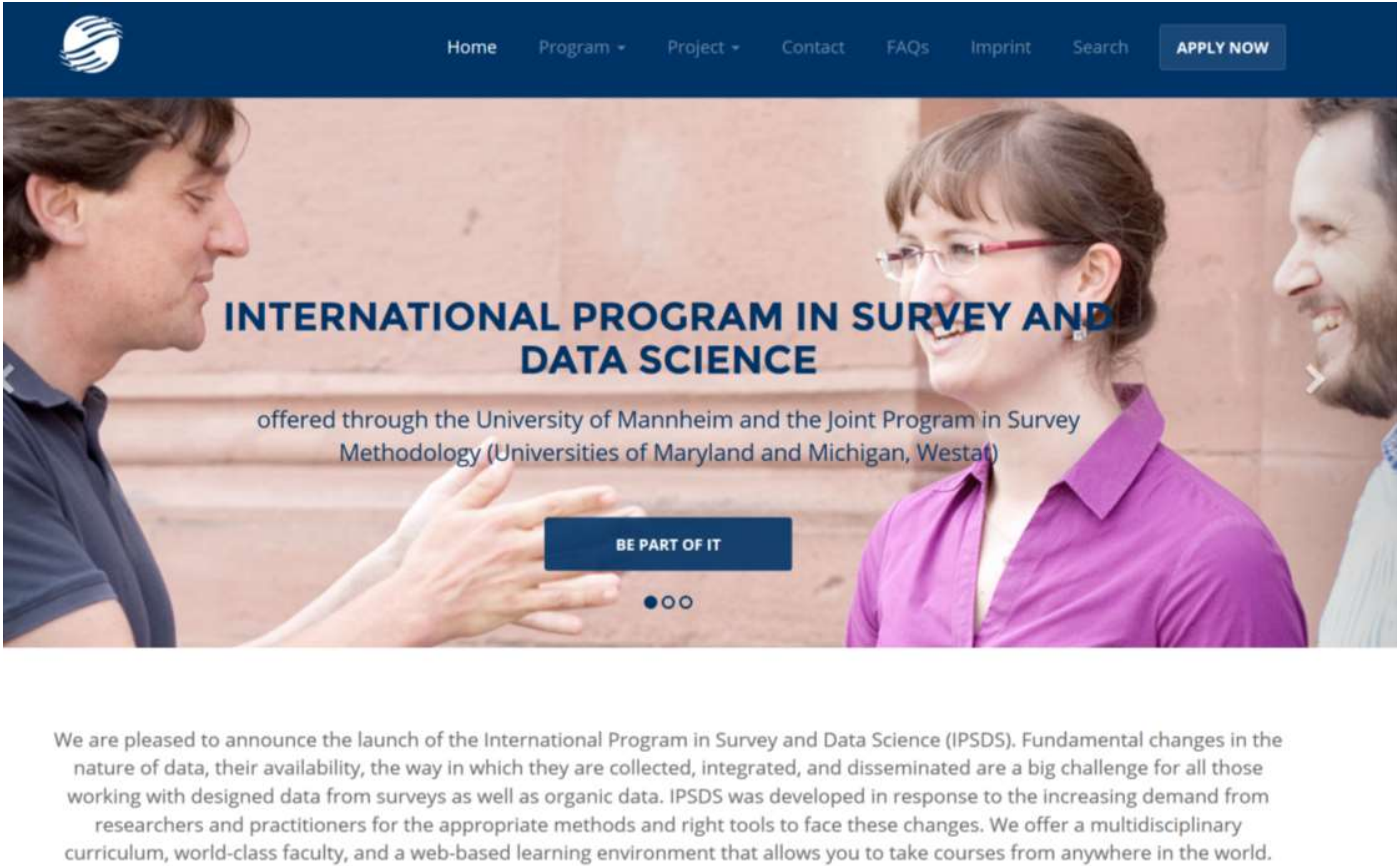
Training – Survey and Data Science

Frauke Kreuter

JPSM – Uni Mannheim – IAB

Canberra 2018

survey-data-science.net



INTERNATIONAL PROGRAM IN SURVEY AND DATA SCIENCE

offered through the University of Mannheim and the Joint Program in Survey Methodology (Universities of Maryland and Michigan, Westat)

We are pleased to announce the launch of the International Program in Survey and Data Science (IPSDS). Fundamental changes in the nature of data, their availability, the way in which they are collected, integrated, and disseminated are a big challenge for all those working with designed data from surveys as well as organic data. IPSDS was developed in response to the increasing demand from researchers and practitioners for the appropriate methods and right tools to face these changes. We offer a multidisciplinary curriculum, world-class faculty, and a web-based learning environment that allows you to take courses from anywhere in the world.

Problem we tried to solve – In brief

- Allow for multidisciplinary curriculum
- Modularized – adapt to prior skills and work needs
- Relevant methods and tools
- Mix of faculty from academia and industry

Key elements:

- Flexible web-based learning environment
- Live (video) interaction with faculty and students
- Face-to-face networking meetings

Why regular Data Science courses don‘t work

- Little discussion of data quality
- Data Science happens in context
- Single data sources unlike to be sufficient
- Combination of surveys and other data sources needed

Partners and Funding

University Partners

- University of Maryland
- University of Mannheim

- Catholic University of Santiago de Chile
- Australian National Unversity
- Beijing University
- Ashoka University (expressed interest)
- U. of Capetown (planned)

Other Partners

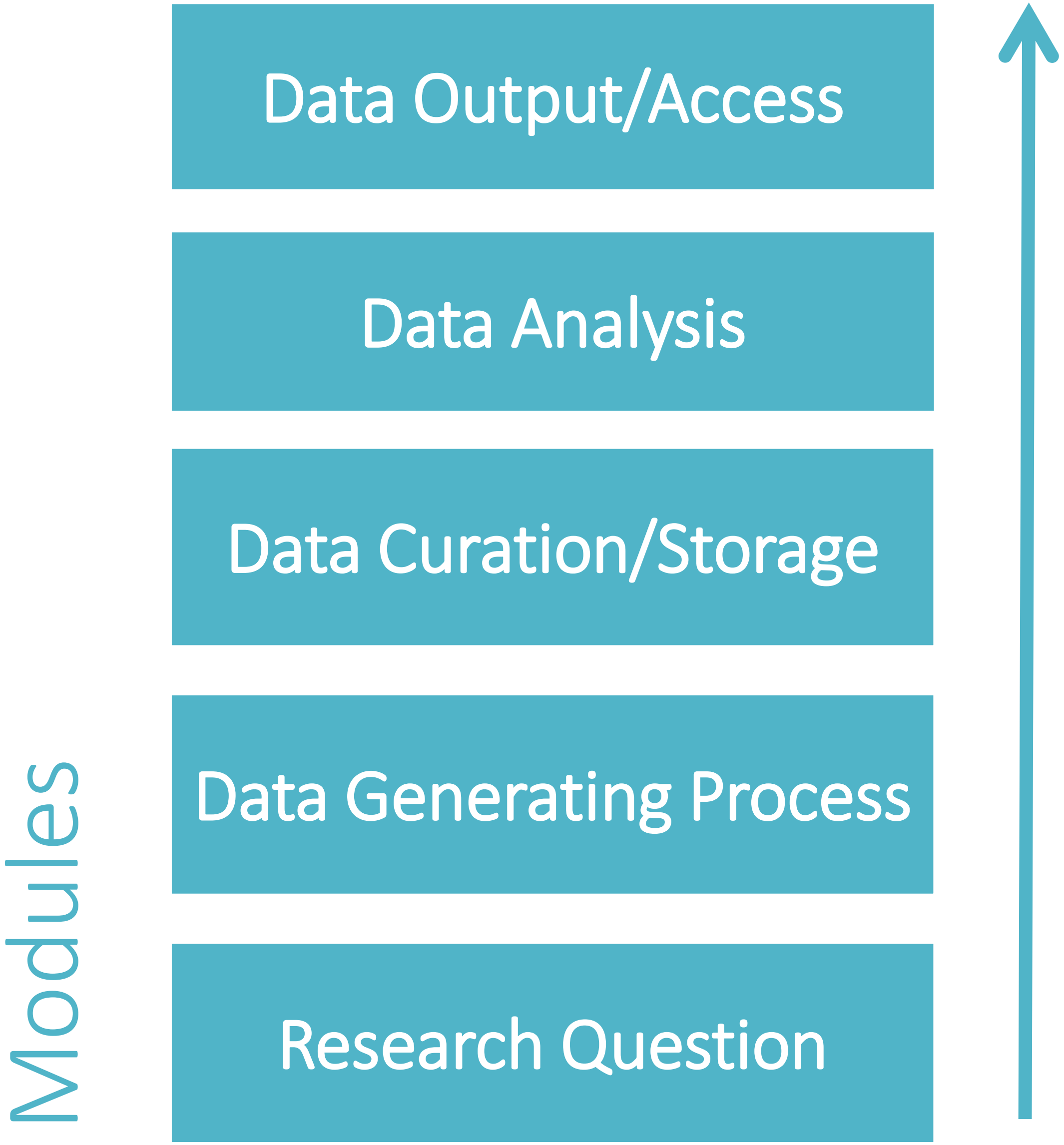
- SRO - Michigan
- PEW
- German Record Linkage Center
- GESIS
- Bureau of Labour Statistics
- U.S. Census Bureau
- Statistics Netherlands

SPONSORED BY THE

Federal Ministry
of Education
and Research

AUFSTIEG DURCH
BILDUNG
OFFENE HOCHSCHULEN

The project on which this report is based was promoted with funds from the Federal Ministry of Education and Research under the reference number (16OH22064]. Responsibility for the contens of this publication lies with the author.



Modules

Data Output/Access

Learn how to communicate results, distribute and store your data; Ethics

Data Analysis

Learn a variety of analysis methods suited for different data types

Data Curation/Storage

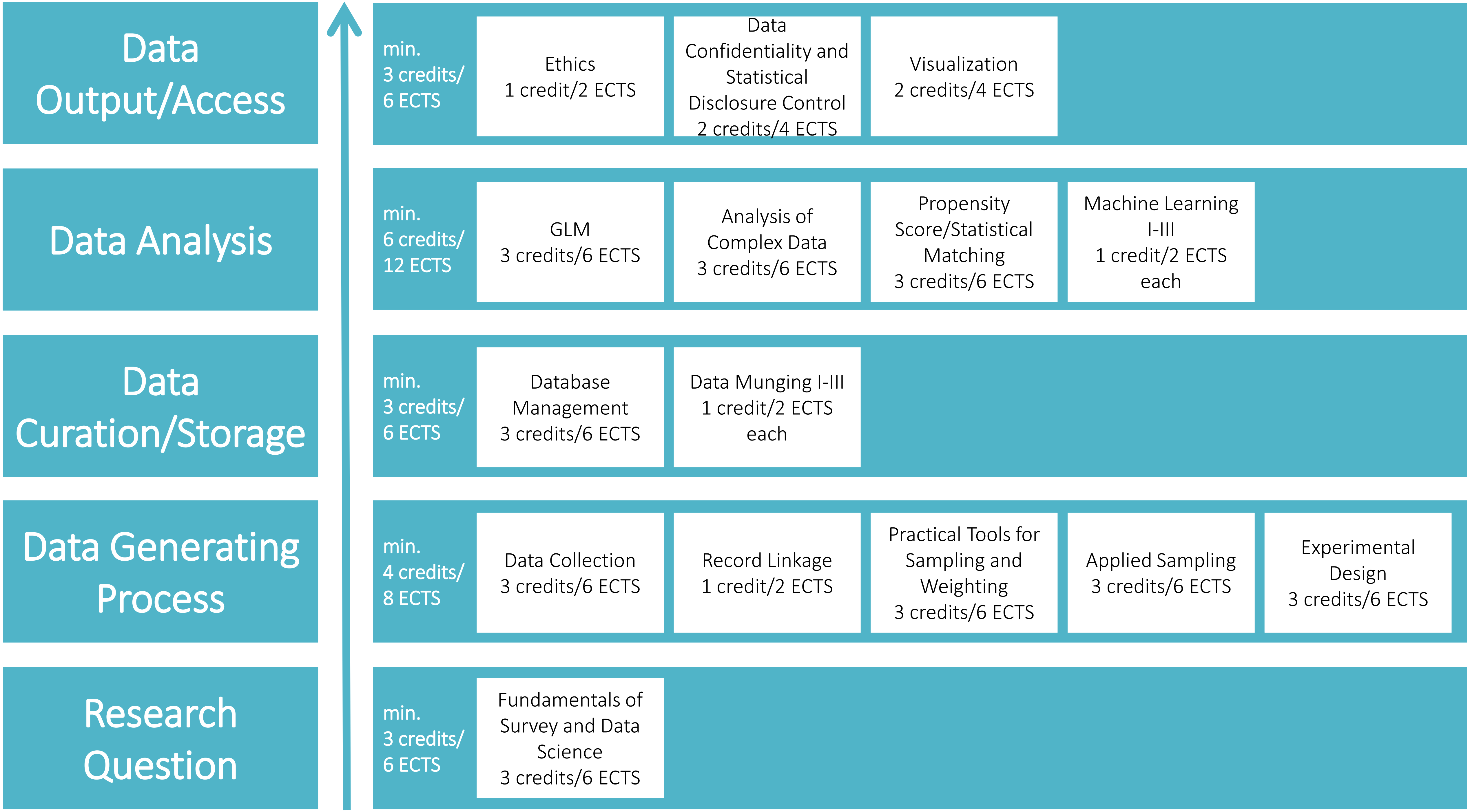
Learn how to curate and manage data

Data Generating Process

Understand how to collect data, and how data are generated through administrative and other processes.

Research Question

Learn how to ask the right question and evaluate which data can/should be used to answer it



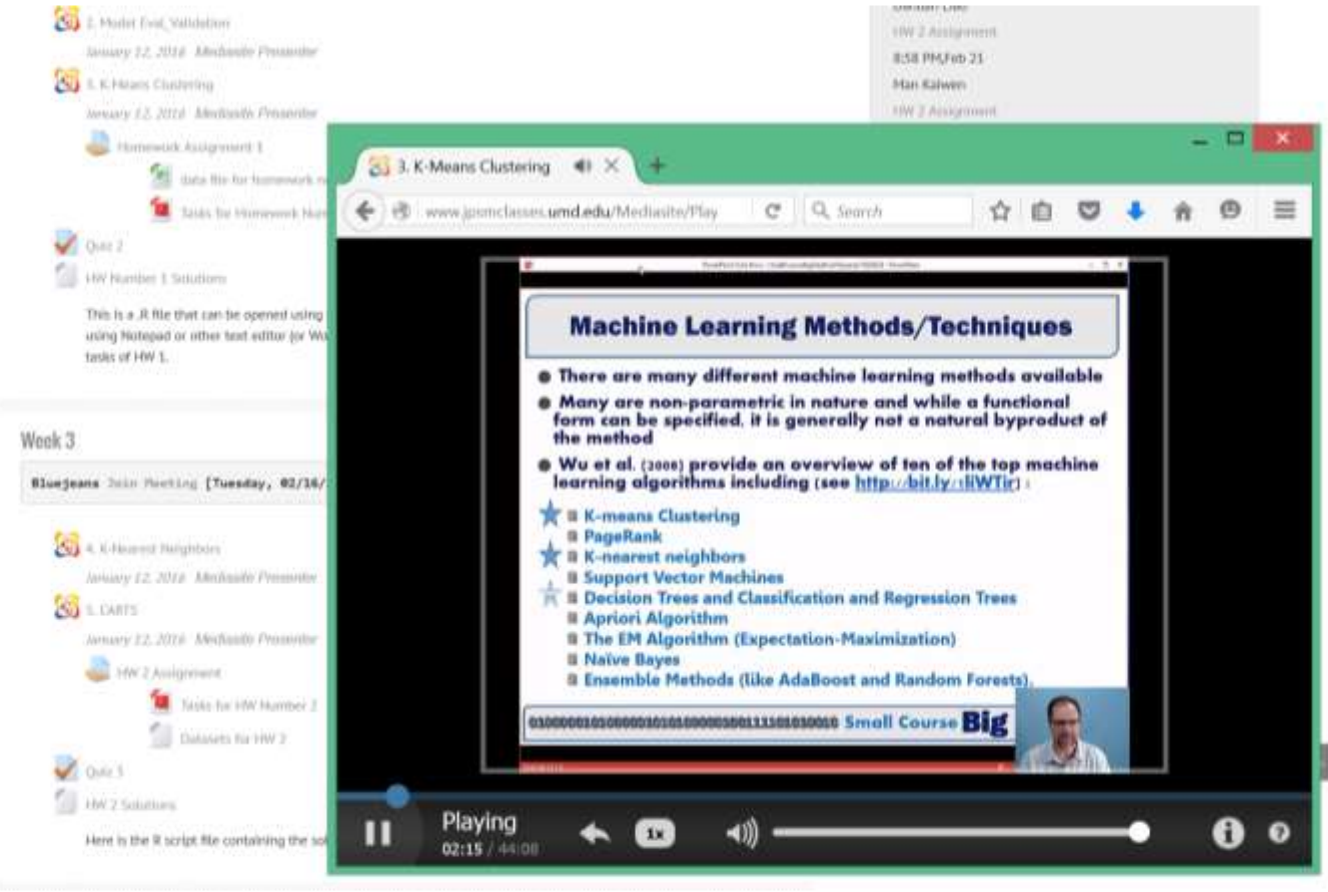
Data Output/Access	min. 3 credits/ 6 ECTS	Ethics 1 credit/2 ECTS	Data Confidentiality and Statistical Disclosure Control 2 credits/4 ECTS	Visualization 2 credits/4 ECTS		
Data Analysis	min. 6 credits/ 12 ECTS	GLM 3 credits/6 ECTS	Analysis of Complex Data 3 credits/6 ECTS	Propensity Score/Statistical Matching 3 credits/6 ECTS	Machine Learning I-III 1 credit/2 ECTS each	
Data Curation/Storage	min. 3 credits/ 6 ECTS	Database Management 3 credits/6 ECTS	Data Munging I-III 1 credit/2 ECTS each			
Data Generating Process	min. 4 credits/ 8 ECTS	Data Collection 3 credits/6 ECTS	Record Linkage 1 credit/2 ECTS	Practical Tools for Sampling and Weighting 3 credits/6 ECTS	Applied Sampling 3 credits/6 ECTS	Experimental Design 3 credits/6 ECTS
Research Question	min. 3 credits/ 6 ECTS	Fundamentals of Survey and Data Science 3 credits/6 ECTS				

Format

Each week set of videos (pre-recorded)

Lectures are broken into easily digestible sessions to help participants to better focus on the material

Engage with the material at their own pace



2. Model Eval_Validation

January 12, 2016 Mediasite Presenter

3. K-Means Clustering

January 12, 2016 Mediasite Presenter

Homework Assignment 1

data file for homework n

Tasks for Homework Num

Quiz 2

HW Number 1 Solutions

This is a .R file that can be opened using using Notepad or other text editor (or Wo tasks of HW 1.

Week 3

Bluejeans Join Meeting **[Tuesday, 02/16/**

4. K-Nearest Neighbors

January 12, 2016 Mediasite Presenter

5. CARTS

January 12, 2016 Mediasite Presenter

HW 2 Assignment

Tasks for HW Number 2

Datasets for HW 2

Quiz 3

HW 2 Solutions

Here is the R script file containing the so

HW 2 Assignment

8:58 PM,Feb 21

Han Kaiwen

HW 2 Assignment

3. K-Means Clustering

www.jpsmclasses.umd.edu/Mediasite/Play

Machine Learning Methods/Techniques

- There are many different machine learning methods available
- Many are non-parametric in nature and while a functional form can be specified, it is generally not a natural byproduct of the method
- Wu et al. (2008) provide an overview of ten of the top machine learning algorithms including (see http://bit.ly/1liWTir) :
  - K-means Clustering
  - PageRank
  - K-nearest neighbors
  - Support Vector Machines
  - Decision Trees and Classification and Regression Trees
  - Apriori Algorithm
  - The EM Algorithm (Expectation-Maximization)
  - Naïve Bayes
  - Ensemble Methods (like AdaBoost and Random Forests).

010000010100000101010000010011110101010010 Small Course **Big**

Annual „Connect“ Event

http://coleridgeinitiative.org
http://survey-data-science.net/

fkreuter@umd.edu