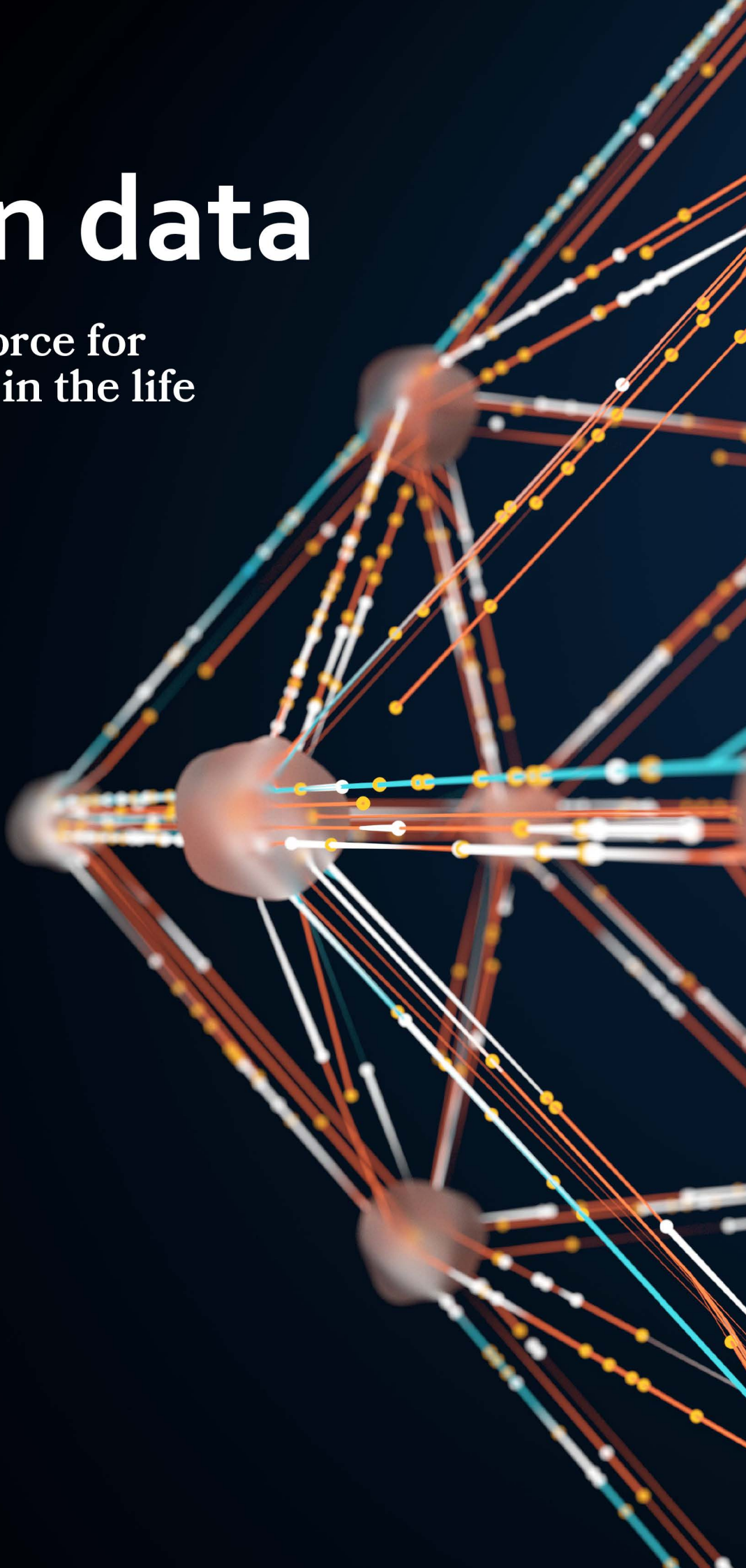
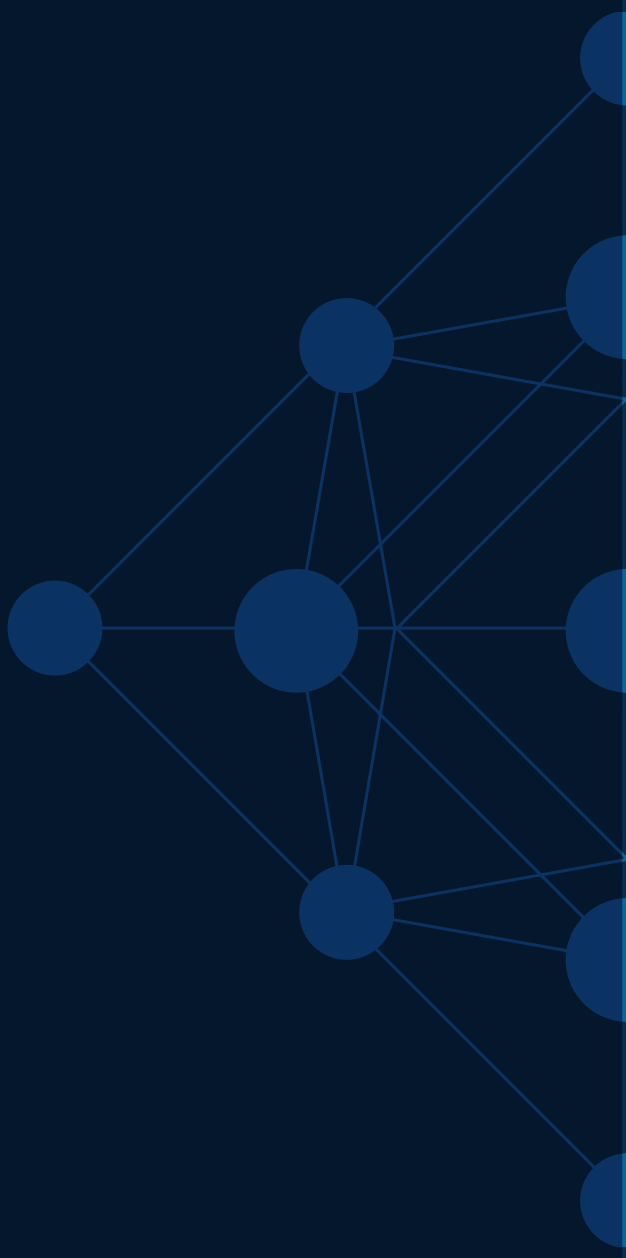
Open data

A driving force for innovation in the life sciences





Authors

ELIXIR

Dr Katharina B Lauer
Andrew Smith
Dr Niklas Blomberg
Xènia Pérez Sitjà
Callum Talbot-Cooper

Freie Universität Berlin

Prof Hannes Rothe
Daniel Sivizaca Conde

With a special thanks to the contributors from companies presented in this report and to the members of the ELIXIR Industry Advisory Committee.

For further details on this report or the ELIXIR Industry programme visit:

www.elixir-europe.org/industry

or contact Katharina Lauer:
katharina.lauer@elixir-europe.org

Contents

Abstract

'In the long history of humankind (and animal kind, too) those who learned to collaborate and improvise most effectively have prevailed.'

Charles Darwin

The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, 1859

Openness in the life sciences facilitates collaboration and increases innovation potential for industry

Undoubtedly, at one point in their career, every scientist will come across the work of Charles Darwin and his groundbreaking discoveries and theories. Already in 1859, he argued that collaborative use and reuse of existing resources enable novel outcomes. His work has been regularly used to explain success through innovation and entrepreneurship.

Two centuries later we have numerous examples that prove this is true – scientific discovery and groundbreaking products never appear on green fields in isolation, they are underpinned by knowledge generated by others[1].

But how can we foster these collaborative environments and what are the key components of success, particularly after years of digital transformation in the life science sector?

Our research has shown that access to open, free resources, such as data and software, is the underlying basis for breakthrough discoveries, scientific excellence and entrepreneurial endeavours.

Translating zeros and ones into real-world medicines, food or energy sources has become a cornerstone of economic activity, promising benefits for societies across the globe. To realise this potential, we need to understand the economic environment in the life sciences today and the actors therein, including entrepreneurs that bridge academia and industry, companies that use open resources and public infrastructure providers that engage in public-private partnerships.

1 Mariana Mazzucato. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*. (Penguin, 2018).

Open data enable innovation in the bioinformatics industry

Research data is a major asset for life sciences industry[2,3,4]. Exploiting these data drives development within health, agriculture and biotechnology sectors, thereby leading to economic and societal benefits. For example, to accelerate drug development, produce efficient washing powder, and even design better feed for livestock and pets.

Such innovative life sciences projects require the essential use of open data, open-source software and common standards. Together, open resources allow industry, academia and the wider community to create novel services and products.

For today's digital life sciences, data standards and data quality are major bottlenecks for data accessibility and data sustainability, thereby essential enablers of effective data utilisation. Research infrastructures, such as ELIXIR, ensure that life science innovators and researchers globally can utilise open data. They facilitate access to open resources such as databases, repositories, software tools, registries, compute and cloud resources, while promoting standards and enabling collaboration through training and networking events[5].

This report sheds light on the importance of continued access to open data and related resources, along with their underpinning infrastructures that contribute to a thriving life science ecosystem.

We showcase the growing European bioinformatics industry and its local environment, and demonstrate the importance of long-term investment in digital research infrastructures as bedrocks for innovation.

'The more freely available data that is compatible with other data, the better for us. Without open data resources, large companies, such as pharmaceuticals, would be hampered in their work but would not cease to exist. In contrast, small and medium-sized enterprises, as well as the startups, would be severely impacted as in many cases their business models are dependent on the availability of open data resources.'

Hans Garritzen
Sales Director, Medisapiens

'I think Europe, by investing in that framework [ELIXIR], is doing the right thing to support the long term sustainability of the Life Sciences knowledge economy. Imagine if all the data resources that are freely available today would disappear from tomorrow. It would be tough for the life sciences companies to sustain their innovation. They would have to work really, really hard and spend a large portion of their Research and Development (R&D) spending on replacing that common good.'

Abel Ureta-Vidal
Co-founder & Chief Data Officer at Sofi Health,
Investment director at CMS Ventures

'Open data is absolutely crucial for us. As soon as we want to explore new ways of finding molecules, because this is our business, we first need to have some proof of concept to make sure that we are heading in the right direction. For that, we are actually heavily reliant on open data.'

Florent Villiers
Lab Leader for Biochemistry & Data Science, Bayer

2 Beagrie, N. & Houghton, J. *The Value and Impact of the European Bioinformatics Institute Full Report.* http://www.beagrie.com/EBI-impact-report.pdfandhttp://www.beagrie.com/EBI-impact-summary.pdf (2016).

3 Jefferson, O. A. et al. Mapping the global influence of published research on industry and innovation. *Nat. Biotechnol.* **36**, 31–39 (2018).

4 Westbrook, J. D., Soskind, R., Hudson, B. P. & Burley, S. K. Impact of the Protein Data Bank on antineoplastic approvals. *Drug Discovery Today* vol. **25** 837–850 (2020).

5 Garcia, P. R. et al. *Public data resources as a business model for SMEs. The Role of Public Bioinformatics Infrastructure in supporting innovation in the life sciences [version 1; not peer reviewed]. F1000Research* vol. 7 https://doi.org/10.7490/f1000research.1115445.1 (2018)

Innovation ecosystems in bioinformatics

Ecosystems, such as Silicon Valley in California (USA), have been forerunners in supporting entrepreneurship and innovation. In Europe, attempts have been made to create and nurture ecosystems that attract cutting-edge firms. Several factors play a critical role in fostering social and technical innovation in such regions: high-quality education systems leading to large pools of graduates, well-funded research programmes, access to financial and human resources, and an open mindset towards innovation and entrepreneurship.

While research has provided extensive insights into the interrelationship between technological change and innovation in ecosystems, digital research infrastructures, such as ELIXIR, play a crucial role in amplifying the development of innovation ecosystems, particularly in the data-driven life science sector[6,7].

What is an innovation ecosystem?

Innovation ecosystems are complex, localised networks of individuals and organisations that facilitate an adequate flow of knowledge, talent and financial resources to create the basis for sustained value co-creation[8,9]. For knowledge-based industries, ecosystems are a prerequisite in allowing researchers and entrepreneurs to transform scientific insights from research publications into products and services that create value for the society.

Many thriving ecosystems harbour both innovative technology companies and customers of those companies. Our survey showed that bioinformatics Small and Medium Enterprises (SMEs) rarely sell their products directly to individual customers (13%). The majority of companies rely on large corporates (83%), other SMEs (63%), universities (46%) or clinics and hospitals (48%) to sell their services and products.

Proximity to customers and resources is an important factor for many growing companies to allow innovators to start and run a company from anywhere in the world. This is true even for open data resources and digital infrastructures, including cloud storage and federated data processing capabilities. In our survey, companies stated that half of their staff were recruited in the

How does an innovation ecosystem support bioinformatics companies?

Providing easy access to:

- Highly qualified personnel through bioinformatics institutes
  e.g. European Bioinformatics Institute (EMBL-EBI), the Barcelona Supercomputing Centre, the Swiss Institute of Bioinformatics (SIB)
- Customers
  e.g. pharma, biotech industry
- Facilities and collaborators
  e.g. biobanks, labs, research groups
- Companies that provide services to the company
  e.g. lawyers, media experts
- Opportunities to showcase their work
  e.g. ELIXIR Bioinformatics Industry Forum
- Mentors
  e.g. serial entrepreneurs, technology experts
- Support systems
  e.g. joint labs, incubators
- Funding from public and private investors
- Physical infrastructure
  e.g. transport, accommodation/offices, communications
- Digital infrastructure
  e.g. computing facilities, or data resources

6 Constantinides, P., Henfridsson, O. & Parker, G. G. Platforms and infrastructures in the digital age. *Inf. Syst. Res.* **29**, 381–400 (2018).
7 Adner, R. & Kapoor, R. Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. *Strateg. Manag. J.* **31**, 306–333 (2010).
8 Russell, M. G. & Smorodinskaya, N. V. Leveraging complexity for ecosystemic innovation. *Technol. Forecast. Soc. Change* **136**, 114–131 (2018).
9 Russell, M. G., Huhtamäki, J., Still, K., Rubens, N. & Basole, R. C. Relational capital for shared vision in innovation ecosystems. *Triple Helix* **2**, 1–36 (2015).

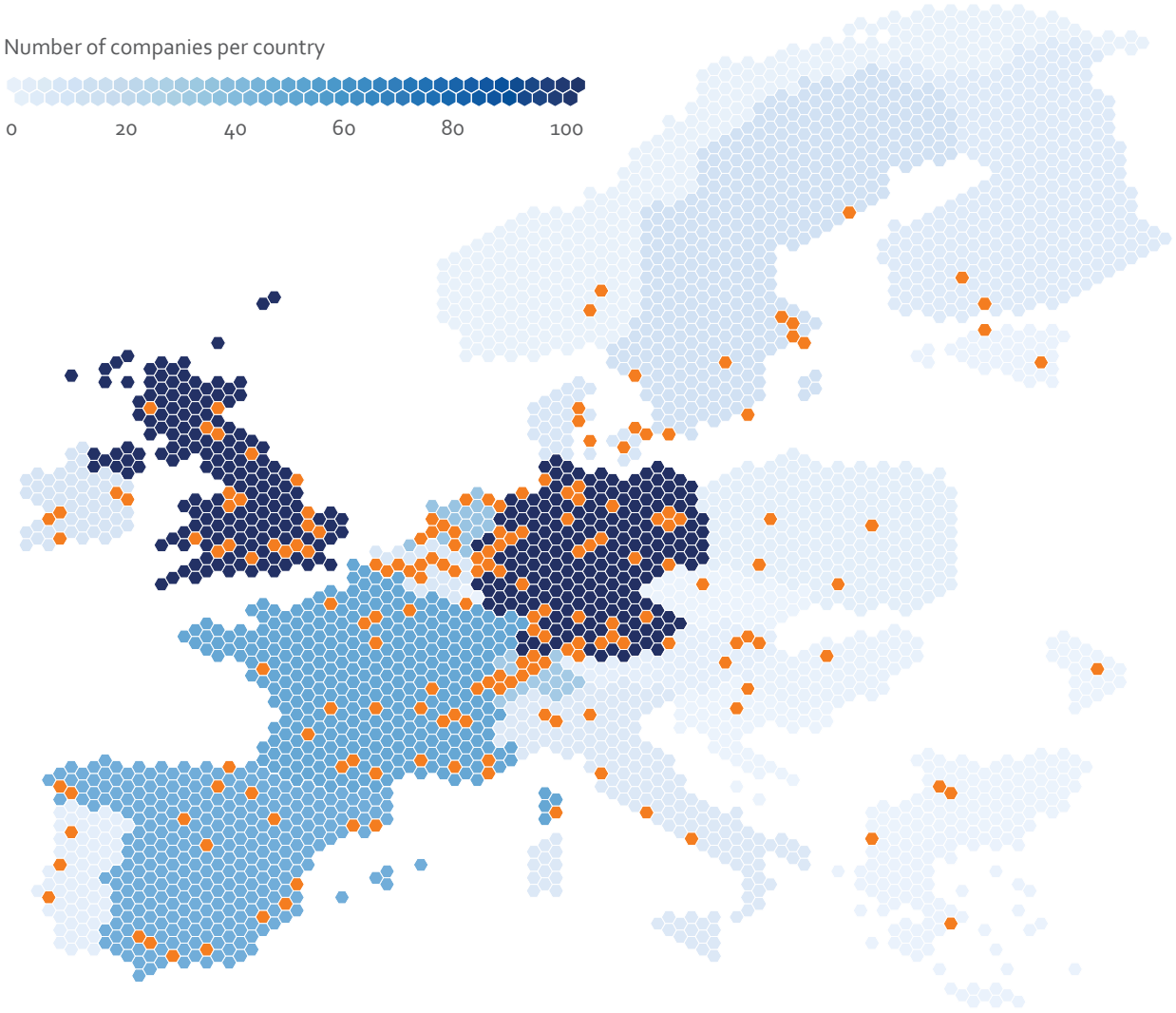


Figure 1. Number of bioinformatics-related start-ups and SMEs identified per country within Europe. Orange hexagons represent regions with high concentrations of companies (countries shown where data was available).

year, with over 75% having a science or technical degree – people and expertise matter.

Take Arcticzymes, located in Tromsø[10,11]. Arcticzymes extensively uses public databases to identify novel cold-adapted enzymes from arctic organisms, and formulates these for use in molecular research, *in vitro* diagnostics and therapeutics. Companies that base their business model on digital technologies and data can be based anywhere – but our research has shown that being close to an innovation ecosystem helps many to thrive and go further. Because of the knowledge-intensive nature of the industry, bioinformatics companies face a talent-acquisition challenge. They need to employ experts in specialised life science areas and, at the same time, bring in experts in cutting-edge technologies, such as machine learning and artificial intelligence. Access to people, personal networks, partners and collaborators that provide open resources remains an important factor for the concentration of bioinformatics ventures in certain regions (Figure 1).

In this report, we will explore the innovation ecosystems in Europe that harbour most bioinformatics SMEs.

10 Novel enzyme technologies | ArcticZymes Technologies ASA. https://arcticzymes.com/.
11 ELIXIR – Value of open bioinformatics resources to industry and SMEs – YouTube. https://youtu.be/YWYKVzxxeZQ

To get a better understanding of ecosystems with high innovation potential, we take a closer look at Cambridge (UK), Berlin (Germany), and Barcelona (Spain).

The majority of companies in the domain of bioinformatics have been founded in western Europe. Cambridge sums 27 companies, while Berlin hosts 21 of them, closely followed by Barcelona, with 19 companies. Our deep dive into the European innovation ecosystems with the highest concentration of bioinformatics companies suggests that strong existing science infrastructures and past successes in growing ventures in other domains have contributed to the concentration and growth of ventures in bioinformatics.

A critical aspect of our analysis is the capacity of the European innovation ecosystems for attracting science entrepreneurs.

Deep Dive: Cambridge

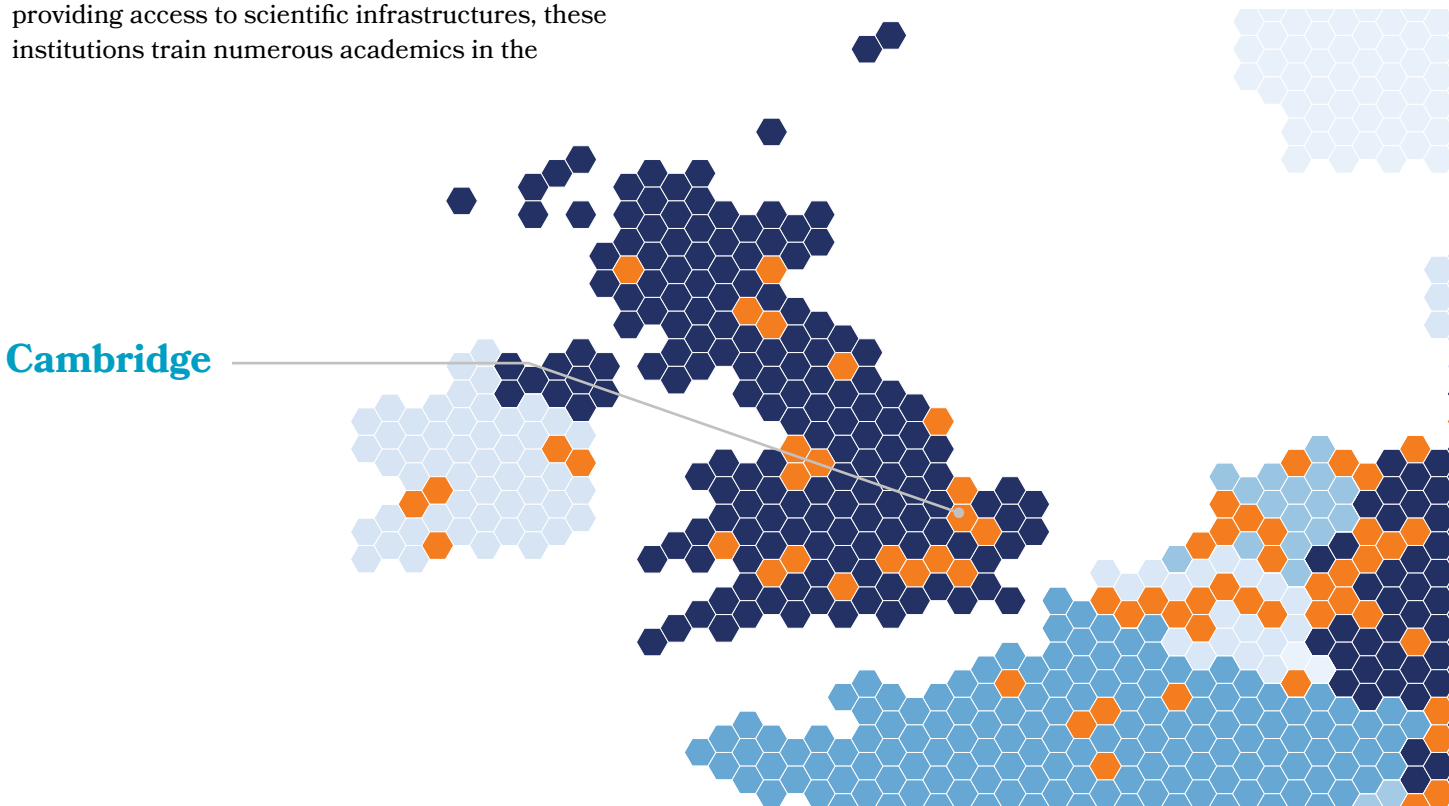
The Cambridge region is the most successful ecosystem for launching and growing bioinformatics companies in Europe.

The region is the fourth most populous in the UK, contributing 8.4% of the UK's Gross Domestic Product (GDP). Cambridge is one of the fastest-growing areas in the UK, both in terms of population and economy. The Science & Innovation Audit for the East of England has identified four key technology themes for the region: Life Sciences, Agricultural Technology, Advanced Materials and Manufacturing, and Information and Communications Technology. According to the '2019 Regional Innovation Scoreboard', the East of England is an innovation leader, where innovation performance has increased over time. The region outperforms national and international peers in business sector R&D spending – East of England records a performance of 58% above the national average and 44% above the EU average.

Cambridge has a strong research heritage with world-class research facilities and capabilities at the University of Cambridge, attracting many multinational companies and many tech startups. In addition to the University of Cambridge (23,247 students in 2019 and 7,913 researchers in 2019), the ecosystem consists of research institutes with scientific links to bioinformatics and related fields, including the European Bioinformatics Institute, the MCR Laboratory of Molecular Biology, Cancer Research UK or the Sanger Institute. Beyond providing access to scientific infrastructures, these institutions train numerous academics in the region. Multiple ventures, such as *Eagle Genomics*, have been founded by former members of research groups from local academic organisations. Growing ventures can settle in more than eight science parks. In order to attain guidance and financial capital, new ventures profit from knowledge and local capital that have been reinvested into the ecosystem from past successes in raising unicorn ventures in the area. Finally, the ecosystem has an established structure of existing general life science companies (seven publicly traded companies and 345 SMEs)[12].

Among many ventures in the domain, *Cambridge Cancer Genomics* stands out for its use of open data to train machine learning models. The venture provides precision oncology solutions that detect relapse of patients earlier than the standard, predict responses to cancer therapy and reduce ineffective treatment regimens. *Sano Genetics* complements public data infrastructures by providing researchers with access to a platform of high-quality genomic and medical data from engaged cohorts consisting of individuals, patient groups and biobanks. The genomics research platform has data privacy and transparency at its core, while giving data sponsors and cohorts the ability to access and manage their own data.

12 Regional Innovation Scoreboard, East of England – Internal Market, Industry, Entrepreneurship and SMEs – European Commission. https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/east-england 9 (visited in 2019)



Deep Dive: Berlin

Berlin is both the capital and the largest city in Germany, with more than 3.6M inhabitants at the end of 2019. While the German industry primarily resides in southern parts of the country, many German and international companies are headquartered in Berlin. The region has one of the highest proportions of SMEs: 181,581 SMEs compared to 633 companies with more than 250 employees[14].

According to the 'Regional Innovation Scoreboard 2019', Berlin is characterised as 'Innovation Leader +', and has been so classified since 2012.

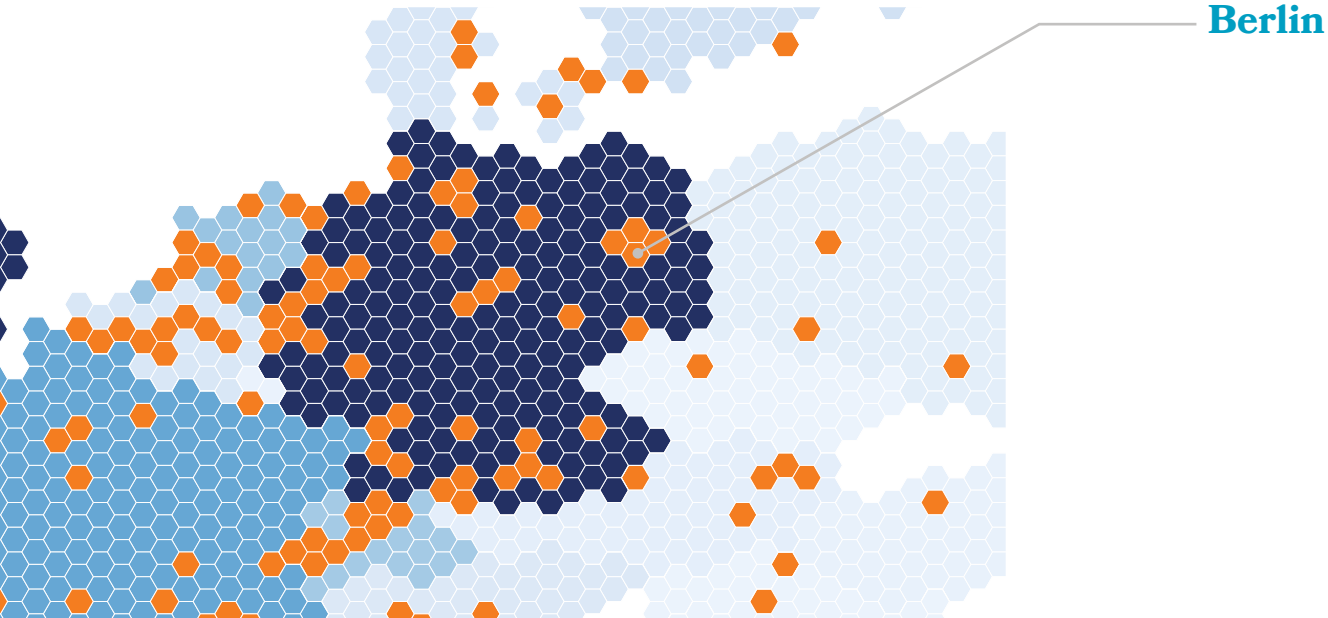
In recent years, the city has become a well-known centre for technology-oriented ventures, particularly in e-commerce, financial technologies, data sciences and analytics. The region provides broad access to scientific insights and highly-trained talent at 11 public universities, including Europe's largest university hospital, *Charité University Medicine Hospital*, and more than 40 non-university research institutions, at least 17 of which focus on the life sciences.

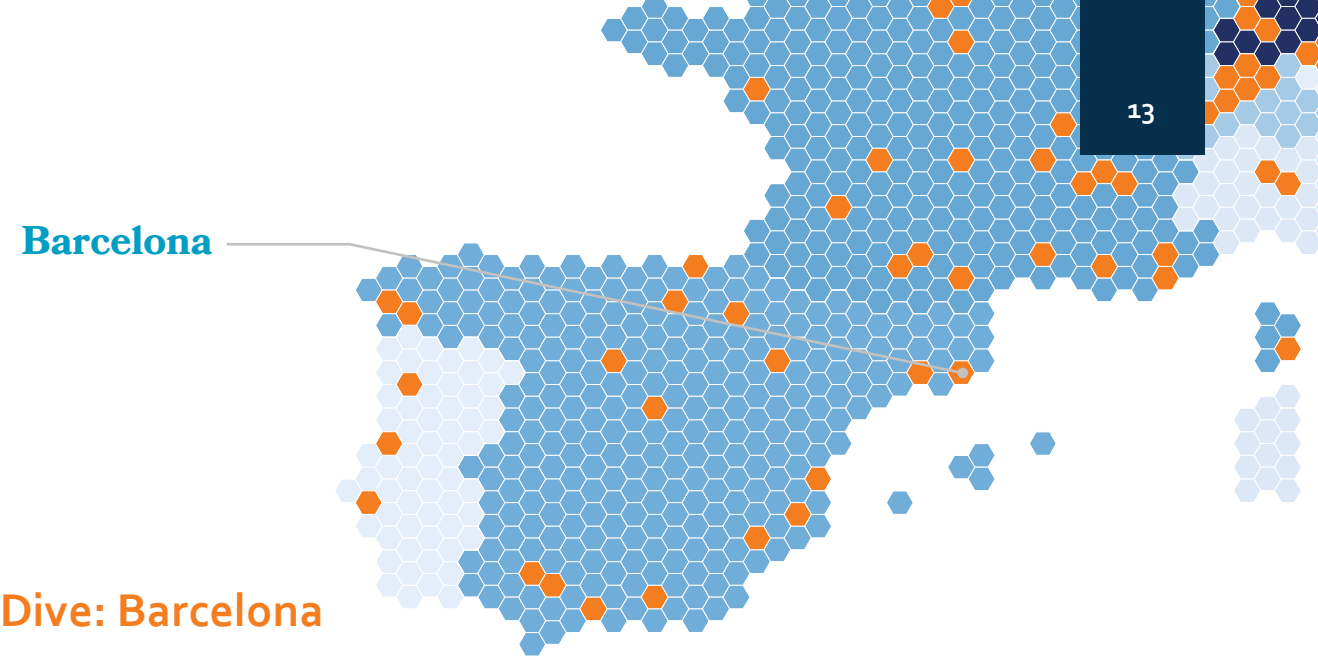
In detail, Berlin's particular strengths lie in public sector R&D spending at 1.5 times of the German and 1.8 times of the European average and non-R&D innovation spending. This is reflected in excellent public-private co-publications, at 137% of the German and 169% of the European average[13].

Ventures in the domain typically reside in one of nine current science parks, with many more under development at the time of this report. In comparison to Cambridge, the ecosystem has seen fewer companies with large scale exits in the past. Thus even though more and more capital flows into the region by venture capital funds, early-stage and growth investment are heavily complemented by public investors. This being said, early-stage venture rounds are larger than the global average[14].

Finally, the capital region also provides established structures for life science companies (34 large life science companies, 255 biotech companies, and 332 medical technology companies)[15]. Among the bioinformatics ventures of the German capital region, we find *Genetek Biopharma* and *Biotx.ai*. *Genetek Biopharma* manufactures genetic diagnostic kits based on QF-PCR technique – for detection of genetic diseases and chromosomal anomalies and human DNA profiling kits used in forensics and kinship issues. *Biotx.ai* delivers predictive biomarkers as a basis for decision support in precision medicine. The venture applies machine learning techniques to automatically select hypotheses from life science data and eventually diagnose previously undiagnosed patients using polygenic analysis.

13 Regional Innovation Scoreboard, Berlin – Internal Market, Industry, Entrepreneurship and SMEs – European Commission. https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/berlin (visited in 2019).
14 Startup Genome – Berlin. https://startupgenome.com/ecosystems/berlin.
15 Life Sciences Report 2019/2020 – Berlin Partner for Business and Technology https://www.berlin-partner.de/fileadmin/user_upload/01_chefredaktion/02_pdf/publikationen/Life-Sciences-Report_en.pdf (2020).





Deep Dive: Barcelona

In 2019, Catalonia had about 7.5M residents, making it the second-most populous region in Spain, accounting for 16.1% of the total population[16]. At that time, Catalonia reached a GDP of €228.68 billion.

The Barcelona ecosystem represents a dense and industrial community of SMEs with an active presence of large multinationals, especially in the biomedical, agro-food, automotive and telecommunications sectors. Barcelona is a growing ecosystem for new ventures, with particular strengths in gaming and life sciences, the majority of which in the digital health domain[17].

It has a long-standing research tradition in the field of life sciences and all major research areas. 12 universities and more than 40 research institutes, with at least 15 focusing on the life sciences, provide scientific research and training. According to the 'Regional Innovation Scoreboard 2019', Catalonia is classified as a 'Moderate+ Innovator' that exceeds innovation performance compared with the rest of Spain and is continuously growing. Ventures start and grow in more than 14 science parks.

While lagging behind Cambridge and Berlin in overall early-stage capital being invested in the region, successful funding rounds are similar in size to the global average[17]. With more than 280 biotechnology, 180 medical technology, and 125 pharma companies, the region provides established structures for new ventures in the life sciences.

Among the many bioinformatics ventures that have launched in Barcelona are *Seqera Labs* and *Minoryx*. *Seqera Labs* provides an open-source and cloud-based high-performance computing solution that helps researchers analyse large biomedical datasets and manage their analysis workflows. *Minoryx* focuses on the discovery and development of novel therapies for severe, orphan genetic diseases of the central nervous system, a group of rare diseases often characterised by neurodegeneration and no or few treatments available.

> *'BITAC's decision to locate its headquarters in Barcelona was inspired by the city's strong entrepreneurial and digital health scene and the first-class scientific infrastructures of Catalonia.'*
>
> **Mireia Rodriguez-Naque**
> BITAC, Barcelona

16 Region Innovation Scoreboard, Catalonia – Internal Market, Industry, Entrepreneurship and SMEs – European Commission. https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/catalonia (visited in 2019).
17 Startup Genome – Barcelona. https://startupgenome.com/ecosystems/barcelona

Mobility of science entrepreneurs for bioinformatics venture creation

Building an efficient, highly talented founder team is a critical factor for forming successful ventures.

Innovation ecosystems that exhibit a healthy innovation performance attract outstanding talents, such as founders. At the same time, thriving innovation ecosystems increase the attractiveness of an ecosystem, leading to increased mobility of exceptional talents to work and study there.

We assess the mobility of foreign and local bioinformatics founders of innovation ecosystems. Comparing the number of founders that studied in a country with the number of those who started a venture in the same country, we obtain an entrepreneur export ratio.

If the ratio is below one, we consider the ecosystem an exporter of founders, while a ratio above one indicates an importer.

Exploring the entrepreneur export ratio, we consider that countries with moderate and neutral ratios – close to one – are 'sticky'. These countries invest in training bioinformaticians who stay and establish a venture. Thus, retaining valuable talent.

While founding a venture is just one way to create economic value amongst many others, creating a bioinformatics venture stands out for its ability to generate new jobs in a specialised domain.

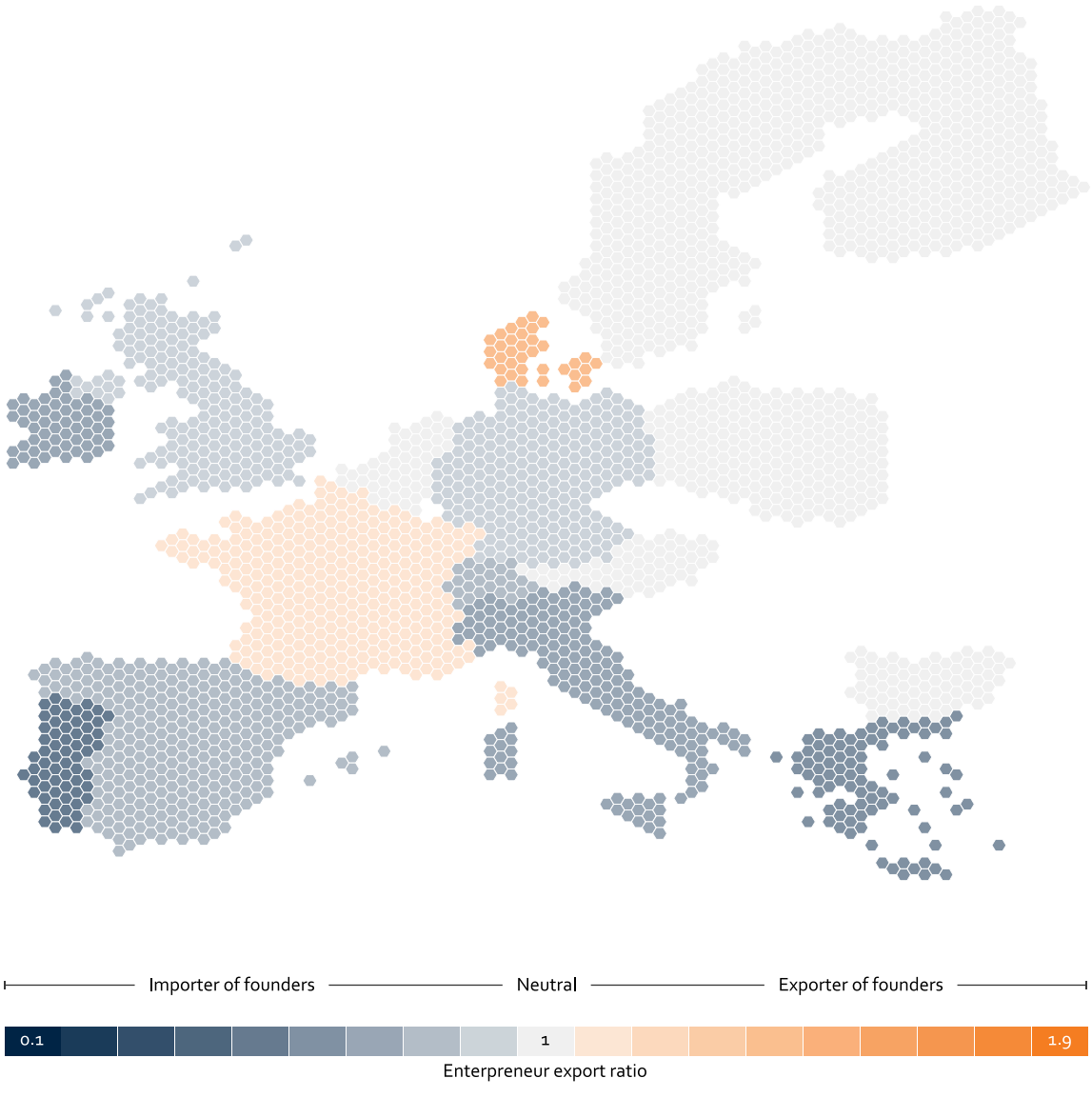


Figure 2. Some countries in Europe excel in being 'sticky'. These countries show moderate and neutral entrepreneur export ratios – close to one –, therefore excelling at retaining talent. (only shown countries where data were available)

Creating customer value from open data

Ventures that operate in the bioinformatics sector can either cater for many different life science application areas or be highly specialised and focused. We have surveyed a range of companies in this sector to understand how they operate, which public resources they use in their business and how they contribute to the economy.

Our survey of over 50 bioinformatics SMEs has indicated that all surveyed companies use openly available resources, such as data, software or standards at varying degrees to create value for clients. Figure 3 gives a general overview of firms operating in this industry segment. Business input can consist of open data, such as publications (e.g. from Europe PMC), protein sequence information (e.g. from UniProt), or genome sequences (e.g. from the European Genome-Phenome Archive (EGA)), as well as open software or algorithmic workflows.

76% stated that without data shared on open repositories, they would not be able to offer their product or service.

89% stated that a product or service has more features because of access to shared or open repositories.

63% stated that without access to registries, ontologies, and dictionaries published on open repositories, they would not be able to offer their product or service.

92% stated that a product or service has more features because of access to registries, ontologies, and dictionaries shared on open repositories.

68% of respondents indicated that large scale data analysis lies at the core of their services, and 62% considered access to open software to be critical for doing their business efficiently. Our survey confirmed the importance of open resources for bioinformatics companies to improve existing or develop new products and services.

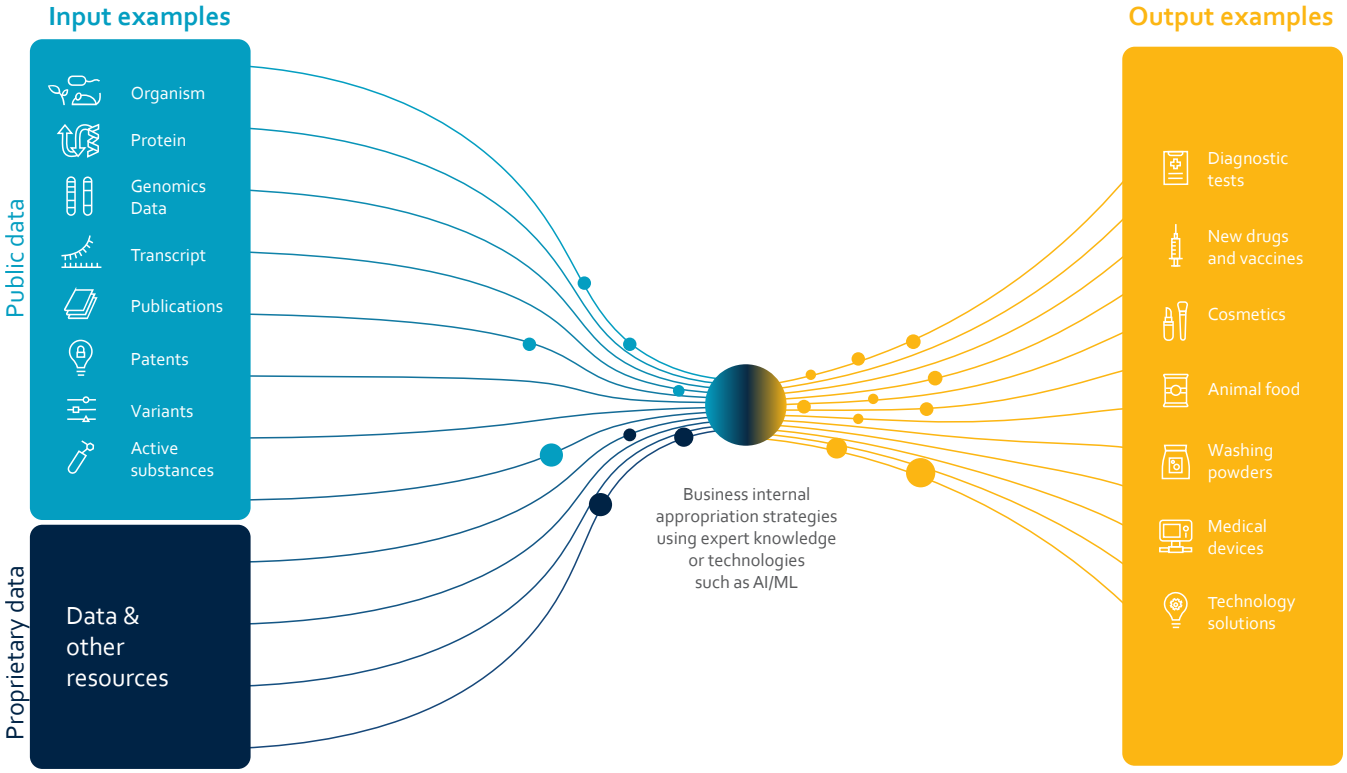


Figure 3. Bioinformatics companies often combine open and proprietary resources to generate a product or service.

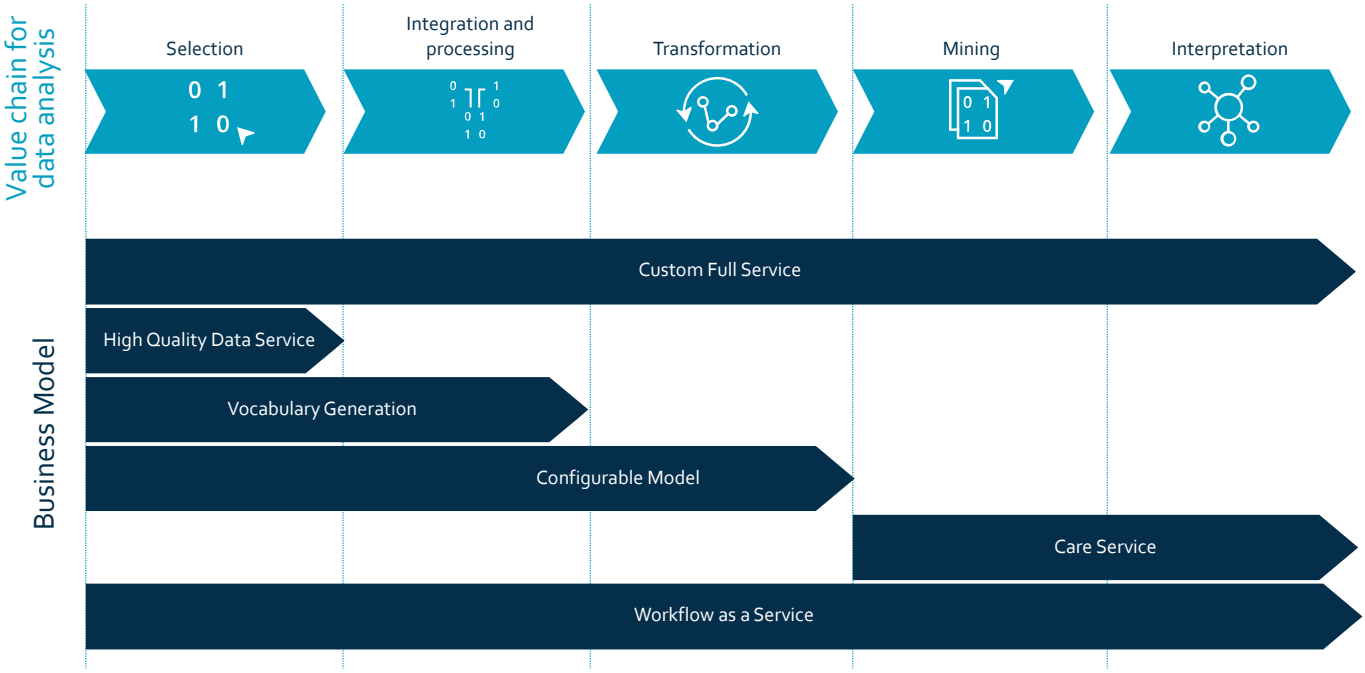


Figure 4. Appropriation strategies of companies along the data analysis value chain

These openly available resources are mostly combined with proprietary resources created by bioinformatics ventures themselves or their clients in the private sector. Outputs of those companies can vary considerably and range from drug discovery or diagnostic tests, to shaping new technology solutions or even domestic products such as washing powders.

With more open and proprietary life science data resources available, the bioinformatics industry has brought forth companies on multiple stages of the value chain. Figure 4 provides an overview of the value chain of businesses making use of open life science resources:

1. Collecting, selecting, and providing data resources.
2. Integrating data and pre-processing these resources for later use either by the company itself or its customers.
3. Transforming data into models such as vocabularies or ontologies.
4. Enabling customers to create their own insights from these models through mining.
5. Providing interpretations from the data on behalf of a client, for example in the form of reports.

Depending on what stage(s) a company operates along the value chain, it can employ different appropriation strategies[18]. A firm can either focus on one or more core activities or include the whole value chain in its business strategy. Particularly at early stages, our survey has shown that bioinformatics ventures tend to provide a full-service model for clients on a custom and hardly scalable basis. The majority of mature ventures rely on technology to either provide a product at a particular stage, for instance, to provide access to curated repositories of open and proprietary data, or generate tools for creating ontologies that will integrate new data resources and answer specific research questions.

To provide a scalable product along the entire value chain, some companies created platforms and private infrastructures that help customise otherwise automated workflows to process raw data into visualisations and insights.

18 Rothe, H., Jarvenpaa, S. & Penninger, A. How do entrepreneurial firms appropriate value in bio data infrastructures: an exploratory qualitative study. *Res. Pap.* (2019).

scailyte®

Core activities

Founded in 2017, Swiss company *Scailyte* uses single-cell data and artificial intelligence (AI) to identify novel disease biomarkers. Unlike methods that capture average measurements of bulk populations of cells, single-cell technologies generate data at the level of individual cells, offering unparalleled insight into biological diversity between cells. Single-cell data contains biomedically relevant information not captured by other technologies but requires innovative computational approaches to extract this information.

At *Scailyte*, single-cell data is fed into their proprietary AI developed pipeline (ScaiVision) to associate patterns in single-cell data with disease status.

Using this approach, *Scailyte* can identify molecular profiles (biomarkers) from specific cell types associated with a disease. These biomarkers are of value for developing early diagnostic tools and may represent potential therapeutic targets.

The company has raised about €5.5M. The venture relies on highly specialised knowledge to make effective use of data. Of 17 team members, 65% hold a PhD in the life sciences, and a third have a computer science or bioinformatics background.

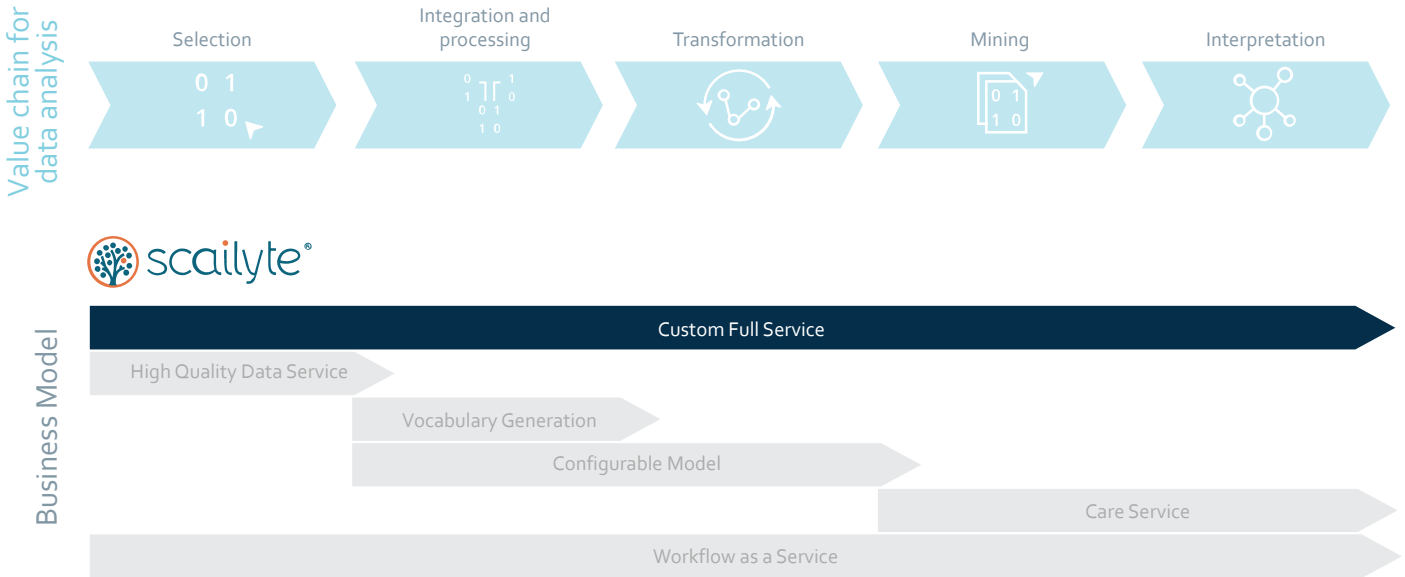
How do they use open data?

While collecting proprietary data themselves and with research groups, *Scailyte* takes relevant single-cell data from the open domain for their analysis.

Open data resources, including annotations, genomes and transcriptomes, play an important role in providing a baseline for single-cell data processing. During this data processing stage, *Scailyte* uses several community-developed tools.

Position along the value chain

Scailyte operates across the entire length of the value chain – performing data selection, pre-processing and analysis, and data interpretation to identify profitable novel insights. For *Scailyte*, custom data inputs are dependent on the customer.



Core activities

Founded in 2012, UK-company *SciBite* specialises in ontology management and text analytics. Their software tackles the challenge to extract, analyse and integrate, knowledge locked within unstructured biomedical text to release its full potential.

Using an ontology-led approach powered by machine learning, their software can quickly extract scientific terminology from unstructured text and transform it into a rich, semantically annotated machine-readable dataset. Ready-made application programming interfaces (API) integration of data into private infrastructures. The resulting data can be used for downstream processes, such as big data analysis and enterprise-wide search.

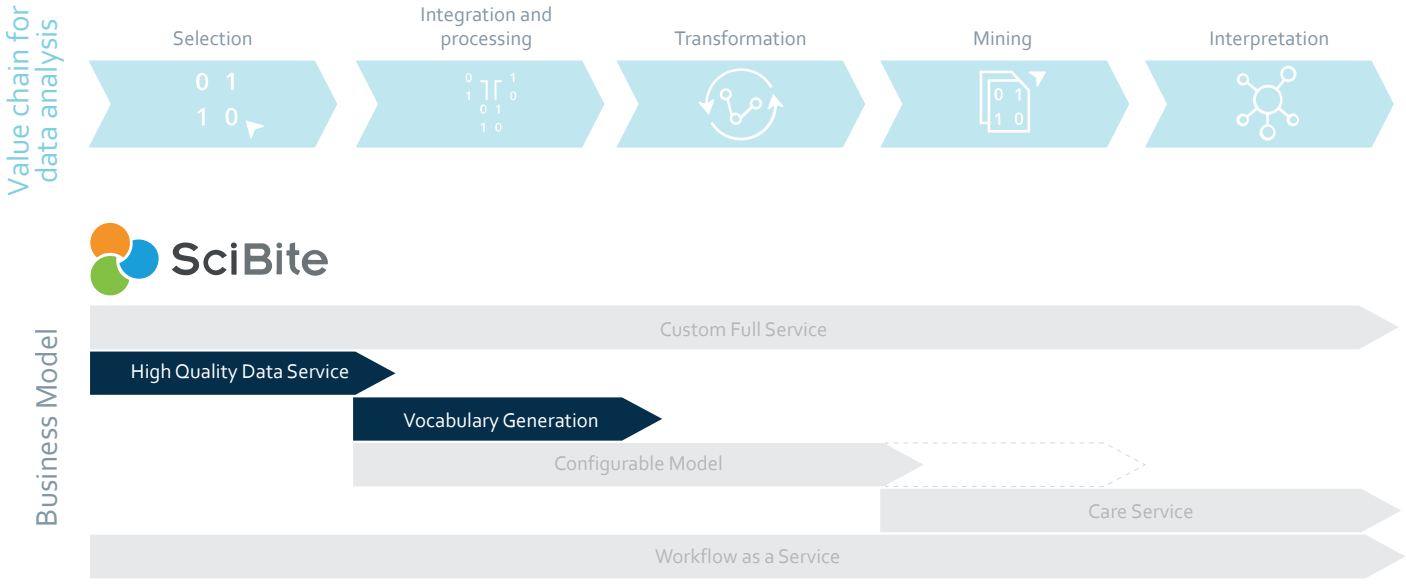
Elsevier acquired the company in 2020 for over €70M. Similar to *Scailyte*, the venture relies on highly specialised knowledge. From 59 team members, the company equally employs life scientists (31%) or computer scientists and bioinformaticians (38%).

How do they use open data?

Publicly available ontologies form one of the key pillars of *SciBite*'s software solutions. *SciBite* enriches these ontologies through manual curation and machine learning to make them more applicable to industry-specific challenges. This enrichment process has been performed for 120 publicly available ontologies. In addition, *SciBite* works closely with ontology communities to feedback issues they may discover as part of their effort to improve these public resources continually.

Position along the value chain

Using their specialised domain knowledge, *SciBite* operates at a specific stage along the value chain – data integration, pre-processing and data transformation. Their suite of APIs can be picked up by large corporations and integrated into existing infrastructures.



Spotlight on open data

Open data resources are essential for today's life science community in both academia and industry. They act as archives of experimental research data, inform future research studies' direction, and help link and bring together a vast wealth of interoperable information.

Many new databases are essential to the broad research community, but only a few have a sustained funding model, as highlighted in a recent study. In an 18-year period, more than 60% of life science research databases ceased to exist, and a further 14% were no longer updated[19].

Research infrastructures such as ELIXIR work to increase the long-term sustainability of their resources and carry out targeted actions so that these feed into innovation and discovery in the data-driven life science sector. Our study highlighted the significance of open data resources, particularly for small and medium-sized life science companies, exemplified by 73% of respondents indicating that data resources with information on rare diseases (e.g. from Orphanet) are important to the business. Repositories holding information on scientific publications present another good example (e.g. from EuropePMC) or human genomics data (e.g. from EGA). 84% and 71% of respondents highlighted the importance of open data their business operations.

Interestingly all companies in the survey indicatezzwas relevant to the business (no individual resource stands out as most important). Figure 6 shows how ventures combine data from different databases.

They combine open data resources, often with proprietary information, to create value for customers. In doing so, ventures combine open data resources, often with proprietary information. To enable new insights to be garnered, or to facilitate a diverse range of customers to gain new knowledge, an assortment of data needs to be connected and analysed. We, therefore, found that companies that grew more rapidly were among those who used many different resources in combination (Figure 5).

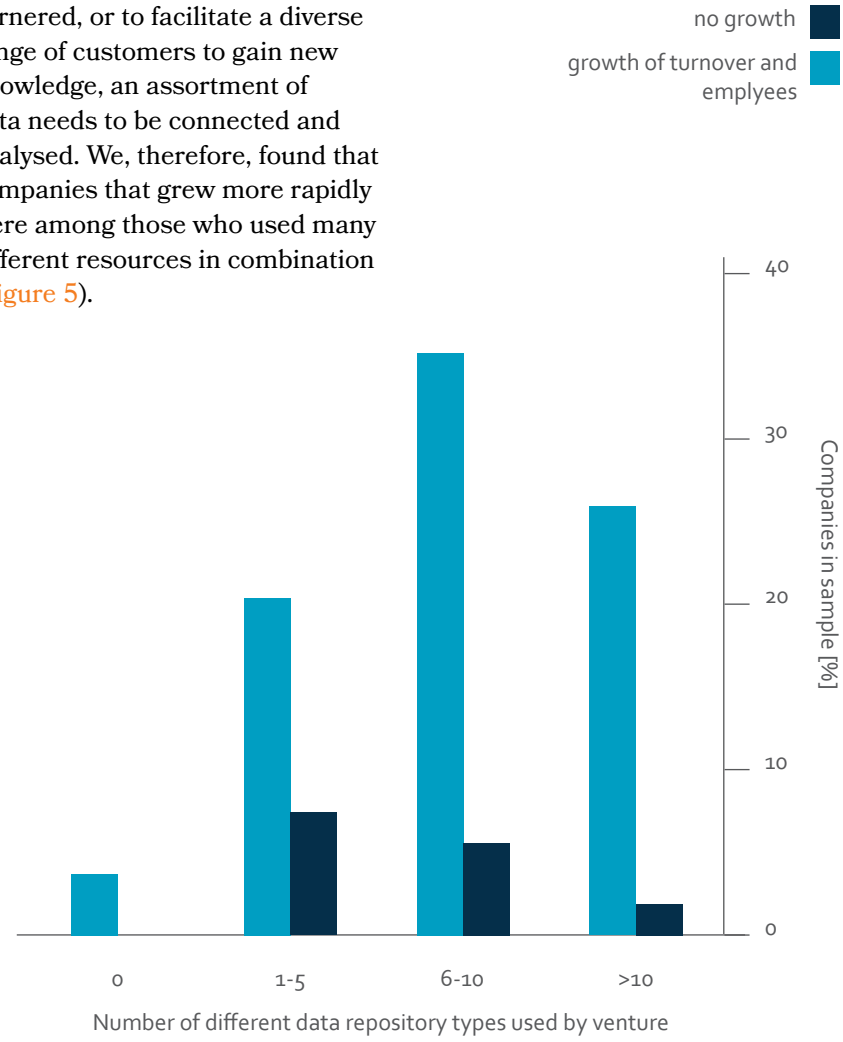


Figure 5. Percentage of companies reporting either growth of turnover and number of employees or no change, related to the number of open data repositories used by a company.

19 zAttwood, T. K., Agit, B. & Ellis, L. B. M. Longevity of Biological Databases. *EMBnet.journal* **21**, (2015).

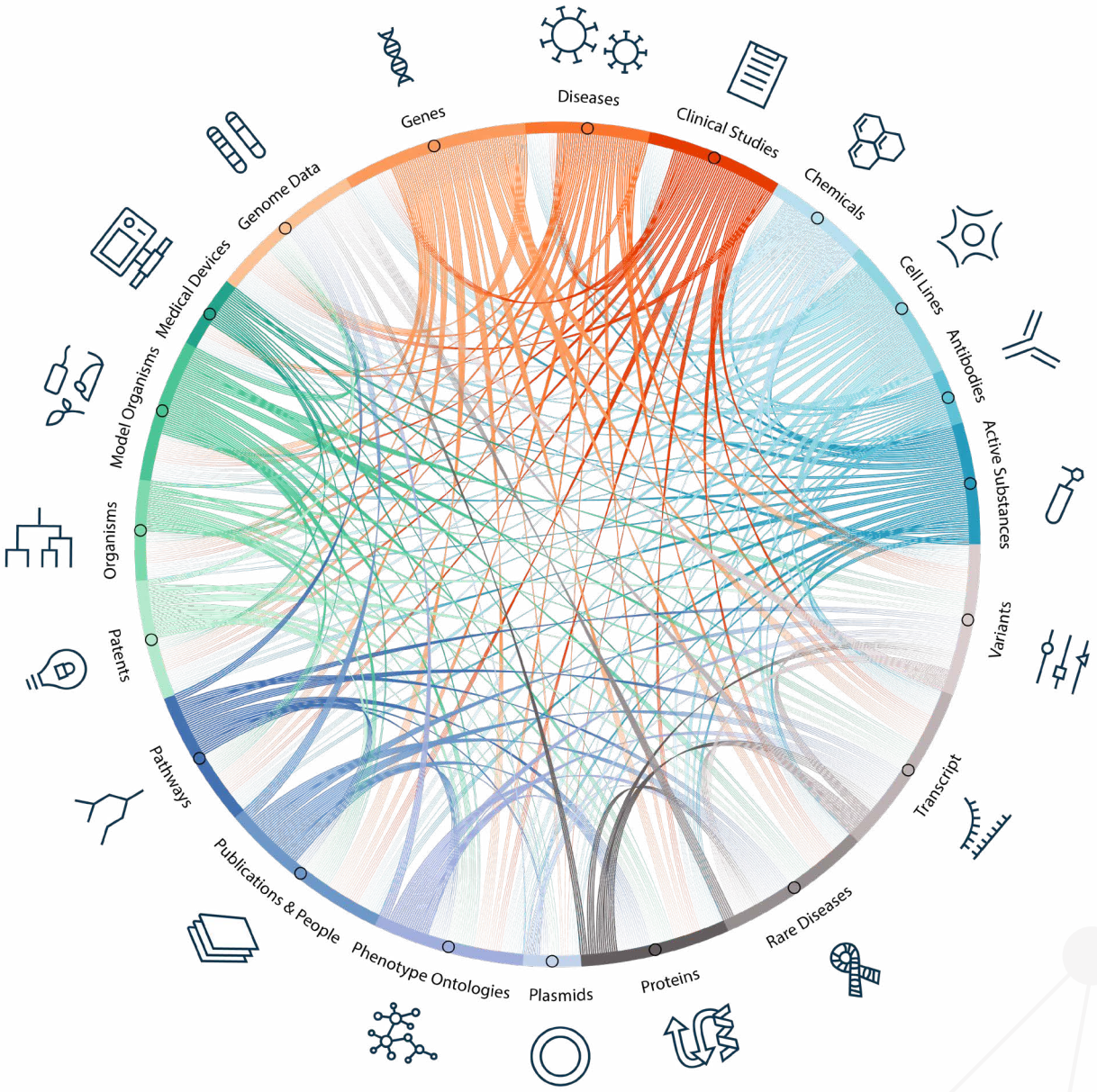


Figure 6. Data resources that are commonly used together by bioinformatics ventures. There is no individual resource that stands out as most important – the value resides in the interconnection and combination of resources.

Spotlight on talent

Who are the founders?

It is indisputable that education, knowledge and skills translate into economic value. Human capital has long been regarded as one of the driving forces of innovation and creating entrepreneurial actions. A deep understanding of science and technology is needed to develop a bioinformatics venture. For that reason, it is vital to understand where talent in Europe is forged. Universities and academic institutes increasingly equip academics with entrepreneurial competencies.

Those entrepreneurs move to set up their ventures wherever they find the best support system and supporting network. We looked at several founders of bioinformatics companies. We investigated how research infrastructures, such as the national Nodes in ELIXIR, are hotbeds of collaboration and networking and have a strong focus on training that have influenced their career paths.

Abel
Ureta-Vidal

Born in Spain

I am a Spanish-born, French-educated and British-inspired entrepreneur – that gives a very short overview of my career over the past 25 years.

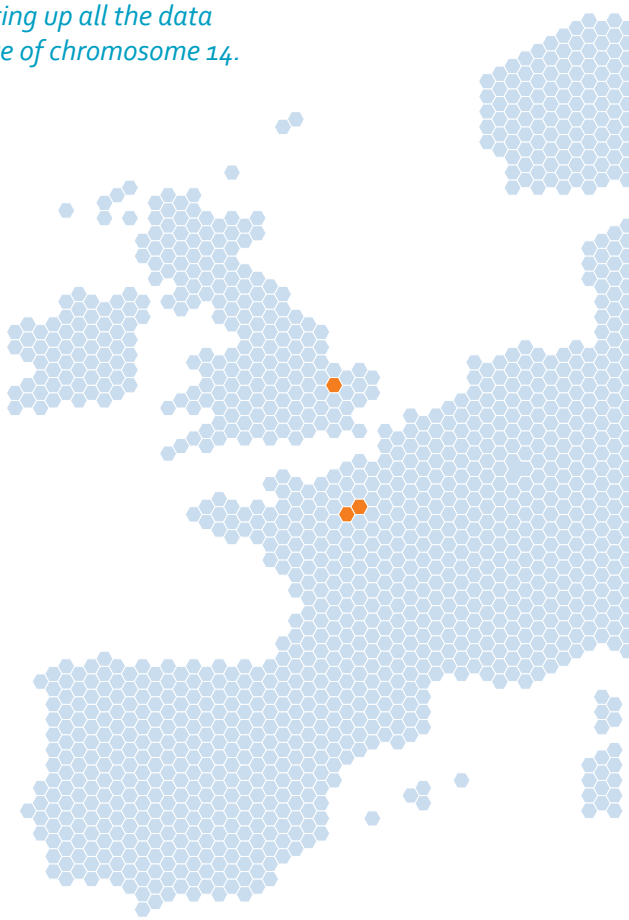
1999
University in France and PhD at Universite de Paris

I was very fortunate to join the Human Genome Project straight after my PhD. I was put in charge of setting up all the data analysis to predict genes on the sequence of chromosome 14.

1999-2001
Genoscope – French contribution to the human genome project

For a French man, it was a dream job, i.e. a job for life in an academic institution, but I wanted a bit more international exposure and came to Cambridge and joined the Ensembl project at EMBL-EBI.

2001-2007
EMBL-EBI – Project leader at Ensembl (Cambridge)



2007-2008
MBA (Judge Business School Cambridge)

I was analysing genomes to find particular genes that determine if you develop a disease with a small team of about five or six people. In 2007 I felt a bit too comfortable there, and I decided to leave EMBL-EBI to learn about business by studying for a MBA at the Judge Business School.

2008-2020
Founder of Eagle Genomics

During that time, I was exposed to the entrepreneurial path and all entrepreneurship activities around Cambridge. I then founded Eagle Genomics in May 2008, which took me on a journey for about 12 years – moving from bioinformatics consultancy into developing an enterprise software platform for data and knowledge management.

2018-Now
Member of the ELIXIR Industry Advisory Board (Chair since 2020)

2020-Now
Investment Director at CMS Ventures, Co-founder & Chief Data Officer of Sofi Health

Maria
Chatzou Dunford

Born in Greece

2008
BSc in Computer Science and Biomedical Informatics University of Thessaly (Volos)

2010
MSc in Bioinformatics (University of Athens/University of Stockholm)

2010-2011
Bioinformatics Engineer

2014
Founding member of the grass-roots network Innovation Forum

2016
PhD in Bioinformatics co-developed Nextflow (University Pompeu Fabra, Barcelona)

2016-2017
Postdoctoral researcher at CRG

2017
Founder of Lifebit (Cambridge)

I am originally from Greece, but I have spent two-thirds of my life abroad. Every scientist has a similar trajectory – you move for your work.

I have had incredible people helping me throughout my entrepreneurial journey. I am not from a privileged family, my mum is not Marie Curie, and my dad is not Steve Jobs or someone equivalent.

So what is my secret? I choose to focus on opportunities. I would encourage everyone to become part of an organisation, such as ELIXIR or the Innovation Forum, as a hobby to give you a sandbox, where you can get training in business, communication, networking, etc.

I did move to Cambridge when I founded Lifebit. When you are trying to build a deep-tech company, then you need to be very close to the pharmaceutical clients, hospitals, biotech companies because they're our predominant clients. Then, you need to be close to a cluster of big research institutions – take the Golden Triangle between London, Cambridge and Oxford.

About ELIXIR

ELIXIR is a research infrastructure. And just as physical infrastructures, such as roads and bridges, support a flourishing economy, digital infrastructures are the underlying basis for a flourishing bioinformatics industry.

Through 183 current affiliated bioinformatics institutes, ELIXIR offers infrastructure resources including databases, software tools, cloud storage and supercomputers, as well as training and opportunities for knowledge exchange, collaboration, and networking.

ELIXIR aims to make it easier for researchers and entrepreneurs in academia and industry to find and share data, exchange expertise and agree on best practices to move science forward and accomplish breakthrough discoveries.

The vision of the ELIXIR Industry Programme

Through the ELIXIR Industry Programme, ELIXIR actively encourages the creation of an open innovation ecosystem in the life sciences, driven by collaboration between industry and ELIXIR Nodes. In this ecosystem, existing companies increasingly use and integrate public life science resources, leading to new services, economic growth, and new SMEs.

Funders and policymakers will, in turn, acknowledge that providing free bioinformatics resources represents a public good and support innovation and the transition to a knowledge-based economy. Activities from the industry programme include networking events (e.g. the Innovation and SME Forum), funding mechanisms for small scale collaborative projects, or opportunities for knowledge exchange.

Methods

This report has been based on the qualitative-exploratory field study[20] conducted in 2018-2019 to understand the challenges that digital entrepreneurs face to appropriate value from bio-data infrastructures.

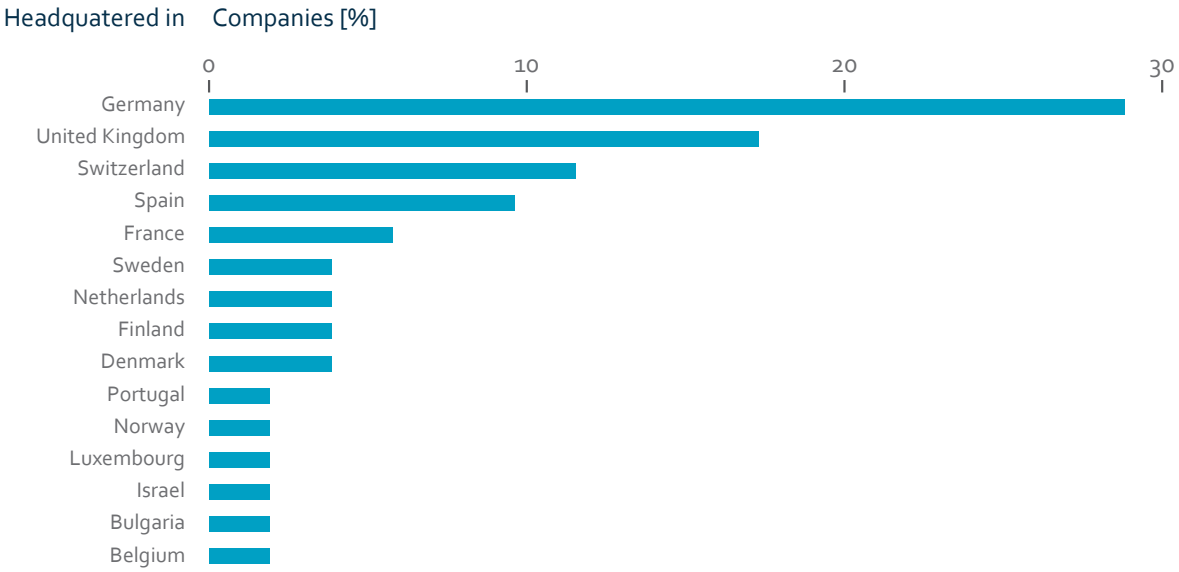
Further, an open questionnaire with 54 participating bioinformatics companies was used to get a deeper understanding of business models, scalability and resources of importance. Databases, such as Crunchbase, Thomson Reuters, Dealroom, Amadeus and LinkedIn, were used to support this report and complement the field studies. The ecosystem analysis presented in this report focused on SMEs that have been founded after 2001, and that have neither been dissolved nor sold or integrated into large corporations.

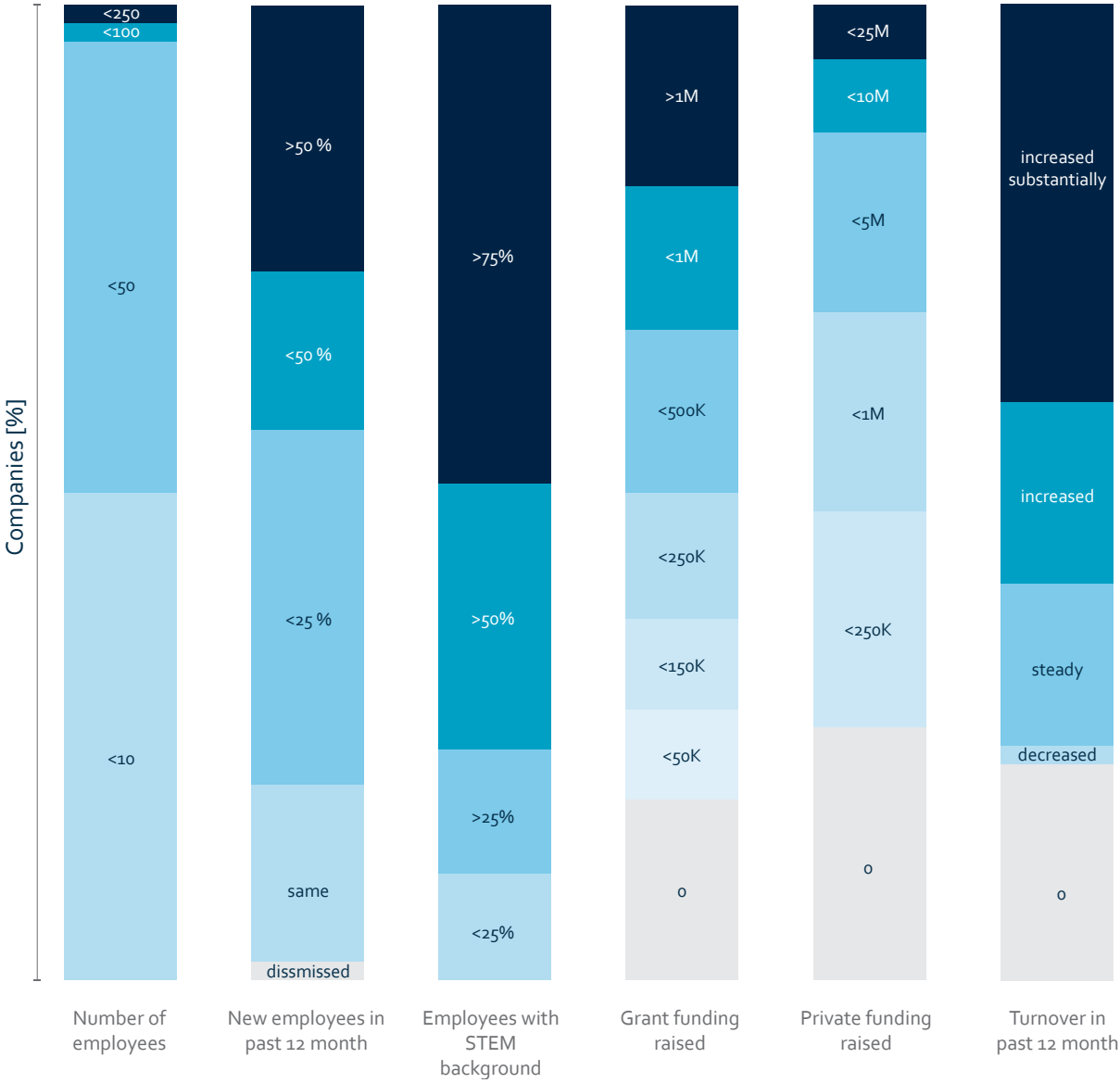
Complementary to our own data and analysis, this report also refers to insights from the 'Regional Innovation Scoreboard (RIS)'[21], which provides a comparative assessment of the regional research and innovation performance of the EU Member States and the relative strengths and weaknesses of their innovation ecosystems.

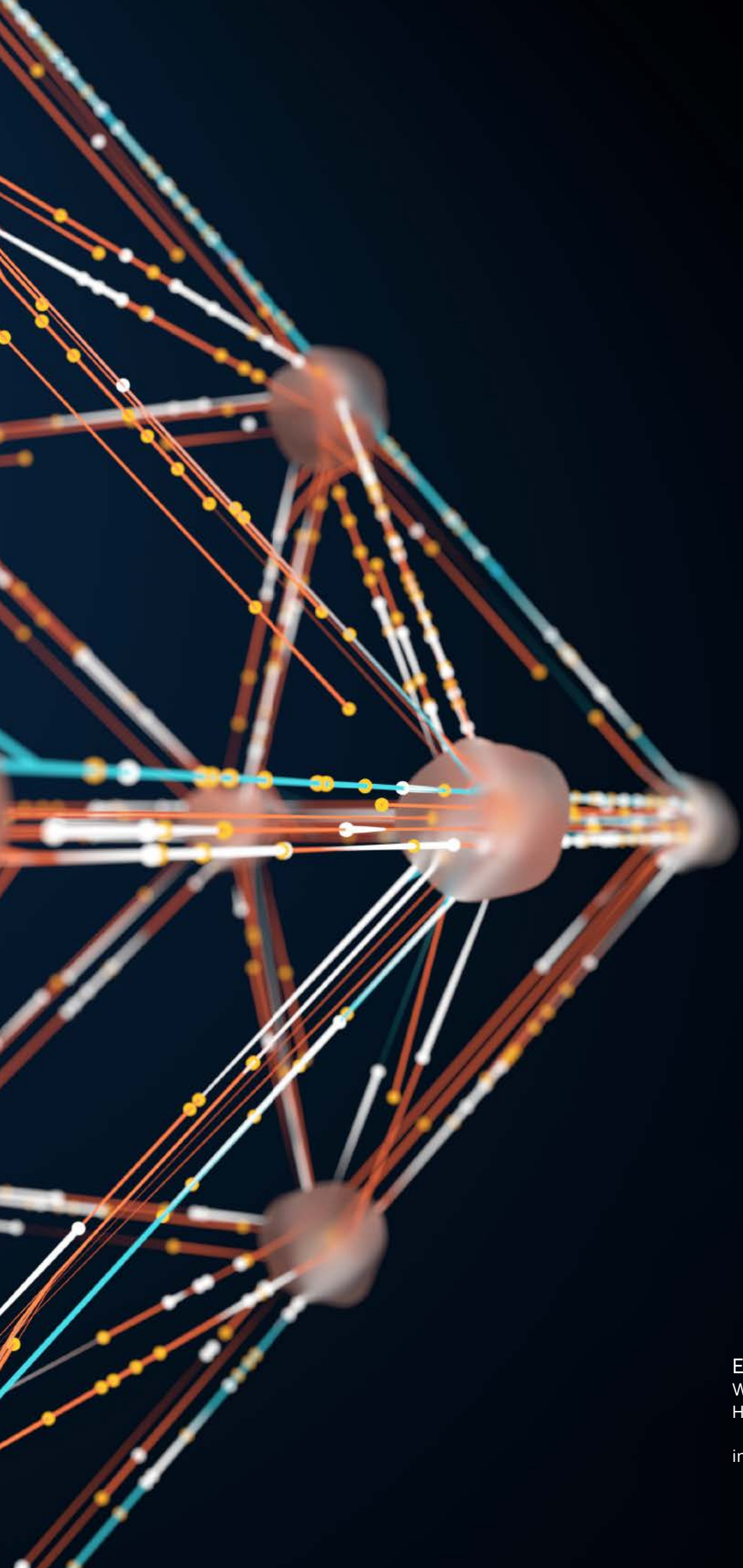
Together, we compared our results to explain the attractiveness of individual ecosystems for bioinformatics companies within each ecosystem.

20 Rothe, H., Jarvenpaa, S. & Penninger, A. How do entrepreneurial firms appropriate value in bio data infrastructures: an exploratory qualitative study. *Res. Pap.* (2019).

21 Regional innovation scoreboard | Internal Market, Industry, Entrepreneurship and SMEs. https://ec.europa.eu/growth/industry/policy/innovation/regional_en.


Headquatered in
Companies [%]
0
10
20
30
Germany
United Kingdom
Switzerland
Spain
France
Sweden
Netherlands
Finland
Denmark
Portugal
Norway
Luxembourg
Israel
Bulgaria
Belgium


Companies [%]
<250
<100
<50
<10
>50 %
<50 %
<25 %
same
dissmissed
>75%
>50%
>25%
<25%
>1M
<1M
<500K
<250K
<150K
<50K
0
<25M
<10M
<5M
<1M
<250K
0
increased substantially
increased
steady
decreased
0
Number of employees
New employees in past 12 month
Employees with STEM background
Grant funding raised
Private funding raised
Turnover in past 12 month



ELIXIR Hub, South Building
Wellcome Genome Campus
Hinxton, Cambridge, CB10 1SD, UK

info@elixir-europe.org