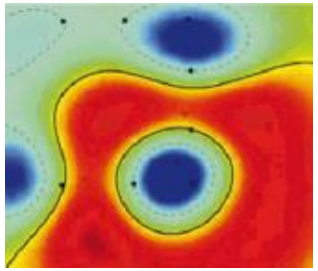
Dimension Reduction & Visualization

MDS, Isomap, SNE, t-SNE, Non-Metricity, Multiple t-SNE

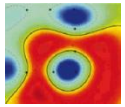


Lecture by Klaus-Robert Müller, TUB 2023

Linear methods of reducing dimensionality

- PCA finds the directions that have the most variance.
  - By representing where each datapoint is along these axes, we minimize the squared reconstruction error.
  - Linear autoencoders are equivalent to PCA

- Multi-Dimensional Scaling arranges the low-dimensional points so as to minimize the discrepancy between the pairwise distances in the original space and the pairwise distances in the low-D space.



[From Hinton]

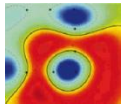
Metric Multi-Dimensional Scaling

- Find low dimensional representatives, y, for the high-dimensional data-points, x, that preserve pairwise distances as well as possible.
- An obvious approach is to start with random vectors for the y's and then perform steepest descent by following the gradient of the cost function.
- Since we are minimizing squared errors, maybe this has something to do with PCA?
  - If so, we don't need an iterative method to find the best embedding.

$$Cost = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2$$

$$d_{ij} = \| x_i - x_j \|^2$$

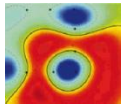
$$\hat{d}_{ij} = \| y_i - y_j \|^2$$



[From Hinton]

Converting metric MDS to PCA

- If the data-points all lie on a hyperplane, their pairwise distances are perfectly preserved by projecting the high-dimensional coordinates onto the hyperplane.
  - So in that particular case, PCA is the right solution.
- If we "double-center" the data, metric MDS is equivalent to PCA.
  - Double centering means making the mean value of every row and column be zero.
  - But double centering can introduce spurious structure.



[From Hinton]

Other non-linear methods of reducing dimensionality

- Non-linear autoencoders with extra layers are much more powerful than PCA but they can be slow to optimize and they get different, locally optimal solutions each time.

- Multi-Dimensional Scaling can be made non-linear by putting more importance on the small distances. A popular version is the Sammon mapping:

high-D distance

low-D distance

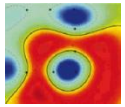
$$Cost = \sum_{i\,j} \left(\frac{\| \mathbf{x}_i - \mathbf{x}_j \| - \| \mathbf{y}_i - \mathbf{y}_j \|}{\| \mathbf{x}_i - \mathbf{x}_j \|} \right)^2$$

- Non-linear MDS is also slow to optimize and also gets stuck in different local optima each time.

[From Hinton]

Problems with Sammon mapping

- It puts too much emphasis on getting very small distances exactly right.
- It produces embeddings that are circular with roughly uniform density of the map points.

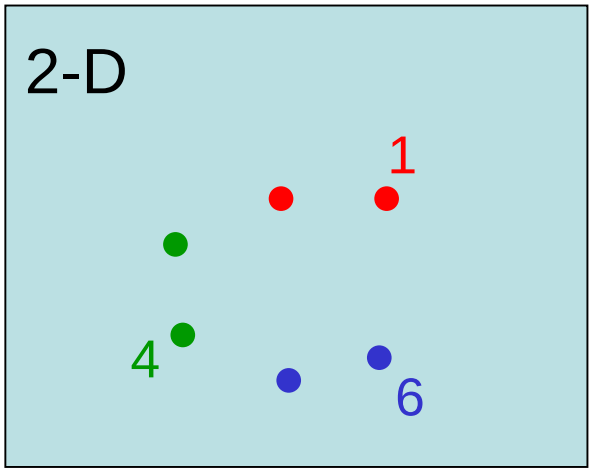


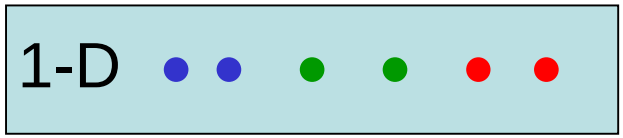
[From Hinton]

IsoMap: Local MDS without local optima

- Instead of only modeling local distances, we can try to measure the distances along the manifold and then model these intrinsic distances.
  - The main problem is to find a robust way of measuring distances along the manifold.
  - If we can measure manifold distances, the global optimisation is easy: It's just global MDS (i.e. PCA)

If we measure distances along the manifold,
d(1,6) > d(1,4)

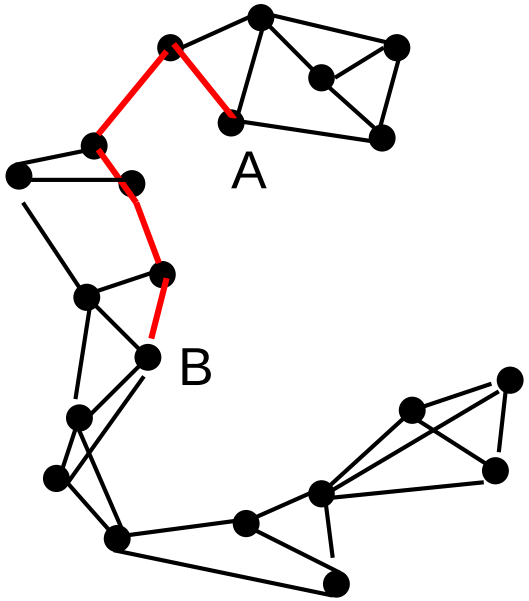




[From Hinton]

How Isomap measures intrinsic distances

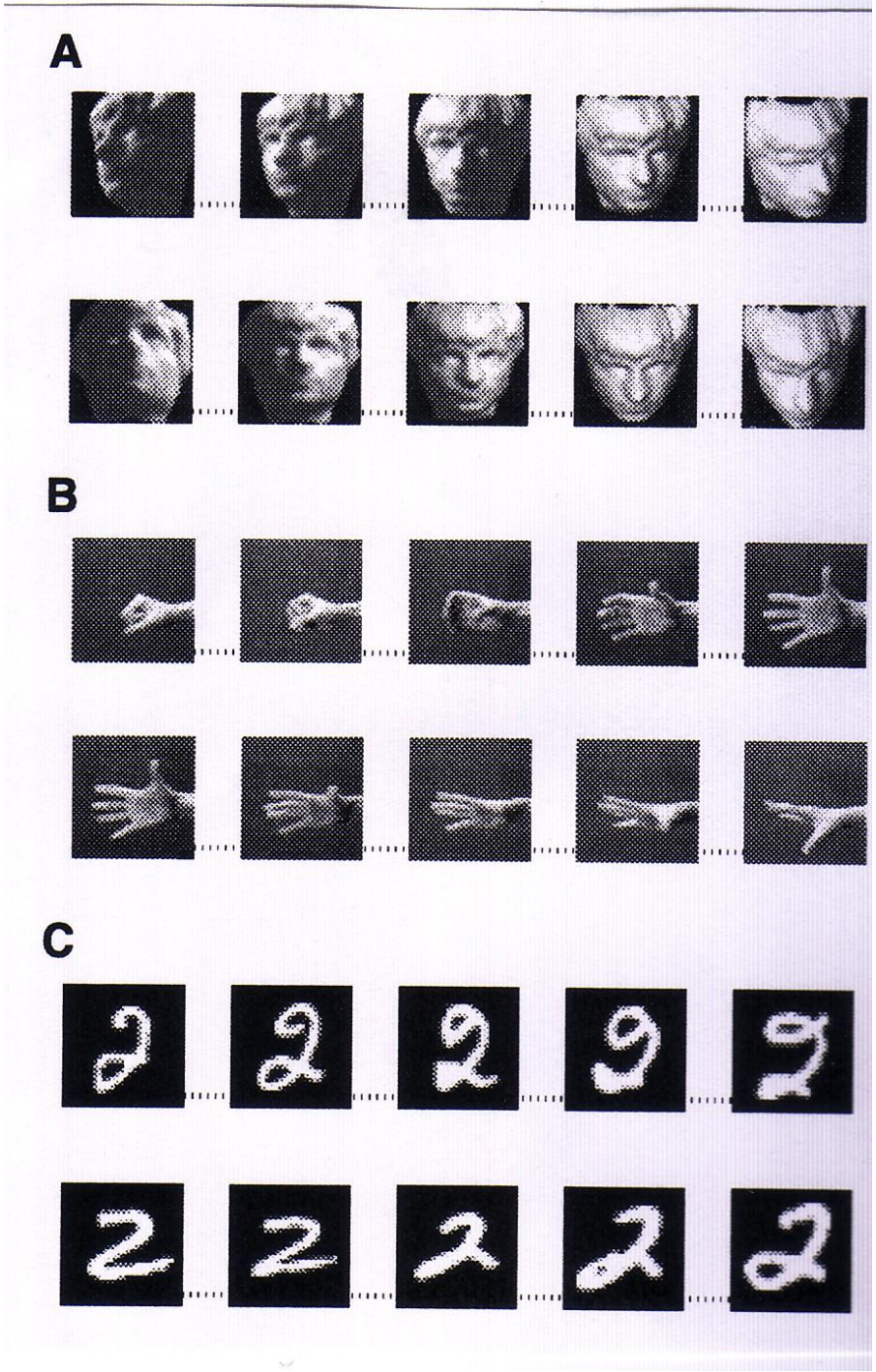
- Connect each datapoint to its K nearest neighbors in the high-dimensional space.
- Put the true Euclidean distance on each of these links.
- Then approximate the manifold distance between any pair of points as the shortest path in this “neighborhood graph”.



[From Hinton]

Using Isomap to discover the intrinsic manifold in a set of face images

[From Hinton]



Linear methods cannot interpolate properly between the leftmost and rightmost images in each row.

This is because the interpolated images are NOT averages of the images at the two ends.

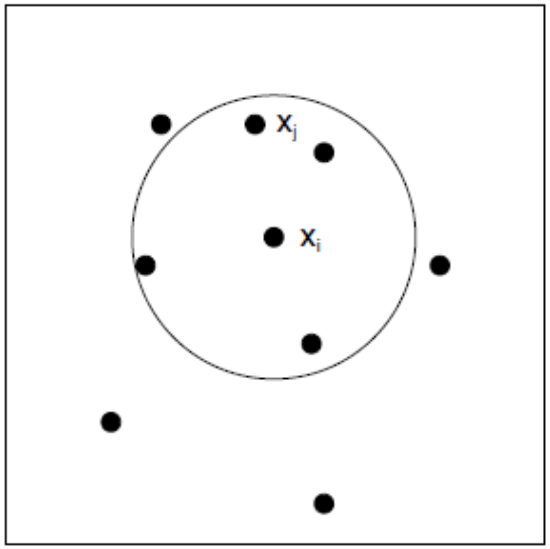
Isomap does not interpolate properly either because it can only use examples from the training set. It cannot create new images.

But it is better than linear methods.

[From Hinton]

(t)-SNEs

- Represents similarities between objects by measuring densities under Gaussians:



$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)}$$

- This is so-called *stochastic neighbor selection**

* If we are lucky, our data already has the form of p_{ij}'s!

[From vd Maaten]

SNE

- Defines similar probabilities in the map:

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)}$$

- Minimizes the KL divergence between the distributions in the high-dimensional space and the map:

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

[From vd Maaten]

SNE optimization

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j).$$

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right),$$

Physically, the gradient may be interpreted as the resultant force created by a set of springs between the map point y_i and all other map points y_j. All springs exert a force along the direction $(y_i - y_j)$. The spring between y_i and y_j repels or attracts the map points depending on whether the distance between the two in the map is too small or too large to represent the similarities between the two high-dimensional datapoints. The force exerted by the spring between y_i and y_j is proportional to its length, and also proportional to its stiffness, which is the mismatch $(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$ between the pairwise similarities of the data points and the map points.

Crowding Problem

- Suppose we have datapoints that are sampled uniformly from a hypercube
- Also suppose we perfectly modeled the local structure of this data in the map (which is usually impossible)
- Result: dissimilar points have to be modeled as **too far apart** in the map
- Resulting forces 'crush together' the map

[From vd Maaten]

t-SNE

- Uses a heavy-tailed distribution in the map

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

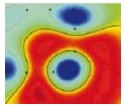
- As a result, dissimilar objects are allowed to be modeled too far apart
- This eliminates the crowding problem!

[From vd Maaten]

Algorithm

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for t=*1* **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end



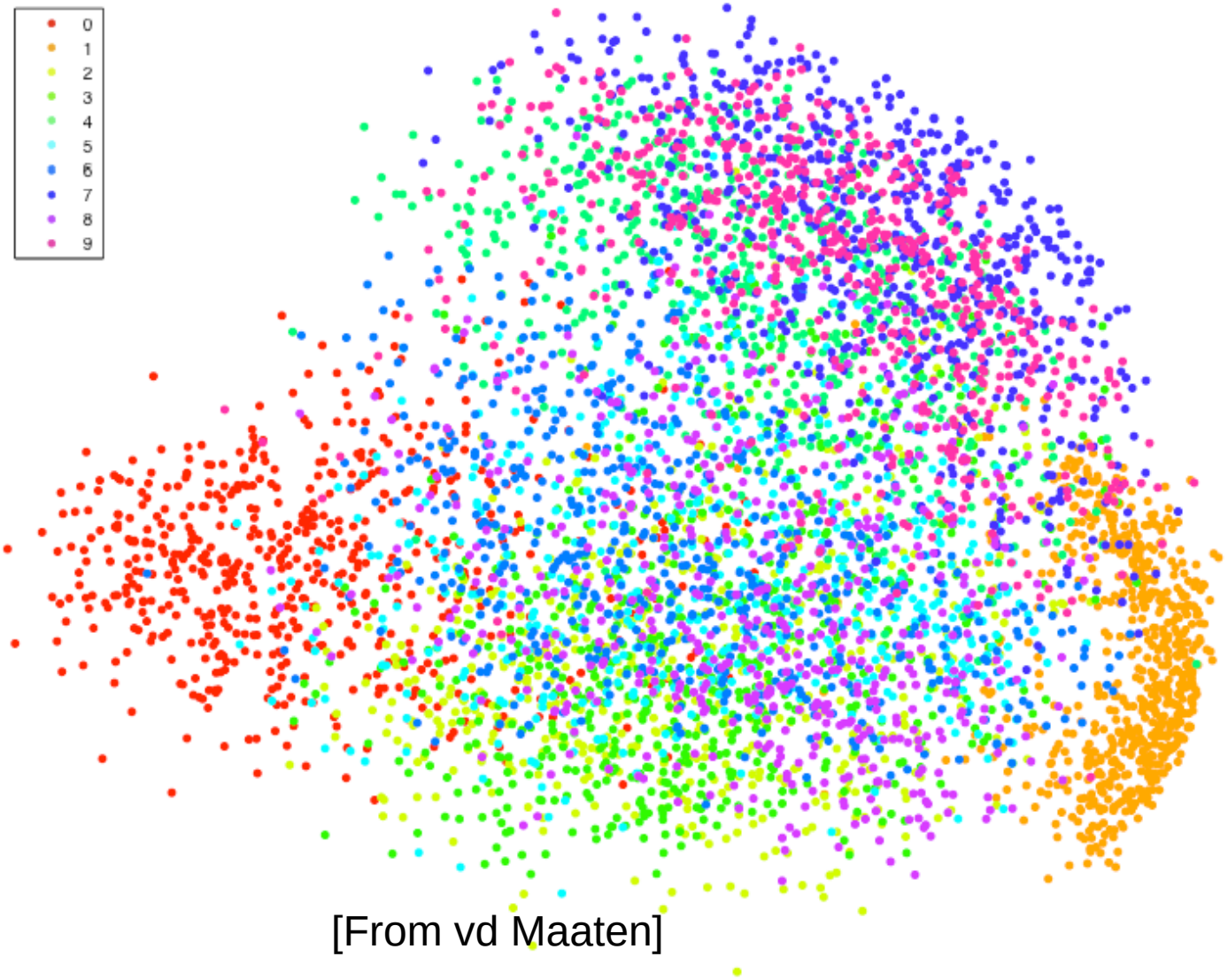
[From vd Maaten]

Experiment

- MNIST dataset (10 classes, 70,000 images)

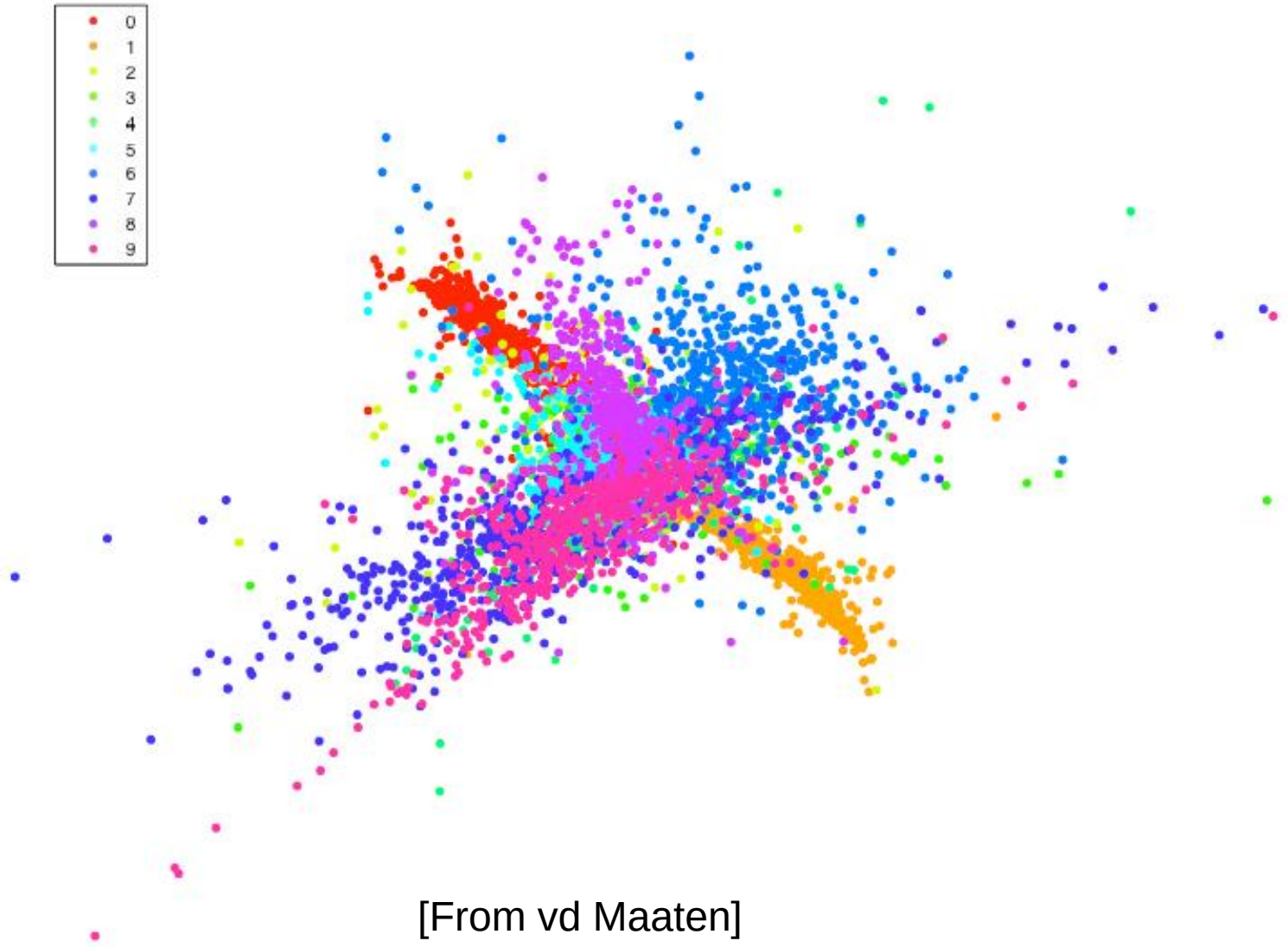
[From vd Maaten]

PCA



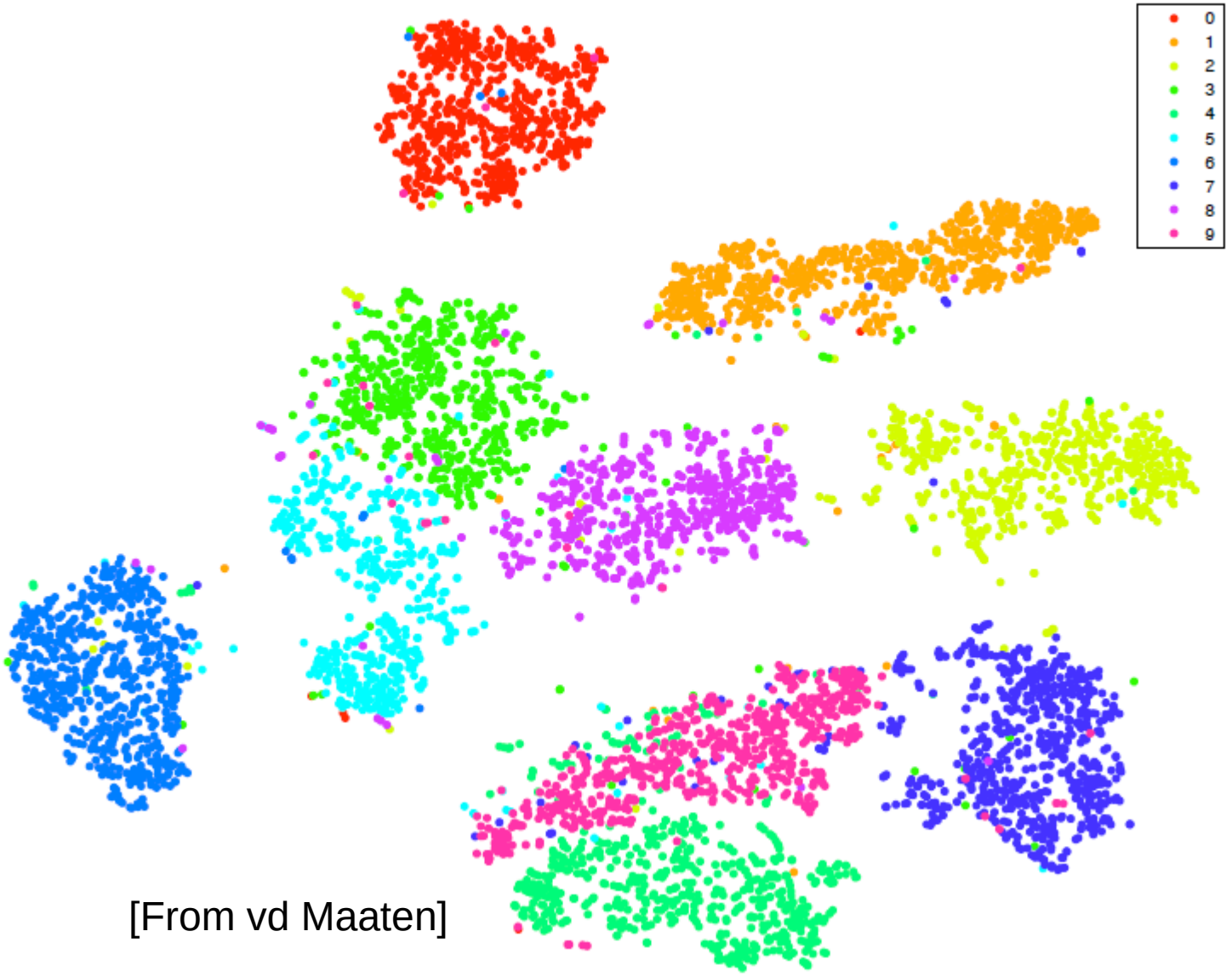
[From vd Maaten]

LLE

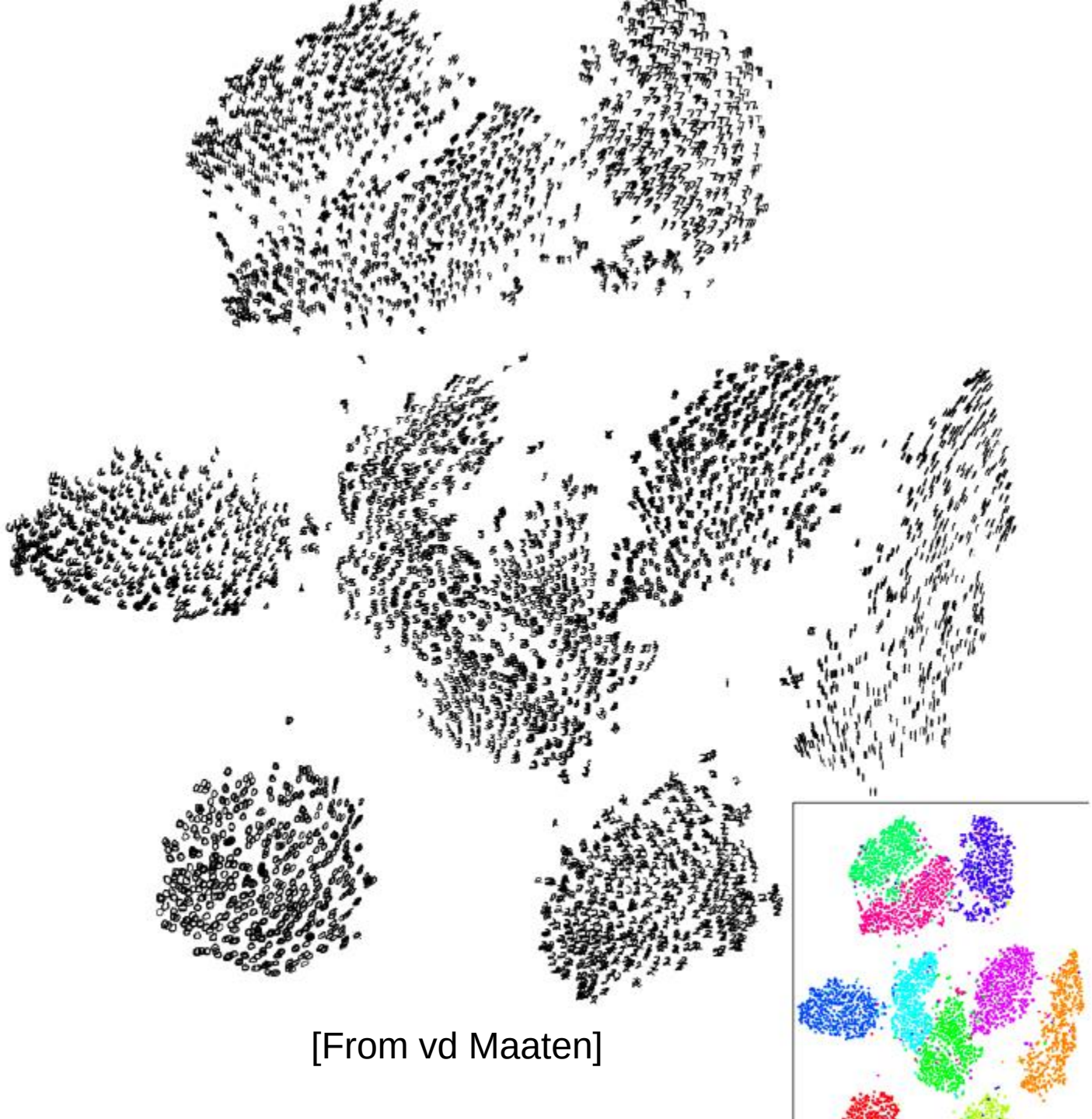


[From vd Maaten]

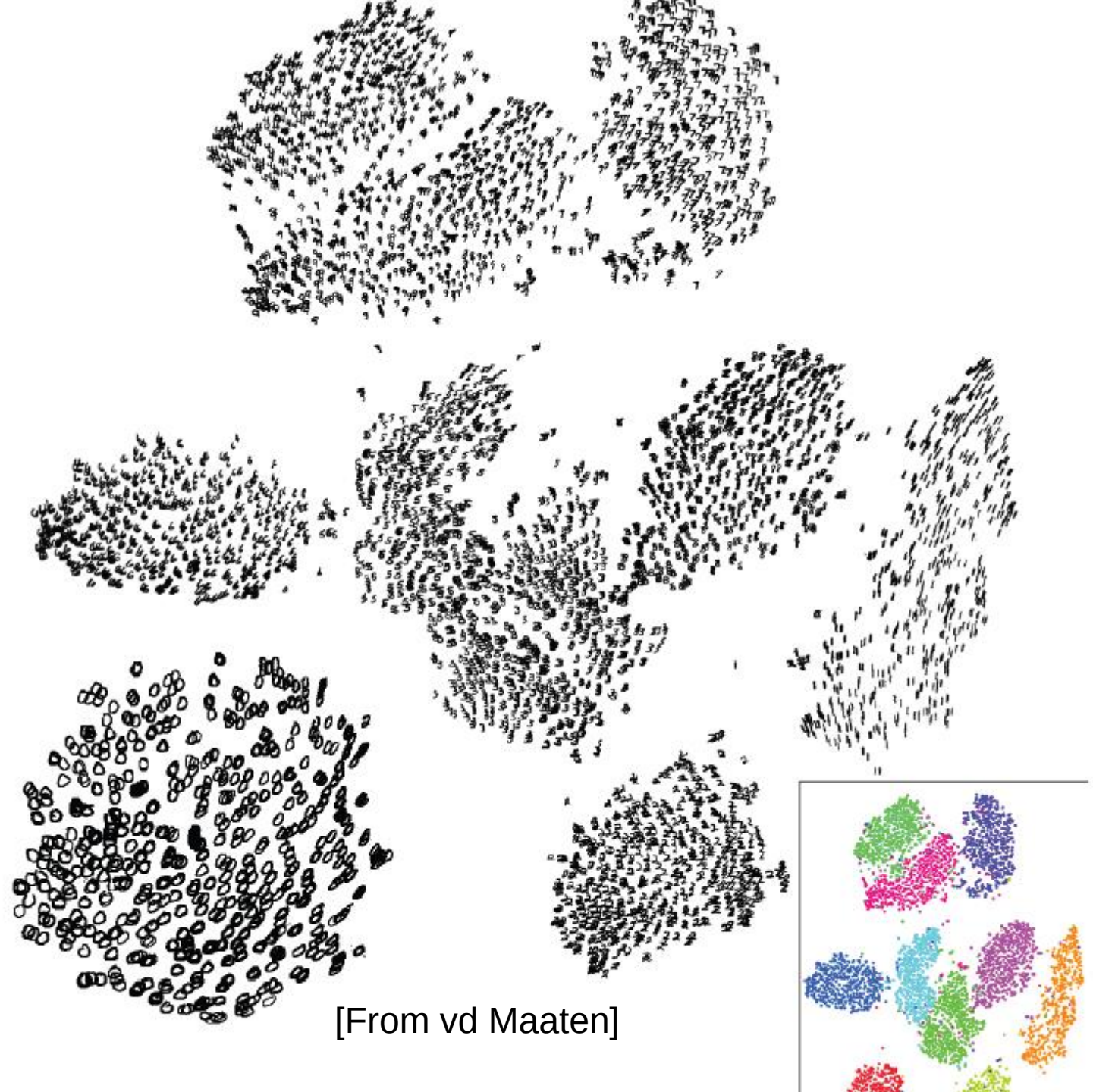
t-SNE



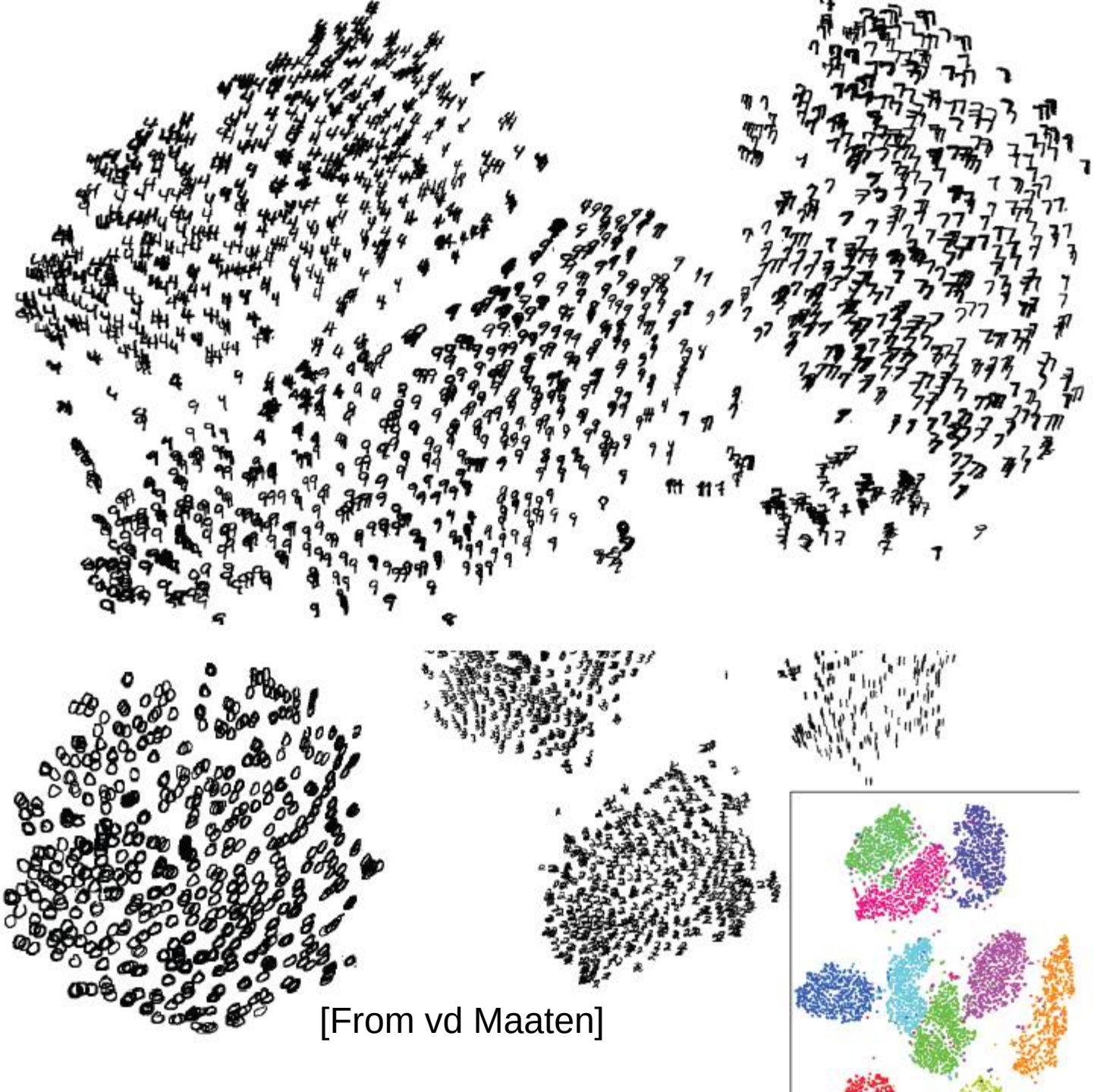
[From vd Maaten]



[From vd Maaten]

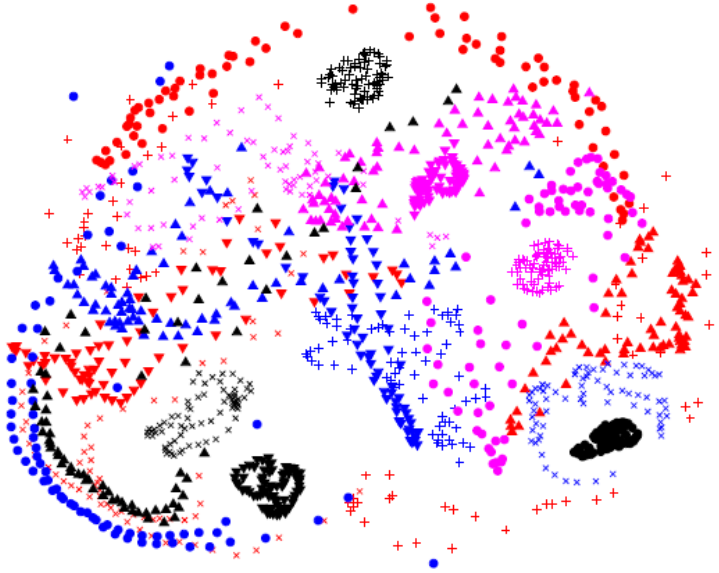


[From vd Maaten]

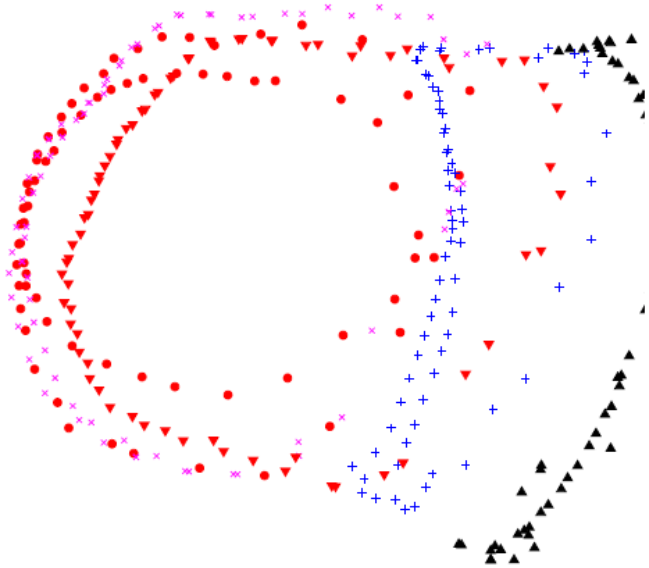


[From vd Maaten]

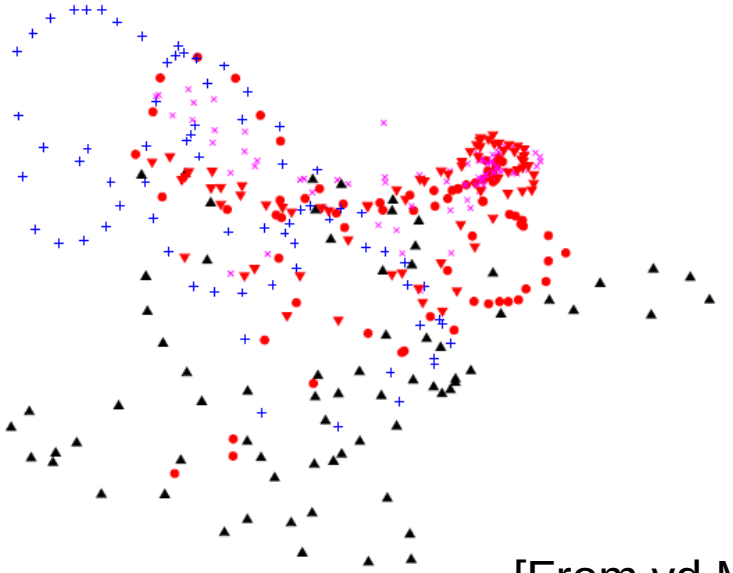
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



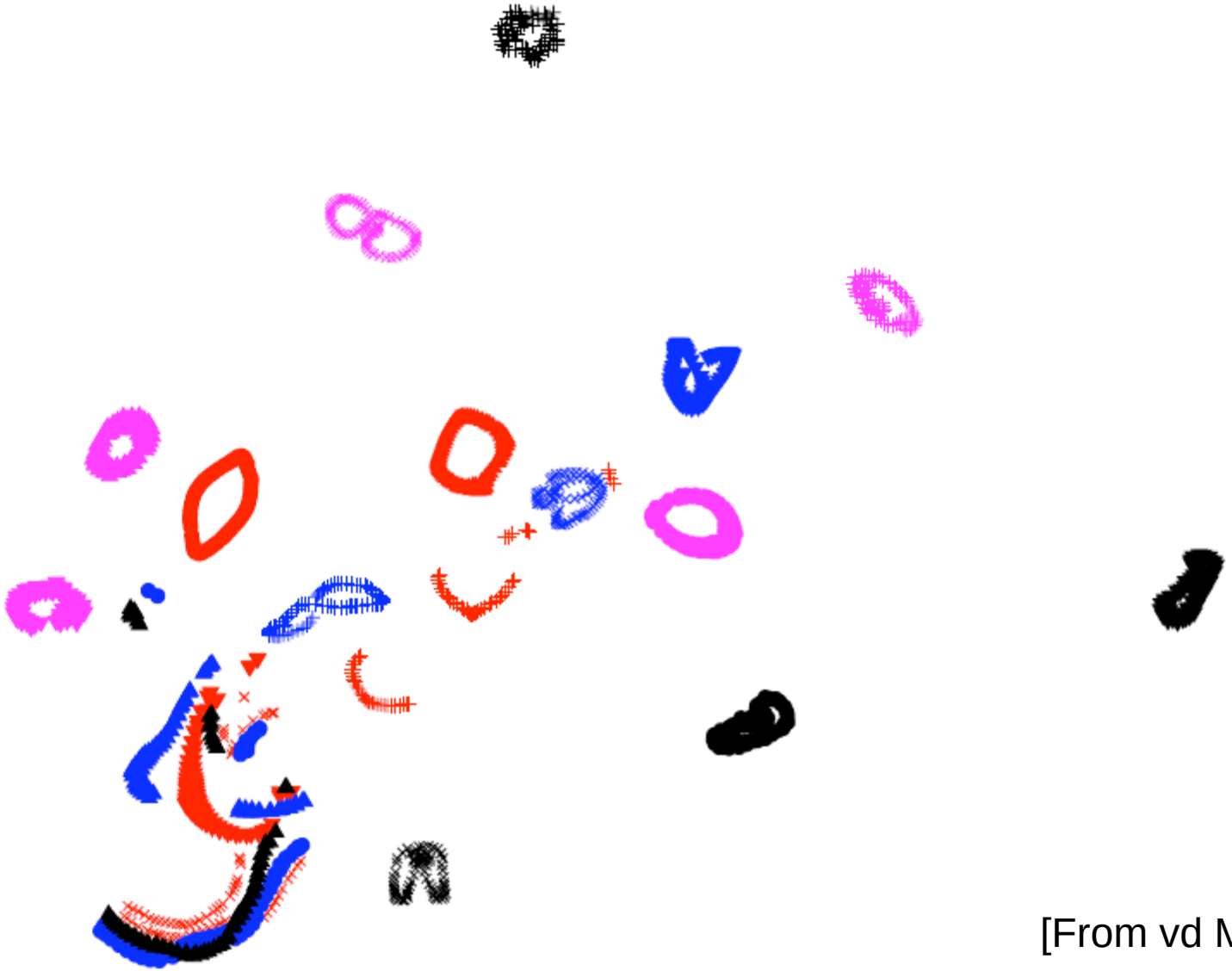
(c) Visualization by Isomap.



(d) Visualization by LLE.

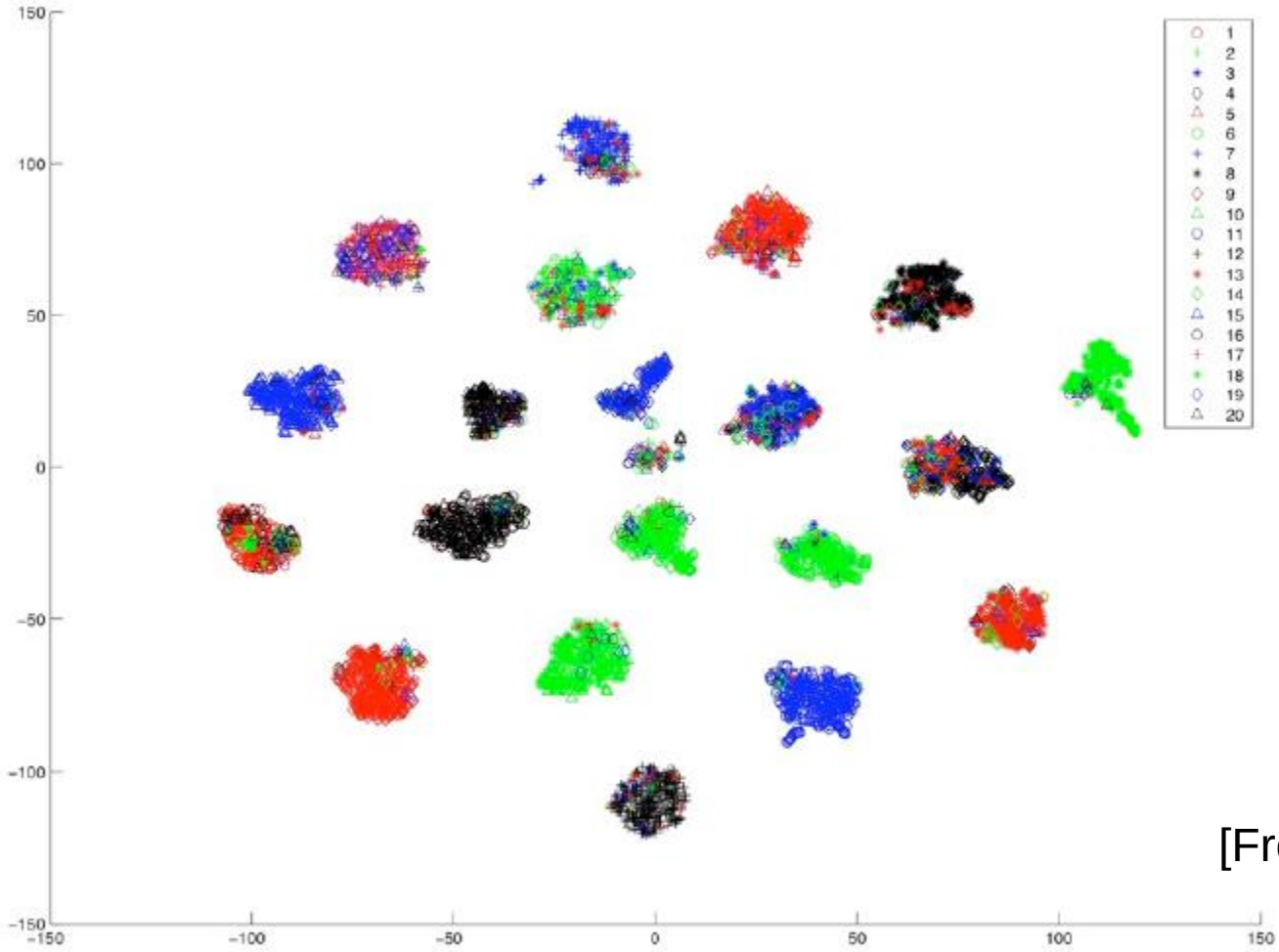
[From vd Maaten]

COIL-20



[From vd Maaten]

20-Newsgroups



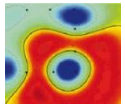
[From vd Maaten]

* Based on discLDA features (by Simon Lacoste-Julien).

Non-Metricity

Analyzing Non-Euclidean Pairwise Data

Julian Laub[1] & Klaus-Robert Müller[1,2]

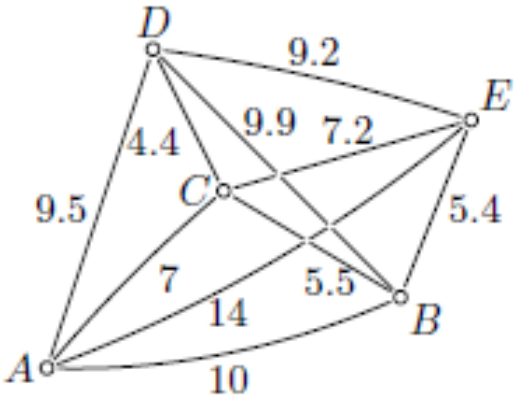


Pairwise Similarity Data

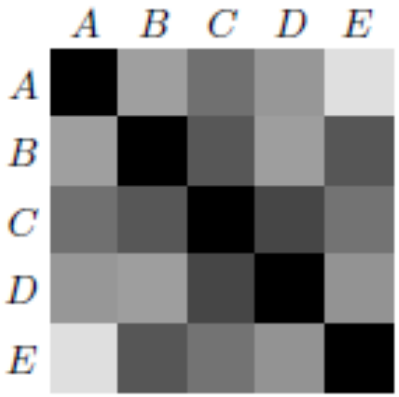
1. Pairwise data occur in many fields: genomics, text mining, cognitive psychology, social sciences...
2. Pairwise data can be represented as undirected graphs, as tables ("matrices") or as checkerboard patterns.
3. For a mathematical treatment, pairwise proximity data will be represented as a similarity matrix $S = (s_{ij})$, $s_{ij} \in \mathbb{R}$, or a dissimilarity matrix $D = (d_{ij})$, $d_{ij} \in \mathbb{R}$.
4. There are no formalized mathematical requirements on S or D. They may be very general and very hard to interprete.

Pairwise Data

Pairwise Data: Overview



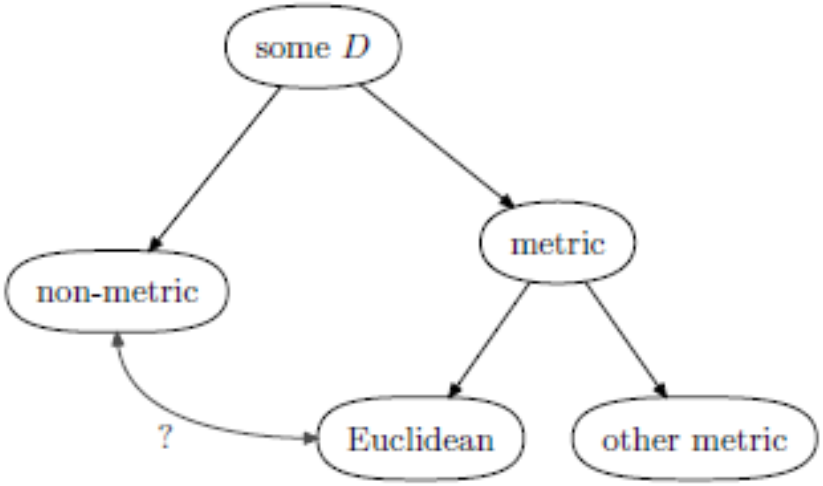
	A	B	C	D	E
A	0	10	7	9.5	14
B	10	0	5.5	9.9	5.4
C	7	5.5	0	4.4	7.2
D	9.5	9.9	4.4	0	9.2
E	14	5.4	7.2	9.2	0



Representation of pairwise data as a graph (left), a table (middle) or a checkerboard pattern (right).

Pairwise Data: metric violations

1. Overview:



2. Example of metric violations:
   - Noisy measurements.
   - Intrinsic non-metricity in human similarity judgments, text-mining
3. Non-metric pairwise data *cannot* be represented isometrically as vectors, even in high dimensions.

Pairwise Data: metric violations

1. Let $D = (d_{ij})$ be a dissimilarity matrix. D is called *metric* if the d_{ij}'s satisfy the following conditions:

   (a) $d_{ij} \geqslant 0 \, \forall i, j$,

   (b) $d_{ij} = 0$ iff $i = j$,

   (c) $d_{ij} = d_{ji} \, \forall i, j$ and

   (d) $d_{ij} \leqslant d_{ik} + d_{ik} \, \forall i, j, k$.

2. A dissimilarity matrix $D = (d_{ij})$ is called *squared Euclidean* if and only if there exist vectors $x_1, x_2, \ldots x_n \in \mathbb{R}^p$ such that $d_{ij} = \|x_i - x_j\|_2^2$, where $\| \cdot \|_2$ denotes Euclidean norm.

Visualizing Metric Violations

1. Metric violations translate into indefinite pseudo-covariance matrices.
2. Algorithm:

$$D \xrightarrow{C = -\frac{1}{2}QDQ} C \text{ with } p \text{ positive and } q \text{ negative eigenvalues}$$

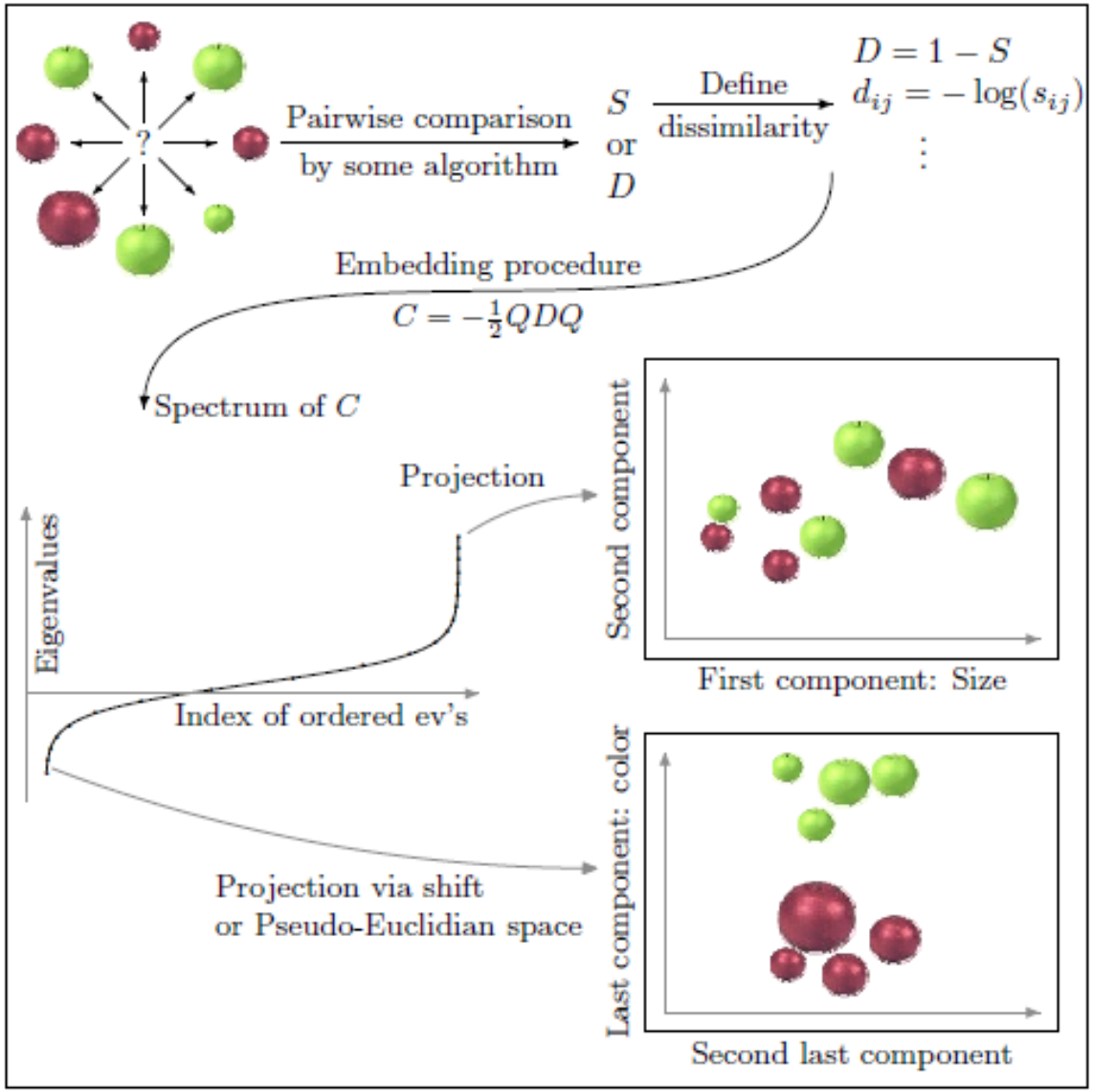
$$C \xrightarrow{\text{spectral decomposition}} V\Lambda V^\top = V|\Lambda|^{\frac{1}{2}} M |\Lambda|^{\frac{1}{2}} V^\top$$

$$X^* = |\Lambda|^{1/2} V^\top$$

where $Q = \left(I - \frac{1}{n} ee^t\right)$ and $M = \begin{pmatrix} I_{p\times p} & & \\ & -I_{q\times q} & \\ & & 0_{k\times k} \end{pmatrix}$

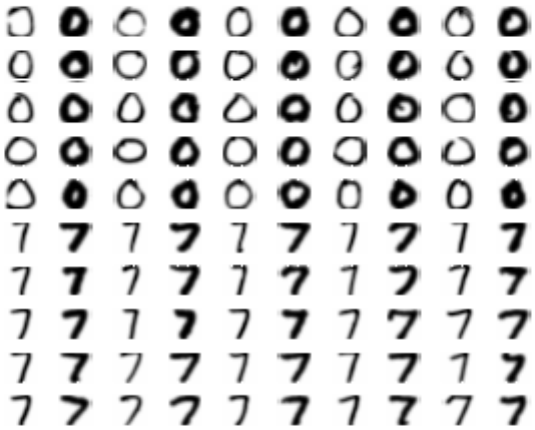
3. Project onto $\{v_1, v_2\}$.
4. Project onto $\{v_n, v_{n-1}\}$.

Metric Violations Summary



Feature Discovery: Examples

1. **USPS handwritten digits.** The similarity matrix is obtained from binary image matching on the digits 0 and 7 of the USPS data set.



2. Binary image matching:

$$s_{rs} = \frac{a}{\min\left(a+b, a+c\right)}. \qquad (1)$$

Feature Discovery: Examples

1. Projection onto the positive and projection onto the negative eigenspaces yield results different in nature.

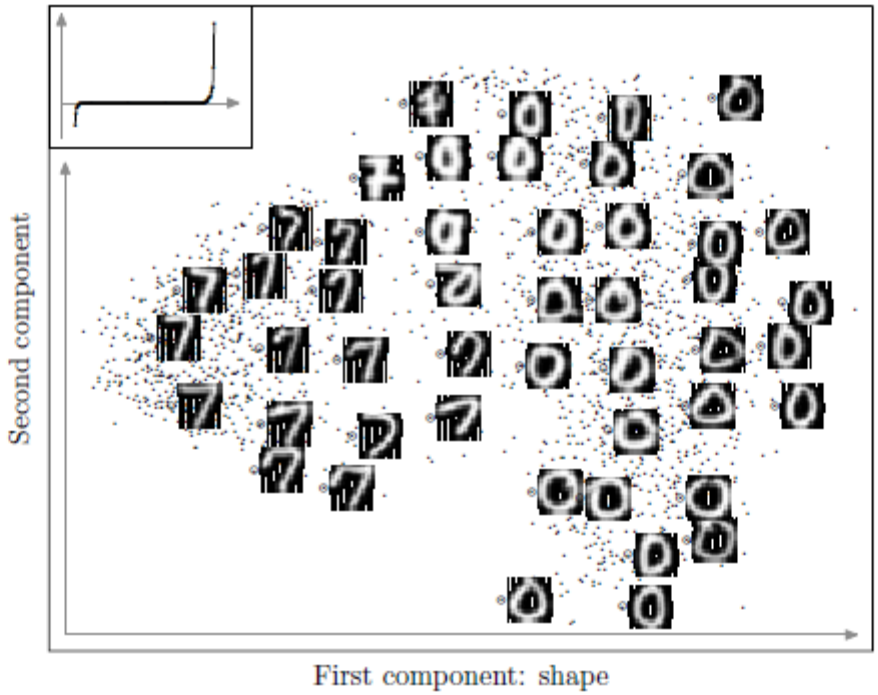


Figure 2: Information coded by metricity.

Feature Discovery: Examples

1. Projection onto the positive and projection onto the negative eigenspaces yield results different in nature.

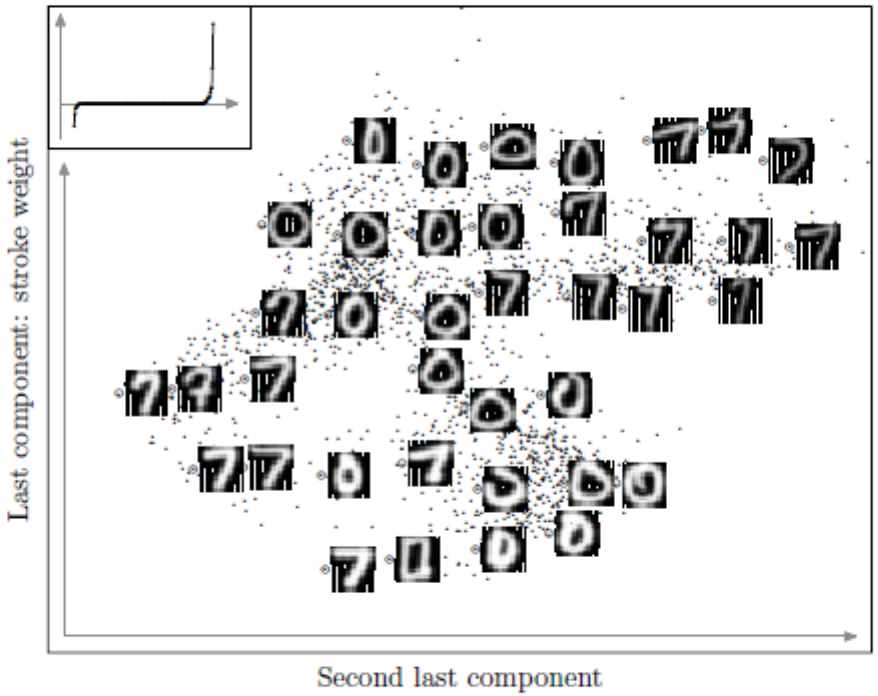


Figure 3: Information coded by metric violations.

- We are interested in the semantic structure of nouns and adjectives from different text sources.
- A subset of 120 nouns and adjectives has been selected, containing both very specific and very general terms out of two topically unrelated sources: Grimm's Fairy Tales[a], and popular science articles about space exploration[b].
- Both sources contributed 60 documents containing roughly between 500 and 1500 words each.
- For p documents and a choice of n keywords construct the contingency table, by simply indicating whether word i ($1 \leq i \leq n$) appears in document k ($1 \leq k \leq p$) or not.

[a] Project Gutenberg `http://promo.net/pg/`

[b] Science at Nasa articles `http://science.nasa.gov/headlines/news_archive.`

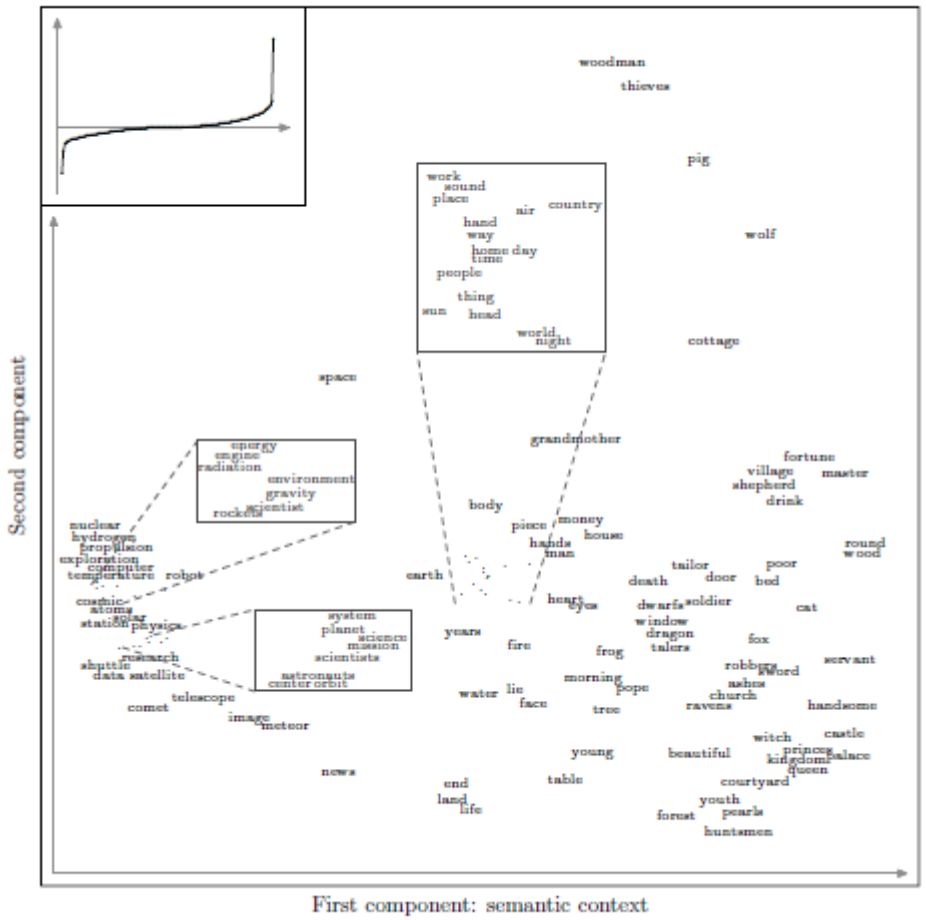
– This yields an $p \times n$ boolean matrix X, with $X_{ki} = 1$ if word i appears in document k and 0 else. Let X_i denote the ith column of X (associated to word i).

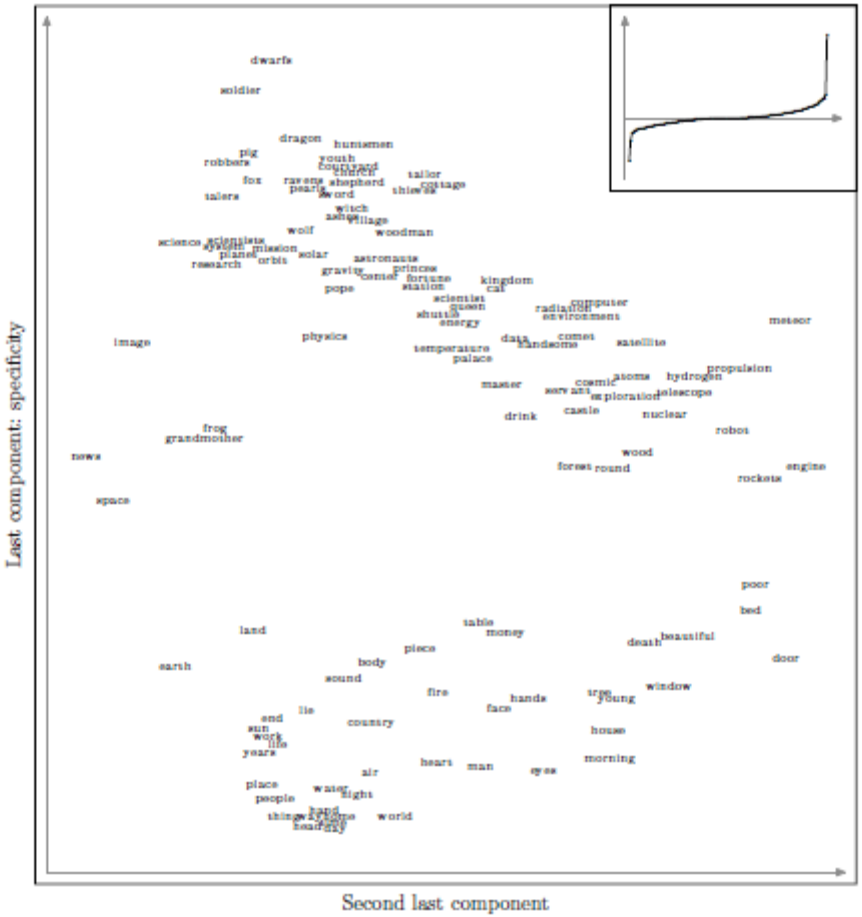
– *Keyword Semantic Proximity*:

$$s_{ij} = \frac{\#\{\text{documents where word } i \textit{ and } \text{word } j \text{ appear}\}}{\#\{\text{documents where word } i \textit{ or } \text{word } j \text{ appear}\}}$$

$$= \frac{\sum_{X_i + X_j = 2} 1}{\sum_{X_i = 1} 1 + \sum_{X_j = 1} 1 - \sum_{X_i + X_j = 2} 1}$$

– From this similarity measure, we obtain a dissimilarity matrix via, e.g. $d_{ij} = -\log(s_{ij})$.

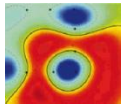

woodman
thieves
pig
work
sound
place
air
country
hand
way
home day
time
people
thing
sun
head
world
night
wolf
cottage
space
grandmother
fortune
village
master
shepherd
drink
energy
engine
radiation
environment
gravity
scientist
rocket
body
piece
money
house
hands
man
round
wood
nuclear
hydrogen
propulsion
exploration
computer
temperature
robot
earth
tailor
poor
death
door
bed
heart
eyes
dwarfs
soldier
cat
window
dragon
talers
fox
servant
system
planet
science
mission
scientists
astronauts
center orbit
years
fire
frog
robbers
sword
ashes
shuttle
research
data satellite
morning
pope
church
ravens
handsome
water
lie
face
tree
telescope
comet
image
meteor
witch
castle
young
beautiful
princess
kingdom
palace
queen
news
table
courtyard
end
land
life
youth
forest
pearls
huntsmen
Second component
First component: semantic context


Last component: specificity
Second last component
dwarfs
soldier
dragon
humanmen
pig
robbers
youth
fox
ravens
shepherd
cottage
tailor
pearls
sword
witch
village
woodman
wolf
science
scientists
mission
planet
orbits
research
solar
astronauts
gravity
prince
fortune
kingdom
pope
station
scientist
queen
shuttle
energy
computer
radiation
environment
meteor
image
physics
temperature
palace
data
comet
handsome
satellite
propulsion
manner
cosmic
atoms
hydrogen
servant
exploration
telescope
drink
castle
nuclear
robot
frog
grandmother
news
wood
forest
round
engine
rockets
space
poor
bed
table
money
land
death
beautiful
piece
door
earth
body
sound
window
fire
hands
face
young
lie
end
country
sun
work
life
house
years
heart
man
eyes
morning
air
place
water
light
people
world

Conclusion: non-metricity in general

Julian Laub, Klaus-Robert Müller Feature Discovery in Non-Metric Pairwise Data, JMLR ; 5(Jul):801--818, 2004

1. The current paradigm is wrong/incomplete.
2. Metric violations *can* carry relevant information.
3. A complete data exploratory research must specifically study this information.



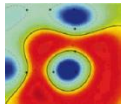
Non-Metric Similarities: continued

- Cannot be modeled faithfully in a metric map
- Metric space has three limitations:
  - Cannot model intransitive similarities
  - Cannot model objects with high centrality
  - Cannot model asymmetric similarities
- Lead Tversky (among others) to reject MDS as a model for semantic representation

[From vd Maaten]

Multiple Maps t-SNE

- Circumvents limitations of most MDS variants by constructing multiple maps instead of a single map:
  - Each object has a point in all maps
  - Each point has a weight in each map
  - The weights sum up to 1 for an object
- Addresses all three limitations!



[From vd Maaten]

Multiple Maps t-SNE

- The input is a collection of $p_{j|i}$'s
- The similarity between i and j under the model is given by $q_{j|i}$:

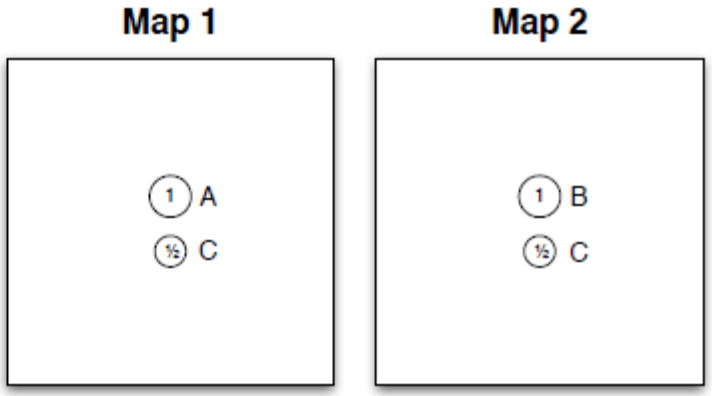
$$q_{j|i} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} \left(1 + ||y_i^{(m)} - y_j^{(m)}||^2\right)^{-1}}{\sum_{m'} \sum_{i \neq k} \pi_i^{(m')} \pi_k^{(m')} \left(1 + ||y_i^{(m')} - y_k^{(m')}||^2\right)^{-1}}$$

- Minimize the sum of KL divergences:

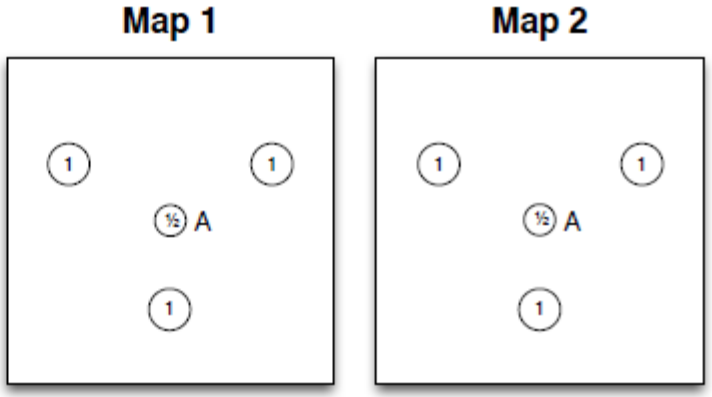
$$\sum_i KL(P_i||Q_i) = \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

[From vd Maaten]

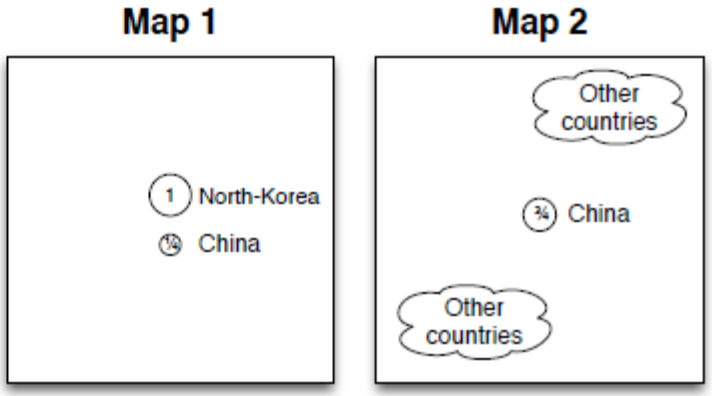
Intransitive similarities:



High centrality:

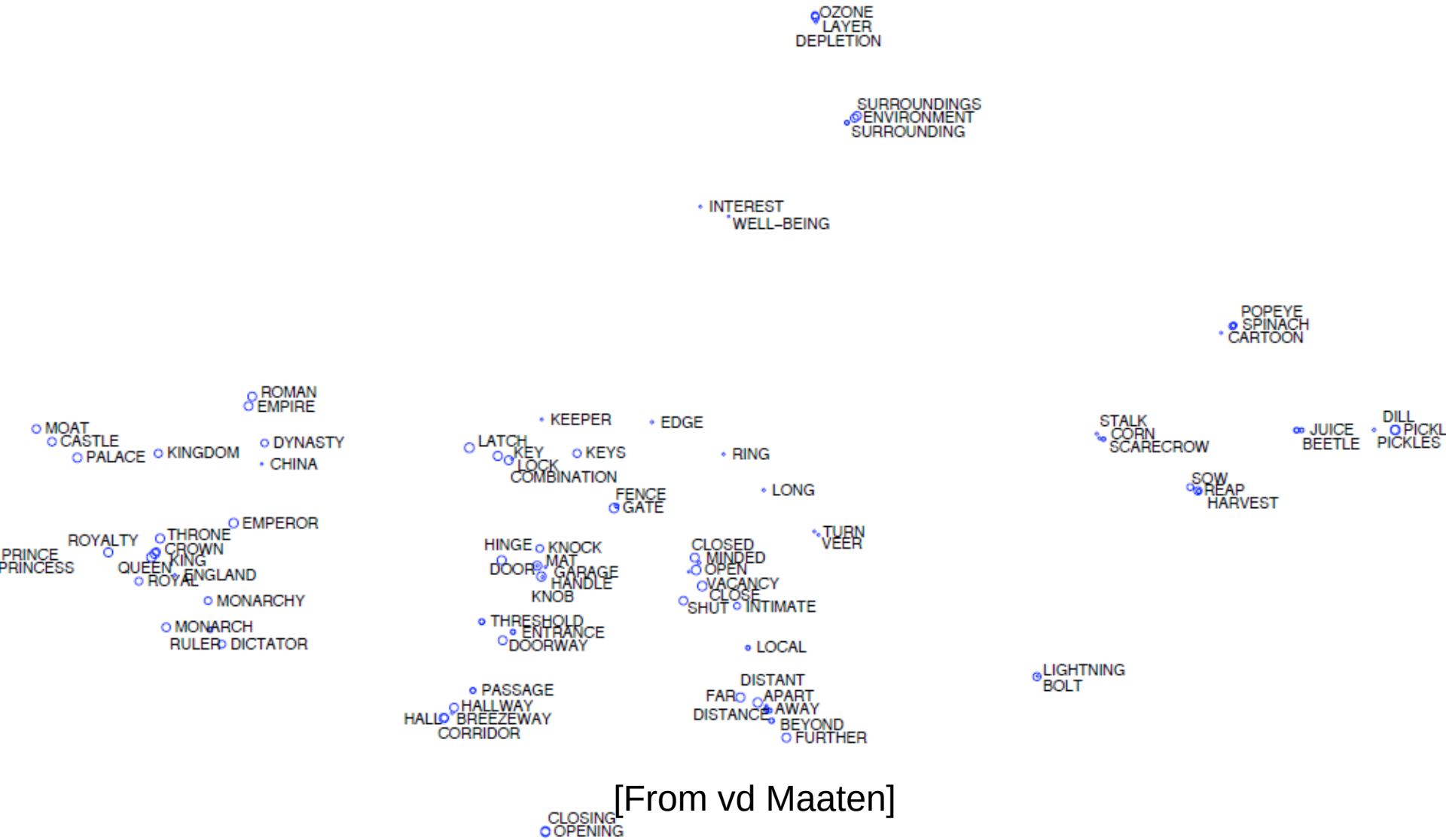


Asymmetries:

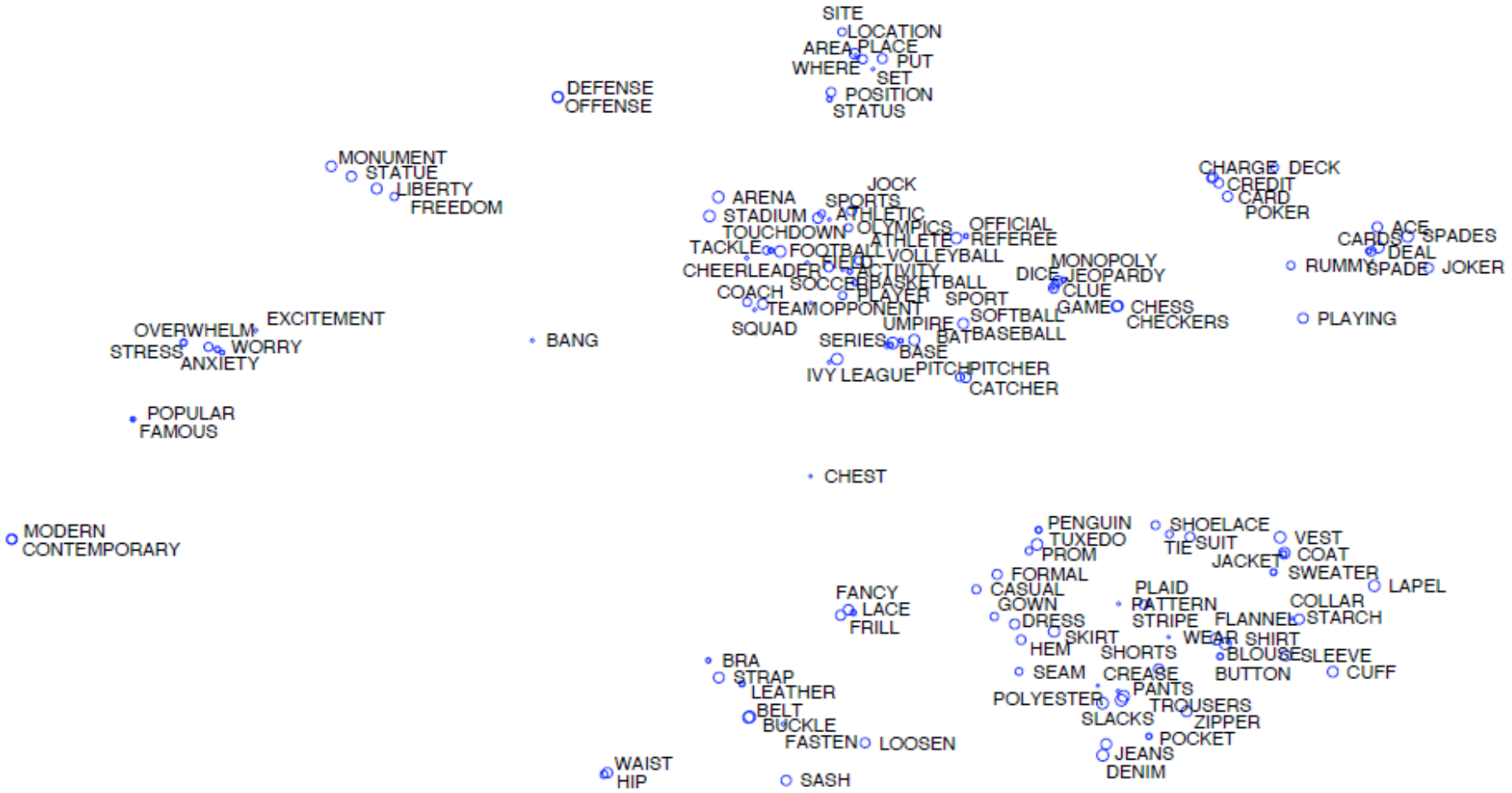


[From vd Maaten]

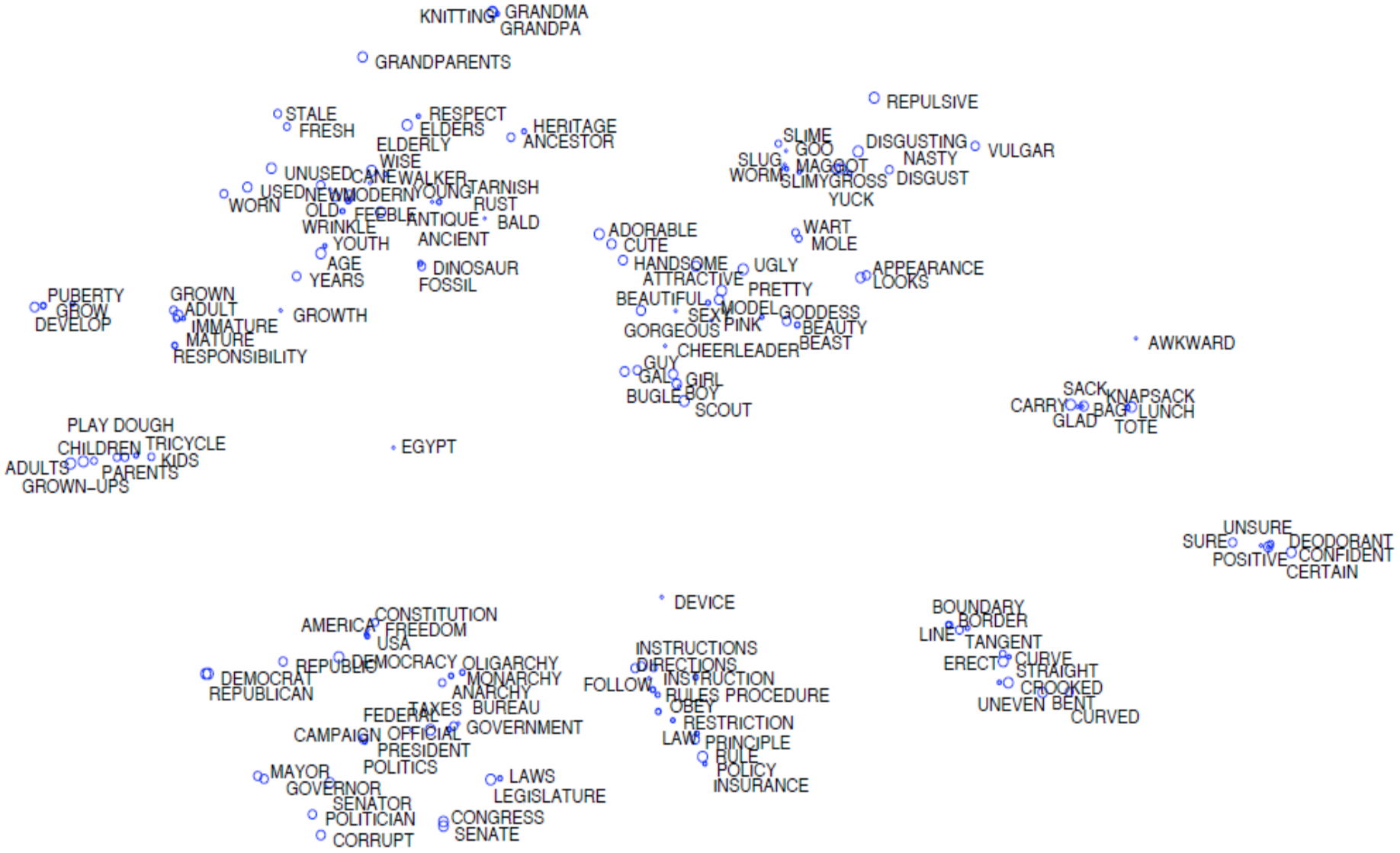
Multiple Maps t-SNE



[From vd Maaten]



[From vd Maaten]



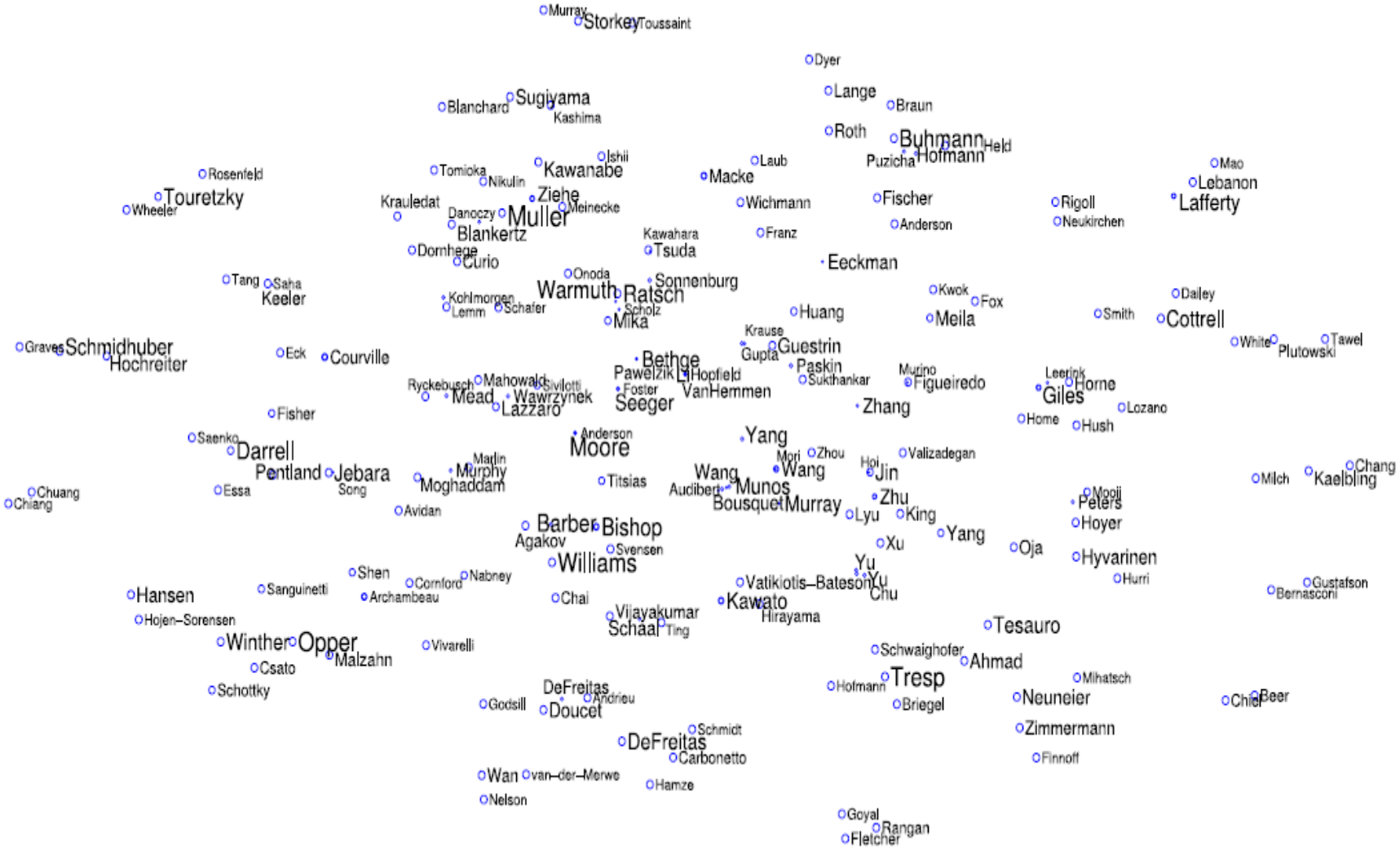
[From vd Maaten]

Modeling NIPS Authors

- Gather the authors of all papers of NIPS volume 1-20
- Remove all authors with only one paper and authors without co-authors
- Compute $p_{j|i}$ as the probability that j is author of a paper of which i is an author
- The similarities are likely to be intransitive and asymmetric: use multiple maps t-SNE!

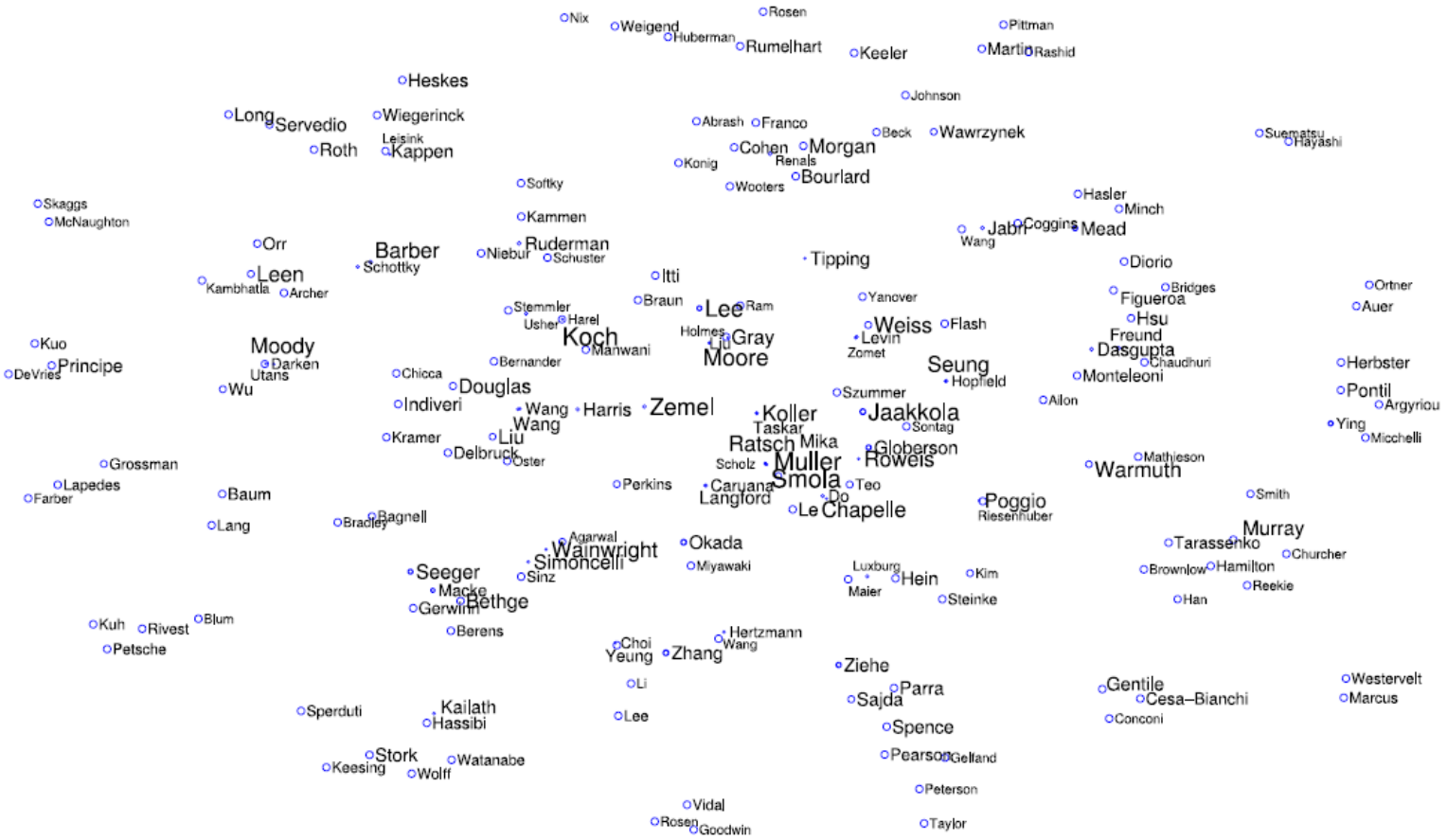
[From vd Maaten]

Map 1



[From vd Maaten]

Map 3



[From vd Maaten]

Quote

From vd Maaten & Hinton Machine Learning 2012

to collaborators from UC Irvine (where he is currently a professor) in map 2. As a second example, Martin Wainwright has collaborated extensively with both Eero Simoncelli and Michael Jordan, but on different topics and at different times. He appears with Simoncelli in map 3 and with Jordan in map 4 thus allowing their representations to remain far apart. As a third example, Klaus-Robert Müller's collaborations until 2000 (with, among others, Alex Smola and Gunnar Rätsch) are visualized in map 3, whereas his collaborations after 2000 (for instance, with Benjamin Blankertz) are shown in map 1.

Figure 5 shows the neighborhood preservation ratio obtained by aspect maps and multiple maps t-SNE for increasing numbers of maps. The results presented in the figure are in

[From vd Maaten]

Refs

- Google: t-SNE (van der Maaten)
- vd Maaten & Hinton Machine Learning 2012
- Laub & Müller JMLR 2004
- Slides on SNE & MDS adapted from Hinton and vd Maaten

- L.J.P. van der Maaten and G.E. Hinton. *Visualizing Data using t-SNE*. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.
- L.J.P. van der Maaten and G.E. Hinton. *Visualizing Similarities with Multiple Maps*.