LECTURER: NGHIA DUONG-TRUNG

DATA SCIENCE

TOPIC OUTLINE

Introduction to Data Science	1
Use Cases and Performance Evaluation	2
Data Preprocessing	3
Processing of Data	4
Selected Mathematical Techniques	5
Selected Artificial Intelligence Techniques	6

UNIT 5

SELECTED MATHEMATICAL TECHNIQUES

STUDY GOALS

On completion of this unit, you will have learned …

- how to apply principal component analysis to data.
- how to perform cluster analysis on a dataset.
- how to describe the linear regression model and compute its coefficients.
- how to describe the important features of time-series data.
- the popular models for forecasting future values in time-series data.
- the common approaches for dataset transformation.

EXPLAIN SIMPLY

1. Explain when to use the Principal Component Analysis (PCA) in practice.
2. Describe the concept of linear regression models and its coefficients using your own words.
3. Identify when the use of clustering techniques is helpful for business.

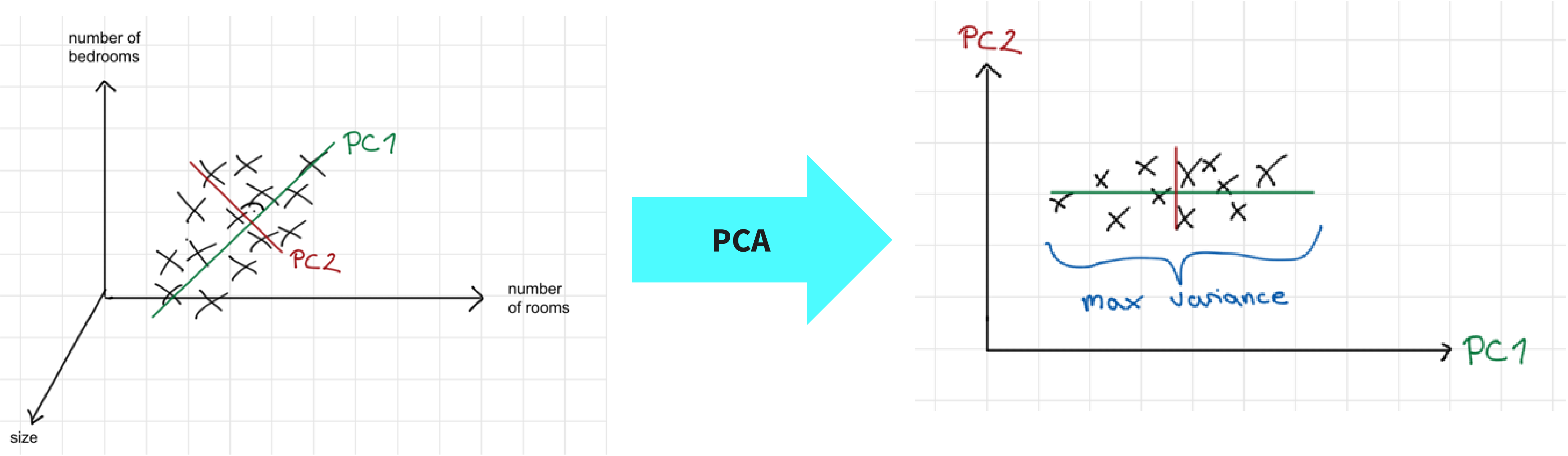
PREPARE THE CODES

- Download Session5_codes on Github
- Sub-folder 01

PRINCIPAL COMPONENT ANALYSIS

Transform potentially correlated variables into fewer uncorrelated variables (PCs).

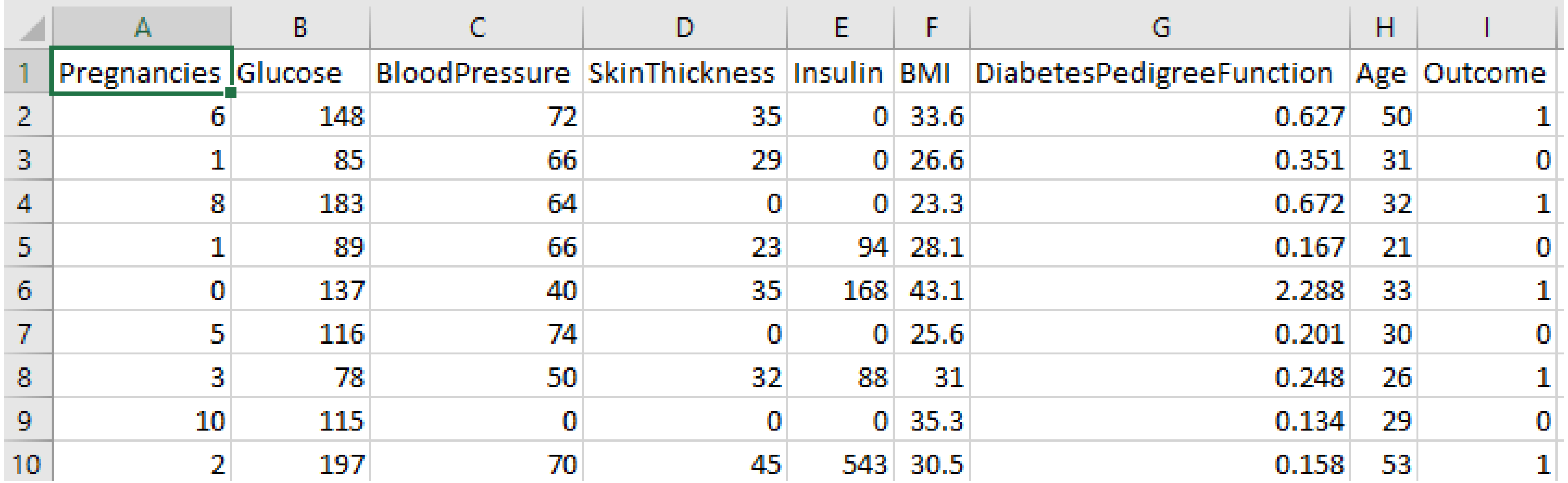
→ dimensionality reduction of the dataset while loosing only a small amount of information.



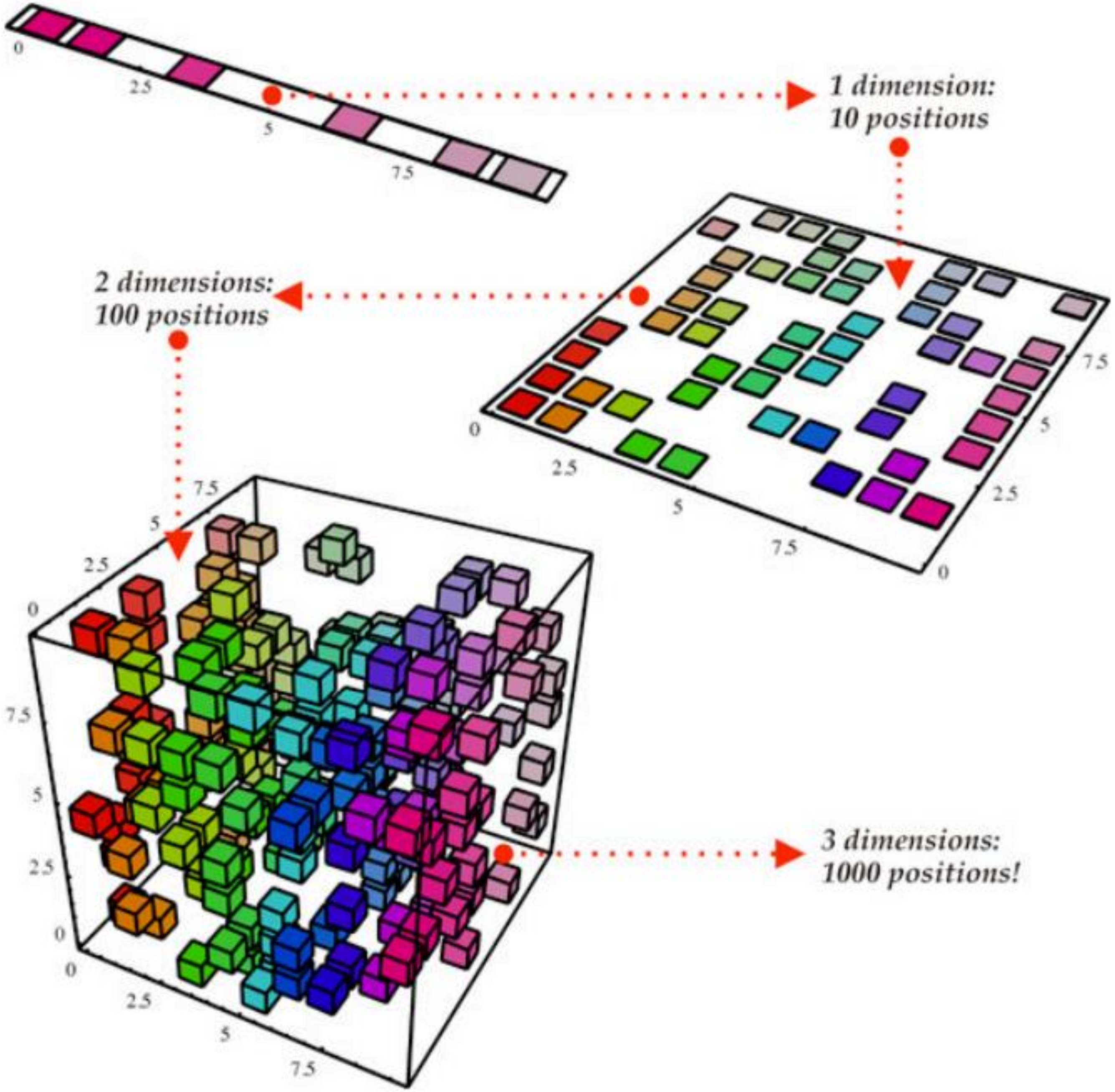
Source of the graphics: Drawings by Antonia Schulze, 2021.

KEY CONCEPTS

- Dimension
- Dimensionality reduction
- Feature scaling



	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1



1 dimension:
10 positions

2 dimensions:
100 positions

3 dimensions:
1000 positions!

Further references:

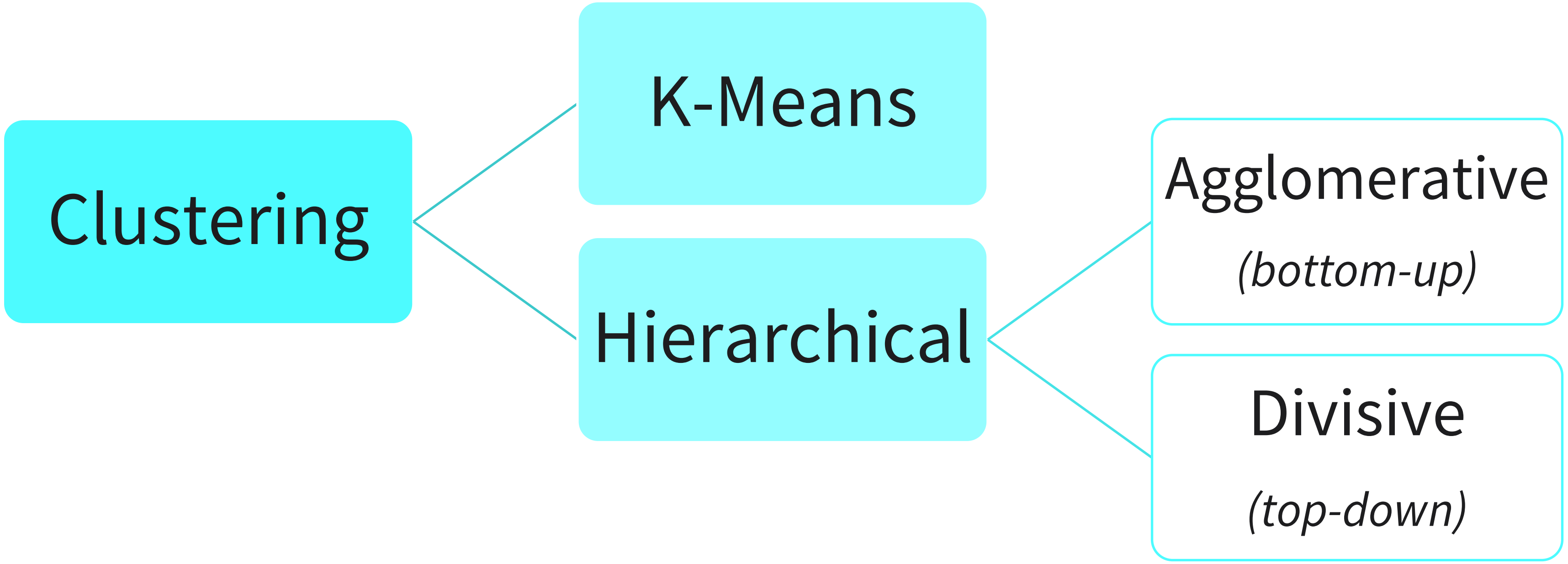
- https://www.youtube.com/watch?v=BsJJXQ10ayM
- https://setosa.io/ev/principal-component-analysis/
- https://builtin.com/data-science/step-step-explanation-principal-component-analysis

PCA codes

- Sub-folder: Session5_codes\02

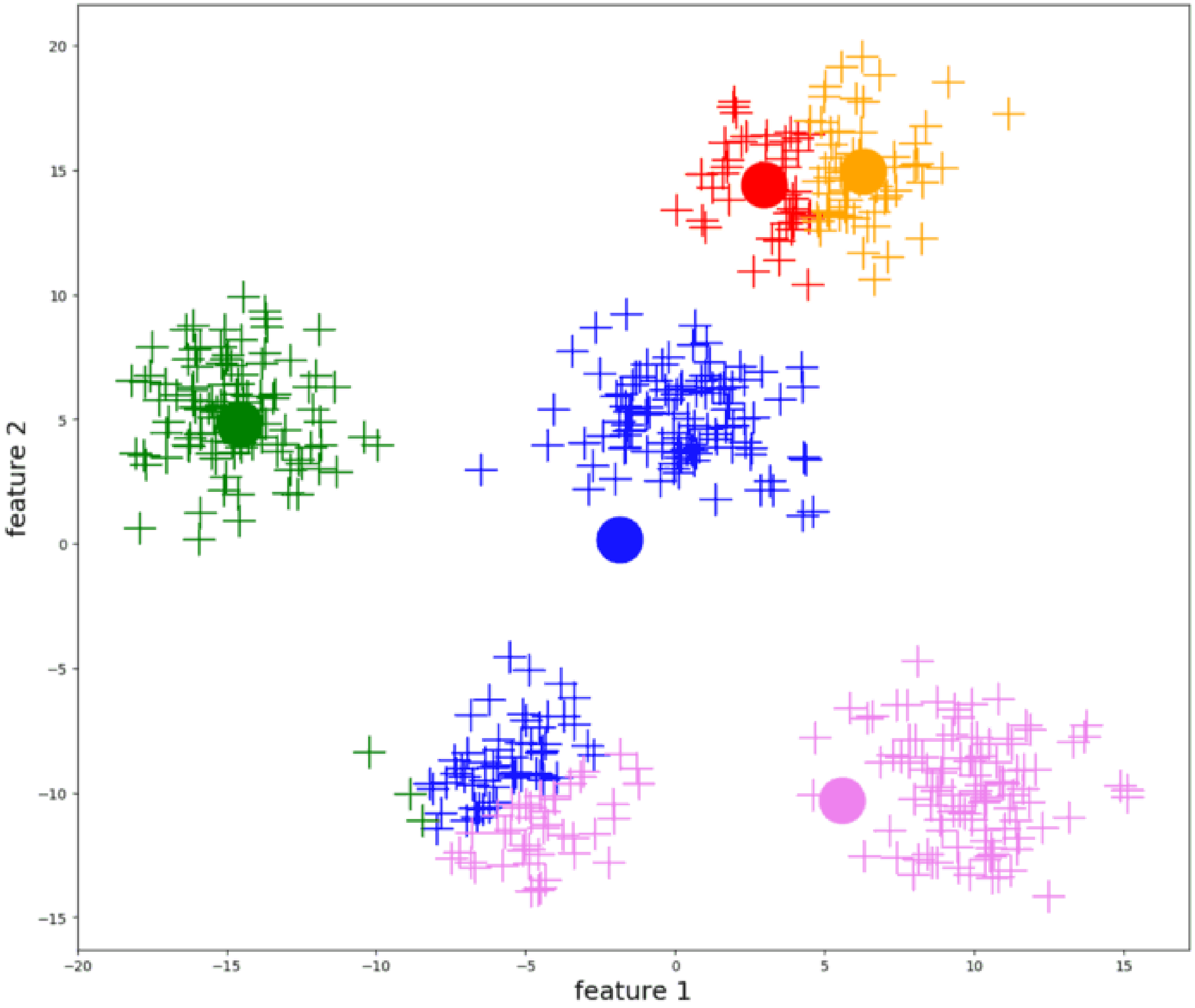
Grouping objects into unlabeled, meaningful clusters

- maximize similarity within a cluster (distance to centroids)
- maximize dissimilarity between clusters



K-MEANS CLUSTERING

- select # of clusters (k)
- choose random centroids
- assign data points to clusters based on minimal distance to centroid
- calculate new centroid
- start over until no changes made to centroids



Source of the animation: Brilenkov, 2021.

Further references:

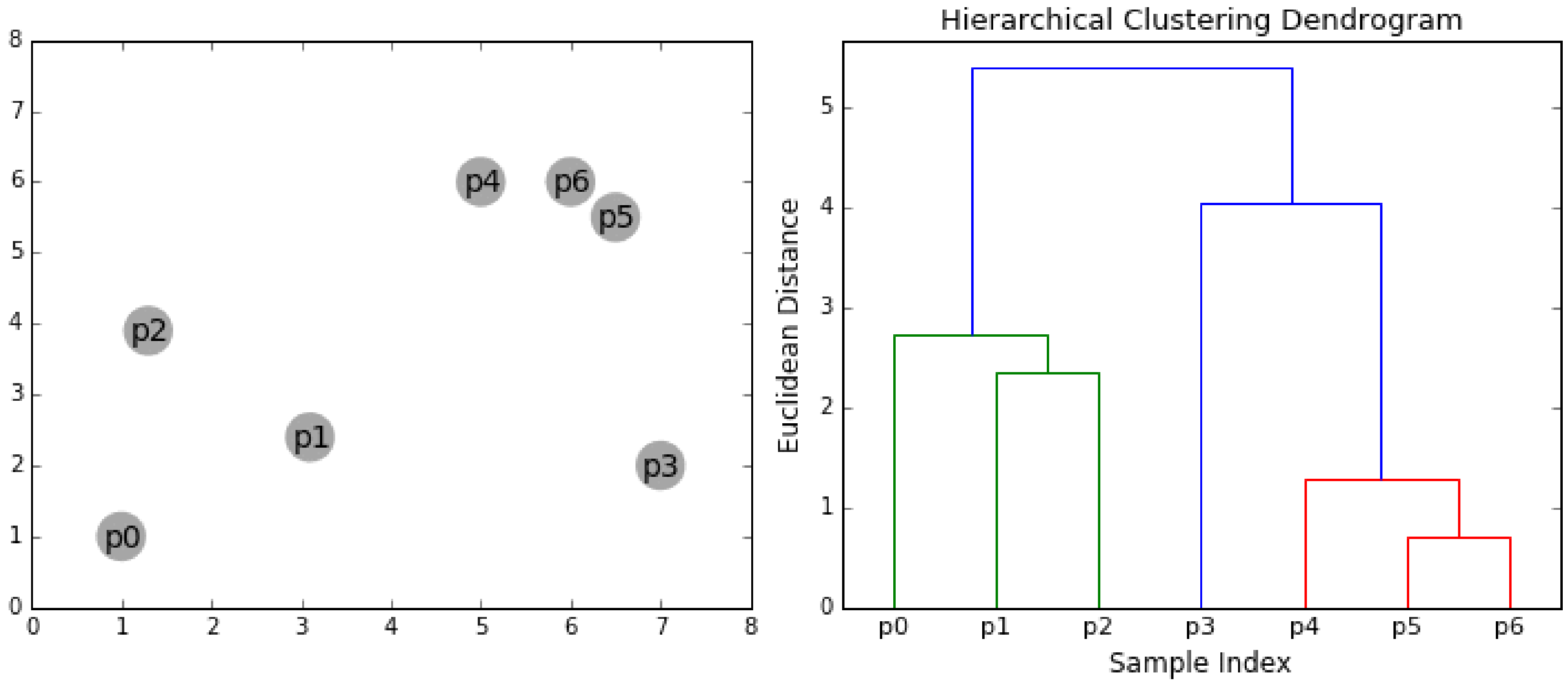
- https://www.youtube.com/watch?v=SeswFFdH03U
- https://www.analyticsvidhya.com/blog/2021/02/simple-explanation-to-understand-k-means-clustering/
- https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm

Codes:

- Sub-folder: Session5_codes\03

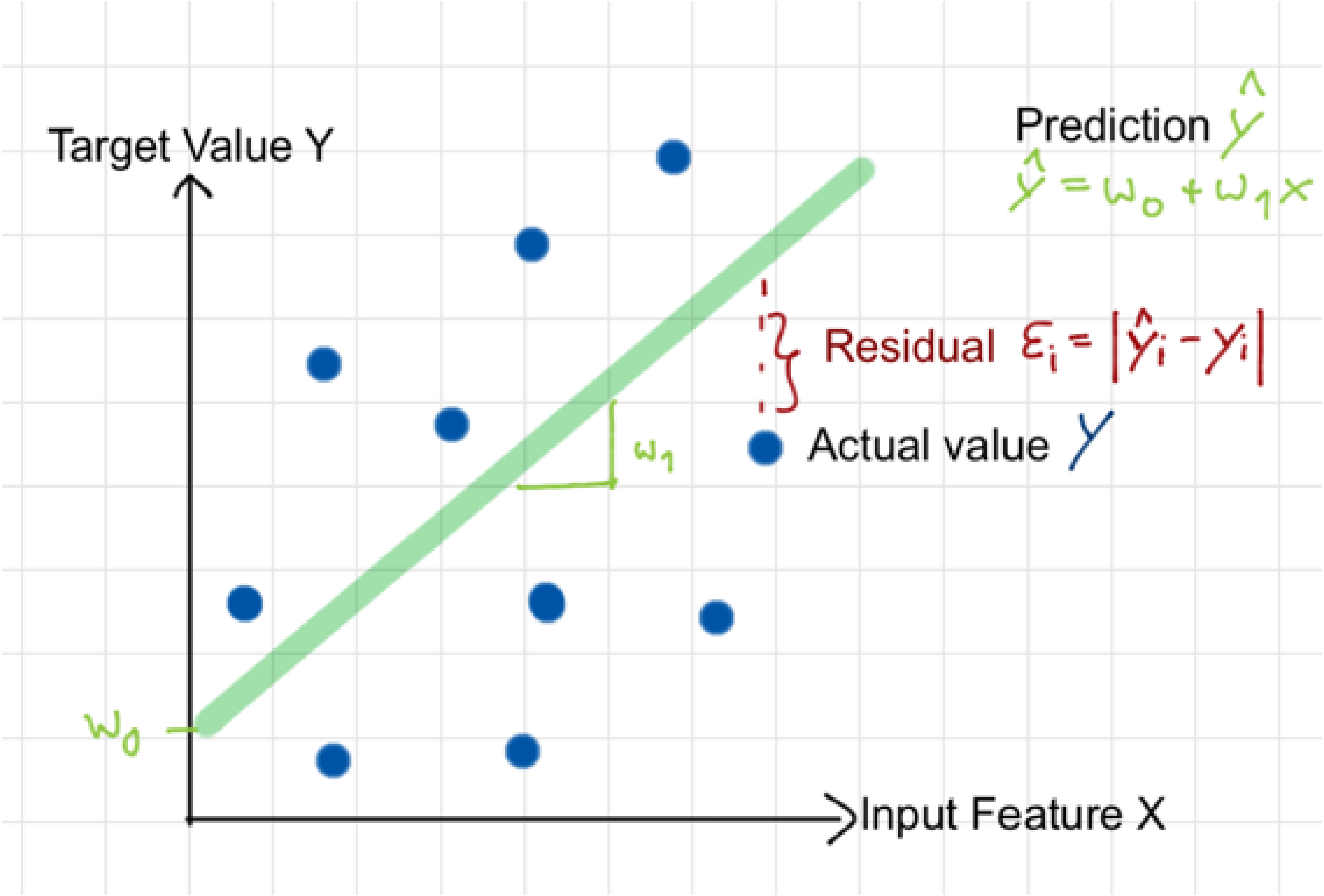
HIERARCHICAL CLUSTERING

- assign each record to a unique cluster
- merge clusters with minimum distance
- repeat until only one cluster left



Source of the animation: Sheenan, 2017.

- predict value of dependent (*target*) variable given independent (*predictor*) variables
- assumption: linear relationship between variables



Source of the graphic: Drawing by Antonia Schulze, 2021.

Further references:

- https://www.youtube.com/watch?v=Kxw1AjAN1GA
- https://towardsdatascience.com/linear-regression-made-easy-702e5dc01f03
- https://www.analyticsvidhya.com/blog/2020/10/linear-regression-for-absolute-beginners-with-implementation-in-python/

Codes:

- Sub-folder Session5_codes\04

MODEL PERFORMANCE

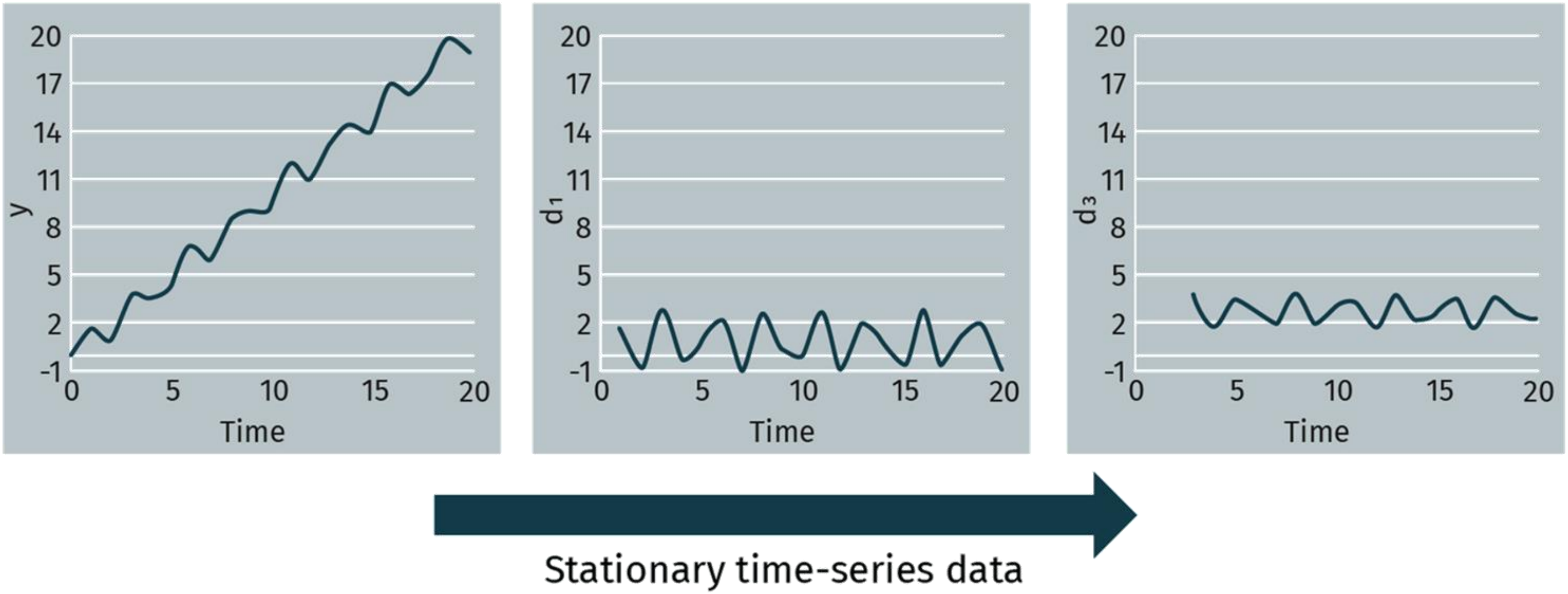
- Mean Absolute Error (MAE) is the mean of the absolute value of the errors.

- Mean Squared Error (MSE) is the mean of the squared errors.

- Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors.
  - RMSE is the most popular because it tell us RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
  - RMSE is better than MSE in most cases because it accounts for large errors.

TIME-SERIES FORECASTING – VOCABULARY

- stationary time series = constant mean and standard deviation over time
- Lag(n) = backshift of a time-series by n time steps



Source of the graphic: IU International University, Course Book DLMBDSA01.

TIME-SERIES FORECASTING – VOCABULARY

- stationary time series = constant mean and standard deviation over time
- Lag(n) = backshift of a time-series by n time steps
- Autocorrelation (ACF) = correlation between variable and previous lags
- Partial Autocorrelation (PACF) = autocorrelation between y_t and y_{t-k} that is not accounted for by the autocorrelations from the 1^{st} to the $(k-1)^{st}$ lags.

TIME-SERIES FORECASTING

Autoregressive Model (AR)

models future values as a function of recent past **sequential values**

Moving Average Model (MA)

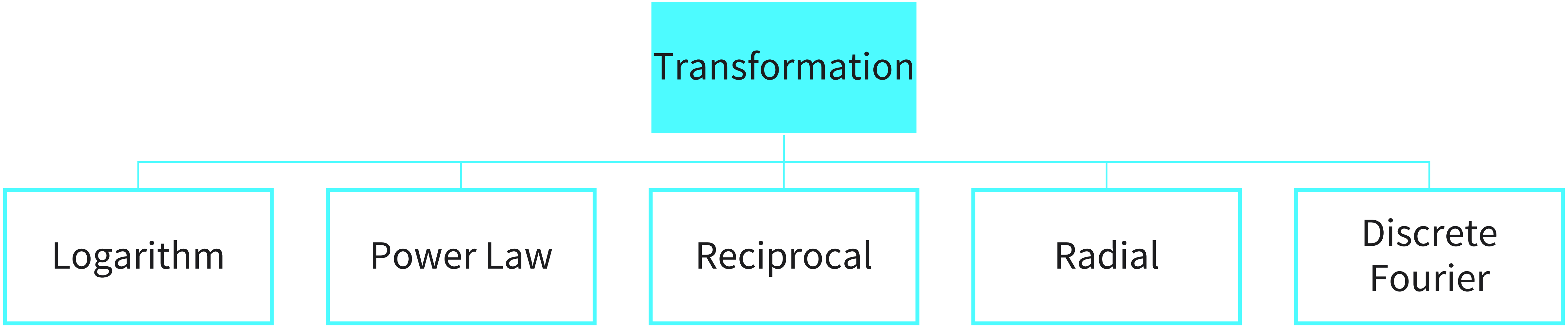
models future values as a function of recent past **sequential error terms**

Autoregressive Integrated Moving Average Model (ARIMA)

combination of **AR** & **MA** models with an **I**ntegration of differencing the time-series until stationarity reached

TRANSFORMATION APPROACHES

Process of transforming variables to improve its interpretability

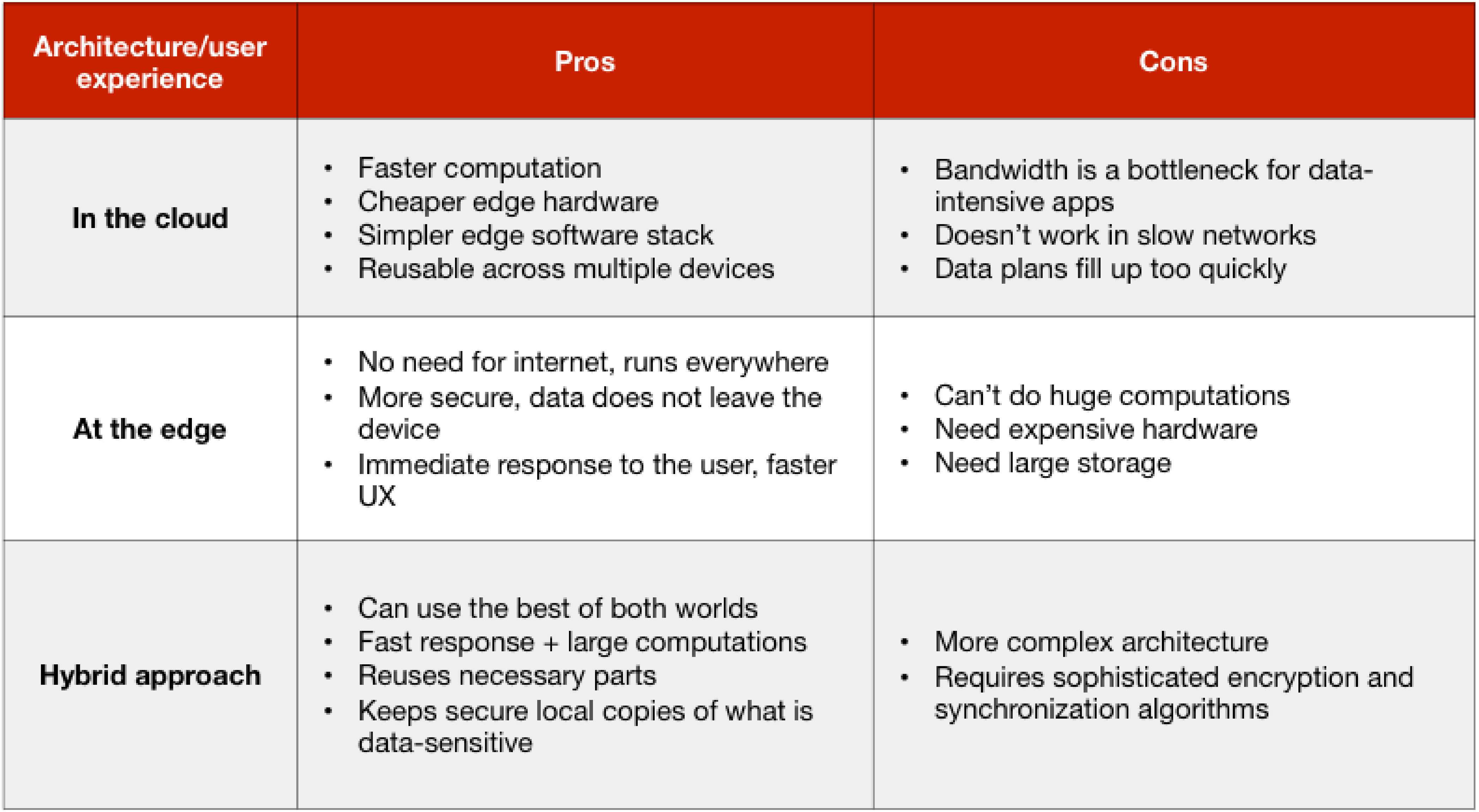


Fourier Transform:

- https://www.youtube.com/watch?v=spUNpyF58BY
- https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/

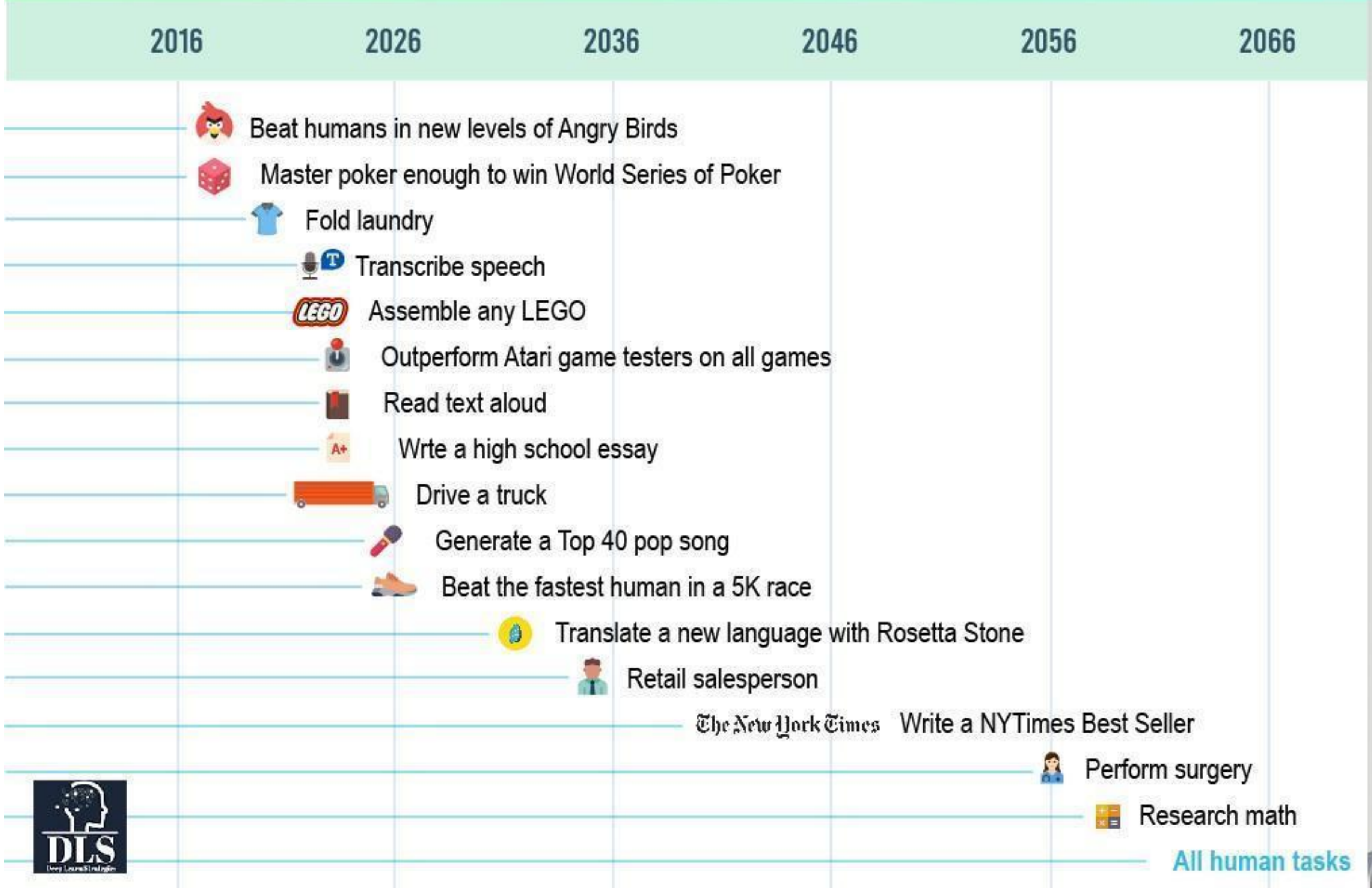
Further references:

- https://www.youtube.com/watch?v=TR6vn4lZ3Mo
- https://www.kaggle.com/code/ryanholbrook/linear-regression-with-time-series



Architecture/user experience	Pros	Cons
In the cloud	• Faster computation • Cheaper edge hardware • Simpler edge software stack • Reusable across multiple devices	• Bandwidth is a bottleneck for data-intensive apps • Doesn’t work in slow networks • Data plans fill up too quickly
At the edge	• No need for internet, runs everywhere • More secure, data does not leave the device • Immediate response to the user, faster UX	• Can’t do huge computations • Need expensive hardware • Need large storage
Hybrid approach	• Can use the best of both worlds • Fast response + large computations • Reuses necessary parts • Keeps secure local copies of what is data-sensitive	• More complex architecture • Requires sophisticated encryption and synchronization algorithms

WHEN JOB\TASKS WILL BE TAKEN OVER BY MACHINES



Beat humans in new levels of Angry Birds

Master poker enough to win World Series of Poker

Fold laundry

Transcribe speech

Assemble any LEGO

Outperform Atari game testers on all games

Read text aloud

Wrte a high school essay

Drive a truck

Generate a Top 40 pop song

Beat the fastest human in a 5K race

Translate a new language with Rosetta Stone

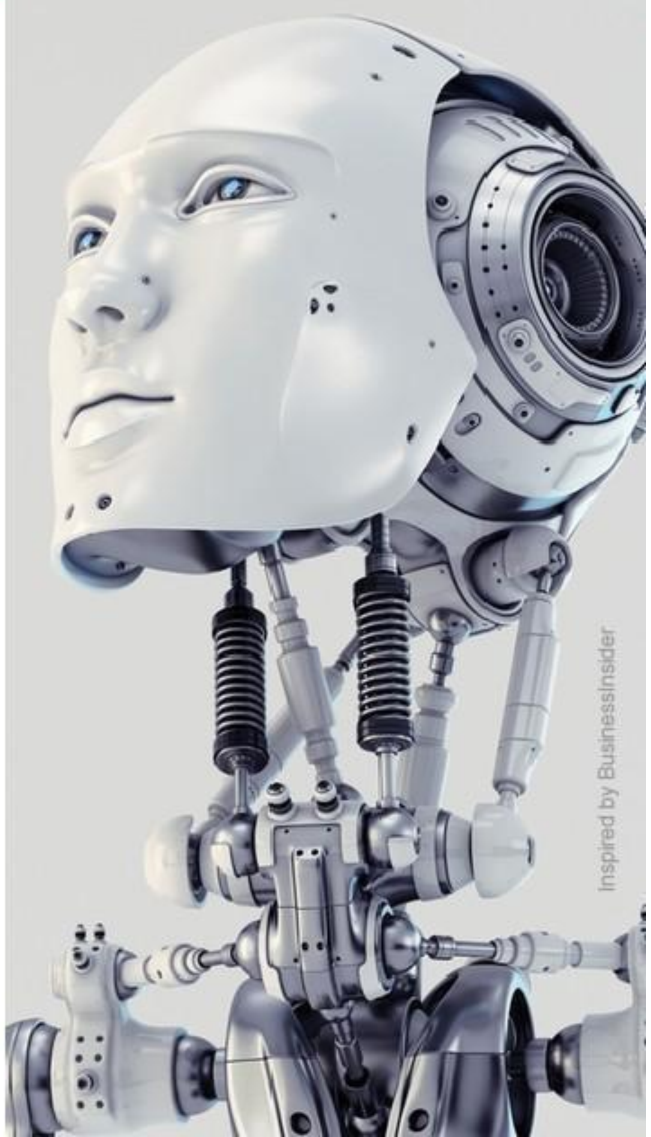
Retail salesperson

Write a NYTimes Best Seller

Perform surgery

Research math

All human tasks


Inspired by Businessinsider

SOURSE: Grace, Salvatier, et al @deeplearn007

REVIEW STUDY GOALSTUDY GOALS

You have learned …

- how to apply principal component analysis to data.
- how to perform cluster analysis on a dataset.
- how to describe the linear regression model and compute its coefficients.
- how to describe the important features of time-series data.
- the popular models for forecasting future values in time-series data.
- the common approaches for dataset transformation.

SESSION 5

TRANSFER TASK

TRANSFER TASK

You are facing a big dataset and want to apply your previous knowledge of PCA to get a smaller, but still informative dataset. Your colleague, however, has some questions that you will need to answer. Prepare a role play.

Inspiration:

- Discuss: More data → more information?
- Analyze: advantages and disadvantages of using PCA

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LEARNING CONTROL QUESTIONS

1. The transformation approach, which transfers data variables to their frequency domain, is called the…

   a) radial transformation.
   b) reciprocal transformation.
   c) Fourier transformation.
   d) logarithm transformation.

LEARNING CONTROL QUESTIONS

2. The auto-regressive model assumes a…

a) linear function between the future output and past outputs.

b) repeated pattern in the time-series data.

c) constant output over time.

d) sinusoidal wave that relates the outputs and the inputs.

LEARNING CONTROL QUESTIONS

3. The operation of sorting data variables according to their level of changeability along data records is part of…

   a) regression modelling.
   b) classification modelling.
   c) clustering analysis.
   d) principal component analysis.

LIST OF SOURCES

Brilenkov, R. (2021). *Understanding K-Means Clustering: Hands-on Visual Approach* [blog post]. Retrieved from: https://ai.plainenglish.io/understanding-k-means-clustering-hands-on-visual-approach-c2dc46f0ed18

Sheenan, D. (2017). *Clustering with Scikit with GIFs.* [blog post]. Retrieved from: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/