Deep Learning on Structured Data
Mark Ryan

1. Tell us about yourself.

- *What are your qualifications for writing this book?*
    - Masters of Science in Computer Science from the University of Toronto
    - 20 years leading teams delivering IBM's premier relational database product: Db2
    - Conference presentations on machine learning:
        - Machine learning bootcamp at CASCON conference (Oct 2018):
        - Machine learning with DSX and Db2 at IDUG conference (April 2018):
    - Extensive hands-on experience applying deep learning to practical business problems involving structured data (described in posts https://medium.com/@markryan_69718)
    - Extensive experience as a consumer of machine learning / deep learning educational material, including:
        - Deep Learning Specialization by Andrew Ng
        - Fast.ai by Jeremy Howard
- *Do you have any unique characteristics or experiences that will make you stand out as the author?*
    - Extensive experience technical writing (e.g. Getting Started book for IBM VisualAge for Java)
    - Professional experience that spans the worlds of relational database and deep learning
    - Empathy for the reader.
        - I anticipate a significant portion of the audience for this book will be coming from a background of traditional structured database. I also come from this background, and because of that I believe that my experience will be both relevant and credible to this audience.

2. Tell us about the book.

- *What is the technology or idea that you're writing about?*
    - Applying deep learning techniques to problems that involve structured data (such as the data found in relational databases)
- *Why is it important now?*
    - Educational material on Deep learning focuses on applications on unstructured data (images, audio, text).
    - In many professional settings, people work primarily with structured data. Examples of deep learning that focus on images aren't as relevant to these people as examples that deal with structured data: tables, rows, columns.
    - Based on the views & feedback I've received on my blogs about

deep learning on structured data, there's an audience for more details on using deep learning on structured data.

- *In a couple sentences, tell us roughly how it works or what makes it different from its alternatives? Feel free to use pictures.*
    - Deep learning (DL) is a machine learning approach that exploits a multi-layered architecture to train a model based on input data. This model then can be used to make predictions on / analyze new data that the model has never seen. Unlike other machine learning approaches, deep learning can take advantage of a high degree of automation and can define a model that accurately identifies extremely complex relationships in the training data.
    - When it comes to analyzing / making predictions based on structured data, there are two significant alternatives to deep learning:
        - Non-DL machine learning: e.g. linear regression, logistic regression, random forest
        - Traditional Business Intelligence (BI)
    - Deep learning provides the following benefits over non-DL machine learning approaches:
        - Performance improves as more data is available for training the model
        - Less demand for feature tuning (time-consuming and error-prone tinkering with the input data in order to get adequate performance from the model)
        - Superior overall performance, given an adequate volume of training data, sufficient compute resources, and parameter tuning
    - Deep learning provides the following benefits over traditional BI techniques:
        - Greater flexibility to adapt to data changes, including schema changes (e.g. new columns) and overall workload changes (e.g. spike in sales in a given jurisdiction)
        - Less code to maintain
        - Potential to automate detection of new benefits (e.g. crawling catalog tables to detect unanticipated opportunities to accurately predict valuable outcomes from available data)
- *What type of book are you planning to write?*
    - Is your book a tutorial or a reference?
        - Tutorial
    - Does this book fall into a Manning series such as In Action, In Practice, Month of Lunches, or Grokking?
        - Deep learning
    - Are there any unique characteristics of the proposed book, such as a distinctive visual style, questions and exercises, supplementary online materials like video, etc?
        - I plan to use a complete, end-to-end code example with

open source data to illustrate the concepts and practice of applying deep learning to structured data
- Because the data & code will be openly available, the reader will be able to follow along and developer their own variations on the code as they are guided through an exploration of the data and the deep learning code that consumes the data

3. *Please give us 5 or 6 representative tasks in the domain of your book.*

1. Classify input columns as one of the following:
   a. continuous (e.g. prices, temperatures),
   b. categorical (e.g. days of the week, states in the US, countries)
   c. text
2. General data cleanup - write and execute code to deal with problems in the data such as: missing data, type mismatches (e.g. string values in a column you expect to be numeric)
3. Prepare categorical columns – for columns with categorical values, replace string values with tokens
4. Prepare text columns – for columns with multi-word text values, perform basic text preparation (e.g. convert to all lower case, remove stop works and punctuation) and replace strings with tokens
5. Define model input data structure: use the continuous, categorical, and text columns that you have prepared to create the input structure for the deep learning model
6. Define deep learning model: use the input data structure you defined in the previous task to generate the embedding, RNN, and other layers for your model.

4. The minimally-qualified reader (MQR)

You need to understand clearly what you're assuming your reader already knows before beginning the book. The Minimally Qualified Reader (MQR) represents a reader who possesses only the lowest level of pre-requisite skills and knowledge. For example, if you're intending that your book is appropriate for a beginning JavaScript programmer, you probably can't assume knowledge of something like promises or closures.

- Beginner-Intermediate skill in Python
- Beginner skill in Jupyter Notebooks
- Beginner-Intermediate skill in non-DL machine learning
- Beginner skill with at least one of the common Cloud environments (e.g. AWS, Azure, IBM Cloud)

What is the primary job role of your readers?

- The minimally-qualified reader will be an IT professional (coder, DBA, data scientist) with at least 2 years experience

- This book would also be appropriate for business stakeholders who are familiar with coding and basic relational database and want to understand the potential of applying deep learning to structured data

E.g.

- *The minimally-qualified reader will be a system administrator with at least four years experience*
- *This book would also be appropriate for developers or DevOps engineers who need to administer Linux systems*

What do you expect your MQR to already know before they start reading?

- Beginner-Intermediate skill in Python, including
  - Type handling
  - Basic constructs (e.g. lists, dictionaries)
  - Control structures (e.g. if, while) and operators
  - Numpy
- Beginner skill in Jupyter Notebooks
  - Code vs. markup cells
  - Executing and debugging code cells
- Beginner skill in non-DL machine learning
  - Familiarity with basic algorithms: e.g. Logistic regression, linear regression, random forest
  - Rough grasp of mathematical concepts (linear regression, basic calculus) sufficient to understand the "secret sauce": define a function to capture the error (delta between the actual outcome for a given set of incomes and the outcome predicted by the model) and use the derivative of that function to adjust the parameters to minimize the error
- Beginner skill with at least one of the common Cloud environments (e.g. AWS, Azure, IBM Cloud)
  - Has gone through the process of setting up an account on one of the mainline Cloud providers

5. Q&A

What are the three or four most commonly-asked questions about this technology? What are others interested in this topic asking in forums?

- How is deep learning different from other forms of machine learning?
- What are the advantages and disadvantages of deep learning compared to non-DL machine learning approaches?
- Why use deep learning with structured data? Aren't there simpler / cheaper ways to get the same result?

- Deep learning needs tons of data: how can you get enough structured data for a given problem to be adequate to train a deep learning model?
- Where can I find an open structured data sources to learn with?
- Do Kaggle competitors use deep learning for problems involving structured data?
- What kinds of problems with structured data lend themselves to a deep learning approach?

6. Tell us about the competition and the ecosystem.

- *What other books are available on this topic?*
  - There are lots of books on Deep Learning (e.g. Deep Learning, Introduction to Deep Learning, Deep Learning with Python) and some books (e.g. Deep Learning: Natural Language Processing in Python) and courses (e.g. Natural Language Processing with Deep Learning in Python) that cover key concepts (for deep learning on structured data) like embeddings
  - However, I am not aware of any other books that focus Deep Learning on structured data specifically
- *How does the proposed book compare to them?*
  - This book is specifically focused on deep learning with structured data, and covers an end-to-end practical example (including open data & code) of applying deep learning to structured data
  - This book is aimed at a professional, rather than academic audience. While it will introduce theoretical topics, the focus of the book is practical application using an accessible stack (Python developed in Notebooks with Keras as the deep learning framework)
- *What resources would you currently recommend to someone wanting to learn this subject?*
  - Sections of the Fast.ai course by Jeremy Howard
- *What are the most important web sites and companies associated with this topic?*
  - I am not aware of any web sites specifically dedicated to deep learning on structured data
  - This article indicates that the majority of commercial applications of deep learning are on unstructured (e.g. image) data
  - This article covers some related ideas about applying ML to yield structured data from unstructured data
- *Where do others interested in this topic gather online?*
  - Medium – posts on this topic include Using Deep Learning for Structured Data with Entity Embeddings and Structured Deep Learning
  - The forum for Jeremy Howard's fast.ai course has some interesting discussions on the topic, including this thread

7. Book size and illustrations

Please estimate:

- The approximate number of published pages to within a 50-page range
    - 300 pages
- The approximate number of diagrams and other graphics
    - 30
- The approximate number of code listings
    - 40

8. Contact information

Formal Name: Mark B.P. Ryan
Name: Mark Ryan
Mailing Address:
1398 Rennie Street
Oshawa, Ontario
Canada
L1K 0H1
Preferred email: ryanmark@rogers.com
Preferred phone: 416 434 9173
Skype: live:ryanmark_3
Website/blogs/Twitter, etc: blogs: https://medium.com/@markryan_69718
LinkedIn: https://ca.linkedin.com/in/mark-ryan-31826743, Twitter: @MarkRyanMkm

9. Schedule

- *Most authors require 2-4 weeks to write each chapter. Please estimate your writing schedule*

    Chapter 1:
    - Mid-Jan 2019

    1/3 manuscript:
    - End of Feb 2019

    2/3 manuscript:
    - Mid-April 2019

    3/3 manuscript:
    - End of May 2019

- *Are there any critical deadlines for the completion of this book?*
    - Not that I am aware of.

10. Table of Contents

While every Table of Contents is different, there are a few common best practices for a typical In Action book.

Chapter 1: The first chapter is typically an overview of the book's topic that tells the reader what the technology is, how it works, and orients the reader to its practical benefits.

Chapter 2: The second chapter is often a brief tutorial example that shows what the technology looks like in actual use. The reader isn't expected to understand everything in this example, and it should be simple enough to show just the important characteristics.

The next several chapters should cover the core aspects of the technology that all users must know.

In later chapters, you may cover topics of interest to only certain segments of your audience.

Formatting the Table of Contents

Your ToC should look something like the sample attached. Please number your Table of Contents in the same way

- *Whenever possible, the first section of every chapter should be a gentle overview of prerequisites for the chapter, a what-you-need-to-know to read the chapter.*
- *The final section of each chapter is a Summary.*
- *You may include a brief annotation for each chapter, but the topics covered in the ToC should be clear without the annotation.*
- *Every chapter should have clear objectives. "In this chapter, the reader will learn how to…"*

Table of Contents
Deep Learning on Structured Data