DATA SCIENCE ACROSS DISCIPLINES
FOCAL POINT PROJECT

Diana Byrne & Halie Rando
October 7, 2015
Graduate students need to learn data science skills

- In 2011, McKinsey Global Institute highlighted the importance of data management and data analysis skills across industries/sectors
- Volume of data is increasing in many fields (e.g., text mining in digital humanities, biological sequencing technologies)
- Data science-oriented jobs require computational skills
- Are Illinois graduate students prepared to meet the expectations of their fields in the 21st century?
Advisors unable to provide support for learning these skills

When each student was asked whether his/her advisor had ever helped them learn a programming skill (including Excel), students responded:

“No – I learned mostly from other grad students… and I’ve been figuring out data management on my own.”

“No – but expected me to know how!”

“Yes(ish)- my old adviser gave me an R script to use but I pretty much unpacked it myself”

“No (well, I guess pivot tables)”
<table>
<thead>
<tr>
<th>How else can graduate students at Illinois learn data science skills?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Take an introductory computer science class</strong></td>
</tr>
<tr>
<td>- Requires extensive time commitment</td>
</tr>
<tr>
<td>- Male-dominated environment</td>
</tr>
<tr>
<td>- Starts at the basics and may not all be relevant</td>
</tr>
<tr>
<td>- Hemorrhage interested students (sources say up to 60%)</td>
</tr>
<tr>
<td><strong>2. Learn from their advisor or another graduate student</strong></td>
</tr>
<tr>
<td>- Requires that someone has these skills</td>
</tr>
<tr>
<td>- Requires that person to be willing to teach them</td>
</tr>
<tr>
<td><strong>3. Enroll in a MOOC or learn on their own</strong></td>
</tr>
<tr>
<td>- Requires student to have enough knowledge to know what he/she needs to learn</td>
</tr>
</tbody>
</table>

New option: participate in the Data Science Across Disciplines (DSAD) Focal Point project
The Data Science Across Disciplines (DSAD) Focal Point

- **Explicit code of conduct**
  Safe environment for people less typically welcome in tech environments

- **Open to graduate students from any discipline**
  Students from: biology, engineering, psychology, linguistics, anthropology, history

- **Speakers from across campus who use data in creative ways**
  Digital humanities, educational technology, spatial mapping of social factors, etc.

- **Data analytic and management-oriented skills**
  Advanced Excel, introductory Matlab and Python, familiarize with Unix terminal

- **Data management strategies**

- **Applications to research through final project**
Is DSAD working?

- “[My advisor] told me to talk to more experienced students, who have been helpful. However, from starting this class, I realize many of the techniques I have been taught from older students are inefficient.”

- “Now I can clean my data or find trends quickly! I was spending so much time on things that literally take 30 seconds now that I format, filter, then sort my data in ways I had never thought of before!”

- “Sorting is awesome! Actually organizing my data is awesome! I wish I had done all my many Excel sheets the way I have them set up now.”

- “I have already used some ‘if’ statements to sort data!”

- “Conditional formatting! I LOVE conditional format ever since we went over it in class. I’m actually pretty obsessed with it.”
Showed improvements with Excel

Baseline Understanding (%)

Understanding after Lesson (%)

Very comfortable
Somewhat comfortable
Not comfortable
What’s next?

- For our participants, introductory data analysis with **Matlab** and **Python**
  - Basic vocabulary and syntax
  - Importing and organizing data
  - If/elseif/else statements, for/while loops
  - Graphing/visualizing data

- Main goal: **Expose students to the possibilities**, while giving them the **resources** and **confidence** to continue gaining and applying these skills
Second semester final project

Choose a dataset: Students can choose one from their research or one that they find interesting.

Come up with a question: Students ask an interesting question of their datasets.

Pick a language: Students identify the language(s) that best fits with their project.

Determine methods and get results: Students write the code they need to complete their project.

Present final work as a poster: Next semester we will showcase student projects at a poster session.
Want to support DSAD?

1. Check out our website (or google “data science across disciplines focal point”): [http://publish.illinois.edu/data-science-across-disciplines/](http://publish.illinois.edu/data-science-across-disciplines/)

2. Join our mailing list using the “Contact” page on the website

3. Attend the speaker series (106 Main Library):
   - October 20, 4 pm: Dr. Chad Lane (Education)
   - November 17, 4 pm: Nicole Brown (Sociology)

4. Attend our final poster session in May (details TBD)
Acknowledgements

• Heidi Imker and Ayla Stein, our faculty organizers
• Jeremy Guest and Anna Kukekova, our advisors
• Graduate School of Library and Information Science (GSLIS)
• University Library and the Research Data Service (RDS)
• Illinois Informatics Institute (I3)
• Graduate College and the Focal Point initiative for funding DSAD