INSTITUT TUTTE INSTITUTE

Embed All the Things

Tutte Institute
for Mathematics and Computing

John Healy

jchealy@gmail.com

Joint work with Leland McInnes and Colin Weir

All the things

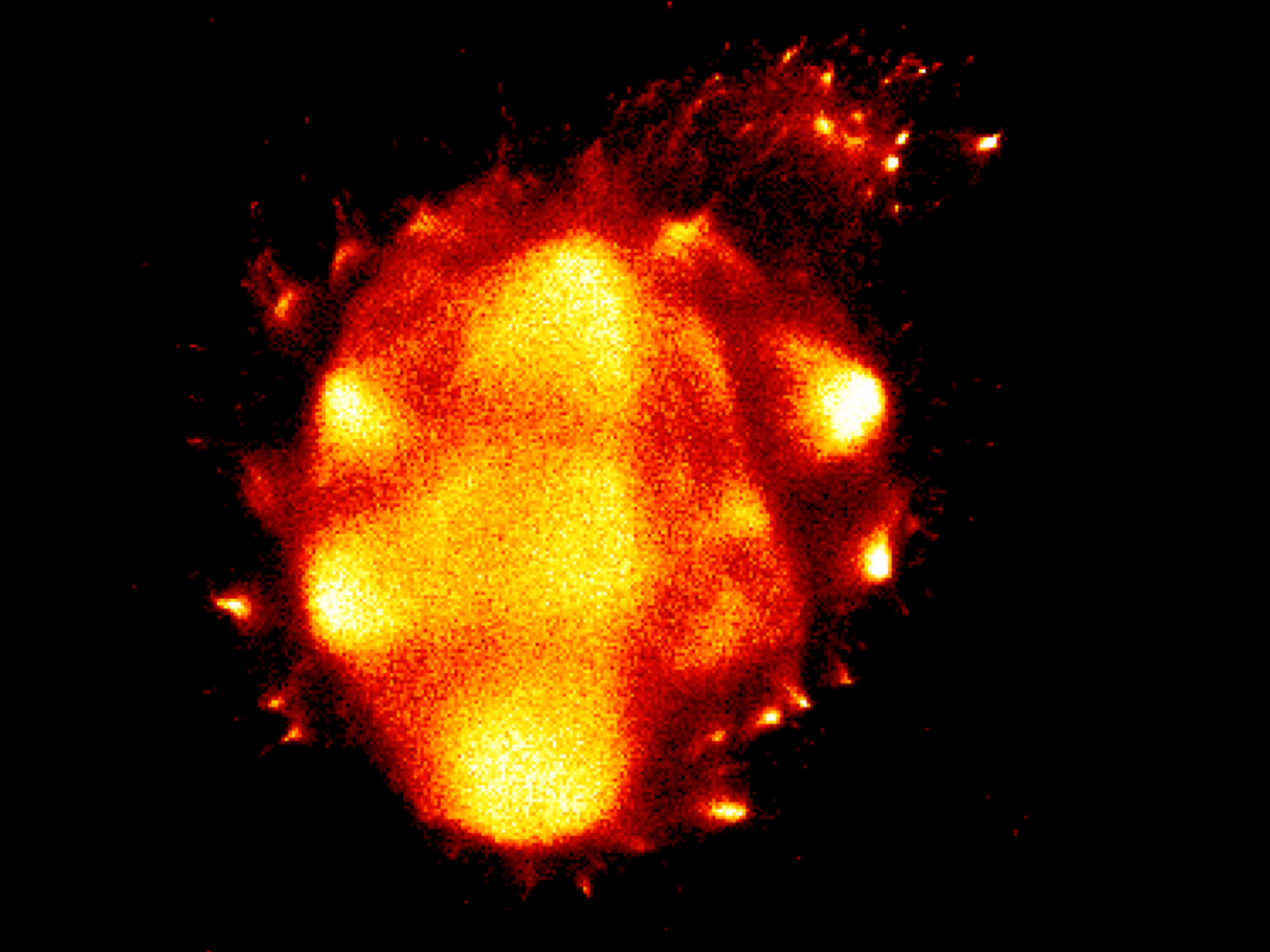
- Fixed width numeric data
- Variable length categorical data
- Documents
- Words

Notebooks and slides can be found @

https://github.com/jc-healy/EmbedAllTheThings

INSTITUT TUTTE INSTITUTE

What is an embedding?



INSTITUT TUTTE INSTITUTE

An embedding is a

numeric representation

of your data

along with a

distance

INSTITUT TUTTE INSTITUTE

Why do I care?

INSTITUT TUTTE INSTITUTE

Clustering
Grouping

Outlier Detection
Anomaly Detection

Unsupervised Learning

Exploratory Data Analysis
Visualization

INSTITUT TUTTE INSTITUTE

Because Math

An embedding is a

numeric representation

of your data

I have no labels

along with a

distance

INSTITUT TUTTE INSTITUTE

Data

INSTITUT TUTTE INSTITUTE

https://www.kaggle.com/rdoume/beerreviews

https://www.kaggle.com/rdoume/beerreviews/data

Search

kaggle Search Competitions Datasets Notebooks Discussion Courses •••

Data Kernels (6) Discussion (1) Activity Metadata Download (172 MB) New Notebook

Usability 7.1 Tags business, computing, food and drink, alcohol, consumer resources

Description

Context

This is the dataset disccused in the talk " How to hire and test for data skills: A one-size-fits-all interview kit" from https://conferences.oreilly.com/strata/strata-ny-2017/public/schedule/detail/59542

INSTITUT TUTTE INSTITUTE

1.5 million rows

```
reviews.head().T
```

	0	1	2	3	4
brewery_id	10325	10325	10325	10325	1075
brewery_name	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Caldera Brewing Company
review_time	1234817823	1235915097	1235916604	1234725145	1293735206
review_overall	1.5	3	3	3	4
review_aroma	2	2.5	2.5	3	4.5
review_appearance	2.5	3	3	3.5	4
review_profilename	stcules	stcules	stcules	stcules	johnmichaelsen
beer_style	Hefeweizen	English Strong Ale	Foreign / Export Stout	German Pilsener	American Double / Imperial IPA
review_palate	1.5	3	3	2.5	4
review_taste	1.5	3	3	3	4.5
beer_name	Sausa Weizen	Red Moon	Black Horse Black Beer	Sausa Pils	Cauldron DIPA
beer_abv	5	6.2	6.5	5	7.7
beer_beerid	47986	48213	48215	47969	64883

INSTITUT TUTTE INSTITUTE

Fixed width numeric data

Part I: Representation

INSTITUT TUTTE INSTITUTE

1.5 million rows

```
reviews.head().T
```

	0	1	2	3	4
brewery_id	10325	10325	10325	10325	1075
brewery_name	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Caldera Brewing Company
review_time	1234817823	1235915097	1235916604	1234725145	1293735206
review_overall	1.5	3	3	3	4
review_aroma	2	2.5	2.5	3	4.5
review_appearance	2.5	3	3	3.5	4
review_profilename	stcules	stcules	stcules	stcules	johnmichaelsen
beer_style	Hefeweizen	English Strong Ale	Foreign / Export Stout	German Pilsener	American Double / Imperial IPA
review_palate	1.5	3	3	2.5	4
review_taste	1.5	3	3	3	4.5
beer_name	Sausa Weizen	Red Moon	Black Horse Black Beer	Sausa Pils	Cauldron DIPA
beer_abv	5	6.2	6.5	5	7.7
beer_beerid	47986	48213	48215	47969	64883

1.5 million rows

```
reviews.head().T
```

	0	1	2	3	4
brewery_id	10325	10325	10325	10325	1075
brewery_name	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Caldera Brewing Company
review_time	1234817823	1235915097	1235916604	1234725145	1293735206
review_overall	1.5	3	3	3	4
review_aroma	2	2.5	2.5	3	4.5
review_appearance	2.5	3	3	3.5	4
review_profilename	stcules	stcules	stcules	stcules	johnmichaelsen
beer_style	Hefeweizen	English Strong Ale	Foreign / Export Stout	German Pilsener	American Double / Imperial IPA
review_palate	1.5	3	3	2.5	4
review_taste	1.5	3	3	3	4.5
beer_name	Sausa Weizen	Red Moon	Black Horse Black Beer	Sausa Pils	Cauldron DIPA
beer_abv	5	6.2	6.5	5	7.7
beer_beerid	47986	48213	48215	47969	64883

But each of those variables is a distribution not a number.

Pandas groupby()

```
beer_style = reviews.groupby('beer_style').agg({
    'beer_name':lambda x: x.mode(),
    'brewery_name':lambda x: x.mode(),
    'beer_abv':'mean',
    'review_aroma':'mean',
    'review_appearance':'mean',
    'review_palate':'mean',
    'review_taste':'mean',
    'review_overall':'mean',
    'review_profilename':len
}).reset_index()

beer_style.columns = """beer_style common_beer common_brewer abv
aroma appearance overall palate taste
num_reviews""".split()
beer_style.shape
```

Pandas groupby()

```
beer_style.tail(2).T
```

104 rows

	103
beer_style	Witbier
common_beer	Hoegaarden Original White Ale
common_brewer	Boston Beer Company (Samuel Adams)
abv	5.47752
aroma	3.64088
appearance	3.6867
overall	3.62575
palate	3.65216
taste	3.77689
num_reviews	30140

INSTITUT TUTTE INSTITUTE

Fixed width numeric data

Part II: Distance

INSTITUT TUTTE INSTITUTE

Choose your distance

Jaccard

Haversine

Hamming

Hellinger

Cosine

Braycurtis

Minkowski

Correlation

Manhattan

Canberra

Euclidean

INSTITUT TUTTE INSTITUTE

Choose your distance

Jaccard

Haversine

Hamming

Hellinger

Cosine

Braycurtis

Minkowski

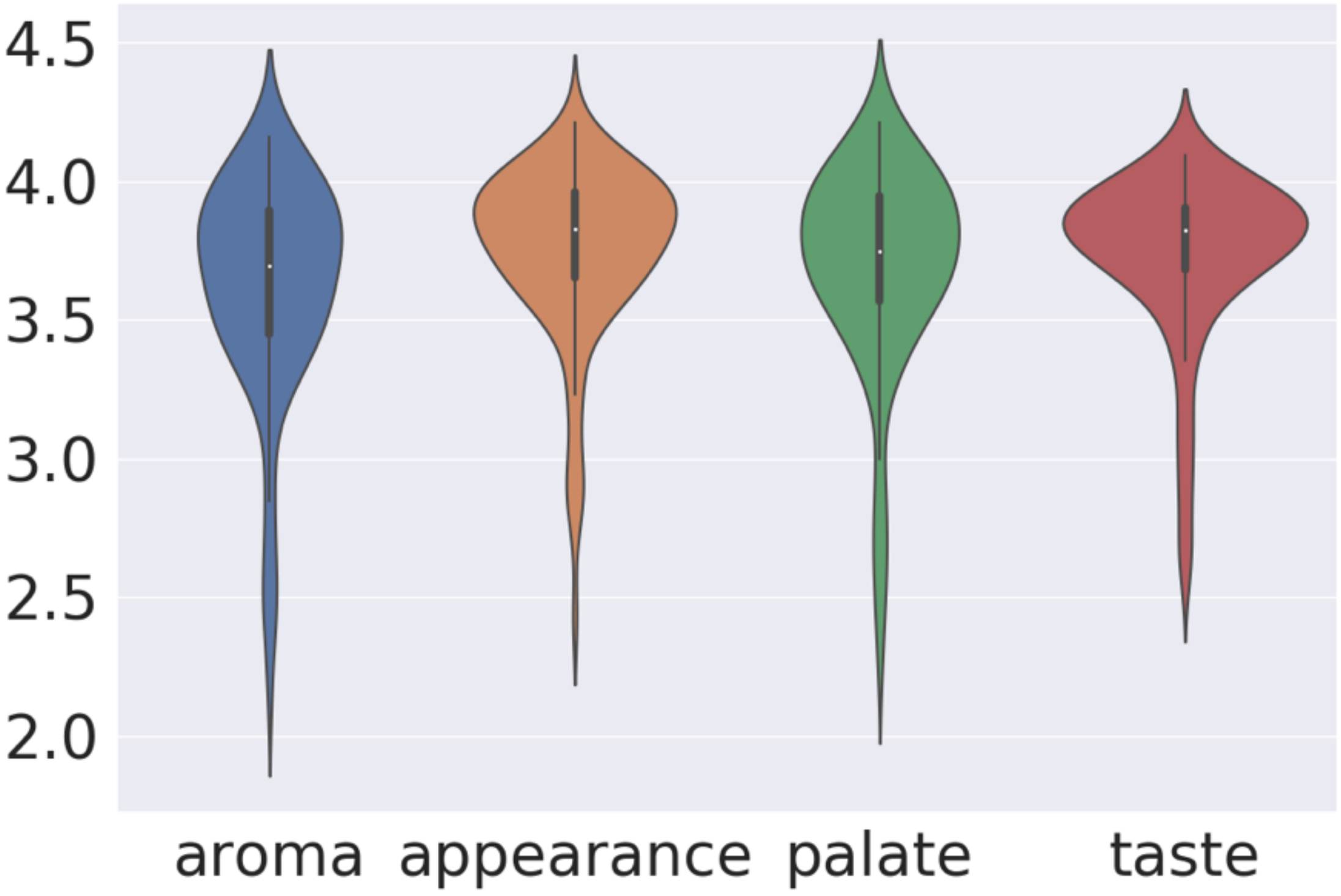
Correlation

Manhattan

Canberra

Euclidean

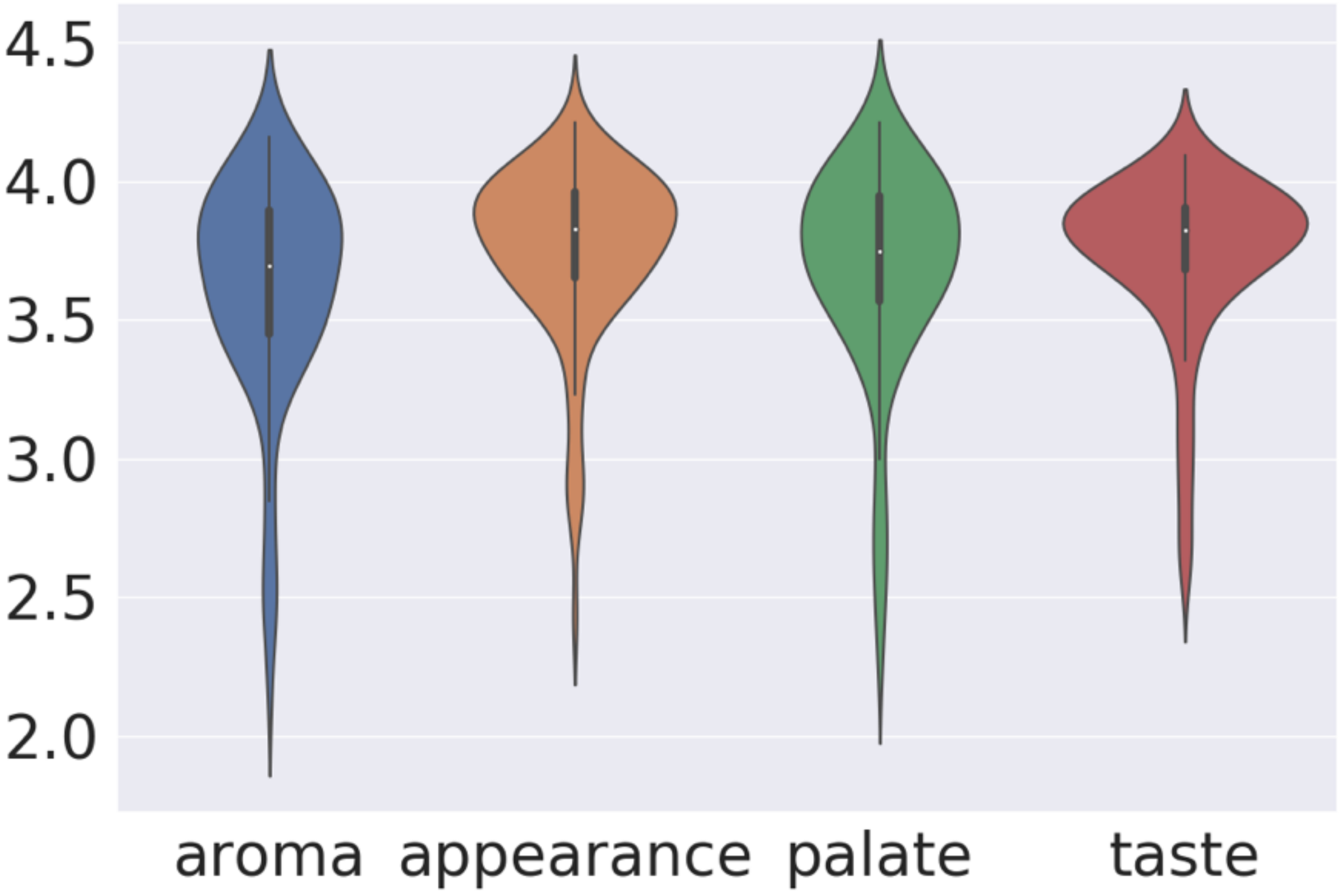
Look at your data



INSTITUT TUTTE INSTITUTE

Look at your data

Look at your data



INSTITUT TUTTE INSTITUTE

Beer styles are a **vector** of their mean scores along with manhattan **distance**

```
import hdbscan
clusterer = hdbscan.HDBSCAN(min_cluster_size=5, metric='manhattan')
clusters = clusterer.fit_predict(beer_style[numeric_columns])
```

INSTITUT TUTTE INSTITUTE

Dimension Reduction

UMAP

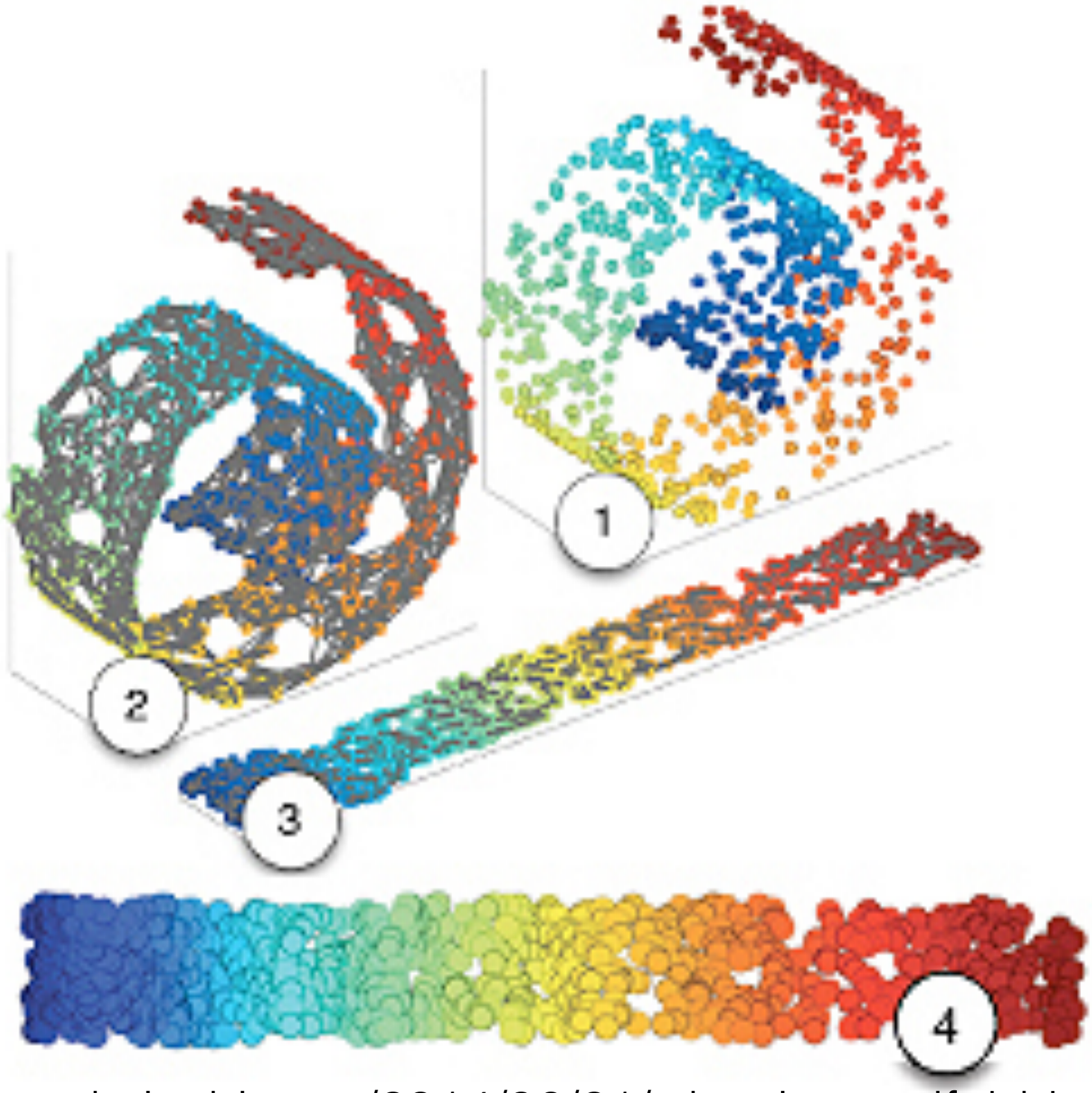
INSTITUT TUTTE INSTITUTE

Uniform Manifold Approximation and Projection

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
Leland McInnes, John Healy, James Melville, 2018
https://arxiv.org/abs/1802.03426

INSTITUT TUTTE INSTIT

Manifold Learning



https://prateekvjoshi.com/2014/06/21/what-is-manifold-learning/

INSTITUT TUTTE INSTITUTE

https://pair-code.github.io/understanding-umap/

UMAP

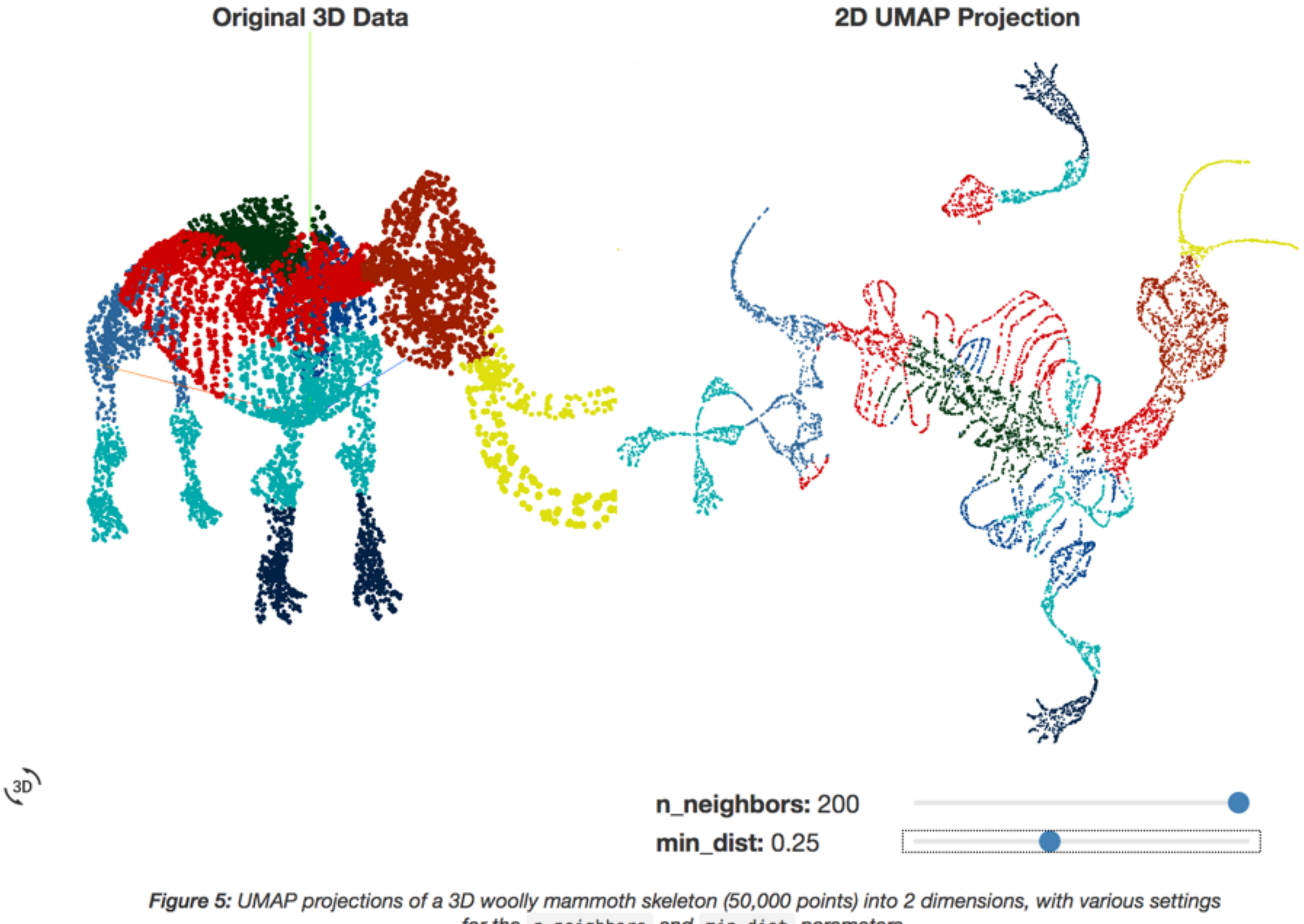


Figure 5: *UMAP projections of a 3D woolly mammoth skeleton (50,000 points) into 2 dimensions, with various settings for the* `n_neighbors` *and* `min_dist` *parameters.*

Read the docs

https://umap-learn.readthedocs.io/en/latest/

Search

latest

Search docs

USER GUIDE / TUTORIAL:

- How to Use UMAP
- Basic UMAP Parameters
- Transforming New Data with UMAP
- UMAP for Supervised Dimension Reduction and Metric Learning
- Using UMAP for Clustering
- Gallery of Examples of UMAP usage
- Frequently Asked Questions

BACKGROUND ON UMAP:

Read the Docs v: latest

Docs » UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Edit on GitHub

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction. The algorithm is founded on three assumptions about the data

1. The data is uniformly distributed on Riemannian manifold;
2. The Riemannian metric is locally constant (or can be approximated as such);
3. The manifold is locally connected.

From these assumptions it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.

The details for the underlying mathematics can be found in our paper on ArXiv:

McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

https://www.youtube.com/watch?v=nq6iPZVUxZU

UMAP Uniform Manifold Approximation and Projection for Dimension Reduction | SciPy 2018 |

27,585 views • Jul 13, 2018 712 4 SHARE SAVE

pip install umap-learn

GitHub, Inc. (US) https://github.com/lmcinnes/umap

Search

lmcinnes / umap

Unwatch 119 | Unstar 3.4k | Fork 337

Code | Issues 97 | Pull requests 4 | Actions | Projects 0 | Wiki | Security | Insights

Uniform Manifold Approximation and Projection

umap | dimensionality-reduction | visualization | machine-learning | topological-data-analysis

569 commits | 15 branches | 0 packages | 14 releases | 32 contributors | BSD-3-Clause

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

lmcinnes Merge pull request #307 from felixdivo/patch-1 ... Latest commit c26bc8d on Oct 9

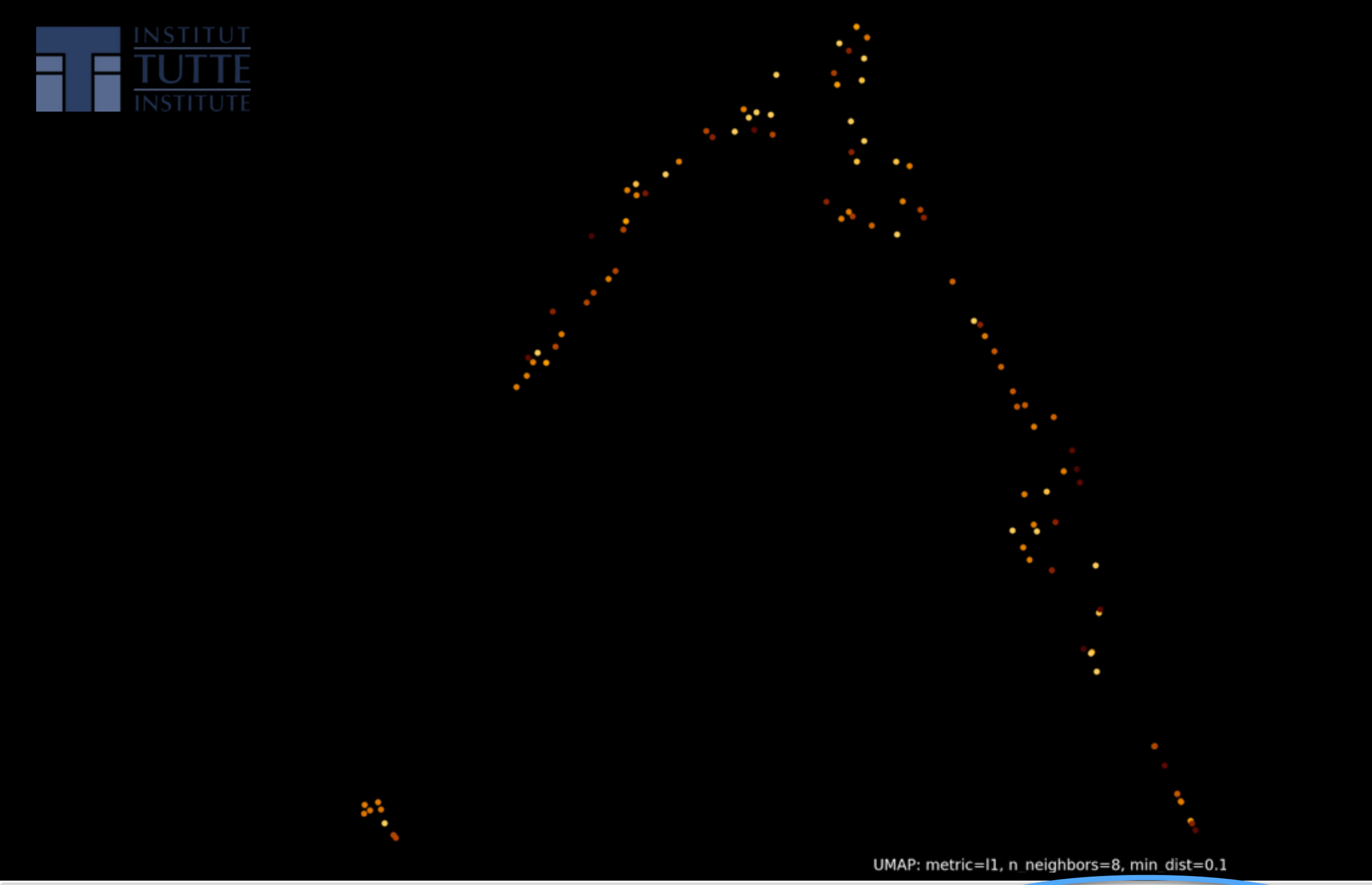
.idea	Code style	2 years ago
ci_scripts	Merge pull request #215 from tomwhite/enforce-black	5 months ago
doc	Fix order of algorithms in caption to be less confusing	2 months ago
examples	fix mnist example with fetch_openml call	6 months ago
images	Update mnist example	2 years ago
notebooks	Add notebook for generating parameter animations	2 years ago
umap	Merge branch 'master' into patch-1	3 months ago

INSTITUT TUTTE INSTITUTE

Fixed width numeric data

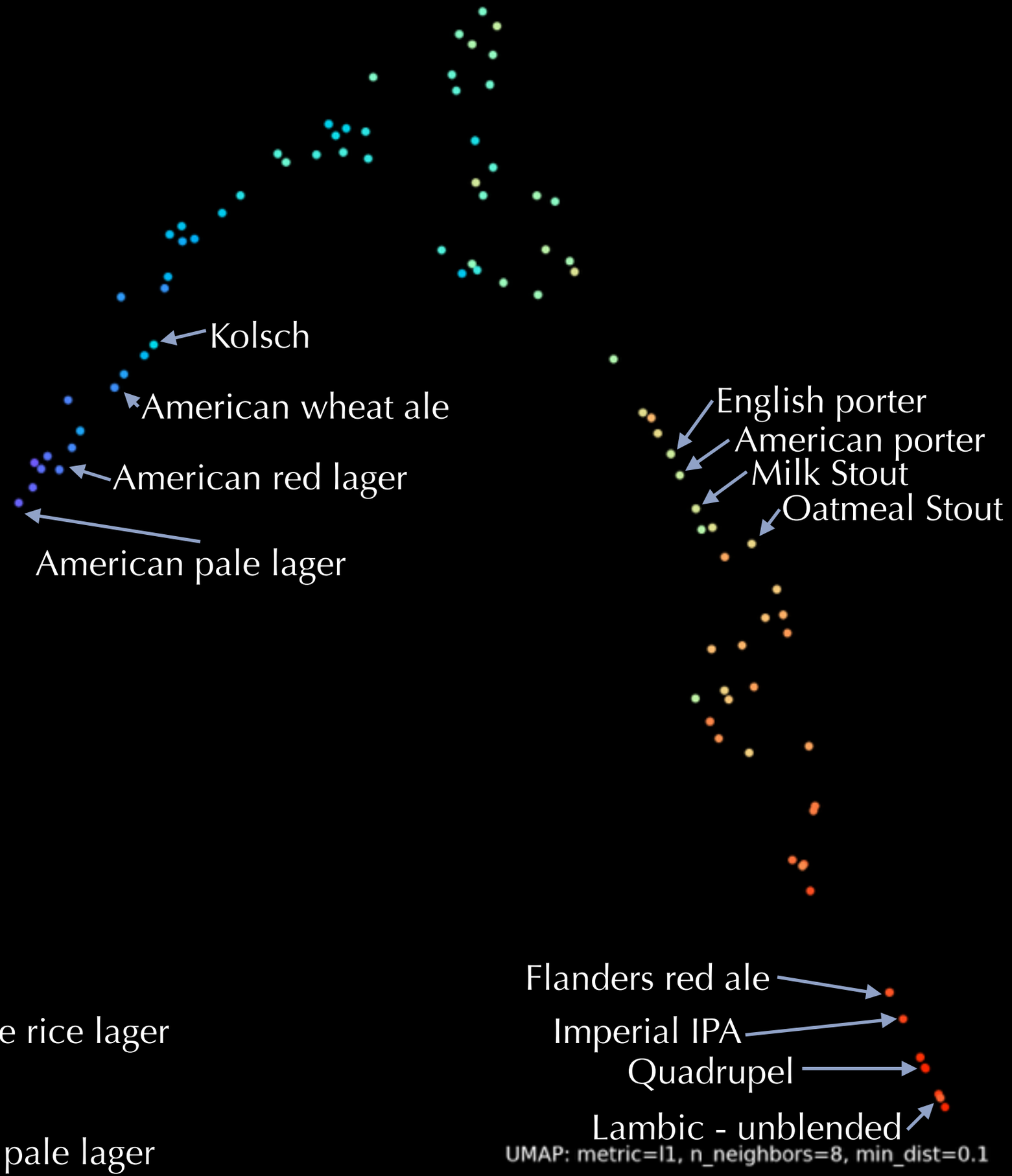
Part III: Look at your data

INSTITUT TUTTE INSTITUTE




```
style_by_reviews_model = umap.UMAP(n_neighbors=6, n_components=2, metric='l1',
                                   unique=False, random_state=42).fit(style_by_reviews)
umap_plot = umap.plot.points(style_by_reviews_model, color_key=beer_style.srm_rgb,
                             labels=beer_style.srm_rgb, theme='fire');
```

INSTITUT TUTTE INSTITUTE



INSTITUT TUTTE INSTITUTE

Variable width categorical data

Part I: Representation

INSTITUT TUTTE INSTITUTE

Choose your data

1.5 million rows

```
reviews.head().T
```

	0	1	2	3	4
brewery_id	10325	10325	10325	10325	1075
brewery_name	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Vecchio Birraio	Caldera Brewing Company
review_time	1234817823	1235915097	1235916604	1234725145	1293735206
review_overall	1.5	3	3	3	4
review_aroma	2	2.5	2.5	3	4.5
review_appearance	2.5	3	3	3.5	4
review_profilename	stcules	stcules	stcules	stcules	johnmichaelsen
beer_style	Hefeweizen	English Strong Ale	Foreign / Export Stout	German Pilsener	American Double / Imperial IPA
review_palate	1.5	3	3	2.5	4
review_taste	1.5	3	3	3	4.5
beer_name	Sausa Weizen	Red Moon	Black Horse Black Beer	Sausa Pils	Cauldron DIPA
beer_abv	5	6.2	6.5	5	7.7
beer_beerid	47986	48213	48215	47969	64883

INSTITUT TUTTE INSTITUTE

Shape your data

```
unique_join = lambda x: join(x.unique(), " ")
beer_style = reviews.groupby('beer_style').agg({
    'beer_name':lambda x: x.mode(),
    'brewery_name':lambda x: x.mode(),
    'beer_abv':'mean',
    'review_aroma':'mean',
    'review_appearance':'mean',
    'review_overall':'mean',
    'review_palate':'mean',
    'review_taste':'mean',
    'review_profilename':[unique_join, len],
    'brewery_id':lambda x: len(x.unique()),
}).reset_index()

beer_style.columns = """beer_style beer_name brewery_name beer_abv
review_aroma review_appearance review_overall review_palate review_taste
review_profilename_list num_reviewers num_ids""".split()
beer_style.shape
```

Shape your data

```
unique_join = lambda x: join(x.unique(), " ")
beer_style = reviews.groupby('beer_style').agg({
    'beer_name':lambda x: x.mode(),
    'brewery_name':lambda x: x.mode(),
    'beer_abv':'mean',
    'review_aroma':'mean',
    'review_appearance':'mean',
    'review_overall':'mean',
    'review_palate':'mean',
    'review_taste':'mean',
    'review_profilename':[unique_join, len],
    'brewery_id':lambda x: len(x.unique()),
}).reset_index()

beer_style.columns = """beer_style beer_name brewery_name beer_abv
review_aroma review_appearance review_overall review_palate review_taste
review_profilename_list num_reviewers num_ids""".split()
beer_style.shape
```

Pandas groupby()

```
beer_style.head(2).T
```

104 rows

	0	
beer_style	Altbier	
beer_name	Alaskan Amber	
brewery_name	Uerige Obergärige Hausbrauerei GmbH / Zum Uerige	
beer_abv	5.82858	
review_aroma	3.62401	
review_appearance	3.80933	
review_overall	3.82405	
review_palate	3.71309	
review_taste	3.74487	
review_profilename_list	charlatan kmacphail BSF foamer JamesS SolomonG...	isualum12 BeerAdvoc
review_profilename_len	7741	

sklearn CountVectorize()

```
popular_beer = beer[beer.review_profilename_len>10].reset_index(drop=True)
beer_by_authors_vectorizer = CountVectorizer(binary=True, min_df=10)
beer_by_authors = beer_by_authors_vectorizer.fit_transform(popular_beer.review_profilename_list)
beer_by_authors
```

```
<13389x10451 sparse matrix of type '<class 'numpy.int64'>'
        with 1376374 stored elements in Compressed Sparse Row format>
```

```
beer_by_authors[:10,:20].todense()
```

```
matrix([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

Beer style

Reviewers

INSTITUT TUTTE INSTITUTE

Beer styles are a bag of **reviewers**

INSTITUT TUTTE INSTITUTE

Variable width categorical data

Part II: Distance

INSTITUT TUTTE INSTITUTE

Question:

Do counts matter?

INSTITUT TUTTE INSTITUTE

Counts don't matter

Beer styles are a ~~bag~~ of **reviewers**

Set

INSTITUT TUTTE INSTITUTE

Beer Styles are a set of **reviewers**

Distance = Jaccard

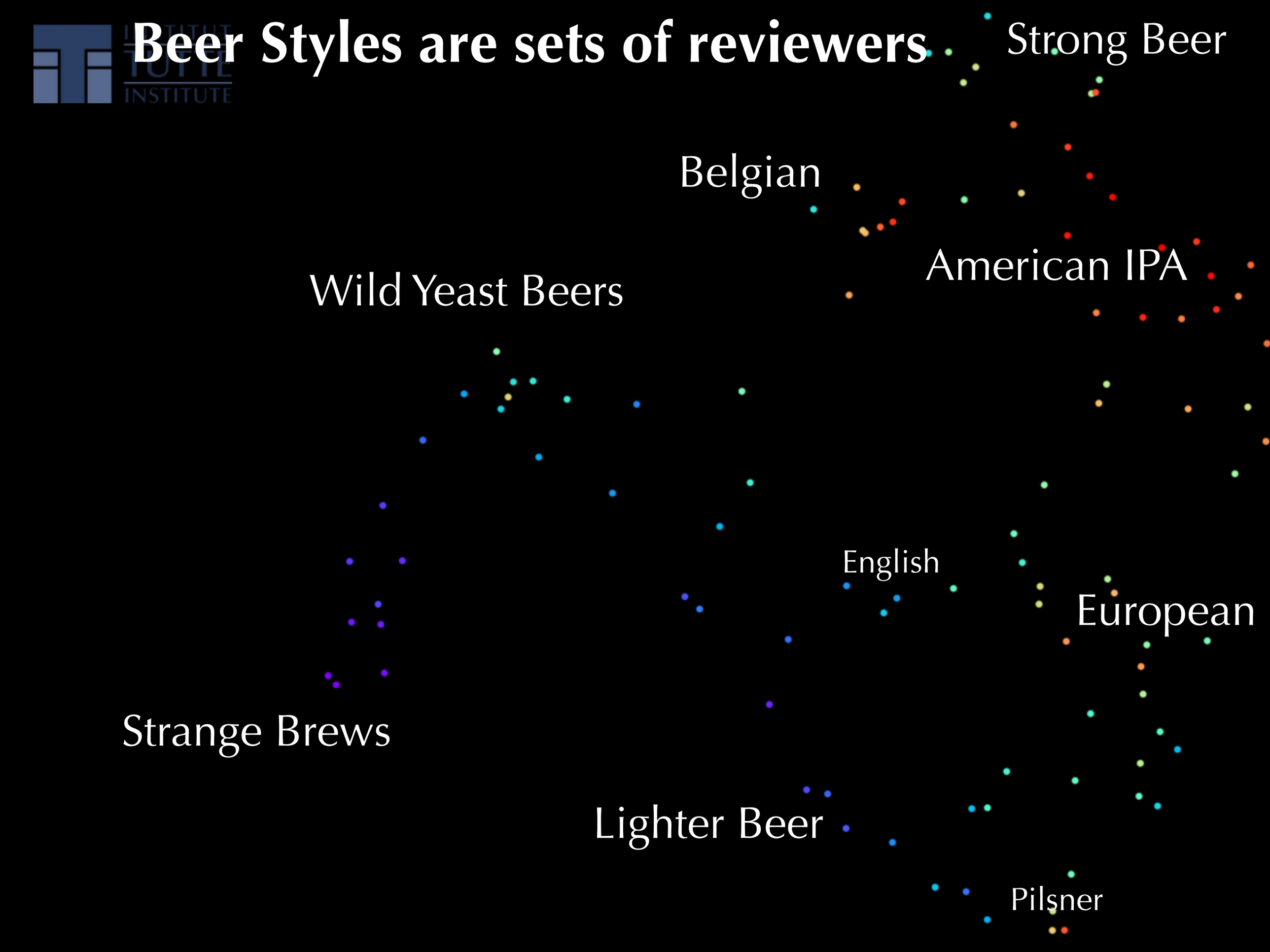
$$J(A,B) = 1 - \frac{A \cap B}{A \cup B}$$

INSTITUT TUTTE INSTITUTE

Variable width categorical data

Part III: Look at your data

Beer Styles are sets of reviewers



INSTITUT TUTTE INSTITUTE

But that was a general process

INSTITUT TUTTE INSTITUTE

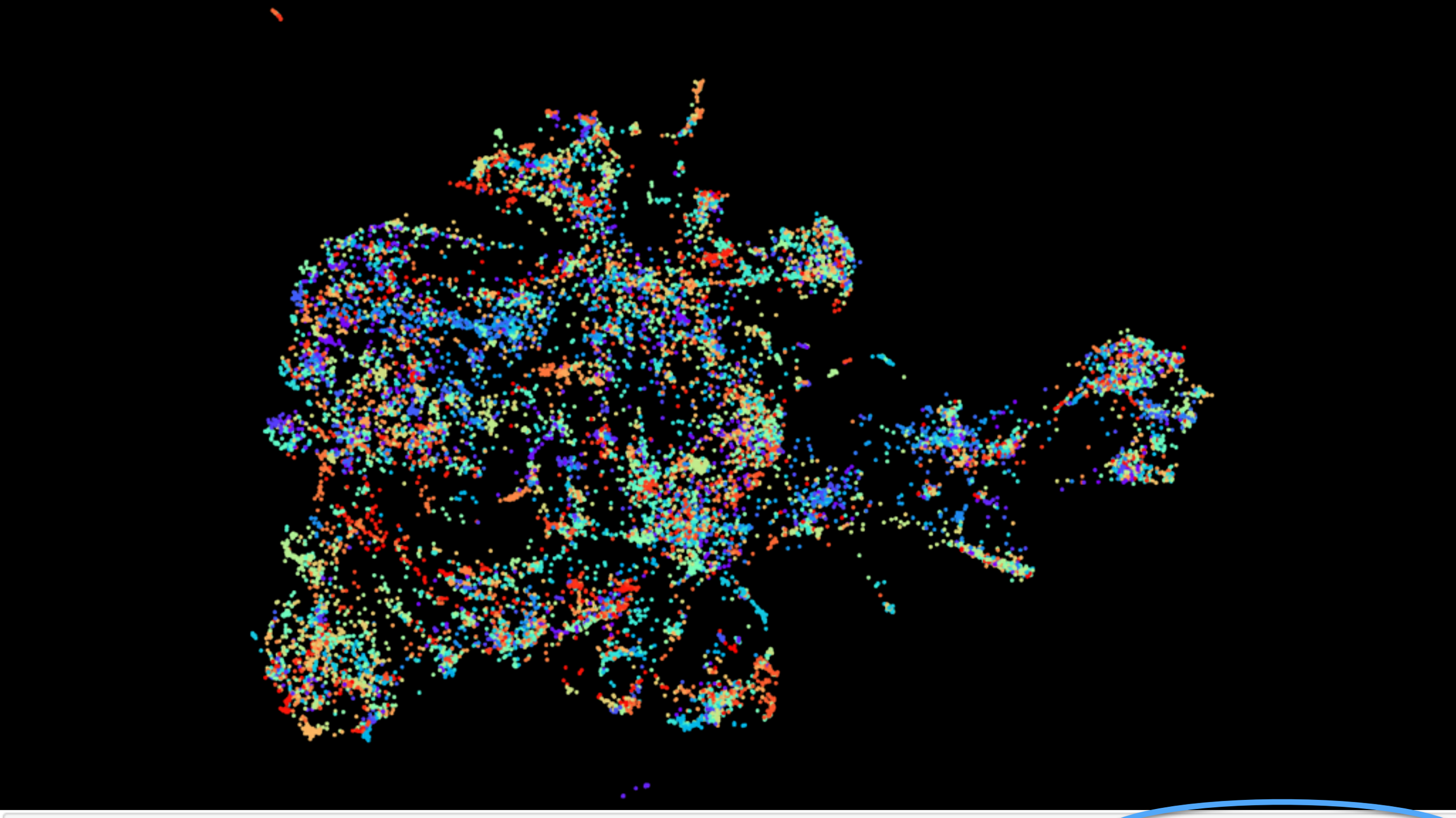
Beers are a set of **reviewers**

Distance = Jaccard

$$J(A,B) = 1 - \frac{A \cap B}{A \cup B}$$

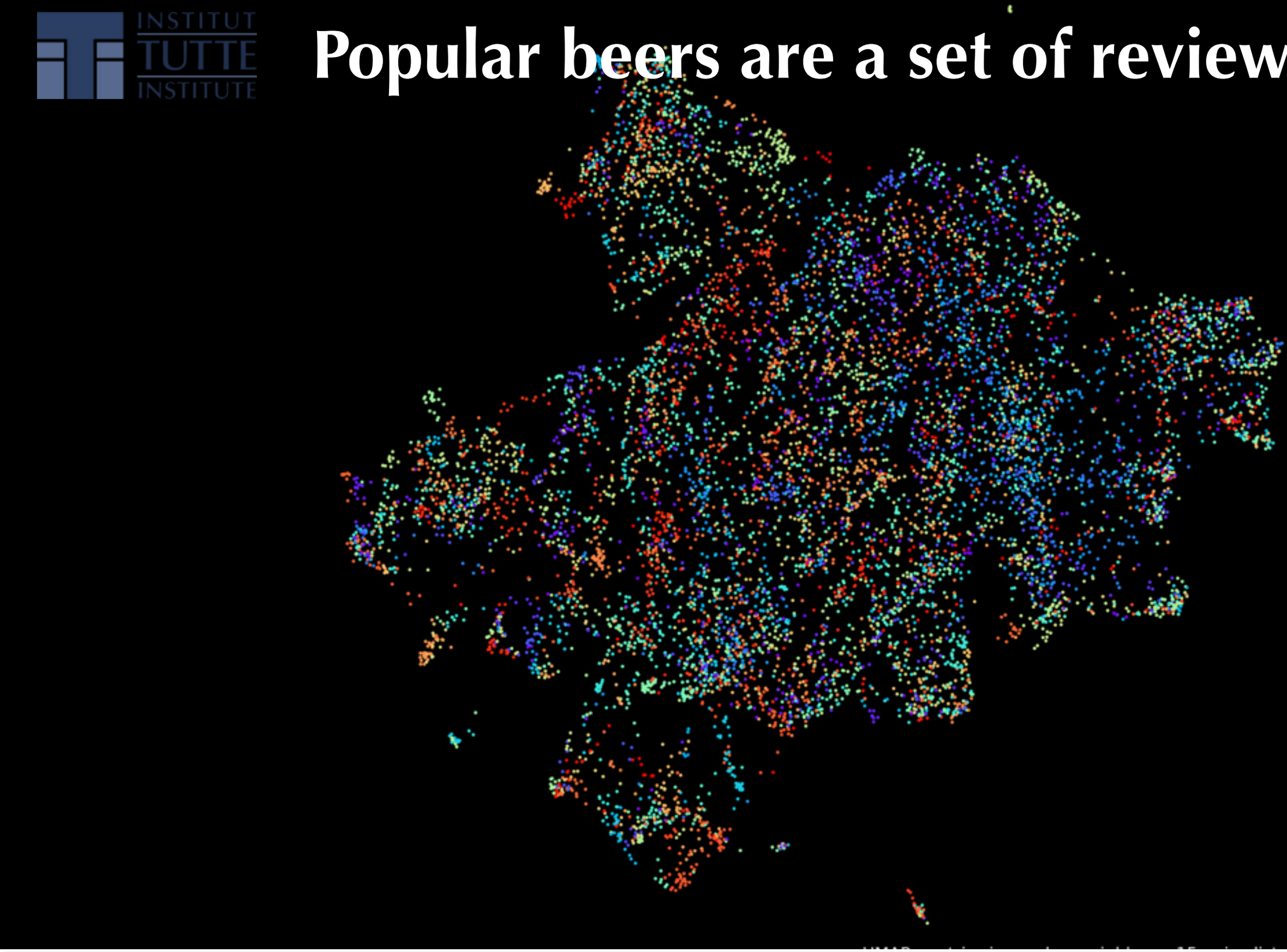
INSTITUT TUTTE INSTITUTE

Beers are a set of reviewers



```
beer_by_authors_model = umap.UMAP(n_neighbors=15, n_components=2, metric='jaccard',
                                  unique=True, random_state=42).fit(beer_by_authors)
```

Popular beers are a set of reviewers



```
beer_by_authors_model = umap.UMAP(n_neighbors=15, n_components=2, metric='jaccard',
                                  unique=True, random_state=42).fit(beer_by_authors)
```

INSTITUT TUTTE INSTITUTE

Breweries are a set of **reviewers**

Distance = Jaccard

$$J(A,B) = 1 - \frac{A \cap B}{A \cup B}$$

Breweries are sets of reviewers

New Zealand and Australian breweries

Central American breweries

Canadian breweries

Midwest breweries

European breweries

Colorado breweries

Texas breweries

California breweries

Washington and Oregon breweries

New England breweries

South-east breweries

New York breweries

UMAP: metric=jaccard, n_neighbors=15, min_dist=0.1

INSTITUT TUTTE INSTITUTE

What if counts matter?

INSTITUT TUTTE INSTITUTE

Variable width categorical data

When counts matter

Part I: Representation

INSTITUT TUTTE INSTITUTE

A **brewery** is a bag of the **beer styles it makes**

INSTITUT TUTTE INSTITUTE

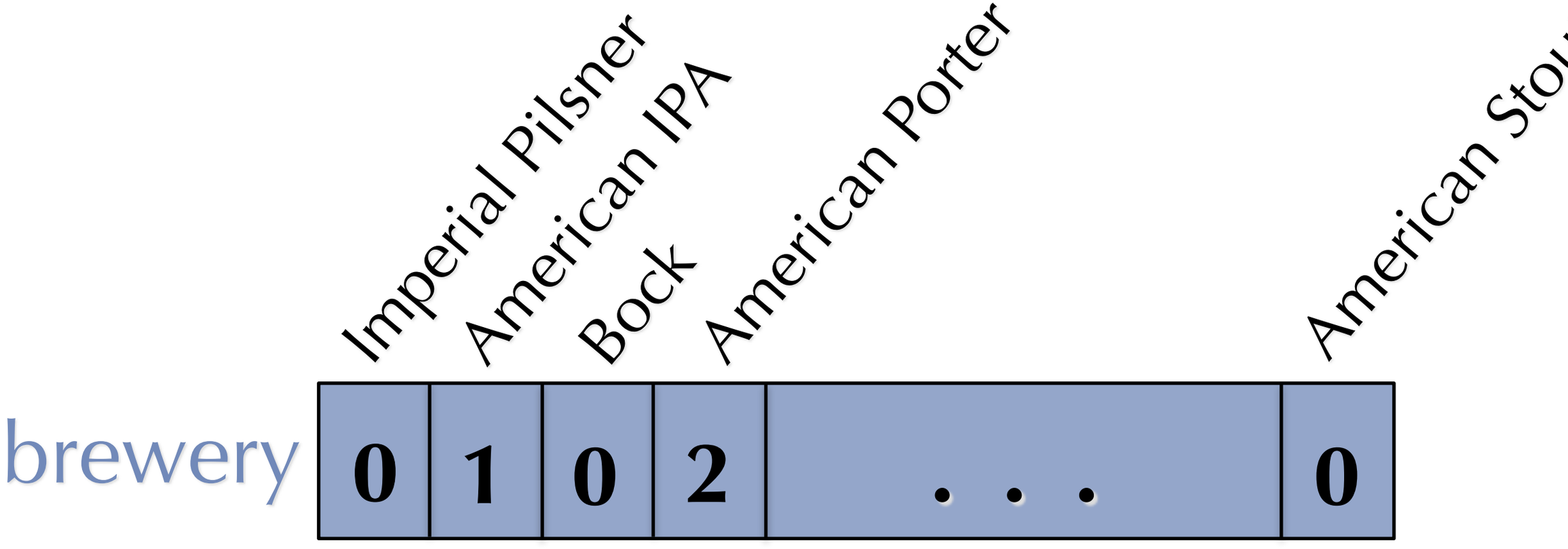
What if many breweries have made most kinds of beer?

INSTITUT TUTTE INSTITUTE

CountVectorize

A brewery is a bag of beer styles reviews

American IPA,American Porter,American Porter,Witbier,Witbier,American Brown Ale,...

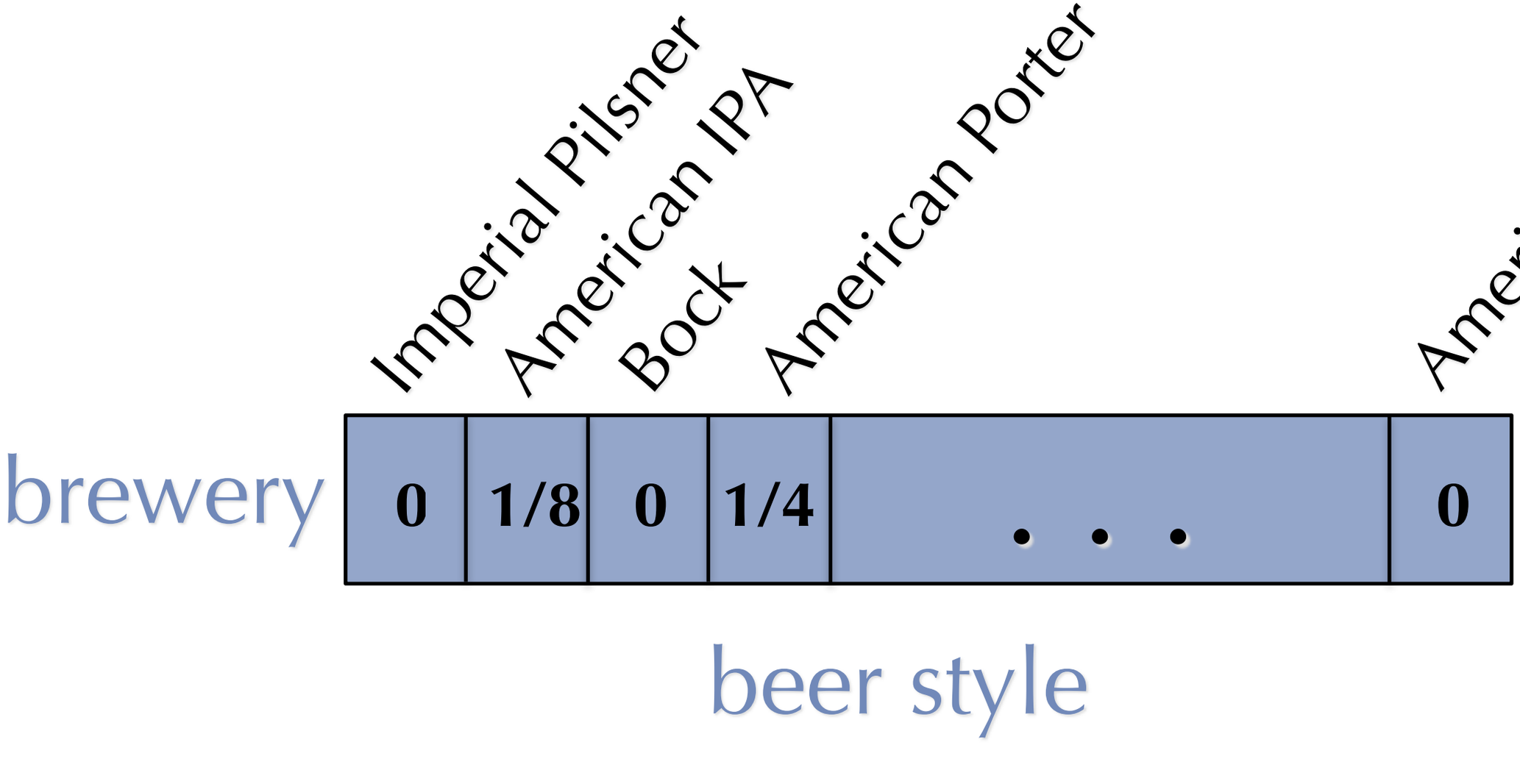


INSTITUT TUTTE INSTITUTE

CountVectorize

A brewery is a bag of ~~beer styles reviews~~ probabilities

American IPA,American Porter,American Porter,Witbier,Witbier,American Brown Ale,...

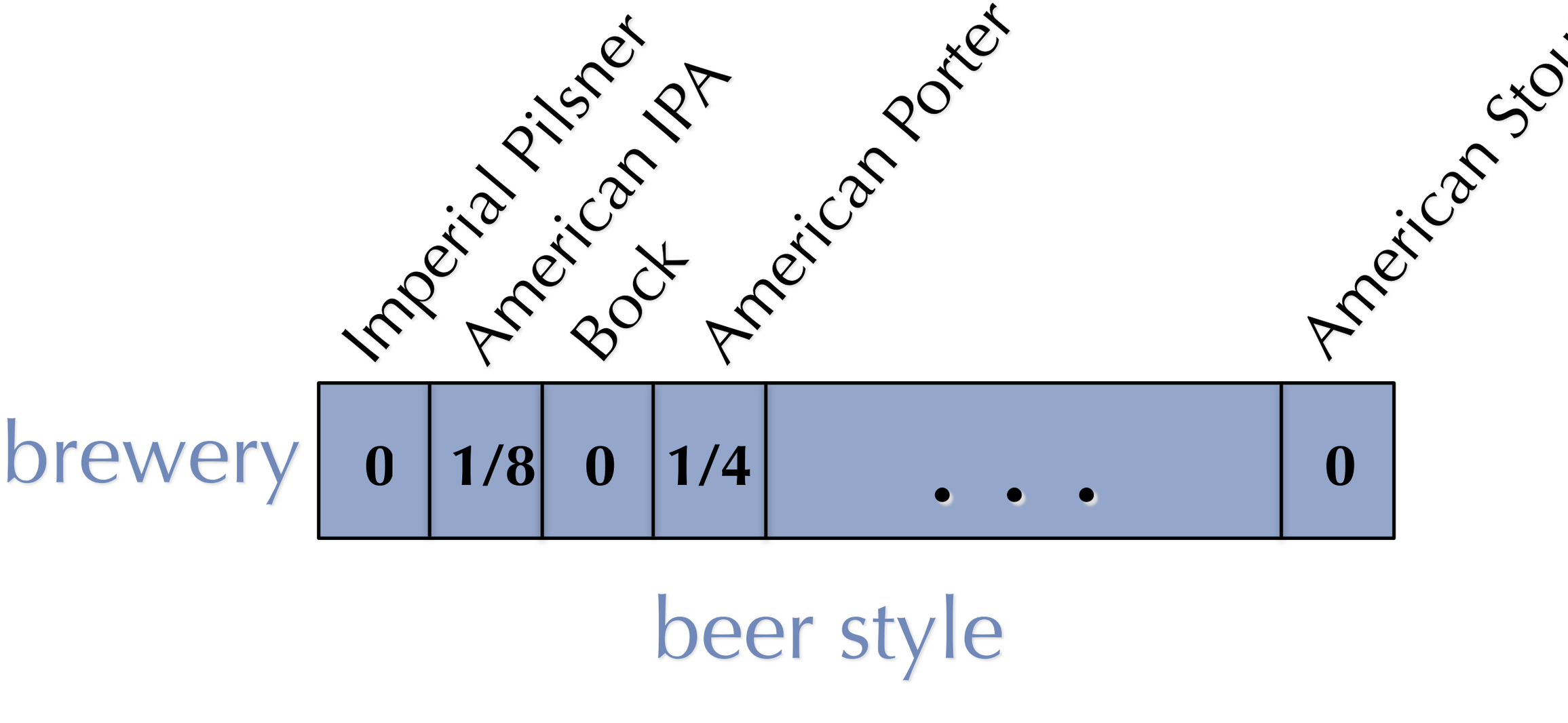


INSTITUT TUTTE INSTITUTE

CountVectorize

A brewery is a multinomial distribution across our beer style space

American IPA, American Porter, American Porter, Witbier, Witbier, American Brown Ale, ...



INSTITUT TUTTE INSTITUTE

Imperial Pilsner
American IPA
Bock
American Porter
American Stout

breweries

				⋮	
0	1/8	0	1/4	. . .	0
1/3	0	0	1/6	. . .	0
				⋮	

beer style

INSTITUT TUTTE INSTITUTE

Variable width categorical data

When counts matter

Part II: Distance

INSTITUT TUTTE INSTITUTE

Theoretical statistics to the rescue!

Carter, Raich, Finn, Hero, 2009
Amari, 2012

Fisher Information Theory

$$(\Delta_{n-1}, f) \to (S^n, \mu)$$

$$x_i \mapsto \sqrt{x_i}$$

$d_a(w_a, w_b) = arccos\left(\frac{\sum_{i=1}^{n}\sqrt{a_i b_i}}{\sqrt{|a|_1|b|_1}}\right)$ approximated with $arccos(\theta) \approx \sqrt{1-\theta}$,

$$d_H(w_a, w_b) = \sqrt{1 - \frac{\sum\sqrt{a_i b_i}}{\sqrt{|a|_1|b|_1}}}$$

Carter, Raich, Finn, Hero, 2009
Amari, 2012

Fisher Information Distance

INSTITUT TUTTE INSTITUTE

Hellinger Distance!

INSTITUT TUTTE INSTITUTE

A **brewery** is a bag of the **beer styles it makes**

Distance = Hellinger

$$d_H(w_a, w_b) = \sqrt{1 - \frac{\sum \sqrt{a_i b_i}}{\sqrt{|a|_1 |b|_1}}}$$

INSTITUT TUTTE INSTITUTE

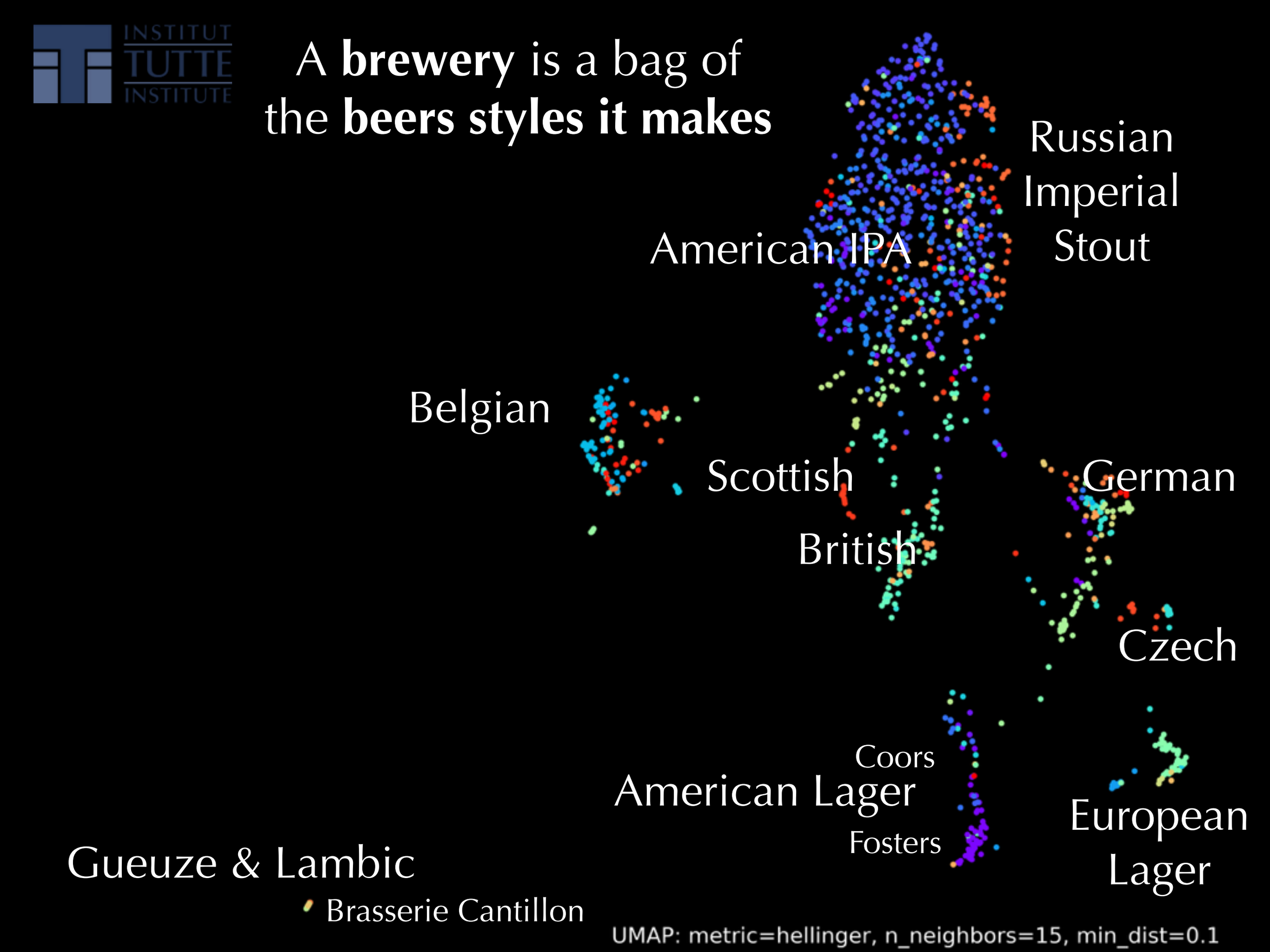
Variable width categorical data

When counts matter

Part III: Look at your data

INSTITUT TUTTE INSTITUTE

A **brewery** is a bag of
the **beers styles it makes**



INSTITUT TUTTE INSTITUTE

Where else have I heard of a bag of counts before…?

Documents!

INSTITUT TUTTE INSTITUTE

Clustering
Grouping

Outlier Detection
Anomaly Detection

Unsupervised Learning

... on documents

Exploratory Data Analysis
Visualization

INSTITUT TUTTE INSTITUTE

Documents are

Variable width categorical data

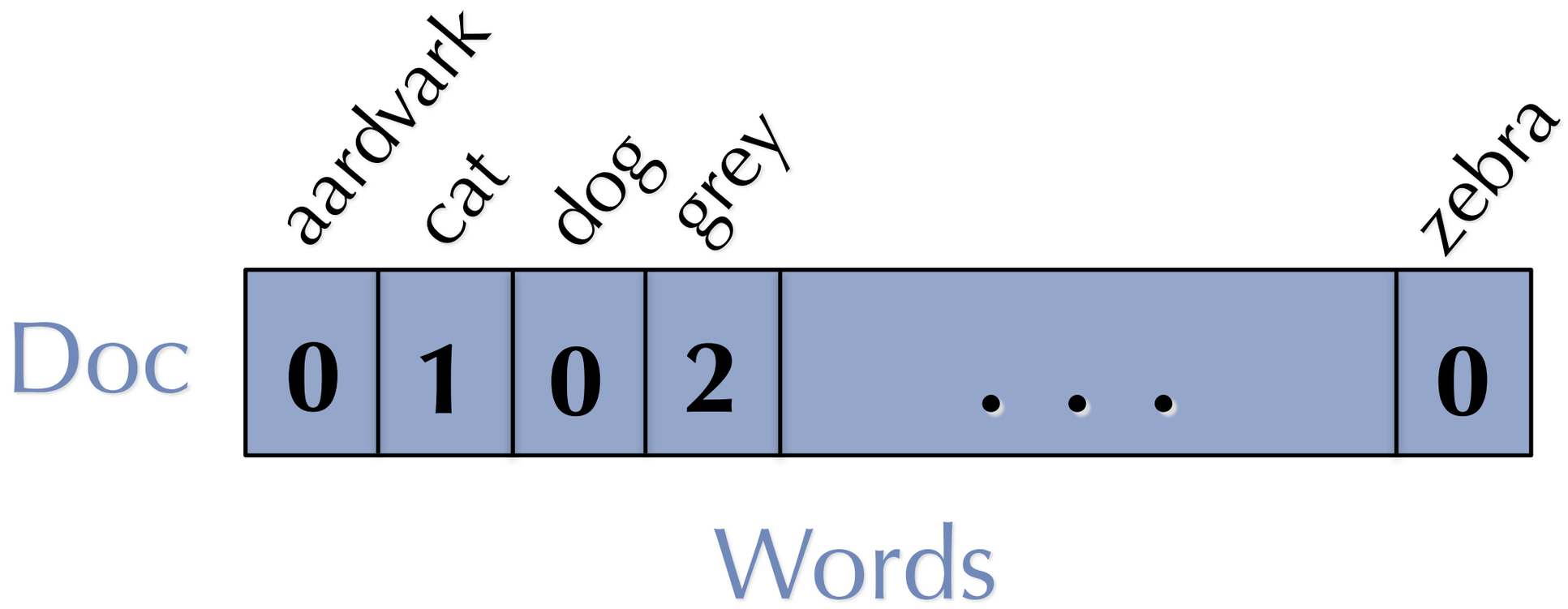
When counts matter

Part I: Representation

CountVectorize

A document is a bag of words

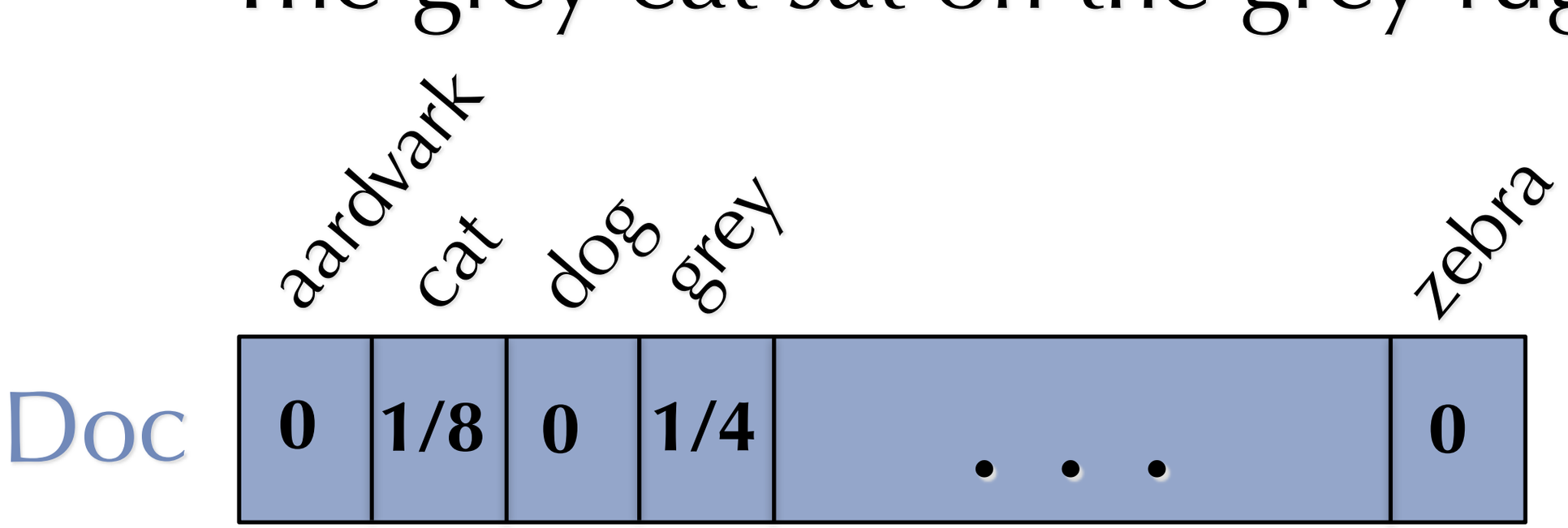
"The grey cat sat on the grey rug"



CountVectorize

A document is a bag of ~~words~~ probabilities

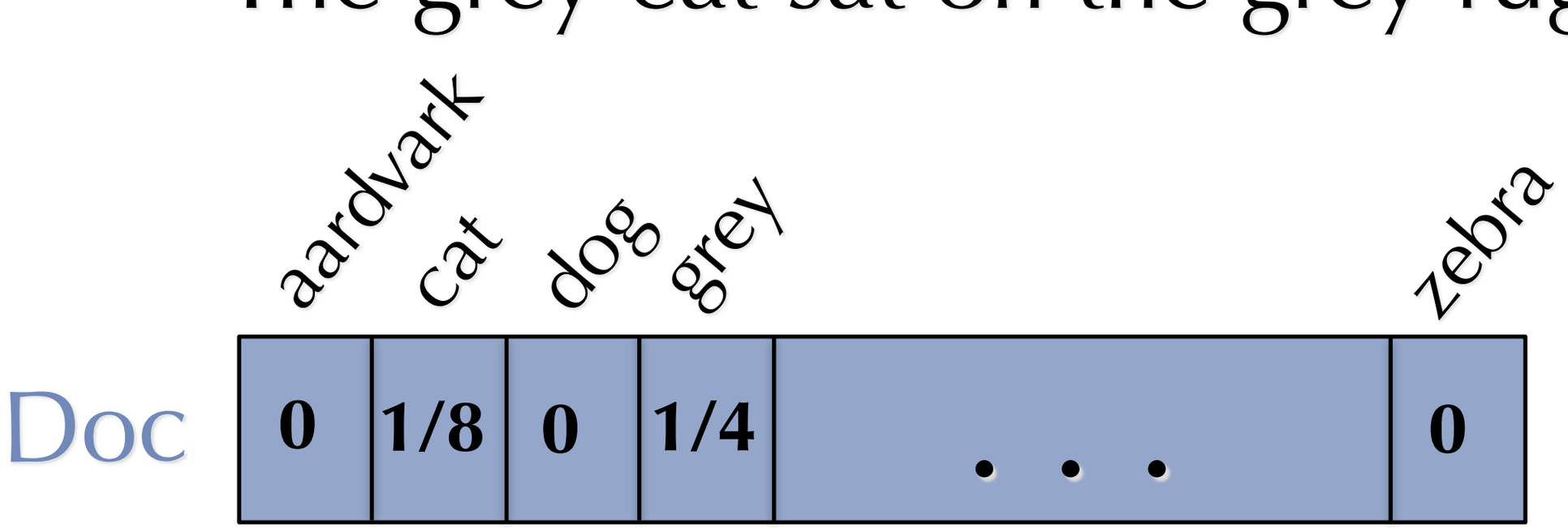
"The grey cat sat on the grey rug"



CountVectorize

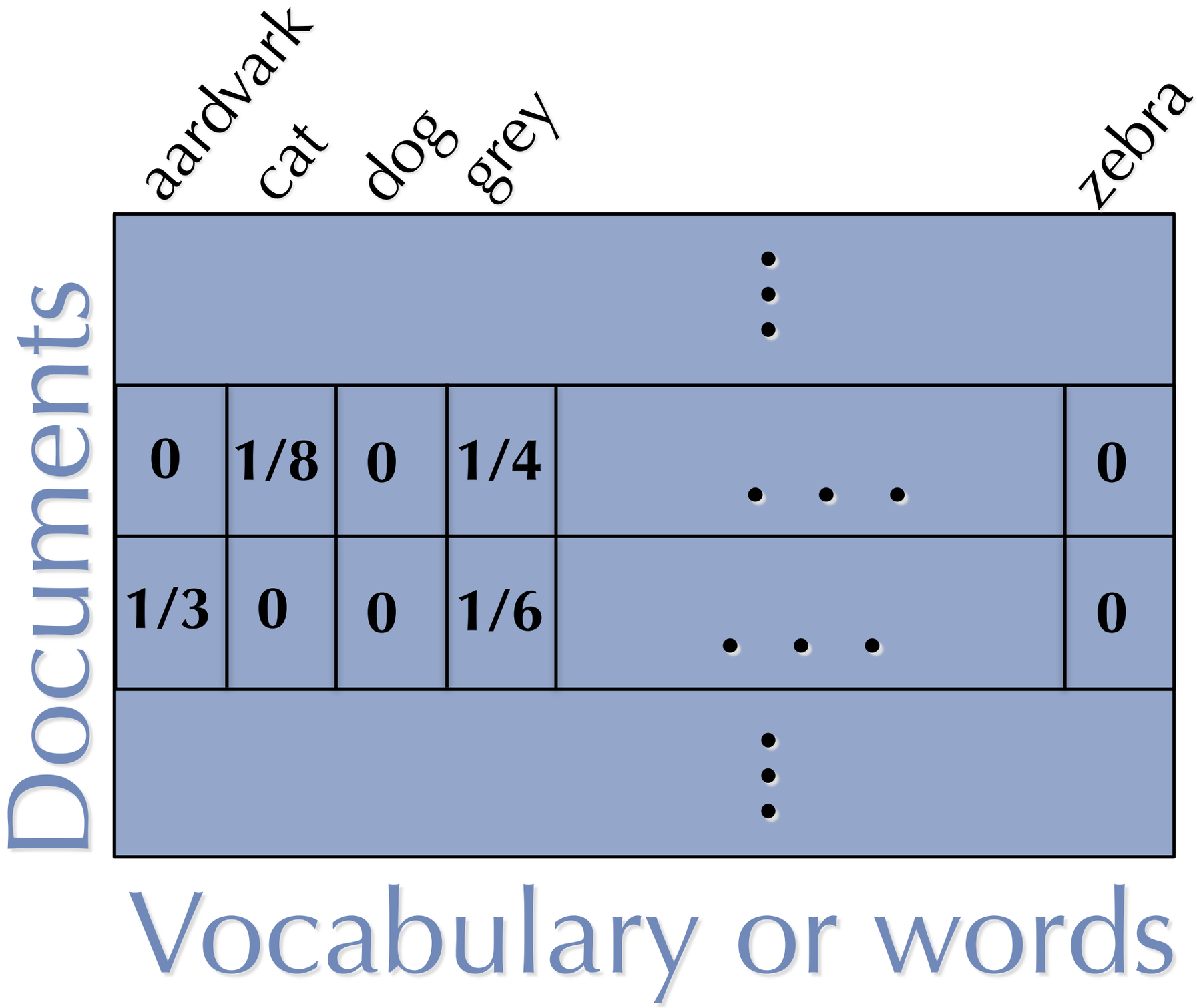
A document is a multinomial distribution across our vocabulary space

"The grey cat sat on the grey rug"



TfidfVectorize ~~CountVectorize~~

A Corpus is a document by word matrix*



INSTITUT TUTTE INSTITUTE

Documents are

Variable width categorical data

When counts matter

Part II: Distance

A **document** is a bag of **words**

Distance = Hellinger

$$d_H(w_a, w_b) = \sqrt{1 - \frac{\sum \sqrt{a_i b_i}}{\sqrt{|a|_1 |b|_1}}}$$

INSTITUT TUTTE INSTITUTE

Documents are

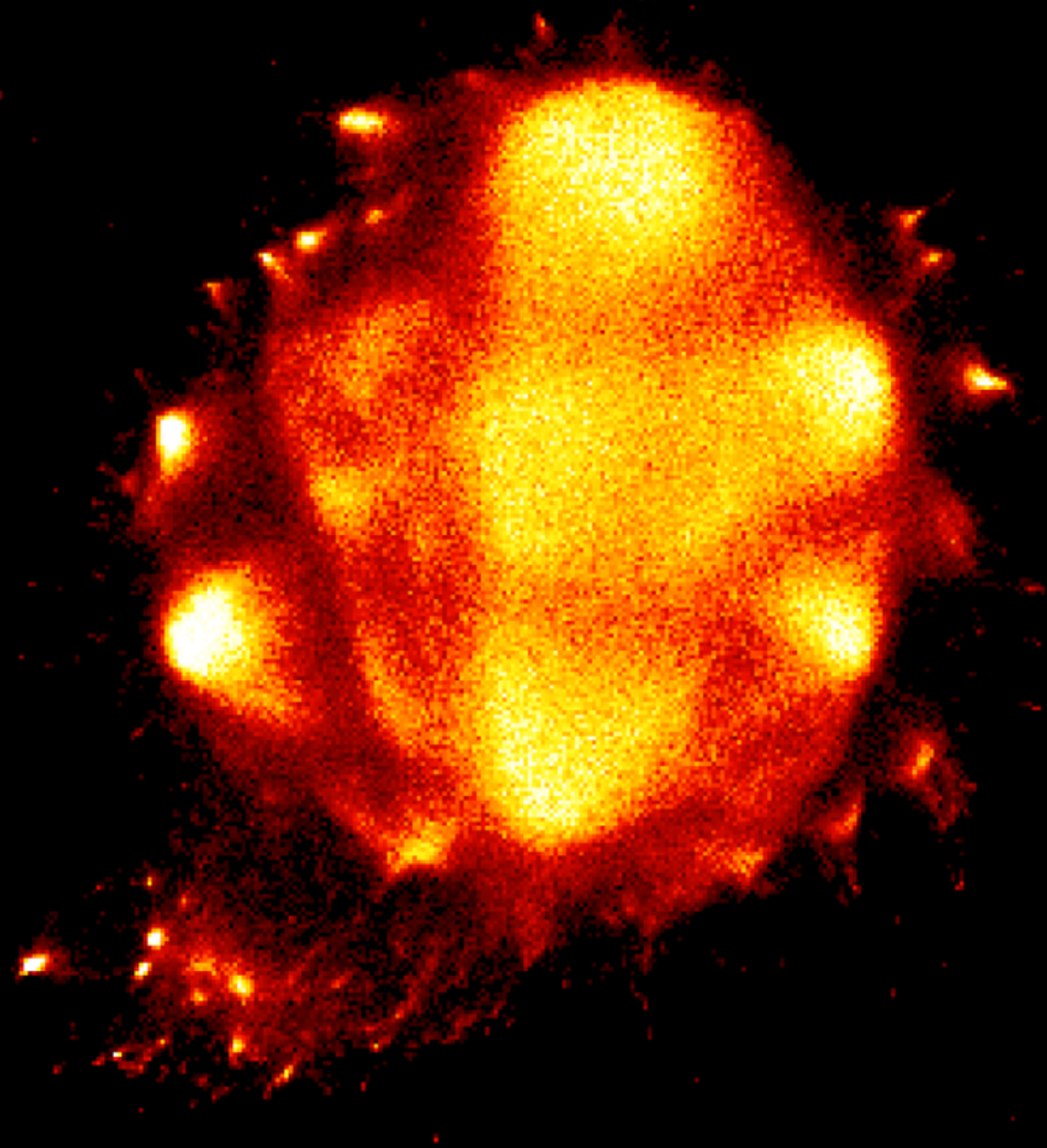
Variable width categorical data

When counts matter

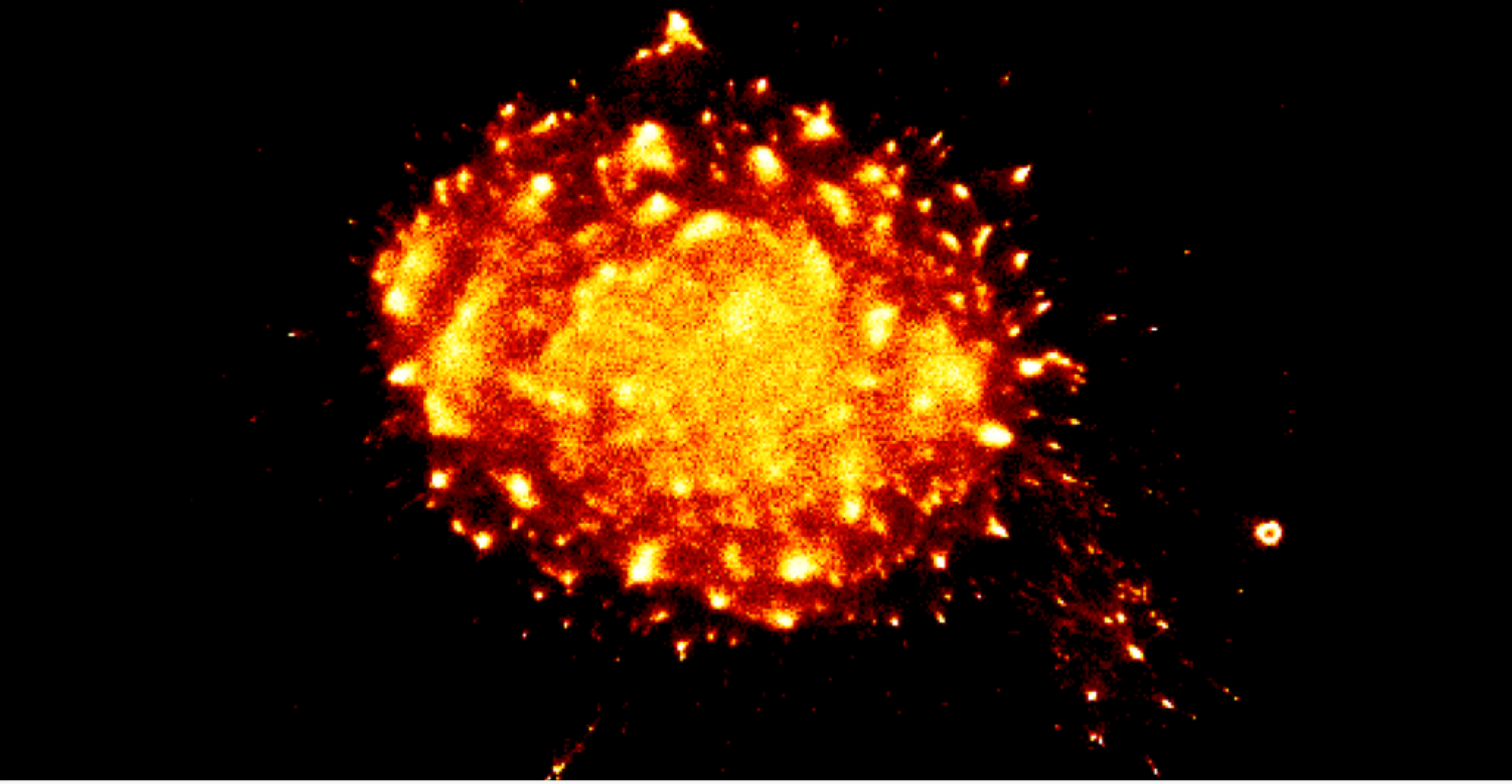
Part III: Look at your data

INSTITUT TUTTE INSTITUTE

Documents are a bag of words



Documents are a bag of bigrams



```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(ngram_range=(2,2))
representation = vectorizer.fit_transform(text_data)
```

That's what we've been calling

DocMAP

i.e. UMAP on documents

INSTITUT TUTTE INSTITUTE

But you said ALL the things!

Words are things

INSTITUT TUTTE INSTITUTE

There are other ways to embed words

Word2Vec

GloVe

FastText

INSTITUT TUTTE INSTITUTE

“You shall know a word by the company it keeps”

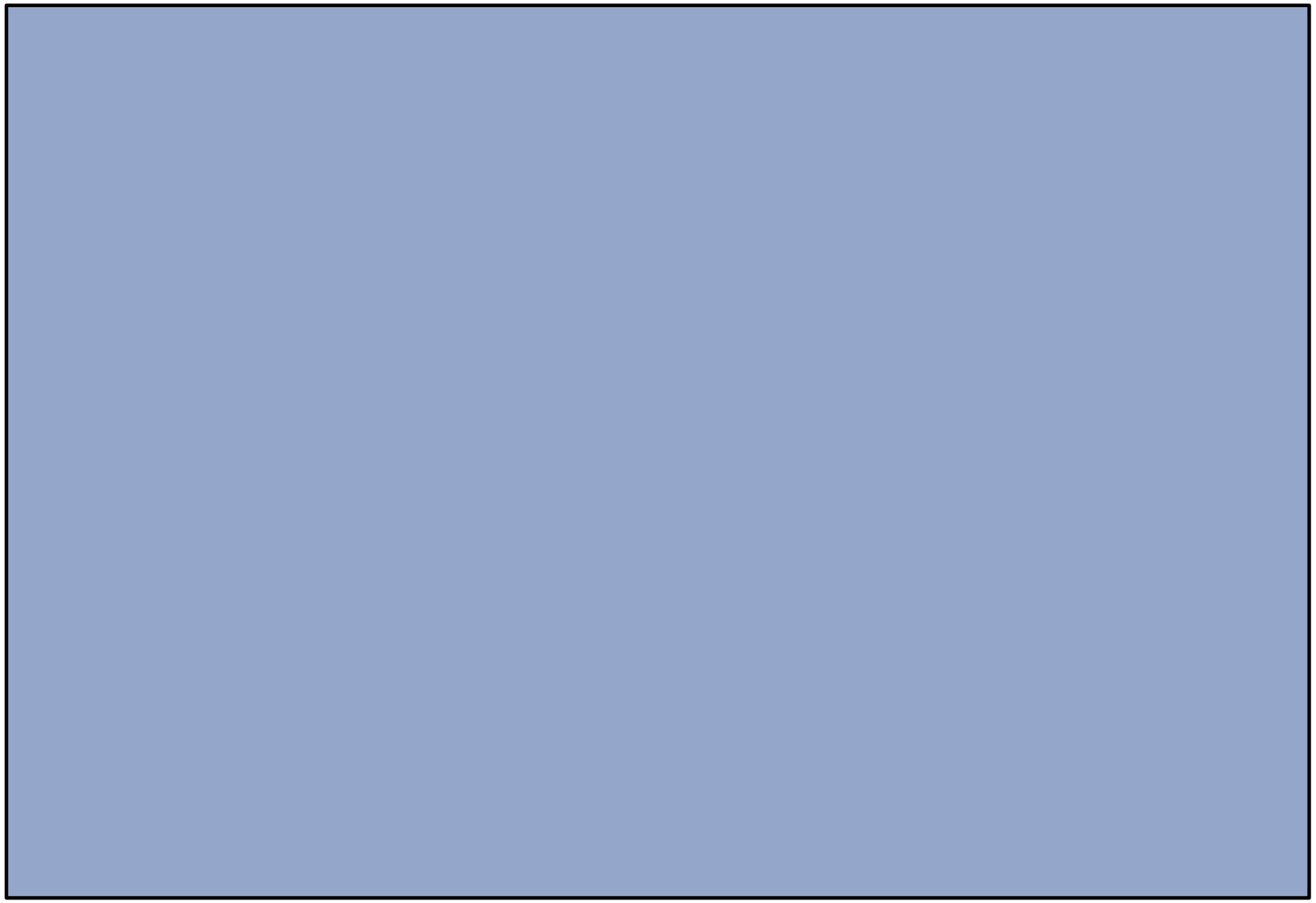
John Rupert Firth, 1957

(a famous linguist)

A word is a document of all ~~sentences~~ containing it ~~contexts~~ windows

Word usage can be represented by a document by word matrix

Word documents



Vocabulary or words

INSTITUT TUTTE INSTITUTE

Words are

Documents are

Variable width categorical data

When counts matter

Representation & Distance

INSTITUT TUTTE INSTITUTE

A **word** is a bag of **nearby words**

Distance = Hellinger

$$d_H(w_a, w_b) = \sqrt{1 - \frac{\sum \sqrt{a_i b_i}}{\sqrt{|a|_1 |b|_1}}}$$

INSTITUT TUTTE INSTITUTE

Words are

Documents are

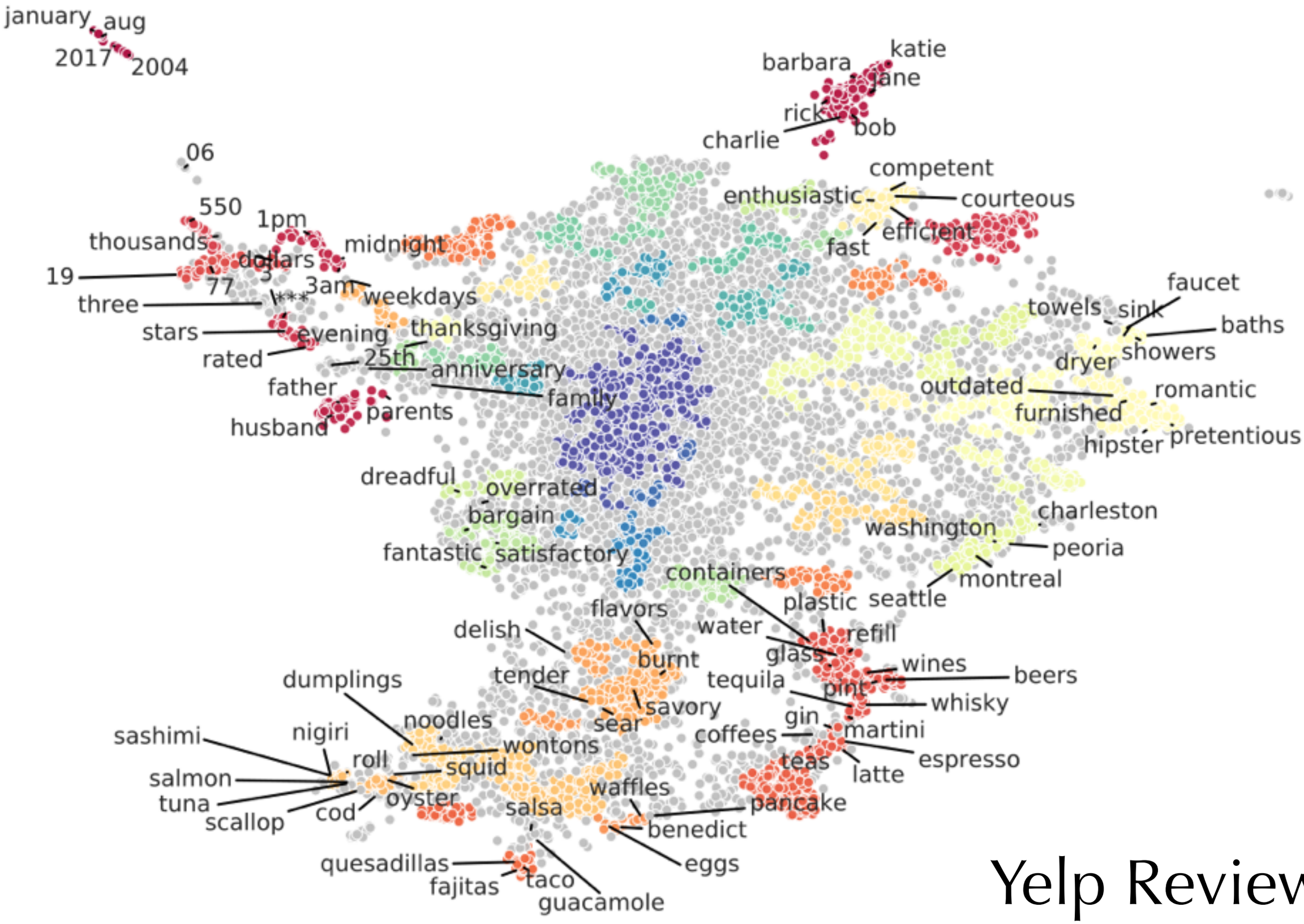
Variable width categorical data

When counts matter

Part III: Look at your data

INSTITUT TUTTE INSTITUTE

WordMAP



Yelp Reviews

That's what we've been calling

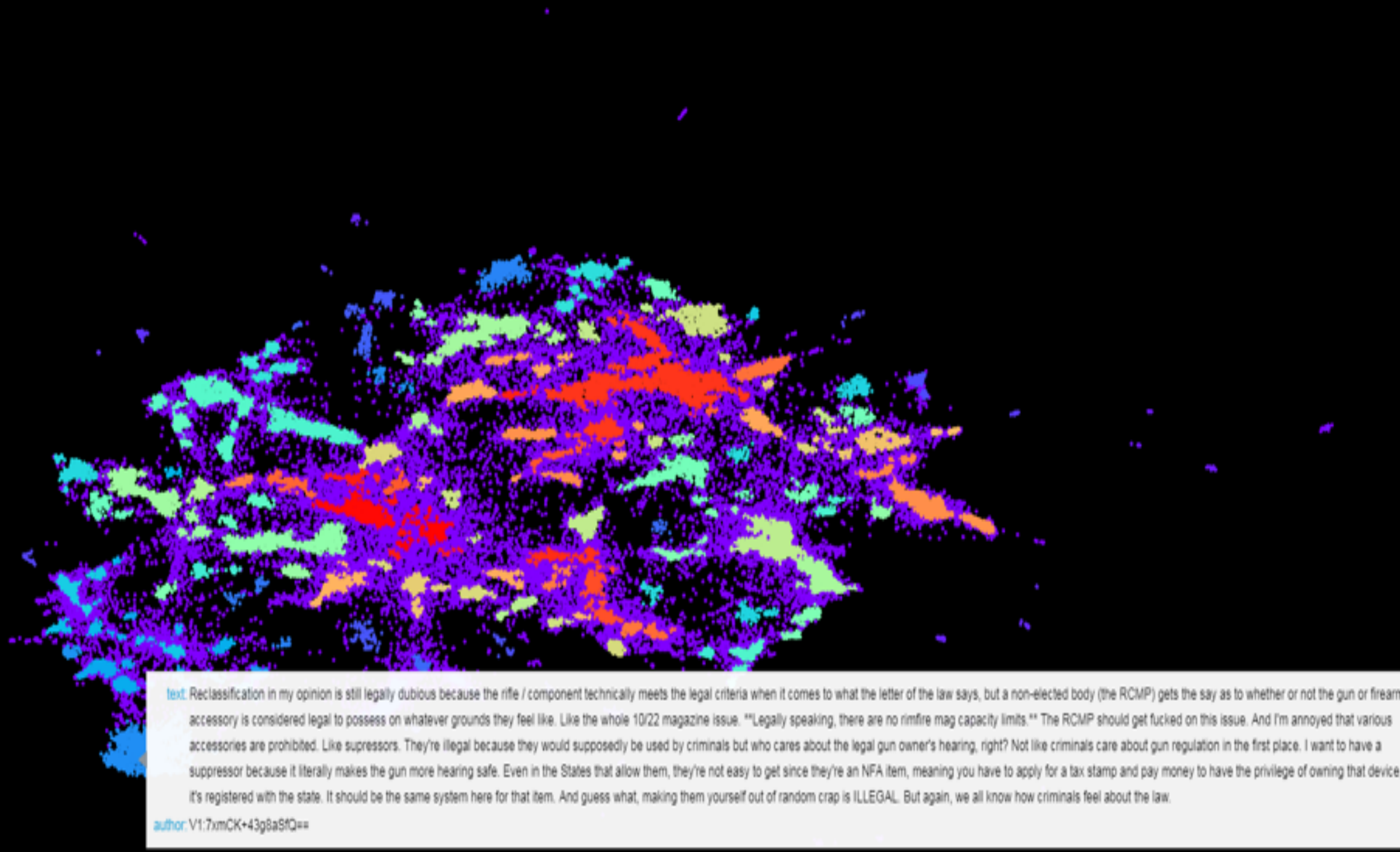
WordMAP

i.e. UMAP on words

INSTITUT TUTTE INSTITUTE

But we aren't limited to beer and language

Documents are a bag of topics

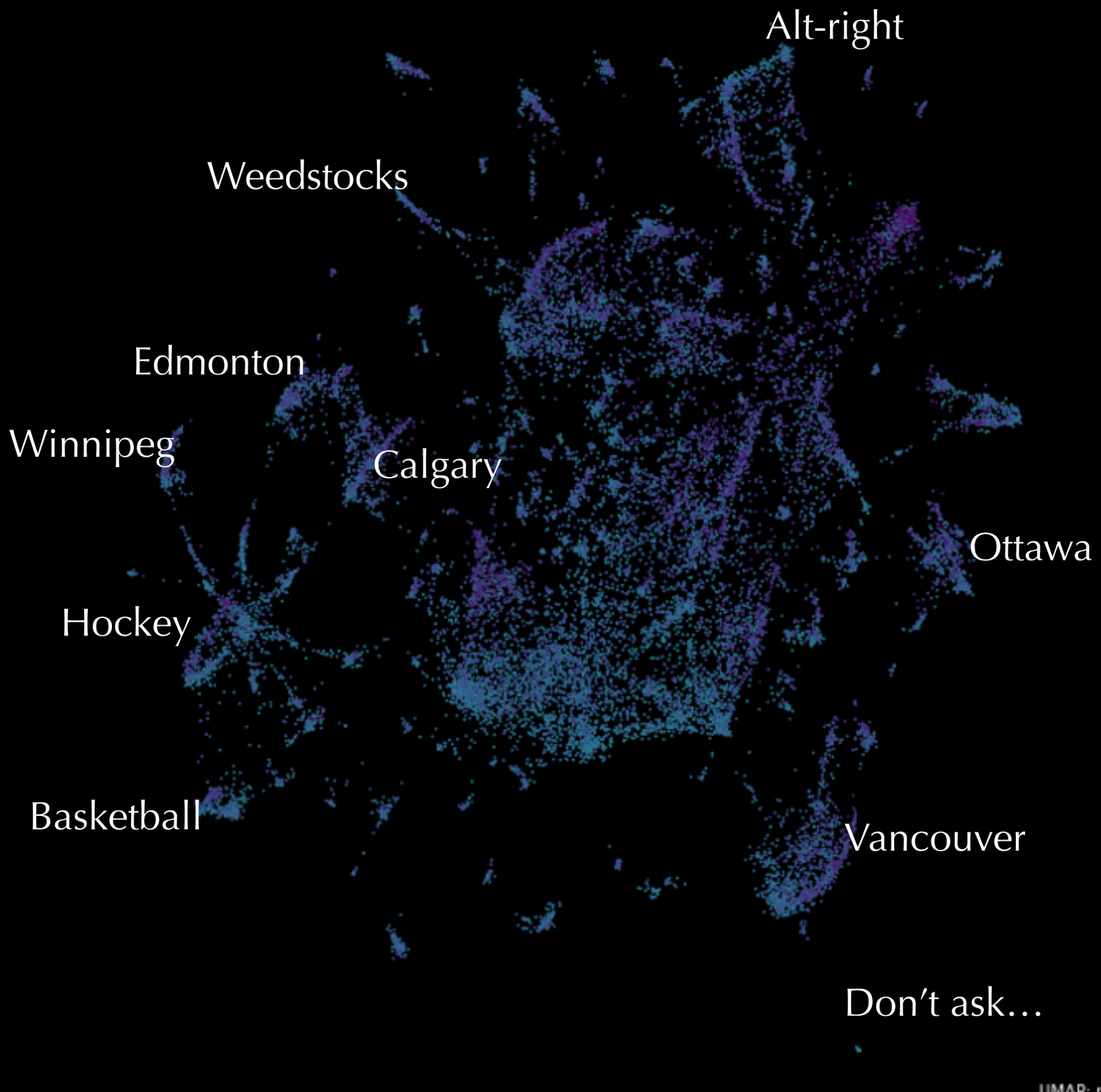


text: Reclassification in my opinion is still legally dubious because the rifle / component technically meets the legal criteria when it comes to what the letter of the law says, but a non-elected body (the RCMP) gets the say as to whether or not the gun or firearm accessory is considered legal to possess on whatever grounds they feel like. Like the whole 10/22 magazine issue. **Legally speaking, there are no rimfire mag capacity limits.** The RCMP should get fucked on this issue. And I'm annoyed that various accessories are prohibited. Like supressors. They're illegal because they would supposedly be used by criminals but who cares about the legal gun owner's hearing, right? Not like criminals care about gun regulation in the first place. I want to have a suppressor because it literally makes the gun more hearing safe. Even in the States that allow them, they're not easy to get since they're an NFA item, meaning you have to apply for a tax stamp and pay money to have the privilege of owning that device it's registered with the state. It should be the same system here for that item. And guess what, making them yourself out of random crap is ILLEGAL. But again, we all know how criminals feel about the law.

author: V1:7xmCK+43g8aSfQ==

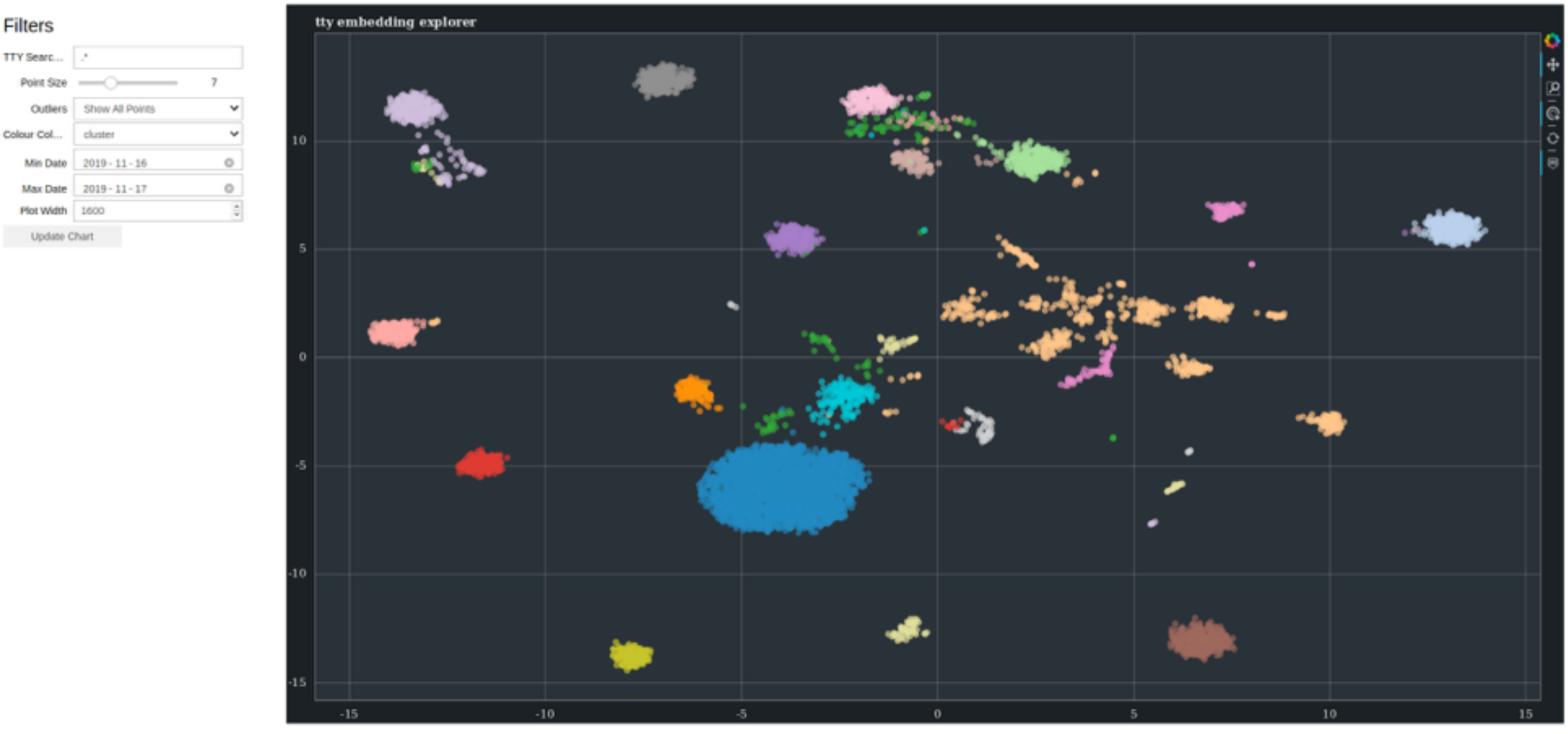
Malware is a bag of commands

Accounts are bags of forums



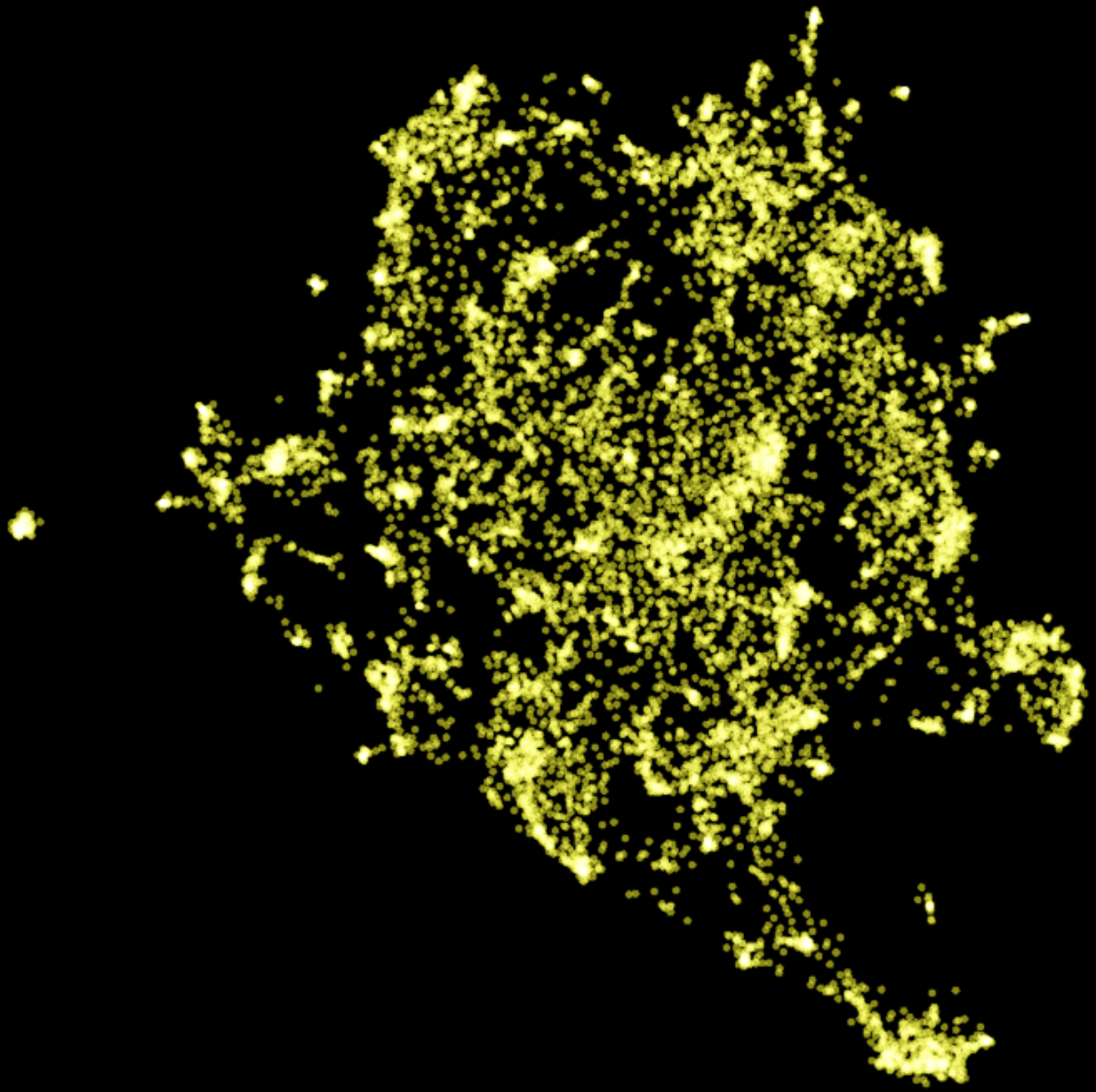
INSTITUT TUTTE INSTITUTE

Honeypots: Sessions are bags of bigrams over commands run



INSTITUT TUTTE INSTITUTE

Subreddits are sets of users



A

insert your thing here

is a

{set, bag, vector}

of

insert your other thing here

Questions?

https://github.com/lmcinnes/umap
pip install umap-learn

as of 2019 for interactive vis try the experimental branch:
pip install datashader
pip install holoviews
pip install git+https://github.com/lmcinnes/umap@0.4dev

Reproducibility Matters

Experiments and notebooks:
https://github.com/jc-healy/EmbedAllTheThings