

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



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.

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

- Founded in 2018

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.

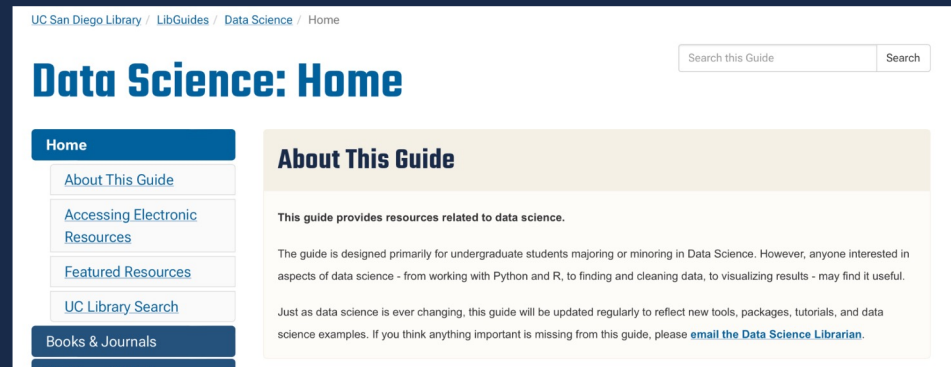
Data science undergrad support - interactions

- 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/subsription 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.)



The screenshot shows the 'Data Science: Home' LibGuide page. At the top, there is a breadcrumb trail: 'UC San Diego Library / LibGuides / Data Science / Home'. To the right is a search box labeled 'Search this Guide' with a 'Search' button. The main heading is 'Data Science: Home' in large blue font. Below the heading is a navigation menu with 'Home' highlighted in blue, and other options: 'About This Guide', 'Accessing Electronic Resources', 'Featured Resources', and 'UC Library Search'. Below the menu is a section titled 'Books & Journals'. The main content area has a heading 'About This Guide' and a sub-heading 'This guide provides resources related to data science.' The text below explains the guide's purpose for undergraduate students and mentions it will be updated regularly. It also includes a link to 'email the Data Science Librarian'.

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

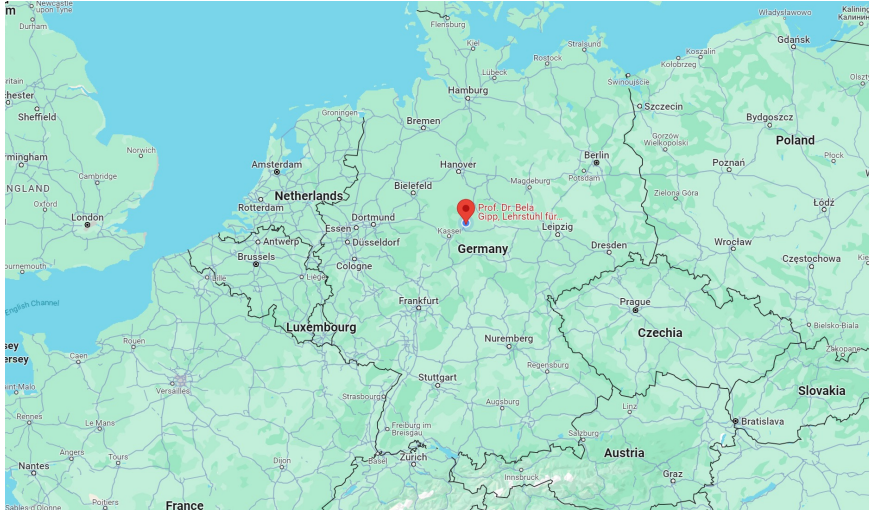
Next steps: Educational Dataset Service

- 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

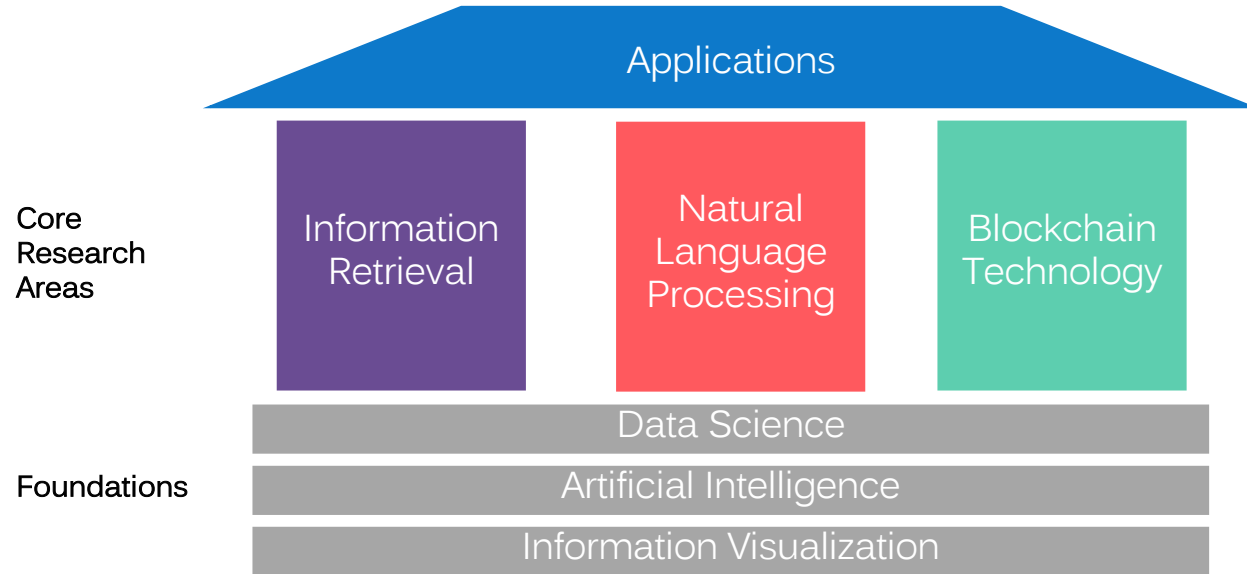




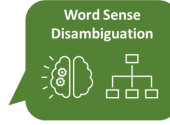
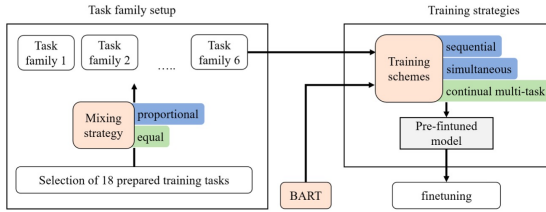
www.gipplab.org/team

Our Research

Natural Language Processing



Natural Language Processing



Natural Language Processing

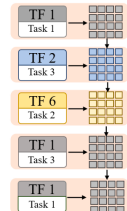


Original Text

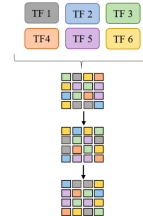
...
 On **April 29, 2017**, **Bill Gates** partnered with Swiss tennis legend **Roger Federer** in playing the "Match for **Africa**" 4, a noncompetitive tennis match at a sold-out Key Arena in Seattle. The event was in support of **Roger Federer** Foundation's charity efforts in **Africa**.
 ...

Paraphrased using CPT-3

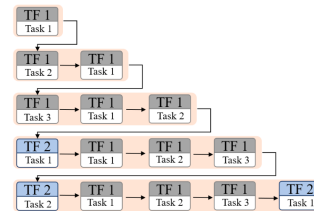
...
Bill Gates teamed up with Swiss tennis player **Roger Federer** to play in the "Match for **Africa** 4" on **April 29, 2017**. The noncompetitive tennis match at a sold-out Key Arena in Seattle was in support of **Roger Federer** Foundation's charity efforts in **Africa**.
 ...



(a) Sequential learning.



(b) Simultaneous learning.



(c) Continual multi-task learning.

Downloads last month
 635



Hosted inference API

Text2Text Generation

Examples

question: which description describes the word "java" best in the following context? descriptions: ["A drink consisting of an infusion of ground coffee beans ", " a platform-independent programming lanugage ", or " an island in Indonesia to the south of Borneo "] context: I like to drink 'java ' in the morning.

Compute

#+Enter

1.7

Computation time on cpu: 1.596 s

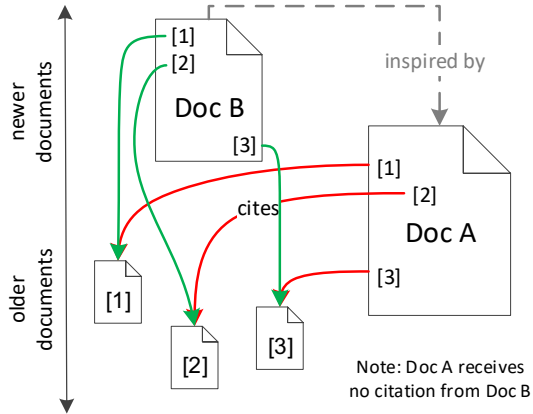
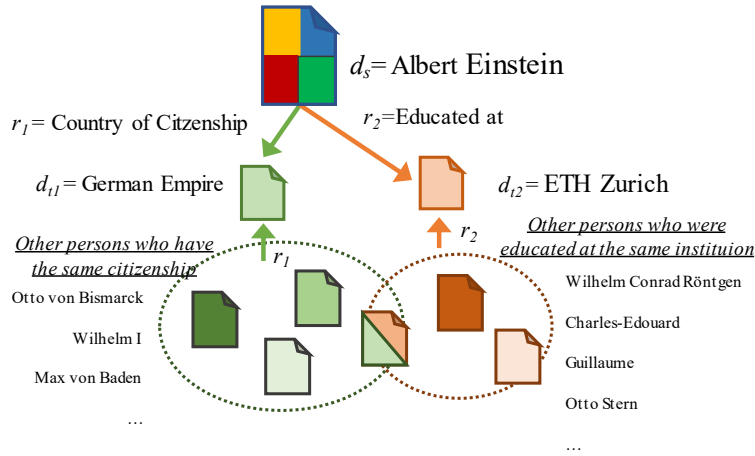
a drink consisting of an infusion of ground coffee beans

<> JSON Output

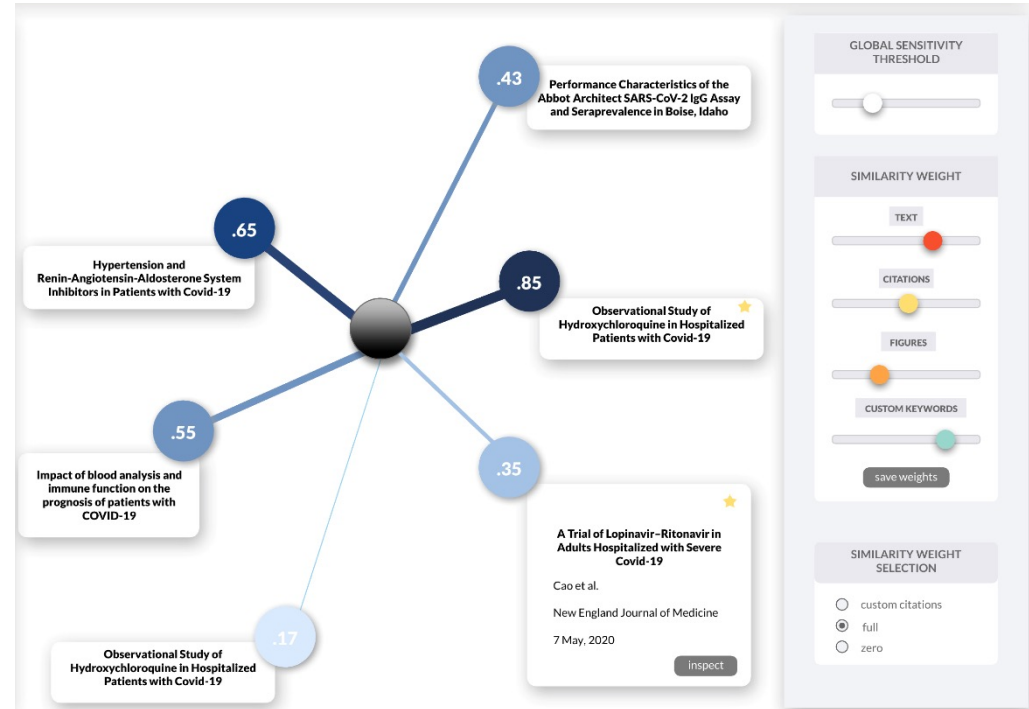
Maximize

Space using jpwahle/t5-word-sense-disambigua..

linpearjun/T5-LARGE



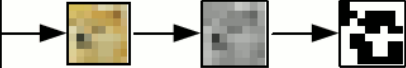
Literature Recommendation



lexical, syntactic and semantic text similarity.

To summarize the contributions of this article, we refer to the four questions [136] suggested to assess the quality of literature reviews:

1. Are the review's inclusion and exclusion criteria described and appropriate?
2. Is the literature search likely to have covered all relevant studies?
3. Did the reviewers assess the quality/validity of the included studies?
4. Were the basic data/studies adequately described?



```
10000000
11011110
10111110
00011011
01110000
01110011
00110011
00111111
```

The term $C(\Phi, \dot{\Phi})$ can be decomposed, according to the joints being actuated or not, into two independent parts of the following forms

$$C(\Phi, \dot{\Phi}) = C_U(\Phi, \dot{\Phi}) + C_{UC}(\Phi, \dot{\Phi}) \quad (37)$$

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

$$C_U(i, j) = \sum_{k=1}^n c_{N_U(i), N_U(j)} \dot{\Phi}_k, \quad (38)$$

$$C_{UC}(i, j) = \sum_{k=1}^n c_{N_U(i), N_C(j)} \dot{\Phi}_k, \quad (39)$$

and likewise $B_U(\Phi)$, $B_{UC}(\Phi)$ have elements

$$B_U(i, j) = b_{N_U(i), N_U(j)}, \quad (40)$$

$$B_{UC}(i, j) = b_{N_U(i), N_C(j)}, \quad (41)$$

and further, $F_U(\Phi, \dot{\Phi})$ has element $F_{U,i}$ given by

$$F_{U,i}(\Phi_{N_U(i)}, \dot{\Phi}_{N_U(i)}) := F_s(\Phi_U) + F_v(\dot{\Phi}_U). \quad (42)$$

hence it has been assumed $G_U(\cdot) = -K_U(\cdot)$. The terms $C_U(\Phi, \dot{\Phi})$ and $C_{UC}(\Phi, \dot{\Phi})$ respectively have elements

$$C_{U,i,j} = \sum_{k=1}^n c_{\mathcal{I}_U(i), \mathcal{I}_U(j), k} \dot{\Phi}_k, \quad C_{UC,i,j} = \sum_{k=1}^n c_{\mathcal{I}_U(i), \mathcal{I}_C(j), k} \dot{\Phi}_k$$

and likewise $B_U(\Phi)$ and $B_{UC}(\Phi)$ have elements

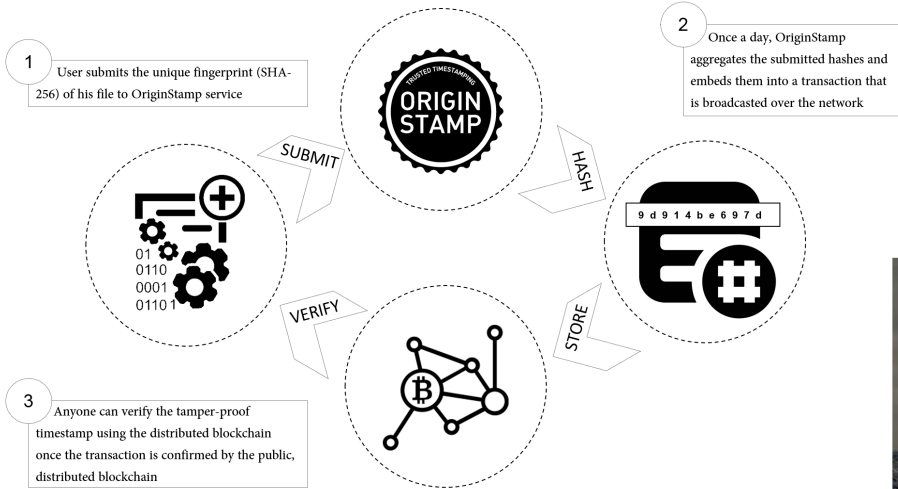
$$B_{U,i,j} = b_{\mathcal{I}_U(i), \mathcal{I}_U(j)}, \quad B_{UC,i,j} = b_{\mathcal{I}_U(i), \mathcal{I}_C(j)} \quad (16)$$

$F_U(\Phi_U, \dot{\Phi}_U)$ has elements $F_{U,i}$ which take the form

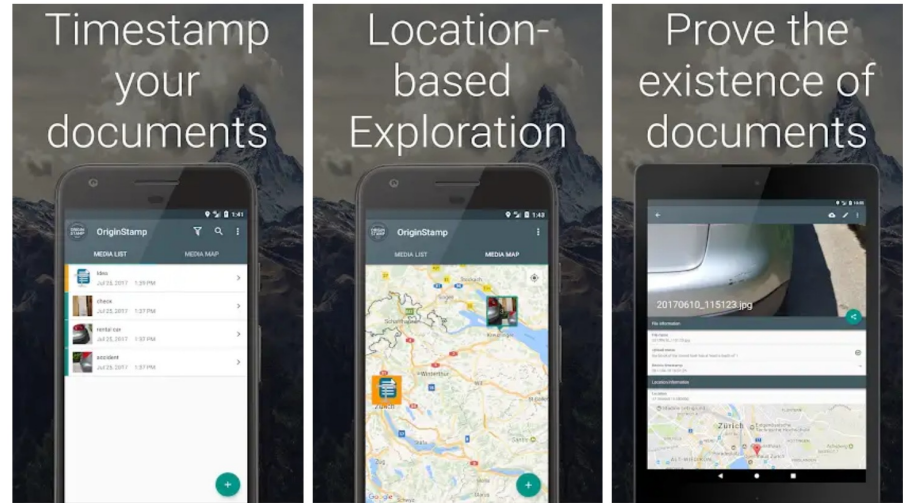
$$F_{\mathcal{I}_U(i)}(\Phi_{\mathcal{I}_U(i)}, \dot{\Phi}_{\mathcal{I}_U(i)}) = F_{s, \mathcal{I}_U(i)}(\Phi_{\mathcal{I}_U(i)}) + F_{v, \mathcal{I}_U(i)}(\dot{\Phi}_{\mathcal{I}_U(i)}) \\ := F_s(\Phi_U) + F_v(\dot{\Phi}_U) \quad (17)$$

Plagiarism Detection

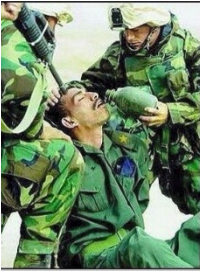
The screenshot shows the HYPLAG plagiarism detection tool. It compares two documents: 'Cascade controller design and stability analysis in FES-aided upper arm stroke rehabilitation robotic system' (Case 01) and 'FES based Rehabilitation of the Upper Limb using Input/Output Linearization and ILC' (Case 01). The interface displays similarity scores for citation, text, and formula. The text similarity score is 44%. The interface also shows a syntax tree for the compared text, with various elements highlighted in different colors (green, yellow, orange) to indicate similarity levels. The text being compared includes mathematical expressions and theorems related to system stability and control.



IP Protection



News Analysis



Objective



Frame 1



Frame 2

Search Terms: Date: 20141107 20141107

Publisher Country: Topic Country:

rusia ukraïne ALL_DOCS ukraine/england germany france

GB	Dozens Of Tanks Enter Ukraine From Russia	Ukraine crisis: Kiev accuses Russia of military invasion after tanks cross border	Ukraine crisis: Kiev accuses Russia of military invasion after tanks...	EU surcharge: UK will only pay back half of £1.7bn bill	EU surcharge: UK will only pay back half of £1.7bn bill	Look Who's Back film: Hitler lookalike 'seen touring Germany'	Peru, new train trips, and Birmingham-New York: Travel agenda
DE	Kyiv: 32 tanks enter Ukraine from Russia	Kyiv calls Berlin amid Russian incursion reports	EANS-Adhoc: Telekom Austria AG / Telekom Austria Group announces capital...	AuRico Gold Reports Third Quarter Financial Results and Ninth Consecutive...	Britain allowed to halve EU budget bill	Germany's east still lags behind	Read my lips: No new taxes, says Hollande
FR	Kyiv accuses Russia of sending more tanks and troops into Ukraine	Kyiv accuses Russia of sending more tanks and troops into Ukraine	Australia's G20 Presidency and the OECD: Joining forces to address the...	FRANCE: Paris school students protest against police violence after Green...	EU denies UK has received a budget bill discount	Germany: train strike causes chaos for commuters	Video: Widow of slain African Che Guevara seeks answers
US	Ukraine accuses Russia of sending dozens of tanks	Ukraine accuses Russia of sending dozens of tanks	Ukraine accuses Russia of sending dozens of tanks	Ukraine accuses Russia of sending dozens of tanks	Britain finds deal with EU over controversial bill	AP WAS THERE: The Berlin Wall crumbles	French president supports a Paris 2024 Olympic bid
RU	Ukraine Crisis: Sanctions Against Russia Not on G20 Agenda in Australia...	NATO Tactics Left Ukraine in 'Gray Zone' Between West and Russia: Atlantic...	Ukraine Crisis: Sanctions Against Russia Not on G20 Agenda in Australia...	Ukraine Crisis: Sanctions Against Russia Not on G20 Agenda in Australia...	Cameron Says Britain Will Pay Only Half of \$ 2.6 Bln EU Surcharge	Berlin wall: the symbol of Cold War as an art object	H... un... cooperation with...

Objective

Frame 1

Frame 2

A

B

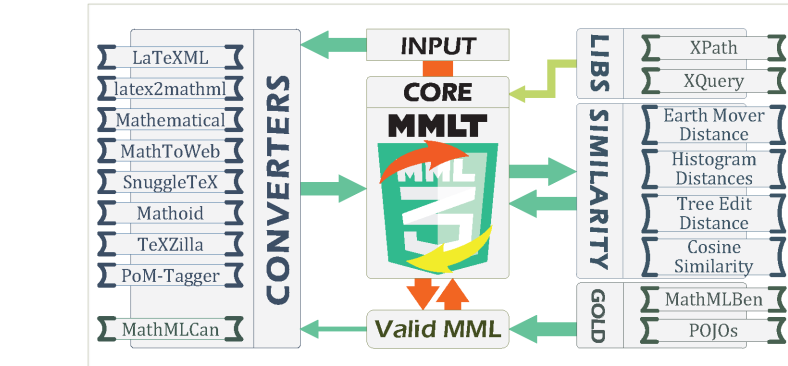
Definition 3.4. (Generalisierte Tauschabbildungen) Die Menge aller Tauschpolynome von \mathbb{T} nennen wir **generalisierte Tauschabbildungen**.

Des Weiteren halten wir einen Knoten $t_0 \in V$ fest und führen den Anfangs-Cluster $\mathbf{x}(t_0)$ aus n unabhängigen Variablen $x_1(t_0), \dots, x_n(t_0)$ ein. Zu jedem Knoten $t \in V$ fügen wir nun ein Cluster $\mathbf{x}(t)$ aus n Elementen $x_1(t), \dots, x_n(t)$ aus dem rationalen Funktionenkörper $\mathbb{A}(x_1(t_0), \dots, x_n(t_0))$ hinzu. Die Elemente $x_i(t)$ sind eindeutig durch die **Tauschrelationen** (3.1) - (3.2) bestimmt. So gilt für jede **Kante**

$$x_i(t) = x_i(t') \quad \text{für alle } i \neq k, \quad (3.1)$$

$$x_k(t)x_k(t') = P(\mathbf{x}(t)). \quad (3.2)$$

Damit ergibt sich ein vollwertiger Cluster. Weiterhin ist zu beachten, dass P nicht von x_k abhängt 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.



Math Information Retrieval

Special page

mass–energy equivalence

physical law

Math Formula Information

Formula: $E = mc^2$

Name: mass–energy equivalence

Type: physical law

Description: mass and energy are proportionate measures of the same underlying property of an object

Elements of the Formula

energy E quantitative physical property transferred to objects to perform heating or work or

mass m measure of the resistance of a physical body and its susceptibility to gravitational attraction

speed of light c speed at which all massless particles and associated fields travel in a vacuum

Data Source

<https://www.wikidata.org/wiki/Q235875>

WIKIPEDIA
The Free Encyclopedia

Article [Talk](#) [Read](#) [Edit](#) [Edit source](#) [View history](#) [More](#)

Mass–energy equivalence

From Wikipedia, the free encyclopedia

"E=MC2" and "E=mc2" redirect here. For other uses, see E=MC2 (disambiguation).

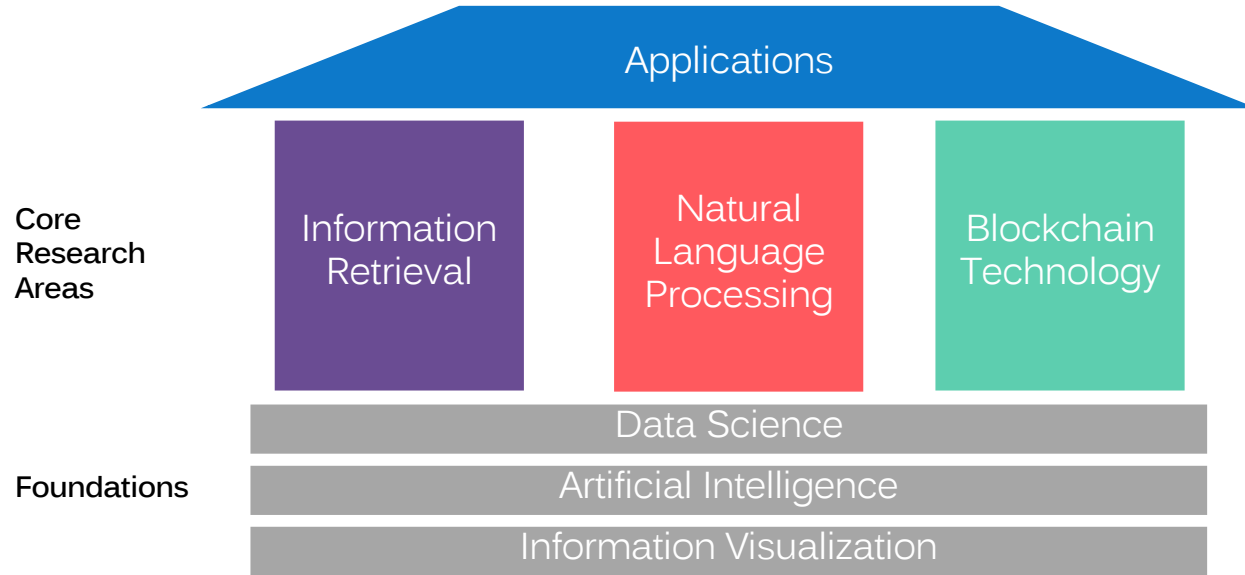
In physics, **mass–energy equivalence** is the principle that anything having mass has an equivalent amount of energy and vice versa, with these fundamental quantities directly relating to one another by Albert Einstein's famous formula:^[1]

$E = mc^2$



WIKIMEDIA
Community User Group
<math>

Applications



Core
Research
Areas

Information
Retrieval

Natural
Language
Processing

Blockchain
Technology

Data Science

Foundations

Artificial Intelligence

Information Visualization

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 motivtation 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

