COM-480 DATA VISUALIZATION COURSE PROJECT

BY DANIEL BUCHER, MARIEM BACCARI AND SYRINE NOAMEN

COMPUTER SCIENCE HALL OF FAME

May 30th 2025

Process book by:
DSM Group

website

Overview:

The CS Hall of Fame is an interactive data-driven website that visualizes the evolution and impact of foundational research in computer science. It celebrates seminal papers, influential authors, and transformative ideas across decades of innovation. Users can explore trends in subfields, track the influence of highly cited papers, and investigate co-authorship networks. The site presents data through dynamic visualizations such as citation bar charts, Sankey diagrams, word clouds, timelines, and collaboration graphs.

It aims to engage visitors with both a historical and analytical perspective on the growth of computer science as a discipline. It also aids students to find their interest and see how a paper can flourish and inspire many domains.

Motivation:

Computer science is a fast-moving field shaped by landmark publications and the contributions of visionary researchers. However, for newcomers and even seasoned professionals, understanding how these contributions interconnect over time and across subfields can be difficult.

The **motivation** for this project stemmed from the desire to:

- Make foundational computer science research accessible and engaging.
- Help students and researchers discover important papers, authors, and subfields.
- Explore how ideas propagate through co-authorship and publication venues.
- Use data visualization to transform abstract bibliometric data into a story of progress and collaboration.

By surfacing historical milestones and citation-driven influence, we wanted to create a compelling learning resource that sparks curiosity and honors the discipline's legacy.

Design Process:

The design process was guided by what we wanted to achieve and what we can get from the data in our hands.

Data sources were Semantic Scholar API which provides scientific publication data via an API and enriched with DBLP, the computer science bibliography.

Iterations and Improvements:

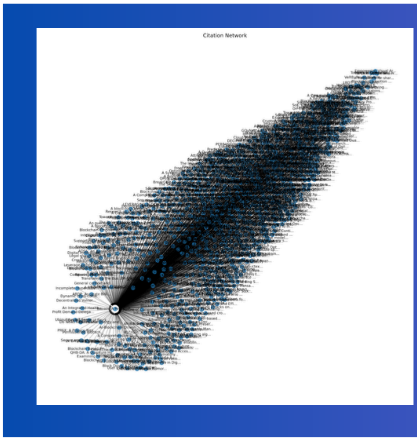
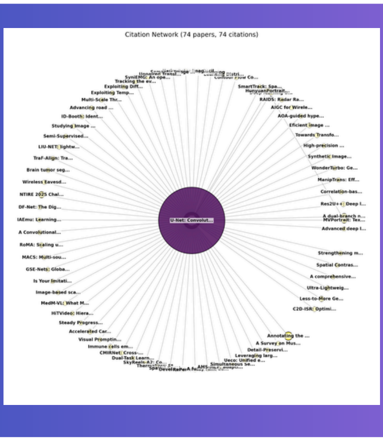
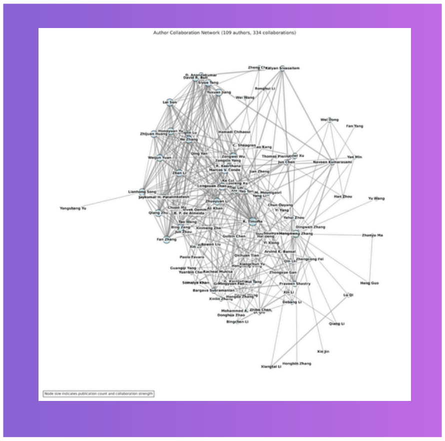
Initially, we used API, querying with keywords (e.g., "attention") and requesting fields such as title, publication type, date, citation count, venue, authors, abstract, and fields of study. However, this keyword-based search proved inadequate for our need to analyze a substantial volume of data to discern research trends. Our initial approach of bulk paper collection via the API was limited by its tendency to return only the most recent publications matching the query, often resulting in a skewed dataset predominantly from 2025. While DBLP offered a broader collection, it lacked the primary field of study, providing only keywords. To address this and chart the growth of research areas over time, we employed Natural Language Processing for keyword extraction across a defined period. By identifying the most frequent keywords or subfields and filtering out generic terms like "Computing," "Research," or "Engineering," we obtained informative field-specific keywords. Subsequently, we quantified the number of papers associated with each keyword annually and calculated its percentage within the total publications for that year. This enabled us to derive insightful longitudinal trends in research subfield evolution.

Challenges and Design Decisions

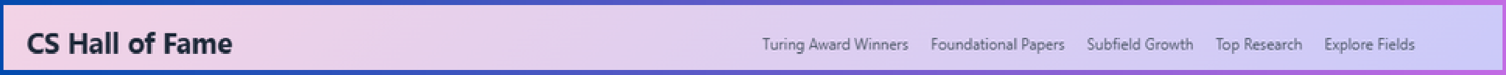
One major issue we faced was obtaining exhaustive data for figures that required ranking (Top Research at a Glance) or making connections (Influential papers in your favourite field) due to the API limits and timeout errors which made our scripts unusable for retrieving large amounts of data and when the API cooperated, the latency of retrieving data for a tremendous number of papers. DBLP imposed challenges as it consumes a lot of memory to process and is not much up to date and has many empty values. It was mostly used for the “Research Subfield Growth” analysis.

To deal with this, we decided to focus on recent individual papers we are interested in. For example, for the subfields section, we decided to handpick one major paper per subfield and build upon that by retrieving the papers that cite it.

For the most cited papers, we decided to only retrieve papers reaching a certain number of citations as a threshold.
To make a graph showing how papers are connected, we needed to find every time a main paper cited another paper. After we made a script to get this network graph, it didn't really make sense because it only went one way (from the main paper to what it cited). Also, going deeper than two levels took forever, like at least four hours of waiting! So, we either had to pick a main paper that wasn't cited a ton so we could see the papers that those papers cited, or we had to change our plan.
We decided to just look at one level of citations and then group the papers and the main paper by where they were published (the venue). For the graph showing authors working together, we made some rules about how many times they had to work together and we also sorted the papers by how relevant they were.

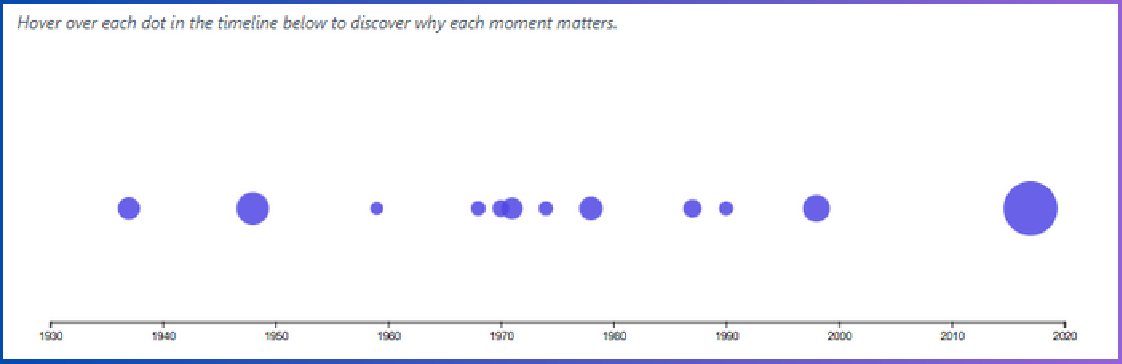
Initial Network graph with 2 levels	**Main paper with papers that cited it**	Author collaboration with rules
		 Author Collaboration Network (109 authors, 334 collaborations)

Final Result
The final design is aiming to be a clean implementation that allows intuitive navigation. We provide a fixed top fixed navigation bat that provides quick access to our main sections:



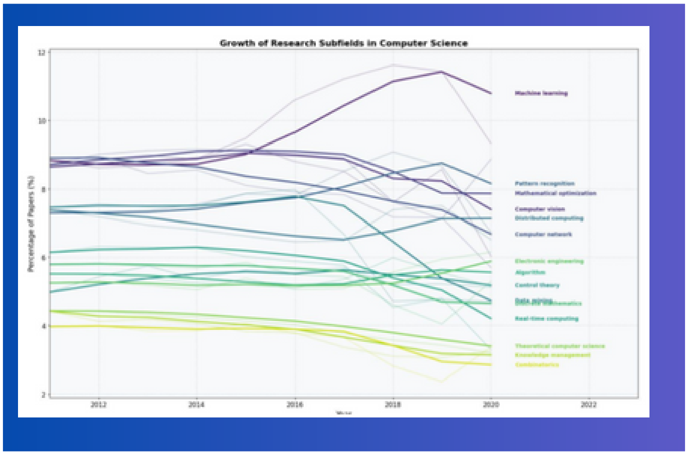
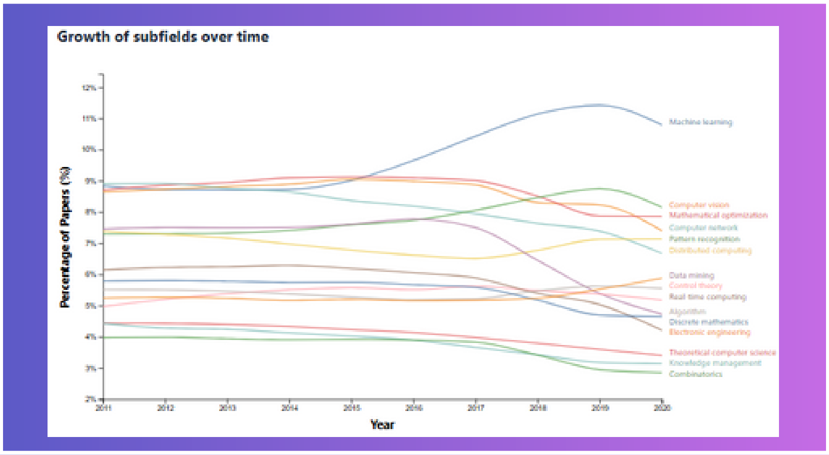
CS Hall of Fame

The sections provide hover effects, clickable elements and seamless scrolling and transitions that maintain user engagement. They were designed with our story in mind: Presenting major breakthroughs, a historical journey of the foundations, current trends then allowing the user to explore their own interests.
In Milestone 2, our core visualization’s goal was to provide a chronological view of the foundation, which we achieved with the interactive timeline that was already implemented at that point:

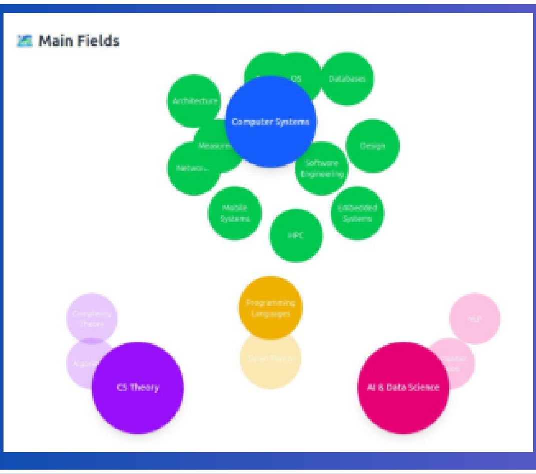
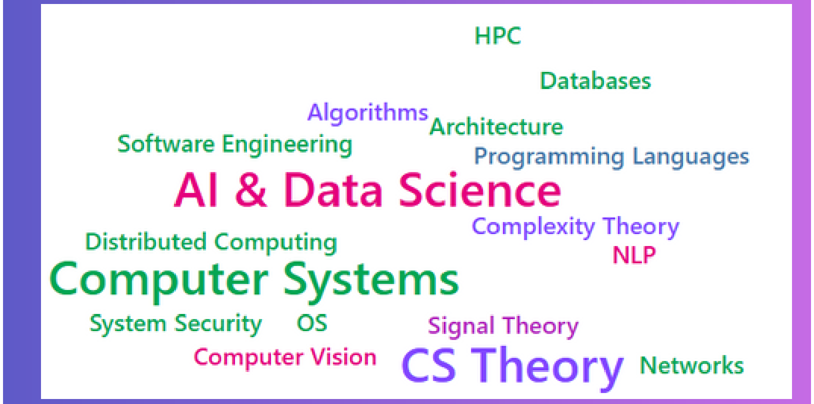


Hover over each dot in the timeline below to discover why each moment matters.

Our second goal was to make academic trends visible. This is achieved using a similar graph as in the Milestone 2 document but made visually better.

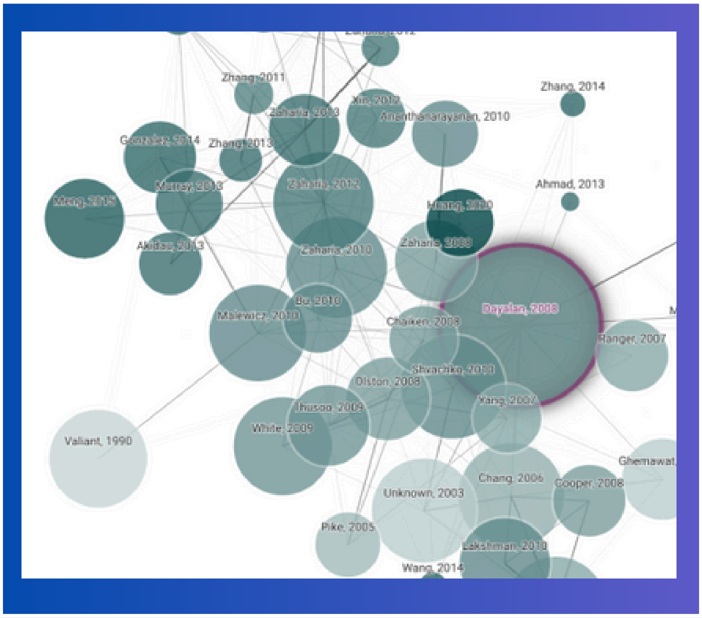
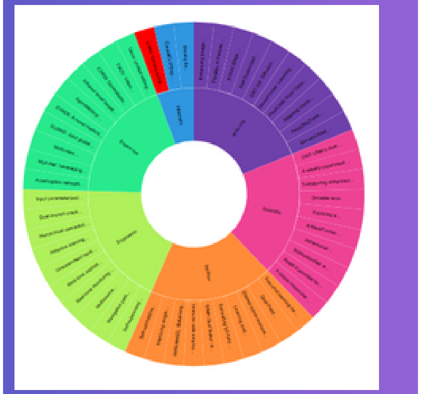
Milestone 2 Figure	Milestone 3 Implementation
 	

The last goal of the core visualization was to allow subfield exploration for users to focus on specific domains. Here, we decided to switch to word cloud visualization:

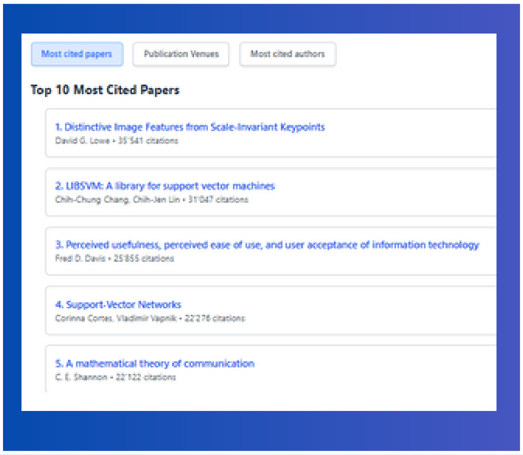
Milestone 2 Figure	Milestone 3 Implementation
 	 HPC Databases Algorithms Architecture Software Engineering Programming Languages AI & Data Science Complexity Theory Distributed Computing NLP Computer Systems System Security OS Signal Theory Computer Vision CS Theory Networks

We found this more aesthetically pleasing and minimalistic, which fits better with the design of the remaining parts. In Milestone 2, we aimed to provide the most cited papers and authors within a specified timeframe from each subfield. However, due to the data retrieval issues mentioned in the

challenges section, this was not feasible and we ended up picking one paper for each subfield to focus on. Our data story would be too repetitive if we include many papers for each subfield. We think picking one "favorite" paper from each subfield works out very well at the end.
Instead, we built a sunburst char to group the papers citing it by venue as well as an interactive author collaboration graph based on those same papers.

Milestone 2 Figure	Milestone 3 Implementation
 Zhang, 2011 Zhang, 2014 Xin, 2012 Zaharia, 2013 Ananthanarayanan, 2010 Gonzalez, 2014 Zhang, 2013 Meng, 2015 Murray, 2013 Zaharia, 2012 Ahmad, 2013 Huang, 2010 Akidau, 2013 Zaharia, 2010 Zaharia, 2008 Bu, 2010 Dayalan, 2008 Malewicz, 2010 Chaiken, 2008 Ranger, 2007 Shvachko, 2010 Olston, 2008 Yang, 2007 Thusoo, 2009 White, 2009 Valiant, 1990 Ghemawat Chang, 2006 Cooper, 2008 Unknown, 2003 Pike, 2005 Lakshman, 2010 Wang, 2014	

These ended up being an implementation of the "Additional Ideas" section of the Milestone 2 except the graph focused on authors instead of papers.
See if you are able to spot authors with great papers from EPFL 👀
We kept the idea of the Top Papers, Authors and Venues but decided to apply it to highly cited papers in computer science in general:



Top 10 Most Cited Papers

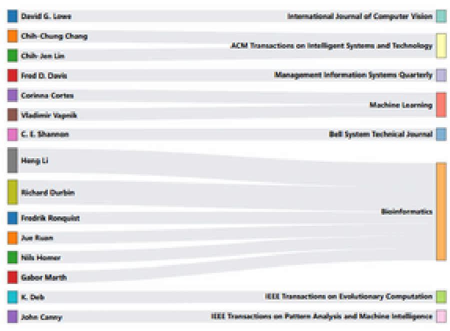
1. Distinctive Image Features from Scale-Invariant Keypoints
David G. Lowe • 35'541 citations

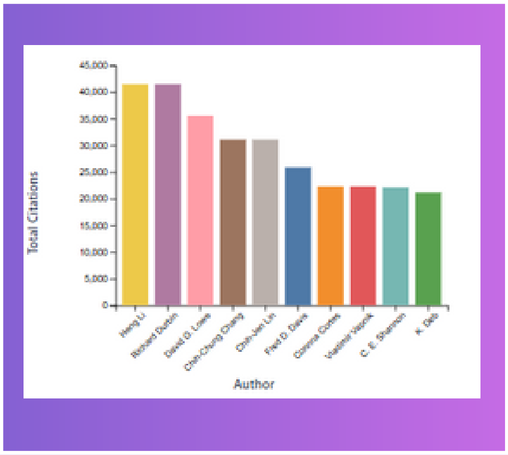
2. LIBSVM: A library for support vector machines
Chih-Chung Chang, Chih-Jen Lin • 31'047 citations

3. Perceived usefulness, perceived ease of use, and user acceptance of information technology
Fred D. Davis • 25'855 citations

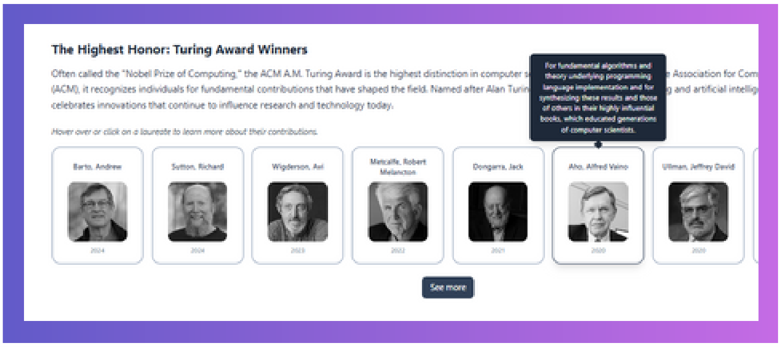
4. Support-Vector Networks
Corinna Cortes, Vladimir Vapnik • 22'276 citations

5. A mathematical theory of communication
C. E. Shannon • 22'122 citations





Finally, we implemented our last "Additional Ideas" visualization which was the Turing award Gallery:

Milestone 2 Figure	Milestone 3 Implementation
	 **The Highest Honor: Turing Award Winners** Often called the "Nobel Prize of Computing," the ACM A.M. Turing Award is the highest distinction in computer s … e Association for Com (ACM), it recognizes individuals for fundamental contributions that have shaped the field. Named after Alan Turin … g and artificial intellig celebrates innovations that continue to influence research and technology today. *Hover over or click on a laureate to learn more about their contributions.* For fundamental algorithms and theory underlying programming language implementation and for synthesizing these results and those of others in their highly influential books, which educated generations of computer scientists. Barto, Andrew 2024 Sutton, Richard 2024 Wigderson, Avi 2023 Metcalfe, Robert Melancton 2022 Dongarra, Jack 2021 Aho, Alfred Vaino 2020 Ullman, Jeffrey David 2020 See more

Reflection:
We are very satisfied with the final outcome and with having reached all of our goals that we set in Milestone 2 with the adequate modifications to create a coherent story and adapt to the datasets. We also managed to split the project work in a way that allowed us to work in parallel.
Next time, we will be more prepared for the reality of the data and the challenges of gathering it. We learned, early in the project, that we needed to be more flexible when choosing our datasets. We were initially set on solely using the Semantics Scholar API but it would have been a big limitation. Instead, we completed our dataset with some web scraping for the ancestral gallery and made design decisions based on what we can get.

Peer assessment:
Syrine was the main responsible of the data analysis, scripts retrieval and the initial network plots and Python based sketches that gave us an idea of what was possible or not with the data. Mariem and Daniel focused on the website code, working on different visualizations. We all participated equally in the overall design and story of the project.