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:

Lisa Terry, Co-Chair, Statistics Sweden
Frank Levenstein, Co-Chair, EPIC at the U. of Maryland, U. of Memphis & LBA
Martin Berg, Stockholm University
Paul Bienert, IIT International
Paul Becker, Mathematics Policy Research
Cliff Lampe, School of Information at the University of Michigan
John Reira, American Institute for Research
Colin O’Neil, Data Research Labs
Abe Usher, HumanGeo Group

Acknowledgment: We are grateful for comments, feedback and editorial help from Lura Ben-Porath, Jason McMillan, and the AAPOR council members.
Economic Indicators

US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI the PriceStats website. 1/1/2015
Data

Designed
- Experiment
- Survey

Organic
- Administrative
- Aspirational
- Transactional

Source: Roberto Rigobon
1\textsuperscript{st} Example – Coleridge Initiative
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
Networks: The first two classes brought together ~40 agencies from city, state, county and federal agencies
Approach

Data on ex-offenders, welfare recipients

Data on housing and transportation

Joined Up Datasets

Trained Staff

New Products

New Networks
Big Data for Federal Agencies
- Fall course: 25 students
- curriculum = book outline

Outlook
- one stop enrollment
- engagement of PI/PR
Collaborative secure environment

- Data Discovery
- Database Browser
- OpenOffice: Writer
- OpenOffice: Calc
- OpenOffice: Impress
- Source Code Editors
- Command Line
- GitLab
- JupyterHub (Data Analysis)

Applications:
- Data Discovery
- Database Browser
- OpenOffice: Writer
- OpenOffice: Calc
- OpenOffice: Impress
- Source Code Editors
- Command Line
- GitLab
- JupyterHub (Data Analysis)
2nd Example – International Program in Survey and Data Science
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
Partners and Funding

University Partners
- University of Maryland
- University of Mannheim
- Catholic University of Santiago de Chile
- Australian National University
- Beijing University
- NCAER -- India
- U. of Capetown (planned)

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

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 contents of this publication lies with the author.
Understand how to collect data yourself, and how data are generated through administrative and processes.

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

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

Learn a variety of analysis methods suited for different data types

Learn how to curate and manage data
<table>
<thead>
<tr>
<th>Data Generating Process</th>
<th>Data Curation/Storage</th>
<th>Data Analysis</th>
<th>Data Output/Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>min. 4 credits/8 ECTS</td>
<td>min. 3 credits/6 ECTS</td>
<td>min. 6 credits/12 ECTS</td>
<td>min. 3 credits/6 ECTS</td>
</tr>
<tr>
<td>Data Collection 3 credits/6 ECTS</td>
<td>Database Management 3 credits/6 ECTS</td>
<td>GLM 3 credits/6 ECTS</td>
<td>Ethics 1 credit/2 ECTS</td>
</tr>
<tr>
<td>Record Linkage 1 credit/2 ECTS</td>
<td>Analysis of Complex Data 3 credits/6 ECTS</td>
<td>Analysis of Complex Data 3 credits/6 ECTS</td>
<td>Data Confidentiality and Statistical Disclosure Control 2 credits/4 ECTS</td>
</tr>
<tr>
<td>Practical Tools for Sampling and Weighting 3 credits/6 ECTS</td>
<td>Propensity Score/Statistical Matching 3 credits/6 ECTS</td>
<td>Propensity Score/Statistical Matching 3 credits/6 ECTS</td>
<td>Machine Learning I-III 1 credit/2 ECTS each</td>
</tr>
<tr>
<td>Applied Sampling 3 credits/6 ECTS</td>
<td>Experimental Design 3 credits/6 ECTS</td>
<td>Machine Learning I-III 1 credit/2 ECTS each</td>
<td>Visualization 2 credits/4 ECTS</td>
</tr>
<tr>
<td>Experimental Design 3 credits/6 ECTS</td>
<td>Fundamentals of Survey and Data Science 3 credits/6 ECTS</td>
<td>Fundamentals of Survey and Data Science 3 credits/6 ECTS</td>
<td>Visualize 2 credits/4 ECTS</td>
</tr>
<tr>
<td>min. 3 credits/6 ECTS</td>
<td>min. 3 credits/6 ECTS</td>
<td>min. 6 credits/12 ECTS</td>
<td>min. 3 credits/6 ECTS</td>
</tr>
<tr>
<td>Data Collection 3 credits/6 ECTS</td>
<td>Database Management 3 credits/6 ECTS</td>
<td>GLM 3 credits/6 ECTS</td>
<td>Ethics 1 credit/2 ECTS</td>
</tr>
<tr>
<td>Record Linkage 1 credit/2 ECTS</td>
<td>Analysis of Complex Data 3 credits/6 ECTS</td>
<td>Analysis of Complex Data 3 credits/6 ECTS</td>
<td>Data Confidentiality and Statistical Disclosure Control 2 credits/4 ECTS</td>
</tr>
<tr>
<td>Practical Tools for Sampling and Weighting 3 credits/6 ECTS</td>
<td>Propensity Score/Statistical Matching 3 credits/6 ECTS</td>
<td>Propensity Score/Statistical Matching 3 credits/6 ECTS</td>
<td>Machine Learning I-III 1 credit/2 ECTS each</td>
</tr>
<tr>
<td>Applied Sampling 3 credits/6 ECTS</td>
<td>Experimental Design 3 credits/6 ECTS</td>
<td>Experimental Design 3 credits/6 ECTS</td>
<td>Visualization 2 credits/4 ECTS</td>
</tr>
</tbody>
</table>
Research: How to interact?

Synchronous

• Reduces social isolation
• Questions answered right away

Asynchronous interaction

• Flexibility
• More time to think
(Hrastinski & Keller, 2007, p. 66)
Research: How to structure the material?

- Video lectures
- Lectures + Interviews

- Expert interviews can show practical application
  (Renninger & List, 2012)
Research: How much flexibility?

Instructor-paced:
11 week
Deadlines for homework

Self-paced:
8 week
No deadlines – all due at the end
Lessons Learning

- Modular approach much appreciated by working professionals
- Learning with application at hand is key
- Participants from all sectors and disciplines
- Very high demand on graduates
- Privacy and confidentiality very important

- Hardest to learn and hardest to teach: Asking the right question!
http://coleridgeinitiative.org
http://survey-data-science.net/
fkreuter@umd.edu