Preface

Domain-specific knowledge graphs have emerged as a field unto their own, steadily and perhaps not so slowly. Graphs have been pervasive in AI for a long period of time, dating back to the earliest eras in the field, but automatically representing large quantities of data as graphs is a relatively modern invention. With the advent of the Web, and the need for smarter search engines, both Google and (over a decade later) the Google Knowledge Graph were born. The Google Knowledge Graph has changed the way we interact with search engines, even though we often do not realize it. For example, it is not uncommon anymore for users to not click on a single link when they are searching for something; generally, the search engine itself is able to provide the solution for the problem the user seems to be facing. Organic integration of the traditional search engine with images, news, and videos has only added an element of richness to these interactions.

For all its success, the Google Knowledge Graph (and other similar efforts) was not designed with a specific domain in mind, although Google has rolled out flavors of "domain-specific search" engines (e.g., Google Scholar) every now and then. One would almost be forgiven for thinking that building domain-specific systems, powered by knowledge graphs, for problems such as geopolitical event forecasting, or academic literature mining, is too esoteric to come into its own as an independent, impactful area of study.

What has changed the game and made researchers (and customers) look at domain-specific knowledge graphs as a viable technology is that it has become *easier* to build such knowledge graphs, starting from data collection all the way to the application interface. This was not always the case. Only a few years ago, if I wanted a domain-specific knowledge graph for the e-commerce domain, for example, I would have to assemble a team and build out a system for months before anything remotely viable would emerge. The DARPA Memex program has had an enormous impact in changing this sad state of affairs, by allowing the *democratization* of domain-specific knowledge graph construction. Technologies that emerged from the Memex program combined both classic and state-of-the-art techniques in fields as diverse as information extraction and entity resolution to produce end-to-end systems that could be used by *nontechnical* domain experts to

build entire search engines powered by knowledge graphs. A lot of the work that we describe here was rediscovered and utilized in the Memex program to build these end-to-end systems.

Some of the fields that I mentioned above, such as information extraction and entity resolution, are entire areas of study in their own right, with numerous surveys and books individually covering them. Thus, I have had to make some necessary trade-offs in writing this book, and I have chosen to focus on breadth, and comprehensiveness, rather than depth and full academic rigor. In other words, what I attempt to provide in this short work is a comprehensive, practical methodology for constructing domain-specific knowledge graphs using the full range of technology that is available today. I do not shy away from the truism that in many cases, there are no right solutions; one has to deal with compromises. This book tries to detail what these compromises are and when it makes sense for someone wishing to construct domain-specific knowledge graphs to adopt a particular technology or technique.

Since the book is largely based on the findings of multiple communities, there is a lot of credit to go around in conveying the content of each chapter. In some cases, such as IE, I have drawn broadly on widely cited reviews of the field by merging and conveying key elements of both classic and modern surveys, to give the reader a sense of both new developments and established techniques. Because this book is only meant to be a condensed, though hopefully practical and relatively comprehensive, introduction to the field, I have not attempted to provide a rigorous citation for every system or statement. Rather, at key junctures, I have provided pointers to the broader sources that provide a much more comprehensive treatment of related work for the more technically oriented researcher.

I am fairly confident that this book will not provide the last word on this subject. All indicators suggest that research on knowledge graph construction is intensifying, and with increasing synergies between natural language processing, deep learning, knowledge discovery, and semantic web, we will likely see some exciting new work emerge in the years to come. At the time of writing, it is safe to conclude that the field stands at an exciting junction.

Marina del Rey, CA, USA Mayank Kejriwal
December 2018