Revamping Student Education with Real-World Data

Prof. Brian D. Davison
January 2016

Glassdoor Ranks the 25 Best Jobs in America

Of no surprise to us, “Data scientist” ranked #1 on Glassdoor’s list of 25 jobs with the best work-life balance last October. (The Washing
Universities are educating more and more data-savvy students

• MS degrees in data science…
• BS degrees…
• Every semester I teach a popular *Introduction to Data Science* course to students from any discipline

• In all of these environments, students need to apply methods to data sets to learn, gain experience, and test intuitions
One machine learning training website says:

### Why Do We Need Practice Datasets?

If you are interested in practicing applied machine learning, you need datasets on which to practice.

This problem can stop you dead.

- Which dataset should you use?
- Should you collect your own or use one off the shelf?
- Which one and why?

For beginners, you can get everything you need and more in terms of datasets to practice on from the UCI Machine Learning Repository.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data Point</th>
<th>Attribute</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>9</td>
<td>6*</td>
</tr>
<tr>
<td>WDBC</td>
<td>569</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>CMC</td>
<td>1473</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Yeast</td>
<td>1484</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Optical Digit</td>
<td>5620</td>
<td>64</td>
<td>10</td>
</tr>
<tr>
<td>Statlog</td>
<td>6435</td>
<td>36</td>
<td>6*</td>
</tr>
<tr>
<td>Thyroid</td>
<td>7200</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Magic Gamma</td>
<td>19020</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>
Why share real-world data with students?

• Improve student education!
  • Modern real-world problems are great motivators for students
  • Help them understand complex and dirty real-world data

• Educate students about your data science problems
  • Industry needs are often different from academic research
  • They will know what you care about, and can jump in when hired

• Increase visibility to students
  • Your brand becomes part of otherwise inaccessible conversations

• Indirectly educate students about your organization and needs
  • The best data science requires an understanding of the domain
<table>
<thead>
<tr>
<th>Competition Description</th>
<th>Prize</th>
<th>Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018 Data Science Bowl</td>
<td>$100,000</td>
<td>3,634</td>
</tr>
<tr>
<td>TalkingData AdTracking Fraud Detection Challenge</td>
<td>$25,000</td>
<td>2,670</td>
</tr>
<tr>
<td>CVPR 2018 WAD Video Segmentation Challenge</td>
<td>$2,500</td>
<td>15</td>
</tr>
<tr>
<td>iMaterialist Challenge (Fashion) at FGVC5</td>
<td>$2,500</td>
<td>28</td>
</tr>
<tr>
<td>iMaterialist Challenge (Furniture) at FGVC5</td>
<td>$2,500</td>
<td>177</td>
</tr>
<tr>
<td>Google Landmark Retrieval Challenge</td>
<td>$2,500</td>
<td>113</td>
</tr>
<tr>
<td>Google Landmark Recognition Challenge</td>
<td>$2,500</td>
<td>231</td>
</tr>
</tbody>
</table>
Why not share data?

• Your organization already has all the answers and staff it needs
• There is already lots of data out there
  • True! And it is growing! But process data is still pretty rare.
• Your competitor might learn something from your data
  • Not all data needs to expose intellectual property (e.g., pump runs)
• Concerns about privacy
Privacy concerns are real
How to share data?

• Publicly through competitions
• Publicly via your website
• Publicly via data repositories or network
• Privately with a usage license
  • E.g., only educational or non-profit use
  • Perhaps with a data broker to enforce
• Privately with a non-disclosure agreement
If you really want to impact students...

- Don’t just share data on the web (although that's a start)!
- Provide background into your industry, your organizational goals, and the problem domain
- Provide support for the data
  - Answer questions about the data and what kinds of analysis is valued
- Connect with the faculty
- Visit classes – either in person or via teleconference
- Sponsor (existing) hackathons that might use your data
Understanding the domain is essential

- How was the data generated or collected?
  - What do the data mean (i.e., how to interpret)?
  - What kinds of errors are likely to be present?
    - What are the sources of noise in the data sources or labels?
    - What does missing data mean?
    - What level of precision is important?
- What aspects of the data are already known (or believed) to be predictive of the desired target?
- What kinds of solutions are valuable to the organization?
A peek into my Thursday morning presentation: Introduction to Time-Series Analysis with PI System and R

Fermentation data was shared through and obtained from the OSIsoft Academic Community Service.
Lehigh University (Bethlehem, PA) is considered one of the hidden ivies. Mission statement: *To advance learning through the integration of teaching, research, and service to others.*

**CHALLENGE**
Should organizations share data to help educate students?

- Concerns about exposing IP
- Concerns about privacy
- Already lots of data being shared

**ADVANTAGES**
Sharing data can impact students and benefit your organization

- Modern problems motivate students
- Students learn about issues and processes that you care about
- New venues to expose your brand

**HOW**
Many options for sharing are available

- PI data can be shared via OSIsoft Academic Community Service
- Share other data via data brokers, repositories, and web sites

Lehigh University (Bethlehem, PA) is considered one of the hidden ivies. Mission statement: *To advance learning through the integration of teaching, research, and service to others.*
Contact Information

Brian D. Davison
davison@cse.lehigh.edu
Associate Prof., Lehigh University

Feel free to share your data with me! 😊
Thank You

Merci
Danke
Gracias
Grazie
Obrigado
감사합니다
谢谢
Спасибо
ありがとうございます