Pablo Javier Barrio | Curriculum Vitæ

112 West 15th Street, Apt #4 – New York, NY 10011
+1 646-353-3057 • pjbarrio@cs.columbia.edu

Education

Columbia University — **New York, USA**
Ph.D. (Computer Science) — *Sep. 2009–Sep. 2015*
Adviser: Prof. Luis Gravano. Area of research: Information Extraction, Information Retrieval, Knowledge Discovery, and Data Mining. Thesis title: "Ranking for Scalable Information Extraction." Thesis committee: Chris Develder, Kathy McKeown, Ken Ross, Luis Gravano, and Panos Ipeirotis.

Columbia University — **New York, USA**
M.Phil. (Computer Science) — *May 2013*
Research Topics Covered: Query Processing in Databases, Transaction Processing in Databases, Database System Projects, Data Warehousing, Data Mining, Web Search, Text and Semistructured Data Retrieval, and Data Structures and Access Methods.

Columbia University — **New York, USA**
M.Sc. (Computer Science) — *Dec. 2011*
Relevant Courses: Natural Language Processing, Search Engine Technologies, Database Systems Implementations, Operating Systems, and Algorithms for Dealing with Massive Data.

National University of Central Buenos Aires — **Tandil, Argentina**
B.S. and M.S. in Software Engineering (Computer Science) — *Aug. 2009*
Relevant Courses: Data Structures and Databases, Operating Systems, Software Design and Architecture, Software Methodologies, Programming Paradigms (Object Oriented, Procedural, Functional, Aspect Oriented, and Logic), and Statistics.

Work experience

Google, Inc. — **New York, NY**
Software Engineer, Local Search — *Feb. 2016–Current*
Helping Local Search understand natural language queries and serve high-quality local results.

Columbia University — **New York, NY**
Researcher – Ph.D. Student, with Prof. Luis Gravano — *Sep. 2009–Sep. 2015*
Focused on information extraction, information retrieval, knowledge discovery, and data mining. (See Research Projects for more details.)

Bloomberg L.P. — **New York, NY**
Researcher (academic collaboration), with Knowledge Engineering Research Group — *Sep. 2014–Sep. 2015*
- Designed and developed algorithm for prioritizing fragments of text worth processing for given information extraction tasks.
- Evaluated algorithm over real news documents and with variety of information extraction tasks across diverse domains (e.g., finance, entertainment, business).

Microsoft Research — **New York, NY**
Research Intern, with Jake Hofman and Dan Goldstein. — *May 2014–Sep. 2014*
- Designed and developed crowdsourced system to generate contextual information around measurements and improve their understanding.
- Designed and performed large-scale evaluation on effectiveness of contextual information in terms of recall, estimation, and error detection.

Tenaris, Techint Group — **Buenos Aires, Argentina**
Internal Software Auditor Intern — *Jan. 2008–Mar. 2008*
- Developed data mining and integration solution for automatically generating auditing reports from various databases, text reports, and file formats.
- Integrated solution into supply chain system to support real-time, on-demand report generation.

ACOTEC Group — **Buenos Aires, Argentina**
Software Engineer — *Sep. 2006–Apr. 2007*
- Developed interactive, user-friendly system for controlling production volume in gear coupling product line.
- Developed system for automatically assembling gear couplings for given gear power, speed, direction demand.

Deloitte. — **Buenos Aires, Argentina**
Software Systems Auditor Intern — *Dec. 2006–Mar. 2007*
- Audited development cycle of software applications for banking within worldwide banking group for Sarbanes-Oxley Act (SOX) compliance.
- Devised and performed first fully automatic Segregation of Duties (SoD) analysis in Information Technology (IT) across all areas in large private pension fund.

Teaching experience

Columbia University — **New York**
Teaching Assistant, Advanced Database Systems — *Spring 2013*

- Assisted master-level students in development of data mining projects.
- Graded midterms and final exams for 80+ students.

National University of Central Buenos Aires — **Tandil, Argentina**
Teaching Assistant, Database Systems I[†]*, Design and Analysis of Algorithms II*[‡] — *Spring 2008*

- [†]Designed semester-long project for management of pharmaceutical industries.
- [†]Taught entity–relation model from theory to practice.
- [‡]Lead practical portion of class: Students were asked to develop multiple projects on C and C++.
- [‡]Designed and graded problem sets for 150+ students.

National University of Central Buenos Aires — **Tandil, Argentina**
Teaching Assistant, Design and Analysis of Algorithms I[†]*, Software Development Methodologies*[‡] — *Fall 2007 and 2008*

- [†]Taught 200+ students data structures in C and C++ and object oriented programming in C++.
- [†]Guided students through the design and development of challenging projects.
- [‡]Lead 20+ groups of students throughout development cycle of software system for administering driving tests.
- [‡]Taught various aspects of Unified Modeling Language (UML) (e.g., use cases, classes, and sequence diagrams) and on Agile Software Development.

Awards and Honors

The 2012 IOM–NAE Health Data Collegiate Challenge — **Institute of Medicine**
Washington D.C., USA — *2012*
Together with the School of Nursing at Columbia University, developed mobile application for tracking milestones, diagnostics, and appointments for women before, during, and after pregnancy. The application also served as a reliable source of information and social tool for interaction with health-care providers.

Academic Excellence — **Deloitte Foundation**
Buenos Aires, Argentina — *2006–2008*
Yearly nationwide award for academic and community commitment of students in economy- and technology-related careers.

Santander Río Academic Merit Award — **Santander Río Bank**
Buenos Aires, Argentina — *2007*
Nationwide award granted to students with the highest GPAs.

Commitment, Dedication, and Achievements in Studies — **Olavarría Intendancy**
Olavarría, Argentina — *2003*
Recognition to the top high-school graduates in all educational institutions.

Recognition to Project Leader — **Business Association of Olavarría**
Olavarría, Argentina — *2003*
Recognition for leading "Knowledge Sharing Academy" project. The project provided the infrastructure to enable highly qualified people in different areas (e.g., construction, manufacturing, life style) to teach—as well as to learn from—others about their areas of expertise.

Recognition to Innovation — **Olavarría Intendancy**
Olavarría, Argentina — *2003*
Together with professors in the National University of Central Buenos Aires, developed fully-functional educational software for primary school (1st through 8th grade) to learn and evaluate knowledge on math, linguistics and literature, natural sciences, and social sciences.

Languages

English: Full professional proficiency.

Spanish: Native speaker.

Technical Skills

Programming Languages: Java, C++, C, C#, Python, Scala, Bash, Prolog, Perl, AspectJ.

Markup Languages and Web: HTML, CSS, XML, JSON, Javascript, Ruby, PHP.

DBMS Languages and Usage: SQL, plSQL, postgreSQL, mySQL, Oracle DB, SQL Server.

Operating Systems: GNU/Linux, Windows, Solaris.

Others: LaTeX (Texts), R, OpenGL (Graphics), Matlab, UML.

Publications

Conference Papers

Pablo Barrio, Jake Hofman, and Dan Goldstein. Improving the comprehension of numbers in the news. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems, CHI 2016, San Jose, CA, May 7-12, 2016*, 2016.

Pablo Barrio, Gonçalo Simões, Helena Galhardas, and Luis Gravano. Learning to rank adaptively for scalable information

extraction. In *Proceedings of the 18th International Conference on Extending Database Technology EDBT 2015, Brussels, Belgium*, 2015.

Pablo Barrio, Jake Hofman, and Dan Goldstein. Improving the comprehension of numbers in the news. In *Proceedings of the Conference on Digital Experimentation, CODE 2015, Cambridge, MA, United States*, 2015.

Pablo Barrio, Jake Hofman, and Dan Goldstein. Improving the comprehension of numbers in the news. In *Computation+Journalism Symposium, C+J 2015, New York, NY, United States*, 2015.

Pablo Barrio, Luis Gravano, and Chris Develder. Ranking deep web text collections for scalable information extraction. In *Proceedings of the Twenty-third ACM International Conference on Conference on Information and Knowledge Management, CIKM 2015, Melbourne, Australia*, 2015.

Pablo Barrio, Gonçalo Simões, Helena Galhardas, and Luis Gravano. REEL: A relation extraction learning framework. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom*, 2014.

Journal Articles

Pablo Barrio and Luis Gravano. Sampling strategies for information extraction over the deep web. *IP&M*, 53(2):309–331, 2017.

Kathy McKeown, Hal Daumé III, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R. Fleischmann, Luis Gravano, Rahul Jha, Ben King, Kevin McInerney, Taesun Moon, Arvind Neelakantan, Diarmuid Ó Séaghdha, Dragomir R. Radev, Thomas Clay Templeton, and Simone Teufel. Predicting the impact of scientific concepts using full-text features. *JASIST*, 67(11):2684–2696, 2016.

Technical Reports

Pablo Barrio, Gonçalo Simões, Helena Galhardas, and Luis Gravano. REEL: A relation extraction learning framework. Technical report, 2014. `http://www.inesc-id.pt/ficheiros/publicacoes/10191.pdf`.

Theses

Pablo Barrio. *Ranking for Scalable Information Extraction*. PhD thesis, Columbia University, New York, NY, 2015.

Pablo Barrio and Juan Pablo Timpanaro. Hand–tracking: A virtual multi–functional mouse. Master's thesis, National University of Central Buenos Aires, Tandil, Argentina, 2009.

Research projects

Sentence Selection and Ranking for Efficient Information Extraction **Columbia University**
Information Extraction, with Dr. Anju Kambadur *2014–2015*
We developed a sentence-level ranking approach for improving the efficiency of running an extraction system over large text collections. We rely on an Orthogonal Matching Pursuit (OMP)-based approach that in addition to targeting recall—to prioritize sentences according to their usefulness—targets novelty—to prioritize sentences according to whether they produce unseen tuples. Furthermore, the ranking approach is able to revise the ranking decisions as it processes more sentences and intrinsic properties of useful sentences are revealed. Finally, the approach is able to learn—and seamlessly incorporate—certain important intuitions about useful sentences, such as the position where they appear in the documents or their expected number in useful documents.

A Crowdsourced System for Improving Numeracy **Microsoft Research**
Computational Social Sciences, with Dr. Jake Hofman and Dr. Dan Goldstein *2014*
During my internship at Microsoft Research NYC, I developed and tested a Web-based crowdsourced system to help people comprehend numerical measurements. In the system, crowd workers are shown actual measurements taken from the news and asked to create and vote on perspectives that employ percentages, ratios, rankings, or other comparisons to make the underlying measurements easier to understand. Based on user voting, the best perspectives are selected to appear within actual news articles as they are read. In principle, the system can annotate the daily news in a matter of hours. We tested the effectiveness of the system through a series of online experiments, which show that augmenting news articles with these perspectives improves the ability of people to understand the magnitude of numerical measurements. In particular, we observed that perspectives improve the ability of participants to recall measurements they have read, to estimate measurements they have not, and detect errors in manipulated measurements. We see this as the first of many steps in leveraging digital platforms to improve numeracy amongst online readers.

Ranking Deep Web Text Collections for Scalable Information Extraction **Columbia University**
Information Extraction, with Prof. Chris Develder and Prof. Luis Gravano *2014–2015*
State-of-the-art approaches for scaling the information extraction over deep web collections process focus on one text collection at a time. These approaches prioritize the extraction effort by learning keyword queries to identify the "useful" documents for the extraction task at hand, namely, those that lead to the extraction of structured "tuples." These approaches, however, do not attempt to predict which text collections are useful for the extraction task—and hence merit further processing—and which ones will not contribute any useful output to the task—and hence should be ignored altogether, for efficiency. In this work, we introduce and address the problem of ranking deep web collections for an extraction task, to prioritize the extraction effort by focusing on collections with substantial relevant information for the extraction task. We introduce statistically sound methods for effectively and efficiently estimating interesting measurements for a given extraction task, such as the number of useful documents, in deep web collections. We perform a large-scale experimental evaluation over real deep web collections, for several different IE tasks, and compare our approaches against baseline adaptations (to our setting) of state-of-the-art strategies in the (related) area of measurement estimations over indexes. Our evaluation shows that our methods enable the accurate computation of rich measurements, as they are able to collect a substantially larger number and diversity of useful documents than those of other approaches.

Learning to Rank in Information Extraction **Columbia University**

Information Extraction, with Gonçalo Simões, Prof. Helena Galhardas, and Prof. Luis Gravano *2013–2014*

Developed a principled, learning-based approach for ranking documents according to their potential usefulness for an extraction task. The low-overhead, online learning-to-rank methods exploit the information collected during extraction, as they process new documents and the fine-grained characteristics of the useful documents are revealed. These methods also decide automatically when the ranking model should be updated, hence significantly improving the document ranking quality over time. Our experiments show that our approach achieves higher accuracy than the state-of-the-art alternatives. More importantly, this approach is lightweight and efficient, and hence is a substantial step towards scalable information extraction.

Predicting Impact of Scientific Concepts using Full Text Features **Columbia University**

Inference on Scientific Literature, with Prof. Kathleen McKeown, Prof. Luis Gravano, among others *2013–2014*

Developed the information extraction building block for a system that predicts the future impact of a scientific concept, represented as a technical term, based on the information available from recently published research articles. The system analyzes the usefulness of rich features derived from the full text of the articles through a wide variety of approaches, including rhetorical sentence analysis, information extraction, and time-series analysis. The results from two large-scale experiments with 3.8 million full-text articles and 48 million metadata records support the conclusion that full-text features are significantly more useful for prediction than metadata-only features, and that the most accurate predictions result from combining the metadata and full text features. Surprisingly, these results hold even when the metadata features are available for a much larger number of documents than are available for the text features.

Learning Information Extraction Systems: The REEL Project **Columbia University**

Information Extraction, with Gonçalo Simões, Prof. Helena Galhardas, and Prof. Luis Gravano *2012–2013*

Together with a colleague from University of Lisbon, we developed REEL (**RE**lation **E**xtraction **L**earning framework), an open source framework that facilitates the development and evaluation of relation extraction systems by providing the code and infrastructure to: (i) handle various input text formats, which enables operations over different text collections; (ii) plug in appropriate text processing steps and tools, which enables diverse processing of the text with minimal effort; (iii) define and combine conceptual relation constraints that are automatically enforced; (iv) decouple learning and extraction from the text processing, which enables the straightforward integration and reusability of different extraction algorithms; and (v) uniformly execute and evaluate relation extraction systems, which enables the testing and fair assessment of these systems. The framework is open source and publicly available at *http://reel.cs.columbia.edu/*.

Sampling for Information Extraction over the Deep Web **Columbia University**

Information Extraction, with Prof. Luis Gravano *2010–2012*

Systematically studied the space of query-based document sampling techniques for information extraction over the deep web. Specifically, we considered (i) alternative document retrieval and processing schedules, which vary on how they deploy the extraction effort over documents, and (ii) alternative query execution schedules, which vary on how they account for the query effectiveness. We conducted the first (to the best of our knowledge) large-scale experimental evaluation over real deep-web databases and using several different extraction tasks. We showed a number of sampling techniques that significantly outperform the state of the art in terms of sample quality and efficiency and can, in turn, help improve the efficiency of the overall extraction process.

Hand–Tracking: A Virtual Multi–Functional Mouse (Master's Thesis) **National University of Central Buenos Aires**

Computer Vision, with Juan Timpanaro, Dr. José Massa, and Dr. Paula Tristán. *2009*

Developed a toolkit for Human–Computer Interaction (HCI) that emulates the mouse—along with its conventional applications—via hand-tracking using an inexpensive Web cam. The toolkit first detects the position (i.e., X and Y coordinates) and form (e.g., closed or open) using a properly trained 3-layered Perceptron Neural Network that receives as input the pixels captured by the Web cam. These coordinates are tracked over time and mapped in real-time to the position of the mouse cursor. Additionally, the toolkit allows for the creation and matching of dynamic gestures, such as slide to the sides or closed circle, that the user can associate with complex functionalities, such as dragging and dropping elements, changing focus of images and maps, and opening specific applications. For this, the system recognizes via Dynamic Time Warping (DTW) whether the last tracked movements correspond with any of the created patterns and, if so, triggers the associated action. The algorithms respond efficiently—to trigger commands in real-time—and generalize well to users with different characteristics and motion capabilities—to provide an accessible and reliable interaction with the computer. The source code is open source and freely available at *https://code.google.com/p/mythesisproject/*