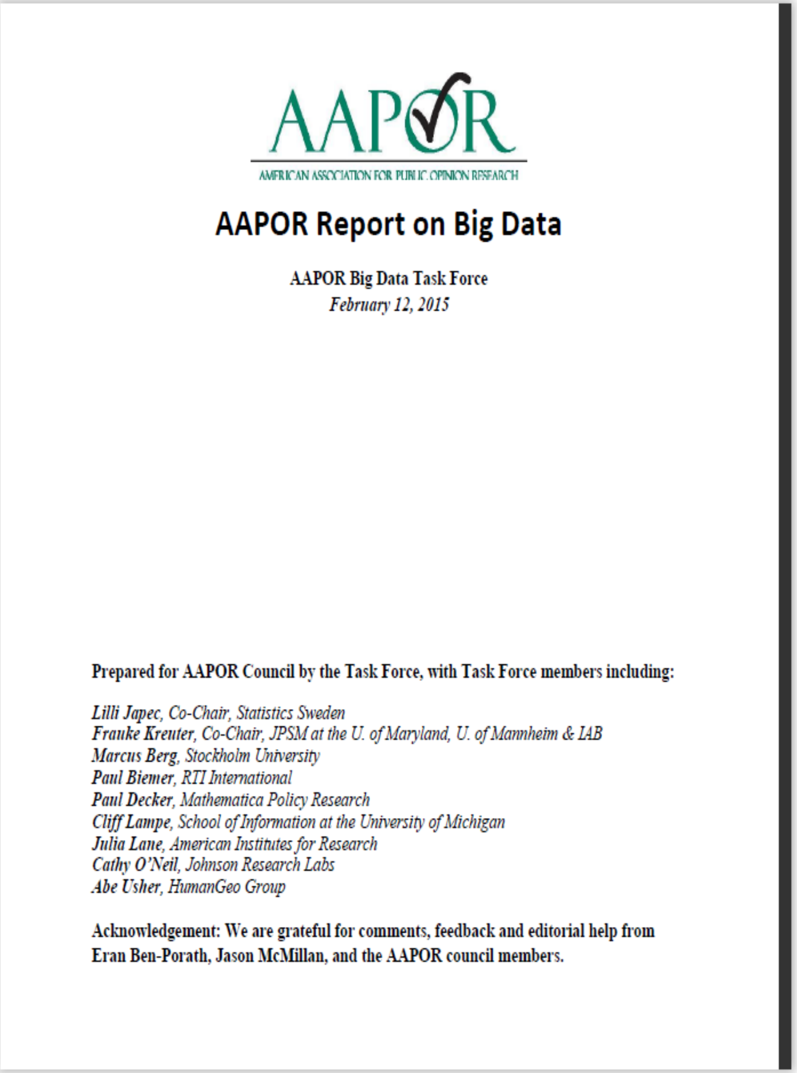
International Program in Survey and Data Science

Frauke Kreuter

JPSM – Uni Mannheim – IAB

ASI 28.06.2017



AAPOR

AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

AAPOR Report on Big Data

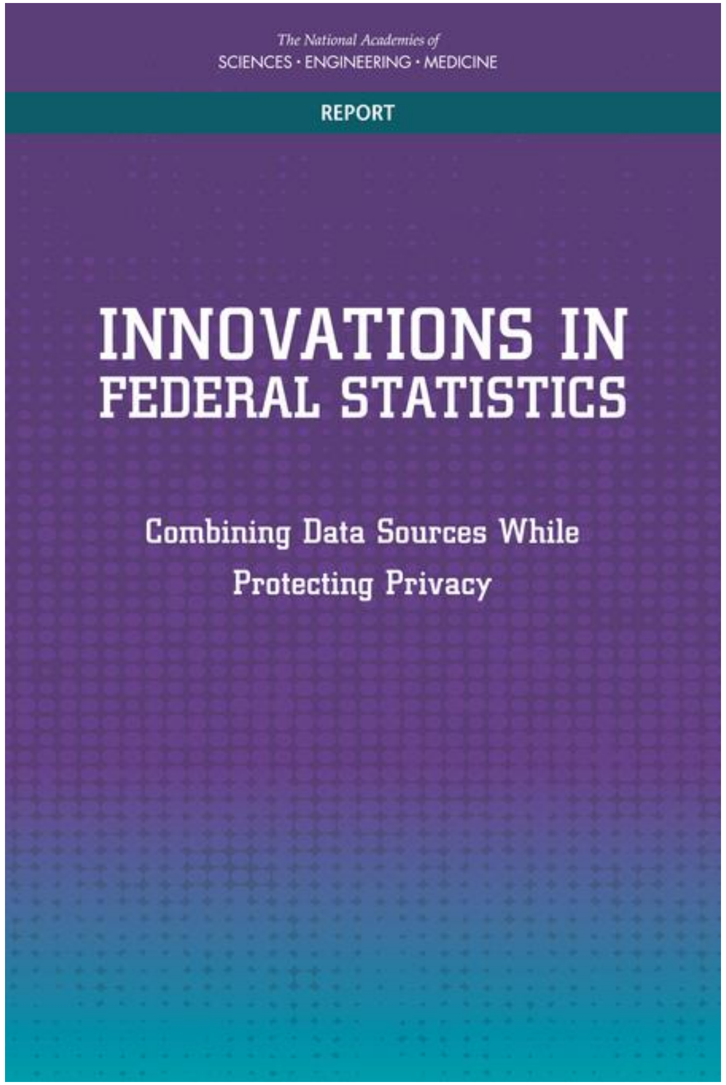
AAPOR Big Data Task Force

February 12, 2015

Prepared for AAPOR Council by the Task Force, with Task Force members including:

Lilli Japec, Co-Chair, Statistics Sweden
Frauke Kreuter, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB
Marcus Berg, Stockholm University
Paul Biemer, RTI International
Paul Decker, Mathematica Policy Research
Cliff Lampe, School of Information at the University of Michigan
Julia Lane, American Institutes for Research
Cathy O'Neil, Johnson Research Labs
Abe Usher, HumanGeo Group

Acknowledgement: We are grateful for comments, feedback and editorial help from Eran Ben-Porath, Jason McMillan, and the AAPOR council members.

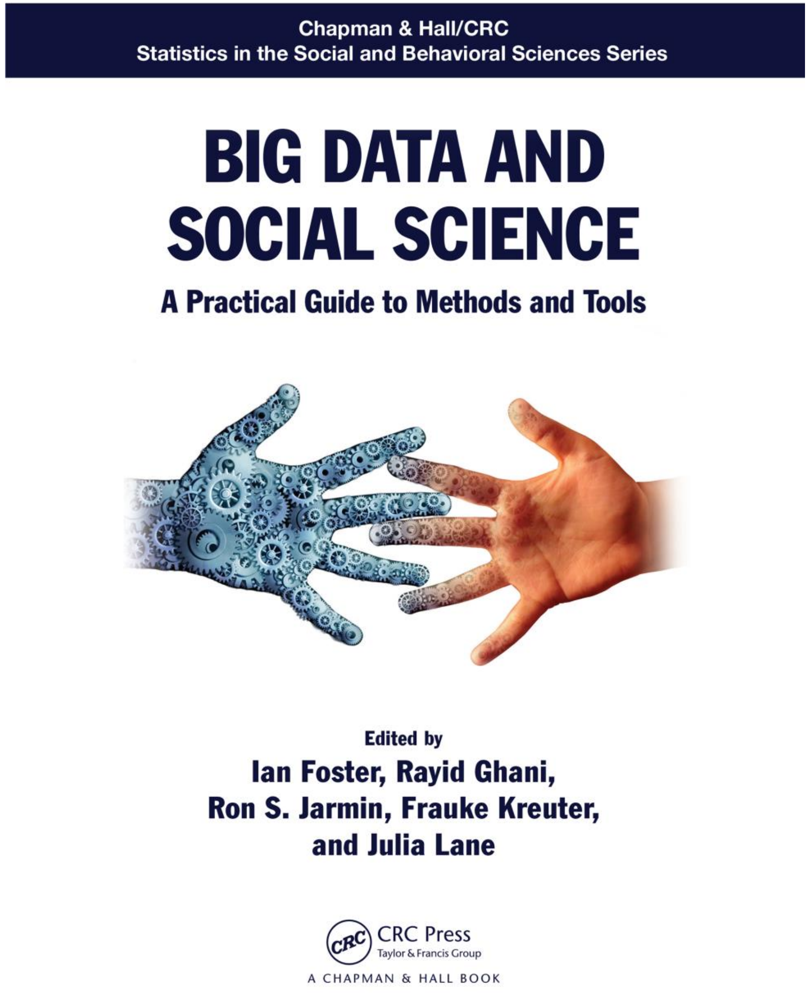


The National Academies of
SCIENCES · ENGINEERING · MEDICINE

REPORT

INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy



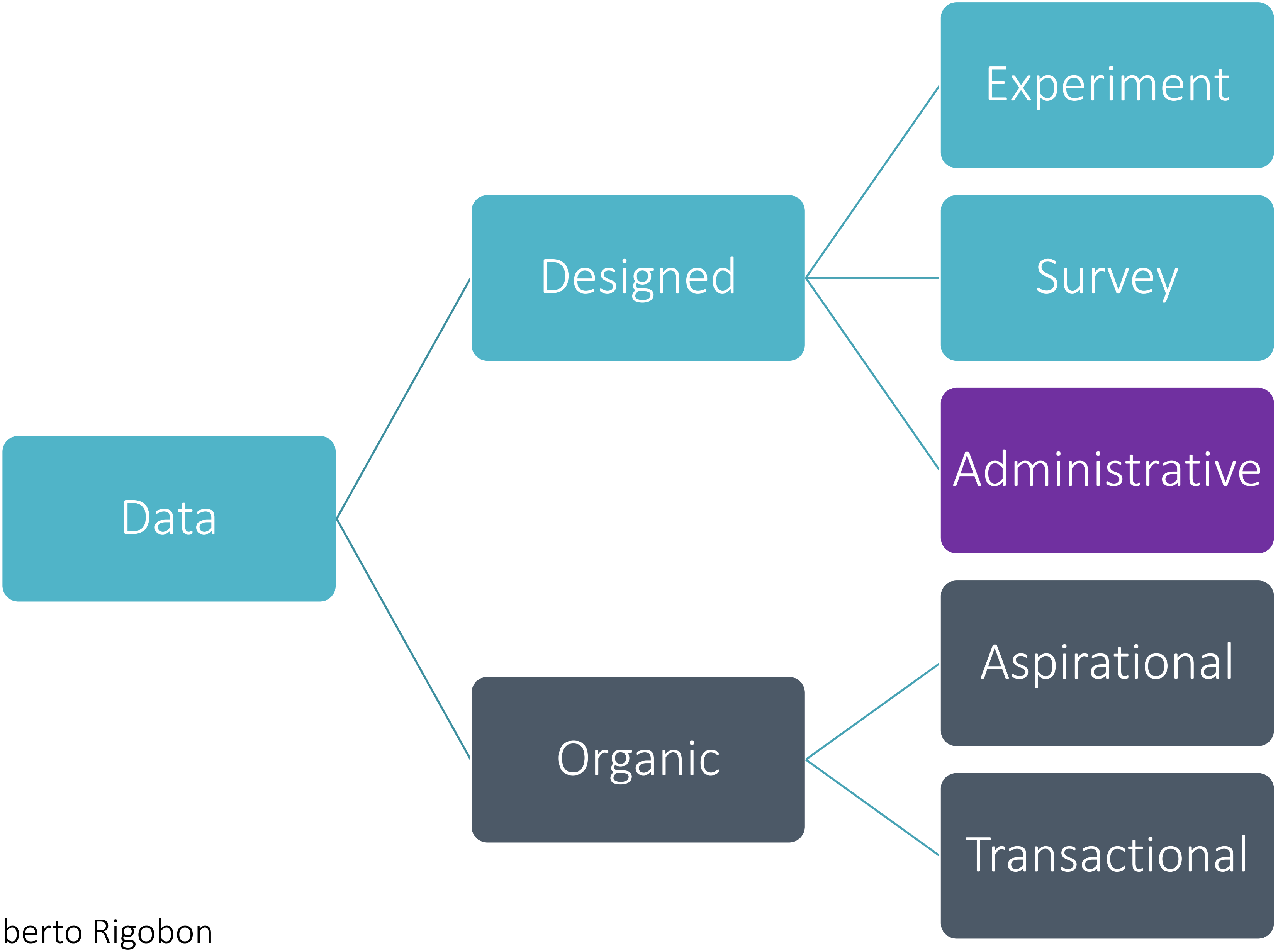
Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools

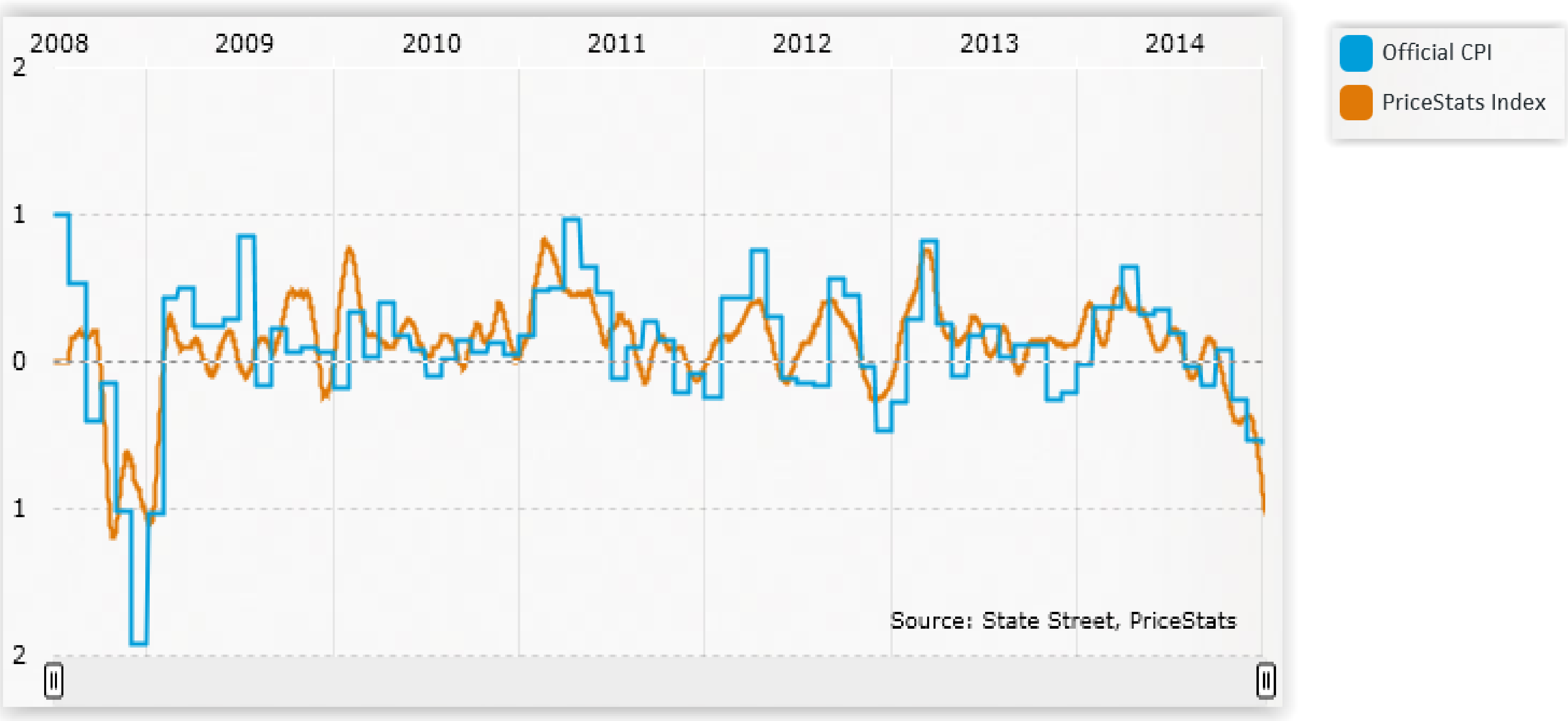
Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



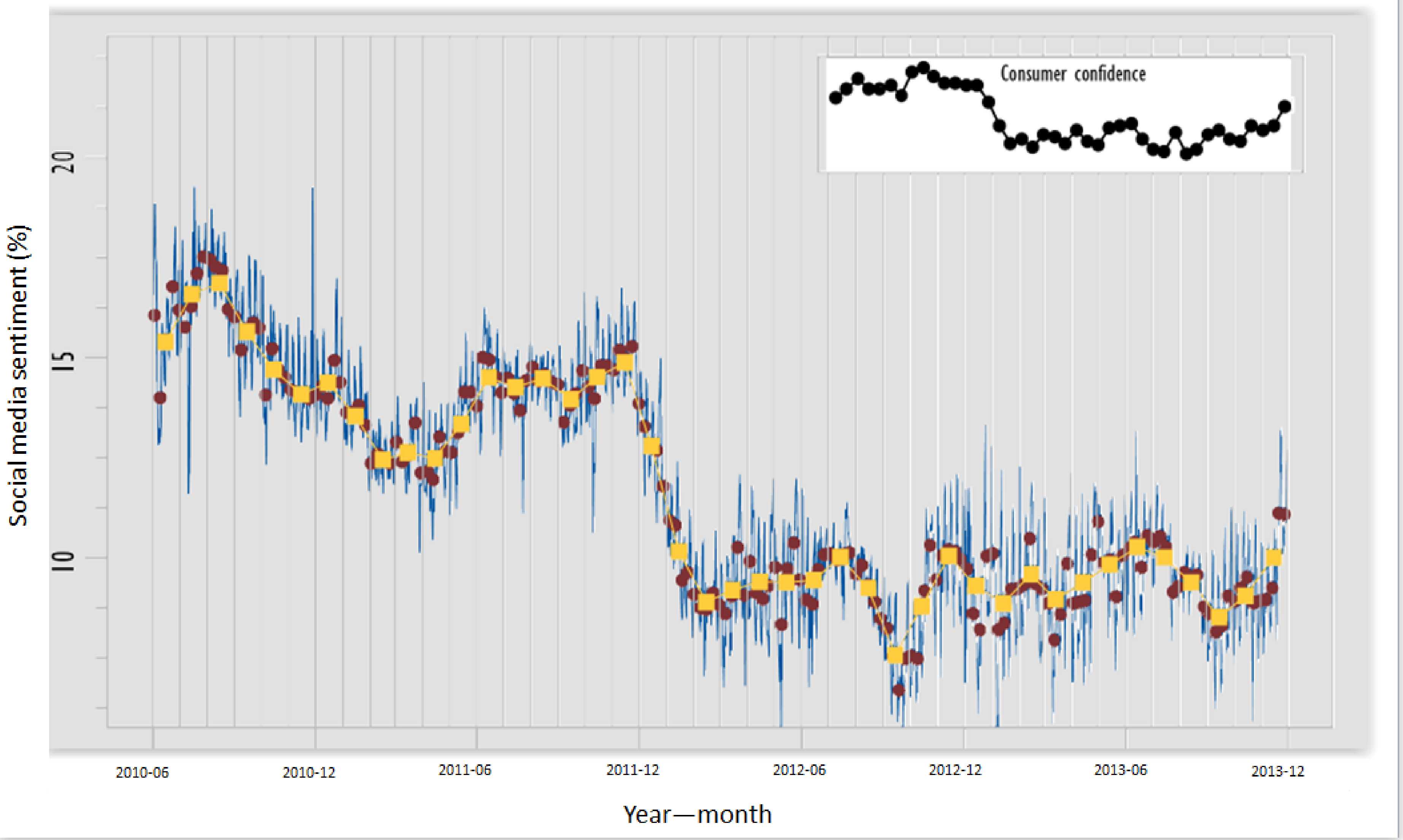
Source: Roberto Rigobon

The Excitement



Source: State Street, PriceStats

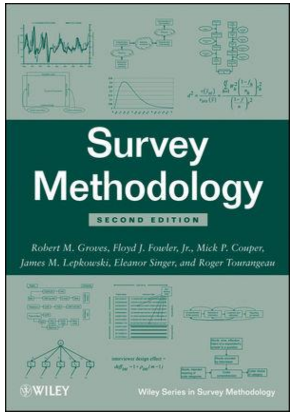
US Aggregated Inflation Series, Monthly Rate,
PriceStats Index vs. Official CPI. Accessed January 18, 2015 from the PriceStats website.



Social media sentiment (daily, weekly and monthly) in the Netherlands, June 2010 - November 2013. The development of consumer confidence for the same period is shown in the insert (Daas and Puts 2014).

The Doubt

Data Generating Process

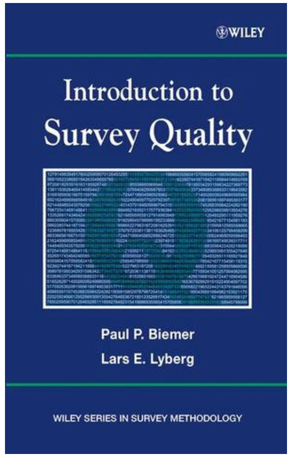

Survey Methodology

SECOND EDITION

Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau

WILEY

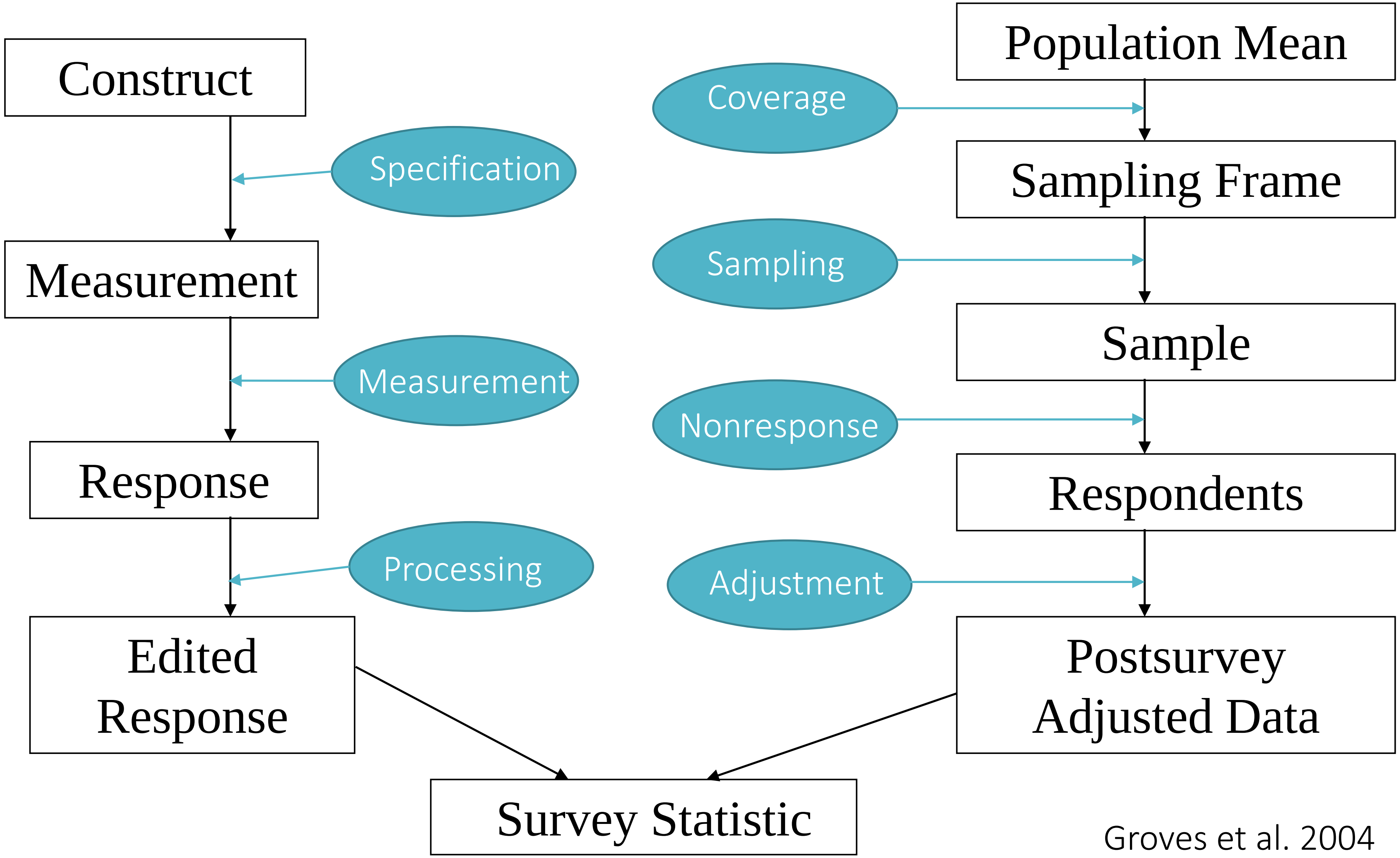
Wiley Series in Survey Methodology


WILEY

Introduction to Survey Quality

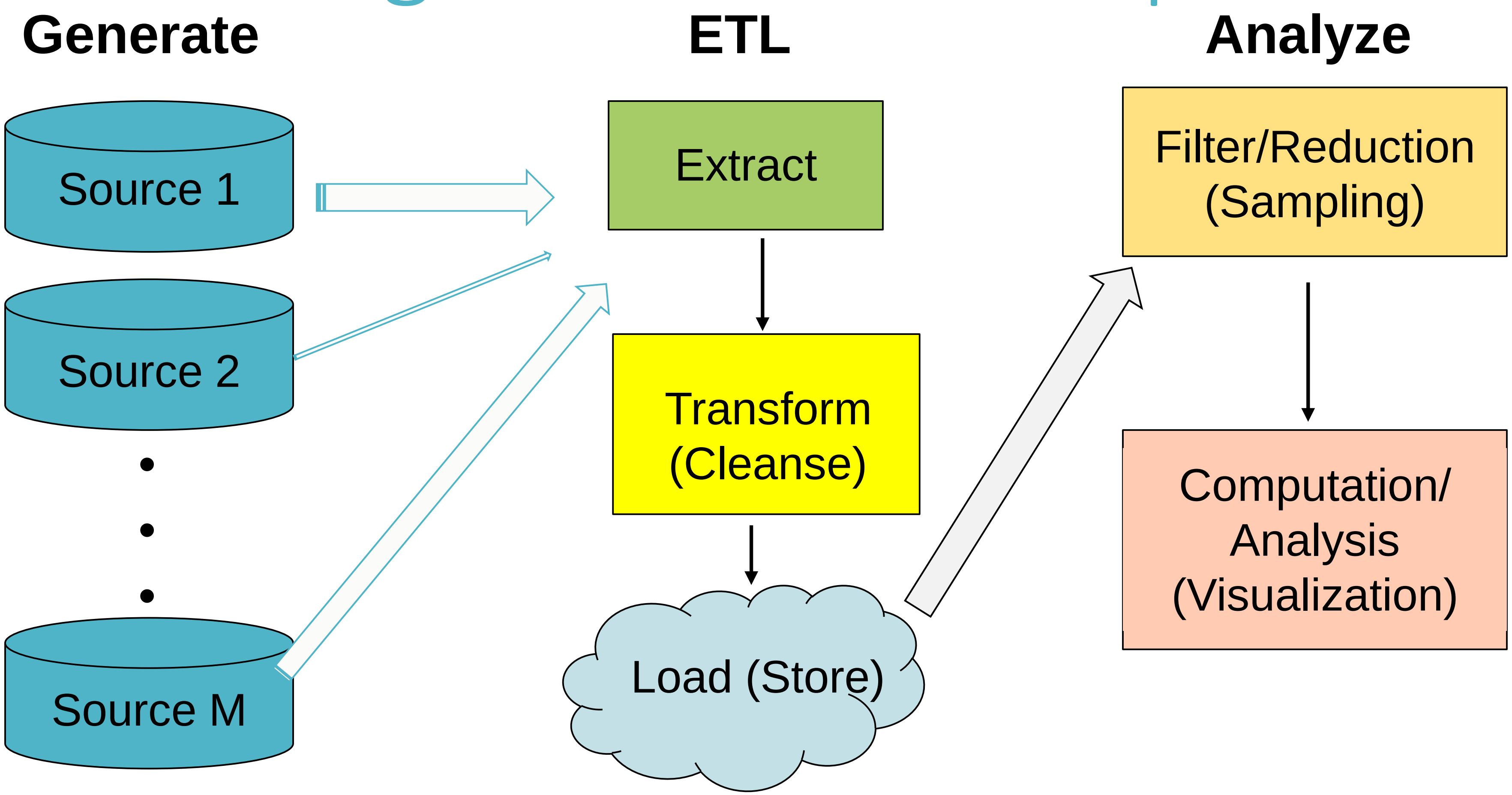
Paul P. Biemer
Lars E. Lyberg

WILEY SERIES IN SURVEY METHODOLOGY



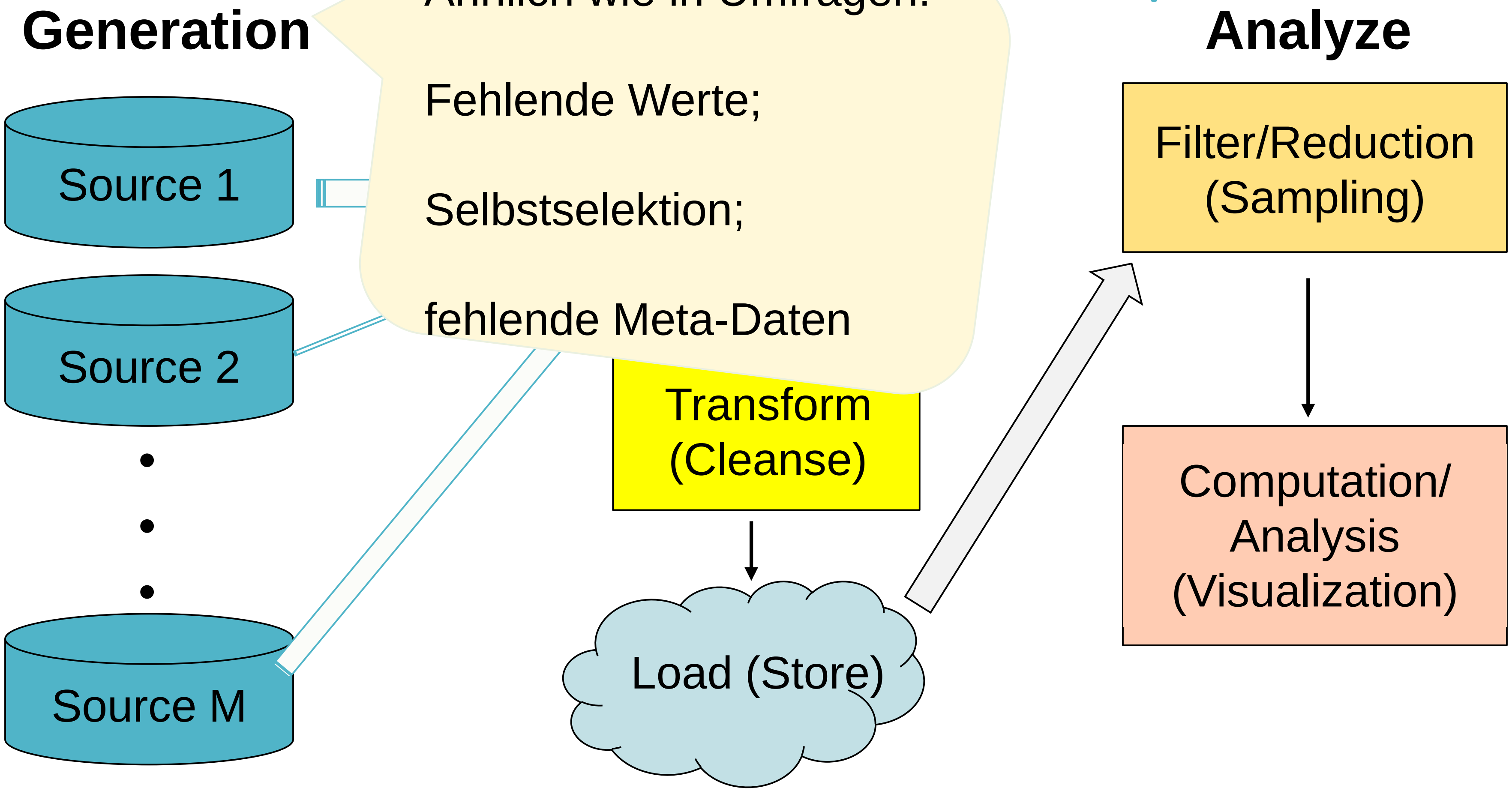
Groves et al. 2004

Big Data Process Map



BIEMER, P. (2017) ERRORS AND INFERENCE, CHAPTER 10 IN BD AND SOCIAL SCIENCE

Bi Process Map



Ähnlich wie in Umfragen:

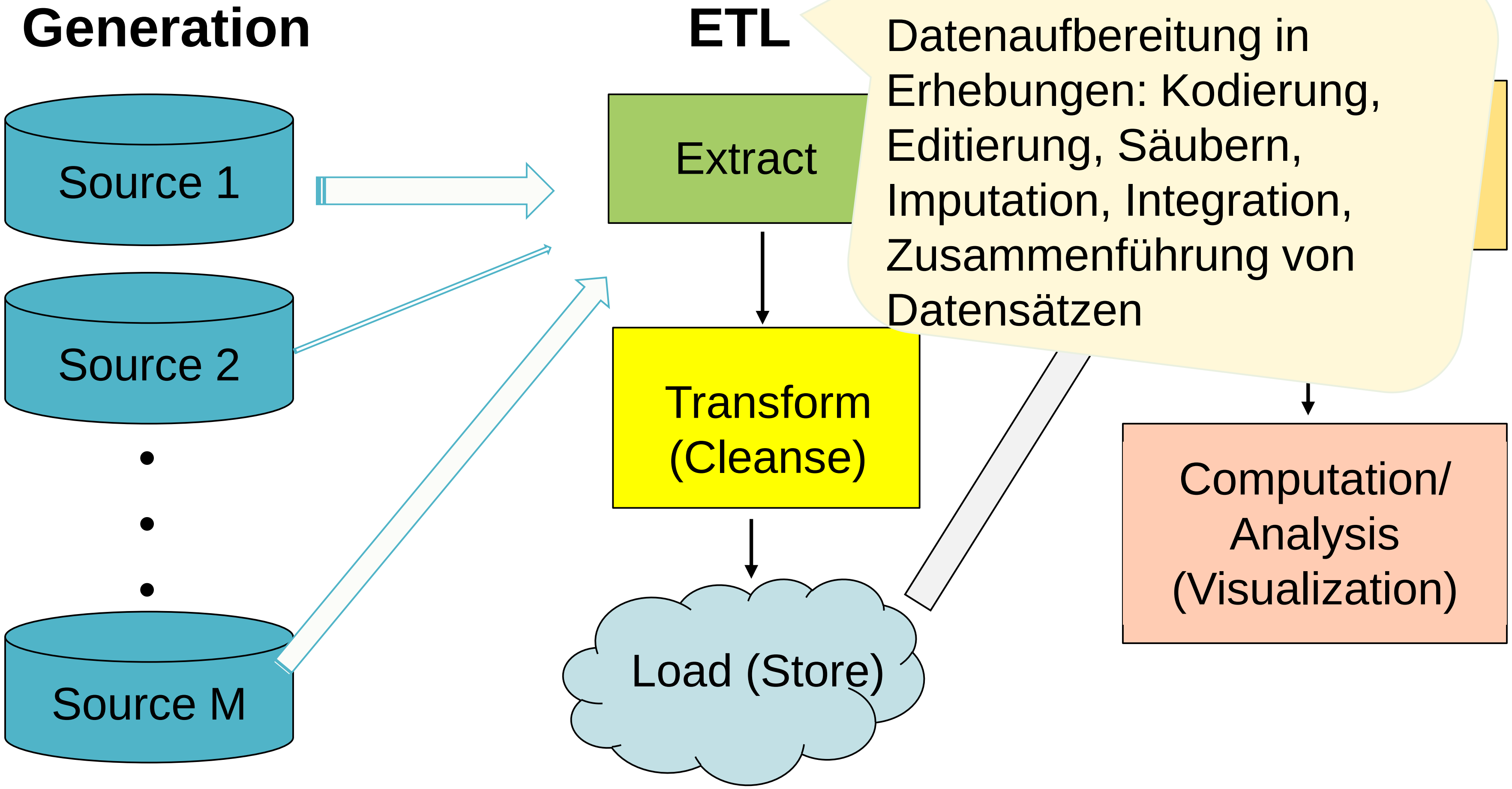
Fehlende Werte;

Selbstselektion;

fehlende Meta-Daten

BIEMER, P. (2017) ERRORS AND INFERENCE, CHAPTER 10 IN BD AND SOCIAL SCIENCE

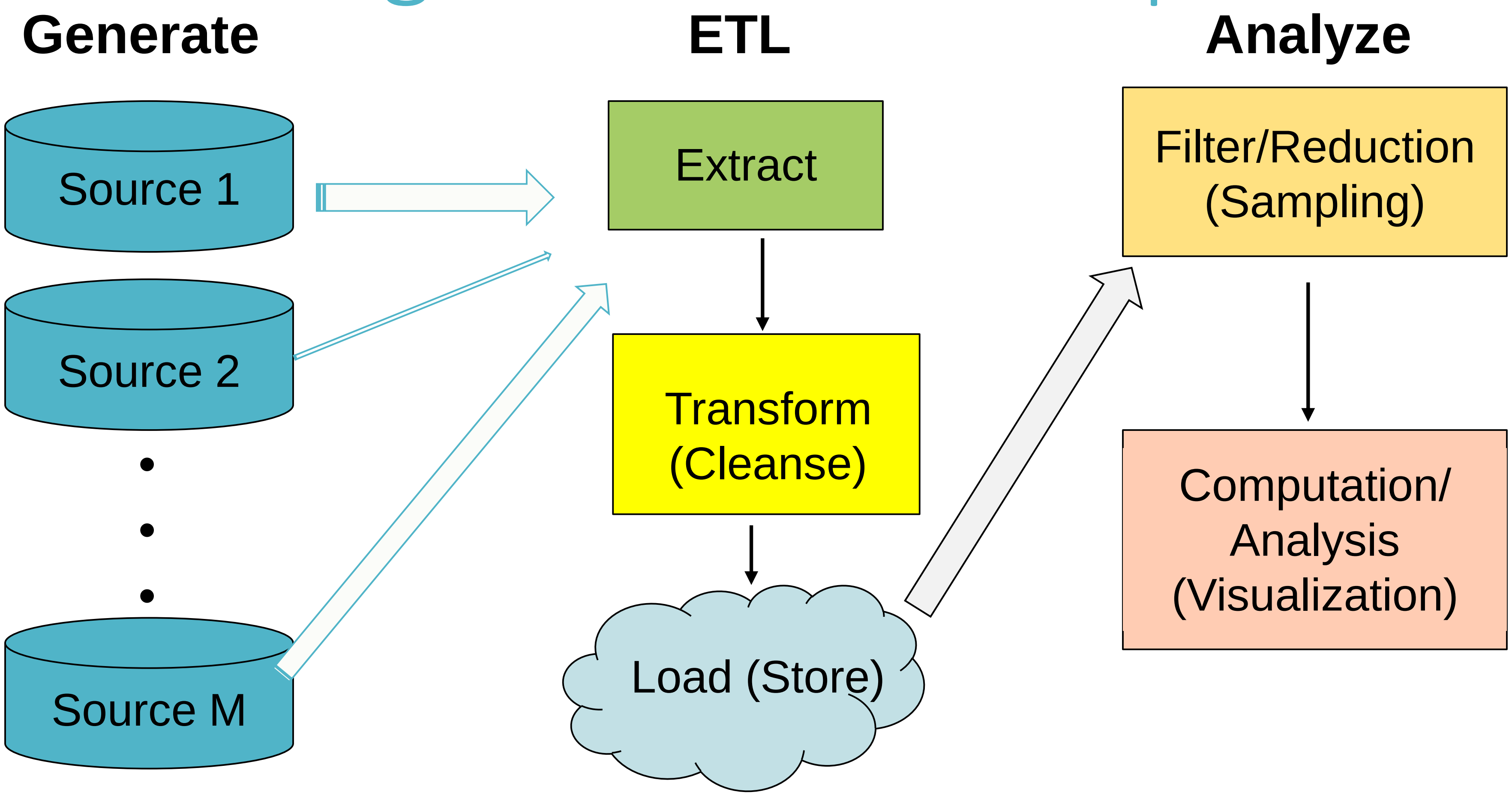
Big Data Proce



Ähnlich wie Datenaufbereitung in Erhebungen: Kodierung, Editierung, Säubern, Imputation, Integration, Zusammenführung von Datensätzen

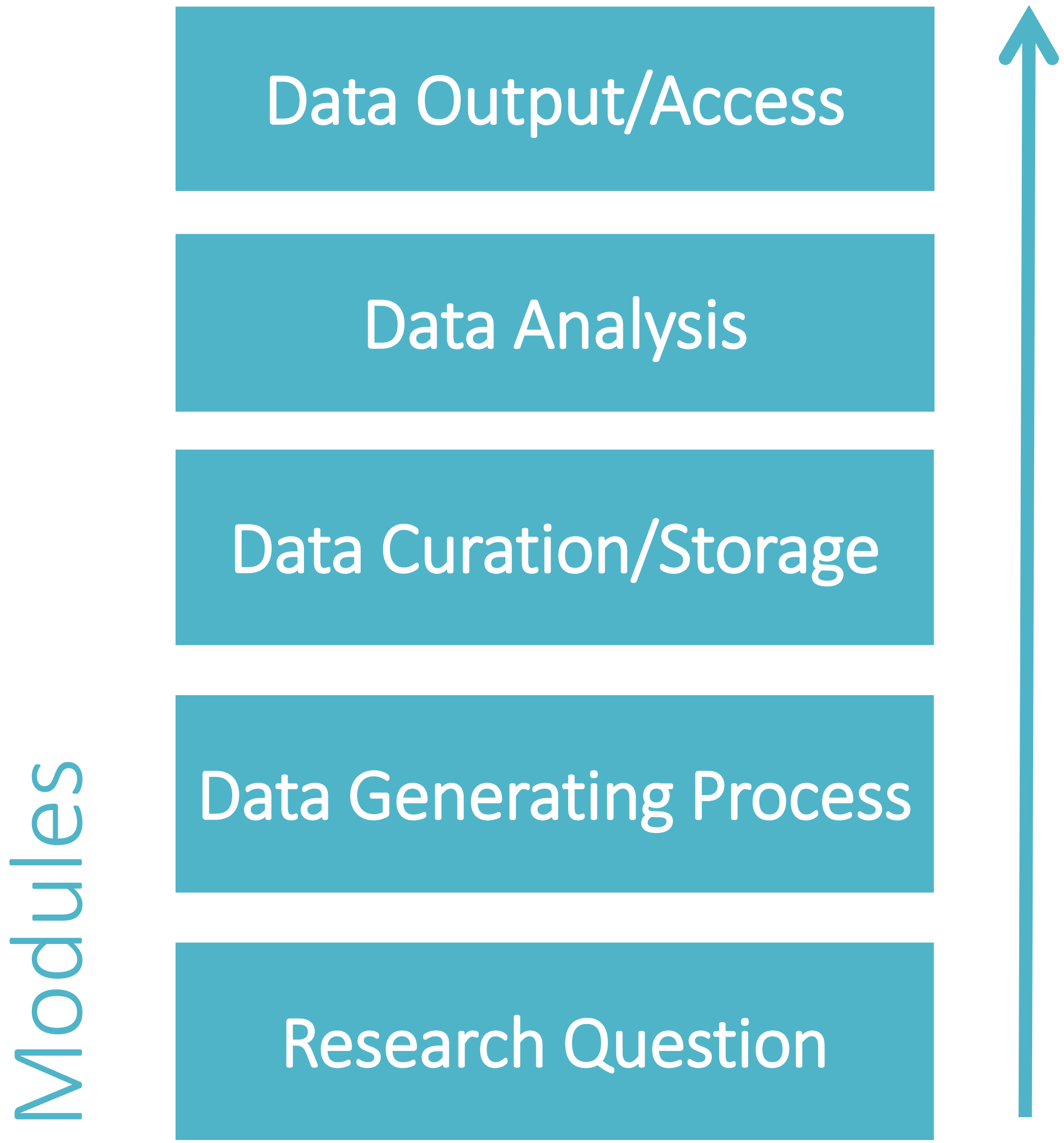
BIEMER, P. (2017) ERRORS AND INFERENCE, CHAPTER 10 IN BD AND SOCIAL SCIENCE

Big Data Process Map



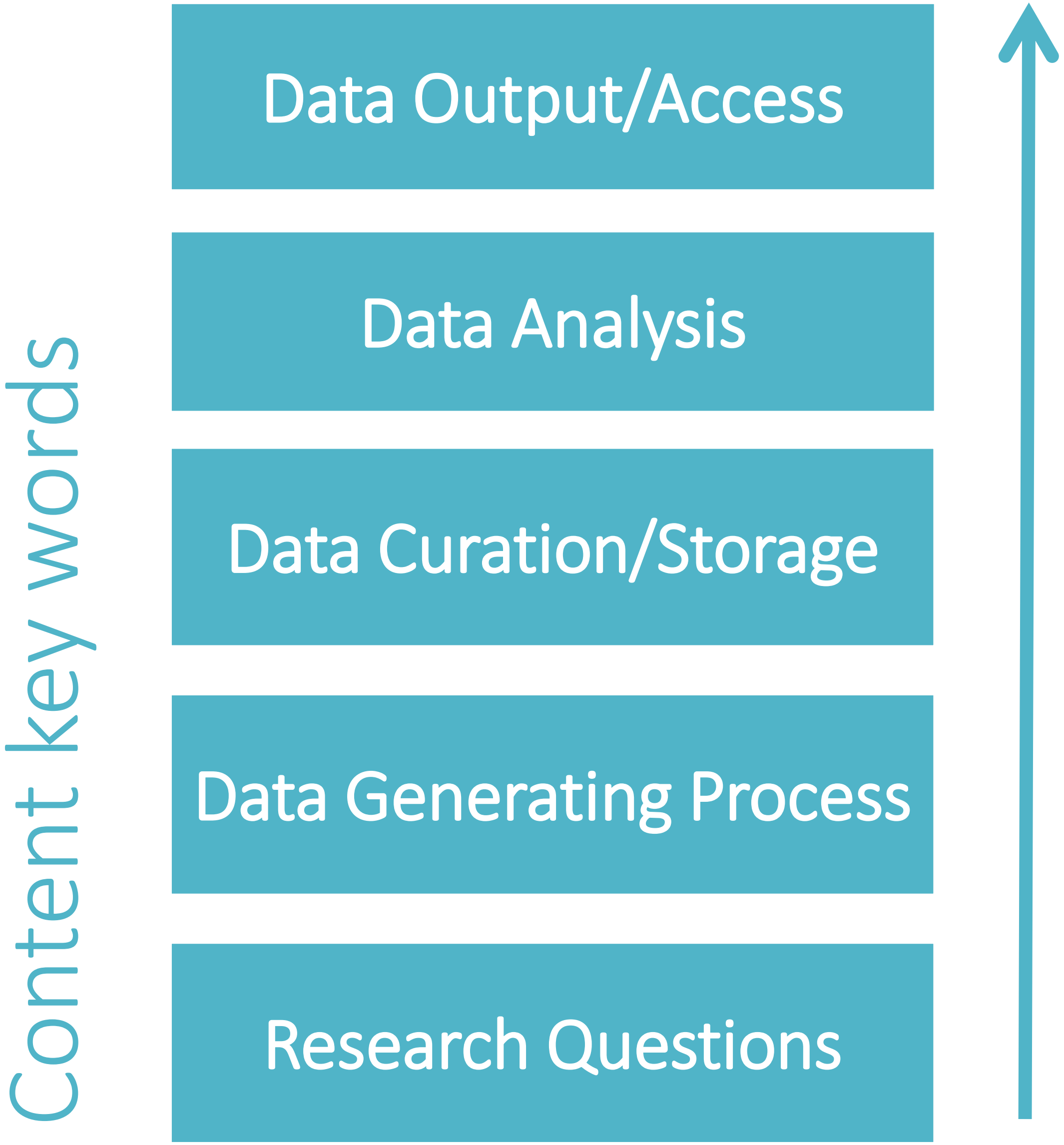
BIEMER, P. (2017) ERRORS AND INFERENCE, CHAPTER 10 IN BD AND SOCIAL SCIENCE

The Skills



Modules

Data Output/Access	Ergebnisse kommunizieren, visualisieren; Daten Weitergabe; Ethische Prinzipien
Data Analysis	Wissen welche Analysen für welche Daten-Typen in Frage kommen. Möglichkeiten und Grenzen.
Data Curation/Storage	Datenaufbereitung und Daten(bank)-Management
Data Generating Process	Einsicht in Daten generierende Prozesse (Transaktionen, Administratives, Web)
Research Question	Fragestellungen formulieren im Hinblick auf Datenerhebung, Analyse und Verarbeitung



Content key words

Data Output/Access

Visualization, disclosure control, ethics, privacy

Data Analysis

Statistical methods, machine learning, Bayesian, hierarchical, small area estimation

Data Curation/Storage

Practical training in data base management, SQL, editing, coding, imputation, etc.

Data Generating Process

Designed (survey and admin) and organic data (transaction and aspirational), linkage, matching

Research Questions

Economics, public policy, criminology, journalism, public health, sociology, etc.

The Program

Project coordinators and funding

JPSM
THE JOINT PROGRAM IN
SURVEY METHODOLOGY

IN OMNIBUS VERITAS
UNIVERSITÄT MANNHEIM

SPONSORED BY THE

Federal Ministry
of Education
and Research

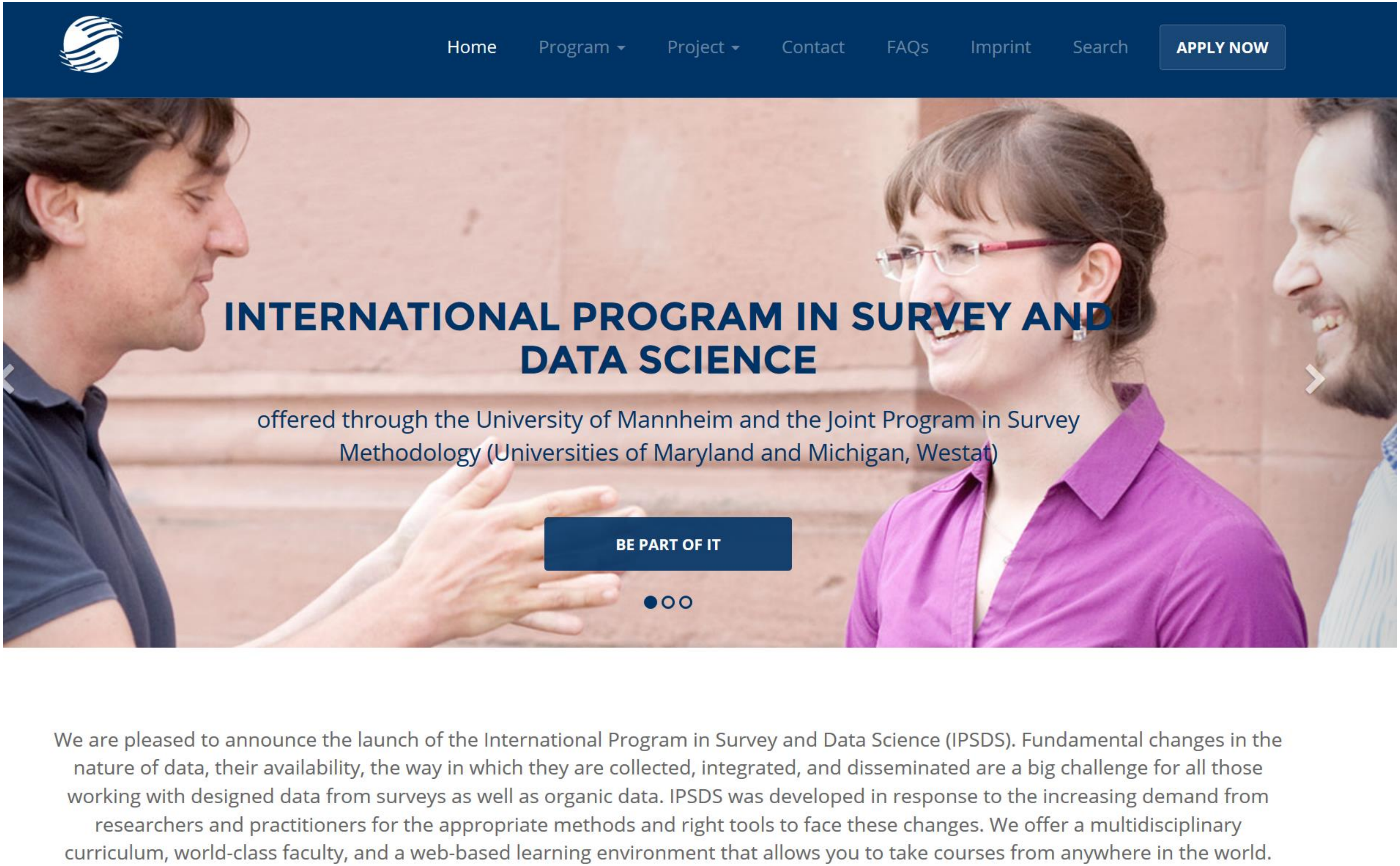
AUFSTIEG DURCH
BILDUNG >>
OFFENE HOCHSCHULEN

New program characteristics – In brief

- Multidisciplinary and modularized curriculum
- Relevant methods and tools
- Faculty from world-leading institutions

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

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.

Cooperation

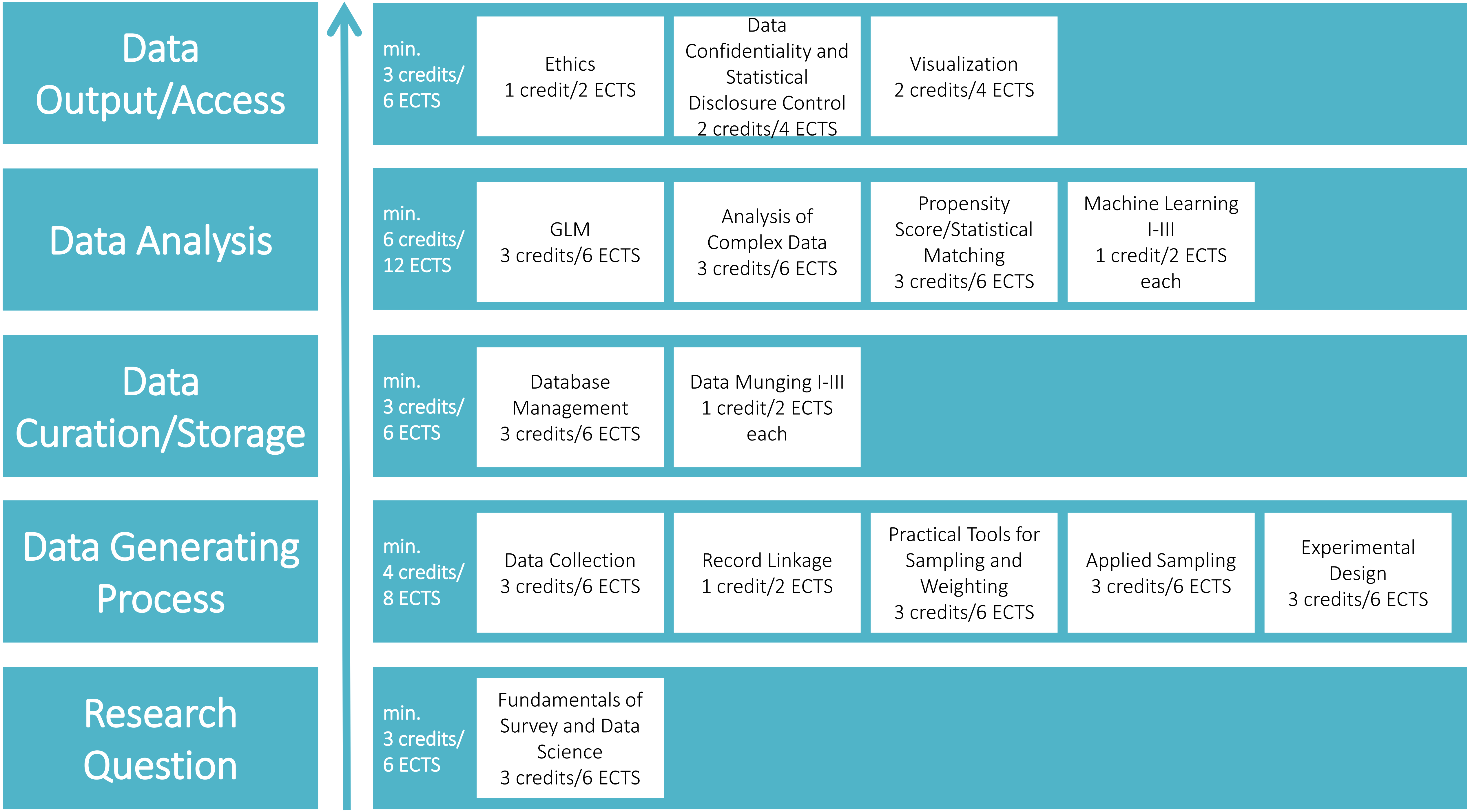
University Partners

- University of Maryland
- University of Michigan

- 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



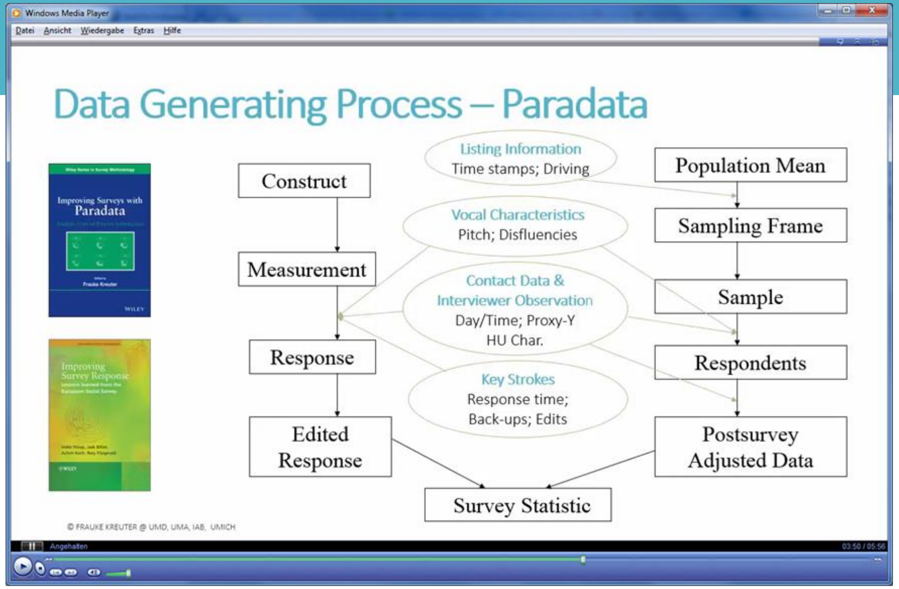
Data Output/Access	min. 3 credits/ 6 ECTS	Ethics 1 credit/2 ECTS	Data Confidentiality and Statistical Disclosure Control 2 credits/4 ECTS	Visualization 2 credits/4 ECTS		
Data Analysis	min. 6 credits/ 12 ECTS	GLM 3 credits/6 ECTS	Analysis of Complex Data 3 credits/6 ECTS	Propensity Score/Statistical 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 students to better focus on the material

Engage with the material at their own pace


Windows Media Player
Datei Ansicht Wiedergabe Extras Hilfe
Improving Surveys with Paradata
WILEY
Improving Survey Response
Angehalten
03:50 / 05:56

Data Generating Process – Paradata

Construct → Measurement → Response → Edited Response → Survey Statistic

Population Mean → Sampling Frame → Sample → Respondents → Postsurvey Adjusted Data → Survey Statistic

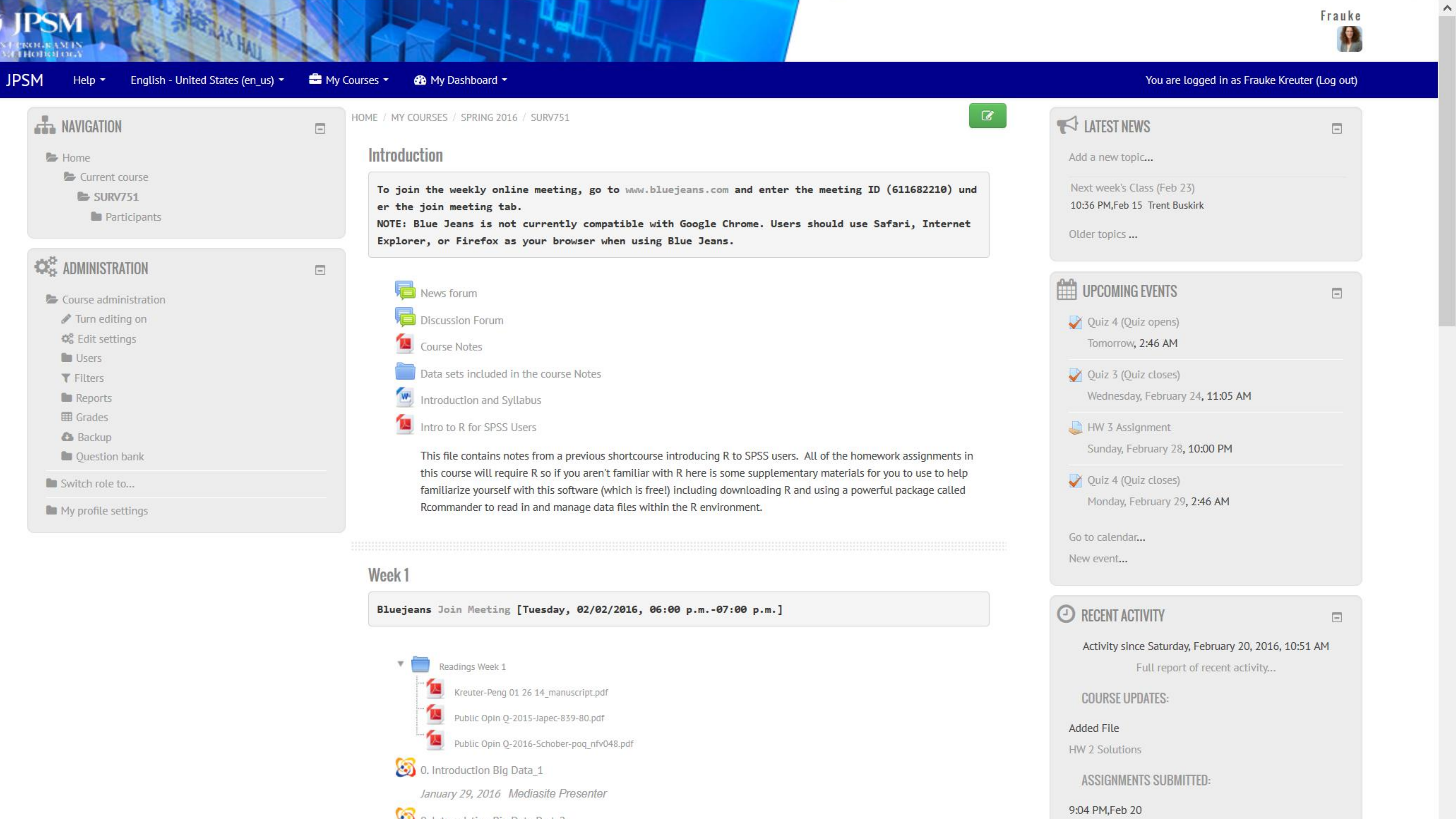
Listing Information
Time stamps; Driving

Vocal Characteristics
Pitch; Disfluencies

Contact Data & Interviewer Observation
Day/Time; Proxy-Y
HU Char.

Key Strokes
Response time;
Back-ups; Edits

© FRAUKE KREUTER @ UMD, UMA, IAB, UMICH



Introduction

To join the weekly online meeting, go to www.bluejeans.com and enter the meeting ID (611682210) under the join meeting tab.
NOTE: Blue Jeans is not currently compatible with Google Chrome. Users should use Safari, Internet Explorer, or Firefox as your browser when using Blue Jeans.

- News forum
- Discussion Forum
- Course Notes
- Data sets included in the course Notes
- Introduction and Syllabus
- Intro to R for SPSS Users

This file contains notes from a previous shortcourse introducing R to SPSS users. All of the homework assignments in this course will require R so if you aren't familiar with R here is some supplementary materials for you to use to help familiarize yourself with this software (which is free!) including downloading R and using a powerful package called Rcommander to read in and manage data files within the R environment.

Week 1

Bluejeans Join Meeting [Tuesday, 02/02/2016, 06:00 p.m.-07:00 p.m.]

- Readings Week 1
  - Kreuter-Peng 01 26 14_manuscript.pdf
  - Public Opin Q-2015-Japec-839-80.pdf
  - Public Opin Q-2016-Schober-poq_nfv048.pdf
- 0. Introduction Big Data_1

January 29, 2016 Mediasite Presenter

LATEST NEWS

Add a new topic...

Next week's Class (Feb 23)
10:36 PM,Feb 15 Trent Buskirk

Older topics ...

UPCOMING EVENTS

Quiz 4 (Quiz opens)
Tomorrow, 2:46 AM

Quiz 3 (Quiz closes)
Wednesday, February 24, 11:05 AM

HW 3 Assignment
Sunday, February 28, 10:00 PM

Quiz 4 (Quiz closes)
Monday, February 29, 2:46 AM

Go to calendar...
New event...

RECENT ACTIVITY

Activity since Saturday, February 20, 2016, 10:51 AM

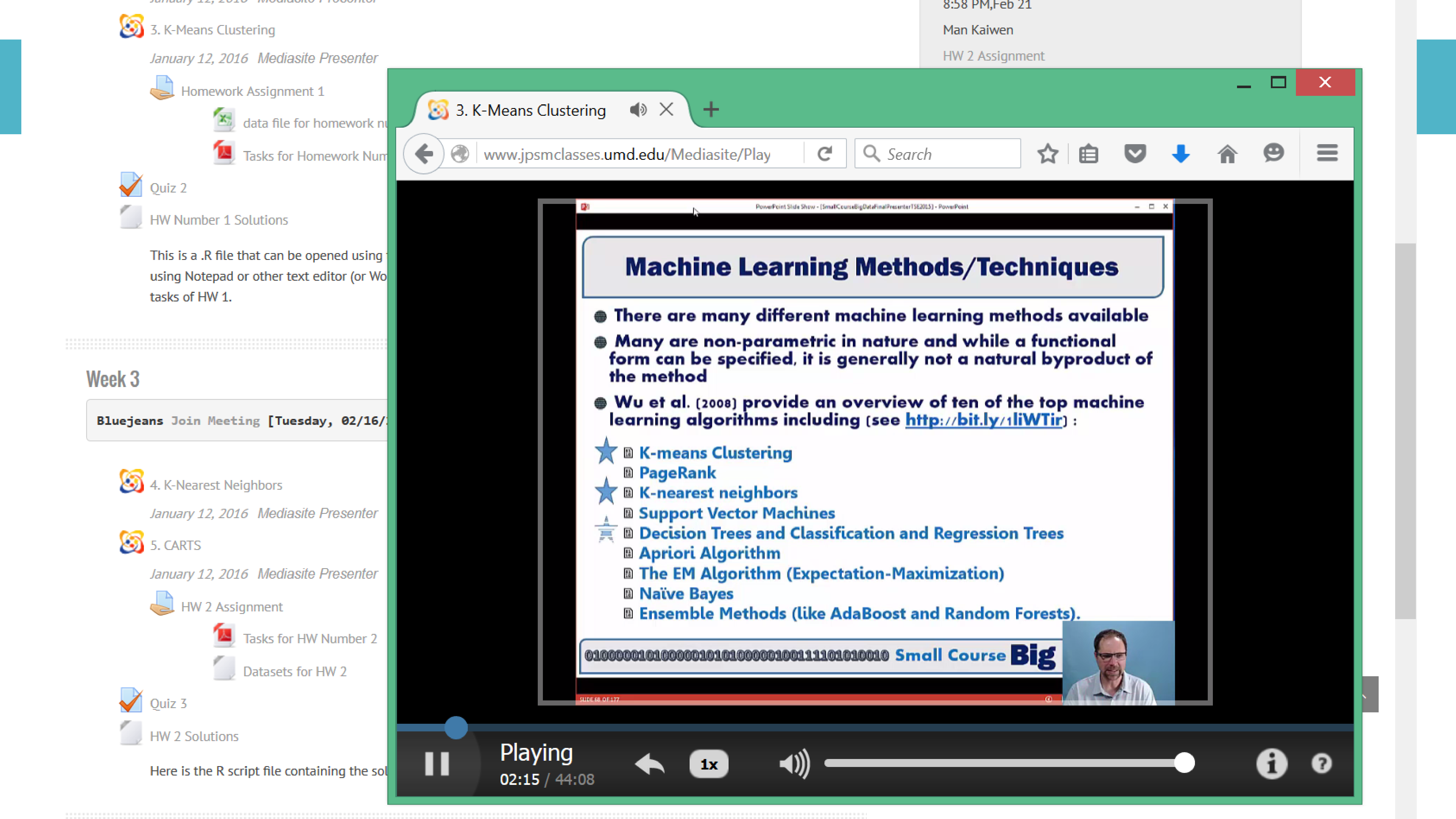
Full report of recent activity...

COURSE UPDATES:

Added File
HW 2 Solutions

ASSIGNMENTS SUBMITTED:

9:04 PM,Feb 20



January 12, 2016 Mediasite Presenter

3. K-Means Clustering

January 12, 2016 Mediasite Presenter

Homework Assignment 1

data file for homework nu

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 sol

8:58 PM,Feb 21

Man Kaiwen

HW 2 Assignment

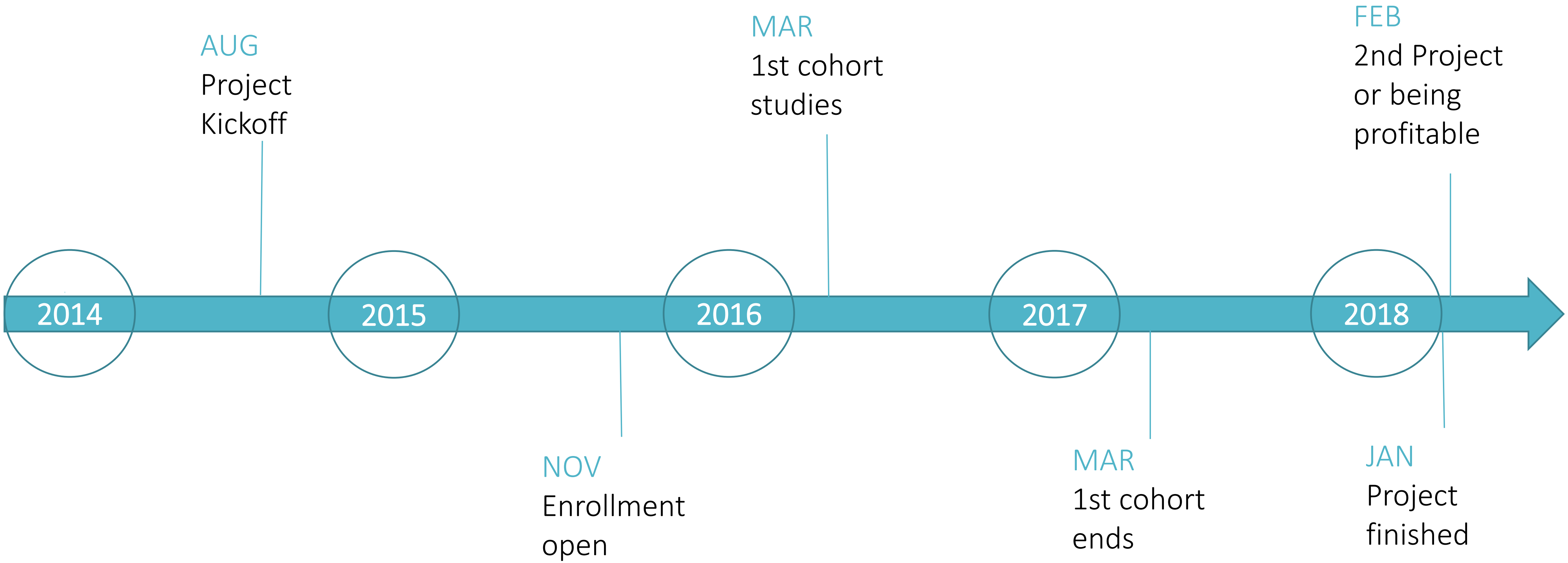
3. K-Means Clustering
www.jpsmclasses.umd.edu/Mediasite/Play
Search
PowerPoint Slide Show - [SmallCourseBigDataFinalPresenterTSE2015] - PowerPoint

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).

01000000101000000101010000001001111010100010 Small Course Big

Timeline – Test Phase at Mannheim



2014

AUG
Project
Kickoff

2015

NOV
Enrollment
open

2016

MAR
1st cohort
studies

2017

MAR
1st cohort
ends

2018

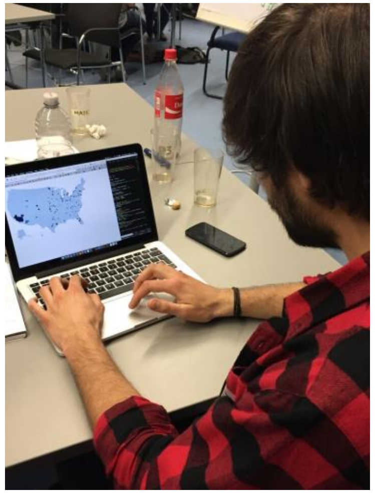
JAN
Project
finished

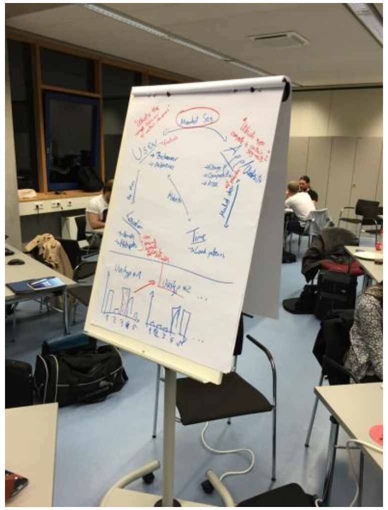
FEB
2nd Project
or being
profitable

Kick-off 2/20/2016

… recruitment and team work

DataFest





Interest

Demand for our students

RTI
INTERNATIONAL

SurveyMonkey

DEUTSCHE
BUNDESBANK
EUROSYSTEM

U.S. BUREAU OF LABOR STATISTICS

CDC
SAFER·HEALTHIER·PEOPLE
CDC 24/7
Saving Lives
Protecting People™

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

fkreuter@umd.edu