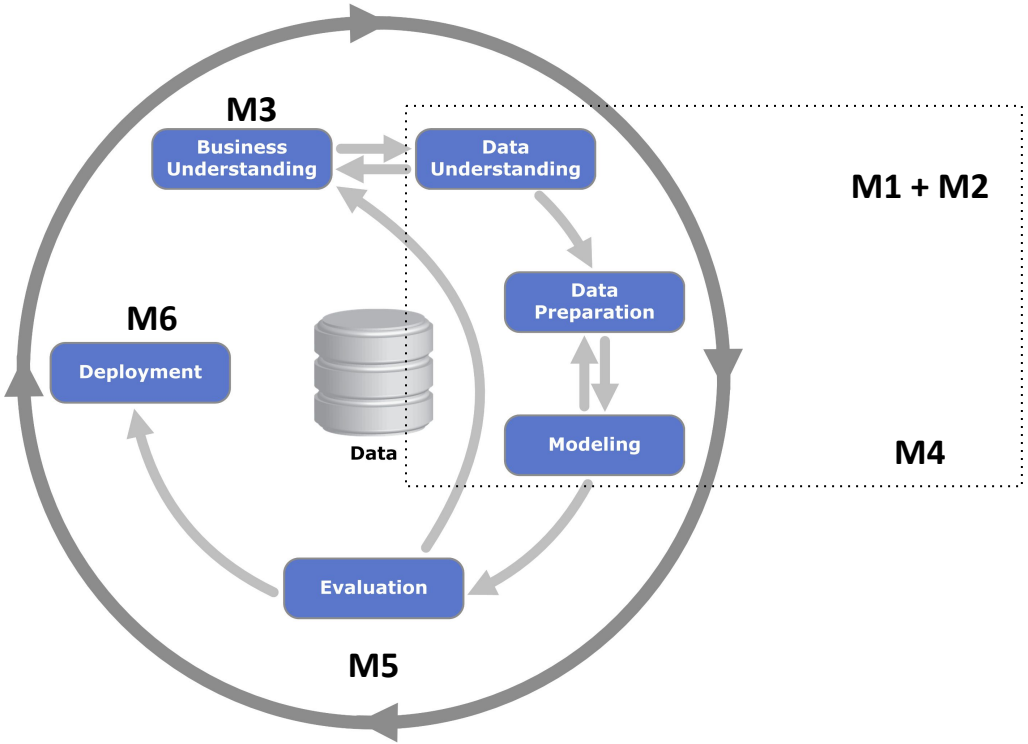
Life2Vec, Market2Vec & Fine-Tuning

Hamid & Milad

Aalborg University Business School
Innovation, Knowledge, and Economic Development (IKE)
AI Denmark (AIDK)
hamidb@business.aau.dk
Feb 02, 2026

CRISP - DM


M3
Business Understanding
Data Understanding
M1 + M2
Data Preparation
M6
Deployment
Data
Modeling
M4
Evaluation
M5

Today's lecture overview

Parts I & II

- Recap lecture 1
- Review the highlights in the BERT paper (Devlin et al. 2019)
- Fine tuning (API and Native Pytorch)

Part III

- Socioeconomic phenomena as sequential processes
- Life2Vec
- Transformers and BDS
- Market2Vec
- Sentence Transformers Finetuning (SetFit)
- Going beyond Market2Vec

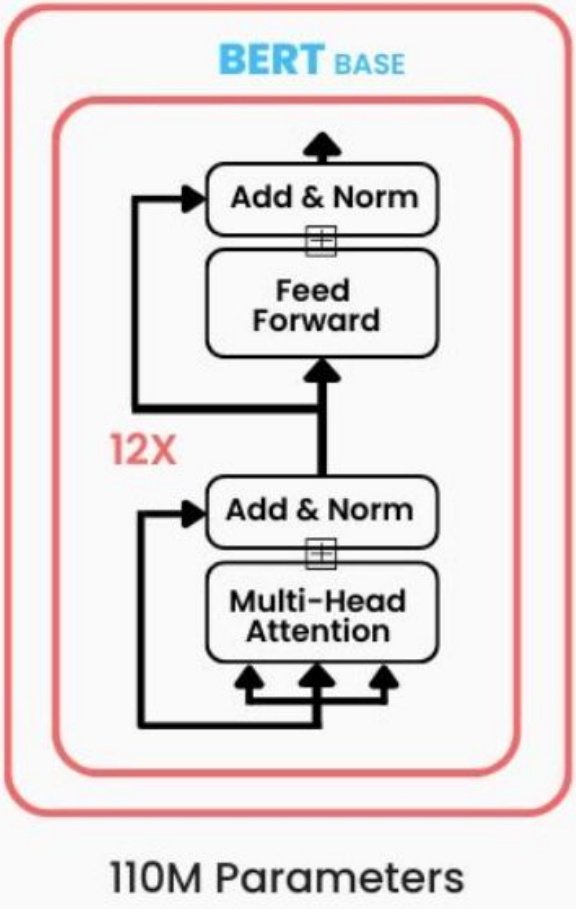
Part IV

- Hands-on exploration: running code to understand how things work in practice

Recap: Lecture 1

Let’s refresh what we learned in the first lecture!

Attention and Contextualization



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.

Comparison	BERT October 11, 2018	RoBERTa July 26, 2019	DistilBERT October 2, 2019	ALBERT September 26, 2019
Parameters	**Base:** 110M **Large:** 340M	**Base:** 125 **Large:** 355	**Base:** 66	**Base:** 12M **Large:** 18M
Layers / Hidden Dimensions / Self-Attention Heads	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 6 / 768 / 12	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16
Training Time	**Base:** 8 x V100 x 12d **Large:** 280 x V100 x 1d	1024 x V100 x 1 day (4-5x more than BERT)	**Base:** 8 x V100 x 3.5d (4 times less than BERT)	[not given] **Large:** 1.7x faster
Performance	Outperforming SOTA in Oct 2018	88.5 on GLUE	97% of BERT-base's performance on GLUE	89.4 on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Trans-former, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced para-meters & SOP (not NSP)

3 BERT

We introduce BERT and its detailed implementation in this section. There are two steps in our framework: *pre-training* and *fine-tuning*. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. The question-answering example in Figure 1 will serve as a running example for this section.

A distinctive feature of BERT is its unified architecture across different tasks. There is mini-

In this work, we denote the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A.[3] We primarily report results on two model sizes: $\textbf{BERT}_{\textbf{BASE}}$ (L=12, H=768, A=12, Total Parameters=110M) and $\textbf{BERT}_{\textbf{LARGE}}$ (L=24, H=1024, A=16, Total Parameters=340M).

$\text{BERT}_{\text{BASE}}$ was chosen to have the same model size as OpenAI GPT for comparison purposes. Critically, however, the BERT Transformer uses bidirectional self-attention, while the GPT Transformer uses constrained self-attention where every token can only attend to context to its left.[4]

[1] https://github.com/tensorflow/tensor2tensor
[2] http://nlp.seas.harvard.edu/2018/04/03/attention.html
[3] In all cases we set the feed-forward/filter size to be $4H$, i.e., 3072 for the $H = 768$ and 4096 for the $H = 1024$.
[4] We note that in the literature the bidirectional Trans-

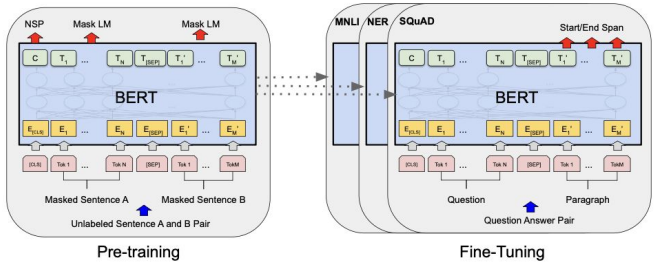
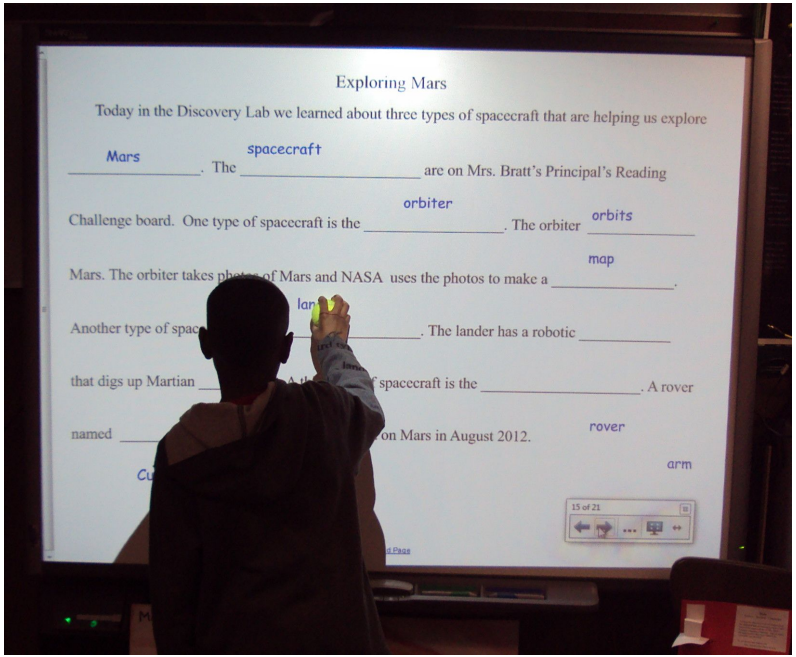


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special

A cloze test (Taylor, 1953) and Masked Language Model (MLM)

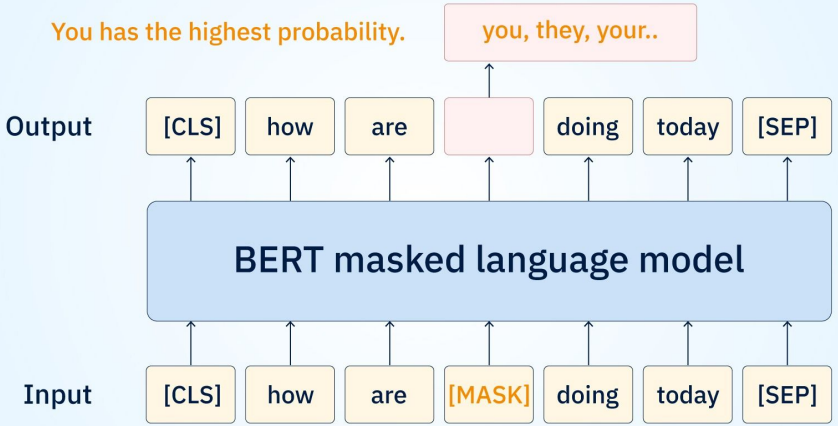
Input/Output Representations To make BERT handle a variety of down-stream tasks, our input representation is able to unambiguously represent both a single sentence and a pair of sentences (e.g., ⟨ Question, Answer ⟩) in one token sequence. Throughout this work, a "sentence" can be an arbitrary span of contiguous text, rather than an actual linguistic sentence. A "sequence" refers to the input token sequence to BERT, which may be a single sentence or two sentences packed together.

In order to train a deep bidirectional representation, we simply mask some percentage of the input tokens at random, and then predict those masked tokens. We refer to this procedure as a "masked LM" (MLM), although it is often referred to as a *Cloze* task in the literature (Taylor, 1953). In this case, the final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary, as in a standard LM. In all of our experiments, we mask 15% of all WordPiece to-



A cloze test is an exercise, test, or assessment in which a portion of text is masked and the participant is asked to fill in the masked portion of text . Source: wikipedia

Masked Language Model (MLM)



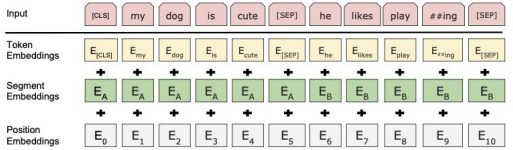


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

The NSP task is closely related to representation-learning objectives used in Jernite et al. (2017) and Logeswaran and Lee (2018). However, in prior work, only sentence embeddings are transferred to down-stream tasks, where BERT transfers all parameters to initialize end-task model parameters.

Pre-training data The pre-training procedure largely follows the existing literature on language model pre-training. For the pre-training corpus we use the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). For Wikipedia we extract only the text passages and ignore lists, tables, and headers. It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the Billion Word Benchmark (Chelba et al., 2013) in order to extract long contiguous sequences.

3.2 Fine-tuning BERT

Fine-tuning is straightforward since the self-attention mechanism in the Transformer allows BERT to model many downstream tasks—whether they involve single text or text pairs—by swapping out the appropriate inputs and outputs. For applications involving text pairs, a common pattern is to independently encode text pairs before applying bidirectional cross attention, such as Parikh et al. (2016); Seo et al. (2017). BERT instead uses the self-attention mechanism to unify these two stages, as encoding a concatenated text pair with self-attention effectively includes *bidirectional* cross attention between two sentences.

For each task, we simply plug in the task-specific inputs and outputs into BERT and fine-tune all the parameters end-to-end. At the input, sentence A and sentence B from pre-training are analogous to (1) sentence pairs in paraphrasing, (2) hypothesis-premise pairs in entailment, (3) question-passage pairs in question answering, and (4) a degenerate text-∅ pair in text classification or sequence tagging. At the output, the token representations are fed into an output layer for token-level tasks, such as sequence tagging or question answering, and the [CLS] representation is fed into an output layer for classification, such as entailment or sentiment analysis.

Compared to pre-training, fine-tuning is relatively inexpensive. All of the results in the paper can be replicated in at most 1 hour on a single Cloud TPU, or a few hours on a GPU, starting from the exact same pre-trained model.[7] We describe the task-specific details in the corresponding subsections of Section 4. More details can be found in Appendix A.5.

4 Experiments

In this section, we present BERT fine-tuning results on 11 NLP tasks.

4.1 GLUE

The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018a) is a collection of diverse natural language understanding tasks. Detailed descriptions of GLUE datasets are included in Appendix B.1.

To fine-tune on GLUE, we represent the input sequence (for single sentence or sentence pairs) as described in Section 3, and use the final hidden vector $C \in \mathbb{R}^H$ corresponding to the first input token ([CLS]) as the aggregate representation. The only new parameters introduced during fine-tuning are classification layer weights $W \in \mathbb{R}^{K \times H}$, where K is the number of labels. We compute a standard classification loss with C and W, i.e., $\log(\text{softmax}(CW^T))$.

[7] For example, the BERT SQuAD model can be trained in around 30 minutes on a single Cloud TPU to achieve a Dev F1 score of 91.0%.

[8] See (10) in https://gluebenchmark.com/faq.

Pre-training objectives

- Masked Language Modeling (MLM): mask ~15% of tokens, then predict the originals using both left and right context.
- Next Sentence Prediction (NSP, BERT): classify whether Sentence B truly follows Sentence A (IsNext vs NotNext).
- Scale matters: train on a large, diverse text corpus to capture broad linguistic patterns.

What the model sees

MLM EXAMPLE

```
Original:  The quick brown fox jumps over
the lazy dog.
```

```
Input:     The quick brown [MASK] jumps over
the lazy dog.
```

Predict: “fox”

NSP EXAMPLE (BERT)

```
Sentence A:  The quick brown fox jumps over
the lazy dog.
```

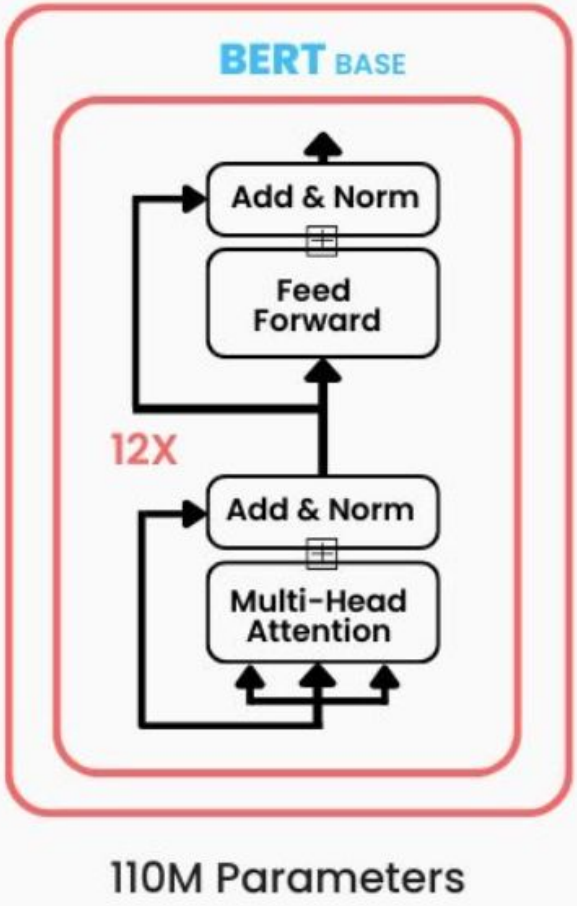
```
Sentence B:  It then swiftly ran up the
hill.
```

Classifier: IsNext / NotNext

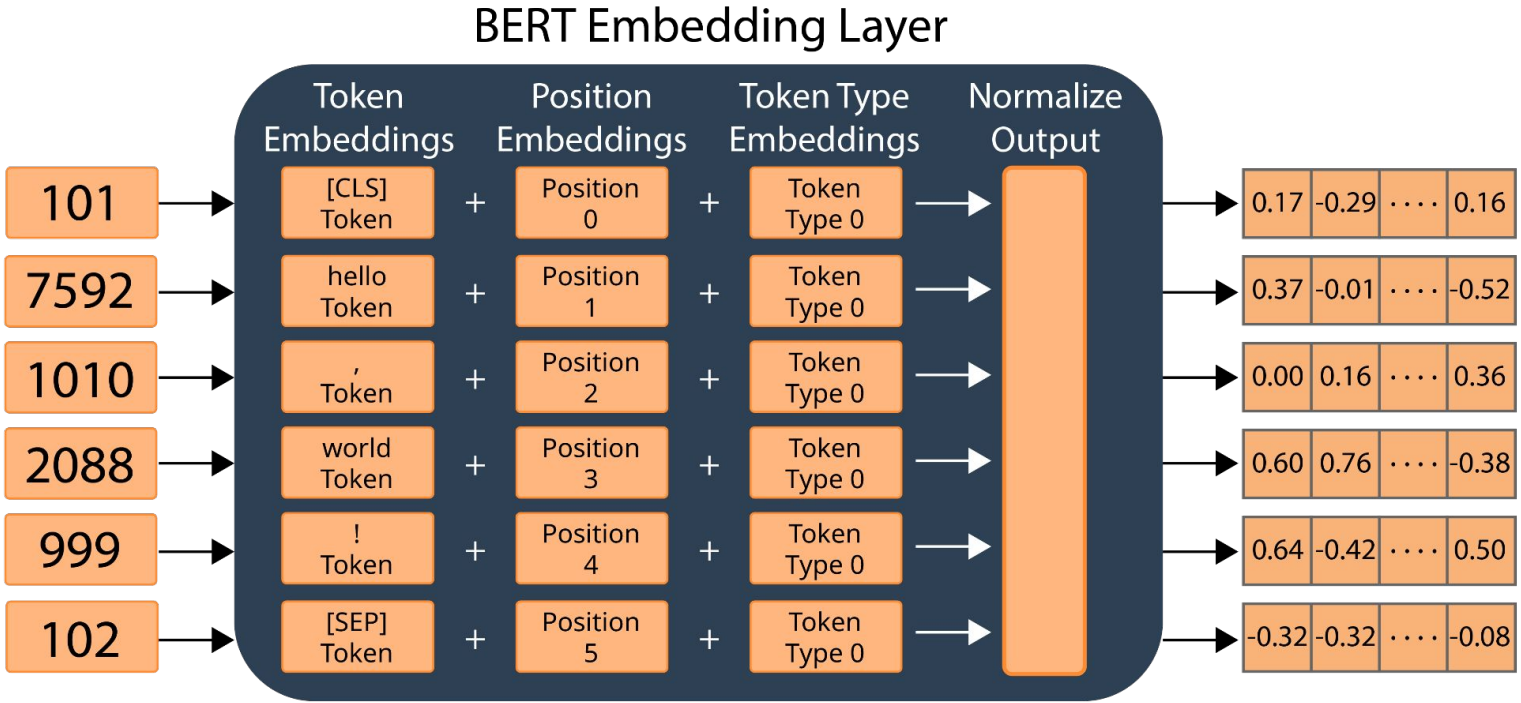
```
BertModel(
  (embeddings): BertEmbeddings(
    (word_embeddings): Embedding(30522, 768, padding_idx=0)
    (position_embeddings): Embedding(512, 768)
    (token_type_embeddings): Embedding(2, 768)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (encoder): BertEncoder(
    (layer): ModuleList(
      (0-11): 12 x BertLayer(
        (attention): BertAttention(
          (self): BertSelfAttention(
            (query): Linear(in_features=768, out_features=768, bias=True)
            (key): Linear(in_features=768, out_features=768, bias=True)
            (value): Linear(in_features=768, out_features=768, bias=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (output): BertSelfOutput(
            (dense): Linear(in_features=768, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
        (intermediate): BertIntermediate(
          (dense): Linear(in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation()
        )
        (output): BertOutput(
          (dense): Linear(in_features=3072, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
```

Lecture 2

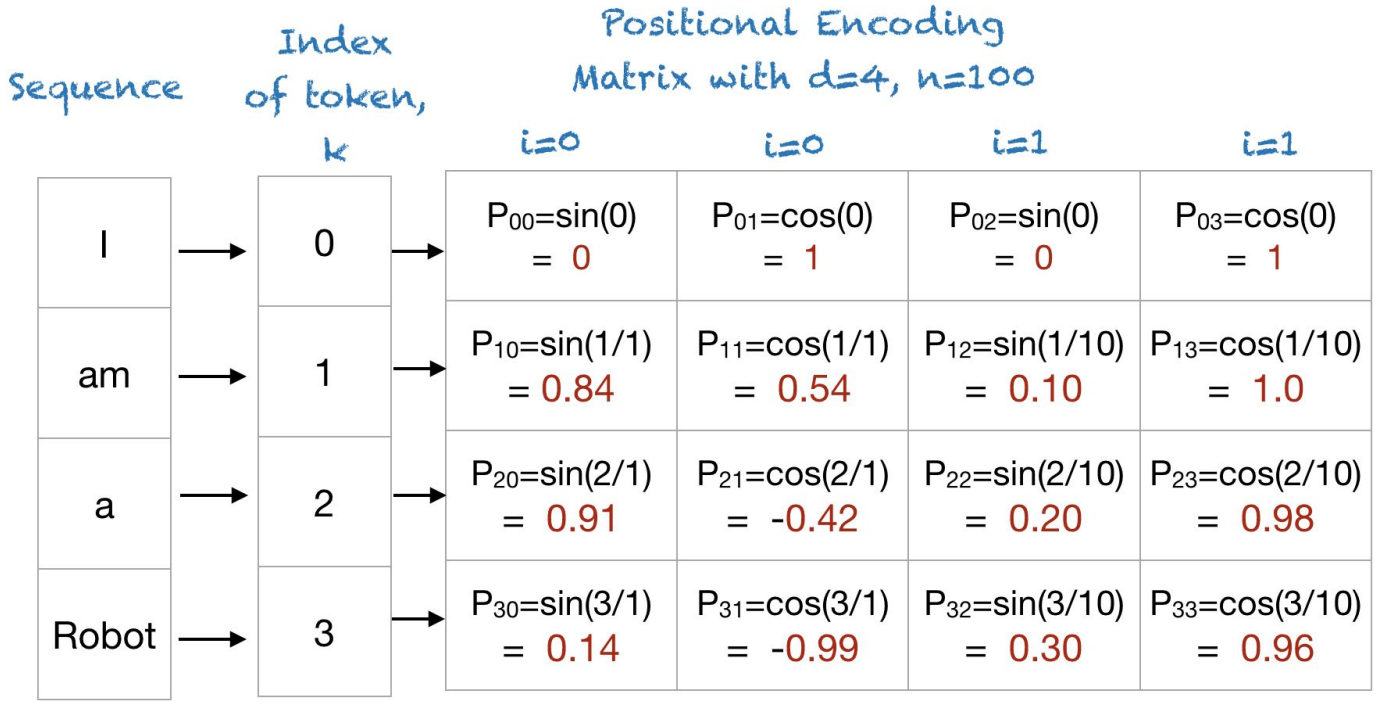
Lecture 1



BERT Embedding Layer

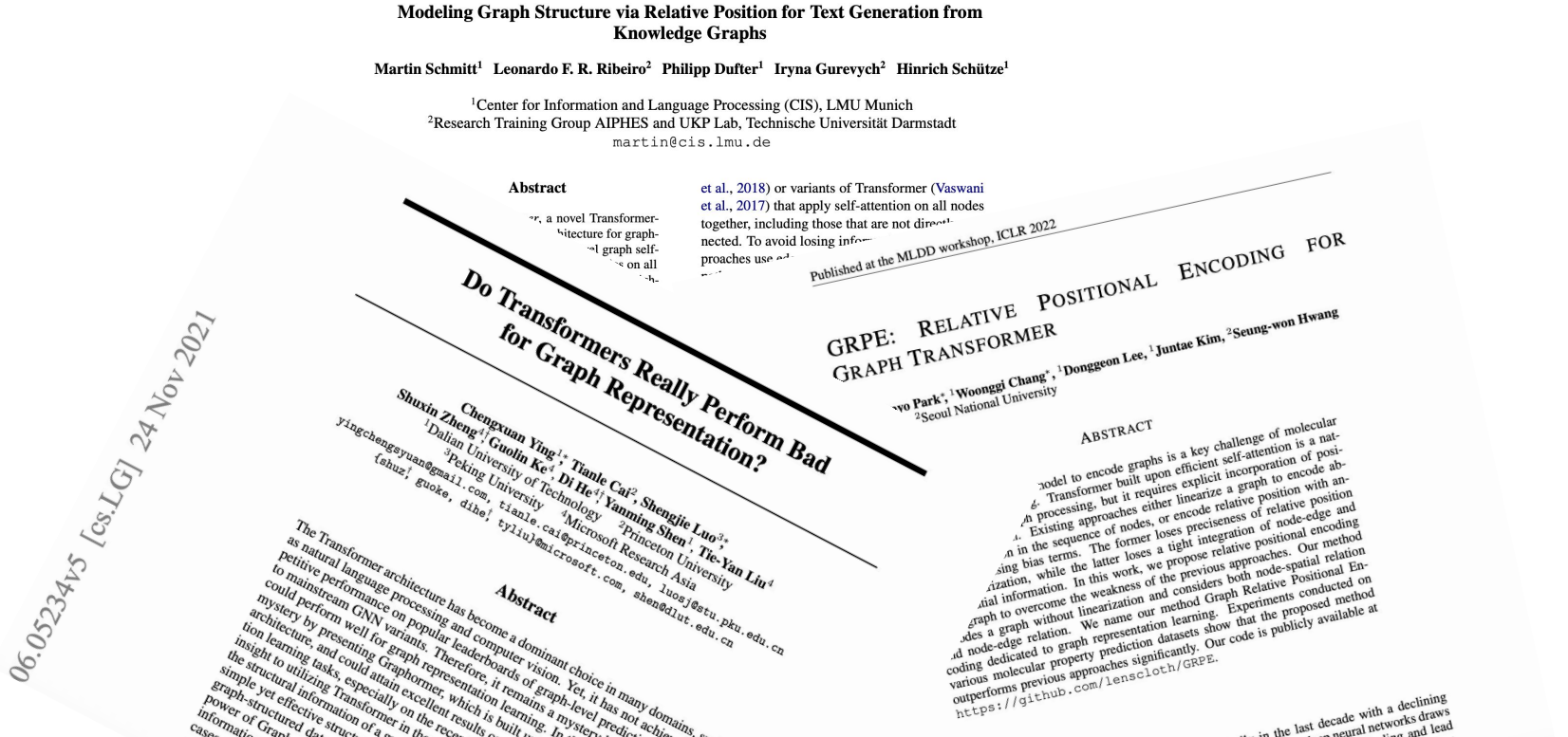


Absolute vs Relative Positioning



Positional Encoding Matrix for the sequence 'I am a robot'

Absolute vs Relative Positioning



Modeling Graph Structure via Relative Position for Text Generation from Knowledge Graphs

Martin Schmitt[1] Leonardo F. R. Ribeiro[2] Philipp Dufter[1] Iryna Gurevych[2] Hinrich Schütze[1]

[1]Center for Information and Language Processing (CIS), LMU Munich
[2]Research Training Group AIPHES and UKP Lab, Technische Universität Darmstadt
martin@cis.lmu.de

Abstract

et al., 2018) or variants of Transformer (Vaswani et al., 2017) that apply self-attention on all nodes together, including those that are not dire[illegible] nected. To avoid losing info[illegible] proaches use a[illegible]

06.05234v5 [cs.LG] 24 Nov 2021

Do Transformers Really Perform Bad for Graph Representation?

Chengxuan Ying[1]*, **Tianle Cai**[2]*, **Shengjie Luo**[3]*,
Shuxin Zheng[4†], **Guolin Ke**[4], **Di He**[4†], **Yanming Shen**[1], **Tie-Yan Liu**[4]
[1]Dalian University of Technology [2]Princeton University
[3]Peking University [4]Microsoft Research Asia
yingchengyuan@gmail.com, tianle.cai@princeton.edu, luosj@stu.pku.edu.cn
{shuz, guoke, dihe, tyliu}@microsoft.com, shen@dlut.edu.cn

Abstract

The Transformer architecture has become a dominant choice in many domains, [illegible] as natural language processing and computer vision. Yet, it has not achi[illegible] petitive performance on popular leaderboards of graph-level predi[illegible] to mainstream GNN variants. Therefore, it remains a mystery how [illegible] could perform well for graph representation learning. In [illegible] mystery by presenting Graphormer, which is built [illegible] architecture, and could attain excellent results [illegible] tion learning tasks, especially on the rec[illegible] insight to utilizing Transformer in the [illegible] the structural information of a [illegible] simple yet effective struc[illegible] graph-structured da[illegible] power of Graph[illegible] informatio[illegible] case[illegible]

Published at the MLDD workshop, ICLR 2022

GRPE: RELATIVE POSITIONAL ENCODING FOR GRAPH TRANSFORMER

[illegible]vo Park*, [1]Woonggi Chang*, [1]Donggeon Lee, [1]Juntae Kim, [2]Seung-won Hwang
[2]Seoul National University

ABSTRACT

[illegible]odel to encode graphs is a key challenge of molecular [illegible]. Transformer built upon efficient self-attention is a nat- [illegible]h processing, but it requires explicit incorporation of posi- [illegible]. Existing approaches either linearize a graph to encode ab- [illegible]n in the sequence of nodes, or encode relative position with an- [illegible]sing bias terms. The former loses preciseness of relative position [illegible]ization, while the latter loses a tight integration of node-edge and [illegible]tial information. In this work, we propose relative positional encoding [illegible]graph to overcome the weakness of the previous approaches. Our method [illegible]des a graph without linearization and considers both node-spatial rela- [illegible]d node-edge relation. We name our method Graph Relative Positional En- coding dedicated to graph representation learning. Experiments conducted on various molecular property prediction datasets show that the proposed method outperforms previous approaches significantly. Our code is publicly available at https://github.com/lenscloth/GRPE.

[illegible] in the last decade with a declining [illegible] neural networks draws [illegible] and lead

Socioeconomic phenomena as sequential processes

Many human phenomena unfold as ordered sequences rather than isolated events
The meaning of an action often depends on what comes before and after it

Examples

Language: Word order shapes meaning and grammar. The same words in a different sequence can change or destroy meaning

Grammar, let's eat! □ Inviting grandma to dinner.

Let's eat grandma! □ Cannibalism 😬

More examples?

Socioeconomic phenomena as sequential processes

Many human phenomena unfold as ordered sequences rather than isolated events
The meaning of an action often depends on what comes before and after it

Examples

Language: Word order shapes meaning and grammar. The same words in a different sequence can change or destroy meaning

Life course events

Typical sequences such as kindergarten, primary school, secondary school, and so on.

Firms and products

Products are developed, launched, and upgraded in specific sequences. Earlier technological choices constrain or enable future innovations

Life2Vec

nature computational science

Article

https://doi.org/10.1038/s43588-023-00573-5

Using sequences of life-events to predict human lives

Received: 6 June 2023

Accepted: 15 November 2023

Published online: 18 December 2023

Check for updates

Germans Savcisens[1], Tina Eliassi-Rad[2,3], Lars Kai Hansen[1], Laust Hvas Mortensen[4,5], Lau Lilleholt[6,7], Anna Rogers[8], Ingo Zettler[6,7] & Sune Lehmann[1,7]

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

We live in the age of algorithm-driven prediction of human behavior. The predictions range from those at the global and population level, with societies allocating vast resources to predicting phenomena such as global warming[1] or the spread of infectious diseases[2], all the way to the constant flow of individual micro-predictions that shape our reality and behavior as we use social media[3]. When it comes to individual life outcomes, however, the picture is more complex. Sociodemographic factors play an important role in human lives[4], but, based on independent analyses of the same dataset, a recent collaboration of 160 teams has recently argued for practical upper limits for the predictions of life outcomes[5].

In this Article we find that, with highly detailed data, a different picture of individual-level predictability emerges. Drawing on a unique dataset consisting of detailed individual-level day-by-day records[6,7] describing the six million inhabitants of Denmark, and spanning a decade interval, we show that accurate individual predictions are indeed possible. Our dataset includes a host of indicators, such as health, professional occupation and affiliation, income level, residency, working hours and education (Dataset section).

The main reason why we are currently experiencing this 'age of human prediction' is the advent of massive datasets and powerful machine learning algorithms[8,9]. Over the past decade, machine learning has revolutionized the image- and text-processing fields by accessing ever larger datasets that have enabled increasingly complex models[10,11]. Language processing has evolved particularly rapidly, and transformer architectures have proven successful at capturing complex patterns in massive and unstructured sequences of words[12–14]. Although these models originated in natural language processing, their ability to capture structure in human language generalizes to other sequences[15–19] that share properties with language, for example, where sequence

[1]DTU Compute, Technical University of Denmark, Lyngby, Denmark. [2]Network Science Institute, Northeastern University, Boston, MA, USA. [3]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [4]Data Science Lab, Statistics Denmark, Copenhagen, Denmark. [5]Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [6]Department of Psychology, University of Copenhagen, Copenhagen, Denmark. [7]Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Copenhagen, Denmark. [8]Computer Science Department, IT University of Copenhagen, Copenhagen, Denmark. e-mail: sljo@dtu.dk

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

Life2Vec

nature computational science

Article

https://doi.org/10.1038/s43588-023-00573-5

Using sequences of life-events to predict human lives

Received: 6 June 2023

Accepted: 15 November 2023

Published online: 18 December 2023

Check for updates

Germans Savcisens[1], Tina Eliassi-Rad[2,3], Lars Kai Hansen[1], Laust Hvas Mortensen[4,5], Lau Lilleholt[6,7], Anna Rogers[8], Ingo Zettler[6,7] & Sune Lehmann[1,7]

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

We live in the age of algorithm-driven prediction of human behavior. The predictions range from those at the global and population level, with societies allocating vast resources to predicting phenomena such as global warming[1] or the spread of infectious diseases[2], all the way to the constant flow of individual micro-predictions that shape our reality and behavior as we use social media[3]. When it comes to individual life outcomes, however, the picture is more complex. Sociodemographic factors play an important role in human lives[4], but, based on independent analyses of the same dataset, a recent collaboration of 160 teams has recently argued for practical upper limits for the predictions of life outcomes[5].

In this Article we find that, with highly detailed data, a different picture of individual-level predictability emerges. Drawing on a unique dataset consisting of detailed individual-level day-by-day records[6,7] describing the six million inhabitants of Denmark, and spanning a decade interval, we show that accurate individual predictions are indeed possible. Our dataset includes a host of indicators, such as health, professional occupation and affiliation, income level, residency, working hours and education (Dataset section).

The main reason why we are currently experiencing this 'age of human prediction' is the advent of massive datasets and powerful machine learning algorithms[8,9]. Over the past decade, machine learning has revolutionized the image- and text-processing fields by accessing ever larger datasets that have enabled increasingly complex models[10,11]. Language processing has evolved particularly rapidly, and transformer architectures have proven successful at capturing complex patterns in massive and unstructured sequences of words[12–14]. Although these models originated in natural language processing, their ability to capture structure in human language generalizes to other sequences[15–19] that share properties with language, for example, where sequence

[1]DTU Compute, Technical University of Denmark, Lyngby, Denmark. [2]Network Science Institute, Northeastern University, Boston, MA, USA. [3]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [4]Data Science Lab, Statistics Denmark, Copenhagen, Denmark. [5]Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [6]Department of Psychology, University of Copenhagen, Copenhagen, Denmark. [7]Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Copenhagen, Denmark. [8]Computer Science Department, IT University of Copenhagen, Copenhagen, Denmark. e-mail: sljo@dtu.dk

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

Life2Vec

nature computational science

Article

https://doi.org/10.1038/s43588-023-00573-5

Using sequences of life-events to predict human lives

Received: 6 June 2023

Accepted: 15 November 2023

Published online: 18 December 2023

Check for updates

Germans Savcisens[1], Tina Eliassi-Rad[2,3], Lars Kai Hansen[1], Laust Hvas Mortensen[4,5], Lau Lilleholt[6,7], Anna Rogers[8], Ingo Zettler[6,7] & Sune Lehmann[1,7]

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

We live in the age of algorithm-driven prediction of human behavior. The predictions range from those at the global and population level, with societies allocating vast resources to predicting phenomena such as global warming[1] or the spread of infectious diseases[2], all the way to the constant flow of individual micro-predictions that shape our reality and behavior as we use social media[3]. When it comes to individual life outcomes, however, the picture is more complex. Sociodemographic factors play an important role in human lives[4], but, based on independent analyses of the same dataset, a recent collaboration of 160 teams has recently argued for practical upper limits for the predictions of life outcomes[5].

In this Article we find that, with highly detailed data, a different picture of individual-level predictability emerges. Drawing on a unique dataset consisting of detailed individual-level day-by-day records[6,7] describing the six million inhabitants of Denmark, and spanning a decade interval, we show that accurate individual predictions are indeed possible. Our dataset includes a host of indicators, such as health, professional occupation and affiliation, income level, residency, working hours and education (Dataset section).

The main reason why we are currently experiencing this 'age of human prediction' is the advent of massive datasets and powerful machine learning algorithms[8,9]. Over the past decade, machine learning has revolutionized the image- and text-processing fields by accessing ever larger datasets that have enabled increasingly complex models[10,11]. Language processing has evolved particularly rapidly, and transformer architectures have proven successful at capturing complex patterns in massive and unstructured sequences of words[12–14]. Although these models originated in natural language processing, their ability to capture structure in human language generalizes to other sequences[15–19] that share properties with language, for example, where sequence

[1]DTU Compute, Technical University of Denmark, Lyngby, Denmark. [2]Network Science Institute, Northeastern University, Boston, MA, USA. [3]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [4]Data Science Lab, Statistics Denmark, Copenhagen, Denmark. [5]Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [6]Department of Psychology, University of Copenhagen, Copenhagen, Denmark. [7]Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Copenhagen, Denmark. [8]Computer Science Department, IT University of Copenhagen, Copenhagen, Denmark. e-mail: sljo@dtu.dk

Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

Life2Vec

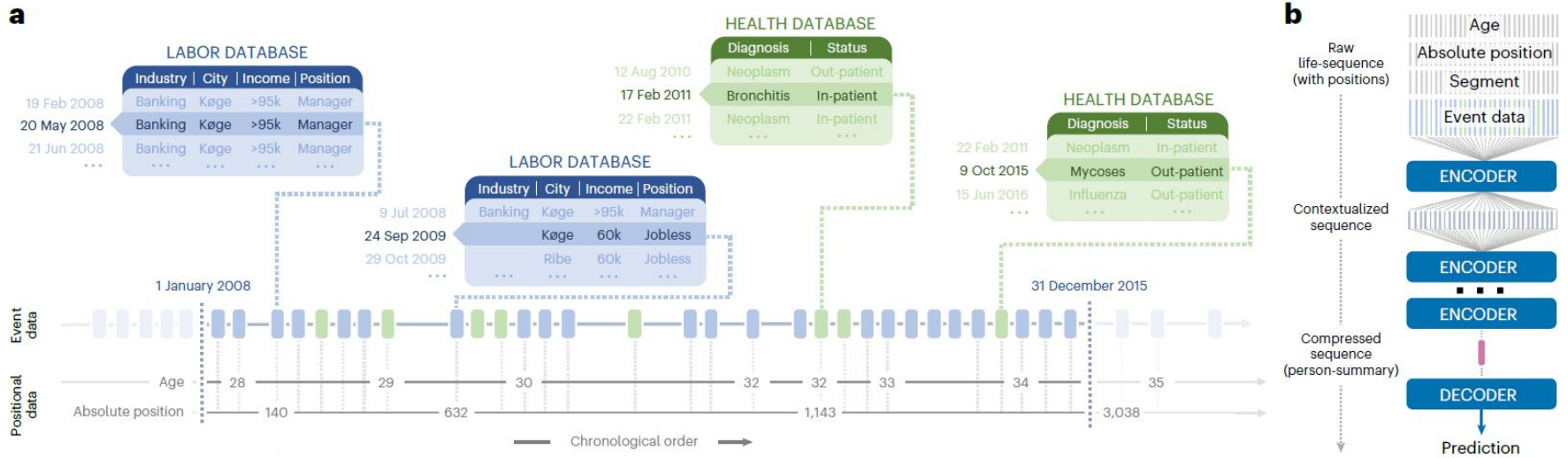


Fig. 1 | A schematic individual-level data representation for the life2vec model. a,**b**, We organize socio-economic and health data from the Danish national registers from 1 January 2008 to 31 December 2015 into a single chronologically ordered life-sequence (**a**). Each database entry becomes an event in the sequence, where an event has associated positional and contextual data. The contextual data include variables associated with the entry (for example, industry, city, income and job type). The positional data include the person's age (expressed in full years) and absolute position (number of days since 1 January 2008). The raw life-sequence is then passed to the model described in **b**. The model consists of multiple stacked encoders. The first encoder combines contextual and positional information to produce a contextual representation of each life-event. The following encoders output deep contextual representations of each life-event (considering the overall content of the life-sequence). The final encoder layer fuses the representations of life-events to produce the representation of a life-sequence. The decoder uses the latter to make predictions.

Life2Vec

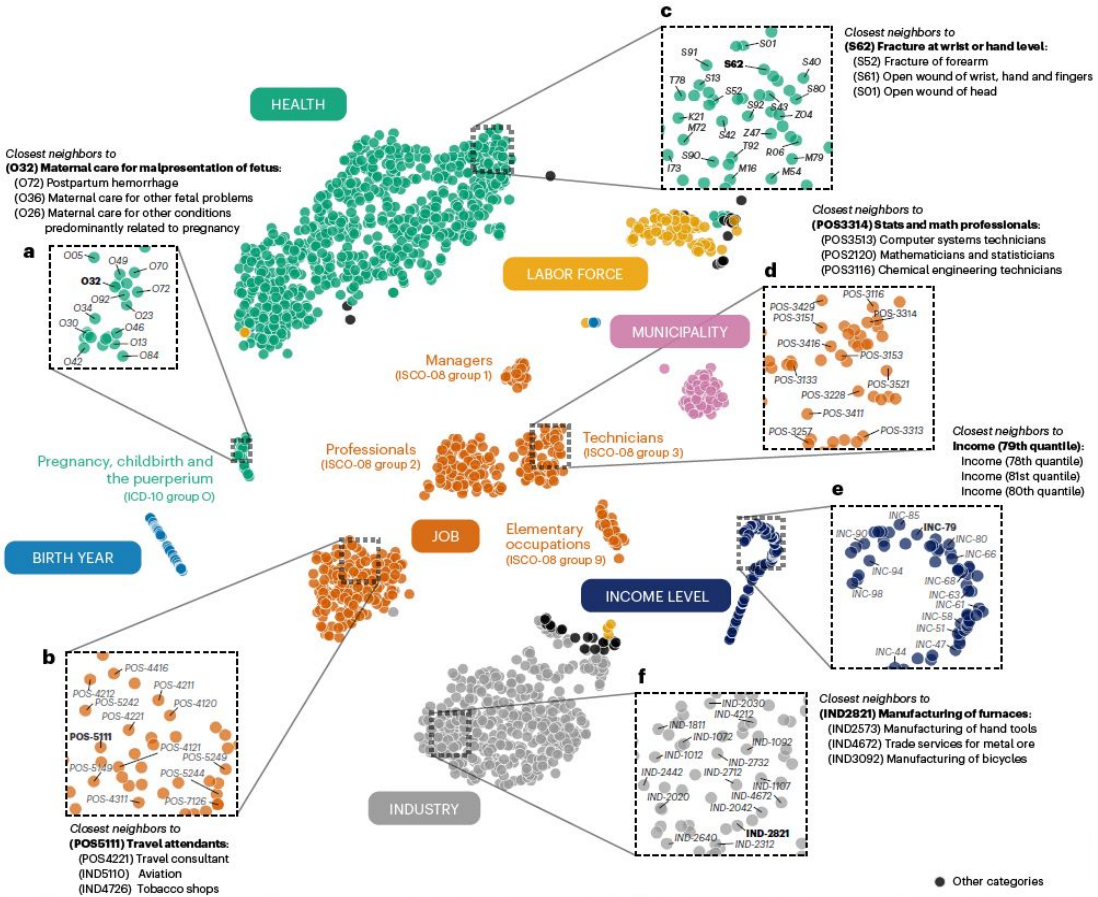


Fig. 2 | Two-dimensional projection of the concept space (using PaCMAP). Each point corresponds to a concept token in the vocabulary ($n = 2{,}043$). Points are colored based on the concept types (infrequent types are represented as black points). Each region provides a zoom of a part of the concept space. The top three closest neighbors for selected tokens (based on the cosine distance) are also displayed. **a**, Diagnoses related to pregnancy, childbirth and the puerperium in ICD-10[27]. **b**, Job concepts related to service and sales workers (corresponds to job category 5 of ISCO-08[25]). **c**, Injury-related diagnoses in ICD-10[27]. **d**, Job concepts related to technicians and associate professionals (corresponds to job category 3 of ISCO-08[25]). **e**, Income-related concepts. life2vec arranges these concepts in increasing ordinal order. **f**, Concepts related to the manufacturing industry in DB07[26].

Life2Vec

Matthews Correlation Coefficient (MCC)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

- MCC considers all four outcomes of the confusion matrix
- Values range from −1 to +1
- +1 indicates perfect prediction
- 0 indicates random prediction
- −1 indicates total disagreement

Life2Vec

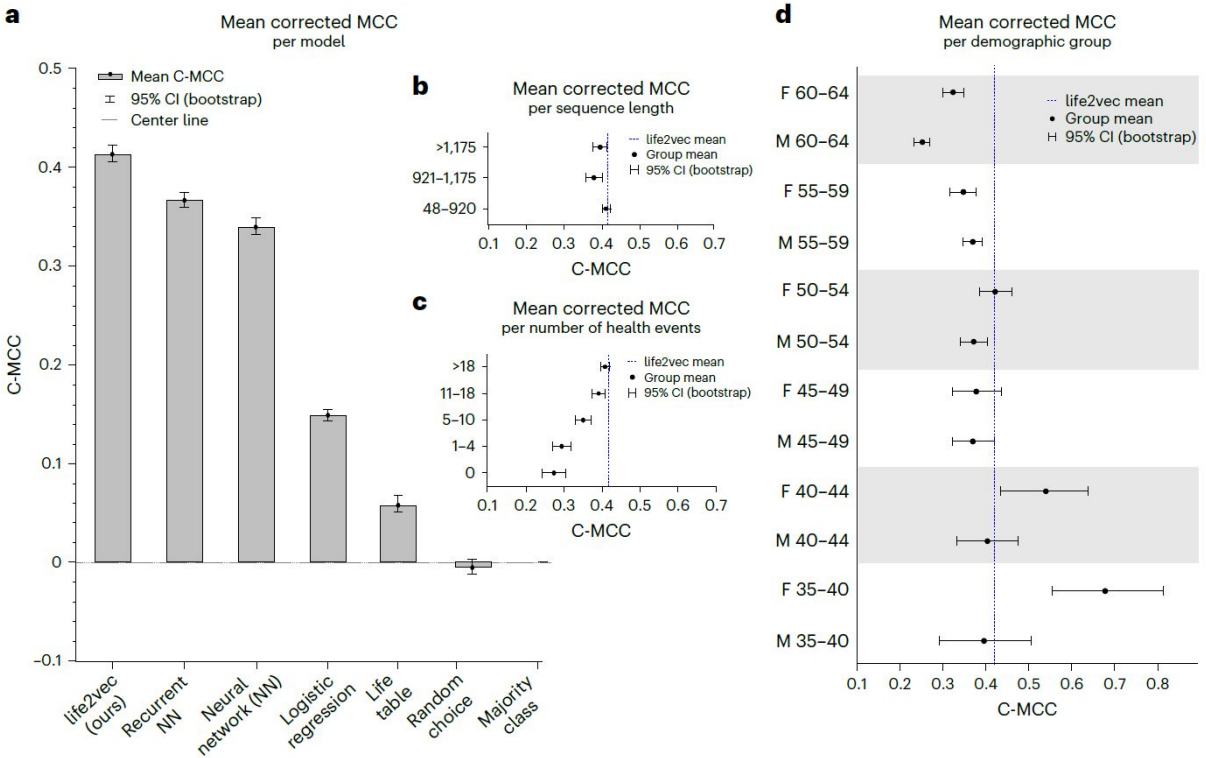
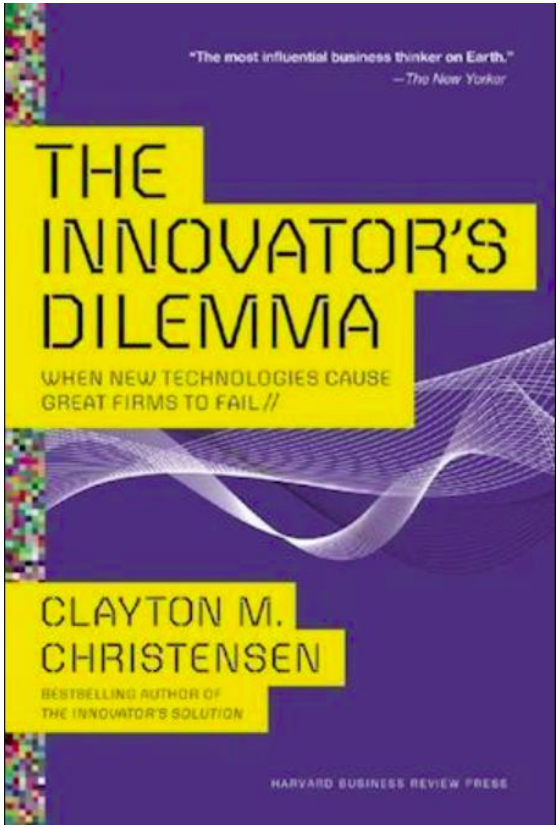


Fig. 3 | Performance of models on the mortality prediction task quantified with the mean C-MCC with 95% confidence interval. a, Comparison of life2vec performance to baselines ($n = 100{,}000$). **b**–**d**, Performance of life2vec on different cohorts of the population: performance of life2vec per sequence length (**b**), performance of life2vec based on the number of health events in a sequence (**c**) and performance of life2vec per intersectional group (based on age group and sex) (**d**). F, female; M, male.

Transformers and BDS



Successful firms often fail:

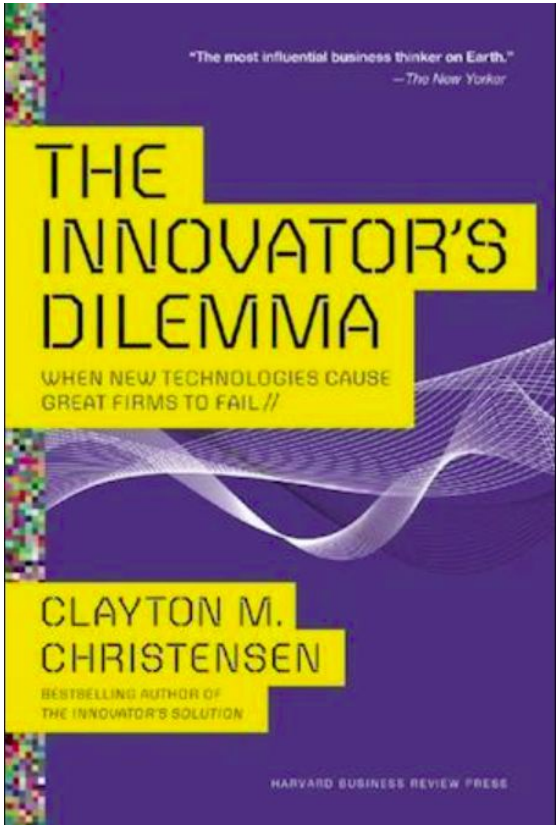
- not because they are poorly managed
- because they focus too strongly on existing customers and current technologies

Disruptive innovations start as simpler, cheaper, or lower-performance alternatives

- Disruptive innovations initially serve niche or low-end markets
- Over time, they improve and displace established technologies and firms
- The book conceptualizes the role of timing, sequencing, and path dependency in innovation

Transformers and BDS

Related and unrelated diversification (remember?)



The Principle of Relatedness

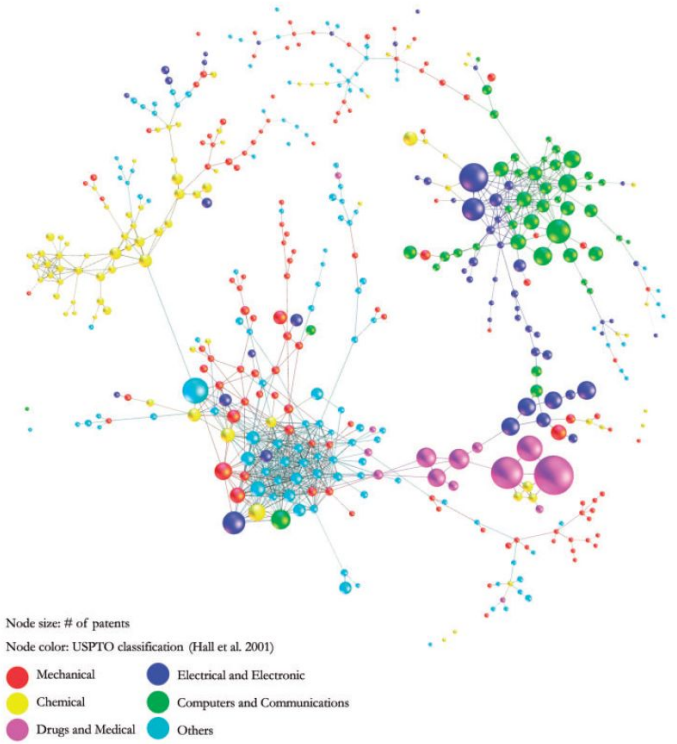
César A. Hidalgo[1(✉)], Pierre-Alexandre Balland[2], Ron Boschma[2,3], Mercedes Delgado[4], Maryann Feldman[5], Koen Frenken[6], Edward Glaeser[7,8], Canfei He[9], Dieter F. Kogler[10], Andrea Morrison[2], Frank Neffke[11], David Rigby[12], Scott Stern[4,8], Siqi Zheng[13,14], and Shengjun Zhu[9]

[1] Collective Learning Group, the MIT Media Lab, Massachusetts Institute of Technology, Cambridge, USA
hidalgo@mit.edu
[2] Department of Human Geography and Planning, Utrecht University, Utrecht, Netherlands
[3] Business School, University of Stavanger, Stavanger, Norway
[4] MIT Sloan School of Management, Cambridge, USA
[5] Department of Public Policy, University of North Carolina at Chapel Hill, Chapel Hill, USA
[6] Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, Netherlands
[7] Department of Economics, Harvard University, Cambridge, USA
[8] National Bureau of Economic Research, Cambridge, USA
[9] College of Urban and Environmental Sciences, Peking University, Beijing, China
[10] Spatial Dynamics Lab and School of Architecture Planning and Environmental Policy, University College Dublin, Dublin, Ireland
[11] Center for International Development, Harvard University, Cambridge, USA
[12] Department of Geography and Department of Statistics, University of California, Los Angeles, USA
[13] Department of Urban Studies and Planning, MIT, Cambridge, USA
[14] School of Civil Engineering, Tsinghua University, Beijing, China

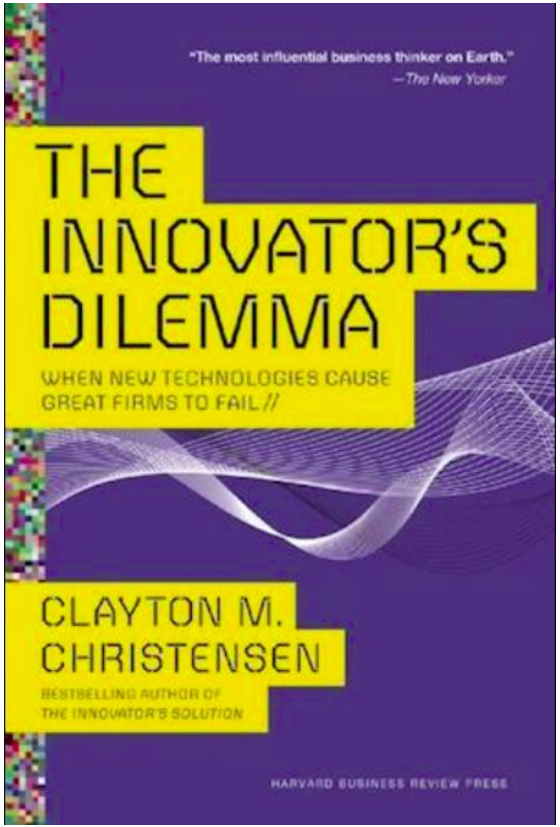
Abstract. The idea that skills, technology, and knowledge, are spatially concentrated, has a long academic tradition. Yet, only recently this hypothesis has been empirically formalized and corroborated at multiple spatial scales, for different economic activities, and for a diversity of institutional regimes. The new synthesis is an empirical principle describing the probability that a region enters—or exits—an economic activity as a function of the number of related activities present in that location. In this paper we summarize some of the recent empirical evidence that has generalized the principle of relatedness to a fact describing the entry and exit of products, industries, occupations, and technologies, at the national, regional, and metropolitan scales. We conclude by describing some of the policy implications and future avenues of research implied by this robust empirical principle.

C. A. Hidalgo and P.-A. Balland—Contributed equally.

A. J. Morales et al. (Eds.): ICCS 2018, SPCOM, pp. 451–457, 2018.
https://doi.org/10.1007/978-3-319-96661-8_46



Transformers and BDS

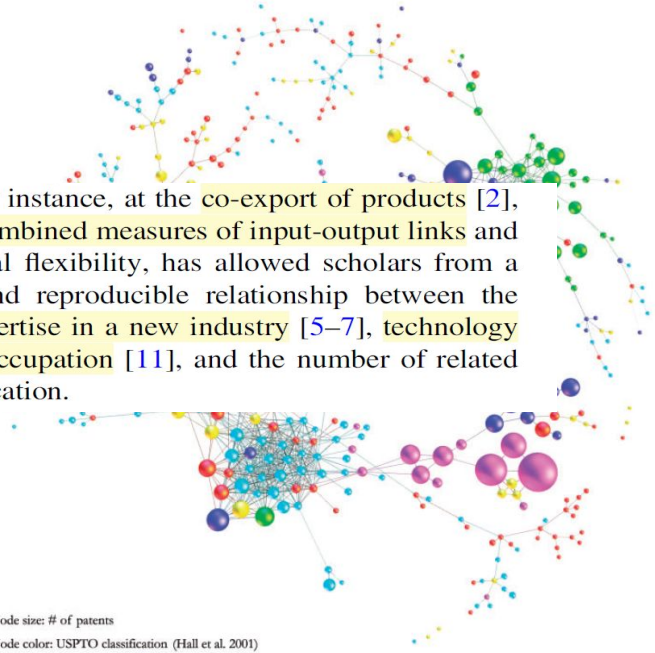


Related and unrelated diversification (remember?)

The Principle of Relatedness

César A. Hidalgo[1(✉)], Pierre-Alexandre Balland[2], Ron Boschma[2,3], Mercedes Delgado[4], Maryann Feldman[5], Koen Frenken[6], Edward Glaeser[7,8], Canfei He[9], Dieter F. Kogler[10], Andrea Morrison[2], Frank Neffke[11], David Rigby[12], Scott Stern[4,8], Siqi Zheng[13,14], and Shengjun Zhu[9]

[1] Collective Learning Group, the MIT Media Lab, Massachusetts Institute of Technology, Cambridge, USA



relatedness among activities. They look, for instance, at the co-export of products [2], the flow of labor among industries [3], or combined measures of input-output links and shared labor pools [4]. This methodological flexibility, has allowed scholars from a variety of fields to document a robust and reproducible relationship between the probability that a location will develop expertise in a new industry [5–7], technology [8, 9], research area [10], product [2], or occupation [11], and the number of related activities that are already present in that location.

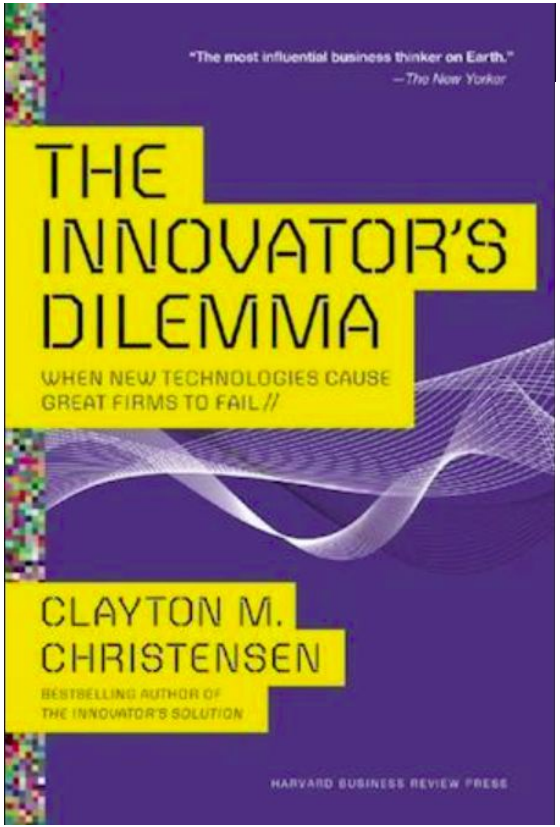
Hidalgo et al. 2018 (p. 452)

Abstract. The idea that skills, technology, and knowledge, are spatially concentrated, has a long academic tradition. Yet, only recently this hypothesis has been empirically formalized and corroborated at multiple spatial scales, for different economic activities, and for a diversity of institutional regimes. The new synthesis is an empirical principle describing the probability that a region enters—or exits—an economic activity as a function of the number of related activities present in that location. In this paper we summarize some of the recent empirical evidence that has generalized the principle of relatedness to a fact describing the entry and exit of products, industries, occupations, and technologies, at the national, regional, and metropolitan scales. We conclude by describing some of the policy implications and future avenues of research implied by this robust empirical principle.

C. A. Hidalgo and P.-A. Balland—Contributed equally.

A. J. Morales et al. (Eds.): ICCS 2018, SPCOM, pp. 451–457, 2018.
https://doi.org/10.1007/978-3-319-96661-8_46

Transformers and BDS



Related and unrelated diversification (remember?)

Related Variety and Regional Development: A Critique

Harald Bathelt
Department of Geography and Planning
University of Toronto
Toronto, Ontario M5S 3G3
Canada
harald.bathelt@utoronto.ca

Michael Storper
Department of Geography and Environment
London School of Economics
London WC2A 2AE
UK
and
Luskin School of Public Affairs
UCLA
Los Angeles, CA 90095-1656
m.storper@lse.ac.uk

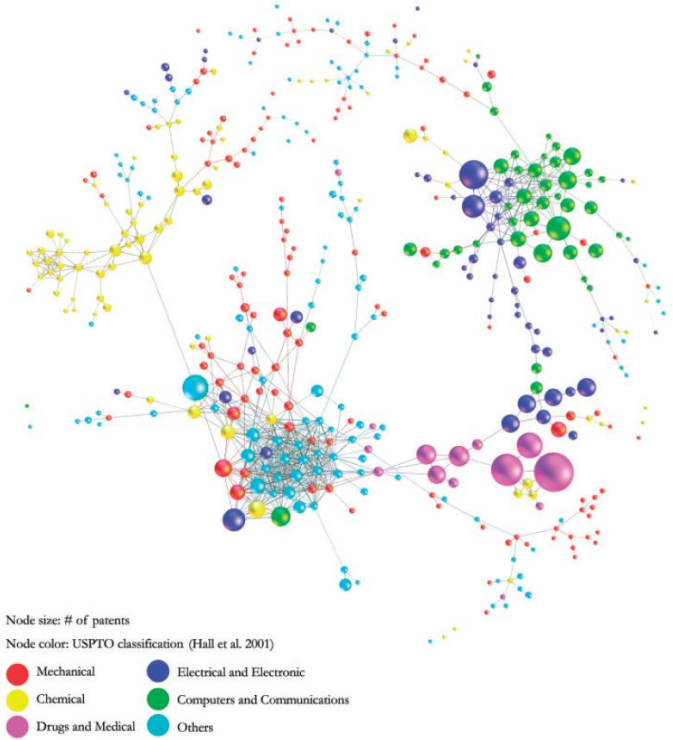
Key words:
economic geographies of places
evolutionary economic geography (EEG)
regional development
regional specialization
related variety

JEL codes:
L23
R11

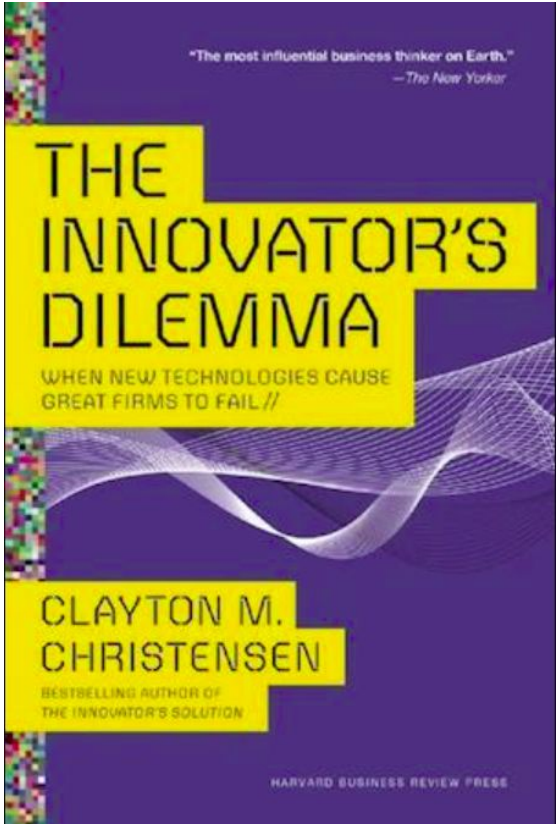
abstract

Evolutionary approaches in economic geography have contributed substantially to the growing body of knowledge of regional development processes and their underlying mechanisms. One key concept in the literature on evolutionary economic geography is that of *related variety*. Herein, regional industry structure is represented through the level of related variety of technologies, skills, or outputs. The related variety concept proposes that regional economic development is favored when an economy diversifies into products or technologies that are closely related to the stock of existing activities. In this article, we raise substantive questions regarding the internal logic of the concept of related variety, its spatial expressions, measurement specifics, empirical regularities and biases, and its possible short- and long-term effects on regional development. Based on this investigation, we make suggestions for improvements to future research.

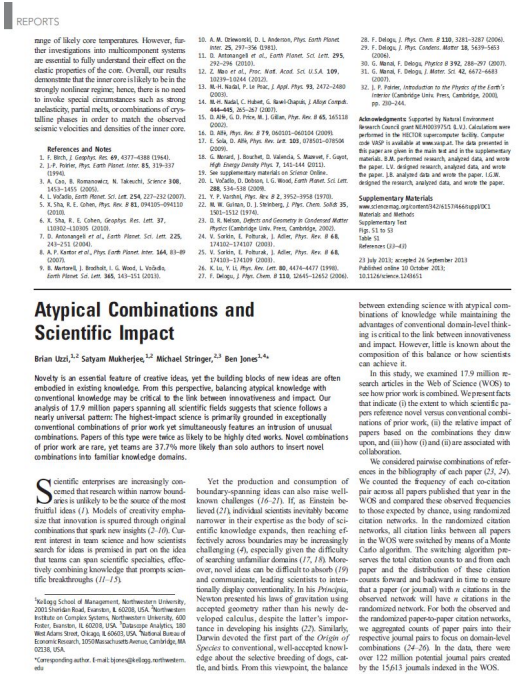
ECONOMIC GEOGRAPHY 441
99(5):441–470 © 2023 Clark University. www.economicgeography.org



Transformers and BDS



A BDS approach identifying novel combinations in firms (remember?)



REPORTS

range of likely core temperatures. However, further investigations into multicomponent systems are essential to fully understand their effect on the elastic properties of the core. Overall, our results demonstrate that the inner core is likely to be in the strongly nonlinear regime; hence, there is no need to invoke special circumstances such as strong anelasticity, partial melts, or combinations of crystalline phases in order to match the observed seismic velocities and densities of the inner core.

References and Notes

1. F. Birch, *J. Geophys. Res.* **69**, 4377–4388 (1964).
2. J.-P. Poirier, *Phys. Earth Planet. Inter.* **85**, 319–337 (1994).
3. A. Cao, B. Romanowicz, N. Takeuchi, *Science* **308**, 1453–1455 (2005).
4. L. Vočadlo, *Earth Planet. Sci. Lett.* **254**, 227–232 (2007).
5. X. Shu, R. E. Cohen, *Phys. Rev. B* **81**, 094105–094110 (2010).
6. X. Shu, R. E. Cohen, *Geophys. Res. Lett.* **37**, L10302–L10305 (2010).
7. D. Antonangeli *et al.*, *Earth Planet. Sci. Lett.* **225**, 243–251 (2004).
8. A. P. Kantor *et al.*, *Phys. Earth Planet. Inter.* **164**, 83–89 (2007).
9. B. Martorell, J. Brodholt, I. G. Wood, L. Vočadlo, *Earth Planet. Sci. Lett.* **365**, 143–151 (2013).
10. A. M. Dziewonski, D. L. Anderson, *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
11. D. Antonangeli *et al.*, *Earth Planet. Sci. Lett.* **295**, 292–296 (2010).
12. Z. Mao *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10239–10244 (2012).
13. M.-H. Nadal, P. Le Poac, *J. Appl. Phys.* **93**, 2472–2480 (2003).
14. M.-H. Nadal, C. Hubet, G. Rawi-Chapsi, *J. Alloys Compds.* **446–445**, 265–267 (2007).
15. D. Alfè, G. D. Price, M. J. Gillan, *Phys. Rev. B* **65**, 165118 (2002).
16. D. Alfè, *Phys. Rev. B* **79**, 060101–060104 (2009).
17. E. Sola, D. Alfè, *Phys. Rev. Lett.* **103**, 078501–078504 (2009).
18. G. Morard, J. Bouchet, D. Valencia, S. Mazevet, F. Guyot, *High Energy Density Phys.* **7**, 141–144 (2011).
19. See supplementary materials on *Science* Online.
20. L. Vočadlo, D. Dobson, I. G. Wood, *Earth Planet. Sci. Lett.* **288**, 534–538 (2009).
21. Y. P. Varshni, *Phys. Rev. B* **2**, 3952–3958 (1970).
22. M. W. Guinan, D. J. Steinberg, *J. Phys. Chem. Solids* **35**, 1501–1512 (1974).
23. D. R. Nelson, *Defects and Geometry in Condensed Matter Physics* (Cambridge Univ. Press, Cambridge, 2002).
24. V. Sarkis, E. Poltorak, J. Adler, *Phys. Rev. B* **68**, 174102–174107 (2003).
25. V. Sarkis, E. Poltorak, J. Adler, *Phys. Rev. B* **68**, 174103–174109 (2003).
26. K. Lu, Y. Li, *Phys. Rev. Lett.* **80**, 4474–4477 (1998).
27. F. Delogu, *J. Phys. Chem. B* **110**, 12645–12652 (2006).
28. F. Delogu, *J. Phys. Chem. B* **110**, 3281–3287 (2006).
29. F. Delogu, *J. Phys. Condens. Matter* **18**, 5639–5653 (2006).
30. G. Morard, F. Delogu, *Physica B* **392**, 288–297 (2007).
31. G. Morard, F. Delogu, *J. Mater. Sci.* **42**, 6672–6683 (2007).
32. J. P. Poirier, *Introduction to the Physics of the Earth's Interior* (Cambridge Univ. Press, Cambridge, 2000), pp. 200–244.

Acknowledgments: Supported by Natural Environment Research Council grant NE/H003975/1 (L.V.). Calculations were performed in the HECTOR supercomputer facility. Computer code VASP is available at www.vasp.at. The data presented in this paper are given in the main text and in the supplementary materials. B.M. performed research, analyzed data, and wrote the paper. L.V. designed research, analyzed data, and wrote the paper. J.B. analyzed data and wrote the paper. I.G.W. designed the research, analyzed data, and wrote the paper.

Supplementary Materials

www.sciencemag.org/content/342/6157/466/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S3
Table S1
References (33–43)

23 July 2013; accepted 26 September 2013
Published online 10 October 2013;
10.1126/science.1243651

Atypical Combinations and Scientific Impact

Brian Uzzi,[1,2] Satyam Mukherjee,[1,2] Michael Stringer,[2,3] Ben Jones[1,4]*

Novelty is an essential feature of creative ideas, yet the building blocks of new ideas are often embodied in existing knowledge. From this perspective, balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact. Our analysis of 17.9 million papers spanning all scientific fields suggests that science follows a nearly universal pattern: The highest-impact science is primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of unusual combinations. Papers of this type were twice as likely to be highly cited works. Novel combinations of prior work are rare, yet teams are 37.7% more likely than solo authors to insert novel combinations into familiar knowledge domains.

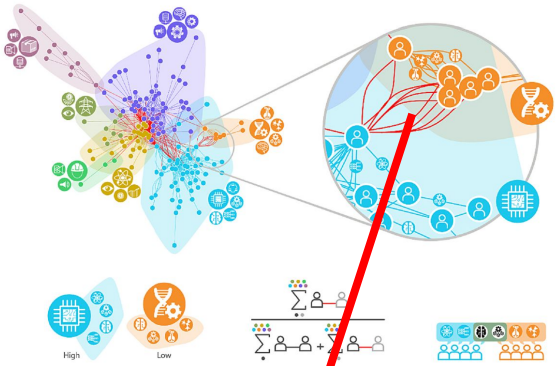
Scientific enterprises are increasingly consumed that research within narrow boundaries is unlikely to be the source of the most fruitful ideas (*1*). Models of creativity emphasize that innovation is spurred through original combinations that spark new insights (*2*–*10*). Current interest in team science and how scientists search for ideas is premised in part on the idea that teams can span scientific specialties, effectively combining knowledge that prompts scientific breakthroughs (*11*–*15*).

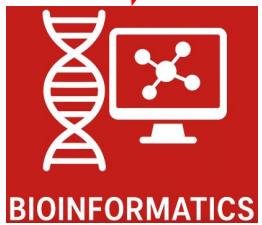
Yet the production and consumption of boundary-spanning ideas can also raise well-known challenges (*16*–*21*). If, as Einstein believed (*17*), individual scientists inevitably become narrower in their expertise as the body of scientific knowledge expands, then reaching effectively across boundaries may be increasingly challenging (*4*), especially given the difficulty of searching unfamiliar domains (*17*, *18*). Moreover, novel ideas can be difficult to absorb (*19*) and communicate, leading scientists to intentionally display conventionality. In his *Principia*, Newton presented his laws of gravitation using accepted geometry rather than his newly developed calculus, despite the latter's importance in developing his insights (*22*). Similarly, Darwin devoted the first part of the *Origin of Species* to conventional, well-accepted knowledge about the selective breeding of dogs, cattle, and birds. From this viewpoint, the balance between extending science with atypical combinations of knowledge while maintaining the advantages of conventional domain-level thinking is critical to the link between innovativeness and impact. However, little is known about the composition of this balance or how scientists can achieve it.

In this study, we examined 17.9 million research articles in the Web of Science (WOS) to see how prior work is combined. We present facts that indicate (i) the extent to which scientific papers reference novel versus conventional combinations of prior work, (ii) the relative impact of papers based on the combinations they draw upon, and (iii) how (i) and (ii) are associated with collaboration.

We considered pairwise combinations of references in the bibliography of each paper (*23*, *24*). We counted the frequency of each co-citation pair across all papers published that year in the WOS and compared these observed frequencies to those expected by chance, using randomized citation networks. In the randomized citation networks, all citation links between all papers in the WOS were switched by means of a Monte Carlo algorithm. The switching algorithm preserves the total citation counts to and from each paper and the distribution of these citation counts forward and backward in time to ensure that a paper (or journal) with n citations in the observed network will have n citations in the randomized network. For both the observed and the randomized paper-to-paper citation networks, we aggregated counts of paper pairs into their respective journal pairs to focus on domain-level combinations (*24*–*26*). In the data, there were over 122 million potential journal pairs created by the 15,613 journals indexed in the WOS.

[1]Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, 600 Foster, Evanston, IL 60208, USA. [3]Datasupp Analytics, 180 West Adams Street, Chicago, IL 60603, USA. [4]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: bjones@kellogg.northwestern.edu



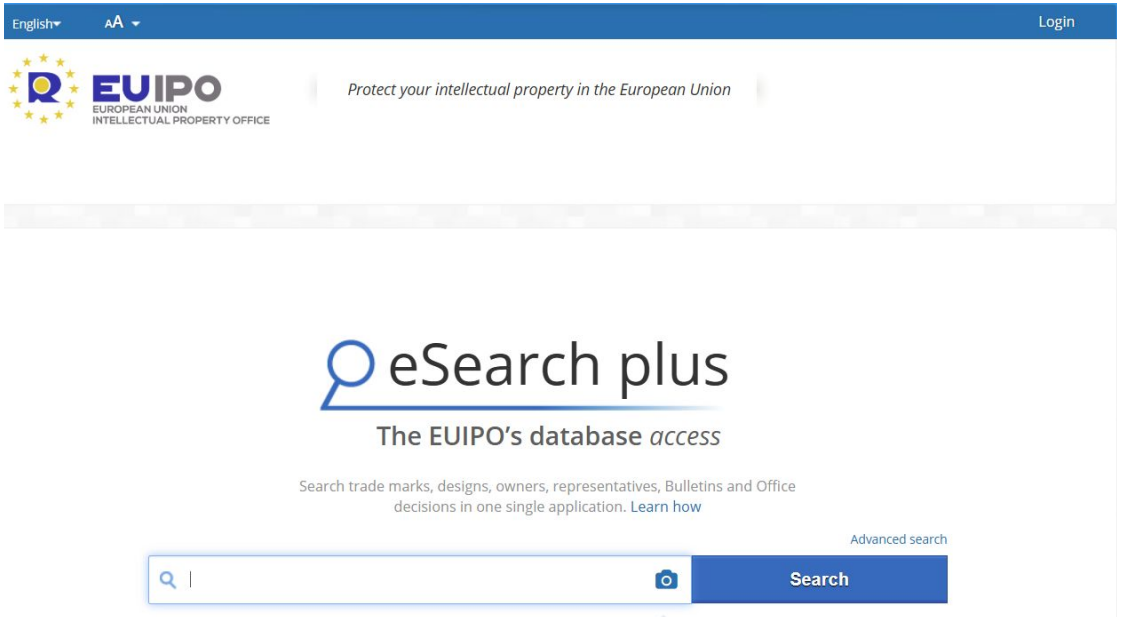


Market2Vec (or MarketBert)

Can we predict a firm’s next products (or product features) based on its past products?

Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?



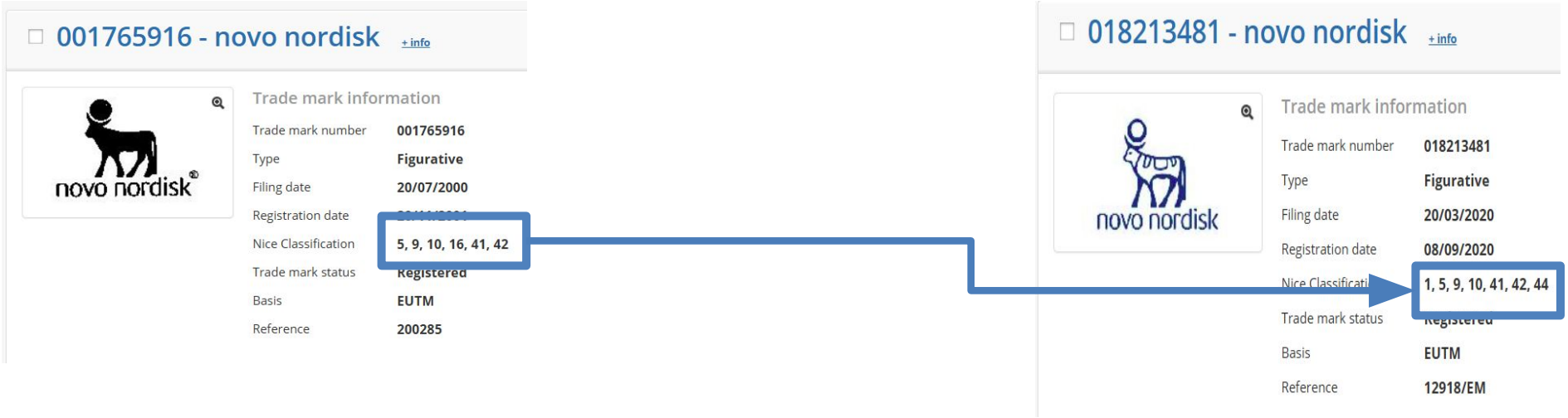
https://euipo.europa.eu/eSearch/

Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?

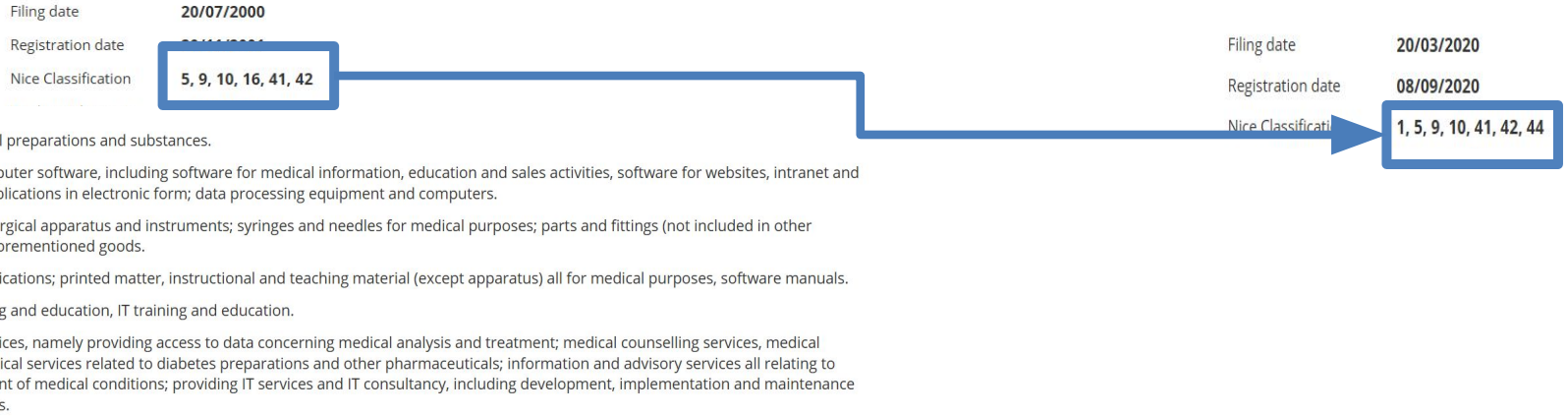
Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?



Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?



5 Pharmaceutical preparations and substances.

9 Recorded computer software, including software for medical information, education and sales activities, software for websites, intranet and Internet; medical publications in electronic form; data processing equipment and computers.

10 Medical and surgical apparatus and instruments; syringes and needles for medical purposes; parts and fittings (not included in other classes) for all the aforementioned goods.

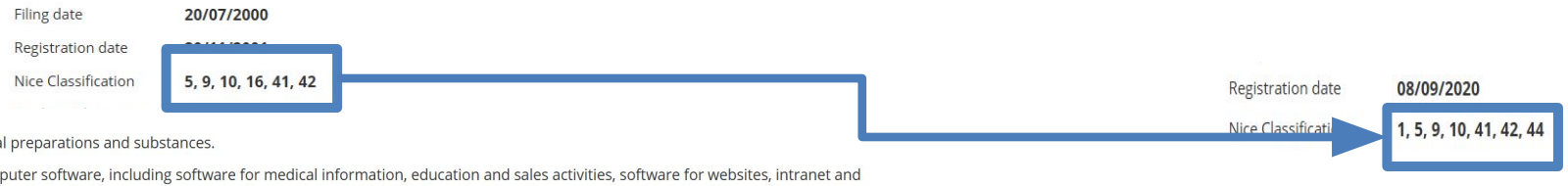
16 Periodical publications; printed matter, instructional and teaching material (except apparatus) all for medical purposes, software manuals.

41 Medical training and education, IT training and education.

42 Computer services, namely providing access to data concerning medical analysis and treatment; medical counselling services, medical clinical services, medical services related to diabetes preparations and other pharmaceuticals; information and advisory services all relating to personal management of medical conditions; providing IT services and IT consultancy, including development, implementation and maintenance of IT system solutions.

Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?



5 Pharmaceutical preparations and substances.

9 Recorded computer software, including software for medical information, education and sales activities, software for websites, intranet and Internet; medical publications in electronic form; data processing equipment and computers.

10 Medical and surgical apparatus and instruments; syringes and needles for medical purposes; parts and fittings (not included in other classes) for all the aforementioned goods.

16 Periodical publications; printed matter, instructional and teaching material (except apparatus) all for medical purposes, software manuals.

41 Medical training and education, IT training and education.

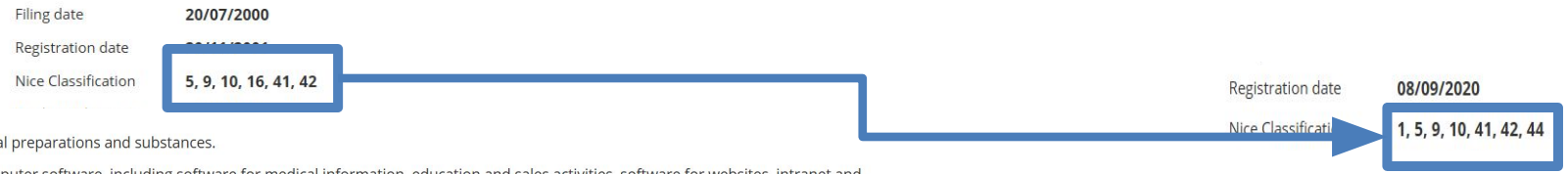
42 Computer services, namely providing access to data concerning medical analysis and treatment; medical counselling services, medical clinical services, medical services related to diabetes preparations and other pharmaceuticals; information and advisory services all relating to personal management of medical conditions; providing IT services and IT consultancy, including development, implementation and maintenance of IT system solutions.

1 Chemical preparations for use in industry; chemical additives; chemical additives for use in the manufacture of pharmaceuticals, cosmetics, hair lotions, disinfectants; chemical diagnostic reagents for industrial use; chemical preparations for use in the oil industry.

44 Providing medical services via websites and mobile apps for the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment; providing medical information and medical counselling the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment.

Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?



5 Pharmaceutical preparations and substances.

9 Recorded computer software, including software for medical information, education and sales activities, software for websites, intranet and Internet; medical publications in electronic form; data processing equipment and computers.

10 Medical and surgical apparatus and instruments; syringes and needles for medical purposes; parts and fittings (not included in other classes) for all the aforementioned goods.

16 Periodical publications; printed matter, instructional and teaching material (except apparatus) all for medical purposes, software manuals.

41 Medical training and education, IT training and education.

42 Computer services, namely providing access to data concerning medical analysis and treatment; medical counselling services, medical clinical services, medical services related to diabetes preparations and other pharmaceuticals; information and advisory services all relating to personal management of medical conditions; providing IT services and IT consultancy, including development, implementation and maintenance of IT system solutions.

1 Chemical preparations for use in industry; chemical additives; chemical additives for use in the manufacture of pharmaceuticals, cosmetics, hair lotions, disinfectants, chemical diagnostic reagents for industrial use; chemical preparations for use in the oil industry.

Concept (or market) token?

44 Providing medical services via websites and mobile apps for the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment; providing medical information and medical counselling the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment.

Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?

5 Pharmaceutical preparations and substances.

9 Recorded computer software, including software for medical information, education and sales activities, software for websites, intranet and Internet; medical publications in electronic form; data processing equipment and computers.

10 Medical and surgical apparatus and instruments; syringes and needles for medical purposes; parts and fittings (not included in other classes) for all the aforementioned goods.

16 Periodical publications; printed matter, instructional and teaching material (except apparatus) all for medical purposes, software manuals.

41 Medical training and education, IT training and education.

42 Computer services, namely providing access to data concerning medical analysis and treatment; medical counselling services, medical clinical services, medical services related to diabetes preparations and other pharmaceuticals; information and advisory services all relating to personal management of medical conditions; providing IT services and IT consultancy, including development, implementation and maintenance of IT system solutions.

Challenge!!
How to construct market token sequences?

1 Chemical preparations for use in industry; chemical additives; chemical additives for use in the manufacture of pharmaceuticals, cosmetics, hair lotions, disinfectants, chemical diagnostic reagents for industrial use; chemical preparations for use in the oil industry.

Concept (or market) token?

44 Providing medical services via websites and mobile apps for the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment; providing medical information and medical counselling the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment.

Market2Vec (or MarketBert)

Can we predict a firm's next products (or product features) based on its past products?

5 Pharmaceutical preparations and substances.

9 Recorded computer software, including software for medical information, education and sales activities, software for websites, intranet and Internet; medical publications in electronic form; data processing equipment and computers.

10 Medical and surgical apparatus and instruments; syringes and needles for medical purposes; parts and fittings (not included in other classes) for all the aforementioned goods.

16 Periodical publications; printed matter, instructional and teaching material (except apparatus) all for medical purposes, software manuals.

41 Medical training and education, IT training and education.

42 Computer services, namely providing access to data concerning medical analysis and treatment; medical counselling services, medical clinical services, medical services related to diabetes preparations and other pharmaceuticals; information and advisory services all relating to personal management of medical conditions; providing IT services and IT consultancy, including development, implementation and maintenance of IT system solutions.

Challenge!!
How to construct market token sequences?

1 Chemical preparations for use in industry; chemical additives; chemical additives for use in the manufacture of pharmaceuticals, cosmetics, hair lotions, disinfectants, chemical diagnostic reagents for industrial use; chemical preparations for use in the oil industry.

Concept (or market) token?

44 Providing medical services via websites and mobile apps for the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment; providing medical information and medical counselling the management of diabetes, obesity, haemostasis management and for the treatment of sickle cell disease, treatment of brain disorders, non-alcoholic steatohepatitis, atherosclerosis, cardiovascular disease, chronic kidney disease, hormone therapy and stem cell treatment.

Market2Vec (o

Can we predict a firm's next products
past products?

s

Filing date	20/07/2000
Registration date	
Nice Classification	5, 9, 10, 16, 41, 42

5 Pharmaceutical preparations and substances.

9 Recorded computer software, including software for medical information, education and sales activities, software for websites, Internet; medical publications in electronic form; data processing equipment and computers.

10 Medical and surgical apparatus and instruments; syringes and needles for medical purposes; parts and fittings (not included in classes) for all the aforementioned goods.

16 Periodical publications; printed matter, instructional and teaching material (except apparatus) all for medical purposes, softwar

41 Medical training and education, IT training and education.

42 Computer services, namely providing access to data concerning medical analysis and treatment; medical counselling services, clinical services, medical services related to diabetes preparations and other pharmaceuticals; information and advisory services all re personal management of medical conditions; providing IT services and IT consultancy, including development, implementation and m of IT system solutions.

1 Chemical preparations for use in industry; c
hair lotions, disinfectants, chemical diagnostic reag

44 Providing medical services via web
treatment of sickle cell disease, treatmer
kidney disease, hormone therapy and st
obesity, haemostasis management and f
atherosclerosis, cardiovascular disease,

Concept (or market) token?

ces?

osmetics,

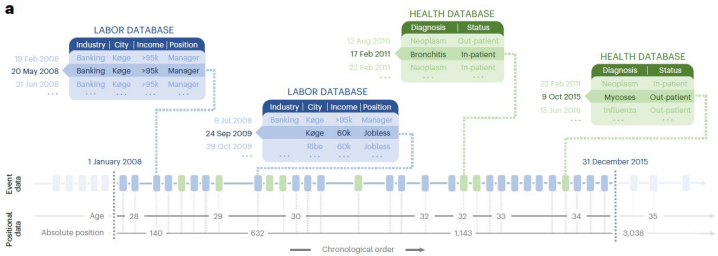
d for the
ronic
diabetes,
itis,

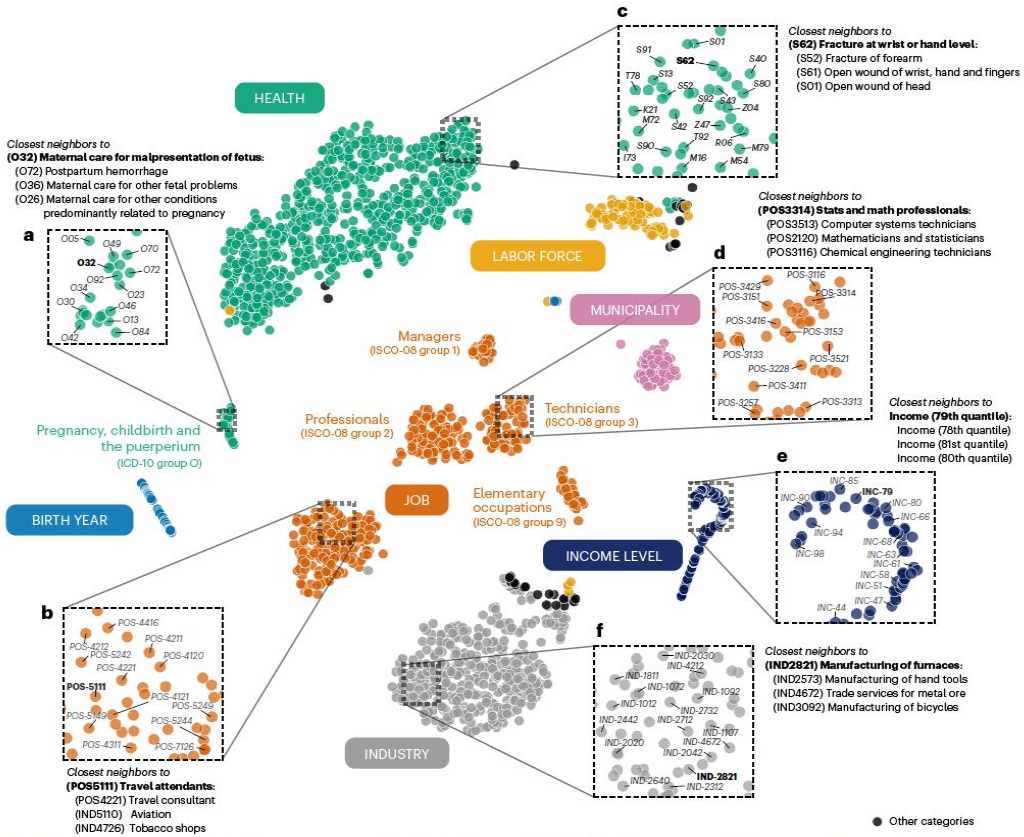
```
# ============================================================
# 4) TRAINING LOOP (4 steps: forward -> loss -> backward -> update)
# ============================================================
def train_one_epoch(model, train_loader, optimizer, scheduler):
    model.train()
    total = 0.0
    pbar = tqdm(train_loader, desc="Train", leave=False)

    for batch in pbar:
        batch = {k: v.to(device) for k, v in batch.items()}

        # (1) Forward pass
        outputs = model(**batch)

        # (2) Loss
        loss = outputs.loss

        # (3) Backward
        loss.backward()
        torch.nn.utils.clip_grad_norm_(model.parameters(), GRAD_CLIP)

        # (4) Update
        optimizer.step()
        scheduler.step()
        optimizer.zero_grad()

        total += loss.item()
        pbar.set_postfix(loss=f"{loss.item():.4f}")

    return total / max(1, len(train_loader))
```

Market2Vec (or MarketBert)

Embedding and clustering

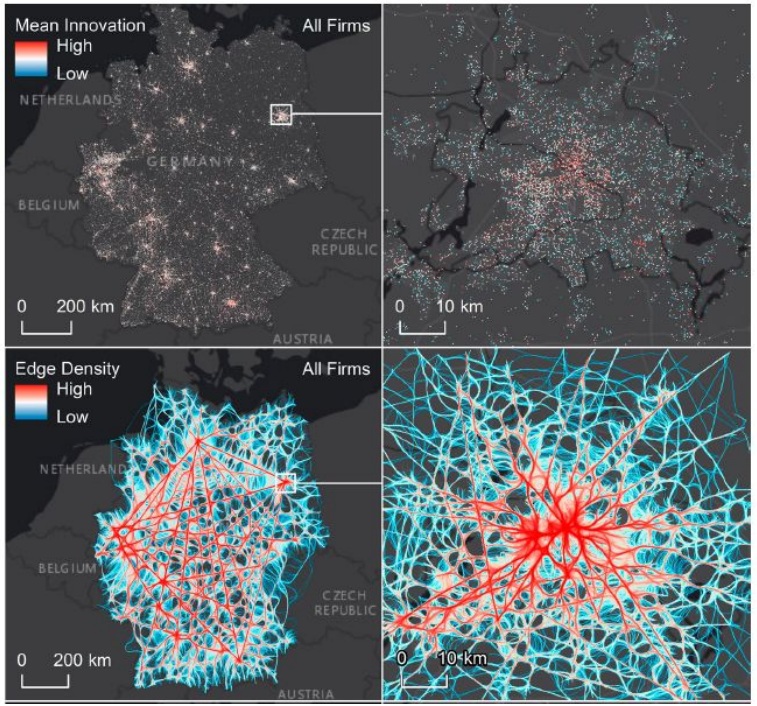
- Text-based pretrained model
- Pre-training a model from scratch





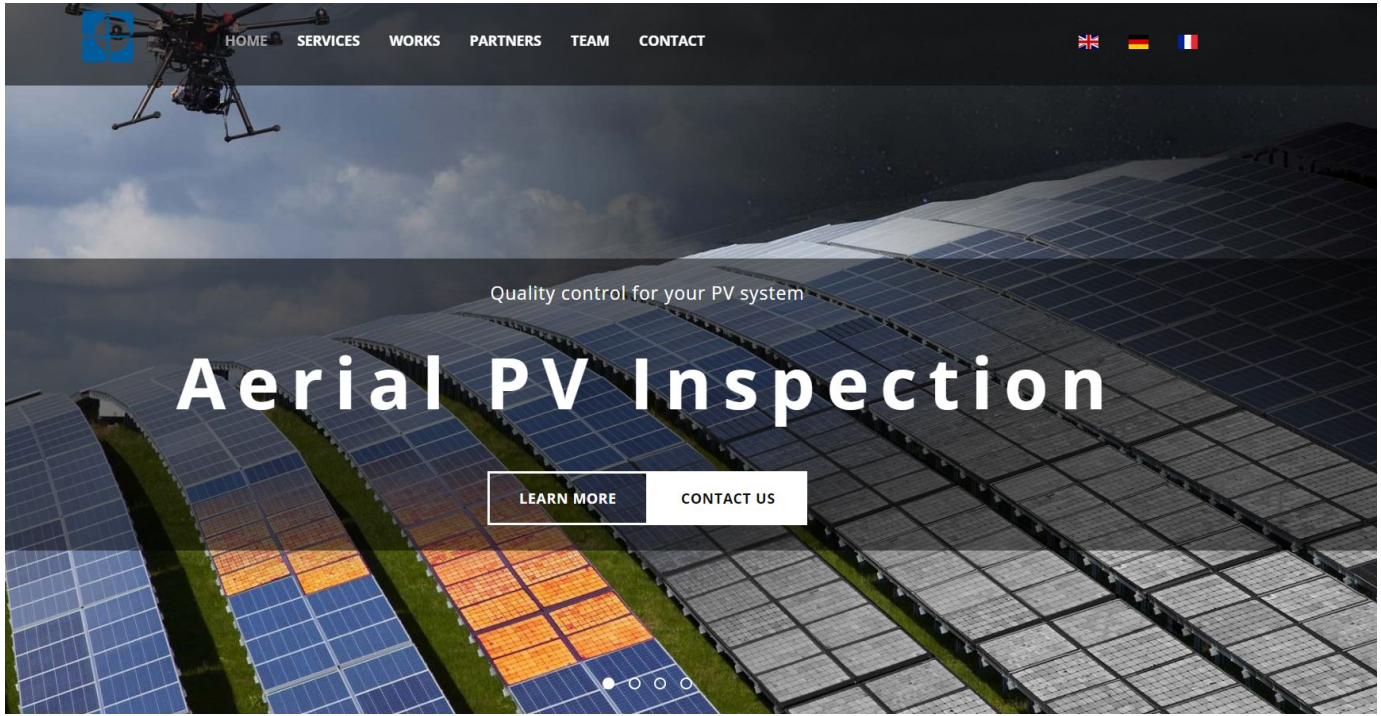
Sentence Transformers Finetuning (SetFit)

Text-based pretrained model (using SetFit for finetuning)



Sentence Transformers Finetuning (SetFit)

Text-based pretrained model (using SetFit for finetuning)



Sentence Transformers Finetuning (SetFit)

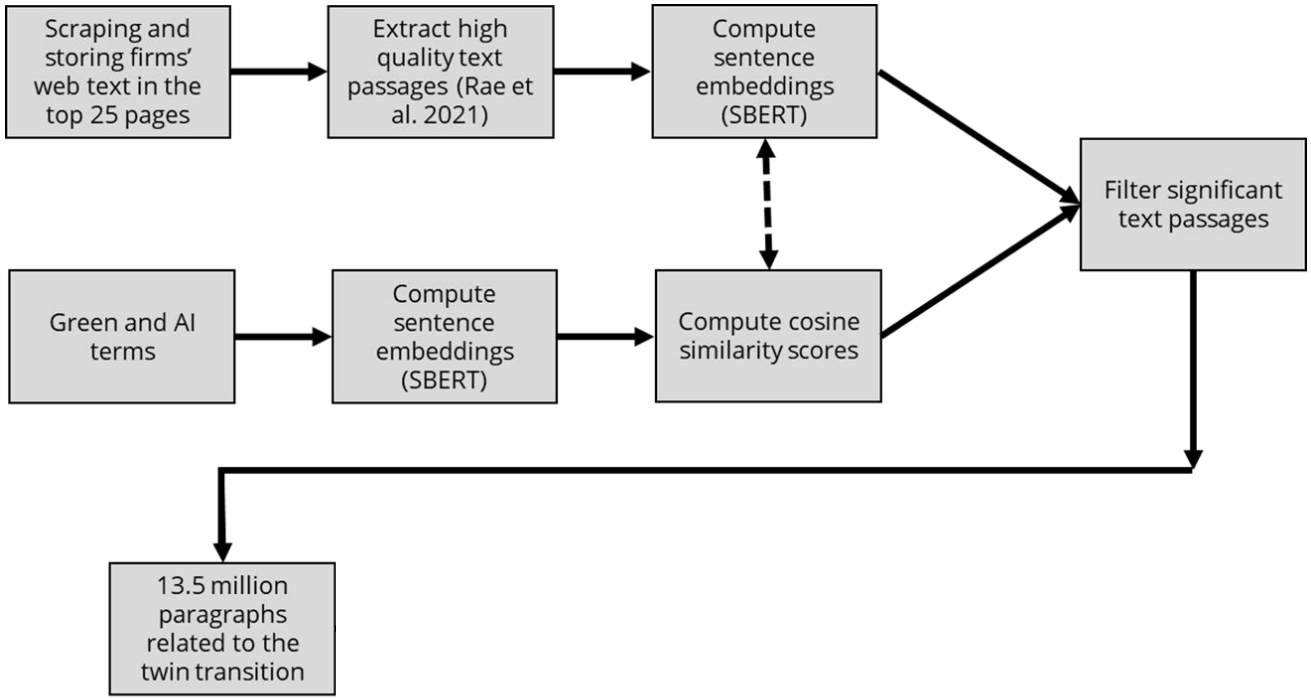
Text-based pretrained model (using SetFit for finetuning)

We examine the quality of photovoltaic systems without dismantling.
Are the modules ok? Were they damaged during assembly? Are the yields too low?
Aerial PV inspection identifies and locates faults precisely.

We combine and automate classic and preventive inspections of solar power plants with image-based methods like thermography, electroluminescence and UV-fluorescence measurements, and with IV-curve recording by day and night, isolation issues detection, serial number scan, and geography data referencing.

To this end we employ tripod- and drone-based sensor systems, providing a cost-effective check-up by systematic sampling or 100% screening. EL measurements offer a comprehensive and exact assessment of faults on cell level and thus represent the tower of strength of the AePVI concept.

Sentence Transformers Finetuning (SetFit)



Sentence Transformers Finetuning (SetFit)

Efficient Few-Shot Learning Without Prompts

Lewis Tunstall[1], **Nils Reimers**[2], **Unso Eun Seo Jo**[1], **Luke Bates**[3], **Daniel Korat**[4], **Moshe Wasserblat**[4], **Oren Pereg**[4]
[1]Hugging Face [2]cohere.ai
[3]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
[4]Emergent AI Lab, Intel Labs
[1]firstname@huggingface.com [2]info@nils-reimers.de
[3]bates@ukp.informatik.tu-darmstadt.de
[4]firstname.lastname@intel.com

Abstract

Recent few-shot methods, such as parameter-efficient fine-tuning (PEFT) and pattern exploiting training (PET), have achieved impressive results in label-scarce settings. However, they are difficult to employ since they are subject to high variability from manually crafted prompts, and typically require billion-parameter language models to achieve high accuracy. To address these shortcomings, we propose SetFit (Sentence Transformer Fine-tuning), an efficient and prompt-free framework for few-shot fine-tuning of Sentence Transformers (ST). SetFit works by first fine-tuning a pretrained ST on a small number of text pairs, in a contrastive Siamese manner. The resulting model is then used to generate rich text embeddings, which are used to train a classification head. This simple framework requires no prompts or verbalizers, and achieves high accuracy with orders of magnitude less parameters than existing techniques. Our experiments show that SetFit obtains comparable results with PEFT and PET techniques, while being an order of magnitude faster to train. We also show that SetFit can be applied in multilingual settings by simply switching the ST body. Our code[1] and datasets[2] are made publicly available.

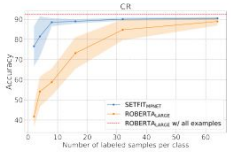


Figure 1: Compared to standard fine-tuning, SetFit is more sample efficient and exhibits less variability when trained on a small number of labeled examples.

1 Introduction

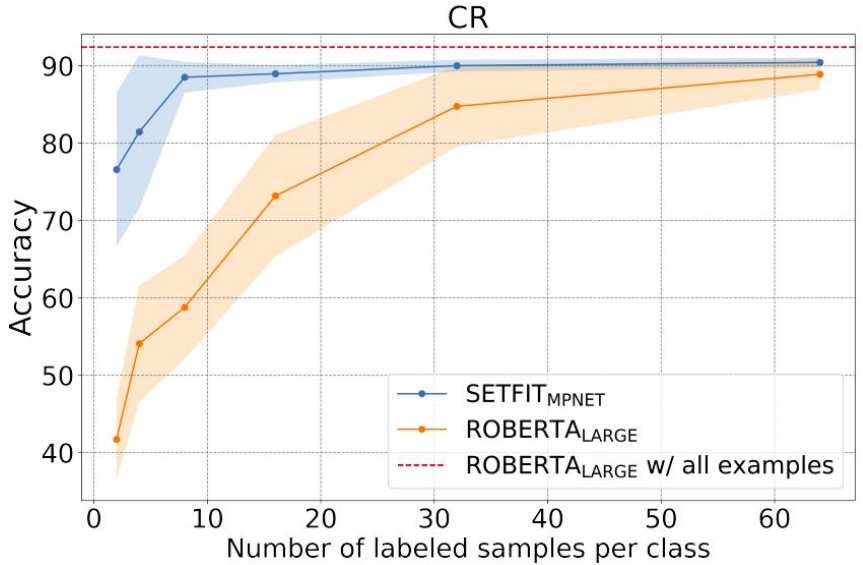
Few-shot learning methods have emerged as an attractive solution to label-scarce scenarios, where data annotation can be time-consuming and costly. These methods are designed to work with a small number of labeled training examples, and typically involve adapting pretrained language models (PLMs) for specific downstream tasks.

Today, there exist several approaches to few-shot learning with PLMs. These include in-context learning (ICL), parameter-efficient fine-tuning (PEFT), and pattern exploiting training (PET). Unfortunately, these approaches can be impractical for many researchers and practitioners. One disadvantage is that these approaches typically rely on the use of large-scale language models to achieve high performance. For example, T-Few (Liu et al., 2022) is based on the 11 billion parameter model T0 (Sanh et al., 2021), while GPT-3 (Brown et al., 2020a) is an order of magnitude larger. Secondly, training and deploying these few-shot methods typically requires specialized infrastructure with limited accessibility. Moreover, PET and the prominent PEFT methods require, as part of their training, the input of manually generated prompts, yielding varying outcomes depending on the level of manual prompt-engineering.

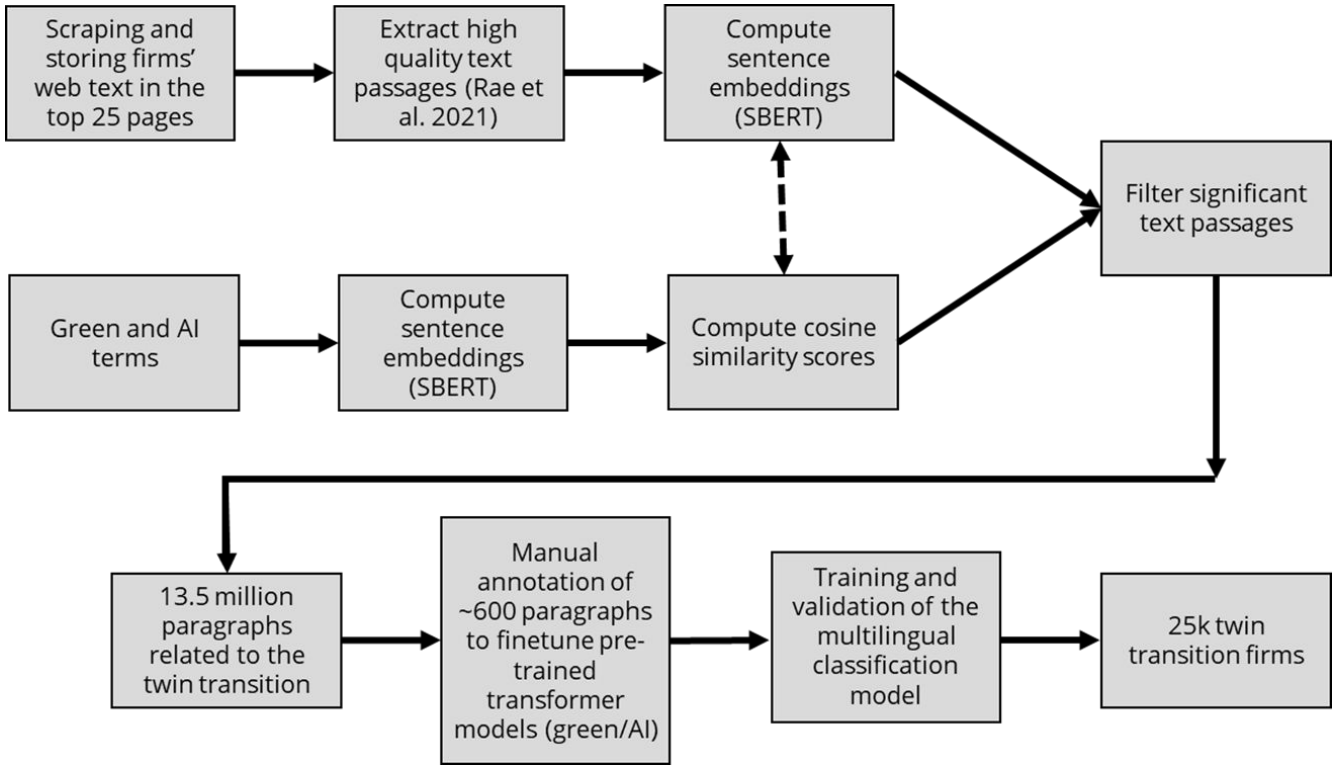
In this paper, we propose SetFit, an approach based on Sentence Transformers (ST) (Reimers and Gurevych, 2019) that dispenses with prompts altogether and does not require large-scale PLMs to achieve high accuracy. For example, with only 8 labeled examples in the Customer Reviews (CR) sentiment dataset, SetFit is competitive with fine-tuning on the full training set, despite the fine-tuned model being three times larger (see Figure 1).

We demonstrate SetFit's efficacy in few-shot text classification over a range of NLP datasets

[1]https://github.com/huggingface/setfit
[2]https://huggingface.co/setfit



Market2Vec (or MarketBert)



Sentence Transformers Finetuning (SetFit)

Text-based pretrained model (using SetFit for finetuning)

REGIONAL STUDIES, REGIONAL SCIENCE
2025, VOL. 12, NO. 1, 513–516
https://doi.org/10.1080/21681376.2025.2510679

Routledge Taylor & Francis Group | RSA Regional Studies Association

REGIONAL GRAPHIC

OPEN ACCESS Check for updates

The geography of digital and green (twin) firms in Germany

Lukas Kriesch [a], Milad Abbasiharofteh [b] and Sebastian Losacker [a,c]

ABSTRACT

The twin transition, which combines green and digital innovation in economic activities, is increasingly central to policy agendas and is also receiving growing attention in regional research. However, accurately mapping green, digital and twin (both green and digital) economic activities across regions remains challenging, particularly due to data constraints. In this study, we advance this research frontier and present a geographic analysis of digital, green and twin economic activities in Germany, using a web-mined dataset of website texts from 678,381 firms, collected through web scraping in 2023. By processing over 44 million text paragraphs from these websites and applying a cosine similarity filter with green and AI-related terms, we filtered firms that are likely engaged in green, digital and twin activities. Based on this subset, 1437 text paragraphs were manually annotated to fine-tune two transformer models within a SetFit framework, accurately classifying firms as green, digital or both. We aggregate this firm-level data into hexagonal cells to reveal the geographic concentration of the twin transition in Germany. The final map shows a higher number of firms involved in green activities, widely spread across Germany, while AI activities are concentrated in urban centres. We identify 23,819 firms engaged in both green and digital activities, with major hubs like Berlin and Munich leading, and peripheral regions potentially being left behind. Our findings offer critical insights into the geography of the twin transition and highlight the need for policies that address potentially induced spatial inequalities.

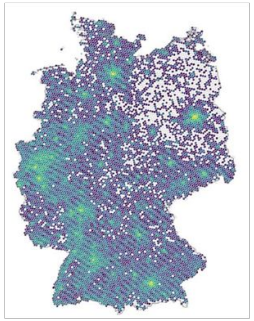
ARTICLE HISTORY
Received 15 November 2024; Accepted 18 May 2025

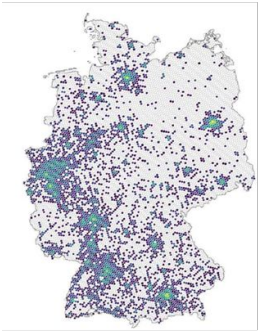
KEYWORDS
Twin transition; web-mining; green transition; digital transition; natural language processing; Germany

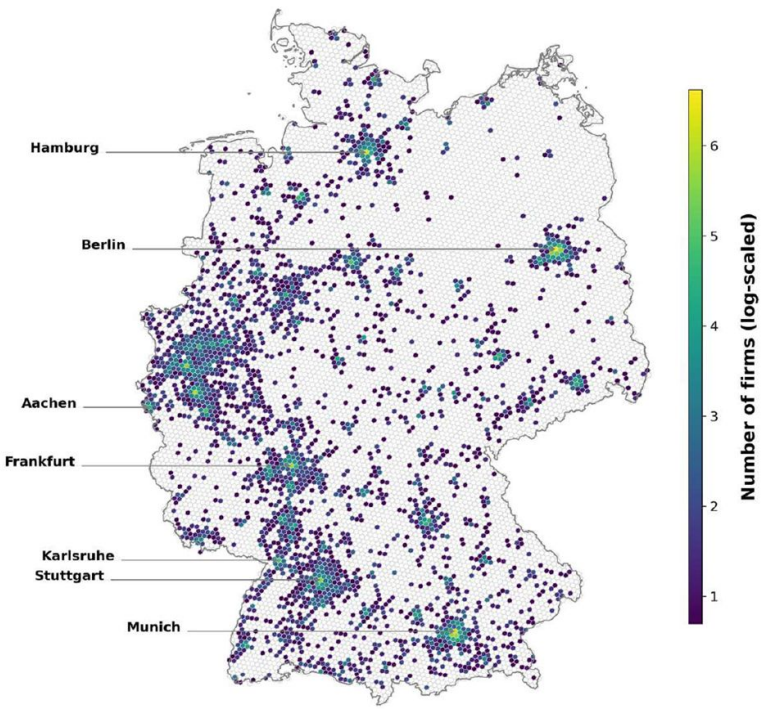
JEL
O31; O33; Q55; R11

The twin transition, which refers to the combined shift of economies toward greener and more digital modes of production, is highly prioritised on the political agenda in Europe

CONTACT Sebastian Losacker sebastian.losacker@geogr.uni-giessen.de
[a] Department of Geography, University of Giessen, Germany
[b] Department of Economic Geography, Faculty of Spatial Sciences, University of Groningen, the Netherlands
[c] CIRCLE – Centre for Innovation Research, Lund University, Sweden
Supplemental data for this article can be accessed online at https://doi.org/10.1080/21681376.2025.2510679

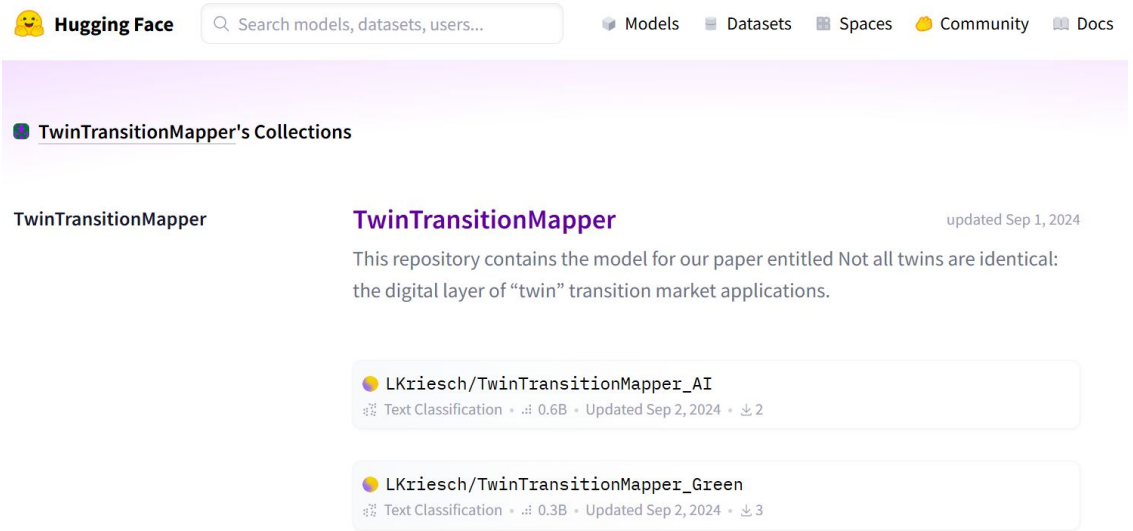






Sentence Transformers Finetuning (SetFit)

Text-based pretrained model (using SetFit for finetuning)



Going beyond Market2Vec

Going beyond Market2Vec as a prediction tool

- What does the bias in a transformer model tell us about firms' next potential products?
- Can we interpret lower levels of predictability in our model, in some cases, as meaningful signals rather than mere noise?

Going beyond Market2Vec

Applying the same approach in a broader context:

- Emerging skills in the job market
- Detecting scientific breakthrough
- Atypical inventions firm’s future market value!
- Investigating team compositions
- etc.