A Social Scientific Approach to Data Science: Building the Penn State PhD in Social Data Analytics

Burt L. Monroe*

Roundtable on Data Science Postsecondary Education Programs and Approaches for Data Science Education at the PhD Level
National Academies of Sciences, Engineering, and Medicine
June 13, 2018, Washington, DC.

*Liberal Arts Professor of Political Science, Social Data Analytics, & Informatics
Head, Program in Social Data Analytics
Director, Center for Social Data Analytics (and BDSS and QuaSSI)
Pennsylvania State University

This work is partially supported by NSF IGERT Grant 1144860, “Big Data Social Science.”
Social Data Analytics @ Penn State

- At Penn State, **SoDA** refers to
  - A *field of study* integrating social science and data science approaches to learning from complex &/or intensive data arising from human interactions.
    - (Conversely, we don’t see SoDA as a *domain* of data science.)
    - (Further … within SoDA we are often siloed into “domains” according to a common structure in social data — network / hierarchy / space / time — or a channel through which human interaction occurs — text / language / image / video.)
  - A **doctoral minor** that PSU PhD students in any program may pursue.
  - An undergraduate **BS degree**.
  - A product of the NSF-funded **Big Data Social Science (BDSS) IGERT**, itself a project of the **Quantitative Social Science Initiative (QuaSSI)**.
  - Beginning Fall 2018, a new **Center for Social Data Analytics**, consolidating that alphabet soup with combined research, education, and community-building missions.
Some reasons to care …

• It’s probably beyond my scope here to convince you of the scientific or societal merits of SoDA.

• So let me highlight some instrumental merits …
  • The SoDA model has resulted in impactful student research.
  • The SoDA model has resulted in attractive student placements.
  • The SoDA model is attractive to students of more diverse backgrounds than is typical in data science.
The MID4 dataset, 2002–2010: Procedures, coding rules and description
G Palmer, V d'Orazio, M Kenwick, M Lane
Conflict Management and Peace Science 32 (2), 222-242

App recommendation: a contest between satisfaction and temptation
P Yin, P Luo, WC Lee, M Wang
Proceedings of the sixth ACM international conference on Web search and data...

Mapping moods: geo-mapped sentiment analysis during hurricane Sandy
C Caragea, A Squicciarini, S Stehle, K Neppalli, A Tapia
Proc. of ISCRAM

Understanding topics and sentiment in an online cancer survivor community
K Portier, GE Greer, L Rokach, N Ofek, Y Wang, P Biyani, M Yu, ...
Journal of the National Cancer Institute Monographs 2013 (47), 195-198

CiteSeerX: AI in a Digital Library Search Engine
J Wu, K William, HH Chen, M Khabsa, C Caragea, S Tuarob, A Ororbia, ...
AI Magazine 36 (3)

Citesearx: AI in a digital library search engine
J Wu, K Williams, HH Chen, M Khabsa, C Caragea, A Ororbia, D Jordan, ...
The Twenty-Sixth Annual Conference on Innovative Applications of Artificial...

Differentially private graphical degree sequences and synthetic graphs
V Karwa, AB Slavkovic
International Conference on Privacy in Statistical Databases, 273-285

A straw shows which way the wind blows: ranking potentially popular items from early votes
P Yin, P Luo, M Wang, WC Lee
Proceedings of the fifth ACM international conference on Web search and data...

Recent awards won by SoDA students include:

- 2017 Data Science for Public Good Fellowship, VPI: Sayali Phadke & Claire Kelling (Statistics)
- 2017 Data Science for Social Good Fellowship, U Wash: Mitch Goist (Political Science)
- 2015 Data Science for Social Good Fellowship, U Chicago: Fridolin Linder (Political Science)
- 2017 NASA PA Space Grant Graduate Fellowship: Carolynne Hultquist (Geography)
- 2017 Best Poster, Political Networks Conference: Matt Denny (Political Science)
- 2016 Joint Statistical Meetings ASA Student Paper Award: Josh Snoke (Statistics)
- 2017 Population Association of America Best Poster Award: Cassie McMillan (Sociology)
- 2015 Sloan Foundation UCEM Award: Alex Ororbia (Information Science)
- 2015 NIH Pathways T32 Predoctoral Fellowship: Rachel Koffer (Human Development)
Alumni placements (BDSS/SoDA)

• Tenure track:
  • Informatics: RIT
  • Political Sci: Delaware, Georgia, Miami (OH), (Minnesota)
  • Statistics: Ohio State

• Postdoctoral positions:
  • Geography: Maynooth, UCLA
  • Human Development & Family Studies: Ohio State
  • Statistics: (Carnegie Mellon), (Harvard)
  • Political Science: Chicago, Concordia, (Harvard), (Johns Hopkins), NYU, Penn

• Data science in industry or government:
  • Computer Sci: IBM Research
  • Geography: NASA, Strava
  • Health Policy / Demog: Ipsos
  • Informatics: Google
  • Political Science: Google, IARPA, Verisk (x3)
  • Sociology / Demography: RTI
  • Statistics: Google, SAIS, RAND

• (indicates initial placement, same student listed again)
SoDA recruits & attracts students from a greater variety of backgrounds than is typical in data science programs.

From 2012-2018, 23% of American IGERT trainees (7/30) were from underrepresented groups, and 50% of funded trainees (18/36) were women.

In 2017-18, 70% of active IGERT trainees (7/10) and 75% of the SoDA cohort (9/12) are women.
In the beginning, there was **SSSP:** The Social Science Statistics Partnership
2006
- SSSP renamed QuaSSI: the Quantitative Social Science Initiative.
- Predoc eligibility expanded college-wide.
2008-

• Predoc eligibility expanded campus-wide.
• QuaSSI participated in multiple cyberinfrastructure projects and the taskforce that created the Institute for Cyberscience.
The IGERT proposal went in with 84 “participating” faculty from 17 departments, and 25 research centers and labs. Matching support was pledged by six colleges, two campus-wide institutes and the VP for Research.
Elements of BDSS-IGERT

• Over a two year Traineeship (typically the 2nd-3rd year of the PhD), Trainees were expected to participate in …
  • **Community-building** (speakers, workshops, poster sessions, hackathons, group projects, facetime with each other).
  • Two academic-year **research rotations** in relevant interdisciplinary projects.
    • Generally expected to “cross the social / non-social science boundary” at least once.
  • Two summer **externships** (which means “go away”).
    • Generally expected to have at least one to be in a nonacademic setting (industry, government, nonprofits).
  • The **SoDA curriculum**, in early years as the boat was built on the open sea.
A place to meet: The Databasement
A place to meet - The Databasement

10 Creepy Places You Never Noticed On Campus

3: Sparks Building

I bet you didn’t know Sparks has a “sub-basement.” I didn’t either until I found it by accident. This isn’t actually a room, it’s an opening in the wall I couldn’t even fit through. I’m not really sure how anything was stuffed into this crawl space without the use of child labor.

Unless you’ve had a class down here, you’d never know about this perfect horror movie office. It was filled with 70’s furniture, toys, and a tv. I am half-convinced this was at one point a bomb shelter. It was weird, there was like a coffee room to the side and it had been used recently except NO ONE was down there. I did hear noises though and I nearly got trapped between two sets of locked doors when I tried dipping out.

https://onwardstate.com/2012/10/29/at-least-10-creepy-places-on-campus/
A place AND TIME to meet
Externships
SoDA is a “Dual-title PhD”

- The dual-title PhD is **Penn State’s mechanism for creating interdisciplinary graduate programs** (without creating departments).
- Students earn a PhD with two titles, e.g. “Sociology and Demography.”
- Examples of organizationally complex intercollege dual-titles are:
  - Operations Research (20 programs in 10 colleges),
  - Human Dimensions of Natural Resources & the Environment (9 in 6),
  - Biogeochemistry (8 programs in 6 colleges),
  - Women’s Studies (10 programs in 4 colleges),
  - Demography (7 programs in 4 colleges),
  - **Social Data Analytics (6 programs in 5 colleges)**.
- More are simpler intracollege dual-titles: e.g., Language Science or Astrobiology.
Generic dual-title PhD rules

• Student must first be admitted to approved home PhD program.
• Student must be admitted to dual-title *before* candidacy. Candidacy committee must have chair/co-chair from dual-title field.
• Comprehensive exam / dissertation committee must have chair/co-chair from dual-title field.
• Dissertation must have “substantial” content from the dual-title field.
• Students must be able to complete the dual-title course requirements with no more than two semesters of delay beyond the home degree requirements.
Challenges in design

- The union of my colleagues’ lists of their “bare-bones minimum” requirements would take 40 years to complete. Must be doable.
  - Conversely, it must be a PhD, and it should be different than what the student would do otherwise.
  - Some departments have almost no specific requirements (Geography) and some have almost no flexibility for 2-2.5 years (HDFS).
- The “social science / non-social-science” boundary.
- Literally no two programs interpret or implement “candidacy” and “comprehensive” “exams” identically.
- If the graduate faculty is too narrowly defined, it will be difficult or impossible for students to form committees. If too broadly defined, it’s meaningless … no one has ownership, no one knows the rules.
PhD Students, Dual-title & Doctoral Minor
SoDA-specific dual-title requirements

- Core seminars (come closer to the end than the beginning)
  - SoDA 501 — Big Social Data: Approaches and Issues
  - SoDA 502 — Social Data Analytics: Approaches and Issues

- 18 credits of approved electives collectively satisfying distribution requirements:
  - 3 credits: A — Analytics
  - 6 credits: S — Social
  - 6 credits: Q — Quantitative [Statistical - S was taken] / Mathematical
  - 6 credits: C — Computational / Informational
  - 3 credits: Departmental cluster 1 (Social science / Statistics)
  - 3 credits: Departmental cluster 2 (Informatics / Geography / Comp Sci / Engineering)
  - 6 credits: Outside home department.

- Students pick up some required A, S, Q, and/or C credits in the course in doing their home PhD. As a result, the SoDA requirements typically amount to four “extra” courses: SoDA 501, SoDA 502, and two courses outside your home program that together satisfy any remaining A, S, Q, C, and cluster requirements.
The Social Data Stack

Structure of SoDA 501/2

Informatics

Computer Science

Geography (Visual Analytics)

Applied Linguistics

Econ

Psych

Crim

Popular Electives

Social Data Analytics @ Penn State

Political Science

Network Analysis for PS

Text as Data

Modern Measurement

Causal Inference

Bayesian Inference

Sociology (& Demography)

Social Network Analysis

Survey Research

Spatial Demography

Observational Data

Health Policy

Rural Sociology

Human Development & Family Studies

Data-Mining for HDFS

Intensive Longitudinal Data

Bayesian Measurement

Wearable Technology

Big Data for Public Health

RPTM

Social Media Mining for Tourism Studies

Network Science

Social Network Analysis for PS

Survey Research

Spatial Demography

Observational Data

Corpus Linguistics

Reproducible Science

Social Media Mining for Demography

Industrial Engineering

Data-Driven Design

Computer Science

Pattern Recognition

Graph Mining

Computer Vision

Computational Semantics

Matrix Computation

Informatics

Big Data Fundamentals

Deep Learning

Artificial Intelligence

Information Retrieval

Computational Psycholinguistics

Network Visualization

Statistics

Computational Statistics

Spatial Statistics

Graphical Models

Statistical Privacy

Experimental Design

Physics

Network Visualization

Reproducible Science

Social Media Mining for Tourism Studies

Network Science

Computational Statistics

Spatial Statistics

Graphical Models

Statistical Privacy

Experimental Design

Physics

Network Visualization

Reproducible Science

Social Media Mining for Tourism Studies

Network Science
Ongoing challenges

• Big universities have very healthy immune systems for rejecting the infection of interdisciplinarity.
  • Cultural, language (jingle/jangle), and funding model differences a constant battle.
  • Every time a staff member, grad director, etc., moves, we lose institutional memory (losing students, requiring retraining or, worse, reselling). Almost nobody really “works for me.”

• Scheduling is really a nightmare.

• Our efforts toward enabling nonacademic / alt-ac have been too successful.

• Many many students express interest, but really need a Master’s or certificate program focused on “skills” development.

• 25% (made up #) insert the word “media” after “social.”

• Transition from IGERT to Center, while ramping up SoDA.
Thanks! For more info on …

- BDSS-IGERT and SoDA:
  - [http://bdss.psu.edu](http://bdss.psu.edu) (coming soon, [http://soda.psu.edu](http://soda.psu.edu)), @BDSS_PSU

- SoDA Graduate Program admissions and requirements:
  - [http://bdss.psu.edu/soda](http://bdss.psu.edu/soda)

- SoDA and BDSS IGERT faculty and students:
  - [http://bdss.psu.edu/people](http://bdss.psu.edu/people)

- BDSS IGERT student research:

- A fun tale of a SoDA team disrupting data science norms and breaking a Kaggle contest:

- For some (mostly historical) information about QuaSSI:
  - [http://qssi.psu.edu](http://qssi.psu.edu)

- The Bachelor’s of Science in SoDA:
  - [http://soda.la.psu.edu](http://soda.la.psu.edu)