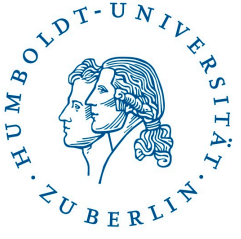


Introduction to Data Science & Visualizations in Learning Analytics

Jakub Kuzilek

Lecture schedule

2 parts (approx. 40 minutes each):

1. Introduction to Data Science in LA
2. OULAD dataset & template solution

In the middle we will have 5 minutes break.

At the end:

- Voting for the additional non-mandatory lecture on data analysis/manipulation using OULAD in R
- Group formation

Data Science in Education

Data Science in Learning Analytics

Data Science is discipline that allows you to turn raw data into understanding, insight, and knowledge.

Grolemund, G., & Wickham, H. (2017). R for Data Science. O'Reilly Media.

DS in education is challenging

Making sense from data about teaching, learning and educational systems.

Data analysis + content knowledge

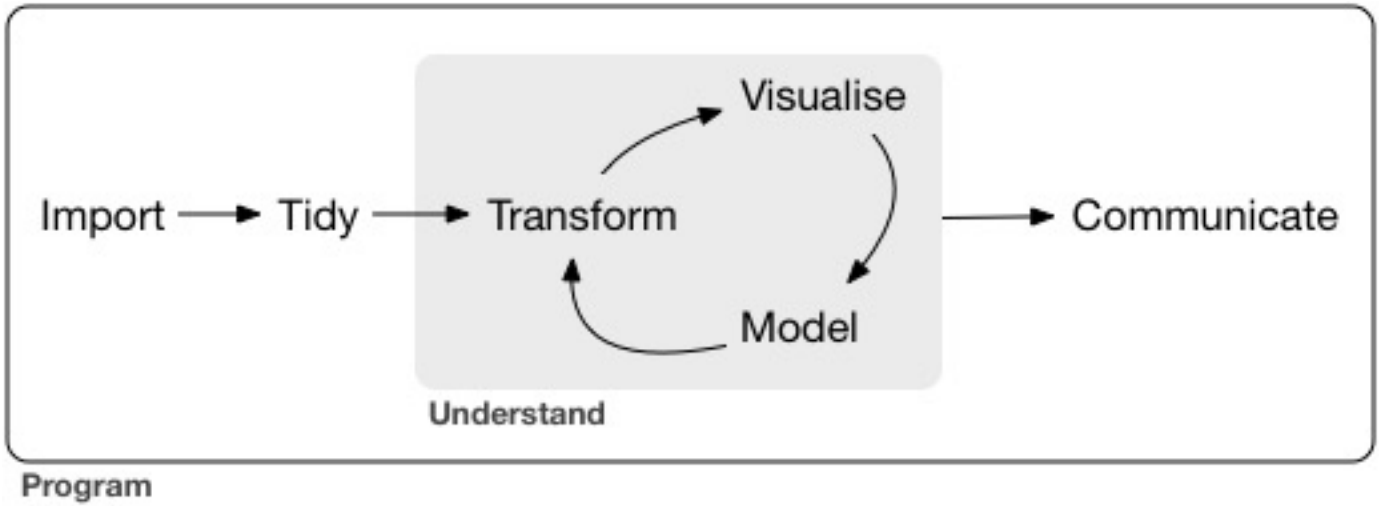
Tasks

- Building systems that get data to the right people
- Measuring impact on the student experience
- Searching for patterns in student data
- Using statistical models in education
- Studying the effects of educational support
- Advancing scientific knowledge about learning and learners

Common methods

- Prediction
- Clustering
- Behaviour modelling
- Recommendation
- Student grouping
- Social network analysis
- Text mining

Data Science process



Tidy + Transform = Wrangle

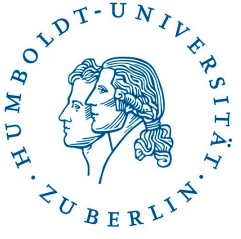
https://r4ds.had.co.nz/

Import

Data sources in education:

- Student record data
- Staff data
- Admissions & applications data
- Financial data
- Alumni data
- Course data
- Estates and facilities data
- Virtual Learning Environments
- Assessment data
- Forum data

Import



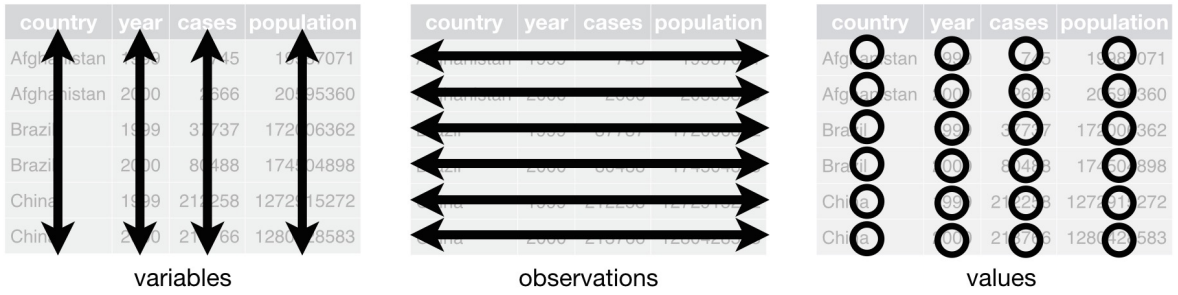
Data sources types:

- Files
  - CSV, XML, JSON
- Databases
  - SQL, NoSQL (key-value, graph-based, document-based,...)
- API (Social media,…)

Tidy

Tidy dataset:

- Each column represents one variable
- Each row represents one observation
- Each cell represent one value

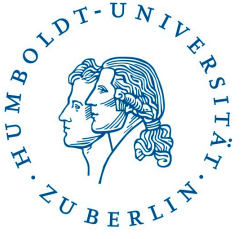


Problems:

- Multiple data sources with different unique identifier
- Values in one column represents multiple variables
- One observation spreads in multiple rows

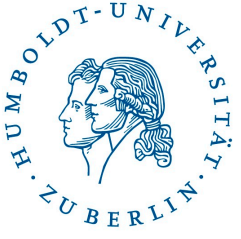
https://r4ds.had.co.nz/tidy-data.html

Transform



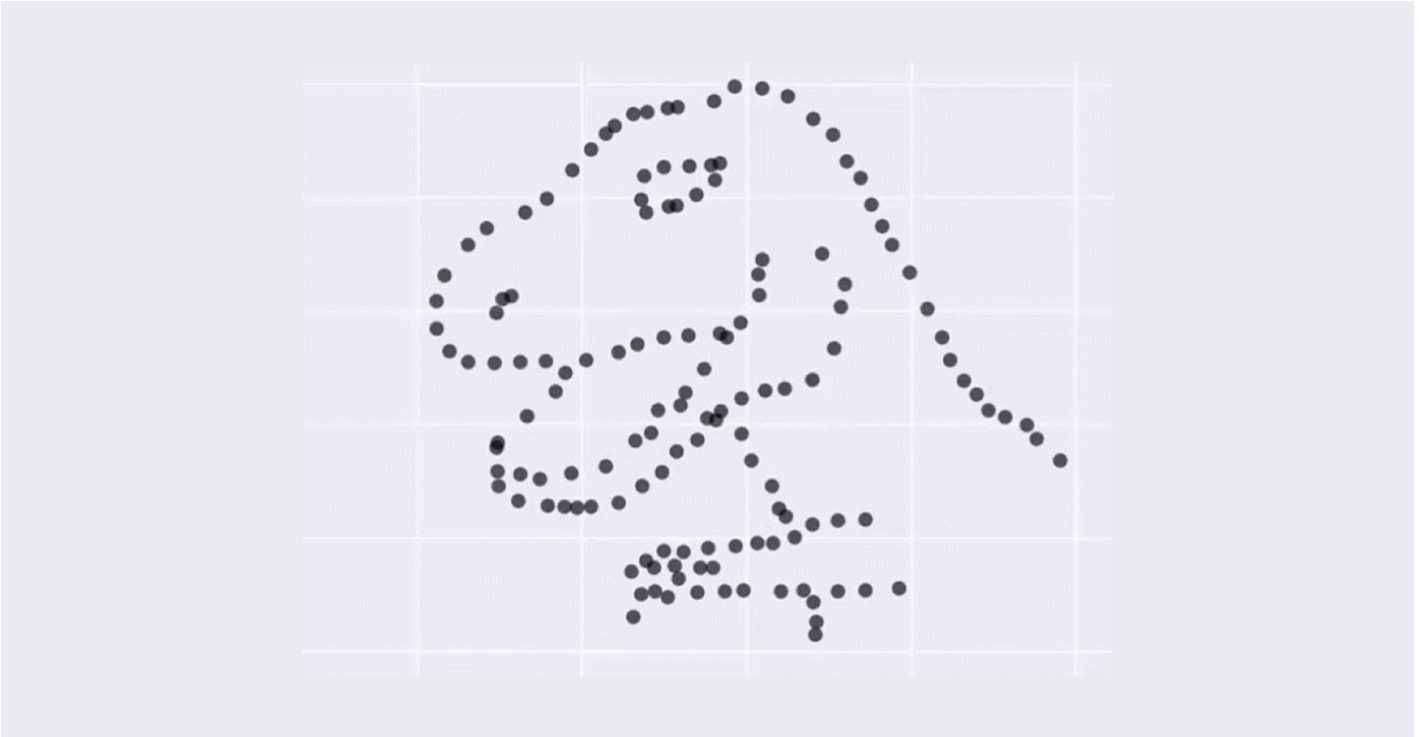
- Handle missing data
- Inconsistent data types
- Outliers
- Encoding
- Filtering the data
- Aggregation of the data
- Transforming values
- Handling texts and dates

Visualize

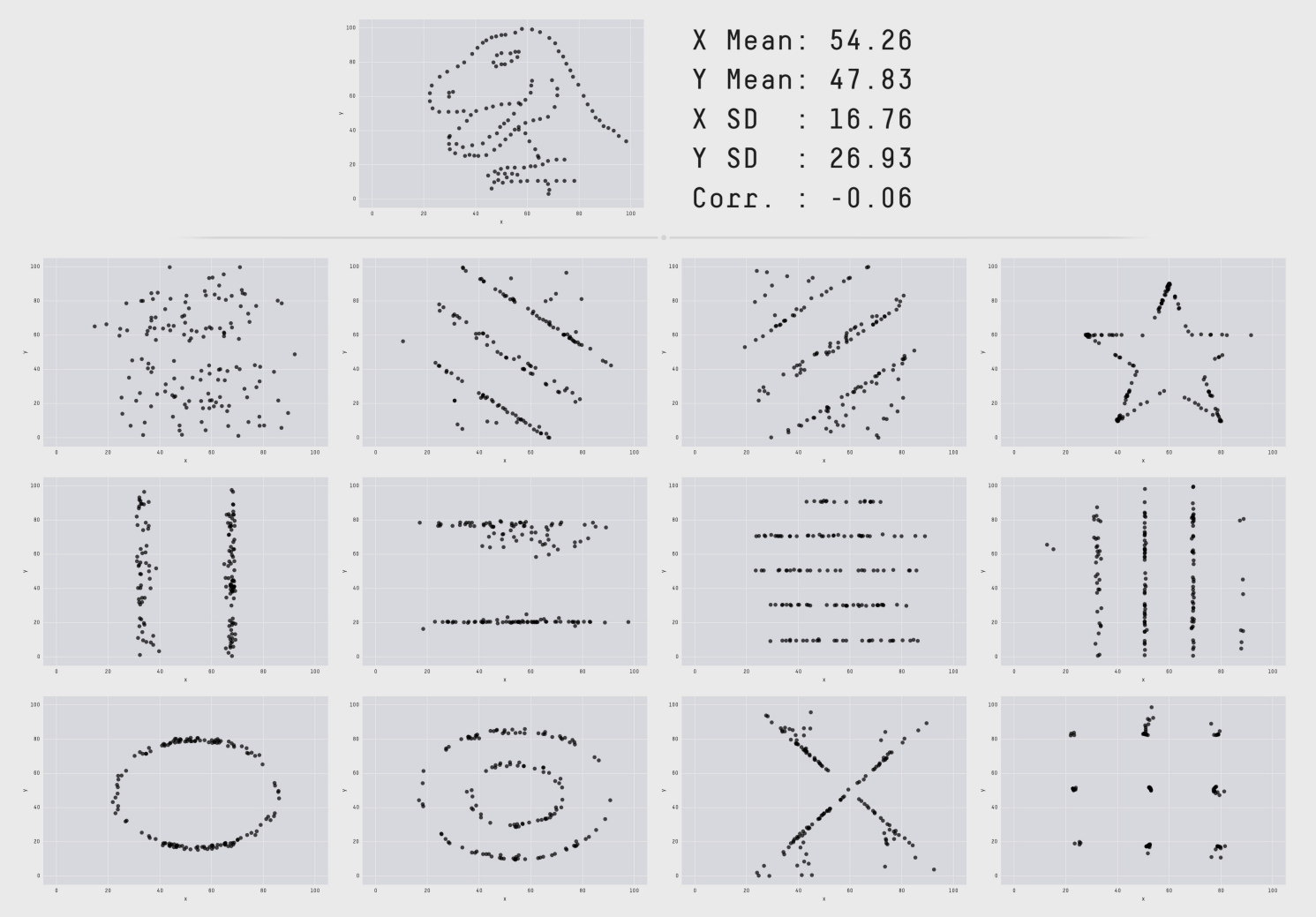


- Visualization is useful tool for providing the information to stakeholder but also during wrangling the data
- Helps to understand issues in the data
- Uncovers outliers
- Helps to identify relationships between variables

Datasaurus



Datasaurus



https://www.autodesk.com/research/publications/same-stats-different-graphs

Do not trust your data blindly

- **Always check data visually.**
- **Statistics can be misleading.**

Model

- Tidy data can be used for creating of model
- For that Machine Learning methods can be used

Machine Learning is:

"Field of study that gives computers the ability to learn without being explicitly programmed"

~ Arthur Samuel, 1959

- Machine Learning is subfield of Computer Science
- Objective: *Generalize from experience*

Model

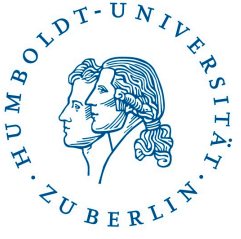
ML task categories based on “feedback” available to learning system:

- Supervised learning
  - We know the right answers
  - Supervised learning algorithm is inferring decision function from labelled training data. The algorithm needs to generalize from training data to unseen data "reasonably".
- Unsupervised learning
  - We do not know right answers
  - Unsupervised learning algorithm is inferring function, which describes hidden structure of unlabelled data. We cannot estimate error of algorithm.
- Reinforcement Learning
  - Machine interacts with dynamic environment in which it needs to achieve certain goal without teacher telling it if it is close to the goal or not.

Model

ML categories based on ”outputs” produced:

- Classification
- Regression
- Clustering
- Density estimation
- Dimensionality reduction



Visualisations

Gestalt Laws of Perceptual Organization

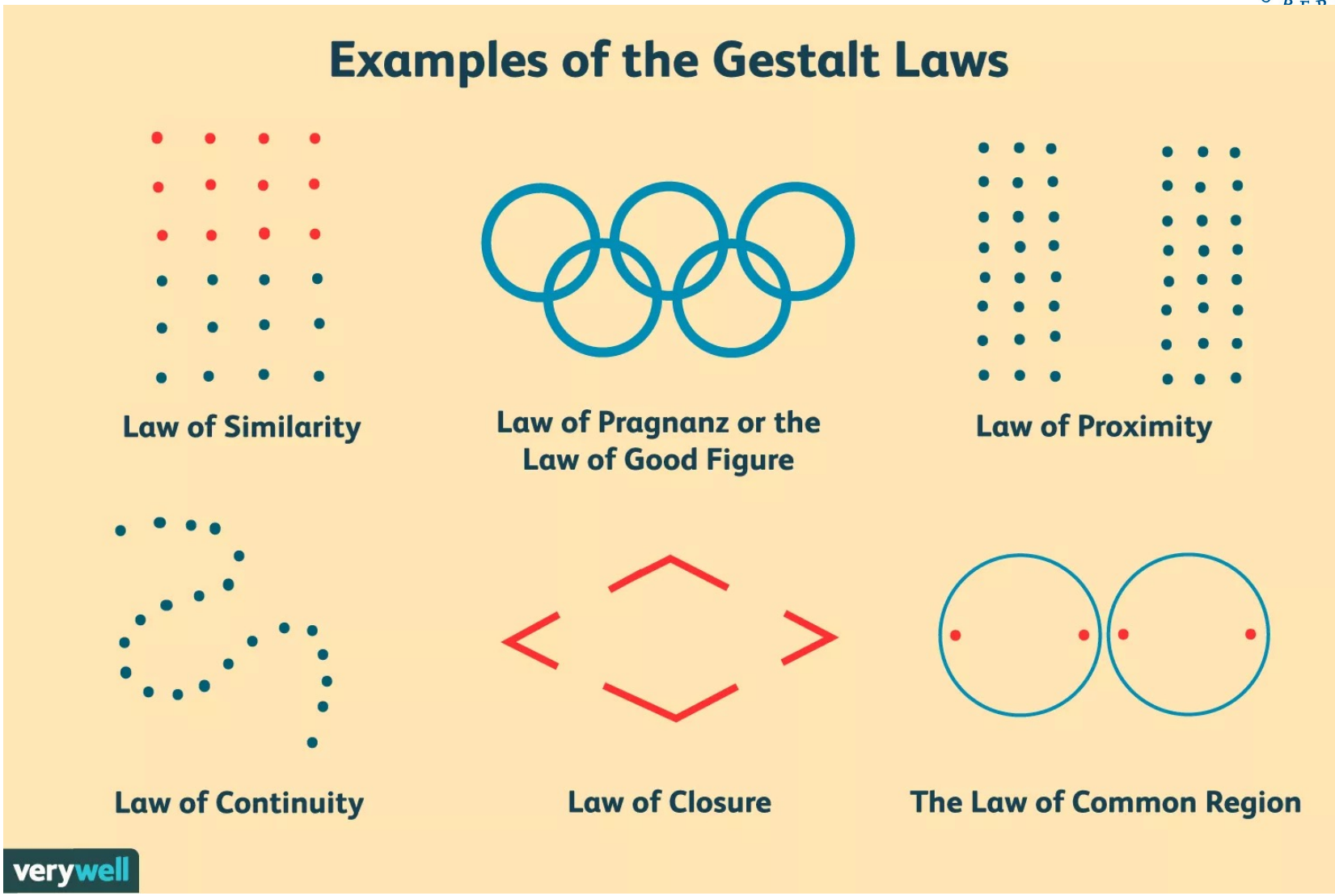
- Gestalt psychology – how human mind organize and interpret visual data
- Humans have advanced perceptual abilities – good at recognizing patterns

- Wolfgang Kohler & Kurt Koffka developed rules how human group small objects to form larger ones (perceptual organization) ~ Gestalt laws

- A set of principles for understanding some of the ways in which perception works.
- Sometimes lead to incorrect perceptions of the world
- Actually heuristics or shortcuts
- Heuristics are usually designed for speed not for accuracy

https://socialsci.libretexts.org/Bookshelves/Psychology/Book%3A_General_Psychology_for_Honors_Students_(Votaw)/01%3A_History_of_Psychology/1.05%3A__Wertheimer_Koffka_Kohler_Aad_Gestalt_Psychology

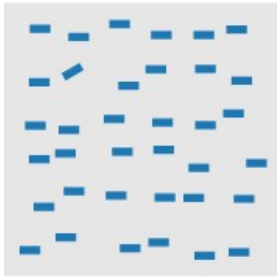
Gestalt Laws of Perceptual Organization

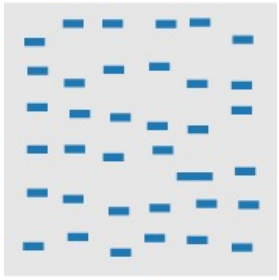


https://www.verywellmind.com/gestalt-laws-of-perceptual-organization-2795835

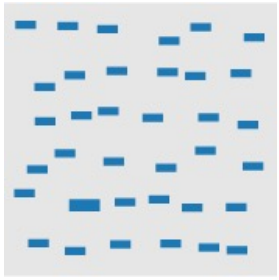
Pre-attentive characteristics

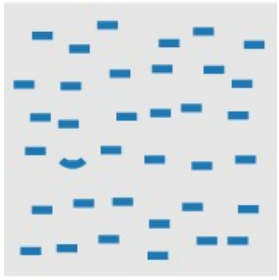
Set of visual properties are detected rapidly (< 250 ms) in multi-element display and accurately by low-level visual system

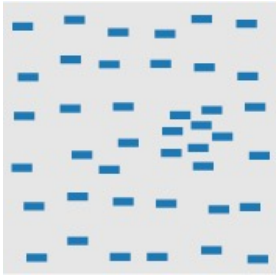

line orientation

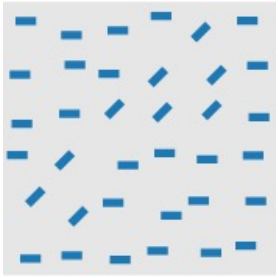

length/width

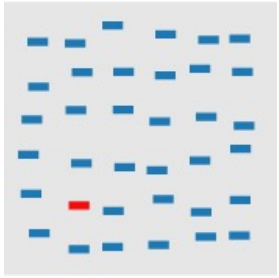

closure

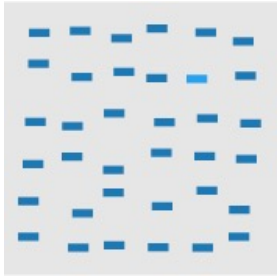

size

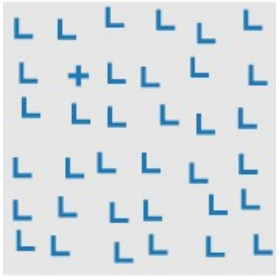

curvature

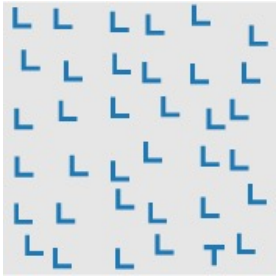

density/contrast

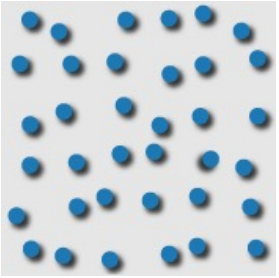

number


color


intensity


intersection


terminators


lighting direction

https://www.csc2.ncsu.edu/faculty/healey/PP/

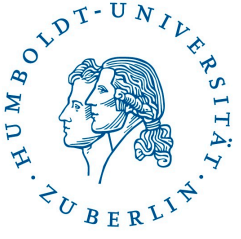
Dashboards

Visualise + Communicate -> Dashboard

- Deliver information to stakeholder in interactive way
- Automation of the analysis
- Includes possibility for user to adjust some parameters

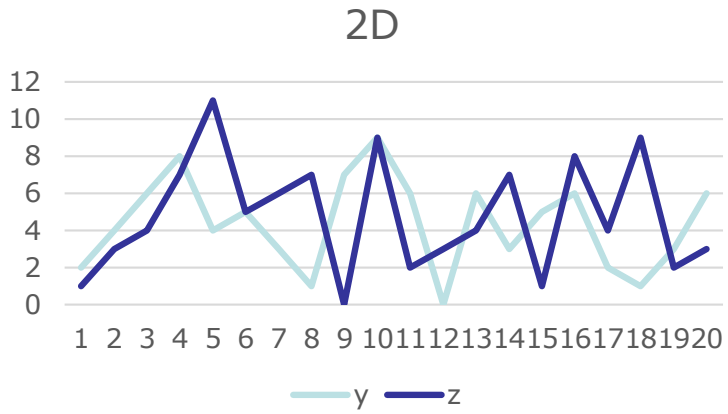
- Challenges:
  - Scalability
  - Data quality
  - User interface
  - Evaluation

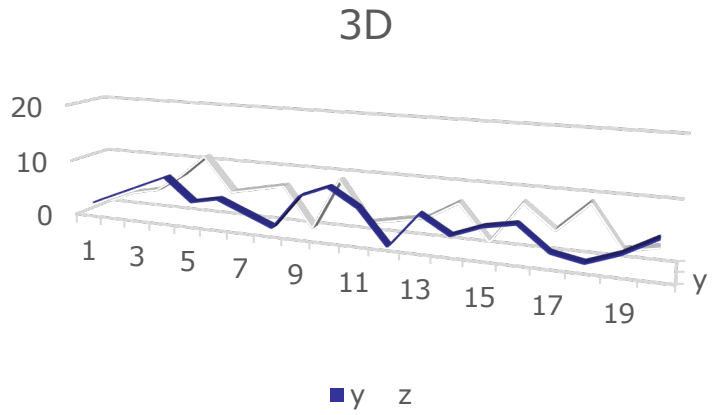
- Issues:
  - Too much colour
  - Too much details
  - Useless decorations
  - Poor visualisations



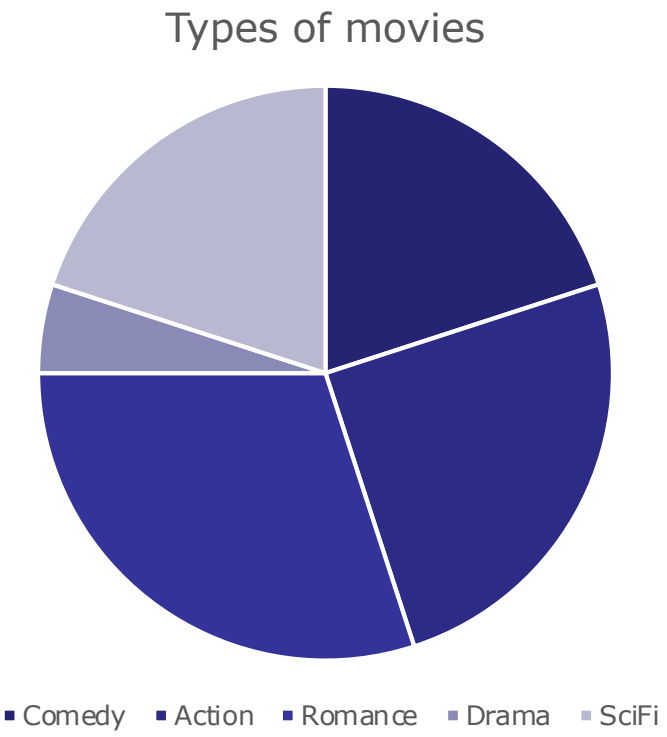
Tips for creating useful visualizations

Tips: No 3D graphs



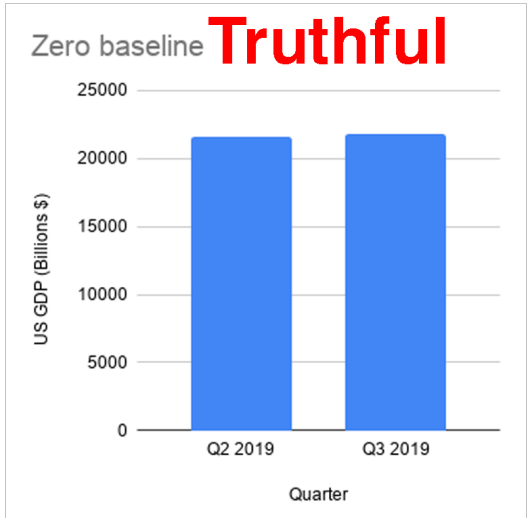


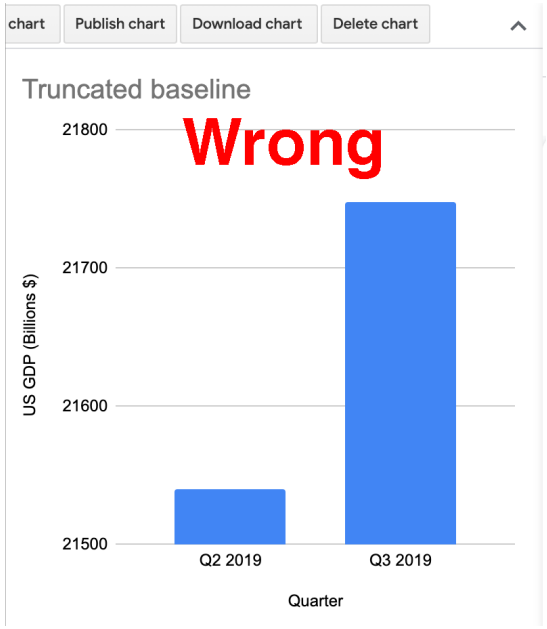
Tips: No pie charts



Pre-attentive characteristics does not help with showing exact quantitative diferences

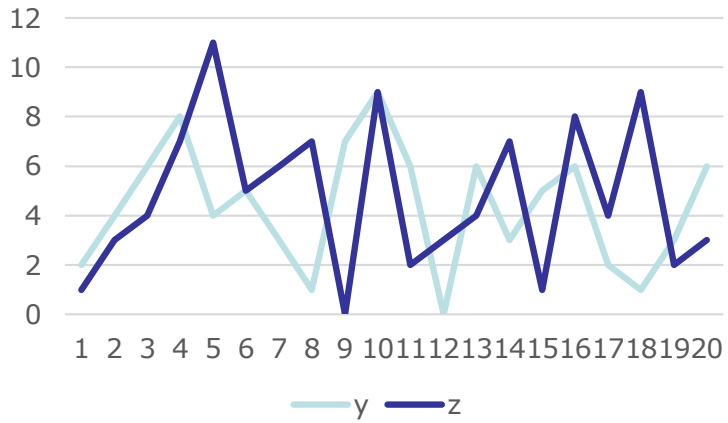
Tips: Do not lie

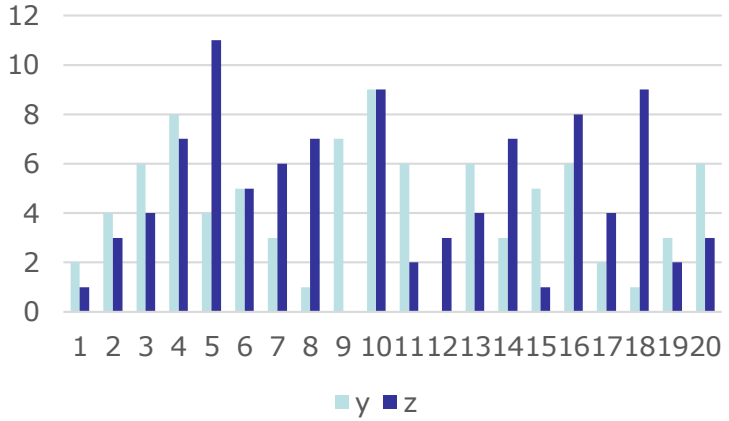




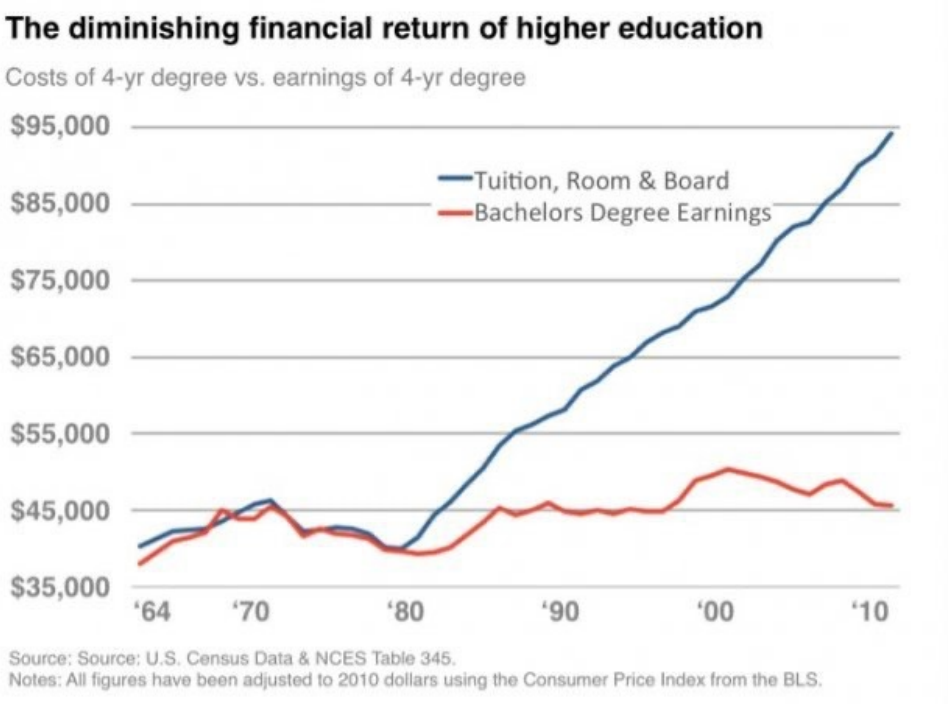
https://handsondataviz.org/how-to-lie-with-charts.html#fig:gdp-baseline-merged

Tips: Use common sense





Tips: Don’t misled the users



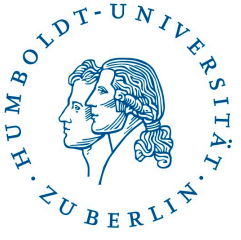
How to start

1. Know your data and audience
2. Formulate questions about you data
3. Apply visual mapping

“Simplicity is the ultimate sophistication”
~ Leonardo da Vinci

References

- Wagemans J, Elder JH, Kubovy M, et al. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychol Bull*. 2012;138(6):1172–1217. doi:10.1037/a0029333
- Vezzani, S, Marino, BF, Giora, E. An early history of the Gestalt factors of organization. *Perception*. 2012;41(2):148-67. doi:10.1068/p7122
- Dresp-Langley B. Principles of perceptual grouping: Implications for image-guided surgery. *Front Psychol*. 2015;6:1565. doi:10.3389/fpsyg.2015.01565
- Ali N, Peebles D. The effect of Gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Hum Factors*. 2013;55(1):183-203. doi:10.1177/0018720812452592



5 minutes break

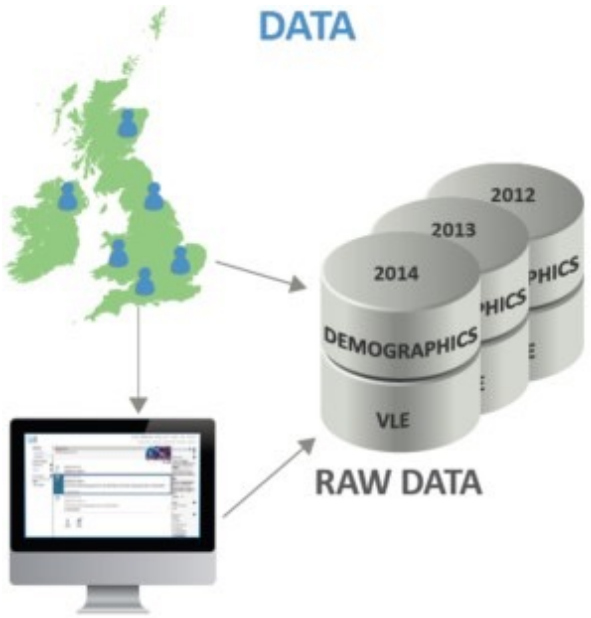
Introduction to OULAD data

https://www.nature.com/articles/sdata2017171

About Open University

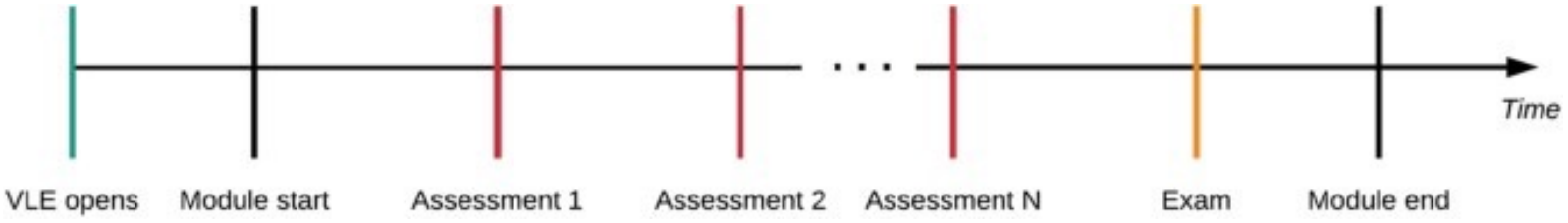
- One of the largest distance based learning universities worldwide
- Approx. 180,000 students
- Teaching delivered remotely via Moodle-like system
- During the degree students participate in several courses (modules)
- 1 module ~ 60 credits
- Wide variety of units/departments exists to support students

The Open University



About courses

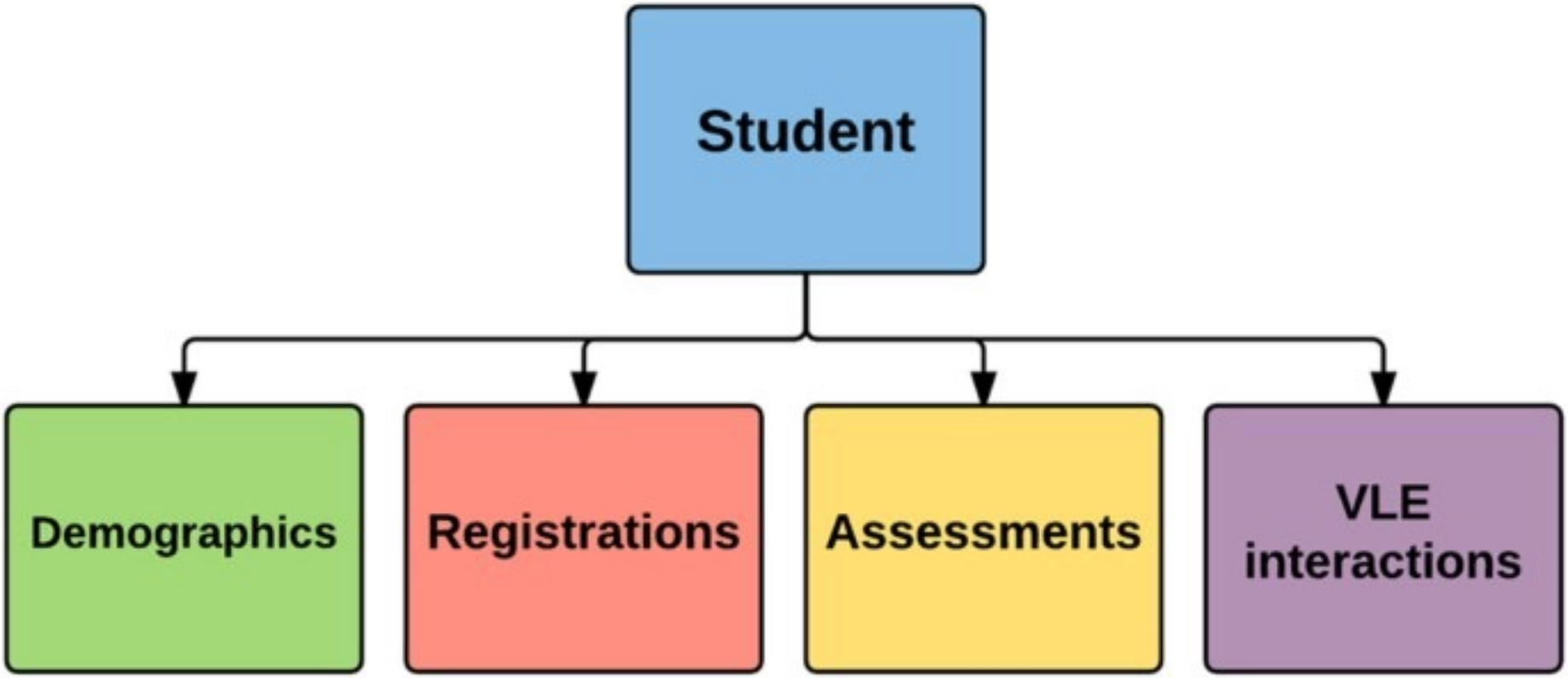
- Each academic year has two semesters – winter (J) and summer (B)
- Typical semester last approximately ¾ of the year
- Students interacts with the Virtual Learning Environment, which contains all the learning resources, assessments and tools for contacting their teachers and peers
- Course-presentation covers multiple topics, each usually ends with the assessment => assessments can be considered as milestones in course-presentation



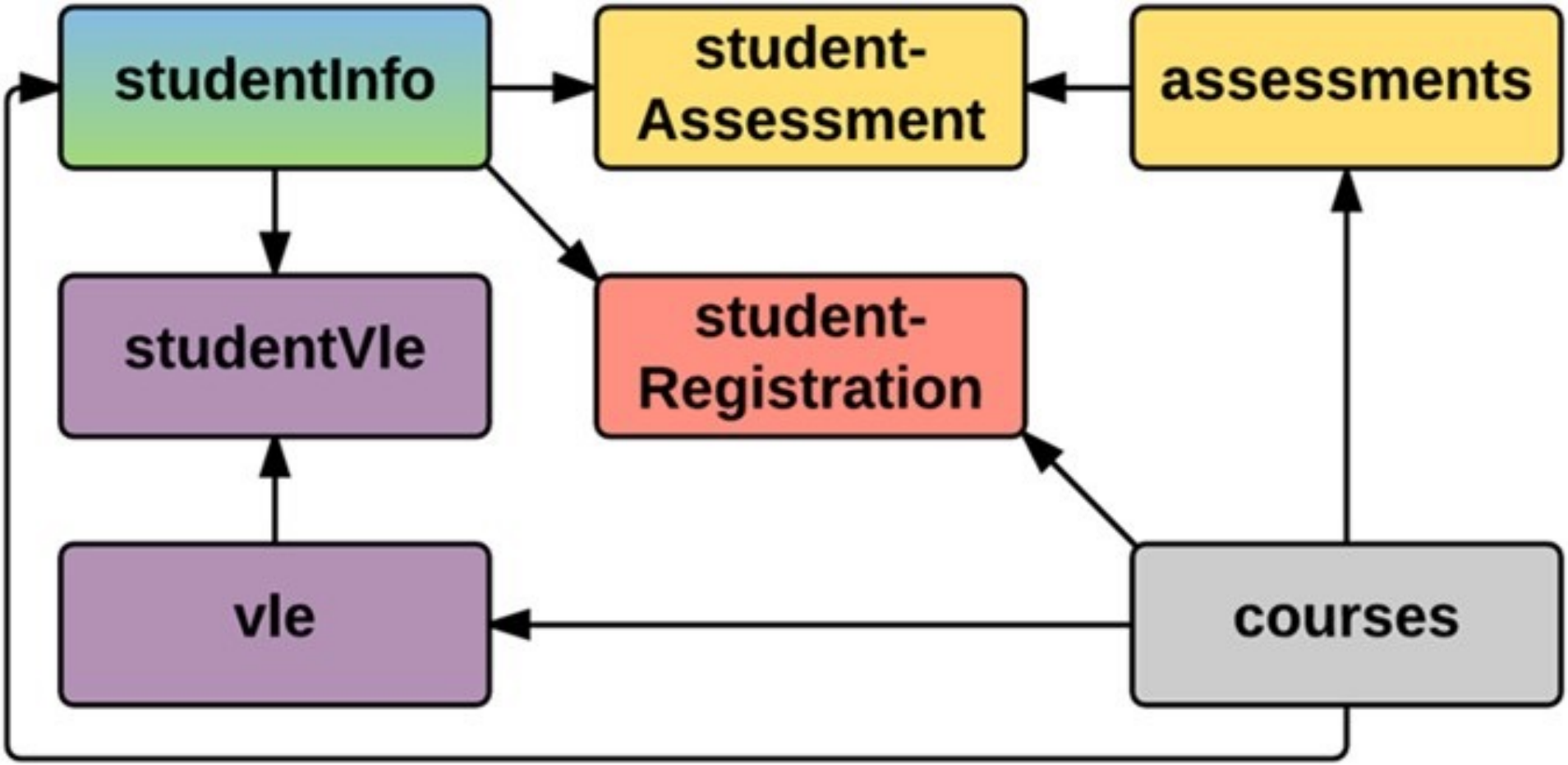
Dataset info

Module	Domain	Presentations	Students
AAA	Social Sciences	2	748
BBB	Social Sciences	4	7,909
CCC	STEM	2	4,434
DDD	STEM	4	6,272
EEE	STEM	3	2,934
FFF	STEM	4	7,762
GGG	Social Sciences	3	2,534

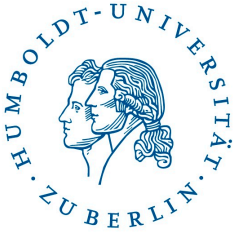
OULAD structure



OULAD structure

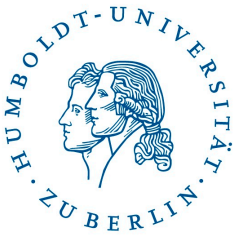


student_info



- ***code_module*** - module identification code on which the student is registered.
- ***code_presentation*** - presentation identification code during which the student is registered on the module.
- ***id_student*** - the unique student identification number.
- *gender* - student's gender.
- *region* - the geographic region, where the student lived while taking the module-presentation.
- *highest_education* - the highest student education level on entry to the module presentation.
- *imd_band* - the IMD band of the place where the student lived during the module-presentation.
- *age_band* - a band of student's age.
- *num_of_prev_attempts* - the number of how many times the student has attempted this module.
- *studied_credits* - the total number of credits for the modules the student is currently studying.
- *disability* - indicates whether the student has declared a disability.
- *final_result* - student's final result in the module-presentation.

assessments

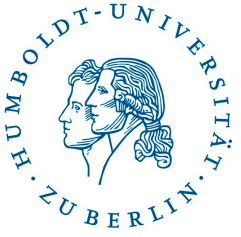


- ***code_module*** - module identification code, to which the assessment belongs.
- ***code_presentation*** - presentation identification code, to which the assessment belongs.
- ***id_assessment*** - assessment identification number.
- assessment_type - a type of assessment. Three types of assessments exist—Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- *date* - information about the cut-off day of the assessment.
- *weight* - the weight of the assessment. Typically, Exams are treated separately and have the weight equal to 100%; the sum of all other assessments is also 100%.

student_assesment

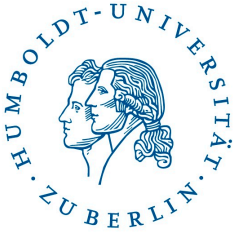
- ***id_assessment*** - the assessment identification number.
- ***id_student*** - the unique student identification number.
- *date_submitted* - the day of assessment submission.
- *is_banked* - the status flag indicating that the assessment result has been transferred from a previous presentation.
- *score* - the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

courses



- ***code_module*** - code name of the module, which serves as the identifier.
- ***code_presentation*** - code name of the presentation.
- *length* - the length of the module-presentation in days from module start date to module end date.

student_registration



- ***code_module*** - the module identification code.
- ***code_presentation*** - the presentation identification code.
- ***id_student*** - the unique student identification number.
- *date_registration* - the day of student's registration for the module presentation.
- *date_unregistration* - the day of student unregistration from the module presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result in the studentInfo table.

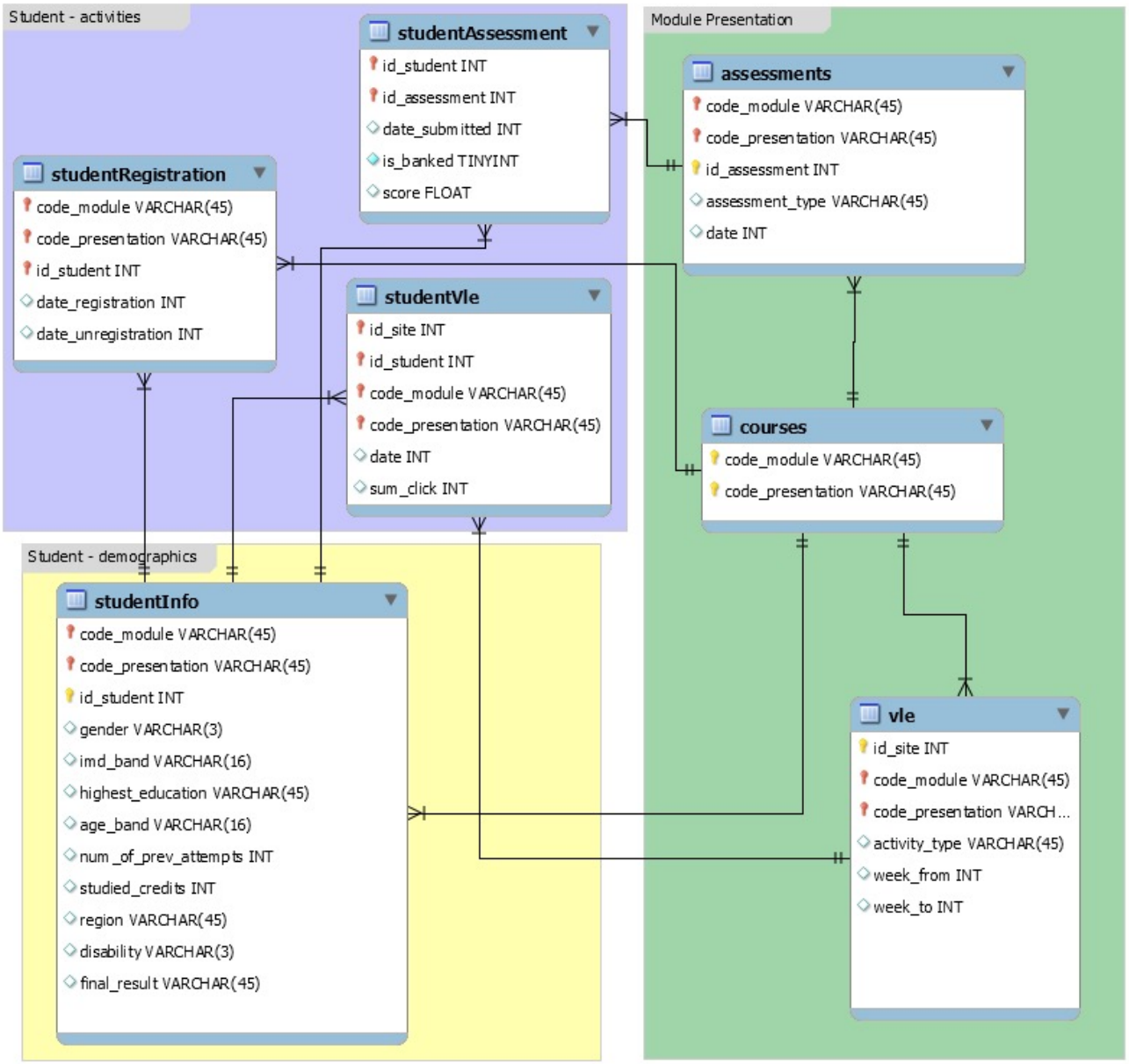
vle

- ***id_site*** - the identification number of the material.
- ***code_module*** - the identification code for the module.
- ***code_presentation*** - the identification code of the presentation.
- *activity_type* - the role associated with the module material.
- *week_from* - the week from which the material is planned to be used.
- *week_to* - the week until which the material is planned to be used.

student_vle

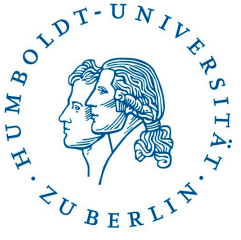
- ***code_module*** - the module identification code.
- ***code_presentation*** - the presentation identification code.
- ***id_student*** - the unique student identification number.
- ***id_site*** - the VLE material identification number.
- *date* - the day of student's interaction with the material.
- *sum_click* - the number of times the student interacted with the material.

E-R diagram


Student - activities
studentAssessment
id_student INT
id_assessment INT
date_submitted INT
is_banked TINYINT
score FLOAT
studentRegistration
code_module VARCHAR(45)
code_presentation VARCHAR(45)
id_student INT
date_registration INT
date_unregistration INT
studentVle
id_site INT
id_student INT
code_module VARCHAR(45)
code_presentation VARCHAR(45)
date INT
sum_click INT
Module Presentation
assessments
code_module VARCHAR(45)
code_presentation VARCHAR(45)
id_assessment INT
assessment_type VARCHAR(45)
date INT
courses
code_module VARCHAR(45)
code_presentation VARCHAR(45)
Student - demographics
studentInfo
code_module VARCHAR(45)
code_presentation VARCHAR(45)
id_student INT
gender VARCHAR(3)
imd_band VARCHAR(16)
highest_education VARCHAR(45)
age_band VARCHAR(16)
num_of_prev_attempts INT
studied_credits INT
region VARCHAR(45)
disability VARCHAR(3)
final_result VARCHAR(45)
vle
id_site INT
code_module VARCHAR(45)
code_presentation VARCH...
activity_type VARCHAR(45)
week_from INT
week_to INT

Template solution

Gitlab project



1. Fork project from:
   https://gitlab.informatik.hu-berlin.de/kuzilekj/la-dashboards-course
2. Add kuzilekj as maintainer to your project
3. Do the development
4. After final presentation create merge request

Docker

1. Install Docker desktop (https://www.docker.com/get-started)
2. Checkout the git project to your local machine
3. Run: docker compose up

Required versions:

- Docker Engine 20.10.7 or higher
- Docker Compose 1.29.2 or higher

Voting & Group formation