Models of support for Data Science

The perspective of two libraries

CNI Spring Meeting 2024

David Minor, UC San Diego
Jan Brase and Bela Gipp, SUB Göttingen
Content

1. Introduction
2. UC San Diego, the situation
3. GippLab: Scientific Information Analytics
4. Student and researcher exchange
5. Lessons learned – next steps
What are we talking about?

Since 2016 the UC San Diego and the State and University Library Goettingen in Germany have been actively cooperating and learned from each other through staff visits and joint projects

MOU (signed in 2016 and renewed in 2019) to compare and contrast research support organizations:

- Comparing different support models
- Relationships between our libraries and respective campuses
- Overall goals and objectives
Future work (2024 - ?)

At CNI fall 2023 we presented our longtime Exchanges on RDM infrastructures and services.

For the next years we will additionally focus on:
• Data / research analytics
• Academic exchanges – students
• Further staff exchanges
• Re-examine technical discussions

December 11, 2023
UC San Diego, the situation
UC San Diego Data Science Program - Founded in 2015

Fall 2022: expanded to include three new graduate degrees: in person MS, online MS (joint with Computer Science & Engineering, first fully online graduate degree program offered by UC San Diego)

Fall 2023: ~1000 undergraduate majors, ~6000 students taking courses, ~200 graduate students

The Data Science Program is primarily focused on education and training students in data science.
Brings together faculty members and researchers from various departments across the university to work on data-driven research projects.

Offers various educational programs, including a PhD in Data Science.

Provides resources and support for data science research and education.

HDSI is focused on advancing the field of data science through research and innovation.
Data science in the Library

Hired Data Science Librarian in 2018
- Data & GIS Lab
- Wide range of services
- Bringing student data into Library Digital Collections
- Forthcoming paper in Journal of eScience Librarianship

Data science support in the Library has to this point been external facing, for students and faculty. We are not yet doing work for internal Library needs.
Data science undergrad support - interactions

- Annual workshops on developing independent (outside of courses) data science projects
- 1-on-1 meetings with students to develop their independent projects, focused on scoping and finding/accessing appropriate data, as well as how to document and present their findings
- Working with capstone student groups to identify data sources for group-based projects
- Working with student groups on their capstone projects - finding and accessing licensed/subscription data; acting as liaison between vendors and student groups when needed
Data science undergrad support - materials

- Data Science LibGuide with general data science resources (Python tutorials, etc.), and content tied to course curriculum (e.g. assigned readings, direct links to purchased and open access books in library catalog)

- Sharing information about hackathons/datathons, new data resources, Library-hosted events (e.g. UC Love Data Week) and other relevant content in weekly departmental student newsletter

- Quarterly newsletter to faculty about Library updates

- Creating and maintaining a web archive of capstone projects (project websites, GitHub repositories, web-based visualizations, etc.)
Data science grad support

- Identifying data suitable for student-organized hackathons/datathons; purchasing requested books for new grad courses; early stages of international student exchanges

- Data Science & Engineering MAS program (not technically HDSI): working directly with students for the annual ingest of group capstone projects

- Computational Social Science grad program (not HDSI, but one of many data science-ish grad programs): working with students to access and analyze data, including leveraging the Library’s Data & GIS Lab virtual machines for large data/compute Windows-based analyses
Next steps: Educational Dataset Service

- Curated datasets for student use in instruction and research:
  - Real research datasets
  - Sanitized and synthetic administrative data (e.g. Student Activity Hub, IR)
  - Datasets commonly used for training in specific disciplines
- Guidance & training for students on working w/associated datasets
  - How to re-use data
  - How to cite sources
  - How to work with metadata
- Support for instructors in developing & assessing meaningful assignments
- Usable in the student Data Science/Machine Learning Platform or elsewhere
GippLab: Scientific Information Analytics

University of Göttingen and State and University Library (SUB) Göttingen

Bela Gipp
Our Research

Core Research Areas
- Information Retrieval
- Natural Language Processing
- Blockchain Technology

Foundations
- Data Science
- Artificial Intelligence
- Information Visualization

Natural Language Processing
On April 29, 2017, Bill Gates partnered with Swiss tennis legend Roger Federer in playing the “Match for Africa” 4, a noncompetitive tennis match at a sold-out Key Arena in Seattle. The event was in support of Roger Federer Foundation’s charity efforts in Africa.

Bill Gates teamed up with Swiss tennis player Roger Federer to play in the “Match for Africa 4” on April 29, 2017. The noncompetitive tennis match at a sold-out Key Arena in Seattle was in support of Roger Federer Foundation’s charity efforts in Africa.
Applications

Plagiarism Detection

The term $C(\phi, \phi_\theta)$ can be decomposed, according to the joint being actuated or not, into two independent parts of the following forms:

$$C(\phi, \phi_\theta) = C_U(\phi, \phi_\theta) + C_{\phi U}(\phi, \phi_\theta)$$

with each element $C_U(i, j)$ and $C_{\phi U}(i, j)$ defined as

$$C_U(i, j) = \sum_{k=0}^{n} c_{n(i), n(j), k} \phi_k$$

(38)

and likewise $B_U(\phi)$ and $B_{\phi U}(\phi)$ have elements

$$B_U(i, j) = b_{n(i), n(j), k} \phi_k$$

(39)

and further, $F_U(\phi_\theta)$ has element $F_U(i)$ given by

$$F_U(i) = F(\phi_i) + F_\phi(\phi_i)$$

(42)

hence it has been assumed $C_U(i) = -K_U(i)$. The terms $C_{\phi U}(\phi, \phi_\theta)$ and $C_{\phi U}(\phi, \phi_\theta)$ respectively have elements

$$C_{\phi U}(i, j) = \sum_{k=0}^{n} c_{n(i), n(j), n(k)} \phi_k$$

and likewise $B_{\phi U}(\phi)$ and $B_{\phi U}(\phi)$ have elements

$$B_{\phi U}(i, j) = b_{n(i), n(j), n(k)} \phi_k$$

(40)

and

$$F_{\phi U}(\phi_\theta) = F_{\phi U}(\phi_\theta)$$

(41)

where $F_{\phi U}(\phi_\theta)$ takes the form

$$F_{\phi U}(\phi_\theta) = F_{\phi U}(\phi_\theta)$$

(42)
Applications

IP Protection

1. User submits the unique fingerprint (SHA-256) of his file to OriginStamp service.

2. Once a day, OriginStamp aggregates the submitted hashes and embeds them into a transaction that is broadcasted over the network.

3. Anyone can verify the tamper-proof timestamp using the distributed blockchain once the transaction is confirmed by the public, distributed blockchain.

- Timestamp your documents
- Location-based Exploration
- Prove the existence of documents
Applications

News Analysis
Definition 3.4. (Generalisierte Tauschabbildungen) Die Menge aller Tauschpolynome von $\mathbb{C}$ nennen wir generalisierte Tauschabbildungen.

Des Weiteren halten wir einen Knoten $v_n \in V$ fest und führen den Anfangs-Cluster $w(v_0)$ aus $n$ unabhängigen Variablen $x_1(v_0), \ldots, x_n(v_0)$ ein. Zu jedem Knoten $v_k \in V$ fügen wir nun ein Cluster $w(v_k)$ aus $n$ Elementen $x_1(v_k), \ldots, x_n(v_k)$ aus dem rationalen Funktionenkörper $\mathbb{Q}(x_1(v_0), \ldots, x_n(v_0))$ hinzu. Die Elemente $u(v_k)$ sind endseitig durch die Tauschrelationen (3.1) bis (3.2) bestimmt. Sie gilt für jede Kante $v_k \rightarrow v_{k+1}$.

\begin{align*}
x_i(t) &= x_i(t') \\
x_i(t)x_i(t') &= P(x_i(t))
\end{align*}

Dann ergibt sich ein vollständiger Cluster. Weiterhin ist zu beachten, dass $P$ nicht von $n$ abhängig und damit (3.2) auch gilt, wenn wir die Richtung (Orientierung einer Kante) ändern. Wir haben uns damit einen Caterpillar-Baum konstruiert. Fassen wir die Eigenschaften eines Caterpillar-Baums noch einmal in einer Definition zusammen.
Researcher and Student exchange

- **Academic exchanges – students**
  - UCSD students mandatory have to work on a project for their data science master - SUB and University of Goettingen have many projects
  - Uni Goettingen students are encouraged to spend a semester abroad and work on projects – Extrensiv motivation for excellent students.

- **Further staff exchanges**
  - Walk in each others shoes, looking over the shoulder in understanding daily work
  - Networking
Funding Programms for Exchange

There are several scholarships by the DAAD for:

- Undergraduates: [LINK](#)
- Graduates, PhDs, Postdocs from the US
  - up to 6 months: [LINK](#)
  - up to 12 months: [LINK](#)
- For jointly supervised PhD candidates: [LINK](#)
- For academics & scientists (up to 3 months): [LINK](#)
Lessons learned - Next steps

UCSD library and SUB Göttingen have benefitted from learning from each other through staff visits.

Additionally UCSD and University of Göttingen will cooperate in data science research and teaching, this time fueled by researchers and student exchanges.

We aim at having the first exchanges in 2024-2025
Thank you

Jan Brase - brase@sub.uni-goettingen.de
David Minor – dminor@ucsd.edu
Bela Gipp- gipp@uni-goettingen.de
www.gipplab.org