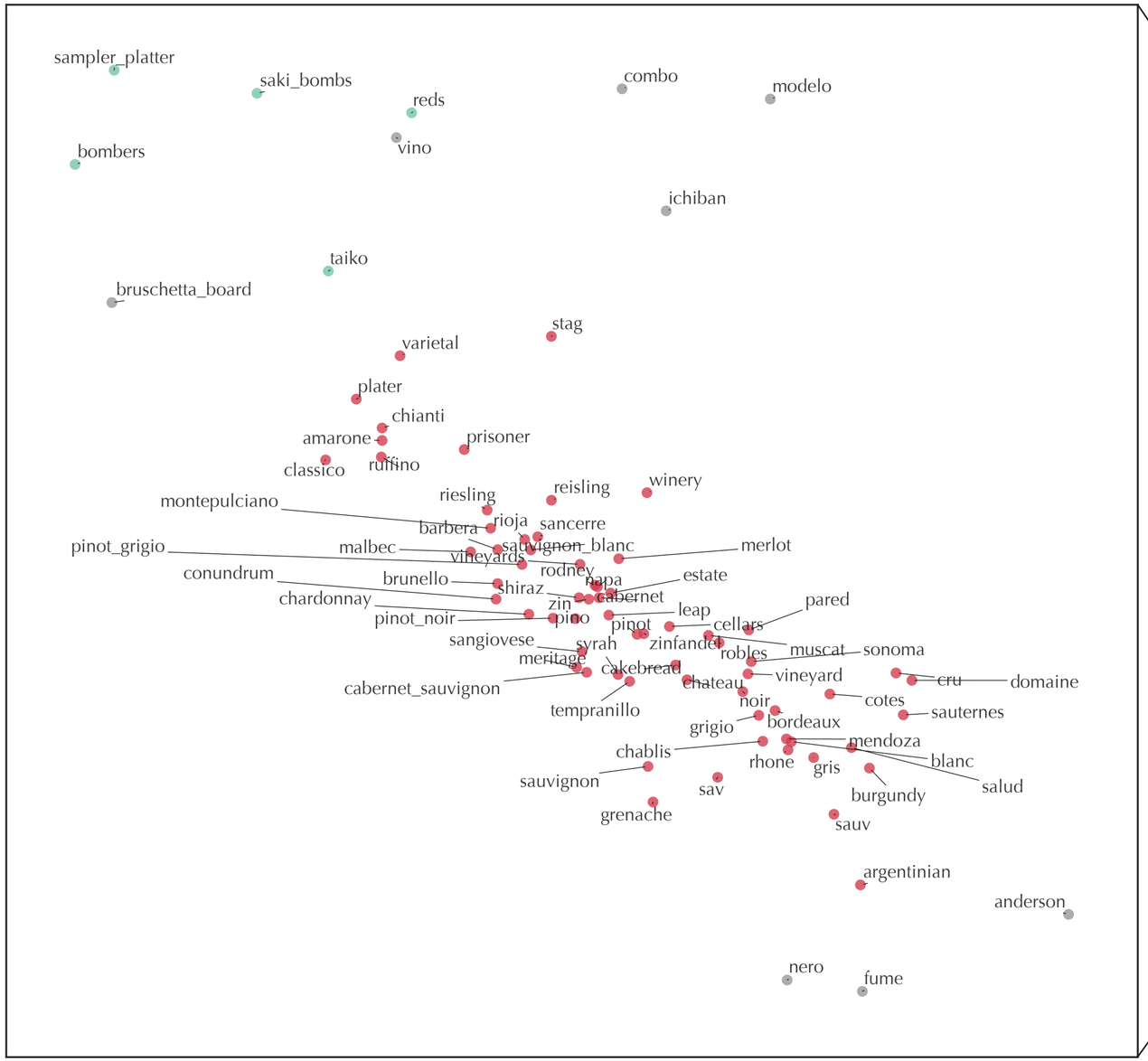
INSTITUT TUTTE INSTITUTE

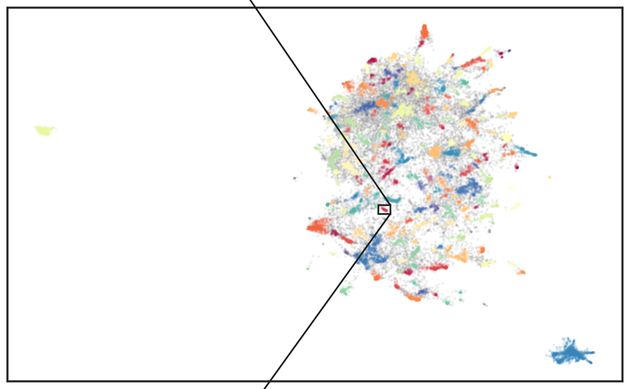


INSTITUT TUTTE INSTITUTE

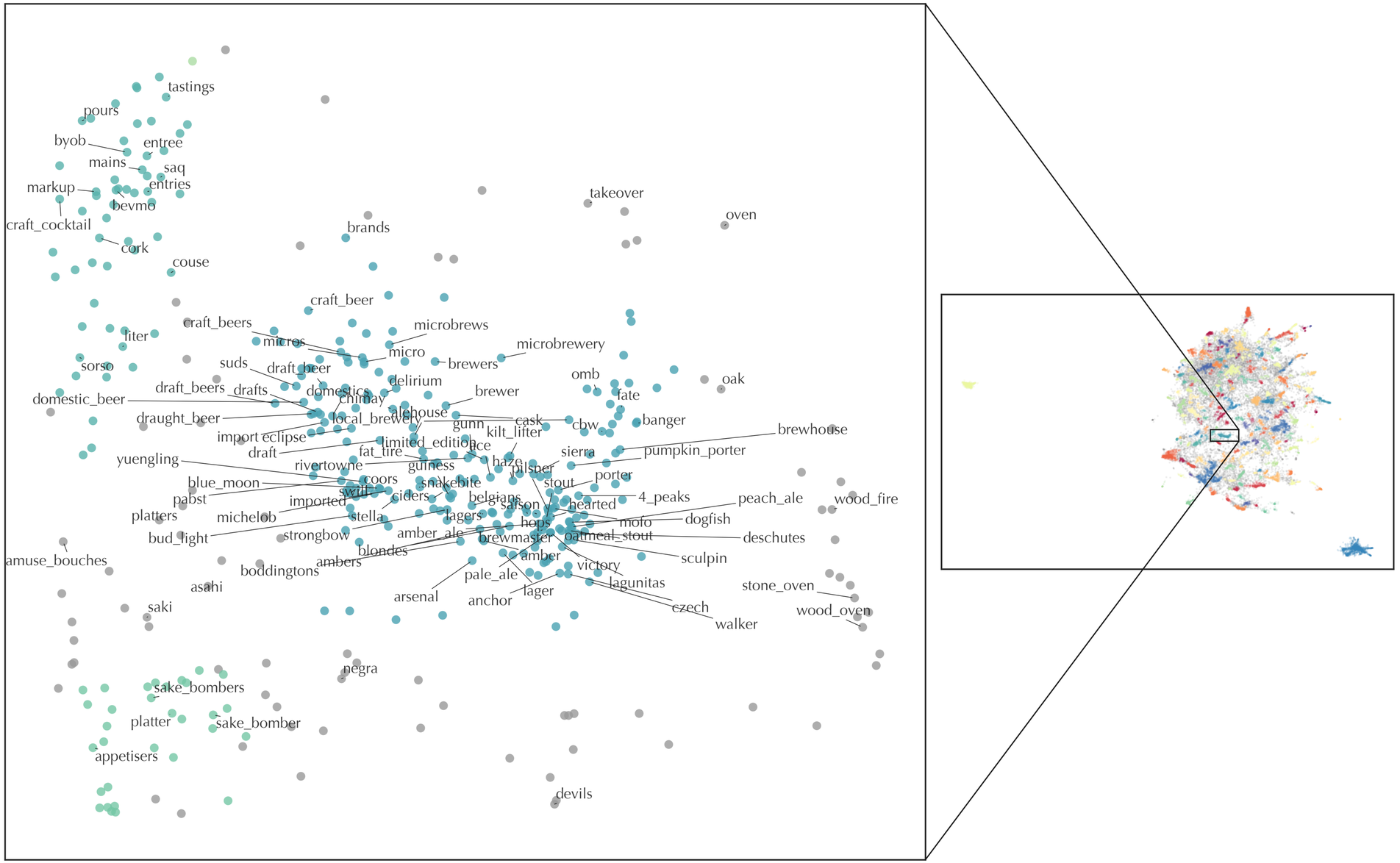


INSTITUT TUTTE INSTITUTE





INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

A lot of structure to explore!

What is the persistent homology of UMAP's simplicial sheaf representation?

INSTITUT TUTTE INSTITUTE

Documents as Distributions

INSTITUT TUTTE INSTITUTE

Represent a document using a “bag-of-words” model

INSTITUT TUTTE INSTITUTE

Represent a document as the multinomial distribution of words occurring in the document

INSTITUT TUTTE INSTITUTE

Information geometry tells us that multinomial distributions live on a manifold

INSTITUT TUTTE INSTITUTE

$$d(p,q) = \arccos\left(\sum_i \sqrt{p_i}\sqrt{q_i}\right)$$

INSTITUT TUTTE INSTITUTE

Some problems:

- Word order matters
- Words carry different amounts of information
- Language use is noisy
- Words are not independent / orthogonal

INSTITUT TUTTE INSTITUTE

We can't fix word order issues this time

(yet)

INSTITUT TUTTE INSTITUTE

Information in words:

Approximate the information of a word in a given context and weight words accordingly

INSTITUT TUTTE INSTITUTE

Language is noisy:

For larger documents we can hope to have the noise wash out in the background

INSTITUT TUTTE INSTITUTE

This red rose smells sweet

That scarlet flower's fragrance is delightful

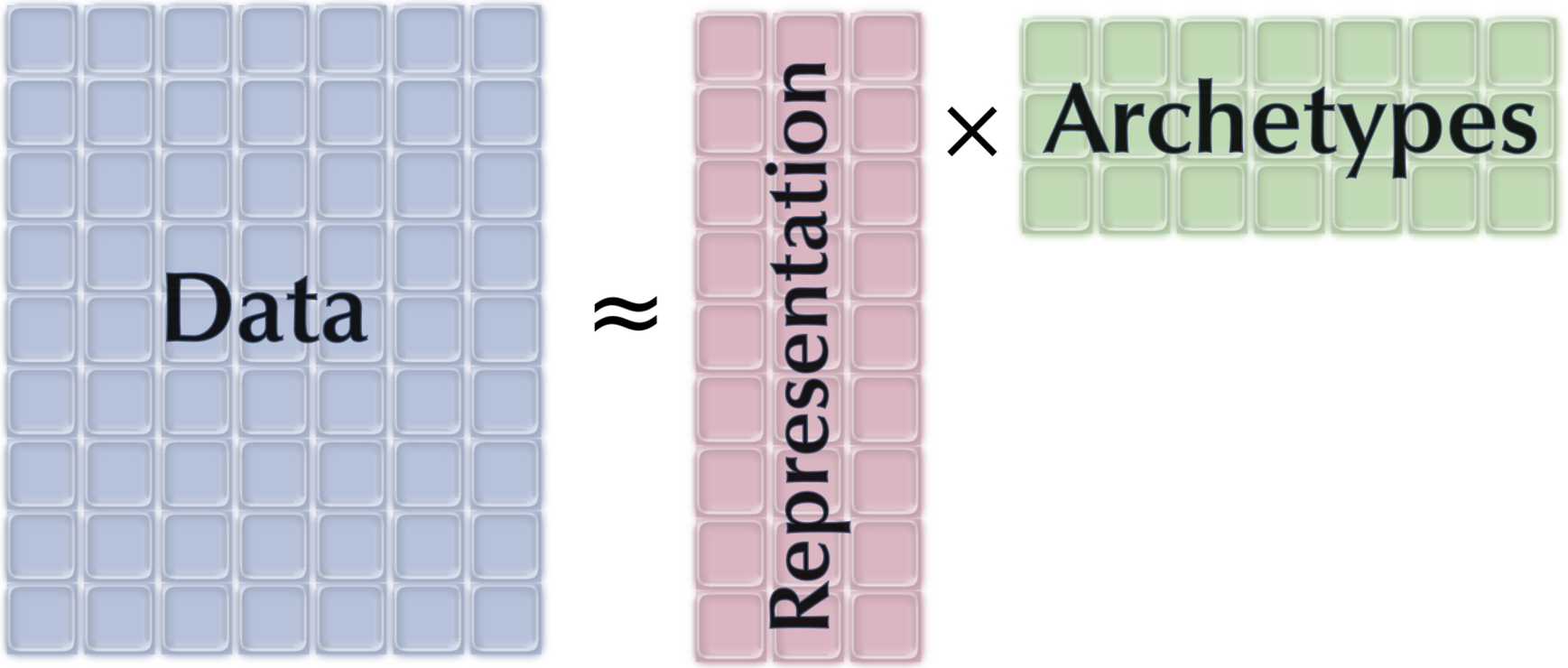
INSTITUT TUTTE INSTITUTE

A document as a multinomial is averaging one-hot encodings of words

INSTITUT TUTTE INSTITUTE

We can perform a change of basis to word vectors

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

$$D' = D \times W_{pLSA}$$

INSTITUT TUTTE INSTITUTE

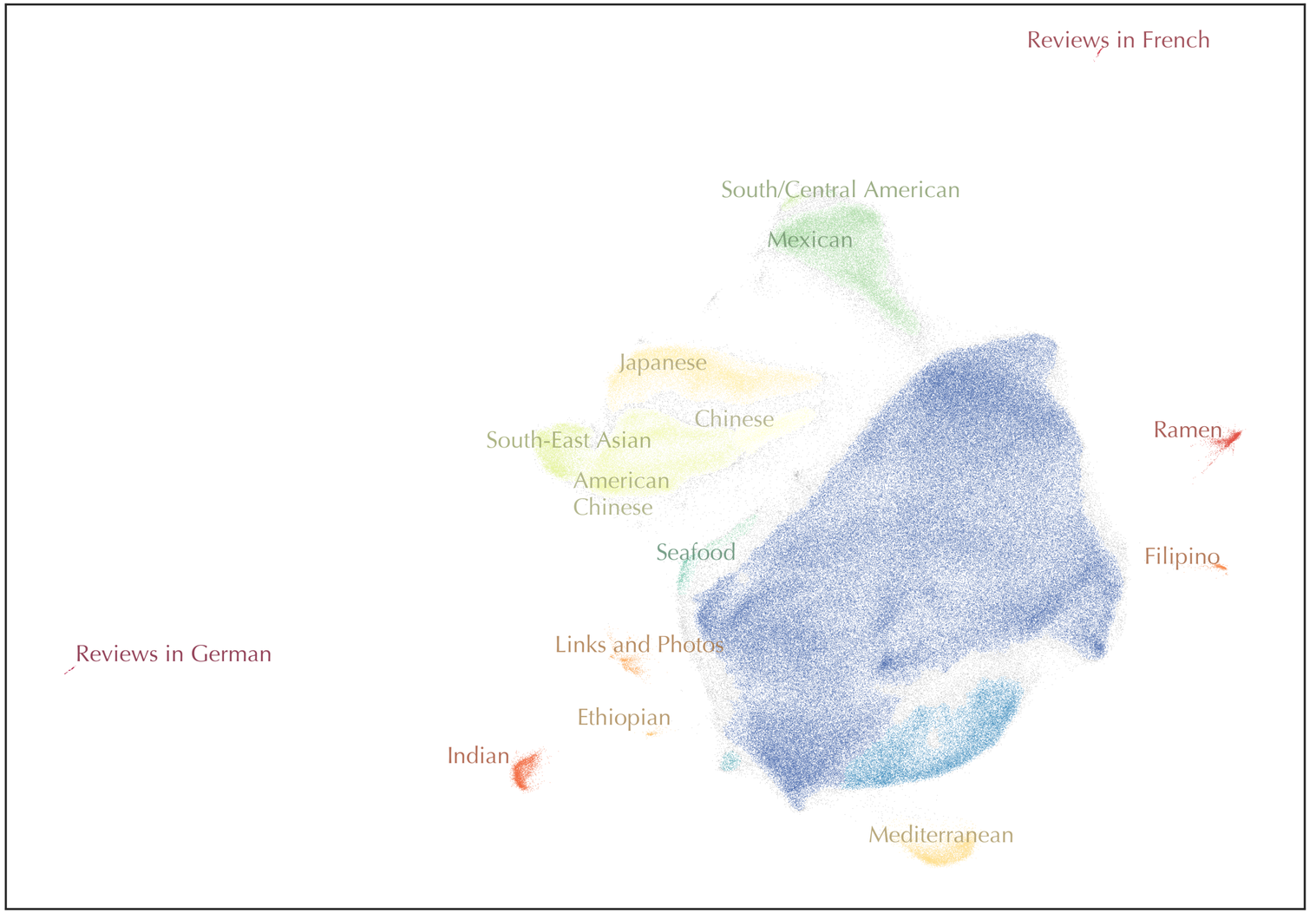
We are taking averages,
so we should remove
noise again …

INSTITUT TUTTE INSTITUTE

Embedding Documents

INSTITUT TUTTE INSTITUTE

UMAP is a topology based dimension reduction algorithm


Reviews in French
South/Central American
Mexican
Japanese
Chinese
South-East Asian
American
Chinese
Seafood
Ramen
Filipino
Reviews in German
Links and Photos
Ethiopian
Indian
Mediterranean

INSTITUT TUTTE INSTITUTE

Since words and documents now live in the same combined space we can embed both together

INSTITUT TUTTE INSTITUTE

This can provide powerful tools for topic modelling

INSTITUT TUTTE INSTITUTE

Conclusions

INSTITUT TUTTE INSTITUTE

We can use topological techniques to explore and represent language

INSTITUT TUTTE INSTITUTE

Many of these ideas provide a mathematical basis for variations on existing techniques

INSTITUT TUTTE INSTITUTE

This remains a rich field for further exploration