FCSM/WSS Workshop on Quality of Blended Data

26. Februar 2018

Summary
Lessons learned

Combining Data Sources
When assessing quality, we need to focus on
We need to get comfortable with proxies in Y and X
We need to remember the initial question
We need to change the way we operate
Lessons not yet learned

Combined data collection
Research Question – Effects of Unemployment
Research Question – Effects of Unemployment

Research-App, that …
  … issues questionnaires
  … collects passive data
  … links to panel survey and administrative data
PASS – Panel (10 years) + Administrative Data

Sample of households with at least one welfare benefit recipient (at reference date)
Refreshed annually
Surveyed annually

Random household sample of resident population
Refreshed annually
Surveyed annually

Pew Research estimates: 77% smart phone user in U.S. in 2016

Ownership by age groups (unweighted PASS estimates)
Predicting ownership and device type

Average Marginal Effect with 95% CIs

- Age in years
- Gender: Female
- Immigrant
- Higher education
- Welfare benefit recipient

Chart indicates the average marginal effect of each variable on device ownership, with 95% confidence intervals.
Lessons offered
Survey and Data Science
Understand how to collect data yourself, and how data are generated through administrative and other processes.

Learn how to formulate your research goal and which data are best suited to achieve this goal.

Learn how to communicate results and distribute and store your data.

Learn a variety of analysis methods suited for different data types.

Learn how to curate and manage data.

Source: Usher in Japec et al 2015
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)

BE PART OF IT

SPONSORED BY THE

Federal Ministry of Education and Research
<table>
<thead>
<tr>
<th>Master Thesis</th>
<th>Data Generating Process</th>
<th>Research Question</th>
<th>Data Curation/Storage</th>
<th>Data Analysis</th>
<th>Data Output/Access</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>min. 6 ECTS</td>
</tr>
<tr>
<td></td>
<td>Fundamentals of Survey and Data Science</td>
<td>Paper Writing / Publishing</td>
<td>Database Management I-III</td>
<td>Generalized Linear Models</td>
<td>Ethics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Analysis of Complex Data I-III</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Practical Tools for Sampling and Weighting</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Machine Learning I-II</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Experimental Design</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data Collection</td>
<td>Web Surveys</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>User Experience</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Questionnaire Design</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Applied Sampling I-II</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Data Collection</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Paper Writing / Publishing</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fundamentals of Survey and Data Science</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Record Linkage</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Multiple Imputation</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Python / SQL</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>User Experience</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Questionnaire Design</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Applied Sampling I-II</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Data Collection</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Paper Writing / Publishing</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fundamentals of Survey and Data Science</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Record Linkage</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Multiple Imputation</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Python / SQL</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>User Experience</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Questionnaire Design</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Applied Sampling I-II</td>
<td>1 credits/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Data Collection</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Paper Writing / Publishing</td>
<td>2 credits/4 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fundamentals of Survey and Data Science</td>
<td>3 credits/6 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Record Linkage</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Multiple Imputation</td>
<td>1 credit/2 ECTS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Python / SQL</td>
<td>1 credit/2 ECTS</td>
</tr>
</tbody>
</table>
Faculty

U. of Maryland / Michigan:
Chris Antoun
Fred Conrad
Steven Heeringa
Partha Lahiri
James Lepkowski
Richard Valliant

University of Mannheim:
Thomas Gautschi
Florian Keusch
Thomas Fetzer
Heiner Stuckenschmidt

Other universities:
Helmut Kuechenhoff (LMU Munich)
Daniel Oberski (Utrecht University)
Trent Buskirk (U. Mass, Boston)
Simon Munzert (HU Berlin)

Government Agencies:
Manfred Antoni (IAB)
Jörg Drechsler (IAB)
Joseph Sakshaug (IAB)
Stefan Bender (Bundesbank)

Jeffrey Gonzalez (BLS)
Carolina Franco (Census)

Private partners:
Mario Callegaro (Google)
Jennifer Romano-Bergstrom (Facebook)

Jill Dever (RTI)
Emily Geisen (RTI)
Raphael Nishimura (Abt)
Roger Tourangeau (Westat)
Onsite (Connect@IPSDS)

Online
Asynchronous

- Pre-recorded lectures (split into smaller video units)
- (Bi)weekly assignments
- Discussion forums

Synchronous

- Small virtual classrooms
- Weekly 50-minute discussions led by the instructor
- Obligatory component
Community is key

Coleridge Initiative
Networks: The first two classes brought together ~40 agencies from city, state, county and federal agencies
### Professional Training Workshops

#### Three Classes

- Different cohorts (ex-offenders, welfare recipients and veterans)
- Joined with housing, transportation and jobs data

#### Class Format

- Module 1: Foundations – Research Questions, Python, SQL
- Module 2: Data Acquisition – Web Scraping, API, Record Linkage
- Module 3: Data Analysis – Machine Learning, Networks, Text, Spatial
- Module 4: Visualization, Inference, Ethics, Privacy

#### Additional Information

- Final reports are all virtual
- Teaching Assistants and facilitators will be at each site for each module
Collaborative secure environment

- Data Discovery
- Database Browser
- Software Version Control
- JupyterHub (Data Analysis)

DATA FROM AGENCIES
- Illinois Department of Human Services (IDHS): documentation link
- Illinois Department of Employment Services (IDES)
  - Individual wage records: documentation link
  - Employer data: documentation link
- Illinois Department of Corrections (IDOC) admissions: link to pdf
- IDOC exits: link to pdf
  - More details on EDUCLVL and HCLASS fields:
    - link to pdf of notes
    - link to Excel of Offense Codes
  - BIRTHPL lookup file
- HUD program data: link to pdf (note descriptions are of unaggregated data and not all fields included due to data use agreements)
- Parole data dictionary: link to Sheet
Big Data for Federal Agencies
- Fall course: 25 students
- curriculum = book outline

Outlook
- one-stop enrollment
- engagement of PI/PR
DOMAIN EXPERT
User, analyst, or leaders with deep subject matter expertise related to the data, its appropriate use, and its limitations

SYS ADMIN
Team member responsible for defining and maintaining a computation infrastructure that enables large scale computation

RESEARCHER
Team member with experience applying formal research methods, including survey methodology and statistics

COMPUTER SCIENTIST
Technically skilled team member with education in computer programming and data processing technology

Source: Abe Usher
Shift in mindset! Dare to experiment (now)!

Thank you!

Frauke Kreuter (fkreuter@umd.edu)

survey-data-science.net
coleridgeinitiative.org