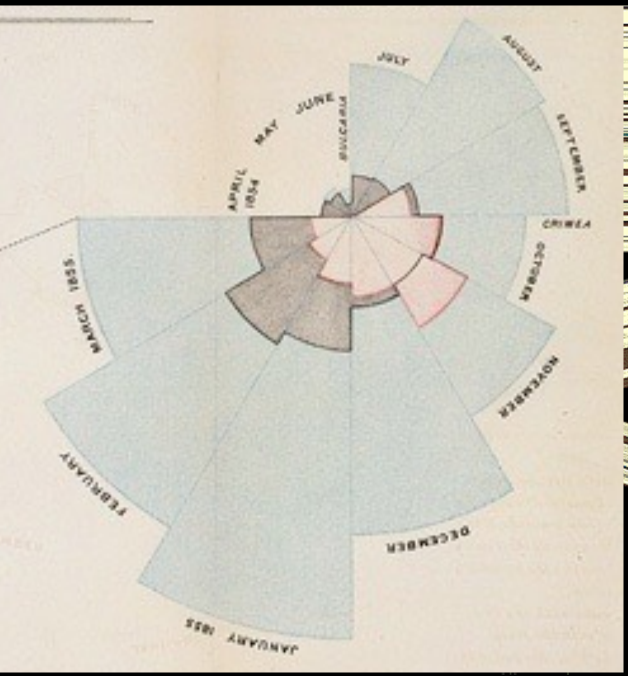
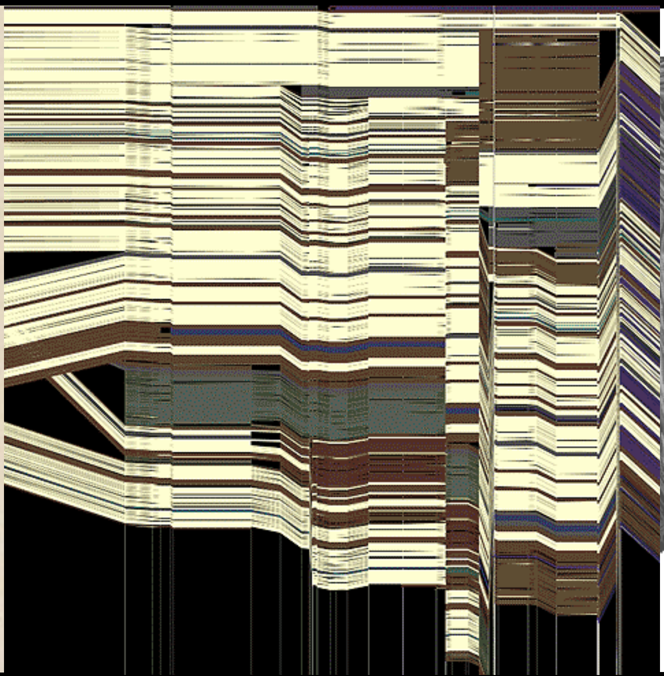
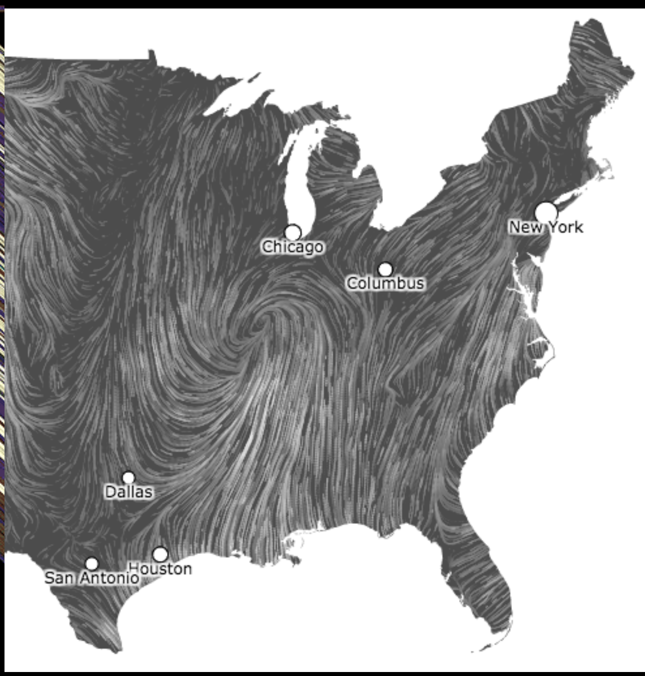
CSE 512 - Data Visualization

Model Interpretability







Jeffrey Heer University of Washington

What do we mean by "interpretable"?

Varied notions of "interpretable"

Causal (why): the degree to which a person can understand the cause of a result.

Predictive (what): the degree to which a person can predict the model's result.

By whom? For what purpose?

Why “interpretable” models?

Why “interpretable” models?

Fairness: assess for bias / discrimination

Privacy: protect sensitive information

Reliability: sensitivity to input changes

Causality: explanatory, not just predictive

Trust: make informed deployment choices

Approaches to Interpretability

“Inherently” interpretable models (?)

- Decision trees, decision lists
- Linear models
- Generalized additive models (GAMs)

Approaches to Interpretability

“Inherently” interpretable models (?)

- Decision trees, decision lists
- Linear models
- Generalized additive models (GAMs)

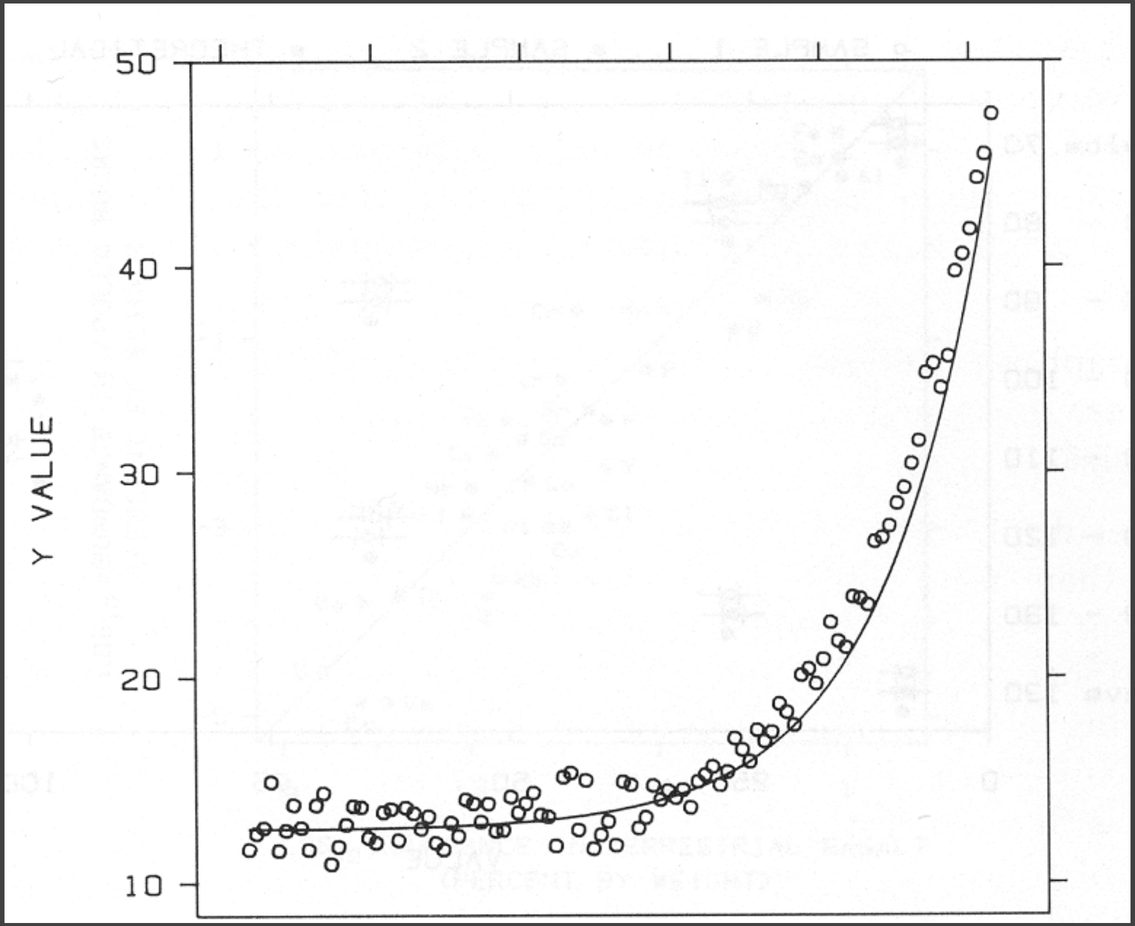
Inspection / analysis of existing models

- Visualize model features, activations
- “Model-agnostic” analysis of behavior

Model Assessment

Transforming Data

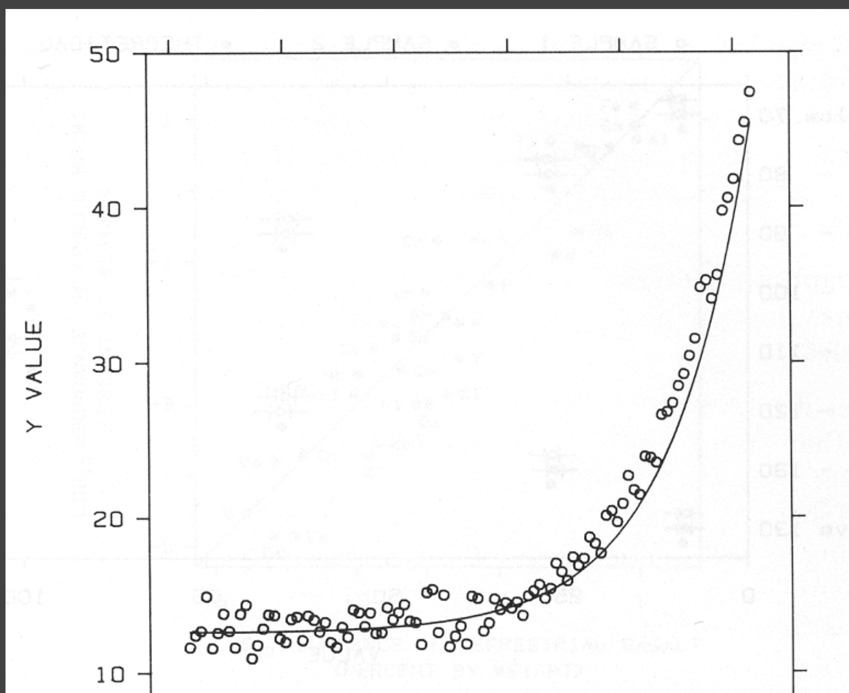
How well does the curve fit the data?

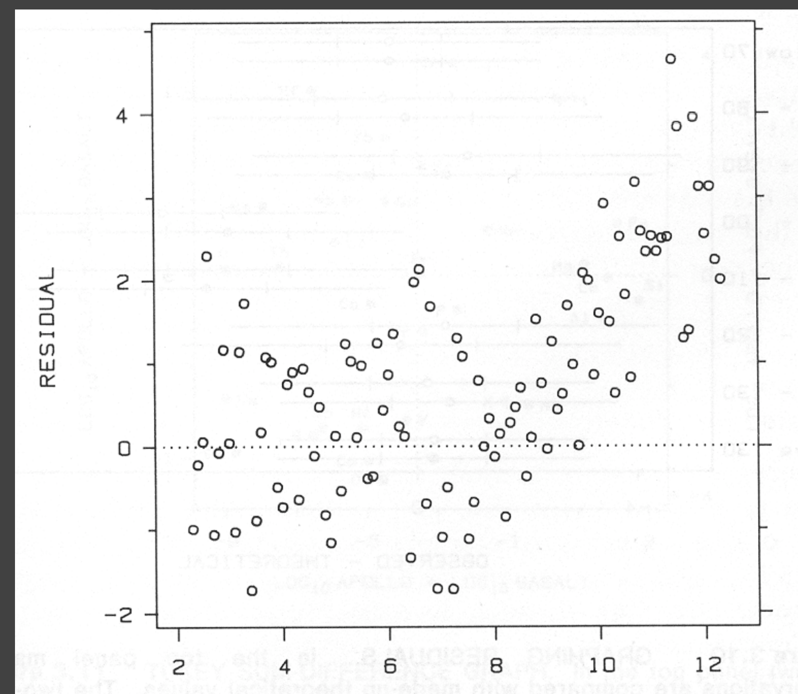


[Cleveland 85]

Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows goodness of fit

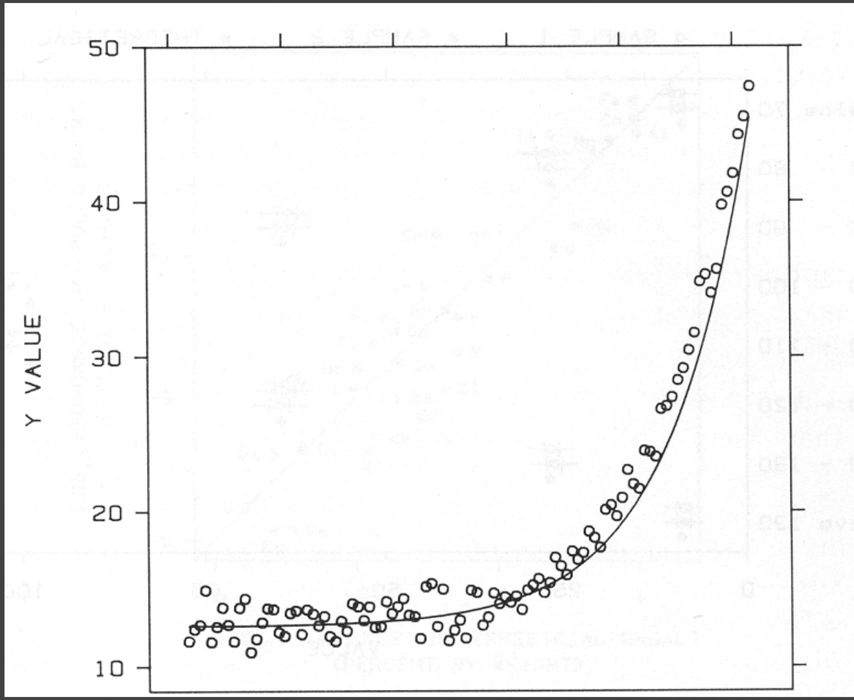


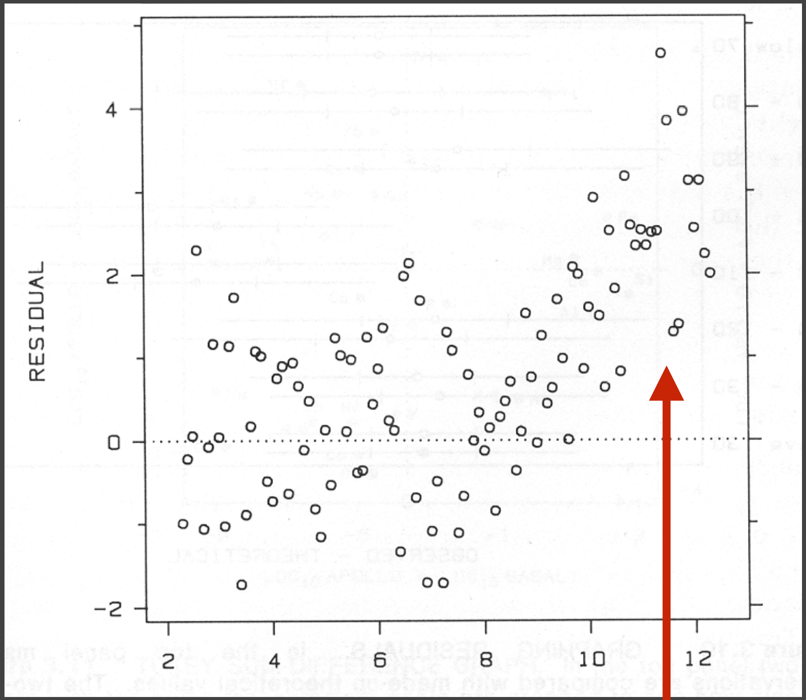


[Cleveland 85]

Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows goodness of fit



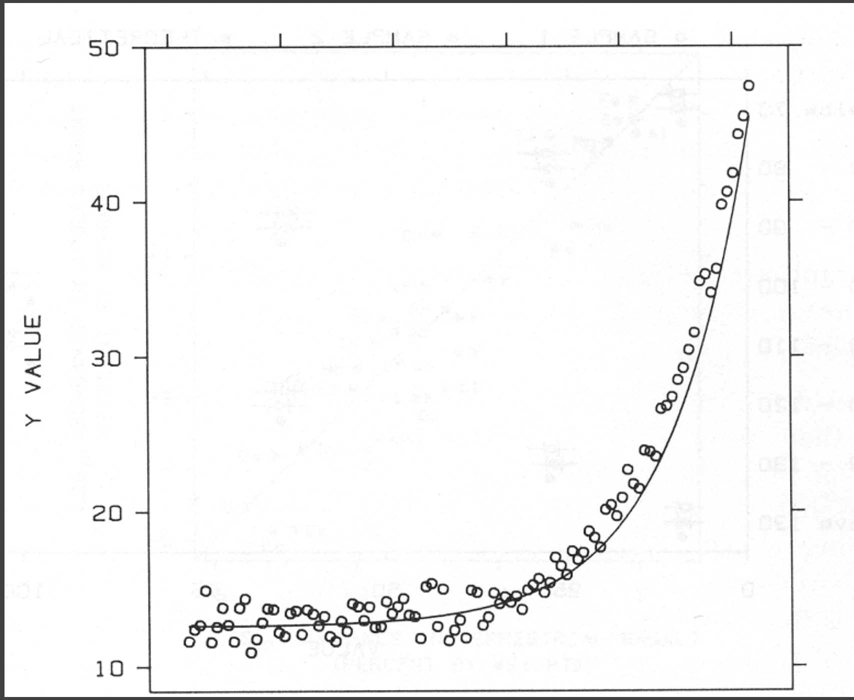


[Cleveland 85]

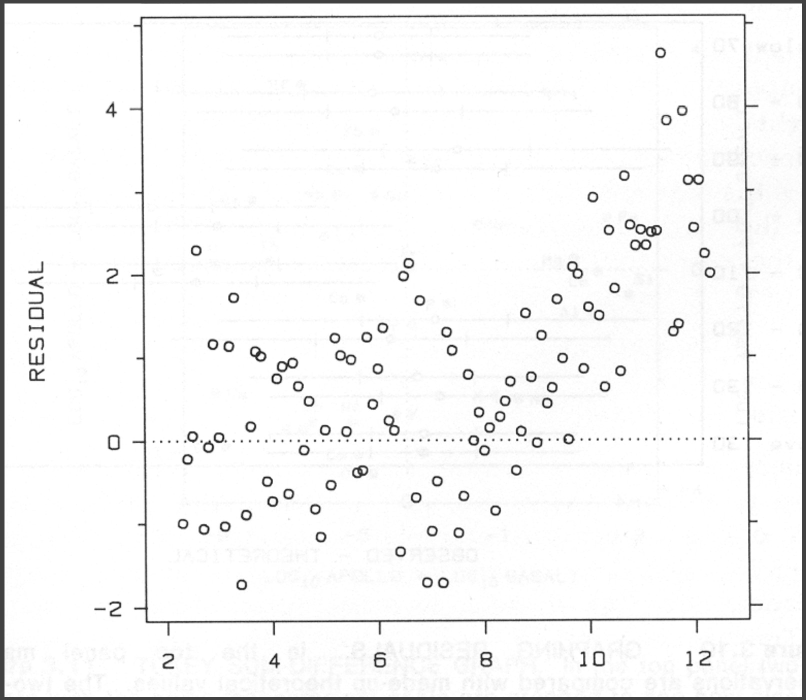
Heteroscedasticity!

Multiple Plotting Options

Plot model in data space

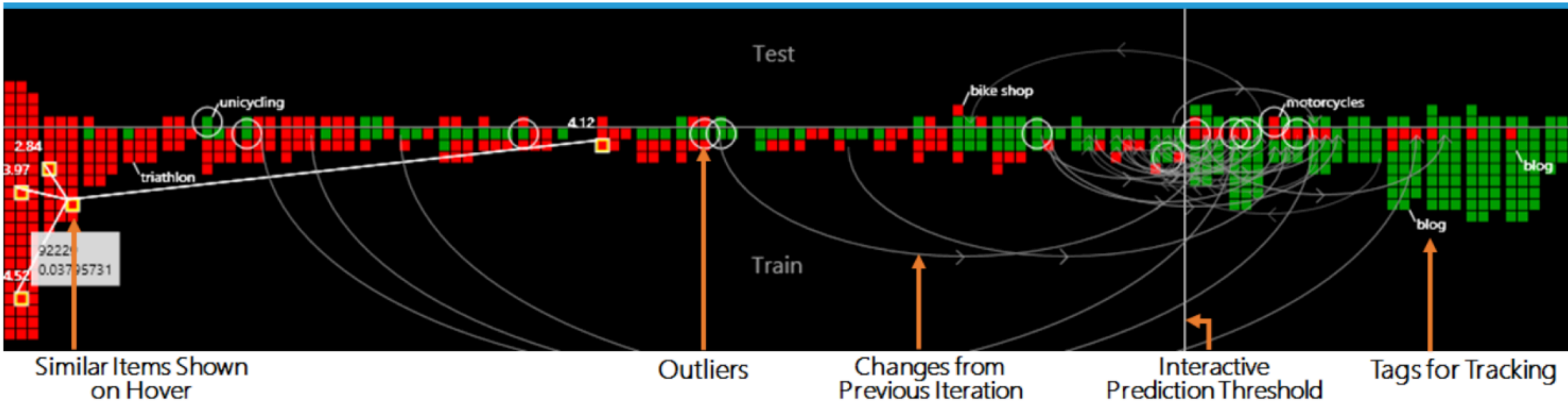


Plot data in model space

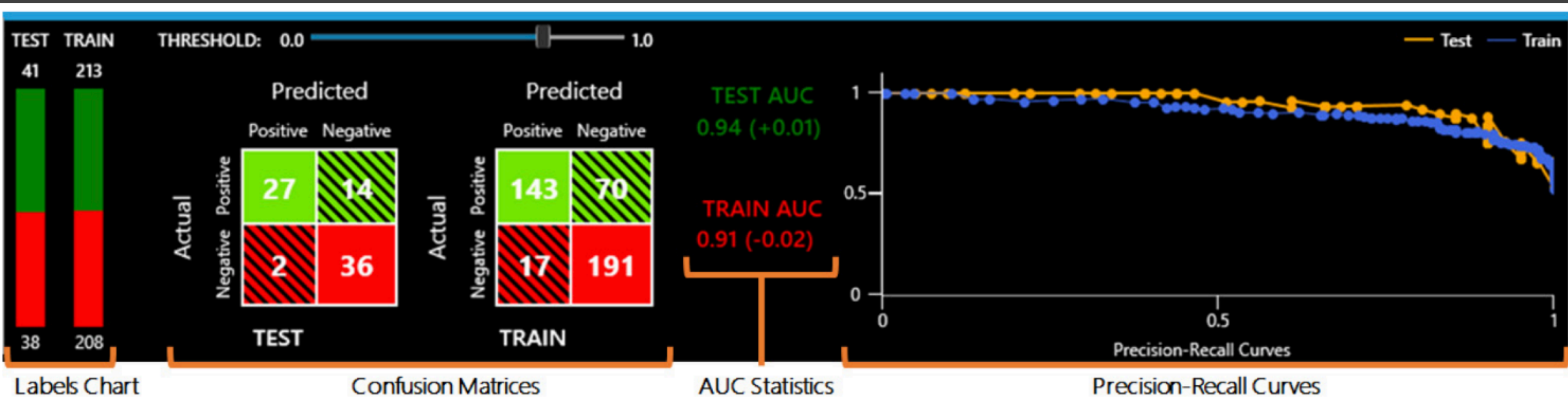


[Cleveland 85]

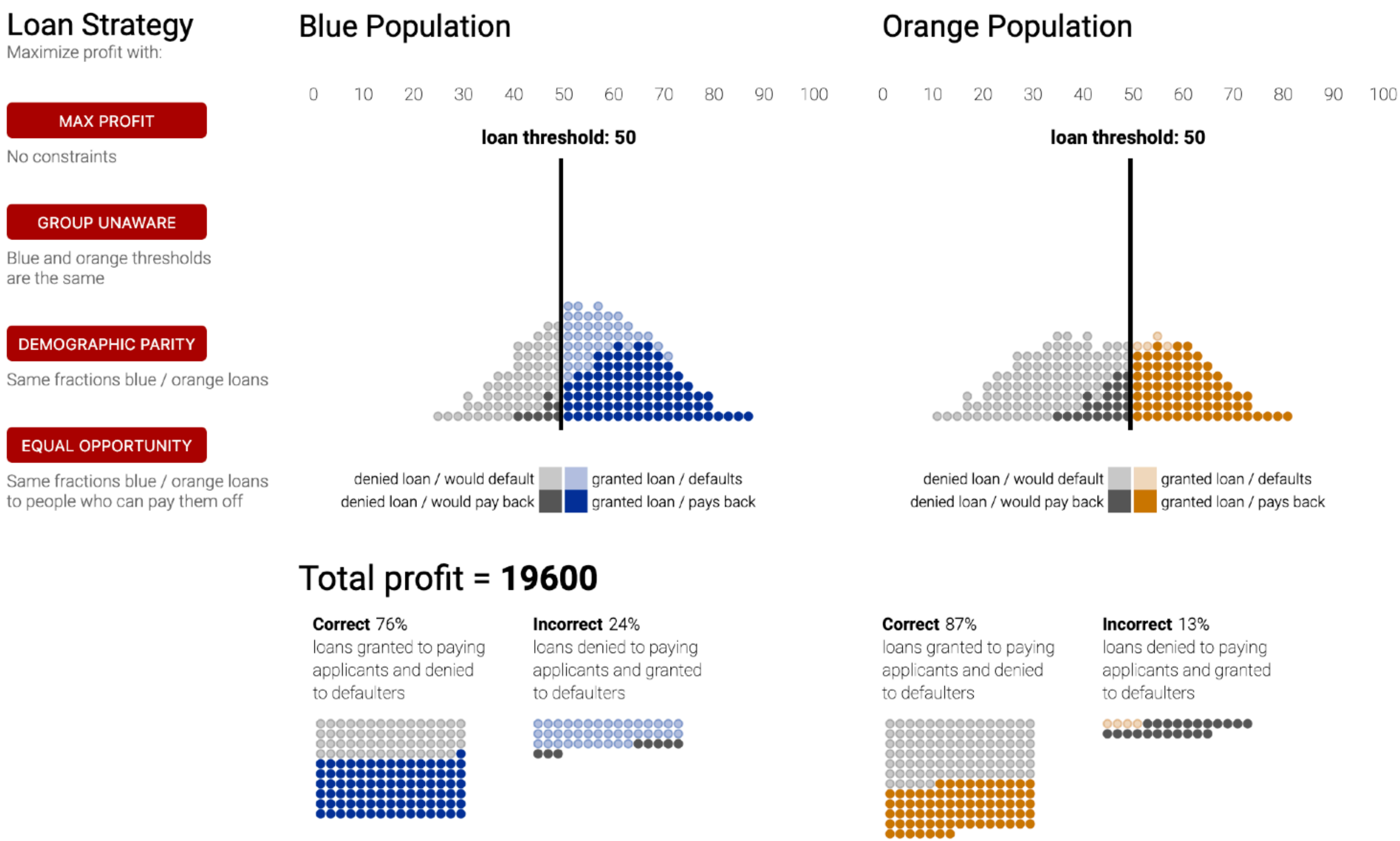
Model Tracker [Amershi et al. 2015]



- Model Prediction Score +



Assessing Fairness [Wattenberg et al. 2016]



Dimensionality Reduction

Dimensionality Reduction

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Dimensionality Reduction

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

DR methods are used to aid interpretation, but are also **subject to their own interpretation issues!**

Dimensionality Reduction

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

DR methods are used to aid interpretation, but are also **subject to their own interpretation issues!**

Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

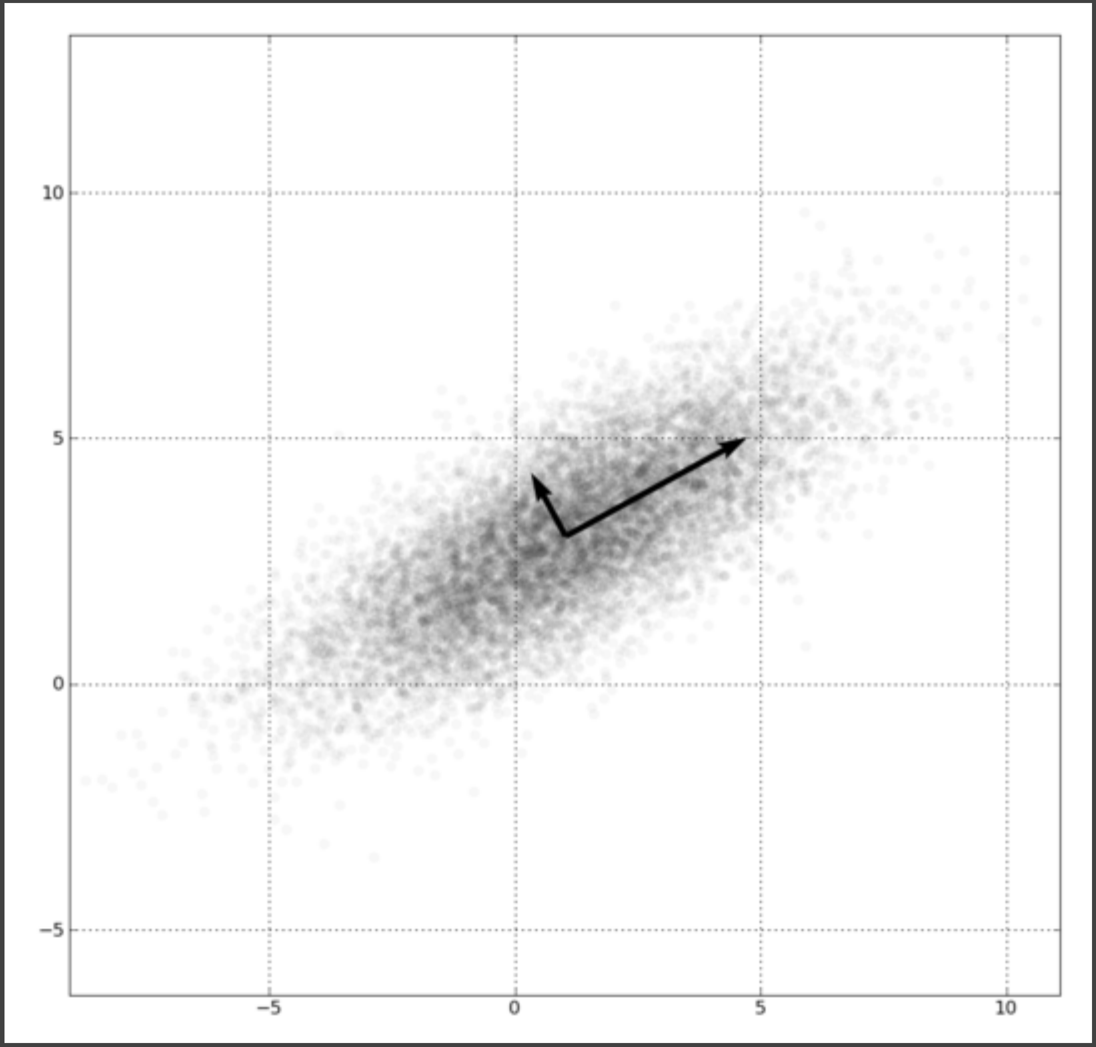
Reduction Techniques

Principal Components Analysis (PCA)
Linear transformation of basis vectors, ordered by amount of data variance they explain.

t-Dist. Stochastic Neighbor Embedding (t-SNE)
Probabilistically model distance, optimize positions.

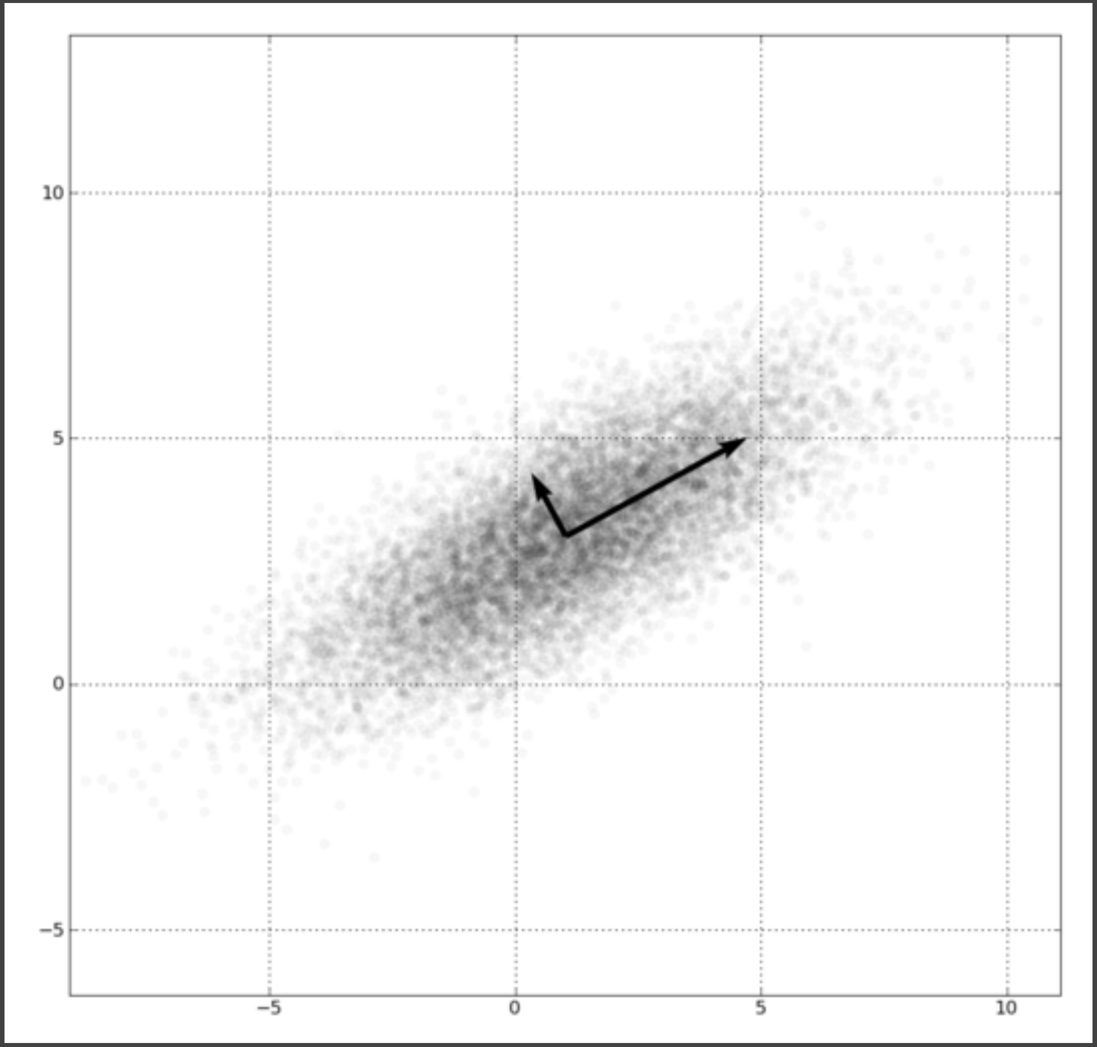
Uniform Manifold Approx. & Projection (UMAP)
Identify local manifolds, then stitch them together.

Principal Components Analysis



1. Mean-center the data.
2. Find ⊥ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

Principal Components Analysis



Linear transform: scale and rotate original space.

Lines (vectors) project to lines.

Preserves global distances.

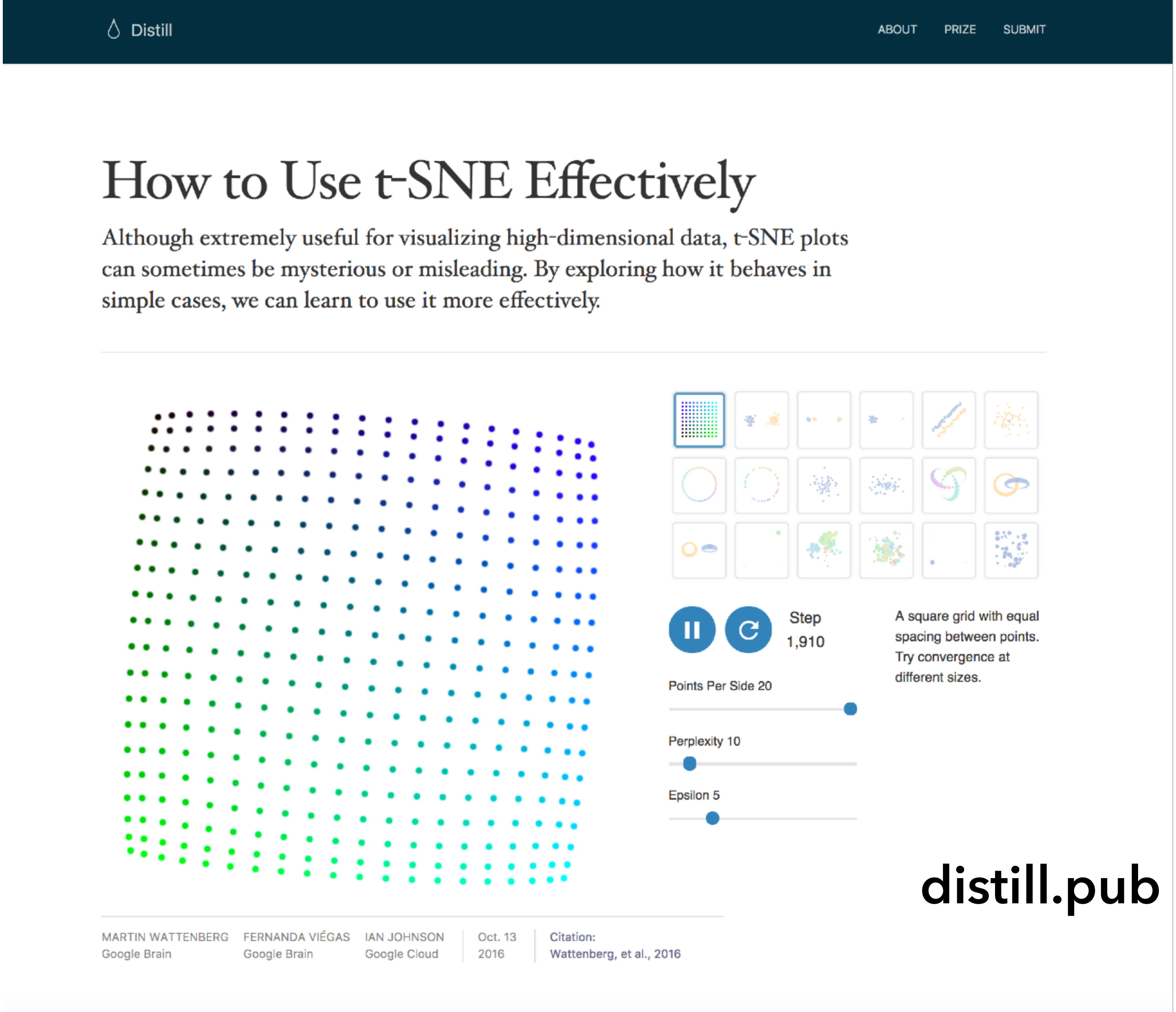
Non-Linear Techniques

Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

t-SNE - probabilistic interpretation of distance

UMAP - tries to balance local/global trade-off



Distill   ABOUT   PRIZE   SUBMIT

How to Use t-SNE Effectively

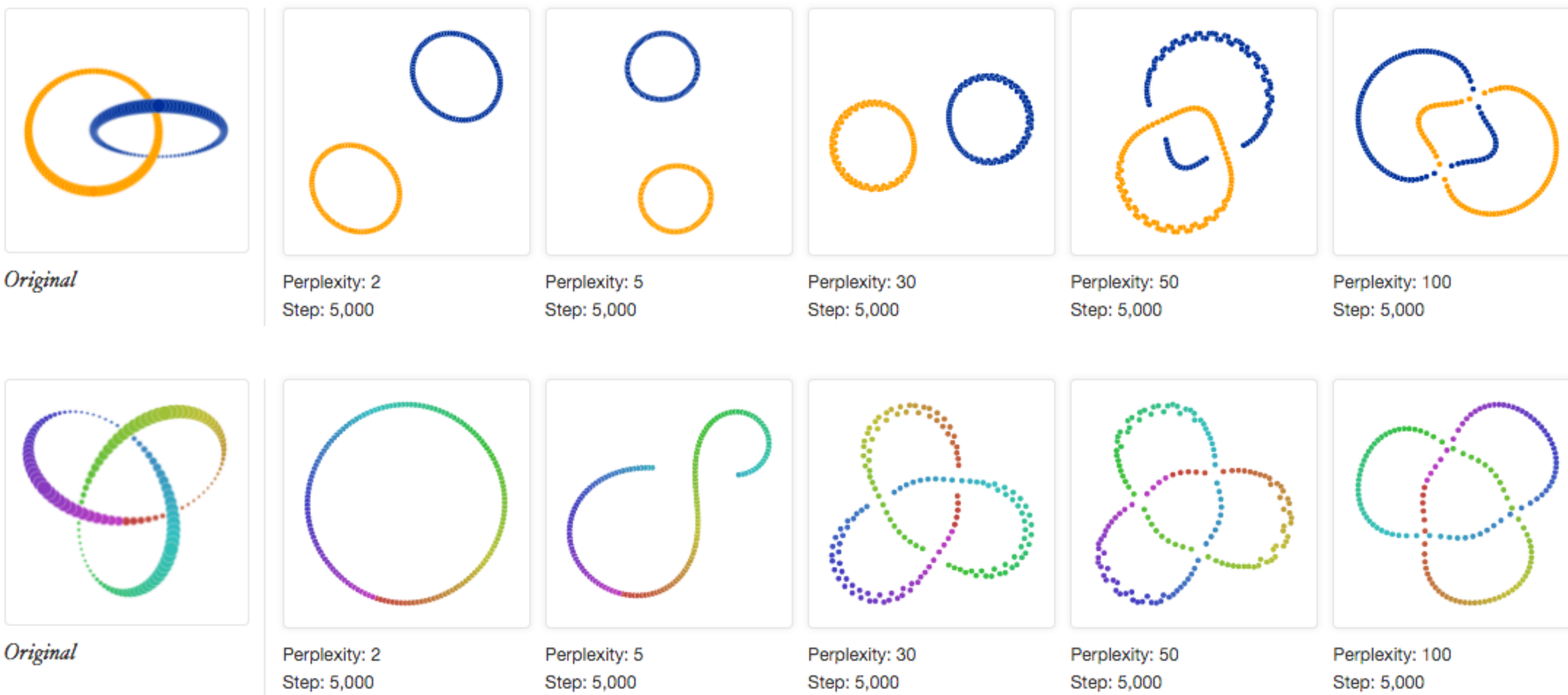
Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.

MARTIN WATTENBERG
Google Brain
FERNANDA VIÉGAS
Google Brain
IAN JOHNSON
Google Cloud

Oct. 13 2016
Citation: Wattenberg, et al., 2016

distill.pub

Visualizing t-SNE [Wattenberg et al. '16]

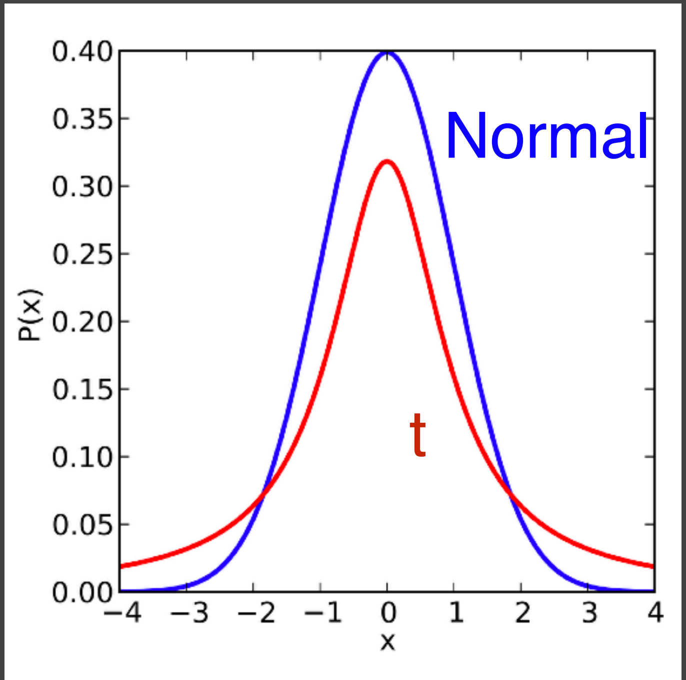


t-SNE [Maaten & Hinton 2008]

1. Model probability **P** of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

t-SNE [Maaten & Hinton 2008]

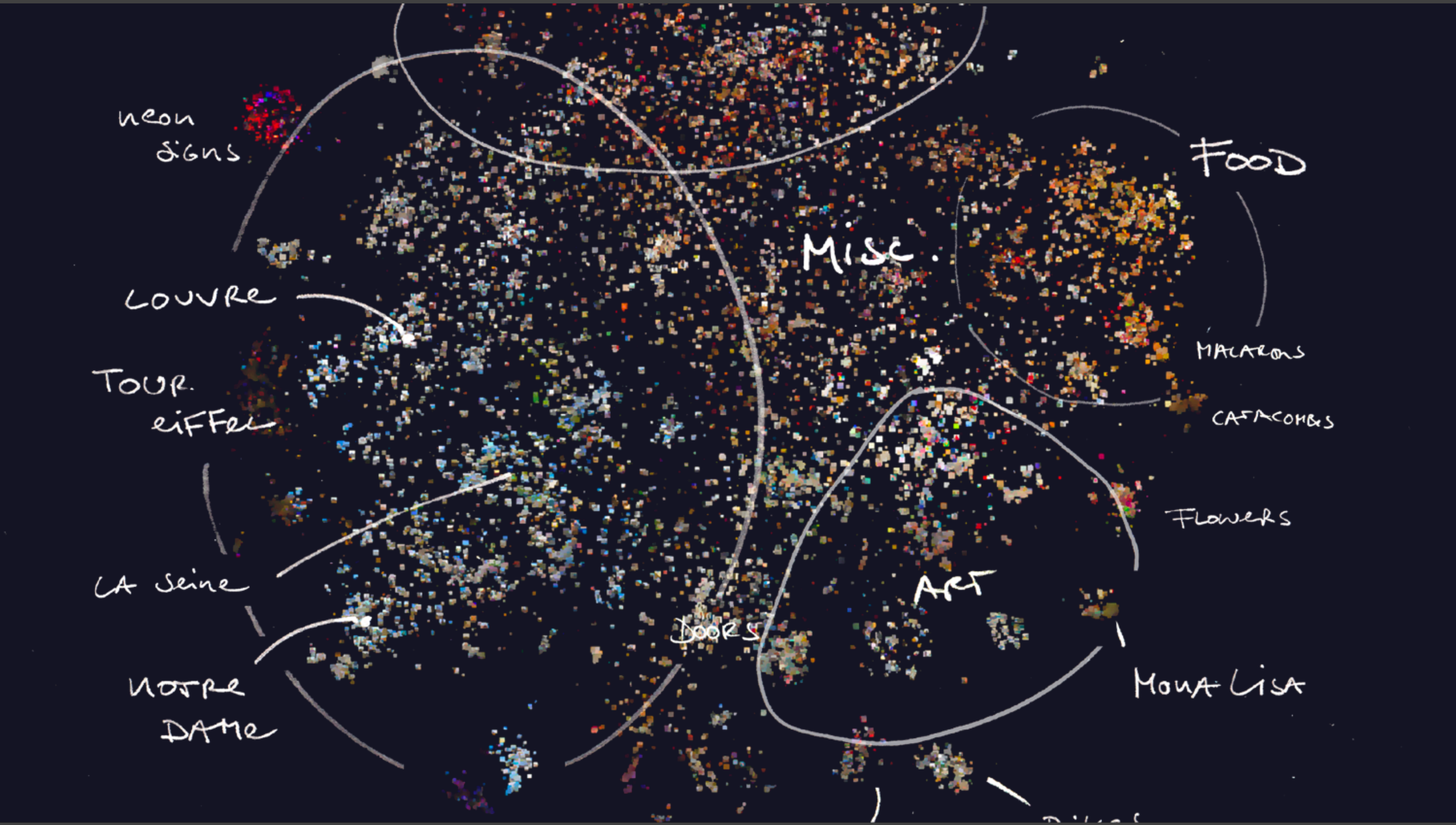
2. Define a similar probability **Q** in the low-dimensional (2D or 3D) embedding space, using a Student's *t* distribution *(hence the "t-" in "t-SNE"!)*. The *t*-distribution is heavy-tailed, allowing distant points to be even further apart.



t-SNE [Maaten & Hinton 2008]

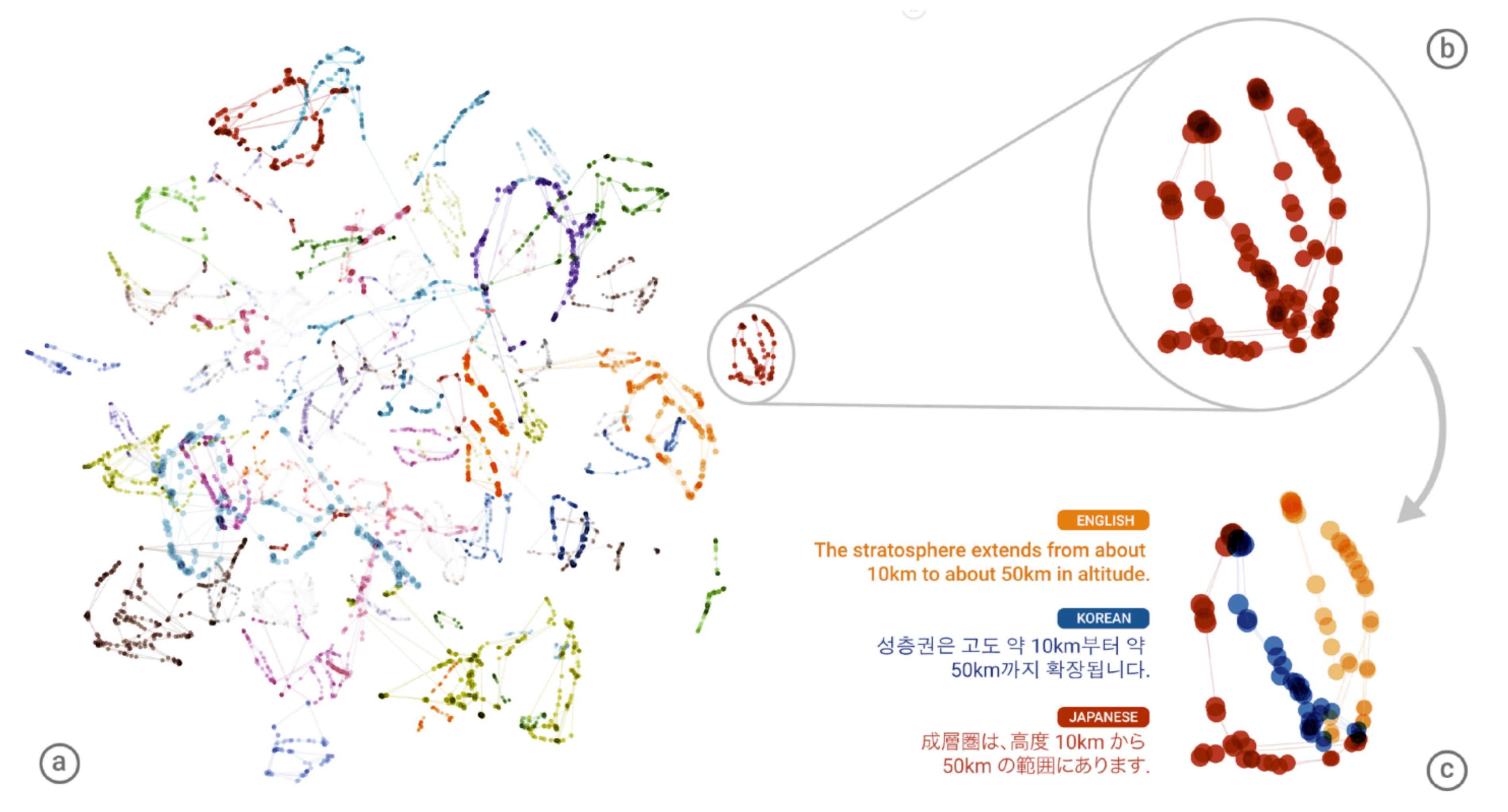
1. Model probability **P** of one point "choosing" another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
2. Define a similar probability **Q** in the low-dimensional (2D or 3D) embedding space, using a Student's *t* distribution *(hence the "t-" in "t-SNE"!)*. The *t*-distribution is heavy-tailed, allowing distant points to be even further apart.
3. Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the **P** and **Q** distributions: *KL(P || Q)*

Multiplicity [Stefaner 2018]



t-SNE projection of photos taken in Paris, France

MT Embedding [Johnson et al. 2018]



t-SNE projection of latent space of language translation model.

UMAP [McInnes et al. 2018]

Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

"Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance." - McInnes et al. 2018

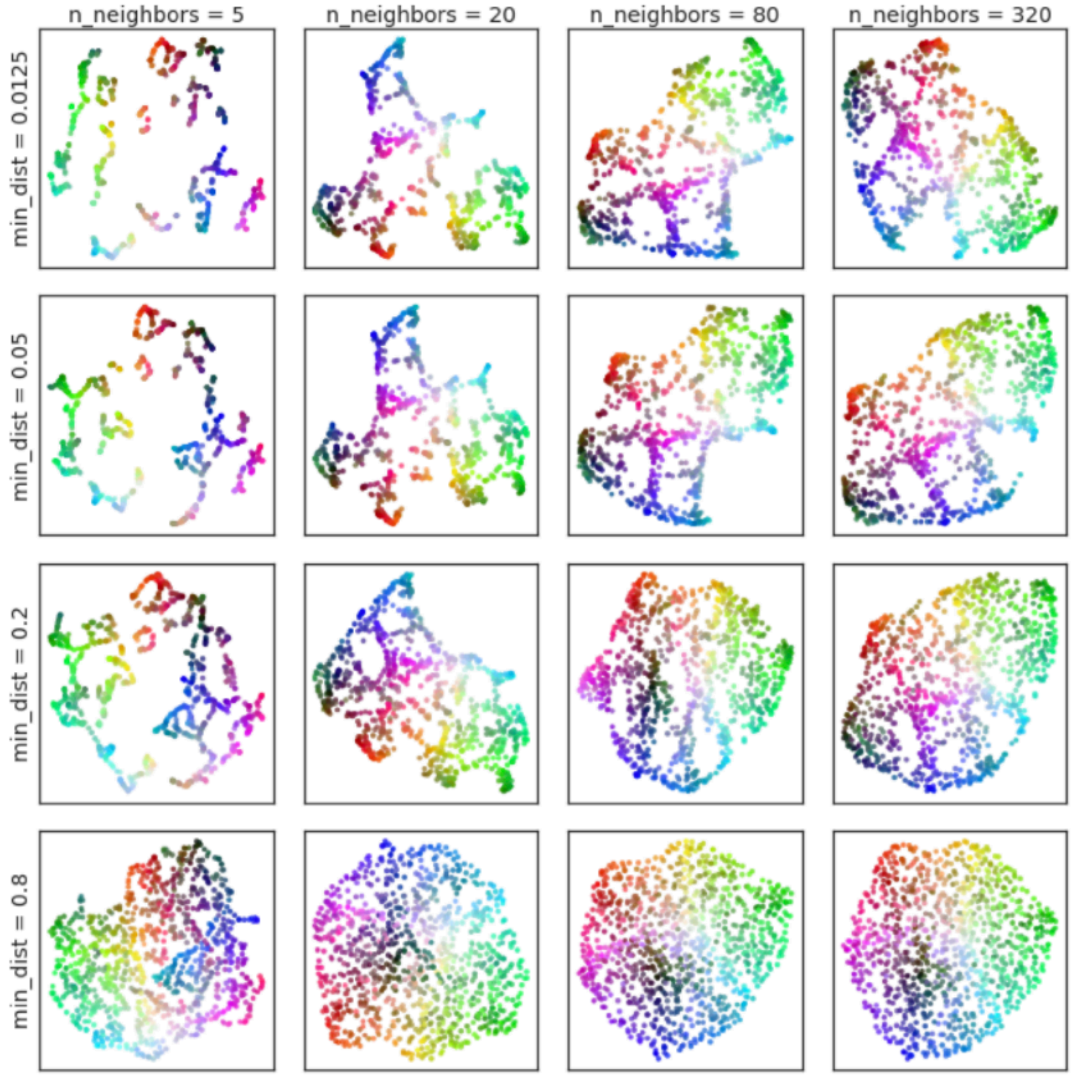
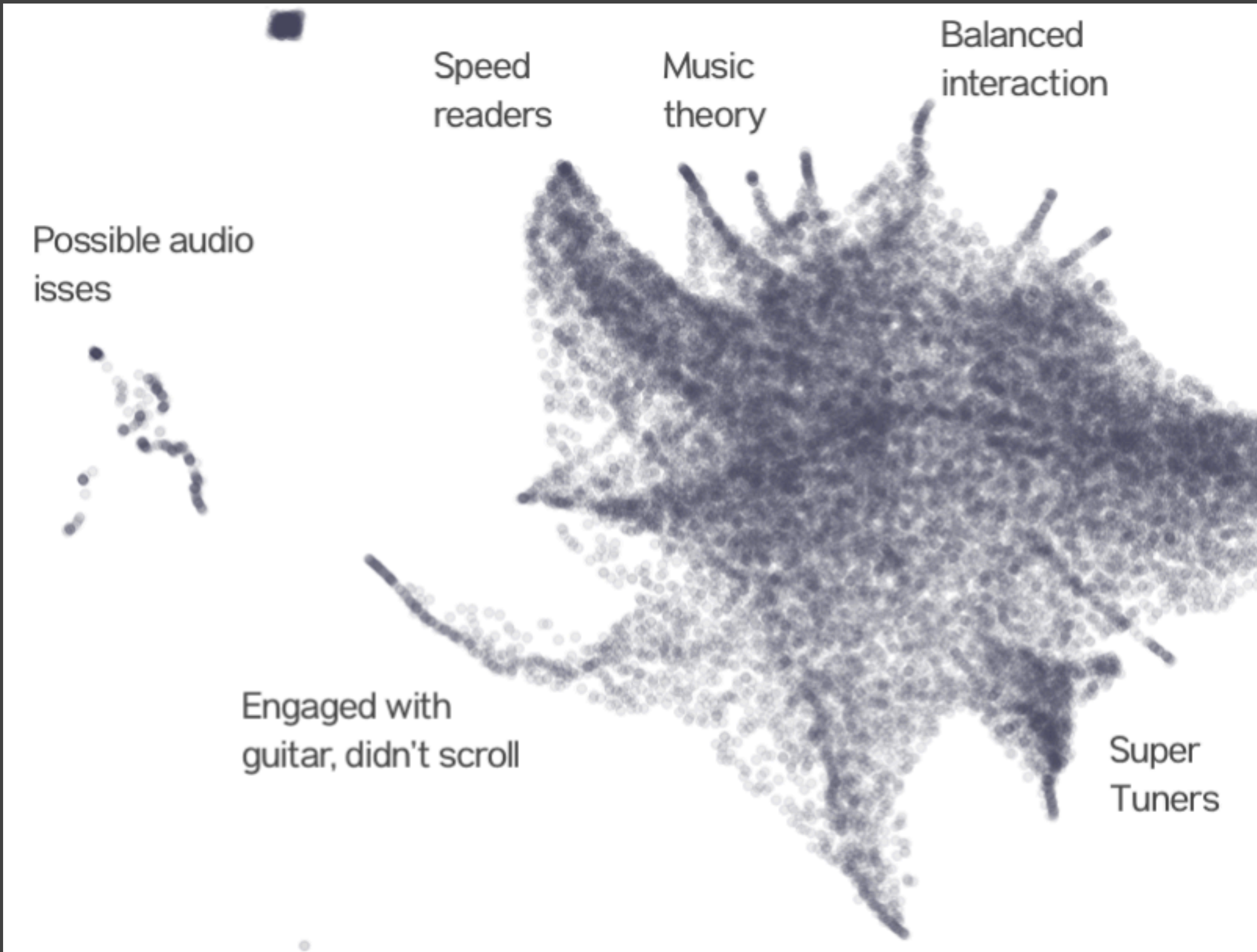


Figure 1: Variation of UMAP hyperparameters n and min-dist result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of n spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.

Reader Behavior [Conlen et al. 2019]



“Tentacles” map to activity archetypes, “blob” body maps to sessions that blend behaviors.

UMAP projection of reader activity for an interactive article.

Visualization of "Deep" Neural Network Models

TensorFlow Graph [Wongsuphasawat et al. 2018]

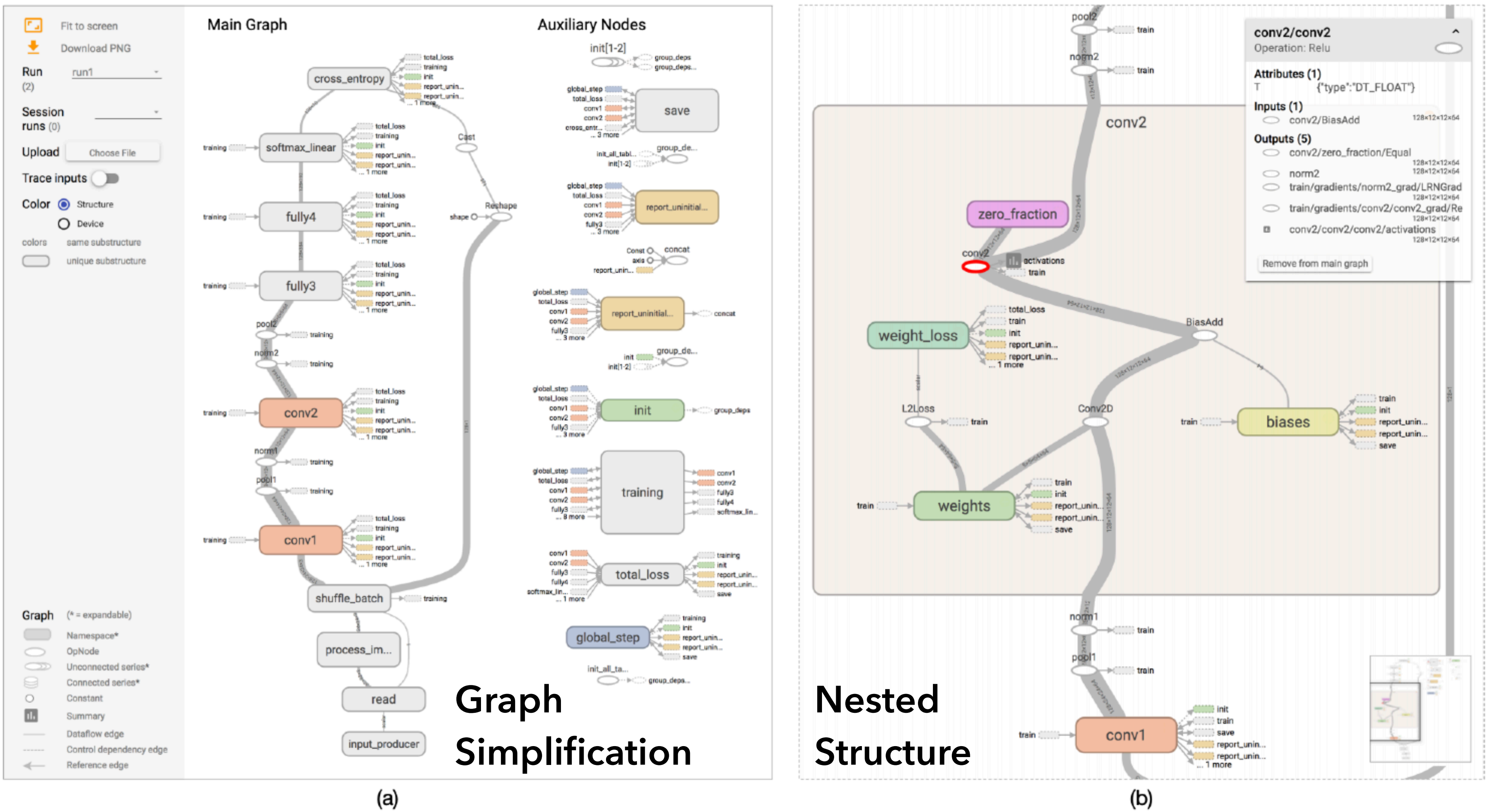


Fig. 1. The TensorFlow Graph Visualizer shows a convolutional network for classifying images (`tf_cifar`) . (a) An overview displays a dataflow between groups of operations, with *auxiliary nodes* extracted to the side. (b) Expanding a group shows its nested structure.

ActiVis [Kahng et al. 2017]

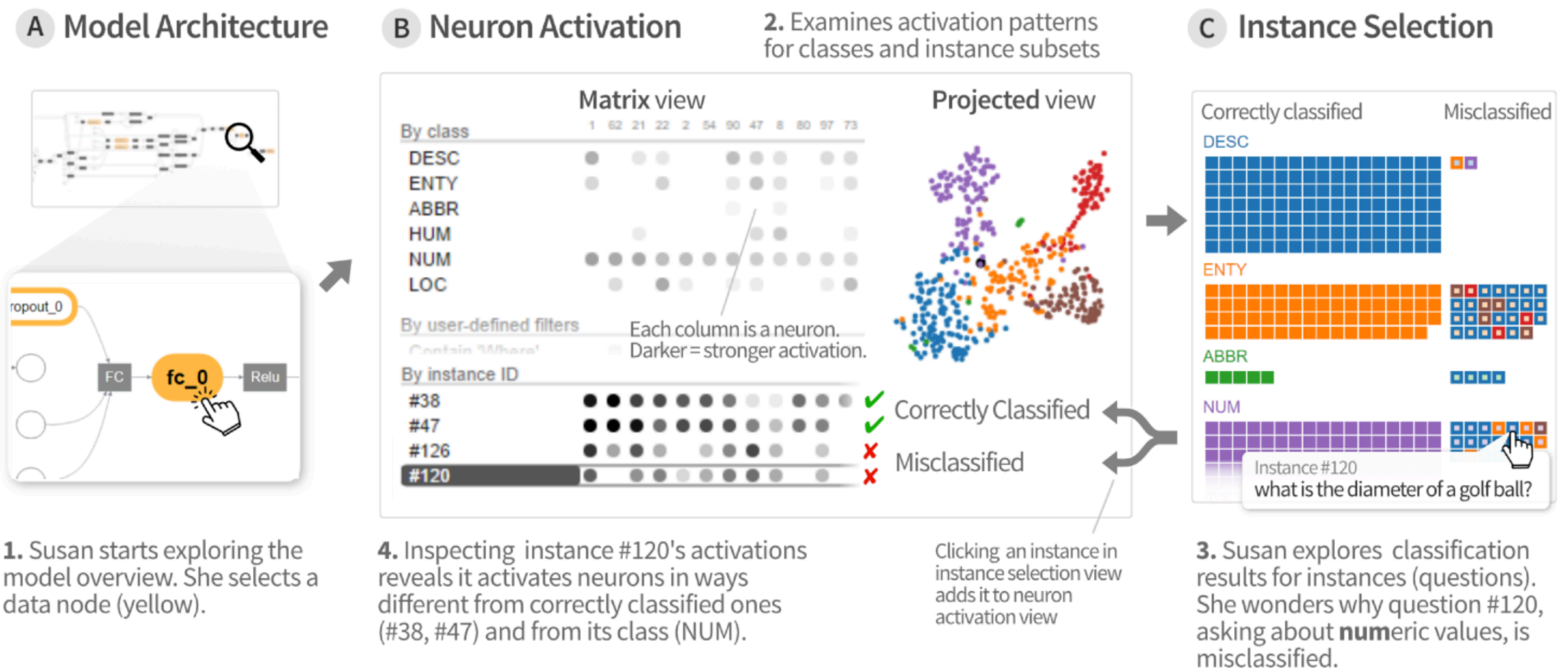


Fig. 1. ActiVis integrates several coordinated views to support exploration of complex deep neural network models, at both instance- and subset-level. **1.** Our user Susan starts exploring the model architecture, through its *computation graph* overview (at A). Selecting a *data node* (in yellow) displays its *neuron activations* (at B). **2.** The *neuron activation matrix view* shows the activations for instances and instance subsets; the *projected view* displays the 2-D projection of instance activations. **3.** From the *instance selection* panel (at C), she explores individual instances and their classification results. **4.** Adding instances to the matrix view enables comparison of activation patterns across instances, subsets, and classes, revealing causes for misclassification.

Seq2Seq-Vis [Strobelt et al. 2018]

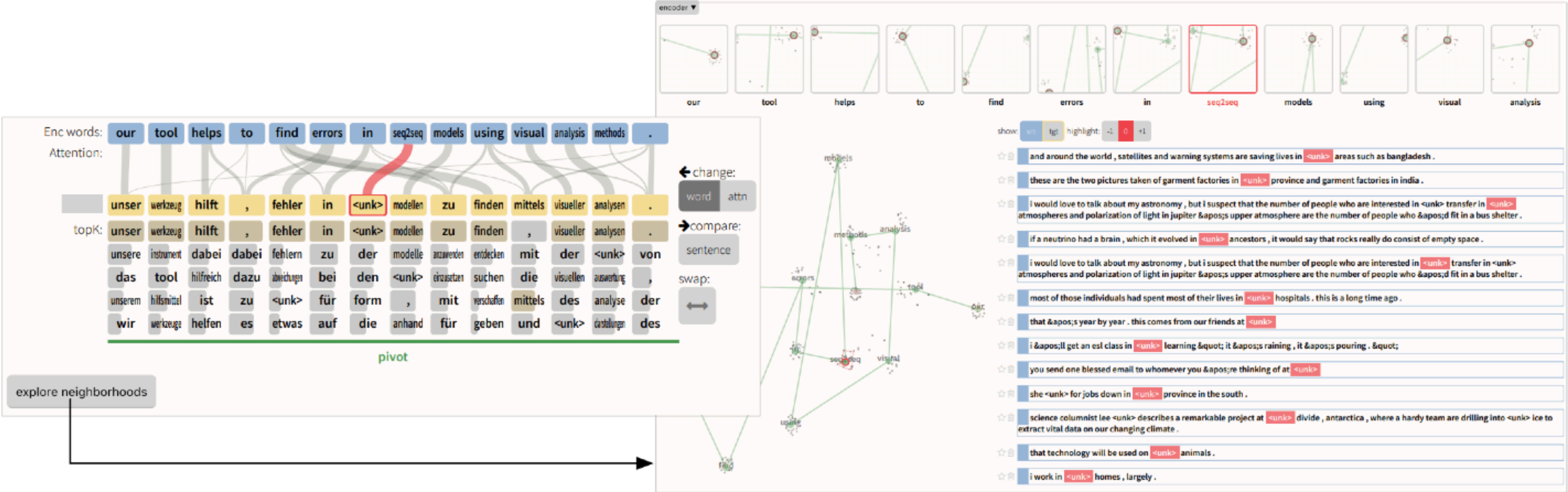
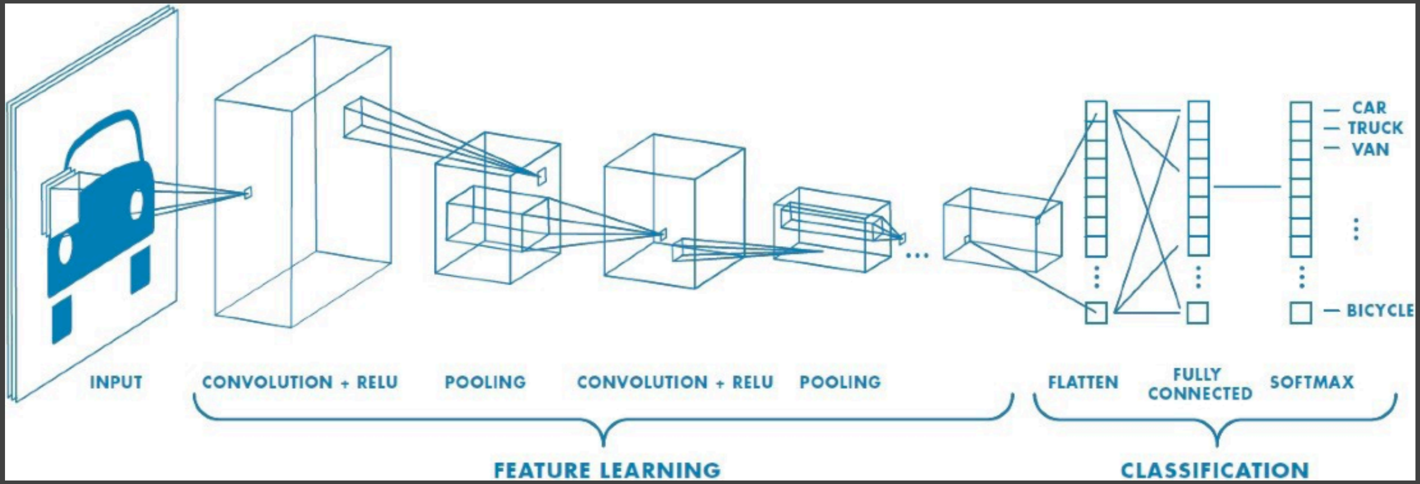


Fig. 1. Example of `Seq2Seq-Vis`. In the translation view (left), the source sequence *"our tool helps to find errors in seq2seq models using visual analysis methods."* is translated into a German sentence. The word *"seq2seq"* has correct attention between encoder and decoder (red highlight) but is not part of the language dictionary. When investigating the encoder neighborhoods (right), the user sees that *"seq2seq"* is close to other unknown words *"⟨unk⟩"*. The buttons enable user interactions for deeper analysis.

Local Explanations (Model Specific)

Convolutional Neural Nets

CNNs for Image Processing



Prototypical CNN Architecture

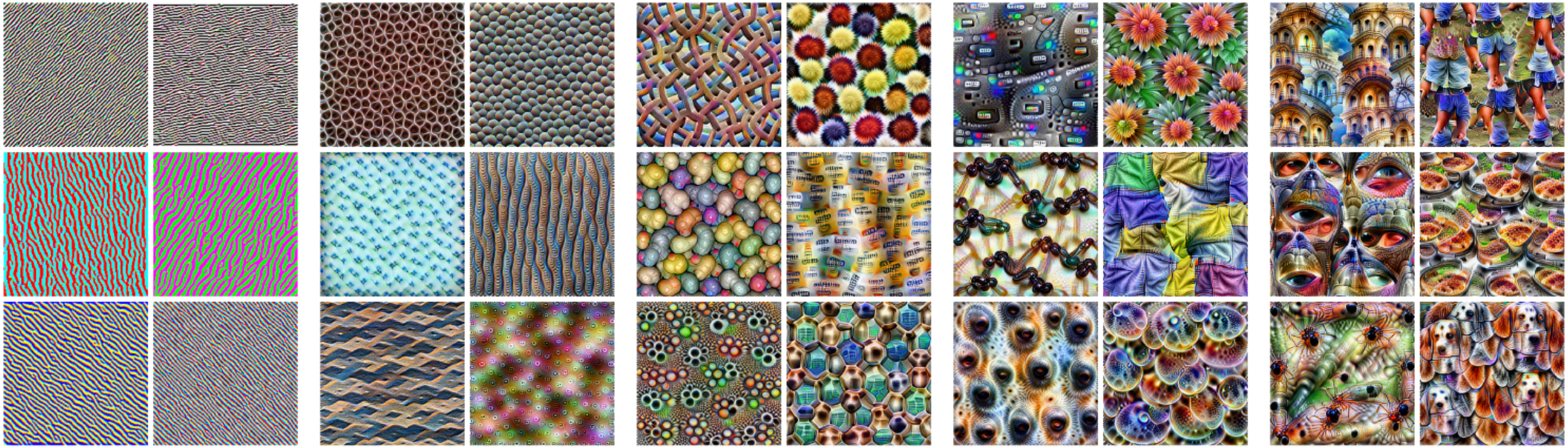
GoogLeNet - 22 layers!

Feature Visualization [Olah et al. 2017]

Convolutional Neural Network (CNN) for Images

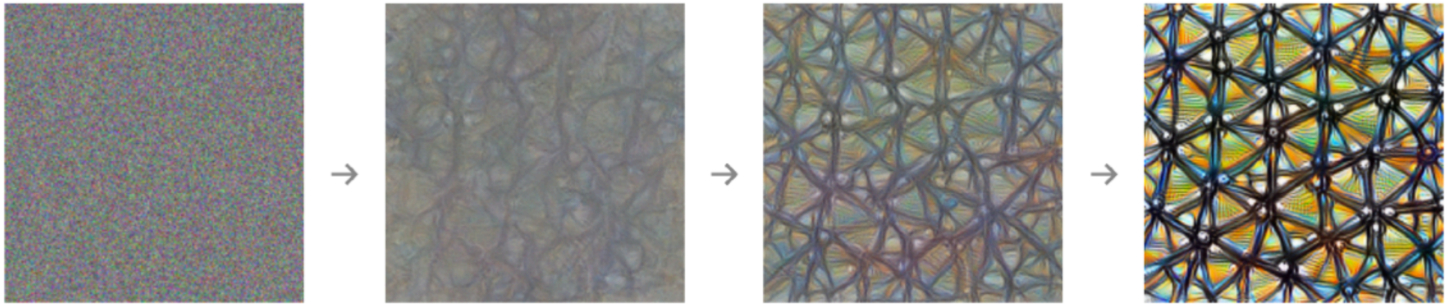
Basic Idea:
Select one or more “neurons” in a network layer
Optimize to find input that maximizes excitation

Feature Visualization for CNNs



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

Starting from random noise, we optimize an image to activate a particular neuron (layer mixed4a, unit 11).



Step 0 → Step 4 → Step 48 → Step 2048

Single Unit Visualizations

Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

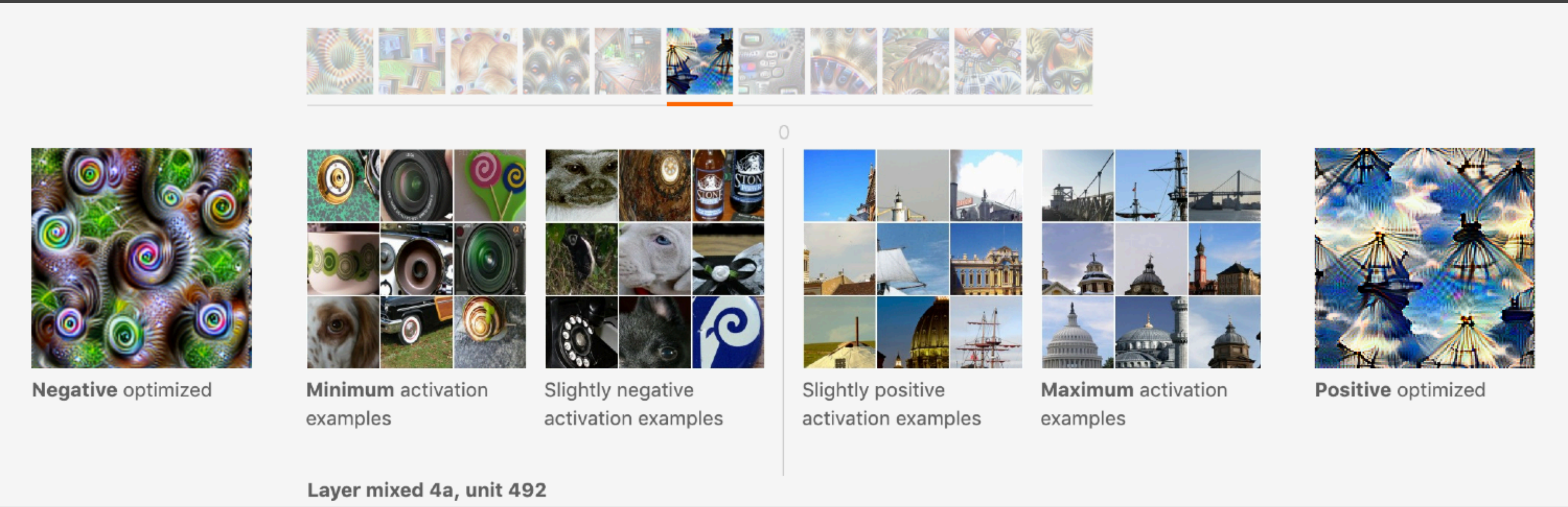
Baseball—or stripes?
mixed4a, Unit 6

Animal faces—or snouts?
mixed4a, Unit 240

Clouds—or fluffiness?
mixed4a, Unit 453

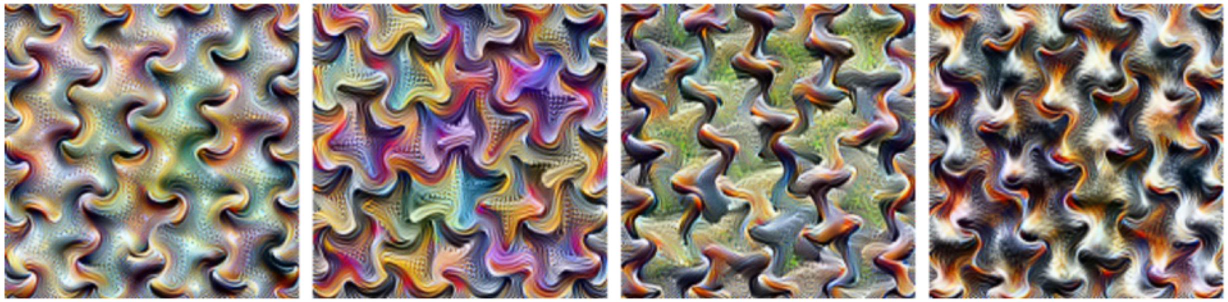
Buildings—or sky?
mixed4a, Unit 492

Single Unit Visualizations



Optimizing for Diversity

Simple Optimization



Dataset examples

Optimization with diversity reveals four different, curvy facets. *Layer mixed4a, Unit 97*

REPRODUCE IN A NOTEBOOK

Simple Optimization

Dataset examples

Optimization with diversity reveals multiple types of balls. *Layer mixed5a, Unit 9*

Multi-Unit Visualization

Dataset examples and optimized examples of **random directions** in activation space. The directions shown here were hand-picked for interpretability.

REPRODUCE IN A NOTEBOOK

mixed3a, random direction *mixed4c, random direction* *mixed4d, random direction* *mixed5a, random direction*

Multi-Unit Visualization

By jointly optimizing two neurons we can get a sense of how they interact.

REPRODUCE IN A NOTEBOOK

Feature Visualization for CNNs

Convolutional Neural Network (CNN) for Images

Basic Idea:
Select one or more “neurons” in a network layer
Optimize to find input that maximizes excitation

Challenges:
Choice of optimization? What dimensions to inspect? How to constrain or regularize?
Unconstrained approach leads to model artifacts.
Applicability to non-image data?

Local Explanations (Model Agnostic)

LIME [Ribeiro et al. 2016]

Local
Interpretable
Model-Agnostic
Explanations

LIME Local Interpretable Model-Agnostic Explanations

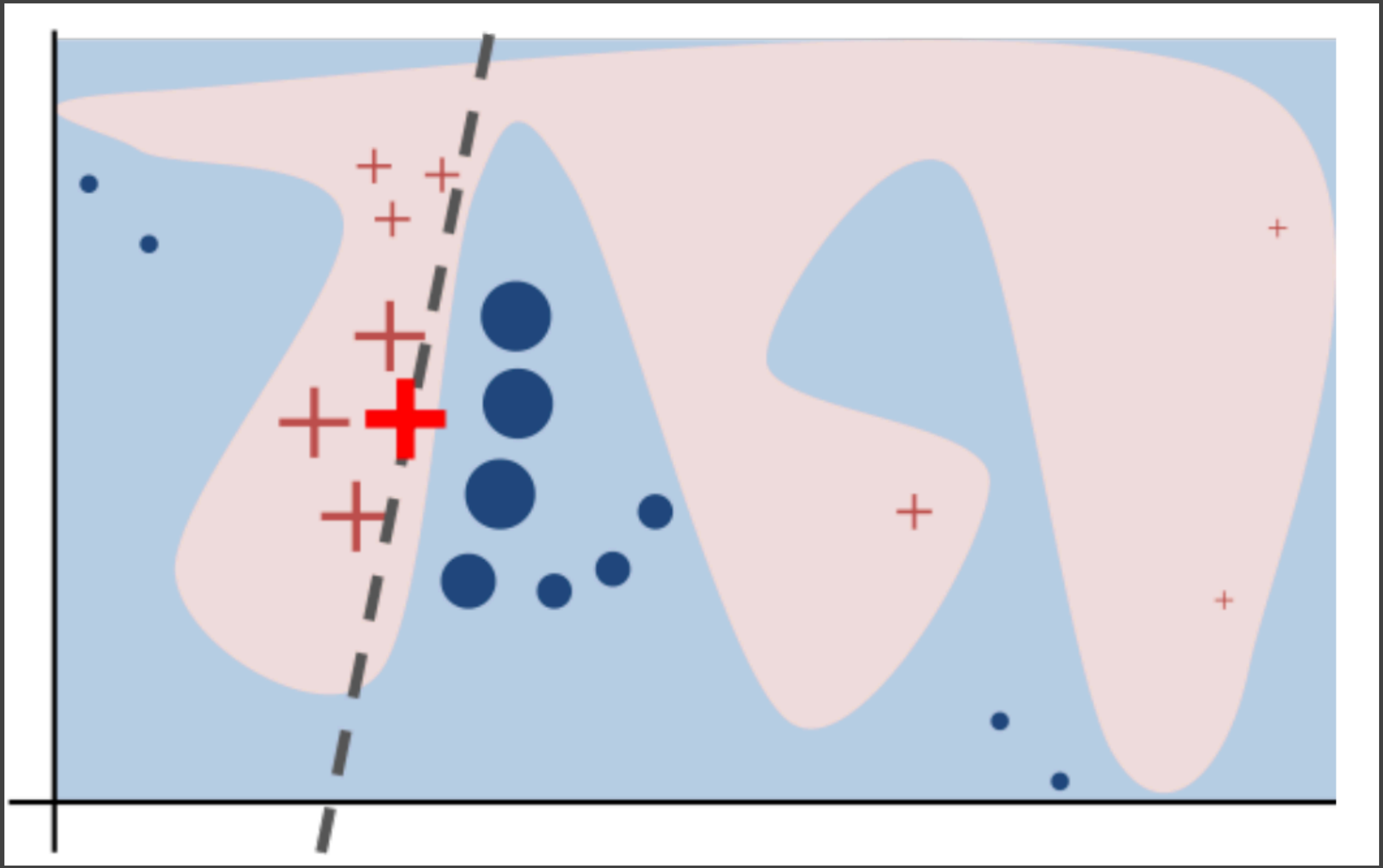
Model-agnostic: take *any* classifier as input

LIME **Local Interpretable Model-Agnostic Explanations**

Model-agnostic: take *any* classifier as input

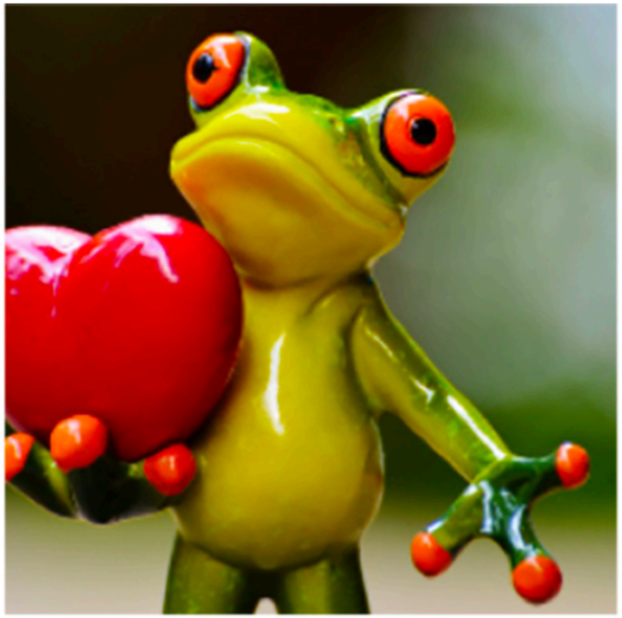
For a given prediction:
Identify aspects meaningful to a person
Perturb those aspects around the prediction
(e.g., remove words or image regions)
Fit local "interpretable" model to the results
(e.g., locally-weighted linear model)

LIME Intuition



Despite complex global structure, a locally weighted linear model may suffice to explain.

LIME Local Interpretable Model-Agnostic Explanations

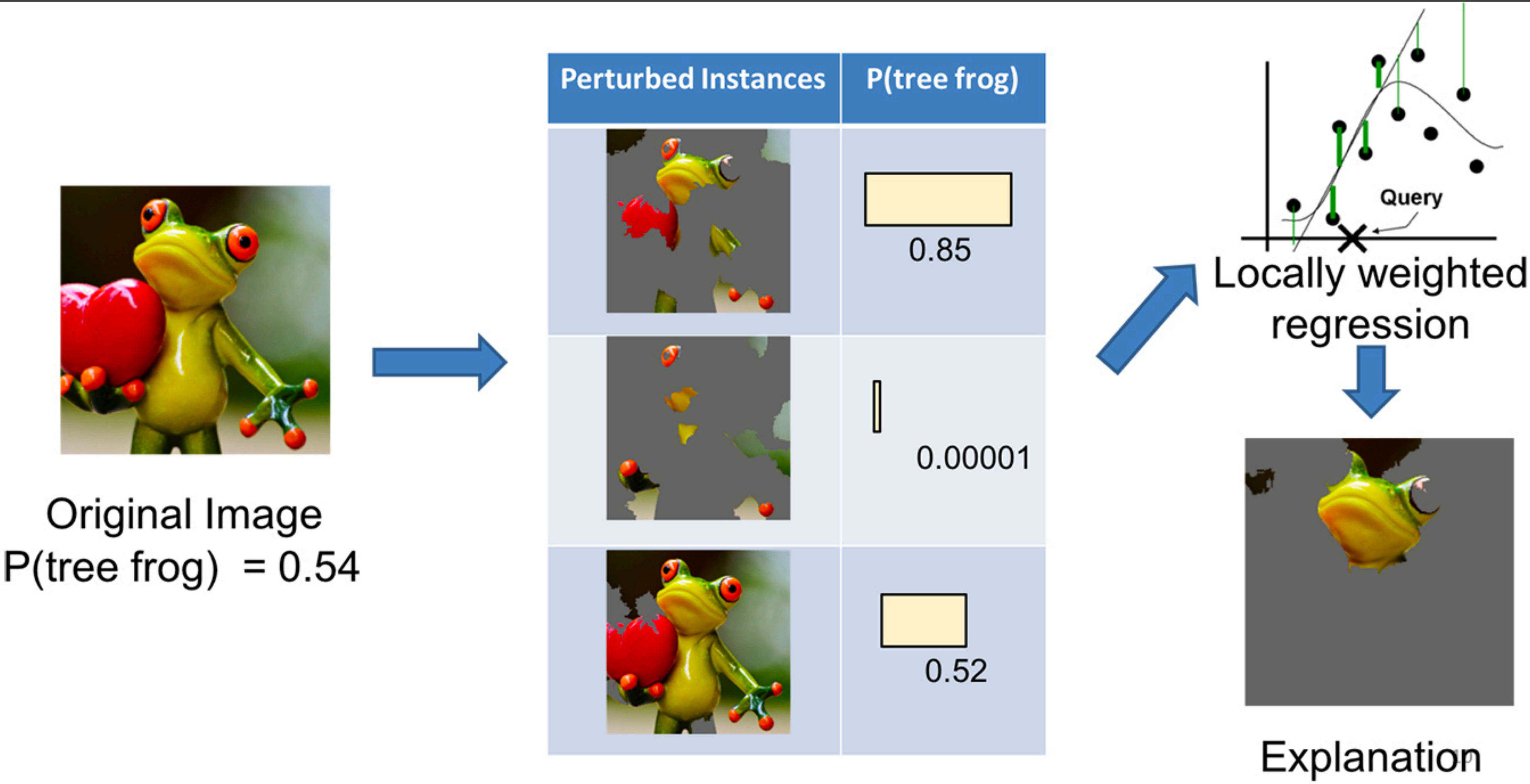


Original Image

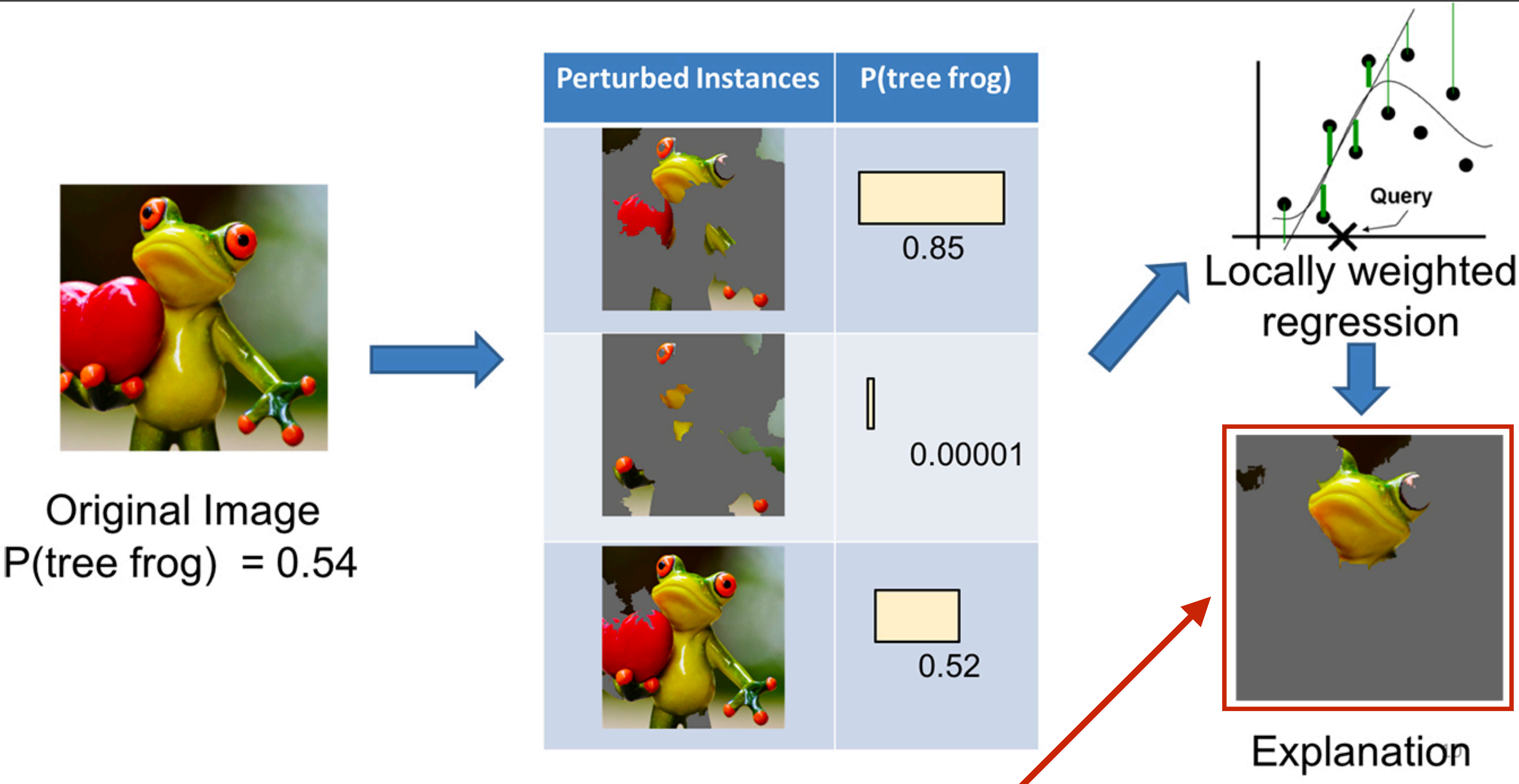
Interpretable Components

Why is this predicted to be a "tree frog"?

LIME Local Interpretable Model-Agnostic Explanations



LIME Local Interpretable Model-Agnostic Explanations



Regions sufficient for “frog” detection

LIME **Local Interpretable Model-Agnostic Explanations**

Model-agnostic: take *any* classifier as input

For a given prediction:

Identify aspects meaningful to a person

Perturb those aspects around the prediction

(e.g., remove words or image regions)

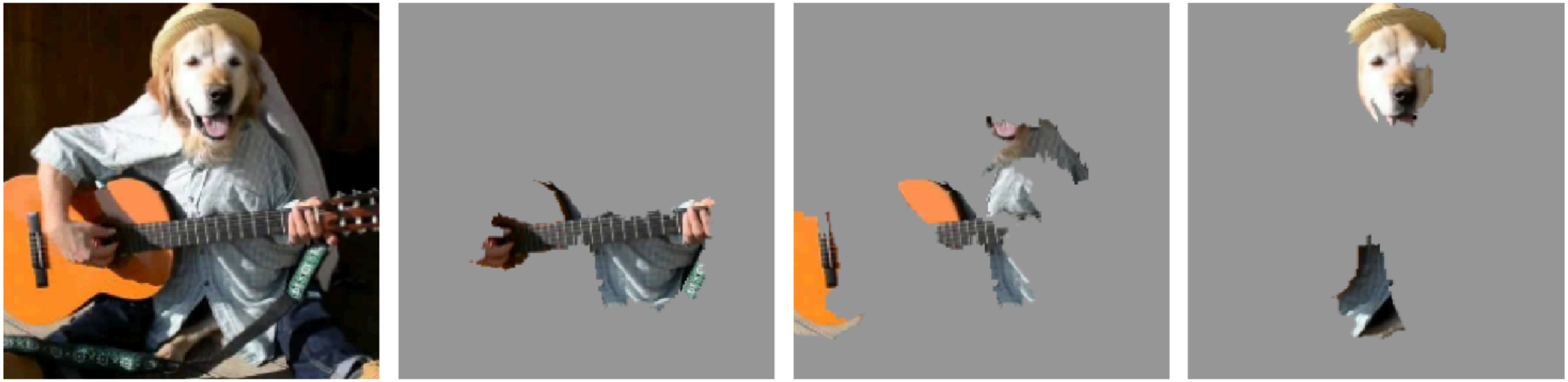
Fit local "interpretable" model to the results

(e.g., locally-weighted linear model)

For an entire model:

Optimize for a set of representative examples

LIME Local Interpretable Model-Agnostic Explanations



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

LIME Local Interpretable Model-Agnostic Explanations

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

LIME Local Interpretable Model-Agnostic Explanations

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Detects snow, not wolves!

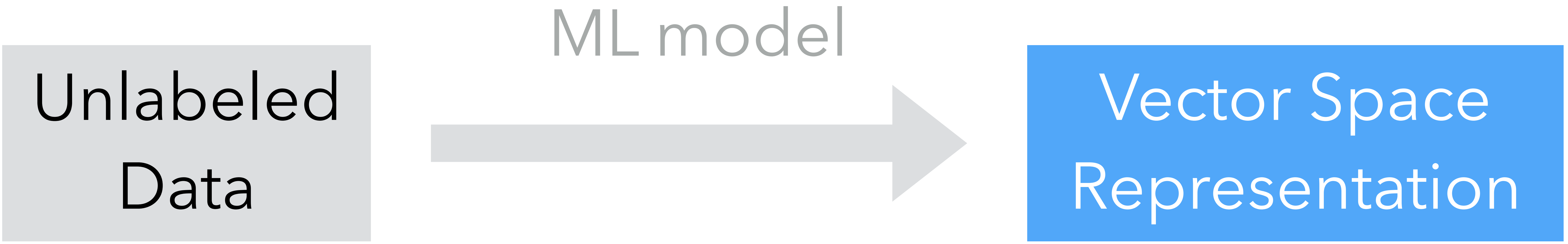
Latent Space Cartography:

Visual Analysis of Vector Space Embeddings

Yang Liu, Eunice Jun, Qisheng Li, Jeffrey Heer
University of Washington
Interactive Data Lab

https://github.com/uwdata/latent-space-cartography

DESIGN USE BUILD


Unlabeled Data
ML model
Vector Space Representation

Vector Space: Example

Word embeddings

represent a word as a vector of numbers

dragon	0	1	0	2	1	0	0	0	1	0

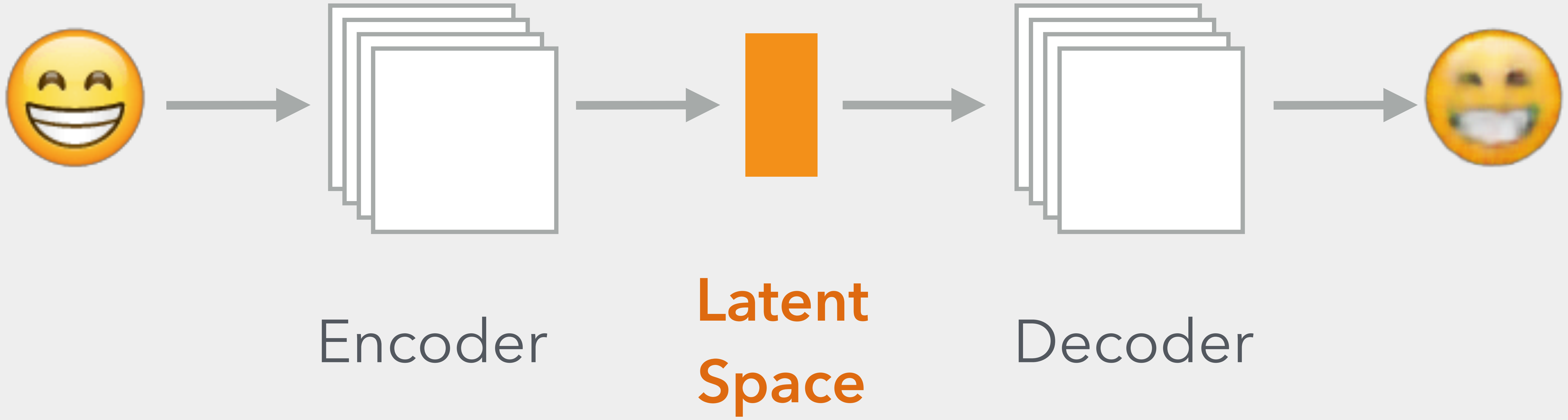
Vector Space: Example

Word embeddings

represent a word as a vector of numbers

dragon | 0 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 0

Latent Spaces in Generative Models



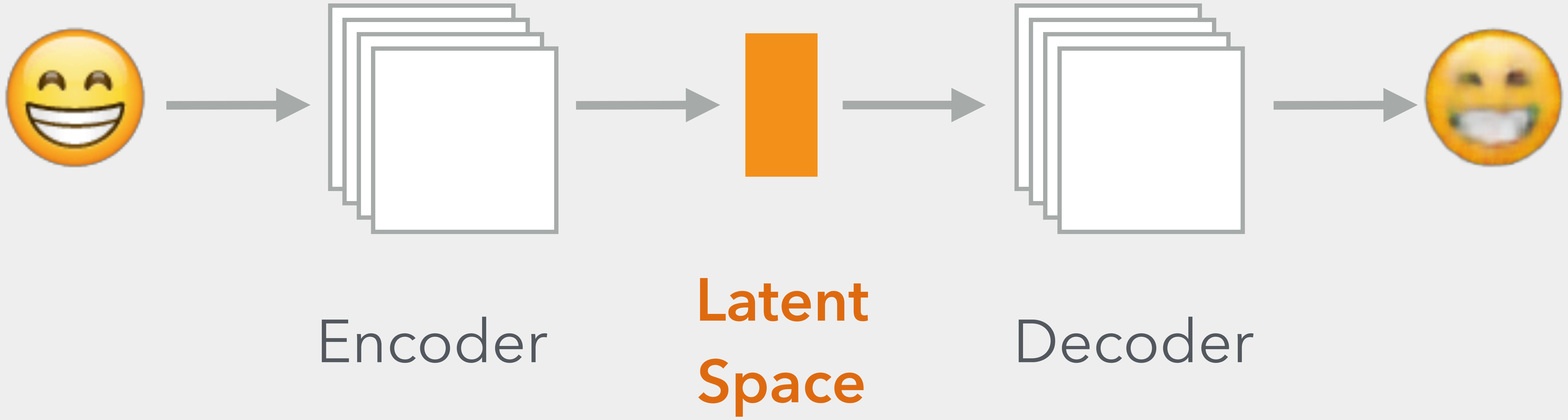
Vector Space: Example

Latent Spaces

Word embeddings

represent a word as a vector of numbers

Latent Spaces in Generative Models

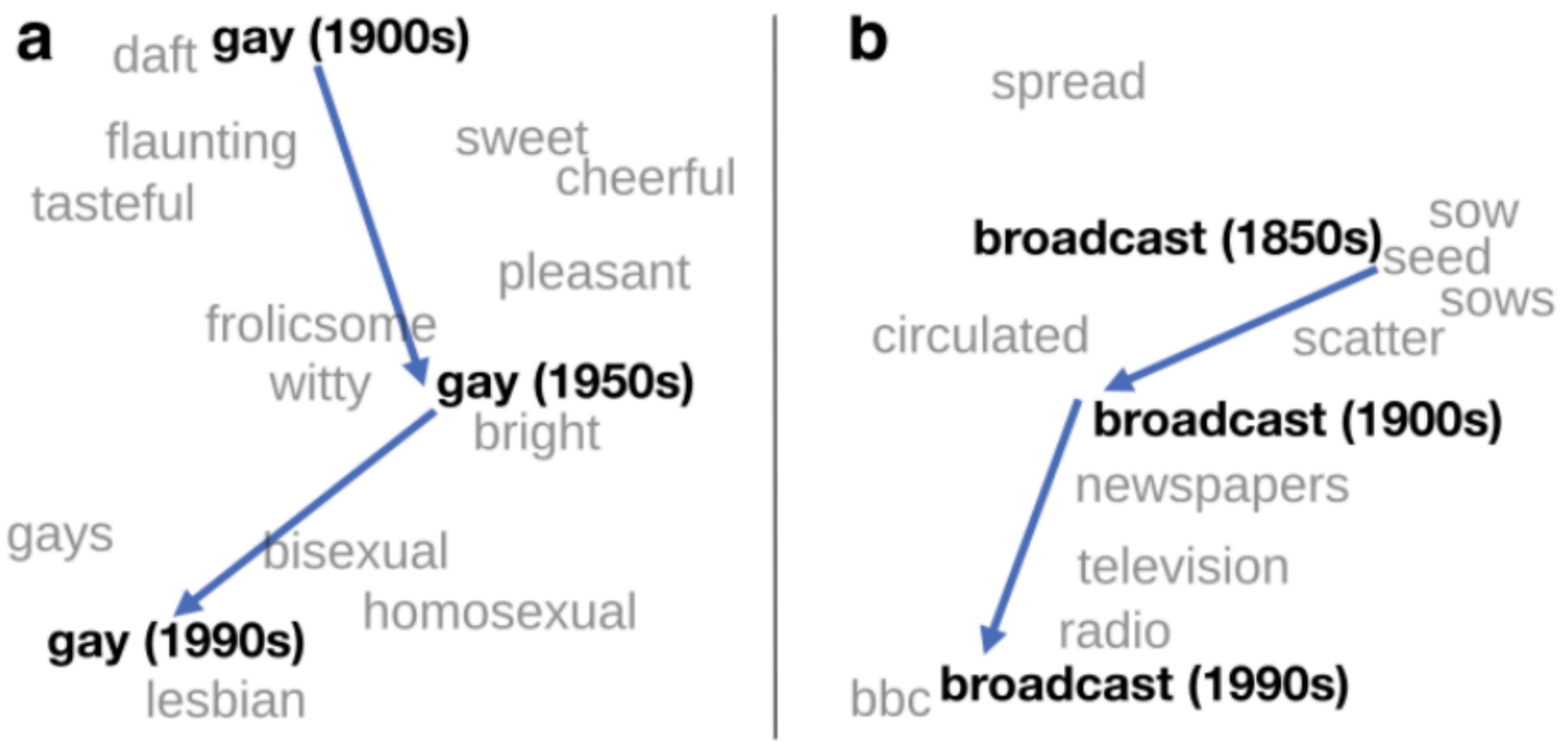


Why are latent spaces important?

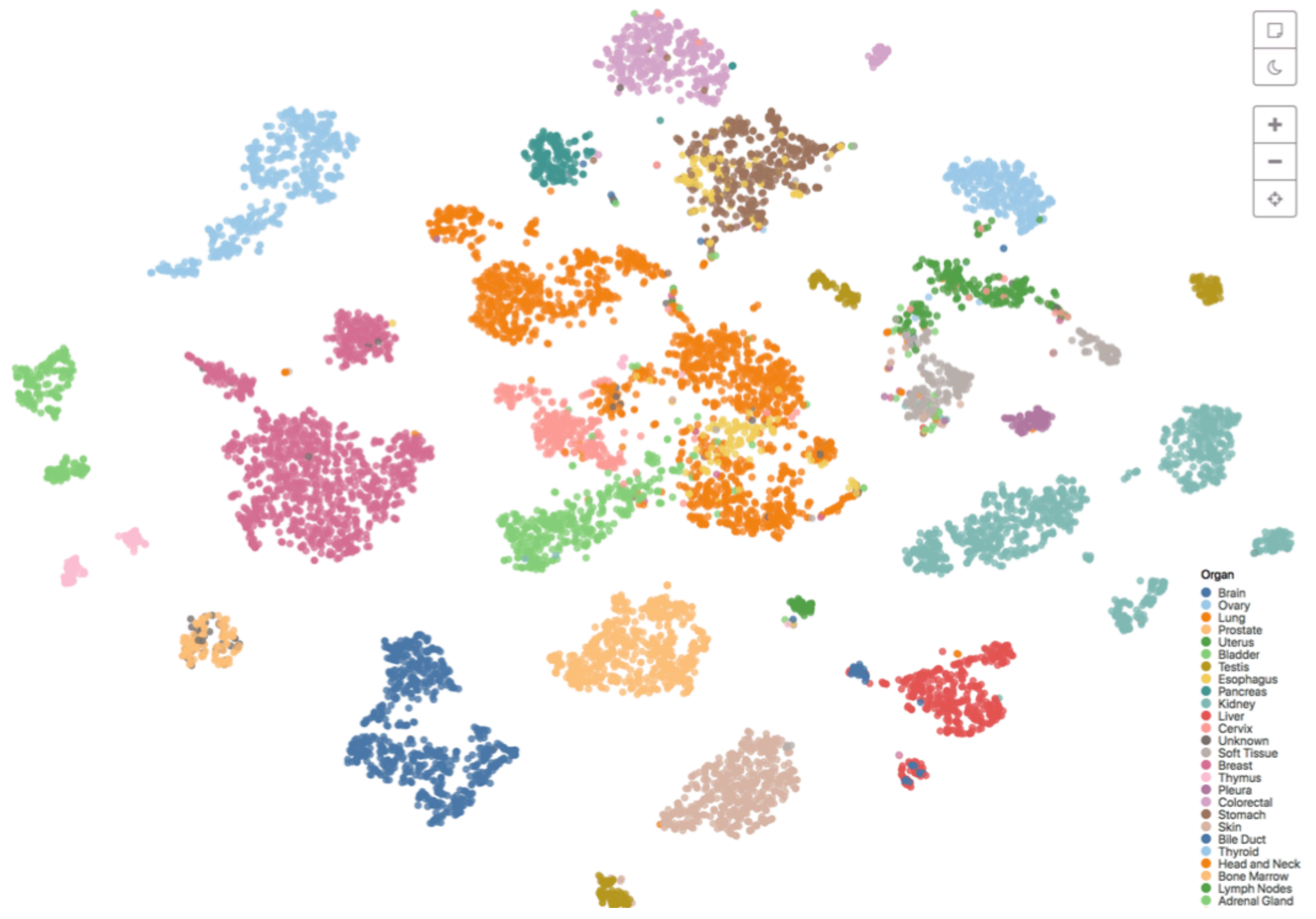
1. Serve as features for downstream ML applications

Why are latent spaces important?

1. Serve as features for downstream ML applications
2. Provide insights into the data



How the meaning of words change over time



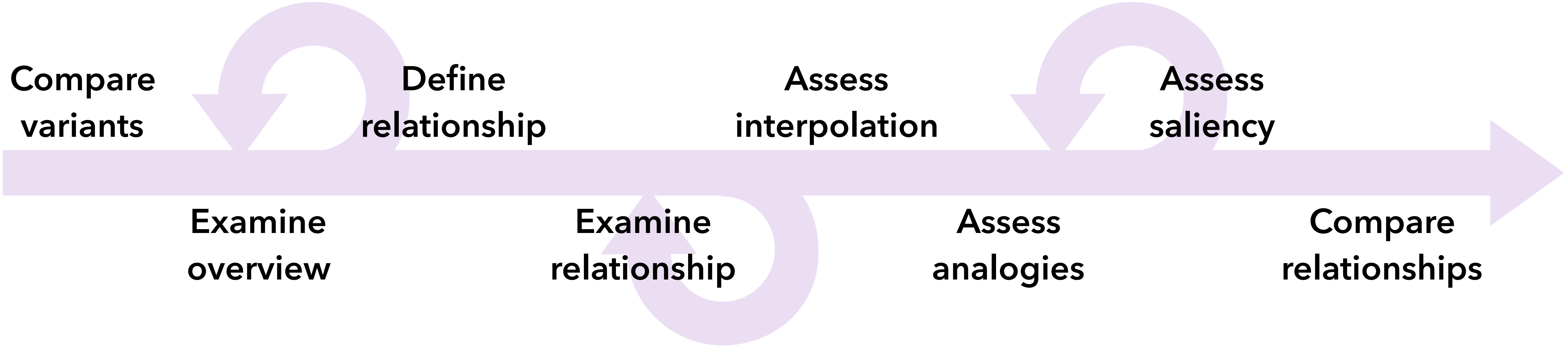
Biologically meaningful latent spaces

Human judgement is essential in **interpreting** latent spaces

Latent Space Cartography

Mapping meaningful dimensions of latent spaces

1. Surveyed 78 papers from ML, NLP and science
2. Extracted most common interpretation tasks
3. Integrated the tasks into a workflow



4. Built a visual analysis system, also named Latent Space Cartography (LSC)

Latent Space Cartography

Visual Analysis of Vector Space Embeddings

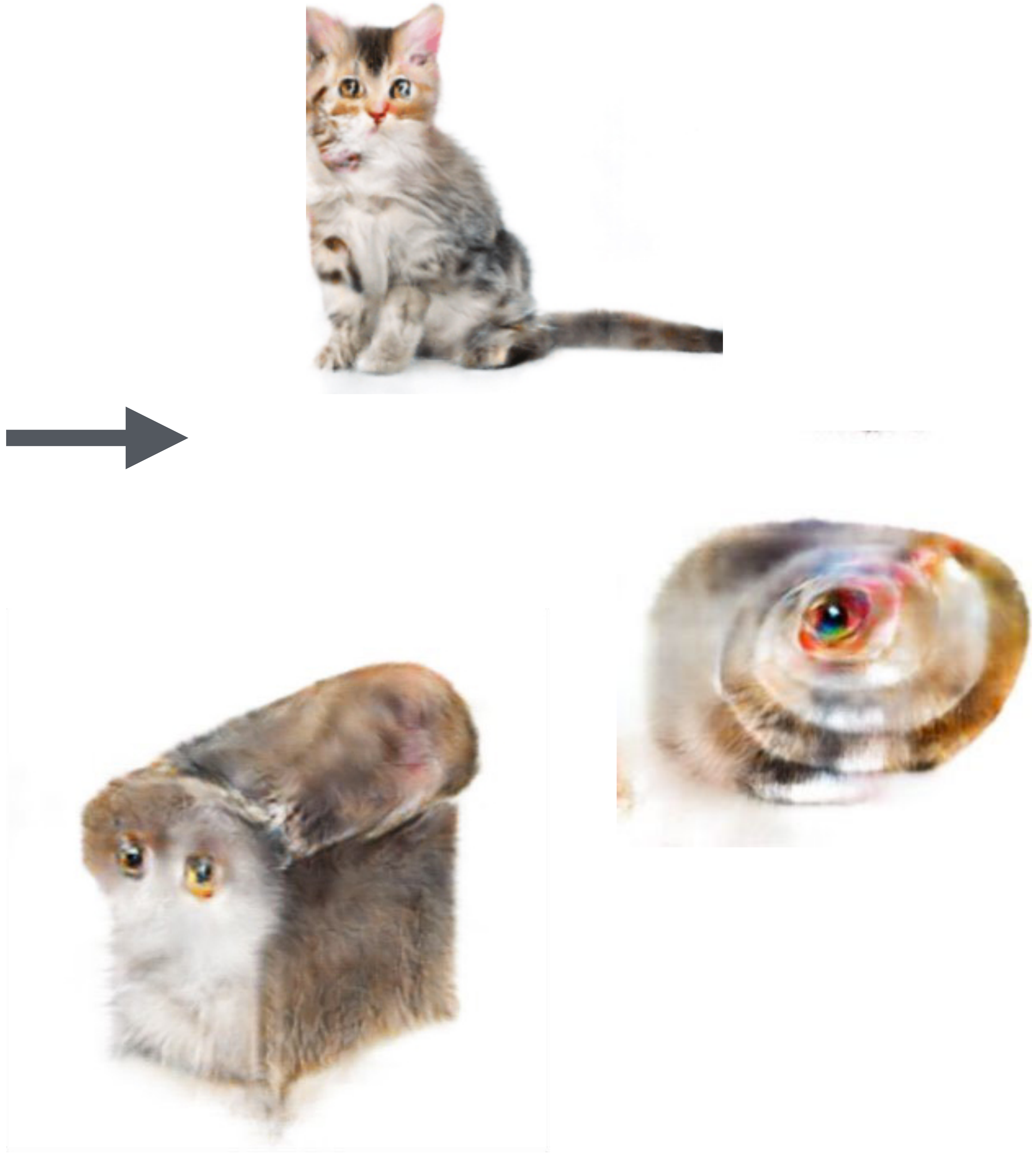
Background: Variational Auto-encoder (VAE)

Background: Variational Auto-encoder (VAE)

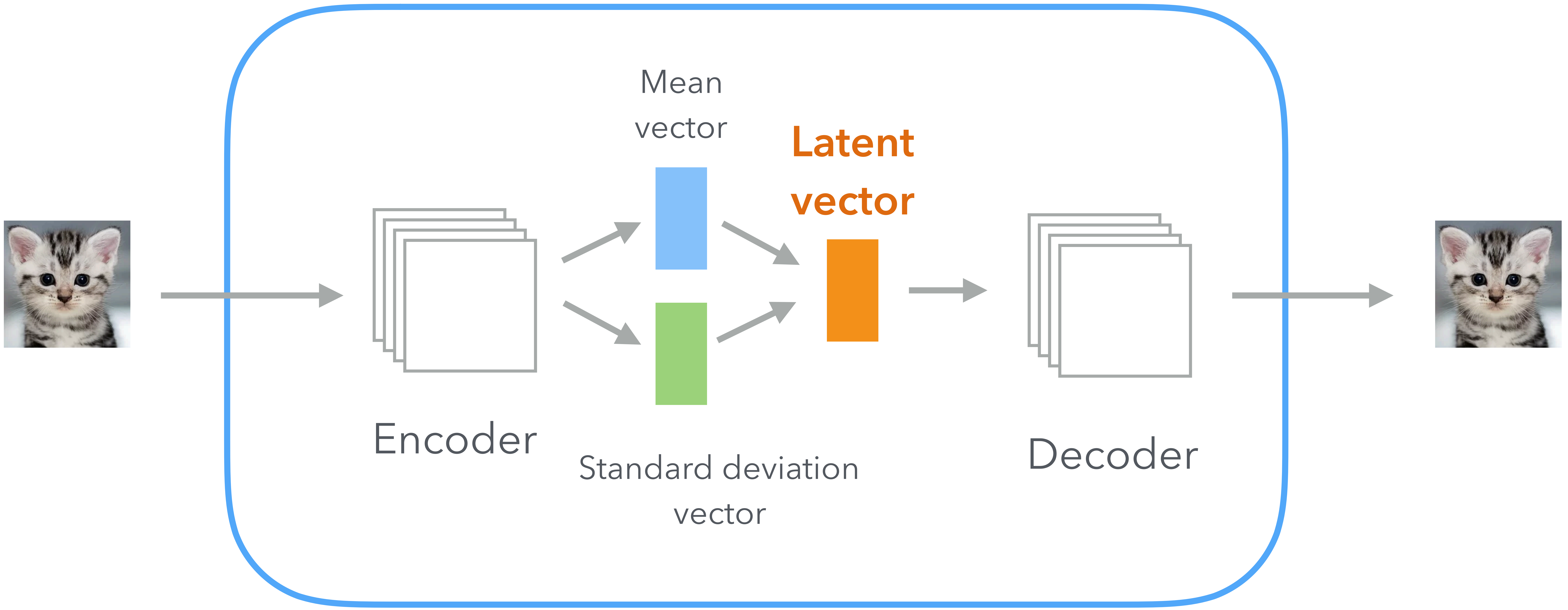
Variational
Auto-encoder
(VAE)

Background: Variational Auto-encoder (VAE)

Variational
Auto-encoder
(VAE)

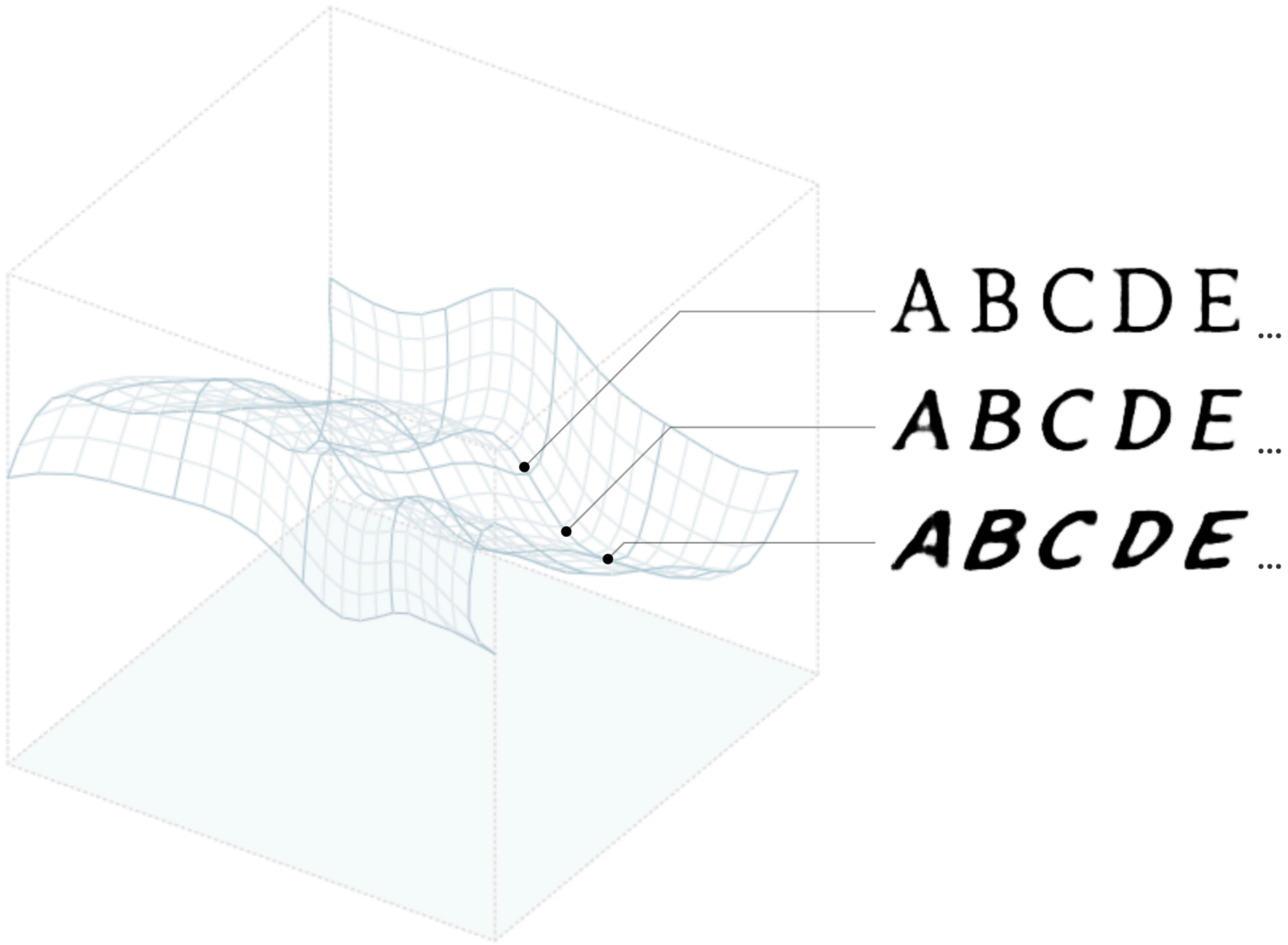


Background: Variational Auto-encoder (VAE)



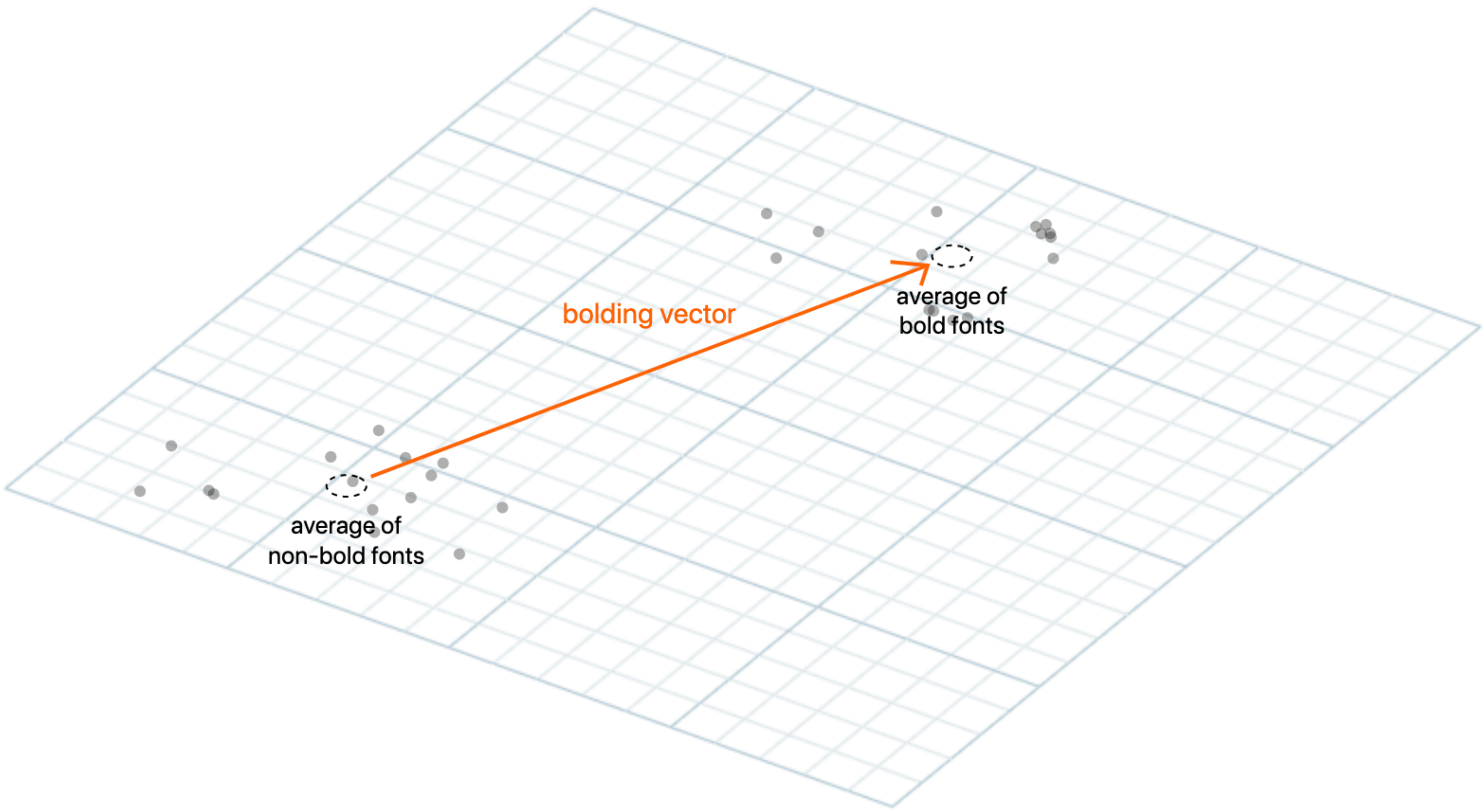
Loss = Reconstruction loss + KL divergence

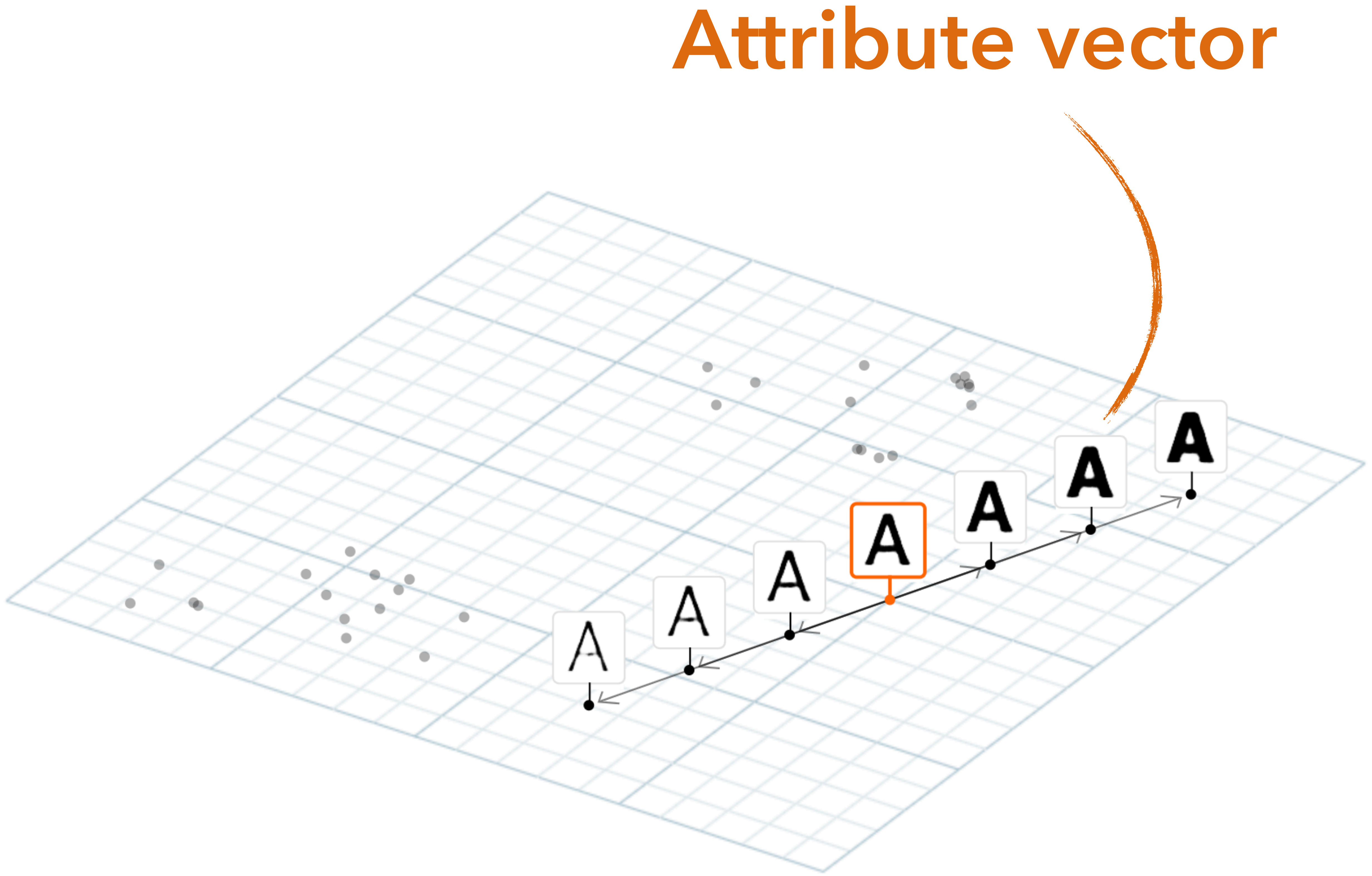
Latent Space in VAE



CARTER S., NIELSEN M.: Using artificial intelligence to augment human intelligence. Distill (2017). https://distill.pub/2017/aia.

Latent Space in VAE

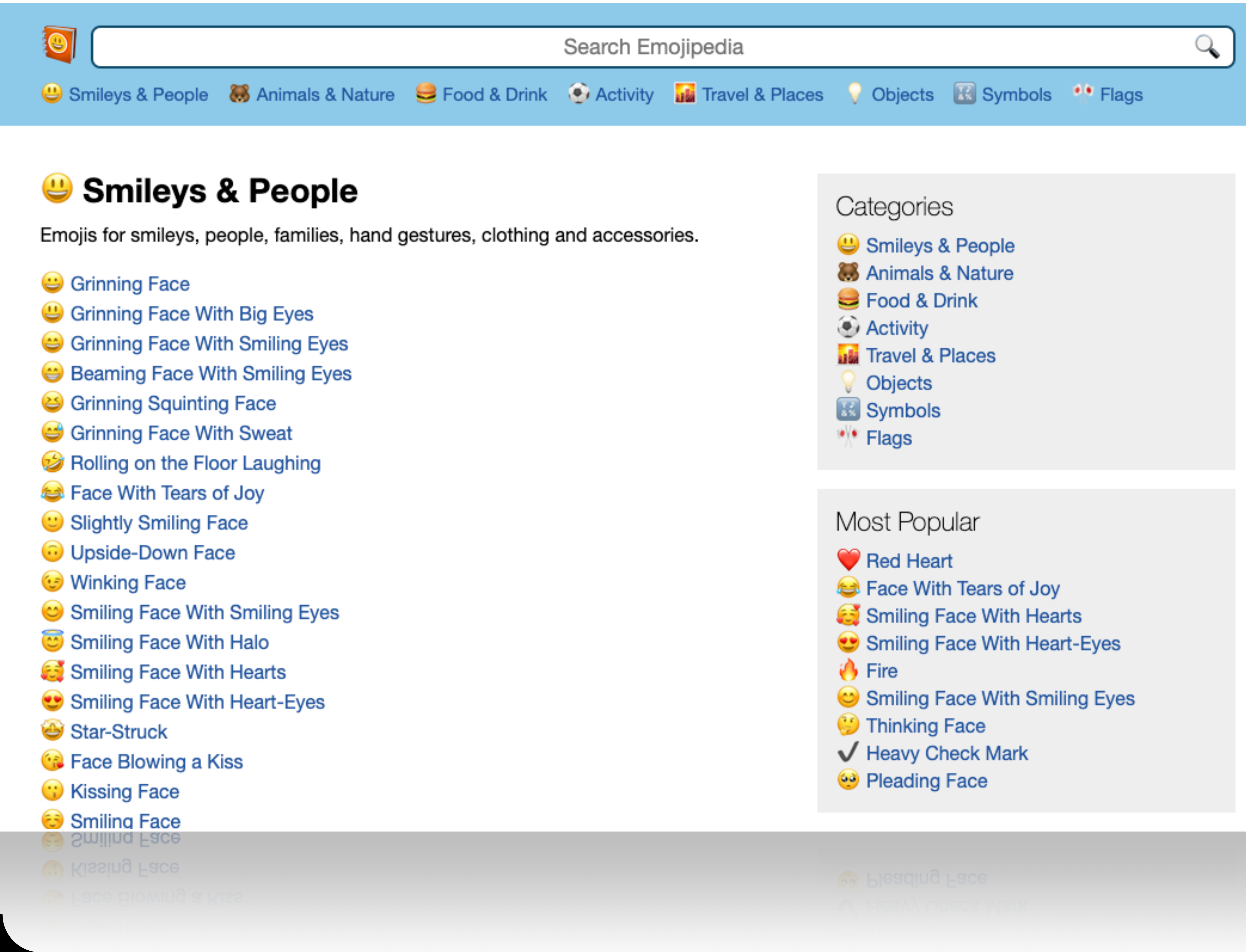




CARTER S., NIELSEN M.: Using artificial intelligence to augment human intelligence. Distill (2017). https://distill.pub/2017/aia.

VAEs might help me understand emojis!

Latent spaces might help me understand emojis!

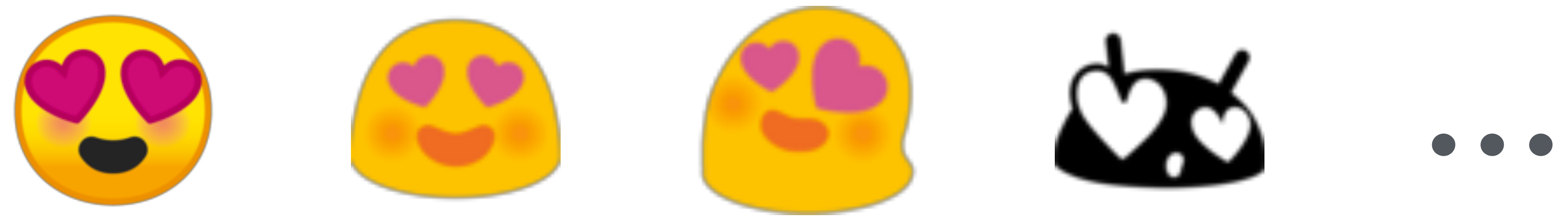


Crawl ~24,000 emojis

Platforms

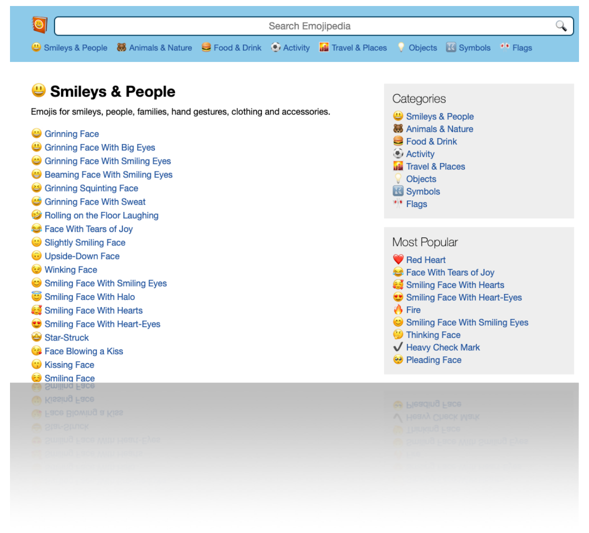
Apple Google Twitter LG Samsung

Versions

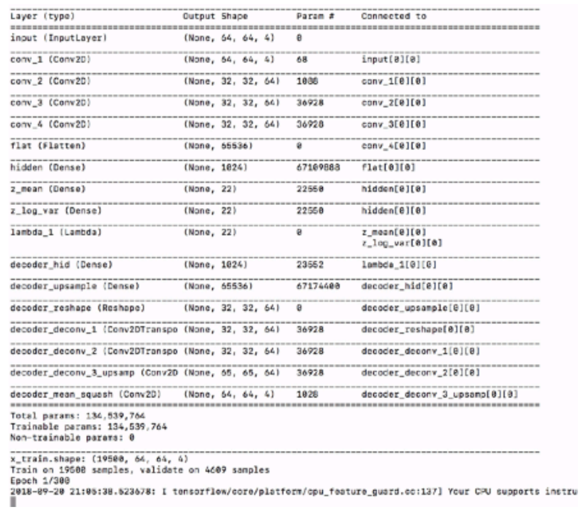


9.0 7.0 4.4 4.3

Latent spaces might help me understand emojis!

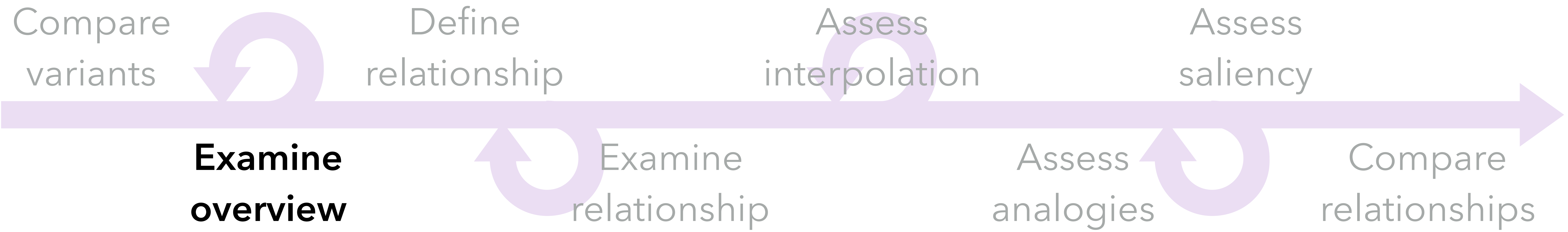


Crawl ~24,000 emojis



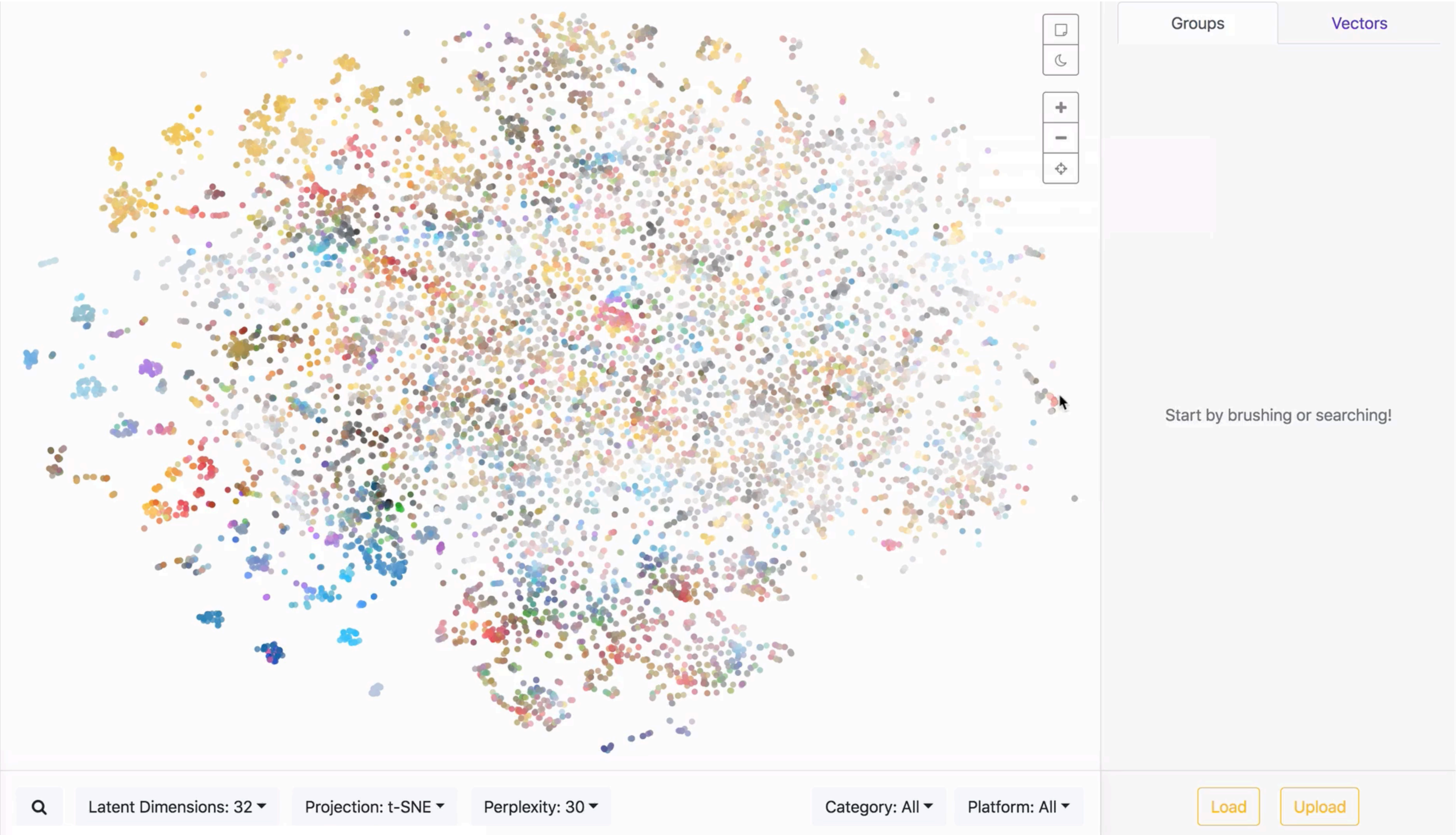
Train 6 variational auto-encoders (VAEs) with varying latent dimensions

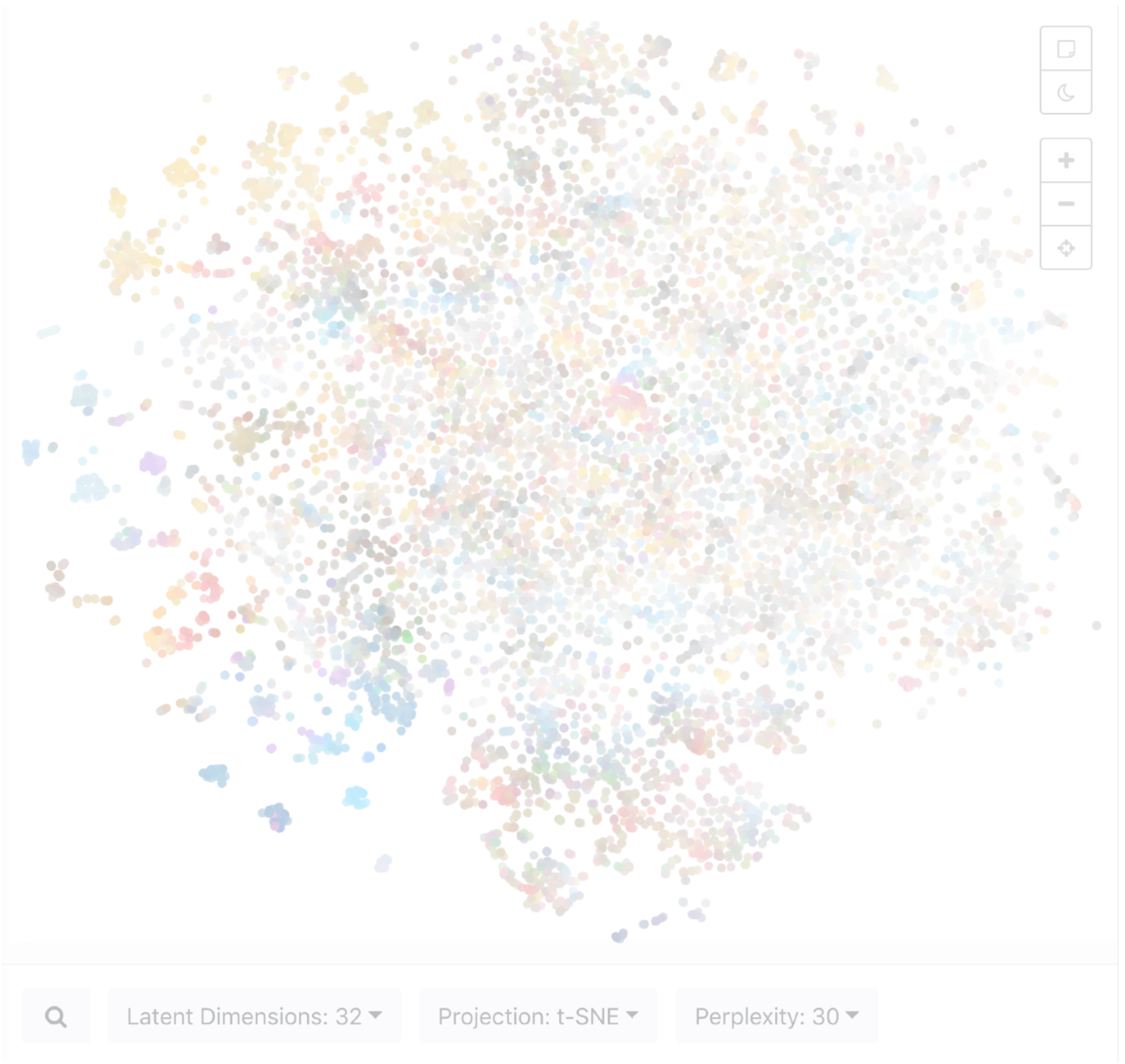
Load latent spaces into LSC!



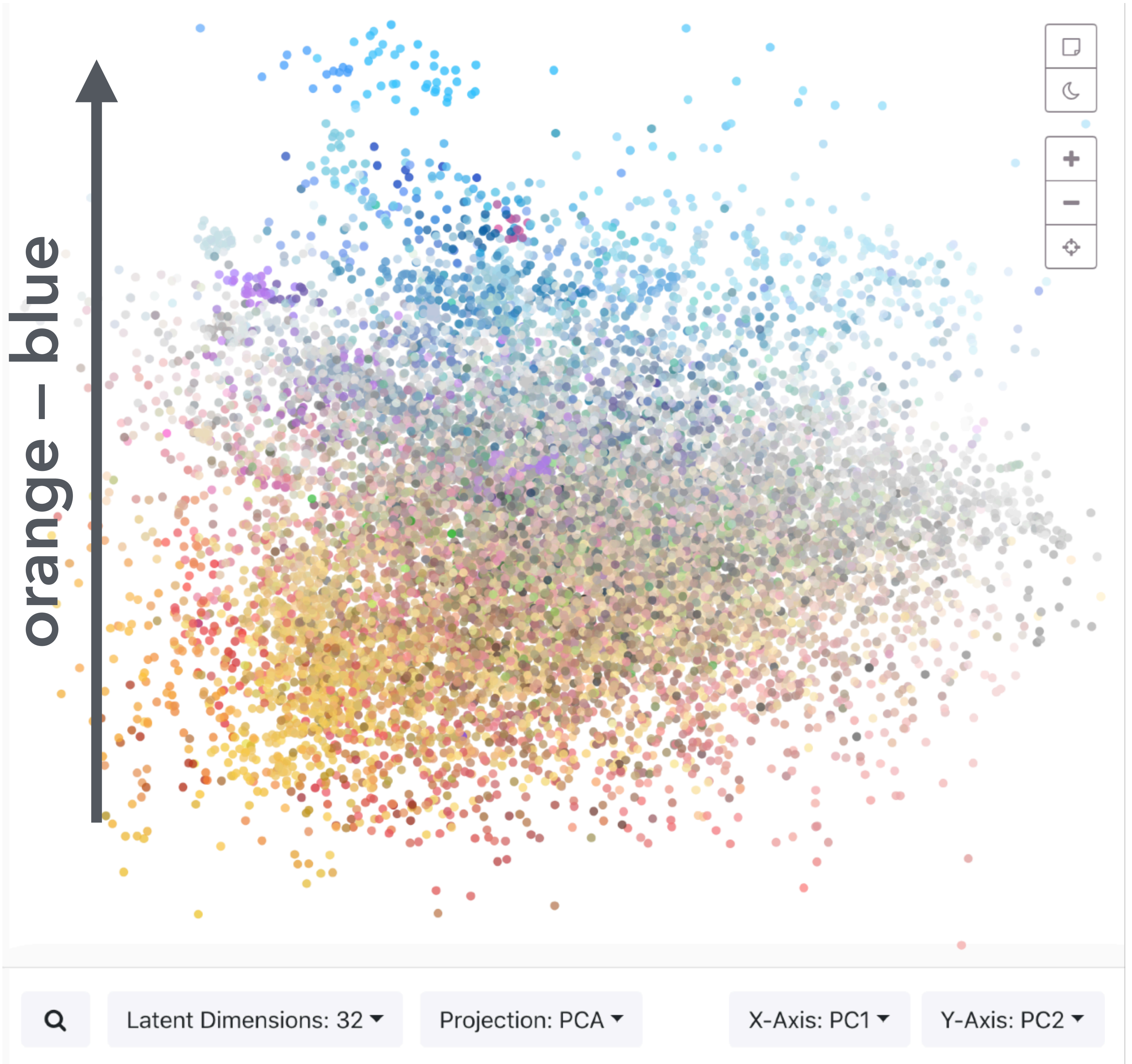
I would like to gain initial familiarity with the latent space …

Examining an overview distribution

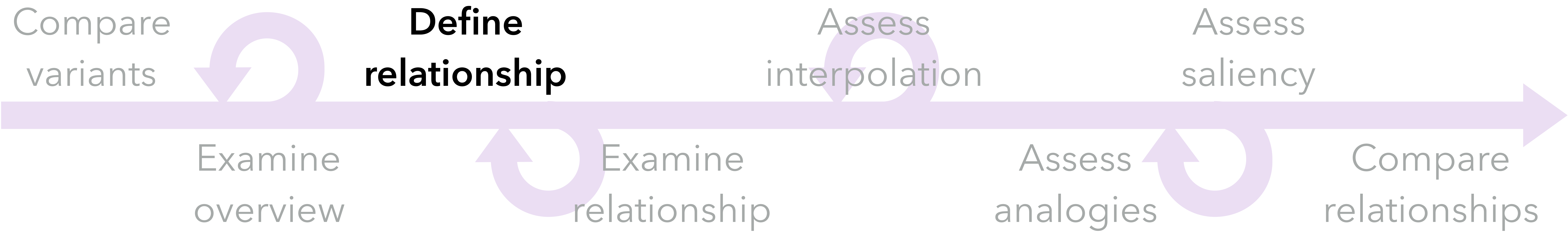



Latent Dimensions: 32
Projection: t-SNE
Perplexity: 30

t-SNE

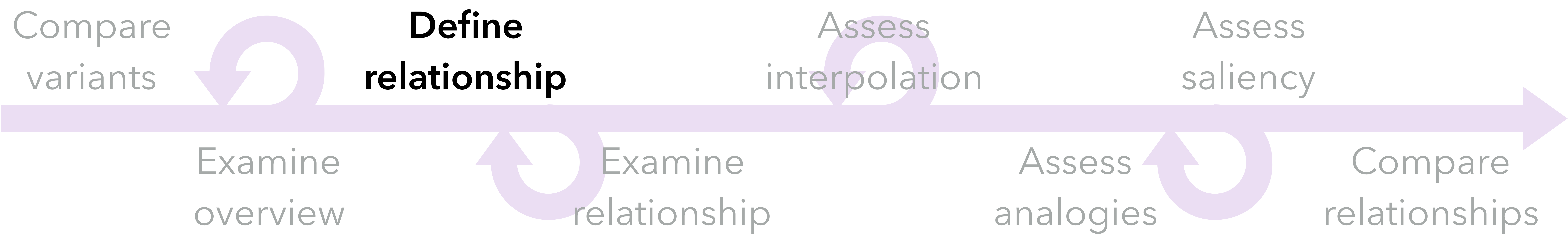

orange – blue
Latent Dimensions: 32
Projection: PCA
X-Axis: PC1
Y-Axis: PC2

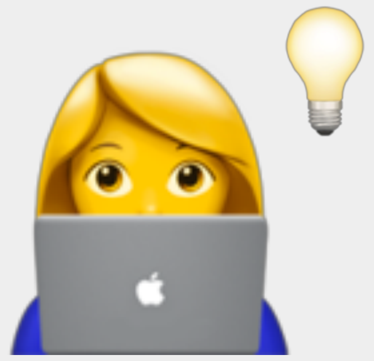
PCA



Android version 9 adopts a distinct style compared to its earlier versions

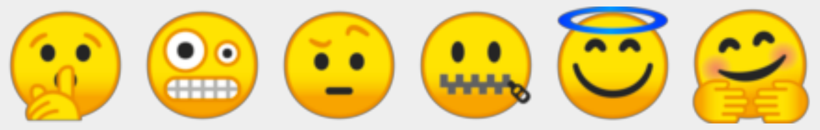
Android 7:
Android 9:

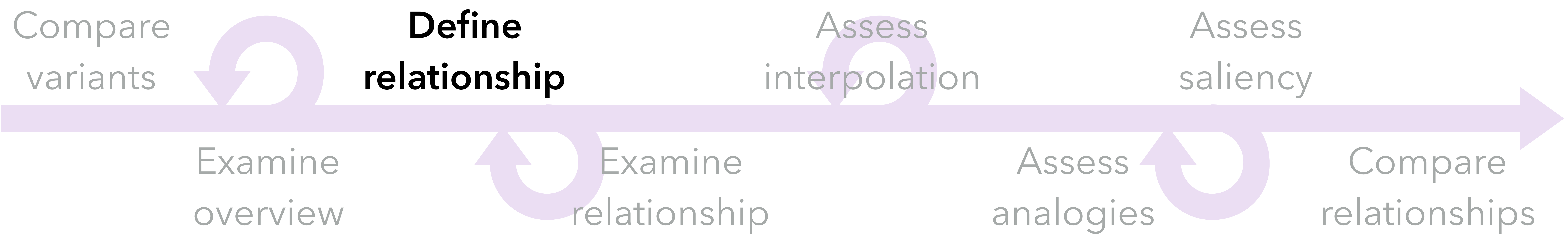




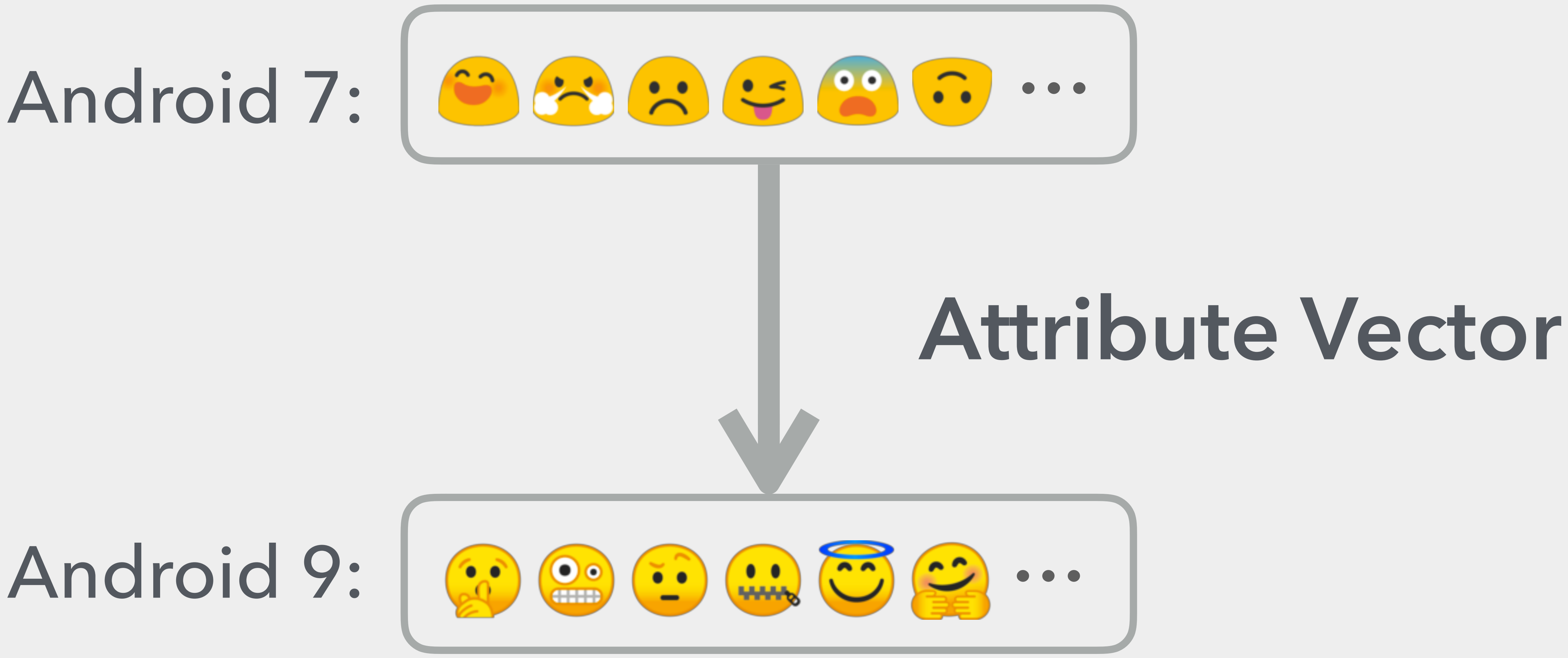
Does the latent space capture this trend?

Android 7:  ...

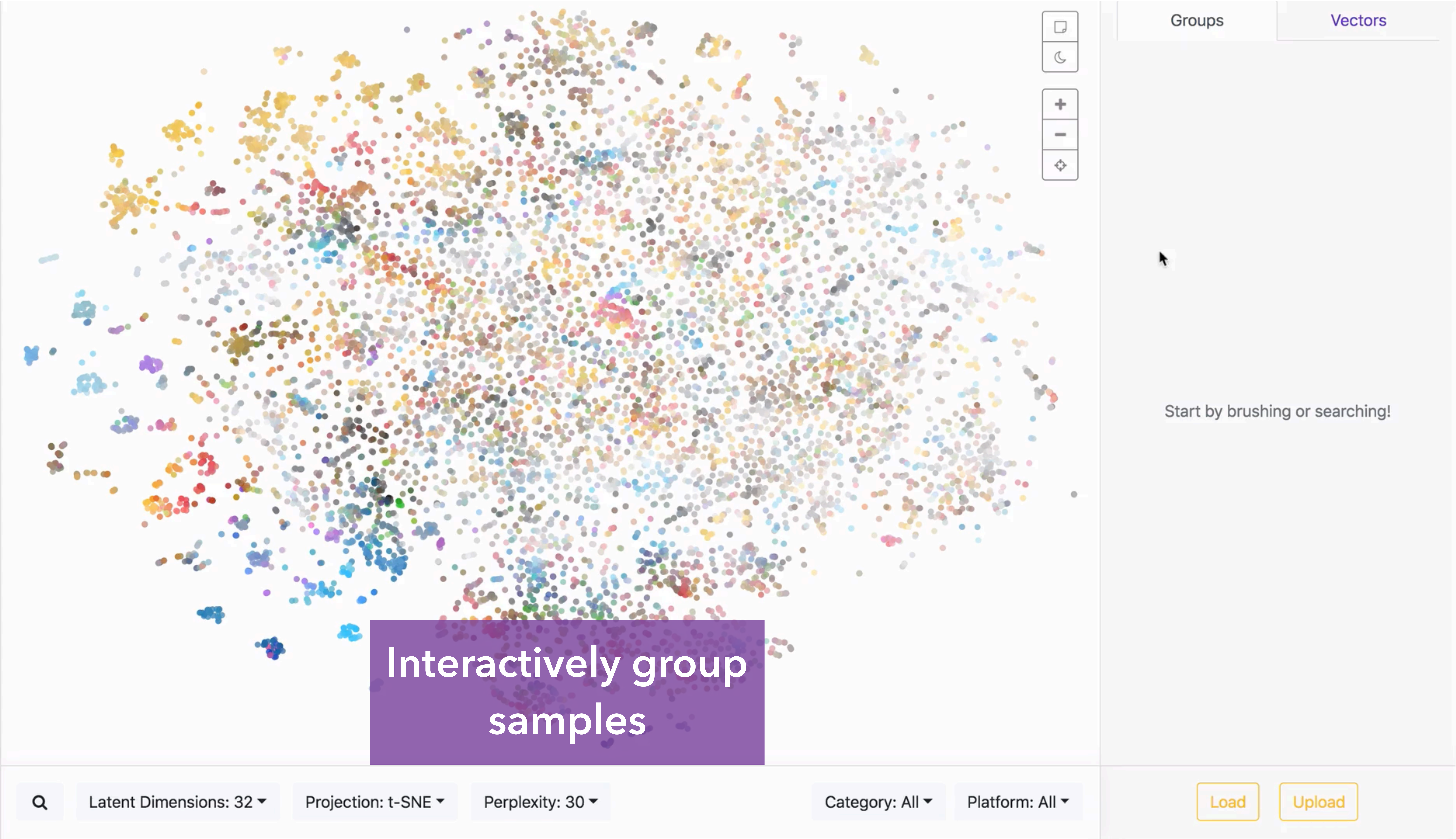
Android 9:  ...



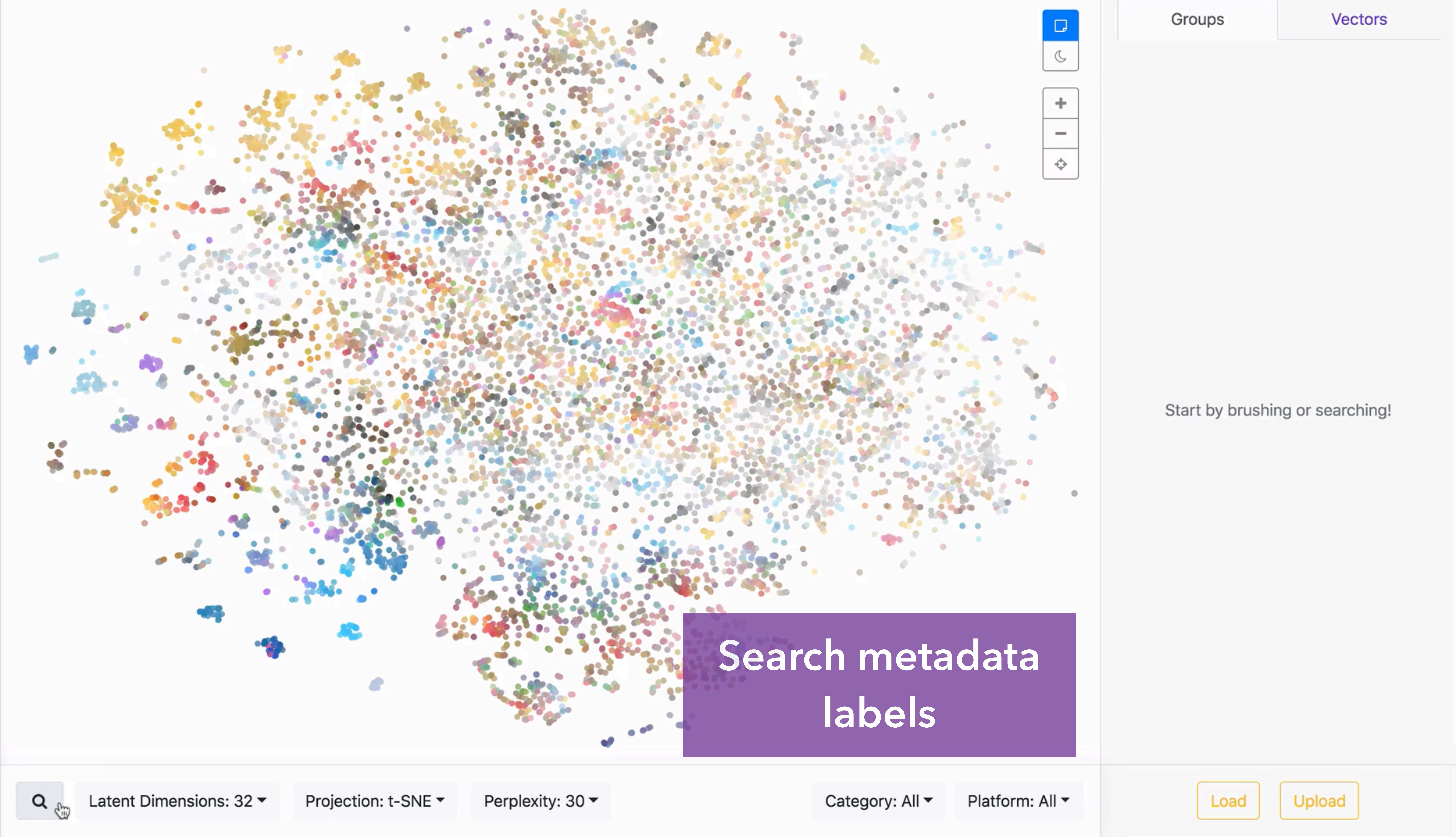
Does the latent space capture this trend?



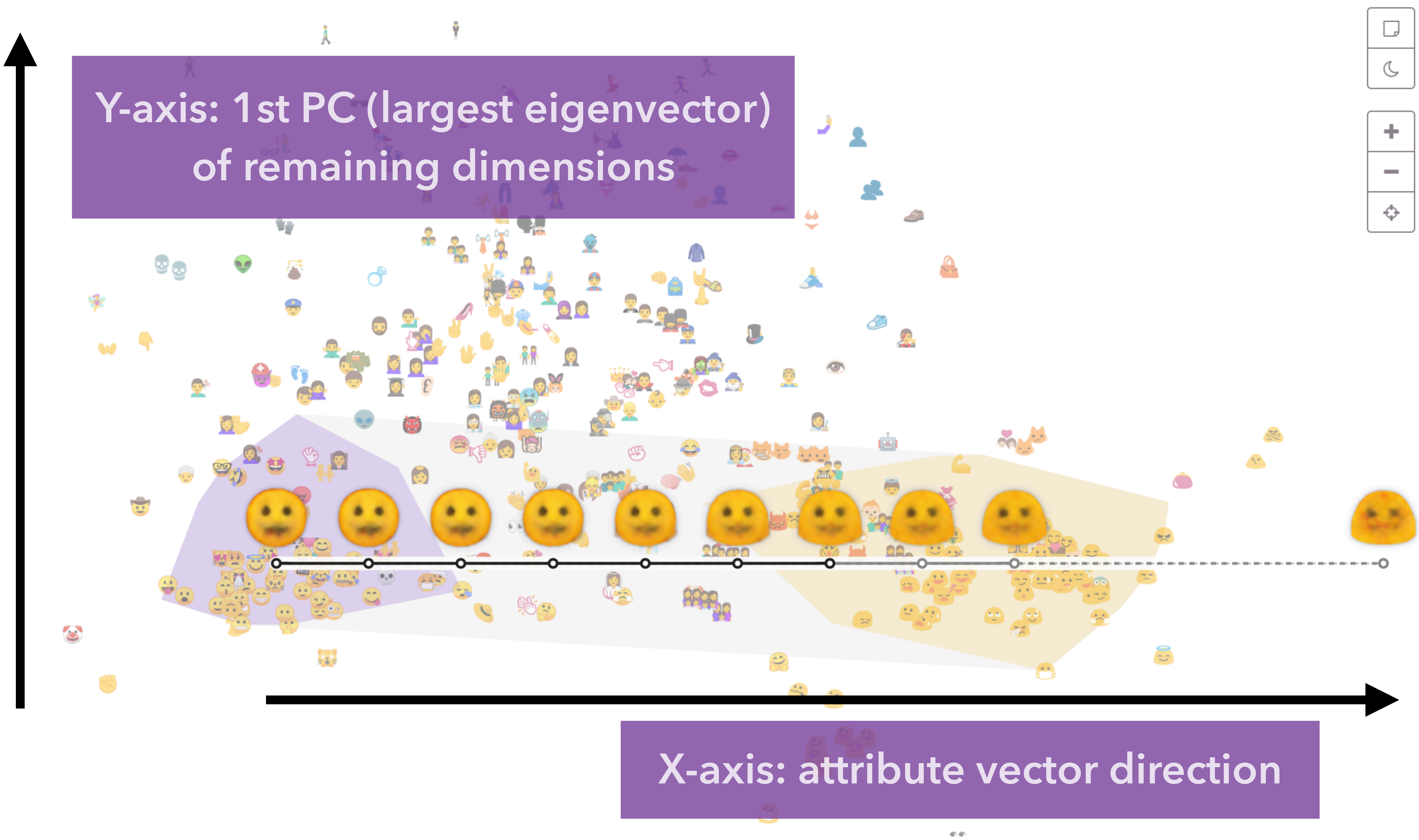
Defining an attribute vector

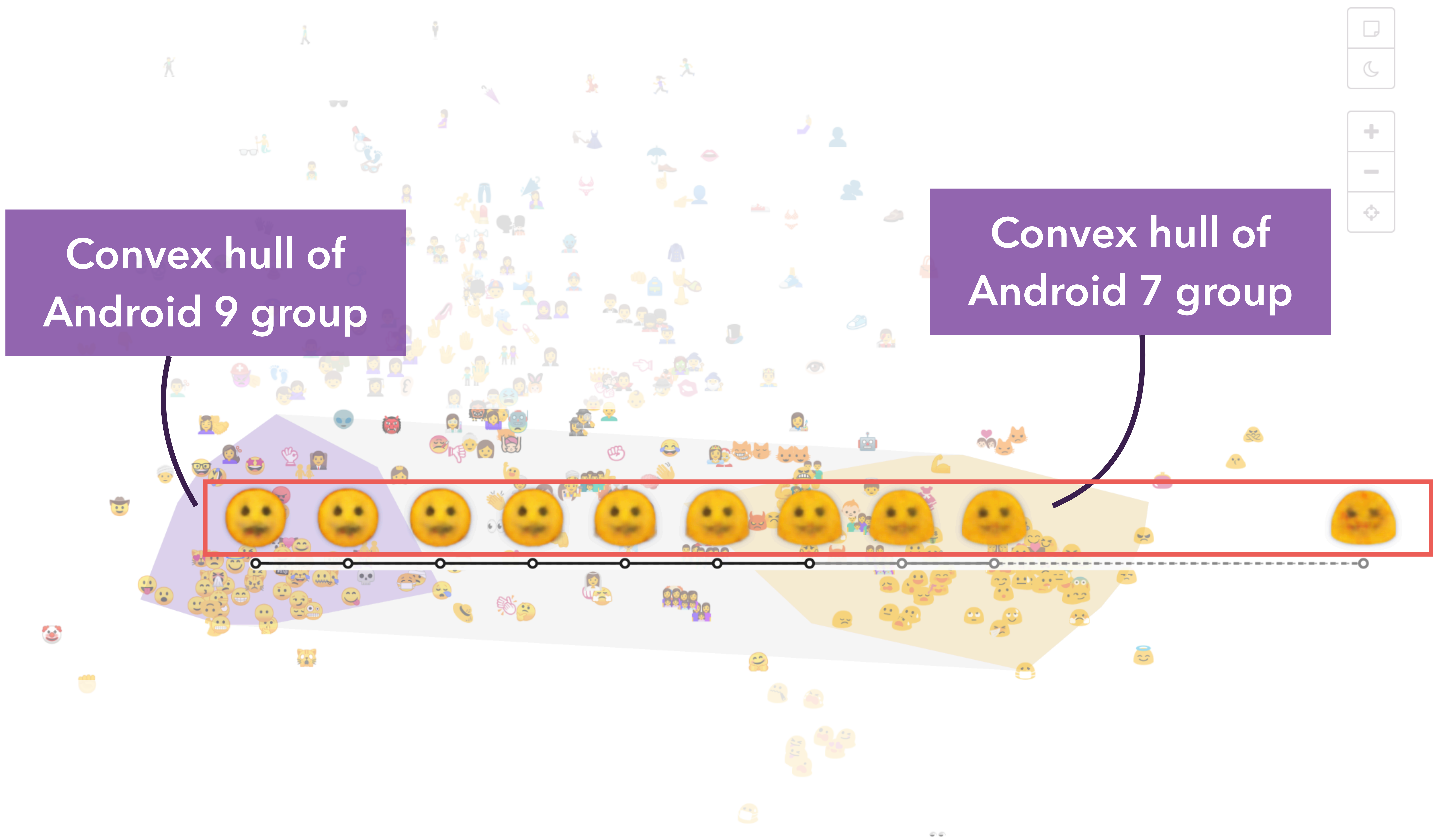


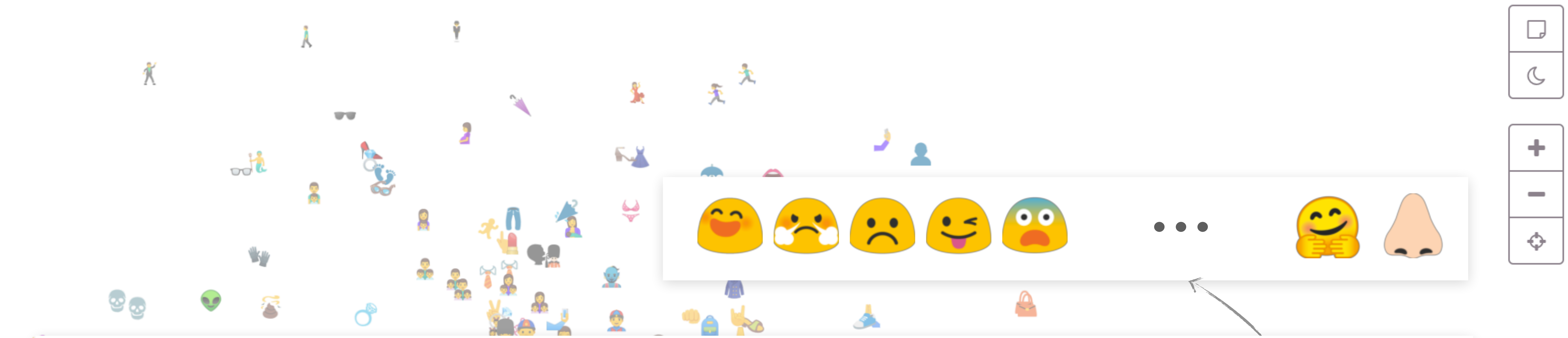
Defining an attribute vector



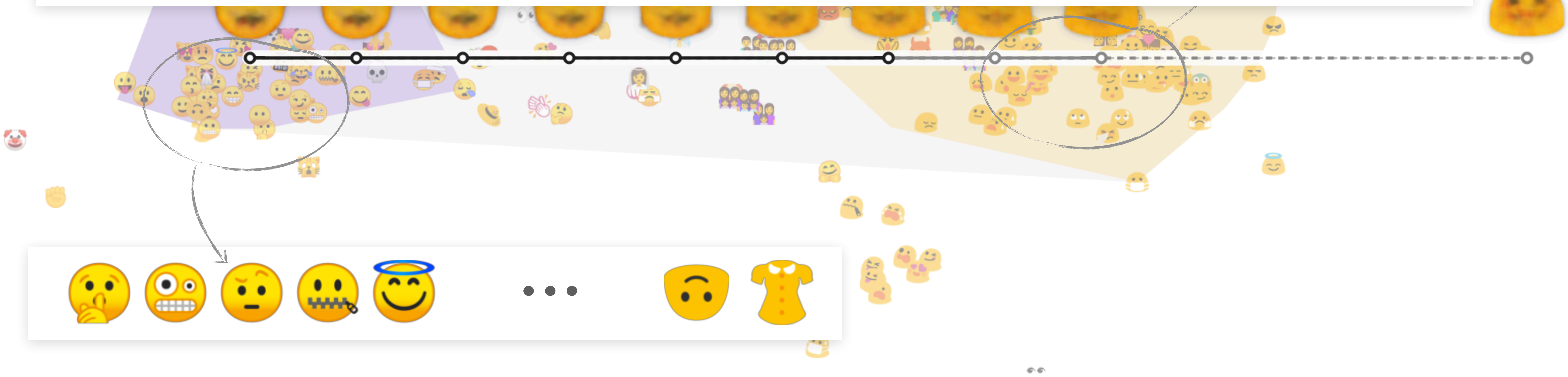
How do emojis fall along the attribute vector spectrum?

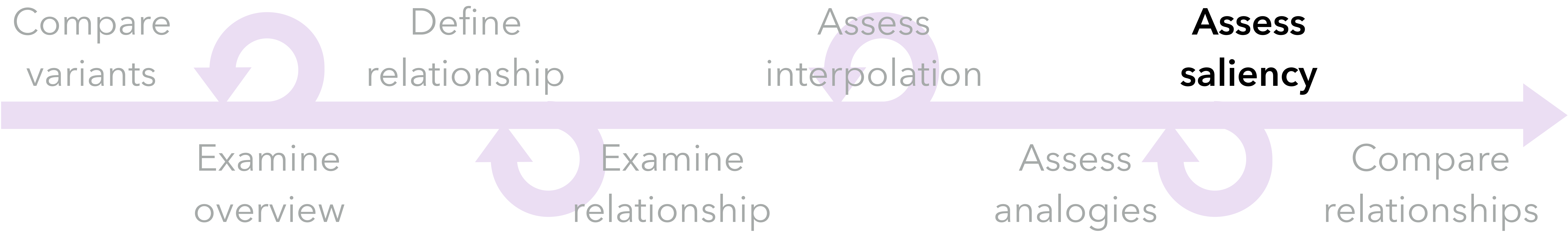

Y-axis: 1st PC (largest eigenvector) of remaining dimensions
X-axis: attribute vector direction


Convex hull of Android 9 group
Convex hull of Android 7 group

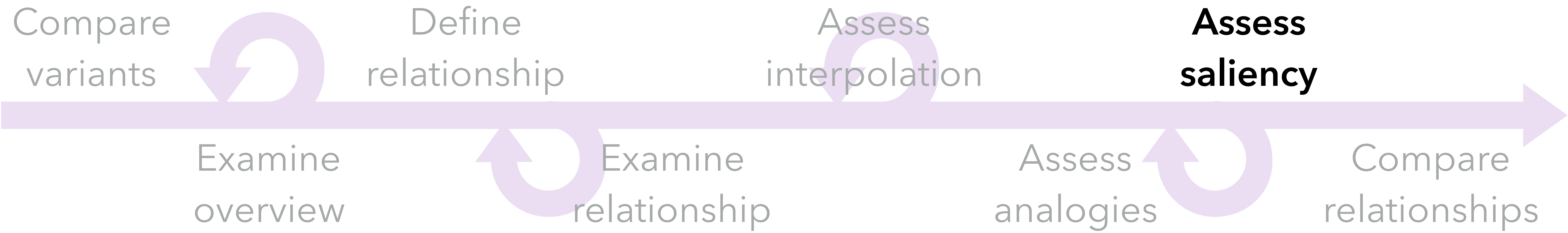


What emojis are considered by the model to be more similar to Android 9 than Android 7?

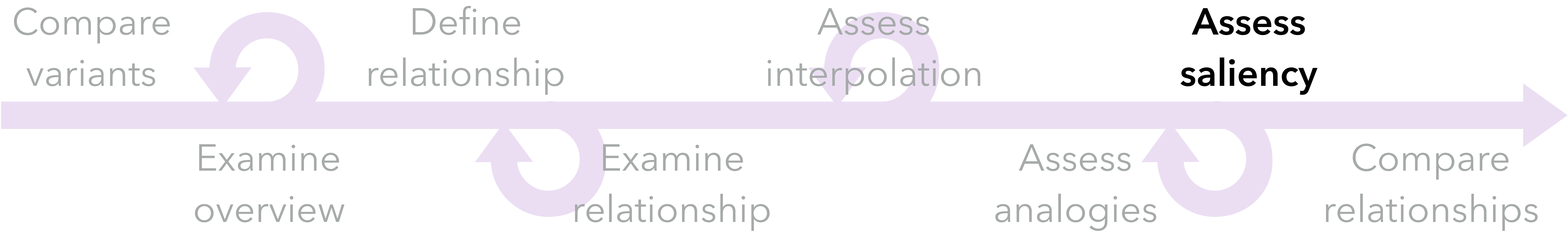




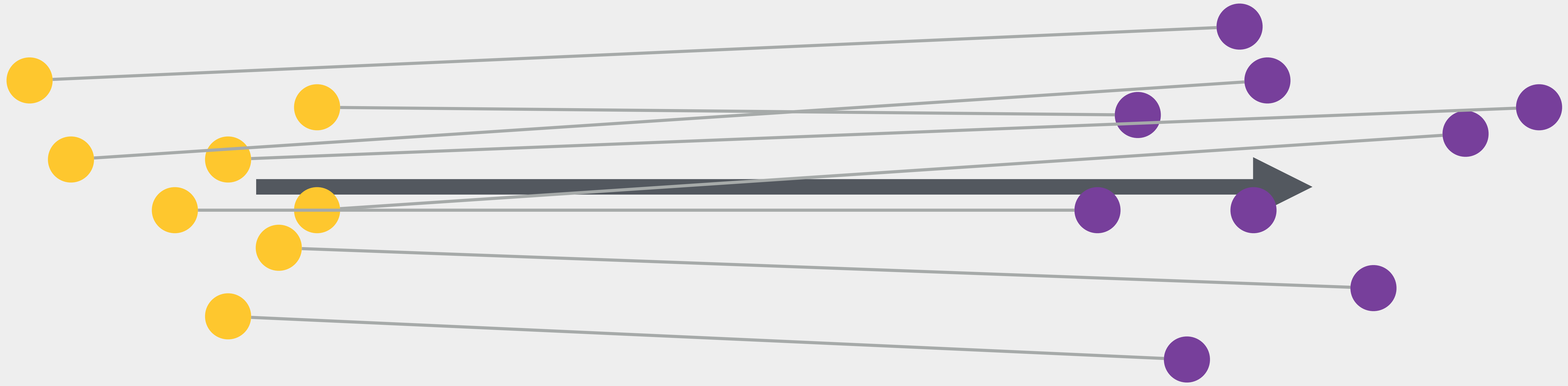
Does the attribute vector reliably represent a salient relationship?

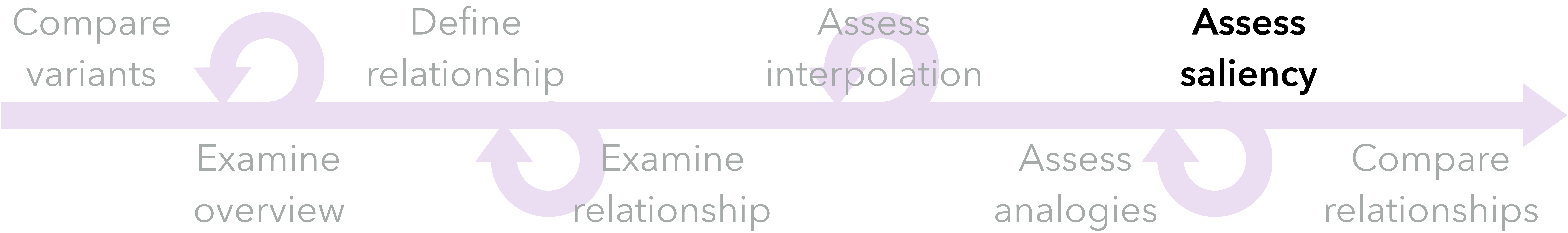


Does the attribute vector reliably represent a salient relationship?

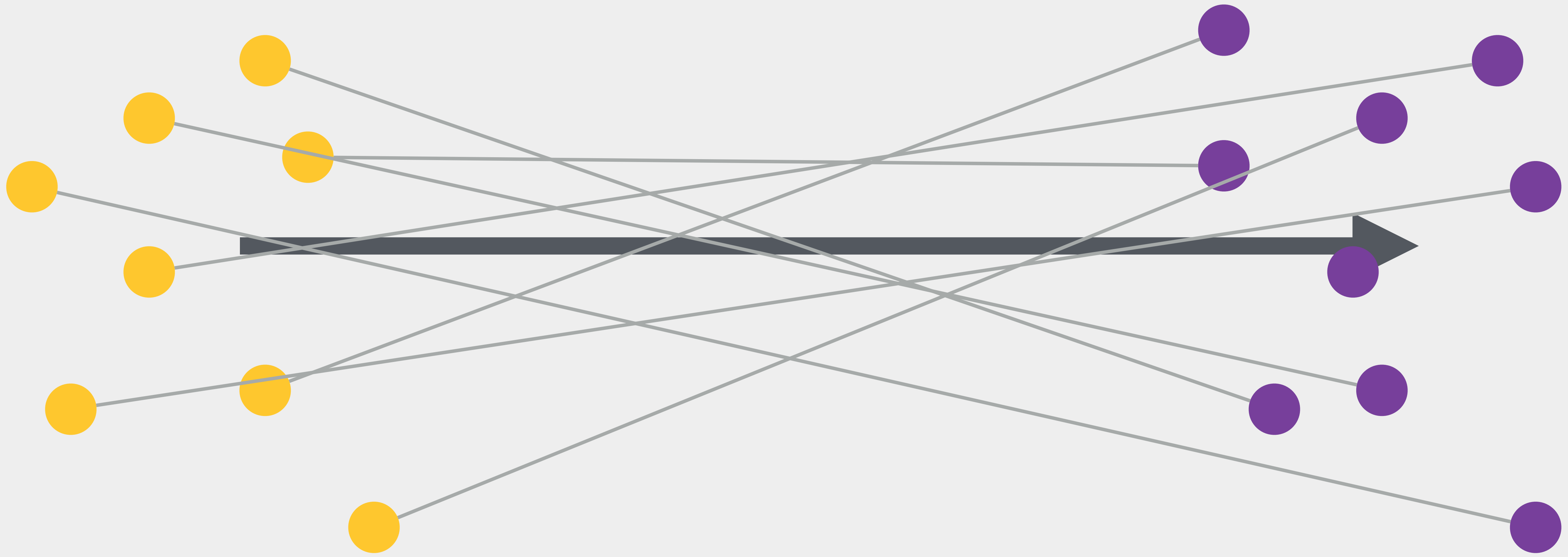


Does the attribute vector reliably represent a salient relationship?

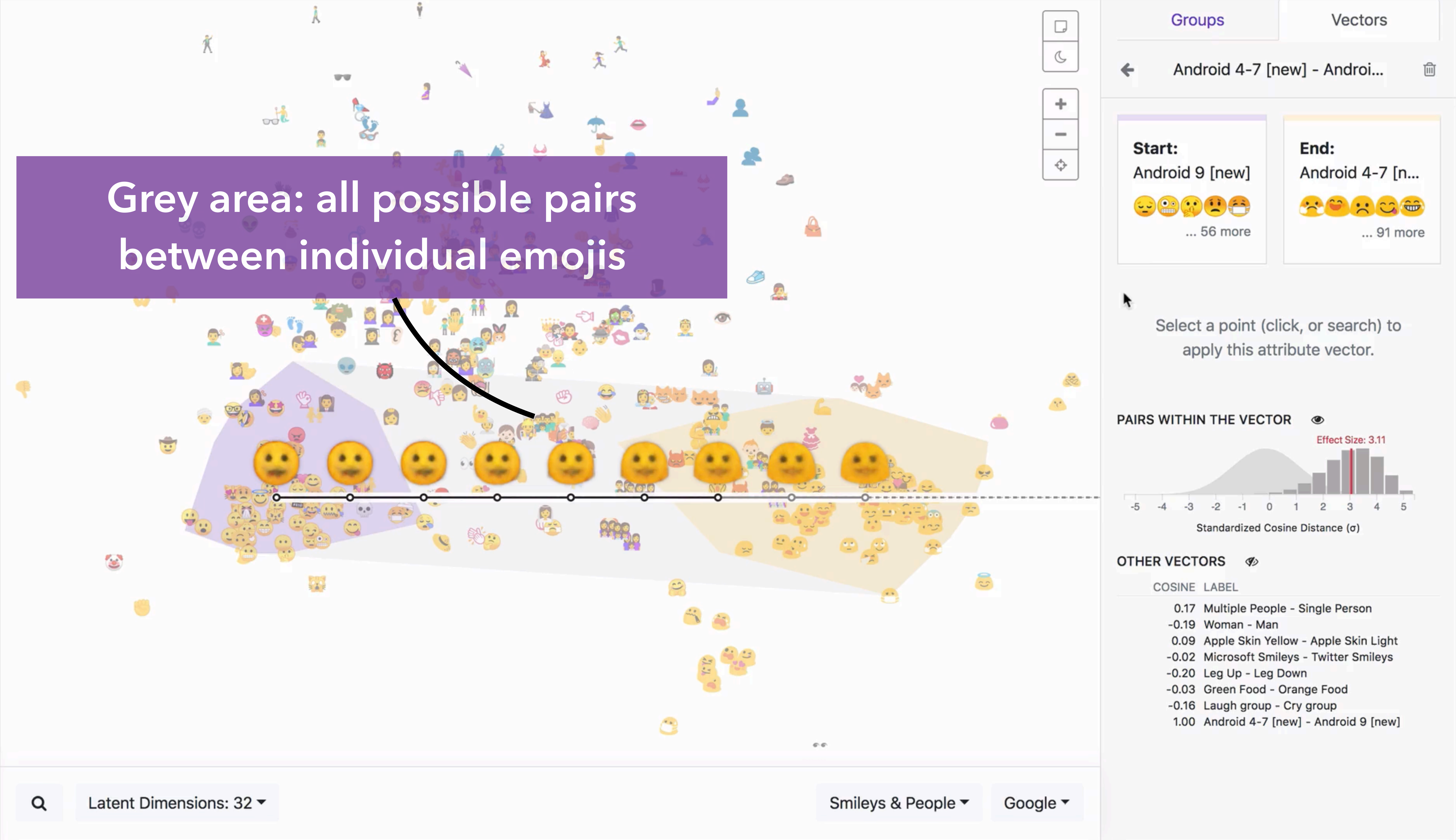




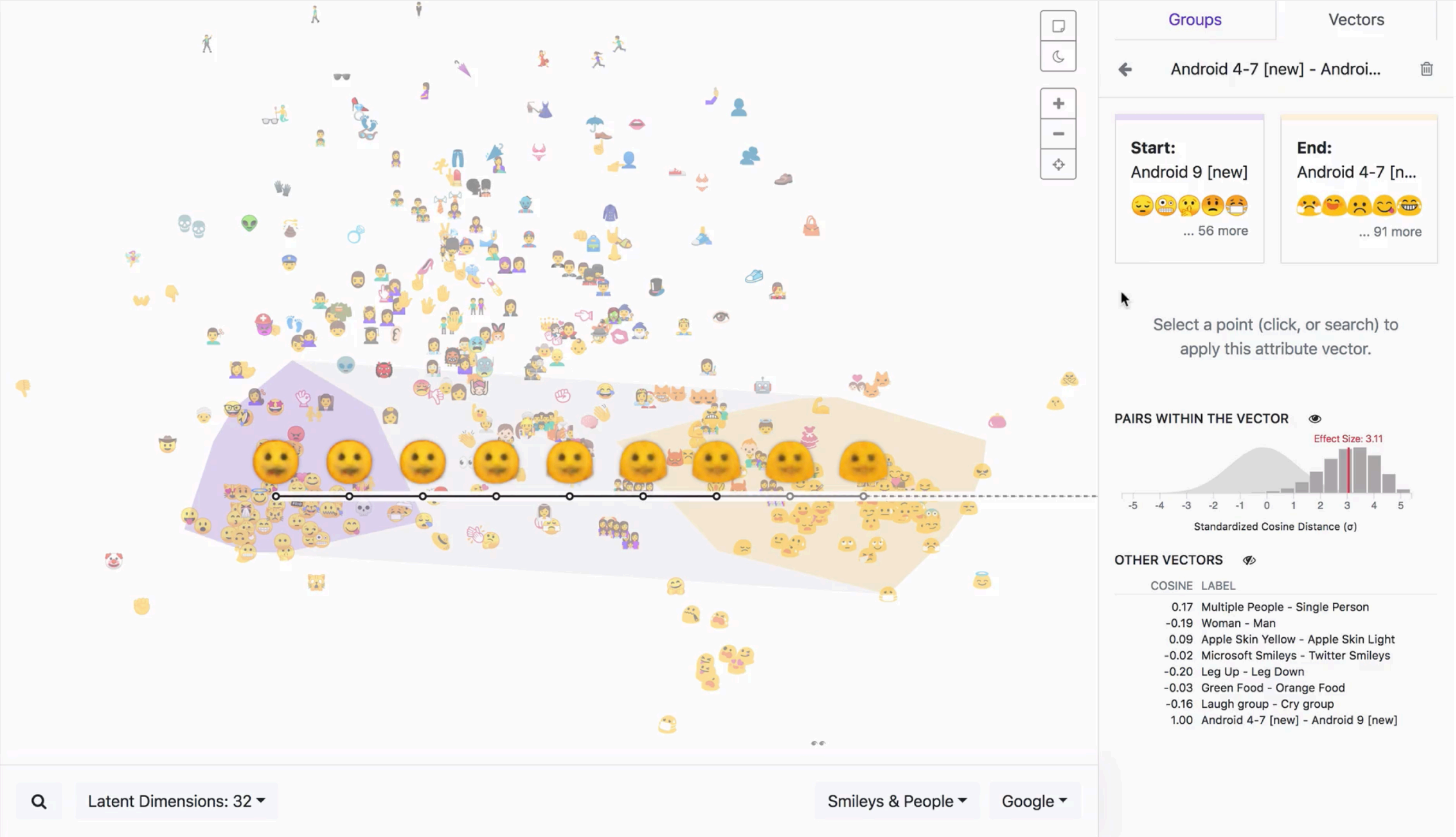
Does the attribute vector reliably represent a salient relationship?



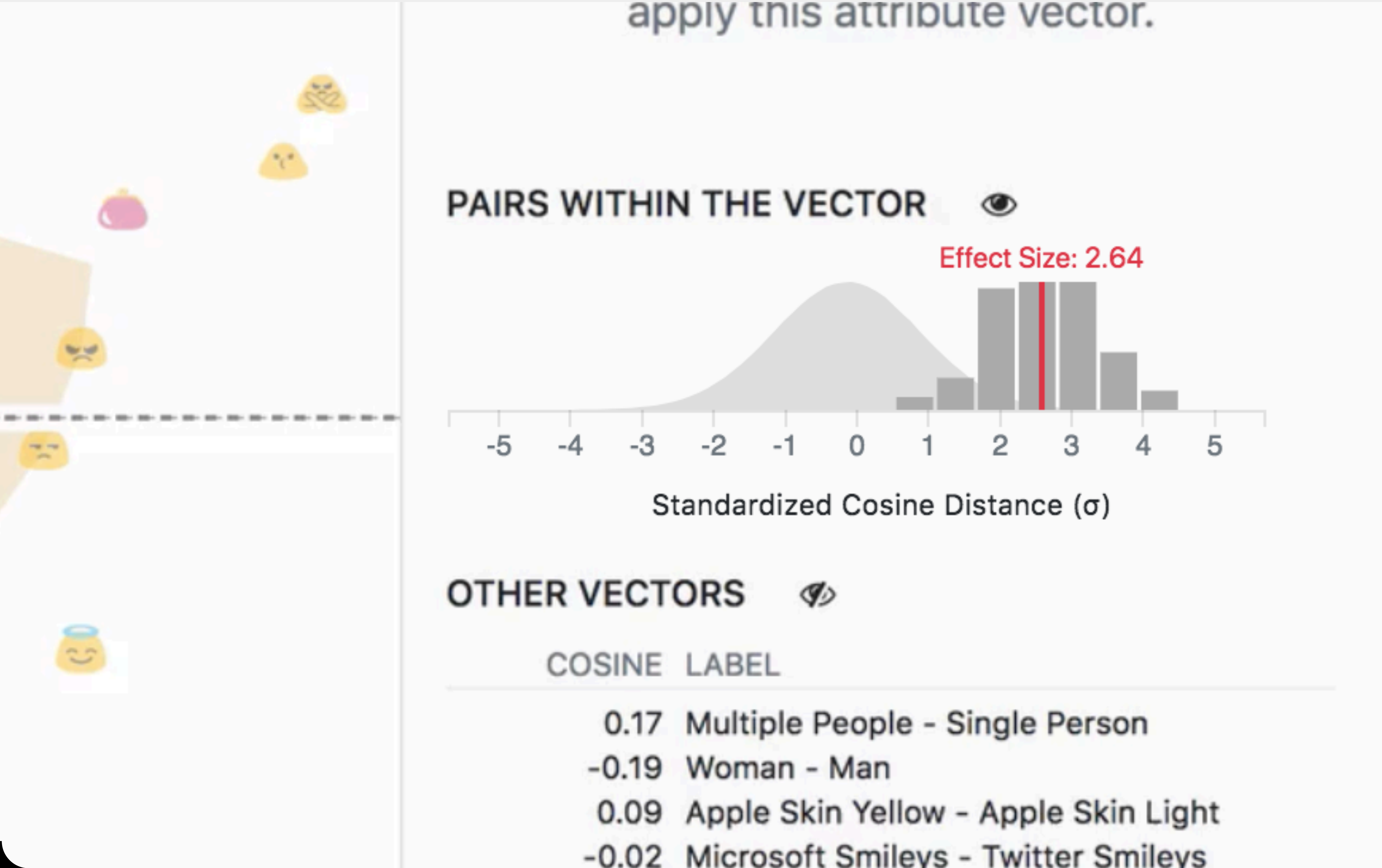
Assessing attribute vector saliency



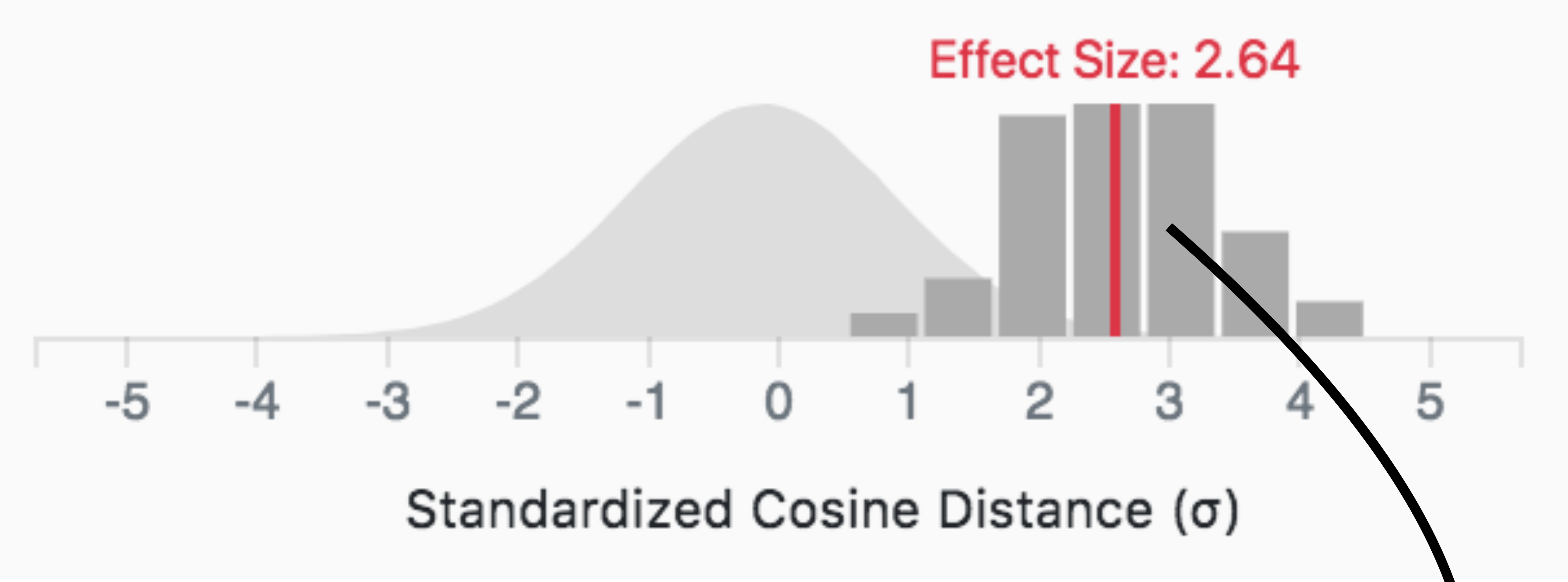
Assessing attribute vector saliency



Assessing attribute vector saliency



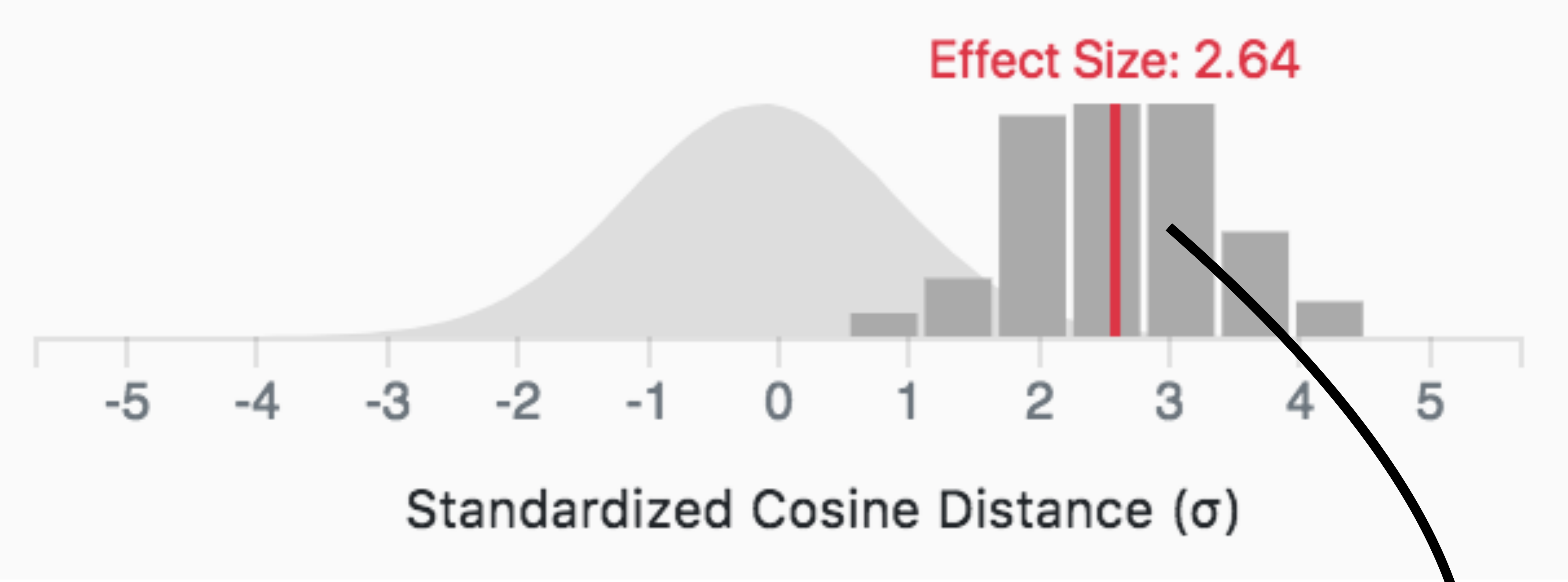
Pair alignment in the original space



$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

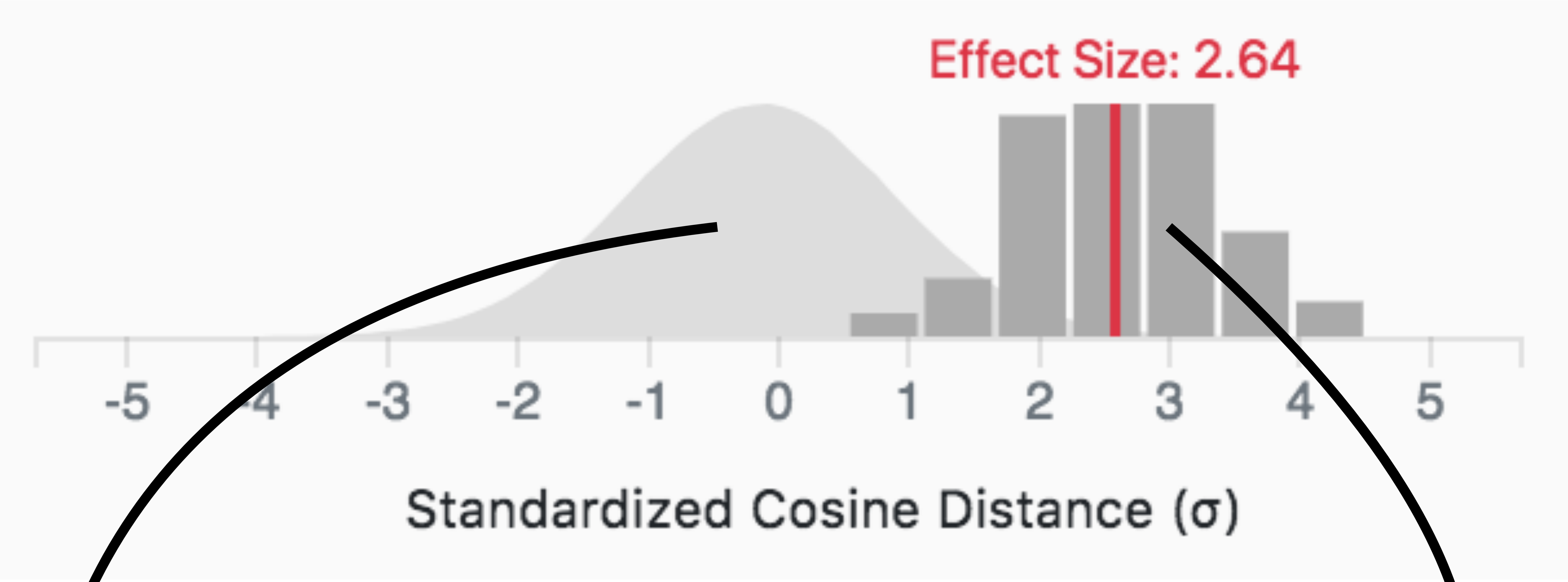
Alignment of individual pairs in the attribute vector

Pair alignment in the original space



😈 **Problem:** as dimensionality increases, random vectors are more likely to be orthogonal!

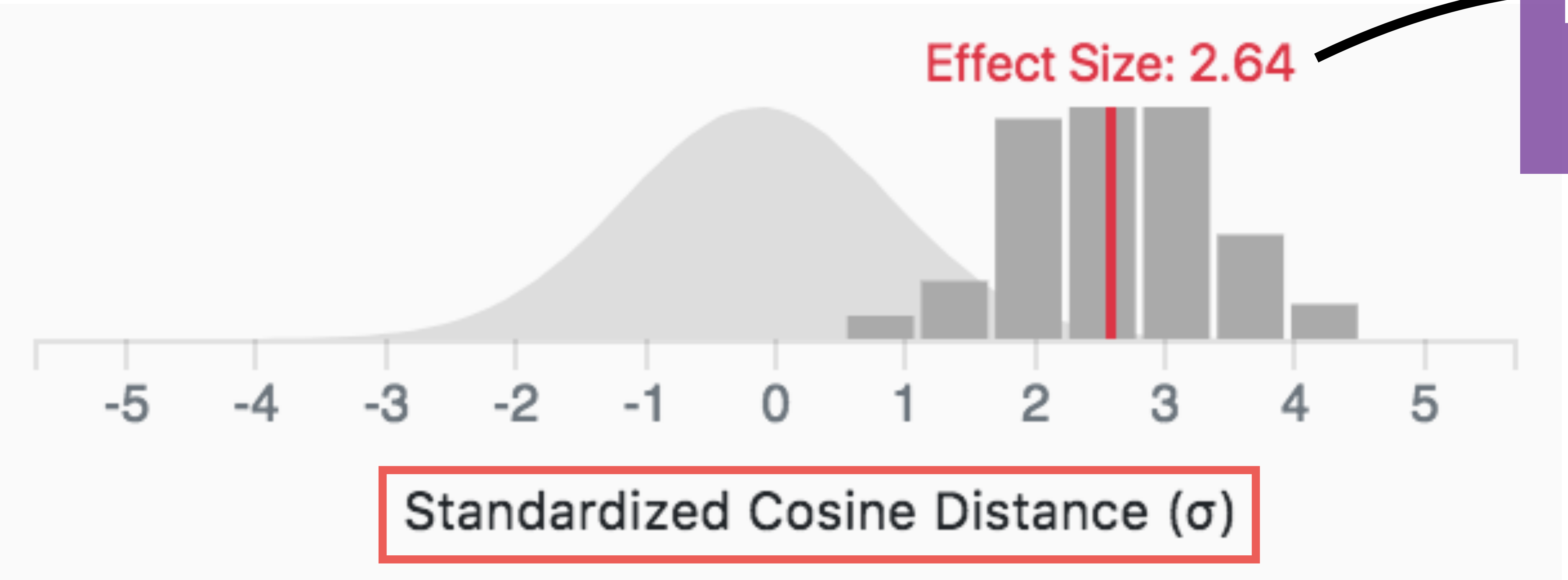
Pair alignment in the original space



Alignment of random pairs

Alignment of individual pairs in the attribute vector

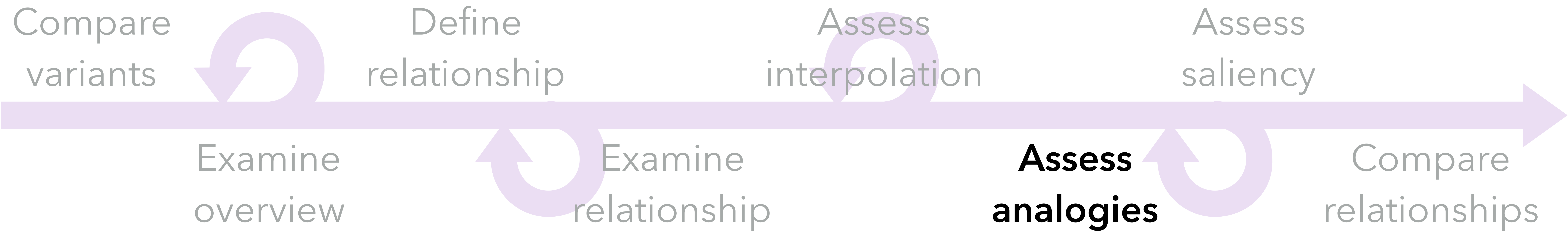
Pair alignment in the original space



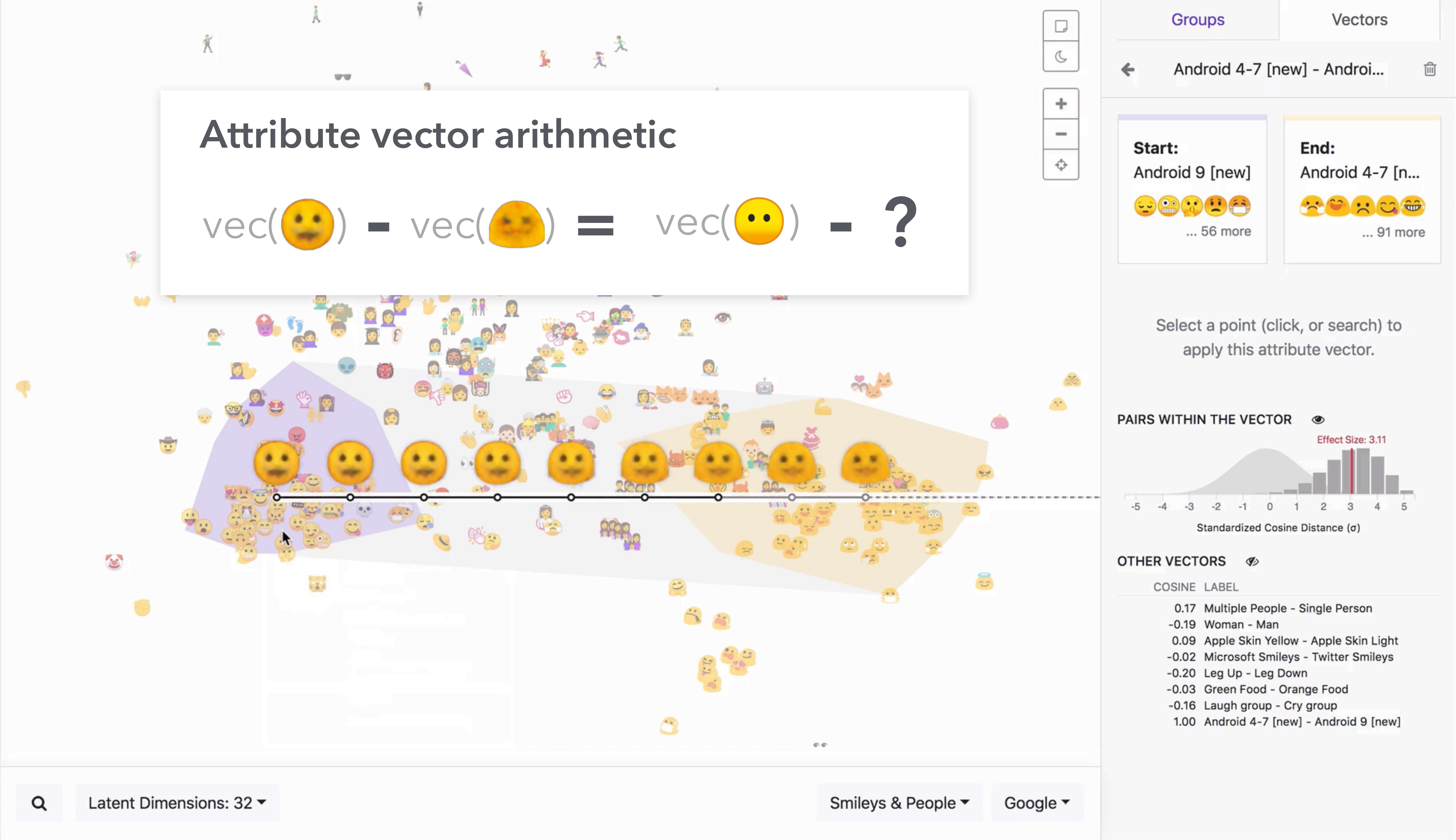
Difficult to observe by chance

Cosine similarity divided by pooled standard deviation

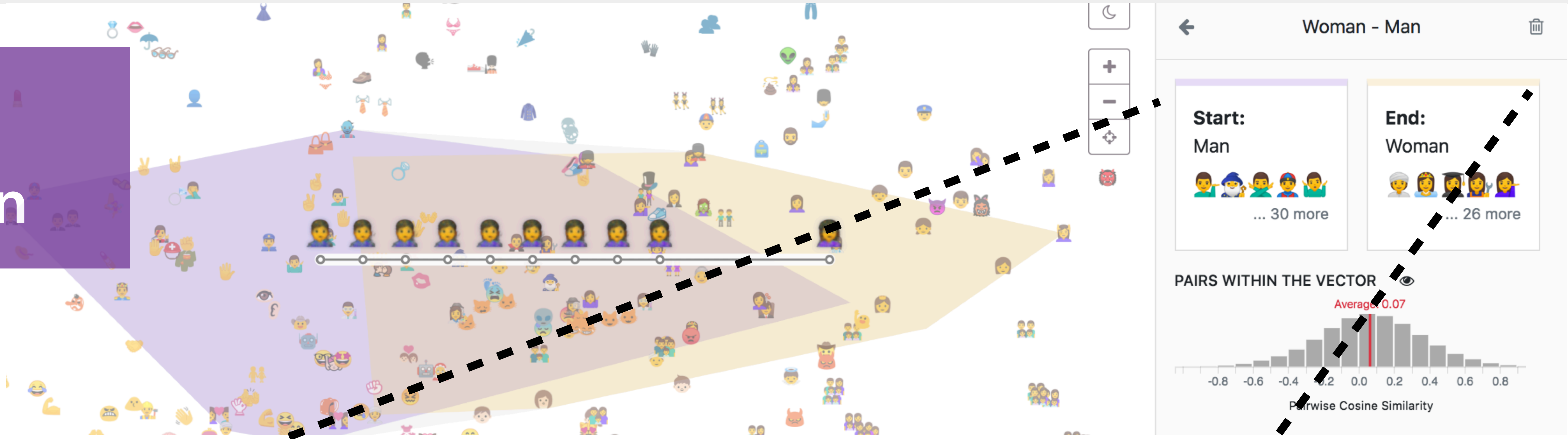
$$Sp = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{(n_1 + n_2 - 2)}}$$

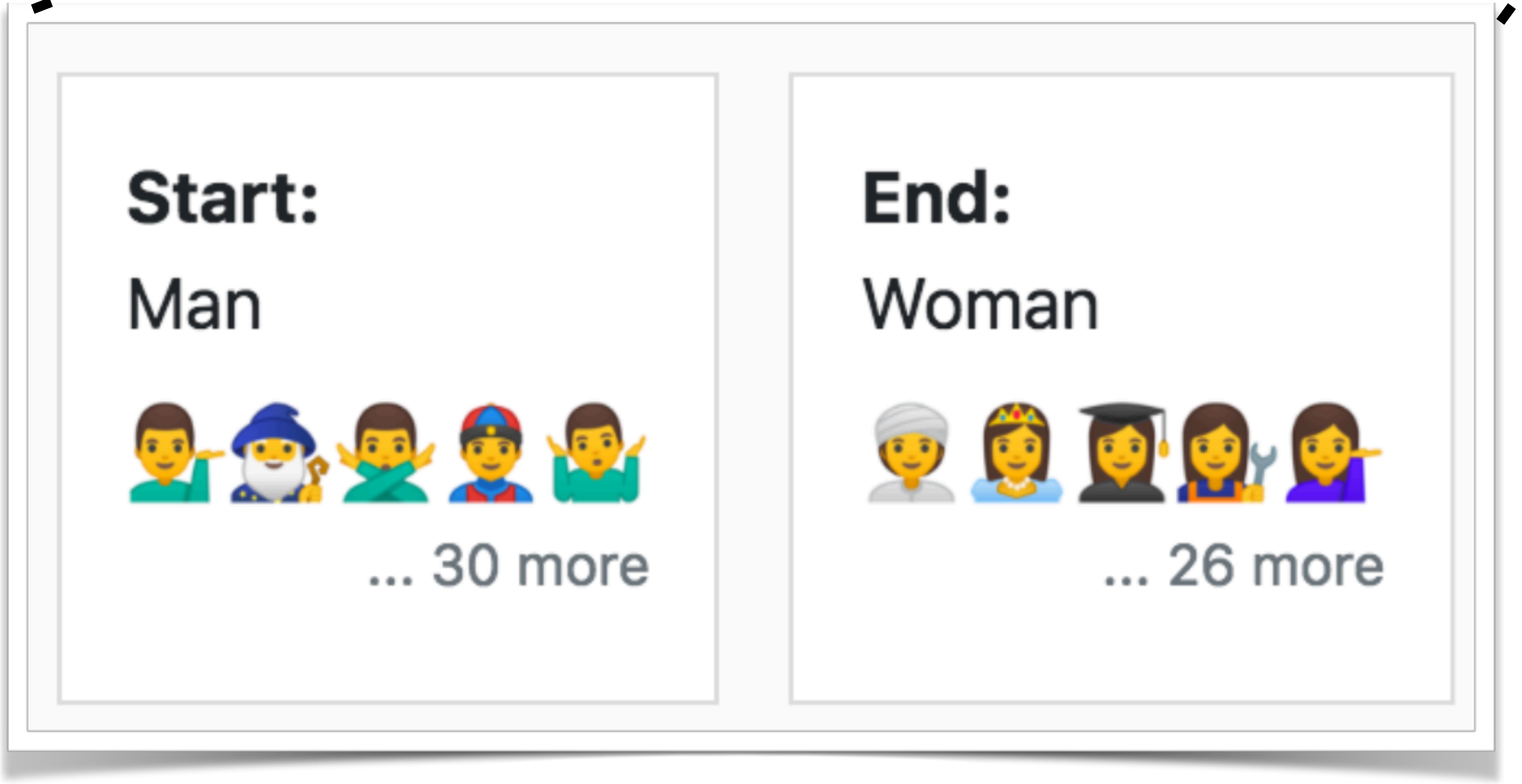


Might our attribute vector transform an arbitrary emoji into Android 7 style?


Attribute vector arithmetic
vec( ) - vec( ) = vec( ) - ?
Groups
Vectors
Android 4-7 [new] - Androi...
Start:
Android 9 [new]
... 56 more
End:
Android 4-7 [n...
... 91 more
Select a point (click, or search) to apply this attribute vector.
PAIRS WITHIN THE VECTOR
Effect Size: 3.11
-5 -4 -3 -2 -1 0 1 2 3 4 5
Standardized Cosine Distance (σ)
OTHER VECTORS
COSINE LABEL
0.17 Multiple People - Single Person
-0.19 Woman - Man
0.09 Apple Skin Yellow - Apple Skin Light
-0.02 Microsoft Smileys - Twitter Smileys
-0.20 Leg Up - Leg Down
-0.03 Green Food - Orange Food
-0.16 Laugh group - Cry group
1.00 Android 4-7 [new] - Android 9 [new]
Latent Dimensions: 32
Smileys & People
Google

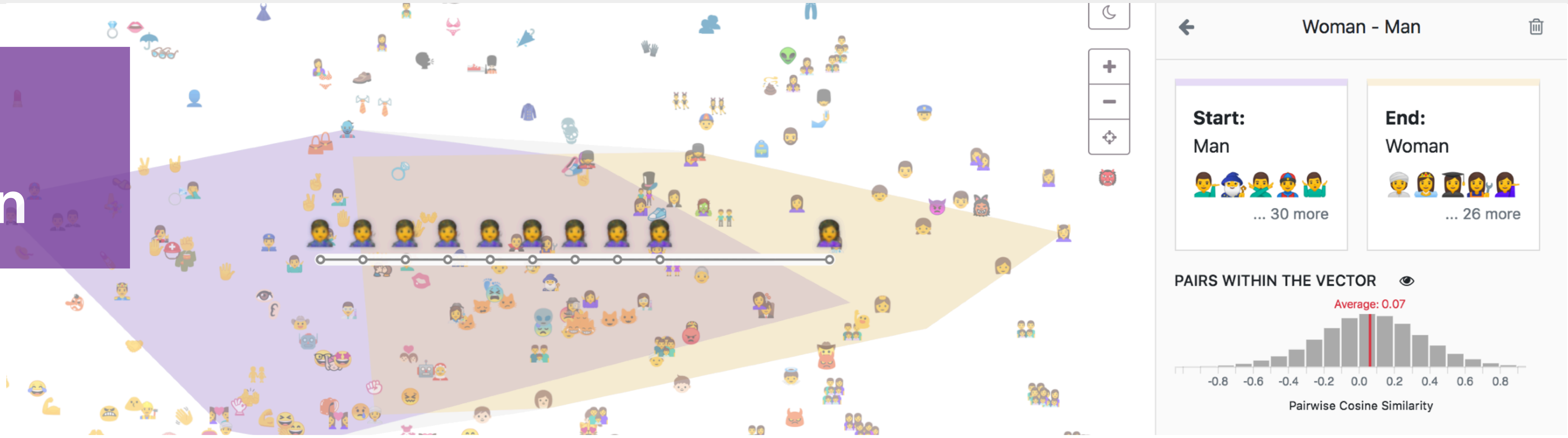
Example: man - woman





Performing attribute vector arithmetic

Example: man - woman



Attribute Vector

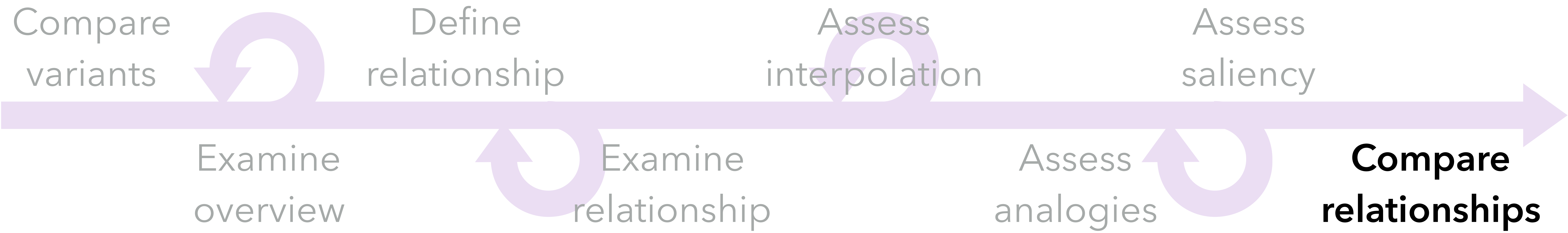
Analogy
Man

Analogy
Man Dancing

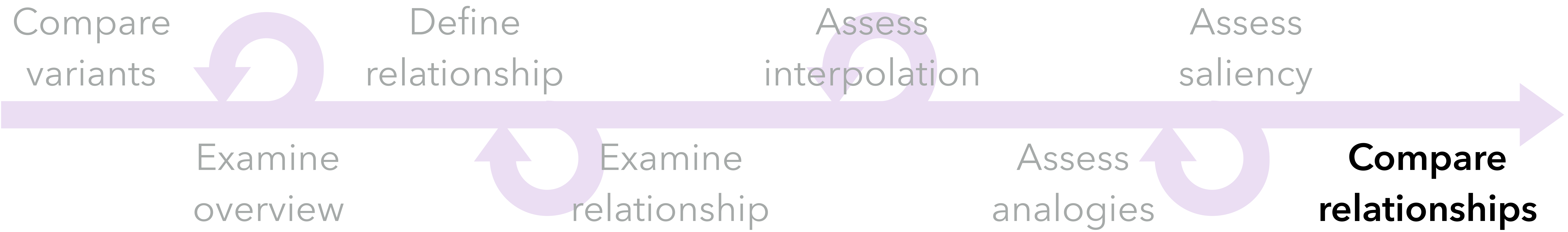
Analogy
Man Wearing Turban

Analogy
Man Pilot

Analogy
Man Health Worker

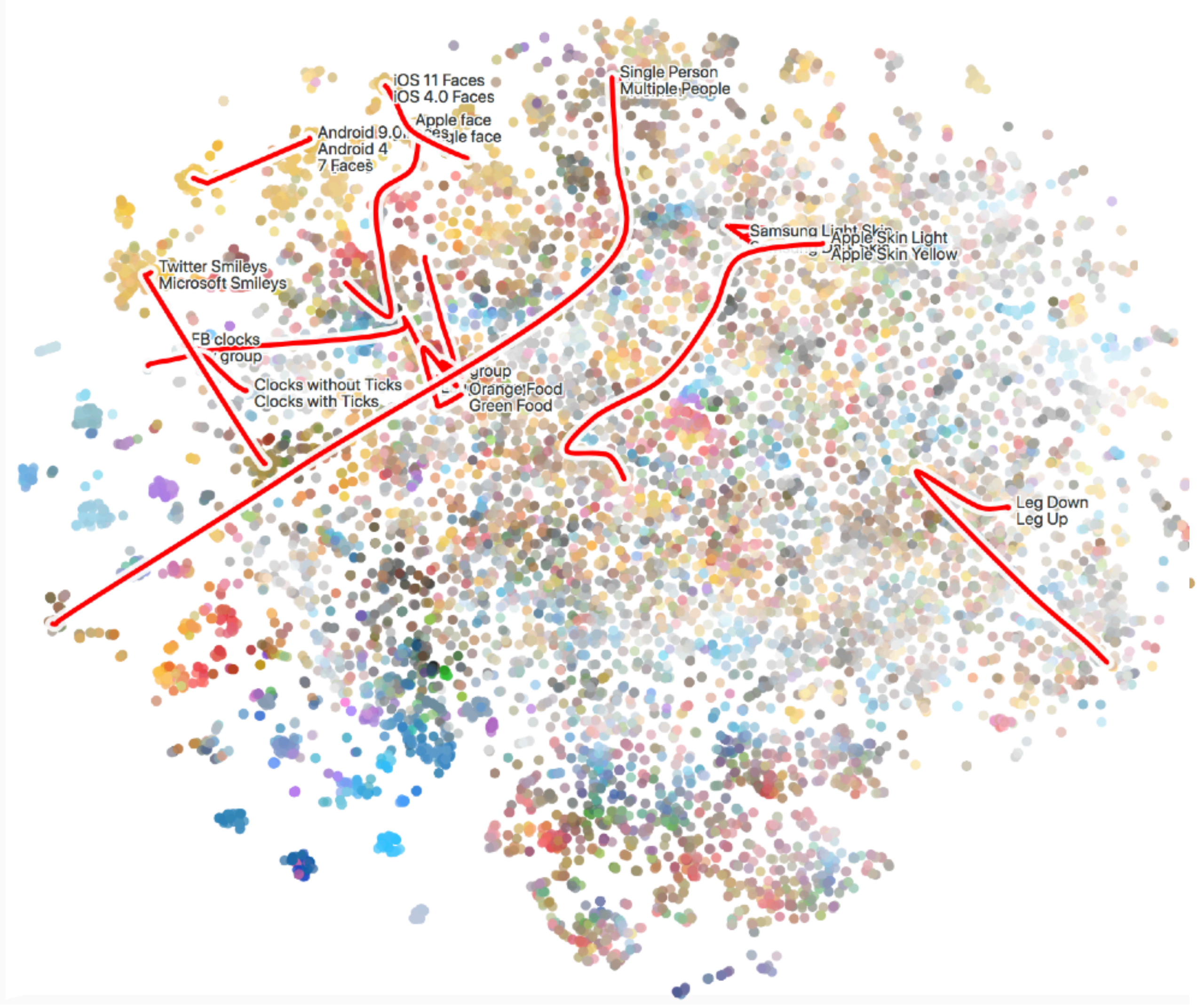


(… investigated more attribute vectors …)

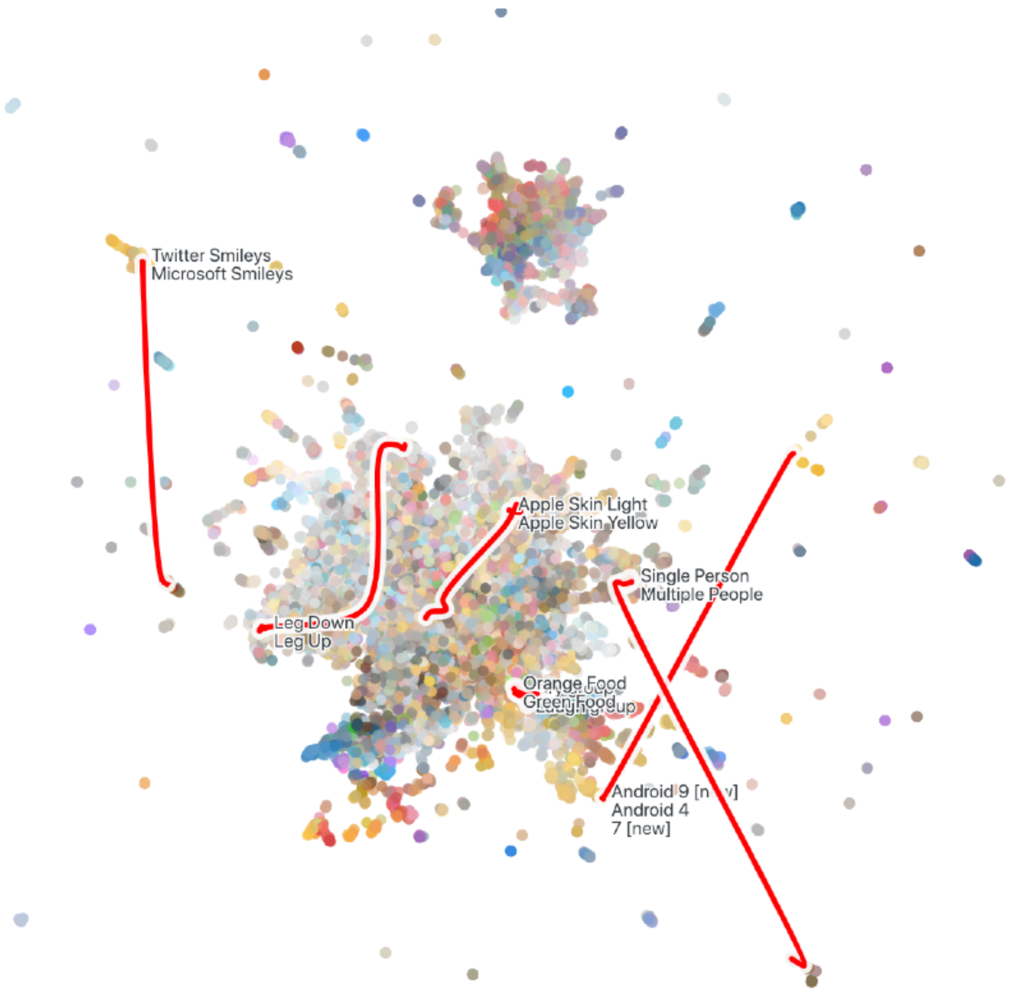


How do multiple attribute vectors relate?

Visualizing attribute vectors in a global view

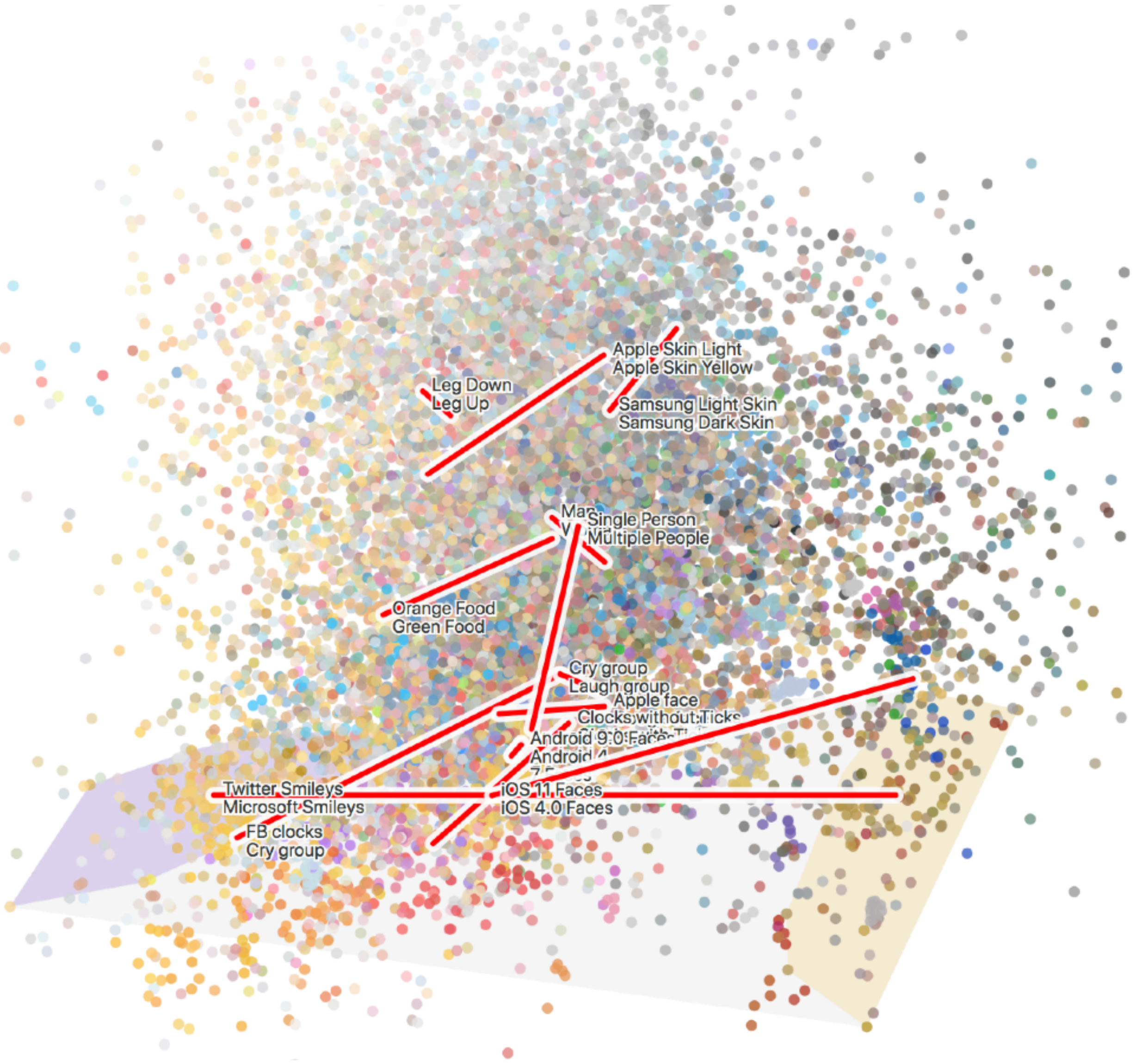


t-SNE



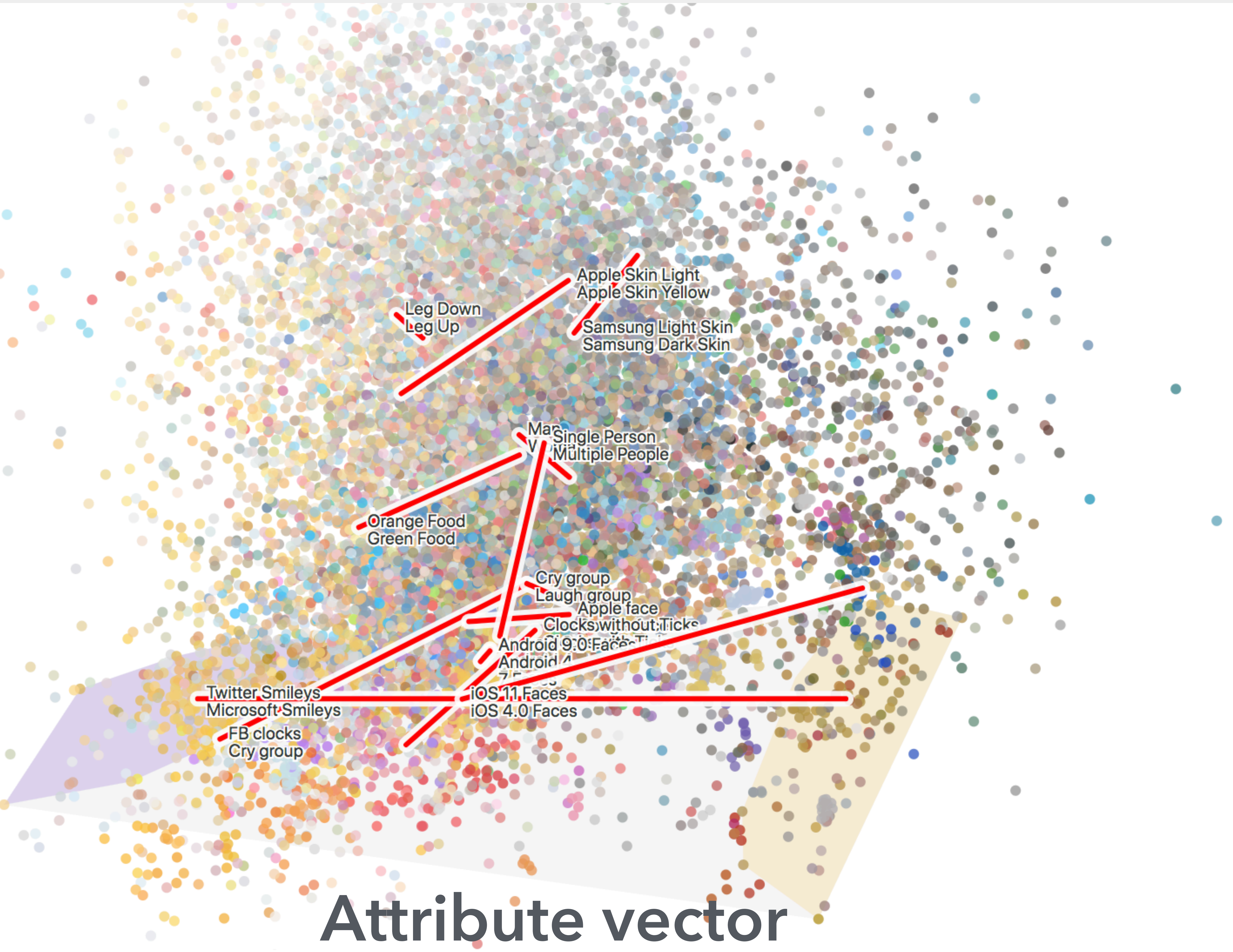
UMAP

PCA

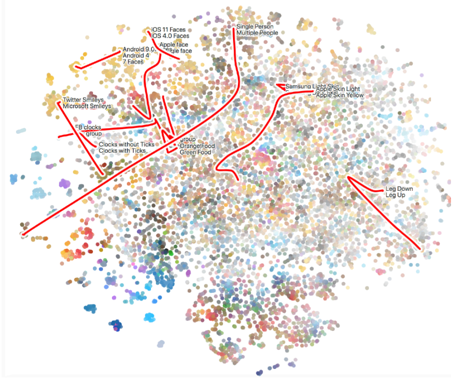


Attribute vector projection

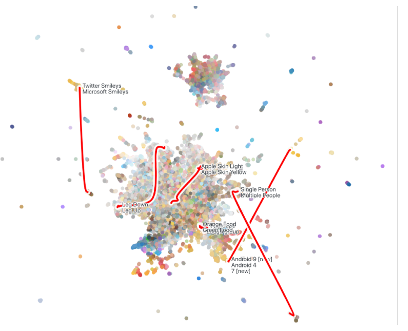
PCA



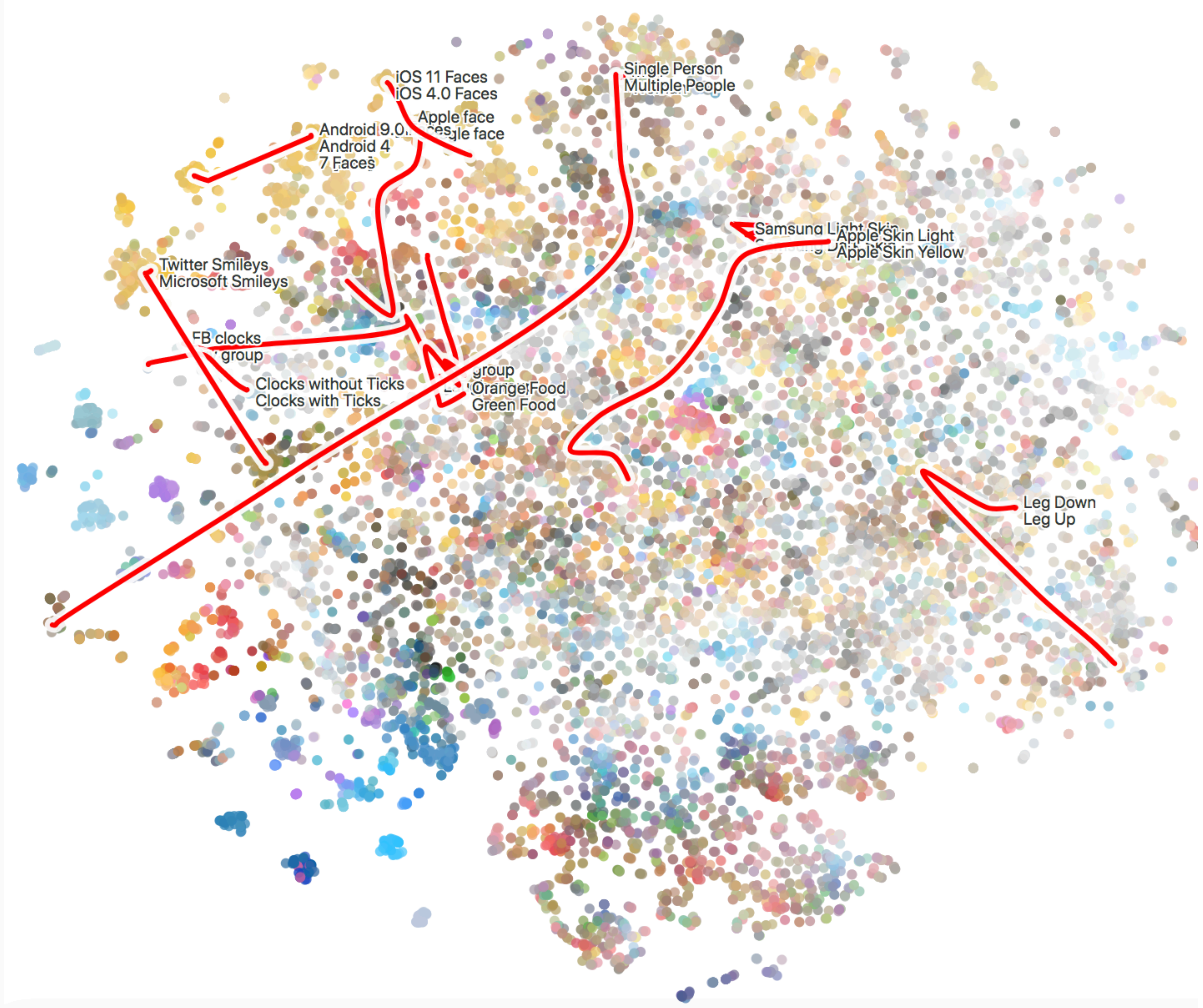
Attribute vector projection



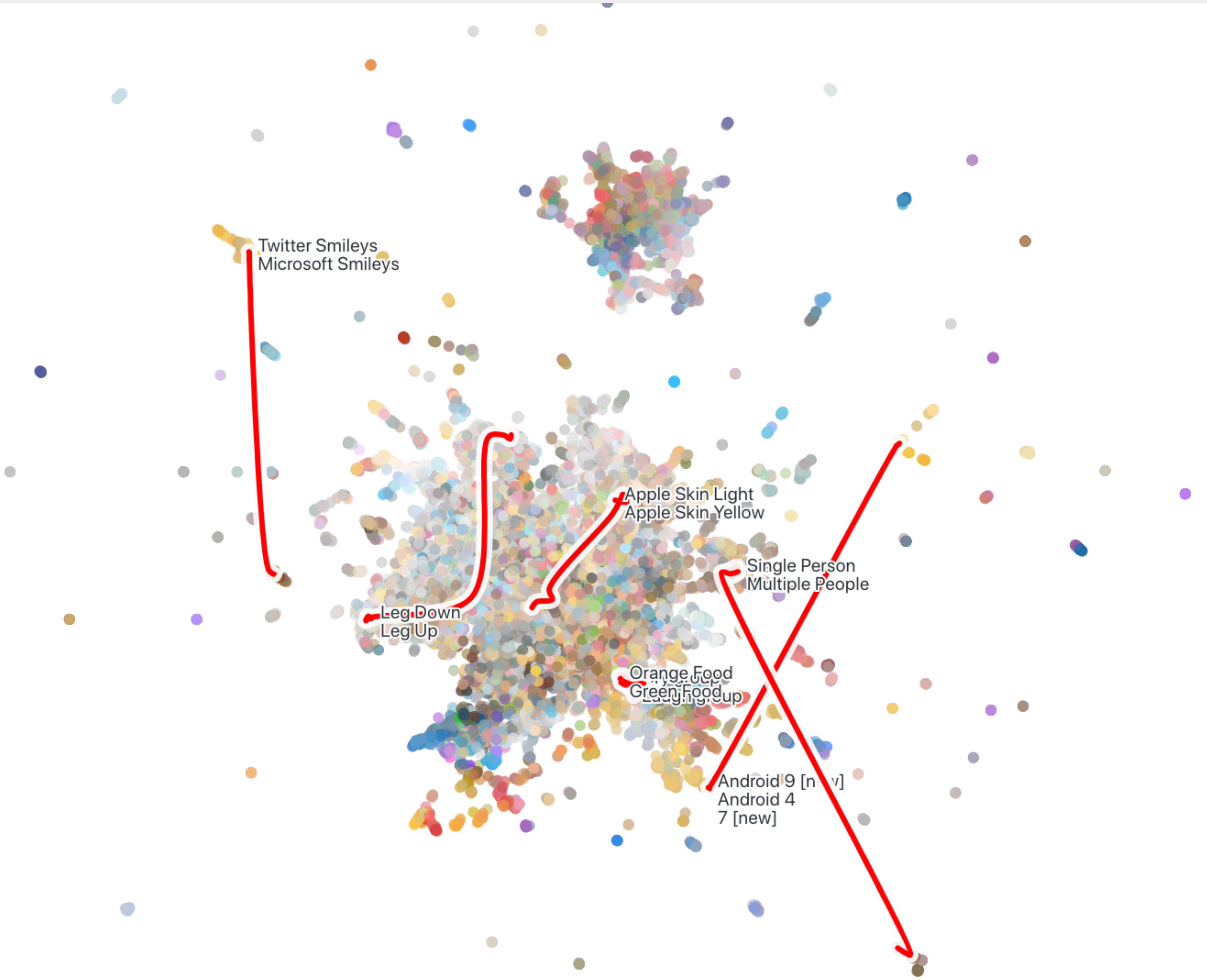
t-SNE



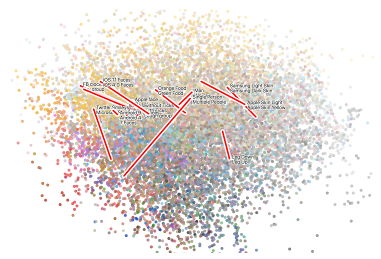
UMAP



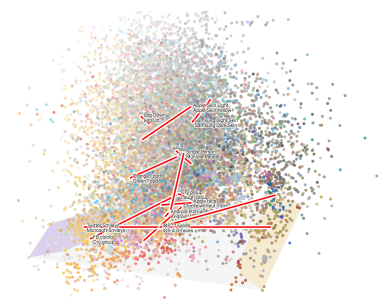
t-SNE



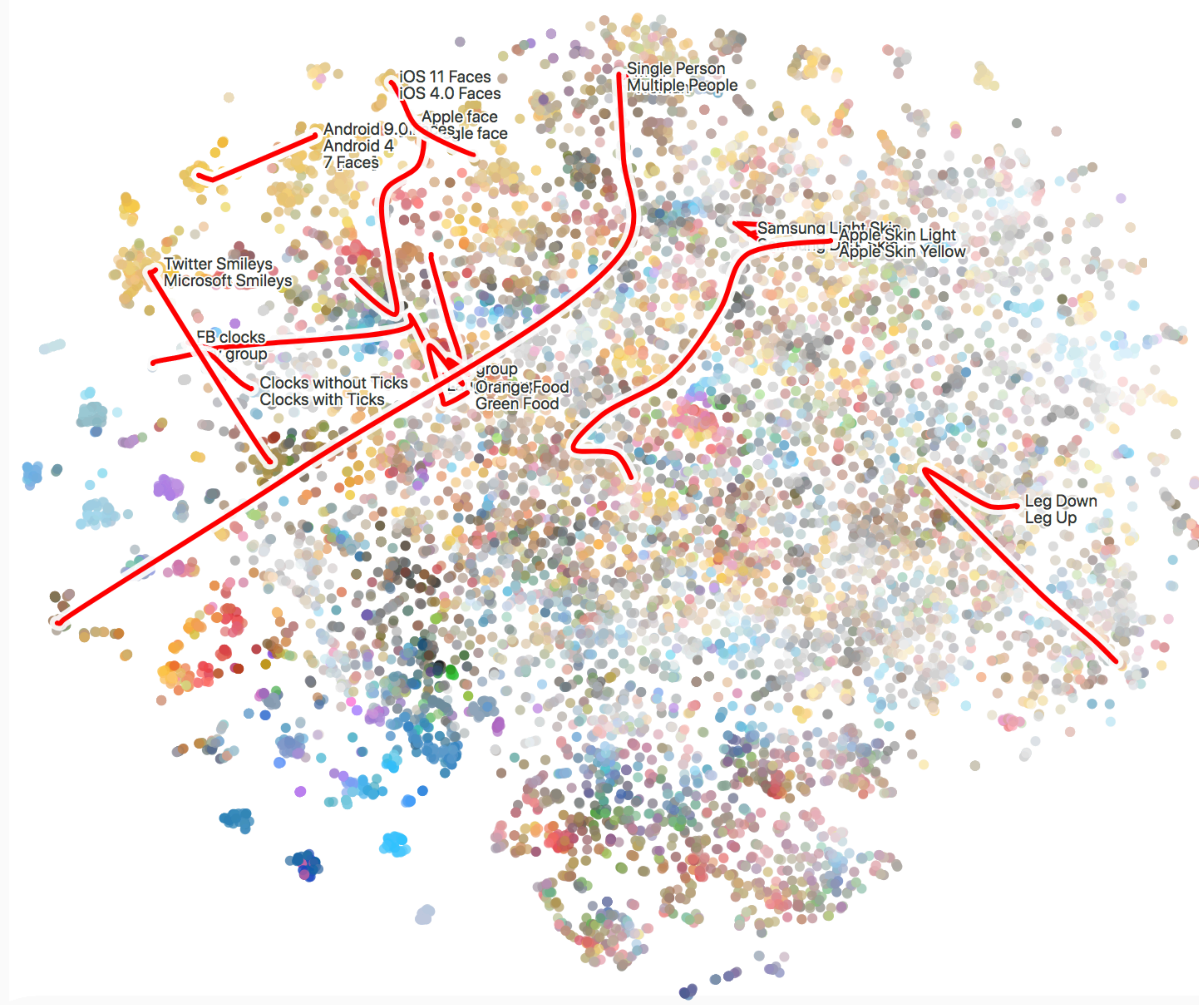
UMAP



PCA

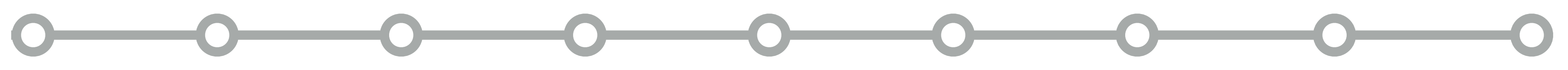


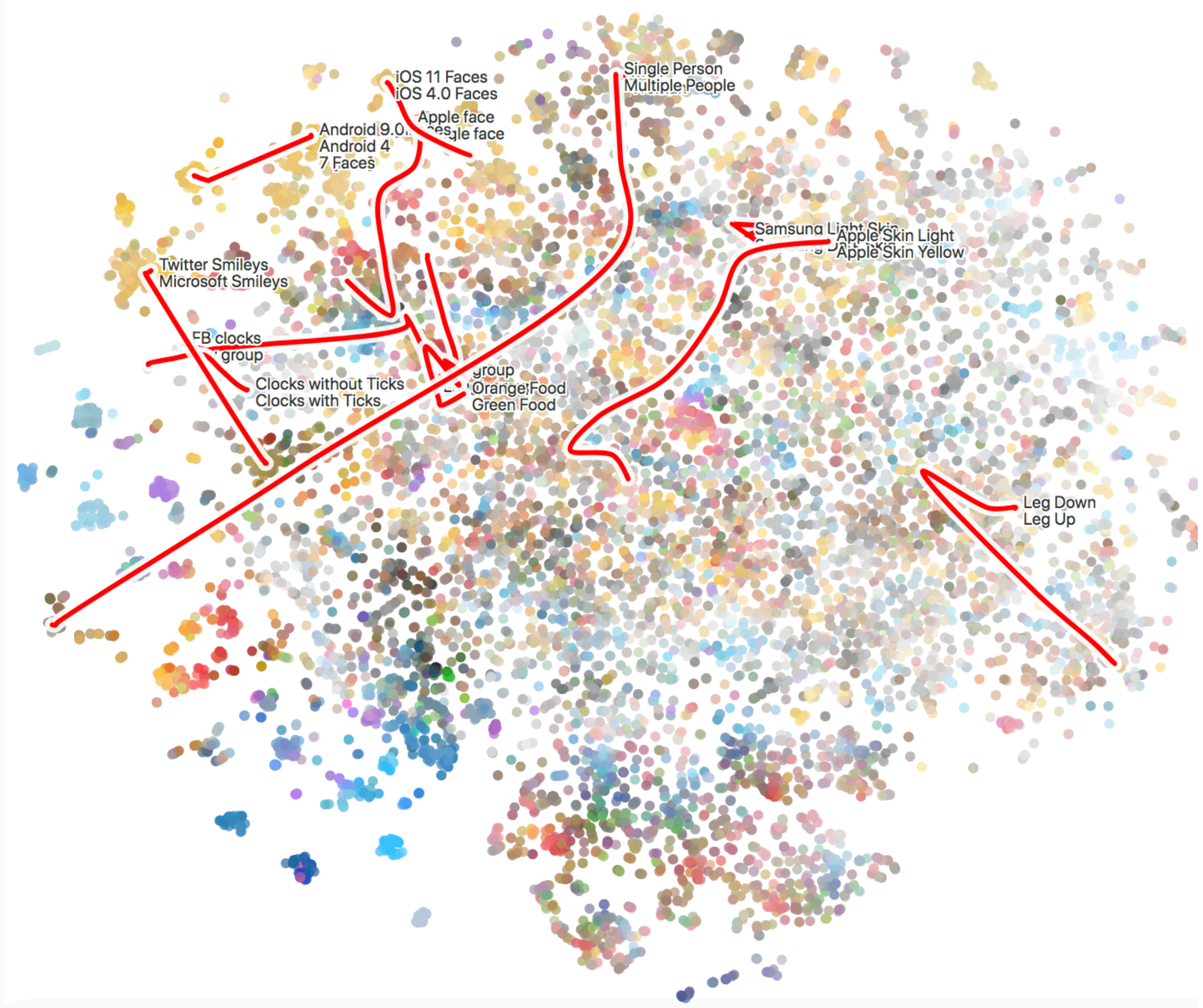
Attribute vector projection



t-SNE

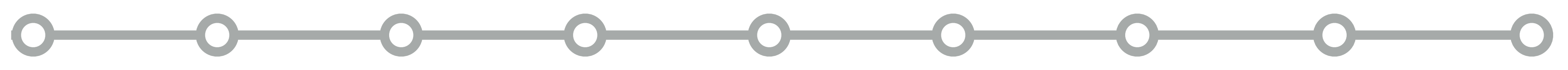
Step 1: sample at regular intervals





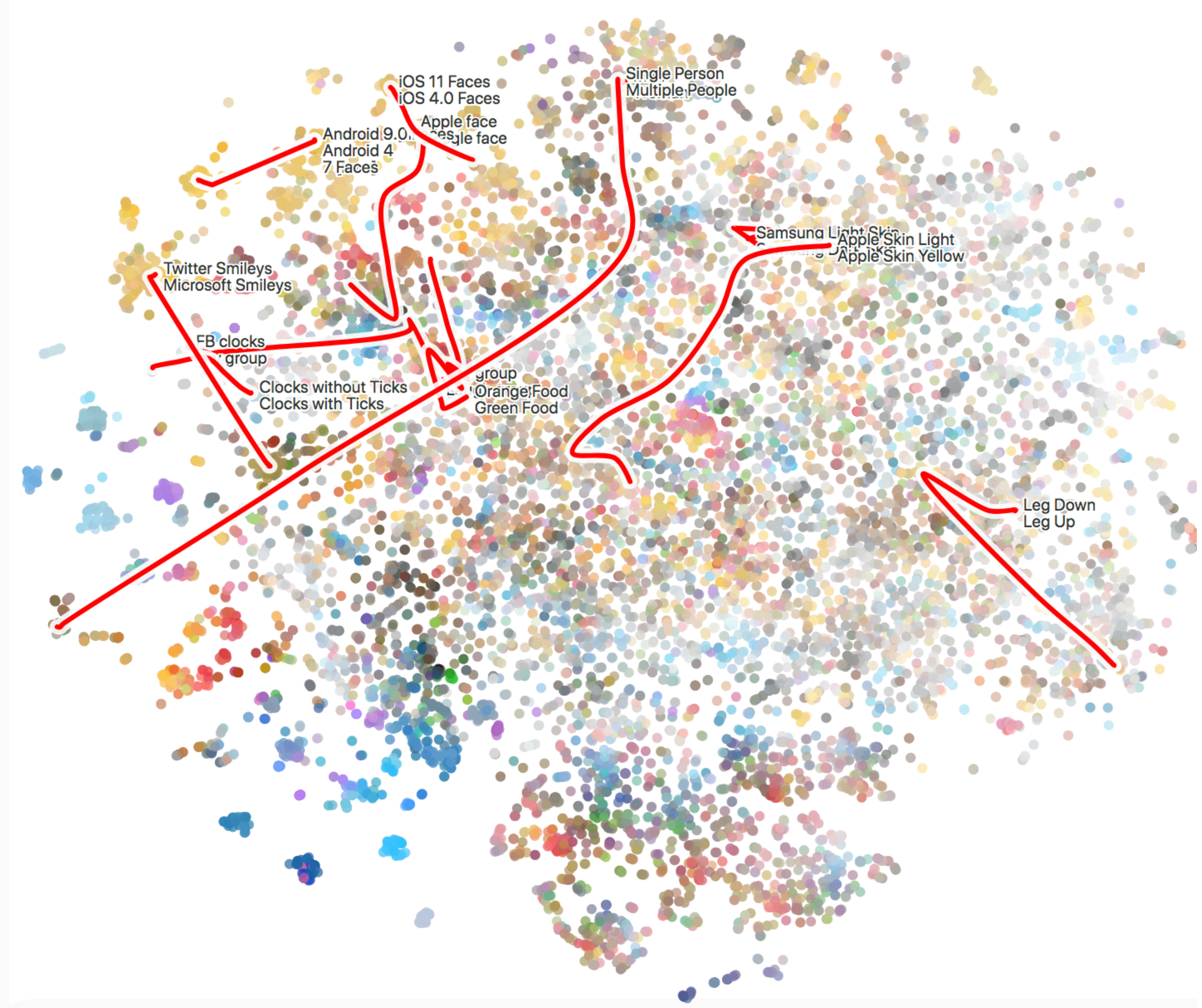
t-SNE

Step 1: sample at regular intervals



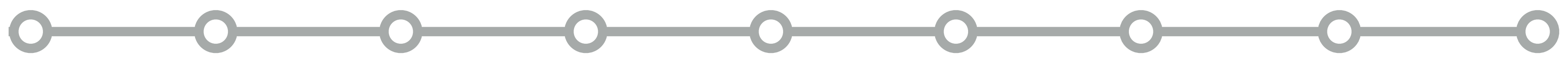
Step 2: map control points to 2D

Step 3: render a Catmull-Rom spline



t-SNE

Step 1: sample at regular intervals



Step 2: map control points to 2D

- Find k nearest neighbors
- Map neighbors to 2D
- Compute a weighted average

Step 3: render a Catmull-Rom spline

Examining how multiple attribute vectors relate

Examining how multiple attribute vectors relate

🤔 Orthogonal vectors represent independent dimensions …

🤯 Semantic axes to re-orient the latent space?

Latent Space Cartography

Visual Analysis of Vector Space Embeddings

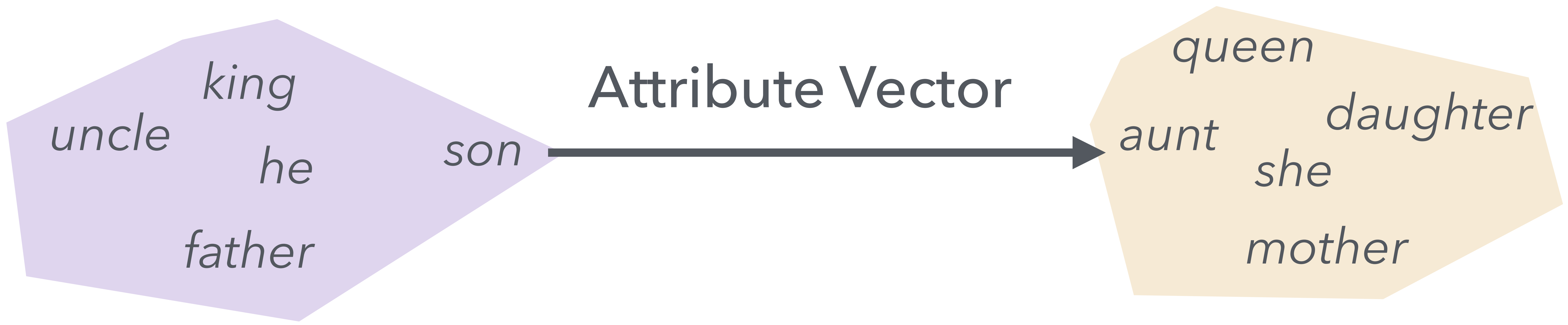
Gender Biases in Word Embeddings

… with a few simple interactions

Gender Biases in Word Embeddings

Bolukubasi *et al.* quantify which words are closer to ***he*** versus ***she*** in word embedding to reveal **gender stereotypes**

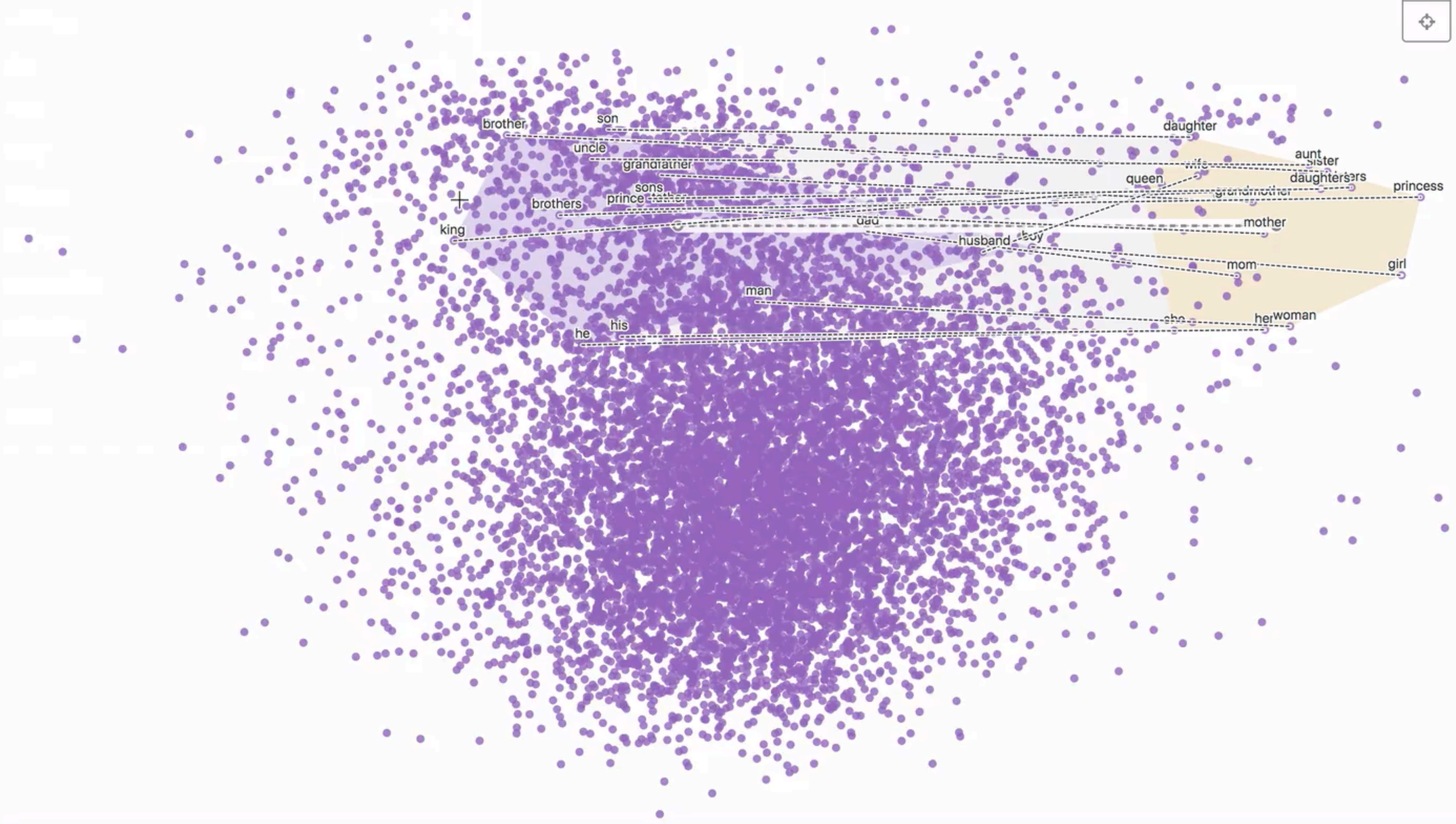
We'll quickly replicate the findings in LSC



BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V., KALAI A. T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (2016), pp. 4349–4357


Toggle Brush
Groups
Vectors
family
Start:
male
15 total
End:
female
15 total
Select a point (click, or search) to apply this attribute vector.
PAIRS WITHIN THE VECTOR
Effect Size: 3.68
-5 -4 -3 -2 -1 0 1 2 3 4 5
Standardized Cosine Distance (σ)
OTHER VECTORS
COSINE LABEL
-0.02 country - capital
0.07 participle - present
-0.19 nationality adjective - nationality
-0.19 past tense - participle 2
0.01 plural - singular
1.00 female - male
0.12 state - city
0.13 adverb - adjective
0.07 negative - positive
0.17 comparative - adjective 2
-0.01 superlative - adjective 3
brother
son
uncle
sons
brothers
prince
king
husband
man
his
he
daughter
aunt
queen
princess
mother
mom
girl
woman
Latent Dimensions: 50


Implicit stereotype in the training corpus
Brushed
mother
spokeswoman
ms.
wedding
pink
mom
nurse
bedroom
ladies
householder
butterfly
swim
raped
brother
son
uncle
brothers
prince
husband
man
he
his
daughter
aunt
queen
daughters
princess
mother
mom
girl
woman
Latent Dimensions: 50
Groups
Vectors
family
Start:
male
15 total
End:
female
15 total
Select a point (click, or search) to apply this attribute vector.
PAIRS WITHIN THE VECTOR
Effect Size: 3.68
-5 -4 -3 -2 -1 0 1 2 3 4 5
Standardized Cosine Distance (σ)
OTHER VECTORS
COSINE LABEL
-0.02 country - capital
0.07 participle - present
-0.19 nationality adjective - nationality
-0.19 past tense - participle 2
0.01 plural - singular
1.00 female - male
0.12 state - city
0.13 adverb - adjective
0.07 negative - positive
0.17 comparative - adjective 2
-0.01 superlative - adjective 3



Latent Dimensions: 50

Groups | Vectors

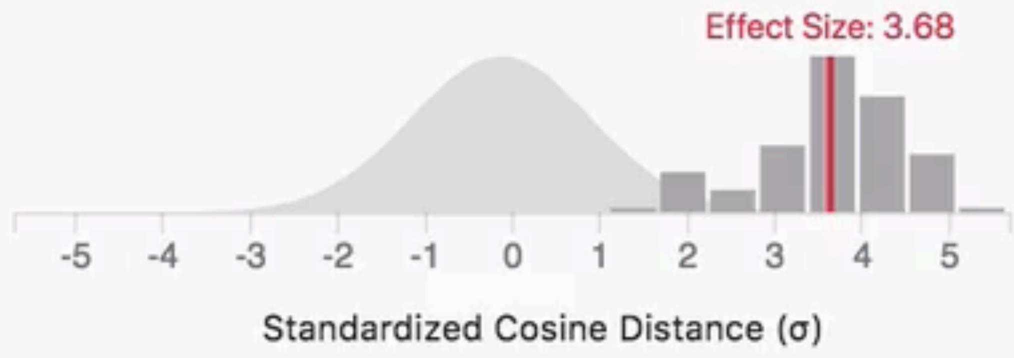
family

Start:
male
15 total

End:
female
15 total

Select a point (click, or search) to apply this attribute vector.

PAIRS WITHIN THE VECTOR



OTHER VECTORS

COSINE	LABEL
-0.02	country - capital
0.07	participle - present
-0.19	nationality adjective - nationality
-0.19	past tense - participle 2
0.01	plural - singular
1.00	female - male
0.12	state - city
0.13	adverb - adjective
0.07	negative - positive
0.17	comparative - adjective 2
-0.01	superlative - adjective 3


Brushed
season
director
player
victory
mark
joined
ball
mayor
w.
owner
brothers
bowl
founder
rugby
gov.
hero
singh
longtime
regiment
jacques
businessman
critic
Latent Dimensions: 50
brother
son
uncle
daughter
aunt
queen
princess
mother
husband
mom
girl
man
woman
his
he
Groups
Vectors
family
Start:
male
15 total
End:
female
15 total
Select a point (click, or search) to apply this attribute vector.
PAIRS WITHIN THE VECTOR
Effect Size: 3.68
-5 -4 -3 -2 -1 0 1 2 3 4 5
Standardized Cosine Distance (σ)
OTHER VECTORS
COSINE LABEL
-0.02 country - capital
0.07 participle - present
-0.19 nationality adjective - nationality
-0.19 past tense - participle 2
0.01 plural - singular
1.00 female - male
0.12 state - city
0.13 adverb - adjective
0.07 negative - positive
0.17 comparative - adjective 2
-0.01 superlative - adjective 3

Words are shifted toward the *male* concept

Analysis of Analogy Benchmark

Google's Analogy Benchmark

Family

king:queen
son:daughter
uncle:aunt
…

Comparative

bad:worse
bright:brighter
high:higher
…

• • •

MIKOLOV T., YIH W.-T., ZWEIG G.: Linguistic regularities in continuous space word representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013), pp. 746–751

Google’s Analogy Benchmark

Family

king:queen
son:daughter
uncle:aunt
…

MIKOLOV T., YIH W.-T., ZWEIG G.: Linguistic regularities in continuous space word representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013), pp. 746–751

Google's Analogy Benchmark

Family

king:queen

son:daughter

uncle:aunt

…

king:queen son:daughter

v = *vec*(king) - *vec*(son) + *vec*(queen)

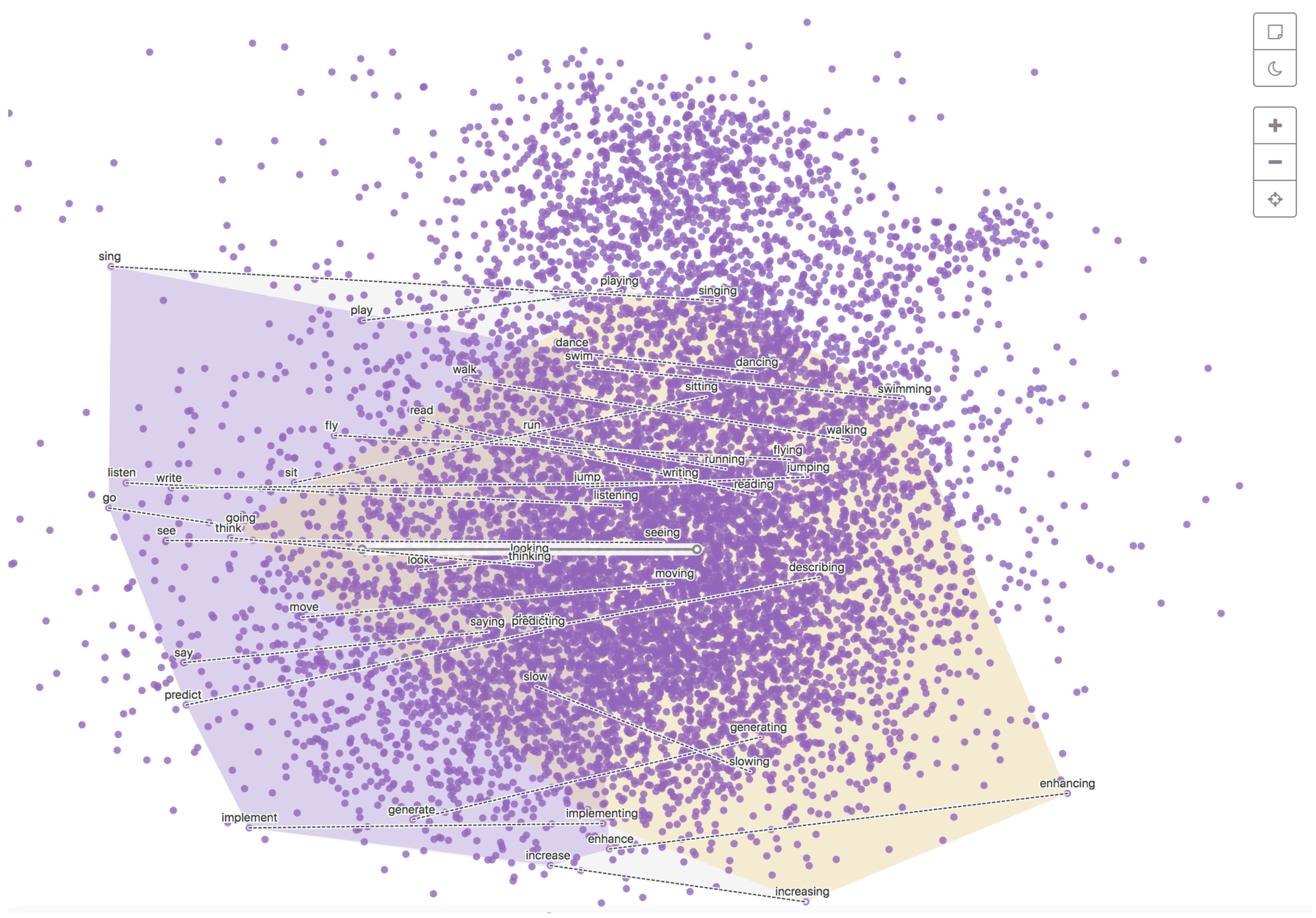
Is ***v*** the **nearest neighbor** of *vec*(daughter)?

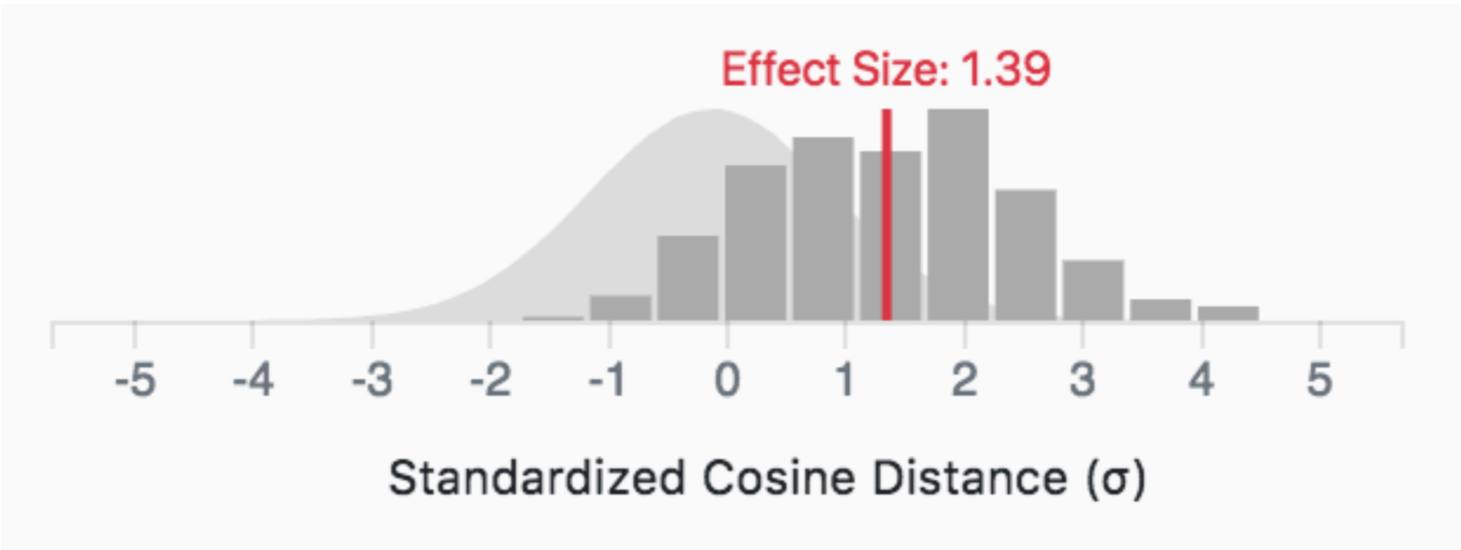
We use words in these analogy groups to define attribute vectors in LSC

MIKOLOV T., YIH W.-T., ZWEIG G.: Linguistic regularities in continuous space word representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013), pp. 746–751

Dimension = 50

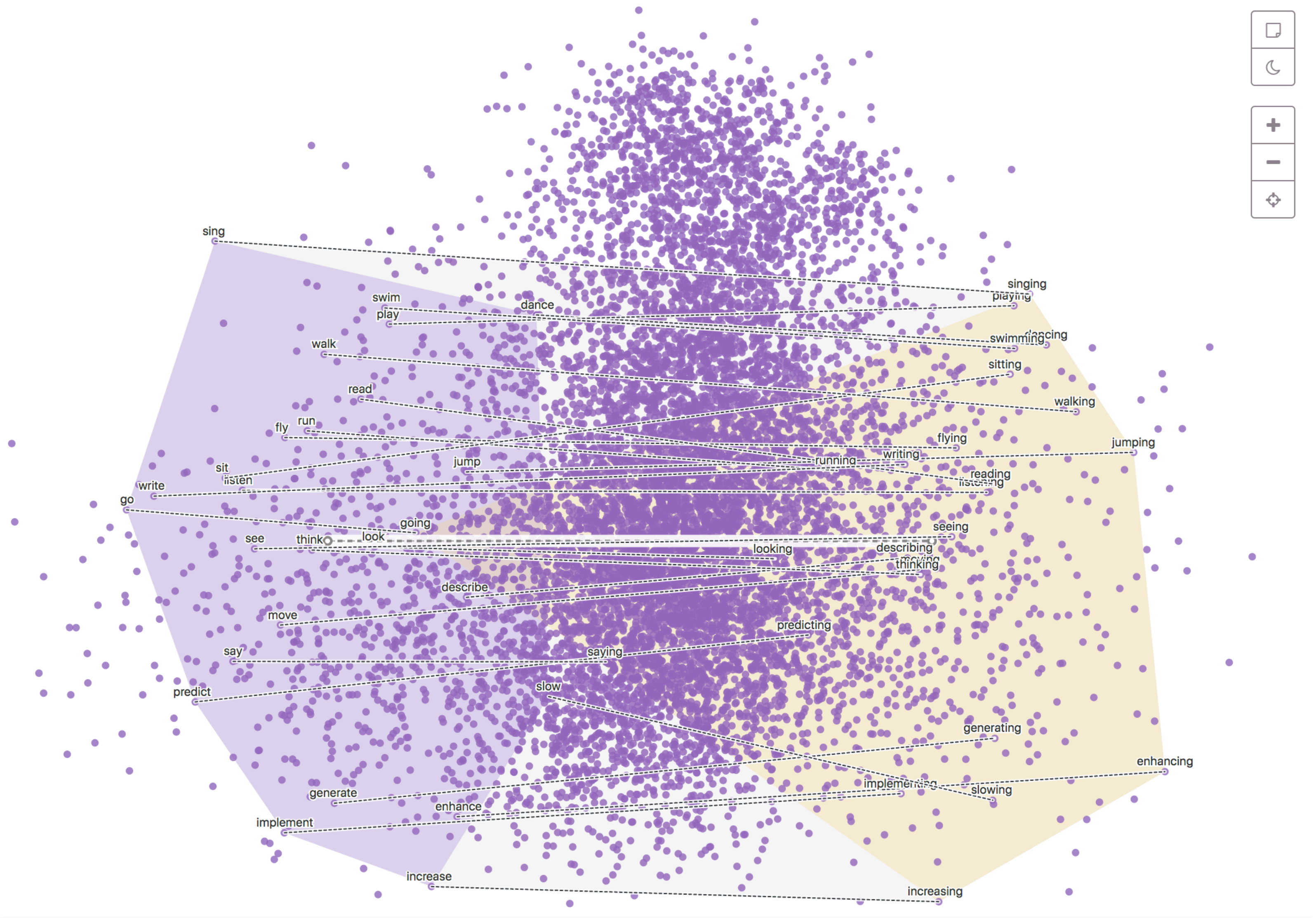
Google's analogy test score:
52.8% (317/600)

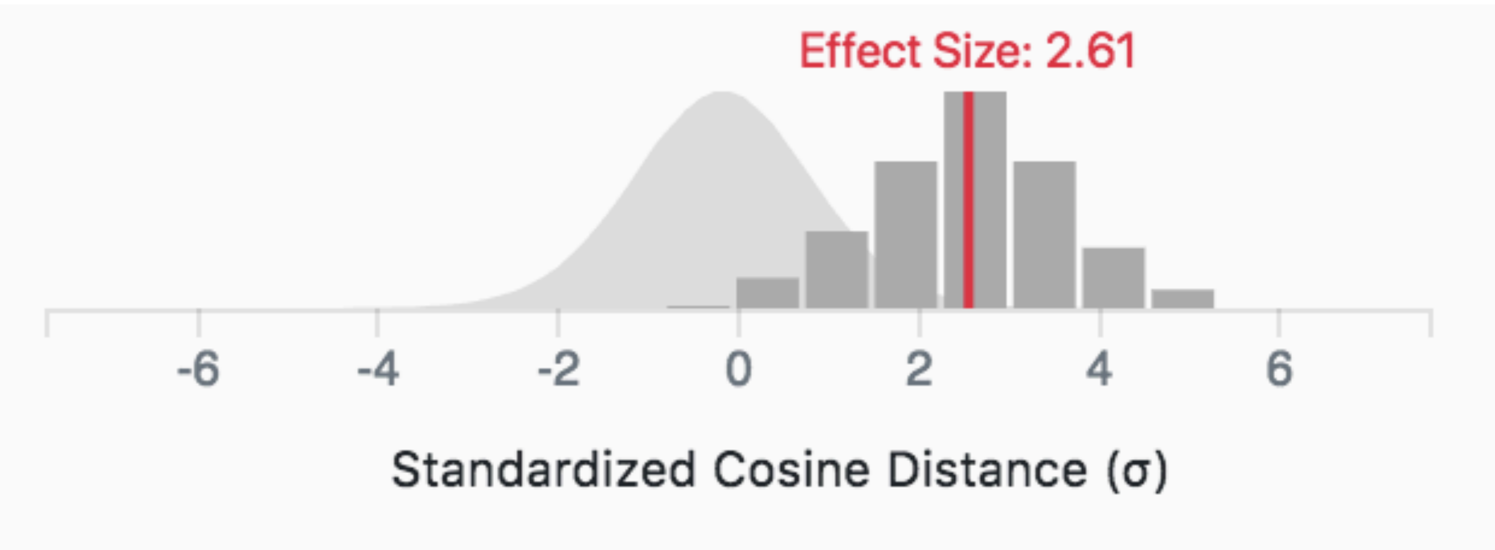




Dimension = 100

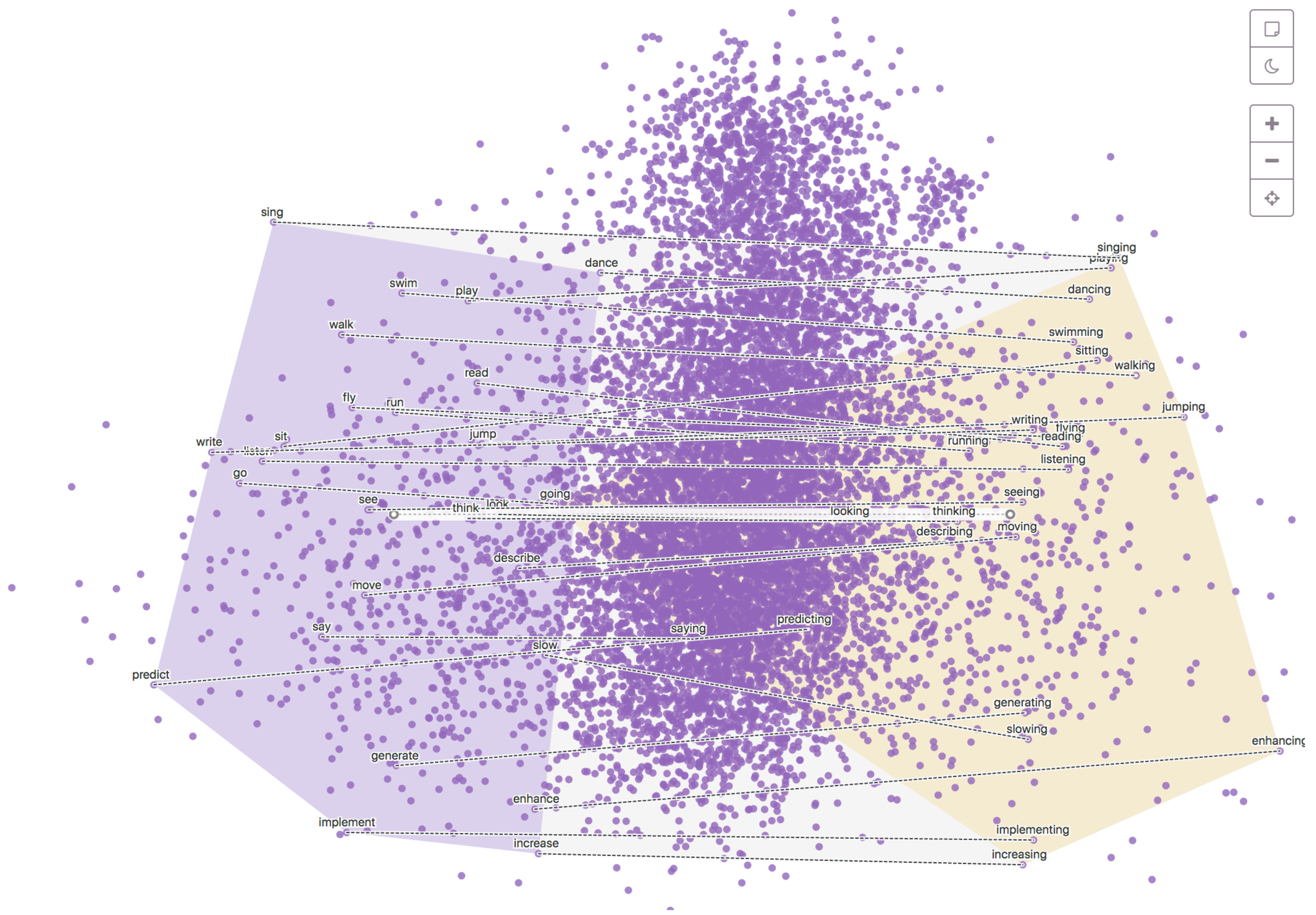
Google's analogy test score:
78.0% (468/600)

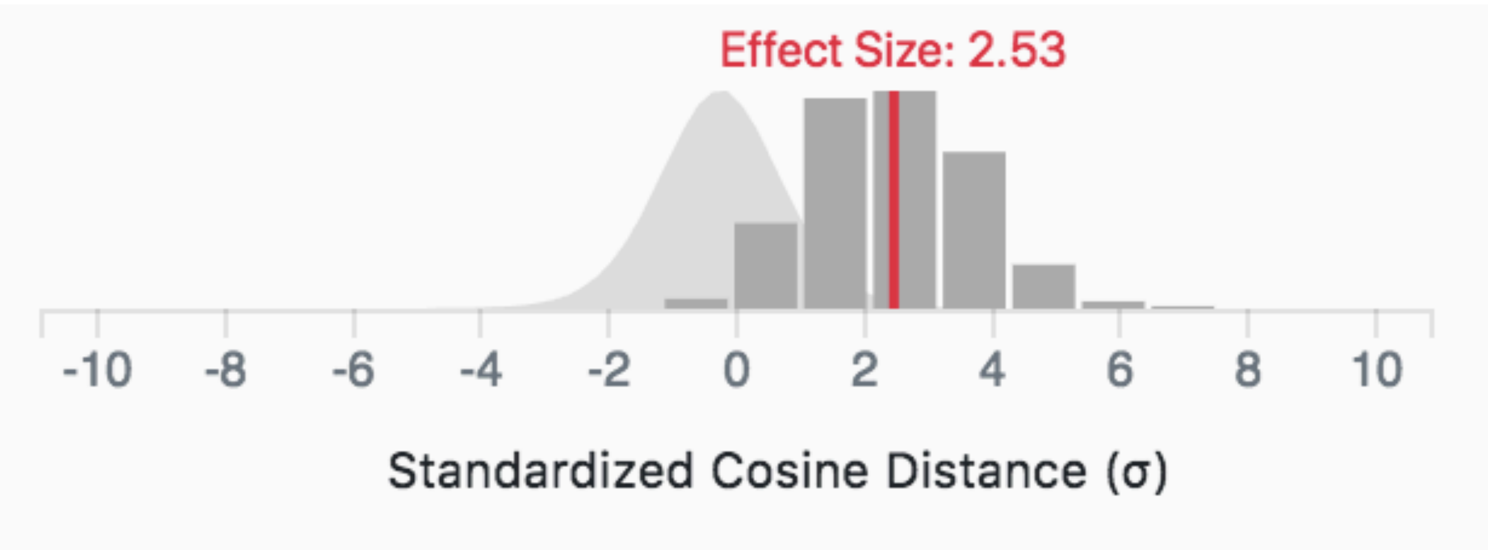




Dimension = 300

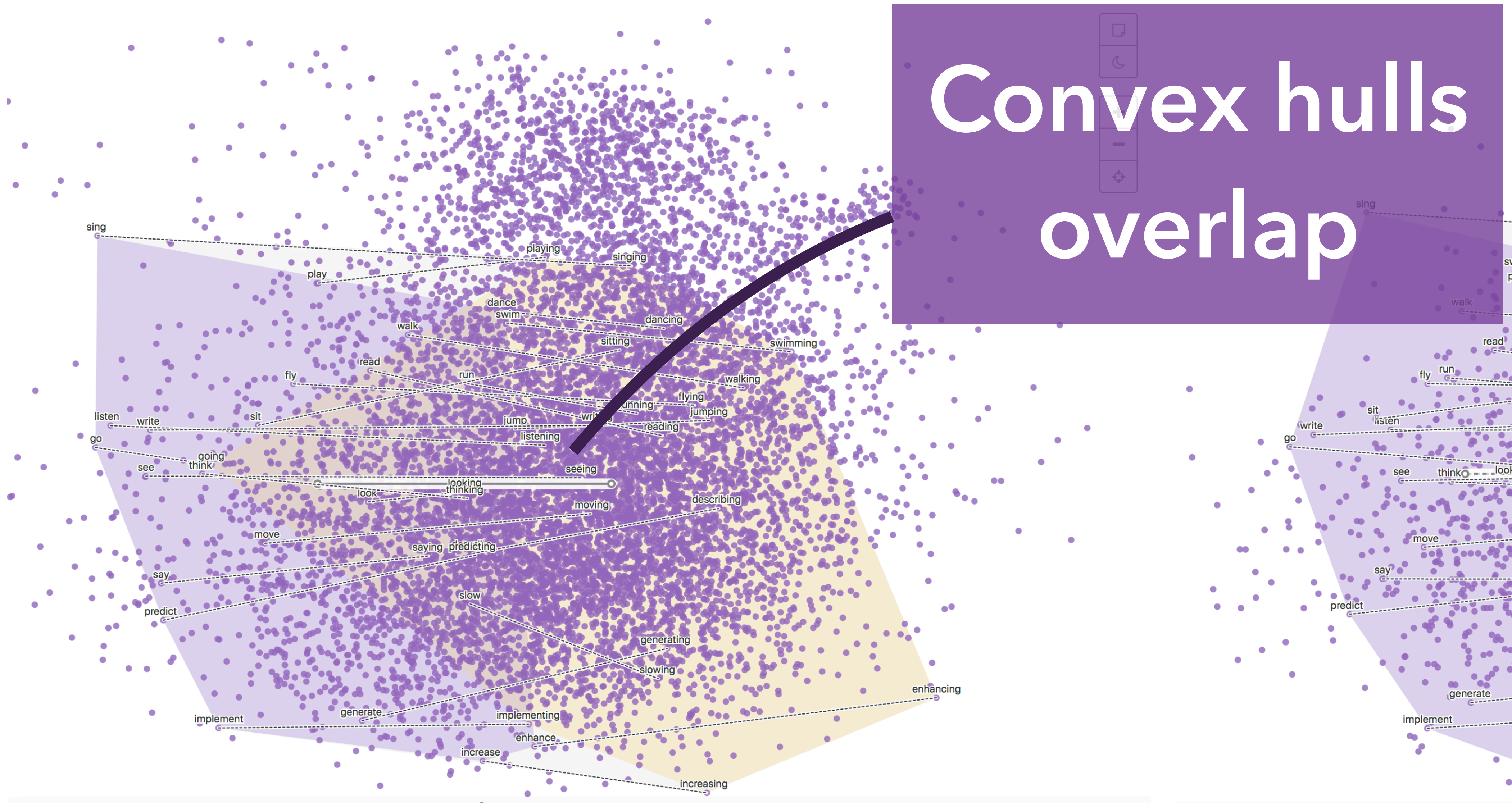
Google's analogy test score:
78.7% (472/600)

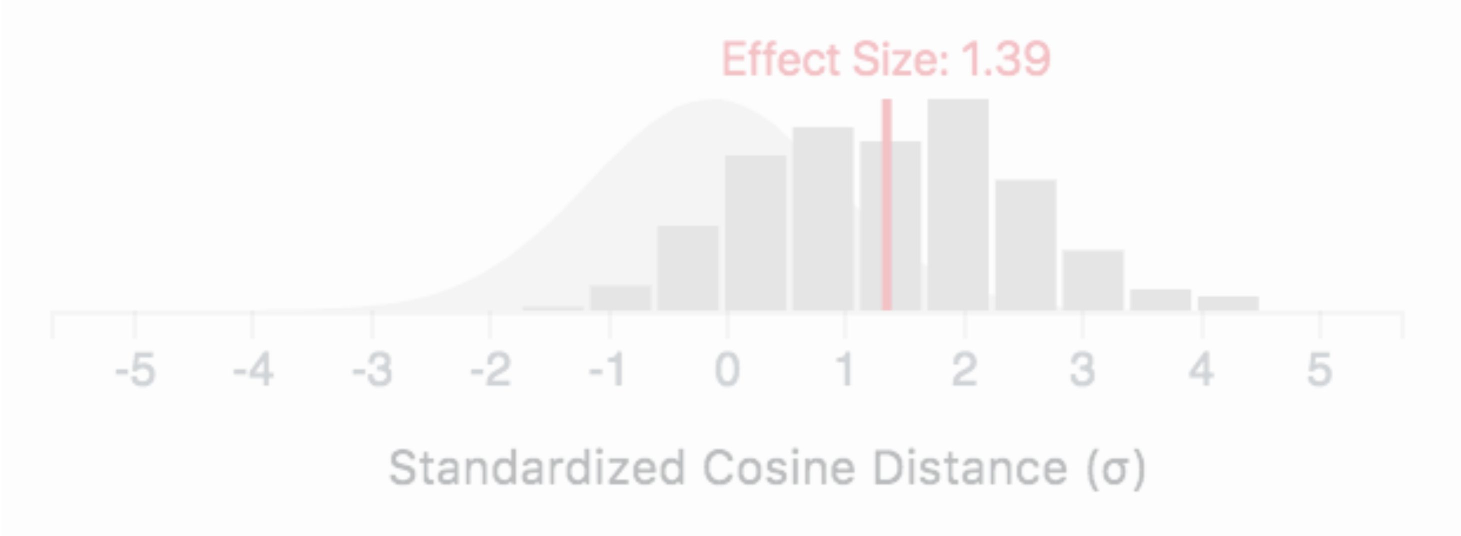




Dimension = 50

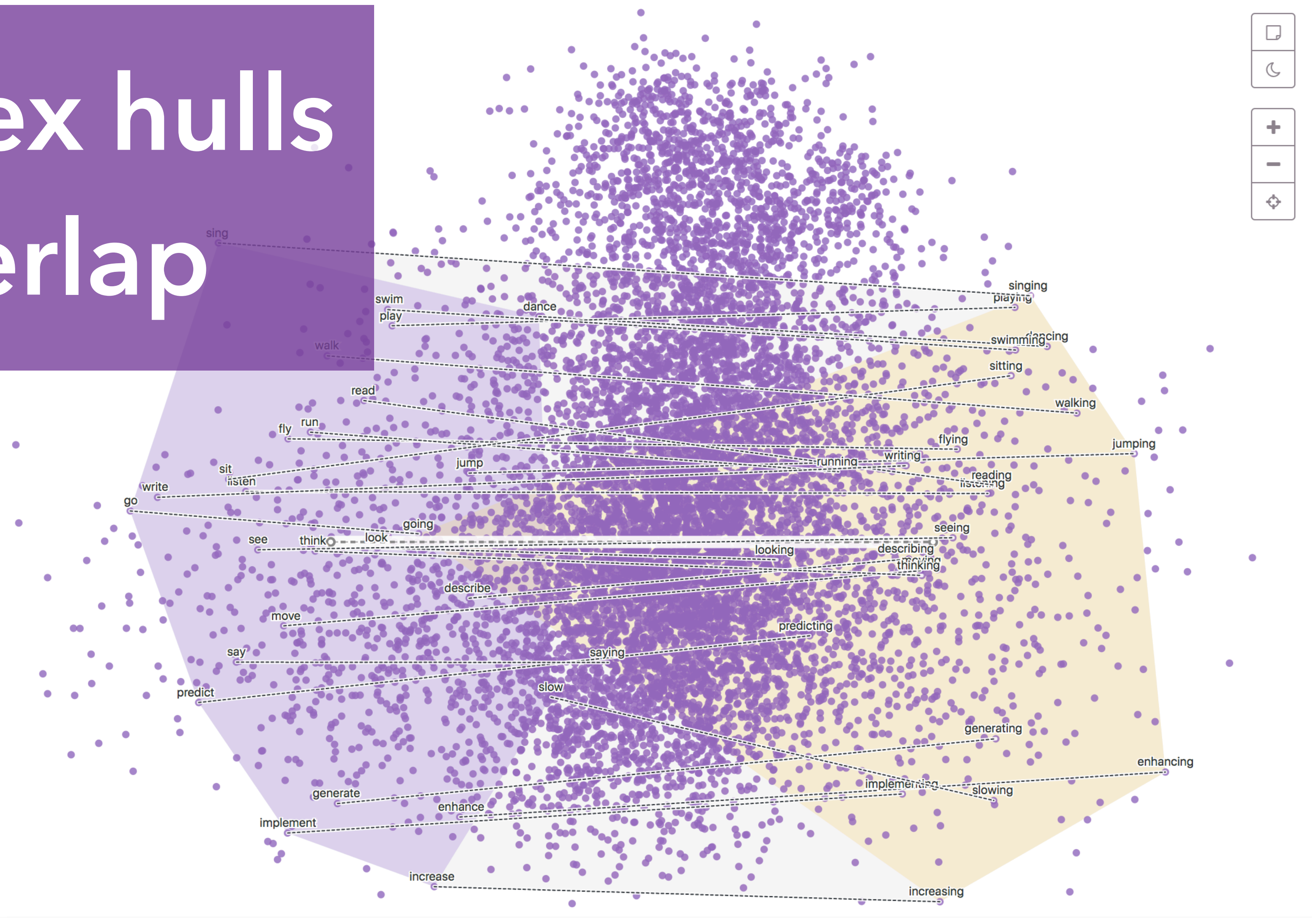
Google's analogy test score:
52.8% (317/600)

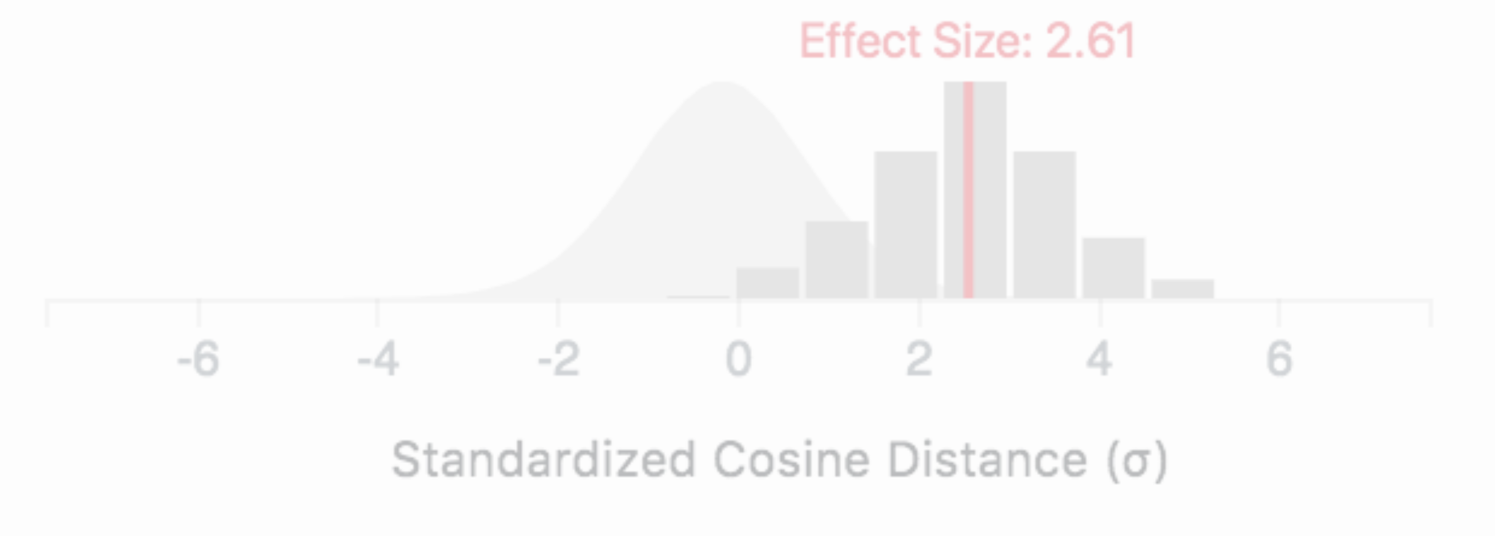




Dimension = 100

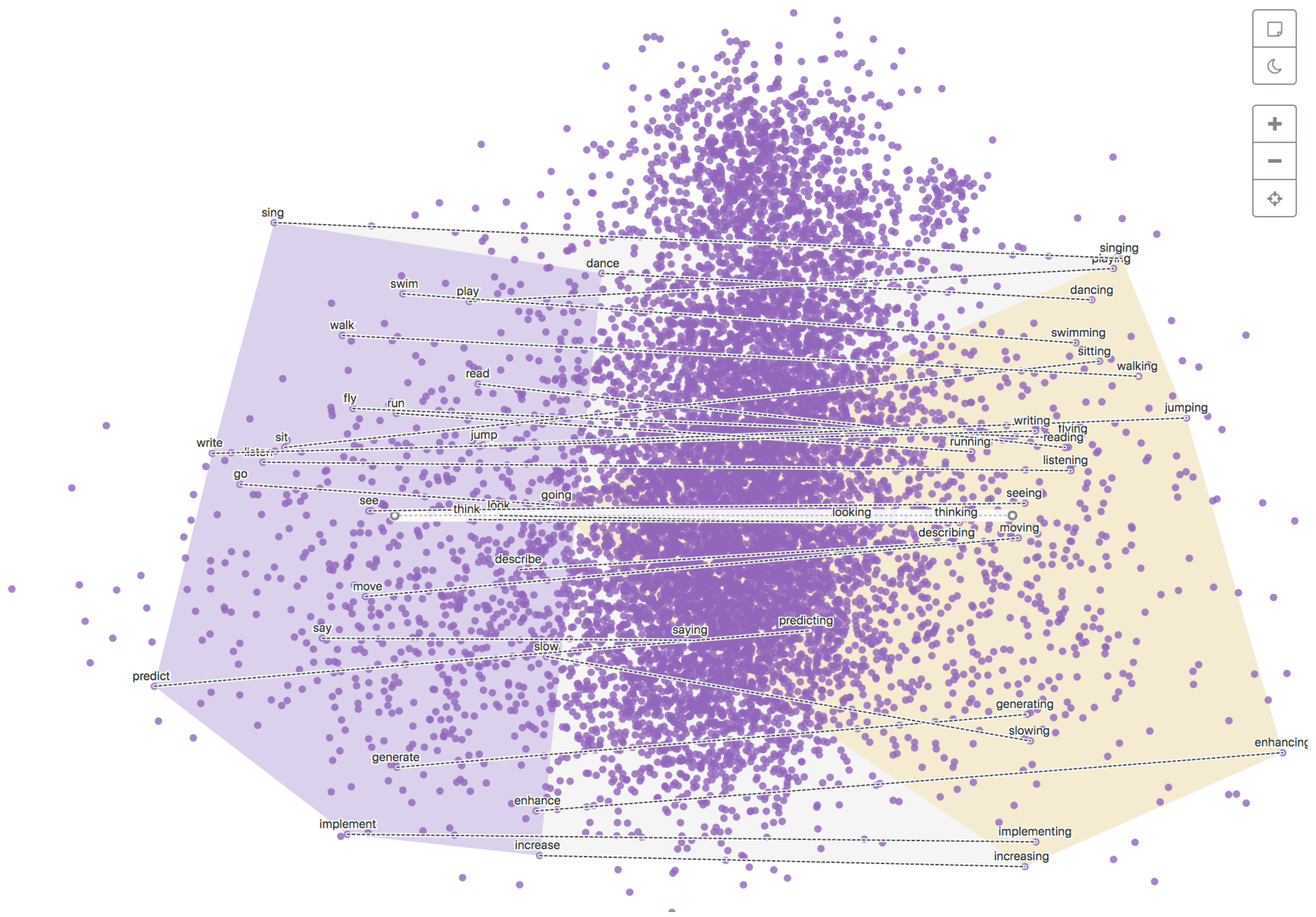
Google's analogy test score:
78.0% (468/600)

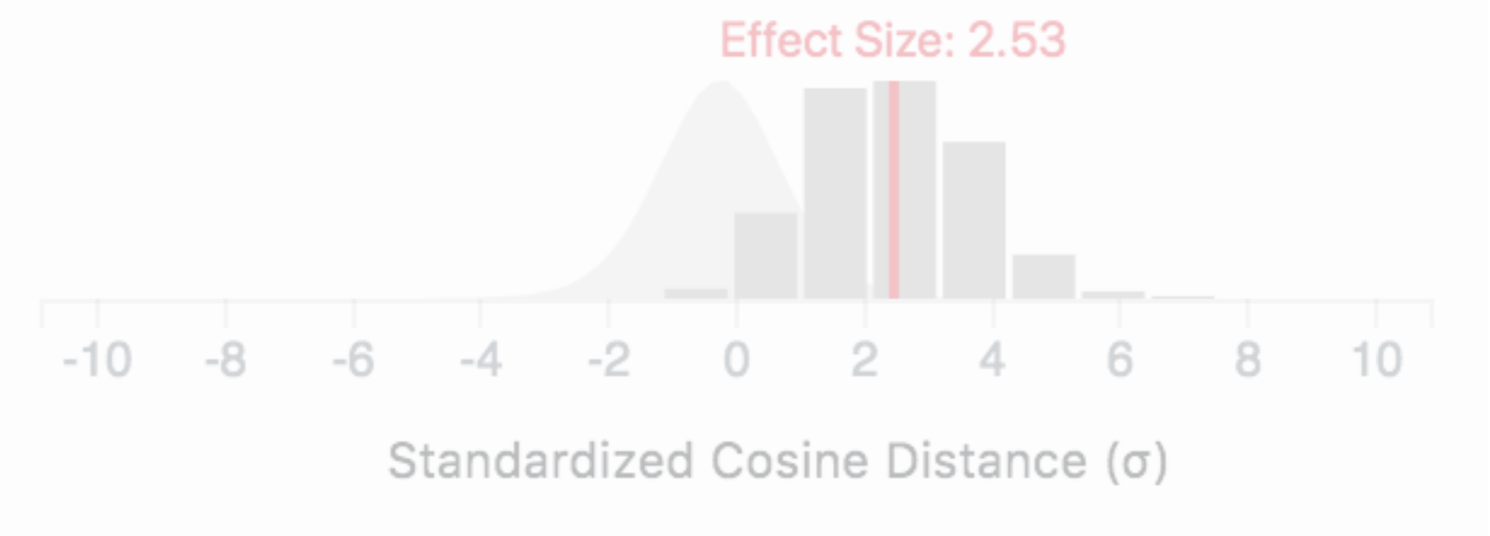




Dimension = 300

Google's analogy test score:
78.7% (472/600)





Dimension = 50

Google's analogy test score:
52.8% (317/600)

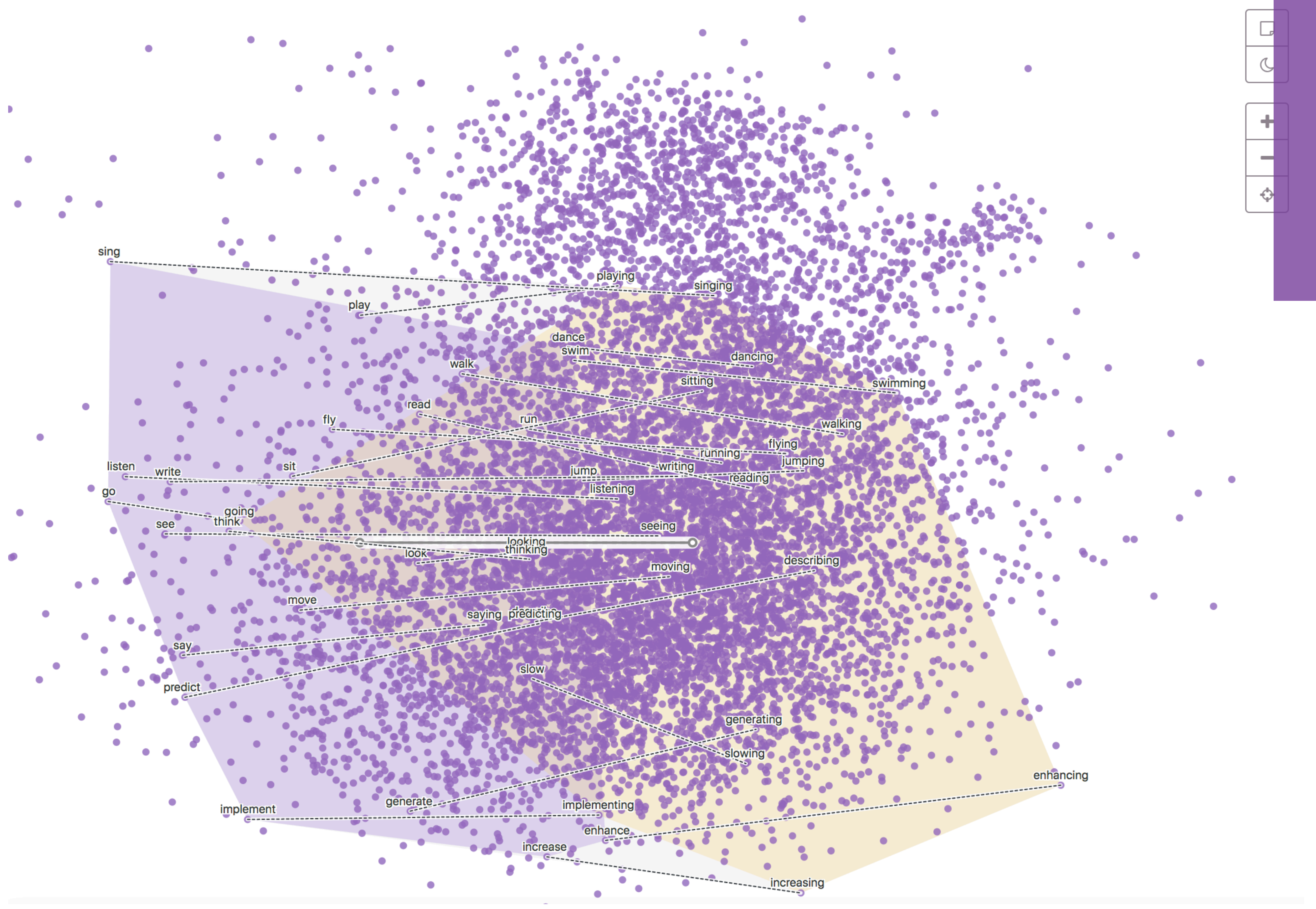
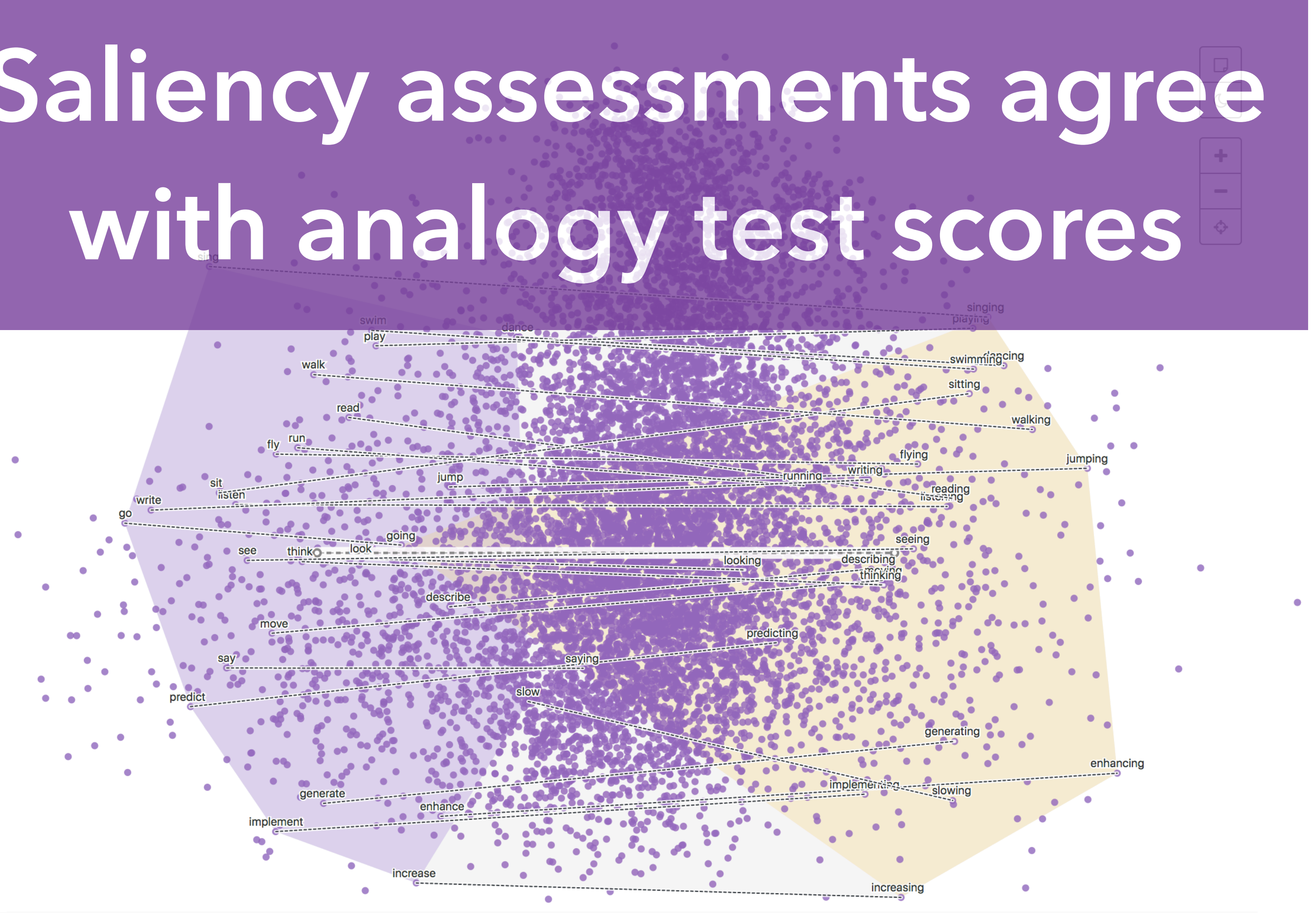
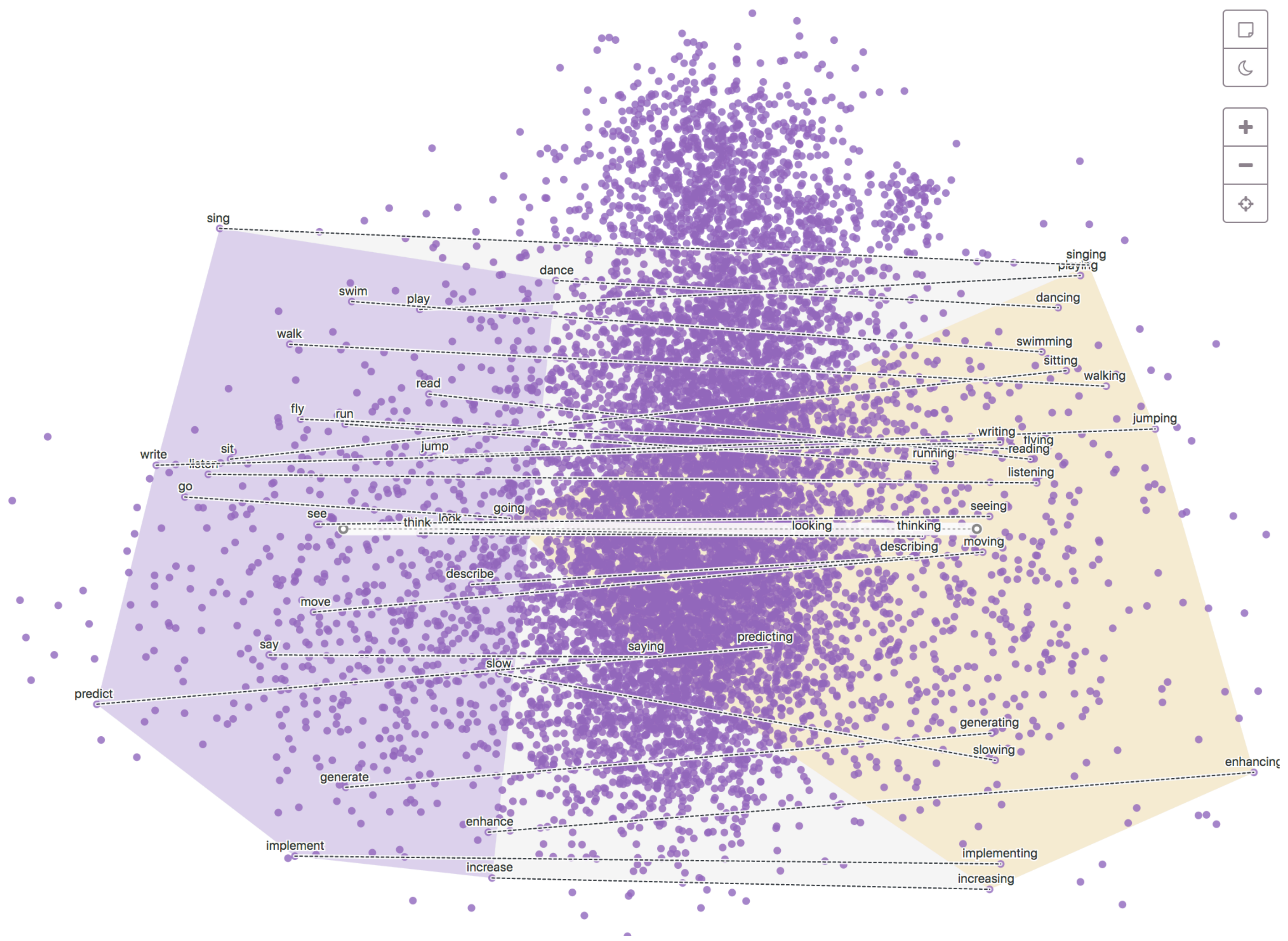
Dimension = 100

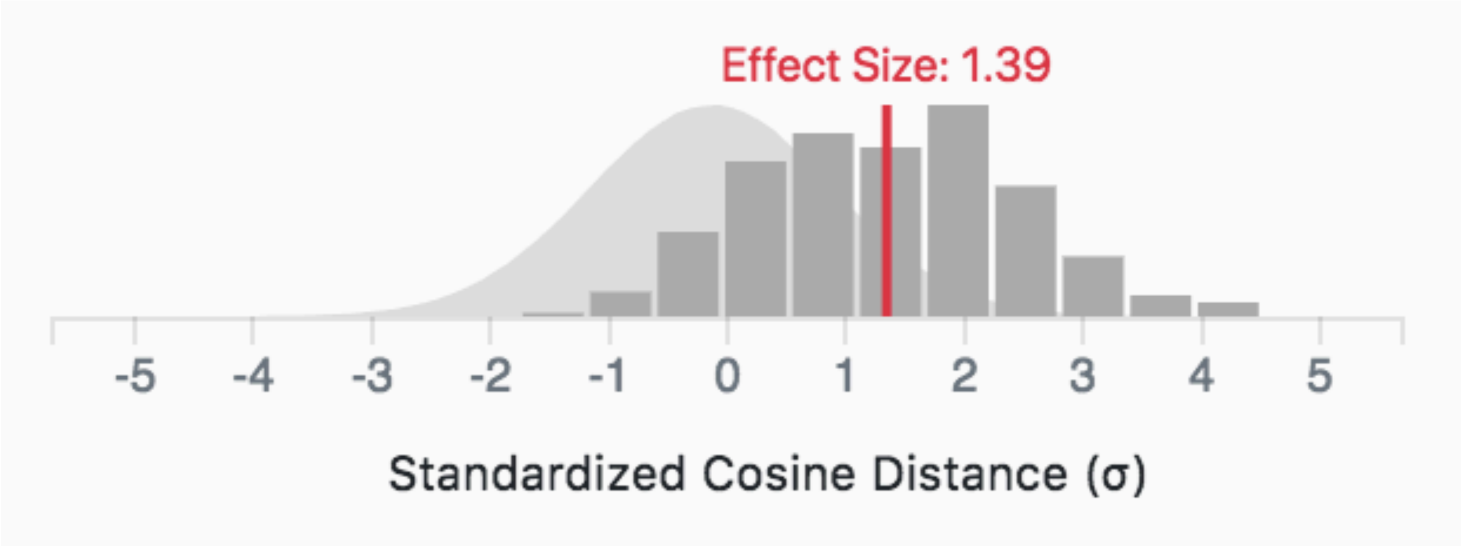
Google's analogy test score:
78.0% (468/600)

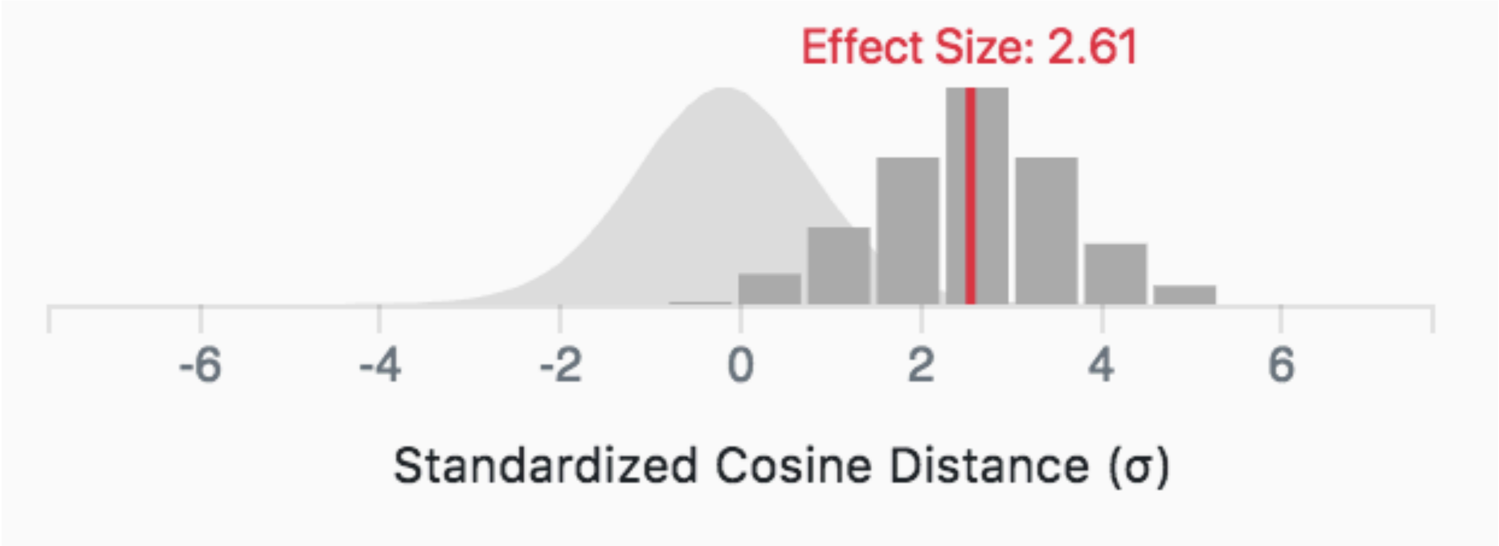
Dimension = 300

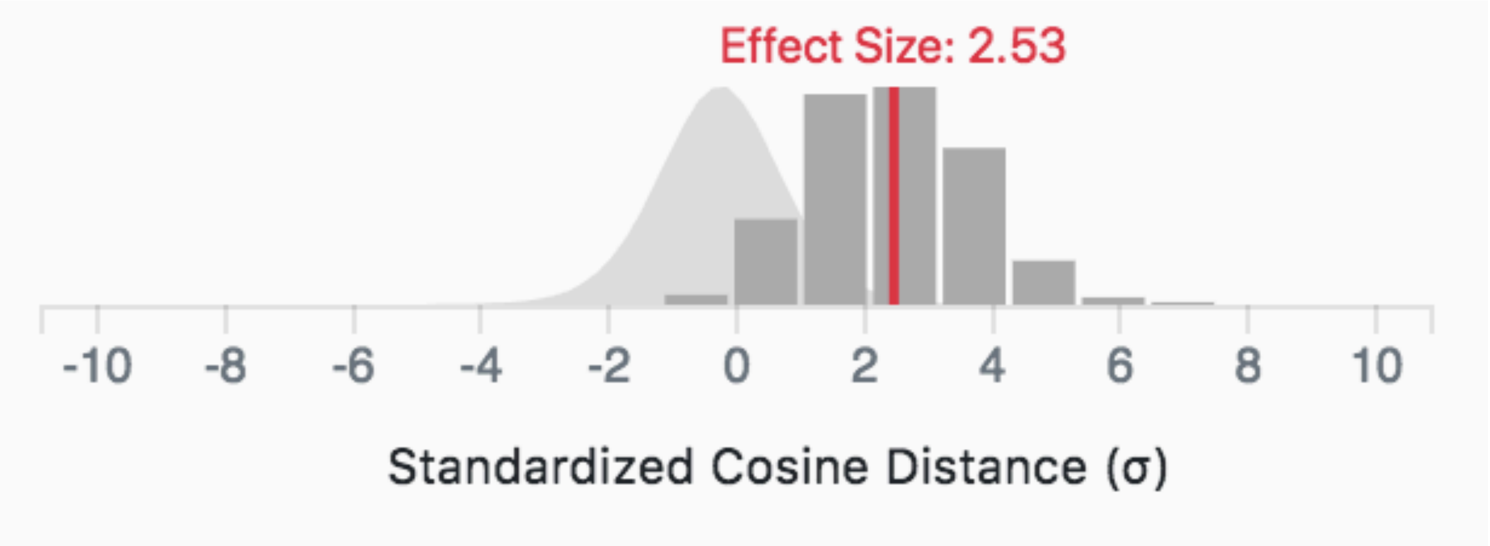
Google's analogy test score:
78.7% (472/600)

Saliency assessments agree with analogy test scores







Case Study: Cancer Transcriptomes

Details in paper

Case Study: Cancer Transcriptomes

Biological latent space

Disagree with prior work!

Our list	Their list	Agreement
Proliferative	38 -	16%
Differentiated	38 +	3%
Differentiated	79 -	0%

“ I agree that vector subtraction makes the most sense to get the full response. ”

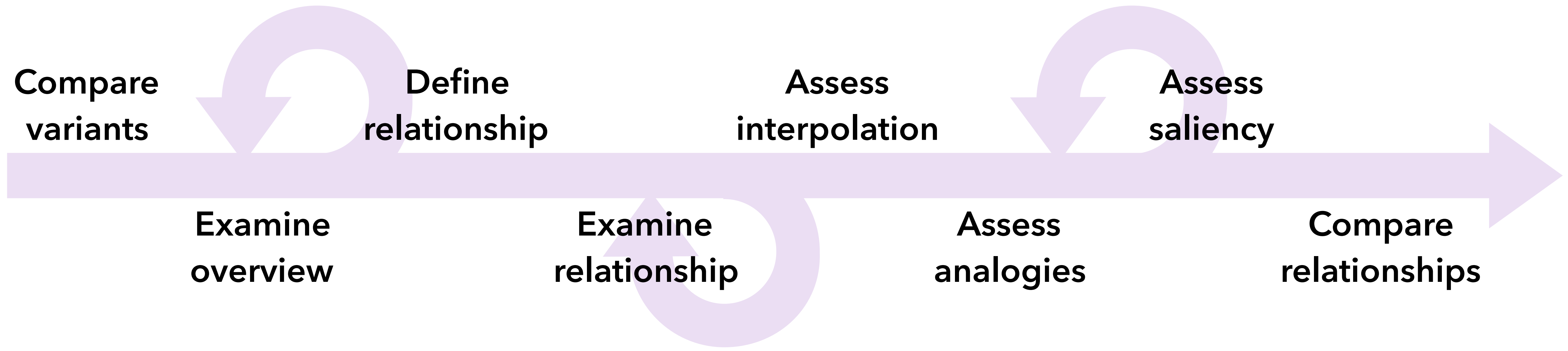
WAY G. P., GREENE C. S.: Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In Proceedings of Pacific Symposium on Biocomputing (2018), vol. 23, pp. 80–91.

Latent Space Cartography

Visual Analysis of Vector Space Embeddings

Latent Space Cartography

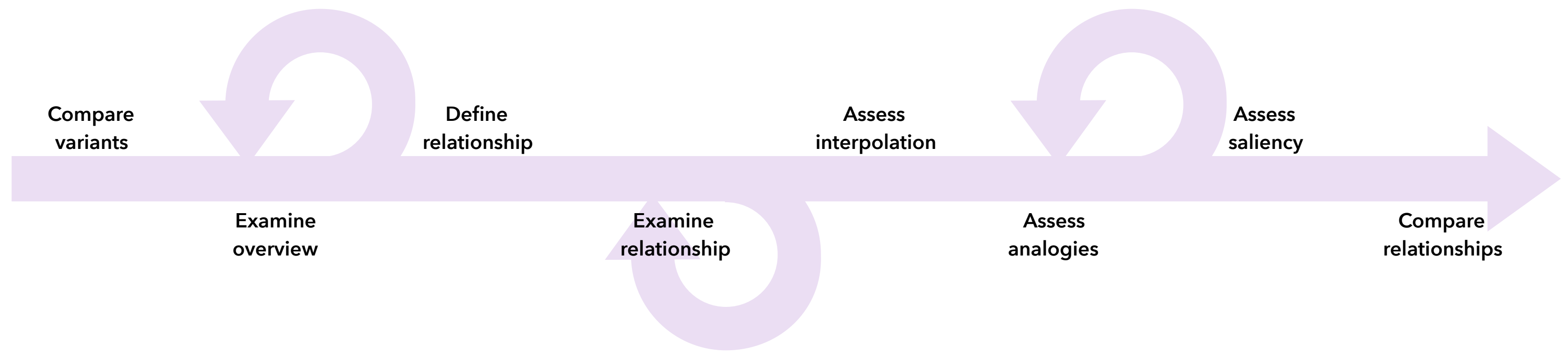
Mapping meaningful dimensions of latent spaces



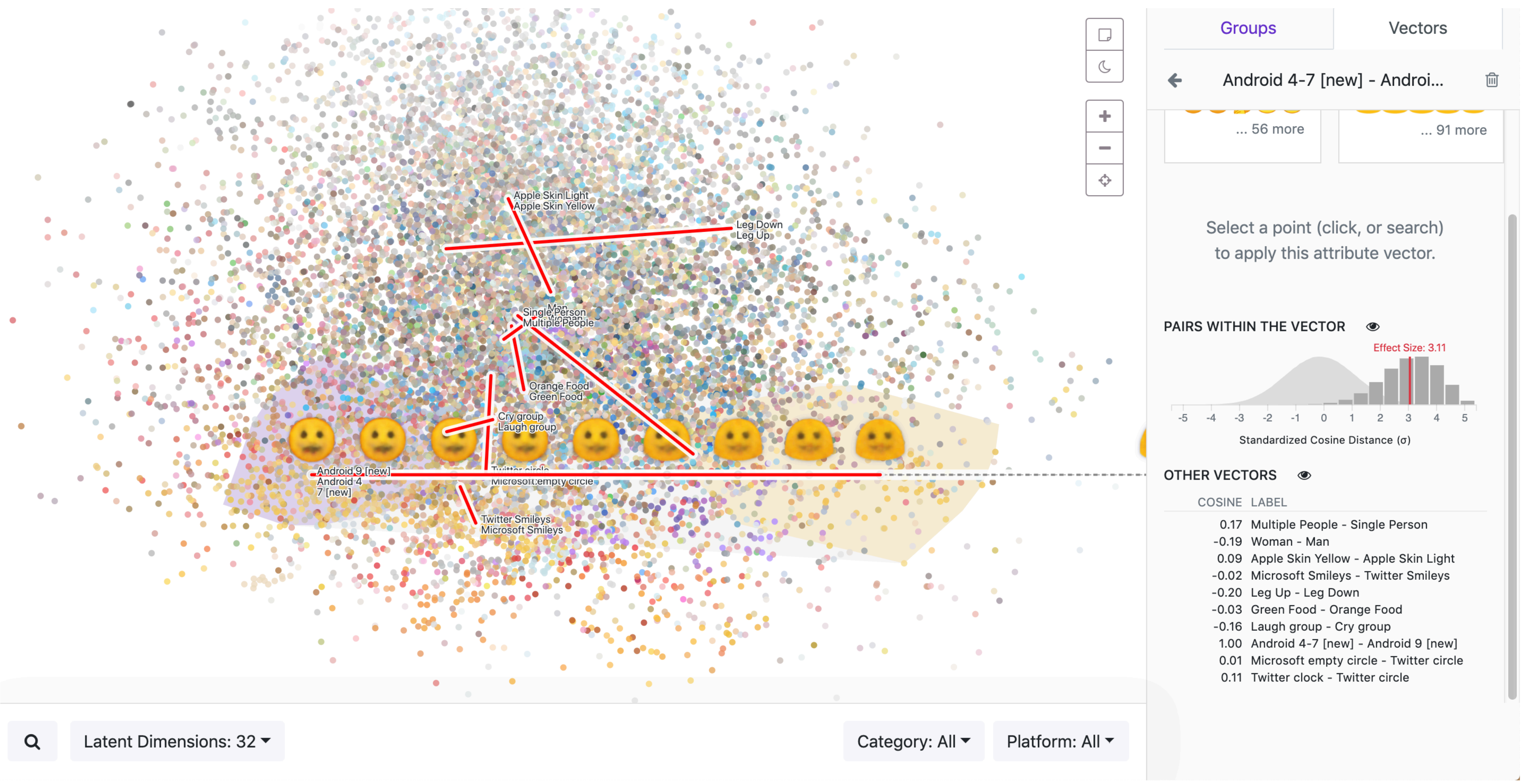
A **workflow** of interpretation tasks

Latent Space Cartography

Mapping meaningful dimensions of latent spaces



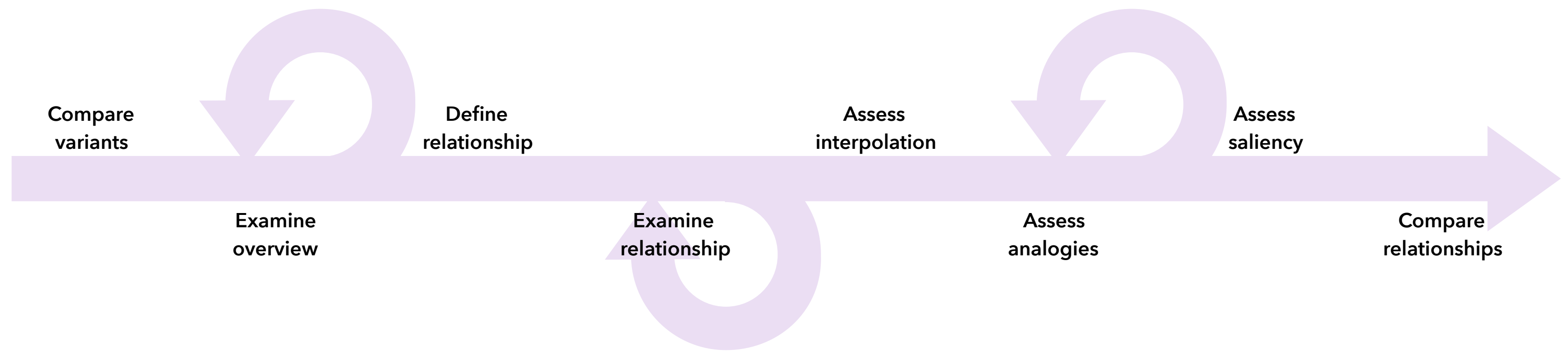
A **workflow** of interpretation tasks



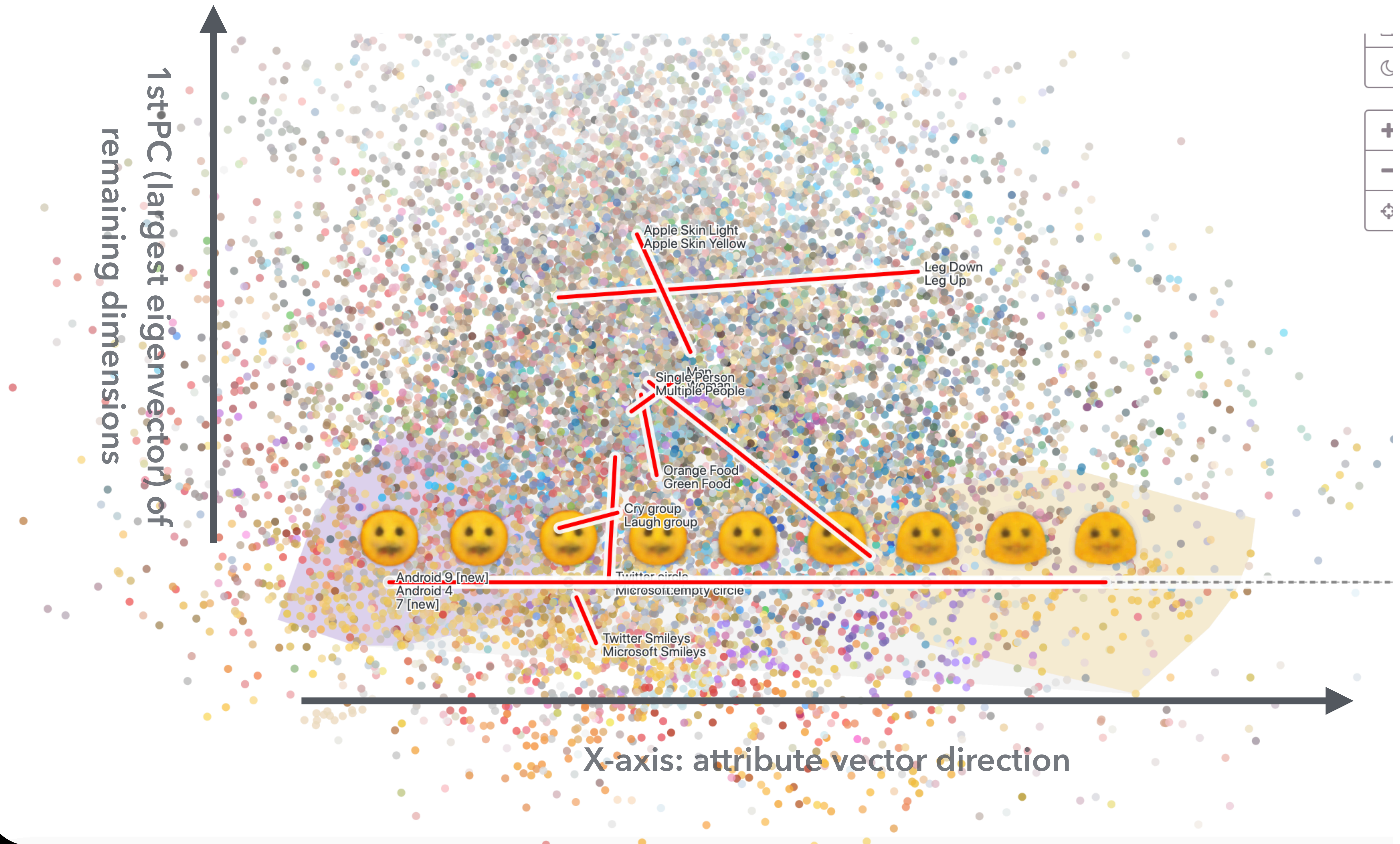
A **visual analysis system** for supporting this workflow

Latent Space Cartography

Mapping meaningful dimensions of latent spaces



A **workflow** of interpretation tasks

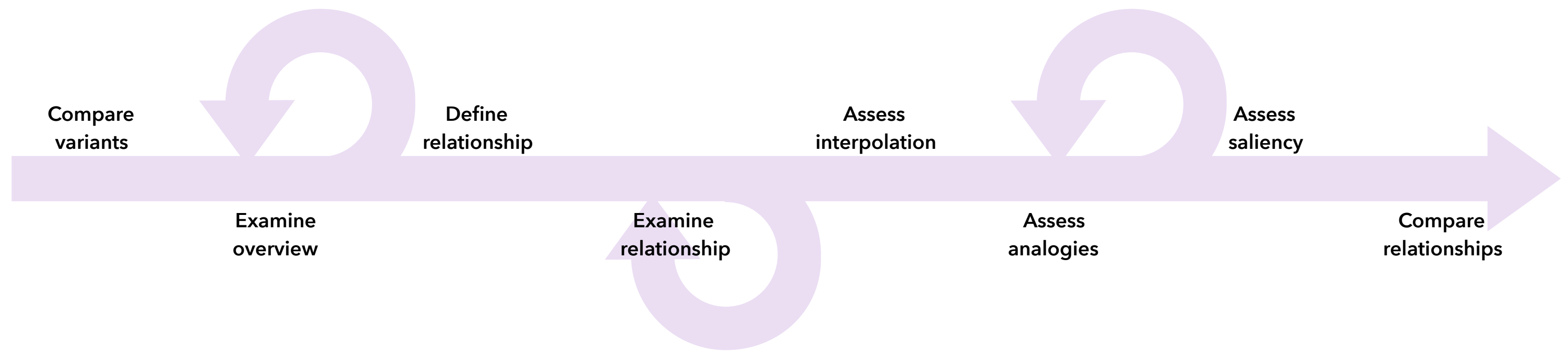


A **visual analysis system** for supporting this workflow

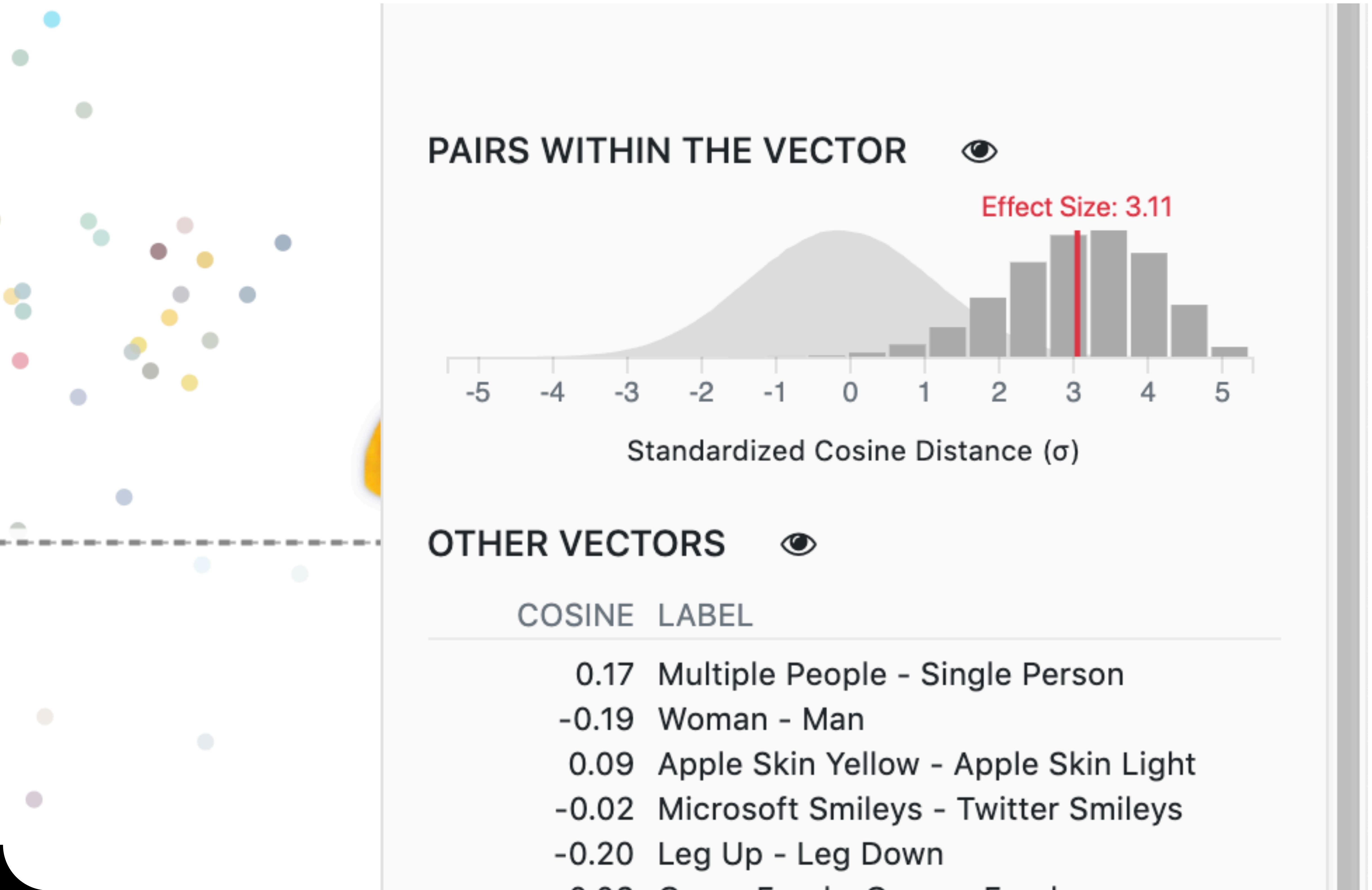
- Linear projection to provide context

Latent Space Cartography

Mapping meaningful dimensions of latent spaces



A **workflow** of interpretation tasks



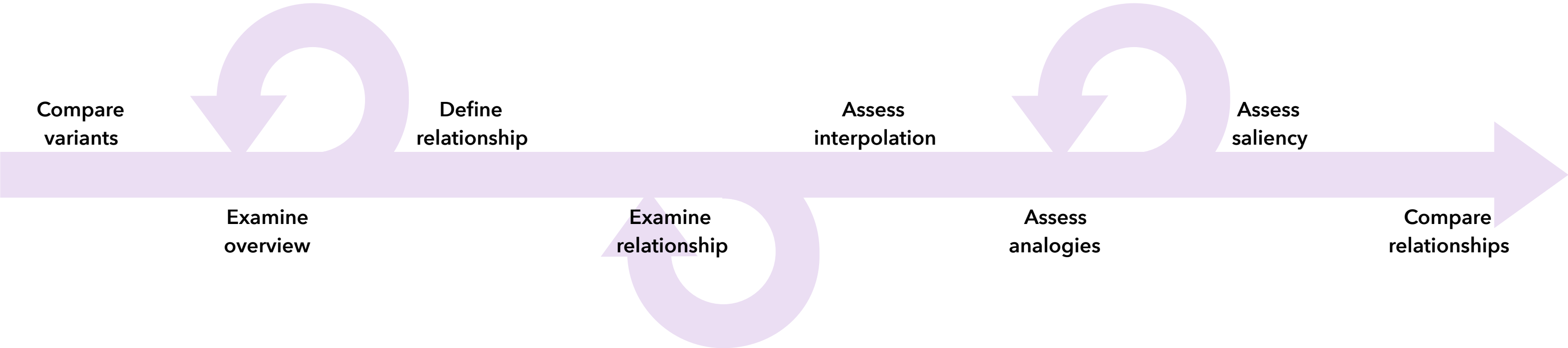
A **visual analysis system** for supporting this workflow

- Linear projection to provide context

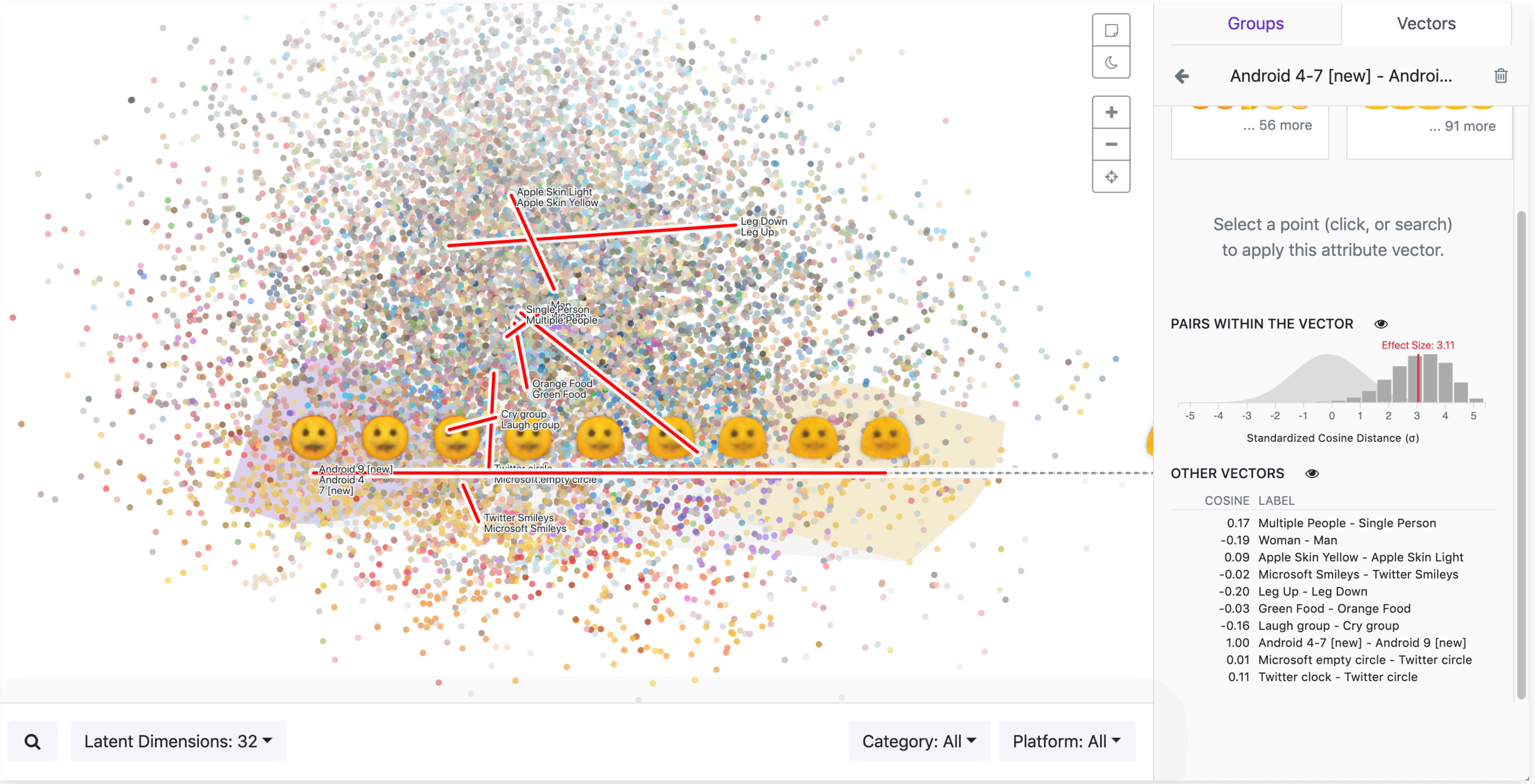
- Methods to assess vector saliency

Latent Space Cartography

Mapping meaningful dimensions of latent spaces



A **workflow** of interpretation tasks

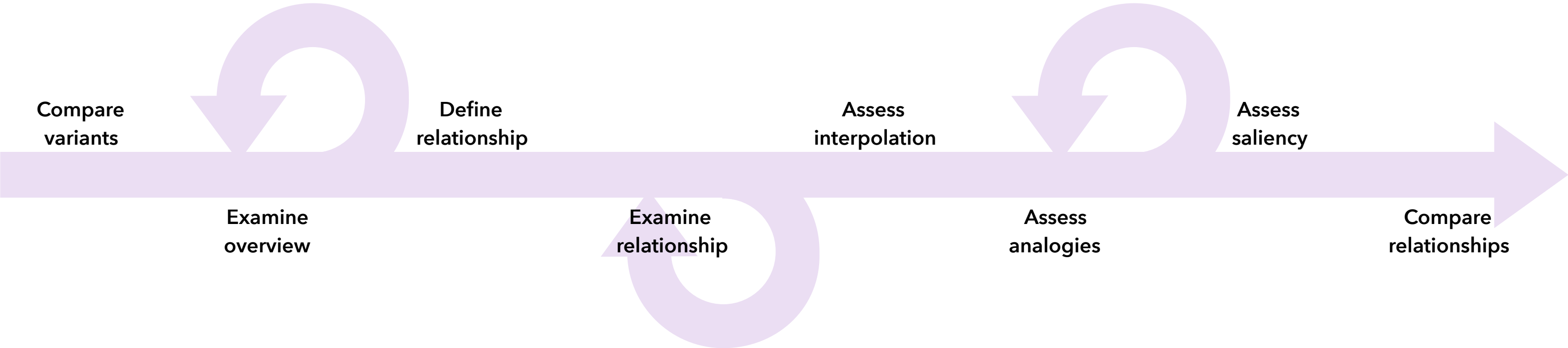


A **visual analysis system** for supporting this workflow

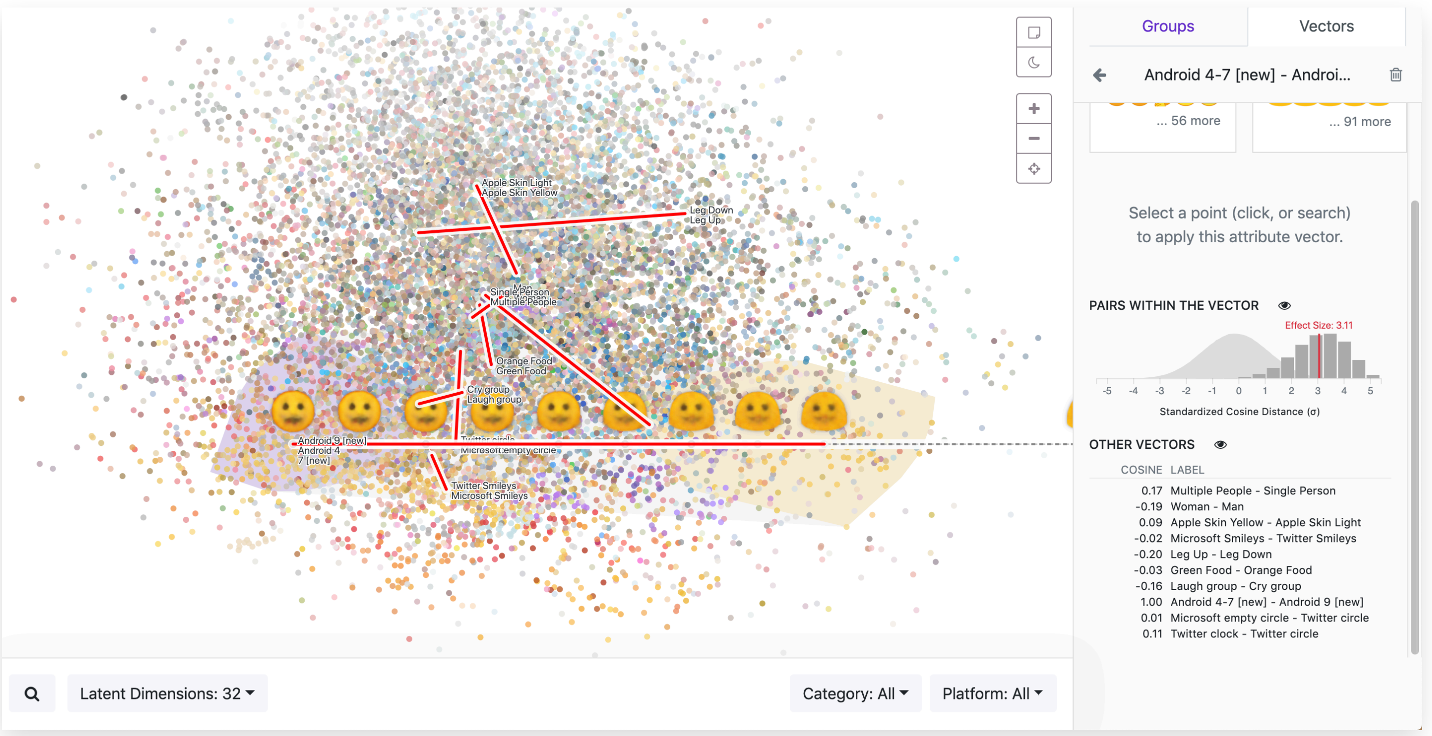
- Linear projection to provide context
- Methods to assess vector saliency
- Methods to compare multiple vectors

Latent Space Cartography

Mapping meaningful dimensions of latent spaces



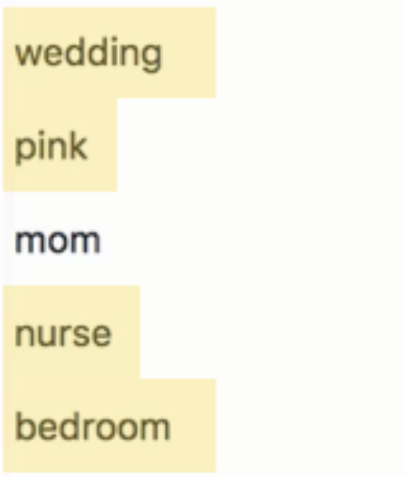
A **workflow** of interpretation tasks



A **visual analysis system** for supporting this workflow

Case studies across multiple domains

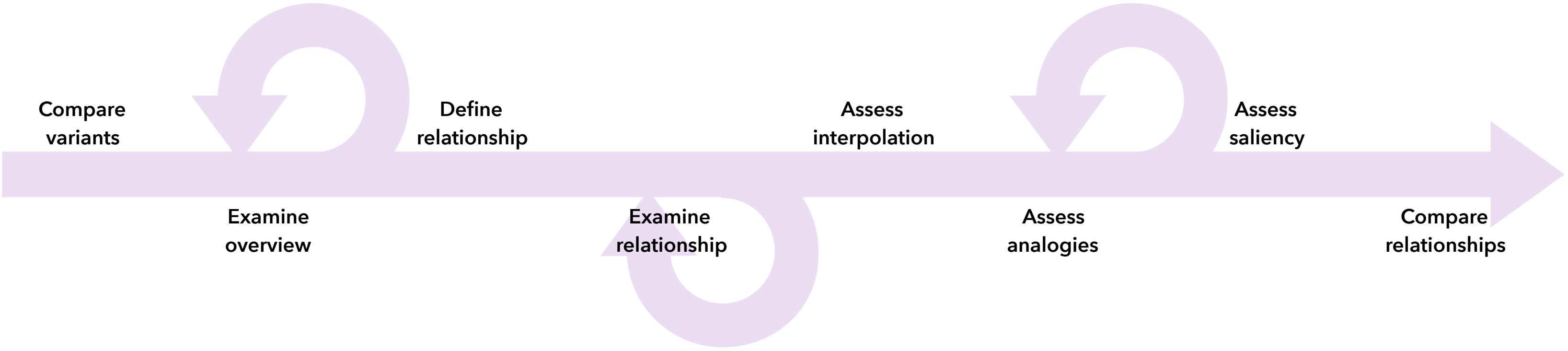
Scientific findings on cancer gene expression



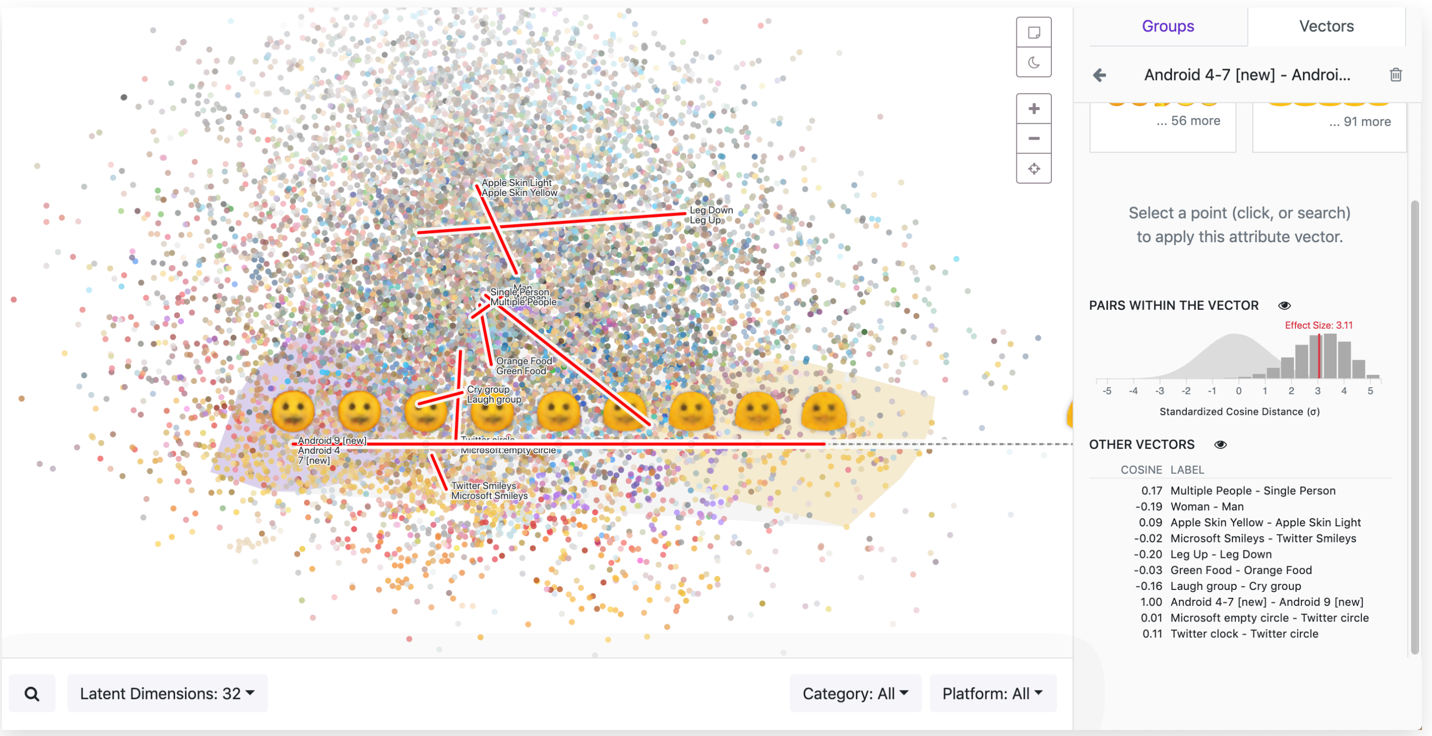
Gender stereotypes in word embeddings

Latent Space Cartography

Mapping meaningful dimensions of latent spaces



A **workflow** of interpretation tasks



A **visual analysis system** for supporting this workflow

Case studies across multiple domains

Available at: https://github.com/uwdata/latent-space-cartography

Latent Space Cartography:

Visual Analysis of Vector Space Embeddings

Yang Liu, Eunice Jun, Qisheng Li, Jeffrey Heer
University of Washington
Interactive Data Lab

https://github.com/uwdata/latent-space-cartography

DESIGN USE BUILD