Supporting the New Techniques of Knowledge Discovery in Databases: TDM & AI

Jake Zarnegar
Chief Product Officer, Silverchair
World’s Dumbest Diagram

Researchers on a noble quest to create new knowledge & better understand our world

Trusted evaluators, publishers, and archivists of research findings

Networked research content hosting and distribution platforms
Discovering our discoveries*

*World’s dumbest catchphrase

Knowledge Discovery in Research Databases:

- **Traditional** knowledge discovery hypothesis: “Some questions are answered by highly-educated human specialists finding, reading, and synthesizing the right few pieces of information”
  - *Meta-Analysis, Systematic Review, Cohort Study*

- **TDM/AI** knowledge discovery hypothesis: “Some questions can only be answered by specialist-directed algorithms that synthesize thousands, millions, or billions of pieces of information”
Today’s Question

How can scientific and scholarly publishers best prepare and deliver their content to assist (and not hinder) TDM/AI knowledge discovery?
But First

How do TDM/AI knowledge discovery projects actually work?
Knowledge Discovery in Databases (KDD) [1] is divided in four main phases: domain exploration, data preparation, data mining, and interpretation of results.

1. The first phase is responsible for understanding the problem and what data will be used in the knowledge discovery process.

2. **The next phase selects, cleans, and transforms the data to a format that is suitable for a specific data mining algorithm.**

3. In the third phase, the chosen data mining algorithm performs some intelligent techniques to discover patterns that can be of potential use.

4. The last phase is responsible for manipulating the extracted patterns to generate interpretable knowledge for humans...
Most of the research carried out in this area focus on the data mining phase, which uses artificial intelligence algorithms like decision trees, artificial neural networks, evolutionary computation, among others [2] to discover knowledge. On the other hand, the data preparation phase, responsible for integration, cleaning, and transformation of data, has not been the subject of much research. In fact, Pyle [3] argues that “data preparation consumes 60 to 90% of the time needed to mine data – and contributes 75 to 90% to the mining project’s success”.

“I downloaded 2TB of Arxiv content last week but I can’t bring myself to open it and start working on analyzing it because I know I have at least 6 months of painstaking content cleanup & preparation ahead of me before I can begin.”

--Mike M., Fast Forward Labs
And Second

What are TDM/AI developers trying to achieve?
Artificial intelligence: don't believe the hype | WIRED UK
www.wired.co.uk/article/sensationalism-ai-hype-innovation
Feb 18, 2017 - Hype could be of significant short-term benefit to artificial intelligence research, but prove fatal in the long run. Luke Dormehl talks to WIRED ...

Myth Busting Artificial Intelligence | WIRED
https://www.wired.com/insights/2015/02/myth-busting-artificial-intelligence/
We've all been seeing hype and excitement around artificial intelligence, big data, machine learning and deep learning. There's also a lot of confusion about ...

The Hype—and Hope—of Artificial Intelligence - The New Yorker
Aug 26, 2016 - Om Malik on the hype about artificial intelligence, and the three stages of A.I.: recognition intelligence, cognitive intelligence, and virtual ...

'Artificial Intelligence' was 2016's fake news • The Register
https://www.theregister.co.uk/2017/01/02/ai_was_the_fake_news_of_2016/
Jan 2, 2017 - As with the most cynical (or deranged) internet hypesters, the current "AI" hype has a grain of truth underpinning it. Today neural nets can ...
Much like “the cloud,” “big data,” and “machine learning” before it, the term “artificial intelligence” has been hijacked by marketers and advertising copywriters.

If the hype leaves you asking “What is A.I., really?,” don’t worry, you’re not alone. I asked various experts to define the term and got different answers. The only thing they all seem to agree on is that artificial intelligence is a set of technologies that try to imitate or augment human intelligence.

To me, the emphasis is on augmentation, in which intelligent software helps us interact and deal with the increasingly digital world we live in.

Om Malik, The New Yorker

Two Types of Augmentation

General Augmentation
- Amazon Alexa
- "Ok Google"
- Cortana

Specific Augmentation
Received November 28, 1769.

VIII. Account of a very remarkable young Musician. In a Letter from the Honourable Daines Barrington, F. R. S. to Mathew Maty, M. D. Sec. R. S.

S I R,

If I was to send you a well attested account of a boy who measured seven feet in height, when he was not more than eight years of age, it might be considered as not undeserving the notice of the Royal Society. The instance which I now desire you will communicate to that learned body, of as early an exertion of most extraordinary musical talents, seems perhaps equally to claim their attention.

Joannes Chrysostomus Wolfgangus Theophilus Mozart, was born at Salzburg in Bavaria, on the 17th of January, 1756.*

- We’re pretty good at this*
- The fundamentals of a well-indexed, permanent scholarly record (DOI, WoS, CLOCKSS, etc.)

*For journals, at least
Also strong in creating interfaces that assist understanding from human readers.
Fostering Software Understanding

In some ways we've got a good foundation – detailed, consistent content tagging to aid with software Understanding

• Structure understanding through normalized XML: what is the title, authors, abstract, where the conclusions are in the paper, etc.

Increased indexing of named entities: understanding what is a gene, what is a clinical trial ID, what is a person
What We Already Know about This Topic

- Glucocorticoids are commonly given to prevent nausea and vomiting
- However, glucocorticoids are immunosuppressive and may promote infection
- The authors conducted a meta-analysis of 56 trials (n = 5,607) that evaluated infection, hospital duration, and intraoperative glucose concentration

What This Article Tells Us That Is New

- Glucocorticoids did not impact on any wound infection (odds ratio, 0.84; 95% CI, 0.62 to 1.15) or length of stay (weighted mean difference, −0.27 days; CI, −1.37 to 0.84)
- Glucocorticoids slightly increased peak postoperative glucose concentrations by 20 mg/dl (CI, 11 to 29; P < 0.001), an amount that is probably not clinically important
- Single-dose steroid administration for prevention of nausea appears safe

Rheumatology key messages

- Clinical manifestations of Ebola viral disease include muscle pain and arthralgia.
- Musculoskeletal complications post-Ebola viral disease include muscle pain, arthritis, enthesitis and tendon ruptures.
- Immunological mechanisms are postulated for the musculoskeletal manifestations of Ebola viral disease.


OK, Back to Today’s Question

How can scientific and scholarly publishers best prepare and deliver their content to assist (and not hinder) TDM/AI knowledge discovery?
4 Obstacles to Accelerated TDM/AI Knowledge Discovery

1: Interfaces still primarily visual w/narrative text

2: Normalized, detailed underlying XML structure & metadata not shared

3: Limited indexing/tagging above the “Understanding” level

4: Limited ability to take away large amounts of content quickly
Consider Providing Structured Content as a New Product

- Full-text normalized XML (or JSON)
- Separate product/subscription for sale
- Enrich it further or sell it “as is”
- Separate delivery mechanism (no human interface) but can piggyback on existing content workflows
- Accesses a new class of customer w/deep pockets (AI creators or implementers)
- Requires new vetting/legal agreements
We’re Piloting This Now

• We have two pilot partners that will be launching TDM products in early 2018 that includes pre-packaged metadata downloads for subscribers and full-text packages for purchase
Remember the Business Justification

“data preparation consumes 60 to 90% of the time needed to mine data – and contributes 75 to 90% to the mining project’s success”.

Or Create Your Own Derived Data Products

- Deliver new on-site content analysis tools for end users using a combination of semantics & new text mining/AI techniques (many of which are free to use from Google/Amazon/Microsoft/others)
- Experiments, experiments, experiments (recommendation, auto-summarization, analogues, image analysis, sentiment, prediction, etc.)

Publishers will either do this themselves or it will be done by someone else on the backs of their content
Thank You

Jake Zarnegar
Chief Product Officer, Silverchair