Dimensionality Reduction

John Joseph Valletta

University of Exeter, Penryn Campus, UK

April 2019

Why reduce the dimensionality of the problem?

- Visualise the data to uncover structure
- Elucidate the best predictors of the underlying process (plausible causal drivers under an experimental setup)
- Improve the model's predictive performance by removing uninformative features/extracting better features
- Decrease computational power

Why reduce the dimensionality of the problem?

- Visualise the data to uncover structure
- Elucidate the best predictors of the underlying process (plausible causal drivers under an experimental setup)
- Improve the model's predictive performance by removing uninformative features/extracting better features
- Decrease computational power

Why reduce the dimensionality of the problem?

- Visualise the data to uncover structure
- Elucidate the best predictors of the underlying process (plausible causal drivers under an experimental setup)
- Improve the model's predictive performance by removing uninformative features/extracting better features
- Decrease computational power

Why reduce the dimensionality of the problem?

- Visualise the data to uncover structure
- Elucidate the best predictors of the underlying process (plausible causal drivers under an experimental setup)
- Improve the model's predictive performance by removing uninformative features/extracting better features
- Decrease computational power

Why reduce the dimensionality of the problem?

- Visualise the data to uncover structure
- Elucidate the best predictors of the underlying process (plausible causal drivers under an experimental setup)
- Improve the model's predictive performance by removing uninformative features/extracting better features
- Decrease computational power

Dimensionality Reduction

Rationale

Although the data may seem high dimensional, the **structure** of the data can be represented by fewer features.

- **Feature extraction**: mapping the original data to a new feature set
- **Feature selection**: selecting a subset of attributes

Dimensionality Reduction

Rationale

Although the data may seem high dimensional, the **structure** of the data can be represented by fewer features.

Dimensionality reduction can be achieved through:

- **Feature extraction**: mapping the original data to a new feature set
- **Feature selection**: selecting a subset of attributes

Dimensionality Reduction

Rationale

Although the data may seem high dimensional, the **structure** of the data can be represented by fewer features.

Dimensionality reduction can be achieved through:

- **Feature extraction**: mapping the original data to a new feature set
- **Feature selection**: selecting a subset of attributes

Dimensionality Reduction

Rationale

Although the data may seem high dimensional, the **structure** of the data can be represented by fewer features.

Dimensionality reduction can be achieved through:

- **Feature extraction**: mapping the original data to a new feature set
- **Feature selection**: selecting a subset of attributes

Principal Component Analysis

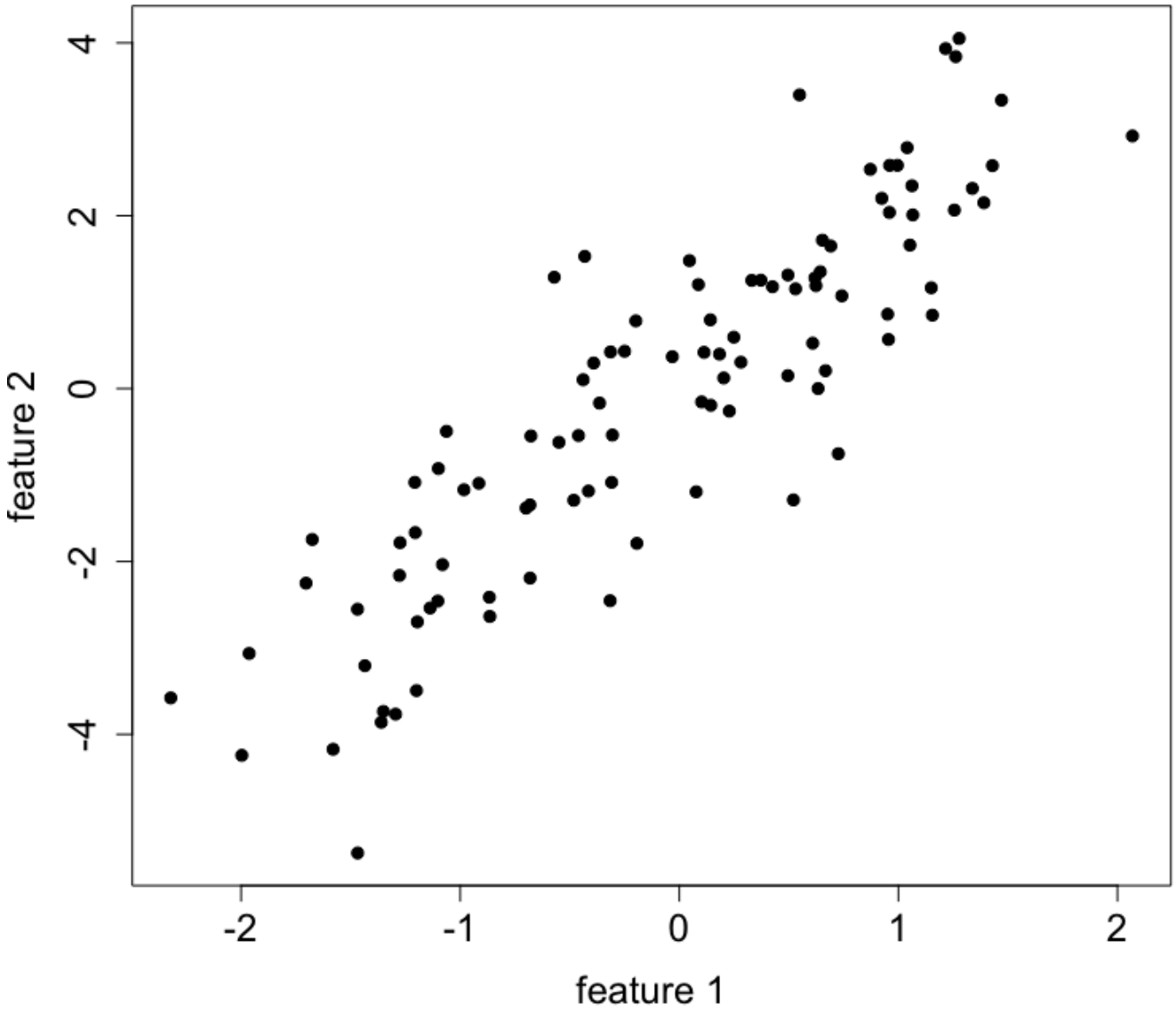
- A **linear** dimensionality reduction method

- The new uncorrelated features (PCA 1, PCA 2,...) are **weighted** (w's) linear combinations of the original data (x's)

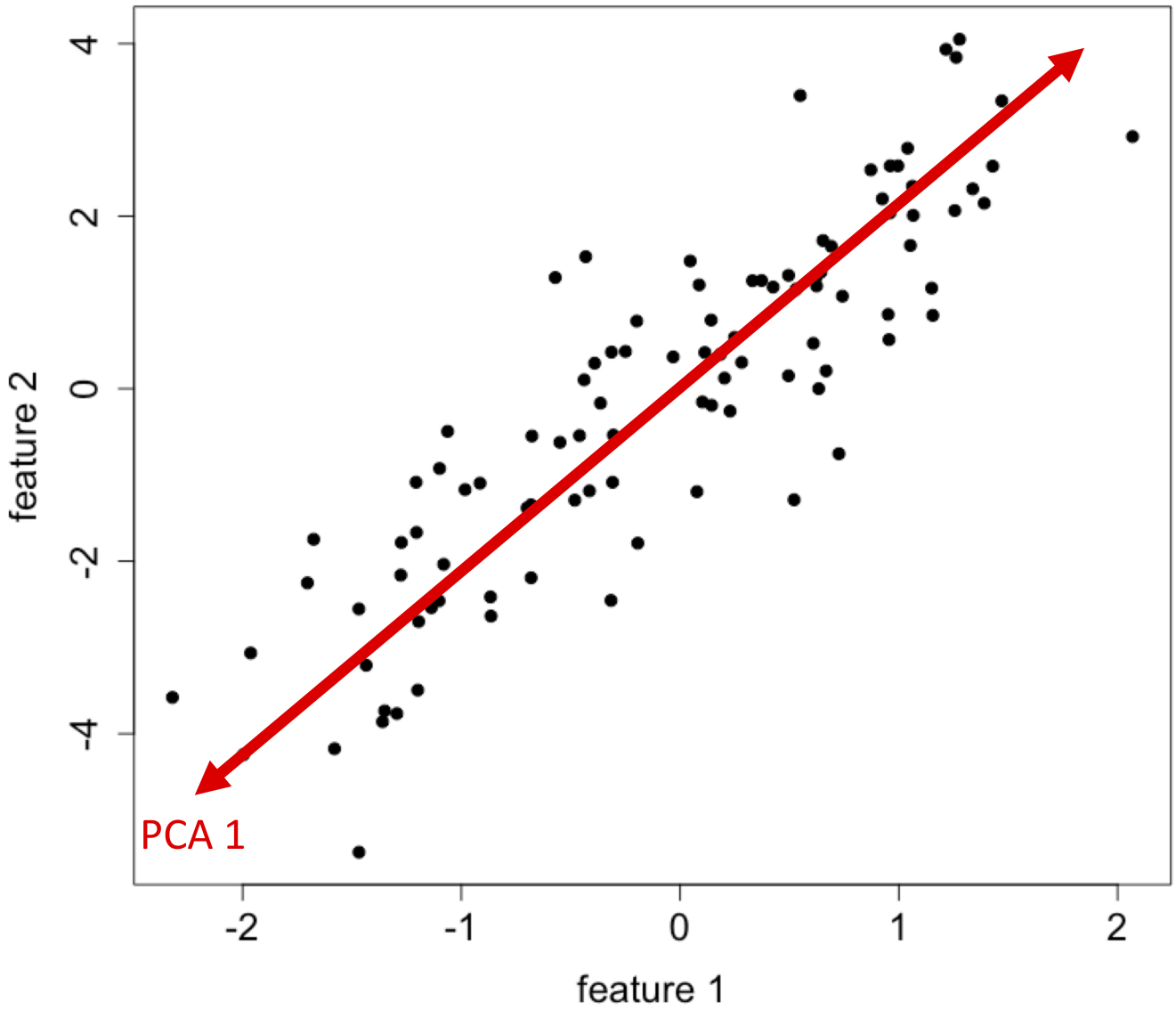
$$\begin{aligned}
\text{PCA 1} &= w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p \\
\text{PCA 2} &= w_{21}x_1 + w_{22}x_2 + \ldots + w_{2p}x_p \\
&\vdots \\
\text{PCA p} &= w_{p1}x_1 + w_{p2}x_2 + \ldots + w_{pp}x_p
\end{aligned}$$

- Objective is to find directions, called principal components, that maximise the variance of the data

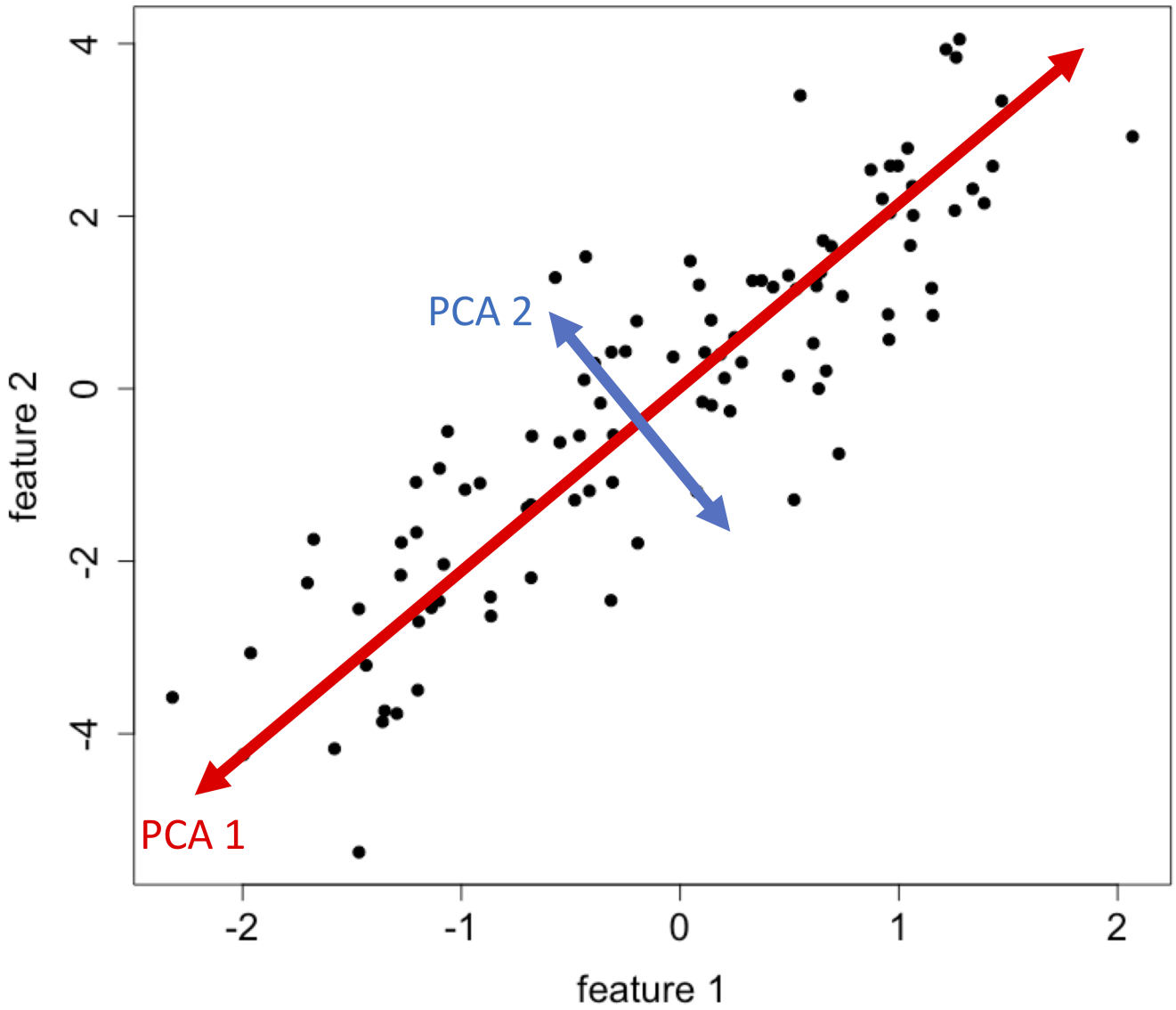
Principal Component Analysis



Principal Component Analysis



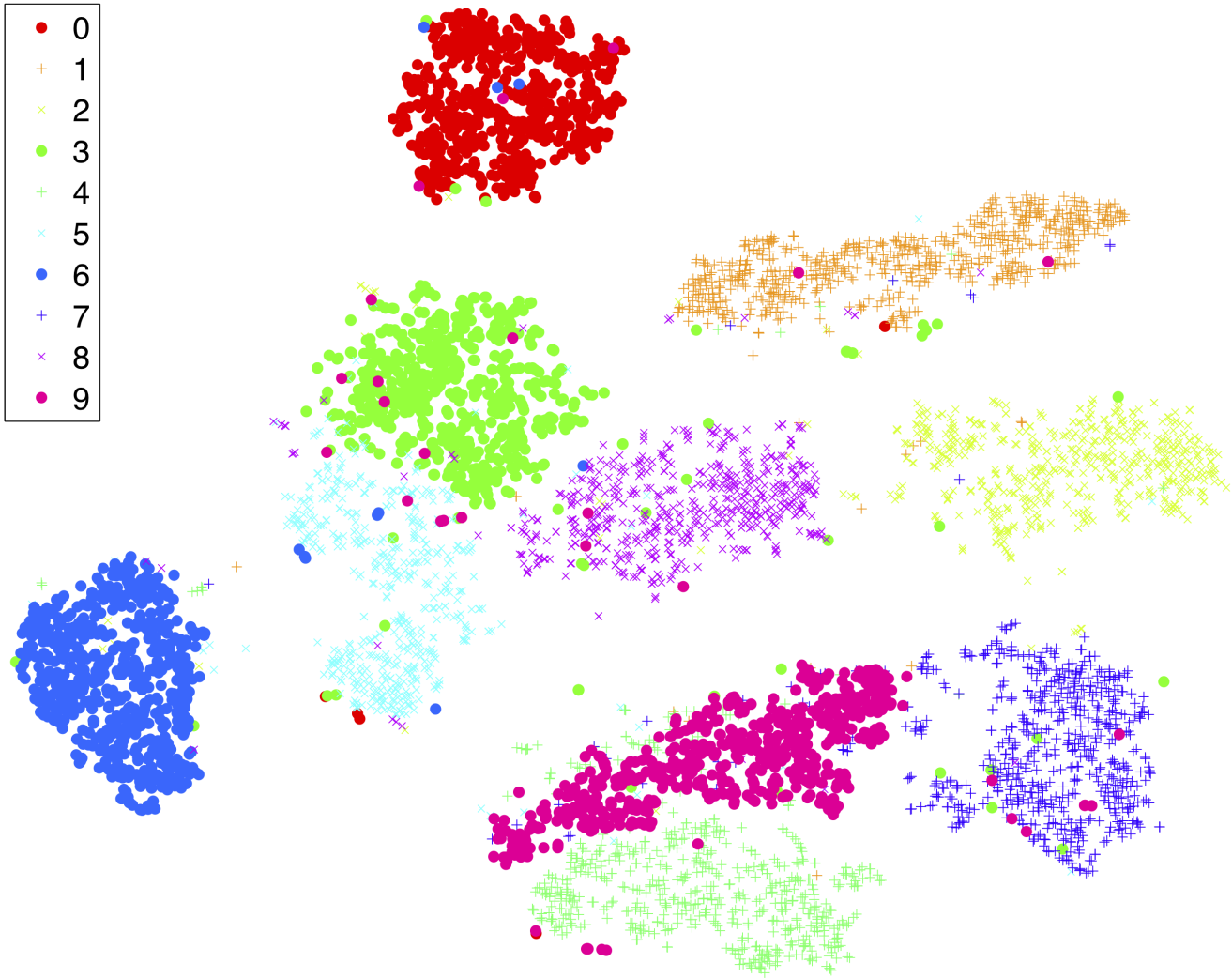
Principal Component Analysis



t-Distributed Stochastic Neighbor Embedding (t-SNE)

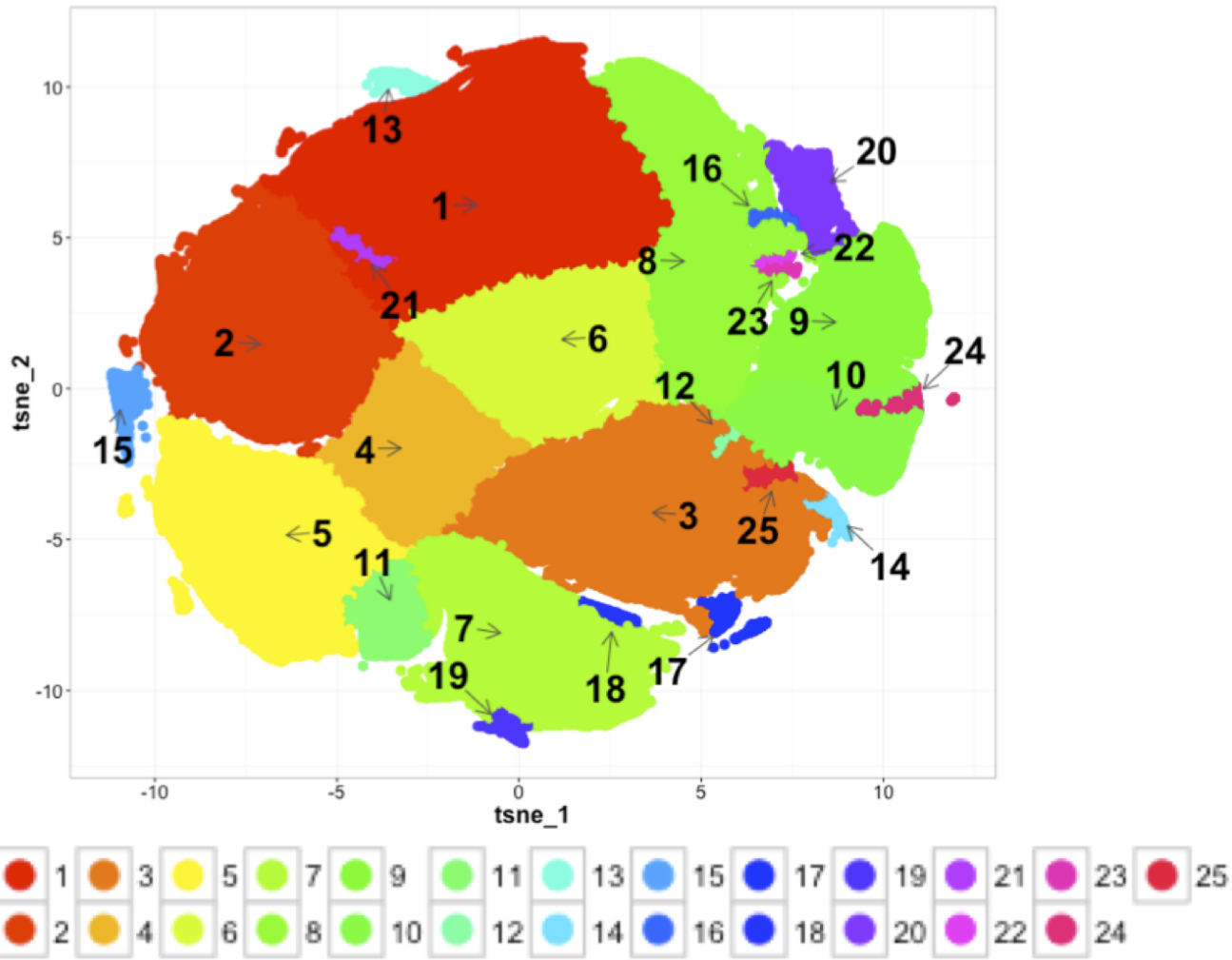
- A **non-linear** dimensionality reduction method
- Projects data into a lower-dimensional space/embedding such that the original high-dimensional clustering is preserved

t-SNE on handwritten digits



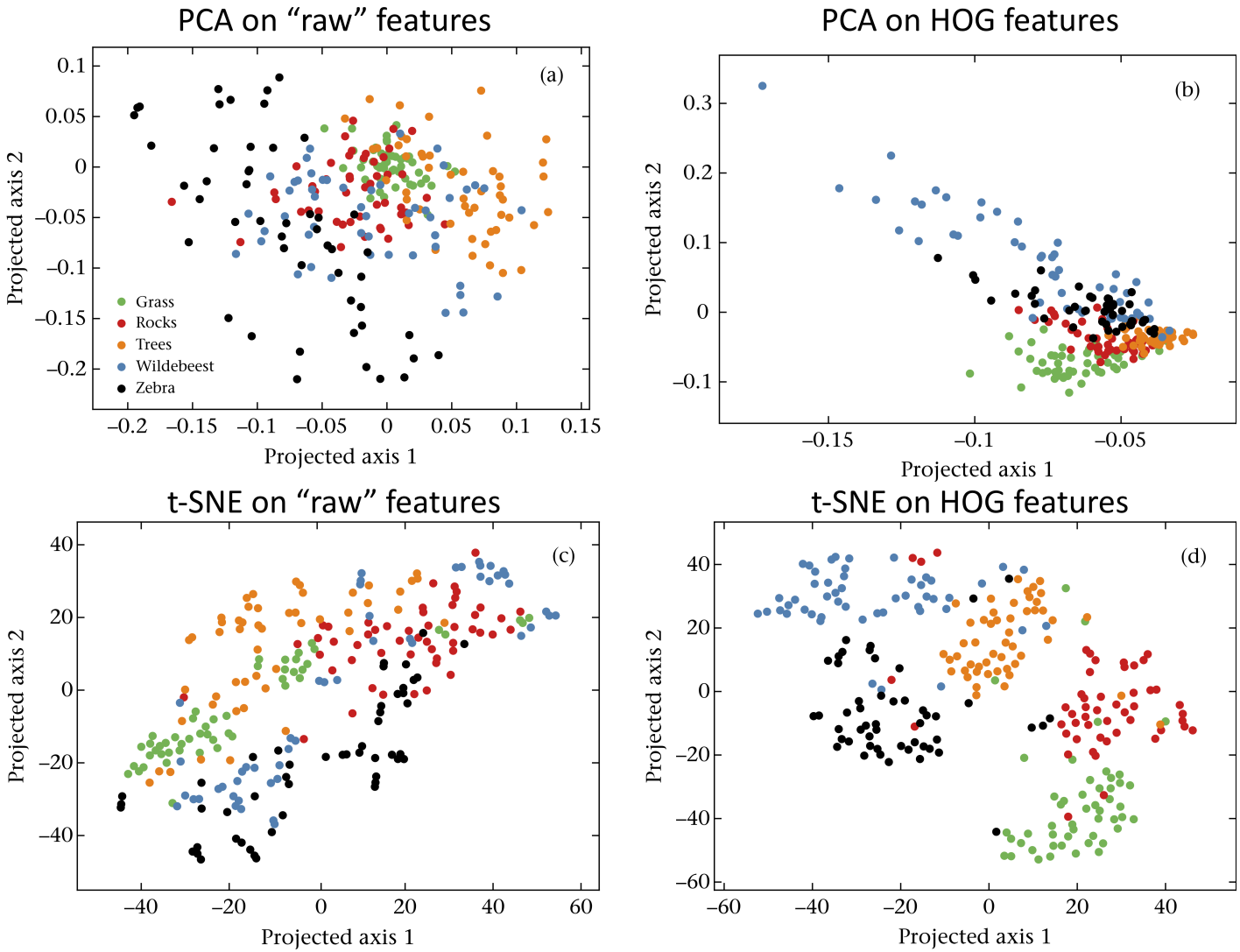
L. Van der Maaten and G. Hinton. (2008) *Journal of Machine Learning Research*

t-SNE on flow cytometry data



Y. Bediako, R. Adams, A. Reid, J.J. Valletta, F. Ndungu *et al.* (2019) *BMC Medicine*

Comparison of methods



J.J. Valletta *et al.* (2017) *Animal Behaviour*