Answers instead of Articles:

Helping Users Understand Scientific Content

Hosein Azarbonyad
September 2022

Data Science in Elsevier

Using new capabilities (machine learning, natural language processing, AI) to increase our content utility

Data Science

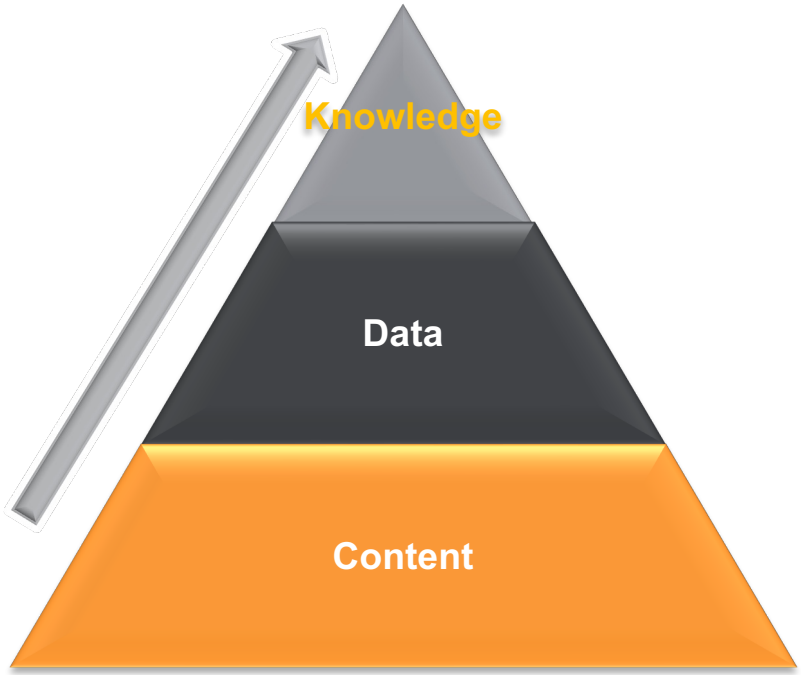
- What we do

Turn Unstructured Content into Structured Content

- Text Mining
- Images
- Video

→ Enabling Data Mining
→ Enabling Data Analytics

Data Science in Elsevier



Answers: *users wanting knowledge – tailor cut to the exact needs of the moment. next-generation search and recommendation Evolved expectations by emergence of AI, Knowledge Graph, new UXes*

Data: *accumulated, structured knowledge. Meta-data around the known entities (authors, articles, geographicals, references, institutions, concepts, relations) – human or machine generated*

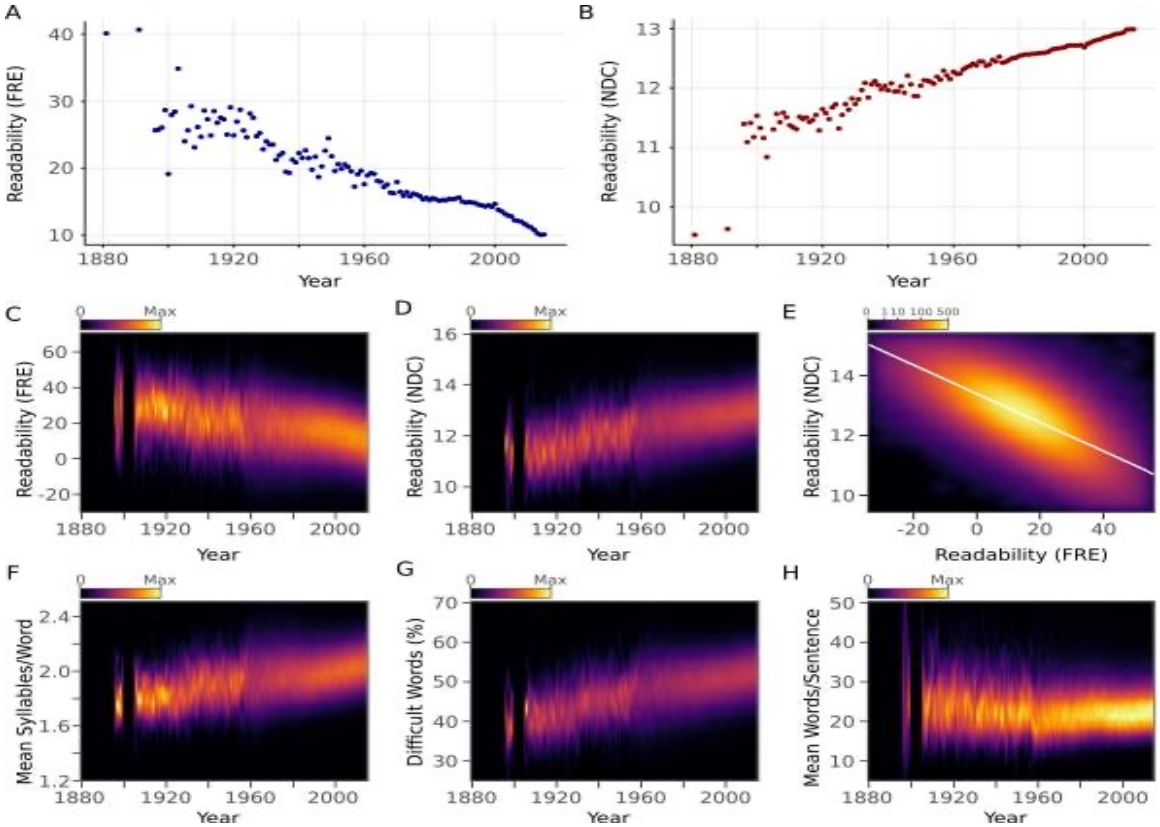
Content: *the underpinning of anything good – published material from Journals, Patents, Web, client data.*

ELSEVIER

What kind of content we are dealing with?

- Technical terms/concepts
- Concise language
- Inherent complexity of scientific language
- Documents not being self-contained
- Ambiguity across or within domains
- Long documents with multi-modal information

What kind of content we are dealing with?



Plavén-Sigray P, Matheson GJ, Schiffler BC, Thompson WH. **The readability of scientific texts is decreasing over time**. Elife. 2017 Sep 5;6:e27725. doi: 10.7554/eLife.27725. PMID: 28873054; PMCID: PMC5584989.

Users of scientific content

- Researchers and scholars
  - Help them to track advancements
- Clicials/Hospitals/Patients
  - Help them track advancements in helath and clinical domains
- Government/Funding agancies
  - Help them to find areas to invest on
- General public
  - Help them understand complex scientific content
- Students
  - Help them to find and understand key learning material

ELSEVIER

Challenges and tasks for scientific document understanding

- Named Entity Recognition
  - Specific to scientific documents
    - Scientific concepts
    - Methods
    - Datasets
    - Equipments
- Maping research outputs to different ta
  - Taxonomy of science
  - Sustainable Development Go
  - Taxonomy of rare diseases

Experiments are conducted on two corpora with different characteristics (Cardoso-Cachopo, 2007), i.e., Reuters-21578 dataset and 20 Newsgroups dataset. More specifically, there are 8 categories in Reuters-21578 dataset, including 5485 training texts and 2189 test texts; 20 categories in 20 Newsgroups dataset, including 11,293 training texts and 7528 test texts. In addition, Reuters-21578 dataset is highly skewed, while the 20 Newsgroups dataset is highly balanced.

problems (Wang et al., 2018; Souery et al., 2007) or lack of effects (Rush, 2007).

The twin support vector machine (TSVM) as a classifier with non-parallel hyper-planes was proposed in [16], which is four times faster than traditional SVM. A

reduce symptoms of depression (Hennings et al., 2013; Wegner et al., 2014; Khanzada et al., 2015; Stanton et al., 2016). However, there are also studies, showing no additional effect of exercise compared to antidepressant medication alone (Danielsson et al., 2013; Kvam et al., 2016) or cognitive behavioral therapy alone (Bernard et al., 2018).

The Fourier transform infrared (FTIR) spectral analysis was carried out on a Thermo Scientific spectrometer (Nicolet iS10) and performed in transmission mode using KBr pellets. Raman spectra were obtained with a WITec Alpha300RA spectrometer using an excitation wavelength of 488 nm. Thermogravimetric analysis (TGA) data was recorded on a Netzsch TG209-F1 instrument at a heating rate of 10 °C min^{-1} in N_2. Thermogravimetric analysis with mass spectrometry

Challenges and tasks for scientific document understanding

- Summarizing (single) scientific articles
  - Highlight extraction/generation

- Creating an inventory of scientific concepts
  - Assisting users by providing contextual global information on unfamiliar scientific concepts as users face them

Science Direct Topic Pages

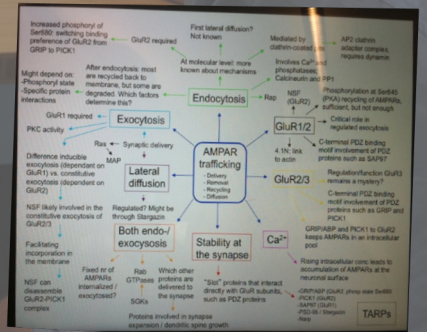
ELSEVIER

Empowering Knowledge

Understanding an article from users perspective

Observation:

When users come across an unknown term in an article, they stop reading, open up Wikipedia and look up the unknown term to get definitions and background information about the concept.



Understanding an article from users perspective

Problem

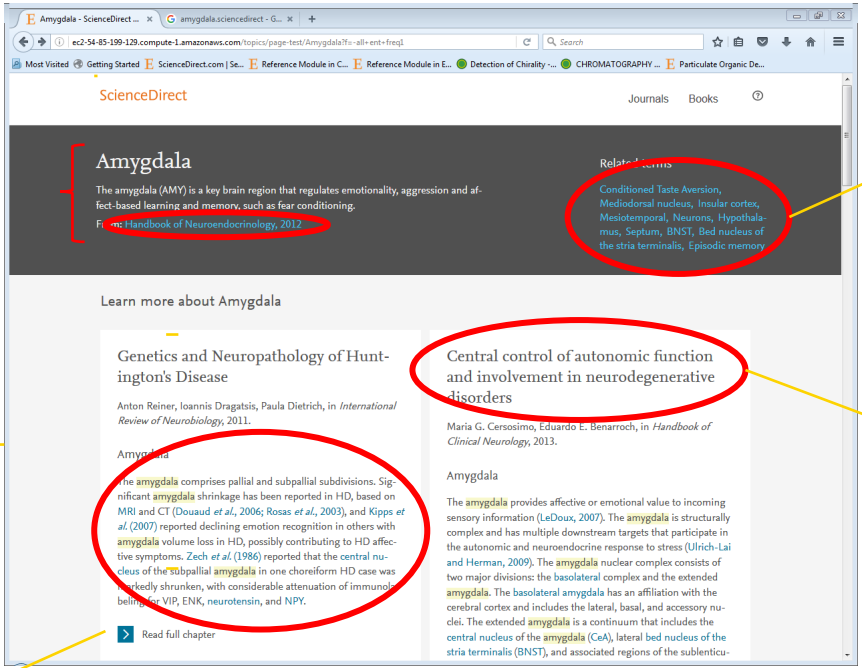
- Academic articles have scientific concepts
- Researchers need information about unfamiliar concepts they encounter
- They lose time searching for foundational information that is trusted and citable

How

- Summarize relevant content from ScienceDirect on *Topic Pages*
- Enrich content with links to the *Topic Pages*
- **Automated** to make processing the content scalable
- Automation presents its own challenges:
  - **Disambiguation** of terms
  - Extraction of **good definitions**

Anatomy of a topic page

Definition, clearly delineated



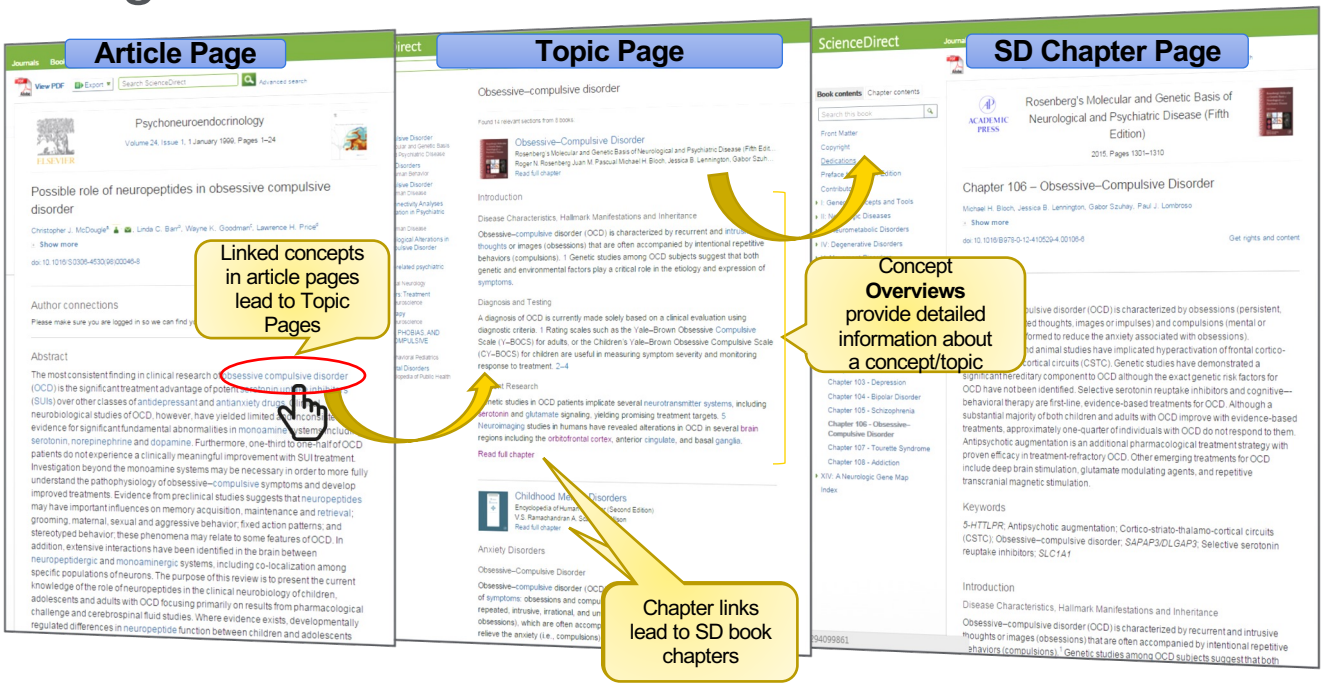
Related terms link to further topic pages *drives serendipity*

Card presentation supports easy scanning and short snippets preferred by users, *saves time*

Title links to chapter, *drives usage*

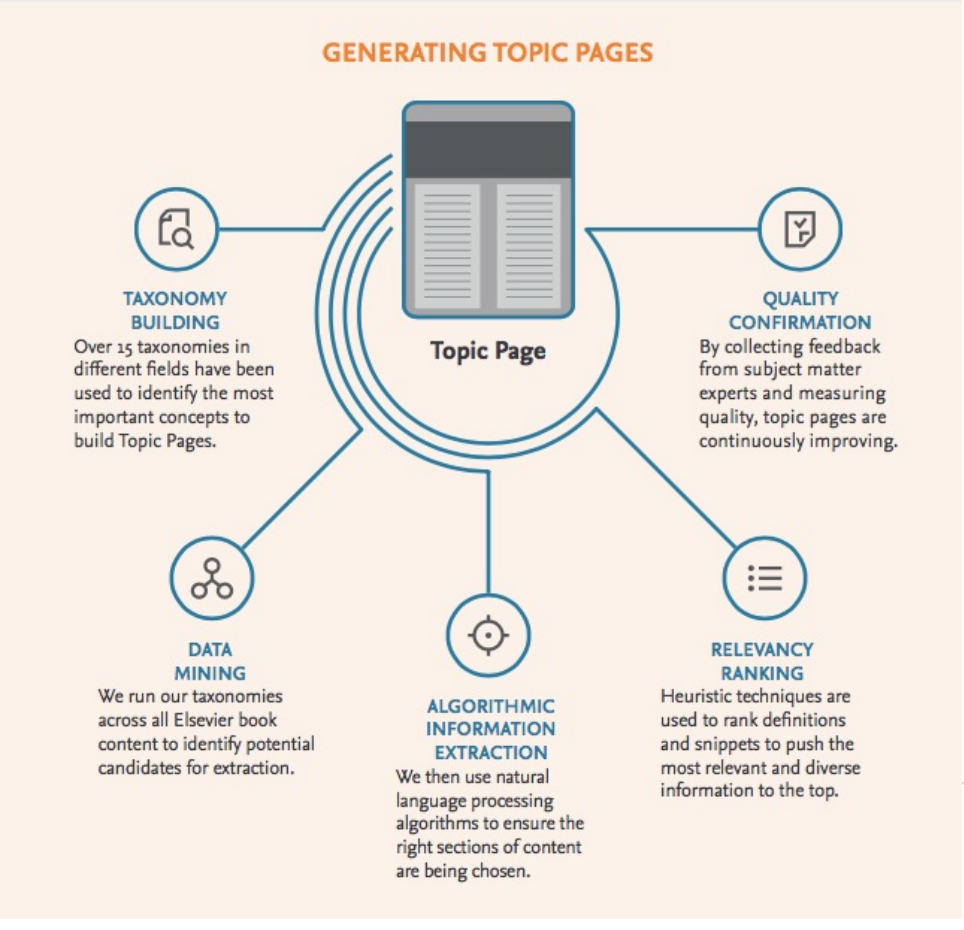
”Read full chapter” links at end of snippet, *drives usage*

The Topic Pages solution

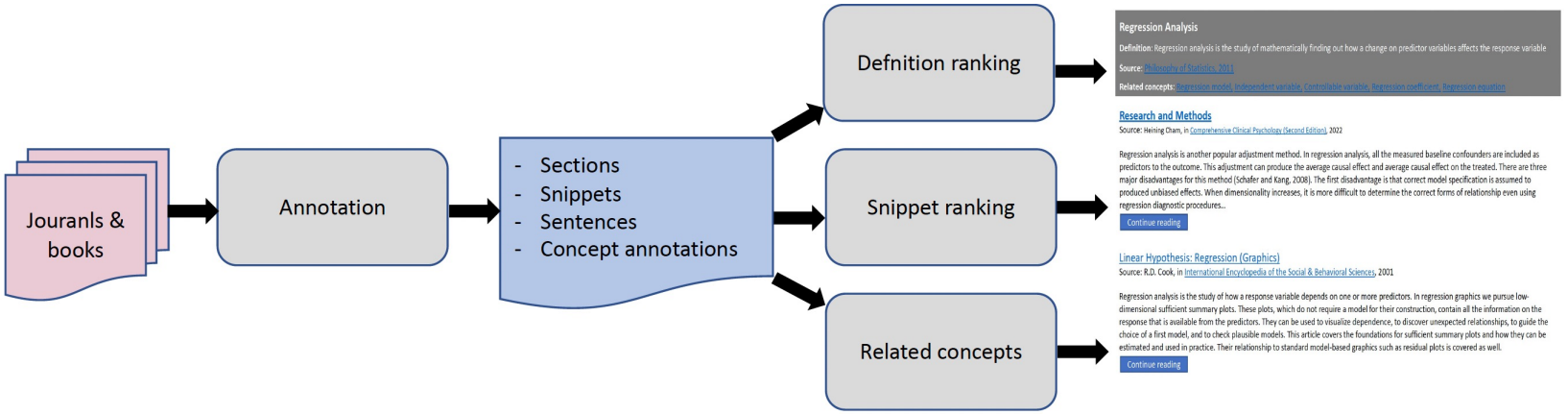


- **Integrates** book content alongside journal articles
- Leverages **user behavior** to deliver content at the **point of need**
- **Free layer** of selected, relevant content
- Links to SD chapter pages from Topic Pages

Used technologies



Generation pipeline



Definition extraction

- Need to automatically identify good definitions from text
- Large amount of data
- Most sentences are not definitions
- Sentences that look like definitions may not be definitions
- Ambiguous concepts

ELSEVIER

Definition ranking

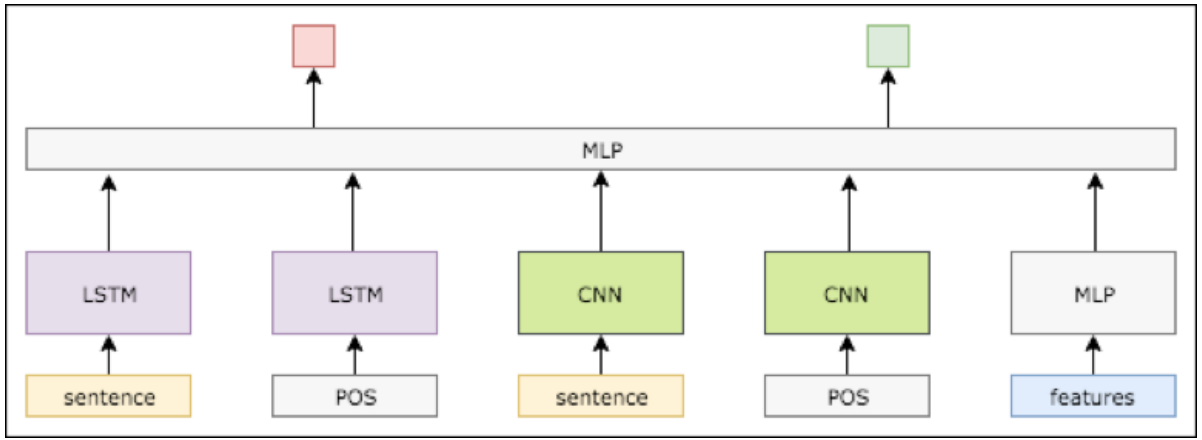
- **Task**
  - Given a pair of <concept, sentence> assign a score reflecting if the sentence provides a good definition for the concept

- **Ranking**
  - Estimate the score for all candidate sentences
  - Rank candidates and pick the top-ranked one

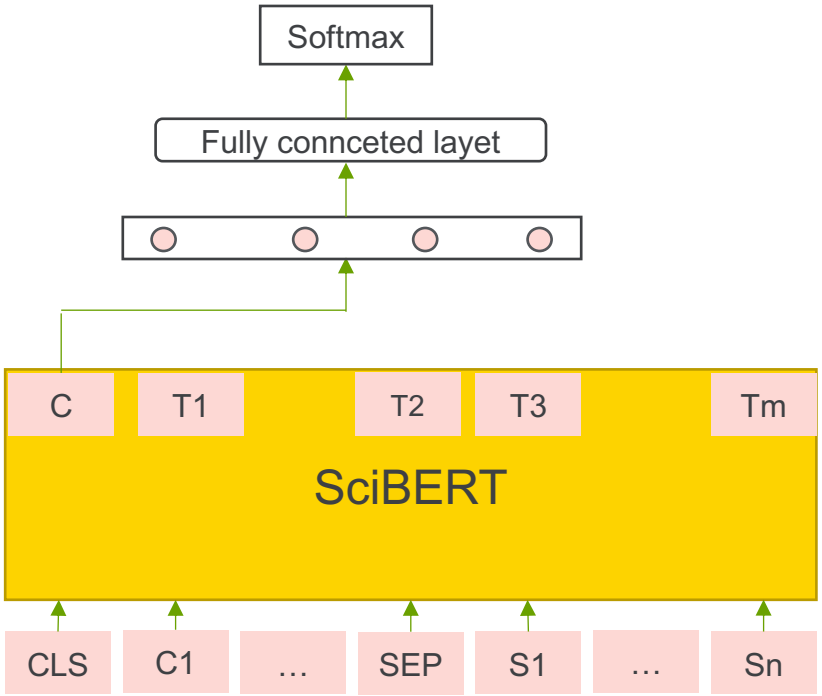
- **Models**
  - LSTM+CNN using structural information
  - SciBERT

LSTM+CNN model

- Captures strctural, sequential, and spatial information inside text
- A set of hand-crafted features are added to inject concept information to the model



SciBERT model



Performance

Results on the WCL dataset

Model	P	R	F1
Jin et al. (2013)	0.92	0.79	0.85
Li et al. (2016)	0.90	0.92	0.91
Navigli and Velardi (2010)	**0.99**	0.61	0.85
LSTM+CNN	0.94	0.91	0.93
SciBERT	0.94	**0.93**	**0.93**

Results on Elsevier dataset

Model	P	R	F1
LSTM+CNN	0.70	0.69	0.69
SciBERT	**0.79**	**0.78**	**0.78**

Types of bad definitions

Concept	Definition	Error source
Association List	An association list is simply a list of name /value pairs.	Too generic
Hierarchical DB	In a hierarchical DB, relationships are defined data by storage structure.	Too generic
Wearable Device	Smart glasses are wearable devices that can be used as AR or VR devices.	Too specific
Habilitation	The acquisition of abilities not possessed previously.	Too specific
TCP	TCP is a popular means of transmitting data through IP packets.	Partially good
Sample Space	the set of all possible outcomes in a probability model	Partially good

Impact of domain difference

domain	SciBERT			LSTM+CNN		
	P	R	F1	P	R	F1
Chem.	0.78	0.80	0.79	0.69	0.68	0.68
Ear.Sc	0.80	0.84	0.82	0.66	0.64	0.65
Mat.SC	0.80	0.88	0.83	0.50	0.49	0.49
Com.Sc	0.56	0.60	0.58	0.43	0.48	0.45
Soc.Sc	0.39	0.43	0.41	0.38	0.46	0.42

Topic pages help students find answers

360,000 topic pages
Across different subject areas

Hyperlinked from 6 million journal articles and book chapters on Science Direct
And highly discoverable in search engines

23 million visits on average every month
With customer research regularly carried out to ensure optimal user experience

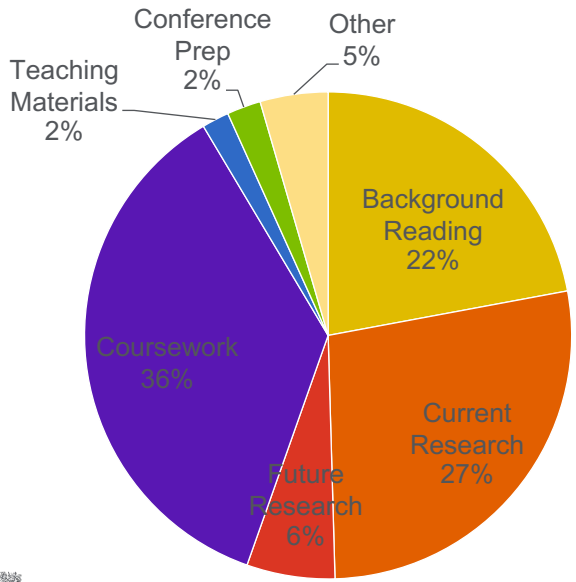
We uncovered that 69% of Topic page users are, in fact, students

> *"As a research student… it helped me achieving the knowledge of a complicated topic more effectively… I am thankful to the entire team for providing such **useful and authentic information at one click***

ELSEVIER

Exit Survey

Why were you interested in [this term] today?



How helpful would this type of page be to you in the following situations?	Very or Quite Helpful
Multidisciplinary work	87%
Investigating a new area	97%
Unfamiliar term in a journal article	84%
Reading in my current research area	87%
Looking up a technique or methodology	81%

Did you find the content helped you for these purposes?	% Yes
Background reading in primary specialty	75%
Background reading in new area	82%
Planning future research	72%
Current research	74%
Writing up current research	68%
Teaching or Coursework	85%
Conference	57%

ELSEVIER

Extracting Article Summaries

ELSEVIER

Empowering Knowledge

Can we use extractive summarization to find the key finding/points within a document?

Available Data

Full Text

Australia, and several other industrialized nations, require an extensive science, technology, engineering, and mathematics (STEM) workforce for economic prosperity, productivity, and global competitiveness. However, the demand for people in STEM outweighs the supply of STEM-trained individuals. One reason for this supply-demand issue is a decline in the proportion of students choosing STEM-related pathways (Ainley, Kos, & Nicholas, 2008). In response to this concern, bourgeoning research has been devoted to identifying predictors of STEM educational and career choices (Shoffner & Dockery, 2015). Among the determinants examined is vocational interests (Bartlett, Perera, & McIlveen, 2016), which is unsurprising, given not only theory positing a central role of interests in choice behaviors (Lent, Brown, & Hackett, 1994) but also extant evidence demonstrating that interests predict choices (Gasser et al., 2007, Larson et al., 2010, Päßler and Hell, 2012). However, existing research, with few exceptions (Leuty et al., 2016, McLarnon et al., 2015), is limited to investigating the unique and additive relations of interests with choices from a variable-centered perspective. This approach assumes that individuals in a sample are from the same population and share the same set of parameters, disregarding the potential existence of multiple latent subpopulations that may show distinct configurations of interests. The near-exclusive focus on unique relations is problematic given work showing that individuals may simultaneously endorse multiple interests (McLarnon et al., 2015, Strahan and Severinghaus, 1992, Tay et al., 2011). From a social cognitive perspective on the career choice process, such interest combinations may be more important for people's educational and vocational choices than interests in isolation and may be a truer representation of individuals' interest profiles, which themselves emerge, in part, from people's dispositional characteristics. However, only little research has been conducted to determine how interests can be combined, and even less is known about how these combinations predict individuals' choices and are predicted by theoretically-meaningful antecedents in the career choice process, such as personality dispositions.

Available Data

Title

A social influence model of consumer participation in network- and small-group-based virtual communities

Abstract

We investigate two key group-level determinants of virtual community participation —group norms and social identity—and consider their motivational antecedents and mediators.

We also introduce a marketing-relevant typology to conceptualize virtual communities, based on the distinction between *network-based* and *small-group-based* virtual communities. Our survey-based study, which was conducted across a broad range of virtual communities, supports the proposed model and finds further that virtual community type moderates consumers' reasons for participating, as well as the strengths of their impact on group norms and social identity. We conclude with a consideration of managerial and research implications of the findings.

Available Data

Keywords

Vocational interests; Interest profiles; STEM career choices; Academic and career choices; Latent profile analysis; Profile invariance; Profile similarity

Article Metrics

Citations

Citation Indexes:	6

Captures

Exports-Saves:	18
Readers:	43

Social Media

Tweets:	13

References

Ainley et al., 2008 J. Ainley, J. Kos, M. Nicholas
Participation in science, mathematics, and technology in Australian education
ACER Research Monograph (No. 63)
(2008)
http://research.acer.edu.au/acer_monographs/4
Google Scholar

Ainley et al., 1990 J. Ainley, W. Jones, K.K. Navaratnam
Subject choice in senior secondary school
Australian Publishing Service, Canberra, ACT (1990)
Google Scholar

Armstrong and Vogel, 2009 P.I. Armstrong, D.L. Vogel
Interpreting the interest–efficacy association from a RIASEC perspective
Journal of Counseling Psychology, 56 (3) (2009), pp. 392-407, 10.1037/a0016407
CrossRef View Record in Scopus Google Scholar

Armstrong et al., 2008 P.I. Armstrong, W. Allison, J. Rounds
Development and initial validation of brief public domain RIASEC marker scales
Journal of Vocational Behavior, 73 (2) (2008), pp. 287-299
https://doi.org/10.1016/j.jvb.2008.06.003
Article Download PDF View Record in Scopus Google Scholar

Asparouhov and Muthén, 2014 T. Asparouhov, B. Muthén
Auxiliary variables in mixture modeling: Three-step approaches using M plus
Structural Equation Modeling: A Multidisciplinary Journal, 21 (3) (2014), pp. 329-341
https://doi.org/10.1080/10705511.2014.915181
CrossRef View Record in Scopus Google Scholar

Available Data – Author Submitted Highlights

- Cover 100% of newly submitted documents
- Cover 8% of all documents
- Covers 25% of traffic

Highlights

- Latent profiles of vocational interests were identified.
- The profiles replicated across subsamples.
- Big-Five personality dimensions differentiated the profiles.
- Profile membership was associated with the probability of STEM major choice.

Greedy Rouge Sampling

1. Select best sentence compared to author highlights
2. Select second sentence, which combined with best set makes biggest increase in score when compared to author highlight
3. Repeat until stop criteria

Culture's impact on institutional investors' trading frequency

Author Highlights

- Culture influences institutions' trading frequency within their own portfolio.
- Institutions' turnover decreases with cultural distance to stocks' home market.
- Cultural ambiguity aversion is negatively related to trading frequency.
- Cultural trust is positively related to trading frequency.

Sampled Data

- In addition, we find evidence that cultural ambiguity aversion is related to lower trading frequency and that cultural trust is related to higher trading frequency.
- In order for the results to be consistent with H1, we expect Cultural distance to be negatively related to institutions' turnover.
- Cultural ambiguity aversion reduces trading frequency.
- Cultural trust increases trading frequency.

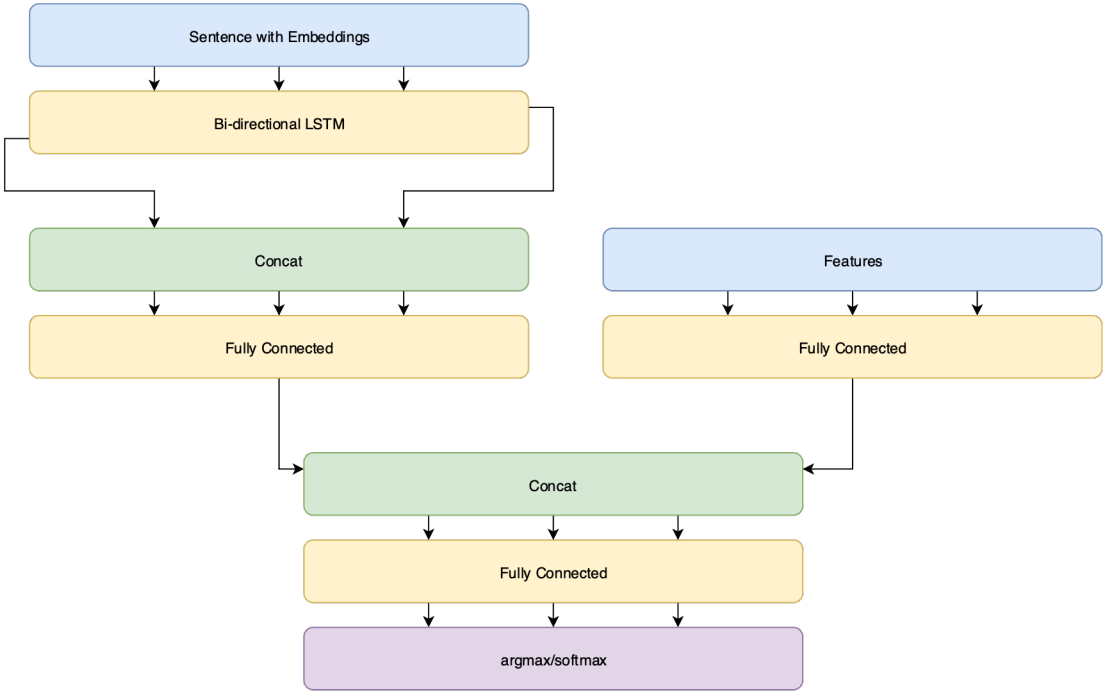
Training Data

138,735 randomly selected documents with author highlights

Split into Train (60%), Test (30%), Validation (10%)

Top 10 sentence labels with Greedy Sampling

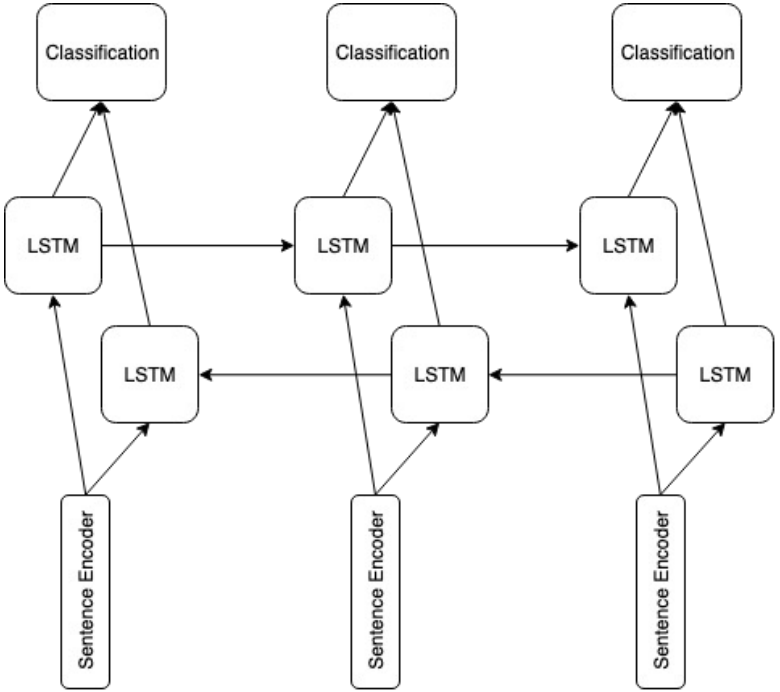
Initial Model – Sentence Classification



- Section Classification
- Content overlap
- Number of numbers
- Sentence length

Collins, E., Augenstein, I., & Riedel, S. (2017). A Supervised Approach to Extractive Summarisation of Scientific Papers. *CoRR*, 195–205. http://doi.org/10.18653/v1/K17-1021

Sequential RNN

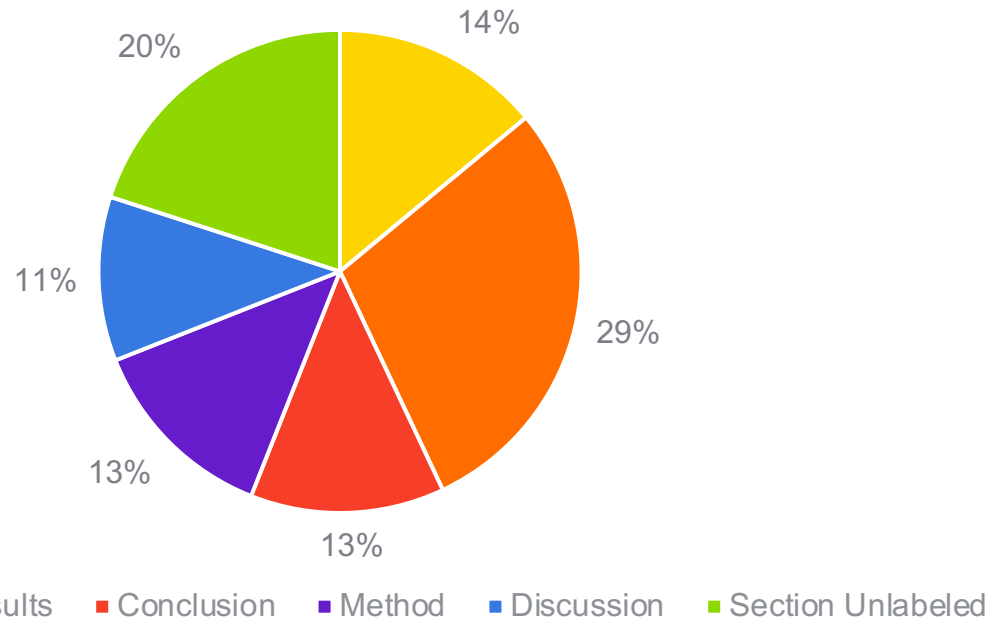


Kedzie, C., McKeown, K. R., & Daumé, H., III. (2018). Content Selection in Deep Learning Models of Summarization. *Emnlp*.

Effectiveness of Sentence Extractors

Sentence Embedding	Word Embedding Trainable	Rouge-l-f
CNN	FALSE	22.13
CNN	TRUE	**22.70**
MEAN	FALSE	22.60
MEAN	TRUE	22.28
RNN	FALSE	22.53
RNN	TRUE	21.36

Section Analysis



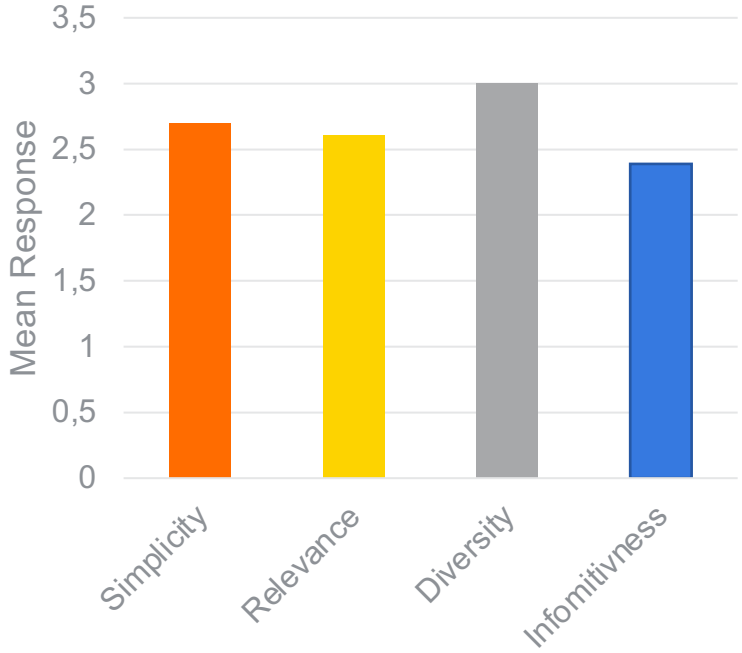
“Human (Editor) in the loop”

Simplicity: are the sentences which have been selected simply to read or are they too long and using over-complicated language.

Informativeness: do the sentences which have been selected inform the user about what is going on within the papers

Relevancy: are the sentences which have been selected relevant to the main findings of the paper

Diversity: are all the sentence which have been selected covering the same points or is their diversity across the sentences.



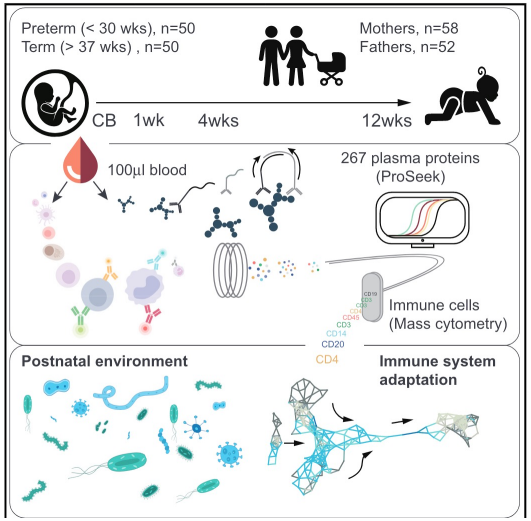
Previous PDF Next PDF

Article

Cell

Stereotypic Immune System Development in Newborn Children

Graphical Abstract



Authors

Axel Olin, Ewa Henckel, Yang Chen, ..., Cheng Zhang, Kajsa Bohlin, Petter Brodin

Correspondence

petter.brodin@ki.se

In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

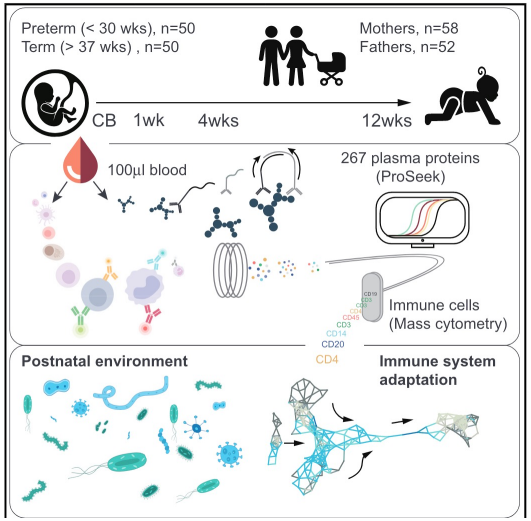
Previous PDF Next PDF

Article

Cell

Stereotypic Immune System Development in Newborn Children

Graphical Abstract



Authors

Axel Olin, Ewa Henckel, Yang Chen, ..., Cheng Zhang, Kajsa Bohlin, Petter Brodin

Correspondence

petter.brodin@ki.se

In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

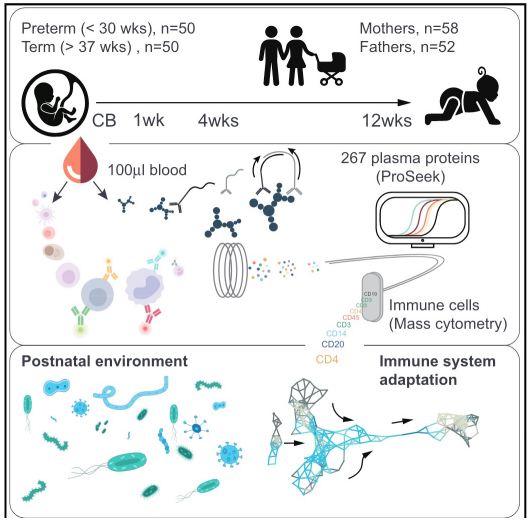
Previous PDF Next PDF

Article

Cell

Stereotypic Immune System Development in Newborn Children

Graphical Abstract



Authors

Axel Olin, Ewa Henckel, Yang Chen, ..., Cheng Zhang, Kajsa Bohlin, Petter Brodin

Correspondence

petter.brodin@ki.se

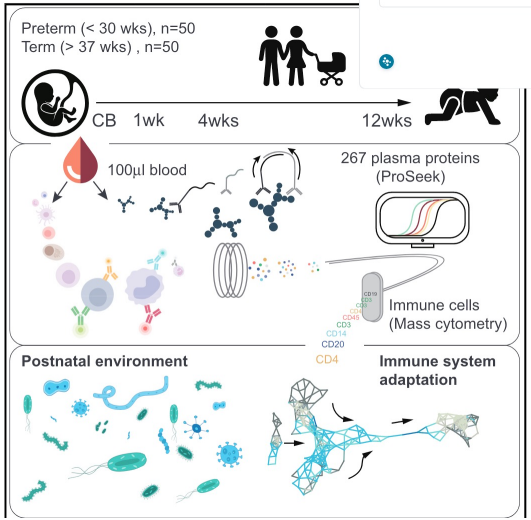
In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

Cell

Stereotypic Immune … Newborn Children

Graphical Abstract



…din

petter.brodin@ki.se

In Brief

Longitudinal profiling of blood immune cells from 100 newborns provides a systemic view on the ontogeny of the human neonatal immune system.

Article info Hide

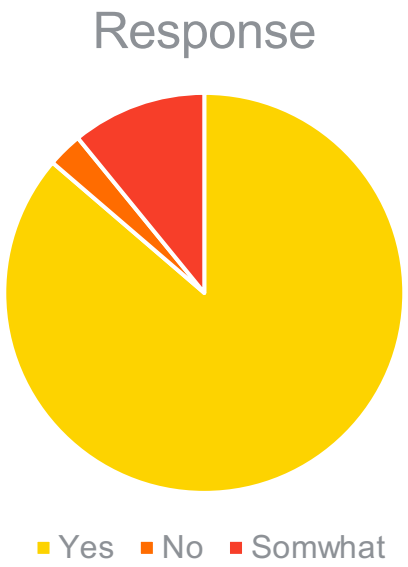
Highlights

- We also describe evidence of a critical period in the development of B, NK, and DCs during the first 3 months of life, as these cell populations reach adult-like phenotypes during this period, suggesting that environmental influences imprinting on these cells during this time window could have long-term consequences.
- If microbial stimuli present during the first 100 days have similar effects on DC development, this might establish an individual's DCs on a trajectory associated with reduced disease risk.
- We also propose that in-depth analyses during early life adaptation to environmental influences provides a unique opportunity for better understanding the molecular mechanisms of immune system adaptation to environmental influences in humans.
- These results show that immune cell compositional changes after birth follow a stereotypic pattern of development in all children, preterm and terms alike, despite their differences in both maturity and postnatal environmental conditions.
- This also suggests that specific cell populations and pathways have different critical periods of calibration when they would be most amenable to environmental imprinting, allowing specific exposures at specific time points in the context of a given genetic makeup to contribute to an individual's risk of individual immune-mediated diseases.
- This converged 3-month immune system state might therefor represent the real set point from which human immune system variation is shaped by environmental exposures over the course of life.

Recommended Articles

B cell alterations during BAFF inhibition with

Usabilla Results



Provides a good quick summary of the research easier to take in at a glance than an abstract. Probably useful as a first step to decide whether the paper is of interest.

Highlights are too similar to an abstract.

Summary

- Scientific document processing poses new challenges and tasks that are unique for such documents
  - Specific named entities, technical jargon, long multi-modal documents and much more
- Summarization is found to be helpful (especially by students) to understand scientific articles
  - Students can potentially benefit from text simplification and relevant tasks
- Limitations and future steps
  - Search result diversification for snippet ranking
  - Logical ordering of snippets rather than ranking by relevance
  - A high level summary of the topic page
    - More informative than the definition and less complex than snippets

Zubair Afzal

George Tsatsaronis

Emma Bruun

Janneke van de Loo

Rob Koeling

Daneil Kershaw

ELSEVIER

ELSEVIER

Thank you