Essays on the Application of

Statistical Learning

in

Empirical Economic Research

Julian Dörr[1,2]

[1]Justus Liebig University Giessen
[2]former: ZEW - Leibniz Centre for European Economic Research

November 16, 2022

Outline of talk

Change in the nature of data

*The phrase **'big data'** emphasizes a change in the scale of data. But there has been an equally important **change in the nature of** this **data**. Machine learning can deal with unconventional data that is too **high-dimensional** for standard estimation methods, including **image and language information** that we conventionally **had not even thought of as data we can work with, let alone include in a regression**.*

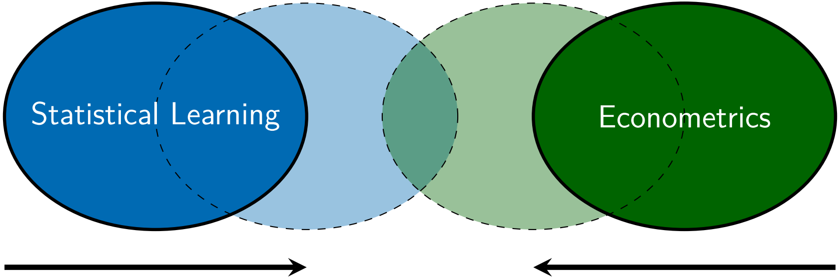
Mullainathan et al. (2017)

Calls to extend the econometric toolkit

Calls to extend the econometric toolkit

Varian (2014), Mullainathan et al. (2017), Athey et al. (2019), Gentzkow et al. (2019), . . .

Convergence of methodological fields



Statistical Learning in Economics

Statistical learning thrives in modeling $f(\mathbf{x})$:

$$\hat{f}(\mathbf{x}) = \begin{cases} \hat{y} & \text{given pairs of } (y, \mathbf{x}) & \textbf{supervised learning} \\ cluster_j & \text{given } \mathbf{x} & \textbf{unsupervised learning} \end{cases}$$

Economists are typically interested in **parameter estimation** and put more structure on their models:

$$f(\mathbf{x}) = \beta\mathbf{x} + \epsilon \longrightarrow \hat{f}(\mathbf{x}) \longrightarrow \hat{\beta}$$

My research

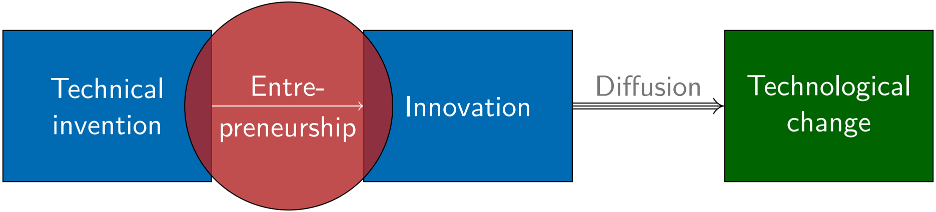
Obtain $\hat{y}$ or $cluster_j$ from typically unstructured, high-dimensional $\mathbf{x}$ to open new perspectives on economic research questions of the kind $y = \beta\mathbf{x} + \epsilon$.

Mapping Technologies to Business Models: An Application to Clean Technologies and Entrepreneurship

published as part of the *the 26th International Conference on Science, Technology and Innovation Indicators (STI2022)* Conference Proceedings

Technological innovation and its measurement

Technological innovation and its measurement



Technological innovation and its measurement

***Patents** have become a surrogate for measuring the innovation process.*

Jaffe (2021)

Technological innovation and its measurement

Patents *have become a surrogate for measuring the innovation process.*

Jaffe (2021)

Patent *subclasses provide a [. . .] reliable picture of a firm's technological capabilities.*

Aharonson et al. (2016)

Patents and start-ups

A measurement problem

Start-ups barely file patents (Mann, 2005; Graham et al., 2008; Graham et al., 2009; Helmers et al., 2011):

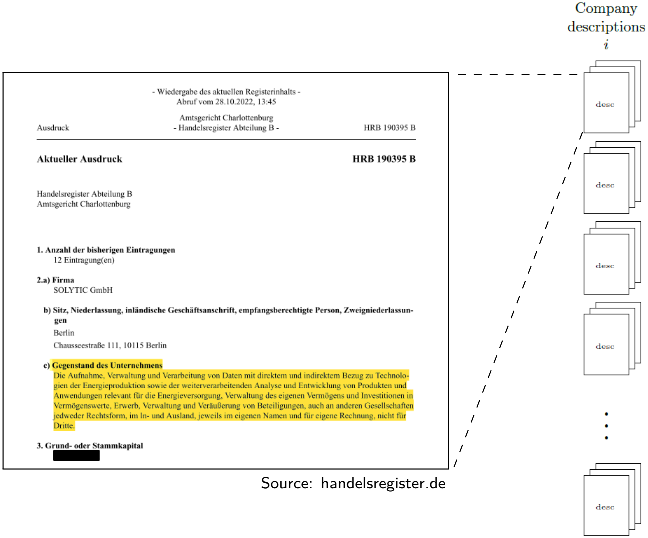
- distracting engineers/managers from key functions
- costs of patenting/patent litigation too high
- disclosure through patent allows ‘design around’

Research question I

How to capture the role of start-ups in the technological innovation process?

Textual innovation data

New ventures legally obliged to publish business purpose at business registration



- Wiedergabe des aktuellen Registerinhalts -
Abruf vom 28.10.2022, 13:45

Ausdruck | Amtsgericht Charlottenburg - Handelsregister Abteilung B - | HRB 190395 B

Aktueller Ausdruck **HRB 190395 B**

Handelsregister Abteilung B
Amtsgericht Charlottenburg

1. Anzahl der bisherigen Eintragungen
12 Eintragung(en)

2.a) Firma
SOLYTIC GmbH

b) Sitz, Niederlassung, inländische Geschäftsanschrift, empfangsberechtigte Person, Zweigniederlassungen
Berlin
Chausseestraße 111, 10115 Berlin

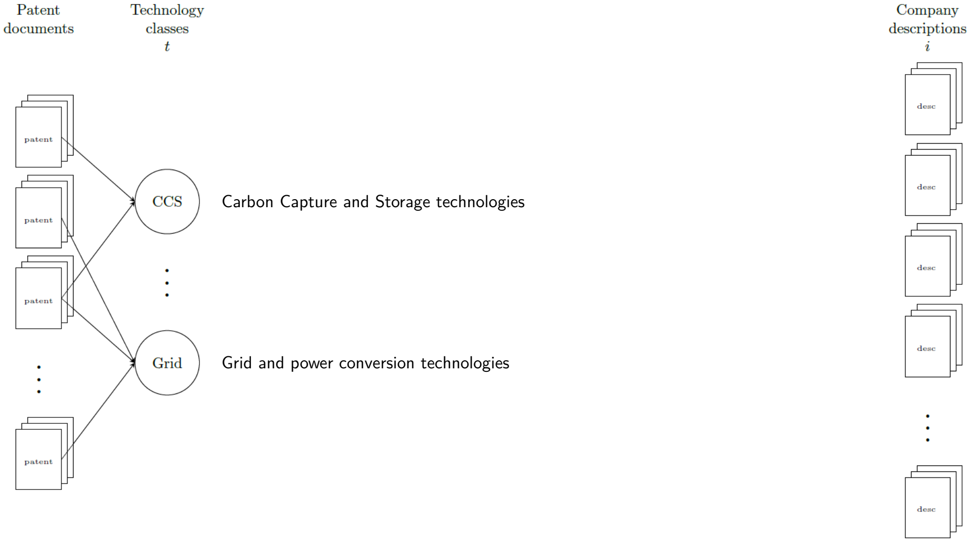
c) Gegenstand des Unternehmens
Die Aufnahme, Verwaltung und Verarbeitung von Daten mit direktem und indirektem Bezug zu Technologien der Energieproduktion sowie der weiterverarbeitenden Analyse und Entwicklung von Produkten und Anwendungen relevant für die Energieversorgung, Verwaltung des eigenen Vermögens und Investitionen in Vermögenswerte, Erwerb, Verwaltung und Veräußerung von Beteiligungen, auch an anderen Gesellschaften jedweder Rechtsform, im In- und Ausland, jeweils im eigenen Namen und für eigene Rechnung, nicht für Dritte.

3. Grund- oder Stammkapital
[redacted]

Source: handelsregister.de

Textual innovation data

Patent texts and assigned technology classes

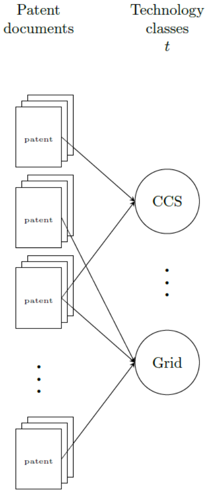


Carbon Capture and Storage technologies

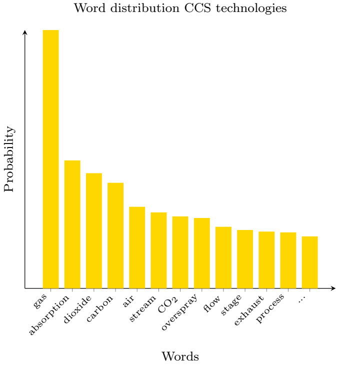
Grid and power conversion technologies

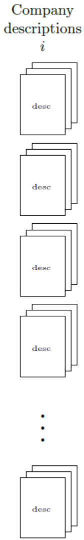
From patents to technology descriptions

L-LDA (Ramage et al., 2009)



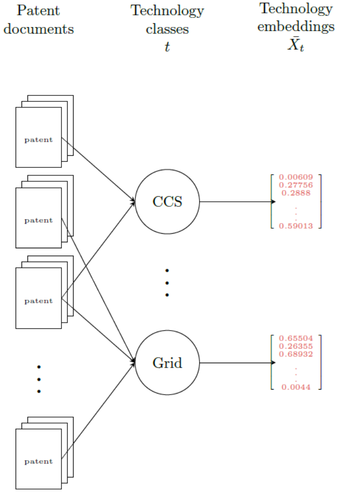
Goal: Derive technology-word distributions from expert-labeled corpus of patent docs





Contextualized vector representations

BERT (Devlin et al., 2018)

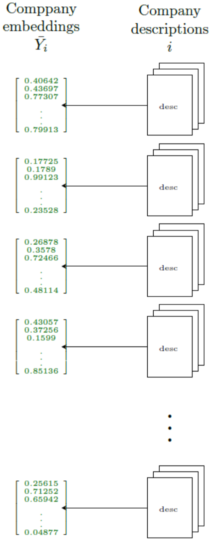


<u>Goal:</u> Derive contextualized vector representations of technology & business descriptions

$$\underset{(1\times Q)}{X_{CCS}} = \langle \text{gas, absorption, dioxide}, \ldots, \text{scrub}, \ldots \rangle$$

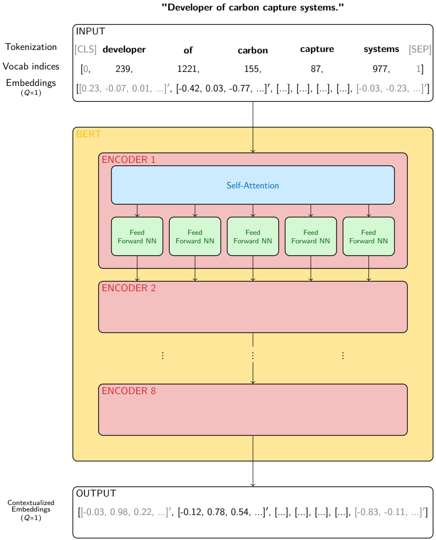
$$\Big\downarrow \text{SBERT}$$

$$\underset{(1\times 384\ \forall Q)}{X_{CCS}} = [0.006, 0.277, 0.288, \ldots 0.590]'$$



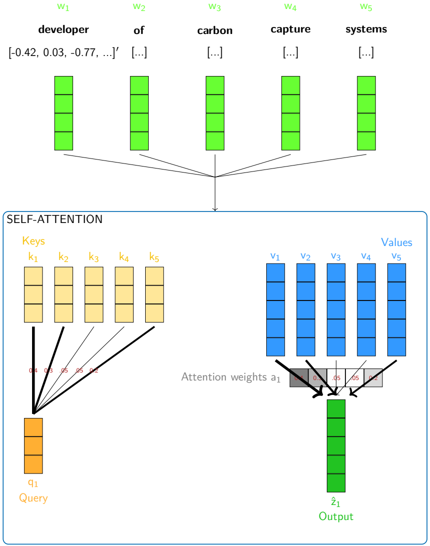
Excursus: BERT

Model architecture



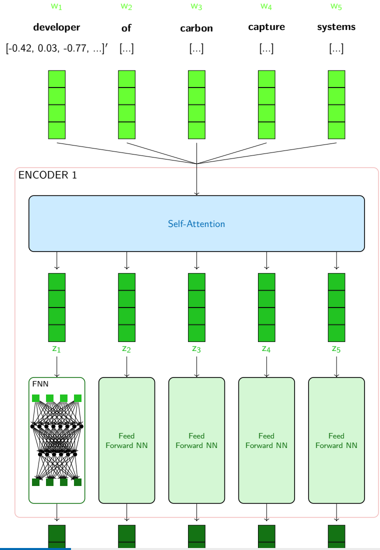
Attention Is All You Need (Vaswani et al., 2017)

Let tokens 'look around' the whole input, and decide how to update its representation based on on what it sees



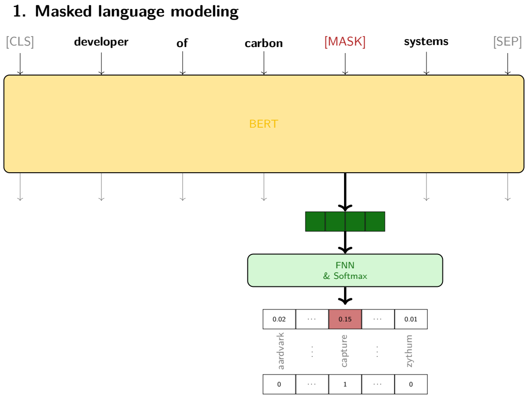
Encoder

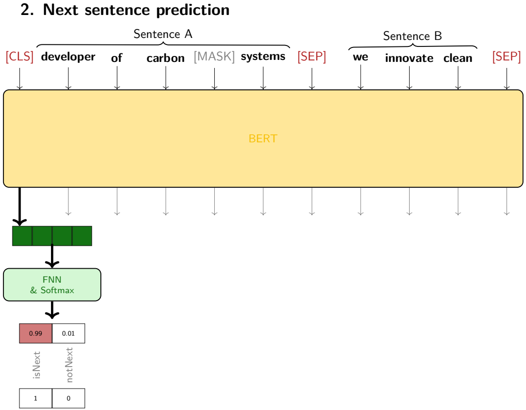
After Attention, each token pondering for itself about what it has observed previously



Training BERT

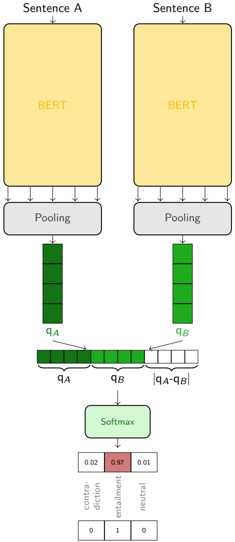
Self-supervised learning based on English Wikipedia





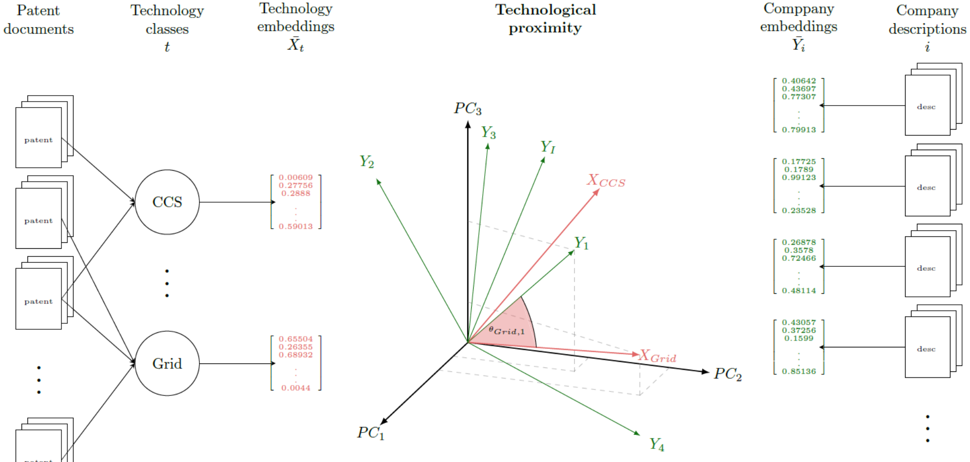
Finetuning BERT: SBERT (Reimers et al., 2019)

Finetuning based on collection of sentence pairs labeled for entailment, contradiction, and semantic independence



Mapping framework

Cosine similarity as measure of a company's technological capability



$$\textsc{TechProx}_{t,i} := max\left(0, \frac{\bar{X}_t \bar{Y}_i}{||\bar{X}_t|| ||\bar{Y}_i||}\right)$$

Application

Role of start-ups in clean technology diffusion

Adaptation to and mitigation of climate change requires new technological pathways and radical innovations (e.g. United Nations (2015), European Commission (2019))

- but: technological path dependencies and system/innovation inertia among incumbents (Patel et al., 1997; Aghion et al., 2016)
- costly: delay in redirecting innovation towards clean technologies (Benner, 2009; Dijk et al., 2016; Sick et al., 2016)

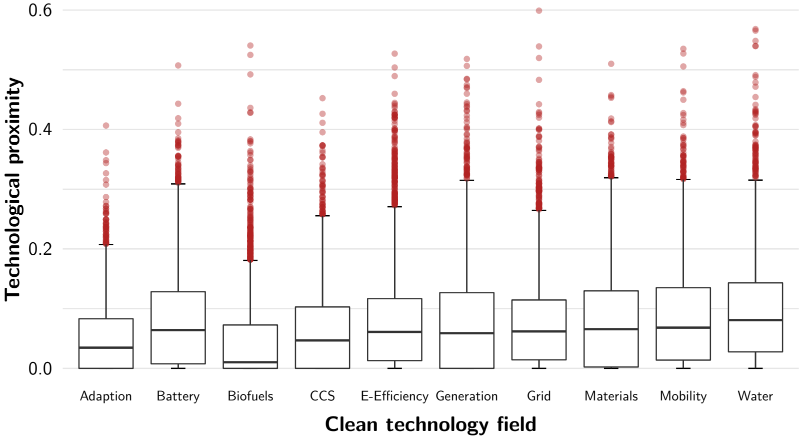
$\Rightarrow$ special role of new (path-independent!) ventures in triggering clean technology change (Cohen et al., 2007; Hockerts et al., 2010; Horne et al., 2022)

Research question II

Which role do start-ups play in the diffusion of clean technologies?

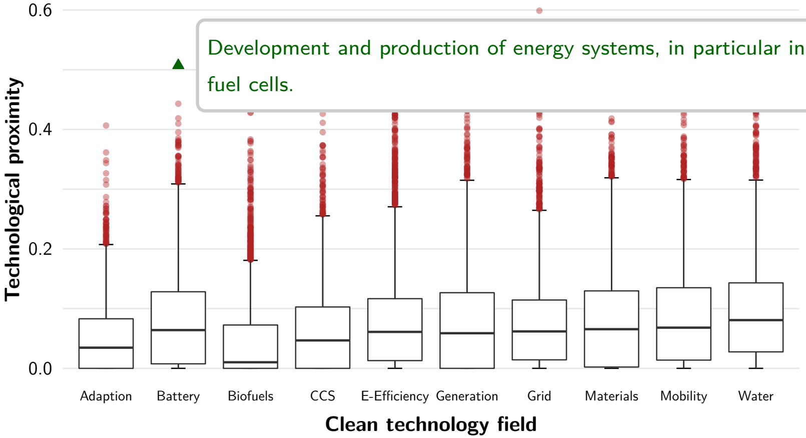
Application

TechProx in survey of German start-ups



Application

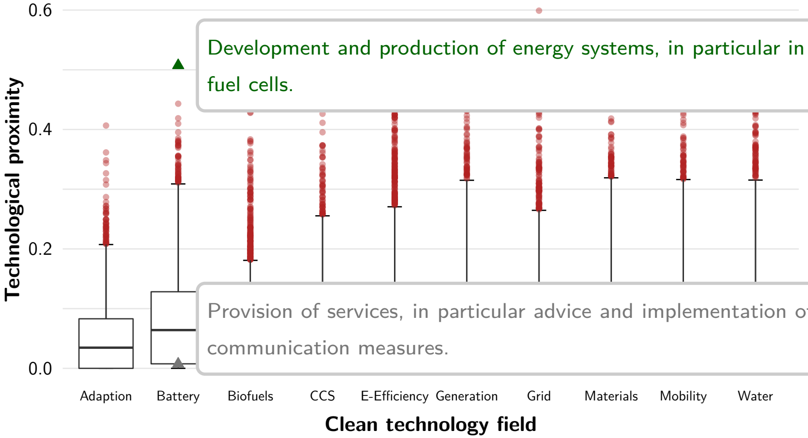
A glance at the 'outliers'



Development and production of energy systems, in particular in connection with fuel cells.

Application

A glance at the 'outliers'

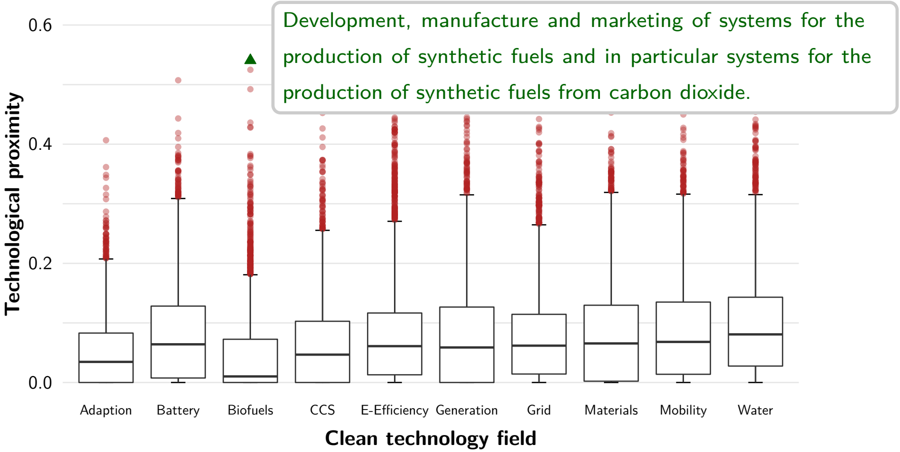


Development and production of energy systems, in particular in connection with fuel cells.

Provision of services, in particular advice and implementation of marketing and communication measures.

Application

A glance at the 'outliers'



Characteristics of clean technology start-ups

Cleantech start-ups show a higher propensity to eco-innovate

	EInno					
	(1)	(2)	(3)	(4)	(5)	(6)
TechProx (0-1)	1.339***	1.328***	1.325***	1.295***	1.288***	1.383***
log(size)		1.191***	1.154***	1.129***	1.191***	1.187***
subsidy			1.304***	1.352***	1.411***	1.445***
R&D			1.334***	1.411***	1.574***	1.595***
returns				1.773***	1.665**	1.616**
break even				1.299***	1.232**	1.257**
team size					0.901**	0.891**
university					0.612***	0.627***
Sector controls	Y	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	N	Y
N	3,269	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.021	0.025	0.029	0.033	0.040	0.047

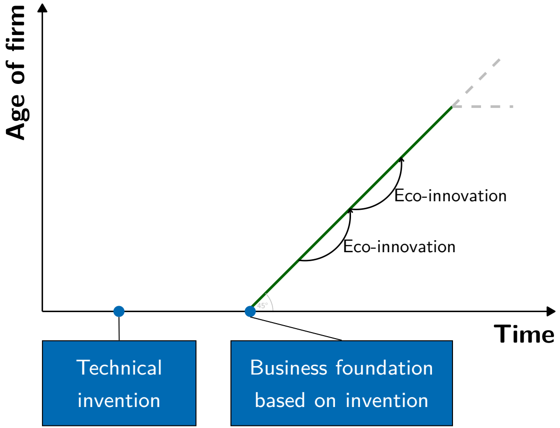
EInno := Introduction of environmental innovation?

- no environmental innovation
- environmental innovation with moderate environmental effect
- environmental innovation with substantial environmental effect

Coefficient estimates reported as proportional odds ratios.

Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Entrepreneurial process and innovation



Summary

- Latest evolutions in the field of NLP allow fine granular determination of a firm's technological profile
- Legal obligation to publish a business purpose makes the technology mapping possible for start-ups even w/o traditional innovation data

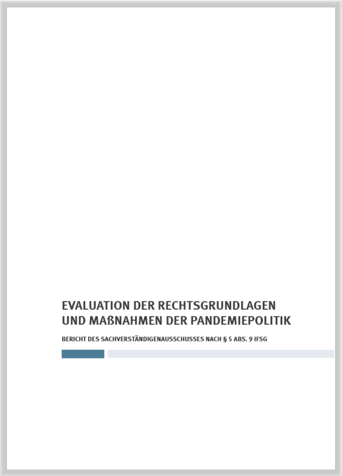
- Leveraging the introduced technology mapping to the field of clean technologies suggests:
  - a high propensity of cleantech start-ups to introduce eco-innovations
  - supporting their special role in the transition to a green economy derived from theory
  - both by virtue of their business models as well as a high propensity to adopt additional environmental innovations

Small firms and the COVID-19 insolvency gap

joint with: Georg Licht and Simona Murmann

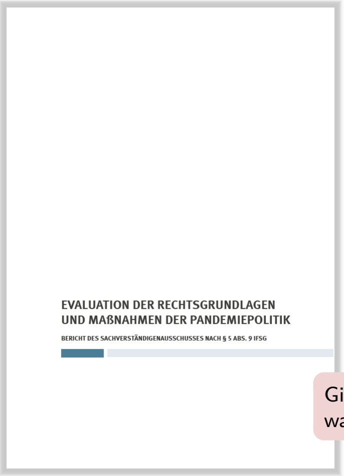
published in: Small Business Economics

Motivation - pragmatic and timely policy evaluation



Source: Sachverständigenausschuss (2022)

Motivation - pragmatic and timely policy evaluation


EVALUATION DER RECHTSGRUNDLAGEN
UND MAßNAHMEN DER PANDEMIEPOLITIK

BERICHT DES SACHVERSTÄNDIGENAUSSCHUSSES NACH § 5 ABS. 9 IFSG

Source: Sachverständigenausschuss (2022)

Bei der Evaluierung von Maßnahmen und Maßnahmenpaketen geht es darum, die richtigen Fragen nach deren Wirkung zu stellen und ein ebenso sorgfältiges wie angesichts der meist lückenhaften Datenlage ***pragmatisches Studiendesign*** *zu wählen, das es erlaubt, diese Fragen zumindest näherungsweise zu beantworten.*

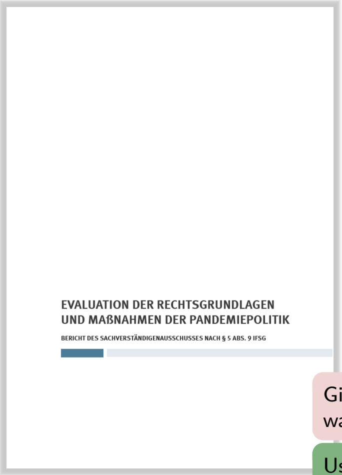
Allerdings muss über politisches Handeln und dessen ***Nachsteuerung in Echtzeit*** *entschieden werden. [...] Daher können bereits* ***indikative Aussagen*** *zu Teilen des [Maßnhahmen-]Bündels, die dem Prinzip genügen, das Vergleichbare zu vergleichen,* ***von erheblichem Wert*** *sein.*

Im Fall der Corona-Pandemie müssen sie vor allem mit dem Problem umgehen, dass ***Vorher-Nachher- oder Differenz-in-Differenzen-Ansätze*** *völlig* ***von der hohen Infektionsdynamik überlagert*** *sein können.*

Sachverständigenausschuss (2022)

Given the **dynamics of the pandemic** and the **bundle of policy measures** to prevent a wave of corporate insolvencies, can we still evaluate the policy measures' effectiveness?

Motivation - pragmatic and timely policy evaluation


EVALUATION DER RECHTSGRUNDLAGEN
UND MAßNAHMEN DER PANDEMIEPOLITIK

BERICHT DES SACHVERSTÄNDIGENAUSSCHUSSES NACH § 5 ABS. 9 IFSG

Source: Sachverständigenausschuss (2022)

Bei der Evaluierung von Maßnahmen und Maßnahmenpaketen geht es darum, die richtigen Fragen nach deren Wirkung zu stellen und ein ebenso sorgfältiges wie angesichts der meist lückenhaften Datenlage ***pragmatisches Studiendesign*** *zu wählen, das es erlaubt, diese Fragen zumindest näherungsweise zu beantworten.*

Allerdings muss über politisches Handeln und dessen ***Nachsteuerung in Echtzeit*** *entschieden werden. [...] Daher können bereits* ***indikative Aussagen*** *zu Teilen des [Maßnhahmen-]Bündels, die dem Prinzip genügen, das Vergleichbare zu vergleichen,* ***von erheblichem Wert*** *sein.*

Im Fall der Corona-Pandemie müssen sie vor allem mit dem Problem umgehen, dass ***Vorher-Nachher- oder Differenz-in-Differenzen-Ansätze*** *völlig* ***von der hohen Infektionsdynamik überlagert*** *sein können.*

Sachverständigenausschuss (2022)

Given the **dynamics of the pandemic** and the **bundle of policy measures** to prevent a wave of corporate insolvencies, can we still evaluate the policy measures' effectiveness?

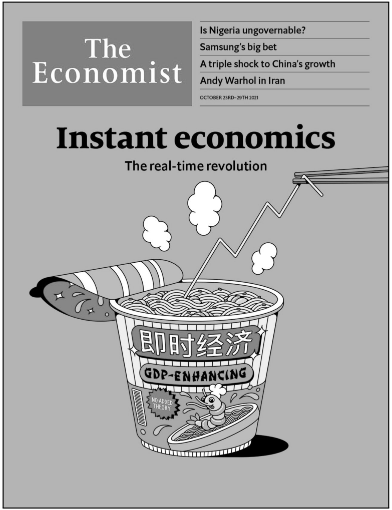
Using kNN as supervised learning algorithm to find for each rating update observed after the COVID outbreak the k nearest control units from the pre-COVID period and compare their insolvency states.

An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers

joint with Jan Kinne, David Lenz, Georg Licht, Peter Winker

published in: PLoS ONE

Motivation - lack of real-time economic data



Source: The Economist (2021a)

Motivation - lack of real-time economic data

Source: The Economist (2021a)

Does anyone really understand what is going on in the world economy? The pandemic has made plenty of observers look clueless.

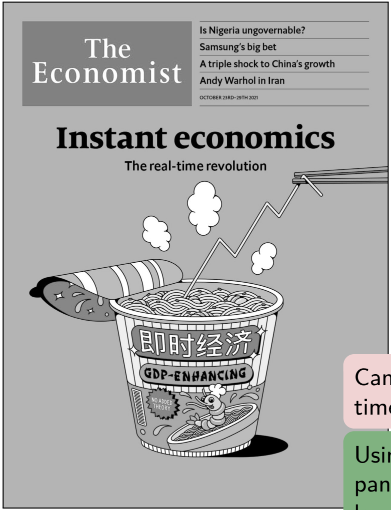
Especially in times of rapid change, policymakers have operated in a fog.

The gap between official data and what is happening in the real economy can still be glaring.

The Economist (2021a, 2021b)

Can we assist policy makers with **timely** and **insightful** firm level data in times of dynamic economic shocks such as COVID-19?

Motivation - lack of real-time economic data



Source: The Economist (2021a)

Does anyone really understand what is going on in the world economy? The pandemic has made plenty of observers look clueless.

Especially in times of rapid change, policymakers have operated in a fog.

The gap between official data and what is happening in the real economy can still be glaring.

The Economist (2021a, 2021b)

Can we assist policy makers with **timely** and **insightful** firm level data in times of dynamic economic shocks such as COVID-19?

Using firm communication patterns from corporate websites about the pandemic's effects on their business and classify these with a fine-tuned language model to obtain leading indicators at near real-time.

Contributions of this thesis

Statistical Learning in Empirical Economics as

1. Fine-granular technology indicator
2. Policy evaluation tool
3. Early indicator development during economic shocks

Acs, Z. J., & Audretsch, D. B. (2005). Entrepreneurship, Innovation and Technological Change. Foundations and Trends® in Entrepreneurship, 1(4), 149–195. https://doi.org/10.1561/0300000004

Aghion, P., Dechezleprêtre, A., Hémous, D., Martin, R., & van Reenen, J. (2016). Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. Journal of Political Economy, 124(1), 1–51. https://doi.org/10.1086/684581

Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. Research Policy, 45(1), 81–96. https://doi.org/10.1016/j.respol.2015.08.001

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I. (2013). Predicting financial distress of companies: revisiting the Z-Score and ZETA® models (A. R. Bell, C. Brooks, & M. Prokopczuk, Eds.). In A. R. Bell, C. Brooks, & M. Prokopczuk (Eds.), Handbook of research methods and applications in empirical finance. Edward Elgar Publishing. https://doi.org/10.4337/9780857936097.00027

Athey, S., & Imbens, G. W. (2019). Machine learning methods economists should know about. Annual Review of Economics, 11, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3, 1137–1155. https://doi.org/10.1080/1536383X.2018.1448388

Benner, M. J. (2009). Dynamic or static capabilities? Process management practices and response to technological change. Journal of Product Innovation Management, 26(5), 473–486. https://doi.org/10.1111/j.1540-5885.2009.00675.x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.1145/2133806.2133826

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5arXiv 1607.04606, 135–146. https://doi.org/10.1162/tacl_a_00051

Caballero, R. J., Hoshi, T., & Kashyap, A. K. (2008). Zombie lending and depressed restructuring in Japan. American Economic Review, 98(5), 1943–1977. https://doi.org/10.1257/aer.98.5.1943

Cohen, B., & Winn, M. I. (2007). Market imperfections, opportunity and sustainable entrepreneurship. Journal of Business Venturing, 22(1), 29–49. https://doi.org/10.1016/j.jbusvent.2004.12.001

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv 1911.02116. https://doi.org/10.48550/arXiv.1911.02116

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dijk, M., Wells, P., & Kemp, R. (2016). Will the momentum of the electric car last? Testing an hypothesis on disruptive innovation. Technological Forecasting and Social Change, 105, 77–88. https://doi.org/10.1016/j.techfore.2016.01.013

European Commission. (2003). Commission recommendation concerning the definition of micro, small and medium-sized enterprises. Official Journal of the European Union, L124, 36–41. http://data.europa.eu/eli/reco/2003/361/oj%7B%5C%%7D0A%7B%5C%%7D0A
European Commission. (2019). The European Green Deal (tech. rep.). University of California Press.
Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. http://annabellelukin.edublogs.org/files/2013/08/Firth-JR-1962-A-Synopsis-of-Linguistic-Theory-wfihi5.pdf
Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574. https://doi.org/10.1257/jel.20181020
Gottschalk, S. (2013). The Research Data Centre of the Centre for European Economic Research (ZEW-FDZ). Schmollers Jahrbuch: Journal of Applied Social Science Studies, 133(4), 607–618. https://doi.org/10.3790/schm.133.4.607
Graham, S. J., Merges, R. P., Samuelson, P., & Sichelman, T. (2009). High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey. Berkeley Technology Law Journal, 24(4), 1255–1327. http://www.jstor.org/stable/24120583
Graham, S. J., & Sichelman, T. (2008). Why Do Start-ups Patent? Berkeley Technology Law Journal, 23(3), 1063–1097. http://www.jstor.org/stable/24118267
Helmers, C., & Rogers, M. (2011). Does patenting help high-tech start-ups? Research Policy, 40(7), 1016–1027. https://doi.org/10.1016/j.respol.2011.05.003
Hockerts, K., & Wüstenhagen, R. (2010). Greening Goliaths versus emerging Davids - Theorizing about the role of incumbents and new entrants in sustainable entrepreneurship. Journal of Business Venturing, 25(5), 481–492. https://doi.org/10.1016/j.jbusvent.2009.07.005
Horne, J., & Fichter, K. (2022). Growing for sustainability: Enablers for the growth of impact startups – A conceptual framework, taxonomy, and systematic literature review. Journal of Cleaner Production, 349(February 2021), 131163. https://doi.org/10.1016/j.jclepro.2022.131163
Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th annual meeting of the association for aomputational linguistics. https://doi.org/10.48550/arXiv.1801.06146
Jaffe, A. B. (2021). Patent Metrics for Innovation Research: Overview, Update and Speculation.
Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of the 15th conference of the European chapter of the association for computational linguistics, 2(Short Papers), 427–431. https://doi.org/10.1176/appi.ps.201500423
Mann, R. J. (2005). Do Patents Facilitate Financing in the Software Industry? Texas Law Review, 83(4), 961–1030.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations ofWords and Phrases and their Compositionality. Advances in neural information processing systems, arXiv 1606.08359, 3111–3119. https://doi.org/10.18653/v1/d16-1146
Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106. https://doi.org/10.1257/jep.31.2.87
Patel, P., & Pavitt, K. (1997). The technological competencies of the world's largest firms: Complex and path-dependent, but not much variety. Research Policy, 26(2), 141–156. https://doi.org/10.1016/S0048-7333(97)00005-X

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation Jeffrey. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543. https://doi.org/10.1080/02688697.2017.1354122

Peters, M. E., Neumann, M., Iyyer, M., & Gardner, M. (2018). Deep contextualized word representations. Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, 1(Long Paper), arXiv 1802.05365, 2227–2237.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training (tech. rep.).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248–256. http://www.aclweb.org/anthology/D09-1026

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language p Stroudsburg, PA, USA, Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. Biometrics, 29(1), 159–183. https://doi.org/10.2307/2529684

Sachverständigenausschuss. (2022). Evaluation der Rechtsgrundlagen und Maßnahmen der Pandemiepolitik.

Schumpeter, J. A. (1942). Capitalsim, Socialism and Democracy. New York: Harper.

Sick, N., Nienaber, A. M., Liesenkötter, B., vom Stein, N., Schewe, G., & Leker, J. (2016). The legend about sailing ship effects – Is it true or false? The example of cleaner propulsion technologies diffusion in the automotive industry. Journal of Cleaner Production, 137, 405–413. https://doi.org/10.1016/j.jclepro.2016.07.085

The Economist. (2021a). A real-time revolution will up-end the practice of macroeconomics. Retrieved October 25, 2021, from https://www.economist.com/leaders/2021/10/23/a-real-time-revolution-will-up-end-the-practice-of-macroeconomics

The Economist. (2021b). Enter third-wave economics. Retrieved October 25, 2021, from https://www.economist.com/briefing/2021/10/23/enter-third-wave-economics

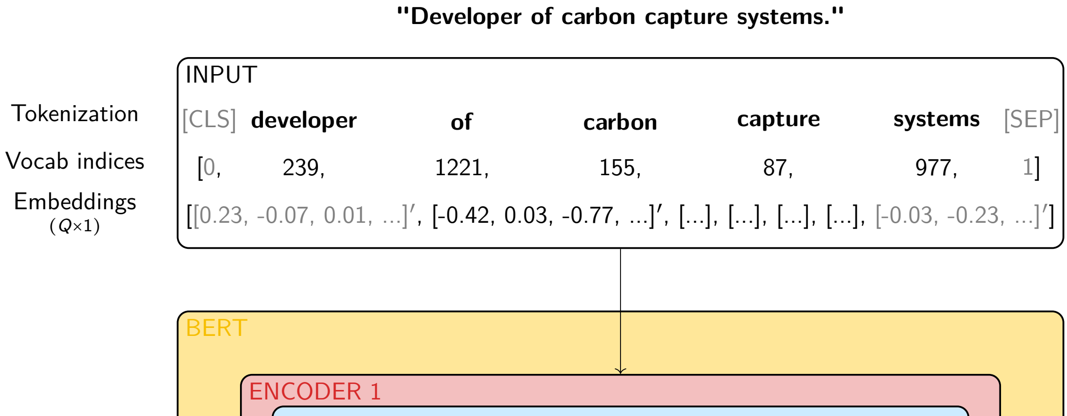
United Nations. (2015). Paris Agreement. Retrieved October 1, 2020, from https://unfccc.int/sites/default/files/english%7B%5C_%7Dparis%7B%5C_%7Dagreement.pdf

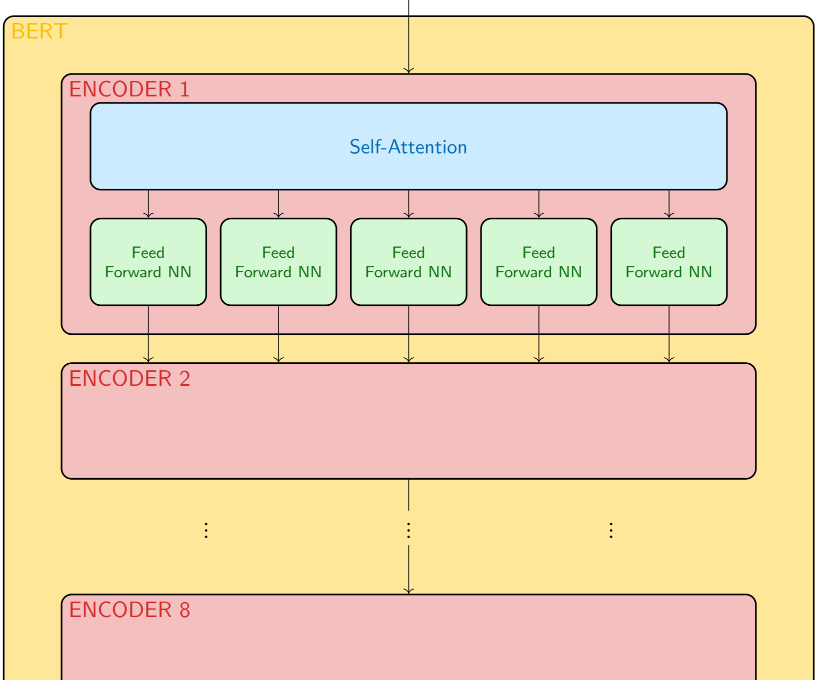
Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3–28. https://doi.org/10.1257/jep.28.2.3

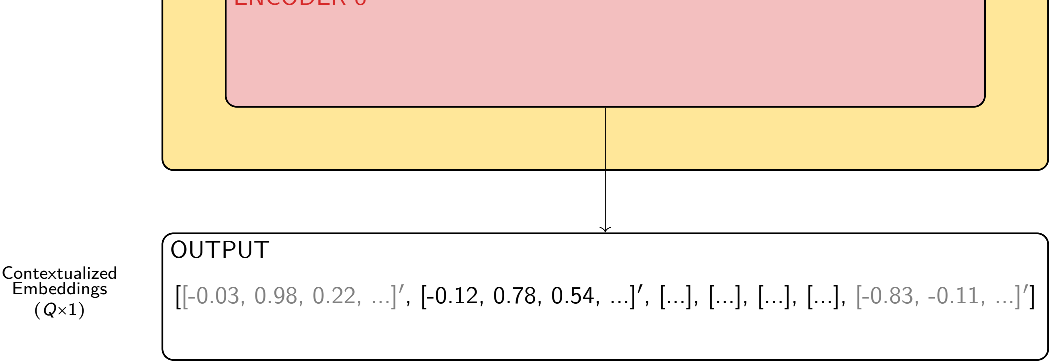
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), arXiv 1706.03762. http://arxiv.org/abs/1706.03762

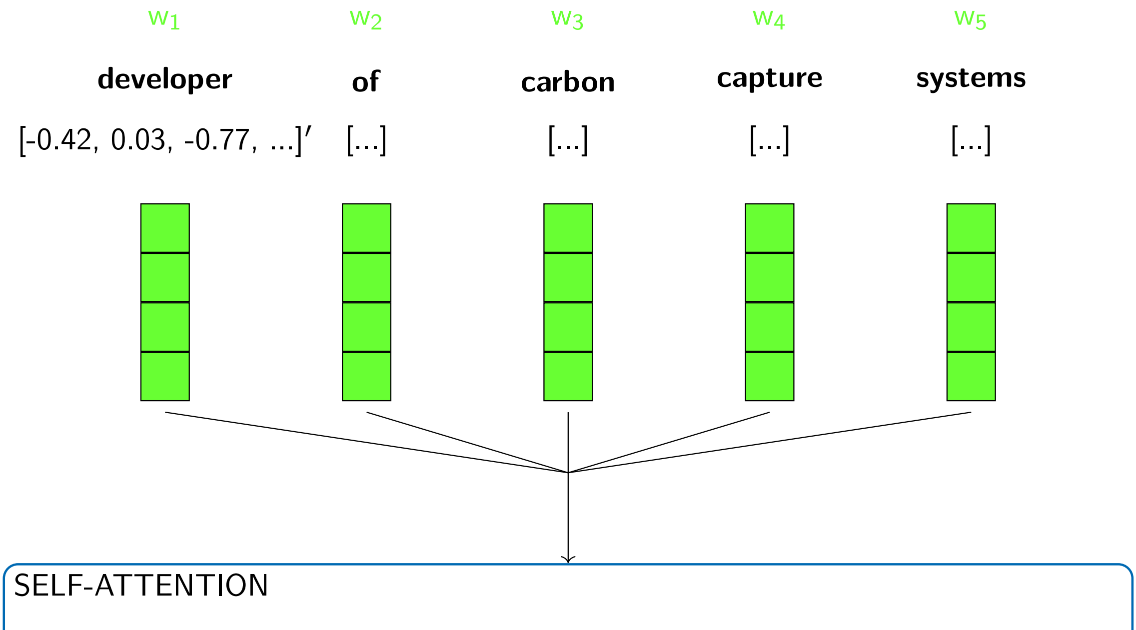
Wang, Y., Hou, Y., Che, W., & Liu, T. (2020). From static to dynamic word representations: a survey. International Journal of Machine Learning and Cybernetics, 11(7), 1611–1630. https://doi.org/10.1007/s13042-020-01069-8

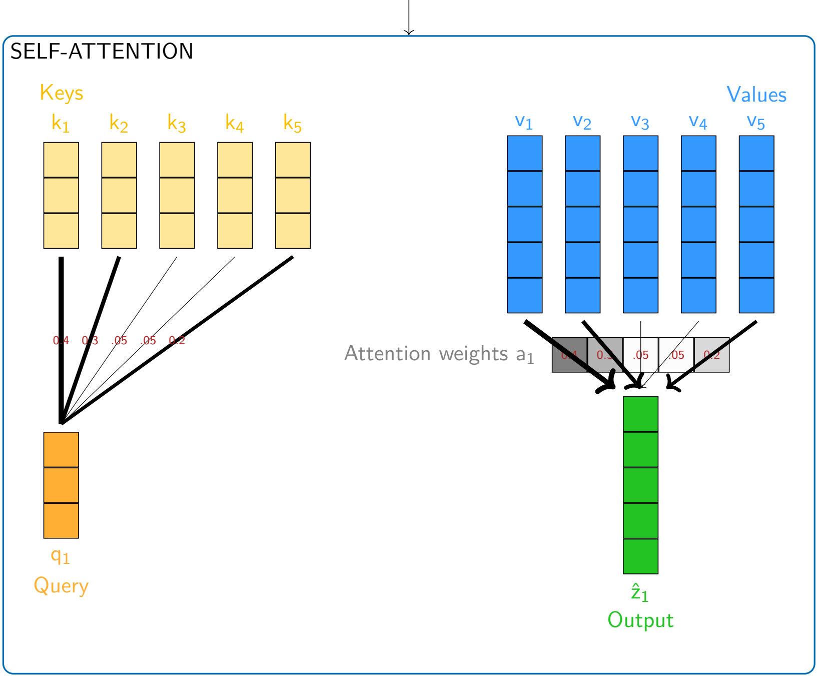
Appendix
Technology-company mapping framework

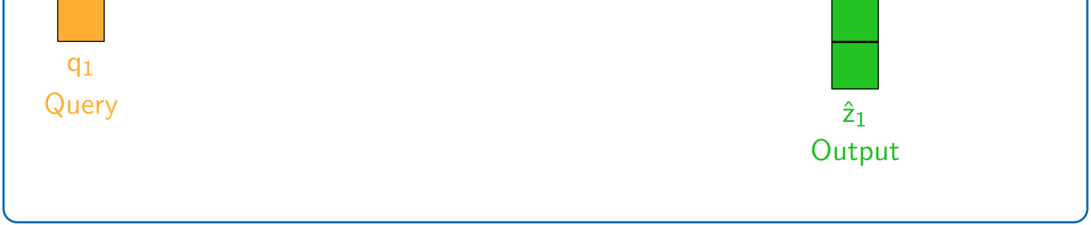

"Developer of carbon capture systems."
INPUT
Tokenization
[CLS] developer of carbon capture systems [SEP]
Vocab indices
[0, 239, 1221, 155, 87, 977, 1]
Embeddings
(Q×1)
[[0.23, -0.07, 0.01, ...]′, [-0.42, 0.03, -0.77, ...]′, [...], [...], [...], [...], [-0.03, -0.23, ...]′]
BERT
ENCODER 1


BERT
ENCODER 1
Self-Attention
Feed Forward NN
Feed Forward NN
Feed Forward NN
Feed Forward NN
Feed Forward NN
ENCODER 2
ENCODER 8


ENCODER 6
OUTPUT
[[-0.03, 0.98, 0.22, ...]′, [-0.12, 0.78, 0.54, ...]′, [...], [...], [...], [...], [-0.83, -0.11, ...]′]
Contextualized Embeddings
(Q×1)


w_1
w_2
w_3
w_4
w_5
developer
of
carbon
capture
systems
[-0.42, 0.03, -0.77, ...]′
[...]
[...]
[...]
[...]
SELF-ATTENTION


SELF-ATTENTION
Keys
k1 k2 k3 k4 k5
0.4 0.3 .05 .05 0.2
q1
Query
Values
v1 v2 v3 v4 v5
Attention weights a1
0.4 0.3 .05 .05 0.2
ẑ1
Output



1. Attention weights $a_{1:5}$ are query-key similarities:

   $\hat{a}_i = \mathbf{q}_i \times \mathbf{k}_i$

   Normalized via softmax: $a_i = e^{\hat{a}_i} / \sum_j e^{\hat{a}_j} \ \in [0, 1]$

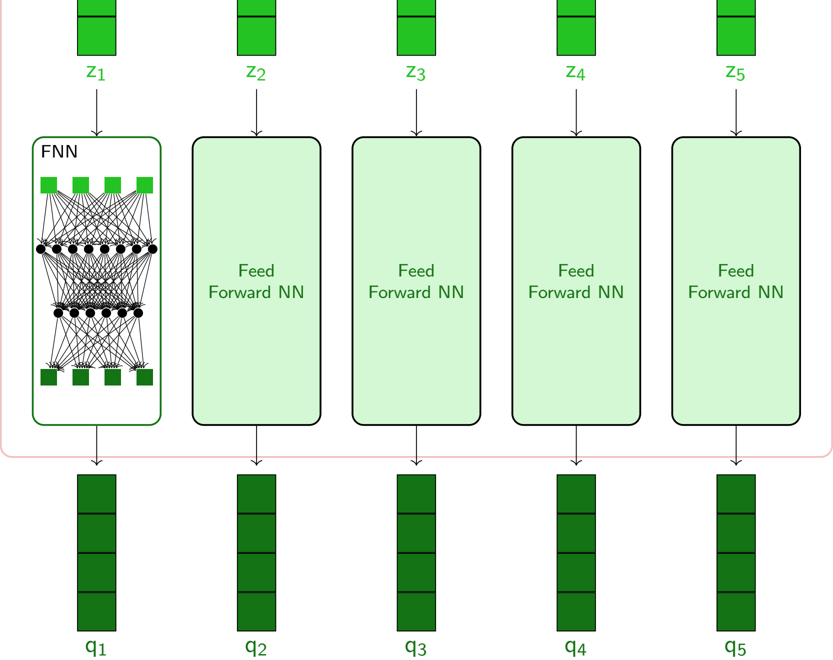
2. Output $\underset{(1 \times Z)}{\widehat{\mathbf{z}}_i}$ is attention-weighted average of value vectors $\mathbf{v}_{1:5}$:

   $\widehat{\mathbf{z}}_i = \sum_i a_i \mathbf{v}_i$

3. $\mathbf{k}$, $\mathbf{v}$ and $\mathbf{q}$ are derived from the entire input $\mathbf{w}$:

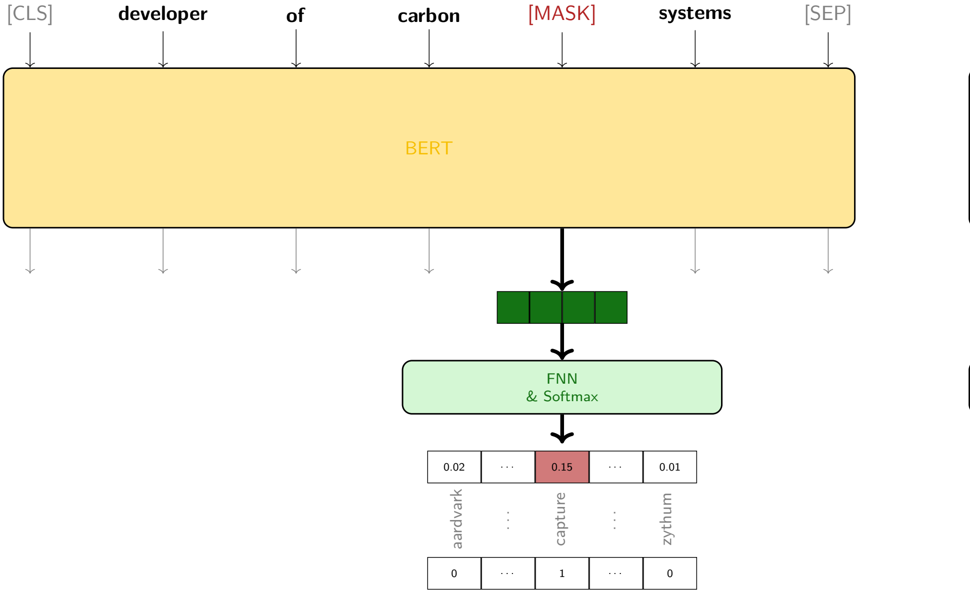
   $\mathbf{k} = W_k \times \mathbf{w} \qquad \mathbf{v} = W_v \times \mathbf{w} \qquad \mathbf{q} = W_q \times \mathbf{w}$

Note: Self-attention is repeated H times (multi-head attention) and the resulting vectors are concatenated along the feature dimension. Multiplying with a weight matrix $\underset{(Q \times HZ)}{W_z}$ yields the final output vector that is passed to the FNN.

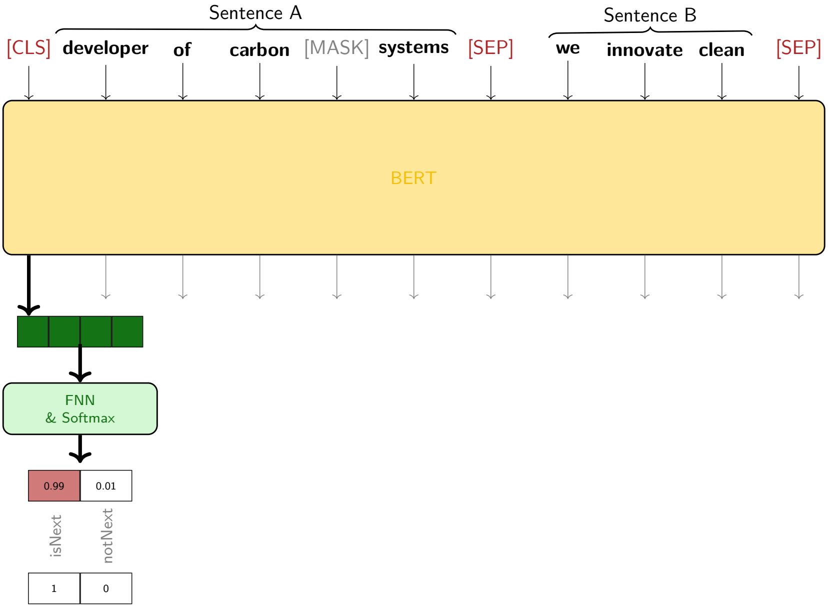


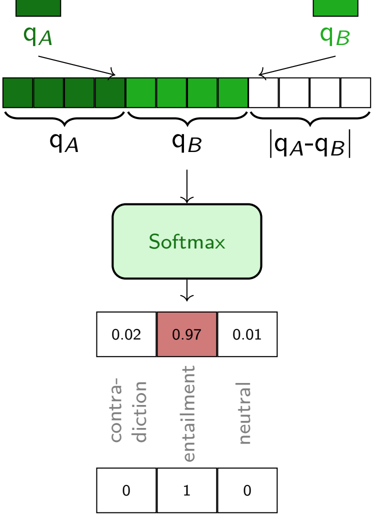
z_1
z_2
z_3
z_4
z_5
FNN
Feed Forward NN
Feed Forward NN
Feed Forward NN
Feed Forward NN
q_1
q_2
q_3
q_4
q_5

1. Masked language modeling



2. Next sentence prediction





qA
qB
qA
qB
|qA-qB|
Softmax
0.02
0.97
0.01
contra-
diction
entailment
neutral
0
1
0

Invention, Innovation & Entrepreneurship

A patent reflects new technical knowledge, but it does not indicate whether this knowledge has a positive economic value. Only those inventions which have been successfully introduced in the market can claim that they are innovations as well. While innovations and inventions are related, they are not identical.

Acs et al. (2005)

Text preprocessing

1. translation of non-English texts to English
2. Part of Speech (PoS) tagging
   2.1 remove punctuation, numbers and unknown tags
   2.2 lemmatization
3. stop word deletion

A labeled corpus of patent abstracts

Patent	Technology class	Abstract
1	B, C, Y02C, Y02P	Catalyst, comprising one or more compounds of the perovskite-type as catalytically active component, is new, where the catalytically active component in the form of at least one layer is applied on a support body from an open cell foam ceramic material ...
2	A, Y02A, Y02C, Y02E	Absorber fluid, comprises a carbon dioxide binding absorbent and an ionic additive in a concentration, which is greater than a minimum concentration, so that the activity of the products formed by the connection of carbon dioxide to the absorbent is reduced ...
⋮	⋮	⋮
P	B, F, Y02C	The invention relates to a power plant for generating electrical energy, comprising a combustion chamber for producing steam, at least one waste gas purification stage that is connected downstream, a separation stage for CO2 ...

Note: Corpus comprises $P \sim 560,000$ patents (all patents filed by German firms after 1990) and a vocabulary size of $V \sim 370,000$ (after text preprocessing).

Clean technology classes by European Patent Office (EPO)

	Clean technology field		Technology example
1	Adaption	Technologies for the adaption to climate change	Genetically modified plants resistant to drought
2	Battery	Battery storage and fuel cells	Fuel cell technologies in production processes
3	Biofuels	Biofuel technologies	Algae biomass
4	CCS	Carbon capture, storage and sequestration	Enhanced coal bed methane recovery
5	E-efficiency	Energy efficiency	Insulation technologies inhibiting radiant heat transfer
6	Generation	Renewable energy generation	Generation of geothermal energy
7	Grid	Grid and power conversion	Smart grids
8	Materials	Low carbon materials and manufacturing	Technologies to replace cement by fly ash in concrete production
9	Mobility	Electric vehicles and low carbon mobility solutions	Ultracapacitors for efficient electric vehicle charging
10	Water	Water and wastewater treatment	Technologies for the production of fertilisers from the organic fraction of waste or refuse

Note: Clean technology fields form the basis for deriving a mapping between specific clean technologies and business models. Patent documents labeled with the corresponding CPC classes by the EPO as listed in the last column are used to derive semantic representations of the respective clean technology field.

Vertical differentiation in technology classes

Classification system of the European Patent Office using the example of **carbon capture and storage technologies**:

CPC	COOPERATIVE PATENT CLASSIFICATION
Y	New technological developments
Y02	Climate change mitigation technologies
Y02C	Carbon capture and storage technologies
Y02C20	Capture and disposal of greenhouse gases
Y02C20/10	- of N_2O

Latent Dirichlet Allocation

Core idea in Blei et al. (2003) seminal work on Latent Dirichlet Allocation (LDA):
Model the generative process that led to the creation of a text corpus incorporating both:

- the observed words in the corpus' documents
- *and* the hidden topic structure within the corpus

in the imaginary data generating process.

The latter includes the distribution of topics over documents and the word distributions over topics.

Latent Dirichlet Allocation

Core idea in Blei et al. (2003) seminal work on Latent Dirichlet Allocation (LDA):
Model the generative process that led to the creation of a text corpus incorporating both:

- the observed words in the corpus' documents
- *and* the hidden topic structure within the corpus

in the imaginary data generating process.

The latter includes the distribution of topics over documents and the word distributions over topics.

L-LDA (Ramage et al., 2009) extents upon LDA by taking into consideration document labelsin the generative process.

L-LDA in patent corpus:

- document $\hat{=}$ patent, p
- labels/topics $\hat{=}$ technology classes, t
- word distributions over topics $\hat{=}$ semantic technology description, δ_t

Statistical Learning in L-LDA

Patent corpus D consisting of P distinct patent abstracts each of length N_p, generative process can be modeled as follows:

1. For each technology class $t \in \{1, \dots, T\}$: generate word distribution $\delta_t \sim Dir(\beta)$
2. For each patent $p \in \{1, \dots, P\}$: generate technology class distribution $\lambda_p \sim Dir(\alpha_p)$
3. For each of the word positions p, n, with $p \in \{1, \dots, P\}$ and $n \in \{1, \dots, N_p\}$:
   3.1 generate technology class assignment $z_{p,n} \sim Multinomial(\lambda_p)$
   3.2 and choose word $w_{p,n} \sim Multinomial(\delta_{z_{p,n}})$

$$p(\delta_{1:T}, \lambda_{1:P}, z_{1:P}, w_{1:P}) = \prod_{t=1}^{T} p(\delta_t) \prod_{p=1}^{P} p(\lambda_p) \left(\prod_{n=1}^{N_p} p(z_{p,n}|\lambda_p) p(w_{p,n}|\delta_{1:T}, z_{p,n}) \right)$$

Goal: Derive word distribution over technology class δ_t from joint distribution $p(\delta_{1:T}, \lambda_{1:P}, z_{1:P}, w_{1:P})$

Gibbs Sampling (1)

$$\underbrace{p(z_{p,n} = t|\mathbf{z}_{p,-n})}_{\substack{\text{Probability that technology } t \\ \text{is chosen for position } n \text{ in patent } p \\ \textbf{conditioned} \text{ on all other} \\ \text{technology-position assignments in the patent}}} \propto \frac{\overbrace{C^{WT}_{w_n,t,-n}}^{\substack{\text{Count of word } w_n \\ \text{in technology } t \\ \text{not including the} \\ \text{current assignment } z_n}} + \beta}{\sum_{i=1}^{N_p} C^{WT}_{w_i,t,-n} + V\beta} \times \frac{\overbrace{C^{WP}_{p,t,-n}}^{\substack{\text{Count of technology } t \\ \text{having already been assigned} \\ \text{to some position in patent } p \\ \text{not including the} \\ \text{current assignment } z_n}} + \alpha_p}{\sum_{j}^{T} C^{WP}_{p,j,-n} + T\alpha_p}$$

- C^{WT}: Word-technology count matrix
- C^{WP}: Word-patent count matrix
- V: Vocabulary size
- T: Number of distinct technologies

Gibbs Sampling (2)

Iteratively draw new technology position attributions according to the above probability and update the topic assignment list with the newly sampled topic for token z_n and re-increment the word-topic and document-topic count matrices with the new sampled topic for token z_n.

After sufficient iterations the probability of a word given a technology can be calculated as follows:

$$\hat{\delta}_{n,t} = \frac{C^{WT}_{w_n,t} + \beta}{\sum_{i=1}^{V} C^{WT}_{w_i,t} + V\beta}$$

Importance of capture contextual meaning of words

- **technical terms** in technology descriptions:
  $X_t = \langle$ gas, absorb, carbon, dioxide, desorption $\ldots \rangle$
- **non-technical terms** in company descriptions:
  '*Developer of direct air capture technology that safely and permanently removes CO2 from the air.*
  $\rightarrow Y_c =$ (developer, direct, air, technology, safe, permanent, remove, co2)'
- **But**: high semantic overlap between x_t and y_c as captured by token embeddings
  $\bar{X}_t(carbon) \approx \bar{Y}_c(co2)$
  $\bar{X}_t(absorb) \approx \bar{Y}_c(remove)$
- **Goal**: Exploit these relations to capture adopters of a technology

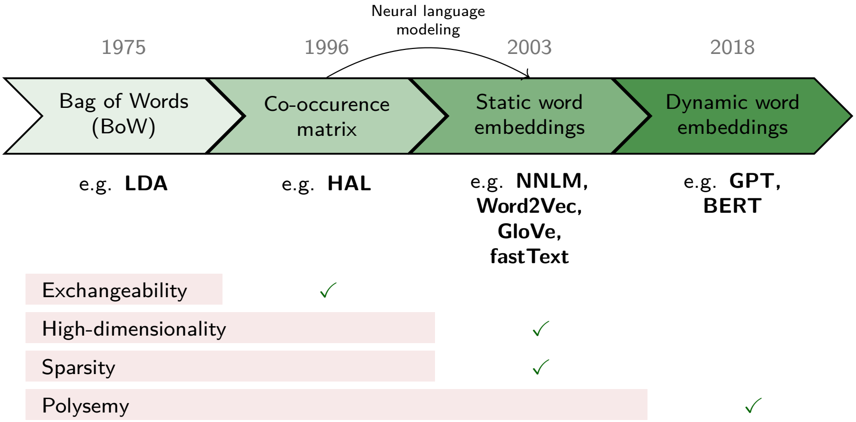
Classification performance of TechProx

Table: Performance of TechProx in distinguishing cleantech from non-cleantech firms

Label	**Precision**	**Recall**	**F1-Score**	**Support**
Cleantech	0.87	0.86	0.86	284
Non-cleantech	0.83	0.84	0.83	233
			0.85	517

Note: Performance measured on random test set with optimal values of $Q = 15$ and $\text{TechProx}_{min} = 0.27$. Optimal values for Q and TechProx_{min} have been determined on the validation set by tuning F1-Score.

Evolution of NLP



Word embeddings (1)

You shall know a word by the company it keeps!

Firth (1957)

General idea: exploit information on co-occurrence of words in large text corpora in order to learn the semantic meaning of a word as represented by a low-dimensional, dense vectors ($E \ll V$).

Natural Language Processing (NLP) as highly active field of research with major advances in recent years (see Wang et al. (2020)):

Neural Network Language Models

- 'distributed representation for words' (Bengio et al., 2003)
  - learn model that predicts next word given previous words
  - word embeddings carrying semantic meaning of a word as by-product

Word embeddings (2)

Static word embeddings

- Word2Vec (Mikolov et al., 2013)
  - neural network architecture specifically designed to learn word embeddings
  - Continuous Bag-of-Words (CBOW): predict word given its surrounding context words
  - Skipgram: predict context words given central word
- GloVe (Pennington et al., 2014)
  - direct exploitation of co-occurence statistics from large text corpora
- fastText (Bojanowski et al., 2017; Joulin et al., 2017)
  - learning embeddings for character n-grams and representing words as the sum of the n-gram embeddings (towards multi-language models)

Word embeddings (3)

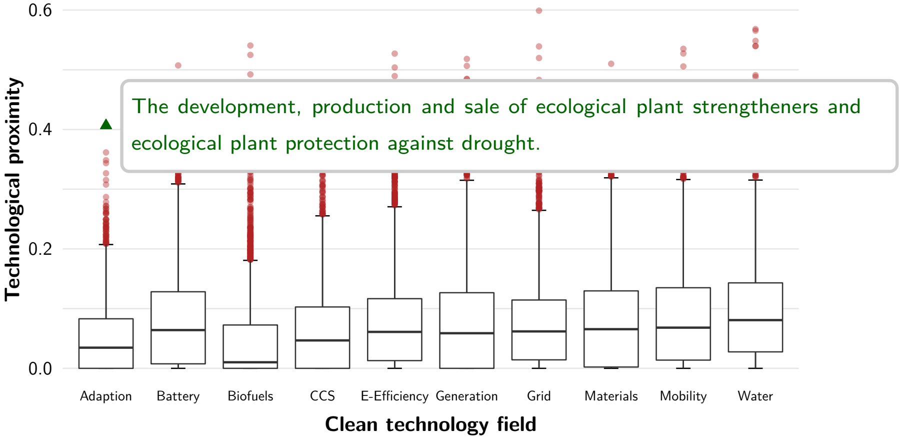
Contextualized word embeddings

Tackle the issue that words have different meanings in different contexts (polysemy)

- ELMo (Peters et al., 2018)
  - use bidirectional LSTM to capture whole sentence (context!) in order to model embeddings of words in sentence
- ULMFit (Howard et al., 2018)
  - introduce a general language model and a process to fine-tune to domain-specific NLP tasks
- GPT (Radford et al., 2018)
  - use transformer decoders to learn linguistic long-term dependencies
- BERT (Devlin et al., 2018)
  - Consider bidirectional contexts and relation of sentence pairs based on transformer encoders

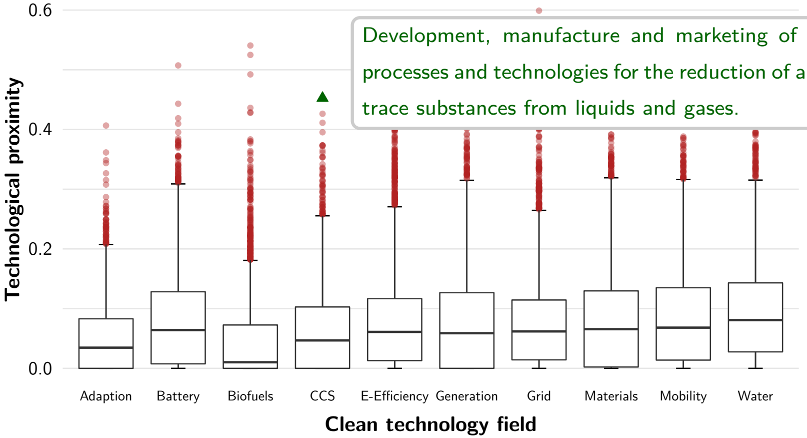
Application

A glance at the 'outliers'



Application

A glance at the 'outliers'

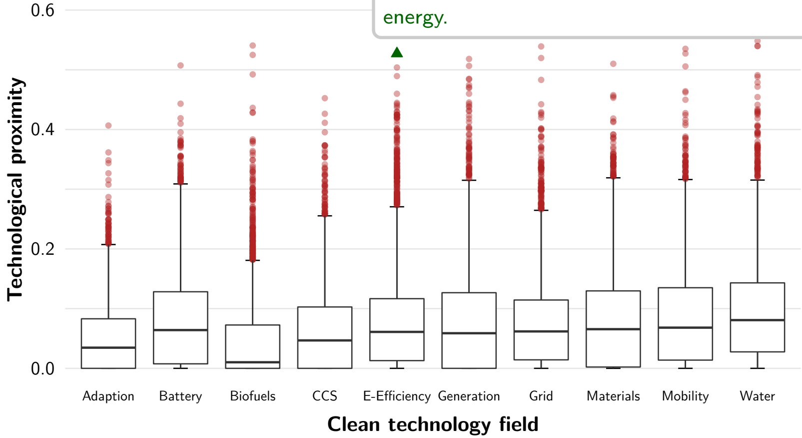


Development, manufacture and marketing of devices, processes and technologies for the reduction of anthropogenic trace substances from liquids and gases.

Application

A glance at the 'outliers'

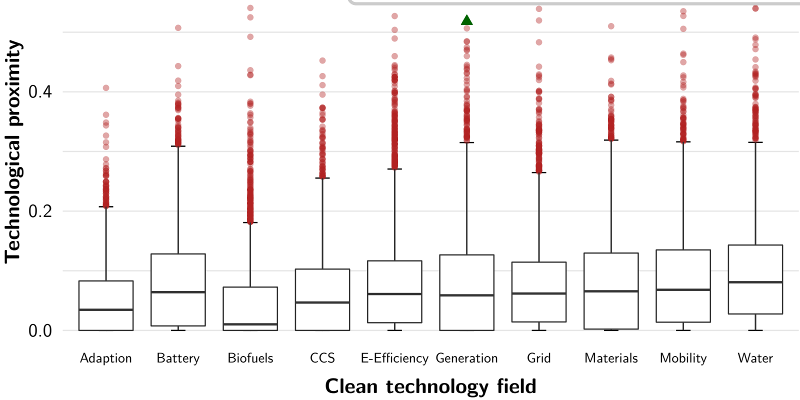
Manufacture, sale, maintenance and repair of heat pumps and other technical equipment, in particular for generating thermal energy.



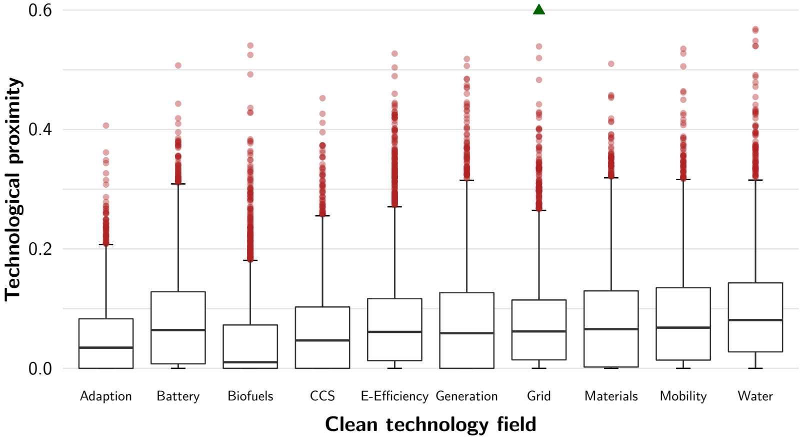
Application

A glance at the 'outliers'

The construction, distribution and sale of turnkey photovoltaic systems and other activities of all kinds in the field of renewable energies.



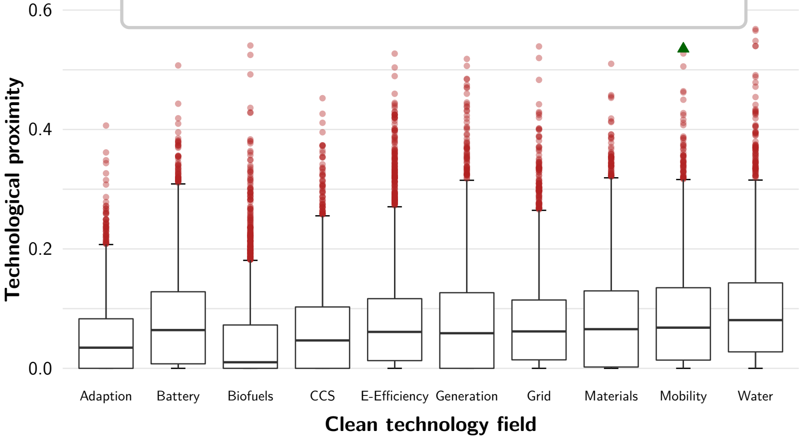
Manufacture of electrode foils, lithium accumulators and energy storage systems and the provision of services in this area.



Application

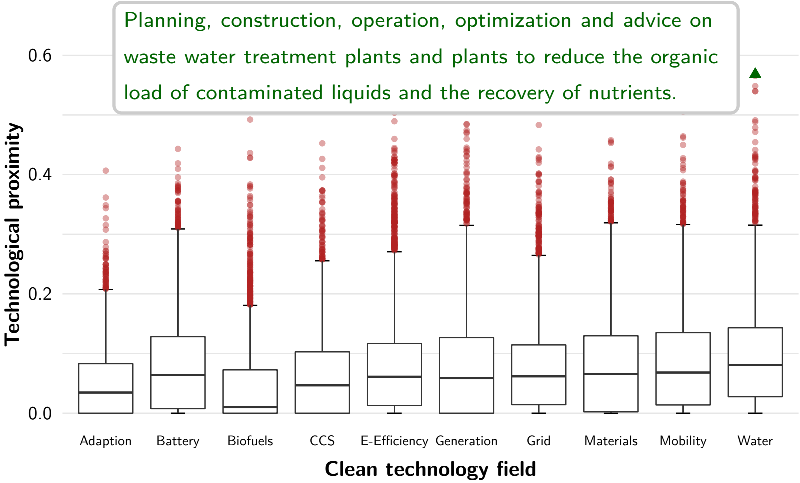
A glance at the 'outliers'

Engine and software optimization on motor vehicles of all kinds.



Application

A glance at the 'outliers'



IAB/ZEW Start-up survey

A representative sample of German start-up companies (Gottschalk, 2013)

Table: 2018 IAB/ZEW Start-up survey questions on environmental impacts and environmental innovation

Environmental impact
Does your company offer products or services which have the following environmental effects on the customer or the end user?
1. Reduction of energy consumption or CO_2 footprint for the customer.
2. Reduction of other emissions to the air, water, soil or noise for the the customer.
3. Reduction of material or resource consumption, for instance water, for the customer.
4. Improvement of recyclability of customer's products.
5. Improvement of durability of customer's products.
Environmental innovation
Since its inception, has your company introduced innovations that have impacted the environment as follows?
1. Reduction of energy consumption or the overall CO_2 balance in your company.
2. Reduction of other emissions to the air, water, soil or noise in your company.
3. Reduction of material or resource consumption, for instance water, in your company.
4. Improvement of recyclability of your own products.
5. Improvement of durability of your own products.

Note: The questions have been asked on a Likert response scale with the following response possibilities. (1) No; (2) Yes, somewhat; (3) Yes, substantial.

Characteristics of clean technology start-ups

Cleantech start-ups show a higher propensity to eco-innovate

	EInno					
	(1)	(2)	(3)	(4)	(5)	(6)
TechProx	1.015***	1.014***	1.013***	1.013***	1.012***	1.014***
log(size)		1.190***	1.140***	1.125***	1.186***	1.175***
age		1.001	1.010	1.001	1.005	1.012
subsidy			1.317***	1.353***	1.413***	1.456***
R&D			1.427***	1.434***	1.605***	1.675***
R&D intensity			0.780	0.910	0.904	0.815
returns				1.743***	1.633**	1.551**
break even				1.295***	1.226**	1.237**
team size					0.899**	0.887**
university					0.614***	0.627***
Sector controls	Y	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	N	Y
N	3,269	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.022	0.026	0.030	0.033	0.041	0.047

EInno := Introduction of environmental innovation?
- no environmental innovation
- environmental innovation with moderate environmental effect
- environmental innovation with substantial environmental effect

Coefficient estimates reported as proportional odds ratios.
Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Appendix

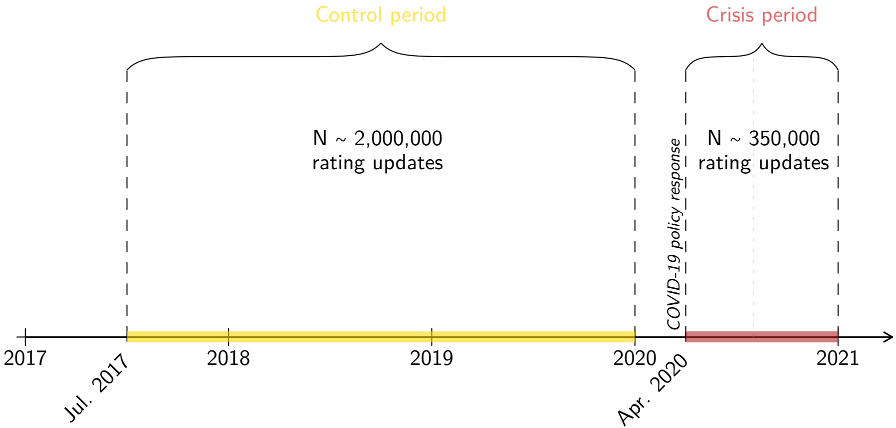
Policy evaluation tool

Motivation

- **COVID-19** caused many companies to fall **short of liquidity** (lockdown measures, drop in demand, logistical difficulties, . . .).
- Federal government temporarily **suspend**ed the **firms' obligation to file for insolvency** (COVInsAG)
- as a result corporate insolvencies dropped substantially despite the worsened economic conditions
- but COVInsAG has been launched largely **indiscriminately** with little control on firms' pre-crisis conditions
  - risk that close to bankrupt firms remain in the market possibly absorbing aid measures as windfall gains
  - counterfactual scenario hard to construct (no controls)

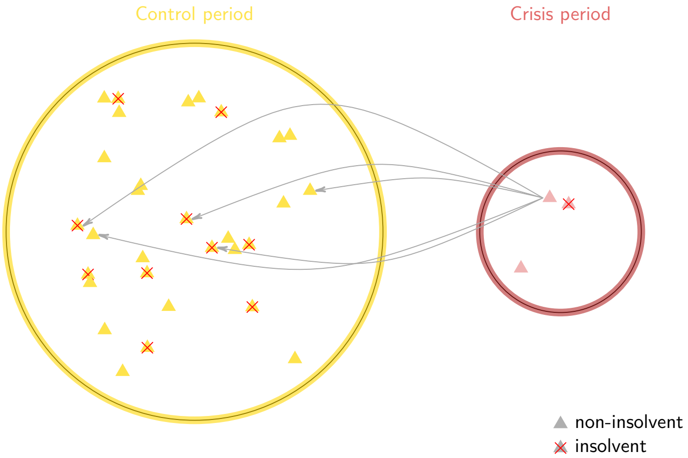
Control and crisis period

Towards a statistical learning framework



Statistical learning task

For each rating update from the crisis period find the k nearest neighbors (kNN) from the pre-crisis period that experienced the very similar rating updates and observe their insolvency state Insolvency rates



Insolvency Gap on the sector-size level

Substantial among micro-enterprises (≤ 10 employees) but vanishes with increasing firm size

Sector affiliation	Size of company		
	Micro $\hat{IG}$	Small $\hat{IG}$	Medium $\hat{IG}$
Manufacturing	+1.0330***	+0.0192	−0.0413
Business-related services	+0.7037***	−0.0072	−0.0530
Food production	+0.2741	+0.2418	−0.1881
Others	+0.3703***	−0.0183	0.0000
Manufacturing of data processing equipment	+0.4419*	−0.0904	0.0000
Mechanical engineering	+0.0325	+0.1768	-0.2458***
Accommodation & catering	+1.1474***	+0.0531	+0.2755
Creative industry & entertainment	+0.1225	+0.1718	0.0000
Health & social services	+0.3698***	+0.0529	−0.1148
Insurance & banking	+0.3696***	0.0000	0.0000
Logistics & transport	+0.7042***	+0.0207	+0.2981
Chemicals & pharmaceuticals	+0.3279*	+0.0299	0.0000
Wholesale & retail trade	+1.0747***	+0.0404	+0.0070

Note: Estimates presented in pp. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$ based on χ^2-Test for equality in the insolvency proportions using Rao-Scott corrections to account for matching weights.

Insolvency Gap and pre-crisis credit rating

Insolvency gap driven by firms with weak pre-crisis conditions

Policy Response in Germany

'Largest assistance package in the history of the Federal Republic of Germany' (Federal Ministry of Finance)

Liquidity provision

- Subsidies and government guarantees
  - 'Soforthilfen'
  - 'Überbrückungshilfen'
  - 'KfW-Schnellkredite'
  - ...
- Labor cost subsidies: 'Kurzarbeitergeld'
- Tax deferrals

Policy Response in Germany

'Largest assistance package in the history of the Federal Republic of Germany' (Federal Ministry of Finance)

Liquidity provision

- Subsidies and government guarantees
  - 'Soforthilfen'
  - 'Überbrückungshilfen'
  - 'KfW-Schnellkredite'
  - ...
- Labor cost subsidies: 'Kurzarbeitergeld'
- Tax deferrals

Change in insolvency regime

Act to Mitigate the Consequences of the COVID-19 Pandemic under Civil, Insolvency and Criminal Procedure Law

of 27 March 2020

The Bundestag has adopted the following Act:

Article 1
Act to Temporarily Suspend the Obligation to File for Insolvency and to Limit Directors' Liability in the Case of Insolvency Caused by the COVID-19 Pandemic (COVID-19-Insolvenzaussetzungsgesetz – COVInsAG)

Source: Federal Ministry of Justice

Zombification of Economy?

The Economist

Menu

Finance & economics

Sep 26th 2020 edition

The corporate undead

Why covid-19 will make killing zombie firms off harder

Easier access to credit and government support means they will stumble on

Zombification of Economy?

T
Econ

Finan

The corp

Wh
firn

Easier a

The New York Times

Europe's Bankruptcies Are Plummeting. That May Be a Problem.

Governments have extended national programs to keep troubled businesses afloat, but the aid may only be postponing a painful reckoning.

By Liz Alderman

Jan. 25, 2021

Zombification of Economy?

The Econ

Finan

The corp

Wh firn

Easier a

The New York Times

Europ

Governme reckoning.

By L

Jan. 25, 2021

Handelsblatt

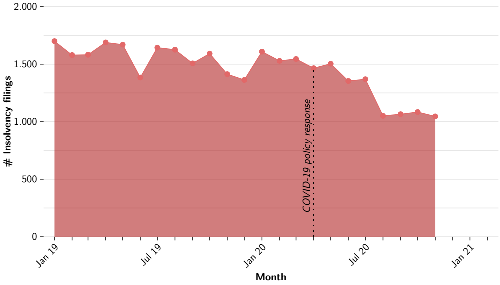
FIRMENPLEITEN

Insolvenzverwalter warnen vor Zombie-Unternehmen

von: Heike Anger • Kirsten Ludowig
Datum: 10.08.2020 16:53 Uhr

Die Regierung will überschuldeten Firmen in der Cororonakrise mehr Luft verschaffen. Doch Experten fürchten massive Schäden für die Wirtschaft.

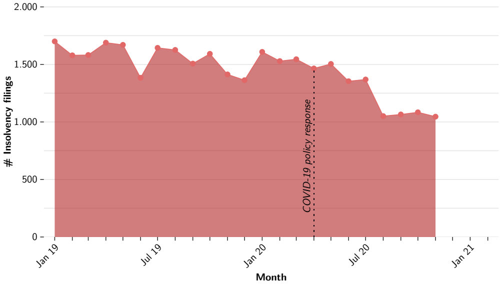
Corporate insolvencies and economic shocks



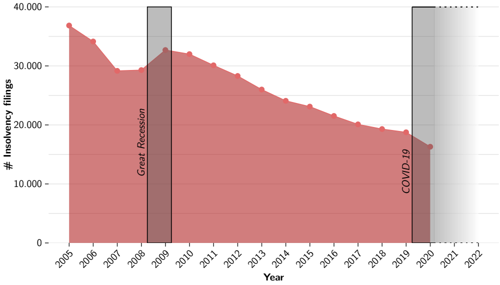
Source: Destatis (2020)

In 2020, 16% decrease in corporate insolvencies compared to 2019. latest insolvency numbers

Corporate insolvencies and economic shocks



Source: Destatis (2020)



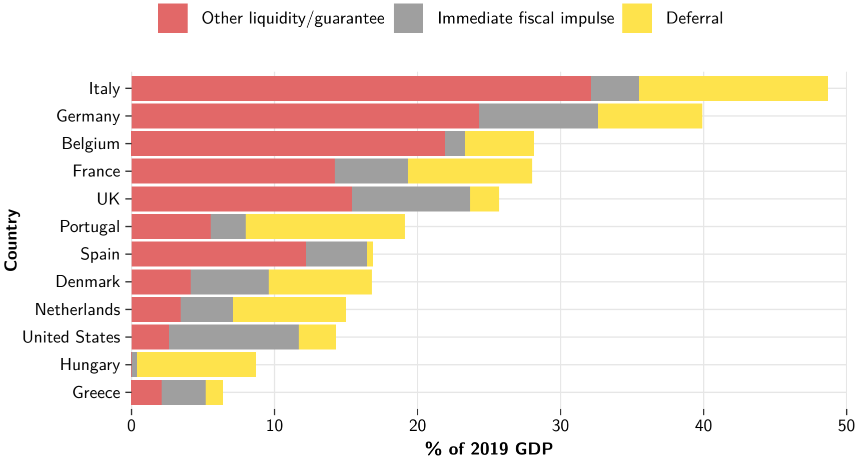
Source: Destatis (2020)

In 2020, 16% decrease in corporate insolvencies compared to 2019. latest insolvency numbers

Typically, corporate insolvencies rise in times of economic crisis (cleansing mechanism).

COVID-19 Fiscal Policy Response

By international comparison



Source: Bruegel

Cleansing mechanism of economic crises

Efficient resource reallocation:

- crises force unproductive companies out of the market
- freeing up resources
- that find more productive use elsewhere

(Schumpeter, 1942; Caballero et al., 2008)

Has the COVID-19 policy response impaired the cleansing effect typically observed in economic crises?

Data

Credit ratings

Insolvency information

Firm characteristics

Data

Credit ratings

Scoring index by Creditreform incorporating

- payment discipline
- legal form
- credit line limits
- financial account indicators
- ...

$r_{it} \in [100, 500]$

Insolvency information

Firm characteristics

Data

Credit ratings

Scoring index by Creditreform incorporating

- payment discipline
- legal form
- credit line limits
- financial account indicators
- ...

$$r_{it} \in [100, 500]$$

Insolvency information

Business insolvency declarations at German insolvency courts including

- firm identification
- filing date

$$f_{it} = \begin{cases} 0 & \text{if } i \text{ non-insolvent at } t \\ 1 & \text{if } i \text{ insolvent at } t \end{cases}$$

Firm characteristics

Data

Credit ratings

Scoring index by Creditreform incorporating

- payment discipline
- legal form
- credit line limits
- financial account indicators
- ...

$$r_{it} \in [100, 500]$$

Insolvency information

Business insolvency declarations at German insolvency courts including

- firm identification
- filing date

$$f_{it} = \begin{cases} 0 & \text{if } i \text{ non-insolvent at } t \\ 1 & \text{if } i \text{ insolvent at } t \end{cases}$$

Firm characteristics

Firm information from Mannheim Enterprise Panel

- industry sector
- firm size
- ...

$$\mathbf{X}_{it}$$

Data

Credit Ratings and Insolvency Information

Credit rating information incorporating

- payment discipline,
- legal form,
- credit line limits,
- financial account indicators,
- ...

Data

Credit Ratings and Insolvency Information

Credit rating information incorporating

- payment discipline,
- legal form,
- credit line limits,
- financial account indicators,
- ...

used as basis for firms' solvency & liquidity state both

- in practice: Probability of default (PD)

Data

Credit Ratings and Insolvency Information

Credit rating information incorporating

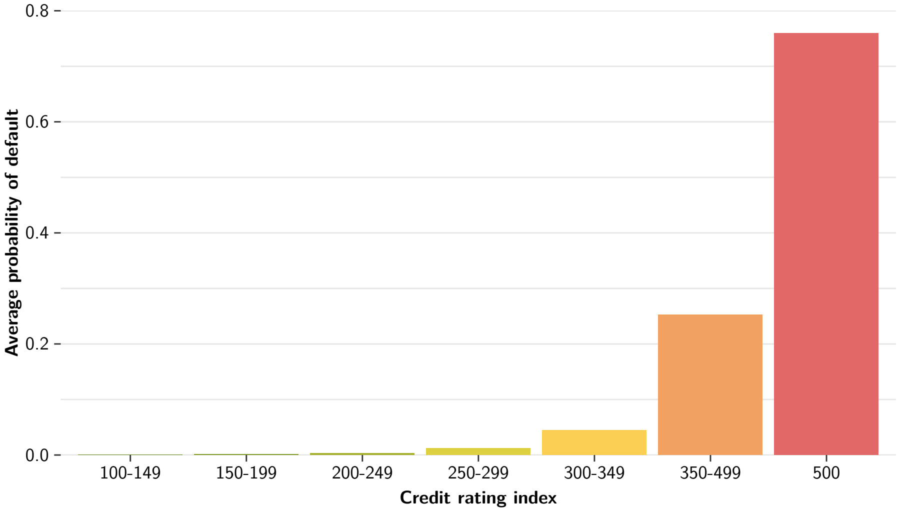
- payment discipline,
- legal form,
- credit line limits,
- financial account indicators,
- ...

used as basis for firms' solvency & liquidity state both

- in practice: Probability of default (PD)
- in research: Insolvency risk (Altman, 1968, 2013)

Credit Rating Data

Commonly used by banks (probability of default of debtors) and by research (insolvency risk estimation)



Source: Creditreform

Lack of controls

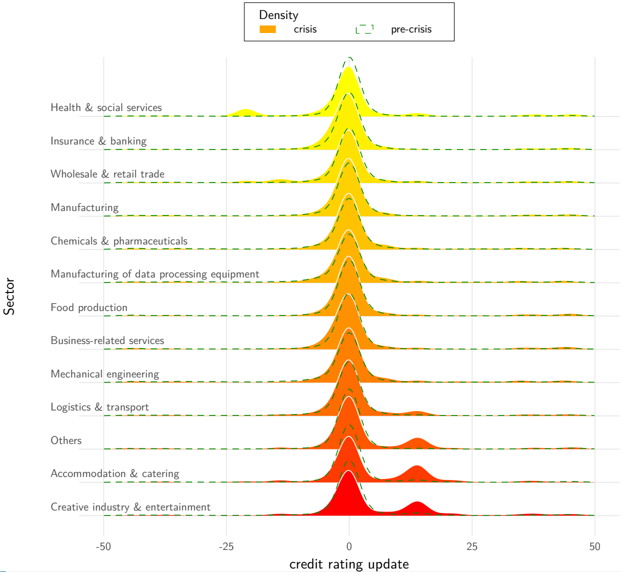
- COVInsAG has been granted indisriminately
- lack of (contemporaneous) control units
- hard to asses policy effect on cleansing mechanism empirically
- can we still ‘construct’ a counterfactual scenario?

Lack of controls

- COVInsAG has been granted indisriminately
- lack of (contemporaneous) control units
- hard to asses policy effect on cleansing mechanism empirically
- can we still ‘construct’ a counterfactual scenario?
  - a look at credit rating updates

Lack of controls

- COVInsAG has been granted indisriminately
- lack of (contemporaneous) control units
- hard to asses policy effect on cleansing mechanism empirically
- can we still 'construct' a counterfactual scenario?
  - a look at credit rating updates



Nearest neighbor matching

Some more details

- only match control units, j, from the same sector-size stratum
- within sector-size stratum calculate Mahalanobis distance (MD) between each possible pair of control and crisis unit, i, on

Nearest neighbor matching

Some more details

- only match control units, j, from the same sector-size stratum
- within sector-size stratum calculate Mahalanobis distance (MD) between each possible pair of control and crisis unit, i, on
  - rating update (with caliper!): Δr_{it}
  - rating prior to update: $r_{i,t-x}$
  - number of downgrades preceding the update: d_{it}
  - average rating before the update: $\bar{r}_{it}$
  - company age: a_{it}

Nearest neighbor matching

Some more details

- only match control units, j, from the same sector-size stratum
- within sector-size stratum calculate Mahalanobis distance (MD) between each possible pair of control and crisis unit, i, on
  - rating update (with caliper!): Δr_{it}
  - rating prior to update: $r_{i,t-x}$
  - number of downgrades preceding the update: d_{it}
  - average rating before the update: $\bar{r}_{it}$
  - company age: a_{it}

$$MD_{ij} = \begin{cases} (\mathbf{X}_i - \mathbf{X}_j)'\Sigma^{-1}(\mathbf{X}_i - \mathbf{X}_j) & \text{if} \quad |\Delta r_{it} - \Delta r_{jt}| \leq c \\ \infty & \text{if} \quad |\Delta r_{it} - \Delta r_{jt}| > c \end{cases}$$

with $\mathbf{X} = (\Delta r_t \; r_{t-x} \; d_t \; \bar{r}_t \; a_t)'$, Σ as the variance-covariance matrix of $\mathbf{X}$ in the pooled sample of in-crisis and all pre-crisis observations and c a predefined caliper on the rating update.

From insolvency rates to insolvency gap

Actual insolvency rate

$$IR_s^{actual} = \frac{N_s^{insolvent}}{N_s}$$

Counterfactual insolvency rate

$$IR_s^{counterfactual} = \frac{\sum_{j=1}^{\tilde{N}_s} w_{j,s} \mathbf{1}(f_{j,t+4}=1)}{\sum_{j=1}^{\tilde{N}_s} w_{j,s}}$$

Insolvency gap

From insolvency rates to insolvency gap

Insolvency gap as the deviation between expected and observed insolvency rates

Actual insolvency rate

$$IR_s^{actual} = \frac{N_s^{insolvent}}{N_s}$$

Counterfactual insolvency rate

$$IR_s^{counterfactual} = \frac{\sum_{j=1}^{\tilde{N}_s} w_{j,s} \mathbf{1}(f_{j,t+4}=1)}{\sum_{j=1}^{\tilde{N}_s} w_{j,s}}$$

$$IG_s = IR_s^{counterfactual} - IR_s^{actual}$$

Insolvency gap

Statistical Learning in kNN (1)

$$IR_s^{counterfactual} = \frac{\sum_{j=1}^{\tilde{N}_s} w_{j,s} \mathbf{1}(f_{j,t+4} = 1)}{\sum_{j=1}^{\tilde{N}_s} w_{j,s}} \quad \text{with} \quad \tilde{N}_s = \sum_{j=1}^{\tilde{N}_s} w_{j,s}$$

$$= \frac{1}{N_s} \sum_{i=1}^{N_s} Pr\left(f_{i,t+4} = 1 \mid X_i\right)$$

Find k_s observations from pre-crisis control group which are closest to X_i and average their survival status:

$$\hat{f}(X_i) = Pr\left(f_{i,t+4} = 1 \mid X_i\right) = \frac{1}{k_s} \sum_{j \in N_k(X_i)} \mathbf{1}(f_{j,t+4} = 1)$$

Closeness is defined by Mahalanobis distance:

$$MD_{ij} = \begin{cases} (\mathbf{X}_i - \mathbf{X}_j)' \Sigma^{-1} (\mathbf{X}_i - \mathbf{X}_j) & \text{if} \quad |\Delta r_{it} - \Delta r_{jt}| \leq c \\ \infty & \text{if} \quad |\Delta r_{it} - \Delta r_{jt}| > c \end{cases}$$

Variables:

i	crisis observation
j	control observation
s	sector-size stratum
N_s	number of observed firms in s
$\tilde{N}_s$	number of matched pre-crisis obs. in s
$w_{j,s}$	matching weight on j in s
$f_{j,t+4}$	survival status of j 4 month after rating update
X_i	observed firm characteristics of i
k_s	matched number of NNs in s
$N_k(X_i)$	k closest points in neighborhood of X_i
Δr_{it}	rating update of i in t

Matching details:

$k_s = \frac{N_s^{control}}{N_s^{crisis}}$

caliper on Δr_t $(= 8)$

matching with replacement

crisis units w/o match neglected in IR_s^{actual}

Statistical Learning in kNN (2)

Why not simply calculate $IR_s^{counterfactual}$ on *all* pre-crisis observations that fall in the respective sector-size stratum?

- control for credit rating update!
- remove bias in comparing IR_s^a and IR_s^c due to additional firm characteristics whose distribution differs in control and crisis sample (Rubin, 1973)
  - e.g. larger firms more often evaluated by rating agency → these are more likely observed in the first months after the crisis

Company Size Definition

	Size of company			
	Micro	Small	Medium	Large
Number of employees	≤ 10	11 – 49	50 – 249	≥ 250
Annual tunover in M €	≤ 2	2 – 10	10 – 50	> 50
Annual balance sheet total in M €	≤ 2	2 – 10	10 – 43	> 43

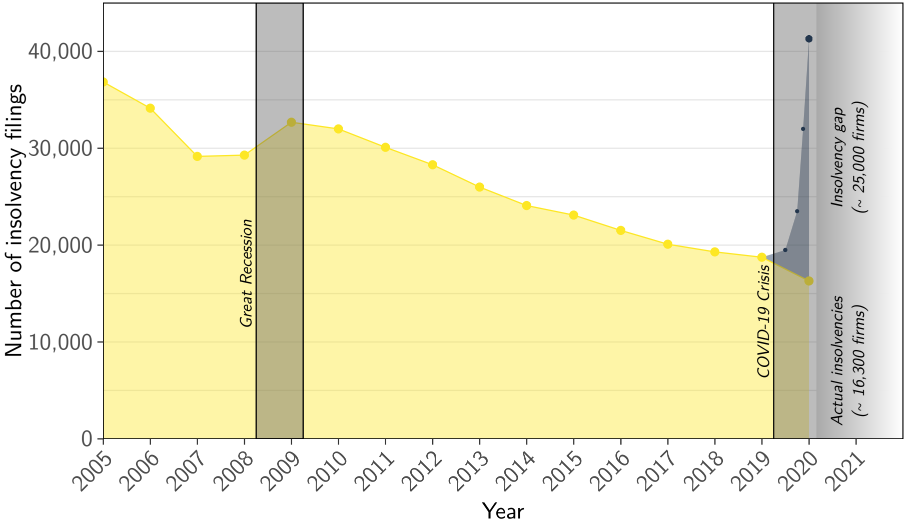
Note: Table shows translation of firm characteristics into company size classes used in this study as defined by European Commission (2003).

Insolvency Gap in Absolute Numbers (1)

Sector	Size of company						$\sum$
	Micro		Small		Medium		
	N_s	IG_s (in %)	N_s	IG_s (in %)	N_s	IG_s (in %)	
Accommodation & catering	37,633	0.0115	4,852	0.0005	810	0.0028	
Creative industry & entertainment	16,057	0.0012	1,910	0.0017	476	0.0000	
Food production	8,191	0.0027	3,674	0.0024	1,962	-0.0019	
Health & social services	69,029	0.0037	12,331	0.0005	4,269	-0.0011	
Insurance & banking	46,670	0.0037	2,583	0.0000	1,290	0.0000	
Logistics & transport	43,899	0.0070	10,756	0.0002	2,773	0.0030	
Chemicals & pharmaceuticals	5,170	0.0033	3,980	0.0003	2,342	0.0000	
Manufacturing of data proc. eq.	4,270	0.0044	2,449	-0.0009	1,057	0.0000	
Mechanical engineering	10,567	0.0003	6,828	0.0018	3,386	-0.0025	
Business-related services	287,115	0.0070	40,448	-0.0001	9,871	-0.0005	
Manufacturing	251,027	0.0103	50,447	0.0002	12,399	-0.0004	
Others	37,695	0.0037	5,381	-0.0002	2,398	0.0000	
Wholesale & retail trade	201,838	0.0107	46,342	0.0004	10,549	0.0001	
Weighted insolvency gap (in %)	0.0080		0.0003		-0.0003		
Number of active firms (official statistics)	3,109,261		293,610		63,928		3,466,799
Insolvency gap (absolute)	24,933		90		-19		25,004

Note: Insolvency gap in absolute terms is calculated as product between the weighted insolvency gap and the total number of active German firms within the respective size class.

Insolvency Gap in Absolute Numbers (2)

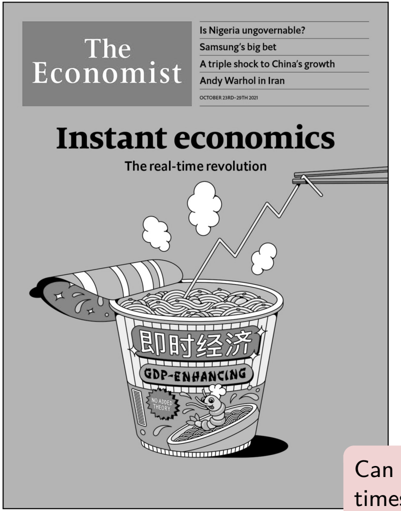


Main findings

- policy response allowed to prevent large-scale business insolvencies ...
- at the cost of saving firms that would have likely ended insolvent *regardless* of the COVID-19 shock ...
- possibly impeding efficient resource reallocation during the COVID-19 crisis

Appendix
Leading indicator development

Motivation - lack of real-time economic data



Source: The Economist (2021a)

Does anyone really understand what is going on in the world economy? The pandemic has made plenty of observers look clueless.

Especially in times of rapid change, policymakers have operated in a fog.

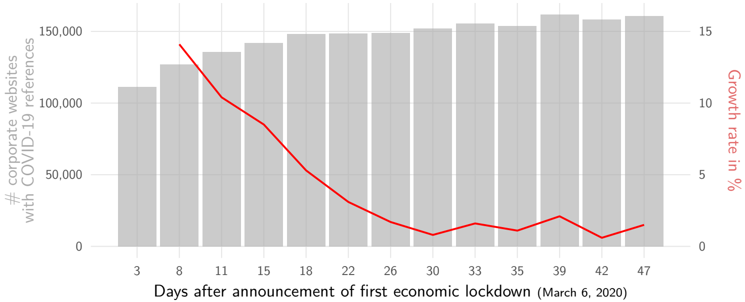
The gap between official data and what is happening in the real economy can still be glaring.

The Economist (2021a, 2021b)

Can we assist policy makers with **timely** and **insightful** firm level data in times of dynamic economic shocks such as COVID-19?

Early firm communication and corporate websites

- accessed corporate websites of ~ 1.18M German companies from Mar 20 - May 20 twice a week searching for references related to the pandemic
- finding: companies used their websites intensively to report about the pandemic



Turn website references into knowledge

But: context of Corona references greatly differed across firms:

'The Corona pandemic is not only
affecting ongoing [redacted] projects,
but also the current selection
rounds of the 13th and 14th funding seasons.'

* * *

'We have therefore decided to adapt
our services to the current situation
and to limit them until further notice.
Although we want to continue to
provide you with all indispensable
services, we also want to meet the
recommendations of the federal government
on how to deal with the corona virus.'

* * *

'Your [redacted] advisor stands by your side
- also in times of COVID-19.'

Turn website references into knowledge

But: context of Corona references greatly differed across firms:

'The Corona pandemic is not only affecting ongoing [redacted] projects, but also the current selection rounds of the 13th and 14th funding seasons.'

* * *

'We have therefore decided to adapt our services to the current situation and to limit them until further notice. Although we want to continue to provide you with all indispensable services, we also want to meet the recommendations of the federal government on how to deal with the corona virus.'

* * *

'Your [redacted] advisor stands by your side - also in times of COVID-19.'

Statistical learning approach:

1. introduced 5 meaningful & distinguishable classes

   (1) problem, (2) confidence, (3) adaption, (4) information, (5) unclear

Turn website references into knowledge

But: context of Corona references greatly differed across firms:

'The Corona pandemic is not only affecting ongoing [redacted] projects, but also the current selection rounds of the 13th and 14th funding seasons.'

* * *

'We have therefore decided to adapt our services to the current situation and to limit them until further notice. Although we want to continue to provide you with all indispensable services, we also want to meet the recommendations of the federal government on how to deal with the corona virus.'

* * *

'Your [redacted] advisor stands by your side - also in times of COVID-19.'

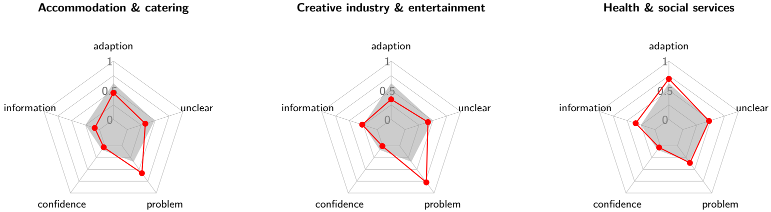
Statistical learning approach:

1. introduced 5 meaningful & distinguishable classes

   (1) problem, (2) confidence, (3) adaption, (4) information, (5) unclear
2. manually annotated ~ 4,000 references
3. fine-tuned pre-trained language model (XLM-R by Conneau et al. (2019))

Early insights from website analysis

- classified *firm level* communication on websites revealed impact heterogeneity at *sector level*
- insights generated at near-real time (right after shutdown announcement in Mar 20)



Classified website references as leading indicators

for later credit rating updates

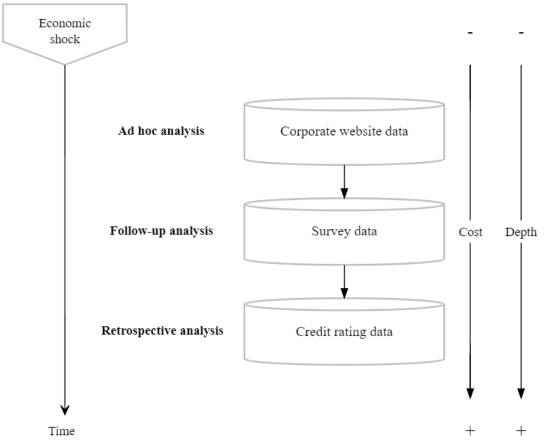
$$\Delta r_{i,\bar{t}+z} = \alpha + \beta_1 \text{Problem}_{i,\bar{t}} + \beta_2 \text{Confidence}_{i,\bar{t}} + \beta_3 \text{Adaption}_{i,\bar{t}} + \beta_4 \text{Information}_{i,\bar{t}} + \beta_5 \text{Unclear}_{i,\bar{t}} + \gamma r_{i,\bar{t}-x} + \boldsymbol{\delta FE}_i + \epsilon_i$$

	(1) $\Delta r_{\bar{t}+z}$	(2) $\Delta r_{\bar{t}+z}$	(3) $\Delta r_{\bar{t}+z}$	(4) $\Delta r_{\bar{t}+z}$
Problem$_{\bar{t}}$	+1.66***	+1.68***	+1.62***	+0.42**
Confidence$_{\bar{t}}$	-1.70***	-1.69***	-1.73***	-0.69
Adaption$_{\bar{t}}$	-0.46***	-0.47***	-0.33***	-0.13
Information$_{\bar{t}}$	-0.24***	-0.24***	-0.23***	-0.17***
Unclear$_{\bar{t}}$	-0.42***	-0.42***	-0.10	-0.08
$r_{\bar{t}-x}$	-0.09***	-0.10***	-0.11***	-0.13***
Age FE	No	Yes	Yes	Yes
Size FE	No	No	Yes	Yes
Sector FE	No	No	No	Yes
N	61,228	61,138	57,343	57,343

Δr_i: credit rating update (**down**grade, **up**grade) of firm i
$\bar{t}$: 01/03/20 - 31/05/20, $\bar{t}+z$: z days after 01/06/20, $\bar{t}-x$: x days before 01/03/20
FE: fixed effects. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Main contributions

- **proposed** a **data framework for policy guidance** in times of economic shocks
  Follow-up surveys Outcome analysis
- to **overcome information deficits** policy makers are confronted with in highly dynamic situations
- possibly allowing more targeted liquidity injections to support affected companies instead of choosing the 'bazooka' as policy instrument



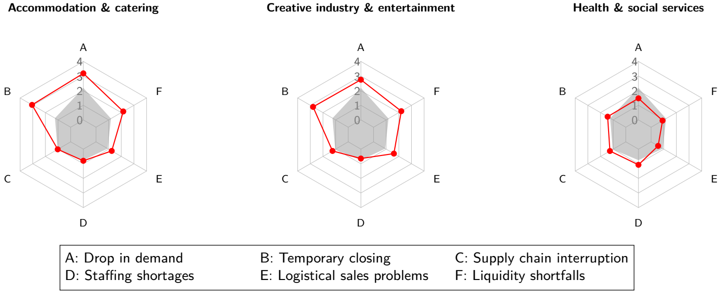
Classification of COVID-19 web references

Categories	Description	Examples (translated)
Problem	Firm reports about adverse impacts of the pandemic on its business operations.	Due to the Corona pandemic, [redacted] & [redacted] are closed. [redacted] has been cancelled due to the increasing concerns and escalated circumstances surrounding the recent coronavirus (COVID-19) outbreak.
Confidence	Firm indicates that the pandemic has no negative impacts on its business operations.	We are there for you 24/7 as usual despite Corona! Your [redacted] advisor stands by your side - also in times of COVID-19.
Adaption	Firm reports that it is adapting to the new economic circumstances.	We have also upgraded our IT and telecommunications system. Our employees are now also able to ensure that you are looked after from home, should this be necessary. Since we receive new information on the development of the coronavirus, the measures and the safety precautions every day, we will continue to monitor the development and react to it. Within our emergency opening times, we particularly take care of those who are currently performing at their best for our society in view of the coronavirus crisis and who depend on their glasses for their work.
Information	Firm reports generally, not necessarily in a business-context, about the pandemic.	The corona pandemic affects each of us now and in the near future. There are many uncertainties and resulting (insurance) issues. What about entitlement to holiday cancellations, health protection abroad and coverage in the event of business interruption are just a few of the questions. In cooperation with the software provider [redacted], the Bundesverband Pflegemanagement (Federal Association of Care Management) is launching a platform to recruit former care professionals to cope with the currently dramatic challenges facing care against the background of the Corona crisis.
Unclear	COVID-19 reference does not come with further clearly distinguishable content.	Current situation COVID-19. COVID-19 and how it affects us.

Follow up surveys

- based on the early findings, construct targeted businesses surveys
- gain more detailed understanding about the sort of impact in order to design counter measures most effectively
- here: surveyed ~ 1,500 companies consecutively (Apr, Jun, Sep 2020) with targeted impact questions

Figure: Targeted impact questions at sector level



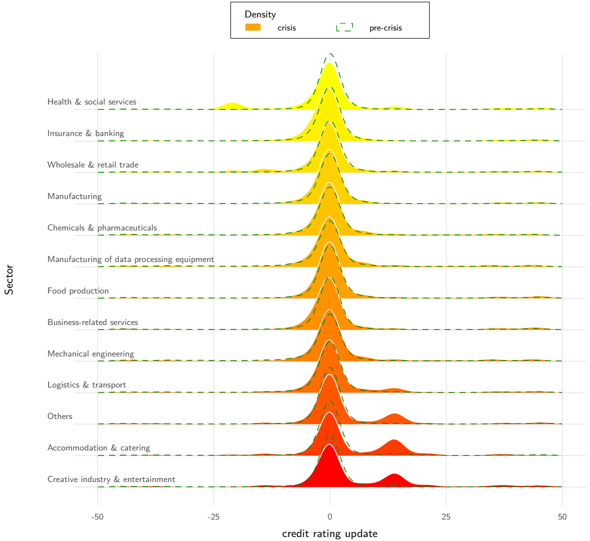
Retrospective analysis of firm outcomes

- after economic shock has materialized in economy, analyze firm outcomes
- understand possible long-term consequences and design stimulus programs
- here: examined credit rating updates of $\sim$ 870,000 companies (between Jun 20 - Apr 21)

Retrospective analysis of firm outcomes

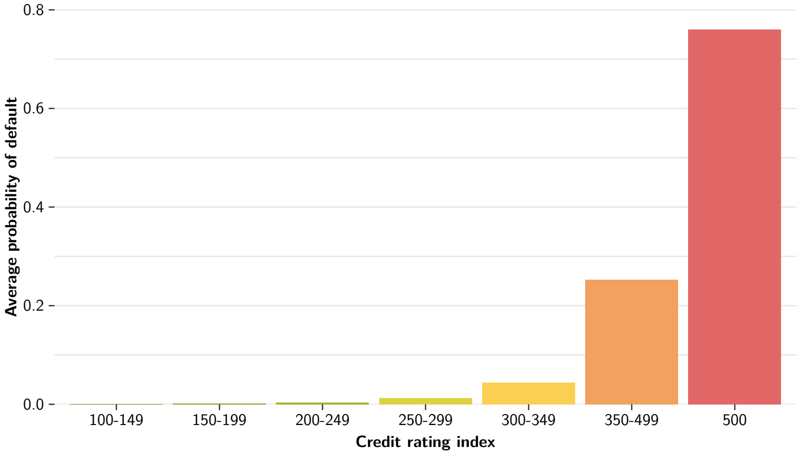
- after economic shock has materialized in economy, analyze firm outcomes
- understand possible long-term consequences and design stimulus programs
- here: examined credit rating updates of ~ 870,000 companies (between Jun 20 - Apr 21)

Figure: Credit rating movements at sector level



Credit Rating Data

Commonly used by banks (probability of default of debtors) and by research (insolvency risk estimation)



Source: Creditreform