with Sweden ranking first. Thus it does not come as a surprise that Denmark has a high number of patents relative to its population. These patents are often within fields such as life sciences and pharmaceuticals, biomedicine, environmental science, and food and agriculture, (Denmark.dk, n.d.). Fields that do not come as a surprise when taking a quick view of Denmark's most prominent firms; Novo Nordisk, Leo Pharma, Novozymes, Chr. Hansen and Grundfos. In addition, Denmark has been a hub for innovation within robotics. An innovation hub that is known for having companies that develop robotics with humans in mind, creating better safety and increased usability in production lines, (Denmark.dk, n.d.). Despite this, Denmark is not perceived as one of the most promising countries for VC investments in the nearest future, (Kraemer-Eis et al., 2022).

4 Data

4.1 Data Sources

This analysis builds on market data from 3 sources; PreQin, Orbis and Vækstfonden.

- PreQin is a platform that contains comprehensive global research on the investment industry. Preqin covers institutional investors, fund performance, fundraising, deals, fund managers, and fund terms for all fund types, (Copenhagen Business School, 2022). PreQin empowers financial professionals who invest in or allocate to alternatives with essential data and insight to make confident decisions. It supports them throughout the investment lifecycle with critical information and leading analytics solutions, (PreQin, n.d.). This paper draws data on approximately 60,000 venture deals from Europe and North America.

- Orbis is a company database comprising information about more than 300 million companies worldwide. The database also covers the banking and insurance sectors. This paper uses approximately 888,000 patents within specific technology fields from Orbis, (Copenhagen Business School, n.d.).

- Vækstfonden is the Danish state's investment fund. Vækstfonden monitors the Danish venture capital market to give an up-to-date view of the venture investments in Danish companies. Their list on venture deals in Denmark is updated by monitoring data sources such as Pitchbook and Crunchbase and going through business newspapers. The list, therefore, does not catch investments that are not public, (O. B. Jensen, 2022a). This list is used for data on the Danish venture capital market.

4.2 Data Description

Table 2 summarizes the data size from Orbis, PreQin and Vækstfonden and shows a great number of venture deals in North America, Europe and Denmark used for this paper's analysis. Furthermore, it shows an extensive amount of companies with a WIPO technology patent. Note that in this paper, Europe refers to geographical Europe without Denmark. Furthermore, North America comprises Canada and the US. This paper concentrates on the venture market in the period from 2018 to the first-half year of 2022 (2022H1).

Table 2: Number of observations from the data sources Vækstfonden, PreQin and Orbis

	Vækstfonden	PreQin	Orbis
Denmark	374		12,289
Europe		20,398	541,152
North America		40,739	335,035
Total	374	61,137	888,476

Note: *Observations from Vækstfonden and PreQin are number of deals, while observations from Orbis are number of companies with a WIPO patent. Moreover, observations from Vækstfonden and PreQin are limited to 2018-2022H1*

A descriptive summary of the distribution of the data used on venture deals has been collected in table 3. For all geographical areas, Denmark, Europe and North America, the distribution of venture deals are highly right-skewed, as expected. This right skewness implies that a few number of firms get large amounts in a single venture capital round. Moreover, the distributions for Denmark and Europe are highly similar, with percentiles up until 90% being relatively close. However, the distribution of venture deals in Europe has a higher average and a notably higher standard deviation, implying a far greater tail in the highest decile compared to Denmark.

Even though the North American distribution is highly right-skewed, as is seen for the European and Danish distribution, there is a clear difference between these. The North American distribution is generally on a higher level, with all of the percentiles being greater than the European and Danish percentiles. One thing to notice is that, even though the North American distribution in general is on a higher level, the European distribution has a greater tail in the highest decile, as implied by the coefficient of variation.

Table 3: Descriptive summary of data in million dollars

	Vækstfonden	PreQin		
	Denmark	Europe	North America	Total
Average	12.05	18.88	31.89	27.50
Standard Dev.	24.46	96.13	128.56	118.77
25% percentile	1.47	1.14	2.50	1.95
Median	3.66	3.16	7.50	5.50
75% percentile	11.40	10.43	24.33	19.99
90% percentile	30.91	34.18	66.00	55.00
Max	238.43	8010.73	7750.00	8010.73
Coefficient of variation	2.03	5.09	4.03	4.32

Note: *The coefficient of variation is the standard deviation divided by the average and is a measure of the relative variation in the distribution. Source: PreQin and Vækstfonden.*

4.3 Methodology: Data Preparation

This paper prepares the data before analyzing by matching the primary data sets to create a meaningful deep tech classification. Typically this matching would be performed based on a firm-specific number, such as the CVR number for Danish registered firms. However, such a number does not exist internationally or at least does not come with the collected data sets. Data, thus, has to be matched based company names.

With more than 60,000 venture deals and 888,000 WIPO patented companies, it is incomprehensible to perform manually. Thus, this paper uses an automated program. However, since computers only identifies perfect fits, names like ('Qvest' and 'Qvest.io aps') and ('Good Monday' and 'GoodMonday aps') would not get matched, resulting in an insufficient detection of deep tech companies. This paper sets up an automated programming system to capture relations in company names to perform better matchings. The automated system contains three main concepts; 1) N-grams, 2) TF-IDF vectorization, and 3) Cosine similarity, as suggested by

(Van Den Berg, 2017) and (Deep, 2020). Matching code is attached at the end of the Appendix, page 47.

4.3.1 N-grams

The concept of N-grams is to collect relations of letters within each name. The person building the matching chooses a number N (in this paper, N=3) that splits each name into all possible combinations of N=3 sequential letters.

From the example in the beginning of section 4.3, the name 'Qvest' is split into the so-called Gram-list of the name, which is given by ('Qve', 'ves', 'est'). The number of sequential letters, N, can be chosen as any positive integer. N=3 is chosen in this assignment because it creates the longest possible Gram-list for every name while capturing as much relation between letters as possible.

4.3.2 TF-IDF Vectorization

The idea of TF-IDF vectorization is to transform the Gram-list, explained in subsection 4.3.1, into a vector that algorithms can manage. The TF-IDF statistical measure, defined below, is used to perform this vectorization.

$$TF - IDFmeasure = TF(g,d) * IDF(g,D)$$

$$TF(g,d) = log(1 + f(g,d)) \quad , \quad f(g,d) = \frac{g}{d}$$

$$IDF(g,D) = log\Big(\frac{M}{f(g,D)}\Big).$$

Where g is the element in the Gram-list, d is the number of elements in the Gram-list for the given name, D is the total number of elements in all Gram-lists in the data sets, and M is the number of names in the data sets. The above calculation is done for each name's elements in the Gram-list, and the vectorization results are composed in a sparse matrix holding the TF-IDF vector for each name.

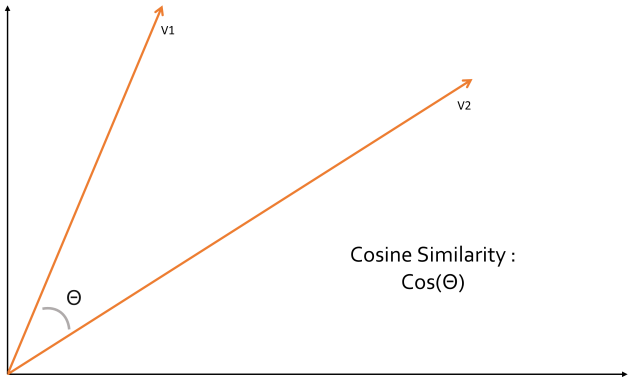
4.3.3 Cosine Similarity

After all occurring names in the datasets have been transformed into numerical values in vector form, see section 4.3.2, the next step is to calculate the similarity between these vectors in the P-dimensional vector space.[1] Various methods can be used to calculate the similarity between two vectors (the similarity between the underlying names). Here we rely on the Cosine similarity between two vectors, as defined below.

$$CosineSimularity = cos(\theta) = \frac{A \cdot B}{||A||\,||B||},$$

where A and B are the two vectors evaluated, and $||A||$ is the Euclidian norm of A. The cosine similarity calculates the cosine of the angle between the two vectors and is bounded in the range $[0; 1]$, see figure 9. Note that illustration 9 is made to show the idea about the cosine similarity. It is not representing data.

Figure 9: Illustration of the cosine similarity in the 2-dimensional space



The cosine similarity is quite intuitive. Depending on the direction the vectors are pointing, the more similar they are. When equal to 1, the vectors point in the same direction. One nice feature of the cosine similarity compared to other vector similarity measures, such as the Euclidian distance is that, the length of the vectors does not matter for the measure, only the direction the vectors are pointing. Said differently, the measure only depends on how many elements in the Gram-list the names have in common and not how often the elements in the Gram-list occur.

[1] P being the total number of unique elements in all Gram-lists

4.3.4 Evaluating the Critical Value

After having constructed a measure that calculates a single number for the similarity between names, a critical value has to be chosen for when two names are similar enough, in the sense of cosine similarity, to be categorized as the same. To choose this critical value, this paper conducted a trial where 1.000 unique company names were drawn from the Preqin data set and matched with the WIPO patented companies. The matches with a similarity greater than 0.5 were then manually assessed for which were correctly matched. The matches could fall into four categories as shown in figure 10. Note that illustration 10 is made to show the idea about the matching procedure. It is not representing data.

Figure 10: Possible outcomes from the matching procedure

Deep tech		Real	
		True	False
Predicted	True	True Deep tech (True positive)	False Deep tech (False positive)
	False	False Non-Deep tech (False negative)	True Non-Deep tech (True negative)

The critical value of the similarity measure was chosen to minimize the sum of false matches, i.e. false positives and false negatives. It resulted in a critical value of 0.8, with an overall matching error of 1.5%. Looking into this overall matching error at the critical value of 0.8 shows that all the errors occurred on false negatives, being companies that were Deep tech but did not get categorized as one. The error matchings are all false negatives when minimizing the total number of false matches because the dataset is unbalanced regarding the category of interest (deep tech/non-deep tech). As seen in figure 10 false positives comes from the non-deep tech category and false negative comes from the deep tech category. Since the non-deep tech category is significantly larger than the deep tech category in our data sets, lowering the critical value slightly introduces a high number of false positives compared to the number of true positives gained. A visualisation of the evaluation metric for the critical value can be seen in appendix 8.2.

This paper's results contains approximately 0% false positives (in the experiment, it was 0.5%) in the matches above the critical value of 0.8. In other words, nearly 100% of the companies categorized as deep tech are deep tech companies, according to this paper's definition. However,

the downside of the unbalanced data set results in all the matching errors falling under the false negatives. Again, since the deep tech category is considerably smaller, the error percentage for this group becomes relatively large, at 5.5%. Thus, the deep tech category is missing 5.5% of the companies that are deep tech according to this paper.

This paper accepts the loss of 5.5% of the deep tech companies because this is this paper's category of interest. Accepting this creates a very high level of truth for companies categorized as deep tech. Thus, when analyzing aggregated sums or shares for the deep tech category, it can be seen as a lower bound. However it is worth noting that all main conclusions in this paper are robust to the choice of the critical value. For example, a more balanced critical value, as seen in the appendix 8.2, does not affeact the main results in this paper.

5 Analysis

The analysis will dive into three aspects of the venture capital market:

1. It will look into the total volume of investments and deals in Denmark, Europe and North America.
2. It will analyze the share of deep tech companies and the amount of invested capital in deep tech versus non-deep tech companies.
3. The analysis will shed light on whether there is a difference in the deep tech industries that get venture capital funding in Denmark, Europe and North America.

5.1 Activity in the Venture Markets Across Denmark, Europe and North America

This section analyzes the activity across all venture capital investments in Denmark, Europe and North America by first looking at an Index Chart, analyzing differences in the development of invested amount and number of deals. Secondly, this section looks at the total volume of invested venture capital and the number of deals. Lastly, it analyzes the average and median investment rounds across stages and geography.