



December 2021

Thematic Basket Incubator

Leveraging Alternative Data with Machine Learning and Knowledge Graph

Helen Krause

Yehuda Dayan

Brian Yeung

Andreas Theodoulou

Citi Global Insights (CGI) is Citi's premier non-independent thought leadership curation. It is not investment research, however it may contain thematic content previously expressed in an Independent Research report. For the full CGI disclosure, click [here](#).

Authors

Helen Krause, CFA

+44 (20) 7986-8653

helen.krause@citi.com

Yehuda Dayan, PhD

+44 (20) 7986-5502

yehuda.dayan@citi.com

Brian Yeung

+44 (20) 7986-8692

brian.yeung@citi.com

Andreas Theodoulou

+44 (20) 7986-1139

andreas.theodoulou@citi.com

Table of Contents

Executive Summary	4
Motivation.....	5
Thematic Investing	5
AI vs. Human-based Approach.....	7
Identify Alternative Data Required	8
Data Sources	8
Thematic Basket Incubator Pipeline	10
Wearable Technology Theme	11
Revenue Eligibility Filter	11
Knowledge Graph – Patents and News Exposure	15
Principal Components Analysis on Concept Exposure Scores.....	18
Thematic Classification Model: In Search of an Enhanced Aggregate Score.....	20
Hiring Exposure – Occupation Group Profiling.....	23
Hiring as a Forward-Looking Measure	24
Bringing It All Together	26
Conclusion	29
Appendix	30

Executive Summary

Thematic investing has gained strong momentum over the last few years in terms of asset gathering, as investors seek investment opportunities through different lenses, outside the traditional country and sector angles. Along with the strong appetite from investors in themes, they also want to be able to examine a theme of interest quickly as often they need to respond to the overarching geopolitical, macro and micro driving forces in the market to adjust their portfolios accordingly.

In order to meet such needs where investors are looking for thematic basket construction solutions that are customisable with a shorter turnaround time, we have developed our “Thematic Basket Incubator” process in which we leverage alternative data and combine that with advanced machine learning models, together with knowledge graph technology. Our approach incorporates the strengths of fundamental analysis with data science insights, providing a systematic and data-driven solution that is highly scalable and flexible to accommodate investors’ specific requirements.

Using the “Wearable Technology” theme as an example, we demonstrate how the process works from identifying relevant data sources, deriving thematic exposures in them to finally combining them to arrive at the final ranking of companies that are deemed to be exposed to the theme. Our framework is capable of distinguishing both positive and negative exposure in public-listed companies as well as private entities, all of which are important to the thematic-minded investors. Specifically, it shows companies which are investing in new themes and driving the latest trends. For corporates which are interested in understanding their peers’ activities in certain themes, our approach can also be very helpful, especially the hiring trends.

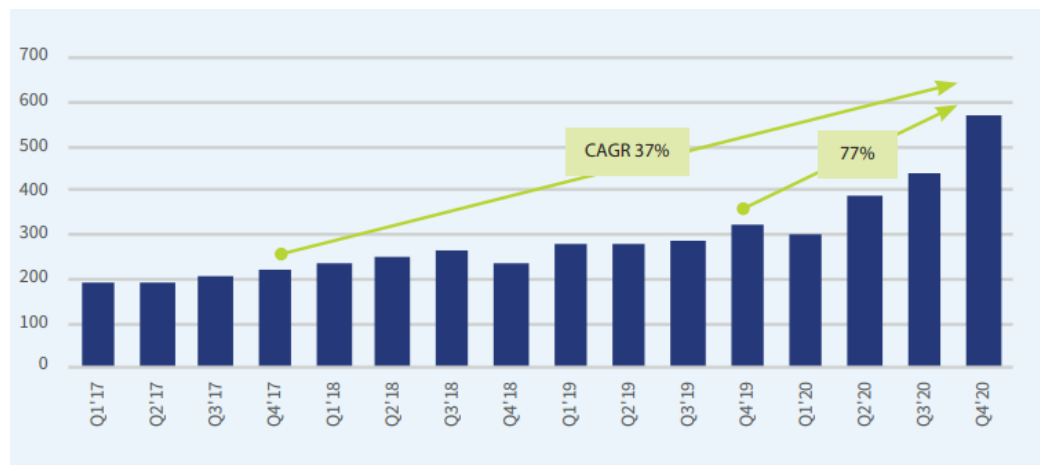
Our Thematic Basket Incubator bespoke solution is available through our Data Science Insights which offers completely customisable data science services to our clients.

Motivation

Thematic Investing

Thematic investing as a concept is not new, but over the last few years it has taken centre stage in the investment industry. The growing popularity of thematic investing is rooted in the recognition of the impact of longer term economic, environmental, political, and social trends which cut across regions and sectors could present themselves as interesting investable opportunities. Indeed, based on a recent report published by Broadridge in 2021, asset gathering in this category has enjoyed an annualised growth rate of 37% since 2018, with a total of Euro 570bn AUM as of end of 2020, and thematic portfolios account for almost 40% of all equity fund net sales since 2017. As depicted in Figure 1, over the last year, the demand for thematic shot up by 77% as technology-related themes gained strong momentum partly as a result of the pandemic.

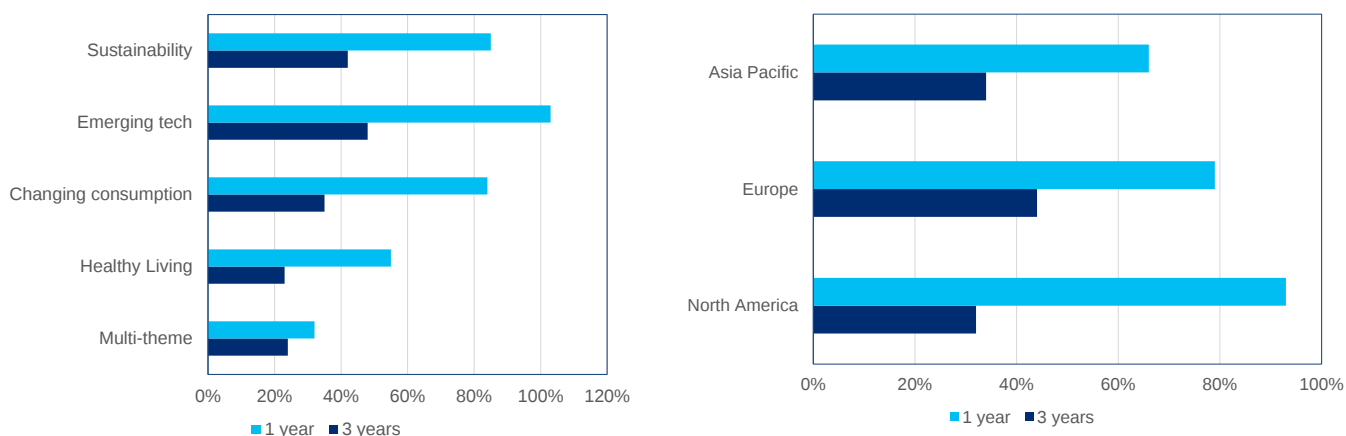
Figure 1. Thematic Fund Growth – Global AUM (Euro bn)



Source: Broadridge

Delving deeper, Figure 2 shows the AUM growth by theme type and by region based on Broadridge's data. Over the past 3 years, the highest growth by theme type was "Emerging Tech" where its popularity soared by over 100+% in 2020. As the world was put in lockdown and people were forced to work from home due to the pandemic, technology advances were accelerated on multiple fronts in order to cope with the sudden increase of demand on better connectivity for work, school and daily life necessities. Closely followed were "Sustainability" and "Changing Consumption" where the former saw the surging interest in social aspect of sustainability while the latter captured consumers' changing behaviour, both also chiefly as a result of the pandemic. Regionally, Europe enjoyed the highest growth in AUM over the three-year period but was overtaken by North America last year when Covid hit. Nonetheless, all three regions witnessed over 60% increases in AUM last year.

Figure 2. Thematic AUM CAGR by Theme Type (LHS) and Region (RHS)



Source: Broadridge, Citi Global Data Insights

Citi has long been active in the thematic space – back in January 2012, Citi GPS¹ published the 2012 Investment themes report where it discussed the potential changing landscape in macro and geopolitical environment that year and its impact on markets, in addition to more micro themes overarching corporates. Since then, Citi has continuously been publishing on new themes with the latest being “Disruptive Innovation VIII”² where themes such as “Alternative Proteins”, “Carbon Capture and “Metaverse” are discussed. In 2013, Citi Research launched its “Global Theme Machine” product where it mapped individual stocks, by leveraging fundamental analysts’ deep understanding of their sectors and companies, to a list of themes which strategists, directors of research, sector heads and the quant team all contributed towards. It recognised the investment potential in themes as country, region and sector in isolation could no longer explain the market movements succinctly as before and cross-effects from these dimensions were observed.

Citi’s Global Theme Machine (GTM) is grounded in deep knowledge about companies and sectors covered by analysts which is then combined with quantitative factor screening in order to assess the attractiveness of each theme. While this approach has distinctive advantages of combining fundamental assessments with quantitative factors, we believe alternative data and AI/machine learning can further enhance the thematic baskets within the GTM or can offer standalone solutions for themes which might be more transient in nature or are dominated by SMEs as these often are the pure plays that are not on analysts’ radar due to the company size.

¹ <https://www.citivelocity.com/citigps/>

² <https://www.citivelocity.com/citigps/disruptive-innovations-viii/>

AI vs. Human-based Approach

The GTM is rebalanced once a year as it takes time and effort to refresh all the mappings across stocks and sectors globally and the product is also aiming for an investment horizon of 5 years which means more frequent rebalance is not deemed to be necessary. Typically analysts have a month to map stocks within their covered sector to the annually updated themes. Once preliminary mappings are completed, the quality assurance process which is run by the quant research team then kicks in where consistency within the same sector and across time is sought. That process can take up to a month to check about 30,000³ data points, covering 87 themes, from 5,500 companies mapped, and involves discussions with respective analyst teams to ensure data accuracy⁴ is achieved.

As thematic investing has become more popular, there is increased demand from investors who want to cast the net as wide as possible and find companies that are smaller in size with purer or higher exposures – both positive and negative – to themes they care about. Also on the rise is the interest in private companies which could present a fruitful hunting ground for returns through direct investments, especially if they bring about disruptive innovations. These are harder to achieve at scale with a human-based approach due to the fact that research analysts in bulge bracket banks typically cover larger public-listed companies within their respective sectors and have deep knowledge about these companies. Smaller companies often feature less in their coverage universe and there is little research on private companies as research analysts' focus is on large liquid public-listed names, which remain at the core of investment opportunities, to service the institutional investor base. Additionally, negative screening is not currently provided within the GTM.

In order to address some of the aforementioned shortcomings and meet investors' demand, we have developed a systematic framework which we call "Thematic Basket Incubator". It leverages several alternative data sources, combined with AI/machine learning models, to provide 'candidate' companies that are deemed to have exposures, positive and negative, to a desired theme. Another advantage of this data-driven, systematic approach is the speed of basket generation and the flexibility of basket update frequency, covering both public and private companies. After the substantial development effort in the platform's data feeds, processing and signal optimization, creation of a basket can take only a couple of days and rebalancing on demand.

Having said that, it is important to recognise that this framework is not infallible as machine learning models tend to come with some degree of noise that cannot be completely eliminated. The key foundation of the GTM is that the thematic exposures are mapped based on the in-depth knowledge of highly trained fundamental analysts and that knowledge is not easily replicable by the machine. This is why a fundamental layer, via the Factset RBICS data, has been incorporated into our thematic basket generation process to provide solid grounding and the GTM data is also used to train our classification models, both of which are designed to mitigate the noise issue. We will describe the building blocks and the process in detail in the following sections.

³ Based on the latest 2021 annually rebalanced GTM, [Global Theme Machine: Quantitative Thematics Come of Age - and New Themes for 2021](#)

⁴ As in reflecting analysts' mapping intention correctly

Identify Alternative Data Required

Data Sources

The first step of our thematic basket incubator process is to identify the types of alternative data needed in order to fulfil the requirements highlighted in the previous section. News sources could be an interesting starting point for screening companies with exposures to certain themes. For example, if one is interested in the “Electric Vehicle” theme and is not fully aware of all companies which have exposure to the theme, it is intuitive to look at news articles where companies are mentioned alongside this topic. The co-occurrence frequency between companies and a specific topic gives us a good indication of how closely related they are to the topic.

However, at times the exposures are not explicitly mentioned in the news articles but can be inferred indirectly. Hence, a simple natural language processing (NLP) pipeline that filters for certain terms would not be sufficient. Instead, we consider “knowledge graphs” in the semantic neural network space which represents a knowledge base that utilises a graph-structured model to capture the interconnectedness of the data points. Through knowledge graph, objects, companies, events and abstract concepts are captured in the network which illustrates the relationships among them. Most commonly used internet search engines deploy knowledge graphs in order to efficiently fetch the most relevant results from their data sources and enrich the search results. For our thematic basket incubator process, we have partnered with Yewno⁵ which leverages a proprietary knowledge graph to infer these connections from data sources such as news, patents corporate filings and others.

In addition, we mentioned earlier the importance of reducing noise from machine learning models. To that end, we have added the GTM historical data from Citi Research as the training set for our classification model which helps to reduce statistical errors. When viewed as a training set, we believe the GTM is unique in the application of machine learning models in the financial markets. It combines size (5,000 companies covered), time (7 years) and quality (labelling is conducted by Citi’s well-ranked Research department). Furthermore, we need a fundamental layer to cross-validate the terms, or ‘concepts’, which define the theme of interest to further reduce ‘noise’ in the process. Factset’s Revenue Business Industry Classification System (RBICS)⁶ provides detailed revenue breakdown of companies by dividing them into 1,700+ sector groups and 6 levels to reflect each company’s main line of business. This system is put together by Factset’s research team where they scrape corporate filings and map the revenue streams of companies into their classification system. In contrast to traditional industry classification systems such as GICS, RBICS allows a company to have exposures to multiple industries rather than a binary labelling into one particular industry. The depth and breadth of this data source helps us have a more transparent

⁵ <http://yewno.com>

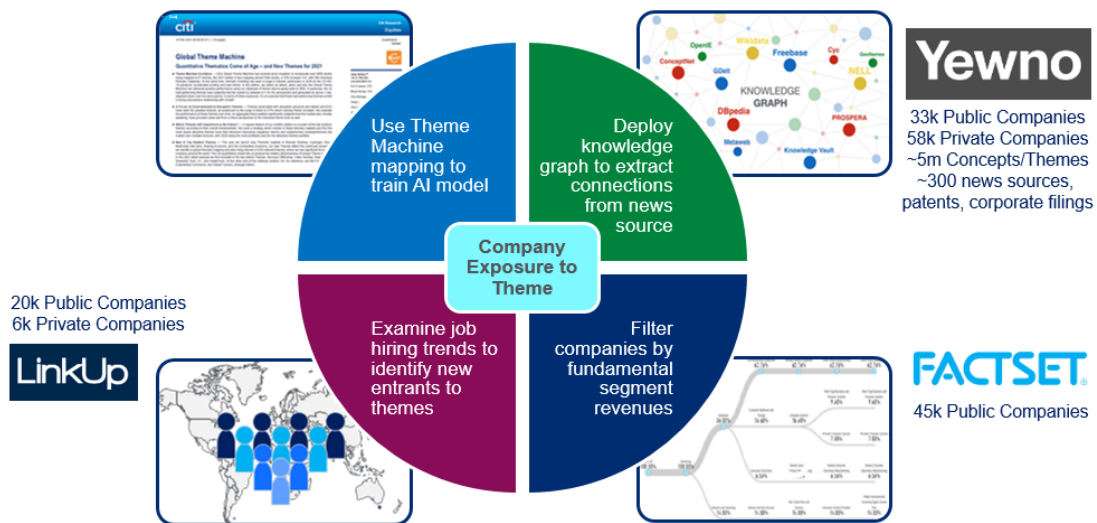
⁶ <https://insight.factset.com/resources/factset-revere-business-industry-classifications-datafeed>

view of the commonalities of companies with similar exposures to themes, adding a fundamental lens that provides the important cross-validation checks.

Finally, job postings data is another source of alternative data that warrants thoughtful consideration in our thematic basket process. While traditionally the first iteration of a thematic exposure screening process is about which companies have exposures (often measured in terms of revenues), investors increasingly are also interested in up-and-coming companies, who have yet to make substantial revenues related to a specific theme, rather than those well-established names, as they present much more attractive future growth opportunities. LinkUp⁷ is our chosen partner for providing the job postings data source which consists of job specifications in their entirety for all the jobs scraped from companies' career websites.

These four alternative data sources for the building blocks of our thematic basket incubator which is depicted in Figure 3.

Figure 3. Thematic Basket Incubator's Building Blocks



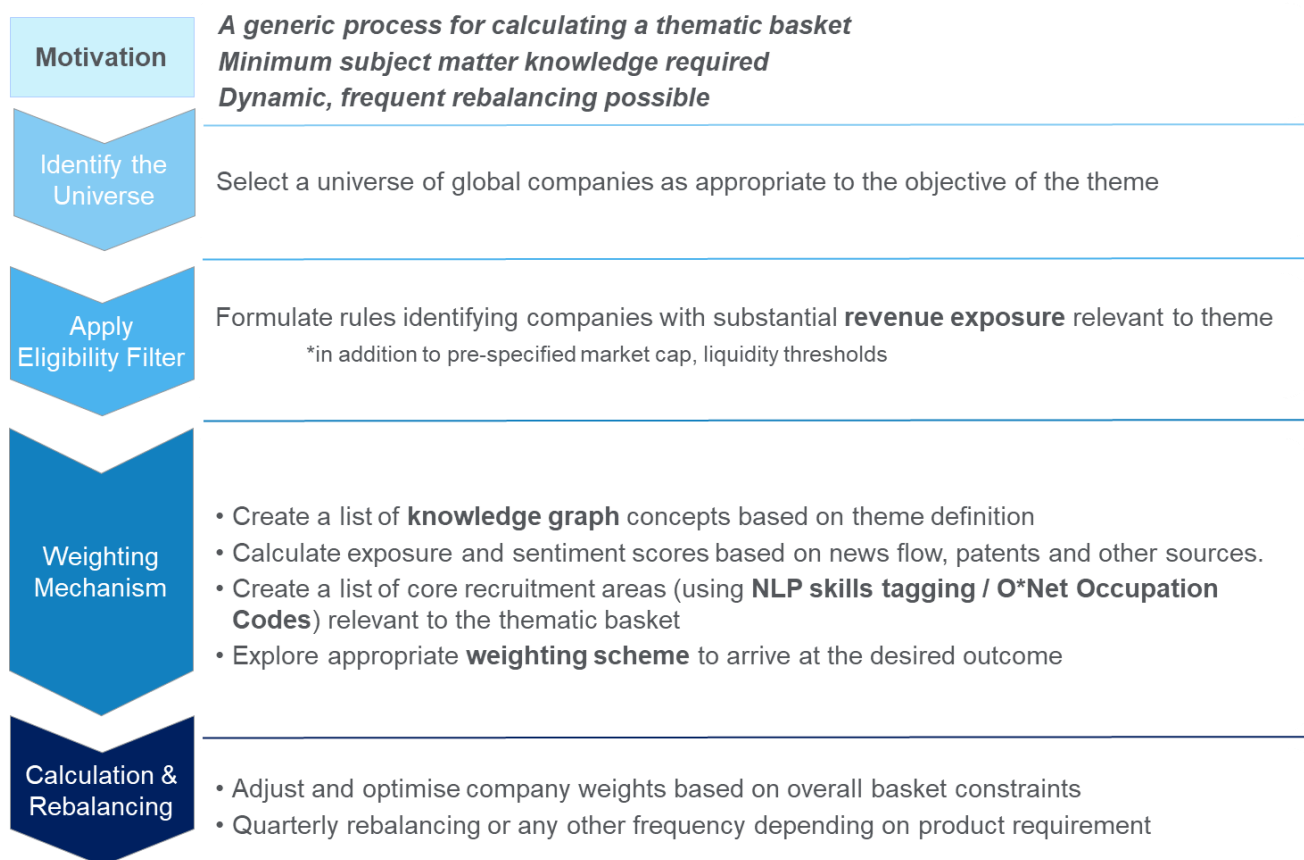
Source: Citi Global Data Insights, Citi Research, Yewno, Factset, LinkUp

⁷ <https://www.linkup.com/>

Thematic Basket Incubator Pipeline

With alternative data sources needed for our process identified, we describe below in terms of how the process flows from start to finish to generate a thematic basket of our choice. Each data source discussed in the earlier section provides an important angle to our process. We believe that through combination of these sources and AI/machine learning techniques embedded in our process, our thematic basket incubator pipeline is robust and well-balanced.

Figure 4. Thematic Basket Incubator – Process Flow



Source: Citi Global Data Insights

In the following sections, we will use “Wearable Technology” as an example to demonstrate how our process works in terms of identifying companies which have exposures and quantifying their connections to this theme.

Wearable Technology Theme

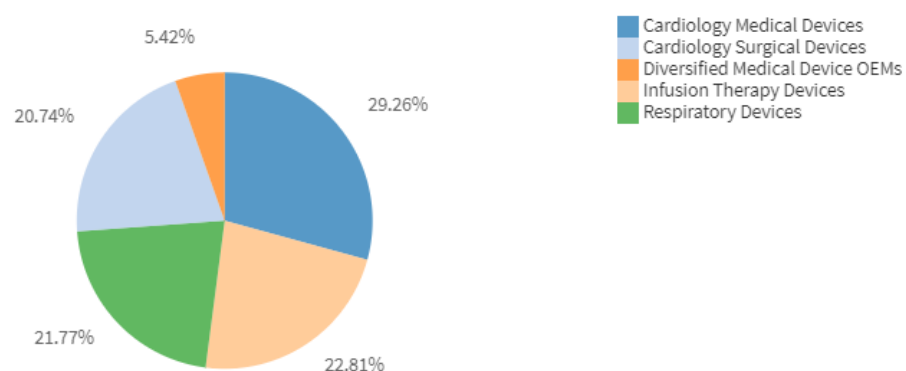
As previously described in our pipeline, an important step to create a thematic basket is to identify a set of companies which bear potential relevance to the theme of interest after choosing the relevant starting universe. In this section, we explore a number of ways that Factset RBICS data could be utilised in this process.

Revenue Eligibility Filter

As with any theme, we first select a small set of companies that bear significant relevance to the theme as a starting point, whether that is through Citi Research's GTM 'high exposure' constituents, public indices or a list of commonly known names to the theme. In the case of the Wearable Technology theme, we begin by studying in details the revenue breakdown of the highly exposed companies from GTM's Wearable Technology theme. This comprises 11 companies in total – with a noticeable distinction between health care and non-health care related companies. This is in line with the initial intuition of the wearable technology theme, where one dimension of theme points towards medical wearable devices, and the other dimension relates to general consumer wearable electronic devices.

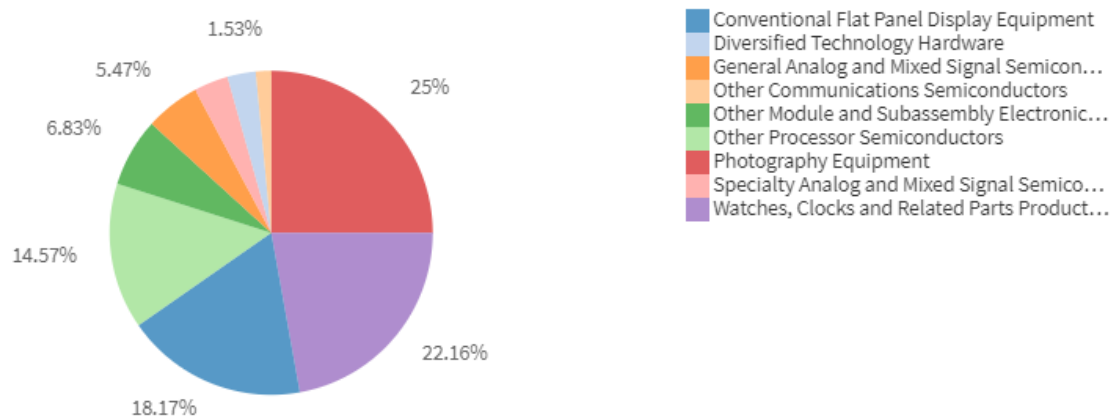
Consequently, in the following company revenue breakdown analysis we separate the 11 highly exposed companies into the health care and non-health care groups. Within both groups, we extract the list of RBICS L6 revenue sectors each company is exposed to, along with the revenue percentage attributed to each of these L6 revenue sectors. These L6 sectors and revenue percentages are then aggregated across the whole group to give the following results (shown in Figure 5 and Figure 6).

Figure 5. Wearable Technology Theme – Revenue Breakdown of Health Care Companies



Source: Citi Global Data Insights, Factset

Figure 6. Wearable Technology Theme – Revenue Breakdown of Non-Health Care Companies



Source: Citi Global Data Insights, Factset

In the Health Care group, we can see that over 90% of revenue on average are generated from either of the top 4 RBICS L6 sectors: *Cardiology Medical Devices*, *Cardiology Surgical Devices*, *Diversified Medical Device OEMs* and *Respiratory Devices*. Similarly for the non-Health Care group, over 75% of revenue on average are generated from either of: *Photography Equipment*, *Watches, Clocks and Related Parts Production*, *Conventional Flat Panel Display Equipment* and *Other Processor Semiconductors*.

There are two major considerations in utilising RBICS L6 revenue sectors to construct an eligibility filter for the thematic basket incubator pipeline. Firstly, these revenue sectors capture the fundamental characteristics of companies that are deemed to have high exposures to the Wearable Technology theme by our analysts based on the definition in the GTM. However, not all companies with such revenue profiles would be relevant and qualify to be included in our final thematic basket output. This issue could be alleviated through the downstream processes in our pipeline. Effectively, the revenue eligibility filter is designed to screen out irrelevant companies with incompatible fundamental profiles, rather than to pinpoint precisely relevant companies to a theme.

Secondly, this list of RBICS L6 revenue sectors are not necessarily comprehensive either. In fact, we have identified through further analysis that among the companies in the Wearable Technology GTM with “medium” and “low” exposures, there are additional business products and revenue segments related to the Wearable Technology theme that are mapped to RBICS L6 revenue sectors, which are not included in the list derived from companies with “high” exposure above. The list of additional RBICS L6 revenue sectors is as follows: *Wearable Technology*, *Diversified Electronic Components*, *Industrial Robots and Robotic Assembly Line makers*, *Other Peripherals* and *RF Analog and Mixed Signal Semiconductors*.

Figure 7. RBICS L6 Revenue Sectors Identified from Medium/Low Exposure GTM Companies

entity_type	exposure	bus_seg_name	I6_name	I6_desc
PUB	Low	Wearable and Industrial Products Segment - Inter-segment revenue	Wearable Technology	Wearable Technology
PUB	Low	Wearable and Industrial Products Segment - Wearable products	Wearable Technology	Wearable Technology
PUB	Low	Wearable and Industrial Products Segment - Micro-devices, Other	Diversified Electronic Components	Electronic components that serve as the building blocks for all electronic systems; including semiconductors
PUB	Low	Wearable and Industrial Products Segment - Robotics solutions	Industrial Robots and Robotic Assembly Line Makers	Robots and related equipment (e.g pallet exchanger) used in a variety of industrial applications, such as welding, painting, assembly, pick and place, packaging and palletizing, product inspection, testing, etc.
PUB	Med	Audio & Wearables	Other Peripherals	The Peripherals subsector is comprised of computer hardware devices that are connected to a computer and are designed to allow data input, display data output, or to enhance the functionality of computer systems.
PUB	Med	Wearables, Home and Accessories	Wearable Technology	Wearable Technology
PUB	Low	Wireless components - Wearables	RF Analog and Mixed Signal Semiconductors	Integrated circuits specialized for radio frequency applications.

Source: Citi Global Data Insights, Factset

Taken altogether, we have collected 13 RBICS L6 revenue sectors in total for the Wearable Technology theme. In the revenue eligibility filtering process, we apply a threshold so that any company 1) covered by the RBICS data 2) has revenue exposure in these L6 revenue sectors, and above the specified threshold would be considered relevant to the theme.

The first group of RBICS L6 revenue sectors are derived from the GTM Wearable Theme Health Care companies with “high” exposure – a 10% threshold is applied for these revenue sectors.

Eligibility Criteria 1. Any company with at least 10% revenue exposure to the following RBICS L6 revenue sectors:

- *Cardiology Medical Devices*
- *Cardiology Surgical Devices*
- *Diversified Medical Device OEMs*
- *Respiratory Devices*

The second group of RBICS L6 revenue sectors were derived from the GTM Wearable Theme non-Health Care companies with “high” exposure – a 10% threshold is also applied for these revenue sectors.

Eligibility Criteria 2. Any company with at least 10% revenue exposure to the following RBICS L6 revenue sectors:

- *Photography Equipment*
- *Watches, Clocks and Related Parts Production*
- *Conventional Flat Panel Display Equipment*
- *Other Processor Semiconductors*

The third group of RBICS L6 revenue sectors were derived from a relevant subset of the GTM Wearable Theme companies with “medium” or “low” exposure. Given the material relevance (by definition) of the *Wearable Technology* RBICS L6 revenue sector, a lower threshold of 5% is applied. The remaining four RBICS L6 revenue sectors are given a 10% threshold.

Eligibility Criteria 3. Any company with at least 5% revenue exposure to the following RBICS L6 revenue sector:

- *Wearable Technology*

Eligibility Criteria 4. Any company with at least 10% revenue exposure to the following RBICS L6 revenue sectors:

- *Diversified Electronic Components*
- *Industrial Robots and Robotic Assembly Line makers*
- *Other Peripherals*
- *RF Analog and Mixed Signal Semiconductors*

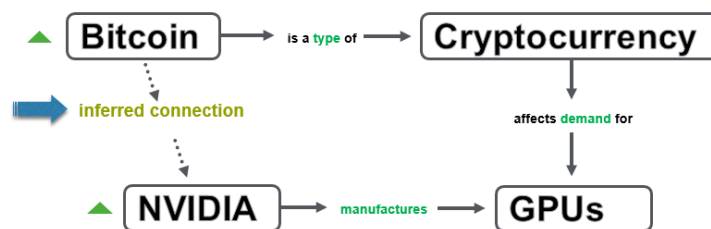
In summary, any company passing either of Eligibility Criteria 1-4 are deemed to have ‘passed’ the revenue eligibility filter, and would progress to the next stage in our thematic basket incubator pipeline.

Knowledge Graph – Patents and News Exposure

In order to capture companies with exposures to a given theme at scale, we have partnered with Yewno which utilises knowledge graph technology to construct a data framework for extracting knowledge.

A knowledge graph is a collection of relationships between entities that uses a network representation to create a graph-structured data model. This representation can be generated in a deterministic fashion, such as the case of a data ontology, or statistically, where the connections are inferred. The graph is a network of nodes connected by edges. In the YewNo graph, nodes are “concepts” – an atomic unit of knowledge such as a company, a person, a technology, among others – and the edges are the inferred levels of association based on the textual database. Figure 8 illustrates how concepts of “Bitcoin” and “NVIDIA” are associated through a statistically inferred connection derived from the knowledge graph.

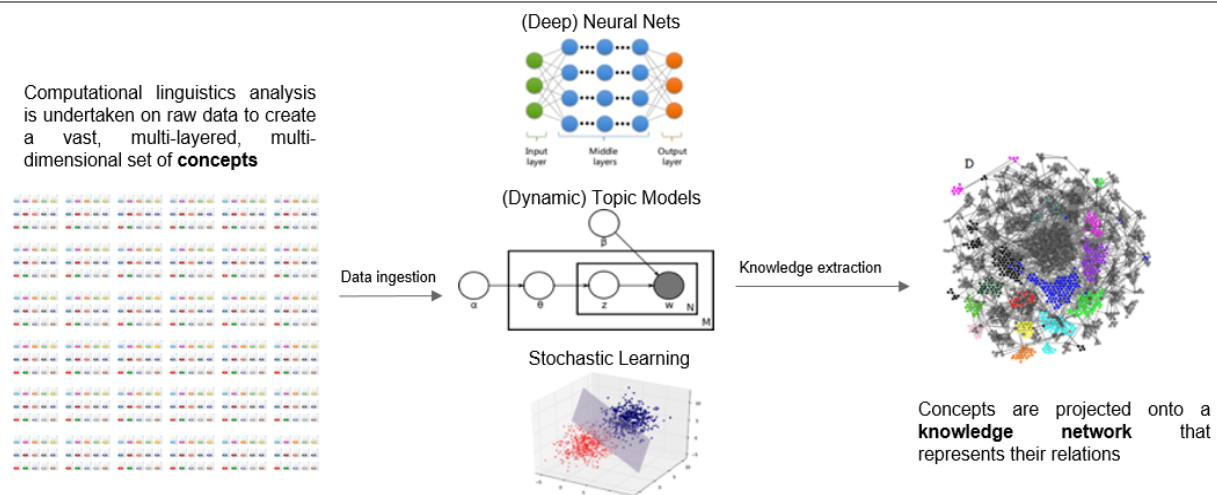
Figure 8. Knowledge Graph Illustration



Source: Yewno

Nodes in the network are represented by concepts extracted from document snippets and edges between nodes are computed based on strong Subject-Verb-Object (SVO) connections calculated over the text sources processed.

Figure 9. Knowledge Graph Illustration



Source: Yewno

The construction of the knowledge graph is done separately for each data source including newsfeeds, corporate filing and patents etc, and a graph-like network is induced for each. How it works in practice is that content gets ingested into the Yewno pipeline, then their concept extraction pipeline split raw data into text snippets, followed by concept extraction being performed on each snippet and finally computing different metrics for each extracted concept.

Utilising knowledge graph technology has the advantage of enriching the search results through a network of 'concepts' on the graph rather than just keywords. In the thematic exposure context, it offers flexibility and transparency by anchoring around concepts connected to a given theme based on text snippets and their edges on the graph. The inferred connections from the knowledge graph are what is missing in the traditional textual analysis where it often relies on a bag of words or terms.

The output from Yewno's knowledge graph consists of five scores which measures a concept's exposure to a given theme (target concept) through various angles. Yewno's definitions of these scores are as follows –

- **Contribution Score** is a measure of how much the target concept was co-mentioned with the source concept in the news, or published documents relative to all the mentions of the concept source. The Contribution ratio aims to measure the contribution of the target concept to the source concept.
- **Pure Play Score** is a measure of the percentage of co-mentions between the target concept and the source concept relative to all mentions of the target concept. The Pureplay ratio aims to measure the concentration of the target concept in the source concept.
- **Centrality Score** is based on centrality diffusion of the network constructed from mentions between concepts. Incorporates second order connections that favour central nodes connected to other central nodes.
- **Similarity score** is based on how close are the companies and concepts projections in the semantic space.
- **Aggregated Score** is a weighted linear combination of the aforementioned scores normalised by the maximum.

In addition, Yewno provides concept-to-concept sentiment which is derived by measuring the polarity of how directly and indirectly connected concepts impact each other, positive or negative and to what extent – this facilitates our negative screening of names to a given theme. Contribution and Pure Play scores are importance scores which are based on the number of co-occurrences between concepts. For example, using the knowledge graph based on news, if our target concept is 'Tesla' and the source concept is 'Electric Vehicle', then the Pure Play score is the percentage of instances where “Electric Vehicle” is mentioned amongst all “Tesla” related news articles. The Contribution Score is the percentage of times where “Tesla” is mentioned in all “Electric Vehicle” articles.

For our Wearable Technology theme, we have utilised both knowledge graphs based on news as well as patents from Yewno, as intuitively patents data is highly relevant for technology themes. The news knowledge graph is derived by processing over 350,000 articles each month from a curated list of approximately 300 global websites that are in English. The patent knowledge graph is constructed by processing millions of patents from global patent websites such as the World Intellectual Property Organisation (WIPO) and United States Patents and Trademark Office (USPTO) which receive 20.5K new patents per week.

To construct a company thematic news or patents exposure signal, we proceed as follows: first, we start with the theme named concept, e.g. 'Wearable Technology' and search our Knowledge Graph for related concepts to create a rich description of the topic. For this step, the platform uses pairwise similarity scores between concepts uses an ensemble of two metrics: 1) BERT Language Model – to capture static semantic relationships between concepts definitions; 2) News Concept Exposure – to incorporate dynamics of similarity metrics between concepts based on media attention and graph embedding. The result is a taxonomy of concepts representing the theme.

Figure 10. Wearable Technology Example Concepts Based on Yewno's News & Patents Knowledge Graphs

Wearable Health Care Concepts	Wearable General (non-Health Care) Concepts
Activity tracker Blood glucose monitoring Glucose meter Health informatics Heart rate monitor Internet of things Pedometer Telemedicine Virtual reality therapy Wearable computer Wearable technology Hearing aid Digital hearing aid Bone-anchored hearing aid	Activity tracker Android Operating System Apple Watch Google Assistant Internet of things Web of Things Smartwatch Virtual reality headset Wearable computer Wearable technology Portable audio player Portable media player Portable DVD player 5G

Source: Citi Global Data Insights, Yewno

Figure 10 shows some of the concepts from Yewno's news and patents knowledge graphs related to the Wearable Technology theme, splitting into two sub-groups "health care" concepts and "non-health care", consistent with our revenue eligibility filters discussed in the earlier section. From these concepts, we can further extract a list of companies (both public and private), their corresponding scores which measure the strength of their connections to the concepts and sentiment (both positive and negative).

Principal Components Analysis on Concept Exposure Scores

As we go through the iterative process of building a ‘concept bank’ which characterises various aspects of the central theme, the list of relevant concepts continues to grow and helps us more comprehensively articulate important elements of a theme. On the other hand, we inadvertently face the issues of high dimensionality and potentially highly correlated concept exposure scores. This needs to be carefully addressed through a rigorous statistical approach, with the goal to identify various aspects of a theme.

Principal Components Analysis (PCA) is a very popular statistical technique that can be useful in our case. Intuitively, PCA takes the concept list as input, groups correlated concepts together and produces orthogonal principal components (PCs). These PCs are essentially weighted sums of the underlying concept exposure scores adjusting for their correlations, so that each PC spans in different dimensions that would explain the biggest proportion of variance in the data, thereby capturing the most important elements of a theme.

Figure 11. Wearable Technology Example Concepts Based on Yewno’s News & Patents Knowledge Graphs

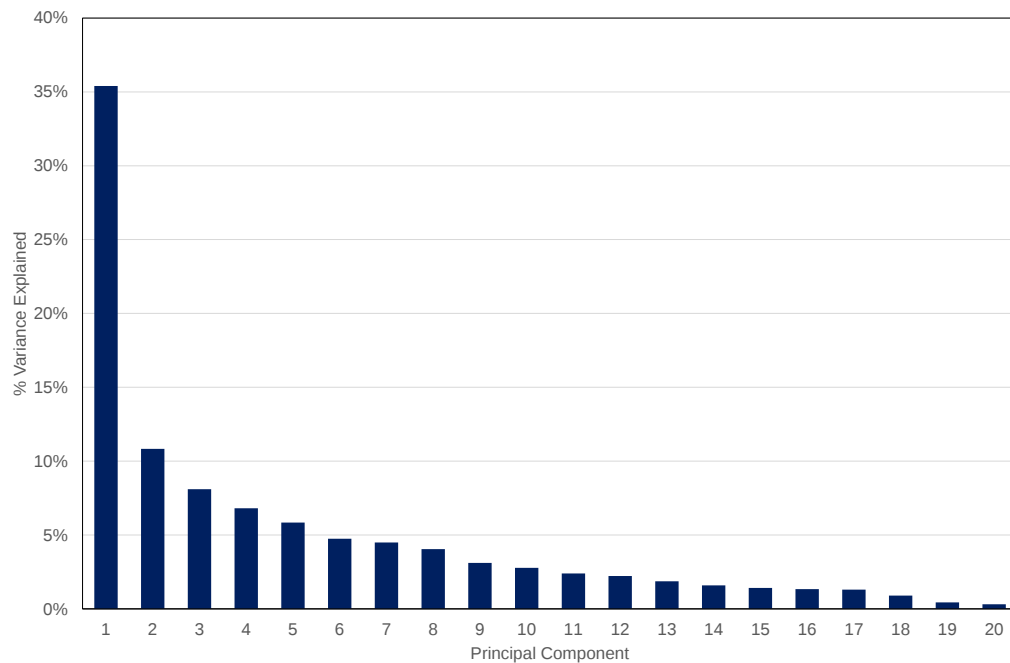
Concept correlation	5G	Activity tracker	Android Operating System	Apple Watch	Blood glucose monitoring	Glucose meter	Google Assistant	Health informatics	Hearing aid	Heart rate monitor	Internet of things	Pedometer	Smartwatch	Telemedicine	Virtual reality headset	Virtual reality therapy	WatchOS	Wearable computer	Wearable technology	Web of Things
5G	100%	37%	62%	36%	0%	1%	43%	14%	5%	5%	64%	9%	50%	12%	37%	17%	30%	33%	13%	0%
Activity tracker	37%	100%	55%	72%	5%	20%	66%	20%	29%	61%	33%	49%	73%	25%	47%	19%	54%	46%	25%	0%
Android Operating System	62%	55%	100%	70%	0%	13%	78%	26%	23%	28%	40%	24%	80%	9%	66%	14%	65%	49%	17%	0%
Apple Watch	36%	72%	70%	100%	15%	27%	73%	24%	48%	57%	19%	46%	81%	19%	65%	19%	86%	65%	28%	0%
Blood glucose monitoring	0%	5%	0%	15%	100%	49%	0%	27%	35%	11%	0%	27%	6%	32%	0%	7%	9%	16%	14%	9%
Glucose meter	1%	20%	13%	27%	49%	100%	18%	19%	29%	20%	4%	25%	22%	21%	14%	12%	24%	26%	19%	0%
Google Assistant	43%	66%	78%	73%	0%	18%	100%	25%	32%	42%	28%	33%	67%	16%	56%	18%	61%	42%	17%	0%
Health informatics	14%	20%	26%	24%	27%	19%	25%	100%	24%	21%	14%	23%	29%	47%	14%	23%	9%	16%	6%	5%
Hearing aid	5%	29%	23%	48%	35%	29%	32%	24%	100%	24%	2%	28%	36%	26%	31%	20%	46%	41%	22%	14%
Heart rate monitor	5%	61%	28%	57%	11%	20%	42%	21%	24%	100%	3%	56%	50%	22%	26%	17%	32%	26%	15%	4%
Internet of things	64%	33%	40%	19%	0%	4%	28%	14%	2%	3%	100%	8%	35%	23%	25%	21%	13%	25%	8%	7%
Pedometer	9%	49%	24%	46%	27%	25%	33%	23%	28%	56%	8%	100%	49%	15%	16%	17%	36%	25%	9%	21%
Smartwatch	50%	73%	80%	81%	6%	22%	67%	29%	36%	50%	35%	49%	100%	11%	66%	26%	70%	53%	19%	0%
Telemedicine	12%	25%	9%	19%	32%	21%	16%	47%	26%	22%	23%	15%	11%	100%	0%	16%	2%	19%	9%	5%
Virtual reality headset	37%	47%	66%	65%	0%	14%	56%	14%	31%	26%	25%	16%	66%	0%	100%	39%	64%	56%	24%	0%
Virtual reality therapy	17%	19%	14%	19%	7%	12%	18%	23%	20%	17%	21%	17%	26%	16%	39%	100%	18%	17%	9%	20%
WatchOS	30%	54%	65%	86%	9%	24%	61%	9%	46%	32%	13%	36%	70%	2%	64%	18%	100%	64%	36%	0%
Wearable computer	33%	46%	49%	65%	16%	26%	42%	16%	41%	26%	25%	36%	53%	19%	56%	17%	64%	100%	43%	0%
Wearable technology	13%	25%	17%	28%	14%	19%	17%	6%	22%	15%	8%	9%	19%	9%	24%	9%	36%	43%	100%	0%
Web of Things	0%	0%	0%	0%	9%	0%	0%	5%	14%	4%	7%	21%	0%	5%	0%	20%	0%	0%	0%	100%

Source: Citi Global Data Insights, Yewno

The graph above illustrates correlation phenomenon across concept scores. This is expected, for example, between the concepts “WatchOS”, “Smartwatch” and “Apple Watch” – these are different enough to be defined as two concepts but naturally have a high co-occurrence in news, and hence the highly correlated concept exposure.

As we apply PCA on the company concept exposure score data, we define the cut-off for the optimal number of principal components to use by inspecting the scree plot, which shows the percentage of variance explained by each PC. In the plot overleaf, we see that almost 50% of the total variance in the data are explained by the first two PCs alone. The largest jump in the percentage variance explained is also observed in the first two PCs, meaning that the amount of additional variance explained by adding more PCs is diminishing after PC2. Both evidences suggest that we can use PC1 and PC2 as a basis to reduce the concept exposure dimensionality.

Figure 12. Percentage of Variance Explained by Each Principal Component



Source: Citi Global Data Insights, Yewno

Next, we inspect the top concepts and their loadings within the first two PCs. In Figure 13, PC1 contains high weights for smartwatch related concepts, such as Apple Watch, Android Operating System, WatchOS and Activity tracker. These relate to the consumer-centric wearable devices designed for convenience in daily life, fitness tracking and easy access to information. This PC would likely have a heavier focus on the technology sector. PC2 features heavier weights on medical related concepts that are more specialised and technical, such as blood glucose monitoring, glucose meter and telemedicine. These are the wearable devices designed for specific medical usage and monitoring purposes, which naturally tilt towards the two principal components which encapsulate the majority of concepts for the theme involving the health care sector.

Figure 13. Wearable Technology Example Concepts Based on Yewno's News & Patents Knowledge Graphs

Concept	PC1	PC2
Apple Watch	0.343	-0.009
Smartwatch	0.334	-0.106
Android Operating System	0.302	-0.248
WatchOS	0.302	-0.072
Activity tracker	0.298	-0.021
Google Assistant	0.298	-0.131
Virtual reality headset	0.273	-0.183
Wearable computer	0.263	0.004
5G	0.197	-0.293
Internet of things	0.147	-0.197
Wearable technology	0.134	0.070
Virtual reality therapy	0.127	0.076
Web of Things	0.022	0.145
Heart rate monitor	0.212	0.175
Pedometer	0.197	0.249
Hearing aid	0.191	0.285
Health informatics	0.135	0.267
Telemedicine	0.105	0.338
Glucose meter	0.128	0.361
Blood glucose monitoring	0.075	0.474

Source: Citi Global Data Insights, Yewno

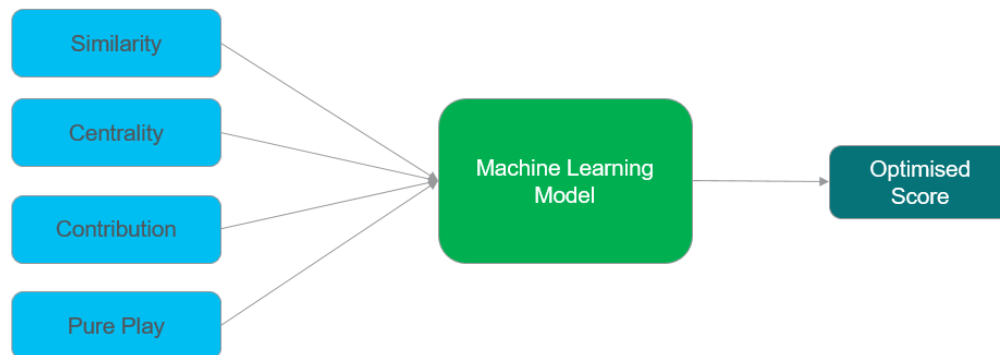
In summary, through the use of PCA we have identified two essential components of the Wearable Technology theme – smartwatches for lifestyle-conscious consumers, and medical wearable devices for monitoring purposes. This helps overcome the problem of high dimensionality by coupling correlated concepts into two primary dimensions, and serves as a simplified input to our thematic basket incubator workflow measuring the relevance of companies to the theme through their news exposure.

Thematic Classification Model: In Search of an Enhanced Aggregate Score

The scores Yewno provide are very useful as they measure to what extent a company is related to a theme from different aspects of the connections. However, these scores are not normalised which makes a straight aggregation difficult. Yewno uses their respective maximums to scale each score category accordingly and combines them linearly.

Since we have a very rich thematic mapping dataset from the GTM with long history, we employ the GTM to test different supervised machine learning models to optimise the weighting scheme to be applied to the knowledge graph scores. It is important to emphasise that this classification model is designed to enhance the weighting scheme of the four Yewno scores based on a wide range of theme families in the GTM and the core characteristics of the analysts' mappings. Once the best machine learning model is determined, the weighting scheme can then be deployed to any theme regardless of whether or not that theme is in the GTM.

Figure 14. Thematic Classification Model on Yewno Scores



Source: Citi Global Data Insights, Yewno

The specific steps we have taken are as the following –

Data Preparation

Exposure signals are scaled for a specific concept separately to a normal distribution (using the Yeo-Johnson transformation). We then take the average across all concepts within the desired time window (e.g. monthly) to estimate the theme-level exposure of a company to the theme within our time window of interest.

Predictions

The model uses the aggregated version of the news-sourced exposure signals as features from the data preparation stage. It is trained on a set of 43 GTM Themes⁸ with a label of 1 if the company is ranked is either “Low”, “Medium” or “High” exposure to theme and 0 otherwise. The trained model is then used for predictions across any given theme.

We explore various classification models which would give us the best outcome as measured in their precision scores. As can be seen in Figure 15, the XGBoost model of combining the four scores from Yewno has the highest precision score of 65.6% based on the training set of GTM themes. However, it also come with relatively high variance compared to the other models indicated by the 10 fold cross validation process. Consequently, we have chosen the Random Forest model which has achieved a respectable precision level with reasonably tight variance.

⁸ Out of 87 themes in total, the chosen themes have the longest history which is important for machine learning models.

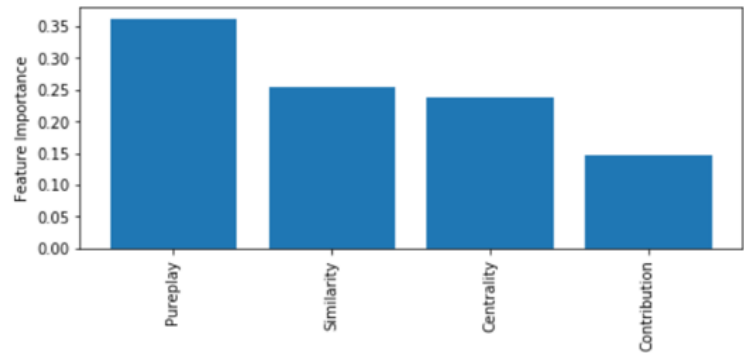
Figure 15. Thematic Classification Model on Yewno Scores

Model	Precision score (10 fold cross validation)
Random Forest	0.504 (± 0.078)
XGBoost	0.656 (± 0.578)
Logistic Regression	0.476 (± 0.090)

Source: Citi Global Data Insights, Yewno

Figure 16 shows the feature importance of the scores in the best performing model, which constitutes to the optimal weighting scheme of the scores. Pure Play score appears to be the most important feature, followed by Similarity, Centrality and Contribution scores.

Figure 16. Thematic Classification Model on Yewno Scores



Source: Citi Global Data Insights, Yewno

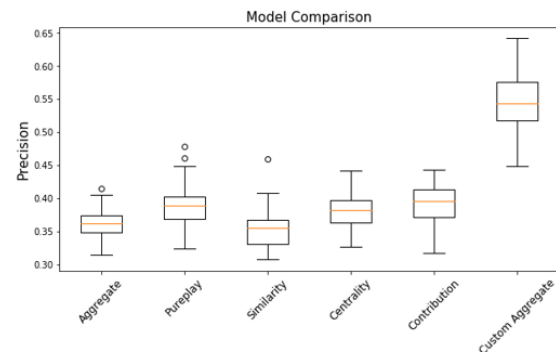
Equipped with the optimal weighting scheme identified by our classification models, we can then compare the realised precisions based on Yewno four scores, the aggregate score vs our optimised or custom score. As depicted in Figure 17, our custom aggregate score performs substantially better, improving precision by 50%. This insight helps us shape the weighting scheme applied to our final outputs from our thematic basket process.

Figure 17. Custom Aggregate Score vs Yewno Scores

Compare Models

- **Custom Aggregate:** Model based on the 4 Scores (Similarity, Centrality, Pureplay, Contribution)
- **Other:** Model based on single scores (e.g. Aggregate is based on Aggregate score only)

	Precision	Precision Pct. Chg.
Aggregate	0.36	
Custom Aggregate	0.54	50%



Source: Citi Global Data Insights, Yewno

Hiring Exposure – Occupation Group Profiling

In the previous section we have illustrated the importance of measuring a company's relevance to a theme through news and patents data and established a framework to score and rank companies using a knowledge graph approach. In this section, we will look into job postings data, which is another crucial method to capture a company's relevance to a theme through measuring how aligned its recruitment strategy and what its distribution of talent demand looks like.

The idea of studying the occupation groupings of companies highly exposed to the Wearable Technology theme is to identify a set of common occupational groups of which these companies are exposed to, and use them as a proxy to measure how relevant other companies passing the RBICS revenue eligibility filter are to the theme. The underlying assumption is that investing in talent is the first step to develop a product or business, and the companies that are highly exposed to the theme should see hiring activities in similar occupational groups that are related to the theme.

Similar to the RBICS revenue eligibility analysis, we will break the occupation group profiling analysis by Health Care and non-Health Care related companies. Below we list the top 10 O*Net broad occupations for both groups.

Figure 18. Wearable Technology: Top 10 Occupation Group Profiles

Health Care Companies		Non-Health Care Companies	
ONet Occupation Group L3	% Jobs	ONet Occupation Group L3	% Jobs
Industrial Engineers, Including Health and Safety	10.18%	Software and Web Developers, Programmers, and Testers	22.06%
Sales Representatives, Wholesale and Manufacturing	10.00%	Marketing and Sales Managers	10.67%
Marketing and Sales Managers	8.25%	Industrial Engineers, Including Health and Safety	6.46%
Natural Sciences Managers	4.77%	Market Research Analysts and Marketing Specialists	5.41%
Software and Web Developers, Programmers, and Testers	3.81%	Mechanical Engineers	3.72%
Medical Scientists	2.44%	Electrical and Electronics Engineers	2.99%
Compliance Officers	2.42%	Miscellaneous Computer Occupations	2.90%
First-Line Supervisors of Production and Operating Workers	2.27%	Designers	2.71%
Computer and Information Analysts	2.19%	Human Resources Workers	2.26%
Registered Nurses	2.09%	Computer Hardware Engineers	2.01%

Source: Citi Global Data Insights, LinkUp

For Health Care companies, apart from Compliance Officers and Registered Nurses, all other occupations could be considered relevant to the Wearable Technology theme, which in total accounts for just over 45% of the jobs. For Non-Health Care companies, apart from Human Resources Workers, all other occupations are relevant and the total is close to 60% of the jobs.

Hiring as a Forward-Looking Measure

So far we have identified the following O*Net broad occupations from the highly exposed companies in the GTM Wearable Technology theme which we deem are relevant to the theme:

- Industrial Engineers, Including Health and Safety
- Sales Representatives, Wholesale and Manufacturing
- Marketing and Sales Managers
- Natural Science Managers
- Software and Web Developers, Programmers, and Testers
- Medical Scientists
- Compliance Officers
- First-Line Supervisors of Production and Operation Workers
- Computer and Information Analysts
- Marketing Research Analysts and Marketing Specialists

- Mechanical Engineers
- Electrical and Electronics Engineers
- Miscellaneous Computer Occupations
- Designers
- Computer Hardware Engineers

For any company passing the revenue eligibility filter, we calculate the percentage of job postings that fall under the above list and transform that to a score. This score will be combined with the news and patents exposure scores to create a final ranking for each company. Examining the hiring activities in relevant talent is a meaningful method to capture a company's relevance to the theme, but not the only determinant.

It is often useful to tie the hiring surges in occupations relevant to the theme to the spike in news exposure data, as a way to cross-validate whether companies are putting their words (according to news announcements) into actions. Hiring data also serves as an alternative way for one to identify up-and-coming companies in the theme that has not necessarily been able to generate significant revenue streams from the relevant RBICS L6 revenue sectors yet, but has already made significant investments in hiring for creating the products or services related to the theme.

Bringing It All Together

In previous sections, we have presented and discussed our 3 core data sources: Factset RBICS, Yewno's knowledge graphs (based on newsfeeds and patents) and LinkUp's job postings data. From these data sources, we then construct the four building blocks of our thematic basket incubator, consisting of revenue eligibility filter, knowledge graph scores, aggregate knowledge graph score based on the findings of classification models trained on the GTM dataset and hiring exposure. What comes out of the process from these building blocks is a list of companies which have passed all the filters and set thresholds with significant exposures to a given theme based on knowledge graph connections. These companies could be public or private and could have positive or negative exposures to the theme concerned, depending on the specification of the theme to begin with.

Compared to the GTM Wearable Technology theme (which consists of 167 constituents with 11 being high, 27 being medium and the rest being low exposure names), our process has yielded over 5000 public-listed names and close to 900 private companies which have exposures to this theme. Filtering further based on revenues (as described on page 11-14) gives us about 150 public companies which are deemed to have high exposures, vs 11 in the GTM. This demonstrates the far-reaching nature of our process and being able to capture both public and private companies, our output is a solid incubator for identifying companies in a given theme. The list of names can then be sliced and diced based on one's specifications. Figure 19 shows a partial snapshot of the dashboard, which summarises the scores for the qualified companies with the snippets from the knowledge graph, providing transparency on the inclusions of these companies.

Figure 19. Wearable Technology Example Dashboard

Company	2018-08	2018-09	2018-10	2018-11	2018-12	2019-01	2019-02	2019-03	2019-04	2019-05	2019-06	2019-07	2019-08	2019-09	2019-10	2019-11	2019-12	2020-01
Company A	1.55	1.52	1.49	1.46	1.33	1.29	1.39	1.46	1.58	1.56	1.30	1.29	1.50	1.58	1.62	1.55	1.44	1.43
Company B	1.30	1.23	1.25	1.29	1.32	1.37	1.41	1.49	1.59	1.59	1.24	1.38	1.38	1.36	1.42	1.56	1.41	1.42
Company C	1.52	1.43	1.38	1.43	1.38	1.43	1.48	1.73	1.78	1.75	1.47	1.62	1.52	1.62	1.70	1.58	1.45	1.56
Company D	1.68	1.63	1.55	1.45	1.21	0.97	1.39	1.57	1.68	1.68	1.45	1.60	1.65	1.58	1.18	1.28	1.35	1.37
Company E	1.41	1.46	1.48	1.46	1.32	1.28	1.27	1.30	1.26	1.29	1.03	1.20	1.27	1.24	1.22	1.07	0.92	0.88
Company F	1.39	1.33	1.27	1.31	1.24	1.24	1.25	1.21	1.36	1.25	1.12	0.98	0.96	0.99	1.18	0.91	0.93	1.06
Company G	1.66	1.68	1.63	1.60	1.39	1.25	1.27	1.33	1.53	1.43	1.10	1.21	1.14	1.19	0.97	1.17	1.11	1.01
Company H	1.50	1.37	1.46	1.47	1.35	1.31	1.28	1.13	1.09	0.94	0.54	0.98	1.20	1.14	1.12	0.93	0.86	0.75
Company I	1.58	1.56	1.51	1.58	1.39	1.37	1.34	1.35	1.47	1.41	1.17	1.27	1.20	0.83	0.78	0.85	0.60	0.95
Company J	1.42	1.44	1.40	1.40	1.29	1.29	1.20	1.11	1.21	1.14	0.85	0.99	0.85	0.85	0.66	0.80	0.45	0.93
Company K	1.51	1.49	1.31	1.35	1.27	1.10	1.13	1.17	1.04	1.21	0.83	0.65	0.63	0.52	0.40	0.94	1.00	1.01
Company L	1.57	1.51	1.53	1.56	1.32	1.19	1.18	1.30	1.40	1.26	0.91	1.17	1.15	1.27	1.15	1.07	0.89	1.01

Concept	Aggregate	Similarity	Centrality	Pureplay	Contribution	Title	Company	Concept	Snippet	URL	m
Telemedicine	17,803.65	122,278.68	597.92	18	0	4-WAY HANDSHAKE OPTIMIZATION	XYZ	Internet of things	a desktop computer, a mobile computer, a laptop computer, an Ultrabook, a computer, a notebook computer, a tablet computer, a server computer, a handheld computer, a handheld device, an internet of things (IoT) device, a sensor device, a personal digital assistant (PDA) device, a handheld PDA device, an on-board device, an off-board device, a hybrid device (e.g. combining cellular phone functionalities with PDA device functionalities), a consumer device, a vehicular device, a non-vehicular device, a mobile or portable device, a non-mobile or non-portable device, a mobile phone, a cellular telephone, a personal communications service (PCS) device, a PDA device which incorporates a wireless communication device, a mobile or portable global positioning system (GPS) device, a digital video broadcasting (DVB) device, a relatively small computing device, a	http://aofu.usdo.gov/zhetao/zhetao-Parser2_Sect1=PT02&Sect2=H1IOFF&S=1&u=%2Fstahml%2Fsearcv.html&r=1&f=G&l=50&d=PG01&S1=20210345105&OS=20210345105&RS=20210345105	20
Web of Things ...	17,054.90	118,775.69	160.51	0	0						
Blood glucose ...	16,915.34	116,264.37	461.65	0	3						
Wearable techn...	16,810.86	116,097.73	510.63	0	0						
Glucose meter	16,402.28	112,886.25	442.00	10	0						
Pedometer	16,374.95	112,510.92	537.91	0	33						
Hearing aid	13,896.33	94,856.70	462.62	6	0						
Portable audio ...	12,612.91	86,703.42	338.75	0	32						
Health informat...	12,379.33	84,074.57	639.20	2	0						
Heart rate moni...	12,282.30	84,061.68	474.90	0	2						
Virtual reality h...	11,814.06	80,250.70	656.27	0	0						
Wearable comp...	10,958.83	72,679.76	479.45	0	20						
Virtual reality th...	10,858.36	75,803.67	53.62	0	0						
Digital hearing ...	10,651.77	74,117.41	100.45	0	0						
Activity tracker	10,499.63	71,061.63	561.60	0	17						
Portable media ...	10,459.10	70,917.16	396.51	0	44						

Source: Citi Global Data Insights, Yewno

Taking a step further, we explore the following strategies to examine their backtest performance – News, Patents, RBICS (in isolation), News & Patents, and News & Patents & RBICS. Note that we did not include hiring exposure in this part of the analysis as our focus is to look at how successful our pipeline is able to identify well-established public-listed names in the Wearable Technology theme, as we want to compare our basket to the existing GTM theme and MSCI World index. Each of the strategies tested is based on the signal of the relevant source(s). For the strategies where more than one source (e.g. News & Patents) is used, we scale the signal of each source to a 0-1 range using min-max scaling and combine them. We test two approaches of combining the signals for each case: 1) a simple average approach and 2) a multiplication approach between the combined signals to arrive at the final score.

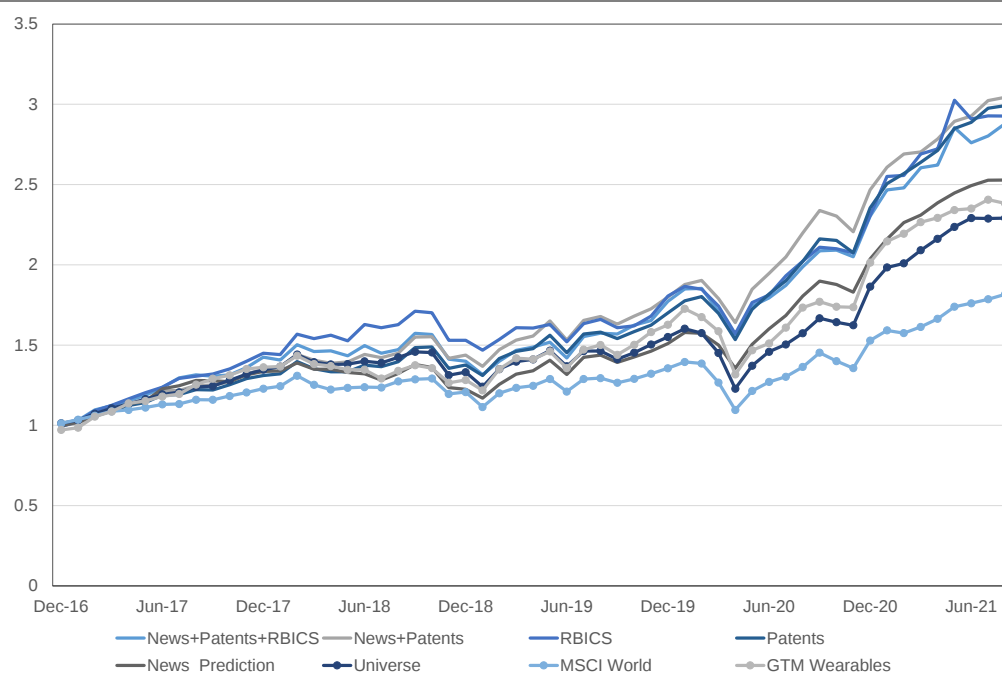
Thematic basket performance is then measured by performing a backtest which goes long the top 50 ranked companies every month with respect to the signal of each strategy. The weighting schemes used are 1) weighting proportionally to their signal strength, 2) equal weighting. As the signal combination with different weight schemes yield quite similar results empirically, we show the performance and stats for the multiplication approach weighted by signal strength below for simplicity's sake.

We benchmark these thematic baskets against MSCI World Index, GTM Wearable Technology basket and Yewno's stock universe. The universe is essentially going long (equally weighted) all companies that have been picked up by the data sources we are using for each strategy, i.e. union of companies picked up by news, patents and RBICS. For example for the news part (of the union), this would involve the companies where exposure signals have been picked up from the news sources during each month. These companies would be deemed to have exposures to our theme based on the newsfeeds or through their 2nd and 3rd order inferred relationships, such as Centrality scores or predicted values from our own classification model. In other words, this would be the broadest and the most rudimentary version of our thematic strategy.

Signal Efficacy in Wearable Technology

Compared to MSCI World, all strategies including the naïve Universe and the GTM Wearable Technology theme have outperformed since Jan 2017. Strategies that are based on knowledge graph exposures derived from news and patents data with or without RBICS have achieved the best backtest performance. While the return and volatility profiles are similar, those combined with RBICS as the eligibility filter have substantially lower turnover.

Figure 20. Wearable Technology Strategy Performance



Source: Citi Global Data Insights, Yewno

The stats table below summarises the signal backtest performance from our signals derived from the building blocks. The signal based on patents data has the highest Sharpe ratio amongst the others, but it also comes with relatively high turnover which in reality would incur significant trading costs. In contrast, the combined signal performance is similar to that of the patents data but with markedly lower turnover, especially when combined with signals from the RBICS dataset. Interestingly, all strategies barring “Universe” have beta lower than 1 and yet carried higher returns than MSCI World. Compared to the GTM Wearables theme, our combined signals also performed better both in terms of returns and risks, with lower drawdown.

Figure 21. Wearable Technology Signal Stats (Dec 2017 – Aug 2021)

	Annualised Return	Annualised Volatility	Sharpe Ratio	Beta*	Maximum Drawdown	Avg Turnover
News + Patents + RBICS	25%	15%	1.67	0.90	16%	9%
News + Patents	26%	16%	1.63	0.91	15%	31%
Patents	26%	15%	1.73	0.95	15%	43%
News	24%	16%	1.50	0.86	22%	55%
RBICS	25%	16%	1.56	0.90	15%	6%
Universe	19%	16%	1.19	1.07	23%	19%
GTM Wearables	20%	18%	1.11	1.05	24%	
MSCI World	13%	15%	0.87	1.00	21%	

* with respect to MSCI World

Source: Citi Global Data Insights, Yewno, Factset

The back-testing results show the potential benefits of combining signals from the selected building blocks of our thematic basket incubator process. In addition to being scalable and able to capture a much wider set of companies (thanks to news and

patents knowledge graphs) with fundamental underpinning from RBICS, the example shown in Figure 21 achieved respectable returns in terms of ex post performance across all metrics.

Conclusion

In this report, we have discussed how thematic investing demand has been changing in the recent years as investors seek themes that present good growth and return opportunities but do not have a scalable solution to identify companies in the theme of interests quickly.

To address such needs, we have developed our Thematic Basket Incubator pipeline which combines fundamental data and alternative data and extracts the thematic exposures through detailed revenue segmentations, machine learning models and knowledge graphs based on various sources. We have showcased the flexibility and transparency of our process as it can be modular and designed to investors' exact specifications.

Our approach also has the advantages of shorter turnaround time, incorporating both positive and negative exposures to a given theme and covering public and private companies – all of these are the clear distinction from traditional methods where substantial subject matter expertise is required to identify companies with desired theme exposures. This service is available through Citi Global Data Insights (CGDI) which offers customised solutions to clients' specific needs.

Appendix

