**1) Challenges in Data Extraction**

* Balancing Time, Cost, and Technical Sophistication:
* Developing custom scrapers, highly time-consuming due to anti-bot measures
* Web-Scraping Services offer clean, structured data via APIs. Subscription-based and can become very expensive for large-scale data volumes
* Direct scraping is unreliable, often leading to incomplete data or access denial.
* Need headless browsers, proxy rotation, user-agent spoofing, and rate limiting.
* Lack of a single, universally adopted unique identifier for researchers across all academic platforms and publications. Variations in their name or aliases.
* Linking profiles/publications across different sources, data matching.

**2) Data We CAN Extract (or Access with APIs) with Unlimited Budget:**

**ACL Anthology (via acl-anthology library or direct dataset access):**

Highly structured, domain-specific (Computational Linguistics/NLP) open archive. Its strength is in the quality and consistency of its metadata.

Researcher Data:

* Canonical Name: The primary, standardized name for a researcher.
* Other Names/Aliases: Variations or former names.
* Custom ID: Unique identifier within the ACL Anthology.
* Affiliation: Current institutional affiliation.
* List of Publications: Direct links to their papers within the Anthology.

Paper Data:

* Title, Abstract, Author List (with IDs): Standard publication metadata.
* DOI (Digital Object Identifier): Persistent identifier for the paper.
* Publication Type: e.g., conference paper, journal article, workshop paper, thesis.
* Venue: Specific conference (e.g., ACL, EMNLP) or journal.
* Year: Publication year.
* Awards: not typically found on Google Scholar.
* Code Links: direct links to code repositories or datasets if submitted with the paper.
* Key Limitation: Doesn't provide aggregated citation counts, H-index, or i10-index, need to cross-reference with other sources.

**Google Scholar (via SerpApi):**

Comprehensive indexing of scholarly literature, strong on citations and researcher profiles, but limited in advanced analytics.

Researcher Profile Data:

* Google Scholar ID: Good identifier for individuals on Google Scholar.
* Canonical Name: As identified by Google Scholar.
* Affiliation: Current and potentially past affiliations.
* Verified Email Domain: If available and verified.
* Core Citation Metrics: Total citations, H-index, i10-index.
* Interests/Keywords: Self-declared (for research direction analysis).
* List of Publications: Title, authors, venue, specific citation count for each paper, publication year. Default top 100 sorted by cites, can use other filters
* List of Co-authors: Name, Google Scholar ID, affiliation…

Paper-specific Data (from general search results):

* Title, snippet/abstract, authors, source (journal/conference), year, citation count.
* Key Limitation: Indirect access to Impact Factors

**Scholarly (direct scraping of Google Scholar, need to address anti-bot measures):**

Open-source Python library. With an unlimited budget, you'd fund significant proxy infrastructure and CAPTCHA-solving integrations to overcome Google Scholar's anti-bot measures.

Researcher Profile Data:

* Same to what SerpApi provides from Google Scholar, as it scrapes the same source, less structured

Key Differences/Limitations:

* Proxy dependency, active maintenance of the scraping logic

**3) Data We Generally CANNOT/difficult to Extract even with unlimited budget**

* Journal Impact Factor (JIF): From Clarivate Analytics (Web of Science).
* CiteScore/SNIP/SJR: From Scopus.
* Specific Co-authorship Role
* Comprehensive Academic Service Records:
* Editorial Board Memberships: Specific journal roles.
* Conference Program Committee/Reviewer Roles: Detailed involvement in conference organization.
* Internal University Committee Work: Participation in departmental or university-wide committees.
* Non-Publication Awards and Honors:
* Specific courses taught, teaching evaluations, mentorship numbers (for professors, beyond co-authored student papers).
* Full patent details, claims, legal status, licensing info.

**4) Based on the Data We Can Realistically Gather, What Insights Can We Obtain?**

Individual Research Impact and Influence:

* Overall Impact using aggregated citation metrics (total citations, H-index, i10-index) from Google Scholar profiles.
* Most Influential Works: Identify a researcher's seminal contributions by sorting their publications by citation count.

Research Directions and Evolution over time

* Core Research Areas: Extract interests/keywords from Google Scholar profiles.
* Topic Modeling/Keyword Analysis to publication titles and abstracts to identify dominant research themes for individual researchers or groups.
* Research Transition/Progression: By analyzing topics over time (using publication year), observe shifts in a researcher's focus
* Niche vs. Broad Focus: Evaluate if a researcher's publications cluster around a narrow set of topics or span multiple distinct areas.
* Productivity

Collaboration Networks and Dynamics:

* Co-authorship Network Mapping
* Identifying Key Collaborators
* Community Structure: Identifying research groups or clusters within a larger network.
* Interdisciplinary and Institutional Collaboration: Analyze the affiliations of co-authors to identify collaborations across different departments, universities, or even countries.
* Collaboration Breadth vs. Depth: Quantify the number of unique co-authors (breadth) versus the average number of papers co-authored with each collaborator (depth).

Aggregated Insights (for groups/institutions):

* Departmental/Institutional Research Strengths: Aggregate interests/keywords and topic models from all researchers within a department/university to identify its core research specializations and comparative advantages.
* Collective Impact: Sum or average citation metrics for a group of researchers to assess the collective influence of a department or institution.
* Co-authorship across institutions to understand broader research ecosystems and partnerships.