Xin Luna Dong

http://lunadong.com E-mail: lunadong@gmail.com
Seattle, WA Phone: (201) 650-3494

Highlights

- World renowned expert in **knowledge graph** and **data quality**. Ten years of experiences in building the world's largest generic and product knowledge graphs at Google and Amazon.
- Experience in starting a project from scratch and growing a team from 3 to 70+ people; collaborating across orgs at Amazon Retail, Digital, Voice to support 30+ downstream applications and industry first experiences.
- Science leader for end-to-end science innovation cycles including identifying customer needs, inventing state-of-the-art ML techniques, implementing and deploying production features, and launching new customer experiences.
- High visibility in multiple science communities including **data mining, NLP,** and **databases**. Co-authored 2 books on knowledge graph and big data integration; co-chaired 3 top-tier conferences; 60+ keynotes/invited-talks/tutorials; 100+ papers in top academic conferences and journals. ACM Distinguished Scientist and VLDB Early Career Research Contribution Award.

Experiences

Senior Principal Scientist, Amazon 7/2016-Present

- Founded the Product Graph project to build an authoritative knowledge graph for retail and digital products, grew the team from 3 to 70+ people, and playing a key role in technical vision, architect, and data deliveries. The product graph has been used in 30+ downstream applications across Retail, Amazon Digital, and Alexa, increased Amazon revenue by hundreds of millions of dollars, and supported an industry-first experience. [Podcast]
- Founded and playing a key role in leading the efforts on web-scale knowledge extraction and integration for Alexa knowledge graph. Improved web extraction accuracy from 60% in state-of-the-art to 90%+ to support production data deliverables. [Keynote]
- Growing a science team to conduct cutting-edge research in knowledge extraction, integration, cleaning, and mining. The team has published 20+ papers and 8 tutorials in top-tier conferences in NLP, Web, data mining, and, and filed 1 patent.
- General chair for Amazon Machine Learning Conference in 2017.

Senior Research Scientist, Google 2013/1-2016/6

- Built a scalable system to collect long-tail knowledge. The system interacts with ~20 other systems, collecting tail knowledge from raw data and triggering knowledge panels in Google Search. Leading ML efforts in the project and playing a key role in data quality control.
- Invented Knowledge-Based Trust. Tech lead for a system to decide truthfulness of facts and trustworthiness of web sources, and implemented the core algorithms. The system is serving multiple products, ranging from global and personal knowledge extraction, fact verification, to source recommendation. [Washington Post article]
- Initiated the Knowledge Fusion research area and contributed to the Knowledge Vault project, which automatically collects knowledge from various types of web contents on billions of web pages, and fuses them into a probabilistic knowledge base containing billions of knowledge facts.

Senior Member of Technical Staff, AT&T Research. 2007/10-2012/12

- Initiated the Data Fusion research area--resolving conflicts from multiple data sources. Built the system SOLOMON, which decides correctness of facts, trustworthiness of data sources, and copying relationship between sources.
- Researched on record linkage where tolerance to high value variety is required, such as linking records incrementally, linking business listings that belong to the same business chain, linking temporal records that describe the same entity over time.

Research Assistant, Univ. of Washington 2002/10-2017/8

- One of the first people researching on Dataspaces--data integration methodology that automatically bootstraps approximate schema-mapping and entity-resolution services, and evolves the performance over time in a pay-as-you-go fashion. Published a paper in VLDBJ Special Issue of "Best papers of VLDB 2007".
- Built Semex Personal Information Management system, which generates and manages personal knowledge graph from personal data. The system won Best Demo Award in Sigmod'2005.
- Built the Woogle Web-Service Search Engine to provide novel mechanisms for searching and composing web services. The paper has 750+ citations.

Education

Ph.D. in Computer Science	**University of Washington,** 2001-2007
M.S. in Computer Science	**Peking University,** 1998-2001
B.S. in Computer Science, A.S. in International Finance	**Nankai University,** 1994-1998

Honors

- ACM Distinguished Member, 2018 *(For "Significant contributions to data and knowledge integration")*
- VLDB Early Career Research Contribution Award, 2016 *(For "Advancing the state of the art of knowledge fusion")*
- Best Demo Award in ACM Int. Conf. on Management of Data (SIGMOD), 2005 *(One of the top 3 demos in that year)*

Major Professional Activities and Services

- Board of Trustees of the *VLDB Endowment.*
- Member of *PVLDB Advisory Committee, VLDB Strategy Committee, TCDE Award 2017-2019, CIKM Best Paper Award 2017*
- Chair of DBCares, with the mission to create an inclusive and diverse DB community with zero tolerance for abuse, discrimination, or harassment.
- PC co-chair: *WSDM'2022, VLDB'2021, SigKDD'2020 Applied Data Science Invited Talks, Sigmod'2018*

Major Talks

- "Building the Product Knowledge Graph at Amazon w. Luna Dong". Podcast at *TWIML (This week in Machine Learning)*
- "Ceres: Harvesting Knowledge from Semi-Structured Web". Keynote at *CIKM'2020*.
- "Building a Broad Knowledge Graph for Products". Keynote at *ICDE'2019*, Invited talk at *KDD'2018 Applied Data Science Invited Talk Series.*
- "How Far Are We from Collecting the Knowledge in the World". Keynote at *ICWE'2016*.

Latest Major Publications

Book:

- G. Weikum, Xin Luna Dong, S. Razniewski, F. Suchanek. Machine Knowledge: Creation and Curation of Comprehensive Knoweldge Bases. To appear in 2021.
- Xin Luna Dong, and D. Srivastava: Big Data Integration. Morgan Claypool Publisher, 2015.

Tutorials

- C. Lockard, P. Shiralkar, Xin Luna Dong, H. Hajishirzi. Web-scale knowledge collection. In *WSDM'2020, ACL'2020, KDD'2020.*
- Xin Luna Dong, C. Faloutsos, A. Kan, J. Ma, S. Mukherjee. Graph and tensor mining: for fun and for profit. In *SigKDD'2018*.
- Xin Luna Dong, C. Faloutsos, X. Li, S. Mukherjee, Prashant Shiralkar. Fact checking: theory and practice. In *SigKDD'2018*.
- Xin Luna Dong, T. Rekatsinas: Data integration and machine learning: a natural synergy. In *SigKDD'2019, VLDB'2018, Sigmod'2018.*

Selected Papers (H-Index: 42; Full list)

- [eComm Knowledge Graph] Xin Luna Dong, X. He, A. Kan, X. Li, Y. Liang, J. Ma, Y. Xu, C. Zhang, T. Zhao, G. Saldana, S. Deshpande, A. Manduca, J. Ren, S. Singh, F. Xiao, H. Chang, G. Karamanolakis, Y. Mao, Y. Wang, C. Faloutsos, A. McCallum, J. Han. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. In *SigKDD'2020*.
- [Knowledge extraction] C. Lockard, P. Shiralkar, H. Hajishirzi, Xin Luna Dong. ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages. In *ACL'2020.*
- [Knowledge integration] Qi Zhu, Hao Wei, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos, Jiawei Han. Collective multi-type entity alignment between knowledge graphs. In *WebConf'2020*.
- [Knowledge fusion] Xin Luna Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: estimating the trustworthiness of web sources. In *VLDB'2015*.
- [Knowledge fusion] Xin Luna Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In *SIGKDD'2014.*

Patents (8 in total)

- "Similar but Different (SBD): Presenting Item Recommendations in Dynamically Generated Groups with Explanations," Andrey Kan, Christos Faloutsos, and Xin Dong, United States Patent 10.891.676, issued 1/2021.