FROM SIS TO DATA-DRIVEN INSIGHTS

EDUCATION ANALYTICS AT SCALE
What is your role?
Introduction
SISs have interesting data, but aren’t aimed at analysis

- **Situation**
  - Student Information Systems (SISs) contain a lot of information about students
  - However, focus is on operational support
  - Issues with data quality and combining data

- **Effect**
  - Doing an analysis demands a lot of data wrangling

- **Result**
  - Analyses take a lot of time and are often inconsistent with each other
## CONTENT

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>1. Building data assets</td>
<td>6</td>
</tr>
<tr>
<td>2. Report quickly and reliable</td>
<td>14</td>
</tr>
<tr>
<td>3. Further together</td>
<td>20</td>
</tr>
<tr>
<td>Discussion</td>
<td>24</td>
</tr>
</tbody>
</table>
Introduction

SISs have interesting data, but aren’t aimed at analysis

• **Source**
  - External (DUO, CROHO)
  - Internal (SIS, Data warehouse, policy-makers)

• **Aggregation level**
  - Student-year-programme (Registration)
  - Process logs (BSA)

• **Availability**
  - Automated exports (SIS, DWH)
  - Manual (policy makers)

• **Frequency**
  - Daily (pre-enrollment)
  - Weekly (results)
  - Monthly (registrations)
  - Yearly (study programmes)
Building data assets: Data pipeline
Data is first cleaned and aggregated, then analysed and reported

01. Audit  02. Prepare  03. Aggregate  04. Analyse  05. Report

Data Assets
Building data assets: Auditing

Ensure data is technically correct and column names are interpretable

Checks
✓ Column type (numeric, character)
✓ Range of values (2010 – 2022, pre-defined subsets)
✓ Usage (TRUE or FALSE)
✓ Percentage of missing data per column (10%)
Building data assets: Preparation

Improve data quality and usability based on context and domain knowledge

Changes
- Improve in-record consistency
  - The date of a result (a grade) can’t be in 2022 if it is entered in SIS in 2021
- Improve cross-record consistency
  - Update study programmes to current names
- Creating categoric variables with mapping tables
  - Summarise ~100 types of previous education to ~6 categories
- Calculate a new variables based on others
  - Year in programme based on the number of registrations of a student in a programme
- Removing columns that aren’t useful for analysis
  - The date of result entered in SIS
- Remove rows with low data quality
  - Rows without student number
Building data assets: Aggregation

Data is aggregated and combined into four related data assets

- Data is aggregated and combined into four data assets

- A data asset is high quality, very complete data set that is ready for analysis (Kruhse-Lethonen & Hofmann, 2020)

- Most data is aggregated to the enrollment level

- For specific analyses, data can be aggregated to other levels, for instance:
  - student level (all registrations of a student)
Building data assets: Analysis
All structural analyses share the same general method, more on June 17th

Requirements
✓ Assessable
✓ Adaptable
✓ Extensible
✓ Explainable
✓ Interactive

June 17th 2022 → HOlink conference
➢ Prediction of registrations
➢ Link to program
Building data assets: Reporting

Reports are based on the data assets, privacy of individuals is ensured

- Selection and possible combination of the data assets
  - Programme characteristics added to registrations data
- Visualisations almost always made in Tableau
  - For a specific use-case R Shiny is used
- Interactive dashboards
  - Users have several filtering options available
- Shared with all colleagues
  - Every VU professional can request access
- Privacy by design
  - Individual data itself is not visible (only aggregated trends)
  - Unique identifiers are hashed
Building data assets: Question
<table>
<thead>
<tr>
<th>CONTENT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>1. Building data assets</td>
<td>6</td>
</tr>
<tr>
<td>2. Report quickly and reliable</td>
<td>14</td>
</tr>
<tr>
<td>3. Further together</td>
<td>20</td>
</tr>
<tr>
<td>Discussion</td>
<td>24</td>
</tr>
</tbody>
</table>
Report quickly and reliable: Study progress

Automated general analysis
Report quickly and reliable: Study progress
Automated general analysis, with a range of indicators
Report quickly and reliable: Study progress
Automated general analysis, with a range of indicators, at multiple levels
Report quickly and reliable: Impact Covid
High-quality ad-hoc analyses are possible, which proved useful during Covid

- The above analysis grouped cohorts together, but what happened with study success during Covid?
- Since the necessary information was already present in the data assets, such ad-hoc analysis are a matter of dashboard design, without worries about data-engineering and data-quality issues
Report quickly and reliable: Participation
## CONTENT

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>1. Building data assets</td>
<td>6</td>
</tr>
<tr>
<td>2. Report quickly and reliable</td>
<td>14</td>
</tr>
<tr>
<td>3. Further together</td>
<td>20</td>
</tr>
<tr>
<td>Discussion</td>
<td>24</td>
</tr>
</tbody>
</table>
Further together: Collaborations on project level
Partnering on joint challenges

Forecasting student population
• (Partly) national data
• Informal gatherings
• Show and discuss code, modelling and hiccups along the way

Programme feasibility research
• Vrije Universiteit Amsterdam and University of Twente are starting technical programmes together
• Provide high school graduates in region of Amsterdam with more technical programmes
• More information on general collaboration: https://vu-utwente.nl/

Applied Research and dashboards
• Plan for Success
• National Cohort Research Higher Education (NCO-HO)
Further together: Intensive collaboration
Two-year+ collaboration with Applied University the Hague

Disseminate (way of) work
• Export of project documentation
• Hands-on support to set-up tooling
• Export of all code

Pilot
• Guidance, mentoring and code review in pilot project
• Working step-by-step through data pipeline

Further support and collaboration
• Export of new code for two years
• Support-on-demand for two years
• Collaboration as equals into the future
Further Together: Acceleration Plan

Participation in zone Secure and reliable use of education data

Privacy and Ethics Reference Framework for Education data
• Based on early version of Code of Practice VU
• More information on site Acceleration Plan

Quickscan Education data
• https://www.versnellingsplan.nl/en/Kennisbank/quickscan-education-data/
• Sessions Round II, Through the looking glass, by Tom Konings & Justian Knobbout

Simulation dataset
• Together with Erasmus University we made a simulation dataset of education data
• More information on site Acceleration Plan and even more information (in Dutch)

Data science teams in HE
• Coming soon
## CONTENT

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>1. Building data assets</td>
<td>6</td>
</tr>
<tr>
<td>2. Report quickly and reliable</td>
<td>14</td>
</tr>
<tr>
<td>3. Further together</td>
<td>20</td>
</tr>
<tr>
<td>Discussion</td>
<td>24</td>
</tr>
</tbody>
</table>
Discussion

Moving forward in data analytics in higher education

• Set up a pipeline for the creation of data assets

• Extend the pipeline to (real-time) modelling and visualisations

• Learn from and with each other

• How do you think we need to move forward?
Version History

Versie 0.1 – 19-05-2022, CdH, Part 1 and 2
Versie 1.0 – 27-05-2022, CdH, Part 3 and finishing touches