TOOLS AND BEST PRACTICES IN QUANTITATIVE RESEARCH

Mercè Crosas, IQSS, Harvard University
@mercecrosas

XXXVII JORNADAS DE ECONOMÍA DE LA SALUD,
BARCELONA, 6-8 SEPTIEMBRE
SOME EXAMPLES
OF WHAT YOU CAN DO USING
QUANTITATIVE DATA AND METHODS

- quantitative and geospatial data
- unstructured text as data
Imagine that you have data for all the deaths of all Medicare beneficiaries in the US 2000-2012 (~half a million person-years) and want to model the effect of air pollution levels on death, controlling for other factors that also affect death (such as smoking, BMI).
Concludes that levels of PM2.5 below the current standard are still harmful
WHAT IS USED TO COMPUTE?

- Medicare data:
  - 5 TB
  - Privacy requirements
- Air pollution grids
  - 50 TB
- Statistical model from survival analysis (Cox proportional hazard, R package)
- WorldMap (GIS) for geospatial visualizations
- Massive computations performed on a secure cluster (1.3 years of combined runtime, ~700 CPU's, 24TB of memory, on IQSS Research Computing Environment)
FROM TEXT TO QUANTITATIVE DATA

Documents

Vector-space representation

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
</tr>
</thead>
<tbody>
<tr>
<td>complexity</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>algorithm</td>
<td>3</td>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>entropy</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>traffic</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>network</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Term-document matrix

source: https://www.linkedin.com/pulse/text-mining-its-applications-industry-subhajit-mukherjee
**Consilience**: a tool that enables you to quickly read, understand, categorize, and derive insights from large quantities of unstructured text.
WHAT IS USED TO COMPUTE?

- On Microsoft Azure cloud
- Using Java, Scala, Python, R
- MongoDB (document database)
- Apache Spark (large-scale data processing engine)
- Elasticsearch (search and analytics engine)
- Machine Learning Library (MLlib) for Spark
- 6 worker VMs with 112 GB memory each (total 672 GB), 16 cores each (total 64 cores)
SKILLS, LANGUAGES, AND TOOLS
USING QUANTITATIVE METHODS IN TODAY'S DATA-INTENSIVE RESEARCH

- Computer Science
- Software Engineering
- Statistics
- Machine Learning
- Domain Expertise
# Program comparison

<table>
<thead>
<tr>
<th>Program</th>
<th>Statistics</th>
<th>Visualization</th>
<th>Machine learning</th>
<th>Ease of use</th>
<th>Power/flexibility</th>
<th>Fun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stata</td>
<td>Good</td>
<td>Serviceable</td>
<td>Limited</td>
<td>Very easy</td>
<td>Low</td>
<td>Some</td>
</tr>
<tr>
<td>SPSS</td>
<td>OK</td>
<td>Serviceable</td>
<td>Limited</td>
<td>Easy</td>
<td>Low</td>
<td>None</td>
</tr>
<tr>
<td>SAS</td>
<td>Good</td>
<td>Not great</td>
<td>Good</td>
<td>Moderate</td>
<td>Moderate</td>
<td>None</td>
</tr>
<tr>
<td>Matlab</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Moderate</td>
<td>Good</td>
<td>Some</td>
</tr>
<tr>
<td>R</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Good</td>
<td>Moderate</td>
<td>Excellent</td>
<td>Yes</td>
</tr>
<tr>
<td>Python</td>
<td>Good</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Moderate</td>
<td>Excellent</td>
<td>Yes</td>
</tr>
<tr>
<td>Julia</td>
<td>OK</td>
<td>Excellent</td>
<td>Good</td>
<td>Hard</td>
<td>Excellent</td>
<td>Yes</td>
</tr>
</tbody>
</table>

# Language specific editors and IDEs

<table>
<thead>
<tr>
<th>Editor</th>
<th>Features</th>
<th>Ease of use</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>RStudio</td>
<td>Excellent</td>
<td>Easy</td>
<td>R</td>
</tr>
<tr>
<td>Spyder</td>
<td>Excellent</td>
<td>Easy</td>
<td>Python</td>
</tr>
<tr>
<td>Stata do file editor</td>
<td>OK</td>
<td>Easy</td>
<td>Stata</td>
</tr>
<tr>
<td>SPSS syntax editor</td>
<td>OK</td>
<td>Easy</td>
<td>SPSS</td>
</tr>
</tbody>
</table>

POPULARITY OF TOOLS/LANGUAGES

BASED ON JOB POSTINGS

Top: SQL, Python, Java, Hadoop, R

BASED ON SCHOLARLY ARTICLES

Top: SPSS, R, SAS, Stata

Robert Muenchen: http://r4stats.com/articles/popularity
R VS PYTHON BASED ON JOB POSTINGS

Popularity of SPSS, Stata, SAS is decreasing compared to Python, R, and Julia.
R Packages developed at Harvard's Institute for Quantitative Social Science (IQSS)
zeligproject.org
BEST PRACTICES
"Convert the raw results of any specific statistical procedure into expressions that:
1) convey numerically precise measurements of the quantities of greatest substantive interest,
2) include reasonable measurements of uncertainties about those estimates,
3) require little specialized knowledge to understand"
Enhancing reproducibility for computational methods

Victoria Stodden¹, Marcia McNutt², David H. Bailey³, Ewa Deelman⁴, Yolanda Git⁴, Brooks Hanson⁵, Michael...

+ See all authors and affiliations

Science 09 Dec 2016:
Vol. 354, Issue 6317, pp. 1240-1241
DOI: 10.1126/science.aah6168
BEST PRACTICES FOR REPRODUCIBILITY

• Share data and code in open trusted repositories
• Use persistent links from publication to data and code
• Citation to data and code should be a standard
• Document data, code, workflows, and computational environment
• Use open license for your code and data
WHEN WRITING METHODS, YOUR CODE SHOULD BE:

1. Informatively documented
2. Open source
3. Comprehensible and automatically tested
4. Developed using version control
5. Stored in a public repository
6. Clearly citable

Christopher Gandrud, from IQSS Software Best Practices workshop
Thanks

@mercecrosas

Harvard's Institute for Quantitative Social Science

@IQSS

iq.harvard.edu