Data Science and Machine Learning in Python

Stephan Weyers

Fachhochschule Dortmund
University of Applied Sciences and Arts

Topics covered in the online lectures

Fachhochschule Dortmund
University of Applied Sciences and Arts

Part 1: Data Science

	Date	Topics covered
1	Apr 13th	Course introduction Data Science motivation How to use Jupyter Notebook Python types and lists Loops, if/else, functions
2	Apr 20th	Python tuples, lists, dictionaries Functions Numpy basics, operations Image processing
3	Apr 27th	Pandas Series, DataFrame Pandas basic operations Import/export files
4	May 4th	Principles of data visualization Data cleaning and preparation Join, combine and reshape data
5	May 11th	Volkswohl Bund dataset Data visualization in Python How to write Data Science reports Data aggregation and grouping

Part 2: Machine Learning

	Date	Topics covered
6	Jun 1st	Introduction to supervised learning Classification and regression scikit-learn k-Nearest Neighbors Linear regression (ridge and lasso)
7	Jun 8th	Linear classification models Decision trees Random forests and gradient boosting
8	Jun 15th	Kernel support vector machines Neural networks
9	Jun 22nd	Introduction to unsupervised learning Preprocessing and scaling Dimensionality reduction Principal component analysis
10	Jun 29th	k-means clustering Hierarchical clustering DBSCAN
11	Jul 6th	Representing data Engineering features Model evaluation and improvement Text data analysis

Deadlines for Submission and Distribution of Grading

Fachhochschule Dortmund
University of Applied Sciences and Arts

Student task	Deliverables	Deadline	Work	Share of grade
W01 Assignment	Code and results	Apr 26th	Team A	5.0%
W02 Case Study	Code / presentation slides	May 22nd	Team B	18.0%
W02 Case Study	Peer review*	May 31st	Individual	2.0%
W03 Assignment	Code and results	May 29th	Team B	5.0%
W04 Assignment	Code and results	Jun 12th	Team C	10.0%
W05 Assignment	Code and results	Jun 26th	Team D	7.0%
W06 Assignment	Code and results	Jul 8th	Team D	13.0%
W07 Case Study	Code / presentation slides	Jul 17th	Team D	22.0%
W07 Case Study	Peer review*	Jul 31st	Individual	3.0%
DataCamp 1	Finish course	May 9th	Individual	2.5%
DataCamp 2	Finish course	May 30th	Individual	2.5%
DataCamp 3	Finish course	Jun 20th	Individual	2.5%
DataCamp 4	Finish course	Jul 11th	Individual	2.5%

* Peer review is mandatory. Quality of peer review itself is graded. Not providing peer review at all would result in high point deduction

Teams for assignment W05-W07

Team	Univ.	Name
D1	UV	Paula Piña
D1	UBA	Facundo Ignacio Zanalda
D1	UBA	Manuel Cabeza Galucci
D1	FHDO	Daniel Tobien
D2	UV	Adonis Nicola Cruz Navarrete
D2	UBA	Manuel Durán
D2	UBA	Lucas Trabanco
D2	FHDO	Bedirhan Abaz
D3	UV	Felipe Galdames
D3	UBA	Victoria Marquez
D3	FHDO	Minh Quan Dinh
D3	ESAN	Juan Jose A. Velasquez Leon
D4	UBA	Gian Franco Lancioni
D4	UBA	Kevin Michalewicz
D4	FHDO	Mohamed Elbaraka
D4	ESAN	Nayely Mayli Ore Ichpas
D5	UV	Nilari Berger Díaz
D5	UBA	Daniel Kundro
D5	UBA	Belen Ticona
D5	FHDO	Celine Cramer
D6	UBA	Francisco Rossi
D6	FHDO	René Frackmann
D6	FHDO	Jessica Heilig
D6	FHDO	Marius Meiners
D7	UV	Valentina Andrea Acuña Ponce
D7	UBA	Sofía Nieva
D7	FHDO	Fabian Herberholt
D7	FHDO	Arnold Urbanio Olympio

Team	Univ.	Name
D8	UV	Manuel Orellana Hinojosa
D8	UV	Dian Arriagada
D8	UGTO	Abraham Morales Iturriaga
D8	ESAN	María Ximena Latorre Guzmán
D9	UV	Luis Martinez
D9	UV	Paula Riquelme
D9	UDEM	Jordana L.M. Apolinario Simon
D9	FHDO	Robin Drabon
D10	UV	Jaime Godoy
D10	UDEM	Mariana Gómez Gómez
D10	UBA	Francisco Alan Luna
D10	FHDO	Marco Vom Bovert
D11	UV	Joel Santana
D11	UV	Paula Toro
D11	UBA	Lucía Ailén Kasman
D11	UBA	Rocío Palacín Roitbarg
D11	FHDO	Intissar Boudi
D12	UV	Marcelo Leiton
D12	UV	Emmanuel Cuevas Parra
D12	UBA	Matías Nicolás Pereyra
D12	FHDO	Mamadama Cherif
D12	ESAN	Luiggy Johan Zea Guzman
D13	UV	Dietrich Ganz
D13	UV	Rodrigo Llano Orellana
D13	UBA	Juan Cruz Camacho
D13	FHDO	Justin Skupsch
D13	FHDO	Marco Kusnierek

Team	Univ.	Name
D14	UV	Jorge Rodriguez
D14	UV	Alejandra Valencia
D14	UTTEC	Hugo Isaac Vázquez Gutiérrez
D14	UDEM	Dilan Stiven Correa López
D14	UBA	Andrómeda P. Ovalles Castro
D15	UV	Diego Del Rio
D15	UV	Franco Garrido
D15	UTTEC	José Luís Godínez Vázquez
D15	FHDO	Tegar Fathir Muhammad
D15	ESAN	Jhossy J. Vargas Saldaña
D16	UV	Jose Ignacio Meneses Castillo
D16	UV	Benjamin Serra
D16	UV	Sofia Contreras Figueroa
D16	UGTO	Andrea Rodriguez Sotelo
D16	FHDO	Jakub Bogusz
D17	UV	Amaya Arroyo
D17	UV	Catalina Escobar
D17	UGTO	Frida Martinez Flores
D17	UBA	Victoria Cambriglia
D17	FHDO	Jannick Bröring
D18	UV	Maximiliano Arancibia Santana
D18	UV	Fernando Parada
D18	UGTO	Andrea Ortiz Alvarado
D18	UBA	Joaquin Ceppi
D18	ESAN	Angela Karin Paredes Solano

Agenda for online lecture 9

Fachhochschule Dortmund
University of Applied Sciences and Arts

Session	Topic	Mode	Materials used	Minutes	End
14:30-16:00	Organizational questions	Q&A		10	14:40
	Scaling of Data	Lecture / Q&A	Lecture slides	10	14:50
	Curse of Dimensionality	Lecture / Q&A	Lecture slides	5	14:55
	Projections and PCA	Lecture / Q&A	Lecture slides	20	15:15
	Questions for discussion	Team work in break-out rooms	Lecture slides	25	15:40
	Olivetti Faces	Lecture / Q&A	Lecture slides	10	15:50
16:10-17:40	Olivetti Faces	Lecture / Q&A	Lecture 09a notebook	15	16:25
	Manifold learning, t-SNE	Lecture / Q&A	Lecture slides	5	16:30
	MNIST with PCA / t-SNE	Joint coding in main room	Lecture 09b notebook	60	17:30
17:50-19:20	Questions for discussion	Team work in break-out rooms	Lecture slides	30	18:20
	OCEAN Personality Traits	Lecture / Q&A	Lecture slides	15	18:35
	OCEAN Personality Traits	Lecture / Q&A	Lecture 09c notebook	15	18:50
	Organizational questions	Q&A		10	19:00

Types of problems

Supervised Approaches

- Labeled data
- Target values known

Classification

- Predict category

Regression

- Predict numeric value

Unsupervised Approaches

- Unlabeled data
- No target value provided

Cluster Analysis

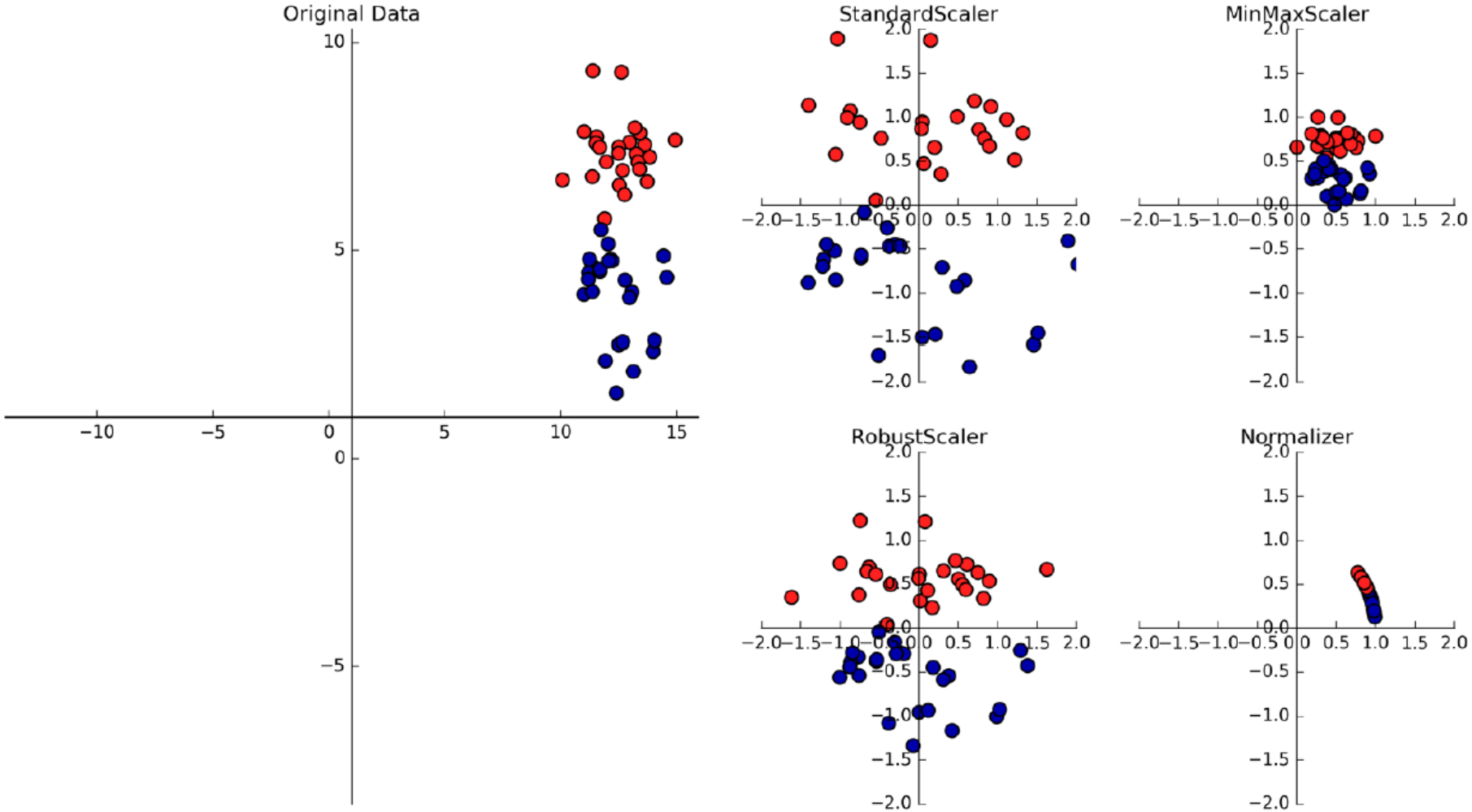
- Organize similar cases into segments

Dimensionality reduction

- Reduce number of features

Scaling of Data

Fachhochschule Dortmund
University of Applied Sciences and Arts



Source: Müller, A. C., Guido, S. (2016) Introduction to Machine Learning with Python : A Guide for Data Scientists, O'Reilly Media

Python Code StandardScaler

Fachhochschule Dortmund
University of Applied Sciences and Arts

In [7]:
```
matrix = [[0,4,-1],[10,6,-1],[20,2,2],[10,4,2]]
df = pd.DataFrame(matrix, columns=["X1","X2","X3"])
df
```

Out[7]:

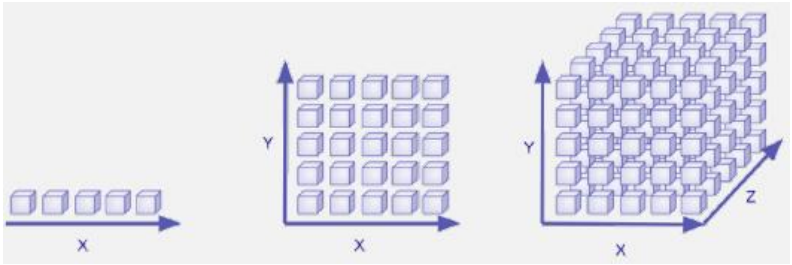
	X1	X2	X3
0	0	4	-1
1	10	6	-1
2	20	2	2
3	10	4	2

In [8]:
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(df)
df_scaled = scaler.transform(df)
df_scaled = pd.DataFrame(df_scaled,columns=df.columns)
df_scaled
```

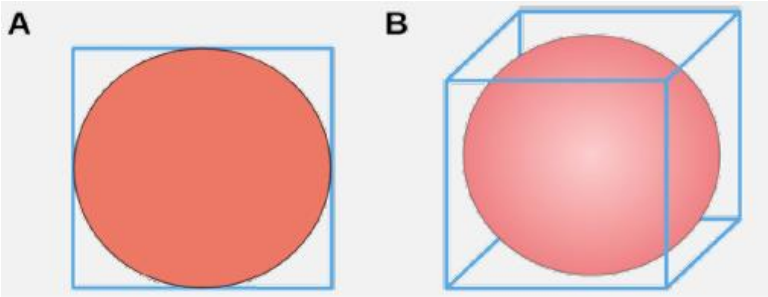
Out[8]:

	X1	X2	X3
0	-1.414214	0.000000	-1.0
1	0.000000	1.414214	-1.0
2	1.414214	-1.414214	1.0
3	0.000000	0.000000	1.0

Curse of Dimensionality – Too Many Features



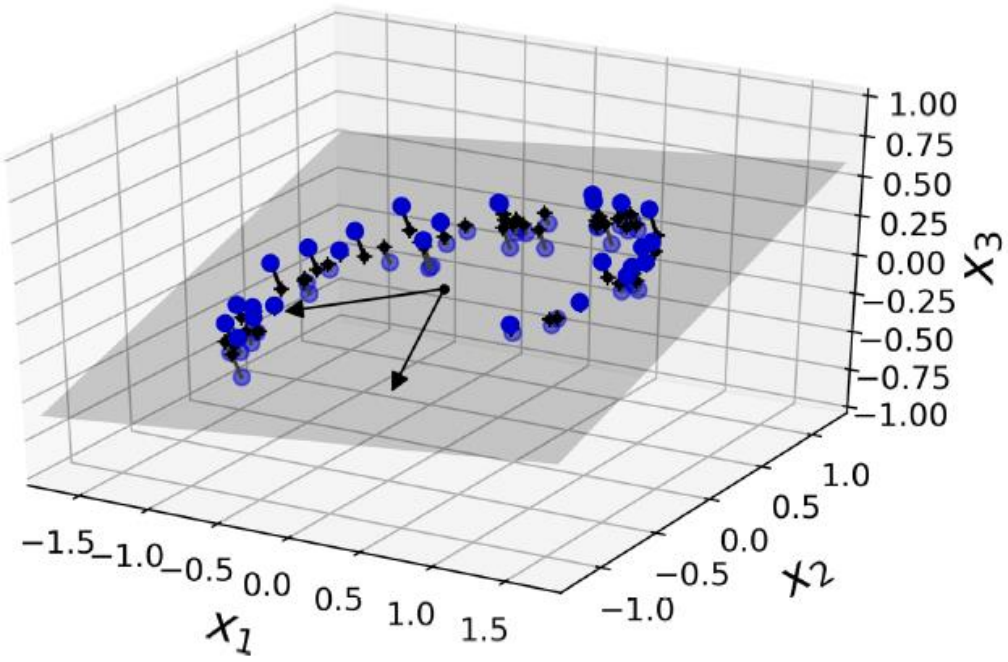
p	# combinations for 5 groups in p-dim
1	5
2	25
3	125
5	3,125
10	9,765,625
15	30,517,578,125

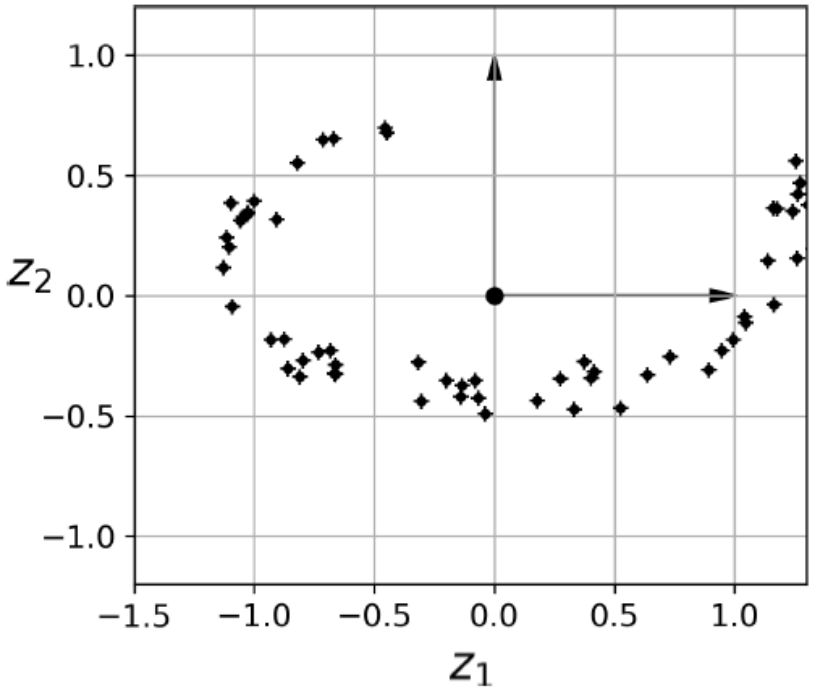


p	p-dim volume of p-ball with diameter 1
1	100.000%
2	78.540%
3	52.360%
4	30.843%
5	16.449%
10	0.249%
15	0.001%

Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

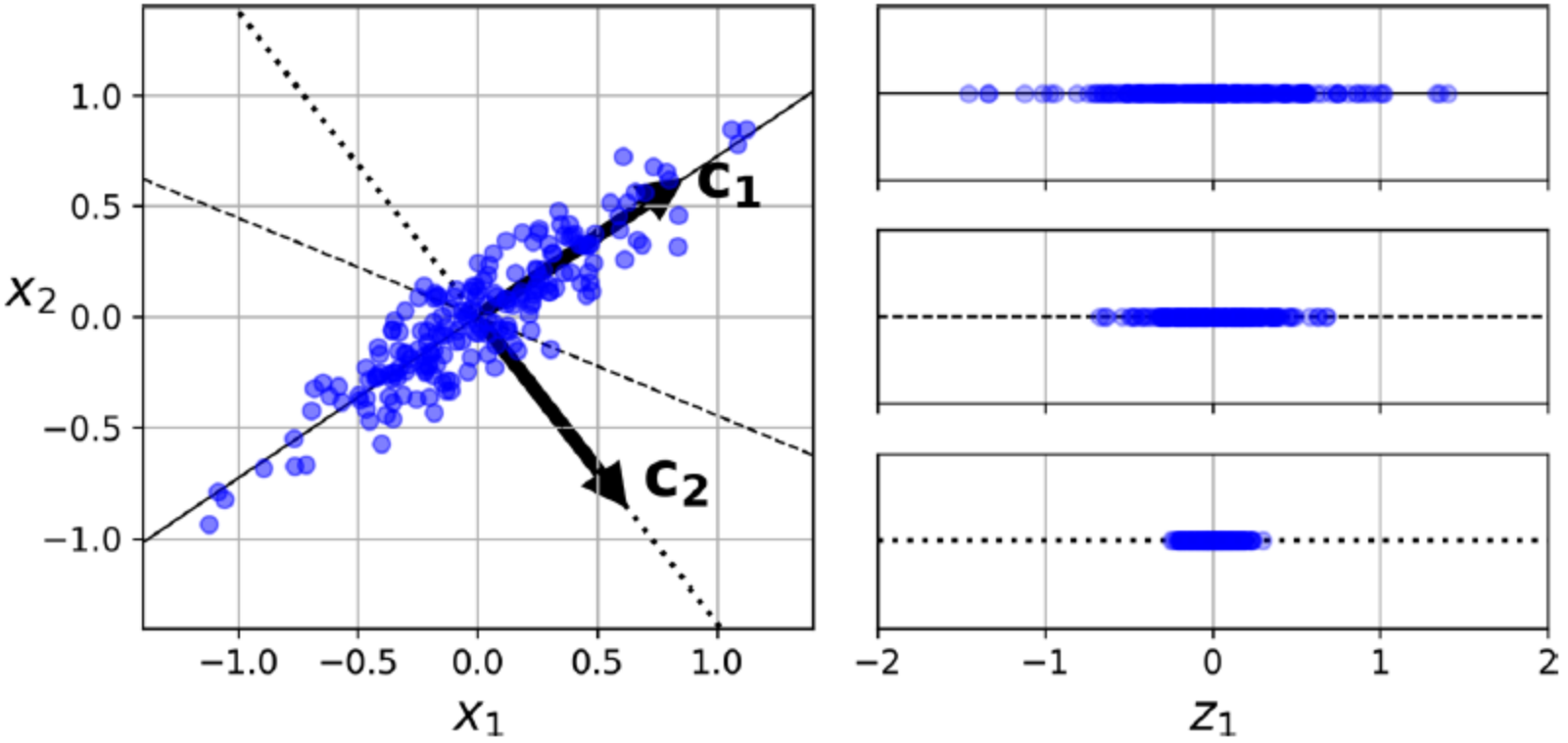
Projections





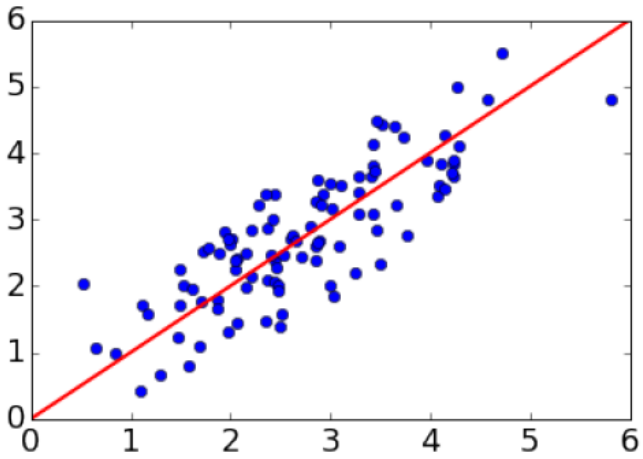
Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Projection components

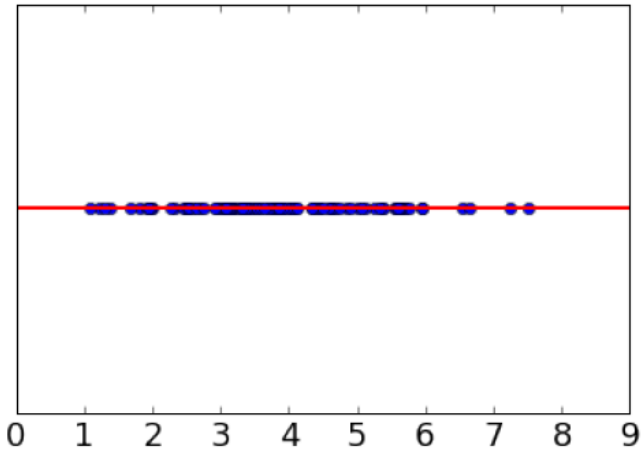


Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

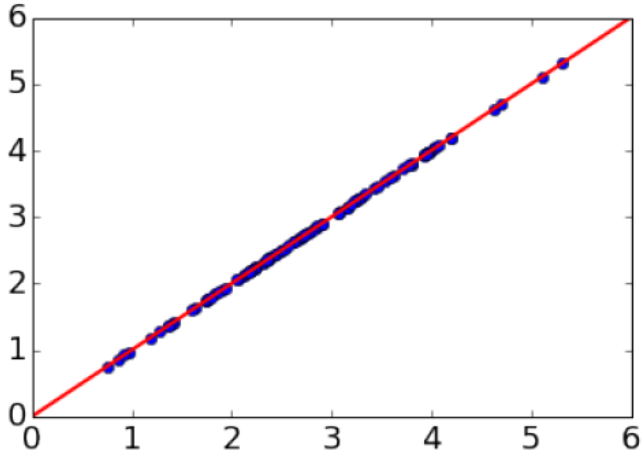
Projection and Reconstruction



Projection onto $\mathbb{R}$:



Reconstruction in $\mathbb{R}^2$:



Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals

Principal Components Mechanics (1/2)

Original data	Adjusted by mean	Projection
$\begin{pmatrix} 2 & 4 & -1 \\ -3 & 2 & 0 \\ 8 & -2 & 6 \\ 1 & -3 & 4 \\ 4 & 4 & 2 \end{pmatrix}$	$\begin{pmatrix} -0.4 & 3.0 & -3.2 \\ -5.4 & 1.0 & -2.2 \\ 5.6 & -3.0 & 3.8 \\ -1.4 & -4.0 & 1.8 \\ 1.6 & 3.0 & -0.2 \end{pmatrix}$	$\begin{pmatrix} -3.45 & -2.57 & 0.96 \\ -5.41 & 2.37 & -0.35 \\ 7.37 & -0.73 & 0.11 \\ 1.91 & 4.19 & 0.13 \\ -0.42 & -3.27 & -0.87 \end{pmatrix}$
Mean $(2.4 \quad 1 \quad 2.2)$	$(0 \quad 0 \quad 0)$	$(0 \quad 0 \quad 0)$
Var $(13.0 \quad 8.8 \quad 6.7)$	$(13.0 \quad 8.8 \quad 6.7)$	$(19.8 \quad 8.2 \quad 0.4)$

Principial components in rows

$$\begin{pmatrix} 0.69 & -0.47 & 0.55 \\ -0.66 & -0.72 & 0.21 \\ 0.30 & -0.50 & -0.81 \end{pmatrix}$$

Principal Components Mechanics (2/2)

Original data

$$\begin{pmatrix} 2 & 4 & -1 \\ -3 & 2 & 0 \\ 8 & -2 & 6 \\ 1 & -3 & 4 \\ 4 & 4 & 2 \end{pmatrix}$$

Projection

$$\begin{pmatrix} -3.45 & -2.57 \\ -5.41 & 2.37 \\ 7.37 & -0.73 \\ 1.91 & 4.19 \\ -0.42 & -3.27 \end{pmatrix}$$

Reconstructed data

$$\begin{pmatrix} 1.71 & 4.48 & -0.22 \\ -2.90 & 1.83 & -0.28 \\ 7.97 & -1.94 & 6.09 \\ 0.96 & -2.93 & 4.11 \\ 4.26 & 3.56 & 1.30 \end{pmatrix}$$

Mean

$$\begin{pmatrix} 2.4 & 1 & 2.2 \end{pmatrix}$$

Principial components in rows

$$\begin{pmatrix} 0.69 & -0.47 & 0.55 \\ -0.66 & -0.72 & 0.21 \end{pmatrix}$$

Python Code Principal Components (1/2)

Fachhochschule Dortmund
University of Applied Sciences and Arts

```
data
```

	X1	X2	X3
0	2	4	-1
1	-3	2	0
2	8	-2	6
3	1	-3	4
4	4	4	2

```
from sklearn.decomposition import PCA
pca = PCA(n_components=3)
pca.fit(data)
```

```
PCA(n_components=3)
```

```
pca.components_
```

```
array([[ 0.69, -0.47,  0.55],
       [-0.66, -0.72,  0.21],
       [ 0.3 , -0.5 , -0.81]])
```

```
pca.transform(data)
```

```
array([[-3.45, -2.57,  0.96],
       [-5.41,  2.37, -0.35],
       [ 7.37, -0.73,  0.11],
       [ 1.91,  4.19,  0.13],
       [-0.42, -3.27, -0.87]])
```

```
pca.explained_variance_ratio_
```

```
array([0.7 , 0.29, 0.01])
```

Python Code Principal Components (2/2)


```
data
```

	X1	X2	X3
0	2	4	-1
1	-3	2	0
2	8	-2	6
3	1	-3	4
4	4	4	2

```
pca = PCA(n_components=2).fit(data)
print(pca.components_)
print("\n")
print(pca.transform(data))
print("\n")
print(pca.inverse_transform(pca.transform(data)))
```

```
[[ 0.69 -0.47  0.55]
 [-0.66 -0.72  0.21]]


[[-3.45 -2.57]
 [-5.41  2.37]
 [ 7.37 -0.73]
 [ 1.91  4.19]
 [-0.42 -3.27]]


[[ 1.71  4.48 -0.22]
 [-2.9   1.83 -0.28]
 [ 7.97 -1.94  6.09]
 [ 0.96 -2.93  4.11]
 [ 4.26  3.56  1.3 ]]
```

Questions

Questions for discussion

1. What are the main motivations for reducing a dataset's dimensionality? What are the main drawbacks?

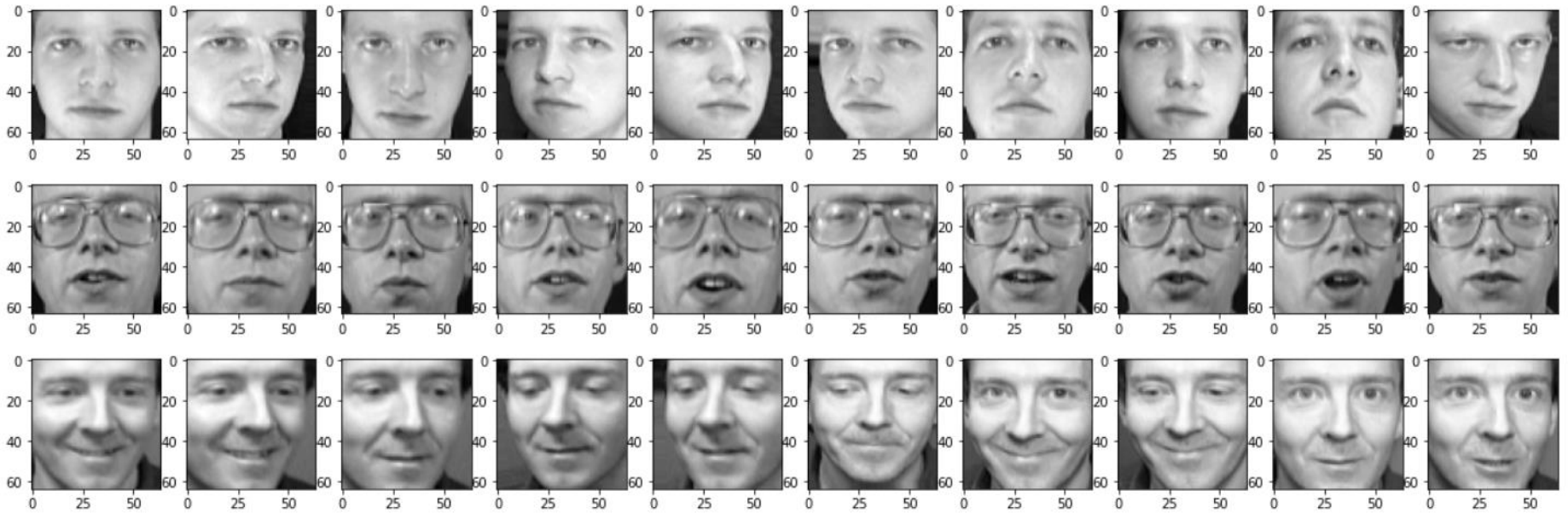
2. What is the curse of dimensionality?

3. Once a dataset's dimensionality has been reduced, is it possible to reverse the operation? If so, how? If not, why?

Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

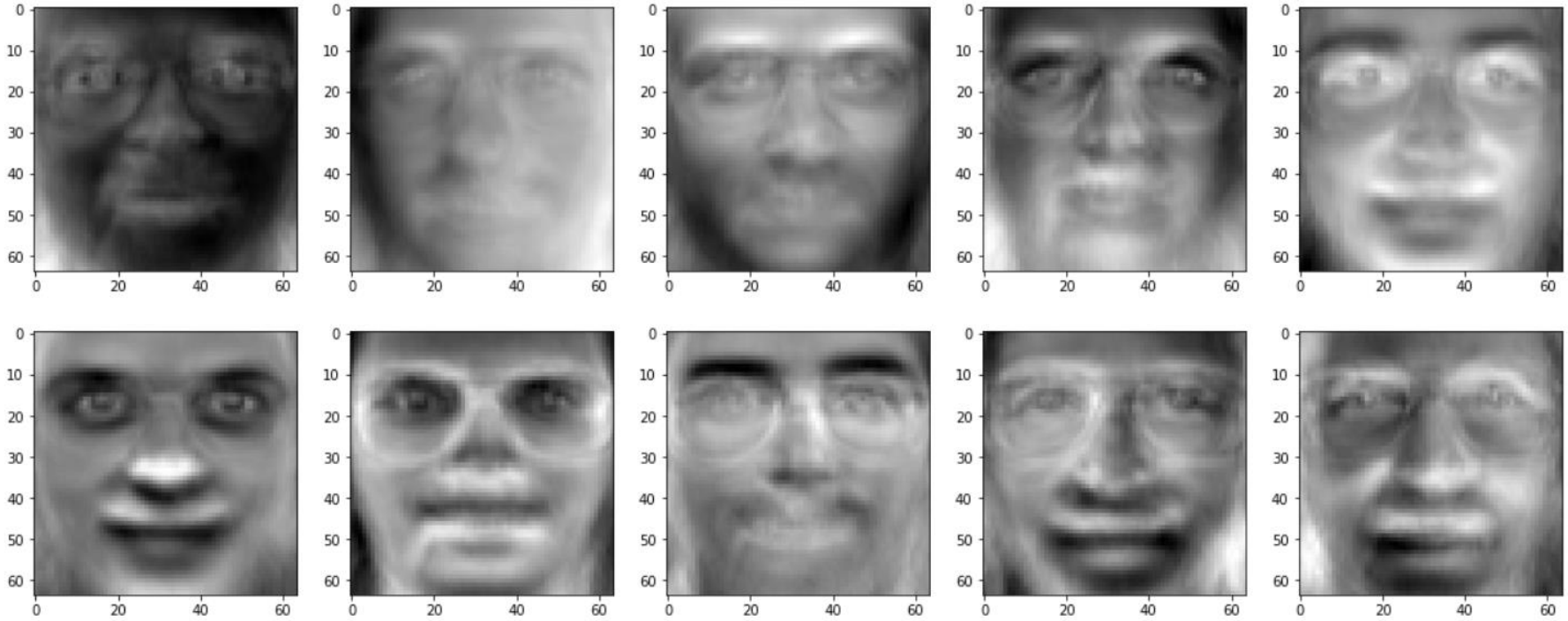
Olivetti Faces – Dataset

Fachhochschule Dortmund
University of Applied Sciences and Arts



Olivetti Faces – EigenFaces

Fachhochschule Dortmund
University of Applied Sciences and Arts

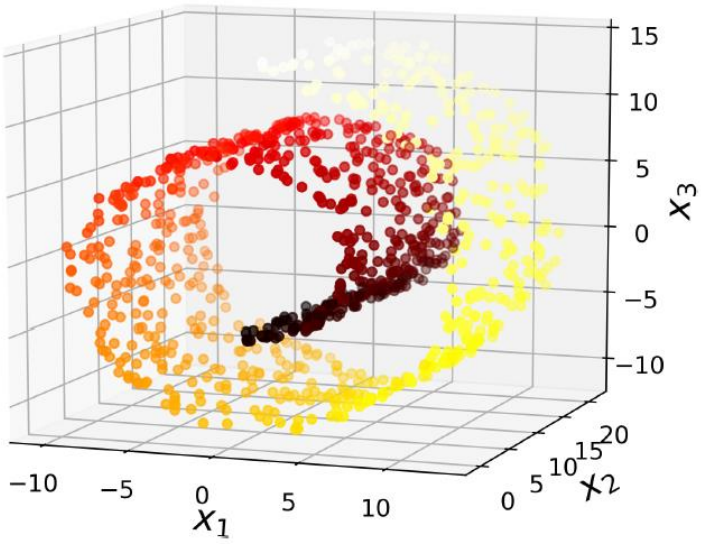


Olivetti Faces – Reconstructed

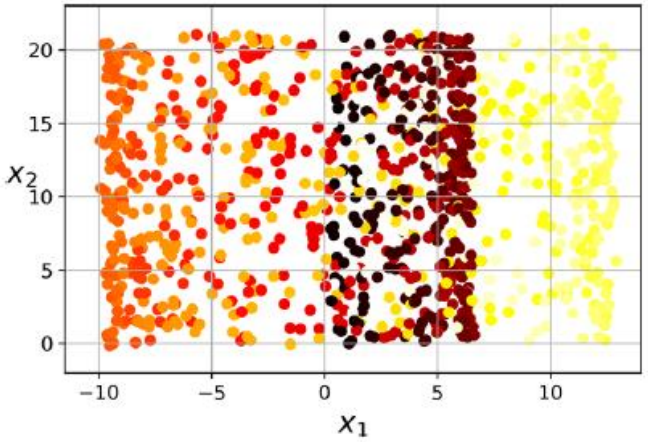
Manifold Learning

Fachhochschule Dortmund
University of Applied Sciences and Arts

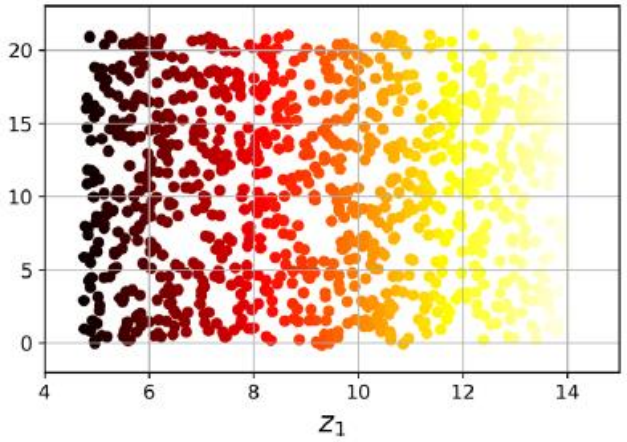
Original data



Projection

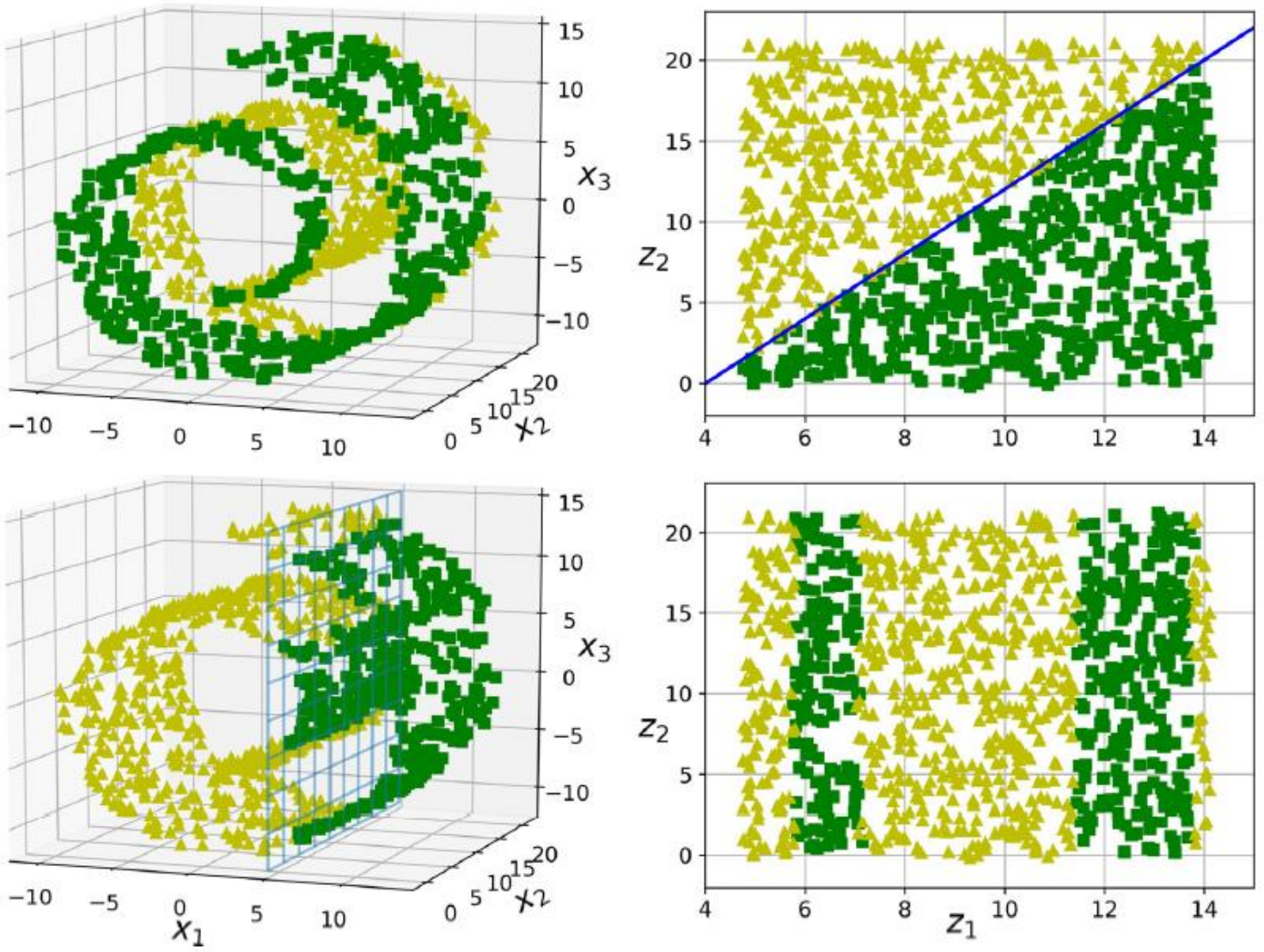


Unrolling manifold



Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

The decision boundary may not always be simpler with lower dimensions



Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Questions

Fachhochschule Dortmund
University of Applied Sciences and Arts

Questions for discussion

1. Can PCA be used to reduce the dimensionality of a highly nonlinear dataset?

2. Suppose you perform PCA on a 1,000-dimensional dataset, setting the explained variance ratio to 95%. How many dimensions will the resulting dataset have?

3. How can you evaluate the performance of a dimensionality reduction algorithm on your dataset?

4. Does it make any sense to chain two different dimensionality reduction algorithms?

Sources: Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Quantifying personality

- Lexical hypothesis: most important personality characteristics have become encoded in natural language.
- Allport and Odbert (1936): identified 4500 words describing personality traits.
- Group these words into (approximate) synonyms, by manual clustering
- Data collection: Ask persons whether these words describe them.

Spirit	Jolly, merry, witty, lively, peppy
Talkativeness	Talkative, articulate, verbose, gossipy
Sociability	Companionable, social, outgoing
Spontaneity	Impulsive, carefree, playful, zany
Boisterousness	Mischievous, rowdy, loud, prankish
Adventure	Brave, venturous, fearless, reckless
Energy	Active, assertive, dominant, energetic
Conceit	Boastful, conceited, egotistical
Vanity	Affected, vain, chic, dapper, jaunty
Indiscretion	Nosey, snoopy, indiscreet, meddlesome
Sensuality	Sexy, passionate, sensual, flirtatious

	shy	merry	tense	boastful	forgiving	quiet
Person 1	4	1	1	2	5	5
Person 2	1	4	4	5	2	1
Person 3	2	4	5	4	2	2
		⋮				

Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals

Big Five Personality Traits – OCEAN

Correlation of first five principle components with personality traits

Extraversion
−: quiet (-.83), reserved (-.80), shy (-.75), silent (-.71)
+: talkative (.85), assertive (.83), active (.82), energetic (.82)

Agreeableness
−: fault-finding (-.52), cold (-.48), unfriendly (-.45), quarrelsome (-.45)
+: sympathetic (.87), kind (.85), appreciative (.85), affectionate (.84)

Conscientousness
−: careless (-.58), disorderly (-.53), frivolous (-.50), irresponsible (-.49)
+: organized (.80), thorough (.80), efficient (.78), responsible (.73)

Neuroticism
−: stable (-.39), calm (-.35), contented (-.21)
+: tense (.73), anxious (.72), nervous (.72), moody (.71)

Openness
−: commonplace (-.74), narrow (-.73), simple (-.67), shallow (-.55)
+: imaginative (.76), intelligent (.72), original (.73), insightful (.68)

Source: UCSanDiegoX DSE220x – Machine Learning Fundamentals