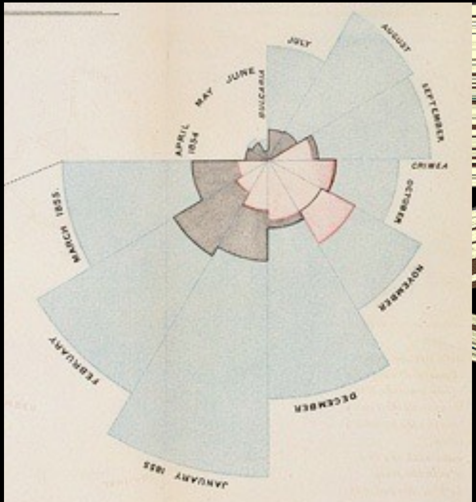
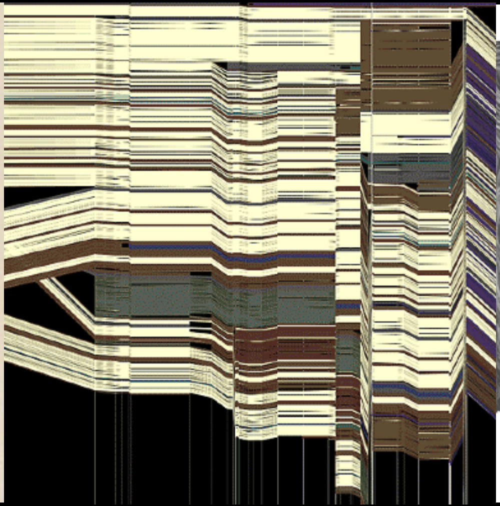
CSE 442 - Data Visualization

Dimensionality Reduction





Jeffrey Heer University of Washington

Dimensionality Reduction

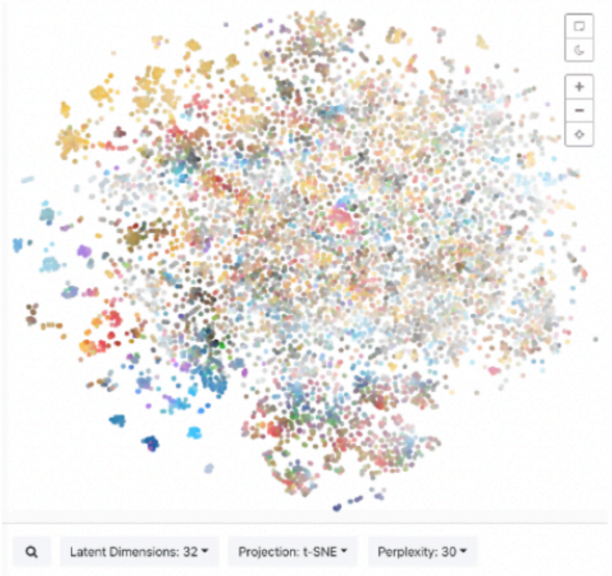
Dimensionality Reduction (DR)

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

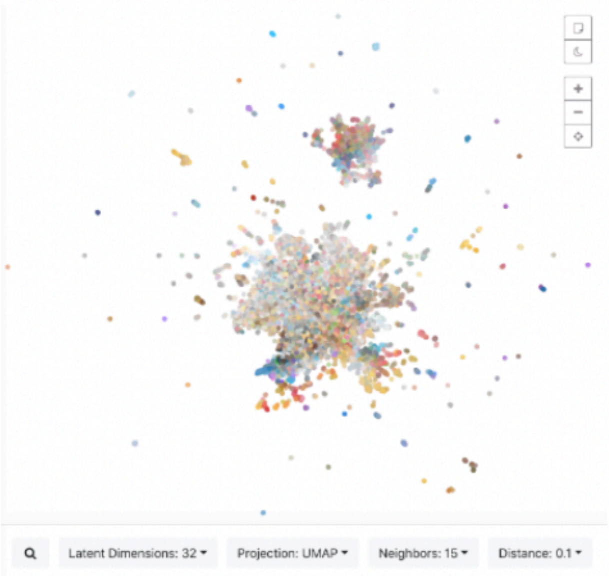
Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

In contrast, multidimensional scaling (MDS) attempts to **preserve pairwise distances**.

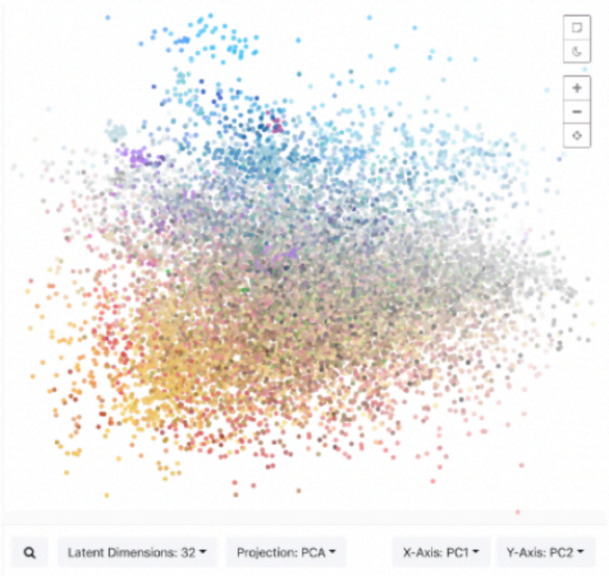
Mapping Emoji Images



t-SNE



UMAP



PCA

Reduction Techniques

LINEAR - PRESERVE GLOBAL STRUCTURE

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

NON-LINEAR - PRESERVE LOCAL TOPOLOGY

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

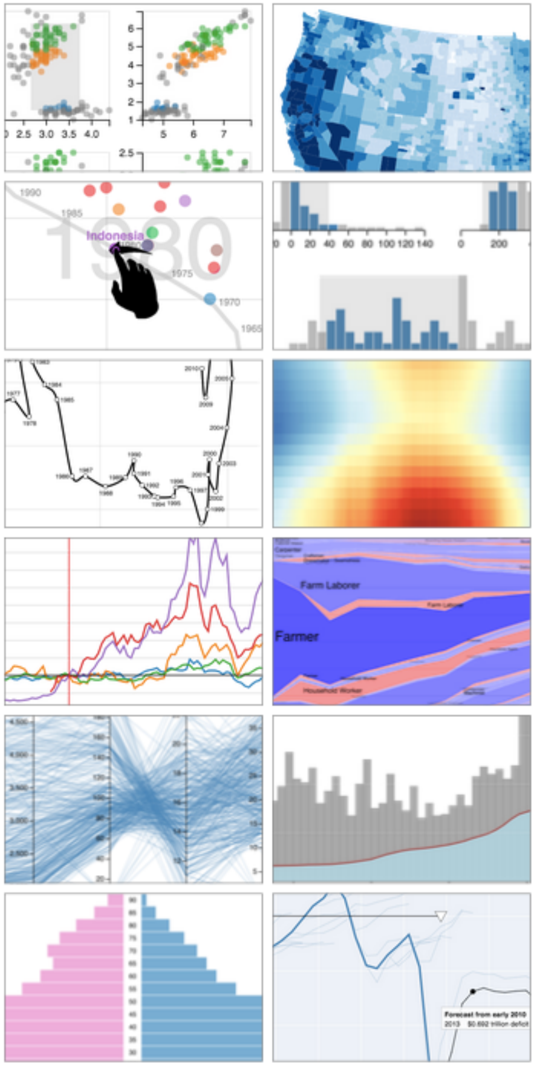
Uniform Manifold Approx. & Projection (UMAP)
Identify local manifolds, then stitch them together.

Dimensionality Reduction Issues

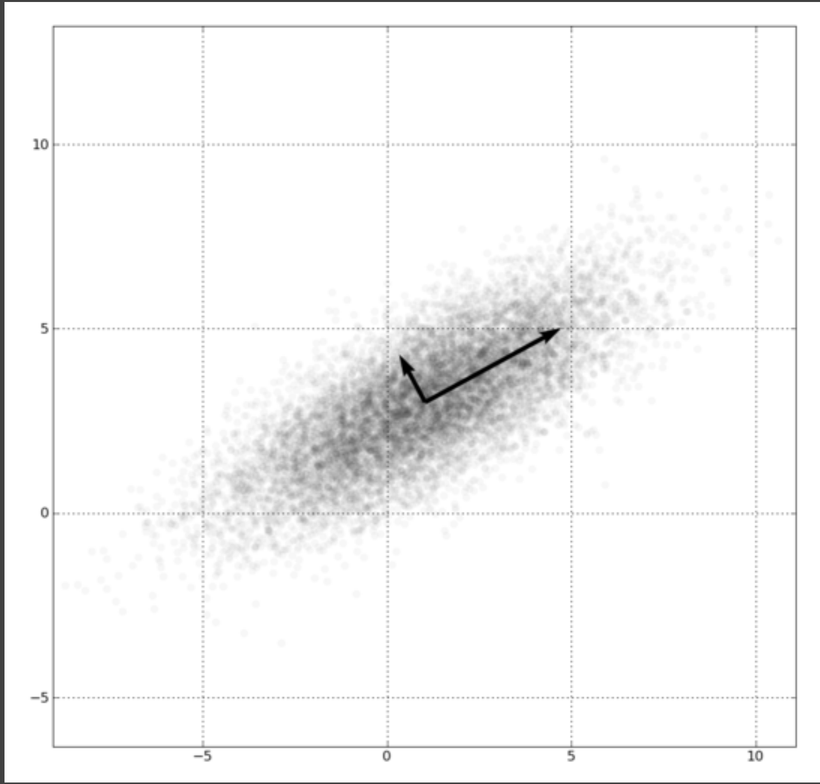
Reproducible?
Projections are *data-dependent*. Fitting a new projection with different data can give rise to different results.

Reusable?
PCA and UMAP provide reusable projection functions that can map new points from high-D to low-D. t-SNE (and others, like MDS) do not provide this.

Interpretable?
DR plots are hard to interpret! Try multiple methods and hyperparameter settings. Inspect via interaction!

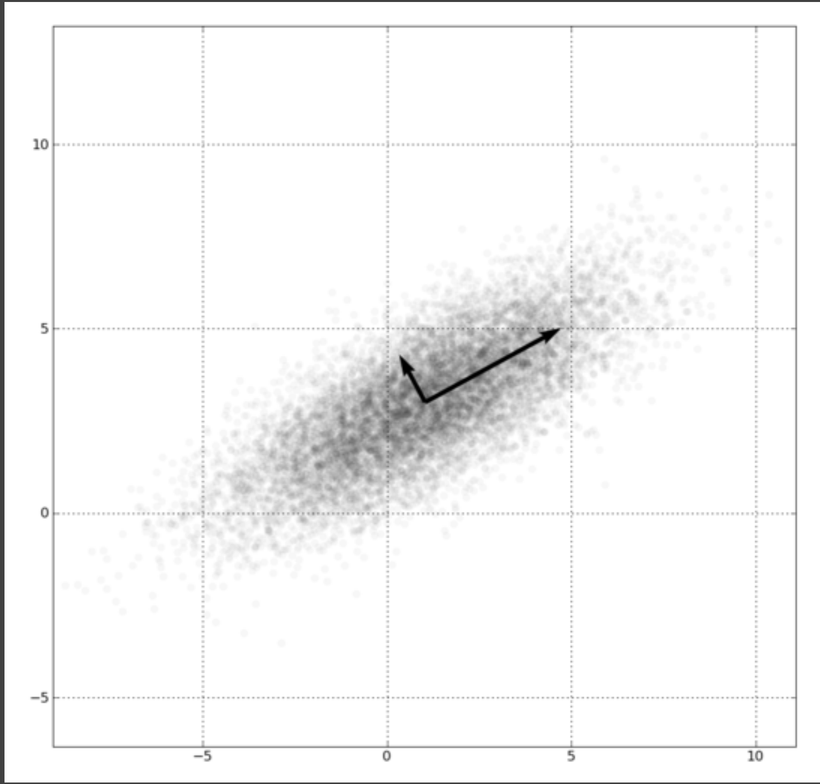


Principal Components Analysis



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

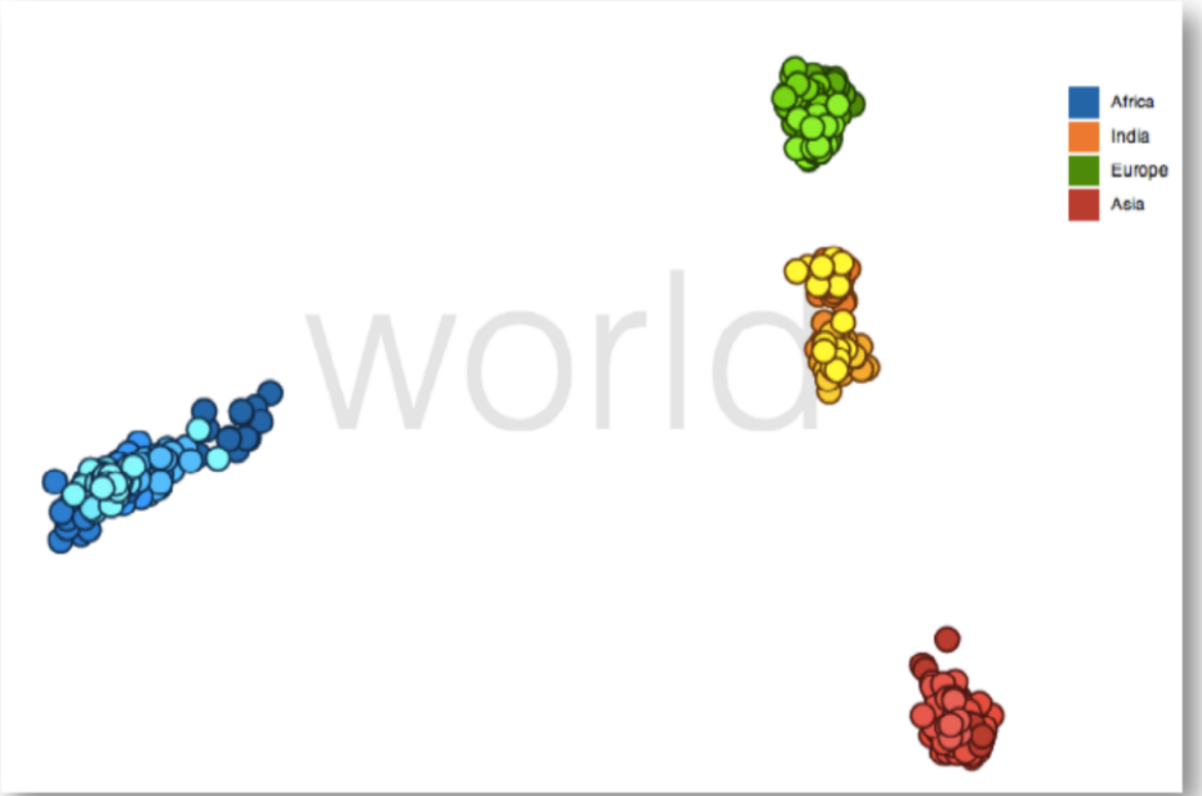


Linear transform: scale and rotate original space.

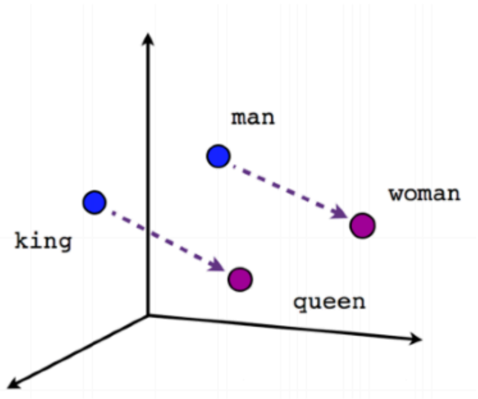
Lines (vectors) project to lines.

Preserves global distances.

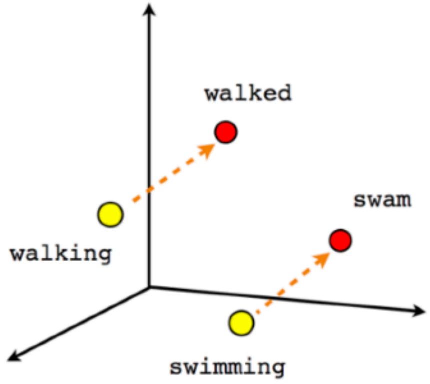
PCA of Genomes [Demiralp et al. '13]



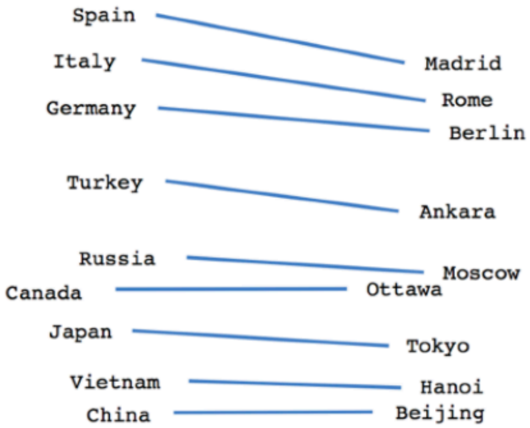
Vector Space Word Embeddings (word2vec, GloVe)



Male-Female



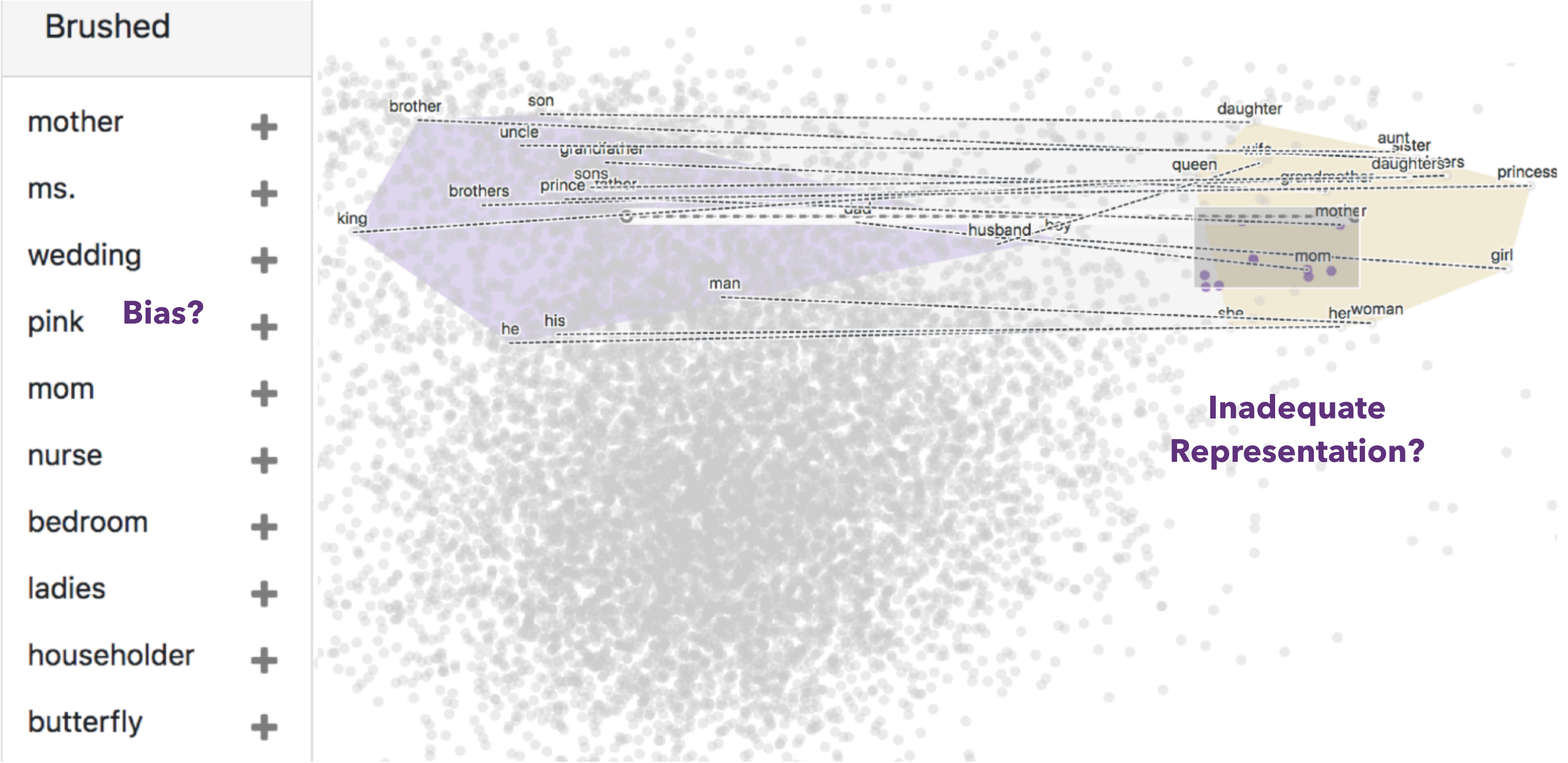
Verb tense



Country-Capital

Mapping Machine-Learned Latent Spaces

[Liu et al. 2019]



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

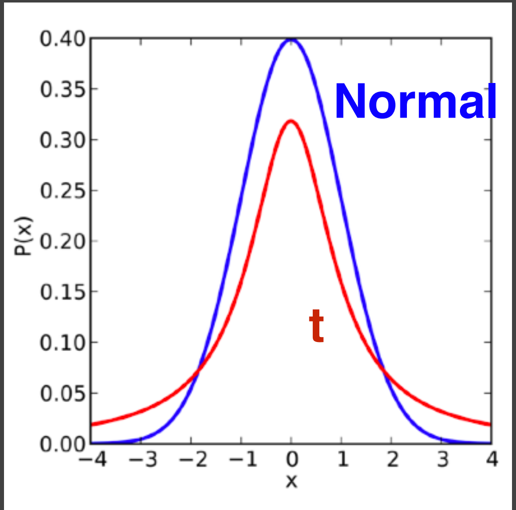
UMAP - tries to balance local/global trade-off

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]

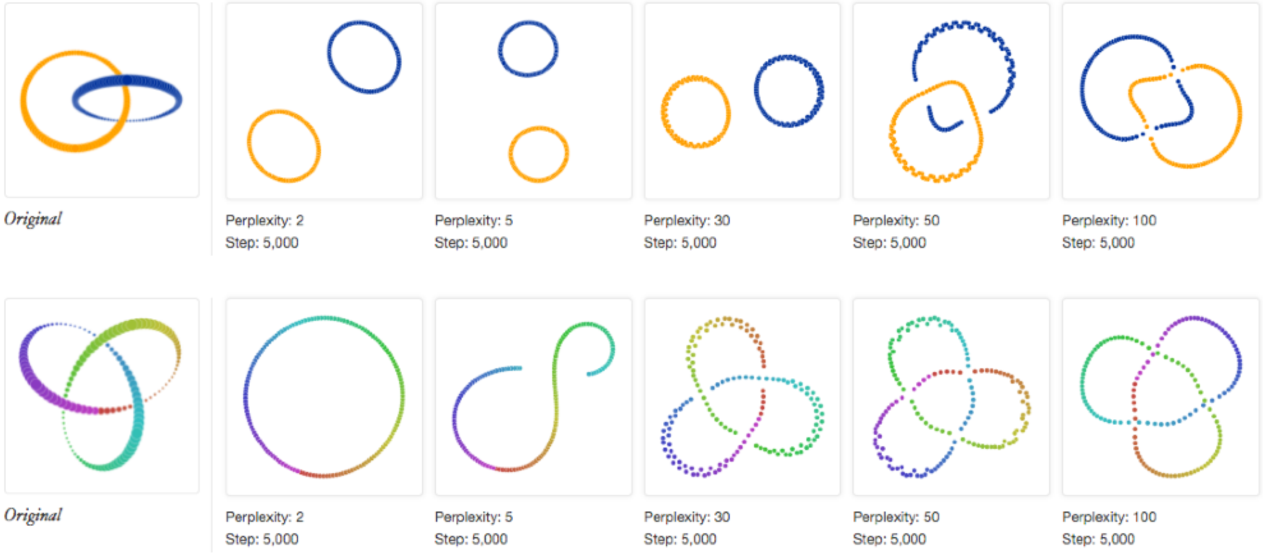
2. Define a similar probability **Q** in the low-dimensional (2D or 3D) embedding space, using a Student's *t* distribution *(hence the "t-" in "t-SNE"!)*. The *t*-distribution is heavy-tailed, allowing distant points to be even further apart.



t-SNE [Maaten & Hinton 2008]

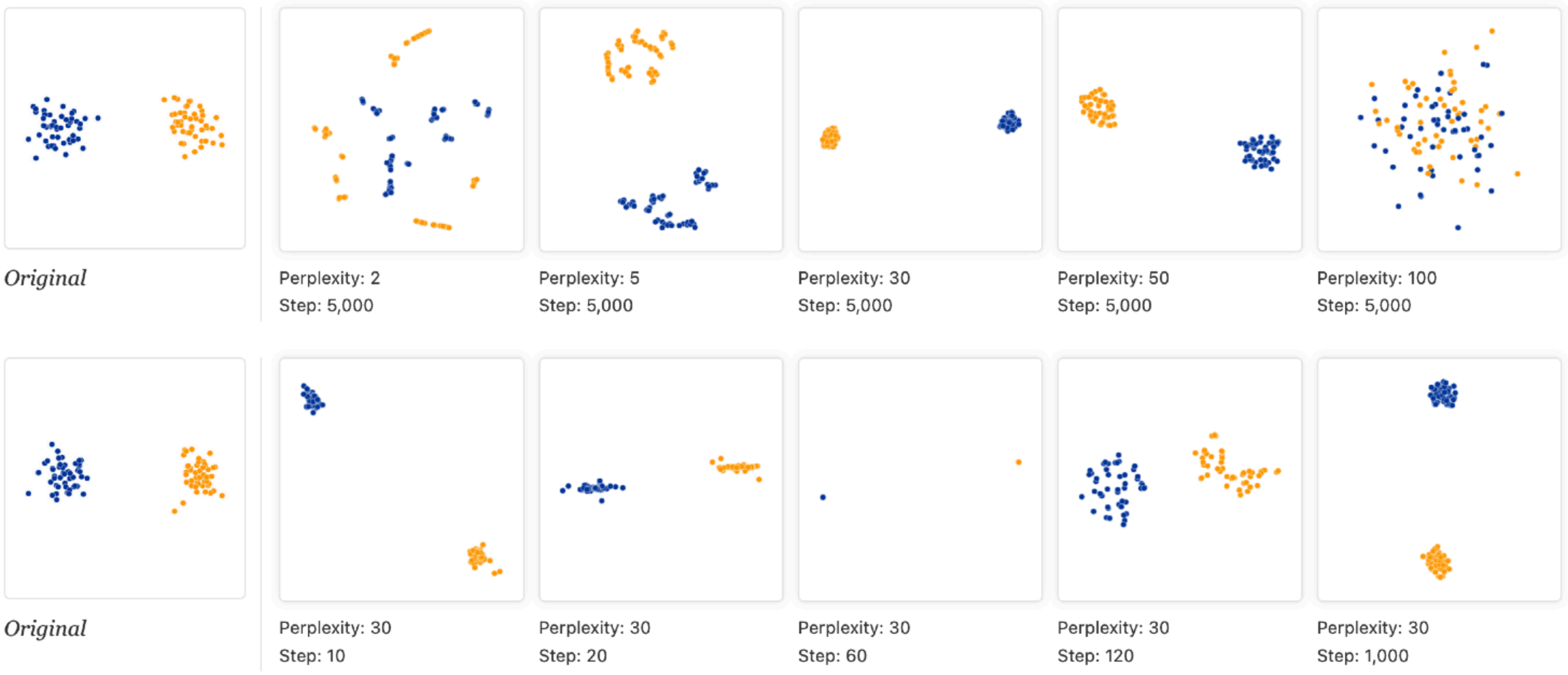
1. Model probability **P** of one point "choosing" another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
2. Define a similar probability **Q** in the low-dimensional (2D or 3D) embedding 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: $KL(P \,||\, Q)$

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

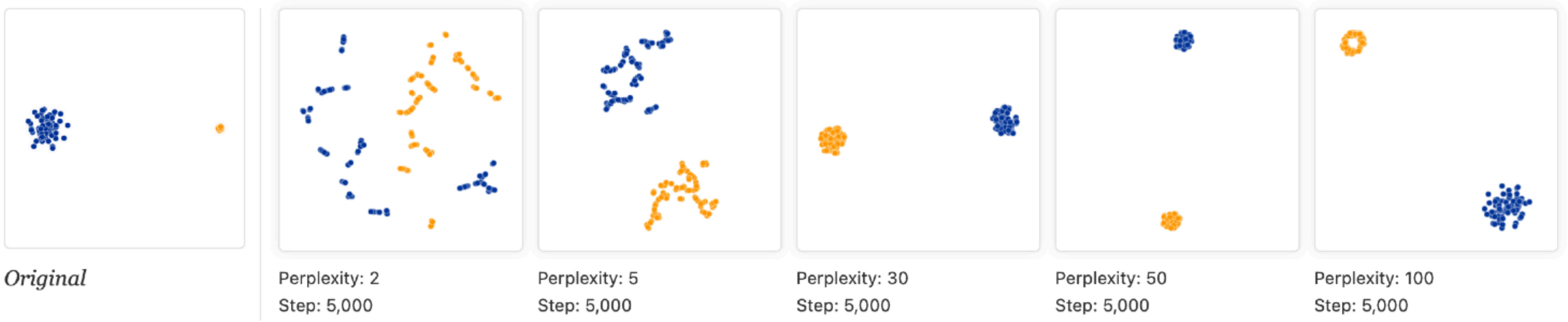


Results can be highly sensitive to the algorithm parameters!
Are you seeing real structures, or algorithmic hallucinations?

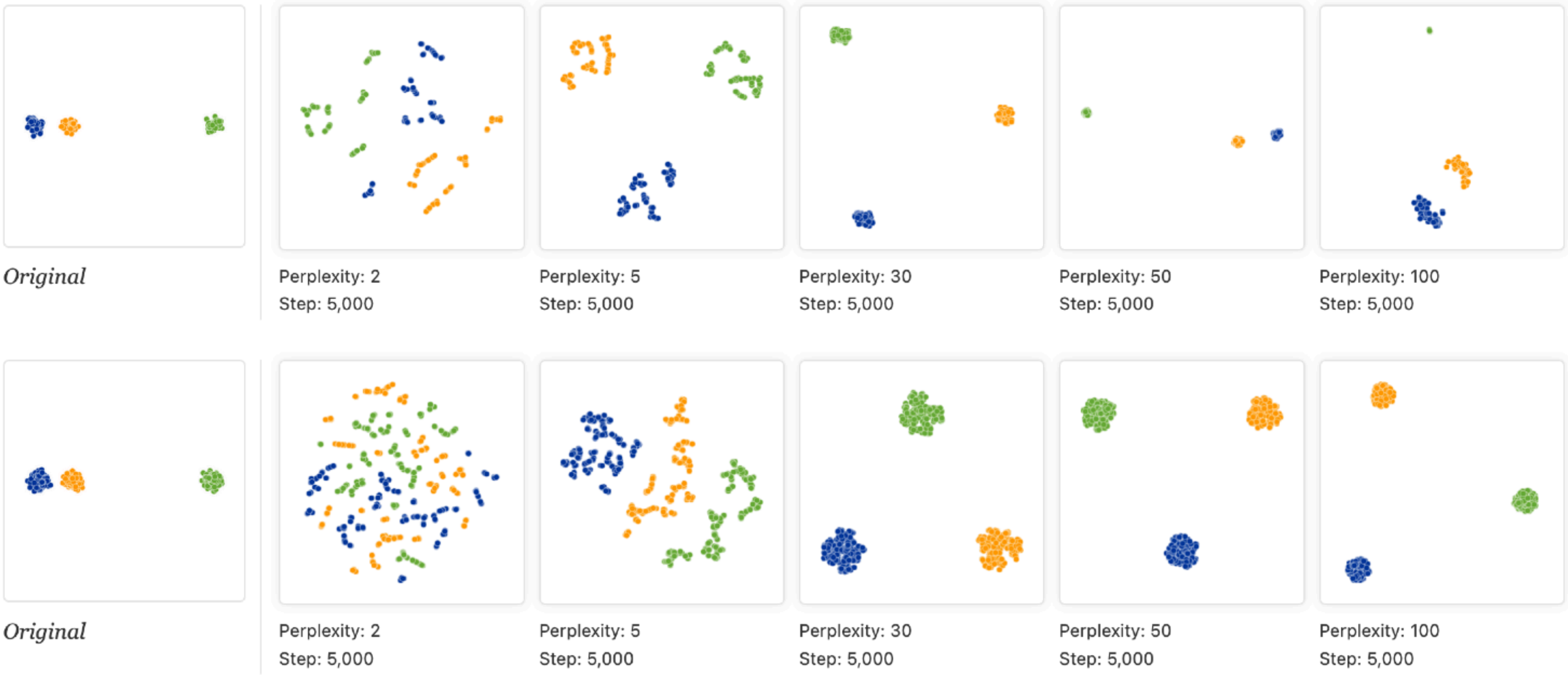
Hyperparameters matter!



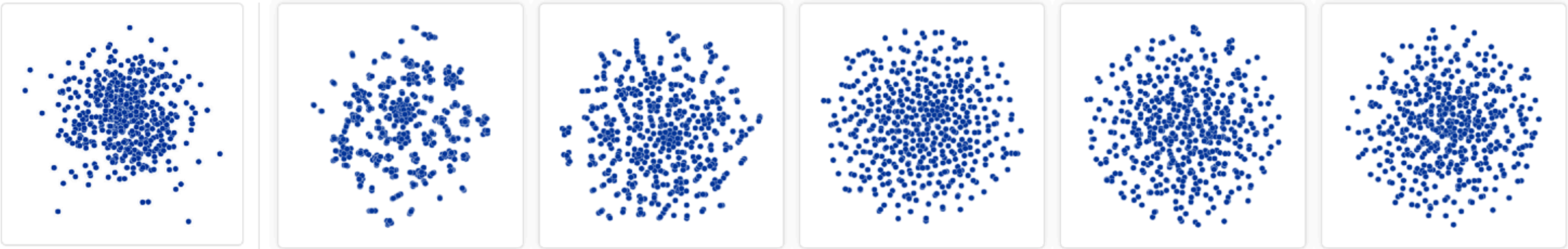
Cluster sizes mean nothing…



Cluster distances are illusive



Random noise may not look like it



Original

Perplexity: 2
Step: 5,000

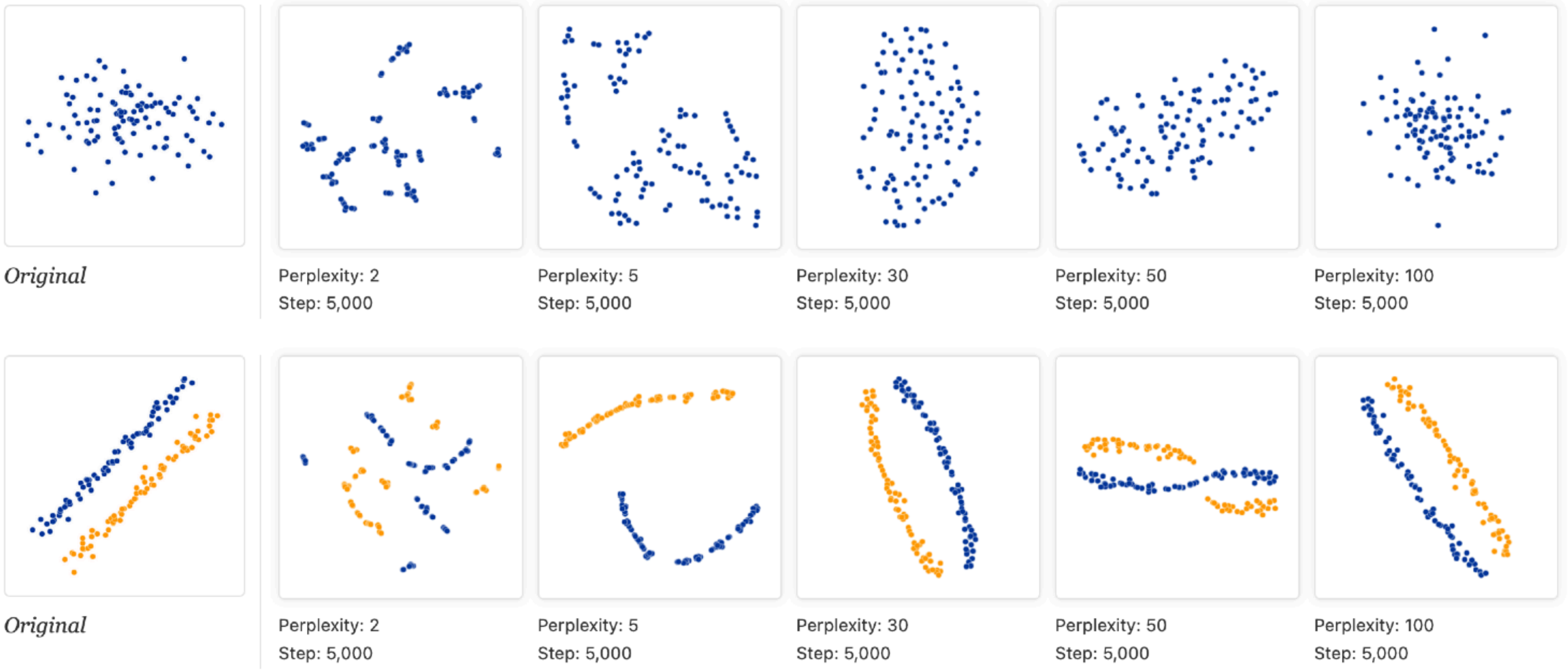
Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

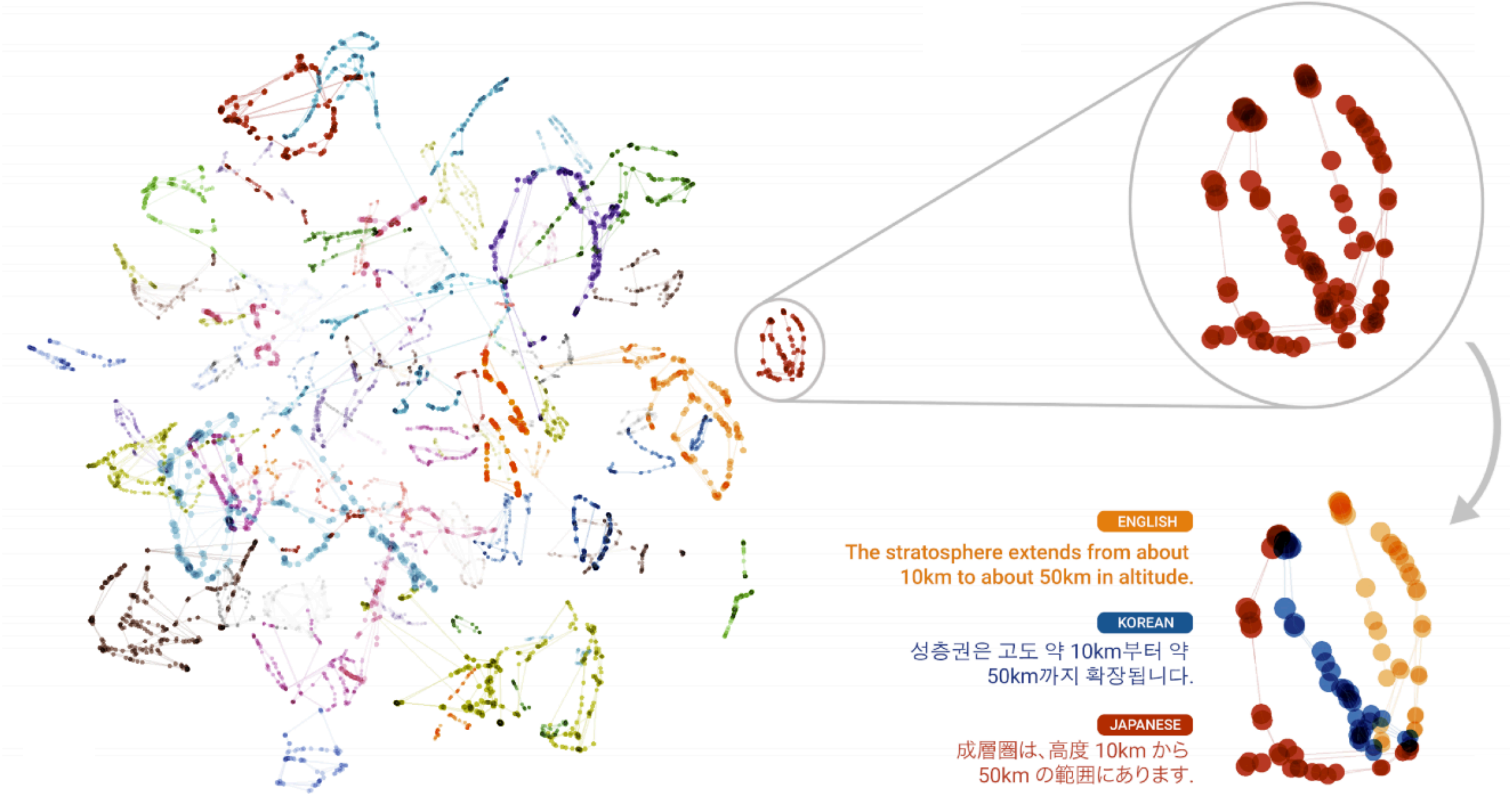
Perplexity: 50
Step: 5,000

Perplexity: 100
Step: 5,000

You can see shapes, sometimes



Multi-Lingual Word Embedding [Google 2016]



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

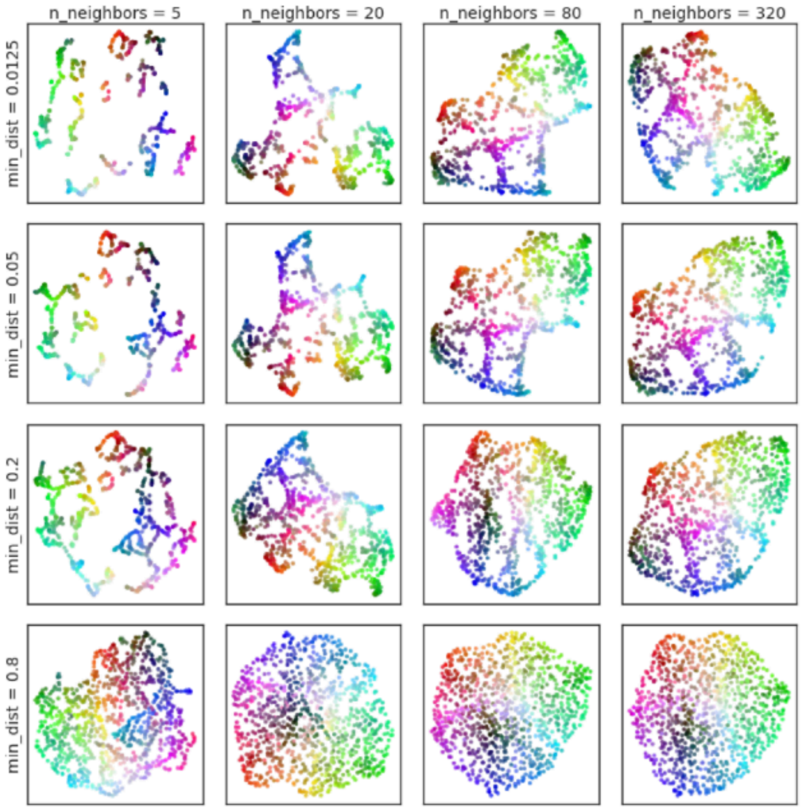


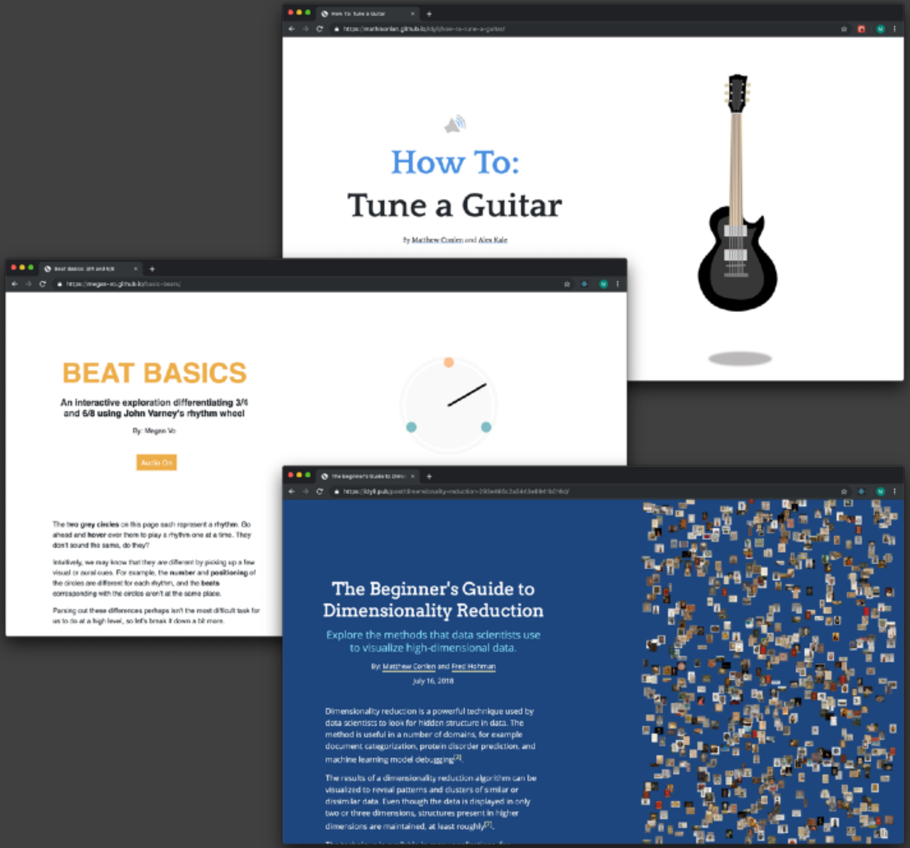
Figure 1: Variation of UMAP hyperparameters n and min-dist result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values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

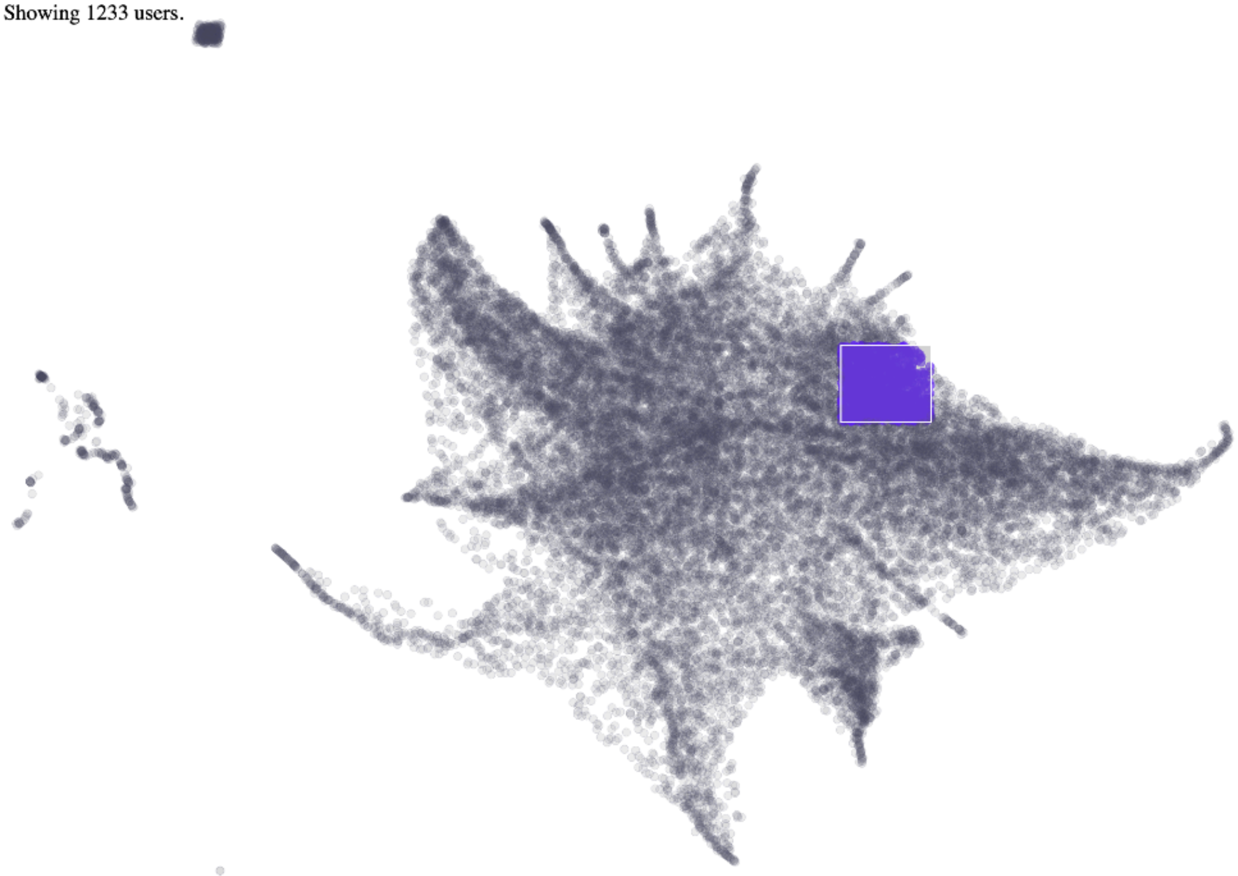
User Engagement with Interactive Articles

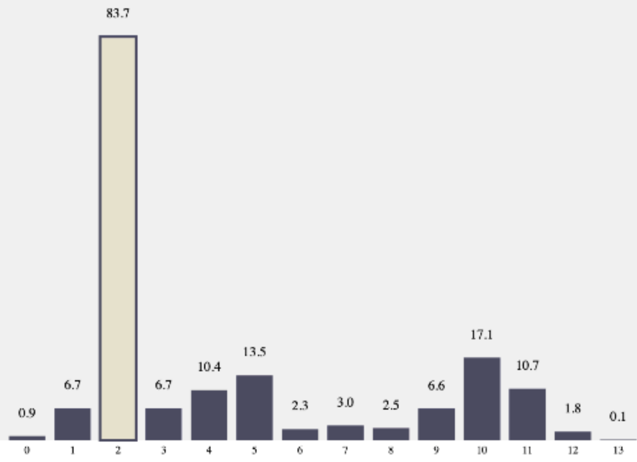
Provide an overview of usage patterns of interactive features.

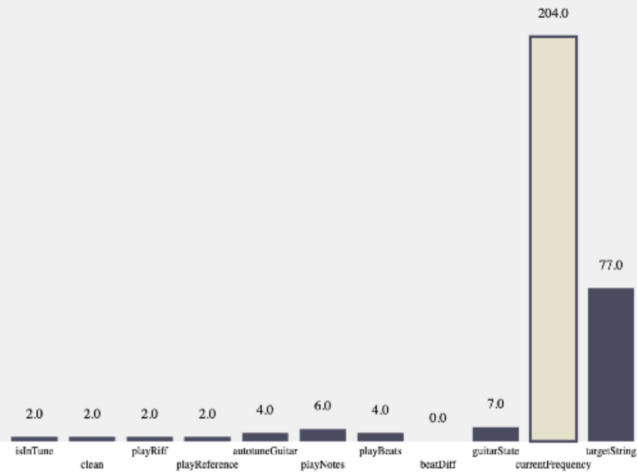
Identify variations in usage

Represent reader sessions as a feature vector with:
- time spent in each section
- count of variable changes



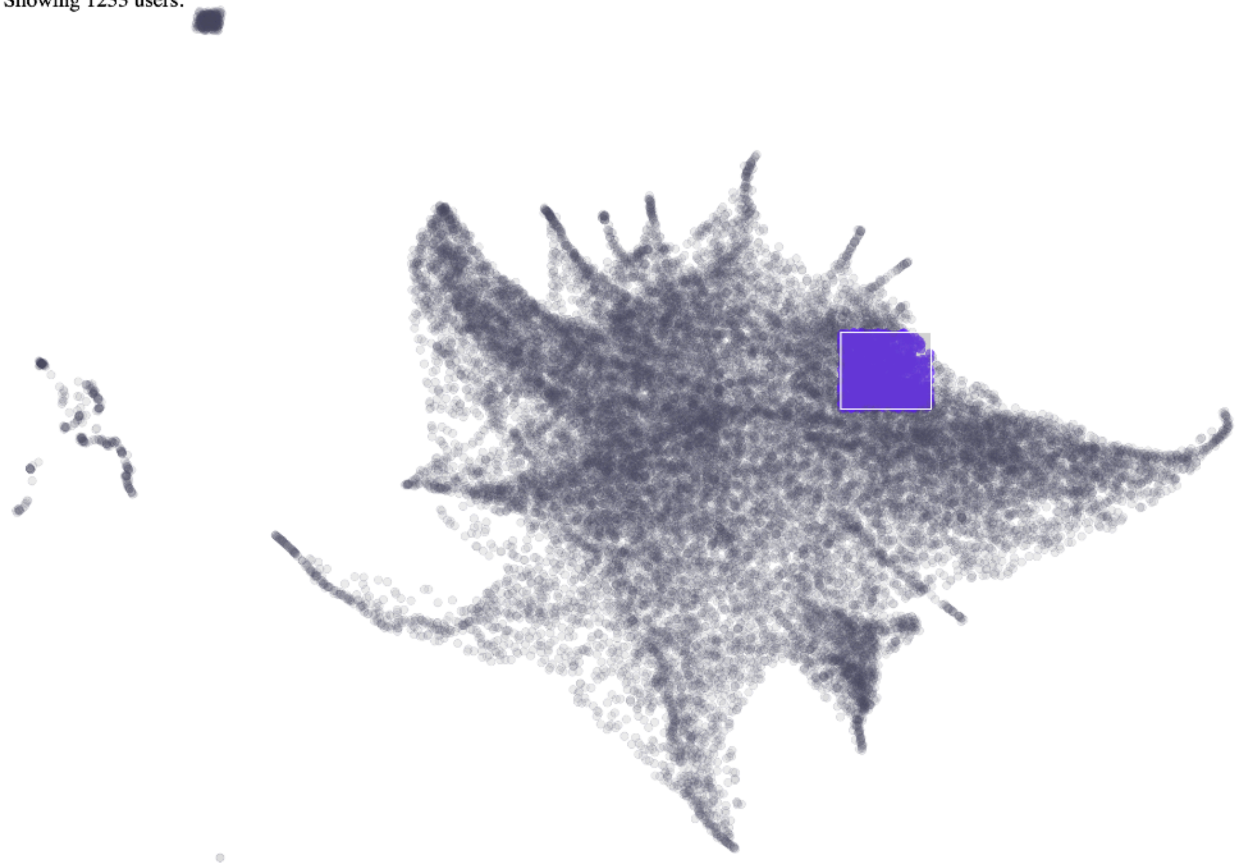


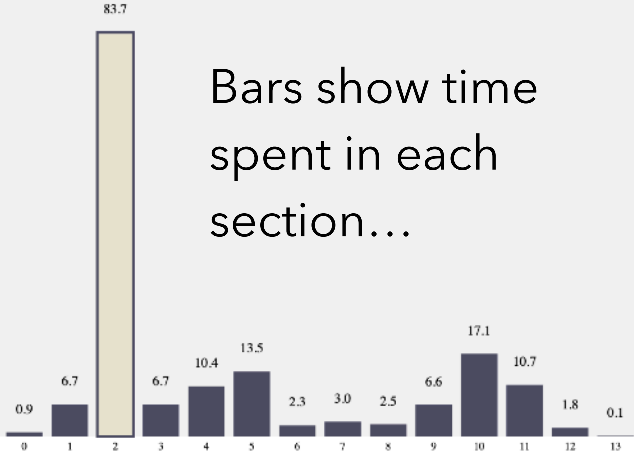




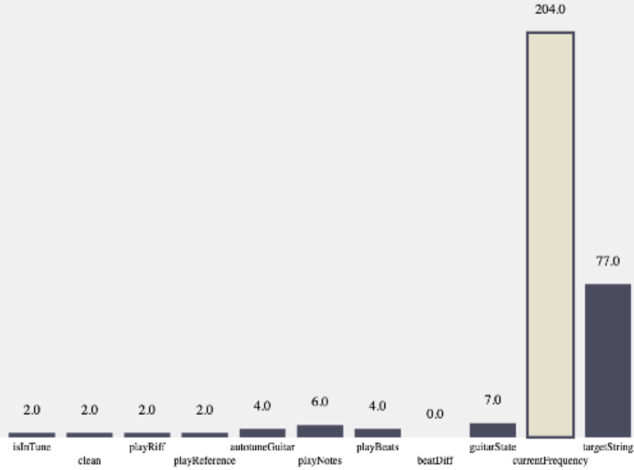
Each point represents a readers session, projected via UMAP.

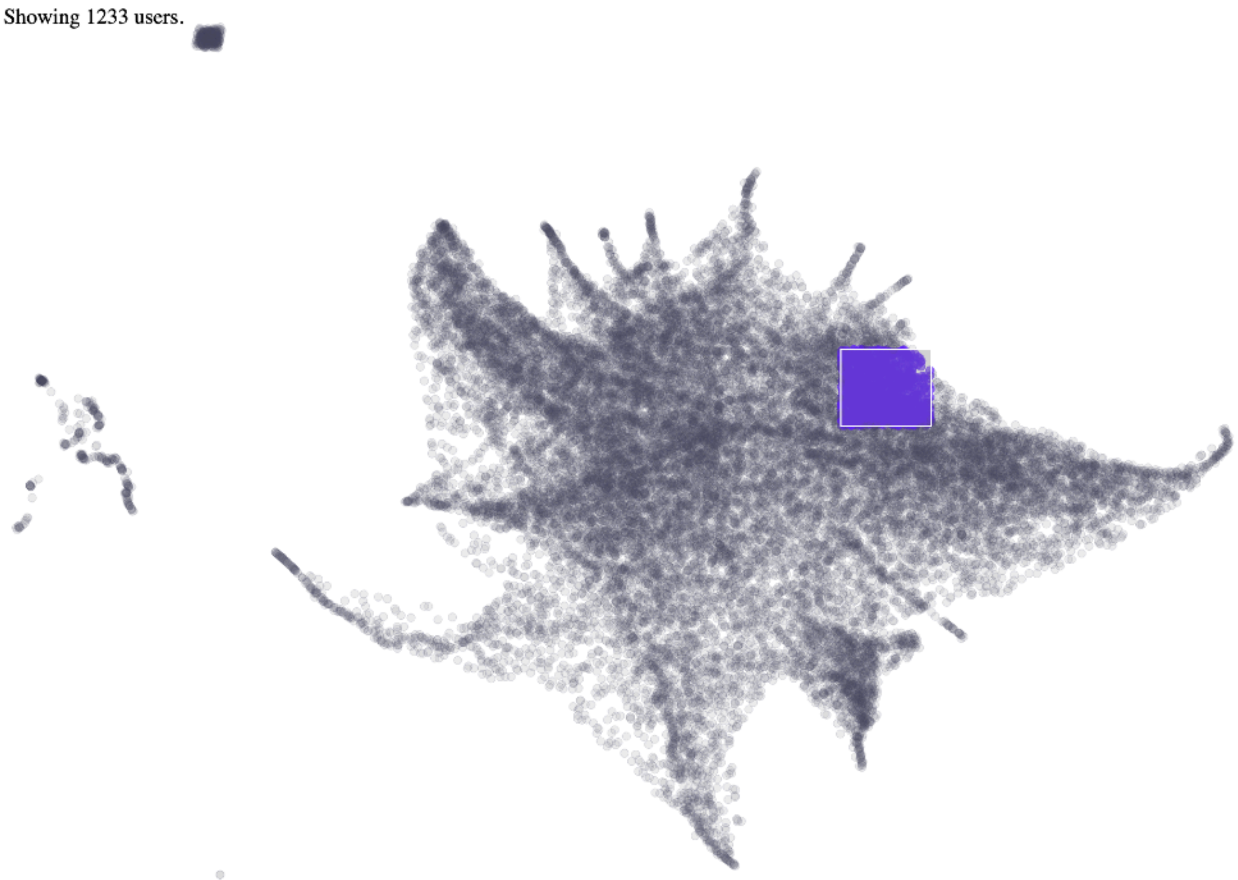
Showing 1233 users.

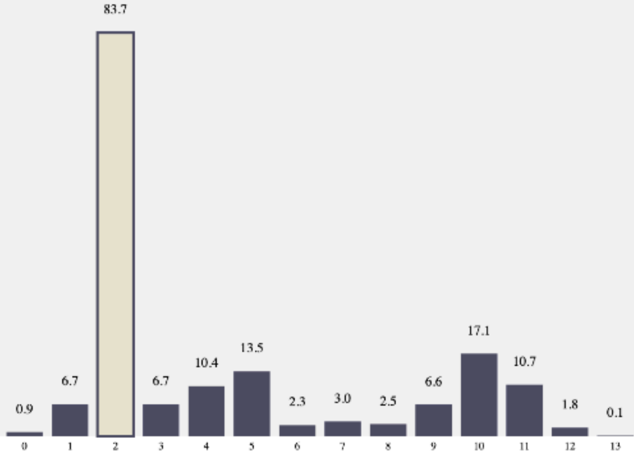




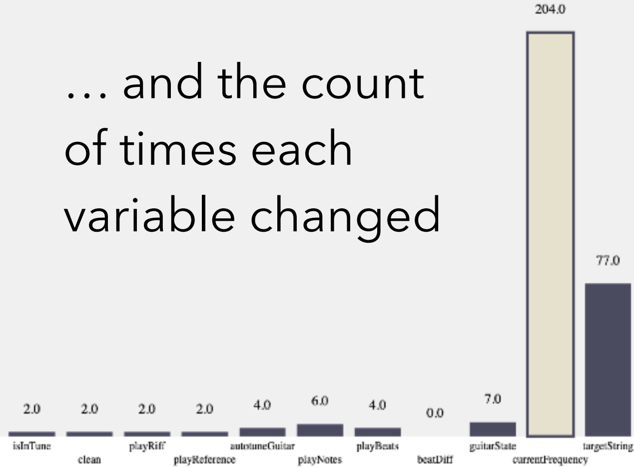
Bars show time spent in each section…



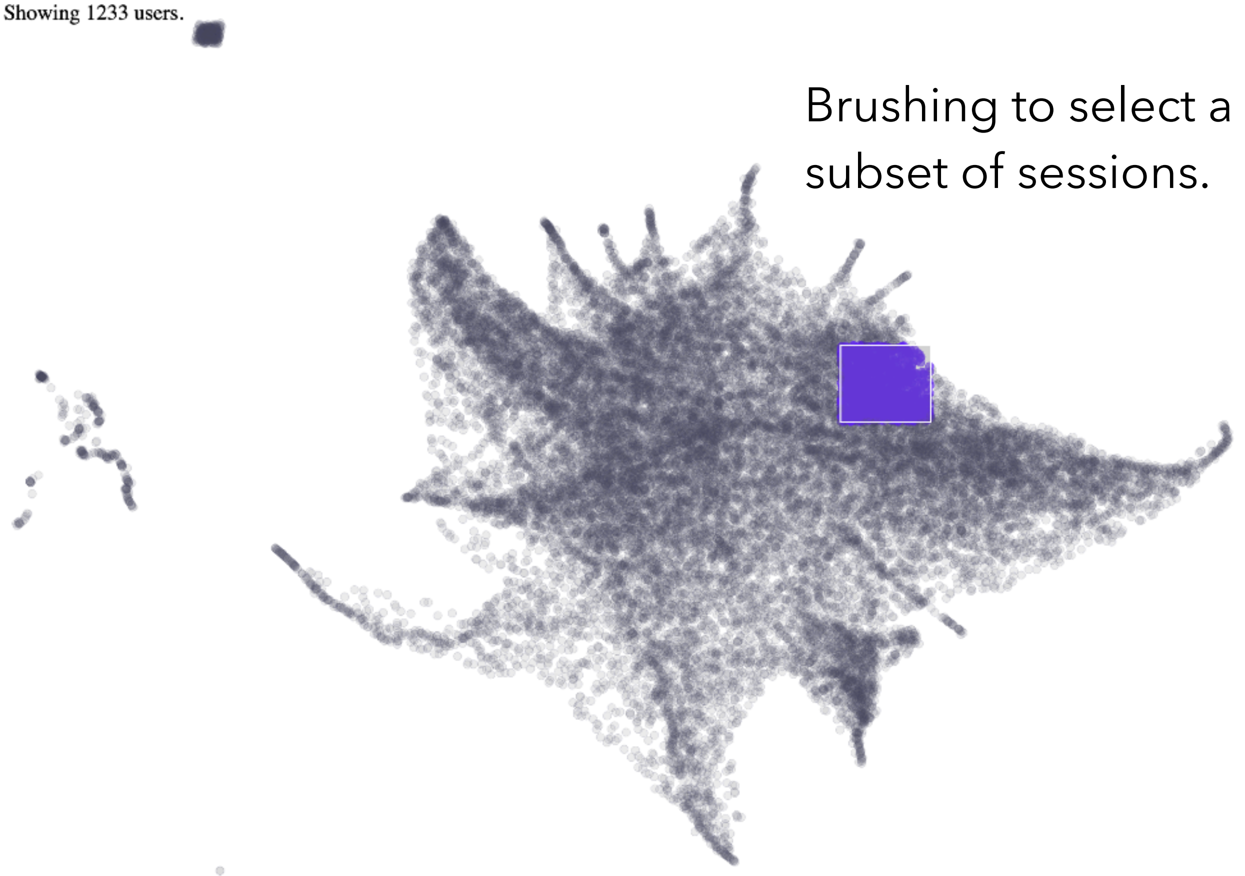


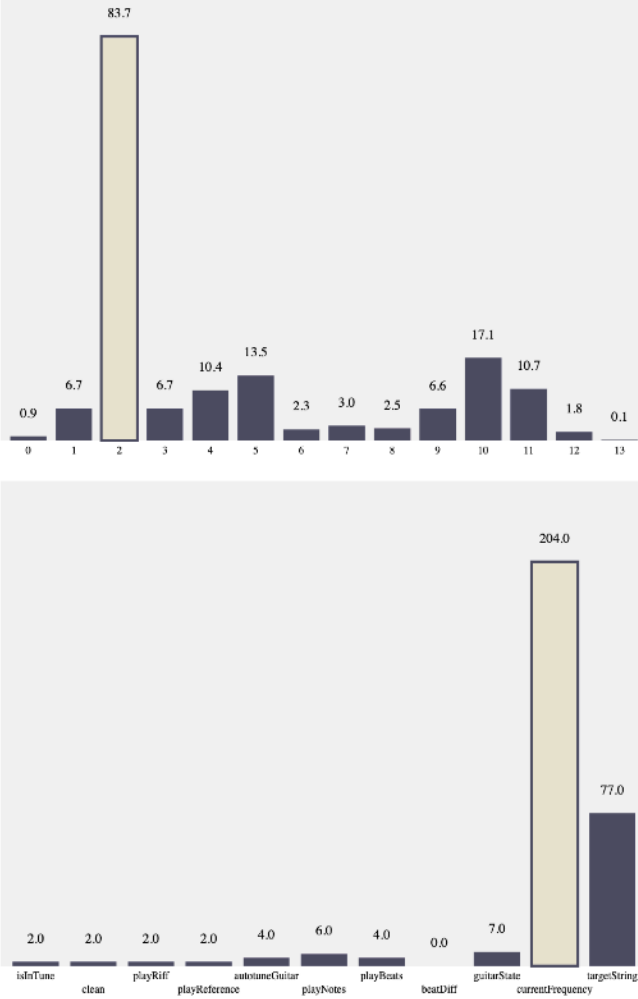


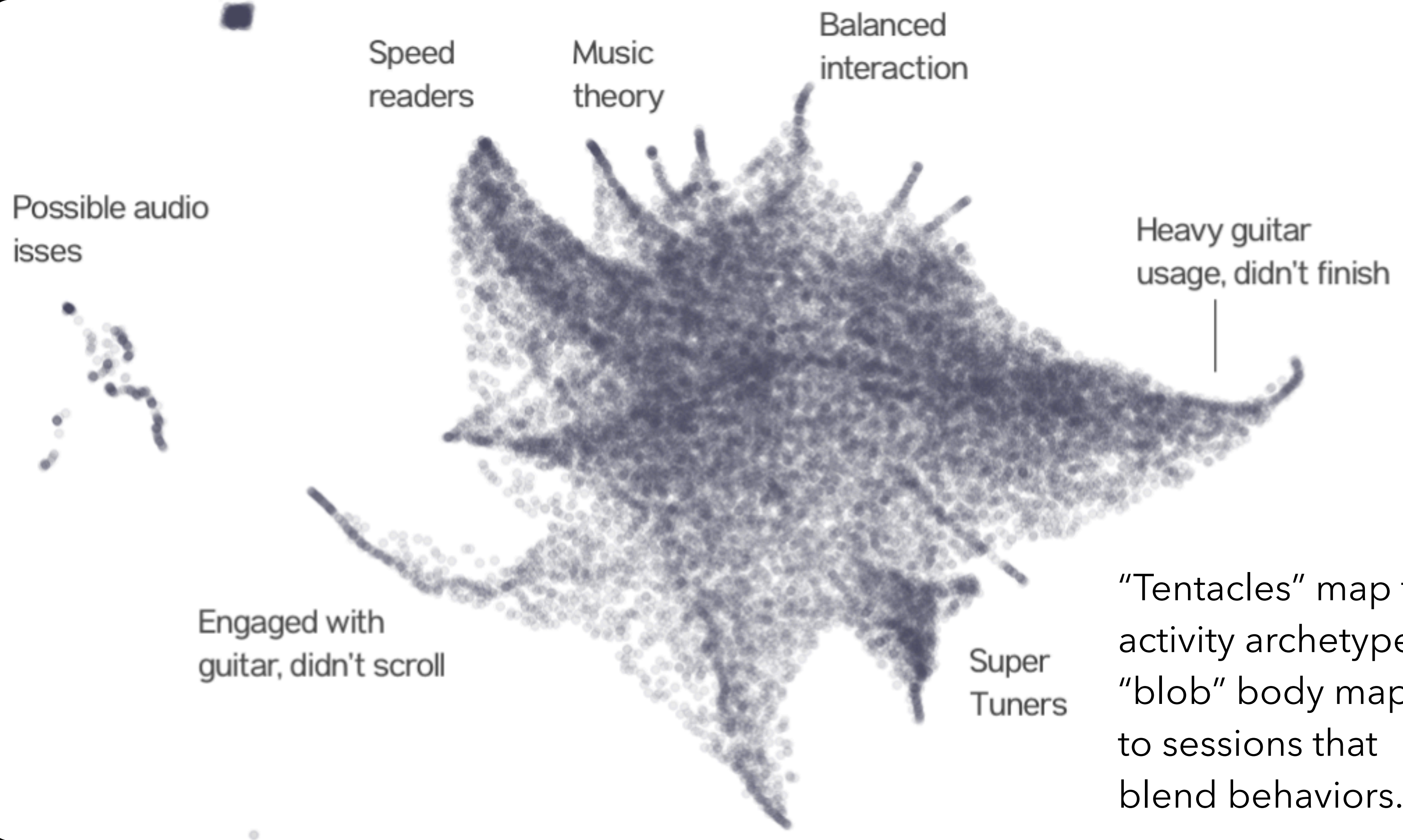
… and the count of times each variable changed



Brushing to select a subset of sessions.

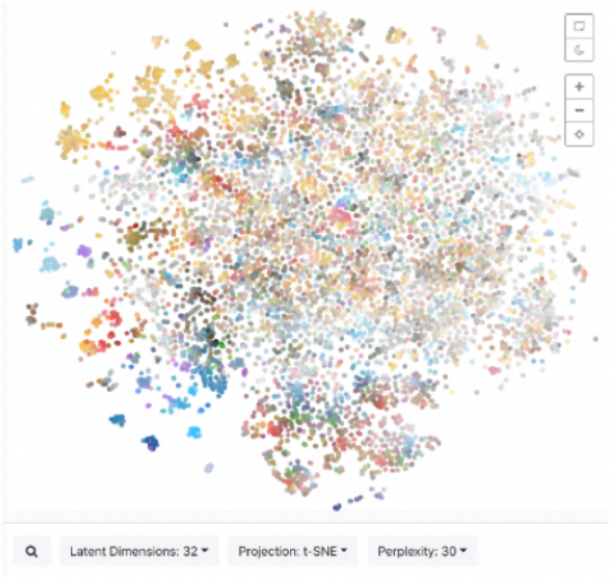




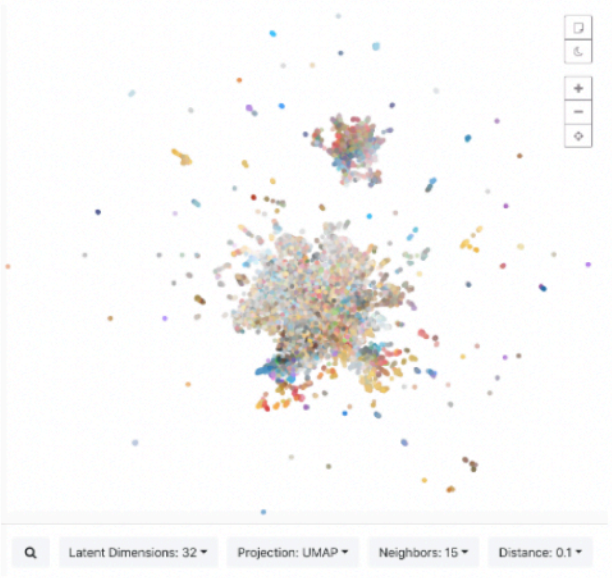


"Tentacles" map to activity archetypes, "blob" body maps to sessions that blend behaviors.

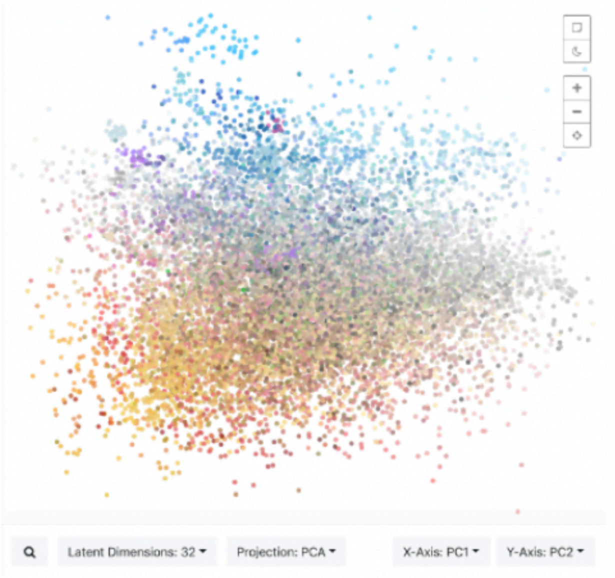
Mapping Emoji Images



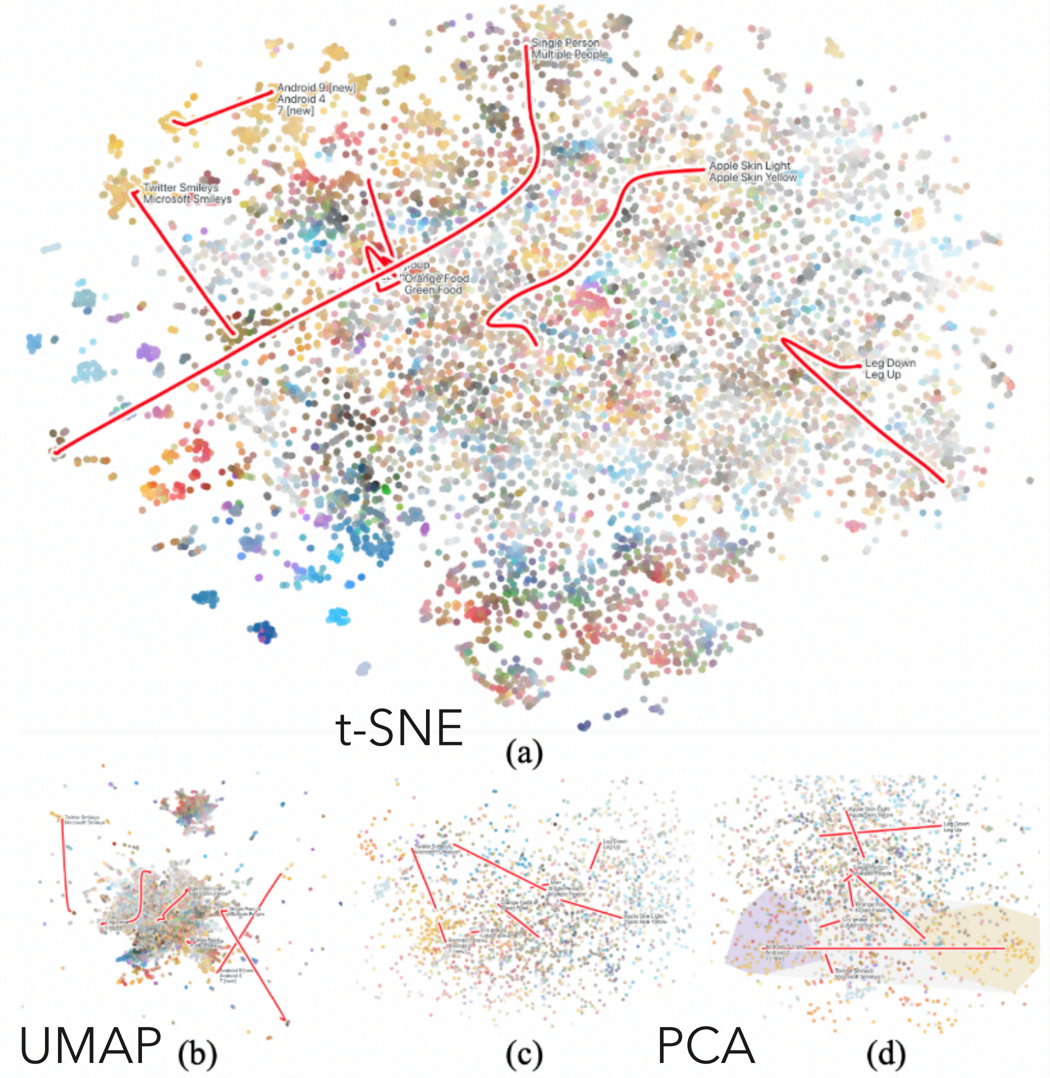
t-SNE



UMAP



PCA



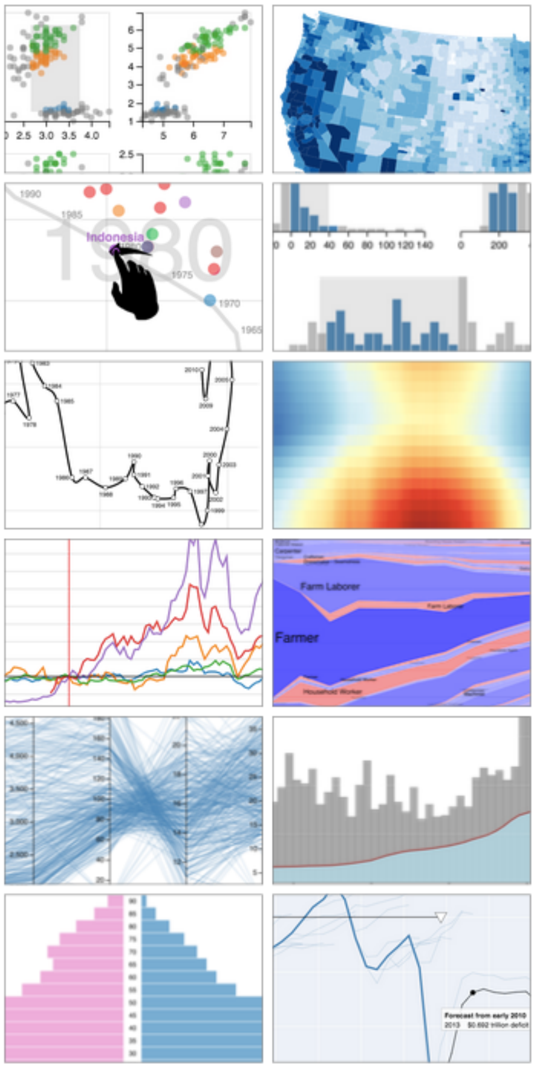
Single Person
Multiple People
Android 9 [new]
Android 4
7 [new]
Twitter Smileys
Microsoft Smileys
Apple Skin Light
Apple Skin Yellow
Orange Food
Green Food
Leg Down
Leg Up
t-SNE
(a)
UMAP
(b)
(c)
PCA
(d)

Dimensionality Reduction Issues

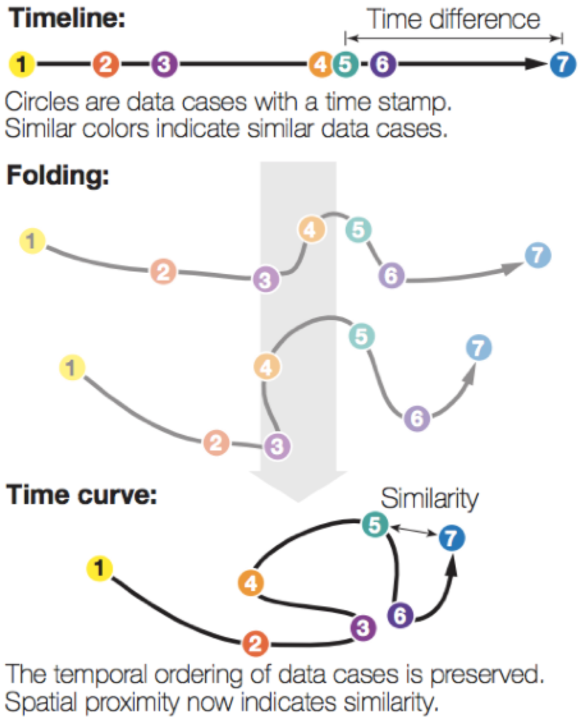
Reproducible?
Projections are *data-dependent*. Fitting a new projection with different data can give rise to different results.

Reusable?
PCA and UMAP provide reusable projection functions that can map new points from high-D to low-D. t-SNE (and others, like MDS) do not provide this.

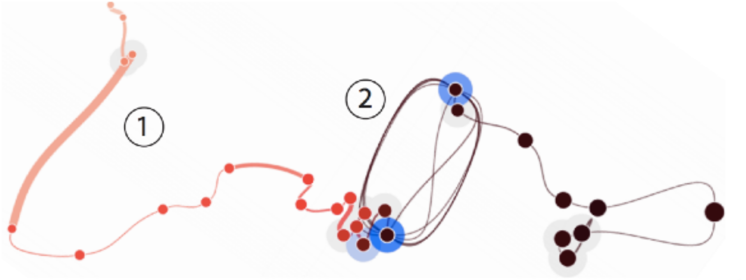
Interpretable?
DR plots are hard to interpret! Try multiple methods and hyperparameter settings. Inspect via interaction!



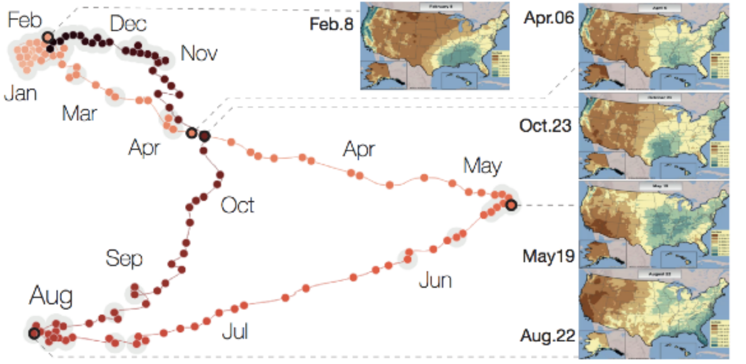
Time Curves [Bach et al. '16]



(a) Folding time



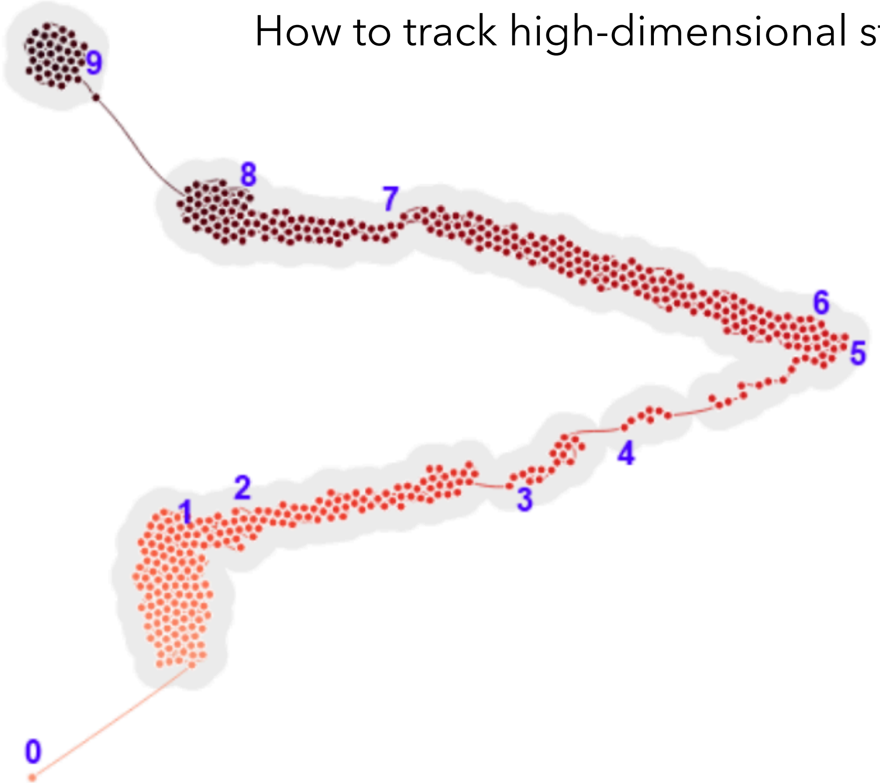
Wikipedia "Chocolate" Article



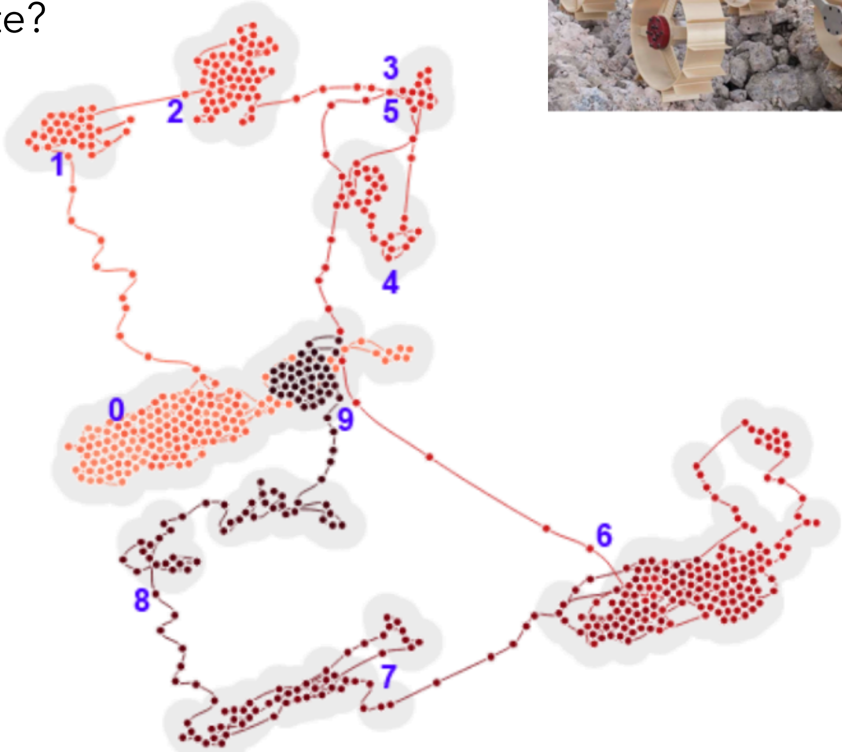
U.S. Precipitation over 1 Year

Rover Telemetry [Guy '16]

How to track high-dimensional state?



Using Raw Multi-D Data

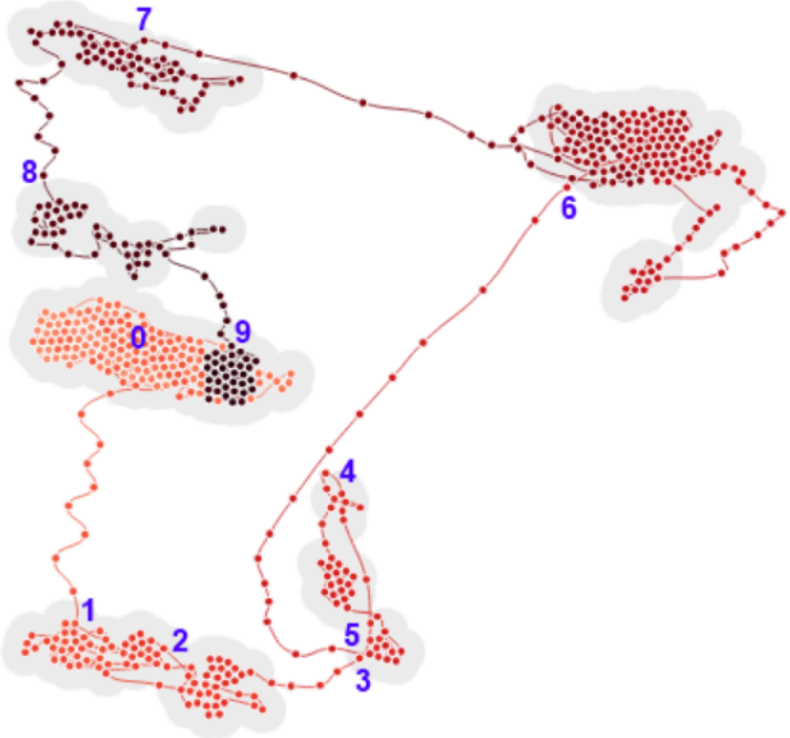


Using Pearson Correlation Matrix

Rover Telemetry [Guy '16]

How to track high-dimensional state?

Event Number	Timestamp	Event Description
0	0s	Rover begins traveling forward along smooth terrain.
1	188s	Rover begins descending into crater.
2	223s	Rover loses line of sight with lander and packet drops begin.
3	287s	Rover enters shade, causing temp, comms, and power drops.
4	300s	Rover begins traversing smooth bottom of crater.
5	330s	Rover begins climbing out of crater.
6	343s	Rover exits shade; continues uphill.
7	534s	Rover emerges from crater and enters smooth terrain.
8	594s	Rover enters choppy terrain.
9	643s	Rover wheel has fault; rover stops moving.

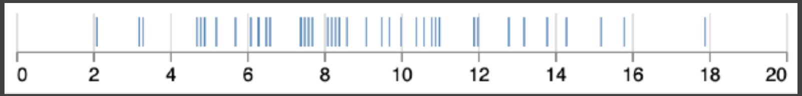


Using Spearman Correlation Matrix

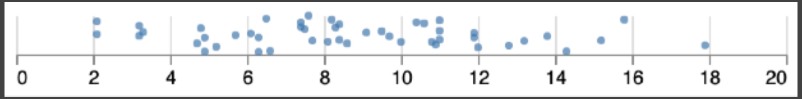
Visualizing Distributions
(Review)

Distribution Visualizations

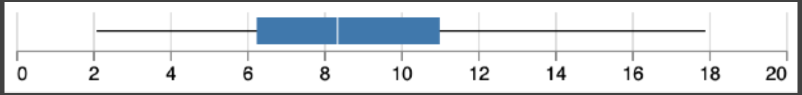
Strip Plot



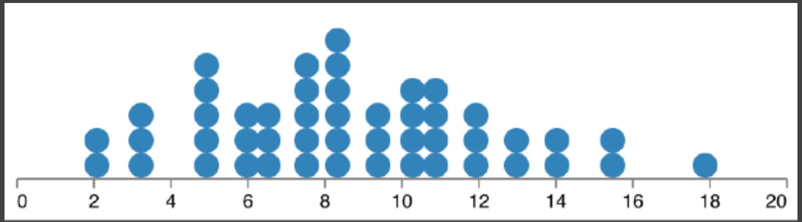
Jittered Plot



Box Plot



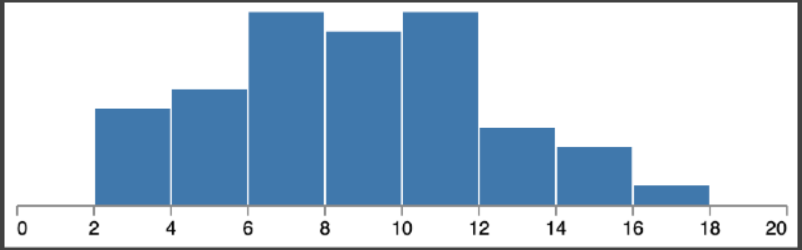
Dot Plot



Distribution Visualizations

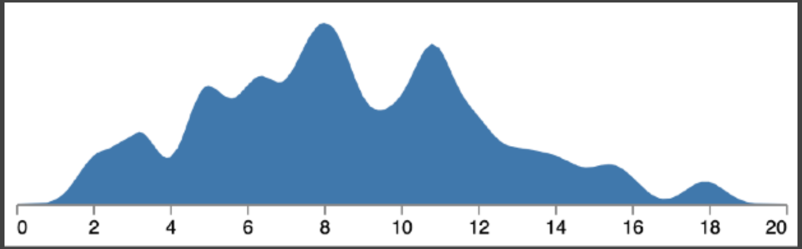
Histogram

bin size = 2



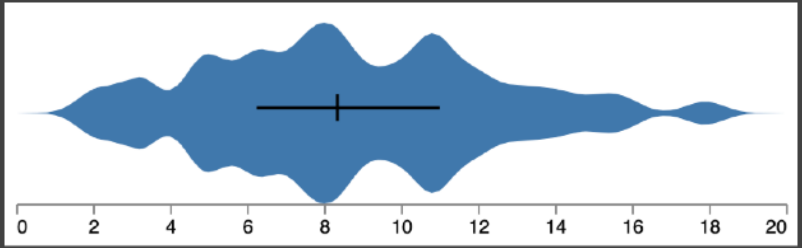
Density Plot

kde, $\sigma = 0.5$



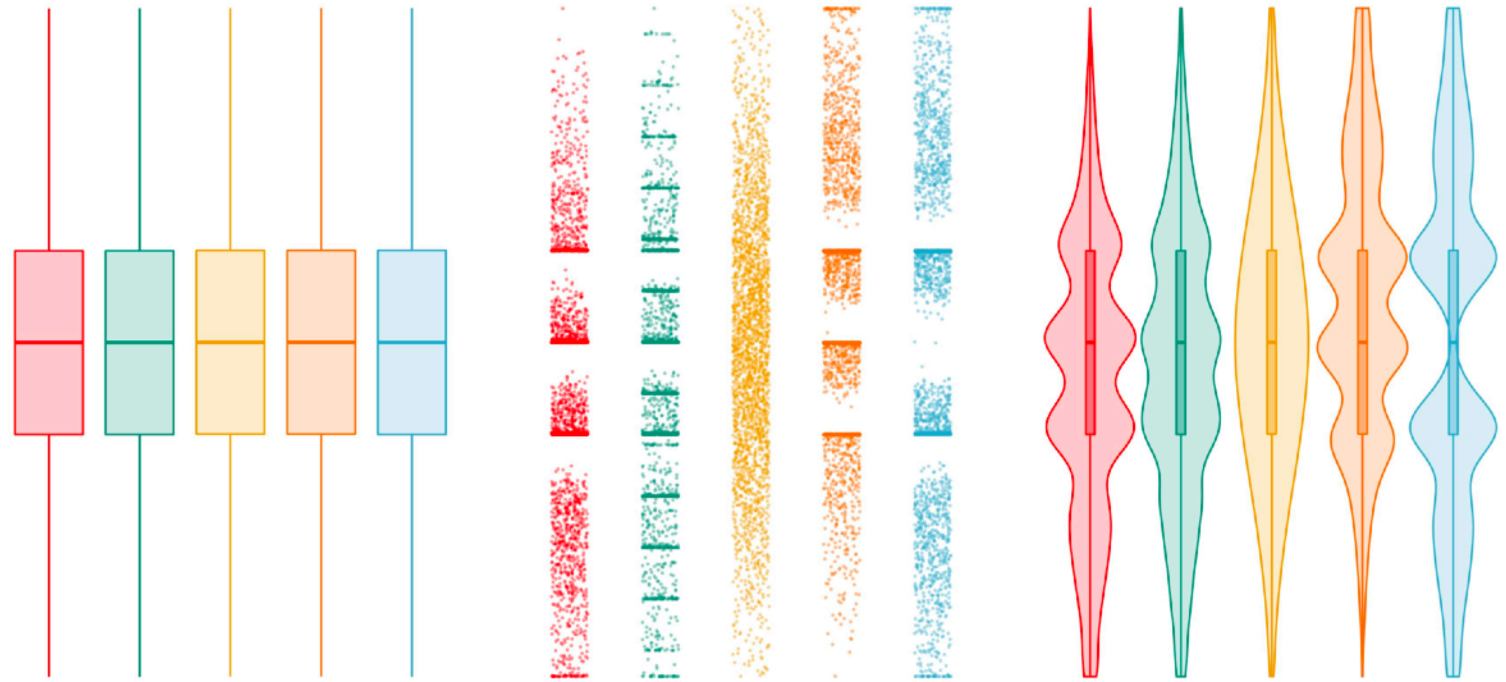
Violin Plot

kde, $\sigma = 0.5$



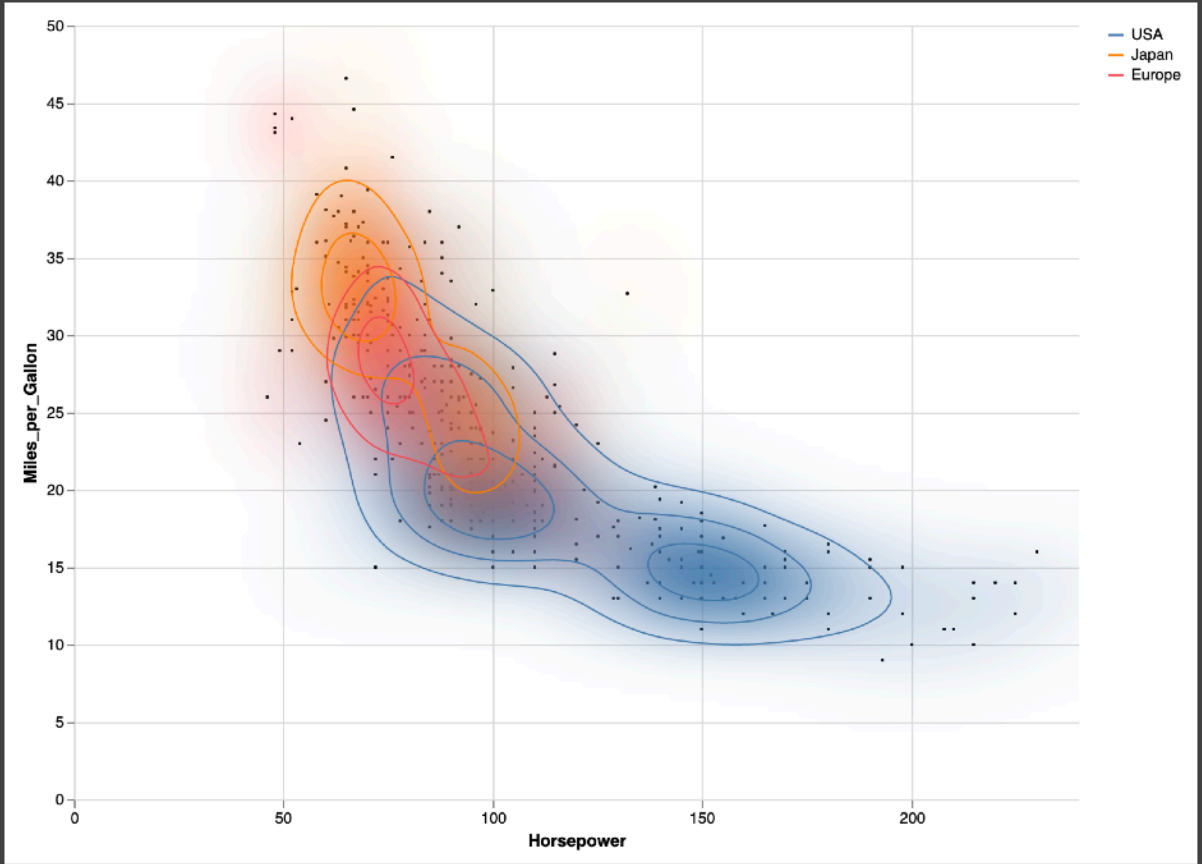
Identical boxplots, different distributions

Boxplots are great. They show medians and ranges and enable comparison of different groups. However, boxplots can be misleading.
Different datasets can have the same descriptive statistics (left), but quite different underlying distributions (middle).
Therefore, it is crucial to visualize the distribution in addition to descriptive statistics. Violin plots with integrated boxplots are great for this.



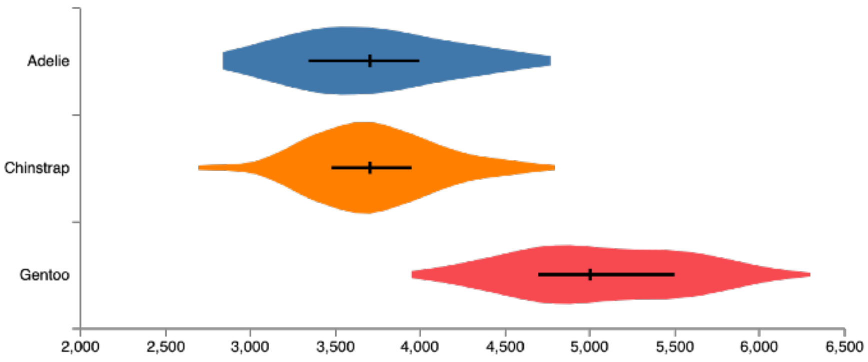
@MBozukova | #30DayChartChallenge | Day 27 - educational
Data source: {datasauRus} R package

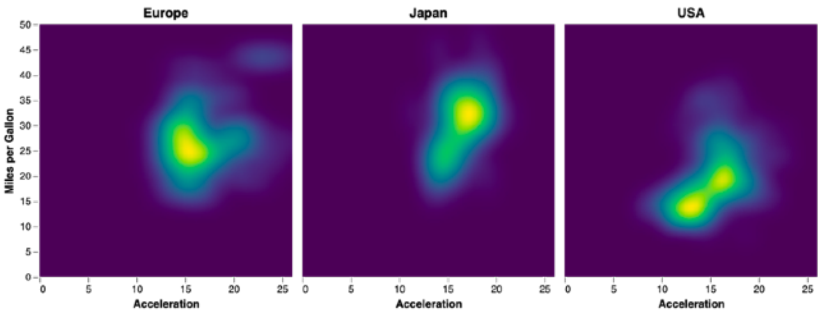
Now in 2D! Heatmaps, Contours

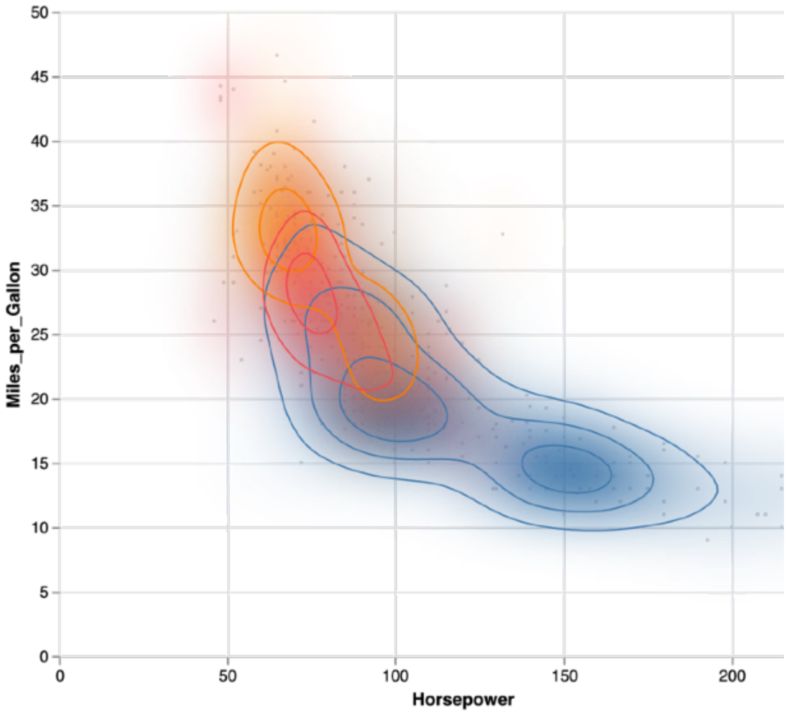


Kernel Density Estimation (KDE)

Enables violin plots, heat maps, contour plots…

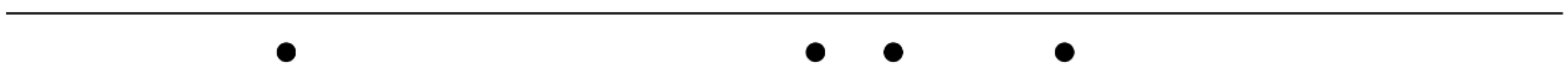






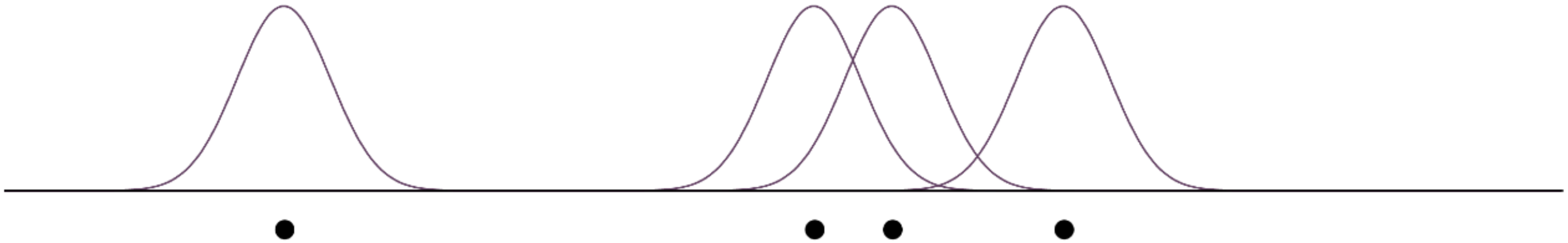
Kernel Density Estimation

For a set of input data points…



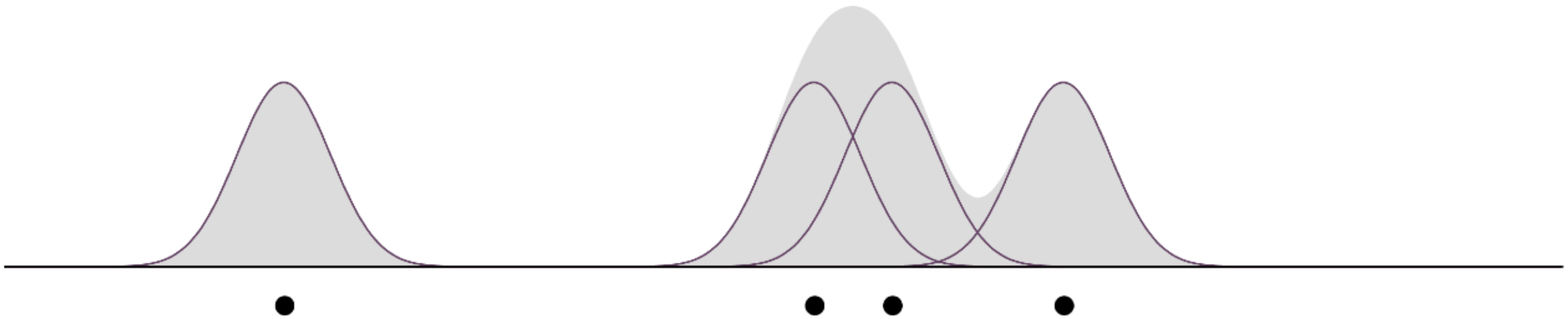
Kernel Density Estimation

Represent each point with a “kernel” distribution



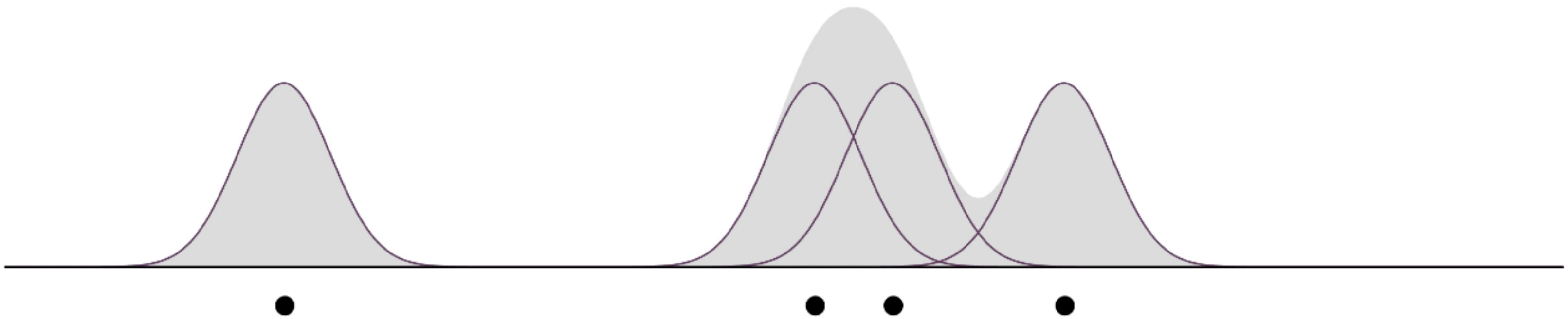
Kernel Density Estimation

Sum the kernels to form a density estimate



Kernel Density Estimation

Sized by bandwidth (standard deviation)



1D Distribution Exercise

This week you will think through how to communicate a distribution of values.

One column of numbers, a couple categories – *simple, right?* But there are many ways to transform and visualize distributions, which can lead to different insights as well as misleading omissions.

See Ed for links to the exercise document and online submission form.

We're here to help!