Data sciences as essential job skill!

A social science, economics, and public policy training perspective

Frauke Kreuter

JPSM – Uni Mannheim – IAB

Paderborn 11/12/17
Data

Designed

Experiment

Survey

Administrative

Organic

Aspirational

Transactional

Source: Roberto Rigobon
Social Science

Marienthal study:
• social inclusion,
• coping behavior,
• day structure of the unemployed

MoDeM @ IAB

Sensor data replace time intensive observations in original study
• Radius of action
• Walking speed, sports activities
• Social networks
• Media use

Linkage with survey and admin data
Full-time employed

- App use past 5pm

Part-time employed

- App use at noon

Job seekers

- Continuous app use
Economic Indicators

US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI (the PriceStats website) - 1/1/2015
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).
Public Policy

11 million people move through 3,100 Jails

$22 Billion in costs

- 64% suffer from mental illness,
- 68% have a substance abuse disorder
- 44% suffer from chronic health problems
Hope: Combined data and predictive systems can support **targeted, preventative interventions** to help people at **risk of interactions** with the criminal justice system.

Of the top 200 predicted individuals:
- **104** went to jail within 1 year
- **19 years** total jail time
Skills
DOMAIN EXPERT
User, analyst, or leaders with deep subject matter expertise related to the data, its appropriate use, and its limitations

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

SYS ADMIN
Team member responsible for defining and maintaining a computation infrastructure that enables large scale computation
1st 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
Approach

Taught using **hands-on projects** with **real microdata** in a **secure environment** so that participants can learn the basics of how to:

- Code and collect new data
- Work with spatial data
- Manage complex data,
- Apply machine learning, text and network analysis
- Visualize relationships
- Address inference issues
- Manage privacy and confidentiality
Problem

In this phase, you take the predictors from your test set and apply your model to them, then assess the quality of the model by comparing the predicted values to the actual values for each record in your testing data set.

- **Performance Estimation**: How well will our model do once it is deployed and applied to new data?

Now let’s use the model we just fit to make predictions on our test dataset, and see what our accuracy score is:

Python’s `scikit-learn` is a commonly used, well documented Python library for machine learning. This library can help you split your data into training and test sets, fit models and use them to predict results on new data, and evaluate your results.

We will start with the simplest `LogisticRegression` model and see how well that does.

You can use any number of metrics to judge your models (see `model evaluation`), but we’ll use `accuracy_score()` (ratio of correct predictions to total number of predictions) as our measure.

```python
# Let's fit a model
from sklearn import linear_model
model = linear_model.LogisticRegression(penalty='l1', C=1e5)
model.fit(X_train, y_train)
print(model)
```
Isochrone API

- Back to the [Table of Contents](#)

Gives the area (as a polygon) a traveler can reach from a
the user can set if so desired, [description here](#)

```python
# set start location
start_point = [40.693856, -73.985754]  # Doc

# set travel time - in seconds
travel_time = 60*20  # 20 minutes

# create query URL
qry_url = '{isochrone?fromPlace={},{},&mode[0],
vel_time)

# get json request
iso_json = json.loads(requests.get(qry_url)

## load isochrone into a geopandas dataframe
iso_gdf = gpd.GeoDataFrame.from_features(iso_json['features'])```
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

GitHub

JupyterHub (Data Analysis)
What our participants say about the program

"Love the Jupyter notebooks!! ... I love how the code snippets and explanations are set up in the Jupyter notebooks. The format of going through it individually and discussing questions/challenges in your group, with the experts available when needed, worked really well for my learning style."

I could see our agency benefiting potentially from something like this in that, as the system builds out and collects additional resources/datasets that impact criminal justice system practices, this may be an option for a place for us to look for the results of studies using evidence based practices.

Danielle Fulmer
Director of Business Analytics

Katy Fitzgerald
Management Analyst
"Thanks for your email. I will be away from the office starting on April 18th and returning Monday, April 24th.

Where did I go? University of Maryland - College Park, Maryland
Why so long? I am at the most awesome, intense, hands-on training for using data to improve public policy.
Why? Noemi Reyes and I were accepted into the program – AND received a very generous scholarship from the Laura and John Arnold Foundation.
What can I do? If you have data you would like to contribute, send me an email with the Subject: Dane Data – that will get the attention of the little robot I set up to prioritize that message! We can set up a time to chat when I get back."
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
- 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
<table>
<thead>
<tr>
<th>Modules</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Generating</td>
<td>Understand how to collect data yourself, and how data are generated</td>
</tr>
<tr>
<td>Data Generating</td>
<td>through administrative and processes.</td>
</tr>
<tr>
<td>Data Curation/Storage</td>
<td>Learn how to curate and manage data</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>Learn a variety of analysis methods suited for different data types</td>
</tr>
<tr>
<td>Data Output/Access</td>
<td>Learn how to communicate results, distribute and store your data; Ethics</td>
</tr>
<tr>
<td>Research Question</td>
<td>Learn how to formulate your research goal and which data are best</td>
</tr>
<tr>
<td></td>
<td>suited to achieve this goal.</td>
</tr>
</tbody>
</table>
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
Annual „Connect“ Event (next June 9, 2018)
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!